Measure-Driven Visual Analytics of Categorical Data

Doctoral thesis for obtaining the academic degree of Doctor of Natural Sciences (Dr. rer. nat.)

submitted by

Frederik Dennig

at the

Universität
Konstanz

Faculty of Sciences

Department of Computer and Information Science

Konstanz, 2024

Doctoral Thesis of the University of Konstanz
Day of the oral examination: July 17th, 2024

1. Referee: Prof. Dr. Daniel A. Keim, University of Konstanz
2. Referee: Prof. Dr. Tobias Schreck, Graz University of Technology

(This page intentionally left blank)

Dennig, Frederik
Measure-Driven Visual Analytics of Categorical Data
Dissertation, University of Konstanz, 2024

Abstract

Visual Analytics (VA) enables data analysts and domain experts to engage in analytical reasoning through interactive visual interfaces. One type of data often encountered in data analysis tasks is categorical data. Unlike numerical data, categorical data with nominal attributes has no inherent order or scale and, therefore, does not lend itself to the application of common arithmetic operations. However, many data mining and visualization techniques are predominantly based on numerical data. Notwithstanding these challenges, the analysis of categorical data is crucial in various domains, including linguistics and software engineering. This dissertation addresses the challenges posed by categorical data, including difficulties in establishing an order of attributes for visualization and defining numerical abstractions. This work bridges the qualitative-quantitative divide in the visual analysis of categorical data by introducing abstractions that improve the readability of categorical data visualizations, developing new strategies for applying methods typically designed for numerical data, and exploring their interplay with numerical data. This thesis is structured in three parts: The first part introduces quality measures for the Parallel Sets visualization. In addition, we present measures that guide the exploration of categorical data projections by suggesting attributes that differentiate groups of data items. The second part presents measure-driven approaches for expressing categorical data properties and deriving numerical representations for the domains of linguistics and software engineering, demonstrating the power of measure-driven approaches in real-world applications. The third part addresses the joint analysis of categorical attributes and numerical data dimensions. It offers strategies for the use of categorical data for model training and exploratory data analysis in supervised and unsupervised frameworks. Finally, this thesis outlines the limitations and lessons learned from the explored measure-driven approaches and suggests future directions for more effectively integrating categorical data into VA with the goal of improving the readability of visualization, pattern quantification and user guidance. In conclusion, this work improves the analysis and visualization of categorical data by proposing new measure-driven approaches, improving readability and interpretability of visualizations, providing domain-agnostic and domain-specific support for exploratory data analysis, and their integration into supervised and unsupervised VA frameworks.

Zusammenfassung

Visual Analytics (VA) ermöglicht es Datenanalysten und Fachexperten, über interaktive visuelle Schnittstellen analytische Überlegungen anzustellen. Eine bei der Datenanalyse häufig anzutreffende Art von Daten sind kategoriale Daten. Im Gegensatz zu numerischen Daten haben kategoriale Daten mit nominalen Attributen keine inhärente Ordnung oder Skala und eignen sich daher nicht für die Anwendung gängiger arithmetischer Operationen. Viele Data-Mining- und Visualisierungstechniken beruhen jedoch überwiegend auf numerischen Daten. Trotz dieser Herausforderungen ist die Analyse kategorischer Daten in verschiedenen Bereichen wie Linguistik und Software-Engineering von entscheidender Bedeutung. Diese Dissertation befasst sich mit den Herausforderungen, die kategoriale Daten mit sich bringen, einschließlich der Schwierigkeiten bei der Festlegung einer Reihenfolge von Attributen für die Visualisierung und der Definition numerischer Abstraktionen. Diese Arbeit überbrückt die Lücke zwischen Qualität und Quantität in der visuellen Analyse kategorialer Daten, indem sie Abstraktionen einführt, die die Lesbarkeit kategorialer Datenvisualisierungen verbessern, neue Strategien für die Anwendung von Methoden zur Analyse numerischer Daten vorstellt, und das Zusammenspiel kategorischer und numerischen Daten untersucht. Diese Dissertation ist in drei Teile gegliedert: Der erste Teil führt Qualitätsmaße für die Parallel Sets Visualisierung ein. Darüber hinaus stellen wir Maße vor, die die Exploration von Projektionen kategorialer Daten leiten, indem wir Attribute vorschlagen, die Gruppen von Datenelementen unterscheiden. Der zweite Teil stellt maßgetriebene Ansätze vor, um Eigenschaften kategorialer Daten auszudrücken und numerische Darstellungen für die Bereiche Linguistik und Software-Engineering abzuleiten und demonstriert deren Stärke in praktischen Anwendungen. Der dritte Teil befasst sich mit der Analyse kategorialer Attribute in Kombination mit numerischen Datendimensionen. Er beschreibt Strategien für die Nutzung von kategorialen Daten, für das Trainieren von Modellen und für die explorative Datenanalyse, in überwachten und unüberwachten Frameworks. Abschließend werden in dieser Arbeit die Grenzen und Lehren aus den untersuchten maßgetriebenen Ansätzen aufgezeigt und Vorschläge für eine effektivere Integration kategorialer Daten in VA gemacht, mit dem Ziel, die Lesbarkeit von Visualisierungen zu verbessern sowie die Mustererkennung und User Guidance zu optimieren. Zusammenfassend verbessert diese Arbeit die Analyse und Visualisierung kategorialer Daten, indem sie neue maßgetriebene Ansätze vorschlägt,

die Lesbarkeit und Interpretierbarkeit von Visualisierungen verbessert, domänenunabhängige und domänenspezifische Unterstützung für die explorative Datenanalyse bietet und deren Integration in überwachte und unüberwachte VA-Frameworks ermöglicht.

Acknowledgement

This dissertation is the result of the help and continued support of many people who have generously shared their time and expertise with me - I thank all of them for making this thesis possible! I am immensely grateful to my advisor, Daniel Keim, for allowing me to join an outstanding research group to pursue exciting ideas and provide valuable guidance when needed. I would also like to thank my second advisor, Tobias Schreck, for his support throughout my journey, especially at the beginning of my research career, by providing helpful insights during my first, now published, research project. I would also like to thank Michael Grossniklaus for chairing the oral examination committee, as well as for his critical questions and comments as a student, which helped refine my research.

I was fortunate to have excellent mentors who showed me the beauty of research before and during my Ph.D. journey. I thank Michael Behrisch and Michael Blumenschein for their constructive feedback, knowledge, experience, and perspective. Their support significantly influenced this thesis and helped me develop my scientific and personal skills on a broader level. I thank all members of the Data Analysis and Visualization group for their help and support, the lively debates, and the warm conversations around the coffee machine that turned busy days into delightful and sometimes hilarious moments. I would also like to thank Matthias Miller, Thilo Spinner, Maximilian Fischer, Julius Rauscher, Lucas Joos, Rita Sevastjanova, and Daniel Seebacher for our meaningful conversations outside of work. I appreciate the support of our system administration team, Giovanna Ratini, Matt Sharinghousen, Philipp Traber, Felix Dobler, Niclas Rohn, Benedikt Bäumle, and Timon Kilian, who helped me with all my technical challenges. I would also like to thank Wolfgang Jentner, who generously improved our technical setup and computing infrastructure even after he left our research group. Johannes Fuchs gave me valuable tips on how to design user studies. Evanthia Dimara showed me how to improve my writing by being critical and transparent about every word, sentence, and paragraph. I thank both of them for their help. Finally, thank you, Sabine Kuhr, for helping me with administrative issues.

Christine Beck from the Linguistics Department has been a great source of support. I thank her for sharing perspectives from outside computer science and advancing our research areas. Working with the outstanding students Daniela Blumberg, who recently joined our research group, Benjamin Halbritter, and Nina

Geyer has been a privilege over the past few years. They have supported my research with their programming, discussions, and innovations. This thesis is greatly indebted to the funding and excellent research network of the DFG-funded SFB-TRR 161 ”Quantitative Methods for Visual Computing” and the “Strategic Programs for Advanced Research and Technology in Europe” project funded by the European Union’s Horizon 2020 research program.

I would also like to thank my family and friends for their love, support, and encouragement. I appreciate your help finding my direction in life and sometimes challenging my decisions. Finally, I sincerely thank God, my loving Father, for carrying me through every challenge. Your guidance was evident each day. It is by Your grace that I was able to complete my doctoral studies. I will continue to rely on You for all that is to come. Thank You, Lord.

Contents

1 Introduction to Visual Analytics of Categorical Data

Contents

Visual Analytics (VA) [180, 176, 303, 179, 266] facilitates analytical reasoning through interactive visual interfaces combining the fields of *Information Visualization (InfoVis)* and *Data Mining* with the primary goal to enable humans to effectively understand and analyse complex datasets. VA uses visual representations and interaction methods of *Human-Computer Interaction (HCI)* [209, 93] to help users discover patterns, correlations, and other relationships within large and complex datasets that would be difficult to uncover using traditional analytical methods. This interdisciplinary approach leverages the cognitive processing capabilities of humans and the computational power of computers to support decision making, problem solving and discovery in diverse domains such as business intelligence, healthcare, security and research. VA emphasizes the importance of visualization in the analytical process, not only for presenting results, but as an integral part of exploration, hypothesis generation, and validation. Complex datasets often contain a wide variety of data types. These can range from qualitative and quantitative variables found in high-dimensional datasets to other representations such as geographic [234] and network data [51], all of which may be in a temporal context [151, 6].

The domain of qualitative variables is defined by the term *categorical data* [44, 90, 5, 36], which describes non-numerical properties of a data item representing characteristics such as gender, nationality, brand preference or type of cuisine. More specifically, categorical data is further divided into two subtypes, nominal and ordinal. *Nominal data* describes categories that have no inherent order or ranking among themselves. Nominal attributes are used to label or name properties of items in a dataset. Examples include colors (e.g., “red”, “blue”, “green”) or types of animals (e.g., “dog”, “cat”, “bird”). *Ordinal data* consists of categories that do have a natural order or ranking, but the intervals between the categories are not necessarily consistent or defined. Examples of ordinal data include educational levels (e.g., “high school”, “undergraduate”, “graduate”), satisfaction ratings (e.g., “satisfied”, “neutral”, “dissatisfied”), or stages of disease (e.g., “mild”, “moderate”, "severe"). Categorical data is fundamental to statistical analysis and research because it allows you to classify and compare different groups or entities based on qualitative characteristics. It is widely used in surveys, polling, marketing research, and many other disciplines where data needs to be categorized, analyzed, and interpreted. Analyzing categorical data presents several challenges due to its unique characteristics that differ from numerical data, mainly due to the fact that common arithmetic operations such as addition, subtraction, multiplication, and division cannot be applied.

Challenges in the Visual Analysis of Categorical Data

The visual analysis of categorical data faces several general challenges, particularly related to the limitations in encoding variables. These challenges stem from the inherent nature of categorical data and the need to convey meaningful information visually without the aid of numerical scales. Here we describe general challenges in categorical data analysis that also affect VA and InfoVis.

In many cases, particularly with nominal categorical data, there is *no natural ordering* among the categories. For example, categories representing colors, genders, or types of vehicles do not have a mathematical order that can be universally agreed upon. Categorical data does not support meaningful distance or interval measurements between its categories [5, 297]. While numerical data allows for the calculation of difference or distance (e.g., the difference in temperature between two days), categorical data does not provide a way to quantify the "distance" between its categories, e.g., how much "different" or "farther away" is "blue" from "red". Because of the lack of order and distance, most mathematical operations that apply to numerical data do not apply to categorical data. Operations such as addition, subtraction, or averaging do not make sense for categorical data, limiting the types of statistical analyses that can be applied. For ordinal data, where there is a sense of order, the *lack of a consistent scale* poses a challenge in determining spacing or size in visual representations.

Categorical variables can have a large number of categories, i.e., *high cardinality* [217], which makes them difficult to display in a visual format. Visualizing a variable with many categories can lead to cluttered, confusing, or overly complex visualizations that are difficult to read and interpret. Designing a visualization that accommodates all categories without sacrificing readability and insight is a significant challenge. Similarly, color is a standard way to encode categorical attributes. However, there are limitations to the number of distinct, easily distinguishable colors that can be used, which is particularly problematic for attributes with many categories [137, 223, 204]. In addition, relying on color alone can make visualizations inaccessible in certain cases, such as for color-blind users or when printed.

Visualizing multiple categorical variables at the same time can make it difficult for viewers to compare across different categories or variables because the *lack of a common metric or scale* across different categorical variables makes it difficult to compare effectively, such as when using many small visualizations (i.e., small multiples) [307]. For datasets with a large number of different category combinations, visualizing categorical data in a single visualization can result in overplotting, where parts of the visualization overlap, making it difficult to discern groups, relationships, or even individual categories [228]. Thus, ensuring the interpretability of categorical data and providing context through visual means requires careful consideration of visualization design. Labels, legends, and annotations are critical, especially when dealing with abstract categories that may require additional explanation for the viewer to understand.

Finally, categorical data is often present in conjunction with quantitative data, called *mixed data* [14]. However, most data mining and visualization methods are designed for one data type only, requiring categorical data to be encoded for analysis with numerical data methods, or vice versa. One-hot encoding, label encoding, and other methods have different advantages and disadvantages, and their suitability varies depending on the analysis or modeling technique used [261, 57, 134]. In particular, designing visualizations that effectively integrate and display both types of data without overshadowing or misrepresenting one or the other can be challenging.

1.1 Supporting Categorical Data Analysis and Research Questions

This thesis seeks to harness the potential of categorical data within the realm of visual data analysis, with the goal of improving the analysis of such data through the contribution of measure-driven visual analysis techniques that address all transformation steps of the InfoVis Pipeline [62] (see figure 1.1). The following is a description of three sub-research questions addressed in this dissertation that deal with categorical data from three different perspectives: (1) categorical data visualizations, (2) categorical data in domain-specific applications, and (3) categorical data and its interactions with numerical data. However, these sub-research questions and perspectives address *one* overarching research question:

RQ0 *How can we improve the measure-driven visual analysis of categorical data?*

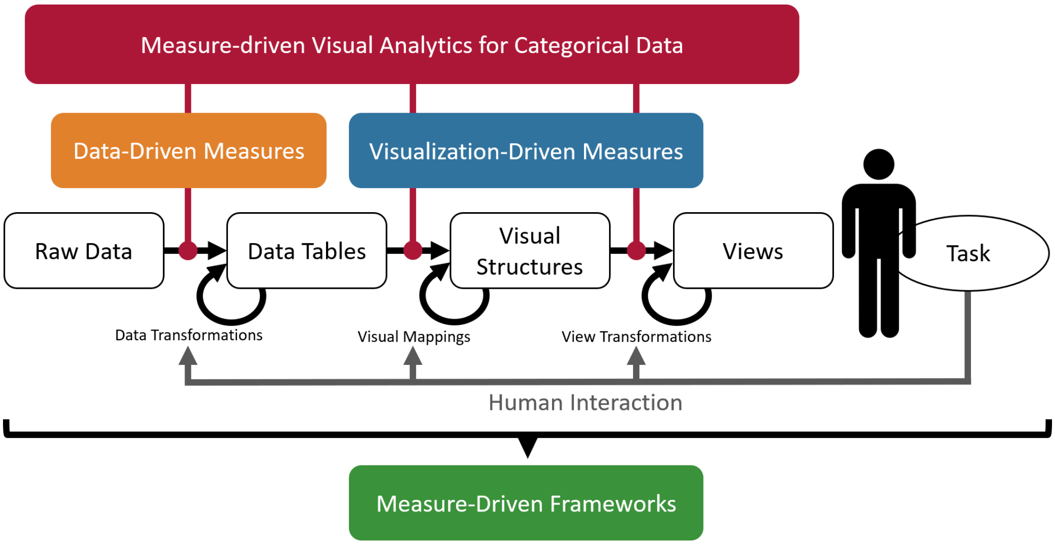


Figure 1.1: This dissertation contributes measure-driven VA approaches for categorical data, thereby addressing the transformation steps of the visualization reference model by Card et al. [62] (i.e. the Information Visualization (InfoVis) Pipeline) and is structured along the topics *data- and visualization-driven measures* as well as *measures-driven frameworks*.

Contributing Measures for Quality and Patterns in Categorical Data Visualizations

In the field of data visualization, significant progress has been made in the development of measures for evaluating quality and identifying patterns, especially for numerical data visualizations. Classic examples include Scagnostics [330], which are used to identify interesting patterns in scatterplots, and Pargnostics [78], tailored for assessing the presence of patterns in Parallel Coordinate Plots (PCPs), both visualization techniques are designed for numeric data. These measures have proven invaluable in guiding the design and interpretation of visualizations, facilitating a deeper understanding of the underlying data patterns and relationships [41, 33]. However, there is a notable gap in the visualization analytics landscape: the lack of measures specifically tailored to the evaluation of categorical data visualizations. When visualization designers set out to create effective diagrams, they have two main options: they can either rely on their own expertise or seek out knowledge about visualization design [325, 228, 291, 306].

In addition, categorical data, with its unique characteristics and challenges, requires different approaches to quality assessment and pattern detection. The lack of such measures hinders the ability of analysts and researchers to systematically evaluate and optimize the design of categorical data visualizations, limiting the effectiveness of these tools in any kind of data analysis task. We therefore focus on these challenges, which are summarized in the following research question:

RQ1 *How can we quantify quality and patterns in categorical data visualizations?*

Demonstrating the Effectiveness of Measure-Driven Applications for Categorical Data

Given the advances in data visualization and analysis techniques in computer science, there is a need for their application and evaluation in real-world contexts, especially outside the confines of theoretical or computational studies [246, 109, 169, 69]. Theoretical advances and practical implementation have limited validity and impact unless their effectiveness is demonstrated in real-world settings. This dissertation addresses this gap by demonstrating the effectiveness of measure-driven VA approaches tailored for categorical data in a domain-specific context. A goal of our work is to develop and deploy methodologies driven by measures designed for nuanced analysis of categorical data. These measures are derived from the unique characteristics and needs of specific domains, providing a more relevant and impactful set of analysis tools that can directly contribute to domain-specific challenges and goals. We summarize this goal in the following research question:

RQ2 *How can we support domain experts in the analysis process by leveraging domain-specific measures for categorical data?*

Leveraging Measures for Categorical Data in Supervised and Unsupervised Frameworks

In the fields of VA and InfoVis, various abstract frameworks outline how different components interact to promote insight and knowledge generation [62, 329, 125, 178, 266]. These frameworks provide a structured approach to understanding VA and InfoVis in an abstract and comprehensive way. In addition, some frameworks are designed for specific purposes and are evaluated by reference implementations [34, 265]. However, the development and application of visualization- and data-driven measures within supervised and unsupervised learning frameworks have predominantly focused on quantitative data, leaving a noticeable gap in the treatment and analysis of categorical data. This gap is particularly evident in the lack of generalized frameworks that effectively incorporate measures for categorical data. This dissertation presents frameworks designed to apply measures in both supervised and unsupervised learning contexts for the combined analysis of categorical and numerical data. We address the following research question:

> **RQ3** *How can we use measures for the joint analysis categorical and numerical data?*

1.2 Contributions and Thesis Structure

To address the challenges described in the previous section centered around the primary research question (R0), we focus on the visual analysis of categorical data through the application of measures, which involves the quantification of various elements within the VA workflow. The structure of this thesis is organized into three main parts (see figure 1.2) that address: (1) the visualization of categorical data, (2) the use of categorical data in different domain-specific applications, and (3) the exploration of how categorical data interacts with numerical data.

Part I: Visualization-Driven Measures

This part focuses on quantifying categorical data visualizations, specifically quality and patterns, and addresses research question (R1). To address this research question, our work introduces measures to quantify the quality of Parallel Sets visualizations, a popular visualization technique for categorical data [35, 190]. In chapter 2, we contribute Parsetgnostics, a set of eight quality measures to improve the visualization of parallel sets, quantifying visual clutter. ParSetgnostics quantifies key properties of Parallel Sets, such as overlap, orthogonality, ribbon width variance, and mutual information. Our measures are intended to provide objective criteria for evaluating the effectiveness of Parallel Sets visualization designs in representing categorical data, and thereby serve as a guide for their creation and refinement. We conducted a systematic correlation analysis between the individual ParSetgnostics measures to ensure that each measure quantifies a unique property and, thus, is distinct from the others. We also applied ParSetgnostics to reconstructions of six datasets previously visualized using Parallel Sets to demonstrate the effect of clutter reduction. By optimizing visual designs based on the proposed measures, ParSetgnostics achieves a clutter reduction of up to 81% compared to original Parallel Sets visualizations, improving the clarity and usability of visualizations. These

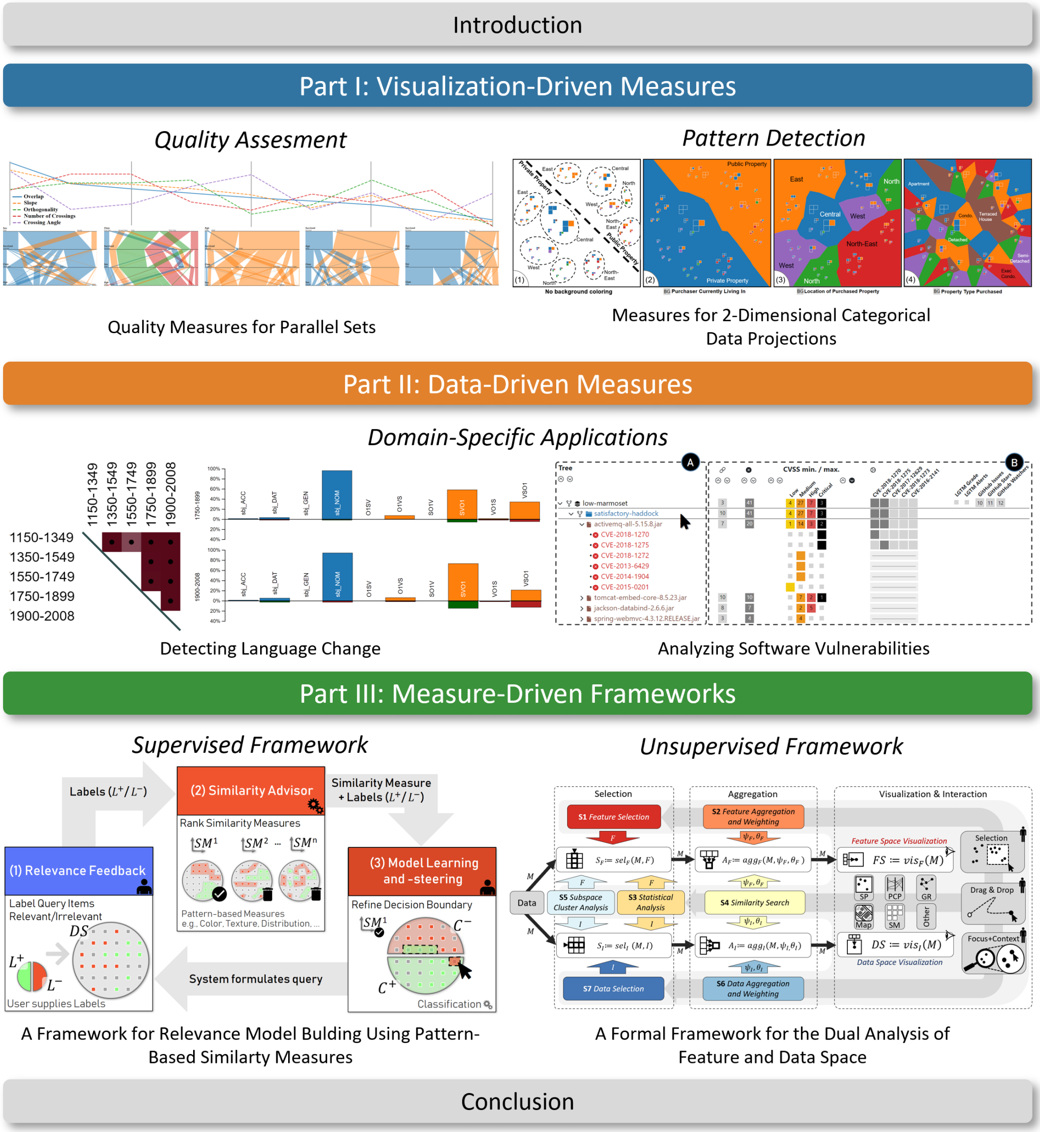

Introduction
Part I: Visualization-Driven Measures
Quality Assesment
Pattern Detection
Quality Measures for Parallel Sets
Measures for 2-Dimensional Categorical Data Projections
Part II: Data-Driven Measures
Domain-Specific Applications
1150-1349
1350-1549
1550-1749
1750-1899
1900-2008
Detecting Language Change
Analyzing Software Vulnerabilities
Part III: Measure-Driven Frameworks
Supervised Framework
Unsupervised Framework
Labels (L^+/L^-)
(1) Relevance Feedback
Label Query Items Relevant/Irrelevant
User supplies Labels
(2) Similarity Advisor
Rank Similarity Measures
Pattern-based Measures e.g., Color, Texture, Distribution, ...
Similarity Measure + Labels (L^+/L^-)
(3) Model Learning and -steering
Refine Decision Boundary
Classification
System formulates query
Selection
Aggregation
Visualization & Interaction
S1 Feature Selection
S2 Feature Aggregation and Weighting
S3 Statistical Analysis
S4 Similarity Search
S5 Subspace Cluster Analysis
S6 Data Aggregation and Weighting
S7 Data Selection
Data
Feature Space Visualization
Data Space Visualization
Selection
Drag & Drop
Focus+Context
A Framework for Relevance Model Bulding Using Pattern-Based Similarty Measures
A Formal Framework for the Dual Analysis of Feature and Data Space
Conclusion

Figure 1.2: This dissertation is structured into three parts of two chapters each. Part I covers *visualization-driven measures*, while Part II focuses on *data-driven measures*. Part III addresses the interplay with numerical data in *measure-driven frameworks*.

measures are crucial for optimizing the ordering of categories and dimensions within Parallel Sets, aiming to improve readability and facilitate pattern quantification. By establishing clear quality standards, we aim to improve the interpretability and utility of Parallel Sets, facilitating their use in communicating complex relationships to diverse audiences.

In addition to quality assessment, we also delve into the area of pattern quantification within categorical data projections [55]. Chapter 3 introduces methods for abstracting categorical data, enabling its representation in a "map metaphor" that facilitates easier orientation, navigation, and exploration. We contribute a new visualization technique for categorical data that overcomes the limitations of set-based or frequency-based analysis (e.g., Euler diagrams or Parallel Sets). The technique uses dimensionality reduction, based on defining the distance between two data elements as the number of different attributes, to allow for more nuanced exploration of data. It allows users to pre-attentively identify groups of similar data elements within the visualization. This feature is particularly valuable for exploring and understanding the structure and clustering within large categorical datasets. The technique allows to observe which attributes have a strong influence on the data embedding. This aspect is crucial for analysts to understand the factors that drive the grouping and separation of data points in the visualization. We propose two graph-based measures to quantify the visual quality of the plot. These measures rank attributes according to their contribution to cluster cohesion, providing metrics for evaluating the effectiveness of the visualization in revealing meaningful relationships between clusters and attributes. We evaluate our approach by comparing to traditional methods like Euler diagrams and Parallel Sets in terms of visual scalability. This comparison highlights the advantages of the new approach in handling large datasets with many category combinations. In addition, the usefulness and effectiveness of the Categorical Data Map is demonstrated by an expert study involving data scientists analyzing complex datasets (e.g., the Titanic and Mushroom datasets). The study confirms the advantages of the method, especially when analyzing large datasets with a high number of category combinations.

By automatically recommending views that highlight meaningful insights, these measures enhance the exploratory analysis process, allowing analysts to uncover and communicate key aspects of the data more efficiently. By providing tools for systematically evaluating and optimizing visualizations, we pave the way for more effective and insightful exploratory analysis of categorical data, improving the decision-making process and fostering a deeper understanding of complex datasets.

Part II: Data-driven Measures

This part details the application of measure-driven approaches for expressing properties of categorical data by deriving numerical measures through aggregation and domain knowledge to represent categorical items within a given application area, addressing research question (R2) by providing data-driven measures in domain-specific applications. Specifically, we focus on two distinct but equally complex domains: diachronic linguistics and software engineering. Both fields are rich in categorical data, from categorizing linguistic phenomena and language patterns to classifying software vulnerabilities and user feedback.

In chapter 4, we present the HistoBankVis application in the field of diachronic linguistics for the interactive analysis of large and complex categorical datasets. It is a novel visualization method specifically designed to support the interactive analysis of complex and multidimensional data within the context of linguistic research. HistoBankVis is tailored to facilitate the exploration and analysis of language change. Its design is geared towards uncovering the diachronic interactions among various linguistic factors, such as word order and subject case, particularly demonstrated through a case study on Icelandic. One of the technical contributions is the application of the Parallel Sets technique within HistoBankVis. This technique models complex interrelationships among linguistic factors, showcasing the system's ability to visualize and analyze multidimensional data effectively. Through the application of HistoBankVis to Icelandic linguistic data, the tool has demonstrated its powerful potential in aiding the understanding of the interaction among case, grammatical relations, and word order throughout the history of the Icelandic language. By enabling separate annotation, extraction, and comparison of linguistic data elements in a more streamlined and interactive manner, HistoBankVis contributes to improving the methodology of historical linguistics research. It reduces the need for painstaking pairwise comparisons by providing visual insights into complex data relationships.

In chapter 5, we present VulnEx (Vulnerability Explorer), a tool aimed at auditing software development organizations for third-party security risks associated with Open Source Software (OSS) use. VulnEx targets the crucial issue of managing and resolving potential security risks posed by third-party OSS components. Given the widespread use of OSS, identifying Common Vulnerabilities and Exposures (CVEs) within large software ecosystems is a pressing need for ensuring software security. CVEs are commonly classified into ordinal categories: “Low”, “Medium”, “High”, and “Critical”. The proposal of VulnEx as a tool to audit entire software development organizations represents a comprehensive approach to security analysis. Unlike more limited tools that may only analyze individual projects or components, VulnEx is designed to provide a holistic view of vulnerability exposures across the entire organization. A key contribution of VulnEx is its introduction of three complementary table-based representations specifically designed to facilitate the identification and assessment of OSS vulnerabilities. These representations allow users to effectively navigate and understand the landscape of vulnerability exposures within their organization. The design of VulnEx was conducted in close collaboration with security analysts. This collaborative approach ensures that the tool meets the practical needs of its users and addresses real-world challenges in software security analysis. VulnEx enables the examination of problematic projects and applications (repositories), third-party libraries, and specific vulnerabilities. This capability is crucial for prioritizing security efforts and directing resources toward the most critical areas of concern. We demonstrate the applicability of VulnEx through a use case and includes preliminary expert feedback. This feedback highlights the tool's potential effectiveness and value in identifying and managing security vulnerabilities within software organizations.

By demonstrating the effectiveness of measure-driven applications in the real-world contexts of software engineering and linguistics, we address the critical gap in the evaluation and application of data analysis techniques outside the domain of computer science. The adoption and application of measure-driven methodologies for categorical data in specific domains such as software engineering and linguistics underscores the potential of tailored data analysis tools to provide meaningful, real-world insights.

Part III: Measure-Driven Frameworks

This part examines how the techniques developed for categorical data can be integrated with and complement analyses involving other types of data, enhancing overall analytical capabilities. By addressing the interaction between categorical and numerical data, this part addresses (RQ3).

In chapter 6, the FDive approach contributes to the field of data analysis and pattern recognition in several ways: FDive introduces a novel approach that combines visual analytics with active learning. This integration assists users in creating relevance models that are not only accurate but also visually explorable, enhancing the interpretability of high-dimensional data. By employing a pattern-based similarity learning mechanism, FDive advances the methodology for assessing the relevance of data points. This allows for a more nuanced differentiation between relevant and irrelevant data based on user-provided categorical labels. Utilizing the best-ranked similarity measure, FDive calculates an interactive Self-Organizing Map (SOM)-based relevance model. This model classifies data according to cluster affiliations, providing a clear, visual representation of data groupings and their relevance. The approach includes a mechanism for soliciting additional user feedback by requesting labels for data elements with uncertain relevance classification. This feature enables the continuous refinement and improvement of the relevance model's accuracy based on user input. FDive identifies and highlights uncertain areas, especially near decision boundaries within the data. This allows users to focus their attention and refinement efforts on the most ambiguous parts of the model, enhancing model precision through targeted feedback. We demonstrate the effectiveness of our approach through a comparative evaluation with state-of-the-art feature selection techniques and a practical case study involving the classification of Electron Microscopy (EM) images of brain cells. The approach is shown to enhance both the quality and understanding of relevance models, potentially leading to new insights in specific research areas such as brain research.

In chapter 7, we provide a comprehensive review of existing dual analysis methods across various domains, such as medicine, crime analysis, and biology. This review helps in understanding the current landscape of dual analysis techniques. A major contribution is the development of a unified theoretical framework for dual analysis. This framework integrates the diverse approaches to dual analysis into a cohesive model, addressing the gap created by the varying definitions and implementations of dual analysis in existing research. We formalize the interactions between the three key components of dual analysis: the visualization of feature summaries, the visualization of data records, and the bidirectional linking of both visualizations through human interaction. This formalization clarifies how each component contributes to the overall analysis process and enhances the understanding of dual analysis. By categorizing existing dual analysis approaches within the proposed framework, our approach offers a structured overview of the field. This categorization not only helps in identifying the strengths and weaknesses of current methods but also in understanding how they fit within the broader context of dual analysis. Our framework reveals multiple components and processing steps in which the analysis of feature and data space can leverage categorical data, specifically for unsupervised learning methods where using categorical labels is less common. We identify and outline future research directions to further advance dual analysis. Specifically, it suggests incorporating state-of-the-art visual analysis techniques, such as user guidance and subspace detection algorithms, to improve data exploration. This contribution can guide subsequent research efforts and for push the boundaries of what is currently possible with dual analysis. Through its contributions, the framework aims to improve the exploration of large high-dimensional data by enabling more effective interactions for the joint analysis of features and data.

Both frameworks underscore the versatility and necessity of incorporating categorical data-driven measures into analytical models, whether the approach is

Computer Science Fields	Ch. 2	Ch. 3	Ch. 4	Ch. 5	Ch. 6	Ch. 7
Visual Analytics	●○○	●●○	●●○	●●○	●●●	●●●
Information Visualization	●●○	●●●	●●○	●●○	●○○	●○○
Evaluation	●●○	●○○	●○○	●○○	●●○	●●○
Applications	●○○	●○○	●●●	●●●	●●○	●○○

Table 1.1: The relevance of the contributions of each chapter of this thesis for different sub-fields of computer science. Rating scale: some relevance ●○○, medium relevance ●●○, high relevance ●●●.

supervised or unsupervised. The supervised framework, with its application in neurology, demonstrates the potential of these measures to improve the specificity and sensitivity of classification tasks in highly specialized domains. Meanwhile, the unsupervised framework illustrates how these measures can transform exploratory data analysis, enabling the discovery of new insights and patterns that can inform further research or operational strategies. The introduction of these frameworks not only fills a critical gap in the current data analysis landscape, but also sets the stage for future research and development.

In summary, this thesis contributes measures and frameworks for improving the visual analysis of categorical data. We demonstrate the effectiveness of our approaches through quantitative and qualitative evaluations, while user-centered approaches are evaluated through domain expert studies, typically conducted in pair-analytic sessions [63, 157, 169]. This dissertation is primarily focused on VA and InfoVis. However, it also makes significant contributions to other areas within computer science. A comprehensive comparison between the individual chapters with regard to the focus of the contributions can be found in table 1.1.

Citation Rules and Good Scientific Practices

This dissertation adheres to the established scientific practices and standards of the computer science research community. The main contributions have been disseminated through publications in journals, conferences and workshop proceedings and have thus undergone the peer review process. I retain the copyright to these publications, which form the basis of this dissertation. Sections of the dissertation that reflect content from my publications were either directly written by me or adapted by me in the course of writing the papers or the dissertation itself.

I am committed to maintaining complete transparency regarding the origins of each chapter of my thesis in order to allay any concerns regarding plagiarism and self-plagiarism. In section 1.3, I have listed all publications that I have either authored or co-authored. In addition, for each publication, I identify the contributors and describe how tasks were distributed among all contributors to ensure clarity about the collaborative nature of the work. At the beginning of each chapter, I acknowledge the publications from which text and figures have been used or modified. In incorporating these works, I adhere to the following principles:

1. Paragraphs in quotation marks are not authored by me, but contain contributions from other authors, and clearly indicate material that is not my original work.

2. Certain chapters of this dissertation are *taken from* my own publications, where I was the author or made significant revisions. These sections are slightly modified in their wording to fit the overall context of the thesis, but they emphasize my original contributions to the field of research.

3. This dissertation contains chapters *based on* my publications, or parts of them, for which I was either the sole author or co-author. These sections have been extensively rewritten and modified to fit the goals and scope of the dissertation. Nevertheless, they represent my authentic contributions to the field of research.

I aim to strike a balance between developing a thesis that is both clear and reader-friendly, achieved through careful editing and revision of my peer-reviewed articles, and adhering to rigorous citation practices, namely by accurately citing all material that comes from a publication. My decision to prioritize content, the contributions within it, and the reader experience stems from my belief that these are the fundamental aspects of scholarly work.

1.3 Publications

In this section, I outline and clarify my specific contributions to the publications that support this thesis. Recognizing that these publications were the result of teamwork, I will break down my individual contributions for each. These publications and manuscripts are the foundational elements of this dissertation:

- [87] **Frederik L. Dennig**, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel A. Keim, and Maximilian T. Fischer. “The Categorical Data Map: A Multidimensional Scaling-Based Approach”. In: *Proceedings of the 2024 IEEE Visualization in Data Science Symposium (to appear)*. IEEE, 2024.

  Contribution Clarification: I developed the idea and concepts of the manuscript. I directed the project and formulated the overall structure and research goals. In addition, I have outlined the fundamental research questions, outlined the contributions, and detailed the relevant background literature. I drafted and consolidated the manuscript, devising the conceptual framework, its discoveries, and the subsequent discussions. The replication data associated with this publication has been curated by me and is accessible in Data Repository of the University of Stuttgart (DaRUS) [86]. Lucas Joos provided feedback on the prototype and ideas regarding the glyph representations. Daniela Blumberg implemented an early prototype and provided feedback on the manuscript. Patrick Paetzold and Maximilian T. Fischer provided critiques and recommendations on the text during several revisions of the manuscript. Oliver Deussen and Daniel A. Keim commented on drafts and helped to revise the sections by providing helpful feedback. I am the sole author of all sections, incorporating and revising the suggestions of my co-authors several times. Thus, I reuse the text of that manuscript in chapter 3.

- [88] **Frederik L. Dennig**, Matthias Miller, Daniel A. Keim, and Mennatallah El-Assady. "FS/DS: A Theoretical Framework for the Dual Analysis of Feature Space and Data Space". In: *IEEE Transactions on Visualization and Computer Graphics* 30.8 (2024), pp. 5165–5182. DOI: 10.1109/TVCG.2023.3288356.

  Contribution Clarification: This paper is the result of a discussion between Mennathallah El-Assady and myself. I took the lead on the project and created the overall framework. In addition, I formulated the research question and delineated the contributions. I conducted the survey, analyzed the results, and discussed the findings. I developed the conceptual framework in its entirety, including all related formalizations. I also evaluated our framework. Matthias Miller provided parts of an early draft of the related work chapter and feedback on paper drafts. Daniel A. Keim and Mennathallah El-Assady provided feedback on the general idea and commented on paper drafts. I wrote and revised all sections myself, incorporating feedback from the co-authors. I restructured and adapted the related work section during the writing process. Therefore, I use this text in chapter 7.

- [84] **Frederik L. Dennig**, Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. "ParSetgnostics: Quality Metrics for Parallel Sets". In: *Computer Graphics Forum* 40.3 (2021), pp. 375–386. DOI: 10.1111/cgf.14314.

  Contribution Clarification: This project was initiated after a discussion between Johannes Fuchs and myself. I formulated the research question, identified the contributions, and conducted the evaluations. I developed all components, including the quality measures and the analytical front-end, and executed all necessary analyses. The data associated with this publication was collected or created by me and is accessible in DaRUS [83]. I wrote the entire manuscript, with each paragraph going through several rounds of revision throughout the writing process. Feedback on the development of the measures, especially their mathematical expressions, was provided by Maximilian T. Fischer. Michael Blumenschein and Evanthia Dimara provided guidance on the project of this paper and contributed insights on the overall concept and design of the evaluation. Johannes Fuchs and Daniel A. Keim also shared their perspectives on the overall concept and provided comments on drafts of the paper. Therefore, I use this text in chapter 2.

- [82] **Frederik L. Dennig**, Eren Cakmak, Henrik Plate, and Daniel A. Keim. "VulnEx: Exploring Open-Source Software Vulnerabilities in Large Development Organizations to Understand Risk Exposure". In: *Proceedings of the IEEE Symposium on Visualization for Cyber Security*. IEEE, 2021, pp. 79–83. DOI: `10.1109/VizSec53666.2021.00014`.

  Contribution Clarification: This paper emerged from a collaboration between Eren Cakmak, Henrik Plate, and myself. I was responsible for determining the research question and defining the contributions. I developed the demonstrator. I wrote the entirety of the manuscript, from structuring and initial drafting to revising sections, while incorporating feedback from the co-authors. Eren Cakmak prepared an initial draft of the related work section and provided critical feedback on the draft. Henrik Plate provided a compelling use case and facilitated discussions with several SAP domain experts to support our evaluation and provide input on drafts of the paper. Daniel A. Keim provided feedback on the general idea and commented on paper drafts. The evaluation process was a collaborative effort involving Eren Cakmak, Henrik Plate and myself. I did all the writing, including the evaluation. As a result, I incorporate this content into Chapter chapter 5.

- [89] **Frederik L. Dennig**, Tom Polk, Zudi Lin, Tobias Schreck, Hanspeter Pfister, and Michael Behrisch. "FDive: Learning Relevance Models Using Pattern-based Similarity Measures". In: *Proceedings of the 14th IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2019, pp. 69–80. DOI: `10.1109/VAST47406.2019.8986940`.

  Contribution Clarification: This paper was a result of a close collaboration between Michael Behrisch and myself. I wrote all the sections, identified the contributors, and performed the evaluations. I designed and implemented a research prototype that elaborated on a project I created during my undergraduate studies and addressed new research questions that arose. I evaluated the results by conducting both an expert study and a quantitative analysis. Tom Polk provided feedback on the draft of the paper. Hanspeter Pfister and Tobias Schreck shared their perspectives on the concept and provided feedback on several drafts. Michael Behrisch introduced the initial research challenge and has overseen this project, providing guidance on conceptual direction and reviewing drafts of the paper. Zudi Lin supplied the data used for evaluation and provided feedback on the evaluation section regarding the domain. I was responsible for writing and revising the paper throughout the drafting phase. Therefore, I am including this content in chapter 6.

- [271] Christin Schätzle, **Frederik L. Dennig**, Michael Blumenschein, Daniel A. Keim, and Miriam Butt. "Visualizing Linguistic Change as Dimension Interactions". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Aug. 2019, pp. 272–278. DOI: `10.18653/v1/W19-4734`.

  Contribution Clarification: This publication is the result of a close collaboration with Christin Schätzle and Michael Blumenschein. I developed the research prototype, provided technical descriptions of our approach, and provided feedback on the paper. The paper was written by Christin Schätzle. Michael Blumenschein and I contributed to the technical explanations and provided input on the paper. Daniel A. Keim and Miriam Butt provided feedback on the general idea and commented on paper drafts. As a result, I rewrote and adapted all the sections, adding additional content and elaborating on the technical contributions. Chapter 4 contains concepts from this publication, rephrased and expanded.

- [272] Christin Schätzle, Michael Hund, **Frederik L. Dennig**, Miriam Butt, and Daniel A. Keim. "HistoBankVis: Detecting Language Change via Data Visualization". In: *Proceedings of the NoDaLiDa 2017 Workshop Processing Historical Language*. NEALT Proceedings Series 32. Association for Computational Linguistics, 2017, pp. 32–39.

  Contribution Clarification: This work was created in close collaboration with Christin Schätzle and Michael Blumenschein. I developed the research prototype, detailed the technical aspects of our methodology, and reviewed the paper. Christin Schätzle was the main author, focusing on the historical development of the Icelandic language. Therefore, in this dissertation, I have restructured and rewritten all sections and integrated them with additional content. Michael Blumenschein and I were involved in fleshing out the technical descriptions and providing feedback on the manuscript. Daniel A. Keim and Miriam Butt provided insight into the overall concept and critiqued several drafts. Consequently, I have rephrased and expanded the content in chapter 4, drawing on that paper, but with an increased emphasis on the contributions to computer science.

During my time at the Data Analysis and Visualization group, I have co-authored and contributed to nine other publications. These papers contribute to areas and topics related to the research questions explored in this thesis. However, as they do not directly contribute to the main discourse of the thesis, they have not been included. The publications are listed below:

- [161] Mohsen Jenadeleh, **Frederik L. Dennig**, Rene Cutura, Quynh Quang Ngo, Daniel A. Keim, Michael Sedlmair, and Dietmar Saupe. "An Image Quality Dataset with Triplet Comparisons for Multi-dimensional Scaling". In: *Proceedings of the 16th International Conference on Quality of Multimedia Experience*. IEEE, 2024, pp. 278–281. DOI: 10.1109/QoMEX61742.2024.10598258.

- [45] Daniela Blumberg, Yu Wang, Alexandru Telea, Daniel A. Keim, and **Frederik L. Dennig**. "Inverting Multidimensional Scaling Projections Using Data Point Multilateration". In: *Proceedings of the 15th International EuroVis Workshop on Visual Analytics*. Eurographics, 2024. DOI: 10.2312/eurova.20241112.

- [58] Raphael Buchmüller, Bastian Jäckl, Michael Behrisch, Daniel A. Keim, and **Frederik L. Dennig**. "cPro: Circular Projections Using Gradient Descent". In: *Proceedings of the 15th International EuroVis Workshop on Visual Analytics*. Eurographics, 2024. DOI: 10.2312/eurova.20241111.

- [115] Johannes Fuchs, **Frederik L. Dennig**, Maria-Viktoria Heinle, Daniel A. Keim, and Sara Di Bartolomeo. "Exploring the Design Space of BioFabric Visualization for Multivariate Network Analysis". In: *Computer Graphics Forum* 43.3 (2024). DOI: 10.1111/CGF.15079.

- [255] Nils Rodrigues, **Frederik L. Dennig**, Vincent Brandt, Daniel A. Keim, and Daniel Weiskopf. "Comparative Evaluation of Animated Scatter Plot Transitions". In: *IEEE Transactions on Visualization and Computer Graphics* 30.6 (2024), pp. 2929–2941. DOI: 10.1109/TVCG.2024.3388558.

- [168] Lucas Joos, Karsten Klein, Maximilian T. Fischer, **Frederik L. Dennig**, Daniel A. Keim, and Michael Krone. "Exploring Trajectory Data in Augmented Reality: A Comparative Study of Interaction Modalities". In: *Proceedings of the 2023 ISMAR International Symposium on Mixed and Augmented Reality*. IEEE, 2023, pp. 790–799. DOI: 10.1109/ISMAR59233.2023.00094

- [230] Quynh Quang Ngo, **Frederik L. Dennig**, Daniel A. Keim, and Michael Sedlmair. "Machine learning meets visualization – Experiences and lessons learned". In: *it - Information Technology* 64.4-5 (2022), pp. 169–180. DOI: 10.1515/itit-2022-0034.

- [111] Maximilian T. Fischer, **Frederik L. Dennig**, Daniel Seebacher, Daniel A. Keim, and Mennatallah El-Assady. "Communication Analysis through Visual

Analytics: Current Practices, Challenges, and New Frontiers”. In: *Proceedings of the 2022 IEEE Visualization in Data Science Symposium*. IEEE, Oct. 2022. DOI: 10.1109/VDS57266.2022.00006.

- [250] David Pomerenke, **Frederik L. Dennig**, Daniel A. Keim, Johannes Fuchs, and Michael Blumenschein. “Slope-Dependent Rendering of Parallel Coordinates to Reduce Density Distortion and Ghost Clusters”. In: *Proceedings of the 30th IEEE Visualization Conference*. IEEE, 2019, pp. 86–90. DOI: 10.1109/VISUAL.2019.8933706.

1.4 Open Science and Replication Data

To increase transparency, reproducibility, and trust in scientific research, I openly share the data and code associated with my publications, wherever copyright considerations permit. *Replication data* allows researchers to verify the results of studies by re-analyzing the data using the original methods. This is a critical step in confirming the reliability and validity of scientific findings, as it helps to uncover potential errors or biases. *Open science* aims to make scientific research and dissemination accessible to all levels of an inquiring society, whether amateur or professional. It fosters collaboration and information sharing, ensuring that scientific knowledge is accessible to all, accelerating discovery and innovation. Together, replication data and open science are fundamental to building a more robust, inclusive, and democratic scientific community, where knowledge is shared openly and credibility is ensured through collective verification of results. Thus, I have created and contributed the following replication data associated with the above mentioned publications:

- [86] **Frederik L. Dennig**, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel Keim, and Maximilian T. Fischer. *The Categorical Data Map - Replication Data*. Version V1. https://osf.io/jzd46/ (alternative repository). 2024. DOI: 10.18419/darus-3372.
- [83] **Frederik L. Dennig**, Maximilian T. Fischer, Michael Blumenschein, Daniel Fuchs Johannes; Keim, and Evanthia Dimara. *Replication Data for: "ParSetgnostics: Quality Metrics for Parallel Sets"*. Version V1. https://osf.io/rwhf5/ (alternative repository). 2022. DOI: 10.18419/darus-2869.
- [249] David Pomerenke, **Frederik L. Dennig**, Daniel A. Keim, Johannes Fuchs, and Michael Blumenschein. *Replication Data for: "Slope-Dependent Rendering of Parallel Coordinates to Reduce Density Distortion and Ghost Clusters"*. Version V2. https://osf.io/sy3dv/ (alternative repository). 2022. DOI: 10.18419/darus-3060

Part I

Visualization-Driven Measures

Meten is weten
(To measure is to know)

— **Dutch Proverb**

2 Quality Measures for Parallel Sets Visualizations

Contents

While there are many visualization techniques for exploring numeric data, only a few work with categorical data. One prominent example is Parallel Sets, showing data frequencies instead of data points – analogous to parallel coordinates for numerical data. As nominal data, a subset of categorical data, does not have an intrinsic order, the design of Parallel Sets is sensitive to visual clutter due to overlaps, crossings, and subdivision of ribbons hindering readability and pattern quantification. In this chapter, we propose a set of quality measures, called ParSetgnostics, which aim to improve Parallel Sets by reducing clutter. These quality measures quantify important properties of Parallel Sets such as overlap, orthogonality, ribbon width variance, and mutual information to optimize the attribute and category ordering. By conducting a systematic correlation analysis between the individual measures, we ensure their distinctiveness. Further, we evaluate the clutter reduction effect of ParSetgnostics by reconstructing six datasets from previous publications using Parallel Sets measuring and comparing their respective properties. Our results show that ParSetgnostics facilitates multi-attribute analysis of categorical data by automatically providing optimized Parallel Set designs with a clutter reduction of up to 81% compared to the originally proposed Parallel Sets visualizations.

This chapter is *taken from* the following publication:

- [84] **Frederik L. Dennig**, Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. "ParSetgnostics: Quality Metrics for Parallel Sets". In: *Computer Graphics Forum* 40.3 (2021), pp. 375–386. DOI: 10.1111/cgf.14314.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

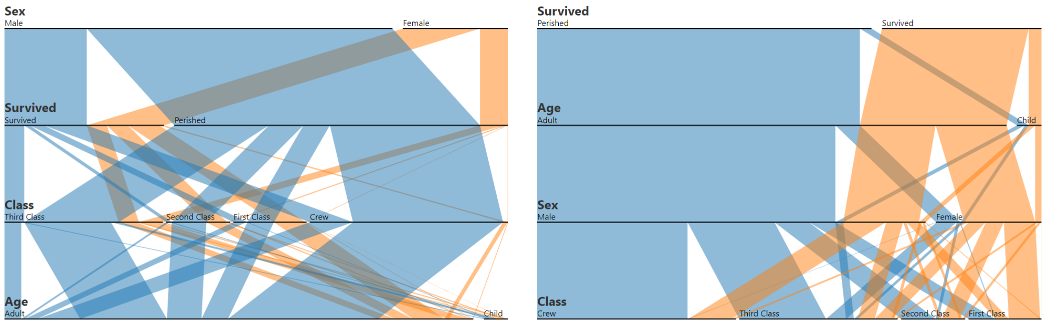


Figure 2.1: Two Parallel Sets of the Titanic dataset [80]. The right visualization has less ribbon overlap than the one on the left and, thus, is easier to read because of the less clutter.

2.1 Challenges Using Parallel Sets for Categorical Data Visualization

Categorical data is an inherent data type in many real-world datasets. Examples include business intelligence, when assigning personnel to tasks and resources, or inventory data, when describing product qualities like color. However, most multi-dimensional visualization techniques, such as scatterplot matrices [139, 74], parallel coordinates [156], and projections [64], are designed for numerical data, where data values come with a meaningful scale or ordering. In contrast, nominal data, a subset of categorical data, does not have an intrinsic ordering or distance between the values. Instead, it describes properties in name only, requiring context for analysis. Frequency-based visualizations [148, 331, 292] are a possible solution mapping categorical variables to their corresponding frequencies. Yet, for most techniques, the frequency information is often not visible or imposes a hierarchical structure. On the other hand, solutions that treat dimensions independently [261, 300, 164], mapping categories to numbers, follow a continuous design model which deviates from the discrete mental user model of the data [190]. The Parallel Sets visualization is a hybrid solution combining the strengths of frequency-based designs with the independent treatment of dimensions, which is essential for multi-attribute analysis of categorical data [35, 189].

To support multi-attribute analysis of categorical data, Parallel Sets appropriate the layout of parallel coordinates [156]. They replace the polylines representing numerical data points with parallelograms, called ribbons, with their sizes representing the frequency of the categories. Parallel Sets serve as an interaction framework used in various fields that require user-driven analysis of heterogeneous and multi-attribute categorical data. Compared to other visualization types, Parallel Sets offer fewer degrees of freedom with respect to design considerations, making them a compelling solution for the challenging representation of categorical data. In contrast, Sankey diagrams [181] exhibit more degrees of freedom, such as placing attribute axes or sections of them freely on the chart, while stacked bars can present the same data without the explicit links between the individual values of each attribute. However, as nominal data does not have an intrinsic order, the readability of Parallel Sets depends on the chosen ordering of attributes, as well as the ordering of categories within each attribute. Certain attribute and category orderings are more challenging to read than others. Figure 2.1 shows two Parallel Sets of the same data. The left Parallel Sets visualization appears is to read due to the high degree of clutter. On the right, an alternative reordering with minimized ribbon overlap is easier to read.

Identifying the optimal data representation with Parallel Sets can be challenging. Parallel coordinates can be applied to categorical data. However, the frequency information is lost. For exploratory scenarios, choosing an adequate Parallel Sets configuration for the dataset is key to the understanding and knowledge gained in the process. Manual reordering of attributes is not always feasible due to the large set of possible attribute and category orderings. We note that the number of possible configurations exceeds those of parallel coordinates because the order of categories can be chosen freely. There are $|\mathcal{C}_i|!$ possible category orderings of an attribute axis, where $\mathcal{C}_i$ are the category values of attribute $a_i \in \mathcal{A}$. The attribute axes themselves can be reordered and allow for $|\mathcal{A}|!$ orderings. Thus, there are a total of $|\mathcal{A}|! \cdot \prod_{a_i \in \mathcal{A}} |\mathcal{C}_i|!$ possible Parallel Sets visualizations. Existing approaches focus on interaction [338], which requires user interaction and suffers from summarization that loses information and imposes a biased first view [143] by reducing the number of attributes and categories. Automatic solutions for designing Parallel Sets do not sufficiently support data analysis in fully exploratory scenarios because they limit the number of attributes for the displayed subsets [9]. Thus, these approaches often exclude possibly relevant information beforehand. In this chapter, we contribute:

Contributions

- Eight *quality measures* for Parallel Sets aiding their comparison and ranking in-terms of clutter and readability.
- A *formalization* of geometric properties of Parallel Sets underlying our measures and a discussion of parameters.
- An *evaluation* of our measures by applying our technique to six datasets from previous publications.
- For *accessibility*, we provide the ParSetgnostics Explorer at dennig.dbvis.de/parsetgnostics making our results interactively explorable.
- For *reproducibility*, we make all our statistical analysis, results, and source code available on OSF (osf.io/rwhf5) and DaRUS [83].

2.2 Related Work

In this section, we discuss related techniques and differentiate our approach from other strategies and methods for the optimization, quality assessment and visualization of categorical data.

2.2.1 Visualizations for Categorical Data

There are many visualization techniques for categorical data. In the context of flows, sets, and subsets, Sankey diagrams are common in the analysis of categorical data [205, 253]. Their first appearance described the optimization of steam engines, visualizing the energy flow of a steam engine [181]. They have also been used to visualize relations between subsets [3, 205]. However, Sankey diagrams are very flexible in their layout and design, often requiring a visualization designer to determine a useful visualization. In contrast, Parallel Sets, which can be classified as a type of Sankey diagram, is more restricted in their layout, limiting the degrees of freedom of their design space to the category- and dimension-ordering and color choice, but still requiring manual interaction. The main focus of Parallel Sets is on sets and subset relationships [190]. Other types of visualizations for sets and subsets exist, such as the Icicle plot [195] and the Sunburst [294] diagram. However, they are limited to the visualization of hierarchical or clustered data, i.e., predefined subsets. Other visualization techniques are purely frequency-based by mapping categorical variables to their corresponding frequencies, missing the capability for analyzing subset relationships [148, 331, 292]. All these types of visualization techniques are common in the visualization community [50]. Hammock Plots [275] combine parallel coordinates with Parallel Sets to allow for the analysis of datasets with numeric and categorical data.

2.2.2 Improvements of Parallel Sets

Parallel Sets can be improved through visual approaches. These techniques change the representation of ribbons to make them easier to follow. A common visual method for improving the readability of Parallel Sets in this way is to curve the ribbons of Parallel Sets [256]. Another technique is to draw ribbons with a fixed angle, called Common Angle Plots [150], yielding better readability. This technique addresses the effects of a class of perceptual illusions, called Müller-Lyer illusion [81, 121], where lines appear to have a different distance or length. Our approach differs from these techniques in that we propose a different layout of coordinate axes and categories. Techniques changing the representation of ribbons can be applied after our quality measures have been used to determine a useful dimension- and category ordering, further improving the readability. There also exist a set of dimension ordering strategies for parallel coordinates [46], which can apply to Parallel Sets if modified. Parallel Sets can also be improved in a semi-automatic way, using machine learning or statistical methods. The interactive approach by Zhang et al. [338] uses association rule mining to reduce the number of dimensions and categories, requiring user interaction. The approach by Alsakran et al. [9] changes the layout and ordering of dimension axes but restricts the dimensionality of the subgroups, i.e., ribbons, to two dimensions. This approach simultaneously uses mutual information [278] to measure the dependence of two variables. Both techniques remove dimension information or data from the visualization. Our approach differs in that it does not remove any data and does not restrict the dimensionality of the displayed ribbons but tries to optimize a set of target properties.

2.2.3 Quality Measures for Visualization Techniques

Screen-space quality measures describe a set of measures specifically designed measures or features that measure the quality of visualization and can be used to optimize them for readability or quantify the appearance of specific patterns [33]. They do not remove any information from the visualization. They rather measure properties of the visualization, which can be used to compare and rank them. Examples of those approaches are: Magnostics for matrix visualizations [31], Scagnostics for scatterplots [330], Pargnostics for parallel coordinate plots [78], Visualgnostics projections of high-dimensional data [197], and Pixgnostics for pixel-based visualizations [274]. We contribute to this area of information visualization by providing a set of eight measures for the quantification of visual properties of Parallel Sets. In this way, we improve the quality of Parallel Sets without performing any sampling or dimensionality reduction of the underlying data.

2.3 Parameters of Parallel Sets

In this section, we provide the necessary definitions to describe the properties of Parallel Sets formally. We also discuss the parameters of Parallel Sets in light of semi-automatic and fully automatic reordering of attributes and categories.

2.3.1 Background

Parallel Sets are a visualization type for categorical data. An example of a Parallel Sets visualization is shown in figure 2.2. Parallel Sets show flow-paths that divide

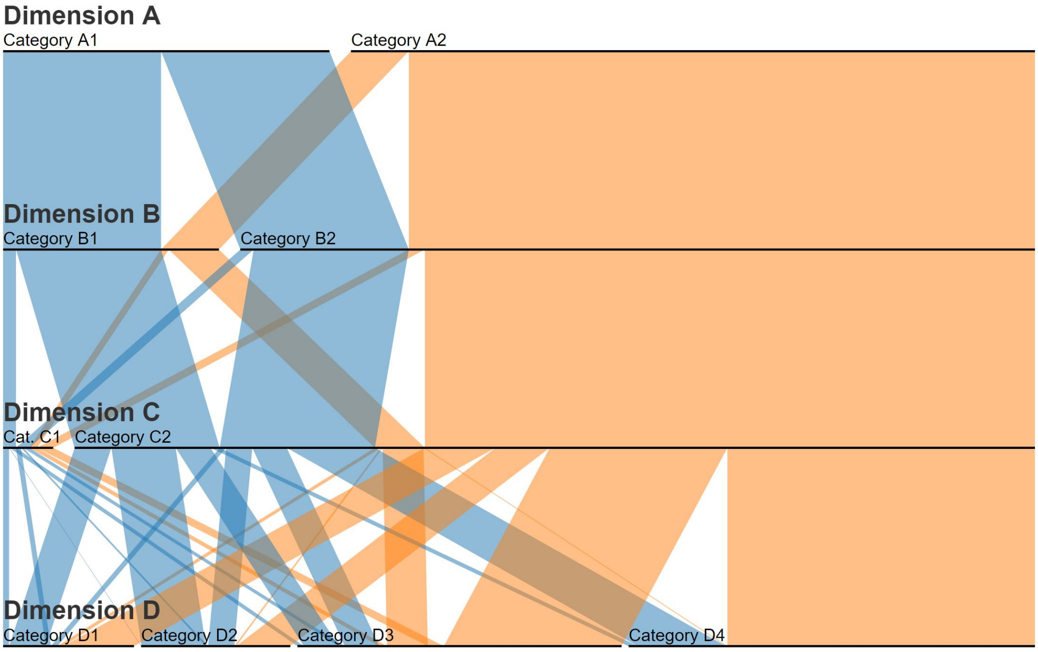


Figure 2.2: A Parallel Sets visualization of a generic dataset with four attributes (A-D) and their respective categories (of cardinality two for attributes A-C and four for attribute D).

the flow into smaller and smaller subsets at each category if an attribute splits the subset into multiple categories. This introduces a direction or flow, in the case of figure 2.2 from top to bottom, while also increasing granularity with each attribute axis splitting the dataset into smaller subsets. Every attribute is represented by an axis and a set of ribbons. Each ribbon represents a subset defined by the categories above and the one category connected to the following attribute axis. Compared to parallel coordinates, the individual categories on the attribute axis are not discrete points. Instead, the axis and the width of the ribbon are proportional in size to their flow, i.e., the number of data items with the corresponding categories they represent. They can be compared to stacked bars. However, stacked bars only display attributes that can show the same data without the explicit links between. Sankey diagrams exhibit more degrees of freedom, such as placing attribute axes or sections of them freely on the chart.

2.3.2 Definitions

This work aims to optimize a Parallel Sets visualization by ordering the attributes and categories to conform better to the design considerations described in the following. We developed our measures with the general idea of quality measures for information visualization described by Behrisch et al. [33] in mind. With the definition of a quality criterion (see equation (2.1)) provided in their work, the problem is described formally:

$$\arg \underset{\phi \in \Phi}{\overset{\min}{\max}} \; q(\phi | D, U, T) \tag{2.1}$$

D denotes the data, U the user, and T the task. ϕ denotes a specific configuration of a visualization of the set of all possible configurations of a given visualization type Φ. q describes a quality criterion and $\arg\max/\min_{\phi\in\Phi}$ optimization strategy. In this work, we focus on defining quality criteria q for Parallel Sets visualizations, i.e., a set of objective functions (see section 2.4). We test our quality measures with six datasets, which in this definition corresponds to D (see section 2.5.1). The measures can be task and user-dependent. The user can choose which quality measures he aims to minimize or maximize or even how to weight them. It is also possible to limit Φ by choosing a set of constraints, e.g., filtering or sampling. In our work, we consider the task T to be an exploration task with no prior knowledge of the specifics of the dataset. The result is ϕ, in our case, the configuration of a Parallel Sets visualization, defined by the order of attributes $\mathcal{A}$ and the order of categories $\mathcal{C}_i$ of all attributes $a_i \in \mathcal{A}$. We define $\mathcal{A}_{\mathsf{ord}}$ as the tuple of all attributes of a purely categorical dataset:

$$\mathcal{A}_{\mathsf{ord}} := (a_1, a_2, \dots, a_i) \tag{2.2}$$

Similarly, we define the ordering of the category values $\mathcal{C}_i$ of the i-th element a_i of $\mathcal{A}_{\mathsf{ord}}$ as:

$$\mathcal{C}_{\mathsf{ord}}(i) := (c_i^1, c_i^2, \dots, c_i^j) \tag{2.3}$$

where each element is a single category of a specific attribute. This is consistent with the tree-like structure of Parallel Sets [189], separating the dataset into smaller subsets while descending the tree levels, where each level represents an attribute axis defined by the order of elements of tuple $\mathcal{A}$. Ribbons are representatives of edges between two levels, i.e., connections between two adjacent attribute axes a_n and a_{n+1}. Thus, we can define the possible ribbons R_n^* between two adjacent attributes for $n \in [1, |\mathcal{A}| - 1]$ as:

$$R_n^* := \bigtimes_{i=1}^{n+1} \mathcal{C}_{\mathsf{ord}}(i) \tag{2.4}$$

Since R_n^* denotes all possible ribbons between two attributes, it includes empty subsets. Parallel Sets do not visualize empty or non-existent subsets. Thus, we remove such ribbons by verifying that at least one entry exists that belongs to a subset defined by a ribbon r, i.e., $|r| > 0$. This yields the set of all existing ribbons between two attribute axes, which we define as:

$$R_n := \{r \mid r \in R_n^* \ \wedge |r| > 0\} \tag{2.5}$$

Finally, we can define the set of all existing ribbons $\mathbf{R}$ and analogous the set of all possible ribbons $\mathbf{R}^*$ as:

$$\mathbf{R} := \bigcup_{i=1}^{|\mathcal{A}|-1} R_n \qquad\qquad \mathbf{R}^* := \bigcup_{i=1}^{|\mathcal{A}|-1} R_n^* \tag{2.6}$$

2.3.3 Parameter Space

In the next section, we will discuss the specific parameters and caveats of Parallel Sets related to the choice of the category and attribute ordering, dataset-dependent properties, and ribbon parameters. We will use those parameters to explain our measures described in section 2.4.

Selection of the First Attribute: The analysis task is *the* determining factor for the axes ordering. The first attribute and its categories determine the ribbon color, and thus the main aspects the analysis focuses on. In case there exists a formulated analysis question or hypothesis, we suggest determining this attribute beforehand or interactively. A partial ordering is possible. The user with domain knowledge can decide best which attributes are more important than others. In the case of an exploratory scenario, we suggest a fully automatic approach, generating multiple clutter reduced and readability improved versions with different axes orderings to allow for an overview of the dataset. We suggest choosing the first attribute based on the attribute with the highest entropy for a fully automatic approach, thus focusing on the attribute with the most significant amount of information. Thus, it is a attribute with balanced category sizes. Attributes with low entropy will contain more categories of less size, making them hard to read.

Ordering of Remaining Attributes: The following axes split the ribbons into increasingly fine-grained subsets, each split according to a attribute's categories. With the increasing amount of ribbons, clutter is likely to increase. The strength of this effect is ultimately dependent on the dataset. We identified two effects on the ribbons linked to this parameter: the number of ribbons and the ribbon widths. Firstly, the number of ribbons should be kept as low as possible to avoid premature splitting into subsets. Secondly, the ribbon widths should be kept as large as possible to keep them easy to follow. This properties is also influenced by the slope of the ribbon, dependent on the ordering of categories. In a fully automatic approach, the order can be determined by three strategies: (1) Order the attributes by ascending number of categories, minimizing the number of ribbons. (2) Maximize the ribbon width, lowering the number of thin ribbons, which are hard to perceive. (3) Ordering the attribute based on a information-theoretic property, such as mutual information [278].

Ordering of Categories: While there is no natural order among nominal values and the order of categories on each attribute can be chosen freely [35], not every category ordering is intuitive, useful, or supportive for exploratory or confirmatory data analysis. Some category orderings lead to a high degree of clutter by increasing the slope and overlap of ribbons. Therefore, the category ordering can be optimized such that the Parallel Sets visualization is readable and shows patterns inside the data, even with an increasing amount of ribbons caused by splits according to attribute axes. Since this parameter offers the most potential for improvement, five of the eight measures we define are sensitive to category reordering and are designed to help analysts in their choice of attribute and category ordering. However, given that some categorical data is ordinal, e.g., time, the sequence is fixed by the inherent order and should not be changed.

Impact of Number and Size of Categories: Attributes with many categories split the data into many small ribbons that are hard to follow. Additionally, since the number of ribbons monotonously increases with every attribute axis, this leads to an increased number of ribbons in every following attribute. The data distribution is the determining factor, i.e., attributes having a few categories of equal size, or the many small categories or a mixture thereof. The issue can be addressed by delaying splits yielding thin ribbons to later attributes, i.e., prioritizing attributes with large equal-sized categories. Such an attribute should be placed at the beginning of the attribute ordering.

Influence of the Distance Between Attribute Axes: A short distance increases the slope of diagonal ribbons, which increases the overlap of ribbons and clutter. Since ribbons are parallelograms, this reduces the perceived width [250]. In contrast, an excessively large distance makes ribbons, especially thin ones, hard to follow since they are visually less prominent due to their small surface area. Additionally, it decreases the crossing angle of ribbons, which makes them also harder to follow [155, 326]. This parameter is ultimately dependent on the available screen-space and its aspect ratio. Four category ordering-dependent measures, namely *Overlap*, *Slope*, *Orthogonality*, *Crossing Angle* are sensitive to this parameter. We fixed the distance between the attributes for all our measurements.

Impact of Ribbon Width and Plot Width: The width of the ribbons is dependent on the available plot space. In the case of a vertical ribbon flow, it depends on the plot width. For a horizontal ribbon flow, it will depend on the plot height. The width of all ribbons remains relative, as with the number of ribbons, the ribbons width decreases. The plot size should be chosen accordingly. All ribbons, especially those representing small subsets, should have a large enough width such that they can be visually compared and easy to follow. With increasing plot size, the distance between the attribute axes also increases. Four of our category ordering-dependent measures are sensitive to this parameter. Thus, we also choose a constant plot size for all our measurements.

Selection of Ribbon Colors: The ribbon color is not considered by our measures. However, we suggest choosing colors according to common criteria, i.e., easy to differentiate colors [223, 54]. Since the number of colors is equal to the number of categories of the first attribute, it is beneficial to reduce the number of colors by selecting an attribute with a low number of categories that is still pertaining to the analysis question. In exploratory tasks, we suggest an attribute with a category count no larger than nine based on Miller's Law [218, 219]. Parallel Sets are intrinsically "2.5D," meaning that the ribbons can have an ordering along the depth direction. The typical solution to avoid occlusion is to use transparency to show the path and area of overlapping ribbons. In this case, the colors of ribbons need to be chosen such that the mixtures of colors produce a distinguishable color that still implies which ribbons are crossing. If no transparency is used, we suggest ordering the ribbon, such that the thinner ribbons are on top to minimize occlusion.

2.4 Quality Measures

This section describes and discusses a set of eight quality measures that measure different properties of Parallel Sets. These properties are dependent on the attribute and category ordering. These properties are either desirable or undesirable, and thus, our measures can be used to compare Parallel Sets and help adjust them to be more readable and interpretable. For explanation and comparability, we use the Titanic dataset [80] to show-case their effects.

2.4.1 Category Ordering-Dependent Measures

We present five category ordering-dependent measures, which means that they are sensitive to the reordering of attributes and individual categories of an attribute. Small changes in the order of categories can already have a large impact on the appearance of a Parallel Sets visualization.

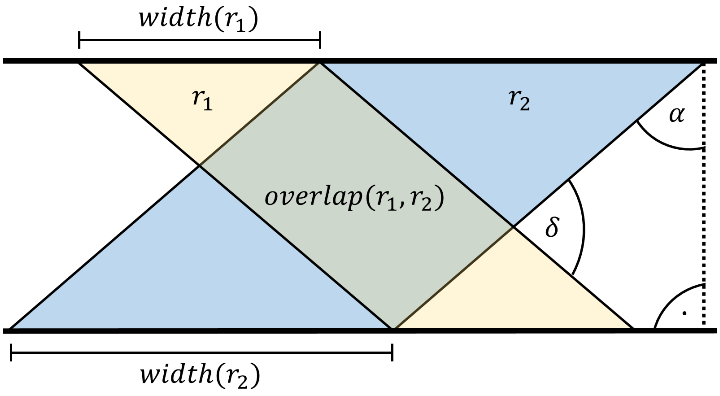


Figure 2.3: This figure shows the geometric variables required for our measures. It shows two ribbons r_1 and r_2 between two attribute axes. $overlap(r_1, r_2)$ defines as their shared area. The angle α denotes the slope of a ribbon. An orthogonal ribbon has a slope of $\alpha = 0$. δ describes the crossing angle of r_1 and r_2. The width of a ribbon is the distance of the intersections with an attribute axis.

Three category ordering-dependent measures consider the relationships of pairs of ribbons between two attribute axes. We describe this set as follows:

$$P_i := \{(r_1, r_2) | (r_1, r_2) \in R_i \times R_i \wedge r_1 \neq r_2\} \tag{2.7}$$

The set P_i describes all possible pairs of ribbons between the attributes a_i and a_{i+1} and is required for the *Overlap* and *Number of Crossings* and *Crossing Angle* measures.

Overlap O measures the overlapping area of all ribbons. A high overlap is indicative of clutter since overlapping areas are harder to interpret, since crossing ribbons are harder to follow [155, 326]. Furthermore, there is a connection to the slope of a ribbon as only sloped ribbons contribute to overlap. The overlap is especially high if large subsets overlap in their ribbon representation. We formally describe this measure in equation (2.8).

Definition: Overlap Measure

$$\textsc{Overlap} := \frac{1}{S} \sum_{i=1}^{|\mathcal{A}|-1} \sum_{(r_1, r_2) \in P_i} overlap(r_1, r_2) \tag{2.8}$$

The set of tuples P_i defines all possible pairs of ribbons between two neighboring attribute axes of the Parallel Sets. S denotes the area of the Parallel Sets visualization. The factor $\frac{1}{S}$ allows for the comparability of different Parallel Sets visualizations on different resolutions. $overlap(r_1, r_2)$ with $r_1, r_2 \in \mathbf{R}$ defines the overlapping area of two ribbons as described in figure 2.3. The examples shown in figure 2.4 show the effects of reducing the overlap of ribbons yielding a Parallel Sets visualization with a low degree of clutter.

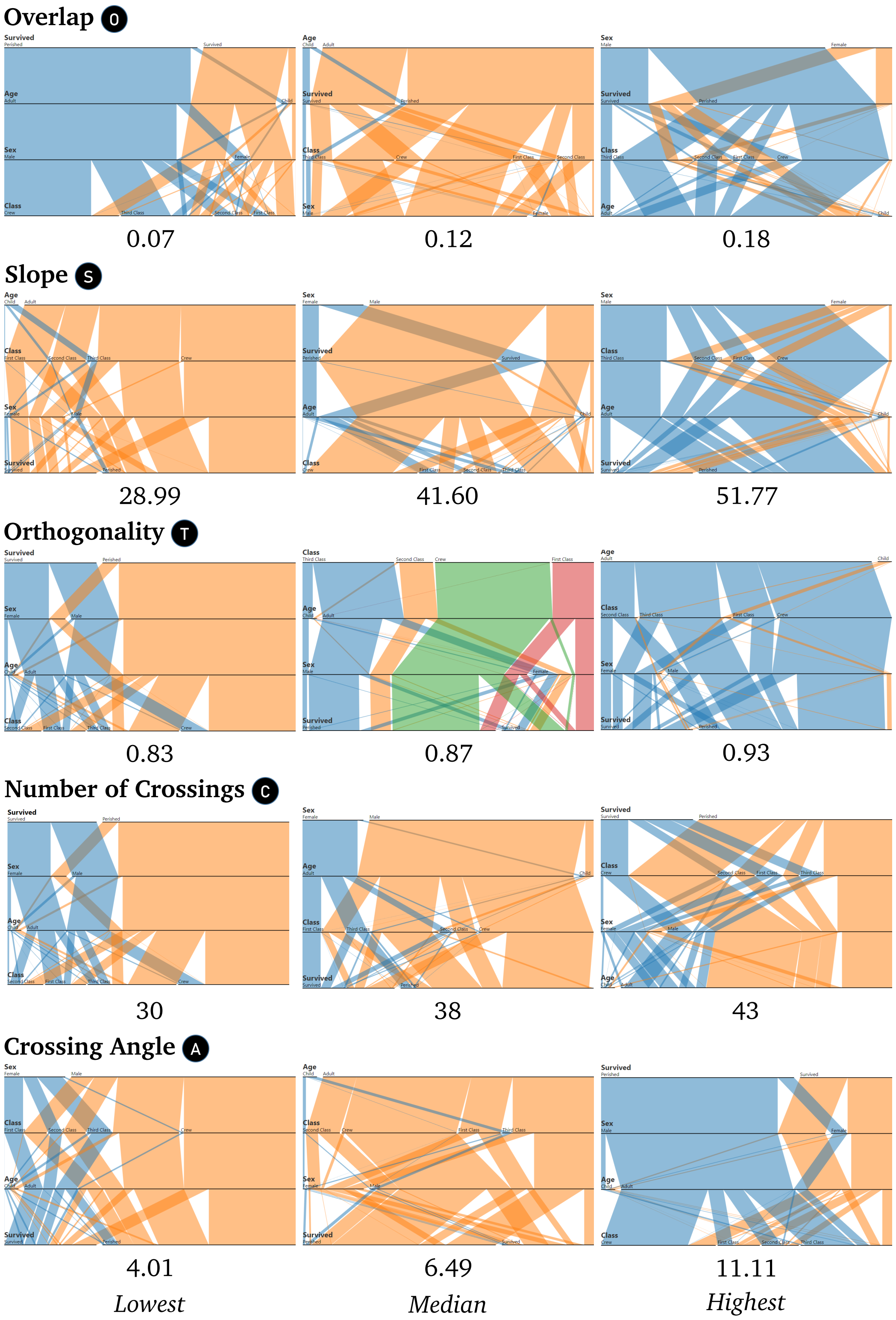


Figure 2.4: We show three Parallel Sets visualizations for each of the five category ordering-dependent measures: *Overlap* Ⓞ, *Slope* Ⓢ, *Orthogonality* Ⓣ, *Number of Crossings* Ⓒ, and *Crossing Angle* Ⓐ. We show the Parallel Sets corresponding to the *lowest*, *median*, and *highest* measure value. Lower values signify less clutter and thus improved readability, presenting a good starting point for exploratory data analysis.

Slope Ⓢ measures the average slope of all ribbons. A low average slope is preferable since ribbons that have a high angle to the attribute axes are easier to follow [155, 326]. This is grounded in the area preserving geometrical properties of parallelograms. Highly sloped ribbons get thinner and longer [250]. Only sloped ribbons contribute to overlap. The *Slope* measure differs from the *Overlap* measure in that it is not affected by the ribbon width, meaning that the *Slope* measure is not weighting the slope by the size of the subset that the ribbon represents. We formally describe this measure in equation (2.9).

Definition: Slope Measure

$$\textsc{Slope} := \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \alpha(r) \tag{2.9}$$

In this equation, the slope of a ribbon is denoted by angle α, which is geometrically defined as depicted in figure 2.3. The effects of minimizing the *Slope* measure can be observed in figure 2.4. A low average slope reduces clutter, while high *Slope* introduces a noticeable zigzag pattern which is hard to interpret.

Orthogonality Ⓣ leverages the concept to the *Slope* measure but explicitly focuses on the orthogonality of ribbons. This focus restricts the layout of ribbons to enforce a close to a perpendicular angle to the attribute axis. This property increases readabilty [155, 326]. It measures the average number of ribbons with a slope α smaller than a threshold value τ. We formally describe this measure in equation (2.10).

Definition: Orthogonality Measure

$$\begin{aligned} \textsc{Orthogonality} &:= \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} f_{\text{orthogonal}}(r_1, r_2) \\ \text{where} \quad f_{\text{orthogonal}}(r_1, r_2) &:= \begin{cases} 1 : \alpha(r) > \tau \\ 0 : \alpha(r) \leq \tau \end{cases} \end{aligned} \tag{2.10}$$

A group of ribbons that is perpendicular to the attribute axes shows a categorical correlation. Therefore, we choose $\tau = 0$. However, τ can be chosen with respect to the target orthogonality, such that slightly sloped ribbons are also considered. In figure 2.4, we can see that enforcement of perpendicular ribbons, forming rectangles, reduces clutter. In the example of the Titanic dataset [80] it improves the Parallel Sets visualization even more than the *Slope* measure, significantly differing from it.

Number of Crossings Ⓒ measures the number of ribbon crossings. This measure is analogous to the *Number of Line Crossings* measure of the Pargnostics [78] measure set for parallel coordinates. A high number of crossing produces similar patterns like dissimilarity orderings for parallel coordinates, which can be used to detect patterns [46]. In Parallel Sets visualizations a high degree of ribbon crossings can lead to visual clutter, making ribbons hard to follow. This effect has been observed for parallel coordinates [100]. Thus, a very high and very low value for *Number of Crossings* can indicate an interesting Parallel Sets for exploratory analysis. The value $\mathscr{C}$ in equation (2.12) describes the absolute number of crossings.

$$f_{\text{crossing}}(r_1, r_2) := \begin{cases} 1 : overlap(r_1, r_2) > 0 \\ 0 : overlap(r_1, r_2) \leq 0 \end{cases} \tag{2.11}$$

$$\mathscr{C} := \sum_{i=1}^{|\mathcal{A}|-1} \sum_{(r_1, r_2) \in P_i} f_{\text{crossing}}(r_1, r_2) \tag{2.12}$$

We formally describe this measure in equation (2.13), which provides a relative number of crossing proportional to the number of ribbons contained in a Parallel Sets visualization.

Definition: Number of Crossings Measure

$$\textsc{Crossings} := \frac{\mathscr{C}}{|\mathbf{R}|} \tag{2.13}$$

The examples depicted in figure 2.4 show that a minimization of the number of crossings progressively reduces the amount of clutter. A Parallel Sets visualization with a maximum number of is likely to exhibit zigzag patterns.

Crossing Angle Ⓐ quantifies the average crossing angle of crossing ribbons of a Parallel Sets visualization. This measure is motivated by the *Angels of Crossing* measure of the Pargnostics [78] measure set for parallel coordinates. A very high or very low angle of crossing benefits the readability of the Parallel Sets visualization. Ribbons crossing at a flat angle are hard to follow compered to ribbons crossing at close to right angles. This effect has already been observed for lines [155, 326]. We formally describe this measure in equation (2.14).

Definition: Crossing Angle Measure

$$\textsc{CrossingAngle} := \frac{1}{\mathscr{C}} \sum_{i=1}^{|\mathcal{A}|-1} \sum_{(r_1, r_2) \in P_i} \delta(r_1, r_2) \tag{2.14}$$

In this equation, the crossing angle of two ribbons is denoted by angle δ. The factor $\frac{1}{\mathscr{C}}$ based on equation (2.13) provides a value relative to the total number of crossings. The concept of a crossing angle and how it is described by δ is depicted in figure 2.3. In the examples shown in in figure 2.4 this measure offers Parallel Sets visualizations with a low amount of clutter for a high and low value, while the median exhibits a zigzag pattern and clutter. In general, a high crossing angle is preferred since it supports readability.

2.4.2 Attribute Ordering-Dependent Measures

This section describes three attribute ordering-dependent measures, which means that they are only sensitive to the reordering of attributes and are not affected by changes in the order of categories of any attribute axes. These measures can be used to limit the search space by fixing the order of attribute axes.

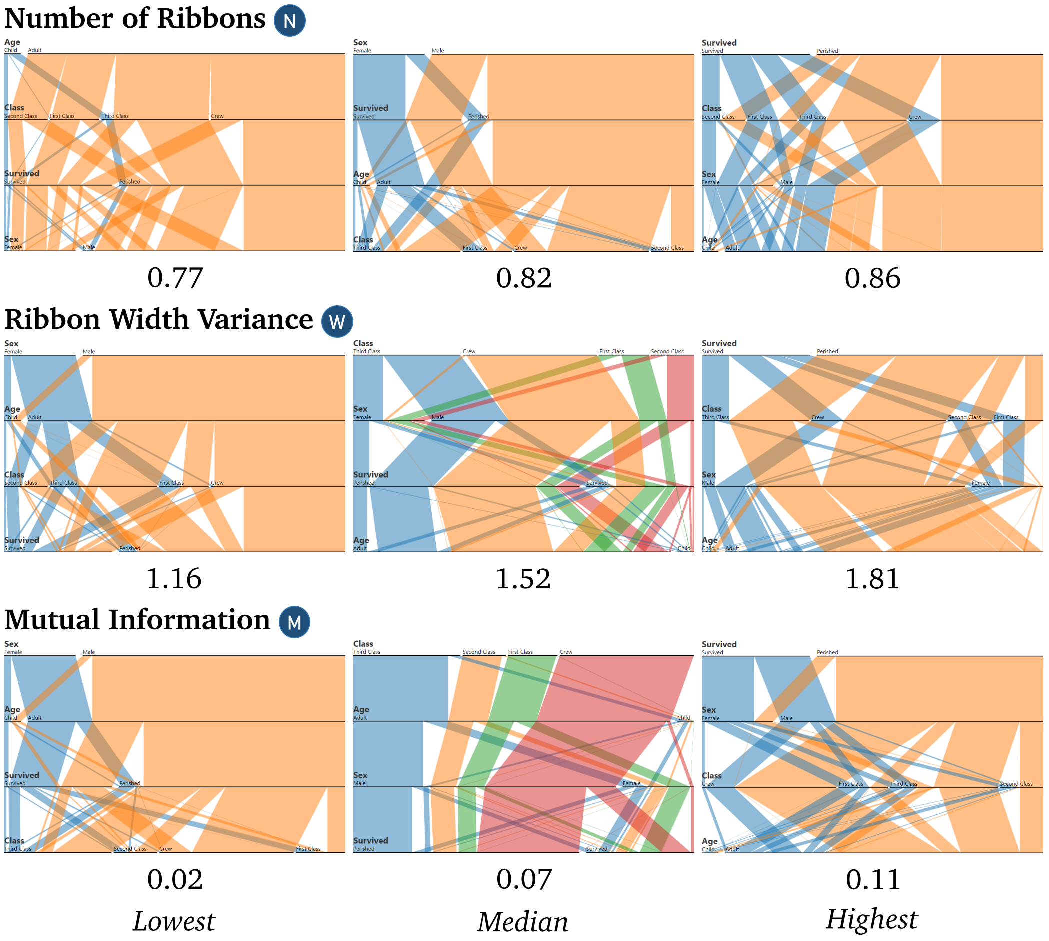


Figure 2.5: We show three Parallel Sets visualizations for each of the three attribute ordering-dependent measures: *Number of Ribbons* N, *Ribbon Width Variance* W, and *Mutual Information* M. The results shows a reduction of clutter for a reordering of the attributes, which can serve as a basis for further improvements.

Number of Ribbons N measures the number of ribbons. The number of ribbons determine the number of ribbon splits according to the categories of attribute axes. In general, a low number of splits is preferable since a high number of ribbons increase the likelihood of sloped and overlapping ribbons. Furthermore, splits reduce the ribbon width, creating thin ribbons, which are hard to follow. Thus, splits into subcategories should be avoided and only occur where the analysis question requires it. The only exception is when the analyst wants to determine the number of subsets created by a specific category or attribute.

Definition: Number of Ribbons Measure

$$\textsc{Ribbons} := \frac{|\mathbf{R}|}{|\mathbf{R}^*|} \tag{2.15}$$

The equation measures the ratio of all exiting ribbons to all possible ribbons, allowing for comparability between different attribute orderings. The effects of minimizing the number of ribbons is shown in figure 2.5. A low amount of ribbons reduces clutter and improves readability.

Ribbon Width Variance Ⓦ measures the variance of ribbon widths. A low ribbon width variance is preferable, splits that create very small categories should be delayed. Very broad ribbons hide smaller ones. We calculate the standard deviation σ of the ribbons widths, allowing for comparability of different Parallel Sets. To avoid absolute widths, we define $maxWidth = max(\{width(r) \mid r \in \mathbf{R}\})$, which we use to normalize the ribbons widths. We formally describe this measure in equation (2.16).

Definition: Width Variance Measure

$$\textsc{WidthVariance} := \sigma(\{width(r)/maxWidth \mid r \in \mathbf{R}\}) \tag{2.16}$$

The effect is shown in figure 2.5. We found that a ribbon with variance can reduce clutter of Parallel Sets, showing that a uniform ribbon width improves readability.

Mutual Information Ⓜ measures the average mutual information of neighboring attribute axes. It was first proposed by Shannon [278]. Mutual information measures the dependence between two variables, in the case of Parallel Sets, two neighboring attributes. It measures the amount of information gained about one variable by observing another variable. Mutual information is formally defined as:

Definition: Mutual Information Measure

$$\begin{gathered}\textsc{MutualInfo} := \frac{1}{|\mathcal{A}|-1} \sum_{i=1}^{|\mathcal{A}|-1} I(\mathcal{C}_i, \mathcal{C}_{i+1}) \\ \text{where } \; I(X,Y) := \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}\end{gathered} \tag{2.17}$$

In this equation, $p(x, y)$ is the probability of the values $x \in X$ and $y \in Y$ occurring together. Since mutual information only measures the distribution of categories between two attributes without considering the category ordering, it does not change by reordering categories. Thus, it can only be used to determine an ordering of the attributes axes. It is used by Dasgupta and Kosara [78] in the reordering of parallel coordinate axes and by Alsakran et al. [9] where it is combined with binning or dimensionality reduction. In figure 2.5, this measure shows an improvement of readability for high and low values. In general, it should be maximized to improve Parallel Sets visualizations.

2.4.3 Combining Quality Measures

Our measures can be combined since they measure different aspects of Parallel Sets. Two or more measures can be minimized or maximized simultaneously, or they can be optimized successively. This especially applies to the combination of a attribute ordering dependent-measure and a category ordering-dependent measure. The order of categories of an axis in Parallel Sets is the most flexible parameter. Therefore, we are free to maximize or minimize the category ordering for one or multiple of the category ordering-dependent measures, each reducing different artifacts. They can also be combined in frameworks for the weighting of features [239]. The ordering of attributes is not as flexible as the ordering of categories because of the following reasons: (1) The number of attributes is usually lower than the number of categories. (2) The categories of the first attribute axis determine the ribbon colors, and thus, the primary target of analysis. (3) All remaining axes split the ribbons into finer and finer subcategories according to their ordering. We suggest minimizing the number of ribbons to reduce the possibility of crossings and overlap. However, this may lead to thin ribbons in the visualization. Alternatively, we propose to reduce the ribbon width variance to avoid excessively thin or broad ribbons, which does not enforce the minimum amount of ribbons. The mutual information measure tries to place related attributes close to each other, independent of ribbon sizes. We propose the use of those types of measures as a filtering step.

2.5 Evaluation

To show the effectiveness of our approach, we perform a quantitative evaluation based on visualizations used in previous publications. We perform single-measure and multi-measure optimizations of the Parallel Sets visualizations and conduct a correlation analysis to validate the distinctiveness of our measures.

Source	Description	Type	Num. Attr.
Hassan et al. [140]	Cyber-security	Application	4
Koh et al. [186]	Property sales	Design study	3
Kosara et al. [190]	Titanic dataset [80]	Technique	4
Rodgers et al. [256]	HCI study; 2 datasets	Application	3
Schätzle et al. [271]	Linguistics	Application	2

Table 2.1: We found five papers from different domains using Parallel Sets yielding six datasets for our evaluation of ParSetgnostics.

2.5.1 Reconstruction of Datasets from Parallel Set Visualizations

To evaluate our approach, we performed a literature search with the terms “Parallel Sets” and “ParSets.” Additionally, we performed a forward search on the foundational publication on Parallel Sets by Bendix et al. [35], and Kosara et al. [190]. Both searches were performed using the digital libraries of ACM, IEEE, and Eurographics. This yields a set of five publications using Parallel Sets listed in table 2.1. The Titanic dataset is available online [80]. We reconstructed the other remaining five datasets manually. To this end, we measured the width of the ribbons in the lowest level to get the size of the subset and traced the ribbon from top to bottom to determine the categories determining the subset. After a visual comparison to the original Parallel Sets visualization, we estimate the reconstructions to be mostly accurate. The most challenging to reconstruct was the dataset by Koh et al. [186] since it contains many small ribbons only one to five pixels wide and a high slope. We choose these datasets because they have published Parallel Sets, implying that they are suitable targets for comparison. To determine the optimized values of our measures, we calculated all measures for all possible configurations.

2.5.2 Single-Measure Optimization

To show the usefulness of each measure, we perform a single-measure optimization on two visualizations using visualizations provided by Hassan et al. [140] and Rogers et al. [256]. Firstly, one by Hassan et al. [140], which is a regular Parallel Sets visualization. Secondly, a visualization by Rogers et al. [256] with curved ribbons.

Hassan et al.: In figure 2.6, we perform an optimization using all measures individually on the Parallel Sets published by Hassan et al. [140]. This visualization aims to analyze the security and cost of data storage, determining the location where data storage should be bought with a high security level. We can see that all category ordering-dependent measures produce visualizations with lower clutter. The *Overlap*

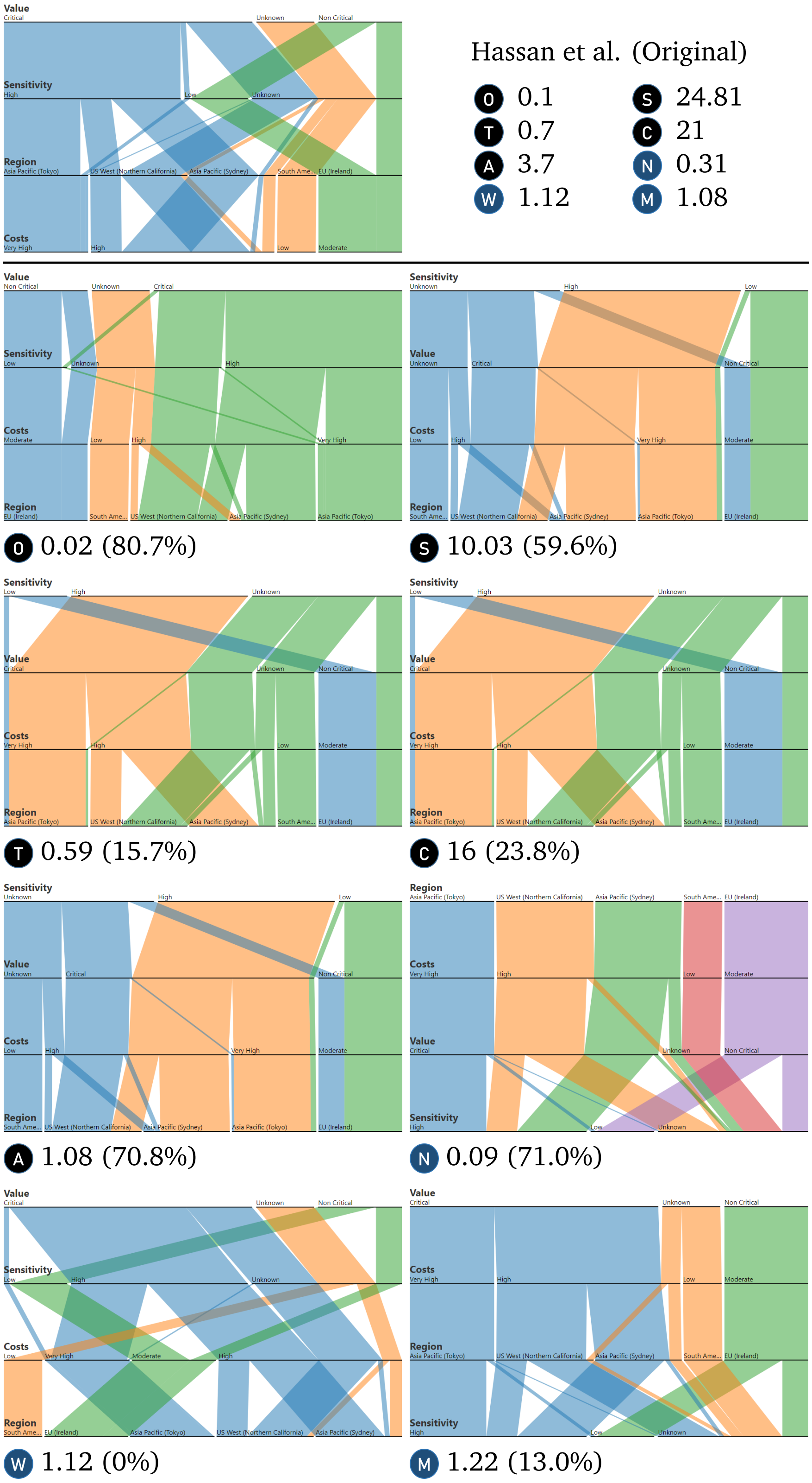


Figure 2.6: We show the optimization results for the visualizations provided by Hassan et al. [140]. The original Parallel Sets visualization is shown at the top along with measure values. All single-measure optimizations are shown below with the percentage improvement compared to the original. All category ordering-dependent measures have lower clutter. All measures are lower in comparison to the original Parallel Sets visualization. The *Ribbon Width Variance* W measure yields the worst result.

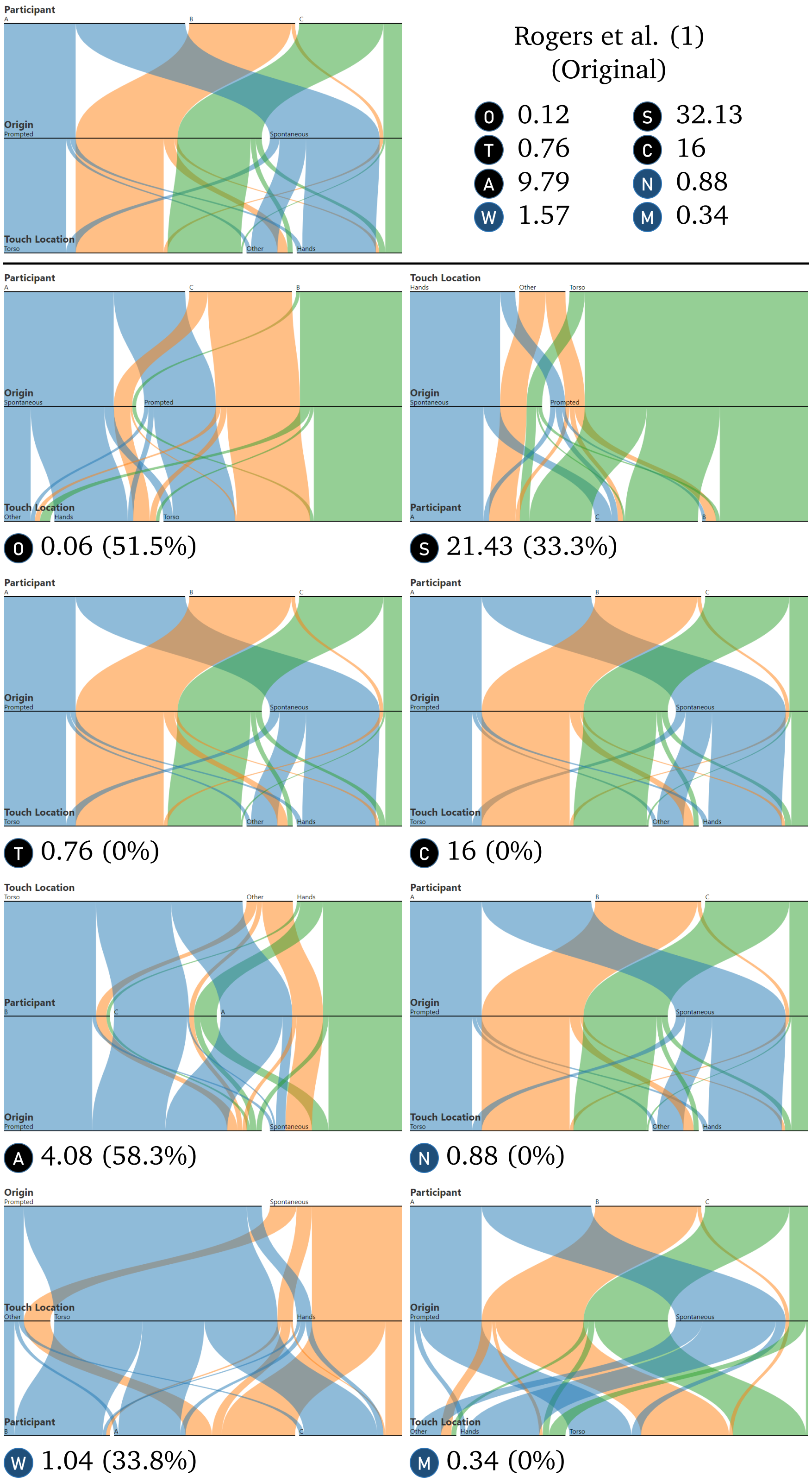


Figure 2.7: We show the optimization results for Rogers et al. (1) [256] with curved ribbons. The original Parallel Sets visualizations are shown at the top along with measure values. All single-measure optimizations are shown below with the percentage improvement compared to the original. All category ordering-dependent measures have lower clutter and are lower in comparison to the original Parallel Sets visualization. The *Number of Ribbons* N measure performs worst.

Ⓞ measure reduces the overlap of ribbons by 80.7% compared to the original. If we assume the overlap as an objective measure of clutter [337] the *Slope* Ⓢ and *Crossing Angle* Ⓐ measure reduce overlap by 70.8%. These measures improve by 59.6% and 70.8%. The attribute ordering-dependent measures reduce clutter as well, with the exception of the *Ribbon Width Variance* Ⓦ measure. All measures are lower in comparison to the original Parallel Sets, showing that the original visualization was not optimized according to any property of the Parallel Sets. We note that *Slope* Ⓢ and *Crossing Angle* Ⓐ create the same visualizations, as well as *Orthogonality* Ⓣ and *Number of Crossings* Ⓝ.

Rogers et al. (1): We perform an optimization using all measures individually on the Parallel Sets published by Rogers et al. [256] showing the more complex dataset of this publication's datasets with curved ribbons. We determine the angles of the ribbons based on the underlying straight ribbons. The task for this visualization is to present the result of a Human-Computer Interaction (HCI) study. The optimization results are shown in figure 2.7. *Orthogonality* Ⓣ, *Number of Crossings* Ⓒ, and *Number of Ribbons* Ⓝ are already optimized in the original visualization. Thus, there is no improvement by these measures. We can see that the *Slope* Ⓢ measure produces large contiguous ribbons and focuses the smaller ribbons in the center. Considering the overlap as a measure of the degree of clutter [337], the *Slope* Ⓢ, measure reduces clutter by 50.9% and the *Crossing Angle* Ⓐ measure by 45.1%. The attribute ordering-dependent measures yield the same ordering for *Number of Ribbons* Ⓝ and *Mutual Information* Ⓜ than the original visualization and thus optimal in those aspects.

2.5.3 Multi-Measure Optimization

We evaluate the multi-measure optimization capabilities by selecting the attribute ordering that two out of three attribute ordering-dependent measures agree on. Based on this ordering, we choose a Parallel Sets visualization according to the measure that improved the most compared to the original visualization.

Koh et al.: We perform this optimization using the visualization in the publication by Koh et al. [186] dealing with property sales analysis. Each step is shown in figure 2.8. First, we analyze the attribute ordering. The *Number of Ribbons* Ⓝ and *Ribbon Width Variance* Ⓦ yield the same attribute ordering, while *Ribbon Width Variance* Ⓦ is reduced by 12.4%. For the category ordering we fix the attribute ordering accordingly. We apply all category ordering-dependent measures to the visualization. We can observe that the *Orthogonality* Ⓣ and *Number of Crossings* Ⓒ yield the identical visualization. By assessing the *Crossing Angle* Ⓐ, reducing its value by 15.0% and choosing the overlap as an objective measure for clutter [337] we can see that clutter is reduced by 0.8%. Observing the result, we can also see a cleared-up top level compared to the original.

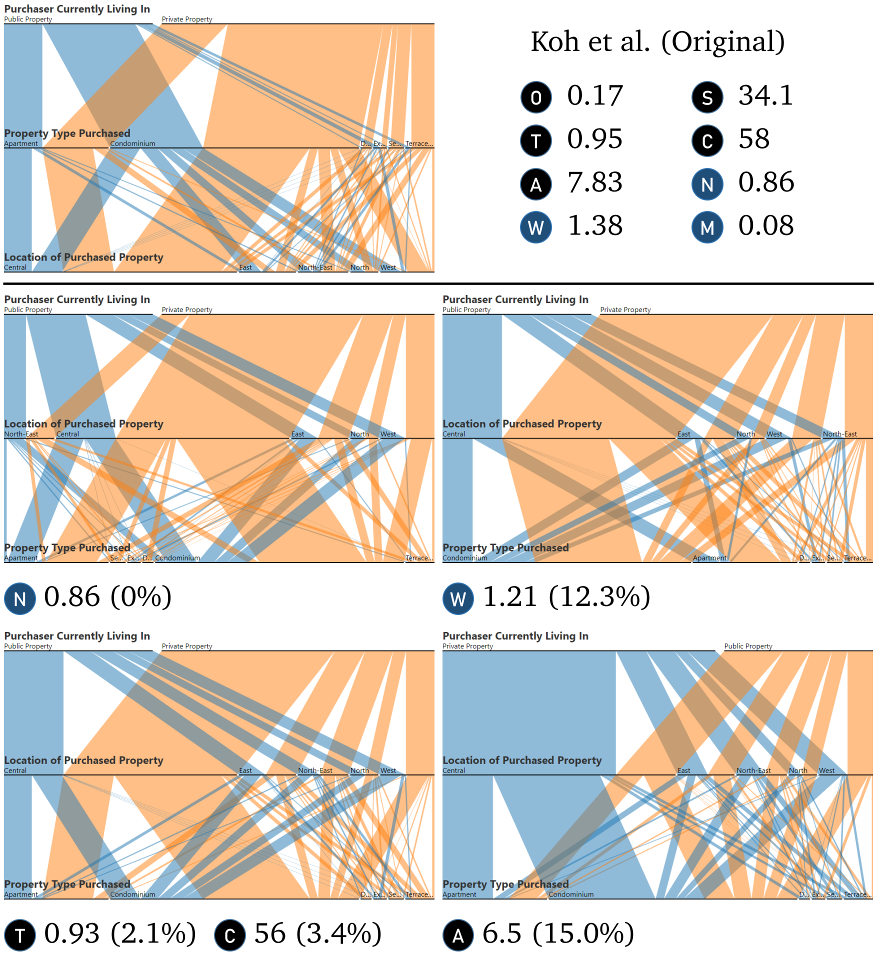


Figure 2.8: We optimize the dataset supplied by Koh et al. [186]. In this case, the optimization is based on the attribute ordering derived from the *Number of Ribbons* N and *Ribbon Width Variance* W measures.

Rogers et al. (2): The visualization presented by Rogers et al. [256] describes the result of an HCI study with curved ribbons. We determine the ribbon angles based on the underlying straight ribbon. The steps are shown in figure 2.9. The *Number of Ribbons* N and *Ribbon Width Variance* W provide the same attribute ordering. Thus, we only consider layout with this ordering. To determine the order of categories, which influences the appearance of the ribbons. We find that optimum of *Overlap* O and *Slope* S have the attribute ordering as suggests by the attribute ordering-dependent measures. The visualization suggested by the *Slope* S measure reduces the clutter by 53.2% considering overlap as an objective measure [337]. This visualization focuses all splits and crossings on one the left half of the visualization.

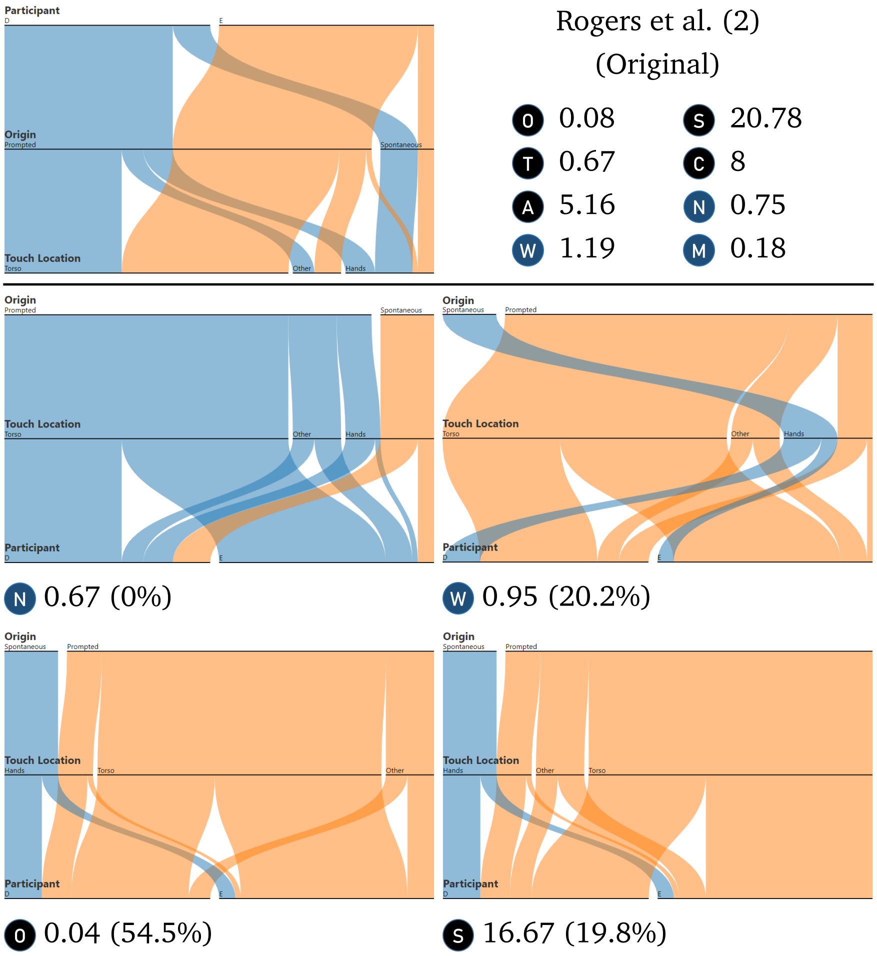
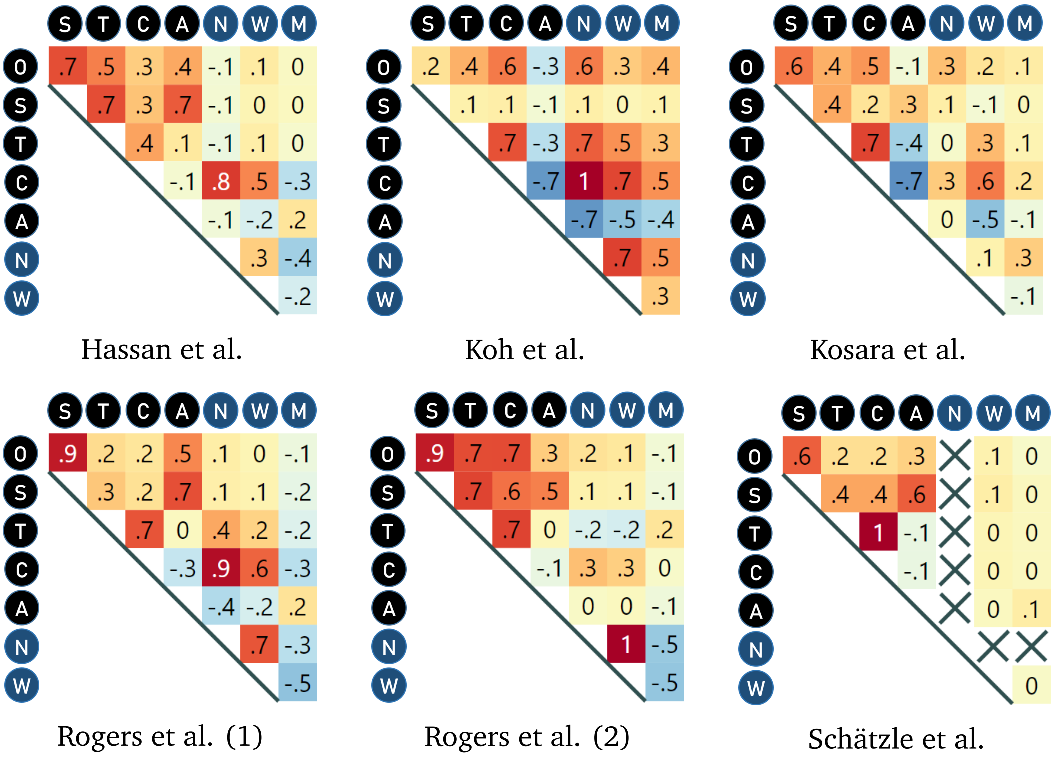


Figure 2.9: We apply our measures to improve the second visualization of Rogers et al. [256]. In this case, the optimization is also based on the attribute ordering derived from the *Number of Ribbons* N and *Ribbon Width Variance* W measures.

2.5.4 Correlation Analysis

In order to evaluate that our measures quantify different properties of Parallel Sets we performed a Pearson correlation analysis [183] of the measures. We calculated the value of all measures for all attribute and category layouts for all available datasets. The results of the analysis are summarized in figure 2.10. The measure *Crossing Angle* A shows a weak negative correlation for the Koh et al. [186] dataset and *Mutual Information* M shows a weak negative correlation for the Rogers et al. (1) [256] dataset. The *Number of Ribbons* N measure could not be analyzed for the data by Schätzle et al. [271] because it only has two attributes and thus a fixed number of ribbons for all configurations. The correlation analysis shows that the correlations between measures is dependent on the dataset. This is shown by the differing Pearson correlations. Figure 2.10 provides the correlations between the measures for all datasets. We found that no measure correlates with any other measure for all analyzed datasets. This shows that all measures are independent and measure distinct properties, and are mutually independent.



	S	T	C	A	N	W	M
O	.7	.5	.3	.4	-.1	.1	0
S		.7	.3	.7	-.1	0	0
T			.4	.1	-.1	.1	0
C				-.1	.8	.5	-.3
A					-.1	-.2	.2
N						.3	-.4
W							-.2

Hassan et al.

	S	T	C	A	N	W	M
O	.2	.4	.6	-.3	.6	.3	.4
S		.1	.1	-.1	.1	0	.1
T			.7	-.3	.7	.5	.3
C				-.7	1	.7	.5
A					-.7	-.5	-.4
N						.7	.5
W							.3

Koh et al.

	S	T	C	A	N	W	M
O	.6	.4	.5	-.1	.3	.2	.1
S		.4	.2	.3	.1	-.1	0
T			.7	-.4	0	.3	.1
C				-.7	.3	.6	.2
A					0	-.5	-.1
N						.1	.3
W							-.1

Kosara et al.

	S	T	C	A	N	W	M
O	.9	.2	.2	.5	.1	0	-.1
S		.3	.2	.7	.1	.1	-.2
T			.7	0	.4	.2	-.2
C				-.3	.9	.6	-.3
A					-.4	-.2	.2
N						.7	-.3
W							-.5

Rogers et al. (1)

	S	T	C	A	N	W	M
O	.9	.7	.7	.3	.2	.1	-.1
S		.7	.6	.5	.1	.1	-.1
T			.7	0	-.2	-.2	.2
C				-.1	.3	.3	0
A					0	0	-.1
N						1	-.5
W							-.5

Rogers et al. (2)

	S	T	C	A	N	W	M
O	.6	.2	.2	.3	×	.1	0
S		.4	.4	.6	×	.1	0
T			1	-.1	×	0	0
C				-.1	×	0	0
A					×	0	.1
N						×	×
W							0

Schätzle et al.

Figure 2.10: The results of the correlation analysis of the measures for all reconstructed datasets. We found that no measure correlates with any other measure for all analyzed datasets. This shows that all measures are independent and measure distinct properties.

2.6 Discussion and Future Work

The calculation of all quality measures is dependent on the number of ribbons of a Parallel Sets visualization. All measures are described in terms of vector graphics. Our measures can be applied before the ribbons are curved since the straight ribbons approximate the properties of the curved ribbons. All attribute ordering-dependent measures are directly applicable since they are not dependent on the ribbon shape. All category ordering-dependent measures, except the *Number of Crossings* measure, will provide an approximate result, which can improve the visualization. All quality measures, except the angle-related measures (i.e., *Slope*, *Orthogonality*, and *Crossing Angle*) can be directly applied to Common Angle Plots since they enforce the angle a ribbon has in-between two attribute axes. Our measures can be used to measure the quality increase or decrease in cases where the underlying data changes. This is also true for streaming scenarios, where new categories might be encountered.

However, determining an optimal ordering of attributes and categories would require a more efficient optimization strategy, other than calculating the measures for all possible configurations. The measures are calculated reasonably fast, such that in an interactive design process, they can be used to compare and rank different manually created Parallel Sets visualizations instantly. Our correlation analysis shows that all measures quantify distinct properties and thus are mutually independent. We derive the set of measures from our discussion on parameters of Parallel Sets related to the choice of the category and attribute ordering, dataset-dependent properties, and ribbon parameters. Our measures address all parameters, and thus, we argue for completeness in terms of geometric properties. We plan a user study as an additional validation of completeness.

Guidelines: We found the following design guidelines for the layout of attributes and categories of Parallel Sets visualizations: (1) Choose the first attribute according to the analysis question or well-known categories. In exploratory tasks, choose a attribute with a category count no larger than nine. We suggest following Miller's Law, which states to limit the number of shown items to seven plus or minus two [218, 219]. We also suggest choosing a attribute with a high entropy leading to equal-sized categories. (2) Filter the set of all configurations by attribute ordering-dependent measures. These measures can be used in a voting system as we do in section 2.5.3. (3) Minimize/Maximize a category ordering-dependent measure. In our experiments, we found some suggestions: Parallel Sets with a low number of ribbon splits, i.e., a low number of ribbons in the lower levels of Parallel Sets show better results when optimized with the *Overlap* and *Slope* measures. Parallel Sets with a high number of ribbons are optimized with the *Orthogonality*, *Number of Crossings* and *Crossing Angle*. Curved ribbons are easier to read. This is based on the fact that curved lines have a larger crossing angle, which makes lines easier to follow [155, 326].

Limitations and Future Work: Our measures quantify the visual appearance of Parallel Sets. They do not provide a reordering strategy. The next step is to assess the properties of our measures and derive a reordering algorithm. Another possible direction is an extension towards local measures since our measures only describe Parallel Sets globally. We plan to study the connection between specific measures with general tasks and data set characteristics through a user-study. A user study would also verify whether the set of measures is exhaustive. This work does not describe an efficient strategy to determine the minimum and maximum value of a measure. Additionally, we plan to study the effects of the measures in the interactive design of Parallel Sets suggesting and validating user choices. One drawback of our approach is that the measures need to be recalculated if the aspect ratio of the plot changes. In the case of simple zooming with a fixed aspect ratio, the values can be reused. Our quality measures could potentially be transferred to the quantification of properties of Sankey diagrams since many desirable proprieties of Parallel Sets are also desirable for Sankey diagrams, e.g., a low overlap of ribbons.

2.7 Conclusion

Determining a useful attribute and category ordering for Parallel Sets is challenging. We propose a set of eight distinct quality measures for Parallel Sets, called ParSetgnostics. They provide a new model for quantifying properties of Parallel Sets visualizations, which can be used as a quality criterion as described by Behrisch et al. [33]. Our measures allow us to improve the readability of Parallel Sets visualizations by optimizing a specific measure or a combination thereof or even determining the presence of undesirable patterns. We argue for our measures' effectiveness by applying them to Parallel Sets in previous publications, showing their applicability in a single- and multi-measure optimization approach. We perform a correlation analysis on all datasets and quality measures combinations and validate that no measure correlates with any other measure for all datasets, showing each measure's distinctiveness. We published the results online where users can explore our results and test the quality measures' properties interactively. Our work provides a more meaningful way to analyze categorical data with Parallel Sets, especially in exploratory scenarios.

3 Measures for 2-Dimensional Categorical Data Projections

Contents

Categorical data does not have an intrinsic definition of distance or order, and thus, established visualization techniques for categorical data only allow for a set-based or frequency-based analysis, e.g., through Euler diagrams or Parallel Sets, and do not support a similarity-based analysis. We present a dimensionality reduction-based visualization for categorical data based on defining the distance of two data items as the number of varying attributes. Our technique enables users to pre-attentively detect groups of similar data items and observe the properties of the projection, such as attributes strongly influencing the embedding. Our prototype visually encodes data properties in an enhanced scatterplot-like visualization, visualizing attributes in the background to show the distribution of categories. We propose two graph-based measures to quantify the plot's visual quality for ranking attributes according to their contribution to cluster cohesion. To demonstrate the capabilities of our method, we compare it to Euler diagrams and Parallel Sets regarding visual scalability and evaluate it quantitatively on seven real-world datasets using a range of common quality measures. We conducted an expert study with five data scientists analyzing the Titanic and Mushroom datasets with up to 23 attributes and 8124 category combinations. Our results indicate that our Categorical Data Map is an effective analysis method for large datasets with a high number of category combinations.

This chapter is *taken from* the following manuscript:

- [87] **Frederik L. Dennig**, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel A. Keim, and Maximilian T. Fischer. "The Categorical Data Map: A Multidimensional Scaling-Based Approach". In: *Proceedings of the 2024 IEEE Visualization in Data Science Symposium (to appear)*. IEEE, 2024.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

3.1 The Need for Similarity-Based Analysis of Categorical Data

Categorical data can be encountered in numerous domains, such as representing inventory data describing product properties like color in sales or bioinformatics, encoding the genes formed by nucleotide sequences [5]. In contrast to numeric and ordinal data, categorical data does *not* have an intrinsic order or distance associated with each value pair. The visual analysis of categorical data is challenging since categorical data describes an attribute by name only, with the only supported operators being *equality*, *set membership*, and *mode*.

Currently, there are two widespread methods of visualizing categorical data: (1) *Frequency-based visualizations* [149, 331, 292] map the categorical values to their frequencies, for example, through bar charts, pie charts, or enhanced variants, such as stacked bar charts. In contrast, (2) *set visualizations* solely focus on the set nature of categorical data items, specifically their intersections [10]. Examples include such as Euler diagrams [237] and UpSet plots [200]. Set visualizations like Euler diagrams do not scale well for sets with many intersections because visual clutter is detrimental to their readability. Other, less common solutions treat dimensions independently and map data to a continuous design model [300, 164, 261], leveraging visualization types that initially have been designed for numerical data, such as scatterplots or parallel coordinate plots. However, these approaches deviate from the *discrete* nature of categorical data and suffer from visual clutter and overplotting, limiting their readability [190]. Approaches, such as Parallel Sets [35] and Sankey diagrams [181], follow the frequency and set-based paradigms. These approaches trade effectiveness in visualizing the presence of small subsets for the presentation of frequency information. These approaches require additional design considerations since they tend to emphasize preselected attributes over others [85].

None of the previously described techniques support the similarity-based analysis of categorical data, i.e., deriving the *similarity* of categorical data items as distances such that similar data items are placed close to each other while differing data items are positioned far apart. Analyzing categorical data based on a group or subset similarity is useful, e.g., visually clustering data items only differing in a few attributes can help us better understand important characteristics of the group. Generally, this would allow us to apply methods from cluster analysis to categorical data.

We follow the suggestion by Broeksema et al. [55] to investigate multidimensional scaling to generate *visual mappings* that enable the interpretation of distances and simultaneously convey the properties of data items, i.e., effectively visualizing an item's attributes by using color and position to visually encode attributes. Through this, we address the combinatorial problem of categorical data, i.e., that with the increasing number of attributes and categories, the number of required colors to represent a category with distinguishable colors becomes increasingly difficult. Tackling these challenges, we contribute the following:

Contributions

- A technique applying multidimensional scaling to categorical data while *visually encoding* the category distribution into the background. Through *layout enrichment*, we enable the exploration of the category distribution, enhancing orientation and navigation. Additionally, we contribute *four glyph designs* to represent categorical subsets.
- *Quality measures* based on subset distribution to guide the analysis, recommending layout enriched views on attributes contributing strongly to clusters and subset separation.
- A *quantitative comparison* to multiple correspondence analysis-based projections and a *qualitative expert study* validating the effectiveness of our approach.
- An *online demonstrator* (https://dennig.dbvis.de/categorical-data-map) making the acquired results accessible. To further aid reproducibility, we *openly publish* all our *datasets* and *source code* via OSF (osf.io/jzd46) and DaRUS [86].

3.2 Related Work

Our approach is related to visualization and dimensionality reduction methods for categorical data. Furthermore, we propose a layout enrichment for multidimensional projections and contribute visual quality measures for categorical data projections.

3.2.1 Visualization Techniques for Categorical Data

Set visualization is one of the core techniques for categorical data. To visualize the members of sets and their intersections, Venn and Euler diagrams are the two most prevalent representations [28]. Multiple adaptations of both techniques mitigate challenges, e.g., to preserve semantics [174], draw area-proportional diagrams [243], or incorporate glyphs to show additional information [216]. Other set visualization techniques use lines to indicate set intersections [254] and matrices to show the cardinality of intersection sets [200], or include the semantic context to visualize sets [215]. Alsallakh et al. presented a comprehensive survey on set visualizations [10]. There are also *frequency-based visualization* methods that focus on attribute frequencies, such as Mosaic plots [149] and Parallel Bargrams [331] by mapping data item occurrences to one or multiple attributes, e.g., a rectangle's area. Other methods map data to a continuous design model, such that they are compatible with visualization for numeric data, e.g., Rosario et al. [261] describe the mapping of categorical data to numeric values for the visualization in Parallel Coordinates [156]. Hybrid methods consider both aspects, e.g., Parallel Sets [35, 190] and Sankey diagrams [181]. However, Parallel Sets and Sankey diagrams can suffer from the Müller-Lyer and Sine illusions [81, 316] where lines seem to vary in distance or length, affecting the accurate interpretation of frequencies and proportions.

While plenty of approaches visualize categorical data, to the best of our knowledge, none allows identifying groups of similar data items. Thus, we propose a visualization that focuses on similarity.

3.2.2 Dimensionality Reduction for Categorical Data

Our approach makes use of Dimensionality Reduction (DR). However, there exist DR methods for categorical data that do not focus on similarity but rather describe the central oppositions in the data [126]. When needing to reduce the dimensionality of categorical data, Correspondence Analysis (CA), similar to Principal Component Analysis (PCA) [166] for numerical data, extracts the standard coordinates, yielding a Biplot [118] of the reduced space. In case of more than two categorical variables, Multiple Correspondence Analysis (MCA) can be used to reduce the number of dimensions showing the central oppositions [126]. Factor Analysis of Mixed Data (FAMD) is a principal component technique for continuous and categorical variables [238]. The continuous variables are scaled to unit variance, and the categorical variables are transformed into a disjunctive data table and then scaled using the specific scaling of MCA to balance the influence of both continuous and categorical variables in the analysis. Multiple Factor Analysis (MFA) combines these methods for mixed data: It uses PCA when variables are quantitative, MCA when variables are qualitative, and FAMD when the active variables belong to both of the two types. The Data Context Map [71] visualizes *mixed-data* using an Multidimensional Scaling (MDS)-based plot and displays categorical attributes on top of the projection while also coloring points and regions according to the predominant category. The approach by Thane et al. [301] uses force-directed graph layouts to visualize categorical datasets representing categories as nodes while edges represent their co-occurrence.

MCA can embed categorical data but, like PCA, is a linear dimensionality reduction technique and thus not able to detect non-linear relationships [126, 55]. We propose using MDS to visualize the similarity of categorical data points in a scatterplot-like layout.

3.2.3 Layout Enrichment for 2-Dimensional Data Projections

The idea to enrich scatterplot layouts by encoding additional information in the background of a projection is not new [235]. The main usage occurs for the visualization of distortions in the topology of the embedding resulting from DR [19]. The following approaches make use of Voronoi diagrams [22] to encode additional information in the background of a projection. Lespinats and Aupetit proposed CheckViz [199], visualizing the presence of *tears* (i.e., missing neighborhood) and *shuffled data* (i.e., wrong neighborhood). Broeksema et al. explored the visualization of categorical data, combining MCA with an enhanced treeview to integrate data record information visualizing user-selected categories. However, they did not address the high redundancy of categorical datasets [55]. Sohns et al. followed a similar approach; however, they used non-linear DR methods to project *mixed data* while using categorical attributes to highlight areas of the embedding space. However, this approach excludes all categorical attributes from the DR process altogether [285]. DICON enables the analysis of multidimensional clusters with an interactive icon-based visualization that encodes additional statistical information visually using space-filling methods, including Voronoi diagrams [61]. Aside from using Voronoi diagrams, other methods for layout enrichment exist [45]. Morariu et al. encode the projection's quality into the plot's background using contours showing the embedding of projections called the metamap [225].

Layout enrichment methods largely focus on visualizing distortions of the projection. The approach by Broeksema et al. [55] does not address the analysis of a single attribute, so we propose a new enrichment that encodes the category of an attribute using color.

3.2.4 Measures for Quality and Patterns in Visualizations

Quality measures for visualizations describe a set of measurements designed to optimize visualizations in terms of *readability* and *clutter reduction* [33]. Other measures quantify the presence of *patterns* in a visualization. Instead of measuring quality, pattern measures can be used to compare and rank different visualizations based on specific properties. Examples are: Magnostics for matrix visualizations [31], Scagnostics for general patterns and trends on scatterplots of numeric data

[330], Pargnostics for parallel coordinate plots [78], Visualgnostics for projections of high-dimensional data [197], Pixgnostics for pixel-based visualizations [274], and ParSetgnostics for Parallel Sets [85]. SepMe is a machine-learning-based approach to quantify the presence of clusters in scatterplots [21], while ClustMe quantifies the visual separation of classes in scatterplots [1]. Aupetit and Catz [20] addressed the analysis of high-dimensional labeled data using graphs, including Voronoi diagrams. However, this approach does not address categorical data analysis, i.e., where no numerical attributes are present.

We contribute two novel measures for quantifying visual quality for 2-dimensional projections of categorical data. In this way, we improve the exploration of categorical data by recommending layout-enriched views according to their visual structure.

3.3 Constructing the Categorical Data Map

Typically, categorical datasets exhibit inherent *sparsity*, i.e., only a fraction of all possible category combinations is present in a dataset, e.g., for the Mushroom dataset, only 8124 out of 243.799.621.632.000 possible combinations. Thus, we assume that there are relationships among the existing categories restricting their combinations. Additionally, categorical datasets can be *highly redundant*, e.g., the Titanic dataset contains 2201 data items but only 24 unique entries, i.e., all data items can be assigned to one of 24 subsets. Thus, we focus on categorical subsets as subsets of unique attribute values. These subsets are our main representations, enabling us to assign a frequency. We leverage these properties in the design of the Categorical Data Map as an analytical approach for the similarity-based analysis of categorical subsets with the following *constraints*:

Requirements

(C1) Distances of categorical subsets in a scatterplot should indicate similarity, i.e., subsets with a smaller distance should differ in fewer attributes than subsets with a larger distance.

(C2) Allow analysts to find groups of subsets by clustering similar categorical subsets and separating outliers.

(C3) Highlight attributes contributing to the clustering of subsets enabling navigation and orientation in the projection.

(C4) Provide a recommendation for attributes to explore first, linked to the distribution of categories in the plot.

An example of our approach is shown in figure 3.1. (C1) and (C2) are described further in section 3.3.1. We address (C3) by evaluating different glyph designs and layout enrichments for subsets of categorical data (see section 3.3.2). We address

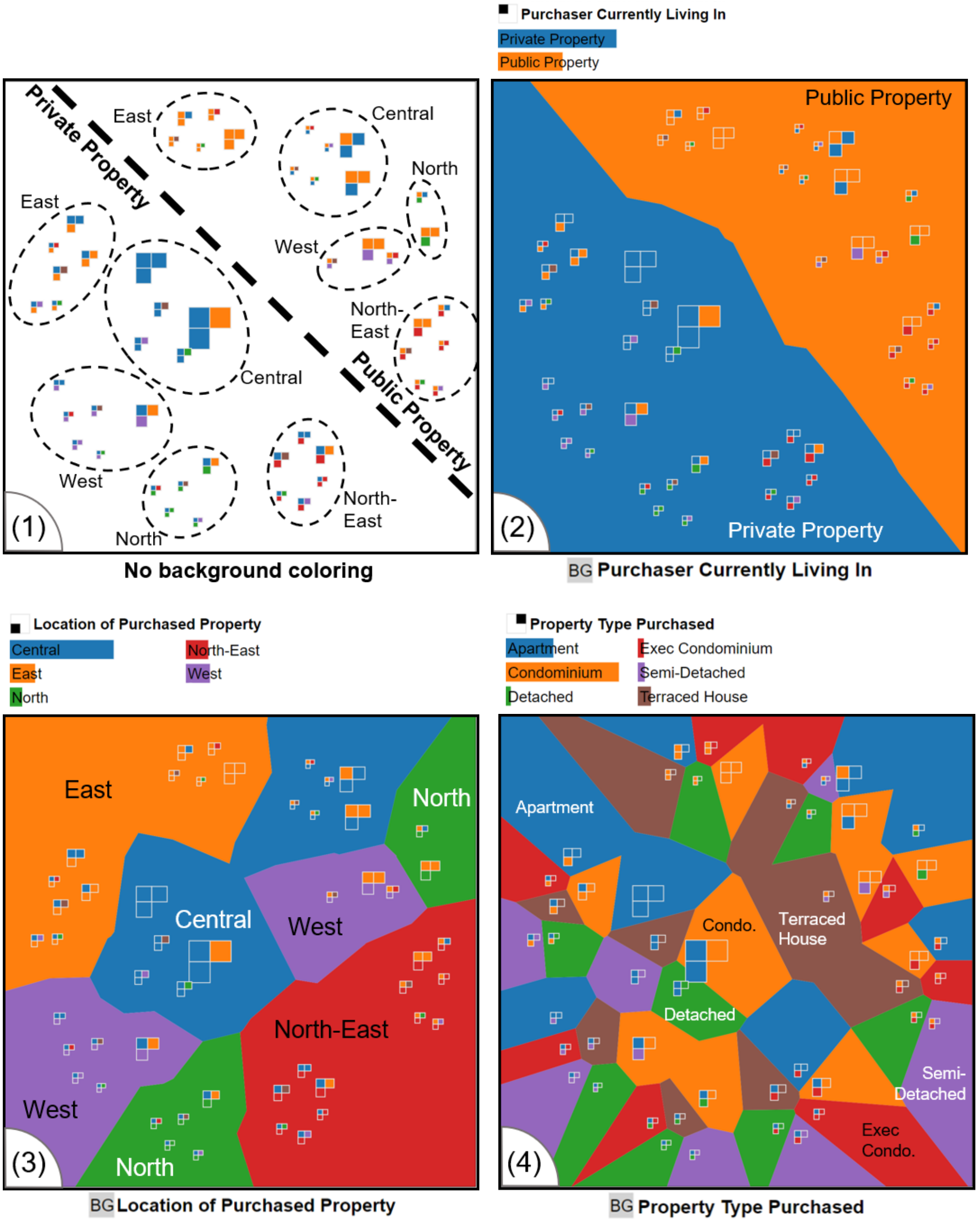


Figure 3.1: The Categorical Data Map enables projection-based analysis of categorical data here exemplified by the *Property Sales* dataset [186] with MDS [194] using the *Jaccard coefficient* [159]: (1) shows 10 groups without layout enrichment. Our method reveals the patterns annotated in (1) in plots (2)-(4). (2) shows a clear separation between Private Property vs Public Property. (3) indicates boundaries and symmetries for the Location of Purchased Property attribute, while in (4), the Property Type Purchased contributes the least to the clusters. The glyph sizes encode the subset sizes, revealing that categories Private Propriety and Central often occur together.

(C4) in section 3.3.3, describing measures to rank attributes according to their degree of splitting the embedding into connected areas. In the following, we describe how we derive distance relations of categorical data and how a projection-based approach, i.e., the Categorical Data Map, is constructed.

3.3.1 Projecting Categorical Data

The Categorical Data Map enables the visual clustering of similar categorical subsets and separating outliers, addressing (C1) and (C2). At the core, we rely on DR to create a scatterplot-like visualization. In general, we describe encoding E, distance measure M, DR method P, and overlap reduction method O to project a categorical dataset x by applying $O(P(M(E(x))))$.

Encoding (E): We convert all data items into a set representing their categorical data values. We define the set of all attributes as $\mathcal{A} := \{a_1, a_2, \ldots, a_{|\mathcal{A}|}\}$ and the possible categories associated with attribute a_i as the set $\mathcal{C}_i := \{c_i^1, c_i^2, \ldots, c_i^{|\mathcal{C}_i|}\}$ with $i \in \mathbb{N}$. $|\mathcal{A}|$ is the cardinality of a set representing a data item, i.e., the number of attributes since a data item has one category associated with each attribute. We denote a data item as $x_n = (c_1^{n_1}, c_2^{n_2}, ..., c_{|\mathcal{A}|}^{n_{|\mathcal{A}|}})$. From a practical point of view, we make sure that all categories have a unique descriptor across all attributes. We then create a representation compatible with the distance measure. We explored the *set representation* and two variants of *one-hot encoding* [57, 134].

Distance Measure (M): With the set representation, we can describe the categories of a data item to define similarity. Based on surveys on distance measures for categorical data [65, 48, 297], we chose and evaluated three set-based distance measures: *Overlap coefficient* [318], *Jaccard Similarity Index* [159], and *Sørenson-Dice coefficient* [288]. By including one-hot encoding, converting each categorical value to a new binary dimension enables us to use classical distance measures, such as *Euclidean* or *Manhattan distance*, to describe a dissimilarity relationship.

Projection Method (P): DR techniques are a set of non-/linear transformation methods with which a dataset's dimensionality can be reduced. We compared the following two DR methods.

Multiple Correspondence Analysis (MCA): This method is the categorical equivalent of PCA. MCA creates groups of items that are similar according to their categories. Objects sharing the same categories are placed close together, and objects with differing categories are placed far apart [126]. To our knowledge, MCA is the only existing technique that directly uses the set representation of categorical data.

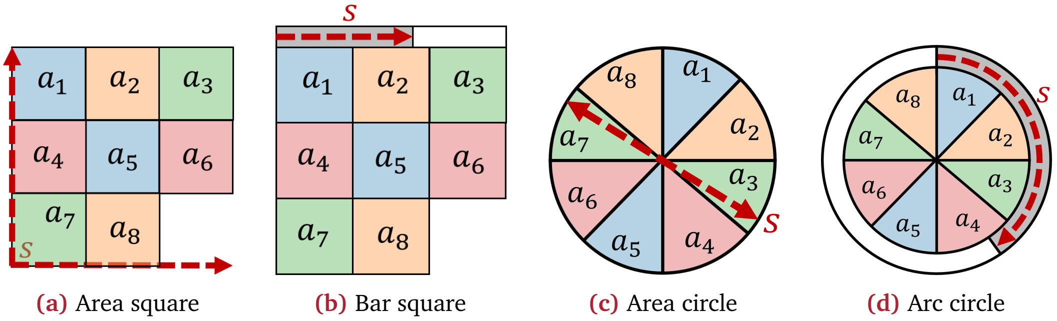
Multidimensional Scaling (MDS): This method describes a set of linear and nonlinear DR techniques that attempt to preserve pairwise distances. Multiple criteria are possible; Kruskal's stress optimization criterion is usually used [194]. We create a dissimilarity matrix to compute the projection given one of the described distance measures.

We chose the two methods based on their popularity and common usage in visual data analysis [55, 88] and compared MDS and MCA as DR methods for categorical data. However, a key difference between both methods is that MCA reduces the number of projected points to the number of unique subsets, while MDS, applied naively, would result in a number of projected points equal to the number of categorial data items. Since categorical datasets can contain many duplicates, projecting each data point individually and using a DR method for numeric data (e.g., MDS) could lead to multiple data points being projected to the same position. The main reason is that the distance of identical points is zero. To achieve a comparable result, i.e., the same number of projected points, we remove all duplicates and project one data point for each unique combination of attribute values, i.e., for each categorical subset, describing the prototype of the represented data subset. A second reason is that we want to show the subsets represented by a point irrespective of the method (e.g., MCA or MDS). We visually represent a subset's size (see section 3.3.2). Reducing the number of data points also improves the runtime of projection algorithms for datasets with duplicate items.

Overlap Reduction (O): Given that some subsets in the categorical data may differ in only one or a few attributes, these subsets will be projected close to each other. This property is desirable in the design of a map by keeping the distances representing similarity coherent. However, it may also introduce overlap if the projected point visually encodes the subset categories through a glyph representation. Additionally, points that are close together will yield small or narrow-shaped Voronoi cells. Thus, we allow users to reduce the overlap after projecting the data using a method based on force-directed graph drawing. This type of layout applies forces to the nodes and edges of a graph [185]. We add a repulsive force to all points with a strength equal to the radius of the glyph while all points are vertices of a fully connected graph, forcing all points into a configuration without overlap but with minimal space in between the glyphs.

3.3.2 Representing Categorical Data Subsets in Scatterplots

We implemented the visual components of the Categorical Data Map using D3 [50]. To represent categorical subsets, we developed four glyph representations and the layout enrichment based on experiences gained during the design phase, addressing (C3). To visualize individual categories, we use the `d3.schemeCategory10` color scale, a well-established color scale for categorical data.

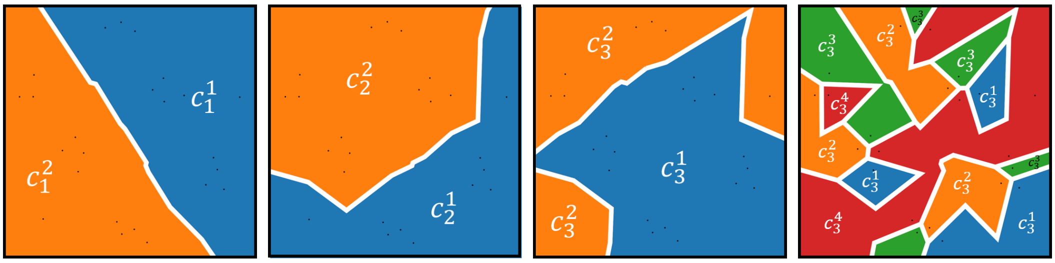


(a) Area square (b) Bar square (c) Area circle (d) Arc circle

Figure 3.2: Representation of subsets for a dataset with eight attributes. (a) shows the eight attributes in four segments with the same area while the size encodes the overall subset size. (b) shows a similar glyph, but instead, the size is encoded by a bar at the top, and all glyphs have the same size. (c) Encodes the attributes similar to the area square but is circle-shaped. (d) encodes the size by an arc filled according to the subset size.

Glyph Representation: To represent categorical subsets, we developed four glyph representations. All glyphs visualize the attributes and their respective values by dividing a square or circle into segments of equal size, such that each segment represents one attribute. This square-based glyph is inspired by pixel visualizations pioneered by Keim et al. [175]. In figure 3.2, this is represented by the categories a_1 to a_8 for the case of a dataset with eight attributes. For all glyphs, the segments are colored according to the respective category of the attribute. However, we discuss some limitations in section 6.8. The area-based glyphs represent the relative size of a subset $s \in \mathbb{N}$ by the area (see figure 3.2 (a) and (c)). Thus, we calculate the width and height accordingly. The bar- and arc-based glyphs have a fixed size to minimize space requirements and overlap issues with neighboring glyphs (see figure 3.2 (b) and (d)). To reduce overlap while preserving the relative proximity of the projected points, we decided to map a subset's size $s \in \mathbb{N}$ to a bar at the top or an arc surrounding the glyph as an alternative encoding for the subset size. Hence, each unique subset is represented by a square or circle sized relative to the percentage of data points the subset represents or an indicator filled accordingly. This enables users to perceive similar subsets and assess the size of each group.

Layout Enrichment: To enable the observation of cluster characteristics and explore attributes in the projected space, we show a Voronoi diagram [22] for a selected attribute (see figure 3.1). The Voronoi diagram automatically partitions the map into polygons such that each polygon contains exactly one subset. By selecting one attribute of interest, the partition for the selected attribute gets displayed in the background of the projection. The color of the polygon then encodes the category of the selected attribute. Thereby, it is possible to directly spot cluster regions for the selected attribute and to identify cluster boundaries and outlying data points. The appearance of the background can differ a lot across attributes (see figure 3.3).



$\mathcal{F}_{\text{edge}}(a_1) = 0.17$ $\mathcal{F}_{\text{comp}}(a_1) = 0$

$\mathcal{F}_{\text{edge}}(a_2) = 0.17$ $\mathcal{F}_{\text{comp}}(a_2) = 0$

$\mathcal{F}_{\text{edge}}(a_3) = 0.42$ $\mathcal{F}_{\text{comp}}(a_3) = 0$

$\mathcal{F}_{\text{edge}}(a_4) = 0.79$ $\mathcal{F}_{\text{comp}}(a_4) = 0.69$

Figure 3.3: The *fracturedness* of attributes differs a lot and can imply an order, i.e., increasing from left to right. The examples are derived from the Titanic dataset [80]. The edge-based (i.e., $\mathcal{F}_{\text{edge}}$) and component-based fracturedness (i.e., $\mathcal{F}_{\text{comp}}$) values are provided below for each attribute.

Attributes form distinct contiguous areas of different sizes, indicating a neighborhood or larger area of subsets of the same category. We added *detail-on-demand* using tooltips, allowing users to see the respective category for each polygon directly.

3.3.3 Measuring Fracturedness

We quantify *fracturedness*, generally defined as the strength with which the Voronoi partitioning of an attribute appears disjointed and fractured (see figure 3.3). We use *fracturedness* to suggest attributes for analysis, e.g., the lower the fracturedness value, the larger the contiguous areas of categories and thus the more straightforward to orient along, addressing (C4). We use the Delaunay triangulation of the Voronoi diagram [22] as a basis for our measures. In contrast to Aupetit and Catz [20], we describe measures for purely categorical datasets. Before describing the measures, we define the common notations following established notations [73, 20]. Let $G \coloneqq (V, E)$ be the Delaunay triangulation of the discrete set of points P resulting from the projection (see section 3.3.1). Thus, G is an undirected graph and the dual graph of the Voronoi diagram of the points P. Therefore, there exists exactly one $v \in V$ for every $p \in P$ defining its x,y-location and categories. Each vertex $v \in V$ has exactly one associated category $\mathscr{C}_n(v) \in \mathcal{C}_n$ for each attribute $a_n \in \mathcal{A}$.

Edge-based Fracturedness: We measure the number of edges in G that connect cells with different associated attributes. This concept is shown in figure 3.4. We define an edge $e \in E$ as $\{v_1, v_2\}$ with $v_1, v_2 \in V$ and $v_1 \neq v_2$. An edge contributes to *fracturedness*, if the category for the analyzed attribute a_n and its associated categories in $\mathcal{C}_n$ differ for the connected vertices, i.e., $\mathscr{C}_n(v_1) \neq \mathscr{C}_n(v_2)$ for

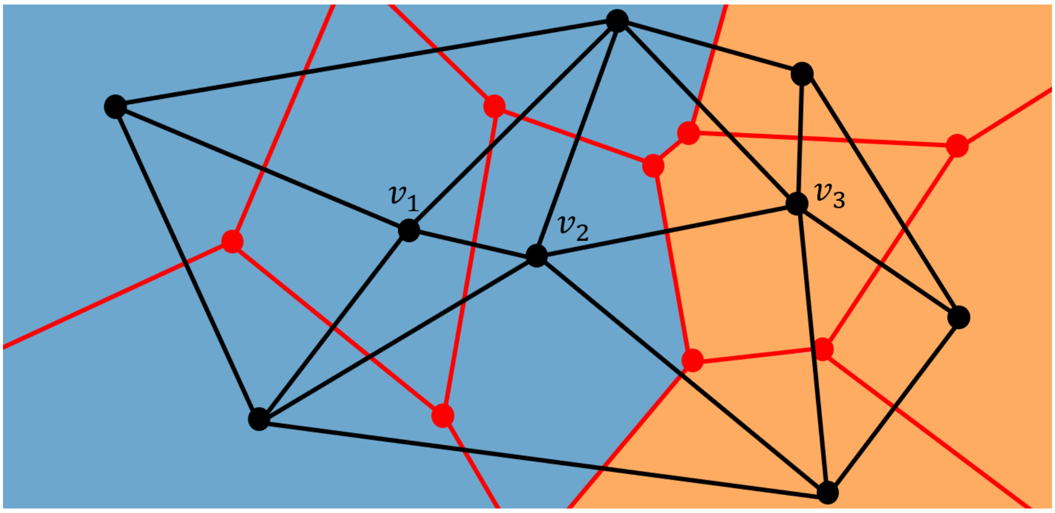


Figure 3.4: We illustrate *edge-based fracturedness* with a Delaunay triangulation shown in black, and a Voronoi partitioning with cell borders shown in red. The cells are colored according to the categories of an attribute. v_1, v_2 and v_3 are vertices of the Delaunay triangulation. The edge v_1, v_2 will not contribute to edge-based fracturedness, since it connects cells representing the same category of a given attribute. Edge v_2, v_3 contributes to edge-based fracturedness because it connects cells representing different categories.

$\{v_1, v_2\} \in E$. *Edge-based fracturedness* is defined as $\mathcal{F}_{edge} : \mathcal{A} \mapsto [0, 1]$ and calculated using equation (3.1).

$$\mathcal{F}_{edge}(a_n) := \frac{|\{v_1, v_2\} \in E : \mathscr{C}_n(v_1) \neq \mathscr{C}_n(v_2)|}{|E|} \text{ with } a_n \in \mathcal{A} \tag{3.1}$$

Component-based Fracturedness: This measure quantifies the number of continuous areas an attribute produces in the plot through its categories. We show the concept of *component-based fracturedness* in figure 3.5. Each category $c \in \mathcal{C}_n$ defines an induced subgraph $G[S(c)]$ of G, with $S(c) \subset V$ for all $c \in \mathcal{C}_n$ of an attribute $a_n \in \mathcal{A}$. The induced subgraph $G[S(c)]$ is a graph with the vertices $S(c)$ and the edges in E with both of its vertices in $S(c)$. We formally define $S(c)$ for a category $c \in \mathcal{C}_n$ in equation (3.2).

$$S(c) := \{v \mid v \in V, \mathscr{C}_n(v) = c\} \text{ for } c \in \mathcal{C}_n \text{ of } a_n \in \mathcal{A} \tag{3.2}$$

With this definition, a category defines a partition of V, i.e., $\bigcup_{c \in \mathcal{C}_n} S(c) = V$ and a vertex $v \in V$ can only have one category $\mathscr{C}_n(v)$, thus $\bigcap_{c \in \mathcal{C}_n} S(c) = \emptyset$ for a given attribute a_n. Therefore, there exits $|\mathcal{C}_n|$ subgraphs of G for attribute $a_n \in \mathcal{A}$. Let $\omega(G)$ be the number of connected components of any graph G. The *component-based fracturedness* is dependent on the number of connected components of all subgraphs $\omega(G[S(c)])$ for each $c \in \mathcal{C}_n$ (see s_1 to s_6 in figure 3.5). We define the sum of the

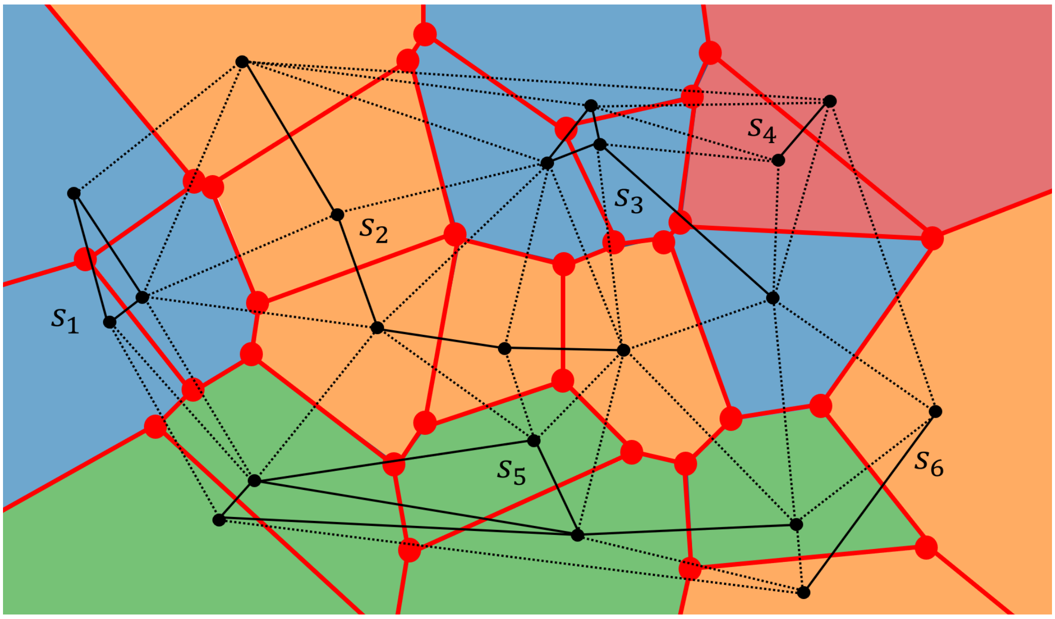


Figure 3.5: We describe *component-based fracturedness* with a Voronoi partitioning with cell borders shown in red. The associated Delaunay triangulation is shown in black. The cells are colored according to the categories of an attribute. s_1 to s_6 are six components induced by an attribute through the subgraphs associated with a category. Solid lines connect each subgraph, while dashed lines are not part of any subgraph. With six components $\mathcal{F}_{comp} = 0.33$ for the attribute (see equation (3.5)).

number of components of all induced subgraphs as $\Omega(a_n)$ for an attribute $a_n \in \mathcal{A}$. $\Omega(a_n)$ is formally defined in equation (3.3):

$$\Omega(a_n) := \sum_{c \in \mathcal{C}_n} \omega(G[S(c)]) \text{ with } a_n \in \mathcal{A} \tag{3.3}$$

We can also quantify the fracturedness a single category contributes to the overall measure. This allows us to differentiate categories forming contiguous areas and highly fractured ones. The fracturedness $f_{\text{comp}}(c)$ of a single category $c \in \mathcal{C}_n$ is defined in equation (3.4):

$$f_{\text{comp}}(c) := \frac{\omega(G[S(c)]) - 1}{\Omega(a_n)} \text{ with } c \in \mathcal{C}_n \text{ of } a_n \in \mathcal{A} \tag{3.4}$$

Component-based fracturedness is defined as $\mathcal{F}_{comp} : \mathcal{A} \mapsto [0, 1]$ and calculated using equation (3.5). It allows us to compare different attributes and is an alternative measure to $\mathcal{F}_{edge}(a_n)$.

$$\mathcal{F}_{\text{comp}}(a_n) := 1 - \frac{|\mathcal{C}_n|}{\Omega(a_n)} \text{ with } a_n \in \mathcal{A} \tag{3.5}$$

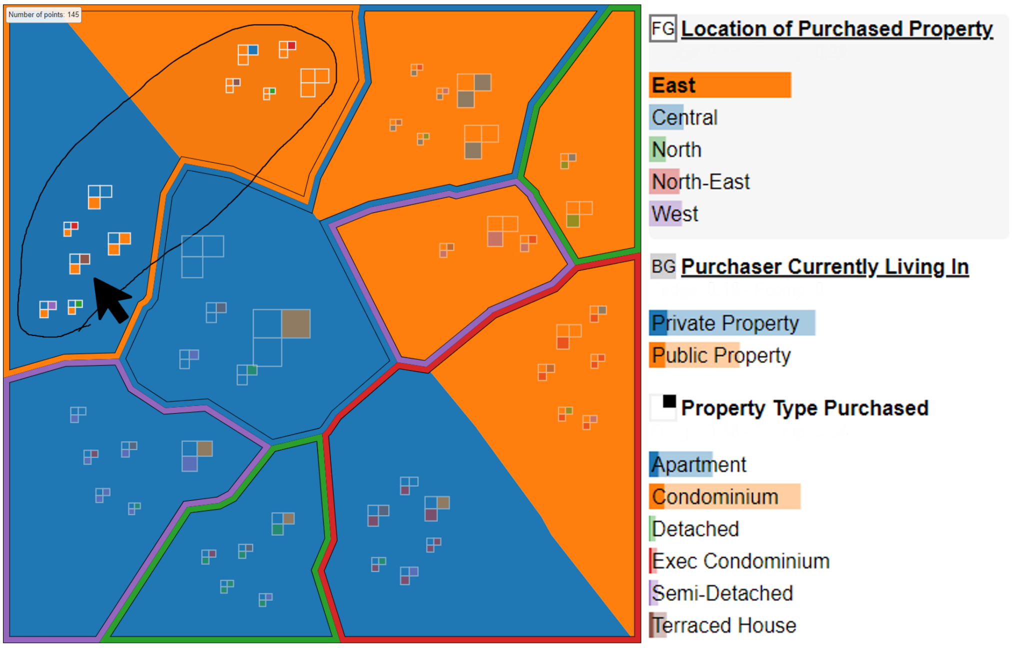


Figure 3.6: Through user selection, the borders of a second attribute can be added to the foreground of the plot, e.g., Purchaser Currently Living In is shown in the background as the primary attribute, and Location of Purchased Property is shown in the foreground.

The sum of all component-based fracturedness values of individual categories $c \in \mathcal{C}_n$ is equal to the fracturedness of the attribute $a_n \in \mathcal{A}$. We express this relationship in equation (3.6):

$$\mathcal{F}_{\text{comp}}(a_n) = \sum_{c \in \mathcal{C}_n} f_{\text{comp}}(c) \text{ with } a_n \in \mathcal{A} \tag{3.6}$$

We proof of the equivalence described in equation (3.6).

Proposition: For any $a_n \in \mathcal{A}$, $\mathcal{F}_{comp}(a_n) = \sum\limits_{c \in C_n} f_{comp}(c)$.

Proof:

$$\begin{aligned}
\sum_{c \in C_n} f_{comp}(c) &= \sum_{c \in C_n} \frac{\omega(G[S(c)]) - 1}{\Omega(a_n)} \\
&= \sum_{c \in C_n} \frac{\omega(G[S(c)])}{\Omega(a_n)} - \sum_{c \in C_n} \frac{1}{\Omega(a_n)} \\
&= \frac{\sum_{c \in C_n} \omega(G[S(c)])}{\Omega(a_n)} - \frac{|C_n|}{\Omega(a_n)} && \text{use eq. 3.3} \\
&= 1 - \frac{|C_n|}{\Omega(a_n)} \\
&= \mathcal{F}_{comp}(a_n)
\end{aligned}$$

Hence, $\mathcal{F}_{comp}(a_n) = \sum\limits_{c \in C_i} f_{comp}(c)$ for any $a_n \in \mathcal{A}$. ■

3.3.4 Interacting with Attributes and Subsets

Our prototype allows interactions on the attributes of the dataset shown in the side panel and projected subsets.

Attribute Selection: Users can change the attribute visualized through layout enrichment. We also show the outline for categories of a second selected attribute (see figure 3.6). We add the borders of categories to the foreground if another attribute is already selected and visualized in the background. This visual cue does allow for the observation of one main attribute and a second attribute, similar to the outline of MosaicSets [262]. This introduces less clutter and thus requires less effort to perceive. We initially used textures with different colors to represent different categories. However, using textures of different colors to fill each cell in the Voronoi portioning introduced excessive clutter, and the interpretation of common regions was difficult.

Subset Selection: We allow for the selection and highlighting of groups of subsets. Once the user has selected data items, we show the common categories of the selection using Lasso selection and highlight all data items outside of the selection with the same combination of categories in the side panel on the left, similar to the proximity visualization for continuous data proposed by Aupetit and Catz [19]. This interaction enables cluster analysis since all common categories among the selected items are highlighted (see side panel in figure 3.6). Thus, visual groupings can be compared with respect to the categories and attributes contributing to cluster cohesion. Additionally, all subsets matching the common categories of the selection are also highlighted (see plot in figure 3.6). Together, this allows analysts to observe and judge group cohesion along with the contributing attributes.

Attribute and Category Ordering: A user can select attributes of the dataset listed on the side panel to change the attribute encoded in the foreground and background of the plot. By default, attributes are sorted by their edge-based fracturedness in ascending order, and categories are ordered by their individual contributions to component-based fracturedness in ascending order, allowing for a focus on attributes forming clear splits in the projection space. When selecting subsets (see previous paragraph), the lists of common attributes and distinct attributes are also ordered similarly.

3.4 Interpreting the Categorical Data Map

In the following, we perform a case study on cluster and attribute analysis, using the Property Sales dataset [186] (see figure 3.1) to show how to interpret emerging patterns for cluster, outlier, and similarity analysis. We chose this dataset because of its relative simplicity. However, it lacks the complexity of large categorical datasets, which we will address in an expert study (see section 6.7).

Cluster Analysis: There exist a total of $\Pi_{n\in\{1,\ldots,|\mathcal{A}|\}} |\mathcal{C}_n|$ possible data items, given that all combinations of attributes are allowed, resulting in an exponential growth in the number of possible and unique data items. Hence, we can assume that there are dependencies and relationships among the categories contained in a dataset impacting their distribution. This means that groups of subsets that share a set of attributes should form perceivable structures (i.e., clusters) when projected using DR methods. Thus, our approach benefits from and leverages the sparsity of categorical data.

For the Property Sales dataset, there are ten clusters (see figure 3.1 (1)). There is a symmetric split along the center of the projection. Given the size of this dataset, we can observe that the two attributes Purchaser Currently Living In and Location of Purchased Property dominate the appearance of the projection. The glyph sizes indicate that the categories Private Propriety and Central often occur together while {Private Propriety, Central, Condominium} is the largest unique subset. Thus, we can see that most private property is purchased in the central areas, and in this general group, the large majority are condominiums.

Attribute Analysis: By encoding the attribute values in the background, we enable users to analyze the distribution of subsets in the projection with respect to one or two attributes. For the Property Sales dataset, we found that the attribute Purchaser Currently Living In creates a clear and straight division between subsets (see figure 3.1 (2)). We can also see a second level of grouping by the Location of Purchased Property attribute forming a close to orthogonal split in the projection, which can be spotted with our visualizations (see figure 3.1 (3) and figure 3.6). Thus, Purchaser Currently Living In and Location of Purchased Property are the primary attributes. This finding is substantiated when checking the side panel entry of the attribute Property Type Purchased, which has three categories with low frequency. The appearance of the partitioning depends a lot on the selected attributes. When observing the layout enrichment, attributes present themselves on a spectrum from a few clearly separated groups to intermingled and highly fractured appearances. Property Type Purchased does not contribute to elements' clustering (or cluster cohesion) since most groups contain subsets of the majority of its categories. Thus, the areas of the categories are disjointed, which reflects the fact that most property types are sold as both private and public property, as well as most of the geographic locations.

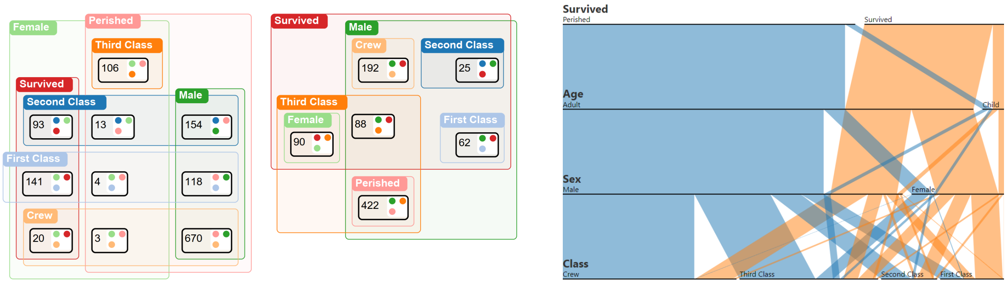


Figure 3.7: Two visualizations of the Titanic dataset [80]. A *split* Euler diagram without the Age attribute (left) and an overlap reduced Parallel Sets visualization (right) with *very thin ribbons*. Both have drawbacks with a small dataset and do not scale with an increasing number of attributes.

3.5 Evaluation

We qualitatively compare our Categorical Data Map to existing visualizations for categorical data and quantitatively compare our approach to MCA used by Broeksema et al. [55]. Additionally, we performed an expert study on two representative datasets with five data scientists.

3.5.1 Comparison to Euler Diagrams and Parallel Sets

For categorical data, each data point has exactly one category for each attribute, while in Euler diagrams, the number of sets an element is included in is not restricted, i.e., it could be in less. Thus, to truthfully represent categorical data in Euler diagrams, there need to be $\Sigma_{a_i \in \mathcal{A}} |\mathcal{C}_i|$ sets, i.e., one set for each category of all attributes. Euler diagrams may require the selection of specific subsets of attributes and, therefore, are less suitable for exploratory data analysis. For highly intersecting sets, automatic layout methods might not create a single diagram [237]. We show an example of an automatically generated split Euler diagram for the Titanic dataset in figure 3.7 (left). The attribute Age was removed to reduce the diagram's complexity. The Titanic dataset requires ten sets. However, even with eight sets, the visualization is disjointed. Parallel Sets are alternative categorical sets visualization, combining principles from stacked bars and parallel coordinate plots [35, 190]. Figure 3.7 (right) shows the Titanic dataset in a Parallel Sets visualization, where the readability is improved through overlap reduction. Small subsets are represented as very thin ribbons on the lowest level, which can be hard to perceive. Visualizing the Mushroom dataset with classical Parallel Sets is not visually feasible since it will have 22 ribbon layers and 8123 subsets on the lowest level (see figure 3.8). Alsakran et al. [9] addressed this issue by only visualizing 2-dimensional subsets in a modified Parallel Sets visualization. However, the relation between 2-dimensional

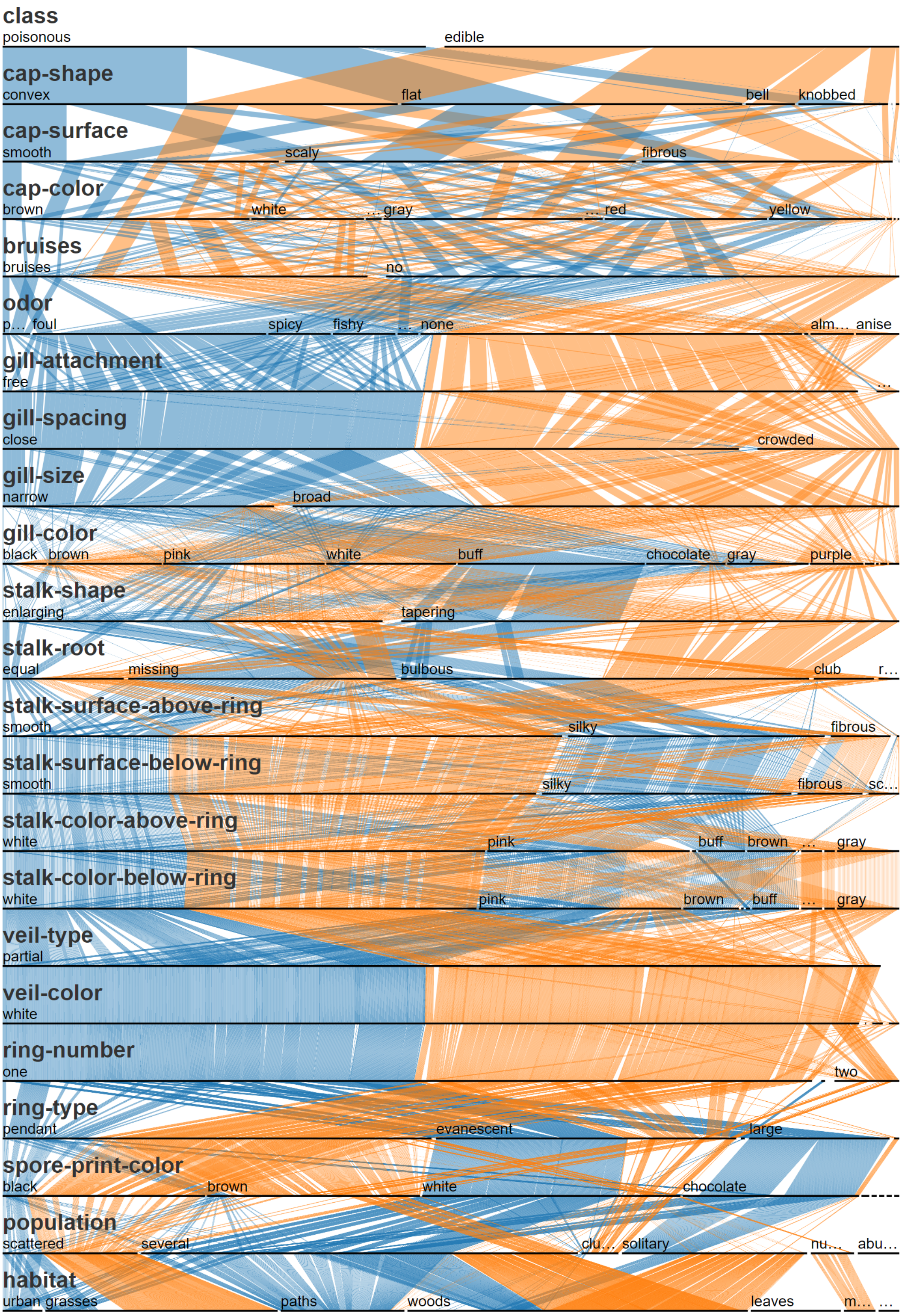


Figure 3.8: A basic Parallel Sets visualization of the Mushroom dataset. The visualization exhibits a high ribbon overlap, with many thin ribbons, especially in the lower levels, contributing to low readability [85].

Source	Description	Num. Attr.
Bareiss et al. [27]	Audiology	70
Lincoff et al. [202]	Mushrooms	23
Dawson et al. [80]	Titanic dataset	4
Hassan et al. [140]	Cyber-security	4
Koh et al. [186]	Property sales	3
Rodgers et al. [256]	HCI study; 2 datasets	3

Table 3.1: The original sources of the seven datasets we used for evaluation, a description, and the number of attributes.

subsets is lost. Thus, we argue that Euler diagrams and Parallel Sets, as examples of established visualizations for categorical data, do not scale with an increasing number of attributes.

3.5.2 Quantitative Evaluation of Projection Quality

We use five quality measures commonly used in related work for DR to evaluate and compare the quality of our categorical data projections [104]. The result of comparing MDS with Overlap coefficient (*MDS+O*) and Jaccard distance (*MDS+J*) to MCA are shown in table 3.2. We briefly describe each measure below and use them to compare our MDS-based method to MCA using seven real-world categorical datasets (see table 3.1).

Trustworthiness (TW) [317] quantifies the proportion of points that remain close in the lower-dimensional representation to assess how accurately local patterns in the projection represent the data patterns. This is linked to the occurrence of "false neighbors" in the protection. The **TW** quality measures, as presented by Venna and Kaski, is defined as:

$$\mathbf{TW}(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k) \tag{3.7}$$

In the definition given by Equation 3.7, $U_k(i)$ denotes the k nearest neighbors of a point i in the 2D projection that are not neighboring in the original space. $r(i,j)$ represents the rank of the 2D point j within the ordered nearest neighbors of i in 2D (i.e., projection space).

Dataset	**TW** (↑)			**CT** (↑)			**SC** (↑)			**NS** (↓)			**Avg. NH** (↑)			**Med. NH** (↑)		
	MDS+O	**MDS+J**	**MCA**	**MDS+O**	**MDS+J**	**MCA**	**MDS+O**	**MDS+J**	**MCA**	**MDS+O**	**MDS+J**	**MCA**	**MDS+O**	**MDS+J**	**MCA**	**MDS+O**	**MDS+J**	**MCA**
Audiology [27]	.58	**.89**	.83	.64	**.90**	.89	.29	**.77**	.69	.17	**.09**	.81	.89	**.92**	**.92**	.98	.98	.98
Mushroom [202]	.96	**.97**	.91	.92	.93	**.97**	.78	.77	**.79**	**.08**	.09	.64	.89	**.90**	.84	.90	**.92**	.87
Titanic [80]	**.86**	**.86**	.76	**.84**	**.84**	.81	**.76**	.75	.59	**.07**	.07	.28	**.68**	**.68**	.63	.74	**.75**	.60
Cyber-Security [332]	.84	**.87**	.79	.83	**.86**	.81	**.87**	.82	.68	**.04**	.06	.30	**.56**	.55	.54	**.66**	.62	.58
Property Sales [186]	**.91**	.89	.73	**.86**	.85	.81	**.70**	.66	.46	**.09**	.10	.24	**.65**	**.65**	.51	**.76**	**.76**	.39
HCI Study (1) [256]	**.84**	.77	.71	**.81**	.79	.73	**.72**	.69	.61	**.07**	**.07**	.18	**.59**	.58	.58	**.53**	**.53**	.52
HCI Study (2) [256]	1.0	1.0	1.0	1.0	1.0	1.0	.86	.85	**.89**	**.03**	**.03**	.11	.56	.56	.56	.57	.57	.57

Table 3.2: We compare projections of MDS using the Overlap coefficient (*MDS+O*) and the Jaccard distance (*MDS+J*) to MCA by applying them to seven real-world datasets. The MDS outperforms MCA for most datasets and quality measures. In the case of the *Audiology* dataset with high category overlap, usually present in datasets with many attributes, we found that MDS combined with the Jaccard distance always outperforms both alternatives.

Continuity (CT) [317] measures the ratio of points in the projection that remain close in the original space. This is related to the "missing neighbors" of a projected point. The **CT** measure, as defined by Venna and Kaski, is formulated as:

$$\mathbf{CT}(k) = 1 - \frac{2}{Nk(2N-3k-1)} \sum_{i=1}^{N} \sum_{j \in V_k(i)} (\hat{r}(i,j) - k) \tag{3.8}$$

In Equation 3.8, $V_k(i)$ represents the set of points that are nearest neighbors in the original space but not among the k nearest neighbors in the 2D projection. $\hat{r}(i,j)$ is point j's rank in the ordered set of nearest neighbors of point i in the original space.

Normalized Stress (NS) [165] quantifies how well the distances between pairs of points are preserved when mapping from the original space to the projected space. This measure should be as low as possible. The **NS**, as presented by Joia et al., is formulated as:

$$\mathbf{NS}(x) = \frac{\sum_{i,j}(d(x_i,x_j) - d(P(x_i),P(x_j)))^2}{\sum_{i,j} d(x_i,x_j)^2} \tag{3.9}$$

The function d denotes the Euclidean distance. P denotes the transformation of x_i and x_j into the projection space (here 2D) and using MDS or MCA.

Shepard Diagram Correlation (SC) [103] measures the rank correlation of all distances of the original and the projected space, assessing the quality of distance preservation globally using Spearman's ρ [290]. The correlation coefficient of the Shepard diagram, as presented by Espadoto et al., is formulated as:

$$\begin{aligned}\mathbf{SC}(x) = \rho(&R(\{d(x_i,x_j) | x_i,x_j \in x\}),\\ &R(\{d(P(x_i),P(x_j)) | x_i,x_j \in x\}))\end{aligned} \tag{3.10}$$

This definition uses Spearman's ρ for calculating the rank correlation. The rank of distances is denoted by R, while the distances are all pairwise distances x_i, x_j of dataset x. As for **NS**, P denotes the transformation of a data point into the projection space.

Neighborhood Hit (NH) [241] measures the proportion of a point's neighbors in the projection space that share the same label as the point itself, averaged across all points in its neighborhood. This measure is related to the separation of labeled data in the projection. In our case, we evaluate every attribute as a set of labels. Thus, we calculate the mean and median values of **NH** across all attributes of a dataset. Paulovich et al. define **NH** as:

$$\mathbf{NH}(k) = \sum_{i=1}^{N} \frac{\left| j \in N_i^k : l_j = l_i \right|}{Nk} \tag{3.11}$$

This measure quantifies the proportion of a point's k neighbors N_i^k in the projection space that shares the same label l as the point itself, averaged across all points in N_i^k (see Equation 3.11). This measure is related to the separation of labeled data in the projection.

TW, **CT**, **NH** require a parameter k defining a neighborhood size. We set $k = 7$, a commonly used value [103]. We found that our approach generally outperforms MCA quantitatively.

3.5.3 Qualitative Expert User Study

To evaluate the Categorical Data Map we performed a paired analytics study [169]. We conducted an expert study with five data scientists, **E1–E5**, with varying backgrounds. All participants were Ph.D. candidates and students. All were male, and the age range was 25 to 30 years. All experts had experience in the area of information visualization and visual analytics. During the study, we asked the experts to verbalize their thought process to capture it. The following studies are set up using MDS projections of the Mushroom and Titanic dataset using the Overlap coefficient (*MDS*+*O*). Table 3.2 shows that these projections are higher quality than MCA-based ones regarding most quality measures.

All trials followed a predefined structure and took between 43 and 57 minutes. The study was conducted in German. The study started with an introduction to the Categorical Data Map using the Property Sales dataset by Hassan et al. [140] shown in figure 3.1 and included a description of the square area glyph, layout enrichment, and interactions to introduce the expert to the prototype. After the introduction, the experts had the opportunity to ask questions regarding our approach.

Titanic Dataset: The experts had to analyze the Titanic dataset [80] using the Categorical Data Map shown in figure 3.9. **E1–E5** were able to locate the largest subset {Male, Perished, Adult, Crew} by looking at the visualization without any additional interaction (figure 3.9 (1)). **E1–E5** used Lasso selection to find and validate that the largest subset regarding three attributes is {Male, Perished, Adult} (figure 3.9 (2)). Additionally, **E1–E5** were able to find six clusters and two outliers. **E1**, **E3**, and **E4** found that the outliers represent the subsets defined by the categories {Perished, Child, Crew} (figure 3.9 (3)). **E1**, **E3**, and **E5** commented on the high number of perished males and the large number of casualties among the {Male, Crew}. **E1–E5** used the layout enrichment to navigate and reason about the location of subsets, including the Class attribute (figure 3.9 (4)). **E2** commented on the close to orthogonal split in the projection between Sex and Survived shown in figure 3.9 (1) and (2).

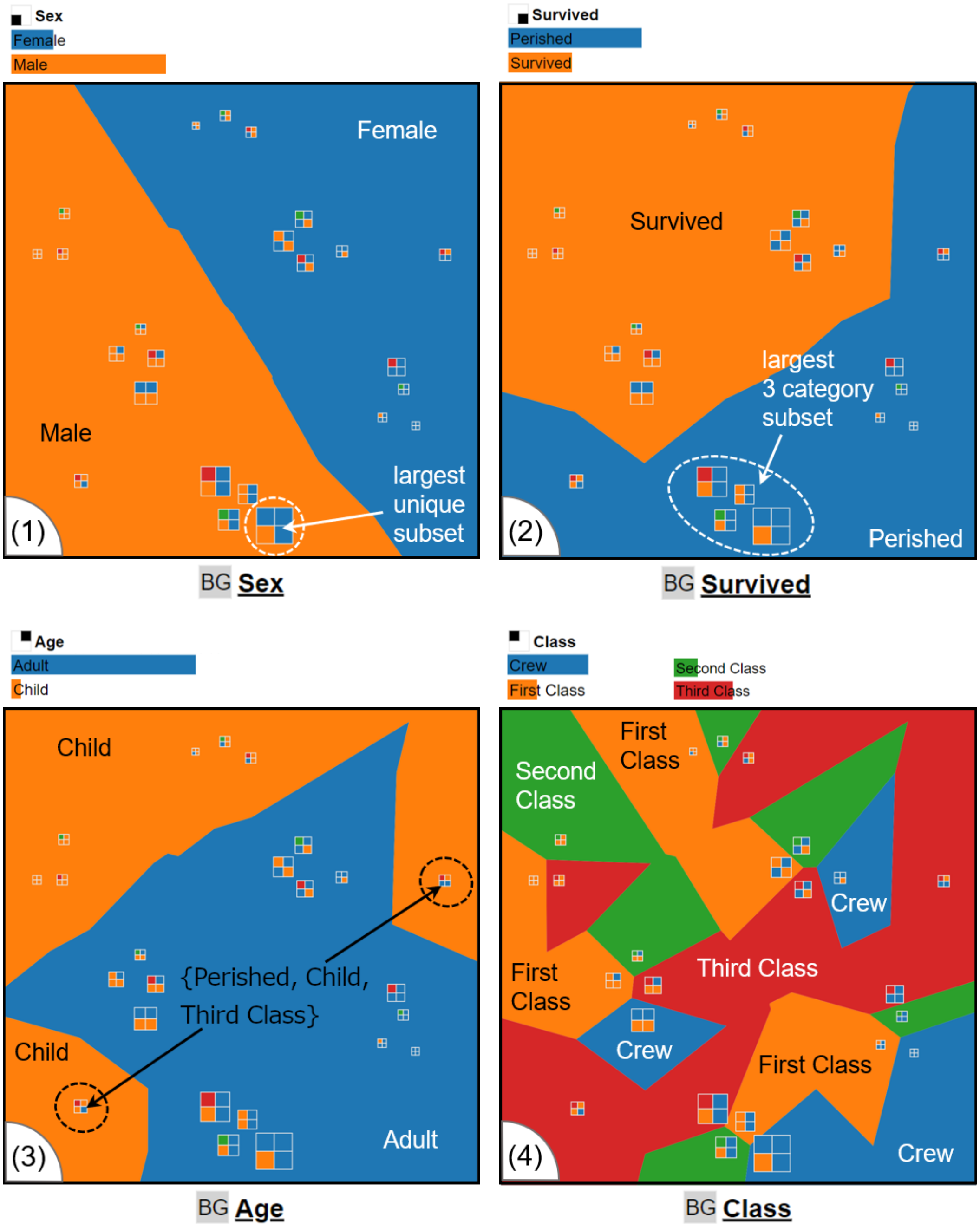


Figure 3.9: Categorical Data Map visualizations of the *Titanic* dataset [80] using MDS [194] and *Overlap coefficient* [298]. (1) The visualization shows six clusters and two outliers. The largest cluster is the subset of Adult, Male, Perished (at the bottom). The background encoding shows that the Survived and Sex attributes are relevant for this dataset, clearly separating the data items. For Sex, the separation is left and right. (2) For Survived, the separation is bottom-right/top-left. (3) The Age dimension also yields a separation, while (4) Class shows no clear structure.

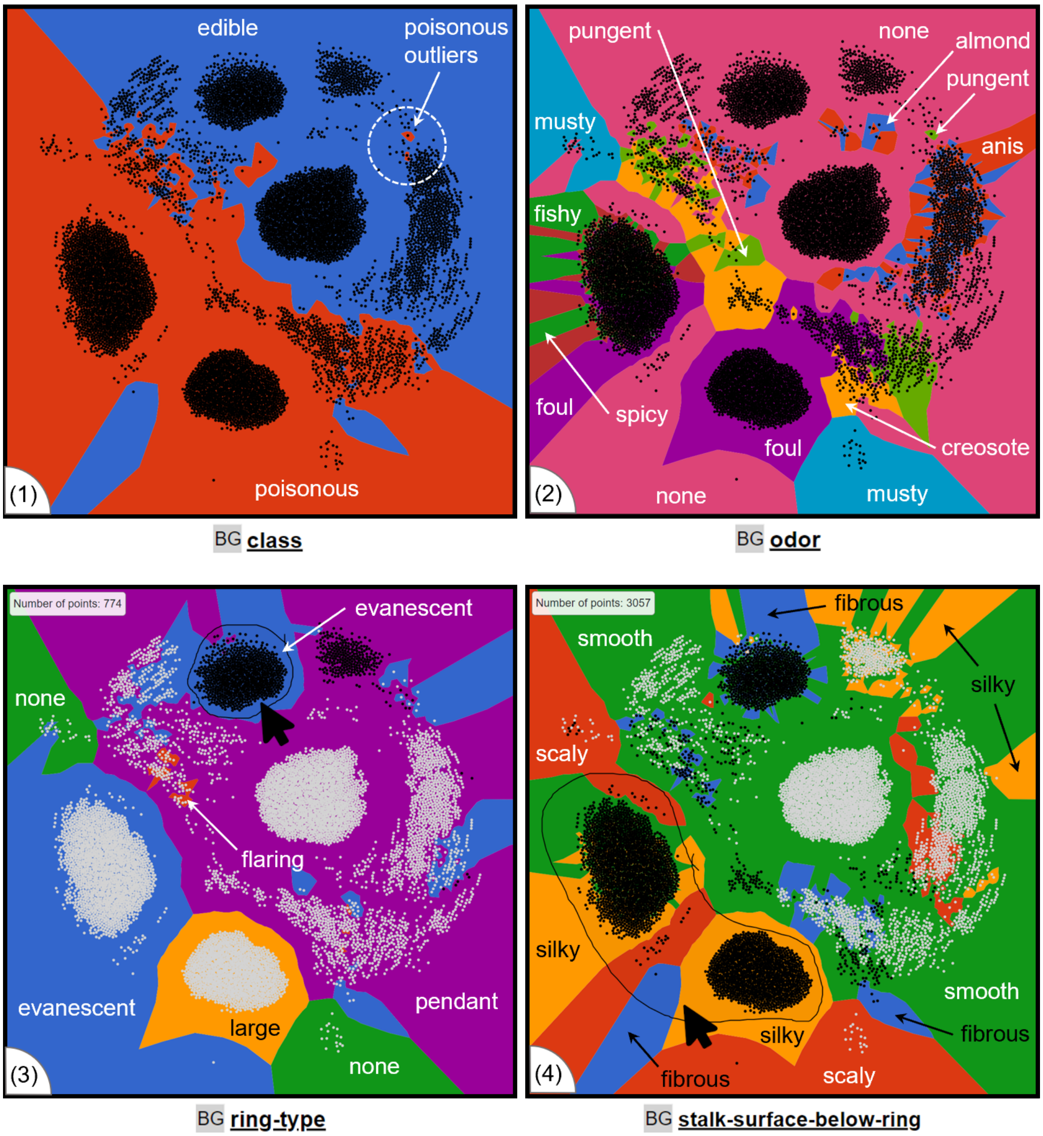


Figure 3.10: Categorical Data Map visualizations of the *Mushroom* dataset [202] using the MDS [194] and *Overlap coefficient* [298]. (1) Two poisonous mushrooms very similar to edible mushrooms. (2) Comparing class and odor reveals that the poisonous outlier has a pungent odor. Continued analysis reveals that mushrooms with an unpleasant smell are poisonous. (3) After the selection of a cluster, the ring-type is identified as a defining characteristic for the cohesion of visible clusters and is used as a property for the classification of mushrooms. (4) Selecting two poisonous clusters, reveals that the vast amount of poisonous mushrooms are silky at the stalk-surface-below-ring, while there exist very few silky mushrooms that are edible.

Mushroom Dataset: E1–E5 had the opportunity to perform an open exploration task and were only given the information that the dataset is about mushrooms and that the class attribute indicates their poisonousness. The glyph was replaced with a simple black dot to reduce the visual complexity. **E1–E5** perceived five clusters right at the outset. **E2** and **E3** used the Lasso selection together with layout enrichment to determine differentiating categories for cluster separation, e.g., evanescent, large, and pendent for the ring-type attribute (figure 3.10 (3)). **E1–E5** found the poisonous outliers nested in the group representing edible mushrooms (figure 3.10 (1)) being poisonous mushrooms very similar to edible ones. **E1–E5** found the general rule that mushrooms with an fishy foul, musty, spicy or other unpleasant smells indicate a poisonous mushroom (figure 3.10 (2)). During the open exploration task, **E1**, **E2**, and **E3** found the rule without additional information. **E4** and **E5** needed help to find the class and odor combination. However, **E4** and **E5** could deduce the rule by only interpreting the plot. **E3**, quickest in exploring the dataset, found that stalk-surface-below-ring is silky for the majority of poisonous mushrooms and the stalk-surface-below-ring is mostly smooth for edible ones (figure 3.10 (4)).

General Comments: Before concluding the study, the participants were asked to comment on their preferences for the available glyph designs. **E1**, **E2**, and **E4** preferred a circular glyph design (figure 3.2 (c) and (d)) over a square design. **E3** and **E5** preferred square glyph designs (figure 3.2 (a) and (b)). **E1** and **E3** found that the area-based glyphs are inferior to the alternative designs for reading off precise subset sizes. **E1** mentioned as a drawback that the glyphs are not rotation invariant. **E1** commented that the layout enrichment is very useful for navigation and orientation and helps to perceive the impact on category groups. However, **E1** also noted that the layout enrichment does not reflect the ratio of data items with a given category. **E3** mentioned a general preference for the map metaphor by being helpful for orientation among different subsets. **E2** mentioned potential scalability issues with the glyph for large datasets, e.g., for a high number of attributes, and proposed semantic zoom as a potential option. **E1–E5** commented that ordering attributes according to their fracturedness was understandable and useful. During the general questions at the end, **E2–E4** freely explored plots created with other distance measures and DR methods. **E3** commented that the result of MCA-based plots was hard to interpret, noticing the disjointed layout enrichment and thus having larger fracturedness. **E1** mentioned issues with the encoding of categories, such as the category North not being located north of the plot or the category brown not having the color brown, and suggested being able to select the color of a category manually.

3.6 Discussion and Future Work

In this section, we discuss the lessons-learned, reflect on the design decisions, and discuss computational complexity and future work.

Visualizing Attributes and Categories: We initially used circular glyphs as shown in figure 3.2 (c) and (d), which had the benefit of using the available space effectively since overlap minimization relative to the radius is straightforward to implement. The subset size encoding by the arc around the circle enables finer-grained distinction of sizes since it offers more space. However, during the design phase, users misinterpreted the circle segments as pie charts, a common method for displaying categorical data. Thus, we decided to circumvent this common misconception by using square-based representation for the categorical subsets. However, three out of five experts preferred a circular glyph design.

There are visual limitations to the number of dimensions and categories that our approach is able to support. The number of visually distinguishable categories is limited by the number of square segments that fit into the glyph, which is limited by the screen space. The number of attributes is limited by the number of colors, which have to be distinguishable and memorizable. Thus, we suggest following Miller's Law [220] for the number of dimensions and attributes, which proposes a maximum of seven plus or minus two. Alternatively, we suggest interactions such as semantic zoom, e.g., removing attributes for which all subsets have the same category after zooming in on a specific area.

Encoding of Subset Sizes: We evaluated four different visual encodings for the size of a categorical data subset (see figure 3.2). The area-based glyph makes it easier to perceive subset sizes at a glance, and thus, a user can spot the distribution of the dataset directly. Still, it suffers from overlap, especially for tight clusters. Thus, there is a benefit to applying methods to reduce overlap. We are able to mitigate the overlap problem with the force-directed overlap reduction largely. Simplifying the representation of a dot requires less space, but the assessment of subset sizes requires interaction. It is possible to remove the subset size information altogether. However, this may limit analysis tasks where the subset sizes it not important, e.g., the Mushroom dataset. All glyph designs benefit from a mouse-over mechanism that moves the currently selected glyph to the top so that all attributes can be observed.

Encoding of Attributes Into the Background: Figure 3.9 shows that encoding an attribute into the visualization gives insight into the topology of the projection. We could also show the benefit of encoding multiple attributes into the background to allow for a more complex representation of the topology. We found that the number of categories of an attribute weakly influences the fracturedness of an attribute. However, the main factor is the number of subsets containing the attribute, i.e., an attribute with two categories and an occurrence roughly equal among all subsets will yield a low fracturedness for that attribute. With increased imbalance between the categories, the fracturedness may increase if other more balanced attributes are present.

We discussed the use of *weighted Voronoi diagrams* [17] to better reflect the subset size in the background encoding. The use of a weighted Voronoi diagram will conflict with local cluster patterns; more specifically, for imbalanced datasets, the area of one Voronoi cell extends below the point of its neighbors, requiring restrictions on the range weights. This behavior makes the layout enrichment hard to interpret since points are placed inside or close to an area representing a category they do not belong to. For datasets with only unique entries, the weight Voronoi diagram will be identical to the regular Voronoi diagram. To organize subsets, we also considered *Voronoi Treemaps* [24]. However, Voronoi Treemaps require a hierarchical structure, just like regular Tree Maps [280] and, thus, cannot be applied to categorical data without additional information to derive a hierarchy of attributes.

Computational Complexity: The number of data records n poses potential limitations. The time complexity of projecting data is determined by the DR methods. However, since categorical data sets are sparse, as discussed in section 3.4, the number of projected subsets is significantly lower than that of data records. The Voronoi diagram calculation and the corresponding Delaunay triangulation are both in $O(n\ log(n))$ [22]. The time complexity of calculating the fracturedness measures depends on the number of vertices and edges of the Delaunay triangulation, which will have n vertices and $3n - 3 - h$ edges, where h is the number of vertices on the convex hull. The time complexity of calculating *edge-based fracturedness* is based on enumerating all edges of the Delaunay triangulation and has a time complexity of $O(|E|)$. The time complexity of calculating *component-based fracturedness* is dependent on the algorithm for determining the number of components. We use a depth-first search-based approach with time complexity of $O(|V| + |E|)$. Thus, the dimensionality reduction method employed poses the highest contribution to the time complexity, i.e., $O(n^3)$ for MDS.

Future Work: We found that Voronoi cells can overrepresent the amount of data associated with a specific category. Thus, there is a need for a new layout enrichment method following these constraints: (1) The global area associated with one category should be relative to the occurrence in the dataset (data-ink ratio), (2) the extent of individual category areas should remain close to their projected data point positions, (3) where meaningful (e.g., among clusters), the layout enrichment should visually enclose the data points with the same category if the data-ink ratio allows. The expert study showed that the color assignment for foreground and background colors could be improved. We suggest assigning attributes to a few sets of colors based on an exploration phase. Later in the analysis, we require one color set for the attribute used in the background, one for the foreground using a distinctive palette, and one for the attribute under focus by the user. All the other attributes would be assigned a neutral color (e.g., grey). In this paper, we studied the use of MDS for categorical data analysis. However, following the approach of encoding categorical data into distances, other DR methods could be used (e.g., t-Distributed Stochastic Neighbor Embedding (t-SNE) [208] or Uniform Manifold Approximation and Projection (UMAP) [213]). These can be evaluated and compared quantitatively, following the evaluation presented in this paper. We think that the concept of *fracturedness* can be transferred to high-dimensional space when analyzing categorical data. Such a measure can be used to compare the low- and high-dimensional representations and provide a quality measure for projections of categorical data.

3.7 Conclusion

We presented a novel projection-based visualization method to address the need for similarity-based analysis techniques for categorical data. We leverage distance relations based on set intersections to create enhanced and interactive glyph-based scatterplot-like visualizations called the Categorical Data Map. We visualized attributes and categories by calculating a Voronoi partitioning and coloring the cells according to the category of the associated attribute. Our method allows for exploring the categorical data space through segmentation, enabling the orientation along an automatic or user-selected attribute. For automatic selection, we rank-order attributes along a visual property we defined as *fracturedness* measures. We quantitively evaluated different distance measures for the projection of categorical data with MDS, suggesting that the Overlap coefficient and Jaccard distance yield results outperforming MCA. Through a case study, we showed that our Categorical Data Map can support the identification of similar subsets and clusters, as well as the detection of attributes with a strong influence on the topology of the embedding. In an expert study, we were able to confirm that our approach facilitates the analysis of categorical data, especially for large datasets, by grouping similar subsets while, through layout enrichment, visualizing the distribution of categories of an attribute. We published a demonstrator and our results online so that users can interactively experiment with our approach and build upon our results. We conclude that the Categorical Data Map effectively analyzes large categorical datasets, especially in exploratory scenarios.

Part II

Data-Driven Measures

The purpose of computing is insight, not numbers.
— **Richard Hamming**, Mathematician

4 Detecting Language Change

Contents

The focus of diachronic linguistics is on understanding the development and evolution of language over time. This chapter presents HistoBankVis, a Visual Analytics (VA) approach developed for diachronic linguistics, aiming to explore and examine relationships found within complex categorical datasets generated from large text corpora. We tackle the challenge of various factors affecting linguistic change over time, necessitating rigorous annotation and exhaustive analysis. HistoBankVis, with its multilayered visual analysis system, allows for an interactive exploration of annotated Penn TreeBank datasets, providing visual overviews through easy-to-interpret histogram and matrix visualizations. Additionally, HistoBankVis enables the analysis of multi-attribute interactions between different linguistic structures that can be analyzed using Parallel Sets visualizations. Through a case study on the evolution of Icelandic, we demonstrate the system's capacity to help generate and validate new hypotheses by visualizing the interplay of multiple linguistic elements across several time periods, revealing a previously unknown link between word order, subject case, and voice.

This chapter is *based on* the following publications:

- [271] Christin Schätzle, **Frederik L. Dennig**, Michael Blumenschein, Daniel A. Keim, and Miriam Butt. "Visualizing Linguistic Change as Dimension Interactions". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Aug. 2019, pp. 272–278. DOI: 10.18653/v1/W19-4734.
- [272] Christin Schätzle, Michael Hund, **Frederik L. Dennig**, Miriam Butt, and Daniel A. Keim. "HistoBankVis: Detecting Language Change via Data Visualization". In: *Proceedings of the NoDaLiDa 2017 Workshop Processing Historical Language*. NEALT Proceedings Series 32. Association for Computational Linguistics, 2017, pp. 32–39.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

4.1 The Benefit of Visualizing Language Change

Quantitative methods have become more popular in diachronic linguistics as more data from historical texts have been digitized, for example the Bibliotheca Augustana [138], TITUS [305], and GRETIL [128]. In addition, more advanced statistical methods are used to analyze these datasets, including the calculation of co-occurrences, correlations, and other techniques [211, 23, 146]. The methods above lend themselves more for hypothesis validation and less for exploratory data analysis. Studying diachronic linguistics requires understanding the complex interactions of many linguistic and non-linguistic factors and structures. Datasets for studying language change are so complex that purely statistical methods may miss important patterns and trends. To explore how historical change can be visualized, we developed HistoBankVis, a new system that combines visual and analytical methods [178]. With HistoBankVis, a researcher can interact with the data and discover the relationships between linguistic attributes and structures. Our system eliminates the need for tedious manual work of searching for patterns in various tables of attributes and statistical significances. Instead, our system allows the researcher to select specific attributes to study and get a visual overview that shows if there are any interesting patterns across attributes over time. The researcher can then examine the relevant patterns more closely by focusing on individual linguistic structures and generating new hypotheses. These hypotheses can then be retested with a new view of the data, taking into account related attributes. Since historical data often presents a data scarcity problem, we also offer several ways to replace statistical significance, such as Euclidean distance, to deal with the small number of data points.

We demonstrate the effectiveness of HistoBankVis by applying it to a specific case study: a syntactic analysis of the Icelandic Parsed Historical Corpus (IcePaHC) [320]. The visualization helps to detect and explore syntactic changes in IcePaHC in a systematic and interactive way, allowing linguists to form and test hypotheses. Furthermore, the visualization connects the annotated values, the statistical analyses and the actual data by allowing the researcher to access the original sentences from IcePaHC during a data filtering and selection process. In summary, we contribute:

Contributions

- The *interactive HistoBankVis tool* to interactively explore language change in annotated Penn TreeBank datasets.
- An *evaluation* with domain experts, yielding uncovering insights into Icelandic concerning word order and the presence of dative subjects.
- For *accessibility*, the HistoBankVis tool and the preprocessed IcePaHC are publicly available at https://dennig.dbvis.de/histobankvis/.

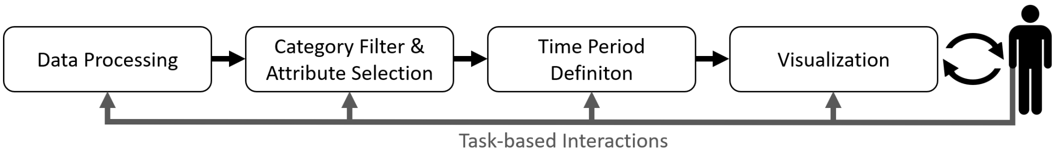


Figure 4.1: The workflow of our updated visual analysis application is as follows: Based on the specific analysis task, the user segments documents into sentences and extracts, and filters for relevant linguistic elements (called attributes), along with either customized or predefined time periods. The visualization provides different levels of detail that the user can switch between as needed. Crucially, the system includes a feedback loop that allows the user to go back and refine the filters or change the data used in the analysis.

4.2 Related Work

Visualizations and VA approaches addressing the analysis of historical linguistic data cover a large diverse set of topics. Lyding et al. [207] researched the evolution on modal verbs in historical academic discourse. Rohrdantz et al. visualized the cross-linguistic spread of new suffixes through mass media [260] and the semantic change of word meaning [259] Theron and Fontanillo [302] visually analyzed the diachronic development of different meanings of as describe in subsequent versions dictionaries. In previous work on Icelandic using IcePaHC, Butt et al. [60] as well as Schätzle and Sacha [273] used glyph visualization to analyze different linguistic factors for syntactic change. HistoBankVis addresses the challenges and shortcomings found when working with glyph visualizations. In general, we found that, the glyph visualizations had difficulties to handle the potentially large amounts of interacting attributes that are relevant to various historical linguistic research questions. The system was also based on certain assumptions regarding the nature of the data and the research questions being addressed. The objective of HistoBankVis is to offer a broadly applicable system for historical linguistic research, as well as a more adaptable approach to the study of linguistic attributes, allowing exploratory access to a wide range of factors. The system allows you to analyze individual factors separately or examine the interplay between related factors as needed.

4.3 The HistoBankVis Application

HistoBankVis is based on an iterative workflow as shown in figure 4.1. Textual data is analyzed by extracting linguistic factors that the researcher considers relevant to the research question. These factors are derived from a careful review of the theoretical literature. We call these factors attributes and their possible values categories because they describe nominal properties of linguistic structures, usually sentences. For example, voice is an attribute that has the categories active, passive, and middle.

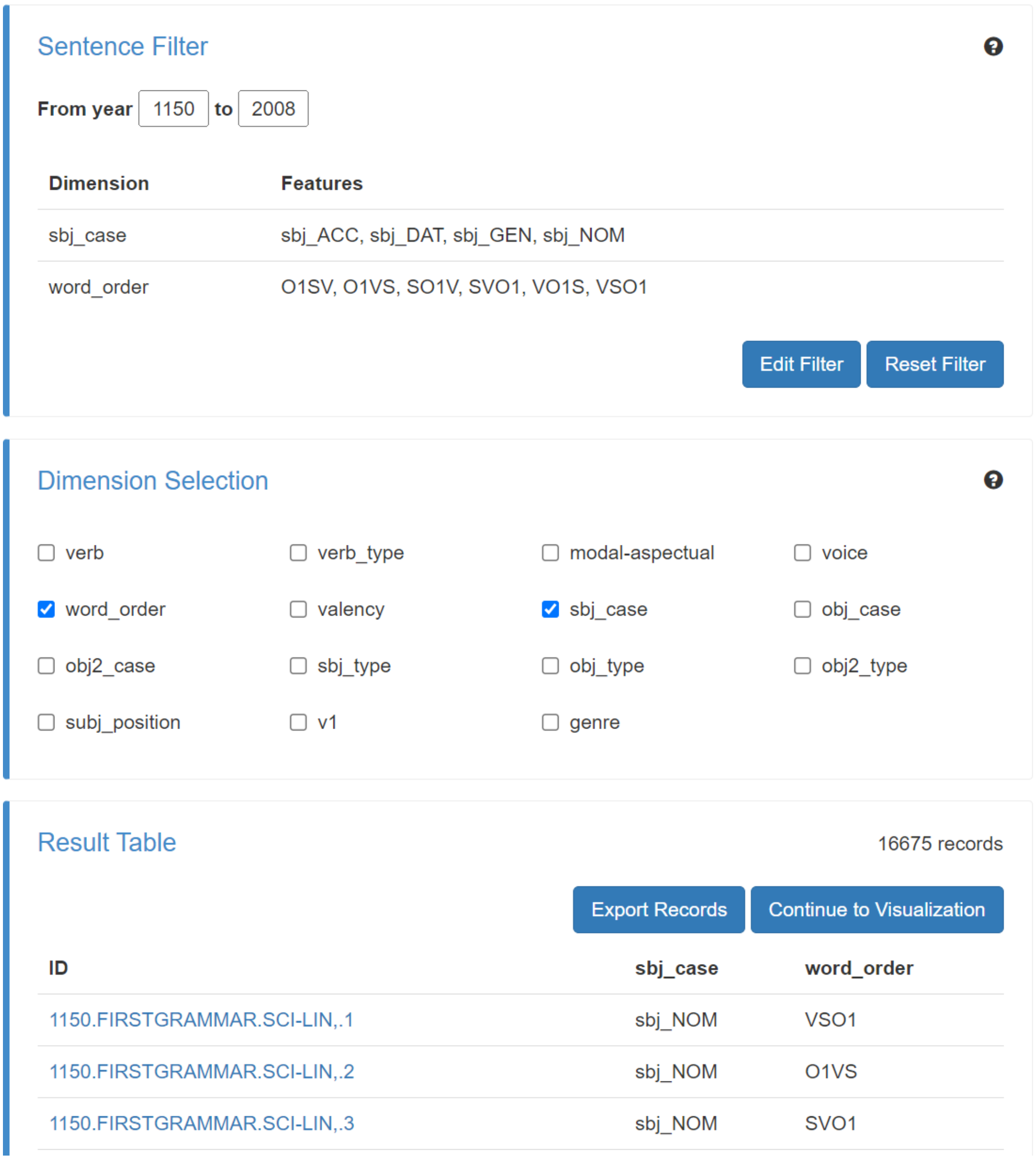
The user can filter the data according to the research question (e.g., by selecting certain attributes/categories or sentences from certain genres or time periods). The user must also define time periods to compare the historical development of the attributes. The visualization allows the user to interactively compare the distribution of the selected attributes and categories across the specified time periods. The user can also access further details of the visualization using mouse interaction techniques. The user can then use the insights from the visualization to test new hypotheses by interacting with the system. This may include changing the selected data by adjusting the filter, changing the time periods, or selecting a different set of attributes or categories to visualize.

4.3.1 Data Processing

We are working with HistoBankVis to study the relation between *subject case* and *word order* as part of a specific case study. Icelandic is generally considered to have little change in syntax and morphology of words [304, 258], but some word order changes have been reported on the shift from OV (Object-Verb) to VO (Verb-Object) [182, 257, 153] and on the decline of V1 (i.e., direct verbs) [114, 281]. Our two main linguistic questions about Icelandic based on the existing literature are: How are grammatical relations represented? Do these representations change over time in Icelandic? In linguistics, these expressions are called markers and describe the functions of words in a phase or sentence. We extracted relevant linguistic attributes from the theoretical literature and used Perl scripts to assign the corresponding categories from the IcePaHC annotation. In this chapter, we look at the historical changes in *word order*, which we represent by codes such as SVO1 (Subject-Verb-Direct Object), VSO1 (Verb-Subject-Direct Object), or VO1S (Verb-Direct Object-Subject). In the same way, we added information about *verb type*, *voice*, *case*, and *valence*. We also linked these attributes to the sentence IDs in IcePaHC. These sentence IDs give information about the year, the name of the work, and the genre in which the sentence is found. We used this information to create a well-structured database that HistoBankVis can use as part of our preprocessing.

4.3.2 Category Filter and Attribute Selection

After processing the data, the researcher can filter the records with relevant properties. In addition to filtering by time period, the researcher can interactively create filters for the categorical attributes in the database. The attributes and individual categories can be combined with logical AND or OR functions, depending on the analysis task. For example, we filtered for sentences with any OVS word order,



Sentence Filter

From year 1150 **to** 2008

Dimension	Features
sbj_case	sbj_ACC, sbj_DAT, sbj_GEN, sbj_NOM
word_order	O1SV, O1VS, SO1V, SVO1, VO1S, VSO1

Dimension Selection

☐ verb ☐ verb_type ☐ modal-aspectual ☐ voice

☑ word_order ☐ valency ☑ sbj_case ☐ obj_case

☐ obj2_case ☐ sbj_type ☐ obj_type ☐ obj2_type

☐ subj_position ☐ v1 ☐ genre

Result Table

16675 records

ID	sbj_case	word_order
1150.FIRSTGRAMMAR.SCI-LIN,.1	sbj_NOM	VSO1
1150.FIRSTGRAMMAR.SCI-LIN,.2	sbj_NOM	O1VS
1150.FIRSTGRAMMAR.SCI-LIN,.3	sbj_NOM	SVO1

Figure 4.2: The researcher can filter the sentences by selecting specific years and specific attributes and categories to create a dataset for visualization and export.

i.e. (direct) object, verb, subject, in texts from 1750 to 1900 Common Era (CE) in figure 4.2. The researcher can then select the attributes to analyze, such as *subject case*, *voice*, *word order*, and *verb* used. The researcher can analyze any sentence that matches the filter by viewing it and its Penn Treebank annotation [212], along with any extracted attributes available on demand. Therefore, the filtering component of HistoBankVis is itself a preprocessing system that allows the researcher to have a finer view of the data by selecting only some attributes and/or records. This helps the researcher become familiar with and explore the data set, as well as better understand the data quality by accessing detailed information about each data point. In addition, the researcher can download the filtered data set as a CSV file and process it with another tool.

4.3.3 Time Period Definition

The researcher must first define relevant time periods to visualize and analyze the selected attributes over time. Our system automatically supports two common time period divisions for Icelandic: (1) Old and Modern Icelandic, i.e., 1150–1550 and 1550–2008 CE [304] and (2) finer periods as per suggested by Haugen [141], i.e., 1150–1350, 1350–1550, 1550–1750, 1750–1900, and 1900–2008 CE. Additionally, the user can define custom time ranges of arbitrary size or split the available range into a variable number of evenly sized ranges, allowing for the exploration of periods independent of historically predefined ranges while also enabling the analysis of other languages. To verify the availability of data for each time period, we show a histogram (see figure 4.3) allowing for a comparison of sentence counts for each time period after applying the user-defined filter, giving researchers insight into the distribution of sentences fitting their filter criteria while also verifying that a large enough sample size is available for further analysis.

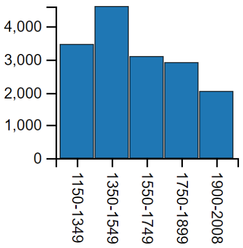
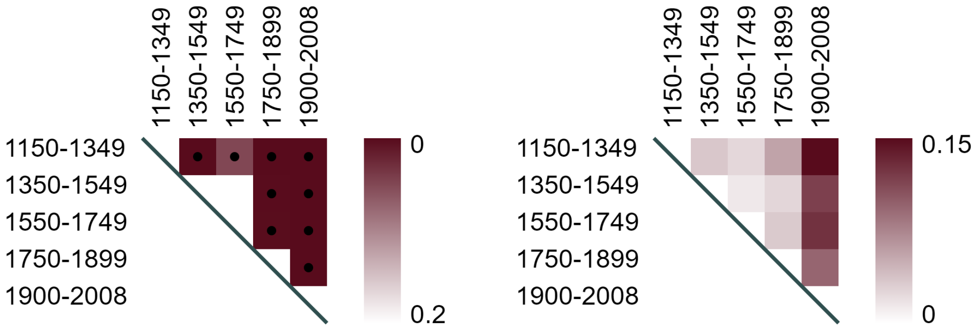


Figure 4.3: Sentence frequency histogram

4.3.4 Visualization

In HistoBankVis we offer three visualizations to explore linguistic attributes and their evolution over defined time periods. Each visualization offers a different perspective on the data, with a different level of aggregation, allowing researchers to get an overview of the current view, while also allowing them to drill down for more detail.

Compact Matrix Visualization: Our tool includes a Compact Matrix Visualization that shows how the selected attributes change with respect to the defined time periods. Each row and column of the matrix represents a period. This makes it easy



(a) $\mathcal{X}^2$-based distance matrix

(b) Euclidean distance matrix

Figure 4.4: In (a), the matrix visualization shows statistically significant differences between data distributions from different time periods. In (b), the relative changes are shown, visualizing a rapid change between 1750-1899 and 1900-2008.

to compare the first period to the rest, and each period to the one before (the entries along the diagonal of the matrix). HistoBankVis shows the difference between two periods using two methods: statistical significance and distance based. In both methods we use a color map (red means high and white means low significance/distance) to show the difference. For statistical significance, we use a χ^2 test and map the p-value to the color map: red for $p = 0$ and white for $p \geq 0.2$. In the of the χ^2 test, a dot • in the cell of the matrix shows if the difference is statistically significant (with $a = 0.05$) and a crossed-out cell (×) that there is no sufficient data to perform a χ^2-test (see figure 4.4a). If the $\mathcal{X}^2$ test is not appropriate, we can use the Euclidean distance instead. A high Euclidean distance means a large difference. This measures how much the frequency of a catagory differs, allowing for a more fine-grained comparison (see figure 4.4b). In general, the matrix view helps us to see the quality and interest of the data regarding the current filter. We can quickly find the periods that have a big difference and interesting patterns in the matrix that indicate the need for more detailed analysis.

Difference Histograms: The Compact Matrix Visualization provides a quick overview, but the difference histograms show more detail on how individual characteristics change over time. Each time period is displayed as a histogram, as in figure 4.5. Each attribute has a different color, e.g., blue for *subject case* and orange for *word order*. The height of a bar indicates the percentage of sentences with the given category. The user can also get more information, such as the sentences, the exact percentages and the relative size of the feature, by using interaction techniques. The user can compare bar heights across time periods to see which attributes and/or category combinations change over time. We also calculate the difference between two adjacent time periods and display it as another histogram below the category percentages in the histograms to provide an indicator of change. The color green

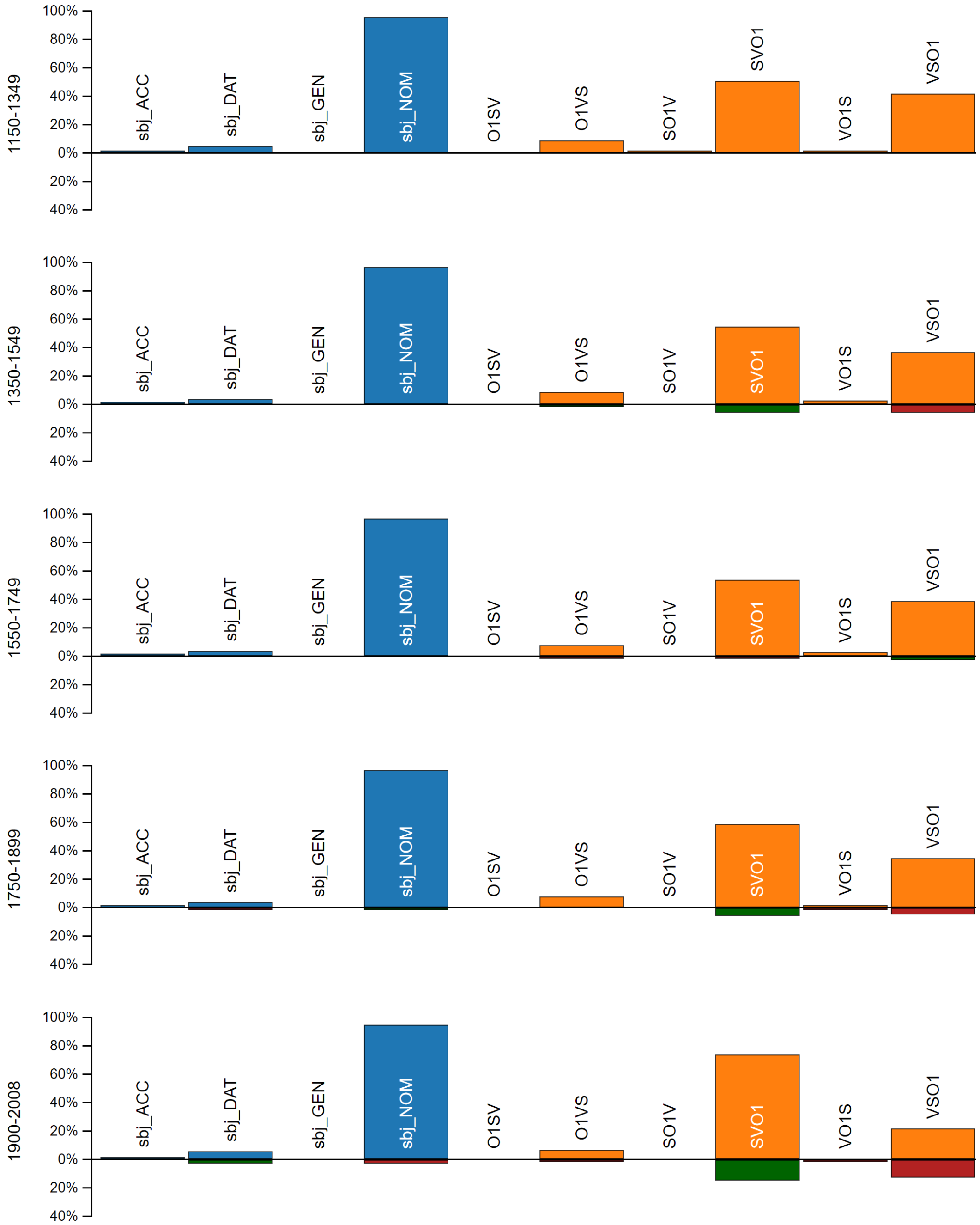


Figure 4.5: To examine the differences in more detail, users can visualize the frequency distribution of each category over different time periods, and can focus on aspects such as word order and case. Blue bars indicate the overall distribution of subject cases within the data set, which includes sentences with a subject, direct object, and verb. Orange bars indicate the different patterns of word order observed in the data. Over time, there is a consistent increase in the SVO pattern (indicated by a green bar), while the VSO pattern shows a decrease (indicated by a red bar). The overall distribution of subject case remains relatively stable until the last period, where there is a noticeable increase in the occurrence of dative subjects.

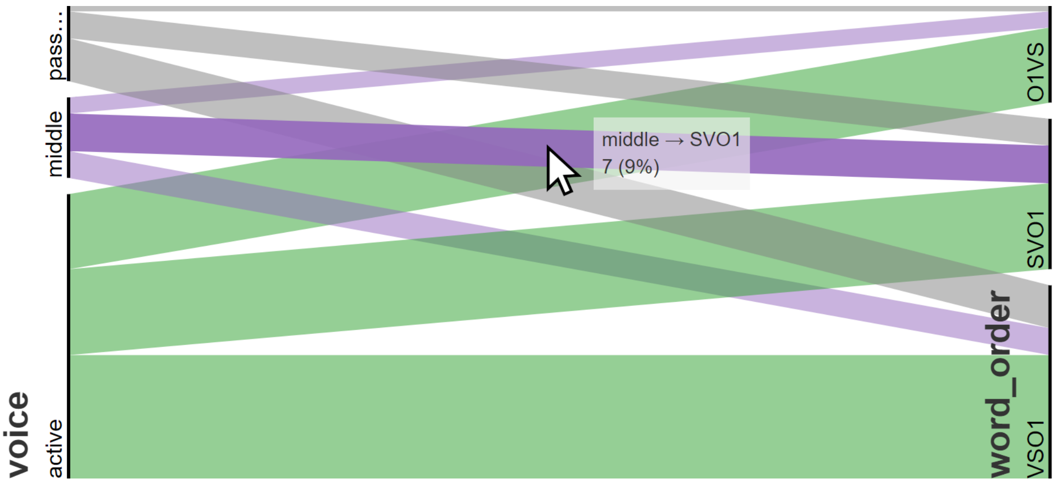


Figure 4.6: Attribute interactions in dative subject sentences from 1750–1899 for the attributes *voice* and *word order*.

means that a category has increased in frequency compared to the previous period and red means that it has decreased, e.g. SVO increased in figure 4.5 while VSO decreased. The system also has other comparison modes, such as comparing each period to the first or last period, to the average of all periods, or to the average of previous periods, to highlight different attributes and observe trends.

Attribute Interaction Visualization: We added a Attribute Interaction Visualization to the HistoBankVis system to help us understand how different categories of attributes interact. This visualization uses the Parallel Sets technique [35, 190], which is a way of displaying categorical data dimensions as frequency-based parallel lines, similar to Parallel Coordinate Plot (PCP) [156]. PCPs show how data points from a high-dimensional dataset are related by connecting points on parallel axes arranged on a 2D plane and representing each dimension as a vertical axis. In this way, each data point is represented as a polyline. This helps us see the patterns and trends in neighboring dimensions. Structured Parallel Coordinates [76] are a special kind of PCP for studying language data. They have been used to look at how words appear together and to explore the meanings of words that express possibility or necessity in old academic texts [207]. The Diachronlex diagrams by Theron and Fontanillo [302] also use PCP to show how meanings change over time, based on historical dictionaries. Parallel Sets is a visualization technique that displays the frequency of each attribute as equally spaced axes like the dimensions axes of PCPs. Through ribbons connecting the axes, they also show how categories of different attributes interact with each other, something PCPs cannot do. For example, in figure 4.6, we can see the relationship between *voice* and *word order*. The width of a ribbon indicates how much a category from one attribute corresponds to a category from another attribute. In figure 4.6 we can see that VSO1 is mostly used with the active voice, while SVO1 is mostly used with the middle voice. Our Parallel Sets implementation allows users to reorder dimensions by dragging and dropping, and to sort categories by size or alphabetically. Users can also get more detail about a category interaction using mouse-over techniques, as shown in figure 4.6. Parallel Sets have not yet been used for linguistic research, but we will show that our Attribute Interaction Visualization is a very useful and powerful tool for historical linguistic analysis, helping us to discover and understand how different categories in a dataset with many attributes influence each other.

4.3.5 Feedback and Hypothesis Generation

The researcher can use the knowledge gained from exploring the data and testing hypotheses to modify any part of the previous configuration system. The researcher can change the filters, try different time periods, or go back to the data processing step and include different or more categories. This creates an iterative analysis process that combines knowledge-based and data-driven hypothesis testing. In addition, each filter and visualization configuration is identified by a unique URL, making it easy to share results in a collaborative environment.

4.4 Evaluation

In the visualization community, we use case studies to evaluate how useful a visualization is for finding significant and novel insights about the data [63, 157]. This section presents two case studies of how HistoBankVis can be used to explore syntactic change in Icelandic, focusing on how subject case and word order interact [257, 26]. Previous studies looking at changes in these phenomena do not consider how they affect each other. By visualizing the data, we discovered multiple phenomena are strongly related. We analyze a real-world dataset of Icelandic in the annotated IcePaHC format with texts from 1550 to 2008 covering different types of text genres.

4.4.1 Case Study: Exploring Correlations Between Word Order and Dative Subjects

We used the visualizations above to study how word order and dative subjects are related. First, we looked at the word order of all subjects in Old and Modern Icelandic by selecting sentences with a subject (S), a verb (V) and a direct object (O/O1). We then looked at the subject case and word order dimensions. The Difference Histograms show that SVO is the most common order for both time

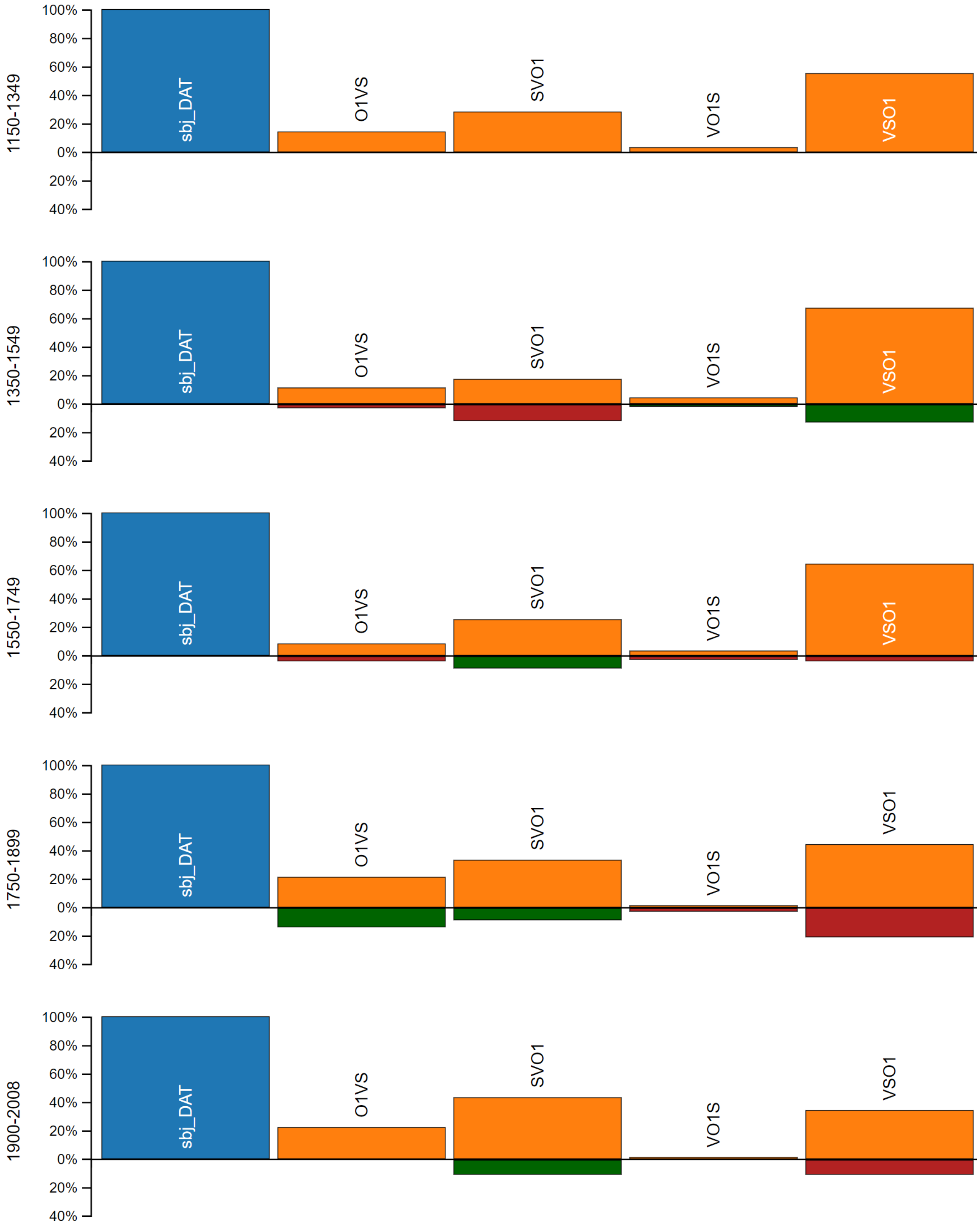


Figure 4.7: We show word order patterns for dative subjects. Initially, VSO was the predominant word order until the final time stage, at which point SVO emerged as the dominant order after consistently increasing throughout the corpus. Additionally, the OVS word order is notably prominent in the penultimate time period.

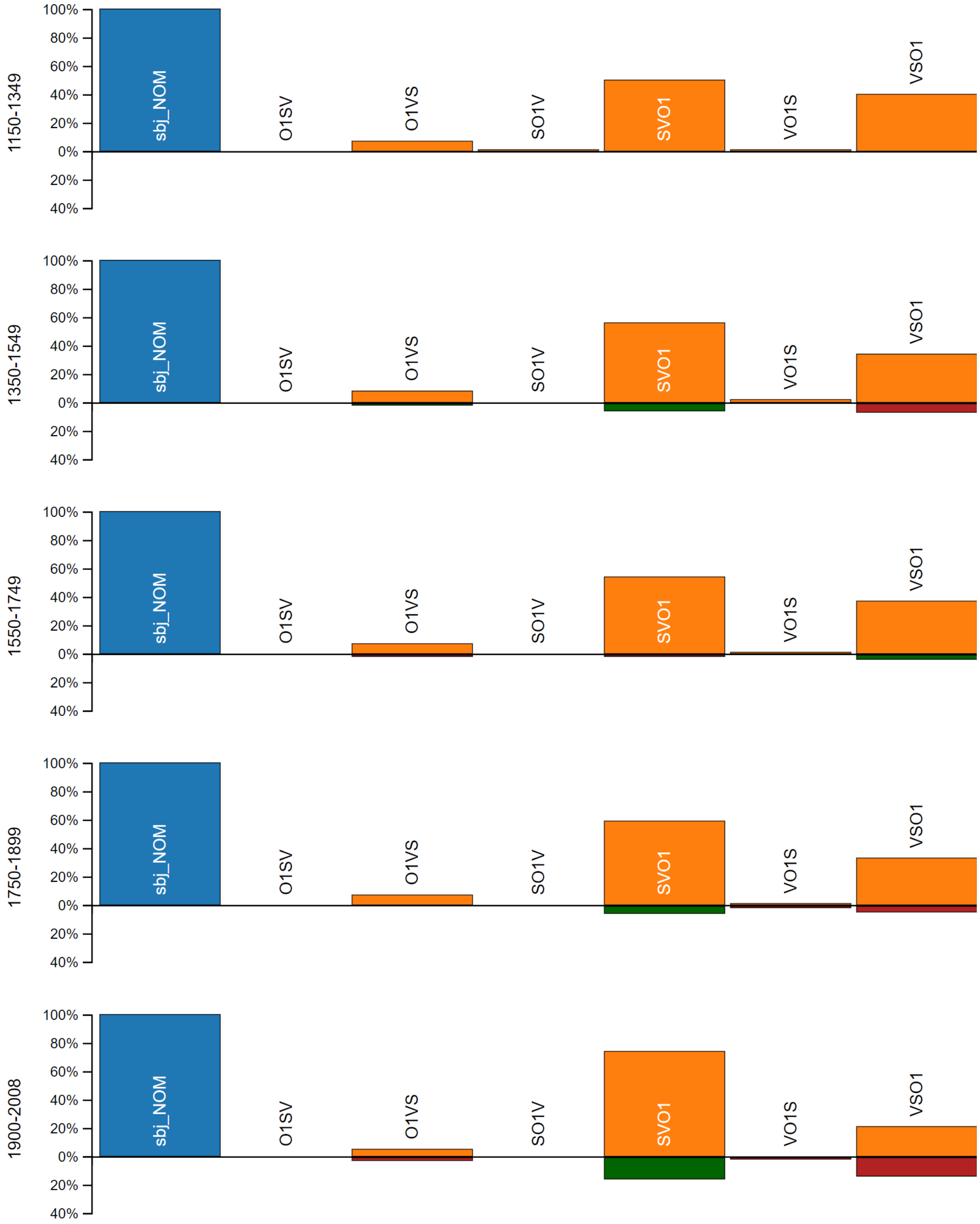


Figure 4.8: We show the word order trends for nominative subjects. The evolution of word order patterns mirrors those observed across all subjects, with VSO decreasing and SVO increasing over the various time periods.

periods, and that it is increasing over time, while VSO is decreasing (see figure 4.5). They also show that most subjects are nominative and some are dative. After this general overview, the Compact Matrix Visualization (see figure 4.4) showed us that there is a big change in the last two time periods. Comparing each range with the previous one, we saw a large increase in SVO in the last time stage (see the green bar under SVO1 in figure 4.7) and a decrease in VSO, as shown by the red bar under VSO1. Dative subjects also increased slightly in the last range (see figure 4.5). Based on these results, we decided to analyze word order separately for dative and nominative subjects. We did this by changing the filter settings to include only dative or nominative subjects. The word order histograms for nominative subjects (see Figure figure 4.8) matched the overall word order trends for all subjects, but dative subjects were different. The Difference Histograms in figure 4.7 show that VSO was the most common word order for dative subject clauses until about 1900, when SVO became more common than VSO.

We found that O1VS was very different in the penultimate period compared to the other periods (see figure 4.8 and figure 4.8). Therefore, we filtered the data again for O1VS only and noticed that the verbs in this period were mostly experiencer verbs, such as líka (to like, to please) as in “I like sunny days”. We think that these experiencer verbs changed over time from a structure where the experiencer/goal was an object to a structure where the experiencer/goal was a subject. For example, “that pleases me,” where the experiencer is an object, is changed to “I like that,” where the experiencer is a subject. This is a common phenomenon in many languages [129], and there are also linguistic principles that explain why experiencers/goal participants tend to be subjects [96]. We believe that the Icelandic pattern is an example of a historical change in which experiencers became more often dative subjects. Our findings are also consistent with recent research by Schätzle et al. [270] on how middle morphology and dative subjects interact.

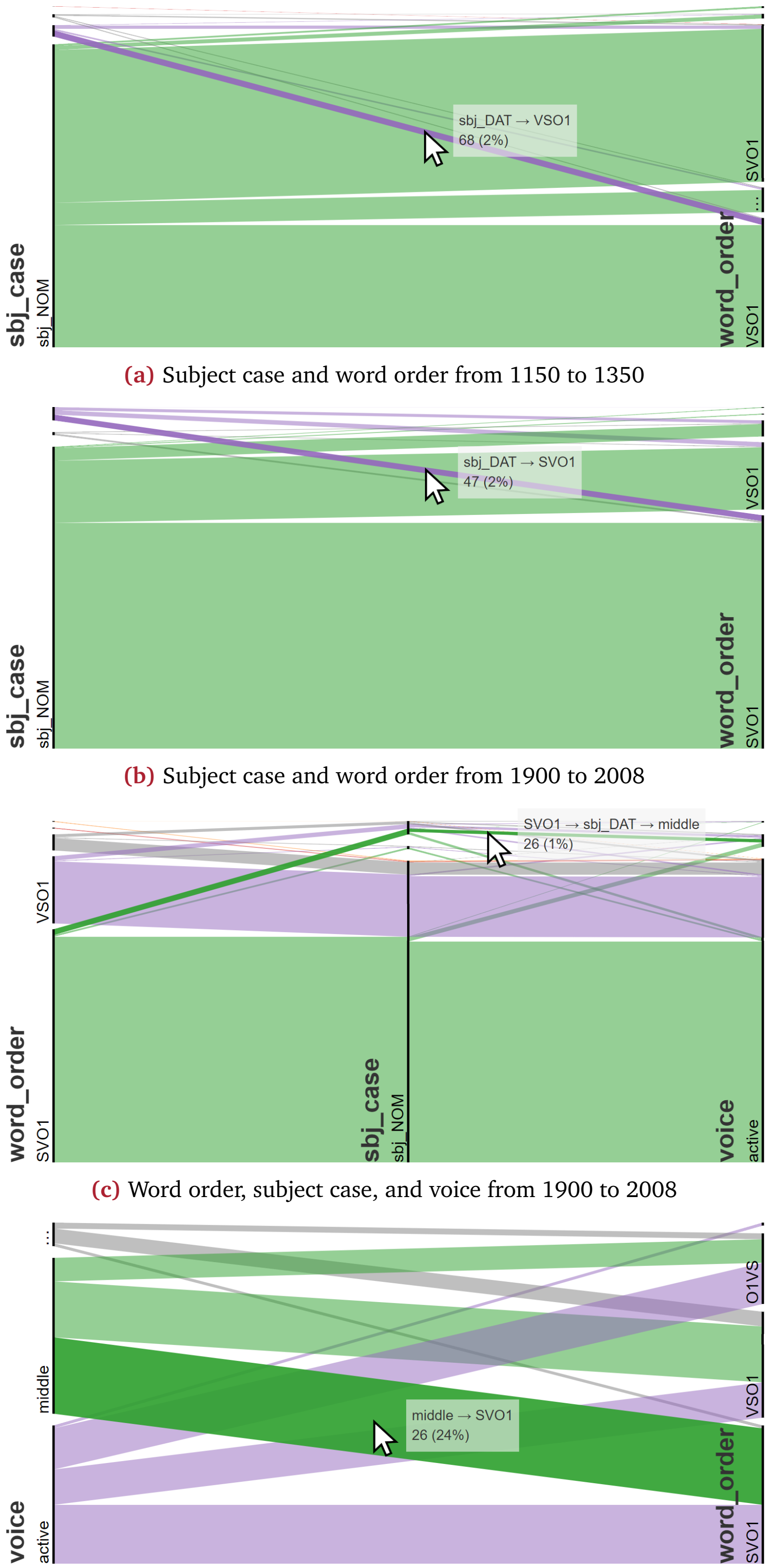
Remember that we also found a general shift towards SVO word order. We think that this means that Icelandic developed a fixed position for subjects before the verb in its history, and that the 19th century was a key moment for this change. Dative subjects followed this change more slowly. We explain this slower change by the fact that experiencer/goal arguments were not typical subjects and many of them changed from object to subject first. Other changes in Icelandic word order occurred around the same time, such as the decline of V1 [281, 60] and the loss of OV [153].

4.4.2 Case Study: Analyzing Subject Case, Word Order and Voice

We used the Difference Histograms to examine how word order and subject case changed over time in transitive sentences, i.e., sentences with a subject (S), a finite verb (V), and a direct object (O1). The Compact Matrix Visualization showed that the distribution of word order and subject case changed a lot after 1900, as seen in figure 4.4. Figure 4.5 shows the difference histogram distributions for subject case and word order in the periods before and after 1900. The most noticeable change in word order is that SVO1 became more common between 1900 and 2008 (green bar), while VSO1 became less common (red bar). At the same time, dative subjects increased slightly. We hypothesized that these two changes were related.

The Attribute Interaction Visualization shows the correlations between the characteristics of any selected attribute to explore possible interactions. Figure 4.9b shows the interaction between subject case and word order in the period 1900–2008. The attributes are sorted by the size of their category, with the largest category at the bottom. The subject case proportions on the left are mapped to the word order proportions on the right. The interaction shows that SVO1 is the most common word order overall. Most nominative subjects go with SVO1, while the share of SVO1 for dative subjects is much smaller. The interactions in the period from 1900 to 2008 are different from the ones in an earlier period (1150–1350), as seen in the top left of figure 4.9a. In contrast to the period after 1900, the proportions of SVO1 and VSO1 are similar for nominative subjects. Dative subjects are also more likely to use VSO1. This means that word order changes over time depending on the subject case. The Difference Histograms in figure 4.5 show that subjects are more often placed before the verb, but the attribute interaction shows that dative subjects are behind this change.

Voice is known to affect the frequency of dative subjects in Icelandic [335, 282]. But the relationship between voice, subject case, and word order has not been studied. HistoBankVis makes it possible to study this relationship by adding the attribute of voice to the analysis of the interactions between subject case, word order, and voice. Figure 4.9c shows the interactions for the period 1900-2008 and shows that nominative subjects mostly go with SVO1 in active constructions. But dative subjects mostly go with SVO1 in middle constructions. We can also look at the interaction between voice and word order for dative subjects separately, as in figure 4.9d for 1900–2008. Dative subjects are more common with middle voice, and SVO1 is the most common word order for both active and middle constructions. However, in earlier stages of the language, voice and word order had different patterns, as shown in figure 4.6 for the interaction from 1750 to 1899. First, dative



(a) Subject case and word order from 1150 to 1350

(b) Subject case and word order from 1900 to 2008

(c) Word order, subject case, and voice from 1900 to 2008

(d) Voice and word order with dative subjects from 1900 to 2008

Figure 4.9: The top figure shows the interplay between case and word order from 1150 to 1350. The second shows subject case and word order from 1900 to 2008. The third figure shows the interactions between word order, subject case, and voice from 1900 to 2008, and the last figure shows the relationship between voice and word order in sentences with dative subjects on the right.

subjects were more common in active clauses than in middle clauses. Also, SVO1 was already the dominant word order for middles, but not for active constructions where VSO1 was more common.

In summary, these results show that dative subjects are more often placed before the verb because they are more often used with middle voice. HistoBankVis helped us to easily discover this new relationship between word order, dative subjects and voice in a few minutes using the dimensional interactions.

4.5 Discussion and Future Work

HistoBankVis streamlines the process of identifying and analyzing complex patterns in historical linguistic data. The tool allows researchers to interactively explore data and quickly identify interesting linguistic patterns across time periods. Through a case study using IcePaHC, the tool demonstrated its ability to help generate findings for changes in word order and subject case over time. We identified specific shifts from older word order forms to more modern configurations, in particular tracing the evolution from object-verb (OV) to verb-object (VO) structures in Icelandic. Since these changes in Icelandic have occurred relatively recently, we can hypothesize that the relatively isolated Icelandic language is experiencing changes due to globalization. The tool supports dynamic hypothesis testing and generation. Researchers can drill down from broad patterns of change to specific category interactions to test new hypotheses and efficiently iterate through the analysis process. By introducing Parallel Sets, we have enhanced the tool's ability to explore and analyze interactions between different linguistic factors in more depth, such as the relationship between word order, case, and grammatical relations in Icelandic. Using the Dimension Interaction Visualization, researchers were able to observe and document specific trends and shifts in linguistic structures over time, providing a more nuanced understanding of how certain linguistic features evolve together. Researchers were able to observe and document specific trends and shifts in linguistic structures over time, providing a more nuanced understanding of how certain linguistic features evolve together. We conclude that HistoBankVis is an effective tool for studying lexical-semantic change and, in principle, is applicable to any type of language change, facilitating the identification of complex feature interactions over time.

Future Work: In computer science terms, HistoBankVis is a VA approach for analyzing large categorical datasets with a segmentable temporal component, and thus could be generalized and applied to other domains. In the field of linguistics, further evaluations, including other languages, could lead to new linguistic insights and further demonstrate the power of HistoBankVis. Currently focused on Icelandic, the system could be adapted to other languages and linguistic corpora, providing a broader tool for comparative historical linguistics, since HistoBankVis relies only on a categorical description of linguistic structures. For the study of language structure in terms of semantic change, graph-based approaches offer a promising direction. The integration of graph visualizations could broaden the scope of HistoBankVis from lexical change to semantic drift.

4.6 Conclusion

We presented HistoBankVis, a novel visualization tool for linguistic research that allows users to explore and analyze language change in annotated corpora. HistoBankVis provides multiple perspectives of the data at different levels of detail on demand, enabling an iterative process of hypothesis testing and generation. One of the key features of HistoBankVis is the use of parallel sets, a visualization technique that shows complex interactions across different dimensions of data. This is the first application of Parallel Sets in linguistic visualization, and we demonstrate its usefulness and flexibility on a case study of changing linguistic categories in Icelandic. HistoBankVis can also be used as a preprocessing and filtering tool, allowing users to export filtered data sets according to their specifications. In addition, HistoBankVis supports collaborative research by allowing users to share their analyses and perspectives on the data through unique identification URLs. HistoBankVis is a powerful and effective tool that can be applied to any Penn Treebank-style annotated corpus or well-structured dataset, facilitating the study of historical language change.

5 Analyzing Software Vulnerabilities

Contents

The prevalent usage of Open Source Software (OSS) has led to an increased interest in resolving potential third-party security risks by fixing Common Vulnerabilities and Exposures (CVEs). However, even with automated code analysis tools in place, security analysts often lack the means to obtain an overview of vulnerable OSS reuse in large software organizations. In this design study, we propose VulnEx (Vulnerability Explorer), a tool to audit entire software development organizations leveraging the ordinal classification of CVEs into different risk classes. We introduce three complementary table-based representations to identify and assess vulnerability exposures due to OSS, which we designed in collaboration with security analysts. The presented tool allows examining problematic projects and applications (repositories), third-party libraries, and vulnerabilities across a software organization. We show the applicability of our tool through a use case and expert feedback.

This chapter is *taken from* the following publication:

- [82] **Frederik L. Dennig**, Eren Cakmak, Henrik Plate, and Daniel A. Keim. “VulnEx: Exploring Open-Source Software Vulnerabilities in Large Development Organizations to Understand Risk Exposure”. In: *Proceedings of the IEEE Symposium on Visualization for Cyber Security*. IEEE, 2021, pp. 79–83. DOI: 10.1109/VizSec53666.2021.00014.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

5.1 The Impact of Vulnerabilities in Open Source Software

The extensive usage of OSS nowadays promotes a straightforward integration of common software features into existing applications [245, 287]. However, software reuse also poses a significant risk as software with disclosed vulnerabilities is often extensively reused, affecting various applications across whole organizations [59]. For instance, the Equifax data breach in 2017 resulted from a missed OSS package update and led to the disclosure of the private data of over 145 million U.S. citizens [224]. Hence, an organization's governance or audit system must identify the organization's overall exposure to OSS vulnerabilities.

Developers and security analysts regularly utilize automated code analysis tools to identify vulnerabilities and investigate the mitigation of OSS security risks. For example, static [286] and dynamic code analysis [247, 251] are applied to execute the developed code and detect inherent vulnerabilities. However, such code analysis tools heavily differ in their detection capabilities. They often only store the vulnerability metadata as text files that do not meet software developers' basic requirements, such as prioritizing the most severe vulnerabilities. Assessing the impact of software vulnerabilities is essential for organizations since the effects of exposures can vary significantly. Code analysis tools are usually used for single software applications and do not show the impact of OSS vulnerabilities across multiple applications in whole software organizations. Additionally, it is crucial to evaluate the quality of libraries and other dependencies if they originate from another source, such as that the source can be trusted [284], and that OSS developers are swift in addressing vulnerabilities [7]. The mentioned points are crucial for deciding whether a software development organization should use an OSS library.

We propose VulnEx (Vulnerability Explorer), a new tailored design to explore and assess the mitigation of OSS vulnerabilities for auditing and governance of whole software development organizations looking beyond individual applications and teams. In our user-centered design study, we designed three complementary table-based representations to identify and assess vulnerabilities across various applications. We demonstrate the applicability of our approach through use cases and initial expert feedback. VulnEx is open-source and accessible online. We present a first study to improve the analysis and mitigation of software vulnerabilities, especially from an organization-wide perspective. In summary, the primary contributions of this chapter are:

Contributions

- A *design study* with problem characterization, findings, and lessons learned for the visual analysis of OSS vulnerabilities.
- The *interactive VulnEx analysis tool* to interactively explore critical vulnerabilities and their impact on dependent projects.
- For *accessibility*, a demonstrator of the VulnEx tool is available online (https://dennig.dbvis.de/vulnex) and VulnEx is open-source (https://github.com/dbvis-ukon/vulnex).

With this work, we hope to improve the analysis and mitigation of software vulnerabilities by addressing the need for an analysis tool for auditing entire software development organizations.

5.2 Related Work

Software visualizations provide a comprehensive overview of complex systems, such as program structures, execution behavior, and the development process [91]. These visualizations are also useful to investigate security aspects, e.g., SecSTAR [107] automatically generates execution diagrams to examine, debug, and test software applications. For an overview of software visualization research, refer to the reviews of Wagner et al. [319], and Chotisarn et al. [72].

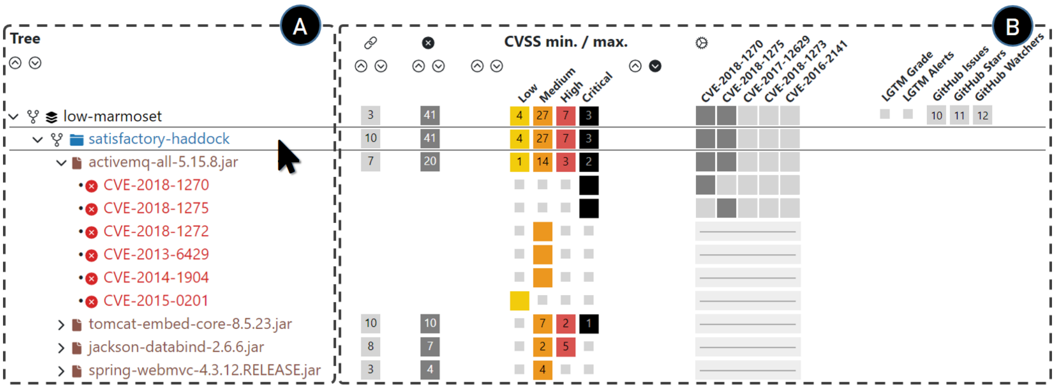
In software security visualization, some approaches for vulnerability exploration have been proposed. Harrison et al. [136] proposed the Nessus vulnerability visualization (NV) to discover and analyze network vulnerabilities of Nessus scans. The system simplifies and displays vulnerability assessment results to support security analysts, using zoomable treemap visualization with linked histograms. In a similar context, Angelini et al. [12] proposed Vulnus, which aims to increase situational awareness of security managers by visually analyzing vulnerability spreads in computer networks. Furthermore, CVExplorer [244] is a visual analytics system for analyzing vulnerability reports and enhancing network security using three linked views. These vulnerability systems differ from our approach since they primarily focus on exposing computer network vulnerabilities. Moreover, Goodall et al. [122] proposed a system to explore vulnerabilities and code weaknesses in software development. The goal is to help users understand their code's security status by displaying code vulnerabilities using an aggregated block metaphor for each file. Goodall et al. [122] approach focuses on identifying false positives, which we reduce in our application by checking whether third-party vulnerabilities are reachable. Assal et al. [18] presented Cesar, a collaborative code analysis system to reduce vulnerabilities and improve code security. The authors utilize a treemap visualization to help security experts and developers collaboratively explore static-code analysis methods' results. The treemap visualization displays a software package, and each leaf node shows a class file. Angelini et al. [11] presented a visual analytics approach to assist users in exploring program execution, describing in a use case of the detection of single vulnerabilities. However, the system is mainly targeted to investigate symbolic execution engine data. Recently, Alperin et al. [8] presented a study for the interpretable visual assessment of vulnerabilities. In their study, the authors focus on local explanations for predictive vulnerability analysis.

In summary, the listed approaches focus on exploring network vulnerabilities and improving the code security of individual software packages, such as investigating potential false positives. In contrast, we propose an initial approach that provides an overview of entire software development organizations. Our design study focuses mainly on the visual analysis of OSS vulnerabilities by supporting auditing teams in assessing OSS dependencies through table-based views to evaluate vulnerabilities in large software organizations.

5.3 Table-Based Vulnerability Exploration

The main goal of this work is to design visualizations to explore security risks in large software organizations. We gathered knowledge about the domain and user requirements in three interviews with two security analysts and a software developer from SAP. The interviews provided valuable insight into the daily workflows and challenges faced by security analysts regarding vulnerability assessment.

Application Background: The essential user task is to understand the overall risk exposure of large development organizations, e.g., commercial software vendors or open-source foundations, due to the consumption of open-source components in a considerable number of development projects or applications. During software development, projects are regularly scanned with code analysis tools. At SAP, the developers regularly utilize *Eclipse Steady* (https://github.com/eclipse/steady) [251], which supports static and dynamic analysis to detect and assess vulnerabilities. *Eclipse Steady* scans projects for CVE, which have a unique identifier in the National Vulnerability Database. *Eclipse Steady* displays the Common Vulnerability Scoring System (CVSS) score to indicate the severity of identified security vulnerabilities. However, the CVSS score only captures the vulnerability severity. Organizations require complementary information from other sources to evaluate the general software quality of the most-used libraries and determine whether they have sufficient quality. The identification of low-quality libraries is a prerequisite for follow-up decisions. However, the visual exploration of applications, consumed



Tree	🔗	⊗	Low	Medium	High	Critical	LGTM Grade	LGTM Alerts	GitHub Issues	GitHub Stars	GitHub Watchers
low-marmoset	3	41	4	27	7	3			10	11	12
satisfactory-haddock	10	41	4	27	7	3					
activemq-all-5.15.8.jar	7	20	1	14	3	2					
CVE-2018-1270											
CVE-2018-1275											
CVE-2018-1272											
CVE-2013-6429											
CVE-2014-1904											
CVE-2015-0201											
tomcat-embed-core-8.5.23.jar	10	10		7	2	1					
jackson-databind-2.6.6.jar	8	7		2	5						
spring-webmvc-4.3.12.RELEASE.jar	3	4		4							

Figure 5.1: VulnEx is a tool for the investigation of exposure to open-source software vulnerabilities on an organization-wide level. The tool shows repositories, modules, libraries, vulnerabilities in a tree representation (A), and meta-information about each entry (B), such as the CVSS score. We can see that the "low-marmoset" repository is exposed to severe vulnerabilities, three critical and seven high. Two of the critical vulnerabilities are originating from the *activemq-all* indicating that the library should be updated swiftly.

libraries, and related vulnerabilities on an organizational level are not supported by any of the tools available to date. From the interviews and further discussions with domain experts, we derived the following requirements for our tool aimed at the organization-wide analysis of software vulnerabilities.

Requirements

(R1) The tool should provide views to detect vulnerable *repositories* and *projects* to apply countermeasures, such as training weaker teams and reallocating resources. For this, repositories need to be represented in a comparable way to estimate relevance and understand how they compare against each other.

(R2) Software projects potentially depend on vulnerable *libraries*, which have to be updated. Thus, the tool needs to convey the overall exposure and allow for the inspection of specific bugs.

(R3) Vulnerabilities need to be explored to address specific exploited known *vulnerabilities*, e.g., OSS vulnerabilities prominently discussed in mainstream media, where organizations may be required or expected to make a statement whether and which of their applications are affected. Thus, the tool needs to enable users to find specific bugs with a high-security risk.

(R4) Vulnerabilities can have different effects depending on the *severity* and how many projects the origin is, and thus need to be prioritized accordingly. Therefore, the tool needs to show the impact of specific bugs on the organization's codebase.

In a two-year process, we applied the guidelines by Chen et al. [69] to perform our design study. Our tool covers the pipeline from scanning to repairing or mitigating a vulnerability. The overall workflow of VulnEx follows the Knowledge Discovery in Databases (KDD) pipeline [108]. The workflow of VulnEx: (1) The security analyst starts a scan of the source code of all software projects. (2) He then selects a type of analysis target, i.e., repositories, libraries, or vulnerabilities, choosing between overviews. (3) Then, the analyst defines criteria he is interested in, i.e., the number or severity of a bug allowing for filtering. (4) The analyst observes the findings and determines their relevance by drilling down to the specific issue. (5) In case of a relevant finding, the analyst can start a repair or mitigation process; this is supported by the detailed report of *Eclipse Steady*. Thus, we follow Shneidermans' mantra: *Overview first, zoom and filter, then details-on-demand* [279]. In figure 5.1 we show the dependency tree (A), allowing users to explore the hierarchy of the software project and the vulnerability information (B) to give insight into the exposure to vulnerabilities.

5.3.1 Dependency Tree

The dependency tree representation in figure 5.1 (A) shows the relationship of all repositories ☰, modules 📁, libraries 📄, and bugs ⓧ. VulnEx is inspired by the tree+table approach by Nobre et al. [232, 233]. We choose this because tree structures are common and known by domain experts and allow us to leverage the hierarchy inherent in software projects while supplying additional information about vulnerabilities, keeping a high level of detail. The tree representation allows for the analysis of vulnerabilities in three ways.

Repository-Centered: ☰ → 📁 → 📄 → ⓧ
This order of levels allows for a repository-focused analysis. Starting with a repository, then showing information about modules and sub-modules, enabling analysts to locate severe vulnerabilities. If a module uses a vulnerable library, this can be quickly detected. Finally, we show the vulnerabilities caused by a library, allowing for detailed analysis and estimation of the impact.

Library-Centered: 📄 → ⓧ → ☰ → 📁
Beginning with a library, displaying its vulnerabilities allows analysts to estimate the risk associated with a library. If a repository uses a vulnerable library, this repository is shown on the next level. Finally, we present the associated module or sub-module exposed to the CVE of that library, allowing for inspecting it in detail.

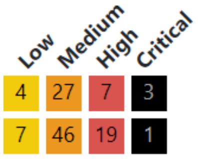
Bug-Centered: ⓧ → 📄 → ☰ → 📁
Starting with a CVE, then showing the affected library allows analysts to find specific bugs quickly. If a repository uses a vulnerable library, this repository is shown on the

next level. Finally, we display the associated module or sub-module impacted by the CVE of that library, allowing for detailed analysis.

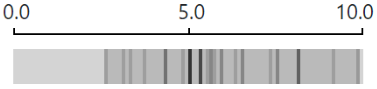
5.3.2 Vulnerability Information

We provide additional information about the vulnerabilities of a repository, module, and libraries, shown in figure 5.1 (B), through which we support the detection and analysis of critical vulnerabilities, as well as the assessment of the quality of OSS dependencies, e.g., Looks Good To Me (LGTM) grade and score. The 🔗 column shows the number of entities on the next level of the tree, indicating the number of related entities on the following hierarchy level. The ⊗ column shows the number of vulnerabilities a repository, module, or library is exposed to. The absence of an element indicates that the information is not available or applicable to the entity of the row.

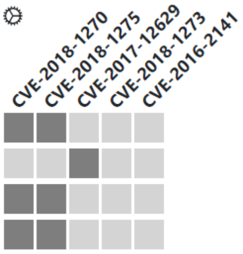
CVSS Score: The CVSS score column shows the number of CVEs with a given score. We use the common ordinal classification: Low (0.1 - 3.9), Medium (4.0 - 6.9), High (7.0 - 8.9), Critical (9.0 - 10.0), which is also mirrored in the coloring from the *National Security Database* (https://nvd.nist.gov/). The number in each square encodes the number of occurrences in the given range.



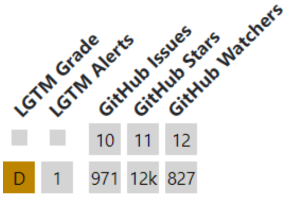
To inspect the distribution of CVSS scores in a more finely-grained way, we offer a representing of each CVE and its CVSS score with its precise numerical value. It also indicates the range of CVSS scores.



CVE Matrix: The CVE matrix indicates the presence of a specific vulnerability. Each column shows the presence of a CVE with dark gray squares, while a light gray square indicates the absence of the CVE. We adopted this encoding from Nobre et al. [233]. Columns can be added and removed to highlight specific CVEs dependent on the user. The CVE matrix allows users to get an overview of the presence of specific vulnerabilities in repositories, modules, and libraries. It also enables the analysis of the co-occurrence of CVEs.

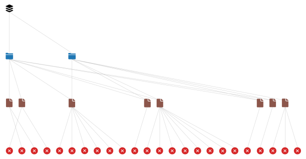


Meta-Information: We provide additional information from LGTM (https://lgtm.com/), a code analysis platform, and GitHub (https://github.com/). The columns describe the LGTM grade, LGTM score, GitHub issues, GitHub stars, GitHub watchers. The LGTM grade and score provide an additional measure for the quality of software artifacts. The number of GitHub issues



provides an indicator for active development, while the GitHub stars and watchers provide an indicator for the popularity of a repository.

Dependency Graph: The user can view the structure of a software project by clicking ⑂. The repository is shown at the top, its modules and libraries in the middle, and the bugs at the bottom.

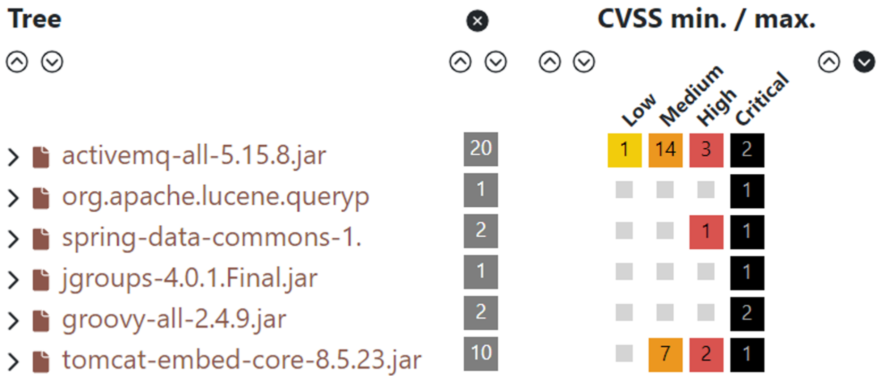


5.3.3 Filter and View Options

Based on expert feedback, we offer filter options to reduce the number of entries in the table and allow for a focused analysis. The user can search for a name of a given repository, library, or bug. We enable users to filter by the minimum and the maximum number of dependencies, vulnerabilities, and the CVSS score. Users can hide all repositories and modules that do not contain any vulnerability, as well as CVEs without a CVSS score.

5.4 Evaluation

We analyze all public GitHub repositories of the *Eclipse Foundation* (https://github.com/eclipse) that are *Apache Maven* (https://maven.apache.org/) projects in the *Java* programming language. All projects are scanned using the *Eclipse Steady* tool. We scanned these repositories from January 21 to February 4, 2020. This yields a total of 295 projects that we analyze for common libraries and vulnerabilities. We replace the original repository and module names with pseudonyms not to blame the individual projects. At SAP, the analysis of an individual application follows a defined process, starting from the automated scanning with tools like *Eclipse Steady* to discover vulnerable open-source dependencies, the assessment of findings by a security expert, and finally, depending on the assessment result, the remediation of the vulnerability or the dismissal of the finding. However, open-source software analysis across multiple applications for an entire organization does not follow a defined process. To show the usefulness of VulnEx we analyze the gathered data to answer the following four questions. Questions (Q1-Q3) are examples for exploratory analysis, while (Q4) addresses a need when vulnerabilities in open-source components get a lot of public attention, even in mainstream media. In such cases, commercial vendors like SAP are expected to state to what extent and which of their applications are affected by a given vulnerability. Thus, we include (Q4) as a search task.



Tree		CVSS min. / max.			
		Low	Medium	High	Critical
activemq-all-5.15.8.jar	20	1	14	3	2
org.apache.lucene.queryp	1				1
spring-data-commons-1.	2			1	1
jgroups-4.0.1.Final.jar	1				1
groovy-all-2.4.9.jar	2				2
tomcat-embed-core-8.5.23.jar	10		7	2	1

Figure 5.2: The analyst detects the most vulnerable libraries. *activemq-all* contains one low, 14 medium, three high, and two critical severity vulnerabilities, affecting 20 repositories.

(Q1) Which repositories contain most severe vulnerabilities?

Security analysts utilize the *Repository Table* to analyze all repositories, depicted in figure 5.1. They find the “low-marmose” repository, which has three critical bugs. We can see that all critical vulnerabilities are in the “satisfactory-haddock” module by expanding the entry. They inspect the module and see that the *tomcat-embed-core* library contains *CVE-2018-8014* and *activemq-all* contains *CVE-2018-1270* and *CVE-2018-1270*. They find that all three CVEs are critical, which should be addressed promptly.

(Q2) Which dependencies contain severe vulnerabilities and are often used across different applications?

The security analysts use the *Library Table* (see figure 5.2). They sort the table by the most severe vulnerability. The libraries *activemq-all*, *org.apache.lucene.queryparser*, *spring-data-commons*, *jgroups*, *groovy-all*, and *tomcat-embed-core* all contain critical bugs.

(Q3) Which severe vulnerabilities are present?

Using the *Bug Table*, the analysts find that eight critical bugs (see figure 5.3) are present, one in *activemq-all* affecting 20 repositories, one in *org.apache.lucene.queryparser* affecting 14 repositories, one in *spring-data-commons* affecting seven repositories, one in *jgroups* affecting five repositories, two in *groovy-all* affecting seven repositories, and one in *tomcat-embed-core* affecting eight repositories. They remark that these six libraries should be updated and fixed or replaced swiftly since they contain critical vulnerabilities.

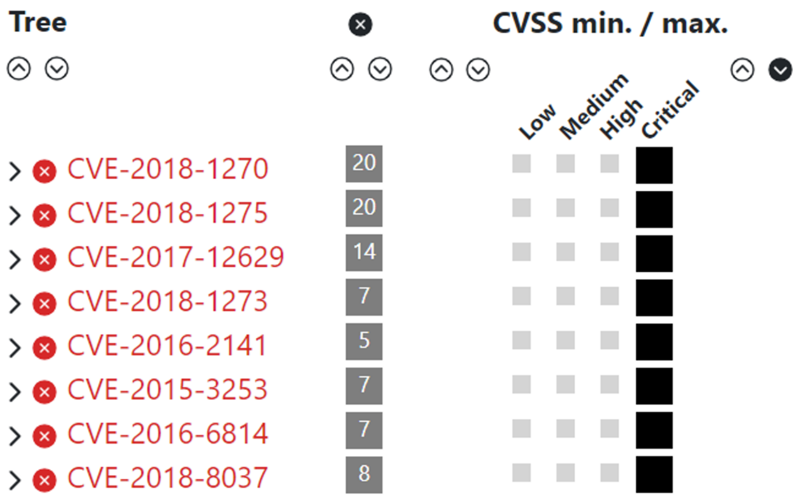
(Q4) Are specific vulnerabilities present in any of the projects?

Analysts searched for the oldest bug for the severities medium, high, and critical. For this task, they use the *Bug Table*. CVEs encode the years that they were detected. To find the oldest unfixed bug, they searched for the different years before 2019. They found *CVE-2009-2625*, a medium severity bug, present in *org.apache.xerces*, which affects 27 repositories. The oldest high severity bug they found was *CVE-2013-1768* in *openjpa-asm-shaded*, affecting three repositories and *CVE-2015-3253*, a critical bug, affecting seven repositories.

Preliminary User Feedback

We conducted an initial preliminary user feedback session with three software security analysts from SAP. All three participants (P1–P3) have five to ten years of experience in software security and work in dedicated security teams. Two participants support developers of mature applications regarding software security, including assessing the relevance and severity of vulnerabilities in open-source components. One participant acted as program manager for open-source security and Development, Security, and Operations (DevSecOps), determining requirements, developing tools, and standardizing the secure consumption and publication of open-source components at SAP managing the Software Development Lifecycle (SDLC). We adopted the pair analytics guidelines of Kaastra and Fisher [169] to structure our interviews conceptually. During the one-hour interviews, we gathered regular user tasks, related employed visual interfaces, familiarity with information visualization for cyber-security, and afterward reviewed and compared in a live session their initial expectations to the proposed VulnEx tool. All three participants approved the usefulness of VulnEx to visually explore the use of open-source libraries in large software organizations. P1 and P2 appreciated that the tool displays how often libraries and their potential vulnerabilities are used in the whole organization. P3 liked that the CVE matrix displays the top five bugs in the organization as it highlights the affected packages, including other prevalent vulnerabilities with their CVSS scores. Overall, all participants believed that VulnEx tool helps explore software organizations' vulnerabilities from different perspectives, such as in repository, library, and bug table views.

The participants expressed some concerns and outlined some shortcomings of our tool. P1 suggested adding additional information about the open-source libraries to the tool, such as short descriptions of the main functionality and purpose of the library. The participant argued that keeping track of each library's functionality without such additional information remains challenging due to the sheer number of 3rd party libraries. P2 emphasized that the current visual representations might not scale to large-scale organizations, e.g., organizations with more than 1000 repositories. P2 also proposed to enable the annotation of individual repositories, libraries, and bugs. Such annotations let analysts search for particular vulnerabilities and guide the auditing team to potential known solutions. P3 emphasized that his focus is heavily on vulnerabilities with critical CVSS scores above 9.0 that need to be resolved within several hours. Therefore, P3 suggested focusing on such vulnerabilities and recommending appropriate counter-measures. All participants suggested including potential solutions to resolve the vulnerabilities.



Tree	☒	Low	Medium	High	Critical
› ☒ CVE-2018-1270	20	▪	▪	▪	■
› ☒ CVE-2018-1275	20	▪	▪	▪	■
› ☒ CVE-2017-12629	14	▪	▪	▪	■
› ☒ CVE-2018-1273	7	▪	▪	▪	■
› ☒ CVE-2016-2141	5	▪	▪	▪	■
› ☒ CVE-2015-3253	7	▪	▪	▪	■
› ☒ CVE-2016-6814	7	▪	▪	▪	■
› ☒ CVE-2018-8037	8	▪	▪	▪	■

Figure 5.3: The analyst found eight critical vulnerabilities. *CVE-2018-1270* affects 20 repositories and has a critical severity.

5.5 Discussion and Future Work

We found that three security experts approved the usefulness of VulnEx. The experts found the different task-focused views useful. We learned that more detailed representations were less preferred. The domain experts had an easier time working with the categories low, medium, high and critical, rather than the precise values of the heatmap visualization. The CVE matrix gives a helpful overview of specific vulnerabilities. All vulnerability analysis tools at SAP focus on individual applications. Thus, we present VulnEx supporting organization- and enterprise-wide decision making. In terms of scalability, we performed our analyses on all public GitHub repositories of *Eclipse Foundation*. Therefore, we argue that VulnEx can be used for large software organizations since few organizations have more projects than the *Eclipse Foundation*.

Future Work: We plan to address the feedback from the security experts by including a method to annotate repositories, modules, libraries, vulnerabilities and provide additional information for each item which could be taken from *libraries.io* or comparable online services. We also plan to include the temporal component, analyzing multiple "snapshots" to compare projects and understand how the organization's risk exposure develops over time. Another goal is to extend VulnEx for the assessment of libraries before choosing a specific one and provide a feedback loop to inform the open-source community and add the vulnerability information to the original repository. We also plan to evaluate VulnEx with experts external to the design process. Our approach is transferable to other organizations and open-source vulnerability analysis tools, but VulnEx is currently limited to the import and processing of scan results from *Eclipse Steady* allowing for the analysis of *Java* and *Python* code.

5.6 Conclusion

Determining the impact of vulnerabilities on software organizations is challenging due to the missing aggregation of software analysis results. As a solution, we propose the VulnEx (Vulnerability Explorer) tool, which we designed in a user-focused design process, which allows analysts to detect severe and relevant vulnerabilities and determine impacted libraries, modules, and repositories. Three security experts confirmed the effectiveness of VulnEx, appreciating its task-oriented views and finding the *ordinal ratings* (low, medium, high, critical) more user-friendly than detailed heatmap values. The CVE matrix provided a concise overview of specific vulnerabilities, highlighting the utility of VulnEx for enterprise-wide decision making beyond individual applications, as is typical at SAP. Our scalability tests on all of the Eclipse Foundation's public GitHub repositories demonstrate that VulnEx is suitable for large software organizations like SAP, as few have more active projects than the Eclipse Foundation.

Part III

Measure-Driven Frameworks

All models are wrong, but some are useful.

— **George E. P. Box**, Statistician

6 A Framework for Relevance Model Building Using Pattern-based Similarity Measures

Contents

Data analysts require automated support for the extraction of relevant patterns. In this chapter, we present FDive, a visual active learning approach that helps to create visually explorable relevance models, assisted by learning a pattern-based similarity. We use a sparse set of user-provided categorical labels to rank similarity measures, consisting of feature descriptor and distance function combinations, by their ability to distinguish relevant from irrelevant data. Based on the best-ranked similarity measure, we calculate an interactive Self-Organizing Map (SOM)-based relevance model, which classifies data according to the cluster affiliation. It also automatically prompts further relevance feedback to improve its accuracy. Uncertain areas, especially near the decision boundaries, are highlighted and can be refined by the user. We evaluate our approach by comparison to state-of-the-art feature selection techniques and demonstrate the usefulness of our approach by a case study classifying electron microscopy images of brain cells. The results show that FDive enhances both the quality and understanding of relevance models and can thus lead to new insights for brain research.

This chapter is *taken from* the following publication:

- [89] **Frederik L. Dennig**, Tom Polk, Zudi Lin, Tobias Schreck, Hanspeter Pfister, and Michael Behrisch. “FDive: Learning Relevance Models Using Pattern-based Similarity Measures”. In: *Proceedings of the 14th IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2019, pp. 69–80. DOI: 10.1109/VAST47406.2019.8986940.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

6.1 The Benefit of Feedback-Driven Relevance Model Building

A primary challenge when analyzing collected data is to distinguish relevant from irrelevant data items. Large and high-dimensional datasets are not easily analyzed, because of their size, dimensionality, and possible complex patterns. Therefore, analysts need automated support. This support is realized in the form of a relevance model that can help them to make this distinction. Its task is the retrieval of relevant data items from large high-dimensional datasets that are often associated with many types of analysis scenarios. Similarity models are key to effective data clustering and classification. It is crucial that the model reflects the notion of relevance as it pertains to the analysis task. More generally, when we are dealing with high-dimensional datasets, we need to automatically and adaptively assess the relevance of data items. Although analysts interact with data for analysis and exploration purposes, their primary goal is to quickly generate new insights and results. All interactions, such as labeling or relevance feedback, should be focused on yielding insights and need to be as impactful as possible.

The fully automatic creation of relevance models is non-trivial. Deep learning approaches, such as Convolutional Neural Networks (CNNs), have been applied successfully, but typically require a large number of labeled training data to distinguish relevant from irrelevant data [193]. Classic machine learning techniques depend on a predefined set of features and a given distance function, chosen or even designed by experts based on their experience. In most real-world scenarios, these labels do not exist and the manual assignment of labels is time consuming, tedious, and expensive. In many analysis scenarios, this is not a viable solution. Transfer learning could be an alternative solution. These methods reapply a previously learned model for a different task then that for which they were originally trained [240]. While the idea seems intriguing, these models are unable to transfer the complex user understanding between datasets. One reason is that the problem and task definition in exploratory scenarios, particularly the pattern space, is highly specific and non-static. Users' mental model of *what makes up relevance* evolves throughout an analysis, thus requiring adaptive methods for the process. Additionally, the created model needs to be understandable, explorable, and refinable in areas where it is inaccurate.

The feedback-driven view exploration pipeline by Behrisch et al. [34] was an early approach towards a relevance model-guided exploration of large multidimensional datasets using Feature Descriptors (FDs). Complex data items can be abstracted using feature descriptors. The resulting features ideally express the properties of the data items concerning the analysis objective. Features reduce the complexity of comparing data items but are limited in their ability to express all properties of a data item. The approach by Behrisch et al. [34] only uses one *fixed* FD, namely Scagnostics [330], limiting the set of described properties and introducing biases into the analysis process. In this work, we tackle the question of choosing an appropriate FD that models the given dataset, analysis domain, and analysis task. We claim that FDs alone do not express the relationship between data items. We also need a distance function that describes their relationships. Depending on the analysis scenario, other measures than the ubiquitous Euclidean distance may perform significantly better [127], which reflects on the performance of the relevance model learning component, too. In this work, we expand on Behrisch et al.'s static decision tree model, in which exploration decisions are irreversible, with a more flexible and adaptive approach to guide the user through the data space. Our classification results and feature abstraction can be visually explained, making the quality of the model easier to capture and more trustworthy.

In this work, we present FDive, a visual analytics approach for the creation of relevance models. In FDive, we model relevance as a binary classification problem. Since the quality of the underlying classification or ranking model depends on the usefulness of the employed FDs and distance function, we introduce the concept of the *Similarity Advisor* engine, which ranks FD-distance function pairs, according to their ability to distinguish relevant from irrelevant data. This removes the need for an expert choosing an FD and distance function manually. The system uses the best-ranked similarity measure for the creation of the relevance model. To learn fine-grained differences between relevant and irrelevant data, we introduce a SOM-based relevance model that classifies data items according to their cluster membership. To allow the judgment of the model quality and model refinement, the SOM-based model is visually explorable and guides the user towards areas of uncertainty. We embed the *Similarity Advisor* and the model learning process into an iterative framework, to allow for convergence towards the optimal similarity measure and relevance model. In summary, we contribute the following:

Contributions

- The general FDive framework using the *Similarity Advisor* to determine an effective pattern-based similarity measure.
- An *instantiation* of FDive learning the relevance of Electron Microscopy (EM) images of mouse brain cells.
- An *expert study* in the domain of computer vision for neurology.
- A *quantitative evaluation* comparing FDive to state-of-the-art feature selection techniques.

6.2 Related Work

In this section, we delineate FDive from other approaches. FDive is a relevance model builder, in contrast to image retrieval tools like PixSearcher [231] which enables users to retrieve images through query by example. In the following, we discuss related concepts such as feature selection, visual active learning, and distance function learning. We also discuss similarities and differences in the area of model visualization and understanding.

6.2.1 Feature Selection for Dimensionality Reduction

Feature selection algorithms typically try to approximate the usefulness of a given feature. These techniques determine a subset of relevant feature dimensions based on feature-ranking and feature-weighting [160, 130]. Although prior studies show how visualizations can support feature selection and optimization in 3D models [276] or exploration of chemical compounds [296, 53], the feature evaluation procedure is reoccurring and potentially exhausting for the user. Thus, we decided to use two purely automatic statistical feature selection algorithms in the evaluation of FDive. First, *ReliefF* [188, 323] is a state-of-the-art extension of the *Relief* algorithm for multi-class problems [214]. It ranks features based on how well they distinguish an instance from its k-nearest neighbors. If a neighbor is from a different class, the weights of features that separate both instances are increased, and all others are decreased accordingly. In case the neighbor is from the same class, the weights of features that separate both instances are decreased, and all others increased. Second, *Linear Ranking Ensembles* combine multiple ranking classifiers, such as the *Recursive Elimination Support Vector Machine Support Vector Machine (SVM)*, into one ranking ensemble. They are, thus, more stable than other approaches [268]. *Recursive Elimination SVMs* iteratively reduce the feature dimensions size using linear SVMs [201]. Attributes are ranked, and the worst performing dimension is removed. This process, including the SVM training, continues until only one feature dimension remains.

The quality of a feature selection depends on the number of available labels and is computationally expensive in scenarios that require continuous reevaluation. With FDive, we provide a solution for this scenario by keeping the feature descriptions while ranking a set of similarity measures, consisting of an FD and a distance function combination, based on how well it separates relevant from irrelevant data. We embed this technique in an iterative process, allowing for an adaptation to the best-suited similarity measure.

6.2.2 Visual Active and Interactive Machine Learning

In a visual Active Learning (AL) approach, users are provided with auxiliary information about the learning process and model state, specifically decision boundaries of the classification model, query choice, and learned instances. Bernard et al. [39] present a visual AL method to assess the well-being of prostate cancer patients from the patient's history, describing interesting biological and therapy events. The tool suggests a set of candidates to label, as well as allowing for the visual verification of the validity of learned instances. Heimerl et al. [145] present a visual AL system as an SVM classifier for text. The tool supports the visualization of the decision boundary, including instances on it, and user-based instance selection for labeling. Eaton et al. [99] adjust the underlying data space by describing it with manifold geometry, allowing users to label data items, serving as control points leading to improved learning performance.

In contrast to AL, the sample selection in Interactive Machine Learning (IML) is driven by the user. Dudley et al. [97] describe a general approach to interface design for IML providing an overview of challenges and common guiding principles. Arendt et al. [15] present an IML interface with model feedback after every interaction by updating the items shown for each class. The users can drag misplace data items to the appropriate class and, if needed, create a new one. Both actions update and improve the model.

FDive is a visual active learning approach that learns a relevance model based on the user's notion of relevance. We propose a SOM-based model, which is interactively explorable, guiding the user to areas of uncertainty and decision boundaries. The model creation and inspection are combined in an iterative workflow that allows the user to observe and judge model change, leading to a more understandable relevance model and learning process.

6.2.3 Distance Function Learning

Another requirement to represent the relationship of data items is a distance function. A distance function can include a feature weighting. The Mahalanobis metric [210] measures the standardized distance of a data point to the estimated mean of its population. Relevant Component Analysis [25] uses a parameterized Mahalanobis distance. This technique adapts the feature space by assigning large weights to relevant dimensions and low weights to irrelevant dimensions through equivalency constraints, describing the similarity of data items. As opposed to purely algorithmic approaches, there are also visual and interactive approaches to the generation of suitable distance functions. Brown et al. [56] learn a distance function from a 2-dimensional projection of the data space where the user drags the data point to the desired position, thus describing similarity relations. The underlying distance function is updated accordingly by the adaptation of feature weights. The work by Gleicher [120] demonstrates the learning of multiple distance functions, each describing the relationship of the data based on different features, capable of describing abstract concepts, such as socio-cultural properties of cities. Fogarty et al. [113] present an image retrieval system that determines the weights of a distance metric based on user-supplied feedback to learn concepts.

In contrast, FDive unifies many concepts mentioned above. It ranks arbitrary feature descriptors and similarity measure combinations by their ability to discriminate relevant from irrelevant data. FDive removes the limitation on a pre-defined set of features through the comparison of multiple FDs describing a diverse set of data properties. Also, a set of similarity coefficients is used, thus removing the limitation of a single similarity coefficient or feature weighting. This makes FDive a generalized relevance model builder for different types of data.

6.2.4 Model Visualization and Understanding

Visual Analytics (VA) aims to provide the analyst with visual user interfaces that tightly integrate automatically obtained results with user feedback [179]. The knowledge generation model [266] describes an iterative process of exploration and verification activities of both human and machine. Results are presented visually to analysts, who interpret obtained patterns and provide feedback to steer the exploration process or form and refine hypotheses. The understanding and interpretation of machine-learned models is key for the effective incorporation of user feedback in such scenarios. Several prior works have studied model visualizations and interactions. BaobabView [101] presents a model where the structure of a decision tree is augmented with data distributions and data flows. Liu and Salvendy [203] and Ankerst et al. [13] use icicle plots [195, 184] to visualize decision trees. Visual interactive approaches for cluster evaluation and understanding were presented by Nam et al. for general high-dimensional data [229] and by Ruppert et al. [264] for the clustering of text documents. Sacha et al. present SOMFlow [265], an exploration system that uses SOM to guide the user through an iterative cluster refinement task, leveraging the proximity-preserving property of SOMs [315, 38] for clustering and data partitioning tasks.

In a model creation task, the user needs to be guided towards areas of high uncertainty. Thus, FDive steers the data exploration to specific parts of the model, such as the decision boundaries. The SOM-based model of FDive is capable of providing the necessary information about uncertain areas and automatic refinement.

6.3 Similarity Advised Model Learning

The key idea of our approach is to iteratively and interactively create relevance models, where a useful feature description is unknown, and no or only few labels are available. Our proposed *Similarity Advisor* allows approaching the question which feature descriptor and similarity measure combination is useful to distinguish relevant from irrelevant data items. In a scenario where labels are sparse, the quantitative validation of classification models with performance measures is inexpressive. Thus, there is a need for techniques that allow for model assessment without test data. Classifiers, such as SVMs, have been used in visual active learning approaches [145]. However, the representation of the data space created by SVMs does not allow the user to judge the quality of a classifier visually. Decision trees are more intuitively interpretable.

We propose a SOM-based classification model which is embedded in an iterative workflow to allow for observable learning steps. In each step, the model is explorable and refinable to judge and improve its quality. Both, the *Similarity Advisor* and the SOM-based classification model constitute FDive, a generalized model builder. In the following, we provide an introduction to SOMs.

Self-Organizing Map (SOM): FDive relies on a *neural network architecture*, called SOM or Kohonen Network. SOMs are the basic building block of our relevance model and are one of the classical neural network structures, created by Kohonen to derive topologically coherent feature maps [187]. SOMs can be visualized as a grid of cells representing the neurons of the network. The cells contain prototype vectors representing data clusters. In the learning phase of the network, the most similar prototype vector (best-matching-unit) to the training input is identified and adjusted towards the input vector. Spatially close neighbors are also adapted, depending on a learning rate and radius parameter. The latter gives rise to the self-organization property of the map. The final result is a topology, where data items are clustered. Clusters can consist of single or multiple cells, and cluster similarity can be captured by spatial proximity of clusters on the SOM grid [38, 265].

We extend this algorithm into a tree-like classifier to allow for the representation of fine-grained similarity differences. This concept is based on the idea that items can "flow" from a parent SOM node into a child SOM for further analysis, as presented by Sacha et al. [265]. In our work, we extend this idea to create a classification model that *automatically* partitions the high-dimensional data space into relevant and irrelevant data item clusters. We will detail this approach in section 6.6. We use an interactive SOM visualization to allow for the visual inspection of the currently learned model, e.g., where groups of relevant or irrelevant data elements are located, and how well decision boundaries can distinguish known groups.

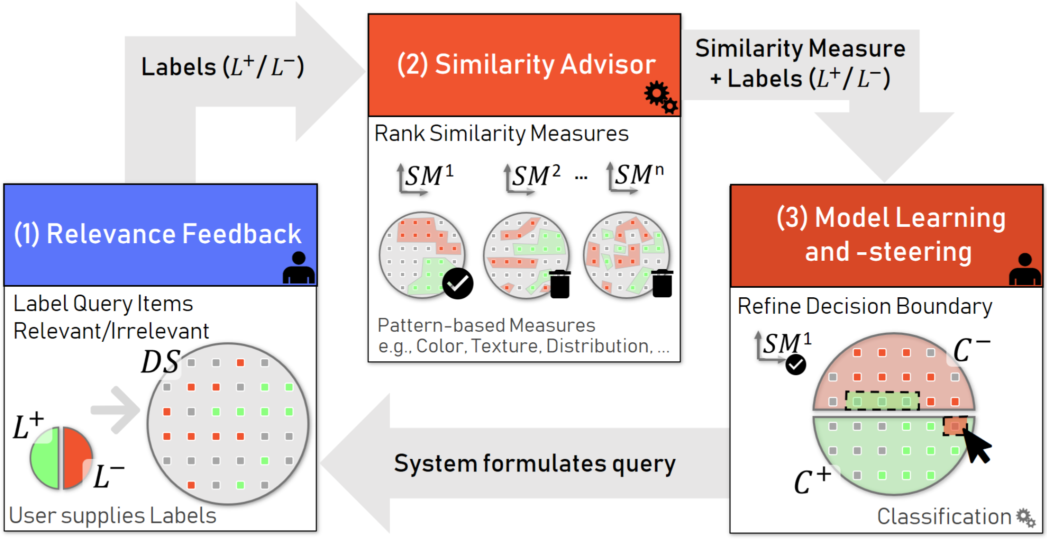


Figure 6.1: (1) Users provide categorical labels (relevant, neutral, irrelevant) to express their idea of relevance. (2) This selection is used to automatically determine the best-fitting similarity measure, which distinguishes relevant from irrelevant data. (3) The system adapts the model using the relevance labels and similarity measure. The model is explorable and refinable by the users, to improve its accuracy.

Workflow for Iterative Relevance Model Learning

FDive is inspired by the feedback-driven interactive exploration tool by Behrisch et al. [34], which propose an iterative and FD-based exploration framework. A central principle is to represent an arbitrary dataset with the help of visualizations to make it accessible for an analyst. This visualization needs to be translated into a language understood by a computer, which uses this *proxy* to guide through the information- and pattern space which is achieved by a single fixed FD introducing bias into the analysis process. We expand this body of work by changing the focus from an exploratory approach to a model building technique. The validation of relevance models, though, is a challenging task, due to the following reasons. We need to define a useful definition of similarity, but a metric for separating classes can only be determined during the learning process. What is needed are flexible and adaptive strategies for determining a useful metric defining similarity. FDive allows for arbitrary data modeling through the *Similarity Advisor*, which ranks a set of FDs and distance functions by their usefulness concerning the current analysis domain and dataset properties. The FD, representing the data modeling, is subsequently used to create a relevance model. Additionally, the model needs to be explorable and refinable to convince an analyst of its usefulness and accuracy.

In FDive, we leverage an iterative workflow to continuously revalidate the similarity measure and improve the relevance model. In the following, we describe each iteration step and its impact. Figure 6.1 shows each step accordingly.

(1) **Relevance Feedback:** The system prompts the user to label a subset of data items of the dataset (DS) using the categories relevant or irrelevant, representing relevance as a binary classification problem. Those data items labeled as relevant are referred to as $\mathcal{L}^+$ and all labeled as irrelevant as $\mathcal{L}^-$. Unlabeled data items are considered neutral. In the first iterations, this step is replaced by a query generated through a representative data sample. In all following iterations, the query is determined by the SOM-based model. FDive supports the user by visual feedback allowing the validity assessment of a currently used similarity measure and classification through visual feedback. This step is described in detail in section 6.4.

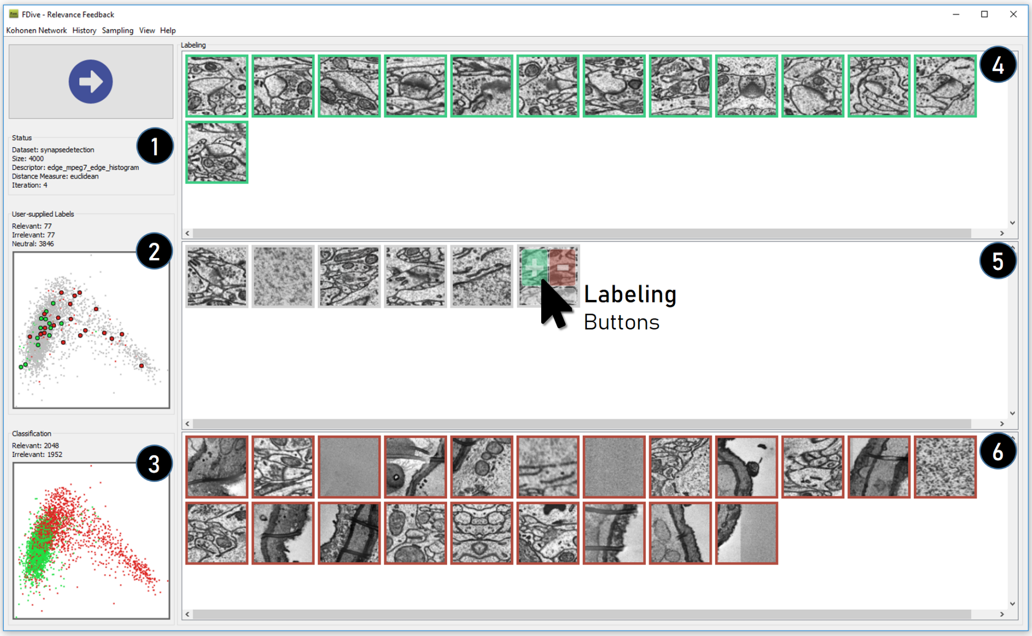
(2) **Similarity Advisor:** The system evaluates all possible pair-wise combinations of FDs and distance functions by their ability to separate relevant ($\mathcal{L}^+$) from irrelevant ($\mathcal{L}^-$) data items. A ranking shows the evaluation result, giving an intuition about the similarity measures. The user can follow the recommendation or choose a different similarity measure. The system uses the FD and distance function for the creation of the relevance model. We describe the algorithmic background of the *Similarity Advisor* in section 6.5.

(3) **Model Learning and Steering:** The system creates a classification model based on the selected similarity measure and available labeled data ($\mathcal{L}^+$ and $\mathcal{L}^-$). The model can be explored to asses its properties and viability for its classification task. The classification result is referred to with $\mathcal{C}^+$ describing all data items classified as relevant and all irrelevant as $\mathcal{C}^-$. The SOM-model creation and interactions are described in section 6.6. Subsequently, the system determines a set of query items which are labeled by the user in the first step of the next iteration.

In the following, we describe the design, user interaction and algorithmic support in FDive.

6.4 Context-Aware Relevance Feedback

Data labeling is the first and reoccurring step in our relevance model learning process from section 6.3. During start-up, this essential bootstrapping step helps us to form a decision basis for the subsequent application of our *Similarity Advisor*. Throughout the learning process the classifier queries relevance labels through this interface to improve its accuracy. We describe this step of FDive in section 6.6.



Status
Dataset: synapsedetection
Size: 4000
Descriptor: edge_mpeg7_edge_histogram
Distance Measure: euclidean
Iteration: 4

User-supplied Labels
Relevant: 77
Irrelevant: 77
Neutral: 3846

Classification
Relevant: 2048
Irrelevant: 1952

FDive - Relevance Feedback
Kohonen Network History Sampling View Help
Labeling
1
2
3
4
5
6
Labeling
Buttons

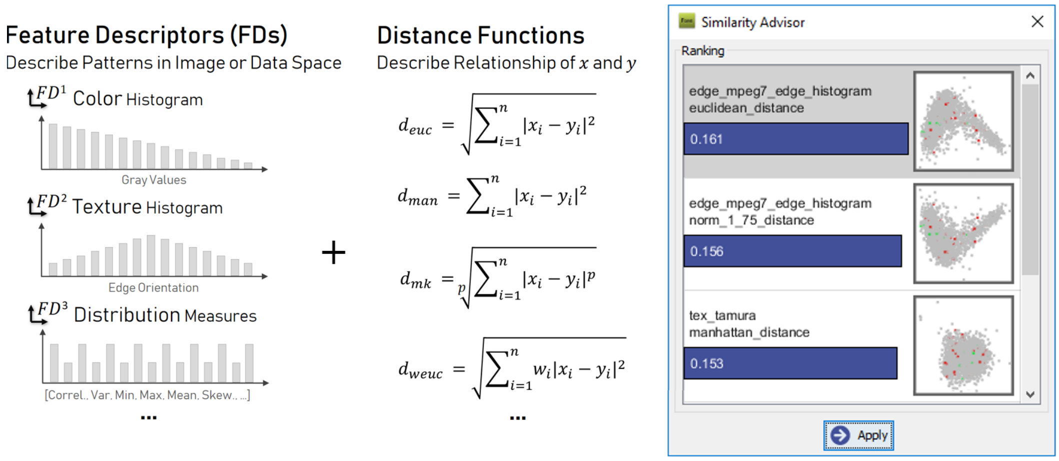
Figure 6.2: Context-aware Relevance Feedback: (1) Status display showing the current analysis state. (2) Scatter plot highlighting newly labeled data. (3) Scatter plot of the current classification result. Both allow judging the impact of new labels. (5) Queried neutral data items. (4) Data items labeled as relevant and irrelevant (6).

Relevance Feedback of Representatives

We sample data items in the first iteration for an initial user labeling. The sampling method can be chosen from the following options: Minimum-Maximum-, Quantile Sampling, Normal-, Stratified Normal Bootstrapping, Normal- or Stratified Subsampling [34]. In all following iterations, the request for labeling is determined by the relevance model, in our case a SOM-based model (see section 6.6). The user can apply three types of labels: relevant, irrelevant and neutral. While the relevant and irrelevant labels express a user preference and have an impact on all steps of FDive, neutral represents an uncertain item. The model may prompt a label for the given element at a later iteration. The user labels a subset of displayed data items by clicking on the mouse-over menu or using a keyboard-shortcut. For visual clarity, all elements are assigned to specific panels (relevant, neutral, irrelevant, from top to bottom in figure 6.2 (3-5), according to their label type, which also allows comparing items with the same relevance label.

Visual Assessment of Labeling Impact

A status display (figure 6.2 (1)) provides information about the current analysis state, such as the current FD and distance function, the number of supplied relevant


Feature Descriptors (FDs)
Describe Patterns in Image or Data Space
FD^1 Color Histogram
Gray Values
FD^2 Texture Histogram
Edge Orientation
FD^3 Distribution Measures
[Correl., Var. Min. Max. Mean. Skew.. ...]
...
+
Distance Functions
Describe Relationship of x and y
$d_{euc} = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$
$d_{man} = \sum_{i=1}^{n} |x_i - y_i|^2$
$d_{mk} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$
$d_{weuc} = \sqrt{\sum_{i=1}^{n} w_i |x_i - y_i|^2}$
...
Similarity Advisor
Ranking
Apply

edge_mpeg7_edge_histogram
euclidean_distance
0.161

edge_mpeg7_edge_histogram
norm_1_75_distance
0.156

tex_tamura
manhattan_distance
0.153

(a) Computational model **(b)** User interface

Figure 6.3: The *Similarity Advisor* uses a set of FDs and distance functions. FDs model the data based on perceptible patterns in the data or image space. Distance functions describe the relationship between two points in the FDs space. In FDive, we consider all pair-wise combinations as potentially useful measures. We call a combination of an FD and a distance function a *pattern-based similarity measure* (see (a)). The *Similarity Advisor* ranks all pair-wise combinations of FDs and distance functions according to their ability to distinguish relevant from irrelevant data. A bar indicates the score and a scatter plot shows the topology of implied data distribution allowing users to judge its usefulness (see (b)).

and irrelevant labels, as well as the number of remaining neutral items. A scatter plot (figure 6.2 (2)) of the dataset using the currently chosen FD and distance function depicts the possible impact of new labels when compared to the projection of the classification result (figure 6.2 (3)). We create both 2D projections with Multidimensional Scaling (MDS). MDS projects the data in a distance-preserving way without the need for additional parameters. The annotation view is also used to refine the labels in the SOM-based model and explore elements assigned to a SOM-neuron (section 6.6). Chegini et al. explored the idea of showing the classification result in a scatter plot [68], while the visual feedback on data labeling was evaluated by Bernard et al. [37]. Combining both approaches allows assessing the impact of newly assigned labels in a natural form. The comparison of both scatter plots shows the effect of new labels, e.g., a relevant label in an area of irrelevant classifications hints at an incomplete reflection of the user's notion of relevance, a matching label hints at a convergence.

6.5 Assessing Pattern-Based Similarity Measures

The goal of the *Similarity Advisor* is to select the most expressive FD and distance function combination from a predefined set of FDs and distance measures to improve the relevance model creation. We claim that a combination of FD and distance measure can define a *pattern-based similarity measure*. To describe the discriminative ability, we need a *quality metric* that reflects the similarity measure's ability to distinguish between our relevant and irrelevant items. We consider a useful similarity measure one that maximizes the distance between both sets $\mathcal{L}^+$ and $\mathcal{L}^-$. We considered other quality metrics, such as metrics that measure distances between elements of a cluster, but found them lacking in performance. We propose the *Similarity Advisor* for the selection of a suitable distance metric; this includes the choice of an FD and a distance function. For this, we require a set of diverse FDs. We use various FDs from the Image Processing and Computer Vision Community because these algorithms are designed to match the human perceptual system.

In essence, the application scenario determines the usefulness of a feature description and distance function. However, the selection of a useful distance function is hard. Thus, we introduce the concept of continuously evaluating a set of *pattern-based similarity measures* for their applicability to the current analysis task, allowing for the convergence to the most useful one. To describe the algorithmic basis of the *Similarity Advisor*, we define all relevant terms.

Feature Descriptor (FD): FDs are modeling specific characteristics of a data item. Examples for low-level FDs are color histogram descriptors, modeling the color distribution, or edge histograms describing edge orientations of an image [263]. Low-level FDs are typically inexpensive to compute and may work robustly. Depending on the type of data at hand, many FDs are applicable. Mathematically, an FD can be described as a function $FD : DS \to \mathbb{R}^n$, where DS denotes the dataset and $\mathbb{R}^n$ the implied vector space. The dimensionality n depends on the Feature Space (FS). Table 6.1 lists all FDs used by FDive. These FDs describe a variety of different image features, such as color, layout, structure, and shape [32].

Feature Vector (FV): An FV is an instantiation of an FD for a specific data item. An FV contains one or multiple components, called feature dimensions or features. A feature vector $FD(x) \in \mathbb{R}^n$ represents a description of a data item $x \in DS$, w.r.t. the properties described by the applied FD.

Feature Space (FS): A feature space describes the set of all feature vectors created by an individual feature descriptor. Additionally, a feature descriptor implies a vector space, called feature space. Thus, each feature descriptor has an associated vector space.

Pattern-based Similarity Measure: We define a *pattern-based similarity measure* as a combination of one feature descriptor and a single distance function (figure 6.3a). The *Similarity Advisor* evaluates the usefulness all possible combinations of an FD

Color	**Color Layout**
Auto Color Correlogram [154]	Cedd [66]
Fuzzy Histogram [133]	Fcth [67]
Fuzzy Opponent Histogram [269]	Jcd [66]
Global Color Histogram [263]	Luminance Layout [206]
Opponent Histogram [269]	MPEG7 Color Layout [173]
Edge	**Structure**
Edgehist [32]	JPEG Coefficient Histogram [206]
MPEG7 Edge Histogram [226]	Phog [49]
Hough [152]	Profile [32]
Texture	**Other**
Gabor [206]	Blocks [32]
Haralick [135]	Compactness [226]
Local Binary Pattern [144]	Magnostics [32]
Tamura [299]	Statistical Noise [32]

Table 6.1: FDive uses 24 feature descriptors. These FDs describe a variety of different image features, such as color, layout, structure and shape [32] allowing for a description of a diverse set of properties.

and a distance function in their ability to separate the clusters of relevant (L^+) and irrelevant (L^-) data items.

In FDive, we use a set of norms as distance functions because the FD learning algorithm requires a similarity measure that can describe a vector space allowing for an adaptation of the cluster prototypes “towards” an input vector. FDive uses the following norms: Euclidean L^2, Manhattan L^1, $L^{1.25}$-norm, $L^{1.5}$-norm and $L^{1.75}$-norm, which are all L^p-norms with $||x||^p = (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ and the implied metric $d(x, y) = ||x - y||$ as a similarity measure.

Comparability of Pattern-based Similarity Measures

Every FD describes a different set of data properties by mapping a data item to a vector representation. To derive useful similarity relations, we need to use a distance function that applies to the vector. We limit ourselves to L^p-norms. However, this approach is extendable to other distance functions and similarity coefficients, including those which do not satisfy the metric axioms.

We leverage the definition of normed vector spaces, which is defined as $(V, || \cdot ||)$ where V is a vector space and $|| \cdot ||$ a norm on V. We use this definition and apply it to the combination of an FD and its FS along with an L^p-norm with $p \in [1, \infty)$. Throughout this chapter, the term distance function refers to the induced metric $d(x, y) = ||x - y||$. In FDive, we define a *pattern-based dissimilarity measure*, a combination of a single FD and a distance function, formally as $dist_d^{FD} : (x, y) \rightarrow$

$[0, \infty)$ with $dist_d^{FD}(x, y) = d(FD(x), FD(y))$ and $x, y \in DS$ data items of the dataset.

We apply a non-standard normalization to transfer a feature space FS and the associated norm into a comparable format. To achieve this outcome, we center the set of all feature vectors $x \in FS$ on the origin, such that the center of each dimension range is located at the origin. This translation does not change vector distances. For this we create a translation vector $t \in \mathbb{R}^n$. The components of t are defined for each dimension i as

$$t_i = \frac{1}{2} \cdot (max_{v \in FS}(v_i) + min_{v \in FS}(v_i)) \tag{6.1}$$

With this, we can formalize the necessary normalization step to transform the feature space into a comparable state as described by the following function.

$$normalize(x) = (x - t)/max_{v \in FS}(||v - t||) \text{ with } x \in FS \tag{6.2}$$

The normalization needs to be performed for all elements x of feature space to convert it into a comparable format. This normalization can be implemented with a complexity of $O(N \cdot M)$ for the full dataset of size N and M *pattern-based similarity measures* implied by the similarity measures, leveraging the mathematical definition of a norm. In essence, this transformation translates all vectors such that the center of each dimension range is located at the origin and scales all vectors such that $||x|| \in [0, 1]$ for all vectors x, while preserving relative distances between all vectors according to the norm. This normalization allows us to compare the different topologies created by different feature descriptor and norm combinations.

This approach can extend to non-norm similarity coefficients, under the following implications. (1) Ideally, the subsequently applied classification model is compatible with the similarity coefficient, e.g., SOMs require a norm as an internal distance function since prototype vectors need to be updated “towards” an input vector. (2) With non-norm similarity coefficients, the following non-standard normalization needs to be performed. Non-norm similarity coefficients define the difference purely by the distance of data items. This requires the normalization of the full distance matrix of the feature space. This leads to a significant complexity increase since all pair-wise distances need to be computed in $O(N^2 \cdot M)$.

Quality of Pattern-based Similarity Measures

In this section, we discuss a set of heuristic quality metrics that we designed to estimate the applicability of a similarity measure for a given analysis task. All quality metrics are calculated based on the transformed features space and the associated distance function, according to the previous section. We measure two concepts, *Inter-Group-Distance*, and *Intra-Group-Distance*. A group is defined as a set data items sharing an identical label, i.e., relevant or irrelevant. Thus one group is formed by all elements in $\mathcal{L}^+$ and another by $\mathcal{L}^+$. An intuition is given in figure 6.4.

Inter-Group-Distance measures the similarity of the groups, by calculating synthetic centroids of $\mathcal{L}^+$ and $\mathcal{L}^-$, and subsequently determining the distance between both centroids or short $Q_{inter}(\mathcal{L}^+, \mathcal{L}^-) = dist(\mathcal{L}_c^+, \mathcal{L}_c^-)$. A large *Inter-Group-Distance* is highly desirable.

Intra-Group-Distance measures the maximum distance between distinct elements one of label, i.e. $\mathcal{L}^+$ and $\mathcal{L}^-$. Thus, we can say $Q_{intra}(\mathcal{L}) = max_{i,j \in \mathcal{L}}(dist(\mathcal{L}_i, \mathcal{L}_j))$, where $i \neq j$. We will apply the above heuristic for every dissimilarity measure.

We experimented with different combinations of *Inter-* and *Intra-Group-Distance* and variants also involving mean and median values instead of the maximum for the *Intra-group-distance*. We also combined both measures into $Q_{comb}(\mathcal{L}^+, \mathcal{L}^-) = Q_{inter}(\mathcal{L}^+, \mathcal{L}^-) - w \cdot (Q_{intra}(\mathcal{L}^+) + Q_{intra}(\mathcal{L}^-))$, with a weighting w. In general, we found that the *Inter-Group-Distance* performed the best on its own, i.e., with $w = 0$.

Other metrics in the context of internal cluster quality metrics use similar notions to *Inter-* and *Intra-Group-Distance*. Cutting et al. [77] describe internal cluster metrics such as the cluster self-similarity defined as the average distance of all cluster members or the average distance of all cluster members to the centroid. We found that this measure did not describe the group separation very well since the ideal case describes a cluster concentrated on a small region. This case rarely occurs in real-world scenarios, without all points of both L^+ and L^- clusters being concentrated at the same location. We looked at internal cluster quality metrics such as the Dunn Index [98] which measures the ratio of minimum cluster distance to the maximum cluster extent. Another measure is the Davies-Bouldin index [79] describing the sum of cluster extents to the centroid distances. Both approaches include the notion similar to the *Intra-Group-Distance*. We found that both measures were sensitive to outliers and thus where not as useful as the *Inter-Group-Distance* heuristic.

We use and suggest the *Inter-Group-Distance* on its own in all applications and evaluations of FDive. This distance-based score is used to rank the set of similarity in descending order, as shown in figure 6.3b. The *Similarity Advisor* shows the score as bar. Additionally, we display the topology of the associated features space. Labeled data items are highlighted, allowing users to verify the separation of relevant and irrelevant data items. With the *Inter-Group-Distance* we found a heuristic that is intuitive, easy to calculate and performs well, as we will show in section 6.7.2.

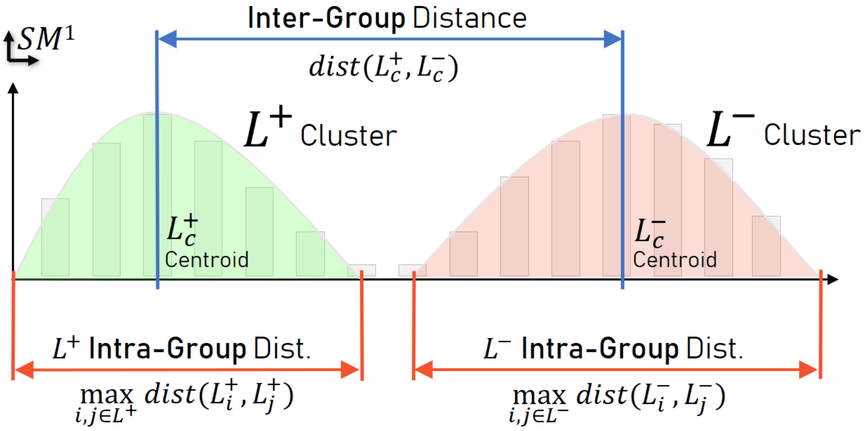


Figure 6.4: We propose two quality measures. *Inter-Group-Distance* describes the distance between the centroids of the relevant and irrelevant data, measuring how well a similarity measure separates both groups. The *Intra-Group-Distance* is defined as the maximum distance in the relevant or irrelevant data, measuring whether a similarity measure describes elements of the same group to be dissimilar.

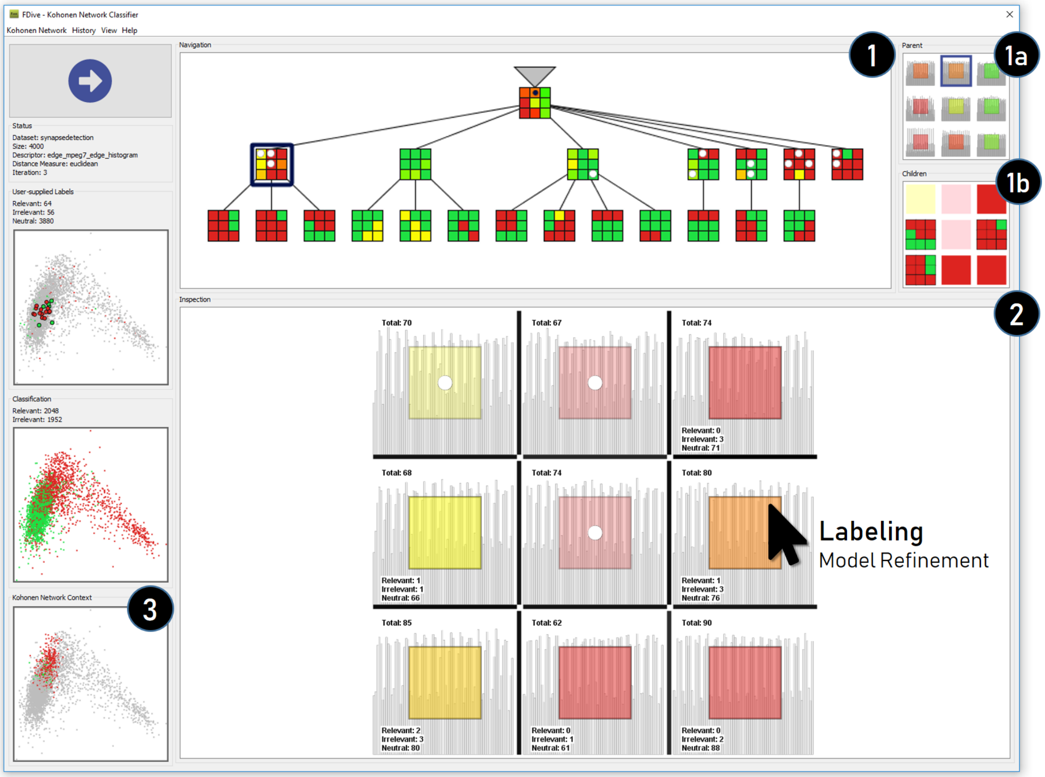
6.6 Learning Relevance of Data Points with Self-Organizing Maps

FDive features a SOM-based classifier, which is used to classify data items by their assignment to a SOM-neuron, and to learn decision planes in the high-dimensional space discriminating $\mathcal{L}^+$ and $\mathcal{L}^-$. We introduce a set of visual encodings to guide the user to potentially interesting data subsets, or regions of classifier uncertainty.

6.6.1 Self-Organizing Maps as Visual Classifier

SOMs cluster similar items in cells, which provides users with an intuition about the classification process. SOMs preserve distance relations between cells allowing for orientation in the data space. The tree-structure and SOM cell exploration allow for a drill-down from the data space to clusters and individual data items. SOM cells are arranged in a grid which is directly visualizable, which also applies and our tree-like classifier model. Additionally, our SOM classifier conveys areas of uncertain classifications by highlighting cells with mixed labeling and cells with a low amount of labeled data items. Additional labels improve the classification. Labels can be added in those specific areas. The grid size is a user parameter, and 3×3 is the default setting.

We use SOMs as a basis for our model because it is visually explorable; it partitions the feature space and the data space, which provides the user with analyzable chunks. The supplied relevance labels and the selected dissimilarity



Status
Dataset: synapsedetection
Size: 4000
Descriptor: edge_mpeg7_edge_histogram
Distance Measure: euclidean
Iteration: 3

User-supplied Labels
Relevant: 64
Irrelevant: 56
Neutral: 3880

Classification
Relevant: 2048
Irrelevant: 1952

Kohonen Network Context

Figure 6.5: Visual Exploration of SOM Model: 1) Classifier tree. 1a) Parent of the currently observed SOM. 1b) Children of the current SOM. 2) Detailed SOM Display. 3) Scatter plot highlighting data of the SOM node.

definition are used to calculate a SOM-based relevance model that separates relevant and irrelevant data items. The model can be explored for visual model understanding. Moreover, the model visually conveys areas of uncertainty. The user is then able to refine the relevance feedback in areas of uncertainty, namely the decision boundary. Since our approach is focused on the creation of a relevance model reflecting the user's notion of relevance and thus in essence, not for exploratory analysis, we limit our approach to the representation of a user's fixed notion of relevance.

Classifier Training: A regular SOM is likely to create cells in which relevant and irrelevant items are mixed. We resolve this by proposing a hierarchical SOMs that allows for the expression of fine-grained differences in the user's notion of relevance. For this reason, we merge the concept of a tree with the concept of child SOMs presented by Sacha et al. [265], where a new SOM is calculated only with a subset of the dataset determined by the cell selection of a parent SOM. However, our algorithm creates a classifier automatically without any user interaction other than supplied labels. We automatically calculate a child SOM only from the data items assigned to the given cell c if this cell exhibits a mixture of relevant of irrelevant greater then a threshold m_t, i.e., $MixRatio(c) > m_t$ with

$$MixRatio(c) = min(|\mathcal{L}_c^+|/|\mathcal{L}_c|, |\mathcal{L}_c^-|/|\mathcal{L}_c|) \tag{6.3}$$

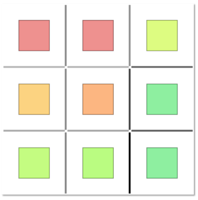
The cell needs to contain enough data items in order for a child SOM to be useful. We model this circumstance by another threshold value c_t, such that the number of items in a cell $|E_c|$ must exceed c_t. Thus c_t determines the split criterion. In FDive, the creation of SOM models is based on the supplied similarity measure, as determined by the *Similarity Advisor*, and the relevance labels. The resulting SOM-based model can exhibit a tree structure (figure 6.5 (1)). We limit the layout to 3×3 to leverage the projection of a SOM into 2D but not handle an excessive amount of children for a given parent in the classification tree.

Classification of Data Items: A classification of a given data item is performed recursively, similar to a decision tree. (1) Find the most similar neuron in the root SOM; (2) If the node has a child, perform the same action recursively on the child SOM; (3) If the SOM node has no child, classify the item as the predominant label of the respective cell, i.e., the categories relevant or irrelevant; (4) If no label information is available for this node, use the next most similar cell with label information in that specific SOM.

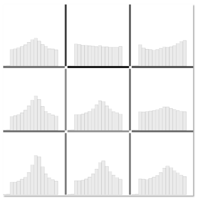
6.6.2 Classifier Exploration and Refinement

Our SOM-based visual classifier is visually explorable. It conveys its relevance decisions through multiple visual and interactive techniques. The main navigation happens in the visual classifier tree (figure 6.5 (1)). Each SOM can be selected to examine it in detail. The currently active SOM is marked with a purple border. A purple dot highlights the parent SOM of the selected child SOM. The color coding of the grid in each SOM is intuitive, green signals a predominance of relevant items, red a predominance of irrelevant items. Yellow signals a mixture of relevant and irrelevant items, according to the $MixRatio$ of a cell. Such cells are likely to be recursively refined, as described in the previous section. We deliberately chose this encoding since it intuitively signals the relevance of data items from green over yellow to red gradient. Figure 6.5 (3) shows the classification outcome for data items assigned to a child SOM or individual cell. This allows us to detect whether a cell is on decision boundary.

To provide insight into the data items assigned to each node, we provide a range of stackable cell visualizations that can be selected in a user-defined order.



Relevance Label Quality: The label quality is depicted as colored squares on top of each node. We use the $MixRatio$ to determine the color and create a gradient from red over yellow to green; red is representing only irrelevant, green only relevant items within the cell and yellow implies an uncertain cell, i.e., decision boundary. A white dot signalizes that the cell contains not enough categorically labeled data items, visually prompting users for more labels.



Feature Histogram: This layer displays the trained vector of the node. It can be used to judge the differences of SOM cells according to the currently recommended feature description. If the currently active FD is interpretable, like an FD derived from a color histogram, describing the color spectrum of an image, it can also hint at the properties of the contained data items.

The user can also utilize two other layers, the Quantization Error (QE) [248] and the U-Matrix [315], to explore clusters of nodes that should be treated similarly by the model. Also, we support the user with detailed information about the number of assigned data items, relevant, irrelevant, and neutral data items. This information allows the user to judge the importance of a given node and the amount of information available to the model.

Visual Active Learning with SOMs

Cells with a low amount of labeled data items are uncertain. We measure this uncertainty with the $LabelRatio$ of a cell c. $|E_c|$ defines the number of items in a cell. Thus, we define the $LabelRatio$ as

$$LabelRatio(c) = (|\mathcal{L}_c^+| + |\mathcal{L}_c^-|)/|\mathcal{E}_c| \tag{6.4}$$

The model marks cells that do not have a child SOM with a white dot if $LabelRatio(c) < q_t$, where q_t defines a threshold. A white dot signals uncertain neurons with a low label count to prompt the user to supply additional labels in these uncertain data regions. If the user does not supply an additional label by the suggested SOM node, the query formulated by an active learning system is generated from those marked nodes. For every node, the user can request details-on-demand in the form of a model-refinement dialog, similar to the annotation view, presented in section 6.4.

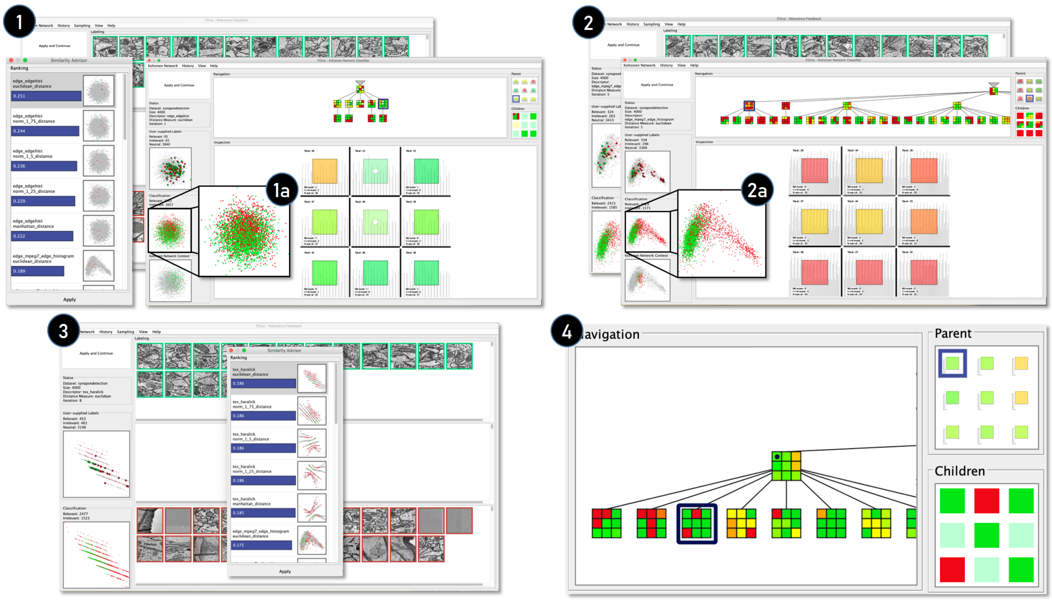


Figure 6.6: FDive learns to differentiate EM images containing synapses from images that do not. The domain expert found the classification model to be satisfactory after nine iterations. We show four key events in the model learning process. (1) The initial model is classifying the data very poorly, as presented by the scatter plot (1a) being very noisy and mixed. (2) The scatter plot shows a cleaner decision boundary (2a) and the model gets more complex, while the expert labels requested data items. (3) In the seventh iteration, the domain expert noticed that the HARALICK [135] FD combined with the Euclidean distance is recommended for the third time in a row, hinting at convergence for the similarity measure. (4) The last two iterations were spent exploring the model, observing and refining decision boundaries.

6.7 Evaluation

Approaches involving relevance feedback are not straightforward to evaluate, as the results depend on both hidden and explicit user preferences and the definition of the learning components [246]. Therefore, we show its usefulness by applying it to a real-world use-case. We evaluate the general workflow, including the *Similarity Advisor*, through a comparison to multiple feature selection techniques.

6.7.1 Case Study: Synapse Detection

The goal of connectomics is to reconstruct the neural wiring diagram from EM images of the animal brain to improve the understanding of neuropathology and intelligence. A synapse is a functional structure that enables signal transfer from one neuron to the other, which connects individual neurons into a complex network. Manual labeling of synapses can be extremely hard because (1) there are approximately one billion synapses in a $1mm^3$ cube of a mouse brain, and (2) the labeling of synapses requires expertise and cannot be crowdsourced. Therefore, a good labeling system of synapses should be semi-automatic and only provide informative samples to the domain experts to improve the labeling efficiency. To showcase the effectiveness of our proposed approach, we applied the annotation system to a high-resolution EM image dataset generated by a multi-beam scanning electron microscope In total, there are 4,000 image patches, half of them containing a synapse at the center of the image, while the other half do not contain synapses. In this study, we show how our system helps experts classify synapse images and non-synapse images without any labeled training set and pre-specified domain knowledge.

CNN-based approaches have achieved state-of-the-art performance on image classification tasks [193, 142]. However, there are still two main shortcomings of CNN-based methods. First, because the model space of CNNs can be huge, the model can easily overfit the training set and have poor performance on the test set, which requires a large training set. Second, the features extracted over convolutional layers are hard to interpret, which restricts the understanding of the discriminative features, especially for scientific applications where the expert wants to have a full understanding of the model.

Thus, we perform a case study involving the classification of EM images of brain cells. A domain expert is tasked with the creation of a relevance model able to distinguish images depicting neuronal synapses. The domain expert has experience in the area of connectomics and the interpretation of EM images, including the identification of cell structures such as cell organelles and neuronal synapses. The study was conducted as a semi-structured interview. The case-study was performed after a training period. The expert performed a total of nine iterations to teach our relevance model the difference between EM images containing synapses and those which do not. Figure 6.6 shows four key events in the model learning process. After the initial annotation of 40 data items, the system suggested the EDGEHIST FD. The expert finished the first iteration by labeling data items in cells with a white dot. A total of 95 images were annotated as relevant and 65 as irrelevant. In the second iteration, the system suggested the TAMURA FD. The expert labeled 63 images as relevant and 57 as irrelevant. In the third iteration, the system suggested the TAMURA FD again. In the fourth and fifth iteration, the MPEG7 EDGE HISTOGRAM FD was suggested. In iteration six to nine the system consistently suggested the HARALICK FD point at convergence on this specific FD. The expert followed the recommendations of the *Similarity Advisor* in every iteration, finishing after the ninth iteration.

In the first three iterations, the system indicated uncertain cells. In later iterations, we are able to check the distribution of samples in a SOM on the scatter plots to see if they are still mixed up. In the end, it notified the expert that it has enough labels, such that no further inspection or labeling is necessary. After several iterations of labeling, the expert noticed that samples are separated in the classification scatter plot, and, when inspecting the individual nodes pertaining to a data region, the labels of similar data items were matching. From the root node to the leaf nodes, he was able to see a trend towards purity. Therefore, when the uncertainty indicators (i.e., white dot) disappears, the nodes with mixed colors are more appealing to be labeled. The inspection of nodes was helpful to the expert to validate whether a set of samples spread out on the scatter plots and thus do not form a coherent cluster. When inspecting a cell colored in yellow, the expert was able to see decision boundaries. Subsequently, the expert labeled ten queried samples to refine the decision boundary. After labeling one node, the color of the node itself and its sibling nodes may change, and the expert was able to verify the impact. The expert noted that the appearance of the scatter plot changed several times at the initial iteration and that the relevant and irrelevant samples on the scatter plots were mixed and not forming a coherent cluster. However, after several iterations, the model converges to a specific similarity measure, and samples become more separable on the scatter plots.

With FDive we can learn to distinguish and extract relevant data items, in this case, EM images depicting synapses, using a sparse amount of categorical labels. Whenever a new label is applied, the system conveys its impact visually. The relevance model is visually explorable and refinable such that the expert was able to assess the model quality and the convergence towards a useful relevance model.

6.7.2 Quantitative Framework Evaluation

This evaluation compares the best best-breed-competitor generated by 3 algorithms and 4 different FD sizes against our “one-shot” *Similarity Advisor* result. Comparing a recombination of all features with the *Similarity Advisor* using only the predefined feature descriptors make this evaluation biased against our approach. However, we were still able to outperform the best best-breed-competitor in 36 out of 75 cases. We evaluate FDive on the following options and parameter settings with the central goal to show the usefulness of ranking pattern-based similarity measures for model learning. We provide a comprehensive overview of the results in Table 6.2. The basis for all experiments is the *Quick, Draw!* dataset [167]. We reduced the dataset to 4500 images consisting of 150 sketches for each of the 30 labels, describing the depicted objects. We choose the labels *square*, *circle*, *banana*, *crayon*, and *monkey*. These labels cover a variety of shapes with different complexity. We assume each label as a specific analysis target. For each of the target labels, we label progressively more items as relevant and irrelevant. The progression is 25/25, 50/50, 75/75, 100/100, and 125/125 for $\mathcal{L}^+$ / $\mathcal{L}^-$. This sequence represents an increase in the

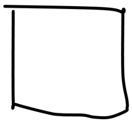
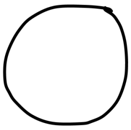
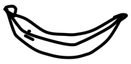
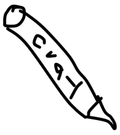
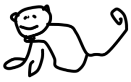
Labeling		**Results** (larger is better)					
Target Example	**#Labels**	**Best Selected FD**			**Best Ranked Original FD**		
	$\mathcal{L}^+$ / $\mathcal{L}^-$	Baseline			FDive		
	each	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
	25	**.359**	**.397**	**.410**	.268	.317	.312
	50	**.398**	**.464**	**.449**	.238	.330	.330
	75	**.326**	**.436**	**.490**	.215	.295	.328
	100	**.350**	**.407**	**.465**	.239	.321	.347
	125	**.437**	**.516**	**.494**	.250	.328	.368
	25	**.363**	**.368**	**.320**	.272	.264	.239
	50	**.399**	**.444**	**.426**	.296	.292	.279
	75	**.461**	**.533**	**.542**	.286	.309	.292
	100	**.539**	**.611**	**.567**	.306	.338	.323
	125	**.507**	**.600**	**.602**	.304	.357	.345
	25	.212	.222	.224	**.556**	**.566**	**.490**
	50	.303	.310	.306	**.561**	**.574**	**.578**
	75	.323	.351	.362	**.605**	**.619**	**.626**
	100	.473	.526	.507	**.529**	**.595**	**.586**
	125	.363	.447	.469	**.522**	**.585**	**.606**
	25	.152	.170	**.187**	**.166**	**.174**	.153
	50	.175	.157	.171	**.192**	**.216**	**.222**
	75	.180	.192	.184	**.197**	**.205**	**.202**
	100	.160	.179	.186	**.192**	**.203**	**.194**
	125	**.173**	**.186**	**.181**	.135	.142	.145
	25	**.179**	**.169**	**.173**	.096	.105	.101
	50	.162	.165	.176	**.183**	**.247**	**.253**
	75	**.197**	.201	.215	.180	**.222**	**.254**
	100	.180	.176	.191	**.186**	**.245**	**.273**
	125	**.193**	.209	.210	.180	**.235**	**.262**

Table 6.2: We compare the F_1 scores for different k-Nearest Neighbors (k-NN) classifiers. Our heuristic approach performs better for analysis targets with a higher complexity (i.e. *banana*, *crayon* and *monkey*) than state-of-art feature selection algorithms that can draw features from all available feature descriptors (4694 features). It performs worse for less complex patterns (i.e. *square* and *circle*).

available labels through the iteration cycle. To verify the validity of the similarity measure ranking, we train a k-NN classification model. We chose k-NN, because it is fully automatic and represents an intuitive classification model. We select three parameters for k, namely 1, 3, and 5. To make our results invariant to the feature selection technique, we conducted our experiments using the ReliefF algorithm, a Linear Ranking Ensemble consisting of ten Recursive Elimination SVMs, and a regular Recursive Elimination SVM. These techniques are described in Section 6.2.1. Those algorithms rank features according to their significance. We choose subsets of different lengths, namely 5, 10, 15, and 20. We perform a feature selection on the concatenation of all FDs (4694 features), resulting in recombination of different features, according to the significance assigned by the feature selection algorithm. This approach creates 12 (= 3 algorithms × 4 sizes) recombined FDs for each label and label count (i.e., table row). We determine the F_1 score of the trained k-NN for each k with all recombined FDs and all distance functions, yielding 60 (= 12 selected FD × 5 similarity coefficients) F_1 scores for each k parameter of the k-NN classifier. Table 6.2 shows the best score out of 60 for a given k in the three columns titled "Best Selected FD", serving as the benchmark. We compare this score to the single one resulting from a classification based on the best-ranked similarity measure according to the *Similarity Advisor*. All FDs are in their original state and combined with all available distance functions. The *Similarity Advisor* ranks the similarity measures based on the same label information as available to the feature selection. Table 6.2 shows the F_1 score for a given k for the best ranked similarity measure in the three rightmost columns titled "Best Ranked Original FD".

Generally, we found that our the suggested similarity measure performs on a similar level than the best feature selection created by the feature selection algorithms. It outperforms the feature selected FD in all scenarios involving the *banana* label and in 11 out of 15 scenarios pertaining to the *crayon* label. The best-ranked *Similarity Measure* is outperformed in scenarios where the analysis target is a less complex shape (i.e., *square* and *circle*. In case of the *monkey* label, our ranked FD can achieve similar result than the selected FD with 50 or more labeled instance for each of $\mathcal{L}^+$ / $\mathcal{L}^-$. Given that we compare 60 feature selection-based similarity measures to our single best ranked fixed-FD similarity measure, we can say that the similarity advisor is an efficient and effective method for the evaluation of similarity measures and that the best-ranked measure helps in the creation of a relevance model.

6.8 Discussion and Future Work

With FDive, we provide a technique which allows for the iterative learning of a relevance model, including the definition of a useful similarity measure. In the case of FDive, a similarity measure comprised of a feature descriptor and a distance function. The visual guidance of the SOM-based relevance model to uncertain classification near decision boundaries improved the understanding and quality of the model. We show that the continuous evaluation of the similarity measure benefits the iterative creation of relevance models, helping them to converge towards increasingly useful results.

One area of improvement noted by the expert was that, upon change of the similarity measure, the relevance model changes its layout, requiring the analyst to relearn it. For this reason, the mapping of different model representations into various feature spaces would allow us to explore the impact of a changed feature space on the model. Making this effect accessible would further the understanding of the feature space and underlying data distribution.

We plan to extend the general concept of the *Similarity Advisor* to other types of distance functions, removing the limitation to vector spaces implied by the L^p Minkowski family of distance measures. This extension would allow us to use other distance functions, such as Cosine, Canberra, or Clark distance. Analysts apply these measures often in specific scenarios and domains. The automatic detection of a distance function would replace the need for an expert, removing the bias introduced through the single fixed distance function. Additionally, we want to explore the application of the *Similarity Advisor* in different contexts, such as the validation of feature weightings or the design of feature descriptors based on prototypical representations of the described properties. In this instance, the *Similarity Advisor* could serve as a *concept validator*. Feature descriptors can be linked to visualization types. Through a technique similar to the *Similarity Advisor*, it should be possible to suggest other data representations, such as switching from a scatter plot representation to a parallel coordinate plot. An automatic suggestion of a useful visualization would add another step to a generalized analysis workflow, where many choices an analyst or even system designer can make is automatically assessed and supported. We layout the SOM-based relevance model in a tree structure, because it is explainable and an intuitive way of reading a classifier. Techniques introduced by Sacha et al. [265] can be used to enhance its descriptive ability. This addition can lead to novel SOM interactions focused on classification rather than exploratory cluster analysis.

We discuss scalability on two levels. First, we discuss the computational effort of *Similarity Advisor*. The main computational effort lies in the required preprocessing to transform the feature spaces and distance functions into a comparable format. The transformations are parallelizable. The complexity is determined by the dataset size. The complexity of the *Inter-Group-Distance* calculation is determined by the number of supplied labels. However, this relationship is linear. Second, we discuss the scalability limit of the complete FDive approach. The main limit approach is the creation of the SOM-based relevance model. However, the results of a previous iteration cycle can be reused in the subsequent cycles. One issue that we found was that the tree representation of the SOM-based model can become very wide. Here we have to consider a tradeoff between the size of the SOM and the associated data partitioning properties and the number of child SOMs leading to a broad tree. We found that a 3×3 SOM is an acceptable size for the SOMs since it is a size where the 2D projection property has a notable effect.

6.9 Conclusion

The extraction of interesting patterns from large high-dimensional datasets is a challenging task. With FDive, we present a workflow for the creation of relevance models based on *pattern-based similarity measures*. The system ranks similarity measures according to how well they separate relevant from irrelevant data. Our SOM-based relevance model is interactively explorable and guides the user to uncertain areas, i.e., decision boundaries. We evaluated our technique with a real-world case study in which we show that FDive can reflect the complex differences between electron microscopy images showing synapses of neurons or other brain cell structures. Our comparison to feature selection shows that FDive's *Similarity Advisor* serves as a useful metric to evaluate the discriminative ability of feature descriptor and distance function combinations. With FDive, we introduce the concept of continuous *Similarity Advisor* assessment during the learning process of a relevance model. The *Similarity Advisor* concept is applicable to areas where the user expresses his relevance for specific data items and can improve the results of the given task. The full FDive approach allows the creation of relevance models for a complex task while providing the user with valuable insights about the learning process, such as the underlying similarity measure and the model properties, including the judgment of classification results in areas of high uncertainty.

7 A Formal Framework for the Dual Analysis of Feature and Data Space

Contents

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 or categorical attributes). 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 for data exploration.

This chapter is *taken from* the following publication:

- [88] **Frederik L. Dennig**, Matthias Miller, Daniel A. Keim, and Mennatallah El-Assady. "FS/DS: A Theoretical Framework for the Dual Analysis of Feature Space and Data Space". In: *IEEE Transactions on Visualization and Computer Graphics* 30.8 (2024), pp. 5165–5182. DOI: 10.1109/TVCG.2023.3288356.

Please refer to Sections 1.2 and 1.3 for the citation rules and contribution clarification.

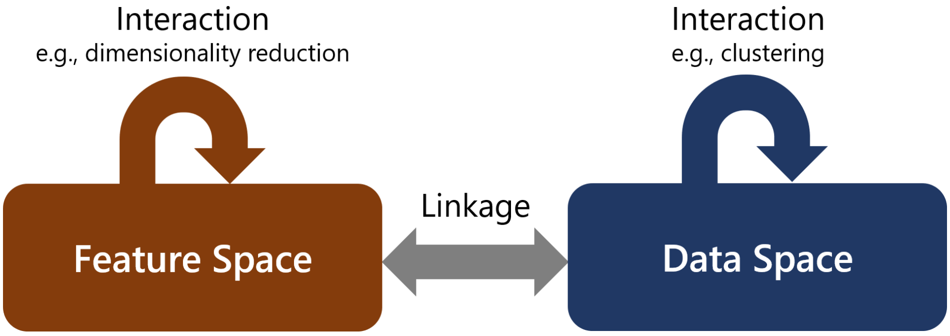


Figure 7.1: Dual analysis leverages the interactions on the feature space and data space by linking the visualizations of both spaces. Both spaces are tightly coupled, allowing for joint analysis with an immediate response.

7.1 The Need for a Generalized Model of Dual Analysis

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. [308] for the analysis of Deoxyribonucleic Acid (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) [166], 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 (see figure 7.1). In recent years, approaches solved problems in other domains, such as medicine [311, 252, 158, 119, 227, 312], crime analysis [191, 162, 117, 289], and finance [334, 309, 289]. Other approaches exchanged the visualizations for feature and data space, e.g., Parallel Coordinate Plots (PCPs) [156], and also used different interaction techniques on these visualizations, such as Drag & Drop interactions [277, 95] or subspace selection [110, 334, 191, 162, 227, 309], 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.

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 [266]. The visualization of feature and data space symmetrically leverages the preference of humans for symmetry [106]. 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 [308, 75, 95], 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 features, 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 [30] and feature selection algorithms [188, 214], layout enrichment for scatterplots [235], analytical provenance [147], and guidance mechanisms [242]. 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:

Contributions

- 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.

7.2 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.

7.2.1 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 [329], 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. [125, 124] 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. [179]. 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. [266]. 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.

7.2.2 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. [333] 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. [198] 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 [52] extends the ideas by Yi et al. [333]. 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 [235] applied the taxonomy by Brehmer and Munzner [52] to dimensionality reduction, formalizing tasks specific for dimensionality reduction.

Landesberger et al. [196] 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. [42].

Endert et al. [102] 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.

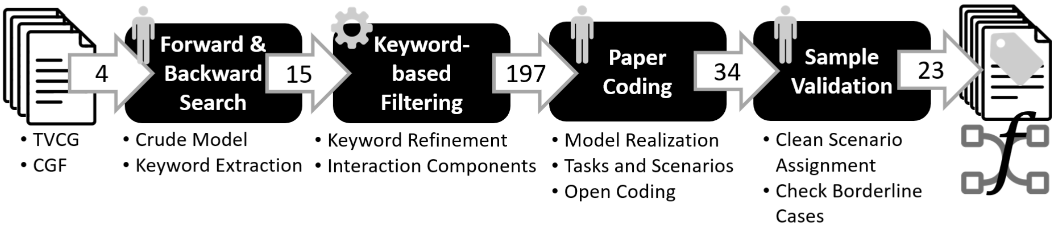


Figure 7.2: Paper Selection Process: 1.) Landmark papers 2.) Forward and Backward Search 3.) Automated Keyword-based Filtering, 4.) Paper Filtering, and 5.) Sample Validation. The numbers in the arrows describe the number of papers retained after each step.

Dimara and Perin [92] 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.

7.3 Literature Survey Methodology

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)
(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 figure 7.1). 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 figure 7.2. First, we manually identified a small subset of four publications [308, 334, 75, 95] 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 section 7.3.2). Then, we performed a detailed qualitative analysis of the selected papers, extracting and refining dual analysis characteristics, yielding a set of keywords (see section 7.3.3). 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 section 7.3.2). In general, we follow a methodology described by Snyder [283] as a systematic review to create a theoretical model or framework.

7.3.1 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 table 7.1), 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. [308]. I^F, F^I-Tables [75], SIRUS [95], and the Dimensions Projection Matrix/Tree [334]. We chose these publications since they are referenced by other publications in table 7.1 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 table 7.1. Turkay et al.'s publications [308, 311, 310, 312, 309, 313] can be viewed as fundamental to dual analysis, as they introduced the concept and established the foundation for this approach.

7.3.2 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 figure 7.2) 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.

7.3.3 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 figure 7.2). 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 figure 7.2). 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.

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. figure 7.3 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.

7.3.4 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.

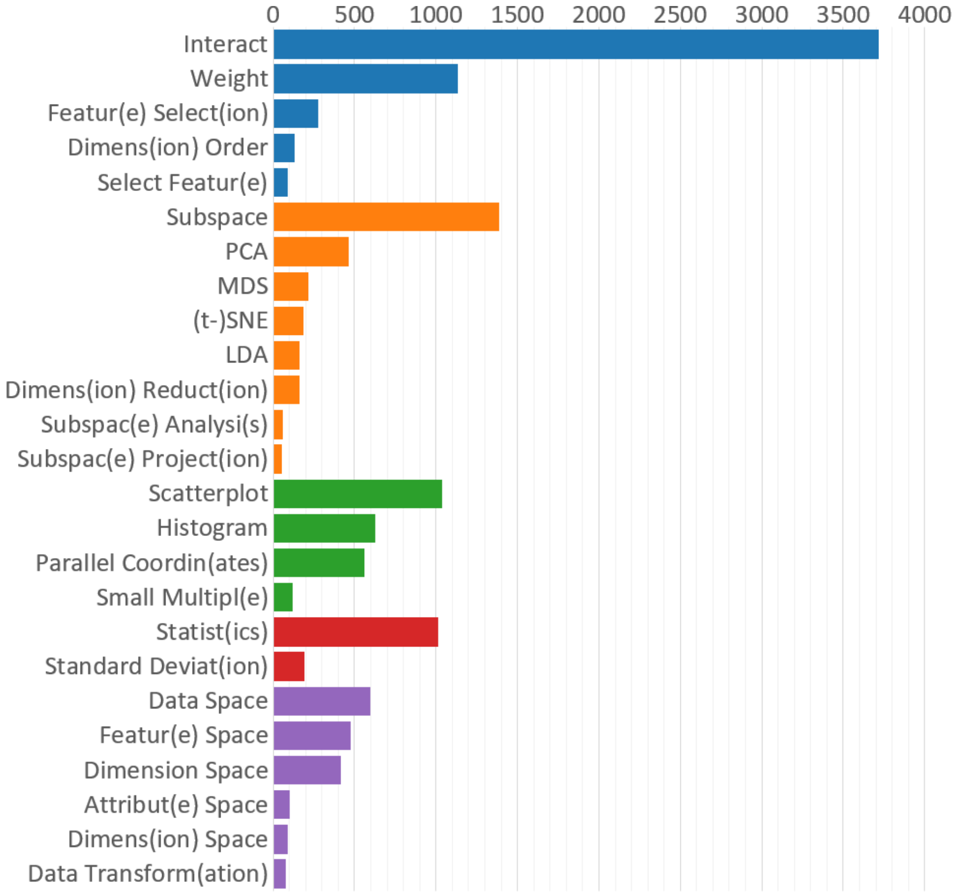


Figure 7.3: The frequency of the top 25 concepts we extracted from the 197 automatically selected papers (see section 7.3.3). Five colors contextualize each concept: ● Interaction, ● Dimensionality Reduction (DR), ● Visualization, ● Statistics, ● Analysis Space.

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 [43] 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 figure 7.1 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 section 7.4.2). 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.

7.3.5 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 section 7.3. 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. [339] presents a feature space visualization but is not linking it with the data space. The approach by Wei et al. [328] 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 [322, 321]. 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 [314, 94, 70, 327], since both views constitute a data space visualization. Our final set consists of 23 relevant publications, which we present in table 7.1. We transformed the table into a set of feature vectors to present similarities (see figure 7.4). We cleaned the encoding and grouped the identified approaches into high-level scenarios (see section 7.4.2).

Paper	Year	Feature Space					Data Space						Scenarios							Domain
		Visual.			Transf.		Visual.				Transf.									
		SP	SM	Other	Lossy	Lossless	SP	PCP	Map	Other	Lossy	Lossless	S1	S2	S3	S4	S5	S6	S7	
[308] Turkay et al. (1)*	2011	●			μ, σ		●				PCA				●					🧬
[311] Turkay et al. (2)	2012	●			7		●				MDS		●	●				●		🩺
[334] Yuan et al.*	2013	●			MDS		●				MDS						●			⚛ 🍎
[110] Fernstad et al.	2013			PCP	5		●	●			PCA				●				●	🦠
[310] Turkay et al. (3)	2014		●			Ord			●			Id			●					🔗
[312] Turkay et al. (4)	2014	●			6		●				$\Delta_\mu, \Delta_\sigma$				●		●			🩺 🧬
[191] Krause et al.	2016		●			Ord	●	●			PCA		●	●					●	👮 🍎
[75] Corput et al.*	2016			DT		Ord				DT		Ord				●		●		🔗
[336] Zanabria et al.	2016			SC		Ord	●				iStar			●				●		🏃
[277] Self et al.	2016			LG		Ord	●				MDS					●				🦠
[158] Itoh et al.	2017			GR	MDS			●				Ord	●						●	🩺 ⚙
[313] Turkay et al. (5)	2017	●			MDS				●			Id			●		●			🔗
[309] Turkay et al. (6)	2017			HG		Sel	●		●		PCA				●		●			🦠 🔄
[162] Jentner et al.	2018	●			MDS		●				3		●					●		👮
[252] Rauber et al.	2019	●			MDS		●				t-SNE		●			●			●	🩺 🦠
[95] Dowling et al.*	2019	●			MDS		●				MDS			●				●		👮 🦠
[16] Artur and Minghim	2019	●			RAD		●				RAD			●				●		🩺
[340] Zhao et al.	2019			HG		Ord	●					Sel	●	●					●	🦠
[117] Fujiwara et al.	2020			PIX		Id	●				PCA		●					●		👮 🍎
[289] Soriano-Vargas et al.	2020		●		Clu		●				IDMAP				●	●		●		👮 ⚛ 🔄
[119] Garrison et al.	2021			PCP	FAMD			●				Sel	●					●		🩺
[227] Müller et al.	2021	●			μ, σ			●				Sel			●					🩺
[221] 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	

Table 7.1: 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 section 7.4.1. 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 section 7.4.2. 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*.

7.4 Existing Dual Analysis Approaches

This section covers all dual analysis approaches, which we selected following the definition and criteria we described in section 7.3. All approaches are listed in table 7.1, 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 [235]. We proposed seven descriptive scenarios in section 7.4.2 to categorize different tasks for dual analysis structure along the three questions *Why*, *What*, and *How* proposed by Brehmer and Munzner [52]. 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 [267]. We also list the evaluation and application domain to give an overview of the addressed areas.

7.4.1 Visualizations and Transformations

We categorize all dual analysis approaches by their individual representations of feature space and data space (see table 7.1). 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 [308, 311, 334, 312, 313, 162, 252, 95, 16, 227]. Most approaches encode information by using the visual variables color and size [40] 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 Multidimensional Scaling (MDS) [194], 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 [310, 191, 117]. 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 [156] visualize data by plotting a polyline crossing parallel coordinate axis [110, 119]. Zanabria et al. [336] use *Star Coordinates* SC [172] to visualize features. Corput et al. [75] 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 [277]. A *Graph* GRA visualizes a network with a node link-diagram. Itoh et al. [158] visualize dimensions and their relations using a graph. *Histograms* HG [309, 340] are used to display statistical analysis results [309] and results of features selection. Miller et al. [221] use a *Pixel visualization* PIX [177] 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 DR techniques. Seven approaches [334, 110, 313, 162, 252, 95] use the *Multidimensional Scaling* MDS technique [194], 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 [47]. For the feature space, this is often a measure of correlation [95, 110]. A particular case is Weighted Multidimensional Scaling (WMDS), which allows the weighting of individual features and the estimation of the weight of features according to their position in the reduced space [95]. *RadViz* RAD [236] 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 [16]. 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 [308] of all values of a feature. The approaches by Garrison et al. [119] and Müller et al. [227] deal with mixed data and, thus, employ statistical measures for categorical data, like *factor analysis for mixed data* FAMD [238] and the *coefficient of unalikeability* and a definition for standard deviation thereof μ, σ [227]. Three approaches use more than five values, i.e., *mean*, *median*, *standard deviation*, *variance*, *skewness*, and *kurtosis* N in table 7.1 shows the number of measures) [311, 110, 312]. The approach by Sariano-Vargas et al. [289] 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 [310, 191, 75, 336, 277, 340, 221]. No reduction or change to the data is marked as *identity* Id [117], 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 [313], 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 DR techniques [308, 311, 334, 110, 191, 75, 336, 277, 309, 162, 252, 95, 16, 340, 117, 289, 221]. Another way of visualizing the data space is *Parallel Coordinates Plots* PCP, which are only used in four approaches to represent the data space [110, 191, 158]. PCPs are used as an auxiliary view to show the dataset. The approach by Itho et al. [158] uses PCPs to select subspaces manually. Three approaches by Turkay at al. use a glyph and *geographical Map* MAP [312, 309, 313] combination, which deal with social and census data. In the case of Corput et al. [75], 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 [166] is used six approaches [308, 110, 191, 309, 162, 117]. Six approaches [311, 334, 277, 162, 95, 221] use *Multidimensional Scaling* MDS [194]. The *t-Distributed Stochastic Neighborhood Embedding* t-SNE [208] is employed twice [252, 162], including the approach by Jentner et al., which allows the user to choose between PCA, MDS, and t-SNE denoted by 3. Another approach for dimensionality reduction is *RadViz* RAD [236], which we already described as a feature space transformation. It is used by Artur and Minghim [16] to create a symmetric dual analysis approach for aggregating features and data items. The iStar [336] 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) [312] 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. [221] applies a DBSCAN clustering [105] 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.

7.4.2 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 table 7.1). These scenarios are also linked to our formal framework (see figure 7.5), 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 [52].

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., by a category of an attribute (i.e., categorical feature) or interactively using a Lasso selector. The selected features are then available for further analysis. This scenario

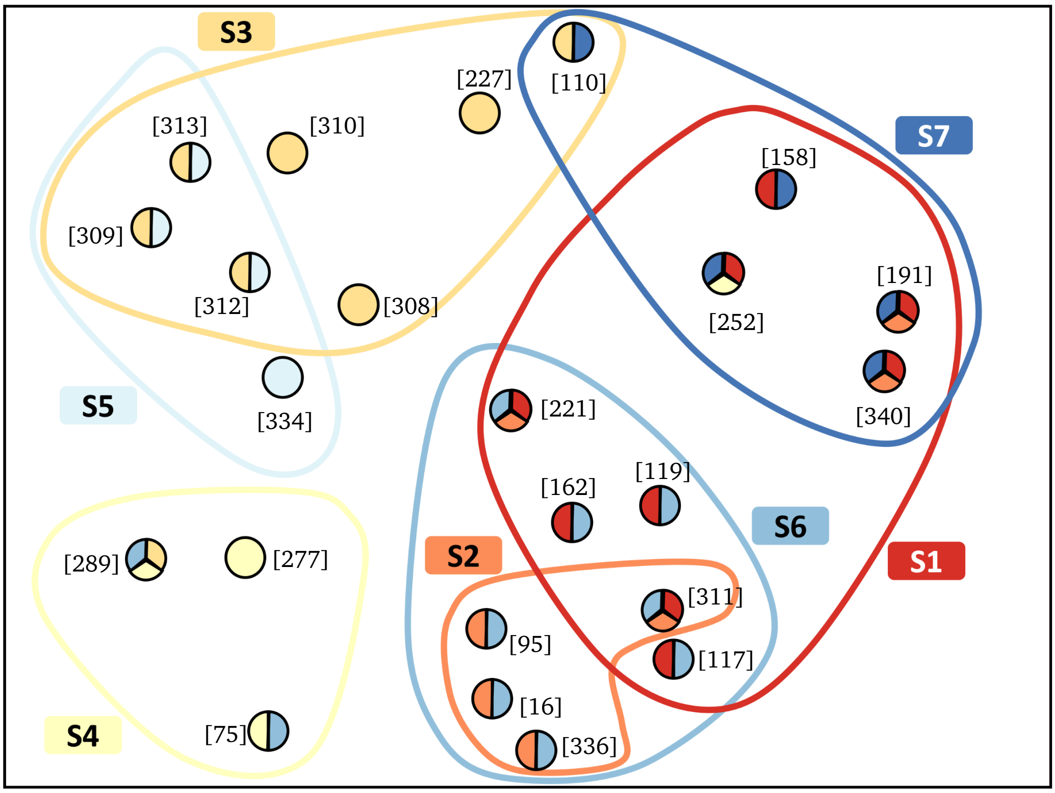


Figure 7.4: Similarity-based projection of the 23 papers in table 7.1. The similarity is defined by one-hot encoding the columns of table 7.1, excluding name, year, and domain using the Manhattan distance to create an MDS projection. We weight the scenarios three times higher, yielding a scenario-based grouping. Glyphs are colored according to their scenarios (see section 7.4.2) and grouped showing the relation between them.

never occurs alone since it would only correspond to changes in the data space visualization. A common partner is S7 *Data Selection* [191, 158, 252, 16].

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. [162], 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) [311] 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 [110]. 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 [311]. Dual analysis also supports the creation and validation of classifiers [252]. 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. [95] 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) [162]. 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 [95]. 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) [308]. 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. [227], which analyzes variance and attribute variability using Factor Analysis of Mixed Data (FAMD). 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) [308], 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. [75], 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 [277]. 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 [192]. 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. [334] 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. [162] allows for exploring subspace clusters, specifically enabling analysts to understand cluster characteristics, develop alternative clusterings and verify cluster robustness. Turkay et al. (4) [312] 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 [334].

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 [310, 191, 252]. 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 or by selecting a category of an attribute (i.e., a categorical feature) in the feature space. The only data manipulation process we found in the set of works is labeling data items with a classification algorithm [252]. 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 table 7.1 (see figure 7.4). We transformed the entries of table 7.1 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. [334]. 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.

7.4.3 Application and Evaluation Domains

Dual analysis has found application in many domains, most notably in *Medicine* (🩺), where we found seven approaches [311, 312, 158, 252, 16, 119, 227], 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* (🦠) [110, 277, 309, 252, 95, 340] and *Genomics* (🧬) [308, 312], where we found seven approaches combined. *Crime Analysis* (👮) with five approaches [191, 162, 95, 117, 289], 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*, (👥) [310, 75, 313] analyzing different aspects of society, such as the comparison of households in different geographic regions. Three publications address the analysis of *Nutrition* (🍎)[334, 191, 117], by analyzing the nutritional contents of food items. Two papers deal with problems in *Finance* (💱) [309, 289]. *Physics and Chemistry* (⚛) [334, 289], *Engineering* (⚙) [158], *Sports* [336] (🏃), and *Musicology* (🎵) [221] are each addressed once.

7.5 Theory and Formalization

Our formalization encompasses all previous work (see table 7.1) 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. [75]. 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

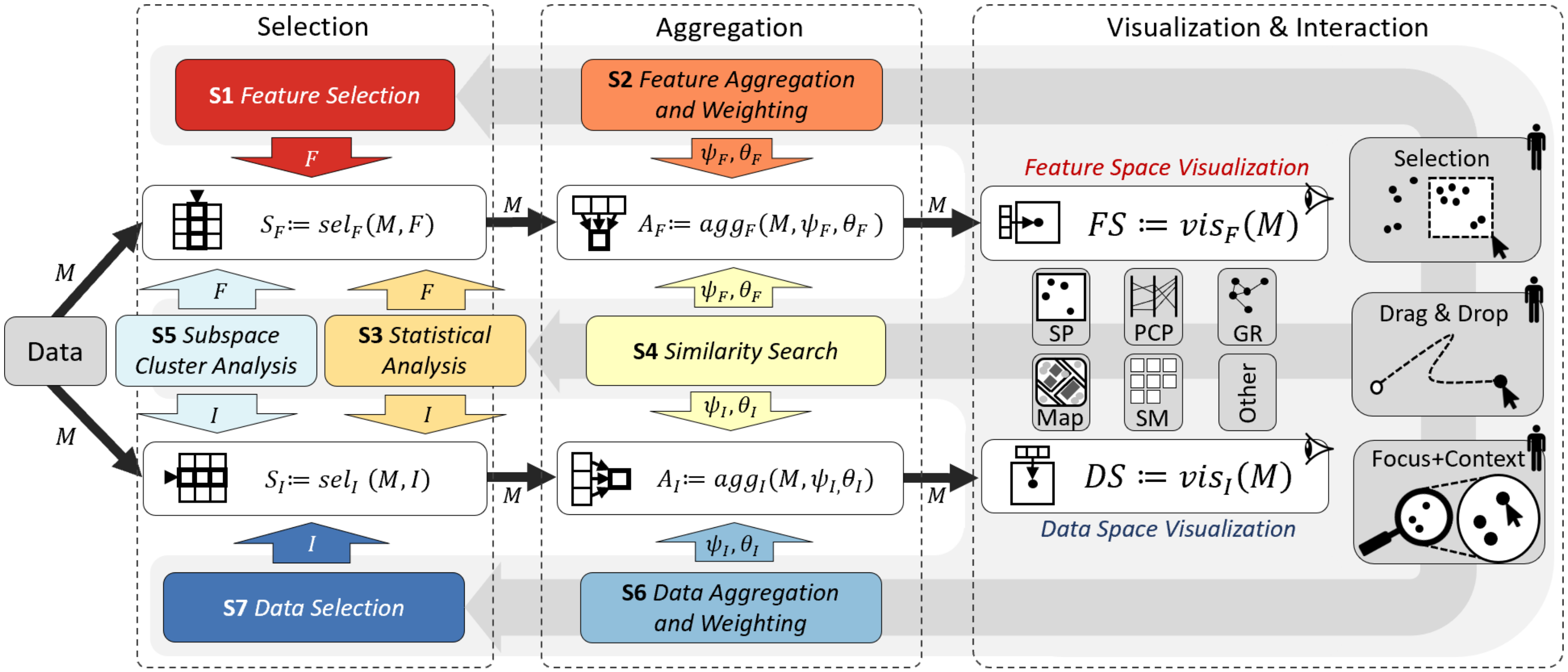


Figure 7.5: In our framework for dual analysis, the dataset $\mathcal{D}$, is interpreted as a matrix. The matrix can then be transformed by a selection step, where the data can be reduced. Secondly, the result of this step is used for an aggregation step, which can be used to create representatives. 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 (equation (7.7)), agg_I (equation (7.8)), vis_I (equation (7.9)), for data space operations, and sel_F (equation (7.1)), agg_F (equation (7.2)), vis_F (equation (7.3)). 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 section 7.5.4.

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.

7.5.1 Feature and Data Types

Dual analysis has been applied to quantitative and qualitative variables, i.e., mixed data [227, 119]. Thus, our formalization has to describe data analysis for all common features and data types, e.g., numeric and categorical data [228]. 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.

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 [119, 227].

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 [57, 134].

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 [57]. This data type is common in social science [313, 309].

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).

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 equation (7.1).

$$sel_F : (M, F) \rightarrow \mathbb{R}^{r \times |F|} \tag{7.1}$$

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 (i.e, columns) to aggregate, and θ_F defines the aggregation and reduction which we formalize in equation (7.2).

$$\begin{gathered} agg_F : (M, \psi_F, \theta_F) \to \mathbb{R}^{d \times |\psi_F|} \\ \text{where } M \in \mathbb{R}^{m \times n},\ \psi_F \text{ a partition of } [\![1..n]\!] \\ \text{with } e \in \psi_F \text{ a set of column indices of } M, \\ \text{and } \theta_F : \mathbb{R}^{m \times |e|} \to \mathbb{R}^d \text{ with } e \in \psi_F \text{ and } d \in \mathbb{N} \end{gathered} \tag{7.2}$$

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 [95, 110]. 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 equation (7.3). For example, to detect large groups of features in S5 *Subspace Analysis* Yuan et al. [334] use scatterplots, while for a more fine-grained analysis of relatedness between a few features Garrison et al. [119]

use parallel coordinate plots. To describe the visualization of the feature space, we define vis_F in equation (7.3).

$$vis_F : M \rightarrow FS \tag{7.3}$$

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 equation (7.4).

$$glyph_F : \mathbb{R}^m \rightarrow G_F \tag{7.4}$$

One example of G_F is a pixel-based visualization [289]. The position of the glyph is determined in equation (7.5).

$$pos_F : \mathbb{R}^m \rightarrow (x, y) \in \mathbb{R}^2 \tag{7.5}$$

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 equation (7.6), which gives a complete definition of the feature space scatterplot.

$$\begin{gathered} vis_F := \forall i \in [\![1..n]\!]. \\ \rho_i = (glyph_F(M_{*,i}), pos_F(M_{*,i})) \text{ with } M \in \mathbb{R}^{m \times n} \end{gathered} \tag{7.6}$$

Most approaches that use a scatterplot to visualize the feature space relying on a dimensionality reduction method utilize MDS or variations thereof (see table 7.1). 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.

7.5.3 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 equation (7.7), yielding a reduced data set or, ultimately, a smaller matrix by reducing the number of rows.

$$sel_I : (M, I) \rightarrow \mathbb{R}^{|I| \times f} \tag{7.7}$$

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. Common techniques are linking & brushing [324] or selecting a category of an attribute that defines a subset of the dataset. However, other methods are possible, such as the selection of data items based on class labels, cluster affiliation, filtering, sampling [132, 4] or grouping instances [198, 2]. 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 equation (7.8)). 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 equation (7.8).

$$\begin{gathered} agg_I : (M, \psi_I, \theta_I) \rightarrow \mathbb{R}^{|\psi_I| \times d} \\ \text{where } M \in \mathbb{R}^{m \times n},\ \psi_I \text{ a partition of } [\![1..m]\!] \\ \text{with } e \in \psi_I \text{ a set of row indices of } M, \\ \text{and } \theta_I : \mathbb{R}^{|e| \times n} \rightarrow \mathbb{R}^d \text{ with } e \in \psi_I \text{ and } d \in \mathbb{N} \end{gathered} \tag{7.8}$$

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. ψ_I can also be determined by selecting data items that share common properties, such as one or more categories in their attributes (i.e., categorical features). 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 equation (7.9).

$$vis_I : M \rightarrow DS \tag{7.9}$$

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. [116]. 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 equation (7.10)).

$$glyph_I : \mathbb{R}^n \rightarrow G_I \tag{7.10}$$

Second, we also define a function to determine the position of the glyph in the scatterplot in equation (7.11).

$$pos_I : \mathbb{R}^n \rightarrow (x, y) \in \mathbb{R}^2 \tag{7.11}$$

Thus, with these two functions, we can cover the scatterplot-based visualization of the data space in equation (7.12), 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.

$$\begin{gathered} vis_I := \forall i \in [\![1..n]\!]. \\ \rho_i = (glyph_I(M_{i,*}, pos_I(M_{i,*})) \text{ with } M \in \mathbb{R}^{m \times n} \end{gathered} \tag{7.12}$$

To determine a position (see equation (7.11)), 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 equation (7.11). We found that PCA [166], MDS [194], t-Distributed Stochastic Neighbor Embedding (t-SNE) [208], or Interactive Document

Map (IDMAP) [222] 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.

7.5.4 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 [228, 112]. 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 [227]. The interaction of feature and data space constitutes a form of Linking & Brushing [308, 334] 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 [102]. 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 [277, 95]. 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* [75] reflecting the different goals.

Focus+Context: Another concept in the dual analysis is Focus+Context [62]. 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. [308] 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.

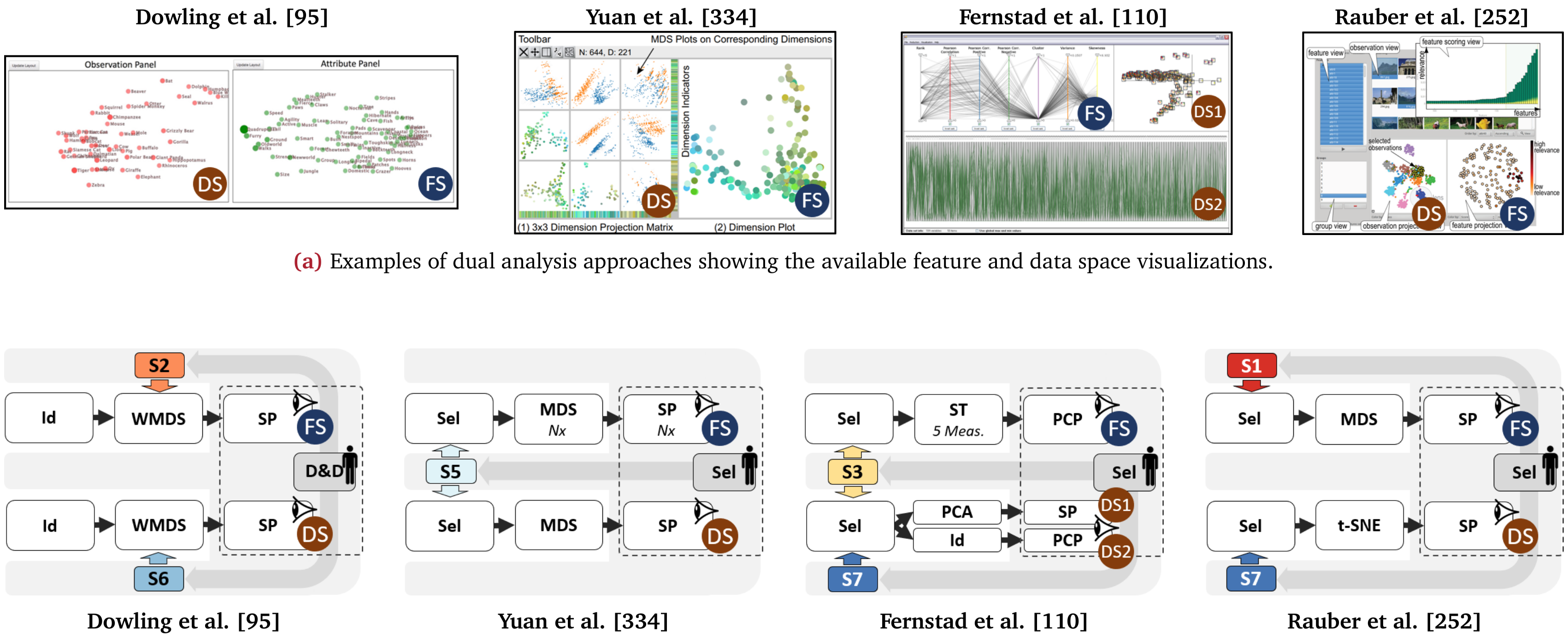
7.6 Evaluation

To evaluate our approach, we apply an evaluation strategy inspired by Sacha et al. [267]. 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 [95, 334, 110, 252] to be representative of the set of papers described in table 7.1 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).

7.6.1 Descriptive Use: Examples

In this section, we describe four representative approaches.

Dowling et al.: The system by Dowling et al. [95] 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



(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.

Figure 7.6: We describe four approaches in section 7.6.1 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. [95], Yuan et al. [334], Fernstad et al. [110], and Rauber et al. [252] as a representative set.

feature or data subsets, except the dataset is pre-processed. Our model 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.: [334] 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. [110] 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. [110] 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. [252] 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.

7.6.2 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 [116]. 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 figure 7.5), which we already extend by defining specific functions for glyph-based visualizations $glyph_I$ and $glyph_F$ (see equation (7.10) and equation (7.4)).

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 [235]. Since dual analysis approaches make extensive use of dimensionality reduction and scatterplot visualizations, even manipulating parameters of the dimensionality reduction [277, 95], we see a clear need for additional visual feedback. An example of this idea is uPCA [123] and uMDS [131], 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 equation (7.9) and equation (7.3)).

Subspace Detection Algorithms: Four approaches we found during our review mainly address the analysis of subspaces and subspace clusters [334, 309, 312, 309]. 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 [192]. In particular, SURFING [30], SUBCLU [170], and RIS [171]. 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 figure 7.5). 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 figure 7.5) at each step of the dual analysis pipeline allows for a nuanced tracking of the analysis state. Steinparz et al. [295] and Hinterreiter et al. [147] 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 figure 7.5) with guidance scenarios to find interesting ways to help analysts in their analysis tasks through guidance[242]. Practical guidance frameworks such as Lotse by Sperrle et al. [293] 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.

7.7 Discussion and Future Work

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. [75] 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. [308] 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 figure 7.1). 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 section 7.5, 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 table 7.1. 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 [84]. The framework by Bertini et al. [41], later extended by Behrisch et al. [33], describes an enriched VA pipeline with quality-measure-driven automation. Quality can be measured at each analysis step (i.e., upon a view update) while the analyst steers the process. Quality measures can aid user interactions with automatic configurations or recommendations at each step. However, quality measures 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 [252, 89]. 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 derived from the existing literature, while focusing on incorporating visual analysis methods for categorical data, which is currently only addressed by two approaches based on mixed data. 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.

7.8 Conclusion

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.

Conclusion

8

Contents

This thesis explored the fields of Visual Analytics (VA) and Information Visualization (InfoVis), focusing on the analysis of categorical data in multiple measure-driven approaches and frameworks. In this final chapter, we look back at our contributions to the field and their broader implications. By reflecting on our work, we aim to summarize the advances we've made, while positioning and contextualizing them within the larger framework of VA. We also highlight potential avenues for future research and offer perspectives on methods for analyzing categorical data using measure-driven approaches.

8.1 Summary

This dissertation explored the power of VA to facilitate analytical reasoning through interactive visual interfaces, specifically addressing the complexities associated with analyzing categorical data. VA combines the innate strengths of humans, such as domain expertise and cognitive pattern recognition, with the computational power of computers. This integration is especially critical in the context of big data analytics and addressing complex societal challenges. Societal challenges such as public health crises, environmental sustainability, and urban planning often involve complex systems with interdependent variables and vast amounts of data. Visual analytics plays a critical role in unraveling this complexity by enabling the visualization of social phenomena in an accessible and compelling way. In public health, for example, VA can be used to track the spread of disease in real time, identify risk factors, and optimize resource allocation [163]. The ability to integrate and visualize data from diverse sources-social media, satellite imagery, census data-allows policymakers, researchers, and the public to engage directly with the data, fostering a collaborative approach to problem solving. In addition, by highlighting trends, disparities, and outcomes, VA supports the development of targeted interventions and policies that address the root causes of social problems and promote equity and sustainability.

In this context, *categorical data*, including nominal attributes with no inherent order or measurable distance, presented significant challenges to traditional data mining and visualization techniques tailored for numerical data. These challenges, which arise from the arbitrary order of attributes and categories, hinder the application of traditional analysis methods. Yet categorical data analysis is essential in fields as diverse as business intelligence, financial risk assessment, software engineering, and linguistics. Aiming to bridge the gap between qualitative and quantitative analysis, this research introduced novel approaches for categorical data. These approaches improved the quality of visualizations for categorical data, provided methods for applying numerical analysis techniques, and facilitated the study of their interaction with numerical data dimensions.

The *contributions* of this dissertation are as follows: The first part focused on quality improvement and pattern quantification in categorical data visualizations. In chapter 2 we introduced quality measures for Parallel Sets visualizations. We were able to define measures for properties such as ribbon overlap and crossings, which create visual clutter and reduce readability. We also found that published visualizations of Parallel Sets could be improved by optimizing their properties along our measures. In chapter 3, we were able to apply projection methods to numerical data by abstracting from categorical data. In addition, we contributed two measures to support orientation, navigation, and exploration by quantifying the property of fracturedness.

The second part explored measure-driven methods for articulating the properties of real-world data and deriving numerical measures for categorical entities, with a focus on linguistics and software engineering. More specifically, in chapter 4 we presented an approach to studying language change. Where we have been able to help linguists to find periods of relevant change with overviews that show drift in language structure as expressed by categorical attributes. In chapter 5, we provided an approach for investigating the risk that vulnerable open source software poses to software products from large development organizations such as SAP. By aggregating information from a set of categorical properties, we were able to find several serious vulnerabilities affecting the Eclipse Foundation's open source projects.

The third part broadened the scope to include the integration of categorical data into supervised and unsupervised analysis frameworks, suggesting effective strategies for model development and exploratory analysis. This part also explored the interactions between categorical and numerical data to provide a comprehensive view of integrating categorical data into VA workflows. In chapter 6, we introduced a supervised process for developing models that use categorical data to derive a descriptive abstraction of data item properties. In addition, our classification model based on Self-Organizing Maps (SOMs) provides insight into the learning mechanism, highlighting measures that are similar across the dataset and pinpointing areas where classification uncertainty is particularly high. In chapter 7, we conducted a systematic review of the visual analytics literature to identify and classify methods that employ dual analysis, a method for simultaneously examining feature and data space. We found that most approaches only address the analysis of numerical data. We have included categorical data in our framework, thereby extending its applicability to interactive data exploration tasks and addressing unsupervised techniques such as clustering algorithms. Our findings were presented through seven descriptive scenarios, contextualized within a formally established framework for dual analysis.

8.2 Future Research

Throughout this thesis, the concluding sections of each chapter have already presented challenges and suggested directions for future research, focusing primarily on the techniques and methods explored in each respective chapter. Throughout this body of work, specific recurring themes and broad challenges emerge that are relevant to the entire field of study. In the following paragraphs, we will highlight the overarching research challenges and the potential of measure-driven approaches in the visual analysis of categorical data.

Local and Global Measures for Quality and Patterns in Categorical Data Visualizations: There is an ongoing need to measure the quality and patterns in categorical data visualizations. In general, quantifying the quality of visualizations establishes objective evaluation criteria that are essential for systematically assessing the effectiveness of different visualization techniques [41]. This enables a standardized approach to comparing visualizations, ensuring that they accurately convey the intended data insights. With Parsetgnostics, we addressed one of many categorical data visualizations. Other categorical data visualizations, such as Sankey diagrams, could be improved using measures derived from Parsetgnostics. Mosaic plots [148, 149], however, follow a different paradigm and require new, tailored measures for quality qualification. Additionally, Parsetgnotics provides a set of *global* measures that consider the entire visualization to quantify quality in visualizations. However, measures that provide specific feedback on visual quality and patterns *locally*, i.e., within a particular visualization area, could guide designers to make targeted improvements and alleviate problems by pinpointing areas where the visualization may lack clarity or fail to effectively represent data, leading to more impactful visualizations.

Improved Methods for Automatic Optimization of Categorical Data Visualizations: By automating the visualization design process, non-expert users gain access to high-quality visualizations without requiring deep knowledge of visualization techniques. However, visualization optimization tools require robust algorithms capable of generating or suggesting optimal visualizations. These algorithms guide these tools in selecting the most appropriate visualization types and configurations, tailored to the underlying data and analysis objectives [33]. However, using measures and brute force optimization is not feasible because there can be more than an exponential number of visualization configurations, i.e., as in the case of Parallel Sets factorial many. Thus, for purely automatic optimization, it is crucial to develop algorithms that optimize a specific type of visualization. Parallel Sets are layered graphs for which efficient algorithms exist to reduce edge crossings [29]. This allows Parallel Sets to be optimized without having to try every possible combination. While Parallel Sets are just one example, the same principle applies to other categorical data visualizations. A fast, automated optimization process reduces the time required to produce effective visualizations. This speed is essential in dynamic environments where timely data analysis and decision making are critical.

Guiding Categorical and Mixed Data Analysis using Quantitative Measures: Quantitative measures can guide exploratory data analysis by highlighting areas of interest or concern within the visualization. They help analysts prioritize their areas of focus by directing attention to a portion of a dataset or visualization that exhibits a quantitative property or pattern [242, 293]. Unlike using measures to design a visualization, they can reveal subtle patterns and relationships within the data that may not be immediately apparent through manual analysis and inspection alone. In this thesis, we contributed measures to guide the analysis of categorical data by providing a quantification for the degree to which an attribute differentiates groups of observations in a domain-agnostic approach. However, this approach is limited to this type of visualization and also only captures a property of a single attribute. Therefore, extending to other visualizations while supporting multiple attributes is a worthwhile goal by defining new measures and extending the visualization to support multiple attributes. This is a viable approach, as we have also explored domain-specific approaches, using measures for multiple categorical attributes to guide users to relevant time periods, in the case of HistoBankVis, and guide to high-impact software vulnerabilities with VulnEx.

Adapting Numerical Data Analysis Methods and Visualizations to Categorical Data: In this dissertation, we presented an approach that specifically adapts methods for numerical data analysis to categorical data. By adapting numerical methods for categorical data, analysts gain access to a broader set of analytical tools that enable more sophisticated and nuanced examination of datasets that contain qualitative information [261]. However, work in this area remains limited due to challenges in processing categorical data to be compatible with existing techniques. Common methods like One-Hot Encoding (OHE) can create artifacts in the analysis and visualization if not handled correctly [48, 57]. Therefore, one research direction we propose is to systematically explore the limitations of numerical analysis methods and to determine what kind of biases occur under which conditions. As we have shown in this thesis, aggregation can be an effective basis to transform categorical data into a numerical representation. This adaptation enhances the scope and depth of data analysis, leading to richer insights and more effective decision-making across various domains.

8.3 Closing

In this thesis, we addressed the challenges in the field of VA posed by categorical data, focusing in particular on ways to bridge the qualitative-quantitative divide through measure-driven methods, summarized by the overarching research question (R0, see page 4). Our work has been guided by three sub-questions that have guided our exploration and contributions to the field to more precisely capture the esence of (R0). By tackling (R1), we have successfully developed and implemented a set of measures designed to evaluate and optimize Parallel Sets visualizations for categorical data. These measures not only allow for a more objective assessment of visualization quality, but also provide a basis for creating optimized Parallel Sets of visualizations. We also contributed a visualization method for exploring categorical data in a "map metaphor" made possible by transforming categorical data into numerical representations and quantifying patterns in the categorical data projection. In addition, exploration is aided by measures that quantify visualization properties, guide users in analyzing attributes, and differentiate groups of observations. By addressing (R2), our research has illustrated the practical application and utility of these measure-driven methodologies in the fields of linguistics and software engineering. By applying the measures we have developed for real-world problems and datasets, we have demonstrated their ability to systematically uncover insights and patterns that traditional analytical methods might miss. This not only validates the effectiveness of our approach, but also highlights the versatility and potential of our approaches to contribute to domain-specific challenges, foster a deeper understanding of complex datasets, and improve the decision-making process. Finally, in tackling (R3), we have introduced novel frameworks that integrate these measures into supervised and unsupervised learning contexts. These frameworks improve the classification and exploration of mixed data by leveraging categorical data and providing interfaces for custom measures that improve model accuracy, interpretability, and exploratory analysis capabilities. In conclusion, this thesis contributes to the advancement of visual analytics by addressing the unique challenges posed by categorical data through the development of measure-driven methodologies and frameworks. Our work not only enhances the analytical capabilities available to researchers and practitioners, but also opens new avenues for future research, promising to increase the usefulness of categorical data and improve our ability to communicate complex insights effectively.

List of Figures

List of Tables

List of Acronyms

ACM	Association for Computing Machinery
AL	Active Learning
CA	Correspondence Analysis
CE	Common Era
CNN	Convolutional Neural Network
CVE	Common Vulnerabilities and Exposure
CVSS	Common Vulnerability Scoring System
DaRUS	Data Repository of the University of Stuttgart
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DevSecOps	Development, Security, and Operations
DNA	Deoxyribonucleic Acid
DR	Dimensionality Reduction
EM	Electron Microscopy
FAMD	Factor Analysis of Mixed Data
FD	Feature Descriptor
FS	Feature Space
FV	Feature Vector
HCI	Human-Computer Interaction
IcePaHC	Icelandic Parsed Historical Corpus
IDMAP	Interactive Document Map
IEEE	Institute of Electrical and Electronics Engineers
IML	Interactive Machine Learning
InfoVis	Information Visualization
KDD	Knowledge Discovery in Databases
k-NN	k-Nearest Neighbors
LGTM	Looks Good To Me
MCA	Multiple Correspondence Analysis
MDS	Multidimensional Scaling
MFA	Multiple Factor Analysis
OHE	One-Hot Encoding
OSF	Open Science Foundation
OSS	Open Source Software
PCA	Principal Component Analysis

PCP	Parallel Coordinate Plot
QE	Quantization Error
SDLC	Software Development Lifecycle
SOM	Self-Organizing Map
SVM	Support Vector Machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VA	Visual Analytics
WMDS	Weighted Multidimensional Scaling

Bibliography

[1] Mostafa M. Abbas, Michaël Aupetit, Michael Sedlmair, and Halima Bensmail. "ClustMe: A Visual Quality Measure for Ranking Monochrome Scatterplots based on Cluster Patterns". In: *Computer Graphics Forum* 38.3 (2019), pp. 225–236. DOI: 10.1111/CGF.13684 (cit. on p. 52).

[2] Ala Abuthawabeh and Michaël Aupetit. "Toward an Interactive Voronoi Treemap for Manual Arrangement and Grouping". In: *Proceedings of the 21st Eurographics Conference on Visualization*. Eurographics, 2021, pp. 97–101. DOI: 10.2312/evs.20211062 (cit. on p. 160).

[3] Shivam Agarwal and Fabian Beck. "Set Streams: Visual Exploration of Dynamic Overlapping Sets". In: *Computer Graphics Forum* 39.3 (2020), pp. 383–391. DOI: 10.1111/cgf.13988 (cit. on p. 24).

[4] Charu C. Aggarwal and Chandan K. Reddy, eds. *Data Clustering: Algorithms and Applications*. CRC Press, 2014. ISBN: 978-1-46-655821-2 (cit. on p. 160).

[5] Alan Agresti. *An Introduction to Categorical Data Analysis*. 3rd. Wiley. ISBN: 978-1-119-40528-3 (cit. on pp. 1, 2, 48).

[6] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. 2nd. Human-Computer Interaction Series. Springer, 2023. ISBN: 978-1-4471-7526-1. DOI: 10.1007/978-1-4471-7527-8 (cit. on p. 1).

[7] Shinji Akatsu, Ayako Masuda, Tsuyoshi Shida, and Kazuhiko Tsuda. "A study of quality prediction for large-scale open source software projects". In: *Journal of Artificial Intelligence Research* 10.1 (2021), pp. 34–42. DOI: 10.5430/air.v10n1p34 (cit. on p. 96).

[8] Kenneth B. Alperin, Allan B. Wollaber, and Steven R. Gomez. "Improving Interpretability for Cyber Vulnerability Assessment Using Focus and Context Visualizations". In: *Proceedings of the 17th IEEE Symposium on Visualization for Cyber Security*. IEEE, 2020, pp. 30–39. DOI: 10.1109/VizSec51108.2020.00011 (cit. on p. 98).

[9] Jamal Alsakran, Xiaoke Huang, Ye Zhao, Jing Yang, and Karl Fast. "Using Entropy-Related Measures in Categorical Data Visualization". In: *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2014, pp. 81–88. DOI: 10.1109/PacificVis.2014.43 (cit. on pp. 23, 25, 37, 63).

[10] Bilal Alsallakh, Luana Micallef, Wolfgang Aigner, Helwig Hauser, Silvia Miksch, and Peter J. Rodgers. "The State-of-the-Art of Set Visualization". In: *Computer Graphics Forum* 35.1 (2016), pp. 234–260. DOI: 10.1111/cgf.12722 (cit. on pp. 48, 50).

[11] Marco Angelini, Graziano Blasilli, Luca Borzacchiello, Emilio Coppa, Daniele Cono D'Elia, Camil Demetrescu, Simone Lenti, Simone Nicchi, and Giuseppe Santucci. "SymNav: Visually Assisting Symbolic Execution". In: *Proceedings of the 16th IEEE Symposium on Visualization for Cyber Security*. IEEE, 2019, pp. 1–11. DOI: 10.1109/VizSec48167.2019.9161524 (cit. on p. 98).

[12] Marco Angelini, Graziano Blasilli, Tiziana Catarci, Simone Lenti, and Giuseppe Santucci. "Vulnus: Visual Vulnerability Analysis for Network Security". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 183–192. DOI: 10.1109/TVCG.2018.2865028 (cit. on p. 97).

[13] Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. "Towards an effective cooperation of the user and the computer for classification". In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2000, pp. 179–188. DOI: 10.1145/347090.347124 (cit. on p. 114).

[14] Nancy L. Leech Anthony J. Onwuegbuzie John R. Slate and Kathleen M. T. Collins. "Mixed data analysis: Advanced integration techniques". In: *International Journal of Multiple Research Approaches* 3.1 (2009), pp. 13–33. DOI: 10.5172/mra.455.3.1.13 (cit. on p. 3).

[15] Dustin Arendt, Emily Saldanha, Ryan Wesslen, Svitlana Volkova, and Wenwen Dou. "Towards rapid interactive machine learning: evaluating tradeoffs of classification without representation". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Mar. 2019, pp. 591–602. DOI: 10.1145/3301275.3302280 (cit. on p. 113).

[16] Erasmo Artur and Rosane Minghim. "A novel visual approach for enhanced attribute analysis and selection". In: *Computer Graphics* 84 (2019), pp. 160–172. DOI: 10.1016/j.cag.2019.08.015 (cit. on pp. 146, 148–151, 154).

[17] Peter F. Ash and Ethan D. Bolker. "Generalized Dirichlet tessellations". In: *Geometriae Dedicata* 20.2 (Apr. 1986), pp. 209–243. ISSN: 1572-9168. DOI: 10.1007/BF00164401 (cit. on p. 73).

[18] Hala Assal, Sonia Chiasson, and Robert Biddle. "Cesar: Visual representation of source code vulnerabilities". In: *Proceedings of the 13th IEEE Symposium on Visualization for Cyber Security*. IEEE, 2016, pp. 1–8. DOI: 10.1109/VIZSEC.2016.7739576 (cit. on p. 97).

[19] Michaël Aupetit. "Visualizing distortions and recovering topology in continuous projection techniques". In: *Neurocomputing* 70.7-9 (2007), pp. 1304–1330. DOI: 10.1016/J.NEUCOM.2006.11.018 (cit. on pp. 51, 61).

[20] Michaël Aupetit and Thibaud Catz. "High-dimensional labeled data analysis with topology representing graphs". In: *Neurocomputing* 63 (2005), pp. 139–169. DOI: 10.1016/j.neucom.2004.04.009 (cit. on pp. 52, 57).

[21] Michaël Aupetit and Michael Sedlmair. "SepMe: 2002 New visual separation measures". In: *Proceedings of the IEEE Pacific Visualization Symposium*. Ed. by Chuck Hansen, Ivan Viola, and Xiaoru Yuan. IEEE, 2016, pp. 1–8. DOI: 10.1109/PACIFICVIS.2016.7465244 (cit. on p. 52).

[22] Franz Aurenhammer. "Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure". In: *ACM Computing Surveys* 23.3 (1991), pp. 345–405. DOI: 10.1145/116873.116880 (cit. on pp. 51, 56, 57, 73).

[23] R. Harald Baayen. *Analyzing linguistic data: a practical introduction to statistics using R*. 1st. Cambridge University Press, 2008. ISBN: 978-0-521-70918-7 (cit. on p. 78).

[24] Michael Balzer and Oliver Deussen. "Voronoi Treemaps". In: *Proceedings of the IEEE Symposium on Information Visualization*. Ed. by John T. Stasko and Matthew O. Ward. IEEE, 2005, pp. 49–56. DOI: 10.1109/INFVIS.2005.1532128 (cit. on p. 73).

[25] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. "Learning a Mahalanobis Metric from Equivalence Constraints". In: *Journal of Machine Learning Research* 6 (2005), pp. 937–965 (cit. on p. 113).

[26] Jóhanna Barðdal. "The rise of Dative Substitution in the history of Icelandic: A diachronic construction grammar account". In: *Lingua* 121.1 (2011), pp. 60–79. DOI: https://doi.org/10.1016/j.lingua.2010.07.007 (cit. on p. 86).

[27] Ray Bareiss, Bruce W. Porter, and Craig C. Wier. "Protos: An Exemplar-Based Learning Apprentice". In: *International Journal of Man-Machine Studies* 29.5 (1988), pp. 549–561. DOI: 10.1016/S0020-7373(88)80012-9 (cit. on pp. 65, 66).

[28] Margaret E. Baron. "A Note on the Historical Development of Logic Diagrams: Leibniz, Euler and Venn". In: *The Mathematical Gazette* 53.384 (1969), pp. 113–125 (cit. on p. 49).

[29] Oliver Bastert and Christian Matuszewski. "Layered Drawings of Digraphs". In: *Drawing Graphs, Methods and Models*. Ed. by Michael Kaufmann and Dorothea Wagner. Vol. 2025. Lecture Notes in Computer Science. Springer, 1999, pp. 87–120. DOI: 10.1007/3-540-44969-8_5 (cit. on p. 174).

[30] Christian Baumgartner, Claudia Plant, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. "Subspace Selection for Clustering High-Dimensional Data". In: *Proceedings of the 4th IEEE International Conference on Data Mining*. IEEE, 2004, pp. 11–18. DOI: 10.1109/ICDM.2004.10112 (cit. on pp. 137, 167).

[31] Michael Behrisch, Benjamin Bach, Michael Hund, Michael Delz, Laura von Rüden, Jean-Daniel Fekete, and Tobias Schreck. “Magnostics: Image-Based Search of Interesting Matrix Views for Guided Network Exploration”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 31–40. DOI: 10.1109/TVCG.2016.2598467 (cit. on pp. 25, 51).

[32] Michael Behrisch, Benjamin Bach, Michael Hund, Michael Delz, Laura von Rüden, Jean-Daniel Fekete, and Tobias Schreck. “Magnostics: Image-Based Search of Interesting Matrix Views for Guided Network Exploration”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 31–40. DOI: 10.1109/TVCG.2016.2598467 (cit. on pp. 120, 121).

[33] Michael Behrisch, Michael Blumenschein, Nam Wook Kim, Lin Shao, Mennatallah El-Assady, Johannes Fuchs, Daniel Seebacher, Alexandra Diehl, Ulrik Brandes, Hanspeter Pfister, Tobias Schreck, Daniel Weiskopf, and Daniel A. Keim. “Quality Metrics for Information Visualization”. In: *Computer Graphics Forum* 37.3 (2018), pp. 625–662. DOI: 10.1111/cgf.13446 (cit. on pp. 4, 25, 26, 46, 51, 169, 174).

[34] Michael Behrisch, Fatih Korkmaz, Lin Shao, and Tobias Schreck. “Feedback-driven interactive exploration of large multidimensional data supported by visual classifier”. In: *Proceedings of the 9th IEEE Conference on Visual Analytics Science and Technology*. IEEE, Oct. 2014, pp. 43–52. DOI: 10.1109/VAST.2014.7042480 (cit. on pp. 5, 110, 111, 116, 118).

[35] Fabian Bendix, Robert Kosara, and Helwig Hauser. “Parallel Sets: Visual Analysis of Categorical Data”. In: *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 133–140. DOI: 10.1109/INFVIS.2005.1532139 (cit. on pp. 6, 22, 29, 38, 48, 50, 63, 85).

[36] Boris Beranger, Huan Lin, and Scott A. Sisson. “New models for symbolic data analysis”. In: *Advances in Data Analysis and Classification* 17.3 (2023), pp. 659–699. DOI: 10.1007/S11634-022-00520-8 (cit. on p. 1).

[37] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter W. Fellner, and Michael Sedlmair. “Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 298–308. DOI: 10.1109/TVCG.2017.2744818 (cit. on p. 119).

[38] Jürgen Bernard, Tobias Ruppert, Maximilian Scherer, Tobias Schreck, and Jörn Kohlhammer. “Guided discovery of interesting relationships between time series clusters and metadata properties”. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*. ACM, Sept. 2012. DOI: 10.1145/2362456.2362485 (cit. on pp. 114, 115).

[39] Jürgen Bernard, David Sessler, Andreas Bannach, Thorsten May, and Jörn Kohlhammer. “A visual active learning system for the assessment of patient well-being in prostate cancer research”. In: *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare*. ACM, Oct. 2015, pp. 1–8. DOI: 10.1145/2836034.2836035 (cit. on p. 113).

[40] Jacques Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. Esri Press, 2010. ISBN: 978-1-589-48261-6 (cit. on p. 148).

[41] Enrico Bertini. “Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2203–2212. DOI: 10.1109/TVCG.2011.229 (cit. on pp. 4, 169, 173).

[42] Enrico Bertini and Denis Lalanne. “Surveying the complementary role of automatic data analysis and visualization in knowledge discovery”. In: *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*. ACM, 2009, pp. 12–20. DOI: 10.1145/1562849.1562851 (cit. on p. 140).

[43] Adrien Bibal, Antoine Clarinval, Bruno Dumas, and Benoît Frénay. “IXVC: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees”. In: *Array* 11 (2021), p. 100080. DOI: 10.1016/j.array.2021.100080 (cit. on p. 144).

[44] Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis Theory and Practice*. Springer, 2007. ISBN: 978-0-387-72806-3. DOI: 10.1007/978-0-387-72806-3 (cit. on p. 1).

[45] Daniela Blumberg, Yu Wang, Alexandru Telea, Daniel A. Keim, and Frederik L. Dennig. “Inverting Multidimensional Scaling Projections Using Data Point Multilateration”. In: *Proceedings of the 15th International EuroVis Workshop on Visual Analytics*. Eurographics, 2024. DOI: 10.2312/eurova.20241112 (cit. on pp. 17, 51).

[46] Michael Blumenschein, Xuan Zhang, David Pomerenke, Daniel A. Keim, and Johannes Fuchs. “Evaluating Reordering Strategies for Cluster Identification in Parallel Coordinates”. In: *Computer Graphics Forum* 39.3 (2020), pp. 537–549. DOI: 10.1111/cgf.14000 (cit. on pp. 25, 34).

[47] Ingwer Borg and Patrick Groenen Groenen. *Modern multidimensional scaling: Theory and applications*. 1st. Springer Series in Statistics. Springer, 2005. ISBN: 978-1-4757-2711-1. DOI: https://doi.org/10.1007/978-1-4757-2711-1 (cit. on p. 148).

[48] Shyam Boriah, Varun Chandola, and Vipin Kumar. “Similarity Measures for Categorical Data: A Comparative Evaluation”. In: *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2008, pp. 243–254. DOI: 10.1137/1.9781611972788.22 (cit. on pp. 54, 174).

[49] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. “Representing shape with a spatial pyramid kernel”. In: *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM, July 2007, pp. 401–408. DOI: 10.1145/1282280.1282340 (cit. on p. 121).

[50] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. “D^3 Data-Driven Documents”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2301–2309. DOI: 10.1109/TVCG.2011.185 (cit. on pp. 24, 55).

[51] Ulrik Brandes and Michael Sedlmair. “Network Visualization”. In: *Network Science: An Aerial View*. Springer, 2019, pp. 5–21. DOI: 10.1007/978-3-030-26814-5_2 (cit. on p. 1).

[52] Matthew Brehmer and Tamara Munzner. “A Multi-Level Typology of Abstract Visualization Tasks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2376–2385. DOI: 10.1109/TVCG.2013.124 (cit. on pp. 140, 147, 150).

[53] Sebastian Bremm, Tatiana von Landesberger, Jürgen Bernard, and Tobias Schreck. “Assisted Descriptor Selection Based on Visual Comparative Data Analysis”. In: *Computer Graphics Forum* 30.3 (2011), pp. 891–900. DOI: 10.1111/j.1467-8659.2011.01938.x (cit. on p. 112).

[54] Cynthia A. Brewer, Geoffrey W. Hatchard, and Mark A. Harrower. “ColorBrewer in Print: A Catalog of Color Schemes for Maps”. In: *Cartography and Geographic Information Science* 30.1 (Jan. 2003), pp. 5–32. ISSN: 1523-0406. DOI: 10.1559/152304003100010929 (cit. on p. 30).

[55] Bertjan Broeksema, Alexandru C. Telea, and Thomas Baudel. “Visual Analysis of Multi-Dimensional Categorical Data Sets”. In: *Computer Graphics Forum* 32.8 (2013), pp. 158–169. DOI: 10.1111/cgf.12194 (cit. on pp. 8, 48, 51, 55, 63).

[56] Eli T. Brown, Jingjing Liu, Carla E. Brodley, and Remco Chang. “Dis-function: Learning distance functions interactively”. In: *Proceedings of the 7th IEEE Conference on Visual Analytics Science and Technology*. IEEE, Oct. 2012, pp. 83–92. DOI: 10.1109/VAST.2012.6400486 (cit. on p. 113).

[57] Jason Brownlee. *Why One-Hot Encode Data in Machine Learning?* Online. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/, accessed 2022-10-19. 2017 (cit. on pp. 3, 54, 157, 174).

[58] Raphael Buchmüller, Bastian Jäckl, Michael Behrisch, Daniel A. Keim, and Frederik L. Dennig. “cPro: Circular Projections Using Gradient Descent”. In: *Proceedings of the 15th International EuroVis Workshop on Visual Analytics*. Eurographics, 2024. DOI: 10.2312/eurova.20241111 (cit. on p. 17).

[59] Benjamin L. Bullough, Anna K. Yanchenko, Christopher L. Smith, and Joseph R. Zipkin. “Predicting Exploitation of Disclosed Software Vulnerabilities Using Open-source Data”. In: *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*. ACM, 2017, pp. 45–53. DOI: 10.1145/3041008.3041009 (cit. on p. 96).

[60] Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehé, and Daniel A. Keim. “V1 in Icelandic: A Multifactorical Visualization of Historical Data”. In: *Proceedings of VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*. Ed. by Annette Hautli-Janisz. 2014, pp. 33–40. ISBN: 978-2-9517408-8-4 (cit. on pp. 79, 89).

[61] Nan Cao, David Gotz, Jimeng Sun, and Huamin Qu. “DICON: Interactive Visual Analysis of Multidimensional Clusters”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2581–2590. DOI: 10.1109/TVCG.2011.188 (cit. on p. 51).

[62] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision To Think*. 1st. Morgan Kaufmann, 1999. ISBN: 978-1-55860-533-6 (cit. on pp. 3–5, 163).

[63] Sheelagh Carpendale. “Evaluating Information Visualizations”. In: *Information Visualization - Human-Centered Issues and Perspectives*. Vol. 4950. Lecture Notes in Computer Science. Springer, 2008, pp. 19–45. DOI: 10.1007/978-3-540-70956-5_2 (cit. on pp. 12, 86).

[64] Marco Cavallo and Çagatay Demiralp. “A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–13. DOI: 10.1145/3173574.3174209 (cit. on p. 22).

[65] Sung-Hyuk Cha. “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions”. In: *International Journal of Mathematical Models Methods Applied Science* 1.4 (2007), pp. 300–307 (cit. on p. 54).

[66] Savvas A. Chatzichristofis and Yiannis S. Boutalis. “CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval”. In: *Computer Vision Systems*. Springer, 2008, pp. 312–322. ISBN: 978-3-540-79547-6. DOI: 10.1007/978-3-540-79547-6_30 (cit. on p. 121).

[67] Savvas A. Chatzichristofis and Yiannis S. Boutalis. “FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval”. In: *Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services*. IEEE, May 2008, pp. 191–196. DOI: 10.1109/WIAMIS.2008.24 (cit. on p. 121).

[68] Mohammad Chegini, Jürgen Bernard, Philip Berger, Alexei Sourin, Keith Andrews, and Tobias Schreck. “Interactive labelling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning”. In: *Visual Informatics* 3.1 (2019), pp. 9–17. DOI: 10.1016/j.visinf.2019.03.002 (cit. on p. 119).

[69] Min Chen and David S. Ebert. “An Ontological Framework for Supporting the Design and Evaluation of Visual Analytics Systems”. In: *Computer Graphics Forum* 38.3 (2019), pp. 131–144. DOI: 10.1111/cgf.13677 (cit. on pp. 5, 100).

[70] Wei Chen, Zi’ang Ding, Song Zhang, Anna MacKay-Brandt, Stephen Correia, Huamin Qu, John Allen Crow, David F. Tate, Zhicheng Yan, and Qunsheng Peng. “A Novel Interface for Interactive Exploration of DTI Fibers”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1433–1440. DOI: 10.1109/TVCG.2009.112 (cit. on p. 145).

[71] Shenghui Cheng and Klaus Mueller. “The Data Context Map: Fusing Data and Attributes into a Unified Display”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 121–130. DOI: 10.1109/TVCG.2015.2467552 (cit. on p. 50).

[72] Noptanit Chotisarn, Leonel Merino, Xu Zheng, Supaporn Lonapalawong, Tianye Zhang, Mingliang Xu, and Wei Chen. “A systematic literature review of modern software visualization”. In: *Journal of Visualization* 23.4 (2020), pp. 539–558. DOI: 10.1007/s12650-020-00647-w (cit. on p. 97).

[73] John Clark and Derek Allan Holton. *A First Look at Graph Theory*. World Scientific, 1991. ISBN: 978-981-02-0490-7. DOI: 10.1142/1280 (cit. on p. 57).

[74] William S. Cleveland. *The Elements of Graphing Data*. 2nd. Hobart Press, 1994. ISBN: 978-0-9634884-1-1 (cit. on p. 22).

[75] Paul van der Corput and Jarke J. van Wijk. “Exploring Items and Features with I^F, F^I-Tables”. In: *Computer Graphics Forum* 35.3 (2016), pp. 31–40. DOI: 10.1111/cgf.12879 (cit. on pp. 137, 142, 146, 148, 149, 151, 153, 155, 163, 168).

[76] Chris Culy, Verena Lyding, and Henrik Dittmann. “Structured Parallel Coordinates: a visualization for analyzing structured language data”. In: Editorial Universitat Politècnica de València, 2011, pp. 525–533 (cit. on p. 85).

[77] Douglas R. Cutting, Jan O. Pedersen, David R. Karger, and John W. Tukey. “Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections”. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, June 1992, pp. 318–329. DOI: 10.1145/133160.133214 (cit. on p. 123).

[78] Aritra Dasgupta and Robert Kosara. “Pargnostics: Screen-Space Metrics for Parallel Coordinates”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1017–1026. DOI: `10.1109/TVCG.2010.184` (cit. on pp. 4, 25, 33, 34, 37, 52).

[79] David L. Davies and Donald W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1.2 (1979), pp. 224–227. DOI: `10.1109/TPAMI.1979.4766909` (cit. on p. 123).

[80] Robert J. MacG. Dawson. *The "Unusual Episode" Data Revisited*. `http://jse.amstat.org/v3n3/datasets.dawson.html`, last accessed 2023-11-14. 1995 (cit. on pp. 22, 30, 33, 38, 57, 63, 65, 66, 68, 69).

[81] Ross H. Day and Erica J. Stecher. “Sine of an Illusion”. In: *Perception* 20 (1 1991), pp. 49–55. DOI: `10.1068/p200049` (cit. on pp. 24, 50).

[82] Frederik L. Dennig, Eren Cakmak, Henrik Plate, and Daniel A. Keim. “VulnEx: Exploring Open-Source Software Vulnerabilities in Large Development Organizations to Understand Risk Exposure”. In: *Proceedings of the IEEE Symposium on Visualization for Cyber Security*. IEEE, 2021, pp. 79–83. DOI: `10.1109/VizSec53666.2021.00014` (cit. on pp. 15, 95).

[83] Frederik L. Dennig, Maximilian T. Fischer, Michael Blumenschein, Daniel Fuchs Johannes; Keim, and Evanthia Dimara. *Replication Data for: "ParSetgnostics: Quality Metrics for Parallel Sets*". Version V1. `https://osf.io/rwhf5/` (alternative repository). 2022. DOI: `10.18419/darus-2869` (cit. on pp. 14, 18, 23).

[84] Frederik L. Dennig, Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. “ParSetgnostics: Quality Metrics for Parallel Sets”. In: *Computer Graphics Forum* 40.3 (2021), pp. 375–386. DOI: `10.1111/cgf.14314` (cit. on pp. 14, 21, 169).

[85] Frederik L. Dennig, Maximilian T. Fischer, Michael Blumenschein, Johannes Fuchs, Daniel A. Keim, and Evanthia Dimara. “ParSetgnostics: Quality Metrics for Parallel Sets”. In: *Computer Graphics Forum* 40.3 (2021), pp. 375–386. DOI: `10.1111/cgf.14314` (cit. on pp. 48, 52, 64).

[86] Frederik L. Dennig, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel Keim, and Maximilian T. Fischer. *The Categorical Data Map - Replication Data*. Version V1. `https://osf.io/jzd46/` (alternative repository). 2024. DOI: `10.18419/darus-3372` (cit. on pp. 14, 18, 49).

[87] Frederik L. Dennig, Lucas Joos, Patrick Paetzold, Daniela Blumberg, Oliver Deussen, Daniel A. Keim, and Maximilian T. Fischer. “The Categorical Data Map: A Multidimensional Scaling-Based Approach”. In: *Proceedings of the 2024 IEEE Visualization in Data Science Symposium (to appear)*. IEEE, 2024 (cit. on pp. 13, 47).

[88] Frederik L. Dennig, Matthias Miller, Daniel A. Keim, and Mennatallah El-Assady. "FS/DS: A Theoretical Framework for the Dual Analysis of Feature Space and Data Space". In: *IEEE Transactions on Visualization and Computer Graphics* 30.8 (2024), pp. 5165–5182. DOI: 10.1109/TVCG.2023.3288356 (cit. on pp. 14, 55, 135).

[89] Frederik L. Dennig, Tom Polk, Zudi Lin, Tobias Schreck, Hanspeter Pfister, and Michael Behrisch. "FDive: Learning Relevance Models Using Pattern-based Similarity Measures". In: *Proceedings of the 14th IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2019, pp. 69–80. DOI: 10.1109/VAST47406.2019.8986940 (cit. on pp. 15, 109, 169).

[90] Edwin Diday. "Thinking by classes in data science: the symbolic data analysis paradigm". In: *WIREs Computational Statistics* 8.5 (2016), pp. 172–205. DOI: 10.1002/wics.1384 (cit. on p. 1).

[91] Stephan Diehl. *Software Visualization - Visualizing the Structure, Behaviour, and Evolution of Software*. Springer, 2007. ISBN: 978-3-540-46504-1. DOI: 10.1007/978-3-540-46505-8 (cit. on p. 97).

[92] Evanthia Dimara and Charles Perin. "What is Interaction for Data Visualization?" In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 119–129. DOI: 10.1109/TVCG.2019.2934283 (cit. on p. 141).

[93] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. *Human-Computer Interaction*. 7th. Pearson Prentice Hall, 2003. ISBN: 978-0-13-046109-4 (cit. on p. 1).

[94] Helmut Doleisch, Helwig Hauser, Martin Gasser, and Robert Kosara. "Interactive Focus+Context Analysis of Large, Time-Dependent Flow Simulation Data". In: *Simulation* 82.12 (2006), pp. 851–865. DOI: 10.1177/0037549707078278 (cit. on p. 145).

[95] Michelle Dowling, John E. Wenskovitch, J. T. Fry, Scotland Leman, Leanna House, and Chris North. "SIRIUS: Dual, Symmetric, Interactive Dimension Reductions". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 172–182. DOI: 10.1109/TVCG.2018.2865047 (cit. on pp. 136, 137, 142, 146, 148–152, 155, 158, 162–164, 167).

[96] David Dowty. "Thematic Proto-Roles and Argument Selection". In: *Language* 67.3 (1991), pp. 547–619 (cit. on p. 89).

[97] John J. Dudley and Per Ola Kristensson. "A Review of User Interface Design for Interactive Machine Learning". In: *ACM Transactions on Interactive Intelligent Systems* 8.2 (2018). DOI: 10.1145/3185517 (cit. on p. 113).

[98] Joseph C. Dunn. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3.3 (1973), pp. 32–57. DOI: 10.1080/01969727308546046 (cit. on p. 123).

[99] Eric Eaton, Gary Holness, and Daniel McFarlane. "Interactive Learning Using Manifold Geometry". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. Vol. 24. 1. July 2010, pp. 437–443. DOI: 10.1609/aaai.v24i1.7688 (cit. on p. 113).

[100] Geoffrey P. Ellis and Alan J. Dix. "Enabling Automatic Clutter Reduction in Parallel Coordinate Plots". In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006), pp. 717–724. DOI: 10.1109/TVCG.2006.138 (cit. on p. 34).

[101] Stef van den Elzen and Jarke J. van Wijk. "BaobabView: Interactive construction and analysis of decision trees". In: *Proceedings of the 6th IEEE Conference on Visual Analytics Science and Technology*. IEEE, Oct. 2011, pp. 151–160. DOI: 10.1109/VAST.2011.6102453 (cit. on p. 114).

[102] Alex Endert. "Semantic Interaction for Visual Analytics: Toward Coupling Cognition and Computation". In: *IEEE Computer Graphics and Applications* 34.4 (2014), pp. 8–15. DOI: 10.1109/MCG.2014.73 (cit. on pp. 140, 162).

[103] Mateus Espadoto, Nina Sumiko Tomita Hirata, and Alexandru C. Telea. "Deep learning multidimensional projections". In: *Information Visualization* 19.3 (2020), pp. 247–269. DOI: 10.1177/1473871620909485 (cit. on pp. 67, 68).

[104] Mateus Espadoto, Rafael Messias Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. "Toward a Quantitative Survey of Dimension Reduction Techniques". In: *IEEE Transactions on Visualization and Computer Graphics* 27.3 (2021), pp. 2153–2173. DOI: 10.1109/TVCG.2019.2944182 (cit. on p. 65).

[105] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery & Data Mining*. AAAI Press, 1996, pp. 226–231 (cit. on p. 150).

[106] David W. Evans, Patrick T. Orr, Steven M. Lazar, Daniel Breton, Jennifer Gerard, David H. Ledbetter, Kathleen Janosco, Jessica Dotts, and Holly Batchelder. "Human preferences for symmetry: subjective experience, cognitive conflict and cortical brain activity". In: *PLoS One* 7.6 (2012). DOI: 10.1371/journal.pone.0038966 (cit. on p. 137).

[107] Wenbin Fang, Barton P. Miller, and James A. Kupsch. "Automated tracing and visualization of software security structure and properties". In: *Proceedings of the 9th International Symposium on Visualization for Cyber Security*. ACM, 2012, pp. 9–16. DOI: 10.1145/2379690.2379692 (cit. on p. 97).

[108] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD Process for Extracting Useful Knowledge from Volumes of Data". In: *Communications of the ACM* 39.11 (1996), pp. 27–34. DOI: 10.1145/240455.240464 (cit. on p. 100).

[109] Sara Johansson Fernstad and Jimmy Johansson. "A Task Based Performance Evaluation of Visualization Approaches for Categorical Data Analysis". In: *Proceedings of the 15th International Conference on Information Visualisation*. IEEE, 2011, pp. 80–89. DOI: 10.1109/IV.2011.92 (cit. on p. 5).

[110] Sara Johansson Fernstad, Jane Shaw, and Jimmy Johansson. "Quality-based guidance for exploratory dimensionality reduction". In: *Information Visualization* 12.1 (2013), pp. 44–64. DOI: 10.1177/1473871612460526 (cit. on pp. 136, 146, 148, 149, 151, 152, 155, 158, 163–166).

[111] Maximilian T. Fischer, Frederik L. Dennig, Daniel Seebacher, Daniel A. Keim, and Mennatallah El-Assady. "Communication Analysis through Visual Analytics: Current Practices, Challenges, and New Frontiers". In: *Proceedings of the 2022 IEEE Visualization in Data Science Symposium*. IEEE, Oct. 2022. DOI: 10.1109/VDS57266.2022.00006 (cit. on p. 17).

[112] Danyel Fisher and Miriah D. Meyer. *Making Data Visual - A Practical Guide to Using Visualization For Insight*. O'Reilly, 2018. ISBN: 978-1-491-92846-2 (cit. on p. 162).

[113] James Fogarty, Desney S. Tan, Ashish Kapoor, and Simon A. J. Winder. "CueFlik: interactive concept learning in image search". In: *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2008, pp. 29–38. DOI: 10.1145/1357054.1357061 (cit. on p. 114).

[114] Irene Franco. "V1, V2 and criterial movement in Icelandic". In: *Studies in Linguistics* 2 (2008), pp. 141–164 (cit. on p. 80).

[115] Johannes Fuchs, Frederik L. Dennig, Maria-Viktoria Heinle, Daniel A. Keim, and Sara Di Bartolomeo. "Exploring the Design Space of BioFabric Visualization for Multivariate Network Analysis". In: *Computer Graphics Forum* 43.3 (2024). DOI: 10.1111/CGF.15079 (cit. on p. 17).

[116] Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, and Daniel A. Keim. "A Systematic Review of Experimental Studies on Data Glyphs". In: *IEEE Transactions on Visualization and Computer Graphics* 23.7 (2017), pp. 1863–1879. DOI: 10.1109/TVCG.2016.2549018 (cit. on pp. 161, 166).

[117] Takanori Fujiwara, Oh-Hyun Kwon, and Kwan-Liu Ma. "Supporting Analysis of Dimensionality Reduction Results with Contrastive Learning". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 45–55. DOI: 10.1109/TVCG.2019.2934251 (cit. on pp. 136, 146, 148–151, 155).

[118] Kuno R. Gabriel. "The biplot graphic display of matrices with application to principal component analysis". In: *Biometrika* 58.3 (Dec. 1971), pp. 453–467. DOI: 10.1093/biomet/58.3.453 (cit. on p. 50).

[119] Laura Garrison, Juliane Müller, Stefanie Schreiber, Steffen Oeltze-Jafra, Helwig Hauser, and Stefan Bruckner. "DimLift: Interactive Hierarchical Data Exploration Through Dimensional Bundling". In: *IEEE Transactions on Visualization and Computer Graphics* 27.6 (2021), pp. 2908–2922. DOI: 10.1109/TVCG.2021.3057519 (cit. on pp. 136, 146, 148, 149, 151, 154, 157, 158).

[120] Michael Gleicher. "Explainers: Expert Explorations with Crafted Projections". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2042–2051. DOI: 10.1109/TVCG.2013.157 (cit. on p. 114).

[121] E. Bruce Goldstein. *Sensation and Perception*. 10th. Wadsworth ISE, 2016. ISBN: 978-1-305-58029-9 (cit. on p. 24).

[122] John R. Goodall, Hassan Radwan, and Lenny Halseth. "Visual analysis of code security". In: *Proceedings of the 7th International Symposium on Visualization for Cyber Security*. ACM, 2010, pp. 46–51. DOI: 10.1145/1850795.1850800 (cit. on p. 97).

[123] Jochen Görtler, Thilo Spinner, Dirk Streeb, Daniel Weiskopf, and Oliver Deussen. "Uncertainty-Aware Principal Component Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 822–831. DOI: 10.1109/TVCG.2019.2934812 (cit. on p. 167).

[124] Tera Marie Green, William Ribarsky, and Brian D. Fisher. "Building and applying a human cognition model for visual analytics". In: *Information Visualization* 8.1 (2009), pp. 1–13. DOI: 10.1057/ivs.2008.28 (cit. on p. 138).

[125] Tera Marie Green, William Ribarsky, and Brian D. Fisher. "Visual analytics for complex concepts using a human cognition model". In: *Proceedings of the 3rd IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2008, pp. 91–98. DOI: 10.1109/VAST.2008.4677361 (cit. on pp. 5, 138).

[126] M. Greenacre and Blasius. *Multiple Correspondence Analysis and Related Methods*. 1st ed. Chapman & Hall/CRC, 2006. DOI: 10.1201/9781420011319 (cit. on pp. 50, 51, 54).

[127] Robert Gregor, Andreas Lamprecht, Ivan Sipiran, Tobias Schreck, and Benjamin Bustos. "Empirical evaluation of dissimilarity measures for 3D object retrieval with application to multi-feature retrieval". In: *Proceedings of the 13th International Workshop on Content-Based Multimedia Indexing*. IEEE, June 2015, pp. 1–6. DOI: 10.1109/CBMI.2015.7153629 (cit. on p. 111).

[128] *GRETIL (Göttingen Register of Electronic Texts in Indian Languages)*. Online. https://gretil.sub.uni-goettingen.de/gretil.html, accessed 2024-04-17. 2001 (cit. on p. 78).

[129] Jane Grimshaw. *Argument Structure*. Linguistic Inquiry Monographs. The MIT Press, Apr. 1992. ISBN: 978-0-262-57090-9 (cit. on p. 89).

[130] Isabelle Guyon and André Elisseeff. “An Introduction to Variable and Feature Selection”. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182 (cit. on p. 112).

[131] David Hägele, Tim Krake, and Daniel Weiskopf. “Uncertainty-Aware Multidimensional Scaling”. In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2023), pp. 23–32. DOI: 10.1109/TVCG.2022.3209420 (cit. on p. 167).

[132] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000. ISBN: 1-55860-489-8 (cit. on p. 160).

[133] Ju Han and Kai-Kuang Ma. “Fuzzy color histogram and its use in color image retrieval”. In: *IEEE Transactions on Image Processing* 11.8 (2002), pp. 944–952. DOI: 10.1109/TIP.2002.801585 (cit. on p. 121).

[134] John T. Hancock and Taghi M. Khoshgoftaar. “Survey on categorical data for neural networks”. In: *Journal of Big Data* 7.1 (2020), p. 28. DOI: 10.1186/S40537-020-00305-W (cit. on pp. 3, 54, 157).

[135] Robert M. Haralick, K. Sam Shanmugam, and Its’hak Dinstein. “Textural Features for Image Classification”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 3.6 (1973), pp. 610–621. DOI: 10.1109/TSMC.1973.4309314 (cit. on pp. 121, 128).

[136] Lane Harrison, Riley Spahn, Michael D. Iannacone, Evan Downing, and John R. Goodall. “NV: Nessus vulnerability visualization for the web”. In: *Proceedings of the 9th International Symposium on Visualization for Cyber Security*. ACM, 2012, pp. 25–32. DOI: 10.1145/2379690.2379694 (cit. on p. 97).

[137] Mark Harrower and Cynthia A. Brewer. “ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps”. In: *The Cartographic Journal* 40.1 (2003), pp. 27–37. DOI: 10.1179/000870403235002042 (cit. on p. 3).

[138] Ulrich Harsch. *Bibliotheca Augustana*. Online. https://www.hs-augsburg.de/~harsch/augustana.html, accessed 2024-04-17. 1997 (cit. on p. 78).

[139] J. A. Hartigan. “Printer graphics for clustering”. In: *Journal of Statistical Computation and Simulation* 4.3 (1975), pp. 187–213. DOI: 10.1080/00949657508810123 (cit. on p. 22).

[140] Sabri Hassan and Günther Pernul. “Efficiently Managing the Security and Costs of Big Data Storage using Visual Analytics”. In: *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services*. 2014, pp. 180–184. DOI: 10.1145/2684200.2684333 (cit. on pp. 38, 39, 65, 68).

[141] Einar Haugen. *Die skandinavischen Sprachen: Eine Einführung in ihre Geschichte*. German. Buske, 1984. ISBN: 978-3-87118-551-9 (cit. on p. 82).

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 129).

[143] Christopher G. Healey and Brent M. Dennis. “Interest Driven Navigation in Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.10 (2012), pp. 1744–1756. DOI: 10.1109/TVCG.2012.23 (cit. on p. 23).

[144] Marko Heikkilä and Matti Pietikäinen. “A Texture-Based Method for Modeling the Background and Detecting Moving Objects”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.4 (2006), pp. 657–662. DOI: 10.1109/TPAMI.2006.68 (cit. on p. 121).

[145] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. “Visual Classifier Training for Text Document Retrieval”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2839–2848. DOI: 10.1109/TVCG.2012.277 (cit. on pp. 113, 115).

[146] Martin Hilpert and Stefan Th. Gries. “Quantitative approaches to diachronic corpus linguistics”. In: *The Cambridge Handbook of English Historical Linguistics*. Ed. by Merja Kytö and Päivi Pahta. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2016, pp. 36–53. DOI: 10.1017/CBO9781139600231.003 (cit. on p. 78).

[147] Andreas P. Hinterreiter, Christian Alexander Steinparz, Moritz Schöfl, Holger Stitz, and Marc Streit. “Projection Path Explorer: Exploring Visual Patterns in Projected Decision-making Paths”. In: *ACM Transactions on Interactive Intelligent Systems* 11.3-4 (2021), 22:1–22:29. DOI: 10.1145/3387165 (cit. on pp. 137, 167).

[148] Heike Hofmann. “Exploring categorical data: interactive mosaic plots”. In: *Metrika* 51.1 (July 2000), pp. 11–26. ISSN: 0026-1335. DOI: 10.1007/s001840000041 (cit. on pp. 22, 24, 173).

[149] Heike Hofmann, Arno Siebes, and Adalbert F. X. Wilhelm. “Visualizing association rules with interactive mosaic plots”. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2000, pp. 227–235. DOI: 10.1145/347090.347133 (cit. on pp. 48, 50, 173).

[150] Heike Hofmann and Marie Vendettuoli. “Common Angle Plots as Perception-True Visualizations of Categorical Associations”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2297–2305. DOI: 10.1109/TVCG.2013.140 (cit. on p. 24).

[151] Petter Holme and Jari Saramäki. *Temporal Network Theory*. Computational Social Sciences. Springer, 2019. ISBN: 978-3-031-30398-2. DOI: 10.1007/978-3-031-30399-9 (cit. on p. 1).

[152] P. V. C. Hough. *Method and means for recognizing complex patterns*. U.S. Patent No. 30696541962. Dec. 1962 (cit. on p. 121).

[153] Thorbjörg Hróarsdóttir. *Word Order Change in Icelandic: From OV to VO*. Vol. 35. Linguistik Aktuell/Linguistics Today. John Benjamins Publishing Company, 2000. ISBN: 978-90-272-2756-0 (cit. on pp. 80, 89).

[154] Jing Huang, Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. "Image Indexing Using Color Correlograms". In: *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*. IEEE, June 1997, pp. 762–768. DOI: 10.1109/CVPR.1997.609412 (cit. on p. 121).

[155] Weidong Huang, Seok-Hee Hong, and Peter Eades. "Effects of Crossing Angles". In: *IEEE VGTC Pacific Visualization Symposium*. IEEE, 2008, pp. 41–46. DOI: 10.1109/PACIFICVIS.2008.4475457 (cit. on pp. 29, 31, 33, 34, 45).

[156] Alfred Inselberg. "The plane with parallel coordinates". In: *The Visual Computer* 1.2 (Aug. 1985), pp. 69–91. DOI: 10.1007/bf01898350 (cit. on pp. 22, 50, 85, 136, 148).

[157] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. "A Systematic Review on the Practice of Evaluating Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2818–2827. DOI: 10.1109/TVCG.2013.126 (cit. on pp. 12, 86).

[158] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman Kim. "High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots". In: *Journal of Visual Languages & Computing* 43 (2017), pp. 1–13. DOI: 10.1016/j.jvlc.2017.03.001 (cit. on pp. 136, 146, 148, 149, 151, 154, 155).

[159] Paul Jaccard. "The Distribution of The Flora In The Alpine Zone.1". In: *New Phytologist* 11.2 (1912), pp. 37–50. DOI: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x (cit. on pp. 53, 54).

[160] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. 2nd. Springer, 2021. ISBN: 978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1 (cit. on p. 112).

[161] Mohsen Jenadeleh, Frederik L. Dennig, Rene Cutura, Quynh Quang Ngo, Daniel A. Keim, Michael Sedlmair, and Dietmar Saupe. "An Image Quality Dataset with Triplet Comparisons for Multi-dimensional Scaling". In: *Proceedings of the 16th International Conference on Quality of Multimedia Experience*. IEEE, 2024, pp. 278–281. DOI: 10.1109/QoMEX61742.2024.10598258 (cit. on p. 17).

[162] Wolfgang Jentner, Dominik Sacha, Florian Stoffel, Geoffrey P. Ellis, Leishi Zhang, and Daniel A. Keim. "Making machine intelligence less scary for criminal analysts: reflections on designing a visual comparative case analysis tool". In: *The Visual Computer* 34.9 (2018), pp. 1225–1241. DOI: `10.1007/s00371-018-1483-0` (cit. on pp. 136, 146, 148–153, 155).

[163] Wolfgang Jentner, Fabian Sperrle, Daniel Seebacher, Matthias Kraus, Rita Sevastjanova, Maximilian T. Fischer, Udo Schlegel, Dirk Streeb, Matthias Miller, Thilo Spinner, Eren Cakmak, Matthew Sharinghousen, Philipp Meschenmoser, Jochen Görtler, Oliver Deussen, Florian Stoffel, Hans-Joachim Kabitz, Daniel A. Keim, Mennatallah El-Assady, and Juri F. Buchmüller. "Visualisierung der COVID-19-Inzidenzen und Behandlungskapazitäten mit CoronaVis". In: *Resilienz und Pandemie: Handlungsempfehlungen anhand von Erfahrungen mit COVID-19*. Ed. by Andreas Karsten and Stefan Voßschmidt. Kohlhammer, 2022, pp. 176–189. ISBN: 978-3-17-039930-3 (cit. on p. 171).

[164] Dean F. Jerding and John T. Stasko. "The Information Mural: A Technique for Displaying and Navigating Large Information Spaces". In: *IEEE Transactions on Visualization and Computer Graphics* 4.3 (1998), pp. 257–271. DOI: `10.1109/2945.722299` (cit. on pp. 22, 48).

[165] Paulo Joia, Danilo Barbosa Coimbra, José Alberto Cuminato, Fernando Vieira Paulovich, and Luis Gustavo Nonato. "Local Affine Multidimensional Projection". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2563–2571. DOI: `10.1109/TVCG.2011.220` (cit. on p. 67).

[166] Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 1986. ISBN: 978-1-4757-1906-2. DOI: `10.1007/978-1-4757-1904-8` (cit. on pp. 50, 136, 149, 161).

[167] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. *Quick, Draw!* Online. `https://experiments.withgoogle.com/quick-draw`, last accessed 2023-11-14. May 2017 (cit. on p. 130).

[168] Lucas Joos, Karsten Klein, Maximilian T. Fischer, Frederik L. Dennig, Daniel A. Keim, and Michael Krone. "Exploring Trajectory Data in Augmented Reality: A Comparative Study of Interaction Modalities". In: *Proceedings of the 2023 ISMAR International Symposium on Mixed and Augmented Reality*. IEEE, 2023, pp. 790–799. DOI: `10.1109/ISMAR59233.2023.00094` (cit. on p. 17).

[169] Linda T. Kaastra and Brian D. Fisher. "Field experiment methodology for pair analytics". In: *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM, 2014, pp. 152–159. DOI: `10.1145/2669557.2669572` (cit. on pp. 5, 12, 68, 104).

[170] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. “Density-Connected Subspace Clustering for High-Dimensional Data”. In: *Proceedings of the Fourth SIAM International Conference on Data Mining*. SIAM, 2004, pp. 246–256. DOI: 10.1137/1.9781611972740.23 (cit. on p. 167).

[171] Karin Kailing, Hans-Peter Kriegel, Peer Kröger, and Stefanie Wanka. “Ranking Interesting Subspaces for Clustering High Dimensional Data”. In: *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Vol. 2838. Lecture Notes in Computer Science. Springer, 2003, pp. 241–252. DOI: 10.1007/978-3-540-39804-2\textunderscore23 (cit. on p. 167).

[172] Eser Kandogan. *Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions*. https://people.cs.vt.edu/~north/infoviz/starcoords.pdf, last accessed 2023-11-14 (cit. on p. 148).

[173] Eiji Kasutani and Akio Yamada. “The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval”. In: *Proceedings of the 2001 International Conference on Image Processing*. IEEE, Oct. 2001, pp. 674–677. DOI: 10.1109/ICIP.2001.959135 (cit. on p. 121).

[174] Rebecca Kehlbeck, Jochen Görtler, Yunhai Wang, and Oliver Deussen. “SPEULER: Semantics-preserving Euler Diagrams”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2022), pp. 433–442. DOI: 10.1109/TVCG.2021.3114834 (cit. on p. 49).

[175] Daniel A. Keim. “Designing Pixel-Oriented Visualization Techniques: Theory and Applications”. In: *IEEE Transactions on Visualization and Computer Graphics* 6.1 (2000), pp. 59–78. DOI: 10.1109/2945.841121 (cit. on p. 56).

[176] Daniel A. Keim. “Information Visualization and Visual Data Mining”. In: *IEEE Transactions on Visualisation and Computer Graphics* 8.1 (2002), pp. 1–8. DOI: 10.1109/2945.981847 (cit. on p. 1).

[177] Daniel A. Keim. “Pixel-oriented Database Visualizations”. In: *SIGMOD Record* 25.4 (1996), pp. 35–39. DOI: 10.1145/245882.245896 (cit. on p. 148).

[178] Daniel A. Keim, Gennady L. Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. “Visual Analytics: Definition, Process, and Challenges”. In: *Information Visualization - Human-Centered Issues and Perspectives*. Vol. 4950. Lecture Notes in Computer Science. Springer, 2008, pp. 154–175. DOI: 10.1007/978-3-540-70956-5_7 (cit. on pp. 5, 78).

[179] Daniel A. Keim, Jörn Kohlhammer, Geoffrey P. Ellis, and Florian Mansmann. *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010. ISBN: 978-3-905673-77-7 (cit. on pp. 1, 114, 138).

[180] Daniel A. Keim, Wolfgang Müller, and Heidrun Schumann. "Visual Data Mining". In: *23rd Annual Conference of the European Association for Computer Graphics*. Ed. by Dieter W. Fellner and Roberto Scopigno. Eurographics, 2002. DOI: 10.2312/EGST.20021052 (cit. on p. 1).

[181] Alexander B. W. Kennedy and Henry R. Sankey. "The Thermal Efficiency Of Steam Engines". In: *Minutes of the Proceedings of the Institution of Civil Engineers* 134 (1898), pp. 278–312. DOI: 10.1680/imotp.1898.19100 (cit. on pp. 23, 24, 48, 50).

[182] Paul Kiparsky. "The Shift to Head-Initial VP in Germanic". In: *Comparative Germanic Syntax* 2 (1996), pp. 140–179 (cit. on p. 80).

[183] "Pearson's Correlation Coefficient". In: *Encyclopedia of Public Health*. Ed. by Wilhelm Kirch. Springer, 2008, pp. 1090–1091. ISBN: 978-1-4020-5614-7. DOI: 10.1007/978-1-4020-5614-7_2569 (cit. on p. 43).

[184] B. Kleiner and J. A. Hartigan. "Representing Points in Many Dimensions by Trees and Castles". In: *Journal of The American Statistical Association* 76.374 (June 1981), pp. 260–269. DOI: 10.1080/01621459.1981.10477638 (cit. on p. 114).

[185] Stephen G. Kobourov. *Spring Embedders and Force Directed Graph Drawing Algorithms*. 2012. arXiv: 1201.3011 [cs.CG] (cit. on p. 55).

[186] Lian Chee Koh, Aidan Slingsby, Jason Dykes, and Tin Seong Kam. "Developing and Applying a User-Centered Model for the Design and Implementation of Information Visualization Tools". In: *Proceedings of the 15th International Conference on Information Visualisation*. IEEE, 2011, pp. 90–95. DOI: 10.1109/IV.2011.32 (cit. on pp. 38, 41–43, 53, 61, 65, 66).

[187] Teuvo Kohonen. "Self-organized formation of topologically correct feature maps". In: *Biological Cybernetics* 43.1 (Jan. 1982), pp. 59–69. ISSN: 1432-0770. DOI: 10.1007/BF00337288 (cit. on p. 115).

[188] Irena Koprinska. "Feature Selection for Brain-Computer Interfaces". In: *New Frontiers in Applied Data Mining*. Vol. 5669. Lecture Notes in Computer Science. Springer, 2009, pp. 106–117. DOI: 10.1007/978-3-642-14640-4_8 (cit. on pp. 112, 137).

[189] Robert Kosara. "Turning a Table into a Tree: Growing Parallel Sets into a Purposeful Project". In: *Beautiful Visualization: Looking at Data through the Eyes of Experts*. O'Reilly Media, 2010, pp. 193–204 (cit. on pp. 22, 27).

[190] Robert Kosara, Fabian Bendix, and Helwig Hauser. "Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data". In: *IEEE Transactions on Visualization and Computer Graphics* 12.4 (2006), pp. 558–568. DOI: 10.1109/TVCG.2006.76 (cit. on pp. 6, 22, 24, 38, 48, 50, 63, 85).

[191] Josua Krause, Aritra Dasgupta, Jean-Daniel Fekete, and Enrico Bertini. “SeekAView: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces”. In: *Proceedings of the 6th IEEE Symposium on Large Data Analysis and Visualization*. IEEE, 2016, pp. 11–19. DOI: 10.1109/LDAV.2016.7874305 (cit. on pp. 136, 146, 148, 149, 151, 154, 155).

[192] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. “Subspace clustering”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.4 (2012), pp. 351–364. DOI: 10.1002/widm.1057 (cit. on pp. 153, 167).

[193] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105 (cit. on pp. 110, 129).

[194] Joseph B. Kruskal. “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis”. In: *Psychometrika* 29.1 (1964), pp. 1–27. DOI: 10.1007/BF02289565 (cit. on pp. 53, 55, 69, 70, 148, 150, 161).

[195] Joseph B. Kruskal and James M. Landwehr. “Icicle Plots: Better Displays for Hierarchical Clustering”. In: *The American Statistician* 37.2 (1983), pp. 162–168. DOI: 10.1080/00031305.1983.10482733 (cit. on pp. 24, 114).

[196] Tatiana von Landesberger, Sebastian Fiebig, Sebastian Bremm, Arjan Kuijper, and Dieter W. Fellner. “Interaction Taxonomy for Tracking of User Actions in Visual Analytics Applications”. In: *Handbook of Human Centric Visualization*. Springer, 2014, pp. 653–670. DOI: 10.1007/978-1-4614-7485-2_26 (cit. on p. 140).

[197] Dirk J. Lehmann, Fritz Kemmler, Tatsiana Zhyhalava, Marco Kirschke, and Holger Theisel. “Visualnostics: Visual Guidance Pictograms for Analyzing Projections of High-dimensional Data”. In: *Computer Graphics Forum* 34.3 (2015), pp. 291–300. DOI: 10.1111/cgf.12641 (cit. on pp. 25, 52).

[198] Fritz Lekschas, Xinyi Zhou, Wei Chen, Nils Gehlenborg, Benjamin Bach, and Hanspeter Pfister. “A Generic Framework and Library for Exploration of Small Multiples through Interactive Piling”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2021), pp. 358–368. DOI: 10.1109/TVCG.2020.3028948 (cit. on pp. 140, 160).

[199] Sylvain Lespinats and Michaël Aupetit. “CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings”. In: *Computer Graphics Forum* 30.1 (2011), pp. 113–125. DOI: 10.1111/J.1467-8659.2010.01835.X (cit. on p. 51).

[200] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. “UpSet: Visualization of Intersecting Sets”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1983–1992. DOI: 10.1109/TVCG.2014.2346248 (cit. on pp. 48, 49).

[201] Xiaohui Lin, Fufang Yang, Lina Zhou, Peiyuan Yin, Hongwei Kong, Wenbin Xing, Xin Lu, Lewen Jia, Quancai Wang, and Guowang Xu. “A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information”. In: *Journal of Chromatography B* 910 (2012). Chemometrics in Chromatography, pp. 149–155. DOI: 10.1016/j.jchromb.2012.05.020 (cit. on p. 112).

[202] Gary Lincoff and National Audubon Society. *National Audubon Society field guide to North American mushrooms*. Audubon Society field guide series. Knopf: Distributed by Random House New York, 1981. ISBN: 978-0-394-51992-0 (cit. on pp. 65, 66, 70).

[203] Yan Liu and Gavriel Salvendy. “Design and evaluation of visualization support to facilitate decision trees classification”. In: *International Journal of Man-Machine Studies* 65.2 (2007), pp. 95–110. DOI: 10.1016/j.ijhcs.2006.07.005 (cit. on p. 114).

[204] Kecheng Lu, Khairi Reda, Oliver Deussen, and Yunhai Wang. “Interactive Context-Preserving Color Highlighting for Multiclass Scatterplots”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Ed. by Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson. ACM, 2023, 823:1–823:15. DOI: 10.1145/3544548.3580734 (cit. on p. 3).

[205] Richard C. Lupton and Julian M. Allwood. “Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use”. In: *Resources, Conservation and Recycling* 124 (2017), pp. 141–151. ISSN: 0921-3449. DOI: https://doi.org/10.1016/j.resconrec.2017.05.002 (cit. on p. 24).

[206] Mathias Lux and Savvas A. Chatzichristofis. “Lire: lucene image retrieval: an extensible java CBIR library”. In: *Proceedings of the 16th International Conference on Multimedia*. ACM, Oct. 2008, pp. 1085–1088. DOI: 10.1145/1459359.1459577 (cit. on p. 121).

[207] Verena Lyding, Ekaterina Lapshinova-Koltunski, Stefania Degaetano-Ortlieb, Henrik Dittmann, and Chris Culy. “Visualising Linguistic Evolution in Academic Discourse”. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Ed. by Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokić, and Michael Cysouw. Association for Computational Linguistics, Apr. 2012, pp. 44–48 (cit. on pp. 79, 85).

[208] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605 (cit. on pp. 74, 150, 161).

[209] Frank Maddix. *Human-computer Interaction: Theory and Practice*. Ellis Horwood series in computers and their applications. Ellis Horwood, 1990. ISBN: 978-0-13-446220-2 (cit. on p. 1).

[210] Prasanta Chandra Mahalanobis. “On the generalized distance in statistics”. In: *Proceedings of the National Institute of Sciences* 2 (1936), pp. 49–55 (cit. on p. 113).

[211] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, 2001. ISBN: 978-0-262-13360-9 (cit. on p. 78).

[212] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. “Building a Large Annotated Corpus of English: The Penn Treebank”. In: *Computational Linguistics* 19.2 (1993), pp. 313–330 (cit. on p. 82).

[213] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: `1802.03426 [stat.ML]` (cit. on p. 74).

[214] Maroua Mehri, Ramzi Chaieb, Karim Kalti, Pierre Héroux, Rémy Mullot, and Najoua Essoukri Ben Amara. “A Comparative Study of Two State-of-the-Art Feature Selection Algorithms for Texture-Based Pixel-Labeling Task of Ancient Documents”. In: *Journal of Imaging* 4.8 (2018), p. 97. DOI: `10.3390/jimaging4080097` (cit. on pp. 112, 137).

[215] Wouter Meulemans, Nathalie Henry Riche, Bettina Speckmann, Basak Alper, and Tim Dwyer. “KelpFusion: A Hybrid Set Visualization Technique”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.11 (2013), pp. 1846–1858. DOI: `10.1109/TVCG.2013.76` (cit. on p. 49).

[216] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. “Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2536–2545. DOI: `10.1109/TVCG.2012.199` (cit. on p. 49).

[217] Daniele Micci-Barreca. “A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems”. In: *ACM SIGKDD Explorations Newsletter* 3.1 (2001), pp. 27–32. DOI: `10.1145/507533.507538` (cit. on p. 2).

[218] George A. Miller. “Human memory and the storage of information”. In: *IRE Transactions on Information Theory* 2.3 (1956), pp. 129–137. DOI: `10.1109/TIT.1956.1056815` (cit. on pp. 30, 45).

[219] George A. Miller. “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. In: *Psychological Review* (1956), pp. 81–97. DOI: `10.1037/h0043158` (cit. on pp. 30, 45).

[220] George A. Miller. “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. In: *Psychological Review* (1956), pp. 81–97 (cit. on p. 72).

[221] Matthias Miller, Julius Rauscher, Daniel A. Keim, and Mennatallah El-Assady. “CorpusVis: Visual Analysis of Digital Sheet Music Collections”. In: *Computer Graphics Forum* 41.3 (2022), pp. 283–294. DOI: 10.1111/cgf.14540 (cit. on pp. 146, 148–151, 155).

[222] Rosane Minghim, Fernando Vieira Paulovich, and Alneu de Andrade Lopes. “Content-based text mapping using multi-dimensional projections for exploration of document collections”. In: *Visualization and Data Analysis*. Vol. 6060. SPIE Proceedings. 2006, 60600S. DOI: 10.1117/12.650880 (cit. on p. 162).

[223] Sebastian Mittelstädt, Dominik Jäckle, Florian Stoffel, and Daniel A. Keim. “ColorCAT: Guided Design of Colormaps for Combined Analysis Tasks”. In: *Proceedings of the 17th Eurographics Conference on Visualization*. Eurographics, 2015, pp. 115–119. DOI: 10.2312/eurovisshort.20151135 (cit. on pp. 3, 30).

[224] Tyler Moore. “On the harms arising from the Equifax data breach of 2017”. In: *International Journal of Critical Infrastructure Protection* 19 (2017), pp. 47–48. DOI: 10.1016/j.ijcip.2017.10.004 (cit. on p. 96).

[225] Cristina Morariu, Adrien Bibal, Rene Cutura, Benoît Frénay, and Michael Sedlmair. “Predicting User Preferences of Dimensionality Reduction Embedding Quality”. In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (Jan. 2023), pp. 745–755. DOI: 10.1109/TVCG.2022.3209449 (cit. on p. 51).

[226] Bryan S. Morse. *Lecture 9: Shape Description (Regions)*. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/region-props-and-moments.pdf, last accessed 2023-11-14. Brigham Young University (cit. on p. 121).

[227] Juliane Müller, Laura Garrison, Philipp Ulbrich, Stefanie Schreiber, Stefan Bruckner, Helwig Hauser, and Steffen Oeltze-Jafra. “Integrated Dual Analysis of Quantitative and Qualitative High-Dimensional Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.6 (2021), pp. 2953–2966. DOI: 10.1109/TVCG.2021.3056424 (cit. on pp. 136, 146, 148, 149, 151, 152, 154, 157, 162).

[228] Tamara Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. A K Peters, 2014. ISBN: 978-1-466-50891-0 (cit. on pp. 3, 4, 157, 162).

[229] Eun Ju Nam, Yiping Han, Klaus Mueller, Alla Zelenyuk, and Dan Imre. “ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data”. In: *Proceedings of the 2nd IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 75–82. DOI: 10.1109/VAST.2007.4388999 (cit. on p. 114).

[230] Quynh Quang Ngo, Frederik L. Dennig, Daniel A. Keim, and Michael Sedlmair. “Machine learning meets visualization – Experiences and lessons learned”. In: *it - Information Technology* 64.4-5 (2022), pp. 169–180. DOI: 10.1515/itit-2022-0034 (cit. on p. 17).

[231] Dang Tuan Nhon and Leland Wilkinson. “PixSearcher: Searching Similar Images in Large Image Collections through Pixel Descriptors”. In: *Proceedings of Advances in Visual Computing - 10th International Symposium*. Springer, 2014, pp. 726–735. ISBN: 978-3-319-14364-4. DOI: 10.1007/978-3-319-14364-4_70 (cit. on p. 112).

[232] Carolina Nobre, Nils Gehlenborg, Hilary Coon, and Alexander Lex. “Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.3 (2019), pp. 1543–1558. DOI: 10.1109/TVCG.2018.2811488 (cit. on p. 100).

[233] Carolina Nobre, Marc Streit, and Alexander Lex. “Juniper: A Tree+Table Approach to Multivariate Graph Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 544–554. DOI: 10.1109/TVCG.2018.2865149 (cit. on pp. 100, 101).

[234] Martin Nöllenburg. “Geographic Visualization”. In: *Human-Centered Visualization Environments*. Springer, 2007, pp. 257–294. ISBN: 978-3-540-71949-6. DOI: 10.1007/978-3-540-71949-6_6 (cit. on p. 1).

[235] Luis Gustavo Nonato and Michaël Aupetit. “Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2019), pp. 2650–2673. DOI: 10.1109/TVCG.2018.2846735 (cit. on pp. 51, 137, 140, 147, 167).

[236] Lenka Nováková and Olga Štěpánková. “Visualization of Trends Using RadViz”. In: *Journal of Intelligent Information Systems*. Springer, 2009, pp. 56–65. ISBN: 978-3-642-04125-9. DOI: 10.1007/s10844-011-0157-4 (cit. on pp. 149, 150).

[237] Patrick Paetzold, Rebecca Kehlbeck, Hendrik Strobelt, Yumeng Xue, Sabine Storandt, and Oliver Deussen. “RectEuler: Visualizing Intersecting Sets using Rectangles”. In: *Computer Graphics Forum* 42.3 (2023), pp. 87–98. DOI: 10.1111/cgf.14814 (cit. on pp. 48, 63).

[238] J. Pagès. *Multiple Factor Analysis by Example Using R*. 1st ed. The R Series. Chapman & Hall/CRC, 2014. ISBN: 978-0-429-17108-6. DOI: 10.1201/b17700 (cit. on pp. 50, 149).

[239] Stephan Pajer, Marc Streit, Thomas Torsney-Weir, Florian Spechtenhauser, Torsten Möller, and Harald Piringer. “WeightLifter: Visual Weight Space Exploration for Multi-Criteria Decision Making”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 611–620. DOI: 10.1109/TVCG.2016.2598589 (cit. on p. 37).

[240] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191 (cit. on p. 110).

[241] Fernando Vieira Paulovich, Luis Gustavo Nonato, Rosane Minghim, and Haim Levkowitz. “Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.3 (2008), pp. 564–575. DOI: 10.1109/TVCG.2007.70443 (cit. on p. 67).

[242] Ignacio Pérez-Messina, Davide Ceneda, Mennatallah El-Assady, Silvia Miksch, and Fabian Sperrle. “A Typology of Guidance Tasks in Mixed-Initiative Visual Analytics Environments”. In: *Computer Graphics Forum* 41.3 (2022), pp. 465–476. DOI: 10.1111/cgf.14555 (cit. on pp. 137, 168, 174).

[243] José G. Pérez-Silva, Miguel Araujo-Voces, and Víctor Quesada. “nVenn: generalized, quasi-proportional Venn and Euler diagrams”. In: *Bioinformatics* 34.13 (2018), pp. 2322–2324. DOI: 10.1093/bioinformatics/bty109 (cit. on p. 49).

[244] Vung Pham and Tommy Dang. “CVExplorer: Multidimensional Visualization for Common Vulnerabilities and Exposures”. In: *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2018, pp. 1296–1301. DOI: 10.1109/BigData.2018.8622092 (cit. on p. 97).

[245] Mike Pittenger. “Open source security analysis - The state of open source security in commercial applications”. In: *Black Duck Software, Technical Report* (2016). https://www.vojtechruzicka.com/bf4dd32d5823c258c319cced38727dce/OSSAReport.pdf, last accessed 2023-11-14 (cit. on p. 96).

[246] Catherine Plaisant. “The challenge of information visualization evaluation”. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM, May 2004, pp. 109–116. DOI: 10.1145/989863.989880 (cit. on pp. 5, 128).

[247] Henrik Plate, Serena Elisa Ponta, and Antonino Sabetta. “Impact assessment for vulnerabilities in open-source software libraries”. In: *Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution*. IEEE, 2015, pp. 411–420. DOI: 10.1109/ICSM.2015.7332492 (cit. on p. 96).

[248] Georg Pölzlbauer. “Survey and comparison of quality measures for self-organizing maps”. In: *Proceedings of the Fifth Workshop on Data Analysis*. 2004, pp. 67–82 (cit. on p. 127).

[249] David Pomerenke, Frederik L. Dennig, Daniel A. Keim, Johannes Fuchs, and Michael Blumenschein. *Replication Data for: "Slope-Dependent Rendering of Parallel Coordinates to Reduce Density Distortion and Ghost Clusters"*. Version V2. https://osf.io/sy3dv/ (alternative repository). 2022. DOI: 10.18419/darus-3060 (cit. on p. 18).

[250] David Pomerenke, Frederik L. Dennig, Daniel A. Keim, Johannes Fuchs, and Michael Blumenschein. “Slope-Dependent Rendering of Parallel Coordinates to Reduce Density Distortion and Ghost Clusters”. In: *Proceedings of the 30th IEEE Visualization Conference*. IEEE, 2019, pp. 86–90. DOI: 10.1109/VISUAL.2019.8933706 (cit. on pp. 18, 29, 33).

[251] Serena Elisa Ponta, Henrik Plate, and Antonino Sabetta. “Detection, assessment and mitigation of vulnerabilities in open source dependencies”. In: *Empirical Software Engineering* 25.5 (2020), pp. 3175–3215. DOI: 10.1007/s10664-020-09830-x (cit. on pp. 96, 98).

[252] Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. “Projections as visual aids for classification system design”. In: *Information Visualisation* 17.4 (2018), pp. 282–305. DOI: 10.1177/1473871617713337 (cit. on pp. 136, 146, 148–152, 154, 155, 163, 164, 166, 169).

[253] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. “Interactive Sankey Diagrams”. In: *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2005, pp. 233–240. DOI: 10.1109/INFVIS.2005.1532152 (cit. on p. 24).

[254] Peter J. Rodgers, Gem Stapleton, and Peter Chapman. “Visualizing Sets with Linear Diagrams”. In: *ACM Transactions on Computer-Human Interaction* 22.6 (2015), 27:1–27:39. DOI: 10.1145/2810012 (cit. on p. 49).

[255] Nils Rodrigues, Frederik L. Dennig, Vincent Brandt, Daniel A. Keim, and Daniel Weiskopf. “Comparative Evaluation of Animated Scatter Plot Transitions”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.6 (2024), pp. 2929–2941. DOI: 10.1109/TVCG.2024.3388558 (cit. on p. 17).

[256] Kristopher Rogers, Janet Wiles, Scott Heath, Kristyn Hensby, and Jonathon Taufatofua. “Discovering Patterns of Touch: A Case Study for Visualization-Driven Analysis in Human-Robot Interaction”. In: *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE, 2016, pp. 499–500. DOI: 10.1109/HRI.2016.7451825 (cit. on pp. 24, 38, 40–43, 65, 66).

[257] Eiríkur Rögnvaldsson. “Word order variation in the VP in Old Icelandic”. In: *Working Papers in Scandinavian Syntax* 58 (1996), pp. 55–86 (cit. on pp. 80, 86).

[258] Eiríkur Rögnvaldsson, Anton Karl Ingason, and Einar Freyr Sigurðsson. “Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC)”. In: *Language Variation Infrastructure, Oslo Studies in Language* 3 (2 2011), pp. 97–112 (cit. on p. 80).

[259] Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. “Towards Tracking Semantic Change by Visual Analytics”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. The Association for Computer Linguistics, 2011, pp. 305–310 (cit. on p. 79).

[260] Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. “Lexical Semantics and Distribution of Suffixes - A Visual Analysis”. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Ed. by Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokić, and Michael Cysouw. Association for Computational Linguistics, Apr. 2012, pp. 7–15 (cit. on p. 79).

[261] Geraldine E. Rosario, Elke A. Rundensteiner, David C. Brown, and Matthew O. Ward. "Mapping Nominal Values to Numbers for Effective Visualization". In: *Proceedings of the 9th IEEE Symposium on Information Visualization*. IEEE, 2003, pp. 113–120. DOI: 10.1109/INFVIS.2003.1249016 (cit. on pp. 3, 22, 48, 50, 174).

[262] Peter Rottmann, Markus Wallinger, Annika Bonerath, Sven Gedicke, Martin Nöllenburg, and Jan-Henrik Haunert. "MosaicSets: Embedding Set Systems into Grid Graphs". In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2023), pp. 875–885. DOI: 10.1109/TVCG.2022.3209485 (cit. on p. 61).

[263] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. "Image Retrieval: Current Techniques, Promising Directions, and Open Issues". In: *Journal of Visual Communication and Image Representation* 10.1 (1999), pp. 39–62. DOI: 10.1006/jvci.1999.0413 (cit. on pp. 120, 121).

[264] Tobias Ruppert, Michael Staab, Andreas Bannach, Hendrik Lücke-Tieke, Jürgen Bernard, Arjan Kuijper, and Jörn Kohlhammer. "Visual Interactive Creation and Validation of Text Clustering Workflows to Explore Document Collections". In: *Visualization and Data Analysis*. Society for Imaging Science and Technology, 2017, pp. 46–57. DOI: 10.2352/ISSN.2470-1173.2017.1.VDA-388 (cit. on p. 114).

[265] Dominik Sacha, Matthias Kraus, Jürgen Bernard, Michael Behrisch, Tobias Schreck, Yuki Asano, and Daniel A. Keim. "SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 120–130. DOI: 10.1109/TVCG.2017.2744805 (cit. on pp. 5, 114, 115, 125, 133).

[266] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey P. Ellis, and Daniel A. Keim. "Knowledge Generation Model for Visual Analytics". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1604–1613. DOI: 10.1109/TVCG.2014.2346481 (cit. on pp. 1, 5, 114, 137, 139).

[267] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John Aldo Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 241–250. DOI: 10.1109/TVCG.2016.2598495 (cit. on pp. 147, 163).

[268] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. "Robust Feature Selection Using Ensemble Feature Selection Techniques". In: *Machine Learning and Knowledge Discovery in Databases*. Vol. 5212. Lecture Notes in Computer Science. Springer, 2008, pp. 313–325. DOI: 10.1007/978-3-540-87481-2_21 (cit. on p. 112).

[269] Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. "Evaluating Color Descriptors for Object and Scene Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1582–1596. DOI: 10.1109/TPAMI.2009.154 (cit. on p. 121).

[270] Christin Schätzle, Miriam Butt, and Kristina Kotcheva. "The Diachrony of Dative Subjects and the Middle in Icelandic : A Corpus Study". In: *Proceedings of the LFG15 Conference*. CSLI Publications, 2015, pp. 357–377 (cit. on p. 89).

[271] Christin Schätzle, Frederik L. Dennig, Michael Blumenschein, Daniel A. Keim, and Miriam Butt. "Visualizing Linguistic Change as Dimension Interactions". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics, Aug. 2019, pp. 272–278. DOI: 10.18653/v1/W19-4734 (cit. on pp. 16, 38, 43, 77).

[272] Christin Schätzle, Michael Hund, Frederik L. Dennig, Miriam Butt, and Daniel A. Keim. "HistoBankVis: Detecting Language Change via Data Visualization". In: *Proceedings of the NoDaLiDa 2017 Workshop Processing Historical Language*. NEALT Proceedings Series 32. Association for Computational Linguistics, 2017, pp. 32–39 (cit. on pp. 16, 77).

[273] Christin Schätzle and Dominik Sacha. "Visualizing Language Change: Dative Subjects in Icelandic". In: *Proceedings of the LREC 2016 Workshop VisLR II: Visualization as added value in the development, use and evaluation of Language Resources*. 2016 (cit. on p. 79).

[274] Jörn Schneidewind, Mike Sips, and Daniel A. Keim. "Pixnostics: Towards Measuring the Value of Visualization". In: *Proceedings of the 1st IEEE Symposium On Visual Analytics Science And Technology*. IEEE, 2006, pp. 199–206. DOI: 10.1109/VAST.2006.261423 (cit. on pp. 25, 52).

[275] Matthias Schonlau. "Visualizing categorical data arising in the health sciences using hammock plots". In: *Proceedings of the Joint Statistical Meetings, Section on Statistical Graphics* (2003). http://www.schonlau.net/publication/03jsm_hammockplot.pdf, last accessed 2020-09-30 (cit. on p. 24).

[276] Tobias Schreck, Dieter W. Fellner, and Daniel A. Keim. "Towards automatic feature vector optimization for multimedia applications". In: *Proceedings of the 2008 ACM Symposium on Applied Computing*. Mar. 2008, pp. 1197–1201. DOI: 10.1145/1363686.1363964 (cit. on p. 112).

[277] Jessica Zeitz Self, Radha Krishnan Vinayagam, J. T. Fry, and Chris North. "Bridging the gap between user intention and model parameters for human-in-the-loop data analytics". In: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. ACM, 2016, p. 3. DOI: 10.1145/2939502.2939505 (cit. on pp. 136, 146, 148–151, 153, 155, 162, 167).

[278] Claude E. Shannon. "A mathematical theory of communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on pp. 25, 29, 36).

[279] Ben Shneiderman. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*. IEEE, 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307 (cit. on p. 100).

[280] Ben Shneiderman. “Tree Visualization with Tree-Maps: 2-d Space-Filling Approach”. In: *ACM Transactions on Graphics* 11.1 (1992), pp. 92–99. DOI: 10.1145/102377.115768 (cit. on p. 73).

[281] Halldór Ármann Sigurðsson. “V1 Declaratives and Verb Raising in Icelandic”. In: *Modern Icelandic Syntax*. Brill, 1990, pp. 41–69. ISBN: 978-90-04-37323-5 (cit. on pp. 80, 89).

[282] Halldór Ármann Sigurðsson. “Verbal Syntax and Case in Icelandic. In a Comparative GB Approach”. PhD thesis. University of Lund, 1989 (cit. on p. 90).

[283] Hannah Snyder. “Literature review as a research methodology: An overview and guidelines”. In: *Journal of Business Research* 104 (2019), pp. 333–339. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2019.07.039 (cit. on p. 142).

[284] Amitpal Singh Sohal, Sunil Kumar Gupta, and Hardeep Singh. “Trust in Open Source Software Development Communities: A Comprehensive Analysis”. In: *International Journal of Open Source Software and Processes* 9.4 (2018), pp. 1–19. DOI: 10.4018/IJOSSP.2018100101 (cit. on p. 96).

[285] Jan-Tobias Sohns, Michaela Schmitt, Fabian Jirasek, Hans Hasse, and Heike Leitte. “Attribute-based Explanation of Non-Linear Embeddings of High-Dimensional Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2022), pp. 540–550. DOI: 10.1109/TVCG.2021.3114870 (cit. on p. 51).

[286] Ian Sommerville. *Software Engineering*. 9th ed. it : Informatik. Addison-Wesley, 2010. ISBN: 978-3-8273-7257-4 (cit. on p. 96).

[287] Sonatype Inc. *State of the Software Supply Chain (2020)*. Technical Report. https://de.sonatype.com/resources/white-paper-state-of-the-software-supply-chain-2020, last accessed 2021-06-24. 2020 (cit. on p. 96).

[288] Thorvald Sørenson. “A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons”. In: *Kongelige Danske Videnskabernes Selskab*. Biologiske skrifter 5.4 (1948), pp. 1–34 (cit. on p. 54).

[289] Aurea Soriano-Vargas, Bernd Hamann, and Maria Cristina Ferreira de Oliveira. “TV-MV Analytics: A visual analytics framework to explore time-varying multivariate data”. In: *Information Visualization* 19.1 (2020). DOI: 10.1177/1473871619858937 (cit. on pp. 136, 146, 149, 151, 155, 159).

[290] Charles Spearman. “The Proof and Measurement of Association between Two Things”. In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101. DOI: 10.2307/1412159 (cit. on p. 67).

[291] Robert Spence. *Information Visualization - An Introduction*. Springer, 2014. ISBN: 978-3-319-07340-8. DOI: 10.1007/978-3-319-07341-5 (cit. on p. 4).

[292] Michael Spenke and Christian Beilken. "Visualization of Trees as Highly Compressed Tables with InfoZoom". In: *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2003, pp. 122–123 (cit. on pp. 22, 24, 48).

[293] Fabian Sperrle, Davide Ceneda, and Mennatallah El-Assady. "Lotse: A Practical Framework for Guidance in Visual Analytics". In: *IEEE Transactions on Visualization and Computer Graphics* 29.1 (2023), pp. 1124–1134. DOI: 10.1109/TVCG.2022.3209393 (cit. on pp. 168, 174).

[294] John T. Stasko and Eugene Zhang. "Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations". In: *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, 2000, pp. 57–65. DOI: 10.1109/INFVIS.2000.885091 (cit. on p. 24).

[295] Christian Alexander Steinparz, Andreas P. Hinterreiter, Holger Stitz, and Marc Streit. "Visualization of Rubik's Cube Solution Algorithms". In: *Proceedings of the 10th International EuroVis Workshop on Visual Analytics*. Eurographics, 2019, pp. 19–23. DOI: 10.2312/eurova.20191119 (cit. on p. 167).

[296] Hendrik Strobelt, Enrico Bertini, Joachim Braun, Oliver Deussen, Ulrich Groth, Thomas U. Mayer, and Dorit Merhof. "HiTSEE KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform". In: *BMC Bioinformatics* 13.S-8 (2012), S4. DOI: 10.1186/1471-2105-13-S8-S4 (cit. on p. 112).

[297] Zdenek Sulc and Hana Rezanková. "Comparison of Similarity Measures for Categorical Data in Hierarchical Clustering". In: *Journal of Classification* 36.1 (2019), pp. 58–72. DOI: 10.1007/S00357-019-09317-5 (cit. on pp. 2, 54).

[298] Dezydery Szymkiewicz. "Une conlribution statistique à la géographie floristique". In: *Acta Societatis Botanicorum Poloniae* 11.3 (1934), pp. 249–265. DOI: https://doi.org/10.5586/asbp.1934.012 (cit. on pp. 69, 70).

[299] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. "Textural Features Corresponding to Visual Perception". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 8.6 (1978), pp. 460–473. DOI: 10.1109/TSMC.1978.4309999 (cit. on p. 121).

[300] Soon Tee Teoh and Kwan-Liu Ma. "PaintingClass: interactive construction, visualization and exploration of decision trees". In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2003, pp. 667–672. DOI: 10.1145/956750.956837 (cit. on pp. 22, 48).

[301] Michael Thane, Kai M. Blum, and Dirk J. Lehmann. "CatNetVis: Semantic Visual Exploration of Categorical High-Dimensional Data with Force-Directed Graph Layouts". In: *Proceedings of the 25th Eurographics Conference on Visualization*. Ed. by Thomas Hoellt, Wolfgang Aigner, and Bei Wang. Eurographics, 2023. ISBN: 978-3-03868-219-6. DOI: `10.2312/evs.20231049` (cit. on p. 50).

[302] Roberto Therón and Laura Fontanillo. "Diachronic-information visualization in historical dictionaries". In: *Information Visualization* 14.2 (2015), pp. 111–136. DOI: `10.1177/1473871613495844` (cit. on pp. 79, 85).

[303] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE, 2005. ISBN: 978-0-7695-2323-1 (cit. on p. 1).

[304] Höskuldur Thráinsson. "Icelandic". In: *The Germanic Languages*. Routledge Language Family Descriptions. Routledge, 1994, pp. 142–189. ISBN: 0-415-05768-X (cit. on pp. 80, 82).

[305] *TITUS (Thesaurus Indogermanischer Text- und Sprachmaterialien)*. Online. `https://titus.uni-frankfurt.de/indexd.htm`, accessed 2024-04-17. 1987 (cit. on p. 78).

[306] Christian Tominski and Heidrun Schumann. *Interactive Visual Data Analysis*. AK Peters Visualization Series. CRC Press, 2020. ISBN: 978-1-4987-5398-2. DOI: `10.1201/9781315152707` (cit. on p. 4).

[307] Edward R. Tufte. *Envisioning information*. Graphics Press, 1990. ISBN: 978-0-9613921-1-6 (cit. on p. 3).

[308] Cagatay Turkay, Peter Filzmoser, and Helwig Hauser. "Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2591–2599. DOI: `10.1109/TVCG.2011.178` (cit. on pp. 136, 137, 142, 146, 148, 149, 151, 152, 155, 162, 163, 168).

[309] Cagatay Turkay, Erdem Kaya, Selim Balcisoy, and Helwig Hauser. "Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 131–140. DOI: `10.1109/TVCG.2016.2598470` (cit. on pp. 136, 142, 146, 148, 149, 151, 155, 157, 167).

[310] Cagatay Turkay, Alexander Lex, Marc Streit, Hanspeter Pfister, and Helwig Hauser. "Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX". In: *IEEE Computer Graphics and Applications* 34.2 (2014), pp. 38–47. DOI: `10.1109/MCG.2014.1` (cit. on pp. 142, 146, 148, 149, 151, 154, 155).

[311] Cagatay Turkay, Arvid Lundervold, Astri J. Lundervold, and Helwig Hauser. “Representative Factor Generation for the Interactive Visual Analysis of High-Dimensional Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2621–2630. DOI: 10.1109/TVCG.2012.256 (cit. on pp. 136, 142, 146, 148–152, 154).

[312] Cagatay Turkay, Aidan Slingsby, Helwig Hauser, Jo Wood, and Jason Dykes. “Attribute Signatures: Dynamic Visual Summaries for Analyzing Multivariate Geographical Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2033–2042. DOI: 10.1109/TVCG.2014.2346265 (cit. on pp. 136, 142, 146, 148–151, 153–155, 167).

[313] Cagatay Turkay, Aidan Slingsby, Kaisa Lahtinen, Sarah Butt, and Jason Dykes. “Supporting theoretically-grounded model building in the social sciences through interactive visualisation”. In: *Neurocomputing* 268 (2017), pp. 153–163. DOI: 10.1016/j.neucom.2016.11.087 (cit. on pp. 142, 146, 148, 149, 151, 155, 157).

[314] Fan-Yin Tzeng, Eric B. Lum, and Kwan-Liu Ma. “A Novel Interface for Higher-Dimensional Classification of Volume Data”. In: *Proceedings of the 14th IEEE Visualization Conference*. IEEE, 2003, pp. 505–512. DOI: 10.1109/VISUAL.2003.1250413 (cit. on p. 145).

[315] Alfred Ultsch. “Data mining and knowledge discovery with emergent Self-Organizing Feature Maps for multivariate time series”. In: *Kohonen Maps*. Elsevier, July 1999 (cit. on pp. 114, 127).

[316] Susan VanderPlas and Heike Hofmann. “Signs of the Sine Illusion—Why We Need to Care”. In: *Journal of Computational and Graphical Statistics* 24.4 (2015), pp. 1170–1190. DOI: 10.1080/10618600.2014.951547 (cit. on p. 50).

[317] Jarkko Venna and Samuel Kaski. “Visualizing gene interaction graphs with local multidimensional scaling”. In: *14th European Symposium on Artificial Neural Networks*. 2006, pp. 557–562 (cit. on pp. 65, 67).

[318] William M. Waggener. *Pulse Code Modulation Techniques*. Springer, 1994. ISBN: 978-0-442-01436-0 (cit. on p. 54).

[319] Markus Wagner, Fabian Fischer, Robert Luh, Andrea Haberson, Alexander Rind, Daniel A. Keim, and Wolfgang Aigner. “A Survey of Visualization Systems for Malware Analysis”. In: *Proceedings of the 17th Eurographics Conference on Visualization*. Eurographics, 2015, pp. 105–125. DOI: 10.2312/eurovisstar.20151114 (cit. on p. 97).

[320] Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. *Icelandic Parsed Historical Corpus (IcePaHC) 0.9*. Version: 0.9, http://hdl.handle.net/20.500.12537/62, last accessed 2023-11-17. 2011 (cit. on p. 78).

[321] Lei Wang, Xin Zhao, and Arie E. Kaufman. “Modified Dendrogram of Attribute Space for Multidimensional Transfer Function Design”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.1 (2012), pp. 121–131. DOI: 10.1109/TVCG.2011.23 (cit. on p. 145).

[322] Yunhai Wang, Wei Chen, Jian Zhang, Tingxin Dong, Guihua Shan, and Xuebin Chi. “Efficient Volume Exploration Using the Gaussian Mixture Model”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.11 (2011), pp. 1560–1573. DOI: 10.1109/TVCG.2011.97 (cit. on p. 145).

[323] Zhi Wang, Yan Zhang, Zhichao Chen, Huan Yang, Yaxin Sun, Jianmin Kang, Yong Yang, and Xiaojun Liang. “Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image”. In: *Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, July 2016, pp. 755–758. DOI: 10.1109/IGARSS.2016.7729190 (cit. on p. 112).

[324] Matthew O. Ward. “Linking and Brushing”. In: *Encyclopedia of Database Systems*. Springer, 2009, pp. 1623–1626. DOI: 10.1007/978-0-387-39940-9_1129 (cit. on p. 160).

[325] Colin Ware. *Information Visualization: Perception for Design*. 3rd. Morgan Kaufmann, 2012. ISBN: 978-0-12-381464-7 (cit. on p. 4).

[326] Colin Ware, Helen C. Purchase, Linda Colpoys, and Matthew McGill. “Cognitive measurements of graph aesthetics”. In: *Information Visualization* 1.2 (2002), pp. 103–110. DOI: 10.1057/palgrave.ivs.9500013 (cit. on pp. 29, 31, 33, 34, 45).

[327] Chris E. Weaver. “Cross-Filtered Views for Multidimensional Visual Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.2 (2010), pp. 192–204. DOI: 10.1109/TVCG.2009.94 (cit. on p. 145).

[328] Jishang Wei, Hongfeng Yu, Ray W. Grout, Jacqueline H. Chen, and Kwan-Liu Ma. “Dual space analysis of turbulent combustion particle data”. In: *Proceedings of the IEEE Pacific Visualization Symposium*. IEEE, 2011, pp. 91–98. DOI: 10.1109/PACIFICVIS.2011.5742377 (cit. on p. 145).

[329] Jarke J. van Wijk. “The Value of Visualization”. In: *Proceedings of the 16th IEEE Visualization Conference*. IEEE, 2005, pp. 79–86. DOI: 10.1109/VISUAL.2005.1532781 (cit. on pp. 5, 138).

[330] Leland Wilkinson, Anushka Anand, and Robert L. Grossman. “Graph-Theoretic Scagnostics”. In: *Proceedings of the IEEE Symposium on Information Visualization*. IEEE, Oct. 2005, pp. 157–164. DOI: 10.1109/INFVIS.2005.1532142 (cit. on pp. 4, 25, 52, 111).

[331] Kent Wittenburg, Tom Lanning, Michael Heinrichs, and Michael Stanton. “Parallel bargrams for consumer-based information exploration and choice”. In: *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2001, pp. 51–60. DOI: 10.1145/502348.502357 (cit. on pp. 22, 24, 48, 50).

[332] Yuki Yano, Raula Gaikovina Kula, Takashi Ishio, and Katsuro Inoue. “VerXCombo: an interactive data visualization of popular library version combinations”. In: *Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension*. IEEE, 2015, pp. 291–294. DOI: 10.1109/ICPC.2015.43 (cit. on p. 66).

[333] Ji Soo Yi, Youn ah Kang, John T. Stasko, and Julie A. Jacko. “Toward a Deeper Understanding of the Role of Interaction in Information Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1224–1231. DOI: 10.1109/TVCG.2007.70515 (cit. on p. 140).

[334] Xiaoru Yuan, Donghao Ren, Zuchao Wang, and Cong Guo. “Dimension Projection Matrix/Tree: Interactive Subspace Visual Exploration and Analysis of High Dimensional Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2625–2633. DOI: 10.1109/TVCG.2013.150 (cit. on pp. 136, 142, 146, 148–151, 153–155, 158, 162–165, 167).

[335] Annie Zaenen, Joan Maling, and Höskuldur Thráinsson. “Case and grammatical functions: The Icelandic passive”. In: *Natural Language & Linguistic Theory* 3 (4 1985), pp. 441–483. DOI: 10.1007/BF00133285 (cit. on p. 90).

[336] Germain Garcia Zanabria, Luis Gustavo Nonato, and Erick Gomez Nieto. “iStar (i*): An interactive star coordinates approach for high-dimensional data exploration”. In: *Computer Graphics* 60 (2016), pp. 107–118. DOI: 10.1016/j.cag.2016.08.007 (cit. on pp. 146, 148–151, 155).

[337] David Cheng Zarate, Pierre Le Bodic, Tim Dwyer, Graeme Gange, and Peter J. Stuckey. “Optimal Sankey Diagrams Via Integer Programming”. In: *IEEE Pacific Visualization Symposium*. IEEE, 2018, pp. 135–139. DOI: 10.1109/PacificVis.2018.00025 (cit. on pp. 41, 42).

[338] Chong Zhang, Yang Chen, Jing Yang, and Zhengcong Yin. “An association rule based approach to reducing visual clutter in parallel sets”. In: *Visual Informatics* 3.1 (2019), pp. 48–57. DOI: 10.1016/j.visinf.2019.03.006 (cit. on pp. 23, 25).

[339] Zhiyuan Zhang, Kevin T. McDonnell, Erez Zadok, and Klaus Mueller. “Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map”. In: *IEEE Transactions on Visualization and Computer Graphics* 21.2 (2015), pp. 289–303. DOI: 10.1109/TVCG.2014.2350494 (cit. on p. 145).

[340] Jieqiong Zhao, Morteza Karimzadeh, Ali Masjedi, Taojun Wang, Xiwen Zhang, Melba M. Crawford, and David S. Ebert. “FeatureExplorer: Interactive Feature Selection and Exploration of Regression Models for Hyperspectral Images”. In: *Proceedings of the 30th IEEE Visualization Conference*. IEEE, 2019, pp. 161–165. DOI: `10.1109/VISUAL.2019.8933619` (cit. on pp. 146, 148, 149, 151, 155).