Bridging Gaps in Exploratory Data Analysis using Dimensionality Reduction

by

Aindrila Ghosh

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Organizational transactions generate immense amounts of data every day. The decisions made using such data are not only important for their financial impacts on the business; they also regulate the relationships with other businesses in their supply chain. There has been much research that focuses on facilitating more efficient data-driven decision making. As a result, in the past years, researchers have explored several directions of research that range from business to technical areas, for this purpose. Such directions include, understanding specific business disciplines in order to identify their challenges and gaps in decision making, creating Exploratory Data Analysis (EDA) tools to help with better visual interpretation of data, and producing algorithms that can assist with compressing and summarizing high-dimensional industrial datasets to analyze them using spatial techniques. However, in each of these explored areas there exist many open challenges. For example, despite of their financial importance, data generating processes from many business units, such as the Sales-and-Subscriptions (S&S) renewal, have received limited attention from researchers. Moreover, with the abundance of EDA tools and data compression algorithms analysts often struggle with the selection of the most appropriate solution for their analytical context. Furthermore, the highly technical nature of data summarization techniques makes their evaluation, interpretation, and usage challenging for both novice and expert data analysts. Following an action research method, this research attempts to bridge several gaps in all the above mentioned areas. Firstly, a longitudinal study across multiple organizations is performed, that identifies the state-of-the-art industrial process of data-driven decision making in the business unit of Sales-and-Subscriptions (S&S). The analysis of the business unit shows that, analyzing customers' experiences with the seller organization can help mitigate renewal risks. Hence, in the next part of the research, 50 cutting edge visual EDA tools are investigated for their ability to assist with visually exploring large industrial datasets. Then, the focus is shifted to popular data summarization and visual EDA area of Dimensionality Reduction (DR). More specifically, three different challenges associated with the DR process are addressed namely: selection of the most appropriate algorithm, interpretation of its outcome, and evaluation of the quality of the reduced dimensions. In order to achieve the research goals, at first a large-scale experimental study is performed, where 15 of the most popular DR techniques are statistically analyzed and the first ever practitioners' guideline for selecting DR algorithms in a given analytical context, is created. Next, two novel algorithms namely Local Approximation of Preserved Structure (LAPS) and Global Approximation of Projection Space (GAPS) are presented that help with the interpretation of the structural quality of the outcome of any DR technique. Finally, to enable a user driven evaluation of DR methods, a visual interactive toolkit namely: Visual Explanations of Preserved Structure (VisExPreS) is presented with Proactively Guided LAPS and GAPS. The value and novelty of the presented solutions are demonstrated using extensive evaluations throughout the thesis.

Preface

This thesis is an original work by Aindrila Ghosh submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Software Engineering & Intelligent Systems. The thesis is written in paper-based format with the details as follows:

Chapter 2 of this thesis contains the article: Aindrila Ghosh, Mona Nashaat, and James Miller, "The Current State of Software License Renewals in the IT Industry", In Information and Software Technology, Elsevier, Volume 108, April 2019, Pages: 139-152, DOI: 10.1016/j.infsof.2019.01.001

Chapter 3 of this thesis contains the article: Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets", In Visual Informatics, Elsevier, Volume 2, Issue 4, Pages 235-253, December 2018, DOI: 10.1016/j.visinf.2018.12.004.

Chapter 4 of this thesis contains an extended version of the article: Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, "Context Based Evaluation of Dimensionality Reduction Algorithms – Experiments and Statistical Significance Analysis", accepted at the ACM Transactions on Knowledge Discovery from Data.

Chapter 5 of this thesis contains the article: Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, "Interpretation of Structural Preservation in Low-dimensional Embeddings", In IEEE Transactions on Knowledge and Data Engineering (Early Access), DOI: 10.1109/TKDE.2020.3005878

Chapter 6 of this thesis contains the article: Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, "VisExPreS: A Visual Interactive Framework for User-driven Evaluations of Embeddings", Under review at the IEEE Transactions on Visualization and Computer Graphics.

This thesis is dedicated to my husband, Soumalya, without whose endless support and encouragement getting a Ph.D. would have been just another dream.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my supervisor Prof. James Miller for his patience, motivation, guidance and endless support. His constant encouragement and immense knowledge helped me throughout the years of my Ph.D. research. I could not have hoped for a better advisor and teacher.

Secondly, I would like to thank the members of my examining and supervisory committee: Dr. Kamran Sedig, Dr. Marek Reformat, Dr. Scott Dick, Dr. Cor-Paul Bezemer, and Dr. Hamzeh Khazaei for their thoughtful comments and constructive feedbacks during my doctoral candidacy and final exams.

Further, I would like to thank my friend and colleague Ms. Mona Nashaat. Since the day we started our doctoral journey together, her continued support and motivation were the things I could always count on. I am glad that I met such a wonderful person at the UofA and I wish her all the success she deserves. I would also like to thank our industrial liaison Mr. Shaikh Quader from IBM. It has been an immense pleasure being a part of this team and to have worked alongside. Moreover, I would like to express my sincere gratitude to Ms. Pinder Bains for all her help and support during the entire period of my graduate studies at the UofA.

Finally, I would like to thank my family, especially, my husband Mr. Soumalya Ghosh. None of this would have been possible without his constant encouragement and support. I cannot thank him enough for his patience and care during all the frustrations and difficulties that came our way during the last four years. I would also like to thank my father Mr. Pradip Kumar Basak, my parents-in-law Mr. Tapan Kumar Ghosh and Mrs. Gopa Ghosh, and my brother Nirmalya for their love, support, and understanding.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As a detailed list of articles is provided in the preface, this section elaborates on the motivation and a brief overview of this research. Moreover, this section also summarizes the primary contributions of the overall thesis followed by a discussion on the organization of the thesis documentation.

1.1 Motivation

In today's data-centric world, organizational transactions generate immensely large amounts of data every day. The insights obtained from analyzing this data is often used by organizations to make important business decisions. These data-driven decisions are not only important for their financial impacts on businesses, but also for the way they regulate the relationships of organizations with other businesses in their supply chain. Hence, conducting efficient preprocessing, analysis, and summarization of such real-world industrial data is of paramount importance. However, the challenges of storing, managing, and analyzing these large high-dimensional [1] datasets are well-known [2] in both industry and academia. In order to avert these challenges as well as to manage the expenses of analyzing and processing this large amount of data, organizations are trying to make efficient choices. For example, whilst different disparate data sources are being synchronized to collect more insight from the data, automated data analysis pipelines are being incorporated to pre-process and examine the data. Additionally, members of different business units (e.g., the sales and marketing teams) are being trained to perform various data analysis tasks as a part of their day-to-day decision making. Nevertheless, an in-depth analysis of high-dimensional datasets often involves [3] complex data manipulation (e.g., data transformation, feature extraction) techniques that require significant knowledge and expertise from the analyst. In these situations, any misguided decisions made by automated data-analysis pipelines or non-expert data analysts can be catastrophic for an organization. As a result, the trade-off between cost and accuracy of analyzing high-dimensional data still remains an open [3] challenge.

In order to facilitate efficient data-driven decision making, over the years both academic and industrials [3]–[10] have explored several directions of research. As some of them [4], [5] have examined different organizational units and have suggested strategies for better decision making, others [6]–[8] have presented Exploratory Data Analysis (EDA) tools that help with summarizing and analyzing real-world data. On the other hand, in order to enable a more in-depth analysis of high-dimensional data, some researchers [3], [9], [10] have proposed data compression techniques (e.g., dimensionality reduction). These techniques attempt to represent high-dimensional datasets using lower dimensions (i.e., smaller sized feature vectors) while retaining as much of the original information as possible. Nevertheless, each of these research directions has its own gaps and limitations. For example, whereas there exists several business units [4], [5] that have never been studied, most existing EDA tools [11]–[13] lack in scalability and support for complex data analysis tasks (e.g., analysis of multivariate relationships). On the other hand, although facilitating in-depth data analysis data compression techniques remain highly mathematical and black-box, making their selection, interpretation, and evaluation challenging [14]–[16].

1.2 Research Overview

Following an action research method, this thesis explores the different research areas that enable efficient data-driven decision making. At first, the organizational unit of Sales-and-Subscriptions (S&S) renewal is studied and its limitations and risk areas are identified. Next, a detailed survey of 50 academic and industrial EDA tools is performed to identify their gaps and opportunities for improvements. Finally, focusing on the black-box data compression technique of Dimensionality Reduction (DR in short) at first, a detailed statistical analysis and a practitioners' guideline is presented for an efficient selection of DR algorithms, followed by two novel algorithms, their enhanced versions, and a visual interactive toolkit that assist with interpretation and user-driven evaluation of DR techniques.

There has been much research [4], [5] that focuses on understanding specific business disciplines (e.g., customer relationship management, marketing strategies) and resolving their challenges. However, some business processes have received limited attention from researchers, despite their financial importance. One such sector is Sales-and-Subscriptions (S&S) renewal teams. S&S teams are responsible to make sure that (sold) software licenses of any I.T. organization are renewed on time. Hence, they are equally important for the organization's revenue as the sales or marketing teams. One of the biggest challenges of such renewal management teams is to make informed decisions about the upcoming renewals. Despite incorporating several human and technological resources in these teams, due to the lack of research in identification of the pain-points of these teams, some process-related uncertainties remain open. In order to bridge this gap, in this research, at first the results of the ***longitudinal study across multiple organizations*** is presented that identify the "state-of-the-art industrial process" of software license renewal and the challenges associated with it. In order to assist with mitigation of these challenges and to help renewal teams to analyze customer data and make more informed business decisions, next, at existing solutions that enable visual Exploratory Data Analysis (EDA) [17] techniques are explored for obtaining detailed insights from large industrial datasets. A ***comprehensive review of 50 visual data analytics tools*** is performed for this purpose and discuss their utilities in each step of the EDA process. From this analysis, some research opportunities are discovered that can help to enhance these tools in order to perform a more detailed multivariate analysis of the data.

The next phase of this research investigates popular data compression method Dimensionality Reduction (DR in short) [3] that are commonly used for big-data analytics in domains [18] such as biochemistry, medicine, and biotechnology. DR algorithms transform high-dimensional data into low-dimensional embeddings while attempting to maximally preserve the structural properties of the input dataset. During this transformation, most DR algorithms attempt to retain [19] the local structure (i.e., the neighbourhood of individual data-points) as well as the overall global structure (i.e., the relative distances between all data-points) of the original data. Despite their popularity, DR techniques come with a set of major caveats. Firstly, in recent years, a plethora of DR techniques have been proposed [3] with their respective parameter combinations that significantly influence the embedding structure. The non-intuitive nature [20] of these parameters hinder the interpretability of these techniques making the selection of the most appropriate DR algorithm challenging. Secondly, the dimensions derived using such techniques lack in a clear-to-interpret mapping [14] with the original attributes in the data. As a result, data analysts with limited experiences with DR are often forced to blindly trust [15] the embeddings without truly understanding the meaning of the projection axes or the positioning of data-points. Finally, the above-mentioned issues lead to the challenges associated with evaluation [16] of DR algorithms.

This research attempts to address the mentioned challenges of the DR algorithms in the same order as discussed above. Given a plethora of dimensionality reduction algorithms and metrics [3] for their quality analysis [16], there is a long-existing need for guidelines on; ***"how to select the most appropriate algorithm in a given scenario?"*** In order to bridge this gap, at first, five analytical contexts for DR are identified and 12 state-of-the-art quality metrics are categorized into those contexts. Furthermore, 15 most popular dimensionality reduction algorithms are assessed on the chosen quality metrics using a systematic experimental study. Later, using a set of robust nonparametric statistical tests [21], the generalizability of the evaluation of the algorithms is assessed using 40 real-world datasets. Finally, based the results a ***practitioners' guideline*** for the selection of an appropriate DR algorithm is presented in the discussed analytical contexts.

Next, the focus is shifted towards the challenges associated with the interpretability [22] of DR algorithms. Interpreting the quality of a low-dimensional embedding is crucial as it enables trust [23] on the transformed data. Here, two novel interactive explanation techniques are proposed for low-dimensional embeddings obtained from any DR algorithm. The first method & data-type agnostic [24] technique ***LAPS (Local Approximation of Preserved Structure)*** provides explanations on the preserved local structure of a low-dimensional embedding that justify the fidelity of the relative positioning [22] of any individual data-point by approximating a neighbourhood [24] locally around that point. The second technique ***GAPS (Global Approximation of Projection Space)*** presents explanations on the preserved global structure in a low dimensional embedding, by combining non-redundant local approximations from a coarse discretization of the projection space [25]. Using a comprehensive evaluation, the proposed techniques are assessed for their flexibility (with 10 DR algorithms on 16 datasets), applicability (with tabular, text, image, and audio data) and reliability.

Finally, focusing on the challenges of evaluating DR algorithms, this research unifies the benefits of both [16], [26] quantitative and qualitative evaluation of DR techniques by presenting an interactive toolkit and visual tool that enables a user-driven quantitative analysis of preserved structures in any embedding. Towards achieving this goal, the enhanced versions of LAPS and GAPS namely ***PG-LAPS (Proactively Guided LAPS)*** and ***PG-GAPS (Proactively Guided GAPS)*** are composed into a visual toolkit [1] named ***VisExPreS (Visual Explanations of Preserved Structure)*** such that, users not only have control over the quality analysis of DR but also can focus on the aspects of the analysis that are the most interesting from their perspective.

1.3 Summary of Contributions

The primary contributions of this research are as follows:

- This research represents the first ever study on the 'state-of-the-art' industrial practice of software license renewal and the challenges & risks associated with it (cf. Section 2.3).

- A comprehensive review of 50 EDA tools presented in this research, is unique in terms of being recent, voluminous, and focused on the utility of the tools in each step of the EDA process (cf. Section 3.2).
- For the first time in academia, this research composes 12 most popular DR quality metrics and categorizes them into the five identified analytical contexts. The metrics are then used to perform a systematic comparison among 15 popular DR algorithms. The results identify the best, mediocre, and worst-performing algorithms in a given analytical context. Furthermore, this novel research performs a thorough statistical significance analysis of the performance of DR algorithms using 40 real-world datasets. Finally, this work presents the first generic guideline for practitioners to select the most appropriate DR algorithms in any scenario (cf. Section 4.4).
- This research presents LAPS, a novel algorithm that provides interpretable and faithful explanations on the retained local structures in any low-dimensional embedding, by locally approximating the neighborhoods. This research also presents GAPS, a novel technique that provides explanations on the preserved global structure of a manifold in its low-dimensional embedding, by combining local approximations of discrete non-redundant neighborhoods into a global approximation (cf. Section 5.3).
- Finally, this research presents VisExPreS, an interactive visual toolkit that enables a user-driven computation of local and global-divergence metrics using proactively guided versions of LAPS and GAPS, while enabling side-by-side comparison of multiple embeddings (cf. Section 6.3).

1.4 Thesis Organization

This thesis has been prepared in a paper-based format and is organized as follows:

Chapter 2 of this thesis presents a longitudinal study across organizations for identifying the state-of-the-art, challenges, and risk factors in the industrial software license renewals process. The study is performed using the Grounded theory method. To implement the method, semi-structured, cross-sectional, anonymous, self-reported interviews are carried out with 20 professionals from multiple organizations, later the Constant Comparative Method is used to analyze the collected data.

Chapter 3 presents a comprehensive survey of the recent advancements in the emerging field of Exploratory Data Analysis. It presents the results of the investigations on 50 academic and non-academic visual data exploration tools with respect to their utility in the six fundamental steps of the exploratory data analysis process. It also reveals the extent to which these modern data exploration tools fulfil the identified additional exploratory requirements of analyzing large datasets.

Chapter 4 presents the investigations and statistical analysis of 15 most popular dimensionality reduction algorithms on 12 state-of-the-art DR quality metrics using a large scale and systematic experimental study for five analytical contexts or DR. The final result presents a practitioners' guideline for the selection of an appropriate dimensionally reduction algorithm in the presented analytical contexts.

Chapter 5 presents two novel interactive explanation techniques for low-dimensional embeddings obtained from *any* dimensionality reduction algorithm. The first technique LAPS produces a local approximation of the neighborhood structure to generate interpretable explanations on the preserved locality for a single instance. The second method GAPS explains the retained global structure of a high-dimensional dataset in its embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space.

Chapter 6 enhances the LAPS and GAPS methods into proactively guiding users with the selection of representative data-points for analysis and incorporates the two techniques into VisExPreS, a visual interactive toolkit that enables a user-driven assessment of low-dimensional embeddings. Using a set of examples, it demonstrates the utility of VisExPreS in interpreting, analyzing, and comparing derived embeddings from different dimensionality reduction algorithms.

Finally, Chapter 7 concludes the thesis and presents a set of directions for future work.

References

[1] S. Liu, D. Maljovec, B. Wang, P. -t Bremer, and V. Pascucci, “Visualizing High-Dimensional Data: Advances in the Past Decade,” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 21–30, 2016.

[2] I. M. Johnstone and D. M. Titterington, “Statistical challenges of high-dimensional data,” *Proc. R. Soc. A*, vol. 367, no. 1906, pp. 4237–4253, Nov. 2009, doi: 10.1098/rsta.2009.0159.

[3] L. van der Maaten, E. O. Postma, and H. J. van den Herik, “Dimensionality Reduction : A Comparative Review,” *J Mach Learn Res 10*, vol. 66, no. 71, p. 13, 2008.

[4] R. Agarwal and C. E. Helfat, “Strategic renewal of organizations,” *Organization Science*, vol. 20, pp. 281–293, 2009.

[5] A. Schmitt, S. Raisch, and H. W. Volberda, “Strategic Renewal: Past Research, Theoretical Tensions and Future Challenges: Strategic Renewal,” *International Journal of Management Reviews*, vol. 20, no. 1, pp. 81–98, Jan. 2018, doi: 10.1111/ijmr.12117.

[6] M. A. Yalcin, N. Elmqvist, and B. B. Bederson, “Keshif: Rapid and Expressive Tabular Data Exploration for Novices,” *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 8, pp. 2339–2352, Aug. 2018.

[7] T. Kraska, “Northstar: an interactive data science system,” *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 2150–2164, Aug. 2018.

[8] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit, “Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 2023–2032, Dec. 2014, doi: 10.1109/TVCG.2014.2346260.

[9] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *arXiv:1705.07874 [cs, stat]*, Nov. 2017, Accessed: Mar. 24,

2020. [Online]. Available: http://arxiv.org/abs/1705.07874.

[10] J. B. Tenenbaum, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[11] Z. Cui, S. K. Badam, A. Yalçin, and N. Elmqvist, “DataSite: Proactive Visual Data Exploration with Computation of Insight-based Recommendations,” *arXiv:1802.08621 [cs]*, Sep. 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1802.08621.

[12] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit, “Taggle: Combining Overview and Details in Tabular Data Visualizations,” *Information Visualization*, vol. 19, no. 2, pp. 114–136, Apr. 2020, doi: 10.1177/1473871619878085.

[13] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, J. Heer, “Voyager 2: Augmenting Visual Analysis with Partial View Specifications,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, May 2017, pp. 2648–2659, doi: 10.1145/3025453.3025768.

[14] M. Cavallo and Ç. Demiralp, “A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration,” *arXiv:1811.12199 [cs]*, Nov. 2018, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1811.12199.

[15] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, “Visual analysis of dimensionality reduction quality for parameterized projections,” *Computers & Graphics*, vol. 41, pp. 26–42, Jun. 2014.

[16] J. A. Lee and M. Verleysen, “Quality assessment of dimensionality reduction: Rank-based criteria,” *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, Mar. 2009, doi: 10.1016/j.neucom.2008.12.017.

[17] S. Tufféry, *Data Mining and Statistics for Decision Making: Tufféry/Data Mining and Statistics for Decision Making*. Chichester, UK: John Wiley & Sons, Ltd, 2011.

[18] 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*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[19] J. M. Lewis and V. R. de Sa, "A Behavioral Investigation of Dimensionality Reduction," *In Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, p. 7, 2012.

[20] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality Reduction in the Wild: Gaps and Guidance," Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, Jun. 2012.

[21] J. Demśar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[22] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, "SIRIUS: Dual, Symmetric, Interactive Dimension Reductions," *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, Jan. 2019.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *arXiv:1602.04938 [cs, stat]*, Feb. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1602.04938.

[24] G. Plumb, D. Molitor, and A. Talwalkar, "Model Agnostic Supervised Local Explanations," *arXiv:1807.02910 [cs, stat]*, Jul. 2018, Accessed: Aug. 12, 2019. [Online]. Available: http://arxiv.org/abs/1807.02910.

[25] J. P. Boyd, "Additive blending of local approximations into a globally-valid approximation with application to the dilogarithm," *Applied Mathematics Letters*, vol. 14, no. 4, pp. 477–481, 2001.

[26] D. Sacha, L. Zhang, M. Sedlmair, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, D.A. Keim, "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis," *IEEE Trans. Visual. Comput. Graphics*, vol. 23, no. 1, pp. 241–250, Jan. 2017, doi: 10.1109/TVCG.2016.2598495.

Chapter 2

The Current State of Software License Renewals in the I.T. Industry

In the continuously evolving software industry, it is no longer common that organizations create and sell products directly to their customers; instead, customers are given access to these products in terms of license agreements [1]. These agreements not only include conditions on using the features of the purchased software but also a promise of assistance and support from the selling organization. Hence, for large multinational companies that sell licenses worth millions of dollars to other businesses, successful renewal of software licenses makes a key impact on the selling organization's revenue. In fact, customer and renewal acquisitions are considered as the two primary revenue sources for subscription-based organizations [2]. Whereas customer acquisition only occurs once during a customer's lifetime, the renewal of software licenses is, on average, an annual event. Alongside, research [3] shows that acquiring new customers not only can be 5 to 25 times more expensive than retaining existing customers, but a 5% increase in customer retention can also increase profits by 25%-125% [4]. Moreover, high customer renewal rates can provide an organization with a stable base for profitable growth.

Software licensing has been in practice for the last few decades [5]. However, with technology rapidly changing, open-source and cloud-based products are becoming more popular in the market [6]. These products allow customers to avail themselves of the same technological benefits with less initial cost. Hence, it is getting difficult for organizations to maintain an increasing, or even a stable, subscription renewal rate [7]. Also, the industrial practice of license renewals is directly driven by the changing consumption patterns of customers [2]. Therefore, the product licenses that were once purchased as cutting-edge technology are not as appealing to the same customers today. Despite being a key contributor to organizational revenue, the topic of subscription renewal strategies receives very limited attention in the literature. Although, there is some research that looks into the industrial practice of Business-to-Business (B2B) marketing [8], studies that particularly focus on understanding the end-to-end license renewal process and identifying the challenges and risks that motivates renewal decisions, are sparse. Moreover, on the one hand, there is much research that describes the processes and challenges associated to different business units [9], [10], but most of this work focus on the strategic renewal of organizations [9] and overall process improvements. On the other hand, there is ample research that considers managing customer relations [2], [11]–[13], addressing the challenges of predicting customer churn [4], [7], [14]–[16], and to trying to mitigate customer churn risks [12], [13], [17]; nevertheless, there is hardly any work that looks at the challenges with customer retention from the perspective of the subscription renewal process.

To bridge the gap in literature, in this research, we perform a longitudinal study with 20 participants across multiple organizations and multiple locations to attempt to distil the current practice of software license renewals in industry. Following the steps of the Grounded Theory method [18], we performed semi-structured, cross-sectional, anonymous, self-reported interviews [19], [20] with individuals from large multinational organizations, who are related to the subscription renewal process. We analyzed the information obtained from the interviews using the Constant Comparative Method (CCM) [21], [22]. An in-depth qualitative analysis of the data using CCM helped us to identify and refine the core ideas of the interviews. From this input, we formed a set of theories (cf. Section 2.2.1.1) depicting a comprehensive picture of "the most common practice" of software license renewals in today's I.T. industry. We also identified the challenges and risk factors associated with the license renewal process from our data analysis. We validated our analysis results with content validity expert(s) from the participating organizations, where we also identified a set of strategies and research directions for mitigating the identified risks and challenges by engaging in structured brainstorming activities [23] with the experts. It is important to mention that, the results presented in this chapter does not reflect the practice of an individual organization, instead, it is a montage of the most popular strategies followed by a group of large multinational organizations. The primary contributions of this research are three-fold and can be summarized as follows:

- This research represents the first to our knowledge that synthesizes the current industrial practice of the end-to-end software license renewal process (cf. Section 2.3.2).
- This study identifies a set of challenges (cf. Section 2.3.3) and risk factors (cf. Section 2.3.4) in the license renewal process, that impact on renewal decisions of customers, and hence on the overall revenue of seller organizations.
- Finally, this work presents a list of immediate action plans (cf. Section 2.4.1) and a set of directions for future research (cf. Section 2.4.2), that can help organizations with mitigation of the risks in the license renewal process.

The chapter is organized as follows: in Section 2.1 we present the theoretical background and discuss the fundamental terminology used in this chapter, whereas in Section 2.2 we give a detailed description of our research methodology. The results of our study are presented in Section 2.3 with a list of proposed action plans and research directions being discussed in Section 2.4. In Section 2.5, we discuss some limitations and future work opportunities for our study, while Section 2.6 concludes the chapter.

2.1 Theoretical Background

2.1.1 Background and Research Questions

Software license agreements are legal contracts between end users and software publishers, that inform the end users of their privileges when using a software product and restricts what the users can or cannot do with the product [1]. From an end user's perspective, software license agreements specify the users' privilege to download and install software [5]. From a publisher's perspective, these contracts specify the details of promised technical support associated with the licensed software products.

Software licensing arguably came into practice in mid-1980s [5]. For the last few decades, several researchers [1], [5], [24]–[26] have analyzed different aspects of software license agreements. Much research has been done on the design [1], significance, completeness [24], and comprehensiveness of license agreements. However, in the 21st century, with software being offered as services hosted on cloud [16], the practice of software licensing has seen some change. Nevertheless, large organizations such as Microsoft [27] and IBM [28] continue to provide license and support for software products to their customers. According to the 2017 annual report of Microsoft Corporation [27], $44 billion (USD) were defined as unearned revenue from selling software license agreements. Similarly, for IBM Corporation [28] this amount was approximately $21 billion (USD). Hence, it is clear that successful renewal of these license agreements is a matter of paramount importance for these organizations.

In large companies, the end-to-end license renewal process is a collaborative work of multiple teams. Among these teams, the sales and subscription representatives (S&S reps) or, more commonly known as renewal reps, work at the front end with the customer. At the back end, there are brand representatives or leaders who are responsible for managing products from specific brands. Apart from these, there are worldwide reps or global sales reps that take care of customers across different regions and continents. However, without proper strategies, processes, tools, and support, it can get very difficult for the different teams to collaborate and work together [29]. According to Agarwal et al., [9], in order for multiple teams to successfully co-operate with each other, a dedicated process of organizational management is required in every department. Hence, for this study, we formulate our first research question as:

> ***RQ 1:*** *What is the end-to-end industrial process of software license renewal?*

Much research [2], [12], [13], [30] has been done on identifying the challenges of customer relationship management. Research has shown that, in any collaborative practice, effective communication [31], [32] between the associated teams is an absolute necessity for providing high-quality customer service. As per Suter et al. [29], despite an efficient performance from each individual department, in absence of effective role understanding and communication, the risks and challenges in collaborative work can increase dramatically [33]. Hence, based on the above statements, we construct the following research questions for our study:

> ***RQ 2:*** *How does the communication among stakeholders impact the end-to-end software license renewal process?*

> ***RQ 3:*** *How is information exchange in a distributed workforce associated with the challenges of the identification of customers that are likely to not renew their contracts?*

Apart from effective communication within the departments, the seller organizations also need to ensure the trust [11] and satisfaction of their customers [34], [35]. Research shows that customer trust and satisfaction is directly related to the risks associated with customer retention [36], [37]. As mentioned by Hannan et al. [34], often customers choose to move to competitors because of unsatisfactory customer service and experience. Research [10] also shows that the amount of business value generated from the purchased licenses can impact the renewal decisions from customers. Hence, we form the final two research questions for this research as:

> ***RQ 4:*** *How do customer satisfaction and trust impact on the risks associated with the license renewal process?*

RQ 5: *How does the value generated from purchased licenses impact on the renewal decisions of customers?*

With a primary focus on the broad perspective of elicitation of the license renewal process and its challenges, the research questions discussed in this section formulate the basis of this research. These questions were developed in close collaboration with the content validity experts from different organizations that participated in our study. Over the past years, much research has been done that attempt to identify customers at churning risks [2], [12]–[14] by looking into behavioral traits such as software usage data and the number of comments from customers. However, as pointed out by Haenlein et al. [7], these features usually vary from domain to domain [6], [11]. For example, factors relevant for software-publisher [38] companies, may be invalid for the telecommunication domain. The research questions defined in this section attempt to distil such factors for the domain of software license renewal.

2.1.2 Fundamentals and Terminology

This section provides the necessary background on the key roles and terminologies for the license renewal process. The section is divided into two parts; the first part identifies the key people who take part in the renewal process. The second part defines the industrial terminologies used during license renewals.

2.1.2.1 Key Roles in the Renewal Process

Following are the key people who take part in the license renewal process:

i. **Sales & Subscriptions Representatives (S&S Reps):** These are the representatives in the seller organization (commonly known as renewal reps); who are the first line of contact for customers in the subscription renewal process. The renewal reps are responsible for guiding customers through the end-to-end renewal process.

ii. **Brand Representatives or Brand Managers:** These are representatives of specific brands in the seller organizations; who are responsible for resolving technical difficulties (e.g., feature incompetence, pricing problems) with products belonging to these brands.

iii. **Worldwide Representatives or Global Sales Reps:** Also known as worldwide leaders or the global sales leaders, these are representatives of the seller organization that work across multiple brands and territories.

iv. **Business Partners:** These are industrial partners of the seller organization, that help with marketing, selling, and renewing of subscriptions. Since they are not a part of the seller organization, they usually have their own profit margin. Such partner organizations can sell product licenses for other companies, including competitors, as well.

2.1.2.2 Terminologies Related to Renewals

Key industrial terminologies related to the subscription renewal process are described as follows:

i. **Quotes / Price Quotes:** These are documents that include customer license information such as the product code, the purchased quantity, due dates of renewal, and pricing information.

ii. **Purchase Orders:** These are documents that are sent to customers by the renewal reps when both parties agree on renewal pricing and volumes.

iii. **Partial Renewal:** This is a final renewal status, that occurs when a customer decides to not renew the subscriptions for some of their pending licenses.

iv. **Reinstatement:** It is a penalty, that customers need to pay when they decide to renew their licenses after the renewal due date is passed.

v. **Product Migration:** This is a situation that occurs when a customer decides to change their business location and wants to move some of their licenses to the new location.

vi. **Product Evolution:** Evolution of products happen when the seller organization decides to modify and market an existing product in a different form. This may include combining many products into one unified product or splitting one product into distributed solutions.

2.2 Methodology

As per our background analysis, communication gaps, lack of customer trust, and deficiency of business value on the customers' side can cause serious challenges in any business process [29], [31], [32]. Therefore, along with the end-to-end software license renewal process in the I.T. industry, this study also aims at determining the impacts of the above-mentioned factors on the renewal decisions of customers. In order to achieve our goals, we implemented the Grounded Theory method developed by Glaser and Strauss [39], [40]. Grounded theory is an inductive research technique [41] that is commonly used for qualitative data analysis [21]. With the help of systematic data collection and analysis methodologies, the grounded theory approach allows its users to construct theoretical propositions from data. As shown in Figure 2.1, the grounded theory method is primarily composed of steps namely: (i) systematic data collection, (ii) data analysis, (iii) theoretical integration [18], [42], and (iv) validation of developed theories. In order to implement the grounded theory approach, in this research, at first, we collected the required data using semi-structured, cross-sectional, self-reported interviews [19], [20] with 20 professionals from multiple organizations. Next, the Constant Comparative Method (CCM) [18], [21], [22], [43], [44] was used to qualitatively analyze the information obtained from the interviews. Finally, we developed a set of theoretical propositions regarding the end-to-end process of software license renewals, along with the various challenges and risks involved in the license renewal process. We validated our analysis results using a quantitative measure of Inter-rater Reliability (IRR) [45], [46]. In this section, we give detailed descriptions of our research context, data collection, data analysis, and validation methodologies.

2.2.1 Research Context

The results presented in this chapter are the outcome of our research project that started at the end of the year 2017 and lasted for approximately 8 months. The project was carried out with the aim to understand the end-to-end process for software license renewals and to identify the risks and problems associated to the process. Overall, 20 employees across multiple organizations participated in this research. All organizations chosen for our study are headquartered in North America and are classified as Software Publishers (code: 511210) by the North American Industry Classification System [38]. These are organizations that design, produce, and distribute computer software along with an assurance for support services, in terms of license agreements. As for other firmographic variables [47], we chose organizations that have over 10,000 employees and operate worldwide. Each organization has its own distributed license renewal teams, which interact internally and with their clients via modern communication media. The departments we interacted with handle individual renewal contracts worth between approximately $100,000 (USD) and $5,000,000 (USD). The primary focus of our research has been Business-to-Business (B2B) [48], including Business-to-Government (B2G), subscription renewal agreements, where both the seller and the buyer are organizations. It is worth mentioning that, the B2B renewal contracts investigated in this study, have no impact on their Business-to-Consumer (B2C) equivalents. This research presents an aggregation of the collected information from all the participating organizations. Following strict anonymity requirements, this chapter strategically avoids mentioning the names and/or the detailed locations of the organizations that participated in our study.

2.2.1.1 Data Collection and Analysis using Grounded Theory Method

In this section, we elaborate on our application of the grounded theory method and discuss different steps of the process from the context of this research. This section is primarily divided into four subsections. We begin with a detailed description of the semi-structured interviews conducted with different stakeholders of the software license renewal process. Next, we discuss the steps we followed while analyzing the obtained information using the Constant Comparative Method. Later, we present our theoretical propositions that emerged from our data analysis. This section ends with a brief discussion on the reliability of our applied research methodology.

Data Collection and Intervention

In order to obtain the perspective of each individual study participant, we chose semi-structured cross-sectional anonymous self-reported interviews [19], [20] for data collection. It is a well-accepted approach in the literature, that has been used by several information systems researchers [49]–[51], in different contexts. Unlike a fully structured interview [52] that consists of a standard set of questions and follow-up questions, a semi-structured interview is a qualitative research technique that allows the interviewees to openly express their ideas and views on the problem. On the other hand, a cross-sectional survey enables researchers to analyze information at a specific point in time [53]. According to researchers, semi-structured interviews are ideal for cross-sectional cases, where the interviewers get to interview each participant only once. In par with our approach, an extensive amount of research [45], [54], [55] exists that not only relies on anonymous surveys [19], but also uses semi-structured cross-sectional studies that involve self-reported behaviors [19], [49], [50].

Our interaction in each organization started with a key contact person, typically program directors, who enabled our access to the right personnel in their organizations. From each organization, we interacted with individuals from different roles such as renewal reps, brand reps, global-sales leaders, data analysts, and program directors. With the help of our key contact person, we scheduled meetings with employees from at least two different roles in each organization. While some of these interactions took place in a face-to-face environment, the rest were carried out remotely. During our interactions with the participants, we used an interview guideline composed of open-ended questions discussed in Table 2.1. Like any other semi-structured interview [56], we prepared a set of common questions (see Table 2.1) that was used to initiate a conversation with each interview participant. However, depending on the answer of any given question from a participant, the follow-up questions were asked. Our questions were strategically prepared to obtain general information about the overall process flow of subscription renewals in the organizations, along with the personal experiences

Table 2.1: Common Questionnaire for All Interview Participants

Topic	Questions
Overall Process of License Renewals	What is the overall process for the software license renewals in your organization?
Stakeholders of the License Renewal Process	Who are the different stakeholders involved in the license renewal process? Can you elaborate on the roles and activities of these stakeholders? What is your role in the license renewal process? Can you elaborate on the day to day activities of someone in your role? How do you communicate with the other stakeholders of the renewal process? Do you think there are any communication gaps between the involved stakeholders? What do you think can be changed to improve the level of communication amongst the stakeholders?
Customers of the License Renewal Process	Do you interact directly with the customers? On a quarterly basis, how many renewals are handled by you? What are the approximate price ranges for the software licenses that are handled by your team? Do you think renewals belonging to different price ranges receive equal attention?
Other Aspects of the License Renewal Process	Do you use any software applications to manage renewals? If so, what are the useful and challenging aspects of these applications? What are the biggest challenges you have faced while doing your job? Can you share your experiences with the customers that you felt were challenging? In your experience, what are the reasons for customers to not renew their license agreements? Without communication from the customer's end, is there a way for you to know that they might not renew some pending licenses? If you knew about the possible non-renewal earlier, could you save the renewal? How do you locate the key people associated with a specific renewal?

Note: Additional follow-up questions were asked, to each participant depending on their role and answers to the previous questions. Also, during the theoretical sampling, the interview questions were updated based on the analysis that was already done.

of the interviewees on factors that influence the renewal process. The responses of the participants were literally transcribed, allowing the destruction of the original material, on the same day of the interviews; in addition, all identifying remarks were perpetually removed and destroyed to protect all the participants. Beyond this summarized content, no other information was taken outside the organizations protecting their confidentially. Content validity expert(s) in every organization was allowed to delete all information which they believed was unique or sensitive to their organization. The aim of this step was to obtain a complete description of "the industrial practice of software license renewal", where the details are limited to reflect only common practice and avoid unique undertakings from an organization to avoid the risk of the inference of their identity.

Data Analysis using the Constant Comparative Method

Data analysis using the Constant Comparative Method (CCM) [18], [22] constitutes the core of the grounded theory method [39], [40]. CCM has been used by researchers [22], [42], [43], [56] in different domains for developing concepts and theories from qualitative data. The overall steps of CCM [18], [22], [43] are depicted in the data analysis step, in Figure 2.1. Data analysis in CCM primarily consists of coding and theoretical sampling [18], [43], where coding involves three levels of analyses [18], [22]: (i) open coding, (ii) axial coding, and (iii) selective coding, and theoretical sampling involves collecting additional information to gather new insights to refine the identified concepts. Table 2.2 summarizes the aims, asked questions, and obtained results from different steps of the coding process of CCM followed in this research. In order to apply CCM, we followed the guidelines of Boeije et al. [22], where we coded each of our collected documents such as interview transcripts and observation notes, into categories. The coding process involved using the tools ATLAS.ti and Microsoft Excel and was carried out by reading each of these documents and attributing codes to sentences, paragraphs, and sections. These codes were then associated with a theme or idea, from which our conclusions on the license renewal process were drawn.

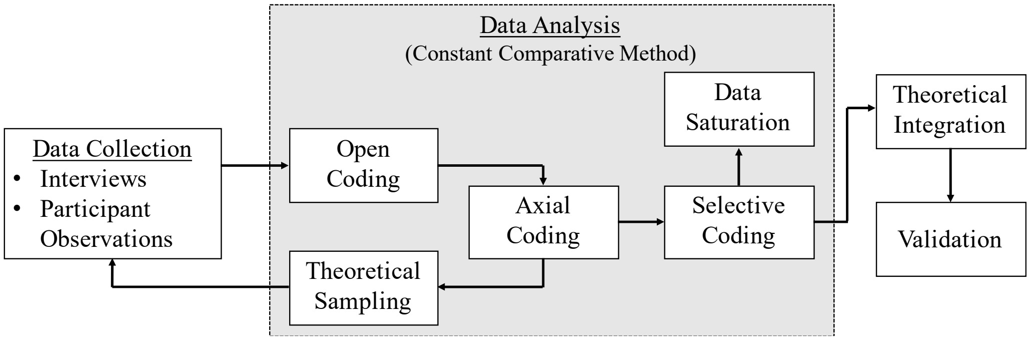


Figure 2.1: Steps Followed in this Research for Data Collection and Analysis using Grounded Theory Method

Right after the first interview, we started analyzing the data with open coding [18], [44], using the qualitative data analysis tool ATLAS.ti we scrutinized the interview transcript line by line and attributed categories to sentences, paragraphs, and sections. These categories represented themes or concepts for each of these parts of the data, with which they are associated. For example, the category 'internal challenges' (cf. Figure 2.2) was attributed to sentences that mentioned challenges within the organization that affected the license renewal process. As our next step, we performed axial coding [44], where we analyzed, compared, and characterized the interview fragments with the same category; in addition, we found relationships among all the different categories that were assigned to the data. During this step, we also started creating memos [22], [43] that defined each category along with its properties and demonstrated the relationship of this category to other categories [18]. At this stage, theoretical sampling [43] occurred as we interviewed more and more stakeholders of the license renewal process from different organizations. Theoretical sampling helped us to refine and check the properties of our developed theoretical characterization from the categories; see Figure 2.2 for details. Finally, we performed selective coding [22], again see Figure 2.2, in order to identify the most significant and frequent categories, systematically connected them to our developed theoretical categories, until the point where data saturation [56] occurred. Data saturation is a situation where the information collected from the interviews become redundant and no new information could be obtained from

Table 2.2: Detailed Steps of the Coding Process using the Constant Comparative Method

Coding Activities	Aim	Questions	Outcomes
Open Coding	Systematic development of categories	• What is the core message of this interview? • Is this interview consistent? • Are there interview fragments that are coded with the same categories? • What are the relationships between the identified categories?	• Interview summaries • Categories • "Code-tree" showing interrelations among categories
Axial Coding	Conceptualize the categories and produce a typology	• Are the participants from two interviews talking about the same things? • Do the same categories and the combination of categories occur in both the interviews? • How are the categories related in both the interviews?	• Extended memos • Expansion of categories • Relevant themes among categories • Clusters of interviews
Selective Coding	Integrate categorical findings	• What are the central concepts in all the interviews? • What are the relationships between the most significant and frequent categories in all the interviews?	• Central concepts • Extended memos • Conceptual profile of relationships among themes

further interviews.

Theoretical Integration

Following the guidance of Boeije et al. [22] and Sbaraini et al. [41], we present the results of our grounded theory analysis using a set of five theoretical propositions regarding the end-to-end process of the software license renewals in the I.T. service industry. These theories emerged from our systematic analysis of the collected data using the constant comparative method. Figure 2.2 presents a 'code tree' [22] depicting a mapping between of the first order codes to initial theoretical categories [18] and then to a set of aggregated theoretical dimensions [42], [43] obtained from the open, axial, and selective coding approaches of CCM. The code tree not only summarizes the identified patterns of relationships between the category

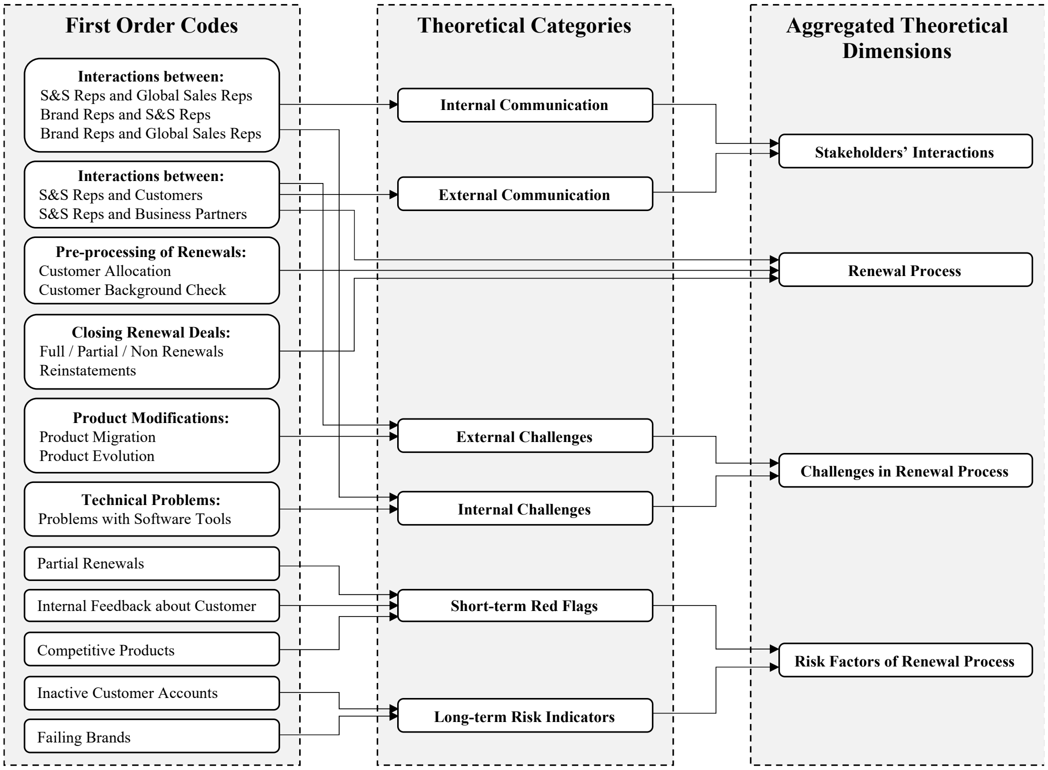


Figure 2.2: Code tree output of Data Analysis using Constant Comparative Method

characterizations in the data, but it also forms the basis of the theories that emerged from our data analysis.

As shown in Figure 2.2, the responses of all the interview participants were focused on four aspects of the license renewal process namely: (a) stakeholder interactions, (b) process steps, (c) challenges, and (d) risks. From this synthesized information we answer our research questions (cf. Section 2.1.1) regarding the end-to-end license renewal process, along with the effects of communication, customer satisfaction, and value generated from the purchased licenses on the renewal decisions of customers. In this section, we present a set of five Theoretical Propositions (TP-1 to TP-5) that emerged from our implementation of the grounded theory method, answering our research questions RQ1 to RQ5 discussed in Section 2.1.1.

Research [9] shows that, departments of large multinational organizations need to follow dedicated processes in order to operate successfully. Moreover, communication [31], [32] among stakeholders is an important factor that influences the successful operations of any organizational department. In support of these facts, on the one hand, our data analysis using CCM shows, during the interviews each study participant not only mentioned the roles of different stakeholders, but they also mentioned several gaps in stakeholders' interactions. On the other hand, our analysis of the second theoretical category (cf. Figure 2.2) shows that most industry practitioners mentioned more or less similar set of steps in the license renewal process. Hence, as answers of our research questions RQ1 and RQ2 (cf. Section 2.1.1), we develop the following theoretical propositions:

> ***TP-1:*** *(Effective) communication is (positively) associated with the end-to-end license renewal process. That is, the smaller the communication gap is, the more effective the process.*
>
> ***TP-2:*** *The need to (successfully) close the pending renewals on time is (positively) associated with the need for a dedicated process for the practice of software license renewal.*

Moreover, during our analysis of the interview transcripts, we realized that the factors such as communication gaps among stakeholders, lack of customer trust, and scarcity of value generated from the purchased licenses act as the primary sources for the challenges and risks in the license renewal process. The code tree depicted in Figure 2.2 confirms this finding of ours. Hence, as answers to our research questions RQ3, RQ4, and RQ5, we develop the following theories:

> ***TP-3:*** *Information exchange in a distributed workforce is (negatively) associated with the challenges of (effective) identification of customers who are likely to not renew the contracts. That is, the higher the information exchange, the lower the challenges in the renewal process.*
>
> ***TP-4:*** *The renewal risk is (negatively) associated with customer satisfaction and trust. That is, the (more) satisfied the customers are*

> *with the services from the seller organization, the lower the risks are in the renewal process.*
>
> ***TP-5:*** *The renewal risk is (negatively) associated with the business value. That is, the (more) value the customers can generate from the purchased licenses, the lower the non-renewal risks.*

It is important to note that, the presented theories are derived from an in-depth analysis of the consensus of the stakeholders' opinions on the license renewal process, hence, supporting the qualitative validity of the emerged theories. An empirical study with a larger population of stakeholders that could statistically validate the proposed theories, is beyond the scope of this research.

Validation of Analysis Results

We performed a three-level validation of our analysis results from CCM. Firstly, we validated each step of our analysis internally across the set of authors through team meetings and discussions. Secondly, we validated our derived concepts by presenting our analysis results to a team of content validity experts [19] from the surveyed organizations. Finally, to quantitatively justify the reliability of our analysis results we calculated the Inter-rater Reliability (IRR) [42], [57] of our findings. IRR is a common measure [46], [58] to evaluate the reliability of qualitative studies. IRR involves multiple researchers independently coding, clarifying, and re-coding the obtained data until a specific level of accordance is achieved [59]. It enhances the fidelity of the analysis by answering the question of whether different researchers code the same data in the same way or not [60]. Researchers [61], [62] have proposed several metrics to measure IRR, among which Cohen's Kappa [61], that calculates the percentage of agreement among coders is commonly used [42]. As formally defined by Cohen et al. [61], Kappa can be computed as:

$$K = \frac{P(a) - P(e)}{1 - P(e)} \tag{2.1}$$

Where P(a) represents the observed percentage of agreement between coders, and P(e) represents the probability of agreement between coders, due to chance. Possible values for Kappa can range between -1 to 1 [63], where 1 signifies perfect agreement, 0 indicates completely random agreement, and -1 signifies perfect disagreement.

During the validation of our analysis results, we adopted IRR as a tool to validate the reliability of our results and used Cohen's Kappa as the metric to measure IRR. Using two coders to analyze qualitative data is a common approach among researchers [42], [46], [58] to increase the validity and reliability of the study results. Hence, in order to implement IRR, two coders were involved in independent analysis and coding the transcripts from the interviews and the convergence of their findings was evaluated at the end of each open, axial, and selective coding phases. In cases of conflicts between the decisions made by these two coders, a third coder was involved in the discussions for resolving the conflicts. At the end of each coding phase, we merged the coding files from ATLAS.ti and exported the coding results of each researcher into Microsoft Excel. We used Microsoft Excel to calculate Kappa as a measure of IRR. Table 2.3 shows the list of Kappa values for the theoretical categories presented in Figure 2.2.

Later, after complete anonymization and aggregation, we presented the summarized results of our analysis along with our defined theoretical propositions to a team of content validity experts.

Table 2.3: Kappa output for Theoretical Categories

Theoretical Category	Coder 1	Coder 2	Kappa (K)
Internal Communication	79	80	0.82
External Communication	33	31	0.77
External Challenges	41	41	1.00
Internal Challenges	53	48	0.67
Short-term Red Flags	28	31	0.79
Long-term Indicators	22	24	0.91

During this discussion, we not only presented our final coding-tree (cf. Figure 2.2) but also discussed disagreements among the two coders. Due to the nature of the participating organizations, while some interactions happened face-to-face, others took place over teleconference.

2.3 Results

This section elaborates on the code-tree depicted in Figure 2.2 and discusses the results of our data analysis in detail. In this section, we present the detailed rationale behind the five theoretical propositions that emerged from our implemented grounded theory method. The section is primarily divided into four subsections based on the four aggregated theoretical dimensions presented in Figure 2.2. We begin with an analysis of the interactions between different stakeholders involved in the renewal process and identify the level of communication between the teams. Next, we discuss the end-to-end license renewal process along with the challenges and risks associated with the process.

2.3.1 Stakeholder Interactions in the License Renewal Process

In order to find answers to our five research questions (cf. Section 2.1.1), we started with the identification of the level of communication among the key stakeholders involved in the subscriptions renewal process. Our findings are depicted in Figure 2.3. During our study, we found that in each participating organization, the S&S reps (cf. Section 2.1.2) act as the first line of contact for the customers, the brand reps manage products from specific brands with global sales leaders overseeing the renewal process across territories. Figure 2.3 shows the communication gaps among the involved teams. During our study, we observed that in many cases, the S&S reps not only have limited access to the archived license renewal data from previous years, but also they face a hard time to find the right representatives from other teams, who might help them with brand or price specific challenges with a renewal. Moreover, since different teams focus on different aspects of the renewals, each team usually possesses different types of information about the customers. However, a lack of proactive information sharing has been observed among these

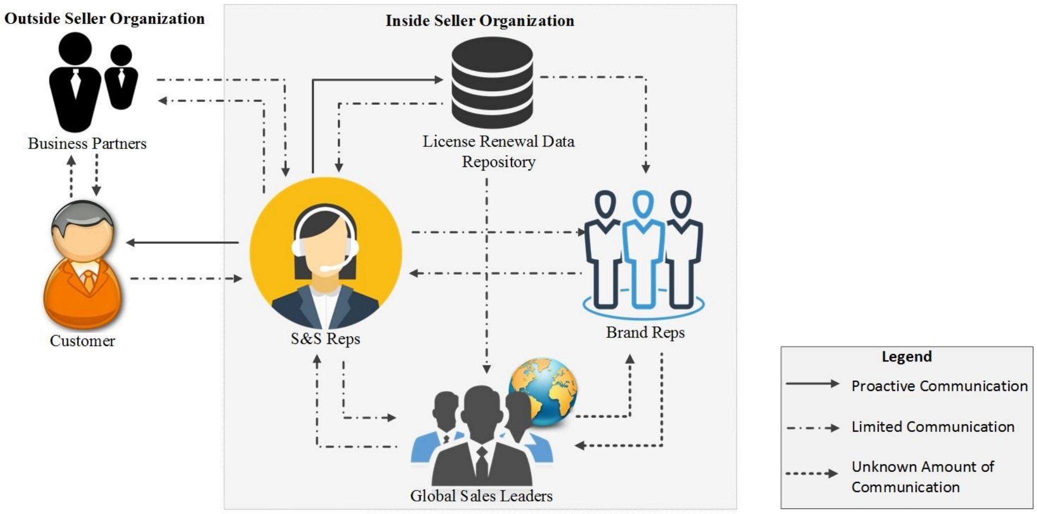

Outside Seller Organization
Inside Seller Organization
Business Partners
License Renewal Data Repository
Customer
S&S Reps
Brand Reps
Global Sales Leaders
Legend
Proactive Communication
Limited Communication
Unknown Amount of Communication

Figure 2.3: Stakeholder Interactions During the License Renewal Process

teams. For example, the brand reps often have specific information on product features that are not being liked by customers, however, this information is not voluntarily shared with the renewal reps. A similar gap in communication has been observed between the renewal reps and the global sales leaders. On the other hand, business partners, being independent workforces outside the selling organizations, also do not share much information with the S&S reps. Hence, from our analysis, we developed the theoretical proposition TP-1 (cf. Section 2.2.1.1) and conclude that there is a significant lack in effective communication among the stakeholders, that increases the challenges in the renewal process.

2.3.2 End-to-End License Renewal Process

This section supports our theoretical proposition TP-2 (cf. Section 2.2.1.1) and gives an overview of our findings on the overall license renewal process followed by today's I.T. service industry. The detailed renewal process is depicted in Figure 2.4 using a BPMN collaboration diagram. The figure depicts all the different stakeholders of the license renewal process and shows the interactions between them using BPMN connecting objects. For ease of presentation, we divide the process into four distinct steps. In the next subsections, we elaborate on Figure 2.4 and give a step by step overview of the process.

2.3.2.1 Pre-Processing of Renewals

The renewal process starts at the beginning of each quarter about three to four months prior to the renewal due dates. At this point, system generated price quotes (cf. Section 2.1.2) are sent to the customers via automated emails. Concurrently, renewal reps also get assigned to their designated customers for the quarter. Sometimes the renewal reps get assigned to customers they have previously worked with, and at other times they get assigned to new customers. Hence, to equally assist all their customers with the renewal process, the reps analyze background information for each customer before they initiate any communication with them. At this time, the reps collect information on previous renewal transactions of their customers and locate the key people involved in these transactions.

2.3.2.2 Contacting the Customer

After the background analysis of customers, the renewal reps initiate personal communication with them. The reps often start with the customers who have the most dollar amount of pending renewals and work their way down the list. In their first direct communication, the reps send a personalized email to the customer with the same previously sent renewal quote as a reminder of closing the renewal before its due date.

2.3.2.3 Customer Responses

Usually, very few customers respond to the first system generated email, and the first personal email from renewal reps often triggers the communication between the two parties. Nevertheless, customer responses can be very unpredictable. As shown in Figure 2.4, customers react to renewal reminders from the reps in the following three different ways:

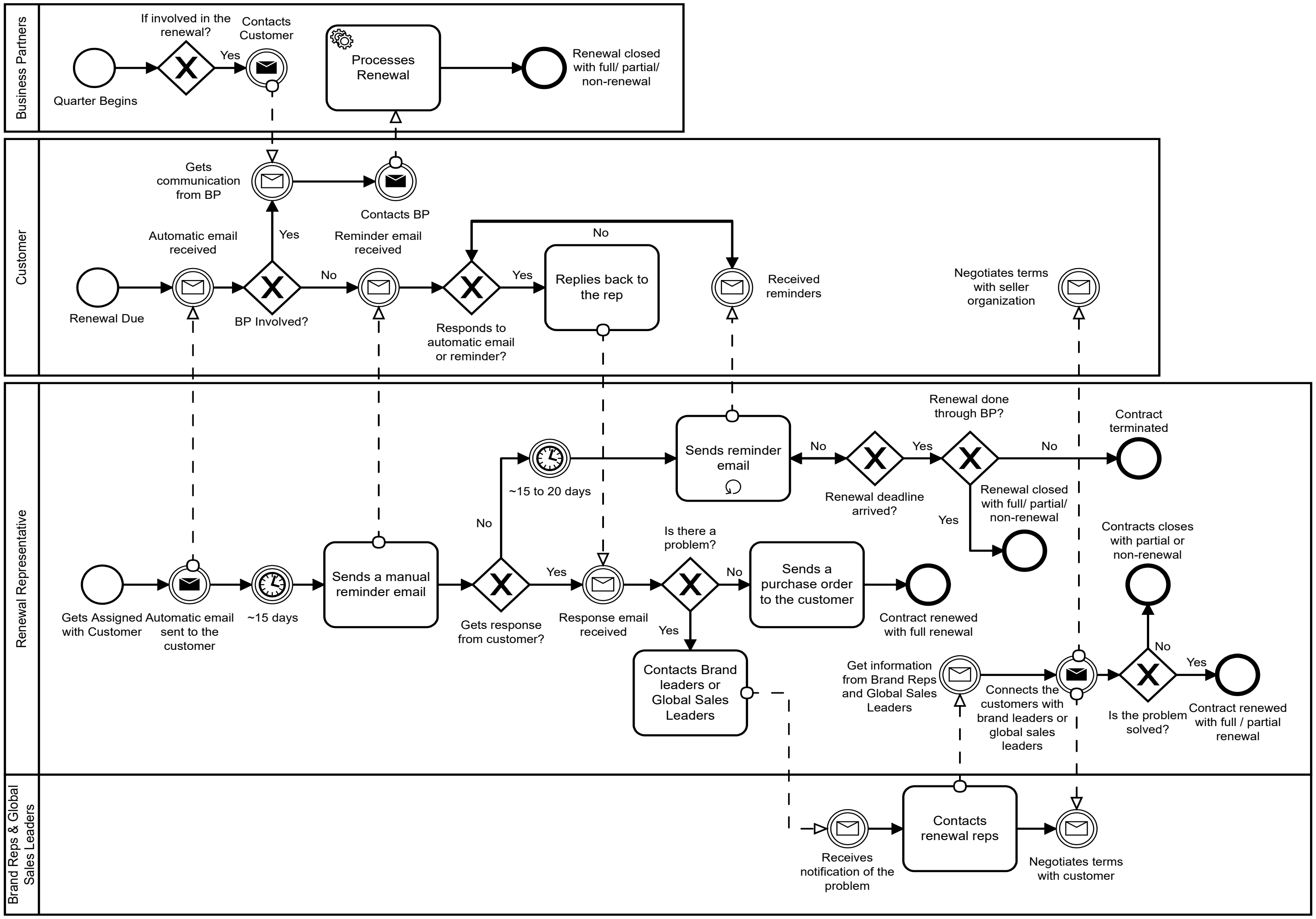

Business Partners
Quarter Begins
If involved in the renewal?
Yes
Contacts Customer
Processes Renewal
Renewal closed with full/ partial/ non-renewal
Customer
Gets communication from BP
Contacts BP
Automatic email received
Yes
Reminder email received
No
Renewal Due
No
BP Involved?
Yes
Replies back to the rep
Received reminders
Negotiates terms with seller organization
Responds to automatic email or reminder?
Renewal Representative
Renewal done through BP?
Contract terminated
Sends reminder email
No
Yes
No
~15 to 20 days
Renewal deadline arrived?
Renewal closed with full/ partial/ non-renewal
Yes
No
Is there a problem?
Contracts closes with partial or non-renewal
Yes
No
Sends a purchase order to the customer
Sends a manual reminder email
Gets Assigned with Customer
Automatic email sent to the customer
~15 days
Gets response from customer?
Response email received
Contract renewed with full renewal
Yes
Contacts Brand leaders or Global Sales Leaders
Get information from Brand Reps and Global Sales Leaders
Connects the customers with brand leaders or global sales leaders
No
Yes
Is the problem solved?
Contract renewed with full / partial renewal
Brand Reps & Global Sales Leaders
Receives notification of the problem
Contacts renewal reps
Negotiates terms with customer

Figure 2.4: The End-to-end License Renewal Process Followed by the Service Based I.T. Industry

i. **No Response:** In some cases, even after multiple reminders, customers do not respond to the emails from reps. Hence, the reps do not get a clear picture of the probable renewal outcome and are forced to assume that either the customer is not interested in renewing the licenses, or they are in contact with business partners for the renewal. In both the cases, reps try to contact the key people in the customer organization and in the business partner organizations.

ii. **Negative Response:** Customers often respond to the reps with unwillingness to renew some or all pending licenses due to concerns regarding product pricing or usefulness. In such cases, reps often try to analyze the reasons behind the customer's decision and engage the right people from the selling organization to assist the customers with their challenges.

iii. **Positive Response:** These are the scenarios, where a customer informs the rep about their willingness to renew the pending licenses without any concerns. In such cases, reps send out the purchase orders (cf. Section 2.1.2) to the customers before the renewal due date and endorse new product licenses.

2.3.2.4 Closing the Renewal

Prior to the renewal due date, the reps try to get a final decision from the customers regarding the renewals. For the customers who decide to renew all or some of their licenses, the reps send legal purchase orders. Often customers, who do not communicate to the renewal reps at all, choose to renew their licenses via business partners, in such cases once the licenses are renewed the renewal reps get notified in the system. However, if the reps do not get any notifications on the renewal status of some customers even after their renewal due dates are passed, cancellation letters are sent to these customers for their pending licenses. As shown in Figure 2.4, customers often choose to renew all their pending licenses (full renewal) or to renew some of the licenses and cancel the rest (partial renewal).

2.3.3 Challenges in the Renewal Process

In this section, we present the rationale behind our theoretical proposition TP-3 (cf. Section 2.2.1.1), as we list out the challenges that were identified in the subscription renewal process during our study. As per our analysis, we classify the challenges into two categories namely: external and internal challenges, based on their source of origin. In the next two subsections, we discuss these two categories in detail.

2.3.3.1 External Challenges

External challenges are those that originate from outside the organization. During our survey, we identified the following challenges that fall into this category:

i. **Unresponsive Customers:** These are among the biggest challenges the reps face during the renewal process, as customers who do not communicate back often drop most of the licenses. Even though some customers prefer to communicate with business partners instead of renewal reps, the reps never know until the last moment if the customer is having any trouble with the licenses. In these cases, there is no way for the reps to help the customers.

ii. **Migrations and Product Evolutions:** Migrations can acutely affect the renewal, as they involve a set of licenses being moved to a different physical location with diverse business conditions. On the other hand, product evolution may require pricing changes of the product. In both these situations, customers usually evaluate the value generated by their subscriptions. Hence, they may decide to add or drop some of their licenses or maybe even switch to another product from a different organization.

iii. **Getting Real Feedback from Customers:** Most renewal reps confirmed that there is no way to know what exactly is happening on the customer's side. For example, in most organizations, although there are online portals that show if a customer attempted to download a software, there is no record of whether the software download and install were successful. This affects customer satisfaction and trust. Hence, without a real feedback from their customers, it becomes impossible for the reps to anticipate the renewal outcome.

iv. **Renewal via Business Partners:** Not all contracts are sold directly through the organization; some contracts are sold through business partners. In such cases, renewal reps usually do not have access to any information about the customer's experience. Hence, when the renewal due date arrives, the reps only get to see the final sale records with no additional information. Therefore, the reps never get to step into the renewal process and help the customers with any challenges they might be facing.

2.3.3.2 Internal Challenges

Internal challenges arise from inside the selling organization. Challenges in this category are discussed as follows:

i. **Internal Communication Problems:** As identified in Section 2.3.1, there is a significant communication gap among the stakeholders that causes several challenges in the renewal process. During our study, we discovered some reasons behind this gap. Firstly, renewal contact points for customers change frequently due to people leaving the organizations or moving to different teams, however rarely any update is made in the internal contact database. Secondly, different stakeholders of license renewals have different perspectives of the process, with enough workload of their own. Finally, there is no standard way of documenting the details of interactions with customers. Often reps document valuable customer information in the form of hand-written notes or emails, but once these reps leave the organization or move to different teams, this information is lost forever.

ii. **False Estimation of Renewal Risks:** Many factors that indicate renewal risks can be misleading. For example, although unresponsive clients are considered the number one risk by most reps, the contracts with these customers may end up with a full renewal with the help of business partners. Furthermore, customers with multiple partial renewals can purchase more licenses based on their needs. Reps often make mistakes in identifying the customers, who are at risk of not renewing their licenses, as there is no formalized mechanism to do this task.

iii. **Technical Problems with Tools:** Often renewal monitoring tools do not perform efficiently; as they freeze due to the high volume of data. Also, reps often need to work with multiple tools at the same time to monitor multiple aspects of the renewal process, as no one tool provides all the necessary information. As a result, even though organizations provide renewal reps with significant software resources, reps often perform much of the analysis manually.

2.3.4 Risk Indicators in the Renewal Process

This section presents our list of identified factors that indicate a customer is at a risk of non-renewal and supports our theoretical propositions TP-4 and TP-5 (cf. Section 2.2.1.1). We classify the risk factors into two categories, namely: Short-term Red Flags and Long-term Indicators. While the short-term red flags are risk factors that suddenly show up during the renewal process, the long-term risk indicators are risks that have been there for some time but were never mitigated.

2.3.4.1 Short-term Red Flags

Short-term red flags could be driven by a keyword mentioned in a conversation with the customer, or from the final state of a previous renewal. However, the short-term red flags are often ignored as they often turn out to be false positive risk indicators. Nevertheless, the following red flags, if ignored continuously, can evolve into long-term risk indicators over time.

i. **Partial Renewals:** Our study shows that when a customer opts for a partial renewal, it is an indicator of risk. This could mean that; the customer might not be generating enough value from their purchased licenses. This sends a signal to the reps that the customer may be dropping more licenses in future.

ii. **Internal Negative Feedback about Customer:** Often global sales leaders have specific information about customers such as: if a customer organization is downsizing, or if the customer has been interacting with competitors. On the other hand, sometimes the brand reps know if a customer is having issues with specific products of their brand. To renewal reps, this information can indicate that this customer might not renew many of their licenses. We consider this as a short-term red flag, as there is no way to track such information over a period.

iii. **Competitive Products:** Market competition is a major renewal risk. During our study, we observed that accounts with smaller revenue streams are usually more prone to move to competitors, as they continuously assess the value being generated by their investments. We list this indicator as a short-term red flag, as the competition moves fast and in an unpredictable way.

2.3.4.2 Long-term Indicators

Long-term risk indicators are analyzed over a period. They are more severe than the short-term indicators as they have already been there for some time but were not alleviated. Following are some risk factors we found that can be long-term indicators:

i. **Consecutive Partial Renewals:** If the results of the previous renewal processes show that the customer is gradually dropping some of their licenses over the last years, a rep usually considers this trend as a serious risk factor.

ii. **Inactive Accounts:** If the online portal of customer accounts does not show any changes over time, such as, no service requests, no download attempts; this is considered as a risk.

iii. **Failing Long-term Relationship with Customers:** Sometimes reps notice that the customers they have been assisting for years, reduce the amount of communication. For example, customers who used to be proactive with their renewals, begin to respond after several emails from the reps. If this goes on for a few consecutive renewals, it can be a serious risk indicator for the reps.

iv. **Brand-specific Indicators:** Renewal reps often find that some products are experiencing significant cancellations over a number of years, despite multiple versions being released over this time period. Sometimes, customers keep complaining about pricing or features of products from a specific brand. This is a sign that the customers will eventually drop the licenses for these products.

2.4 Proposed Research Solutions and Future Research Directions

At the time of presenting our analysis results to a team of content validity experts, we engaged in brainstorming [23] with the experts in order to identify open research directions that can help alleviate the risks and challenges in the license renewal process. Brainstorming is a common group activity that is used by both academics [64], [65] and industry professionals, in order to identify solutions for problems. Brainstorming sessions primarily consist of a facilitator, a scribe, and a number of team members [64], where the facilitator leads the session, the scribe notes down the core concepts of the discussions, and the other team members contribute their insights in solving the problem. Research [64], [65] shows that the key benefit of group brainstorming over individual thinking is, brainstorming helps to elicit many more ideas than an individual can think of. As per the guidelines by Shi et al. [65], we followed the following steps:

Step 1: The facilitator stated the problem – center of Figure 2.5
Step 2: The participants proposed as many solution ideas as possible
Step 3: Ideas were discussed among participants
Step 4: The subset of ideas was evaluated with respect to acceptability, demand, and implementation until a conclusion was decided upon.
Step 5: The steps 1 to 4 were repeated until all the sub-issues were discussed.

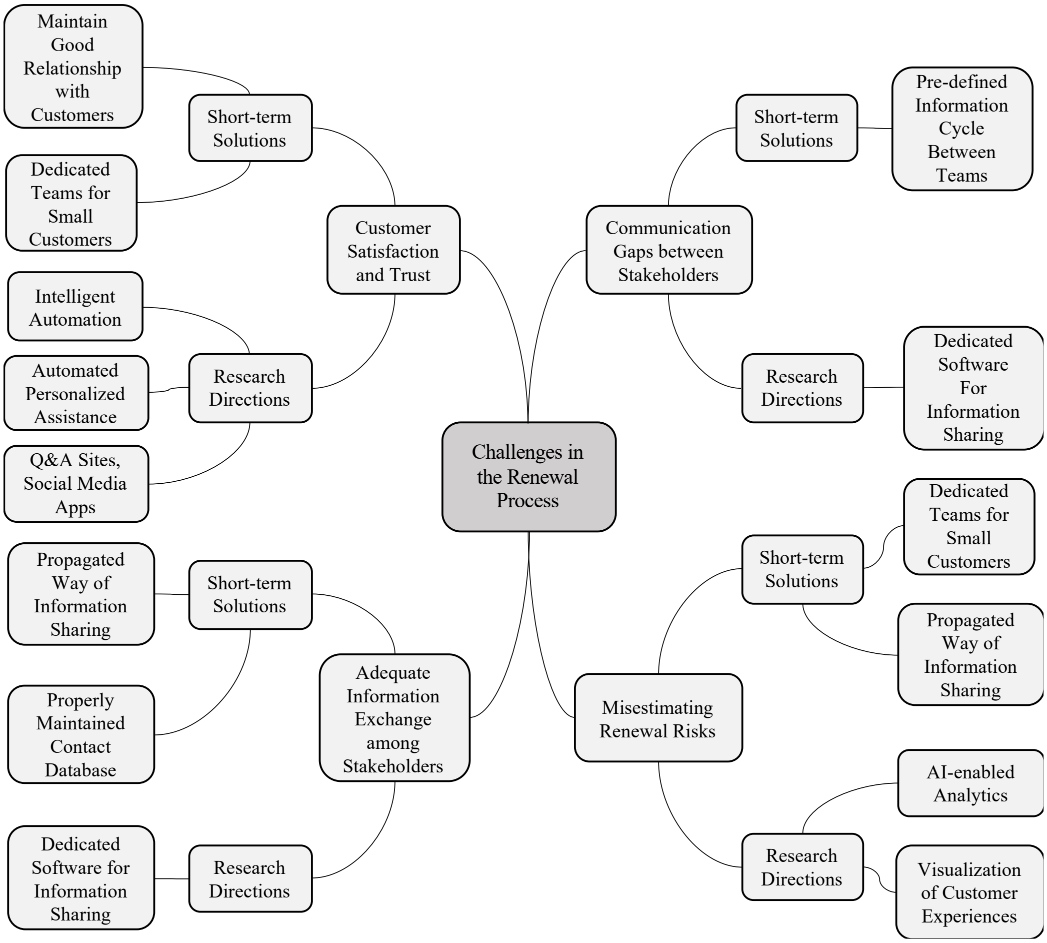

Maintain Good Relationship with Customers
Short-term Solutions
Dedicated Teams for Small Customers
Customer Satisfaction and Trust
Intelligent Automation
Automated Personalized Assistance
Research Directions
Q&A Sites, Social Media Apps
Challenges in the Renewal Process
Propagated Way of Information Sharing
Short-term Solutions
Properly Maintained Contact Database
Adequate Information Exchange among Stakeholders
Dedicated Software for Information Sharing
Research Directions
Short-term Solutions
Pre-defined Information Cycle Between Teams
Communication Gaps between Stakeholders
Research Directions
Dedicated Software For Information Sharing
Dedicated Teams for Small Customers
Short-term Solutions
Propagated Way of Information Sharing
Misestimating Renewal Risks
AI-enabled Analytics
Research Directions
Visualization of Customer Experiences

Figure 2.5: Mind Map of Action Plans and Research Directions for the Challenges in the Renewal Process

As the discussion progressed, the scribe in our team gradually created a Mind-map [64] (see Figure 2.5) of ideas for solving the problems. A mind map is a hierarchical visualization that is used to depict the relationships among conclusions formed during a brainstorming session. Figure 2.5 shows our final mind map with all the challenges and risks and their possible solutions. We divide the conclusions into immediate actions and long-term plans.

2.4.1 Immediate Action Plans

In order to assist the renewal reps in resolving some of the identified challenges within a short time, we came up with the following list of immediate actions.

i. **Maintaining Good Relationship with Customers:** To retain customer trust, it is necessary for the renewal reps to maintain good relationships with customers. The process of building a good relationship can start with renewal reps sending more frequent emails to the customers, to check if the licensed products are satisfactory or not. Alongside, the reps could proactively share information on upcoming feature improvements of the already licensed products of their customers. Research [66], [67] shows that taking these extra steps can help the reps to gain their customers' trust, and eventually can reduce the number of non-responsive customers. Most stakeholders confirmed that the proposed idea is both acceptable and practical. However, in order to integrate the idea in the day-to-day schedule of renewal reps, organizations would need more planning.

ii. **Maintaining Information Cycle Between Teams:** To bridge the communication gap between the different teams involved in the license renewal process, a pre-defined information cycle can be created among them. Valuable customer-specific information from the brand leaders and worldwide leaders can be shared with renewal reps via a quarterly teleconference meeting. So that the reps can be extra careful with these renewals and alleviate any churning risks ahead in time. Researchers [29], [31], [32] have demonstrated that a premeditated information cycle between teams can help in improving the core competencies of each team. As per the experts, this idea can be implemented relatively quickly.

iii. **Dedicated Teams for Small Customers:** Sometimes renewal reps put more focus on customers that generate higher revenue. As a result, often customers with smaller revenue feel neglected and choose to move to competitors. Nevertheless, as cumulative revenue generated by all these customers can be high, teams can be restructured, and dedicated teams can be created for customers belonging to different revenue groups. According to Hannan et al. [34] committed teams can not only help with information building but also can increase customer satisfaction and trust. During our presentations, experts have confirmed that the idea is both acceptable and demanding. However, for a practical implementation of the idea, companies would need to invest in more planning and resources.

iv. **Propagated Way of Collecting and Sharing Information:** A propagated mechanism is required to share information among stakeholders. Such that, for every new customer assigned to a rep, the background analysis for these customers can be done by going through the previously recorded information. The idea seemed acceptable and practical to the stakeholders. However, creating an organization specific tool has its own long-term planning and budget requirements [54]. Nevertheless, in order to implement the idea in a short time, organizations can encourage employee interaction the via social intranet [68] (e.g., Slack, Igloo). Additionally, they can make use of existing industrial information sharing applications, such as the Kanban tool for sharing team workflow information, also common note-taking tools for sharing customer information.

v. **Properly Maintained Contact Database:** Reps often have a hard time to locate the right person who is responsible for a specific client. Hence, as a quick fix to the problem, the organizations can make use of existing industrial contact management applications, such as Pipedrive or Salesforce Marketing Cloud. During our interactions with the stakeholders, we noticed that many organizations already are using some of these solutions, however, they still need to be integrated with the business case of license renewals.

2.4.2 Long-term Research Directions

During our study, we realized that, although the immediate action plans can assist the renewal reps to quickly resolve some of the problems, they may not be sufficient to provide long-lasting solutions. Hence, our suggestion to the participating organizations was to make use of the rapidly changing technology and invest in the future research directions for enhancing the overall renewal process. Hence, we proposed the following set of long-term research directions to the experts, that would not only keep the renewal revenue high but also would be cost-effective for the selling organizations.

i. **Intelligent Automation:** Making use of Intelligent automation (IA) [69] can help with avoiding the challenges of maintaining a high quality of service for all the customers, irrespective of the revenue. Automation of business processes will not only save time on customer onboarding, but it will also provide greater flexibility, efficiency, and security to the organizations. Companies can take advantage of processing the customer data by cloud-based IA representatives [69] at a much lower cost than current manual processing. Nevertheless, research [69] shows that only 35% of all North American businesses have invested in implementing automotive solutions, among them only 19% are I.T. service-based organizations. Most companies are still reluctant in applying IA due to budget, planning, and adaptability reasons [69]. Hence, we think organizations need to invest in further research for implementing IA in the service sector.

ii. **Automated Personalized Assistance:** For the purpose of cost-effectiveness, businesses can make use of automated personalized assistance to provide support to a wider range of accounts. Automated personal assistants [70] are intelligent computer systems, that can perform tasks, or provide services based on a combination of user input such as text, voice commands, location etc. Currently, several multinational I.T. companies, banks, and insurance corporations are using this technology for more efficient marketing and customer support. However, there is a necessity to use this technology in the domain of customer retention.

iii. **Q&A Sites and Social Media Applications:** Q&A sites are websites that follow question-and-answer format to assist end-users to solve problems in different domains. Organizations often use domain-specific Q&A sites for assisting their employees and customers with different technical challenges. Social media applications [71] are web applications that allow their end-users to socially interact and share information with other users of the same application. Although Q&A sites and social media applications exist for almost every multinational organization, they are rarely used for assisting customers with renewing their software licenses. Hence, we propose the use of Q&A sites and social media applications to assist customers with the issues associated with license renewals. For example, while Q&A sites can provide answers to the most frequently asked questions on the renewal process, Social media applications can be helpful for sharing upcoming product features and initiating discussions among the stakeholders on product usability problems.

iv. **Artificial Intelligence Enabled Analytics:**

"By 2020, 85% of customer interactions will be managed without a human" – Gartner [3]

Analysis of real-time customer experience data can help to identify customers at risk [17]. However, manual business analytics, being both expensive and time-consuming [12], cannot provide sufficient insight into customer churn risks. Hence, we propose making use of Artificial intelligence (AI) [4] enabled analytics [2], [14], [15] to identify the customers at non-renewal risk. There has been much research that predicts customer churn using AI enabled predictive models [4], [15] in various domains such as telecommunication, banking, and insurance, among others. Researchers have performed studies that compare [4], [15], [16], [30], [48] and combine [11], [14] multiple AI methodologies, and have identified techniques that focus on predicting customers at churning risk. However, the proposed predictive models suffer from several limitations. Firstly, although the existing models usually are validated using real-life datasets [4], [12], [16], they are seldom used in real-life B2B scenarios [4]. Secondly, since the end-users of the predictions generated by these models are not always data scientists, the interpretation of the models' outputs is often a challenge [72]. Finally, predictive models often generate many false positive predictions [72] that can create additional problems for the renewal reps in real-life B2B scenarios. Hence, we propose that more research is required that addresses the challenges associated to machine learning based predictive models so that they can be used in the industrial license renewal process to identify the customers who are at risk of non-renewal.

v. **Visualization of Customer Experiences:** In order to understand the end-to-end license renewal experience from a customer's point-of-view, the organizations can make use of Visual Analytics (VA) [73]. Mapping the customers' journey [74] using data visualization tools [75], has several benefits. Firstly, it can help to unify all known information regarding a customer that will be useful for analyzing customer satisfaction. Trends can be visible in historical customer transactions that will help the reps to identify renewal risks. Secondly, it will assist organizations in performing behavioral segmentation among customers so that targeted offers and loyalty programs [12] can be created. Much research [73] has been done that presents visualization frameworks for deriving more insights from data. However, most traditional data analytics tools [75] present a static and inflexible model of the data in an incomprehensive way. Hence, we think further research is required in order to develop tools to visualize customer experiences throughout their journey.

vi. **Dedicated Software for Information Sharing:** Our analysis shows that there is a necessity for developing a dedicated software tool to document and share experiences of the stakeholders across the teams that are involved in the renewal process. Hence, in case there is a problem with any renewal, all the stakeholders can be aware at once and act on the problem ahead in time. As discussed in Section 2.4.1, there exist industrial applications that can be used for sharing information within organizations. However, as per Shahzad et al. [54], in many cases, there is a necessity to develop an organization specific tool that can be designed as per the protocols of the organization. Hence, organizations can invest in further research to develop knowledge management tools of their own.

2.5 Limitations and Future Work

Just like any other study, this research also has its own limitations that offer opportunities for future work and threats to the validity of the experimental process. Firstly, research [76] shows that, semi-structured self-reported interviews (cf. Section 2.2.1.1), that were used for collecting our research data, are often subjected to respondent biases [76]. In order to address this limitation, we interviewed 20 participants from different organizations who play different roles in the license renewal process. Hence, we obtained viewpoints from most of the possible stakeholder groups of the renewal process; this allowed us to perform data-source triangulation [77]. On the other hand, our analysis of the obtained information using the Constant Comparative Method (CCM) allowed us to identify any possible inconsistencies and biases in the responses of the interview participants. Moreover, we shared our analysis results with content validity experts in the participating organizations in order to validate the information obtained from the interviews. Nevertheless, as future work, researchers, including the authors, could perform methodological triangulation [77] where the survey data could be collected using multiple methods or instruments [78] other than only semi-structured interviews. This would further reduce these threats.

Secondly, another possible threat to our research is content validity [78]. Content validity is a subjective assessment of completeness of any survey instrument, that refers to the fact: if an instrument or questionnaire used for any study contains all the information it should. In order to mitigate this threat, the five research questions (cf. Section 2.1.1) addressed by this chapter were produced in close collaboration with content validity experts from the participating organizations. However, for future work, one could look into quantitative aspects (such as, variables associated with the Technology Acceptance Model (TAM) [2]) that impact license renewal decisions from customers, in conjunction with the explored qualitative factors.

Moreover, researcher bias [76] of our work could be avoided with investigator triangulation [77], where investigators from both industry and academia could collaboratively explore the different qualitative and quantitative aspects associated with the challenges and risks of software license renewals.

Thirdly, the validity of findings, in qualitative research, is a known threat among researchers [42]. In order to enhance the validity of our analysis results we followed the guidelines of Boeije et al. [22] and implemented systematic data collection and analysis techniques using Grounded theory [21] and Constant Comparative Method (CCM) [18]. These methods are well-known inductive content analysis techniques that are commonly practiced by researchers [21], [22], [39], [47], [48] for qualitative data analysis. Researchers [18], [22] often argue that the coding techniques in CCM are not very well defined in the literature. In order to address this challenge, we performed both internal (i.e., among the researchers) and external (i.e., with content validity experts) validation of our characterizations, theoretical categories, and propositions. Finally, we used Inter-rater Reliability (IRR) [42] as a quantitative measure to justify (cf. Section 2.2.1.1) the validity of the findings. Nevertheless, in future, an empirical study [78] could be performed with a larger population of stakeholders, in order to validate the theories that emerged from our analysis.

Finally, in order to collect information on the end-to-end license renewal process, we interacted with professionals from organizations headquartered in North America. Although, we have clearly distinguished the firmographic [40] variables (cf. Section 2.2.1) of the organizations that participated in our study, in case one or more of these firmographic variables (e.g., location, size of the organization, monetary value of the licenses) are changed, the results might include more factors that impact on renewal decisions from customers. Hence, in future further investigations could be performed to look into organizations with different firmographic backgrounds to generalize the license renewal process and its challenges further.

2.6 Conclusions

Successful renewal of software licenses makes a huge impact on the steady and profitable growth of organizational revenue in the software industry. Therefore, this research tries to answer the question: What is the current state of industrial practice for software license renewals in I.T. service-based organizations? The chapter synthesizes the most common practice of license renewal and the challenges associated with it. The research implements the Grounded theory method, where it adopts a semi-structured, cross-sectional, anonymous, self-reported study across 20 participants from multiple organizations and analyses the obtained information using Constant Comparative Method (CCM). The participants of our study were carefully chosen from several multinational organizations headquartered in North America, with various roles in the renewal process such as, sales and subscription (S&S) representatives, brand leaders, global sales leaders, program directors, and data analysts. Our analysis not only presents the end-to-end license renewal process, but it also shows that lack of effective communication among the stakeholders, scarcity of customer satisfaction, and absence of value generated from the purchased licenses, are among the primary drivers that influence the renewal decisions from customers. To validate our findings from CCM, we used the quantitative measure of inter-rater reliability, where multiple researchers analyzed the same data independently at the same time. Finally, we presented our analysis results to a team of content validity experts in the participating organizations.

We also identified 11 possible risk mitigation strategies by engaging in structured brainstorming with the team of experts. As per the opinions of the experts, the proposed risk mitigation strategies can be classified into short-term action plans and future research directions. For future research, we think that the organizations can take advantage of applying intelligent automation either in the form of chat-bots or as predictive models. We also think that an effective visualization of customers' journey with an organization, can help renewal reps to analyze the overall experience and satisfaction of their customers. This research is, to our knowledge, the first that presents the current state of the license renewal process for software publishing organizations and identifies the risks and challenges associated with it. We think that our identified challenges and proposed research directions can assist software publishing companies to identify the pain-points in their software renewal processes and enhance the procedure to improve the overall renewal rates.

References

[1] F. Marotta-Wurgler, “What’s in a Standard Form Contract? An Empirical Analysis of Software License Agreements,” *Journal of Empirical Legal Studies*, vol. 4, no. 4, pp. 677–713, 2007, doi: 10.1111/j.1740-1461.2007.00104.x.

[2] F. v. Wangenheim, N. V. Wünderlich, and J. H. Schumann, “Renew or cancel? Drivers of customer renewal decisions for IT-based service contracts,” *Journal of Business Research*, vol. 79, pp. 181–188, Oct. 2017, doi: 10.1016/j.jbusres.2017.06.008.

[3] “CRM Strategies and Technologies to Understand, Grow and Manage Customer Experiences,” Gartner, Los Angeles, CA, Gartner Customer 360 Summit, 2011. Accessed: May 25, 2020. [Online]. Available: https://www.gartner.com/imagesrv/summits/docs/na/customer-360/C360_2011_brochure_FINAL.pdf.

[4] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

[5] D. Einhorn, “The Enforceability of Tear-Me-Open Software License Agreements,” *Journal of the Patent and Trademark Office Society*, vol. 67, p. 509, 1985.

[6] K. Jeffery, “Clouds for Science Tackle Challenges Facing Industry and Society,” *IEEE Cloud Computing*, vol. 5, no. 2, pp. 4–6, Mar. 2018, doi: 10.1109/MCC.2018.022171660.

[7] M. Haenlein, “How to date your clients in the 21 st century: Challenges in managing customer relationships in today’s world,” *Business Horizons*, vol. 60, no. 5, pp. 577–586, Sep. 2017, doi: 10.1016/j.bushor.2017.06.002.

[8] F. Wiersema, “The B2B agenda: The current state of B2B marketing and a look ahead,” vol. 4, no. 43, pp. 470–488, 2013.

[9] R. Agarwal and C. E. Helfat, “Strategic renewal of organizations,” *Organization Science*, vol. 20, pp. 281–293, 2009.

[10] A. Schmitt, S. Raisch, and H. W. Volberda, “Strategic Renewal: Past Research, Theoretical Tensions and Future Challenges: Strategic Renewal,” *International Journal of Management Reviews*, vol. 20, no. 1, pp. 81–98, Jan. 2018, doi: 10.1111/ijmr.12117.

[11] N. Lu, H. Lin, J. Lu, and G. Zhang, “A Customer Churn Prediction Model in Telecom Industry Using Boosting,” *IEEE Trans. Ind. Inf.*, vol. 10, no. 2, pp. 1659–1665, May 2014, doi: 10.1109/TII.2012.2224355.

[12] A. Aluri, B. S. Price, and N. H. McIntyre, “Using Machine Learning To Cocreate Value Through Dynamic Customer Engagement In A Brand Loyalty Program,” *Journal of Hospitality & Tourism Research*, vol. 43, no. 1, pp. 78–100, Jan. 2019, doi: 10.1177/1096348017753521.

[13] E. Ascarza, S.A. Neslin, O. Netzer, Z. Anderson, P.S. Fader, S. Gupta, B.G. Hardie, A. Lemmens, B. Libai, D. Neal, and F. Provost, “In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions,” *Cust. Need. and Solut.*, vol. 5, no. 1–2, pp. 65–81, Mar. 2018, doi: 10.1007/s40547-017-0080-0.

[14] K. Coussement and D. Van den Poel, “Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,” *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, Jan. 2008, doi: 10.1016/j.eswa.2006.09.038.

[15] M. Hassouna, A. Tarhini, T. Elyas, and M. S. Abou Trab, “Customer Churn in Mobile Markets: A Comparison of Techniques,” *IBR*, vol. 8, no. 6, p. p224, May 2015, doi: 10.5539/ibr.v8n6p224.

[16] Yizhe Ge, Shan He, Jingyue Xiong, and D. E. Brown, “Customer churn analysis for a software-as-a-service company,” in *2017 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, Apr. 2017, pp. 106–111, doi: 10.1109/SIEDS.2017.7937698.

[17] W. Bi, M. Cai, M. Liu, and G. Li, “A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn,” *IEEE Trans. Ind. Inf.*, vol. 12, no. 3, pp. 1270–1281, Jun. 2016, doi: 10.1109/TII.2016.2547584.

[18] S. M. Kolb, “Grounded Theory and the Constant Comparative Method: Valid Research Strategies for Educators,” *Journal of emerging trends in educational research and policy studies,* vol. 3, no. 1, pp. 83-89, Feb 2012.

[19] N. Guhr, B. Lebek, and M. H. Breitner, “The impact of leadership on employees’ intended information security behaviour: An examination of the full-range leadership theory,” *Info Systems J*, vol. 29, no. 2, pp. 340–362, Mar. 2019, doi: 10.1111/isj.12202.

[20] P. B. Lowry, J. Zhang, C. Wang, and M. Siponen, “Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model,” *Information Systems Research*, vol. 27, no. 4, pp. 962–986, Dec. 2016, doi: 10.1287/isre.2016.0671.

[21] B. G. Glaser, “The Constant Comparative Method of Qualitative Analysis,” *Social Problems*, vol. 12, no. 4, pp. 436–445, 1965, doi: 10.2307/798843.

[22] H. Boeije, “A Purposeful Approach to the Constant Comparative Method in the Analysis of Qualitative Interviews,” *Quality and quantity*, vol. 36, no. 4, pp.391-409, 2002

[23] J. G. Rawlinson, *Creative thinking and brainstorming*. Routledge, 2017.

[24] F. Marotta-Wurgler, “Competition and the Quality of Standard Form Contracts: The Case of Software License Agreements,” *Journal of Empirical Legal Studies*, vol. 5, no. 3, pp. 447–475, 2008, doi: 10.1111/j.1740-1461.2008.00130.x.

[25] T. Tuunanen, J. Koskinen, and T. Kärkkäinen, “Automated software license

analysis," *Autom Softw Eng*, vol. 16, no. 3–4, pp. 455–490, Dec. 2009, doi: 10.1007/s10515-009-0054-z.

[26] F. M. Kifetew, M. Morandini, D. Munante, A. Perini, A. Siena, and A. Susi, "Goal-aware Analysis of Software License Risks," p. 6.

[27] M. Corporation, "Annual Report," Microsoft Corporation, 2017. [Online]. Available: https://www.microsoft.com/investor/reports/ar17/index.html.

[28] I. Corporation, "Annual Report," IBM Corporation, 2017. [Online]. Available: https://www.ibm.com/annualreport/index.html.

[29] E. Suter, J. Arndt, N. Arthur, J. Parboosingh, E. Taylor, and S. Deutschlander, "Role understanding and effective communication as core competencies for collaborative practice," *Journal of Interprofessional Care*, vol. 23, no. 1, pp. 41–51, Jan. 2009, doi: 10.1080/13561820802338579.

[30] P. S. H. Leeflang, P. C. Verhoef, P. Dahlström, and T. Freundt, "Challenges and solutions for marketing in a digital era," *European Management Journal*, vol. 32, no. 1, pp. 1–12, Feb. 2014, doi: 10.1016/j.emj.2013.12.001.

[31] M. Leonard, S. Graham, and D. Bonacum, "The human factor: the critical importance of effective teamwork and communication in providing safe care," *Qual Saf Health Care*, vol. 13, no. Suppl 1, pp. i85–i90, Oct. 2004, doi: 10.1136/qshc.2004.010033.

[32] M. A. Campion, G. J. Medsker, and A. C. Higgs, "Relations Between Work Group Characteristics and Effectiveness: Implications for Designing Effective Work Groups," *Personnel Psychology*, vol. 46, no. 4, pp. 823–847, 1993, doi: 10.1111/j.1744-6570.1993.tb01571.x.

[33] N. S. Hill and K. M. Bartol, "Empowering Leadership and Effective Collaboration in Geographically Dispersed Teams," *Personnel Psychology*, vol. 69, no. 1, pp. 159–198, 2016, doi: 10.1111/peps.12108.

[34] S. Hannan, B. Suharjo, K. Kirbrandoko, and R. Nurmalina, "International Journal of Economic Perspectives, 2017, Volume 11, Issue 1, 344-353.," vol. 11, no. 1, p. 10, 2017.

[35] G. Walsh, M. Schaarschmidt, and S. Ivens, "Effects of customer-based

corporate reputation on perceived risk and relational outcomes: empirical evidence from gender moderation in fashion retailing," *Journal of Product & Brand Management*, vol. 26, no. 3, pp. 227–238, Jan. 2017, doi: 10.1108/JPBM-07-2016-1267.

[36] P. Guenzi, L. L. M. De, and R. Spiro, "The combined effect of customer perceptions about a salesperson's adaptive selling and selling orientation on customer trust in the salesperson: a contingency perspective," *Journal of Business & Industrial Marketing*, vol. 31, no. 4, pp. 553–564, Jan. 2016, doi: 10.1108/JBIM-02-2015-0037.

[37] L. Marakanon and V. Panjakajornsak, "Perceived quality, perceived risk and customer trust affecting customer loyalty of environmentally friendly electronics products | Elsevier Enhanced Reader," *Kasetsart Journal of Social Sciences*, vol. 38, pp. 24–30, 2017, doi: https://doi.org/10.1016/j.kjss.2016.08.012.

[38] U. S. C. Bureau, "North American Industry Classification System 2017," *U.S. Department of Commerce*, 2017. https://www.census.gov/cgi-bin/sssd/naics/naicsrch?code=511210&search=2017%20NAICS%20Search (accessed May 25, 2020).

[39] B. G. Glaser, A. L. Strauss, and E. Strutzel, "The discovery of grounded theory; strategies for qualitative research," *Nursing research*, vol. 17, no. 4, p. 364, 1968.

[40] J. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology*, vol. 13, no. 1, pp. 3-21, 1990.

[41] A. Sbaraini, S. M. Carter, R. W. Evans, and A. Blinkhorn, "How to do a grounded theory study: a worked example of a study of dental practices," *BMC Med Res Methodol*, vol. 11, no. 1, p. 128, Dec. 2011, doi: 10.1186/1471-2288-11-128.

[42] J. D. Olson, C. McAllister, L. D. Grinnell, K. G. Walters, and F. Appunn, "Applying Constant Comparative Method with Multiple Investigators and Inter-Coder Reliability," *Qualitative Report,* vol. 21, no. 1, Jan 2016.

[43] S. M. Fram, “The Constant Comparative Analysis Method Outside of Grounded Theory,” *Qualitative Report*, vol. 18, no. 1, 2013.

[44] J. Hewitt-Taylor, “Use of constant comparative analysis in qualitative research,” *Nursing Standard*, vol. 15, no. 42, pp. 39–42, Jul. 2001, doi: 10.7748/ns2001.07.15.42.39.c3052.

[45] T. Mettler and R. Winter, “Are business users social? A design experiment exploring information sharing in enterprise social systems,” *Journal of Information Technology*, vol. 31, no. 2, pp. 101–114, Jun. 2016, doi: 10.1057/jit.2015.28.

[46] M. Lombard, J. Snyder-Duch, and C. C. Bracken, “Content Analysis in Mass Communication: Assessment and reporting of intercoder reliability,” *Human communication research*, vol. 28, no. 4, pp.587-604, 2002.

[47] B. Bodenmann and K.W. Axhausen, “Synthesis report on the state of the art on firmographics,” Institute for Transport Planning and Systems, ETH, Zurich, p. 31, 2010.

[48] A. Tamaddoni Jahromi, S. Stakhovych, and M. Ewing, “Managing B2B customer churn, retention and profitability,” *Industrial Marketing Management*, vol. 43, no. 7, pp. 1258–1268, Oct. 2014, doi: 10.1016/j.indmarman.2014.06.016.

[49] G. D. Moody and M. Siponen, “Using the theory of interpersonal behavior to explain non-work-related personal use of the Internet at work,” *Information & Management*, vol. 50, no. 6, pp. 322–335, Sep. 2013, doi: 10.1016/j.im.2013.04.005.

[50] A. Vance, M. Siponen, and S. Pahnila, “Motivating IS security compliance: Insights from Habit and Protection Motivation Theory,” *Information & Management*, vol. 49, no. 3–4, pp. 190–198, May 2012, doi: 10.1016/j.im.2012.04.002.

[51] V. Venkatesh, J. Y. L. Thong, and X. Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” *MIS Quarterly*, vol. 36, no. 1, pp. 157–178, 2012, doi:

10.2307/41410412.

[52] D. L. Segal, F. L. Coolidge, A. O'Riley, and B. A. Heinz, "Structured and semistructured interviews," *Clinician's handbook of adult behavioral assessment*, pp. 121–144, 2006.

[53] A. Rindfleisch, A. J. Malter, S. Ganesan, and C. Moorman, "Cross-Sectional versus Longitudinal Survey Research: Concepts, Findings, and Guidelines," *Journal of Marketing Research*, vol. 45, no. 3, pp. 261–279, Jun. 2008, doi: 10.1509/jmkr.45.3.261.

[54] B. Shahzad, A. M. Abdullatif, N. Ikram, and A. Mashkoor, "Build Software or Buy: A Study on Developing Large Scale Software," *IEEE Access*, vol. 5, pp. 24262–24274, 2017, doi: 10.1109/ACCESS.2017.2762729.

[55] M. Westergaard and F. M. Maggi, "Looking into the future," *In OTM Confederated International Conferences on the Move to Meaningful Internet Systems*, Springer, Berlin, Heidelberg, pp. 250-267, Sep 2012.

[56] C. Urquhart and W. Fernández, "Using Grounded Theory Method in Information Systems: The researcher as blank slate and other myths," *In Enacting Research Methods in Information Systems*, vol. 1, Palgrave Macmillan, Cham, pp. 129-156, 2016.

[57] H. E. Tinsley and D. J. Weiss, "Interrater reliability and agreement of subjective judgments.," *Journal of Counseling Psychology*, vol. 22, no. 4, pp. 358–376, 1975, doi: 10.1037/h0076640.

[58] K. S. Kurasaki, "Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data," *Field Methods*, vol. 12, no. 3, pp. 179–194, Aug. 2000, doi: 10.1177/1525822X0001200301.

[59] K. A. Neuendorf and A. Kumar, "Content analysis," *The International Encyclopedia of Political Communication*, vol. 1, pp. 221–230, 2002.

[60] J. L. Campbell, C. Quincy, J. Osserman, and O. K. Pedersen, "Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement," *Sociological Methods & Research*, vol. 42, no. 3, pp. 294–320, Aug. 2013, doi: 10.1177/0049124113500475.

[61] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[62] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968, doi: http://dx.doi.org/10.1037/h0026256.

[63] K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” *Tutor Quant Methods Psychol*, vol. 8, no. 1, pp. 23–34, 2012.

[64] M. R. Shirey, “Brainstorming for Breakthrough Thinking:,” *JONA: The Journal of Nursing Administration*, vol. 41, no. 12, pp. 497–500, Dec. 2011, doi: 10.1097/NNA.0b013e3182378a53.

[65] Y. Shi, “Brain storm optimization algorithm,” presented at the In International conference in swarm intelligence, Berlin, Heidelberg, 2011, pp. 303–309.

[66] F. L. Eichorn, “Internal Customer Relationship Management (IntCRM) A Framework for Achieving Customer Relationship Management from the Inside Out,” *Problems and Perspectives in Management*, p. 24.

[67] N. Umashankar, M. K. Ward, and D. W. Dahl, “The Benefit of Becoming Friends: Complaining after Service Failures Leads Customers with Strong Ties to Increase Loyalty,” *Journal of Marketing*, vol. 81, no. 6, pp. 79–98, Nov. 2017, doi: 10.1509/jm.16.0125.

[68] A. Sharma and J. Bhatnagar, “Enterprise social media at work: web-based solutions for employee engagement,” *Human Resource Management International Digest*, vol. 24, no. 7, pp. 16–19, Jan. 2016, doi: 10.1108/HRMID-04-2016-0055.

[69] I. Hawkins, “The future is happening: a global snapshot of IA,” PEX Network, May 2018.

[70] K. A. Olsen and A. Malizia, “Automated Personal Assistants,” *Computer*, vol. 44, no. 11, pp. 112–111, Nov. 2011, doi: 10.1109/MC.2011.329.

[71] T. Mettler, “Contextualizing a professional social network for health care:

Experiences from an action design research study," *Info Systems J*, vol. 28, no. 4, pp. 684–707, Jul. 2018, doi: 10.1111/isj.12154.

[72] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 3504–3512, Accessed: May 25, 2020. [Online]. Available: http://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism.pdf.

[73] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, Oct. 2000, pp. 69–75, doi: 10.1109/INFVIS.2000.885092.

[74] K. N. Lemon and P. C. Verhoef, "Understanding Customer Experience Throughout the Customer Journey," *Journal of Marketing*, vol. 80, no. 6, pp. 69–96, Nov. 2016, doi: 10.1509/jm.15.0420.

[75] G. Andrienko, N. Andrienko, and S. Wrobel, "Visual analytics tools for analysis of movement data," *SIGKDD Explor. Newsl.*, vol. 9, no. 2, pp. 38–46, Dec. 2007, doi: 10.1145/1345448.1345455.

[76] E. G. Guba, "Naturalistic inquiry," *Improving Human Performance Quarterly*, vol. 8, no. 4, pp. 268–76, 1979.

[77] N. Carter, D. Bryant-Lukosius, A. DiCenso, J. Blythe, and A. J. Neville, "The Use of Triangulation in Qualitative Research," *Oncology Nursing Forum*, vol. 41, no. 5, pp. 545–547, Sep. 2014, doi: 10.1188/14.ONF.545-547.

[78] F. Shull, J. Singer, and D. I. K. Sjøberg, Eds., *Guide to advanced empirical software engineering*. London: Springer, 2008.

Chapter 3

A Comprehensive Review of Exploratory Data Analysis Tools

In today's digital world, insights obtained from Exploratory Data Analysis (EDA) are used in strategic business decision making. EDA [1] is a fundamental procedure that makes use of statistical techniques and graphical representations in order to obtain insights from data [2]. EDA not only assists with the identification of hidden patterns and correlations among attributes in data, but also helps with the formulation and validation of hypotheses from the data. Over the last few decades, interactive visualization strategies have become an integral part of data exploration and analysis techniques [3]. With a picture being worth a thousand words, academics have proposed several tools and techniques [4]–[9] to visualize complex relationships among data attributes using diagrams and charts. Whilst some of these visual data analysis tools [10]–[14] assist with domain-specific analysis (for example, analysis of genome-sequence data [10], meteorological data [11], results of predictive analysis [14] etc.), other tools [3], [15], [16] focus on general purpose exploratory browsing of tabular data. In either case, since the beginning of visual interactive data analysis [3], almost all visual EDA tools perform a few common analytics tasks. In their work, Heer et al. [17], as well as Amar et al. [18], have identified these basic data exploration tasks as sort, filter, aggregate, correlate, group, and derive attributes.

Nevertheless, in recent years, the requirements for exploratory data analysis have changed significantly. With the ever-growing size and types of data to be analyzed, scalability and analysis duration [3], [5] of the EDA tools are now among the primary concerns of researchers. Moreover, with data being used to train predictive models [14] for making strategic business decisions, analysts are in need of data exploration tools that can help to accurately analyze complex multivariate relationships [1], [19] in datasets, with limited available analytical expertise. To address these challenges, EDA tools are constantly evolving [15], [20]. In the last few years, many advancements have taken place with the design of data visualization tools [5], [21]–[23] in order to address different challenges [24]–[27] of analyzing large datasets [2], [27]–[30]. However, the trade-off between the depth and the breadth of analysis supported by the modern exploratory data visualization tools is still a challenge [3]. As, on the one hand, despite covering the breadth of basic exploration tasks [18], general purpose data exploration tools [9], [31] often do not fulfill the in-depth analysis requirements of their users. On the other hand, tools [21] that focus on highly scalable and in-depth multivariate analysis, often lack in interpretability and require significant knowledge of the problem domain.

To identify the current state of research in the emerging field of EDA, at first, we examine a real-world dataset with 3.4 million records obtained from our industrial partner IBM. From this investigation, we identify a set of additional exploratory requirements specific to resolving the challenges of analyzing such enormous business data. Later, we investigate 50 visual interactive EDA tools (cf. Section 3.2.2) for their ability to assist with the traditional EDA process steps, along with their fulfillment of the identified additional exploratory requirements for large scale EDA. Among the 50 analyzed tools, 43 are proposed by academic researchers and the remaining 7 are commercial tools used in industry. Since, performing a complete survey of each and every existing EDA tool would be too large to cover in a single research, we carefully define precise selection criteria (cf. Section 3.1.1) for the selected tools. For example, whilst for academic tools we only look at the ones that were presented within the last five years and help with general purpose exploration of tabular data, for commercial tools we follow the guidelines of Gartner Inc. and select the business intelligence platforms that received Gartner Customer Choice Awards in the year 2017. During our evaluation of the selected tools, we identify some gaps and research opportunities in the emerging field of visual EDA.

Although there has been much research [10], [11], [14], [32]–[35] that aims at surveying the state-of-the-art in visual data analytics, in most cases the studies consider exploration tools for specific domains [10]–[12]. Moreover, as per our knowledge, this novel work is at least more than a year ahead of its closest competitors [3], [15], [16], [35]–[37] as it also considers tools that were proposed in the last one year. It also presents a list of 50 visualization tools that were analyzed for the first time from the perspective of the steps followed in EDA [1], [26]. The primary contributions of this research are as follows:

- This novel work presents the current state of research on visual EDA tools for exploring tabular data by investigating 50 tools for their utility in the EDA process steps (cf. Section 3.2.2.1).
- The research also evaluates the selected tools for their abilities to fulfill different additional exploratory requirements of large industrial datasets (cf. Section 3.2.2.2).
- The work identifies open research opportunities for the domain of visual EDA tools (cf. Section 3.3).

The rest of the chapter is organized as follows: in Section 3.1, we present the scope and methodology of this research. In Section 3.2, with a discussion on different aspects of our analyzed industrial dataset, we summarize the findings of our survey. Section 3.3 presents identified research opportunities and gaps in the field of EDA tools, whilst Section 3.4 presents surveys that are similar to our research. In Section 3.5, we list the limitations of this research, while in Section 3.6 we conclude the chapter.

3.1 Research Scope and Methodology

To precisely define the scope of our research, in this section, at first we present our primary research questions for this work. Next, we outline a specific set of inclusion and exclusion criteria of the visual data analytics tools included in our study. Finally, we discuss the detailed steps that were followed to analyze the industrial dataset and to perform the state-of-the-art survey of EDA tools.

3.1.1 Research Scope

In this section we outline the boundaries of our study in terms of investigation time-frame, purpose, and popularity of the analyzed EDA tools. Therefore, we enlist our research questions as follows.

> ***RQ1:*** *What are the additional exploratory requirements for EDA tools to investigate large industrial datasets?*
>
> ***RQ2:*** *What research activities have taken place in last five years in the domain of visual EDA tools for general purpose exploration of tabular data?*
>
> ***RQ3:*** *What are the most popular commercial EDA tools in industry?*
>
> ***RQ4****: To what extent do modern EDA tools assist with the steps of the EDA process and fulfill the additional exploratory requirements of analyzing large datasets (i.e., answers of RQ1)?*
>
> ***RQ5:*** *What are the gaps and future directions for the current state of research on visual EDA tools?*

Based on the known challenges [34], [38]–[40] of analyzing large datasets, researchers [34], [38] have proposed a set of additional requirements for analyzing such data. However, work that addresses all possible challenges of large datasets, is sparse. Hence, with RQ1, we investigate a real-world tabular dataset and identify different challenging aspects of this dataset. Later, based on existing literature [21], [23], [34], [37], [40]–[42] that relates the identified aspects to specific data analysis requirements, we identify four additional exploratory requirements for analyzing large industrial datasets.

Figure 3.1 illustrates our decisions for RQ2 and RQ3 in detail. As shown in the figure, for RQ2, limiting the analysis timeframe for academic EDA tools to five years was one of our very first decisions in this research. The reasons behind this decision was: technology trend analysis for five years is a common industrial practice. Moreover, we fixed our focus on investigation of tabular data stored in relational databases, because as discussed by researchers [3], most business data are stored in relational databases despite of being initially recorded as plain text, XML, or graphs. We also narrowed our focus on tools used for general purpose exploration of tabular data. The reason being, due to the existence of large number of EDA tools in every research field (such as, time-series, geo-spatial, or genomic data etc.), it would not be feasible to cover all these fields in one research. Additionally, with a focus of investigating tools with a smooth learning curve for novice users, we chose to exclude data analytics libraries, frameworks, and packages that require programming skills from end-users.

In today's data-centric world, almost all businesses make use of general-purpose commercial Business Intelligence (BI) and analytics tools for performing EDA tasks to gather insights from data. As shown in Figure 3.1, with respect to RQ3, we selected the seven most popular tools that were awarded by Gartner Inc. in 2017. Our primary purpose of investigating commercial tools were to identify the similarities and differences between the current state of academic research and industrial practice.

To summarize, tools fulfilling the following criteria were included in our study (cf. Figure 3.1):

- Presented within the last five years. (criterion applicable only for academic tools)
- Focused on analyzing tabular data stored in relational databases.
- Focused on general purpose exploratory analysis of data.
- Most popular and widely used (criterion applicable only for commercial tools)

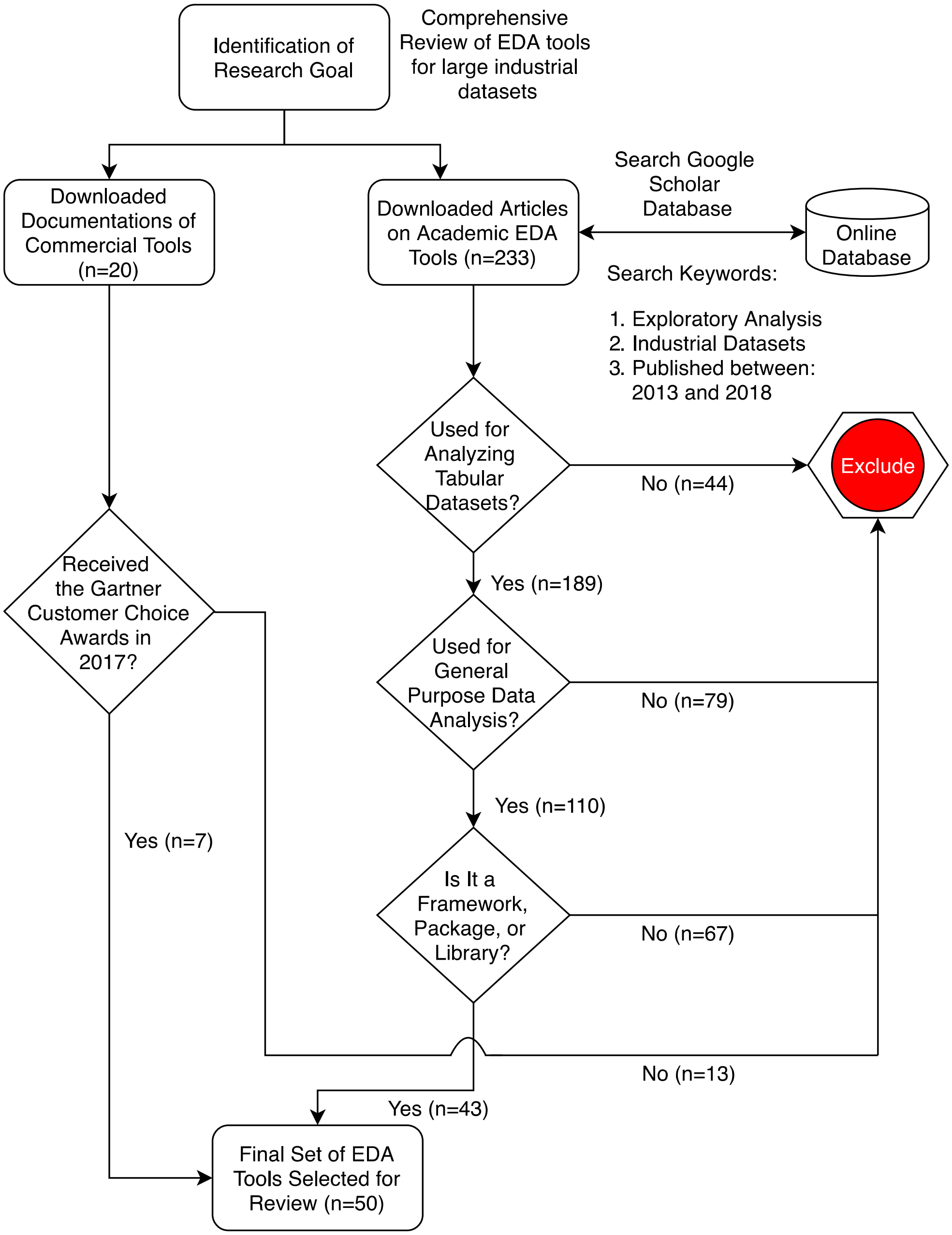

Identification of Research Goal
Comprehensive Review of EDA tools for large industrial datasets
Downloaded Documentations of Commercial Tools (n=20)
Downloaded Articles on Academic EDA Tools (n=233)
Search Google Scholar Database
Online Database
Search Keywords:
1. Exploratory Analysis
2. Industrial Datasets
3. Published between: 2013 and 2018
Used for Analyzing Tabular Datasets?
No (n=44)
Exclude
Yes (n=189)
Received the Gartner Customer Choice Awards in 2017?
Used for General Purpose Data Analysis?
No (n=79)
Yes (n=110)
Yes (n=7)
Is It a Framework, Package, or Library?
No (n=67)
No (n=13)
Yes (n=43)
Final Set of EDA Tools Selected for Review (n=50)

Figure 3.1: Flow-diagram of Selection Criteria for the State-of-the-Art EDA Tools

On the other hand, tools were excluded from the study based on the following criteria:

- Domain specific visual exploratory analysis tools.

- Frameworks, packages, or libraries for performing visual EDA tasks.

Like every other process, EDA consists of a set of steps (cf. Section 3.2.2). With RQ4 we aimed at investigating the utility of the selected tools at these different steps. Also, we intended to investigate the extent to which the tools fulfill the additional exploratory requirements (i.e., answers of RQ1) for analyzing large datasets. At the end of our study, we aimed at seeking answers for RQ5 and identifying gaps and research opportunities in the current state of research on visual EDA.

3.1.2 Research Methodology

In this section, we discuss different steps of our research methodology in detail. This section is primarily divided into three subsections. The first subsection presents our analysis methodology for the real-world dataset, whilst in the next two subsections, we discuss the detailed processes of collection and analysis of the selected EDA tools for this research.

3.1.2.1 Background Analysis of an Industrial Dataset

The industrial dataset analyzed in this work is comprised of 3.4 million records with 27 attributes and contains product license renewal information. It is important to note that we only had access to a completely anonymized version of the dataset. The dataset was provided to us in Comma Separated Values (CSV) format and was created by joining five different DB2 tables from an IBM data server. These tables contained information such as sales figures, product details, customer interaction details, and types of product licenses. The tabular dataset was investigated by a group of two researchers (the first and second authors) using Microsoft Excel. At this time, we performed different data manipulation tasks such as: plotting the value distributions of attributes and finding correlations among attributes. We also generated a pivot table from the data that enabled us to sort and filter the attribute values, so that we could compare the maximum, minimum, mean, and standard deviations [1] of each attribute. During these tasks, we identified a set of challenging aspects (cf. Section 3.2.1) of the analyzed dataset. Once, these challenges were identified, the two researchers looked into the literature [21], [23], [34], [37], [39], [41], [42] that is associated with these identified challenges [34, 38, 39]. Based on the literature evidence, we created a list (cf. Section 3.2.1) of additional exploratory requirements for large scale EDA tools.

3.1.2.2 Data Collection for Systematic Literature Review

In order to address RQ2, we carried out a manual search of conference proceedings and journals that are known to publish novel ideas on data visualization techniques. The article sources were chosen not only based on their impact factors in the EDA community, but also because they have been popularly chosen by researchers [3], [15], [34] for performing similar studies. As the next step, the last five years' archive for each of the identified journals and conferences were scrutinized by the two researchers. As shown in Figure 3.1, during this task, the researchers collected each and every article from the identified journals into a pool of 233 articles that were relevant to EDA. Later the collected articles were filtered by the researchers based on the inclusion and exclusion criteria (cf. Figure 3.1) defined for the tools. During this step, 190 articles were excluded from our study. In cases of conflicts between the two researchers regarding an article's eligibility to be included in the study, a third researcher was brought in to resolve the disagreements. In parallel, for addressing RQ3, the two researchers started investigating on the most popular commercial exploratory data analysis tools in industry (cf. Figure 3.1). Later, following the judgement of Gartner Inc. the researchers selected 7 commercial EDA tools. For this study, we considered both the winners and the honorable mentions of the customer choice awards. Once the selected EDA tools were finalized, a quality assessment was performed by a team of three researchers (the first three authors) involved in this work, where the fulfillment of the inclusion and exclusion criteria for each of the selected tools was validated. During the quality assessment session, the team also confirmed if the systematic review has covered all relevant EDA tools from the selected journals and conference proceedings that it should. Once the tools to be analyzed were finally chosen, the following information was extracted regarding the tools:

- The source journal or conference proceedings of the tool and its year of publication.
- The research questions addressed by the tool and its primary focus.
- The EDA steps supported by the tool along with its additional supported features.

3.1.2.3 Data Analysis for Systematic Literature Review

While reviewing the chosen EDA tools, following the guidelines of Kitchenham et al. [43], at first both the researchers thoroughly read the articles for each tool, later for the tools [4], [6], [8], [9], [30], [44], [45] that provide open source access to their implementations, the two researchers independently executed the source code of these tools. Among the tools that were executed, whilst some [8], [21], [30], [44]–[46] allowed their source code to be downloaded to our local systems, some other tools [4], [6] only presented a live executable version that require users to upload their datasets on the tools' servers. Due to the strict Data Access Policy requirements from IBM, we applied our analyzed industrial dataset only to those academic tools [8], [21], [30], [44]–[46] that allowed us to download their source code. For the tools [4], [6], [9] that did not enable us to download any code, we executed the tools using the sample datasets on the tools' websites. In case of the tools [5], [25], [31] that did not share any source code information, the two researchers thoroughly reviewed the main articles of the tools. For commercial tools however, we could download all seven of the tools [20], [47]–[52] and applied them on our industrial dataset. At the end of the analysis, both researchers discussed their findings to derive a final evaluation for each tool. In case of disagreements, the third researcher helped to resolve the conflicts. Finally, the group of three researchers collaboratively derived a summary table (cf. Table 3.1) with the evaluation of the identified EDA tools. Later, the researchers identified the gaps and open research areas (cf. Section 3.3) in the emerging field of visual EDA.

3.2 Results

In this section, we discuss the primary results of our research in detail. We begin with a brief description of findings from analyzing the industrial dataset, followed by a detailed discussion on the results of our systematic literature review of the chosen EDA tools.

3.2.1 Elicitation of Additional Exploratory Requirements for Large Industrial Datasets

This section presents our analysis results of the industrial dataset obtained from IBM, where we first highlight the challenging aspects of the dataset, then we present a list of additional exploratory requirements for large scale EDA tools.

i. **High Dimensionality:** The dimensionality of a tabular dataset usually refers to the number of independent variables or attributes in the data. High dimensionality of large business datasets is a known challenge among researchers [40], [53]. As, firstly, the computational workload for analyzing a dataset increases as the number of dimensions grows [54]. Secondly, as a result of dimensional redundancy [54], some attributes in a high-dimensional dataset might not be as useful as others. For example, in our industrial dataset, there were three attributes representing the country code of customers from three different viewpoints. In these situations, strong correlations can be noticed [33] among the redundant dimensions that can be difficult to visualize. Finally, high-dimensional datasets cause "geometrical insanity" [40] when visually exploring the data. For example, as the dimension changes only from 2D to 3D, the data that could initially be represented by a 1-dimensional line now becomes a 2-dimensional surface. Hence, when the dimension increases from 3D to 4D and further, it gets extremely challenging to visualize such dimensionality in the data.

ii. **Categorical Attributes:** The second primary aspect of an industrial tabular dataset is the large number of categorical attributes in the data (precisely, in our dataset 19 among the 27 attributes were categorical). Research [1] shows that analysis of categorical features in a dataset can be a primary challenge due to reasons such as:

a. Performing statistical analysis on categorical attributes is more challenging than the numeric attributes, as some of the measures of centrality (such as, mean and median) and dispersion (such as, variance) applies only to numerical data. Also, in case the categories are not relative, sorting them according to an ascending or descending order can be a challenging task. Hence, it becomes difficult for data analysts to perform any normality tests [1] on the categorical features.

b. Analysis of categorical features with too many categories can result in performance challenges [40] for any data analysis tool. Also, often for these features, there are some categories that are more dominant; such that, whilst the dominant categories account for the majority of the data points, the remaining categories represent extremely small portion of the data in comparison to the dominant categories. In such situations, it gets immensely challenging [55] to perform univariate analysis [1] of the categorical features.

iii. **Missing or Aberrant Values and Outliers:** The data points with missing values show the incompleteness of the data. As we discovered in the dataset, many data points had missing values for attributes that did not have a NOT NULL constraint in the original database tables. On the other hand, the records with outliers or aberrant values show inconsistency in the data. In our dataset, some aberrant values (such as, 9999 in place of date values) represented some undocumented codes for missing data. The outliers in the dataset were either results of human errors in data input or indicated calculation errors when deriving attribute values. In either case, given the enormous size of the dataset, the outliers were among our main challenges for exploring this dataset.

iv. **Data Sanity:** In the industrial dataset, we noticed that the dataset being created by merging different tables not only had some columns with ambiguous names but also had columns with inconsistent values. For example, whilst some ambiguous column names represented abbreviations of long sentences (e.g., FYCA standing for: First Year of Contract Agreement), some other column names represented organization specific terminology with internal meaning. On the other hand, the inconsistency in values for some columns resulted from different tables storing the values for the same attribute in different formats. We noticed these inconsistencies in, attributes containing date information and financial details. The data sanity problems made us realize that a significant amount of expertise is required to understand the values of each attribute in the data.

v. **Multivariate Relationships:** Attributes in business datasets contain complex multivariate relationships that are not easily visible in tabular data. Whilst, in some cases the values of an attribute depend on two or more other attributes, in some other cases combined exploration of several attributes can provide more meaningful insights than exploration of an individual attribute. For example, in our dataset, the attribute containing the information on the next purchase date depended on the attributes: previous purchase date, product type, and business value of customers. On the other hand, combined exploration of customer industry, type of purchased products, and product pricing information gave us insights on the pricing requirements of customers in different industries. So, it can often get challenging to identify the attributes that are related to each other without appropriate domain knowledge and training.

vi. **Anonymity:** Another aspect of a real-world industrial dataset is anonymity that can cause challenges during the data analysis process. In large multinational organizations, much data is classified business information that is only shared with specific teams and individuals. In such cases, even the data analysts do not get access to the entire information about the dataset. For example, in case of our dataset, attributes such as product pricing or customer firmographic information were anonymized that lead us to some misinterpretation of the data.

vii. **Large Scale of Data Points:** One of the primary aspects of real-business data is the large scale of data points in the datasets. In our case, the dataset with 3.4 million rows and 27 columns resulting into 91.8 million data values took hours to be extracted from the database into CSV. Hence, we think it will take longer time for any EDA tool to visualize such amount of data.

From our analysis, we believe that in order to efficiently analyze any industrial dataset, in addition to supporting the EDA process steps [1], EDA tools need to address the above-mentioned challenges. Research [34], [38], [39] shows that, each of these identified challenges of big-data analytics can be associated to specific exploratory requirements of modern EDA tools. Following the existing research results [19], [21], [23], [34], [37], [41], [42], we identify four additional exploratory requirements namely: (i) scalability, (ii) reduced analytical expertise, (iii) user-engagement, and (iv) interpretability. Figure 3.2 summarizes the relations between the identified challenges and the additional exploratory requirements.

As shown in Figure 3.2, researchers such as Najafabadi et al. [38] and Chan et al. [19] have associated the aspects of high-dimensionality, and large scale of data points to the scalability requirements of EDA tools. The reason being, both the aspects refer to the size and complexity of a dataset [23], [37], [42], and hence signify the necessity for scalability in EDA tools. According to Wang et al. [34], in the absence to support for scalability, a tool cannot be used for analyzing large industrial datasets. Hence, we consider scalability as our first additional exploratory requirement. Moreover, according to Tufféry et al. [1], the analysis of categorical attributes and multivariate relationships among attributes can require significant analytical expertise. Hence, based on existing research [34], [38] reduced analytical expertise is chosen as our next requirement for EDA tools. Researchers [26], [34] have also shown that, whilst the results of multivariate analysis can be challenging to interpret, the presence of poor data sanity, anonymity, missing values, and

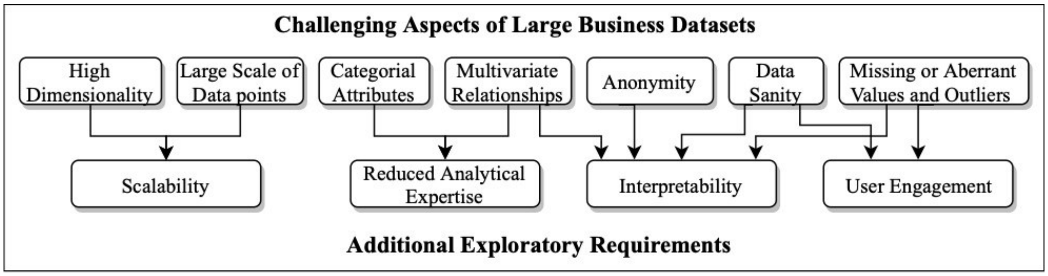

Challenging Aspects of Large Business Datasets
High Dimensionality
Large Scale of Data points
Categorial Attributes
Multivariate Relationships
Anonymity
Data Sanity
Missing or Aberrant Values and Outliers
Scalability
Reduced Analytical Expertise
Interpretability
User Engagement
Additional Exploratory Requirements

Figure 3.2: Elicitation of Additional Exploratory Requirements for Large Scale EDA Tools

outliers require additional support for interpretability from EDA tools. Finally, in order to rectify the sanity issues of large datasets, EDA tools need enable user engagement [34], [38], [39] in the form of user feedback.

3.2.2 Survey of Exploratory Data Analysis Tools

In this section, we present the results of our systematic literature review that answers our research question RQ4 (cf. Section 3.1.1). We begin this section with the evaluation results of the chosen EDA tools for their ability to assist with the EDA process [1]. Later, we discuss our findings on the tools' fulfillment of the additional exploratory requirements for large scale EDA (cf. Section 3.2.1).

3.2.2.1 Support for Traditional EDA Process Steps

According to Tufféry et al. [1] and Demiralp et al. [26], the EDA process usually follows six distinct steps (cf. Figure 3.3) namely: (i) Distinguish Attributes, (ii) Univariate Data Analysis, (iii) Detect Interactions Among Attributes, (iv) Detect Missing & Aberrant Values, (v) Detect Outliers, and (vi) Feature Engineering. As shown in Figure 3.3, the analysis begins with identification of attributes in a dataset that gives a clear understanding of the data to be analyzed. Next, in order to understand individual attributes and their relationships with each other, univariate, bivariate, and multivariate analyses are performed. Later, cleaning and data preparation tasks are carried out, where missing, aberrant values and outliers [24] are detected and imputed. The process ends with feature engineering, where

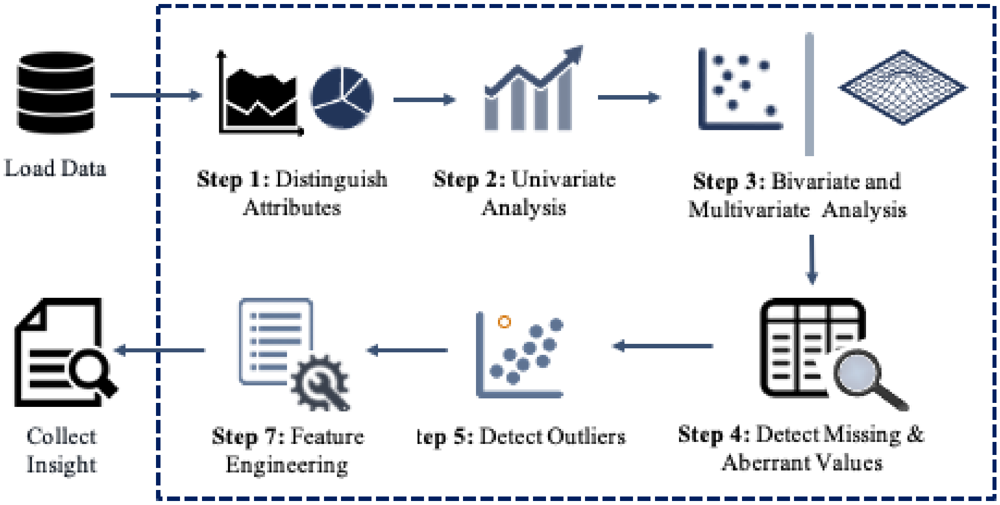

Load Data
Step 1: Distinguish Attributes
Step 2: Univariate Analysis
Step 3: Bivariate and Multivariate Analysis
Step 4: Detect Missing & Aberrant Values
tep 5: Detect Outliers
Step 7: Feature Engineering
Collect Insight

Figure 3.3: Fundamental Steps of the Exploratory Data Analysis Process

features are transformed or combined to generate new features. We summarize our analysis results in Table 3.1.

i. **Distinguish Attributes:** Exploratory data analysis begins with identification of the attributes in a dataset. This is an essential step at the beginning of the EDA process that not only helps with the "Cold-start" [2], [20] problem of data analysis, but it also assists users to formulate clear analysis goals. According to researchers [3], datasets commonly have numerical (or quantitative) or categorical (or qualitative) attributes [1]. However, not all statistical analysis techniques can be applied to all the attributes in a dataset [56]. Hence, it is important for data analysts to clearly distinguish and understand the meaning of each attribute in a dataset prior to analyzing the data.

   Most existing commercial data visualization tools such as Microsoft Power BI [48] and IBM Watson Analytics [20], show the entire dataset in a tabular format and allow users to see and modify the data in terms of attribute

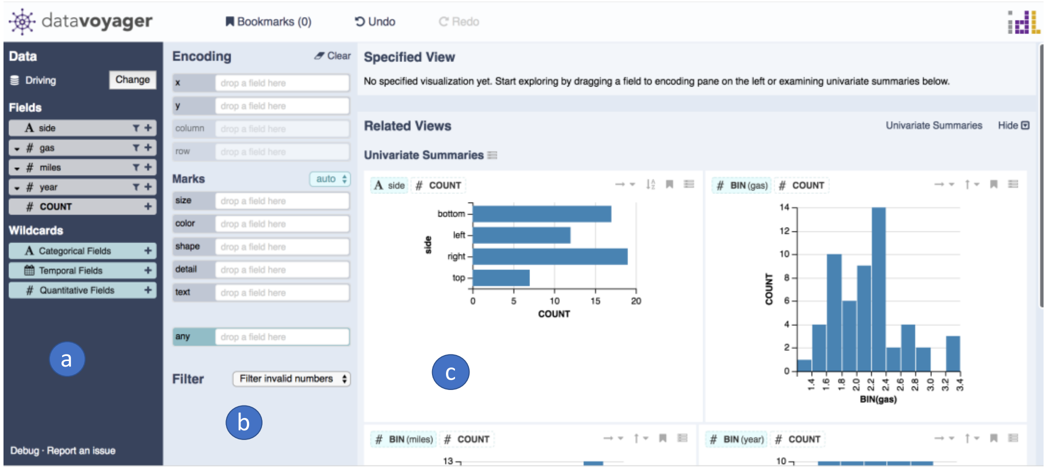
Table 3.1: Summary of Investigated Exploratory Data Analysis Tools

Note: In the following table 'y' represents 'supports the operation' and blank spaces represent the opposite. In case of commercial tools[1]: GA – Gold Award, SA – Silver Award, BA – Bronze Award, HM – Honorary Mention

				Traditional Steps of EDA Process							Additional Focus Areas			
						Detect relations hips								
No.	Type	Tools	Ordered by	Distinguish Attributes	Univariate Analysis	Bivariate Analysis	Multivariate Analysis	Detect Missing Value	Detect Outliers	Feature Engineering	Scalability	Interpretability	Reduced Domain Expertise	User Engagement
1	Academic Tools	DataScope (Iyer, 2017)	2018	y	y	y				y	y			
2		DataSite (Cui, 2018)	2018		y	y						y	y	
3		Duet (Law, 2018)	2018	y	y	y		y				y	y	
4		FastMatch (Macke, 2018)	2018		y						y			
5		InfoNice (Wang Y. Z., 2018)	2018		y	y				y				y
6		Keshif (Yalcin, 2016)	2018	y	y	y	y			y	y			
7		NorthStar (Kraska, 2014)	2018	y	y	y	y	y	y	y	y		y	
8		Podium (Wall, 2018)	2018	y	y	y	y		y	y				y
9		RCLens (Lin, 2018)	2018	y	y	y	y		y	y			y	
10		Taco (Niederer, 2018)	2018		y						y			
11		Taggle (Furmanova, 2017)	2018		y	y		y		y				y
12		VisComposer (Mei, 2018)	2018	y	y	y				y				y
13		Voder (Srinivasan, 2018)	2018		y	y			y	y	y		y	y
14		Zenvisage (Siddiqui, 2016)	2018	y	y	y				y			y	y
15		Analyza (Dhamdhere, 2017)	2017	y	y	y	y			y	y		y	y
16		ChartAccent (Ren D. B., 2017)	2017		y	y				y		y		
17		ForeSight (Demiralp, 2017)	2017	y	y	y			y	y	y	y	y	y
18		GaussianCubes (Wang Z. F., 2017)	2017	y	y	y	y	y		y	y			
19		HindSight (Feng, 2017)	2017		y	y				y	y			y
20		MyBrush (Koytek, 2017)	2017		y	y		y	y	y				
21		VisFlow (Yu, 2017)	2017		y	y	y		y	y	y	y		y
22		Voyager 2 (Wongsuphasawat, 2017)	2017	y	y	y	y		y	y			y	y
23		AggreSet (Yalcin, 2016)	2016		y	y					y			y
24		DimScanner (Xia, 2016)	2016	y	y	y	y			y	y			

[1] https://www.gartner.com/reviews/customer-choice-awards/analytics-business-intelligence-platforms

No.	Type	Tools	Ordered by	Traditional Steps of EDA Process							Additional Focus Areas			
						Detect relations hips								
				Distinguish Attributes	Univariate Analysis	Bivariate Analysis	Multivariate Analysis	Detect Missing Value	Detect Outliers	Feature Engineering	Scalability	Interpretability	Reduced Domain Expertise	User Engagement
25	Academic Tools	ForeCache (Battle, 2016)	2016		y	y	y			y	y		y	y
26		VisTrees (El-Hindi, 2016)	2016		y	y				y	y			
27		SeeDB (Vartak, 2015)	2015	y	y	y	y		y	y	y	y	y	y
28		Sketch (Budiu, 2015)	2015		y	y				y	y			
29		Bertifier (Perin, 2014)	2014	y	y	y			y		y			
30		Domino (Gratzl S. G., 2014)	2014	y	y	y	y	y		y	y			
31		Ellipsis (Satyanarayan, 2014)	2014		y	y				y				y
32		iVisDesigner (Ren D. H., 2014)	2014		y	y			y	y				
33		Lyra (Satyanarayan, 2014)	2014	y	y	y				y				
34		PanoramicData (Zgraggen, 2014)	2014	y	y	y				y				y
35		Progressive Insights (Stolper, 2014)	2014		y	y	y			y	y			
36		UpSet (Lex, 2014)	2014		y	y					y			
37		ExPlates (Javed, 2013)	2013	y	y	y	y			y				y
38		imMens (Liu, 2013)	2013		y	y	y		y		y			
39		LineUp (Gratzl S. L., 2013)	2013		y	y	y	y	y	y	y			y
40		PivotSlice (Zhao, 2013)	2013	y	y	y	y			y				
41		SketchStory (Lee, 2013)	2013		y	y				y			y	y
42		VisDeck (Perry, 2013)	2013		y	y							y	y
43		VisReduce (Im, 2013)	2013		y	y				y	y			y
44	Commercial Tools	Alteryx (Sallam, 2014)	GA	y	y	y	y	y	y	y	y		y	y
45		Tableau (Tableau, n.d.)	SA	y	y	y	y	y	y	y	y		y	y
46		Domo (Domo, n.d.)	BA	y	y	y	y						y	y
47		Watson Analytics (Kelly, 2015)	HM	y	y	y	y	y	y	y	y	y	y	y
48		MS Power BI (Corp., n.d.)	HM	y	y	y	y	y	y	y	y		y	
49		QlikView (QlikView, n.d.)	HM	y	y	y				y				y
50		Sisence (Sisense, n.d.)	HM	y	y	y	y			y	y		y	y



No specified visualization yet. Start exploring by dragging a field to encoding pane on the left or examining univariate summaries below.

Figure 3.4: Dashboard of the Tool Voyager 2 (Wongsuphasawat et al. [30])

Note: (a) the dashboard shows the names of attributes while dividing them in categories such as quantitative, categorical and temporal. (b) the panel assists with bivariate and multivariate analysis by allowing users to choose filters and embellishments. (c) the panel shows univariate summaries of all attributes.

names, attribute values, and data types. Among academic EDA tools, while some tools such as Keshif [57], Explates [28], NorthStar [6], DimScanner [58], and Analyza [59] present only a list of attribute names to the user, tools such as Podium [60], ForeSight [26] and Bertifier [61] present a portion of data in tabular format at the beginning of the analysis process. On the other hand, tools such as Voyager 2 [30], Taggle [44], Zenvisage [62], and LineUp [63] provide visual overviews of all attributes immediately at the beginning of the analysis. In most cases [30], [44], [59], [60], an initial summary uses a variety of interactive histograms to present an overview of each attribute. For example, Figure 3.4 shows a snapshot of the tool Voyager 2 [30], where the parts (a) and (c) are relevant to distinguishing attributes. In the figure, the section marked by (a) gives an example of distinguishing attributes at the beginning of the analysis process. Whereas, the section (c) depicts the visual summaries of each attribute.

Moreover, some existing EDA tools [64] provide more detailed summaries of attributes. For example, while the tool Domino [9] summarizes attribute information such as datatypes, number of records, and dimensions, Taggle [44] provides a short description of the dataset with an HTML link to the data source. Among the commercial EDA tools, Domo [49] provides a brief summary of datatypes and groups attributes based on their types. On the other hand, some data exploration tools such as, Taco [45], [65] and Domino [9] do not describe any attributes in the dataset at all, and begin with complex data exploration tasks (e.g., join, merge, aggregate etc.) right after the data is loaded.

ii. **Univariate Data Analysis:** Once the attributes in a dataset are identified, it is necessary to perform univariate analysis [1] in order to get a deeper understanding of each attribute. Univariate analysis also allows the determination of attribute combinations for subsequent analysis. It helps with detection of details such as: centrality (i.e., mean, median, and mode) and dispersion (i.e., range, variance, standard deviation, skewness, and kurtosis) of attributes in the data. While the centrality measures help us determine an approximate average for the attribute values, the dispersion measures help us identify the spread of the value between its lowest and highest bounds. Univariate analysis is also used to identify missing values or outliers in a dataset and to discretize continuous variables [1], [66].

Most recent advancements in data exploration tools facilitate univariate data analysis. Typically, in both academic and commercial tools [26], [57], interactive histograms and boxplots are used to depict value distributions of the variables. For example, as shown in the part (c) of Figure 3.4, the tool Voyager 2 uses interactive histograms to depict the value distributions of the attributes. Additionally, commercial EDA tools such as IBM Watson Analytics [67], [68], Microsoft Power BI [48], QlikView [50], Alteryx [52], and academic tools such as Voyager 2 [30], DataSite [2], Northstar [6], and ForeSight [26], let users choose from a set of optional visual representations (such as, heat-maps, pie-charts, line-graphs etc.) and in some cases visual embellishments [69] (such as, color, texture etc.) to better analyze each attribute. For example, as depicted in the part (b) of Figure 3.4, Voyager 2 allows users to select details such as shape, size, color etc. for the visualizations. In most cases, modern visualization tools such as Keshif [4], NorthStar [6], Voder [24], Tableau [47] allow end-users to interactively brush [4], [25], hover [4], [70], and zoom [9], [21], [71] on the visualizations. While, aggregation [44] of feature values is one of the most common ways [1] of performing univariate analysis, most EDA tools also support sorting and filtering [45], [71], [72] of attribute values.

iii. **Detect Interactions Among Attributes:** After the univariate analysis of each attribute, the next step is to understand the relationships among different attributes in the dataset. This not only helps to identify incompatibilities among attribute values, but it also enables analysts to generate optimal feature combinations [6], [49] for future analysis. Analysis of attribute relationships can be performed in two different ways: bivariate and multivariate statistics [1]. Whereas, bivariate statistics only analyses the association of a chosen pair of attributes, the intersection of more than two variables are analyzed using multivariate statistics. As per Tufféry et al. [1], bivariate analysis needs to be performed prior to multivariate analysis. This way, once the users have a clear idea of the compatibility of an attribute pair, they can combine more attributes with them, for further analysis.

   a. **Bivariate Statistics:** In modern EDA tools, interactive filtering and aggregation of attributes are the most common ways [5], [8], [24], [47], [57], [67], [73] of performing a combined analysis of two attributes. Typically, the vast majority of all the exploratory data visualization tools perform bivariate data analysis. In some tools (e.g., Voder [24], Taggle [44], Domino [9], MyBrush [69], DataScope [42], ForeSight [26]), filtered and aggregated attribute values are usually obtained by interactive brush-and-link [8], [31]

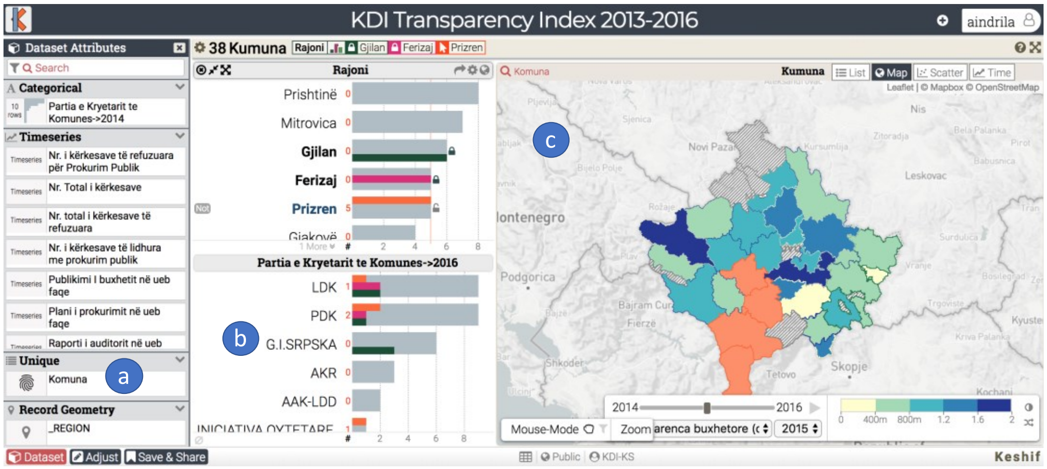


Figure 3.5: Dashboard of the Tool Keshif (Yalçın et al. [4])

Note: (a) Keshif enlists the attributes in the dataset in groups such as categorical, quantitative, time-series data. (b) For bivariate and multivariate analysis Keshif allows users to lock histograms of up to three attributes. (c) Attribute relationships are also shown on visual representations that allow users to switch to different visuals and/or filter the data.

operations and are presented using highlighted and interactive histograms [8], [44], [45], [74]. These histograms use different colors and/or textures to represent correlations among attributes. Moreover, tools like Keshif [4] allow users to lock histograms of specific variables and compare them to other variables. Figure 3.5 shows a snapshot of both univariate and bivariate analysis using Keshif. Whereas, part (a) in Figure 3.5 presents individual attributes in groups based on their datatype, the upper half of part (b) depicts bivariate relationships among attributes using overlapped and locked histograms, and part (c) shows univariate analysis with filter operation.

Some tools perform different variations of the brush-and-link operations in order to correlate attributes. For example, VisTrees [5] requires users to explicitly link two attributes prior to performing the brush and filter operations. VisFlow [8] makes users select two attributes and pass them through a binder component before the brush and filter operations can be performed. NorthStar [6] links the two attributes and creates a scatter plot that shows the correlations among a pair of attributes. The tool MyBrush [69] on the other hand, focuses entirely on brushing and linking attributes. It provides a unified interface for interactively configuring different components of the brush-and-link operation namely: source, link, and target. Some tools such as PanoramicData [25], Tableau [47], iVisDesigner [70], Voder [24], DataSite [2], IBM Watson Analytics [20], [67] allow users to compose different visuals (such as, scatter-plots and pie-charts) other than just histograms to perform bivariate data analysis. Tools such as Tableau [47], IBM Watson Analytics [68], Alteryx [52] also enable users to perform join operations on multiple related tables in the same database.

Some EDA tools [8], [21], [44], [75] analyze horizontal subsets of data. These subsets are often created either based on user-driven selections [44], or algorithmic analysis [21]. Horizontal data subsets are used in many different visualization tools to achieve different goals. For example, whereas FastMatch [74] uses subset sampling to analyze the histograms of all attributes in a dataset and finds the top-k similar histograms among them. Taggle [44] allows end-users to create hierarchical aggregations of data subsets in order to create nested attributes. Domino [9], on the other hand, describes data subsets as ‘blocks’ and depicts relationships (e.g., strong or weak) among the blocks. Duet [72] makes use of data subsets to perform pairwise comparison among tabular data. Figure 3.6 depicts an example of bivariate analysis using the tool Domino [9], where the relationships between data subsets are presented by parallel coordinates and scatter plots.

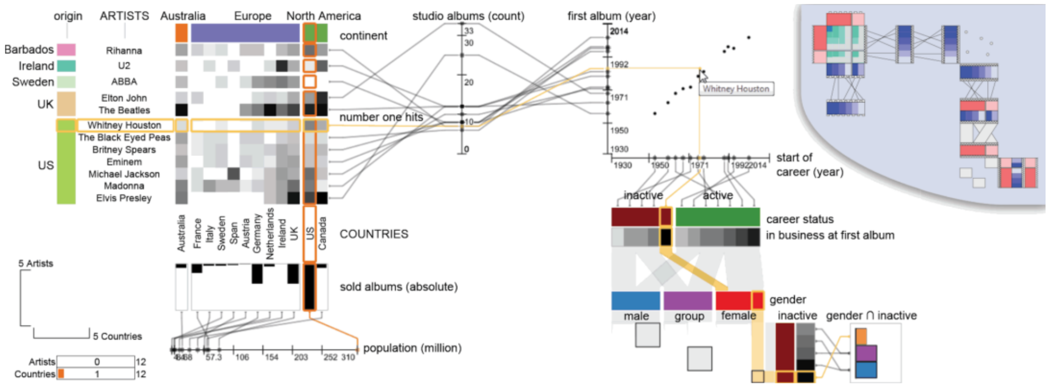


Figure 3.6: Domino (cf. Gratzl et al. Fig. 1 [9]) showing the relationship between data subsets using parallel coordinates and scatter plots

b. **Multivariate Statistics:** Once a pair of relevant attributes in a dataset are analyzed, the next step is to perform a deeper investigation, where more attributes are added with the analyzed pair for a combined exploration. Research [1] shows that analysis of the correlation among more than two attributes is a complex and time-consuming task, that can only be achieved by factor analysis techniques such as clustering [61] and dimensionality reduction [2]. As a result, in order to avoid the complexity of visualizing the results of factor analysis, most of the modern EDA tools depict relationships among multiple attributes using group and filter operations. For example, in cases of tools such as PanoramicData [25], Keshif [4], iVisDesigner [70], bivariate histograms and scatter plots are filtered using one or more attributes to show the relationships among all these features. An example of multivariate analysis using 2-dimensional histograms is presented in Figure 3.5 (i.e., the lower half of the part (b)), where three histograms of different colors are used to compare the values of three different attributes.

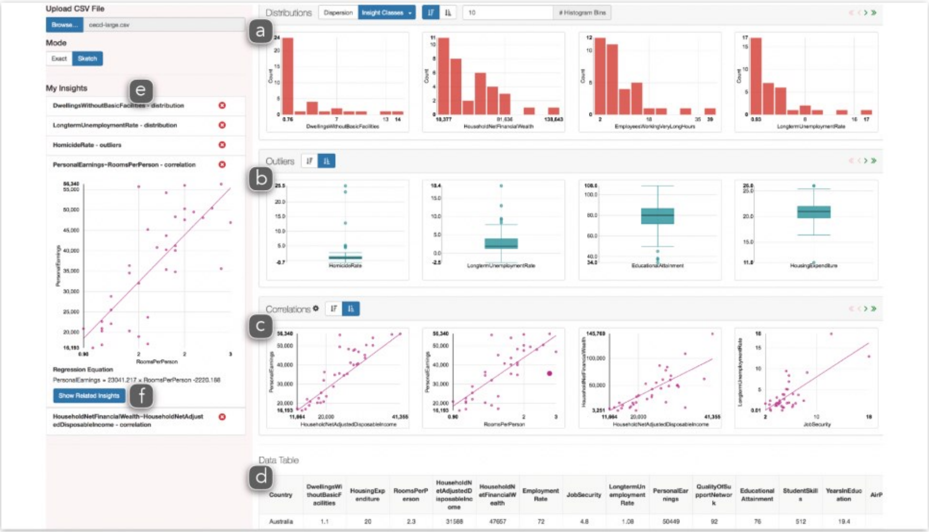
Nevertheless, despite of the complexity of multivariate statistics, some of the analyzed EDA tools implement factor analysis tasks.

For example, PivotSlice [7] uses multi- dimensional query mechanisms to generate faceted exploratory visualizations; and VisTrees [5] allows users to create multi-dimensional indexes in order to combine feature subsets with each other. Moreover, tools such as GaussianCube [23], imMens [21], Podium [60] and LineUp [63] enhanced the scalability of EDA process with the use of dimensionality reduction [2] techniques. For example, imMens generates data cubes [21] from the binned aggregation of data that is further transformed into multi-variate data tiles; whereas GaussianCube [23] improves on imMens by precomputing the best multivariate Gaussian distribution among attributes. On the other hand, LineUp [63] and Podium [60] make use of multi-attribute rankings based on attribute combinations. Finally, using multi-attribute ranking, LineUp [63] allows end-users to alter attribute combinations or column rankings to compare the differences in the relationships.

iv. **Detect Aberrant & Missing Values:** Aberrant and missing values may result in biased analysis of data [1]. Aberrant values are erroneous values which occur as a result of incorrect user inputs or calculation errors, whilst missing values occur in a dataset during data extraction and/or data collection. Detection of such values in a dataset usually happens right after multivariate analysis, when the user has a clear idea about the value ranges of the attributes and their compatibilities. In case of a large dataset, the search for missing and aberrant values begins when any abnormality is noticed in the univariate, bivariate, or multivariate visualizations. Once data-points with aberrant or missing values are detected, usually the first action of the data analysts is to remove these data points [1]. However, removing data-points can have its own consequences. Firstly, there can be a large number of data-points for which at least one attribute value is missing. Secondly, the dataset might have special significance for the data-points with missing values. Hence, removal of observations with missing and/or aberrant values can add further bias into the analysis. According to Tufféry et al. [1], alternatives to deletion of records with missing values are: to perform value imputations, or to include the data with missing values in the analysis with a known margin of error. Imputations of missing values can either be user driven [1] or automatically performed with the help of predictive models [14].

Although, some tools such as Keshif [4] and AggreSet [57] allow the user to temporarily remove some attributes from analysis, except for a few tools such as IBM Watson Analytics [20], GaussianCubes [23], and MyBrush [69] most of the analyzed tools do not allow users to detect or modify aberrant values in the dataset. Tools such as Podium [60], ForeSight [26] (cf. Figure 3.7), and Bertifier [61] allow users to visualize missing values in the data in tabular format, however, these tools require users to manually scroll through the entire table in order to identify the missing values. Despite scalability challenges, these tools allow users to perform user driven imputations on the missing values. None among our analyzed the tools perform any automatic imputation of missing or faulty data.

v. **Detect Outliers:** The detection of outliers usually happens during or after the univariate, bivariate, or multivariate analysis. An outlier is an observation that deviates further away [1] from other observations in the dataset. Like aberrant values, outliers can also add bias to the analysis leading to misinterpretation of attribute properties. According to researchers [76], outliers in a dataset can be primarily of three types namely: univariate, bivariate, and multivariate outliers. Therefore, usually after multivariate analysis and detection of aberrant values, users focus on the detection of outliers. While univariate outliers can be detected by calculation of the Inter-Quartile Range (IQR) [56] of individual variables, to detect bivariate and multivariate outliers, analysts need to inspect correlations among different attributes. For example, bivariate outliers can be detected using



Upload CSV File

oecd-large.csv

Mode

My Insights

DwellingsWithoutBasicFacilities - distribution

LongtermUnemploymentRate - distribution

HomicideRate - outliers

PersonalEarnings~RoomsPerPerson - correlation

Regression Equation

PersonalEarnings = 23041.217 × RoomsPerPerson -2220.188

HouseholdNetFinancialWealth~HouseholdNetAdjustedDisposableIncome - correlation

Data Table

Country	DwellingsWithoutBasicFacilities	HousingExpenditure	RoomsPerPerson	HouseholdNetAdjustedDisposableIncome	HouseholdNetFinancialWealth	Employment Rate	JobSecurity	LongtermUnemployment Rate	PersonalEarnings	QualityOfSupportNetwork	Educational Attainment	StudentSkills	YearsInEducation	AirP
Australia	1.1	20	2.3	31588	47657	72	4.8	1.08	50449	92	76	512	19.4	

Figure 3.7: Dashboard of ForeSight (cf. Demiralp et al. – Fig. 1 [26])

combining two attributes and calculating their correlation coefficient [26], whereas multivariate outliers can be detected using factor analysis [1]. The complexity of visualization of outliers in a dataset also depends on the type of outlier. Whilst, univariate and bivariate outliers can be easily depicted using boxplots, interactive histograms and scatter plots [6], it is often challenging for visual EDA tools to depict multivariate outliers. Figure 3.7 shows an example of the tool ForeSight [26], where the detection of univariate and bivariate outliers is depicted in parts (b) and (c) respectively.

As per our analysis, some of the modern EDA tools such Inflow [8], ForeSight [26] (cf. Figure 3.7), Podium [60], RCLens [77], DimScanner [58], HindSight [78], and IBM Watson Analytics [67], allow their users to detect univariate and bivariate outliers in a dataset. Just like the missing and aberrant values, once the outliers in a dataset are detected, they can be rectified by either removing the observations, performing automatic or user-driven imputations, or transformation of variables [9], [44].

vi. **Feature Engineering:** Finally, after obtaining detailed insights about the dataset, as the last step of the EDA process feature engineering is carried out. Feature engineering is a core step of exploratory data visualization [1] that is performed by almost all EDA tools [2], [5], [6], [57]. It is primarily divided into two parts: variable creation and transformation. The creation of derived variables often happens to ease the data analysis process. Derived variables not only summarize linear relationships among many attributes, but they also help to simplify the understanding of complex attributes in the dataset. Variable transformations convert complex non-linear relationships into linear relationships; and standardize values to obtain a better understanding. Normalization [1] is a type of variable transformation that helps to convert skewed distributions into more symmetric distributions. Among the tools we analyzed, FastMatch [74] identifies similarities between different distributions by comparing the relative values. Most visualization tools, such as Keshif [4], NorthStar [6], Voyager 2 [30] use binning or categorization strategies to split up continuous variables into categories. This operation is known as discretization [1]. Nevertheless, none of the analyzed tools propose any mechanism to analyze the error [66] added by the discretization tasks in the EDA process.

3.2.2.2 Support for Additional Exploratory Requirements

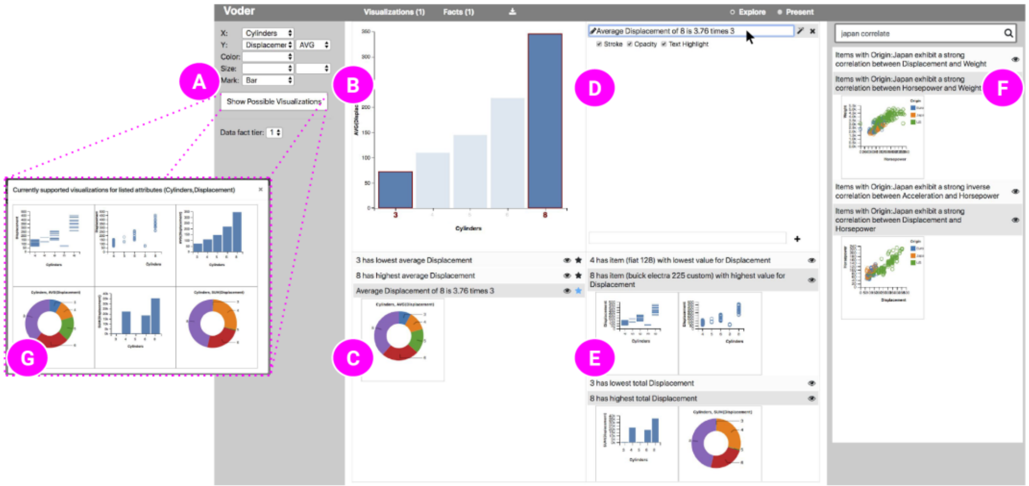
In this section, we present our evaluation of each of the analyzed tools with respect to their fulfilment of the four additional exploratory requirements (cf. Section 3.2.1). For each requirement, we discuss the different ways each of the analyzed tools have addressed this requirement. We summarize the results of our analysis as follows:

i. **Scalability:** Scalability of exploratory visualization tools primarily has two aspects: firstly, loading the entire dataset into the main memory, secondly, processing the data and producing visual representations of the attribute relationships (i.e., the response time of the tool). In the case of academic tools, researchers have attempted to address both of these aspects. For example, in order to address the challenge of a large set of raw data that does not fit into main memory, tools like ForeCache [22] use a client-server architecture, where a middleware layer fetches portions of data ahead in time based on the analysis history of the user. On the other hand, EDA tools make use of several different techniques to assist with the response time for processing very large datasets. For example, tools such as ProgressiveInsights [79], NorthStar [6], and VisReduce [41] progressively create incremental visual representations of the data and provide incremental updates to notify the user of the wait time. Alternatively, tools such as ForeSight [26] and ProgressiveInsights [79] provide approximate visualizations with a known boundary of error. Other tools make use of creating subsets from the data in on order to achieve scalability. For example, tools such as Taggle [44], Domino [9], GaussianCubes [23], FastMatch [74] and imMens [21] make use of horizontal data subsets for this purpose. In case of commercial tools, almost all the EDA tools [47], [51], [52] analyzed during this work, support highly scalable analytics.

With respect to the scalability of the tools in each step of the EDA process; only a few tools [6], [23], [26], [59] focus on distinguishing attributes. For example, tools such as ForeSight [26] and Microsoft Power BI [48] present attribute names in a tabular form, and the tools such as NorthStar [6], Tableau [47], and Domino [9] group attributes into categories. Scalability in univariate and bivariate analysis is supported by most EDA tools that allow large scale analysis. For this purpose, the tools such as ProgressiveInsights [79], NorthStar [6], and VisReduce [41] constantly refine partially loaded univariate and bivariate analysis charts of attributes. Moreover, to provide support for scalable multivariate analysis, tools such as imMens [21], and GaussianCubes [23] precompute multivariate data tiles. On the other hand, scalable identification of missing, aberrant values, and outliers is supported by some EDA tools [6], [9], [23], [63]. Whilst in most cases [6], [24], the outliers are presented using graphical representations such as box-plots or scatter plots, the missing values are presented either in tabular form [6] or using visual encodings [44]. Finally, the scalability of feature engineering [1] depends on the scalability of univariate and bivariate analysis in the analyzed EDA tools.

ii. **Reduced Analytical Expertise:** In order to help non-expert users to explore data, researchers [2], [26], [30] have proposed proactive visual recommender systems that can ease the learning curve for novice users. During this study, we noticed three different types of recommendations: (i) recommendation of charts [46], (ii) recommendation of actions [2], and (iii) recommendation of questions [68]. Among these, recommendation of charts is the most common and is offered by many tools such as SeeDB [46], Voyager 2 [30], VizDeck [80], Tableau [47], Analyza [59], Alteryx [52], Microsoft Power BI [48], among others. Recommendation of action is less common; however, it is offered by tools such as DataSite [2] and ForeSight [26] that suggest users with subsequent steps of analysis. Recommendation of possible questions that can be asked from data is offered by Voder [24] (cf. Figure 3.8) and IBM Watson Analytics [68] that performs natural-language-processing for the task. Apart from proactive recommendations, tools such Zenvisage [62] automatically search for user specified patterns in data; while the tool SketchStory [27] (cf. Figure 3.9(b)) identifies specific partial sketches drawn on the user interface using a digital pen, and automatically completes the graphical representation. Moreover, tools such as Lyra [81] and iVisDesigner [70] facilitate users to explore data without any programming knowledge.

To reduce the required analytical expertise in each step of the EDA process, tools such as Voyager2 [30] and ForeSight [26] proactively provide visual summaries to distinguish attributes; whereas, the tool Analyza [59] guides users through the data discovery (i.e., distinguish attributes and univariate analysis) and the detection of relations between attributes (i.e., bivariate and multivariate analysis). Moreover, the proactive chart recommendations by some academic [30], [46], [59], [80] and commercial tools [47], [48], [52]



Currently supported visualizations for listed attributes (Cylinders,Displacement)

Average Displacement of 8 is 3.76 times 3

3 has lowest average Displacement

8 has highest average Displacement

Average Displacement of 8 is 3.76 times 3

4 has item (fiat 128) with lowest value for Displacement

8 has item (buick electra 225 custom) with highest value for Displacement

3 has lowest total Displacement

8 has highest total Displacement

Items with Origin:Japan exhibit a strong correlation between Displacement and Weight

Items with Origin:Japan exhibit a strong correlation between Horsepower and Weight

Items with Origin:Japan exhibit a strong inverse correlation between Acceleration and Horsepower

Items with Origin:Japan exhibit a strong correlation between Displacement and Horsepower

Figure 3.8: Explore view of the interface of the tool Voder (cf. Srinivasan et al. - Fig. 4 [24])

Note: (A) shows specification of visualization, (B) shows active visualization, (C) automatically generated data facts, (D) starred data facts about the current visualization, (E) System generated visuals for other data facts that can be explored, (F) Query panel for data facts, (G) possible visualizations for the chosen attributes.

also help with univariate and bivariate analysis. Nevertheless, we noticed a lack of proactive guidance for multivariate analysis among the EDA tools. For the identification of outliers, tools such as ForeSight [26], RCLens [77], Voyager2 [30], and SeeDB [46] proactively highlight apparently abnormal values in the dataset. With the help of a live keyword search, some tools [7], [24] allow the user to impute missing and aberrant data. For assistance with feature engineering, some tools [20], [26] proactively recommend feature combinations and derivations of new features.

iii. **User Engagement:** In recent years, visual EDA tools are used in different domains to make informed decisions from data. Hence, in order to enhance the users' trust on the visual representations provided by these EDA tools, researchers have proposed several mechanisms to engage end-users. For example, tools such as NorthStar [6], PanoramicData [25], SketchStory [27], and ExPlates [28] use interactive pen and touch features of the graphical user interface to engage users. Other tools such as LineUp [63], Voder [24], Duet [72], RCLens [77], ForeSight [26], InfoNice [82] (cf. Figure 3.9(a)) allow users to provide feedback on the visual representations, embellishments, and proactive recommendations. Additionally, tools such as ExPlates [28], Voyager [83], ForeCache [22], and HindSight [78] allow users to see a history of the performed analysis tasks, so that not only undo operations can be permitted but also the new EDA results can be compared with previously obtained results. Finally, tools like Voder [24] (cf. Figure 3.8) and PivotSlice [7] allow users to execute live search operations on the data that produce transformed or derived results.

In these EDA tools, user engagement with the EDA process usually starts from the very beginning of the analysis. For example, the drag and drop feature in NorthStar [6], PanoramicData [25], and SketchStory [27] engages users in distinguishing attributes and performing univariate, bivariate, and multivariate analysis. This feature lets users combine two or more attributes together simply by drawing a line between them. On the other hand, the interactive feedback allowed by Duet [72], RCLens [77], ForeSight [26] engages users in the detection and imputation of outliers and missing data. For engagement with feature engineering [20], showing historical interactions from users assists with more informed decision making.

iv. **Interpretability:** Due to the large volume of data being analyzed, visualizations showing inter-relations among attributes can be difficult to interpret. In order to assist with this challenge, recent visual EDA tools attempt to help users with interpretations of the generated visualizations. For example, tools such as Voder [24], DataSite [2], ExPlates [28], Ellipsis [29], and ChartAccent [84] present users with natural language annotations alongside the visualizations. These annotations discuss details such as the distribution, value range, and most common values of an attribute. However, for the tools that we analyzed, comprehensive annotations are only offered for univariate [72], bivariate [26], and multivariate [46]

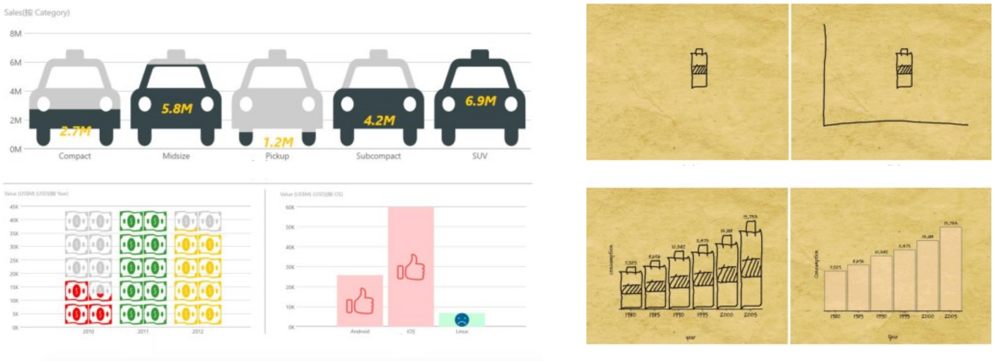


Figure 3.9: User Engagement initiatives: (a) On the left tool InfoNice (cf. Wang et al. Fig. 7 [69]) allows users to customize traditional visualizations. (b) On the right tool SketchStory (cf. Lee et al. Fig. 5 [27]) autocompletes the visuals based on sketches of the users.

analysis. Whereas, the other steps of the EDA process such as distinguishing attributes, and identification of missing values and outliers are rarely addressed by the interpretable EDA tools.

3.3 Research Opportunities

The results of our analysis show that based on changes in data analysis requirements [3], modern EDA tools have included support for some additional features (e.g., scalability, interpretability etc.). However, we have identified some potential research opportunities that can enhance the abilities of visual EDA tools. We believe, in order to make informed decisions from data, deeper statistical analysis is required to understand the complex relationships among its attributes. Our analysis shows, the trade-off between the breadth and depth of supported operations in the visual EDA tools still remains open. Whereas, most EDA tools designed for a generic target audience do not perform complex statistical analysis of data, tools that support such operations are either domain specific or are challenging to interpret. Hence, we identify and list a set of potential research opportunities in the domain of exploratory data analysis as follows.

i. **Detailed Analysis and Visualization of Bivariate & Multivariate Statistics:** In statistical analysis, the strength of a bivariate relationship between two attributes is usually obtained using correlation coefficients [56]. On the other hand, for accurate multivariate analytics, factor analysis (e.g., PCA [85]) techniques are used. However, the visualization of the results for these statistical tests can be complicated [33] for non-expert users. Currently, most of our analyzed visual EDA tools only perform brush-link and filter operations to show correlations among attributes. Although some tools [23] do perform dimensionality reduction of attributes, the reduced dimensions are not depicted in a comprehensive way [75]. Hence, there is a need for visual EDA tools to perform more complex statistical analysis (e.g., performing factor analysis for multivariate attributes instead of brush-link and filter) and to provide more comprehensive visualizations of the results. Additionally, during our analysis we also noticed that although some of the investigated EDA tools [24] allow users to visualize univariate and bivariate outliers, identification and visualization of multivariate outliers is still not performed by any tool. Moreover, the tools that detect outliers in data do not support any automated imputation of these values. It is important for researchers to consider automated strategies for outlier imputation in visual EDA tools.

ii. **Advanced Discretization of Continuous Variables:** Almost all the tools that were investigated during this work perform discretization [1] of continuous variables. Discretization is a process where continuous variables are split into bins or categories based on ranges in their values. Research [66] shows that the task of discretization can add error in data analysis as the selection of optimal bin-value ranges for continuous variables is often challenging. Our analysis shows, although most of the recent visual EDA tools discretize continuous variables into histograms, none of our analyzed tools consider any error or confidence [66] of the discretization process. Hence, more research is required that considers minimizing the discretization error in order to perform a more accurate analysis. Moreover, some values of a continuous variable might have higher importance than some other values of the same variable. There is a need for EDA tools to accommodate this fact. Although some tools [26], [60] support weighing attributes values based on their importance, there is a need for further research in this direction.

iii. **Proactive Guidance for Multivariate Relationships:** As discussed in Section 3.2.1, in high-dimensional industrial datasets there can be complex multivariate relationships among attributes that would require much analytical expertise to understand. Moreover, in case of datasets with large number of attributes, it can be immensely challenging to identify the features that are related together and influence each other. During our analysis, we noticed a significant gap in EDA tools with respect to proactive grouping and depiction of related attributes in the data. Although some tools such as Microsoft Power BI [48][48] visualizes the relationships among different data-sources using entity-relationship diagrams, none of the analyzed EDA tools perform any proactive grouping among the related attributes (apart from grouping them with respect to their data types [24]).

iv. **Scalability Vs. Data Visualization:** Scalability of visual EDA systems is a known challenge [3]. In order to deal with this challenge, many of our analyzed EDA tools [6], [9], [22], [44] have suggested several scalability measures that can visualize billions of records within an acceptable time limit. Nevertheless, the concern of scalable visual analysis is twofold: firstly, despite of several existing visualization approaches, the reduced dimensions of a dataset are difficult to interpret. Secondly, the number of data points to display is often much larger than the number of pixels available in one screen [3]. Researchers have proposed the use of data reduction techniques such as filtering, aggregation, sampling, and clustering in order to address the challenges. However, whilst data reduction techniques can solve visual scalability challenges, they can induce additional error in the analysis process. Moreover, outputs of data reduction tasks such as binned aggregation of data [21], or data split into data cubes

[23] are difficult to visually interpret. Hence, there is a need for researchers to investigate more comprehensive visual techniques for data reduction.

3.4 Limitations and Future Work

In this section, we enlist a set of limitations of this research, that provides opportunities for future work. First of all, in this chapter, we perform a comprehensive review of visual EDA tools based on a selection of 43 academic and 7 commercial tools used for general purpose data analysis. Although, we precisely define and justify our selection criteria (cf. Section 3.1), many existing visual EDA tools were excluded. In order to avoid any biases in the selection criteria, we performed data source triangulation [86], where the selected tools were chosen from both academia and industry. Moreover, the academic tools were selected from multiple reputed journals and conferences. Nevertheless, as the analysis of each and every existing EDA tool is beyond the capacity of any individual research article, we had to limit the scope of this research. Future work needs to focus on extending our study and include more tools in the analysis.

Moreover, apart from the utility for the analyzed tools in each step of the EDA process, we also evaluated the tools for the extent to which they meet the list of additional exploratory requirements (cf. Section 3.2.1) for analyzing large industrial datasets. To elicit these additional requirements, we mapped the identified challenging aspects (cf. Section 3.2.1) of our analyzed industrial dataset to the known big-data analysis requirements [19], [21], [23], [34], [37], [41], [42]. In order to add more exploratory requirements in the evaluation of EDA tools, future work could perform a cross-sectional study [86] across industry and academia to identify more requirements for large scale EDA. Finally, researcher bias [86] is a known challenge [35] in systematic literature reviews. To avoid any kind of researcher biases, in this study a group of two researchers independently performed all the data analysis tasks. In case of conflicts among these two researchers, a third researcher stepped in to alleviate the disagreements. Nevertheless, in future investigator triangulation [86] could be performed, where researchers from both industry and academia could collaboratively explore the utilities of different EDA tools, to generalize the decisions made during this research even further.

3.5 Related Work

Identification of the state-of-the-art in exploratory data visualization is a well-researched area [16], [35], [36], [55]. However, a common challenge with such research is that with every new advancement in the research community, the work gets outdated quickly. Visual analysis of data is a large umbrella that spreads over several different perspectives and applications of data analysis. Numerous surveys exist that focus on identification of visualization libraries [32], packages [34], and tools [33]. For example, whereas, surveys [87] on visual data mining tools are commonly published within research community, surveys [19] also exist that focus on presenting multivariate data visualization techniques. Moreover, many surveys have been performed on tools and techniques used to analyze big data [34], [37]. However, most of these surveys focus on specific aspects of big data analysis, such as indexing techniques for big data, or visualization of high-dimensional [33], [53] data. Among these, some of the surveys [35] focus on the advancements only in commercial data analysis tools. Also, other surveys [88] have looked into visualization recommender systems. However, in most cases, the state-of-the-art surveys for visualization tools focus on applications of the visualization. For example, surveys exist that present visualization of biological data [10], or visual sentiment analysis tools [86], or visualization of meteorological data [11]. In recent years, researchers have been focusing on combinations of visualization techniques and machine learning models [12] to enhance interpretability of the machine learning process [12]. Surveys presented by Liu et al. [14] and Endert et al. [12] focus on techniques that are used to integrate machine learning and visual analytics together. Several surveys [13] have been performed by researchers that classify visualization tools based on their utilities with respect to data analysis steps for various purposes [55]. However, unlike existing surveys on visualization tools, this work focuses on 50 visual EDA tools that are used for exploration of tabular data and were developed within the last 5 years. Our novel analysis classifies the existing tools for their abilities to assist with each steps of exploratory visualization of large industrial data.

3.6 Conclusions

In this research, we identify the primary focus areas of visually exploring industrial tabular datasets by analyzing a real-world dataset of 3.4 million records. Later, we present a systematic literature review of 50 state-of-the-art visual data analytics tools and their utility in six distinct steps of the Exploratory Data Analysis (EDA) process. We also investigate the extent to which these modern visual EDA tools address scalability, interpretability, and analytical expertise challenges of analyzing large datasets. Our analysis shows, most modern EDA tools assist with the fundamental steps of the EDA process, whilst only some tools consider addressing the challenges of big-data analytics. Among the analyzed tools however, the trade-off between breadth of supported features and in-depth analysis of data is still remaining. Even the most advanced tools in both academia and industry do not depict complex multivariate relationships among attributes. The reason behind this is, most tabular data analysis tools are primarily designed for a generic audience who might need more training to perform complex statistical analysis with the data. Moreover, some academic EDA tools that perform factor analysis or use complex diagrams to show relationships between multiple attributes, often suffer from interpretability and scalability issues. Incorporation of domain expertise is another challenge in most modern EDA tools. As in most cases for both commercial and academic tools, the user gets to take only the viewer's role in the data analysis process. Especially for the EDA tools that proactively generate visual recommendations; the absence of any feedback process can cause users to lose their confidence on the suggestions provided by the tools. Overall, we think there are many research opportunities in this emerging field that can be looked into for enhancing the performance and user experience of visual EDA tools.

References

[1] S. Tufféry, *Data Mining and Statistics for Decision Making: Tufféry/Data Mining and Statistics for Decision Making*. Chichester, UK: John Wiley & Sons, Ltd, 2011.

[2] Z. Cui, S. K. Badam, A. Yalçin, and N. Elmqvist, “DataSite: Proactive Visual Data Exploration with Computation of Insight-based Recommendations,” *arXiv:1802.08621 [cs]*, Sep. 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1802.08621.

[3] P. Godfrey, J. Gryz, and P. Lasek, “Interactive Visualization of Large Data Sets,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2142–2157, Aug. 2016, doi: 10.1109/TKDE.2016.2557324.

[4] M. A. Yalcin, N. Elmqvist, and B. B. Bederson, “Keshif: Rapid and Expressive Tabular Data Exploration for Novices,” *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 8, pp. 2339–2352, Aug. 2018.

[5] M. El-Hindi, Z. Zhao, C. Binnig, and T. Kraska, “VisTrees: fast indexes for interactive data exploration,” in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA ’16*, San Francisco, California, 2016, pp. 1–6, doi: 10.1145/2939502.2939507.

[6] T. Kraska, “Northstar: an interactive data science system,” *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 2150–2164, Aug. 2018, doi: 10.14778/3229863.3240493.

[7] Jian Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, “Interactive Exploration of Implicit and Explicit Relations in Faceted Datasets,” *IEEE Trans. Visual. Comput. Graphics*, vol. 19, no. 12, pp. 2080–2089, Dec. 2013.

[8] B. Yu and C. T. Silva, “VisFlow - Web-based Visualization Framework for Tabular Data with a Subset Flow Model,” *IEEE Trans. Visual. Comput. Graphics*, vol. 23, no. 1, pp. 251–260, Jan. 2017.

[9] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit, “Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 2023–

2032, Dec. 2014, doi: 10.1109/TVCG.2014.2346260.

[10] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M.R. Speicher, J. Zschocke, and Z. Trajanoski, “A survey of tools for variant analysis of next-generation genome sequencing data,” *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 256–278, Mar. 2014.

[11] M. Rautenhaus, M. Böttinger, S. Siemen, R. Hoffman, R.M. Kirby, M. Mirzargar, N. Röber, R. Westermann, “Visualization in Meteorology - A Survey of Techniques and Tools for Data Analysis Tasks,” *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 12, pp. 3268–3296, Dec. 2018.

[12] A. Endert, W. Ribarsky, C. Turkay, B.W. Wong, I. Nabney, I.D. Blanco, and F. Rossi, “The State of the Art in Integrating Machine Learning into Visual Analytics,” *Computer Graphics Forum*, vol. 36, no. 8, pp. 458–486, Dec. 2017, doi: 10.1111/cgf.13092.

[13] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, “Tools for Educational Data Mining: A Review,” *Journal of Educational and Behavioral Statistics*, vol. 42, no. 1, pp. 85–106, Feb. 2017.

[14] S. Liu, X. Wang, M. Liu, and J. Zhu, “Towards better analysis of machine learning models: A visual analytics perspective,” *Visual Informatics*, vol. 1, no. 1, pp. 48–56, Mar. 2017, doi: 10.1016/j.visinf.2017.01.006.

[15] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, “Overview of Data Exploration Techniques,” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD ’15*, Melbourne, Victoria, Australia, 2015, pp. 277–281.

[16] M. Khan and S. S. Khan, “Data and Information Visualization Methods, and Interactive Mechanisms: A Survey,” *International Journal of Computer Applications*, vol. 34, p. 15.

[17] J. Heer and B. Shneiderman, “Interactive Dynamics for Visual Analysis,” *Queue*, vol. 10, no. 2, pp. 30-55, 2012.

[18] R. Amar, J. Eagan, and J. Stasko, “Low-Level Components of Analytic Activity in Information Visualization,” *In IEEE Symposium on Information*

Visualization, INFOVIS 2005, pp. 111-117. 2005.

[19] W. Chan, “A Survey on Multivariate Data Visualization,” Department of Computer Science and Engineering. Hong Kong University of Science and Technology, vol. 8, no. 6, pp.1-29, 2006.

[20] R. High, “The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works.” RedBooks, 2012, Accessed: May 24, 2020. [Online].

[21] Z. Liu, B. Jiang, and J. Heer, “imMens: Real-time Visual Querying of Big Data,” *Computer Graphics Forum*, vol. 32, no. 3pt4, pp. 421–430, Jun. 2013.

[22] L. Battle, R. Chang, and M. Stonebraker, “Dynamic Prefetching of Data Tiles for Interactive Visualization,” in *Proceedings of the 2016 International Conference on Management of Data - SIGMOD ’16*, San Francisco, California, USA, 2016, pp. 1363–1375, doi: 10.1145/2882903.2882919.

[23] Z. Wang, N. Ferreira, Y. Wei, A. S. Bhaskar, and C. Scheidegger, “Gaussian Cubes: Real-Time Modeling for Visual Exploration of Large Multidimensional Datasets,” *IEEE Trans. Visual. Comput. Graphics*, vol. 23, no. 1, pp. 681–690, Jan. 2017, doi: 10.1109/TVCG.2016.2598694.

[24] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko, “Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication,” *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 672–681, Jan. 2019, doi: 10.1109/TVCG.2018.2865145.

[25] E. Zgraggen, R. Zeleznik, and S. M. Drucker, “PanoramicData: Data Analysis through Pen & Touch,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 2112–2121, Dec. 2014, doi: 10.1109/TVCG.2014.2346293.

[26] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati, “Foresight: Rapid Data Exploration Through Guideposts,” *arXiv:1709.10513 [cs]*, Sep. 2017, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1709.10513.

[27] Bongshin Lee, R. H. Kazi, and G. Smith, “SketchStory: Telling More Engaging Stories with Data through Freeform Sketching,” *IEEE Trans. Visual. Comput. Graphics*, vol. 19, no. 12, pp. 2416–2425, Dec. 2013.

[28] W. Javed and N. Elmqvist, "ExPlates: Spatializing Interactive Analysis to Scaffold Visual Exploration," *Computer Graphics Forum*, vol. 32, no. 3pt4, pp. 441–450, Jun. 2013, doi: 10.1111/cgf.12131.

[29] A. Satyanarayan and J. Heer, "Authoring Narrative Visualizations with Ellipsis: Authoring Narrative Visualizations with Ellipsis," *Computer Graphics Forum*, vol. 33, no. 3, pp. 361–370, Jun. 2014.

[30] K. Wongsuphasawat, , Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, J. Heer, "Voyager 2: Augmenting Visual Analysis with Partial View Specifications," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA, May 2017, pp. 2648–2659, doi: 10.1145/3025453.3025768.

[31] H. Mei, W. Chen, Y. Ma, H. Guan, and W. Hu, "VisComposer: A Visual Programmable Composition Environment for Information Visualization," *Visual Informatics*, vol. 2, no. 1, pp. 71–81, Mar. 2018, doi: 10.1016/j.visinf.2018.04.008.

[32] N. Bikakis and T. Sellis, "Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art," *arXiv:1601.08059 [cs]*, Jan. 2016, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1601.08059.

[33] W. Dunn, A. Burgun, M.-O. Krebs, and B. Rance, "Exploring and visualizing multidimensional data in translational research platforms," *Brief Bioinform*, p. bbw080, Sep. 2016, doi: 10.1093/bib/bbw080.

[34] L. Wang, G. Wang, and C. A. Alexander, "Big Data and Visualization: Methods, Challenges and Technology Progress," *Digital Technologies*, vol. 1, no. 1, pp.33-38, 2015.

[35] M. Behrisch, Streeb, F. Stoffel, D. Seebacher, B. Matejek, S.H. Weber, S. Mittelstaedt, H. Pfister, and D. Keim, "Commercial Visual Analytics Systems–Advances in the Big Data Analytics Field," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 10, pp. 3011–3031, Oct. 2019, doi: 10.1109/TVCG.2018.2859973.

[36] M. Diamond and A. Mattia, “Data visualization: an exploratory study into the software tools used by businesses,” *Journal of Instructional Pedagogies*, vol. 18, p. 7, 2018.

[37] S. M. Biju and A. Mathew, “Comparative analysis of selected big data analytics tools,” vol. 26, no. 2, p. 23, 2017.

[38] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” *Journal of Big Data*, vol. 2, no. 1, p. 1, Dec. 2015, doi: 10.1186/s40537-014-0007-7.

[39] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, “Big Data: Issues and Challenges Moving Forward,” in *2013 46th Hawaii International Conference on System Sciences*, Wailea, HI, USA, Jan. 2013, pp. 995–1004, doi: 10.1109/HICSS.2013.645.

[40] I. M. Johnstone and D. M. Titterington, “Statistical challenges of high-dimensional data,” *Proc. R. Soc. A*, vol. 367, no. 1906, pp. 4237–4253, Nov. 2009, doi: 10.1098/rsta.2009.0159.

[41] J.-F. Im, F. G. Villegas, and M. J. McGuffin, “VisReduce: Fast and responsive incremental information visualization of large datasets,” in *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, Oct. 2013, pp. 25–32, doi: 10.1109/BigData.2013.6691710.

[42] G. R. Iyer, S. Duttaduwarah, and A. Sharma, “DataScope: Interactive visual exploratory dashboards for large multidimensional data,” PeerJ Preprints, preprint, Jan. 2018. doi: 10.7287/peerj.preprints.26441v1.

[43] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[44] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit, “Taggle: Combining Overview and Details in Tabular Data Visualizations,” *Information Visualization*, vol. 19, no. 2, pp. 114–136, Apr.

2020, doi: 10.1177/1473871619878085.

[45] C. Niederer, H. Stitz, R. Hourieh, F. Grassinger, W. Aigner, and M. Streit, "TACO: Visualizing Changes in Tables Over Time," *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 1, pp. 677–686, Jan. 2018.

[46] M. Vartak, S. Rahman, S. Madden, and A. Parameswaran, "SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics," *In Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, vol. 8, no. 13, p. 2182. NIH Public Access, 2015.

[47] Tableau, "Tableau." https://www.tableau.com/products. (accessed Oct. 01, 2018).

[48] Microsoft Corp., "Microsoft Power BI." https://powerbi.microsoft.com/ (accessed Oct. 01, 2018).

[49] Domo, "Domo." https://www.domo.com/ (accessed Oct. 01, 2018).

[50] QlikView, "QlikView." https://www.qlik.com/us/ (accessed Oct. 01, 2018).

[51] Sisense, "Sisense." https://www.sisense.com/product/ (accessed Oct. 01, 2018).

[52] R. L. Sallam, C. Howson, C. J. Idoine, T. W. Oestreich, J. Laurence, and J. Tapadinhas, "Magic Quadrant for Business Intelligence and Analytics Platforms," p. 87.

[53] S. Liu, D. Maljovec, B. Wang, P. -t Bremer, and V. Pascucci, "Visualizing High-Dimensional Data: Advances in the Past Decade," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 21–30, 2016.

[54] J. Fan and R. Li, "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," *arXiv:math/0602133*, Feb. 2006, Accessed: May 25, 2020. [Online]. Available: http://arxiv.org/abs/math/0602133.

[55] J. C. Roberts, "State of the Art: Coordinated & Multiple Views in Exploratory Visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, Zurich, Switzerland, Jul. 2007, pp. 61–71, doi: 10.1109/CMV.2007.20.

[56] M. Smith, *Statistical Analysis Handbook*. Edinburgh,: The Winchelsea Press, Drumlin Security Ltd, 2018.

[57] M. A. Yalçin, N. Elmqvist, and B. B. Bederson, “AggreSet: Rich and Scalable Set Exploration using Visualizations of Element Aggregations,” *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 688–697, Jan. 2016, doi: 10.1109/TVCG.2015.2467051.

[58] Jing Xia, Wei Chen, Yumeng Hou, Wanqi Hu, Xinxin Huang, and D. S. Ebertk, “DimScanner: A relation-based visual exploration approach towards data dimension inspection,” in *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Baltimore, MD, USA, Oct. 2016, pp. 81–90, doi: 10.1109/VAST.2016.7883514.

[59] K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan, “Analyza: Exploring Data with Conversation,” in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, Limassol Cyprus, Mar. 2017, pp. 493–504, doi: 10.1145/3025171.3025227.

[60] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, “Podium: Ranking Data Using Mixed-Initiative Visual Analytics,” *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 1, pp. 288–297, Jan. 2018, doi: 10.1109/TVCG.2017.2745078.

[61] C. Perin, P. Dragicevic, and J.-D. Fekete, “Revisiting Bertin Matrices: New Interactions for Crafting Tabular Visualizations,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 2082–2091, Dec. 2014.

[62] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran, “Effortless Data Exploration with zenvisage: An Expressive and Interactive Visual Analytics System,” *arXiv:1604.03583 [cs]*, Jan. 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1604.03583.

[63] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual Analysis of Multi-Attribute Rankings,” *IEEE Trans. Visual. Comput. Graphics*, vol. 19, no. 12, pp. 2277–2286, Dec. 2013.

[64] A. Schulz, “Discriminative Dimensionality Reduction: Variations,

Applications, Interpretations," *Doctoral Thesis - Publikationen an der Universität Bielefeld*, 2017.

[65] R. Hourieh, H. Stitz, N. Gehlenborg, and M. Streit, "TaCo: Comparative Visualization of Large Tabular Data," *Detail*, vol. 1000, no. 1, p.1, 2016.

[66] N. Kamat and A. Nandi, "InfiniViz: Interactive Visual Exploration using Progressive Bin Refinement," *arXiv:1710.01854 [cs]*, Oct. 2017, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1710.01854.

[67] J. Kelly, "Computing, cognition and the future of knowing How humans and machines are forging a new age of understanding," IBM Corporation. Accessed: May 25, 2020. [Online]. Available: https://cloud.report/Resources/Whitepapers/e55108d4-92bd-428a-b432-64530b50c6b9_Computing_Cognition_WhitePaper.pdf.

[68] F. Anderson, "Getting Started Tutorial for IBM Watson Analytics," IBM Corporation, New York, NY, USA, 2012.

[69] P. Koytek, C. Perin, J. Vermeulen, E. Andre, and S. Carpendale, "MyBrush: Brushing and Linking with Personal Agency," *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 1, pp. 605–615, Jan. 2018, doi: 10.1109/TVCG.2017.2743859.

[70] D. Ren, T. Hollerer, and X. Yuan, "iVisDesigner: Expressive Interactive Design of Information Visualizations," *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 2092–2101, Dec. 2014, doi: 10.1109/TVCG.2014.2346291.

[71] M. Budiu, R. Isaacs, D. Murray, G. Plotkin, P. Barham, S. Al-Kiswany, Y. Boshmaf, Q. Luo, A. Andoni, "Interacting with large distributed datasets using Sketch," UW-Madison Department of Computer Sciences, CS Technical Reports, 2015.

[72] P.-M. Law, R. C. Basole, and Y. Wu, "Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification," IEEE transactions on visualization and computer graphics, vol. 25, no. 1, pp. 427-437, 2018.

[73] A. Mokalis and J. Davis, *Google Analytics Demystified*. CreateSpace

Independent Publishing Platform, 2018.

[74] S. Macke, Y. Zhang, S. Huang, and A. Parameswaran, “Adaptive Sampling for Rapidly Matching Histograms,” *arXiv:1708.05918 [cs]*, May 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1708.05918.

[75] E. Isaacs, K. Damico, S. Ahern, E. Bart, and M. Singhal, “Footprints: A Visual Search Tool that Supports Discovery and Coverage Tracking,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 1793–1802, Dec. 2014, doi: 10.1109/TVCG.2014.2346743.

[76] A. F. Zuur, E. N. Ieno, and C. S. Elphick, “A protocol for data exploration to avoid common statistical problems: Data exploration,” *Methods in Ecology and Evolution*, vol. 1, no. 1, pp. 3–14, Mar. 2010, doi: 10.1111/j.2041-210X.2009.00001.x.

[77] H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao, “RCLens: Interactive Rare Category Exploration and Identification,” *IEEE Trans. Visual. Comput. Graphics*, vol. 24, no. 7, pp. 2223–2237, Jul. 2018, doi: 10.1109/TVCG.2017.2711030.

[78] M. Feng, C. Deng, E. M. Peck, and L. Harrison, “HindSight: Encouraging Exploration through Direct Encoding of Personal Interaction History,” *IEEE Trans. Visual. Comput. Graphics*, vol. 23, no. 1, pp. 351–360, Jan. 2017, doi: 10.1109/TVCG.2016.2599058.

[79] C. D. Stolper, A. Perer, and D. Gotz, “Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics,” *IEEE Trans. Visual. Comput. Graphics*, vol. 20, no. 12, pp. 1653–1662, Dec. 2014, doi: 10.1109/TVCG.2014.2346574.

[80] D. Perry, B. Howe, A. M. F. Key, and C. Aragon, “VizDeck: Streamlining Exploratory Visual Analytics of Scientific Data,”, *In iSchools*, 2013, DOI: https://doi.org/10.9776/13206.

[81] A. Satyanarayan and J. Heer, “Lyra: An Interactive Visualization Design Environment: Lyra: An Interactive Visualization Design Environment,”

Computer Graphics Forum, vol. 33, no. 3, pp. 351–360, Jun. 2014, doi: 10.1111/cgf.12391.

[82] Y. Wang *et al.*, "InfoNice: Easy Creation of Information Graphics," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Montreal QC, Canada, 2018, pp. 1–12, doi: 10.1145/3173574.3173909.

[83] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations," *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 649–658, Jan. 2016, doi: 10.1109/TVCG.2015.2467191.

[84] D. Ren, M. Brehmer, Bongshin Lee, T. Hollerer, and E. K. Choe, "ChartAccent: Annotation for data-driven storytelling," in *2017 IEEE Pacific Visualization Symposium (PacificVis)*, Seoul, South Korea, Apr. 2017, pp. 230–239, doi: 10.1109/PACIFICVIS.2017.8031599.

[85] R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, 2014, doi: 10.1039/C3AY41907J.

[86] F. Shull, J. Singer, and D. I. K. Sjøberg, Eds., *Guide to advanced empirical software engineering*. London: Springer, 2008.

[87] D. A. Keim, "Information visualization and visual data mining," *IEEE Trans. Visual. Comput. Graphics*, vol. 8, no. 1, pp. 1–8, Mar. 2002, doi: 10.1109/2945.981847.

[88] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran, "Towards Visualization Recommendation Systems," *SIGMOD Rec.*, vol. 45, no. 4, pp. 34–39, May 2017, doi: 10.1145/3092931.3092937.

Chapter 4

Context-Based Evaluation of Dimensionality Reduction Algorithms

Real-world industrial datasets are often high dimensional, where some of the attributes are redundant or correlated. Hence, due to the curse of dimensionality [1], it is challenging to analyze the underlying patterns of the data and to obtain insights from it. In order to address this challenge, researchers have proposed Dimensionality Reduction (DR in short) [1]–[3], a procedure that aims at transforming any high-dimensional dataset into its low-dimensional representation while retaining as much of the original structural relationships in the data as possible. Ideally, the low-dimensional representation should reflect the intrinsic dimensionality [4] of the original dataset, which is the minimum number of attributes required to present the original data. Since DR can decrease the computation time and required resources while facilitating better visual analytics of the data, it is a widely used technique in the domains of biotechnology [5], biology [5], medicine[5], pattern-recognition [6], among many others.

DR has been an open research area for over a century [7]. As a result, throughout the past years, numerous algorithms [1], [8], [9] have been proposed by researchers. Initially, DR methods such as Principal Component Analysis (PCA) [7] and classical metric Multidimensional Scaling (MDS) [1] primarily focused on generating simple linear projections of the input datasets. However, linear techniques cannot preserve nonlinear structural relationships within any high-dimensional dataset. As a result, in the past few years, several new nonlinear DR (NLDR) methods have been proposed. Such methods include: Isomap [1], non-metric Multidimensional Scaling (nMDS) [10], Locally Linear Embedding (LLE)

[11], t-distributed Stochastic Neighbor Embedding (t-SNE) [9], Uniform Manifold Approximation and Projection (UMAP) [8], Trimap [12] etc. Hence, given the plethora of existing techniques, users constantly face the challenge of selecting the most appropriate algorithm for their specific analytical context. In practice, different DR algorithms can produce qualitatively dissimilar results for the same input dataset. The reason being: not all high-dimensional relationships can be preserved [1] in the low-dimensional representations and the relationships in the data that must be preserved remains unclear. As a result, different DR algorithms use different objective functions to preserve specific aspects of the existing structural relationships in data. Moreover, multiple combinations of hyperparameter values can lead to very different outcomes [13] for the same algorithm.

Besides, each DR method has its strengths and weaknesses [1]. That is, for any given DR algorithm there exists a perfectly reasonable metric [8], [9], [12] for which it performs better than others. For example, if experiments were carried out to identify the algorithm that captures the maximum covariance in the original data, PCA [1] would be the winner. Or, in case we evaluated the extent to which the overall distances among data-points were preserved, non-metric Multidimensional Scaling (nMDS) [10] could be the chosen algorithm. As a result, a large number of potential quantitative metrics; for example, that measure the accuracy of DR[14] or preservation of proximity relations [10], [11], have been proposed in the past years that could be applied to determine the most suitable DR method in a given context. However, due to the lack of systematic comparisons among the existing DR techniques, no generic guidelines exist [12] that can help users understand the trade-off between the performance of the same DR method in different analytical contexts. In order to bridge this gap, in this chapter, at first, we identify five most popular analytical contexts for DR, subsequently, we categorize 12 most popular DR quality metrics into these five analytical contexts. These metrics are then used to perform a systematic comparison between 15 state-of-the-art DR techniques for the identified contexts. Our primary objective is to produce a generic guideline for the long-existing open [1], [12] research question- **“Which DR algorithm should be used in a given scenario?”** Investigations were performed on 40 real-world datasets, among which 39 were compiled from open-source data repositories and one was obtained from our industrial partner IBM.

Furthermore, in this research, we perform statistical significance analysis [15] to validate our obtained experimental results. The primary reason being, traditionally DR algorithms [16] are compared against a set of other closely related algorithms over the same set of test datasets to prove the superiority of one algorithm over its competitors. However, such comparisons do not guarantee statistical significance [15]–[17] of the performance of the algorithms. That is, the question- **“How reliable and replicable is the performance of any DR method for a given metric?”** remains unanswered. Hence, to generalize our experimental results, following the guidelines of Demšar et al. [17], for each of the 12 quality metrics, we perform null hypothesis significance testing to implement both pairwise and overall comparisons among the set of algorithms. For our experiments, we used robust non-parametric statistical tests [15] such as Wilcoxon’s sign rank, McNemar, and Friedman tests. The contributions of this work are as follows:

- For the first time in the field of DR research, this chapter identifies, describes, and analyzes five analytical contexts in which DR algorithms are commonly applied.
- The chapter, to the first of our knowledge, composes 12 most popular DR quality metrics and categorizes them into the five identified analytical contexts. The metrics are then used to perform a systematic comparison among 15 state-of-the-art DR algorithms. The results identify the best, mediocre, and worst-performing algorithms in a given analytical context.
- Furthermore, this novel research performs a thorough statistical significance analysis of the performance of DR algorithms. This analysis statistically validates the replicability and reliability of our obtained results using 40 real-world datasets.

- Finally, this chapter presents the first generic guideline for practitioners to select the most appropriate DR algorithms in a given scenario. On the one hand, this guideline categorizes the DR algorithms and quality metrics for their utility in a given analytical context. On the other hand, it reduces the required analytical time and expertise from the data analyst.

The overall chapter is organized as follows: Section 4.1 presents the preliminary background information while introducing the identified analytical contexts and our chosen contextual metrics. It also provides an overview of the statistical significance tests used in our experiments. Section 4.2 discusses the experimental procedure along with the selected algorithms and datasets chosen for our experiments. Section 4.3 presents the detailed experimental results along with the outcome of the statistical significance testing. Section 4.4 presents a practitioner's guideline for users; Section 4.5 discusses the threats to the validity of our work. Section 4.6 presents a discussion on related work; and finally, Section 4.7 concludes the chapter.

4.1 Analysis and Problem Characterization

The process of DR can be formally defined as follows: assuming we have a dataset represented by a matrix X of size $n \times D$, where $X = [x_1, x_2, \dots., x_n] \in \mathbb{R}^{D\times n}$. Where each row in X represents a data vector x_i, such that the size of the vector x_i is D. DR transforms X to a low-dimensional embedding Y of size $n \times d$, where $Y = [y_1, y_2, \dots., y_n] \in \mathbb{R}^{d\times n}$. Each row y_i in Y represents a low-dimensional mapping for x_i so that size of the vector y_i is d. Where d is the intrinsic dimensionality of the dataset and ideally $d \ll D$. The identification of d for the input dataset makes a key impact on the amount of information loss [12] in Y. Most modern NLDR techniques are based on the concept of manifold learning [1]. Manifold learning assumes that the vectors in X are sampled from a smooth manifold [2]. Hence, the goal of any NLDR method is to embed each data vector from X in a space with dimensionality d while keeping the topological properties of the original manifold intact. Nevertheless, the identification of both intrinsic dimensionality and topological properties of real-life datasets are extremely challenging [1], [3]. Therefore, using an objective [4] function, most NLDR algorithms attempt to preserve the original structure of X that is defined by proximity relationships [1], [5] among the data-points [6] of X.

Quality analysis of DR has been a focus of the scientific community [1], [4], [5], [7], [8] for many years. DR can be considered as an optimization problem with the two simplest ways to evaluate [1], [5] its outcome being: (1) an assessment of the value of the objective function of an NLDR algorithm upon convergence; and (2) an inverse transformation on the embedding Y to investigate how accurately X can be retrieved. However, although the first approach can be useful for comparing multiple runs of the same algorithm (e.g., with different input datasets or with different hyperparameter combinations), it would cause an unfair comparison among multiple algorithms. On the other hand, the retrieval of X from Y is only applicable when the structure of the original high-dimensional manifold is known. However, for most real-world datasets this is not the case [1], [3], [5]. As a result, the quality of DR is often assessed using measures that look into specific analytical contexts. For example, by analyzing hidden patterns in X using Y or by evaluating the quality of Y with limited number of records in X. To perform a systematic evaluation of DR techniques, in this chapter, we have compiled five most popular application contexts for DR along with 12 DR quality analysis metrics (summarized in Table 4.1) that can be used in those contexts. In the next three sub-sections, at first, we present our identified analytical contexts for DR followed by a discussion on their associations with the 12 DR evaluation metrics. Finally, we formally introduce different statistical significance tests that were used to validate the results of our experiments.

4.1.1 Identification of Analytical Contexts for DR

In this research, the term ‘analytical context’ refers to the *purpose* of applying DR on any high-dimensional dataset. In real-world scenarios, such purposes [9] include identification of patterns and regularities in data or efficiently training predictive models [10] with the data. In this section, we identify and discuss *five* such analytical contexts in which DR techniques are commonly used. The identification of these analytical contexts is performed via a detailed investigation of relevant literature [3], [7], [11]–[17] that discusses the circumstances for examining, applying, and proposing DR algorithms. Our analysis revealed that, since DR helps with visualizing high dimensional data using traditional spatial techniques (i.e., 2D or 3D representations) [13], one of the most common [7], [12], [15] uses for DR is to identify patterns and similarities among the data-points in the input dataset. Hence, in this work, we identify *pattern analysis and similarity search* as our first analytical context for DR. At the same time, DR techniques are also commonly used [3], [16], [17] for pre-processing datasets before training machine learning models [17] with them. Since DR helps with removing redundancies in data, such pre-processing is performed to manage the execution time of these predictive models and to improve their performances. As a result, due to its popularity with DR, we select *predictive modelling* as our second analytical context for this research. Our investigation of relevant literature [7], [12], [16], [17] also revealed poor quality and limited input data to be some well-known challenges [9] of exploratory analysis of high-dimensional datasets. In their work, Becht et al. [16] and Amid et al. [17] have shown that such inconsistencies in data can have a significant impact on the performances of DR algorithms. Since, real-world datasets commonly [9] have such inconsistencies in them, in this work we distinguished *poor quality* and *limited input data* as our third and fourth analytical contexts respectively. Finally, limitations on computational resources are amongst well-known challenges [2], [16]–[18] of real-life data analysis. In order to mitigate such limitations, over the years researchers [2], [18] have presented DR algorithms that are faster and more efficient than their predecessors. Due to its importance in real-life data analysis, in this research we designate *limited computational resources* as our fifth analytical context for DR. In the following, we discuss the objectives and characteristics of our five identified analytical contexts in detail.

i. **DR for Pattern Analysis and Similarity Search:** One of the primary objectives of analyzing high-dimensional datasets is to discover previously unknown patterns and regularities in data [19]. Such patterns can help with the summarization and classification of data-points [10] and hence with decision making. Hence, in this analytical context, the primary goal of any DR technique becomes the reliable retention of the original proximity relationships in $\boldsymbol{Y}$ [7]. That is, how accurately can the DR technique project similar data-points in $\boldsymbol{X}$, to clusters in $\boldsymbol{Y}$; and dissimilar data-points in $\boldsymbol{X}$ to "remote locations" in $\boldsymbol{Y}$. The same objectives for DR techniques also apply when similarity search [20] and range queries are performed using $\boldsymbol{Y}$. In such cases, the analysis results depend not only on the preserved original proximities in data but also on the distance ranking [1] among the data-points. As the goals of this analytical context do not depend on the existence of labels in the data, it consists of common analytical tasks such as classification [14], clustering [21], summarization, and nearest neighbor [3] search queries.

ii. **DR for Predictive Modelling:** DR techniques are commonly used [3], [16], [17] during predictive modeling [17] where, they not only assist with reducing the analytical complexity but also with removing the redundancy in $\boldsymbol{X}$. In such an analytical context, we assess whether a model trained with $\boldsymbol{Y}$ can make equally accurate predictions as the same model trained with $\boldsymbol{X}$. Hence, the analysis depends on the quality of distance and neighborhood preservations [7] by the DR algorithms. This context is only applicable to supervised learning scenarios where labeled training data is available.

iii. **DR with Poor Quality Input Data:** Missing values and outliers are known challenges in real-world datasets [9] that violate data quality [22] characteristics such as completeness, accuracy, and consistency. During DR, such inconsistencies in the values of a dataset can impact on proximity scores [23], [24] of any data-point with respect to other points. Hence, this analytical context needs the DR techniques to handle such inconsistencies in the data better than their competitors.

iv. **DR with Limited Input Data:** DR is often applied to inadequate volumes of data samples that may result in misleading interpretations of the dataset. The reason being, with limited input data relatively dissimilar data-points, may seem similar [7] to each other in the embedding. Moreover, in some cases, DR is executed on datasets with unknown characteristics. For example, datasets with ambiguous attribute names or irrelevant attribute values. In this research, we combine the two situations under the analytical context of limited input data problem for DR. In this analytical context tuning the hyperparameters of DR algorithms can be challenging due to the incompleteness of the information. Hence, here the primary goal becomes the identification of the DR technique that performs equally well with and without having all necessary information regarding the input data.

v. **DR with Limited Computational Resources:** Optimizing an objective function as a part of DR is often a computationally expensive [3] process. Although, some algorithms are inherently faster than others [5], [7], [31], the overall execution time for DR algorithms depends on the amount of available computational resources [5]. In this analytical context, the primary need for a DR technique is to produce better quality results than other algorithms given the same resource constraints.

It is important to note that: the identified popular contexts are not an exhaustive list of analytical contexts where DR techniques can be applied. In order to define a finite scope for this research, we limit our analysis to the above-mentioned contexts. Further investigation of literature may reveal more contexts for using DR.

4.1.2 Classification of DR Quality Metrics into Analytical Contexts

In this section, we formally introduce the 12 most popular DR evaluation metrics and categorize them into the five identified analytical contexts. This categorization

Table 4.1: DR Quality Metrics with their Associated Analytical Contexts

Metric No.	Chosen Quality Metrics	Derived Analytical Contexts
1.	Residual variance ($\hat{\sigma}^2$)	DR for Pattern Analysis and Similarity Search
2.	Spearman rank correlation (ρ_s)	
3.	Mean K-ary neighborhood agreement ($\mu_{R_{nX}}$)	
4.	Local quality criteria (Q_{local})	
5.	k_{max} neighborhood loss ($\lambda_{K_{max}}$)	
6.	Global quality criteria (Q_{global})	
7.	Area under the R_{nX} curve ($AUC_{\ln K}(R_{nX}(K))$	DR for Predictive Modelling
8.	KNN prediction accuracy (ACC_{ψ})	
9.	Normalized Mutual Information (nMI)	DR with Poor Quality Input Data
10.	Structural Similarity Index (SSI)	
11.	Logarithmic loss of multi-class classification	DR with Limited Input Data
12.	Mean accuracy with constraints $\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}} \mid 0.1 * default\ iter(g_i)$	DR with Limited Computational Resources

of the quality metrics is performed by detailed analysis of background literature [1], [4], [5], [8], [12], [25] for each metric. In Table 4.1 we summarize the metrics and their categorizations. Assuming, we compare G DR algorithms on N datasets for M metrics, in the formal definitions of the metrics we adopt the notations presented above.

4.1.2.1 DR for Pattern Analysis and Similarity Search

In par with our discussion in Section 4.1.1, here we identify 6 most popular structural preservation metrics [4], [5], [8] for DR algorithms for the analytical context of pattern analysis and similarity search. As the metrics presented in this section, are agnostic to the existence of training labels in the datasets, they can be applied to both supervised and unsupervised analysis tasks (e.g., clustering and classification). The existing quality metrics for the structural preservation of DR can be broadly categorized as **distance-based** metrics [25] and **rank-based** metrics [1]. Whilst the former analyses the preserved structure by comparing pairwise distances among data points, the latter performs the same by comparing the ranks of the relative distances between points. It is proven [1], [5] that rank-based metrics are more stable to scaling of pairwise distances among data points in X caused due to the unfolding of the manifold. In our experiments, we use a combination of both distance and rank-based metrics to compare the structural preservation of different

DR algorithms.

Most distance-based quality metrics use the Euclidean distance [1], [3], [5], [26] as their underlying proximity measure. Formally, assuming x_i and x_j are any two different data points in X such that $x_i = [u_{1,}u_2, \dots, u_D]$ and $x_j = [v_{1,}v_2, \dots, v_D]$, the Euclidean distance between x_i and x_j can be defined as:

$$dist_\varepsilon\left(x_i, x_j\right) = \sqrt{\sum_{a=1}^{D}(u_a - v_a)^2} \qquad (4.1)$$

Formally, considering $dist_{\varepsilon_{ij}} = dist_\varepsilon(x_i, x_j)$ be the Euclidean distance between any two points x_i and x_j in X and $\widehat{dist}_{\varepsilon_{ij}} = d_\varepsilon(y_i, y_j)$ be the same for two points y_i and y_j in Y. Based on Eq. 4.1 we define Residual Variance [25] as our **first metric**, that is the first quality measure for analyzing preserved local structure as:

$$\hat{\sigma}^2 = 1 - \hat{R}^2(\{dist_{\varepsilon_{i0}}, \dots., dist_{\varepsilon_{in}}\}, \{\widehat{dist}_{\varepsilon_{i0}}, \dots., \widehat{dist}_{\varepsilon_{in}}\}) \qquad (4.2)$$

where, $\hat{R}$ represents a linear correlation coefficient [16]. As a local quality measure, $\hat{\sigma}^2$ measures the complement of the explained variance between all $dist_{\varepsilon_{ij}}$ and $\widehat{dist}_{\varepsilon_{ij}}$ using $\hat{R}$.

Among the rank-based quality measures for DR, the Spearman's rank correlation [8] has been one of the traditional techniques [8], [25]. The rank of any $x_i \in X$ with respect to any $x_j \in X$ can be defined as:

$$r_{ij} = \left|\left\{l : dist_{\varepsilon_{il}} < dist_{\varepsilon_{ij}} \; or \; \left(dist_{\varepsilon_{il}} = dist_{\varepsilon_{ij}} \; and \; 1 \le \mathrm{i} \le l \le j \le n\right)\right\}\right| \qquad (4.3)$$

Analogously, the rank for any $y_i \in Y$ with respect to any $y_j \in Y$ can be defined as:

$$\hat{r}_{ij} = \left|\left\{l : \widehat{dist}_{\varepsilon_{il}} < \widehat{dist}_{\varepsilon_{ij}} \; or \; \left(\widehat{dist}_{\varepsilon_{il}} = \widehat{dist}_{\varepsilon_{ij}} \; and \; 1 \le \mathrm{i} \le l \le j \le n\right)\right\}\right| \qquad (4.4)$$

Where, the notation $|A|$ in the equations 4.3 and 4.4 denote the size of set A. Using the above definitions for r_{ij} and $\hat{r}_{ij}$, our **second metric** the Spearman rank correlation coefficient ρ_s can be defined as:

$$\rho_s = 1 - 6\sum_{i=1}^{n}\sum_{j=1}^{n}\left(r_{ij} - \hat{r}_{ij}\right)/n(n^2 - 1) \qquad (4.5)$$

where n denotes the number of samples in X as $r_{ij,}$ and $\hat{r}_{ij}$ represent distinct integer ranks in high and low-dimensional spaces.

Over the past years, several [1], [5] other rank-based quality metrics have been proposed for DR. These include, Local Continuity Meta-Criterion (LCMC) [1], Trustworthiness and Continuity (T&C) [1], Mean Relative Rank Errors (MRRE) [1]. All these metrics analyze the ranks of sorted distances in K-ary neighborhoods [5] before and after DR. Later, these metrics were unified by Lee and Verleysen [1] under the *co-ranking matrix* framework. This framework analyzes the average agreement between K-ary neighborhoods in high and low-dimensional spaces using a matrix consisting of the ranks of the distances among data-points. Formally, defining the K-ary neighborhoods of the points $x_i \in X$ and $y_i \in Y$ as $ng_i^K = \{j: 1 \leq r_{ij} \leq K\}$ and $\nu\gamma_i^K = \{j: 1 \leq \hat{r}_{ij} \leq K\}$ respectively, using the co-ranking framework mean K-ary neighborhood preservation after DR can be defined [1] as:

$$Q_{nX}(K) = \frac{1}{Kn} \sum_{i=1}^{n} \left| ng_i^K \cap \nu\gamma_i^K \right| \quad \text{where } 1 \leq K \leq n-1 \tag{4.6}$$

The value of $Q_{nX}(K)$ varies from 0 to 1 implying an empty intersection and a perfect agreement between the same neighborhoods in X and Y respectively. To fairly compare and combine the values of $Q_{nX}(K)$ for different neighborhood sizes, Lee et al. [21] later defined a scaled version of $Q_{nX}(K)$ as:

$$R_{nX}(K) = \frac{(n-1)Q_{nX}(K) - K}{n-1-K} \tag{4.7}$$

Where, n is the number of points in X and $\forall\ 1 \leq K \leq n-2$, $R_{nX} = 0$ represents a random embedding as $R_{nX} = 1$ represents a perfect embedding. As $R_{nX}(K)$ can be better evaluated using visual analytics with a curve for different R_{nX} values for varying size of K, in our experiments we use $\mu_{R_{nX}}$ [5] as our **third metric** that is a quantitative scalar rank based quality metric for DR. Any DR algorithm that aims at preserving the local structure should return high values for $R_{nX}(K)$ for small values of K. On the other hand, a DR algorithm attempting to preserve both the local and the global structure of X in Y should attempt to keep all the values of $R_{nX}(K)$ as high as possible. In general, smaller values of K represent the locality of X as larger values of K represent the global structure of the manifold. Hence, as an estimate for K that represents locality in X, K_{max} [4] is defined as:

$$K_{max} = arg \max_{k}(Q_{nX}(K) - \frac{K}{n-1}) \tag{4.8}$$

Considering K_{max} as the splitting point, a local quality measure for an embedding can be defined [4] as our **fourth metric** as:

$$Q_{local} = \frac{1}{K_{max}} \sum_{K=1}^{K_{max}} Q_{nX}(K) \tag{4.9}$$

As our final quality metric for preserved local proximity, in our experiments, we used another popular metric *Neighborhood Loss* [25] as our **fifth metric**. We define the metric as:

$$\lambda_{K_{max}} = \sum_{i=1}^{n} 1 - \left|ng_i^K \cap \nu\gamma_i^K\right| / K \qquad \text{where } 1 \leq K \leq K_{max} \tag{4.10}$$

where, ng_i^K and $\nu\gamma_i^K$ represents an enumeration of the K nearest neighbors for each point in the original and embedded spaces respectively. In our experiments we use K_{max} as the value for K.

On the other hand, for analysis of preserved global structure in an embedding again considering K_{max} as the splitting point in the $Q_{nX}(K)$ curve, Lee and Verleysen [4] have defined the metric Q_{global} as:

$$Q_{global} = \frac{1}{n-K_{max}} \sum_{K=K_{max}}^{\mathrm{n}-1} Q_{nX}(K) \tag{4.11}$$

We use Q_{global} as the **sixth metric** in our experiments. Both the quality measures presented in Eq. 4.9 and 4.11 range from 0 to 1. Nevertheless, in the literature [4] Q_{local} is given more importance than Q_{global}. As a common consensus among researchers [4] specifies that for pattern analysis the preservation of small K-ary neighborhoods is more important [4], [5] than the preservation of the overall global structure of the data.

Among the five rank-based metrics discussed above, the primary difference between the Spearman's correlation [8] and the co-ranking framework [1] based metrics is that the latter performs more detailed comparisons among X and Y by looking into both the *intrusions* and *extrusions* [1] in the embedding. Despite belonging to the co-ranking matrix framework, the four metrics $\mu_{R_{nX}}$, $\lambda_{K_{max}}$, Q_{local}, and Q_{global} assess embeddings from different aspects. For example, $\mu_{R_{nX}}$ assesses the average agreement among all K-ary neighborhoods in the data $Q_{local;}$ and Q_{global} assess the same for smaller (i.e., local neighborhood structure) and larger (i.e., global structure of the dataset) values of K respsctively. On the other hand, $\lambda_{K_{max}}$ analyzes the neighborhood loss instead of agreement. The applicability and utility of the selected metrics in the analytical context of pattern analysis and similarity search are further discussed in Section 4.4.2.

4.1.2.2 DR for Accurate Predictive Modelling

The accuracy metrics for DR can be classified based on the aspect of information loss that they focus upon. Such metrics are of two types: **distance preserving**[2] and **neighborhood preserving** [7] accuracy metrics. Whilst, the distance preserving metrics have been used as objective functions in DR algorithms (e.g., Kruskal's stress function [8] in non-metric Multidimensional Scaling [4]), the neighborhood preserving metrics are more widely deployed [7] for quality analysis. One of the most widely used neighborhood preserving accuracy metric is $AUC_{\ln K}(R_{nX}(K))$ [5]. The metric computing the area under the $R_{nX}(K)$ curve was defined by Lee and Verseylen [5] as a sum of neighborhood preservation for all neighborhood sizes in a logarithmic scale. Formally, the metric $AUC_{\ln K}(R_{nX}(K))$ can be defined as our **seventh metric** as:

$$AUC_{\ln K}\big(R_{nX}(K)\big) = \left(\sum_{K=1}^{n-2} R_{nX}(K)/K\right) \Bigg/ \left(\sum_{K=1}^{n-2} 1/K\right) \qquad (4.12)$$

In our experiments we use $AUC_{\ln K}(R_{nX}(K))$ as our first and direct measure for DR accuracy. Furthermore, following popular research [3], [14], [17], [19], [28], [29],

[2] The primary difference between the **distance preserving** quality metrics and the **distance based** quality metrics discussed in Section 2.2.1 is, the distance based metrics merely compare the relative distances among data-points to evaluate embeddings. Whereas, the distance preserving metrics are used as objective functions that attempt to minimize the discrepancies among relative distances (i.e., the structural properties) in the original dataset and in the embeddings.

we use the K-Nearest Neighbor[3] (KNN) classification [14] accuracy as our second and indirect quality measure for DR accuracy. Formally, assuming the feature matrix and the label matrix for X to be defined as a combination of X_F and X_L so that, the data-points in X can be represented as $\{(x_{F_1}, x_{L_1}), .., (x_{F_n}, x_{L_n})\}$. Here x_{F_i} represents only the feature vector of the data-point x_i with x_{L_i} representing only the training label. Similarly, defining Y as a combination of Y_F and Y_L, where $Y_L = X_L$, the *classification accuracy* ACC_ψ of the KNN classifier ψ for multi-class classification can be defined as our **eighth metric** as follows:

$$ACC_\psi = \frac{\psi(y_{F_i})/y_{L_i}}{n} \tag{4.13}$$

where, $\psi(y_{F_i})/y_{L_i}$ represents the number of predicted correct labels and n represents the number of data-points in the test dataset. The KNN classifier is an appropriate DR accuracy metric as: most NLDR algorithms are based on geometric methods that exploit the concept of locality in the neighborhoods of data-points using nearest neighbor (NN) distances. Similarly, the main idea of KNN is based on the assumption of locality in the data space [19]. We note that ACC_ψ is an indirect quality measure for DR and only applies to labelled data. However, we follow popular practice in academia and include ACC_ψ in the list of our metrics.

4.1.2.3 DR with Poor Quality Input Data

Missing values for data attributes is a common challenge in real-life data analytics. As DR algorithms are usually affected [12] by unattributable missing values in input data, in this research we compare the chosen DR algorithms for their stability

[3] In our experiments we used KNN classifier for both metrics 8 and 11. The primary reason being: similarly, as KNN classifier most NLDR algorithms exploit the concept of locality in the neighborhoods of data-points. Moreover, the same experiments with Random Forest classifier [30] as well as K-means clustering [23] algorithms revealed that for all datasets KNN classifiers produced better results for the same DR techniques. Additionally, as per Hastie et al. [31] the bias for the KNN classifiers remain low during our experiments because following the guidelines of Maaten et al. [3] we train our models with only 1-nearest neighbor. As pointed out by as per Hastie et al. [31], although in this case the variance of the classifiers may remain high, a model with high variance but low bias can make better quality predictions on an average [31] than a model with low variance and high bias.

with datasets containing missing values. Following guidelines from researchers [8], [12], we have used the *normalized mutual information* (nMI) score [12] as our **ninth metric** to assess the stability of the DR methods with missing data. The rank-based metric nMI is computed using entropy [8] and mutual information (MI) [12] metrics. In our experiments, nMI quantitatively assesses the differences between two embeddings for the same dataset one with and one without missing input values. Formally, assuming a data matrix $\hat{X}$ containing B% missing values from X, and $\hat{Y}$ and Y being the low-dimensional embeddings of $\hat{X}$ and X respectively. In our experiments, we design $\hat{X}$ such that both $\hat{X}$ and X are of size $n \times D$ but, $\exists\ \hat{x}_i \in \hat{X}$ where, some $\hat{x}_{ij} = \emptyset$. In $\hat{x}_{ij}$, j represents any position in the vector $\hat{x}_i$ and $j \in 1,2,3, \dots, D$. In order to compute the MI, using the co-ranking framework at first we computed the joint probabilities for Y and $\hat{Y}$ using:

$$p(Y) = \sum_{\hat{Y}} p(\hat{Y}, Y) \qquad and \qquad p(\hat{Y}) = \sum_{Y} p(Y, \hat{Y}) \tag{4.14}$$

where the joint probabilities are computed as $p(\hat{Y} \cap Y) = p(\hat{Y}) \cdot p(Y)$. This is followed by the computation of entropy [12] for both $\hat{Y}$ and Y as:

$$\hat{H}(Y; \hat{Y}) = -\sum_{Y} \sum_{\hat{Y}} p(Y, \hat{Y}) \,.\, \log\left(p(Y, \hat{Y})\right) \tag{4.15}$$

where, entropy $\hat{H}(Y) = -\sum_{Y} p(Y) \,.\, \log\left(p(Y)\right)$ and $\hat{H}(\hat{Y}) = \sum_{\hat{Y}} p(\hat{Y}) \,.\, \log\left(p(\hat{Y})\right)$. Finally, the nMI between $\hat{Y}$ and Y is obtained [8] as:

$$nMI(Y; \hat{Y}) = MI(Y, \hat{Y}) / \tfrac{1}{2} [\hat{H}(Y) + \hat{H}(\hat{Y})] \tag{4.16}$$

where, $MI(Y, \hat{Y}) = \hat{H}(Y) + \hat{H}(\hat{Y}) - \hat{H}(Y; \hat{Y})$ represents the mutual information between Y and $\hat{Y}$.

On the other hand, outlier values in the input data are a known challenge in data analytics. Ideally, a low-dimensional embedding should not be affected by outliers as usually outlier data-points in a high-dimensional manifold reside far away from other points. However, very little research has been done that aims at detecting the sensitivity of DR algorithms to outliers. Formally, assuming a data matrix $\hat{X}$, such that $|\hat{X}| = |X|$. Assuming $B\%$ records in $\hat{X}$ contains outlier values[4] from X and $\hat{Y}$ and Y being the low-dimensional representations of $\hat{X}$ and X respectively. In order to assess the stability of DR algorithms with outliers, we select samples $S_{\hat{Y}}$ from $\hat{Y}$ and S_Y from Y containing only the $(100 - B)\%$ of low-dimensional representations for the non-outlier data-points in $\hat{X}$ so that $|S_Y| = |S_{\hat{Y}}|$. Then, we measured the similarity between S_Y and $S_{\hat{Y}}$ using the Structural Similarity Index [33] (our **tenth metric**) as:

$$SSI(S_Y, S_{\hat{Y}}) = \frac{(2\mu_{S_Y}\mu_{S_{\hat{Y}}}+C_1)(2\sigma_{S_Y}\sigma_{S_{\hat{Y}}}+C_2)}{(\mu_{S_Y}^2+\mu_{S_{\hat{Y}}}^2+C_1)(\sigma_{S_Y}^2+\sigma_{S_{\hat{Y}}}^2+C_2)} \quad (4.17)$$

where, μ_Y represents the mean of Y and $\mu_{S_{\hat{Y}}}$ represents the mean of $S_{\hat{Y}}$. Similarly, σ_Y signifies the standard deviation of the population and $\sigma_{S_{\hat{Y}}}$ signifies the standard deviation of $S_{\hat{Y}}$. Finally, $C_1$and C_2 are constants that are used to stabilize the impact of the division by a weak denominator. As per Wang et al. [33] considering small constants $\kappa_1 = 0.01$ and $\kappa_2 = 0.03$, the constants in Eq. 17, i.e., C_1 and C_2 are defined as $C_1 = (\kappa_1 L)^2$ and $C_2 = (\kappa_2 L)^2$. With L being the dynamic range of the pixel values, in our experiments the values of C_1 and C_2 were 6.553 and 58.982 respectively.

4.1.2.4 DR with Limited Input Data

A stable DR algorithm should be invariant to a subset of data-points or features in the dataset [2], [17]. That is the placement of the points in the low dimensional embedding should be relatively unchanged even for a subset of the data. Researchers have often performed partial observation tests on DR algorithms as their measure of reliability [2], [17]. In our experiments, each dataset X is randomly sampled into equal-sized A horizontal subsets S_i, where $i = 1,2, \dots A$ and $|S_1| = |S_2| = \cdots = |S_A|$. Then the DR algorithms are applied on each subset. Next, in order to check the reproducibility of the algorithms with respect to partial records, we

[4] For our experiments, for each dataset we manually created the $\hat{X}$ as we added the outliers for each D using their Median Absolute Deviation (MAD) [22]. This is a common approach [22] for detecting outliers in data, as [32] that the detection of outliers using the mean and standard deviations of the sample can be highly impacted by the outlier themselves.

assessed the obtained low-dimensional embedding by measuring the logarithmic loss of multi-class classification [30] using the KNN algorithm defined as:

$$logloss = -\frac{1}{|S_i|}\sum_{o=1}^{|S_i|}\sum_{c=1}^{Cl} yt_{o,c} log\,(yp_{o,c}) \tag{4.18}$$

$logloss$ represents our **eleventh metric** where, $|S_i|$ represents the number of samples in each horizontal subset of X. Moreover, assuming Cl is the total number of actual categories of class labels for multiclass classification and o represents each observation in S_i, $yt_{o,c}$ represents the occurrence of the true label for any observation o. That is, $yt_{o,c}$ is 1 if the observation o belongs to class c; otherwise 0. On the other hand, $yp_{o,c}$ represents the probability estimate of the KNN classifier for the observation o belonging to class c. The logarithmic loss is a popular metric [3], [30] for assessing the preserved structure of the data.

4.1.2.5 DR with Limited Computational Resources

Traditionally, computational resources are limited in terms of either time or space. However, with DR being a transformation technique, it can be challenging to experimentally limit the allocated time and space for its execution as – with limited CPU time, the algorithm will crash without completely transforming X, and with limited access to memory, the DR technique will take much longer time than with optimal memory access but will generate the same transformation for X. As a result, considering the amount of memory usage by any DR technique depends on its hyperparameter settings [34], [35] (more specifically, on the number of iterations [3], [11]), in our experiments we simulate this restriction on the availability of resources by limiting the number of iterations [18], [30], [35] for each DR technique. More specifically, $\forall g_i \in G$, we put additional constraints on their optimization functions as follows:

$$\underset{Y \in \mathbb{R}^{d \times n}}{\text{argmin}} \quad f_{g_i}(Y; X) \tag{4.19}$$
$$subject\ to\text{: iter}(g_j)\text{=0.1}*default\ iter(g_j)$$

Then, $\forall N_i \in N$, we compute our **twelfth metric** mean DR accuracy $\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}}$. Here, we assess the level of accuracy of the embedding given the number of iterations $iter(g_j)$ being set to 10% of default $iter(g_j)$ for each of the DR techniques.

4.1.3 Overview of Statistical Significance Testing

Null hypothesis significance testing forms the core of inferential statistics [31], [36]. It is a form of *reductio ad absurdum* that tries to discredit an idea by assuming the idea is true and then showing a contradiction to the idea. For our experiments, assuming $Sc_{G_i N_a M_b}$ and $Sc_{G_j N_a M_b}$ are the scores of the G_i^{th} and G_j^{th} algorithms on the same N_a^{th} dataset and for the same M_b^{th} metric (in our case, $a \in 1, 2, .. 40$, $b \in 1, 2, .. 12$, $i, j \in 1, 2, .. 15$, and $i \neq j$). Also, without loss of generality, assuming in this particular context $Sc_{G_i N_a M_b} > Sc_{G_j N_a M_b}$. The primary goal of statistical significance testing is to determine whether there is enough empirical evidence to claim that the difference in the performances of the G_i^{th} and G_j^{th} algorithms is random or statistically significant [37]. To enable statistical analysis, it is important to run all G algorithms a large enough number of times [36] on different samples of all N datasets for all the M metrics, so that the probability distribution of the scores for each algorithm on each metric can be understood. In order to perform the null hypothesis significance testing, a *null hypothesis* (denoted by H_0) is defined that states that there is no statistically significant difference between the performances of the G_i^{th} and G_j^{th} algorithms. The alternative hypothesis (represented by H_a) on the other hand, states the exact opposite of the null hypothesis H_0; that is, there is a statistically significant difference between the performances of the two algorithms for the selected dataset samples, and hence we must reject H_0. In our case, when comparing G DR algorithms, the null and alternative hypothesis can be defined as:

$$H_0 = \mu_{\overline{Sc}_{G_1 M_b}} = \mu_{\overline{Sc}_{G_2 M_b}} = \cdots = \mu_{\overline{Sc}_{G_{15} M_b}}$$

$$H_a = \mu_{\overline{Sc}_{G_i M_b}} \neq \mu_{\overline{Sc}_{G_j M_b}} \qquad \text{for at least one pair of i and j (where } i \neq j) \qquad (4.20)$$

In equation 4.20, $\mu_{\overline{Sc}_{G_i M_b}}$ represents the mean performance scores for each algorithm G_i for each metric M_b for all N datasets. In order to decide on whether or not to reject the null hypothesis for a statistical test, the $p - value$ is computed along with the test statistics [38]. The p-value is the probability of obtaining a result equally or even more extreme [31], considering H_0 is true. In case the p-value is smaller than a predefined nominal significance level α, H_0 is rejected. A typical value for α that is commonly used [37], [39] in statistical experiments is 0.05. When performing statistical significance testing, one needs to consider two types of possible errors [38] that could occur during the tests. The *Type I error* represents the situation when H_0 is rejected despite being true, and the *Type II error* relates to the situation when H_0 is accepted despite being false. Here, an important point to note is, the two errors contradict each other, as reducing the probability of one error would increase the probability of the other. However, a common consensus among researchers [36], [38], [39] specify that it is more important to prevent Type I errors than Type II.

Analysis of statistical significance can be done using either parametric or non-parametric statistical tests. Among the two, whilst the former make presumptions about the underlying distribution of the data (i.e., the data must follow a parametric distribution [31]), the latter make no such assumptions. On the other hand, parametric statistical tests are known [37] to be more powerful than their non-parametric counterparts when limited data is available. However, it is often challenging to run parametric tests on samples as the normality of the underlying data cannot be guaranteed. Although according to the central-limit theorem one can claim that the collection of large enough number of random variables converge [40] to a normal distribution, often the access to a large number of data samples is limited. Moreover, although normality tests [31] can be used to assess whether the distribution of the test samples is normal, often these tests produce unreliable results [37]. Hence, in case the amount of available analysis data is not a challenge, Arcuri et al. [36] suggest the use of non-parametric statistical tests with at least $N = 30$ samples. As a result, in this chapter, for each evaluation metric, we use non-parametric statistical tests[5] to perform both pairwise and overall comparisons

[5] It can be argued that, *mean* being a parametric measure for a sample of N random variables, might not be appropriate for non-parametric statistics. However, research shows that in most cases a

between the algorithms.

4.1.3.1 Pairwise comparisons of Algorithms

The most common way to perform a pairwise comparison of two algorithms is to use the paired t-test. Although being a parametric statistical test, the t-test forms the basis of pairwise comparisons among algorithms [31], that examines whether the mean difference between the performances of the two algorithms is significantly greater than 0. Formally, assuming $Sc_{G_iN_aM_b}$ and $Sc_{G_jN_aM_b}$ are the scores of the G_i^{th} and G_j^{th} algorithms for the M_b^{th} metric on the N_a^{th} dataset and the difference between the two be represented by $\delta_{N_a} = Sc_{G_iN_aM_b} - Sc_{G_jN_aM_b}$, then the t-statistic is computed as:

$$t = \overline{\delta_{N_a}}/\sigma_{\delta_{N_a}} \tag{4.21}$$

Where, $\overline{\delta_{N_a}}$ represents the mean difference of all N datasets, where $a \in 1, 2, .. N$ and $\sigma_{\delta_{n_a}}$represents their standard deviation. The t-statistic follows the Student's t-distribution with $N - 1$ degrees of freedom for N datasets and in case the p-value for the t-statistic is less than the α, the null hypothesis H_0 is rejected.

To avoid the normality assumptions of the paired t-test, the Wilcoxon signed ranks test is chosen [31] as its most appropriate non-parametric counterpart. The Wilcoxon signed ranks test [37] is used to compare the absolute differences in performances of the two algorithms to determine if the mean ranks of the positive and negative differences between the performance scores are statistically

normal approximation of the original distribution of the sample values is used with a reference to the central limit theorem. Moreover, researchers such as Sherman et al. [40] have proven that, in case of large enough number of random samples the error for approximation can be minimal. In our experiments, in order to avoid any inconsistencies, (1) we use 10,000 random samples of each dataset to calculate each metric; so that the scaled mean of the random samples can converge into a normal distribution with minimal approximation error. (2) In order to ensure a symmetric skewness of the distribution of each random sample, we calculated the Median Absolute Deviation (MAD) of the sample. In our experiments, we did not encounter a situation where the MAD showed any asymmetry in the distributions of the random samples, possibly due to the largeness of the sample population. As a result, in our experiments, following the footsteps of Demśar et al. [39] and Mohammadi et al. [37] we used the sample *means* for the non-parametric statistical tests.

significant. Formally, following the same notations for $Sc_{G_i N_a M_b}$, $Sc_{G_j N_a M_b}$, and δ_{N_a}as discussed above, ranking the differences based on their absolute values, and assuming r^+ be the sum of ranks where $\delta_{N_a} > 0$ and r^- be the sum of ranks where $\delta_{N_a} < 0$; and the cases for $Sc_{G_i N_a M_b} = Sc_{G_j N_a M_b}$are split equally among the two groups, we can define $r^+$and r^-as:

$$r^+ = \sum_{\delta_{N_a}>0} rank(\delta_{N_a}) + \tfrac{1}{2}\sum_{\delta_{N_a}=0} rank(\delta_{N_a}) \tag{4.22}$$

$$r^- = \sum_{\delta_{N_a}<0} rank(\delta_{N_a}) + \tfrac{1}{2}\sum_{\delta_{N_a}=0} rank(\delta_{N_a})$$

Considering, $T = \min(r^+, r^-)$, the test statistic z for the Wilcoxon signed ranks test for the number of datasets N > 25 is calculated as:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{4.23}$$

The test statistic for Wilcoxon signed ranks test approximately follows the normal distribution. The null hypothesis H_0 is rejected in case the p-value corresponding to the test statistic is less than the threshold α. As per Demšar et al. [39], Wilcoxon signed ranks test is safer than paired t-tests as not only it does not make any assumptions regarding the normal distribution of the data, but also, outliers do not impact the Wilcoxon test as much as the paired t-test. Following the guidelines of Demšar et al. [39], in our experiments alongside t-tests, we use the Wilcoxon signed ranks test for pairwise comparisons among the DR algorithms.

In order to validate the efficiency of the Wilcoxon test, we also calculate the McNemar's test statistic [37] for pairwise comparisons among the algorithms. When comparing two algorithms G_i and G_j for metric M_b over N datasets, considering $N_{01}$is the number of datasets where the performance of G_j is better than G_i and N_{10} is otherwise, The McNemar's asymptotic test statistic is calculated as:

$$\chi^2 = \frac{(N_{01} - N_{10})^2}{N_{01} + N_{10}} \tag{4.24}$$

The statistic follows a χ^2 distribution with one degree of freedom under H_0.

4.1.3.2 Comparisons of Multiple Algorithms

When comparing the performances of all G algorithms for a given metric M_b, the null hypothesis H_0 is there are no differences in the overall performances of any of the G algorithms for the metric M_b. The alternative hypothesis H_a is, there is at least one algorithm among all the algorithms that behaves differently than others. In order to identify such algorithms, researchers often use repeated measures of pairwise comparisons (e.g., repeated paired t-tests) among each pair of G_i and G_j algorithms [39]. However, this approach often suffers from the challenge of *multiple hypothesis testing* [31], where the probability of the Type I error increases with the number of comparisons. In order to avoid such problems, Demśar et al. [39] have suggested making adjustments to the threshold α so that the p-value of the test statistic is also adjusted to accommodate all pairs of hypotheses. However, often these adjustments lack statistical power due to their conservative nature [37]. As a result, more specialized algorithms exist to evaluate the statistically significant differences between multiple algorithms. Among such null hypothesis significance testing techniques, initially, omnibus tests such as the ANOVA (parametric) or Friedman's test (non-parametric) tests are used to identify if there is at least one algorithm whose performance is significantly different than the others. These omnibus tests are then followed by post-hoc [37] tests, as the omnibus tests do not identify the algorithm that is different.

While comparing multiple algorithms, the most common statistical test used is repeated measures of ANOVA (Analysis of Variance) [31] to compare the differences between the mean performances of multiple algorithms on a chosen metric. ANOVA considers the variability within each sample (i.e., for $Sc_{G_iN_aM_b}$ it is the variability in the scores for the same G_i) and between the samples (i.e., for $Sc_{G_iN_aM_b}$it is the variability in the scores of all G_i in G) in order to distinguish the statistical significance between their means. ANOVA being a parametric test, the underlying assumptions regarding the normality of the sample values determine the reliability of the results of the test. Hence, researchers such as Demśar et al. [39] and Mohammadi et al. [37] suggest, the non-parametric counterpart of repeated measures of ANOVA, the Friedman test should be used.

In the Friedman test instead of using the actual performance scores (i.e., $Sc_{G_iN_aM_b}$) for evaluating the algorithms, at first, the scores are ranked based on their value with 1 being the highest rank. For algorithms that produce the same scores, mean ranks are assigned. Formally, assuming $Fr_{G_iN_aM_b}$be the Friedman rank of the G_i^{th} algorithm on the N_a^{th} dataset for the M_b^{th} metric. The Friedman statistic compares the mean ranks $\mu_{\overline{Fr}_{G_iN_aM_b}}$ of all algorithms on the M_b^{th} metric. Considering H_0 the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{G(G+1)}\left[\sum_i \mu_{\overline{Fr}_{G_iN_aM_b}}{}^2 - \frac{G(G+1)^2}{4}\right] \tag{4.25}$$

Although the test statistic reduces the probability of the Type I error, it increases the probability of the Type II error. Hence, the test statistic was updated [39] by Iman and Davenport as:

$$Fr = \frac{(N-1)\chi_F^2}{N(G-1)-\chi_F^2} \tag{4.26}$$

In situations where the null hypothesis is rejected for comparisons among multiple algorithms, to identify the algorithms that differ from others, post-hoc tests for pairwise comparisons among the ranks of algorithms are performed. However, in this situation due to the challenge of multiple hypotheses testing, adjustments to the p-values and the threshold α are made. The most common adjustment to the p-values is the Nemenyi correction [37]. In this adjustment, the threshold α and the p-values are divided by the total number of pairwise comparisons among the algorithms. Nevertheless, Nemenyi adjustment being conservative, in some cases it can increase the probability of the Type II error. Hence, an alternative to the Nemenyi correction is the Holm procedure [37]. This procedure iteratively selects the most significant p-value among all the test statistics of pairwise combinations and divides the p-value with $\frac{\alpha}{G-1}$. The corresponding null hypothesis is rejected in case the selected p-value is less than $\frac{\alpha}{G-1}$. In the next iteration, the next most significant p-value is selected. In our experiments we perform both Nemenyi and

Holm adjustments and compare the results.

4.2 Experimental Setup

In our experiments, we compared 15 DR techniques by empirically evaluating them on 40 high-dimensional real-life datasets. Subsequently, we assessed the quality of the low dimensional embedding obtained from applying the algorithms for 12 different contextual metrics. In this section, we discuss the setup for our experiments. At first, we discuss the algorithms that are compared during our experiments, then the datasets that were chosen for our study, and finally, we explain our experimental procedure in detail.

4.2.1 Algorithms

DR being a popular research area, a large number of techniques exist that assist with transforming high-dimensional datasets into their low dimensional embedding. On the one hand, whilst some of the existing DR algorithms have existed for more than twenty years (e.g., PCA [41], non-metric MDS or nMDS [42]) that are still used [11], some are newly proposed algorithms (e.g., t-SNE [26], UMAP [2], LargeVis [13]) that have become exceedingly popular in the past few years. On the other hand, as discussed earlier, different DR algorithms have different properties (e.g., linear or non-linear) that approach the problem of reducing data dimensions differently. Making empirical comparisons among all existing DR algorithms would be beyond the scope of any single research. Hence, for our experiments, we included 15 state-of-the-art DR algorithms[6] from a large umbrella of categories. The algorithms were selected as a blend of both long-existing and newly proposed algorithms with different properties. Table 4.2 presents a list of the selected algorithms along with their properties and respective parameter settings used in our experiments. It is important to note that, in our

[6] PCA and most NLDR techniques (e.g., t-SNE, UMAP, MDS, Isomap, KernelPCA etc.) are unsupervised [24], [43]. That is the existence of labels in the input data neither has any impact on the proximity detection among points nor on the overall transformation. Nevertheless, in the past few years in order enrich the embeddings with class-separations, supervised versions of some popular NLDR techniques (e.g., t-SNE, nMDS, Isomap UMAP, LEM, LLE) were proposed. However, in the scope this research, we focus on the traditional unsupervised versions of the NLDR algorithms.

experiments only one linear and 14 NLDR techniques were considered. The primary reason being as we only work with real-world datasets and linear DR techniques cannot effectively handle [3] the highly non-linear nature of such data, following the guidelines of Maaten et al. [3] we mainly focus on NLDR techniques. However, PCA being the most popular linear DR technique, we also include PCA in our experiments.

Alongside the properties, in Table 4.2 we also discuss the computational complexity of each algorithm. The computational complexity of DR is an important aspect of an algorithm as it helps us to determine the feasibility of using the algorithm on large datasets. Table 4.2 shows that in the case of only PCA the complexity depends on the number of original dimensions in the input dataset. On the other hand, KPCA [44] and nMDS [42] seem to be the most expensive algorithms in terms of computational cost and LargeVis [13], 8 Core-t-SNE [45], and FIt-SNE [18] seem to be the least expensive with linear complexity. The other three columns presenting the properties of the algorithms specify whether the algorithms are linear or non-linear, locally or globally focused, and matrix factorization or neighborhood analysis based. The parameter settings of the listed algorithms specify the range of values used for their hyper-parameters in our experiments. These parameter settings were determined using an exhaustive grid-search. Such grid-search techniques have been popularly used [3], [46] by researchers when determining parameters of different algorithms. The target dimensionality of all the algorithms were decided for the datasets using the maximum likelihood intrinsic dimensionality estimator [47] following a common practice among researchers [3] for evaluating DR algorithms.

4.2.2 Datasets

In order to evaluate the selected DR algorithms, we have compiled 39 real-world datasets, from a wide range of open-source data repositories such as UCI Machine Learning Library [48], Kaggle Data Repository [49], National Cancer Registration

Table 4.2: Properties and Parameter Settings for the Chosen DR Algorithms

Technique	Properties				Parameter Settings of the Algorithms for Experiments
	Complexity	Linear/ nonlinear	Local/ global	Approach	
UMAP	$O(dn^{1.14})$	nonlinear	local	neighbor analysis	$5 \leq k \leq 25, 0.5 \leq \text{min\_dist} \leq 0.99$
t-SNE	$O(n^2)$	nonlinear	local	neighbor analysis	$5 \leq \text{perplexity} \leq 30$
FIt-SNE	$O(n)$	nonlinear	local	neighbor analysis	$5 \leq \text{perplexity} \leq 30$
PCA	$O(D^3)$	linear	local	matrix factorization	none
Trimap	$O(n^{1.14})$	nonlinear	global	triplet mapping	$25 \leq k \leq 50$
McoretSNE	O(logn)	nonlinear	local	neighbor analysis	$5 \leq \text{perplexity} \leq 30, \text{n\_jobs} = 8$
Isomap	$O(n^3)$	nonlinear	global	neighbor analysis	$5 \leq k \leq 15$
KPCA	$O(n^3)$	nonlinear	global	matrix factorization	$\kappa = (XX^T + 1)^5$
LEM	$O(dn^2)$	nonlinear	local	neighbor analysis	$5 \leq k \leq 15, \sigma = 1$
LTSA	$O(dn^2)$	nonlinear	local	neighbor analysis + PCA	$5 \leq k \leq 15$
nMDS	$O(n^3)$	nonlinear	global	matrix factorization	$300 \leq \text{max\_iter} \leq 500$
HLLE	$O(dn^2)$	nonlinear	local	neighbor analysis	$5 \leq k \leq 15$
LLE	$O(dn^2)$	nonlinear	local	neighbor analysis	$5 \leq k \leq 15$
LargeVis	$O(n)$	nonlinear	global	neighbor analysis	$5 \leq k \leq 25$
MVU	$O((nk)^3)$	nonlinear	global	neighbor analysis	$5 \leq k \leq 15$

Note: In the table above the symbols n, D, d, and k represent the number of samples in the data, the dimensionality of the high-dimensional dataset, the intrinsic dimensionality of the dataset, the number of nearest neighbors respectively.

Full forms of acronyms are as follows: UMAP: Uniform Manifold Approximation and Projection, t-SNE: t-distributed Stochastic Neighbor Embedding, FIt-SNE: Fast Fourier Transform Accelerated Interpolation Based t-SNE, PCA: Principal Component Analysis, Trimap: Triplet Mapping, McoretSNE: Multicore t-SNE with 8 cores, Isomap: Isometric Feature Mapping, KPCA: Kernel PCA, LEM: Laplacian Eigenmap, LTSA: Local Tangent Space Alignment, nMDS: non-metric Multidimensional Scaling, LLE: Local Linear Embedding, HLLE: Hessian LLE, LargeVis: Visualizing Large Scale and High Dimensional Data, MVU: Maximum Variance Unfolding, min_dist: minimum distance, max_iter: maximum number of iterations.

and Analysis Service (NCRAS)[7], European Structural and Investment Funds[8], Taylor & Francis [9,] Open Government Canada[10], Australian Government Open Datasets[11], data.world repository, and Figshare data repository. Besides, we obtained one (closed) real-world dataset from our industrial partner IBM. Table 4.3 summarizes the overall statistics and associated analytical tasks of the selected

[7] http://www.ncin.org.uk/

[8] https://ec.europa.eu/

[9] https://www.tandfonline.com/

[10] https://open.canada.ca

[11] https://data.gov.au/

datasets. 14 of the chosen open-source datasets belong to the OpenML [50] dataset compilation. To strategically simulate our experiments, we decided to build a common set of selection criteria for the datasets that could be used in our study. Firstly, we chose tabular datasets with numeric or categorical data, as DR techniques [2], [17], [26], [51] are not directly applicable to textual data. Secondly, as shown in Table 4.3, we selected datasets that contained at least 10,000 records. The primary reason behind this selection criterion was, Mohammadi et al. [37] show that, for statistical significance tests to be reliable, they not only require enough number of data samples to compare, but also require a sufficient number of executions [36] for different samples from the same dataset. Finally, based on the guidelines of Demšar et al. [39], our last selection criteria was that we only choose real-world datasets and ignored artificial data. As artificially created datasets usually make certain assumptions regarding the data distributions of real-world datasets, this could add further bias to the analysis. A point to note is, all the selected datasets belong to business, computer, physical, medicine, geological, or social sciences domain and are not image data. We purposefully ignored working with image datasets as a large portion of the existing research [3], [17] on the evaluation of DR methods are focused only on image datasets. Moreover, the datasets used in our experiments were designed for classification, regression, or clustering purposes.

4.2.3 Experimental Methodology

Once the algorithms and the datasets were finalized, we executed all the algorithms for each dataset and empirically evaluated them for each of the contextual metrics discussed in Section 4.1.2. Next, in order to identify the replicability of the obtained results, we performed statistical significance testing. Formally, as discussed in Section 4.1.2, in this experimental study, $G = 15, N = 40, and\ M = 12$. Algorithm 4.1 presents the overall procedure of our experimental methodology. As shown in Algorithm 4.1, our entire experimental process is divided into two phases. In the first phase, we record the performance of each of the G algorithms on all datasets for each of the contextual metrics M (cf. Algo:4.1, lines: 1-15). In order to

Table 4.3: Statistics of the Datasets

No.	Dataset	# Records	# Dimensions	Domain	Source	Associated Tasks (Classification - Clsf, Clustering – Clst, Regression - Reg)
1	Renewal Sales	1,354,704	15	Business	IBM	Clsf, Clst
2	Poker-hands	1,025,010	11	Social	Kaggle	Clsf
3	SUSY	1,000,000	18	Physical	UCI repository	Clsf
4	Online Retail	541,909	8	Business	UCI repository	Clsf, Clst
5	Geo Unit Area	399,787	162	Geological	https://data.world	Clst
6	Black Friday	166,821	10	Social	Kaggle	Clsf, Clst
7	Weather Australia	142,193	24	Social	Kaggle	Clsf
8	Postures	78,095	38	Computer	Kaggle	Clsf, Clst
9	Connect-4	67,557	42	Social	UCI repository	Clsf
10	Travel Insurance	63,326	11	Social	Kaggle	Clsf, Clst
11	Aps-Failure	60,000	171	Computer	UCI repository	Clsf
12	Cancer Diagnosis	58,972	88	Medicine	NCRAS	Clst
13	Shuttle	58,000	9	Physical	UCI repository	Clsf
14	TAFL_LTAF	51,749	34	Business	https://open.canada.ca	Clsf, Clst
15	Adult	48,842	14	Social	Kaggle	Clsf
16	Virome Assessment	48,804	68	Medicine	Figshare	Clst
17	Electricity	45,312	9	Social	UCI repository	Clsf, Reg
18	Bank	45,211	17	Business	UCI repository	Clsf
19	ESIF 2014-20	43,592	64	Business	https://ec.europa.eu/	Clsf
20	News	39,797	61	Business	UCI repository	Clsf, Reg
21	Nomao	34,465	120	Computer	UCI repository	Clsf
22	Amazon Emp-Access	32,769	10	Social	Kaggle	Clsf, Clst
23	Credit Card	30,000	24	Business	Kaggle	Classification
24	RBM10 lung cancer	26,021	112	Medicine	Figshare	Clst
25	E-commerce	23,486	10	Social	Kaggle	Clsf, Clst
26	Electricity Australia	22,952	49	Social	https://data.gov.au/	Clsf, Clst
27	Occupancy Detection	20,560	7	Computer	UCI repository	Clsf
28	Letter Recognition	20,000	16	Computer	UCI repository	Clsf
29	Magic	19,020	11	Physical	Kaggle	Clsf
30	HTRU	17,898	9	Physical	UCI repository	Clsf, Clst
31	Mozilla4	15,546	6	Computer	OpenML	Clsf
32	Brain Tumor	15,273	20	Medicine	Taylor & Francis	Clst
33	Bankruptcy	15,002	65	Business	Kaggle	Clsf, Clst
34	EEG_eye_state	14,980	15	Social	UCI repository	Clsf
35	Sylva Agnostic	14,395	217	Computer	OpenML	Clst
36	Gas_Drift	13,910	129	Computer	UCI repository	Clsf
37	Shoppers Intension	12,330	18	Business	UCI repository	Clsf, Clst
38	Epileptic Seizure	11,500	179	Medicine	UCI repository	Clsf, Clst
39	Phishing Websites	11,056	32	Computer	UCI repository	Clsf
40	JM1	10,885	22	Computer	OpenML	Clsf, Clst

do this, before transforming each dataset X to its embedding Y, we estimate the intrinsic dimensionality d of X using the Levina–Bickel's technique [47] for maximum likelihood intrinsic dimensionality estimator defined as:

$$\hat{d} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{d}_k \qquad where,\ \hat{d}_k = \frac{1}{n} \sum_{i=1}^{n} d_k(X) \tag{4.27}$$

where, $\hat{d}$ represents a unit vector with an estimation for d, and $(k_2 - k_1)$ signifies the range of nearest neighbors that were considered while estimating d. This preprocessing is necessary [3], [47], [52] as, the estimation of d prior to obtaining Y not only ensures noise reduction [52], [53] in Y, but also enhances the stability [47] of Y. Next, we execute the algorithms on 10,000 simple random samples (i.e., without replacement) of each dataset, calculate the scores $S_{G_i N_a M_b}$ for each context, and record the mean scores $\mu_{\bar{S}_{G_i N_a M_b}}$ for each metric. The sample size u_a for each dataset is selected in the preprocessing step using the chosen confidence level (cf. Algo:4.1, line: iii), with the formula defined by Daniel et al. [54] as:

$$u_a = \frac{|N_a| * E}{(E + |N_a| - 1)} \tag{4.28}$$

$$\text{where, } E = {Z_{\alpha/2}}^2 * pr * (1 - pr) \Big/ MOE^2$$

and, $Z_{\alpha/2}$ is the critical value of the normal distribution at α/2 (e.g. for a confidence level of 95% and α of 0.05, the predefined critical value is 1.96), MOE is the margin of error, pr is the sample proportion, and $|n_a|$ is the population size (i.e., the cardinality of the original dataset $n_a \in N$). One might argue that the viability scores for datasets from different domains might not commensurate. Following the guidelines of Demśar et al. [39], we address this challenge by restricting the domains, size, and type of the selected datasets. Moreover, researchers [37] have also argued that sometimes the mean score over many datasets might susceptible to outliers. That is, an algorithm's poor performance on one dataset can affect the mean of its overall performance. To avoid this problem, as discussed by Arcuri et al. [36] we selected large enough number (i.e., 40) of datasets to regularize the impact of an algorithm's unusual performance on only one dataset. Apart from the mean scores $\mu_{\bar{S}_{G_i N_a M_b}}$ for each metric (cf. Algo:4.1, line:13), we also record that standard deviation $\sigma_{S_{G_i N_a M_b}}$ for the performances of each algorithm on each dataset.

ALGORITHM 4.1: Experimental Procedure

Input: Datasets $N = \{N_1, N_2, \dots, N_{40}\}$, DR algorithms $G = \{G_1, G_2, \dots, G_{15}\}$, and Metrics $M = \{M_1, M_2, \dots, M_{12}\}$

Output: Values for M for all G, Test statistics for null hypothesis significance testing for all G

Preprocessing:

```
foreach $N_i \in N$ do:
    Identify intrinsic dimensions using maximum likelihood intrinsic dimensionality estimator
    Given the confidence score identify sample size for experiments
    foreach $G_i \in G$ do:
        Exhaustive grid-search to identify the values for their hyper-parameters
    end
end
```

Process:

```
foreach $N_i \in N$ do:
    for $i$ in range (1 to 10000):
        Select $S_i$ samples from $N_i$, where, $S_i \subseteq N_i$
        foreach $G_i \in G$ do:
            Set hyper-parameters of $G_i$ for $N_i$
            Execute $G_i$ with $N_i$ as input
            foreach $M_i \in M$do:
                Calculate contextual metric $M_i$ (cf. Table 4.1)
            end
        end
    end
    foreach $G_i \in G$, $N_i \in N$, and $M_i \in M$do:
        Calculate mean $\mu_{\bar{S}_{G_iN_aM_b}}$and standard deviation $\sigma_{S_{G_iN_aM_b}}$
    end
end
foreach $G_i \in G$ and $N_i \in N$ do:
    Calculate Friedman's rank for all algorithms (cf. Section 4.1.3)
    Determine selection bias $v$ for all datasets (cf. Equation 4.29)
end
for $v$ in range (0 to 20) do:
    for i in range (0 to 1000):
        Select 40 datasets at random with given selection bias $v$
        Run all statistical tests on each selected dataset (cf. Section 4.1.3 - Equations 4.21 to 4.26)
    end
end
```

A small $\sigma_{S_{G_iN_aM_b}}$helps us to identify the stability of the algorithm's performance on all N datasets.

In the second phase of the experiments (cf. Algo:4.1, lines:16-25), we perform null hypothesis significance testing of the experimental results obtained from phase one. Here, we consider the previously recorded $\mu_{\bar{S}_{G_iN_aM_b}}$and $\sigma_{S_{G_iN_aM_b}}$values for the algorithms to define our hypothesis for the tests. Firstly, we rank the performances of each algorithm (cf. Algo:4.1, line:17) using Friedman's ranking method discusses in Section 4.1.3. The primary goal of this phase is to identify the best, mediocre, and worst-performing algorithms for a given metric, and test if the obtained results are statistically significant. To do this, we simulate the next phase of our experiments to identify the advantages or disadvantages of a specific algorithm over the others. Since, no algorithm can be optimal for all forty datasets, for this simulation, following Demšar et al. [39] and Mohammadi et al. [37], we iteratively select 40 datasets with replacement[12] for each pairwise and overall comparisons (cf. Section 4.1.3) among the algorithms. In this selection process, we add an additional bias v (cf. Algo:1, line: 18) to each dataset using a logistic function [39], so the probability of each dataset being selected is proportional to:

$$\frac{1}{1+e^{-v\left(S_{G_iN_aM_b}-S_{G_iN_cM_b}\right)}}, \mathit{where}\ a,c \in 1,2,\dots,40\ \mathit{and}\ a \neq c \qquad (4.29)$$

and, v is the bias through which we regulate the selection of the datasets, and $S_{G_iN_aM_b} - S_{G_iN_cM_b})$ is the positive or negative differences between the performances of each dataset. The introduction of additional selection bias in the form of weighted random sampling [55] is a commonly practiced approach [55]–[58] in data analysis. The primary reason behind adding weights (or bias) to experimental input is: often as the input data is not uniformly distributed, a uniform random selection of samples may produce irrelevant results [56]. Similarly, in our case, the performance scores of the algorithms on none of our chosen metrics are

[12] The repetitive sampling of 40 datasets with replacement resembles bootstrapping [31]. However, the fundamental difference between our sampling technique and bootstrapping is the introduction of the selection bias. In bootstrapping all samples in the population have the same probability for being selected [31], whereas in our case for each non-zero value for bias, some datasets have more probability of being selected over others.

distributed uniformly, hence introducing a selection bias on the datasets for statistical significance analysis can help us put more weight on datasets that are in favor of the better performing algorithm. Moreover, a gradual increase in the selection bias can show [39], [59], [60] the differences among the two samples more prominently. Among several existing temporal bias functions [57], in case of statistical significance analysis, the most common practice [37], [39], [59], [60] is to use the exponential function to simulate a biased selection of datasets. Hence, we follow the footsteps of Demśar et al. [39], Garćia et al. [59], [60], and Mohammadi et al. [37], and use the exponential bias function in our experiments.

For each pair of algorithms, the selection of datasets was repeated 1000 times (cf. Algo:4.1, line:21). We varied the value of bias v from 0 to 20 and the same statistical experiments were performed on the selected datasets for each bias. For example, when $v = 0$, each dataset has the same probability to be picked as the selection is performed uniformly at random. However, as the bias increases, the datasets in favor of a specific algorithm have a higher probability of being selected. It is important to note that, as this type of simulated experiment is a common practice among researchers [37], [39], we employed this practice only for the simulation purposes for measuring the statistical significance of one algorithm's performance over the others. Whilst overall comparisons were performed among all algorithms, for the pairwise comparisons, in a single iteration, we only compare the best performing algorithm with the worst-performing algorithm.

In order to validate our performed statistical tests, we calculated the power of each statistical test. As discussed by Demšar et al. [39], the power of any null hypothesis significance test is determined by its ability to reject a false null hypothesis. This power is usually associated with replicability [37] of statistical significance tests. Researchers often follow two different ways to analyze the replicability of the statistical tests. Firstly, by comparing the number of rejected null hypothesis during a large number of experiments; and secondly, by assessing the average p-value for each test statistic for a substantial number of experiments. The first type of replicability measure was defined by Bouckaert et al. [61] based on the average number of rejected null hypotheses in a series of experiments as:

$$Rep(e) = \sum_{1 \leq i \leq j \leq n} \frac{I(e_i - e_j)}{n(n-1)/2} \tag{4.30}$$

Where I is the indicator function and e_i is the (binary) outcome of the ith experiment (i.e., $e_i = 0$ if the H_0 is rejected for the ith experiment, and $e_i = 1$ otherwise). The second type of replicability measure was defined by Demšar et al. [39] based on the average p-value for each experiment as:

$$Rep(p) = 1 - 2\frac{\sum_i (p - \bar{p})^2}{n-1} \tag{4.31}$$

Where, $\bar{p}$ is the mean p-value and p_i is the p-value for the ith experiment. In our experiments, we use both the replicability measures.

4.3 Experimental Results

In this section, we present our detailed experimental results. This section is primarily divided into two parts; in the first part, we present the scores of each of the 12 metrics on the 40 datasets using Table 4.4.a to 4.4.l. Then, we summarize and discuss the overall performances of the 15 algorithms (cf. Section 4.2.1) for the 12 derived metrics in Table 4.5. We also analyze the impact of additional factors (i.e., the input datasets and the hyperparameter combinations) that may have influenced the performances of the DR techniques. In the second part of this section, we present a thorough statistical significance analysis of our results.

4.3.1 Performance Analysis of DR Algorithms

The experimental results are summarized in Table 4.5. Since scale varies across the quality metrics used for our experiments, in Table 4.5 we rank the performance of each DR algorithm using the Friedman ranking [37] method discussed in Section 4.1.3. In Table 4.5, each row represents a contextual metric as each column signifies a different DR algorithm. In the table, the metrics are grouped based on their respective analytical contexts. Overall, Table 4.5 shows that: in terms of preservation of local proximity relationships, t-SNE shows the most robust performance with UMAP following closely behind. For metrics such as residual

Table 4.4: Performance scores and Friedman ranking for all twelve evaluation metrics for each algorithm; The performance scores for each dataset represents the mean score for its 10,000 samples.

Note: In the tables below, the datasets follow the same order as presented in Table 4.3

(a) Metric 1: Residual Variance ($\hat{\sigma}^2$)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	1.00	1.00	0.88	0.94	0.46	0.73	0.50	0.97	0.58	0.66	0.26	0.42	0.59	0.96	0.59
2	1.00	1.00	0.91	0.92	0.08	0.84	0.74	0.77	0.66	0.77	0.84	0.69	0.39	0.83	0.24
3	1.00	1.00	0.94	0.85	0.02	0.89	0.79	0.94	0.34	0.79	0.82	0.58	0.39	0.91	0.26
4	0.96	0.95	0.87	0.85	0.02	0.92	0.65	0.73	0.54	0.64	0.45	0.28	0.48	0.78	0.08
5	1.00	1.00	0.02	0.04	0.05	0.74	0.03	1.00	0.01	0.50	1.00	0.27	0.46	0.40	0.19
6	0.98	0.98	0.96	0.82	0.00	0.86	0.39	0.95	0.37	0.68	0.44	0.38	0.40	0.94	0.71
7	1.00	1.00	0.94	0.86	0.01	0.91	0.82	0.86	0.05	0.95	0.24	0.20	0.61	0.81	0.82
8	1.00	1.00	0.93	0.84	0.16	0.94	0.78	0.80	0.65	0.95	0.22	0.59	0.23	0.73	0.03
9	1.00	1.00	0.97	0.89	0.25	0.67	0.74	0.95	0.70	0.59	0.90	0.79	0.65	0.92	0.57
10	1.00	1.00	0.92	0.85	0.09	0.97	0.77	0.87	0.23	0.06	0.80	0.24	0.46	0.38	0.17
11	1.00	1.00	0.24	0.56	0.06	0.77	0.12	0.97	0.49	0.04	0.56	0.59	0.46	0.30	0.23
12	1.00	1.00	0.13	0.70	0.03	0.41	0.56	0.96	0.71	0.77	0.03	0.98	0.01	0.29	0.19
13	1.00	1.00	0.94	0.87	0.11	0.93	0.78	0.86	0.46	0.96	0.55	0.65	0.53	0.70	0.96
14	1.00	1.00	0.95	0.95	0.15	0.66	0.74	0.88	0.77	0.72	0.94	0.77	0.56	0.91	0.12
15	1.00	1.00	0.63	0.41	0.05	0.80	0.52	0.94	0.96	0.28	1.00	0.64	0.34	0.62	0.50
16	1.00	1.00	0.79	0.74	0.01	0.83	0.25	0.94	0.03	0.23	0.69	0.59	0.29	0.52	0.18
17	1.00	1.00	0.89	0.80	0.05	0.98	0.73	0.89	0.59	0.97	0.88	0.39	0.42	0.66	0.72
18	1.00	1.00	0.92	0.64	0.02	0.65	0.81	0.94	0.27	0.01	0.99	0.42	0.60	0.83	0.03
19	1.00	1.00	0.61	0.75	0.01	0.62	0.35	0.95	0.57	0.66	0.99	0.49	0.73	0.99	0.11
20	1.00	1.00	0.82	0.73	0.03	0.66	0.77	0.84	0.88	0.73	0.09	0.64	0.38	0.96	0.40
21	0.98	0.99	0.89	0.44	0.07	0.90	0.71	0.74	0.76	0.45	0.83	0.33	0.36	0.67	0.05
22	0.87	0.85	0.79	0.78	0.17	0.84	0.79	0.54	0.32	0.14	0.75	0.50	0.42	0.52	0.04
23	0.98	0.99	0.59	0.47	0.09	0.95	0.51	0.94	0.40	0.38	0.64	0.67	0.27	0.36	0.03
24	1.00	1.00	0.86	0.46	0.01	0.47	0.07	0.98	0.64	0.49	0.30	0.83	0.00	0.99	0.53
25	1.00	1.00	0.89	0.83	0.06	0.96	0.70	0.87	0.39	0.48	0.89	0.55	0.30	0.59	0.15
26	1.00	1.00	0.96	0.94	0.13	0.73	0.71	0.79	0.61	0.00	0.91	0.64	0.63	0.92	0.01
27	1.00	1.00	0.87	0.84	0.14	0.92	0.72	0.91	0.77	0.11	0.36	0.13	0.46	0.57	0.22
28	1.00	1.00	0.94	0.80	0.10	0.91	0.70	0.86	0.34	0.92	0.85	0.47	0.49	0.78	0.36
29	1.00	1.00	0.85	0.74	0.04	0.95	0.75	0.84	0.61	0.97	0.52	0.64	0.59	0.72	0.13
30	1.00	1.00	0.86	0.80	0.05	0.96	0.73	0.88	0.12	0.93	0.88	0.27	0.57	0.66	0.07
31	1.00	1.00	0.74	0.58	0.01	0.97	0.69	0.95	0.97	0.12	0.66	0.24	0.40	0.61	0.18
32	0.94	0.93	0.65	0.52	0.11	0.88	0.47	0.88	0.55	0.48	0.90	0.13	0.31	0.57	0.10
33	1.00	1.00	0.93	0.93	0.24	0.81	0.69	0.85	0.76	0.49	0.86	0.53	0.65	0.90	0.25
34	1.00	0.37	0.34	1.00	0.35	0.36	0.20	0.46	0.96	0.96	0.83	0.81	1.00	0.21	0.82
35	1.00	1.00	0.96	0.91	0.18	0.47	0.81	0.95	0.80	0.18	0.99	0.41	0.65	0.96	0.16
36	0.99	1.00	0.73	0.61	0.04	0.83	0.46	0.95	0.01	0.00	0.97	0.80	0.44	0.67	0.06
37	0.96	0.97	0.93	0.89	0.07	0.90	0.80	0.75	0.43	0.65	0.81	0.31	0.26	0.72	0.00
38	0.90	0.90	0.85	0.72	0.10	0.58	0.66	0.92	0.60	0.30	0.92	0.62	0.57	0.87	0.31
39	1.00	1.00	0.93	0.84	0.14	0.79	0.65	0.94	0.80	0.89	0.91	0.61	0.33	0.86	0.15
40	0.85	0.81	0.79	0.93	0.06	0.77	0.69	0.43	0.47	0.77	0.75	0.18	0.62	0.31	0.58
R_j	1.74	**1.44**	5.46	7.10	14.41	6.05	9.18	5.33	10.18	9.46	7.33	10.46	11.38	7.64	12.82

(b) Metric 2: Spearman Rank Correlation (ρ_s)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.99	0.99	0.78	0.95	0.94	0.95	0.88	0.31	0.34	0.51	0.52	0.88	0.54	0.57	0.65
2	1.00	1.00	0.88	0.79	0.82	0.92	0.95	0.90	0.64	0.27	0.81	0.59	0.75	0.76	0.46
3	1.00	1.00	0.90	0.94	0.91	0.95	0.90	0.87	0.53	0.11	0.79	0.28	0.52	0.81	0.37
4	0.97	0.94	0.91	0.72	0.82	0.89	0.85	0.42	0.29	0.03	0.62	0.29	0.60	0.48	0.53
5	0.61	0.66	0.68	0.93	0.67	0.69	0.50	0.90	0.67	0.02	0.75	0.88	0.75	0.51	0.23
6	0.96	0.95	0.86	0.90	0.90	0.91	0.87	0.54	0.34	0.01	0.66	0.66	0.58	0.45	0.37
7	1.00	1.00	0.93	0.87	0.80	0.94	0.93	0.29	0.22	0.21	0.96	0.82	0.17	0.82	0.61
8	1.00	1.00	0.97	0.83	0.75	0.92	0.88	0.31	0.54	0.29	0.97	0.51	0.61	0.80	0.31
9	1.00	1.00	0.80	0.95	0.92	0.96	0.95	0.95	0.76	0.39	0.80	0.64	0.75	0.76	0.70
10	1.00	1.00	0.98	0.88	0.52	0.92	0.92	0.81	0.30	0.19	0.72	0.09	0.48	0.77	0.48
11	0.93	0.92	0.93	0.96	0.51	0.57	0.42	0.94	0.48	0.17	0.83	0.59	0.55	0.34	0.40
12	0.99	0.99	0.44	0.89	0.86	0.86	0.76	0.76	0.41	0.14	0.96	0.79	0.86	0.83	0.46

13	1.00	1.00	0.95	0.89	0.73	0.94	0.91	0.64	0.58	0.24	0.97	0.97	0.48	0.80	0.55
14	1.00	1.00	0.76	0.88	0.91	0.94	0.97	0.95	0.74	0.20	0.81	0.67	0.85	0.77	0.65
15	1.00	1.00	0.50	0.83	0.87	0.87	0.72	0.99	0.51	0.21	0.88	0.82	0.91	0.85	0.50
16	1.00	1.00	0.72	0.96	0.77	0.86	0.82	0.61	0.31	0.01	0.39	0.38	0.09	0.43	0.43
17	1.00	1.00	0.99	0.92	0.71	0.87	0.86	0.93	0.35	0.18	0.98	0.81	0.63	0.77	0.51
18	1.00	1.00	0.94	0.96	0.82	0.91	0.61	0.99	0.45	0.14	0.91	0.94	0.56	0.85	0.67
19	0.99	0.98	0.76	0.82	0.96	0.93	0.74	0.94	0.45	0.05	0.78	0.69	0.68	0.74	0.50
20	1.00	0.99	0.75	0.87	0.92	0.88	0.77	0.24	0.56	0.09	0.90	0.53	0.89	0.79	0.52
21	0.98	0.99	0.94	0.77	0.69	0.88	0.56	0.86	0.40	0.14	0.82	0.43	0.74	0.72	0.45
22	0.87	0.84	0.85	0.57	0.52	0.80	0.76	0.75	0.47	0.17	0.07	0.18	0.44	0.79	0.45
23	0.97	0.98	0.93	0.92	0.57	0.74	0.70	0.84	0.53	0.04	0.64	0.65	0.39	0.69	0.39
24	0.98	0.97	0.69	0.91	0.90	0.94	0.67	0.71	0.54	0.04	0.77	0.68	0.73	0.57	0.46
25	1.00	1.00	0.99	0.91	0.67	0.88	0.87	0.94	0.52	0.21	0.80	0.37	0.48	0.73	0.38
26	1.00	0.99	0.79	0.81	0.91	0.96	0.95	0.93	0.62	0.22	0.73	0.43	0.68	0.71	0.68
27	1.00	1.00	0.98	0.96	0.73	0.87	0.78	0.43	0.25	0.16	0.60	0.33	0.81	0.76	0.50
28	1.00	1.00	0.93	0.87	0.77	0.94	0.86	0.90	0.46	0.13	0.94	0.54	0.49	0.71	0.55
29	1.00	1.00	0.98	0.86	0.74	0.86	0.79	0.56	0.58	0.14	0.98	0.78	0.66	0.80	0.61
30	1.00	1.00	0.98	0.93	0.70	0.86	0.84	0.92	0.27	0.08	0.96	0.70	0.24	0.80	0.64
31	1.00	1.00	0.99	0.98	0.75	0.73	0.59	0.81	0.24	0.07	0.53	0.02	0.99	0.72	0.44
32	0.91	0.91	0.92	0.86	0.66	0.76	0.60	0.87	0.18	0.13	0.64	0.45	0.51	0.60	0.37
33	1.00	1.00	0.86	0.85	0.90	0.94	0.97	0.91	0.51	0.25	0.65	0.47	0.76	0.71	0.72
34	0.86	0.78	0.78	0.44	0.49	0.80	0.96	0.79	0.15	0.04	0.78	0.68	0.45	0.72	0.64
35	1.00	1.00	0.94	0.95	0.95	0.96	0.95	0.99	0.43	0.32	0.90	0.70	0.84	0.86	0.66
36	0.97	0.98	0.81	0.91	0.73	0.81	0.77	0.92	0.61	0.02	0.43	0.65	0.07	0.60	0.46
37	0.96	0.97	0.92	0.78	0.72	0.93	0.91	0.86	0.34	0.11	0.88	0.40	0.48	0.79	0.30
38	0.89	0.88	0.89	0.94	0.86	0.84	0.74	0.91	0.65	0.03	0.75	0.70	0.61	0.68	0.58
39	1.00	1.00	0.85	0.94	0.85	0.93	0.90	0.94	0.57	0.25	0.92	0.66	0.79	0.65	0.44
40	0.75	0.75	0.95	0.78	0.69	1.00	1.00	0.87	0.74	0.75	0.34	0.59	0.64	0.90	1.00
R_j	2.08	**1.87**	5.90	5.67	7.74	5.00	7.28	6.77	12.51	14.85	7.33	10.85	10.38	9.49	12.28

(c) Metric 3: Mean K-ary neighborhood agreement ($\mu_{R_{nX}}$)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.92	0.87	0.24	0.63	0.42	0.68	0.70	0.50	0.26	0.29	0.33	0.62	0.57	0.31	0.28
2	1.00	0.98	0.71	0.77	0.27	0.67	0.73	0.65	0.46	0.62	0.56	0.52	0.45	0.37	0.18
3	0.99	0.97	0.67	0.67	0.27	0.73	0.80	0.67	0.41	0.57	0.59	0.76	0.23	0.15	0.10
4	0.73	0.73	0.23	0.64	0.34	0.49	0.66	0.57	0.27	0.40	0.29	0.43	0.21	0.40	0.02
5	0.40	0.39	0.51	0.32	0.16	0.47	0.48	0.49	0.38	0.10	0.34	0.56	0.06	0.38	0.02
6	0.85	0.78	0.35	0.71	0.24	0.66	0.69	0.63	0.27	0.50	0.29	0.55	0.51	0.45	0.02
7	0.99	0.97	0.25	0.71	0.42	0.63	0.79	0.69	0.20	0.81	0.60	0.66	0.59	0.18	0.15
8	1.00	0.95	0.26	0.65	0.21	0.55	0.76	0.66	0.45	0.76	0.56	0.51	0.36	0.39	0.16
9	1.00	0.99	0.79	0.73	0.46	0.76	0.88	0.61	0.54	0.59	0.58	0.78	0.55	0.60	0.24
10	0.96	0.90	0.57	0.67	0.33	0.31	0.76	0.65	0.33	0.49	0.48	0.55	0.17	0.36	0.14
11	0.68	0.69	0.61	0.33	0.25	0.38	0.42	0.66	0.34	0.16	0.32	0.61	0.16	0.39	0.05
12	0.88	0.88	0.57	0.52	0.32	0.64	0.66	0.21	0.27	0.71	0.62	0.65	0.26	0.51	0.12
13	0.99	0.95	0.43	0.68	0.35	0.55	0.79	0.66	0.45	0.79	0.53	0.53	0.76	0.41	0.17
14	0.98	0.98	0.78	0.83	0.42	0.73	0.85	0.59	0.55	0.67	0.59	0.67	0.05	0.61	0.14
15	0.99	0.99	0.90	0.60	0.30	0.80	0.82	0.30	0.39	0.02	0.65	0.63	0.00	0.69	0.16
16	0.96	0.93	0.45	0.71	0.33	0.66	0.80	0.53	0.23	0.25	0.23	0.74	0.25	0.09	0.06
17	0.99	0.92	0.62	0.63	0.33	0.46	0.73	0.69	0.31	0.78	0.57	0.53	0.50	0.42	0.15
18	0.98	0.94	0.84	0.44	0.44	0.66	0.76	0.71	0.37	0.05	0.61	0.66	0.07	0.43	0.11
19	0.95	0.95	0.79	0.54	0.33	0.76	0.81	0.47	0.38	0.50	0.61	0.55	0.32	0.39	0.03
20	0.93	0.94	0.29	0.55	0.32	0.76	0.74	0.54	0.37	0.30	0.60	0.61	0.26	0.61	0.01
21	0.89	0.84	0.57	0.39	0.31	0.46	0.70	0.62	0.33	0.55	0.45	0.43	0.07	0.50	0.11
22	0.64	0.62	0.51	0.54	0.29	0.35	0.54	0.52	0.33	0.01	0.50	0.27	0.00	0.26	0.06
23	0.87	0.83	0.52	0.55	0.25	0.42	0.58	0.71	0.43	0.06	0.52	0.55	0.03	0.24	0.02
24	0.79	0.81	0.41	0.50	0.30	0.64	0.72	0.44	0.30	0.51	0.28	0.63	0.39	0.43	0.01
25	0.98	0.93	0.63	0.65	0.23	0.42	0.71	0.70	0.45	0.57	0.48	0.51	0.01	0.28	0.13
26	0.91	0.91	0.65	0.73	0.46	0.73	0.82	0.57	0.48	0.01	0.50	0.61	0.01	0.47	0.15
27	0.96	0.89	0.31	0.57	0.31	0.42	0.70	0.69	0.27	0.41	0.54	0.57	0.06	0.50	0.11
28	0.99	0.97	0.63	0.63	0.37	0.64	0.79	0.69	0.38	0.74	0.52	0.62	0.42	0.28	0.11
29	0.98	0.95	0.31	0.55	0.41	0.54	0.65	0.70	0.49	0.83	0.53	0.51	0.52	0.38	0.12
30	0.97	0.91	0.68	0.57	0.40	0.51	0.72	0.68	0.29	0.71	0.55	0.56	0.43	0.31	0.06
31	0.96	0.85	0.55	0.48	0.24	0.42	0.57	0.65	0.30	0.39	0.47	0.61	0.06	0.66	0.07
32	0.75	0.72	0.67	0.49	0.24	0.51	0.58	0.69	0.26	0.43	0.45	0.56	0.35	0.38	0.05
33	0.76	0.74	0.58	0.72	0.38	0.64	0.65	0.49	0.38	0.39	0.45	0.51	0.31	0.43	0.16
34	0.64	0.60	0.50	0.82	0.44	0.30	0.56	0.46	0.21	0.45	0.43	0.20	0.38	0.26	0.05
35	0.95	0.94	0.90	0.75	0.44	0.86	0.89	0.80	0.28	0.73	0.66	0.81	0.49	0.69	0.20

36	0.88	0.85	0.70	0.56	0.34	0.61	0.72	0.52	0.48	0.00	0.47	0.60	0.02	0.20	0.04
37	0.86	0.86	0.57	0.74	0.20	0.52	0.75	0.66	0.34	0.59	0.53	0.48	0.03	0.34	0.11
38	0.76	0.76	0.77	0.58	0.42	0.73	0.72	0.71	0.48	0.11	0.50	0.74	0.13	0.42	0.04
39	0.99	0.99	0.79	0.69	0.33	0.69	0.82	0.60	0.50	0.74	0.53	0.75	0.55	0.60	0.18
40	0.74	0.70	1.00	1.00	0.99	0.32	0.69	0.68	0.46	0.52	0.65	0.58	0.51	0.38	0.71
R_j	**2.13**	3.41	4.92	6.74	12.49	5.87	3.33	6.64	8.95	9.41	8.46	11.54	11.64	9.92	14.54

(d) Metric 4: Local quality criteria (Q_{local})

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.71	0.72	0.43	0.50	0.18	0.67	0.60	0.37	0.33	0.33	0.55	0.31	0.36	0.48	0.32
2	0.99	0.77	0.90	0.89	0.13	0.59	0.95	0.40	0.36	0.65	0.55	0.72	0.72	0.57	0.30
3	0.96	0.75	0.80	0.82	0.09	0.68	0.87	0.60	0.21	0.75	0.59	0.86	0.76	0.54	0.31
4	0.54	0.66	0.43	0.42	0.02	0.48	0.50	0.28	0.42	0.17	0.52	0.34	0.37	0.50	0.22
5	0.19	0.60	0.30	0.28	0.11	0.48	0.17	0.24	0.23	0.11	0.38	0.12	0.35	0.39	0.12
6	0.90	0.73	0.70	0.53	0.03	0.73	0.70	0.39	0.38	0.56	0.64	0.60	0.39	0.50	0.25
7	0.96	0.75	0.84	0.78	0.12	0.69	0.88	0.46	0.39	0.78	0.60	0.87	0.71	0.44	0.32
8	0.98	0.73	0.78	0.65	0.08	0.66	0.84	0.31	0.37	0.67	0.67	0.79	0.66	0.61	0.31
9	0.99	0.76	0.96	0.93	0.17	0.71	0.96	0.56	0.82	0.93	0.64	0.89	0.62	0.50	0.33
10	0.90	0.68	0.75	0.51	0.10	0.32	0.75	0.33	0.51	0.54	0.62	0.58	0.56	0.42	0.34
11	0.41	0.52	0.45	0.39	0.05	0.38	0.40	0.31	0.30	0.04	0.43	0.04	0.39	0.38	0.19
12	0.74	0.79	0.56	0.25	0.17	0.67	0.63	0.38	0.36	0.10	0.50	0.48	0.53	0.58	0.21
13	0.95	0.73	0.81	0.67	0.12	0.70	0.87	0.32	0.70	0.74	0.61	0.82	0.62	0.64	0.30
14	0.81	0.73	0.67	0.81	0.10	0.71	0.80	0.48	0.78	0.03	0.67	0.69	0.57	0.55	0.35
15	0.94	0.80	0.65	0.30	0.11	0.71	0.91	0.35	0.47	0.01	0.56	0.03	0.56	0.53	0.27
16	0.82	0.78	0.60	0.68	0.05	0.72	0.73	0.46	0.33	0.21	0.58	0.19	0.16	0.55	0.31
17	0.93	0.74	0.81	0.53	0.10	0.54	0.74	0.29	0.57	0.54	0.59	0.71	0.60	0.49	0.33
18	0.90	0.69	0.71	0.62	0.10	0.69	0.81	0.43	0.55	0.01	0.62	0.01	0.55	0.54	0.42
19	0.89	0.79	0.73	0.51	0.05	0.70	0.84	0.29	0.43	0.18	0.55	0.34	0.61	0.61	0.27
20	0.68	0.72	0.54	0.65	0.01	0.75	0.67	0.43	0.37	0.21	0.50	0.09	0.59	0.46	0.30
21	0.88	0.70	0.64	0.50	0.09	0.58	0.69	0.26	0.43	0.10	0.59	0.43	0.54	0.56	0.33
22	0.42	0.59	0.39	0.43	0.02	0.59	0.43	0.15	0.33	0.01	0.53	0.01	0.38	0.50	0.27
23	0.62	0.59	0.41	0.47	0.02	0.49	0.59	0.28	0.21	0.01	0.54	0.00	0.41	0.45	0.21
24	0.46	0.66	0.44	0.48	0.13	0.69	0.50	0.42	0.35	0.26	0.54	0.43	0.29	0.41	0.32
25	0.87	0.65	0.80	0.79	0.11	0.50	0.77	0.27	0.34	0.01	0.62	0.44	0.59	0.51	0.30
26	0.63	0.68	0.61	0.63	0.11	0.69	0.66	0.44	0.56	0.01	0.49	0.00	0.51	0.48	0.27
27	0.85	0.67	0.61	0.53	0.10	0.36	0.68	0.28	0.46	0.08	0.65	0.32	0.49	0.55	0.40
28	0.96	0.76	0.85	0.74	0.10	0.71	0.87	0.46	0.34	0.84	0.60	0.85	0.67	0.56	0.39
29	0.89	0.74	0.68	0.70	0.12	0.64	0.81	0.30	0.42	0.52	0.59	0.80	0.64	0.62	0.32
30	0.83	0.68	0.75	0.68	0.04	0.53	0.80	0.31	0.66	0.33	0.61	0.77	0.56	0.53	0.33
31	0.87	0.55	0.65	0.47	0.09	0.36	0.69	0.30	0.45	0.26	0.62	0.37	0.42	0.56	0.33
32	0.59	0.69	0.62	0.58	0.01	0.65	0.58	0.31	0.49	0.39	0.57	0.29	0.55	0.50	0.19
33	0.68	0.64	0.54	0.57	0.11	0.58	0.60	0.33	0.53	0.51	0.63	0.55	0.50	0.33	0.34
34	0.62	0.70	0.41	0.28	0.05	0.60	0.35	0.08	0.19	0.25	0.59	0.32	0.33	0.47	0.34
35	0.68	0.68	0.60	0.61	0.15	0.68	0.65	0.51	0.57	0.42	0.61	0.55	0.56	0.38	0.32
36	0.58	0.68	0.56	0.41	0.08	0.57	0.58	0.31	0.41	0.00	0.57	0.00	0.51	0.48	0.24
37	0.66	0.74	0.63	0.66	0.09	0.56	0.63	0.31	0.31	0.00	0.60	0.50	0.58	0.53	0.18
38	0.52	0.58	0.54	0.49	0.16	0.61	0.52	0.48	0.46	0.14	0.46	0.13	0.45	0.42	0.26
39	0.97	0.75	0.92	0.81	0.13	0.68	0.94	0.50	0.78	0.64	0.64	0.91	0.68	0.54	0.41
40	0.80	0.58	0.97	0.60	0.70	0.41	0.80	0.80	0.59	0.41	0.97	0.32	0.35	0.41	0.96
R_j	**1.36**	3.69	7.03	6.46	14.59	6.97	2.00	6.95	10.54	12.46	6.36	8.82	9.21	11.56	12.00

(e) Metric 5: k_{max} neighborhood loss ($\lambda_{K_{max}}$)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	1.00	0.84	0.89	0.69	0.81	0.62	0.87	0.86	0.83	0.77	0.79	0.76	0.80	0.75	0.60
2	1.00	1.00	1.00	1.00	0.58	1.00	0.99	0.85	0.87	0.90	0.81	0.73	0.80	0.60	0.56
3	1.00	1.00	1.00	1.00	0.84	0.90	0.97	0.90	0.89	0.88	1.00	0.77	0.78	0.41	0.54
4	0.99	0.77	0.70	0.61	0.64	0.61	0.77	0.97	0.84	0.66	0.57	0.69	0.57	0.25	0.47
5	1.00	0.47	0.79	0.73	0.76	0.52	0.48	1.00	0.56	0.58	0.87	0.59	0.64	0.67	0.39
6	1.00	1.00	0.78	0.85	0.98	0.84	1.00	0.89	0.79	0.78	0.82	0.83	0.96	0.22	0.48
7	1.00	1.00	1.00	0.86	0.74	1.00	1.00	0.88	0.83	0.79	0.89	0.80	0.83	0.52	0.59
8	1.00	1.00	0.87	0.78	0.71	1.00	0.90	0.86	0.79	0.66	0.96	0.85	0.77	0.36	0.59
9	1.00	1.00	1.00	1.00	0.81	1.00	1.00	0.88	0.79	1.00	0.97	0.84	0.83	0.56	0.59
10	0.99	1.00	0.87	0.91	0.78	0.94	0.89	0.80	0.91	0.72	0.80	0.84	0.57	0.39	0.65
11	0.60	0.66	0.97	0.74	0.66	0.65	0.66	0.68	0.58	0.50	0.81	0.62	0.61	0.21	0.53
12	1.00	1.00	0.77	0.51	0.78	0.70	0.80	1.00	0.71	0.72	0.77	0.77	0.80	0.55	0.47
13	1.00	0.98	0.89	0.79	0.71	0.99	0.99	0.84	0.78	0.82	0.83	0.81	0.77	0.46	0.52

14	0.70	0.95	0.84	0.93	0.73	0.82	0.94	0.89	0.75	0.89	0.85	0.85	0.84	0.48	0.68
15	1.00	1.00	1.00	0.48	0.67	0.83	1.00	0.89	0.79	0.67	0.96	0.74	0.84	0.43	0.51
16	1.00	0.89	0.82	0.85	0.75	0.75	0.91	0.88	0.64	0.82	0.87	0.82	0.87	0.26	0.58
17	1.00	1.00	0.90	0.70	0.76	0.90	0.92	1.00	0.74	0.89	0.77	0.79	0.74	0.42	0.58
18	1.00	0.95	0.88	0.77	0.77	0.81	0.89	0.82	0.74	0.69	0.91	0.81	0.82	0.42	0.75
19	1.00	0.94	0.53	0.66	0.53	0.87	0.90	0.86	0.79	0.67	0.57	0.76	0.83	0.24	0.55
20	1.00	0.89	0.83	0.84	0.70	0.75	0.86	0.83	0.85	0.77	0.81	0.71	0.88	0.28	0.56
21	1.00	1.00	0.68	0.70	0.54	0.77	1.00	0.83	0.74	0.67	0.99	0.82	0.95	0.44	0.56
22	0.72	0.65	0.95	0.90	0.44	0.60	0.71	1.00	0.59	0.48	0.96	0.84	0.76	1.00	0.56
23	0.65	0.84	0.82	0.73	0.63	0.62	0.81	0.76	0.60	0.40	0.83	0.76	0.65	0.51	0.45
24	1.00	0.68	0.85	1.00	0.88	0.66	0.77	0.80	0.60	0.77	0.86	0.77	0.82	0.52	0.59
25	0.99	0.91	0.62	0.96	0.67	0.87	0.91	0.78	0.72	0.63	0.55	0.82	0.63	0.50	0.62
26	0.70	0.87	0.80	0.79	0.70	0.75	0.86	0.79	0.70	0.71	0.99	0.74	0.81	0.53	0.63
27	0.93	0.95	0.65	0.49	0.68	0.71	0.79	0.79	0.68	0.50	0.67	0.87	0.45	0.44	0.68
28	1.00	1.00	1.00	0.88	0.72	0.90	0.93	0.88	0.79	0.84	1.00	0.80	0.82	0.36	0.66
29	1.00	0.93	0.87	0.83	0.67	1.00	1.00	0.85	0.76	0.73	0.80	0.80	0.77	0.49	0.59
30	1.00	0.88	0.86	0.78	0.68	0.88	1.00	0.80	0.73	0.92	0.81	0.83	0.67	0.28	0.56
31	0.91	1.00	0.59	1.00	0.70	0.86	0.79	0.66	0.62	0.55	0.68	0.80	0.50	0.33	0.60
32	1.00	0.74	0.48	0.71	0.61	0.76	0.77	0.81	0.72	0.65	0.56	0.75	0.77	1.00	0.47
33	1.00	0.80	0.67	0.67	0.53	0.75	0.70	0.78	0.65	0.71	0.72	0.85	0.76	0.46	0.69
34	1.00	1.00	0.53	0.51	0.40	0.57	0.64	1.00	0.57	0.47	0.48	0.78	1.00	0.39	0.68
35	0.59	0.90	0.80	0.87	0.79	0.86	0.91	0.90	0.82	0.81	0.64	0.83	0.89	0.60	0.61
36	1.00	0.73	0.36	0.59	0.65	0.76	0.76	0.80	0.73	0.62	0.54	0.76	0.71	0.32	0.54
37	1.00	0.81	0.69	0.79	0.64	0.80	0.84	1.00	0.78	0.73	0.69	0.78	0.69	0.45	0.46
38	0.71	0.76	0.84	0.78	0.79	0.81	0.76	0.79	0.71	0.69	0.81	0.72	0.80	0.40	0.66
39	1.00	1.00	1.00	1.00	0.79	1.00	1.00	0.85	0.81	1.00	0.83	0.82	0.80	0.52	0.69
40	0.52	0.69	0.73	0.61	0.59	0.99	0.88	0.87	0.74	0.82	0.73	0.99	0.81	0.87	0.99
R_j	**3.60**	3.88	6.90	7.53	10.62	6.92	4.49	5.36	9.79	10.03	7.32	8.08	8.44	13.85	13.21

(f) Metric 6: Global quality criteria (Q_{global})

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.35	0.26	0.14	0.42	0.17	0.12	0.24	0.34	0.12	0.19	0.36	0.24	0.33	0.15	0.14
2	0.37	0.25	0.39	0.48	0.29	0.34	0.38	0.36	0.07	0.16	0.47	0.26	0.31	0.18	0.22
3	0.35	0.33	0.33	0.48	0.28	0.31	0.37	0.37	0.06	0.13	0.44	0.15	0.34	0.07	0.22
4	0.33	0.19	0.14	0.34	0.12	0.18	0.26	0.37	0.01	0.17	0.31	0.10	0.26	0.18	0.14
5	0.19	0.18	0.19	0.14	0.18	0.03	0.17	0.27	0.00	0.07	0.14	0.02	0.23	0.18	0.15
6	0.35	0.15	0.22	0.40	0.17	0.24	0.24	0.34	0.01	0.13	0.38	0.23	0.33	0.26	0.15
7	0.35	0.28	0.19	0.48	0.30	0.42	0.35	0.37	0.07	0.21	0.47	0.31	0.32	0.11	0.10
8	0.34	0.20	0.20	0.48	0.29	0.38	0.32	0.34	0.08	0.11	0.45	0.24	0.29	0.21	0.23
9	0.36	0.32	0.42	0.48	0.28	0.35	0.38	0.39	0.11	0.22	0.48	0.30	0.37	0.34	0.26
10	0.35	0.20	0.30	0.47	0.27	0.26	0.30	0.37	0.07	0.17	0.42	0.14	0.17	0.19	0.18
11	0.20	0.23	0.26	0.26	0.18	0.02	0.24	0.24	0.02	0.09	0.26	0.05	0.17	0.17	0.14
12	0.27	0.25	0.30	0.40	0.29	0.28	0.10	0.37	0.05	0.16	0.38	0.09	0.34	0.23	0.12
13	0.34	0.21	0.27	0.47	0.28	0.39	0.32	0.37	0.08	0.18	0.46	0.36	0.28	0.24	0.25
14	0.40	0.28	0.38	0.44	0.28	0.34	0.34	0.39	0.07	0.21	0.45	0.02	0.36	0.33	0.28
15	0.31	0.25	0.38	0.48	0.29	0.00	0.13	0.40	0.07	0.14	0.47	0.00	0.39	0.30	0.20
16	0.33	0.31	0.23	0.43	0.13	0.13	0.23	0.40	0.04	0.17	0.39	0.14	0.33	0.07	0.10
17	0.33	0.19	0.31	0.48	0.29	0.38	0.29	0.37	0.07	0.17	0.43	0.24	0.23	0.23	0.18
18	0.24	0.26	0.41	0.46	0.30	0.02	0.33	0.37	0.06	0.21	0.44	0.03	0.31	0.23	0.21
19	0.28	0.24	0.38	0.44	0.32	0.21	0.23	0.40	0.02	0.14	0.44	0.12	0.37	0.18	0.19
20	0.28	0.27	0.19	0.40	0.29	0.12	0.27	0.37	0.00	0.16	0.43	0.13	0.34	0.27	0.19
21	0.27	0.18	0.28	0.42	0.23	0.24	0.27	0.32	0.06	0.16	0.40	0.03	0.26	0.23	0.18
22	0.23	0.11	0.23	0.29	0.24	0.00	0.25	0.30	0.02	0.12	0.25	0.00	0.21	0.13	0.18
23	0.27	0.21	0.26	0.36	0.25	0.02	0.29	0.31	0.02	0.13	0.36	0.01	0.23	0.09	0.21
24	0.25	0.25	0.18	0.33	0.13	0.19	0.24	0.36	0.01	0.15	0.34	0.18	0.33	0.17	0.13
25	0.34	0.19	0.31	0.47	0.26	0.26	0.33	0.34	0.06	0.11	0.43	0.01	0.22	0.15	0.25
26	0.34	0.25	0.33	0.38	0.25	0.00	0.32	0.39	0.06	0.22	0.39	0.00	0.34	0.26	0.23
27	0.27	0.21	0.17	0.45	0.26	0.17	0.31	0.35	0.05	0.16	0.41	0.01	0.20	0.22	0.17
28	0.33	0.26	0.34	0.48	0.27	0.41	0.36	0.37	0.06	0.18	0.46	0.28	0.33	0.15	0.20
29	0.30	0.20	0.20	0.47	0.27	0.40	0.33	0.32	0.05	0.19	0.46	0.25	0.27	0.18	0.24
30	0.30	0.22	0.34	0.46	0.25	0.36	0.32	0.35	0.03	0.20	0.44	0.22	0.26	0.19	0.17
31	0.24	0.23	0.31	0.46	0.19	0.20	0.27	0.28	0.05	0.12	0.38	0.06	0.19	0.27	0.19
32	0.24	0.22	0.30	0.32	0.25	0.18	0.32	0.33	0.02	0.11	0.33	0.18	0.28	0.18	0.16
33	0.35	0.23	0.29	0.35	0.22	0.23	0.27	0.31	0.07	0.17	0.34	0.18	0.29	0.23	0.20
34	0.39	0.07	0.24	0.30	0.19	0.19	0.19	0.28	0.03	0.18	0.23	0.16	0.17	0.11	0.13
35	0.36	0.35	0.40	0.42	0.33	0.35	0.39	0.39	0.09	0.21	0.41	0.24	0.38	0.32	0.13
36	0.27	0.23	0.29	0.39	0.24	0.00	0.23	0.34	0.03	0.15	0.36	0.01	0.29	0.14	0.22

37	0.35	0.19	0.28	0.38	0.28	0.27	0.33	0.38	0.04	0.10	0.39	0.01	0.27	0.14	0.18
38	0.27	0.31	0.33	0.34	0.25	0.04	0.32	0.33	0.02	0.19	0.34	0.05	0.34	0.21	0.21
39	0.34	0.31	0.41	0.48	0.26	0.41	0.35	0.39	0.08	0.18	0.48	0.31	0.34	0.33	0.26
40	0.36	0.49	0.36	0.49	0.18	0.20	0.33	0.25	0.31	0.49	0.48	0.21	0.28	0.33	0.17
$\mathbf{R_j}$	6.87	7.26	7.15	**1.26**	9.22	7.29	5.86	3.82	14.74	12.69	2.10	10.01	8.18	10.95	12.59

(g) Metric 7: Area under the R_{nX} curve ($AUC_{\ln K}(R_{nX}(K))$)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.36	0.80	0.75	0.27	0.15	0.59	0.66	0.22	0.34	0.30	0.51	0.70	0.30	0.37	0.38
2	0.84	1.00	0.79	0.34	0.10	0.63	0.98	0.71	0.60	0.27	0.83	0.60	0.67	0.49	0.44
3	0.79	0.99	0.77	0.16	0.08	0.64	0.94	0.75	0.51	0.26	0.79	0.72	0.72	0.48	0.64
4	0.36	0.62	0.72	0.40	0.02	0.60	0.55	0.32	0.15	0.24	0.47	0.50	0.30	0.39	0.28
5	0.34	0.24	0.57	0.26	0.01	0.38	0.21	0.02	0.01	0.10	0.27	0.49	0.35	0.35	0.29
6	0.60	0.85	0.73	0.41	0.02	0.64	0.73	0.61	0.56	0.24	0.58	0.68	0.37	0.40	0.35
7	0.62	0.99	0.77	0.30	0.11	0.64	0.94	0.88	0.73	0.32	0.77	0.68	0.69	0.31	0.50
8	0.61	1.00	0.75	0.40	0.09	0.66	0.90	0.81	0.57	0.22	0.67	0.64	0.65	0.52	0.33
9	0.91	1.00	0.82	0.75	0.15	0.66	0.99	0.80	0.79	0.34	0.85	0.75	0.62	0.50	0.60
10	0.70	0.95	0.73	0.46	0.09	0.63	0.83	0.54	0.39	0.29	0.56	0.33	0.57	0.37	0.35
11	0.49	0.47	0.52	0.32	0.01	0.42	0.46	0.06	0.07	0.16	0.44	0.40	0.37	0.34	0.34
12	0.60	0.74	0.74	0.39	0.06	0.47	0.74	0.56	0.11	0.22	0.23	0.70	0.56	0.38	0.40
13	0.69	0.98	0.77	0.61	0.11	0.63	0.92	0.82	0.76	0.29	0.68	0.66	0.61	0.55	0.34
14	0.72	0.86	0.77	0.74	0.09	0.72	0.86	0.70	0.03	0.31	0.77	0.73	0.58	0.56	0.52
15	0.74	0.98	0.83	0.52	0.11	0.58	0.96	0.01	0.01	0.24	0.30	0.78	0.56	0.47	0.38
16	0.56	0.89	0.82	0.21	0.05	0.57	0.81	0.18	0.20	0.27	0.62	0.72	0.16	0.37	0.52
17	0.76	0.97	0.76	0.52	0.09	0.61	0.83	0.76	0.54	0.29	0.58	0.52	0.62	0.42	0.30
18	0.79	0.95	0.74	0.52	0.08	0.56	0.88	0.02	0.02	0.36	0.65	0.69	0.56	0.47	0.44
19	0.78	0.93	0.83	0.41	0.04	0.53	0.90	0.38	0.23	0.24	0.50	0.74	0.62	0.49	0.35
20	0.46	0.73	0.75	0.37	0.00	0.51	0.77	0.14	0.18	0.28	0.62	0.77	0.54	0.44	0.44
21	0.65	0.86	0.70	0.47	0.08	0.54	0.75	0.46	0.08	0.28	0.56	0.54	0.51	0.45	0.29
22	0.42	0.49	0.61	0.32	0.01	0.48	0.48	0.01	0.00	0.22	0.46	0.51	0.41	0.43	0.16
23	0.44	0.69	0.61	0.21	0.03	0.53	0.65	0.01	0.02	0.19	0.51	0.49	0.45	0.44	0.31
24	0.43	0.55	0.70	0.33	0.01	0.49	0.57	0.40	0.24	0.28	0.47	0.70	0.27	0.35	0.40
25	0.75	0.93	0.69	0.32	0.09	0.65	0.84	0.49	0.00	0.23	0.73	0.47	0.57	0.51	0.28
26	0.64	0.71	0.75	0.56	0.10	0.54	0.74	0.00	0.01	0.27	0.64	0.72	0.51	0.48	0.48
27	0.51	0.91	0.70	0.46	0.09	0.58	0.78	0.39	0.03	0.32	0.58	0.39	0.52	0.44	0.32
28	0.80	0.98	0.80	0.27	0.09	0.62	0.94	0.85	0.72	0.34	0.76	0.72	0.64	0.47	0.49
29	0.55	0.94	0.73	0.35	0.10	0.56	0.90	0.83	0.51	0.31	0.70	0.62	0.62	0.55	0.33
30	0.75	0.91	0.71	0.54	0.04	0.57	0.86	0.76	0.37	0.32	0.67	0.53	0.54	0.43	0.34
31	0.61	0.92	0.58	0.51	0.07	0.54	0.76	0.38	0.20	0.25	0.48	0.36	0.45	0.45	0.36
32	0.63	0.64	0.68	0.45	0.01	0.54	0.64	0.33	0.38	0.17	0.63	0.63	0.53	0.42	0.35
33	0.55	0.72	0.65	0.50	0.11	0.65	0.65	0.51	0.45	0.29	0.56	0.60	0.48	0.38	0.37
34	0.46	0.64	0.61	0.21	0.04	0.66	0.40	0.35	0.27	0.32	0.33	0.44	0.33	0.37	0.10
35	0.65	0.75	0.72	0.59	0.13	0.64	0.70	0.59	0.43	0.31	0.65	0.73	0.56	0.33	0.57
36	0.59	0.70	0.71	0.35	0.05	0.54	0.67	0.00	0.01	0.23	0.45	0.60	0.48	0.49	0.34
37	0.61	0.75	0.77	0.28	0.09	0.62	0.70	0.52	0.01	0.16	0.67	0.57	0.55	0.44	0.35
38	0.57	0.56	0.61	0.43	0.04	0.47	0.56	0.04	0.05	0.25	0.49	0.63	0.44	0.42	0.51
39	0.88	0.99	0.79	0.73	0.11	0.65	0.98	0.89	0.64	0.32	0.77	0.70	0.63	0.52	0.53
40	0.99	0.64	0.44	0.69	0.79	0.78	0.58	0.35	0.98	0.99	0.59	0.36	0.43	0.38	0.78
$\mathbf{R_j}$	5.59	**1.79**	3.03	10.03	14.62	6.41	2.79	9.15	11.97	12.82	6.10	6.05	8.49	10.05	11.10

(h) Metric 8: KNN prediction accuracy (ACC_{ψ})

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.64	0.68	0.52	0.68	0.68	0.48	0.56	0.64	0.48	0.60	0.68	0.64	0.64	0.64	0.60
2	0.48	0.44	0.48	0.60	0.64	0.48	0.36	0.32	0.52	0.40	0.32	0.40	0.64	0.60	0.40
3	0.44	0.44	0.56	0.64	0.60	0.40	0.56	0.64	0.52	0.28	0.40	0.56	0.60	0.56	0.36
4	0.76	0.68	0.76	0.80	0.80	0.80	0.80	0.76	0.80	0.68	0.84	0.80	0.76	0.76	0.84
5	0.67	0.56	0.45	0.65	0.66	0.59	0.67	0.51	0.60	0.57	0.56	0.56	0.58	0.53	0.56
6	0.60	0.56	0.56	0.52	0.48	0.56	0.36	0.52	0.56	0.48	0.68	0.60	0.68	0.52	0.56
7	0.52	0.64	0.60	0.60	0.72	0.56	0.48	0.60	0.52	0.48	0.60	0.56	0.60	0.60	0.60
8	0.12	0.24	0.28	0.16	0.12	0.28	0.16	0.28	0.16	0.20	0.20	0.20	0.28	0.24	0.32
9	0.52	0.48	0.52	0.52	0.32	0.48	0.48	0.64	0.44	0.52	0.48	0.48	0.52	0.40	0.52
10	0.20	0.20	0.12	0.16	0.12	0.12	0.16	0.16	0.16	0.16	0.12	0.28	0.12	0.16	0.20
11	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	0.96	1.00	1.00	1.00
12	0.72	0.68	0.68	0.68	0.68	0.76	0.68	0.64	0.72	0.68	0.64	0.72	0.64	0.76	0.72
13	0.60	0.68	0.72	0.68	0.72	0.60	0.60	0.64	0.72	0.68	0.60	0.72	0.72	0.60	0.72
14	0.80	0.64	0.68	0.80	0.28	0.72	0.84	0.76	0.76	0.84	0.92	0.68	0.80	0.88	0.48

15	0.80	0.80	0.80	0.80	0.68	0.80	0.80	0.80	0.80	0.80	0.72	0.80	0.80	0.76	0.76
16	0.08	0.04	0.04	0.12	0.08	0.16	0.08	0.16	0.04	0.12	0.08	0.04	0.08	0.00	0.08
17	0.56	0.60	0.60	0.56	0.56	0.60	0.52	0.64	0.56	0.56	0.48	0.56	0.52	0.64	0.48
18	0.88	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
19	0.60	0.48	0.44	0.36	0.28	0.52	0.56	0.48	0.56	0.64	0.56	0.40	0.48	0.56	0.40
20	0.76	0.72	0.76	0.76	0.68	0.72	0.64	0.72	0.72	0.76	0.80	0.72	0.68	0.64	0.72
21	0.20	0.36	0.24	0.24	0.16	0.24	0.28	0.24	0.32	0.24	0.28	0.20	0.36	0.20	0.16
22	0.88	1.00	0.96	1.00	1.00	0.96	0.88	0.88	0.88	0.96	0.96	1.00	0.80	0.96	1.00
23	0.56	0.64	0.56	0.60	0.64	0.52	0.52	0.60	0.72	0.44	0.52	0.68	0.48	0.52	0.52
24	0.56	0.84	0.68	0.76	0.64	0.84	0.76	0.64	0.68	0.60	0.80	0.80	0.56	0.44	0.68
25	0.80	0.76	0.72	0.84	0.80	0.80	0.68	0.68	0.76	0.76	0.76	0.72	0.60	0.80	0.76
26	0.80	0.68	0.76	0.88	0.64	0.76	0.88	0.84	0.84	0.76	0.76	0.76	0.80	0.80	0.76
27	0.72	0.80	0.76	0.68	0.76	0.80	0.64	0.80	0.76	0.72	0.68	0.64	0.76	0.76	0.72
28	0.04	0.04	0.16	0.04	0.03	0.04	0.03	0.03	0.04	0.04	0.03	0.04	0.04	0.04	0.08
29	1.00	0.96	1.00	0.76	0.84	0.80	1.00	0.96	0.80	1.00	1.00	0.80	0.72	0.92	0.76
30	0.96	0.96	0.96	0.96	0.92	0.88	0.88	0.92	0.92	0.92	0.88	0.92	0.96	0.92	0.92
31	0.72	0.80	0.52	0.76	0.64	0.64	0.84	0.68	0.64	0.84	0.80	0.72	0.48	0.72	0.68
32	0.68	0.80	0.80	0.80	0.76	0.76	0.72	0.76	0.76	0.76	0.84	0.76	0.76	0.80	0.76
33	0.33	0.33	0.33	0.34	0.33	0.32	0.32	0.33	0.33	0.33	0.33	0.33	0.33	0.34	0.33
34	0.52	0.60	0.56	0.48	0.40	0.64	0.40	0.56	0.60	0.72	0.92	0.36	0.52	0.52	0.60
35	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.84
36	0.36	0.32	0.20	0.32	0.20	0.36	0.40	0.40	0.20	0.24	0.52	0.20	0.32	0.44	0.36
37	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
38	0.20	0.32	0.40	0.24	0.20	0.32	0.24	0.40	0.12	0.28	0.16	0.40	0.20	0.24	0.28
39	0.44	0.44	0.40	0.52	0.40	0.44	0.48	0.64	0.60	0.40	0.52	0.48	0.44	0.48	0.52
40	0.89	0.98	0.86	0.88	0.89	0.85	0.89	0.85	0.85	0.97	0.89	0.81	0.89	0.85	0.89
R_j	7.76	6.88	7.91	6.74	9.18	7.79	9.13	7.35	8.28	8.35	7.69	8.31	8.50	7.91	8.22

(i) Metric 9: Normalized Mutual Information (nMI)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.15	0.21	0.20	0.26	1.55	0.21	0.25	0.20	0.20	0.10	0.20	0.19	0.20	0.20	0.99
2	0.20	0.22	0.22	0.19	0.18	0.22	0.07	0.22	0.22	0.20	0.23	0.21	0.22	0.22	0.19
3	0.18	0.20	0.21	0.18	0.32	0.21	0.16	0.22	0.21	0.17	0.21	0.20	0.22	0.21	0.20
4	0.13	0.18	0.18	0.17	0.82	0.20	0.30	0.16	0.17	0.16	0.18	0.16	0.13	0.18	0.26
5	0.04	0.06	0.17	0.17	0.18	0.17	0.04	0.17	0.17	0.11	0.08	0.20	0.14	0.17	0.17
6	0.15	0.18	0.19	0.15	0.56	0.18	1.25	0.04	0.18	0.05	0.19	0.17	0.19	0.19	0.13
7	0.18	0.21	0.21	0.18	0.40	0.21	0.14	0.20	0.21	0.17	0.21	0.20	0.20	0.21	0.19
8	0.14	0.17	0.19	0.13	0.64	0.18	0.33	0.21	0.19	0.13	0.19	0.17	0.20	0.20	0.20
9	0.21	0.22	0.22	0.20	0.02	0.22	0.19	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
10	0.12	0.16	0.18	0.13	0.76	0.15	1.19	0.20	0.15	0.09	0.19	0.16	0.20	0.18	0.14
11	0.20	0.23	0.22	0.24	0.01	0.22	0.09	0.23	0.23	0.22	0.22	0.22	0.23	0.22	0.15
12	0.17	0.17	0.18	0.17	0.20	0.17	0.03	0.17	0.18	0.07	0.05	0.17	0.15	0.17	0.17
13	0.12	0.18	0.19	0.14	0.64	0.18	0.33	0.20	0.18	0.05	0.19	0.16	0.19	0.19	0.14
14	0.17	0.17	0.17	0.17	0.17	0.17	0.06	0.17	0.17	0.12	0.03	0.17	0.15	0.17	0.17
15	0.22	0.23	0.23	0.22	0.13	0.23	0.16	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.20
16	0.17	0.17	0.17	0.17	0.18	0.17	0.01	0.17	0.17	0.08	0.01	0.17	0.16	0.17	0.17
17	0.12	0.17	0.18	0.12	0.51	0.16	0.09	0.21	0.16	0.10	0.19	0.16	0.21	0.19	0.16
18	0.17	0.21	0.21	0.18	0.29	0.21	0.65	0.22	0.21	0.19	0.21	0.20	0.22	0.21	0.03
19	0.17	0.17	0.17	0.17	0.17	0.16	0.09	0.17	0.17	0.06	0.05	0.17	0.15	0.17	0.18
20	0.17	0.22	0.22	0.16	0.15	0.22	0.04	0.22	0.22	0.22	0.22	0.21	0.23	0.22	0.06
21	0.14	0.20	0.18	0.85	0.51	0.20	0.22	0.20	0.20	0.01	0.19	0.17	0.19	0.19	0.41
22	0.06	0.15	0.18	0.07	0.68	0.17	0.28	0.14	0.14	0.06	0.18	0.16	0.17	0.18	0.26
23	0.14	0.22	0.21	0.21	0.31	0.23	0.09	0.23	0.22	0.22	0.21	0.21	0.22	0.21	0.13
24	0.17	0.17	0.20	0.17	0.23	0.18	0.02	0.17	0.17	0.12	0.04	0.17	0.16	0.17	0.17
25	0.11	0.18	0.18	0.12	0.76	0.18	0.50	0.19	0.18	0.10	0.18	0.17	0.19	0.18	0.51
26	0.17	0.17	0.17	0.17	0.18	0.17	0.01	0.17	0.17	0.08	0.01	0.17	0.16	0.17	0.17
27	0.10	0.18	0.17	0.09	0.91	0.18	0.78	0.19	0.17	0.06	0.17	0.15	0.19	0.17	0.03
28	0.18	0.20	0.21	0.17	0.21	0.20	0.01	0.21	0.20	0.16	0.21	0.20	0.21	0.21	0.21
29	0.14	0.19	0.19	0.15	0.85	0.20	0.61	0.21	0.19	0.13	0.19	0.18	0.21	0.19	0.19
30	0.13	0.19	0.19	0.13	0.60	0.19	0.79	0.21	0.18	0.11	0.19	0.17	0.21	0.19	0.17
31	0.08	0.13	0.17	0.11	0.99	0.17	1.66	0.15	0.14	0.11	0.18	0.15	0.15	0.18	0.48
32	0.17	0.17	0.17	0.17	0.21	0.18	0.00	0.17	0.17	0.08	0.03	0.17	0.16	0.17	0.20
33	0.17	0.17	0.17	0.17	0.17	0.17	0.07	0.17	0.17	0.11	0.04	0.17	0.16	0.17	0.17
34	0.11	0.18	0.19	0.18	1.00	0.17	2.02	0.04	0.18	0.11	0.19	0.18	0.18	0.19	4.32
35	0.20	0.23	0.23	0.23	0.17	0.21	1.41	0.22	0.23	0.22	0.23	0.22	0.22	0.23	1.92
36	0.18	0.23	0.23	0.23	0.19	0.22	0.84	0.23	0.23	0.22	0.23	0.22	0.23	0.23	3.20
37	0.14	0.22	0.21	0.19	0.19	0.21	0.44	0.22	0.22	0.21	0.21	0.20	0.22	0.21	0.49

38	0.17	0.17	0.17	0.17	0.17	0.17	0.03	0.17	0.17	0.12	0.04	0.17	0.14	0.17	0.17
39	0.20	0.22	0.22	0.20	0.03	0.22	0.07	0.22	0.22	0.19	0.22	0.21	0.22	0.22	0.22
40	0.18	0.22	0.22	0.21	0.17	0.22	0.81	0.22	0.22	0.17	0.22	0.21	0.22	0.22	0.95
R_j	3.46	**3.01**	3.49	9.56	13.90	4.04	13.10	8.79	9.42	9.71	12.03	4.49	4.35	5.68	14.97

(j) Metric 10: Structural Similarity Index (*SSI*)

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	8.6E+02	6.0E+00	7.0E+00	2.1E+12	9.9E+11	1.6E+13	7.6E+10	2.0E+12	7.1E+03	6.7E+02	9.7E+10	4.2E+03	1.4E+01	1.0E+00	2.0E+02
2	1.3E+04	2.0E+00	1.3E+01	3.6E+10	8.6E+12	8.1E+13	1.1E+12	3.0E+12	7.9E+03	6.4E+02	7.6E+12	4.6E+03	3.2E+01	2.0E+00	8.3E+01
3	4.2E+03	1.0E+00	4.7E+01	3.8E+11	5.7E+10	4.0E+09	7.0E+11	4.2E+11	8.8E+03	2.4E+02	1.3E+12	3.5E+03	1.6E+01	2.0E+00	1.5E+02
4	2.1E+03	2.0E+00	3.8E+01	1.7E+10	1.4E+10	1.4E+11	2.2E+10	1.8E+10	8.1E+03	1.7E+02	4.3E+09	1.7E+03	1.1E+02	3.0E+00	1.6E+02
5	2.5E+03	2.0E+00	1.4E+01	5.9E+10	1.0E+00	2.8E+04	5.9E+10	4.0E+10	1.1E+03	3.5E+02	8.4E+11	1.0E+00	7.9E+03	4.0E+00	1.8E+02
6	2.2E+03	1.0E+00	4.5E+01	4.8E+10	6.6E+09	1.1E+11	4.5E+09	5.2E+10	8.5E+03	2.0E+02	4.5E+09	5.1E+03	2.3E+02	3.0E+00	1.1E+02
7	5.4E+03	2.0E+00	2.6E+01	7.6E+08	3.6E+08	5.2E+09	5.4E+07	1.7E+08	7.1E+03	4.5E+02	4.6E+08	4.1E+03	2.0E+02	2.0E+00	1.4E+02
8	2.4E+03	2.0E+00	4.9E+01	1.8E+08	2.3E+08	5.6E+08	5.2E+07	1.5E+08	8.3E+03	4.2E+02	1.1E+08	2.2E+03	2.9E+02	1.2E+01	1.3E+02
9	2.7E+03	2.0E+00	3.8E+01	1.7E+08	1.9E+08	4.2E+08	1.7E+07	1.7E+08	8.2E+03	3.1E+02	2.5E+07	2.5E+03	2.1E+02	8.0E+00	1.4E+02
10	4.1E+03	2.0E+00	4.4E+01	1.8E+08	7.6E+07	3.5E+08	1.2E+07	1.0E+08	7.0E+03	5.6E+02	1.5E+08	2.6E+03	2.8E+02	7.0E+00	1.5E+02
11	6.3E+03	2.0E+00	2.6E+01	4.6E+13	7.3E+16	2.8E+17	2.6E+13	6.8E+16	7.3E+03	5.2E+02	6.9E+14	2.0E+03	1.1E+02	1.0E+00	1.1E+02
12	3.6E+03	2.0E+00	3.2E+01	6.5E+04	6.3E+05	2.3E+03	3.5E+06	2.4E+07	2.5E+03	4.1E+02	6.8E+06	7.5E+03	2.5E+02	9.0E+00	1.3E+02
13	4.9E+02	1.0E+00	2.1E+01	8.9E+07	9.9E+07	2.1E+08	3.8E+06	5.0E+07	6.9E+03	2.8E+02	1.1E+08	6.2E+02	6.1E+01	1.0E+00	1.2E+02
14	5.4E+03	1.0E+00	9.0E+00	6.7E+02	3.8E+05	3.8E+03	5.7E+01	5.0E+02	3.1E+03	5.4E+02	1.7E+02	8.1E+03	2.3E+02	2.0E+00	1.3E+02
15	2.2E+03	1.0E+00	2.0E+00	1.7E+09	4.0E+09	1.5E+10	4.4E+05	8.4E+08	7.6E+03	3.2E+02	1.4E+08	3.3E+03	9.6E+01	9.0E+00	1.1E+02
16	9.0E+02	2.0E+00	1.5E+01	2.3E+07	4.5E+05	3.6E+04	3.5E+05	1.8E+07	4.1E+02	1.8E+02	1.9E+07	7.1E+03	6.1E+03	3.0E+00	9.1E+01
17	4.7E+03	2.0E+00	3.4E+01	1.8E+07	8.0E+07	7.4E+08	1.3E+07	3.0E+07	8.3E+03	1.8E+02	2.0E+07	2.9E+03	7.4E+01	3.0E+00	8.7E+01
18	6.2E+02	2.0E+00	1.6E+01	8.9E+05	6.6E+06	4.9E+07	2.8E+05	1.8E+07	7.2E+03	2.9E+02	2.9E+06	7.2E+02	1.1E+02	2.0E+00	1.2E+02
19	1.1E+04	1.0E+00	6.0E+00	1.1E+07	7.9E+05	2.6E+03	1.1E+08	1.1E+09	4.2E+03	1.6E+02	4.5E+08	7.5E+03	2.3E+02	2.0E+00	1.6E+02
20	4.1E+03	2.0E+00	4.0E+00	3.8E+09	7.3E+16	1.4E+12	1.1E+08	4.7E+09	7.7E+03	4.0E+02	3.6E+09	2.6E+04	1.3E+03	2.0E+00	1.8E+02
21	3.5E+03	1.0E+00	1.8E+01	9.1E+05	2.1E+07	1.7E+08	7.1E+05	1.0E+07	8.0E+03	1.4E+02	1.5E+07	2.3E+03	8.5E+01	4.0E+00	1.2E+02
22	4.5E+03	1.0E+00	1.3E+02	5.2E+09	8.7E+09	5.9E+10	6.7E+08	5.8E+09	7.8E+03	1.2E+02	3.3E+09	2.3E+03	1.9E+01	4.0E+00	2.4E+02
23	2.2E+03	1.0E+00	1.5E+01	2.5E+09	3.6E+10	2.4E+11	4.4E+07	2.0E+09	7.6E+03	2.0E+02	1.6E+10	1.3E+03	6.4E+01	2.0E+00	1.5E+02
24	2.1E+04	1.0E+00	6.4E+01	3.2E+09	6.0E+05	2.7E+03	2.0E+09	8.8E+10	3.4E+03	3.8E+02	2.3E+09	7.4E+03	4.7E+02	7.0E+00	8.5E+01
25	4.5E+04	1.0E+00	1.0E+01	2.0E+05	1.1E+07	1.9E+08	6.2E+05	5.1E+06	6.7E+03	3.9E+02	5.2E+06	1.9E+03	2.4E+02	8.0E+00	8.1E+01
26	2.0E+03	2.0E+00	1.0E+01	1.9E+07	2.8E+05	3.4E+03	2.0E+07	1.1E+07	3.4E+03	3.3E+02	4.8E+07	7.1E+03	2.6E+02	4.0E+00	3.0E+02
27	2.7E+03	2.0E+00	2.9E+01	1.0E+06	2.5E+06	9.0E+06	1.6E+05	2.9E+06	7.5E+03	2.0E+02	4.7E+06	1.9E+03	3.2E+02	4.0E+00	1.1E+02
28	4.0E+03	2.0E+00	2.2E+01	3.1E+06	1.5E+07	3.5E+07	3.4E+06	7.7E+06	7.7E+03	3.9E+02	2.6E+06	2.2E+03	2.8E+02	9.0E+00	1.2E+02
29	8.4E+02	2.0E+00	1.5E+01	1.3E+07	9.9E+06	3.3E+07	3.5E+06	8.0E+06	8.3E+03	1.9E+02	7.5E+06	1.1E+03	2.9E+01	1.0E+01	9.5E+01
30	2.6E+03	2.0E+00	2.9E+01	1.4E+07	1.3E+07	2.9E+07	4.7E+05	2.5E+06	7.1E+03	4.5E+02	1.7E+07	5.2E+03	4.2E+02	7.0E+00	1.1E+02
31	3.0E+03	2.0E+00	6.7E+01	6.3E+09	1.8E+11	2.5E+12	1.6E+10	3.5E+10	8.5E+03	4.2E+02	1.6E+10	3.8E+03	9.0E+01	2.0E+00	4.9E+01
32	3.4E+03	2.0E+00	6.3E+01	8.5E+04	7.4E+05	2.9E+03	2.5E+04	1.7E+05	3.1E+03	3.5E+02	8.6E+05	7.1E+03	3.9E+02	1.0E+00	1.4E+02
33	3.8E+03	1.0E+00	5.7E+01	1.2E+02	7.7E+05	4.0E+04	5.5E+02	2.5E+02	2.4E+03	3.9E+02	2.6E+02	7.3E+03	4.7E+03	6.0E+00	1.8E+02
34	1.7E+04	1.0E+00	6.1E+01	5.9E+05	1.8E+05	8.4E+11	8.9E+05	7.6E+06	7.6E+03	8.4E+02	3.2E+07	3.6E+04	1.8E+03	1.0E+01	7.7E+01
35	1.4E+03	2.0E+00	1.5E+01	9.2E+02	9.6E+06	7.7E+07	1.2E+05	8.4E+05	7.1E+03	2.6E+02	1.0E+07	2.7E+03	8.0E+01	3.0E+00	5.7E+01
36	1.2E+04	2.0E+00	2.2E+01	2.1E+09	9.7E+05	2.7E+09	1.4E+07	4.1E+08	8.7E+03	1.9E+02	1.7E+08	4.0E+04	3.7E+03	1.0E+00	6.5E+01
37	1.1E+03	1.0E+00	5.8E+01	9.7E+04	8.5E+06	6.7E+07	3.5E+05	4.0E+06	7.8E+03	7.1E+02	1.0E+07	1.6E+03	3.5E+01	1.0E+00	8.2E+01
38	6.6E+02	4.0E+00	1.9E+01	2.2E+05	4.7E+05	6.1E+10	8.8E+04	1.1E+06	7.9E+03	5.3E+02	1.4E+05	2.8E+04	2.8E+03	6.0E+00	3.4E+02
39	1.1E+03	2.0E+00	5.8E+01	5.0E+05	5.1E+06	1.0E+07	1.1E+05	1.6E+06	7.1E+03	1.0E+02	1.9E+06	3.4E+03	2.5E+02	1.4E+01	7.7E+01
40	4.1E+02	5.8E+04	3.6E+04	8.4E+11	6.1E+03	1.1E+05	1.0E+07	2.6E+12	1.8E+03	6.6E+04	1.0E+00	1.9E+04	7.6E+03	3.8E+05	8.1E+03
R_j	4.77	10.71	7.97	**2.67**	12.74	12.74	3.54	2.08	12.74	5.90	2.15	12.74	12.74	9.55	6.95

(k) Metric 11: Logarithmic loss of multi-class classification

Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.95	0.56	0.91	0.57	0.56	0.63	0.42	0.37	0.36	0.43	0.27	0.59	0.32	0.64	0.28
2	1.00	0.72	0.99	0.57	0.73	0.72	0.29	0.30	0.36	0.71	0.45	0.62	0.71	0.53	0.60
3	1.00	0.73	0.99	0.34	0.63	0.81	0.29	0.22	0.40	0.54	0.37	0.67	0.67	0.75	0.60
4	0.77	0.56	0.71	0.31	0.66	0.59	0.38	0.18	0.41	0.39	0.29	0.41	0.36	0.43	0.31
5	0.43	0.55	0.39	0.48	0.35	0.47	0.24	0.16	0.36	0.36	0.36	0.43	0.39	0.44	0.30
6	0.87	0.58	0.80	0.39	0.75	0.66	0.26	0.17	0.42	0.56	0.30	0.56	0.39	0.56	0.30
7	1.00	0.73	0.98	0.57	0.66	0.83	0.44	0.26	0.24	0.83	0.22	0.55	0.36	0.66	0.64
8	1.00	0.65	0.95	0.43	0.59	0.77	0.26	0.28	0.46	0.77	0.40	0.45	0.37	0.49	0.63
9	1.00	0.73	1.00	0.54	0.72	0.90	0.48	0.34	0.69	0.74	0.53	0.75	0.82	0.82	0.53
10	0.99	0.60	0.89	0.33	0.70	0.72	0.33	0.25	0.39	0.53	0.30	0.32	0.57	0.48	0.53
11	0.71	0.69	0.69	0.40	0.43	0.50	0.29	0.14	0.38	0.51	0.33	0.37	0.54	0.52	0.40
12	0.88	0.27	0.88	0.51	0.47	0.69	0.36	0.21	0.40	0.70	0.26	0.67	0.61	0.69	0.60
13	1.00	0.61	0.96	0.75	0.62	0.77	0.38	0.28	0.35	0.80	0.40	0.43	0.47	0.50	0.52
14	1.00	0.73	0.99	0.61	0.87	0.86	0.43	0.25	0.67	0.78	0.56	0.70	0.84	0.67	0.57
15	1.00	0.37	1.00	0.70	0.57	0.83	0.28	0.26	0.65	0.81	0.38	0.81	0.89	0.62	0.64
16	0.97	0.47	0.93	0.33	0.74	0.80	0.35	0.22	0.24	0.55	0.25	0.56	0.39	0.73	0.33
17	1.00	0.59	0.92	0.54	0.60	0.69	0.36	0.26	0.39	0.80	0.31	0.39	0.57	0.51	0.51
18	0.99	0.68	0.94	0.77	0.54	0.75	0.47	0.22	0.36	0.71	0.40	0.56	0.85	0.61	0.47
19	0.96	0.48	0.96	0.46	0.52	0.83	0.27	0.20	0.31	0.52	0.33	0.74	0.74	0.58	0.63
20	0.95	0.60	0.96	0.42	0.52	0.74	0.35	0.15	0.59	0.52	0.32	0.74	0.38	0.64	0.66

21	0.89	0.62	0.83	0.29	0.22	0.67	0.32	0.25	0.46	0.51	0.32	0.41	0.55	0.43	0.50
22	0.64	0.48	0.63	0.22	0.50	0.52	0.30	0.13	0.31	0.16	0.35	0.35	0.47	0.30	0.50
23	0.86	0.71	0.85	0.44	0.58	0.61	0.28	0.15	0.28	0.39	0.41	0.41	0.52	0.51	0.51
24	0.84	0.52	0.83	0.49	0.45	0.70	0.31	0.16	0.37	0.46	0.30	0.58	0.33	0.65	0.42
25	0.98	0.58	0.93	0.39	0.62	0.69	0.27	0.25	0.32	0.56	0.44	0.39	0.57	0.42	0.50
26	0.94	0.67	0.94	0.49	0.75	0.84	0.48	0.26	0.52	0.63	0.42	0.73	0.67	0.67	0.50
27	0.97	0.64	0.86	0.17	0.51	0.67	0.31	0.24	0.44	0.41	0.27	0.39	0.33	0.45	0.54
28	0.99	0.75	0.97	0.44	0.61	0.82	0.35	0.21	0.37	0.79	0.36	0.56	0.59	0.59	0.56
29	0.99	0.65	0.95	0.56	0.54	0.63	0.41	0.25	0.37	0.83	0.45	0.45	0.36	0.50	0.51
30	0.98	0.64	0.90	0.50	0.46	0.69	0.44	0.21	0.29	0.72	0.27	0.43	0.64	0.51	0.45
31	0.98	0.70	0.81	0.30	0.43	0.49	0.32	0.22	0.53	0.45	0.30	0.35	0.61	0.49	0.38
32	0.73	0.64	0.70	0.37	0.50	0.61	0.28	0.14	0.33	0.41	0.29	0.47	0.63	0.50	0.48
33	0.74	0.56	0.73	0.38	0.73	0.64	0.41	0.24	0.53	0.49	0.33	0.59	0.56	0.52	0.43
34	0.67	0.50	0.60	0.41	0.83	0.54	0.44	0.18	0.32	0.41	0.27	0.34	0.48	0.25	0.48
35	0.98	0.90	0.97	0.56	0.75	0.91	0.42	0.29	0.62	0.80	0.30	0.88	0.92	0.87	0.62
36	0.88	0.53	0.84	0.43	0.42	0.66	0.36	0.19	0.25	0.29	0.45	0.57	0.63	0.58	0.45
37	0.88	0.70	0.90	0.35	0.77	0.76	0.25	0.23	0.36	0.58	0.29	0.48	0.58	0.53	0.60
38	0.81	0.74	0.80	0.56	0.61	0.76	0.44	0.19	0.49	0.59	0.51	0.76	0.80	0.79	0.51
39	1.00	0.72	1.00	0.66	0.63	0.82	0.33	0.29	0.60	0.82	0.47	0.69	0.80	0.74	0.50
40	0.82	0.49	1.25	0.49	0.26	0.33	0.45	0.42	0.49	0.49	0.42	0.48	1.30	0.81	0.41
$\mathbf{R_j}$	**1.69**	6.15	2.26	12.08	5.38	3.41	12.28	14.62	10.41	9.05	11.54	6.72	6.31	9.44	8.67

(l) Metric 12: Mean accuracy with constraints

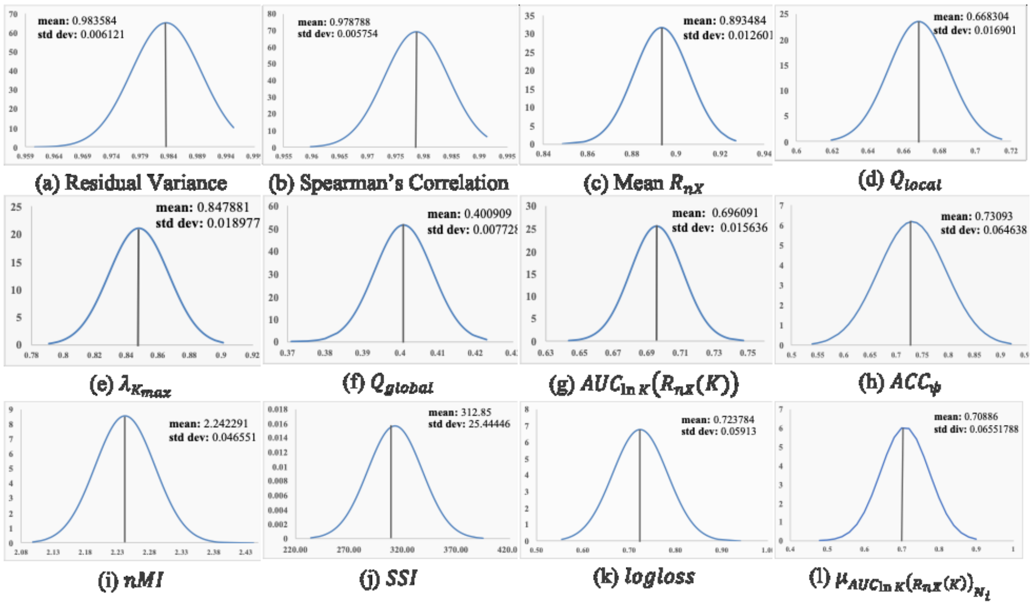
Dataset	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
1	0.50	0.66	0.03	0.81	0.04	0.18	0.07	0.10	0.61	0.05	0.28	0.32	0.54	0.02	0.11
2	0.48	0.68	0.05	0.72	0.06	0.18	0.07	0.10	0.58	0.06	0.28	0.32	0.23	0.01	0.02
3	0.57	0.65	0.05	0.74	0.04	0.18	0.07	0.10	0.60	0.06	0.29	0.34	0.19	0.11	0.02
4	0.56	0.66	0.04	0.69	0.02	0.18	0.06	0.10	0.65	0.04	0.27	0.39	0.39	0.10	0.01
5	0.46	0.51	0.46	0.50	0.47	0.51	0.47	0.49	0.47	0.49	0.48	0.47	0.45	0.44	0.45
6	0.60	0.68	0.05	0.76	0.04	0.18	0.06	0.10	0.65	0.04	0.27	0.35	0.46	0.06	0.11
7	0.48	0.67	0.06	0.79	0.05	0.20	0.06	0.10	0.63	0.06	0.27	0.34	0.22	0.10	0.01
8	0.53	0.66	0.04	0.72	0.04	0.20	0.07	0.10	0.63	0.05	0.27	0.37	0.27	0.10	0.11
9	0.45	0.66	0.03	0.74	0.08	0.18	0.06	0.06	0.59	0.05	0.27	0.37	0.29	0.10	0.02
10	0.42	0.68	0.05	0.74	0.02	0.18	0.06	0.10	0.62	0.07	0.27	0.37	0.21	0.10	0.11
11	0.56	0.75	0.07	0.74	0.40	0.19	0.07	0.06	0.86	0.05	0.31	0.39	0.57	0.10	0.02
12	0.48	0.75	0.05	0.83	0.16	0.19	0.06	0.02	0.71	0.07	0.30	0.34	0.56	0.03	0.11
13	0.56	0.70	0.04	0.74	0.86	0.18	0.06	0.10	0.63	0.05	0.27	0.38	0.28	0.11	0.01
14	0.50	0.69	0.03	0.74	0.33	0.18	0.09	0.10	0.56	0.05	0.27	0.41	0.20	0.10	0.02
15	0.53	0.67	0.06	0.83	0.13	0.18	0.06	0.06	0.59	0.05	0.28	0.35	0.58	0.10	0.11
16	0.52	0.67	0.03	0.81	0.19	0.18	0.07	0.10	0.63	0.05	0.28	0.34	0.56	0.10	0.11
17	0.55	0.68	0.04	0.23	0.51	0.18	0.06	0.10	0.59	0.06	0.27	0.33	0.34	0.02	0.01
18	0.50	0.67	0.05	0.79	0.11	0.18	0.06	0.10	0.82	0.04	0.31	0.31	0.52	0.10	0.02
19	0.57	0.69	0.04	0.72	0.32	0.18	0.06	0.10	0.83	0.04	0.27	0.34	0.27	0.10	0.02
20	0.57	0.67	0.03	0.76	0.15	0.18	0.06	0.02	0.63	0.04	0.30	0.38	0.41	0.02	0.02
21	0.52	0.67	0.06	0.67	0.02	0.18	0.06	0.10	0.64	0.04	0.27	0.34	0.22	0.11	0.11
22	0.57	0.67	0.04	0.74	0.03	0.18	0.07	0.10	0.56	0.04	0.29	0.33	0.39	0.01	0.11
23	0.47	0.65	0.03	0.79	0.06	0.19	0.06	0.10	0.53	0.05	0.29	0.31	0.52	0.11	0.11
24	0.59	0.66	0.04	0.83	0.06	0.19	0.07	0.10	0.58	0.05	0.29	0.31	0.55	0.10	0.03
25	0.53	0.69	0.06	0.74	0.79	0.18	0.06	0.10	0.56	0.05	0.27	0.30	0.28	0.11	0.02
26	0.55	0.71	0.03	0.76	0.16	0.18	0.06	0.10	0.65	0.04	0.28	0.42	0.47	0.02	0.11
27	0.58	0.66	0.03	0.69	0.56	0.18	0.07	0.10	0.63	0.05	0.28	0.30	0.43	0.10	0.02
28	0.53	0.68	0.03	0.76	0.14	0.18	0.06	0.10	0.62	0.05	0.27	0.37	0.19	0.02	0.03
29	0.55	0.69	0.07	0.76	0.88	0.18	0.06	0.10	0.59	0.05	0.27	0.35	0.23	0.11	0.11
30	0.55	0.70	0.03	0.74	0.78	0.18	0.06	0.10	0.56	0.06	0.26	0.34	0.30	0.01	0.02
31	0.57	0.69	0.03	0.72	0.02	0.18	0.06	0.10	0.56	0.05	0.28	0.31	0.50	0.10	0.11
32	0.57	0.66	0.03	0.21	0.06	0.18	0.06	0.10	0.60	0.05	0.28	0.33	0.53	0.11	0.11
33	0.42	0.66	0.04	0.79	0.27	0.18	0.08	0.10	0.56	0.05	0.26	0.35	0.27	0.10	0.11
34	0.42	0.68	0.04	0.79	0.04	0.18	0.06	0.10	0.62	0.04	0.27	0.35	0.56	0.10	0.11
35	0.43	0.71	0.04	0.76	0.36	0.19	0.08	0.03	0.57	0.07	0.28	0.43	0.57	0.10	0.04
36	0.49	0.70	0.04	0.88	0.33	0.19	0.06	0.04	0.60	0.08	0.27	0.40	0.58	0.11	0.02
37	0.55	0.69	0.05	0.69	0.04	0.18	0.06	0.10	0.59	0.05	0.28	0.34	0.29	0.01	0.11
38	0.55	0.68	0.05	0.86	0.43	0.18	0.08	0.11	0.60	0.06	0.31	0.39	0.44	0.03	0.11
39	0.56	0.68	0.04	0.69	0.13	0.18	0.07	0.10	0.63	0.05	0.27	0.65	0.27	0.11	0.11
40	0.56	0.16	0.99	0.07	0.14	0.20	0.52	0.10	0.85	0.13	0.22	0.29	0.13	0.73	0.13
$\mathbf{R_j}$	11.46	13.56	5.51	**1.00**	9.69	14.64	7.21	2.49	12.56	6.05	8.79	10.10	10.21	3.08	3.64

Table 4.5: Friedman's rank of all algorithms for each of the contextual metrics

Context	Metric	Umap	t-sne	FItsne	PCA	Tmp	Mt-sne	Isomap	KPCA	LEM	Ltsa	nMDS	HLLE	LLE	LVis	MVU
Pattern analysis	$\hat{\sigma}^2$	1.74	**1.44**	5.46	7.10	14.4	6.05	9.18	5.33	10.1	9.46	7.33	10.4	11.3	7.64	12.8
	ρ_s	2.08	**1.87**	5.90	5.67	7.74	5.00	7.28	6.77	12.5	14.8	7.33	10.8	10.3	9.49	12.2
	$\mu_{R_{nX}}$	**2.13**	3.41	4.92	6.74	12.4	5.87	3.33	6.64	8.95	9.41	8.46	11.5	11.6	9.92	14.5
	Q_{local}	**1.36**	3.69	7.03	6.46	14.5	6.97	2.00	6.95	10.5	12.4	6.36	8.82	9.21	11.5	12.0
	$\lambda_{K_{max}}$	**3.60**	3.88	6.90	7.53	10.6	6.92	4.49	5.36	9.79	10.3	7.32	8.08	8.44	13.8	13.2
	Q_{global}	6.87	7.26	7.15	**1.26**	9.22	7.29	5.86	3.82	14.7	12.6	2.10	10.1	8.18	10.9	12.5
Pred. modeling	$AUC_{\ln K}(R_{nX}(K))$	5.59	**1.79**	3.03	10.3	14.6	6.41	2.79	9.15	11.9	12.8	6.10	6.05	8.49	10.5	11.1
	ACC_{ϕ}	7.76	6.88	7.91	**6.74**	9.18	7.79	9.13	7.35	8.28	8.35	7.69	8.31	8.50	7.91	8.22
Poor quality data	nMI	3.46	**3.01**	3.49	9.56	13.9	4.04	13.1	8.79	9.42	9.71	12.3	4.49	4.35	5.68	14.9
	SSI	4.77	10.7	7.97	2.67	12.7	12.7	3.54	2.08	12.7	5.90	**2.15**	12.74	12.7	9.55	6.95
Limited data	$logloss$	**1.69**	6.15	2.26	12.0	5.38	3.41	12.2	14.6	10.4	9.05	11.5	6.72	6.31	9.44	8.67
Limited resources	$\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}}$	11.4	13.5	5.51	**1.00**	9.69	14.64	7.21	2.49	12.5	6.05	8.79	10.1	10.2	3.08	3.64

Note: In the above table, LVis refers to LargeVis, Tmp refers to Trimap, and Mt-sne signifies Multicore t-SNE with 8 cores. The highest rank for the best performing algorithms in each metric is highlighted in bold.

variance ($\hat{\sigma}^2$) and Spearman rank correlation (ρ_s), t-SNE and its different variations (e.g., FIt-SNE and 8-Core t-SNE) can be good choices for a practitioner. Nevertheless, for metrics such as $\mu_{R_{NX}}$, Q_{local}, and $\lambda_{K_{max}}$, UMAP and Isomap have outperformed t-SNE. The reason being: whilst t-SNE effectively preserves small neighborhoods in an input dataset, it ignores the overall structure of the data. Hence, t-SNE delivers best results when the value of K is relatively small. For example, t-SNE is ranked in third position for the metric $\mu_{R_{NX}}$ as it computes an average of the agreement between the dataset and its embedding for all possible K values (i.e., $1 \leq K \leq N-1$) in the dataset. Moreover, for Q_{local} and $\lambda_{K_{max}}$ the value of K_{max} is automatically computed from the input data. Although, K_{max} can successfully define the size of 'locality' [4] in a dataset, research shows that a single value for K_{max} is often suboptimal [4] for all neighborhoods in a dataset. As a result, for Q_{local} and $\lambda_{K_{max}}$ UMAP outperforms t-SNE. On the other hand, algorithms such as PCA, KernelPCA, and non-metric MDS outperform t-SNE and its variants for



(a) Residual Variance (b) Spearman's Correlation (c) Mean R_{nX} (d) Q_{local}

(e) $\lambda_{K_{max}}$ (f) Q_{global} (g) $AUC_{\ln K}\left(R_{nX}(K)\right)$ (h) ACC_{ψ}

(i) nMI (j) SSI (k) $logloss$ (l) $\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}}$

Figure 4.1: Sampling distribution for12 metric scores from 10,000 sample of the Credit card dataset

the metric Q_{global}. This shows that, while PCA effectively separates different neighborhoods in an embedding, it perturbs the internal structure of the neighborhoods.

In terms of DR accuracy metrics, both t-SNE and PCA perform well with KNN prediction accuracy. However, for $AUC_{\ln K}\left(R_{NX}(K)\right)$, the local structure preserving algorithms, namely UMAP, t-SNE, and different versions of LLE, outperform non-metric MDS and also both linear and non-linear PCA. We note that KNN prediction accuracy is an indirect quality metric that depends on the quality of input data (e.g., non-noisy labels or class imbalance in data). However, the performance of t-SNE for $AUC_{\ln K}(R_{NX}(K))$ validates our outcome for the KNN prediction accuracy.

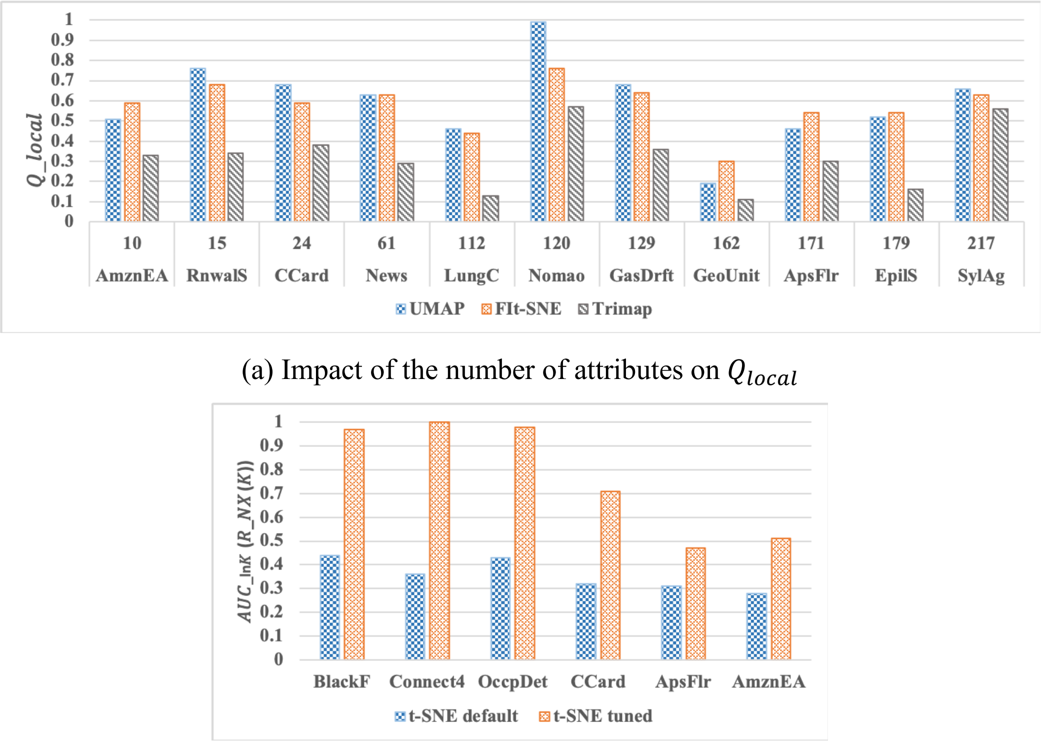
Table 4.5 shows, overall locally focused NLDR techniques (e.g., UMAP, t-SNE, LLE) handle missing values in data better than globally-focused techniques nMDS and KernelPCA. However, with outliers in the input data, the exact opposite thing happens (i.e., PCA, KernelPCA, and nMDS outperform UMAP and t-SNE)! These results show that, while outliers impact the preservation of the global structure of a dataset more than its local neighborhood structures, missing values in the data influence the conservation of the local neighborhoods. In terms of reproducibility with limited input data, UMAP outperforms all the globally focused NLDR algorithms with different variants of t-SNE (i.e., FIt-SNE and 8-core t-SNE) holding the second-best positions. Nevertheless, in terms of accuracy with limited resources both linear and non-linear versions of PCA outperform all of the other algorithms. In order to justify the fidelity of the rankings presented in Table 4.5, in Figure 4.1 we present the sampling distribution of the metric scores for the Credit card dataset. In the Figure we depict the distribution of the metric scores for 10,000 random samples from the dataset. The samples were selected with replacement. The charts also show the mean and standard deviation of the metric scores for the 10,000 random samples. The normality of the distribution proves the legitimacy of computing the mean score from all the samples for each dataset.

Overall, Table 4.5 shows that for any specific analytical context and chosen quality metric, some DR algorithms perform better than others. However, Amid et al. [17] and Shiming et al. [34] discuss that the performances of DR algorithms also depend on factors such as the input datasets and the hyperparameter combinations of the algorithms. Hence, we analyze the impact of these two factors on the obtained results.

4.3.1.1 Impact of Input Datasets

High-dimensional datasets have some unique characteristics [9] that influence any kind of exploratory analysis performed on them. In this section, we discuss the effect of three such characteristics that might influence the performance of any DR technique.

i. **Number of Attributes:** The computation of proximity relationships [11] among data-points forms the basis of almost all NLDR [62] techniques.



(a) Impact of the number of attributes on Q_{local}

(b) Impact of parameter tuning on $AUC_{\ln K}\left(R_{nX}(K)\right)$

Figure 4.2: Influence of additional factors on the performance of DR algorithms

Such proximity computation is highly impacted [24] by the size of the data vector (i.e., attributes in the dataset). The reason being: the larger the size of each data vector the more components are considered in the expressions for proximity calculation. In this section, we report our analysis on the influence of the number of attributes on the retention of local structure after DR. As shown in Figure 4.2(a), we focus on the metric Q_{local}. In Figure 4.2 (a), we compare the Q_{local} scores for the best, mediocre, and worst performing algorithms namely: UMAP, FIt-SNE, and Trimap for 11 datasets. Only these datasets are considered because, either their Q_{local} scores were less than 70% (cf. Table 4.4.d) for the best performing algorithm or they have more than 50 attributes (cf. Table 4.3). The figure shows for datasets (i.e., News, Gas-Drift, Aps-Failure, Sylva-Agnostic, Lung-Cancer, Geographical-Unit, and Epileptic-Seizure) with a relatively high number of attributes the local structure is retained around 10% to 65% by all the algorithms. However, there are some exceptions. On the one hand, for the Nomao dataset with 120 attributes UMAP could retain around 99% of its local structure. On the other hand, for the Amazon-Employee-Access dataset with merely 10 attributes all the three algorithms retained between 30% to 60% of its local structure. This confirms that the structural retention was impacted by other aspects of high-dimensional data, not just the attribute count.

ii. **Categorical Features in Data:** Performing any kind of statistical analysis with categorical attributes can be a challenging [9] task. Primarily because, it is not only impossible to compute statistics like mean, median with such attributes but also in most cases, such attributes cannot be represented ordinally. As a result, they impact on the calculation of the distance measures used in different DR algorithms. Although there are similarity measures (e.g., Gower Similarity) that can effectively handle categorical data, DR algorithms considered for our experiments are designed [2], [3] only to work with numeric continuous attributes. For example, during our analysis of the poor Q_{local} scores for the Amazon-Employee-Access and Renewal Sales datasets (cf. Figure 4.2(a)) with as low as 10 and 15 attributes respectively, we discovered that whilst all the attributes in the former were categorical, 14 out of 15 attributes in the later had categorical data.

iii. **Multivariate Relationships among Attributes:** Multivariate relationships [63] among attributes may cause redundancy in data [9]. Our experiments show that such redundancies also impact on the proximity computation among data-points. For example, during our experiments, we noticed that for datasets such as Credit-card, Aps-Failure, SUSY, Phishing-Websites, Sylva-Prior, and Gas-drift almost all DR algorithms performed poorly in terms of computing local quality metrics such as $\mu_{R_{NX}}$ and Q_{local}. We think that, the high correlation among the attributes in these datasets affected the performances of our chosen DR techniques.

4.3.1.2 Impact of Hyperparameters

Each DR algorithm has a set of hyperparameters that are tuned uniquely for the dataset under consideration [3]. Depending on their construction, DR algorithms have different number of parameters where the parameters vary in their level of importance. For example, the ***neighborhood size*** (i.e., K) is one of the most influential parameters for many DR algorithms. As our experiments show (cf. Table 4.5) most locally focused DR algorithms (e.g., t-SNE, UMAP, FIt-SNE) outperform their competitors for smaller values of K. Hence, in case the value of K is not optimally chosen, the structural retention after DR can deteriorate even for the best performing algorithms. Among other parameters, the importance of the ***perplexity*** [26] parameter for t-SNE is well known among researchers [62], [64]. The parameter signifies the number of closest neighbors to consider when determining the local structure of neighborhoods before the transformation. Tang et al. [13] show, with a slight change in the perplexity value (traditionally ranged between 5 and 50) can highly impact the relative positioning of data-points in the embedding. The same concern is also valid for other t-SNE based algorithms namely McoretSNE and FIt-SNE. The number of ***iterations*** in the DR algorithms is also known for its significance [3]. The iterations allow the refinement of the relative positioning of datapoints after DR, hence reducing the optimization error. Our experiments show that a reduced number of iterations can negatively influence the accuracy and structural retention after DR. Finally, the ***distance function*** chosen for any DR algorithm makes a key impact on the identified proximities among data points. Although most of our chosen DR algorithms consider the Euclidean distance to be their default distance function, the distance functions can be altered depending on the dataset. Since the performance of the distance function depends on the underlying structure of the original manifold, in our experiments, however, alerting Euclidean distance to other distance functions (e.g., 'Cosine' distance for t-SNE and 'Manhattan' distance for UMAP) did not show any significant changes in the output.

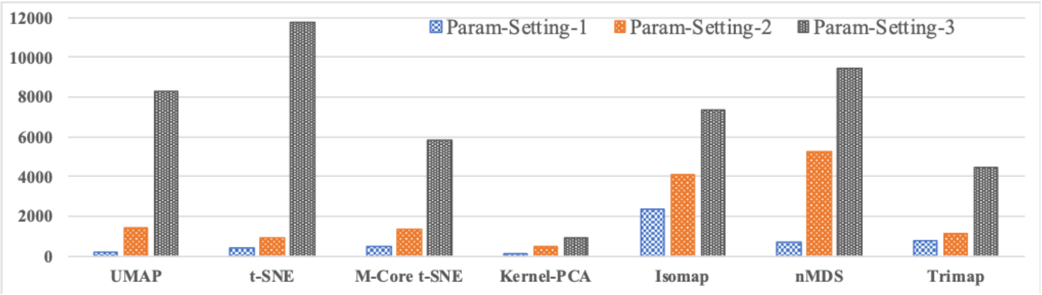


Figure 4.3: Execution times for grid-search with different hyper-parameter combinations discussed in Table 4.5.

In Figure 4.2(b) we show the impacts of tuned versus default hyperparameters for t-SNE on 6 datasets. Here, we focus on the accuracy of the embedding (i.e., for metric $AUC_{\ln K}(R_{nX}(K))$)). Figure 4.2(b) shows that, with the appropriate tuning of the hyperparameters, t-SNE performs with much higher accuracy for all 6 datasets than with its default parameter combinations. During the parameter tuning for our experiments, we learned that the hyperparameters of some algorithms (e.g., PCA, Isomap, nMDS, Trimap, KernelPCA, LLE) are easier to tune than others (e.g., t-SNE, FIt-SNE, McoretSNE, and UMAP). We identified the reasons to be (1) the number of parameters used in grid-search, (2) the range of parameter values, and (3) the computational complexities of DR techniques. For example, the tuning-duration of only the perplexity parameter for t-SNE is significantly lower than the duration for tuning both perplexity and number of iterations. Similarly, for UMAP tuning only neighborhood size and number of target dimensions could be done in much lower time than tuning the two parameters along with minimum distance among neighbors. Moreover, when determining the best kernel function for KernelPCA, tuning with only three functions require much less time than tuning with six functions. Finally, for the algorithms that are known for their speed limitations (e.g., t-SNE [2]) the overall tuning duration was higher than others (e.g., KernelPCA). In the case of some algorithms (e.g., t-SNE, UMAP) increasing the range or number of the tuning parameters exponentially increased the tuning duration. We consider these algorithms like the ones that are difficult to tune. A

Table 4.6: Parameter Settings for the Grid Search for Algorithms Presented in Figure 4.3

Algorithm	Param-Setting 1	Param-Setting 2	Param-Setting 3
UMAP	n_neighbors: 5, 20 n_components: 15, 25, 50	n_neighbors: 2, 5, 10, 20, 50, 100, 200 n_components: 1, 5, 15, 20, 25, 50	n_neighbors: 2, 5, 10, 20, 50, 100, 200 n_components: 15, 20, 25, 50 min_dist: 0.0, 0.1, 0.25, 0.5, 0.8, 0.99
t-SNE	Perplexity: 5, 30	Perplexity: 5, 15, 25, 30	Perplexity: 5, 10, 15, 20, 25, 30 n_iter: 250, 500, 750, 1000
MCore-t-SNE	Perplexity: 5, 30 n_jobs: 2, 4	Perplexity: 5, 15, 25, 30 n_jobs: 2, 4, 6, 8	Perplexity: 5, 10, 15, 20, 25, 30 n_jobs: 2, 4, 6, 8 n_iter: 250, 500, 750, 1000
KernelPCA	Kernel: "linear", "poly"	Kernel: "linear", "poly", "rbf", "sigmoid"	Kernel: "linear", "poly", "rbf", "sigmoid", "cosine", "precomputed"
Isomap	n_neighbors: 5, 20 n_components: 15, 25, 50	n_neighbors: 2, 5, 10, 20 n_components: 1, 5, 15, 20	n_neighbors: 2, 5, 10, 20, 50, 100, 200 n_components: 1, 5, 15, 20, 25, 50
nMDS	n_neighbors: 5, 20 n_components: 15, 25, 50	n_neighbors: 2, 5, 10, 20 n_components: 1, 5, 15, 20	n_neighbors: 2, 5, 10, 20, 50, 100, 200 n_components: 1, 5, 15, 20, 25, 50
Trimap	n_inliers: 5, 10 n_outliers: 5, 10 n_iters: 200, 400	n_inliers: 5, 8, 10,12 n_outliers: 5, 8, 10, 12 n_iters: 200, 400, 600	n_inliers: 5, 8, 10, 12, 15, 18, 20 n_outliers: 5, 8, 10, 12, 15, 18, 20 n_iters: 200, 300, 400, 500, 600

more detailed analysis of our assessment of our hyperparameter tuning is presented in Figure 4.3 and Table 4.6.

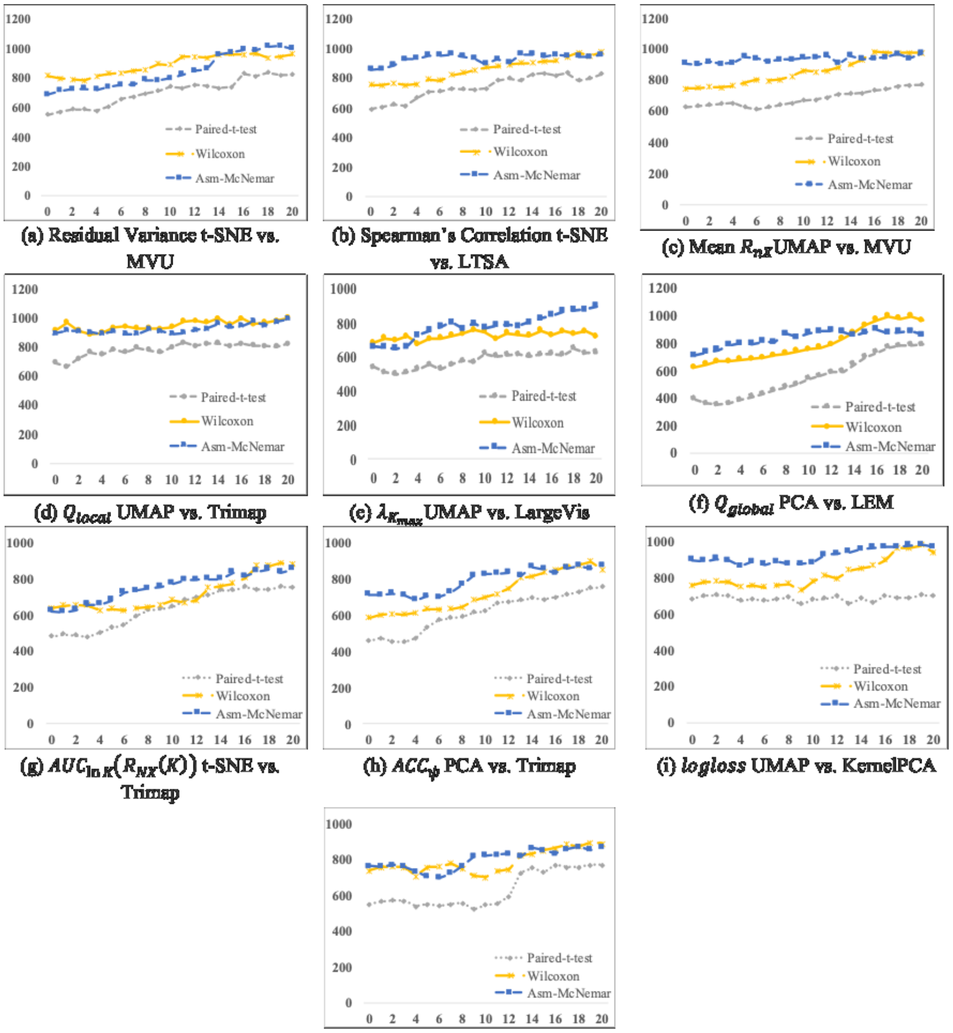
Figure 4.3 plots the differences in training-time with different hyperparameter settings with 7 DR algorithms that performed better than others in different contextual metrics. The exhaustive grid-search of hyperparameters were performed on a system with 192 GB RAM, 12 processor cores and 350 hard-drive space. The Figure 4.3 shows that when training with different value ranges for hyperparameters or different number of hyperparameters, the tuning time varies for each and every algorithm. Whereas for some algorithms, the computational time varies more significantly (e.g., t-SNE) than for others. We think the reason behind this is the executional complexity of these algorithms.

4.3.2 Statistical Analysis of Results

In the next two sub-sections, we summarize the results and discuss our analysis for both pairwise and overall statistical comparisons among the chosen algorithms. In our experiments, we use one parametric (i.e., paired t-test – Eq. 4.21) and two non-parametric (i.e., Wilcoxon signed rank – Eq. 4.23 and asymptotic McNemar's – Eq. 4.24) statistical tests. In our experiments, following the guidelines of Hastie et al. [31], Demśar et al. [39] and Mohammadi et al. [37] we consider more than 70% rejection of H_0 in 1000 experiments as a sign of statistical significance. The line-plots presented in Figure 4.4 comparatively present the results of all the three tests for each metric. Based on the test results we draw our conclusions for the statistical significance of pairwise differences. For overall comparisons of algorithms, we execute Friedman's test (cf. Eq. 4.26) along with Nemenyi and Holm corrections.

4.3.2.1 Pairwise Statistical Comparisons of Algorithms

In this section, for each evaluation metric, we perform pairwise statistical comparisons among the best and the worst performing algorithms identified in Table 4.5. The primary goal of this analysis is to assess whether the difference in the performance of the algorithms might be a good candidate for generalization beyond this study. For these pairwise comparisons, as discussed in Algorithm 1 (line 22), we randomly sampled 40 datasets with replacement and repeated this process 1000 times. Figure 4.4 summarizes our overall analysis results for 10 out of 12 metrics. The results for the remaining two metrics (i.e., nMI and SSI) are presented in Tables 4.6 and 4.7. All the line charts in Figure 4.4 (i.e., Fig. 4.4.a to Fig. 4.4.j) show the combined results of the Wilcoxon signed rank test, paired t-test, along with asymptotic McNemar's test for their number of the rejected null hypothesis in all 1000 experiments. The results of the three tests are presented together to ease the assessment of their similarities and contradictions. The x-axis



(a) Residual Variance t-SNE vs. MVU

(b) Spearman's Correlation t-SNE vs. LTSA

(c) Mean R_{nX}UMAP vs. MVU

(d) Q_{local} UMAP vs. Trimap

(e) $\lambda_{K_{max}}$ UMAP vs. LargeVis

(f) Q_{global} PCA vs. LEM

(g) $AUC_{\ln K}(R_{NX}(K))$ t-SNE vs. Trimap

(h) ACC_{ψ} PCA vs. Trimap

(i) *logloss* UMAP vs. KernelPCA

(j) Mean accuracy with resource constraint PCA vs. 8-Core-tSNE

Figure 4.4: Comparisons of Statistical Significance Tests (Wilcoxon signed rank, paired t-test, and asymptotic McNemar's test) for the number of the rejected null hypothesis in 1000 experiments

of all the line graphs in Figure 4.4 represents the additional bias ν varying from 0 to 20 (cf. Equation 4.29). Ideally, when $\nu = 0$ the plots should depict an unbiased statistical analysis between the sample datasets. However, in case there is indeed any statistically significant differences among the algorithms under inspection, with an increasing ν we should expect to see an upward trend in the plots.

Figure 4.4.a presents our analysis of DR techniques for Residual variance ($\hat{\sigma}^2$) and compares the best performing algorithm t-SNE to the worst-performing algorithm MVU. On the other hand, Figure 4.4.b depicts the results for Spearman's Rank Correlation (ρ_s) and compares t-SNE with LTSA the highest and the lowest scorer for ρ_s respectively. Similarly, Figure 4.4.c compares UMAP with MVU for the metric $\mu_{R_{NX}}$. In all the three charts we see an upward trend for all the statistical tests with increasing bias. In Fig 4.4.a, although for 0 bias the t-test has rejected only about 55% of the null hypotheses, beyond $\nu = 2$, the Wilcoxon, paired-t, and asymptotic McNemar's tests show competitive performances with the number of rejected H_0 being between 70% and 90%. In Figures 4.4.b and 4.4.c, the number of rejected (H_0) hypotheses remain (more-or-less) similar (i.e., 80% to 100%) for both the non-parametric tests. Whereas, in Figure 4.4.b, considering all of the bias values, the asymptotic McNemar's test rejected most (82%, to 100%) of the hypotheses. From this analysis we make three observations: firstly, in terms of residual variance (i.e., Fig. 4.a), t-SNE and MVU are indeed significantly different from each other for $\nu \geq 2$. However, for $\nu < 2$, only the non-parametric statistical tests releveled any statistical significance. Secondly, in terms of Spearman's correlation (i.e., Fig. 4.b), considering the results of the asymptotic McNemar's test for all bias values, we conclude that the performance differences of t-SNE and LTSA are statistically significant. Finally, for Figure 4.4.c, when $< 10,$ the paired t-test, cannot identify any statistical significance in the differences between UMAP and MVU. However, McNemar's asymptotic test and the Wilcoxon signed ranks test have constantly rejected around 70% to 100% of the null hypotheses, hence confirming the statistical significance.

Figures 4.4.d, 4.4.e, and 4.4.f present the results of the statistical comparisons among the best and the worst performing algorithms for the metrics Q_{local}, $\lambda_{K_{max}}$, and Q_{global}. In Figure 4.4.d we compare the differences between UMAP and Trimap for Q_{local}; in Figure 4.4.e, the $\lambda_{K_{max}}$ scores of UMAP and LargeVis and in Figure 4.4.f, the Q_{global} scores of PCA and LEM. Likewise, in figures 4.4.e and 4.4.f, in Figure 4.4.d the number of rejected null hypothesis increased with a raising value for bias for all the three tests. Moreover, for all the tests, for $\nu > 4$, the number of rejected null hypotheses was between 75% -100%. Similarly, in Figure 4.4.e, since the asymptotic McNemar's test has consistently rejected more than 80% of the null hypothesis for $\nu > 6$, we conclude that the difference in loss of K-ary neighborhood information due to the size of K is statistically significant between UMAP and LargeVis. Finally, in Figure 4.4.f, although the paired t-test rejected between only 40% to 60% of H_0 when $\nu < 14$, the asymptotic McNemar's test consistently rejected more than 70% of the null hypothesis for all values of ν. Hence, we think that the differences between PCA and LEM for Q_{global} are statistically significant.

In Figures 4.4.g, 4.4.h, 4.4.i, 4.4.j we illustrate the results of the statistical comparisons for $AUC_{\ln K}(R_{NX}(K))$, ACC_{ψ}, $logloss$ and accuracy with limited computational resources. Among these figures, Figure 4.4.g compares t-SNE with Trimap that is the best and worst-performing algorithms in terms of $AUC_{\ln K}(R_{NX}(K))$; Figures 4.4.h and 4.4.i similarly compare PCA with Trimap and UMAP with KernelPCA for ACC_{ψ} and $logloss$ respectively. Finally, Figure 4.4.j compares PCA with 8-Core t-SNE for $\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}}$. In figures 4.4.g, 4.4.h, 4.4.i, and 4.4.j we have seen an upward trend in the number of rejected H_0 with an increasing ν. Nevertheless, in the figures 4.4.g, and 4.4.h, we can see that with lower values for biases (i.e., when $\nu < 12$) the number of rejected null hypothesis have been between 40%-65% for Wilcoxon signed ranks and Paired-t tests. However, in Figure 4.4.h for the asymptotic McNemar's test, the number of rejected null hypotheses were constantly more than 70% for all bias values. In these cases, even if there is a statistical significance in the differences between the compared algorithms for the chosen metrics, the Wilcoxon signed ranks and Paired-t tests could not find them. In the case of Figure 4.4.i, both the non-parametric tests rejected around 70% to 100% of the null hypotheses with all bias values. The charts show that the introduction of bias was indeed useful in revealing the statistical significance in the differences in the algorithms. For Figure 4.4.j however, the

Table 4.7: Comparisons of Wilcoxon signed rank, paired t-test, with asymptotic McNemar's test for the impact of missing values with bias = 15; (below diagonal: average p-value/Rep(p), above diagonal: rejected null hypothesis/Rep(e))

	t-SNE	KPCA	MVU
t-SNE		675/0.5	895/0.8
KPCA	0.17/0.8		794/0.6
MVU	0.05/0.9	0.13/0.8	

(a) Wilcoxon signed-rank test

	t-SNE	KPCA	MVU
t-SNE		509/0.49	728/0.6
KPCA	0.19/0.82		679/0.56
MVU	0.07/0.93	0.08/0.92	

(b) Paired t-test

	t-SNE	KPCA	MVU
t-SNE		624/0.53	974/0.94
KPCA	0.18/0.83		619/0.52
MVU	0.04/0.96	0.18/0.83	

(c) McNemar's test

Table 4.8: Comparisons of Wilcoxon signed rank, paired t-test, with asymptotic McNemar's test for outlier values with bias = 15; (below diagonal: average p-value/Rep(p), above diagonal: rejected null hypothesis/Rep(e))

	nMDS	MVU	HLLE
nMDS		644/0.5	924/0.85
MVU	0.17/0.8		601/0.51
HLLE	0.04/0.9	0.17/0.8	

(a) Wilcoxon signed-rank test

	nMDS	MVU	HLLE
nMDS		589/0.51	792/0.67
MVU	0.16/0.85		633/0.53
HLLE	0.14/0.87	0.16/0.85	

(b) Paired t-test

	nMDS	MVU	HLLE
nMDS		672/0.55	952/0.9
MVU	0.18/0.83		627/0.53
HLLE	0.03/0.97	0.18/0.83	

(c) McNemar's test

number of the rejected null hypothesis for all the tests was between 50% - 80%. However, for bias less than 12 the Paired t-test could not find any statistical significance in the performance differences of PCA and 8-Core t-SNE. Moreover, for $6 \leq \nu \leq 12$, the statistical significance of the dissimilarities of the results from PCA and 8-Core t-SNE due to resource constraints cannot be strongly concluded even using the non-parametric tests. Here, we think apart from the resource constraints additional factors (e.g., quality of input data, other hyperparameters of the algorithms, etc.) are in play.

For the metrics evaluating the impact of missing and outlier values on DR algorithms, we also present the replicability values of the statistical tests. In Tables 4.6 and 4.7, we tabulate the comparisons among the best, mediocre, and worst-performing algorithms in terms of the impact of missing (cf. Table 4.7) and outlier values (cf. Table 4.8). Like Mohammadi et al. [37], in both Tables 4.6 and 4.7, we report the results for bias $\nu = 15$. For both the tables, the numbers below the diagonals represent the average p-value and their corresponding replicability measure Rep(p) (cf., Equation 4.31), as the numbers above the diagonals signify the number of rejected null hypotheses and their corresponding replicability measure Rep(e) (cf., Equation 4.30). In both Tables 4.6 and 4.7, the statistical significance in the differences of the best and worst-performing algorithms are visible with average p-values being less than 0.05 (i.e., our predefined threshold) for both the non-parametric tests. However, from both the tables the differences between the best and mediocre performing algorithms could not be determined to be statistically significant. On the other hand, the replicability Rep(p) of the tests being above 80% we think the statistical tests to be reliable.

In support of our pairwise statistical analysis of alogrithms presented above, we present the detailed results of the McNemar's asymptotic test used for the statistical analysis of the DR algorithm performances in Table 4.9. The primary reason behind presenting the results of only McNemar's asymptotic test among the three other tests (Wilcoxon's Signed Rank test, Paired t-test, and Exact McNemar's test) is: our experiments have proven the McNemar's Asymptotic test to be the most powerful when comparing DR algorithms. In Tables 4.9.a to 4.9.l each row presents the pairwise comparisons between the two algorithms marked by the row and column headers. Hence, the diagonals in the Tables 4.9.a to 4.9.l remain blank as they indicate the statistical comparison of an algorithm to itself. In the Tables 4.9.a to 4.9.l below the diagonal we present the average p-value of McNemar's Asymptotic test and its replicability measure Rep(p) separated by a '/'. Similarly, above the diagonals in Tables 4.9.a to 4.9.l we show the number of rejected null hypothesis in 1000 experiments using the McNemar's Asymptotic test and its corresponding replicability measure Rep(e).

4.3.2.2 Statistical Comparisons of Multiple Algorithms

In this section, we briefly discuss the results of the statistical comparisons among multiple DR algorithms with the Friedman test along with Nemenyi and Holm corrections of the p-values and depict in Figure 4.5. As per our assumptions, the Friedman test has been the most conservative of the three and has rejected the least number of hypotheses for all 12 metrics. Hence, in our analysis, we primarily focus on the results obtained from Nemenyi and Holm corrections. Among the 12 metrics,

Table 4.9: Statistical Significance Analysis for Pairwise Combinations of All Algorithms using McNemar's Asymptotic Test for bias value of 15

(a) Metric 1: Residual variance ($\hat{\sigma}^2$)

	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		330/0.55	542/0.5	677/0.56	907/0.83	585/0.51	683/0.56	621/0.52	892/0.8	855/0.75	721/0.59	922/0.85	845/0.73	751/0.62	893/0.8
t-SNE	0.18/0.83		428/0.5	723/0.59	674/0.56	396/0.52	582/0.51	623/0.52	764/0.63	698/0.57	523/0.5	669/0.55	601/0.51	546/0.5	637/0.53
FIt_SNE	0.13/0.88	0.21/0.8		503/0.49	759/0.63	473/0.5	697/0.57	584/0.51	716/0.59	771/0.64	682/0.56	629/0.53	634/0.53	726/0.6	582/0.51
PCA	0.09/0.91	0.19/0.82	0.18/0.83		752/0.62	589/0.51	397/0.52	471/0.5	712/0.58	757/0.63	595/0.51	693/0.57	658/0.54	691/0.57	648/0.54
Trimap	0.07/0.93	0.09/0.91	0.26/0.76	0.07/0.93		679/0.56	725/0.6	784/0.66	657/0.54	592/0.51	743/0.61	685/0.56	672/0.55	649/0.54	812/0.69
mTSNE	0.07/0.93	0.28/0.74	0.27/0.75	0.08/0.92	0.08/0.92		697/0.57	741/0.61	784/0.66	716/0.59	794/0.67	768/0.64	713/0.59	701/0.58	682/0.56
Isomap	0.11/0.89	0.17/0.84	0.14/0.87	0.19/0.82	0.06/0.94	0.14/0.87		471/0.5	592/0.51	746/0.62	387/0.52	795/0.67	647/0.54	628/0.53	576/0.51
KPCA	0.14/0.87	0.11/0.89	0.16/0.85	0.28/0.74	0.07/0.93	0.17/0.84	0.27/0.75		914/0.84	714/0.59	367/0.53	628/0.53	597/0.51	577/0.51	542/0.5
LEM	0.05/0.95	0.02/0.98	0.07/0.93	0.04/0.96	0.01/0.99	0.05/0.95	0.06/0.94	0.08/0.92		539/0.5	780/0.65	487/0.49	560/0.5	573/0.51	589/0.51
LTSA	0.08/0.92	0.06/0.94	0.09/0.91	0.06/0.94	0.06/0.94	0.09/0.91	0.08/0.92	0.05/0.95	0.21/0.8		795/0.67	667/0.55	528/0.5	667/0.55	576/0.51
MDS	0.13/0.88	0.04/0.96	0.11/0.89	0.16/0.85	0.03/0.97	0.07/0.93	0.19/0.82	0.02/0.98	0.01/0.99	0.08/0.92		914/0.84	967/0.93	843/0.73	855/0.75
HLLE	0.05/0.95	0.08/0.92	0.05/0.95	0.09/0.91	0.09/0.91	0.18/0.83	0.09/0.91	0.06/0.94	0.13/0.88	0.19/0.82	0.02/0.98		348/0.54	675/0.56	472/0.5
LLE	0.06/0.94	0.08/0.92	0.06/0.94	0.04/0.96	0.08/0.92	0.06/0.94	0.08/0.92	0.02/0.98	0.15/0.86	0.17/0.84	0.18/0.83	0.28/0.74		597/0.51	575/0.51
LVis	0.09/0.91	0.07/0.93	0.08/0.92	0.05/0.95	0.07/0.93	0.15/0.86	0.01/0.99	0.03/0.97	0.09/0.91	0.24/0.77	0.17/0.84	0.19/0.82	0.18/0.83		648/0.54
MVU	0.17/0.84	0.06/0.94	0.09/0.91	0.02/0.98	0.06/0.94	0.19/0.82	0.06/0.94	0.09/0.91	0.1/0.9	0.16/0.85	0.21/0.8	0.21/0.8	0.23/0.78	0.17/0.84	

(b) Metric 2: Spearman rank correlation (ρ_s)

	UMAP	t-SNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		578/0.51	649/0.54	745/0.61	812/0.69	647/0.54	875/0.78	694/0.57	947/0.89	912/0.83	793/0.67	891/0.8	843/0.73	877/0.78	845/0.73
t-SNE	0.21/0.8		375/0.53	485/0.49	596/0.51	374/0.53	729/0.6	671/0.55	725/0.6	784/0.66	618/0.52	799/0.67	738/0.61	801/0.68	844/0.73
FIt_SNE	0.18/0.83	0.23/0.78		504/0.49	675/0.56	317/0.56	791/0.66	602/0.52	836/0.72	877/0.78	745/0.61	862/0.76	874/0.77	906/0.82	948/0.9
PCA	0.15/0.86	0.18/0.83	0.21/0.8		643/0.54	588/0.51	576/0.51	297/0.58	755/0.62	719/0.59	511/0.49	728/0.6	719/0.59	768/0.64	743/0.61
Trimap	0.09/0.91	0.11/0.89	0.18/0.83	0.21/0.8		539/0.5	277/0.59	633/0.53	687/0.56	643/0.54	313/0.56	624/0.53	614/0.52	679/0.56	642/0.53
mTSNE	0.17/0.84	0.24/0.77	0.24/0.77	0.18/0.83	0.16/0.85		775/0.65	506/0.49	836/0.72	843/0.73	729/0.6	869/0.77	884/0.79	872/0.77	901/0.82
Isomap	0.08/0.92	0.17/0.84	0.15/0.86	0.19/0.82	0.24/0.77	0.15/0.86		439/0.5	641/0.53	624/0.53	239/0.63	579/0.51	547/0.5	598/0.51	546/0.5
KPCA	0.14/0.87	0.16/0.85	0.19/0.82	0.26/0.76	0.18/0.83	0.19/0.82	0.26/0.76		682/0.56	679/0.56	274/0.6	732/0.6	742/0.61	746/0.62	711/0.58
LEM	0.05/0.95	0.05/0.95	0.09/0.91	0.09/0.91	0.08/0.92	0.09/0.91	0.19/0.82	0.18/0.83		314/0.56	517/0.5	328/0.55	241/0.63	216/0.66	208/0.67
LTSA	0.07/0.93	0.03/0.97	0.07/0.93	0.11/0.89	0.09/0.91	0.1/0.9	0.21/0.8	0.19/0.82	0.24/0.77		544/0.5	326/0.56	384/0.52	319/0.56	312/0.57
MDS	0.09/0.91	0.11/0.89	0.17/0.84	0.21/0.8	0.18/0.83	0.08/0.92	0.29/0.73	0.26/0.76	0.15/0.86	0.18/0.83		548/0.5	562/0.5	514/0.49	546/0.5
HLLE	0.03/0.97	0.04/0.96	0.06/0.94	0.07/0.93	0.07/0.93	0.07/0.93	0.19/0.82	0.19/0.82	0.26/0.76	0.26/0.76	0.18/0.83		284/0.59	258/0.61	319/0.56
LLE	0.05/0.95	0.03/0.97	0.04/0.96	0.08/0.92	0.06/0.94	0.06/0.94	0.18/0.83	0.21/0.8	0.21/0.8	0.27/0.75	0.14/0.87	0.21/0.8		391/0.52	401/0.51
LVis	0.07/0.93	0.05/0.95	0.05/0.95	0.1/0.9	0.08/0.92	0.08/0.92	0.21/0.8	0.17/0.84	0.26/0.76	0.21/0.8	0.16/0.85	0.26/0.76	0.24/0.77		352/0.54
MVU	0.06/0.94	0.06/0.94	0.06/0.94	0.06/0.94	0.04/0.96	0.05/0.95	0.24/0.77	0.16/0.85	0.24/0.77	0.25/0.77	0.14/0.87	0.23/0.78	0.28/0.74	0.19/0.82	

(c) Metric 3: Mean K-ary neighborhood agreement ($\mu_{R_{nX}}$)

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		426/0.51	518/0.5	549/0.5	863/0.76	549/0.5	316/0.56	625/0.53	816/0.69	849/0.74	743/0.61	846/0.73	852/0.74	719/0.59	892/0.8
tSNE	0.21/0.8		523/0.5	579/0.51	839/0.72	571/0.5	269/0.6	647/0.54	894/0.81	794/0.67	746/0.62	875/0.78	823/0.7	901/0.82	871/0.77
FIt_SNE	0.18/0.83	0.24/0.77		364/0.53	713/0.59	275/0.6	574/0.51	379/0.52	974/0.94	874/0.77	369/0.53	867/0.76	812/0.69	825/0.71	879/0.78
PCA	0.17/0.84	0.21/0.8	0.28/0.74		745/0.61	349/0.54	582/0.51	269/0.6	781/0.65	746/0.62	529/0.5	748/0.62	713/0.59	749/0.62	785/0.66
Trimap	0.08/0.92	0.08/0.92	0.14/0.87	0.11/0.89		872/0.77	874/0.77	719/0.59	369/0.53	347/0.54	617/0.52	319/0.56	369/0.53	482/0.5	316/0.56
mTSNE	0.18/0.83	0.22/0.79	0.34/0.68	0.28/0.74	0.05/0.95		492/0.49	318/0.56	675/0.56	598/0.51	588/0.51	762/0.63	782/0.65	792/0.67	810/0.69
Isomap	0.25/0.77	0.35/0.67	0.19/0.82	0.19/0.82	0.05/0.95	0.24/0.77		471/0.5	795/0.67	748/0.62	694/0.57	947/0.89	975/0.95	826/0.71	819/0.7
KPCA	0.27/0.75	0.18/0.83	0.21/0.8	0.36/0.66	0.11/0.89	0.21/0.8	0.21/0.8		759/0.63	617/0.52	639/0.53	862/0.76	874/0.77	659/0.55	747/0.62
LEM	0.07/0.93	0.12/0.88	0.02/0.98	0.09/0.91	0.24/0.77	0.18/0.83	0.08/0.92	0.18/0.83		349/0.54	317/0.56	547/0.5	514/0.49	429/0.5	547/0.5
LTSA	0.05/0.95	0.16/0.85	0.08/0.92	0.08/0.92	0.25/0.77	0.15/0.86	0.09/0.91	0.21/0.8	0.22/0.79		369/0.53	482/0.5	496/0.49	348/0.54	527/0.5
MDS	0.09/0.91	0.17/0.84	0.24/0.77	0.11/0.89	0.18/0.83	0.19/0.82	0.12/0.88	0.2/0.81	0.21/0.8	0.24/0.77		462/0.5	501/0.49	395/0.52	522/0.5
HLLE	0.07/0.93	0.07/0.93	0.09/0.91	0.19/0.82	0.24/0.77	0.17/0.84	0.04/0.96	0.13/0.88	0.18/0.83	0.19/0.82	0.18/0.83		314/0.56	544/0.5	406/0.51
LLE	0.06/0.94	0.09/0.91	0.08/0.92	0.08/0.92	0.23/0.78	0.11/0.89	0.03/0.97	0.12/0.88	0.17/0.84	0.18/0.83	0.14/0.87	0.21/0.8		492/0.49	529/0.5
LVis	0.12/0.88	0.05/0.95	0.09/0.91	0.12/0.88	0.21/0.8	0.11/0.89	0.07/0.93	0.19/0.82	0.2/0.81	0.27/0.75	0.21/0.8	0.18/0.83	0.18/0.83		523/0.5
MVU	0.07/0.93	0.07/0.93	0.07/0.93	0.11/0.89	0.23/0.78	0.09/0.91	0.08/0.92	0.18/0.83	0.18/0.83	0.15/0.86	0.17/0.84	0.19/0.82	0.17/0.84	0.17/0.84	

(d) Metric 4: Local quality criteria (Q_{local})

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		316/0.56	489/0.49	578/0.51	846/0.73	587/0.51	347/0.54	542/0.5	874/0.77	843/0.73	643/0.54	629/0.53	593/0.51	766/0.64	719/0.59
tSNE	0.26/0.76		488/0.49	589/0.51	816/0.69	655/0.54	398/0.52	629/0.53	749/0.62	762/0.63	622/0.52	647/0.54	744/0.61	718/0.59	743/0.61
FIt_SNE	0.22/0.79	0.22/0.79		547/0.5	699/0.57	366/0.53	589/0.51	496/0.49	682/0.56	647/0.54	485/0.49	719/0.59	426/0.51	659/0.55	643/0.54
PCA	0.18/0.83	0.19/0.82	0.17/0.84		633/0.53	602/0.52	522/0.5	392/0.52	655/0.54	626/0.53	418/0.51	492/0.49	466/0.5	651/0.54	756/0.63
Trimap	0.09/0.91	0.09/0.91	0.15/0.86	0.16/0.85		596/0.51	792/0.67	644/0.54	492/0.49	503/0.49	671/0.55	618/0.52	598/0.51	397/0.52	397/0.52
mTSNE	0.16/0.85	0.17/0.84	0.24/0.77	0.16/0.85	0.18/0.83		509/0.49	492/0.49	526/0.5	579/0.51	452/0.5	462/0.5	583/0.51	601/0.51	623/0.52
Isomap	0.26/0.76	0.27/0.75	0.16/0.85	0.18/0.83	0.14/0.87	0.19/0.82		563/0.5	756/0.63	756/0.63	613/0.52	649/0.54	677/0.56	712/0.58	719/0.59
KPCA	0.17/0.84	0.15/0.86	0.21/0.8	0.22/0.79	0.17/0.84	0.22/0.79	0.19/0.82		588/0.51	598/0.51	267/0.6	492/0.49	584/0.51	577/0.51	628/0.53
LEM	0.11/0.89	0.13/0.88	0.15/0.86	0.16/0.85	0.22/0.79	0.19/0.82	0.15/0.86	0.21/0.8		564/0.5	411/0.51	655/0.54	519/0.5	629/0.53	614/0.52
LTSA	0.11/0.89	0.13/0.88	0.16/0.85	0.16/0.85	0.2/0.81	0.2/0.81	0.15/0.86	0.21/0.8	0.19/0.82		576/0.51	599/0.51	546/0.5	512/0.49	501/0.49
MDS	0.13/0.88	0.15/0.86	0.22/0.79	0.2/0.81	0.17/0.84	0.23/0.78	0.18/0.83	0.29/0.73	0.22/0.79	0.21/0.8		419/0.51	456/0.5	581/0.51	532/0.5
HLLE	0.14/0.87	0.14/0.87	0.14/0.87	0.19/0.82	0.17/0.84	0.23/0.78	0.18/0.83	0.24/0.77	0.18/0.83	0.2/0.81	0.28/0.74		328/0.55	493/0.49	574/0.51
LLE	0.18/0.83	0.13/0.88	0.2/0.81	0.19/0.82	0.18/0.83	0.18/0.83	0.18/0.83	0.21/0.8	0.19/0.82	0.22/0.79	0.28/0.74	0.31/0.71		549/0.5	532/0.5
LVis	0.15/0.86	0.13/0.88	0.18/0.83	0.18/0.83	0.24/0.77	0.17/0.84	0.16/0.85	0.21/0.8	0.17/0.84	0.22/0.79	0.22/0.79	0.25/0.77	0.19/0.82		492/0.49
MVU	0.15/0.86	0.12/0.88	0.17/0.84	0.15/0.86	0.24/0.77	0.17/0.84	0.16/0.85	0.19/0.82	0.18/0.83	0.23/0.78	0.25/0.77	0.21/0.8	0.19/0.82	0.22/0.79	

(e) Metric 5: k_{max} neighborhood loss ($\lambda_{K_{max}}$)

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU

UMAP		319/0.56	490/0.49	549/0.5	724/0.59	522/0.5	346/0.54	293/0.58	697/0.57	853/0.74	562/0.5	719/0.59	758/0.63	901/0.82	896/0.81
tSNE	0.28/0.74		496/0.49	557/0.5	719/0.59	526/0.5	324/0.56	367/0.53	716/0.59	882/0.79	527/0.5	765/0.64	693/0.57	924/0.85	893/0.8
FIt_SNE	0.24/0.77	0.26/0.76		386/0.52	627/0.53	419/0.51	564/0.5	543/0.5	635/0.53	716/0.59	346/0.54	675/0.56	632/0.53	743/0.61	719/0.59
PCA	0.21/0.8	0.24/0.77	0.28/0.74		328/0.55	405/0.51	517/0.5	523/0.5	628/0.53	719/0.59	322/0.56	617/0.52	649/0.54	749/0.62	716/0.59
Trimap	0.18/0.83	0.21/0.8	0.19/0.82	0.28/0.74		549/0.5	744/0.61	695/0.57	293/0.58	517/0.5	543/0.5	347/0.54	366/0.53	549/0.5	573/0.51
mTSNE	0.21/0.8	0.24/0.77	0.25/0.77	0.26/0.76	0.19/0.82		571/0.5	562/0.5	649/0.54	718/0.59	328/0.55	614/0.52	647/0.54	801/0.68	792/0.67
Isomap	0.28/0.74	0.32/0.7	0.21/0.8	0.21/0.8	0.16/0.85	0.18/0.83		283/0.59	624/0.53	963/0.92	581/0.51	716/0.59	697/0.57	839/0.72	906/0.82
KPCA	0.29/0.73	0.29/0.73	0.21/0.8	0.21/0.8	0.18/0.83	0.19/0.82	0.32/0.7		792/0.67	953/0.91	558/0.5	762/0.63	719/0.59	903/0.82	893/0.8
LEM	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.29/0.73	0.17/0.84	0.17/0.84	0.14/0.87		583/0.51	703/0.58	318/0.56	364/0.53	583/0.51	549/0.5
LTSA	0.11/0.89	0.16/0.85	0.17/0.84	0.15/0.86	0.18/0.83	0.15/0.86	0.08/0.92	0.07/0.93	0.18/0.83		749/0.62	536/0.5	416/0.51	317/0.56	423/0.51
MDS	0.21/0.8	0.24/0.77	0.29/0.73	0.29/0.73	0.18/0.83	0.28/0.74	0.19/0.82	0.19/0.82	0.15/0.86	0.15/0.86		638/0.53	649/0.54	853/0.74	847/0.74
HLLE	0.18/0.83	0.15/0.86	0.18/0.83	0.25/0.77	0.28/0.74	0.17/0.84	0.16/0.85	0.14/0.87	0.27/0.75	0.19/0.82	0.19/0.82		273/0.6	516/0.5	547/0.5
LLE	0.17/0.84	0.18/0.83	0.18/0.83	0.24/0.77	0.28/0.74	0.16/0.85	0.17/0.84	0.14/0.87	0.26/0.76	0.25/0.77	0.19/0.82	0.29/0.73		543/0.5	593/0.51
LVis	0.07/0.93	0.09/0.91	0.15/0.86	0.18/0.83	0.18/0.83	0.14/0.87	0.14/0.87	0.08/0.92	0.18/0.83	0.29/0.73	0.15/0.86	0.8/0.26	0.19/0.82		493/0.49
MVU	0.09/0.91	0.1/0.9	0.15/0.86	0.17/0.84	0.18/0.83	0.15/0.86	0.09/0.91	0.1/0.9	0.18/0.83	0.25/0.77	0.15/0.86	0.18/0.83	0.18/0.83	0.21/0.8	

(f) Metric 6: Global quality criteria (Q_{global})

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		364/0.53	419/0.51	536/0.5	614/0.52	349/0.54	397/0.52	514/0.49	896/0.81	843/0.73	567/0.5	694/0.57	679/0.56	669/0.55	849/0.74
tSNE	0.29/0.73		318/0.56	543/0.5	628/0.53	379/0.52	319/0.56	544/0.5	793/0.67	846/0.73	542/0.5	675/0.56	694/0.57	617/0.52	849/0.74
FIt_SNE	0.26/0.76	0.28/0.74		346/0.54	644/0.54	319/0.56	347/0.54	542/0.5	846/0.73	895/0.81	573/0.51	642/0.53	681/0.56	677/0.56	859/0.75
PCA	0.22/0.79	0.19/0.82	0.29/0.73		659/0.55	543/0.5	519/0.5	296/0.58	843/0.73	867/0.76	342/0.54	655/0.54	682/0.56	619/0.52	852/0.74
Trimap	0.18/0.83	0.18/0.83	0.17/0.84	0.18/0.83		572/0.5	549/0.5	716/0.59	843/0.73	895/0.81	744/0.61	315/0.56	328/0.55	349/0.54	684/0.56
mTSNE	0.28/0.74	0.28/0.74	0.31/0.71	0.21/0.8	0.19/0.82		294/0.58	512/0.49	746/0.62	782/0.65	546/0.5	595/0.51	544/0.5	513/0.49	716/0.59
Isomap	0.27/0.75	0.29/0.73	0.29/0.73	0.21/0.8	0.19/0.82	0.29/0.73		549/0.5	743/0.61	752/0.62	519/0.5	543/0.5	547/0.5	533/0.5	746/0.62
KPCA	0.21/0.8	0.19/0.82	0.22/0.79	0.29/0.73	0.17/0.84	0.22/0.79	0.21/0.8		856/0.75	813/0.69	317/0.56	655/0.54	649/0.54	628/0.53	849/0.74
LEM	0.11/0.89	0.15/0.86	0.11/0.89	0.11/0.89	0.12/0.88	0.18/0.83	0.16/0.85	0.08/0.92		276/0.59	843/0.73	512/0.49	562/0.5	547/0.5	317/0.56
LTSA	0.12/0.88	0.12/0.88	0.1/0.9	0.1/0.9	0.09/0.91	0.16/0.85	0.18/0.83	0.09/0.91	0.29/0.73		316/0.56	549/0.5	576/0.51	544/0.5	269/0.6
MDS	0.21/0.8	0.18/0.83	0.21/0.8	0.28/0.74	0.14/0.87	0.22/0.79	0.2/0.81	0.29/0.73	0.09/0.91	0.28/0.74		719/0.59	695/0.57	673/0.55	845/0.73
HLLE	0.18/0.83	0.16/0.85	0.16/0.85	0.18/0.83	0.28/0.74	0.2/0.81	0.2/0.81	0.18/0.83	0.19/0.82	0.22/0.79	0.16/0.85		245/0.62	748/0.62	556/0.5
LLE	0.18/0.83	0.15/0.86	0.15/0.86	0.18/0.83	0.26/0.76	0.22/0.79	0.19/0.82	0.18/0.83	0.18/0.83	0.21/0.8	0.19/0.82	0.28/0.74		362/0.53	576/0.51
LVis	0.18/0.83	0.16/0.85	0.15/0.86	0.19/0.82	0.26/0.76	0.23/0.78	0.18/0.83	0.19/0.82	0.19/0.82	0.22/0.79	0.19/0.82	0.15/0.86	0.29/0.73		549/0.5
MVU	0.12/0.88	0.11/0.89	0.11/0.89	0.12/0.88	0.18/0.83	0.18/0.83	0.17/0.84	0.09/0.91	0.28/0.74	0.31/0.71	0.12/0.88	0.19/0.82	0.22/0.79	0.21/0.8	

(g) Metric 7: Area under the R_{nX} curve ($AUC_{\ln K}(R_{nX}(K))$)

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		516/0.5	538/0.5	647/0.54	816/0.69	355/0.54	546/0.5	602/0.52	803/0.68	849/0.74	346/0.54	328/0.55	673/0.55	645/0.54	823/0.7
tSNE	0.19/0.82		294/0.58	683/0.56	859/0.75	546/0.5	344/0.54	618/0.52	846/0.73	855/0.75	548/0.5	512/0.49	672/0.55	668/0.55	843/0.73
FIt_SNE	0.18/0.83	0.32/0.7		649/0.54	843/0.73	577/0.51	316/0.56	695/0.57	829/0.71	846/0.73	554/0.5	588/0.51	643/0.54	647/0.54	825/0.71
PCA	0.16/0.85	0.16/0.85	0.16/0.85		634/0.53	579/0.51	798/0.67	257/0.61	644/0.54	675/0.56	528/0.5	594/0.51	347/0.54	328/0.55	625/0.53
Trimap	0.11/0.89	0.11/0.89	0.11/0.89	0.15/0.86		753/0.62	942/0.89	501/0.49	334/0.55	318/0.56	725/0.6	795/0.67	554/0.5	542/0.5	365/0.53
mTSNE	0.27/0.75	0.21/0.8	0.21/0.8	0.19/0.82	0.13/0.88		529/0.5	647/0.54	758/0.63	716/0.59	314/0.56	367/0.53	679/0.56	648/0.54	718/0.59
Isomap	0.18/0.83	0.27/0.75	0.27/0.75	0.13/0.88	0.07/0.93	0.21/0.8		758/0.63	942/0.89	947/0.89	514/0.49	584/0.51	732/0.6	736/0.61	948/0.9
KPCA	0.16/0.85	0.16/0.85	0.16/0.85	0.31/0.71	0.21/0.8	0.16/0.85	0.13/0.88		463/0.5	475/0.5	519/0.5	578/0.51	349/0.54	377/0.52	496/0.49
LEM	0.12/0.88	0.12/0.88	0.11/0.89	0.16/0.85	0.28/0.74	0.13/0.88	0.07/0.93	0.24/0.77		354/0.54	716/0.59	795/0.67	522/0.5	549/0.5	347/0.54
LTSA	0.1/0.9	0.12/0.88	0.11/0.89	0.16/0.85	0.28/0.74	0.14/0.87	0.07/0.93	0.24/0.77	0.29/0.73		728/0.6	762/0.63	549/0.5	563/0.5	374/0.53
MDS	0.27/0.75	0.21/0.8	0.19/0.82	0.19/0.82	0.13/0.88	0.27/0.75	0.21/0.8	0.21/0.8	0.14/0.87	0.14/0.87		355/0.54	574/0.51	549/0.5	786/0.66
HLLE	0.27/0.75	0.16/0.85	0.19/0.82	0.19/0.82	0.12/0.88	0.27/0.75	0.2/0.81	0.2/0.81	0.14/0.87	0.14/0.87	0.28/0.74		578/0.51	529/0.5	764/0.63
LLE	0.16/0.85	0.16/0.85	0.16/0.85	0.27/0.75	0.2/0.81	0.16/0.85	0.14/0.87	0.29/0.73	0.21/0.8	0.21/0.8	0.21/0.8	0.21/0.8		375/0.53	596/0.51
LVis	0.16/0.85	0.16/0.85	0.16/0.85	0.27/0.75	0.2/0.81	0.16/0.85	0.14/0.87	0.28/0.74	0.21/0.8	0.21/0.8	0.21/0.8	0.21/0.8	0.27/0.75		558/0.5
MVU	0.1/0.9	0.12/0.88	0.12/0.88	0.16/0.85	0.27/0.75	0.13/0.88	0.06/0.94	0.24/0.77	0.29/0.73	0.27/0.75	0.14/0.87	0.14/0.87	0.19/0.82	0.21/0.8	

(h) Metric 8: KNN prediction accuracy (ACC_{ψ})

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		548/0.5	346/0.54	547/0.5	918/0.84	315/0.56	907/0.83	347/0.54	768/0.64	718/0.59	317/0.56	745/0.61	764/0.63	342/0.54	749/0.62
tSNE	0.19/0.82		549/0.5	314/0.56	916/0.84	546/0.5	932/0.87	562/0.5	746/0.62	754/0.62	549/0.5	798/0.67	746/0.62	532/0.5	719/0.59
FIt_SNE	0.27/0.75	0.19/0.82		519/0.5	765/0.64	316/0.56	795/0.67	378/0.52	627/0.53	655/0.54	347/0.54	625/0.53	695/0.57	375/0.53	627/0.53
PCA	0.19/0.82	0.28/0.74	0.21/0.8		803/0.68	548/0.5	862/0.76	579/0.51	628/0.53	645/0.54	528/0.5	647/0.54	653/0.54	512/0.49	667/0.55
Trimap	0.09/0.91	0.09/0.91	0.14/0.87	0.11/0.89		734/0.6	346/0.54	785/0.66	564/0.5	543/0.5	759/0.63	541/0.5	568/0.5	749/0.62	562/0.5
mTSNE	0.29/0.73	0.27/0.75	0.27/0.75	0.21/0.8	0.14/0.87		813/0.69	319/0.56	618/0.52	627/0.53	327/0.55	659/0.55	634/0.53	322/0.56	375/0.53
Isomap	0.09/0.91	0.08/0.92	0.13/0.88	0.19/0.82	0.27/0.75	0.09/0.91		635/0.53	597/0.51	584/0.51	674/0.56	543/0.5	552/0.5	627/0.53	528/0.5
KPCA	0.28/0.74	0.27/0.75	0.27/0.75	0.21/0.8	0.14/0.87	0.21/0.8	0.15/0.86		594/0.51	563/0.5	357/0.54	597/0.51	542/0.5	367/0.53	528/0.5
LEM	0.14/0.87	0.14/0.87	0.15/0.86	0.18/0.83	0.21/0.8	0.18/0.83	0.21/0.8	0.21/0.8		345/0.54	529/0.5	368/0.53	321/0.56	509/0.49	378/0.52
LTSA	0.14/0.87	0.14/0.87	0.15/0.86	0.18/0.83	0.21/0.8	0.18/0.83	0.21/0.8	0.21/0.8	0.28/0.74		355/0.54	367/0.53	326/0.56	526/0.5	578/0.51
MDS	0.27/0.75	0.21/0.8	0.27/0.75	0.21/0.8	0.14/0.87	0.27/0.75	0.16/0.85	0.27/0.75	0.21/0.8	0.27/0.75		536/0.5	519/0.5	375/0.53	594/0.51
HLLE	0.14/0.87	0.14/0.87	0.15/0.86	0.18/0.83	0.21/0.8	0.18/0.83	0.21/0.8	0.21/0.8	0.28/0.74	0.27/0.75	0.21/0.8		378/0.52	547/0.5	369/0.53
LLE	0.15/0.86	0.14/0.87	0.15/0.86	0.18/0.83	0.21/0.8	0.18/0.83	0.21/0.8	0.21/0.8	0.28/0.74	0.27/0.75	0.21/0.8	0.28/0.74		594/0.51	384/0.52
LVis	0.27/0.75	0.21/0.8	0.27/0.75	0.21/0.8	0.14/0.87	0.27/0.75	0.16/0.85	0.27/0.75	0.21/0.8	0.21/0.8	0.27/0.75	0.21/0.8	0.21/0.8		512/0.49
MVU	0.14/0.87	0.14/0.87	0.15/0.86	0.18/0.83	0.21/0.8	0.27/0.75	0.21/0.8	0.16/0.85	0.28/0.74	0.21/0.8	0.21/0.8	0.28/0.74	0.27/0.75	0.21/0.8	

(i) Metric 9: Normalized Mutual Information (nMI)

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		346/0.54	294/0.58	647/0.54	901/0.82	342/0.54	803/0.68	649/0.54	647/0.54	634/0.53	867/0.76	366/0.53	349/0.54	679/0.56	896/0.81
tSNE	0.27/0.75		349/0.54	627/0.53	872/0.77	347/0.54	876/0.78	624/0.53	679/0.56	637/0.53	895/0.81	394/0.52	385/0.52	628/0.53	974/0.94
FIt_SNE	0.31/0.71	0.28/0.74		634/0.53	875/0.78	395/0.52	874/0.77	658/0.54	694/0.57	628/0.53	874/0.77	324/0.56	397/0.52	652/0.54	875/0.78
PCA	0.18/0.83	0.18/0.83	0.12/0.88		895/0.81	604/0.52	933/0.87	375/0.53	369/0.53	319/0.56	875/0.78	605/0.52	688/0.57	375/0.53	874/0.77
Trimap	0.09/0.91	0.12/0.88	0.11/0.89	0.11/0.89		875/0.78	362/0.53	685/0.56	694/0.57	587/0.51	365/0.53	865/0.76	874/0.77	625/0.53	374/0.53
mTSNE	0.27/0.75	0.27/0.75	0.27/0.75	0.18/0.83	0.11/0.89		895/0.81	627/0.53	639/0.53	745/0.61	855/0.75	362/0.53	974/0.94	629/0.53	874/0.77
Isomap	0.11/0.89	0.12/0.88	0.11/0.89	0.09/0.91	0.27/0.75	0.12/0.88		648/0.54	601/0.51	687/0.56	357/0.54	719/0.59	874/0.77	624/0.53	384/0.52
KPCA	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75	0.18/0.83	0.18/0.83	0.18/0.83		312/0.57	355/0.54	645/0.54	542/0.5	584/0.51	379/0.52	619/0.52
LEM	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75		374/0.53	712/0.58	624/0.53	684/0.56	384/0.52	794/0.67
LTSA	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75	0.21/0.8	0.14/0.87	0.18/0.83	0.27/0.75	0.27/0.75		365/0.53	628/0.53	697/0.57	347/0.54	718/0.59

MDS	0.12/0.88	0.12/0.88	0.11/0.89	0.11/0.89	0.27/0.75	0.11/0.89	0.27/0.75	0.18/0.83	0.15/0.86	0.27/0.75		847/0.74	876/0.78	624/0.53	394/0.52
HLLE	0.27/0.75	0.27/0.75	0.27/0.75	0.18/0.83	0.12/0.88	0.27/0.75	0.15/0.86	0.21/0.8	0.18/0.83	0.18/0.83	0.11/0.89		971/0.94	684/0.56	354/0.54
LLE	0.27/0.75	0.26/0.76	0.126/0.88	0.18/0.83	0.11/0.89	0.08/0.92	0.11/0.89	0.21/0.8	0.18/0.83	0.18/0.83	0.11/0.89	0.08/0.92		384/0.52	524/0.5
LVis	0.18/0.83	0.16/0.85	0.18/0.83	0.27/0.75	0.18/0.83	0.15/0.86	0.18/0.83	0.27/0.75	0.27/0.75	0.27/0.75	0.18/0.83	0.17/0.84	0.27/0.75		587/0.51
MVU	0.12/0.88	0.04/0.96	0.12/0.88	0.12/0.88	0.27/0.75	0.11/0.89	0.27/0.75	0.18/0.83	0.15/0.86	0.15/0.86	0.27/0.75	0.27/0.75	0.21/0.8	0.21/0.8	

(j) Metric 10: Structural Similarity Index (SSI)

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		365/0.53	675/0.56	349/0.54	846/0.73	847/0.74	347/0.54	342/0.54	849/0.74	314/0.56	379/0.52	875/0.78	825/0.71	645/0.54	647/0.54
tSNE	0.27/0.75		374/0.53	647/0.54	716/0.59	784/0.66	624/0.53	682/0.56	713/0.59	685/0.56	622/0.52	749/0.62	752/0.62	349/0.54	347/0.54
FIt_SNE	0.18/0.83	0.27/0.75		659/0.55	746/0.62	723/0.59	674/0.56	628/0.53	791/0.66	527/0.5	647/0.54	342/0.54	743/0.61	369/0.53	374/0.53
PCA	0.27/0.75	0.18/0.83	0.18/0.83		853/0.74	845/0.73	325/0.56	301/0.57	906/0.82	374/0.53	369/0.53	904/0.82	826/0.71	674/0.56	685/0.56
Trimap	0.12/0.88	0.15/0.86	0.16/0.85	0.12/0.88		352/0.54	862/0.76	874/0.77	348/0.54	957/0.91	934/0.87	345/0.54	385/0.52	601/0.51	687/0.56
mTSNE	0.12/0.88	0.15/0.86	0.16/0.85	0.12/0.88	0.27/0.75		874/0.77	852/0.74	345/0.54	862/0.76	874/0.77	325/0.56	362/0.53	647/0.54	697/0.57
Isomap	0.27/0.75	0.18/0.83	0.18/0.83	0.27/0.75	0.12/0.88	0.12/0.88		352/0.54	862/0.76	347/0.54	369/0.53	895/0.81	865/0.76	663/0.55	647/0.54
KPCA	0.27/0.75	0.18/0.83	0.18/0.83	0.28/0.74	0.12/0.88	0.12/0.88	0.27/0.75		374/0.53	385/0.52	375/0.53	895/0.81	924/0.85	685/0.56	612/0.52
LEM	0.12/0.88	0.15/0.86	0.15/0.86	0.09/0.91	0.27/0.75	0.27/0.75	0.11/0.89	0.27/0.75		895/0.81	826/0.71	328/0.55	316/0.56	627/0.53	698/0.57
LTSA	0.27/0.75	0.18/0.83	0.21/0.8	0.27/0.75	0.09/0.91	0.12/0.88	0.27/0.75	0.27/0.75	0.1/0.9		385/0.52	825/0.71	875/0.78	647/0.54	632/0.53
MDS	0.27/0.75	0.18/0.83	0.18/0.83	0.27/0.75	0.09/0.91	0.11/0.89	0.27/0.75	0.27/0.75	0.12/0.88	0.27/0.75		952/0.9	794/0.67	622/0.52	672/0.55
HLLE	0.11/0.89	0.15/0.86	0.27/0.75	0.08/0.92	0.27/0.75	0.27/0.75	0.1/0.9	0.12/0.88	0.27/0.75	0.12/0.88	0.03/0.97		219/0.65	524/0.5	627/0.53
LLE	0.12/0.88	0.14/0.87	0.15/0.86	0.12/0.88	0.27/0.75	0.27/0.75	0.12/0.88	0.08/0.92	0.27/0.75	0.11/0.89	0.13/0.88	0.31/0.71		629/0.53	674/0.56
LVis	0.18/0.83	0.28/0.74	0.27/0.75	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.21/0.8	0.18/0.83		306/0.57
MVU	0.17/0.84	0.27/0.75	0.27/0.75	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75	

(k) Metric 11: Logarithmic loss of multi-class classification

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		648/0.54	325/0.56	942/0.89	645/0.54	328/0.55	901/0.82	845/0.73	852/0.74	693/0.57	855/0.75	607/0.52	746/0.62	863/0.76	624/0.53
tSNE	0.18/0.83		685/0.56	832/0.72	425/0.51	647/0.54	895/0.81	852/0.74	846/0.73	473/0.5	862/0.76	419/0.51	425/0.51	875/0.78	425/0.51
FIt_SNE	0.27/0.75	0.18/0.83		712/0.58	582/0.51	349/0.54	795/0.67	746/0.62	714/0.59	521/0.5	785/0.66	549/0.5	513/0.49	744/0.61	574/0.51
PCA	0.08/0.92	0.12/0.88	0.17/0.84		524/0.5	714/0.59	325/0.56	374/0.53	418/0.51	621/0.52	376/0.53	649/0.54	627/0.53	417/0.51	529/0.5
Trimap	0.18/0.83	0.26/0.76	0.21/0.8	0.21/0.8		743/0.61	526/0.5	547/0.5	599/0.51	374/0.53	549/0.5	379/0.52	395/0.52	578/0.51	378/0.52
mTSNE	0.28/0.74	0.18/0.83	0.27/0.75	0.17/0.84	0.17/0.84		875/0.78	795/0.67	716/0.59	529/0.5	846/0.73	521/0.5	596/0.51	785/0.66	588/0.51
Isomap	0.09/0.91	0.12/0.88	0.16/0.85	0.27/0.75	0.21/0.8	0.12/0.88		364/0.53	355/0.54	576/0.51	394/0.52	562/0.5	687/0.56	317/0.56	574/0.51
KPCA	0.12/0.88	0.12/0.88	0.17/0.84	0.27/0.75	0.21/0.8	0.14/0.87	0.27/0.75		319/0.56	649/0.54	347/0.54	576/0.51	512/0.49	368/0.53	541/0.5
LEM	0.12/0.88	0.12/0.88	0.18/0.83	0.24/0.77	0.2/0.81	0.15/0.86	0.27/0.75	0.27/0.75		294/0.58	358/0.53	599/0.51	651/0.54	348/0.54	579/0.51
LTSA	0.18/0.83	0.24/0.77	0.21/0.8	0.18/0.83	0.27/0.75	0.21/0.8	0.21/0.8	0.18/0.83	0.31/0.71		648/0.54	325/0.56	366/0.53	679/0.56	322/0.56
MDS	0.12/0.88	0.12/0.88	0.18/0.83	0.27/0.75	0.21/0.8	0.12/0.88	0.27/0.75	0.27/0.75	0.27/0.75	0.18/0.83		677/0.56	648/0.54	395/0.52	647/0.54
HLLE	0.18/0.83	0.24/0.77	0.21/0.8	0.18/0.83	0.27/0.75	0.21/0.8	0.21/0.8	0.21/0.8	0.2/0.81	0.27/0.75	0.18/0.83		201/0.67	677/0.56	345/0.54
LLE	0.14/0.87	0.24/0.77	0.21/0.8	0.18/0.83	0.26/0.76	0.2/0.81	0.18/0.83	0.21/0.8	0.18/0.83	0.27/0.75	0.18/0.83	0.32/0.7		674/0.56	327/0.55
LVis	0.12/0.88	0.12/0.88	0.18/0.83	0.24/0.77	0.21/0.8	0.14/0.87	0.27/0.75	0.27/0.75	0.27/0.75	0.18/0.83	0.27/0.75	0.18/0.83	0.18/0.83		301/0.57
MVU	0.18/0.83	0.24/0.77	0.21/0.8	0.21/0.8	0.26/0.76	0.21/0.8	0.21/0.8	0.21/0.8	0.21/0.8	0.27/0.75	0.18/0.83	0.27/0.75	0.27/0.75	0.28/0.74	

(l) Metric 12: Mean accuracy with constraints

	UMAP	tSNE	FIt_SNE	PCA	Trimap	mTSNE	Isomap	KPCA	LEM	LTSA	MDS	HLLE	LLE	LVis	MVU
UMAP		345/0.54	826/0.71	872/0.77	648/0.54	365/0.53	659/0.55	874/0.77	319/0.56	628/0.53	644/0.54	395/0.52	368/0.53	876/0.78	749/0.62
tSNE	0.27/0.75		812/0.69	863/0.76	625/0.53	379/0.52	648/0.54	876/0.78	341/0.55	695/0.57	674/0.56	419/0.51	398/0.52	875/0.78	861/0.76
FIt_SNE	0.12/0.88	0.11/0.89		319/0.56	687/0.56	864/0.76	698/0.57	312/0.57	945/0.89	646/0.54	624/0.53	995/0.99	884/0.79	341/0.55	357/0.54
PCA	0.11/0.89	0.1/0.9	0.27/0.75		328/0.55	836/0.72	642/0.53	377/0.52	896/0.81	621/0.52	647/0.54	901/0.82	893/0.8	356/0.54	317/0.56
Trimap	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75		648/0.54	311/0.57	519/0.5	628/0.53	298/0.58	322/0.56	674/0.56	651/0.54	514/0.49	589/0.51
mTSNE	0.27/0.75	0.27/0.75	0.11/0.89	0.11/0.89	0.18/0.83		544/0.5	876/0.78	398/0.52	516/0.5	529/0.5	374/0.53	365/0.53	849/0.74	796/0.67
Isomap	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.27/0.75	0.21/0.8		749/0.62	503/0.49	318/0.56	365/0.53	579/0.51	518/0.5	698/0.57	752/0.62
KPCA	0.12/0.88	0.11/0.89	0.27/0.75	0.27/0.75	0.21/0.8	0.11/0.89	0.14/0.87		812/0.69	562/0.5	589/0.51	849/0.74	855/0.75	369/0.53	374/0.53
LEM	0.27/0.75	0.27/0.75	0.05/0.95	0.1/0.9	0.18/0.83	0.27/0.75	0.21/0.8	0.11/0.89		584/0.51	514/0.49	375/0.53	345/0.54	395/0.52	749/0.62
LTSA	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.32/0.7	0.21/0.8	0.27/0.75	0.21/0.8	0.21/0.8		375/0.53	643/0.54	625/0.53	511/0.49	597/0.51
MDS	0.18/0.83	0.18/0.83	0.18/0.83	0.18/0.83	0.28/0.74	0.21/0.8	0.27/0.75	0.2/0.81	0.21/0.8	0.27/0.75		678/0.56	648/0.54	544/0.5	539/0.5
HLLE	0.27/0.75	0.24/0.77	0.02/0.98	0.09/0.91	0.18/0.83	0.27/0.75	0.21/0.8	0.1/0.9	0.27/0.75	0.18/0.83	0.18/0.83		219/0.65	586/0.51	564/0.5
LLE	0.27/0.75	0.27/0.75	0.11/0.89	0.1/0.9	0.18/0.83	0.27/0.75	0.21/0.8	0.1/0.9	0.27/0.75	0.18/0.83	0.18/0.83	0.32/0.7		876/0.78	825/0.71
LVis	0.12/0.88	0.12/0.88	0.27/0.75	0.27/0.75	0.21/0.8	0.11/0.89	0.18/0.83	0.27/0.75	0.27/0.75	0.21/0.8	0.21/0.8	0.21/0.8	0.11/0.89		394/0.52
MVU	0.14/0.87	0.12/0.88	0.27/0.75	0.27/0.75	0.2/0.81	0.14/0.87	0.14/0.87	0.27/0.75	0.14/0.87	0.21/0.8	0.21/0.8	0.21/0.8	0.12/0.88	0.27/0.75	

for the first 6 (i.e., $\hat{\sigma}^2$, ρ_s, $\mu_{R_{nX}}$, Q_{local}, $\lambda_{K_{max}}$, Q_{global}) the number of rejected null hypotheses have remained approximately constant for all the bias values. For Spearman's rank correlation, the number of rejected H_0 is 60% to 80%; for the rest of the metrics, it is 80% to 100%. For the remaining 6 metrics (i.e., $AUC_{\ln K}(R_{nX}(K))$, ACC_{ψ}, nMI, SSI, $logloss$, $\mu_{AUC_{\ln K}(R_{nX}(K))_{N_i}}$), the count of rejected H_0 varied more than the first 6 metrics. Among these for ACC_{ψ} and nMI score, the count varied from 60% to 100%; as for the remaining scores, the number of rejected H_0 varied from 30% to 90%. For these metrics, with a low bias (i.e., between 0 and 8) the majority of the H_0 could not be rejected. However, as the bias increased the count of rejected null hypotheses raised to 90%. The overall statistical comparisons among all the algorithms confirm the fact that not every algorithm performs equally for any particular quality metric. Also, there is at least one algorithm for each metric that is whose performance is statistically significant from the others.

4.4 Discussion - A Guideline for Practitioners

In this section, we use the results presented in Section 4.3 to produce practitioners' guidelines for the analytical contexts presented in Section 4.1.1. The guideline aims at assisting data analysts with making efficient decisions while reducing the dimensionality of datasets. Especially, with the selection of the most appropriate DR algorithm in a given scenario. DR being a complex technique, the selection of an algorithm is influenced by factors such as the analytical context, the quality metrics, the input data, the intrinsic dimensionality of the input data, the chosen hyperparameter values, among others. On the one hand, the selection of the appropriate quality metric makes a significant impact on the quality analysis of embeddings in any analytical context. For example, when training a predictive model with a labeled business dataset an analyst might wish to investigate the data prior to determining the parameters of the model. Such an investigation can reveal the cohesion within individual clusters or class imbalance in the dataset. During this investigation, DR techniques would be used in the context of pattern analysis. Hence, in this case, it would be more beneficial to select quality metrics that can quantitatively evaluate the structural preservation after DR (e.g., Residual variance, $\mu_{R_{nX}}$, Q_{local} and other pattern analysis metrics) over the metrics that can determine the accuracy of the projection ($AUC_{\ln K}(R_{nX}(K))$) or classification error (ACC_{ψ}) after DR. However, as shown in Table 4.4, the same DR algorithm that performs well with respect to pattern analysis might not be the best technique for reducing the dataset dimensions before actually training the model. Hence, in this case, the analysts need to evaluate algorithms using the accuracy metrics for DR. On the other hand, researchers [15], [47] have often argued that an appropriate estimation of the intrinsic dimensionality of the dataset under investigation makes a key impact on the performance of the DR techniques. The reason being, whilst an underestimation of this parameter forces important attributes of the dataset to collapse onto the same dimension, an overestimation of the same increases the noise in the projection while reducing its stability [47]. Hence, when analyzing real-world datasets, it is of utmost importance for any data analyst to accurately estimate the intrinsic dimensionality d (cf. Eq. 4.27) for any given dataset prior to performing DR.

Based on these two aspects of DR, we divide the guidelines presented in this section into two parts. The first part gives an overview of different mechanisms for estimating the intrinsic dimensionality of any dataset in a given scenario. The second part emphasizes on individual analytical contexts for DR (cf. Section 4.1.1) and presents our suggestions regarding the appropriate analytical directions.

4.4.1 Estimating Intrinsic Dimensionality of Real-world Datasets

The intrinsic dimensionality d of the manifold passed as a hyperparameter to most DR algorithms, plays an important role in the amount of information loss after DR. In our experiments, we have used Levina-Bickel's technique [47] for maximum likelihood intrinsic dimensionality estimation (cf. Section 4.2.3) for the chosen datasets. We applied this algorithm in our experiments following the footsteps of Maaten et al. [3] who have also performed large scale experimental studies with DR techniques. Nevertheless, several other techniques can be used for the purpose. Whilst detailed surveys of the techniques for estimating d can be found in [65], [66], in this section, we give a brief overview of some of the most popular methods for determining d and discuss their benefits and limitations.

On a high level, the methods for determining intrinsic dimensionality for datasets can be [15], [67] classified into two types namely: *projection-based* methods and *geometric* methods. The projection-based methods estimate d using the number of eigenvalues that are greater than a given threshold. On the other hand, the geometric methods exploit the nearest neighbor distances among the datapoints for estimating d. A more detailed classification of the techniques was presented by Camastra et al. [66] that divided them into three groups namely *global*, *local*, and *point-wise* methods. The global methods examine an entire input dataset and present a unified estimate for d, whereas local methods approximate d by independently investigating specific subsets of the input data. The point-wise methods on the other hand generate both global and local estimates of d for any given dataset. As discussed by Camastra et al. [66] all the above mentioned three categories include both projection-based and geometric methods for estimating d. In the following, we discuss the categories and their strengths and weaknesses in detail.

i. **Global Methods for Intrinsic Dimensionality Estimation:** The global methods for estimating d assume that all data-points in the input dataset lie on a uniform manifold with a constant dimensionality. The global methods can be further classified into [66]: projection-based, Multidimensional Scaling (MDS) based, and fractal-based methods. Among the three, the projection-based global methods attempt to identify the dimensionality of the optimum subspace by minimizing projection error. Principal Component Analysis (PCA) [41] is the most common among the projection-based techniques that projects the input data along the direction of the maximum variance. When estimating d using PCA, at first the eigenvalues of the covariance matrix for the input dataset are normalized. Then using a threshold for the topmost Eigenvalues, an integer estimate for d is obtained. Some alternatives to traditional PCA include techniques [66] such as probabilistic, Bayesian, or non-linear PCA are used for estimating d. On the other hand, the MDS based global methods for estimating d focus on *distance preservation* among the data-points and approximate the optimum dimensionality of the projection-space by computing the minimum *stress* [42] for different values of d. Here an optimum value for d is obtained from the saturation point in the graph of minimum stress values for all d. Some examples of MDS based techniques include the Bennett's algorithm [66] and the Sammon's mapping [66] technique. In some cases, the neural network-based Curvilinear Component Analysis (CCA) [66] is used instead of the MDS based techniques. Finally, the fractal-based methods are known for generating non-integer estimates for d. Such methods popularly include Camastra and Vinciarelli's correlation-based method [66], Kegl's algorithm [66], Grassberger–Procaccia algorithm [66], and Taken's method [66] that produce a lower bound for the cardinality of the dataset that must be fulfilled to obtain an accurate estimate for d. Nevertheless, global methods for estimating d have their limitations. For example, despite their simplicity and computational feasibility, most PCA based techniques being linear tend to overestimate [66] d. On the other hand, due to their non-robustness [66] with very high-dimensional data, the determination of d can become infeasible [66] with both the MDS based and fractal-based techniques for such datasets.

ii. **Local Methods for Intrinsic Dimensionality Estimation:** In contrast with the global methods, the local techniques for estimating d assume that a dataset does not lie on a uniform manifold with a fixed dimensionality rather it lies on several different manifolds [65] that have their unique intrinsic dimensionalities. Overall, these algorithms consider the neighborhoods of each data-point to be locally linear and estimate the topological [66] dimensions of the underlying manifold. Local methods for estimating d include Fukunaga–Olsen's algorithm [66] that estimates d for any linear subspace within a manifold from the number of non-zero eigenvalues from its covariance matrix. On the other hand, there are local MDS based methods that approximate d similarly as the global MDS based methods discussed earlier. Except the local MDS based methods usually work with subsets of the data instead of the entire dataset. Moreover, the local methods for intrinsic dimensionality estimation include techniques that address multi-scaling problems [66] in datasets. Such techniques include the Brand's method [66] that make assumptions regarding the data-points being only distributed in the directions of the manifold's local tangent space and estimate d based on the scaling of individual neighborhood radiuses [66] with the addition of new data-points in the neighborhoods. Despite being computationally feasible and locally accurate, local methods for estimating d also suffer from their own challenges. On the one hand, computing d with such methods can be infeasible for very high-dimensional datasets; however, on the other hand, the estimation for the dimensionality of the local tangent space using these techniques might not be close to the underlying manifold dimensionality for non-linear manifold. Hence, such an estimate can mislead the subsequent executions of DR algorithms on the entire dataset.

iii. **Point-wise Methods for Intrinsic Dimensionality Estimation:** The algorithms [66] that belong to this category are capable of generating both global and local estimates of d. However, in contrast to local methods that estimate topological dimensions of the manifold (i.e., for smaller subsets of the data), with these techniques a local estimation of d provides pointwise dimensions [66] (i.e., for each data point) of the dataset. Moreover, the global estimates using the point-wise methods are obtained as the mean of pointwise dimensions [66] for all points in a dataset. Examples of point-wise techniques include Levina-Bickel's algorithm [66] along with the approaches proposed by Farahmand et al. [66] and Mordohai & Medioni [66]. Among these, the method proposed by Farahmand et al. [66] uses a nearest-neighbor method to estimate d locally around individual data-points, as the Mordohai &Medioni's algorithm [66] uses tensor voting to identify geometric relationships among data-points. Levina-Bickel's algorithm [66] is arguably the most popular technique among the point-wise methods, that derives the maximum likelihood estimator of d for any given dataset. The Levina-Bickel's technique also has a logarithmic computational complexity [66]. Despite their popularity, one of the challenges of point-wise methods is that their robustness with very high-dimensional datasets cannot be guaranteed in all situations.

When analyzing real-world data, selecting an appropriate technique for estimating d can be challenging due to several reasons. For example, the existence of a diverse range of techniques for the purpose, with each technique having its strengths and limitations can make the selection difficult for a practitioner. Moreover, for real-world industrial datasets, the original structure of the underlying manifold (hence, the actual value of d) being unknown, evaluating different methods for estimating d and making an informed decision remains a challenge [66]. Furthermore, many techniques belonging to each category discussed above, lack in robustness for very high-dimensional data, that is often the case for real-world datasets. Finally, most existing methods for estimating d being highly technical [66], their interpretation and implementation can be a challenge for any novice data analyst.

From the description of methods belonging to the three categories it can be noted that when analyzing smaller subsets of data, local methods for estimating dimensionality can give a more accurate approximation of d. On the other hand, the PCA based global methods can be more computationally feasible for large industrial datasets. At the same time, the point-wise methods can be more robust with high-dimensional datasets [66] for both local and global data analysis. Hence, for such real-life situations, Camastra et al. [66] have suggested forming an ensemble of different estimators for d and compute an average of their outcomes.

4.4.2 Selecting Algorithms and Metrics Based on Analytical Contexts

In this section, highlighting individual analytical contexts for DR we discuss different aspects that influence the selection of an appropriate DR algorithm in a given scenario.

i. **DR for Pattern Analysis and Similarity Search:** As discussed earlier, the preservation of proximity relationships acts as the primary quality criteria of any DR algorithm. In real-life scenarios, high-dimensional datasets are often analyzed for recognizing useful patterns in the data [16]. For example, analysis of gene sequences, investigation of customer behavioral patterns, social media influence analysis, or simple range queries with data among others. In such cases, the quality metrics for structural preservation (especially the local quality criteria) can be the most appropriate way to evaluate DR algorithms. To be more specific, the metrics such as residual variance, Spearman's correlation, and $\boldsymbol{Q_{local}}$ can be more suitable over metrics like KNN-accuracy or $\boldsymbol{nMI}$ scores because they directly evaluate the retention of neighborhood patterns in the embeddings. In our experiments (cf. Table 4.5), t-SNE and UMAP have shown the most robust performance in terms of preservation of local structure after DR. Hence, when in need for a detailed pattern analysis of the input dataset, practitioners can start with computing the $\boldsymbol{Q_{local}}$ metric for the embeddings obtained from t-SNE, UMAP, Isomap, or non-metric MDS. As the metric $\boldsymbol{Q_{local}}$ does not depend on the user defined value of $\boldsymbol{K}$ [4], [5], it can effectively indicate the best performing algorithm for the input dataset without any further assumptions (i.e., on the value of $\boldsymbol{K}$) from the user. On the other hand, for visual quality analysis of the embedding, users can investigate the $\boldsymbol{R_{NX}}$ curve for a user defined range of $\boldsymbol{K}$. Where $\boldsymbol{Q_{local}}$ can help with shortlisting the most appropriate algorithms, visual investigation of the $\boldsymbol{R_{NX}}$ curve can enhance the interpretability of the embedding quality for the chosen algorithm. In this regard, the scalar value of $\boldsymbol{\mu_{R_{NX}}}$ can also provide useful information on the preserved overall quality of the embedding for all values of $\boldsymbol{K}$. More traditional [25] local quality metrics such as the residual variance and spearman's correlation can be used to cross-check and validate the outcome of $\boldsymbol{Q_{local}}$ and $\boldsymbol{\mu_{R_{NX}}}$. When analyzing the local neighborhood quality, $\boldsymbol{\lambda_{K_{max}}}$ metric can be quite informative but requires more analytical expertise from the user [8]. As for algorithms that perform well for small values of $\boldsymbol{K}$ but compromise the neighborhood structure for larger $\boldsymbol{K}$ values, users can be easily mislead [8] to trust a poor-quality embedding as a good quality one. In case all these local quality metrics generate indecisive conclusions [4], then the global quality metric $\boldsymbol{Q_{global}}$ can be useful for selecting the best algorithm.

ii. **DR for Predictive Modelling:** In real-life scenarios, DR is often used as a part of data-preparation prior to training supervised Machine Learning (ML) [3] models. In these cases, ML models are trained using the obtained embedding as their predictive performances [17] are assessed on previously unseen data. In such an analytical context, the DR Accuracy metrics (cf. Section 4.1.2) can assist users to select the most suitable algorithm. For example, although the metric ACC_{ψ} (cf. Eq. 4.13) highly depends on the quality of input data, it can be a useful quality metric [3] in such a scenario. Hence, researchers [3], [14], [17], [19], [28], [29] over the past many years have successfully used ACC_{ψ} as a quality metric for DR. Depending on their analytical expertise, the practitioner can also use $AUC_{\ln K}(R_{NX}(K))$ (cf. Eq. 4.12) while visually investigating the R_{NX} curve for evaluating the accuracy of DR algorithms. As for the most accurate algorithms, Table 4.3 shows PCA has performed with the highest ACC_{ψ} with t-SNE, Kernel PCA, and UMAP following right behind. Similarly, for $AUC_{\ln K}\big(R_{NX}(K)\big)$, t-SNE (and its variants), UMAP and Isomap have proven to be the best performing algorithms. Since the best performing algorithms closely follow each other, practitioners can compare the best 3 to 5 algorithms on their respective input datasets using ACC_{ψ} or $AUC_{\ln K}(R_{NX}(K))$ to make a decision.

iii. **DR with Poor Quality Input Data:** The quality of input data makes a huge impact on the selection of a DR algorithm. Firstly, DR algorithms only work with numeric or at least ordinal data [3]. Hence, in situations where a dataset contains nominal data, DR becomes inapplicable. Secondly, when datasets have too many missing or outlier values, irrespective of the analytical context, no DR algorithm performs optimally [12], [17]. Our experimental results in Table 4.5 show, whereas t-SNE, UMAP, and LLE handle datasets with missing values better than other algorithms, Kernel PCA, nMDS, and PCA minimized the impact of outlier values in the input data. However, as discussed in Section 4.3.1, whilst the missing values impact the preserved local structure in an embedding, outliers perturb its preserved global structure. Hence, in cases of datasets with both missing and outlier values, users might need to select an algorithm (e.g., FIt-SNE, LargeVis) that mediocrely addresses both the inconsistencies.

iv. **DR with Limited Input Data:** Our experiments have proven UMAP to be the most reproducible algorithm with limited amounts of data with different variations of t-SNE following closely behind. Since, both UMAP and t-SNE have proven their robustness with other quality metrics such as $AUC_{\ln K}(R_{NX}(K))$ and Q_{local}, when only limited input data is available users can select either of these algorithms for their analysis.

v. **DR with Limited Computational Resources:** In some cases, practitioners encounter situations where limited computation time or resources are available for executing DR algorithms. This often happens when DR is performed a part of an automated data analysis pipeline or as a part of an interactive visual analytics tool. In such scenarios, the speed of execution becomes an important factor [16] along with the quality of the obtained embedding. However, computational complexity and embedding quality are two different aspects [1] of evaluating DR. Hence, in case of limited resources, practitioners can select any one of the algorithms that perform well with resource constraints (i.e., PCA, KernelPCA, Isomap, or non-metric MDS) prior to assessing them with other metrics for their analytical contexts.

It is important to note that, in and beyond the above-mentioned contexts, DR techniques are most commonly [14], [25] used for visually analyzing high-dimensional data using traditional spatial techniques (e.g., 3D scatter plots). Although, this is one of the most important analytical contexts for DR, in this research we do not consider it as one of our focus areas. The reasons being: (1) visual analytics and interpretations of embeddings obtained from DR is an entire research area [7], [13], [21], [25] on its own that is beyond the scope of this research. (2) Since visual analytics of DR is a qualitative evaluation process for DR that primarily depends on the analysts' perception and expertise, in such a case it is hard to quantitatively compare different DR algorithms and to draw generic inferences about the most suitable technique. Furthermore, it is also important to note that, DR being an extremely popular technique in big-data analytics, several other DR methods exist (e.g., techniques for Independent Component Analysis [3], [11], fractal-based DR methods [10], [63], Linear Discriminant Analysis [3], [11] and many more [11], [23]). Nevertheless, in order to maintain a finite scope for our study, following the guidelines of Maaten et al. [3] and Sorzano et al. [23] we had to exclude several such relevant DR techniques from consideration. Nevertheless, we share our implementation on GitHub[13] so that any DR techniques can be included in the study and quickly compared by analysts.

4.5 Threats to Validity

In experimental studies like ours that involves statistical significance analysis, a set of threats exists that can raise questions about the validity of the research outcome. In this section, we present a set of such threats to the validity of our research that were addressed during this work. The first of such possible threats can be *content validity* [68]. In the scope of this research, content validity refers to a subjective assessment that checks whether all possible DR algorithms and all reasonable evaluation metrics were considered during our study or not. In order to mitigate this threat, following the footsteps of Maaten et al. [3], we selected a wide-spread combination of algorithms for this study. As for the evaluation metrics, we composed them from an extensive review of the related literature. Next, in terms of statistical significance analysis two possible threats could raise questions regarding the validity of our statistical conclusions [69]. The first one is the *construct validity* [69] of the statistical tests that required all the assumptions behind our statistical tests to be fulfilled. Hence, in our experiments, to avoid this threat, we used non-parametric statistical tests that do not make strong assumptions regarding the underlying distribution of the sample data. The second one is *conclusion validity* [69] due to low statistical power of the tests. In order to mitigate this threat, following the guidelines of Demšar et al. [39], we not only selected the most

[13] https://github.com/aindrila-ghosh/SmartReduce

powerful statistical tests, but also, we assessed the reliability of each test.

Finally, in this research we also mitigated the risks for Type I and Type II errors [36]. As the Type I error rate is more critical than the Type II error rate [31], we avoided any dredging[14] or multiple hypothesis testing [31] on the data without making the necessary adjustments to the p-value and the significance threshold α. Moreover, following Demśar et al. [37], [39], we predetermined the statistical tests that would be used in our study. In order to maintain the *internal validity* [69] of the experiments, we not only formally defined the evaluation metrics, but we also carefully controlled the experimental environment and simulate them to fulfill all their desired assumptions. Also, to mitigate the threats to the *external validity* [50] of the experiments, that is to safeguard that our experimental results could be generalized outside the scope of this study, following the guidelines of Demšar et al. [39] we selected a large number of data samples (N=40) and large enough number of runs [36] of the algorithms.

4.6 Related Work

This research is primarily focused on two areas namely: empirical comparisons of DR algorithms over multiple contextual metrics and statistical significance analysis of the evaluation results. Although, we could not find any evidence in literature that combines these two areas, individually, ample amount of research work has been done on both. For example, empirical analysis and comparisons among DR algorithms [3], [23], [62], [64] has been commonly performed by researchers to determine the supremacy of any DR algorithms over its opponents. On the other hand, analysis of existing DR quality metrics [8] and proposal of new metrics [1], [5], [27] have also received a lot of attention from researchers. However, assessment of DR algorithms based on the analytical context and generic guidelines for selecting the most appropriate DR technique in any given context needs more investigation. Moreover, null-hypothesis significance testing of the differences

[14] dredging refers to running different tests on the same data only to select the test that returns a significant difference among samples.

between algorithms is commonly carried out in academia. Nevertheless, in the field of DR such an analysis has never been performed! In this section, we recognize the related work in the two above mentioned areas and discuss the novelty of our work.

4.6.1 Survey of DR Algorithms and DR Quality Metrics

With the abundance of DR techniques, surveys and comparisons among these techniques is a well-studied area among researchers. Whilst some of these surveys have compared techniques belonging to any specific category (e.g., linear [11], non-linear [14], or local [70] DR techniques), other surveys have made comparisons among a set of algorithms from multiple categories [1], [3], [5], [23], [27], [62], [64]. However, most of these surveys have looked into the techniques from a specific perspective. For example, Hou et al [70] have compared three linear DR techniques and presented them as semi-definite programs as a part of a unified framework that can help with solving their complex Eigen-problem. Vlachos et al. [14] have compared only non-linear DR methods to assess their level of accuracy in capturing the user's perception of similarity between data-points in the low dimensional embedding, using visualizations. Additionally, Silva et al. [62] have also looked into only nonlinear methods and have combined the benefits of global and local techniques into a new method. Among other work, Cunningham et al. [11] proposed an optimization framework for linear DR methods, where they have discussed eight such methods in detail and presented the normalized improvement [11] and improved execution time of their framework. On the other hand, Lee et al. [1], [5], [27] and Meng et al. [64] have presented detailed comparisons of different DR algorithms using their proposed DR quality metrics. Another survey of DR that explains a set of linear and nonlinear methods is performed by Maaten et al. [3]. This research classified the techniques into different categories and compared the obtained low dimensional embedding with respect to their generalization errors [3].

DR has a long-existing application in the domain of medicine and cell biology [16]. As a result, in the last few years there has been much research that compares more recent DR methods such as UMAP [2], and t-SNE [26] with more traditional methods like PCA [51]. For example, Becht et al. [16] have performed a detailed comparison primarily between UMAP and t-SNE from different perspectives such as classification accuracy with the low dimensional embedding, preservation of local structure, and reproducibility of the algorithms. Apart from research that primarily focuses on comparing multiple methods, academics have often compared a set of closely related DR methods [24], [34], [43], whenever a new method is proposed. For example, Amid et al [17] have compared their newly proposed algorithm Trimap with PCA, t-SNE, and LargeVis to present the improvement of Trimap over existing methods in terms of outliers and the preservation of global structure.

DR being a well-practiced technique, in the past years several quality criteria [1], [4], [5], [27], [64] were proposed to assess the quality of the obtained embedding. In some cases, researchers have compared multiple DR algorithms for their proposed quality criteria such the co-ranking matrix [1] or Q_{NX}, some researchers have compared multiple quality metrics to identify the most suitable criteria for evaluating DR. Nevertheless, given the fact that the selection of a DR algorithm highly depends on the analytical context, there is still a need for guidelines to select from the existing algorithms using quality metrics. Hence, in contrast to the existing research, we bridge the gap between the DR algorithms and quality metrics and produce a guideline for practitioners for selecting the most appropriate DR method in a given context.

4.6.2 Statistical Significance Tests for Comparing Algorithms

Null hypothesis significance testing is a common strategy [31] used in empirical research to assess if the results obtained from a comparative evaluation are statistically significant. In the past few years, the data mining research community has implemented such statistical analysis of results in several different contexts. For example, one of the most popular research papers that compares different classification algorithms used in supervised machine learning, is presented by Demšar et al. [39]. The work not only presented different possible statistical tests that can be performed to assess the performances of the algorithms, but also formally introduced metrics to evaluate the power of the statistical tests [39].

Mohammadi et al. [37] presented similar experiments with the performances of a set of ontology matching systems, where they have added more statistical tests into the comparisons and have suggested the best performing statistical tests that can be used in specific contexts (e.g., with or without accessibility to large amount of data). Dror et al. [38] have employed statistical tests to evaluate natural language processing algorithms and have presented a generic guideline for selecting the right statistical tests for the purpose. Similarly, Arcuri et al. [36] have produced guidelines for using statistical tests on assessing randomized algorithms by presenting comparisons among parametric and non-parametric tests and by discussing the effects of number of *runs* of these algorithms. Enhancing this work, later Dror et al. [38] have performed replicability analysis of NLP algorithms using statistical significance testing. In this research, for the first time ever we employ statistical significance testing on the results of DR algorithms. We use the results of the null significance tests to generalize and validate our findings from the experimental study.

4.7 Conclusions

Dimensionality Reduction (DR) being a common technique in data analytics several such algorithms have been proposed over the years. While some of these algorithms generate a simple linear projection of the input dataset, other algorithms perform complex non-linear transformations on the data. Nevertheless, the quality analysis of the embedding obtained from DR has been an open research area among academics. The reasons being: (1) for nonlinear DR, no direct mapping exists between the original attributes in the high-dimensional dataset and the dimensions in the embedding. (2) Often the relationships in the high-dimensional data that should be preserved by DR are not clearly identifiable. (3) For real-world datasets the intrinsic dimensionality and the topology of the original manifold is usually unknown. As a result, several quality metrics have been proposed over the past years to evaluate the outcome of DR. Generally, these quality metrics evaluate the extent of preserved proximities among the data-points in a high-dimensional dataset after DR. Among these metrics, whereas some consider the actual proximities (distances) among the data-points, others compare the ranks of the distances between these points. Hence, given a plethora of DR algorithms and quality metrics to evaluate the outcome of DR, it often gets challenging for any practitioner to select the most appropriate quality metrics and DR algorithms in their analytical context. Hence, in this research at first, we identify *five* most popular analytical contexts for DR. Next we categorize the 12 most popular DR quality metrics into the identified analytical contexts followed by a systematic comparison between 15 state-of-the-art DR algorithms in those contexts. Then, after a statistical significance analysis of our obtained results we present an answer to the long-open research question on: "how to determine the most appropriate DR algorithm in a given scenario?". Our results identify t-SNE and UMAP to be the most robust algorithms in terms of metrics that evaluate the preservation of small neighborhoods in the original data. However, our results also indicate that the performance of t-SNE starts to deteriorate as the neighborhood size grows larger. We also found that for datasets with unattributable missing values algorithms such as t-SNE, UMAP, LEM, LLE (i.e., DR techniques that attempt to preserve local structure of data) perform better than the globally focused algorithms. However, in case of datasets with outliers globally focused algorithms such as non-metric MDS, Kernel PCA, PCA perform better than the locally focused methods. In the statistical significance analysis of our results we use 40 real-world datasets (39 open source and 1 from our industrial partner IBM). The null hypothesis significance tests confirm that: the difference in the performances of the best, mediocre, and worst performing algorithms for our chosen 12 quality metrics are indeed statistically significant. Moreover, the analysis also indicates although not every algorithm performs equally well on every DR quality metrics, there is a perfectly reasonable metric for every algorithm where it performs better than its competitors. Finally, based on our experimental results we present a practitioner's guideline for five different analytical contexts in which DR algorithms are commonly used.

References

[1] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction:

Rank-based criteria," *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, Mar. 2009, doi: 10.1016/j.neucom.2008.12.017.

[2] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv:1802.03426*, Feb. 2018, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1802.03426.

[3] L. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality Reduction : A Comparative Review," *J Mach Learn Res 10*, vol. 66, no. 71, p. 13, 2008.

[4] J. A. Lee and M. Verleysen, "Scale-independent quality criteria for dimensionality reduction," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2248–2257, Oct. 2010, doi: 10.1016/j.patrec.2010.04.013.

[5] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, "Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure," *Neurocomputing*, vol. 169, pp. 246–261, Dec. 2015, doi: 10.1016/j.neucom.2014.12.095.

[6] S. Xiang, F. Nie, Y. Song, C. Zhang, and C. Zhang, "Embedding new data points for manifold learning via coordinate propagation," Knowledge and Information Systems, vol. 19, no. 2, pp.159-184, 2009.

[7] A. Bibal and B. Frenay, "Measuring Quality And Interpretability Of Dimensionality Reduction Visualizations," *In Safe Machine Learning Workshop at ICLR*, 2019.

[8] J. Johannemann and R. Tibshirani, "Spectral Overlap and a Comparison of Parameter-Free, Dimensionality Reduction Quality Metrics," *arXiv:1907.01974 [cs, stat]*, Jul. 2019, Accessed: Sep. 30, 2019. [Online]. Available: http://arxiv.org/abs/1907.01974.

[9] A. Ghosh, M. Nashaat, J. Miller, S. Quader, and C. Marston, "A comprehensive review of tools for exploratory analysis of tabular industrial datasets," *Visual Informatics*, vol. 2, no. 4, pp. 235–253, Dec. 2018, doi: 10.1016/j.visinf.2018.12.004.

[10] A. C. Fraideinberze, "Effective and unsupervised fractal-based feature selection for very large datasets: removing linear and non-linear attribute correlations," *In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 615-622, 2016.

[11] J. P. Cunningham and Z. Ghahramani, "Linear Dimensionality Reduction: Survey, Insights, and Generalizations," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2859–2900, 2015.

[12] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, A. Sim, and A. Todd, "Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization," *J. Data and Information Quality*, vol. 11, no. 2, pp. 1–22, Mar. 2019, doi: 10.1145/3301294.

[13] J. Tang, J. Liu, M. Zhang, and Q. Mei, "Visualizing Large-scale and High-dimensional Data," *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pp. 287–297, 2016, doi: 10.1145/2872427.2883041.

[14] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-Linear Dimensionality Reduction Techniques for Classification and Visualization," presented at the Proceedings of the eighth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD), Edmonton, Alberta, Canada, 2002.

[15] G. Navarro, R. Paredes, N. Reyes, and C. Bustos, "An empirical evaluation of intrinsic dimension estimators," *Information Systems*, vol. 64, pp. 206–218, Mar. 2017, doi: 10.1016/j.is.2016.06.004.

[16] 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*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[17] E. Amid and M. K. Warmuth, "A more globally accurate dimensionality reduction method using triplets," *arXiv:1803.00854*, Mar. 2018, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1803.00854.

[18] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, "Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding," *Nature Methods*, vol. 16, no. 3, pp. 243–245, Mar. 2019, doi: 10.1038/s41592-018-0308-4.

[19] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang, "Neighborhood preserving embedding," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Beijing, China, 2005, pp. 1208-1213 Vol. 2, doi: 10.1109/ICCV.2005.167.

[20] C. C. Aggarwal, "On the effects of dimensionality reduction on high dimensional similarity search," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '01*, Santa Barbara, California, United States, 2001, pp. 256–266, doi: 10.1145/375551.383213.

[21] J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, S. Leman, and C. North, "Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 131–141, Jan. 2018, doi: 10.1109/TVCG.2017.2745258.

[22] K. M. Sunderland, D. Beaton, J. Fraser, D. Kwan, P.M. McLaughlin, M. Montero-Odasso, A.J. Peltsch, F. Pieruccini-Faria, D.J. Sahlas, R.H. Swartz, and S.C. Strother, "The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project," *BMC Medical Research Methodology*, vol. 19, no. 1, p. 102, May 2019, doi: 10.1186/s12874-019-0737-5.

[23] C. O. S. Sorzano, J. Vargas, and A. Pascual, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, p. 35.

[24] C. Zhang, S. Xiang, F. Nie, and Y. Song, "Nonlinear dimensionality reduction with relative distance comparison," *Neurocomputing*, vol. 72, no. 7–9, pp. 1719–1731, Mar. 2009, doi: 10.1016/j.neucom.2008.08.003.

[25] B. Rieck and H. Leitte, "Agreement Analysis of Quality Measures for

Dimensionality Reduction," in *Topological Methods in Data Analysis and Visualization IV*, H. Carr, C. Garth, and T. Weinkauf, Eds. Cham: Springer International Publishing, 2017, pp. 103–117.

[26] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[27] J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation," *Neurocomputing*, vol. 112, pp. 92–108, Jul. 2013, doi: 10.1016/j.neucom.2012.12.036.

[28] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood Components Analysis," *Advances in neural information processing systems*, pp. 513–520, 2005.

[29] O. Kramer, "Dimensionality Reduction by Unsupervised K-Nearest Neighbor Regression," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, Honolulu, HI, USA, Dec. 2011, pp. 275–278, doi: 10.1109/ICMLA.2011.55.

[30] X. Zhao and S. S.-U. Guan, "A subspace recursive and selective feature transformation method for classification tasks," *Big Data Anal*, vol. 2, no. 1, p. 10, Dec. 2017, doi: 10.1186/s41044-017-0025-5.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, vol. 1. New York: Springer series in statistics, 2001.

[32] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, Jul. 2013, doi: 10.1016/j.jesp.2013.03.013.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[34] Shiming Xiang, Feiping Nie, Changshui Zhang, and Chunxia Zhang,

"Nonlinear Dimensionality Reduction with Local Spline Embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1285–1298, Sep. 2009, doi: 10.1109/TKDE.2008.204.

[35] M. Gashler, D. Ventura, and T. Martinez, "Iterative Non-linear Dimensionality Reduction by Manifold Sculpting," p. 8.

[36] A. Arcuri and L. Briand, "A practical guide for using statistical tests to assess randomized algorithms in software engineering," in *Proceeding of the 33rd international conference on Software engineering - ICSE '11*, Waikiki, Honolulu, HI, USA, 2011, p. 1, doi: 10.1145/1985793.1985795.

[37] M. Mohammadi, W. Hofman, and Y.-H. Tan, "A Comparative Study of Ontology Matching Systems via Inferential Statistics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 615–628, Apr. 2019, doi: 10.1109/TKDE.2018.2842019.

[38] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing," presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, vol. 1, pp. 1383–1392.

[39] J. Demśar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[40] R. Sherman, "Error of the normal approximation to the. sum of N random variables," *Biometrika*, vol. 58, no. 2, pp. 396–398, 1971.

[41] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[42] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, Jun. 1964, doi: 10.1007/BF02289694.

[43] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, "Local and Global Regressive Mapping for Manifold Learning with Out-of-Sample

Extrapolation," In Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

[44] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.

[45] D. Ulyanov, "Multicore t-SNE." Github Repository, 2016, Accessed: May 14, 2020. [Online]. Available: https://github.com/DmitryUlyanov/Multicore-TSNE.

[46] Q. Hu and C. S. Greene, "Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics," *Pac Symp Biocomput*, vol. 24, pp. 362–373, 2019.

[47] E. Levina and P. J. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," *Advances in neural information processing systems*, pp. 777–784, 2005.

[48] D. Dheeru and G. Casey, "UCI Machine Learning Repository." University of California, Irvine, School of Information and Computer Sciences, 2017, [Online]. Available: http://archive.ics.uci.edu/ml.

[49] Kaggle, "Kaggle Data Repository." https://www.kaggle.com/datasets, Accessed: Jun. 12, 2019. [Online]. Available: https://www.kaggle.com/datasets.

[50] B. Bischl, G. Casalicchio, M. Feurer, F. Hutter, M. Lang, R.G. Mantovani, R.G., J.N. van Rijn, and J. Vanschoren, "OpenML Benchmarking Suites and the OpenML100," *arXiv:1708.03731*, Aug. 2017, Accessed: Jun. 12, 2019. [Online]. Available: http://arxiv.org/abs/1708.03731.

[51] R. Bro and A. K. Smilde, "Principal component analysis," *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, 2014, doi: 10.1039/C3AY41907J.

[52] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A.K. Tung, "LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 236–245, Jan. 2018, doi:

10.1109/TVCG.2017.2744098.

[53] K. Bunte, M. Biehl, and B. Hammer, “Dimensionality reduction mappings,” in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Paris, France, Apr. 2011, pp. 349–356, doi: 10.1109/CIDM.2011.5949443.

[54] W. W. Daniel and C. L. Cross, *Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition*, Eleventh. Wiley, 2018.

[55] P. S. Efraimidis, “Weighted Random Sampling over Data Streams,” *arXiv:1012.0256*, Dec. 2010, Accessed: Jun. 24, 2019. [Online]. Available: http://arxiv.org/abs/1012.0256.

[56] J. Gama, *Knowledge Discovery from Data Streams*, 1st ed. Chapman and Hall/CRC, 2010.

[57] C. C. Aggarwal, “On Biased Reservoir Sampling in the Presence of Stream Evolution,” in *Proceedings of the 32nd international conference on Very large data bases VLDB Endowment*, Sep. 2006, pp. 607–618.

[58] R. Tortolani, “Introducing Bias Intentionally into Survey Techniques,” *Journal of Marketing Research*, vol. 2, no. 1, pp. 51–55, 1965.

[59] S. Garćıa and F. Herrera, “An Extension on ‘Statistical Comparisons of Classifiers over Multiple Data Sets’ for all Pairwise Comparisons,” *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694.

[60] S. García, A. Fernández, J. Luengo, and F. Herrera, “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,” *Information Sciences*, vol. 180, no. 10, pp. 2044–2064, May 2010, doi: 10.1016/j.ins.2009.12.010.

[61] R. R. Bouckaert, “Estimating Replicability of Classifier Learning Experiments,” in *Proceedings of the Twenty-first International Conference on Machine Learning*, New York, NY, USA, 2004, pp. 15–, doi: 10.1145/1015330.1015338.

[62] V. D. Silva and J. B. Tenenbaum, “Global Versus Local Methods in Nonlinear

Dimensionality Reduction," *Advances in neural information processing systems*, pp. 721–728, 2003.

[63] C. T. Jr, A. Traina, L. Wu, and C. Faloutsos, "Fast Feature Selection using Fractal Dimension," *Journal of Information and data Management*, vol. 1, no. 1, p. 14, 2010.

[64] D. Meng, Y. Leung, and Z. Xu, "A new quality assessment criterion for nonlinear dimensionality reduction," *Neurocomputing*, vol. 74, no. 6, pp. 941–948, Feb. 2011, doi: 10.1016/j.neucom.2010.10.011.

[65] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern Recognition*, vol. 36, no. 12, pp. 2945–2954, Dec. 2003, doi: 10.1016/S0031-3203(03)00176-6.

[66] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, Jan. 2016, doi: 10.1016/j.ins.2015.08.029.

[67] S. Gong, V. N. Boddeti, and A. K. Jain, "On the Intrinsic Dimensionality of Image Representations," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3982–3991, doi: 10.1109/CVPR.2019.00411.

[68] A. Ghosh, M. Nashaat, and J. Miller, "The current state of software license renewals in the I.T. industry," *Information and Software Technology*, vol. 108, pp. 139–152, Apr. 2019, doi: 10.1016/j.infsof.2019.01.001.

[69] R. Feldt and A. Magazinius, "Validity Threats in Empirical Software Engineering Research - An Initial Survey," *Seke*, pp. 374–379, Jul. 2010.

[70] C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable local dimensionality reduction approaches," *Pattern Recognition*, vol. 42, no. 9, pp. 2054–2066, Sep. 2009, doi: 10.1016/j.patcog.2008.12.009.

Chapter 5

Interpretation of Structural Preservation in Embeddings

Dimensionality reduction algorithms transform high-dimensional datasets into low-dimensional embeddings while attempting to retain most of the original structural relationships (i.e., relative distances) among the data points. On a high level, all dimensionality reduction algorithms perform complex mathematical optimizations to obtain the low-dimensional projection of a dataset that is often hard to interpret. The primary reason behind this is, the dimensions derived by such algorithms do not have any directly interpretable mappings to the original attributes of the high-dimensional data [1]. Hence, dimensionality reduction being one of the first steps in big-data analytics, a vital concern remains [2]: *if the users do not understand the quality of the low dimensional embedding, they will not make efficient decisions during subsequent analysis.* Moreover, the lack of interpretability in dimensionality reduction algorithms also leads to the challenge of selecting the most appropriate algorithm in a given scenario. In their work, Maaten et al. [3] and Becht et al. [4] have shown that different dimensionality reduction methods perform differently on the same dataset. Also, for every such algorithm, there exists a perfectly reasonable metric [4] for which it is superior to its competitors. For example, in case of the maximum amount of preserved variance in an embedding, Principal Component Analysis (PCA) [3] could perform better than others. Or, for the maximum retention of overall distances among data points, Multidimensional Scaling (MDS) [3] could be the best choice. However, given the fact that there is no established way [5] to evaluate the performance of dimensionality reduction methods, data-scientists often follow their intuitions to use any one of these algorithms, without really understanding their behavior.

The quality [3], [6], [7] of a low-dimensional embedding depends on the extent to which an algorithm can preserve the local structural relationships (i.e., the structural similarities in individual neighborhoods) as well as the global structural associations (i.e., the relative differences in overall neighborhoods) from the original dataset. Hence, an interactive assessment of the preserved structure [1] can not only help users to *trust* the relative positioning of individual data points in a projection but also to have confidence in the overall embedding. In recent years, interactive exploration of low-dimensional embeddings has become an increasingly popular [1], [8], [9], [10] mechanism for evaluating the quality dimensionality reduction. However, our investigation shows, the existing research [1], [8], [9], [10], [11] primarily enables visual exploration of embeddings and rarely compare the embedding to the original data [12]. Also, the majority of the existing techniques do not allow simultaneous comparisons of multiple algorithms to evaluate their outcome on a specific dataset. Most importantly, the research of Adadi et al. [2] and Guidotti et al. [13] confirm that there is still a need for a well-defined mechanism for explaining the structural preservation after dimensionality reduction.

To bridge these gaps, firstly, we present *LAPS* - ***L****ocal* ***A****pproximation of* ***P****reserved* ***S****tructure*, a method & data-type agnostic technique that provides explanations on the preserved local structure of a low-dimensional embedding. The explanations presented by LAPS justify the fidelity of the relative positioning of any individual data-point in an embedding by approximating a neighborhood locally around that point. Secondly, we present *GAPS* - ***G****lobal* ***A****pproximation of* ***P****rojection* ***S****pace*, an interactive technique that presents explanations on the preserved global structure in a low dimensional embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. As a part of an extensive and comprehensive evaluation, we assess both of the proposed techniques for their flexibility (with 10 different dimensionality reduction algorithms on 16 real-life datasets), applicability (i.e., with tabular, text, image, and audio data), utility (i.e., with a user-study that examines their ability to explain the quality [7] of a projection), and reliability (i.e., to assist with the selection of the most appropriate dimensionality reduction algorithm). Our experiments also reveal the roles of different user-defined parameters in the outcome of the proposed techniques. Moreover, they uncover the ability of the techniques in discovering feature correlations in high-dimensional data.

Our primary contributions in this work are as follows:

1. LAPS, a novel algorithm that can provide interpretable and faithful explanations on the retained local structures in any low-dimensional embedding, by locally approximating the neighborhoods.
2. GAPS, a novel technique that provides explanations on the preserved global structure of a manifold in its low-dimensional embedding, by combining local approximations of discrete non-redundant neighborhoods into a global approximation.
3. An extensive 5-phase experimental evaluation of the proposed methods LAPS and GAPS.

The rest of the chapter is organized as follows: Section 5.1 provides an overview of related work as Section 5.2 introduces the necessary background information and design requirements for the proposed techniques. Next, whilst Section 5.3 presents the proposed algorithms, Section 5.4 describes our experimental evaluations of the presented techniques in detail. Finally, Section 5.5 concludes the chapter with a brief discussion on future work.

5.1 Related Work

When it comes to interpretability, visual interaction with low-dimensional embeddings [1], [14], [15] has been the most commonly proposed approach by researchers. In the past few years, several tools [8], [11], techniques [9], frameworks [1], and essays [15] have been presented that aim at making the complex procedure of dimensionality reduction more understandable to its users. Whilst detailed surveys of different interaction paradigms for low-dimensional embeddings can be found in [14] and [16], in this section we highlight the most closely related work to our proposed algorithms.

Covering different aspects of interaction with dimensionality reduction, some existing techniques (e.g., Embedding Projector [8]) allow users to visually explore the neighborhood structures in embeddings. Some other techniques (e.g., Probing Projections [11], CheckViz [6]) visualize the amount of approximation errors in relative distances between the data points in a projection. Among these, whilst Probing Projections [11] assists users to perform distance corrections within neighborhoods, CheckViz [6] enables visualization of false neighborhoods in the projection. Taking the scope for interactivity one step further, some techniques (e.g., Praxis [1], DimStiller [17], LAMP [18]) allow users to interact with the dimensionality reduction process itself. For example, Praxis [1] lets users interactively modify the input feature values for a data-point to see the change in its projection, as well as to alter the position of a point in an embedding to see the changes in original feature values. DimStiller [17] represents the transformation performed during dimensionality reduction as a series of events in a pipeline. The technique allows users to interactively add or remove dimensions in the input and visualize any step in the pipeline at any point in time. The interactive multidimensional projection technique LAMP [18] allows users to interactively steer a projection by enabling them to select the control points that build a family of affine mappings.

To facilitate an efficient selection of hyper-parameters for dimensionality reduction, some techniques such as LDSScanner [19] enable the exploration of the neighborhood structures in the high-dimensional datasets. On the other hand, some tools like SIRIUS [9] enable interactive symmetric dual exploration of the most correlated attributes and neighborhoods in data. At the same time, to explain the quality of embeddings, techniques such as DimReader [10] enable visual exploration of the newly generated axis lines in the projections. Identifying the need to quantify the structural preservation in embeddings, researchers such as Martins et al. [20] propose mechanisms to both visually and quantitatively assess low-dimensional embeddings using false and missing neighbors. In an attempt to explain the relative positioning of data points in an embedding, researchers such as

Pagliosa et al. [21], Silva et al. [22], and Self et al. [23] present techniques that identify the influences of the original attributes in the formation of neighborhood structures.

Nevertheless, our investigation of related research showed that very few researchers (e.g., Kodali et al. [12]) have considered both the aspects of neighborhood preservation and the retention of attribute influences when quantifying the structural quality of an embedding. Even so, a majority of these approaches are designed for only a specific set of dimensionality reduction algorithms (e.g., the approach proposed by Kodali et al. [12] is designed for Weighted Multidimensional Scaling). As a result, these techniques rarely provide an opportunity for a side-by-side comparison among embeddings obtained from different dimensionality reduction algorithms or to perform an interactive selection of the most appropriate algorithm for any given dataset. Also, very few approaches [12] enable any interactive comparisons between the original high-dimensional data and their low-dimensional embeddings to explain the quality of the obtained projections. Hence, there is still a need for a well-defined technique that would visually and quantitatively explain [2], [13] the extent of the preserved local and global structures in reduced dimensions and consider the impacts of both neighborhoods and attribute influence preservations in embeddings.

5.2 Problem Characterization

The overall procedure of dimensionality reduction can be formally defined as: assuming a matrix X of size n × D, that represents a high-dimensional dataset with n records and D attributes so that $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^{D \times n}$. That is, each x_i represents a data-vector for an individual record in X and cardinality of $x_i = |x_i| = D$. Dimensionality reduction can be defined as a mapping function:

$$f: X \to Y \tag{5.1}$$

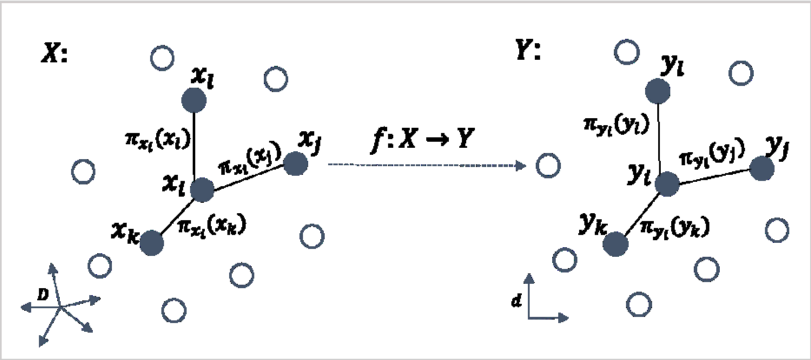


Figure 5.1: An overview of the dimensionality reduction process

where, f transforms X into a low-dimensional embedding Y of size $n \times d$, where $Y = \{y_1, y_2, \dots, y_n\} \in \mathbb{R}^{d \times n}$ and $|y_i| = d$. On a high level, any f can also be formulated [3] as an optimization problem as:

$$\underset{Y \in \mathbb{R}^{d \times n}}{\operatorname{argmin}} f(Y; X, \theta) \tag{5.2}$$

where f represents the objective function that attempts to minimize the structural differences between X and Y as θ represents the hyper-parameters of the dimensionality reduction algorithm. Ideally, the dimension d for Y is the intrinsic dimensionality [3] of the dataset X. The intrinsic dimensionality d represents an estimation of the minimum number of dimensions that can be used to represent X with minimum information loss.

For most real-world datasets, $d \ll D$. This means, as D is reduced to d, the points in the dataset are relocated to a much smaller space than the original high-dimensional manifold. Fig. 5.1 shows such a transformation, where the preservation of the structure of X in Y refers to the fact that the points that were close to each other in X should remain close in Y as well. Also, the points that are far from each other in X should remain the same in Y. The notion of closeness among the data points lying on a manifold is defined using proximity measures [5], [6]. Considering, for any data-point x_i lying on a manifold represented by X, the neighborhood [19], [24] of x_i is a subset Z of X containing x_i, so that $x_i \in Z \subseteq X$.

In this case, Z contains the data points that are closest to x_i. In Fig. 5.1, we define the proximity between the point x_i and its nearest neighbors, i.e., $\forall\ x' \in Z$ and $x_i \neq x'$ as $\pi_{x_i}(x')$. Also, in Fig. 5.1 as f (cf. Eq. 5.2) attempts to minimize the overall divergence between X and Y and preserve the neighborhood structure, in an ideal case [24] the following inequalities should hold:

$$\pi_{y_i}(y_j) \begin{cases} < \pi_{y_i}(y_k) & if \pi_{x_i}(x_j) < \pi_{x_i}(x_k) \\ > \pi_{y_i}(y_k) & otherwise \end{cases} \tag{5.3}$$

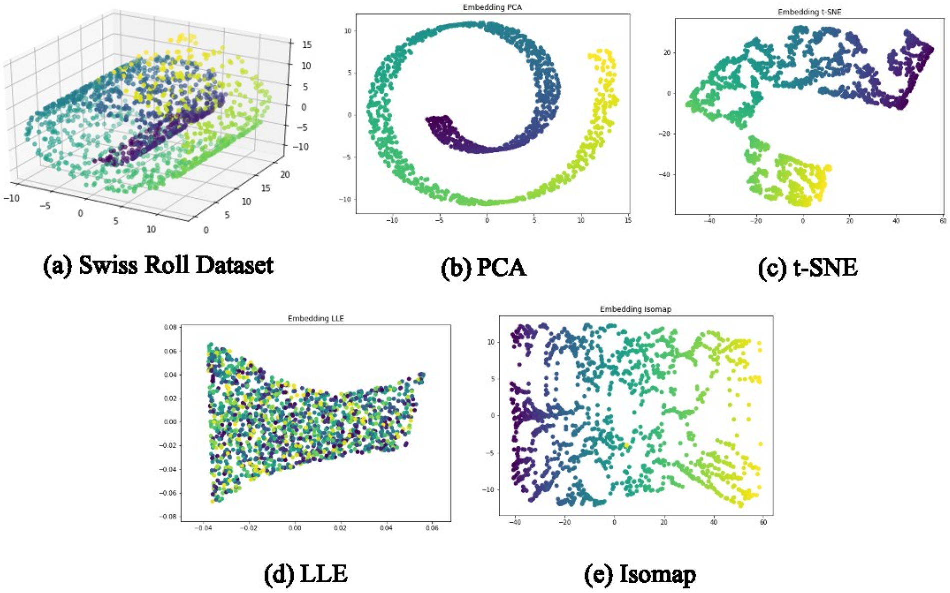
where, $y_i = f(x_i),\ y_j = f(x_j)$, and $y_k = f(x_k)$ and $i \neq j \neq k$. Also, the points x_j and x_k belong to the neighborhood of x_i as y_j and y_k belong to the neighborhood of y_i. Nevertheless, dimensionality reduction being an optimization problem, research [6], [11] shows, its outcome is often likely to converge to a local-optima leading to the inequalities between relative distances not being retained for every data-point in X after the transformation.

A large number of non-linear dimensionality reduction algorithms (e.g., MDS, Isomap [3]) rely on the neighborhood geometry of a manifold to recognize its overall structure. Among them, some algorithms (e.g., MDS [3]) use the Euclidean distance [9] as their proximity measure for the data points, considering the manifold to be locally isometric to a Euclidean space [6]. The Euclidean distance $dist_\varepsilon$ between two points x_i and x_j can be defined as:

$$dist_\varepsilon(x_i, x_j) = \sqrt{\sum_{a=1}^{D} (u_a - v_a)^2} \tag{5.4}$$

where, u_a and v_a represent individual features in x_i and x_j respectively. On the other hand, some algorithms (e.g., Isomap [25]) use pairwise Geodesic distance [3] among points to measure the global intrinsic feature of the manifold. The Geodesic distance $dist_\gamma$ among the points x_i and x_j in X, can be defined using the infimum over the lengths of all the smooth paths connecting the two points as:

$$dist_\gamma(x_i, x_j) = \inf\{L(\sigma)\} \tag{5.5}$$



(a) Swiss Roll Dataset

(b) PCA

(c) t-SNE

(d) LLE

(e) Isomap

Figure 5.2: Structural preservation with different dimensionality reduction algorithms on the artificial Swiss Roll dataset

where σ is a smooth path from x_i and x_j. The smoothness of σ is measured by the number of continuous derivatives along σ. Formally, assuming S_σ as a set of all points along σ and every $s_\sigma \in S_\sigma \in \mathbb{R}$, then σ is considered to be smooth if it has derivatives of all orders for every $s_\sigma \in S_\sigma$.

Due to the use of different proximity measures and optimization functions, different dimensionality reduction algorithms perform differently [5] on the same dataset. For example, as shown in Fig. 2.2, the artificial Swiss-roll dataset (cf. Fig. 2.2.a) is transformed using four different dimensionality reduction algorithms. The Figures 2.2.b., 2.2.c, 2.2.d, and 2.2.e, clearly show the differences in preservation of the local and global structures of the original dataset. Hence, it can be noted that the structural preservation plays an important role in selecting the most suitable dimensionality reduction algorithm for a given dataset. Moreover, as dimensionality reduction is performed prior to any deeper analysis with a high-dimensional dataset (e.g., training a predictive model with Y), the lack of preserved structure can result in poor subsequent analysis. In this chapter, we propose interpretable explanations about Y as a solution to the above-mentioned problems.

5.2.1 Requirements Analysis for Explanations

In this chapter, we define *explanations of preserved structure* as a *set* of meaningful textual and visual artifacts that describes the ability of an algorithm to retain the original relative distances among the data points in the low-dimensional space. More specifically, the explanations aim to answer a range of questions regarding *what* happened during the transformation of a dataset using the equation 5.2. In this section, we define a set of requirements for the explanations of reduced dimensions.

First of all, we need the explanations to be **interpretable** [26], [27] for both expert and novice users. According to Ribeiro et al. [28], and Yang et al. [29], as humans, we relate to meaningful names much faster than numeric values or complex graphical representations. Hence, along with displaying the relative distances among data points [8], [9], [11], there is a need to present the explanations in terms of the attributes of the original dataset. Moreover, to prevent the explanations from overwhelming users in cases of very high-dimensional (e.g., $D > 100$) datasets, there needs to be a way for the users to regulate the amount of information they would want to see.

Secondly, we expect **local fidelity** and **global legitimacy** from the explanations. For example, whilst it is important for LAPS to be locally faithful (i.e., for the data point being explained), for GAPS it is essential to be globally legitimate (i.e., accurate for the entire dataset). We note that it is often impossible to achieve complete global legitimacy of the explanations unless all the data points in a dataset are considered. The explanations need to incorporate this fact.

Thirdly, the explanations should be **algorithm & data-type agnostic**. That is, the selection of a suitable DR algorithm requires the explanations to be applicable for a variety of such algorithms. Moreover, to explore the full potential of the explanations, they should be flexible enough to incorporate any type of data.

Finally, the explanations for local and global structures should be **consistent**. That is, not only the look and feel of the explanations but also the user-interactions with them should be made consistent for both LAPS and GAPS.

5.3 Explaining Reduced Dimensions

In this section, we present the overall ideas of our proposed methods *LAPS (Local Approximation of Preserved Structure)* and *GAPS (Global Approximation of Projection Space)*.

Prior to formally defining the methods, we introduce some notations that would be used later in the chapter. Considering a random point $x_i \in X$, that is represented using a feature vector $U = [u_{1,} u_2, \dots, u_D] \in \mathbb{R}^D$, where $\forall\, u_a \in U,\ a$ represents an individual feature. The locality around x_i is defined using a set $Z \subseteq X$, containing k-nearest neighbors of x_i. We define the local explanation for x_i as a *set* containing feature influence explanations $fie(x_i)$ and the local divergence λ_{x_i} for x_i. Whilst the feature influence explanation $fie(x_i)$ represents an interpretable function that approximates the contribution of each feature in the relative proximity between x_i and its k-nearest neighbors, the local-divergence λ_{x_i} represents the disagreements in the feature influence explanations and neighborhood structures of x_i and y_i . In this case, $y_i \in Y$ represents the low-dimensional counterpart of $x_i \in X$, i.e., $y_i = f(x_i)$. In par with our requirements for **consistency**, we compose the global explanations using $fie(X_S)$ and λ_{X_S}. Here, $fie(X_S)$ represents the unification of local feature explanations of a user-defined subset X_S of X. λ_{X_S} presents the global structural divergence between the original data points in X_S and their counterparts in the embedding Y_S. Formally, we define the local and global explanations as:

$$loc\_expl_{x_i} = \{\, fie(x_i), \lambda_{x_i} \mid \exists x_i \in X\}$$

(5.6)

$$glob\_expl_{X_S} = \{\, fie(X_S), \lambda_{X_S} \mid \exists X_S \subseteq X\}$$

where both $loc\_expl_{x_i}$ and $glob\_expl_{X_S}$ are *sets* of textual and visual artifacts used for interpreting the embeddings.

5.3.1 Motivating Example

To facilitate a better understanding of the concept of *explanation* defined above, in this section, we demonstrate the idea with a toy-example. In this example, we have our analyst Alice analyze the Animals[15] dataset [9] that contains 30,475 images and distinguishes 50 animal classes using 85 numeric attributes. In this case, X represents the dataset, where $n = 30475$ and $\mathrm{D} = 85$. After transforming the data into a 2D embedding Y using *any* dimensionality reduction algorithm, Alice wants to interpret the preserved local structure in the embedding. To obtain a local explanation using LAPS, Alice selects a single point-of-interest x_i (say, $x_i =$ *rabbit*) from X. The $loc\_expl_{x_i}$ of preserved structure for *rabbit* contains the following: (1) $fie(x_i)$ and $fie(y_i)$: the positive and negative influence scores for all the 85 attributes in the construction of the neighborhood of *rabbit* in X as well as in Y (where, $y_i = f(x_i)$). (2) λ_{x_i}: the local-divergence score for the data-point *rabbit*. Here, λ_{x_i} is computed as a weighted sum of the disagreements between $fie(x_i)$ and $fie(y_i)$ and the disparities in the neighborhood structures for the point *rabbit* in X and Y. Similarly, to obtain an explanation on the preserved global structure in the embedding, Alice interactively selects a subset X_S from X that contains the data points *rabbit, mouse, hamster, mole,* and *squirrel*. The $glob\_expl_{X_S}$ obtained using GAPS consists of (1) $fie(X_S)$ and $fie(Y_S)$: the overall influences of the original attributes in the relative positioning of the neighborhoods of the points in X_S and Y_S. (2) λ_{X_S}: the global divergence computed by adding the scaled local divergences of the points $\lambda_{x_i} \in \lambda_{X_S}$.

5.3.2 Local Approximation of Preserved Structure

In order to generate **data-type agnostic** local explanations using LAPS, we avoid making any assumptions about X. Next, as a pre-processing before transforming X to Y, we estimate the intrinsic dimensionality d of X using the maximum likelihood

[15] https://cvml.ist.ac.at/AwA2/

Algorithm 5.1 – The *LAPS* Procedure

```
Input: dataset $X$, embedding $Y$, instance $x_i$, neighbor count $k$
Output: $fie(x_i)$, $fie(y_i)$ and local divergence $\lambda_{x_i}$

Step 1: Obtain nearest neighbors for $x_i$ and $y_i$
  for all $j \in \{0,1,\ldots\ldots,k\}$ do:
    $Z_{x_i} \leftarrow nn_{x_{ij}}$ , $Z_{y_i} \leftarrow nn_{y_{ij}}$                    (Eq. 5.8)

Step 2: Approximate the local neighborhoods for $x_i$ and $y_i$
  for all $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$ do:
    $\overline{Z_{x_i}}$ ← sample_around ($x'$), $\overline{Z_{y_i}}$ ← sample_around ($y'$)     (Eq. 5.9)

Step 3: Compute relative proximities among $x_i$, $y_i$ and their respective neighbors
  for all $x' \in \overline{Z_{x_i}}$ and $y' \in \overline{Z_{y_i}}$ do:
    $\overline{\pi_{x_i}} \leftarrow \pi_{x_i}(x')$, $\overline{\pi_{y_i}} \leftarrow \pi_{y_i}(y')$

Step 4: Order data-vectors in terms of ascending proximity
  $\overline{\overline{Z_{x_i}}}$ ← sort ($\overline{Z_{x_i}}$), $\overline{\overline{Z_{y_i}}}$ ← sort($\overline{Z_{y_i}}$)

Step 5: Compute feature distance contributions for $x_i$ and $y_i$
  Compute $FC_{Z_{x_i}}$ ← feature_contribution ($\overline{\overline{\overline{Z_{x_i}}}}$)          (Eq. 5.11)
  Compute $FC_{Z_{y_i}}$ ← feature_contribution ($\overline{\overline{\overline{Z_{y_i}}}}$)          (Eq. 5.11)

Step 6: Compute feature influence explanations for $x_i$ and $y_i$
  $fie(x_i)$ ← corr($FC_{Z_{x_i}}$,$\overline{\pi_{x_i}}$ ), $fie(y_i)$ ← corr($FC_{Z_{y_i}}$,$\overline{\pi_{y_i}}$ )     (Eq. 5.12)

Step 7: Compute the local divergence score for $x_i$
  Compute $\lambda_{x_i}$                                             (Eq. 5.13)
```

intrinsic dimensionality estimator [30] defined as:

$$\hat{d} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{d}_k \quad where,\ \hat{d}_k = \frac{1}{n} \sum_{i=1}^{n} d_k(X) \qquad (5.7)$$

where, $\hat{d}$ represents a unit vector with an estimation for d and $(k_2 - k_1)$ signifies the range of nearest neighbors to consider while estimating d. This pre-processing is necessary [3], [19], [30] as the estimation of d prior to obtaining Y not only ensures noise reduction [19], [24] in Y, such an estimation also enhances the stability [30] of Y. Next, with d as a parameter to f, we obtain Y as $f(X)$. In order for the explanations to be **algorithm-agnostic**, we also avoid making any assumptions about f.

Once Y is obtained, the LAPS process is initiated (cf. Algorithm 5.1). As the first step, the user interactively selects a single data-point $x_i \in X$. Considering, $y_i \in Y$ being the low-dimensional counterpart of x_i i.e., $f(x_i) = y_i$, LAPS begins with the identification of the localities (i.e., neighborhood structure) around x_i and y_i by performing an unsupervised k-nearest neighbor search using the *ball-tree* [31] algorithm. Where the nearest neighbors nn_{x_i} for x_i and nn_{y_i} for y_i be defined as:

$$\begin{aligned} nn_{x_i} &= \{\forall x' \in X | \forall x'' \in X, x' \neq x'' \colon \pi_{x_i}(x') \leq \pi_{x_i}(x'')\} \\ nn_{y_i} &= \{\forall y' \in Y | \forall y'' \in Y, y' \neq y'' \colon \pi_{y_i}(y') \leq \pi_{y_i}(y'')\} \end{aligned} \tag{5.8}$$

After identification of the indexes of the k-nearest neighbors for both x_i and y_i, the original features vectors from X for the closest neighbors of x_i and y_i are selected and combined into feature vector matrices Z_{x_i} and Z_{y_i} respectively, where, $x_i \in Z_{x_i}$ and $y_i \in Z_{y_i}$. To enhance user-interactions with the LAPS process, we allow the value of k to be user-defined. The primary reasons behind using the *ball-tree* algorithm for an unsupervised k-nearest neighbor search are firstly, the algorithm is well-known [31] for its efficiency with the fast discovery of nearest neighbors in high-dimensional manifolds. Secondly, f being a data-transformation technique, the neighborhood structure after using f has no direct impact from the training labels associated with the data points.

Next, to approximate the local neighborhoods for x_i and y_i, LAPS samples instances around each $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$. During this step, a constant number of data-point samples are drawn uniformly at random having a normal distribution centered around each $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$. Formally, the sampling of each perturbed neighbor for any data point $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$ can be defined as:

$$\begin{aligned} \bar{x}_i &= \forall x' \in Z_{x_i}, \forall u'_a \in x', \delta \in [0,1] \colon \{\delta \times \sigma_{u'_a} + \mu_{u'_a}\} \\ \bar{y}_i &= \forall y' \in Z_{y_i}, \forall v'_a \in y', \delta \in [0,1] \colon \{\delta \times \sigma_{v'_a} + \mu_{v'_a}\} \end{aligned} \tag{5.9}$$

where δ represents a random perturbation, as $\sigma_{u'_a}$, $\sigma_{v'_a}$ represent the standard deviation[16] of an individual feature value in Z_{x_i} and Z_{y_i} respectively. At the same time, $\mu_{u'_a}$and $\mu_{v'_a}$ signify the means of individual feature values in Z_{x_i} and Z_{y_i} respectively. To ensure **local fidelity,** such an approximation of local structure of data points is a commonly practiced [28], [32] approach among researchers. The perturbed neighborhood for each point in both Z_{x_i} and Z_{y_i} are combined into feature vector matrices $\overline{Z_{x_\iota}}$ and $\overline{Z_{y_\iota}}$ respectively.

In the next step, the relative proximities: $\pi_{x_i}(x')$ is calculated between the points x_i and $\forall x' \in \overline{Z_{x_\iota}}$ and $\pi_{y_i}(y')$ is computed between y_i and $\forall y' \in \overline{Z_{y_\iota}}$. In case of the feature vectors for x_i and y_i containing only continuous values, the Euclidean distance (cf. Eq. 5.4) is used as the proximity measures $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$. In contrast, in the case of feature vectors with a mixture of both continuous and categorical values, the Gower dissimilarity [9], [11], [33] is used as the proximity measures $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$ between instances. The Gower dissimilarity [33] $dist_\omega$ between any pair of data points x_i, x' can be defined as:

$$dist_\omega(x_i, x') = \sum_{u=1}^{D} \delta_{x_i x' u} \times dist_{\omega_{x_i x' u}} / \sum_{u=1}^{D} \delta_{x_i x' u} \qquad (5.10)$$

where u represents an individual attribute in $\overline{Z_{x_\iota}}$ as $dist_\omega$ signifies the distance between x_i and x' for attribute u. In case of continuous variables $dist_\omega$ is calculated as $abs|x_{iu} - x'_u|/range(u)$. For categorical variables, $dist_\omega$ is 0 if $x_{iu} = x'_u$, otherwise 1. In Eq. 5.10, $\delta_{x_i x' u}$ is 1 if x_{iu} and x'_u are comparable, otherwise 0. In our experiments, we do not consider weighted proximity measures [9] in par with our **algorithm-agnostic** design goal because not every dimensionality reduction method considers additional feature weights in their process [3].

[16] The number of data points in the perturbed neighborhood of every data point is fixed to 5000, Hence, due to the effect of central limit theorem, the distribution of feature values was noticed to be very close to a Gaussian distribution.

Next, every data-vector $x' \in \overline{Z_{x_\iota}}$ and $y' \in \overline{Z_{y_\iota}}$ are ordered in terms of their (ascending) proximity with the original (data) points x_i and y_i respectively. We represent these ordered feature-vectors matrices as $\overline{\overline{Z_{x_\iota}}}$ and $\overline{\overline{Z_{y_\iota}}}$ respectively. Also, as the ascending proximity values between x_i and the data points $x' \in \overline{Z_{x_\iota}}$ are stored in a set $\overline{\pi_{x_\iota}}$, the same for y_i and the data points $y' \in \overline{Z_{y_\iota}}$ are stored as $\overline{\pi_{y_\iota}}$. The feature-vector matrices $\overline{\overline{Z_{x_\iota}}}$ and $\overline{\overline{Z_{y_\iota}}}$, are then used compose two feature distance contribution [34] matrices namely, $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ respectively. Each element in a feature distance contribution matrix holds the impact of each attribute in the overall distances between a pair of consecutive points. We define the elements in $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ as:

$$\begin{aligned} &\forall x'_i, x'_{i+1} \in \bar{\bar{Z}}_{x_i}, \forall u \in x'_i, x'_{i+1}: \pi_{x'_{i_u}}(x'_{i+1_u}) / \pi_{x'_i}(x'_{i+1}) \\ &\forall y'_i, y'_{i+1} \in \bar{\bar{Z}}_{y_i}, \forall v \in y'_i, y'_{i+1}: \pi_{y'_{i_u}}(y'_{i+1_u}) / \pi_{y'_i}(y'_{i+1}) \end{aligned} \tag{5.11}$$

where u represents an individual attribute in $\overline{\overline{Z_{x_\iota}}}$ and v represents the same in $\overline{\overline{Z_{y_\iota}}}$. The points x'_i, x'_{i+1} and y'_i, y'_{i+1} signify two consecutive data-vectors in $\overline{\overline{Z_{x_\iota}}}$ and $\overline{\overline{Z_{y_\iota}}}$ respectively. We build the two matrices based on the concept of feature distance contribution [34], which represents a ratio showing the importance (i.e., contribution) of an individual feature in the overall distance between two points.

Finally, from the feature distance contribution matrices, LAPS generates the first component of the local explanations: feature influence explanations $fie(x_i)$ for x_i using the Pearson's correlation[17] [4] between each column (representing the distance contribution for each feature) in $FC_{Z_{x_i}}$ with ordered overall distances $\overline{\pi_{x_\iota}}$ between the data points in $\overline{Z_{x_\iota}}$. Similarly, feature influence explanations $fie(y_i)$ are calculated for y_i, the embedding counterpart of x_i from $FC_{Z_{y_i}}$ and $\overline{\pi_{y_\iota}}$. Formally, the feature influence explanations $fie(x_i)$ and $fie(y_i)$ can be defined as:

[17] Experiments with Spearman's correlation (non-parametric) returned the same values with Pearson's (parametric) until the second decimal point

$$
\begin{aligned}
fie(x_i) &= \left\{\forall fc_{xa} \in FC_{Z_{x_i}}: \frac{cov(fc_{xa}\,, \overline{\pi}_{x_i})}{\sigma_{fc_{xa}} \sigma_{\,\overline{\pi}_{x_i}}}\right\} \\
fie(y_i) &= \left\{\forall fc_{ya} \in FC_{Z_{y_i}}: \frac{cov(fc_{ya}\,, \overline{\pi}_{y_i})}{\sigma_{fc_{ya}} \sigma_{\,\overline{\pi}_{y_i}}}\right\}
\end{aligned}
\qquad (5.12)
$$

Where, fc_{xa} and $\overline{\pi}_{x_i}$ as well as fc_{ya} and $\overline{\pi}_{y_i}$ are the pairs of random variables under consideration. In Eq. 5.12, $cov\left(fc_{xa}\,, \overline{\pi}_{x_i}\right) = \sum_{j=1}^{n}(fc_{xa_j} - \mu_{fc_{xa}})(\overline{\pi}_{x_i}(x_j) - \mu_{\overline{\pi}_{x_i}})$ and $\sigma_{fc_{xa}} = \sqrt{\sum_{j=1}^{n}(fc_{xa_j} - \mu_{fc_{xa}})^2}$ as $\sigma_{\,\overline{\pi}_{x_i}} = \sqrt{\sum_{j=1}^{n}(\overline{\pi}_{x_i} - \mu_{\overline{\pi}_{x_i}})^2}$. The overall $fie(x_i)$ and $fie(y_i)$ are represented using a *set* of key-value pairs, where, the keys fc_{xa} and fc_{ya} represent individual attributes in matrices $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$ respectively. Finally, we compute the local divergence λ_{x_i} for x_i as:

$$
\lambda_{x_i} = w_1 \pi_{fie(x_i)}\big(fie(y_i)\big) + w_2 \frac{nn_{x_i} \cap nn_{y_i}}{\left|nn_{x_i}\right|} + w_3 d_{r_{nn_{x_i}, nn_{y_i}}} \qquad (5.13)
$$

where, $\pi_{fie(x_i)}\left(fie(y_i)\right)$ signifies the *cosine* distance between $fie(x_i)$ and $fie(y_i)$. As $d_{r_{nn_{x_i}, nn_{y_i}}}$represents the difference between the relative orders of neighborhoods of x_i and y_i. In Eq. 5.13, w_1, w_2, and w_3 signify user-defined scalar weights of the three components of λ_{x_i}. By default, w_1, w_2, and w_3 are equal, i.e., 0.33.

Since Algorithm 5.1 produces explanations for a single data-point in X, its complexity does not depend on the size of X, but on the user-defined size of the sampled neighborhood $\overline{Z_{x_\iota}}$ (Eq. 5.8) for the selected instance. As per our analysis, the run-time complexity of Algorithm 5.1 is $O(n^2)$, n being the number of samples in $\overline{Z_{x_\iota}}$. In practice, on a personal computer (i.e., with 4 cores and 8 GB main memory) LAPS executes in less than 15 seconds for $n = 5000$ data points.

Algorithm 5.2 – The *GAPS* Procedure

```
Input: dataset $X$, data subset $X_S$, budget $B$, neighbor count $k$
Output: global divergence $\lambda_X$

Step 1: Generate local feature influence explanations
  for all $x_i \in X_S$ and $y_i \in Y_S$ do:
    $LFI_{X_S} \leftarrow fie(x_i), LFI_{Y_S} \leftarrow fie(y_i)$, $\lambda_{X_S} \leftarrow \lambda_{x_i}$    (Algo. 5.1)

Step 2: Approximate the local neighborhoods for the nearest neighbors of all $x_i \in X_S$ and $y_i \in Y_S$
  for all $x_i \in X_S$ and $y_i \in Y_S$ do:
    for all $j \in \{0,1, \dots \dots, k\}$ do:
      $Z_{X_S} \leftarrow$ sample_around $(nn_{x_{ij}})$          (Eq. 5.8, 5.9)
      $Z_{Y_S} \leftarrow$ sample_around $(nn_{y_{ij}})$

Step 3: Compute pairwise proximities between each pair of points in feature vectors in $Z_{X_S}$ and $Z_{Y_S}$ followed by an estimation of the overall feature distance contribution and global feature influence explanations
  for all $x_i, x_j \in Z_{X_S}$ and $y_i, y_j \in Z_{Y_S}$ do:          (Eq. 5.10, 5.11)

    Compute $\pi_{x_i}(x_j)$, $\pi_{y_i}(y_j)$, Compute $fie(X_S), fie(Y_S)$

Step 4: Obtain an approximation of the global divergence $\lambda_{\widehat{X_S}}$ for the selected subset $X_S$
  Compute $\lambda_{\widehat{X_S}}$          (Eq. 5.12)

Step 5: Calculate the overall global divergence for $X$ using the Global-Local Approximation (GLA) approach
  Compute $\lambda_X$          (Eq. 5.13)
```

5.3.3 Global Approximation of Projection Space

To ensure the **global legitimacy** of explanations, we now propose the algorithm *GAPS* that generates an estimation for the retained global structure of the projection space. The preserved global structure is explained as a unification of preserved local structures [28], [35] for a subset X_S of non-redundant data points in X. Acknowledging the importance of a judicious selection of X_S for an accurate global approximation, *GAPS* enables two different ways for formulating X_S from X. Here, either the users interactively pick data points around the manifold or GAPS selects a fixed number of instances uniformly at random belonging to each training label around X. In either case, GAPS lets the users determine the number of instances that they are willing to investigate, and represents it with a budget B. Potentially, the formation of X_S can also be represented using an Exhaustive Subset Enumeration [36] problem. We envision the maximization of diversified sample selection for X_S as future work.

Once the data points in X_S are selected, as shown in Algorithm 5.2, GAPS obtains a set of local explanations for $x_i \in X_S$ and $y_i \in Y_S$, where, $y_i = f(x_i)$. Next, as the local feature contributions for the instances in X_S and Y_S are used to compose two $B \times D$ dimensional matrices LFI_{X_S} and LFI_{Y_S} respectively, the local divergences for each point in X_S and Y_S are represented using the sets λ_{X_S} and λ_{Y_S} respectively. In parallel, GAPS obtains a global estimate of the structural relations among the data points in X_S. As the first step towards obtaining this, the approximated local neighborhoods (cf. Eq. 5.9) for each $x_i \in X_S$ and $y_i \in Y_S$ are combined into two feature vector matrices namely Z_{X_S} and Z_{Y_S} respectively. Next, pairwise proximities between each pair of points in feature vectors in Z_{X_S} and Z_{Y_S} are calculated. Considering the proximities among data points around a high-dimensional manifold, in GAPS, we use the Geodesic distances (cf. Eq. 5.5) among the pairs of points in Z_{X_S} and Z_{Y_S}. After ordering the data point pairs in ascending order of proximity, using equations 5.11 and 5.12 an estimation of the overall feature distance contribution and global feature influence explanations are obtained. Finally, similarly as LAPS, an approximation of the global divergence $\lambda_{\widehat{X_S}}$ for the selected subset X_S is obtained as a weighted sum of the disagreements in the overall estimation of the feature influences, and the disagreements in the neighborhood structures for X_S and Y_S.

Boyd et al. [35] and Haftka et al. [37] show, on the one hand, the local approximation of divergence for each data-point in X_S is the most effective near the point where it was calculated. However, the accuracy of such local approximations can deteriorate [37] as it moves away from the point where it was constructed. In contrast, a global approximation may not be accurate for every data-point in the manifold, its quality does not deteriorate with distance. Hence, in GAPS we follow the Global-Local Approximation (GLA) [37] approach. GLA allows an additive blending of local approximations to form a globally valid approximation. Here, before the unification of the local-approximations, the ratio of the global estimate to each of the local approximation is used as a scaling factor [35] to multiply the local-approximations. Hence, we define the overall divergence in the preserved global structure λ_X as:

$$\lambda_X = \sum_{j=1}^{B} \frac{\lambda_{X_{S_j}}}{\lambda_{\hat{X}_S}} \lambda_{X_{S_j}} \tag{5.14}$$

where, λ_X represents an additive blending of scaled local divergence scores $\lambda_{x_i} \in \lambda_{X_S}$.

Although Algorithm 2 presents a unified approximation for B instances in X, it has a run-time complexity of $O(n^2)$, n being the number of row vectors in the unified perturbed neighborhood matrix Z_{X_S}.

5.4 Experimental Evaluation

In this section, we present the results of our experimental evaluations of the two proposed techniques. This section aims at answering the following questions:

i. Do LAPS and GAPS fulfill their design requirements discussed in Section 5.2.1?
ii. Can the proposed methods instill confidence in the projection and enable the selection of a suitable dimensionality reduction algorithm?
iii. Are the explanations able to effectively explain the structural preservations in embeddings?
iv. Can the explanations be considered as an improvement over the most closely related work?
v. Do the explanations remain consistent for different user-selected parameter combinations?

Based on the above-mentioned questions, our evaluation of the proposed methods was performed in five phases. Firstly, we applied the techniques on 16 real-world datasets to assess the applicability of the methods. Secondly, we investigated the role of the local and global divergence scores in the selection of the most suitable dimensionality reduction algorithms. Thirdly, we executed a user-study to assess the utility of the two techniques. Fourthly, we compared the proposed techniques to the most closely related research that aims at interpreting embeddings. Finally, we analyzed the impact of different user-defined parameter combinations of the proposed techniques. Overall, this section is divided into two sub-sections; first, we define the algorithms and datasets that were used in our experiments, followed by a detailed analysis of the answers to the above-mentioned questions.

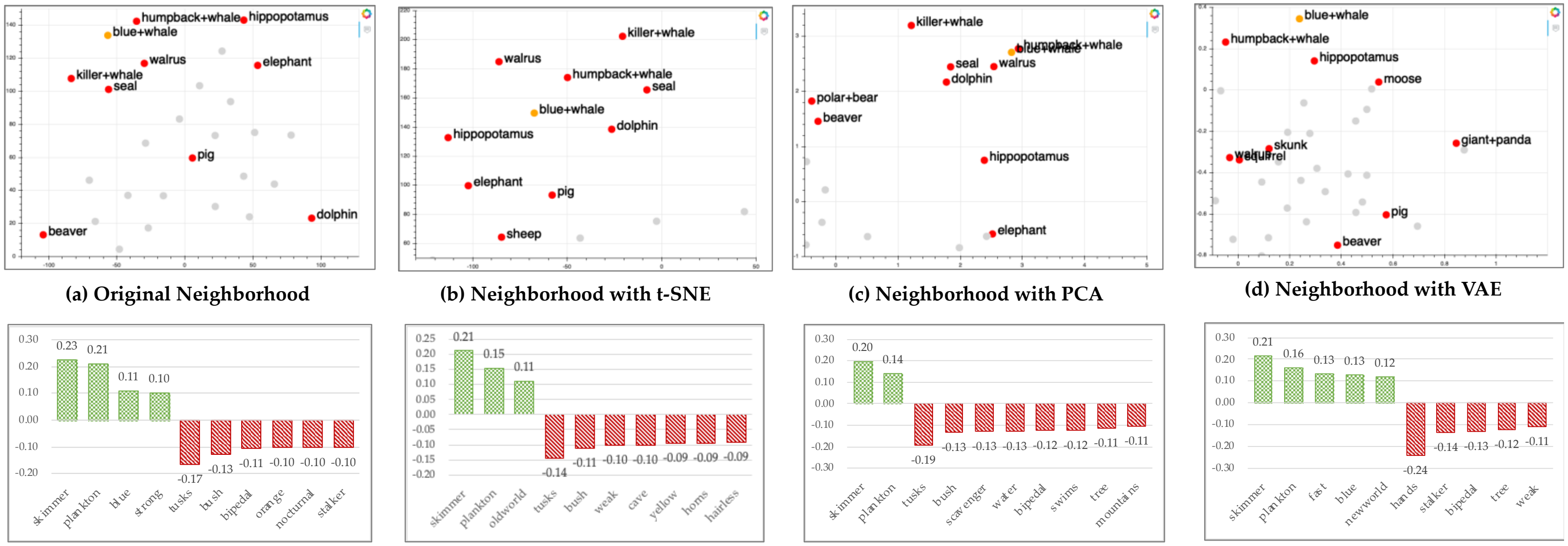
5.4.1 Experimental Setup

To ensure the **model & data-type agnostic** nature of the proposed algorithms, we compare the structural retention of 10 state-of-the-art linear and non-linear dimensionality reduction methods for 16 real-world datasets. The algorithms include popular techniques such as PCA [38], t-distributed Stochastic Neighbor Embedding (t-SNE) [39], Uniform Manifold Approximation and Projection (UMAP) [40], openTSNE [41], MDS [42], Isomap [25], Locally Linear Embedding (LLE) [43], Variational Autoencoder (VAE) [44], Local tangent space analysis (LTSA) [45], and KernelPCA [46]. The 16 high-dimensional datasets[18] used in our experiments belong to four different datatypes namely tabular, text, audio, and images. As 14 out of the 16 datasets were selected from the Kaggle[19] data repository, the Animals image dataset [9] and the UrbanSound8k[20] dataset were selected from related literature. For our experiments, we decided to consider labeled datasets only. Moreover, since dimensionality reduction only works with tabular numeric data [3],

[18] Breast Cancer, Adult, Wine Quality, Credit Card, Animals, MNIST, Flower17, Fashion-MNIST, UrbanSound8K, ESC50, GTZAN, Free-Spoken-Digits, Sentiment140, BBC-Text, SMS Spam Collection, Quora Question Pairs

[19] https://www.kaggle.com/datasets

[20] https://urbansounddataset.weebly.com/urbansound8k.html



(a) Original Neighborhood

(b) Neighborhood with t-SNE

(c) Neighborhood with PCA

(d) Neighborhood with VAE

(e) Original Feature Influence

(f) Feature Influence with t-SNE

(g) Feature Influence with PCA

(h) Feature Influence with VAE

Original	blue whale	humpback whale	walrus	seal	dolphin	hippopotamus	killer whale	elephant	pig	beaver
t-SNE	blue whale	humpback whale	walrus	dolphin	hippopotamus	pig	elephant	seal	killer whale	sheep
PCA	blue whale	humpback whale	walrus	seal	dolphin	killer whale	hippopotamus	elephant	polar bear	beaver
VAE	blue whale	humpback whale	hippopotamus	walrus	giant panda	moose	squirrel	pig	beaver	skunk

(i) Order to data-points in Neighborhood

	t-SNE	PCA	VAE
Discrepancy in Feature Influence	0.78	0.79	0.73
Discrepancy in Neighborhood Contents	0.10	0.10	0.20
Discrepancy in Neighborhood Order	0.70	0.30	0.70

Overall Local Divergences	
t-SNE	0.5268
PCA	0.3983
VAE	0.5434

Figure 5.3: Application of LAPS on the Animals dataset. Scatter plot in (a) shows the original neighborhood structure of data-point blue-whale, as the plots in (b to d) represent the same after running different dimensionality reduction algorithms. The bar-graphs in (e to h) show the feature influence explanations for the relative distances in the neighborhood of the point. In the above figure, each of the three components of the local divergence has a weight of 0.33 (i.e. the default weight).

we pre-processed the non-tabular datasets before the experiments. For example, as the audio datasets were converted to time-series data of sound amplitudes using the *librosa*[21] library, the text datasets were converted into word embeddings using *Word2Vec*[22] models.

5.4.2 Experimental Results

This section presents the experimental results for LAPS and GAPS. We divide the section into four parts based on the questions defined at the beginning this Section.

5.4.2.1 Applicability Analysis of LAPS and GAPS

To assess whether LAPS and GAPS fulfill their design requirements discussed in Section 5.2.1, now we present four case-studies applying the techniques on the image, text, audio, and tabular datasets.

Case Study 1: Image data - Animal Dataset

Fig. 5.3 shows an application of LAPS on the Animal dataset [9]. The Animal dataset contains 30,475 images and is composed of 85 numeric attributes and 50 animal classes. In Fig. 5.3, we explain the use of LAPS with a subset of the original dataset to enhance visual clarity. Here we select the data-point with the label *blue-whale* as our point of interest. Fig. 5.3.a. shows the 10 nearest-neighbors of *blue-whale* in the original 85-dimensional dataset on a two-dimensional projection. As shown in Fig. 5.3.a, in the original dataset some of the most closely related data points to *blue-whale* are the *humpback-whale*, *walrus, seal*, *dolphin*, and *killer-whale,* whilst points such as *pig* and *elephant* are also considered neighbors of *blue-whales* for resemblances in their values for the attribute *strong* (cf. Fig. 5.3.e). Showing our analysis with LAPS, Fig. 5.3.e explains the most influential attributes for the neighborhood structure shown in Fig. 5.3.a. As the green bars in Fig. 5.3.e represent a positive correlation of an attribute's contribution to the relative distances in the neighborhood, the red bars show the negatively influencing

[21] https://librosa.github.io/librosa/

[22] https://pypi.org/project/gensim/

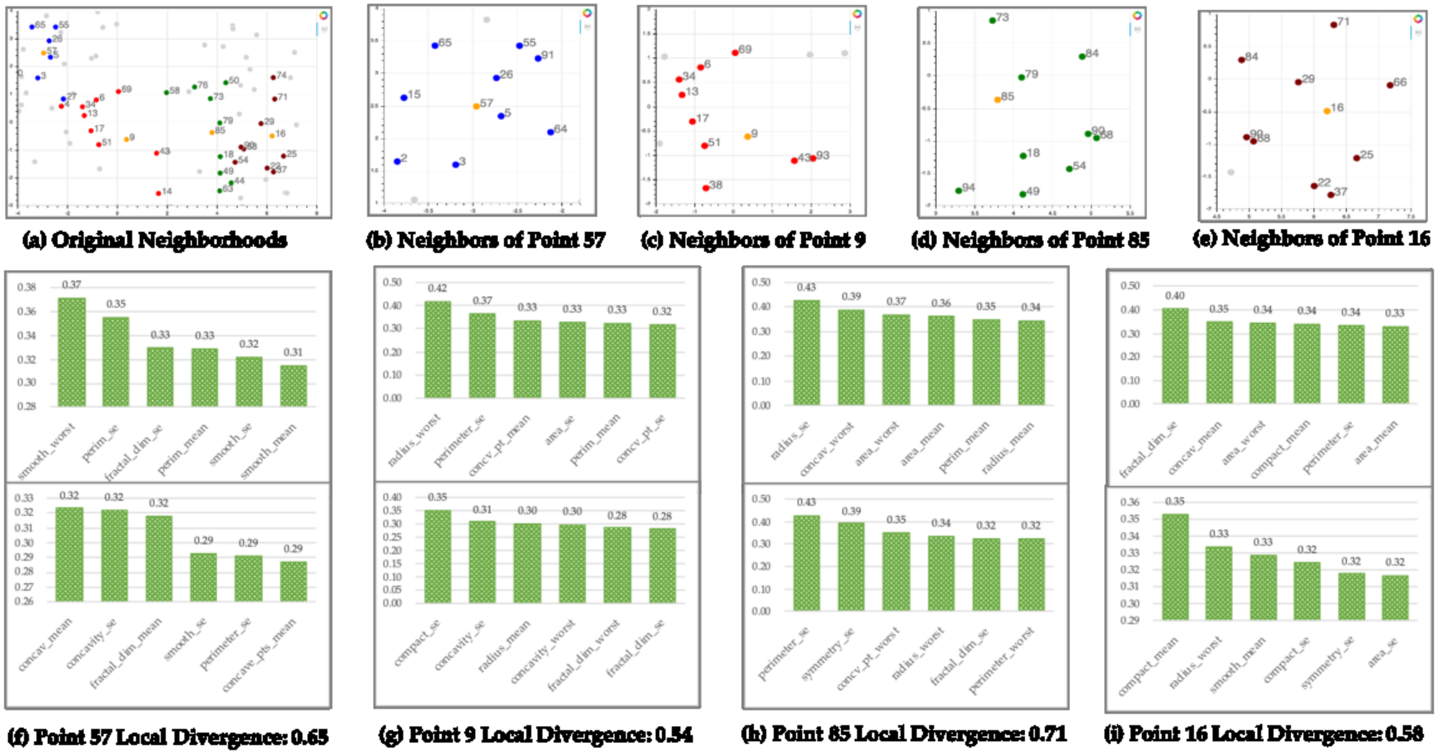
attributes for the same. From Fig. 5.3.e. it can be seen that for the neighborhood of *blue-whale* the most positively influential attributes are *skimmer, plankton, blue,* and *strong*. On the other hand, the most negatively influential attributes include *tusks* and *bush* that separate *blue-whale* from some of its closest neighbors such as the *elephant*. Fig. 3.e. also shows the attributes that are positively or negatively influential with similar magnitude. Due to this similarity, we consider them to be *highly correlated* with each other.

Fig. 5.3.b and 5.3.c show the neighborhoods for *blue-whale* after the application of t-SNE and PCA on the dataset respectively. As shown in the Figures 5.3.b and 5.3.c as well in Fig. 5.3.i, both t-SNE and PCA preserved 9 out of 10 neighbors for *blue-whale* in the embedding, while replacing the original neighbor *beaver* with *sheep* and *pig* with *polar bear* respectively. Nevertheless, the attribute influencing in the neighborhood structures are significantly changed for both t-SNE and PCA (cf. Fig. 5.3.f and Fig. 5.3.g). Looking into Figure 5.3.i it can be seen that VAE preserved 6 out of 10 neighbors in the projection and considered points such as *giant panda*, *moose*, *squirrel*, and *skunk* as the neighbors of *blue whale*. As a result, with VAE additional attributes such as *hands, tree,* and *weak* show significant negative contribution on the neighborhood structure.

Overall, Fig. 5.3 shows that all algorithms have performed poorly in terms of preserving the attribute influences in the embeddings. In terms of preserving the neighborhood components and neighborhood orders, t-SNE and PCA have performed relatively better than VAE. Hence, the neighborhood order contributed the most in the comparison of their local divergences. Here, PCA has performed much better than both the other algorithms, resulting in the lowest local divergence score among the three.

Case Study 2: Tabular data - Breast Cancer Dataset

Our second case study focuses on the Breast Cancer dataset [9] that classifies tumors into malignant and benign. This tabular numeric dataset is composed of 32



(a) Original Neighborhoods

(b) Neighbors of Point 57

(c) Neighbors of Point 9

(d) Neighbors of Point 85

(e) Neighbors of Point 16

(f) Point 57 Local Divergence: 0.65

(g) Point 9 Local Divergence: 0.54

(h) Point 85 Local Divergence: 0.71

(i) Point 16 Local Divergence: 0.58

Original	57	5	3	65	55	26	4	27	0	6
Embedding	57	5	26	15	64	3	91	65	55	2
Original	9	17	43	14	6	13	51	69	4	34
Embedding	9	51	43	17	38	93	69	6	13	34
Original	85	49	79	44	18	73	50	58	63	76
Embedding	85	79	18	73	84	90	68	54	94	49
Original	16	22	29	25	37	68	54	71	74	90
Embedding	16	29	25	66	22	68	37	90	71	84

Original Distance Ordering: [[16, 57], [16, 9], [9, 57], [85, 9], [16, 85], [85, 57]]
Embedding Distance Ordering: [[16, 57], [85, 57], [9, 57], [85, 9], [16, 85], [16, 9]]

(j) Original Vs. Embedding Neighborhoods

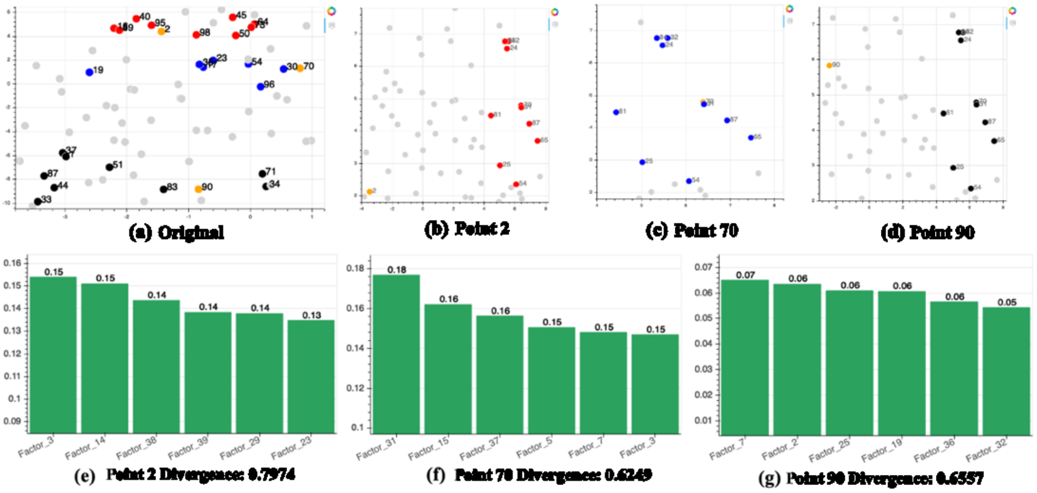
Global Divergence with default weights (i.e., w_1, w_2, w_3 = 0.33) = 0.32074

Figure 5.4: Application of GAPS on the Breast Cancer dataset. As the scatter plot in (a) shows the original neighborhoods of the four data points, the plots in (b to e) show their neighborhoods using t-SNE. In the scatter plots, the points selected for analysis are colored in 'orange' and the neighborhoods of the four points 57, 9, 85, and 16 are colored in 'blue', 'red', 'green', and 'maroon' respectively. The bar-graph pairs (f to i) show the feature influence explanations of the points. The tabular representation in (j) shows the original versus the projected neighborhoods of the selected points.

attributes and 569 data points. Fig. 5.4 primarily shows the utility of GAPS with the breast cancer data using *t-SNE* as the used dimensionality reduction technique. The analysis begins with an interactive selection of four[23] non-redundant data points (with indexes: 9, 16, 57, 85) from the dataset. In the original neighborhood structure of these points (cf. Fig. 5.4.a) we can see the points are far away from each other on the manifold and have no overlaps in neighbors.

From Figures 5.4.b, 5.4.c, 5.4.d, 5.4.e, and 5.4.j it can be seen that after t-SNE, as two of the original neighbors of points 9 and 16 are replaced by two different points in the embedding, for the points 57 and 85 four and five points are replaced respectively. In terms of attribute influences, in the original neighborhood of point 57 (Fig. 5.4.f), the most influencing feature *smoothness_worst* is replaced by *concavity-mean* in the embedding. Moreover, some of the highly influencing attributes such as *fractal_dimension_mean*, *perimeter_mean,* and *smoothness_mean* are not included in the group of six most influential attributes in the embedding neighborhood. Similar discrepancies (cf. Figures 5.4.g, 5.4.h, 5.4.i) in feature influences are noticed for all the points. Nevertheless, from all the original attributes influence bar-graphs, it can be seen that for the chosen points attributes such as *perimeter-mean*, *area-se*, *fractal_dimension_mean*, and *area-mean,* are *highly correlated* attributes in the neighborhoods of all the four selected data points. Along with the divergence in the preserved global structure, in Fig. 5.4 the analyst can also see the disagreement in the order of preserved relative distances between the four selected data points in the original and their low-dimensional embedding. This disagreement shows that t-SNE has failed to preserve the original relative proximities among the four selected points in the embeddings. We think the reason for this is, t-SNE being a locally focused dimensionality reduction technique has preserved 50% to 80% of the nearest neighbors for each of the selected points whilst

[23] Here we select only four data points only to enhance visual clarity. A detailed discussion on appropriate budget size is presented in Section 5.2.5



(a) Original (b) Point 2 (c) Point 70 (d) Point 90

(e) Point 2 Divergence: 0.7974 (f) Point 70 Divergence: 0.6249 (g) Point 90 Divergence: 0.6557

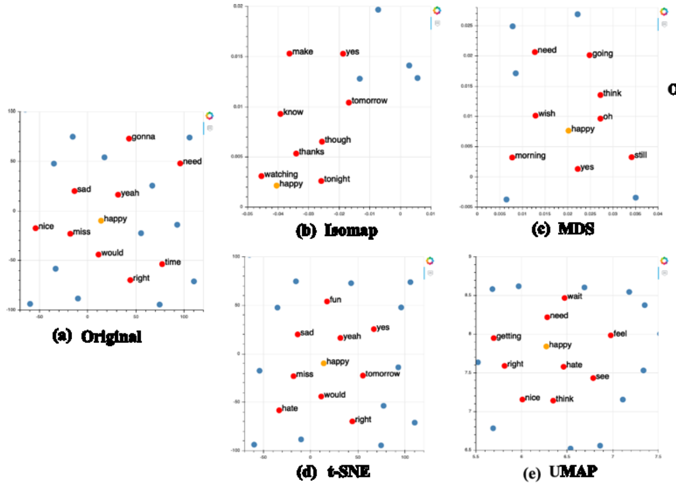
Figure 5.5: LAPS and GAPS on the UrbanSound8K dataset. The scatter plot in (a) shows the original neighborhood of the three points as the plots in (b, c, and d) show the neighborhoods of the points after t-SNE. The bar-graphs below show the feature influence explanations for the three points after t-SNE. In the scatter, plots the neighborhoods of points 2, 70, and 90 are presented using colors “red”, “blue” and “black”.

disrupting the global distances among the neighborhoods of the four points in the embedding.

Case Study 3: Audio data – UrbanSound8k Dataset

The UrbanSound8k[24] dataset contains 8732 audio files containing sounds belonging to 10 classes namely: air_conditioner, car_horn, children_playing, dog_bark, drilling, enginge_idling, gun_shot, jack_hammer, siren, and street_music. After pre-processing the data, we extracted 39 features in terms of time-slices 0 to 38. Figure 5.5 combines our analysis of the transformation of the UrbanSound8k dataset using t-SNE on multiple data-points. Fig. 5.5.a shows the selected data-points with the indexes 2, 70, and 90 in the original dataset, with their original labels as enginge_idling, dog_bark, and drilling respectively. As the figures 5.5.b, 5.5.c, and 5.5.d show their neighborhood structures in the embedding, LAPS shows the disagreements of the four neighbors for data-point index 2 and two neighbors for both indexes 70 and 90 from the original dataset. An analysis of attribute

[24] https://urbansounddataset.weebly.com/urbansound8k.html



Order of neighbors in the original space vs. embeddings:

Original	Isomap	MDS	tSNE	UMAP
thanks	watching	make	today	nice
twitter	thanks	time	thanks	thanks
make	know	hope	though	watching
nice	make	much	yes	make
still	tonight	like	cant	hate
got	though	know	know	hey
getting	oh	go	nice	sad
yes	tomorrow	week	wish	know
right	yes	morning	night	yes
Divergence	**0.5215**	**0.5919**	**0.5411**	**0.4874**

(f)

Figure 5.6: Application of LAPS on the Sentiment140 Dataset. The scatterplot (a) shows the original neighborhood of the word happy in the word-embedding as the scatter plots in (b to e) show the same after executing four different dimensionality reduction algorithms. Whereas, (f) shows a comparison between the neighborhood order for the point happy in the original data and the embeddings.

contributions for the (data)points shows that, as for engine-idling (cf. Fig 5.5.e), the time-slices at the beginning (i.e., slice 8, 9, and 10) have more positive influence than others. Similarly, for a dog-bark (cf. Fig. 5.5.f) the time slices in the middle (e.g., 25, 27, 29) make a more significant impact than others. For drilling (cf. Fig. 5.5.g) however, time slices amplitudes are more uniformly distributed throughout the 0 to 38 attributes. In par with the number of misplaced neighbors in the neighborhoods, Fig. 5.5 shows the local divergence of the data-point 2 (engine_idling) to be the highest among the three. This means that the algorithm has preserved the locality of the points 70 and 90 better than the point 2 and this has compromised the overall global divergence of the projection.

Case Study 4: Text data – Sentiment140 Dataset

Our second case study presents the application of LAPS on Stanford's Sentiment140[25] dataset with 1.6 million tweets, where each tweet is associated with a numeric label of 0, or 2, or 4 representing negative, neutral, and positive sentiments in the tweet. To pre-process the text data for dimensionality reduction,

[25] https://www.kaggle.com/kazanova/sentiment140

after removing the stop-words, hashtags, URLs, and twitter usernames, we converted the original dataset into word embeddings using Word2Vec. Fig. 5.6 shows our application on LAPS on the Stanford140 dataset with a zoomed-in version of the embedding to visualize our analysis. In Fig. 5.6, using LAPS we compare the preserved local neighborhood structure for the point *happy* using four different dimensionality reduction algorithms (cf. Fig. 5.6.b, 5.6.c, 5.6.d, 5.6.e). From Fig. 5.6, it can be seen that UMAP (cf. Fig. 5.6.e) has preserved most of the 10 original (cf. Fig. 5.6.a) neighbors (i.e., 5 out of 10 neighbors) for the word *happy* after the transformation. Whereas, MDS has preserved the least number of original neighbors (i.e., only 1 out of 10) in the embedding with both Isomap and t-SNE preserving only 3 of the 10 original neighbors for the word *happy*. Nevertheless, the divergence score of t-SNE is higher than Isomap as the order of the neighbors is better preserved using Isomap than t-SNE. That is, as shown in Fig 5.6.a, the neighbor *yes* for the word *happy* comes after the neighbor *makes* in the original neighborhood and Isomap has preserved this order as depicted in Fig. 5.6.b. Whereas, even though the neighbor *nice* comes before *yes* in Fig 5.6.a, t-SNE has reversed their order in the embedding presented in Fig. 5.6.c.

5.4.2.2 Selection of Appropriate Algorithm using LAPS and GAPS

In this section, we present our assessment of LAPS and GAPS for their ability to instill confidence in a single projection within a group of projections and to assist with the selection of an appropriate dimensionality reduction algorithm for any dataset. Table 5.1 presents the local divergence scores obtained using LAPS for all the 16 datasets. In Table 5.1, we compute the mean local divergence scores for 100 random data points from every dataset. To validate the local divergences obtained from LAPS, following the guidelines of Maaten et al. [3], we use the generalization error of the 1-nearest neighbor classification algorithm. 1-NN generalization error being a popular metric [3] for the validation of retained local structure in an embedding. In Table 5.1, we present a side-by-side comparison between the divergence scores obtained from LAPS and the 1-NN generalization errors for the 10 algorithms. Similar to Maaten et al. [3], we compute the generalization errors

Table 5.1: Divergence Scores Of Dimensionality Reduction Algorithms Using Laps Vs. 1-NN Generalization Errors

		LAPS Divergence Scores										**1-NN Generalization Errors**									
Datasets	Type	tSNE	PCA	UMAP	oTSNE	MDS	ISMP	LLE	KPCA	LTSA	VAE	tSNE	PCA	UMAP	oTSNE	MDS	ISMP	LLE	KPCA	LTSA	VAE
Animal	Img	0.472	0.308	0.298	**0.275**	0.334	0.457	0.492	0.308	0.432	0.417	0.418	0.292	**0.263**	0.298	0.353	0.342	0.362	0.290	0.327	0.332
MNIST	Img	**0.357**	0.399	0.458	0.362	0.382	0.362	0.386	0.477	0.362	0.391	0.225	0.388	0.545	0.265	**0.213**	0.622	0.604	0.846	0.323	0.47
FLOWER17	Img	0.560	0.556	0.559	0.560	0.407	**0.375**	0.406	0.407	0.523	0.544	0.632	0.812	0.716	0.628	0.791	**0.577**	0.813	0.818	0.880	0.742
F-MNIST	Img	0.490	0.527	**0.419**	0.489	0.525	0.521	0.528	0.427	0.546	0.443	**0.249**	0.529	0.320	0.255	0.493	0.450	0.529	0.357	0.640	0.387
Brst Cancer	Tabl	**0.469**	0.667	0.535	0.497	0.526	0.593	0.662	0.56	0.66	0.524	**0.046**	0.088	0.056	0.047	0.090	0.056	0.070	0.088	0.903	0.159
Magic	Tabl	**0.313**	0.61	0.446	0.313	0.613	0.646	0.711	0.612	0.541	0.398	**0.186**	0.369	0.285	0.186	0.338	0.372	0.380	0.336	0.350	0.247
Wine Qlty	Tabl	**0.503**	0.537	0.537	0.603	0.604	0.56	0.57	0.537	0.504	0.518	**0.404**	0.415	0.412	0.495	0.470	0.452	0.541	0.415	0.545	0.46
Crdt Card	Tabl	**0.46**	0.587	0.592	0.581	0.527	0.529	0.662	0.587	0.556	0.507	**0.258**	0.296	0.311	0.293	0.305	0.289	0.370	0.290	0.280	0.281
ESC50	Aud	**0.597**	0.664	0.629	0.631	0.696	0.63	0.695	0.664	0.632	0.641	**0.705**	0.898	0.826	0.717	0.907	0.935	0.895	0.898	0.886	0.887
Urban8k	Aud	**0.329**	0.596	0.463	0.329	0.663	0.529	0.629	0.596	0.658	0.415	**0.122**	0.486	0.276	0.128	0.448	0.524	0.594	0.486	0.840	0.413
Spkn Digits	Aud	**0.496**	0.694	0.591	0.502	0.688	0.588	0.69	0.682	0.601	0.577	**0.005**	0.273	0.009	0.005	0.191	0.122	0.089	0.273	0.286	0.122
GTZAN	Aud	0.495	0.487	0.46	0.523	**0.457**	0.523	0.558	0.504	0.558	0.624	0.459	0.326	0.256	0.454	**0.299**	0.311	0.324	0.345	0.425	0.351
SMS Spam	Txt	**0.474**	0.508	0.513	0.499	0.475	0.475	0.508	0.508	0.474	0.499	**0.256**	0.325	0.544	0.260	0.345	0.360	0.311	0.300	0.446	0.379
Quora	Txt	**0.57**	0.57	0.537	0.558	0.603	0.536	0.603	0.57	0.57	0.535	**0.021**	0.211	0.199	0.022	0.054	0.164	0.254	0.337	0.185	0.154
BBC-Text	Txt	**0.447**	0.678	0.546	0.571	0.571	0.646	0.621	0.675	0.685	0.688	**0.149**	0.270	0.449	0.149	0.356	0.225	0.297	0.315	0.333	0.27
Sntmt140	Txt	0.575	0.572	**0.502**	0.593	0.569	0.638	0.615	0.567	0.571	0.603	0.658	0.717	0.682	**0.655**	0.745	0.803	0.811	0.752	0.699	0.721

Description of Acronyms in Table 5.1: Img: Image, Tabl: Tabular, Aud: Audio, Txt: Text, oTSNE: openTSNE, ISMP: Isomap, F-MNIST: Fashion MNIST, Brst Cancer: Breast Cancer, Wine Qlty: Wine Quality, Crdt Card: Credit Card, Spkn Digits: Spoken Digits, Sntmt140: Sentiment140

Note: In the above table, the left-hand side shows an average of LAPS divergence scores for 100 data points from the 16 datasets using 10 different dimensionality reduction algorithms. The right-hand side shows the 1-NN generalization errors using the same algorithms on the same datasets. The algorithms with lowest local-divergences and 1-NN gen. errors are highlighted in bold & red. In both the sides, the column representing the algorithm with the lowest local divergence and 1-NN gen. error for more than 50% of the datasets are highlighted in grey.

Table 5.2: Statistical Comparison of Suggested Algorithms

Dataset	**Suggested algorithms**		**p-value paired t-test**	
	LAPS	1-NN GE	Dim-1	Dim-2
Animal	oTSNE	UMAP	0.404	0.667
MNIST	tSNE	MDS	0.826	0.847
FashionMNIST	UMAP	tSNE	0.186	0.108
Sentiment140	UMAP	oTSNE	0.088	0.095

Note: GE: Generalization Error, Dim-1: first target dimension obtained from dimensionality reduction, Dim-2: second target dimension after dimensionality reduction. For the paired t-tests, the threshold α is considered to be 0.05 (i.e., the most commonly used value for α [27]). The results show, all obtained p-values are more than α accepting the null hypothesis that embeddings are not statistically significantly different from each other.

using leave-one-out cross-validation. The results in Table 5.1 show that in 75% of cases LAPS agree with the 1-NN generalization error scores on the algorithm that preserved most of the local structure. For the remaining 25% cases, where the lowest divergences from LAPS do not agree with lowest 1-NN generalization errors, we perform paired t-tests[26] [47] to compare the embeddings obtained from the algorithms suggested by LAPS and 1-NN generalization error. As shown in Table 5.2, the p-values obtained from statistical significance analysis show no significance differences between the results of the two techniques. Hence, from Table 5.1 it can be seen that multiple iterations of LAPS can help users to select an algorithm that has preserved most of the local structure of the original dataset in its embedding.

Next, we graphically analyze the comparisons between the local divergence scores and the 1-NN generalization errors discussed in Section 5.2.2 of the main manuscript. The table presented as a part of this Section (cf. Table 5.1, main manuscript) performs a side-by-side comparison among the local divergence scores obtained from 16 different datasets using 10 dimensionality reduction algorithms

[26] The paired t-test [47] is the most common parametric statistical test to compare the mean of two sample populations.

with 1-NN generalization errors of the same datasets transformed using the same 10 algorithms. As discussed earlier, The algorithms include popular techniques such as PCA [38], t-distributed Stochastic Neighbor Embedding (t-SNE) [39], Uniform Manifold Approximation and Projection (UMAP) [40], openTSNE [41], MDS [42], Isomap [25], Locally Linear Embedding (LLE) [43], Variational Autoencoder (VAE) [44], Local tangent space analysis (LTSA) [45], and KernelPCA [46]. The 16 high-dimensional datasets[27] used in our experiments belong to four different datatypes namely tabular, text, audio, and images.

The results of the graphical analysis are presented in Figure 5.7. The main idea of the graphical analysis is to visually analyze the utility of local divergence scores for their assistance with the selection of the most appropriate dimensionality reduction algorithms. The figure shows, overall local divergence scores, and 1-NN generalization errors show the same patterns for most of the datasets. The best examples are the Animals dataset, the Credit Card dataset, the Wine Quality dataset, and the ESC50 dataset. In a majority of the cases, the algorithms such as t-SNE and UMAP that have the lowest local divergence in most datasets also have the lowest 1-NN generalization errors. In the case of some datasets, such as Breast Cancer, Free-Spoken-Digits, Quora Question Pairs, the gap between the algorithm with the highest 1-NN generalization error and the lowest 1-NN generalization error is much bigger than the same for local-divergence scores. One could argue that in this case, 1-NN generalization error distinguishes the best and worst-performing algorithms better than the local-divergence scores. However, since we have only considered the default weights for different components of the local divergence scores in Figure 5.7, as discussed in Section 5.2.5, we think that altering the weight combinations might show the results more clearly and increase the gaps between the divergence scores for the best and worst-performing algorithms.

[27] Breast Cancer, Adult, Wine Quality, Credit Card, Animals, MNIST, Flower17, Fashion-MNIST, UrbanSound8K, ESC50, GTZAN, Free-Spoken-Digits, Sentiment140, BBC-Text, SMS Spam Collection, Quora Question Pairs

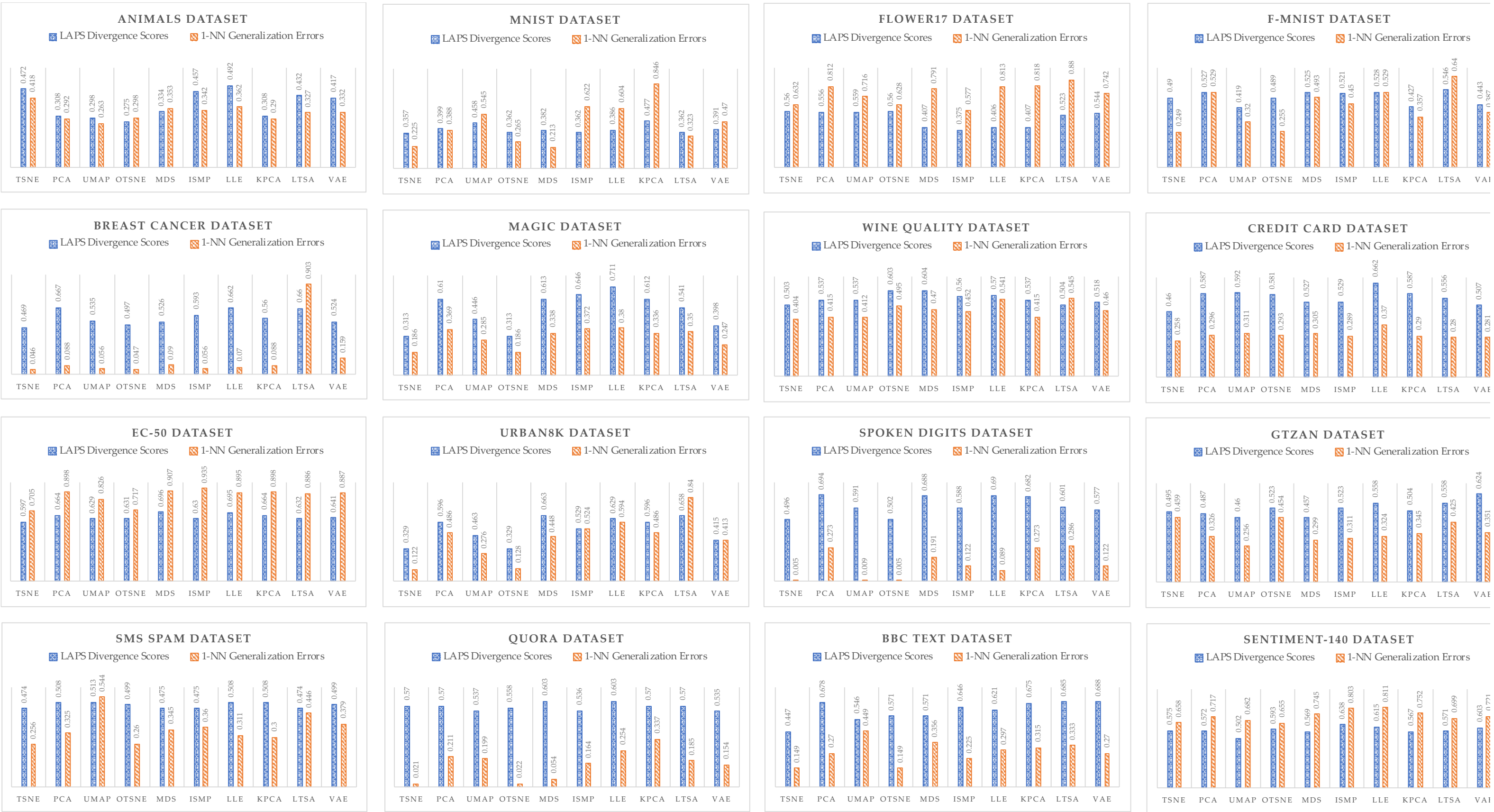


Figure 5.7: Graphical Analysis of Local Divergence Scores Vs. 1-NN Generalization errors of 16 real-world datasets using 10 dimensionality reduction algorithms. The bars on the left (blue color) represent the LAPS divergence scores, while the bars on the right (orange color) represent 1-NN generalization errors.

Table 5.3: Divergence Scores of Dimensionality Reduction Algorithms using GAPS

Datasets	Type	tSNE	PCA	UMAP	oTSNE	MDS	ISMP	LLE	KPCA	LTSA	VAE
Animal	Image	0.319	0.279	0.368	0.365	**0.198**	0.323	0.324	0.279	0.336	0.393
MNIST	Image	0.401	0.262	0.399	0.361	**0.207**	0.381	0.614	0.376	0.427	0.345
FLOWER17	Image	0.414	0.403	**0.316**	0.331	0.321	0.366	0.425	0.365	0.427	0.412
Fashion MNIST	Image	0.312	0.362	**0.206**	0.287	0.244	0.381	0.394	0.283	0.377	0.346
Breast Cancer	Tabular	0.290	0.250	0.245	0.266	**0.242**	0.275	0.251	0.250	0.258	0.317
Magic	Tabular	0.294	0.282	0.309	0.433	**0.265**	0.304	0.799	0.277	0.499	0.377
Wine Quality	Tabular	0.328	**0.245**	0.303	0.304	0.279	0.306	0.279	0.245	0.431	0.391
Credit Card	Tabular	0.303	0.261	0.275	0.419	0.430	0.607	0.495	0.440	**0.250**	0.312
ESC50	Audio	0.338	0.864	**0.286**	0.348	0.303	0.288	0.313	0.291	0.864	0.621
UrbanSound8k	Audio	0.291	0.662	**0.276**	0.318	0.283	0.731	0.254	0.244	0.244	0.504
Spoken Digits	Audio	0.403	0.335	0.397	0.403	**0.308**	0.336	0.345	0.335	0.418	0.482
GTZAN	Audio	0.382	0.331	0.425	0.447	0.447	0.332	**0.283**	0.333	0.338	0.419
SMS Spam	Text	0.694	0.330	0.314	0.300	0.320	0.354	0.370	0.330	**0.315**	0.329
Quora	Text	0.407	0.313	0.313	0.436	**0.301**	0.322	0.300	0.313	0.303	0.318
BBC-Text	Text	0.400	0.328	0.295	**0.283**	0.324	0.302	0.332	0.328	0.351	0.362
Sentiment140	Text	0.260	0.293	**0.237**	0.288	0.276	0.265	0.284	0.293	0.292	0.304

Note: The above table, presents the average GAPS divergence scores for 10 random data-points selected 100 times from the 16 datasets. The algorithms with lowest global divergences highlighted in bold & red. The idea is to evaluate whether GAPS produces an acceptable global-divergence score for the chosen datasets that can help with selecting appropriate dimensionality reduction algorithm for the data.

Next, we present the results of applying GAPS on the 16 real-world datasets used in our experiments and compare the global divergence scores for 10 dimensionality reduction algorithms. In Table 5.3, we present the global divergence scores obtained from GAPS. For the scores in Table 5.3, we selected 10 data-points for each dataset 100 times and computed the mean of their global divergences. Although reconstruction errors of dimensionality reduction algorithms could be used to validate the preserved global structure in embeddings (i.e., the global divergence scores in Table 5.3), we do not compute the reconstruction errors of the algorithms. The primary reason being, as pointed by Maaten et al. [3] there is no way to compute accurate reconstruction errors for real-world datasets, as their true underlying manifold is unknown. Overall, from Table 5.3 it can be seen that GAPS allows for a user-driven quantification of the global structural retention in embeddings. As shown in Table 5.3, for most of the datasets UMAP and MDS have preserved the maximum of the global structure in comparison with the other algorithms. The user study presented in the main structural quality helps users in making decisions regarding the most appropriate dimensionality reduction algorithm in a given scenario.

One could argue on the superiority of LAPS over 1-NN generalization error for evaluating the preserved local structure in an embedding [3]. However, the computation of the 1-NN generalization error is black-box to its users, as users cannot interact with the computation of the metric. Whereas, LAPS allows users to interpret and interact with each component of the metric local-divergence. Moreover, by allowing users to define weights for different components of the divergence calculation, users can decide the importance of the feature influences over the neighborhood structures in divergence calculation, based on their domain expertise.

5.4.2.3 Evaluation with Human Subjects

To assess the utility of LAPS and GAPS in real-world data analysis, we have performed a user study with 10 human subjects. In this section, we present an overview of our study that includes detailed information regarding the study questions, gathered insights, and analysis of obtained information.

The participants of our study included both industry professionals and Ph.D. candidates with strong analytical backgrounds. During our study, the subjects were divided into two groups namely novice and expert participants. The novice-group contained 6 individuals with no prior knowledge of dimensionality reduction. Whereas, the expert-group comprised 4 individuals with moderate experience with dimensionality reduction techniques. In this study, we asked the subjects to analyze the Wine Quality dataset that classifies 4898 wine samples into 10 quality categories using 12 attributes. This section is primarily divided into four sub-sections that focus on the four primary steps that were followed during our study with the subjects.

Step 1 - Define Study Objectives and Methods

The first step of the study was conducted in three phases. At first, the study participants were briefed about the characteristics of the dataset under investigation, in detail. During this briefing, a set of graphical overviews of the dataset was presented to the participants. These graphical representations included the following: (1) tabular representations demonstrating the dimensions of the original dataset (i.e., 4898 x 12), (2) histograms of value distributions of each of the 12 attributes in the data, (3) bar-charts showing correlations of individual attributes with the wine quality labels, and (4) balances of the classes (i.e., the quality labels) in the dataset. Next, the users were presented with three 2D scatter plots of the embeddings of the Wine Quality dataset obtained by executing t-SNE, Isomap, and UMAP on the data. The interactive scatter plots allowed users to zoom into the neighborhoods of the data points and using mouse-hover operations to determine the indexes and labels of the individual points in the plots. The users were given 10 minutes to observe these scatter plots and note down their insights, but not yet disclose them with other study participants or the conductors of the study. Finally, the users were presented with an overview of the objectives and expected outcomes of LAPS and GAPS. During the process, some sample outcomes of the processes on other open-source datasets were presented to the participants. Also, the subjects were given a walkthrough of how to execute the two algorithms on the different projections of the Wine Quality dataset. The first and the third phases of this stage were interactive, where the participants were encouraged to ask about any confusion regarding the dataset or the study process.

Step 2 - Conduct User Evaluation

At this step of the study, the users were allowed to perform independent analysis on the Wine Quality dataset. During this time, no study participants were allowed to interact with the conductors or any other participant in the study. Each individual was given 10 minutes to analyze the Wine Quality Dataset and execute the algorithms LAPS and GAPS on their selected data points. During their analysis, the percipients were asked to execute each of the two algorithms at least once and compare the three embeddings of the Wine Quality dataset that were presented to them. Also, the participants were asked to keep track of their decisions regarding their interpretation of the structural quality of the embeddings, the required analysis time, and their choice of the best performing algorithms for the Wine Quality dataset.

Step 3 - Gather User Feedback

After the independent evaluation of the algorithms, each participant was individually asked the four questions that were mentioned in the original manuscript. The questions included: (i) Can the two techniques efficiently explain the structural preservations in the embeddings? (ii) Can executing the two techniques enhance user-trust on the embeddings? (iii) Can the techniques help users with decision making regarding the best performing algorithm? (iv) Can the techniques reduce the analytical time?

When answering the above questions, the subjects were asked to answer with a firm *yes* or *no* at first and then to elaborate on their answer. For example, in case of the questions regarding the ability of the two algorithms to explain structural

Table 5.4: User-agreement Analysis on LAPS and GAPS

Analytical Aspects	**Fleiss' Kappa**	
	Novice	**Experts**
Efficiently Explains	0.66	0.47
Enhances Trust	0.24	0.33
Helps with decision making	-0.17	0.11
Reduces analytical time	1.00	1.00

Note: The table above summarizes the results of our user study. Here we analyze four different aspects of utility with LAPS and GAPS using 10 human subjects. Among them 6 are novice and 4 were expert participants.

preservations or enhance user-trust in the embeddings, the participants were asked to describe their understanding of the structural quality of each of the embeddings and also their rationale behind selecting one embedding over another. Moreover, for the questions regarding the reduced analytical time and assistance of the two techniques in instilling confidence on any specific embedding over another and helping with the decision of the most appropriate algorithm for the given dataset, the participants were asked to discuss their observed differences between not using LAPS and GAPS (cf. Section 2.4.1) and using the two algorithms to analyze the same projections. At this time, the subjects were also asked for any identified limitations in the two techniques and their suggestions for improvements, if there were any.

Step 4 - Analyze and Summarize Insights

Upon collecting the feedbacks, the conductors of the study (i.e., the authors) analyzed the obtained positive and negative responses from both the novice and expert participants and computed the agreements between their responses To quantitatively summarize the results of this study, following the idea of Lewis et al. [7] we computed Fleiss' Kappa consistency measure κ to assess the participant agreements on the feedbacks for LAPS and GAPS. The value of κ ranges from -1 to +1, where -1 represents no observed agreement, +1 signifies perfect agreement and 0 denotes agreement due to random chance. The results of our study are summarized in Table 5.4. The table shows that in terms of reduction of analytical time both novice and expert users had a perfect agreement about the utility of LAPS and GAPS over manual analysis. Besides, due to their prior experience with dimensionality reduction, the expert users could trust the results of the two techniques more easily (i.e., κ=0.33) than the novice users (i.e., κ=0.24). Regarding the ability of the techniques to provide an efficient explanation of embedding quality, both novice and expert users had a moderate agreement with κ of 0.66 and 0.47 respectively. In terms of decision making, approximately 50% of both novice and expert users agreed on the utility of LAPS and GAPS making κ close to 0.

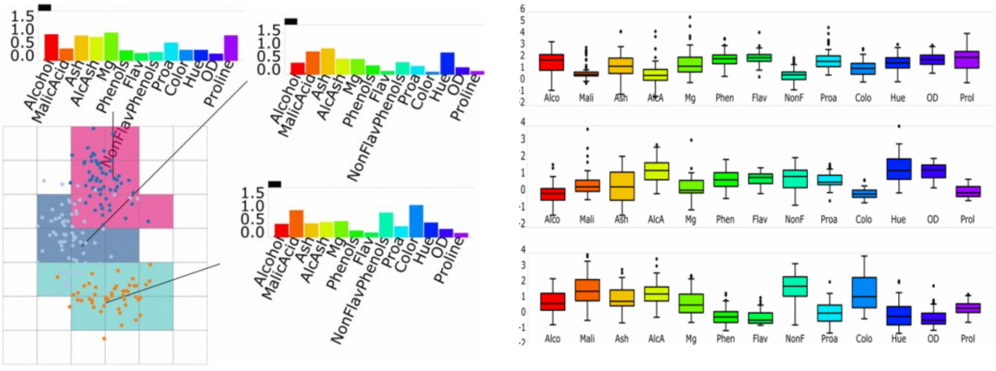
Overall, the participants agreed on the utility of the two techniques in all four aspects of our analysis with some suggestions for improvements. For example, only 33% of the novice participants altered the relative weights during their analysis of the local and global divergence scores. At the same time, only 50% of the expert subjects could easily understand the discrepancies in all the individual components of local and global divergence scores, whilst the rest needed more assistance. As a proposed solution for both the problems, both novice, and expert users have proposed to integrate the LAPS and GAPS procedures as a part of a visual interactive framework. We envision this integration as future work.

5.4.2.4 Detailed Comparison with Related Research

In this section, we present detailed comparisons of our proposed methods with the two most closely related research. At first, we compare our approach with the approach proposed by Pagliosa et al. [21] followed by a comparison with the approach SIRIUS proposed by Dowling et al. [9]. In both the approaches, we have used the examples shared by the original authors in their original manuscripts and compared the outcomes of our approaches in the same scenarios.

Comparison with Pagliosa et al.

The approach presented by Pagliosa et al. [21] attempts to identify the contribution of each attribute in the similarity (i.e., proximity) among the data-points that belong to the same label. For example, Figure 4 shows the application of the approach on the Wine Quality dataset presented by the original authors [21]. The dataset is



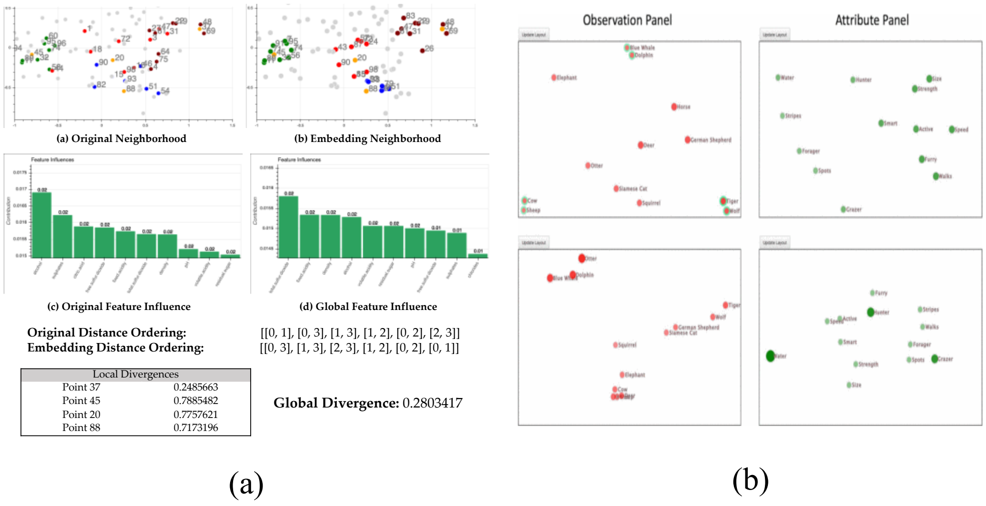
(a) Vector-based clustering & Attribute Variances

(b) Box-plots showing interquartile ranges and outliers

Figure 5.8: Analysis of the Wine Quality dataset by the approach presented by Pagliosa et al. [21]. The figures are taken from Figures 15 and 16 of the original manuscript of Pagliosa et al. [21].

composed of 178 instances that are classified into 3 wine categories using 13 attributes. In this approach, the authors have identified the variance within the attributes to distinguish their contributions on the relative proximities among data-points in a selected region. In Fig. 5.8.a, the three different regions of interest, representing data-points belonging to 3 different wine categories, are highlighted within the uniform grids. The bar charts show the variance of the attributes within the grids highlighted in the same color. Whereas, the box-plots in Fig. 5.8..b presents the interquartile range, and outliers, per dimension. The authors identify the attributes with the highest variance to be the most contributing attributes in that region and suggest users that the attributes with a large number of outliers are not relevant enough to represent the data-points belonging to different class labels.

Our approach looks into the same research question however from a different perspective. Here, we combine the impacts of both neighborhood structure and attribute influences and identify the algorithms that can preserve most of both. A user-driven computation of local and global divergences helps users to quantify the overall quality of the preserved local and global structure. For example, as shown in Figure 5.9.a, for the Wine quality dataset, we allow users to analyze each data-points as well as a region of data-points in the embeddings. However, the class



Original Distance Ordering: [[0, 1], [0, 3], [1, 3], [1, 2], [0, 2], [2, 3]]
Embedding Distance Ordering: [[0, 3], [1, 3], [2, 3], [1, 2], [0, 2], [0, 1]]

Local Divergences	
Point 37	0.2485663
Point 45	0.7885482
Point 20	0.7757621
Point 88	0.7173196

Global Divergence: 0.2803417

Figure 5.9: (a) Analysis of the Wine Quality dataset using GAPS. (b) An Analysis of the Animals dataset by Dowling et al. [12]. The figure is taken from Figure 5 of the original manuscript of Dowling et al.

labels of the data-points are not considered relevant in our approach as we look deeper into the pairwise distances within the neighborhoods of the points irrespective of their class labels. Next, our proposed techniques allow users to visualize the neighborhoods along with the influences of the original data attributes in their structural formation of the neighborhood. Here the feature influences are computed as the correlation of the feature contribution scores with the relative proximities among the data-points. Also, to better understand the impacts of each component of the local and global divergence scores, users are allowed to alter the relative weights of the neighborhood structure and the attribute influences when computing the metrics. Overall, our approach focuses on a diverse group of aspects of structural preservation and allows for interactive exploration of embeddings.

Comparison with SIRIUS

In this section, we compare our proposed approach with the work of Dowling et al. [9]. The work focuses on dual, symmetric, observation, and attribute level interactions of the low-dimensional embeddings. In Figure 5.8.b we show the original example presented by the authors [9] in their chapter. In this example, the authors analyzed the Animals dataset [9]. SIRIUS allows users to select multiple pairs of data-points and identifies the most influential attributes that distinguish the different pairs of points. For example, in Fig. 5.9.a the authors have analyzed the differences between the point pairs: blue-whale, dolphin with cow, sheep and tiger, wolf. Their attribute panel shows that the attributes water, hunter, and grazer are responsible for the relative differences in the point pairs. In their work, the authors have primarily focused on embeddings created using weighted-Multidimensional Scaling (WMDS).

Summary of Comparison with Related Research

In this section, we present the results of a behavioral comparison between our proposed techniques and the most closely related research. The results of our analysis are summarized in Table 5.5. The table compares the design requirements of explanations presented in Section 5.2.1, with related techniques such as SIRIUS [9], Andromeda [23], along with the approaches presented by Pagliosa et al. [21], Martins et al. [20] and Silva et. al [22]. As shown in Table 5.5, the results of this analysis validate our claim from Section 5.1 which stated that the related research primarily focuses on individual aspects of interpreting and evaluating embeddings. For example, SIRIUS, Andromeda, and the techniques proposed by Pagliosa et al. [21] and Silva et al. [22] focus on identifying the most influential attributes in certain points or regions in the embeddings. Nevertheless, they do not attempt to quantify the different aspects of local and global structural preservations. At the same time, Martins et al. [20] present a diverse set of visualizations for neighborhood quality analysis they do not look into the aspect of the contributions of the original attributes in the structural preservations of embeddings. Moreover, none of the existing work considers model or data-type agnosticism among their design goals. Overall, our analysis shows that LAPS and GAPS indeed unifies the different aspects of analyzing structural preservation in embeddings. As a result, they provide a more elaborate visual and quantitative interpretations for the low-dimensional embeddings than its closely related research.

Table 5.5: Behavioral Comparison of Laps and Gaps with Closely Related Research

Requirements for Explanations	**LAPS & GAPS**	**SIRIUS**	**Andromeda**	**Pagliosa et al.**	**Martins et al.**	**Silva et al.**
Interpretability (Present attribute influences)	●	○	●	●		●
Local Fidelity (Explain preserved local structure)	●	○	○	○	●	○
Global Legitimacy (Preserved global structure)	●			○	●	○
Model Agnostic (Applicable to any algorithm)	●					
Datatype Agnostic (Applicable to any datatype)	●					
Consistency (In local and global explanations)	●			●	●	●

Note: In the table above ‘●’represents complete support ‘○’ represents partial support for the requirement.

5.4.2.5 User-defined Parameter Analysis

From our description of the LAPS and GAPS in Section 5.3, it can be noticed that two user-defined parameters can significantly influence the outcome of the proposed methods. These parameters include: (1) user-defined scalar *weights* (cf. Eq. 13) for the components of local and global divergence scores and (2) the *selection budget B* (cf. Section 5.3.3) in the computation of global divergence. In this section, we analyze whether the selection of these parameters impacts the consistency of outcome for LAPS and GAPS.

The role of weight in local and global divergences:

Since the scalar weights of each sub-component of local and global divergence can be user-defined, it can be argued whether a strategic selection of these weights can help in manipulating the results or hiding any imperfections in the embeddings. To find answers to this question, we further investigate the local divergence scores

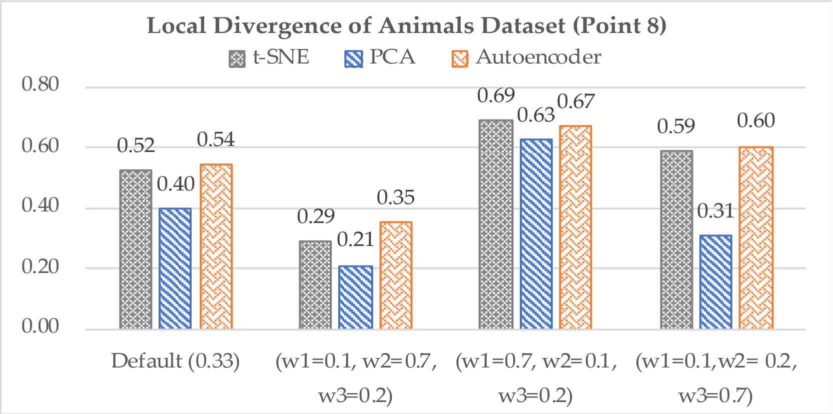


Figure 5.10: Analysis of weight-combinations in the computation of local-divergence using LAPS on Animals dataset.

Note: In the graph above, the left-most set of bars represent the local divergence with a default value of 0.33 for each of the three components of λ_{x_i}(Eq.13). The second, third, and fourth set of bars from the left represent the weight combinations of (0.1, 0.7, 0.2), (0.7, 0.1,0.2), and (0.1, 0.2, 0.7) on the three components of λ_{x_i}.

presented in our first case study (on the Animals dataset) discussed in Section 5.2.1. As shown in Fig. 5.10, in this study we analyze the data-point *blue-whale* from the Animals dataset [9] and compare the local divergence for the point in the embeddings obtained using the algorithms t-SNE, PCA, and VAE. Fig. 5.3 shows that, with the default relative weights (i.e., 0.33) among the three algorithms, PCA has the lowest local-divergence for *blue-whale*. Fig. 5.3 computes the local divergence scores for the algorithms with the default weights of 0.33 for each of its components. In this section, we analyze the impact of any changes in the weights of the individual components of the local divergences for the point *blue-whale*. Our analysis results are summarized in Fig. 5.10. The results show that different combinations of weights for the individual components of local divergence does not allow users to manipulate the results but only shows the differences between the embeddings more clearly. For example, as shown in Fig. 5.3, in all the embeddings, the discrepancy for attribute influence explanations have been the highest (i.e., >73%) among the three components. Whilst the false and missing neighborhood has the lowest discrepancy (i.e., ≤20%), the inconsistency in the order of neighbors is high for t-SNE and VAE (i.e., ~70%) but low for PCA (~30%). Hence, in Fig. 5.10, when increasing the weight of attribute influence explanation component (i.e., w1) to 0.70 the local divergence for all the algorithms is increased by 25% to 58% from their original local divergences obtained using default weight combinations. Similarly, increasing the weight of the neighborhood order component (i.e., w2) to 0.70 reduces the local divergence scores for all the algorithms by 50% to 80%. However, as shown in Fig. 5.10 in all the cases, PCA still has the lowest local divergence among the three algorithms in every case.

The role of budget in global-divergence computation:

In this section, we investigate the impact of the budget (i.e., the data subset size) in the computation of global divergence. For this analysis, we extend the results of our second case study presented in Section 5.2.1 and investigate the impact of a budget size 5, 10, 15, 20, and 25 on the global-divergence scores of the Breast Cancer dataset using t-SNE, MDS, and UMAP. We summarize the results of our analysis in Figure 5.11. As shown in Fig. 5.11, with a gradual increase in the selection budget, only the absolute value of the overall global divergence steadily dropped for all the three algorithms with MDS being the best performing algorithm in all the 5 cases. Hence, from Fig. 5.11, it can be seen that GAPS appropriately shows the best performing algorithm in terms of preservation of global structure for a given dataset. However, the absolute value of the divergence might get more accurate with a higher budget size.

5.5 Future Work

There are several avenues of future work that we would like to explore. For instance, in any interactive technique, one of the most important aspects is scalability. Although, both the proposed algorithms have a computational complexity of $O(n^2)$, for our current design of LAPS and GAPS, we restrict the user-defined neighborhood size (cf. Eq. 9) to be as large as 10 and the number of perturbed samples (cf. Eq. 10) to be a maximum of 5000. These design constraints are inspired by Ribeiro et al. [28] who confirm the adequacy of 10 nearest neighbors and 5000 sampled instances in determining the local properties of a data-point. However, we leave experimenting with different sizes of neighborhoods (i.e., > 10)

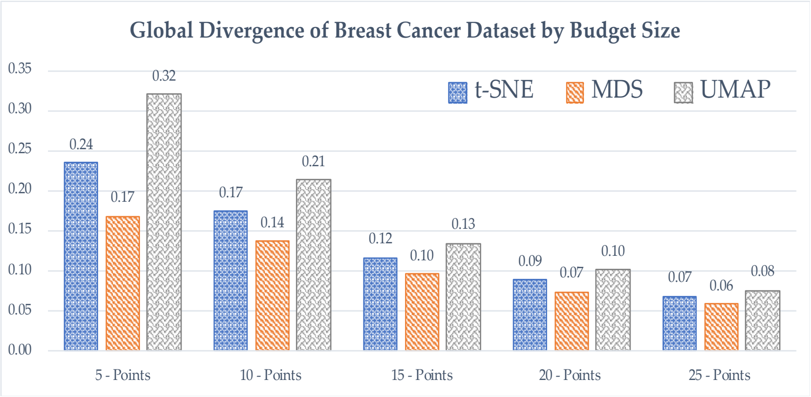


Figure 5.11: Analysis of global divergence scores by budget size.

Note: The bars in the above graphs represent the global divergence λ_X (Eq.14) computed using t-SNE, MDS, and UMAP respectively. The graphs show that with a gradual increase in the selection budget, the global divergence steadily dropped for all three algorithms.

to future work. Although, improving any inherent open challenges [3], [4] of dimensionality reduction techniques (e.g., computational complexity [4], optimization of hyperparameters [41]) is beyond the scope of this research.

Apart from scalability, we think there are a few more aspects where the proposed work can be improved. Firstly, although both the proposed algorithms allow for user interactions with the processes, the overall interactivity of the approaches can be improved by integrating them as a part of a unified visual framework. As ongoing work, we are working on creating such a framework. To enhance the overall scalability of the framework, we are currently exploring parallel processing for LAPS and GAPS. Secondly, to enhance the fidelity of GAPS, as discussed in Section 3.2, our ongoing work also includes defining the diversified sample selection for GAPS as an Exhaustive Subset Enumeration [36] problem.

5.6 Conclusions

In this chapter, we propose two interactive explanation techniques for low-dimensional embeddings obtained from *any* dimensionality reduction algorithm. The first technique LAPS produces a local approximation of the neighborhood structure to generate interpretable explanations on the preserved locality for a single instance in an embedding. The second method GAPS explains the retained global structure of a high-dimensional dataset in its embedding, by unifying non-redundant local-approximations from a coarse discretization of the projection space. Our experimental evaluation of the techniques with tabular, image, text, and audio data demonstrates the flexibility of these techniques. Moreover, our extensive experiments show the utility of the proposed techniques in demonstrating the preserved structural relationships in lower dimensions, as well as determining the most correlated attributes in a dataset, along with an interactive selection of the most appropriate dimensionality reduction algorithm for any given dataset.

References

[1] M. Cavallo and Ç. Demiralp, "A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration," *arXiv:1811.12199 [cs]*, Nov. 2018, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1811.12199.

[2] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[3] L. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality Reduction : A Comparative Review," *J Mach Learn Res 10*, vol. 66, no. 71, p. 13, 2008.

[4] 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*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[5] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality Reduction in the Wild: Gaps and Guidance," Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, Jun. 2012.

[6] S. Lespinats and M. Aupetit, "CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings," *Computer Graphics Forum*, vol. 30, no. 1, pp. 113–125, Mar. 2011, doi: 10.1111/j.1467-8659.2010.01835.x.

[7] J. M. Lewis and V. R. de Sa, "A Behavioral Investigation of Dimensionality

Reduction," *In Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, p. 7, 2012.

[8] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding Projector: Interactive Visualization and Interpretation of Embeddings," *arXiv:1611.05469 [cs, stat]*, Nov. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1611.05469.

[9] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, "SIRIUS: Dual, Symmetric, Interactive Dimension Reductions," *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, Jan. 2019, doi: 10.1109/TVCG.2018.2865047.

[10] R. Faust, D. Glickenstein, and C. Scheidegger, "DimReader: Axis lines that explain non-linear projections," *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 481–490, Jan. 2019, doi: 10.1109/TVCG.2018.2865194.

[11] J. Stahnke, M. Dork, B. Muller, and A. Thom, "Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions," *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 629–638, Jan. 2016, doi: 10.1109/TVCG.2015.2467717.

[12] L. Kodali, J. Wenskovitch, N. Wycoff, L. House, and C. North, "Uncertainty in Interactive WMDS Visualizations," *In Visualization in Data Science (VDS at IEEE VIS)*, 2019.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Aug. 2018, doi: 10.1145/3236009.

[14] D. Sacha, L. Zhang, M. Sedlmair, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, and D.A. Keim, "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis," *IEEE Trans. Visual. Comput. Graphics*, vol. 23, no. 1, pp. 241–250, Jan. 2017, doi: 10.1109/TVCG.2016.2598495.

[15] M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively," *Distill*, 2016, doi: 10.23915/distill.00002.

[16] S. Liu, D. Maljovec, B. Wang, P. -t Bremer, and V. Pascucci, "Visualizing

High-Dimensional Data: Advances in the Past Decade," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 21–30, 2016.

[17] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Moller, "DimStiller: Workflows for dimensional analysis and reduction," in *2010 IEEE Symposium on Visual Analytics Science and Technology*, Salt Lake City, UT, USA, Oct. 2010, pp. 3–10, doi: 10.1109/VAST.2010.5652392.

[18] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, "Local Affine Multidimensional Projection," *IEEE Trans. Visual. Comput. Graphics*, vol. 17, no. 12, pp. 2563–2571, Dec. 2011, doi: 10.1109/TVCG.2011.220.

[19] J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A.K. Tung, "LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 236–245, Jan. 2018, doi: 10.1109/TVCG.2017.2744098.

[20] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, "Visual analysis of dimensionality reduction quality for parameterized projections," *Computers & Graphics*, vol. 41, pp. 26–42, Jun. 2014, doi: 10.1016/j.cag.2014.01.006.

[21] L. Pagliosa, P. Pagliosa, and L. G. Nonato, "Understanding Attribute Variability in Multidimensional Projections," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Sao Paulo, Brazil, Oct. 2016, pp. 297–304, doi: 10.1109/SIBGRAPI.2016.048.

[22] R. R. O. D. Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea, "Attribute-based Visual Explanation of Multidimensional Projections," *EuroVis Workshop on Visual Analytics (EuroVA)*, p. 5 pages, 2015, doi: 10.2312/EUROVA.20151100.

[23] J. Z. Self, L. House, S. Leman, and C. North, "Andromeda: Observation-Level and Parametric Interaction for Exploratory Data Analysis." *Technical report, Department of Computer Science*, Virginia Tech, Blacksburg, Virginia, 2015.

[24] K. Bunte, M. Biehl, and B. Hammer, “Dimensionality reduction mappings,” in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Paris, France, Apr. 2011, pp. 349–356, doi: 10.1109/CIDM.2011.5949443.

[25] J. B. Tenenbaum, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[26] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *arXiv:1705.07874 [cs, stat]*, Nov. 2017, Accessed: Mar. 24, 2020. [Online]. Available: http://arxiv.org/abs/1705.07874.

[27] A. A. Freitas, “Comprehensible classification models: a position paper,” *SIGKDD Explor. Newsl.*, vol. 15, no. 1, pp. 1–10, Mar. 2014, doi: 10.1145/2594473.2594475.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” *arXiv:1602.04938 [cs, stat]*, Feb. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1602.04938.

[29] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang, “Correlation Judgment and Visualization Features: A Comparative Study,” *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 3, pp. 1474–1488, Mar. 2019, doi: 10.1109/TVCG.2018.2810918.

[30] E. Levina and P. J. Bickel, “Maximum Likelihood Estimation of Intrinsic Dimension,” *Advances in neural information processing systems*, pp. 777–784, 2005.

[31] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, “Neighbourhood Components Analysis,” *Advances in neural information processing systems*, pp. 513–520, 2005.

[32] G. Plumb, D. Molitor, and A. Talwalkar, “Model Agnostic Supervised Local Explanations,” *arXiv:1807.02910 [cs, stat]*, Jul. 2018, Accessed: Aug. 12, 2019. [Online]. Available: http://arxiv.org/abs/1807.02910.

[33] J. C. Gower, “A General Coefficient of Similarity and Some of Its Properties,” *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971, doi: 10.2307/2528823.

[34] H. Wang and M. Hong, “Distance Variance Score: An Efficient Feature Selection Method in Text Classification,” *Mathematical Problems in Engineering*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/695720.

[35] J. P. Boyd, “Additive blending of local approximations into a globally-valid approximation with application to the dilogarithm,” *Applied Mathematics Letters*, vol. 14, no. 4, pp. 477–481, 2001, doi: 10.1016/S0893-9659(00)00180-4.

[36] F. Pan, A. Roberts, L. McMillan, D. Threadgill, and W. Wang, “Sample Selection for Maximal Diversity,” in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, USA, Oct. 2007, pp. 262–271, doi: 10.1109/ICDM.2007.16.

[37] R. Haftka, “Combining global and local approximations,” *AIAA journal*, vol. 29, no. 9, p. 1523, Sep. 1991.

[38] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

[39] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[40] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv:1802.03426*, Feb. 2018, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1802.03426.

[41] P. G. Poličar, M. Stražar, and B. Zupan, “openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding,” Bioinformatics, preprint, Aug. 2019. doi: 10.1101/731877.

[42] W. S. Torgerson, “Multidimensional scaling: I. Theory and method,” *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952, doi:

10.1007/BF02288916.

[43] S. T. Roweis, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.

[44] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *arXiv:1312.6114 [cs, stat]*, May 2014, Accessed: Mar. 24, 2020. [Online]. Available: http://arxiv.org/abs/1312.6114.

[45] Z. Zhang and H. Zha, “Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment,” *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, Jan. 2004, doi: 10.1137/S1064827502419154.

[46] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.

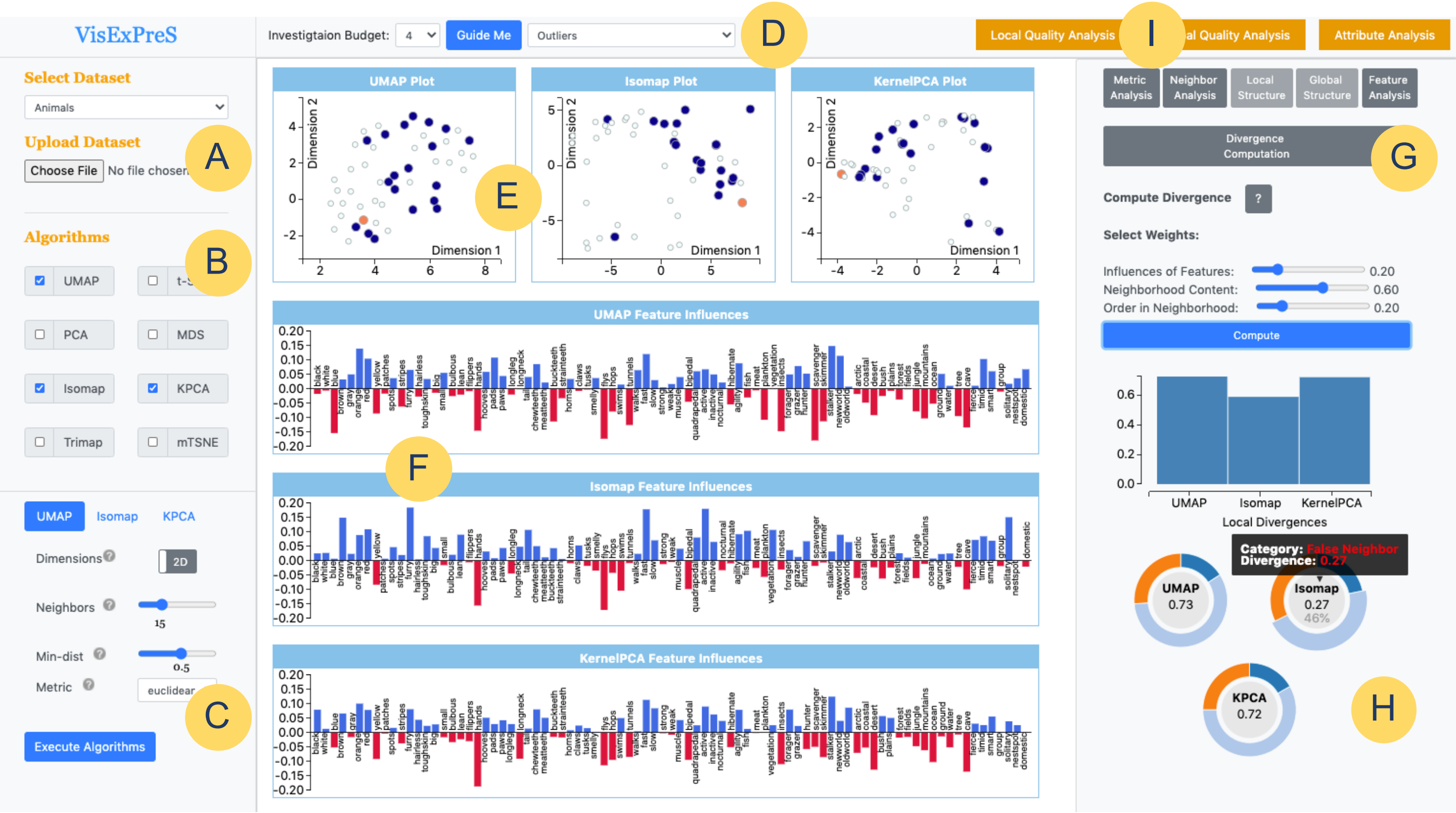
[47] J. Demśar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Chapter 6

VisExPreS: A Visual Interactive Toolkit for Evaluations of Embeddings

Embeddings are low-dimensional representations of high-dimensional data that are obtained using Dimensionality Reduction (DR in short). DR algorithms transform high-dimensional data into embeddings while attempting to maximally preserve their structural properties. Most DR algorithms identify the structure of the original data using the relative proximities among their data-points [1]. DR as a transformation technique not only reduces the computational overhead of high-dimensional data analysis, but it also makes the visualization of such datasets possible with traditional spatial techniques (i.e., 2D or 3D plots).

Despite their utility, DR techniques come with a set of major caveats. Firstly, the dimensions derived using such techniques lack a clear-to-interpret mapping with the original features in the data [2], [3]. As a result, novice data analysts are often forced to blindly trust the embeddings without truly understanding the meaning of the projection axes or the positioning of data-points. Secondly, there exists a plethora of DR techniques with their own respective hyperparameter combinations that significantly influence the embedding structure. The non-intuitive nature of these parameters also hinders the interpretability of these techniques making the selection of an appropriate DR algorithm for any dataset, difficult [4]. Thirdly, in most cases, embeddings derived from DR do not make existing errors and distortions [5] prominent to the users. In some cases [5]–[8], where such distortions are visually exposed, the users are not allowed to control or interact with them [6], [7]. All these limitations make an efficient evaluation of embeddings obtained from



Investigtaion Budget: 4

Outliers

Select Dataset: Animals

Upload Dataset: Choose File No file chosen

Algorithms:

☑ UMAP ☐ t-S

☐ PCA ☐ MDS

☑ Isomap ☑ KPCA

☐ Trimap ☐ mTSNE

Dimensions: 2D

Neighbors: 15

Min-dist: 0.5

Metric: euclidean

Compute Divergence ?

Select Weights:

Influences of Features: 0.20

Neighborhood Content: 0.60

Order in Neighborhood: 0.20

Category: False Neighbor
Divergence: 0.27

Figure 6.1: The Structural Quality Analysis View of VisExPreS. The interface is divided into nine regions that enable simultaneous assessment of preserved structure in a set of embeddings.

DR algorithms extremely challenging [9].

Traditionally, the quality of embeddings is interpreted and evaluated using two different methods namely: (i) metric based quantitative analysis; and, (ii) qualitative or visual analysis of the obtained embedding. Being reliable and repeatable [9]–[11], metric-based quantitative evaluation of embeddings can effectively assist users to compare DR algorithms by associating numeric identifiers to their qualitative characteristics. **Nevertheless, such techniques being formally defined, do not allow users to have much control over the analysis process.** As a result, users do not get the opportunity to apply their knowledge and expertise in the assessment process. On the other hand, several qualitative analysis techniques for embeddings allow users to visually explore the neighborhood structures [12]–[14], errors & distortions [5]–[8], and feature variances [15], [16] within the neighborhoods of the projections. **Nevertheless, using such techniques making any decision regarding the best performing DR algorithm in a given context entirely depends on the analyst's perception and understanding of the embeddings.** The reason being, such techniques [2], [17], [18] often do not proactively guide users with the analysis process. For example, for very large datasets, most existing techniques [2], [17], [18] do not assist users with the selection of influential data points [13] or representative data subsets [19], [20] that provide a good representation of the original data and can reveal the overall quality of the embeddings better than other points. As a result, **novice data analysts often fail to utilize the complete potential of such interactive visualization techniques** when exploring low-dimensional embeddings.

This research aims to bridge the gap between these two traditional techniques for evaluating embeddings. In this work, unifying the benefits of both, we present a visual interactive toolkit that enables a proactively guided and user-driven analysis of preserved structure in any embeddings obtained from any DR algorithm. Towards achieving this goal, at first, we present two novel interactive embedding quality analysis methods. The first technique PG-LAPS (Proactively Guided Local

Approximation of Preserved Structure) enables the computation of the local-divergence, that examines the fidelity of the relative positioning of any individual data-point in an embedding by approximating a neighborhood locally around that point. Moreover, to assist novice users with the analysis process, PG-LAPS proactively guides users with the selection of representative data points from the input dataset. The second technique PG-GAPS (Proactively Guided Global Approximation of Projection Space) computes the global-divergence, that explains the preserved global structure in a low dimensional embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. To facilitate a proactively guided exploration of embeddings, as a part of PG-GAPS we present RepSubset, a novel algorithm that generates representative subsets from the original data based on the notions of density [13], [21] and dissimilarity [22], [23] in the dataset. The two techniques are then composed into a visual toolkit that we named VisExPreS (Visual Explanations of Preserved Structure). An overview of the VisExPreS interface is presented in Fig. 6.1. The presented toolkit not only gives users more control over the quality assessment process but also allows them to focus on the aspects of the analysis that are the most interesting to them. Moreover, VisExPreS enables side-by-side (both visual and quantitative) comparison of the performances of multiple DR algorithms on the same dataset. We evaluate VisExPreS, PG-LAPS, PG-GAPS, and RepSubset using extensive evaluation. Our primary contributions in this article are as follows:

1. *PG-LAPS*: a novel user-driven embedding quality analysis method for proactively guided investigation of the retained local structure in an embedding.

2. *PG-GAPS*: a novel, proactively guided, and user-driven DR quality assessment method for examining the preserved global structure in an embedding.

3. *RepSubset:* A novel algorithm for selecting representative subsets from the original dataset based on the notions of density and dissimilarity in the data.

4. *VisExPreS*: an interactive visual toolkit that enables a user-driven computation of metrics *local* and *global-divergence* metrics, while enabling side-by-side comparison of multiple embeddings.

The chapter is organized as follows: Section 6.1 presents the required background and related work; Section 6.2 discusses our design goals for VisExPreS interface. Section 6.3 presents PG-LAPS, PG-GAPS, and RepSubset and elaborates on the different views of VisExPreS. Section 6.4 presents an extensive evaluation of the framework whilst Section 6.5 discusses some limitations and opportunities for future work. Section 6.6 concludes the article.

6.1 Background and Related Work

In its most general setting, non-linear DR (NLDR) can be formally defined as assuming a matrix X of size $n \times D$, that represents a high-dimensional dataset with n records and D features, DR algorithms map into an embedding Y of size $n \times d$. Ideally, for most real-world datasets d represents the intrinsic dimensionality of X that is the minimum number of dimensions that can be used to represent X [5]. Normally, $d \ll D$. In their most general settings, NLDR techniques can be formally defined [5] as an optimization problem:

$$\underset{Y \in \mathbb{R}^{d \times n}}{\operatorname{argmin}} f(Y; X, \theta) \tag{6.1}$$

where the objective function f attempts to preserve the relative proximities among the data points from X to Y. In Eq. 6.1, θ represents the hyperparameters of the function f. In manifold learning [24], the vectors in X are assumed to be sampled from a non-linear manifold where the notion of proximity among the data-points is traditionally defined using distance measures [1], [25]. In this research, for any data-point $x_i \in X$ where the neighborhood [26], [27] of x_i is a subset Z of X containing x_i, we define the proximity between the point x_i and its nearest neighbors as $\pi_{x_i}(x')$, where, $x' \in Z$ and $x_i \neq x'$. Here, $\pi_{x_i}(x') \in \mathbb{R}$. Ideally, for any NLDR, the preservation of the relative proximities among data-points refers to the following: considering a set of data-points $x_i, x_j, x_k \in X$ where, $i \neq j \neq k$, if

$\pi_{x_i}(x_j) < \pi_{x_i}(x_k)$ then after the transformation $\pi_{y_i}(y_j) < \pi_{y_i}(y_k)$ should hold. Here, $y_i = f(x_i)$, $y_j = f(x_j)$, and $y_k = f(x_k)$. However, NLDR being an optimization problem, its outcome often converges to a local-optima leading to the relative proximities not being retained for every data-point in embedding Y. For most NDLR techniques, such relative proximities form the basis of both quantitative and visual evaluations of their resulting embeddings [9], [10]. In this section, at first, we elaborate on the existing mechanisms for quality assessments of embeddings. Next, we present the closely related work that focuses on obtaining representative subsets from datasets for further analysis.

6.1.1 Quality Assessments of Embeddings

In this section, we elaborate on related work that evaluates embeddings using quantitative or qualitative (visual) methods and discuss the novelty of VisExPreS, our proposed visual interactive toolkit.

6.1.1.1 Quantitative Evaluation of Embeddings

In theory, for most NLDR techniques, a quantitative analysis of the resulting embedding should be possible using two simple mechanisms. Firstly, by examining the value of the objective function f upon convergence; and secondly, by performing an inverse transformation from Y to X. Nevertheless, in real-life scenarios often neither of the above-mentioned techniques is applicable. The reasons being: as the former technique can only be used for comparing different executions of the same algorithm [9], the later becomes infeasible for real-world datasets whose underlying manifold structures are unknown [4]. As a result, most existing techniques that attempt to quantify the characteristics of an embedding [28], either by examining the absolute (i.e., the actual distances among data-points) or the relative (i.e., distance ranking among points) proximities among the data-points after the transformation. As per Rieck et al. [29], rank-based DR quality metrics are more popular due to their stability with the scaling of pairwise distances among the data points.

Popular rank-based DR quality metrics include Local Continuity Meta-Criterion (LCMC) [9], Trustworthiness and Continuity (T&C) [9], Mean Relative Rank Errors (MRRE) [29] among others. To evaluate the embedding quality, these metrics compare the ranks of the sorted distances in K-ary neighborhoods in X and Y [10]. Due to their similarities, Lee and Verleysen have unified these three metrics under a co-ranking matrix framework [9]. Also, the authors proposed several other rank-based DR quality metrics namely: mean K-ary neighborhood preservation (Q_{nX} [9]), local quality criteria for K-ary neighborhoods (Q_{local} [10]) and global quality criteria for the embedding (Q_{global} [10]). The co-ranking framework primarily examines the average agreement between all K-ary neighborhoods in X and Y based on a matrix containing the ranks of pairwise distances [10] between the data-points. Apart from these, the metrics entropy and mutual information [30] and Spearman rank correlation [29], [30] are also popularly used to determine the preservation of topology in embeddings. On the other hand, Residual Variance [29] is a popularly used distance-based DR quality metric that computes a linear correlation between the absolute distances between any pairs of data-points in X and Y. Other distance based metrics that focus on dissimilaries of neighborhoods [24] include neighborhood hits [24], projection precision score [6], and Spectral Overlap [30]. All these metrics examine the proportion of the neighborhood that is preserved in an embedding. Apart from the distance and rank based metrics, some other quality measures analyze the stress (i.e., distortion) of the objective function for DR. Such metrics popularly include normalized stress [7] and Krushkal Stress [6], [29].

As shown by Lee et al. [10] and Johannemann et al. [30], such quantitative embedding quality measures can be useful for comparing the performances of multiple DR algorithms. However, these metrics being formally defined, do not allow much flexibility in the analysis process. Moreover, it can often get challenging for users to actively engage with their computations that will enhance the user's trust on the metric's value. For example, in most cases such metrics present a single quantitative value for the embedding quality, without providing any rationale behind the computation of the metric. In such cases, the users remain unable to intervene or interact with the metric computation process and are bound to trust the presented result blindly without really understanding how it was computed. As a result, in most cases, the calculation of these metrics does not incorporate the user's perception, and expertise in the assessment process.

6.1.2 Visual Evaluation of Embeddings

In order to make the analysis of embeddings more engaging for users, several visual interactive mechanisms [2], [8], [17], [18], [31] have been proposed in the past few years. Among these, some techniques only visualize the neighborhood structures of the data-points after DR [12]. Some depict the distortions in the embeddings [7], [13], [14] with the help of false [13] and missing [7] neighbors. Taking the interaction one step further, some techniques even allow users to intervene and fix (i.e., reposition or remove) [2], [16] any misplaced data-points [5] in the embeddings. For example: as Smilkov et al. [12] effectively visualizes the neighborhoods of any selected data-point in the embedding, Lai et al. [13], Cutura et al. [14], Martins et al. [7], and Aupetit et al. [31] identify and depict false and missing neighbors in embeddings. Moreover, using graphical representations, Heimerl et al. [8] show the distribution of neighborhood distances for all data points, whilst France et al. [32] depict the *agreements* in K-ary neighborhoods. On the other hand, in order to improve distortions, Stahnke et al. [5] and Joia et al. [33] automatically reposition misplaced datapoints; whereas, Pagliosa et al. [16], Cavallo et al. [2] allow users to interact with the DR process and remove or relocate points in embeddings. Among further existing techniques, some assist users to interactively modify the hyperparameter combinations for DR algorithms [12]. Some other techniques [13], [15]–[18] help to visualize the contributions of the features in the relative positioning of data-points using feature weights [13], [18], feature correlations [17], or the variance in feature values [15], [16].

The case-studies [5], [12], [17] and user-studies [5] performed by the authors of the existing techniques show that such visual embedding quality analysis techniques can indeed help with an interactive exploration of the projection. However, as these techniques are not designed to quantify the behavioral characteristics of the embeddings, the decision on the best performing algorithm relies on the perception and expertise of the analyst. Moreover, most of these techniques focus on partial aspects of embedding quality. For example, either their focus lies on investigating false and missing neighbors [7], [14], or on the contributions of original features [15]–[18], or on the impact of hyperparameters [7] on the chosen DR algorithm.

In this research, we attempt to unify the different focus areas of visual analysis of embeddings and address the challenges that we have discussed for both quantitative and qualitative assessment methods for DR.

6.1.3 Selecting Representative Subsets from Data

The existing methods for representative subset selection can be broadly categorized into two groups namely [19], [34]: clustering-based design and uniform design approaches. The clustering-based techniques aim at generating clusters from the original data followed by selecting a diversified subset from the generated clusters [20]. Such techniques can be further classified into [19], [34]: hierarchical, non-hierarchical, and density based techniques. Over the past years, for clustering-based subset selection popular methods such as K-means [20], OPTICS [33], DBSCAN [20] have commonly been used by researchers. For example, in their work, Lai et al. [13] have used DBSCAN to suggest representative data points as well as data subsets, whereas Daszykowski et al. [20] have suggested representative points using a hybrid of DBSCAN and K-Means clustering methods.

With the uniform design approaches [20]–[22], [35], the representative data-points are selected in such a way that they uniformly cover the data-space. Being more popular than the clustering-based subset selection methods, over the past few decades several approaches have been proposed for selecting uniform representative subsets. One of the most popular technique in this category is the Kennard-Stone [23] algorithm that is based on the notion of *dissimilarity* between the data-points. The Kennard-Stone is an iterative method that aims at minimizing the pairwise Euclidean distances [17] between the data-points that are already a part of the representative subset and the remaining points in the dataset. Another popular method dissimilarity based uniform subset selection technique is OptiSim, presented by Clark et al. [22]. The technique generalizes the maximum and minimum dissimilarity-based approaches. Also based on the notion of dissimilarity, Tominaga et al. [35] presented a genetic algorithm to select representative subsets. The algorithm uses the pairwise Euclidean distances among the data-points within the selected subset along with the mean of the product moment correlation as its two fitness functions. On the other hand, instead of focusing on dissimilarity among data points, researchers such as Chaudhuri et al. [21], [36] and Mall et al. [20] have focused on the idea of *density* of each point the dataset for them to be considered as a part of the representative subset. For example, in a multidimensional space, Chaudhuri et al. [21] have presented a technique that selects data points with the highest density in the dataset. Whereas for connected graphs, Mall et al. [20] have presented the technique FURS that selects representative data-points with the highest measure of 'degree centrality' in the network. In both these iterative techniques, to enhance the diversity of the selected subset, the authors have ignored the K-nearest neighbors of the data points that have already been added to the representative subsets.

The representative subset selection technique RepSubset, presented in this research, addresses multiple aspects that are not considered by the traditional iterative subset selection algorithms [20]–[22], [35]. First of all, to enable a diversified subset selection RepSubset combines the notions of density and dissimilarity in the original dataset while ignoring the impact of K-nearest neighbors of the already selected points. Secondly, in order to capture the nonlinearity in the underlying manifold of the original dataset, in contrast to the existing approaches, the proposed technique computes the pairwise geodesic distances [4] to measure the dissimilarities among data points.

6.2 Design Goals for the VisExPreS Toolkit

The primary goal for the VisExPreS toolkit is to assist users with an interactive and engaging quality assessment of low-dimensional embeddings. DR being a complex and black-box technique, this toolkit should not only guide users with the assessment process but also should allow users to have the driver's seat in the interactive and quantitative quality analysis of embeddings. In this section, we elaborate on our primary design goals for VisExPreS.

Goal 1: Proactive guidance for representative data: In order to assist novice data scientists with their analysis, the toolkit needs to provide proactive guidance to its users in terms representative data points [13] as well as, representative data subsets [22] for analyzing the preserved local and global structures of the embeddings. In case of representative data-points, the toolkit should consider multiple perspectives from which a data-point may seem interesting and offer users with a few such perspectives to select from. For representative subsets, the toolkit should automatically generate diverse representative subsets from the input data and present users with multiple subset options. Moreover, the toolkit should also suggest appropriate hyperparameter combinations for the chosen DR methods.

Goal 2: Simultaneous investigations of embeddings: To address the challenges of selecting the most appropriate DR algorithm in a given scenario, we require the toolkit to be algorithm agnostic and assist with side-by-side comparisons among multiple embeddings.

Goal 3: Contributions of the original features: Features of a dataset play the most important role in computing the proximities among the points in a dataset. Moreover, as humans, we relate to meaningful names [37] more easily than numeric values or complex visuals. Hence, to enhance the interpretability of the analysis, we need the toolkit to interactively present the influences of the original features in the formation of the neighborhoods in the dataset.

Goal 4: Interpretable assessment of structural quality: We need the toolkit to address the non-transparent nature of the embeddings. Here, we require the toolkit to assist with interpretable explanations of preserved local and global structures in embeddings for both novice and expert data analysts. For this purpose, we need the toolkit to present multiple aspects of structural quality in embeddings annotated with textual descriptions.

Goal 5: User-driven computation of quality criteria: In this research, we define the term user-driven as follows: the computation of any quality metrics as a part of the interactive toolkit, should be completely user-steering. That is, the toolkit should put the users in charge of entire quality analysis process. It should not only let them select data points of their interests for the assessment, but it must also allow users to actively participate in determining the contributions (i.e., weights) of each component of the defined quality metrics. Based on this definition, to bridge the gaps between the quantitative and visual assessment methods for embeddings (cf. Section 6.1.1), we need the presented toolkit to be *user-driven* in nature.

6.3 The VisExPreS Toolkit

At the core of the *VisExPreS* (***Vis**ual **Ex**planations of **Pre**served **S**tructure*) toolkit, we incorporate two novel quality analysis techniques for embeddings. The first technique *PG-LAPS* ***(P**roactively **G**uided **L**ocal **A**pproximation of **P**reserved **S**tructure)* allows for an interactive computation of the metric *local-divergence* that both visually and quantitatively assesses the local neighborhoods of individual data-points in embeddings. The second technique *PG-GAPS* ***(P**roactively **G**uided **G**lobal **A**pproximation of **P**rojection **S**pace)* helps with the computation of a global quality criterion *global-divergence*. Th metric global-divergence quantifies the preserved global structure for a set of non-redundant data points in the embedding. In VisExPreS, the set of non-redundant points is recommended by the toolkit itself in the form of a representative subset of the original data.

It is important to note that, the techniques PG-LAPS and PG-GAPS are presented as enhancements of our previously proposed methods LAPS and GAPS [38]. In this research, we have added proactive guidance for selection of representative data-points (also data-subsets) and hyperparameters for the DR algorithms and have incorporated the two techniques into the visual interactive toolkit VisExPreS. Also, as apart of PG-GAPS we present RepSubset an novel method for selecting representative data subsets with high coverage. In the next two sections, at first, using Figures 6.2 and 6.3 we discuss the PG-LAPS and PG-GAPS processes along with the proposed RepSubset algorithm, in detail. Further discussions on the LAPS and GAPS methods can be found in [38]. Next, we justify the interface design for VisExPreS with respect to its design goals discussed in Section 6.2.

6.3.1 Proactively Guided Computation of Local-Divergence in Embeddings

As depicted in Figure 6.2, the Proactively Guided LAPS or PG-LAPS process can be divided into seven distinct steps. Overall, PG-LAPS investigates a single data-point in the dataset and enables user to quantify its preserved local structure using the output metric local-divergence. In the following, we discuss each step of PG-LAPS in detail.

Step 1: Pre-process Input Data and Obtain Embedding:

As shown in Fig. 6.2- Step 1, to avoid additional noise in the obtained embeddings and also to enhance their stability [26], [27], [39], the pre-processing of our input data begins with an estimation of the intrinsic dimensionality d (cf. Section 6.1). Here, following a popular practice in academia [4], [26], [39], we use the maximum likelihood intrinsic dimensionality estimator for this purpose [39]. The estimator can be defined as:

$$\hat{\mathrm{d}} = \frac{1}{\mathrm{k}_2 - \mathrm{k}_1 + 1} \sum_{\mathrm{k}=\mathrm{k}_1}^{\mathrm{k}_2} \hat{\mathrm{d}}_\mathrm{k} \quad \text{where, } \hat{\mathrm{d}}_\mathrm{k} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{d}_\mathrm{k}(\mathrm{X}) \qquad (6.2)$$

where, $\hat{\mathrm{d}}$ represents a unit vector with an estimation for d and $(\mathrm{k}_2 - \mathrm{k}_1)$ signifies the range of nearest neighbors to consider while estimating d. At the same time, we

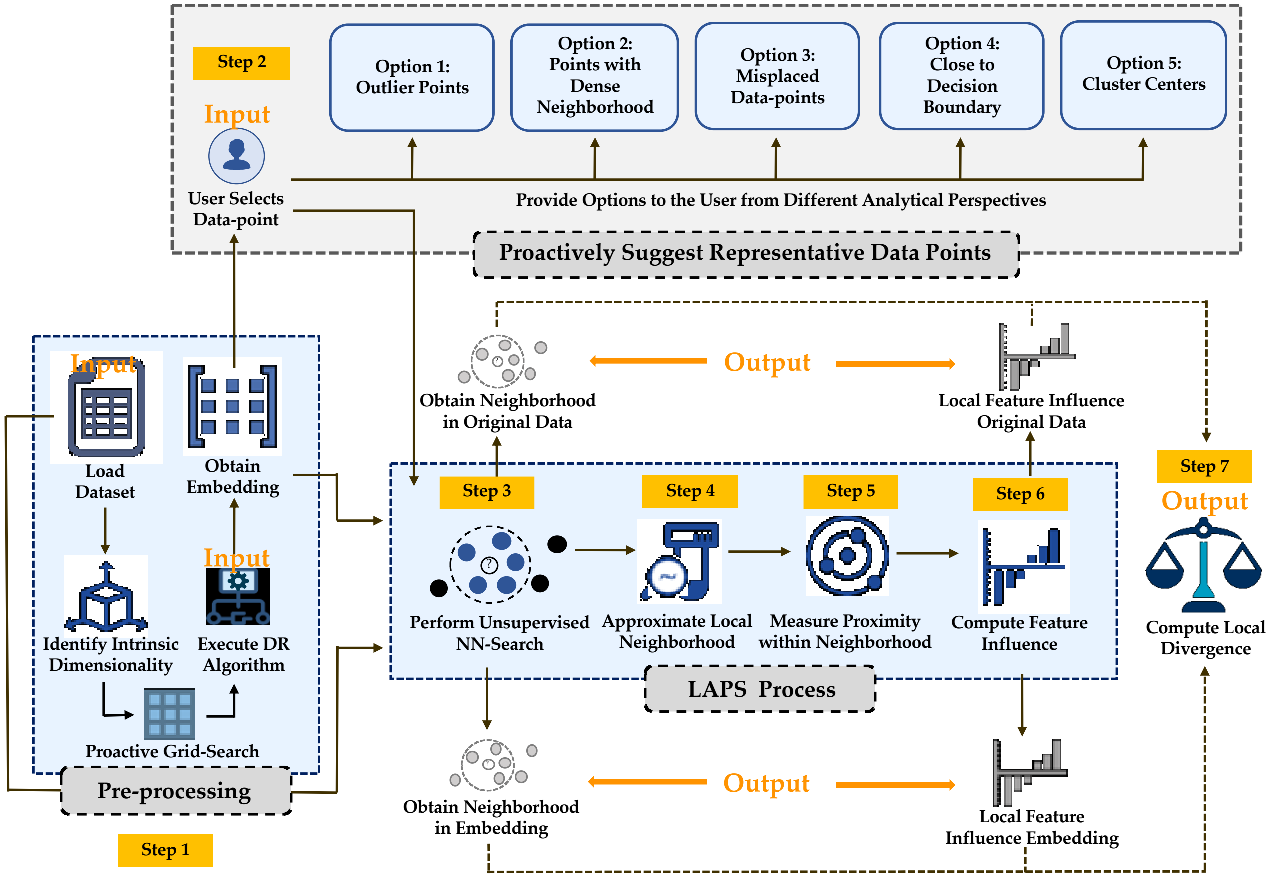

Step 2
Input
User Selects Data-point
Option 1: Outlier Points
Option 2: Points with Dense Neighborhood
Option 3: Misplaced Data-points
Option 4: Close to Decision Boundary
Option 5: Cluster Centers
Provide Options to the User from Different Analytical Perspectives
Proactively Suggest Representative Data Points
Input
Load Dataset
Obtain Embedding
Input
Identify Intrinsic Dimensionality
Execute DR Algorithm
Proactive Grid-Search
Pre-processing
Step 1
Obtain Neighborhood in Original Data
Output
Local Feature Influence Original Data
Step 3
Step 4
Step 5
Step 6
Perform Unsupervised NN-Search
Approximate Local Neighborhood
Measure Proximity within Neighborhood
Compute Feature Influence
LAPS Process
Step 7
Output
Compute Local Divergence
Obtain Neighborhood in Embedding
Output
Local Feature Influence Embedding

Figure 6.2: Overview of the PG-LAPS technique. The technique primarily contains three components namely: Pre-processing, Proactive Guidance, and the LAPS process. The overall process steps for PG-LAPS are numbered in an ascending order and are highlighted in yellow.

also perform an exhaustive and proactive grid search [4] to identify optimum hyperparameter combinations for the chosen DR method given the data. Next, the embedding for the input data is obtained by executing any DR algorithm chosen by the user using the estimated value for d and the identified hyperparameter combinations.

Step 2: Proactively Suggest Representative Data Points:

To enable the evaluation of preserved local structure in an embedding, PG-LAPS allows users to interactively select a single data point for subsequent analysis. Alternatively, as shown in Fig. 6.2 - Step 2, the proactive guidance component of PG-LAPS suggests users with a set of representative data-points that might be interesting from five different analytical perspectives. Here, data-points are considered to be interesting if the point (1) is an outlier, (2) has highly dense neighborhood, (3) is misplaced (i.e., has false or missing neighbors in the projection), (4) is close to the decision boundary[28] , (5) is the center of any cluster. All these options are well-known aspects in academia for identifying influential data-points in any dataset [14].

Next, we discuss how the PG-LAPS system identifies representative data points from these five analytical perspectives, in detail.

- **Option 1:** The outliers in the original data are identified using a combination of the notions of density [13], [21] and dissimilarity [22], [23] among the data points. Here, at first, we measure the pairwise geodesic distances [5] among all data points. The geodesic distance $dist_\gamma$ among any two points x_i and x_j in X, can be defined as:

  $$dist_\gamma(x_i, x_j) = inf\{L(\sigma)\} \tag{6.3}$$

  where $inf\{L(\sigma)\}$ represents the infimum over the lengths of all the smooth paths σ connecting the two points x_i and x_j. Next, we arrange the points in

[28] Only applicable if the input dataset has labelled data.

a descending order of dissimilarity as we measure the neighborhood density of 50% data-points with the maximum dissimilarity. The points with the lowest density are highlighted in the projections as potential outliers in data.

- **Option 2:** Here, we use the DBSCAN [13], [19] algorithm to compute the density of all data-points and the data points with the highest density are presented to the user based on their chosen threshold. Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a commonly used technique among researchers [13], [20] to identify data-points with dense neighborhoods.
- **Option 3:** For this option, we present users with the points with the lowest trustworthiness [9] in each embedding and allow them to further investigate these points. The metric trustworthiness [9] quantifies the number of false and missing neighbors for any data-point in its embedding.
- **Option 4:** Here, the data-points with the minimum distance from the decision boundaries are presented as representative points of interest to the users. It is important to note that, this option is only applicable for labelled datasets, where we use Support Vector Machine (SVM) for multi-class classification [40] in order to measure to the Euclidean distance of all data points from the decision boundaries.
- **Option 5:** In this option, the K-means clustering [19] technique is used to present only the cluster centers as representative data points to the users.

As shown in Fig 6.2, once the user interactively selects an $x_i \in X$ (either among the proactively guided points or interactively from the dataset) for further analysis, the next steps of PG-LAPS are initiated.

Step 3: Perform Unsupervised Nearest Neighbors Search:

In this step, as shown in Fig. 6.2, PG-LAPS simultaneously identifies the localities around the chosen point x_i and its low-dimensional counterpart y_i using unsupervised k-nearest neighbor search. For this purpose, PG-LAPS uses the *ball-tree* [41] algorithm. The reason being, ball-tree is well-known [41] for its efficiency with the fast discovery of nearest neighbors in high-dimensional manifolds. To assist with the understanding of our future computations, we formally define the identified local neighborhoods of size k for x_i and y_i as:

$$\begin{aligned} nn_{x_i} &= \{\forall x' \in X | \forall x'' \in X, x' \neq x'' : \pi_{x_i}(x') \leq \pi_{x_i}(x'')\} \\ nn_{y_i} &= \{\forall y' \in Y | \forall y'' \in Y, y' \neq y'' : \pi_{y_i}(y') \leq \pi_{y_i}(y'')\} \end{aligned} \tag{6.4}$$

where, $|nn_{x_i}|$=$|nn_{y_i}| = k$. Once the indexes of the k-nearest neighbors for both x_i and y_i, are identified, the original feature vectors from X for the data-points in nn_{x_i} and nn_{y_i} are combined into matrices that we name Z_{x_i} and Z_{y_i} respectively.

Step 4: Approximate Local Neighborhood:

As the next step, the local neighborhood of the point x_i and its low-dimensional counterpart y_i are approximated by sampling a constant number of data-point samples uniformly at random with a normal distribution centered around each $x' \in Z_{x_i}$ and $y' \in Z_{y_i}$. The primary reason for performing such an approximation is two-fold. Firstly, the process ensures local fidelity by amplifying the locality of the points x_i and y_i without having the need to consider an extremely large value[29] for k. Secondly, such an approximation also ensures normality in the distribution of the feature values in the neighborhood. As a result, such perturbation of local neighborhoods is popular practice in academia and is used by authors such as Ribeiro et al. [37], Plumb et al. [42], and Guidotti et al. [43]. The approximated perturbed neighborhoods for x_i and y_i are combined into feature two vector matrices $\overline{Z_{x_\imath}}$ and $\overline{Z_{y_\imath}}$ respectively.

Step 5: Measure Proximities in the Local Neighborhood:

Next, as shown in Fig. 6.2, the relative proximities $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$ (cf. Section

[29] As shown by Lee and Verleysen [10], with very large values of k, the trustworthiness of the embedding reduces significantly. As the larger the value of k grows, noisy data-points are included in the identified neighborhoods.

6.1) are calculated between the points x_i, y_i and their perturbed neighbors respectively. For feature vectors with continuous values, the Euclidean distance [17] is used as the proximity measure. Alternatively, in the case of feature vectors with a mixture of continuous and categorical values, the Gower dissimilarity [5], [17], [44] is used to measure the proximity between them. The Gower dissimilarity [44] is a popular distance measure for such purposes [17] and can be defined as:

$$dist_\omega(x_i, x') = \sum_{u=1}^{D} \delta_{x_i x' u} \times dist_{\omega_{x_i x' u}} / \sum_{u=1}^{D} \delta_{x_i x' u} \tag{6.5}$$

where u represents an individual feature in $\overline{Z_{x_\imath}}$. For continuous data $dist_\omega$ is calculated as $abs|x_{iu} - x'_u|/range(u)$. For categorical data, $dist_\omega$ is 0 if $x_{iu} = x'_u$, otherwise 1.

Step 6: Compute Feature Influence Scores:

In the following step (cf. Fig. 6.2), we compute the influences of the original features in the formation of the neighborhoods of x_i and y_i. For this purpose, at first, we generate ordered feature-vector matrices $\overline{\overline{Z_{x_\imath}}}$ and $\overline{\overline{Z_{y_\imath}}}$ from the perturbed neighborhoods of x_i and y_i. In these matrices, we order the data vectors in Z_{x_i} and Z_{y_i} according to their descending proximities with x_i and y_i respectively. In parallel, we store the ascending proximity values between x_i, y_i and every point in its perturbed neighborhoods (i.e., the actual $\pi_{x_i}(x')$ and $\pi_{y_i}(y')$ values) in two sets namely $\overline{\pi_{x_\imath}}$. and $\overline{\pi_{y_\imath}}$. Subsequently, from $\overline{\overline{Z_{x_\imath}}}$ and $\overline{\overline{Z_{y_\imath}}}$ we generate two feature distance contribution [45] matrices $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$. These matrices are built based on the concept of feature distance contribution [45], that represents a ratio of the differences in each single feature value in the overall distance between two points. Thereafter, as shown in Fig. 6.2, as the first result of the proposed local quality analysis technique, from $FC_{Z_{x_i}}$ and $FC_{Z_{y_i}}$we generate feature influence explanations

$inf(x_i)$ for x_i and $inf(y_i)$ for y_i using Pearson's correlation[30],[31] [46] between each column in matrix $FC_{Z_{x_i}}$ and set $\overline{\pi_{x_i}}$. Similarly, we also compute $inf(y_i)$ from $FC_{Z_{y_i}}$ and $\overline{\pi_{y_i}}$. In this research, $inf(x_i)$ and $inf(y_i)$ signify influences of the original features of the dataset on the relative dissimilarities between the data-points in the same neighborhood. Features can have either positive, negative, or no influence on the relative proximities among data. Whilst features with positive influences push the data-points further and hence, have a positive Pearson's correlation with the increasing pairwise distances among data-points in the neighborhood. Features with negative influences bring the data-points close to each other and have a negative Pearson's correlation with the same. Similarly, the non-influential features are those that show extremely low or no linear correlation with the increasing pairwise dissimilarities in the neighborhood.

Step 7: Compute Local Divergence:

Finally, local-divergence λ_{x_i} for the selected data-point x_i is computed as:

$$\lambda_{x_i} = w_1 \pi_{\text{inf}(x_i)}(\text{inf}(y_i)) + w_2 \frac{nn_{x_i} \cap nn_{y_i}}{|nn_{x_i}|} + w_3 d_{r_{nn_{x_i}, nn_{y_i}}} \qquad (6.6)$$

As shown in Eq. 6.6, local-divergence is composed as a weighted sum of three components. These include:

- $\pi_{inf(x_i)}\left(inf(y_i)\right)$: Signifies the *cosine* distance between $inf(x_i)$ and $inf(y_i)$ (cf. Step 6). That represents the discrepancy between the feature influences scores for the neighborhood of x_i in the original dataset versus in the embedding.

- $nn_{x_i} \cap nn_{y_i} / |nn_{x_i}|$: Represents the false and missing neighbors for x_i in the embeddings. Here we compute a ratio between the number of the

[30] Our experiments with Spearman's correlation (non-parametric) returned the same values up-until the second decimal point.

[31] Perason's correlation is a well-known effect size estimator.

preserved neighbors with the total number of k nearest neighbors of x_i considered for the analysis.

- $d_{r_{nn_{x_i},nn_{y_i}}}$: Represents the difference between the relative orders (or ranks [9]) of neighbors in the neighborhoods of x_i and y_i. This component measures whether an embedding could preserve the ordinal relationships among the data points in the original neighborhood.

In Eq. 6.6, w_1, w_2, and w_3 signify user-defined (scalar) weights of the three components of λ_{x_i}, by default in our computations, we consider each component of local divergence to be equally weighted (i.e., a weight of 0.33). However, during a user-driven computation of local-divergence the users are enabled to alter these weights in any way they seem fit. In either case w_1, w_2, and w_3 sum up to 1.

6.3.2 Proactively Guided Computation of Global-Divergence in Embeddings

The process of Proactively Guided GAPS (PG-GAPS) can be divided into five distinct steps as shown in Fig. 6.3. Overall, the process aims at examining the preserved global structure in an embedding. For this purpose, PG-GAPS suggests two possible types of representative subsets of the data and enables a user-driven computation of the metric global-divergence for any embedding. The detailed steps of PG-GAPS are as follows:

Step 1: Pre-process Input Data and Obtain Embedding:

At the very beginning of the analysis, PG-GAPS performs similar pre-processing on the input dataset as discussed in Section 6.3.1. Here, as a part of the pre-processing, PG-GAPS estimates the intrinsic dimensionality d of the input dataset (cf. Eq. 6.2) followed by performing a proactive grid search of hyperparameters for the selected DR algorithms.

Step 2: Proactively Suggest a Representative Subset:

The technique PG-GAPS analyzes the quality of the global structure in embeddings

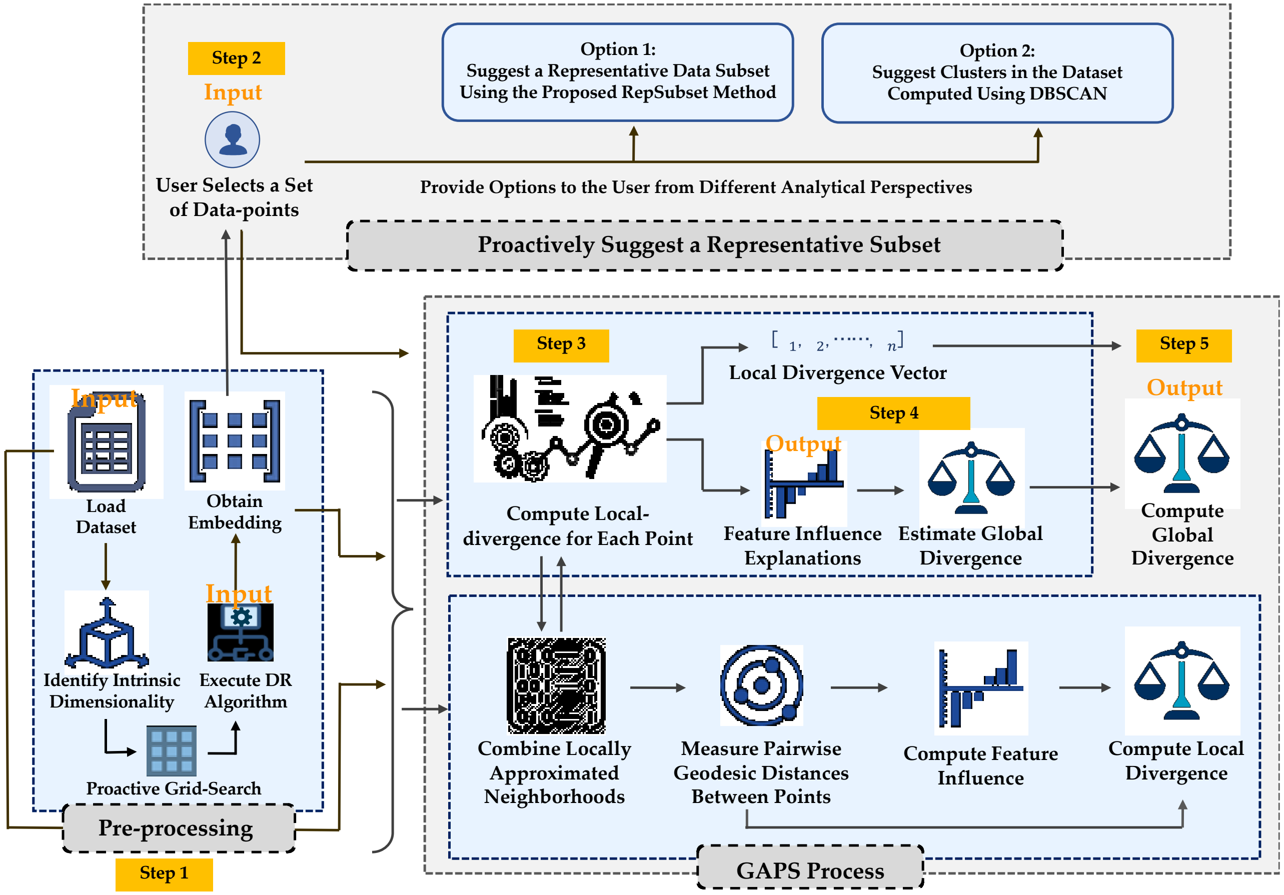

Step 2
Input
User Selects a Set of Data-points
Option 1:
Suggest a Representative Data Subset Using the Proposed RepSubset Method
Option 2:
Suggest Clusters in the Dataset Computed Using DBSCAN
Provide Options to the User from Different Analytical Perspectives
Proactively Suggest a Representative Subset
Input
Load Dataset
Obtain Embedding
Identify Intrinsic Dimensionality
Input
Execute DR Algorithm
Proactive Grid-Search
Pre-processing
Step 1
Step 3
Compute Local-divergence for Each Point
$[_1, _2, \cdots, _n]$
Local Divergence Vector
Step 4
Output
Feature Influence Explanations
Estimate Global Divergence
Step 5
Output
Compute Global Divergence
Combine Locally Approximated Neighborhoods
Measure Pairwise Geodesic Distances Between Points
Compute Feature Influence
Compute Local Divergence
GAPS Process

Figure 6.3: Overview of the PG-GAPS technique. The technique primarily contains three components namely: Pre-processing, Proactive Guidance, and the GAPS process. The overall process steps for PG-GAPS are numbered in an ascending order and are highlighted in yellow.

from a subset X_S of non-redundant data-points in X. In order to perform a judicious selection of X_S, PG-GAPS presents users with proactive guidance for possible options. Although, users are allowed to interactively select data-points of their choice around the manifold ignoring these guidelines. As shown in Fig. 6.3 PG-GAPS, as a part of the proactive guidance PG-GAPS recommends users with a representative subset of the original data that provides maximum coverage [20] of the input dataset. For this purpose, as a part of PG-GAPS we present a novel representative subset selection technique RepSubset for diversified representative subset selection that is based on the notion of density and dissimilarities among the data points in the dataset. Here, we formally define the representative subset as $X_S \subseteq X$, such that X_S provides a good representation of X and $n_S \ll n$, where n_S and n are the sizes of X_S and X respectively. The step-by-step process of RepSubset is presented in Fig. 6.4 and is discussed below:

Step i: Compute density of all data points in X. Order the points in terms of their decreasing density. Assign a status label of "active" to all points.

Step ii: Add the "active" point with the highest density to X_S. Update the status label of that point to be "inactive".

Step iii: Find the k-NN of the recently added data point. Update their status labels to be "inactive".

Step iv: Compute pairwise geodesic distances between the last added data point to X_S and the remaining "active" points.

Step v: Select the point with the highest geodesic distance with the last added data point to X_S and add the selected point to the X_S.

Step vi: Identify the k nearest neighbors of the recently added data point to X_S and update their status labels to be "inactive".

Step vii: If $n_S < B$, go to step ii. Else, turn the status of all inactive points that do not already belong to X_S to "active" and go to step ii.

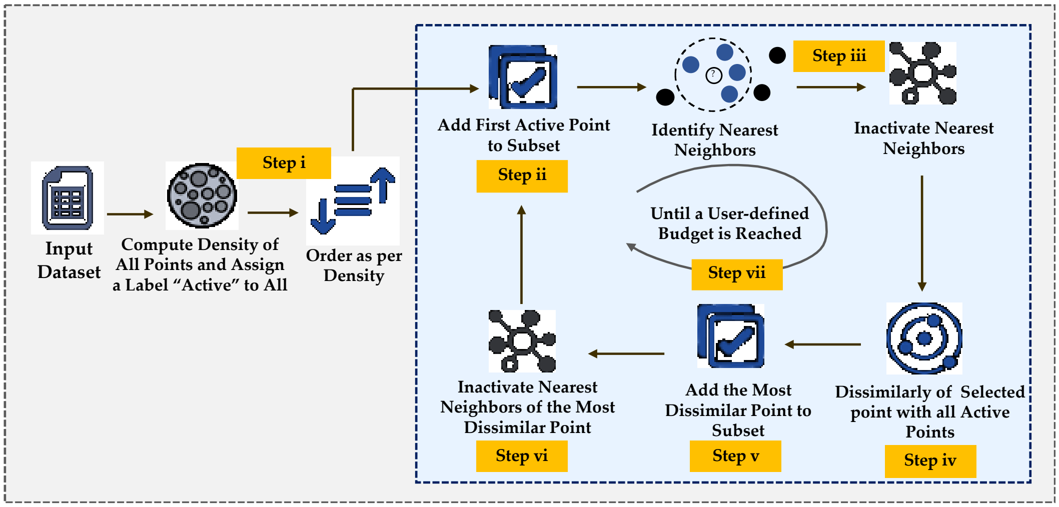

Input Dataset
Compute Density of All Points and Assign a Label "Active" to All
Step i
Order as per Density
Add First Active Point to Subset
Step ii
Identify Nearest Neighbors
Step iii
Inactivate Nearest Neighbors
Until a User-defined Budget is Reached
Step vii
Dissimilarly of Selected point with all Active Points
Step iv
Add the Most Dissimilar Point to Subset
Step v
Inactivate Nearest Neighbors of the Most Dissimilar Point
Step vi

Figure 6.4: Overview of the RepSubset technique. This novel iterative technique presents a representative subset of the input data as a part of the PG-GAPS process. The overall process steps for RepSubset are numbered in roman letters and are highlighted in yellow. Each of these process steps (i.e., Step i to Step ii shown in the figure) are discussed in detail in the Step 3 of Section 6.3.2.

As discussed in Section 6.1.3, the RepSubset technique is presented as an enhancement over the state-of-the-art subset selection methods. A detailed evaluation of the proposed algorithm is presented in Section 6.4.3. As an alternative to RepSubset, PG-GAPS also presents users with subsets of data-points that form clusters in the dataset. The purpose of this alternative is to provide users with multiple perspectives for selecting subsets from data. Here, following popular research [13], [20], the algorithm DBSCAN is used to identify clusters in the data. In both the cases, we let the users determine the number of instances they are willing to investigate and represent this number with an investigation budget B. Once the user selects their desired representative subset from the data, the next steps of the PG-GAPS process are initiated.

Step 3: Compute Local-divergences of Points in Subset:

As shown in Figure 6.3, after obtaining the data-points in X_S, we individually assess the local neighborhoods for each $x_i \in X_S$ and $y_i \in Y_S$ (where, $y_i = f(x_i)$) using the PG-LAPS process (cf. Section 6.3.1) and compute their local divergence scores. This step is necessary as, for maintaining an accurate global structure in the embeddings, not only the relative closeness among the points in the same neighborhoods should also be retained. These computed local-divergence scores for all the data points in X_S and their low-dimensional counterparts in Y_S are then composed into two sets that we name λ_{X_S} and λ_{Y_S} respectively. Subsequently, we combine the perturbed local neighborhoods for each $x_i \in X_S$ and $y_i \in Y_S$ obtained from their local structural analysis (cf. Section 6.3.1, Step 4) are combined into two feature vector matrices Z_{X_S} and Z_{Y_S} respectively.

Step 4: Estimate the Global Divergence of the Subset:

In order to estimate the global divergence of X_S, as our first step, we compute pairwise proximities among the points in the two feature vector matrices Z_{X_S} and Z_{Y_S}. Here as our measure of proximity, the pairwise geodesic distances among the data-points is used. Then, the indexes of the data-points in both high and low dimensions are ordered in terms of descending proximity. From these ordered indexes, the global feature distance contributions [45] are obtained in the same way as discussed in the Step 6 of Section 6.3.1. From the feature distance contributions, the influence of each feature in the relative proximities among the data-points are computed using the Pearson's correlation among the feature distance contributions and the ordered proximity values among the data-points Finally, we obtain an estimate of the global divergence $\lambda_{\widehat{X_S}}$ for X_S as a weighted sum of the disagreements in the overall estimation of the feature influences, and the disagreements in the neighborhood structures of all points in X_S and Y_S.

Step 5: Compute Global Divergence Score:

Finally, a Global-Local Approximation (GLA)[32] [48] is performed to obtain an additive blending of local approximations to form a globally-valid approximation. To compute global-divergence, prior to the unification of the local-approximations, the ratios of the estimate of global divergence with the local-divergences for each

[32] Research [47], [48] shows the local approximation of divergence for each data-point in X_S is the most effective near the point where it was calculated. However, the accuracy of such local approximations can deteriorate [48] as we move away from the point where it was constructed. Alternatively , a global approximation may not be accurate for every data-point in the manifold, however, its quality does not deteriorate with distance.

points in X_S are used as scaling factors [47] for each local-divergence score. As shown in Fig. 6.3, we define global-divergence as:

$$\lambda_X = \sum_{j=1}^{B} \frac{\lambda_{X_{S_j}}}{\lambda_{\hat{X}_S}} \lambda_{X_{S_j}} \tag{6.7}$$

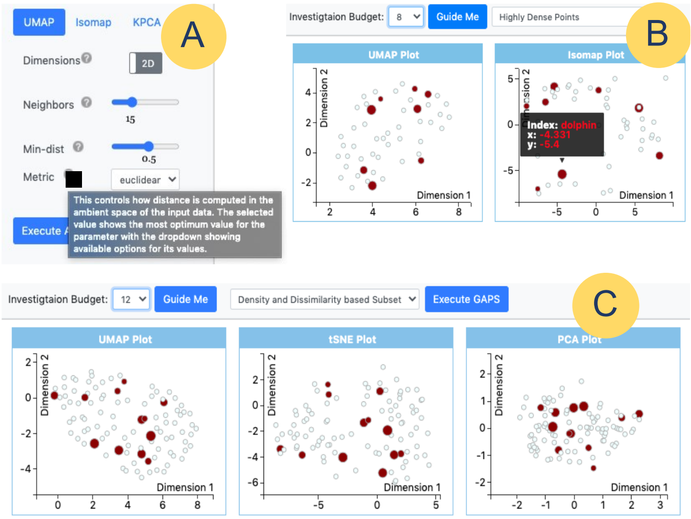
where, λ_X represents an additive blending of scaled local-divergence scores.

6.3.3 The VisExPreS Interface Design

As the final contribution of this research, we incorporated the two methods into a visual analysis toolkit that we named ***Vis**ual **Ex**planations of **Pre**served **S**tructure (VisExPreS)*. It is important to note that, although the VisExPreS toolkit can be used to compute other quality metrics for evaluating embeddings, the interface is primarily designed for PG-LAPS and PG-GAPS. The views in VisExPreS can be broadly categorized into two groups namely: (1) Structural Quality Analysis View (cf. Fig. 6.1); and (2) Feature Analysis View. In this section, we justify the system design of the VisExPreS toolkit based on our design goals described earlier.

6.3.3.1 Proactive Guidance for Representative Data

As shown in Fig. 6.5, the interface of the VisExPreS toolkit is designed to provide proactive guidance in both the PG-LAPS and PG-GAPS processes as a part of its Structural Quality Analysis Views (cf. Fig. 6.1). This guidance is presented to the users in the following two ways. First of all, as depicted in Fig. 6.5a, the interface assists users with the selection of hyperparameter values for the chosen DR methods. Here, with a mouseover operation on the '?' icon next to the hyperparameter name, users are presented with a brief description of the hyperparameter itself. Alongside, using visual elements such as sliders and dropdown boxes (cf. Fig. 6.5), VisExPreS guides users with the possible values of the hyperparameter. Here, the most optimum values of the hyperparameter for any chosen DR method and the corresponding dataset (i.e., obtained from the proactive grid search mentioned in Section 6.3.1) is presented as the default value of the hyperparameter by the interface. Whereas, the value ranges in the sliders and



Dimensions: 2D

Neighbors: 15

Min-dist: 0.5

Metric: euclidear

This controls how distance is computed in the ambient space of the input data. The selected value shows the most optimum value for the parameter with the dropdown showing available options for its values.

Investigtaion Budget: 8

Highly Dense Points

Index: dolphin
x: -4.331
y: -5.4

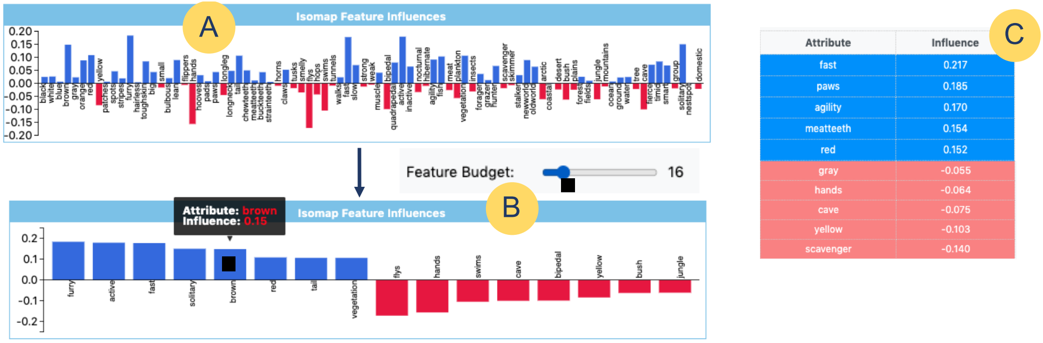
Investigtaion Budget: 12

Density and Dissimilarity based Subset

Figure 6.5: Proactive guidance in the VisExPreS interface. The part A of the figure shows guidance on hyperparameter values in the embeddings. As the parts B and C of the figure show proactive guidance with representative data-points and representative data-subsets respectively.

dropdown boxes represent other suitable values for the hyperparameter for the chosen dataset. At this point, the users can choose to proceed with the suggested value for the hyperparameters or can alter the values as per their choice.

Secondly, as shown in Figures 4b and 4c, VisExPreS proactively guides users with the selection of representative data-points (i.e., for PG-LAPS) and data subsets (i.e., for PG-GAPS) when analyzing an embedding. Here, at first, users are enabled to select an investigation budget to limit the number of data-points that they want to see in the suggestions. This step is necessary as it can prevent the user from being overwhelmed with the data-point options to select from. Once the investigation budget is selected, as depicted in Fig. 6.5b, for PG-LAPS, VisExPreS presents five different types (cf. Section 6.3.1) of representative data-points to the users with the help of a drop-down list. Once the user makes a selection from one of these options, the respective representative data-points are visualized on the embeddings using different color and radius sizes. That is, the represented data-points are highlighted with the color dark-red (cf. Fig. 6.5b and 6.5c) and the most representative data-point in the chosen category is presented with the largest radius size. The radius



Attribute	Influence
fast	0.217
paws	0.185
agility	0.170
meatteeth	0.154
red	0.152
gray	-0.055
hands	-0.064
cave	-0.075
yellow	-0.103
scavenger	-0.140

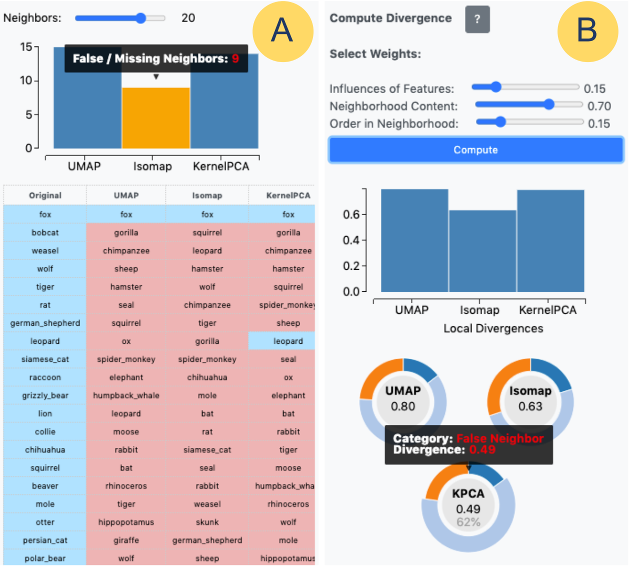
Figure 6.6: Comparison of feature influences between the embeddings and the original dataset. The parts A and B show that by reducing the feature budget the users can investigate only most highly influential features in the dataset and compare them with the feature influences in the original dataset as shown in part C.

sizes of the other representative data-points gradually decrease based on their representativeness of their respective categories.

In the case of PG-GAPS, as shown in Fig. 6.5c, users are presented with two options for receiving suggestions on representative subsets (cf. Section 6.3.2). Similarly, as PG-LAPS, for PG-GAPS the data-points in the suggested data-subsets are highlighted with the color dark-red and the sizes of their radius show their representativeness in the subset. However, in the case of both PG-LAPS and PG-GAPS, users can choose to ignore the proactively suggested data-points, and interactively select their own point(s) of interest from the scatter plots of the embeddings for further analysis.

6.3.3.2 Simultaneous Investigations of Embeddings

In order to facilitate a side-by-side evaluation of embeddings, the VisExPreS interface presents the users with the option of making a selection of their choice of DR algorithms to compare on their choice of dataset (cf. Fig. 6.1, region-B). The design of the VisExPreS interface requires users to select at least three algorithms for this comparison. As shown in the regions C and E of Fig, 1, upon making a selection of the DR methods and deciding on the choice of their respective hyperparameter values, VisExPreS simultaneously presents the user with interactive scatter plots of the embeddings for the input dataset obtained from the



Neighbors: 20

Select Weights:

Influences of Features: 0.15
Neighborhood Content: 0.70
Order in Neighborhood: 0.15

Original	UMAP	Isomap	KernelPCA
fox	fox	fox	fox
bobcat	gorilla	squirrel	gorilla
weasel	chimpanzee	leopard	chimpanzee
wolf	sheep	hamster	hamster
tiger	hamster	wolf	squirrel
rat	seal	chimpanzee	spider_monke
german_shepherd	squirrel	tiger	sheep
leopard	ox	gorilla	leopard
siamese_cat	spider_monkey	spider_monkey	seal
raccoon	elephant	chihuahua	ox
grizzly_bear	humpback_whale	mole	elephant
lion	leopard	bat	bat
collie	moose	rat	rabbit
chihuahua	rabbit	siamese_cat	tiger
squirrel	bat	seal	moose
beaver	rhinoceros	rabbit	humpback_wha
mole	tiger	weasel	rhinoceros
otter	hippopotamus	skunk	wolf
persian_cat	giraffe	german_shepherd	mole
polar_bear	wolf	sheep	hippopotamus

Figure 6.7: (A) Neighborhood Analysis (B) Divergence computation in the VisExPreS Interface.

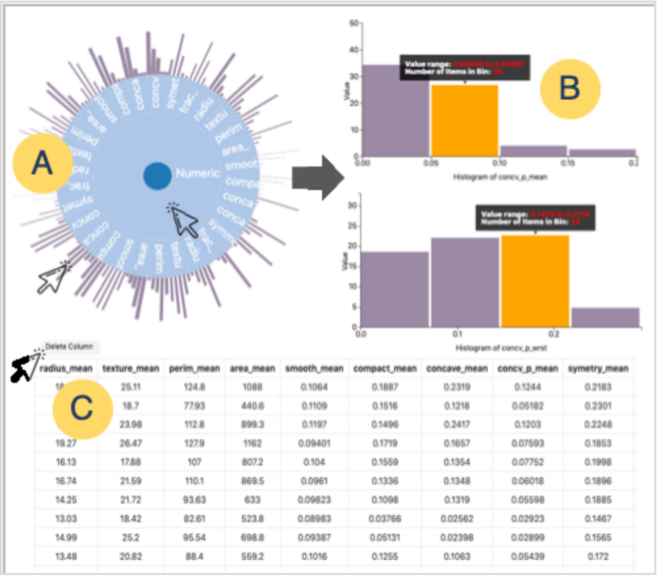
chosen algorithms. Also, a mouseover operation on the embeddings show further details of individual data-points. Apart from the interactive scatter plots, all analysis results from PG-LAPS and PG-GAPS are presented simultaneously for the chosen DR methods. This allows users to perform a side-by-side comparison among the neighborhood structures of the data-points under investigation in multiple embeddings, as well as the influences of the original features (cf. Fig. 6.1, region-F) in the neighborhoods.

6.3.3.3 Contributions of the Original Features

As shown in Fig. 6.6, the influences or the contributions of the original features in the dataset in the structural formation of the neighborhood of the chosen data-point is presented using back-to-back bar graphs for each of the chosen DR algorithm. The reason behind using back-to-back bar graphs for this purpose is to depict both positive and negative influences (cf. Section 6.3.1) of the features simultaneously. At the same time, the VisExPreS interface also shows the influences of the features in the original dataset (cf. Fig. 6.6c) allowing users to compare the differences between the original dataset and the embeddings. Upon looking at the Fig. 6.6a and the at the Fig. 6.6c, it can be argued that comparing the differences in feature influences can be difficult for datasets with a large number (i.e., <20) features. As shown in Fig. 6.6b, the VisExPreS interface provides a solution for this problem by allowing users to control the number of features that they are willing to investigate in the embeddings. Here, upon reducing the budget, the back-to-back bar graphs only show those features with the highest positive and negative influences in the structural formation of the neighborhood(s). At this point, the users can compare only those features with highest positive and negative influences in the original data and in the embeddings and check whether the embeddings have the same influences for the same features or not.

6.3.3.4 Interpretable Assessment of Structural Quality

The design of the VisExPreS interface aims to assist users to make the evaluation of the embeddings as interpretable as possible for both novice and expert users. For this purpose, VisExPreS not only presents textual annotations for multiple aspects of analysis (cf. Fig. 6.6a) but also makes use of the visual interactive interface to assist users in a better engagement with the analysis process (cf. Fig. 6.7). For example, firstly, the colors and radius sizes of the points in the interactive scatter plots are chosen carefully to effectively highlight the representative data-points (and subsets) and their respective neighborhoods (cf. Fig. 6.7a). Secondly, the colors in the back-to-back bar graphs are intentionally chosen to be bright enough so that they stand out of the text showing the attribute names (cf. Fig. 6.6a and 6.6b). Also, it is made sure that in the back-to-back bar graphs the y-axes of each graph shows the same value range (cf. Fig. 6.6a and 6.6b). Thirdly, the neighborhoods of the selected points are enabled to be effectively compared using the tabular representation (cf. Fig. 6.7a) highlighting the order of the neighbors in the original dataset and in the embeddings. Here, with only a single glance at the table the users can notice the discrepancies in the neighborhood. For example, instead of looking at the indexes of the neighbors, users need to only check the colors in the table. As shown in Fig. 6.7a, the left-most column in the table shows the neighborhood of the chosen data-point in the original data and it is highlighted in light purple. In this table, the user only needs to check in which other columns in



radius_mean	texture_mean	perim_mean	area_mean	smooth_mean	compact_mean	concave_mean	concv_p_mean	symetry_mean
[illegible]	25.11	124.8	1088	0.1064	0.1887	0.2319	0.1244	0.2183
[illegible]	18.7	77.93	440.6	0.1109	0.1516	0.1218	0.05182	0.2301
[illegible]	23.98	112.8	899.3	0.1197	0.1496	0.2417	0.1203	0.2248
19.27	26.47	127.9	1162	0.09401	0.1719	0.1657	0.07593	0.1853
16.13	17.88	107	807.2	0.104	0.1559	0.1354	0.07752	0.1998
16.74	21.59	110.1	869.5	0.0961	0.1336	0.1348	0.06018	0.1896
14.25	21.72	93.63	633	0.09823	0.1098	0.1319	0.05598	0.1885
13.03	18.42	82.61	523.8	0.08983	0.03766	0.02562	0.02923	0.1467
14.99	25.2	95.54	698.8	0.09387	0.05131	0.02398	0.02899	0.1565
13.48	20.82	88.4	559.2	0.1016	0.1255	0.1063	0.05439	0.172

Figure 6.8: The Feature Analysis View in the VisExPreS Toolkit.

the table there is another light purple cell. Such a cell means that, the exact same neighbor and its respective order in the neighborhood was preserved in this embedding. Here, with just one glance at the table the user can see which algorithm has preserved the neighborhood order (i.e., the third component of the local divergence metric) the best among all. As for the false and missing neighbors in the embeddings, VisExPreS presents a bar graph as shown in Fig. 6.7a. Finally, in case the differences are still not easily visible for the user, or if the users are not sure how the metrics local and global divergence are computed, the VisExPreS interface provides its users with two solutions. Firstly, as shown in Fig. 6.7b, upon a mouseover operation over the '?' icon next to the "Compute divergence" label, VisExPreS explains the rationale behind the computation of local or global divergences in detail. Secondly, as shown in Fig. 6.7b, VisExPreS presents the users with a set of three donut charts that quantify the discrepancies of the individual aspects (cf. Section 6.3.1) of the metrics local or global divergences.

6.3.3.5 User-driven Computation of Quality Criteria

To support a user-driven assessment of embeddings, the VisExPreS interface puts the user in-charge of the analysis in several different ways. First of all, as shown in

Fig. 6.1c, VisExPreS allows users to select the DR algorithms and their respective hyperparameter values of their choice. Secondly, as shown in Fig. 6.5b, the point(s)-of-interest to be chosen for the analysis of local and global structures completely depend on the user's preference. Thirdly, the computation of the metric local and global divergence primarily depends on the user's decision of the relative weights for the three components of the metrics (cf. Section 6.3.1). As shown in Fig. 6.7b, with the help of the interactive sliders the users are allowed to modify the default weights of the different components of the metrics, this alters the values of the two-output metrics of PG-LAPS and PG-GAPS. Finally, the feature analysis view of the VisExPreS interface (cf. Fig. 6.8) allows users to visually analyze and interact with the features in the original dataset. Figure 7 shows an example of the feature analysis view in VisExPreS. The view can be divided into three primary regions as shown in this figure. The region A depicts a histogram amalgamated sunburst diagram that groups features based on their types (i.e., numeric or categorical) and shows the value distribution of each feature around the perimeter of the sunburst diagram. Here, the visual effectiveness of the sunburst is obtained by grouping the features into numeric and categorical ones. To utilize the visual effectiveness of the sunburst even more, we leave a more advanced grouping of the attributes [49] as our future work. The region B in Fig. 6.8 zooms into the distributions of the individual features in the dataset and shows the histograms of their value distributions. Finally, region C shows the actual records in the data. The region C also allows users to interactively remove one or more features from the analysis that may seem less influential to the user.

6.4 Experimental Evaluation

In this section, we present the results of our detailed experimental evaluation of the VisExPreS toolkit as well as the proposed RepSubset algorithm. This section is primarily divided into two parts. In the first part, we perform an exhaustive evaluation of VisExPreS whilst in the second part we compare RepSubset with some of its closest competitors. Further details regarding our experimental evaluation of VisExPreS are presented as supplemental materials.

6.4.1 Case Study of Data Analysis with VisExPreS

The VisExPreS interface is primarily designed to evaluate the quality of embeddings obtained by executing different DR methods on the same dataset. As a result, the target user type for the toolkit are data scientists who execute DR algorithms as a part of their day-to-day analysis of datasets and need to make decisions regarding which DR method would be more suitable for their subsequent analysis[33] given their input dataset and analytical context.

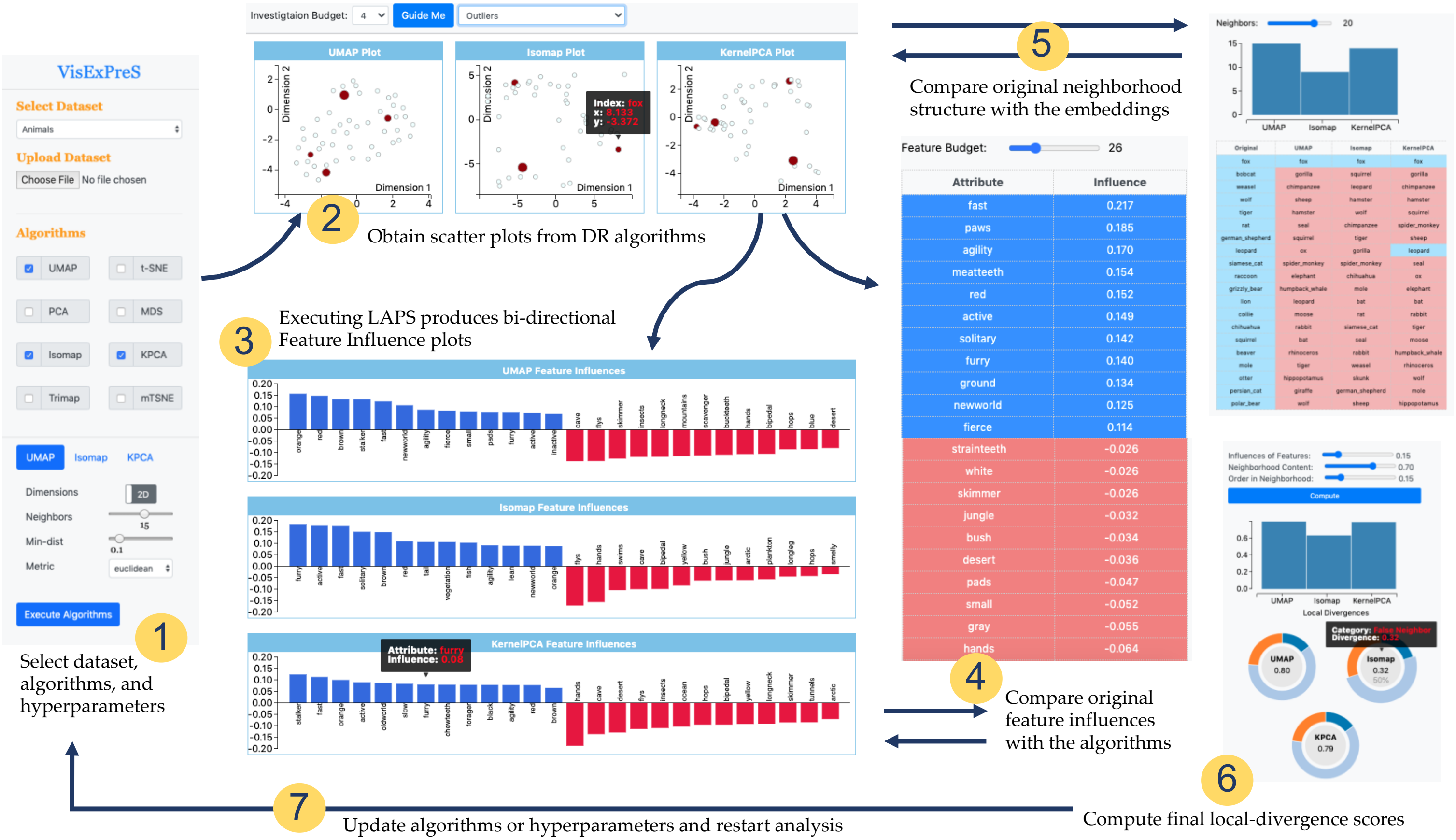
In this section, we present an example of data analysis using the VisExPreS interface that demonstrates its utility in a user-driven evaluation of embeddings. For a better understanding of the toolkit's utility, another such example can be found in our supplemental material and a detailed demonstration video of embedding quality analysis with VisExPreS can be found at: https://bit.ly/3fQsBD0

Analysis of the Animals Dataset using VisExPreS

Alice is a zoological data scientist who studies the behavioral patterns of wild animals and performs predictive modelling of their appearances in the wild. For this purpose, Alice has obtained an open source image dataset[34] [17] that contains 30,475 images of 50 animals that are classified using 85 numeric features. Alice decides to perform DR on the input data prior to training a predictive model with it. Alice has limited experience with DR, but she is aware of that several algorithms exist for this purpose. At this point, Alice decides to explore the data using the VisExPreS interface (cf. Fig. 6.9). In order to enhance visual clarity of the embeddings she uses a subset of 100 points from the dataset for her analysis.

[33] An example of such subsequent analysis can be training a machine learning model with the selected embedding.

[34] https://cvml.ist.ac.at/AwA/



VisExPreS

Select Dataset: Animals

Upload Dataset: Choose File No file chosen

Algorithms: ☑ UMAP □ t-SNE □ PCA □ MDS ☑ Isomap ☑ KPCA □ Trimap □ mTSNE

UMAP Isomap KPCA

Dimensions: 2D

Neighbors: 15

Min-dist: 0.1

Metric: euclidean

Execute Algorithms

1 Select dataset, algorithms, and hyperparameters

Investigtaion Budget: 4 Guide Me Outliers

2 Obtain scatter plots from DR algorithms

Index: fox
x: 8.133
y: -3.372

3 Executing LAPS produces bi-directional Feature Influence plots

Attribute: furry
Influence: 0.08

5 Compare original neighborhood structure with the embeddings

Neighbors: 20

Feature Budget: 26

Attribute	Influence
fast	0.217
paws	0.185
agility	0.170
meatteeth	0.154
red	0.152
active	0.149
solitary	0.142
furry	0.140
ground	0.134
newworld	0.125
fierce	0.114
strainteeth	-0.026
white	-0.026
skimmer	-0.026
jungle	-0.032
bush	-0.034
desert	-0.036
pads	-0.047
small	-0.052
gray	-0.055
hands	-0.064

4 Compare original feature influences with the algorithms

Original	UMAP	Isomap	KernelPCA
fox	fox	fox	fox
bobcat	gorilla	squirrel	gorilla
weasel	chimpanzee	leopard	chimpanzee
wolf	sheep	hamster	hamster
tiger	hamster	wolf	squirrel
rat	seal	chimpanzee	spider_monkey
german_shepherd	squirrel	tiger	sheep
leopard	ox	gorilla	leopard
siamese_cat	spider_monkey	spider_monkey	seal
raccoon	elephant	chihuahua	ox
grizzly_bear	humpback_whale	mole	elephant
lion	leopard	bat	bat
collie	moose	rat	rabbit
chihuahua	rabbit	siamese_cat	tiger
squirrel	bat	seal	moose
beaver	rhinoceros	rabbit	humpback_whale
mole	tiger	weasel	rhinoceros
otter	hippopotamus	skunk	wolf
persian_cat	giraffe	german_shepherd	mole
polar_bear	wolf	sheep	hippopotamus

Influences of Features: 0.15

Neighborhood Content: 0.70

Order in Neighborhood: 0.15

Compute

Category: False Neighbor
Divergence: 0.32

6 Compute final local-divergence scores

7 Update algorithms or hyperparameters and restart analysis

Figure 6.9: Local structural quality analysis of the Animals Cancer dataset with PG-GAPS using VisExPreS. Each step of the assessment for the interactively chosen data-point ‘fox’ is highlighted with numbers and annotated with the details of the respective step.

Fig. 6.9 shows the flow of Alice's analysis on the Animals dataset. Once Alice loads her dataset into the VisExPreS system and selects three DR algorithms she has heard the most about (i.e., UMAP [46], Isomap [4], and KernelPCA [4]), VisExPreS executes a proactive grid search on her chosen algorithms for the input data and presents her with some suggestions for the hyperparameter values for the algorithms. Being a novice data scientist, Alice hovers her mouse pointer over the '?' icons next to these hyperparameters that explain the purpose of the parameter using tooltips (cf. Fig. 6.5). The tooltips also state the optimum values for these hyperparameters for her dataset is already selected for her by the toolkit. Alice decides to proceed with the suggested values of these hyperparameters for her analysis.

Upon choosing to execute the DR algorithms, the VisExPreS interface presents Alice with three embeddings for her input dataset. Here, Alice decides to explore the proactive guidance from the toolkit in order to select a representative data-point from the dataset for further analysis (cf. Fig 6.8.1). With her investigation budget set at 4, Alice looks for the outliers in the dataset. VisExPreS identifies the points *dolphin*, *tiger*, *sheep*, and *fox* to be the outliers. However, Alice chooses to investigate the point *fox* for further analysis. Now, upon executing LAPS on point *fox*, the VisExPreS interface shows her with the neighborhood for the chosen point on the scatter plots as the feature influence explanation bar-graphs for the point are shown just below the scatter plots (cf. Fig. 6.9.3). Here, the VisExPreS interface allows Alice to compare the neighborhood and feature influences for *fox* in the embeddings and in the original dataset. Here, Alice notices that, the original feature influences presented in the right-side bar (cf. Fig. 6.9.4) show that, for *fox* the features *fast*, *paws*, *agility*, *meatteeth, red, active, furry,* and *solitary* have the highest positive influences. On the other hand, features *scavenger*, *yellow*, *cave* and *hands* have the most negative influences on its neighborhood structure. Some features such as *tunnel*, *bipedal*, and *forager* have little or no influences on the relative proximities in the neighborhood of *fox*. To investigate the back-to-back bar graphs, Alice reduces the feature budget in Fig. 6.9.4 to 16. Here, Alice only wants to investigate which of the embeddings have preserved similar influences for the most positive and most negatively influencing features in the original dataset for the neighborhood of *fox.* Upon carefully looking at the bar-graphs, Alice notices that Isomap has preserved the positive feature influences for features such as *fast, furry, red, agility,* and *solitary.* Also*,* in terms of the most negative feature influences Isomap has preserved the influences of *hands*, *yellow*, and *cave*. Whereas, in the embeddings produced by UMAP and KernelPCA, Alice cannot find many common features with high positive and negative influences as in the original data.

Upon comparing the neighborhood of *fox* in the original dataset (cf. Fig. 6.9.5) with the neighborhoods in the embeddings, Alice notices among the 20 neighbors visible in the scatter plots, Isomap has only 9 false and missing neighbors in its embedding, whereas both UMAP and KernelPCA has 14 false or missing neighbors. In terms of the preserved orders of neighbors, KernelPCA has preserved only the point *leopard* in its actual position in the original dataset. The two remaining algorithms have completely messed up the orders of the points in the neighborhoods.

At this point, Alice decides to compute the *local-divergence* score (cf. Fig. 6.9.6) for the chosen point. Hence, at first Alice looks into the local-divergence scores presented by the VisExPreS interface, where equal weights were allocated for all three components of the metric (cf. Section 6.3.1). Here, Alice hovers her mouse pointer on the '?' icon next to the "Compute divergence" label and learns about the calculation of the local-divergence for the chosen point. She also observes the donut charts, where she can see the contributions of each of these components in the final value of local divergence score for the three DR methods. Here, Alice notices that Isomap has lowest local di-vergence score with the default weights. Upon increasing the weight for neighborhood content component, Alice notices that Isomap still performs better than others. Even with a higher weight to the feature influence component Isomap performs the best among the three. At this point Alice decides to analyze the point dolphin and compare UMAP, Isomap, PCA for this point. Alice repeats the analysis for 50 points in the dataset and notices that on an

Table 6.1: Results of Usability Analysis on VisExPreS

Analytical Aspects	Novice	Experts
Effectiveness - mean (SD)	0.89 (0.06)	0.96 (0.03)
Efficiency - mean (SD)	55.92 (4.20)	41.09 (8.32)
SUS Score - mean (SD)	69.29 (6.07)	83.00 (4.47)
Design Goals (Fleiss' Kappa)	0.27	0.48

Note: The table above summarizes the results of our user study. Here we compute the usability metrics defined in ISO 9241-11 standard using 12 human subjects. Additionally, the last row presents the users' agreements on the fulfillment of the design goals of the VisExPreS interface.

average Isomap has performed better than all other algorithms. So, she chooses to execute Isomap on her dataset prior to training her predictive model with the data.

6.4.2 Usability Analysis of the VisExPreS Interface

In this section, we perform a detailed user study and analyze the usability of the VisExPreS interface. Based on the guidelines of Georgsson et al. [50], we used metrics in the International Organization for Standardization (ISO) 9241-11 standard[35] and quantify the usability using **effectiveness**, **efficiency**, and **satisfaction**. In this section, at first we discuss the setup of our user study followed by quantitative analysis of its results. More detailed results of our analysis are presented as supplemental materials.

6.4.2.1 Experimental Setup and Participants

Following the guidelines of Stahnke et al. [5], Ribeiro et al. [37], and Georgsson et al. [50] we performed our user-study on VisExPreS using 12 human subjects. The study was performed under controlled conditions [50] where the study organizers observed and assessed the interactions of the study participants with the VisExPreS interface. The participants were carefully chosen as a group of graduate students

[35] https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

and industry professionals with a strong background in computing. However, among the participants, only five members had some understanding of DR, as the remaining seven were completely unfamiliar with the topic. In this study, we considered the former five as expert users and the latter seven as non-experts or novice users. During this study the participants were asked to analyze the Breast Cancer dataset [17] using the VisExPreS interface. Prior to the study, all participants were given a 60-minutes walk-through of the dataset along with the different functionalities of the VisExPreS toolkit. Next the participants were given 75 minutes time to analyze at least 18 data-points[36] from the proactively guided categories of PG-LAPS (cf. Section 6.3.1) and two rounds of global quality analysis using the representative subsets suggested by PG-GAPS. At the end of their analysis, the responses from the study participants were recorded and the usability metrics from the ISO 9241-11 standard were computed from them.

6.4.2.2 Measuring Effectiveness of VisExPreS

According to the ISO 9241-11 standard, effectiveness is one of the most important attributes of usability that is measured using the metric *degree of task completion* [50]. Here we measured given the total number of tasks, how many of the tasks could be successfully completed by the participants. In our case, the total number of tasks that was given to each study participants were 20. Following the footsteps of Georgsson et al. [50], we encoded the task completion in three different ways as: (i) a score of 1, in case a task was completed by a participant without any assistance, (ii) a score of 0.5, for situations when a participant needed minor assistance from the study organizers to complete their task, (iii) a score of 0, when a participant could not complete a task. We summarize our analysis results in Table 6.1. Here, similarly as Georgsson et al. [50], we present the *mean* and *standard deviation (SD)* of task completion rate for novice and expert users. Considering, a task completion rate of 100% to be an ideal case. However, as a common consensus among

[36] The participants were asked to investigate 4 points from each proactively guided category of PG-LAPS except for the cluster centers. Where, the users were asked to investigate 2 cluster centers only as the dataset had only 2 labels 'malignant' and 'benign'.

researchers [50], any score above 78% signifies an acceptable rate of effectiveness in a system. As shown in Table 6.1, for the novice and expert users the mean task acceptance rate were 88% and 95% respectively. In both cases, the standard deviation was less than 0.06. This confirmed that the VisExPreS interface enabled a high level of efficiency for novice and expert users.

6.4.2.2 Measuring Effectiveness of VisExPreS

ISO 9241-11 standard has identified efficiency as the second important measure of a system's usability that is measured using the level of effort and the amount of resources that were used by study participants to complete as task. Again, following the guidelines Georgsson et al. [50], we measured efficiency using the amount of time (in minutes) that was spent by the study participants in order to complete each of the given tasks. For this purpose, we asked the participants to use a timer and record their overall duration of performing each individual task. It is important to note that, 15 minutes of analytical time were ignored from this computation as the response time for the VisExPreS interface. The second row in Table 6.1 shows the *mean* and *SD* of analysis time reported by the novice and expert users for all 20 tasks. Our results show that, overall the novice users took much longer than the expert users to finish the same analytical tasks. However, all users could complete the respective tasks within the given timeframe. A more detailed analysis of the efficiency of VisExPreS can be found in supplemental materials.

6.4.2.2 Measuring User Satisfaction from VisExPreS

The third metric for usability that is identified by the ISO 9241-11 standard is user satisfaction. In our study, following the footsteps of Georgsson et al. [50] the System Usability Scale (SUS) developed and designed by Brooke [51]. SUS contains 10 pre-defined standard utility questions with a provision to answer each question on a 5-point Likert scale. In our study, the questions of SUS were scored based on the guidelines presented in its original definition [51]. That is, the score contribution of each item was designated from 0 to 4, where for the questions 1, 3, 5, 7, and 9, the score was considered to be the selected scale position minus 1. For

Table 6.2: Evaluation of Coverage for RepSubset

Datasets	Data Statistics		Coverage (Cov)		
	#Rows	#Feat.	RepSubset	FURS	OptiSim
Breast Cancer	569	32	0.86	0.81	0.72
Wine Quality	4898	12	0.93	0.88	0.56
Magic	19020	11	0.83	0.84	0.67
Credit Card	30000	24	0.87	0.84	0.63
Animals	30475	85	0.91	0.89	0.76
MNIST	60000	784	0.64	0.51	0.37

questions 2, 4, 6, 8, and 10, the score allocated was five points minus the scale position. The sum of the scores of all 10 questions was then multiplied by 2.5 to get the overall satisfaction value ranging between 0 and 100. As per Georgsson et al. [50], a SUS score above 70 is considered as good whilst, a score of 85 or above is considered as an indicator of excellent usability. Table 6.1 shows the results of our analysis where we present the *mean* and *SD* of SUS scores for the novice and expert study participants. Our results show that although the novice users had some difficulty in using the VisExPreS interface with a SUS score just close to 70, the expert users found the interface to be extremely useful for analyzing embeddings.

As an additional step of the usability analysis, we also quantified the users' agreement on the fulfillment of the pre-defined design goals for the VisExPreS interface that were discussed in Section 6.2. For this purpose, following the guidelines of Lewis et al. [52], we computed the Fleiss' Kappa (κ) consistency measure for the five pre-defined design goals. The value of κ ranges from -1 to +1, where -1 represents no agreement, +1 signifies perfect agreement and 0 denotes agreement due to random chance. The results of our study are summarized in Table 6.1. The table shows that although the novice users had low positive agreement (i.e., 0.27) on the complete fulfillment of the design goals due to their prior experience with DR, the expert users had a moderate agreement of 0.48 regarding the same. Detailed analysis of user agreements on the design goals is presented in our supplemental materials.

6.4.3 Evaluation of Coverage for RepSubset

In this section, we evaluate the presented RepSubset algorithm and compare its 'Coverage' [20] with two state-of-the-art representative subset selection methods. As mentioned by Mall et al. [20], coverage is a simple evaluation metric for subset selection algorithms that measures a ratio between the total number of unique points that can be directly reached from the points in the subset and the total number of points in the dataset. In this research, to measure the coverage of the RepSubset algorithm, at first we set our required subset size to be 10% of the original dataset and counted the number of unique data-points that are in the neighborhood (for a k of size 15) of all the points in the subset. Then we computed the ratio of these unique data-points with the size of the entire dataset. We compared the coverage of RepSubset with FURS [20] and OptiSim [22], two well-known representative subset selection methods. The results of our comparisons along with the statistics for the 6 datasets that were used in our analysis are presented in Table 6.2. The table shows that RepSubset has consistently shown a higher coverage (ranging between 64% to 93%) than both FURS and OptiSim for all the 6 datasets.

6.5 Discussion

In order to provide further clarity on the usability of the VisExPreS toolkit; in this section, at first we analyze the scalability of the presented PG-LAPS and PG-GAPS methods. Next, we identify the limitations of the proposed toolkit and present some ideas for future work.

6.5.1 Scalability Analysis of PG-LAPS and PG-GAPS

In order to enable user-engagement, analytical speed is one of the primary requirements for any graphical user interface. When designing VisExPreS we kept this in mind. This section presents a detailed scalability analysis of the two techniques PG-LAPS and PG-GAPS that are at the core of the VisExPreS interface.

Table 6.3: Scalability Analysis of the PG-LAPS and PG-GAPS Processes

Datasets	Mean (SD) of execution time for 100 points using PG-LAPS	Mean (SD) of execution time for 100 executions using PG-GAPS				
		5 points	10 points	50 points	100 points	500 points
B-Cancer	38.01 (1.95)	66.80 (6.48)	71.40 (5.32)	129.30 (12.85)	219.60 (18.15)	217.00 (17.84)
W- Quality	12.83 (0.32)	23.00 (3.15)	24.10 (4.57)	42.40 (6.24)	81.30 (7.22)	77.60 (8.44)
Magic	11.42 (0.23)	22.90 (4.85)	24.60 (4.62)	34.10 (5.74)	38.40 (5.22)	49.20 (6.21)
Credit Card	32.62 (0.38)	44.20 (4.23)	50.90 (5.32)	167.70 (25.85)	294.20 (21.78)	335.50 (27.03)
Animals	65.54 (4.66)	31.80 (4.19)	34.30 (5.48)	51.70 (7.88)	128.50 (12.32)	211.40 (18.62)
MNIST	70.87 (9.89)	24.10 (4.23)	24.80 (4.99)	37.50 (6.41)	71.60 (9.47)	133.70 (17.44)

Our analysis results are presented in Table 6.3. In this table, for PG-LAPS we present the *mean* and *SD* of the required end-to-end computation time for 100 points from 6 different datasets. On the other hand, for PG-GAPS, we have also presented the mean and SD for 100 executions of global quality analysis using for sample sizes of 5, 10, 50, 100, and 500 of the same six datasets. Overall, our results show that understandably PG-GAPS has a much higher execution time than the PG-GAPS process for all the six datasets. However, for PG-GAPS the execution time does not increase significantly for large changes in the sample size (i.e., especially for when the sample sizes increase from 100 to 500). The reason is, as mentioned in [38], in the step 3 of the PG-GAPS process the size of the overall perturbed neighborhoods is kept fixed to 5000. Here, in case there are 100 samples in the subset, the PG-GAPS randomly generates 50 perturbed neighbors for each point. Whereas for 500 points in the sample, PG-GAPS randomly generates only 10 perturbed neighbors for each point. As a result, we can see that on a computer with 8 GB RAM and a processor with 4 cores the VisExPreS interface has shown response time of a minimum 49.2 seconds and a maximum of 5.5 minutes for 500 points in the selected subset. However, we think that this range of the response time for PG-GAPS also depends on the number of features in the dataset.

6.5.2 Limitations and Future Work

VisExPreS being primarily based on spatial visualizations (i.e., 2D or 3D plots), all visual scalability limitations associated with special techniques also become applicable to the toolkit. For example, with more than 50 features in the input datasets, the bi-directional feature influences bar graphs or the histogram amalgamated sunbursts for feature analysis can become challenging to comprehend. Although this limitation cannot be completely avoided, the current implementation of VisExPreS allows users to reduce the feature budget to only investigate the most influential features. Also, the feature analysis view enables users to delete uninfluential or redundant features. However, we are currently working on enhancing the visual effectiveness of the sunburst in the feature analysis view by performing a more advanced grouping among the features in the dataset and allowing users to remove or alter groups of features as a whole. Finally, although VisExPreS successfully enables an interactive user-driven assessment of embeddings, it does not provide much assistance with driving the DR algorithms to enhance the embedding quality. Hence, as our ongoing work we are investigating on optimizing the local and global divergence metrics in order to improve the quality of the obtained embeddings.

6.6 Conclusions

This chapter presents VisExPreS, a visual interactive toolkit that assists with a user-driven quality analysis of preserved local and global structures in the embeddings. At the core of VisExPreS, there are two novel techniques that are also introduced in this article. The first technique PG-LAPS generates interpretable explanations of the preserved locality of a single data-point in an embedding. PG-LAPS obtains the explanations regarding the structural preservation by approximating the local neighborhood around the single data point that is proactively recommended by the toolkit. On the other hand, the second technique PG-GAPS explains the preserved global structure in an embedding by unifying the local approximations for a set of non-redundant data-points interactively selected from the projection space into a global approximation. To provide proactive guidance for the non-redundant data points required in the analysis as a part of PG-GAPS, we present RepSubset. A novel algorithm that uses the notions of density and dissimilarity among data-points to generate a representative subset from the data. We demonstrate the utility and usability of the proposed VisExPreS toolkit, PG-LAPS and PG-GAPS techniques, along with the RepSubset algorithm using an exhaustive evaluation. A large amount of our experimental results is presented as supplemental material.

References

[1] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, “Dimensionality Reduction in the Wild: Gaps and Guidance,” Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, Jun. 2012.

[2] M. Cavallo and Ç. Demiralp, “A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration,” *arXiv:1811.12199 [cs]*, Nov. 2018, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1811.12199.

[3] L. G. Nonato and M. Aupetit, “Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment,” *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 8, pp. 2650–2673, Aug. 2019, doi: 10.1109/TVCG.2018.2846735.

[4] L. van der Maaten, E. O. Postma, and H. J. van den Herik, “Dimensionality Reduction : A Comparative Review,” *J Mach Learn Res 10*, vol. 66, no. 71, p. 13, 2008.

[5] J. Stahnke, M. Dork, B. Muller, and A. Thom, “Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions,” *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 629–638, Jan. 2016, doi: 10.1109/TVCG.2015.2467717.

[6] T. Schreck, T. von Landesberger, and S. Bremm, “Techniques for Precision-Based Visual Analysis of Projected Data,” *Information Visualization*, vol. 9, no. 3, pp. 181–193, Sep. 2010, doi: 10.1057/ivs.2010.2.

[7] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, “Visual analysis of dimensionality reduction quality for parameterized projections,” *Computers & Graphics*, vol. 41, pp. 26–42, Jun. 2014.

[8] F. Heimerl, C. Kralj, T. Möller, and M. Gleicher, “embComp: Visual Interactive Comparison of Vector Embeddings,” *arXiv:1911.01542 [cs]*, Nov.

2019, Accessed: Nov. 11, 2019.

[9] J. A. Lee and M. Verleysen, “Quality assessment of dimensionality reduction: Rank-based criteria,” *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, Mar. 2009, doi: 10.1016/j.neucom.2008.12.017.

[10] J. A. Lee and M. Verleysen, “Scale-independent quality criteria for dimensionality reduction,” *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2248–2257, Oct. 2010, doi: 10.1016/j.patrec.2010.04.013.

[11] A. Bibal and B. Frenay, “Measuring Quality And Interpretability Of Dimensionality Reduction Visualizations,” p. 7, 2019.

[12] D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, “Embedding Projector: Interactive Visualization and Interpretation of Embeddings,” *arXiv:1611.05469 [cs, stat]*, Nov. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1611.05469.

[13] C. Lai, Y. Zhao, and X. Yuan, “Exploring high-dimensional data through locally enhanced projections,” *Journal of Visual Languages & Computing*, vol. 48, pp. 144–156, Oct. 2018, doi: 10.1016/j.jvlc.2018.08.006.

[14] R. Cutura, S. Holzer, M. Aupetit, and M. Sedlmair, “VisCoDeR: A Tool for Visually Comparing Dimensionality Reduction Algorithms,” *Computational Intelligence*, p. 6, 2018.

[15] R. R. O. D. Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea, “Attribute-based Visual Explanation of Multidimensional Projections,” *EuroVis Workshop on Visual Analytics (EuroVA)*, p. 5 pages, 2015.

[16] L. Pagliosa, P. Pagliosa, and L. G. Nonato, “Understanding Attribute Variability in Multidimensional Projections,” in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Sao Paulo, Brazil, Oct. 2016, pp. 297–304, doi: 10.1109/SIBGRAPI.2016.048.

[17] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, “SIRIUS: Dual, Symmetric, Interactive Dimension Reductions,” *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, Jan. 2019.

[18] J. Z. Self, L. House, S. Leman, and C. North, “An-dromeda: Observation-

Level and Parametric Inter-action for Exploratory Data Analysis." *Technical re-port, Department of Computer Science, Virginia Tech*, Blacksburg, Virginia; 2015.

[19] M. Daszykowski, B. Walczak, and D. L. Massart, "Representative subset selection," *Analytica Chimica Acta*, vol. 468, no. 1, pp. 91–103, Sep. 2002.

[20] R. Mall, R. Langone, and J. A. K. Suykens, "FURS: Fast and Unique Representative Subset selection retaining large-scale community structure," *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 1075–1095, Dec. 2013.

[21] B. B. Chaudhuri, "How to choose a representative subset from a set data in multi-dimensional space," *Pattern Recognition Letters*, vol. 15, no. 9, pp. 893–899, 1994, doi: 10.1016/0167-8655(94)90151-1.

[22] R. D. Clark, "OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets," *J. Chem. Inf. Comput. Sci.*, vol. 37, no. 6, pp. 1181–1188, Nov. 1997, doi: 10.1021/ci970282v.

[23] R. W. Kennard and L. A. Stone, "Computer Aided Design of Experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, Feb. 1969.

[24] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, "Graph-based measures to assist user assessment of multidimensional projections," *Neurocomputing*, vol. 150, pp. 583–598, Feb. 2015.

[25] S. Lespinats and M. Aupetit, "CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings," *Computer Graphics Forum*, vol. 30, no. 1, pp. 113–125, Mar. 2011, doi: 10.1111/j.1467-8659.2010.01835.x.

[26] J. Xia *et al.*, "LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 236–245, Jan. 2018.

[27] K. Bunte, M. Biehl, and B. Hammer, "Dimensionality reduction mappings," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Paris, France, Apr. 2011, pp. 349–356.

[28] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Towards a Quantitative Survey of Dimension Reduction Techniques," *IEEE*

Trans. Visual. Comput. Graphics, pp. 1–1, 2019.

[29] B. Rieck and H. Leitte, "Agreement Analysis of Quality Measures for Dimensionality Reduction," in *Topological Methods in Data Analysis and Visualization IV*, H. Carr, C. Garth, and T. Weinkauf, Eds. Cham: Springer International Publishing, 2017, pp. 103–117.

[30] J. Johannemann and R. Tibshirani, "Spectral Overlap and a Comparison of Parameter-Free, Dimensionality Reduction Quality Metrics," *arXiv:1907.01974 [cs, stat]*, Jul. 2019, Accessed: Sep. 30, 2019.

[31] M. Aupetit, "Visualizing distortions and recovering topology in continuous projection techniques," *Neurocomputing*, vol. 70, no. 7–9, pp. 1304–1330, Mar. 2007, doi: 10.1016/j.neucom.2006.11.018.

[32] S. L. France and U. Akkucuk, "A Review, Framework and R toolkit for Exploring, Evaluating, and Comparing Visualizations," *arXiv:1902.08571 [cs, stat]*, Feb. 2019, Accessed: Nov. 17, 2019.

[33] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, "Local Affine Multidimensional Projection," *IEEE Trans. Visual. Comput. Graphics*, vol. 17, no. 12, pp. 2563–2571, Dec. 2011.

[34] W. Gani and M. Limam, "A kernel distance-based representative subset selection method," *Journal of Statistical Computation and Simulation*, vol. 86, no. 1, pp. 135–148, Jan. 2016, doi: 10.1080/00949655.2014.996758.

[35] Y. Tominaga, "Representative subset selection using genetic algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 43, no. 1–2, pp. 157–163, Sep. 1998, doi: 10.1016/S0169-7439(98)00085-9.

[36] D. Chaudhuri, C. A. Murthy, and B. B. Chaudhuri, "Finding a subset of representative points in a data set," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 9, pp. 1416–1424, Sep. 1994, doi: 10.1109/21.310520.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, Feb. 2016, pp. 1135–1144, Accessed: Jul. 12, 2019.

[38] A. Ghosh, M. Nashaat, J. Miller, and S. Quader, "Interpretation of Structural Preservation in Low-dimensional Embeddings," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2020, doi: 10.1109/TKDE.2020.3005878.

[39] E. Levina and P. J. Bickel, "Maximum Likelihood Estimation of Intrinsic Dimension," *Advances in neural information processing systems*, pp. 777–784, 2005.

[40] C. Angulo, X. Parra, and A. Català, "K-SVCR. A support vector machine for multi-class classification," *Neurocomputing*, vol. 55, no. 1–2, pp. 57–77, Sep. 2003, doi: 10.1016/S0925-2312(03)00435-1.

[41] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood Components Analysis," *Advances in neural information processing systems*, pp. 513–520, 2005.

[42] G. Plumb, D. Molitor, and A. Talwalkar, "Model Agnostic Supervised Local Explanations," *arXiv:1807.02910 [cs, stat]*, Jul. 2018, Accessed: Aug. 12, 2019. [Online]. Available: http://arxiv.org/abs/1807.02910.

[43] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local Rule-Based Explanations of Black Box Decision Systems," *arXiv:1805.10820 [cs]*, May 2018, Accessed: Aug. 12, 2019.

[44] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971, doi: 10.2307/2528823.

[45] H. Wang and M. Hong, "Distance Variance Score: An Efficient Feature Selection Method in Text Classification," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/695720.

[46] 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*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[47] J. P. Boyd, "Additive blending of local approximations into a globally-valid approximation with application to the dilogarithm," *Applied Mathematics Letters*, vol. 14, no. 4, pp. 477–481, 2001, doi: 10.1016/S0893-

9659(00)00180-4.

[48] R. Haftka, “Combining global and local approximations,” *AIAA journal*, vol. 29, no. 9, p. 1523, Sep. 1991.

[49] S. Mahmood and K. Mueller, “Taxonomizer: Interactive Construction of Fully Labeled Hierarchical Groupings from Attributes of Multivariate Data,” *IEEE Trans. Visual. Comput. Graphics*, vol. 26, no. 9, pp. 2875–2890, Sep. 2020, doi: 10.1109/TVCG.2019.2895642.

[50] M. Georgsson and N. Staggers, “Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics,” *J Am Med Inform Assoc*, vol. 23, no. 1, pp. 5–11, Jan. 2016, doi: 10.1093/jamia/ocv099.

[51] J. Brooke, “SUS - A quick and dirty usability scale,” Usability evaluation in industry. Jun 11 1996, pp. 189.

[52] J. M. Lewis and V. R. de Sa, “A Behavioral Investigation of Dimensionality Reduction,” *In Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, p. 7, 2012.

Chapter 7

Conclusions and Future Work

The primary objectives of this thesis are to examine the research gaps and improve the techniques that are commonly used during exploratory analysis of large-scale industrial data for making data-driven decisions. With this goal in mind, following an action research method, this thesis presents five novel solutions. To facilitate the understanding of the impact of exploratory data analysis on real-life business decision making, the first solution focuses on identifying the industrial process of software license renewals and the risks & challenges associated with it. Next, to distinguish the benefits and limitation of existing options for performing exploratory analysis on business datasets, the second solution investigates 50 cutting edge exploratory data analysis tools that are commonly used for analyzing large industrial data. Then the focus of this thesis is narrowed to Dimensionality Reduction (DR) techniques and several limitations of such techniques are addressed. For example, by performing a large-scale experimental study and statistical significance analysis, the third solution presented in this thesis examines 15 state-of-the-art dimensionality reduction techniques from the perspectives of 5 common analytical contexts for DR. The results of this investigation provide assistance with the selection of the most appropriate DR method for any dataset in the analytical contexts considered during the study. Next, the fourth solution in this thesis addresses the black-box nature of DR techniques and proposes two novel approaches for generating interpretable explanations regarding the quality of embeddings obtained from DR. Finally, the fifth solution presents a visual interactive toolkit that facilitates a proactively guided and user-driven evaluation of embeddings obtained by applying *any* DR techniques on *any* large high-dimensional dataset. In this section, at first, a detailed summary of the contributions in this thesis is presented followed by a discussion on opportunities for future work.

7.1 Overall Contributions

In par with the structural configuration of this thesis, the primary contributions of this thesis are categorized in the following five sub-sections:

7.1.1 Identification of Industrial Process of S&S

The software industry has changed significantly in the 21st century; no longer is it dominated by organizations seeking to sell products directly to customers, instead most multinational organizations nowadays provide services via licensing agreements. These licenses are for a fixed duration; and hence, the question of their renewal becomes of paramount importance for the selling organization's revenue. Despite its financial impact, the topic of license renewal strategies, processes, tools, and support receives very limited attention in the research literature. Hence, it is believed that an interesting research question is: What is the state of current industrial practice in this essential field? To initially explore the topic of license renewals, this research implements the Grounded Theory method [1]. To implement the method, semi-structured, cross-sectional, anonymous, self-reported interviews are carried out with 20 professionals from multiple organizations, later the Constant Comparative Method [2] is used to analyze the collected data. The participants of the study were carefully chosen from several multinational organizations headquartered in North America, with various roles in the renewal process such as, sales and subscription (S&S) representatives, brand leaders, global sales leaders, program directors, and data analysts. From this analysis, this research presents a synthesized picture of the current industrial practice of the end-to-end software license renewal process. Alongside, it also identifies a set of challenges and risk factors that impact on renewal decisions of customers, hence on the overall revenue of seller organizations. For example, the study shows that lack of effective communication among the stakeholders, scarcity of customer satisfaction, and absence of value generated from the purchased licenses, are among the primary drivers that influence the renewal decisions from customers. The results of the study were validated using the quantitative measure of inter-rater reliability [3], where multiple researchers analyzed the same data independently at the same time. Before presenting the analysis results to a team of content validity experts in the participating organizations. Finally, using structured brainstorming techniques, this research identifies 11 possible risk mitigation strategies, that can help organizations with the mitigation of the risks in the license renewal process. The proposed risk mitigation strategies can be classified into immediate action plans and future research directions. The immediate action plans include enhancing effective communication with customers and introducing new propagated ways of collecting customer information. On the other hand, for future research, the organizations can take advantage of applying intelligent automation either in the form of chat-bots or as predictive models. The analysis shows that an effective visualization of customers' journey with an organization can help renewal reps to analyze the overall experience and satisfaction of their customers.

7.1.2 Detailed Survey of Popular EDA tools

Exploratory data analysis plays a major role in obtaining insights from data. Over the last two decades, researchers have proposed several visual data exploration tools that can assist with each step of the analysis process. Nevertheless, in recent years, data analysis requirements have changed significantly. With constantly increasing size and types of data to be analyzed, scalability and analysis duration are now among the primary concerns of researchers. Moreover, in order to minimize the analysis cost, businesses require analysis tools that can be used with limited analytical knowledge. To address these challenges, traditional data exploration tools have evolved within the last few years. In this research, initially with a detailed analysis of an industrial tabular dataset of 3.4 million records, a set of additional exploratory requirements for large datasets are identified. Later, a systematic [4] and comprehensive survey of the recent advancements the emerging field of exploratory data analysis is presented. Here, 50 academic and non-academic visual data exploration tools are investigated with respect to their utility in the six fundamental steps of the exploratory data analysis process. The extent to which these modern data exploration tools fulfill the identified additional exploratory requirements of analyzing large datasets namely: scalability, interpretability, reduced analytical expertise, and user engagement, is also examined. The analysis shows, most modern EDA tools assist with the fundamental steps of the EDA process, whilst only some tools consider addressing the challenges of big-data analytics. Among the analyzed tools however, the trade-off between breadth of supported features and in-depth analysis of data is remaining. Even the most advanced tools in both academia and industry do not depict complex multivariate relationships among attributes. The reason behind this is, most tabular data analysis tools are primarily designed for a generic audience who might need more training to perform complex statistical analysis with the data. Moreover, some academic EDA tools that perform factor analysis or use complex diagrams to show relationships between multiple attributes, often suffer from interpretability and scalability issues. Incorporation of domain expertise is another challenge in most modern EDA tools. As in most cases for both commercial and academic tools, the user gets to take only the viewer's role in the data analysis process. Especially for the EDA tools that proactively generate visual recommendations; the absence of any feedback process can cause users to lose their confidence on the suggestions provided by the tools. From this analysis, a set of research opportunities are identified that include: (1) detailed analysis and visualization of bivariate & multivariate statistics, (2) advanced discretization of continuous variables, (3) proactive guidance for multivariate relationships, (4) addressing the scalability challenges in the data visualization tools.

7.1.3 Practitioners Guidelines for Selecting DR Algorithms

Dimensionality reduction is a commonly used technique in data analytics. Reducing the dimensionality of datasets not only helps with managing their analytical complexity but also with removing redundancy. Over the years, several such algorithms have been proposed with their aims ranging from generating simple linear projections to complex non-linear transformations of the input data. Subsequently, researchers have defined several quality metrics in order to evaluate the performances of different algorithms. Generally, these quality metrics evaluate the extent of preserved proximities among the data-points in a high-dimensional dataset after DR. Among these metrics, whereas some consider the actual proximities (distances) among the data-points, others compare the ranks of the distances between these points. Nevertheless, the quality analysis of the embedding obtained from DR has been an open research area among academics. The reasons being: (1) for nonlinear DR, no direct mapping exists between the original attributes in the high-dimensional dataset and the dimensions in the embedding. (2) Often the relationships in the high-dimensional data that should be preserved by DR are not clearly identifiable. Also, (3) for real-world datasets the intrinsic dimensionality and the topology of the original manifold is usually unknown. Hence, given a plethora of dimensionality reduction algorithms and metrics for their quality analysis, there is a long-existing need [5], [6] for guidelines on; how to select the most appropriate algorithm in every scenario. In order to bridge this gap, in this research, 12 state-of-the-art quality metrics are composed and categorized into five identified analytical contexts. Furthermore, 15 most popular dimensionality reduction algorithms are assessed on the chosen quality metrics using a large scale and systematic experimental study. The results identify t-SNE and UMAP to be the most robust algorithms in terms of metrics that evaluate the preservation of small neighborhoods in the original data. However, the results also indicate that the performance of t-SNE starts to deteriorate as the neighborhood size grows larger. It is also found that for datasets with unattributable missing values algorithms such as t-SNE, UMAP, LEM, LLE (i.e., DR techniques that attempt to preserve local structure of data) perform better than the globally focused algorithms. However, in case of datasets with outliers globally focused algorithms such as non-metric MDS, Kernel PCA, PCA perform better than the locally focused methods. Later, using a set of robust non-parametric statistical tests, the generalizability of the evaluation on 40 real-world datasets (39 open source and 1 from our industrial partner IBM) was assessed. The null hypothesis significance tests confirm that: the difference in the performances of the best, mediocre, and worst performing algorithms for the chosen 12 quality metrics are indeed statistically significant. Moreover, the analysis also indicates although not every algorithm performs equally well on every DR quality metrics, there is a perfectly reasonable metric for every algorithm where it performs better than its competitors. Finally, based on the results a practitioners' guideline is presented for the identification of intrinsic dimensionality in real-world datasets and the selection of an appropriate dimensionally reduction algorithm in the presented analytical contexts.

7.1.4 Two Novel Algorithms for Interpreting the Outcome of DR

Despite being commonly used in big-data analytics; the outcome of dimensionality reduction remains a black-box to most of its users. The quality of a low-dimensional embedding depends on the extent to which an algorithm can preserve the local structural relationships (i.e., the structural similarities in individual neighborhoods) as well as the global structural associations (i.e., the relative differences in overall neighborhoods) from the original dataset. Understanding the quality of a low-dimensional embedding is important as not only it enables trust in the transformed data, but it can also help to select the most appropriate dimensionality reduction algorithm in each scenario. As existing research primarily focuses on the visual exploration of embeddings, there is still a need for enhancing interpretability of such algorithms. To bridge this gap, two novel interactive explanation techniques are proposed for low-dimensional embeddings obtained from any dimensionality reduction algorithm. The first method & data-type agnostic technique *LAPS* - ***L****ocal* ***A****pproximation of* ***P****reserved* ***S****tructure* produces a local approximation of the neighborhood structure of any individual data-point in an embedding to generate interpretable explanations on the preserved locality for that single instance. The second method *GAPS* - ***G****lobal* ***A****pproximation of* ***P****rojection* ***S****pace* explains the retained global structure of a high-dimensional dataset in its embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. The explanations generated by LAPS and GAPS helps with associative reasoning [7] and answers a range of questions regarding *what* happened during the transformation of a dataset. Moreover, in association with the definition of *explanation* obtained from social sciences [7], the explanations generated by LAPS and GAPS are presented to be contrastive (i.e., why a certain event occurred instead of another event) and selective (users are allowed to adjust the amount of information that they would like to see). As a part of an extensive and comprehensive evaluation, both of the proposed techniques are assessed for their flexibility (with 10 different dimensionality reduction algorithms on 16 real-life datasets), applicability (i.e., with tabular, text, image, and audio data), utility (i.e., with a user-study that examines their ability to explain the quality [7] of a projection), and reliability (i.e., to assist with the selection of the most appropriate dimensionality reduction algorithm). The experiments also reveal the roles of different user-defined parameters in the outcome of the proposed techniques. Moreover, they uncover the ability of the techniques in discovering feature correlations in high-dimensional data.

7.1.5 A Visual Interactive Toolkit for Evaluation of DR

Embeddings are complex and black-box representations of high-dimensional datasets that are difficult to interpret and evaluate. The reasons being, firstly, the dimensions derived using such techniques lack a clear-to-interpret mapping with the original attributes in the data. As a result, novice data analysts are often forced to blindly trust the embeddings without truly understanding the meaning of the projection axes or the positioning of data-points. Secondly, in recent years, a plethora of DR techniques have been proposed with their own respective parameter combinations that significantly influence the embedding structure. The non-intuitive nature of these parameters also hinders the interpretability of these techniques making the selection of an appropriate DR algorithm for any dataset, difficult. Thirdly, in most cases, embeddings derived from DR do not make the existing errors and distortions prominent to the users. In some cases, where such distortions are visually exposed, the users are not really allowed to control or interact with them. All these limitations make an efficient evaluation of embeddings obtained from DR algorithms extremely challenging. In recent years, several quantitative and visual methods have been proposed for analyzing low-dimensional embeddings. On the one hand, the quantitative methods associate numeric identifiers to the qualitative characteristics of these embeddings; on the other hand, the visual techniques allow users to interactively explore these embeddings and make decisions. However, in the former case users do not have much control over the analysis, as the later leaves the assessment decisions entirely to the user's perception and expertise. The reason being, such techniques [8]–[10] often do not proactively guide users with the analysis process. For example, for very large datasets, most existing techniques [8]–[10] do not assist users with the selection of influential data points [11] or representative data subsets [12], [13] that provide a good representation of the original data and can reveal the overall quality of the embeddings better than other points. As a result, novice data analysts often fail to utilize the complete potential of such interactive visualization techniques when exploring low-dimensional embeddings. In order to bridge this gap, in this work, the benefits of both are unified. Here, a visual interactive toolkit ***VisExPreS (Vis**ual **Ex**planations of **Pre**served **S**tructure)* is presented that enables a user-driven assessment of low-dimensional embeddings. At the core of VisExPreS there are two novel techniques that are also presented in this work. The first technique ***PG-LAPS (P**roactively **G**uided **L**ocal **A**pproximation of **P**reserved **S**tructure)* enables the computation of the *local-divergence*, that examines the fidelity of the relative positioning of any individual data-point in an embedding by approximating a neighborhood locally around that point. For this purpose, PG-LAPS proactively guides users with the selection of representative data points from the input dataset for analysis. The second technique ***PG-GAPS (P**roactively **G**uided **G**lobal **A**pproximation of **P**rojection **S**pace)* computes the *global-divergence*, that explains the preserved global structure in a low dimensional embedding, by combining non-redundant local-approximations from a coarse discretization of the projection space. In order to enable the proactively guided representative subset selection in PG-GAPS, a novel algorithm ***RepSubset*** is presented that generates representative subsets from the original data based on the notions of density [14], [15] and dissimilarity [16] in the dataset. With the help of PG-LAPS and PG-GAPS, VisExPreS not only gives users more control over the quality assessment of dimensionality reduction algorithms but also allows them to focus on the aspects of the analysis that are the most *interesting* to them. Using a comprehensive evaluation, the utility of VisExPreS is demonstrated for interpreting, analyzing, and comparing derived embeddings from different dimensionality reduction algorithms. The evaluation with both novice and expert users shows that VisExPreS can effectively assist with the selection of an appropriate dimensionality reduction technique for a given dataset.

7.2 Opportunities for Future Work

In this thesis, several avenues have been explored that aims at facilitating a more comprehensive data-driven decision making and analysis. Each research has been presented in detail in Chapters 2 to 6 of this thesis. However, like any experimental work, all the solutions presented in this thesis can be further pursued and improved in different ways. In this section, a set of recommendations for future work is presented for each of the presented solutions in this thesis.

- Chapter 2 of this thesis presents a longitudinal study across multiple organizations and identifies the state-of-the art and research gaps in the business units of Sales and Subscriptions Renewal. Just like any other study, this research offers some opportunities for future work and threats to the validity of the experimental process. In order to avoid any respondent biases [17] in the semi-structured self-reported interviews used for collecting the research data, methodological triangulation [18] could be performed. Here, the survey data could be collected using multiple methods or instruments other than only semi-structured interviews.
- Moreover, researcher bias [17] in the study presented in Chapter 2 could be avoided with investigator triangulation [18]. Here it can be suggested that, investigators from both industry and academia could collaboratively explore the different qualitative and quantitative aspects associated with the challenges and risks of software license renewals.

- Finally, an empirical study [19] could be performed with a larger population of stakeholders from organizations with different firmographic [20] backgrounds, in order to validate the theories that emerged from the analysis in Chapter 2.
- The survey of Exploratory Data Analysis (EDA) [21] tools presented in Chapter 3 can be further improved by extending the study to include more academic and industrial tools in the analysis. Also, any researcher bias [17] in the study presented in Chapter 3 could be avoided with researcher triangulation [18].
- The experimental evaluation of Dimensionality Reduction (DR) methods presented in Chapter 4 identifies five most popular analytical contexts in which DR algorithms are commonly used. Nevertheless, as mentioned in Chapter 4 (cf. Section 4.1.1), the list of analytical contexts is not exhaustive, and could be further improved by adding more analytical contexts. This would also contribute in increasing the applicability of the practitioners' guideline presented in the Chapter 4.
- Moreover, DR being a popular technique there are several methods that exist for the purpose. Although 15 most popular DR algorithms are included in the study, the scope of the study can be further enhanced by adding other techniques (e.g., Autoencoder based methods [22], factor analysis techniques [23], fractal-based methods [24] etc.) in the experiments.
- Also, a recommender system can be generated using the experiments on the dimensionality reduction techniques where the characteristics of real-world datasets can be identified using meta-learning [25] strategies.
- In Chapter 5, there are several avenues of future work that can be explored. For instance, in any interactive technique, one of the most important aspects is scalability [26], [27]. Although, both the proposed algorithms have a computational complexity of $O(n^2)$, for the current design of LAPS and GAPS, the user-defined neighborhood size is restricted to be as large as 10 and the number of perturbed samples to be a maximum of 5000. However,

further experiments can be performed with different sizes of neighborhoods (i.e., > 10) in future. Moreover, to enhance the overall scalability of the algorithms, parallel processing can be implemented for LAPS and GAPS.

- Furthermore, since the LAPS and GAPS processes provide two numeric metrics namely local and global divergences in the embeddings, experiments can be performed to test whether if these metrics were used as objective functions (i.e., if local and global divergences of embeddings were minimized), could better quality projections be generated.
- For the VisExPreS toolkit presented in Chapter 6, the visual scalability limitations associated with spatial techniques [28] could be better addressed in future. For example, with more than 50 attributes in the input datasets, the bi-directional feature influences bar graphs or the histogram amalgamated sunbursts for attribute analysis can become challenging to comprehend. In future, different types of visual representations [29] could be added in the toolkit to better accommodate large number of attributes in the dataset under investigation.
- Finally, although VisExPreS toolkit presented in Chapter 6 successfully enables an interactive user-driven assessment of embeddings, it does not allow users to improve or alter [30] the projections in anyway. As future work, proactive guidance could be provided with respect to the enhancing the projection qualities by interactively modifying the relative positioning of the data-points on the embeddings.

References

[1] B. G. Glaser, A. L. Strauss, and E. Strutzel, “The discovery of grounded theory; strategies for qualitative research,” *Nursing research*, vol. 17, no. 4, p. 364, 1968.

[2] B. G. Glaser, “The Constant Comparative Method of Qualitative Analysis,” *Social Problems*, vol. 12, no. 4, pp. 436–445, 1965, doi: 10.2307/798843.

[3] M. Lombard, J. Snyder-Duch, and C. C. Bracken, “Content Analysis in Mass

Communication: Assessment and reporting of intercoder reliability," *Human communication research*, vol. 28, no. 4, pp.587-604, 2002.

[4] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[5] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, Mar. 2009, doi: 10.1016/j.neucom.2008.12.017.

[6] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, A. Sim, and A. Todd, "Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization," *J. Data and Information Quality*, vol. 11, no. 2, pp. 1–22, Mar. 2019, doi: 10.1145/3301294.

[7] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

[8] M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, "SIRIUS: Dual, Symmetric, Interactive Dimension Reductions," *IEEE Trans. Visual. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, Jan. 2019, doi: 10.1109/TVCG.2018.2865047.

[9] J. Z. Self, L. House, S. Leman, and C. North, "Andromeda: Observation-Level and Parametric Interaction for Exploratory Data Analysis." *Technical report, Department of Computer Science*, Virginia Tech, Blacksburg, Virginia, 2015.

[10] M. Cavallo and Ç. Demiralp, "A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration," *arXiv:1811.12199 [cs]*, Nov. 2018, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1811.12199.

[11] C. Lai, Y. Zhao, and X. Yuan, "Exploring high-dimensional data through locally enhanced projections," *Journal of Visual Languages & Computing*, vol. 48, pp. 144–156, Oct. 2018, doi: 10.1016/j.jvlc.2018.08.006.

[12] M. Daszykowski, B. Walczak, and D. L. Massart, “Representative subset selection,” *Analytica Chimica Acta*, vol. 468, no. 1, pp. 91–103, Sep. 2002.

[13] R. Mall, R. Langone, and J. A. K. Suykens, “FURS: Fast and Unique Representative Subset selection retaining large-scale community structure,” *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 1075–1095, Dec. 2013.

[14] B. B. Chaudhuri, “How to choose a representative subset from a set data in multi-dimensional space,” *Pattern Recognition Letters*, vol. 15, no. 9, pp. 893–899, 1994, doi: 10.1016/0167-8655(94)90151-1.

[15] D. Chaudhuri, C. A. Murthy, and B. B. Chaudhuri, “Finding a subset of representative points in a data set,” *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 9, pp. 1416–1424, Sep. 1994, doi: 10.1109/21.310520.

[16] R. D. Clark, “OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets ‡,” *J. Chem. Inf. Comput. Sci.*, vol. 37, no. 6, pp. 1181–1188, Nov. 1997, doi: 10.1021/ci970282v.

[17] E. G. Guba, “Naturalistic inquiry,” *Improving Human Performance Quarterly*, vol. 8, no. 4, pp. 268–76, 1979.

[18] N. Carter, D. Bryant-Lukosius, A. DiCenso, J. Blythe, and A. J. Neville, “The Use of Triangulation in Qualitative Research,” *Oncology Nursing Forum*, vol. 41, no. 5, pp. 545–547, Sep. 2014, doi: 10.1188/14.ONF.545-547.

[19] F. Shull, J. Singer, and D. I. K. Sjøberg, Eds., *Guide to advanced empirical software engineering*. London: Springer, 2008.

[20] A. Strauss and J. Corbin, “Grounded Theory Methodology,” *Handbook of qualitative research*, vol. 17, pp. 273–85, 1994.

[21] S. Tufféry, *Data Mining and Statistics for Decision Making: Tufféry/Data Mining and Statistics for Decision Making*. Chichester, UK: John Wiley & Sons, Ltd, 2011.

[22] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *arXiv:1312.6114 [cs, stat]*, May 2014, Accessed: Mar. 24, 2020.

[23] J. P. Cunningham and Z. Ghahramani, “Linear Dimensionality Reduction: Survey, Insights, and Generalizations,” *The Journal of Machine Learning*

Research, vol. 16, no. 1, pp. 2859–2900, 2015.

[24] F. Camastra and A. Staiano, “Intrinsic dimension estimation: Advances and open problems,” *Information Sciences*, vol. 328, pp. 26–41, Jan. 2016, doi: 10.1016/j.ins.2015.08.029.

[25] Y. Peng, P. A. Flach, C. Soares, and P. Brazdil, “Improved Dataset Characterisation for Meta-learning,” in *Discovery Science*, vol. 2534, S. Lange, K. Satoh, and C. H. Smith, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 141–152.

[26] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit, “Taggle: Combining Overview and Details in Tabular Data Visualizations,” *Information Visualization*, vol. 19, no. 2, pp. 114–136, Apr. 2020, doi: 10.1177/1473871619878085.

[27] M. A. Yalçin, N. Elmqvist, and B. B. Bederson, “AggreSet: Rich and Scalable Set Exploration using Visualizations of Element Aggregations,” *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 688–697, Jan. 2016, doi: 10.1109/TVCG.2015.2467051.

[28] J. Tang, J. Liu, M. Zhang, and Q. Mei, “Visualizing Large-scale and High-dimensional Data,” *Proceedings of the 25th International Conference on World Wide Web - WWW ’16*, pp. 287–297, 2016, doi: 10.1145/2872427.2883041.

[29] F. Heimerl, C. Kralj, T. Möller, and M. Gleicher, “embComp: Visual Interactive Comparison of Vector Embeddings,” *arXiv:1911.01542 [cs]*, Nov. 2019, Accessed: Nov. 11, 2019. [Online]. Available: http://arxiv.org/abs/1911.01542.

[30] J. Stahnke, M. Dork, B. Muller, and A. Thom, “Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions,” *IEEE Trans. Visual. Comput. Graphics*, vol. 22, no. 1, pp. 629–638, Jan. 2016, doi: 10.1109/TVCG.2015.2467717.

Bibliography

S. Liu, D. Maljovec, B. Wang, P. -t Bremer, and V. Pascucci, “Visualizing High-Dimensional Data: Advances in the Past Decade,” IEEE transactions on visualization and computer graphics, vol. 23, no. 1, pp. 21–30, 2016.

I. M. Johnstone and D. M. Titterington, “Statistical challenges of high-dimensional data,” Proc. R. Soc. A, vol. 367, no. 1906, pp. 4237–4253, Nov. 2009, doi: 10.1098/rsta.2009.0159.

L. van der Maaten, E. O. Postma, and H. J. van den Herik, “Dimensionality Reduction : A Comparative Review,” J Mach Learn Res 10, vol. 66, no. 71, p. 13, 2008.

R. Agarwal and C. E. Helfat, “Strategic renewal of organizations,” Organization Science, vol. 20, pp. 281–293, 2009.

A. Schmitt, S. Raisch, and H. W. Volberda, “Strategic Renewal: Past Research, Theoretical Tensions and Future Challenges: Strategic Renewal,” International Journal of Management Reviews, vol. 20, no. 1, pp. 81–98, Jan. 2018, doi: 10.1111/ijmr.12117.

M. A. Yalcin, N. Elmqvist, and B. B. Bederson, “Keshif: Rapid and Expressive Tabular Data Exploration for Novices,” IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 8, pp. 2339–2352, Aug. 2018.

T. Kraska, “Northstar: an interactive data science system,” Proc. VLDB Endow., vol. 11, no. 12, pp. 2150–2164, Aug. 2018.

S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit, “Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets,” IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 2023–2032, Dec. 2014, doi: 10.1109/TVCG.2014.2346260.

S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” arXiv:1705.07874 [cs, stat], Nov. 2017, Accessed: Mar. 24, 2020. [Online]. Available: http://arxiv.org/abs/1705.07874.

J. B. Tenenbaum, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” Science, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

Z. Cui, S. K. Badam, A. Yalçin, and N. Elmqvist, "DataSite: Proactive Visual Data Exploration with Computation of Insight-based Recommendations," arXiv:1802.08621 [cs], Sep. 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1802.08621.

K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit, "Taggle: Combining Overview and Details in Tabular Data Visualizations," Information Visualization, vol. 19, no. 2, pp. 114–136, Apr. 2020, doi: 10.1177/1473871619878085.

K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, J. Heer, "Voyager 2: Augmenting Visual Analysis with Partial View Specifications," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver Colorado USA, May 2017, pp. 2648–2659, doi: 10.1145/3025453.3025768.

M. Cavallo and Ç. Demiralp, "A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration," arXiv:1811.12199 [cs], Nov. 2018, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1811.12199.

R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea, "Visual analysis of dimensionality reduction quality for parameterized projections," Computers & Graphics, vol. 41, pp. 26–42, Jun. 2014.

J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," Neurocomputing, vol. 72, no. 7–9, pp. 1431–1443, Mar. 2009, doi: 10.1016/j.neucom.2008.12.017.

S. Tufféry, Data Mining and Statistics for Decision Making: Tufféry/Data Mining and Statistics for Decision Making. Chichester, UK: John Wiley & Sons, Ltd, 2011.

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, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

J. M. Lewis and V. R. de Sa, "A Behavioral Investigation of Dimensionality Reduction," In Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 34, no. 34, p. 7, 2012.

M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality Reduction in the Wild: Gaps and Guidance," Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03, Jun. 2012.

J. Demśar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.

M. Dowling, J. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, “SIRIUS: Dual, Symmetric, Interactive Dimension Reductions,” IEEE Trans. Visual. Comput. Graphics, vol. 25, no. 1, pp. 172–182, Jan. 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” arXiv:1602.04938 [cs, stat], Feb. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1602.04938.

G. Plumb, D. Molitor, and A. Talwalkar, “Model Agnostic Supervised Local Explanations,” arXiv:1807.02910 [cs, stat], Jul. 2018, Accessed: Aug. 12, 2019. [Online]. Available: http://arxiv.org/abs/1807.02910.

J. P. Boyd, “Additive blending of local approximations into a globally-valid approximation with application to the dilogarithm,” Applied Mathematics Letters, vol. 14, no. 4, pp. 477–481, 2001.

D. Sacha, L. Zhang, M. Sedlmair, J.A. Lee, J. Peltonen, D. Weiskopf, S.C. North, D.A. Keim, “Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis,” IEEE Trans. Visual. Comput. Graphics, vol. 23, no. 1, pp. 241–250, Jan. 2017, doi: 10.1109/TVCG.2016.2598495.

F. Marotta-Wurgler, “What’s in a Standard Form Contract? An Empirical Analysis of Software License Agreements,” Journal of Empirical Legal Studies, vol. 4, no. 4, pp. 677–713, 2007, doi: 10.1111/j.1740-1461.2007.00104.x.

F. V. Wangenheim, N. V. Wünderlich, and J. H. Schumann, “Renew or cancel? Drivers of customer renewal decisions for IT-based service contracts,” Journal of Business Research, vol. 79, pp. 181–188, Oct. 2017, doi: 10.1016/j.jbusres.2017.06.008.

“CRM Strategies and Technologies to Understand, Grow and Manage Customer Experiences,” Gartner, Los Angeles, CA, Gartner Customer 360 Summit, 2011. Accessed: May 25, 2020. [Online]. Available: https://www.gartner.com/imagesrv/summits/docs/na/customer-360/C360_2011_brochure_FINAL.pdf.

T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

D. Einhorn, “The Enforceability of Tear-Me-Open Software License Agreements,” Journal of the Patent and Trademark Office Society, vol. 67, p. 509, 1985.

K. Jeffery, “Clouds for Science Tackle Challenges Facing Industry and Society,” IEEE Cloud Computing, vol. 5, no. 2, pp. 4–6, Mar. 2018, doi: 10.1109/MCC.2018.022171660.

M. Haenlein, “How to date your clients in the 21 st century: Challenges in managing customer relationships in today’s world,” Business Horizons, vol. 60, no. 5, pp. 577–586, Sep. 2017, doi: 10.1016/j.bushor.2017.06.002.

F. Wiersema, “The B2B agenda: The current state of B2B marketing and a look ahead,” vol. 4, no. 43, pp. 470–488, 2013.

N. Lu, H. Lin, J. Lu, and G. Zhang, “A Customer Churn Prediction Model in Telecom Industry Using Boosting,” IEEE Trans. Ind. Inf., vol. 10, no. 2, pp. 1659–1665, May 2014, doi: 10.1109/TII.2012.2224355.

A. Aluri, B. S. Price, and N. H. McIntyre, “Using Machine Learning To Cocreate Value Through Dynamic Customer Engagement In A Brand Loyalty Program,” Journal of Hospitality & Tourism Research, vol. 43, no. 1, pp. 78–100, Jan. 2019, doi: 10.1177/1096348017753521.

E. Ascarza, S.A. Neslin, O. Netzer, Z. Anderson, P.S. Fader, S. Gupta, B.G. Hardie, A. Lemmens, B. Libai, D. Neal, and F. Provost, “In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions,” Cust. Need. and Solut., vol. 5, no. 1–2, pp. 65–81, Mar. 2018, doi: 10.1007/s40547-017-0080-0.

K. Coussement and D. Van den Poel, “Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,” Expert Systems with Applications, vol. 34, no. 1, pp. 313–327, Jan. 2008, doi: 10.1016/j.eswa.2006.09.038.

M. Hassouna, A. Tarhini, T. Elyas, and M. S. Abou Trab, “Customer Churn in Mobile Markets: A Comparison of Techniques,” IBR, vol. 8, no. 6, p. p224, May 2015, doi: 10.5539/ibr.v8n6p224.

Yizhe Ge, Shan He, Jingyue Xiong, and D. E. Brown, “Customer churn analysis for a software-as-a-service company,” in 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, Apr. 2017, pp. 106–111, doi: 10.1109/SIEDS.2017.7937698.

W. Bi, M. Cai, M. Liu, and G. Li, “A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn,” IEEE Trans. Ind. Inf., vol. 12, no. 3, pp. 1270–1281, Jun. 2016, doi: 10.1109/TII.2016.2547584.

S. M. Kolb, “Grounded Theory and the Constant Comparative Method: Valid Research Strategies for Educators,” Journal of emerging trends in educational research and policy studies, vol. 3, no. 1, pp. 83-89, Feb 2012.

N. Guhr, B. Lebek, and M. H. Breitner, “The impact of leadership on employees’ intended information security behaviour: An examination of the full-range leadership theory,” Info Systems J, vol. 29, no. 2, pp. 340–362, Mar. 2019, doi: 10.1111/isj.12202.

P. B. Lowry, J. Zhang, C. Wang, and M. Siponen, “Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model,” Information Systems Research, vol. 27, no. 4, pp. 962–986, Dec. 2016, doi: 10.1287/isre.2016.0671.

B. G. Glaser, “The Constant Comparative Method of Qualitative Analysis,” Social Problems, vol. 12, no. 4, pp. 436–445, 1965, doi: 10.2307/798843.

H. Boeije, “A Purposeful Approach to the Constant Comparative Method in the Analysis of Qualitative Interviews,” Quality and quantity, vol. 36, no. 4, pp.391-409, 2002

J. G. Rawlinson, Creative thinking and brainstorming. Routledge, 2017.

F. Marotta-Wurgler, “Competition and the Quality of Standard Form Contracts: The Case of Software License Agreements,” Journal of Empirical Legal Studies, vol. 5, no. 3, pp. 447–475, 2008, doi: 10.1111/j.1740-1461.2008.00130.x.

T. Tuunanen, J. Koskinen, and T. Kärkkäinen, “Automated software license analysis,” Autom Softw Eng, vol. 16, no. 3–4, pp. 455–490, Dec. 2009, doi: 10.1007/s10515-009-0054-z.

F. M. Kifetew, M. Morandini, D. Munante, A. Perini, A. Siena, and A. Susi, “Goal-aware Analysis of Software License Risks,” p. 6.

M. Corporation, “Annual Report,” Microsoft Corporation, 2017. [Online]. Available: https://www.microsoft.com/investor/reports/ar17/index.html.

I. Corporation, "Annual Report," IBM Corporation, 2017. [Online]. Available: https://www.ibm.com/annualreport/index.html.

E. Suter, J. Arndt, N. Arthur, J. Parboosingh, E. Taylor, and S. Deutschlander, "Role understanding and effective communication as core competencies for collaborative practice," Journal of Interprofessional Care, vol. 23, no. 1, pp. 41–51, Jan. 2009, doi: 10.1080/13561820802338579.

P. S. H. Leeflang, P. C. Verhoef, P. Dahlström, and T. Freundt, "Challenges and solutions for marketing in a digital era," European Management Journal, vol. 32, no. 1, pp. 1–12, Feb. 2014, doi: 10.1016/j.emj.2013.12.001.

M. Leonard, S. Graham, and D. Bonacum, "The human factor: the critical importance of effective teamwork and communication in providing safe care," Qual Saf Health Care, vol. 13, no. Suppl 1, pp. i85–i90, Oct. 2004, doi: 10.1136/qshc.2004.010033.

M. A. Campion, G. J. Medsker, and A. C. Higgs, "Relations Between Work Group Characteristics and Effectiveness: Implications for Designing Effective Work Groups," Personnel Psychology, vol. 46, no. 4, pp. 823–847, 1993, doi: 10.1111/j.1744-6570.1993.tb01571.x.

N. S. Hill and K. M. Bartol, "Empowering Leadership and Effective Collaboration in Geographically Dispersed Teams," Personnel Psychology, vol. 69, no. 1, pp. 159–198, 2016, doi: 10.1111/peps.12108.

S. Hannan, B. Suharjo, K. Kirbrandoko, and R. Nurmalina, "International Journal of Economic Perspectives, 2017, Volume 11, Issue 1, 344-353.," vol. 11, no. 1, p. 10, 2017.

G. Walsh, M. Schaarschmidt, and S. Ivens, "Effects of customer-based corporate reputation on perceived risk and relational outcomes: empirical evidence from gender moderation in fashion retailing," Journal of Product & Brand Management, vol. 26, no. 3, pp. 227–238, Jan. 2017, doi: 10.1108/JPBM-07-2016-1267.

P. Guenzi, L. L. M. De, and R. Spiro, "The combined effect of customer perceptions about a salesperson's adaptive selling and selling orientation on customer trust in the salesperson: a contingency perspective," Journal of Business & Industrial Marketing, vol. 31, no. 4, pp. 553–564, Jan. 2016, doi: 10.1108/JBIM-02-2015-0037.

L. Marakanon and V. Panjakajornsak, "Perceived quality, perceived risk and customer trust affecting customer loyalty of environmentally friendly electronics products | Elsevier Enhanced Reader," Kasetsart Journal of Social Sciences, vol. 38, pp. 24–30, 2017, doi: https://doi.org/10.1016/j.kjss.2016.08.012.

U. S. C. Bureau, "North American Industry Classification System 2017," U.S. Department of Commerce, 2017. https://www.census.gov/cgi-bin/sssd/naics/naicsrch?code=511210&search=2017%20NAICS%20Search (accessed May 25, 2020).

B. G. Glaser, A. L. Strauss, and E. Strutzel, "The discovery of grounded theory; strategies for qualitative research," Nursing research, vol. 17, no. 4, p. 364, 1968.

J. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," Qualitative sociology, vol. 13, no. 1, pp. 3-21, 1990.

A. Sbaraini, S. M. Carter, R. W. Evans, and A. Blinkhorn, "How to do a grounded theory study: a worked example of a study of dental practices," BMC Med Res Methodol, vol. 11, no. 1, p. 128, Dec. 2011, doi: 10.1186/1471-2288-11-128.

J. D. Olson, C. McAllister, L. D. Grinnell, K. G. Walters, and F. Appunn, "Applying Constant Comparative Method with Multiple Investigators and Inter-Coder Reliability," Qualitative Report, vol. 21, no. 1, Jan 2016.

S. M. Fram, "The Constant Comparative Analysis Method Outside of Grounded Theory," Qualitative Report, vol. 18, no. 1, 2013.

J. Hewitt-Taylor, "Use of constant comparative analysis in qualitative research," Nursing Standard, vol. 15, no. 42, pp. 39–42, Jul. 2001, doi: 10.7748/ns2001.07.15.42.39.c3052.

T. Mettler and R. Winter, "Are business users social? A design experiment exploring information sharing in enterprise social systems," Journal of Information Technology, vol. 31, no. 2, pp. 101–114, Jun. 2016, doi: 10.1057/jit.2015.28.

M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content Analysis in Mass Communication: Assessment and reporting of intercoder reliability," Human communication research, vol. 28, no. 4, pp.587-604, 2002.

B. Bodenmann and K.W. Axhausen, "Synthesis report on the state of the art on firmographics," Institute for Transport Planning and Systems, ETH, Zurich, p. 31, 2010.

A. Tamaddoni Jahromi, S. Stakhovych, and M. Ewing, “Managing B2B customer churn, retention and profitability,” Industrial Marketing Management, vol. 43, no. 7, pp. 1258–1268, Oct. 2014, doi: 10.1016/j.indmarman.2014.06.016.

G. D. Moody and M. Siponen, “Using the theory of interpersonal behavior to explain non-work-related personal use of the Internet at work,” Information & Management, vol. 50, no. 6, pp. 322–335, Sep. 2013, doi: 10.1016/j.im.2013.04.005.

A. Vance, M. Siponen, and S. Pahnila, “Motivating IS security compliance: Insights from Habit and Protection Motivation Theory,” Information & Management, vol. 49, no. 3–4, pp. 190–198, May 2012, doi: 10.1016/j.im.2012.04.002.

V. Venkatesh, J. Y. L. Thong, and X. Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” MIS Quarterly, vol. 36, no. 1, pp. 157–178, 2012, doi: 10.2307/41410412.

D. L. Segal, F. L. Coolidge, A. O’Riley, and B. A. Heinz, “Structured and semistructured interviews,” Clinician’s handbook of adult behavioral assessment, pp. 121–144, 2006.

A. Rindfleisch, A. J. Malter, S. Ganesan, and C. Moorman, “Cross-Sectional versus Longitudinal Survey Research: Concepts, Findings, and Guidelines,” Journal of Marketing Research, vol. 45, no. 3, pp. 261–279, Jun. 2008, doi: 10.1509/jmkr.45.3.261.

B. Shahzad, A. M. Abdullatif, N. Ikram, and A. Mashkoor, “Build Software or Buy: A Study on Developing Large Scale Software,” IEEE Access, vol. 5, pp. 24262–24274, 2017, doi: 10.1109/ACCESS.2017.2762729.

M. Westergaard and F. M. Maggi, “Looking into the future,” In OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Springer, Berlin, Heidelberg, pp. 250-267, Sep 2012.

C. Urquhart and W. Fernández, “Using Grounded Theory Method in Information Systems: The researcher as blank slate and other myths,” In Enacting Research Methods in Information Systems, vol. 1, Palgrave Macmillan, Cham, pp. 129-156, 2016.

H. E. Tinsley and D. J. Weiss, “Interrater reliability and agreement of subjective judgments.,” Journal of Counseling Psychology, vol. 22, no. 4, pp. 358–376, 1975, doi: 10.1037/h0076640.

K. S. Kurasaki, “Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data,” Field Methods, vol. 12, no. 3, pp. 179–194, Aug. 2000, doi: 10.1177/1525822X0001200301.

K. A. Neuendorf and A. Kumar, “Content analysis,” The International Encyclopedia of Political Communication, vol. 1, pp. 221–230, 2002.

J. L. Campbell, C. Quincy, J. Osserman, and O. K. Pedersen, “Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement,” Sociological Methods & Research, vol. 42, no. 3, pp. 294–320, Aug. 2013, doi: 10.1177/0049124113500475.

J. Cohen, “A Coefficient of Agreement for Nominal Scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” Psychological bulletin, vol. 70, no. 4, p. 213, 1968, doi: http://dx.doi.org/10.1037/h0026256.

K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” Tutor Quant Methods Psychol, vol. 8, no. 1, pp. 23–34, 2012.

M. R. Shirey, “Brainstorming for Breakthrough Thinking:,” JONA: The Journal of Nursing Administration, vol. 41, no. 12, pp. 497–500, Dec. 2011, doi: 10.1097/NNA.0b013e3182378a53.

Y. Shi, “Brain storm optimization algorithm,” presented at the In International conference in swarm intelligence, Berlin, Heidelberg, 2011, pp. 303–309.

F. L. Eichorn, “Internal Customer Relationship Management (IntCRM) A Framework for Achieving Customer Relationship Management from the Inside Out,” Problems and Perspectives in Management, p. 24.

N. Umashankar, M. K. Ward, and D. W. Dahl, “The Benefit of Becoming Friends: Complaining after Service Failures Leads Customers with Strong Ties to Increase Loyalty,” Journal of Marketing, vol. 81, no. 6, pp. 79–98, Nov. 2017, doi: 10.1509/jm.16.0125.

A. Sharma and J. Bhatnagar, “Enterprise social media at work: web-based solutions for employee engagement,” Human Resource Management International Digest, vol. 24, no. 7, pp. 16–19, Jan. 2016, doi: 10.1108/HRMID-04-2016-0055.

I. Hawkins, “The future is happening: a global snapshot of IA,” PEX Network, May 2018.

K. A. Olsen and A. Malizia, “Automated Personal Assistants,” Computer, vol. 44, no. 11, pp. 112–111, Nov. 2011, doi: 10.1109/MC.2011.329.

T. Mettler, “Contextualizing a professional social network for health care: Experiences from an action design research study,” Info Systems J, vol. 28, no. 4, pp. 684–707, Jul. 2018, doi: 10.1111/isj.12154.

E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, “RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism,” in Advances in Neural Information Processing Systems 29, 2016, pp. 3504–3512, Accessed: May 25, 2020. [Online]. Available: http://papers.nips.cc/paper/6321-retain-an-interpretable-predictive -model-for-healthcare-using-reverse-time-attention-mechanism.pdf.

E. H. Chi, “A taxonomy of visualization techniques using the data state reference model,” in IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, Oct. 2000, pp. 69–75, doi: 10.1109/INFVIS.2000.885092.

K. N. Lemon and P. C. Verhoef, “Understanding Customer Experience Throughout the Customer Journey,” Journal of Marketing, vol. 80, no. 6, pp. 69–96, Nov. 2016, doi: 10.1509/jm.15.0420.

G. Andrienko, N. Andrienko, and S. Wrobel, “Visual analytics tools for analysis of movement data,” SIGKDD Explor. Newsl., vol. 9, no. 2, pp. 38–46, Dec. 2007, doi: 10.1145/1345448.1345455.

E. G. Guba, “Naturalistic inquiry,” Improving Human Performance Quarterly, vol. 8, no. 4, pp. 268–76, 1979.

N. Carter, D. Bryant-Lukosius, A. DiCenso, J. Blythe, and A. J. Neville, “The Use of Triangulation in Qualitative Research,” Oncology Nursing Forum, vol. 41, no. 5, pp. 545–547, Sep. 2014, doi: 10.1188/14.ONF.545-547.

F. Shull, J. Singer, and D. I. K. Sjøberg, Eds., Guide to advanced empirical software engineering. London: Springer, 2008.

P. Godfrey, J. Gryz, and P. Lasek, “Interactive Visualization of Large Data Sets,” IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 2142–2157, Aug. 2016, doi: 10.1109/TKDE.2016.2557324.

M. El-Hindi, Z. Zhao, C. Binnig, and T. Kraska, “VisTrees: fast indexes for interactive data exploration,” in Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA ’16, San Francisco, California, 2016, pp. 1–6, doi: 10.1145/2939502.2939507.

Jian Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, “Interactive Exploration of Implicit and Explicit Relations in Faceted Datasets,” IEEE Trans. Visual. Comput. Graphics, vol. 19, no. 12, pp. 2080–2089, Dec. 2013.

B. Yu and C. T. Silva, “VisFlow - Web-based Visualization Framework for Tabular Data with a Subset Flow Model,” IEEE Trans. Visual. Comput. Graphics, vol. 23, no. 1, pp. 251–260, Jan. 2017.

S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, B. Krabichler, M.R. Speicher, J. Zschocke, and Z. Trajanoski, “A survey of tools for variant analysis of next-generation genome sequencing data,” Briefings in Bioinformatics, vol. 15, no. 2, pp. 256–278, Mar. 2014.

M. Rautenhaus, M. Böttinger, S. Siemen, R. Hoffman, R.M. Kirby, M. Mirzargar, N. Röber, R. Westermann, “Visualization in Meteorology - A Survey of Techniques and Tools for Data Analysis Tasks,” IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 12, pp. 3268–3296, Dec. 2018.

A. Endert, W. Ribarsky, C. Turkay, B.W. Wong, I. Nabney, I.D. Blanco, and F. Rossi, “The State of the Art in Integrating Machine Learning into Visual Analytics,” Computer Graphics Forum, vol. 36, no. 8, pp. 458–486, Dec. 2017, doi: 10.1111/cgf.13092.

S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, “Tools for Educational Data Mining: A Review,” Journal of Educational and Behavioral Statistics, vol. 42, no. 1, pp. 85–106, Feb. 2017.

S. Liu, X. Wang, M. Liu, and J. Zhu, “Towards better analysis of machine learning models: A visual analytics perspective,” Visual Informatics, vol. 1, no. 1, pp. 48–56, Mar. 2017, doi: 10.1016/j.visinf.2017.01.006.

S. Idreos, O. Papaemmanouil, and S. Chaudhuri, “Overview of Data Exploration Techniques,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD ’15, Melbourne, Victoria, Australia, 2015, pp. 277–281.

M. Khan and S. S. Khan, “Data and Information Visualization Methods, and Interactive Mechanisms: A Survey,” International Journal of Computer Applications, vol. 34, p. 15.

J. Heer and B. Shneiderman, “Interactive Dynamics for Visual Analysis,” Queue, vol. 10, no. 2, pp. 30-55, 2012.

R. Amar, J. Eagan, and J. Stasko, “Low-Level Components of Analytic Activity in Information Visualization,” In IEEE Symposium on Information Visualization, INFOVIS 2005, pp. 111-117. 2005.

W. Chan, “A Survey on Multivariate Data Visualization,” Department of Computer Science and Engineering. Hong Kong University of Science and Technology, vol. 8, no. 6, pp.1-29, 2006.

R. High, “The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works.” RedBooks, 2012, Accessed: May 24, 2020. [Online].

Z. Liu, B. Jiang, and J. Heer, “imMens: Real-time Visual Querying of Big Data,” Computer Graphics Forum, vol. 32, no. 3pt4, pp. 421–430, Jun. 2013.

L. Battle, R. Chang, and M. Stonebraker, “Dynamic Prefetching of Data Tiles for Interactive Visualization,” in Proceedings of the 2016 International Conference on Management of Data - SIGMOD ’16, San Francisco, California, USA, 2016, pp. 1363–1375, doi: 10.1145/2882903.2882919.

Z. Wang, N. Ferreira, Y. Wei, A. S. Bhaskar, and C. Scheidegger, “Gaussian Cubes: Real-Time Modeling for Visual Exploration of Large Multidimensional Datasets,” IEEE Trans. Visual. Comput. Graphics, vol. 23, no. 1, pp. 681–690, Jan. 2017, doi: 10.1109/TVCG.2016.2598694.

A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko, “Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication,” IEEE Trans. Visual. Comput. Graphics, vol. 25, no. 1, pp. 672–681, Jan. 2019, doi: 10.1109/TVCG.2018.2865145.

E. Zgraggen, R. Zeleznik, and S. M. Drucker, “PanoramicData: Data Analysis through Pen & Touch,” IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 2112–2121, Dec. 2014, doi: 10.1109/TVCG.2014.2346293.

Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati, “Foresight: Rapid Data Exploration Through Guideposts,” arXiv:1709.10513 [cs], Sep. 2017, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1709.10513.

Bongshin Lee, R. H. Kazi, and G. Smith, “SketchStory: Telling More Engaging Stories with Data through Freeform Sketching,” IEEE Trans. Visual. Comput. Graphics, vol. 19, no. 12, pp. 2416–2425, Dec. 2013.

W. Javed and N. Elmqvist, “ExPlates: Spatializing Interactive Analysis to Scaffold Visual Exploration,” Computer Graphics Forum, vol. 32, no. 3pt4, pp. 441–450, Jun. 2013, doi: 10.1111/cgf.12131.

A. Satyanarayan and J. Heer, “Authoring Narrative Visualizations with Ellipsis: Authoring Narrative Visualizations with Ellipsis,” Computer Graphics Forum, vol. 33, no. 3, pp. 361–370, Jun. 2014.

H. Mei, W. Chen, Y. Ma, H. Guan, and W. Hu, “VisComposer: A Visual Programmable Composition Environment for Information Visualization,” Visual Informatics, vol. 2, no. 1, pp. 71–81, Mar. 2018, doi: 10.1016/j.visinf.2018.04.008.

N. Bikakis and T. Sellis, “Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art,” arXiv:1601.08059 [cs], Jan. 2016, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1601.08059.

W. Dunn, A. Burgun, M.-O. Krebs, and B. Rance, “Exploring and visualizing multidimensional data in translational research platforms,” Brief Bioinform, p. bbw080, Sep. 2016, doi: 10.1093/bib/bbw080.

L. Wang, G. Wang, and C. A. Alexander, “Big Data and Visualization: Methods, Challenges and Technology Progress,” Digital Technologies, vol. 1, no. 1, pp.33-38, 2015.

M. Behrisch, Streeb, F. Stoffel, D. Seebacher, B. Matejek, S.H. Weber, S. Mittelstaedt, H. Pfister, and D. Keim, “Commercial Visual Analytics Systems–Advances in the Big Data Analytics Field,” IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 10, pp. 3011–3031, Oct. 2019, doi: 10.1109/TVCG.2018.2859973.

M. Diamond and A. Mattia, “Data visualization: an exploratory study into the software tools used by businesses,” Journal of Instructional Pedagogies, vol. 18, p. 7, 2018.

S. M. Biju and A. Mathew, “Comparative analysis of selected big data analytics tools,” vol. 26, no. 2, p. 23, 2017.

M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," Journal of Big Data, vol. 2, no. 1, p. 1, Dec. 2015.

S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," in 2013 46th Hawaii International Conference on System Sciences, Wailea, HI, USA, Jan. 2013, pp. 995–1004, doi: 10.1109/HICSS.2013.645.

J.-F. Im, F. G. Villegas, and M. J. McGuffin, "VisReduce: Fast and responsive incremental information visualization of large datasets," in 2013 IEEE International Conference on Big Data, Silicon Valley, CA, Oct. 2013, pp. 25–32, doi: 10.1109/BigData.2013.6691710.

G. R. Iyer, S. Duttaduwarah, and A. Sharma, "DataScope: Interactive visual exploratory dashboards for large multidimensional data," PeerJ Preprints, preprint, Jan. 2018. doi: 10.7287/peerj.preprints.26441v1.

B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," Information and Software Technology, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

C. Niederer, H. Stitz, R. Hourieh, F. Grassinger, W. Aigner, and M. Streit, "TACO: Visualizing Changes in Tables Over Time," IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 1, pp. 677–686, Jan. 2018.

M. Vartak, S. Rahman, S. Madden, and A. Parameswaran, "SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics," In Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, vol. 8, no. 13, p. 2182. NIH Public Access, 2015.

Tableau, "Tableau." https://www.tableau.com/products. (accessed Oct. 01, 2018).

Microsoft Corp., "Microsoft Power BI." https://powerbi.microsoft.com/ (accessed Oct. 01, 2018).

Domo, "Domo." https://www.domo.com/ (accessed Oct. 01, 2018).

QlikView, "QlikView." https://www.qlik.com/us/ (accessed Oct. 01, 2018).

Sisense, "Sisense." https://www.sisense.com/product/ (accessed Oct. 01, 2018).

R. L. Sallam, C. Howson, C. J. Idoine, T. W. Oestreich, J. Laurence, and J. Tapadinhas, “Magic Quadrant for Business Intelligence and Analytics Platforms,” p. 87.

J. Fan and R. Li, “Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery,” arXiv:math/0602133, Feb. 2006, Accessed: May 25, 2020. [Online]. Available: http://arxiv.org/abs/math/0602133.

J. C. Roberts, “State of the Art: Coordinated & Multiple Views in Exploratory Visualization,” in Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007), Zurich, Switzerland, Jul. 2007, pp. 61–71, doi: 10.1109/CMV.2007.20.

M. Smith, Statistical Analysis Handbook. Edinburgh,: The Winchelsea Press, Drumlin Security Ltd, 2018.

M. A. Yalçin, N. Elmqvist, and B. B. Bederson, “AggreSet: Rich and Scalable Set Exploration using Visualizations of Element Aggregations,” IEEE Trans. Visual. Comput. Graphics, vol. 22, no. 1, pp. 688–697, Jan. 2016, doi: 10.1109/TVCG.2015.2467051.

Jing Xia, Wei Chen, Yumeng Hou, Wanqi Hu, Xinxin Huang, and D. S. Ebertk, “DimScanner: A relation-based visual exploration approach towards data dimension inspection,” in 2016 IEEE Conference on Visual Analytics Science and Technology (VAST), Baltimore, MD, USA, Oct. 2016, pp. 81–90, doi: 10.1109/VAST.2016.7883514.

K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan, “Analyza: Exploring Data with Conversation,” in Proceedings of the 22nd International Conference on Intelligent User Interfaces, Limassol Cyprus, Mar. 2017, pp. 493–504, doi: 10.1145/3025171.3025227.

E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, “Podium: Ranking Data Using Mixed-Initiative Visual Analytics,” IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 1, pp. 288–297, Jan. 2018, doi: 10.1109/TVCG.2017.2745078.

C. Perin, P. Dragicevic, and J.-D. Fekete, “Revisiting Bertin Matrices: New Interactions for Crafting Tabular Visualizations,” IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 2082–2091, Dec. 2014.

T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran, “Effortless Data Exploration with zenvisage: An Expressive and Interactive Visual Analytics

System," arXiv:1604.03583 [cs], Jan. 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1604.03583.

S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual Analysis of Multi-Attribute Rankings," IEEE Trans. Visual. Comput. Graphics, vol. 19, no. 12, pp. 2277–2286, Dec. 2013.

A. Schulz, "Discriminative Dimensionality Reduction: Variations, Applications, Interpretations," Doctoral Thesis - Publikationen an der Universität Bielefeld, 2017.

R. Hourieh, H. Stitz, N. Gehlenborg, and M. Streit, "TaCo: Comparative Visualization of Large Tabular Data," Detail, vol. 1000, no. 1, p.1, 2016.

N. Kamat and A. Nandi, "InfiniViz: Interactive Visual Exploration using Progressive Bin Refinement," arXiv:1710.01854 [cs], Oct. 2017, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1710.01854.

J. Kelly, "Computing, cognition and the future of knowing How humans and machines are forging a new age of understanding," IBM Corporation. Accessed: May 25, 2020. [Online]. Available: https://cloud.report/Resources/Whitepapers/e55108d4-92bd-428a-b432-64530b50c6b9_Computing_Cognition_WhitePaper.pdf.

F. Anderson, "Getting Started Tutorial for IBM Watson Analytics," IBM Corporation, New York, NY, USA, 2012.

P. Koytek, C. Perin, J. Vermeulen, E. Andre, and S. Carpendale, "MyBrush: Brushing and Linking with Personal Agency," IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 1, pp. 605–615, Jan. 2018, doi: 10.1109/TVCG.2017.2743859.

D. Ren, T. Hollerer, and X. Yuan, "iVisDesigner: Expressive Interactive Design of Information Visualizations," IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 2092–2101, Dec. 2014, doi: 10.1109/TVCG.2014.2346291.

M. Budiu, R. Isaacs, D. Murray, G. Plotkin, P. Barham, S. Al-Kiswany, Y. Boshmaf, Q. Luo, A. Andoni, "Interacting with large distributed datasets using Sketch," UW-Madison Department of Computer Sciences, CS Technical Reports, 2015.

P.-M. Law, R. C. Basole, and Y. Wu, "Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification," IEEE transactions on visualization and computer graphics, vol. 25, no. 1, pp. 427-437, 2018.

A. Mokalis and J. Davis, Google Analytics Demystified. CreateSpace Independent Publishing Platform, 2018.

S. Macke, Y. Zhang, S. Huang, and A. Parameswaran, “Adaptive Sampling for Rapidly Matching Histograms,” arXiv:1708.05918 [cs], May 2018, Accessed: May 24, 2020. [Online]. Available: http://arxiv.org/abs/1708.05918.

E. Isaacs, K. Damico, S. Ahern, E. Bart, and M. Singhal, “Footprints: A Visual Search Tool that Supports Discovery and Coverage Tracking,” IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 1793–1802, Dec. 2014, doi: 10.1109/TVCG.2014.2346743.

A. F. Zuur, E. N. Ieno, and C. S. Elphick, “A protocol for data exploration to avoid common statistical problems: Data exploration,” Methods in Ecology and Evolution, vol. 1, no. 1, pp. 3–14, Mar. 2010, doi: 10.1111/j.2041-210X.2009.00001.x.

H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao, “RCLens: Interactive Rare Category Exploration and Identification,” IEEE Trans. Visual. Comput. Graphics, vol. 24, no. 7, pp. 2223–2237, Jul. 2018, doi: 10.1109/TVCG.2017.2711030.

M. Feng, C. Deng, E. M. Peck, and L. Harrison, “HindSight: Encouraging Exploration through Direct Encoding of Personal Interaction History,” IEEE Trans. Visual. Comput. Graphics, vol. 23, no. 1, pp. 351–360, Jan. 2017, doi: 10.1109/TVCG.2016.2599058.

C. D. Stolper, A. Perer, and D. Gotz, “Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics,” IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 1653–1662, Dec. 2014, doi: 10.1109/TVCG.2014.2346574.

D. Perry, B. Howe, A. M. F. Key, and C. Aragon, “VizDeck: Streamlining Exploratory Visual Analytics of Scientific Data,”, In iSchools, 2013, DOI: https://doi.org/10.9776/13206.

A. Satyanarayan and J. Heer, “Lyra: An Interactive Visualization Design Environment: Lyra: An Interactive Visualization Design Environment,” Computer Graphics Forum, vol. 33, no. 3, pp. 351–360, Jun. 2014, doi: 10.1111/cgf.12391.

Y. Wang et al., “InfoNice: Easy Creation of Information Graphics,” in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18, Montreal QC, Canada, 2018, pp. 1–12, doi: 10.1145/3173574.3173909.

K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, “Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations,” IEEE Trans. Visual. Comput. Graphics, vol. 22, no. 1, pp. 649–658, Jan. 2016, doi: 10.1109/TVCG.2015.2467191.

D. Ren, M. Brehmer, Bongshin Lee, T. Hollerer, and E. K. Choe, “ChartAccent: Annotation for data-driven storytelling,” in 2017 IEEE Pacific Visualization Symposium (PacificVis), Seoul, South Korea, Apr. 2017, pp. 230–239, doi: 10.1109/PACIFICVIS.2017.8031599.

R. Bro and A. K. Smilde, “Principal component analysis,” Anal. Methods, vol. 6, no. 9, pp. 2812–2831, 2014, doi: 10.1039/C3AY41907J.

D. A. Keim, “Information visualization and visual data mining,” IEEE Trans. Visual. Comput. Graphics, vol. 8, no. 1, pp. 1–8, Mar. 2002, doi: 10.1109/2945.981847.

M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran, “Towards Visualization Recommendation Systems,” SIGMOD Rec., vol. 45, no. 4, pp. 34–39, May 2017, doi: 10.1145/3092931.3092937.

L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv:1802.03426, Feb. 2018, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1802.03426.

J. A. Lee and M. Verleysen, “Scale-independent quality criteria for dimensionality reduction,” Pattern Recognition Letters, vol. 31, no. 14, pp. 2248–2257, Oct. 2010, doi: 10.1016/j.patrec.2010.04.013.

J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, “Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure,” Neurocomputing, vol. 169, pp. 246–261, Dec. 2015, doi: 10.1016/j.neucom.2014.12.095.

S. Xiang, F. Nie, Y. Song, C. Zhang, and C. Zhang, “Embedding new data points for manifold learning via coordinate propagation,” Knowledge and Information Systems, vol. 19, no. 2, pp.159-184, 2009.

A. Bibal and B. Frenay, “Measuring Quality And Interpretability Of Dimensionality Reduction Visualizations,” In Safe Machine Learning Workshop at ICLR, 2019.

J. Johannemann and R. Tibshirani, “Spectral Overlap and a Comparison of Parameter-Free, Dimensionality Reduction Quality Metrics,” arXiv:1907.01974 [cs, stat], Jul. 2019, Accessed: Sep. 30, 2019. [Online]. Available: http://arxiv.org/abs/1907.01974.

A. Ghosh, M. Nashaat, J. Miller, S. Quader, and C. Marston, “A comprehensive review of tools for exploratory analysis of tabular industrial datasets,” Visual Informatics, vol. 2, no. 4, pp. 235–253, Dec. 2018, doi: 10.1016/j.visinf.2018.12.004.

A. C. Fraideinberze, “Effective and unsupervised fractal-based feature selection for very large datasets: removing linear and non-linear attribute correlations,” In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 615-622, 2016.

J. P. Cunningham and Z. Ghahramani, “Linear Dimensionality Reduction: Survey, Insights, and Generalizations,” The Journal of Machine Learning Research, vol. 16, no. 1, pp. 2859–2900, 2015.

A. Lazar, L. Jin, C. A. Spurlock, K. Wu, A. Sim, and A. Todd, “Evaluating the Effects of Missing Values and Mixed Data Types on Social Sequence Clustering Using t-SNE Visualization,” J. Data and Information Quality, vol. 11, no. 2, pp. 1–22, Mar. 2019, doi: 10.1145/3301294.

J. Tang, J. Liu, M. Zhang, and Q. Mei, “Visualizing Large-scale and High-dimensional Data,” Proceedings of the 25th International Conference on World Wide Web - WWW ’16, pp. 287–297, 2016, doi: 10.1145/2872427.2883041.

M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, “Non-Linear Dimensionality Reduction Techniques for Classification and Visualization,” presented at the Proceedings of the eighth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD), Edmonton, Alberta, Canada, 2002.

G. Navarro, R. Paredes, N. Reyes, and C. Bustos, “An empirical evaluation of intrinsic dimension estimators,” Information Systems, vol. 64, pp. 206–218, Mar. 2017, doi: 10.1016/j.is.2016.06.004.

E. Amid and M. K. Warmuth, “A more globally accurate dimensionality reduction method using triplets,” arXiv:1803.00854, Mar. 2018, Accessed: Apr. 08, 2019. [Online]. Available: http://arxiv.org/abs/1803.00854.

G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, “Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding,” Nature Methods, vol. 16, no. 3, pp. 243–245, Mar. 2019, doi: 10.1038/s41592-018-0308-4.

Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang, "Neighborhood preserving embedding," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 2005, pp. 1208-1213 Vol. 2, doi: 10.1109/ICCV.2005.167.

C. C. Aggarwal, "On the effects of dimensionality reduction on high dimensional similarity search," in Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '01, Santa Barbara, California, United States, 2001, pp. 256–266, doi: 10.1145/375551.383213.

J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, S. Leman, and C. North, "Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics," IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 131–141, Jan. 2018, doi: 10.1109/TVCG.2017.2745258.

K. M. Sunderland, D. Beaton, J. Fraser, D. Kwan, P.M. McLaughlin, M. Montero-Odasso, A.J. Peltsch, F. Pieruccini-Faria, D.J. Sahlas, R.H. Swartz, and S.C. Strother, "The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project," BMC Medical Research Methodology, vol. 19, no. 1, p. 102, May 2019, doi: 10.1186/s12874-019-0737-5.

C. O. S. Sorzano, J. Vargas, and A. Pascual, "A survey of dimensionality reduction techniques," arXiv preprint arXiv:1403.2877, p. 35.

C. Zhang, S. Xiang, F. Nie, and Y. Song, "Nonlinear dimensionality reduction with relative distance comparison," Neurocomputing, vol. 72, no. 7–9, pp. 1719–1731, Mar. 2009, doi: 10.1016/j.neucom.2008.08.003.

B. Rieck and H. Leitte, "Agreement Analysis of Quality Measures for Dimensionality Reduction," in Topological Methods in Data Analysis and Visualization IV, H. Carr, C. Garth, and T. Weinkauf, Eds. Cham: Springer International Publishing, 2017, pp. 103–117.

L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605, 2008.

J. A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen, "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation," Neurocomputing, vol. 112, pp. 92–108, Jul. 2013, doi: 10.1016/j.neucom.2012.12.036.

J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood Components Analysis," Advances in neural information processing systems, pp. 513–520, 2005.

O. Kramer, "Dimensionality Reduction by Unsupervised K-Nearest Neighbor Regression," in 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, USA, Dec. 2011, pp. 275–278, doi: 10.1109/ICMLA.2011.55.

X. Zhao and S. S.-U. Guan, "A subspace recursive and selective feature transformation method for classification tasks," Big Data Anal, vol. 2, no. 1, p. 10, Dec. 2017, doi: 10.1186/s41044-017-0025-5.

T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning, vol. 1. New York: Springer series in statistics, 2001.

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," Journal of Experimental Social Psychology, vol. 49, no. 4, pp. 764–766, Jul. 2013, doi: 10.1016/j.jesp.2013.03.013.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Trans. on Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

Shiming Xiang, Feiping Nie, Changshui Zhang, and Chunxia Zhang, "Nonlinear Dimensionality Reduction with Local Spline Embedding," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1285–1298, Sep. 2009, doi: 10.1109/TKDE.2008.204.

M. Gashler, D. Ventura, and T. Martinez, "Iterative Non-linear Dimensionality Reduction by Manifold Sculpting," p. 8.

A. Arcuri and L. Briand, "A practical guide for using statistical tests to assess randomized algorithms in software engineering," in Proceeding of the 33rd international conference on Software engineering - ICSE '11, Waikiki, Honolulu, HI, USA, 2011, p. 1, doi: 10.1145/1985793.1985795.

M. Mohammadi, W. Hofman, and Y.-H. Tan, "A Comparative Study of Ontology Matching Systems via Inferential Statistics," IEEE Transactions on Knowledge and

Data Engineering, vol. 31, no. 4, pp. 615–628, Apr. 2019, doi: 10.1109/TKDE.2018.2842019.

R. Dror, G. Baumer, S. Shlomov, and R. Reichart, “The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing,” presented at the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, vol. 1, pp. 1383–1392.

R. Sherman, “Error of the normal approximation to the. sum of N random variables,” Biometrika, vol. 58, no. 2, pp. 396–398, 1971.

K. Pearson, “On lines and planes of closest fit to systems of points in space,” The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.

J. B. Kruskal, “Nonmetric multidimensional scaling: A numerical method,” Psychometrika, vol. 29, no. 2, pp. 115–129, Jun. 1964, doi: 10.1007/BF02289694.

Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, “Local and Global Regressive Mapping for Manifold Learning with Out-of-Sample Extrapolation,” In Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.

B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” Neural Computation, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.

D. Ulyanov, “Multicore t-SNE.” Github Repository, 2016, Accessed: May 14, 2020. [Online]. Available: https://github.com/DmitryUlyanov/Multicore-TSNE.

Q. Hu and C. S. Greene, “Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics,” Pac Symp Biocomput, vol. 24, pp. 362–373, 2019.

E. Levina and P. J. Bickel, “Maximum Likelihood Estimation of Intrinsic Dimension,” Advances in neural information processing systems, pp. 777–784, 2005.

D. Dheeru and G. Casey, “UCI Machine Learning Repository.” University of California, Irvine, School of Information and Computer Sciences, 2017, [Online]. Available: http://archive.ics.uci.edu/ml.

Kaggle, “Kaggle Data Repository.” https://www.kaggle.com/datasets, Accessed: Jun. 12, 2019. [Online]. Available: https://www.kaggle.com/datasets.

B. Bischl, G. Casalicchio, M. Feurer, F. Hutter, M. Lang, R.G. Mantovani, R.G., J.N. van Rijn, and J. Vanschoren, “OpenML Benchmarking Suites and the OpenML100,” arXiv:1708.03731, Aug. 2017, Accessed: Jun. 12, 2019. [Online]. Available: http://arxiv.org/abs/1708.03731.

J. Xia, F. Ye, W. Chen, Y. Wang, W. Chen, Y. Ma, and A.K. Tung, “LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 236–245, Jan. 2018, doi: 10.1109/TVCG.2017.2744098.

K. Bunte, M. Biehl, and B. Hammer, “Dimensionality reduction mappings,” in 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, Apr. 2011, pp. 349–356, doi: 10.1109/CIDM.2011.5949443.

W. W. Daniel and C. L. Cross, Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition, Eleventh. Wiley, 2018.

P. S. Efraimidis, “Weighted Random Sampling over Data Streams,” arXiv:1012.0256, Dec. 2010, Accessed: Jun. 24, 2019. [Online]. Available: http://arxiv.org/abs/1012.0256.

J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman and Hall/CRC, 2010.

C. C. Aggarwal, “On Biased Reservoir Sampling in the Presence of Stream Evolution,” in Proceedings of the 32nd international conference on Very large data bases VLDB Endowment, Sep. 2006, pp. 607–618.

R. Tortolani, “Introducing Bias Intentionally into Survey Techniques,” Journal of Marketing Research, vol. 2, no. 1, pp. 51–55, 1965.

S. García, A. Fernández, J. Luengo, and F. Herrera, “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,” Information Sciences, vol. 180, no. 10, pp. 2044–2064, May 2010, doi: 10.1016/j.ins.2009.12.010.

R. R. Bouckaert, “Estimating Replicability of Classifier Learning Experiments,” in Proceedings of the Twenty-first International Conference on Machine Learning, New York, NY, USA, 2004, pp. 15–, doi: 10.1145/1015330.1015338.

V. D. Silva and J. B. Tenenbaum, “Global Versus Local Methods in Nonlinear Dimensionality Reduction,” Advances in neural information processing systems, pp. 721–728, 2003.

C. T. Jr, A. Traina, L. Wu, and C. Faloutsos, "Fast Feature Selection using Fractal Dimension," Journal of Information and data Management, vol. 1, no. 1, p. 14, 2010.

D. Meng, Y. Leung, and Z. Xu, "A new quality assessment criterion for nonlinear dimensionality reduction," Neurocomputing, vol. 74, no. 6, pp. 941–948, Feb. 2011, doi: 10.1016/j.neucom.2010.10.011.

F. Camastra, "Data dimensionality estimation methods: a survey," Pattern Recognition, vol. 36, no. 12, pp. 2945–2954, Dec. 2003, doi: 10.1016/S0031-3203(03)00176-6.

F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," Information Sciences, vol. 328, pp. 26–41, Jan. 2016, doi: 10.1016/j.ins.2015.08.029.

S. Gong, V. N. Boddeti, and A. K. Jain, "On the Intrinsic Dimensionality of Image Representations," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 3982–3991, doi: 10.1109/CVPR.2019.00411.

A. Ghosh, M. Nashaat, and J. Miller, "The current state of software license renewals in the I.T. industry," Information and Software Technology, vol. 108, pp. 139–152, Apr. 2019, doi: 10.1016/j.infsof.2019.01.001.

R. Feldt and A. Magazinius, "Validity Threats in Empirical Software Engineering Research - An Initial Survey," Seke, pp. 374–379, Jul. 2010.

C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable local dimensionality reduction approaches," Pattern Recognition, vol. 42, no. 9, pp. 2054–2066, Sep. 2009, doi: 10.1016/j.patcog.2008.12.009.

A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

S. Lespinats and M. Aupetit, "CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings," Computer Graphics Forum, vol. 30, no. 1, pp. 113–125, Mar. 2011, doi: 10.1111/j.1467-8659.2010.01835.x.

D. Smilkov, N. Thorat, C. Nicholson, E. Reif, F. B. Viégas, and M. Wattenberg, "Embedding Projector: Interactive Visualization and Interpretation of

Embeddings,” arXiv:1611.05469 [cs, stat], Nov. 2016, Accessed: Jul. 12, 2019. [Online]. Available: http://arxiv.org/abs/1611.05469.

R. Faust, D. Glickenstein, and C. Scheidegger, “DimReader: Axis lines that explain non-linear projections,” IEEE Trans. Visual. Comput. Graphics, vol. 25, no. 1, pp. 481–490, Jan. 2019, doi: 10.1109/TVCG.2018.2865194.

J. Stahnke, M. Dork, B. Muller, and A. Thom, “Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions,” IEEE Trans. Visual. Comput. Graphics, vol. 22, no. 1, pp. 629–638, Jan. 2016, doi: 10.1109/TVCG.2015.2467717.

L. Kodali, J. Wenskovitch, N. Wycoff, L. House, and C. North, “Uncertainty in Interactive WMDS Visualizations,” In Visualization in Data Science (VDS at IEEE VIS), 2019.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” ACM Comput. Surv., vol. 51, no. 5, pp. 1–42, Aug. 2018, doi: 10.1145/3236009.

M. Wattenberg, F. Viégas, and I. Johnson, “How to Use t-SNE Effectively,” Distill, 2016, doi: 10.23915/distill.00002.

S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Moller, “DimStiller: Workflows for dimensional analysis and reduction,” in 2010 IEEE Symposium on Visual Analytics Science and Technology, Salt Lake City, UT, USA, Oct. 2010, pp. 3–10, doi: 10.1109/VAST.2010.5652392.

P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, “Local Affine Multidimensional Projection,” IEEE Trans. Visual. Comput. Graphics, vol. 17, no. 12, pp. 2563–2571, Dec. 2011, doi: 10.1109/TVCG.2011.220.

L. Pagliosa, P. Pagliosa, and L. G. Nonato, “Understanding Attribute Variability in Multidimensional Projections,” in 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Sao Paulo, Brazil, Oct. 2016, pp. 297–304, doi: 10.1109/SIBGRAPI.2016.048.

R. R. O. D. Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea, “Attribute-based Visual Explanation of Multidimensional Projections,” EuroVis Workshop on Visual Analytics (EuroVA), p. 5 pages, 2015, doi: 10.2312/EUROVA.20151100.

J. Z. Self, L. House, S. Leman, and C. North, “Andromeda: Observation-Level and Parametric Interaction for Exploratory Data Analysis.” Technical report, Department of Computer Science, Virginia Tech, Blacksburg, Virginia, 2015.

A. A. Freitas, “Comprehensible classification models: a position paper,” SIGKDD Explor. Newsl., vol. 15, no. 1, pp. 1–10, Mar. 2014, doi: 10.1145/2594473.2594475.

F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang, “Correlation Judgment and Visualization Features: A Comparative Study,” IEEE Trans. Visual. Comput. Graphics, vol. 25, no. 3, pp. 1474–1488, Mar. 2019, doi: 10.1109/TVCG.2018.2810918.

J. C. Gower, “A General Coefficient of Similarity and Some of Its Properties,” Biometrics, vol. 27, no. 4, pp. 857–871, 1971, doi: 10.2307/2528823.

H. Wang and M. Hong, “Distance Variance Score: An Efficient Feature Selection Method in Text Classification,” Mathematical Problems in Engineering, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/695720.

F. Pan, A. Roberts, L. McMillan, D. Threadgill, and W. Wang, “Sample Selection for Maximal Diversity,” in Seventh IEEE International Conference on Data Mining (ICDM 2007), Omaha, NE, USA, Oct. 2007, pp. 262–271, doi: 10.1109/ICDM.2007.16.

R. Haftka, “Combining global and local approximations,” AIAA journal, vol. 29, no. 9, p. 1523, Sep. 1991.

P. G. Poličar, M. Stražar, and B. Zupan, “openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding,” Bioinformatics, preprint, Aug. 2019. doi: 10.1101/731877.

W. S. Torgerson, “Multidimensional scaling: I. Theory and method,” Psychometrika, vol. 17, no. 4, pp. 401–419, Dec. 1952, doi: 10.1007/BF02288916.

S. T. Roweis, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” Science, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.

D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” arXiv:1312.6114 [cs, stat], May 2014, Accessed: Mar. 24, 2020. [Online]. Available: http://arxiv.org/abs/1312.6114.

Z. Zhang and H. Zha, “Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment,” SIAM J. Sci. Comput., vol. 26, no. 1, pp. 313–338, Jan. 2004, doi: 10.1137/S1064827502419154.

L. G. Nonato and M. Aupetit, “Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment,” IEEE Trans. Visual. Comput. Graphics, vol. 25, no. 8, pp. 2650–2673, Aug. 2019, doi: 10.1109/TVCG.2018.2846735.

T. Schreck, T. von Landesberger, and S. Bremm, “Techniques for Precision-Based Visual Analysis of Projected Data,” Information Visualization, vol. 9, no. 3, pp. 181–193, Sep. 2010, doi: 10.1057/ivs.2010.2.

F. Heimerl, C. Kralj, T. Möller, and M. Gleicher, “embComp: Visual Interactive Comparison of Vector Embeddings,” arXiv:1911.01542 [cs], Nov. 2019, Accessed: Nov. 11, 2019.

A. Bibal and B. Frenay, “Measuring Quality And Interpretability Of Dimensionality Reduction Visualizations,” p. 7, 2019.

C. Lai, Y. Zhao, and X. Yuan, “Exploring high-dimensional data through locally enhanced projections,” Journal of Visual Languages & Computing, vol. 48, pp. 144–156, Oct. 2018, doi: 10.1016/j.jvlc.2018.08.006.

R. Cutura, S. Holzer, M. Aupetit, and M. Sedlmair, “VisCoDeR: A Tool for Visually Comparing Dimensionality Reduction Algorithms,” Computational Intelligence, p. 6, 2018.

J. Z. Self, L. House, S. Leman, and C. North, “An-dromeda: Observation-Level and Parametric Inter-action for Exploratory Data Analysis.” Technical re-port, Department of Computer Science, Virginia Tech, Blacksburg, Virginia; 2015.

M. Daszykowski, B. Walczak, and D. L. Massart, “Representative subset selection,” Analytica Chimica Acta, vol. 468, no. 1, pp. 91–103, Sep. 2002.

R. Mall, R. Langone, and J. A. K. Suykens, “FURS: Fast and Unique Representative Subset selection retaining large-scale community structure,” Soc. Netw. Anal. Min., vol. 3, no. 4, pp. 1075–1095, Dec. 2013.

B. B. Chaudhuri, “How to choose a representative subset from a set data in multi-dimensional space,” Pattern Recognition Letters, vol. 15, no. 9, pp. 893–899, 1994, doi: 10.1016/0167-8655(94)90151-1.

R. D. Clark, “OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets,” J. Chem. Inf. Comput. Sci., vol. 37, no. 6, pp. 1181–1188, Nov. 1997, doi: 10.1021/ci970282v.

R. W. Kennard and L. A. Stone, “Computer Aided Design of Experiments,” Technometrics, vol. 11, no. 1, pp. 137–148, Feb. 1969.

R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, “Graph-based measures to assist user assessment of multidimensional projections,” Neurocomputing, vol. 150, pp. 583–598, Feb. 2015.

J. Xia et al., “LDSScanner: Exploratory Analysis of Low-Dimensional Structures in High-Dimensional Datasets,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 236–245, Jan. 2018.

K. Bunte, M. Biehl, and B. Hammer, “Dimensionality reduction mappings,” in 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Paris, France, Apr. 2011, pp. 349–356.

M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, “Towards a Quantitative Survey of Dimension Reduction Techniques,” IEEE Trans. Visual. Comput. Graphics, pp. 1–1, 2019.

J. Johannemann and R. Tibshirani, “Spectral Overlap and a Comparison of Parameter-Free, Dimensionality Reduction Quality Metrics,” arXiv:1907.01974 [cs, stat], Jul. 2019, Accessed: Sep. 30, 2019.

M. Aupetit, “Visualizing distortions and recovering topology in continuous projection techniques,” Neurocomputing, vol. 70, no. 7–9, pp. 1304–1330, Mar. 2007, doi: 10.1016/j.neucom.2006.11.018.

S. L. France and U. Akkucuk, “A Review, Framework and R toolkit for Exploring, Evaluating, and Comparing Visualizations,” arXiv:1902.08571 [cs, stat], Feb. 2019, Accessed: Nov. 17, 2019.

P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato, and L. G. Nonato, “Local Affine Multidimensional Projection,” IEEE Trans. Visual. Comput. Graphics, vol. 17, no. 12, pp. 2563–2571, Dec. 2011.

W. Gani and M. Limam, “A kernel distance-based representative subset selection method,” Journal of Statistical Computation and Simulation, vol. 86, no. 1, pp. 135–148, Jan. 2016, doi: 10.1080/00949655.2014.996758.

Y. Tominaga, “Representative subset selection using genetic algorithms,” Chemometrics and Intelligent Laboratory Systems, vol. 43, no. 1–2, pp. 157–163, Sep. 1998, doi: 10.1016/S0169-7439(98)00085-9.

D. Chaudhuri, C. A. Murthy, and B. B. Chaudhuri, “Finding a subset of representative points in a data set,” IEEE Trans. Syst., Man, Cybern., vol. 24, no. 9, pp. 1416–1424, Sep. 1994, doi: 10.1109/21.310520.

A. Ghosh, M. Nashaat, J. Miller, and S. Quader, “Interpretation of Structural Preservation in Low-dimensional Embeddings,” IEEE Trans. Knowl. Data Eng., pp. 1–1, 2020, doi: 10.1109/TKDE.2020.3005878.

C. Angulo, X. Parra, and A. Català, “K-SVCR. A support vector machine for multi-class classification,” Neurocomputing, vol. 55, no. 1–2, pp. 57–77, Sep. 2003, doi: 10.1016/S0925-2312(03)00435-1.

R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local Rule-Based Explanations of Black Box Decision Systems,” arXiv:1805.10820 [cs], May 2018, Accessed: Aug. 12, 2019.

S. Mahmood and K. Mueller, “Taxonomizer: Interactive Construction of Fully Labeled Hierarchical Groupings from Attributes of Multivariate Data,” IEEE Trans. Visual. Comput. Graphics, vol. 26, no. 9, pp. 2875–2890, Sep. 2020, doi: 10.1109/TVCG.2019.2895642.

M. Georgsson and N. Staggers, “Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics,” J Am Med Inform Assoc, vol. 23, no. 1, pp. 5–11, Jan. 2016, doi: 10.1093/jamia/ocv099.

J. Brooke, “SUS - A quick and dirty usability scale,” Usability evaluation in industry. Jun 11 1996, pp. 189.

T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/j.artint.2018.07.007.

A. Strauss and J. Corbin, “Grounded Theory Methodology,” Handbook of qualitative research, vol. 17, pp. 273–85, 1994.

Y. Peng, P. A. Flach, C. Soares, and P. Brazdil, “Improved Dataset Characterisation for Meta-learning,” in Discovery Science, vol. 2534, S. Lange, K. Satoh, and C. H. Smith, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 141–152.

Appendix A

Additional Scholarly Contributions

Journal Papers:

- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, "Transformers Meet Tabular Data: Bidirectional Representation Model for Erroneous Data Detection", Full Research Paper Under Review at: IEEE Transactions on Big Data
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, "Semi-Supervised Ensemble Learning for Dealing with Inaccurate and Incomplete Supervision", Full Research Paper Under Review at: ACM Transactions on Knowledge Discovery from Data
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, "Using Intelligent Active Supervision to Predict Popularity of Mobile News", Full Research Paper Under Review at: International Journal of Mobile Human Computer Interaction (IJMHCI)
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, "Asterisk: Generating Training Data with Automatic Active Supervision", ACM Transactions on Data Science, vol. 1, no. 2, May 202, pp. 1-25
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, Chad Marston. "M-Lean: An end-to-end development framework for predictive models in B2B scenarios", Information and Software Technology, vol. 113, Sep. 2019, pp.131-145

Conference Papers:

- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, "WeSAL: Applying Active Supervision to Find High-quality Labels at Industrial

Scale", 53rd Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, Jan 7, 2020 - Jan 10, 2020

- Mona Nashaat, Aindrila Ghosh, Shaikh Quader, Chad Marston, Jean-Francois Puget, and James Miller, "Hybridization of Active Learning and Data Programming for Labeling Large Industrial Datasets", 2018 IEEE International Conference on Big Data, Seattle, WA, USA.

Patents:

- Shaikh Quader, Aindrila Ghosh, Carmen Stefanita, Patent Invention Disclosure Submitted: "Visual Interpretations of Preserved Structure in Low-dimensional Embeddings", IBM Invention Reference P202001054 - 'PUBLISH' rating from IBM Corporation (Record ID # 96027840)
- Shaikh Quader, James Miller, Aindrila Basak, Mona Nashaat Ali Elmowafy, Patent Invention Disclosure for: "Hybridization of Active Learning and Data Programming for Labelling Large Industrial Datasets", IBM Invention Reference P201804613 – 'PUBLISH' rating from IBM Corporation.

Posters:

- Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston, "Visual Interpretations of Preserved Structure in Low-dimensional Embeddings", Poster presented at: IBM CASCON 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston, "IDLE: Interactive Descriptions for Low-dimensional Embeddings", Poster presented at: IBM CASTLE 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Aindrila Basak, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston, "i-VisMA: Interactive Visualization to Engage End-users in

Multivariate Analysis", Poster presented at: IBM CASCON 2018, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5

- Aindrila Basak, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston, "User-centered Machine Learning Design Process", Poster presented at: IBM CASCON 2017, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, Chad Marston, "WeSAL: Applying Active Supervision to Find High-quality Labels at Industrial Scale", Poster presented at: IBM CASCON 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, Chad Marston, "Smart Learner: Leveraging Latent Label Distributions to Learn from Weak Supervision", Poster presented at: IBM CASCON 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Xingchi Wang, Omar Al-Shamali, Aindrila Ghosh, James Miller and Shaikh Quader, "Classification of Tabular data via Neural Networks", Poster presented at: IBM CASCON 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Mona Nashaat, Aindrila Ghosh, James Miller, Shaikh Quader, Chad Marston, "LaBiD: Automating Weak Supervision to Find Missing Labels for Big Data", Poster presented at: IBM CASTLE 2019, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5
- Mona Nashaat, Aindrila Basak, James Miller, Shaikh Quader, Chad Marston, "Finding Missing Labels for Large Industrial Datasets", Poster presented at: IBM CASCON 2018, 8500 Warden Ave, Markham, Toronto, ON L6G 1A5