DatasetWiz: A Visual Analytics Framework for the Comparative Analysis of Dimensionality Reduction and Clustering Quality

Master's Thesis
of

Aditya Handrale
108489

as part of the study program

curriculum
at the

University of Passau
Faculty of Computer Science and Mathematics

Chair of Cognitive Sensor Systems

Advisor: Prof. Dr. Christoph Heinzl
Assistance: Anja Heim

Passau, DD.MM.20YY

Contents

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Christoph Heinzl, for his invaluable guidance, expertise in visualization research, and unwavering support throughout this thesis journey. His insights into cognitive sensor systems and visual analytics have been instrumental in shaping this work.

I extend my heartfelt appreciation to Anja Heim, my academic assistant, whose patient guidance, constructive feedback, and technical expertise helped me navigate the complexities of visualization design and implementation. Her dedication to helping students succeed is truly commendable.

My deepest gratitude goes to my industry partners at Robert Bosch GmbH: Johannes Mohren and Dr. Sabrina Schmedding. Their real-world perspective on industrial quality assessment challenges provided the essential context that transformed this from a purely academic exercise into a meaningful contribution to industrial practice. The datasets, domain knowledge, and collaborative discussions were invaluable to this research.

I am grateful to the University of Passau for providing the academic environment and resources necessary for this research, and for fostering interdisciplinary collaboration between academia and industry.

Special thanks to my family, my parents and brother, who provided unwavering emotional support and encouragement throughout the challenging periods of this thesis. Their constant reminders to stay focused and not procrastinate (which I may have occasionally ignored) kept me on track during the most demanding phases of this work.

Abstract

In data intensive domains such as manufacturing and medical diagnostics, the success of predictive models is primarily driven by the quality and separability of high-dimensional feature spaces. For practitioners to analyze various properties of these datasets, feature extraction algorithms generate hundreds of features, which are further analyzed using dimensionality reduction (DR) and clustering techniques. An accurate understanding of these high dimensional feature spaces is crucial, especially since data scientists and machine learning engineers make downstream decisions based on the understanding of the feature space. A comparison of various dimensionality reduction techniques, both quantitatively using metrics and qualitatively using visualization is absolutely vital for investigation of the dataset "trainability" and feature space quality assessment. As of now, practitioners rely on "black box" projections and static score tables when analyzing the quality of the high dimensional feature spaces. Visual inspection of lower dimensional projections of these feature spaces is often used to investigate the quality of the dataset and to find various effects, artifacts and outliers. The quality metrics and dimensionality reduction methods must be compared manually, which makes this task time consuming, error prone and cognitively demanding. This thesis aims to support the domain experts in the evaluation of the dataset suitability for downstream classification tasks.

To tackle these challenges, our work introduces DatasetWiz, a comparative visual analytics framework that provides a comprehensive understanding of the feature space quality and projection reliability using summary visualization and two novel visualization techniques. Various dimensionality reduction methods are used to summarize the high dimensional structures and are rendered in a side by side comparison tool called DimCompare (Synchronized dual view scatterplots). Information about why the clusters form is calculated statistically and then visualized using Feature Contribution Glyphs (Cluster annotations that highlight feature level differences). The aggregate performance of the different techniques can be explored in a composite visualisation called Bar-Dar Chart (Summary overview combining Bar chart and Radar chart). The efficacy and usefulness of these visualisations are demonstrated using case studies and a user study with X participants. [The results indicate that DatasetWiz successfully facilitates the identification of structural patterns and artifacts, thereby improving the efficiency of early-stage data diagnostics.]

1 Introduction

In the past few years, high-dimensional data has grown more prevalent in fields like manufacturing, medical diagnostics, environmental monitoring, and quality control. As machine learning has gained popularity across multiple domains, feature extraction techniques, be they statistical, domain-specific, or neural network-oriented, now generate hundreds or even thousands of dimensions/features for a singular data item. These feature vectors often include a lot of information in them, but since they are so complex, it's hard for people to understand, compare, or think about these high-dimensional representations. To address the complexity of these high-dimensional spaces, practitioners make use of Dimensionality Reduction (DR), which is the process of mapping high dimensional data into a lower dimensional representation (typically 2D or 3D), while attempting to preserve the original structural relationships.

1.1 Motivation and problem statement

This problem is especially important in factories and industries, where datasets might not be balanced, might have a lot of noise, or might have been collected under less than optimal settings. Before spending time and money on training a model, machine learning programmers and domain specialists need to perform a trainability assessment to verify whether a dataset is "learnable"or not. This means that the data has enough structure, class separation, or structural signal on its own to make modeling useful. Unfortunately, the technologies we have presently don't assist us in trainability assessment very much. Black-box dimensionality reduction projections, static 2D visualizations like scatterplots, and clustering score tables only show part of the picture. This means that users sometimes have to trust their gut feelings or judgment. Conventional visualization methods do not facilitate the interactive exploration and comparative analysis that is crucial for successful decision-making in industrial settings.

The need for this thesis arose from the disparity between high-dimensional feature representations and human comprehension. A prior research collaboration with Bosch underscored the importance for enhanced feature space diagnostics. The partnership's main goal was to find flaws in industrial hardware parts using images, but the basic problem, i.e. understanding high-dimensional embeddings and their structure, goes much beyond merely images. In fact, the tools that were built in this thesis were tested on conventional, high dimensional CSV-based datasets, such as those that assess air pollution or material strength. This shows that the tools can handle a wide range of data.

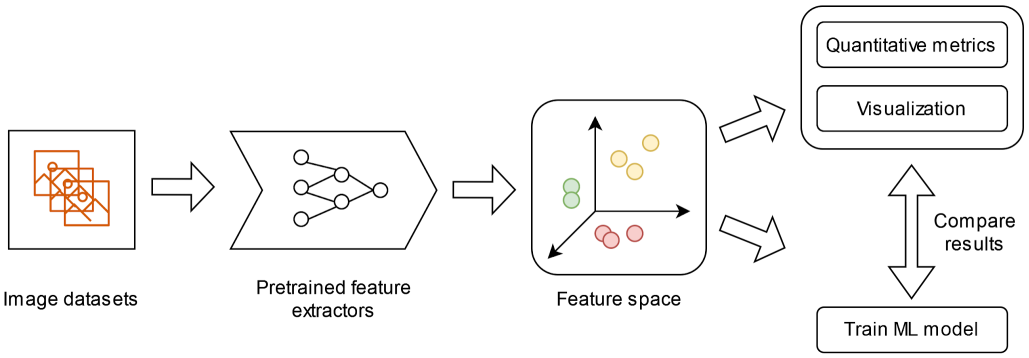

Figure 1: Feature-space analysis pipeline combining dimensionality reduction, visualization, quantitative metrics, and downstream machine learning

More broadly speaking, analyzing and comparing high-dimensional feature spaces is difficult for several reasons:

1. **Projection Instability**: Different Dimensionality Reduction (DR) techniques (PCA, t-SNE, MDS, UMAP) produce inconsistent feature visualizations, each emphasizing different structures. No single method provides a complete picture.
2. **Information loss**: Reducing dimensions often distorts some feature relationships. Clusters may appear well-separated in 2D but actually overlap in the original space, or vice versa.
3. **Lack of visual and quantitative integration**: Clustering metrics like Silhouette Score offer objective numeric values, but interpreting them in context is difficult without visual feedback.

This thesis aims to fill the gap between complex high-dimensional feature spaces and human understanding by introducing novel interactive visualization tools. These tools aim to combine dimensionality reduction, clustering quality metrics, and feature-based visual encodings to support early decision-making in the data science pipeline, for tasks such as evaluating dataset, pointing out anomalies, or identifying the requirement for further preprocessing. Rather than simply illustrating DR results, the goal here is to help the end users, engineers and data scientists, interpret, compare, and assess the quality and separability of high-dimensional data in an intuitive way.

1.2 Research questions

This research is guided by five connected research questions that aim to address both the theoretical and the practical aspects of high-dimensional data visualization in industrial contexts:

Research Question 1 (RQ1): Feature Relationship Understanding

- How can interactive visualization reveal which features drive cluster formation across different dimensionality reduction techniques?

This question seeks to resolve the interpretability challenge that is often posed in high-dimensional feature spaces, where understanding the feature contributions are essential for quality assessment and evaluation, but difficult to achieve with traditional visualization approaches.

Research Question 2 (RQ2): Comparative Evaluation of Dimensionality Reduction

- How can we effectively enable side-by-side comparison of different dimensionality reduction results to support algorithm selection and validation?

This question focuses on the practical needs for practitioners to understand how different dimensionality reduction method choices affect representation of the data and its subsequent analysis.

Research Question 3 (RQ3): Clustering Quality Evaluation and Visualization

- How can composite integrated visualisation effectively facilitate comparative assessment of dimensionality reduction techniques through the aggregation of conflicting quality metrics?

This question addresses the problem of interpreting multiple, and potentially conflicting, quality measures in a visual framework.

Research Question 4 (RQ4): Pattern Identification Across DR Techniques

- How can coordinated multiple views and interactive exploration support the validation of structural patterns and identification of projection induced artifacts across different DR methods?

This question explores the robustness of patterns that are discovered and also the identification of DR technique-specific artifacts.

1.3 Contributions

This thesis makes several novel contributions in the domains of data/information visualization, industrial quality assessment and visual analytics.

1.3.1 Interactive Visualization Techniques

DimCompare Dual-View system: We introduce a novel, dual-view approach for comparing two different dimensionality reduction techniques via synchronized, interactive scatterplots. Compared to traditional single view 2D scatterplots, DimCompare allows for real-time linked interaction between different dimensionality reduction techniques and representations. DimCompare also enables brushing, selection and coordinated exploration.

Dynamic feature contribution glyphs : We developed an innovative cluster annotation "glyphs" that visualize which features contribute most to a cluster formation, dynamically. These cluster annotations provide immediate visual feedback about the feature importance differences between the clusters, which update in real time, as users explore different data subsets.

Bardar Composite Visualization: We created a composite integrated visualiation that utilizes explicit encoding (bar charts) to overcome the well documented perceptual limitations of radar charts (area bias), hence providing an objective ranking of DR method reliability. The integrated design provides users with both detailed metric visualization and also aggregate the performance ranking, enabling more effective metric comparisons across multiple dimensionality reduction techniques.

1.3.2 Technical implementation

Web-based architecture: Development of a complete web-based visualization framework, using Django backend, D3.js frontend, which enables browser-based internet access to visualization capabilities without requiring specialized installation of software.

Scalable Data pipeline: Implementation of efficient algorithms for dimensionality reduction, clustering and metric calculations, that can handle large, industrial scale CSV datasets, while still maintaining, near-real time, interactive performance.

Open source framework: Creation of an open source implementation, that is extensible, and enables reproducible research, while also providing a foundation for future visualization tool development.

1.3.3 Chapter summary

This chapter introduced the motivation, scope and contributions of this thesis. It establishes the core challenge of assessing the "trainability" of a dataset, or separation of high dimensional feature spaces. This is a critical problem in the Industrial AI and quality assessment, as highlighted by the collaboration with Bosch. The key points discussed are :

- **Core problem:** Identified the black box/uninterpretable nature of feature spaces generated as the main problem. The success of a ML model, further in the pipeline, is dependent on this data, hence leads to uncertainty in the data science pipeline.
- **Identified gaps:** We defined some specific limitations of the current methods, like projection instability (different algorithm = different results), information loss (2D projections lose information) and the disconnect between quantitative metrics and qualitative visual inspection.
- **Research questions:** Synthesised the core research questions (RQs) that guide this work, focusing on the feature level understanding, DR comparison, Metric visualization, pattern identification and industrial validation.
- **Novel Contributions:** Introduced the two primary contributions of this thesis, the DimCompare system for visual comparison exploration and the BarDar chart for quantitative metric comparison.

2 Background and related work

This chapter lays the foundation for this thesis by reviewing some basic ideas and existing research that is important for analyzing high-dimensional data, machine learning, and visual analytics. It covers key areas like extracting features from images, methods for dimensionality reduction, clustering algorithms and their related quality metrics, and different methods to visualize and understand High dimensional data.

2.1 Understanding High dimensional data

High-dimensional data refers to the datasets where each of the sample or row is described by a large number of features or variables, often dozens, hundreds, or even thousands. These features can come from a wide variety of sources, including sensor readings, material strength measurements, air quality parameters, or neural network embeddings for images. In this thesis, high-dimensional data is primarily handled in the form of structured CSV files, where each row represents a data instance/sample and each column corresponds to a feature, usually numeric. Such high dimensional data is common in real-world domains like environmental monitoring, quality control, and manufacturing process analysis. However, as the number of dimensions/features increases, it becomes increasingly challenging to explore, visualize or extract meaningful patterns from the data using traditional visualization techniques. This is commonly known as the curse of dimensionality, and it motivates for the use of dimensionality reduction techniques to uncover structure and gives support in interpretation.

2.1.1 Curse of dimensionality

Analyzing high dimensional data has been recognized as one of the fundamental problems in machine learning and data analysis [1]. As also noted by [2], the curse of dimensionality significantly increases computational costs and the storage requirements, while also negatively impacting the accuracy and efficiency of the data analysis methods/algorithms. There is an exponential increase in data sparsity and computational demands as dimensionality grows [3] [4].

This phenomenon is especially evident in image applications where CNN(Convolutional Neural Network) based feature extractors (VGG-19, ResNet-50 etc) are used to generate thousands of features from a single image, creating complex feature spaces that are difficult to interpret as well as computationally intensive.

Anowar et. al. [5] gives a complete comparison of dimensionality reduction algorithms. They categorize them into primarily linear vs non linear and supervised vs unsupervised approaches. The empirical analysis performed by them across challenging datasets clearly demonstrates that different dimensionality reduction techniques excel in different contexts, underscoring a need for a comparative analysis tools that can help practitioners select appropriate methods for their specific applications

2.1.2 Image feature extraction in industrial manufacturing

To evaluate properties like structural integrity of industrial hardware, high dimensional representations are generated by using deep neural network based image feature extractors. In current

manufacturing workflows, pretrained image feature extraction models such as VGG-19 (Visual Geometry Group) and ResNet-50 (Residual Network) are used to convert visual inspection data into numerical feature vectors, consisting of thousands of dimensions. This work also briefly focuses on analyzing the resulting feature spaces to identify production line induced characteristics such as defects and manufacturing inconsistencies.

2.2 Dimensionality Reduction Techniques

Dimensionality reduction is a key set of methods that are used to change the data from a higher dimensional space into a lower dimensional one, while also trying to keep the important properties and structure present in the original data. [6]. These dimensionality reduction methods are generally split into linear and non linear data, where each has different features and compromises. The fact that there exists a vase selection of dimensionality reduction techniques, each having its own biases and compromises, directly indicates why comparative visualization systems are needed. Building such a comparative visualization system is the main goal of this thesis.

2.2.1 Principal Component Analysis (PCA)

This is the most widely used and commonly known Dimensionality reduction method due to its mathematical simplicity and interpretability [7]. It projects the data in a lower dimensional space. PCA does this by finding orthogonal components (principal components) that capture the maximum variation in the data. The linear nature of the technique provides a significant advantage for preserving global structures and relationships in the data [8].

Boileau et al. [9] extend traditional PCA through sparse contrastive PCA, which works by extracting sparse, stable and interpretable features by leveraging control data. Their work demonstrates how PCA variants can be enhanced to address specific domain requirements , particularly in biological applications where interpretability is crucial.

PCA is fast to compute, predictable, and good at keeping the overall structure of the data [8]. However, the downsides of PCA are that it might not capture complex, non-linear relationships present in the data. Which has led to development of kernel PCA and other non-linear extensions [10]. The technique works best when the underlying data is somewhat linear, but it may also miss important patterns in the data that have complex non-linear relationships, such as those found commonly in industrial image analysis applications.

2.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

This is a powerful non-linear method of dimensionality reduction that is really good at showing the local structures present in the data, often revealing tight groups of data points that are similar. t-SNE was introduced by [11], and it revolutionized the visualization of high dimensional data, by preserving local neighborhood structures, while reducing dimensionality.

t-SNE looks at similarities as probabilities and tries to match these probabilities in both the high and low dimensional spaces. It models similarities between data points using probability distributions and minimizes the Kullback-Leibler divergence (KL) between high-dimensional and low-dimensional representations.

It is great at showing clusters that might be hidden in high dimensional space and at preserving local relationships. Making it really valuable for exploratory data analysis (EDA) and pattern discovery. However t-SNE has many limitations that affect the interpretation, it sometimes distorts global structure [12] and the results obtained are sensitive to hyperparameter settings (like perplexity) [13]. Since t-SNE is stochastic in nature, it produces slightly different results in different runs [14]. t-SNE is also computationally expensive O(nˆ2) and can sometimes twist the overall structure, versions like Barnes-hut t-SNE help mitigate this [14].

The non-deterministic nature of this method makes it challenging to reproduce results, increasing the importance of setting random seeds and understanding the impact of hyperparameters upon the final visualisations.

2.2.3 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a classical dimensionality reduction technique that aims at preserving pairwise distances between the different data points, when projecting high dimensional data into lower dimensional space. Originally developed in the field of psychometrics, it is now widely used across various domains. MDS works by transforming a dissimilarity matrix into a geometric configuration in fewer dimensions, while still maintaining the relative spatial relationships of the data. [15]

Unlike linear and variance based methods like PCA, MDS is distance-preserving. It attempts to place each data point in a low dimensional space such that the -inter-object distances are preserved as faithfully as possible. MDS works best when input distances are euclidean, and the data conforms to metric assumptions [16]. There also exist variants of MDS such as non-metric MDS, that relax these assumptions by only preserving the rank of order distances, which makes it more robust to non-linear data structures.

This dimensionality reduction method tried to keep the distances between the data points as much as possible in the lower dimensional space. MDS is good for understanding how far apart items are from each other, but it takes a lot of computing power for larger datasets, (with a computational time complexity of O(n^3) for exact calculation, where N refers to the number of data points) . This makes it less practical for datasets without approximation strategies [17]. Because it needs so much computing, it's usually preferred for smaller datasets. Moreover, it does not explicitly model local or global structure tradeoffs the way t-SNE or UMAP do. Therefore MDS can struggle to highlight cluster boundaries and maintain neighborhood relationships in sparse datasets. MDS provides a valuable contrast to techniques like PCA, t-SNE and UMAP. Its role in this thesis is to primarily serve as a comparative benchmark.

2.2.4 Uniform Manifold Approximation (UMAP)

UMAP was developed by [18], and it addressed several limitations of t-SNE while maintaining the ability to preserve local structure. It is based on manifold learning theory and topological data analysis, and provides faster computation than t-SNE while better preserving global structure alongside local neighborhoods.

This method constructs a high dimensional graph representation of the data and then optimises low dimensional graphs to be as structurally similar to the high dimensional graph as possible.

This kind of approach allows UMAP to handle larger datasets more efficiently than t-SNE, while also producing more stable results across multiple different runs.

UMAP has the ability to preserve both local and global structures, which makes it particularly valuable for industrial applications where understanding both the detailed cluster structure and overall data structure is important, like exploratory analyses in industrial datasets [19]. However, the technique introduces its own set of hyperparameters and assumptions that can impact the final visualization, re- emphasizing the need for competitive analysis tools.

2.2.5 Other techniques : Isomap and Autoencoders

Beyond PCA, t-SNE, UMAP and MDS, there exist additional dimensionality reduction techniques which offer alternative perspectives on high dimensional data, like Isomap and Autoencoders. Isomap [20] extends the classic MDS technique by incorporating geodesic distances, instead of euclidean ones, which allows it to preserve the intrinsic geometry of non-linear manifolds. It constructs a neighborhood graph, and computes the shortest path distances between the points. However, Isomap is sensitive to noise and outliers, and its performance degrades if the neighborhood graph is poorly constructed.

Autoencoders, on the other hand, are unsupervised neural network models that learn to compress the data into lower-dimensional latent representations and reconstruct it back to the original space. [21]. This learned embedding captures the non-linear dependencies and is highly flexible due to the representational power of deep neural networks. There also exist variants such as denoising autoencoders or Variational Autoencoders (VAEs) that have further improved robustness and generative capabilities. However, autoencoders typically require very large training data sets and careful tuning, to learn trivial representations and avoid overfitting.

While these techniques were not the focus of the implementation in this thesis, they offer valuable alternatives and are potential candidates for the future extensions of comparative visualization tools, particularly in deep learning focused workflows.

2.3 Comparative analysis challenges

Espadoto et. al [22] provides a quantitative survey of various dimensionality reduction techniques, it evaluates how these DR methods perform across a wide variety of datasets and metrics. The study shows that no single method always outperforms the others, and each one ever gives only a partial or biased view of the high dimensional data, which underscores the need for comparative analysis tools. This is why the core design of DimCompare is justified.

The authors identified several challenges in evaluating dimensionality reduction techniques:

1. Lack of ground truth, which makes it difficult to assess the quality of the projections.
2. Tradeoffs exist between local and global structure preservation.
3. Dataset characteristics, such as noise and dimensionality, have an influence on the dimensionality reduction performance.
4. Computational scalability becomes important for practical use on high-dimensional datasets.

These challenges emphasize the need for visualization tools, like the ones developed in this thesis. These tools help practitioners navigate the trade-offs and make informed choices based on context-specific requirements.

3 Introduction to LaTeX

Since LaTeX is widely used in academia and industry, there exists a plethora of freely accessible introductions to the language. Reading through the guide at `https://en.wikibooks.org/wiki/LaTeX` serves as a comprehensive overview of most of the functionality and is highly recommended before starting with a thesis in LaTeX.

This is a nice little introduction with a reference to Figure 2.

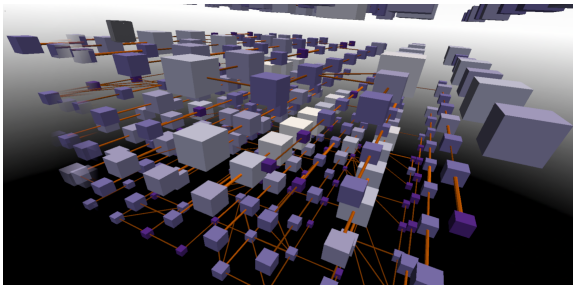

Figure 2: This is a figure with a caption adapted from .

Arguments have to be cited with the cite-command:

A citation may either appear at the end of a sentence or paragraph representing where the idea/research comes from. For example: Scatterplot matrices, parallel coordinate diagrams, and dimensionality reduction techniques are some possibilities to display multiple dimensions simultaneously [23]. Or you can cite the author(s) within the sentence. For example: Further information about Immersive Analytics can be found in the work of Gall et al. [**?**]. Sometimes you may need a direct quotation like: Start of sentence “[...] random citation [...]” as described by [**?**].

4 Content Section 1

Of course, the content chapters of your thesis should be renamed. The number of chapters you need to write depends on the specific task(s) of your thesis and cannot be said in general.

4.1 Subsection 1

You can reference any chapter, section or subsection by its label: looking forward to Section 4.2. You can also reference Appendix A and add footnotes if necessary.[1]

...

[1] This is a footnote ending with a full stop.

4.2 Subsection 2

You can add tables by using the following commands. There is no need for a list of tables or a list of figures in this paper-style thesis. But you should always reference tables and figures in the text (see Table 1).

Item		
Animal	Description	Price ($)
Gnat	per gram	13.65
	each	0.01
Gnu	stuffed	92.50
Emu	stuffed	33.33
Armadillo	frozen	8.99

Table 1: A table that lists items [**?**]

5 Content Section 2

During the writing of your thesis, you may want to highlight some parts or take notes. This is accomplished by using the following command for highlighting or taking notes in the text.

This is a note in the text

If you want to take notes on the side, use the following command. Also make sure that you remove all your notes before submission.

This is a marginal note

5.1 Subsection 1

Adding formulas is easy. Either inline formulas like $E = mc^3$ or unnumbered equations like

$$E = mc^3$$

or a single numbered equation like

$$E = mc^3 \tag{1}$$

or multiple numbered equations that are aligned like

$$E = mc^3 \tag{2}$$
$$a^2 = b^2 + c^2. \tag{3}$$

You can also reference the equations if you set a label. Our approach is captured by the fundamental equations (2) and (3).

5.2 Subsection 2

…

6 Discussion

…

6.1 Subsection 1

…

6.2 Subsection 2

…

7 Conclusion

…

Bibliography

[1] Rizgar R. Zebari, Adnan M. Abdulazeez, Diyar Q. Zeebaree, Dilovan A. Zebari, and Jwan N. Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(1):56–70, 2020.

[2] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: A review. *Complex & Intelligent Systems*, 8:2663–2693, 2022.

[3] Richard Ernest Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[4] Dehua Peng, Zhipeng Gui, and Huayi Wu. Interpreting the curse of dimensionality from distance concentration and manifold effect. *arXiv preprint arXiv:2401.00422*, 2023.

[5] Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378, 2021.

[6] Carlos Oscar S. Sorzano, Juan Vargas, and Alberto P. Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[7] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30, Apr. 2021.

[8] I. T. Jolliffe. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 2016.

[9] Philippe Boileau, Nima S Hejazi, and Sandrine Dudoit. Exploring high-dimensional biological data with sparse contrastive principal component analysis. *Bioinformatics*, 36(11):3422–3430, 2020.

[10] Bernhard Scholkopf, Alexander J. Smola, and Klaus-Robert Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[11] Laurens J.P. van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[12] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 10(1):5416, 2019.

[13] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. Distill, 2016. Online article.

[14] Laurens van der Maaten. Barnes-hut t-sne. *arXiv preprint*, 2013.

[15] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, NY, 2nd edition, 2005.

[16] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. Quantitative Applications in the Social Sciences. SAGE Publications, 1978.

[17] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(3):451–490, 2010.

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[19] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Uniform manifold approximation and projection (umap) and its variants: Tutorial and survey. *arXiv preprint arXiv:2109.02508*, 2021.

[20] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[21] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[22] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Towards a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2019.

[23] Tamara Munzner. *Visualization Analysis and Design*. A K Peters Visualization Series. Taylor & Francis Ltd., Boca Raton, Florida, December 2014.

Appendix

A First Appendix

…

Eidesstattliche Erklärung

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung der Universität Passau zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben. Die Arbeit ist weder von mir noch von einer anderen Person an der Universität Passau oder an einer anderen Hochschule zur Erlangung eines akademischen Grades bereits eingereicht worden.

Passau, den DD.MM.20YY _____

Aditya Handrale

Ich versichere hiermit wahrheitsgemäß, dass

- ❑ die Arbeit ohne Zuhilfenahme von ChatGPT oder anderen generativen KI-Werkzeugen erstellt wurde, oder
- ❑ ich in der nachfolgenden Tabelle vollständig dokumentiert habe, wie solche Systeme bei der Entwicklung der Arbeit verwendet wurden.

Passau, den DD.MM.20YY _____

Aditya Handrale

Generative KI-Werkzeuge, die in der Arbeit verwendet wurden.

Kapitel	KI-Tool	Version	Prompt	Erklärung/Kommentar
1.2	ChatGPT	3.5	Schreibe einen Absatz über den Digital Markets Act.	Der generierte Output wurde in folgender Weise angepasst ...
2.3	ChatGPT	4.0	...	...
3.1	ChatGPT	3.5	...	...