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.

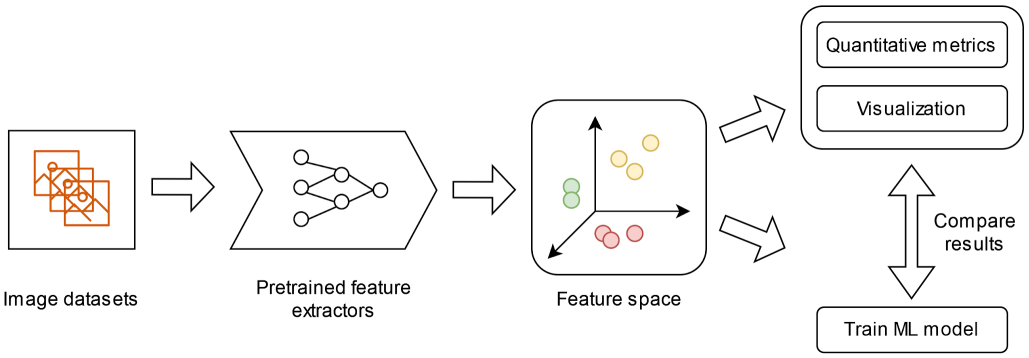

Figure 1: Feature-space analysis pipeline combining dimensionality reduction, visualization, quantitative metrics, and downstream machine learning

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 Jia et. al. [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 $\mathcal{O}(n\hat{}2)$ 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 $\mathcal{O}(n^3)$ 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.

Method	**Linearity**	**Structure Preservation**	**Time Complexity**
PCA	Linear	Global (variance)	$\mathcal{O}(nd^2)$
t-SNE	Nonlinear	Local	$\mathcal{O}(n^2)$
UMAP	Nonlinear	Local + Global	$\mathcal{O}(n \log n)$
MDS	Nonlinear	Distance preserving	$\mathcal{O}(n^3)$

Table 1: Summary of Dimensionality Reduction Techniques

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.

2.4 Clustering algorithms for feature space analysis

Clustering algorithms help us identify groups of similar data points that are present in the feature space and to detect outliers, which is essential for evaluating the structure and separability of high dimensional datasets. Clustering methods are commonly classified into four categories, Partitioning methods (e.g., k-Means, GMM), which assign points into clusters, based on minimizing the within-cluster variance or maximizing likelihood. Density-based methods (e.g., DBSCAN, HDBSCAN), which find clusters by identifying dense regions separated by sparse areas. Graph-based methods (e.g., Spectral clustering), which make use of eigenvectors of similarity matrices to reveal structure. Subspace/high-dimensional methods (e.g., SUBCLU, ENClust) which discover clusters that exist in subsets of dimensions.

2.4.1 k-Means Clustering

k-Means is a centroid-based algorithm that partitions data into k clusters by minimizing within-cluster variance using an iterative refinement method [23]. k denotes the number of clusters, a hyperparameter chosen either based on prior knowledge or quality metrics such as the silhouette score.

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

- n = number of data points,
- k = number of clusters,
- d = dimensionality,
- I = number of iterations.
- Time complexity: Typically $\mathcal{O}(I \cdot n \cdot k \cdot d)$

It is efficient, easy to implement and compatible with many internal cluster metrics. However, the drawbacks of k-Means are that it assumes spherical clusters of similar size, and is sensitive to outliers and initialization while also requiring to specify k. Because k-Means assumes spherical clusters of similar size, visual inspection with DimCompare can help assess whether this assumption holds in practice.

2.4.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

A density-based method of clustering that identifies arbitrary shaped clusters based on density connectivity and marks low-density points as noise [24]. DBSCAN defines clusters as areas of high density separated by sparse regions. A point p is considered a core point if at least $minPts$ neighbors fall within a radius ε. Clusters are formed by density connectivity.

$$|\{q : \|q - p\| \leq \varepsilon\}| \geq \text{minPts}$$

- ε = neighborhood radius,
- $minPts$ = minimum number of neighbors.
- Time complexity: Typically $\mathcal{O}(n \log n)$

It works well for non-convex clusters and clusters of arbitrary shapes and also detects noise. But choosing appropriate distance thresholds can be challenging, especially in high dimensional spaces where distance metrics start becoming less and less meaningful [25].

2.4.3 HDBSCAN (Hierearchical Density-Based Spatial Clustering of Applications with Noise)

An extension of DBSCAN, HDBSCAN constructs a hierarchical clustering structure and extracts clusters based on their stability, and extracts the stable ones [26].

It eliminates the need to choose epsilon and automatically determines the number of clusters. Time complexity : $\mathcal{O}(n^2)$

It is good at discovering clusters of varying densities and identifying outliers without specifying k. This makes it suitable for practical exploratory data analysis scenarios where cluster counts are often unknown or feature spaces contain a lot of noise. Its weaknesses are that it may over split clusters in noisy high dimensional spaces and it is more computationally intensive.

2.4.4 Spectral Clustering

Spectral clustering is a graph based technique of clustering that uses eigenvectors of the similarity matrix's Laplacian to cluster data in reduced dimensions [27]. It is particularly effective at capturing complex cluster structures, but it scales poorly with large scale datasets, due to the costs related to eigen decomposition.

Spectral clustering uses the eigenvectors of a graph Laplacian L=D minus W derived from a similarity matrix W, to embed data before clustering.

$$L = D - W$$

where W is the similarity (adjacency) matrix and D is the degree matrix with $D_{ii} = \sum_j W_{ij}$.

Parameters: Similarity function (e.g., Gaussian kernel), number of clusters k Time Complexity: $\mathcal{O}(n^3)$

2.4.5 Gaussian Mixture Models (GMM)

This is a probabilistic, soft clustering approach that assumes that data is generated from a mixture of various Gaussian distributions. GMM's provide probabilistic class assignments and can model covariance structures well [28]. However, they often do not perform well when clusters deviate significantly from Gaussian shapes and require very careful hyperparameter tuning.

$$p(x \mid \theta) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i)$$

2.4.6 Subspace clustering

These types of methods, like SUBCLU, identify clusters present in specific subsets of dimensions [29]. This is critical in high-dimensional data, where clusters sometimes only exist in particular feature subspaces, and some global clustering methods might miss such latent structures. Their drawback is that they have high computational complexity.

While a wide range of clustering algorithms exist, such as DBSCAN, Spectral Clustering , Gaussian Mixture Models, and Subspace Clustering methods, this thesis focuses primarily on k-Means. This method was selected because k-Means offers a simple, efficient, and widely used baseline, allowing for robust evaluation of feature spaces within the visualization framework.

2.5 Projection and Clustering Quality Metrics

Evaluating the quality of clusters and hence the quality of the feature space, is extremely important for downstream tasks like model training and decision making. There exists methods of quantifying the quality of clusters, known clustering validity indices, a.k.a clustering quality metrics. Cluster validity indices can be classified as internal, known as intrinsic and external, known as extrinsic, each offering a different perspective. It is crucial to understand that no single metric is universally optimal, each metric has biases and limitations of its own. It is important to remember that each metric looks at a slightly different part of "cluster quality" and no single measure is perfect for everything, especially in high dimensional scenarios, where results can often be misleading, without normalization of contextual analysis [30]. High dimensional data can also distort metrics, so it is essential to combine quantitative evaluation with visualization. [31]

2.5.1 Internal Validity Indices (Intrinsic Metrics)

These types of metrics check the cluster quality based only on the data itself, without requiring outside ground truth labels. They were found to be very sensitive to data quality problems like blurry images and wrong labels, making them a good fit for automatic quality checks. While internal clustering metrics assess the visual separation of clusters, they often lack the ability to validate the integrity of underlying dimensionality reduction process.

2.5.1.1 Silhouette coefficient

It measures how similar a data point is to its own cluster (how close it is), compared to other clusters , (how far apart it is). It was originally proposed by Rousseeuw [32]. The values go

from -1 to +1. Higher values mean better-defined clusters. Values above 0.5 are generally good, values above 0.7 indicate strong clustering, whereas negative values suggest that it is in the wrong cluster, i.e. misassignment. Finally, values around 0 indicate that there is some overlap.

The silhouette score for a point i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where $a(i)$ is the average distance between i and all other points in its cluster, and $b(i)$ is the minimum average distance to points in any other cluster.

While it is easy to understand, Silhouette coefficient tends to prefer round, equally sized clusters, and it might not be as useful for clusters that are oddly shaped. One study found it to be a better indicator than Davies-Bouldin and Dunn indices [**?**].

2.5.1.2 Davies Bouldin Index (DBI)

Introduced by [33], it calculates the average similarity ratio of each cluster with its most similar cluster. Similarity is defined as the ratio of how spread out things are within a cluster to how far apart the clusters are. Lower values mean better separation, with 0 being the best possible value, meaning that the clusters are perfectly separate. DBI is defined as

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}, \quad DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} R_{i,j}$$

where S_i is the intra-cluster dispersion for cluster i, and $M_{i,j}$ is the distance between cluster centroids and k is the number of clusters. Lower DBI values indicate better clustering

DBI is sensitive to outliers and differences in cluster shape and density. [34]

2.5.1.3 Trustworthiness (Dimensionality reduction fidelity)

To fill the "trust" gap that's inherent to low-dimensional projections we use the Trustworthiness metric. Originally proposed by Venna and Kaski [**?**], this metric quantifies the degree to which local neighborhood structure is preserved when data is being mapped from a high dimensional space to a lower dimensional 2D projection. Specifically, it is used to measure the presence of false neighbors, i.e. data points that appear closer in visualization, but which are actually distant in the original feature space. This metric is bounded between 0 and 1, which makes for an objective comparison of different projection methods, across diverse datasets and ensures a consistent normalization. If a projection has high trustworthiness, (near 1.0) the user can be confident that the clusters that they see on the screen are real structures and not artifacts of the algorithm.

It is defined as

$$T(k) = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in N_i^k} \max(0, (r(i,j) - k))$$

Where n is the number of samples (data points).k is the number of nearest neighbors considered. N_i^k is the set of k nearest neighbors of sample i in the output (embedded) space. $r(i,j)$ is the rank of sample j in the input (original) space, when ranked by distance from sample i (e.g., the closest neighbor has rank 1).

2.5.1.4 Calinski Harabasz Index (CHI)

It is defined as the ratio of how spread out things are between clusters to how spread out things are within the clusters. It was first introduced in 1974 [35]. A higher CHI score means clusters are dense and well separated. This metric has no upper limit and it is fast to compute. It is often used to find the best number of clusters by looking at the peak value as the value of k (no. of clusters) changes.

It is defined as

$$CHI = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

where $BCSS$ is between-cluster sum of squares, and $WCSS$ is within-cluster sum of squares. Higher values indicate better-defined clusters

However, it is not always linear or the best indicator for feature space quality, no upper limit also makes it challenging to do a clear evaluation.

2.5.1.5 Dunn Index (DI)

The Dunn Index aims to identify dense and well separated clusters.It is defined as the ratio of the minimum inter-cluster distance (the shortest distance between any two points in different clusters) to the maximum intra-cluster distance (the diameter of the largest cluster). It was originally introduced in 1973 by J.C. Dunn [36]. A higher value indicates better separation and compactness.

$$DI = \frac{\min_{i \neq j} \delta(C_i, C_j)}{\max_k \Delta(C_k)}$$

Where $\delta(C_i, C_j)$ is the distance between clusters C_i and C_j, and $\delta(C_k)$ is the diameter of cluster C_k.

Its major drawbacks are that it is sensitive to noise and outliers, which can artificially decrease the inter-cluster distance. It has a high computational complexity on large datasets, as it requires calculating numerous pair wise distances. [37]

2.5.1.6 S_Dbw

S_Dbw is a metric that measures both the compactness and the separation of the clusters. It achieves this by combining two components, an intra-cluster variance term (*Scat*) that measures compactness and an inter-cluster density term ($Dens_{bw}$) that measures separation based on the density of points in the region between clusters.

In research [38], it has been found to be strong across different effects and works well on datasets with noise and complex cluster shapes, it outperforms many traditional indices. Its primary limitation is its implementation complexity compared to simpler metrics.

2.5.1.7 FERM (Feature space Evaluation and Representation Method)

FERM (Feature space Evaluation and Representation Method) is a modern metric designed to evaluate qualities of learned feature spaces for a given classification task. Unlike traditional clustering metrics, which are unsupervised, FERM uses class labels to give a quantitative value to a data representation. It evaluates two properties, class separability (how distinct the representation of the two classes are) and class density (how tightly grouped data points of the same class are). [39]

This makes it particularly suitable for evaluating the feature spaces generated by Deep Neural Networks (DNN), such as ones generated for image analysis in industrial tasks, such as image feature extractors (ex. ResNet, VGG-19 and EfficientNet-B0). Where the goal is to learn a simple representation that eases the task for a downstream classifier.

Metric	**What it Measures**	**Range**	**Better When**
Silhouette Coefficient	Cohesion vs separation of clusters	-1 to +1	Higher
Davies–Bouldin Index	Intra-cluster dispersion relative to inter-cluster separation	0 to ∞	Lower
Calinski–Harabasz Index	Ratio of between vs within cluster variation	$[0, \infty)$	Higher
Dunn Index	Minimum inter-cluster separation over maximum intra-cluster diameter	$[0, \infty)$	Higher
Trustworthiness	Preservation of neighbor ranks in projections	0 to 1	Higher
S_Dbw Index	Compactness and density separation	dataset dependent	Lower*
FERM	Feature space class separability	dataset dependent	Higher*

Table 2: Summary of Internal Clustering and Projection Quality Metrics

2.5.2 External Validity Indices (Extrinsic Metrics)

External Validity indices, evaluate the quality of clustering result by comparing it to a ground truth classification part of the data. This implies that the data should be labelled, making these

types of metrics supervised. This type of comparison allows an objective assessment of how well the clustering algorithm has retained the underlying structure of the data labels. While being good tools for benchmarking, the main drawback is the reliance on pre-existing labels, making them unsuitable for many unsupervised discovery tasks where ground truth is often unknown.

2.5.2.1 Adjusted Rand Index (ARI)

This metric measures the similarity between two partitions, i.e. the clustering results and the ground truth labels provided. It does this by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in both partitions. It then corrects this Rand Index (RI) for randomness/chance, ensuring that the expected value from random clusterings is 0. It has a range of [-1, 1] where 1 indicates a perfect match and values near 0 indicate random agreement, this makes ARI highly interpretable. As noted by [40], the adjustment for chance is the key advantage over the original Rand Index, which prevents inflated scores with a large number of clusters.

Given a contingency table n_{ij} for clustering labels U and V, with row sums a_i and column sums b_j, total n:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{1}{\binom{n}{2}}\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\frac{1}{2}\left(\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right) - \frac{1}{\binom{n}{2}}\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}$$

2.5.2.2 Normalized Mutual Information (NMI)

Originated in information theory, NMI quantifies the statistical information shared between clustering assignment and the ground truth labels. The score is normalized to a [0,1] range, where 1 indicates perfect correlation. A comprehensive analysis by [41] highlights that NMI is a robust and widely used metric, but its value can be influenced by the number of clusters, and different normalization methods yield different results.

Given entropies $H(U)$, $H(V)$ and mutual information $I(U,V)$, it is defined as:

$$NMI(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}}$$

2.5.2.3 V-Measure

This is an entropy based metric that combines two desirable properties, homogeneity and completeness. A clustering is homogeneous if each cluster contains only members of one class. A cluster is considered complete if all the members of a given class are assigned to the same cluster. The V-measure is the harmonic mean of these two values, providing a single, interpretable score between 0 and 1. [42]. This metric was introduced to address the shortcomings in other approaches that might ignore the two properties stated above.

$$V = 2 \cdot \frac{hc}{h+c}$$

Where h and c are normalized information-theoretic scores.

2.5.2.4 Fowlkes-Mallows Index (FMI)

FMI is defined as a geometric mean of precision and recall. FMI computes the similarity between two clusterings by seeing the number of pairs of points that exist in the same cluster in both the partitions. It ranges from [0,1] and can be considered intuitive, however, it can be sometimes misleading. As observed in some studies, FMI assigns high scores even if the clusterings don't align with the ground truth, especially if an algorithm produces a large number of small, pure clusters. It is considered generally to be less discriminative than ARI or NMI due to this. [43] Given true positive (TP), false positive (FP), false negative (FN):

$$FMI = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$$

2.5.3 Additional metrics

Other useful metrics for feature space analysis also exist beyond the standard clustering indices.

2.5.3.1 No. of outliers

This refers to the number of data points which deviate significantly from their respective cluster centers, a.k.a. Outliers. The higher the value of this metric, the worse is the quality of the feature space. This simple metric is especially useful for understanding feature space noise and data integrity.

2.5.3.2 Mahalanobis distance

Mahalanobis distance measures the distance of a data point from a data distribution, it is defined by For a sample x, distribution mean μ and covariance Σ:

$$d_M(x) = \sqrt{(x-\mu)^\top \Sigma^{-1} (x-\mu)}$$

This distance accounts for covariance structure and is very effective for doing multivariate outlier detection/ anomaly analysis. [44]

Bhattacharya distance This metric measures the overlap between two probability distributions. Higher Bhattacharya distances imply more dissimilarity. It also serves as a strong metric for evaluating feature separability and also to perform feature selection. [45] [46] Given two distributions p and q:

$$D_B(p,q) = -\ln\left(\sum_x \sqrt{p(x)\,q(x)}\right)$$

Higher distances implies less overlap.

2.5.3.3 Jefferies Matusita

This is a normalised variant of Bhattacharya distance, whose value is bounded between 0 and 2. It is also used often in feature selection due to its intuitive scale and computational efficiency. [47]

$$JM(p,q) = \sqrt{2\left(1 - e^{-D_B(p,q)}\right)}$$

Metric	Category	Range	Better When
Adjusted Rand Index (ARI)	External	$[-1,1]$	Higher
Normalized Mutual Information (NMI)	External	$[0,1]$	Higher
V-Measure	External	$[0,1]$	Higher
Fowlkes-Mallows Index (FMI)	External	$[0,1]$	Higher
Mahalanobis Distance	Distance	$[0,\infty)$	Lower
Bhattacharyya Distance	Distribution distance	$[0,\infty)$	Higher
Jeffries-Matusita	Distribution distance	$[0,2]$	Higher

Table 3: Summary of External and Distance-Based Clustering/Feature Space Metrics

2.5.4 Need for visualization aided interpretation

It is very crucial to note that simply relying exclusively on metrics is not enough, since they simplify complex underlying structures into a single digestible numeric value. Visualization is still essential for understanding nuance on cluster forms and context. This means that we need some kind of interactive visual exploration for gaining a full understanding of the feature space.

The general observation is that internal metrics match well with classification accuracy and supports using these metrics as a way to guess future model performance (e.g. Silhouette scores often align with model performance), suggesting that intrinsic metrics can serve as a proxy for clustering performance when ground truth is not available, as is often unavailable in real world scenarios [**?**]. This thesis acknowledges the biases and limits of individual clustering metrics. Therefore, it uses a smart approach which focuses on checking multiple metrics, and confirming them with visualization, to give a more reliable and detailed view of the feature space quality.

2.6 Visualization of Feature Space and Interpretability

High dimensional datasets and their feature spaces are hard to interpret with numbers alone, there is an underlying visual spatial structure of the feature spaces generated, that metrics might take into calculation, but they ignore the power of visualization techniques and visual interpretation which can enhance the comprehension of such datasets.

Internal clustering metrics compress rich structures into numeric scores, which usually hide trade-offs like local neighborhood preservation vs. global layout stability. Visualization can complement these internal metrics, by exposing patterns, distortions and outliers that affect the downstream decision making. In this thesis, the goal is to not just show low dimension projects, but

to also help end users connect what they visually see in the 2D representations with quantitative, metric-based evidence. Empirical studies also demonstrate that quality metrics can be used to guide visual analyses and filter our uninteresting and cluttered views. [48]

2.4.1 Why visualization is needed alongside metrics Quality metrics offer many valuable insights, each capturing different aspects of the feature space but they may sometimes conflict with each other on the same data. Prior research shows that no single metric can tell the whole story, since "quality" can encompass things such as clutter, overlap, pattern recognition etc. [48] This is why visual analytics is so powerful, metrics guide the whole process, but in the end, people are needed to inspect, compare and interpret the data.

Visualization theory provides some design guidance for doing it well. Munzner's nested model emphasizes firstly clarifying data and task abstractions, then choosing the appropriate encodings and interactions and finally, validating the result. And for comparison based tasks, Gleicher's taxonomy highlights three patterns, which are, juxtaposition, superposition and explicit encoding of differences. The methods proposed in the thesis, aim to follow these guiding visualization theory principles. DimCompare does this by juxtaposing projections to compare them by showing them side by side and using annotations for clusters (explicit encoding). BarDar uses explicit encoding to show multiple quality metrics in one easy to read chart. [49] [50]

2.6.1 What static plots miss

Static scatterplots are the way to traditionally visualize high dimensional representations, while still useful, they often hide the crucial artifacts that occur after performing dimensionality reduction, like distortions (false neighbors i.e. data points which appear closer together but are not) and tears (data points closer together in the original data are pulled apart in the plot). A body of research highlights these issues, it is argued that analysts must assess first how reliable the projection is first, before making conclusions. These studies show that different dimensionality reduction techniques often optimize for different criterias, which leads to disagreements in the resulting structure. To address this, they propose use of diagnostic overlays and metrics to reveal errors. [51]

Our system adopts this approach, we make the differences visible, highlight the unique features that drive each cluster formation, and also pair visualizations with the quantitative metrics to help prevent misinterpretation.

Traditional radar charts are difficult to interpret because if the order of the axes is changed, the shapes change for the same data and can sometimes get super noisy visually, meaning the order of the axes matters. Moreover, the shapes are hard to read, human brains aren't good at judging and comparing complex irregular shapes [52], it is hard to tell which polygon is smaller or bigger just by visual inspection. This makes just static radar plots unsuitable for intuitive comparative analysis.

Our system BarDar solves these issues, we use a hybrid design that combines simple radar-style overview with a separate bar chart, we tried to use an approach based on well known criticisms for radar plots. This barchart shows the aggregated score of each projection, so the user doesn't have to guess which radar chart is better and do the math in their head.

2.6.2 Visual encodings for feature spaces

Glyphs are a common way to encode multivariate attributes at points or in regions. A recent report discusses when glyphs help and how to design them for readability and detail [15] DimCompare's floating cluster annotations extend this by providing cluster level summaries, which highlight what features deviate from the rest of the data, tying together the visualizations back to the original data space.This design makes use of explicit encodings and the recent literature to enhance projections with context information that is often lost when using standard methods.

2.6.3 How this thesis aims to extend existing tools

There already exist relevant visualization tools like Embedding Projector [53] and Clustrophile 2 [54] which provide good support for explorations of single projections and evaluating clustering, they are not designed to enable and assist with comparative assessment and evaluation of dimensionality reduction methods. This is an important challenge for analysts who have to decide which dimensionality reduction projection best represents their data's structure. Our thesis addresses this by creating a workflow specifically for comparative dimensionality reduction evaluation, with cluster aware explanations and multi-metric summaries, which are optimized to judge dataset "learnability.". Our design addresses the gap noted in recent work by Espadoto et al. that call for techniques like linking projections with high dimensional space, metric quality etc. [22]

2.7 Foundational concepts in Interactive Visualization

Visual exploration plays a key role in understanding high dimensional spaces and the hidden structures present inside them. This thesis makes use of several proven interactivity techniques derived from visual analytics, to make a system tailored to real world, industrial data scenarios. It especially uses ideas from Tamara Munzner's [55] "Visual analysis and design" and how Gleicher et al [50] categorize comparative visualization. This theoretical base makes sure that the design choices that went into the system are not random, but are guided by best practices and aim to solve known problems in the field.

2.7.1 Gleicher et al.'s taxonomy of comparative visualization

This framework identifies three main ways to compare visualization

Juxtaposition

Juxtaposition means putting visualizations side by side for direct comparison. DimCompare view with its two scatterplots which depict two different dimensionality reduction techniques is an example of this strategy. L'Yi et al. [56] revisit comparative layout idioms including juxtaposition, superposition, and explicit encoding. In the BarDar chart, two views help the users compare metrics for different dimensionality reduction methods, one view depicts a radar chart and another shows a bar chart representing the aggregated scores from the radar chart.

Superposition

This means layering different views on top of each other, while not the main method in DimCompare, density contours can be seen as a type of superposition over the scatterplot data points. In the BarDar chart, superposition is used to overlay radar charts of different dimensionality reduction techniques on top of each other.

Explicit encoding

This method involves showing the differences or extra information with special visual elements. Important data is explicitly made known to the user to reduce the cognitive load. The cluster glyphs in DimCompare are an example of explicit encoding; they show high dimensional feature differences. They represent what makes a cluster different in original high dimensional space, bridging the gap between 2D projection and basic feature characteristics. In the BarDar chart, the shapes of radar charts, corresponding to different metrics are explicitly encoded for the user, as a bar chart, since visual calculation of areas of irregular shapes is very challenging.

2.7.2 Munzner's principles

Our system design, both directly and indirectly follows several of Munzner's [55] key ideas for good visualization.

Visual encoding

The design of the system carefully thinks about how visual elements like points, colors, glyphs and bars are used to show data properties and relationships effectively, and following good visualization practices. For example color is always used to show which cluster a point belongs to, transparency is used to manage the density of data, different colour pallets are used for separate visualisations.

Interaction techniques

Munzner stresses how important interaction is for exploring data. Our system includes basic methods like selecting, linking and moving around (panning, zooming etc). In DimCompare, users can use brushing and linking to follow outliers across different projections. Tooltips are also made visible in 2D projection, when the cursor is hovered over a datapoint, which shows original high dimension features.

Overview First, Zoom/Filter, Details on demand

This idea suggests giving a general overview of the data first, letting users zoom or filter to areas they are interested in, and then providing detailed information when asked for. Our visualization framework design follows this, the BarDar chart gives and overall summary with numbers, but users can then decide to zoom in, explore and apply filters for the same data, using the DimCompare view, which also supports brushing, feature selection etc., and finally users can enable cluster annotations to get detailed information about specific clusters on demand.

Scalability

Dealing with clutter in large datasets is very important, the design choices we made, like limiting the number of bars in cluster annotation, having a density aware mode, ability to toggle visual encodings, zoom aware transparency for cluster annotation to avoid occlusion, scrollable list views

for feature selection for really high dimensional datasets etc. While the current system is made for small to medium size datasets, considerations like these set the stage for future scaling to large datasets.

2.7.3 Schneiderman's Mantra

Schneiderman's [57] information seeking mantra, "overview first, zoom and filter, then details on demand, gives foundational guidance for the design of the interaction system. This is implemented through:

Overview: The Cluster feature contribution glyphs and metric summaries provide immediate understanding of overall data quality and structure

Zoom and Filter: Interactive exploration in DimCompare projections allows users to focus on their regions of interest. Users can also filter the features which will be shown in the Cluster annotations.

Details on Demand: Tooltips, visibility toggles for various visual encodings and detailed statistics provide information when needed, without cluttering the display.

2.7.4 Further Concepts

Comparative Visualization The use of multiple synchronized views is common in information visualization. Gliecher et al. categorizes this as juxtaposition (side by side views), superposition (overlay) and explicit encoding of differences. DimCompare mainly makes use of synchronized juxtaposition to help users compare two projections performed using different dimensionality reduction techniques directly.

Linking and Brushing This is a basic interactivity method in exploring the data, it refers to selecting a data point in one view and highlighting the corresponding point in another view, which enables tracking of clusters and outliers across different dimensionality reduction projections. Users can make a custom cluster selection by brushing over desired points. Unfortunately current tools provide little to no support for users to explore data in this way. [58]

Glyphs and Annotations Using glyphs, which are compact visual summaries over clusters, fills the gap between 2D projections and original feature spaces. They are used to make scatterplots richer by giving more information about individual points or clusters. Glyphs can be used to encode anything, such as data labels of individual data points or even used for clusters to summarize key feature differences driving these clusters.

2.7.5 Human in the loop Dimensionality reduction

A study by Sacha et. al. [59] suggests a model for adding human interaction into the dimensionality reduction process. They found out different situations where users can guide the algorithm, like by choosing the features or tuning parameters etc. Our system fits into this model by letting the users explore dimensionality reduction results by changing features, number of clusters etc and use what they learn to drive future decisions.

2.7.6 Overplotting and density management

A huge problem in high dimensional data scatterplots is overplotting, any high density data visualization is bound to suffer from clutter. Ways to target this include use of transparency, zooming, grouping data and changing to density based views. DimCompare provides the user with the option to show contour plots, for when dealing with high density data. These highlight the high-density areas using a 2D Kernel Density Estimate (KDE) over the points. We also make use of semantic zooming and toggling data point visibility, the cluster annotations automatically start getting more transparent when zooming in to reveal to the viewer, points which were occluded by the annotation.

Combining these information visualization methods, like juxtaposition, brushing, glyphs, and cluttering solutions, is a good way of dealing with challenges that high dimensional data usually presents.This work addresses limitations of previous methods and supports a more interpretable, metric driven exploration of high dimensional feature spaces.

2.8 Existing Visualization Frameworks for Comparing and Interpreting Dimensionality Reduction

Many visualization systems have been built to help users interpret the embeddings, compare dimensionality reduction methods, and evaluate the projection quality. These systems have informed our design choices and also show what gaps remain in real situations.

2.8.1 Embedding interpretation and inspection tools

Tensorflow embedding projector is a widely used tool, which was introduced as a web interface for exploring high dimensional embeddings, using PCA, t-SNE and UMAP, it supported search, selection and neighborhood inspection [53] It was helpful in popularizing interactive projection exploration for practitioners but it provides limited comparative assessment across different dimensionality reduction methods, with no support for metric overlays.

2.8.2 DimReader

DimReader is a visual interaction framework by Cavallo [60] which focuses on explaining projections rather than just displaying them. They introduce forward and backward projection and landmarks which let users see how changes in the features move the point in 2D projection. This work is important because it links low dimension data to high dimension data which is in line with our cluster annotation/glyph idea.

Clustrophile2 is also a tool by Cavallo [Cavallo 2019], which provides guided workflows for interactive clustering analysis, it integrates algorithm selection, parameter steering and visual diagnostics . This system shows how to mix semi automated suggestions with a human judging when choosing cluster models and parameters. Unfortunately it was made to explore single projections and still not for comparative analysis across different dimensionality reduction techniques.

2.7.2 Projection comparison and quality assessment Projection inspector is a tool which provides interactive projection space, where users can move between projection methods and interpolate

new layouts between methods, while also inspecting quality metrics. [Pagliosa et al 2015] It explains projection choice as a trade off and combines layout browsing with metric readings, which was a guide for the comparative ideas presented in the thesis like DimCompare and the metric summary in BarDar.

Distortion focused tools such as CheckViz, ProxiLens and other follow ups by Aupetit and colleagues encode false neighbors (data points that are close in low dimension but far apart in high dimension) and missing neighbors (points that are close in high dimension but far apart in the projection) and local stress directly into the visualization. [Lespinats Aupetit, 2011; Heulot et al., 2013; Aupetit, 2007]. This work argues that users should see where the projections are unusual instead of just reading a single numeric value. The DimCompare tries to follow this philosophy, qualitative and quantitative cues and showing structure in two separate dimensionality reduction views with cluster annotations.

Surveys and frameworks have helped generate and refine the ideas that went into this thesis. Nonato and Aupetits survey links dimensionality reduction techniques, distortions in structures, task and enrichment of layouts, which motivated us in adding local quality overlays and explanations to embedding plots [Nonato Aupetit, 2018]. More recently, a review by Behrisch et al., looked at quality metrics for dimensionality reduction, providing us with guidance on which metrics capture which aspects of structure preservation. [Behrisch et al., 2018] which back our choices to combine visual encodings with multiple metrics rather than rely on a single score.

2.7.3 Human interaction and reliability A structured literature analysis by Sacha et al. looks at how users interact with dimensionality reduction and proposes a process model for human in the loop projection analysis. [Sacha et al 2017] Their findings support the importance of brushing and linking, parameter steering and explanatory views, all of which are incorporated in the design of DimCompare. A newer survey by Jeon et al. argues that reliability remains a central issue and advocates for workflows where users can "see" assumptions, uncertainty and distortions. [Jeon et al 2025]

Existing systems let users browse dimensionality reduction projections, inspect neighborhoods and overlay quality cues, but these methods fall short when it comes to cluster level explanations and side by side dimensionality reduction methods comparison with unified multi-metric summaries. DimCompare fulfills the first gap by making use of floating cluster glyphs and BarDar fills the second gap by providing compact multi metric summaries that reduce mental calculation when comparing different dimensionality reduction methods.

2.7.4 What this means for our work We adopt a side by side dimensionality reduction comparison (DimCompare) rather than a heavy superposition to reduce the clutter and while also keeping the context. We add cluster level feature glyphs to bridge the gap that exists between low dimension patterns and high dimensional explanations, which is a gap in many systems. We integrate multiple quality metrics and summarize them visually (BarDar), this along with explicit encodings reduce the cognitive load and over reliance on any single index. We acknowledge distortions and reliability concerns from prior work in this field and address them using brushing, linked selection and optional density encoding, which help reveal distortions like missing and false neighbors. This boosts interpretability and exploration in visual analysis.

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.

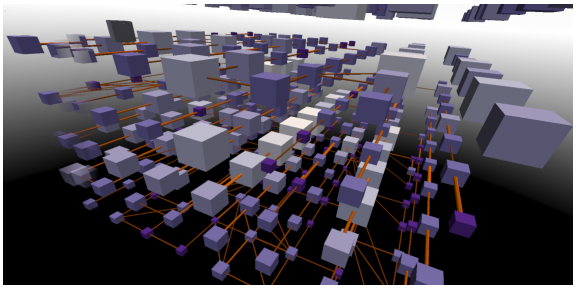

Figure 2: This is a figure with a caption adapted from .

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 [55]. 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]

...

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 4).

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

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 4: 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] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996.

[25] Erich Schubert, Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(19), 2017.

[26] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Proceedings of the 17th Pacific–Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172, 2013.

[27] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[28] Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[29] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. Density-connected subspace clustering for high-dimensional data. In *Proceedings of the 4th SIAM International Conference on Data Mining (SDM)*, pages 246–257, 2004.

[30] Zahid Ansari, M. F. Azeem, Waseem Ahmed, and A. Vinaya Babu. Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. arXiv preprint arXiv:1507.03340, 2015.

[31] M. Gösgens, A. Tikhonov, and L. Prokhorenkova. Systematic analysis of cluster similarity indices. arXiv preprint arXiv:190?, 2019.

[32] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[33] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[34] N. R. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847–857, 1997.

[35] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[36] Joseph C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[37] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

[38] Maria Halkidi and Michalis Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01)*, pages 187–194, San Jose, CA, USA, November 2001. IEEE.

[39] J. P. Morais, M. Adhikari, T. Taha, I. Gemp, and W. G. Macready. Ferm: A feature-space evaluation and representation measure for classification tasks. arXiv preprint arXiv:1909.02699, 2019.

[40] Laurence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[41] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, limits and extensions. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[42] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[43] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486, 2009.

[44] P. C. Mahalanobis. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India, 1936.

[45] A. Bhattacharya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[46] L. Bruzzone and S. B. Serpico. A neural-network approach to the supervised classification of very high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 36(2):558–571, 1998.

[47] Xianmei Zhang, Xiaofeng Lin, Dongjie Fu, Yang Wang, Shaobo Sun, Fei Wang, Cuiping Wang, Zhongyong Xiao, and Yiqiang Shi. Comparison of the applicability of j-m distance feature selection methods for coastal wetland classification. *Remote Sensing*, 2023.

[48] Michael Behrisch, Bruno Bach, Niklas W. Riche, Tobias Schreck, and Jean-Daniel Fekete. Quality metrics for information visualization. *Computer Graphics Forum*, 37(3):625–662, 2018.

[49] Tamara Munzner. A nested model for visualization design and validation. In *IEEE Transactions on Visualization and Computer Graphics*, volume 15, pages 921–928, 2009.

[50] Michael Gleicher, Daniele Albers, Robert Walker, Iftikhar Jusufi, Chris D. Hansen, and James C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[51] Sylvain Lespinats and Michaël Aupetit. Checkviz: Sanity check and topological clues for linear and non–linear mappings. *Computer Graphics Forum*, 30(1):113–125, 2011.

[52] William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[53] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. arXiv preprint arXiv:1611.05469, 2016.

[54] Marco Cavallo and Çağatay Demiralp. Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, 2019.

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

[56] Sehi L'Yi, Jaemin Jo, and Jinwook Seo. Comparative layouts revisited: design space, guidelines, and future directions. arXiv preprint arXiv:2009.00192, 2020.

[57] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[58] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. In *Proceedings of the 2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2018.

[59] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, 2017.

[60] Marco Cavallo and Çağatay Demiralp. Dimreader: Model-agnostic steerable dimensionality reduction for interactive data analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2018.

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	...	...