FS/DS: A Theoretical Framework for the Dual Analysis of Feature Space and Data Space

Frederik L. Dennig, Matthias Miller, Daniel A. Keim, and Mennatallah El-Assady

Abstract—With the surge of data-driven analysis techniques, there is a rising demand for enhancing the exploration of large high-dimensional data by enabling interactions for the joint analysis of features (i.e., dimensions). Such a dual analysis of the feature space and data space is characterized by three components, (1) a view visualizing feature summaries, (2) a view that visualizes the data records, and (3) a bidirectional linking of both plots triggered by human interaction in one of both visualizations, e.g., Linking & Brushing. Dual analysis approaches span many domains, e.g., medicine, crime analysis, and biology. The proposed solutions encapsulate various techniques, such as feature selection or statistical analysis. However, each approach establishes a new definition of dual analysis. To address this gap, we systematically reviewed published dual analysis methods to investigate and formalize the key elements, such as the techniques used to visualize the feature space and data space, as well as the interaction between both spaces. From the information elicited during our review, we propose a unified theoretical framework for dual analysis, encompassing all existing approaches extending the field. We apply our proposed formalization describing the interactions between each component and relate them to the addressed tasks. Additionally, we categorize the existing approaches using our framework and derive future research directions to advance dual analysis by including state-of-the-art visual analysis techniques to improve data exploration.

Index Terms—Visual analytics, dual analysis, feature space, data space, feature exploration, mixed data, high-dimensional data

1 Introduction

One of the major challenges faced by data analysts when exploring and analyzing collected data is the detection of interesting patterns and relationships among data items and features (i.e., dimensions). This is due to multiple reasons. Firstly, the sheer size of the datasets, and secondly, the complexity of patterns that analysts are facing during the investigation. A popular way to explore large high-dimensional datasets is dual analysis. Dual analysis is a technique first introduced by Turkay et al. [1] for the analysis of DNA microarrays. This first instantiation enabled users to perform correlation exploration and hypothesis generation utilizing interactive visual analysis. Turkay et al.'s approach employed three key components: (1) A view visualizing summaries of features, i.e., scatterplots of summary statistics, (2) a view that visualizes the data points, here, a projection based on Principal Component Analysis (PCA) [2], and (3) a bidirectional linkage of both visualizations, in this case, through Linking & Brushing. With those three components, dual analysis allows for simultaneous visual investigation and manipulation of features and data items (Fig. 1). In recent years, approaches solved problems in other domains, such as medicine [3], [4], [5], [6], [7], [8], crime analysis [9], [10], [11], [12], and finance [12], [13], [14]. Other approaches exchanged the visualizations for feature and data space, e.g., Parallel Coordinate Plots (PCPs) [15], and also used different interaction techniques on these visualizations, such as Drag & Drop interactions [16], [17] or subspace selection [7], [9], [10], [13], [14], [18], which necessitates adaptation of the linkage between features and data space. Implementations using the dual analysis paradigm are mainly geared toward specific use cases, while only some are designed for multiple domains.

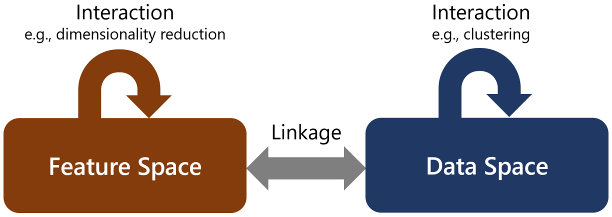


Fig. 1. Dual analysis leverages the interactions on the feature space and data space by linking the visualizations of both spaces. Both spaces are tightly coupled, allowing for joint analysis with an immediate response.

The strength of dual analysis is that the link between the feature and data space visualization allows for an immediate response, which in turn allows for a fast hypothesis generation and validation, ultimately enhancing the knowledge generation process [19]. The visualization of feature and data space symmetrically leverages the preference of humans for symmetry [20]. Since the available approaches are domain-specific, tackling a specific problem, transferring these approaches to solve new problems in other domains is non-trivial. Additionally, many previous works popularized dual analysis for multivariate data analysis, where the data items and attributes are simultaneously shown in two adjacent and symmetric views [1], [17], [21], e.g., two scatterplots using the same dimensionality reduction technique and interaction for feature and data space. These approaches only focus on detecting similarities among data items and fea-

- Frederik L. Dennig and Matthias Miller are with University of Konstanz, Germany. E-Mail: first.last@uni-konstanz.de.
- Daniel A. Keim is with University of Konstanz, Germany. E-Mail: keim@uni-konstanz.de.
- Mennatallah El-Assady is with AI Center, ETH Zürich, Switzerland. E-Mail: melassady@ai.ethz.ch.

tures, or analyzing the impact of a feature on the topology of the dataset. Approaches that do not employ a symmetric design are more flexible. However, the linkage of both visualizations is less straightforward, since both views have other benefits and limitations. Additionally, the number of conceivable combinations is vast. Thus, we provide a formal model that can help structure the development of new dual analysis approaches. Generally, dual analysis approaches lack the capabilities of visual analytics frameworks that employ more sophisticated techniques. For example, machine learning tools, such as interesting subspace recommendation [22] and feature selection algorithms [23], [24], layout enrichment for scatterplots [25], analytical provenance [26], and guidance mechanisms [27]. We argue that the introduction of those techniques into the dual analysis framework to explore, reduce, and transform the data will improve its usefulness since these techniques already improve other visual analytics frameworks. However, the addition of those algorithms is challenging since dual analysis depends on a meaningful interplay between the feature and data space visualizations. Thus, interfaces enabling the integration of machine-learning techniques need to be well-defined.

A comprehensive overview of existing dual analysis approaches is missing in the current literature. Thus, we performed a systematic literature review to get a comprehensive and well-grounded understanding of the area. We present seven scenarios describing ways of applying the dual analysis approaches in addition to their fundamental properties, goals, and use cases, including which techniques have been used to create meaningful feature and data space visualizations and interactions. One challenge faced for future applications is that the state of the feature and data space view need to stay coherent, even with more complex and sophisticated algorithms and interactions. Thus, our FS/DS model presents a unified framework incorporating previously disjunct approaches for dual analysis. Our key contributions include the following:

- A *systematic literature review* describing fundamental properties, goals, and use cases of existing dual analysis approaches.
- A *theoretical model for dual analysis* describing the key components, yielding a formal description of the design space for dual analysis approaches.
- *Validation* of our formal framework through *descriptive and generative use*.

Our contributions enable researchers and developers to include additional analytical capabilities, such as machine learning algorithms and visualization techniques. Finally, we discuss the limitations of our work and present promising research directions.

2 Related Work

This work is related to previous publications in several ways: It is concerned with general theoretical models for visual analytics, specifically proposing one for dual analysis, and it is related to interaction and task taxonomies. Thus, we will cover how they relate to dual analysis and what they are lacking regarding dual analysis interactions. We will briefly describe how our proposed framework will address these shortcomings.

2.1 Theoretical Models in Visual Analytics

Before proposing a formal and theoretical framework for dual analysis, we relate to formal and theoretical models in Visual Analytics (VA) and information visualization.

Jarke J. van Wijk proposed a formal model for visualization [29], which models visualization as a function of data and its specification. The specification can be changed by the user based on the knowledge gained after the perception of the visualization through an exploration process. These interactions are represented as processes or functions (i.e., visualization, perception, and exploration), while the data, the visualization, and its specification are denoted as parameters for the processes. This model was adapted by Green et al. [28], [30] adding interaction between the perception and exploration, as well as the exploration and the users' knowledge. This update highlights that perception directly impacts exploration, and knowledge is also gained through exploration and interaction.

Another high-level model for general VA approaches was published by Keim et al. [31]. It describes the visual analytics process as characterized via interactions between data, visualizations, models about data, and the user to discover knowledge. It defines VA as a combination of automatic and visual analysis techniques with a tight coupling through human interactions, with the primary goal of gaining new insights from data. Thus, the first step in the model is to transform the data to derive different representations for subsequent exploration through automatic or visual analysis. This model makes a clear distinction between automatic and visual analysis and keeps them separated. Also, all transformations are framed as preprocessing. The model describes automated analysis as data mining methods that are used to create models of the data. With these models, the analyst can evaluate and refine the model by interacting with the data through visualization. Visualizations can also allow analysts to parameterize automatic methods. Model visualizations are described as tools for the evaluation of the model itself and the validation of the generated findings. The interplay of automatic and visual techniques is a hallmark of VA. Thus, this model allows for the continuous refinement and adaption of hypotheses.

An extension of this model is the Knowledge Generation Model by Sacha et al. [19]. It takes the model by Keim et al. and extends it with three loops, namely exploration, verification, and knowledge generation. This model places these three loops in the domain of the users, while the model by Keim et al. represents the computation domain. The exploration loop is described with two steps: Action and finding. The verification loop with hypothesis and insight. Most importantly, it describes these steps as nested, e.g., a finding can lead to new insights, which can help create a new hypothesis, which can be tested through an action using a VA approach. Finally, through the exploration and verification of the action, the user can gain new knowledge about the data by verifying the explored hypothesis through multiple perspectives and insights. Thus, the model by Sacha et al. focuses on the user rather than the algorithmic or computer side.

Our work contributes a theoretical and formal framework for the dual analysis of feature and data space. One of the benefits of formalization is the systematization of core operations on the data while describing what tasks are achievable or not with which techniques, such as visualization and interaction techniques. Thus, it provides a more detailed model by focusing on specific properties of dual analysis and is designed explicitly to abstract key properties. Yet, it remains at a high level such that we present our contribution in a way that corresponds to these existing models focusing on the core operations.

2.2 Interaction Techniques and Taxonomies

Dual analysis approaches leverage interaction techniques to enhance opportunities to extract relevant information from the visual representation of the feature and data space. Various taxonomies and generic frameworks explore the design space of visual interaction.

Yi et al. [32] present a framework and taxonomy for information visualization interaction techniques, which categorizes lower-level interactions into seven groups, namely (1) Select: mark something as interesting, (2) Explore: show something else, (3) Reconfigure: show a different arrangement, (4) Encode: show a different representation, (5) Abstract/Elaborate: show more or less detail, (6) Filter: show something conditionally, and (7) Connect: show related items. These categories are focused on the user intent rather than the users' low-level actions. For instance, Lekschas et al. [33] introduced the technique "Interactive Piling" to facilitate the visual organization, exploration, and comparison of numerous small multiple using the pile metaphor to provide visual aggregations.

The taxonomy by Brehmer and Munzner [34] extends the ideas by Yi et al. [32]. It describes a multi-level typology for information visualization tasks. The authors specifically differentiate the ends (i.e., user intent) from the means (i.e., user action), with the primary goal of describing why and how a task is performed. Additionally, Brehmer and Munzner address the inputs and outputs of a given task to create a comprehensive taxonomy. It allows for the expression of complex tasks as sequences of simple, interdependent tasks. All intents, interactions, inputs, and outputs are described in an abstract rather than a domain-specific way, allowing for an application of the taxonomy to a large set of VA systems. Nonato and Aupetit [25] applied the taxonomy by Brehmer and Munzner [34] to dimensionality reduction, formalizing tasks specific for dimensionality reduction.

Landesberger et al. [35] present a new taxonomy for user interaction in VA applications by comparing existing interaction taxonomies. This approach covers three high-level areas, i.e., visualization, reasoning, and data processing. Each area consists of two subcategories, i.e., of data changes and changes in the respective representation. In this taxonomy, changes in the data impact the visualized dataset, and changes in visualizations refer to different forms of interaction. Changes in the dataset are categorized into two subcategories. The first reflects changes that impact the data selection, such as filtering, while the second comprises changes that affect the dataset, such as editing or annotation. The visualization changes are subdivided into changes in the visualization parameters and changes in visualization type or scheme, as described by Bertini et al. [36].

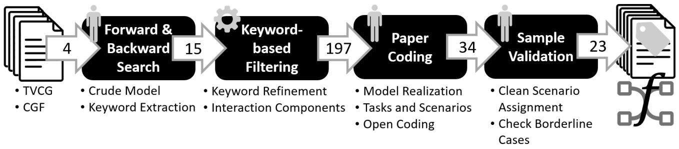


Fig. 2. Paper Selection Process: 1.) Landmark papers 2.) Forward and Backward Search 3.) Automated Keyword-based Filtering, 4.) Paper Filtering, and 5.) Sample Validation. The numbers in the arrows describe the number of papers retained after each step.

Endert et al. [37] specifically focus on the semantic interaction, introducing a visual analytics prototype called FORCESPIRE designed to support diverse forms of semantic interaction. They propose a new design space for interaction in visual analytics, enabling analysts to interact with a visual metaphor leveraging interactions derived from the analytic process, such as searching, repositioning, or highlighting.

Dimara and Perin [38] published a paper about the general concept of interaction for data visualization providing a clear definition that helps to improve understanding of the opportunities that interaction opens to users. Their evaluation identified several crucial factors, such as the computer being a mediator between humans and data, the visualization should invite users to construct a mental model of data concepts, and there can be different intents of why visualization is used at play. Thus, they argue that interaction allows for iterative steps to approach an analysis goal by supporting user intentions while maintaining a high level of flexibility in an application.

Our framework for dual analysis covers interaction in its design by linking them to common analysis scenarios, which internally are connected to a step in the data processing pipeline. Thus, it provides a detailed description of possible interactions linking them to the underlying components facilitating dual analysis.

3 LITERATURE SURVEY

At the outset of this literature review, we present our definition of dual analysis that we use throughout this work.

Defintion of Dual Analysis: Dual analysis facilitates the joint visual analysis of feature and data space through (1) a view visualizing the features (i.e., feature space), (2) a view that shows data points (i.e., data space), and (3) a mechanism to link both views in a *bidirectional* way, meaning that the interaction with one visualization, e.g., the features space, changes the other visualization, i.e., the data space (see Fig. 1). The linkage mechanism can be symmetric, but this is not a requirement.

To present an overview of dual analysis approaches, we performed a systematic literature review. The general process is described in Fig. 2. First, we manually identified a small subset of four publications [1], [13], [17], [21] from the TVCG and Eurographics journals, which we use as landmark papers. From these publications, we found other relevant publications based on a forward- and backward search following the citations (see Sec. 3.2). Then, we performed a detailed qualitative analysis of the selected papers, extracting and refining dual analysis characteristics, yielding a set of keywords (see Sec. 3.3). Finally, to ensure our understanding of dual analysis is comprehensive, we executed a keyword-based search for publications that were not found by following the citation of the landmarks forwards and backward (see Sec. 3.2). In general, we follow a methodology described by Snyder [39] as a systematic review to create a theoretical model or framework.

3.1 Landmark Papers

Before making a contribution towards the topic of dual analysis, i.e., a formal model of the dual analysis paradigm, we started with a few landmark papers that were foundational for this technique (see papers marked with * in Tab. 1), for the primary goal of identifying existing dual analysis approaches implemented by the VA and visualization community. These publications are: The first approach by Turkay et al. [1]. I^F, F^I-Tables [21], SIRUS [17], and the Dimensions Projection Matrix/Tree [13]. We chose these publications since they are referenced by other publications in Tab. 1 and were published in journals with high visibility, more specifically, TVCG and CGF. We also verified later whether they are referenced by other publications in Tab. 1. Turkay et al.'s publications [1], [3], [8], [14], [40], [41] can be viewed as fundamental to dual analysis, as they introduced the concept and established the foundation for this approach.

3.2 Forward and Backward Search

We initiated a forward and backward search of reviewed publications to provide an extensive overview of the existing dual analysis approaches. We reviewed literature citing one of the landmark papers, as well as literature that is cited by landmark papers. This process yields 15 papers (see Fig. 2) from the IEEE, Eurographics, ACM digital libraries, as well as from Elsevier and other literature (i.e., Information Visualization, and The Visual Computer). However, since we also found dual analysis approaches outside the citations of and from the landmark papers, we decided to extend our search range by performing an automated keyword-based search.

3.3 Automated Keyword-based Filtering

From the set of papers that resulted from the forward and backward search, we created a list of relevant keywords by extracting key terms in the papers referencing dual analysis or its components. We combined and cross-referenced the terms to ensure that we did not overlook any relevant terms in the field. We selected the keywords: *Dual analysis*, *dual-analysis*, *dual visual analysis*, *dual-visual analysis*, *dual views*, *dual space*, *dual projections*, *dual scatterplots*, *feature space*, *dimension space*, *dimensions space*, *data space*, *item space*, and *items space*. We used these keywords for our subsequent automated search.

To gain an overview of approaches incorporating dual analysis and also related approaches, we scanned all the available literature (see Fig. 2). We utilized a plain text scanner to accomplish this task, which extracted the plain text from each publication and verified the presence of a given keyword within the paper. The program also generates a frequency count with which single or multiple keywords appear, which provides us with an indication of their relevance. We adjusted our chosen keywords to guarantee that they included all approaches that could be considered dual analysis without any accidental exclusions. We verified that all publications of the previous step also appeared in the result of the automatic keyword-based filtering. This fully automatic scanning resulted in 197 papers (see Fig. 2). We encountered a limitation where the final list of keywords also yielded matches with numerous publications that did not pertain to a dual analysis approach. However, we continued to screen this resulting set of publications.

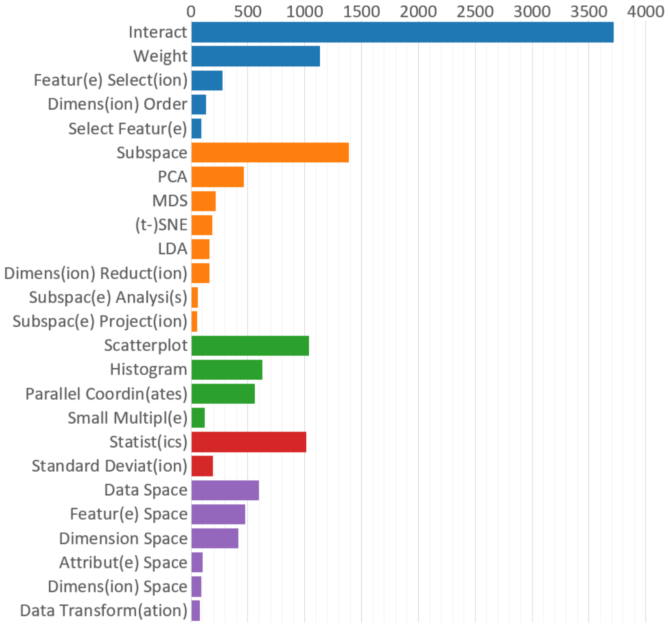


Fig. 3. The top 25 concepts we extracted from the 197 automatically selected papers (see Sec. 3.3). Five colors contextualize each concept: ● Interaction, ● Dimensionality reduction (DR), ● Visualization, ● Statistics, ● Analysis space.

Additionally, from these papers, we extracted core concepts to gain an overview of the used visualizations, techniques, and interactions by stemming all text from all the previously extracted plain text using CoreNLP. Fig. 3 shows the top 25 concepts (i.e., word stems) we extracted from the 197 publications. It shows the number of occurrences on the x-axis. We grouped the concepts into five thematically related groups. This overview helped us create our categorizations and formal framework by highlighting essential topics, such as subspace analysis.

3.4 Paper Coding

We checked the resulting 197 papers manually using the following criteria. Since this is a rather large set to prune, we had to define clear exclusion criteria. First, we checked the paper type. We excluded theory and evaluation papers and papers covering unrelated or tangential areas, such as rendering techniques or physical flow visualizations. Through this filtering, we focus on application or technique papers that analyze high-dimensional data in a domain-specific context. Meaning that these techniques can be applied in very distinct domains.

Second, we checked whether the paper addresses the core components of dual analysis, i.e., a view visualizing summaries of features, a view that visualizes the data records, and linkage of both plots, e.g., through Linking & Brushing. For example, the IXVC pipeline [42] presents an interesting technique for explaining the link between clusters present in lower-dimensional space and the original high-dimensional space with a decision tree missing a dedicated view for the feature space. Based on this, we obtained a candidate set of 34 relevant papers, which we subsequently surveyed in detail.

We open-coded the relevant aspects of the components described in each paper, orienting ourselves along the three key components and their interactions. For each paper, we extracted a brief description of the feature space visualization, data space visualization, feature space transformation, data space transformation, interactions between feature and data space, user tasks, and application domains. Additionally, we iteratively refined the criteria and definition for dual analysis approaches. The general model Fig. 1 and the three key components of dual analysis served as initial criteria to encode which parts are affected by the analysts' feedback.

However, we had to adapt and refine the definition several times. During our study, we discarded several aspects we initially deemed interesting. For example, we classified whether an expert or novice uses a system. Most systems are geared toward domain experts. Thus, we removed this categorization from the review and our model. As a result, we arrived at seven scenarios for dual analysis or, encoding "how the dual analysis approaches can be interacted with" (see Sec. 4.2). We include the used visualizations, the underlying transformations, and the interaction with the components. We describe transformations in the context of lossy and lossless operations describing whether the information is lost during the transformation step.

3.5 Sample Validation

In this final step, we targeted a more fine-grained analysis of edge cases and removed 11 samples, in this case, publications that did not match our definition of dual analysis in Sec. 3. The general reason for their removal was the lack of a *bidirectional* linkage, which is an integral part of our definition of dual analysis. The technique by Zhang et al. [43] presents a feature space visualization but is not linking it with the data space. The approach by Wei et al. [44] allows interaction with a view representing cluster prototypes of particle trajectory. However, there is no second interactive view described. Approaches enabling users to design a transfer function for volume rendering frequently visualize the features space [45], [46]. However, there is no description of direct interaction with the feature or data space visualization. We also exclude approaches that show dimensionality-reduced views of the data alongside other representations [47], [48], [49], [50], since both views constitute a data space visualization. Our final set consists of 23 relevant publications, which we present in Tab. 1. We transformed the table into a set of feature vectors to present similarities (see Fig. 4). We cleaned the encoding and grouped the identified approaches into high-level scenarios (see Sec. 4.2). Finally, we applied our formalization to the approaches to see whether we could describe each with our model, which we provide in the supplementary material.

4 Existing Dual Analysis Approaches

This section covers all dual analysis approaches, which we selected following the definition and criteria we described in Sec. 3. All approaches are listed in Tab. 1, categorizing each approach according to the key components. Thus, each approach is characterized by a feature space visualization and a feature space transformation. Symmetrically, the data space has a visualization and associated data space transformation. The feature and data space transformations are categorized into lossy and lossless representations to reflect that some transformations, such as multidimensional projections, are inherently lossy and cannot be inverted [25]. We proposed seven descriptive scenarios in Sec. 4.2 to categorize different tasks for dual analysis structure along the three questions *Why*, *What*, and *How* proposed by Brehmer and Munzner [34]. Our descriptive scenarios describe goals and tasks addressed by dual analysis approaches similar to those described by Sacha et al.'s literature review on visual interaction for dimensionality reduction [55]. We also list the evaluation and application domain to give an overview of the addressed areas.

4.1 Visualizations and Transformations

We categorize all dual analysis approaches by their individual representations of feature space and data space (see Tab. 1). These representations are formed by a visualization type and a transformation method. However, these techniques do not need to be identical for both spaces.

Feature Space Visualizations: By far, the most common technique to visualize the feature space is *scatterplots* SP, which are used in ten approaches for representing the feature space [1], [3], [4], [7], [8], [10], [13], [17], [41], [52]. Most approaches encode information by using the visual variables color and size [56] in their glyph representations. However, this encoding is limited. The glyphs visualize only one or two attributes, e.g., feature weight, relevance, and category. The position of a glyph often describes the result of a DR method, particularly MDS [57], while some approaches encode statistical properties of the features. Scatterplots are most often used in a symmetric configuration, where the data space is also visualized with a scatterplot.

Small multiples SM are also used more than one time [9], [11], [40]. The features are visualized with a heatmap (i.e., feature thumbnail), where the color of a pixel represents the feature values of data items. An alternative is line charts representing the feature values. The small multiples are ordered by feature weight and feature relevance.

Other visualizations and representation techniques are also used. *Parallel coordinates plots* PCP [15] visualize data by plotting a polyline crossing parallel coordinate axis [6], [18]. Zanabria et al. [51] use *Star Coordinates* SC [58] to visualize features. Corput et al. [21] use a *Data Table* DT to show the feature and data space. *Line Graphs* LG visualize the data by connecting individual points in a plot [16]. A *Graph* GRA visualizes a network with a node link-diagram. Itoh et al. [5] visualize dimensions and their relations using

TABLE 1

We present 23 approaches for dual analysis. We show the feature and data space visualizations: SP *Scatterplot*, PCP *Parallel Coordinate Plot*, SM *Small Multiples*, Map *geographical Map*, DT *Data Table*, SC *Star Coordinates*, LG *Line Graph*, GR *dimension Graph*, HG *Histogram*, PIX *Pixel visualization*. We also show the feature and data space transformations grouped into lossy and lossless transformations: MDS *Multidimensional Scaling*, t-SNE *t-distributed Stochastic Neighbor Embedding*, PCA *Principal Component Analysis*, IDMAP *Interactive Document Map*, RAD *RadViz*, iStar, a technique embedding data using Star coordinates, FAMD *Factor Analysis of Mixed Data*, μ, σ arithmetical mean and standard deviation, the differences relative to them $\Delta_\mu, \Delta_\sigma$, and Clu for clustering, which are lossy methods. Lossless methods are: Id where no transformation is applied, Ord where the order of entities in a view is changed. Sel where the active entities are selected manually. A number N shows the number of distinct measures or methods. We describe all visualizations and transformations on Sec. 4.1. An approach addresses one or multiple scenarios: S1 *Feature Selection*, S2 *Feature Aggregation and Weighting*, S3 *Statistical Analysis*, S4 *Subspace Cluster Analysis*, S5 *Similarity Search*, S6 *Data Aggregation and Weighting*, and S7 *Data Selection*. We describe each scenario in Sec. 4.2. Finally, we show the application or evaluation domain of a given approach: 🩺 *Medicine*, 🦠 *Biology*, 🧬 *Genomics*, 👮 *Crime Analysis*, 👥 *Social Domain*, 🍎 *Nutrition*, 💱 *Financial*, ⚛ *Physics and Chemistry*, ⚙ *Engineering*, 🏃 *Sports*, and 🎵 *Musicology*.

		Feature Space					Data Space						Scenarios							
		Visual.			Transf.		Visual.				Transf.									
Paper	Year	SP	SM	Other	Lossy	Lossless	SP	PCP	Map	Other	Lossy	Lossless	S1	S2	S3	S4	S5	S6	S7	Domain
[1] Turkay et al. (1)*	2011	●			μ, σ		●				PCA				●					🧬
[3] Turkay et al. (2)	2012	●			7		●				MDS		●	●				●		🩺
[13] Yuan et al.*	2013	●			MDS		●				MDS						●			⚛ 🍎
[18] Fernstad et al.	2013			PCP	5		●	●			PCA				●				●	🦠
[40] Turkay et al. (3)	2014		●			Ord			●			Id			●					👥
[8] Turkay et al. (4)	2014	●			6		●				$\Delta_\mu, \Delta_\sigma$				●		●			🩺 🧬
[9] Krause et al.	2016		●			Ord	●	●			PCA		●	●					●	👮 🍎
[21] Corput et al.*	2016			DT		Ord				DT		Ord				●		●		👥
[51] Zanabria et al.	2016			SC		Ord	●				iStar			●				●		🏃
[16] Self et al.	2016			LG		Ord	●				MDS					●				🦠
[5] Itoh et al.	2017			GR	MDS			●				Ord	●						●	🩺 ⚙
[41] Turkay et al. (5)	2017	●			MDS				●			Id			●		●			👥
[14] Turkay et al. (6)	2017			HG		Sel	●		●		PCA				●		●			🦠 💱
[10] Jentner et al.	2018	●			MDS		●				3		●					●		👮
[4] Rauber et al.	2019	●			MDS		●				t-SNE		●			●			●	🩺 🦠
[17] Dowling et al.*	2019	●			MDS		●				MDS			●				●		👮 🦠
[52] Artur and Minghim	2019	●			RAD		●				RAD			●				●		🩺
[53] Zhao et al.	2019			HG		Ord	●					Sel	●	●					●	🦠
[11] Fujiwara et al.	2020			PIX		Id	●				PCA		●					●		👮 🍎
[12] Soriano-Vargas et al.	2020		●		Clu		●				IDMAP				●	●		●		👮 ⚛ 💱
[6] Garrison et al.	2021			PCP	FAMD			●				Sel	●					●		🩺
[7] Müller et al.	2021	●			μ, σ			●				Sel			●					🩺
[54] Miller et al.	2022			PIX		Ord	●				MDS		●		●			●		🎵
	Sum	10	3	10	13	10	17	5	3	1	16	7	9	6	9	4	4	10	5	

a graph. *Histograms* HG [14], [53] are used to display statistical analysis results [14] and results of features selection. Miller et al. [54] use a *Pixel visualization* PIX [59] to display feature values in a matrix configuration.

Feature Space Transformations: We distinguish between lossy and lossless transformations. In contrast to lossless transformations, lossy transformations aggregate and reduce that data such that original values are lost. The most common lossy methods used are dimensionality reduction (DR) techniques. Seven approaches [4], [10], [13], [17], [18], [41] use the *Multidimensional Scaling* MDS technique [57], or derivatives thereof, to create a 2-dimensional projection of the feature space. The well-known combination of visualizing the result of DR with scatterplots is used six times as described for feature space visualizations. The main purpose of dimensionality reduction in dual analysis is to create a two-dimensional representation of the data that can be displayed in a single scatterplot. MDS offers projections where high-dimensional distances are projected into lower-dimensional spaces while trying to preserve global distance relations [60]. For the feature space, this is often a measure of correlation [17], [18]. A particular case is WMDS, which allows the weighting of individual features and the estimation of the weight of features according to their position in the reduced space [17]. *RadViz* RAD [61] offers an alternative approach through a radial layout that presents features as points, i.e., dimensional anchors, which are distributed equally around the perimeter of a circle [52]. The data items are placed according to the influence of each dimensional anchor. For the feature space, the distance is defined as the correlation between pairs of features.

The second most common lossy method is the usage of statistical measures, which represent feature summaries as on the axes of a scatterplot. Values are the *mean* and *standard deviation* μ, σ [1] of all values of a feature. The approaches by Garrison et al. [6] and Müller et al. [7] deal with mixed data and, thus, employ statistical measures for categorical data, like *factor analysis for mixed data* FAMD [62] and the *coefficient of unalikeability* and a definition for standard deviation thereof μ, σ [7]. Three approaches use more than five values, i.e., *mean*, *median*, *standard deviation*, *variance*, *skewness*, and *kurtosis* N in Tab. 1 shows the number of measures) [3], [8], [18]. The approach by Sariano-Vargas et al. [12] uses *clustering* Clu to transform the feature space by aggregating features using the K-means or X-means algorithm, which are also lossy after the aggregation of clusters into prototypes, i.e., centroids.

We also found lossless ways of structuring the feature space, such as domain-specific *orderings* Ord to order features based on a summary in a row or column [9], [16], [21], [40], [51], [53], [54]. No reduction or change to the data is marked as *identity* Id [11], e.g., for a *Data Table* DT and *Parallel Coordinate Plot* PCP all feature values of a data item are displayed. One approach allows for manual *selection* Sel of the visualized features [41], which reflects the user's selection interaction directly.

Data Space Visualizations: Similar to the feature space visualization, the most used technique to visualize the data space are *Scatterplots* SP. A total of 17 publications use scatterplots for the data space and combine them with dimensionality reduction (DR) techniques [1], [3], [4], [9], [10], [11], [12], [13], [14], [16], [17], [18], [21], [51], [52], [53], [54]. Another way of visualizing the data space is *Parallel Coordinates Plots* PCP, which are only used in four approaches to represent the data space [5], [9], [18]. PCPs are used as an auxiliary view to show the dataset. The approach by Itho et al. [5] uses PCPs to select subspaces manually. Three approaches by Turkay at al. use a glyph and *geographical Map* MAP [14], [40], [41] combination, which deal with social and census data. In the case of Corput et al. [21], a *Data Table* DT is used.

Data Space Transformations: We categorize all data space transformations into lossy and lossless transformations. All 15 approaches that use scatterplots to visualize the data space also employ lossy dimensionallity reduction techniques. *Principal Component Analysis* PCA [2] is used six approaches [1], [9], [10], [11], [14], [18]. Six approaches [3], [10], [13], [16], [17], [54] use *Multidimensional Scaling* MDS [57]. The *t-distributed Stochastic Neighbor Embedding* t-SNE [63] is employed twice [4], [10], including the approach by Jentner et al., which allows the user to choose between PCA, MDS, and t-SNEdenoted by 3. Another approach for dimensionality reduction is *RadViz* RAD [61], which we already described as a feature space transformation. It is used by Artur and Minghim [52] to create a symmetric dual analysis approach for aggregating features and data items. The iStar [51] embeds data values relative to star coordinate axes offering an alternative to *RadViz*.

Another lossy way of transforming the data space is the use of statistical measures. The approach by Turkay et al. (4) [8] uses statistical methods to transform the data space by using the difference to the mean and standard deviation of a data point $\Delta_\mu, \Delta_\sigma$. This application of statistics is possible because features are homogeneous, like frequency for the genes, words in a text document, or intensity of pixels in an image. The approach by Miller et al. [54] applies a DBSCAN clustering [64] on the projected data items using a lossy operation on top of the already lossy MDS projection.

Similarly to the feature space transformations, the data space can be transformed using lossless methods. The data table and parallel coordinate plots often show all data items. We this represent by the *identity* Id. In this case, it is combined with a *geographical Map* MAP or *RadViz* RAD. It is also possible to *select* Sel the visualized data items, i.e., manually select or to *order* Ord them in rows or columns.

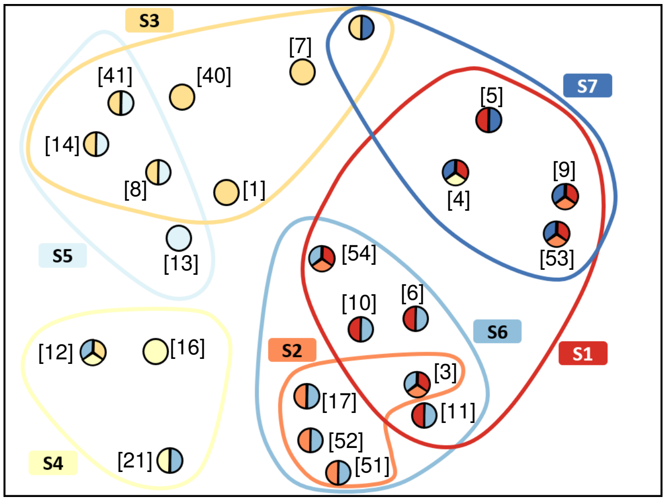


Fig. 4. Similarity-based projection of the 23 papers in Tab. 1. The similarity is defined by one-hot encoding the columns of Tab. 1, excluding name, year, and domain using the Manhattan distance to create an MDS projection. We weight the scenarios three times higher, yielding a scenario-based grouping. Glyphs are colored according to their scenarios (see Sec. 4.2) and grouped showing the relation between them.

4.2 Analysis Scenarios

In this section, we describe the seven scenarios addressed with dual analysis that we found during our literature review. We also assigned each publication in the area of dual analysis to one or more of the identified scenarios (see Tab. 1). These scenarios are also linked to our formal framework (see Fig. 5), where each scenario is addressed by a specific component of the dual analysis workflow. We structured each description along the three main questions, i.e., *Why*, *What*, and *How* by Brehmer and Munzner [34].

S1 **Feature Selection:** The purpose of this scenario is the selection of features for identifying and comparing a set of features relevant to the analyst. In contrast to other scenarios, it is concerned with the original feature values. The primary mechanism for this scenario is to modify the set of active features. The main interaction method is a straightforward selection of the desired features, e.g., through a Lasso Selector. The selected features are then available for further analysis. This scenario never occurs alone since it would only correspond to changes in the data space visualization. A common partner is S7 *Data Selection* [4], [5], [9], [52].

We find this scenario for many different visualization types, as for dual analysis in general, scatterplots are most prevalent. One example is the approach by Jentner et al. [10], where specific features can be selected from a feature space dimensionality reduction-based scatterplot.

S2 **Feature Aggregation and Weighting:** The goal of this scenario is to create different feature summaries. For this purpose, features are aggregated, meaning that a prototype represents groups. Additionally, a feature or feature prototype can be weighted to emphasize or deemphasize it.

There are multiple ways dual analysis approaches create feature aggregations. Most dual analysis approaches make use of dimensionality reduction techniques for the visualization of feature space. For example, Turkay et al. (2) [3] use multidimensional scaling. However, some dual analysis systems allow users to create new features with the primary goal of reducing the number of features of the dataset. This is realized by either combing existing features into a new feature or replacing the original dimensions [18]. This is achieved via the summation of the weighted values or by removing variables that are highly correlated to a representative dimension. In both cases, dual analysis allows for observing the relations of the new features relative to the original dimensions [3]. Dual analysis also supports the creation and validation of classifiers [4]. Generally, dual analysis approaches allow for the creation and subsequent validation of the created features in an iterative loop.

As a secondary way, features can be weighted to give a specific emphasis. The approach by Dowling et al. [17] does this by adjusting the weights of the WMDS for the feature projection. This scenario can require a definition of similarity or dissimilarity for dimensions. The most common way is to define the similarity of features based on a statistical measure (e.g., correlation) [10]. Alternatively, the dimension is condensed to a single numeric statistical value where the difference is meaningful, such as skewness. These measures are adapted to represent distance relations, which can subsequently be used by dimensionality reduction methods to create scatterplot visualizations through projection techniques. Commonly, Drag & Drop interactions change the underlying feature weights [17]. With these interactions, the user can add emphasis to a specific dimension and reduce the impact of dimensions considered less significant. They allow users to observe the effect on the data space, e.g., a change in the general data space patterns.

S3 **Statistical Analysis:** This scenario is focused on different types of statistical analysis. Generally, it allows users to analyze groups of features and data items statistically. For a feature-focused analysis, we found that correlation exploration is the most common type of statistical analysis among all dual analysis approaches. One such approach is the system by Turkay et al. (1) [1]. It has a focus on describing features by their statistical properties, such as the mean and standard deviation. Dual analysis also addresses data-focused statistical analysis, meaning the analysis of data item groups. One such example is the approach by Müller et al. [7], which analyzes variance and attribute variability. In general, this type of analysis focuses on the variance of a subpopulation of the data, with the goal of finding subsets in the data that have either a low variance (i.e., clusters) or high variance (i.e., because of outliers) in their attribute values. The statistical values are used in the feature and data space visualizations, either as a determinant of position (e.g., in a scatterplot) [1], or as a dimension in a PCP.

In terms of interaction, statistical analysis is facilitated by selecting features in the feature space to modify the set of features relevant to the data items in the data space. Similarly, the set of data items is determined through selection by the user determining which values are taken into account for the summary statistics of features.

S4 **Similarity Search:** The goal of this scenario is to find similar features of data items while allowing to change the definition of similarity through parameterization or redefining of similarity functions. A prime example is the approach by Corput et al. [21], which allows for the order-based analysis of features and data items. Generally, dual analysis facilitates similarity search by ordering features and data items or representing dissimilarity as the distance between features or data items [16]. This idea applies to the feature and data space symmetrically.

For this scenario, the selection interaction is most common, either selecting an individual feature or item or a group of both. Through this selection, the definition of similarity is parametrized, yielding updated feature and data space visualizations. More specifically, we find a rerendering of tables, parallel coordinate plots, and scatterplots with updated distance relations.

S5 **Subspace Cluster Analysis:** One main interest of analysts is the detection of subspace structures, e.g., clusters. A subspace cluster is a group of similar data items concerning the subspace dimensions (i.e., features). There are two types of subspaces, axis-parallel subspaces, defined as true subsets of the original data dimensions. In contrast, arbitrarily oriented subspaces are created by freely transforming the data into lower dimensional space, for example, using a dimensionality reduction technique [65]. In this case, the new dimensions are harder to interpret since they can result from a complex transformation (i.e., non-linear projection techniques). Dual analysis supports the interactive user-driven analysis of axis-parallel subspaces and arbitrarily oriented subspaces of linear and non-linear subspaces. For example, the approach by Yuan et al. [13] is purely concerned with the manual analysis of axis-parallel subspaces and subspace clusters. This approach uses MDS to project the analyzed subspaces into 2-dimensional representations, while subspaces are created by selection on the scatter plot or toggled specifically. The approach by Jentner et al. [10] allows for exploring subspace clusters, specifically enabling analysts to understand cluster characteristics, develop alternative clusterings and verify cluster robustness. Turkay et al. (4) [8] visualize statistical properties and enable analysts to select clusters (i.e., groups of data points) and observe their distribution in other subspaces.

In all approaches, selecting subspaces in the feature space visualization plays a key role. The selection of groups and clusters in the data space visualization is less often addressed but needs to be equally covered [13].

S6 **Data Aggregation and Weighting:** Another straightforward scenario is data aggregation and weighting. This scenario describes the data space variant of scenario S2 *Feature Aggregation and Weighting*. This scenario aims to create synthetic and representative group summaries or prototypes of the found groups. Additionally, it is concerned with weighting data items to emphasize or deemphasize them, e.g., for outlier detection and removal.

Since this scenario is linked to scenario S2, the interactions associated with it are identical. Primarily, selection is used to interactively determine groups of data items to aggregate, while the weighting of data items can also be established through Drag & Drop.

S7 **Data Selection:** A basic but essential scenario that is addressed by dual analysis is data selection [4], [9], [40]. This scenario aims to select data items for further analysis. This scenario describes the data space counterpart of scenario S1 *Feature Selection*. This scenario addresses the unconstrained selection of data, as opposed to finding groups and clusters of data items, addressed by S6 *Data Aggregation and Weighting*.

Approaches address this scenario through selection interaction, such as Lasso Selection, in the data space. The only data manipulation process we found in the set of works is labeling data items with a classification algorithm [4]. This technique focuses on the design of classification systems allowing for the observation of feature and data space in dedicated views while allowing for the inspection of different machine learning techniques and their impact on the classification result.

To provide an overview over we also created a similarity-based projection of the 23 papers in Tab. 1 (see Fig. 4). We transformed the entries of Tab. 1 into binary vectors with one-hot encoding the columns and excluded name, year, and domain. We used MDS with the Manhattan distance to create an embedding of the approaches. The glyphs representing each approach are colored according to their scenarios. We can observe the highest overlap between S1 *Feature Selection* and S7 *Data Selection*, as well as S1 *Feature Selection* and S6 *Data Aggregation and Weighting*, due to many approaches allowing the selection of features. Scenario S5 *Subspace Cluster Analysis* always appears with S3 *Statistical Analysis*, except for the approach by Yuan et al. [13]. Also, S4 *Similarity Search* appears to be aspected by the fact that all these approaches use different visualizations for feature and data space compared to the other approaches, mostly using scatterplots.

4.3 Application and Evaluation Domains

Dual analysis has found application in many domains, most notably in *Medicine* (🩺), where we found seven approaches [3], [4], [5], [6], [7], [8], [52], ranging from the analysis of cell abnormalities (e.g., benign or malignant tumor cells) to the results of magnetic resonance imaging (MRI) scans. Next is *Biology* (🦠) [4], [14], [16], [17], [18], [53] and *Genomics* (🧬) [1], [8], where we found seven approaches combined. *Crime Analysis* (👮) with five approaches [9], [10], [11], [12], [17], focuses largely on the analysis of police reports by transforming the data into a high-dimensional feature space. Dual analysis is also applied in the *Social Domain*, (👥) [21], [40], [41] analyzing different aspects of society, such as the comparison of households in different geographic regions. Three publications address the analysis of *Nutrition* (🍎) [9], [11], [13], by analyzing the nutritional contents of food items. Two papers deal with problems in *Finance* (💱) [12], [14]. *Physics and Chemistry* (⚛) [12], [13], *Engineering* (⚙) [5], *Sports* [51] (🏃), and *Musicology* (🎵) [54] are each addressed once.

5 Theory and Formalization

Our formalization encompasses all previous work (see Tab. 1) and offers opportunities for future research directions by revealing new and interesting combinations of methods and analysis scenarios. It serves as a guide for the implementation of dual analysis approaches by formally defining the components and their interactions. Most existing approaches do not include any data manipulations but instead, transform the feature and data space views to reveal patterns through the changed perspective.

Our data model is based on the interpretation of the dataset as one large matrix $\mathcal{D} \in \mathbb{R}^{r \times f}$ where $r \in \mathbb{N}$ is the number of data records (i.e., rows), and $f \in \mathbb{N}$ the number of attributes or features (i.e., columns). This provides a clear distinction between feature and data space and is representative of the two views present in all dual analysis approaches by taking either a column-focused or row-focused perspective. All processing steps that produce additional information (e.g., user interactions or results of a clustering algorithm) can be stored in a data matrix $\mathcal{D}$ as a new column or row. New features, e.g., aggregated and weighted features, are stored as a new column. Symmetrically, a new row is added if synthetic data is created, e.g., a cluster prototype of K-means. Thus, newly created data will also be present in all processing steps of the pipeline. To differentiate functions and operands of the feature and data space, we use the subscript F for the feature space and I for the data space, as this naming is also used by Corput et al. [21]. When referring to a count unrelated to the original dataset matrix, we use $n, m \in \mathbb{N}$. In the following, $M \in \mathbb{R}^{m \times n}$ denotes a matrix with n row and m columns, describing a subselection and aggregation of rows and columns of the dataset matrix $\mathcal{D}$. The matrix M is $\mathcal{D}$ if no selection step exists. Additionally, we use $[\![1..n]\!] \subset \mathbb{N}$ to denote sets of index numbers relative to n, where n is defined in the local context as the number of rows of columns of a matrix.

5.1 Feature and Data Types

Dual analysis has been applied to quantitative and qualitative variables, i.e., mixed data [6], [7]. Thus, our formalization has to describe data analysis for all common features and data types, e.g., numeric and categorical data [66]. To represent each type, the values of column f of the matrix $\mathcal{D}$ denoted by $\mathcal{D}_{*,f} \in \mathbb{R}^r$ are restricted by one of the following definitions to reflect specific properties of feature and data types allowing for the expression of all feature and data types as numeric values.

Binary value: These features are defined by the value set $\{0, 1\}$, reflecting two categories or a binary label. This type is either present in the original dataset or is created through one-hot encoding. This allows for limited analysis with algorithms for numeric data [67].

Discrete values: This data type describes a simple count as values in $\mathbb{N}^0$. Ordinal data dimensions can be converted into this data type by considering their ranked order [67]. This data type is common in social science [14], [41].

Numeric values: This feature type can be divided into two subcategories. Firstly, *bipolar*, which is defined as $[-x, x]$ for $x \in \mathbb{R}^+$. Secondly, *continuous* is simply defined as $\mathbb{R}$ (interval and ratio).

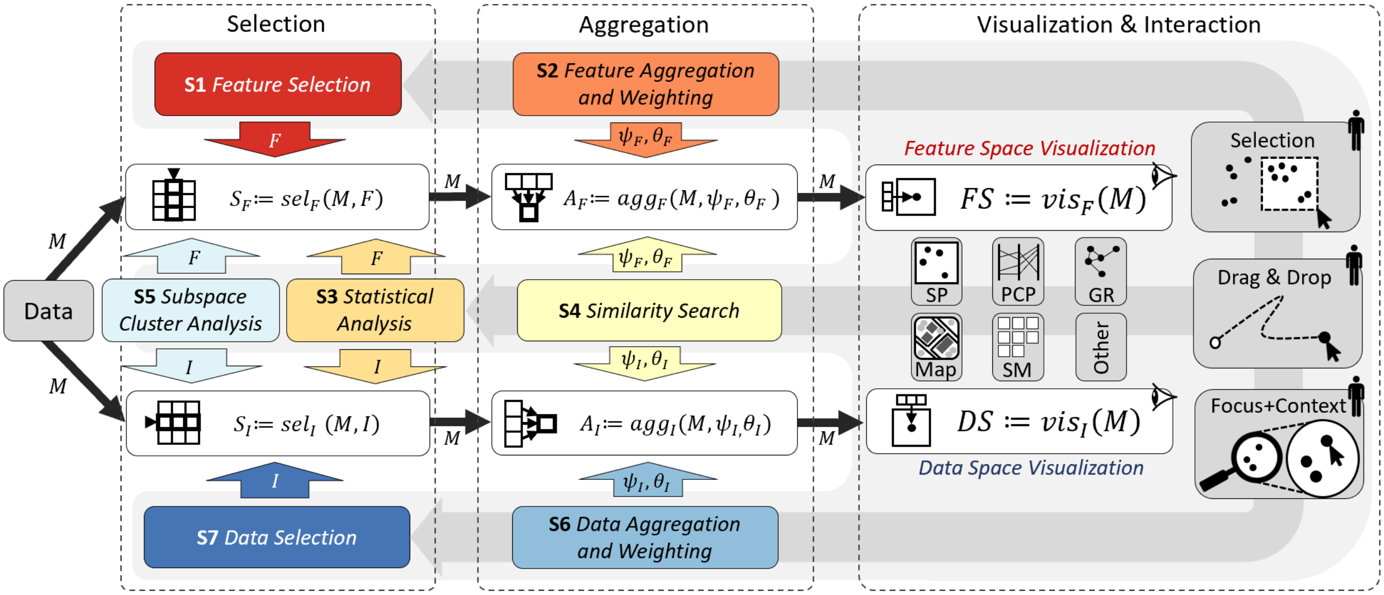


Fig. 5. In our framework for dual analysis, the dataset $\mathcal{D}$, is interpreted as a matrix. The matrix can then be transformed by a selection step, where the data can be reduced. Secondly, the result of this step is used for an aggregation step, which can be used to create representatives. Thirdly and lastly, feature and data space are visualized using distinct but linked visualizations. All three steps take the result of the previous step as input. The scenarios are linked with the different steps of the pipeline by supplying parameterizations to the given operation sel_I (Eq. 7), agg_I (Eq. 8), vis_I (Eq. 9), for data space operations, and sel_F (Eq. 1), agg_F (Eq. 2), vis_F (Eq. 3). The analyst can interact with the visualizations, affecting the previous step and allowing for an immediate response, typical for dual analysis approaches, as described in Sec. 5.4.

Categorical values: This data type can also be represented in two ways. Firstly, *nominal*, which describes a label, and *ordinal*, describing a label with an order. Statistical measures designed for nominal and ordinal data were used in dual analysis [6], [7].

5.2 Feature Space

The *feature space* is a representation of feature or dimensions, i.e., columns of a data table. Features or attributes require different transformations and representations, e.g., showing the distribution of a feature instead of a single value. Even though the formalization of the feature space is symmetric to the data space, the purpose and effect are different by focusing on the columns of the dataset matrix $\mathcal{D}$.

Feature Selection: Many dual analysis approaches allow users to select a subset of features for subsequent analysis. We describe this step in Eq. 1.

$$sel_F : (M, F) \to \mathbb{R}^{r \times |F|} \tag{1}$$

where M is the dataset matrix $\mathcal{D}$ and F is defined as the set of selected features concerning the rows of M. The parameter F is supplied through interactions of the scenarios S1 *Feature Selection*, S3 *Statistical Analysis*, S5 *Subspace Cluster Analysis*.

Feature Aggregation: This step aggregates features items to representatives. Additionally, it allows for the application of an ordering through the definition of the grouping. The aggregation of features supports dimensionality reduction based on the existing features and the calculation of summary statistics. M is the result of the selection step sel_F. To aggregate features, the groups of features are expressed in the tuples of ψ_F with each $e \in \psi_F$ a set of column indices, i.e., features. As for the data space, all existing approaches constrain this step, such that ψ_F is a partition of the of column indices of M. To aggregate groups, we denote the aggregation function with θ_F, which reduces a matrix of selected columns defined by $e \in \psi_F$ by aggregating these columns and reducing the number of rows to d values using dimensionality reduction. Now, ψ_F defines which features to aggregate, and θ_F defines the aggregation and reduction which we formalize in Eq. 2.

$$\begin{gathered} agg_F : (M, \psi_F, \theta_F) \to \mathbb{R}^{d \times |\psi_F|} \\ \text{where } M \in \mathbb{R}^{m \times n}, \psi_F \text{ a partition of } [\![1..n]\!] \\ \text{with } e \in \psi_F \text{ a set of column indices of } M, \\ \text{and } \theta_F : \mathbb{R}^{m \times |e|} \to \mathbb{R}^d \text{ with } e \in \psi_F \text{ and } d \in \mathbb{N} \end{gathered} \tag{2}$$

These sets in ψ_F can be created with a clustering algorithm. For example, k-Means can be used to perform a clustering based on the columns of M. The resulting clusters describe a partitioning of the column indices of M and can be used as ψ_F. Subsequently, the centroids of each cluster could be calculated by defining θ_F as a function that averages all rows of a matrix. To reduce the dimensionality to two dimensions (i.e., $d = 2$), MDS is could be used. However, through the application of agg_F, the original data values are lost. Thus, approaches with an aggregation step are lossy. If a similarity or distance measure is required, e.g., for projection, this is modeled by θ_F.

Techniques combine features by summation and weighting [17], [18]. The parameters ψ_F and θ_F are supplied through interactions of the scenarios S2 *Feature Aggregation and Weighting* and S5 *Similarity Search*.

Feature Visualization: The feature space is visualized using any method that matches the task, as shown in Eq. 3. For example, to detect large groups of features in S5 *Subspace Analysis* Yuan et al. [13] use scatterplots, while for a more fine-grained analysis of relatedness between a few features Garrison et al. use [6] parallel coordinate plots. To describe the visualization of the feature space, we define vis_F in Eq. 3.

$$vis_F : M \rightarrow FS \tag{3}$$

The most frequently used method for visualizing feature space is scatterplots. Therefore, we describe the scatterplot as a combination of a glyph drawing function $glyph_F$ and a function pos_F determining the glyph's position in the plot. For a scatterplot, we have Eq. 4.

$$glyph_F : \mathbb{R}^m \rightarrow G_F \tag{4}$$

One example of G_F is a pixel-based visualization [12]. The position of the glyph is determined in Eq. 5.

$$pos_F : \mathbb{R}^m \rightarrow (x, y) \in \mathbb{R}^2 \tag{5}$$

pos_F usually works by selecting two value form the input vector as x,y-coordinates. Subsequently, we can define the appearance and position for glyphs ρ_i in the feature space scatterplot in Eq. 6, which gives a complete definition of the feature space scatterplot.

$$\begin{gathered} vis_F := \forall i \in [\![1..n]\!]. \\ \rho_i = (glyph_F(M_{*,i}), pos_F(M_{*,i})) \text{ with } M \in \mathbb{R}^{m \times n} \end{gathered} \tag{6}$$

Most approaches that use a scatterplot to visualize the feature space relying on a dimensionality reduction method utilize MDS or variations thereof (see Tab. 1). However, not just dimensionality reduction techniques can be used to determine a position of a feature in the feature space scatterplot. The position of a feature is also determined by statistical properties, such as mean, standard deviation, variance, and skewness, by using them to create scatterplot axes. We do not assign specific scenarios since, for all dual analysis approaches, the visualization type of the feature space does not change during the analysis.

5.3 Data Space

The *data space* represents data items, i.e., rows of a data table $\mathcal{D}$. It focuses on the analysis of individual data items or aggregations thereof. We define the following functions to formalize the processing and relation of steps to create a data space visualization.

Data Selection: Many dual analysis approaches reduce the dataset to a subset of data items. We formalize this mechanic with Eq. 7, yielding a reduced data set or, ultimately, a smaller matrix by reducing the number of rows.

$$sel_I : (M, I) \rightarrow \mathbb{R}^{|I| \times f} \tag{7}$$

where M is the dataset matrix $\mathcal{D}$ and I is defined as the set of selected data items concerning the rows of M. The mechanism for determining the subset of row indices I can be implemented in different ways. A common technique is Linking & Brushing [68]. However, other methods are possible, such as the selection of data items based on class labels, cluster affiliation, filtering, sampling [69], [70] or grouping instances [33], [71]. The parameter I is supplied through interactions of the scenarios S3 *Statistical Analysis*, S5 *Subspace Cluster Analysis*, and S7 *Data Selection*.

Data Aggregation and Weighting: This step aggregates data items to representatives and allows for the application of an ordering through the definition of the grouping ψ_I (see Eq. 8). ψ_I is defined as a tuple of sets with $e \in \psi_I$ describing row indices of the matrix M that are aggregated. θ_I aggregates a selection of rows defined by $e \in \psi_I$ and reduces the dimensionality by reducing the number of columns to d columns. We formalize these functions and operands in Eq. 8.

$$\begin{gathered} agg_I : (M, \psi_I, \theta_I) \rightarrow \mathbb{R}^{|\psi_I| \times d} \\ \text{where } M \in \mathbb{R}^{m \times n}, \psi_I \text{ a partition of } [\![1..m]\!] \\ \text{with } e \in \psi_I \text{ a set of row indices of } M, \\ \text{and } \theta_I : \mathbb{R}^{|e| \times n} \rightarrow \mathbb{R}^d \text{ with } e \in \psi_I \text{ and } d \in \mathbb{N} \end{gathered} \tag{8}$$

For example, to calculate the centroids of clusters, we can apply K-means on the full dataset. K-means is an example algorithm generating ψ_I yielding a partition of the row indices of M with $e \in \psi_I$ corresponding to the data instances assigned to each cluster. The function θ_I can be a method to calculate the centroid of a set. By applying agg_I, information is lost, meaning the original data values are not recoverable. In cases where a similarity or distance measure is used, e.g., for MDS, we express it as a property or parameter of θ_I. Thus, this prototype can represent the dataset or a synthetic data item. Most commonly, ψ_I is a partition of the row indices of M. However, by defining the groups without this constraint, this function can also show the underlying data "as is" after the selection step in the context of their prototype. The parameters ψ_I and θ_I are supplied through interactions of the scenarios S4 *Similarity Search* and S6 *Data Aggregation*.

Data Visualization: Scatterplots are the prevailing data visualization technique in dual analysis. This step involves creating a visual display of data items or aggregations. This is commonly accomplished by utilizing a scatterplot to display a simple glyph, which is then positioned on the screen. Thus, we give it a specific focus in our formalization. However, we also generally address visualizations like parallel coordinate plots and small multiples.

Generally, the visualization DS, is generated from a dataset described as a matrix M. Thus, we define this overarching function in Eq. 9.

$$vis_I : M \rightarrow DS \tag{9}$$

When we deal with scatterplots, we can further specify the generation of the data space visualization by defining how a glyph of the scatterplot will be drawn. Data glyphs can show more information than a simple glyph. [72]. We define a glyph of a scatterplot as a glyph since we do not want to apply unnecessary restrictions on the design of the data point representation (see Eq. 10).

$$glyph_I : \mathbb{R}^n \rightarrow G_I \tag{10}$$

Second, we also define a function to determine the position of the glyph in the scatterplot in Eq. 11.

$$pos_I : \mathbb{R}^n \to (x, y) \in \mathbb{R}^2 \tag{11}$$

Thus, with these two functions, we can cover the scatterplot-based visualization of the data space in Eq. 12, such that the future system can make use of glyphs designed for the given task. The following equation describes the application of these functions to the matrix M by generating a glyph ρ_i for each row and determining the position on the plot.

$$\begin{aligned} vis_I &:= \forall i \in [\![1..n]\!]. \\ \rho_i = (glyph_I(M_{i,*}&, pos_I(M_{i,*})) \text{ with } M \in \mathbb{R}^{m\times n} \end{aligned} \tag{12}$$

To determine a position (see Eq. 11), many approaches employ projection techniques, i.e., dimensionality reduction to two dimensions. We found the following set of commonly used methods in our literature research. They all fit the requirements for Eq. 11. We found that PCA [2], MDS [57], t-SNE [63], or IDMAP [73] are commonly used as a function to determine the position. We refrain from assigning a particular scenario because all dual analysis methods employ a single visualization type for the data space, which remains unchanged throughout the analysis.

5.4 Feature and Data Space Interaction

During our review, we identified *Selection*, *Drag & Drop*, and *Focus+Context* as interaction paradigms of existing dual analysis approaches. We will describe how they facilitate dual analysis by explaining their impact on the feature and data space.

Selection: The most common technique is the selection of data items or features. In general, selection is a common interaction technique [66], [74]. Even techniques that allow for other ways of interaction support this method. Other approaches allow for selecting groups in the feature or data space. Generally, the selection is an interaction component of the feature or data space visualization. Dual analysis approaches realize it through a rectangle or lasso selection on the visualization in scatterplots or axis selection and brushing on parallel coordinate plots [7]. The interaction of feature and data space constitutes a form of Linking & Brushing [1], [13] since selection is used to update feature and data space according to the selection on one view. In our framework, selection parameterizes the sel_F and sel_I functions through their parameters F and I. We refer to selection on one space by the scenarios S1 *Feature Selection* and S7 *Data Selection*. If both parameters are used simultaneously, we enter the realm of S3 *Statistical Analysis* and S5 *Subspace Cluster Analysis*.

Since selection is a very general technique for interaction with dual analysis systems, it also applies to S2 *Feature Aggregation and Weighting*, as well as, S6 *Data Aggregation and Weighting* scenarios. For both scenarios, it determines which features or data items to aggregate. This is expressed by the tuples ψ_F and ψ_I, which hold the selected groups for each space and aggregate them, as formalized by agg_F and agg_I. Thus, we can see that selection is the most applied interaction method in dual analysis.

Drag & Drop: The Drag & Drop interaction is an instance of a direct semantic manipulation [37]. The user modifies the visual-spatial mapping by rearranging elements in the visualization. Drag & Drop is coupled with the weighting of features and data items [16], [17]. Approaches utilizing this interaction modify the underlying definition of similarity. In our framework, we express the similarity of features and data items in the θ_F and θ_I of agg_F and agg_I by parameterizing the dimensionality reduction. Similarly, it can parameterize the ordering implicit in the tuples ψ_F and ψ_I. We refer to the interaction on a single space with the scenarios S2 *Feature Aggregation and Weighting* and S6 *Data Aggregation and Weighting*. If both spaces are used to parameterize agg_F and agg_I simultaneously, users do a S4 *Similarity Search* [21] reflecting the different goals.

Focus+Context: Another concept in the dual analysis is Focus+Context [75]. The analyst can interact with visualization via panning and zooming, allowing for navigation through the visualization. In dual analysis, feature and data space are visualized, and Focus+Context is applicable to both visualizations. The main point is to show a selected region in higher detail (Focus), while preserving the global point of view in a reduced form (Context). Focus+Context predominantly involves a single view, and it does not alter the state of a dual analysis system beyond this scope. Turkay et al. [1] state a modified definition of Focus+Context, which describes a subset of dual analysis fully covered by our selection interactions definition (see above). We state the difference here for the sake of completeness.

6 Evaluation

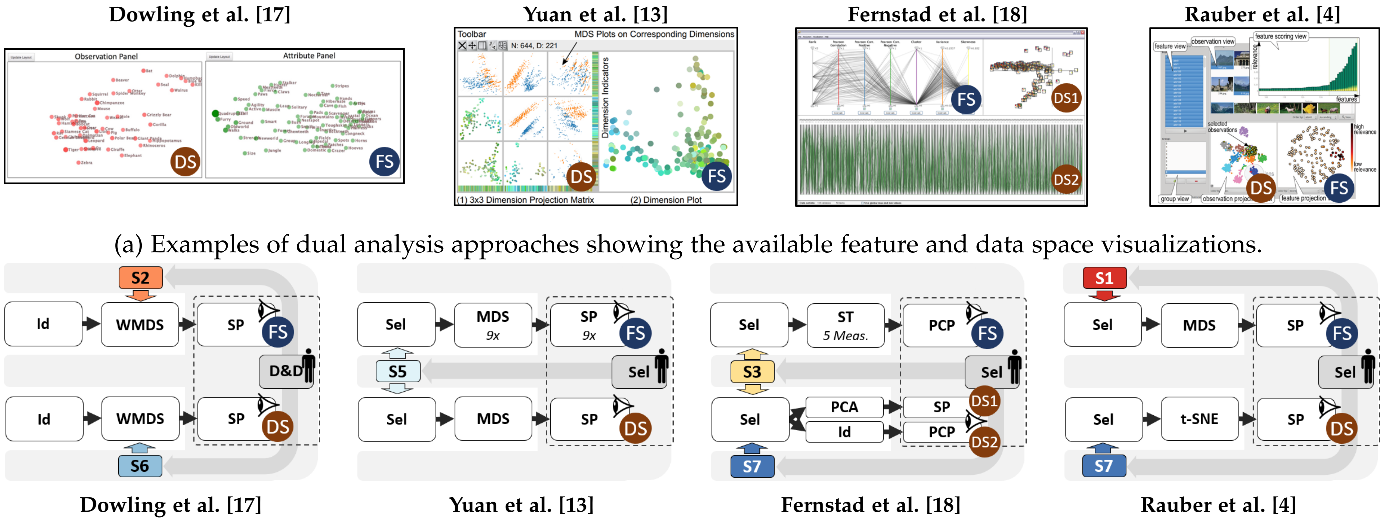
To evaluate our approach, we apply an evaluation strategy inspired by Sacha et al. [55]. We apply our model to existing approaches to show that it offers a consistent method to understand and categorize these systems and analyze their usefulness for the given scenarios (i.e., descriptive use). The presented approaches were either landmark papers or resulted from our literature search and thus also used in the creation process of the model. However, we found the selected four approaches [4], [13], [17], [18] to be representative of the set of papers described in Tab. 1 covering all components of the pipeline. Additionally, we show and discuss gaps that our model revealed that are not addressed in the current research literature (i.e., generative use).

6.1 Descriptive Use — Examples

In this section, we describe four representative approaches.

Dowling et al. The system by Dowling et al. [17] addresses the need for feature and data exploration based on similarity to understand the impact of specific domains on the similarity of data items, as well as the impact of data items on the similarity of features.

Their publication discusses the technique in terms of feature importance. Thus, we assigned S2 *Feature Aggregation and Weighting* as a suitable scenario. Likewise, the paper describes the analysis of data items in terms of finding similar data items after selecting features as less or more important. Here, we also categorize the approach as S4 *Similarity Search*. This approach does not support analyzing feature or data subsets, except the dataset is pre-processed. Our model



(a) Examples of dual analysis approaches showing the available feature and data space visualizations.

(b) The instances of our dual analysis process model for each example depict the different components and scenarios.

Fig. 6. We describe four approaches in Sec. 6.1 to demonstrate that our dual analysis framework applies to existing approaches. We show the application of our model for the systems of Dowling et al. [17], Yuan et al. [13], Fernstad et al. [18], and Rauber et al. [4] as a representative set.

expresses this with the *identity* Id for sel_F and sel_I, since there is no feature or data selection.

A feature and data space *Scatterplot* SP are created using WMDS, which allows weighting dimensions of the projected data but also allows the estimation of the weight once the user alters the scatterplot through drag and drop. The Drag & Drop interactions of the users modify the position of the data and features on their respective scatterplots to modify the perceived similarity to match the user's mental model. This mapping of difference in the perceived distances are realized using WMDS.

The reduction of the vectors describing features and data items to two values is established by the aggregation functions θ_I and θ_F, respectively, which can accommodate dimensionality reduction methods, such as WMDS. The key interaction technique is dragging and dropping of points of feature and data items in the respective scatterplots which parameterizes the functions θ_I and θ_F.

Yuan et al. [13] present an approach for the interactive exploration of subspaces to detect subspace clusters. More generally, the goal of this approach is the detection of interesting structures in subsets of the data. Thus, we assigned the scenario S5 *Subspace Cluster Analysis*. This scenario deals with feature and data item subset selection in a coordinated way. Our framework can express this with sel_F and sel_I, which select feature and data item subsets.

Feature and data space visualizations are visualized using MDS projections with *Scatterplots* SP. In the case of the feature space, this can be multiple views, which are determined interactively by the user through selection on the data space visualization. Distances for features are defined using the Pearson correlation, and distances between data items are calculated using the Euclidean distance. In our model, we express both dimensionality reduction methods through θ_I and θ_F defining each dimensionality reduction. These steps remain static during the analysis process., i.e., they do not have user-steered parameters.

The selection interaction of this approach is realized with a Lasso Selector on the feature and data space projection. A selection on both views directly parametrizes the selection expressed by F for sel_F and I for sel_I. This approach allows for creating multiple features and data space visualizations, enabling the comparison of different spaces.

Fernstad et al. The approach by Fernstad et al. [18] addresses the need for statistical analysis of features and subgroups of data items. Thus, we assign the scenarios S2 *Statistical Analysis* and S7 *Data Aggregation*. The approach is focused on dimensionality reduction using "quality measures," which are five statistical measures such as variance and skewness denoted by 5. The feature space visualization is a parallel coordinate plot showing these five values plus two measures derived from Pearson correlation.

All measures remain static throughout the analysis. We express them in our model through θ_F, which, in this case, comprises all five statistical measures. The approach by Fernstad et al. [18] is one approach that offers two data space visualization to address both scenarios. All views are linked views. The data space is visualized with a scatterplot SP. The approach covers the selection of specific data items, which parametrizes the selection function sel_I using I.

For the scatterplot visualization, the data items' dimensionality is further reduced using *Principal Component Analysis* PCA, denoted as the aggregation function θ_I. Alongside the scatterplot, another *Parallel Coordinate Plot* PCP shows the selected data items without further reduction. The selections on each visualization provide parameters for our selection function, i.e., sel_F and sel_I.

Rauber et al. The approach by Rauber et al. [4] focuses on the design of classification systems using projections. In this case, the components related to dual analysis are embedded in a larger system, where not all parts feedback into the dual analysis components. The approach supports the interactive selection of features, thus enabling S1 *Feature Selection*. Additionally, it allows the selection of data items to be used in the classification process. Thus, we also assign scenario S7 *Data Selection*.

Feature and data space are both visualized using scatterplots SP. However, they differ in the transformation to determine the x and y-coordinates for each view. The feature space uses *Multidimensional Scaling* MDS using the Pearson correlation as distance measure. We map this property to our framework with the function θ_F of $aggr_F$. The data space uses *t-distributed Stochastic Neighborhood Embedding* t-SNE. We express this within our framework by using the two functions θ_I of $aggr_I$.

Both functions are not further parameterized since no user interaction influences them. However, the feature and data item selection is part of our approach. The selection of features is expressed using parameter F of sel_F and I of sel_I for data items. The selection interaction on both views of the user directly determines these two parameters.

6.2 Generative Use – Opportunities

In this section, we highlight and describe future research opportunities which extend components of our proposed framework. We deliberately designed our formalization to encompass these improvements to dual analysis.

Glyph Design and Adaptation: In our review, we found that most approaches use straightforward scatterplots, where a dot visually encodes two data item properties through color and size. Thus, the next logical step, supported by our formalization, is the integration of glyphs into the scatterplots of the feature and data space visualizations. This allows for the representation of more properties of the data [72]. These glyphs can also be adaptive to the data types of the analyzed dataset. This improvement is derived from our definition of feature and data space visualizations, i.e, FS and DS, (see Fig. 5), which we already extend by defining specific functions for glyph-based visualizations $glyph_I$ and $glyph_F$ (see Eq. 10 and Eq. 4).

Scatterplot Layout Enrichment: Our formalization revealed that the visualization of feature and data space remains straightforward, i.e., primarily based on MDS or PCA projections. The remaining task is to expand visualizations using methods encoding manifold properties in the plot [25]. Since dual analysis approaches make extensive use of dimensionality reduction and scatterplot visualizations, even manipulating parameters of the dimensionality reduction [16], [17], we see a clear need for additional visual feedback. An example of this idea is uPCA [76] and uMDS [77], where uncertainty is visualized. We can adapt how the feature and data space are visualized to integrate such a technique. We propose this in the context of the visualization steps vis_I and vis_F (see Eq. 9 and Eq. 3).

Subspace Detection Algorithms: Four approaches we found during our review mainly address the analysis of subspaces and subspace clusters [8], [13], [14], [41]. However, all techniques provide a purely interactive and user-driven way of subspace cluster analysis. Our formalization allows for an integration of machine learning algorithms for the detection of relevant subspace [65]. In particular, SURFING [22], SUBCLU [78], and RIS [79]. They detect potentially interesting subspaces based on data distribution density. These algorithms can be integrated as parameterizations for the steps of our pipeline to support the realization of scenario S5 *Subspace Cluster Analysis* (see Fig. 5). For example, SURFING can be integrated to facilitate the detection of interesting subspaces by suggesting a selection of features represented by parameter F of sel_F in our framework. Similarly, subspace clusters can be detected beforehand determining parameters F of sel_F and I of sel_I, while dual analysis allows for the exploration of the involved features and data items.

Analytical Provenance: The representation of the dataset as a matrix (i.e., S_F, S_I, A_F, and A_I in Fig. 5) at each step of the dual analysis pipeline allows for a nuanced tracking of the analysis state. Steinparz et al. [80] and Hinterreiter et al. [26] systematized the comparison of matrices for analytical provenance allowing for the comparison and visualizations of different analysis paths. Thus, we support the integration of tracking analysis states by formalizing the matrix representations at every step of our framework.

User Guidance: We also found that no approach involves user guidance. Similarly to analytical provenance, our formalization allows for integrating guidance methods since each step's data selection and layout is well-defined. The next logical step is to contrast each stage of the pipeline (see Fig. 5) with guidance scenarios to find interesting ways to help analysts in their analysis tasks through guidance [27]. Practical guidance frameworks such as Lotse by Sperrle et al. [81] require clearly defined data sources and conditions for their guidance strategies, which our framework enables. For example, suggesting the feature selection F of sel_F, based on what the user has already observed.

7 Limitations and Discussion

During our work, we found that the space of dual analysis approaches is vast. We identified two papers providing model sketches for their dual analysis approaches. When comparing them to our framework, we find that both allow for only a subset of scenarios and interactions, i.e., the dual analysis approach by Corput et al. [21] focuses on ordering data table entries according to relevance or similarity metrics of features and data items. This only covers the scenarios S4 *Similarity Search* and S6 *Data Aggregation and Weighting*. The approach by Turkay at al. [1] focuses on S3 *Statistical Analysis* through linking and brushing.

In both publications, the theory behind each approach states the specifics of the approach, i.e., which metrics are used; a generalization allowing for creating a dual analysis toolbox is missing. Although both approaches describe a model of dual analysis, both publications describe dual analysis differently and only converge if generalized to an abstract definition of dual analysis (see Fig. 1). Hence, both publications do not propose a generalized framework. In our work, we provide a formalized framework that offers well-defined interfaces for each described component used in the dual analysis, which covers 23 approaches and thus unifying frameworks of dual analysis.

Our work comes with limitations resulting from the approach we adopted. To keep the study focused on dual analysis, we had to define dual analysis in Sec. 5, limiting the literature analysis to a representative set of examples, explicitly excluding other approaches, such as VA dashboards. We aimed to identify papers that contribute a dual analysis approach for a given analysis problem, offering interactions beyond filtering. We primarily aimed at results with practical relevance, transparency, and reproducibility.

We thoroughly described our method and decision-making process. Thus, we are confident that we analyzed a representative set of publications and that our framework and formalization contribute to future research. It would be interesting to evaluate the stability of our results in the future by performing an expanded "cross-validation" study that would add papers published in the future. We initially started our analysis with landmark publications from all domains and had to limit the number of papers to keep the work manageable. Our literature analysis identified several contributions that offer valuable interactions to explore datasets and validate hypotheses with dual analysis.

We had long discussions about which interactions to include as scenarios, but we finally decided on the seven descriptive scenarios, which cover all 23 approaches listed in Tab. 1. Other aspects may be included in the interactive dual analysis, which can be integrated into many VA frameworks in general. An interesting opportunity, for example, is visualization quality measures, which was a primary concern when we began this study [82]. The framework by Bertini et al. [83], later extended by Behrisch et al. [84], describes an enriched VA pipeline with quality-metric-driven automation. Quality can be measured at each analysis step (i.e., upon a view update) while the analyst steers the process. Quality metrics can aid user interactions with automatic configurations or recommendations at each step. However, quality metrics do not interact with the underlying data, selection, or aggregation but rather the visualizations themselves and can be seen as an add-on to our proposed formal framework.

We also described machine learning algorithms for dimensionality reduction and relevant subspace detection. Yet, incorporating other machine learning techniques, e.g., for classification, might be a worthwhile pursuit as well [4], [85]. Still, as we established the framework, we focused exclusively on analysis scenarios with dual analysis and its three key components with a *bidirectional* linking of feature and data space.

In future work, we want to implement a framework based on the presented model we derived from existing literature. As a general finding, we can state that all dual analysis approaches, indeed, fit into a generalized model, which can be used to categorize existing analysis systems and show other possibilities for combining different components. We also found that even a specific analysis approach, in this case, dual analysis, is challenging to define. First, to find relevant literature amid all visual analytics approaches. Second, to arrange, condense, and organize the different approaches into a coherent and comprehensive overview.

8 Conclusion and Future Work

Enabling users to explore and analyze the data and feature space of a dataset while maintaining the ability for the user to apply their knowledge about the data, task, and domain provide a great benefit. To achieve this, a comprehensive link between the two spaces needs to be established, which often depends on domain specificities. In this study, we systematically analyzed the visual analytics literature to identify and categorize approaches using dual analysis, i.e., the simultaneous analysis of feature and data space.

We presented our findings through seven descriptive scenarios, which we contextualize with a formalized dual analysis framework. Our analysis revealed several ways that dual analysis can be enriched by incorporating other techniques, such as layout-enrichment of the 2-dimensional projections and suggestions for interesting subspaces. We presented how current VA systems and points support existing strategies for future research directions. We hope our contributions help other researchers investigate, design, and evaluate dual analysis approaches. In future work, we plan to develop a system capable of inferring and adapting its settings in a larger design space than current systems for dual analysis. We aim to leverage existing techniques from related domains, such as machine learning and human-computer interaction, to improve dual analysis for more efficient and effective data analysis.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A03), the Federal Ministry of Education and Research of Germany (BMBF) in the framework of PEGASUS under the program "Forschung für die zivile Sicherheit 2018 - 2023" and its announcement "Zivile Sicherheit - Schutz vor organisierter Kriminalität II", and the ETH AI Center.

References

[1] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions - A dual visual analysis model for high-dimensional data," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2591–2599, 2011.

[2] I. T. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer, 1986.

[3] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser, "Representative factor generation for the interactive visual analysis of high-dimensional data," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2621–2630, 2012.

[4] P. E. Rauber, A. X. Falcão, and A. C. Telea, "Projections as visual aids for classification system design," *Inf. Vis.*, vol. 17, no. 4, pp. 282–305, 2018.

[5] T. Itoh, A. Kumar, K. Klein, and J. Kim, "High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots," *J. Vis. Lang. Comput.*, vol. 43, pp. 1–13, 2017.

[6] L. Garrison, J. Müller, S. Schreiber, S. Oeltze-Jafra, H. Hauser, and S. Bruckner, "Dimlift: Interactive hierarchical data exploration through dimensional bundling," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 2908–2922, 2021.

[7] J. Müller, L. Garrison, P. Ulbrich, S. Schreiber, S. Bruckner, H. Hauser, and S. Oeltze-Jafra, "Integrated dual analysis of quantitative and qualitative high-dimensional data," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 2953–2966, 2021.

[8] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes, "Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2033–2042, 2014.

[9] J. Krause, A. Dasgupta, J. Fekete, and E. Bertini, "Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces," in *6th IEEE Symp. Large Data Anal. Vis.* IEEE Computer Society, 2016, pp. 11–19.

[10] W. Jentner, D. Sacha, F. Stoffel, G. P. Ellis, L. Zhang, and D. A. Keim, "Making machine intelligence less scary for criminal analysts: reflections on designing a visual comparative case analysis tool," *Vis. Comput.*, vol. 34, no. 9, pp. 1225–1241, 2018.

[11] T. Fujiwara, O. Kwon, and K. Ma, "Supporting analysis of dimensionality reduction results with contrastive learning," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 45–55, 2020.

[12] A. Soriano-Vargas, B. Hamann, and M. C. F. de Oliveira, "TV-MV analytics: A visual analytics framework to explore time-varying multivariate data," *Inf. Vis.*, vol. 19, no. 1, 2020.

[13] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.

[14] C. Turkay, E. Kaya, S. Balcisoy, and H. Hauser, "Designing progressive and interactive analytics processes for high-dimensional data analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 131–140, 2017.

[15] A. Inselberg, "The plane with parallel coordinates," *Vis. Compt.*, vol. 1, no. 2, pp. 69–91, 8 1985.

[16] J. Z. Self, R. K. Vinayagam, J. T. Fry, and C. North, "Bridging the gap between user intention and model parameters for human-in-the-loop data analytics," in *Proc. Workshop on Human-In-the-Loop Data Anal.* ACM, 2016, p. 3.

[17] M. Dowling, J. E. Wenskovitch, J. T. Fry, S. Leman, L. House, and C. North, "SIRIUS: dual, symmetric, interactive dimension reductions," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 172–182, 2019.

[18] S. J. Fernstad, J. Shaw, and J. Johansson, "Quality-based guidance for exploratory dimensionality reduction," *Inf. Vis.*, vol. 12, no. 1, pp. 44–64, 2013.

[19] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. P. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[20] D. W. Evans, P. T. Orr, S. M. Lazar, D. Breton, J. Gerard, D. H. Ledbetter, K. Janosco, J. Dotts, and H. Batchelder, "Human preferences for symmetry: subjective experience, cognitive conflict and cortical brain activity," *PLoS One*, vol. 7, no. 6, 2012.

[21] P. van der Corput and J. J. van Wijk, "Exploring items and features with i^f, f^i-tables," *Comput. Graph. Forum*, vol. 35, no. 3, pp. 31–40, 2016.

[22] C. Baumgartner, C. Plant, K. Kailing, H. Kriegel, and P. Kröger, "Subspace selection for clustering high-dimensional data," in *Proc. IEEE Int. Conf. Data Min.* IEEE Computer Society, 2004, pp. 11–18.

[23] I. Koprinska, "Feature selection for brain-computer interfaces," in *New Frontiers in Applied Data Mining, PAKDD 2009 International Workshops*, ser. Lecture Notes in Computer Science, vol. 5669. Springer, 2009, pp. 106–117.

[24] M. Mehri, R. Chaieb, K. Kalti, P. Héroux, R. Mullot, and N. E. B. Amara, "A comparative study of two state-of-the-art feature selection algorithms for texture-based pixel-labeling task of ancient documents," *J. Imaging*, vol. 4, no. 8, p. 97, 2018.

[25] L. G. Nonato and M. Aupetit, "Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2650–2673, 2019.

[26] A. P. Hinterreiter, C. A. Steinparz, M. Schöfl, H. Stitz, and M. Streit, "Projection path explorer: Exploring visual patterns in projected decision-making paths," *ACM Trans. Interact. Intell. Syst.*, vol. 11, no. 3-4, pp. 22:1–22:29, 2021. [Online]. Available: https://doi.org/10.1145/3387165

[27] I. Pérez-Messina, D. Ceneda, M. El-Assady, S. Miksch, and F. Sperrle, "A typology of guidance tasks in mixed-initiative visual analytics environments," *Comput. Graph. Forum*, vol. 41, no. 3, pp. 465–476, 2022.

[28] T. M. Green, W. Ribarsky, and B. D. Fisher, "Building and applying a human cognition model for visual analytics," *Inf. Vis.*, vol. 8, no. 1, pp. 1–13, 2009.

[29] J. J. van Wijk, "The value of visualization," in *16th IEEE Vis. Conf.* IEEE Computer Society, 2005, pp. 79–86.

[30] T. M. Green, W. Ribarsky, and B. D. Fisher, "Visual analytics for complex concepts using a human cognition model," in *3rd IEEE Symp. Vis. Anal. Sci. Technol.* IEEE Computer Society, 2008, pp. 91–98.

[31] D. A. Keim, J. Kohlhammer, G. P. Ellis, and F. Mansmann, *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics Association, 2010.

[32] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[33] F. Lekschas, X. Zhou, W. Chen, N. Gehlenborg, B. Bach, and H. Pfister, "A generic framework and library for exploration of small multiples through interactive piling," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 358–368, 2021. [Online]. Available: https://doi.org/10.1109/TVCG.2020.3028948

[34] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[35] T. von Landesberger, S. Fiebig, S. Bremm, A. Kuijper, and D. W. Fellner, "Interaction taxonomy for tracking of user actions in visual analytics applications," in *Handbook of Human Centric Visualization*. Springer, 2014, pp. 653–670.

[36] E. Bertini and D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery," in *Proc. ACM SIGKDD Workshop Vis. Anal. Knowl. Disc.: Integrating Automated Analysis with Interactive Exploration*. ACM, 2009, pp. 12–20.

[37] A. Endert, "Semantic interaction for visual analytics: Toward coupling cognition and computation," *IEEE Computer Graphics and Applications*, vol. 34, no. 4, pp. 8–15, 2014. [Online]. Available: https://doi.org/10.1109/MCG.2014.73

[38] E. Dimara and C. Perin, "What is interaction for data visualization?" *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 119–129, 2020. [Online]. Available: https://doi.org/10.1109/TVCG.2019.2934283

[39] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019.

[40] C. Turkay, A. Lex, M. Streit, H. Pfister, and H. Hauser, "Characterizing cancer subtypes using dual analysis in caleydo stratomex," *IEEE Comput. Graph.*, vol. 34, no. 2, pp. 38–47, 2014.

[41] C. Turkay, A. Slingsby, K. Lahtinen, S. Butt, and J. Dykes, "Supporting theoretically-grounded model building in the social sciences through interactive visualisation," *Neurocomputing*, vol. 268, pp. 153–163, 2017.

[42] A. Bibal, A. Clarinval, B. Dumas, and B. Frénay, "IXVC: an interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees," *Array*, vol. 11, p. 100080, 2021.

[43] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller, "Visual correlation analysis of numerical and categorical data on the correlation map," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 2, pp. 289–303, 2015. [Online]. Available: https://doi.org/10.1109/TVCG.2014.2350494

[44] J. Wei, H. Yu, R. W. Grout, J. H. Chen, and K. Ma, "Dual space analysis of turbulent combustion particle data," in *IEEE Pacific Visualization Symposium*, G. D. Battista, J. Fekete, and H. Qu, Eds. IEEE Computer Society, 2011, pp. 91–98. [Online]. Available: https://doi.org/10.1109/PACIFICVIS.2011.5742377

[45] Y. Wang, W. Chen, J. Zhang, T. Dong, G. Shan, and X. Chi, "Efficient volume exploration using the gaussian mixture model," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1560–1573, 2011. [Online]. Available: https://doi.org/10.1109/TVCG.2011.97

[46] L. Wang, X. Zhao, and A. E. Kaufman, "Modified dendrogram of attribute space for multidimensional transfer function design," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 1, pp. 121–131, 2012. [Online]. Available: https://doi.org/10.1109/TVCG.2011.23

[47] F. Tzeng, E. B. Lum, and K. Ma, "A novel interface for higher-dimensional classification of volume data," in *14th IEEE Visualization Conference*, G. Turk, J. J. van Wijk, and R. J. M. II, Eds. IEEE Computer Society, 2003, pp. 505–512. [Online]. Available: https://doi.org/10.1109/VISUAL.2003.1250413

[48] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara, "Interactive focus+context analysis of large, time-dependent flow simulation data," *Simul.*, vol. 82, no. 12, pp. 851–865, 2006. [Online]. Available: https://doi.org/10.1177/0037549707078278

[49] W. Chen, Z. Ding, S. Zhang, A. MacKay-Brandt, S. Correia, H. Qu, J. A. Crow, D. F. Tate, Z. Yan, and Q. Peng, "A novel interface for interactive exploration of DTI fibers," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1433–1440, 2009. [Online]. Available: https://doi.org/10.1109/TVCG.2009.112

[50] C. E. Weaver, "Cross-filtered views for multidimensional visual analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 2,

pp. 192–204, 2010. [Online]. Available: https://doi.org/10.1109/TVCG.2009.94

[51] G. G. Zanabria, L. G. Nonato, and E. G. Nieto, "istar (i*): An interactive star coordinates approach for high-dimensional data exploration," *Comput. Graph.*, vol. 60, pp. 107–118, 2016.

[52] E. Artur and R. Minghim, "A novel visual approach for enhanced attribute analysis and selection," *Comput. Graph.*, vol. 84, pp. 160–172, 2019.

[53] J. Zhao, M. Karimzadeh, A. Masjedi, T. Wang, X. Zhang, M. M. Crawford, and D. S. Ebert, "Featureexplorer: Interactive feature selection and exploration of regression models for hyperspectral images," in *30th IEEE Vis. Conf.* IEEE, 2019, pp. 161–165.

[54] M. Miller, J. Rauscher, D. A. Keim, and M. El-Assady, "CorpusVis: Visual Analysis of Digital Sheet Music Collections," *Comput. Graph. Forum*, vol. 41, no. 3, pp. 283–294, 2022.

[55] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 241–250, 2017.

[56] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands: Esri Press, 2010.

[57] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[58] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," in *IEEE Inf. Vis. Symp.*, 2000, pp. 9–12.

[59] D. A. Keim, "Pixel-oriented database visualizations," *SIGMOD Record*, vol. 25, no. 4, pp. 35–39, 1996.

[60] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling: Theory and applications.* New York: Springer, 2005.

[61] L. Nováková and O. Štěpánková, "Visualization of trends using radviz," in *J. Intell. Inf. Syst.* Springer, 2009, pp. 56–65.

[62] J. Pagès, *Multiple Factor Analysis by Example Using R*, 1st ed., ser. The R Series. Chapman & Hall/CRC, 2014.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[64] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Second Int. Conf. Knowl. Disc. Data Min.*, E. Simoudis, J. Han, and U. M. Fayyad, Eds. AAAI Press, 1996, pp. 226–231.

[65] H. Kriegel, P. Kröger, and A. Zimek, "Subspace clustering," *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, vol. 2, no. 4, pp. 351–364, 2012.

[66] T. Munzner, *Visualization Analysis and Design*, ser. A.K. Peters visualization series. A K Peters, 2014.

[67] J. Brownlee, "Why one-hot encode data in machine learning?" Online, 2017, https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/, accessed 2022-10-19.

[68] M. O. Ward, "Linking and brushing," in *Encyclopedia of Database Systems.* Springer, 2009, pp. 1623–1626.

[69] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[70] C. C. Aggarwal and C. K. Reddy, Eds., *Data Clustering: Algorithms and Applications.* CRC Press, 2014.

[71] A. Abuthawabeh and M. Aupetit, "Toward an interactive voronoi treemap for manual arrangement and grouping," in *21st Eurographics Conference on Visualization*, M. Agus, C. Garth, and A. Kerren, Eds. Eurographics Association, 2021, pp. 97–101. [Online]. Available: https://doi.org/10.2312/evs.20211062

[72] J. Fuchs, P. Isenberg, A. Bezerianos, and D. A. Keim, "A systematic review of experimental studies on data glyphs," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 7, pp. 1863–1879, 2017.

[73] R. Minghim, F. V. Paulovich, and A. de Andrade Lopes, "Content-based text mapping using multi-dimensional projections for exploration of document collections," in *Visualization and Data Analysis*, ser. Proc. SPIE, vol. 6060, 2006, p. 60600S.

[74] D. Fisher and M. D. Meyer, *Making Data Visual - A Practical Guide to Using Visualization For Insight.* O'Reilly, 2018.

[75] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision To Think.* Acad. Press, 1999.

[76] J. Görtler, T. Spinner, D. Streeb, D. Weiskopf, and O. Deussen, "Uncertainty-aware principal component analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 822–831, 2020.

[77] D. Hägele, T. Krake, and D. Weiskopf, "Uncertainty-aware multidimensional scaling," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 1, pp. 23–32, 2023. [Online]. Available: https://doi.org/10.1109/TVCG.2022.3209420

[78] K. Kailing, H. Kriegel, and P. Kröger, "Density-connected subspace clustering for high-dimensional data," in *Proc. 4th SIAM Int. Conf. Data Min.* SIAM, 2004, pp. 246–256.

[79] K. Kailing, H. Kriegel, P. Kröger, and S. Wanka, "Ranking interesting subspaces for clustering high dimensional data," in *7th Eur. Conf. Principles and Practice of Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, vol. 2838. Springer, 2003, pp. 241–252.

[80] C. A. Steinparz, A. P. Hinterreiter, H. Stitz, and M. Streit, "Visualization of rubik's cube solution algorithms," in *10th Int. EuroVis Workshop Vis. Anal.* Eurographics Association, 2019, pp. 19–23.

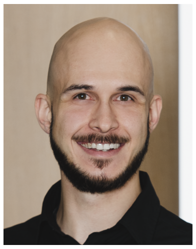
[81] F. Sperrle, D. Ceneda, and M. El-Assady, "Lotse: A practical framework for guidance in visual analytics," *IEEE Trans. Vis. Comput. Graphics*, pp. 1–11, 2022.

[82] F. L. Dennig, M. T. Fischer, M. Blumenschein, J. Fuchs, D. A. Keim, and E. Dimara, "Parsetgnostics: Quality metrics for parallel sets," *Comput. Graph. Forum*, vol. 40, no. 3, pp. 375–386, 2021. [Online]. Available: https://doi.org/10.1111/cgf.14314

[83] E. Bertini, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2203–2212, 2011.

[84] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim, "Quality metrics for information visualization," *Comput. Graph. Forum*, vol. 37, no. 3, pp. 625–662, 2018.

[85] F. L. Dennig, T. Polk, Z. Lin, T. Schreck, H. Pfister, and M. Behrisch, "FDive: Learning relevance models using pattern-based similarity measures," in *14th IEEE Conf. Vis. Anal. Sci. Technol.* IEEE, 2019, pp. 69–80.



Frederik L. Dennig received his Master of Science in Computer and Information Science at the University of Konstanz, Germany, in 2019. As a doctoral researcher, he is part of the Data Analysis & Visualization Research Group at the University of Konstanz. His main research areas are visual analytics, pattern detection, and subspace analysis. A specific focus of his research is on the quantification of quality and visual patterns in visualizations.

Matthias Miller completed his Master of Science in the field of Computer and Information Science at the University of Konstanz, Germany, in 2018. As part of the Data Analysis & Visualization Research Group at the University Konstanz, he started as a Research Associate and Doctoral Researcher who is currently working toward a Ph.D. degree in the topic "Visual Sheet Music Analysis." His interests comprise visual analytics with a special focus on Visual Musicology.

Prof. Dr. Daniel A. Keim leads the Data Analysis and Visualization Research Group in the Computer and Information Science Department at the University of Konstanz, Germany. He was program chair of InfoVis 1999, InfoVis 2000, KDD 2002, VAST 2006, VAST 2019, and VMV 2022; general chair of InfoVis 2003; and associate editor of IEEE TVCG, IEEE TKDE, and Sage Information Visualization Journal. He received his Ph.D. degree in Computer Science from the University of Munich, Germany. He was an associate professor at the University of Halle, Germany, and a Technology Consultant at AT&T Shannon Research Labs, NJ, USA.

Mennatallah El-Assady is a research fellow at the AI Center of ETH Zurich, Switzerland. Prior to that, she was a research associate and doctoral student in the group for Data Analysis and Visualization at the University of Konstanz, Germany, and in the Visualization for Information Analysis lab at the Ontario Tech University, Canada. She works at the intersection of data analysis, visualization, computational linguistics, and explainable artificial intelligence.