A Tale of Two Approaches: Comparing Top-Down and Bottom-Up Strategies for Analyzing and Visualizing High-Dimensional Data

by

Aleksandar Anžel
born in Paraćin, Serbia

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF NATURAL SCIENCES (DR. RER. NAT.)
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
PHILIPPS-UNIVERSITÄT MARBURG

AUGUST, 2023

Reviewers:
Prof. Dr. Dominik Heider
Dr. habil. Georges Hattab

Examiners:
Prof. Dr.-Ing. Bernd Freisleben
Prof. Dr. Bernhard Seeger
Prof. Dr. Thorsten Thormählen

Date of Submission:	11.09.2023
Date of Disputation:	
Place of Publication:	Marburg
Year of Publication:	2023
University Identification Number:	1180

Printed on non-ageing paper according to DIN-ISO 9706.
The digital version of this document is PDF/A-1b compliant.

Declaration of Authorship

I, Aleksandar Anžel, declare that this dissertation entitled, "*A Tale of Two Approaches: Comparing Top-Down and Bottom-Up Strategies for Analyzing and Visualizing High-Dimensional Data*" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signature:

Date:

Supervisors: Prof. Dr. Dominik Heider, Dr. habil. Georges Hattab Aleksandar Anžel

A Tale of Two Approaches: Comparing Top-Down and Bottom-Up Strategies for Analyzing and Visualizing High-Dimensional Data

Abstract

The proliferation of high-throughput and sensory technologies in various fields has led to a considerable increase in data volume, complexity, and diversity. Traditional data storage, analysis, and visualization methods are struggling to keep pace with the growth of modern data sets, necessitating innovative approaches to overcome the challenges of managing, analyzing, and visualizing data across various disciplines.

One such approach is utilizing novel storage media, such as deoxyribonucleic acid (DNA), which presents efficient, stable, compact, and energy-saving storage option. Researchers are exploring the potential use of DNA as a storage medium for long-term storage of significant cultural and scientific materials.

In addition to novel storage media, scientists are also focussing on developing new techniques that can integrate multiple data modalities and leverage machine learning algorithms to identify complex relationships and patterns in vast data sets. These newly-developed data management and analysis approaches have the potential to unlock previously unknown insights into various phenomena and to facilitate more effective translation of basic research findings to practical and clinical applications.

Addressing these challenges necessitates different problem-solving approaches. Researchers are developing novel tools and techniques that require different viewpoints. Top-down and bottom-up approaches are essential techniques that offer valuable perspectives for managing, analyzing, and visualizing complex high-dimensional multi-modal data sets. This cumulative dissertation explores the

challenges associated with handling such data and highlights top-down, bottom-up, and integrated approaches that are being developed to manage, analyze, and visualize this data. The work is conceptualized in two parts, each reflecting the two problem-solving approaches and their uses in published studies. The proposed work showcases the importance of understanding both approaches, the steps of reasoning about the problem within them, and their concretization and application in various domains.

A Tale of Two Approaches: Comparing Top-Down and Bottom-Up Strategies for Analyzing and Visualizing High-Dimensional Data

Zusammenfassung

Die Verbreitung von Hochdurchsatz- und Sensortechnologien in verschiedenen Bereichen hat zu einem erheblichen Anstieg des Datenvolumens, der Komplexität und der Vielfalt geführt. Herkömmliche Methoden der Datenspeicherung, -analyse und -visualisierung können mit dem Wachstum moderner Datensätze nur schwer Schritt halten. Daher sind innovative Ansätze erforderlich, um die Herausforderungen bei der Verwaltung, Analyse und Visualisierung von Daten in verschiedenen Disziplinen zu bewältigen.

Ein solcher Ansatz ist die Verwendung neuartiger Speichermedien wie desoxyribonukleinsäure (DNA), die effiziente, stabile, kompakte und energiesparende Speichermöglichkeiten bieten. Forscher untersuchen den möglichen Einsatz von DNA als Speichermedium für die langfristige Aufbewahrung von bedeutenden kulturellen und wissenschaftlichen Materialien.

Neben neuartigen Speichermedien konzentriert sich die Forschung auch auf die Entwicklung neuer Techniken, die mehrere Datenmodalitäten integrieren und Algorithmen des maschinellen Lernens nutzen können, um komplexe Beziehungen und Muster in riesigen Datensätzen zu erkennen. Diese neu entwickelten Datenverwaltungs- und -analyseverfahren haben das Potenzial, bisher unbekannte Erkenntnisse über verschiedene Phänomene zu erschließen und eine effektivere Umsetzung von Ergebnissen der Grundlagenforschung in praktische und klinische Anwendungen zu ermöglichen.

Die Bewältigung dieser Herausforderungen erfordert unterschiedliche Problemlösungsansätze. Die Forscher entwickeln neue Instrumente und Techniken, die unterschiedliche Sichtweisen erfordern. Top-down- und Bottom-up-Ansätze sind wesentliche Techniken, die wertvolle Perspektiven für die Verwaltung, Analyse und Visualisierung komplexer hochdimensionaler multimodaler Datensätze bieten. Diese kumulative Dissertation untersucht die Herausforderungen, die mit der Handhabung solcher Daten verbunden sind, und beleuchtet Top-Down-, Bottom-Up- und integrierte Ansätze, die zur Verwaltung, Analyse und Visualisierung dieser Daten entwickelt werden. Die Arbeit ist in zwei Teile gegliedert, die jeweils die beiden Problemlösungsansätze und ihre Anwendung in veröffentlichten Studien widerspiegeln. Die vorgeschlagene Arbeit zeigt, wie wichtig es ist, beide Ansätze zu verstehen, die Schritte des Denkens über das Problem innerhalb dieser Ansätze und ihre Konkretisierung und Anwendung in verschiedenen Bereichen.

Contents

List of Figures

List of Tables

List of Abbreviations & Acronyms

AI	**A**rtificial **i**ntelligence
ANOVA	**An**alysis **o**f **va**riance
AUC	**A**rea **u**nder the **c**urve
BD	**B**lu-ray **D**isc
BIRCH	**B**alanced **I**terative **R**educing **C**lustering Algorit**h**m
BWWTP	**B**iological **w**aste**w**ater **t**reatment **p**lant
C	**C**arbon
CD	**C**ompact **D**isc
CF	**C**ompact**F**lash
CH	**C**alinski-**H**arabasz score
CMIP	**C**oupled **M**odel **I**ntercomparison **P**roject
CMU-MOSEI	**CMU M**ultimodal **O**pinion **S**entiment and **E**motion **I**ntensity
CPK	**C**orey-**P**auling-**K**oltun
CPP	**C**ell-**p**enetrating **p**eptide
CPU	**C**entral **p**rocessing **u**nit
CRMSE	**C**entered **r**oot **m**ean **s**quare **e**rror
CSV	**C**omma-**s**eparated **v**alues
Cu	**C**opper
DAT	**D**igital **A**udio **T**ape
DBI	**D**avies–**B**ouldin **i**ndex
DCC	**D**igital **C**ompact **C**assette
DL	**D**eep **L**earning
DNA	**D**eoxyribo**n**ucleic **a**cid
DV	**D**isco**v**ision
DVD	**D**igital **V**ersatile **D**isc
GFF	**G**eneral **f**eature **f**ormat
GNN	**G**raph **n**eural **n**etwork
GO	**G**ene **O**ntology
H	**H**ydrogen
HadCM3	**Had**ley Centre **C**oupled **M**odel version **3**
HDD	**H**ard **D**isk **D**rive
hPa	**H**ecto**pa**scal
HSD	**H**onestly **s**ignificant **d**ifference

HVD	**H**olographic **V**ersatile **D**isc
IDC	**I**nternational **D**ata **C**orporation
K	**K**elvin
KB	**K**ilo**b**yte
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
LD	**L**aser**D**isc
LDDP	**L**imiting **d**ensity of **d**iscrete **p**oints
LOEWE	**L**andes-**O**ffensive zur **E**ntwicklung **w**issenschaftlich-ökonomischer **E**xzellenz
M-DISC	**M**illennial **D**isc
MAE	**M**ean **a**bsolute **e**rror
MCC	**M**atthews **C**orrelation **C**oefficient
MDS	**M**ulti**d**imensional **s**caling
MI	**M**utual **I**nformation
MID	**M**utual **I**nformation **D**iagram
MILC	**M**ultidimensional **I**n-depth **L**ong-term **C**ase studies
ML	**M**achine **l**earning
MOSLA	**Mo**lecular **S**torage for **L**ong-term **A**rchiving
MOVIS	**M**ulti-**O**mics **Vis**ualization
MSE	**M**ean **s**quared **e**rror
N	**N**itrogen
NB	**N**aive **B**ayes
NMI	**N**ormalized **M**utual **I**nformation
NMID	**N**ormalized **M**utual **I**nformation **D**iagram
NVI	**N**ormalized **v**ariation of **i**nformation
O	**O**xygen
OBS	**Obs**ervation
OPTICS	**O**rdering **p**oints **t**o **i**dentify the **c**lustering **s**tructure
PCA	**P**rincipal **c**omponent **a**nalysis
PDF	**P**robability **d**ensity **f**unction
RAID	**R**edundant **A**rray of **I**ndependent **D**isks
RBF	**R**adial **B**asis **F**unction
RMS	**R**oot **m**ean **s**quare
RMSE	**R**oot **m**ean **s**quare **e**rror
RNA	**R**ibo**n**ucleic **a**cid
ROC	**R**eceiver **o**perating **c**haracteristic curve
SBK	**S**even **B**ridges of **K**önigsberg
SOTA	**S**tate **o**f **t**he **A**rt
SD	**S**ecure **D**igital card
SGD	**S**tochastic **G**radient **D**escent
SMI	**S**caled **M**utual **I**nformation
SMID	**S**caled **M**utual **I**nformation **D**iagram
SMILES	**S**implified **m**olecular-**i**nput **l**ine-**e**ntry **s**ystem
SSD	**S**olid-**S**tate-**D**rive

SV	**S**upport **v**ector
t-SNE	**T**-distributed **s**tochastic **n**eighbor **e**mbedding
TAR	**T**ape **Ar**chive
TD	**T**aylor **D**iagram
TSV	**T**ab-**s**eparated **v**alues
UHD	**U**ltra **H**igh **D**efinition
UI	**U**ser **I**nterface
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
USB	**U**niversal **S**erial **B**us
VI	**V**ariation of **i**nformation
WCRP	**W**orld **C**limate **R**esearch **P**rogramme

To my parents and my brother.

Acknowledgments

The completion of this dissertation marks a significant milestone in my academic career, and it is a moment that I am incredibly proud of. I recognize that this accomplishment would not have been possible without the support and guidance of the many individuals who have been instrumental in my success.

The unwavering support and love from my family have been the backbone of my academic journey, and I couldn't have come this far without you.

I owe a great deal of gratitude to my partner, Branka, whose love motivated me and embellished my doctoral life.

To my mentors, Prof. Dr. Dominik Heider and Dr. habil. Georges Hattab, I extend my deepest gratitude and appreciation for the way they guided me throughout this challenging journey and made it not so challenging after all. I will always be amazed by your humility, generosity, open-mindedness, and the atmosphere you created, which enabled me to grow and flourish. I hope to follow in your footsteps and inspire others as you have inspired me.

This research would not have been completed without the help of my colleagues and friends with whom I had many fruitful and stimulating conversations. I will miss our lunches, movie nights, regulars' tables, and other activities that made this journey extremely enjoyable.

I would like to thank Philipps-Universität Marburg, which has provided me with a conducive academic environment to pursue this research. In addition, this work was financially supported by the LOEWE program of the State of Hesse (Germany) in the MOSLA research cluster, to which I also extend my gratitude.

Finally, I am grateful for the support and contributions of the individuals and institutions involved in this work. Below, I extend my sincerest thanks to all those who have been a part of each incorporated study in this dissertation.

- Chapter 3: All authors are members of the MOSLA consortium, which has received funding from the Hessian Ministry for Science and the Arts (LOEWE). Specials thanks to Stefanie Dehnen and Bernhard Seeger for their support and expertise.
- Chapter 4: Specials thanks to Roman Martin for his support and expertise.
- Chapter 6: This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).
- Chapter 7: We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model data set. Support of this data set is provided by the Office of Science, U.S. Department of Energy.

Nothing has such power to broaden the mind as the ability to investigate systematically and truly all that comes under thy observation in life.

Marcus Aurelius Antoninus

1

Introduction

1.1 Motivation and Background

The past decade has seen enormous data acquisition efforts enabled by new sensory technologies, their drop in cost, and the rise of rapidly increasing computational power. Primarily these, but also other factors, allowed humankind to discover, understand, test, and validate new phenomena and existing theories on a scale never imagined before. However, with the advent of an era of *Big Data* *, traditional methods to manage, process and analyze data for testing hypotheses or developing new theories became inadequate. These traditional methods, such as manual data entry, basic statistical analysis, and low-dimensional visualizations, struggled to cope with the volume and variety of data that big data encompasses. The inadequacy of traditional methods holds true, especially when the data is additionally high-dimensional and/or multi-modal (see Section 1.1.2). As a matter of fact, with such data, the diversity and number of tasks that users need to solve has increased significantly. Besides managing, analyzing, and visualizing the data, users now encounter additional challenges such as handling metadata, embedding, clustering, and more. New approaches were and are still required to analyze, process, and visualize such data, which continuously grows in complexity and volume.

This section covers central aspects of problem solving and reasoning as tools for developing new hypotheses or testing existing theories. Two concepts covered here are then used as a foundation for understanding all approaches to managing big, high-dimensional, and multi-modal data.

1.1.1 Reasoning and Problem Solving

One of the first and probably, the most crucial part of the problem-solving process is selecting a strategy for solving the problem. The way the problem is approached likely defines a strategy used to solve it. This step overwhelmingly relies on the type and the quality of reasoning of an individual trying to solve the problem. The selection of a problem-solving strategy is thus intertwined with the

*Big Data is a term commonly used to refer to the massive amount of data with a certain degree of complexity that cannot be easily managed, analyzed, and visualized by traditional data-processing means.

selection and the quality of the reasoning about the problem. Philosophy gives us two main types of reasoning — deductive reasoning and inductive reasoning. The two can be seen as opposites. While deductive reasoning (or deduction) presents the move from general to particular, inductive reasoning (or induction [†]) is the move from particular to general (or universal). Both forms of reasoning are equally important and, as such, have been used unintermittedly throughout human history in various aspects of life. The necessity of using both forms to describe, understand, and predict natural phenomena was first noted in antiques by Aristotle and Plato. The use of reasoning to reach conclusions that humans believe are true, according to Aristotle, is more certain than sense perceptions alone[253].

As mentioned earlier, a deduction is associated with the movement from the general idea to the particularities of it. That movement is strictly ruled by applying the rules of inference — schemas describing the path from a set of premises to a conclusion based only on the logical form of the premises. This is the reason why deduction is closely associated with an experimental approach in science and academia. It is a logical and, thus, straightforward method for checking the validity of the theory and subsequently refining or discarding it. Consequently, deductive reasoning is commonly referred to as "top-down" thinking.

In contrast to the fact that a deductive conclusion is certain given the premises are correct, an inductive conclusion is probable and only based upon the given evidence (sample)[69]. Hence, the property of an induction [‡] where a conclusion can be false, even if all premises are true, is also known as *The Problem of Induction*[139,151]. Despite this property, the necessity for both deduction and induction is commonly seen in science. The development of a theory is regularly done using inductive reasoning. The theory is then adopted and tested by deducing details that must be true if the theory is valid. Similarly to deduction, inductive reasoning is generally known as "bottom-up" thinking.

As mentioned earlier, both deductive and inductive reasoning serve a central purpose in science;

[†]The name comes from Cicero's Latin translation *inductio* of Aristotle's Greek word *epagogé*, which he used in the 300s BCE[113].

[‡]The reader should note that the definition of inductive reasoning presented here differs from mathematical induction. Mathematical induction is, in fact, a form of deductive reasoning.

they allow the development of hypotheses and theories, in addition to validating and testing them. One example of deductive reasoning in science is the use of the scientific method — a step-by-step process used to systematically test hypotheses. One general variant of a scientific method diagram can be seen in Figure 1.1. In addition, mathematics and physics extensively use deduction to validate models and theories. Inductive reasoning is often used in fields where there is little to no existing literature or research on a topic; hence, there is no theory to test. This reasoning is commonly present in ecology (patterns in natural systems are observed and used to make predictions)[316], psychology (experiments are conducted to observe patterns in human behavior)[294], and machine-learning (good models are capable of generalizing the properties underlying the data but also be specific enough when needed)[78], among other fields.

Moreover, combining deductive and inductive approaches can help to identify limitations and errors in scientific theories, and refine them over time. Ultimately, the use of both types of reasoning is essential for advancing scientific knowledge and understanding.

Figure 1.1: The Scientific Method Diagram. Even though the procedure can vary within the fields of inquiry, the general process is usually the same from one domain to another.

1.1.2 Data, Analysis, and Visualization

The International Data Corporation (IDC) estimates that in 2025 we will have 180 zettabytes (ZB) of human and machine-generated data through experimentation, simulation, and observation methods[359]. Besides the obvious problem of managing the sheer amount of data, the new and highly prevalent data properties presents an even more challenging problem — high dimensionality and multi modality.

Analysis and visualization of data with some inherent two- or three-dimensional semantics have been done even before using external computing power. However, as the need to understand and unravel more complex problems increased, the power of analysis and visualization of large amounts of arbitrary multidimensional data stagnated. Researchers in both areas have been extending existing techniques to be helpful for these new large data sets, as well as developed new techniques and tested them in different application domains[85,174,182]. This process is still very much alive and ongoing[26,80,202].

High-Dimensional and Multi-Modal Data

High-dimensional data presents one of the most commonly found data nowadays. It is often defined as data where the number of observations (or samples) is significantly smaller than the number of features (or covariates)[301]. New ways of gathering and creating information enabled high-dimensional data to occupy mainstream statistical research, both in academia and in industry. For example, commonly found gene expression data sets often contain thousands of genes per each independent sample[224].

Another such example can be found in climatology and Earth sciences. Let us consider the UK Hadley Centre Coupled Model version 3 (HadCM3)[45]. This model presented a solid simulation of a climate and was a significant advancement at the time it was developed[255]. However, if we consider that this model contains 133152 spatial points and besides this we only decide to take into ac-

count 19 vertical atmospheric levels with only three pressure levels (850 hPa [§], 500 hPa and 250 hPa), we would still end up with 21024 features. Due to the greater complexity of real-world experiments, this number is usually much greater.

Moreover, in machine-learning and similar domains, an image is often considered a high-dimensional data sample. If there is a data set of K images whose dimensions are $m \times n$, m being the height and n the width in pixels, then that data set is considered high-dimensional if $K < m * n$. If we consider a color image instead of a grayscale one, the inequality becomes $K < 3 * m * n$. If one considers video data sets, which can be regarded as image data sets with a temporal dimension coefficient multiplied by the image dimension, their dimensionality quickly explodes into millions and billions of dimensions. This is why the previous inequality quickly becomes true for many real-world data sets.

From a human perspective, the information we gather about our surroundings is, in the general case, intrinsically multi-modal. Through the five basic human senses (touch, sight, hearing, smell, and taste), humans constantly collect and process data in both conscious and unconscious states. Multi modality of the data presents a challenge but also an opportunity to gather a broader range of information that may not be readily accessible using traditional research methods. Additionally, it provides a more comprehensive picture of the phenomenon being studied, which can lead to more accurate analysis and interpretation of the data. The challenging part comes from the modality itself. As the name suggests, multi-modality means the data comes in different modalities (or types). The modalities include image, audio, video, text, haptic, or other sensory data. Thus, one of the biggest challenges is the sheer volume of data that must be managed, processed, and analyzed. The problem becomes even more complex if one or more modalities contain high-dimensional data. Additionally, the integration of different modalities of data can be highly complex and require specialized expertise, software, and tools to manage.

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) data set is one of the examples of multi-modal data sets. The data set contains three modalities: video, audio,

[§] Hectopascal is a 100x multiple of the pascal, the SI unit for pressure. It is the international unit for measuring atmospheric or barometric pressure. Sea level pressure is around 1000 hPa.

and text. Each sample of the data set presents a transcribed (text modality) video (video modality) where only one person in front of a camera is discussing (audio modality) a certain topic. Since the data set is often processed using deep learning (DL) [¶] methods, it also contains annotations of the person's gender and sentiment throughout the video for the purposes of completing the classification or regression tasks. Besides being multi-modal, the data set is also high-dimensional. The processing of the whole data set can be done in a plethora of ways depending on the problem, as seen in the following papers [191,250,312].

Multi-modal data sets are also commonly found in other fields such as psychology, linguistics, education, biomedicine, and so on. One biological time-series multi-modal data set is extensively covered in Chapter 4. The data set consists of five different modalities from different biological domains: metagenomics, metaproteomics, metatranscriptomics, metabolomics, and physico-chemical data from the Biological WasteWater Treatment Plant (BWWTP) [141]. The data was collected *in situ*, at weekly intervals, and over 14 months. The data was processed, managed, analyzed, and visualized using the tool named *MOVIS* (from Multi-Omics Visualization), created for the study, and presented as the central part of Chapter 4.

Finally, analysis and visualization are only one part of the overall challenge that comes from using big, high-dimensional, multi-modal data. The complete data-management pipeline consists of data storage and transfer as well. Suppose we focus on data storage and consider current data creation and consumption trends. The International Data Corporation (IDC) made a new estimation of the amount of created data for 2025 — 180 ZB [359]. In that case, it becomes apparent that existent data-storage technologies will soon require either more density (more bytes per space or volume) or will take more physical space to store more data. However, with the advancement of novel data-storage media, such as DNA [1,61,73,120,336], the problem of storing ever-growing amounts of data can be solved. Yet, due to its novelty, more research is required in all parts of the data-storage pipeline when using DNA as a storage medium. The pipeline consists of multiple components that are, on

[¶] Deep learning is a subset of machine learning based on artificial neural networks with representation learning containing three or more layers.

their own, very complex and cover various research domains, hence requiring close collaboration and sharing of knowledge across different fields.

Analytical Approaches to High Dimensionality

Managing, processing, and analyzing high-dimensional data poses unique challenges because traditional data analysis methods fail when the number of variables is large. It is crucial to, in some way, preprocess the data before performing analysis. Preprocessing involves many steps, some of which are often excluded depending on the domain from which we get the data and in which we solve the problem. In general, preprocessing steps include data cleaning, normalization, and feature selection. Data cleaning involves identifying and correcting errors or inconsistencies in the data. Normalization is used to scale the data to a common range, while feature selection is used to select the most relevant features to the analysis.

If the preprocessing step is not required, or the data is already preprocessed, the data is analyzed. There are several methods for analyzing high-dimensional data, some of which are dimensionality reduction, clustering, and classification. Dimensionality reduction involves reducing the number of dimensions of the data while retaining the essential information. Clustering involves grouping the data into clusters based on similarities between the variables. Classification involves predicting the class of an unknown sample based on the features.

The main focus of this part is not to explain and describe each of the preprocessing steps and analysis techniques in detail but give a sense of the problem that high-dimensionality creates and the purpose of using some of the presented analysis methods. Many different preprocessing and analysis methods were covered in Chapters 3, 4, 6, and 7 as parts of published studies. For instance, Chapter 4 covers both high-dimensional multi-modal temporal data preprocessing (cleaning and feature selection) and analysis (clustering using OPTICS[10] and K-Means[17] methods, and dimensionality reduction using t-SNE[318], PCA[218,309], and MDS[32] methods)‖.

‖The reader can find the meaning of all abbreviations and acronyms on the List of Abbreviations & Acronyms page at the beginning of the dissertation.

We cluster, classify, or develop reduced dimensionality representations of dynamical systems or data sets they produce to understand them better. By understanding, we mean that we can develop simpler models of the system, allowing us to prove results or gain insight from experimentation that can then be applied to the original system. By moving from complex to manageable, we can apply traditional analytical methods, test our hypotheses, validate results, and eventually apply those new insights back into original high-dimensional models.

When clustering or classifying high-dimensional data, we look for patterns and relationships among the features, which can help us make predictions from the data. Clustering algorithms attempt to group the data into distinct clusters based on the identified patterns and relationships. Classification algorithms try to categorize the data into predefined classes. Multiple examples of cluster analysis have been used for various real-world problems, such as grouping related documents for browsing, finding genes and proteins that have similar functionality, or as a means of data compression [293].

In addition, we want to reduce the dimensionality of systems or data sets in order to aid computational analysis. One phrase, and a significant challenge, is often associated with the high-dimensional data — "the curse of dimensionality". The phrase was first used by Richard E. Bellman [25] and described as difficulty with sampling in spaces when increasing their dimensionality due to their exponential increase in volume during the process. One short example can aid in understanding the problem better. If we want to sample the unit interval $[0, 1]$ with distance no greater than 0.01, we would need only $10^2 = 100$ sample points. However, if we want to do the same thing to a 10-dimensional unit hypercube **, we would need $(10^2)^{10} = 10^{20}$ sample points. Thus, the problem exponentially explodes with the increase of dimensions. Moreover, "the curse of dimensionality" also covers the phenomena known as "the concentration of norms" which presents one of the non-intuitive features in high-dimensional geometry.

Not many studies in the current research landscape represent bottom-up analytical approaches

** Hypercube refers to a cube of length one in n-dimensional spaces, where $n \geq 3$.

(as defined in Section 1.1.1). Many conform to using already developed and established techniques to solve problems in their domains (the top-down approach). In some instances, this is not satisfactory, so there is a need for innovation and the creation of a new method, hence using the bottom-up approach for solving the domain problem.

Visualization Approaches to High Dimensionality

The high dimensionality of data creates several challenges for visualization. One major challenge is that human intuition is limited to three dimensions, making it difficult to interpret and make sense of data with many dimensions. In the pre-computer age, many visualization techniques were successfully used over the years to visualize one- or two-dimensional data, as described in the books by Edward R. Tufte [314,315]. With the introduction of computers, many visualization problems could be tackled for the first time, due to the newly available computing power. Additionally, many traditional visualization techniques were not designed to handle high-dimensional data that was becoming more available. For these reasons, and to address the need for new ways to visualize high-dimensional data, innovative techniques were developed.

The early approaches included parallel coordinates charts [29,159,160], scatterplot matrices [9,62,96], and others [8,37,57,174]. Current researchers in graphics/visualization are working to make these techniques useful for high-dimensional multi-modal data sets as well as creating new ones and using them in different application domains.

One of the newest and probably the most vital research in the visualization domain was done by Tamara Munzner [214,222] by proposing an analysis framework that acts as a structure for the vast design space a researcher is confronted with when trying to visualize any data. The book introduces a model with four cascading levels (known as "the nested model of visualization") that imposes guidelines on good design decisions and importance of choosing appropriate validation strategies at each level. The visual depiction of the model can be seen in Figure 1.2.

The most-outer level of the model, the domain level, defines a problem in some domain (*e.g.*,

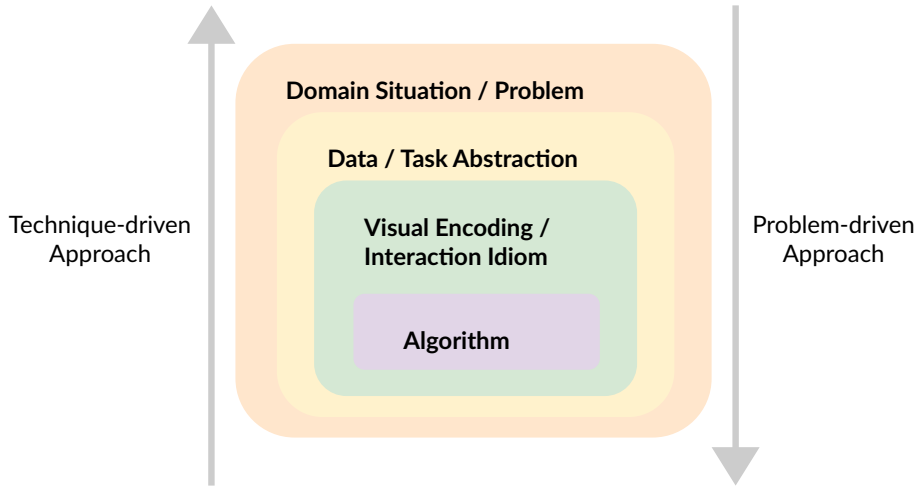


Figure 1.2: The Nested Model of Visualization. Each level builds upon the one before it; the output from one level is used as input for the next. Hence, making bad design choices at the upper level will spill them into lower levels. Two universal visualization design thinking perspectives, technique-driven and problem-driven approaches, are depicted using directional arrows.

physics, mathematics, microbiology, genetics, *etc.*) and incorporates the target users, their field of interest, their questions, and their data. The results of the design process on this level are the sets of requirements or needs of the target audience. The next level deals with the problem of abstracting the domain problem, its data, and requirements. Sometimes, the original, unmodified data is in an appropriate form and does not need any mean of abstraction to complete the design process on this level. Often, however, the original data is unsuitable, and some transformations are required in order to create a more manageable abstraction. At the next level, a proper visual encodings and interaction idioms are selected to present and manipulate the abstractions developed at the previous level. At this level, mixing multiple visual encodings and interaction idioms is often required in order to achieve the best results. Sometimes, there is not one best result but a multitude of them. It is up to the person in charge of the design process to select the one that best suits the requirements by validating his choices on this and upper levels. The deepest level is about creating or selecting an algorithm that implements the choices of all upper levels so that it provides an adequate user experience and is as optimised as possible[222].

The design choice made on each level can, but is not necessarily, the best one that can be made. To be certain of the quality of the resulting visualization, one must validate choices on each level. Because the model in question is nested, the validation of choices made in upper levels has to come after the validation of choices made in inner levels. This co-level dependency presents the biggest obstacle in validating all choices, which is why a single project often addresses only a subset of the levels.

Moreover, one of the other important aspects of the aforementioned research is the use of general reasoning approaches (borrowed from philosophy) as defined in Section 1.1.1 — top-down and bottom-up. In the context of visualization, and by following the nested model of visualization, these two directions of reasoning are often used interchangeably with the following terms: problem-driven (top-down) and technique-driven (bottom-up) approach. When faced with a visualization problem, a person chooses one of these strategies for devising a proper visualization.

If one chooses a problem-driven approach, the starting point of the design study is at the top level of the nested model — a domain situation or a problem. Properly characterizing a problem is paramount and can significantly affect later design choices and the resulting visualization. The formulation of the problem is often devised with the help of domain experts who understand the situation in depth and can give precise requirements for the design. It is up to a visualization expert to refine and translate those requirements into a visualization language. Once the requirements are set, the visualization expert is enabled to traverse into inner levels, hence starting the process of finding the best set of solutions for the given domain problem. Usually, the problem can be solved using one of the existing and well-known abstractions, encodings, idioms, and algorithms. Yet, sometimes the problem requires the design, implementation, and validation of new ones that better solve the problem at hand. This is often the case with data and task abstractions, where a visualization expert has to create a visualization solution for a problem that was not covered extensively before.

However, if a technique-driven approach is more appropriate for a situation, the work starts at one of the inner levels of the nested model and progresses outwards. This approach consists of two

possible paths depending on which level the design process begins. In the case of choosing a visual encoding or an interactive idiom as starting points, the goal of the process is to design and implement new visual encodings or interactive idioms that improve well-known abstractions of the upper level. As a result, these discoveries can help many domains whose problems rely on those abstractions. If the starting point is at the algorithm level, the objective is to create better algorithms that support existing idioms [222].

As mentioned in the previous sections, reasoning about the problem systematically often reveals the best approach to solving that problem. In one case, the domain problem can be solved using existing visualization encodings, idioms, and algorithms. The main challenges with this approach are the proper requirement specification and sometimes the innovation of new abstractions. In another case, the problem rests in the inner levels of the nested model and requires the creation of new algorithms, encodings, or idioms. Often these apparently specific solutions solve more problems than they were designed to solve and apply to multiple fields. Chapters 2 and 5 cover both approaches in more detail and present a few examples relevant to this work. Chapters 3 and 4 are research studies that employ the problem-driven approach, while Chapters 6 and 7 cover the studies that employ the technique-driven approach.

1.1.3 Evaluation

The last part of the chain of challenges that come with the problem-solving process is evaluating the solution. Without evaluation, it is hard to estimate the true quality, power, and extent of the solution. Unfortunately, even this part is not straightforward. Depending on the nature of the solution, some parts of it can be easily evaluated, others with some obstacles, while some cannot be evaluated at all.

On the one hand, evaluating newly developed dimensionality-reduction algorithms is relatively straightforward. One only needs to use the same benchmark data sets other algorithms used and evaluate the newly developed one against them while also benchmarking complexity, performance,

scalability, and other relevant properties. A potential challenge lies in showcasing the enhanced effectiveness of this novel algorithm in comparison to existing ones. However, this step is usually solved since the development of the algorithm is often motivated by solving some specific problem.

All of the before mentioned steps can be clearly recognized in the case of the Uniform Manifold Approximation and Projection algorithm, or UMAP[208]. The authors of this study evaluated the algorithm against t-SNE[318], LargeVis[305], Laplacian Eigenmaps[24], and Principal Component Analysis (PCA)[342] on the COIL20[225], MNIST[185], Fashion-MNIST[343], and GoogleNews[217] data sets. While evaluating different algorithmic properties, they also demonstrated the additional power of the algorithm to create embeddings with a more apparent global and topological structure among the various clusters, than its main competitor t-SNE.

On the other hand, evaluating [††] visualizations is a bit more demanding, as discussed in works by Morse *et al.*[220] and Plaisant *et al.*[244]. Since visualizations are almost always [‡‡] developed to be used by humans, that means humans should also evaluate them. This is the reason why well-designed user studies and surveys are one of the most powerful evaluation tools for visualization. The challenges here are pretty evident — it is hard or even costly to find many participants, it is difficult to design an adequate study, and it is troublesome to control nonessential variables and thus have precise measurements[43]. Moreover, the resulting visualization has to be evaluated on all levels of the nested model to ensure its validity[222]. To overcome this, numerous automatic approaches were presented[20,53,346] in order to relax this constraint on certain levels of the nested model, and streamline evaluating visualizations.

An example of a visualization study where not all levels of the nested model were evaluated is a work by Andreas Noack[228]. The paper does not address the algorithm level at all, hence the lack of an evaluation of the algorithm used to create the proposed visualization. In addition, the selec-

[††] Munzner uses the term *validation* instead of *evaluation* to define the validation of the design choices on each level of the nested model[222]. In the context of this dissertation, using the term *evaluation* feels more natural.

[‡‡] Chapter 6 is one of the exceptions to this rule since the resulting visualizations (*i.e.*, molecule encodings) are meant to be used by machines, for example, in ML experiments.

tion of task and data abstractions is crudely described and, as such, cannot be considered a suitable evaluation of abstraction choices on this level. On the contrary, the choices on the domain level are well discussed by referencing previous work and motivating the need for a new approach. Finally, the evaluation of the selected idiom is done in an unorthodox fashion in the visualization research — using mathematical proof. This provides more than a suitable way of determining the validity of the selection; hence the evaluation of choice on this level is adequately handled.

In the realm of data analysis and visualization, evaluation plays an essential role in determining the quality and credibility of our findings and the results of a problem-solving process. Apart from that, the evaluation process also aids in refining and perfecting our methods, allowing us to obtain better insights and interpretations of the data in the future. However, as shown earlier, the process of evaluation also comes with its own set of challenges. Depending on the nature of the solution, some parts could be easy to evaluate, while others may pose obstacles. Moreover, there may be certain parts of the solution that cannot be evaluated at all. Therefore, it becomes essential to devise a comprehensive evaluation plan, taking into account all of the possible uncertainties and limitations.

1.2 Problem Statement

otivated by a vast amount of valuable information that can be uncovered from high-dimensional and multi-modal data, researchers often skip many vital steps of analytical reasoning and deep-dive into implementing a possible solution. In doing so, many crucial design decisions in analytics and visualization could be omitted for the sake of expediteness. The results of such design process, or the lack of it, to be precise, are almost always not the best and sometimes even wrong for the problem in question. A clear and systematic approach to reasoning about the problem is required to maximize the expressiveness of the data and the transfer of information from data to humans.

In addition, existing methods for analysis and visualization are often not suitable for certain domain-specific problems. More data- and task-specific analysis and visualization methods are

needed for certain research fields. This trend of innovation is an ongoing process. New analysis and visualization techniques are continuously being developed to meet the changing demands and technological advancements across multiple disciplines.

These challenges could be solved in two ways. By employing the nested model [222] for both analysis and visualization, we can systematically approach problems in two directions and avoid skipping important steps in the problem-solving process. On the one hand, existing solutions that span methods, techniques, and tools could be repurposed for a specific problem and setting. The difficulty of this approach lies in the problem and user specification and their abstractions and translations to a vocabulary of analysis and visualization. By adequately formulating what data we analyze and why we do it, we can make suitable abstractions of the tasks and the data. Next, current state-of-the-art (SOTA) algorithms could support those choices, thus ensuring quality in the resulting analyses and visualizations.

On the other hand, the need to further optimize and advance existing SOTA algorithms and methods could prove beneficial for solving problems across various research fields. The main challenge of this approach lies in the necessary innovation and creativity to push the existing methods or algorithms further.

Unfortunately, there is no correct answer regarding choosing the best reasoning approach. Different domains have different problems, different requirements, and different expectations. Sometimes, the conceptualization of the problem is complex on its own and requires significantly more time than the rest of the problem-solving process. Moreover, evaluating choices adequately presents a challenge on its own and requires additional effort. It is up to the researcher, who is in charge of translating and solving the domain problem, to answer the following questions: **what is the best approach**, **how best to tailor methods to the user requirements and needs**, **how to provide the best possible results**, and **how to evaluate those results adequately**?

1.3 Overview

The main scope of this dissertation is to give a theoretical background of problem solving and reasoning and their applications in scientific research, specifically for solving complex analytical and visualization tasks using high-dimensional data in biological, machine-learning, meteorological, medical, and other similar domains. The work is structured and conceptualized according to the approaches described in Section 1.1.2. Therefore, approaches from the visualization domain for solving problems with high-dimensional multi-modal data are highlighted and used as guidance for solving analytical problems as well.

An illustrated overview of the chapters of this dissertation can be seen in Figure 1.3. Chapters 3, 4, 6, and 7 are adaptations of already published studies. The list of studies and overall contributions of each author is covered in Section 1.4. Individual contribution to each study is highlighted at the beginning of the before-mentioned chapters. Comprehensive summaries of each chapter are given in the remainder of this chapter.

Chapter 1 motivates the work and gives a theoretical basis. In addition, it brings an overview of the dissertation both in general and in a specific manner.

Chapter 2 covers in detail one of the approaches mentioned in Chapter 1 — a problem-driven approach. A general overview and description of the approach is given. Moreover, multiple examples are presented and evaluated in the context of this dissertation. Two central examples are given in Chapters 3 and 4, thus corroborating the strengths of this approach for certain tasks.

Chapter 3 presents the first central example of a tool and a review created using problem-driven approach, as introduced in Chapter 2. A self-contained motivation and background are provided on the history and state-of-the-art of data storage media and user interfaces (UIs) used to present their properties. This enables readers to better understand the methods and findings of the underlying

study.

Chapter 4 introduces and covers in detail the study of the second central example mentioned in Chapter 2. As with the previous example, a self-contained motivation and background are provided on the existing tools and methods for managing, processing, embedding, clustering, and visualizing high-dimensional time-series multi-omics data sets. An overview of the resulting tool is presented in detail, along with the case study which was conducted for the purposes of this research.

Chapter 5 covers the second approach motivated and introduced in Chapter 1 — a technique-driven approach. A general overview and description of the approach is given. Moreover, multiple examples are presented and evaluated in the context of this dissertation. Two central examples are given in Chapters 6 and 7, thus corroborating the strengths of this approach for encountered tasks.

Chapter 6 describes in detail the first central example of a technique-driven approach (as mentioned in Chapter 5). A short motivation and an overview are given before an introduction to the research problem covered in this chapter. A multitude of molecular encoding methods and tools are covered and evaluated against the method proposed in this study. An open-source tool implementing the proposed method is presented, along with the evaluation results.

Chapter 7 examines the second central example of a technique-driven approach. As in other examples, a self-contained overview motivates the reader and introduces it to the general problems, ideas, and results of the study. Some essential aspects of the information theory are presented and explained, hence providing a necessary background for the research conducted in this chapter. The resulting tool is described and evaluated using high-dimensional data sets from various domains.

Chapter 8 discusses and establishes the work of this dissertation in the broader context of analytical and visualization tools for high-dimensional data. The use of both problem-driven and

technique-driven approaches is evaluated and discussed. Further research directions are explored and discussed in the context of this dissertation.

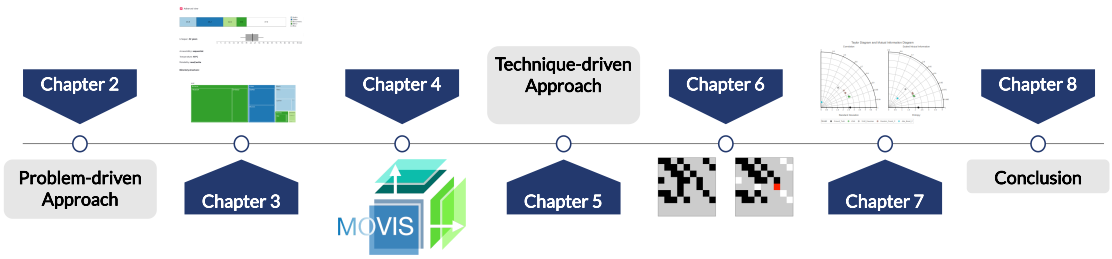


Figure 1.3: Dissertation Overview. The motivation and the background for problem-driven and technique-driven approaches are given in separate chapters. For both problem-driven and technique-driven approaches, two central examples are presented in detail in Chapters 3, 4, 6, and 7, respectively.

1.4 Publications and Contributions

The following section is a comprehensive list of all publications and contributions related to the research presented in this dissertation. The list is organized chronologically and includes the title of the publication, the names of the authors, and the overall contributions of each author. Additionally, any relevant citations or other relevant information is also included. This list is intended to provide readers with a thorough understanding of the research and body of work that has led to the findings presented in this document.

Publication 1 [12] Aleksandar Anžel, Dominik Heider, and Georges Hattab. The visual story of data storage: From storage properties to user interfaces. *Computational and Structural Biotechnology Journal*, 19:4904–4918, 2021. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2021.08.031. URL https://www.sciencedirect.com/science/article/pii/S2001037021003627

Contributions 1 [12] A.A. curated the data, implemented the survey, validated the results, and created the visualizations. A.A., D.H., and G.H. wrote the original draft. G.H. supervised the work. All authors reviewed the manuscript.

Publication 2 [13] Aleksandar Anžel, Dominik Heider, and Georges Hattab. Movis: A multi-omics software solution for multi-modal time-series clustering, embedding, and visualizing tasks. *Computational and Structural Biotechnology Journal*, 20:1044–1055, 2022. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2022.02.012. URL https://www.sciencedirect.com/science/article/pii/S2001037022000526

Contributions 2 [13] A.A. wrote the manuscript, designed and developed the tool. D.H. discussed the results and revised the manuscript. G.H. supervised the project, guided the tool development, proofread, and revised the manuscript. All authors read and approved the final manuscript.

Publication 3 [136] Georges Hattab, Aleksandar Anžel, Sebastian Spänig, Nils Neumann, and Dominik Heider. A parametric approach for molecular encodings using multilevel atomic neighborhoods applied to peptide classification. *NAR Genomics and Bioinformatics*, 5(1), 01 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqac103. URL https://doi.org/10.1093/nargab/lqac103. lqac103

Contributions 3 [136] Conceptualization, G.H. and N.N.; methodology, G.H., N.N., and A.A.; software, N.N., A.A., and S.S.; validation, G.H. and A.A.; investigation, G.H. and A.A.; resources, A.A. and S.S.; data curation, G.H., A.A and S.S.; writing—original draft preparation, G.H.; writing—review and editing, G.H., A.A., and D.H.; visualization, G.H., N.N. and A.A.; supervision, D.H.; project administration, G.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Publication 4 [14] **(Under review)** Aleksandar Anžel, Dominik Heider, and Georges Hattab. Interactive polar diagrams for model comparison. *Computer Methods and Programs in Biomedicine*, 2023. ISSN 1872-7565

Contributions 4 [14] A.A. wrote the manuscript, designed and developed the library, conducted experiments, and evaluated results. D.H. discussed the results and revised the manuscript. G.H. supervised the project, guided the development, proofread, and revised the manuscript. All authors read and approved the final manuscript.

If I had an hour to solve a problem I'd spend fifty-five minutes thinking about the problem and five minutes thinking about solutions.

Albert Einstein

2

Problem-Driven Approach

HE problem-driven approach involves identifying a specific problem or question that needs to be answered using data analysis and visualization. This approach differs from a technique-driven approach, which focuses solely on the data or technology without considering the problem or context. By focusing on a specific problem, the problem-driven approach can help to reduce information overload and improve the relevance and accuracy of the analysis and visualization. It also ensures that the results are actionable and can be used to make informed domain decisions.

Numerous analytical and visualization problems can be addressed using problem-driven strategy. One such example where problem-driven approach marked a historic milestone in mathematics, as it led to the development of graph theory and the emergence of topology, is known as *the Seven Bridges of Königsberg* (SBK). The formulation of the problem and the use of a problem-driven approach to solving it are presented below.

The city of Königsberg had seven bridges connecting the two banks of the river and the island. The story says that its residents were fond of taking walks, and one of the favorite walks was to start from any point in the city, cross each of the seven bridges exactly once, and return to the starting point. However, they realized that no one had been able to complete such a walk, and so the problem was born. Leonhard Euler * was the first person to solve the SBK problem. He realized that the problem could be translated into a mathematical problem that involved graph theory. He transformed the city of Königsberg and its bridges into a graph where the land masses were represented by nodes, and the bridges were represented by edges. By doing this translation, Euler abstracted the real-world problem into a mathematical one, thus completing task and data abstractions. The data abstraction was realized by considering land masses as nodes, and bridges as graphs. In addition, the task of traversing one bridge only once was abstracted into a mathematical problem of traversing each edge only once. Euler then showed that the problem was impossible to solve because it was

*Leonhard Euler was an 18^{th}-century Swiss mathematician, physicist, astronomer, geographer, logician, and engineer. He is credited with establishing the field of graph theory and topology and was influential in many other branches of mathematics, including analytic number theory, complex analysis, and calculus.

impossible to find a path that crossed each edge exactly once if more than two vertices have an odd number of edges connected to them [99]. Such a traversal is now called an *Eulerian path* or *Euler walk* in his honor. Euler's solution formed the basis of graph theory, which has since found numerous applications in computer science, engineering, and social sciences.

In the domain of visualization, problem-driven approach is also extensively utilised for solving real-world problems. Multiple studies have systematically explored problem-driven visualization work, derived methodologies for using it in transdisciplinary teams, and presented guidelines for the domain-problem characterization [128,201,269,273,280]. Two examples of a problem-driven approach in visualization, and published as studies, are extensively covered in Chapters 3 and 4. The remainder of this chapter will introduce them in broader aspects and describe the problem-driven mechanisms used to develop and evaluate the respective studies.

The first example presented in Chapter 3 examines traditional and novel data storage media and their transdisciplinary challenges. Driven by the ever-growing necessity to store data, the domain problem was to research, select, and propose a UI for a novel data storage technology (*i.e.*, DNA) to be implemented at an operating system (OS) level. This problem was then abstracted into multiple tasks, each corresponding to the appropriate domain subtasks mentioned in the previous sentence. With the conducted research on specific properties of traditional and novel data storage media and by surveying domain specialists and non-specialists, data was collected, processed, and analyzed to propose the UI adapted to DNA as a storage medium. The problem-driven approach was used to solve the whole problem in general but also to solve the subtasks of which it consisted (*e.g.*, creating the new UI, conducting the survey, *etc.*).

In the specific case of creating the new UI, a problem-driven approach can be easily demonstrated using the nested model. The domain problem in the case of this task was twofold: the creation of an appropriate visualization of the properties of DNA as a storage medium and its use by experts (in DNA data storage) and non-experts. For each data storage property (*e.g.*, capacity, lifespan, accessibility), the same approach was employed. In the case of capacity and usage, traditional approaches

to its visualization were examined, and new methods were evaluated. The capacity and the available space of the storage medium were encoded using a stacked bar chart while using different colors for each property. The interaction with the visualization was realized using a tooltip that conveys the exact values of each property. The algorithm employed to visualize and interact with the data utilized two *Python* [323] libraries named *Plotly* [275] and *Altair* [324].

Chapter 4 presents an in-depth exploration of a tool specifically designed to manage, process, analyze, and visualize time-series high-dimensional multi-modal data from biomedical domains. The tool, entitled *MOVIS*, presents a novel approach which enables users to analyze and visualize multiple modalities simultaneously. The development of this idea was primarily motivated by a widely cited Shneiderman's mantra: "Overview-first, zoom and filter, then details on demand" [279]. This mantra outlines the importance of providing a broad understanding of the data set (overview), displaying details about each data point, and knowing when to do these activities during an analysis [222].

The domain problem in this study can be defined as follows: what patterns or anomalies can we discover by analyzing and visualizing temporal multi-omics data sets, and do we gain any benefits from considering multiple modalities simultaneously rather than independently? By following the nested model, the task abstraction, in this case, is thus done by translating the problem to finding anomalies and patterns using well-established computer science methods. Furthermore, data abstraction is offered by the tool's various embedding, clustering, and dimensionality-reduction techniques. The tool offers nine visual representations for multi-omics time-series data sets. The type of visual encoding and level of detail in the visualization depend on the data modality and data abstraction procedures performed earlier. All visualizations contain a tooltip as an interactive element, and some also include spanning, zooming, and filtering. The algorithms used to analyze and visualize the data are parts of well-known *Python* data science libraries such as *Sci-Kit Learn* [239], *Pandas* [235], *Altair* [324], *etc.*

The benefits of a problem-driven approach in analysis and visualization are numerous. One of the

main advantages is that it ensures that the analysis and visualization efforts are aligned with the domain problem. This means that the insights generated are more likely to be relevant and actionable. Additionally, this approach can help identify issues that may have been overlooked if the analysis was not problem-driven. Finally, it allows for more efficient use of resources as the focus is on answering specific problems rather than generating general insights. However, the design process for problem-driven work is cyclical, with constant refinement throughout [222]. This property of the design process also presents the biggest challenge that comes from using it. The researcher can only be sure that the right choices were made on the upper levels only when the choices are made on all levels. As a result, this approach can in some cases be quite slow and tedious, due to the continual refinement on all levels according to the domain problem and user specifications.

3

Navigating the Complexity of Data Storage Media

Status

Published as: Aleksandar Anžel, Dominik Heider, and Georges Hattab. The visual story of data storage: From storage properties to user interfaces. *Computational and Structural Biotechnology Journal*, 19:4904–4918, 2021. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2021.08.031. URL https://www.sciencedirect.com/science/article/pii/S2001037021003627

Copyright Notice

Contribution

I collected, processed, curated, and visualized the data. I implemented the survey and validated it. I proposed and implemented the user interface. I co-wrote and revised the original draft.

3.1 Preface

The rise in data consumption and production is creating a variety of challenges for data access and storage. The current data generation and storage trends are leading to increased physical space requirements for storage media. Yet, recent technological advancements uncovered new possible solutions to data storage problems. With the emergence of new storage media came new challenges with representing important data storage properties of those media. It is important to recognize that a technical solution may not always be the answer to these new challenges. Instead, a problem-driven approach is essential for developing effective solutions. For

this reason, we conducted a user survey that assessed different storage properties vital for data management, including accessibility, capacity, usage, mutability, lifespan, addressability, and typology. Our findings allowed us to prioritize user requirements and design a customized user interface that visualizes important data-storage properties and can be tailored to meet the unique needs of various user groups, including experts and the public. Furthermore, our analysis of storage devices over time allowed us to identify three distinct periods of media development: magnetic, optical and electronic, and alternative. By taking a problem-driven approach to solving the real-world challenges of data storage, we can design effective solutions that prioritize user needs and create sustainable storage systems that can adapt to the rapidly evolving tech landscape.

3.2 Introduction

At our current rate, 2.5 quintillion bytes of human and machine-generated data are created every day, and the pace is only accelerating[162]. The amount of created data in 2020 was predicted to reach 35 zettabytes or ZB (*i.e.*, 35 trillion gigabytes or GB). 33 ZB were already reached back in 2018. This lead the International Data Corporation (IDC) to make a new estimation for 2025: 180 ZB of new data will be created worldwide[359]. Although such numbers are not astronomically large, they remain too large to fathom. For comparison, the radius of the observable universe is roughly 43×10^{21} kilometers; which is two orders of magnitude smaller than the aforementioned estimate of 180 ZB or 1.8×10^{23} bytes[117]. With the advent of such a historical era, better storage devices and long-term storage solutions will be required. This especially holds true with a projection putting the world population around 10.88 billion in 2100[264].

A storage device can contain information, process such information, or do both[28]. When the device contains only information, it is called a recording medium. Recording can be done using almost any form of energy, acoustic vibrations (phonographic recording) to electromagnetic energy (magnetic tape and optical discs)[86,173,259]. Today's storage devices contain different types of magnetized media that are usually ferromagnetic materials (*e.g.*, ion or chromium oxides)[66,256]. Ferromagnetic

materials have structurally unpaired spins that are organized into magnetic domains. Many bytes, as the units of information, can be recorded on magnetized media as ”ascending” or ”descending” spin domains. These correspond to ones and zeros in the binary system. Thanks to the unique property of ferromagnetic materials, these magnetized media retain the data, and they can then be read in the same way. To provide a descriptive overview of storage devices and media, we follow their basic properties: (1) Accessibility, defines how data is organized on a device and how it can be accessed: serial or random [44]; (2) Capacity, defines how much capacity a device has (in bytes) [27], (3) Lifespan, defines how long data can be stored in certain conditions (in years) [184,286], (4) Mutability, defines the functions of a device: write, read, or both; and (5) Typology, defines the categories of storage devices: optical, magnetic, semiconductor or electronic, molecular, *etc.* Combined with other properties, such as energy use or data density, they affect the adoption and usage costs of a specific device or medium. For example, data density is a measure of the quantity of information bits that can be stored on a given length of track, area of surface, or in a given volume of a storage medium. A higher density is preferred to optimize the given length, surface, or volume of said medium. Altogether, such properties steer the adoption of storage devices and media not only by commercial companies but also by the public. In the context of long-term data storage, devices with volatile memory are faster than non-volatile memory. However, in the event of a power outage, volatile memory is not retained rendering it unsuitable. In the looming possibility of a digital dark age, this rationale extends to digital libraries and renders them obsolete, *e.g.*, tapes or network storage systems (clouds). This, in turn, led to the development of alternative and novel media.

Striking examples of novel storage media are Ribonucleic acid (RNA) and Deoxyribonucleic acid (DNA) molecules. Because of its greater stability, the DNA molecule presents a better storage medium than its counterpart RNA. It carries the biological information necessary for the proper functioning of cells [150,288,351]. It is not by any means new, as it has been used for billions of years as a carrier for genetic information by living organisms [1]. Compared to other – organic and inorganic – molecules, its capacity puts it in a league of its own [103,149]. DNA has a great potential for infor-

mation storage with a capacity outperforming existing technologies. For instance, traditional storage devices such as magnetic hard drives and flash drives have a data density of 10^{13} and 10^{16} bits per cm^3, respectively[215]. In comparison, DNA reaches a data density up to 10^{19} bits per cm^3. In other words, three orders of magnitude higher and specifically of 1 billion terabyte per gram. Moreover, DNA has a very good molecular stability, which has been shown in sequencing studies of extinct species, referred to as ancient DNA[232,248].

Aside from their novelty, a yet-to-be-considered aspect concerns the User Interface (UI), which of the aforementioned properties are relevant, and how they are visualized. For instance, knowing if a Hard Disk Drive or HDD device is at 20% or 50% of its usage is often visualized using a horizontal stacked bar chart. However, with the advent of new storage media, such as molecular media, there exist no known standards. That is to say, there is no agreement on which properties are more important or relevant for a potential user. Indeed, user preferences impact which information is judged as relevant, then subsequently visualized. In this work, we argue that certain standards should be considered to prepare for the foreseeable future where such media may be available for the task of long-term archiving or even for day-to-day use. For this purpose, we investigated the industry-wide approaches that implement UIs and visualizations to display some of these storage properties. We reported a historical account of their evolution across operating systems and platforms. We also divided the timeline into three distinct periods (magnetic, optical and electronic, alternative) and reported the TOP 3 media and devices for each time period. To accommodate for the shift of storage types into molecular media, we created a survey to rank the storage properties by importance depending on whether the user is a member of the general public or the research community (*i.e.*, domain expert). The expert pool consisted of members of the largest known research consortium on MOlecular Storage for Long-term Archiving (MOSLA). We summarize our findings below:

(a) We conducted a successful literature search of historical, currently in use, novel and experimental data storage media and devices,

(b) we created a survey to determine the most important storage media properties depending on

a target group, and

(c) by relying on our findings, we proposed a user-settable UI to select and display the relevant properties for the right audience.

Year	Data Storage Name	Type	Accessible	Mutable
1932	Drum memory [143,180,200]	●		x
1946	Williams-Kilburn Tube [68,153]	●		x
1949	Magnetic-core memory [108,236]	●	x	x
1952	Magnetic Band (Tape) [18,82,180]	●		x
1956	Hard Disk Drive (HDD) [3,90]	●	x	x
1960	Magnetic stripe card [251,299]	●		x
1969	Floppy Disk [6,311]	●	x	x
1970	Bubble Memory [31,50,56]	●		x
1978	LaserDisc (LD)/Discovision (DV) [123]	●	x	x
1978	Solid-State-Drive (SSD) [70,170]	●	x	x
1981	Compact Disc (CD) [55]	●	x	x
1987	Digital Audio Tape (DAT) [199,304]	●		x
1991	Mini-Disc [199,350]	●	x	x
1992	Digital Compact Cassette (DCC) [145,195]	●		x
1994	Zip Drive [98]	●	x	x
1994	CompactFlash (CF) [187]	●	x	x
1995	Digital Versatile Disc (DVD) [271,345]	●	x	x
2000	USB Flash Disk [2,219]	●	x	x
2001	Secure Digital (SD) card [114]	●	x	x
2002	Blu-ray Disc (BD) [107,278]	●	x	x
2004	Holographic Versatile Disc (HVD) [83,146,147]	●	x	x
2009	Millennial Disc (M-DISC) [197]	●	x	
2010	Synthetic Deoxyribonucleic acid (DNA) [1,33,61,73,97,120,124,231]	○		
2014	Cell cultures [129,298,349]	○		
2014	Colloidal particle clusters [242]	○		x
2016	Chlorine Atomic memory [169]	○	x	x
2019	Synthetic Metabolomes [177]	○		

● Magnetic ● Electronic ● Optical ● Electro-mechanical
● Magneto-Optical ● Cathode-ray tube ○ Atomic ○ Molecular

Table 3.1: Timeline of data storage media. In the beginning, data storage heavily relied on magnetic media. This slowly shifted to optical and electronic media. In recent years, we have seen the rise of novel media that rely on atomic properties, biological molecules, and organisms. The magnetic Band (Tape) storage medium is reported in 1952, although BASF introduced the first magnetic tape for audio in 1934. Although BASF supplied the first 50,000 meters of magnetic audio-tape in 1932, the magnetic tape is reported in 1952 as it corresponds to the first usage of a tape as a data carrier using the IBM 7 track. Mutability refers to read, write, or both. Note that there exists no device with a write-only mutability. We refer to the binary state of read and write (using the letter x) versus the read-only state.

3.3 Results

We report the results in five subsections. First, we focus on presenting specific storage media and devices that relate to the typology, accessibility, and mutability properties. We rely on the typology property to delineate periods of time for making the timeline more tractable. Second, we present storage media and devices by capacity and lifespan. In each subsection, only the TOP 3 is reported for the sake of brevity. Third, we briefly describe novel and alternative media as they bring in their own right a new set of challenges. Fourth, we report the results of our survey on user preferences to visualize storage properties. Fifth, we detail our proposed UI that can be toggled depending on the target audience with historically adapted visualizations and user-settable parameters.

3.3.1 Timeline by typology, accessibility, and mutability

First, we report the timeline of different storage media by their type. Second, and for brevity, we consider the logical condition of reporting the TOP 3 storage devices and media that are only and only if both randomly accessible and mutable (read and write).

In the beginning, data storage heavily relied on magnetic media. As seen in Table 3.1 and Figures 3.1 and 3.2, the magnetic time period is observed in blue. This magnetic era slowly shifted to optical and electronic media (in orange and gray). In addition, experimental storage technologies with hybrid typologies saw the light. For example, electro-mechanical and magneto-optical typologies, DAT (in red) and Mini-Disc (in gray and blue), respectively. This optical and electronic era continues to supply remarkable data storage and media. However, in most recent years, we have seen the rise of the novel and alternative media that effectively rely on atoms and molecules, while continuing to make extensive use of previous media types. Hence, the timeline of data storage devices and media can be divided into three time periods or eras.

The first is the magnetic era, where the arrival of magnetic core memory and Hard Disk Drives

revolutionized data storage as we know it. This era spanned ~ 40 years of almost exclusively magnetic devices. From the introduction of drum memory in 1932 to the bubble memory in 1970, magnetic devices starring this era were the Magnetic-core memory, the Hard Disk Drive (HDD), and the Floppy Disk. **Magnetic-core memory**, or ferrite-core memory, is an early form of computer memory. It uses small magnetic ceramic rings, namely the cores, to store information via the polarity of the magnetic field that they contain. In 1949, the earliest work on core memory was done by Shanghai-born American physicist An Wang, who created the Pulse Transfer Controlling Device. The name referred to the way the magnetic field of the cores could be used to control the switching of the current [7].

A **Hard Disk Drive** (HDD) stores and retrieves digital data from a planar magnetic surface and relies on rigid rotating platters. The information is written to the disk by transmitting an electromagnetic flux through an antenna or write head that is very close to a magnetic material, which in turn changes its polarization due to the flux. The first computer with an HDD as standard was the IBM 350 Disk File, introduced in 1956 with the IBM 305 computer. The **Floppy Disk** is a data storage device that is composed of a circular piece of thin flexible magnetic medium encased in a square or rectangular plastic casing. A Floppy Disk is read and written using a floppy disk drive. In 1967, IBM started developing a practical and inexpensive device for easy loading of microcode into their 370 mainframes and for sending out customer updates. The result of this work was a read-only, 8-inch (20 cm) floppy. Initial floppy disks were designed to hold 80 KB and load microcodes into IBM 3330, making them an intermediate device to fill another storage device, *i.e.*, a disk pack file with a 100 MB capacity. Next disks were 5.25 inches, then dimensions changed to 3.5 inches and with added protection thanks to a sliding metal cover to protect the disk medium from direct physical contact. With new floppy sizes and competitive prices, newer and smaller floppies replaced their predecessors very quickly.

The second is the optical and electronic era. Although other hybrid device and media types were introduced in this era, the optical and electronic types were the most prominent. It spanned 24

years, from 1978 to 2002, with many media and devices that were both accessible and mutable. It all started in 1978 with the LaserDisc. The **LaserDisc** (LD) is a home video format and the first commercial optical disc storage medium. In the same year, in 1978, StorageTek® launched the STC 4305, which was the first semiconductor storage device compatible with a hard drive interface, or the **Solid-State-Drive** (SSD), aimed at the IBM mainframe plug-compatible market. It entered the market as a serious competitor to the IBM 2305 HDD system with seven times the speed and half its price. The LD was the predecessor of the Compact Disc (CD), the Digital Versatile Disc (DVD), and the Blu-ray Disc (BD) which were launched in 1981, 1995, and 2002, respectively. While the SSD was the predecessor of flash memory-based media, namely the Compact Flash (CF), the USB Flash Disk, and the Secure Digital (SD) card, which were launched in 1994, 2000, and 2001, respectively. Introduced in 1994, the **Zip Drive** failed to replace the widely adopted 3.5 inches floppy disks although it was the super-floppy of the era with a much greater capacity and performance.

The third is the molecular and atomic storage era. It spans from 2004 to today. Although this era features novel storage media, only Chlorine Atomic memory and Holographic Versatile Disk (HVD) are considered both accessible and mutable. From 2004 to 2008, the **Holographic Versatile Disc** (HVD) was researched and its development was halted. Compared to DVD technology, holographic memory could have stored information at higher density inside crystals or photopolymers[146]. In DVDs, the upper limit of the data density was reached due to the diffraction limit on the writing beams. Although promising, the HVD was never released. In 2006, **Chlorine Atomic memory** consists of arranging functional atoms into extended and scalable atomic circuits. The idea is to create an atomic-scale memory that can be read and rewritten automatically by means of atomic-scale markers using chlorine vacancies on a copper sheet (Cu)[169]. Such vacancies are found to be stable at temperatures up to 77 K (-196,15°C) and would outperform state-of-the-art HDDs by three orders of magnitude.

Over time, storage devices improved in terms of accessibility. However, preferences for novel upcoming devices and media shifted their adoption and usage. As seen in Figure 3.1, if the timeline is

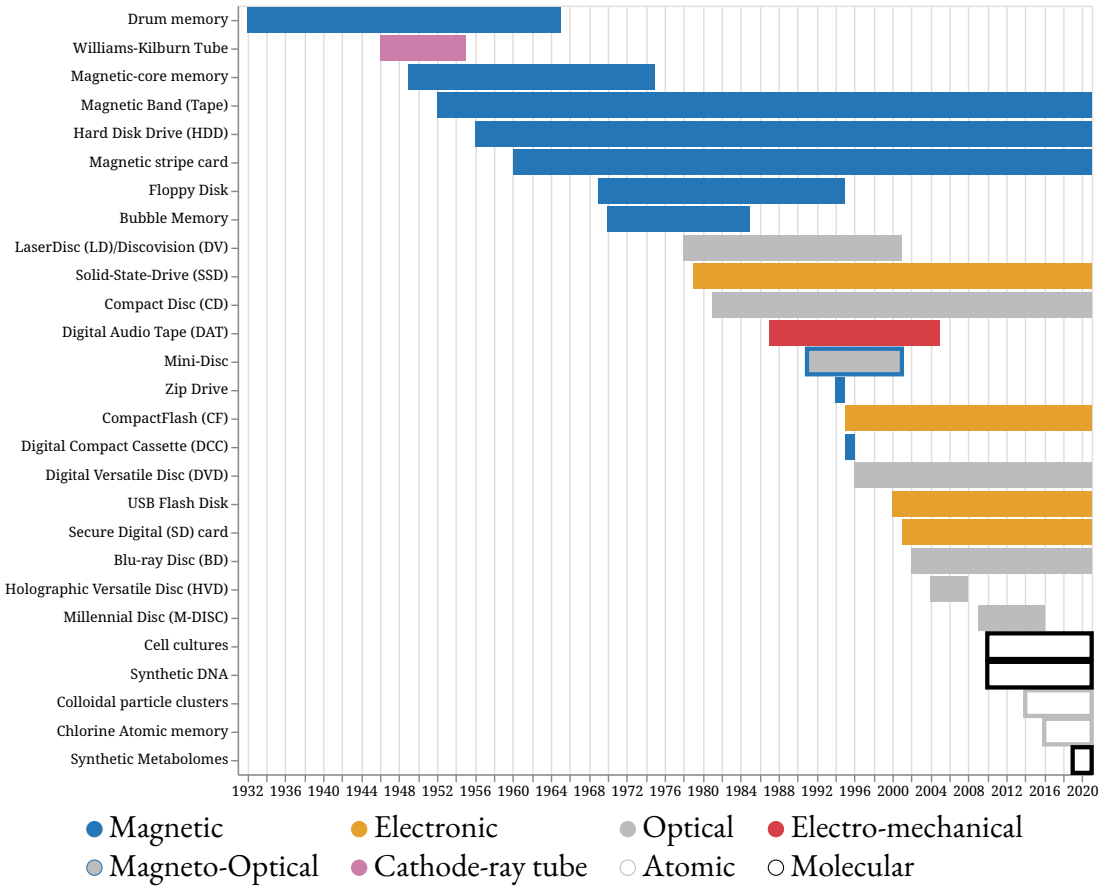


Figure 3.1: Timeline of Storage Media and their Usage. The transition to new storage technologies is observable at multiple time points. Today, various storage media from different eras are still in use.

based solely on the use of storage devices, we observe a clear separation between storage devices that are still in use and those that are obsolete. For example, Floppy Disks became obsolete in the mid-1990s. From the magnetic era, accessible and mutable devices that survived only include HDDs. From the optical and electronic era, all aforementioned optical and electronic media are still being used today with the exception of the Blu-ray Drive which was discontinued in 2019, and later on, replaced with UHD Blu-ray which is now predominantly used. From the molecular and atomic storage era, Chlorine Atomic memory is the only accessible and mutable medium. Although this is the case, many other media have been developed with long-term archiving in mind. This led to a polarization of inaccessible and immutable media (*e.g.*, synthetic DNA, synthetic metabolomes).

3.3.2 Timeline by typology, capacity, and lifespan

In this second subsection, we borrow the aforementioned division of the timeline and report the TOP 3 storage devices in terms of capacity and lifespan properties. Figure 3.2 introduces the capacity of storage media over time on a Log scale.

First, magnetic storage devices proved to be resilient to the ever-growing need for more storage space. However, as electronic technology progressed, the capacity of magnetic-based storage devices became greater and the actual magnetic memory became cheaper. Because of their resilience and price, magnetic devices are widely adopted for archival use. **Hard Disk Drives** (HDDs) are the most developed magnetic-based storage medium. Their capacity increased substantially, from $3,750,000$ bytes of IBM 350 disk storage unit to around 2×10^{12} bytes of modern HDDs[154]. The lifespan of any storage medium is directly influenced by environmental factors in which that medium resides, as well as the frequency in which it is used. It was long thought that higher temperatures might increase the chance of HDD failure, however, recent studies have found that there was no correlation between physical drive temperature and drive failures[243]. Since HDDs rely on mechanical parts for read and write operations, their lifespan is constrained to the quality and durability of those parts. New technologies of disk coating are being introduced to improve the current durability and data density of HDDs. Carbon-based overcoats are replaced with graphene-based ones to achieve a better reduction in friction and provide superior corrosion and wear resistance. This, in turn, enables the potential of increasing data density up to 4 to 10 times[90]. The **Zip drive** was the least long in use compared to all previously mentioned media and devices. Therefore, not much was done to further develop its capacity and lifespan. At the peak of their development, the maximum capacity of the Zip Drive was around 10^8 bytes. The internal structure of the Zip Drive is almost identical to that of the HDD, consisting of read/write heads hovering over a rapidly spinning floppy disk mounted in a sturdy cartridge. That makes the Zip Drive much like the HDD, prone to failure and with a similar lifespan. **Magnetic-core memory** was introduced in the early days of computer systems and engineering. The properties of this storage device, such as non-volatility and random-accessibility,

made it perfect for use as the primary memory of those early systems. The maximum capacity of the Magnetic-core memory was around 10^6 but was not further improved, as the storage medium was replaced with a more technologically advanced static random-access memory.

Second, the optical and electronic era. With the introduction of the first optical storage medium in 1978, the **LaserDisc** (LD) provided a serious alternative for long-term data storage. Its introduction occurred roughly at the same time as the first electronic storage medium, the SSD. As previously mentioned, the LD preceded many optical media. The **Blu-ray disc**, and its modern successor the Ultra HD (UHD) Blu-ray, currently holds the record for highest storage capacity for optical media of this era. When first introduced, the capacity of this storage medium was around 25×10^9, while it now peaks at around 10^{11} bytes. Unlike the organic dye used in optical disc media, found in DVDs and CDs, a different approach is used to encode data on a Blu-ray disk. A combination of silicon and copper is used to create a layer on which the data is engraved, making Blu-ray disks significantly more durable and resilient than DVDs and CDs, with a lifespan of around 150 years. However, this warrants further research since these media did not exist long enough to either confirm or deny such claims. Alternatively, the current record for electronic media or storage devices of this era is currently held by the **SSD** with a maximum capacity of 10^{14} bytes. In second and third place, the **USB flash disk** and the **SD card** share the record with current maximum capacities of around 2×10^{12} bytes. Indeed, electronic-based storage devices do not contain any moving parts and therefore use entirely different methods to read, write, and store data. Research shows that the life span of devices based on flash technology is greatly determined by the usage in terms of data written on those storage media. Furthermore, it has been shown that SSDs were replaced 25% less often than HDDs [272].

The third era is represented by the introduction of many novel media such as cell cultures, synthetic DNA, and synthetic metabolomes. In addition, the Holographic Versatile Disc (HVD) optical medium was meant to compete with Blu-ray disk on capacity and reliability. However, HVD was costly and incompatible with existing or new storage standards, making its adoption problematic. The current storage standard specifies 2×10^{11} bytes of capacity. Yet with no clear demand, such

disks were never put in mass production and the HVD remained in the research phase[93]. This third era is further detailed in the next section.

Over time and across media types, the changes in capacity reached in multiple instances local maximums. For example, optical media have reached a capacity of 10^{11}, as seen in Figure 3.2.

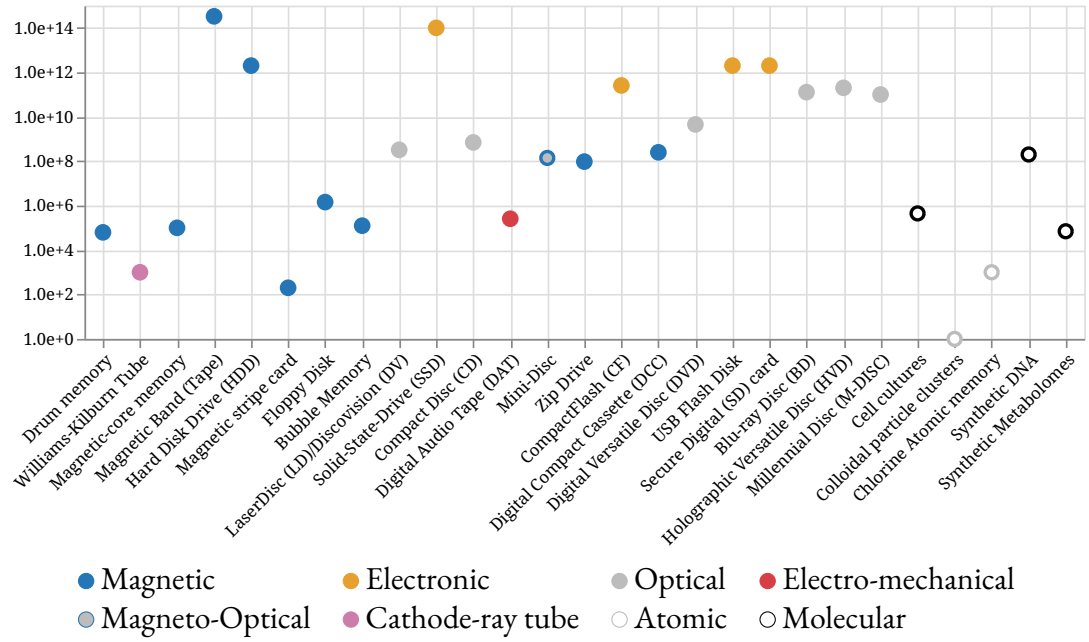


Figure 3.2: The Capacity of Storage Media over Time. Values in bytes are reported on a Log scale with a base 10. It is possible to observe an increase over time as novel media improve in capacity. However, new experimental devices or media are often expensive which limits the storage capacity.

3.3.3 Novel and alternative media

Many novel and alternative media have been proposed in different contexts. They comprise atomic and molecular media. In the case of molecular media, a variety of approaches exist including organic, such as synthetic metabolomes, metallo-organic clusters, synthetic DNA molecules, and inorganic clusters. The most promising molecular medium is **synthetic DNA**, although it requires a variety of steps; *e.g.*, synthesis, encoding, sequencing, error correction, *etc.* Indeed, some approaches figured out that DNA storage systems are too slow to replace HDDs[73]. Yet more recent break-

throughs have made use of hundreds of thousands of short DNA oligonucleotides for encoding small amounts of data [61]. In that approach, naive encoding without error correction was used, which rendered it unsuitable for long-term archiving due to sequencing errors. To avoid problems in sequencing, a Huffman-encoding with a ternary code was employed by another approach [120]. This specifically helped avoid homo-polymeric DNA sequences which are problematic at the sequencing step. Moreover, they introduced redundancy in the encoding, thus enabling a simple error-correcting procedure. Based on these findings, another approach improved data density in the DNA storage medium by employing a kind of RAID (Redundant Array of Independent Disks) system [33]. In addition, a rudimentary random access procedure based on primer sequences was developed. Although it can be used for direct access of parts of the data, it is not possible to perform semantic searches, for which complex index structures are necessary [329]. For this reason, we did not report synthetic DNA as an accessible medium. Further research into improving DNA storage systems made use of different encodings (*e.g.*, fountain codes [97], forward error correction [30], Reed-Solomon codes [124]). In spite of the fact that they are used for error correction, these encodings only correct or compensate until a certain threshold is reached. When the DNA storage molecule(s) are exposed to different reagents or stimuli, error-correcting codes cannot handle any potential degradation (*e.g.*, damaged bases, breaks between individual nucleotides, and fractures in the phosphate backbone). The usage of higher codes such as Galois fields codes could help error- and erasure-correcting codes for reliable DNA storage [168,194]. A comprehensive review of the research literature on synthetic DNA storage and its challenges is addressed by Dong *et al.* [88]. Physical storage of synthetic DNA varies from one laboratory to another. It defines a unit of molecular storage that often relies on sequencing redundancy (*i.e.*, deep sequencing coverage, and having many copies of each sequence) and ranges from, but is not limited to, amber-enclosed spores [1], lyophilized oligonucleotides [120], inorganic silica [124], *etc.* The capacity value for synthetic DNA is experimentally validated and visually reported in Figure 3.2. In this instance, the molecular storage unit is dehydrated synthetic DNA pool.

Significant efforts are being made to use inorganic molecules for data storage. Current efforts are

based on the use of four-color printing of clusters to reveal their extreme nonlinear optical properties. These properties should be distinguishable enough for a spectrometer to read them and thus capture the information they contain. The fundamentals of such molecular clusters are described in several works [89,262,263].

Another aspect of molecular storage is organism-based. This corresponds to maintaining certain **cell cultures**. It effectively relies on integrating short synthetic DNA strands (*i.e.*, oligonucleotides) into a biological organism *in vivo*. In turn, the organisms can store and duplicate the information on an *ad hoc* basis. This special type of medium was researched to insert synthetic DNA fragments encoding data and inserting them into the genome of bacteria, fungi, and plants. This was successfully accomplished in *Escherichia coli*, *Bacillus subtilis*, *Pichia pastoris*, and *Arabidopsis thaliana* [129,298,349]. Estimates for a single gram of bacteria indicated an advertised storage capacity of more than 900 TB. However, an experimentally validated approach confirms using bacterial cell cultures with 445 KB of digital files in synthetic DNA [129].

Synthetic metabolomes are another promising molecular medium. Metabolomes comprise the complete set of small molecules found in a biological system. Unlike DNA and protein molecules, they are small in mass, abundant, and more structurally and energetically diverse [177]. The synthetic metabolomes consist of a mixture of metabolites (*e.g.*, Galactose, Tryptophan, etc). These are spatially arrayed in thousands of nanoliter volumes on a physical multi-well array or plate. In turn, each resulting volume contains a prescribed mixture from a library of purified metabolites, *i.e.*, a synthetic metabolome. This approach demonstrated the storage of many image data. The largest is a 17,424-bit image requiring approximately 70 KB. This image was written into 1,452 mixtures from a 12-metabolite subset of the library [177].

Another noteworthy addition concerns non-genomic molecular media. They have also been demonstrated yet have not been included in this study due to their experimental nature. One type of such media relies on fluorescent dyes on polymer films or the rotaxane molecular architecture [125,335]. Another type relies on creating nano-structures to obtain an etched crystalline quartz or even a thin

diamond layer [172,317].

3.3.4 Industry-based UI and visualizations

From the early days of the computer, the amount of used information in the storage devices was important. Early devices showed the usage of the storage medium in percent or only the occupied space next to the capacity of the medium or both. This information was presented in a textual form and a human-readable format. This refers to encoding the information in ASCII or Unicode text rather than binary data. Many technological breakthroughs, such as nanoscale circuits, enabled more compact and reliable solutions with a higher element density. Various sensors were introduced to different storage types, for example, temperature sensors. This led to the monitoring of meta-information, which in turn allowed the prediction of the device lifespan. In turn, UIs were upgraded with further characteristics such as Temperature or Speed of the storage device. The advancement in graphical processing power also opened up further possibilities. First and foremost, it enabled the visualization of the available information for the capacity property. Later, it was used to visualize partitioning of the storage medium, file structure, usage, *etc.* We divide our investigation into early, pre-modern, and modern operating systems.

Early systems introduced built-in tools that report detailed information about the user's storage devices. These systems comprised the report of textual information for different properties, yet mainly focused on capacity [292]. Such systems align with the pre-modern arrival of the personal computer with Windows 95® or System 7®, and their adoption until Windows XP® and Mac OS X®. The capacity of a storage device was one of the most important properties and was often visualized as a horizontal stacked bar chart displaying used and free storage space. Modern systems have built-in tools and visualize storage capacity with pie charts and some of its variants, like doughnut charts, or even multi-level pie charts, as seen in Figure 3.3.

Along with previously mentioned visualization methods, certain OS offered a built-in representation of used storage space in a tree map chart, as seen in Figure 3.4. In most cases, the visualization of

usage presents an overview first, details second. Generally, the overview is the stacked horizontal bar chart, while details may be observed in visual and textual form simultaneously. Furthermore, supplementary storage device information, for example, SSD temperature or mount point, is presented in textual form. Figure 3.5 depicts this occurrence.

3.3.5 Survey results

As stated earlier, the survey was conducted on two different groups: (a) domain experts in molecular data storage, and (b) the general public. Nineteen responses were collected from the expert group and thirty from the general public.

In the first group, 68.4% of participants identified themselves as male, and 31.6% as female. Exactly 47.4% put computer science as their field of research, followed by biology at 36.8%, then chemistry at 15.8%, then physics and mathematics at 5.3%. Nearly 60% reported having 10 or more years of experience in their respective field, followed by 31.6% who have 0 to 3 years of experience, and the rest having 3 to 5 years of experience. When ranking the TOP 3 properties, and while considering themselves domain experts, participants chose the following in order of importance: 1. lifespan, 2. capacity, and 3. accessibility. These results changed when considering themselves as members of the general public. The properties of accessibility and lifespan switched places, placing accessibility first and lifespan last. All TOP 3 ranking results are also presented in Table 3.2.

Exactly 43.3% of the participants belonging to the general public group identified themselves as male, followed by 50% identified as female, while 6.7% preferred not to state their gender. As for the field of study, 60% reported computer science, 16.7% social science, 13.3% biology, 10% medicine, 6.7% mathematics and the rest was split between economics, and business informatics. The general public was not asked the question about their experience. The results gathered from this group, *i.e.*, the general public, are reported in Table 3.2. In this case, the TOP 3 was: 1. accessibility, 2. capacity, and 3. lifespan.

To support the visualization of the TOP 3 data storage properties, we designed and implemented

Property	The expert pool			The experts as public			The general public		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Accessibility	5	1	**7**	**8**	3	2	**17**	0	6
Capacity	6	**7**	4	5	**8**	2	6	**11**	6
Lifespan	**8**	5	2	2	1	**9**	5	5	**7**
Mutability	0	4	3	1	4	4	2	5	6
Usage	0	2	3	3	3	2	0	9	5

Table 3.2: Survey results. Cells with the highest votes are highlighted. For the expert pool, the TOP 3 properties are lifespan, capacity, and accessibility. When experts are asked about their opinion as members of the public, the order of the TOP 3 changes to: accessibility, capacity, then lifespan. For the general public, the results are the same as when the expert pool stated their opinion as members of the general public.

a User Interface (UI) adapted to both audiences. The methodology for property-based visualizations and the UI is described in the Methods section. The implementation of the UI and the source code used to create all of the figures are uploaded as part of the supplementary material and are also available at the *TVSDS* GitHub repository: `https://github.com/AAnzel/TVSDS`. The survey and its results are provided with this manuscript as part of the supplementary material.

3.4 Methods

The literature search was possible thanks to an array of different scholarly literature databases: Google Scholar, Europe PMC, IEEE Xplore. While the experimental part of the paper consisted of generating a survey, analyzing results, and creating an adequate UI and data visualizations.

3.4.1 Literature search

Relevant search keywords were used for different parts of this paper. A non-exhaustive list of used keywords is: data storage, novel media, storage device, storage medium, DNA storage, storage visualization, molecular medium, molecular storage. Apart from the literature search, our paper also includes knowledge gathered from several archives, for instance, the IBM Archives, The Internet

Archive, and the Museum of Obsolete Media. For certain data storage technologies, like HDDs, HVDs, SSDs, we also incorporated relevant facts stated in their storage standards, that are publicly available online.

3.4.2 Ranking survey

For the survey, we used Google Forms and presented it as follows. First, a brief explanation of the survey was given followed by a one-sentence description of each data storage property: accessibility, capacity, lifespan, mutability, and usage. Second, participants were asked to rank them in a TOP 3 fashion. The expert group was asked to rank properties as experts in the molecular data storage domain and as members of the general public. Third and last, demographics data was collected. We gathered the results of the survey and proposed a new UI for the property-based visualizations. The survey results laid out in Table 3.2 are divided into three groups: (a) the domain expert opinion, (b) their view as a public, and (c) the opinion of the general public.

3.4.3 User Interface

By coupling the results from the survey, the literature search, and the visualization standards to display data storage information, we propose an adapted user interface and state-of-the-art visualizations. The proposed User Interface (UI) consists of two views: the basic view to suit the needs of the general public, and the advanced view for domain experts. The default home view is the basic view. It serves as an overview by presenting the accessibility (sequential versus random), the usage or how much of the media is already in use (in percentage), and the capacity of the media in kilobytes (KB). In addition, a file hierarchy is presented in textual format. The basic view component of the UI can be seen in Figure 3.6.

The advanced view can be toggled by a select box button to display further details. It specifically includes additional properties (*e.g.*, lifespan) and further details for the usage and the file hierarchy. This is shown in Figure 3.7. The advanced view benefits from mouse hover and mouse click

events to provide granular details for each visualization. For example, the mouse hover event for the capacity/usage visualization reports the actual size in kilobytes (KB) of a specific file type category (*e.g.*, audio files), as seen in Figure 3.8. The mouse hover event for lifespan visualization gives detailed information on lifespan estimation, as shown in Figure 3.9.

On the other hand, selecting a directory by a mouse click within the directory hierarchy visualization makes the directory expand to occupy the whole pixel space of the visualization. The chosen directory is now considered a top-level directory, and the visualization shows the hierarchy of its subdirectories, with all previously top-level directories laid on top of the visualization. This is shown in Figure 3.10. The UI reports the properties by relying on the TOP 3 ranking results, as seen in Table 3.2. In light of the survey results, the order of appearance of the visualizations is adjusted to fit the reported importance of the studied properties depending on the audience, and is updated accordingly in its corresponding view (basic vs. advanced). The visualizations present in both UI views dynamically adapt to the current page width, maximizing the ink-to-pixel ratio.

3.4.4 Property-based Visualizations

By considering the state-of-the-art data visualizations for data storage, we propose 3 main visualizations for the ranked properties: usage, capacity, and lifespan. We follow the nested model of visualization to describe the visual encoding[222]. The properties are encoded as follows: capacity as a real number, while usage is reported as a real number and in percentage (%), lifespan as an integer, and accessibility uses the boolean data type (0 or 1). Textual information is represented as UTF-8 text strings of variable lengths.

First, for both usage and capacity, a horizontal bar chart is employed. Bar charts are very effective at displaying part of a whole, and visually comparing metric values across different subgroups of the data at hand. The basic view shows only two parts or two stacks, free and used space. While the advanced view details which kind of file type categories (*i.e.*, audio, video, documents, other) occupy the storage space, and by how much (in percentage and in KB). Except for the free space stack, each

stack is visually encoded using the area and color channels. The area channel represents the used space: each part as a percentage to a whole (*i.e.*, maximum storage capacity in KB). The color channel relies on the categorical or nominal encoding of each file type category. Four file types are considered excluding free space: audio, video, documents, and other are mapped to colorblind-safe categorical colors: `#A6CEE3`, `#1F78B4`, `#B2DF8A`, and `#33A02C`, respectively. The encoding of the color channel follows the state-of-the-art rules to colorize a data visualization [135]. Moreover, each stack benefits from textual overlays to report the used space in percentage (%). This is presented in Figure 3.8.

Second, we used a whisker chart for the lifespan property. It combines a bar and whiskers as well as a vertical line. The latter spans the chart height and encodes the quantitative property (*i.e.*, estimated lifespan) of the medium by relying on the position channel. The estimation of the lifespan is encoded using the area (bar) and the position (line). Since the lifespan of a medium largely depends on various external factors, the whiskers represent the confidence interval of this estimate, while the 95% interval is represented as a gray area. The lifespan property is also reported in a textual form as seen in Figure 3.9: the estimated lifespan is 42 years.

Third and last, and as integrated into different operating systems, we implemented a directory hierarchy visualization by using the tree map visualization. The rationale of this visualization method is to maximize the pixel space at our disposal. It is highly efficient since it uses a space-filling technique and relies on creating multiple rectangular areas [23]. Each directory represents a rectangle, where hierarchies between different directories are encoded with containments to create a nested layout. A rectangle is encoded by its area and color channels. The larger the area, the larger the rectangle, the larger the directory. The color channel follows the aforementioned nominal encoding in the usage/capacity chart with 4 file type categories. The name of a directory is encoded in textual form, with each name contained in a rectangle representing that directory, as seen in Figure 3.10. Additionally, the accessibility property is not visualized but reported in a textual manner. This can be seen in Figure 3.7. The implementation is done using Python 3.9, Altair 4.1, Plotly 4.14.3, and Streamlit 0.82.0 [266,270,275,324].

3.5 Conclusion

Existing storage technologies are insufficient for the long-term storage of the large amounts of data generated in all areas of life. Novel and alternative media propose to extend the existing storage capacity with molecular storage media and devices. Hence, also reducing the risks for information loss in long-term data storage. Although technological advances are being made for molecular storage media types, there has been a lack of efforts for standardization. This is especially relevant before the adoption of upcoming storage media, such as synthetic DNA. Indeed, standards not only improve ways to standardize information storage for long-term archiving (*i.e.*, lifespan), but also help researchers meet certain criteria when developing novel storage media. It is also important to consider the ways we digest the information that is conveyed by storage media properties. By means of ranking, our survey permitted us to identify relevant properties for domain experts and members of the public. Moreover, thanks to an analysis of the industry-based UI and visualizations, we were able to observe and converge to specific design choices. For the bar chart, the use of the color channel was mapped to file type categories (audio, video, documents, and other), while the area channel depicts the used amount of the available capacity (often in percentage or in GB). On one hand, the horizontal stacked bar chart had been a preponderant choice for the display of the usage of a medium's capacity. On the other hand, the tree map had also been a common choice to display file directories and file hierarchies. For the tree map chart, the color channel encodes file type categories, and the area channel encodes the amount used by said files. In addition, the containment provided by this space-filling visualization method helped lay out hierarchies among different directories. Owing to these industry-based and widely accepted charts, we developed a user-settable approach to toggle between an overview and details. This corresponded to the basic vs. advanced view. The latter provided additional data, including further annotations, and unraveled further details about the storage medium.

3.6 Discussion

First, and thanks to our survey, we were able to separate the different storage properties by relevance. Besides this, we found that the TOP 3 properties converged for the general public and when the expert participants considered themselves as members of the general public.

Second, by having the survey results, we were able to propose a new UI that is generalizable and could also be specifically used for molecular storage media or upcoming novel media. Even though accessibility was chosen as the most important data storage property, it might be reasonable to shift its position to second place. The rationale is that visual information is more relevant and has greater power. From a historical point of view, it is more reasonable to present information rather visually than textually. That is to say, a user should be first confronted with some type of graphical or visual representation. That is why we first present a user with the capacity/usage visualization first, then we present the other properties as ranked by the survey results.

Third, the need for standards extends beyond the currently provided example. Indeed, as seen in the proposed UI and visualization, certain properties have an intrinsic uncertainty. This may either be due to the fact that the property relies on estimation, or that a property is measured. In the general case of the lifespan property, estimates vary to include multiple storage and usage conditions such as temperature, number of read/write, recycling of the media, mechanical movements, *etc.* We argue that estimates could benefit from standardization so that precise values with certain confidence intervals may be reported. In this regard, the literature lacks evidence-based estimates. In the example case of the lifespan property of synthetic DNA, the theoretical limit is supported by the oldest known preserved DNA in existence, aging approximately one million years[319]. Moreover, the experimentally validated capacity for synthetic DNA seems small in relation to the theoretical limit. This limit is estimated to reach multiple orders of magnitude higher than the presented capacity value. However, the current costs of storing data inside a DNA molecule greatly limits the experimental validation of this maximum. Furthermore, methods and algorithms developed in the field

of genomics may benefit current data storage approaches using DNA. This includes, but is not limited to, data compression and indexing [48,84,282]. In the example case of a synthetic metabolome, the capacity property depends on the number of metabolites present in a metabolome. That is to say, a measurement is made. Such aspects of uncertainty and the way such information is presented for general consumption remain an open question.

Fourth, a noteworthy example of a novel and experimental storage medium is the Data Sticky. Although very promising we excluded it from our findings of the literature search. Data stickies are inspired by sticky notes, such as Post-Its®. A realistic implementation exists, namely Post-Bit, where a small e-paper device stores multimedia contents and allows for paper-like manipulations [205]. They combine the affordability of physical tiny sticky memos, the digital handling, and the display of information using electron ink or e-ink [204]. Once this ink settles into an image, the display reflects light just like ordinary paper; as a non-volatile medium. Recent advancements in nanotechnology proposed a larger storage capacity using graphene paper and e-ink [245]. Estimates put such upgraded data stickies between 4 and 32 GB. There have also been considerable efforts to create nanoscale data storage using graphene. A promising strategy addressed high precision writing and drawing on graphene nanosheets by manipulating electrons with a one nanometer-based probe [354].

Fifth and last, even though our UI proposal is applicable to molecular data storage media, more visualization research is warranted for medium-specific properties. Our work herein reported properties that are in the broadest sense shared and relevant. Since molecular data storage media is in active development and research, we can expect some new storage medium-specific properties that could be more important than the properties we reported. That means that UIs and visualizations should evolve and adapt to the new storage media types and all of the important, specific properties those media types may have.

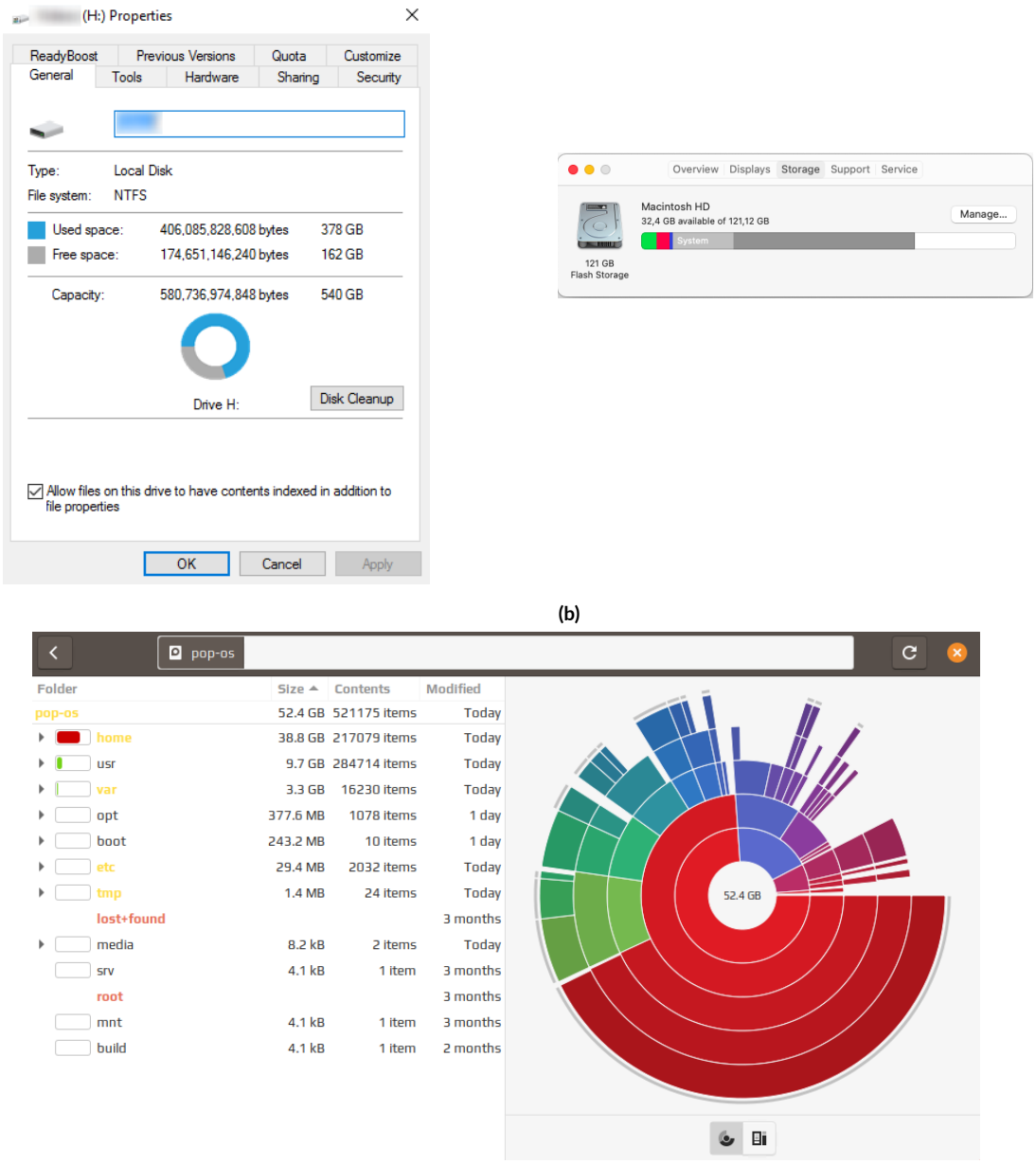


Figure 3.3: Screenshots of data usage visualizations for different Operating Systems. Part-to-a-whole visualization variations are used. (3.3a) Donut chart representing the local disk usage on Windows 10. (3.3b) Stacked bar chart (horizontal) representing the amount of memory used by different file types on Mac OS X Catalina. In the most recent version, Big Sur, only used space is shown in relation to the available space. (3.3c) Sunburst diagram representing the amount of memory used by different directories on Linux (Pop!_OS 20.10).

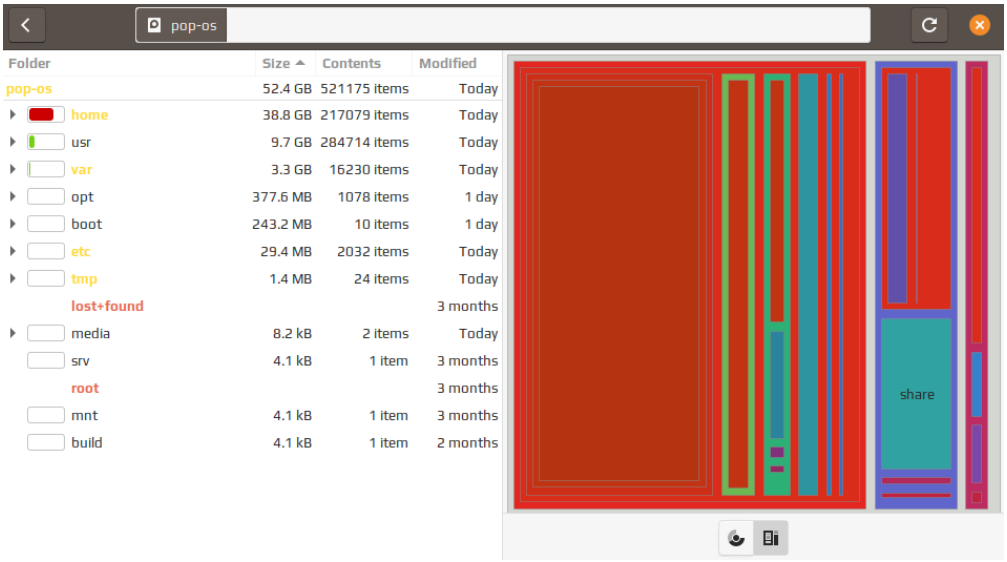


Figure 3.4: Alternative visualisation on a Linux distribution. Tree map chart representing the amount of memory used by different directories on Linux (Pop!_OS 20.10).

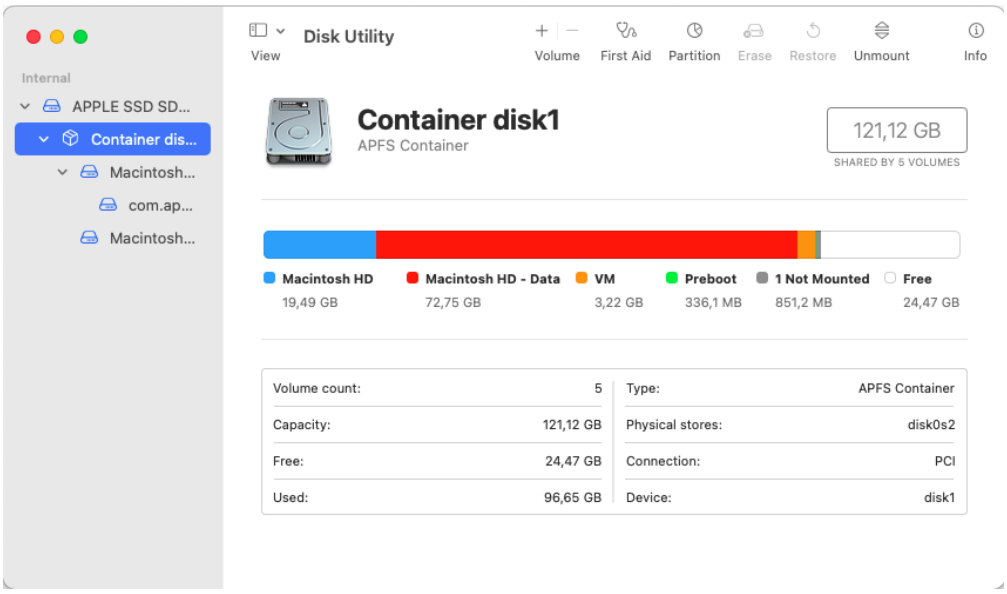


Figure 3.5: Additional information on Mac OS X. Supplementary data of how the storage device is partitioned and other important information such as the type of physical connection and the name of the disk are reported in textual form and tabular format on Mac OS X.

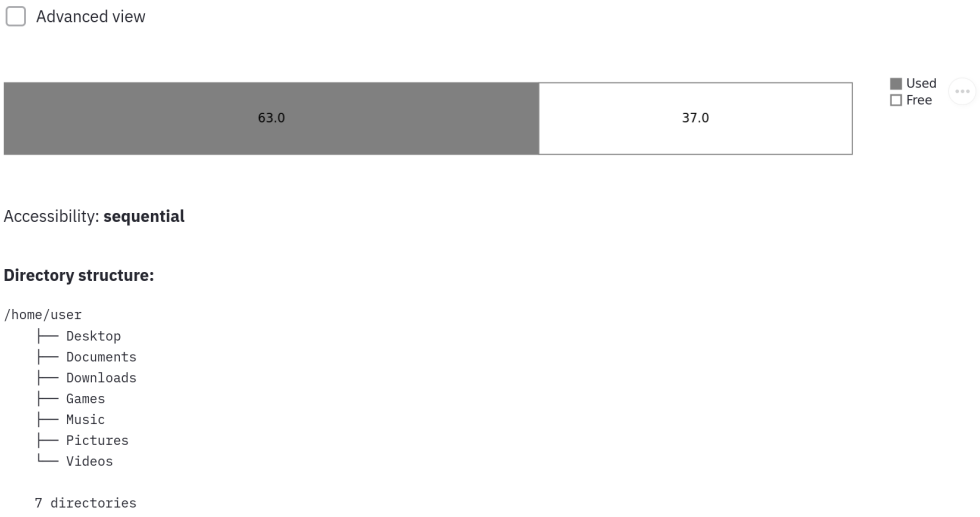


Figure 3.6: The basic view. It provides an overview of the used versus the free storage space as well as the main properties that are required for the general public.

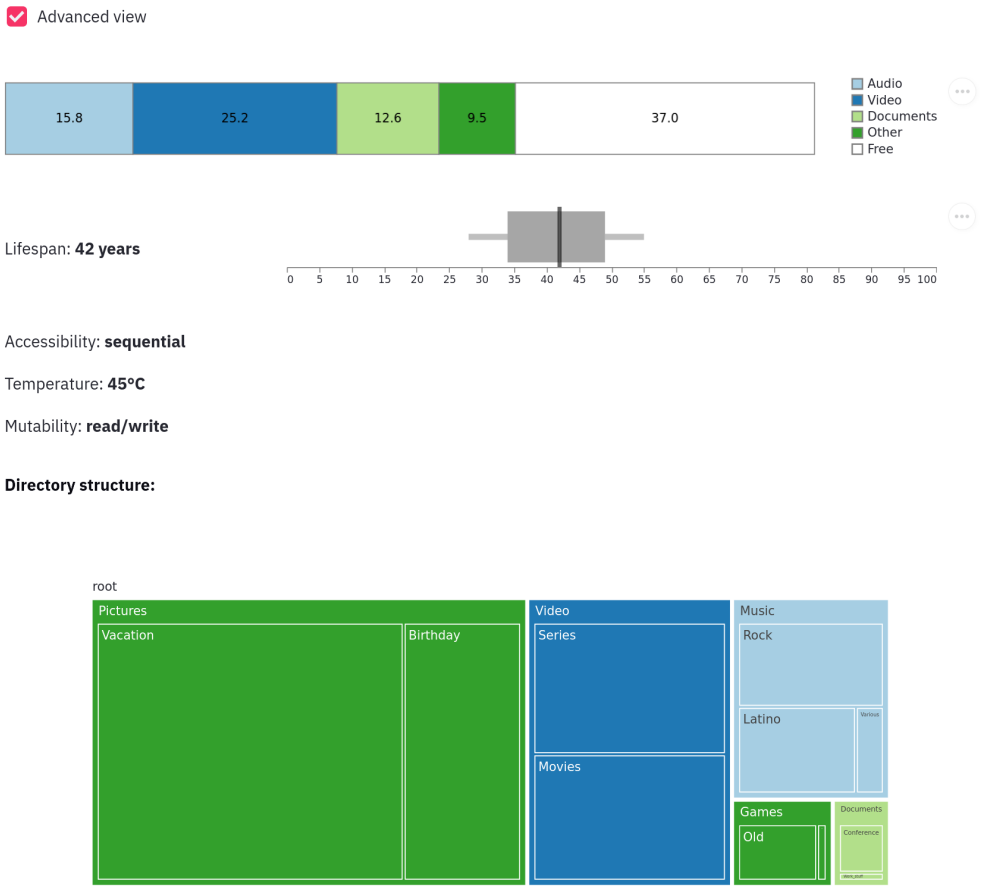


Figure 3.7: The advanced view. If the *Advanced view* checkbox is ticked, the storage medium properties are displayed in order of importance for the expert audience.

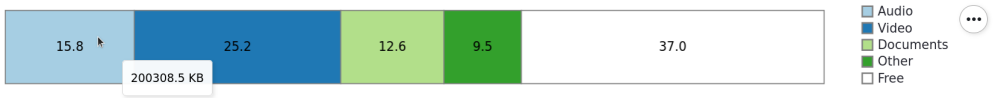


Figure 3.8: Capacity/usage tooltip in the advanced view. Supplementary information is shown while mouse-hovering over stacks.

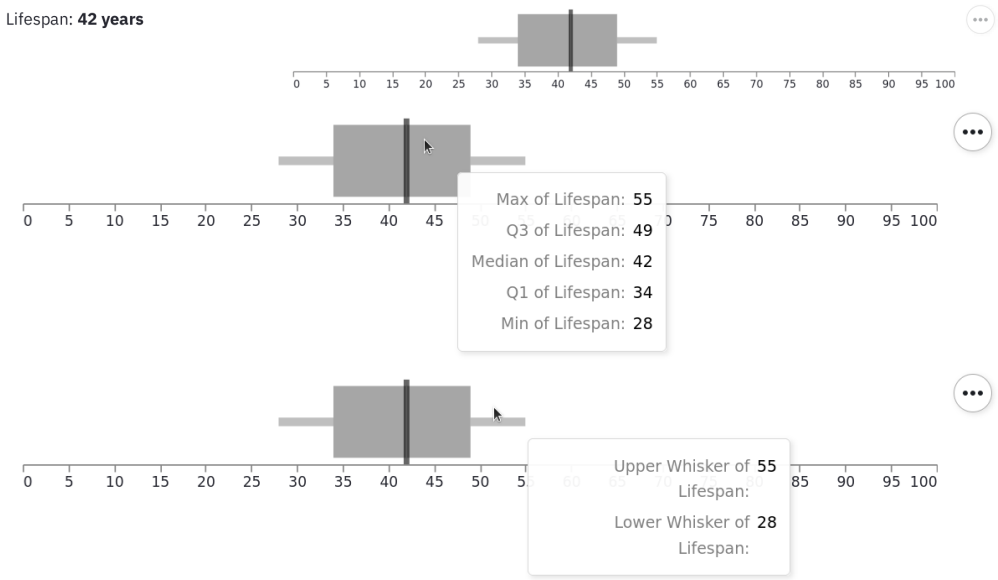


Figure 3.9: Lifespan visualization. The lifespan property is presented with textual and visual information. Detailed information of lifespan estimation is presented in an overlay window while hovering over the gray area, as well as over the whiskers.

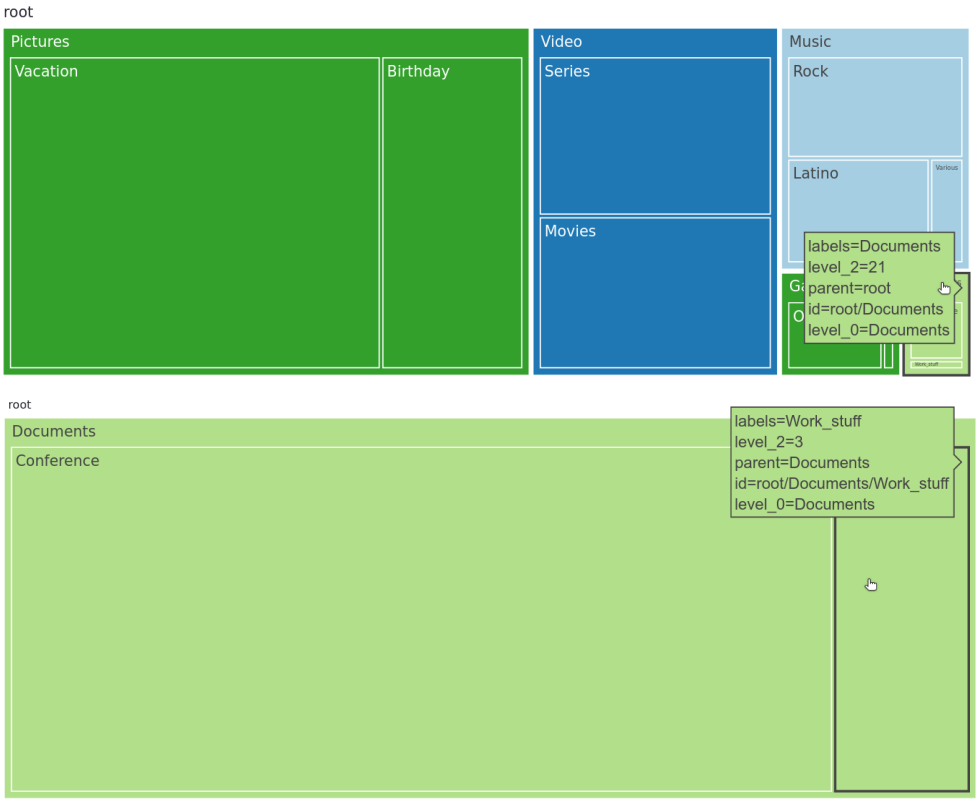


Figure 3.10: Directory hierarchy visualized with a tree map chart. Each rectangle is colored according to the file type category present in the corresponding directory. In the image below, the *Documents* folder is now a top-level directory, with root as its parent directory. Supplementary information is shown while mouse-hovering over rectangles.

4

High-Dimensional Multi-Modal Time-Series Data: Challenges and Opportunities for Analysis, Visualization, and Interpretation

Status

Published as: Aleksandar Anžel, Dominik Heider, and Georges Hattab. Movis: A multi-omics software solution for multi-modal time-series clustering, embedding, and visualizing tasks. *Computational and Structural Biotechnology Journal*, 20:1044–1055, 2022. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2022.02.012. URL https://www.sciencedirect.com/science/article/pii/S2001037022000526

Copyright Notice

Contribution

I designed, developed, documented, and evaluated the tool. I collected, processed, analyzed, and visualized the data. I wrote the manuscript, discussed the results, and revised the manuscript.

4.1 Preface

ecent advancements in sequencing and computational technologies have led to the production of large high-dimensional multi-modal time-series data sets in the biomedical domain. However, exploring such high-dimensional data sets with multiple modalities, discovering anomalies, finding patterns, and understanding their intricacies, is challenging. To effectively explore such data, adopting a problem-driven approach that integrates expertise in biomedicine, bioinformatics and computer science is crucial. To address this need, it is important to

develop modular exploration tools for time-series multi-omics data that provide user-friendly interfaces and facilitate the equal participation of different omics subtypes for analysis and visualization. Such tools should produce task-specific, reproducible, and publication-ready visualizations. By using an open-source software-based framework, such tools can be extended to accommodate different analytical tasks and integrated with existing software. With a problem-driven approach, it is possible to analyze, interpret, and visualize high-dimensional multi-modal time-series data in the biomedical domain, leading to new discoveries and insights.

4.2 Introduction

igh-throughput technologies allow us to generate large amounts of data that could be used for medical and biological research. Multi-omics data sets have been extensively used to provide new insight into certain diseases such as cardiovascular diseases[186], type 2 diabetes[357], cancer[52], and infectious diseases[138]. With the new technologies mentioned above and technical advances in computational power, data sampling could become more granular. Data sets taken at one point in time can now be easily extended by creating new data sets at other time points, providing a more detailed picture of the underlying biological phenomena. These types of time-series data sets are becoming more common and are often explored using machine learning methods[227]. Although the demand for integrative, analytical, and explorative tools for such data is high, only a handful exists, e.g., TIMEOR[67], PyIOmica[87], and Functional Heatmap[340]. Functional Heatmap was developed as a web-based tool for time-series transcriptomics data sets. Analysis results can be exported in textual format or visually thanks to data visualizations, yet only using heatmaps or parallel coordinate plots. TIMEOR was also developed as a web-based tool for defining regulated gene networks from gene-related time-series data sets using RNA-seq, and protein-DNA interaction (such as ChIP-seq[165] and CUT&RUN[285]) techniques. In contrast, PyIOmica was developed as a Python[266] library with the ability to work with different time-series omics data sets, like proteomics, metabolomics, etc. It also includes gene ontology (GO)

and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. PyIOmica currently represents the most complete solution for working with time-series multi-omics data sets. Still, as a library, it does not allow domain specialists like biologists and biophysicists to conduct an exploratory data analysis. Besides that, none of these tools allow a step-wise, easy-to-follow pipeline, providing both control and freedom to explore the data and discover irregularities or trends in the data. With the challenge of multi-modal and temporal data, the analysis of multiple omics should provide multiple aspects that vary in scope and detail. For example, a side-by-side view of the outcome of each analysis integrating the same time span is necessary. To overcome these limitations and reduce the separation between each omic-type data, we developed the Multi-Omics VISualization (MOVIS) tool. MOVIS is a web-based, modular tool that enables easy exploration of time-series multi-omics data sets in a side-by-side fashion. In turn, it enables both developers and domain specialists to formulate and test their hypotheses. While the modularity of the core components enables developers to separate and recombine low-level functionalities, it supports requests for additional functionalities or omics-specific tasks. By means of modularity and using open-source libraries, it can be easily extended to accommodate new use case scenarios. To our knowledge, MOVIS is the first freely available time-series multi-omics data exploration tool and a pipeline for creating publication-ready visualizations.

4.3 Approach

MOVIS consists of three distinct parts: (1) a graphical web interface, (2) a data analysis core, and (3) a visualization canvas. The interactive graphical web interface is built on the open-source framework Streamlit (`https://streamlit.io/`). The user interface (UI) allows the user to split the screen into multiple views. Each view corresponds to one of five omics (genomics, proteomics, transcriptomics, metabolomics, physico-chemical data) available to work with. This functionality is presented in Figure 4.1. UI also consists of a side panel used for navigation and shows basic information about the tool. The whole workflow is divided into five well-

defined parts: (1) the original data set presentation, (2) optional creation of a new data set, (3) optional data set filtering, (4) optional data clustering, and (5) data set visualization. Some steps of the workflow are shown in Figure 4.2. The data analysis core is responsible for five sequential core steps: importing, sanitizing, filtering, analyzing, and visualizing all data sets. The core works with five different types of omics data. In the first core step, i.e., importing, genomics, and proteomics data sets can be provided as archived FASTA files (ZIP, TAR, etc.) or as precalculated tabular (CSV or TSV) files. For the genomics data, archived GFF, KO, and Depth-of-coverage files are also supported. The latter is available for transcriptomics as well. Metabolomics, transcriptomics, and physico-chemical data sets can be provided as tabular files. In the case of transcriptomics data, users can upload multiple tabular data sets. However, each data set must have the same set of columns with precisely the same names. MOVIS concatenates these data sets into one unified data set with one new feature (column) named *Type*. This column contains the names of all user-uploaded files. Multi-tabular functionality is provided to enable more accessible work with multiple biological and technical replicate files simultaneously, which is common when researching gene expression. The sanitizing step is dependant on the data set format provided to the core. For archived data sets, the sanitizing step consists of unpacking the archive, checking the validity of the file names, and correcting them if they do not adhere to the naming rules. We created these rules so that file names could hold standardized temporal information of each file. If a data set is of a tabular format, the sanitizing step consists of various data quality checks. In this case, temporal information must be included as a feature of the data set. The filtering step is present only for tabular data sets. It offers a way to filter a certain time-series period and remove one or more rows or columns from the data set. The data analysis step is the central and most intricate part of the data analysis core. It provides embedding and clustering functionalities, as well as creating a new physico-chemical data set for archived FASTA data sets. The visualization step is responsible for visualizing and additional filtering of the data sets. It also implements interactivity to the resulting visualizations in the form of a tooltip, brushing, and/or spanning and zooming. The visualization canvas provides a unified representational space for all created visu-

alizations. It enables users to export the created visualization in several formats. Each visualization can be exported as a PNG or SVG file or a Vega-Lite [270] source specification. This functionality is presented in Figure 4.3. Currently, MOVIS supports nine different visualizations of time-series data sets: *Correlation heatmap*, *Time heatmap*, *Multiple features parallel chart*, *Scatter-plot matrix*, *Scatter plot*, *Two features plot*, *Feature through time*, *Whisker plot*, and *Top 10 share through time*. The user could use each of the visualizations for any data set type without any restrictions. However, some basic data knowledge is advised in order to choose appropriate visualizations for certain data set types. The visualization canvas, along with multiple visualizations, is shown in Figure 4.4.

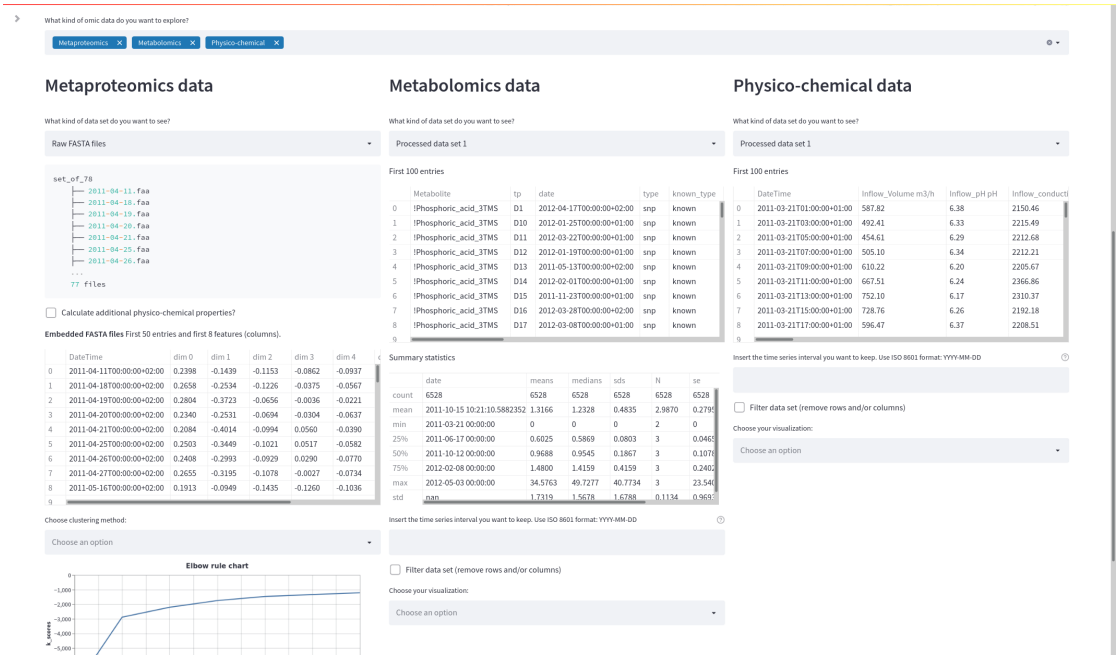


Figure 4.1: Split view of multiple omics. One of the most powerful functionalities of MOVIS is the ability to represent each omic-type in its own view space. This allows the user to inspect and explore multi-omics data sets at the same time.

4.4 Methods

MOVIS is built using Python and additional libraries like Pandas [235], Numpy [131], Scikit-learn [239], Biopython [64], Gensim [254], Altair [324], and Streamlit. Tabular data sets (CSV or TSV) are internally imported as Pandas Data Frame structures. Archived data

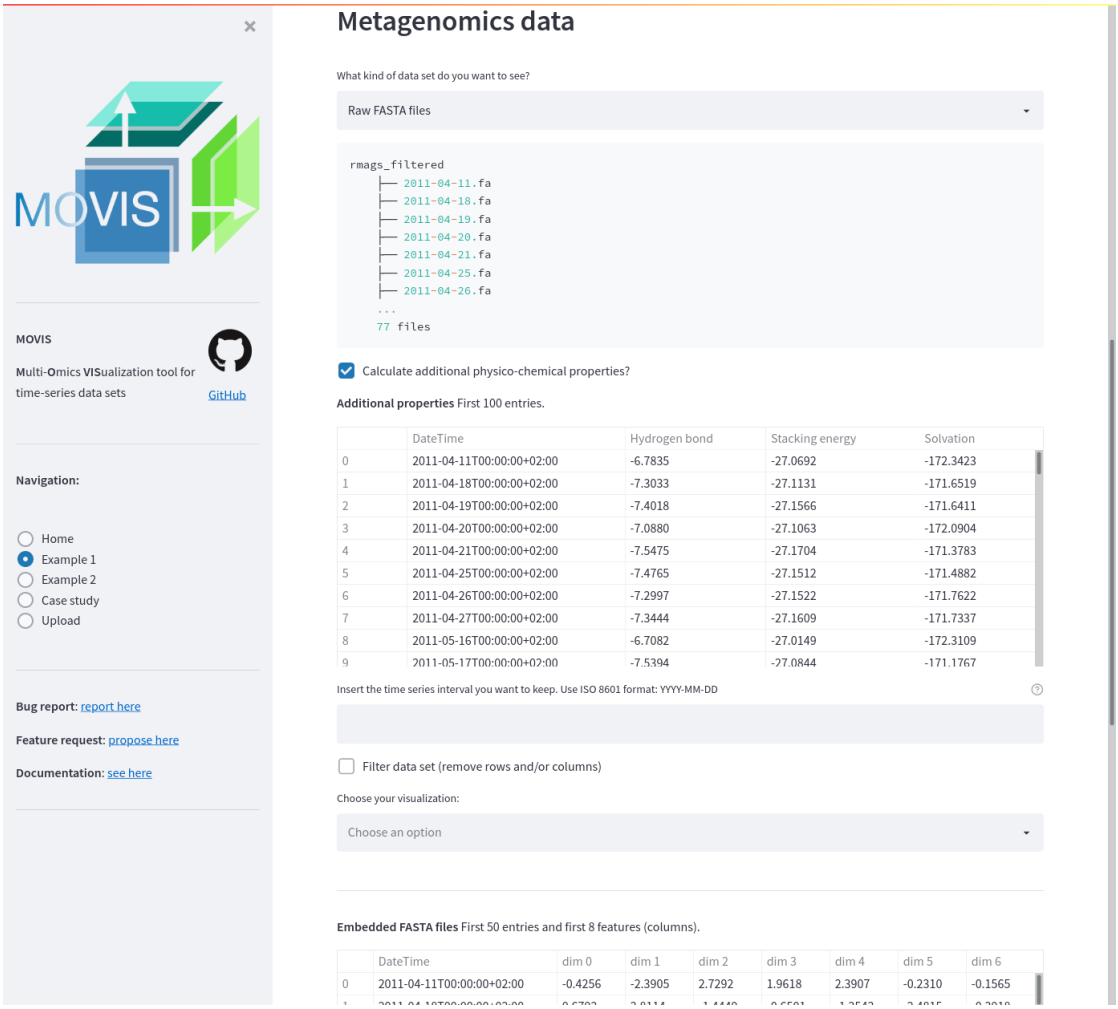


Figure 4.2: Step-wise process of the MOVIS workflow. Data exploration and visualization of each omic are divided into multiple steps. This figure shows steps (1), (2), and (3) for the metagenomics data set.

sets are first unpacked and then cleaned up using built-in Python libraries. If a FASTA data set is selected for proteomics or genomics, the web interface provides an option to calculate an additional physico-chemical data set. These properties are calculated with task-specific in-house algorithms by importing each FASTA sequence, one by one using Biopython, and then processing them. As of the time of writing, the additional data set created out of genomics FASTA data set contains three physico-chemical features — *Hydrogen Bond*, *Stacking Energy*, and *Solvation*. If an option to create an additional physico-chemical data set is selected for the proteomics FASTA data set, the data set

consists of 43 physico-chemical features, e.g., *Molecular Weight*, *Isoelectric Point*, *Instability Index*, etc. The newly created data set incorporates the temporal dimension of the source data and can also be visualized as is. We integrated various methodologies into MOVIS to solve the embedding tasks. One of the central algorithms is the Word2Vec algorithm [216,217] used to embed archived FASTA files into 100-dimensional vectors. Such embedding supports nucleotide- and amino-acid-based sequences. The dimension of the output vector was chosen empirically. Because each FASTA file may contain multiple sequences, we first embedded each sequence in a 100-dimensional vector and then averaged them to create one vector representing the FASTA file containing those sequences. This process results in a tabular data set containing the same number of rows as FASTA files in an archived data set and 101 columns. Each column contains the temporal data, and the other 100 include the 100-dimensional vector embedding. The clustering step is presented with two clustering algorithms as options: KMeans [17] (distance-based) and OPTICS [10] (density-based). By having two different methods that achieve the same goal, different types of inductive principles are covered. The visualizations are created according to the nested model of visualization [221] and follow good data visualization practices [135]. That is to say, design considerations are taken to accommodate color blind users and enhance the accessibility of the visualizations to a broader audience. Other available data visualizations are reported in the supplementary material. Three different dimensionality reduction techniques are included to visualize clusters, namely PCA [218,309], MDS [32], and t-SNE [318]. Although each of these methods aims to reduce the dimensionality of the data, their goal is different. Therefore, more choices provide a greater opportunity to distinguish patterns and anomalies in the data. The effect of choosing different dimensionality reduction techniques to visualize clusters of the same data set can be seen in Figure 4.3. In this Figure, PCA was used for the upper, and MDS for the lower visualization.

4.5 Results

To demonstrate the usability of MOVIS, we present a case study based on one of the available examples. The case study is also built-in into MOVIS as one of the options on the navigation sidebar. As presented in MOVIS, the case study overview can be seen in Figure 4.5 and 4.6.

4.5.1 Case study — Introduction

We focused on the built-in *Example 1* [141] (named as in MOVIS) that contains metagenomics, metaproteomics, metatranscriptomics, metabolomics, and physico-chemical data from the Biological WasteWater Treatment Plant (BWWTP). The data was collected in situ, at weekly intervals, and over 14 months. The end goal of this case study was to reveal if there were any niche types, and if there were, how did they respond to the substrate changes. To limit the scope of this use case, we proposed only using the functional aspects of the metabolomics data [141].

4.5.2 Case study — Main findings

Even though BWWTP operation is a controlled process, factors such as aeration cycles, seasonal changes in temperature, and composition of inflow wastewater fluctuate [164]. The physico-chemical factors may have a meaningful impact on population dynamics and linked process efficiency [344]. Therefore, the first step of our case study was to inspect relevant physico-chemical properties of the wastewater and determine major shifts, if any.

Physico-Chemical data

To demonstrate the function and utility of MOVIS, we selected the *Physico-Chemical data* and, more specifically, the *Processed data set 1*. The selected data set contains 34 different physico-chemical properties with a 2-hour sampling rate. The properties of interest for this case study are *Volume_aeration*

m3/h, *T C* (Temperature in Celsius), and *Inflow_conductivity µS/cm*. We visualized the properties of interest using MOVIS in Figures 4.7a, 4.7b, and 4.7c, respectively. We chose *Feature through time* visualization to accomplish this. Even with a lot of noise present, Figure 4.7a showed the shape of a sine curve with three distinct local maxima (around June 2011, November 2011, and April 2012) and two distinct local minima (around August 2011 and January 2012). Figures 4.7b and 4.7c had less noise, with the latter figure showing a distinct increase in temperature near the end of the sampling period. In order to further examine the irregular behavior of temperature and conductivity, we visualized them using *Time Heatmap* in addition to the existing chart types. The new visualizations are presented in Figures 4.8a and 4.8b, respectively. Provided with more granularity, we were able to see a detailed picture of each property. Figure 4.8a showed a steady increase in temperature that peaks in September of 2011 and then slightly decreases until the start of December 2011. Then, we have a rapid decrease in temperature that lasts until March 2012 followed by a rapid increase that peaks in the beginning of April 2012 with temperatures going as high as 29.62°C. We efficiently inspected values by using the interactive tooltip feature. It appears upon the mouse hovering over the cells of interest. On the other hand, Figure 4.8b showed high but steady values from the beginning of sampling and up until the last week of August 2011. After that, we found a swift decrease and stabilization of values that continues throughout the time series.

Metaproteomics data

Integrated meta-omics approaches hold the potential to resolve niches of microbial populations in situ [141]. Therefore, we shifted our focus to metaproteomics data to identify microbial clusters, if there were any, using only raw FASTA data. To ascertain the presence of microbial clusters, we selected *Metaproteomics data*, and then *Raw FASTA files*. When MOVIS completed embedding FASTA files, we selected the *K-Means* clustering method and using the *Elbow rule chart* we selected three as a number of clusters (centroids) for our clustering method. Then we inspected the evaluation window of the selected clustering method. With a silhouette score [267] of 0.495, we acknowl-

edged that our method was successful, which was further corroborated by other available evaluation scores (e.g., The Davies–Bouldin index (DBI)[81], and The Calinski-Harabasz (CH) score[42]). Then we chose to visualize our data using two different dimensionality reduction techniques in order to determine which one gives a better visual outcome. The selection of *PCA visualization* and *MDS visualization* resulted in Figures 4.9a and 4.9b, respectively. Both figures provided us with a visual way to evaluate chosen clustering method. As can be seen on both Figures 4.9a and 4.9b, their upper-left corner showed a mixture of class-0 (circles) and class-2 (triangles), which indicated that K-Means had problems with clustering data embeddings that occupy that space. However, the clustering was successful since most data embeddings were placed in cloud points that have been determined to be in proximity and define a cluster. Inspecting the color gradient of the visualization marks allowed us to discover even more — samples clustered in the class-1 (rectangles) came in majority from the later time of the sampling period. The same could also be said for the class-2 samples, while class-0 samples came from a more dispersed sampling period. Mouse-hovering over each sample allowed us to determine the exact time that sample was collected. MOVIS also supports calculating amino-acid based physico-chemical properties of the metaproteomics data set, which could uncover an even more detailed picture of the underlying phenomena. For the sake of brevity, we did not select that option.

Metabolomics data

Since a significant shift in substrates of the influent wastewater sludge can alter the community composition[192], we moved our focus to the *Metabolomics data set*, and more specifically the *Processed data set 2*. The selected data set is of composite nature, which means that it contains multi-omic information. Almost 95% of the data set represents metabolomics data, and the rest is physico-chemical data. Pre-combining omics data in such a fashion allows MOVIS to tap into the integrative aspect of the multi-omics nature. That aspect is planned but not yet directly available in MOVIS. Next, we selected *Time heatmap* visualization and chose feature named *value* as a *quantitative*

color feature, *param* as a *y-axis feature*, and *Diverging* for the *color scheme*. Our selection resulted in Figure 4.10. Further inspection of Figure 4.10 revealed substrate shift happening from early to mid-November 2011 and early to mid-December 2011, with noticeably higher values in between. The substrate shift was defined by higher values of mainly non-polar metabolites, as well as polar metabolites, among which are putrescine and various disaccharides. After the end of December 2011, substrate levels normalized, and the community transitioned back to the pre-disturbance state.

Metagenomics data

One way of estimating population abundance is by using metagenomic depth-of-coverage. Since MOVIS is not explicitly designed to work with meta-omics data, no taxa linking is currently enabled. However, by inspecting average depth-of-coverage values, we could get some insights into the overall population dynamics over time. Therefore, we selected *Metagenomics data*, and then *Depth-of-coverage*, which presented us with a directory hierarchy of the underlying data set. After MOVIS automatically calculated important statistical values of the data set in use, we visualized results using *Whisker plot*. The visualization mentioned earlier can be seen in Figure 4.11. The third quartile (Q3) and upper limits form the shape of a sine curve with a period of around one month and a slight discrepancy around the beginning of November 2011. The discrepancy is caused by the increase of outliers (not shown in Figure 4.11) while calculating statistical values. We then visualized *Mean* depth-of-coverage values using *Feature through time* visualization. However, that did not provide us with any new insight.

4.5.3 Case study — Conclusion

The simultaneous exploration of multi-metaomics data sets using MOVIS allowed us to uncover temporal patterns and discrepancies of one metaomic data set and efficiently connect them with other metaomic data sets. Furthermore, a swift visualization of sizable time-series multi-modal data sets revealed significant microbial clusters and temporal points of interest. Withal, we are now em-

powered further to analyze temporal points of interest with metaomic-specific tools and uncover metaomic-specific details.

4.6 Discussion

First and foremost, MOVIS is the only time-series omics data exploration tool that is able to generate publication-ready visualizations of the underlying data by following best practices for user interface and data visualization design. Second, since it was written in a procedural way, it allows quick extensions with minor modifications. For example, the adoption of a new omic-type data requires the addition of one high-level function to the data analysis core. Third, the data analysis core presupposes that the data is preprocessed, that is to say, quality controlled and filtered. While FASTA files may be directly used as input, specific data such as raw microarray-based data cannot be used directly as input. The complexity of each omics domain knowledge makes this a challenging problem. Fourth, to support data exploration and visual analytic tasks, we rely on direct data interaction for tabular data. This interactivity lays the foundation for solving visual analytic tasks[21]. Fifth, clear and concise data analysis guidelines benefit multi-omics time-series analyses. Indeed, further omics-wide guidelines, time-series-specific, and data-specific standards are required. Sixth and last, we plan to integrate and make available many more data sets, omic-types data, clustering algorithms, dimensionality reduction methods, and visualizations. Moreover, the implementation of many other embedding algorithms should provide users with ample choice to accommodate the varying nature of the underlying imported data. In this line of reasoning, we are open to requests to include more data in the tool. Currently, two sample data sets are already integrated and available for exploration[141,241]. Since the scope of the data and its size can have a significant impact on computational performance, we also plan to adapt MOVIS to cloud computing.

4.7 Conclusion

MOVIS is created as a modular and easy-to-use solution that includes state-of-the-art libraries and models to import, embed, cluster, and visualize temporal omics data. We expect that the proposed MOVIS will be a valuable tool to complement and enhance traditional data exploration approaches for temporal omics data and offer further insights into the patterns and anomalies of any of the five available omics types and their potential combination. MOVIS currently supports genomics, transcriptomics, metabolomics, proteomics, physico-chemical data, and metaomic aspects of aforementioned omic types.

4.8 Data availability

We provide MOVIS as a web service at `https://movis.mathematik.uni-marburg.de/` and as a Docker container at `https://hub.docker.com/r/aanzel/movis`. The website version is free and open to all users, without any registration requirements. Source code, help, and documentation can be found at `https://github.com/AAnzel/MOVIS`. MOVIS is licensed under the GNU General Public License, Version 3.0, and can be manipulated, improved, and extended freely by any user.

Figure 4.3: Exported visualizations for example data 1. The user has the ability to export resulting visualizations in several lossless formats with just a few mouse clicks. In this figure, we can see save buttons for the scatter-plot visualization of embedded genomics data. Similar exported publication-ready visualizations could be seen in Figure 4.9.

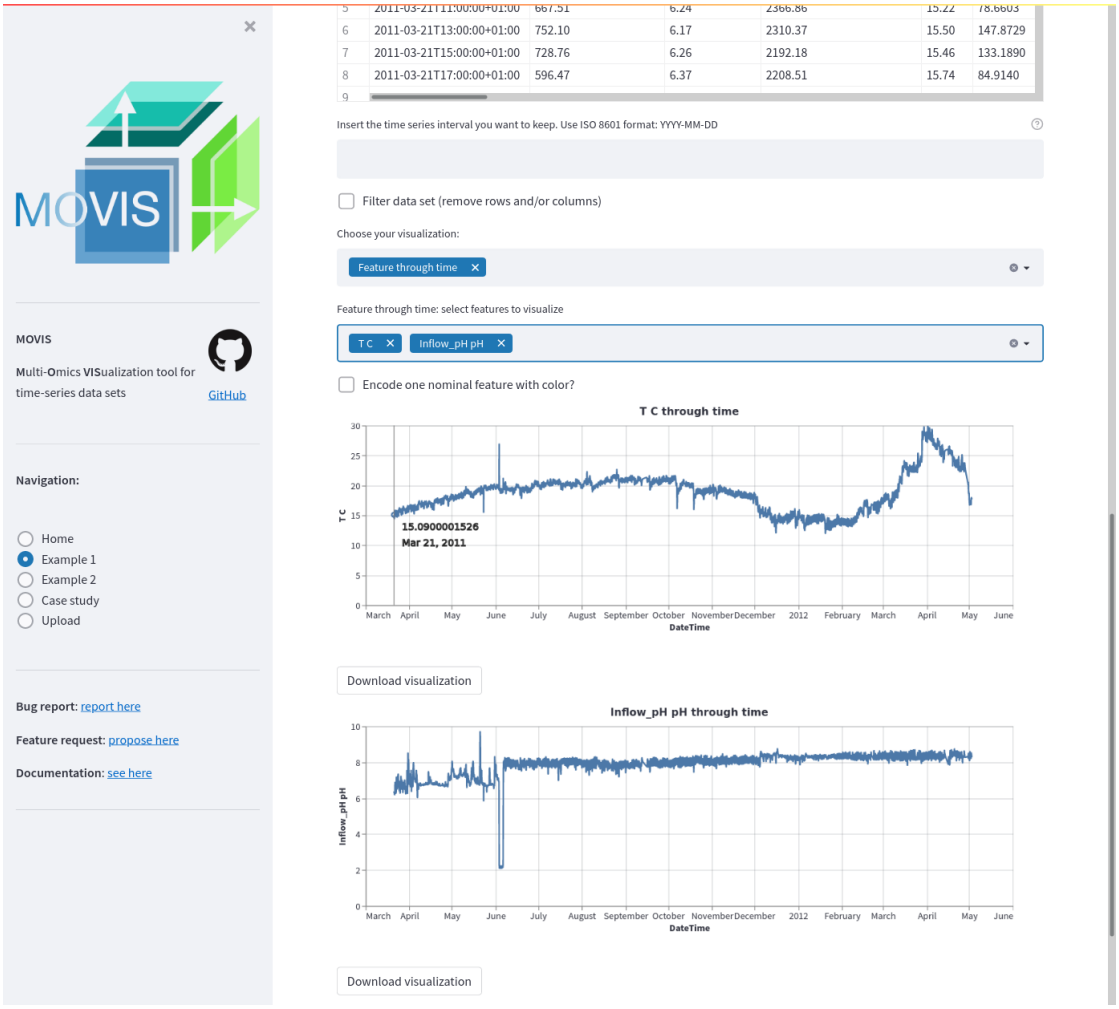


Figure 4.4: Visualization canvas with multiple visualizations. Visualization canvas is the part of the UI that holds visualizations for all data sets in use. It is separated from the data exploration part of MOVIS in order to make the exploration part distraction-free and continuous. This figure also shows the interactive part of every visualization, in this case, a tooltip with additional information of the underlying data point. Further filtering is available to look at the specific time frame.

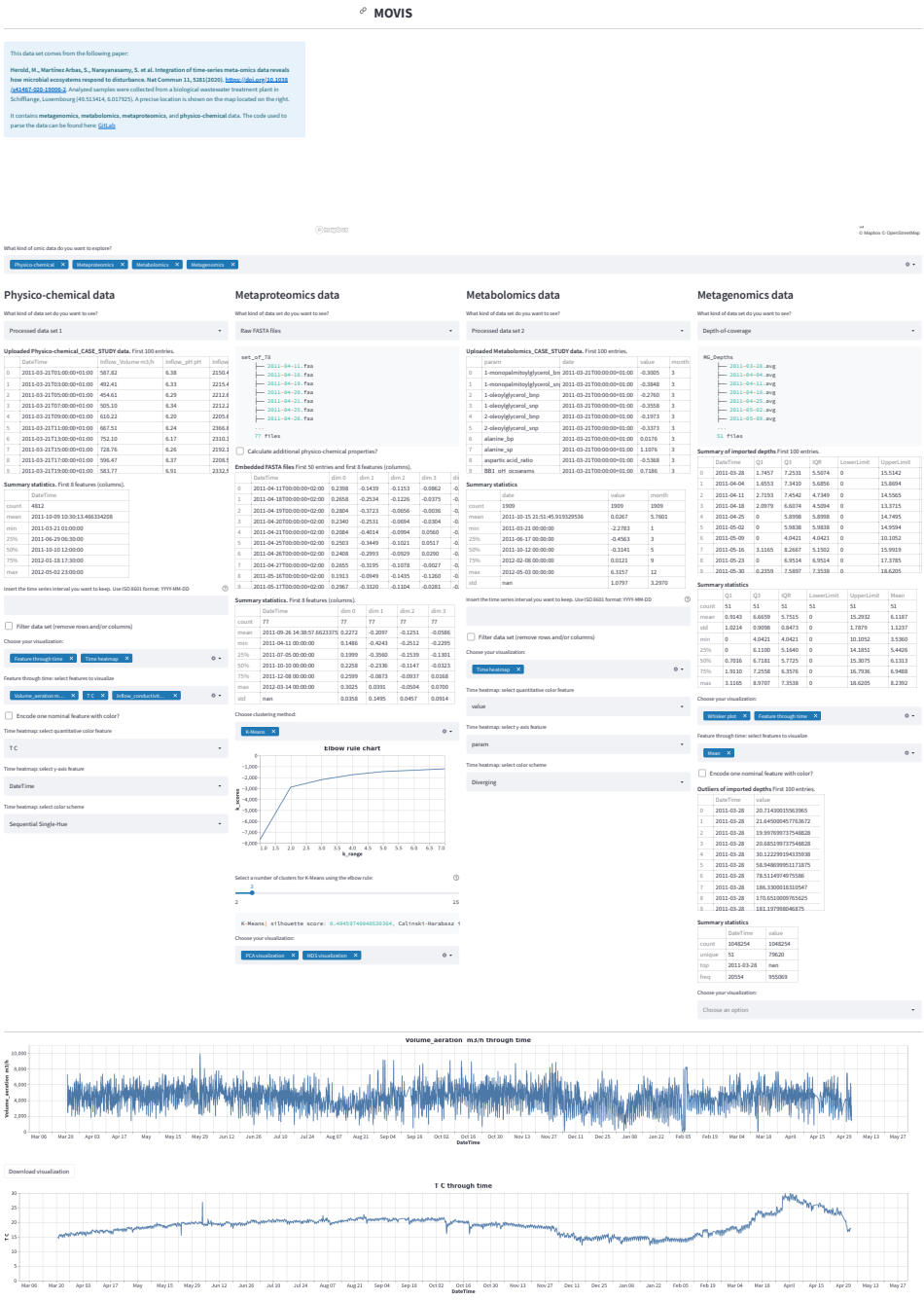


Figure 4.5: Case study overview, part 1. We can see all five parts of the tool's workflow.

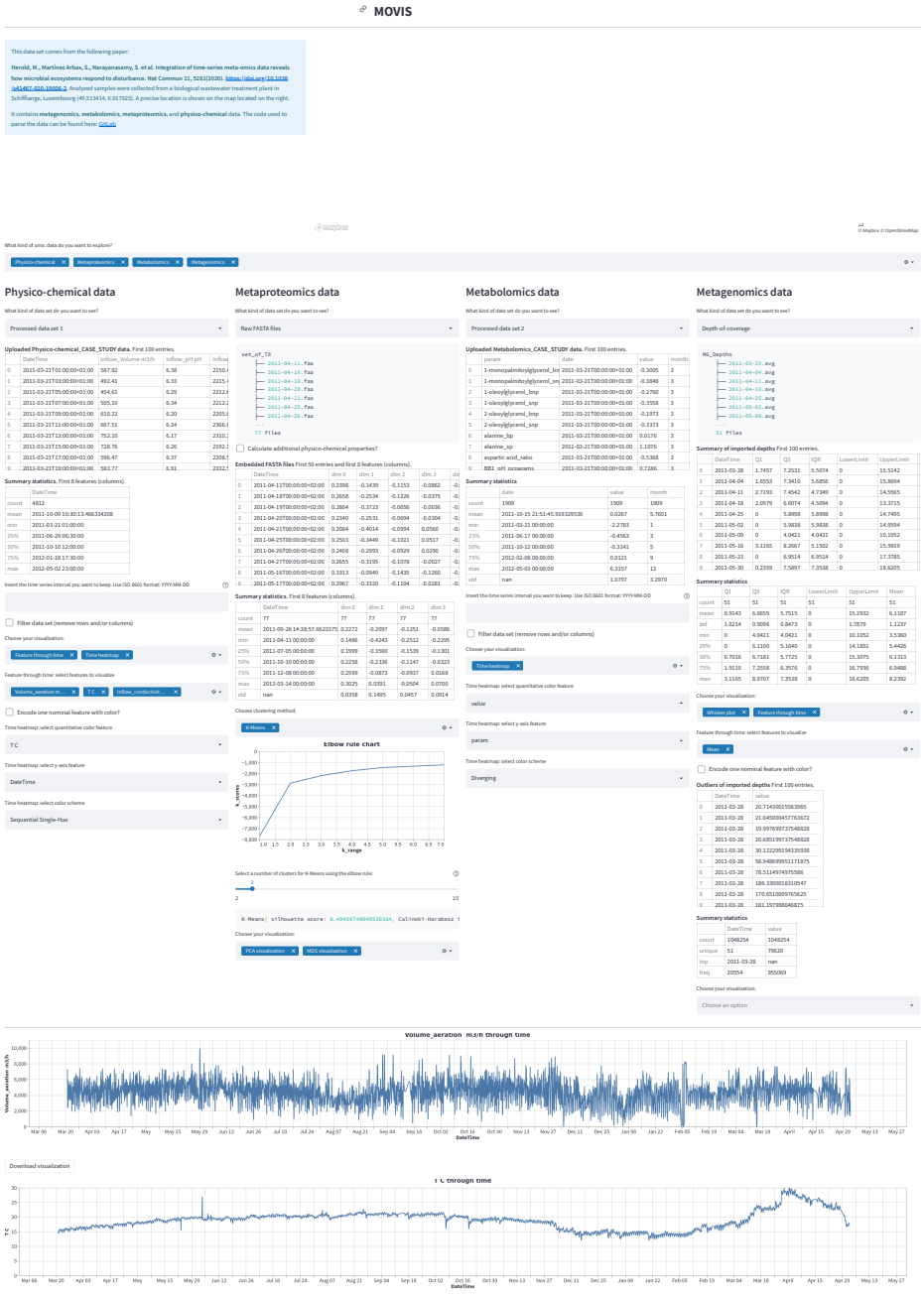
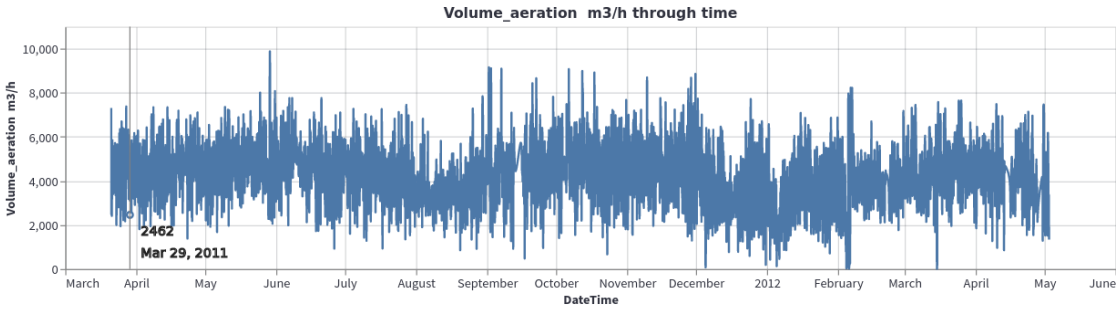
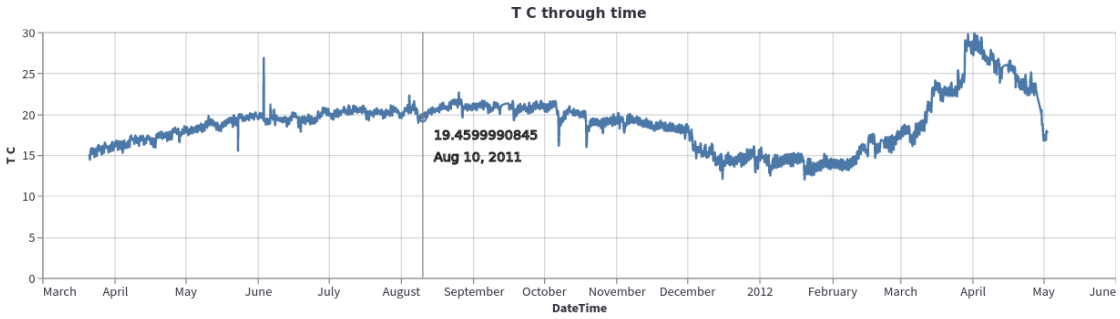


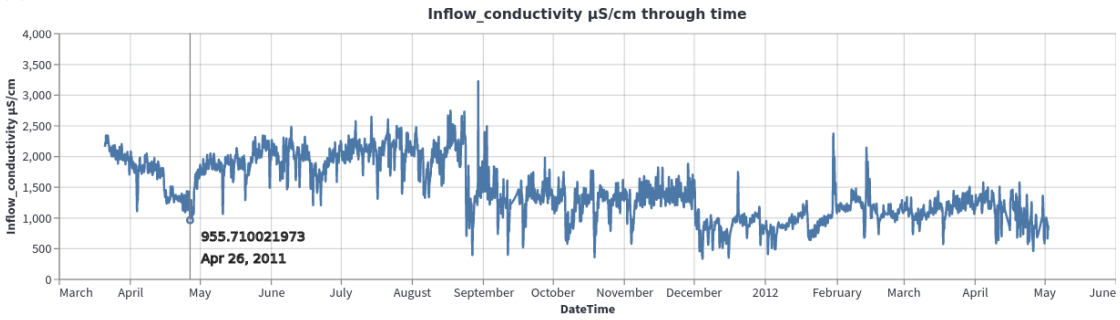
Figure 4.6: Case study overview, part 2. The second part of the overview is consisted only of the visualization canvas, with almost all visualization created for the case study.



(a)

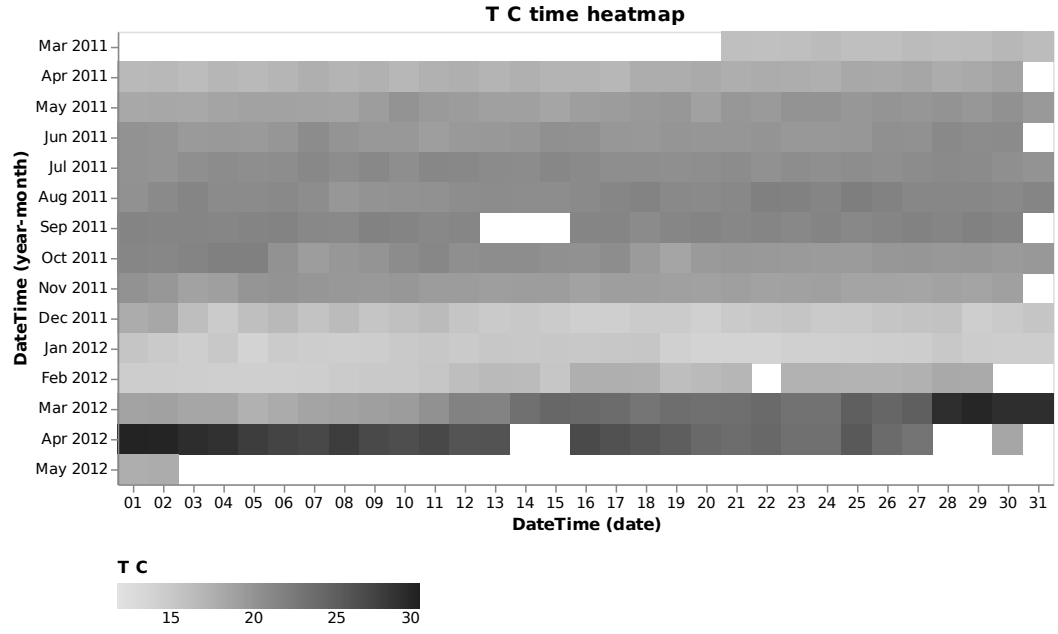


(b)

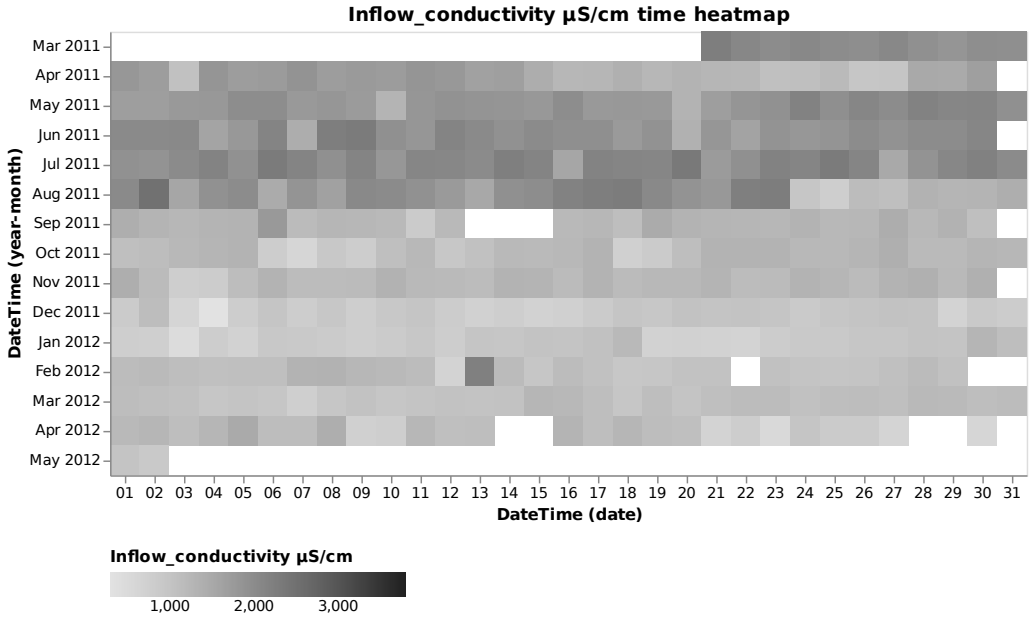


(c)

Figure 4.7: Relevant physico-chemical properties of the wastewater sludge. A seasonal pattern is present in the first figure, while we have more irregular readings in Figure (b) and Figure (c).

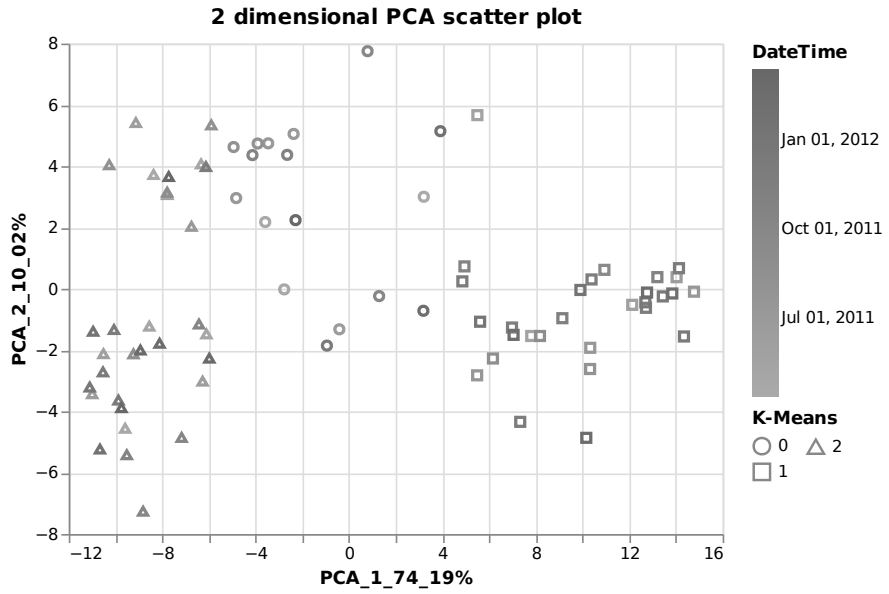


(a)

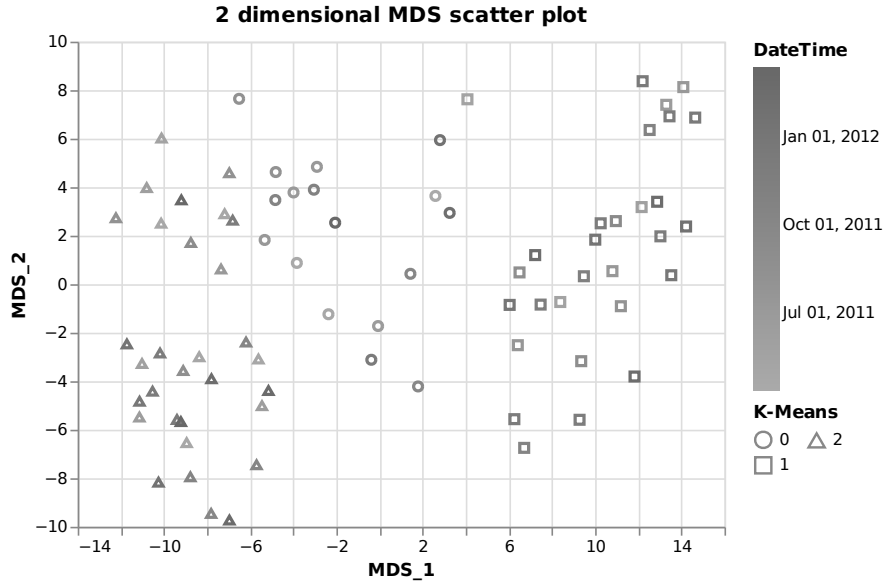


(b)

Figure 4.8: A closer inspection of temperature and inflow conductivity of the wastewater sludge. Figure (a) shows steady values on the upper half of the heatmap and a considerable variation on the lower half. Figure (b) shows a rapid transition from higher to lower conductivity values. The shift happens roughly after one-third of the sampling period.



(a)



(b)

Figure 4.9: Clustered FASTA embeddings of the metaproteomics data set. Figure (a) used the PCA dimensionality reduction technique to visualize embedded data, while Figure (b) used the MDS technique.

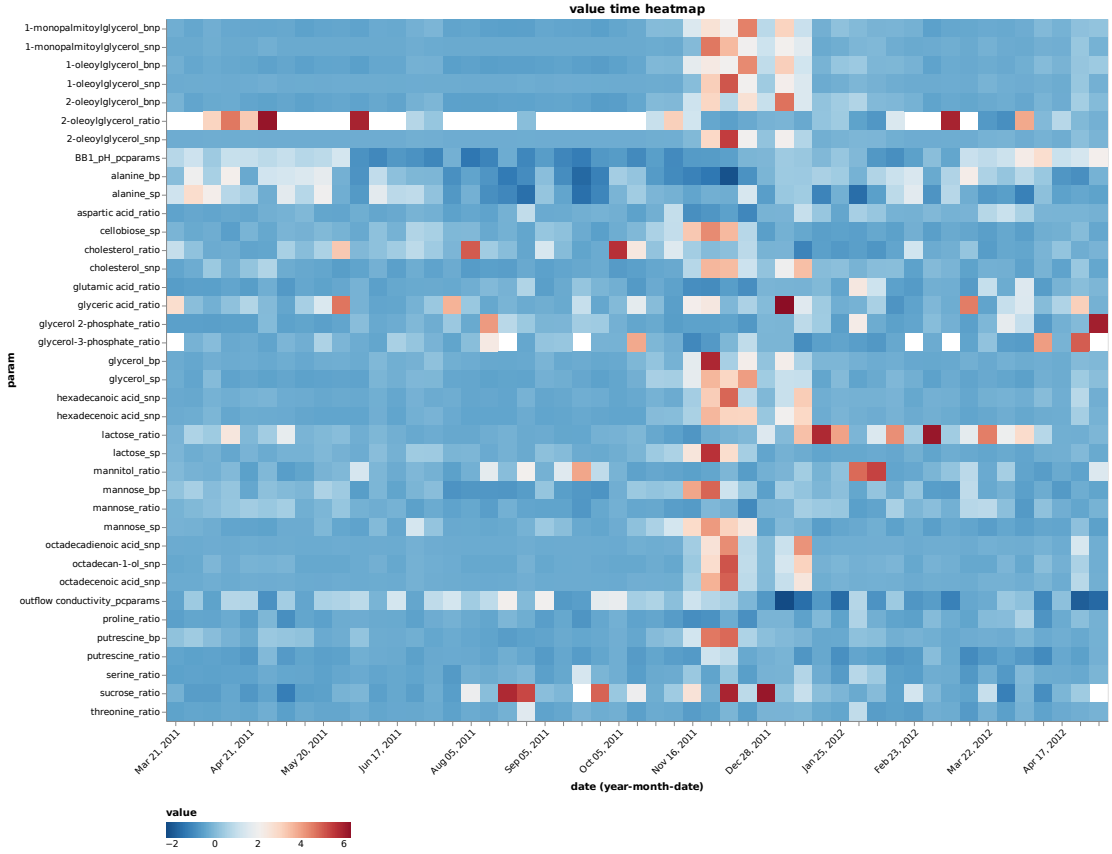


Figure 4.10: Metabolite and physico-chemical values over time. A major shift of multiple parameters can be clearly observed around November 2011. Important abbreviation: **bnp** — intracellular nonpolar metabolites, **bp** — intracellular polar metabolites, **ratio** — metabolite intracellular/extracellular ratio, **snp** — extracellular nonpolar metabolites, **sp** — extracellular polar metabolites.

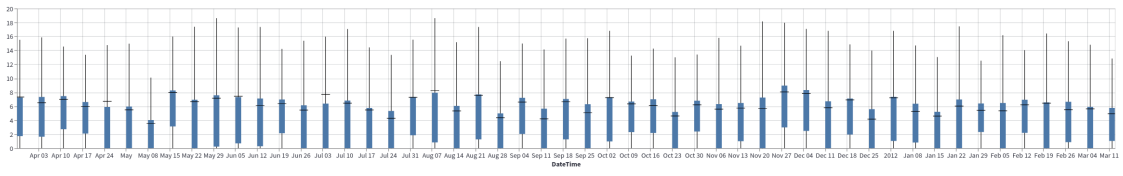


Figure 4.11: Metagenomics depth-of-coverage over time. A sinusoid wave formed by third quartile and upper limit values can be observed.

An algorithm must be seen to be believed, and the best way to learn what an algorithm is all about is to try it.

Donald Knuth

5

Technique-Driven Approach

A technique-driven approach is an approach that emphasizes the technical aspects and provides greater efficiency and effectiveness in data analysis and visualization. With a focused approach to specific techniques and tools, researchers and practitioners gain a deeper understanding of how they can be applied in various contexts. In addition, a technique-driven approach streamlines the data analysis and visualization process. Researchers and practitioners can use well-established techniques and tools, eliminating the need to "reinvent the wheel" and allowing them to concentrate on applying these tools to their own data sets or algorithms. However, even if there is a need for the invention of a new method, the process of classifying the proposed method within the framework of project design is much more streamlined than in problem-driven work. Yet, this requires a significant experience in other fields to find the proper context in which to use the newly-developed method.

The classic example of a technique-driven approach for analysis is the problem of conic sections. Their study began with the ancient Greeks, who explored them from a purely mathematical viewpoint with little practical application. But the research about them eventually turned out to explain planetary orbits in Newtonian physics (2000 years later). Another example, which we will describe in more detail, presents an expansion of a different direction of a problem-driven approach Euler used to prove to the people of Königsberg in 1735 that it is impossible to traverse all seven bridges just once and in one trip.

As mentioned in Chapter 2, where the original problem was explored, by solving the aforementioned problem, Euler invented a new mathematical subfield called topology *. For most of the time since its invention, topology was considered and researched as a part of pure mathematics. Its concepts, like 5-dimensional holes in 11-dimensional spaces, seemed too abstract for real-life applications. Yet, this started to change in the late 19^{th} century with rapid advancements in sensory technologies and computational power [268]. Due to being agnostic to distance, topology is increasingly applied in various fields, such as computer science (*e.g.*, quantum computing [179]), biomedicine (*e.g.*,

*The name of the field was first coined in 1847 by Johann Benedict Listing [190]. The German mathematician also created the term *geoid*, which describes the geometric surface of the planet Earth [189].

protein folding [106], evolutionary biology [291]), robotics [102], *etc.* The two examples of technique-driven approach for the analysis described above are not isolated cases. A plethora of mathematical and physical methods were discovered or constructed without any application in mind. More examples can be found in Peter Rowlett's "Seven Tales" [268], which perfectly illustrates that theoretical work often leads to practical applications, though the process may take centuries.

The study presented in Chapter 6 covers a published study that uses technique-driven approach to analysis. It implements and evaluates an encoding algorithm used to abstract organic molecules into machine-readable language. By traversing their carbon chain, each molecule can be encoded in four different representations — binary, discretized, and their image equivalents. The resulting algorithm was then evaluated using forty-nine encodings of twenty-nine data sets from various biomedical subfields.

This study aimed to enhance existing molecular fingerprinting algorithms through improved flexibility, versatility, and applicability to modern requirements. Major improvements included the capacity to encode any organic molecule, transform molecules into images, and adjust the algorithm to individual use cases. The evaluation of the algorithm also presents its application to specific problems in biomedicine.

UMEROUS examples showcase the use of technique-driven work in visualization research. Moreover, some of them also combine analytical approaches to high dimensionality (see Section 1.1.2). For instance, system DimStiller [158], and algorithms Glimmer [156] and Q-SNE [157] by Ingram *et al.* demonstrate the mixture of dimensionality reduction and visualization methods motivated by the need to improve their interconnection further. In addition, works by Renoust *et al.* [257] and Archambault *et al.* [15] explore the interconnection between graph algorithms and their visualization, while Munzner *et al.* [223] and Liu *et al.* [193] explore similarly an interplay of analysis and visualization in the specific case of graphs known as trees [†].

Chapter 7 discusses a study that employs a technique-driven approach to visualization. The rest

[†] A tree is a graph type consisting of nodes and edges connecting them. The child nodes of each node are the ones that have an incoming edge from the parent node, except for the root node with no incoming edges.

of this chapter is dedicated to introducing this study in the context of its approach and describing the mechanisms used to develop, implement, and evaluate it.

Model comparison is an essential tool in various fields of study, including climatology, biomedicine, and machine learning. By comparing the performance of different models, researchers can choose the most accurate one for a specific task. This can lead to better predictions and decisions, thus advancing our understanding of the world around us. Chapter 7 discusses a *Python* library that is domain-agnostic and can be used to assess models against actual data. The library named *polar-diagrams* utilizes second-order statistics with information theory to create two polar diagrams — the Taylor Diagram (TD) and the Mutual Information Diagram (MID).

The incentive to create a library started with the desire to improve existing implementations of the TD further while also providing the first open-source implementation of the MID. The library employs state-of-the-art algorithms for calculating entropy and mutual information while also providing an interactive resulting diagram that could be exported in various publication-ready formats. The existing visual encoding idioms of these types of charts are further improved by allowing users to visualize two versions of the same model simultaneously (by utilizing marker borders as a property for making a distinction) and one scalar property of each model (by utilizing concentric circles around markers whose size depicts the scalar value). Finally, the library was evaluated using data sets from biomedicine, climatology, and machine learning, thus demonstrating its application to various domain-specific problems.

Besides the mentioned advantages of a technique-driven approach, one major challenge exists. Researchers developing new algorithms, visual encoding, or interactive idioms can easily become too focused on the technical aspects of data analysis and visualization and lose sight of the bigger picture. It is important to remember that data analysis and visualization are ultimately about generating insights and understanding. The technical aspects of these processes are only a means to that end.

6

Organic Molecules in High-Dimensional Spaces

Status

Published as: Georges Hattab, Aleksandar Anžel, Sebastian Spänig, Nils Neumann, and Dominik Heider. A parametric approach for molecular encodings using multilevel atomic neighborhoods applied to peptide classification. *NAR Genomics and Bioinformatics*, 5(1), 01 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqac103. URL https://doi.org/10.1093/nargab/lqac103. lqac103

Copyright Notice

Contribution

I revised the methodology and co-developed the tool. I co-collected and processed the data. I evaluated the tool and visualized the results. I co-wrote, revised, and edited the manuscript.

6.1 Preface

technique-driven approach is essential for identifying the best machine encodings of organic molecules. These encodings facilitate distance and similarity measurements necessary for similarity search or virtual-screening tasks. To achieve this, fingerprinting algorithms can be used to encode the molecules. Therefore, finding new ways to abstract organic

molecules into machine-readable format is crucial for developing new treatments. With a focus on carbon-based multilevel atomic neighborhoods, a walk along the carbon chain of a molecule can be implemented to compute different representations of the neighborhoods in a binary or numerical array that can later be exported into an image. The resulting molecular encodings can then be evaluated using machine learning models against various biomedical data sets. By adopting a domain- and task-agnostic approach, this parametric technique can encode all organic molecules, including unnatural and exotic amino acids and cyclic peptides. The potential for applications and extensions of this approach has been further discussed and corroborated using the peptide classification problem.

6.2 Introduction

omputational approaches to molecular analysis support a range of biologically oriented applications and tasks that are facilitated by the similar property principle[166]. Tasks range from but are not limited to identifying the interactions between drugs and target proteins, to revealing quantitative relationships between structural properties of chemical compounds and biological activities, to screening a handful of membrane proteins for drug delivery[19,46,77,226]. The similarity principle states that similar molecules will also tend to exhibit similar biophysical properties. For example, the virtual screening task is primarily used in drug discovery and allows researchers to find candidate treatments for Alzheimer's disease or HIV[91,246,330]. Virtual screening is carried out by calculating similarity measures of compounds in a database to a reference compound. Using a similarity search, compounds are ranked in descending order and manual screening is performed on the highest ranked compounds[339]. Yet to support the growing number of machine-related tasks, the structure of a molecule must be encoded to a machine-readable format. Indeed, certain structural information may be represented as a numeric feature by means of mapping a large data item to a much shorter bit string. In this context, different types of molecular fingerprints have been proposed: Substructure key-based such as MACCS[46], topological like FP2 OpenBabel[229], circular like MNA[104], pharmacophore, and hybrid. This process leads to a

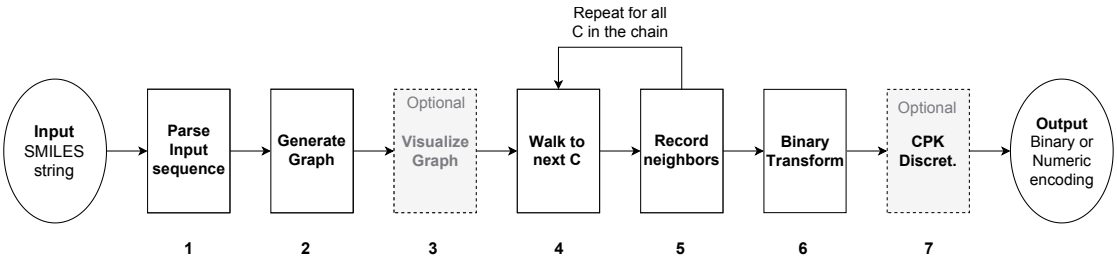


Figure 6.1: Example workflow of the encoding pipeline for a given molecule. C: Carbon. Corey-Pauling-Koltun Discretization: CPK Discret.

molecular fingerprint, which uniquely identifies each molecule through data encoding.

Given such a fingerprint, we can abstract task-specific information at different levels, from the atom, to the neighborhood of an atom, to the amino acid of a protein or even to the base of a DNA molecule. Thanks to this process of abstraction, various biological and chemical aspects may be characterized, similarities and differences may be noted. In the similarity searching example, distances such as Tanimoto or Dice coefficients are calculated between the fingerprint of a certain molecule and its reference during the search [19,233]. Besides previously mentioned measures, researchers have examined many other distance measures and investigated their limitations (*e.g.*, Manhattan, Soergel) [119,258]. Machine learning (ML) has been used in various domain applications (*e.g.*, for predictions, clustering, *etc.*). Different molecular properties can be used as input for training of ML models in order to achieve the best prediction performance. Different molecular properties will have different descriptive power for the source molecule. The molecular properties selected can define the similarity or dissimilarity between molecules. Our proposed approach starts from the question of whether neighborhoods are sufficiently descriptive to characterize organic molecules.

With various bioinformatics tools implementing different types of molecular fingerprints and fingerprinting algorithms [274,289,290], there is an immediate need for adaptable molecular approaches that could accommodate different tasks and specific user needs while respecting different domain standards. That is to say, parametric approaches where users can select and change the values of different parameters, thus adjusting the encoding method according to the, *e.g.*, task, domain, or ML model. In this work, we present a parametric approach to molecular encodings that we apply to the

specific task of peptide classification. The idea is to correctly classify peptides that possess certain features. The concept of depending on the neighborhood hierarchy has not, to our knowledge and despite the existence of several fingerprinting algorithms, been considered.

To implement this concept, we depend on the element carbon (C) to produce various encodings. As the centerpiece of organic life, C is ubiquitous and very good at forming large and stable chains of various organic molecules. Inspired by its central role, we introduce a parametric approach to molecular encodings of carbon-based multilevel atomic neighborhoods as an open source standalone executable and a GitHub source repository; namely cmangoes. It takes as input positional and optional arguments allowing the creation of user-defined molecular encodings. The former include a path to one or more molecular sequences, the type of encoding (binary or discretized), and a padding parameter (centered or offset). The latter include, but are not limited to, a parameter for the upper limit of neighborhood levels to be considered, and whether or not images are required.

This parametric approach paves the way for further efforts to tailor molecular encodings to specific user requirements while taking into account the parameter space of fingerprinting algorithms. Furthermore, since its implementation follows domain-specific standards, the parametric approach can be adopted to address different tasks in a variety of domains. In the following, we introduce the methodology of the proposed parametric approach and showcase its usefulness for the example task of peptide classification via an evaluation on twenty-nine data sets and a comparison to forty-five encodings in the biomedical domain.

6.3 Materials and Methods

The presented work takes into account the ubiquity of the carbon element and its central role in holding together the structure of organic molecules and organizing their neighborhoods. The parametric approach encodes the neighborhoods around the carbon chain of a molecule in multiple levels. Various design considerations are followed to meet established domain standards and create compatible encodings for common similarity measures and distances.

This section describes the parametric approach, design considerations of the underlying algorithm to fit the domain specificity of molecular fingerprinting, the data sets used, and the evaluation of the parametric approach: peptide classification and benchmark.

6.3.1 The parametric approach

The parametric approach handles the input data, generates intermediate data representations as a graph, traverses it to record the relevant neighborhoods according to user-specified parameters, transforms the recorded features to their final output format, and generates the corresponding representations. The walk along the carbon chain iteratively lists the neighboring atoms of each visited atom. The neighborhood of an atom is defined by the neighbors found by direct short paths around it. Figure 6.2 depicts a visual example. The hierarchies are multiple levels of an atom's neighborhood and are defined hierarchically based on their proximity to the carbon chain. By incorporating hierarchies into the encoding, molecules of varying lengths containing different substructures can be appropriately represented. An implementation of this approach is provided as a Python package for easy reproducibility. The core development of the algorithm was performed using Python programming language, version 3.8.5 [230,323]. The chosen language offers high compatibility with existing computational approaches commonly used in bioinformatics and cheminformatics. All core-related dependencies are listed on the official GitHub page. The package accepts FASTA or SMILES file format specifications and follows a seven-step encoding pipeline. Figure 6.1 depicts an example workflow diagram for an input molecule as a SMILES string.

The first step consists of parsing and processing the input data, given in one of the two available formats. To ensure all atoms of a given molecule are present for all following steps of the parametric algorithm, hydrogen atoms (H) are added upon data import.

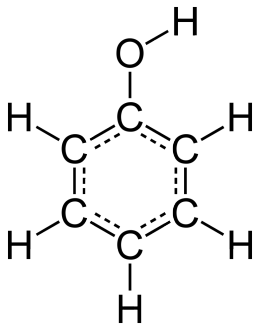
Second, an intermediary molecular graph data structure is employed to efficiently traverse the carbon chain and to record the relevant neighborhoods. To generate the molecular graph, the input data is parsed into an adjacency matrix which is then transformed into a graph. To create a robust

and deterministic encoding, all atoms of a given molecule are represented by nodes in the molecular graph and are numbered with a unique identifier. Each node in the molecular graph stores the type of element they represent using its element symbol from the periodic table. The element labels are required in subsequent steps to generate the feature vectors. To avoid redundancy, the edges of the molecular graph do not store any additional information aside from the nodes they connect, *i.e.*, an unweighted graph.

Third, to aid the identification of the optimal depth, the molecular graph may be visualized. When a data set is used, users select the molecule of their choice in the data set and its intermediary graph is rendered.

Fourth, the walk along the carbon chain corresponds to an iteration over a numbered list of carbon atoms. This list is created by only retaining the nodes that correspond to the carbon element symbol (C). Each carbon node is included exactly once. The filtered nodes are then sorted in ascending order by their unique node identifier. The immediate neighbors include all nodes connected directly by an edge to the respective carbon node. Aside from the immediate neighbors, additional hierarchy levels of a neighborhood can be saved. Figure 6.2 shows an illustrative iteration for the example phenol molecule.

Fifth, neighborhoods along the previously mentioned walk are saved. The additional hierarchy levels are defined as the immediate neighbors of all nodes belonging to the previous level. For instance, the second-level hierarchy includes all nodes with a direct connection to any node from the first-level hierarchy. To avoid redundancy in the encoded information, an additional filter is applied when recording more than one hierarchy. Since neighborhoods are recorded as part of the main iteration, this filter excludes nodes containing carbon atoms. The number of recorded hierarchies can be set using the level parameter. The data structure used for saving the neighborhoods is a dictionary. Only the element symbols belonging to the neighborhood's nodes are saved. The list of element symbols is, by nature of the iteration, automatically sorted according to the unique node identifiers. This ensures that the feature vectors are deterministic across multiple runs of the encoding. To sim-



C_0	C_1	C_2	C_3	C_4	C_5
C	C	C	C	C	C
C	C	C	C	C	C
H	H	H	O	H	H
C	C	C	C	C	C
-	-	-	H	-	-

Table 6.1: The structural formula of the phenol molecule and its recorded neighborhoods using one- and two-level hierarchies. Phenol or C_6H_6O has the SMILES specification: `C1=CC=C(C=C1)O`. Canonical SMILES: `Oc1ccccc1`. C_0 is located at the bottom of the cycle. C_3 is at the top and is connected to the Oxygen O element.

plify subsequent steps of the algorithm and feature-based operations, such as transformation, the dictionary is transformed to a data frame. Table 6.1 reports the resulting hierarchies for the example phenol molecule.

Sixth, feature transformation (binary and discretization) is applied on the hierarchies. Feature transformation enables numeric operations and image generation of the resulting encodings. In the example of the binary encoding, the feature vectors are represented as bit strings, 0 and 1 encode the absence or presence of an atom in the respective neighborhood. The resulting categorical data frame may include missing values depending on the structure of the encoded compound. This can occur when recording more than one hierarchical level, as shown above, when carbon nodes are excluded. To preserve the integrity of the overall data structure and avoid the occurrence of an uneven number of atoms recorded in neighborhoods along the carbon chain, the data frame is automatically filled with missing values in the relevant positions. Since the value 0 represents the absence of information, this procedure does not distort the resulting feature vector.

Seventh, the numerical encodings in the image space follow either a 1-bit coding or the Corey Pauling Koltun color coding (CPK). This optional step exports images with either binary or discretized encodings[135]. Table 6.2 and table 6.3 show the resulting output after feature transformation for the example phenol molecule. Figure 6.3 depicts the image representations of the resulting transformations.

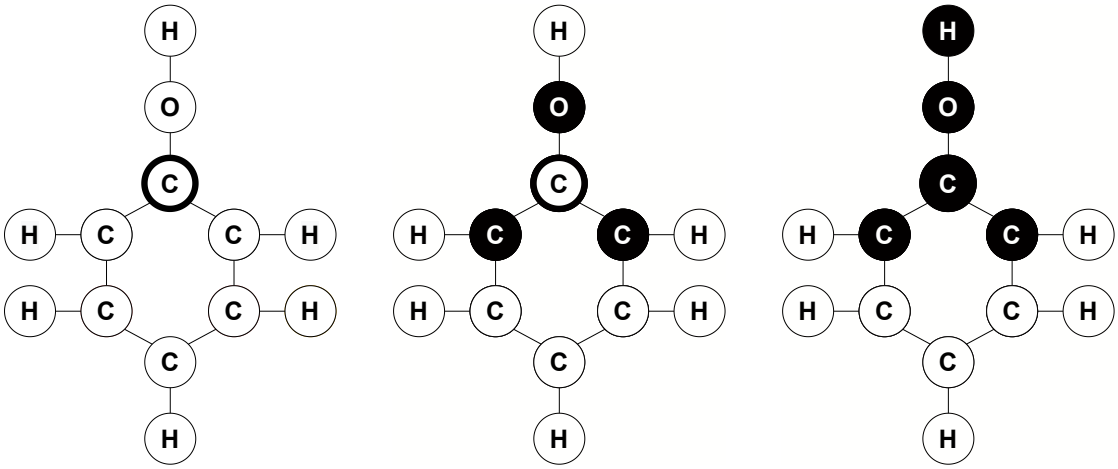


Figure 6.2: Visual demonstration of a computation for two-level hierarchies of the phenol molecule (C_6H_6O). Each figure corresponds to one iteration along the carbon chain. (Left) The algorithm reaches the highlighted carbon atom C at an example iteration. (Center) It records the first-level hierarchy: C,C,O. (Right) Then, the second-level hierarchy: C,H. The resulting hierarchy is C,C,O,C,H. The algorithm iterates onto the next carbon atom.

C_0C	C_0H	C_1C	C_1H	C_2C	C_2H	C_3C	C_3H	C_3O	C_4C	C_4H	C_5C	C_5H
1	0	1	0	1	0	1	0	0	1	0	1	0
1	0	1	0	1	0	1	0	0	1	0	1	0
0	1	0	1	0	1	0	0	1	0	1	0	1
1	0	1	0	1	0	1	0	0	1	0	1	0
0	0	0	0	0	0	0	1	0	0	0	0	0

Table 6.2: Recorded neighborhoods of the phenol molecule using one- and two-level hierarchies after binary transformation. To enable distance-based, similarity searching and machine learning tasks, the categorical encoding is transformed using dummy encoding.

C_0C	C_0H	C_1C	C_1H	C_2C	C_2H	C_3C	C_3H	C_3O	C_4C	C_4H	C_5C	C_5H
3	0	3	0	3	0	3	0	0	3	0	3	0
3	0	3	0	3	0	3	0	0	3	0	3	0
0	2	0	2	0	2	0	0	5	0	2	0	2
3	0	3	0	3	0	3	0	0	3	0	3	0
0	0	0	0	0	0	0	2	0	0	0	0	0

Table 6.3: Recorded neighborhoods for the phenol molecule using one- and two-level hierarchies after CPK-based discretization. The parametric approach transforms the features to integers ranging from 0 to 16 as per the CPK coloring system.

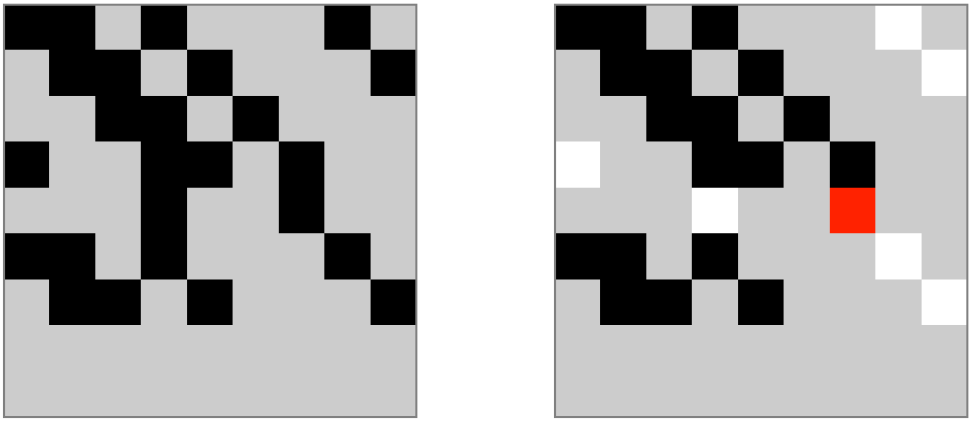


Figure 6.3: Image representations of the encoding for the phenol molecule. (Left) Binary encoding. (Right) CPK-color encoding. The images are created based on the feature vectors found in Table 6.2 and Table 6.3, respectively.

6.3.2 Domain-specific standards

The created feature vectors are domain- and task-agnostic. That is to say, they are compatible with various domain-specific tasks such as database querying or virtual screening[19,46].

In the special case of cyclic molecules, for example aromatic cycles in proteins or cyclic peptides, the unique node identifiers and the sorted filtered list permit the algorithm to bypass cyclical substructures. In turn, this expands the application scope of the proposed approach to include cyclic molecules; such as cyclic peptides often used in therapeutics[297]. Table 6.1 reports the resulting hierarchies. Figure 6.2 shows an example iteration for the phenol molecule.

The image representations complement the mathematical feature vectors to provide an accessible way to understand the resulting encodings and enable additional operations in the image space[175]. Figure 6.3 depicts the image representations of the resulting transformations for the phenol molecule. To avoid dimensional mismatches in the output feature vector and avoid bit collision for different molecule sizes, a padding step is included in the encoding pipeline when applying the encoding to more than one molecule. It includes two padding strategies to either offset (top-left shift) or center

the image representation by introducing new empty pixels around the edges of an image.

6.3.3 Data sets

Twenty-nine data sets comprising peptides and small proteins from various biomedical domains are employed. These include immuno-modulatory and cell-penetrating peptides, but also peptides specifically targeting cancer, fungi, microbes, tuberculosis and viruses. Figure 6.4 lists all the data sets included in this work and reports their class imbalance or imbalance ratio for the evaluation. The properties encoded in the target vectors are represented by ones and zeros, corresponding to the presence or absence of the relevant property, respectively. For example, six data sets are cell-penetrating peptides. Used in research and medicine, they are also known as protein transduction domains and carry a variety of cargoes across the cellular membranes in an intact and functional form [307]. The property encoded in the target vector is whether or not the peptide is cell-penetrating.

6.3.4 Peptide Classification

To evaluate the parametric approach, we adopt the task of peptide classification and rely on the state-of-the-art tool `PEPTIDE REACTOR` [290]. We run a high-throughput comparison of forty-nine encodings on the aforementioned data sets [289]. Based on this work, the Random Forest classifier is used with default parameters as the ML model to address the task of peptide classification. For reproducibility, the complete study details, such as hyperparameter values, data set split sizes, *etc.*, are taken from `PEPTIDE REACTOR`.

To ascertain whether the class-imbalance and the data set size has an effect on the prediction quality, both the class distribution of the respective target vectors and the number of observations contained in each data set vary.

To minimize bias based on the data set choice, the twenty-nine data sets are encoded with four different encodings with the parametric approach. They comprise the first- and second-level hierarchies, with a centered or shifted (offset) padding and are binary or discretized, respectively.

The evaluation is carried out by adding the four encodings to the aforementioned tool. This totals forty-nine encodings. The effective comparison of the classification results relies on the F_β Score metric with $\beta = 1$. It corresponds to the weighted harmonic mean of precision and recall, reaching its optimal value at 1 and its worst value at 0.

The evaluation of the peptide classification task comprises training 1,421 ML models which result from forty-nine encodings applied to twenty-nine data sets. The evaluation was carried out using cloud computing. We relied on the de.NBI Cloud within the German Network for Bioinformatics Infrastructure.

6.3.5 Benchmark

To report the performance results of the proposed parameter approach, it is benchmarked as a fingerprinting algorithm. We consider two parameters for benchmarking the creation of an encoding: the elapsed time in seconds and the amount of encoded data in bytes. Benchmarking is performed for all four encodings on all data sets. For each encoding, six runs are performed and benchmarked.

Benchmarking is conducted using multi-threading on a Linux machine. Kernel: 5.17.5-76051705-generic, CPU: Intel i7-10700 (16) @ 2.90GHz (Turbo 4.90GHz), Thread(s) per core: 2, Core(s) per socket: 8, Memory: 16 GB.

6.4 Results

The parametric approach provided a simplified set of parameters to adapt the encoding step to user-specific needs. It is available as a standalone Linux executable and the source code GitHub repository to create encodings, explore their parameter space, and generate hypotheses and design ML experiments.

By relying on the F_1 Score, we found that the four encodings were consistently providing equivalent results with marginal differences. By conducting a One-Way ANOVA test, we did not find statistically significant differences among the four encodings. At $p < 0.05$ and three degrees of freedom

(df) between-groups and 112 df within-groups, the *F-statistic* value was 0.084 and the *p*-value was 0.969. While the *F-statistic* informed us whether there is an overall difference between the sample mean, the Tukey's range test or Tukey HSD allowed us to determine that there is no significant difference between the various pairs of means. In other words, we found that there is no significant difference in performance if the user chooses a binary or discretized encoding type, and in the padding strategy (center or shifted) for the peptide classification task. Results of the Tukey HSD are reported in Table 6.4.

Pairwise Comparisons		**$HSD_{.05}$ = 0.098** **$HSD_{.01}$ = 0.119**	**Q_{05} = 3.688** **Q_{01} = 4.504**
E1:E2	M1 = 0.62 M2 = 0.62	0.00	Q = 0.08 (p = .99993)
E1:E3	M1 = 0.62 M3 = 0.63	0.01	Q = 0.52 (p = .98262)
E1:E4	M1 = 0.62 M4 = 0.63	0.01	Q = 0.36 (p = .99430)
E2:E3	M2 = 0.62 M3 = 0.63	0.02	Q = 0.61 (p = .97347)
E2:E4	M2 = 0.62 M4 = 0.63	0.01	Q = 0.44 (p = .98948)
E3:E4	M3 = 0.63 M4 = 0.63	0.00	Q = 0.17 (p = .99942)

Table 6.4: Tukey's range test results. E1, E2, E3, E4 are the four encodings `bin_cen`, `bin_shi`, `dis_cen`, `dis_shi`, respectively. M1 to M4 are the means of each encoding group results of the F_1 Score. Q refers to the fact that Tukey's range test is based on a studentized range distribution (q).

The evaluation of the classification task showed that the first- and second-level hierarchies carry enough information to reach acceptable and good classification results. However, the results are fairly sparse across the different data sets and follows the general trend of other existing encodings. A complete overview of the evaluation results using the F_1 score are reported in Figure 6.4. A Jupyter Notebook (`Code/visualize.ipynb`) is made available on GitHub to reproduce all figures and provide interactive visualizations.

The benchmarking of the parametric approach permitted us to report its performance on differ-

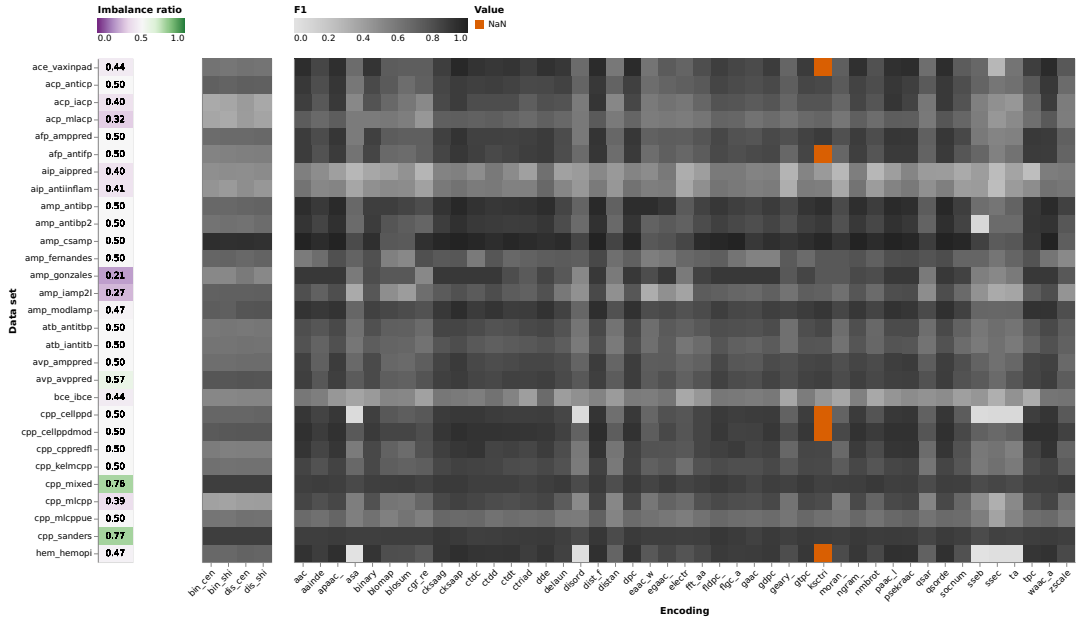


Figure 6.4: Evaluation results of the peptide classification task. The four encodings obtained using the parametric approach are shown on the left versus the 45 sequence- and structure-based encodings. Results are sorted by class imbalance and encoding type. Color coding corresponds to the maximum F_1 Score of the bootstrapped medians for a group. The abscissa is organized by sequence- and structure-based encodings. The ordinate is sorted by class imbalance (cut-off 0.35). Groups are separated by white bars.

ent data sets. Benchmarking results are visualized and complemented with the imbalance ratio of the twenty-nine data sets in Figure 6.5. Overall, the elapsed time (s) shows a linear dependency with the data set size (bytes). Our results indicate that a data size of 1 MB required 2245 median seconds. That is to say, the encoding of 1 byte requires 2.141 median ms.

6.5 Discussion

First, further computational improvements can be made for both the parametric approach and the evaluation pipeline. On one hand, very large data sets can be batched encoded and as such parallel processing of the input molecules is relevant. On the other hand, the large number of training iterations makes computational optimizations especially important. Although we rely on two-dimensional multilevel neighborhoods alone, this work pro-

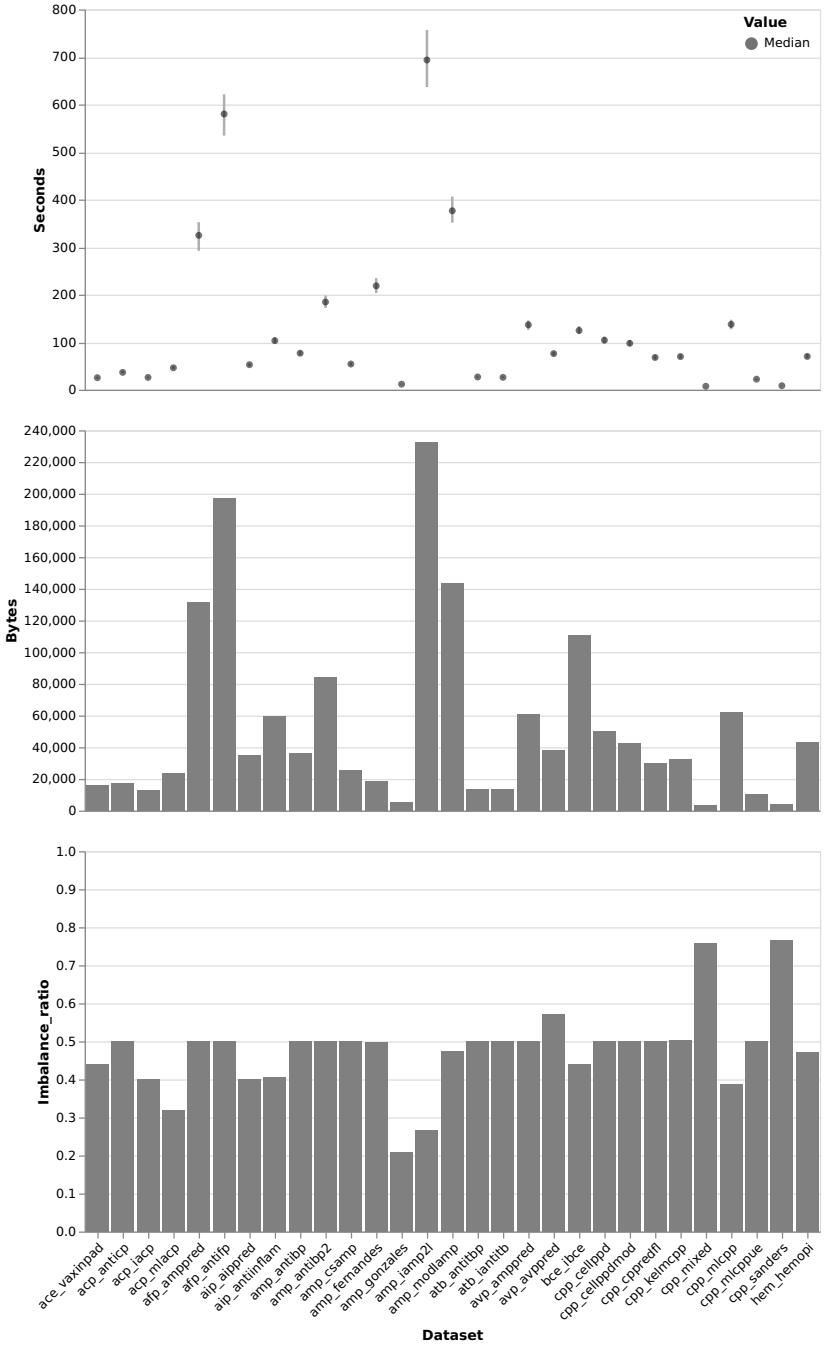


Figure 6.5: Benchmark performance of the four encodings created using the parametric approach. Median values are reported and were calculated for all six runs and encodings. Faceted views of the elapsed time (s), the size of the encoded data (byte), and the imbalance ratio of the data set.

vides a proof of concept and the evaluation of other fingerprinting algorithms for binary classification should be considered, as reported in previous work[290]. Both sequence and structure encodings were included in the evaluation. Indeed, our results can be directly compared to the classification results reported in the PEPTIDE REACToR tool. We hope this effort enables a fair and direct comparison across encodings and data sets.

Second, compared to results reported in the related work, our results were found to be consistent which made the parametric approach a dependable one. Results using the F_1 score showed an acceptable to good separation of the two classes, *i.e.*, robustness. In the example of the six cell-penetrating peptides data sets (*cpp*), it is important to note that in the majority of the original works, both the accuracy and the Matthews Correlation Coefficient (MCC) performance metrics were used and this is in discordance with good practices for binary classification. The AUC usually provides robustness of the resulting classifier and is more discriminative than the MCC, while the accuracy is the measure of the closeness to a specific value and the AUC is the measure across all the possible thresholds[41,127,188]. We chose the F_1 score because it is applicable to any particular point on the ROC curve. While the AUC is the area under the ROC curve, the F_1 score is a measure of precision and recall at a particular threshold value. To maximize this score, both precision and recall must be high. In this ideal case, the model returns many results, all correctly labeled.

Third, we found no significant difference in performance if the user chooses a binary or discretized encoding type as well as in the padding strategy (center or shifted). This may imply that the strongest signal comes from the main positional parameter: the levels to be considered. Although other explanations are possible such as the domain of peptides or the length of a molecule, results portray the robustness of the parametric approach. In this case, the different possible combinations of the feature vector representation (via the optional parameters) did not affect the classification results. In addition, we have discovered that employing only the first or second level leads to subpar performance, albeit this is not covered in this study. In fact, this point has not been addressed because looking at just one level contradicts the logic behind the parametric approach methodology.

Fourth, the image representation of the resulting encodings constitutes an interesting research starting point. It opens up a new space of representation by using the image domain. For example, convolutional neural networks may be used for the same task of classification yet by relying on the images of the resulting encodings. Since such neural networks convolve learned features with input data, and use two-dimensional convolutional layers, their architecture is suited to processing two-dimensional data, such as images. Such methodology would eliminate the need for manual feature extraction required to classify the images.

Fifth, the molecular complexity field is noteworthy. It provides fundamental concepts that underly current fragment-based lead discovery. It considers the general index of molecular complexity, where features that make a molecule more or less complex are taken into account[79,140]. For example, size, symmetry, branching, rings, multiple bonds and heterogeneity in the atoms. Such concepts have been used in various application domains such as chromatography analysis and synthesis pathways. It would be very useful to rely on such features to improve the proposed approach and introduce further parameters such as symmetry, the presence of a cycle, or even the distances among atoms. Such additions may be made at the second step of the parametric approach to enrich the resulting encodings, increase the user-settable parameters, and further vary the resulting performance of an encoding for a specific task or domain.

Sixth, although this parametric approach proved useful for cell-penetrating peptides and achieved acceptable classification results for different data sets, it is important to extend its usage to include larger molecules and more heterogeneous data sets such as membrane proteins[60,133]. For comparability, we successfully evaluated additional data sets, including imbalanced and large data sets that broadened the application scope. Furthermore, it would be valuable to consider correlation results among varying encodings. This could open up the way to build upon the parametric approach and bypass computationally demanding algorithms and move directly to the design of ML experiments.

Seventh, by default the parametric approach produces very sparse encodings. This is especially the case when the encodings are padded or centered. Hence, it is important to develop specialized

methods to address sparsity and evaluate its effects. This relates to the problem of representation and has potential links to data compression. Further considerations are warranted for a more faithful space of representation so to reduce the data and preserve its relevant structure.

Eighth, this work started with the question of whether atomic neighborhoods are descriptive enough to characterize organic molecules. Although this is a naive question, it is related to the basic idea that the neighborhoods created by the carbon atom are not only important but may be sufficient to obtain good classification results. To potentially achieve very good or perfect classification results, the parametric approach can be complemented by the molecular complexity concepts mentioned above. Moreover, the first version of the parametric approach cannot handle other atoms than the carbon atom as the backbone of a molecule. However, heterocyclic compounds can be encoded and the bonds between the different atoms are respected.

Ninth, although the parametric approach is focused on organic compounds or molecules, it is possible to adapt the underlying algorithm to create multilevel atomic neighborhoods of molecules that lack C-H bonds. That is, considering inorganic polymers whose backbone structure does not include carbon atoms, further expanding the application domains and tasks for which the parametric approach could be used.

Tenth and last, evaluating all models using good practices in ML is a standard approach to optimize the prediction performance of the models. Since the parametric approach provides a parameter space, researchers may also move upstream and consider a sensitivity analysis to better fine-tune resulting ML models. Moreover, geometrical deep learning tools, like graph neural networks (GNNs), could be incorporated to further improve the overall ML aspect. This method would exploit the underlying machine representation of molecules using graphs (*i.e.*, adjacency matrices). Work in this direction is already underway and can be seen in [112,338].

6.6 Conclusion

The presented parametric approach is created as an easy-to-use and easy-to-install solution that includes the necessary operations to create custom multilevel encodings of molecular data. Results for the binary peptide classification task were produced by using the PEPTIDE REACToR tool. The F_1 Score reached 0.86 even with a class imbalance of 0.76 and 0.77. The best performance reached 0.93 for the antimicrobial activity prediction in Cysteine-Stabilized peptides (data set: *amp_csamp*)[247]. Moreover, the performance evaluation showed that the first two-level hierarchies carry the most meaningful information for the classification task. Overall, the classification results of the four encodings were consistent with and comparable to the general trend of the state-of-the-art results. Benchmark results indicated that the parametric approach is not computationally intensive and linearly increases with the data set size. Since fingerprint representations decrease computational expenses and enable rapid comparison of different molecules, future work could extend the application of this approach beyond the task of binary classification and peptides. Unlike other fingerprinting algorithms and methods, the intermediate graph data structure makes the parametric approach versatile and permits the usage of organic molecules such as unnatural and exotic amino acids and cyclic peptides. Moreover, we foresee that the proposed work will be a valuable tool to complement and enhance current molecular fingerprinting algorithms and offer further insights into the parameters and the use of hierarchies and their potential combination.

The greatest value of a picture is when it forces us to notice what we never expected to see.

John Wilder Tukey

7

Polar Diagrams for Multi-Dimensional Model Comparison in Complex Systems

Status

Under review as: Aleksandar Anžel, Dominik Heider, and Georges Hattab. Interactive polar diagrams for model comparison. *Computer Methods and Programs in Biomedicine*, 2023. ISSN 1872-7565

Copyright Notice

Contribution

I designed, developed, documented, and evaluated the library. I collected, processed, and curated the data. I wrote, revised, and discussed the original manuscript.

7.1 Preface

valuating the performance of complex models in various domains, such as biology, medicine, climatology, and machine learning, is a challenging task. Using traditional methods for evaluating these models often involves presenting multi-model evaluation scores in a table, which can create difficulties in determining the order of model performance and the similarities between models. The development of juxtaposed Taylor and Mutual Information Diagrams provides a valuable tool for tracking and summarizing the performance of a single model or a collection of different models. These diagrams enable users to determine linear and non-linear relationships between models and are helpful for quickly evaluating model performance. A

technique-driven approach is particularly critical for creating a library that supports both continuous and categorical attributes and enables users to visualize, track, and summarize model performance. Embracing this approach detaches the problem from domain-specific limitations, allowing for enhancements in existing algorithms, encodings, and idioms. As a result, the effectiveness and efficiency of both Taylor and Mutual Information Diagrams are heightened, while also enabling exploration of new domain applications previously unexplored. Leveraging the capabilities of this library facilitates the efficient evaluation of complex models in diverse domains, ultimately facilitating the identification of the optimal model for specific problems.

7.2 Introduction

One of the last steps of any simulation or predictive analytics experiment is to determine the effectiveness of the used models and find the one that best explains the observed phenomenon. The visual comparison of one or two complex models containing multiple variables (dimensions) becomes impractical and often impossible when the number of dimensions exceeds three [29,54]. However, those models are standard in meteorological, medical, biological, and other similar domains. When considering more than two complex models, determining which model is the best becomes unachievable. Although we provide an in-depth examination of model interpretation in Section 7.5, it is imperative to note that any n-dimensional numerical vector is considered a model — hence the definition of a model is not restricted to a specific context within this paper.

To address the task of determining the best model, a quantification of the models' quality is required. The related work relies on the observed data by calculating summary statistics or other types of measures (attributes). To present such attributes and statistics, visualization is needed. By visualizing and representing each model in 2-dimensional (2-D) or 3-dimensional (3-D) plots, reducing the dimensionality of the data is an intrinsic part of the process. Commonly used visualization plot solutions typically rely on scatter plots and heatmaps. Both plot types can be seen in Figure 7.1.

However, these two plot types only allow pairwise comparisons [213,355]. A possible solution is a scatterplot matrix, which is an arrangement of scatter plots organized in a grid or matrix to visualize bivariate relationships among variable combinations. The matrix includes multiple scatter plots, each of which illustrates the relationship between a pair of variables, enabling the examination of several relationships within a single chart. While scatterplot matrix charts are useful for understanding bivariate relationships between multiple variables, they do have limitations. First, they can get cluttered with a large number of variables, making it difficult to distinguish individual plots and trends. Second, outliers can skew the distribution and make it challenging to visualize correlations accurately. Third, it can be difficult to identify cause-and-effect relationships and additional analysis may be required to understand how variables relate to each other [295,333]. As indicated by Figure 7.2, the first drawback becomes evident even with three variables. Alternatively, the parallel coordinates plot is a solution to multivariate analysis where attributes are represented as parallel vertical axes scaled within their data range, as demonstrated in Figure 7.3. However, visual cluttering in this plot type can pose a significant problem for the exploration of relationships between the neighboring axes. Ordering of the axes and visual clutter are limiting factors. This problem has been extensively explored in the past [16,29,152,196,240]. For these plot types, the task of visually comparing large corpora of models becomes intractable. This defines a bottleneck for high-dimensional models' comparison.

Over the last years, many model-comparison visualization solutions have been developed, yet a majority of them are domain-specific and cannot be translated for use in other fields. One example of a domain-specific visualization tool in the field of Machine Learning (ML) by Zhou *et al.* [356] uses a radial-structure approach, which allows for the comparison of ML models with different numbers of features. While this approach is indeed a viable solution for ML models, it is limited by its domain-specificity and the lack of open-source code, preventing its use in other domains. Another example of a domain-specific visualization tool in the field of ML by Talbot *et al.* [303] is *EnsembleMatrix*, an interactive visualization tool that provides a graphical view of confusion matrices to assess ML classifier models. Unfortunately, this visualization solution is heavily domain-specific and cannot be

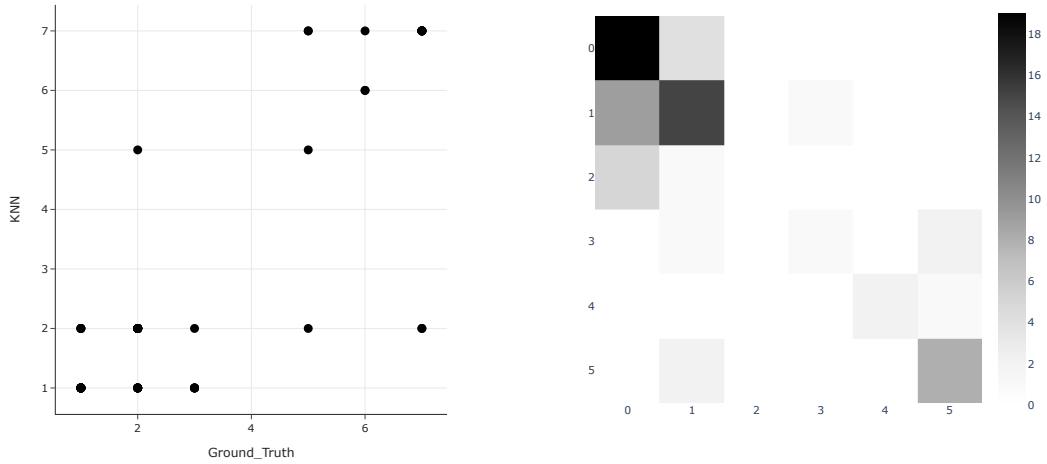


Figure 7.1: Traditional visualization approaches for pairwise comparison. Scatter plot (left) and heatmap chart (right) visualizing the relationship and confusion matrix between *Ground_Truth* and *KNN* model trained and evaluated on the *Glass* [100] data set.

translated for use in other fields. In the field of climatology, besides the Taylor and Mutual Information diagrams, Yatkin *et al.* [347] developed a modified Target Diagram to evaluate the performance of low-cost sensors for air quality monitoring. However, this visualization approach is complicated and requires a level of expertise and extensive training to interpret, making it unsuitable for use in ML or biomedical domains. The limitations of domain-specific visualization methods highlight the need for more generalized techniques that can be applied across different fields.

Previously mentioned limitations were addressed with the publication of the Taylor Diagram [306], which was initially developed for the assessment of climate models. This polar chart efficiently summarizes the model effectiveness according to the observation using three statistical measures: standard deviation, Pearson's correlation coefficient, and centered root mean squared (CRMS) difference or error. However, even though it uses both first- and second-order statistics, the Taylor Diagram cannot capture nonlinear dependencies between models (see Section 7.3.1). Furthermore, if two models are relatively similar but one or both produce outliers, the correlation between them may be low and, in turn, wrongly depict more significant dissimilarity between them.

The Mutual Information Diagram (MID)[71] addresses both issues using information theory. Instead of relying on statistical measures to summarize the models' performance as in the Taylor Diagram, the MID uses entropy, scaled mutual information (SMI), and variation of information (VI). Alternatively, a variant of the diagram incorporates square root of entropy, normalized mutual information (NMI), and the square root of variation of information (RVI). Contrary to the Taylor Diagram, the MID can expose nonlinear dependencies (more in Section 7.3.1), works with both numerical and categorical data, and is far less sensitive to noise (outliers).

Unfortunately, the MID alone does not provide a solution to the original problem because it cannot distinguish between negatively and positively correlated models. Therefore, both diagrams are required in order to get a realistic picture of all model relationships (linear and nonlinear). Moreover, to create the MID, entropy and mutual information have to be calculated for each model. The current implementation requires a domain specialist to tune the parameters for each experiment to calculate both the entropy and the mutual information. The choice of these parameters strongly affects resulting diagram[71], thus presenting a significant obstacle to using the MID without any prior knowledge of information theory.

Furthermore, no publicly available open-source library or tool exists to help users create the MID for uncertainty visualization. The authors of the original paper did not provide any source code or data to reproduce the presented results. Consequently, until now, no publicly available implementation of the MID has been available, even though the need for it was shown[126]. On the other hand, the existing libraries for creating the Taylor Diagram (*MATLAB*[207,261], *Python*[260], *R*[161]) do not provide any interactive aspects of the diagram. The resulting visualizations are static images in `PNG` or `JPEG` format. Moreover, these libraries do not support the raw input — the user has to provide precalculated standard deviations of all models and correlations of all models with the reference models, severely limiting their adoption.

Even though certain Taylor Diagram libraries and tools allow the visualization of multiple model versions, those implementations rely on adding arrows as visual marks that encode the movements of

the models' performances. However, when many models have to be visualized, the diagram quickly becomes overcrowded with visual elements, hence the decreased readability and lower transfer of information from the visualization to the user. Moreover, a set of limiting factors has severely impeded the adoption of polar diagrams until now, including the Taylor Diagram and the Mutual Information Diagram. From static charts to non-scalable graphical formats, to requiring a large set of pre-calculated summary statistics, or even requiring expertise, the adoption and deployment of polar diagrams in the analysis pipeline has suffered greatly.

We show that our library, named *polar-diagrams*, solves completely or partially all of the aforementioned issues. Furthermore, it extends the functionality of both diagrams by allowing users to also visually encode one scalar property of each model or two model versions simultaneously. The resulting diagrams convey more information without overloading the visual space. Moreover, they allow a more granular control by employing multiple interactive techniques such as single- and multi-selection, filter, zoom, and hover. In addition, the back end of *polar-diagrams* employs state-of-the-art methods for calculating mutual information and entropy. As a result, the diagrams become interactive charts that provide accurate information, enable interactivity, and support both discrete and continuous variables.

For the sake of clarity and disambiguation, we adopt the conventional names of the Taylor and the Mutual Information Diagrams for the implemented polar charts, respectively.

7.3 Methods

This section will cover the mathematical aspects of both diagrams, as well as the technological aspect used to design, implement, and present the results. The first part is covered by Section 7.3.1, and the second part by Section 7.3.2. In this section, the terms *variable* and *model* have the same meaning, and we use them interchangeably.

7.3.1 Mathematical Background

As mentioned earlier, the Taylor diagram relies on first- and second-order statistics to summarize model properties, while the MID relies on the information theory. However, both diagrams exploit the same property of the polar diagrams where the position of each point in a diagram is determined by a distance (radial distance, radial coordinate, or radius) from a reference point (pole) and an angle (polar angle, angular coordinate or azimuth) from a reference direction [38]. Similarities between diagrams essentially end here. We will now present the differences in the construction of both diagrams. In addition, we will denote and explain the mathematical deviations from the original works present in our study.

Taylor Diagram

The power of the Taylor Diagram lies in representing each model using three statistical measures: standard deviation, Pearson's correlation coefficient, and centered root mean square error (CRMSE). The "*centered*" aspect of the RMS error definition refers to the subtraction of the respective mean values of both the predicted and observed sets of values before calculating the RMS. This procedure contributes towards rectifying any offset or bias that might have been introduced in the model's predictions, thereby resulting in a more accurate representation of the prediction error [95].

Let us consider a pair of discrete random variables (X, Y), with the cardinality $|\ X\ | = |\ Y\ | = n$, their standard deviations σ_X and σ_Y, and their means μ_X and μ_Y, respectively. We define the mean of a discrete random variable X with the cardinality n as

$$\mu_X = E(X) = \sum_{x \in X} xP(x) = \sum_{i=1}^{n} x_i P(x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7.1}$$

where x represents the values of the random variable X and $P(x)$ represents the corresponding probability. The mean of a discrete random variable X is also known as its expected value and is symbolized as $E(X)$. Furthermore, and for the sake of completeness, we also define the standard deviation of a

discrete random variable X with the cardinality n as

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\sum_{x \in X} (x - \mu_X)^2 P(x)} = \sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 P(x_i)}$$
$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_X)^2} \tag{7.2}$$

where σ_X^2 is also known as the variance of a discrete random variable X. If we define covariance between X and Y as

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y) \tag{7.3}$$

then Pearson's correlation coefficient is

$$R_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}. \tag{7.4}$$

By using the definition of the cosine formula

$$c^2 = a^2 + b^2 - 2ab \cos\theta \tag{7.5}$$

where a, b, and c are the sides of an arbitrary triangle, and the formula for CRMSE

$$CRMSE(X, Y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [(x_i - \mu_X)(y_i - \mu_Y)]^2} \tag{7.6}$$

we get

$$CRMSE(X, Y)^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_X \sigma_Y R_{XY} \tag{7.7}$$

hence $\theta = \arccos(R_{XY})$. The relation between the CRMSE and the total RMSE can be described by

the following expression:

$$\begin{aligned} CRMSE^2 &= RMSE^2 - (\mu_X - \mu_Y)^2 \\ &= \left(\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \right)^2 - (\mu_X - \mu_Y)^2 \end{aligned} \tag{7.8}$$

which also demonstrates how we get Equation 7.6.

The Taylor Diagram can now be easily constructed using the following procedure:

1. calculate standard deviations for all models,
2. pick one model as a reference model (the variable X in all equations),
3. calculate Pearson's correlation coefficient between the reference model and all other models,
4. calculate the angles using Pearson's correlation coefficient,
5. visualize each model using its standard deviation value as the radius and the calculated angle as the polar angle where the reference direction starts from the pole horizontally to the right, and the polar angle increases to positive angles when traversing the diagram in the counter-clockwise direction.

When working with multiple models, it is not uncommon for those models to use different units of measure. That can influence the statistical measures used in the Taylor Diagram. When facing such a situation, CRMSE and standard deviations are normalized ($CRMSE'(X, Y) = CRMSE(X, Y)/\sigma_X$, $\sigma'_Y = \sigma_Y/\sigma_X$), and those "fixed" values are then visualized. As a result, the reference model is now placed on the abscissa with the radius 1.

Mutual Information Diagram

On the other hand, the Mutual Information Diagram exploits information-theoretic properties of measures such as entropy, mutual information, and variation of information for the construction of the polar diagram. Let us again consider a pair of discrete random variables (X, Y), with the cardinal-

ity $|X| = |Y| = n$. We define discrete or Shannon entropy as

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{7.9}$$

and mutual information (MI) between X and Y as

$$\begin{aligned} I(X;Y) &= \sum_{x \in X} \sum_{y \in Y} P_{(X,Y)}(x,y) \log \frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)} \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{7.10}$$

where $H(X, Y)$ is the joint entropy of X and Y defined as

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P_{(X,Y)}(x,y) \log_2 P_{(X,Y)}(x,y) \tag{7.11}$$

The term $P_{(X,Y)}(x,y)$ in Equations 7.10 and 7.11 denotes the joint probability of values $x \in X$ and $y \in Y$ occurring together, and $P_X(x)$ and $P_Y(y)$ are the marginal probability mass functions of X and Y, respectively. Both equations also demonstrate why MI is a robust measure of dependence as it can identify any connections between random variables that deviate from random chance (*i.e.*, it measures general dependence). When two random variables are independent, the sum of their marginal entropies equals their joint entropy. If the joint entropy is less than the sum of the marginal entropies, it reveals some form of dependency. Unlike correlation, MI is non-parametric, and does not require any specific distributions or mathematical forms of dependence to determine the relationship between random variables. This makes it ideal in detecting both linear and nonlinear correlations [183,287].

The last measure, known as the variation of information (VI), unifies entropy and mutual information and enables us to construct the Mutual Information Diagram. This measure, which is also a

metric as shown in [71], is defined as

$$VI(X,Y) = H(X) + H(Y) - 2I(X;Y) \overset{(7.13)}{=} \tag{7.12}$$
$$= I(X;X) + I(Y;Y) - 2I(X;Y)$$

where in Equation 7.12, we used a known property of mutual information where

$$I(X;X) = H(X). \tag{7.13}$$

If we further notice that Equation 7.12 can be written as

$$\sqrt{VI(X,Y)}^2 = \sqrt{H(X)}^2 + \sqrt{H(Y)}^2 \tag{7.14}$$
$$- 2\sqrt{H(X)}\sqrt{H(Y)}\frac{I(X;Y)}{\sqrt{H(X)}\sqrt{H(Y)}}$$

and apply the cosine formula (Equation 7.5) we easily get

$$\theta = \arccos\left(\frac{I(X;Y)}{\sqrt{H(X)}\sqrt{H(Y)}}\right) = \arccos\left(NMI_{XY}\right) \tag{7.15}$$

where NMI_{XY} denotes the normalized mutual information between X and Y[296]. The authors of the paper [71] named the resulting Mutual Information Diagram, which uses the root entropy value for radius and the NMI for calculating the polar angle of a diagram, as *Normalized Mutual Information Diagram* (NMID).

If we square the left side of Equation 7.12, we get the following equation

$$VI^2(X,Y) = (H(X) + H(Y) - 2I(X;Y))^2 = \cdots = \tag{7.16}$$
$$= H^2(X) + H^2(Y) - 2H(X)H(Y) * c_{XY}$$

where

$$c_{XY} = 2I(X;Y)\frac{H(X,Y)}{H(X)H(Y)}. \tag{7.17}$$

Again, by using the cosine formula, we get $\theta = \arccos c_{XY}$. Since $c_{XY} \in [-1,1]$, the authors of[71] proposed using the unbiased version of mutual information for the diagram creation. The new version is called scaled mutual information and is defined as

$$SMI_{XY} = (c_{XY}+1)/2 \tag{7.18}$$

In turn, the resulting diagram is now placed in the range $[0,1] \ni S_{XY}$ and called *Scaled Mutual Information Diagram* (SMID).

Since $NMI_{XY} \in [0,1]$, $SMI_{XY} \in [0,1]$, and $R_{XY} \in [-1,1]$ both positive and negative correlations map to positive mutual information.

Both versions of the Mutual Information Diagram can be constructed similarly to the Taylor Diagram by following the procedure below:

1. calculate

   (SMID) entropies for all models,

   (NMID) square root of entropies for all models,

2. pick one model as a reference model (the variable X in all equations),
3. calculate mutual information between the reference model and all other models,
4. calculate joint entropies between the reference model and all other models using Equation 7.10,
5. calculate

   (SMID) scaled mutual information using Equation 7.18

   (NMID) normalized mutual information using Equation 7.15

   between the reference model and all other models,

6. visualize each model using its

   (SMID) entropy value as radius, and the calculated angle from Equation 7.17

   (NMID) root entropy value as radius, and the calculated angle from Equation 7.15

   as the polar angle where the reference direction starts from the pole horizontally to the right, and the polar angle increases to positive angles when traversing the diagram in the counter-clockwise direction.

As with the Taylor Diagram, measures are often normalized ($I'(X;Y) = I(X;Y)(H(X)/I(X;X))$, $H'(Y) = H(Y)/H(X)$), and those new values are then visualized. As a result, the reference model is now placed on the abscissa with the radius 1, and the property in Equation 7.13 is maintained. All three polar diagrams can be seen in Figure 7.4.

Important Notes and the Deviations from the Original Work

In this section, we will cover some aspects of both diagrams we deem important and describe the deviations from the original MID presented in [71].

We started Sections 7.3.1 and 7.3.1 by considering a pair of discrete random variables (X, Y) and then explained the constructions of both diagrams by relying on that initial condition. However, not all data sets will contain only discrete variables. All statistical measures required for creating the Taylor Diagram and presented in Section 7.3.1 are also applicable to continuous random variables. The situation is far more complex for the MID.

One way to calculate entropy for continuous variables is to estimate the underlying probability density function (PDF) of that variable. The authors of the original MID explored and tested multiple different methods for estimating PDF. Their results show that the choice of a method and its parameters significantly influence the resulting MID, even though the locations of the distributions in the MID are more or less preserved. Yet, each of the continuous data set examples they presented in Section *5. RESULTS* uses different methods for estimating PDF. In essence, Example *5.1 Intercomparison Studies* used histograms, and Example *5.2 Analysis of Climate Ensembles* used kernel density

estimation with the optimal bandwidth for a bivariate normal distribution and Epanechnikov kernels. The lack of consistency in the methodology for the PDF estimation step in the original paper and the lack of transparency in parameter selection shows that each MID may have been tailored to each data set separately. This presents a great obstacle for any user that is not a domain specialist and wants to use the MID.

To address this problem, we designed *polar-diagrams* by considering the results from the original paper and state-of-the-art methods for calculating continuous entropy and MI.

First, we wanted to give as much flexibility to the user as possible by allowing them to input mixed data sets (*i.e.*, data sets with both continuous and discrete models). To accommodate the possibility of mixed data sets as an input we relied on the results of paper [265]. The authors of the study verified the nearest neighbor method as being far more accurate, less computationally, and less memory expensive than binning-based MI estimators. This is why we decided to use the non-parametric method, known as Kraskov's method [181], to calculate only the MI between variables. Our decision is further corroborated by paper [71], that demonstrates the higher accuracy of this method when estimating MI but larger error when estimating entropy.

Second, our library uses different methods for entropy calculations depending on the data type of the models and the parsed optional arguments. If the model in question is discrete, Equation 7.9 is used to calculate entropy. If the model is continuous, the entropy is known as *differential* or *continuous* entropy and measures the average information content of a random variable with a continuous probability distribution. Our library selects different methods to calculate entropy based on the given sample size of the (unknown) distribution. If the data set has less than 10 samples, the *Van Es* estimator [320] is used. In case the sample size is between 11 and 1000, the *Ebrahimi* estimator [92] is used. For larger sample sizes, the *Vasicek* estimator [325] is used with the heuristic value for the *window length* parameter proposed in [76]. The selection behavior and implementation details are described and presented in [4]. Even though our library selects the differential entropy estimation method automatically depending on the sample size, the users are able to override this functionality and manually

select one of the previously mentioned methods.

Third and last, our proposed methodology gives more accurate results. However, since we are not treating continuous variables as discrete (and *vice versa*) and we are using task-specific methods, one new problem arises. Unlike discrete entropy, differential entropy can be negative. Since MID is only able to show models with positive entropies, this means that models with negative differential entropy will not be present on the resulting MID. We do not consider this a flaw but a limitation of the MID. In that case, the user is encouraged to use the Taylor Diagram to evaluate the results. This motivates and supports the coupling of the polar diagrams presented in this work.

7.3.2 Technical Background

We decided to develop *polar-diagrams* using *Python* programming language[322] due to its flexibility and cross-domain popularity. We also relied on multiple well-established libraries for data manipulation, analysis, and visualization. This section will cover all essential libraries on which our library depends and explain the functionalities we use to create the polar diagrams.

The first step in visualizing model results using *polar-diagrams* is to prepare the data set. The user should use *Pandas*[235,337] library to import the raw data into wide-formatted *Pandas* `DataFrame`. The resulting `DataFrame` must have a 1-level index, model names as column names, and model results in rows. Hence, the `DataFrame` has dimensions *nxm* where *n* is the number of rows, and *m* is the number of models (columns). This is the only format of the input data our library accepts. We decided to use *Pandas* because of its ability to parse a plethora of raw data formats (*e.g.,* `XML`, `JSON`, `CSV`, `SQL`, *etc.*), and its wide-spread use across many domains.

Our library also extends the functionalities of both Taylor and MID by enabling users to visualize a scalar property for each model and visualize two different versions of each model on the same diagram simultaneously. However, the user is able to use only one of the extended functionalities at the same time. In case the user wants to visualize one additional scalar property of each model, instead of parsing one `DataFrame` with model results, users should parse a *Python* `list`. The first el-

ement of that list should be a `DataFrame` that contains model results, as described in the previous paragraph. The second element should also be a `DataFrame` that has dimensions $1xm$, where the single row contains the scalar value a user wants to visualize. Column names must be the same for both arguments. All scalar values are internally scaled to $[0, 1]$ domain in order to prevent the "explosions" of scalar markers on a visual canvas. This means that scalar markers can only be double the size of model markers, thus preventing visual clutter. If the user wants to visualize two different versions of models, a *Python* `list` should be parsed as an argument. The list elements are two nxm-dimensional `DataFrames` representing different versions of m models. As with the previous case, column names must be the same for both arguments (`DataFrames`).

To calculate all statistical measures necessary for the creation of the Taylor Diagram, we used *Pandas*, *NumPy*[131], and *Scikit-learn*[39,239] libraries. To calculate the discrete entropy, we implemented an in-house algorithm according to the original Shannon's definition presented in[277]. Differential entropy is calculated as described in Section 7.3.1 using the *SciPy*[328] library. The MI is calculated using Kraskov's method by adopting the implementation provided in the *Scikit-learn* library. Papers[181,265] show the best MI estimation occurs when the number of neighbors for the method is 3. We also use this as a default value, but the user may change it as necessary.

We used *Plotly*[155] library to design and create all diagrams. It also affords all interactive functionalities of both diagrams and the ability to export them in static image formats.

7.4 Results

Our library presents the first open-source implementation of the interactive Taylor Diagram and the first public implementation of the MID. In addition, it extends the functional aspects of both diagrams by enabling users to visualize one scalar property of each model and two different versions of models on the same diagram. The users can take advantage of the first functionality to visualize any scalar value that is important to the experiment. We used it to encode the training and prediction time of a selection of ML models presented in Sec-

tion 7.4.2. The results are presented in Figure 7.9. The second functionality can be exploited for visualizing models in two time points or with the changed (hyper-)parameters, thus allowing the user to examine the shift in model performance. We showcased the latter in Figure 7.8.

The resulting diagrams can be exported in the following formats: `PNG`, `JPEG`, `WebP`, `PDF`, and `SVG`. Before exporting the diagram in a static image format, the user is able to interact with it and explore it. We now follow the *nested model of visualization* [221] to dissect and present all functionalities of our library's resulting charts.

First, incorporating interactive elements into our diagrams presents one of the major advantages over traditional Taylor and Mutual Information diagrams, which are static images. Polar coordinate charts often have multiple axes radiating from a central point, which can make it challenging to accurately assess certain properties without proper interactivity. For example, it can be difficult to compare the magnitude of different data points on the chart or to determine the exact coordinates of a particular point. Similarly, interpreting the distances between the axes may be challenging without the ability to zoom in on specific areas of the chart, *i.e.*, to adjust the scaling of the radial axis. Additionally, given the circular nature of polar coordinate charts, it may be difficult to identify trends or patterns in the data without the ability to interactively adjust various display options such as color-coding or labeling. All of these points were tackled either singularly or simultaneously in works by Burch *et al.* [40], Yee *et al.* [348], Qiang *et al.* [249], and Vehlow *et al.* [326]. Overall, interactivity is essential for accurately interpreting and exploring the properties of polar coordinate charts in a way that is intuitive and meaningful to the user.

By relying on previous studies that researched interactivity in polar coordinates, we incorporated multiple interactive idioms to allow users to explore the data and change the charts before they are exported in a static image format. Hovering the mouse over any model in the diagram reveals a tooltip with additional information about the underlying model. The border of each tooltip is colored the same as the model it refers to. This interactive element can be seen in Figure 7.7. Users are also allowed to click on the models' graphical representation in the legend and exclude them from

the results. If users double-click on the model in the legend, all models except for the selected one are excluded from the diagram. Besides *Single selection*, *Zoom* is the next and default interactive tool a user can employ to navigate the polar diagrams. This allows users to select specific radial intervals or areas to be visible on the diagram. It is important to note that the *Zoom* tool does not actually zoom into the visualization canvas, rather it rescales the radial axis of the diagrams. The upper-right part of the visualization canvas contains two more tools that allow more granular control on which models to highlight — *Box Select* and *Lasso Select*. The first tool allows the creation of rectangular regions outside which the models will be de-emphasized by decreasing the saturation. The latter provides the same functionality by defining the region using any polygon. They both are elements of the multi-selection interactive aspect. The resulting diagram with some models highlighted could then be easily exported in any of the previously mentioned static image formats, thus better conveying the story of the underlying experiment. All interactive tools are shown in the upper-right corner in Figure 7.7.

Second, we used three encoding channels to encode model data. Circles are used as graphical markers that represent models. They represent elements of the shape channel, where each marker has the same size. Hence, when using *polar-diagrams* to visualize model results considering only one version of the models, this channel does not contain any significant information about the models. However, if users invoke the functionality of visualizing two versions of all models at the same time, this channel is used to create a differentiation between model versions. Circles that encode the second-version models have a solid border, while the circles that encode the first-version models are borderless. We also used the shape channel to encode the second extended functionality *polar-diagrams* supports — visualizing a scalar property of each model. When the user wants to visualize the scalar property, the values are encoded using the concentric circle around the model marker (circle) with the same color. The size channel for the concentric circle is used to encode the normalized scalar value. The difference between the models is encoded using the color channel. The reference model is always encoded using the black color, while all other models are encoded using either

Tableau 10 or *Tableau 20* [22] categorical color schemes depending on the number of models. Each color has 60% opacity, thus allowing an easier model distinction when visual markers overlap.

Our library also supports inspecting and exporting intermediary results for diagram creation. Those results are returned as a *Pandas* `DataFrame` object and can be further exported in any tabular format supported by the *Pandas* library.

One of the example arguments presented in [71] for using the MID instead of the Taylor Diagram is Anscombe's data set [11]. This data set is a set of four data sets, and each of them has the same summary statistics (*i.e.,* mean, standard deviation, and correlation). The authors showed that certain components of Anscombe's data set fully overlap on the Taylor Diagram while being dispersed on the MID. We also tested this on a "newer" version of Anscombe's data sets called *The Datasaurus Dozen* data set [203]. Even though this collection of thirteen data sets contains totally different data sets when visualized, they all have the same summary statistics (X/Y mean, X/Y standard deviation, and Pearson's correlation). Indeed, the MID shows better results for this example as well. Despite the fact that the Taylor Diagram produces results where models are hard to differentiate, by using the interactive *Zoom* tool, the user is able to closely inspect if some models are better. We present our results in Figure 7.5.

Even though the phenomenon of overlapping model markers occurs less often in the MID, it still can occur under specific circumstances. To solve this problem, we implemented a *Python* `RuntimeWarning`, which notifies the users if any of the diagrams contains overlapping models. Furthermore, the warning reveals the exact models that are overlapping on the diagram, thus providing the user an insight into the data and motivating them to use the interactive functionalities offered by the polar diagrams.

In the following sections, we present our results using the data from three different domains — climate research, machine learning, and biology/medicine.

7.4.1 Example 1 — Climate Model Evaluation

One of the most important projects of the World Climate Research Programme (WCRP) is the Coupled Model Intercomparison Project (CMIP). The project's objective is to gain insights into past, present, and future climate changes, thus supporting policy-makers and communities worldwide. The understanding of climate phenomena include, among other things, the assessment of various climate models and the quantification of their performance for future projects.

We specifically picked CMIP Phase 3 (CMIP3) data set[209] in an effort to reproduce the results from Section *5.1 Intercomparison Studies* of the original MID paper[71]. However, during this process, we discovered the following pitfalls that prevented us from fully replicating the results:

- the lack of guidelines that specify how to acquire the data from the CMIP3 data repository,
- the lack of details on what data properties (*i.e.,* ensemble runs) were used from CMIP3 data set,
- missing information about the temperature averaging due to the spatial nature of the CMIP3 data set,
- missing information about the source of the reference or observation (OBS) data (model)
- the example-specific probability density estimation method.

Nevertheless, we solved each of the problems mentioned above by following our intuition and commonly used approaches when working with climate data. We acquired CMIP3 data by creating the following link query `https://esgf-data.dkrz.de/esg-search/wget?download_structure=model&project=CMIP3&experiment=historical&ensemble=run1&variable=ts` and downloading the official *wget* script which downloads all model data. As in the original study, the script downloads the data set, which consists of 21 models. The query shows that we selected the data from the CMIP3 project of the *historical* experiment and for the *surface air temperature (ts)* variable. Since the original work is missing the *ensamble* information, we decided to use only *run1* values. Due to the lack of positional information on the data in the original study, we calculated the average of all

temperatures across the globe per year and used those values for each model. Figure 7.6 shows the original and reproduced results.

As we can see in Figure 7.6, both Taylor and MID look very different than those in the original study. The difference is caused by the lack of a step-by-step procedure to replicate the original results and by using a different probability density estimation algorithm. Although the diagrams are not comparable in this way, the overall goal of providing an open-access library is to support the community at large and enable not only static information visualizations but also the creation of interactive data visualizations and the sharing of source code to reproduce the underlying work. Our results reproduce the fundamental principles and analytical steps used in the related work. Although not entirely irreproducible, this also shines a positive light on open tools that facilitate source code adoption, reuse, sharing, and reproducibility. More particularly, positively affecting the advancement of thematic analytics in the field of climate change.

7.4.2 Example 2 — Machine Learning Model Evaluation

The non-parametric and data-type agnostic nature of our library allows us to work with continuous variables without the discretization step and the selection of example-specific PDF estimation method, as opposed to the procedure presented in [71]. To showcase the power of *polar-diagrams*, we selected various traditional ML data sets that contain all discrete, all continuous, or some discrete and some continuous features. Furthermore, we chose eleven classification and nine regression models. The main task of the experiment was to assess model performance using the diagrams our library provides, find the best model, and check if our assessment is in line with commonly used performance metrics. We present the experiment in more detail below.

Data sets

We used the following data sets for our ML experiment: *Iris* [105], *Breast Cancer* [341], *Glass* [100], *E. Coli* [148], *Mushroom* [331], *California Housing* [176], and *Ames Housing* [63]. The first five data sets

contain a discrete target feature (classification task), while the latter two contain a continuous target feature (regression task). The portion of the target feature used as the test data represents our reference model. We conducted the same preprocessing procedure for all data sets. First, we removed columns that contain identification (ID) numbers. Second, we used the label-encoding method to encode categorical columns of each data set. This step allows the use of the Taylor Diagram for model assessment besides the MID. Third, we removed rows or columns that contain `Null` values. Fourth, and only in the case of the *Mushroom* data set, we sampled the data set using stratification and considered only 40% of all samples. Due to its memory requirements, we had to reduce the size of this specific data set. Fifth, we split the data into the training and test parts with proportions $0.67 : 0.33$. For the classification tasks, this procedure was completed in a stratified fashion. Sixth, we scaled both training and test data using *Scikit-Learn*'s `StandardScaler` that was trained on the training data only. We then proceeded with the ML model training.

Machine Learning Models

The ML example includes all commonly used ML classification and regression models implemented in the *Scikit-Learn* [39,239] library. We used the following models for both the classification and regression tasks: *k-Nearest Neighbors* [72], *Linear Support Vector (SV) Machine* [49,75,101], *Kernelized SV Machine* [49,75], *Decision Tree* [121], *Random Forest* [35,36,115], *Multi-layer Perceptron* [118,137,142], *Ada Boost* [109,132,358], *Gradient Boost* [110,111], and *Stochastic Gradient Descent (SGD)* [34,276,313]. Besides these models, we also used *Gaussian Naive Bayes (NB)* [47,352] for the classification tasks and *Gaussian Process Regressor* [252] for the regression tasks. All models were using the default hyper-parameters, as defined in the *Scikit-Learn* library. For the *Kernelized SV Machine*, we used the *Radial Basis Function (RBF)* kernel with default hyper-parameters.

To train the models, we used stratified 5-fold cross-validation on the training data while using the following scoring methods to evaluate the performances of the classification models: *accuracy*, *weighted precision*, *weighted recall*, and *weighted F1-score*. To evaluate regression models during train-

ing, we used 5-fold cross-validation on the training data with the following scoring methods: R^2 *score*[59], *negative mean absolute error*, *negative mean squared error*, *negative mean squared log error*, and *negative root of mean squared error*. The negative values are used due to the nature of the library, where estimators with higher scores are considered better.

The final evaluation of all models was done with the test data set as defined in Section 7.4.2 and using the scoring methods mentioned earlier. Besides visualizing the model performance using our library, we also visualized different scores acquired during the evaluation. For data sets used in the classification task, we visualized *weighted F1-score*, *φ coefficient or Matthews Correlation Coefficient (MCC)*[58,206], and *Cohen's Kappa coefficient*[65]. On the other hand, for data sets used in the regression task, we visualized the R^2 *score*, *mean squared error (MSE)*, and *mean absolute error (MAE)*. It is important to note that when inspecting figures in the paper, higher values are better for all but two metrics: MSE and MAE. For these two metrics, the opposite is true.

Model results for *Breast Cancer*, *E. Coli*, and *Ames Housing*,*Glass*, *Iris*, *Mushroom*, and *California Housing* data sets can be seen in Figures 7.7, 7.8, 7.9, 7.10, 7.11, 7.12, and 7.13, respectively. Figure 7.8 showcases the first extended functionality of *polar-diagrams* that enables users to visualize multiple versions of the same models. Figure 7.9 presents the second extended functionality that enables users to visualize one scalar property of each model.

7.4.3 Example 3 — Biomedical Similarity Assertion

With the rise of electronic medical records and population-level patient profiles, we are getting closer to the widespread use of precision medicine. In order to achieve this goal, it is often required to find similarities between patients, cluster them, and determine the similarity of each new patient to these defined clusters. Besides being comparable to standard medical diagnosis and hence being familiar to physicians, this step also ensures patient privacy and speeds up the decision process[234,237]. On the other hand, comparative studies present an important part of biological research as well. Comparative biology encompasses a plethora of biological sciences (*e.g.*, Ecology, Genomics, Paleontology).

It enables users to identify similarities and more specifically the distance of one organism (or other taxa) in relation to another and derive the phylogeny[171,334]. In this section, and for the sake of consistency, we will use the term *model* when considering organisms (or other taxa) and patients.

More than often, the end goal of asserting similarities and finding clusters is the representation of the results in a 2-D space. Therefore, the traditional approach consists of selecting one of the clustering algorithms (*e.g.*, *K-Means*[17], *OPTICS*[10], *BIRCH*[353]), choosing the distance metric to be used in the algorithm (*e.g.*, Euclidean distance, Manhattan distance), and using some dimensionality reduction technique (*e.g.*, *Principal Component Analysis (PCA)*[309], *Multidimensional Scaling (MDS)*[74], *T-distributed Stochastic Neighbor Embedding (t-SNE)*[318]) to project the data to 2 dimensions and visualize it in a Cartesian plot with colors (or shapes) encoding clusters.

Due to the nature of the Taylor Diagram and the MID, we can skip all these steps and use CRMSE and VI, respectively, to determine similarities between single models, models and clusters, and clusters and clusters (inter-cluster similarity). Multiple studies have shown that VI shows multiple desirable theoretical properties (such as its metric property and its alignment with the lattice of partitions) and, as such, can be used to compare clusters and, by extension, its individual elements[210–212].

To showcase the ability to assert similarities between biomedical models using *polar-diagrams*, we used the *Fertility*[116] (all discrete features) and *Hepatitis*[144] (all continuous features) data sets.

The *Hepatitis* or *HCV* data set consists of patients that are described by demographic properties and laboratory-collected blood values. All patients fall into one of the following categories: *blood donor*, *suspected blood donor*, *hepatitis C patient*, *fibrosis patient*, and *cirrhosis patient*. For the purposes of our study, we included only hepatitis C patients that do not contain `Null` values and without demographic properties. As a result, we were left with twenty patients, each containing ten blood parameters. The results for the *Hepatitis* and the *Fertility* data sets can be seen in Figures 7.14 and 7.15.

Besides visualizing patients using our library, we also visualized them using the traditional approach. First, we used the *Caliński and Harabasz* score[42] to find the best number of clusters for

the *K-Means* clustering algorithm. Second, we clustered patients using *K-Means*. Third, we used *t-SNE* [318] to reduce the dimensionality of our data to two dimensions. Fourth and last, we visualized data using a 2-D scatter plot, with clusters color-coded and patients represented by shapes. The results can be found in the upper part of Figures 7.14 and 7.15.

The overview of all results of our study can be seen in Table 7.1.

Table 7.1: Overview of the results. The *Agreement* column shows whether the diagrams from our library are in agreement with the traditional evaluation approaches. The *Reproducibility* column shows whether the results are reproducible.

Data set	Agreement	Reproducibility	Result
CMIP3	No *	No *	Figure 7.6
Iris	Yes	Yes	Figure 7.11
Breast Cancer	Yes	Yes	Figure 7.7
Glass	Yes	Yes	Figure 7.10
E. Coli	Yes	Yes	Figure 7.8
Mushroom	Yes	Yes	Figure 7.12
California Housing	No	Yes	Figure 7.13
Ames Housing	Yes	Yes	Figure 7.9
Hepatitis	Yes	Yes	Figure 7.14
Fertility	No	Yes	Figure 7.15

[a] Due to the lack of data in the original study [71].

7.5 Discussion

Thanks to our library, we solve multiple hurdles of MID that were mentioned in the original work. However, certain limiting factors remain as "weak" problems, and we take the liberty to discuss them along with other possible improvements.

First, *polar-diagrams* only supports models represented by n-dimensional numerical vectors, which may be perceived as a "hard" constraint. However, representing the data as numerical vectors is a common practice.

Second, the number of models users want to visualize can be perceived as a "soft" constraint. This is caused by the limitation of the human perceptual and cognitive system in its ability to both perceive and retain a large number of categorical colors [130,134,198,221,283]. This is especially true for colorblind-safe schemes/palettes which can be used to make designs more accessible to users with visual impairments. Indeed, our library does not prevent users from parsing more than twenty models since the colors are repeated after the twentieth model. Although a serious problem in color usage, this is why this constraint is considered "soft". Yet, it does not prevent users from exploring the data interactively since they can exclude models that are not of interest by using the interactive legend. When users are faced with more than twenty models, ambiguity is created with repeating colors. Unfortunately, creating a color palette with more than twenty distinguishable colors is unattainable. As an example, *Colorgorical*, a tool by Gramazio *et al.* [122], which creates discriminable color palettes using three color-discriminability scores and a color-preference score, returns a partial palette and an error when more than twenty-one colors are generated due to the exhaustion of the color space. Indeed, more classes require more colors, which are increasingly difficult to distinguish. Depending on the task at hand and the audience, the number of colors varies. For the example of color coding of symbols, Colin Ware suggests using no more than ten colors if reliable identification is required, especially if the symbols are to be used against a variety of backgrounds [332]. Additionally, we explored the possibility of using color harmonies to expand our color palette [302]. However, we have found that this method is insufficient when the number of colors in the palette is more than four. The seven major color schemes are monochromatic, analogous, complementary, split complementary, triadic, square, and rectangle (or tetradic); resulting in a maximum number of four colors. In the first iterations of the library, we tried encoding models using shapes as well in order to increase the number of distinct model encodings. However, that approach yielded diagrams that were cluttered and hard to read. Our results are further corroborated by works [300,321]. We currently give users the freedom and responsibility to decide which models are visualized and control for repeated colors.

Third, another limitation is with models that have a continuous data type. In this case, as we de-

scribed in Section 7.3.1, our library calculates continuous (differential) entropy and MI for the creation of MID. Since these parameters can be negative, the resulting MID might be empty or without some models, as can be seen in Figures 7.13 and 7.6. However, as mentioned earlier, we do not consider this a true limitation since it presents the nature of continuous entropy.

Fourth, the MID can be further improved by incorporating a normalized variation of information (NVI) as presented in paper [327]. However, this implementation was out of the scope of our paper since it requires further research into its application in the form of a diagram.

Fifth, it is important to note that the differential entropy implemented in *SciKit-Learn* library and used for the creation of MID is not the actual continuous analogue of discrete (Shannon) entropy [163]. All methods mentioned in Section 7.3.2 present the limiting case of the actual continuous version of discrete entropy called the limiting density of discrete points (LDDP). To the best of our knowledge, the implementation of this measure in *Python* does not currently exist. However, we plan to include this measure as another option for calculating the differential entropy in one of our future versions.

Sixth, we also acknowledge another model comparison chart type — the Target Diagram [167]. This Cartesian plot type extends the functionalities of the Taylor Diagram by including the sign of the bias to the summary information (which is unbiased) and summarizes how they each contribute to the total RMSE. The need for a summary diagram that also encodes and visualizes the statistical bias was also confirmed in [51]. However, our library solves this problem with the previously introduced functionality to visualize one scalar property of each model. Implementing this plot type in *polar-diagrams* is thus unnecessary and out of scope due to it not being of a polar type.

Seventh and last, it is important to mention that our diagrams do not contain isolines indicating CRMSE, VI, and RVI. Instead, we chose to show these values in a tooltip. One of the reasons behind this decision is the lack of support for such functionality in all high-level visualization libraries we reviewed. The other reason is to make diagrams as visually decluttered and readable as possible. The same rationale applies to our decision not to use the traditional approaches for multi-version

model visualizations. These approaches use arrows to indicate the flow of information. However, when there are ten or more models, the resulting diagrams have twenty or more model markers with ten arrow lines. Depending on the model shift, those lines can have multiple intersections with each other, thus rendering the resulting visualization incomprehensible. This is the reason we decided to encode the second versions with the same markers as the first versions, but with the added solid border for the purpose of differentiating them.

7.6 Conclusion

Our library provides the first public implementation of the MID and the first implementation of the interactive Taylor Diagram. It was developed by following all "good" programming conventions (*e.g., PEP 8, PEP 20, PEP 257, PEP 287*[322]) and with state-of-the-art open-source data manipulation, scientific computing, mathematics, machine learning, and high-level visualization libraries. The resulting diagrams can be exported in publication-ready static-image formats.

Furthermore, our library extends the expressiveness of both diagrams by providing two additional functionalities — the ability to visualize multiple versions of all models and the ability to visualize one scalar property of each model.

By providing the interactive aspects to both diagrams, the users are encouraged to explore results in a way not available until now. We expect *polar-diagrams* to be a valuable tool in climate, biomedical, machine learning, and other domains that produce complex models and offer further insights into interdependencies between such models.

7.7 Code Availability

Source code, help, and documentation can be found at `https://github.com/AAnzel/Polar-Diagrams-for-Model-Comparison`. The repository also contains the source code necessary to reproduce all examples in this work.

Our library is also available on the official third-party software repository for *Python* packages called *PyPI* under the following link https://pypi.org/project/polar-diagrams/. It is licensed under the *GNU General Public License, Version 3.0* and can be manipulated, improved, and extended freely by any user.

7.8 Availability of Data and Materials

he data used in this study can be either found or downloaded using the scripts present at https://github.com/AAnzel/Polar-Diagrams-for-Model-Comparison/tree/master/Data. All data sets are also cited in Section 7.4 and can be downloaded from the originating studies.

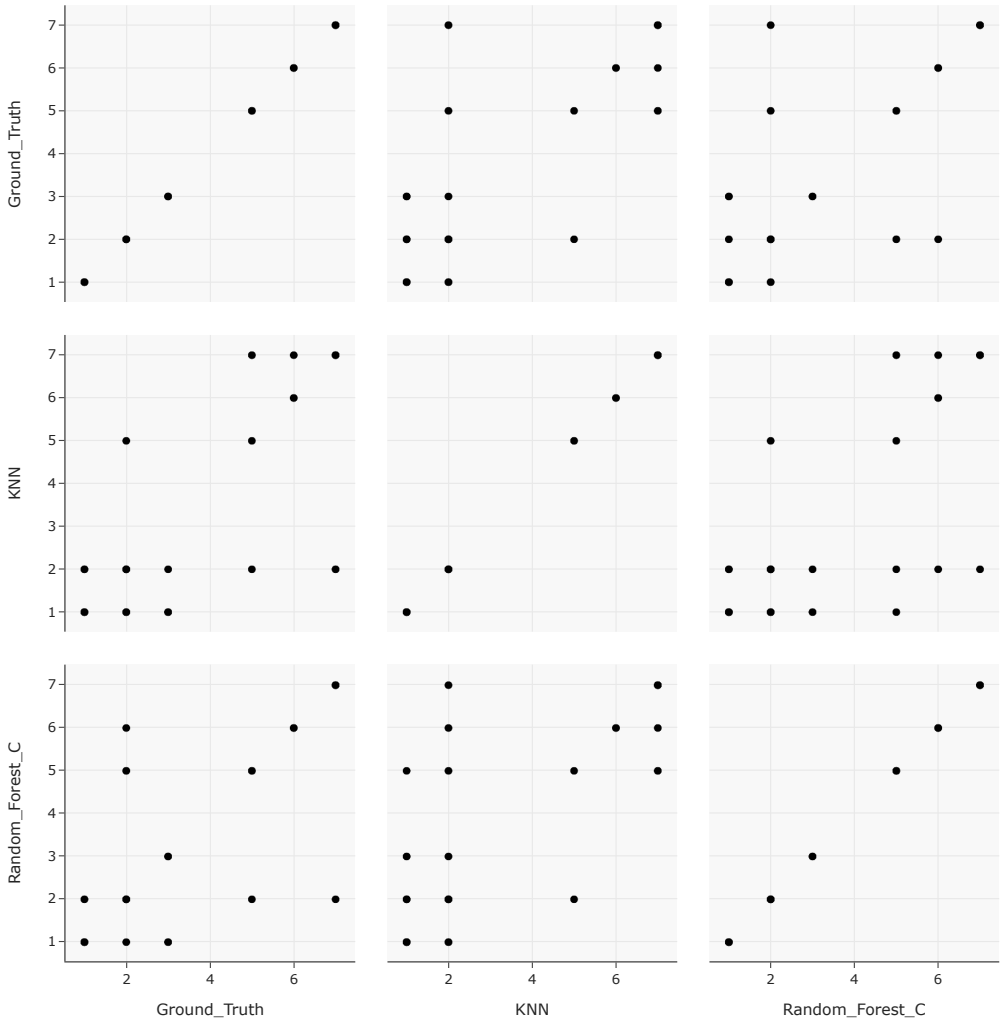


Figure 7.2: Scatterplot Matrix. A scatterplot matrix with three variables taken into account — *Ground_Truth*, *KNN*, and *Random_Forest_C*. The *Glass* [100] data set is used to present this plot type.

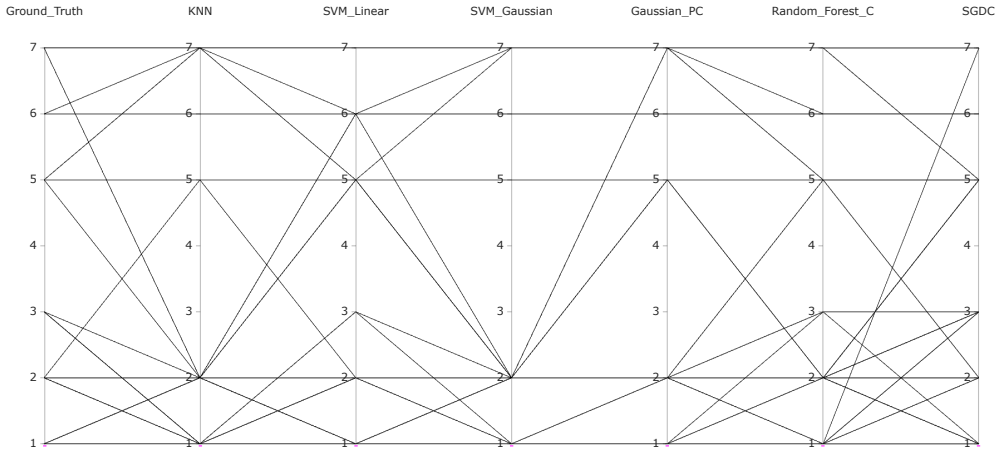


Figure 7.3: Parallel Coordinates Plot. A parallel coordinates chart presents a better alternative to the scatter plot matrix and allows a more compact visualization of more than three variables. After performing an evaluation of machine learning models trained on the *Glass* [100] data set, we visualized the performance of six models — *KNN*, *SVM_Linear*, *SVM_Gaussian*, *Gaussian_PC*, *Random_Forest_C*, and *SGDC*, along with the *Ground_Truth*.

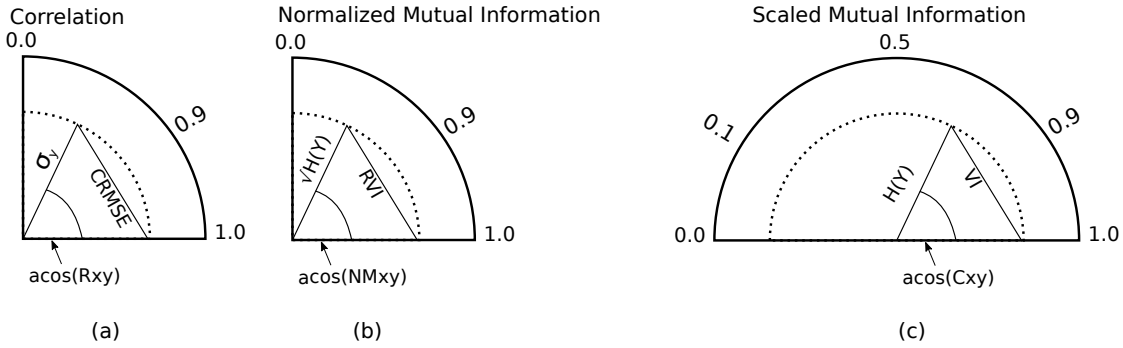


Figure 7.4: Taylor Diagram (a), NMID (b), and SMID (c) are presented. As we can see, while the Taylor Diagram and the SMID span the first and the second quadrants, the NMID spans only the first quadrant. The reader should note that the Taylor Diagram presented in this figure is a trimmed version of the full diagram since the negative correlations are not presented. This procedure is usually applied to the SMID as well.

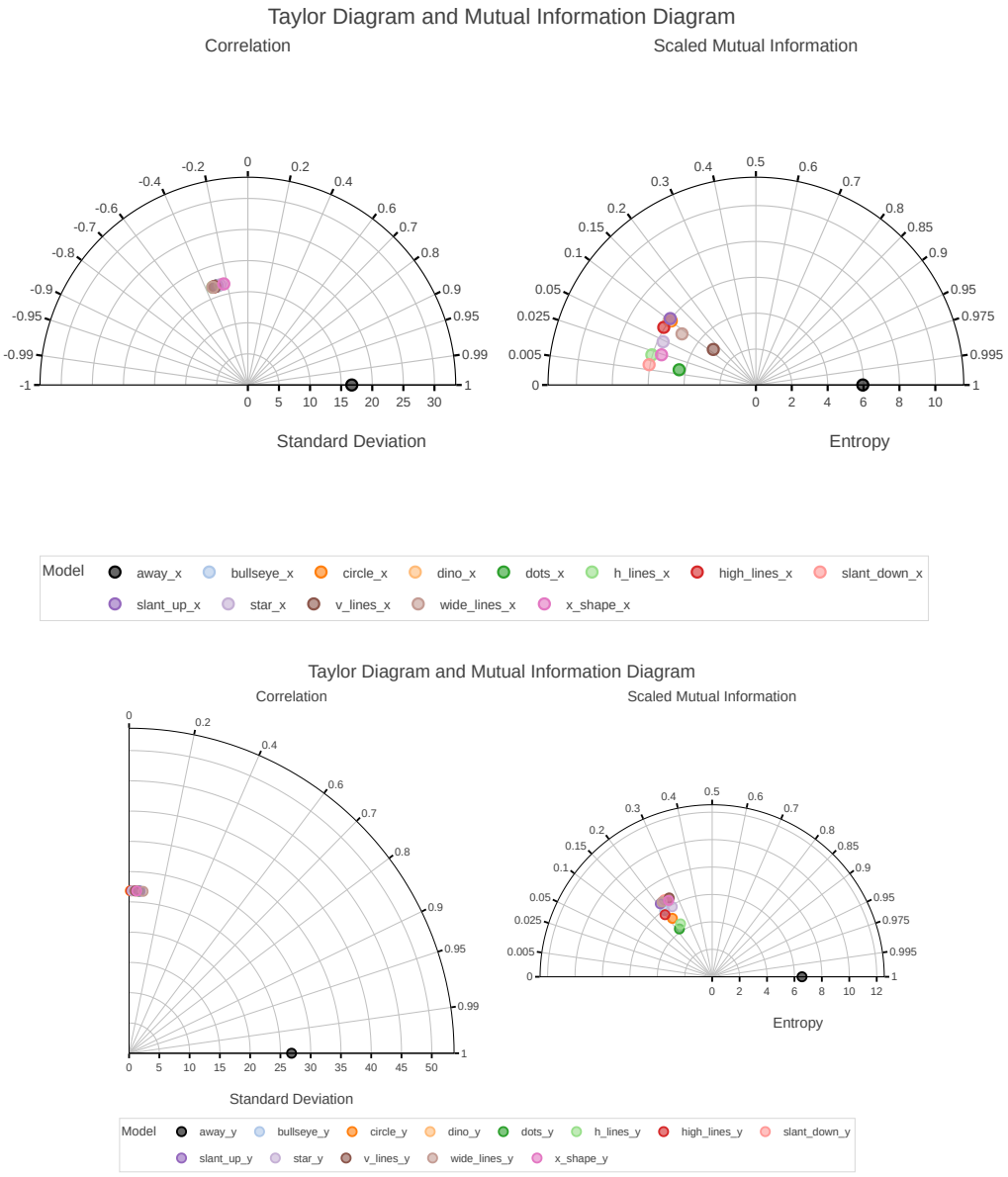


Figure 7.5: *The Datasaurus Dozen* data set. The top row models present x-axis values, while the bottom row models present y-axis values for all thirteen data sets. The models overlap in all diagrams. However, the Taylor Diagrams (top and bottom left) contain models that fully overlap. The user is notified with a *Python* warning about this phenomenon. The MIDs (top and bottom right) give much better results.

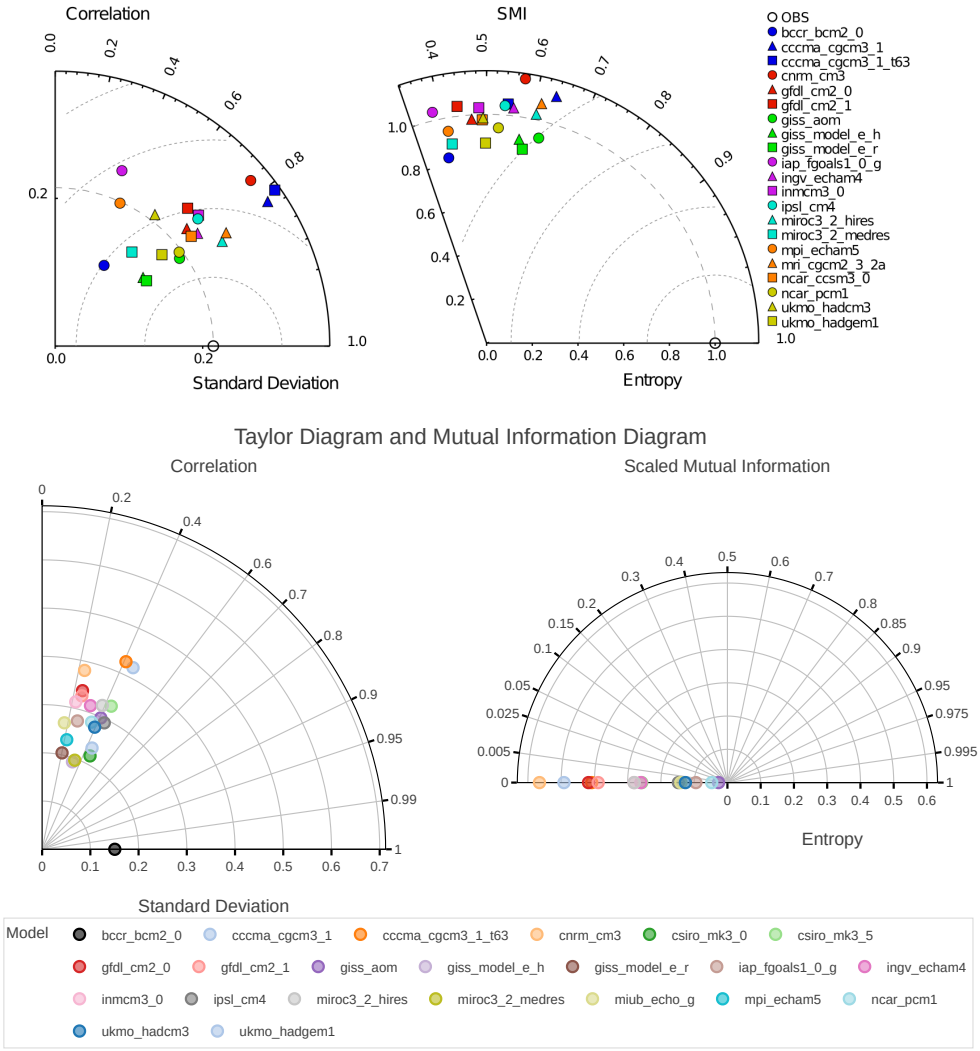


Figure 7.6: *CMIP3* data set. Taylor Diagram and Mutual Information Diagram of CMIP3 air surface temperature data for the *historical* experiment, of only the first *ensamble* run. The top row shows the diagrams adapted from the original study [71], while the bottom row diagrams were created and exported using the library we created — *polar-diagrams*. Model *bccr_bcm2_0* is selected as a reference model. We can see that some models are not visualized on the MID; those models have negative entropies and negative MIs with the reference model. Therefore, by using the Taylor Diagram, we can see the models *miroc3_2_medres* and *csiro_mk3_0* are the most similar to the reference model. This example shows the need to present both diagrams side-by-side. The upper figure is used with permission of Begell House Digital Library, from The mutual information diagram for uncertainty visualization, Correa, C. D., & Lindstrom, P., International Journal for Uncertainty Quantification, 3(3), 2013; permission conveyed through Copyright Clearance Center, Inc.

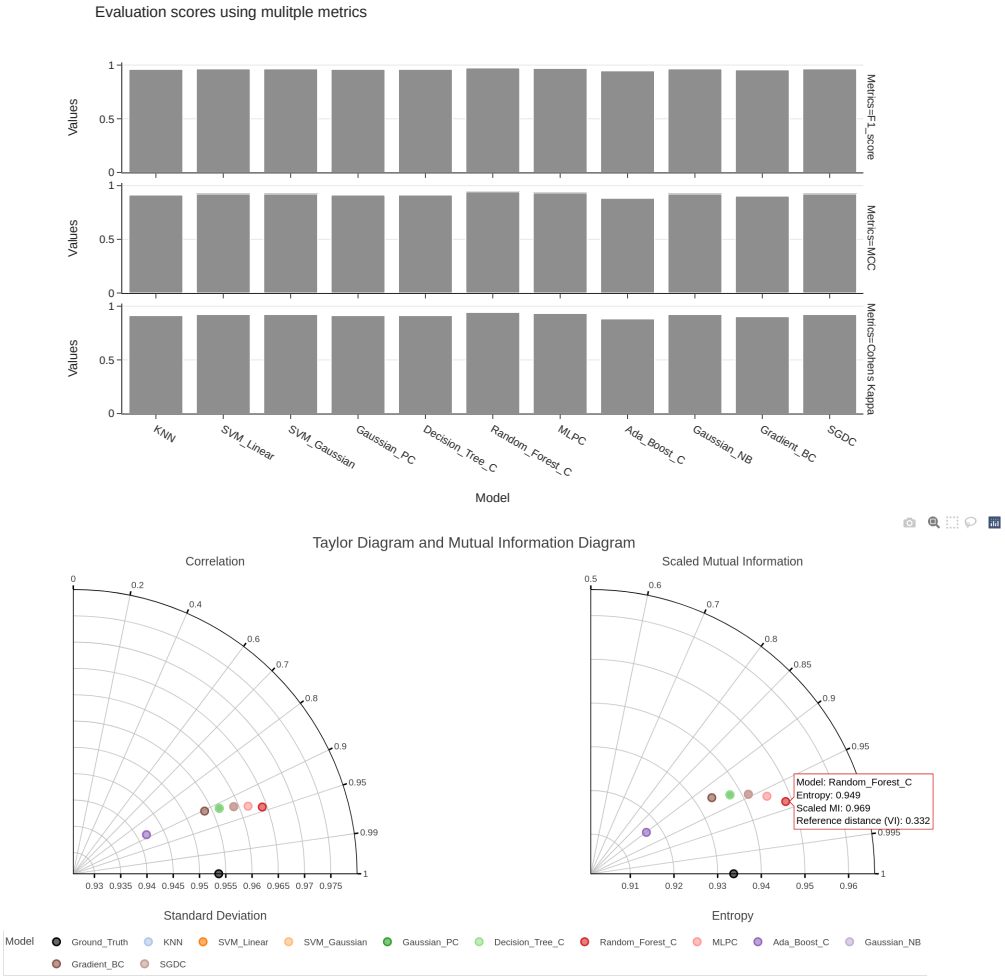
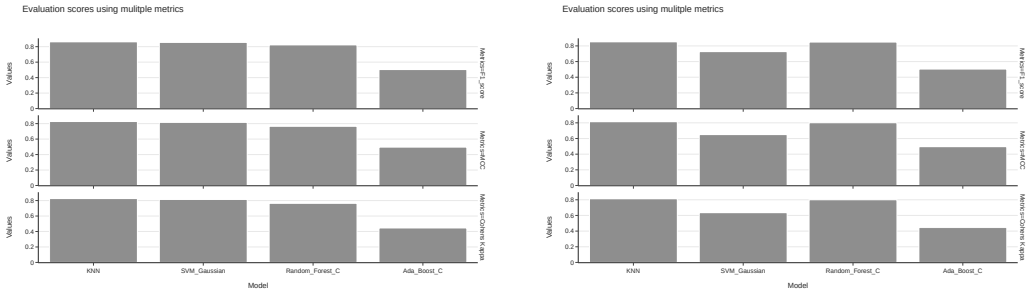


Figure 7.7: *Breast Cancer* data set. Multiple ML models evaluated on *Breast Cancer* data set. The upper part of the figure contains bar charts of commonly used evaluation scores of classification tasks. We can see all models performed similarly well. However, without further inspection of the bar charts using zooming or a tooltip, it is hard to estimate which model performed the best, which model is the second best, and so on. As an alternative to this approach, the user is able to present the performance of each model by creating a table that holds the final scores. On the other hand, Taylor Diagram and MID facilitate clear distinctions between models without any additional work. We can clearly see the *Random_Forest_C* models being the best, *MLPC* being the second best, and *Ada_Boost_C* being the worst model. We can also notice that some models like *KNN*, *SVM_Linear*, and *SVM_Gaussian* are missing. However, this is not the case. The models are overlapping in both diagrams, and the user is notified with a *Python* warning about this phenomenon.



(a) Default hyper-parameters

(b) Changed hyper-parameters

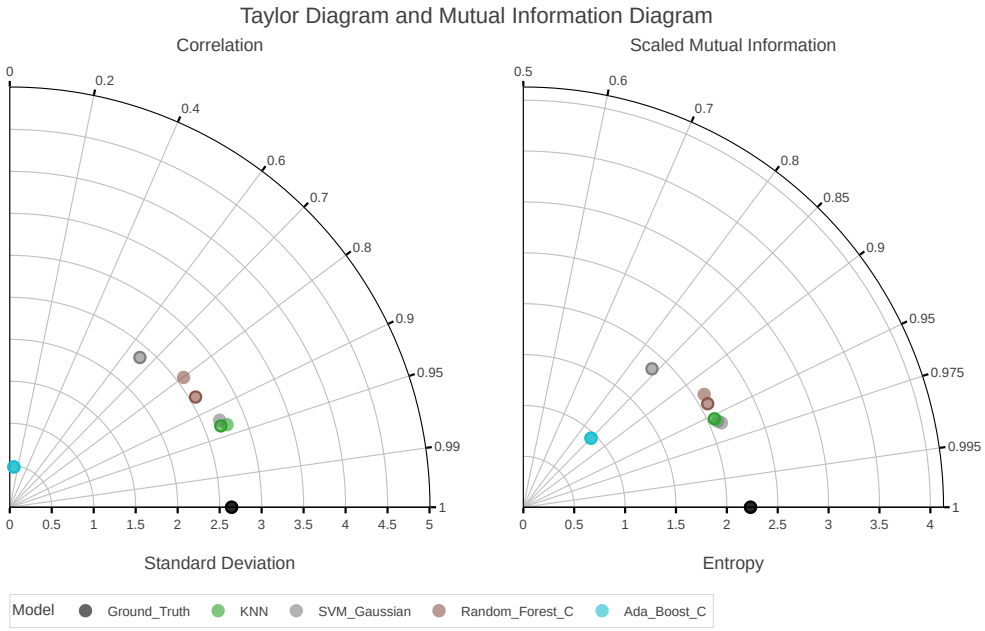


Figure 7.8: *E. Coli* data set. A selection of ML model was used with *E. Coli* data set. To showcase the library's ability to visualize two versions of all models, we conducted two classification experiments using four ML models. In the first experiment, we used models with default hyper-parameters, while in the second experiment, we slightly tweaked hyper-parameters, thus causing some models to perform better and some models to perform worse than in the first run. Models from both experiments were evaluated and visualized, as seen in Figures 7.8a and 7.8b. The visible change in these figures is the decrease in performance of the *SVM_Gaussian* model. This can also be seen in both diagrams since the grey dot with a solid border (which encodes the second version of the model) is further from the *Ground_Truth* than the same borderless grey dot. Moreover, diagrams allow us to easily notice the increase in performance of the *Random_Forest_C* model.

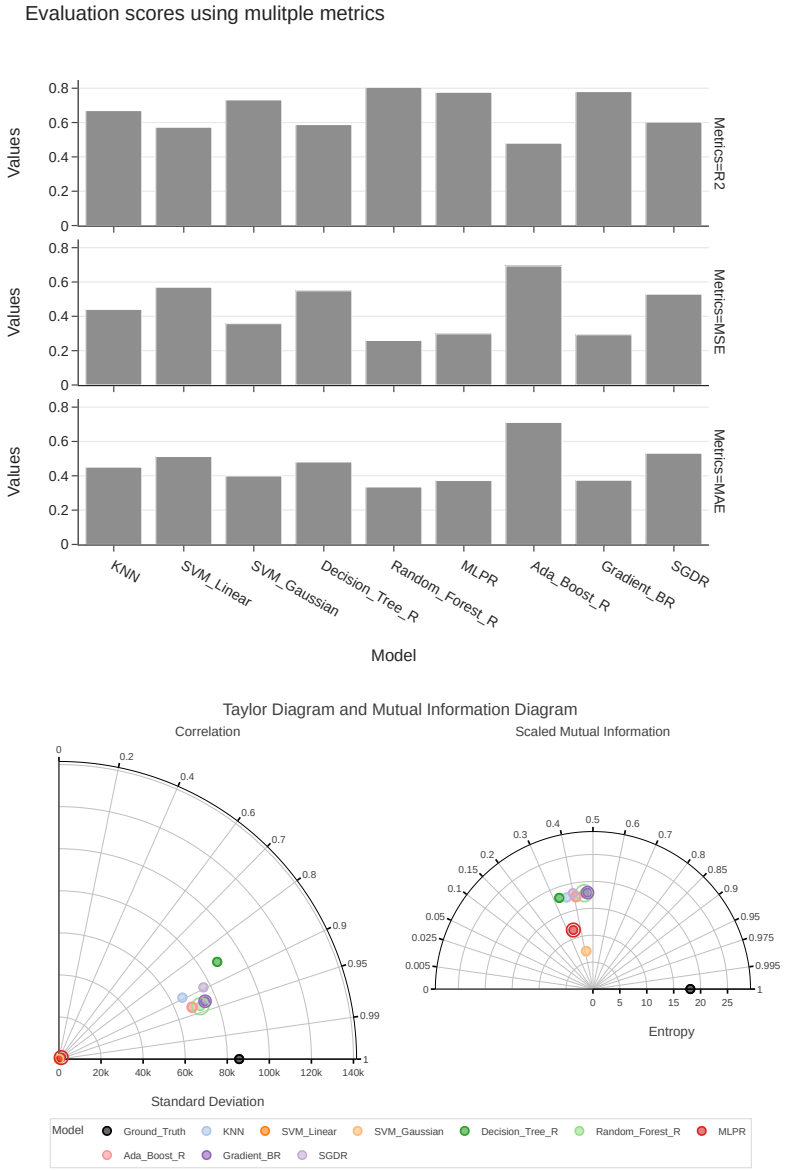


Figure 7.9: *Ames Housing* data set. This example displays the ML models' performances for the regression problem of *Ames Housing* data set. Since the target feature of this data set is continuous, model predictions and ground truth are also continuous. Hence, to visualize all models, the library uses continuous (differential) versions of algorithms for the calculation of entropy and MI. We can see the resulting diagrams are not completely in line, but they agree with both *Random_Forest_R* and *Gradient_BR* being one of the best models for this task. This is completely in line with the commonly used metrics (top row).

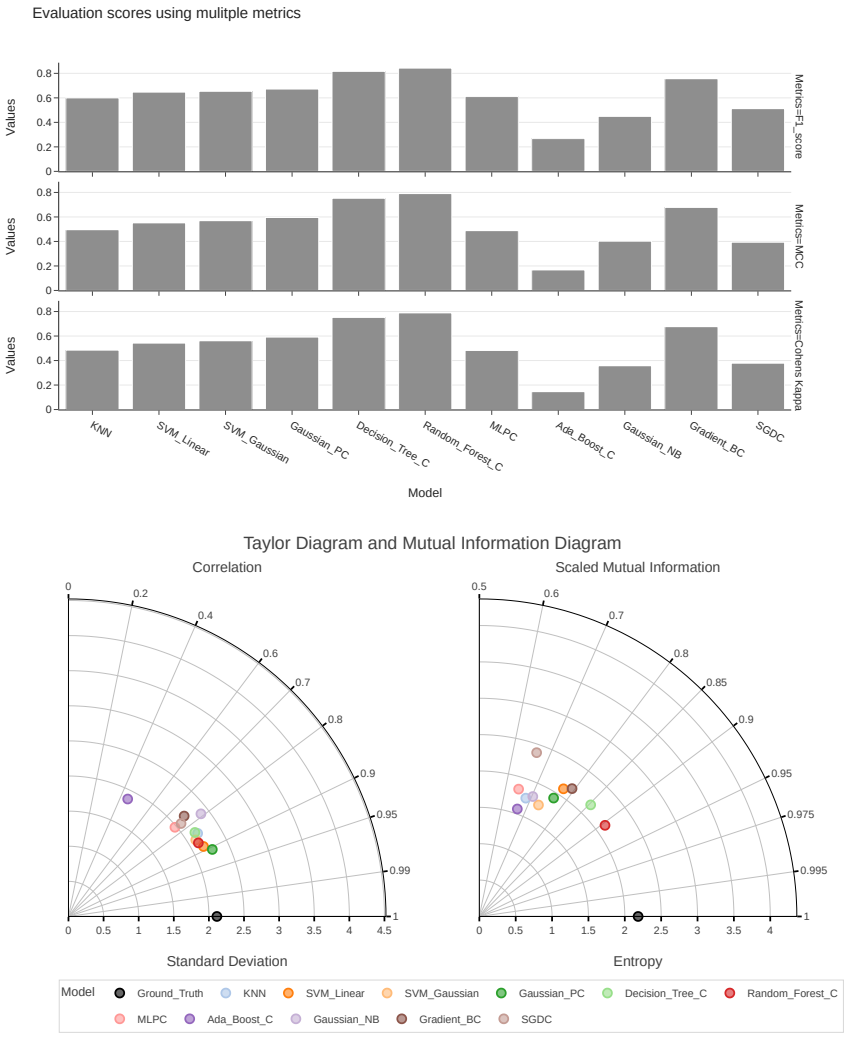


Figure 7.10: *Glass* data set. The performance of ML models while solving the classification task of the *Glass* data set differs greatly from all other results. The traditional way of visualizing ML model performance using bar plots (top row) gives us a clear distinction between model performances instead of being hard to read, as in other examples. Therefore, this approach presents a satisfactory way to visualize ML model performances for this data set. The MID created and exported using *polar-diagrams* (bottom row, right) completely agrees with the results of the previously mentioned approach. The best models are *Random_Forest_C*, *Decision_Tree_C*, and *Gaussian_PC* respectively. The lack of power to capture nonlinear relationships between the models hinders the use of the Taylor Diagram for this example.

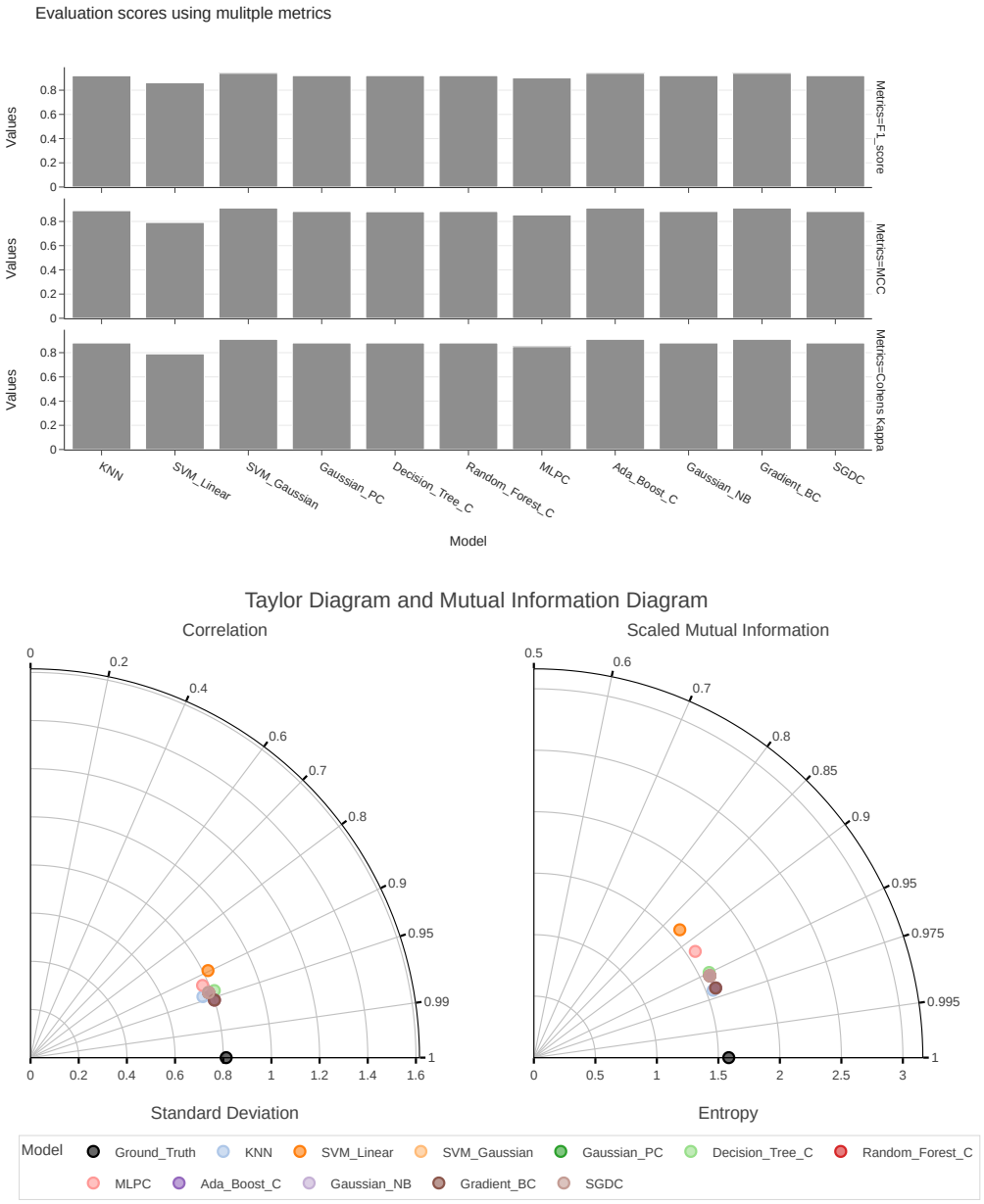


Figure 7.11: *Iris* data set. Both commonly used ML metrics (top row) and diagrams from our library (bottom row) align with *SVM_Linear* and *MLPC* being the worst models on the *Iris* data set, respectively. The best model performances are hard to read from the top-row visualization, but this problem remains in the case of diagrams as well. However, a quick use of the *Zoom* tool provided by *polar-diagrams* would allow us to zoom into the clusters and determine which model is the best.

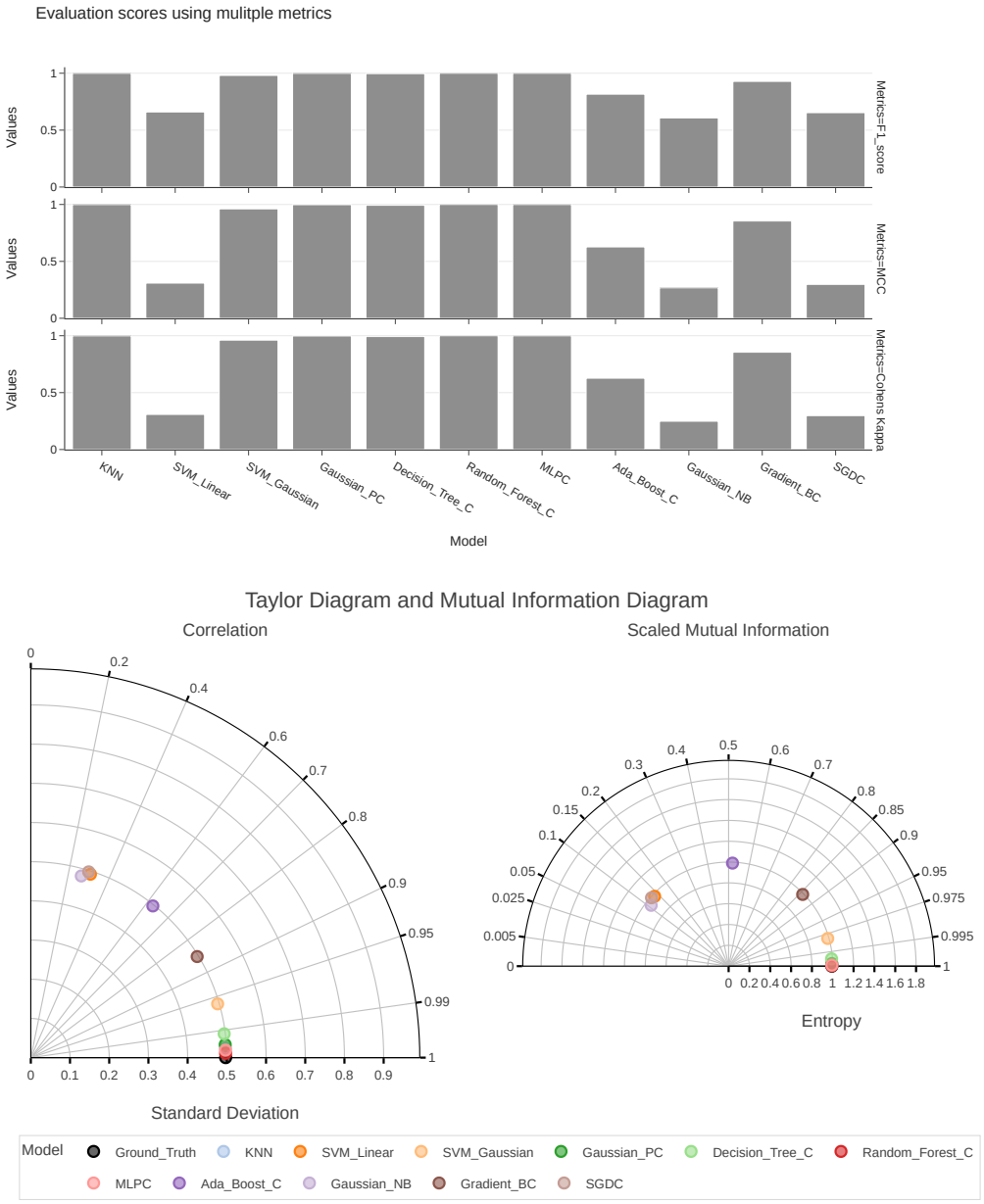


Figure 7.12: *Mushroom* data set. In the case of the *Mushroom* data set, both diagrams from *polar-diagrams* (bottom row) and commonly used ML metrics (top row) align with each other. As with the *Iris* data set, it is hard to assess the best models using the top-row visualization. This is also the case with the diagrams. The interactive *Zoom* tool would help us to single out the best models for this data set quickly. Indeed, it is clear that *Random_Forest_C* and *MLPC* performed the best, which is in alignment with the results from the original study[331].

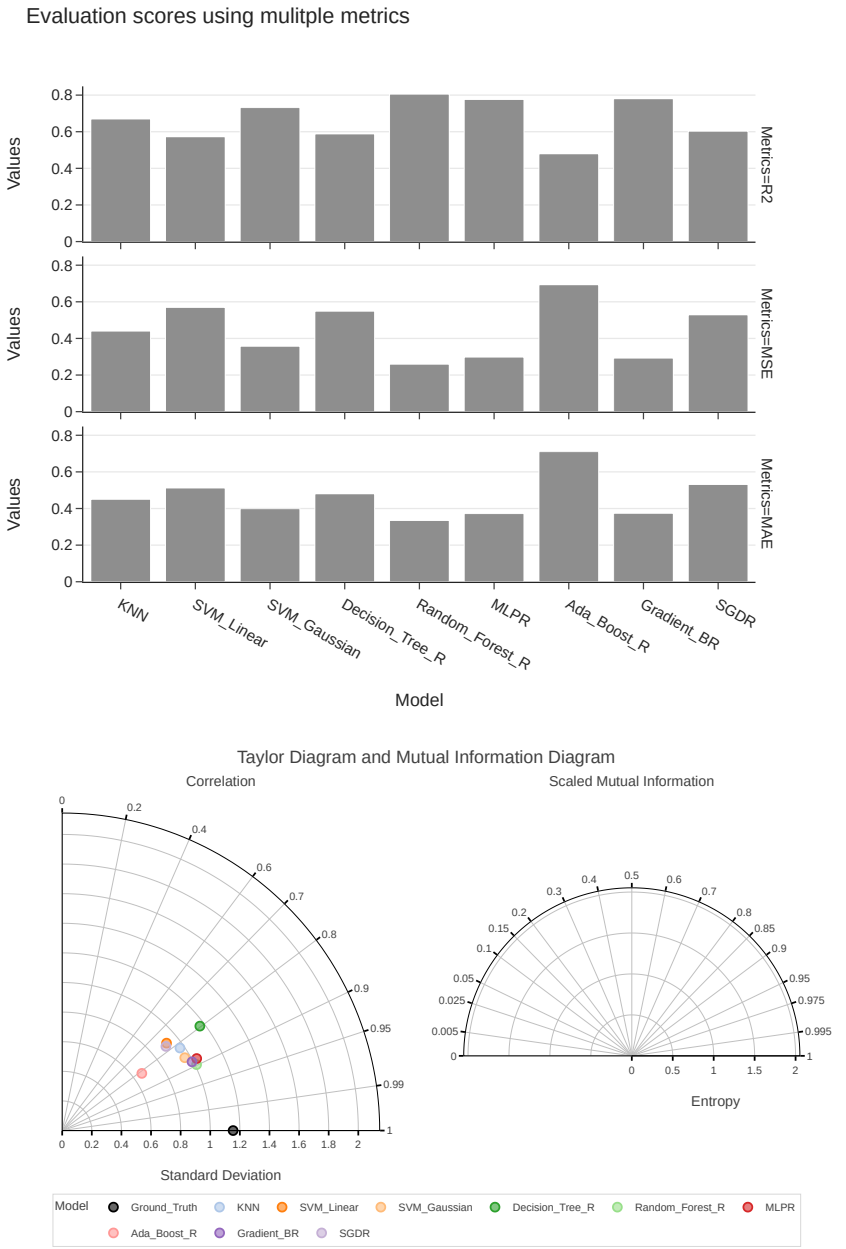


Figure 7.13: *California Housing* data set. In the case of the *California* data set, due to it being a regression problem, all entropies and MIs are negative, hence an empty MID (bottom row, right). Taylor Diagram (bottom row, left) gives better results since it aligns with all commonly used metrics (top row). Both the diagram and bar charts agree with *Random_Forest_R*, *MLPR*, and *Gradient_BR* being the best models for this task, respectively.

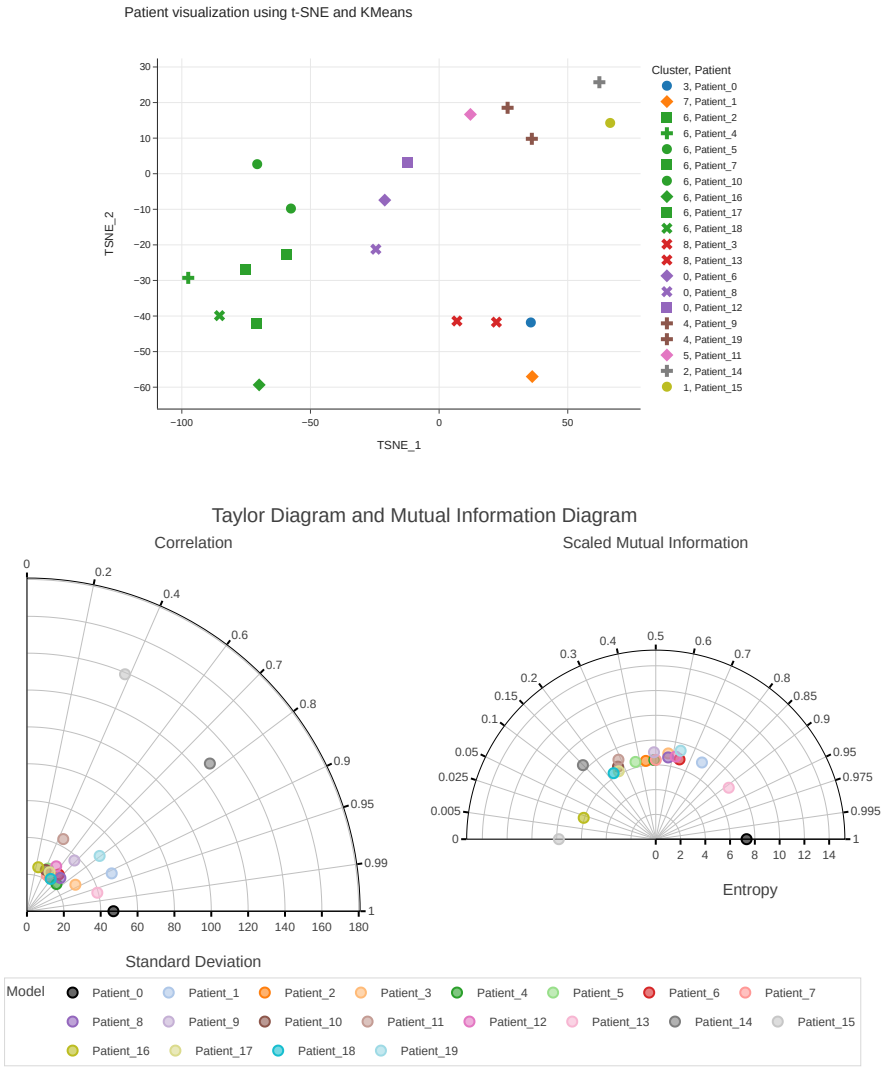


Figure 7.14: *Hepatitis* data set. Hepatitis C patients are visualized using the traditional approach (top row) and *polar-diagrams* (bottom row). While the creation of the top-row scatter plot required algorithm selection, data processing, and computer science knowledge, the bottom-row polar diagrams do not require any domain-specific knowledge or experience. *Patient_0* is selected as a reference model. We can notice the traditional approach (top row) declaring *Patient_0* as the only element of the cluster. However, if we consider distance in a 2-D plot created by *t-SNE*, we see that the most similar patients with the reference patient are *Patient_13* and *Patient_1*. This agrees with both diagrams. However, both diagrams also tell us that *Patient_14* and *Patient_15* are the most dissimilar to the reference model.

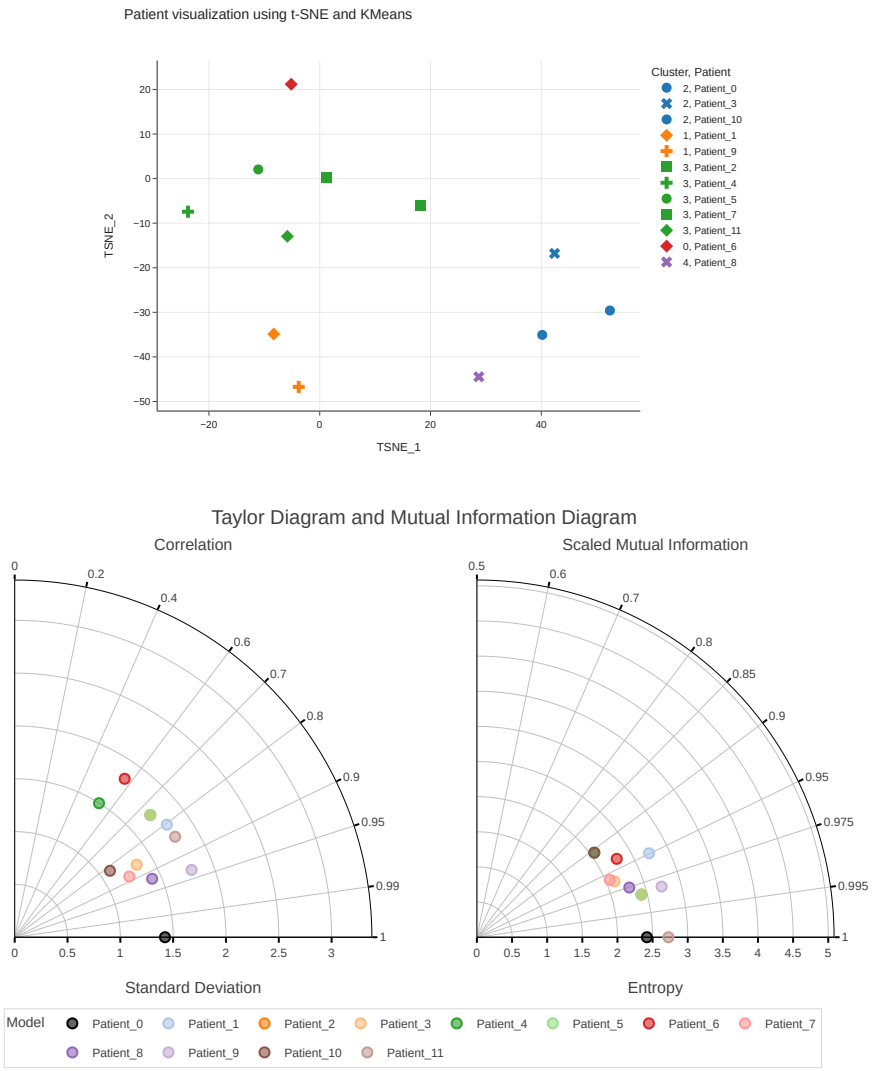


Figure 7.15: *Fertility* data set. *Fertility* data set proves to be one of the examples where the results from the diagrams do not align and disagree with the traditional approach (top row). As we can see, the Taylor Diagram (bottom row, left) depicts *Patient_9*, *Patient_8*, and *Patient_2* as being the most similar to the reference *Patient_0*. On the other hand, MID (bottom row, right) shows *Patient_11* as being the most similar to the reference model. Due to the equidistant nature of models (all having the same VI), the second most similar models are *Patient_5* and *Patient_2*. This is confirmed by the *Python* warning the library produced. The traditional approach places *Patient_10* and *Patient_3* in the same cluster as the reference patient. Due to such a great disagreement, more analysis is required for this data set.

If we knew what it was we were doing, it would not be called research, would it?

Albert Einstein

8

Conclusion

This chapter concludes the dissertation by exploring the implications of previously presented findings for both academics and practitioners. First, the broader conclusions of the work are introduced with a clear focus on problem- and technique-driven approaches for solving intrinsically interdisciplinary problems which were faced during the work on this dissertation. Second, a summary of contributions of each individual study incorporated into this dissertation is given. Achievements and limitations are also covered and discussed in detail. Third and last, the potential of each individual study and the approaches mentioned in this dissertation, are presented and examined comprehensively.

8.1 Broader Conclusions

Even though the approaches from Chapters 2 and 5 were not part of the main bodies of the enclosed studies, they were extensively used to accomplish multiple subtasks in each one of them. Depending on the approach used in each study, these subtasks were different. In the case of the problem-driven approach, some of those subtasks were: formulating the domain problem, getting a sense of the domain data, abstracting and formalizing tasks into other domains (*e.g.*, mathematics, computer science, visualization), abstracting data into more manageable and adequate abstractions, assessing them and finding the most suitable encodings, idioms, and algorithms. In the other case of the technique-driven approach, the subtasks were: inventing or improving an existing algorithm, idiom, or encoding and finding the suitable abstractions or domains in which it is used.

In addition to their valuable role in structuring the problem-solving process, these reasoning approaches represent just one facet of the multi-faceted realm of problem-solving. Recognizing the importance of the problem's origin, typically within a specific research field, presents an opportunity for collaboration and synergy. Practitioners and researchers, armed with their expertise and domain knowledge, come together to tackle the challenges at hand. By pooling their collective understanding and perspectives, they gain a comprehensive grasp of the problem's context, paving the way for

the development of a diverse range of possible solutions. This collaborative approach harnesses the strengths of interdisciplinary collaboration, allowing for a more holistic and innovative problem-solving process.

Moreover, each field produces different kinds of data. Structured or unstructured, high- or low-dimensional, uni- or multi-modal, temporal or spatial, clean or noisy (*i.e.*, with uncertainty), and so on are only some important data properties that one can work with. Besides these, data can also come in low quality, thus further impeding its use for solving certain problems[238]. Because of all these reasons, working on real-world data sets is challenging and requires experience and interdisciplinary knowledge. This is further corroborated when the data analysis and visualization are considered.

Due to the wealth of resources in both steps, researchers are often found overwhelmed when planning how to analyze and visualize the data. This is where domain context, experience, and active collaboration with domain experts come into play as one of the main pillars of interdisciplinary work. If one chooses the problem-driven approach, the need for this is evident right from the start when the problem is being formulated, put into the proper context, and abstracted. With a technique-driven approach, experience comes first when the algorithm, idiom, or encoding is being developed. At the same time, active collaboration and putting everything into the right context are crucial when finding the proper uses of the newly developed solutions.

Last, evaluation is a crucial step in data analysis and visualization as it allows us to determine the effectiveness and reliability of our findings. Through evaluation, we can assess the accuracy of the data, identify potential biases, and determine the degree of confidence we can have in our results. Additionally, evaluation helps us to refine and improve our methods, providing valuable insights into how we can better analyze and visualize data in the future. By thoroughly evaluating our findings, we can ensure that our work is not only useful but also trustworthy.

Besides the impact of each publication in their respective field, all the publications comprising the main body of this dissertation demonstrate the effectiveness of the problem-solving method-

ologies outlined in the paragraphs above and earlier in Chapter 1. Thus, the questions formulated in Section 1.2 were thoroughly considered and answered by carefully designing, implementing, and evaluating the tasks.

Due to the complex interdisciplinary problems that arise from working at the forefront of various domains, including bioinformatics, computer science, biology, mathematics, data science, analysis, and visualization, the proposed reasoning approaches (guidelines) have proven to be an invaluable tool. They provide a systematic and structured approach, breaking down the problem-solving process into manageable steps. They can help practitioners determine the appropriate analytical and visualization techniques to use for a particular problem and evaluate the effectiveness of their solutions. Overall, the proposed reasoning approaches enable practitioners to work efficiently and effectively in interdisciplinary research, providing a roadmap for success. They serve as a valuable tool for researchers and practitioners working at the forefront of multiple domains, enabling them to make significant contributions to their respective fields.

8.2 Summary of Contributions

his dissertation aimed to examine, extend, and improve high-dimensional data analysis, curation, and visualization processes, as well as to suggest new methods for a wider range of scientific problems touching any of these domains. The rest of the section gives chapter-by-chapter summaries, with particular attention to the contributions made by published derivative works that were authored as part of the dissertation.

Chapter 3 investigated previously used and existing data storage technologies and their potential in the data-driven future. The created review uncovered trends and anomalies in data storage technologies over time, but also the lack of standardization in representing various data storage properties to the end user. The review also focused on the UIs implemented in different OSs (historical and existing) which were used to display these storage properties. The gaps in existing approaches

uncovered during the research were documented and used to formulate a user study employed to rank data storage properties of novel media, such as DNA. Thanks to both the survey and the analysis of industry-based UIs and visualizations of data storage properties, a new UI that emphasizes the end user's needs was proposed. The new UI allows users to move between basic and advanced views of data storage properties. Thus, it facilitates greater user engagement with the underlying data storage hardware and fosters a more informed user base. The conducted user study serves as a valuable reminder of the importance of user-centered design in the development of effective and engaging software solutions. Furthermore, the transfer of information from the screen to the user is maximized by employing state-of-the-art visualizations while retaining interactivity, ease of use, and visual appeal.

Chapter 4 examined current integrative tools for managing, processing, analyzing, and visualizing time-series multi-omics data sets. Because of the numerous challenges one can encounter when working with high-dimensional multi-modal data with added temporal dimension, the tools have to be easy to use, provide step-by-step guidance with all tasks of the mentioned pipeline, and allow users to export their findings easily. Due to the lack of such a tool at the time of writing the study, a new tool *MOVIS* was developed to support and integrate all essential tasks when working with multi-omics time-series data sets. The tool can be run online and locally, thus removing the size limitations of these data sets. It enables users to import genomics, proteomics, metabolomics, transcriptomics, and physico-chemical data sets in various data formats with ease. If necessary, users can then preprocess and embed the data to allow further analysis and visualization procedures. Besides previous, multiple state-of-the-art methods for clustering and visualizing the data are available, thus covering data with different properties and structures. With the use of state-of-the-art *Python* front-end libraries, users are empowered to analyze and visualize data from multiple omics simultaneously while retaining visual and analytical integrity. Finally, it is noteworthy that the resulting data visualizations not only provide users with interactive capability but also can be exported in many publication-ready formats. These formats include image files, which can be directly inserted into research papers, pre-

sentations, or reports, and other vector formats, such as PDF, SVG, and EPS, which are great for printed materials. Being able to export the visualizations in various formats enhances the flexibility and usability of the data analysis output, enabling users to more easily and effectively communicate their findings to a broader audience.

Chapter 6 inspected the current research landscape in the domain of molecular fingerprinting techniques, assessed the advantages and limitations of each method, and proposed a new molecular encoding mechanism. The newly proposed method abstracts each molecule into a graph, where atoms are considered nodes while inter-atomic chemical bonds are considered edges. This natural translation from molecules to graphs allowed the usage of a plethora of graph algorithms, developed in mathematics and computer science, to be used to manipulate and process every molecule. Through its intuitive graph-based approach, the proposed method can be easily integrated into existing molecular analysis workflows. By traversing their carbon chains, every molecule is encoded in a graph which is then represented using the matrix where columns represent visited carbon atoms and rows represent neighbors of each visited carbon atom. Due to the traversal of the carbon chains, the method presents the only solution which could be used to encode any organic molecule in a uniform way parametrically. The parameters enable users to encode molecules in binary and discretized tabular and image forms, hence allowing four different types of molecular encodings. The ability to encode molecules in binary and discretized tabular forms makes it easier for researchers to analyze and compare the chemical structures of different molecules. The proposed method performs equally well compared to other fingerprinting algorithms while offering greater flexibility and facilitating the creation of image encodings that could be used for machine learning experiments. Due to the modular development of the resulting library, the proposed method's flexibility extends beyond its current abilities since it can be adapted to suit the needs of different research domains.

Chapter 7 explored the mathematical principles of the Taylor Diagram (TD) and its use for model comparisons in the climate domain. The alternative of this polar diagram rooted in informa-

tion theory, known as Mutual Information Diagram (MID), was inspected as a viable substitute for a more traditional TD for solving problems in other domains besides climatology. The lack of implementation of the second diagram motivated the study further while also proposing and demonstrating new capabilities of both diagrams, thus broadening their benefits and validating their uses in other domains. The presented coupling of the implementations of both diagrams allows users to compare models regardless of their linear or nonlinear relationships. By following the industry standards and conventions, the resulting library offers an innovative visualization solution for both polar diagrams. It can process, analyze, and visualize numerical results from any model or domain given in tabular format. The library's user-friendly interface is designed to accommodate both novice and expert users, providing a low barrier to entry. Due to the support for various interactive elements, the library empowers users to inspect one or both diagrams simultaneously before deciding to save them in a publication-ready image format. The study and the value of the library were further supported by evaluating them against current approaches in assessing model performance in climatology, biomedicine, machine learning, and other relevant fields. The study highlights the value of Taylor and Mutual Information Diagrams in areas where traditional statistical techniques may not be sufficient to provide comprehensive insights. Moreover, the proposed coupling of the diagrams could have significant implications for studying complex systems in fields such as ecology and epidemiology, which were never considered before.

8.3 Future Work and Discussion

Quantifying the value of top-down and bottom-up approaches for a problem-solving process is challenging, if not impossible. Their intuitive nature makes them familiar to all people, hence allowing everyone to apply them in day-to-day tasks. However, without the proper research and structuralization, these approaches would not become as usable as they are today. Being problem-agnostic, they show their usefulness in a plethora of domains. From material science [281], molecular biology [178,360], environmental sciences [94], to business management [308],

ethics [5], and education [284], both approaches proved their value and justified their importance in providing practical solutions to a wide range of problems. They provide a structured approach for problem-solving that can be iterated over and refined as needed, making them an indispensable tool in any problem-solving toolbox.

Although top-down and bottom-up problem-solving approaches are generally applicable, their effectiveness can be limited in specific research fields due to the specialized nature of each field. Researchers often need to refine and tailor these approaches for their particular research area. For example, the scientific papers referenced in the previous paragraph demonstrate how top-down and bottom-up approaches have been adapted to different fields such as material science, molecular biology, environmental sciences, business management, ethics, and education. However, many of these studies have not generalized their approaches to their respective fields, leaving other researchers without the necessary guidelines to apply these approaches to their own work.

Indeed, the lack of per-field generalized guidelines is currently the biggest obstacle to the wider adoption and effective use of these approaches. To address this issue, this dissertation focuses on improving the application of top-down and bottom-up approaches to analytical problems and tasks dealing with high-dimensional data. While the dissertation does not propose a new generalized guideline, it borrows and adapts the existing guidelines from the visualization domain. As previously demonstrated in the enclosed studies, these adapted guidelines can help researchers systematically address high-dimensional analytical problems, leading to solutions of the original domain-specific problems. By providing guidance on how to apply top-down and bottom-up approaches to specific research fields, this dissertation aims to contribute to the wider dissemination and effective use of these approaches.

Moreover, philosophy has played a crucial role in shaping the foundations of both top-down and bottom-up approaches. These methods share many commonalities with philosophical thinking, such as the structured way of approaching complex problems by breaking them down into smaller manageable parts, testing assumptions, and iteratively building new knowledge. In addition, critical

thinking, which is fundamental in philosophy, is also essential for using these approaches effectively. Therefore, it is important to recognize the value of philosophy when working with complex problems, as it can not only inform the use of top-down and bottom-up approaches but also provide additional tools and methods for critical thinking.

The insights from philosophy, psychology and their methods can be particularly useful in any field that deals with human-machine interaction. The study of human nature and its limitations is essential to understanding how people interact with machines and what they need from them. By understanding the philosophical and psychological underpinnings of human thinking and behavior, researchers can develop more effective ways of designing human-machine interaction systems, as well as better ways of evaluating their performance.

In addition to designing better human-machine interaction systems, insights from philosophy and psychology can also help us create better visualizations and data analysis tools. By adapting existing or developing new visual encodings and interactive idioms based on our understanding of human perception and cognition, we can create more effective and easily interpretable visualizations. Similarly, by taking into account the limitations of human perception and cognition, we can create better exploratory data analysis tools that are more streamlined and accessible, improving their adoption and impact. Indeed, additional strategies can be utilized and integrated to enhance the development of more effective human-machine interaction systems. A notable example is the model proposed by Tominski *et al.* [310], which emphasizes minimizing various forms of separation between humans and machines, such as conceptual, spatial, and temporal gaps. By addressing and reducing these gaps, it becomes possible to optimize and achieve the best possible level of interaction between humans and machine systems.

Overall, the combination of bottom-up and top-down approaches provides a powerful toolkit for analyzing and visualizing high-dimensional data. By combining insights from philosophy, psychology, mathematics, computer science, and data science, we can create more powerful and effective tools and methods to tackle complex problems.

References

[1] Leonard M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(5187):1021–1024, 1994. doi: 10.1126/science.7973651. URL https://www.science.org/doi/abs/10.1126/science.7973651.

[2] Deepak Ajwani, Itay Malinger, Ulrich Meyer, and Sivan Toledo. Characterizing the performance of flash memory storage devices and its impact on algorithm design. In Catherine C. McGeoch, editor, *Experimental Algorithms*, pages 208–219, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-68552-4.

[3] Abdullah Al Mamun, GuoXiao Guo, and Chao Bi. *Hard Disk Drive: Mechatronics and Control (Automation and Control Engineering)*. CRC Press, hardcover edition, 11 2006. ISBN 978-0849372537. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0849372534.

[4] Hadi Alizadeh Noughabi. Entropy estimation using numerical methods. *Annals of Data Science*, 2, 06 2015. doi: 10.1007/s40745-015-0045-9.

[5] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155, Sep 2005. ISSN 1572-8439. doi: 10.1007/s10676-006-0004-4. URL https://doi.org/10.1007/s10676-006-0004-4.

[6] Joseph Amankwah-Amoah. Competing technologies, competing forces: The rise and fall of the floppy disk, 1971–2010. *Technological Forecasting and Social Change*, 107:121–129, 2016. ISSN 0040-1625. doi: https://doi.org/10.1016/j.techfore.2016.03.019. URL https://www.sciencedirect.com/science/article/pii/S0040162516000858.

[7] Wang An. Pulse transfer controlling device, May 1955. US Patent 2,708,722.

[8] Edgar Anderson. A semigraphical method for the analysis of complex problems. *Proceedings of the National Academy of Sciences*, 43(10):923–927, 1957. doi: 10.1073/pnas.43.10.923. URL https://www.pnas.org/doi/abs/10.1073/pnas.43.10.923.

[9] D. F. Andrews. Plots of high-dimensional data. *Biometrics*, 28(1):125–136, 1972. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2528964.

[10] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, page 49–60, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130848. doi: 10.1145/304182.304187. URL https://doi.org/10.1145/304182.304187.

[11] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973. doi: 10.1080/00031305.1973.10478966. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966.

[12] Aleksandar Anžel, Dominik Heider, and Georges Hattab. The visual story of data storage: From storage properties to user interfaces. *Computational and Structural Biotechnology Journal*, 19:4904–4918, 2021. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2021.08.031. URL https://www.sciencedirect.com/science/article/pii/S2001037021003627.

[13] Aleksandar Anžel, Dominik Heider, and Georges Hattab. Movis: A multi-omics software solution for multi-modal time-series clustering, embedding, and visualizing tasks. *Computational and Structural Biotechnology Journal*, 20:1044–1055, 2022. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2022.02.012. URL https://www.sciencedirect.com/science/article/pii/S2001037022000526.

[14] Aleksandar Anžel, Dominik Heider, and Georges Hattab. Interactive polar diagrams for model comparison. *Computer Methods and Programs in Biomedicine*, 2023. ISSN 1872-7565.

[15] Daniel Archambault, Tamara Munzner, and David Auber. Grouse: Feature-Based, Steerable Graph Hierarchy Exploration. In K. Museth, T. Moeller, and A. Ynnerman, editors, *Eurographics/ IEEE-VGTC Symposium on Visualization*. The Eurographics Association, 2007. ISBN 978-3-905673-45-6. doi: 10.2312/VisSym/EuroVis07/067-074.

[16] A.O. Artero, M.C.F. de Oliveira, and H. Levkowitz. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement. *Tenth International Conference on Information Visualisation (IV'06)*, pages 707–712, July 2006. ISSN 2375-0138. doi: 10.1109/IV.2006.49.

[17] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

[18] Skipwith W Athey. *Magnetic tape recording*, volume 5038. US Government Printing Office, 1966.

[19] D. Bajusz, A. Rácz, and K. Héberger. 3.14 - chemical data formats, fingerprints, and other molecular descriptions for database analysis and searching. In Samuel Chackalamannil, David Rotella, and Simon E. Ward, editors, *Comprehensive Medicinal Chemistry III*, pages

329–378. Elsevier, Oxford, 2017. ISBN 978-0-12-803201-5. doi: https://doi.org/10.1016/B978-0-12-409547-2.12345-5. URL https://www.sciencedirect.com/science/article/pii/B9780124095472123455.

[20] Kevin Baker, Saul Greenberg, and Carl Gutwin. Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, page 96–105, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581135602. doi: 10.1145/587078.587093. URL https://doi.org/10.1145/587078.587093.

[21] L. Bartram, M. Correll, and M. Tory. Untidy data: The unreasonable effectiveness of tables. *IEEE Transactions on Visualization & Computer Graphics*, 28(01):686–696, January 2022. ISSN 1941-0506. doi: 10.1109/TVCG.2021.3114830.

[22] Lynly Beard and Negeen Aghassibake. Tableau (version 2020.3). *Journal of the Medical Library Association*, 109(1), January 2021. doi: 10.5195/jmla.2021.1135. URL https://doi.org/10.5195/jmla.2021.1135.

[23] Benjamin B. Bederson, Ben Shneiderman, and Martin Wattenberg. Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *ACM Trans. Graph.*, 21 (4):833–854, October 2002. ISSN 0730-0301. doi: 10.1145/571647.571649. URL https://doi.org/10.1145/571647.571649.

[24] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317.

[25] Richard E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 2010. ISBN 9781400835386. doi: doi:10.1515/9781400835386. URL https://doi.org/10.1515/9781400835386.

[26] Afef Ben Brahim and Mohamed Limam. Ensemble feature selection for high dimensional data: a new method and a comparative study. *Advances in Data Analysis and Classification*, 12(4):937–952, December 2018. ISSN 1862-5355. doi: 10.1007/s11634-017-0285-y. URL https://doi.org/10.1007/s11634-017-0285-y.

[27] Wasim Ahmad Bhat. Bridging data-capacity gap in big data storage. *Future Generation Computer Systems*, 87:538–548, 2018. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2017.12.066. URL https://www.sciencedirect.com/science/article/pii/S0167739X17312876.

[28] Bharat Bhushan. *Tribology and Mechanics of Magnetic Storage Devices*. Springer, hardcover edition, 7 1996. ISBN 978-0387946276. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0387946276.

[29] Ayan Biswas, Soumya Dutta, Han-Wei Shen, and Jonathan Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and*

Computer Graphics, 19(12):2683–2692, Dec 2013. ISSN 1941-0506. doi: 10.1109/TVCG.2013.133.

[30] Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, and George M. Church. Forward error correction for dna data storage. *Procedia Computer Science*, 80:1011–1022, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.05.398. URL https://www.sciencedirect.com/science/article/pii/S1877050916308742. International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[31] P. Bonyhard, J. Geusic, A. Bobeck, Yu-Ssu Chen, P. Michaelis, and J. Smith. Magnetic bubble memory chip design. *IEEE Transactions on Magnetics*, 9(3):433–436, September 1973. ISSN 1941-0069. doi: 10.1109/TMAG.1973.1067599.

[32] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Springer, paperback edition, 11 2010. ISBN 978-1441920461. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1441920463.

[33] James Bornholt, Randolph Lopez, Douglas M. Carmean, Luis Ceze, Georg Seelig, and Karin Strauss. A dna-based archival storage system. *SIGPLAN Not.*, 51(4):637–649, March 2016. ISSN 0362-1340. doi: 10.1145/2954679.2872397. URL https://doi.org/10.1145/2954679.2872397.

[34] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD. ISBN 978-3-7908-2604-3.

[35] Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998. ISSN 00905364. URL http://www.jstor.org/stable/120055.

[36] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[37] D.W. Brisson. *Hypergraphics: Visualizing Complex Relationships in Art, Science and Technology*. AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE. Selected symposia. Westview, 1978. URL https://books.google.de/books?id=J6yOzQEACAAJ.

[38] R.G. Brown, A.M. Gleason, and M.A. Brown. *Advanced Mathematics: Precalculus with Discrete Mathematics and Data Analysis*. McDougal Littell, 1994. ISBN 9780395421697. URL https://books.google.de/books?id=0QwOxgEACAAJ.

[39] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, J. Vanderplas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. Api design for machine learning software: experiences from the scikit-learn project. *ArXiv*, abs/1309.0238, 2013.

[40] Michael Burch and Daniel Weiskopf. *On the Benefits and Drawbacks of Radial Diagrams*, chapter On the Benefits and Drawbacks of Radial Diagrams, pages 429–451. Springer New York, New York, NY, 2014. ISBN 978-1-4614-7485-2. doi: 10.1007/978-1-4614-7485-2_17. URL https://doi.org/10.1007/978-1-4614-7485-2_17.

[41] Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, pages 42–53, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74976-9.

[42] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101. URL https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

[43] Sheelagh Carpendale. *Evaluating Information Visualizations*, pages 19–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-70956-5. doi: 10.1007/978-3-540-70956-5_2. URL https://doi.org/10.1007/978-3-540-70956-5_2.

[44] John B Carter, Scott H Davis, Daniel J Dietterich, Steven J Frank, Robert S Phillips, John Woods, David Porter, and Hsin H Lee. System and method for providing highly available data storage using globally addressable memory, June 1999. US Patent 5,909,540.

[45] Gordon CC, Cooper CA, C.A. Senior, Helene Hewitt, J.M. Gregory, Timothy Johns, J. Mitchell, and R.A. Wood. The simulation of sst, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments. *Climate Dynamics*, 16:147–168, 02 2000. doi: 10.1007/s003820050010.

[46] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015. ISSN 1046-2023. doi: https://doi.org/10.1016/j.ymeth.2014.08.005. URL https://www.sciencedirect.com/science/article/pii/S1046202314002631. Virtual Screening.

[47] Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. Technical report, Stanford University, Stanford, CA, USA, 1979.

[48] Shubham Chandak, Kedar Tatwawadi, Idoia Ochoa, Mikel Hernaez, and Tsachy Weissman. SPRING: a next-generation compressor for FASTQ data. *Bioinformatics*, 35(15):2674–2676, 12 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty1015. URL https://doi.org/10.1093/bioinformatics/bty1015.

[49] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), may 2011. ISSN 2157-6904. doi: 10.1145/1961189.1961199. URL https://doi.org/10.1145/1961189.1961199.

[50] Hsu Chang. *Magnetic-bubble Memory Technology*. Marcel Dekker Inc, hardcover edition, 1978. ISBN 978-0824767952. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0824767950.

[51] Joseph Chang and Steven Hanna. Air quality model performance evaluation. *Meteorology and Atmospheric Physics*, 87:167–196, 09 2004. doi: 10.1007/s00703-003-0070-7.

[52] Kumardeep Chaudhary, Olivier B. Poirion, Liangqun Lu, and Lana X. Garmire. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research*, 24(6):1248–1259, 03 2018. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-17-0853. URL https://doi.org/10.1158/1078-0432.CCR-17-0853.

[53] CHAOMEI CHEN and MARY P. CZERWINSKI. Empirical evaluation of information visualizations: an introduction. *International Journal of Human-Computer Studies*, 53(5):631–635, 2000. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.2000.0421. URL https://www.sciencedirect.com/science/article/pii/S107158190090421X.

[54] M. Chen, M. Feixas, I. Viola, A. Bardera, H.W. Shen, and M. Sbert. *Information Theory Tools for Visualization*. AK Peters Visualization Series. CRC Press, 2016. ISBN 9781315352237. doi: https://doi.org/10.1201/9781315369228. URL https://doi.org/10.1201/9781315369228.

[55] Peter Pin-Shan Chen. The compact disk rom: How it works: An offshoot of the compact digital audio disk, this multimegabyte storage technique is revolutionizing database technology. *IEEE Spectrum*, 23(4):43–49, April 1986. ISSN 1939-9340. doi: 10.1109/MSPEC.1986.6370869.

[56] Tien Chi Chen and Hsu Chang. Magnetic bubble memory and logic. In Marshall C. Yovits, editor, *Magnetic Bubble Memory and Logic*, volume 17 of *Advances in Computers*, pages 223–282. Elsevier, 1978. doi: https://doi.org/10.1016/S0065-2458(08)60393-9. URL https://www.sciencedirect.com/science/article/pii/S0065245808603939.

[57] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973. doi: 10.1080/01621459.1973.10482434. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1973.10482434.

[58] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):6, Jan 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7. URL https://doi.org/10.1186/s12864-019-6413-7.

[59] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7, 2021. doi: https://doi.org/10.7717/peerj-cs.623. URL https://doi.org/10.7717/peerj-cs.623.

[60] Kuo-Chen Chou and David W. Elrod. Prediction of membrane protein types and subcellular locations. *Proteins: Structure, Function, and Bioinformatics*, 34(1):137–153, 1999. doi: https://doi.org/10.1002/(SICI)1097-0134(19990101)34:1<137::AID-PROT11>3.0.CO;2-O. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0134%2819990101%2934%3A1%3C137%3A%3AAID-PROT11%3E3.0.CO%3B2-O.

[61] George M. Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012. doi: 10.1126/science.1226355. URL https://www.science.org/doi/abs/10.1126/science.1226355.

[62] William S. Cleveland. *Visualizing data / William S. Cleveland*. At&T Bell Laboratories ; Summit, N.J. : Published by Hobart Press, Murray Hill, N.J., 1993. ISBN 0963488406.

[63] Dean Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19, 11 2011. doi: 10.1080/10691898.2011.11889627.

[64] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL https://doi.org/10.1093/bioinformatics/btp163.

[65] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

[66] R. L. Comstock. Review modern magnetic materials in data storage. *Journal of Materials Science: Materials in Electronics*, 13(9):509–523, September 2002. ISSN 1573-482X. doi: 10.1023/A:1019642215245. URL https://doi.org/10.1023/A:1019642215245.

[67] Ashley Mae Conard, Nathaniel Goodman, Yanhui Hu, Norbert Perrimon, Ritambhara Singh, Charles Lawrence, and Erica Larschan. TIMEOR: a web-based tool to uncover temporal regulatory mechanisms from multi-omics data. *Nucleic Acids Research*, 49 (W1):W641–W653, 06 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab384. URL https://doi.org/10.1093/nar/gkab384.

[68] B. Copeland. The manchester computer: A revised history part 1: The memory. *IEEE Annals of the History of Computing*, 33(1):4–21, January 2011. ISSN 1934-1547. doi: 10.1109/MAHC.2010.1.

[69] Irving Copi, Carl Cohen, and Daniel Flage. *Essentials of Logic*. Routledge, paperback edition, 7 2006. ISBN 978-0132380348. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=013238034X.

[70] Michael Cornwell. Anatomy of a solid-state drive: While the ubiquitous ssd shares many features with the hard-disk drive, under the surface they are completely different. *Queue*,

10(10):30–36, October 2012. ISSN 1542-7730. doi: 10.1145/2381996.2385276. URL https://doi.org/10.1145/2381996.2385276.

[71] Carlos Correa and Peter Lindstrom. The mutual information diagram for uncertainty visualization. *International Journal for Uncertainty Quantification*, 3:187–201, 01 2013. doi: 10.1615/Int.J.UncertaintyQuantification.2012003959.

[72] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967. ISSN 1557-9654. doi: 10.1109/TIT.1967.1053964.

[73] Jonathan P.L Cox. Long-term data storage in dna. *Trends in Biotechnology*, 19(7):247–250, 2001. ISSN 0167-7799. doi: https://doi.org/10.1016/S0167-7799(01)01671-7. URL https://www.sciencedirect.com/science/article/pii/S0167779901016717.

[74] Michael A. A. Cox and Trevor F. Cox. *Multidimensional Scaling*, chapter Multidimensional Scaling, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-33037-0. doi: 10.1007/978-3-540-33037-0_14. URL https://doi.org/10.1007/978-3-540-33037-0_14.

[75] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, mar 2002. ISSN 1532-4435.

[76] Przeniyslaw Crzcgorzewski and Robert Wirczorkowski. Entropy-based goodness-of-fit test for exponentiality. *Communications in Statistics - Theory and Methods*, 28(5):1183–1202, 1999. doi: 10.1080/03610929908832351. URL https://doi.org/10.1080/03610929908832351.

[77] Peter Csermely, Tamás Korcsmáros, Huba J.M. Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics*, 138(3):333–408, 2013. ISSN 0163-7258. doi: https://doi.org/10.1016/j.pharmthera.2013.01.016. URL https://www.sciencedirect.com/science/article/pii/S0163725813000284.

[78] Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito. Inductive learning for the semantic web: What does it buy? *Semantic Web*, 1:53–59, 2010. doi: 10.3233/SW-2010-0007. URL https://doi.org/10.3233/SW-2010-0007. 1-2.

[79] Marius D'Amboise and Michel J. Bertrand. General index of molecular complexity and chromatographic retention data. *Journal of Chromatography A*, 361:13–24, 1986. ISSN 0021-9673. doi: https://doi.org/10.1016/S0021-9673(01)86889-8. URL https://www.sciencedirect.com/science/article/pii/S0021967301868898.

[80] B Dash, Debahuti Mishra, Amiya Rath, and Milu Acharya. A hybridized k-means clustering approach for high dimensional dataset. *International Journal of Engineering, Science and Technology*, 2(2):59–66, 2010. doi: https://doi.org/10.4314/ijest.v2i2.59139.

[81] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. ISSN 1939-3539. doi: 10.1109/TPAMI.1979.4766909.

[82] Richard H. Dee. Magnetic tape for data storage: An enduring technology. *Proceedings of the IEEE*, 96(11):1775–1785, November 2008. ISSN 1558-2256. doi: 10.1109/JPROC.2008.2004311.

[83] G. Deepika. Holographic versatile disc. In *2011 National Conference on Innovations in Emerging Technology*, pages 145–146, February 2011. doi: 10.1109/NCOIET.2011.5738819.

[84] Sebastian Deorowicz. Fqsqueezer: k-mer-based compression of sequencing data. *Scientific Reports*, 10(1):578, January 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-57452-6. URL https://doi.org/10.1038/s41598-020-57452-6.

[85] D.Napoleon and S.Pavalakodi. Article: A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, 13(8):41–46, January 2011. doi: 10.5120/1789-2471. Full text available.

[86] Tsugihiro Doi, Shinji Sakata, Tetsutaro Inoue, and Sadamu Kuse. Magnetic tape, September 2005. US Patent App. 11/052,952.

[87] Sergii Domanskyi, Carlo Piermarocchi, and George I Mias. PyIOmica: longitudinal omics analysis and trend identification. *Bioinformatics*, 36(7):2306–2307, 11 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz896. URL https://doi.org/10.1093/bioinformatics/btz896.

[88] Yiming Dong, Fajia Sun, Zhi Ping, Qi Ouyang, and Long Qian. DNA storage: research landscape and future prospects. *National Science Review*, 7(6):1092–1107, 01 2020. ISSN 2095-5138. doi: 10.1093/nsr/nwaa007. URL https://doi.org/10.1093/nsr/nwaa007.

[89] Eike Dornsiepen, Florian Dobener, Nils Mengel, Olena Lenchuk, Christof Dues, Simone Sanna, Doreen Mollenhauer, Sangam Chatterjee, and Stefanie Dehnen. White-light generation upon in-situ amorphization of single crystals of [(Me3P)3AuSn(phsn)3s6] and [(Et3P)3AgSn(phsn)3s6]. *Advanced Optical Materials*, 7(12):1801793, 2019. doi: https://doi.org/10.1002/adom.201801793. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adom.201801793.

[90] N. Dwivedi, A. K. Ott, K. Sasikumar, C. Dou, R. J. Yeo, B. Narayanan, U. Sassi, D. De Fazio, G. Soavi, T. Dutta, O. Balci, S. Shinde, J. Zhang, A. K. Katiyar, P. S. Keatley, A. K. Srivastava, S. K. R. S. Sankaranarayanan, A. C. Ferrari, and C. S. Bhatia. Graphene overcoats for ultra-high storage density magnetic media. *Nature Communications*, 12(1):2854, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22687-y. URL https://doi.org/10.1038/s41467-021-22687-y.

[91] J. Nikolaj Dybowski, Mona Riemenschneider, Sascha Hauke, Martin Pyka, Jens Verheyen, Daniel Hoffmann, and Dominik Heider. Improved bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Mining*, 4(1):26, November

2011. ISSN 1756-0381. doi: 10.1186/1756-0381-4-26. URL https://doi.org/10.1186/1756-0381-4-26.

[92] Nader Ebrahimi, Kurt Pflughoeft, and Ehsan Soofi. Two measures of sample entropy. *Statistics & Probability Letters*, 20:225–234, 06 1994. doi: 10.1016/0167-7152(94)90046-9.

[93] ECMA. *ECMA-367: Eiffel analysis, design and programming language*. ECMA (European Association for Standardizing Information and Communication Systems), Geneva, Switzerland, June 2006. URL https://www.ecma-international.org/publications-and-standards/standards/ecma-377/. 2006.

[94] Hajo Eicken, Finn Danielsen, Josephine-Mary Sam, Maryann Fidel, Noor Johnson, Michael K Poulsen, Olivia A Lee, Katie V Spellman, Lisbeth Iversen, Peter Pulsifer, and Martin Enghoff. Connecting Top-Down and Bottom-Up Approaches in Environmental Observing. *BioScience*, 71(5):467–483, 04 2021. ISSN 0006-3568. doi: 10.1093/biosci/biab018. URL https://doi.org/10.1093/biosci/biab018.

[95] S. Elvidge, M. J. Angling, and B. Nava. On the use of modified taylor diagrams to compare ionospheric assimilation models. *Radio Science*, 49(9):737–745, 2014. doi: https://doi.org/10.1002/2014RS005435. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014RS005435.

[96] John W. Emerson, Walton A. Green, Barret Schloerke, Jason Crowley, Dianne Cook, Heike Hofmann, and Hadley Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013. doi: 10.1080/10618600.2012.694762. URL https://doi.org/10.1080/10618600.2012.694762.

[97] Yaniv Erlich and Dina Zielinski. Dna fountain enables a robust and efficient storage architecture. *Science*, 355(6328):950–954, 2017. doi: 10.1126/science.aaj2038. URL https://www.science.org/doi/abs/10.1126/science.aaj2038.

[98] Matt Ernst. Digital storage. *Interface: The Journal of Education, Community and Values*, 3(1), 2003.

[99] Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 1741.

[100] I. W. Evett and E. J. Spiehler. Rule induction in forensic science. In *Knowledge Based Systems*, page 152–160, USA, 1989. Halsted Press. ISBN 0470212608.

[101] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9(61):1871–1874, 2008. URL http://jmlr.org/papers/v9/fan08a.html.

[102] Michael Farber. *Invitation to Topological Robotics (Zurich Lectures in Advanced Mathematics)*. Amer Mathematical Society, paperback edition, 2008. ISBN 978-3037190548. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=303719054X.

[103] Ben L. Feringa, Wolter F. Jager, and Ben de Lange. Organic materials for reversible optical data storage. *Tetrahedron*, 49(37):8267–8310, 1993. ISSN 0040-4020. doi: https://doi.org/10.1016/S0040-4020(01)81913-X. URL https://www.sciencedirect.com/science/article/pii/S004040200181913X.

[104] Dmitrii Filimonov, Vladimir Poroikov, Yulia Borodina, and Tatyana Gloriozova. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *Journal of Chemical Information and Computer Sciences*, 39(4): 666–670, July 1999. ISSN 0095-2338. doi: 10.1021/ci980335o. URL https://doi.org/10.1021/ci980335o.

[105] R. A. FISHER. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

[106] Erica Flapan, Adam He, and Helen Wong. Topological descriptions of protein folding. *Proceedings of the National Academy of Sciences*, 116(19):9360–9369, 2019. doi: 10.1073/pnas.1808312116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1808312116.

[107] Jr. Fontana, R. E., G. M. Decad, and S. R. Hetzler. Volumetric density trends (TB/in.3) for storage components: TAPE, hard disk drives, NAND, and Blu-ray. *Journal of Applied Physics*, 117(17), 01 2015. ISSN 0021-8979. doi: 10.1063/1.4906208. URL https://doi.org/10.1063/1.4906208. 17E301.

[108] E. Foss and R. S. Partridge. A 32,000-word magnetic-core memory. *IBM Journal of Research and Development*, 1(2):102–109, April 1957. ISSN 0018-8646. doi: 10.1147/rd.12.0102.

[109] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1504. URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[110] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

[111] Jerome H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, feb 2002. ISSN 0167-9473. doi: 10.1016/S0167-9473(01)00065-2. URL https://doi.org/10.1016/S0167-9473(01)00065-2.

[112] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6790–6802. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/35cf8659cfcb13224cbd47863a34fc58-Paper.pdf.

[113] S. Gattei. *Karl Popper's Philosophy of Science: Rationality without Foundations*. Routledge Studies in the Philosophy of Science. Taylor & Francis, 2008. ISBN 9780203887196. URL https://books.google.de/books?id=oPPu1JvMBFoC.

[114] Warren Gay. *SD Card Storage*, page 274. Apress, paperback edition, 11 2014. ISBN 978-1484208007. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1484208005.

[115] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, Apr 2006. ISSN 1573-0565. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

[116] David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, and Magnus Johnsson. Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*, 39(16):12564–12573, 2012. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2012.05.028. URL https://www.sciencedirect.com/science/article/pii/S0957417412007269.

[117] Jaume Giné. Quantum fluctuations and the slow accelerating expansion of the universe. *Europhysics Letters*, 125(5):50002, April 2019. doi: 10.1209/0295-5075/125/50002. URL https://dx.doi.org/10.1209/0295-5075/125/50002.

[118] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

[119] Jeffrey W. Godden, Ling Xue, and Jürgen Bajorath. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *Journal of Chemical Information and Computer Sciences*, 40(1):163–166, January 2000. ISSN 0095-2338. doi: 10.1021/ci990316u. URL https://doi.org/10.1021/ci990316u.

[120] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos, and Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *Nature*, 494(7435):77–80, February 2013. ISSN 1476-4687. doi: 10.1038/nature11875. URL https://doi.org/10.1038/nature11875.

[121] A. D. Gordon. Classification and regression trees. *Biometrics*, 40(3):874–874, 1984. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530946.

[122] Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, Jan 2017. ISSN 1941-0506. doi: 10.1109/TVCG.2016.2598918.

[123] August E. Grant and Jennifer Meadows, editors. *Communication Technology Update and Fundamentals: 17th Edition*. Routledge, hardcover edition, 6 2020. ISBN 978-0367420130. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0367420139.

[124] Robert N. Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J. Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015. doi: https://doi.org/10.1002/anie.201411378. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201411378.

[125] Jonathan E. Green, Jang Wook Choi, Akram Boukai, Yuri Bunimovich, Ezekiel Johnston-Halperin, Erica DeIonno, Yi Luo, Bonnie A. Sheriff, Ke Xu, Young Shik Shin, Hsian-Rong Tseng, J. Fraser Stoddart, and James R. Heath. A 160-kilobit molecular electronic memory patterned at 1011 bits per square centimetre. *Nature*, 445(7126):414–417, January 2007. ISSN 1476-4687. doi: 10.1038/nature05462. URL https://doi.org/10.1038/nature05462.

[126] Christian A. Gueymard. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. *Renewable and Sustainable Energy Reviews*, 39:1024–1034, 2014. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2014.07.117. URL https://www.sciencedirect.com/science/article/pii/S1364032114005693.

[127] Chongomweru Halimu, Asem Kasem, and S. H. Shah Newaz. Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, ICMLSC 2019, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366120. doi: 10.1145/3310986.3311023. URL https://doi.org/10.1145/3310986.3311023.

[128] Kyle Wm. Hall, Adam J. Bradley, Uta Hinrichs, Samuel Huron, Jo Wood, Christopher Collins, and Sheelagh Carpendale. Design by immersion: A transdisciplinary approach to problem-driven visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):109–118, January 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934790.

[129] Min Hao, Hongyan Qiao, Yanmin Gao, Zhaoguan Wang, Xin Qiao, Xin Chen, and Hao Qi. A mixed culture of bacterial cells enables an economic dna storage on a large scale. *Communications Biology*, 3(1):416, July 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-01141-7. URL https://doi.org/10.1038/s42003-020-01141-7.

[130] Steve Haroz and David Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18 (12):2402–2410, Dec 2012. ISSN 1941-0506. doi: 10.1109/TVCG.2012.233.

[131] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E.

Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[132] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Ensemble Learning*, chapter Ensemble Learning, pages 605–624. Springer New York, New York, NY, 2009. ISBN 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7_16. URL https://doi.org/10.1007/978-0-387-84858-7_16.

[133] Georges Hattab, Dror E. Warschawski, Karine Moncoq, and Bruno Miroux. Escherichia coli as host for membrane protein structure determination: a global analysis. *Scientific Reports*, 5 (1):12097, July 2015. ISSN 2045-2322. doi: 10.1038/srep12097. URL https://doi.org/10.1038/srep12097.

[134] Georges Hattab, Theresa-Marie Rhyne, and Dominik Heider. Ten simple rules to colorize biological data visualization. *PLOS Computational Biology*, 16(10):1–18, 10 2020. doi: 10.1371/journal.pcbi.1008259. URL https://doi.org/10.1371/journal.pcbi.1008259.

[135] Georges Hattab, Theresa-Marie Rhyne, and Dominik Heider. Correction: Ten simple rules to colorize biological data visualization. *PLOS Computational Biology*, 17(4):1–1, 04 2021. doi: 10.1371/journal.pcbi.1008901. URL https://doi.org/10.1371/journal.pcbi.1008901.

[136] Georges Hattab, Aleksandar Anžel, Sebastian Spänig, Nils Neumann, and Dominik Heider. A parametric approach for molecular encodings using multilevel atomic neighborhoods applied to peptide classification. *NAR Genomics and Bioinformatics*, 5(1), 01 2023. ISSN 2631-9268. doi: 10.1093/nargab/lqac103. URL https://doi.org/10.1093/nargab/lqac103. lqac103.

[137] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015 IEEE International Conference on Computer Vision (ICCV), Dec 2015. doi: 10.1109/ICCV.2015.123.

[138] Dominik Heider, Jan Nikolaj Dybowski, Christoph Wilms, and Daniel Hoffmann. A simple structure-based model for the prediction of hiv-1 co-receptor tropism. *BioData mining*, 7: 14, 2014. ISSN 1756-0381. doi: 10.1186/1756-0381-7-14. URL https://europepmc.org/articles/PMC4124776.

[139] Leah Henderson. The Problem of Induction. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

[140] James B. Hendrickson, Ping Huang, and A. Glenn Toczko. Molecular complexity: a simplified formula adapted to individual atoms. *Journal of Chemical Information and Computer Sciences*, 27(2):63–67, May 1987. ISSN 0095-2338. doi: 10.1021/ci00054a004. URL https://doi.org/10.1021/ci00054a004.

[141] Malte Herold, Susana Martínez Arbas, Shaman Narayanasamy, Abdul R. Sheik, Luise A. K. Kleine-Borgmann, Laura A. Lebrun, Benoît J. Kunath, Hugo Roume, Irina Bessarab, Rohan B. H. Williams, John D. Gillece, James M. Schupp, Paul S. Keim, Christian Jäger, Michael R. Hoopmann, Robert L. Moritz, Yuzhen Ye, Sujun Li, Haixu Tang, Anna Heintz-Buschart, Patrick May, Emilie E. L. Muller, Cedric C. Laczny, and Paul Wilmes. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nature Communications*, 11(1), October 2020. doi: 10.1038/s41467-020-19006-2. URL https://doi.org/10.1038/s41467-020-19006-2.

[142] Geoffrey E. Hinton. Connectionist learning procedures. *Artificial Intelligence*, 40(1):185–234, 1989. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(89)90049-0. URL https://www.sciencedirect.com/science/article/pii/0004370289900490.

[143] Albert S. Hoagland. Magnetic drum recording of digital data. *Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics*, 73(4):381–385, September 1954. ISSN 2379-674X. doi: 10.1109/TCE.1954.6372170.

[144] Georg Hoffmann, Andreas Bietenbeck, Ralf Lichtinghagen, and Frank Klawonn. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *Journal of Laboratory and Precision Medicine*, 3:58–58, 06 2018. doi: 10.21037/jlpm.2018.06.01.

[145] A. Hoogendoorn. Digital compact cassette. *Proceedings of the IEEE*, 82(10):1479–1489, October 1994. ISSN 1558-2256. doi: 10.1109/5.326405.

[146] Hideyoshi Horimai and Yoshio Aoki. Holographic versatile disc (hvd). In *International Symposium on Optical Memory and Optical Data Storage*, page ThE6. Optica Publishing Group, 2005. doi: 10.1364/ISOM_ODS.2005.ThE6. URL https://opg.optica.org/abstract.cfm?URI=ISOM_ODS-2005-ThE6.

[147] Hideyoshi Horimai and Xiaodi Tan. Holographic information storage system: Today and future. *IEEE Transactions on Magnetics*, 43(2):943–947, February 2007. ISSN 1941-0069. doi: 10.1109/TMAG.2006.888528.

[148] Paul Horton and Kenta Nakai. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, page 109–115. AAAI Press, 1996. ISBN 1577350022.

[149] Benlin Hu, Chengyuan Wang, Jiangxin Wang, Junkuo Gao, Kai Wang, Jiansheng Wu, Guodong Zhang, Wangqiao Cheng, Bhavanasi Venkateswarlu, Mingfeng Wang, Pooi See Lee, and Qichun Zhang. Inorganic–organic hybrid polymer with multiple redox for high-density data storage. *Chem. Sci.*, 5:3404–3408, 2014. doi: 10.1039/C4SC00823E. URL http://dx.doi.org/10.1039/C4SC00823E.

[150] Yi Huang, Timmy Kendall, Evan S. Forsythe, Ana Dorantes-Acosta, Shaofang Li, Juan Caballero-Pérez, Xuemei Chen, Mario Arteaga-Vázquez, Mark A. Beilstein, and Rebecca A. Mosher. Ancient Origin and Recent Innovations of RNA Polymerase IV and

V. *Molecular Biology and Evolution*, 32(7):1788–1799, 03 2015. ISSN 0737-4038. doi: 10.1093/molbev/msv060. URL https://doi.org/10.1093/molbev/msv060.

[151] D. Hume. *A Treatise of Human Nature*. Dover philosophical classics. Dover Publications, 2003. ISBN 9780486432502. URL https://books.google.de/books?id=zHYOlFh9_JMC.

[152] Catherine Hurley and R. Oldford. Pairwise display of high-dimensional information via eulerian tours and hamiltonian decompositions. *Journal of Computational and Graphical Statistics*, 19, 12 2010. doi: 10.1198/jcgs.2010.09136.

[153] Harry D. Huskey. *Williams Tube Memory*, page 1851–1853. John Wiley and Sons Ltd., GBR, 2003. ISBN 0470864125.

[154] IBM. IBM 350 disk storage unit, 2021. https://www.ibm.com/ibm/history/exhibits/storage/storage_350.html.

[155] Plotly Technologies Inc. Collaborative data science, 2015. URL https://plot.ly.

[156] S. Ingram, M. Olano, and T. Munzner. Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization & Computer Graphics*, 15(02):249–261, March 2009. ISSN 1941-0506. doi: 10.1109/TVCG.2008.85.

[157] Stephen Ingram and Tamara Munzner. Dimensionality reduction for documents with nearest neighbor queries. *Neurocomputing*, 150:557–569, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.07.073. URL https://www.sciencedirect.com/science/article/pii/S0925231214012910. Special Issue on Information Processing and Machine Learning for Applications of Engineering Solving Complex Machine Learning Problems with Ensemble Methods Visual Analytics using Multidimensional Projections.

[158] Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 3–10, October 2010. doi: 10.1109/VAST.2010.5652392.

[159] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, Aug 1985. ISSN 1432-2315. doi: 10.1007/BF01898350. URL https://doi.org/10.1007/BF01898350.

[160] Alfred Inselberg and Bernard Dimsdale. *Parallel Coordinates*, pages 199–233. Springer US, Boston, MA, 1991. ISBN 978-1-4684-5883-1. doi: 10.1007/978-1-4684-5883-1_9. URL https://doi.org/10.1007/978-1-4684-5883-1_9.

[161] Lemon J. Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12, 2006.

[162] Vellingiri Jayagopal and KK Basser. Data management and big data analytics: Data management in digital economy. In *Optimizing Big Data Management and Industrial Systems With Intelligent Techniques*, pages 1–23. IGI Global, 2019. doi: 10.4018/978-1-6684-3662-2.ch078.

[163] Edwin T Jaynes. Information theory and statistical mechanics (notes by the lecturer). *Statistical physics 3*, page 181, 1963.

[164] David R. Johnson, Tae Kwon Lee, Joonhong Park, Kathrin Fenner, and Damian E. Helbling. The functional and taxonomic richness of wastewater treatment plant microbial communities are associated with each other and with ambient nitrogen and carbon availability. *Environmental Microbiology*, 17(12):4851–4860, 2015. doi: https://doi.org/10.1111/1462-2920.12429. URL https://sfamjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1462-2920.12429.

[165] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007. doi: 10.1126/science.1141319. URL https://www.science.org/doi/abs/10.1126/science.1141319.

[166] Mark A. Johnson and Gerald M. Maggiora, editors. *Concepts and Applications of Molecular Similarity*. Wiley-Interscience, hardcover edition, 9 1990. ISBN 978-0471621751. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0471621757.

[167] Jason K. Jolliff, John C. Kindle, Igor Shulman, Bradley Penta, Marjorie A.M. Friedrichs, Robert Helber, and Robert A. Arnone. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, 76(1):64–82, 2009. ISSN 0924-7963. doi: https://doi.org/10.1016/j.jmarsys.2008.05.014. URL https://www.sciencedirect.com/science/article/pii/S0924796308001140. Skill assessment for coupled biological/physical models of marine systems.

[168] Sebastian Kalcher and Volker Lindenstruth. Accelerating galois field arithmetic for reed-solomon erasure codes in storage applications. In *2011 IEEE International Conference on Cluster Computing*, pages 290–298, September 2011. doi: 10.1109/CLUSTER.2011.40.

[169] F. E. Kalff, M. P. Rebergen, E. Fahrenfort, J. Girovsky, R. Toskovic, J. L. Lado, J. Fernández-Rossier, and A. F. Otte. A kilobyte rewritable atomic memory. *Nature Nanotechnology*, 11(11):926–929, November 2016. ISSN 1748-3395. doi: 10.1038/nnano.2016.131. URL https://doi.org/10.1038/nnano.2016.131.

[170] Vamsee Kasavajhala. Solid state drive vs. hard disk drive price and performance study. *Proc. Dell Tech. White Paper*, pages 8–9, 2011.

[171] Gizem Kaya, Chisom Ezekannagha, Dominik Heider, and Georges Hattab. Context-aware phylogenetic trees for phylogeny-based taxonomy visualization. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.891240. URL https://www.frontiersin.org/articles/10.3389/fgene.2022.891240.

[172] Peter Kazansky, Ausra Cerkauskaite, Martynas Beresna, Rokas Drevinskas, Aabid Patel, Jingyu Zhang, and Mindaugas Gecevicius. Eternal 5d data storage via ultrafast-laser writing in glass. *SPIE Optoelectronics & Communications*, 2016.

[173] Richard V Keele, Craig D Mautner, Tracy J Thorpe, Sidney R Thompson, Michael C Goodsell, and Philip J Erdelsky. Virtual addressing of optical storage media as magnetic tape equivalents, October 1995. US Patent 5,455,926.

[174] D.A. Keim, H.-P. Kriegel, and M. Ankerst. Recursive pattern: a technique for visualizing very large amounts of data. In *Proceedings Visualization '95*, pages 279–286, October 1995. doi: 10.1109/VISUAL.1995.485140.

[175] D.A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Tenth International Conference on Information Visualisation (IV'06)*, pages 9–16, July 2006. doi: 10.1109/IV.2006.31.

[176] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(96)00140-X. URL https://www.sciencedirect.com/science/article/pii/S016771529600140X.

[177] Eamonn Kennedy, Christopher E. Arcadia, Joseph Geiser, Peter M. Weber, Christopher Rose, Brenda M. Rubenstein, and Jacob K. Rosenstein. Encoding information in synthetic metabolomes. *PLOS ONE*, 14(7):1–12, 07 2019. doi: 10.1371/journal.pone.0217364. URL https://doi.org/10.1371/journal.pone.0217364.

[178] John R Kettman, Johann R Frey, and Ivan Lefkovits. Proteome, transcriptome and genome: top down or bottom up analysis? *Biomolecular Engineering*, 18(5):207–212, 2001. ISSN 1389-0344. doi: https://doi.org/10.1016/S1389-0344(01)00096-X. URL https://www.sciencedirect.com/science/article/pii/S138903440100096X.

[179] A.Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1):2–30, 2003. ISSN 0003-4916. doi: https://doi.org/10.1016/S0003-4916(02)00018-0. URL https://www.sciencedirect.com/science/article/pii/S0003491602000180.

[180] Dean Klein. The history of semiconductor memory: From magnetic tape to nand flash memory. *IEEE Solid-State Circuits Magazine*, 8(2):16–22, 3 2016. ISSN 1943-0590. doi: 10.1109/MSSC.2016.2548422.

[181] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

[182] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 444–452, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401946. URL https://doi.org/10.1145/1401890.1401946.

[183] Petri Laarne, Martha A. Zaidan, and Tuomo Nieminen. ennemi: Non-linear correlation detection with mutual information. *SoftwareX*, 14:100686, 2021. ISSN 2352-7110. doi:

https://doi.org/10.1016/j.softx.2021.100686. URL https://www.sciencedirect.com/science/article/pii/S2352711021000315.

[184] De-wei Lai, Jen-hung Liao, and Hsiao-te Chang. Methods for measuring usable lifespan and replacing an in-system programming code of a memory device, and data storage system using the same, March 2013. US Patent 8,402,204.

[185] Yann Lecun, L.D. Jackel, Leon Bottou, Corinna Cortes, J. S. Denker, Harris Drucker, I. Guyon, U.A. Muller, Eduard Sackinger, Patrice Simard, and V. Vapnik. *Learning algorithms for classification: A comparison on handwritten digit recognition*, pages 261–276. World Scientific, 1995.

[186] Paola Leon-Mimila, Jessica Wang, and Adriana Huertas-Vazquez. Relevance of multi-omics studies in cardiovascular diseases. *Frontiers in Cardiovascular Medicine*, 6, 2019. ISSN 2297-055X. doi: 10.3389/fcvm.2019.00091. URL https://www.frontiersin.org/articles/10.3389/fcvm.2019.00091.

[187] Adam Leventhal. Flash storage memory. *Commun. ACM*, 51(7):47–51, July 2008. ISSN 0001-0782. doi: 10.1145/1364782.1364796. URL https://doi.org/10.1145/1364782.1364796.

[188] Charles X. Ling, Jin Huang, and Harry Zhang. Auc: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, page 519–524, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[189] Johann B. Listing and Esther von Krosigk. *Über unsere jetzige Kenntnis der Gestalt und Größe der Erde: Enth.: Neue geometrische und dynamische Constanten des Erdkörpers (German Edition)*. VDM Verlag Dr. Müller, paperback edition, 9 2007. ISBN 978-3836422994. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=3836422999.

[190] Johann Benedikt Listing. *Vorstudien zur topologie*. Vandenhoeck und Ruprecht, 1848.

[191] Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423, May 2020. doi: 10.1109/ICASSP40776.2020.9054458.

[192] Tang Liu, Shufeng Liu, Maosheng Zheng, Qian Chen, and Jinren Ni. Performance assessment of full-scale wastewater treatment plants based on seasonal variability of microbial communities via high-throughput sequencing. *PLOS ONE*, 11(4):1–15, 04 2016. doi: 10.1371/journal.pone.0152998. URL https://doi.org/10.1371/journal.pone.0152998.

[193] Zipeng Liu, Shing Hei Zhan, and Tamara Munzner. Aggregated dendrograms for visual comparison between many phylogenetic trees. *IEEE Transactions on Visualization and Computer Graphics*, 26(9):2732–2747, September 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2898186.

[194] Gianluigi Liva, Enrico Paolini, and Marco Chiani. Performance versus overhead for fountain codes over fq. *IEEE Communications Letters*, 14(2):178–180, February 2010. ISSN 1558-2558. doi: 10.1109/LCOMM.2010.02.092080.

[195] G.C.P. Lokhoff. Dcc-digital compact cassette. *IEEE Transactions on Consumer Electronics*, 37 (3):702–706, August 1991. ISSN 1558-4127. doi: 10.1109/30.85589.

[196] Liang Fu Lu, Mao Lin Huang, and Tze-Haw Huang. A new axes re-ordering method in parallel coordinates visualization. *2012 11th International Conference on Machine Learning and Applications*, 2:252–257, Dec 2012. doi: 10.1109/ICMLA.2012.148.

[197] Barry M. Lunt, Matthew R. Linford, Robert C. Davis, Sarah Jamieson, Anthony Pearson, and Hao Wang. Toward permanence in digital data storage. *Archiving Conference*, 10(1): 132–132, 2013. doi: 10.2352/issn.2168-3204.2013.10.1.art00029. URL https://library.imaging.org/archiving/articles/10/1/art00029.

[198] L.W. MacDonald. Using color effectively in computer graphics. *IEEE Computer Graphics and Applications*, 19(4):20–35, July 1999. ISSN 1558-1756. doi: 10.1109/38.773961.

[199] Jan Maes and Marc Vercammen, editors. *Digital Audio Technology: A Guide to CD, MiniDisc, SACD, DVD(A), MP3 and DAT*. Routledge, paperback edition, 9 2001. ISBN 978-0240516547. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0240516540.

[200] W. A. Malthaner and H. E. Vaughan. An automatic telephone system employing magnetic drum memory. *Proceedings of the IRE*, 41(10):1341–1347, October 1953. ISSN 2162-6634. doi: 10.1109/JRPROC.1953.274309.

[201] G. Elisabeta Marai. Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):913–922, January 2018. ISSN 1941-0506. doi: 10.1109/TVCG.2017.2744459.

[202] Giulia Massini, Stefano Terzi, and Massimo Buscema. *Population Algorithm: A New Method of Multi-Dimensional Scaling*, pages 63–74. Springer New York, New York, NY, 2013. ISBN 978-1-4614-4223-3. doi: 10.1007/978-1-4614-4223-3_3. URL https://doi.org/10.1007/978-1-4614-4223-3_3.

[203] Justin Matejka and George Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 1290–1294, 2017. doi: 10.1145/3025453.3025912. URL https://doi.org/10.1145/3025453.3025912.

[204] Takashi Matsumoto and Maribeth Back. Post-bit: multimedia epaper stickies, March 2007. US Patent 7,195,170.

[205] Takashi Matsumoto, Tony Dunnigan, and Maribeth Back. Post-bit: Embodied video contents on tiny stickies. In *Proceedings of the 13th Annual ACM International Conference*

on Multimedia, MULTIMEDIA '05, page 263–264, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930442. doi: 10.1145/1101149.1101197. URL https://doi.org/10.1145/1101149.1101197.

[206] B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

[207] Guillaume MAZE. Taylor diagram, 2023. URL https://www.mathworks.com/matlabcentral/fileexchange/20559-taylor-diagram.

[208] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[209] Gerald A. Meehl, Curt Covey, Thomas Delworth, Mojib Latif, Bryant McAvaney, John F. B. Mitchell, Ronald J. Stouffer, and Karl E. Taylor. The wcrp cmip3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, 88(9):1383 – 1394, 2007. doi: 10.1175/BAMS-88-9-1383. URL https://journals.ametsoc.org/view/journals/bams/88/9/bams-88-9-1383.xml.

[210] Marina Meilă. Comparing clusterings by the variation of information. *Learning Theory and Kernel Machines*, pages 173–187, 2003. doi: https://doi.org/10.1007/978-3-540-45167-9_14.

[211] Marina Meilă. Comparing clusterings: An axiomatic view. *Proceedings of the 22nd International Conference on Machine Learning*, page 577–584, 2005. doi: 10.1145/1102351.1102424. URL https://doi.org/10.1145/1102351.1102424.

[212] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2006.11.013. URL https://www.sciencedirect.com/science/article/pii/S0047259X06002016.

[213] Tauno Metsalu and Jaak Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1): W566–W570, 05 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv468. URL https://doi.org/10.1093/nar/gkv468.

[214] Miriah Meyer, Michael Sedlmair, and Tamara Munzner. The four-level nested model revisited: Blocks and guidelines. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450317917. doi: 10.1145/2442576.2442587. URL https://doi.org/10.1145/2442576.2442587.

[215] Rino Micheloni, Massimiliano Picca, Stefano Amato, Helmut Schwalm, Michael Scheppler, and Stefano Commodaro. Non-volatile memories for removable media. *Proceedings of the IEEE*, 97(1):148–160, January 2009. ISSN 1558-2256. doi: 10.1109/JPROC.2008.2007477.

[216] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[217] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[218] Thomas Minka. Automatic choice of dimensionality for pca. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf.

[219] Olga Mordvinova, Julian Martin Kunkel, Christian Baun, Thomas Ludwig, and Marcel Kunze. Usb flash drives as an energy efficient storage alternative. In *2009 10th IEEE/ACM International Conference on Grid Computing*, pages 175–182, October 2009. doi: 10.1109/GRID.2009.5353062.

[220] E MORSE, M LEWIS, and K.A OLSEN. Evaluating visualizations: using a taxonomic guide. *International Journal of Human-Computer Studies*, 53(5):637–662, 2000. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.2000.0412. URL https://www.sciencedirect.com/science/article/pii/S1071581900904129.

[221] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov 2009. ISSN 1941-0506. doi: 10.1109/TVCG.2009.111.

[222] Tamara Munzner. *Visualization Analysis and Design (AK Peters Visualization Series)*. A K Peters/CRC Press, hardcover edition, 12 2014. ISBN 978-1466508910. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1466508914.

[223] Tamara Munzner, François Guimbretière, Serdar Tasiran, Li Zhang, and Yunhong Zhou. Treejuxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. In *ACM SIGGRAPH 2003 Papers*, SIGGRAPH '03, page 453–462, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137095. doi: 10.1145/1201775.882291. URL https://doi.org/10.1145/1201775.882291.

[224] Naveen Naidu Narisetty. Chapter 4 - bayesian model selection for high-dimensional data. In Arni S.R. Srinivasa Rao and C.R. Rao, editors, *Principles and Methods for Data Science*, volume 43 of *Handbook of Statistics*, pages 207–248. Elsevier, 2020. doi: https://doi.org/10.

1016/bs.host.2019.08.001. URL https://www.sciencedirect.com/science/article/pii/S0169716119300380.

[225] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). *Technical Report CUCS-005-96*, 1996. URL https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[226] Bruno J. Neves, Rodolpho C. Braga, Cleber C. Melo-Filho, José Teófilo Moreira-Filho, Eugene N. Muratov, and Carolina Horta Andrade. Qsar-based virtual screening: Advances and applications in drug discovery. *Frontiers in Pharmacology*, 9, 2018. ISSN 1663-9812. doi: 10.3389/fphar.2018.01275. URL https://www.frontiersin.org/articles/10.3389/fphar.2018.01275.

[227] Giovanna Nicora, Francesca Vitali, Arianna Dagliati, Nophar Geifman, and Riccardo Bellazzi. Integrated multi-omics analyses in oncology: A review of machine learning methods and tools. *Frontiers in Oncology*, 10, 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.01030. URL https://www.frontiersin.org/articles/10.3389/fonc.2020.01030.

[228] Andreas Noack. An energy model for visual graph clustering. In Giuseppe Liotta, editor, *Graph Drawing*, pages 425–436, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-24595-7.

[229] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

[230] Travis E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9 (3):10–20, May 2007. ISSN 1558-366X. doi: 10.1109/MCSE.2007.58.

[231] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z. Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N. Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss. Random access in large-scale dna data storage. *Nature Biotechnology*, 36(3):242–248, March 2018. ISSN 1546-1696. doi: 10.1038/nbt.4079. URL https://doi.org/10.1038/nbt.4079.

[232] Ludovic Orlando, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, Enrico Cappellini, Bent Petersen, Ida Moltke, Philip L. F. Johnson, Matteo Fumagalli, Julia T. Vilstrup, Maanasa Raghavan, Thorfinn Korneliussen, Anna-Sapfo Malaspinas, Josef Vogt, Damian Szklarczyk, Christian D. Kelstrup, Jakob Vinther, Andrei Dolocan, Jesper Stenderup, Amhed M. V. Velazquez, James Cahill, Morten Rasmussen, Xiaoli Wang, Jiumeng Min, Grant D. Zazula, Andaine Seguin-Orlando, Cecilie Mortensen, Kim Magnussen, John F. Thompson, Jacobo Weinstock, Kristian Gregersen, Knut H. Røed, Véra Eisenmann, Carl J. Rubin, Donald C. Miller, Douglas F. Antczak,

Mads F. Bertelsen, Søren Brunak, Khaled A. S. Al-Rasheid, Oliver Ryder, Leif Andersson, John Mundy, Anders Krogh, M. Thomas P. Gilbert, Kurt Kjær, Thomas Sicheritz-Ponten, Lars Juhl Jensen, Jesper V. Olsen, Michael Hofreiter, Rasmus Nielsen, Beth Shapiro, Jun Wang, and Eske Willerslev. Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. *Nature*, 499(7456):74–78, July 2013. ISSN 1476-4687. doi: 10.1038/nature12323. URL https://doi.org/10.1038/nature12323.

[233] Deepak P. and Prasad M. Deshpande. *Operators for Similarity Search: Semantics, Techniques and Usage Scenarios*. Springer Publishing Company, Incorporated, 1st edition, 2015. ISBN 3319212567.

[234] Shraddha Pai and Gary D. Bader. Patient similarity networks for precision medicine. *Journal of Molecular Biology*, 430(18, Part A):2924–2938, 2018. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2018.05.037. URL https://www.sciencedirect.com/science/article/pii/S0022283618305321. Theory and Application of Network Biology Toward Precision Medicine.

[235] The pandas development team. pandas-dev/pandas: Pandas, November 2022. URL https://doi.org/10.5281/zenodo.7344967.

[236] William N. Papian. The mit magnetic-core memory. In *Papers and Discussions Presented at the Dec. 8-10, 1953, Eastern Joint AIEE-IRE Computer Conference: Information Processing Systems—Reliability and Requirements*, AIEE-IRE '53 (Eastern), page 37–42, New York, NY, USA, 1953. Association for Computing Machinery. ISBN 9781450378536. doi: 10.1145/1434878.1434888. URL https://doi.org/10.1145/1434878.1434888.

[237] E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi. Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, 83:87–96, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.06.001. URL https://www.sciencedirect.com/science/article/pii/S1532046418301072.

[238] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2021.100336. URL https://www.sciencedirect.com/science/article/pii/S2666389921001847.

[239] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[240] Wei Peng, M.O. Ward, and E.A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. *IEEE Symposium on Information Visualization*, pages 89–96, Oct 2004. ISSN 1522-404X. doi: 10.1109/INFVIS.2004.15.

[241] Beatriz Merchel Piovesan Pereira, Xiaokang Wang, Ilias Tagkopoulos, and Maia Kivisaar. Short- and long-term transcriptomic responses of escherichia coli to biocides: a systems analysis. *Applied and Environmental Microbiology*, 86(14):e00708–20, 2020. doi: 10.1128/AEM.00708-20. URL https://journals.asm.org/doi/abs/10.1128/AEM.00708-20.

[242] Carolyn L. Phillips, Eric Jankowski, Bhaskar Jyoti Krishnatreya, Kazem V. Edmond, Stefano Sacanna, David G. Grier, David J. Pine, and Sharon C. Glotzer. Digital colloids: reconfigurable clusters as high information density elements. *Soft Matter*, 10:7468–7479, 2014. doi: 10.1039/C4SM00796D. URL http://dx.doi.org/10.1039/C4SM00796D.

[243] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, FAST '07, page 2, USA, 2007. USENIX Association.

[244] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, page 109–116, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138679. doi: 10.1145/989863.989880. URL https://doi.org/10.1145/989863.989880.

[245] Emre O. Polat, Hasan Burkay Uzlu, Osman Balci, Nurbek Kakenov, Evgeniya Kovalska, and Coskun Kocabas. Graphene-enabled optoelectronics on paper. *ACS Photonics*, 3(6):964–971, June 2016. doi: 10.1021/acsphotonics.6b00017. URL https://doi.org/10.1021/acsphotonics.6b00017.

[246] Ignacio Ponzoni, Víctor Sebastián-Pérez, María J. Martínez, Carlos Roca, Carlos De la Cruz Pérez, Fiorella Cravero, Gustavo E. Vazquez, Juan A. Páez, Mónica F. Díaz, and Nuria E. Campillo. Qsar classification models for predicting the activity of inhibitors of beta-secretase (bace1) associated with alzheimer's disease. *Scientific Reports*, 9(1):9102, June 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-45522-3. URL https://doi.org/10.1038/s41598-019-45522-3.

[247] William F. Porto, Állan S. Pires, and Octavio L. Franco. Cs-amppred: An updated svm model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLOS ONE*, 7(12):1–7, 12 2012. doi: 10.1371/journal.pone.0051444. URL https://doi.org/10.1371/journal.pone.0051444.

[248] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Hélène Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence

of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, January 2014. ISSN 1476-4687. doi: 10.1038/nature12886. URL https://doi.org/10.1038/nature12886.

[249] Lu Qiang and Chai Bingjie. Storycake: A hierarchical plot visualization method for storytelling in polar coordinates. In *2016 International Conference on Cyberworlds (CW)*, pages 211–218, Sep. 2016. doi: 10.1109/CW.2016.43.

[250] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214.

[251] Wolfgang Rankl and Wolfgang Effing. *Smart Card Handbook*. Wiley, hardcover edition, 7 2010. ISBN 978-0470743676. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0470743670.

[252] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL https://doi.org/10.7551/mitpress/3206.001.0001.

[253] C.D.C. Reeve. *Metaphysics*. The New Hackett Aristotle. Hackett Publishing Company, Incorporated, 2016. ISBN 9781624664410. URL https://books.google.de/books?id=prGgCwAAQBAJ.

[254] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. doi: 10.13140/2.1.2393.1847.

[255] Thomas Reichler and Junsu Kim. How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89(3):303 – 312, 2008. doi: https://doi.org/10.1175/BAMS-89-3-303. URL https://journals.ametsoc.org/view/journals/bams/89/3/bams-89-3-303.xml.

[256] Günter Reiss and Andreas Hütten. Applications beyond data storage. *Nature Materials*, 4(10):725–726, October 2005. ISSN 1476-4660. doi: 10.1038/nmat1494. URL https://doi.org/10.1038/nmat1494.

[257] B. Renoust, G. Melançon, and T. Munzner. Detangler: Visual analytics for multiplex networks. *Computer Graphics Forum*, 34(3):321–330, 2015. doi: https://doi.org/10.1111/cgf.12644. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12644.

[258] Sereina Riniker and Gregory A. Landrum. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5(1):43, September 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-43. URL https://doi.org/10.1186/1758-2946-5-43.

[259] Wagner Robert. Phonographic recording, February 1962. US Patent 3,023,011.

[260] Peter A. Rochford. Skillmetrics: A python package for calculating the skill of model predictions against observations, 2016. URL http://github.com/PeterRochford/SkillMetrics.

[261] Peter A. Rochford. Skillmetrics: A matlab package for calculating the skill of model predictions against observations, 2016. URL https://github.com/PeterRochford/SkillMetricsToolbox.

[262] Nils W. Rosemann, Jens P. Eußner, Eike Dornsiepen, Sangam Chatterjee, and Stefanie Dehnen. Organotetrel chalcogenide clusters: Between strong second-harmonic and white-light continuum generation. *Journal of the American Chemical Society*, 138(50): 16224–16227, December 2016. ISSN 0002-7863. doi: 10.1021/jacs.6b10738. URL https://doi.org/10.1021/jacs.6b10738.

[263] Nils W. Rosemann, Jens P. Eußner, Andreas Beyer, Stephan W. Koch, Kerstin Volz, Stefanie Dehnen, and Sangam Chatterjee. A highly efficient directional molecular white-light emitter driven by a continuous-wave laser diode. *Science*, 352(6291):1301–1304, 2016. doi: 10.1126/science.aaf6138. URL https://www.science.org/doi/abs/10.1126/science.aaf6138.

[264] Max Roser and Lucas Rodés-Guirao. Future population growth. *Our World in Data*, 2013. https://ourworldindata.org/future-population-growth.

[265] Brian C. Ross. Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):1–5, 02 2014. doi: 10.1371/journal.pone.0087357. URL https://doi.org/10.1371/journal.pone.0087357.

[266] Guido van Rossum. *Python Tutorial: Release 3.6.6rc1*. CreateSpace Independent Publishing Platform, paperback edition, 6 2018. ISBN 978-1721242160. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=1721242163.

[267] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[268] Peter Rowlett. 'the unplanned impact of mathematics' and its implications for research funding: a discussion-led educational activity. *BSHM Bulletin: Journal of the British Society for the History of Mathematics*, 30(1):67–74, 2015. doi: 10.1080/17498430.2014.945136. URL https://doi.org/10.1080/17498430.2014.945136.

[269] Ryo Sakai and Jan Aerts. Card Sorting Techniques for Domain Characterization in Problem-driven Visualization Research. In E. Bertini, J. Kennedy, and E. Puppo, editors, *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association, 2015. doi: 10.2312/eurovisshort.20151136.

[270] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, January 2017. ISSN 1941-0506. doi: 10.1109/TVCG.2016.2599030.

[271] J.P. Scheible. A survey of storage options. *Computer*, 35(12):42–46, December 2002. ISSN 1558-0814. doi: 10.1109/MC.2002.1106178.

[272] Bianca Schroeder, Raghav Lagisetty, and Arif Merchant. Flash reliability in production: The expected and the unexpected. In *Proceedings of the 14th Usenix Conference on File and Storage Technologies*, FAST'16, page 67–80, USA, 2016. USENIX Association. ISBN 9781931971287.

[273] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, December 2012. ISSN 1941-0506. doi: 10.1109/TVCG.2012.213.

[274] Ana Marta Sequeira, Diana Lousa, and Miguel Rocha. Propythia: A python package for protein classification based on machine and deep learning. *Neurocomputing*, 484:172–182, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.07.102. URL https://www.sciencedirect.com/science/article/pii/S0925231221016568.

[275] Shalabh. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*, volume 184. Journal of the Royal Statistical Society Series A: Statistics in Society, 04 2021. doi: 10.1111/rssa.12692. URL https://doi.org/10.1111/rssa.12692.

[276] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. *Pegasos: Primal Estimated Sub-GrAdient SOlver for SVM*. ICML '07. Association for Computing Machinery, New York, NY, USA, 2007. ISBN 9781595937933. doi: 10.1145/1273496.1273598. URL https://doi.org/10.1145/1273496.1273598.

[277] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[278] Noriyoshi Shida, Takanobu Higuchi, Yasuo Hosoda, Hiroko Miyoshi, Akio Nakano, and Katsunori Tsuchiya. Multilayer optical read-only-memory disk applicable to blu-ray disc standard using a photopolymer sheet with a recording capacity of 100 gb. *Japanese Journal of Applied Physics*, 43(7S):4983, July 2004. doi: 10.1143/JJAP.43.4983. URL https://dx.doi.org/10.1143/JJAP.43.4983.

[279] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, September 1996. doi: 10.1109/VL.1996.545307.

[280] Ben Shneiderman and Catherine Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the*

2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization, BELIV '06, page 1–7, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595935622. doi: 10.1145/1168149.1168158. URL https://doi.org/10.1145/1168149.1168158.

[281] Zvi Shtein and Oded Shoseyov. When bottom-up meets top-down. *Proceedings of the National Academy of Sciences*, 114(3):428–429, 2017. doi: 10.1073/pnas.1619392114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1619392114.

[282] Milton Silva, Diogo Pratas, and Armando J Pinho. Efficient DNA sequence compression with neural networks. *GigaScience*, 9(11), 11 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa119. URL https://doi.org/10.1093/gigascience/giaa119. giaa119.

[283] Samuel Silva, Beatriz Sousa Santos, and Joaquim Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2010.11.015. URL https://www.sciencedirect.com/science/article/pii/S0097849310001846. Virtual Reality in Brazil Visual Computing in Biology and Medicine Semantic 3D media and content Cultural Heritage.

[284] Guri Skedsmo and Stephan Gerhard Huber. Top-down and bottom-up approaches to improve educational quality: their intended and unintended consequences. *Educational Assessment, Evaluation and Accountability*, 31(1):1–4, Feb 2019. ISSN 1874-8600. doi: 10.1007/s11092-019-09294-8. URL https://doi.org/10.1007/s11092-019-09294-8.

[285] Peter J Skene and Steven Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *eLife*, 6:e21856, January 2017. ISSN 2050-084X. doi: 10.7554/eLife.21856. URL https://doi.org/10.7554/eLife.21856.

[286] David Slik. Estimating data storage device lifespan, September 2016. US Patent 9,436,571.

[287] Reginald Smith. A mutual information approach to calculating nonlinearity. *Stat*, 4(1):291–303, 2015. doi: https://doi.org/10.1002/sta4.96. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.96.

[288] Ahyun Son, Scott Horowitz, and Baik L. Seong. Chaperna: linking the ancient rna and protein worlds. *RNA Biology*, 18(1):16–23, 2021. doi: 10.1080/15476286.2020.1801199. URL https://doi.org/10.1080/15476286.2020.1801199. PMID: 32781880.

[289] Sebastian Spänig and Dominik Heider. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Mining*, 12(1):7, March 2019. ISSN 1756-0381. doi: 10.1186/s13040-019-0196-x. URL https://doi.org/10.1186/s13040-019-0196-x.

[290] Sebastian Spänig, Siba Mohsen, Georges Hattab, Anne-Christin Hauschild, and Dominik Heider. A large-scale comparative study on peptide encodings for biomedical classification. *NAR Genomics and Bioinformatics*, 3(2), 05 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab039. URL https://doi.org/10.1093/nargab/lqab039. lqab039.

[291] Bärbel M.R. Stadler, Peter F. Stadler, Günter P. Wagner, and Walter Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213(2):241–274, 2001. ISSN 0022-5193. doi: https://doi.org/10.1006/jtbi.2001.2423. URL https://www.sciencedirect.com/science/article/pii/S0022519301924233.

[292] David J. Staley. *Computers, Visualization and History: How New Technology Will Transform Our Understanding of the Past*. Routledge, paperback edition, 12 2002. ISBN 978-0765610959. URL https://lead.to/amazon/com/?op=bt&la=en&cu=usd&key=0765610957.

[293] Michael Steinbach, Levent Ertöz, and Vipin Kumar. *The Challenges of Clustering High Dimensional Data*, pages 273–309. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-662-08968-2. doi: 10.1007/978-3-662-08968-2_16. URL https://doi.org/10.1007/978-3-662-08968-2_16.

[294] Robert J. Sternberg and Michael K. Gardner. Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112:80–116, 1983. doi: 10.1037/0096-3445.112.1.80. URL https://doi.org/10.1037/0096-3445.112.1.80.

[295] S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946. doi: 10.1126/science.103.2684.677. URL https://www.science.org/doi/abs/10.1126/science.103.2684.677.

[296] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303321897735. URL https://doi.org/10.1162/153244303321897735.

[297] Masatake Sugita, Satoshi Sugiyama, Takuya Fujie, Yasushi Yoshikawa, Keisuke Yanagisawa, Masahito Ohue, and Yutaka Akiyama. Large-scale membrane permeability prediction of cyclic peptides crossing a lipid bilayer based on enhanced sampling molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 61(7):3681–3695, July 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00380. URL https://doi.org/10.1021/acs.jcim.1c00380.

[298] Jian Sun, Qiming Wang, Wenyi Diao, Chi Zhou, Bingbing Wang, Liqun Rao, and Ping Yang. Digital information storage on dna in living organisms. *Medical Research Archives*, 7(6), 2019. ISSN 2375-1924. doi: 10.18103/mra.v7i6.1930. URL https://esmed.org/MRA/mra/article/view/1930.

[299] Jerome Svigals. The long life and imminent death of the mag-stripe card. *IEEE Spectrum*, 49 (6):72–76, June 2012. ISSN 1939-9340. doi: 10.1109/MSPEC.2012.6203975.

[300] Danielle Albers Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, Jan 2018. ISSN 1941-0506. doi: 10.1109/TVCG.2017.2744359.

[301] Sahar Tahvili and Leo Hatvani. Chapter three - transformation, vectorization, and optimization. In Sahar Tahvili and Leo Hatvani, editors, *Artificial Intelligence Methods for Optimization of the Software Testing Process*, Uncertainty, Computational Techniques, and Decision Intelligence, pages 35–84. Academic Press, 2022. ISBN 978-0-323-91913-5. doi: https://doi.org/10.1016/B978-0-32-391913-5.00014-2. URL https://www.sciencedirect.com/science/article/pii/B9780323919135000142.

[302] M. Takatsuka and J. Zhou. Automatic transfer function generation using contour tree controlled residue flow model and color harmonics. *IEEE Transactions on Visualization and Computer Graphics*, 15(06):1481–1488, nov 2009. ISSN 1941-0506. doi: 10.1109/TVCG.2009.120.

[303] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 1283–1292, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi: 10.1145/1518701.1518895. URL https://doi.org/10.1145/1518701.1518895.

[304] E. Tan and B. Vermuelen. Digital audio tape for data storage. *IEEE Spectrum*, 26(10):34–38, October 1989. ISSN 1939-9340. doi: 10.1109/6.40682.

[305] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 287–297, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883041. URL https://doi.org/10.1145/2872427.2883041.

[306] Karl E. Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192, 2001. doi: https://doi.org/10.1029/2000JD900719. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JD900719.

[307] Rebecca E. Taylor and Maliha Zahid. Cell penetrating peptides, novel vectors for gene therapy. *Pharmaceutics*, 12(3), 2020. ISSN 1999-4923. doi: 10.3390/pharmaceutics12030225. URL https://www.mdpi.com/1999-4923/12/3/225.

[308] M. Thomas and F. McGarry. Top-down vs. bottom-up process improvement. *IEEE Software*, 11(4):12–13, July 1994. ISSN 1937-4194. doi: 10.1109/52.300121.

[309] Michael E. Tipping and Christopher M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 02 1999. ISSN 0899-7667. doi: 10.1162/089976699300016728. URL https://doi.org/10.1162/089976699300016728.

[310] Christian Tominski and Heidrun Schumann. *Interactive Visual Data Analysis*. AK Peters Visualization Series. CRC Press, 2020. ISBN 9781498753982. doi: 10.1201/9781315152707. URL https://ivda-book.de.

[311] Bindu Trikha. A journey from floppy disk to cloud storage. *International Journal on Computer Science and Engineering*, 2(4):1449–1452, 2010.

[312] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656.

[313] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. *Stochastic Gradient Descent Training for L1-Regularized Log-Linear Models with Cumulative Penalty*. ACL '09. Association for Computational Linguistics, USA, 2009. ISBN 9781932432459.

[314] E.R. Tufte. *The Visual Display of Quantitative Information*. Number Bd. 914 in The Visual Display of Quantitative Information. Graphics Press, 1983. ISBN 9780318029924. URL https://books.google.de/books?id=tWpHAAAAMAAJ.

[315] E.R. Tufte. *Envisioning Information*. Graphics Press, 1990. ISBN 9781930824140. URL https://books.google.de/books?id=I3BQAAAAMAAJ.

[316] Adelinde M Uhrmacher, François E Cellier, and Robert J Frye. Applying fuzzy-based inductive reasoning to analyze qualitatively the dynamic behavior of an ecological system. *AI APPLICATIONS*, 11(2):1–10, 1997.

[317] Thomas Unden, Nikolas Tomek, Timo Weggler, Florian Frank, Paz London, Jonathan Zopes, Christian Degen, Nicole Raatz, Jan Meijer, Hideyuki Watanabe, Kohei M. Itoh, Martin B. Plenio, Boris Naydenov, and Fedor Jelezko. Coherent control of solid state nuclear spin nano-ensembles. *npj Quantum Information*, 4(1):39, August 2018. ISSN 2056-6387. doi: 10.1038/s41534-018-0089-8. URL https://doi.org/10.1038/s41534-018-0089-8.

[318] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9(nov):2579–2605, 2008. ISSN 1532-4435. Pagination: 27.

[319] Tom van der Valk, Patrícia Pečnerová, David Díez-del Molino, Anders Bergström, Jonas Oppenheimer, Stefanie Hartmann, Georgios Xenikoudakis, Jessica A. Thomas, Marianne Dehasque, Ekin Sağlıcan, Fatma Rabia Fidan, Ian Barnes, Shanlin Liu, Mehmet Somel, Peter D. Heintzman, Pavel Nikolskiy, Beth Shapiro, Pontus Skoglund, Michael Hofreiter, Adrian M. Lister, Anders Götherström, and Love Dalén. Million-year-old dna sheds light on the genomic history of mammoths. *Nature*, 591(7849):265–269, March 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03224-9. URL https://doi.org/10.1038/s41586-021-03224-9.

[320] Bert Van Es. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics*, pages 61–72, 1992.

[321] Christian van Onzenoodt, Anke Huckauf, and Timo Ropinski. On the perceptual influence of shape overlap on data-comparison using scatterplots. *Computers & Graphics*, 90:169–181,

2020. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2020.05.028. URL https://www.sciencedirect.com/science/article/pii/S0097849320300881.

[322] Guido van Rossum and Jelke de Boer. Interactively testing remote servers using the python programming language. *CWI Quarterly*, 4(4):283–304, December 1991.

[323] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[324] Jacob VanderPlas, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. Altair: Interactive statistical visualizations for python. *Journal of Open Source Software*, December 2018. doi: 10.21105/joss.01057. URL https://doi.org/10.21105/joss.01057.

[325] Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1):54–59, 1976. doi: https://doi.org/10.1111/j.2517-6161.1976.tb01566.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1976.tb01566.x.

[326] Corinna Vehlow, Michael Burch, Hansjorg Schmauder, and Daniel Weiskopf. Radial layered matrix visualization of dynamic graphs. In *2013 17th International Conference on Information Visualisation*, pages 51–58, July 2013. doi: 10.1109/IV.2013.6.

[327] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854, dec 2010. ISSN 1532-4435.

[328] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[329] Jeffrey Scott Vitter. External memory algorithms and data structures: Dealing with massive data. *ACM Comput. Surv.*, 33(2):209–271, June 2001. ISSN 0360-0300. doi: 10.1145/384192.384193. URL https://doi.org/10.1145/384192.384193.

[330] Jaykant Vora, Shivani Patel, Sonam Sinha, Sonal Sharma, Anshu Srivastava, Mahesh Chhabria, and Neeta Shrivastava. Molecular docking, qsar and admet based mining of natural compounds against prime targets of hiv. *Journal of Biomolecular Structure and Dynamics*, 37(1):131–146, 2019. doi: 10.1080/07391102.2017.1420489. URL https://doi.org/10.1080/07391102.2017.1420489. PMID: 29268664.

[331] Dennis Wagner, Dominik Heider, and Georges Hattab. Mushroom data creation, curation, and simulation to support classification tasks. *Scientific Reports*, 11, 04 2021. doi: 10.1038/s41598-021-87602-3.

[332] Colin Ware. Chapter four - color. In Colin Ware, editor, *Information Visualization (Third Edition)*, Interactive Technologies, pages 95–138. Morgan Kaufmann, Boston, third edition edition, 2013. ISBN 978-0-12-381464-7. doi: https://doi.org/10.1016/B978-0-12-381464-7.00004-1. URL https://www.sciencedirect.com/science/article/pii/B9780123814647000041.

[333] Colin Ware. Chapter ten - interacting with visualizations. In Colin Ware, editor, *Information Visualization (Fourth Edition)*, Interactive Technologies, pages 359–392. Morgan Kaufmann, fourth edition edition, 2021. ISBN 978-0-12-812875-6. doi: https://doi.org/10.1016/B978-0-12-812875-6.00010-4. URL https://www.sciencedirect.com/science/article/pii/B9780128128756000104.

[334] Liping Wei, Yueyi Liu, Inna Dubchak, John Shon, and John Park. Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*, 35 (2):142–150, 2002. ISSN 1532-0464. doi: https://doi.org/10.1016/S1532-0464(02)00506-3. URL https://www.sciencedirect.com/science/article/pii/S1532046402005063.

[335] Peiran Wei, Bowen Li, Al de Leon, and Emily Pentzer. Beyond binary: optical data storage with 0, 1, 2, and 3 in polymer films. *J. Mater. Chem. C*, 5:5780–5786, 2017. doi: 10.1039/C7TC00929A. URL http://dx.doi.org/10.1039/C7TC00929A.

[336] Marius Welzel, Peter Michael Schwarz, Hannah Löchel, Tolganay Kabdullayeva, Sandra Clemens, Anke Becker, Bernd Freisleben, and Dominik Heider. Dna-aeon provides flexible arithmetic coding for constraint adherence and error correction in dna storage. *Nature Communications*, 14, 02 2023. doi: 10.1038/s41467-023-36297-3.

[337] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[338] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020. ISSN 1740-6749. doi: https://doi.org/10.1016/j.ddtec.2020.11.009. URL https://www.sciencedirect.com/science/article/pii/S1740674920300305.

[339] Peter Willett. Similarity-based virtual screening using 2d fingerprints. *Drug Discovery Today*, 11(23):1046–1053, 2006. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2006.10.005. URL https://www.sciencedirect.com/science/article/pii/S1359644606004193.

[340] Joshua R. Williams, Ruoting Yang, John L. Clifford, Daniel Watson, Ross Campbell, Derese Getnet, Raina Kumar, Rasha Hammamieh, and Marti Jett. Functional heatmap: an au-

tomated and interactive pattern recognition tool to integrate time with multi-omics assays. *BMC Bioinformatics*, 20(1):81, February 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2657-0. URL https://doi.org/10.1186/s12859-019-2657-0.

[341] W Wolberg and O Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences of the United States of America*, 87:9193–6, 01 1991. doi: 10.1073/pnas.87.23.9193.

[342] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. ISSN 0169-7439. doi: https://doi.org/10.1016/0169-7439(87)80084-9. URL https://www.sciencedirect.com/science/article/pii/0169743987800849. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[343] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://doi.org/10.48550/arXiv.1708.07747.

[344] Shuang Xu, Junqin Yao, Meihaguli Ainiwaer, Ying Hong, and Yanjiang Zhang. Analysis of bacterial community structure of activated sludge from wastewater treatment plants in winter. *BioMed Research International*, 2018, 2018. doi: https://doi.org/10.1155/2018/8278970.

[345] H. Yamada. Dvd overview [removable storage media]. In *Proceedings IEEE COMPCON 97. Digest of Papers*, pages 287–290, February 1997. doi: 10.1109/CMPCON.1997.584732.

[346] J. Yang-Pelaez and W.C. Flowers. Information content measures of visual displays. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 99–103, Oct 2000. doi: 10.1109/INFVIS.2000.885096.

[347] Sinan Yatkin, Michel Gerboles, Annette Borowiak, Silvije Davila, Laurent Spinelle, Alena Bartonova, Frank Dauge, Philipp Schneider, Martine Van Poppel, Jan Peters, Christina Matheeussen, and Marco Signorini. Modified target diagram to check compliance of low-cost sensors with the data quality objectives of the european air quality directive. *Atmospheric Environment*, 273:118967, 2022. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2022.118967. URL https://www.sciencedirect.com/science/article/pii/S1352231022000322.

[348] Ka-Ping Yee, D. Fisher, R. Dhamija, and M. Hearst. Animated exploration of dynamic graphs with radial layout. In *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.*, pages 43–50, Oct 2001. doi: 10.1109/INFVIS.2001.963279.

[349] Aldrin Kay-Yuen Yim, Allen Chi-Shing Yu, Jing-Woei Li, Ada In-Chun Wong, Jacky F. C. Loo, King Ming Chan, S. K. Kong, Kevin Y. Yip, and Ting-Fung Chan. The essential component in dna-based information storage system: Robust error-tolerating module. *Frontiers in Bioengineering and Biotechnology*, 2, 2014. ISSN 2296-4185. doi: 10.3389/fbioe.2014.00049. URL https://www.frontiersin.org/articles/10.3389/fbioe.2014.00049.

[350] T. Yoshida. The rewritable minidisc system. *Proceedings of the IEEE*, 82(10):1492–1500, October 1994. ISSN 1558-2256. doi: 10.1109/5.326407.

[351] Nikolay Zenkin. Ancient rna stems that terminate transcription. *RNA Biology*, 11(4):295–297, 2014. doi: 10.4161/rna.28342. URL https://doi.org/10.4161/rna.28342. PMID: 24643067.

[352] Harry Zhang. *The Optimality of Naive Bayes*. American Association for Artificial Intelligence, 2004.

[353] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, jun 1996. ISSN 0163-5808. doi: 10.1145/235968.233324. URL https://doi.org/10.1145/235968.233324.

[354] Wei Zhang, Qiang Zhang, Meng-Qiang Zhao, and Luise Theil Kuhn. Direct writing on graphene 'paper' by manipulating electrons as 'invisible ink'. *Nanotechnology*, 24(27):275301, June 2013. doi: 10.1088/0957-4484/24/27/275301. URL https://dx.doi.org/10.1088/0957-4484/24/27/275301.

[355] Yunzhu Zheng, Haruka Suematsu, Takayuki Itoh, Ryohei Fujimaki, Satoshi Morinaga, and Yoshinobu Kawahara. Scatterplot layout for high-dimensional data visualization. *Journal of Visualization*, 18, 02 2015. doi: 10.1007/s12650-014-0230-5.

[356] Jianlong Zhou, Weidong Huang, and Fang Chen. Facilitating machine learning model comparison and explanation through a radial visualisation. *Energies*, 14(21), 2021. ISSN 1996-1073. doi: 10.3390/en14217049. URL https://www.mdpi.com/1996-1073/14/21/7049.

[357] Wenyu Zhou, M. Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R. Leopold, Martin J. Zhang, Varsha Rao, Monika Avina, Tejaswini Mishra, Jethro Johnson, Brittany Lee-McMullen, Songjie Chen, Ahmed A. Metwally, Thi Dong Binh Tran, Hoan Nguyen, Xin Zhou, Brandon Albright, Bo-Young Hong, Lauren Petersen, Eddy Bautista, Blake Hanson, Lei Chen, Daniel Spakowicz, Amir Bahmani, Denis Salins, Benjamin Leopold, Melanie Ashland, Orit Dagan-Rosenfeld, Shannon Rego, Patricia Limcaoco, Elizabeth Colbert, Candice Allister, Dalia Perelman, Colleen Craig, Eric Wei, Hassan Chaib, Daniel Hornburg, Jessilyn Dunn, Liang Liang, Sophia Miryam Schüssler-Fiorenza Rose, Kim Kukurba, Brian Piening, Hannes Rost, David Tse, Tracey McLaughlin, Erica Sodergren, George M. Weinstock, and Michael Snyder. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, 569(7758):663–671, May 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1236-x. URL https://doi.org/10.1038/s41586-019-1236-x.

[358] Ji Zhu, Saharon Rosset, Hui Zou, and Trevor Hastie. Multi-class adaboost. *Statistics and its interface*, 2, 02 2006. doi: 10.4310/SII.2009.v2.n3.a8.

[359] Matt Zwolenski and Lee Weatherill. The digital universe: Rich data and the increasing value of the internet of things. *Journal of Telecommunications and the Digital Economy*, 2 (3):[47.1]–[47.9], 2014. URL https://search.informit.org/doi/10.3316/informit.678436300116927.

[360] Tunahan Çakır and Mohammad Jafar Khatibipour. Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. *Frontiers in Bioengineering and Biotechnology*, 2, 2014. ISSN 2296-4185. doi: 10.3389/fbioe.2014.00062. URL https://www.frontiersin.org/articles/10.3389/fbioe.2014.00062.