Extracting Essential Information and Making Inference from Data

Kjærsgaard, Rune Dodensig

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Extracting Essential Information and Making Inference from Data

Rune Dodensig Kjærsgaard

Kongens Lyngby 2023

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

In recent years, machine learning and artificial intelligence systems have seen great success across a variety of domains. These systems are fueled by their underlying training data, which often stem from extensive historical datasets. Inspired by these advancements, the collection of data has exploded, and significant effort has been funneled into improving the models. Nonetheless, the data sources used to train machine learning models are often unstructured, noisy and encode historical biases. These aspects are frequently neglected during model optimization, and can seep into the trained model, resulting in poor or biased predictive inference.

The goal of this thesis is to shift the focus towards the data by assessing and developing data reduction methods for learning summary data representations, which capture the essential information and are *representative* of the original data source. En route to this, an important consideration rests on defining and evaluating what representative data entail, and how they may be appropriately extracted.

The thesis is divided into two parts. Firstly, part I presents an overview of current practises in data reduction, which demonstrate how dimensionality and numerosity reduction can be used to learn smaller data representations that can lead to reduced computational burdens and improved inference. Part II of the thesis presents the research contributions, which highlight and address various intricacies of the problem by evaluating the ways data can be representative of a target population, and how such data representations can be learned across several scientific domains.

Summary (Danish)

Maskinlæring og kunstig intelligens har haft stor succes over de sidste år på tværs af en række forskellige domæner. Disse systemer er drevet af deres underliggende træningsdata, som typisk stammer fra omfattende historiske datasæt. Fremskridtet har inspireret en tiltagende indsamling af data, og en betydelig indsats for at forbedre modellerne. Den indsamlede træningsdata er dog ofte ustruktureret og indeholder historiske bias. Dette negligeres hyppigt under modeloptimering og kan resultere i dårlige eller uretfærdige forudsigelser fra den trænede model.

Målet med denne afhandling er at flytte fokus mod data ved at vurdere og udvikle datareduktionsmetoder til at lære datarepræsentationer, som fanger den essentielle information og er *repræsentative* for den originale datakilde. Dette inkluderer at definere og vurdere, hvad repræsentative data indebærer, og hvordan de kan udvindes hensigtsmæssigt.

Afhandling er opdelt i to dele. Del I giver et overblik over nuværende praksis i datareduktion, som viser, hvordan dimensionalitet- og numerositetsreduktion kan bruges til at lære mindre datarepræsentationer, der kan reducere computermæssige ressourcer og lede til bedre inferens. Del II af afhandlingen præsenterer forskningsbidragene, som fremhæver og adresserer forskellige problemstillinger ved at evaluere, hvordan data kan være repræsentative for en population, og hvordan sådanne datarepræsentationer kan læres på tværs af en række videnskabelige domæner.

Preface

This thesis was prepared at the Section for Statistics and Data Analysis under the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU) in partial fulfillment of the requirements for the degree of PhD. The project was financed by a university alliance scholarship between the University of Bergen (UiB) and DTU. The project was supervised by Associate Professor Line Katrine Harder Clemmensen and co-supervised by Professor Bjarne Kjær Ersbøll from DTU Compute, Denmark and Professor Saket Saurabh from UiB, Bergen, Norway. The project was conducted at DTU from September 2020 - September 2023.

The thesis consists of two parts. The first part introduces the background, methodology and context of the research. The second part presents the research outcomes and contains 5 research contributions (2 published, 2 submitted for review and 1 preprint) which evaluate and develop techniques for data reduction using methods at the intersection of statistics and machine learning. The first two papers demonstrate techniques for empowering neural network autoencoders with transparent architectures to reduce the dimensionality of the original data and learn a compressed data representation imbued with high interpretability. The third paper concerns data representativity and evaluates what it means for a data summary to be representative. The final two papers demonstrate sampling and clustering approaches for learning data representations which can be used as curated training data to reduce the bias of downstream models.

Lyngby, 31-August-2023

Rune Dodensig Kjærsgaard

Acknowledgements

First, I would like to thank my supervisor Associate Professor Line Katrine Harder Clemmensen for her continued support throughout the project. You have been a constant source of inspiration and a guiding beacon during the crucible of my studies. I would also like to thank Professor Saket Saurabh and Associate Professor Pekka Parviainen for hosting me at the Department of Informatics at the University of Bergen, where I conducted my external research stay. Your hospitality and fundamental theoretical understanding of algorithms have ensured that my stay was personally and professionally enriching. I also thank my co-supvisor Professor Bjarne Kjær Ersbøll for his central role in instigating the collaborations with the university of Bergen.

I extend sincere thanks to all my collaborators, including Professor Lars A. Buchhave and Postdoctoral Fellows Aaron Bello-Arufe and Alexander D. Rathcke from DTU space, California Institute of Technology and Harvard & Smithsonian Center for Astrophysics, Postdoctoral Fellow Ahcène Boubekki from the University of Tromsø, PhD Manja from DTU Compute and PhD student Madhumita Kundu from the University of Bergen. Your domain expertise and our many fruitful discussions have motivated me and helped bridge multiple interdisciplinary gaps. This extends to my colleagues and friends at DTU, who have enriched the daily office days with numerous rewarding and fun conversions and activities.

I give my deep thanks to my family and friends. Your support has been instrumental in everything I have accomplished during the project. A special thanks to my wife Emilie. You have always been by my side, shown unconditional support and provided crucial reinforcement during the hardest times. Last but not least, I am grateful for my son Bertram. You are an invigorating everlasting source of creativity and joy.

Contents

Chapter 1

Introduction

The goal of this thesis is to advance the progress of research within data reduction and data representativity for machine learning and artificial intelligence systems. The thesis develops, discusses and evaluates methods for obtaining representations, which summarize and maintain the characteristics and inherent information in the original data. These surrogate data representations can be viewed as curated *representative samples* or *summary representations*.

The target audience of this thesis is practitioners who are looking to develop techniques for data reduction, or to understand and evaluate the connection between data representativity and appropriate inference. Fundamental understanding of machine learning and statistics is recommended, but not required to understand the thesis.

This chapter gives a summary of the thesis by first outlining the motivation and background of the research. Next, the primary scientific contributions and research objectives are specified. Finally, the remaining sections of the thesis are described.

1.1 Motivation and Background

In recent years the acquisition of data has exploded, causing large datasets to appear in almost every industry of our modern society. The value of extracting information from this data is becoming rapidly apparent, and the insights gained are increasingly used to guide significant decisions that influence people at all levels of society. These insights are largely derived from models trained on extensive datasets, where for the past decade the dominant paradigm has involved a model-centric view [Unb22]. Under this paradigm, the data are kept mostly unchanged, while the models are adjusted to increase performance. This has led to significant improvements on various benchmark problems. However, the underlying data used to train these models are often polluted by undesired noise, imbalance, missing values, and data irrelevant to the problem at hand [KS19]. The previous paradigm fails to sufficiently address these issues [ZBL+23a], which has sparked the advent of a new point of view: data-centric AI, wherein the focus is shifted towards the data. Data-centric AI may be defined as a systematic effort towards developing, engineering, and maintaining data for successful AI systems [ZBL+23a, NG-21, JMG22]. The data-centric perspective encompasses a wide spectrum of tasks that can be broadly categorized into three objectives: inference data development, training data development and data maintenance [ZBL+23b]. Fig. 1.1 illustrates this high-level overview.

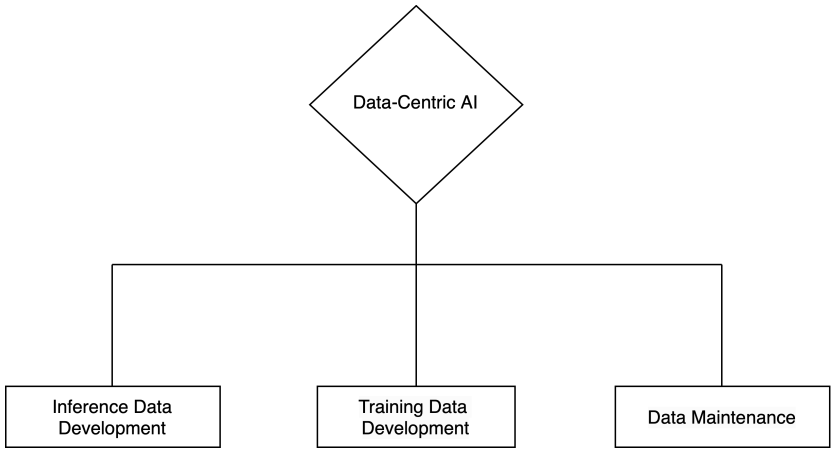


Figure 1.1: High-level overview of data-centric AI.

Inference Data Development: Inference data, also known as test data, are used by machine learning systems to assess model performance. These datasets are developed for both in-distribution and out-of-distribution evaluation. Developing specific inference sets, for instance for sub-populations within the data, allows practitioners to gain more granular insights into model capabilities.

Training Data Development: Training data form the bedrock of machine learning systems. For this reason, an important task is to gather or produce rich training data, which can be leveraged to tune model parameters and ultimately allow inference. This task spans several sub-categories including data collection, labeling, transformation, augmentation, and reduction.

Data Maintenance: Data maintenance refers to monitoring the quality and reliability of data and involves the sub-tasks of data storage and retrieval, data understanding, and data quality assurance. Data quality assurance may be supported by quantitative measures which ensure that the data support the purpose for which they were collected. These tasks act in a supportive role to assess whether the training and inference data are reliable and of high quality [JPN+20].

This thesis studies training data development via construction of reduced data representations. However, in order to evaluate the data quality and downstream inferences made from these representations, the topics of inference data development and data maintenance will also be addressed.

In the modern era we are faced with increasing amounts of data, which accelerate computational costs and may be highly redundant, imbalanced, and unstructured. This motivates the need for data reduction techniques, which lower the complexity by identifying reduced representations that maintain the characteristics and statistical properties of the original data. These techniques not only reduce the computational costs of downstream analysis and storage [FHL14], but also allow models to focus on the essential information, which can enhance accuracy and interpretability [XLZ+19]. This is critical as the original data volume is sometimes so large that it becomes impossible to process the entire data population. Furthermore, while vast amounts of data are generally thought to be representative by definition, this is not necessarily the case [big19]. As we draw conclusions from data, it is critical to evaluate what these data represent, and which inferences they allow us to make. Even large-sized datasets can be non-representative for specific tasks or sub-groups within the data, which translates to lower quality results. Identifying balanced reduced representations may allow us to mitigate these issues [CKLV17].

Data reduction can be broadly categorized into two regimes: numerosity and dimensionality reduction [Gho21, ZBL+23b]. Consider n data instances $\{\boldsymbol{x}_i\}_{i=1}^{n}$ occupying a p-dimensional space $\forall i \in \{1, ..., n\} : \boldsymbol{x}_i \in \mathbb{R}^p$, where the data instances form the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] \in \mathbb{R}^{n \times p}$. Numerosity reduction refers to a reduction in the number of data instances (cardinality) of $\boldsymbol{X}$ from n to $m \in (0, n]$. That is, we are looking to identify (or generate) informative instances to construct a new data representation $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m \times p}$. Numerosity reduction can lead to simpler and yet representative samples that not only alleviate computational costs but can also mitigate inherent data biases [PKDN15]. On the other hand, dimensionality reduction refers to a mapping from the original data $\boldsymbol{X}$ in p-dimensional space to a new representation $\widetilde{\boldsymbol{X}}$ in a lower dimensional space $r \in (0, p]$, where $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times r}$. Dimensionality reduction can bring computational benefits [WBJ+22] while also increasing interpretability [CWY+23] and addressing the curse of dimensionality [VF05].

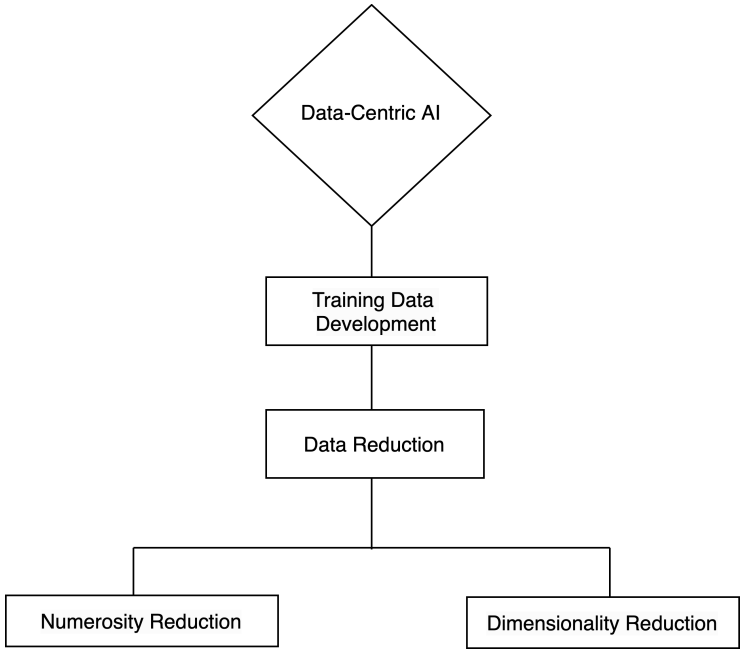


Figure 1.2: Data reduction as a sub-field of data-centric AI.

1.2 Research objectives and contributions

This thesis provides a study on data reduction and data representativity by assessing current methods and their limitations. We study machine learning as well as algorithmic based approaches for obtaining *summary representations*, which maintain the characteristics and inherent information from the population. In line with this, the PhD thesis will address the following research objectives:

1. Research, evaluate and develop methods for obtaining surrogate data representations, which summarize and represent the original data population in a smaller form that may be used as curated training data for downstream models.
2. Evaluate the appropriate application of techniques across different domains by identifying the distinct types of essential information that data reduction seeks to distill.
3. Develop dimensionality reduction techniques that infuse the reduction pipeline and learned representation with high transparency and interpretability.
4. Study data reduction techniques that not only reduce data size, but also reduce inherent data biases existing in the original data.

The research outcomes align with these objectives and take the form of evaluation and development of novel data reduction techniques. The following contributions are a result of the research conducted under this PhD thesis:

1. A highly transparent and interpretable deep learning approach for obtaining a reduced representation of astrophysical spectral data.
2. Introduction of a volumetric loss used in a deep learning setting to learn an expressive data representation for a downstream linear model, which supports and expands current practices in explainable AI.
3. Assessing and expanding the taxonomy of *representative samples* in machine learning and AI literature, which aid in evaluating how different reduction methods may be appropriate for application across various tasks and domains.
4. Alternative sampling approaches to obtain balanced representations with various applications.
5. A fair variant of probabilistic clustering and the introduction of algorithmic approaches for obtaining the fair data representation and solution.

1.3 Paper Contributions

The thesis includes 5 paper contributions. These papers have varying degrees of publication statuses listed below.

Paper A:
TAU: A neural network based telluric correction framework [KBAR+23].
R. Kjærsgaard, A. Bello-Arufe, A. Rathcke, L. Buchhave and L. Clemmensen (2023).
Journal: Astronomy & Astrophysics.
Publication Status: Accepted. Publication forthcoming.

Paper B:
Pantypes: Diverse Representatives for Self-Explainable Models [KBC23].
R. Kjærsgaard, A. Boubekki and L. Clemmensen (2023).
Conference: AAAI 2024.
Publication Status: Submitted for review.

Paper C:
Data Representativity for Machine Learning and AI Systems [CK23].
L. Clemmensen and R. Kjærsgaard (2023).
Journal: ACM Computing Surveys.
Publication Status: In preparation. To be submitted.

Paper D:
Sampling To Improve Predictions For Underrepresented Observations In Imbalanced Data [KGC21a].
Conference: NeurIPS 2021 - Workshop on Data-Centric AI.
R. Kjærsgaard, M. Grønberg and L. Clemmensen (2021).
Publication Status: Published.

Paper E:
Fair Soft Clustering [KPS+23].
R. Kjærsgaard, P. Parviainen, S. Saurabh, M. Kundu and L. Clemmensen (2023).
Journal: Journal of Machine Learning Research.
Publication Status: Submitted for review.

The thesis will discuss the research contribution provided by each paper and put them into context with the overarching data-centric framework. See Fig. 1.3 for context.

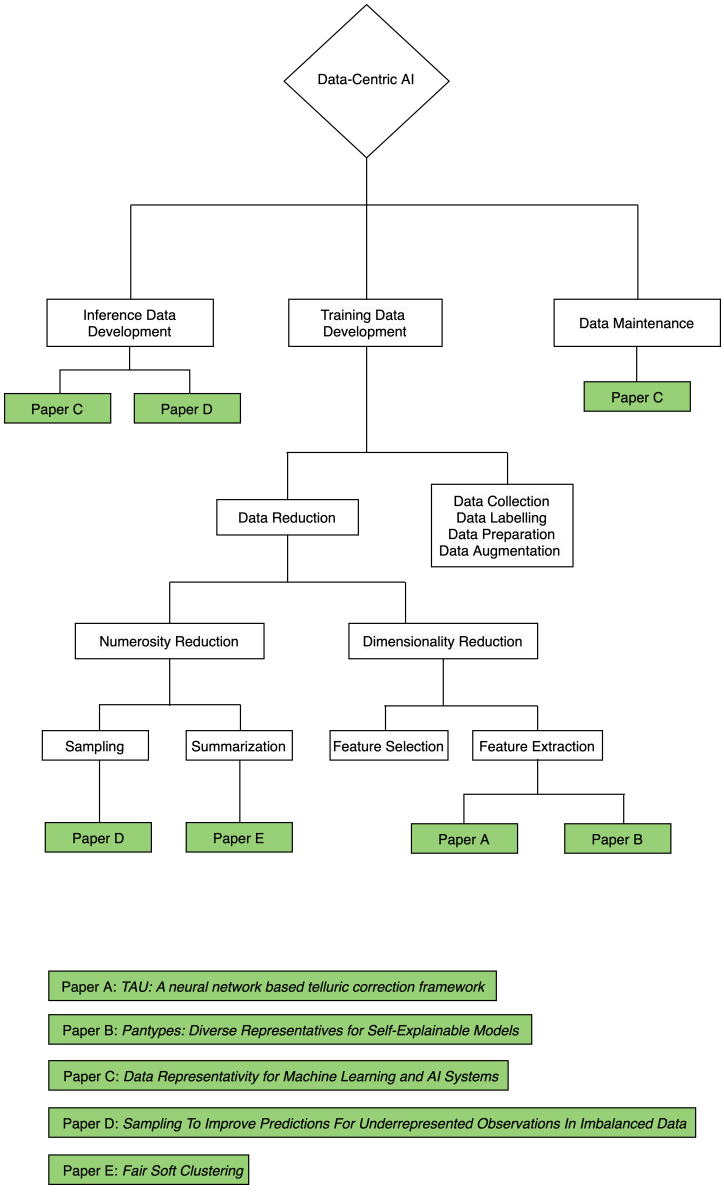


Figure 1.3: Overview of the paper contributions in the landscape of data-centric AI. Feature selection and feature extraction based dimensionality reduction will be explored in Chapter 2 while sampling and summarization based numerosity reduction will be studied in Chapter 3.

An additional paper has been prepared under the PhD project but is not included in the thesis. This paper is an early extended conference abstract, which has since then been significantly expanded upon. The full work is presented in Paper A.

Paper F:
Unsupervised Spectral Unmixing For Telluric Correction Using A Neural Network Autoencoder [KBAR+21].
R. Kjærsgaard, A. Bello-Arufe, A. Rathcke, L. Buchhave and L. Clemmensen (2023).
Conference: NeurIPS 2021 - Workshop on Machine Learning and the Physical Sciences.
Publication Status: Published.

1.4 Outline of the Thesis

The thesis is divided into two parts. First, the methodology and context of the thesis are presented under Part I in chapters 2-3. This part gauges data reduction through the prism of two concepts: dimensionality reduction and numerosity reduction. Part II contains chapters 4-10 and outlines open problems and presents the research outcomes of the thesis. These research contributions are put into context with the research objectives in Sect. 1.2 and the framework of data representativity presented in contribution C.

Chapter 2 introduces Part I by presenting the fundamentals of dimensionality reduction, which targets the variables of the original data to produce compact data representations. Here we assess various approaches that are employed for a number of data modalities, and discuss their advantages and limitations. We identify a common trend across the reviewed methods, that despite the strong expressive power of the derived representations, the reduction pipeline and associated learned representations often lack transparency and interpretability, which hinders their applicability for a number of domains.

Chapter 3 presents numerosity reduction, which directly targets the data instances to obtain smaller representations. Numerosity reduction includes traditional tools like sampling and clustering. We review approaches in the literature and identify a common trend that they are most often used to construct *representative samples*[1] from the original data, which can alleviate computational costs of succeeding models.

[1]This term is ill-defined and obfuscates a web of conflicting ideas. Sect. 4.2 and the associated research contribution C explores this topic in detail.

Chapter 4 introduces Part II of the thesis by outlining open problems and summarizing the research outcomes of the thesis. This section also discusses how the paper contributions relate to the research objectives and how the contributions may find practical use.

Chapter 5 presents the conclusion of the thesis.

Chapter 6 contains research contribution A: "TAU: A neural network based telluric correction framework", which studies data reduction in natural sciences. The paper presents an approach for reducing an extensive dataset of astrophysical observations into a compact form of a few underlying components that can be used for downstream inference. This is enabled by a highly transparent and interpretable reduction pipeline that fuses ideas from neural network autoencoders, factor analysis and hyperspectral unmixing. The approach outperforms the state-of-the-art in inference and is approximately 10,000 times faster.

Chapter 7 contains research contribution B: "Pantypes: Diverse Representatives for Self-Explainable Models". This paper is situated in the domain of explainable AI, which aims to imbue AI systems with higher transparency, interpretability and explainability. The paper develops and evaluates a diversity inducing volumetric approach implemented in a variational autoencoder for reducing image data into a compressed form. The compressed form is used both for visualizing and understanding the input data but is also fed as input to a downstream linear classifier to draw inference. The contribution demonstrates how the construction of the compressed representation affects the performance of the model and impacts the transparency and interpretability of the overall system.

Chapter 8 contains research contribution C: "Data Representativity for Machine Learning and AI Systems". This paper investigates notions of data representativity and concludes that despite a ubiquitous appearance in ML and AI literature, the term *representative sample* is ill-defined and encompasses a range of conflicting perspectives. The paper reviews and surveys the various notions of representativity existing in the literature, and proposes a unified framework of measurable concepts, which may help organize and assess the representativity of existing datasets and derived representations. In light of this, the paper also suggests a guideline of questions related to data representativity to consider when publishing datasets.

Chapter 9 contains research contribution D: "Sampling To Improve Predictions For Underrepresented Observations In Imbalanced Data". This paper studies sampling in manufacturing data, where data imbalance is common. The paper introduces three sampling approaches, which can be used to find smaller balanced representations of the input data. The paper demonstrates how training a model on these reduced representations results in better predictive performance for observations that are underrepresented in the original data space. Such sampling carries implications for manufacturing and production data, but may also find use in demographic data, where underrepresented observations could constitute minority sub-groups of individuals.

Chapter 10 contains research contribution E: "Fair Soft Clustering". This paper extends a recent line of research studying clustering under notions of fairness. Various notions of fairness exist, but in clustering the dominant metric involves a measure of balance, which evaluates if observations in different clusters exhibit balance of protected attributes such as sex or race. Fair clustering research is largely focused on hard clustering, where cluster assignments are fully deterministic. Contribution E aims to extend this line of research into the soft domain, where assignments are probabilistic. The initial dominant paradigm in fair clustering involves modifying the original data into a reduced data representation, which can be fed as input to a traditional clustering algorithm to achieve a balanced solution. We draw on this idea to construct an algorithm to achieve a theoretically bounded fair probabilistic cluster solution from a reduced representation of the data.

Part I

Methodology & Context

Chapter 2

Dimensionality Reduction and Manifold Learning

Dimensionality reduction reduces the number of dimensions in a dataset to mitigate various detrimental properties related to high-dimensional data. Again, consider n data instances $\{\boldsymbol{x}_i\}_{i=1}^n$ occupying a p-dimensional space $\forall i \in \{1, ..., n\}$: $\boldsymbol{x}_i \in \mathbb{R}^p$, where the data instances form the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] \in \mathbb{R}^{n \times p}$. Dimensionality reduction transforms the original data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ into a new representation $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times r}$ by identifying a mapping from the original p-dimensional feature space to a lower dimensional sub-space or manifold of dimension $r \in (0, p]$ [ZAZ+20].

Dimensionality reduction can help mitigate the *curse of dimensionality*, which refers to ubiquitous problems arising from high-dimensional data analysis in various fields of research [D+00]. For instance, as the dimensionality increases, the volume of the space expands rapidly causing data sparsity. Typically, the learning of structure in data relies on identifying groups or clusters of objects with similar characteristics based on for instance density or distance measures. In high-dimensional data, all objects tend to exhibit sparsity and dissimilarity in numerous aspects, rendering conventional data structure learning strategies inefficient. A range of other problems related to combinatorics, optimization and sampling also arise in high-dimensional data [BKH+21].

Fortunately, high-dimensional data often lie on a low-dimensional manifold embedded in the high-dimensional space [GMT20]. Dimensionality reduction works explicitly from this assumption, called the manifold hypothesis [MK20], and seeks to uncover the underlying manifold by discarding non-informative dimensions and thereby identifying (or constructing) relevant features.

Dimensionality reduction is useful for various tasks including visualization, noise reduction and as a general preprocessing step in data analysis [YZL+06]. Dimensionality reduction is especially useful in the context of large unstructured datasets, where visualization is hard and working on the entire data is computationally intractable. For this reason, computing a compressed representation which maintains the most salient information is highly useful. Although dimensionality reduction may bring various benefits and decrease computational complexity, it should not be applied too liberally. It must be ensured that a loss of important information does not occur during the reduction phase. The specific type of information that is important to maintain is problem specific and should be considered in conjunction with the chosen dimensionality reduction approach.

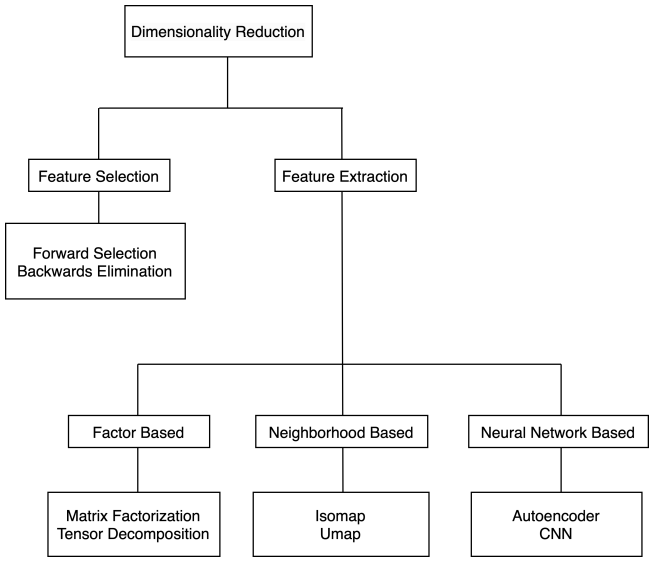


Figure 2.1: A high-level overview of dimensionality reduction. Note that several other techniques exist and that additional distinctions can be made to categorize dimensionality reduction methods. Some of these additional distinctions involve linear versus non-linear, supervised versus unsupervised and global versus local methods.

Dimensionality reduction is a rich field of research, and many approaches exists. Fig. 2.1 shows a simplified overview of the taxonomy of dimensionality reduction by broadly distinguishing methods between feature selection based and feature extraction based approaches. Dimensionality reduction techniques can be further divided into several subcategories not shown in the figure. These sub-categories include among others linear versus non-linear, supervised versus unsupervised as well as global versus local methods.

Linear dimensionality reduction uses a linear mapping of data to a lower dimensional space. Non-linear methods attempt to discover non-linear manifolds, where the data of interest is embedded.

Supervised learning relates to models where structured or labeled data is used to train a model to infer a function that can be used to map new data. For supervised learning each example is a pair of input data and desired output data. In contrast, unsupervised learning refers to methods where no data labels are used and where minimal human supervision is required. Large datasets, where dimensionality reduction is especially useful, are often unstructured with no apriori known labels or patterns in the data.

Global dimensionality reduction tries to preserve geometry at all scales, while local dimensionality reduction is concerned with preserving the local geometry of the data [ST02]. Global dimensionality reduction methods generally find representations that better reflect the global structure of the data but comes with a cost in computational efficiency, which is often higher than in local methods.

Feature selection works by selecting the most informative features without transforming existing features. This allows feature selection to retain the original structure and information of the data [KŁ16]. The most informative features can for example be selected based on a criterion for maximizing correlation to a target variable. In settings with large data volume feature selection faces problems relating to ease of deployment, computational efficiency, non-linearity and universality [BSDG16],[XLZ$^+$19]. Since feature selection involves a combinatorial optimization problem of finding the most informative features, the computational expense can be prohibitively high if the selection is performed on the global dataset. For this reason, research has been conducted on performing feature selection on a local sample space [ARK16]. Feature selection methods are popular for text classification where features might not be numeric. Feature selection methods include wrappers, filters, and embedded techniques.

Feature extraction generates new informative features by transforming existing features to a lower dimensional space. Popular feature extraction methods include principal component analysis (PCA) and linear discriminant analysis

(LDA) [BG98]. Feature extraction has been found to have higher discriminative power than feature selection [HG15]. However, feature extraction can cause a lack of interpretability in the generated features. This is problematic when a clear interpretation of important features is necessary.

There exists an ever-expanding array of dimensionality approaches designed to handle various problems. What follows is a description and evaluation of some of the most well-known and widely applied dimensionality reduction methods.

2.1 Factor Based Dimensionality Reduction

Factor analysis is a family of methods used to explain the variances among correlated variables by means of a smaller set of latent (unobserved) variables known as factors. Factor analysis prescribes to the manifold hypothesis by assuming that the variation in the observed variables predominantly stem from variations in a smaller set of underlying latent variables [Suh05]. Factor analysis aims to identify these shared variations and represent them mathematically as linear combinations of the factors, along with additional terms accounting for residual discrepancies (errors). The general setup for factor analysis can be expressed in the following form:

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{L} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the (typically standardized) original data matrix, $\boldsymbol{F} \in \mathbb{R}^{n \times r}$ is the factor score matrix, $\boldsymbol{L} \in \mathbb{R}^{r \times p}$ is the factor loading matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times p}$ is an error term. This expresses the original observed variables p through a smaller set of r common factors which are directly related to the original variables through the loading matrix $\boldsymbol{L}$.

2.1.1 Principal Component Analysis

PCA [VLG96] is perhaps the most well-known dimensionality reduction technique and is closely related to factor analysis. It works in an unsupervised manner by linearly transforming existing features and thus projecting data onto a lower dimensional sub-space where variation is maximized. This lower dimensional sub-space is formed by principal components, which can be computed from an eigendecomposition of the covariance matrix of $\boldsymbol{X}$, or from a singular value decomposition of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{2.2}$$

where the columns of $\boldsymbol{X}$ have been centered, $\boldsymbol{U}$ is a unitary matrix, $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values and the columns of $\boldsymbol{V}$ are called principal directions (eigenvectors).

Different from factor analysis, PCA does not distinguish between common and unique variance, but simply tries to account for variance in the observed variables [Bro15]. This makes PCA domain-agnostic and highly useful for reducing large sets of variables into a smaller surrogate set of variables across a variety of problems. Unfortunately, PCA is computationally intractable on very large datasets due to a complexity of $O(p^2n + p^3)$ [GVL13], where n is the number of observations and p is the number of variables. Additionally, since PCA is a linear method, it is not suitable if the data lies on a low-dimensional non-linear space. In these cases, non-linear adaptions of the PCA like kernel based PCA can be used. Furthermore, attempts to adapt PCA for large scale dimensionality reduction are also present in the literature [ZY18].

2.1.2 Other Matrix Factorization Techniques

Factor based dimensionality reduction is utilized in a number of related techniques, which all learn linear mixture representations of the original data under various constraints. These techniques aim to decompose the original data and approximate it via a low rank representation. Such techniques include non-negative matrix factorization (NMF) [LS99], independent component analysis (ICA) [Com94], sparse coding (SC) [OF96] and PCA. These methods are constructed with different constraints on the low rank representation, manifesting as learned representations imbued with different properties. For instance, the constraints imposed in PCA allow the model to learn features that account for the directions of greatest variance in the original data, while the non-negativity constraint in NMF allows its representation to capture constituent parts of the data and the constraints in ICA allow its representation to capture underlying statistically independent components of the data. These matrix factorization techniques are all highly flexible but lack interpretable features. A related technique, called archetypal analysis (AA) [CB94a, MH10], merges ideas from matrix factorization with clustering approaches (see section 3.2) to learn an interpretable representation, that captures distinct or archetypal corners of the data space. Some of these matrix factorization methods suffer from lack of a unique solution by consequence of rotational indeterminacy, as well as an unknown optimal number of components.

2.1.3 Tensor Decomposition

Tensors are multi-way generalizations of matrices into many dimensions. These objects serve as a compact way to represent high-dimensional data. A model based on tensors can generalize the 2-dimensional view provided by matrix factorization techniques and is particularly appropriate for real-world data demonstrating couplings across multiple axes.

Tensor decompositions work by extracting sets of smaller representative factor matrices and core tensors [Cic14] and thereby reduce the dimensionality of a dataset. Popular tensor based approaches include the Tucker decomposition [Tuc66] and the more restrictive PARAFAC decomposition [H$^+$70], which imposes a super-diagonal core tensor. Higher-order tensors can also be expressed as a connection of lower-order tensors, called tensor networks [Cic14]. These networks allow for the exploration of hidden structures in high-dimensional data by super compressing the original high-dimensional data to low-rank core tensors. Tensor based data reduction has wide applicability in areas such as anomaly detection, cluster analysis, feature extraction and predictive modeling. Tensor based approaches can however suffer from high computational complexity [uRLA$^+$16].

2.2 Neighborhood-graph Based Dimensionality Reduction

Neighborhood-graph based methods reduce the dimensionality of the original data through a two-step process [MHM18]. Firstly, they construct a weighted k-neighbors graph, which describes the distances between points in the original space and thus captures the underlying topological data structure. Secondly, they compute a projection to a low-dimensional layout of this graph, which constitutes the reduced data representation. This family of methods is popular for visualization of very high-dimensional data. Neighborhood-graph based approaches include among others, Isomap, t-SNE and Laplacian Eigenmaps [TSL00, VdMH08, BN01]. The fundamental differences between these methods lie in the details of how they compute the graph and the lower-dimensional layout [MHM18].

One of these neighborhood-graph based approaches called Uniform Manifold Approximation and Projection (UMAP) has garnered popularity in recent years. UMAP is a general-purpose non-linear manifold learning algorithm for high-dimensional data. It assumes that data is distributed uniformly on a locally connected and constant Riemannian manifold [MHM18]. The algorithm is

primarily used for visualization and is competitive with similar algorithms like t-SNE but can preserve more of the global data structure with less computational expense.

While neighborhood-graph based dimensionality reduction methods are powerful visualization tools, they suffer from a lack of interpretability compared to the related factor based approaches like PCA. This is because the embedding dimensions in for instance UMAP are not imbued with any special meaning. This is contrary to PCA, where the leading principal axes successively describe the directions of greatest variance in the original data. Moreover, distance based methods do typically not directly relate to the original data features, and as such do not provide the equivalent of factor loadings. Furthermore, as methods like UMAP do not directly model or intrinsically regularize against noise, it is possible for them to find manifold structure within the noise. Finally, neighborhood-graph based dimensionality reduction approaches tend to favor the preservation of local structure from local distances among data points. This contrasts with a method like PCA, which tend to prioritize the global structure of pairwise distances among all data points [MHM18].

2.3 Neural Network Based Dimensionality Reduction

The exponential growth of unstructured data has increased the complexity of data analysis and called for methods with strong generalization ability that can work free from human supervision. Advances in artificial neural network (ANN) research has provided such methods. Cleverly designed network architectures can perform dimensionality reduction and help identify and extract the most salient parts of a dataset. ANN approaches of special interest for dimensionality reduction include autoencoders and convolutional neural networks (CNN).

2.3.1 Autoencoder

Autoencoders have a long history [BK88, HZ94] and are acknowledged for their ability to learn efficient data representations through dimensionality reduction [HS06]. Architecturally, the network consists of an input layer, which feeds the data through a bottleneck of low dimensionality, called the encoder function $f(\boldsymbol{x}) = \boldsymbol{z}$. This function maps the input data $\boldsymbol{x} \in \mathbb{R}^p$ to a hidden layer describing an unobserved latent representation $\boldsymbol{z} \in \mathbb{R}^r$. To reconstruct the input data, this

latent representation is then passed through a decoder function $g(\boldsymbol{z}) = \hat{\boldsymbol{x}}$, which generates the reconstruction $\hat{\boldsymbol{x}} \in \mathbb{R}^p$.

Autoencoders are part of the ANN family and can be trained through mini-batch gradient descent; a variant of gradient descent that uses a small portion of the original data to compute an approximate gradient during tanning. Autoencoders may therefore gain knowledge from large datasets in a computationally efficient manner. Moreover, autoencoders are very flexible and can incorporate regularization and architectural constraints to modify the network to learn a specific representation. For instance, autoencoders are typically restricted to ensure they do not learn the identity function $g(f(\boldsymbol{x})) = \boldsymbol{x}$ and may be further restricted to ensure the representation embodies useful properties [GBCB16]. Autoencoders are trained by minimizing the reconstruction error between the input and output, for instance by measuring the squared errors:

$$\mathcal{L}_{\text{reconstruction}}(\boldsymbol{x}) = ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2^2 \tag{2.3}$$

The loss can then be minimized through a gradient descent algorithm using back propagation [RHW86].

A general weakness for autoencoders is the amount of training data needed to learn robust representations. Additionally, autoencoders also face issues with a lack of interpretability of their hidden representation. Adaptions of traditional autoencoders with various properties and constraints have been introduced in the literature to address some of these concerns [WHWW14, FCMR21, VLL$^+$10].

2.3.2 Varational Autoencoder

A variational autoencoder (VAE) [KW13a] is an adaption to the traditional autoencoder structure. VAEs use variational inference to create a probabilistic representation of the input data and to impose a distribution over the latent space. In this way, a variational autoencoder transforms the functions of the traditional autoencoder into probability distributions $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, where θ parameterize the distributions. The joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$ is defined by:

$$p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\theta(\boldsymbol{z}), \tag{2.4}$$

where $p_\theta(\boldsymbol{z})$ is the prior assumed over $\boldsymbol{z}$.

To achieve a model that describes the observed data well, the probability assigned

to the data can be maximized:

$$p_\theta(\boldsymbol{x}) = \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}, \tag{2.5}$$

where the optimal parameters are given by $\theta^* = \underset{\theta}{\operatorname{argmax}}\, p_\theta(\boldsymbol{x})$.

Unfortunately, computing $p_\theta(\boldsymbol{x})$ by marginalizing over $\boldsymbol{z}$ is computationally intractable, and similarly computing the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ is also intractable. The variational inference framework overcomes this problem by introducing an approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{z}|\boldsymbol{x})$. Thus, the variational autoencoder setup involves learning a probabilistic encoder, which computes the approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, and a probabilistic decoder, which computes the conditional likelihood distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$.

To train the network a differentiable loss function is needed. The evidence lower bound (ELBO) provides this:

$$\mathcal{L}_{\phi,\theta}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log\, p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathcal{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})), \tag{2.6}$$

where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback–Leibler divergence, which measures the distributional distance between the approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the prior $p_\theta(\boldsymbol{z})$. Typically the prior is chosen to be an isotropic Gaussian. Maximizing Eq. 2.6 is equivalent to maximizing the quality of the reconstruction and at the same time minimizing the divergence between the approximate posterior and prior distribution. The latter ensures that the latent space is well organized. Eq. 2.6 can be maximized using back propagation and the reparametrization trick [KW13b, RMW14].

The fundamental difference between traditional autoencoders and VAEs is that the VAE latent space is designed to be continuous, which allow for substantial control over the latent representation. For example, imposing isotropic Gaussian distributions can make VAEs useful for finding representations with disentangled properties [HMG+16]. Additionally, the continuous distribution of the latent space makes VAEs useful for generative purposes and the extraction of essential data by allowing sampling from the latent space, which ideally represents the intrinsic structure of the training data. For instance, in [SWXS18] they propose a framework for learning sparse representations of the intrinsic structure of the input space for large-scale and high-dimensional data based on the latent space of a VAE. The compact data representation is then used for anomaly detection.

2.3.3 Convolutional Neural Network

A convolutional neural network (CNN) is a feed-forward feature extraction model consisting of an input and output layer, as well as a number of hidden layers that perform convolutions [FMI83, LBBH98a]. CNNs work by utilizing successive convolutional layers to extract global and local features from the input. The convolutions are typically carried out by taking the dot product of a convolution kernel with the input matrix of a layer. The kernel is then moved along the input matrix to generate a feature map, which is used as input to the proceeding layer. Pooling layers (typically average or max pooling) are often used following convolutional layers. These layers down-sample the feature maps. Max pooling can be interpreted as feature selection, while average pooling can be interpreted as feature extraction. In this way a CNN can combine feature extraction and feature selection. This results in a powerful framework that successfully constructs complex patterns by using relevant filters. Similarities exist between CNNs and the visual cortex of animals and humans, where neurons only react to stimuli from a restricted part from the visual field [VDKDG21].

ANNs utilizing a mix of feed-forward and convolutions layers have seen great success when applied to image and pattern recognition tasks. Even further, these types of layers can be incorporated directly into autoencoder based neural networks to channel the strengths of the various techniques into a combined framework.

Chapter 3

Numerosity Reduction

Numerosity reduction reduces the size of a dataset by reducing the number (numerosity) of data instances. That is, in contrast to dimensionality reduction, numerosity reduction does not operate on the dimensionality of the data, but rather on the cardinality [Gho21]. This is typically accomplished by sampling informative instances, or by summarizing multiple data instances into fewer objects. Again, consider n data instances $\{\boldsymbol{x}_i\}_{i=1}^{n}$ occupying a p-dimensional space, where the data instances form the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] \in \mathbb{R}^{n \times p}$. Numerosity reduction refers to a reduction in the number of data points in $\boldsymbol{X}$ from n to $m \in (0, n]$ [GC19]. In short, we are looking for informative instances that allow us to construct a new representation $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{m \times p}$.

Numerosity reduction is useful when dealing with large-scale datasets containing an excessive number of instances. In this setting, numerosity reduction can lead to smaller representations on which downstream modeling can be performed to alleviate computational costs of training, inference, and storage [HKP12]. Additionally, large datasets usually contain several noisy or redundant instances, that are either in-significant to, or may even hinder the learning process. Finally, data imbalance is common in ML applications and here numerosity reduction may be used to mitigate inherent data biases [PKDN15].

Like dimensionality reduction, numerosity reduction is a rich field of research and encompasses a spectrum of techniques. Numerosity reduction contains among others sampling based approaches and summarization based approaches [HKP12, KDS+22]. Some works also consider numerosity reduction methods that do not directly reduce the number of instances, but rather represents the data in a compact form. These methods include histograms and parametric numerosity reduction, where a model or data distribution is assumed and the parameters of the model are estimated and saved in place of the original data [KTS+12, HKP12]. In this thesis only instance based numerosity reduction is considered, where the number of data instances is directly reduced.

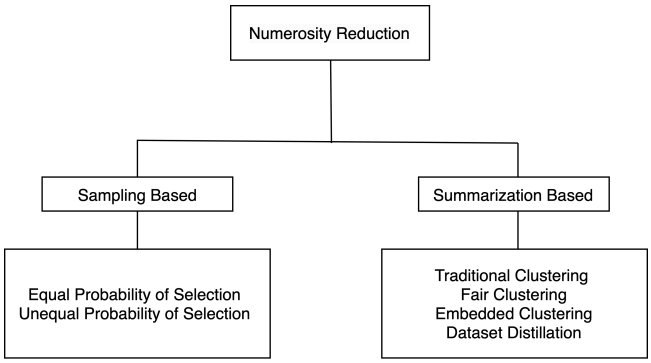


Figure 3.1: A high-level overview of instance numerosity reduction. Note that other types of numerosity reduction methods that do not distinctly reduce the number of instances also exist. These include methods like histograms and parametric numerosity reduction.

Sampling based approaches reduce the data size by selecting a representative[1] subset of data instances [KTS+13]. This type of numerosity reduction may be viewed as instance pruning [KDS+22], where non-informative or redundant instances are removed. Typical sampling based methods include simple random sampling (SRS) and stratified sampling (see Sect. 3.1). The selection of a subset of 'natural' non-modified elements from the original data matrix suggests that sampling based numerosity approaches share similarities with feature *selection* in dimensionality reduction, where an informative subset of the original features is selected.

On the contrary, summarization based approaches create smaller representations by summarizing, aggregating, or collapsing multiple data instances into fewer objects [KDS+22]. This is typically achieved by creating representative prototypes

[1] Here and throughout this chapter, *representative* is a vague term which may cover numerous modalities of representation. The topic of data representativity will be explored further under the research outcomes presented in Part II in Sect 4.2

that capture the characteristics of multiple data instances using for example clustering techniques (see Sect. 3.2). These new objects may be synthetic in nature and thus not exist in the original data matrix. A typical example of this includes k-means clustering, where cluster centers are mean values of data instances. In this way, synthetic numerosity summarization suggests similarities with feature *extraction* in dimensionality reduction, where new informative features are constructed. However, summarization based approaches can also use natural data instances from the original data matrix, for instance in k-medoids clustering, where multiple data instances are collapsed and represented by single data instances (cluster centers) from the original data matrix [KDS+22].

3.1 Sampling Based Numerosity Reduction

Sampling is the idea of selecting a subset of your data, often with the aim that the sample can represent the original data. Sampling is immediately useful for improving computational speed, but optimal sampling has also been found to improve accuracy in various machine learning algorithms [ZZS13]. In the context of large datasets, sampling can be crucial as parameter estimation on the complete dataset is sometimes intractable. For this reason, sampling can be an essential step in reducing the size of the data to make processing and downstream modeling possible. Furthermore, in cases where it is possible to perform analysis on the original dataset, this might not be necessary. The performance of classification algorithms has been found to be only slightly worse on sampled data, while the execution time is greatly reduced [Alb16].

Sampling is not a trivial task, and the type of sampling used can alter the results obtained by introducing different biases and thereby invalidating statistical inference. The combined errors introduced during sampling are often orders of magnitude higher than errors associated with subsequent analytical steps [PME05b]. For this reason, careful steps should be taken to ensure samples are representative[2].

Sampling can be divided into probability and non-probability sampling [Sha17a]. In probability sampling every unit has some well-defined probability of being selected, while in non-probability sampling, like convenience sampling [Sed13], this criterion is not met. Probability sampling is generally preferred when possible [Sha17b].

[2]Again, this should be coupled with an evaluation of what *representative* means, and an estimation of the extent to which the sample meets this requirement (see Sect. 4.2).

Samples are most often designed to represent or reflect the population distribution or certain population characteristics [Tho12]. This is for instance the case when conducting population surveys, where a subset of individuals is selected to represent the overall population and allow for statistical inference on population parameters. Here, the goal is often to obtain an unbiased sample, typically thought of as a sample where all objects share the same selection probability [PM78, Pat16b]. It should be noted that the definition of an unbiased sample can be unclear [Pat16b] and can either refer to a sample of unbiased (equal) selection probability, also known as an equal probability of selection (EPS) sample [KS10, ABM08], or a sample from which an unbiased estimate of a population parameter can be obtained [WSGG12]. A parameter estimate $\hat{y}$ is obtained by applying a statistic ψ on a sample $\boldsymbol{y}$:

$$\hat{y} = \psi(\boldsymbol{y}). \tag{3.1}$$

The bias of the estimate relies on the interaction between the sample and statistic. An estimate is unbiased if its expectation is identical to the true value of the parameter that is being estimated. An unbiased estimate of a population characteristic (such as diabetes prevalence) can be directly computed by finding the corresponding prevalence on a EPS sample [PE95]. While an EPS sample allow direct inference about population parameters, it is also possible to obtain unbiased parameter estimates from samples of unequal probability of selection (UPS samples) [WSGG12]. However, extrapolating findings from a UPS sample to the population is more complicated and typically requires a form of weighting [MK17]. Despite this, UPS sampling can have various benefits for downstream modeling by mitigating data imbalance or by allowing models to focus on certain informative observations [BKH18, WNC05].

What follows is an overview and evaluation of some of the most popular EPS and UPS sampling techniques that reduce the original data into a smaller form which may represent distinct aspects of the original data source.

3.1.1 Equal Probability of Selection Sampling

EPS sampling is useful when the goal is to extrapolate results from a sample to population [PE95], i.e., the goal is to obtain a sample that acts as a miniature of the population, and from which population characteristics can be inferred. The most commonly used EPS sampling technique is simple random sampling (SRS) [SS03], where each unit is selected with equal probability. This sampling technique has the advantage that it requires little knowledge of the whole population and introduces minimal bias [APSN13]. However, other more powerful sampling techniques have been found to generate comparative or better insights

on data [RKRD17], particularly if the focus is on analyzing specific characteristics of the data.

Stratified random sampling involves stratifying the dataset into homogeneous groups (strata) sharing common traits like race or age and drawing a random sample from each stratum [APSN13]. This type of sampling can ensure that all strata are represented in the final sample and can allow estimation and comparisons across strata. However, stratified random sampling is not well suited if limited knowledge of such strata exists. Additionally, when using this sampling scheme, a sampling fraction for each stratum must be defined. If this sampling fraction is common across all strata, then the stratification generates an EPS sample, which ensures that the sample mimics the population with respect to the stratifying factor [PE95].

3.1.2 Unequal Probability of Selection Sampling

Unequal probability of selection (UPS) sampling occurs when units are not sampled with equal probability. This type of sampling complicates inference about population characteristics, but can empower downstream models trained on the sampled data, by mitigating data bias or by focusing on informative observations. The objective of the sampling should be given careful attention, whether it is to create a representative unbiased sample for population inference, or if the sampling is used deliberately to target specific useful data instances.

Data imbalance can manifest both in categorical [HYS+17, KŁ16] and continuous [BTR17, BTR19] data and can be analyzed from the perspective of the target or input distribution. The most studied scenario involves imbalance in categorical targets where the majority class vastly outweigh the minority class, posing problems for machine learning algorithms. Deliberate use of biased sampling to counteract the effects of imbalance in large datasets has been studied in [BKH18], where it was found that random under-sampling techniques could effectively increase accuracy of models trained on the biased samples as opposed to the full dataset. Under-sampling randomly discards data from the majority class, while oversampling randomly over-samples from the minority class. The imbalance problem is widely studied in the data mining and ML communities and learning from these imbalanced sets is becoming increasingly important in various fields such as medical monitoring and fraud detection [CJK04].

Other UPS sampling methods are inspired by the field of active learning, where informative data points are used in the sampling with the aim of achieving high accuracy on ML models. These methods have been found more effective than random sampling especially on imbalanced datasets [WNC05], [ZZS13].

3.1.2.1 Coresets

Coresets are defined as problem specific small sets of (usually weighted) points that accurately represent the original (big) data, such that running a model on the coreset will result in a solution that is provably close to the solution on the full dataset [FSS20]. This means that for the correct coreset construction, applying for instance k-means clustering on the coreset will approximate the results of performing k-means clustering on the full dataset.

Coresets are often designed to be composable, meaning that the union of a pair of coresets is also a coreset for the underlying input [FSS20]. This means that these coresets can be constructed on small subsets of the original data independently, making the method suitable for distributed algorithms and streaming environments. Coresets adhere to the data-centric perspective in the analysis of large data sources, where the focus is moved away from computational efficiency of existing methods towards the reduction of big data. Coresets have been demonstrated for various base problems relating to clustering, classification, and regression, but certain problems have also admitted impossibility results, demonstrating that the specific task does not admit a strong coreset (a specific type of coreset that approximates the cost function loss for every query of a problem, not just the optimal one). One such problem is logistic regression [MS18, MSSW18].

Coresets can be constructed in several ways, but usually involve non-uniform importance sampling, where important data instances are assigned weights and sampled accordingly [Fel20]. However, the first coresets used for covering problems were based on cluster summarization [AP03].

Coresets are typically not employed to derive insights about population parameters or to discover latent patterns in the data, but rather used to reduce the computational complexity of performing certain base tasks on the original data source. This means that coresets are very task dependent and the construction for new tasks can be hard to design [Fel20].

Finally, coresets originate from computational computer science and computational geometry, where strong theoretical guarantees are customary but where there is usually no assumption that a given dataset is a set of independent and identically distributed (i.i.d.) samples from an underlying distribution. This means that unlike in ML, the topic of generalization error is usually given little attention in classical coreset research [Fel20], and consequently they are seldom evaluated on an independent test set. [CGS18] seek to address this gap by introducing Wasserstein measure coresets, which aim to minimize the generalization error on the coreset with respect to the distribution of the original data.

This is achieved by minimizing the distributional distance from the original dataset to the coreset via the Wasserstein metric. However, the effectiveness of the approach is limited to the available statistical knowledge of the underlying (possibly unknown) data distribution.

3.1.2.2 Determinantal Point Process

A determinantal point process (DPP) [KT$^+$12a] is a sampling based technique, which describes a probability distribution over subsets of the original data matrix that can be used to sample diverse sets. DPPs originate from random matrix theory [MG60] and physics [Mac75], where they were first used to describe the repulsive forces existing between fermions (a category of elementary particles containing for instance electrons), which by the Pauli exclusion principle cannot occupy the same quantum state. This repulsive force is modeled precisely by the negative correlations expressed by a DPP.

Recently DPPs have gained attention in the ML community due to their ability to draw highly diverse sets of points across a range of domains and data modalities including videos, images, documents, recommendations systems and sensors [GCGS14a, KT$^+$12a, LB12a, ZKL$^+$10a, KSG08a].

More specifically, DPPs describe a distribution over subsets, such that the sampling probability of a subset is proportional to the determinant (hence determinantal) of an associated sub-matrix of a positive semi-definite kernel matrix. The kernel matrix describes the similarity between feature vectors of the original items through a kernel function $\boldsymbol{G}_{ij} = g(\boldsymbol{v}_i, \boldsymbol{v}_j)$. The kernel is used to model negative correlations between items ensuring that items of similar feature vectors are not likely to co-occur. Various types of kernels are available that each bring their own advantages. One such kernel is the radial basis function (RBF) kernel $\boldsymbol{G}_{ij} = e^{-\gamma||\boldsymbol{v}_i - \boldsymbol{v}_j||^2}$, where $\gamma = \frac{1}{\sigma^2}$ and σ is a free parameter indicating the length scale of the kernel.

The most popular kernel is the linear kernel, which results in a similarity function of inner products known as the Gram matrix $\boldsymbol{G}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. The linear kernel can be constructed from the original data matrix in the following way [Gha21]:

$$\boldsymbol{G} = \boldsymbol{X}^T \boldsymbol{X}, \tag{3.2}$$

where $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ is the Gram matrix and $\boldsymbol{X} \in \mathbb{R}^{p \times n}$ is the original data matrix. Note that the data matrix has the n data items as columns in this setup. The determinant of the Gram matrix expresses the (squared) volume of the parallelotope (a generalization of parallelograms to arbitrary dimensions) formed by the feature vector columns of the original data matrix $\boldsymbol{X}$.

In a DPP, the sampling probability of a subset of items Y from $\boldsymbol{X}$ is proportional to the determinant of an associated sub-matrix $\boldsymbol{G}_Y$ in $\boldsymbol{G}$ [KT$^+$12a]:

$$P_{\boldsymbol{G}}(Y) \propto \det(\boldsymbol{G}_Y), \tag{3.3}$$

where the sub-matrix $\boldsymbol{G}_Y$ refers to the restriction of entries in $\boldsymbol{G}$ indexed by the items in Y. Per definition $\boldsymbol{G}_Y = [\boldsymbol{G}_{ij}]_{i,j\in Y}$ and thus $\det(\boldsymbol{G}_Y)$ is a minor. Due to the geometric nature of the linear kernel, the sampling probability of the subset Y is directly proportional to the $|Y|$-dimensional volume spanned by the feature vectors of the subset. This causes the sampled items to express a high volume and thus a high (geometric) diversity.

3.2 Summarization Based Numerosity Reduction

Summarization based numerosity reduction reduces the original data by representing or collapsing numerous data points into groups, which can be represented by single entities or prototypes. This is typically achieved through clustering algorithms, which seek to identify patterns in the data and to partition the dataspace into groups or clusters with high similarity based on some notion of distance [HPT22]. Each of these clusters may then contain a representative object to which the cluster items are mapped. Each cluster is typically designed to contain a subset of items, where each item expresses high similarity to other items in the same cluster, and high dissimilarity towards items in other clusters. Creating reduced representation in this way can bring benefits related to visualization and interpretation of complex data but can also be used as a preprocessing step for downstream modeling [KDS$^+$22, KU12].

3.2.1 Traditional Clustering Methods

Some of the most common clustering techniques include k-means and k-medoids clustering. These techniques aim to minimize the distances (maximize similarity) between the original items and their assigned cluster centers. K-means clustering determines the representative object $\boldsymbol{\phi}_j$ for the j^{th} cluster based on the mean value of items in the cluster, while k-medoids constrains the representative object to be part of the original dataset. The loss for these algorithms may be expressed as [CY10]:

$$\mathcal{L}_{\text{k}} = \sum_{j=1}^{k} \sum_{\boldsymbol{x}\in C_j} d(\boldsymbol{x}, \boldsymbol{\phi}_j), \tag{3.4}$$

where $d(\cdot)$ is a metric (distance function), $\boldsymbol{x}$ is a data point in cluster C_j, and $\boldsymbol{\phi}_j$ denotes the representative object of C_j to which the data point $\boldsymbol{x}$ is mapped [CY10]. Let D be the set of all data points, then for k-medoids clustering, $\boldsymbol{\phi}_j \in D$ and for k-means clustering $\boldsymbol{\phi}_j \in \mathbb{R}^m$ for $D \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. The distance metric in k-means is the squared Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^2$ [BGK+18].

These techniques can be used as a preprocessing step to reduce the original dataset into a representative set of cluster prototypes on which downstream modeling can be performed. This can significantly reduce the computational load of for instance power market models [KU12]. However, it should be carefully evaluated what trade-offs this reduction entails, both in terms of a potential loss in predictive performance and the efficiency of the reduction.

3.2.2 Fair Clustering

As in sampling, clustering may end up propagating or amplifying inherent data biases from the original population to the reduced representation. This can ultimately cause biased inference for models trained on the representation [CMM21] and is critical if the clustering is performed on demographic data to determine for instance job or loan applications. Bias mitigation efforts have been considered for clustering algorithms, where the clustering objective is changed to incorporate notions of fairness. Various notions exists and are primarily divided between group-level and individual-level fairness [CMM21]. In group-level fair clustering a notion of balance is usually adopted as the fairness metric. This notion considers a given clustering fair if the balance of a protected or sensitive attribute like race is preserved in the final clustering. In fair clustering these sensitive attributes are represented by colors $p \in P$ assigned to each data point. To formulate this, consider a set of points D partitioned into a set of clusters C. Then the balance may be measured by comparing two fractions $r_{D,p}$ and $r_{c,p}$ indicating the color proportion in the overall dataset D and the color proportion in a given cluster $c \in C$. These proportions can be used to construct an overall balance fraction $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ used to define the balance of the complete cluster solution:

$$B = \min_{c \in C, p \in P} \min \left(R_{c,p}, \frac{1}{R_{c,p}} \right), \tag{3.5}$$

where B is the balance and $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ is a fraction for a given cluster c and color p [CMM21]. By construction $B \in [0, 1]$ with higher balance being fairer. Optimal balance ($B = 1$) is achieved when every cluster share the same color fraction $r_{c,p} = r_{D,p}\ \forall c, p$, and complete imbalance ($B = 0$) is obtained when a single cluster becomes monochromatic containing no members of a protected group $r_{c,p} = 0$.

Group-level fairness is considered for k-center and k-median clustering in [CKLV17] using a notion of balance. They provide a solution to the problem by decomposing the original dataset into a reduced representation of fair micro-clusters called fairlets. Traditional k-center or k-median clustering can then be performed on the reduced representation to obtain a final cluster solution that is optimal under the fairness constraint.

3.2.3 Embedded Clustering

The notion of distance (or dissimilarity) is central to clustering algorithms. This notion relies both on the choice of distance metric as well as the chosen representation of the data instances in feature space. A standard choice may consider the Euclidean distance in the original p-dimensional feature space of the data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. However, this choice is not necessarily optimal for data partitioning, as some features may express more discriminative capacity than others [KU12]. Moreover, as alluded to in Chapter 2 on dimensionality reduction, high-dimensional data can cause sparsity in the sample space manifesting as observations that are dissimilar in various ways. These problems can be overcome through dimensionality reduction as a preprocessing step, where clustering may then be performed in an embedded feature space [XGF16, KU12] enjoining numerosity and dimensionality reduction into a combined framework. This approach can be implemented on a deep neural network foundation to jointly optimize the embedding space and clustering objective.

In [XGF16] a deep autoencoder neural network is used on image data to learn an encoding from the original data space $\boldsymbol{X}$ to a lower dimensional latent space $\boldsymbol{Z}$. The decoder is then discarded, and the latent space coordinates are used as the input to a KL divergence based clustering scheme. Other works [GBH+22] based in explainable AI (XAI) adopt a similar structure in a self-explainable VAE model and create a reduced representation of similarity scores between observations and optimized clusters representatives in latent space. This reduced representation is then fed as input data to a linear classifier in a supervised classification scheme to corroborate overall model predictions with explanations based on similarity between input images and learned prototypes. Here the decoder is not discarded, but rather maintained to provide decoded images of prototype appearances, increasing the transparency of the model.

3.2.4 Dataset distillation

Dataset distillation (also referred to as dataset condensation) is a recent domain of research that has attracted attention in the deep learning community. These techniques aim to distill the knowledge of a large complex training dataset (denoted as $\mathcal{T}$) into a smaller dataset of a few informative synthetic training instances (denoted as $\mathcal{S}$) such that training a model on the distilled data yields approximately the same results as on the original data.

Dataset distillation is similar to coreset selection but disregards the typical coreset restriction of uneditable data instances [YLW23], allowing for highly reduced representations to be learned without significant loss in accuracy for downstream models. Dataset distillation is typically performed by either treating the distilled data as a hyperparameter in a data learning network, called the meta-learner framework, or by matching the influence of the distilled and original data on trained models, called the data matching framework [LT23].

In the meta-learner framework, an inner learning algorithm is injected into an outer data learning algorithm. A bilevel optimization scheme is then carried out to simultaneously i) optimize the parameters of the inner algorithm by minimizing the (training) loss on the distilled data $\mathcal{S}$ in standard supervised fashion, and ii) update the distilled dataset $\mathcal{S}$ learned in the outer algorithm to minimize the (validation) loss of the inner model on the original dataset $\mathcal{T}$. The meta-learner framework has demonstrated powerful performance, but the bilevel optimization is computationally expensive by consequence of the required second-order derivative computations.

In the data matching framework, the synthetic dataset $\mathcal{S}$ is optimized to reflect the influence of the original training dataset $\mathcal{T}$ on model training. This influence can be measured in various informative spaces such as parameter, gradient or feature spaces. In [CWT+22] they match the training trajectory on the distilled and original data by minimizing the distance between learned parameters at each learning step. This allows them to learn a small set of informative images that distill popular image datasets such as CIFAR-10 and ImageNet [KH+09, RDS+15]. However, the learned reduced representation from such approaches has been found to exhibit large discrepancies to the distribution of the original data. In light of this, [ZB23] presents an approach to match the data distribution of $\mathcal{T}$ and $\mathcal{S}$ in embedded feature spaces.

While dataset distillation techniques remain powerful tools for speeding up neural network architecture searches [EMH19] and preventing catastrophic forgetting in neural networks [ZNB22] (continual learning [KPR+17]), they pose certain limitations.

Dataset distillation aims to distill knowledge from the original dataset into the learned synthetic representation. However, knowledge is an abstract concept, and the difference in expressed knowledge is often measured indirectly through proxy losses [LT23]. The interpretation of what knowledge entails guides the choice of proxy loss and ultimately affects the results. Under the meta-learner framework, the knowledge difference between $\mathcal{T}$ and $\mathcal{S}$ is measured by the loss on the original dataset $\mathcal{T}$ for a learner trained on $\mathcal{S}$, and in the data matching framework the knowledge difference is measured through similarity of a series of informative spaces such as parameters or features. No consistent definition has been cemented for the concept of knowledge, which remain a limitation for the underlying theory of the techniques [LT23].

Part II

Research Outcomes

Chapter 4

Summary and Discussion of Research

This chapter presents and summarizes the research outcomes of the thesis. The research outcomes fall into the three categories of dimensionality reduction, data representativity and numerosity reduction. Sect. 4.1 presents papers A and B studying dimensionality reduction approaches for spectral and image data. Sect. 4.2 present paper C, which concerns data representativity. Sect. 4.3 presents papers D and E on numerosity reduction in manufacturing and population based data.

A large portion of current research in data reduction (such as coresets and data distillation techniques) centers on ideas from the perspective of reducing the data, while retaining the knowledge, structure, or influence of the original data. This type of knowledge extraction attaches the performance of models trained on the reduced data to the performance of models trained on original data and may not be optimal if the original data is noisy or imbalanced. Here, the original data may encode hidden information, which could be lost if the reduced representation is sought matched to the original data. In such settings, the data reduction pipeline can be imbued with certain properties that accommodate the dataset and task at hand, such as i) for imbalanced data a diverse representation and balanced inference for downstream models is preferred or ii) for noisy data, the data may originate from variations on a few underlying components and the reduced representation should encapsulate these while disregarding noise

(similar to factor analysis). Finally, a third general property may be suitable to address growing concerns against ML transparency; iii) in situations where the data reduction pipeline modifies the input instances, the modification should be transparent, and the learned representation should be interpretable. The research contributions of this thesis aim to develop novel data reduction techniques for noisy or imbalanced datasets that imbue the data reduction pipeline with those properties, so as to reduce the data in a more appropriate manner for the given dataset and situation.

4.1 Dimensionality Reduction Contributions

Dimensionality reduction can be used to encapsulate the original data in a smaller representation by extracting salient structure that is not readily discernible. Such reduced representations can aid not only in data visualization but can also be used as training data for downstream tasks. Dimensionality reduction approaches have been applied to various problems in the literature with great success, particularly using powerful deep learning based frameworks such as autoencoders. Nevertheless, these models have received criticism concerning model opaqueness and a consequent lack of interpretability of the latent space [FLTW20, LLL+23]. This thesis presents two autoencoder based research contributions addressing these concerns. These contributions demonstrate how the specific design as well as the information extracted by the networks differs in accordance with the scientific domain of the application.

Paper A
TAU: A neural network based telluric correction framework

Dimensionality reduction approaches have spread to various scientific domains and proved a vital tool in extracting information from large unstructured datasets. One such field of research is astrophysics, where vast quantities of data are being gathered every second by powerful instruments. These data often contain faint signals submerged in a sea of multifaceted noise. Understanding, modeling, and correcting for this noise is critical in obtaining the precision required to detect the dim signals emitted by distant celestial objects (such as Earth-like exoplanets).

The majority of astrophysical data is gathered by ground-based telescopes, where a significant part of the noise budget originates from absorption of photons (light) in the atmosphere of the Earth. This type of absorption is known as telluric absorption (or telluric contamination) and acts by obscuring the signal of interest.

Sophisticated synthetic approaches based on advanced physical models have been employed to model and remove the telluric contribution from observed spectra [SSN+15, KNS+15]. However, these models are based on our current understanding of the underlying physical interactions and are limited by imprecision in the knowledge we have on external factors to an observation such as molecular line lists, which are required to compute radiative transfer solutions (solutions for the propagation of light through a medium such as an atmosphere). Data-driven empirical initiatives have been explored to circumvent this problem by directly working on the observed data. One such initiative uses PCA [AADD+14] in an effort to decompose observed spectra into their underlying noise and signal components. This approach however has various drawbacks including a lack of flexibility, computational efficiency, and interpretability.

The first research contribution of this thesis is the introduction of a novel autoencoder based approach for telluric correction, which addresses the various intricacies of the problem and outperforms the state-of-the-art synthetic telluric correction approach at a significantly reduced computational expense. We call this framework `TAU` (Telluric AUtoencoder). `TAU` is inspired by the discipline of hyperspectral unmixing, where mixed spectral components are disentangled from hyperspectral data. The network incorporates prior information of the underlying physics and enforces this on the learned representation with various constraints to achieve high transparency and interpretability of the latent space and extracted components.

We train the network on observed solar spectra from the HARPS-N [HAR] (High-Accuracy Radial-velocity Planet Searcher for the northern hemisphere) spectrograph mounted on the 3.6 meter telescope Telescopio Nazionale Galileo (TNG). This instrument has been observing the Sun every 5 minutes of clear skies for the past 5 years amounting to approximately 75,000 observations [DCS+21]. Each of these observations contain signal across the optical wavelength regime [3830 Å - 6930 Å] split into 69 spectral orders each containing 4096 pixels. This amounts to a total of $69 \times 4096 = 282{,}624$ pixels (or features) per observation. We operate spectral order wise and compress the data in the autoencoder network from a dimensionality in the input layer of $p = 4096$ to a dimensionality in the latent space of $r = 3$.

Each observation contains an unknown mix of the inherent solar signal and additional contamination (absorption) from H_2O and O_2 molecules in the atmosphere of Earth. The network is designed such that the three-dimensional latent space represents the individual component abundances of the solar, H_2O and O_2 components. After network training, the telluric H_2O and O_2 spectral signatures can be extracted from the corresponding weights of the decoder.

The approach thus finds a highly reduced representation by structuring the mixed signals in the original data into individual disentangled components and associated abundance weights, which can be used for downstream telluric removal on new observations. We publish the network code and extracted components as an open-source code project to aid scholars in applying `TAU` on their own data.

Paper A mainly addresses research objectives 1 and 3. The paper presents an approach for data reduction in natural sciences where the essential information to extract typically relates to physical phenomena governed by physical laws. We demonstrate how to distill this information by respecting known physical properties of the system and integrating them directly into the design of the network to achieve a high transparency and interpretability of the data reduction pipeline. The H_2O and O_2 components from the reduced representation can be used as training data for downstream models to learn the correct telluric abundances and remove the telluric contamination.

Paper B

Pantypes: Diverse Representatives for Self-Explainable Models

Image classification is another field of research where dimensionality reduction has proven highly effective. Here dimensionality reduction is typically used to learn useful representations in a black-box nature through a cascade of encoding layers. The reduced representation can then be fed as input to a downstream classification algorithm to obtain the class predictions. Such algorithms are increasingly used to support important decisions, which has sparked a growing demand for higher model transparency. To meet this demand, the field of explainable AI (XAI) has emerged. This field uses a variety of approaches to achieve higher model interpretability, transparency and trustworthiness. One such approach, known as a prototypical self-explainable classifier, is based on embedded clustering. Instead of forming predictions directly from the full latent space, these classifiers construct a reduced characterization of prototypical class representative objects and measure the latent space similarity between observations and prototypes. The class representative objects are designed to capture sub-variations within the classes such as different styles of handwritten digits in the MNIST dataset [LBBH98b] or variation in facial features in individuals from facial image datasets. The similarity scores can then be fed as input to a linear classifier, allowing the model to derive and explain inference based on the prototype similarities [GBH+22].

Despite a growing effort in XAI to construct transparent self-explainable models (SEMs), there is still a lack of attention given to the diversity expressed by the reduced representation in latent space.

Diversity is typically ensured by forcing the learned prototypical objects to capture non-overlapping information [VL20] through latent space orthogonality constraints. However, non-overlapping orthogonal information can still be learned in a small region of the input space, leading to reduced interpretability and data representation by the prototypes. Image classification models have historically been trained on biased data, where certain majority sub-populations of the data vastly outweigh minority sub-populations [BG18b]. Without sufficient bias mitigation efforts, this bias can propagate into the learned prototypes and ultimately cause biased inference.

Paper B presents the second research contribution of this thesis, which addresses these issues by introducing a new family of prototypical objects for SEMs designed with high latent space diversity in mind. We call these objects *pantypes* and promote their construction through a novel volumetric loss inspired by the theory of determinantal point processes. We implement the pantypes in a self-explainable variational autoencoder model [GBH+22]. Pantypes can empower prototypical self-explainable models by occupying and representing dissimilar regions of the latent space leading to better model interpretability and fairness.

Pantypes share a number of similarities with archetypal analysis (AA), which seek to learn distinct archetypal aspects of the data, such that data instances can be described by a convex combination of the archetypes. However, the linear combination part of AA reduces its applicability for explaining feature spaces generated from non-linear combinations (such as image transformations). Non-linear reformulations of AA have been presented in the literature [vDBA+19], which is achieved through an AA regularization in a neural network autoencoder structure to learn the data geometry from non-linear input data. The archetypes in this work represent the corners of the data geometry, and the original data can be described by combining these representative objects. On the other hand, our pantypes forego the need for combination, and instead describe distinct and interpretable objects that are directly observed in the data. As such pantypes not only occupy and directly represent the hull (exterior) of the dataspace, but also the distinct interior aspects.

Paper B relates to research objectives 1, 3 and 4. The paper presents an approach for learning a reduced representation of the training image data, where the objective is to distill information that can be fed to an integrated linear classifier. This data representation is learned through a transparent network and is designed to not only allow the classifier to obtain high accuracy, but to also mitigate inherent data biases by covering the input space (in a geometric sense) and representing the distinct aspects of the input distribution.

4.2 Data Representativity Contribution

The various approaches for data reduction share the same overarching goal of finding smaller surrogate data representations which may be used to derive insights about the original data. The nature of these insights varies and may be extracted from different representations imbued with certain notions of representativity. At its core, representativity expresses the capacity for one thing to stand for another. Claims of representation (or mis-, non-, over-, and under-representation) are ubiquitous in scientific as well as non-scientific discussions [CL21] and mask a web of ideas and notions. Ultimately, representativity lie at the heart of inductive science and inferential statistics, where we are imposed to infer knowledge about unobserved entities in the broader world by applying mathematical methods to inevitably finite observed data. We assume the observed may stand for and represent the unobserved. Thus, representativity becomes a matter of generalizability. However, we must recognize that generalizability is itself a term packaging a multitude of meanings from inference of population parameters in samples, generalization of knowledge transferred from one population to another, and extrinsic generalization of learned concepts from experiments to real-world phenomena [CL21].

The setup for supervised ML involves a learning algorithm trained on a training set $\{\boldsymbol{x}_i^{(\text{train})}, y_i^{(\text{train})}\}$. The generalization performance for the learning algorithm on new unseen data is then evaluated on a test set $\{\boldsymbol{x}_i^{(\text{test})}, y_i^{(\text{test})}\}$ assumed to be drawn from the same probability distribution $p(\boldsymbol{x}, y)$ as the training data [WLZ+16, GBCB16]. However, in many real-world situations certain examples may be highly underrepresented or even absent in the training data due to effects such as sample selection bias or distributional shifts. This distributional departure can result in biased inference and decreased generalization performance [CHUK18]. The distribution of the training data may be artificially modified (as in over- or under-sampling efforts) to accommodate predictive inference on sub-populations, or a target population assumed to be generated under a different distribution.

Data representativity and its associated effects on appropriate inference is seldom given sufficient attention. This issue is as relevant to inference drawn from data in physical sciences as it is for appropriate inference in social sciences. Moreover, the problem is particularly prevalent in population sampling and in public datasets where assertive claims of *representative samples* are ubiquitous. As we increasingly harness AI systems to govern important decisions, it is critical to evaluate what the underlying training data of these systems represent, whether they originate from historical datasets or derived representations of the same.

Paper C
Data Representativity for Machine Learning and AI Systems

Paper C presents an overview and survey of data representativity in ML and AI literature. The paper argues that the term *representative sample* is overloaded, and that this term encompasses a range of notions which allow different inferences to be made. The survey identifies a range of notions of data representativity and categorizes them into three measurable underlying concepts: reflection, coverage and representatives. Moreover, the paper links these concepts to mathematical measures, which may help quantify and asses the representativeness of a given dataset.

Reflection: A sample or dataset can be considered representative under the concept of reflection if it reflects, mimics, or matches the distribution of the target population, or the influence of the target population on trained models. This may be ensured by matching the proportional sizes of sub-populations as is done in stratified sampling, or by ensuring that average predictions on sample and population align, such as in meta-learning dataset distillation. The representativity may be evaluated under this notion by comparing average predictions, or the distribution between sample and target with distributional distance measures, such as the general Wasserstein distance [Vas69, Kan60].

Coverage: A sample or dataset can be considered representative under this concept if it covers the target population by representing its heterogeneity. As such, coverage does not require proportions across data partitions between sample and population to match. Coverage may be assessed mathematically through diversity measures such as demographic information entropy or geometric coverage by comparing the data volume expressed by the sample and population.

Representatives: A sample or dataset is representative under this concept if it captures underlying archetypal objects that represent subgroups of the overall population. The representativity of this concept may be assessed through cluster metrics evaluating the average distance to the representative within and between each subgroup or via reconstruction loss in cases where representatives are combined and used to reconstruct original data instances.

Paper C argues that it is impossible to talk about general representativeness and that data collection and representativeness should be considered in tandem with the purpose of the system. The target distribution may be highly complex and may evolve over time. This makes general guarantees of representativeness practically impossible.

In light of this, we propose a framework of guidelines and questions for evaluating and monitoring representativity when publishing and documenting datasets. This framework of questions supports existing efforts like datasheets for datasets [GMV+21a].

Paper C concerns research objectives 1, 2 and 4. The paper surveys various notions of representative samples and identifies core measurable concepts of data representativity. These concepts reflect different objectives in methodology with different appropriate applications. Finally, the paper argues that training data development should not be seen in isolation from inference data development and data maintenance. Rather, these axes should be considered in unison to align the training data with the intended target population. This alignment should be explicitly evaluated by monitoring the representativity and the limits of the same for the given dataset. Just as for published datasets, it is important to evaluate the representativity of data summaries generated through data reduction. This means that the construction of the summary and the trade-offs this entails, should be carefully considered in coherence with the purpose of the study. Data reduction involves distilling the original data by extracting only the most essential information. Exactly which information is essential largely relies on the representativity one aims to achieve. By extracting information in accordance with a specific concept of representativity, one inevitably enters a trade-off and restricts the appropriate inferences the summary allows for. The thesis discussion in Sect. 4.4 explores these concepts further and evaluates how the different research contributions fit into the framework.

4.3 Numerosity Reduction

Numerosity reduction is traditionally applied to generate smaller data samples, which can alleviate computational costs without significantly sacrificing predictive performance [HKP12]. This is usually carried out through the concept of reflection, where the sample is constructed such that the performance of a model trained on the sample is similar to the performance of a model trained on the original data. For many applications this is appropriate. However, the historical data used to train ML models are often imbalanced and encode various biases. Constructing a data summary under the concept of reflection risks propagating and obfuscating these issues. Alternatively, data summaries can be constructed under the concept of coverage to deliberately alter the distribution of the original data, and in turn achieve a more balanced representation and predictive performance. The thesis addresses this matter by introducing and evaluating approaches for obtaining balanced data representations using sampling and clustering based numerosity reduction.

Paper D

Sampling To Improve Predictions For Underrepresented Observations In Imbalanced Data

Paper D demonstrates sampling approaches for obtaining balanced reduced representations. Data imbalance is known to negatively affect the performance of models on underrepresented observations [HYS+17]. This topic has been studied extensively for imbalanced categorical target values, but only sparsely for the input data. In this paper we study data imbalance driven by the input space and demonstrate three sampling approaches to mitigate the effects of this imbalance. We demonstrate the sampling approaches on biopharmaceutical manufacturing data, where data imbalance is common by consequence of the controlled production settings. We find that an inverse density sampling approach can systematically increase the performance for observations in low-density regions without a large reduction in the overall performance. Low-density observations in the biopharmaceutical manufacturing data stem from observations in faulty batches resulting in process aberrations. These examples provide valuable information on the dynamics of the production away from the controlled environment.

The paper emphasizes that sampling of this nature can provide benefits in a range of applications where data imbalance is prevalent and that typical evaluations metrics that only consider overall performance across all observations fail to address the granularity of the performance across sub-groups of observations.

The paper addresses research objectives 1 and 4 by evaluating alternative sampling approaches, which extract samples that diverge from the original data distribution. These techniques are appropriate for obtaining more balanced predictive performance across different input examples and can be used in a variety of domains where inference for minority observations is of importance.

Paper E

Fair Soft Clustering

Paper D considers sampling based methods for obtaining balanced representations that cover the heterogeneity of the input data and mitigate inference bias in supervised models. However, imbalanced data can also be used as training data for unsupervised algorithms. An instance of this is clustering, which may be used as a feature engineering tool to supplement points with a cluster signature ID (label). The cluster IDs can then be used in conjunction with the original data attributes in downstream models to achieve higher expressive power.

If the original data is biased, the cluster solution could propagate this bias into the returned features and lead to biased inference in the resulting model. To avoid this, the field of fair clustering has emerged in the literature.

Fair clustering exists for a variety of fairness notions and has mainly been studied for hard clustering like k-means and k-medoids, where cluster assignments are deterministic [CMM21]. The contribution presented in paper E extends this line of research into the soft clustering setting, where assignments are probabilistic. The paper introduces metrics for fair clustering under probabilistic assignments and demonstrates an approach for learning a fair soft cluster solution from a modified reduced input data representation.

This paper concerns research objective 1 and 4 by providing an approach for obtaining a fair soft cluster solution from a reduced input representation, which mitigates inherent bias encoded in the original data. The reduced representation can be used as curated training data in a traditional clustering setup to ensure the fairness of the solution.

4.4 General Discussion

This section provides a discussion of the research contributions and evaluate how they advance the research objectives introduced in Sect. 1.2. This section also discusses practical applications and the limitations of the presented research contributions.

1. *Research, evaluate and develop methods for obtaining surrogate data representations, which summarize and represent the original data population in a smaller form that may be used as curated training data for downstream models.*

The research contributions advance this objective from various angles. Research contribution A develops a novel technique for obtaining a reduced data representation for astrophysical data, which can be used as curated training data to learn and remove the telluric contribution from new observed spectra. Research contribution B advances this objective by presenting an approach for the reduction of image data in a classification setting from a high-dimensional feature space to a low dimensional embedded space. Similarities with learned prototypical objects in the embedded space are then fed as curated training data to a linear classifier to obtain the class predictions. Paper C advances the objective by evaluating the opposing inherent qualities different data representations provide. Paper D presents alternative sampling approaches for reducing the original data and mitigating the effects of data imbalance on the downstream model.

Finally, research contribution E advances the objective from a soft clustering viewpoint by presenting a technique for modifying and reducing the original data into a new representation, which can be fed as input data to a traditional clustering algorithm such that the learned solution maintains balance of protected attributes.

2. *Evaluate the appropriate application of techniques across different domains by identifying the distinct types of essential information that data reduction seeks to distill.*

The type of information a data representation encapsulates is studied in Paper C. This paper largely summarizes the objectives of the remaining paper contributions into the proposed concepts of representativity. Each of these papers introduce and evaluate methods for extracting essential information from an original dataset to produce a reduced summary representation. The specific type of information the reduced data is representative of is captured by the concepts. For each paper, the type of representativity guides the appropriate methodologies and evaluations of the proposed approaches. Fig. 4.1 depicts a graphical abstract of the concepts of representativity as they relate to the extraction of essential information from a data source.

The research contributions fit into the proposed framework of representativity as follows: In paper A, a constrained autoencoder is trained to extract a reduced representation of astrophysical spectra by decomposing the data into its constituent representative objects. These objects capture the spectral signatures of the intrinsic solar and the telluric H_2O and O_2 components. The reduced representation is thus representative of the original data under the concept of representatives. In the construction of this representation, no effort is made to reflect or match the distribution of the original data instances, and no effort is made to capture the full heterogeneity (diversity) of the original data, as some of this expressed diversity is assumed to originate from noise artifacts from the instrument or cosmic particles. Instead, the representation is constructed to capture a small number of underlying prototypical objects from which the original data instances are assumed to originate. This restricts the valid inferences that can be made from the representation to inference about the physical profile of these objects as observed by the given instrument. The representation is not designed to be used to draw valid inference about population parameters for the population of solar observations. The degree of representativity expressed by the learned components can be assessed with mathematical measures such as the reconstruction loss in regions of the spectrum where the ground-truth spectral components are known to exist or by comparing the learned solar component to the ground-truth solar spectrum as observed by an instrument free from telluric artifacts, for instance from outside the atmosphere of Earth.

Figure 4.1: A graphical illustration of the three proposed overarching concepts of data representativity. The yellow data represent the original data source while the white data depict reduced representations of the original data in accordance with the three proposed concepts of data representativity. Reflection mimics the population distribution, coverage represents the heterogeneity of the population distribution, and representatives represent the original data though underlying prototypical instances. Graphical examples of typical use cases for each concept of representativity is shown. The reflection concept is typically employed in demographic data and used to argue that a data representation acts as a miniature of the population. The coverage concept is typically used in tandem with notions of diversity and associated bias mitigation. The representatives concept is often used to infer underlying archetypal patterns in the original data, for instance in natural sciences.

In paper B, prototypical objects are learned with the goal of covering the distinct aspects of the input distribution. Thus, the reduced representation is constructed under the concepts of representatives and coverage.

The reduced representation aims to represent underlying prototypical patterns of variation in the data, but at the same time aims to express the heterogeneity of the data distribution by capturing its diverse aspects. This allows inference about the underlying archetypical patterns which exist in the training data, but also causes inference from the predictive model to be more balanced across demographic groups. The upside of this type of representativity is met with a corresponding downside on the predictive performance for individuals in majority sub-populations. In this paper, the concepts of coverage and representatives are both assessed with mathematical measures using cluster metrics for the representatives and measures of data coverage and demographic diversity for the concept of coverage.

The last two paper contributions on numerosity reduction both adhere to the coverage concept of representativity. They explore methods for obtaining balanced data representations, which cause models trained on them to achieve more balanced predictions across various input instances. The quality of balanced inference is particularly appropriate in applications based on demographic data.

This thesis supports the idea that the reflection concept is not the only valid option during data reduction. As such, none of the paper contributions in the thesis adopt the concept of reflection. As an example of a reflection based data reduction approach we draw attention to Wasserstein measure coresets [CGS18], which aim to reduce the original data by directly minimizing the distributional distance between the reduced representation and original data. Similarly, [ZB23] perform dataset distillation by learning a reduced representation which matches the data distribution of the original data in various embedded feature spaces.

3. *Develop dimensionality reduction techniques that infuse the reduction pipeline and learned representation with high transparency, and interpretability.*

This research objective is advanced by contributions A and B, which both use autoencoder based frameworks to learn a reduced representation of the training data. Contribution A addresses concerns in the literature about model opaqueness by enjoining ideas from the domains of neural networks, hyperspectral unmixing and factor analysis into a highly transparent and interpretable data reduction pipeline.

Contribution B empowers transparent self-explainable models by presenting a novel volumetric loss, which is used to learn a reduced representation that captures distinct patterns in the input space leading to high interpretability of the representation.

4. *Study data reduction techniques that not only reduce data size, but also reduce inherent data biases existing in the original data.*

This research objective is addressed by contributions B, D and E, which all study reduction techniques under the concept of data coverage. Training a model on the representations proposed in these papers leads to inference that is more balanced across low- and high-density input instances, which can manifest as demographic majority and minority instances.

4.4.1 Practical Application

The thesis has presented a number of research contributions, which may be used in various practical applications outlined in this section.

Paper A presents an empirical data-driven approach, which bridges a multidisciplinary gap between hyperspectral unmixing, theoretical astrophysics and the practical application of machine learning techniques. The data reduction pipeline presented in this article has been made freely available as an open-source code base, in the hope that it may aid scholars in applying the technique on their own data. The computational speed-up over current high-accuracy methods should make it tractable for practitioners to correct and better analyze thousands of observed spectra. The presented work has roots in hyperspectral satellite imaging, which aims to analyze Earth facing spectral data. These ideas are adapted to an outward facing perspective towards the complex domain of astrophysics. Consequently, the presented work may act as a steppingstone to the transfer of knowledge from Earth facing spectral techniques towards the astrophysical realm, allowing us to better peer into our stellar neighborhood and distant celestial origins.

Paper B introduces a volumetric loss into the fabric of a SEM to induce high latent diversity of the learned representation, and argues that not only should the overall model be self-explainable and transparent but the learned reduced representation should also be thoroughly evaluated in terms of representation quality and interpretability. The proposed approach can easily be implemented as a module into any existing SEM to enhance the control over the expressed diversity of the leaned representation. We hope that this may find practical use in existing and future SEMs.

Paper C studies the taxonomy of representative data and identifies a conceptual limitation existing in the literature. We hope that the categorization of existing ideas into overarching measurable concepts, as well as the proposed guidelines of questions to consider when publishing datasets, may find practical use under the design and publishing of datasets.

Paper D studies sampling approaches for obtaining balanced representations from imbalanced data, that allow downstream models to make more balanced predictions across input instances. The density based approach is easily modifiable to change the focus from high-density to low-density observations, and may be used to study the effects of the input representation on downstream inference. These sampling approaches may also find practical use to mitigate biased inference for various sub-groups within the data, particularly in cases where sensitive attribute labels are unavailable.

Paper E studies fair clustering in a probabilistic setting. This is a new domain of research, and as such no measures for soft fairness have been cemented. We hope that the proposed measures may aid in evaluating existing and future approaches for fair probabilistic clustering. The proposed algorithm for constructing a fair probabilistic clustering from a fairlet decomposition is accompanied with theoretical fairness bounds and can be applied to any fairlet decomposition, whether they originate from deterministic or probabilistic clustering algorithms. As such, this algorithm could find practical use in a number of instances. Nonetheless, the introduced minimum cost flow approach for finding a fair Gaussian mixture model fairlet decomposition may not serve much use in practical application by consequence of the high computational complexity of the algorithm. Nevertheless, it serves as an interesting theoretical study that may spark ideas for future study into tractable fairlet decomposition construction for probabilistic clustering.

4.4.2 Limitations

Some of the research contributions in this paper involve bias mitigation and algorithmic fairness. These contributions adhere to a group-level notion of fairness derived from the disparate impact doctrine [Rut87] which prohibits discrimination (disparate predictive accuracy) between different groups categorized by protected attributes (such as race). This is motivated by the fact that victims of discrimination in automated decision frameworks are often part of minority groups [MZP21] that are either affected by historical biases or are underrepresented in the training data [BG18b]. ML and AI systems are typically designed with efficiency and profit in mind and this design usually accepts poor performance for minority groups as a worthy sacrifice of collateral damage to the benefit of improved performance for majority groups of the population. Nonetheless, it is important to note that exactly what fairness entails, and how we may instill these values into our algorithms, is an ongoing scientific and societal debate. Multiple notions and definitions of fairness exist [MZP21], and some of them are incompatible [BHN17]. It is critical to evaluate if a given notion of fairness is suitable for a specific task, as applying an inappropriate notion may result in undesired discrimination.

Some of the contributions presented in the thesis provide a significant decrease in computational expense over existing methods (such as contribution A). However, other contributions (such as contribution E) provide no such speed-up. Particularly, contribution E, like many efforts in fair clustering, scales poorly in the number of data instances and dimensions. However, the proposed soft cluster fairlet algorithm from paper E can be applied to more scalable fairlet decompositions obtained either from current work in deterministic clustering [BIO+19], or in potential future work in probabilistic fairlet decompositions.

Finally, most of the techniques presented in this thesis are not accompanied by theoretical bounds on their expected deviations from the original data, or the associated expected theoretical deviation in costs of models trained on the reduced data. This contrasts with techniques from computational complexity and geometry, such as coresets, which provide provable bounds on the performance of the downstream model. However, the research contributions of the thesis are all empirically evaluated against unseen test data to demonstrate that they have captured underlying structure and allow generalization.

4.5 Proposed Future Research

This thesis investigates data reduction and data representativity by studying, evaluating and developing novel approaches that assess and link the learned summary representations to their intended target population. Nonetheless, the thesis recognizes several directions of research that requires additional attention and may prove as potential future research endeavors.

4.5.1 Dimensionality Reduction

- Dimensionality reduction techniques, such as autoencoders, are trained to represent their training data, but the learned representation is often used to draw inference about an underlying target distribution which may be highly complex, adapt to a specific data collection procedure or exhibit temporal changes. Additional research is required into the effects of applying the learned compressed representation to draw inference on data from different sources (out-of-distribution evaluation). This could for instance be the case in physical sciences, where a representation learned for a specific instrument is applied to draw inference for data observed by another instrument. In such settings the link between learned representations for the different instruments could be guided by transfer learning.

- Self-explainable models in XAI often use dimensionality reduction based pipelines to learn input features for downstream models. The connection between the learned representation and the inference drawn by the succeeding model needs further study and evaluation. Such evaluation may reconcile various notions of diversity expressed in the latent space, and their relation to appropriate inference.

4.5.2 Numerosity Reduction

- Additional research is needed to narrow the gap from algorithmic complexity and computational geometry to ML and AI based perspectives on data reduction. This may aid in harmonizing strong theoretical guarantees with empirical approaches and demonstrations of generalization ability between reduced representation and unseen target data.

4.5.3 Data Representativity

- Additional research is required into sustained monitoring of the representation quality of published data sources, or of in use data sources utilized by deployed AI systems already in production.
- As data complexity increases, we approach a limit for our understanding of the link between input and target distributions. This narrows the window of valid inference and complicates AI alignment. This calls for further research into mathematical measures of data representativity. Particularly measures which allow modelling of joint distributions, and which are suitable in high-dimensional data.

Chapter 5

Conclusion

The purpose of this thesis was to complement an emerging data-centric perspective in ML and AI literature by studying data reduction techniques that modify and reduce the original data into summary representations, achieved by extracting only the most essential information. Data reduction techniques are widespread and are applied across numerous scientific problems most often with the aim of reducing computational complexity for downstream tasks.

In dimensionality reduction, the original data is reduced by identifying or constructing useful features and discarding the remaining features. This can be achieved in a number of ways, but recently powerful neural network architectures have proven especially useful in learning expressive data representations. Nevertheless, neural networks often operate in a black box nature, which obfuscates the dimensionality reduction procedure and makes it hard to interpret the learned latent representation. This thesis has presented two research contributions which address these concerns by proposing neural network architectures that promote high transparency and interpretability of the learned representation, without sacrificing expressive power.

Numerosity reduction is another sub-field of data reduction, where the number of data instances are directly reduced via for instance sampling or clustering approaches. These techniques are widely used in various settings from production data to analysis of demographic population data.

When these data subsets are constructed to match the distribution of the original data, or to produce similar predictions for downstream tasks, they often end up propagating or even enhancing inherent biases existing in the original data population. For certain data sources, this is not problematic. However, for data where underrepresented observations are particularly informative, or represent individuals, such bias propagation should be considered with care. This thesis has presented two research contributions which cause the distribution of the summary to diverge from the original data source and changes the predictions made by succeeding models. This complicates causal inference about the data generating process but allows predictive inference in downstream tasks to be less biased.

Finally, a core conceptual contribution in this work was the identification that claims of representativity are ubiquitous and usually unspecified in the literature. The thesis has addressed this by categorizing data representativity into three measurable concepts, which capture different underlying perspectives and may guide appropriate methodologies for practitioners. Published datasets and data summaries, whether collected or created through numerosity or dimensionality reduction, are often claimed to be representative of a target population, or the original data source, without sufficient specification and assessment of this representativity, or the potential adverse effects this representation may propagate into downstream trained models. This thesis calls for caution on such implicit use on statements of data representativity and advises that published datasets are accompanied by sufficient specification of the intended target population and an estimation of the extent to which the dataset is representative hereof, in the hope that this will align the data collection or creation procedure with the purpose for which the data is later used.

Chapter 6

TAU: A neural network based telluric correction framework

Rune D. Kjærsgaard[1], A. Bello-Arufe[2,3], A. D. Rathcke[2,4], L. A. Buchhave[2,†], Line K. H. Clemmensen[1,†]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Richard Petersens Plads 324, Kgs. Lyngby 2800, Denmark*

[2] *DTU Space, National Space Institute, Technical University of Denmark, Elektrovej 328, DK-2800 Kgs. Lyngby, Denmark*

[3] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, United States*

[4] *Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*

[†] These authors contributed equally to this work

Publication Status: Paper is accepted. Publication forthcoming.

Astronomy & Astrophysics - Journal of worldwide astronomical and astrophysical research.

Abstract: Telluric correction is one of the critically important outstanding issues for extreme precision radial velocities and exoplanet atmosphere observations. Thorough removal of small so-called micro tellurics across the entire wavelength range of optical spectrographs is necessary in order to reach the extreme radial velocity precision required to detect Earth-analog exoplanets orbiting in the habitable zone of solar-type stars. Likewise, proper treatment of telluric absorption will be important for exoplanetary atmosphere observations with high-resolution spectrographs on future extremely large telescopes (ELTs). In this work we introduce the Telluric AUtoencoder (`TAU`). `TAU` is an accurate high-speed telluric correction framework built to extract the telluric spectrum with previously unobtained precision in a computationally efficient manner. `TAU` is built on a neural network autoencoder trained to extract a highly detailed telluric transmission spectrum from a large set of high-precision observed solar spectra. We accomplished this by reducing the data into a compressed representation, allowing us to unveil the underlying solar spectrum and simultaneously uncover the different modes of variation in the observed spectra relating to the absorption from H_2O and O_2 in the atmosphere of Earth. We demonstrate the approach on data from the HARPS-N spectrograph and show how the extracted components can be scaled to remove H_2O and O_2 telluric contamination with improved accuracy and at a significantly lower computational expense than the current state of the art synthetic approach `molecfit`. We also demonstrate the capabilities of `TAU` to remove telluric contamination from observations of the ultra-hot Jupiter HAT-P-70b allowing for the retrieval of the atmospheric signal. We publish the extracted components and an open-source code base allowing scholars to apply `TAU` on their own data.

6.1 Introduction

The absorption of photons by constituents in the atmosphere of Earth (telluric absorption) complicates ground-based observations and is a well-known obstacle for obtaining precise radial velocities (PRVs) in the near-infrared [BSH$^+$10] at the m s^{-1} level. Even in the optical wavelength range, there are several bands of oxygen lines and numerous micro-tellurics originating from shallow water lines. These micro-tellurics can constitute a significant amount of the PRV error budget at the $\sim$20 cm s^{-1} level [CSF$^+$14, WLP$^+$22]. To this end, various methods have been introduced to remove the effects of telluric contamination and the accuracy of these efforts remain a critical challenge on the path to the 10 cm s^{-1} radial velocity (RV) barrier for detecting Earth-like exoplanets [FAA$^+$16].

An acknowledged method, `molecfit` [SSN$^+$15, KNS$^+$15], relies on computing a synthetic transmission spectrum of the atmosphere of Earth by combining

an atmospheric profile from the Global Data Assimilation System (GDAS) website and a line-by-line radiative transfer code [CSM+05] to fit the observed spectrum. `Molecfit` has been found to be more robust than other methods, for instance using airmass [LMCH21]. While `molecfit` is a popular and well-established library, alternative telluric synthesis codes such as TERRASPEC, Transmissions of the AtmosPhere for AStronomical data (TAPAS), and TelFit [BMD+12, BLF+14, GDRK14] also exist. These synthetic approaches have been well-tested, but are inherently reliant on external factors to an observation, such as atmospheric measurements or molecular line lists for computing radiative transfer solutions.

Another realm of methods take a data-driven approach to exploit the modes of variation in a number of observed spectra to uncover the underlying components. By analyzing such a variation, telluric absorption can be modeled without relying on external factors. Additionally, given high-precision data, these methods can uncover the precise spectral location of molecular transitions in the atmosphere, which are otherwise hard to estimate with synthetic models. One such data-driven approach is directly based on principal component analysis (PCA) [AADD+14] while another, the Sys-Rem algorithm [TMZ05], is an extension of PCA which accounts for unequal uncertainties in the data. PCA methods are however ineffective on very large datasets, where the entire data cannot be stored in memory. Additionally, extracted components from PCA can be hard to interpret. Another approach, `wobble` [BHFM+19], uses a linear model in log flux with a convex objective and regularization to model the underlying stellar and telluric components of high-precision observed spectra from bright stars. `Wobble` requires the spectral components to undergo large Doppler shifts with respect to each other to disentangle their components effectively. This means that `wobble` typically requires numerous observations over a large fraction of the year to perform corrections for stars that do not undergo large RV shifts over short timescales. Light-weight data-driven initiatives such as the self-calibrating, empirical, light-weight linear regression telluric (SELENITE) model [LFV19] also exist. This method uses a linear regression fit to observations of rapidly rotating B stars in addition to airmass measurements. However, SELENITE is sensitive to stellar features in the training set and is limited to variation according to airmass and water vapor column density.

All data-driven approaches ultimately exploit that information about the telluric spectrum is encoded within the data. To extract this information, current data-driven initiatives are applied on rather modest sized training sets. We argue that data-driven models can benefit from the introduction of very large high-precision datasets of observations, which encode the telluric spectrum with previously unobtained precision. Inspired by this, we present a novel deep-learning-based approach fueled by recent releases of large-volume datasets and provide the

framework as an open-source code base[1]. The approach is based on a neural network autoencoder.

Autoencoders have seen use in the literature for decades [BK88, HZ94] and have long been known to discover effective compressed data representations through dimensionality reduction [HS06]. Autoencoders can be trained through mini-batch gradient descent, where only a small portion of the entire training data is used to compute an approximate gradient. This means that autoencoders can learn from large datasets. Additionally, autoencoders are highly flexible enabling nonlinear structures to be captured and can be readily modified through regularization and architectural constraints to enforce the learned representation to assume interpretable physical properties.

[PSSU18] present a neural network autoencoder for unmixing of hyperspectral images based on a linear mixing model (LMM). They show that a hyperspectral image can be unmixed into its underlying components called endmembers. We build on this idea by adapting the network architecture to the domain of astrophysical spectral data. This requires various new constraints on the network as well as the introduction of a specialized reconstruction of the input data.

To demonstrate our approach we analyze a large number of observed solar spectra[2] [DCS+21] from the high-resolution (R $\sim$ 115,000) radial velocity cross-dispersed echelle spectrograph HARPS-N [CLP+12] (High-Accuracy Radial-velocity Planet Searcher for the northern hemisphere) mounted on the 3.6 meter Telescopio Nazionale Galileo (TNG) on La Palma, Spain, with the goal of disentangling the observed spectra into their underlying solar and telluric components. The HARPS-N data covers a wavelength range between 3830 Å and 6930 Å. By training on the HARPS-N solar spectra, we let the data speak for itself and through that discover a reduced representation that encapsulates the overall dataset. Data reduction has many uses, but particularly for this data a compressed representation can be used as a way to detect patterns relating to real interpretable physical effects, identified as underlying components (spectra) across all observations. We choose solar data for training since a large quantity of these spectra are available. Moreover, solar observations do not take away observing time from night time observations, and possess high signal-to-noise ratio (S/N) and resolution, allowing the extracted telluric signal to inherit these properties. Training on nonsolar data is also possible, but would require changes to the structural constraints of the network. Finally, by training on observations from a single spectrograph, we capture inherent information to the instrument, such as the point spread function (PSF). This means that our extracted components are specialized for the spectrograph used for training

[1] Our code is publicly available at `https://github.com/RuneDK93/telluric-autoencoder`
[2] `https://dace.unige.ch/dashboard/`

(HARPS-N) but the method can easily be extended to other spectrographs by training on solar observations from these instruments. The extracted telluric components could aid in the detection of faint radial velocity signals of planetary systems by quickly and accurately removing tellurics from observations, leading to an increase in observation quality and hereby a reduction in observing time and cost.

The paper is structured in the following way. Sect. 6.2 describes the physical model of the problem. Sect. 6.3 demonstrates the setup and architecture of the autoencoder neural network. In Sect. 9.4 we show the results of training the network on the HARPS-N data and compare the extracted components with synthetic telluric transmission spectra computed by `molecfit`. Sect. 10.5 discusses our results and evaluates the advantages and limitations of the approach while Sect. 10.6 presents the conclusions of the paper.

6.2 Physical model

Each ground-based observed solar spectrum is a combination of the intrinsic solar spectrum and contamination effects from extrinsic factors like absorption in the line of sight as well as instrumental perturbations. Absorption in the line of sight for solar observations is contaminated by telluric absorption in the Earth's atmosphere. We can express an observed solar spectrum as a convolution between the instrumental profile and the profile of the observed object in the following way [VCR03]:

$$O(\lambda) = [S(\lambda) \cdot T(\lambda)] * I(\lambda) \cdot Q(\lambda), \tag{6.1}$$

where O is the observed spectrum, S is the intrinsic solar spectrum, T is the combined telluric transmission spectrum, I is the instrumental profile, which acts as a line broadening effect, Q is the instrumental throughput, $*$ indicates a convolution, and λ is the wavelength of the observed light.

If we assume perfect throughput and an ideal spectrograph, which maps all light at a particular wavelength to a distinct location on the detector, then we can simplify Eq. 6.1 by representing an observed solar spectrum as an intrinsic solar component and an extrinsic component describing the telluric absorbance occurring in the atmosphere of Earth:

$$O(\lambda) = S(\lambda) \cdot T(\lambda). \tag{6.2}$$

The telluric transmission spectrum T can be described by a combination of a finite set of molecular species acting as absorbers in the atmosphere of Earth. The combined telluric transmission spectrum from K absorbing species can be expressed in the following way:

$$T = \prod_{k=1}^{K} t_k, \tag{6.3}$$

where t_k is the transmission spectrum of an individual molecular species k. Important absorbing species in the atmosphere of Earth include H_2O, O_2, CO_2, CO, CH_4, N_2O, and O_3.

By observing the solar spectrum through varying atmospheric conditions over an extended period of time, the telluric transmission spectrum will naturally show large fluctuations. The overall fluctuation will be comprised of different modes of variation arising from the constituent molecular species making up the telluric spectrum. On the other hand, the solar spectrum does not undergo large changes between observations and can be assumed constant in line depth and shape. This assumption can however be violated due to slight variations arising from solar activity effects like sun-spots, which can cause the shape of a line to change over numerous observations. Such effects are ignored in our analysis and assumed to average out over a larger number of observations. Another important distinction between the solar and telluric components is that the telluric lines will always be positioned at the same location in the wavelength domain (using observer rest frame calibrated spectra). Contrarily, the solar component will exhibit small Doppler shifts between observations relating to the motion and rotation of the Earth during the observation. While this Doppler shift is comparatively small for observations of the Sun, it remains non-negligible and should be accounted for to uncover the true underlying components of the observed spectra.

6.3 Proposed method

We aim to disentangle the observed solar spectrum into the underlying telluric and solar components by using a neural network autoencoder. The autoencoder provides a reduced representation such that the overall data can be described using only a few underlying components. To ensure the learned representation embodies the underlying components, we utilize the physical model of the system and design the architecture and constraints of the neural network to comply with this model.

[PSSU18] present an autoencoder neural network architecture for blind unmixing of hyperspectral images (HSI). These images are a combination of distinct spectra

called endmembers. Spectral unmixing seeks to unmix the endmember spectra and their abundances (defined by relative proportion of an endmember in a pixel) from the observed hyperspectral image by constructing a mixing model of the problem. Endmember unmixing from spectral data is a rich discipline with many existing approaches [SATC11, HPG14]. Mixing models are divided into linear and nonlinear models. The nonlinear variants involve more complicated physical models and for this reason simple strategies are often implemented to remove nonlinear effects. One such strategy is the natural logarithm operation [ZZ19].

Drawing inspiration from the discipline of hyperspectral unmixing, we consider the intrinsic solar spectrum S and the combined telluric spectrum T as endmember spectra with associated abundance weights and use these components to construct a linear mixing model (LMM) in log-space describing each observed solar spectrum:

$$\log O_n = w_{s,n} \log S_n + \sum_{k=1}^{K} w_{k,n} \log t_k, \tag{6.4}$$

where the n^{th} observed solar spectrum O_n is described as a combination between an underlying solar component S_n with weight $w_{s,n}$ and telluric components t_k with weights $w_{k,n}$. We use that the combined telluric spectrum can be described as a combination of individual molecular transmission spectra t_k from each included molecular species k (Eq. 6.3). Each telluric component can then be scaled with an abundance weight $w_{k,n}$ to match the atmospheric conditions for the n^{th} observation.

We denote the number of endmembers as R, with individual endmembers $r = 1, ..., R$. Here $r = 1$ represents the solar endmember and $r = 2, ..., R$ represents the telluric endmembers. We note that in general $R = K + 1$. Using this notation and assuming that the input spectra have been preprocessed (see Sect. 6.3.1), including having the natural logarithm applied to them, we can express the LMM from Eq. 6.4 in the following matrix form:

$$\boldsymbol{x}_n = \sum_{r=1}^{R} w_{r,n} \boldsymbol{m}_r + \boldsymbol{\epsilon}_n = \boldsymbol{M} \boldsymbol{w}_n + \boldsymbol{\epsilon}_n, \tag{6.5}$$

where $\boldsymbol{x}_n$ is the n^{th} preprocessed observed spectrum from a finite set of N observed spectra and $\boldsymbol{m}_r$ is the r^{th} endmember spectrum of R endmembers with individual endmembers $r = 1, ..., R$. Furthermore, $w_{r,n}$ is the abundance of endmember r for observation n, $\boldsymbol{M}$ is the endmember matrix having endmembers as columns, $\boldsymbol{w}_n$ is the abundance vector of the n^{th} observation, and $\boldsymbol{\epsilon}_n$ is an error term accounting for noise artifacts like cosmic rays. The HARPS-N spectra cover the optical wavelength range and for this reason we consider the combined telluric spectrum to be comprised of the two strongest absorbing molecules in this region, namely H_2O and O_2. From this we get $R = 3$ with $r = 1$ representing the

solar endmember, $r = 2$ representing the H_2O endmember and $r = 3$ representing the O_2 endmember.

The goal is to extract the endmember matrix $\boldsymbol{M}$ and abundance vector $\boldsymbol{w}_n$ by training the neural network on a set of preprocessed observed solar spectra, with the purpose of applying the extracted telluric spectrum to nonsolar observations. The endmember matrix $\boldsymbol{M}$ is extracted for training regions with P pixels. Figure 6.1 shows a graphical representation of the approach.

6.3.1 Data preprocessing

The training data consists of 1257 observer rest frame calibrated observations of the solar spectrum each split into 69 spectral orders (hereafter orders) with $P = 4096$ pixels in each order, totaling 69×4096 pixels per observation. Together the 69 orders (which we name from $0, ..., 68$) span the spectral range from 3830 Å - 6930 Å. The data are from 2020 spanning a month of observations from October 22 through November 19. We blaze corrected the observations and associated them with their respective wavelength solutions. We then linearly interpolated the observations to a common wavelength grid constructed by generating 4096 evenly spaced points between the maximum and minimum wavelength values in each order. We removed noisy spectra by filtering observations with a mean flux across the full spectrum 20 per cent lower than the maximum mean flux of the observations. Additionally, to better reflect ordinary observing conditions, we removed any observation with an airmass exceeding 2.0. Finally, we applied the natural logarithm and continuum normalized the observed spectra by iteratively fitting a first degree polynomial to an asymmetrically sigma clipped subset of the data until the clipped pixel selection was stable. This concluded the preprocessing procedure, which ensures that the spectra are corrected for variations in throughput, as required by Eq. 6.2. The described procedure left 838 spectra, of which 75 per cent were used for training and 25 per cent were used for validation to reduce the risk of overfitting. The observations were shuffled before being divided into the training and validation sets.

6.3.2 Neural network autoencoder

Autoencoders are built to reconstruct the input in the output through a low dimensional representation in the middle of the network. The network consists of an encoder function $\boldsymbol{h}_n = f(\boldsymbol{x}_n)$, which maps the input data $\boldsymbol{x}_n \in \mathbb{R}^P$ to a hidden layer describing a latent representation $\boldsymbol{h}_n \in \mathbb{R}^R$ of the input data. This representation is then passed through the decoder function $g(\boldsymbol{h}_n) = \hat{\boldsymbol{x}}_n$,

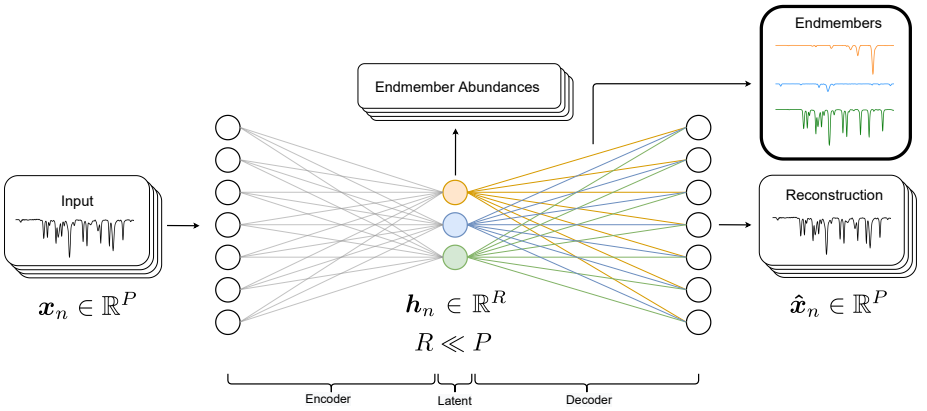


Figure 6.1: Schematic overview of the autoencoder architecture. Observed spectra $\boldsymbol{x}_n$ are given as input and passed through the encoder into a lower dimensional latent space, which is subsequently decoded into the reconstruction $\hat{\boldsymbol{x}}_n$. After training by minimizing the reconstruction error through a gradient descent algorithm, the endmember matrix $\boldsymbol{M}$ is extracted as the weights of the decoder and the abundance vector $\boldsymbol{w}_n$ is extracted as the latent representation $\boldsymbol{h}_n$. P is the number of pixels for each spectral order in the observed spectrum. The network is illustrated for $R = 3$ endmembers representing the solar (orange, top), H_2O (blue, middle) and O_2 (green, bottom) endmembers.

which seeks to reconstruct the input data $\boldsymbol{x}_n$ with the reconstruction $\hat{\boldsymbol{x}}_n \in \mathbb{R}^P$. Autoencoders are typically restricted in various ways to ensure they do not learn the identity function $g(f(\boldsymbol{x}_n)) = \boldsymbol{x}_n$. By utilizing appropriate constraints it is possible to force the latent representation $\boldsymbol{h}_n$ to take on useful properties. One such constraint is to keep $\boldsymbol{h}_n$ at a lower dimension than $\boldsymbol{x}_n$. This will make the autoencoder undercomplete and cause dimensionality reduction, which forces the latent representation to capture only the most salient features of the training data [GBCB16].

We designed the dimension of the latent representation in the telluric autoencoder to match the number of expected endmembers in the trained spectral region such that $\boldsymbol{h}_n \in \mathbb{R}^R$. For the HARPS-N data we considered regions with either $R = 2$ (solar and H_2O endmembers) or $R = 3$ (solar, H_2O and O_2 endmembers) where $R \ll P$. After the network training is complete and the network has learned to reconstruct the input signals, the latent representation $\boldsymbol{h}_n \in \mathbb{R}^R$ can be extracted and interpreted as the endmember abundances $\boldsymbol{w}_n$. While the encoder is responsible for learning the abundances of the underlying components,

the decoder is responsible for learning the spectral shape of these components. To enable this interpretation, the decoder must be restricted to perform a single affine transformation:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{h}_n, \tag{6.6}$$

where $\boldsymbol{W} \in \mathbb{R}^{P \times R}$ are the weights of the decoder. Thus, we constructed the decoder without bias terms. With this restriction the decoder weights $\boldsymbol{W}$ can be extracted and interpreted as the endmember spectra $\boldsymbol{M}$.

6.3.2.1 Network training

Autoencoders are trained by minimizing a loss function $\mathcal{L}$ representing a distance between the input and reconstruction. We used the squared L_2 norm resulting in the mean squared error (MSE) of the input and reconstructed spectra as loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n||_2^2. \tag{6.7}$$

Training was performed with gradient descent by updating the network weights to minimize the loss function. Training was carried out for a number of epochs with early stopping, which has been shown to reduce the risk of overfiting [CLG01]. The network was trained on non-stitched spectra for the 69 orders separately to retain the high fidelity of the observed spectra and to avoid the complications involved in stitching spectra. The loss of signal at the edges of each order is compensated for by information from the high number of observations used to train the network. Training on all orders took approximately 2 hours on an Intel 6 core i7, UHD 630 CPU laptop.

6.3.2.2 Layer architecture

What follows is a detailed description of the network layers. An overview of the layer architecture of the entire network can be seen in Tab. 6.1.

Layer 1 is the input of the network with dimensionality $\boldsymbol{x}_n \in \mathbb{R}^P$ such that it matches the number of pixels P in the observed spectra used for training. Layer 2 applies dropout during training by randomly zeroing entries, with a chosen probability p, according to a Bernoulli distribution. Random dropout has been shown as an effective regularization technique to reduce overfitting by preventing complex coadaptations of feature detectors [HSK+12].

Table 6.1: Summary of autoencoder layers. P is the number of pixels in the observed spectra used for training and $R = 2$ or $R = 3$ is the expected number of endmembers in the spectral region used for training.

Layer	Description	Dimension
1	Input	P
2	Dropout	P
3	Hidden	R
4	Batch Normalisation	R
5	Abundance Normalisation	R
6	Abundance Lower Bound	R
7	Endmember Spectra Clamping	P
8	Solar Doppler Shift	P
9	Output	P

Layer 3 computes a lower dimensional representation by applying the following transformation:

$$\boldsymbol{a}^{(3)} = j(\boldsymbol{W}^{(3)}\boldsymbol{a}^{(2)} + \boldsymbol{b}^{(3)}), \tag{6.8}$$

where $\boldsymbol{a}^{(3)}$ is the activation of layer 3, $\boldsymbol{a}^{(2)}$ is the input given from the previous layer, $\boldsymbol{W}^{(3)}$ are the weights of the layer, $\boldsymbol{b}^{(3)}$ is the bias of the layer, and j is the activation function. The encoder uses a leaky rectified linear unit (LReLU) [XWCL15] as a nonlinear activation function. The encoder can in principle consist of multiple layers with nonlinear activation functions. Similar to [PSSU18], we found that the structural constraints of the decoder, which are expanded upon in the description of layer 9, limits the advantages of a deep encoder with numerous nonlinear activation functions. We tried various encoder depths and found that differences in learned representations were negligible, while training time was increased significantly with a deep encoder structure. For this reason, the final architecture of our presented network consists of a shallow encoder with a single nonlinear activation function in layer 3.

Layer 4 applies batch normalisation, which is known to accelerate learning by reducing internal covariate shift [LBOM12, IS15]. The batch normalisation operation can be expressed as:

$$\boldsymbol{a}_i^{(4)} = \frac{\boldsymbol{a}_i^{(3)} - \mathbb{E}[\boldsymbol{a}^{(3)}]}{\sqrt{Var[\boldsymbol{a}^{(3)}] + \epsilon}}\gamma + \beta, \tag{6.9}$$

where $\boldsymbol{a}_i^{(4)}$ is the activation of unit i in layer 4 for $i = 1, ..., R$, $\boldsymbol{a}_i^{(3)}$ is the activation of unit i from the previous layer, γ and β are learnable parameters,

and ϵ is a small number. The expectation and variance can be expressed as:

$$\mathbb{E}[\boldsymbol{a}^{(3)}] = \frac{1}{R}\sum_{i=1}^{R}\boldsymbol{a}_i^{(3)}, \tag{6.10}$$

$$Var[\boldsymbol{a}^{(3)}] = \frac{1}{R}\sum_{i=1}^{R}(\boldsymbol{a}_i^{(3)} - \mathbb{E}[\boldsymbol{a}^{(3)}])^2, \tag{6.11}$$

Layer 5 enforces a nonnegative abundance constraint (ANC):

$$w_{r,n} \geq 0 \quad \forall r, \forall n, \tag{6.12}$$

and normalizes the batch normalized latent abundance tensor $\boldsymbol{a}^{(4)}$ to interval $[0, 1]$ through the following transformation:

$$\boldsymbol{a}_i^{(5)} = \frac{\boldsymbol{a}_i^{(4)} - min(\boldsymbol{a}^{(4)})}{max(\boldsymbol{a}^{(4)}) - min(\boldsymbol{a}^{(4)})}, \tag{6.13}$$

where $\boldsymbol{a}_i^{(5)}$ is the activation of unit i in layer 5.

Layer 6 enforces a lower abundance bound constraint for each endmember to ensure the correct physical representation is learned. Since it can be assumed that the intrinsic solar spectrum remains constant in line depth over the observations (no abundance variation), the solar endmember abundance w_1 is kept fixed:

$$w_1 = 1. \tag{6.14}$$

The H_2O and O_2 telluric components described by abundance weights w_2 and w_3 change over each observation but are never absent in the observed spectra. This is problematic, since the main feature allowing disentanglement of the solar component from the telluric components is the telluric variability. The constant contribution of each telluric component to the observed spectra complicates this. To ensure the correct representation is learned, the telluric components are normalized to different intervals such that:

$$w_2 \in [c_2, 1], \tag{6.15}$$

where c_2 is a lower bound on the abundance of the H_2O component and:

$$w_3 \in [c_3, 1], \tag{6.16}$$

where c_3 is a lower bound on the abundance of the O_2 component. The lower bounds c_2 and c_3 on the telluric endmember abundances are defined to represent their respective line depth variability in the spectrum. The H_2O component exhibits much larger variability in the observed spectrum than the O_2 component,

and as such $c_2 < c_3$. The value for c_2 was determined by inspecting the relative difference between a known H_2O line at its strongest and weakest instance in the training set. This was similarly done for a known O_2 line. The values used for the HARPS-N training set are $c_2 = 0.03$ and $c_3 = 0.69$.

To incorporate these lower bounds, the abundance tensors are linearly transformed to interval $[c_r, 1]$ in layer 6 of the network using the following transformation:

$$\boldsymbol{a}_i^{(6)} = \boldsymbol{a}_i^{(5)}(1 - c_r) + c_r, \tag{6.17}$$

where $\boldsymbol{a}_i^{(6)}$ is the activation of unit i in layer 6, which are the abundance weights linearly transformed to interval $[c_r, 1]$ with c_r being the lower bound on the abundance for endmember r. This is the final layer in the latent representation and thus $\boldsymbol{a}^{(6)}$ represents $\boldsymbol{h}$, which can be extracted as the endmember abundance vector:

$$\boldsymbol{h}_n = \boldsymbol{w}_n = [w_{1,n}, w_{2,n}, w_{3,n}]^T. \tag{6.18}$$

Layer 7 performs clamping of endmember spectra (decoder weights), which ensures the extracted endmembers remain normalized and interpretable in relation to the continuum normalized input and output spectra. This is achieved by clamping the weights of the decoder at each forward pass to satisfy:

$$\boldsymbol{m}_1 \in [0, 1], \tag{6.19}$$

$$\boldsymbol{m}_2 \in [-1, 0], \tag{6.20}$$

$$\boldsymbol{m}_3 \in [-1, 0], \tag{6.21}$$

where $\boldsymbol{m}_1$ is the solar endmember spectrum, $\boldsymbol{m}_2$ is the H_2O endmember spectrum, and $\boldsymbol{m}_3$ is the O_2 endmember spectrum. The constraint in Eq. 6.19 ensures that the extracted solar endmember remains continuum normalized. The constraints in Eqs. 6.20 and 6.21 allow the extracted telluric endmembers to be interpreted as transmission spectra, which absorb light from the fixed solar endmember at varying weights controlled by the abundance vector $\boldsymbol{w}_n$ from the latent space. The decoder weight constraints act as regularizers guarding against exploding gradients for large learning rates, as the decoder weights can never change significantly between each backward pass.

Layer 8 performs a Doppler shift of the solar component through a shift function. An underlying assumption in the proposed approach is that both the solar spectrum and telluric transmission spectrum remain stationary in the wavelength

domain, such that they can be described by the learned weights of the decoder. The observed spectra are wavelength calibrated in the observer rest frame and hence the telluric lines remain stationary. This is however not the case for the solar component due to a slight Doppler shift of the solar spectrum arising from Earth's rotation and motion around the Sun. The relation between the spectral shift and the velocity is given by:

$$\frac{v}{c} = \frac{\lambda_{\text{shift}} - \lambda_{\text{rest}}}{\lambda_{\text{rest}}}, \tag{6.22}$$

where v is the velocity, c is the speed of light, λ_{shift} is the shifted wavelength, and λ_{rest} is the rest wavelength.

The network architecture does not allow the Solar endmember to change over the various observations and thus the Doppler shift of this component in the observed spectra will be caught in the telluric endmembers and cause noise in the extracted spectra. Since the aim is to extract detailed endmembers, this Doppler shift causes an unacceptable amount of noise, especially in regions with strong solar lines. To account for this, the network performs a Doppler shift of the solar endmember toward the solar component reference frame for each observed spectrum. This allows a single solar endmember spectrum to represent the shifted solar component in each observed spectrum. The shifted solar endmember is created by learning a stationary solar endmember in the decoder weights and Doppler shifting it using the barycentric Earth radial velocity (BERV) to match the solar wavelength reference frame in each observed spectrum. This solution reflects the physical nature of the problem, decreases model noise and allows for a better learned representation. To incorporate this solution, a Doppler shifted wavelength axis is computed for each observation according to:

$$\lambda_{\text{shift}} = \lambda_{\text{rest}} \left(1 + \frac{v}{c}\right), \tag{6.23}$$

where λ_{rest} is the stationary wavelength axis of the solar endmember from the decoder weights and λ_{shift} is the new wavelength axis matching the solar component in each observed spectrum. The solar endmember from the decoder weights is then linearly interpolated to this shifted wavelength axis before being combined with the remaining endmembers in the last layer of the network.

Layer 9 is the final layer and computes the output of the network. It has dimensionality $\hat{\boldsymbol{x}}_n \in \mathbb{R}^P$. To ensure that the autoencoder learns a representation that conforms with our LMM in Eq. 6.5, the decoder function must be restricted to perform a linear transformation from latent representation to reconstruction $g(\boldsymbol{h}_n) : \mathbb{R}^R \to \mathbb{R}^P$. Furthermore, to directly extract endmember spectra from the decoder weights, the decoder layer neurons are constructed without bias terms. The decoder thus consists of a single affine transformation:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}^{(9)} \boldsymbol{a}^{(6)}, \tag{6.24}$$

where $\boldsymbol{a}^{(6)}$ is the activation of layer 6 (the latent abundance representation $\boldsymbol{h}_n$) and $\boldsymbol{W}^{(9)}$ are the weights of the decoder. These weights have dimensionality $\boldsymbol{W}^{(9)} \in \mathbb{R}^{P \times R}$ and can be extracted as the endmember spectra $\boldsymbol{M}$.

6.3.2.3 Network initialization

Initialization of weights in neural networks is known to affect both the convergence time and learned representation [SMDH13]. We initialized the encoder weights according to the Xavier scheme [GB10]. The weights of the decoder can either be initialized in a similar fashion, or can be initialized with weights on a scale similar to the weights the network is expected to learn [GBCB16]. We achieved the latter by initializing the telluric decoder weights with synthetic transmission spectra from `molecfit`. Additionally, the solar decoder weights were initialized from the observed spectrum in the training set that most closely resembles the intrinsic solar spectrum absent from telluric absorption. This specific observed spectrum was identified as the training observation with the largest mean continuum normalized flux. This is because telluric absorption removes light, and thus the observed continuum normalized spectrum with the least telluric contribution has the largest mean flux. Here it is important to emphasize that we operate on continuum normalized spectra, where non-telluric effects causing differences in the mean raw flux are mostly removed. These non-telluric effects include changes in cloud coverage and instrumental response from observation to observation. Our initialization approach biases the solar decoder weights due to the Doppler shift of the initialization observation. We accounted for this bias when computing the Doppler shifted wavelength axis for each observation by using the relative BERV between the n^{th} observation and the initialization observation. This ensures that the solar decoder weights match the solar wavelength reference frame in each observed spectrum. Both Xavier and custom weight initialization schemes were tested and found to lead to similar learned weights by the end of training. The described custom decoder initialization scheme was however found to speed up model convergence over using Xavier initialization.

6.3.2.4 Latent space correlation

Spectral regions with strong telluric signal undergo large variation across the training data. This allows the encoder to confidently learn to disentangle the abundance variation of the underlying components. If stitched 1D spectra from the spectrograph pipeline are used during training, the encoder is able to exploit regions of strong telluric signal to effectively disentangle the individual abundance variation of each endmember. However, we chose to train on non-stitched spectra

for the 69 orders separately to retain the high fidelity of the observed spectra. For a number of these orders the telluric signal is weak, which reduces the amount of abundance variation available to learn from the observed spectra. This can cause latent space abundances to become correlated across various endmembers, potentially resulting in entanglement of the endmember spectra. Fortunately, completely disentangling telluric endmembers is not critical for performing accurate telluric correction, as only the combined telluric spectrum is of interest. However, to obtain interpretable disentangled endmembers, we explored means of circumventing the problem. Firstly, the latent space correlation can be avoided by introducing an orthogonality constraint on the abundance vectors during training. However, this is not advisable, as the different telluric components can be naturally correlated through for instance the airmass. Alternatively, it is possible to exploit strong known telluric bands from key orders, such as the O_2 γ-band in order 60 (spectral range approximately 6270 Å - 6340 Å in air) or the several orders containing significant H_2O absorption, such as order 54 of the HARPS-N data (spectral range approximately 5900 Å - 5970 Å). These orders contain strong telluric variation and can be stitched on an order with weak tellurics during training to guide the encoder in finding the correct abundances. The guiding is exclusively used to encourage the encoder in learning disentangled abundance vectors, and hence the endmembers are only saved for the spectral region of interest during training (i.e. the decoder weights of the stitched on strong telluric order is not saved after training). Exploiting strong telluric variation from separate orders from the current training order is possible since the different orders have been observed simultaneously and thus represent the same atmospheric conditions. We employed this strategy when training on orders with weak telluric signal to effectively learn the individual abundance variations of O_2 and H_2O. The endmember correlation topic is further described in Sect. 6.5.3.

6.3.2.5 Hyperparameters

The hyperparameters of the network including learning rate, dropout, momentum, weight decay, and batch normalisation parameters were tuned based on 50 iterations of a Tree-structured Parzen Estimator (TPE) optimization approach carried out with `optuna` [ASY+19] for the individual orders using the MSE reconstruction loss as objective function for minimization. Hyperparameters such as layer dimensions are constrained and were thus not been included in the hyperparameter tuning procedure.

6.4 Results

The autoencoder has been trained on all 69 orders of the HARPS-N solar observations. We show results from orders containing spectral regions of high interest to the study of exoplanetary atmospheres. These orders exhibit a combination of micro-tellurics in addition to several deep telluric lines originating from H_2O and O_2 absorption and provide an important challenge for telluric correction frameworks. All results are shown for wavelength measured in air. We report results firstly by showing the extracted endmembers and secondly by demonstrating how the extracted components can be applied to new observations to provide accurate and efficient telluric removal.

6.4.1 Extracted endmembers

Endmembers $\boldsymbol{M}$ were extracted for each order by fully training the autoencoder on the spectral region of the order and subsequently saving the tensor $\boldsymbol{W} \in \mathbb{R}^{P \times R}$ representing the final decoder weights. The associated abundance vector $\boldsymbol{w}_n$ for the n^{th} observation was extracted by saving the latent space tensor $\boldsymbol{h}_n$.

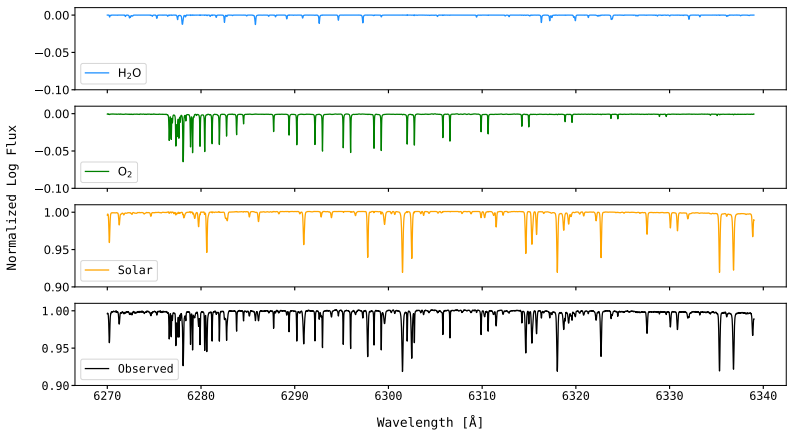


Figure 6.2: Illustration of the extracted endmembers and an observed solar spectrum for order 60. The extracted endmembers represent from top to bottom the H_2O (blue, top), O_2 (green) and solar (orange) components of the observed spectrum (black, bottom).

Fig. 6.2 shows the endmembers extracted from order 60. This spectral range

contains a combination of strong telluric lines from the O_2 γ-band in addition to weaker H_2O tellurics. The autoencoder has disentangled the signals from the observed spectrum into its underlying components consisting of the intrinsic solar spectrum and the telluric absorption from H_2O and O_2. These molecules exhibit different modes of variation in the observed spectra. The network has been allowed to learn their individual abundance variation by using a three-dimensional latent space (extracting three endmembers). The solar endmember is constrained at constant abundance and is not scaled between observations. The telluric components from H_2O and O_2 are scaled individually by their learned abundance vectors for each observation to match the atmospheric conditions at the time of the observation. For visibility, Fig. 6.2 includes an observed spectrum with strong telluric contamination and the H_2O and O_2 endmembers have been scaled with their respective learned abundances from the latent space for the illustrated observation. Note how the network has learned to disentangle the observed spectrum into interpretable components even in highly mixed spectral regions such as 6750 Å - 6850 Å, where significant solar lines are interlaced with deep O_2 tellurics and weaker H_2O lines. This is achieved by exploiting how these spectral components vary individually according to the physical laws that govern them.

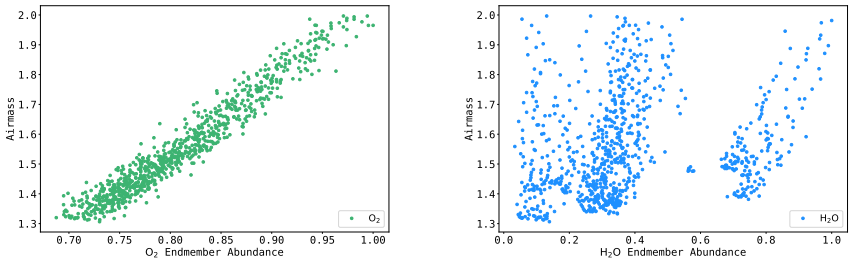


Figure 6.3: Scatter plots of airmass and the telluric abundances learned by the autoencoder for order 60 of the HARPS-N solar observations. The O_2 endmember abundances (left, green) show a clear linear relationship with airmass, while the H_2O endmember abundances (right, blue) show a much weaker correlation with airmass.

We further investigate the interpretability of the extracted components by inspecting learned endmember abundance correlation with airmass. While O_2 abundance is expected to correlate strongly with airmass, H_2O abundance correlates more strongly with atmospheric water vapor content. Fig. 6.3 shows scatter plots of the learned abundances with the corresponding airmass for each observation. As expected from the physical model of the system, the learned abundance of the O_2 component shows a strong linear correlation with airmass, while the H_2O component shows a much weaker correlation.

Additional extracted endmembers for order 53 and 54 can be found in Appendix 6.7 in Figs. 6.10 and 6.11. Firstly, order 53 contains the prominent Na doublet in addition to a wealth of micro-tellurics from H_2O. The network has disentangled the observed spectra into a fixed solar endmember and a telluric H_2O endmember. Order 53 contains a number of lines where H_2O tellurics are positioned on top of strong solar lines. This is for instance the case between 5880 Å and 5900 Å, where strong H_2O tellurics occur in the same spectral region as the prominent solar Na lines. We initially expected such mixed lines to compromise the quality of the extracted endmembers, but as we demonstrate in Sect. 6.4.2, the extracted solar and telluric components have been effectively disentangled allowing for accurate telluric correction even in spectral regions of mixed stellar and telluric signal. Fig. 6.11 shows endmembers for order 54, which contains numerous water tellurics as well as strong solar lines.

6.4.2 Telluric correction

Telluric correction removes tellurics by dividing the observed spectrum with a combined telluric transmission spectrum. In this section we demonstrate how the extracted endmembers can be applied to new observations to carry out accurate and efficient telluric correction. We do this by performing telluric correction on a collection of observations from the HARPS-N spectrograph originating both from solar observations as well as stellar observations of relevance to the detection of exoplanetary atmospheric features.

6.4.2.1 Telluric correction on validation observation

We start by demonstrating telluric correction on a validation solar observation with strong telluric contamination from HARPS-N. The validation observation is from November 2020 and `TAU` has not seen the observation during training. As a baseline for comparison we perform telluric correction of the same observation using the state-of-the-art synthetic telluric correction approach `molecfit`. We computed the `molecfit` telluric transmission spectrum using version 1.5.9 of `molecfit` on an HPC cluster [DTU23] using a 10 core Intel Xeon E5-2660v3, Huawei XH620 V3 node and utilized atmospheric measurements from the time of the observation in addition to a fit to the stitched version of the observation from the HARPS-N pipeline. We computed the `TAU` correction on an Intel 6 core i7, UHD 630 CPU laptop using the extracted telluric components, which were converted back from log-space to represent standard transmission spectra. Autoencoder telluric abundance weights were found using a least squares fit to known telluric lines in the spectrum. No external atmospheric measurements were

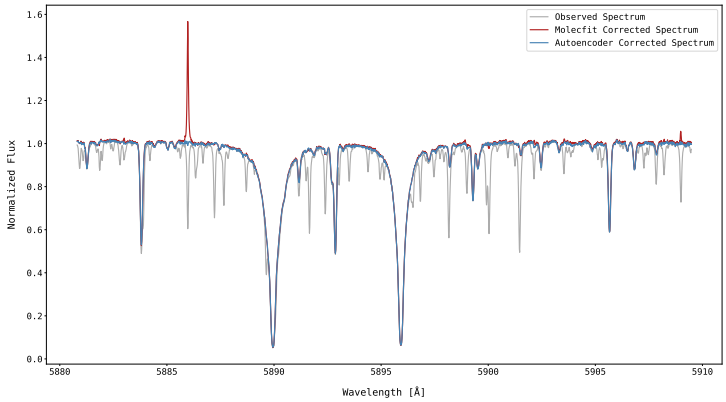


Figure 6.4: Comparison of corrected spectra around the solar Na doublet (order 53). The corrections are computed using either autoencoder extracted tellurics or `molecfit`.

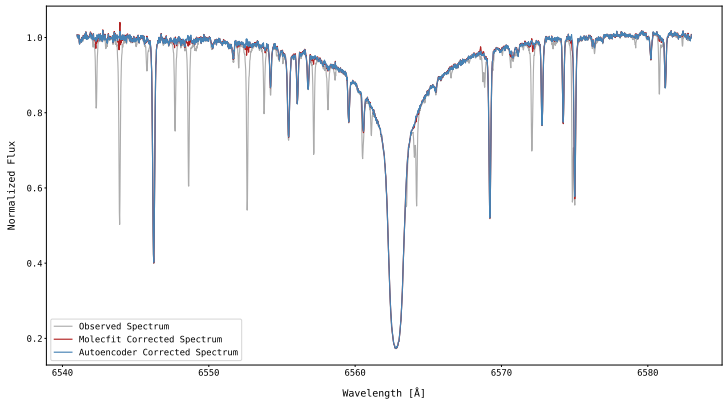


Figure 6.5: Comparison of corrected spectra in order 64. The corrections are computed using either autoencoder extracted tellurics or `molecfit` for H_2O lines in the spectral region around Hα.

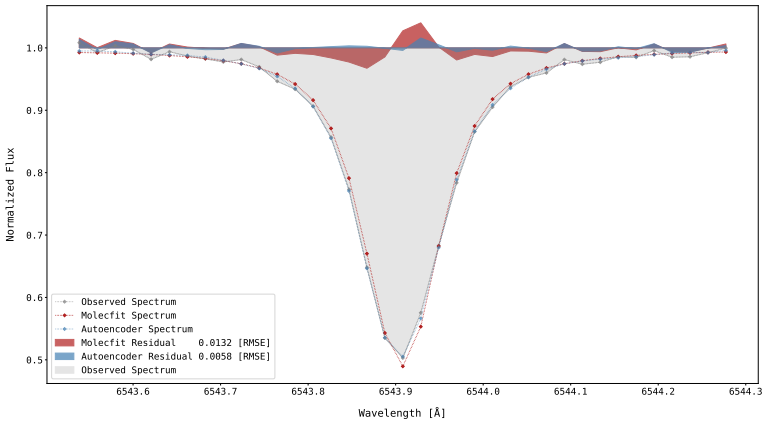


Figure 6.6: Residuals from telluric correction. The correction is performed on the validation spectrum with `TAU` and `molecfit` for a prominent H_2O telluric close to the Hα solar line shown in Fig 6.5. An ideal correction would result in no residual (flat line at continuum level).

used during the autoencoder correction. We interpolated the telluric transmission spectra of `TAU` and `molecfit` to the wavelength axis of the observed spectrum to represent standard telluric correction procedure. The `molecfit` correction took approximately 30 minutes to compute (separate from the additional time needed to optimize `molecfit` parameters to improve the quality of the fit), while the autoencoder correction took less than 0.2 seconds.

Fig. 6.4 shows the correction around the solar Na doublet in order 53. This region contains numerous deep isolated H_2O tellurics as well as telluric lines interwoven with intrinsic solar lines. Both `TAU` and `molecfit` have managed to correct for the telluric contamination by removing most of the telluric lines to continuum level. For an isolated telluric line, a perfect correction would be to the continuum level leaving no residual from the correction. In this spectral region `molecfit` is clearly over-correcting a H_2O line at 5886 Å by correcting it above continuum level and leaving a large residual. The problem with this particular line is discussed in the literature [CMW+20], and is possibly caused by an extra entry in the line list used by `molecfit`. `TAU` does not make use of external factors like molecular line list and has managed to correct this line to continuum level. A similar, but less significant over-correction by `molecfit` can be seen on the H_2O line at 5909 Å. While we have observed a number of similar small over-corrections by `molecfit` across the spectrum, they are not indicative

of the general correction performance of `molecfit` on lines where the correct external parameters are used.

Fig. 6.5 shows telluric correction around the solar Hα line in order 64. This particular spectral range is of interest for exoplanetary research and poses a difficult challenge for telluric correction frameworks due to the large number of deep telluric lines. Both `molecfit` and `TAU` have corrected the observed spectrum uncovering the intrinsic solar spectrum without indications of over-corrections. However, as is visible for a number of the deepest telluric lines, `molecfit` slightly under-corrects and leaves residuals from the correction, while `TAU` generally removes the telluric lines with higher accuracy.

Fig. 6.6 shows a closer look at the fitted telluric transmission spectra and corresponding corrections of the isolated deep 6543.9 Å telluric line also visible in Fig. 6.5. This telluric line is particularly challenging due to its significant depth. Since an ideal correction would leave no residual, we can measure the residual of the correction and use it as a metric for correction performance. The residuals are measured as the root mean squared errors (RMSE) from continuum level. The autoencoder residual (blue) is significantly smaller than the `molecfit` residual (red). As can be seen from the overlaid `molecfit` and autoencoder transmission spectra, the differences in residuals emerge due to discrepancies both in the depth as well as the exact wavelength position of the line center in the transmission spectra used for the correction.

6.4.2.2 Temporal generalizability of correction performance

The validation spectrum has been observed at a point close in time to the training data of the autoencoder. It is possible that seasonal atmospheric effects can impact the correction performance of the extracted tellurics. For this reason we performed numerous corrections on a large number of HARPS-N solar observations across various observing seasons to demonstrate the general correction performance of the autoencoder.

We chose 100 random observations from each observing season between 2015-2018 resulting in 400 observations of the solar spectrum. We then performed telluric correction on each of them using the autoencoder tellurics. All 400 corrections (across the entire observed spectrum) were computed in a total time of less than 1 minute on an Intel 6 core i7, UHD 630 CPU laptop. We have not computed `molecfit` corrections on these observations due to the significant computational expense involved in performing `molecfit` corrections on several hundred observations (and the additional time for optimizing parameters). Tab. 6.2 shows the mean RMSE of the residuals for various telluric lines across each

Table 6.2: Correction performance in three spectral regions with strong tellurics. The order 54 tellurics lie in [5945.1 Å, 5946.3 Å], order 60 in [6289.0 Å, 6290.4 Å] and order 64 in [6543.5 Å, 6544.3 Å] with all wavelengths in observatory rest frame measured in air. Val `Mol` and val `TAU` are the performances for `molecfit` and `TAU` for the single 2020 validation spectrum. 2015 `TAU`, 2016 `TAU`, 2017 `TAU` and 2018 `TAU` are the performances for the annual observing seasons computed as the mean RMSE and standard deviation of the autoencoder residuals from 100 random solar observations from each year.

Data	Order 54 [10^{-3} RMSE]	Order 60 [10^{-3} RMSE]	Order 64 [10^{-3} RMSE]
Val `Mol`	6.4	5.8	13.2
Val `TAU`	3.9	4.5	5.8
2015 `TAU`	4.0 ± 0.5	4.8 ± 0.6	6.7 ± 1.3
2016 `TAU`	3.9 ± 0.4	5.0 ± 0.6	6.6 ± 1.4
2017 `TAU`	4.0 ± 0.5	5.0 ± 0.6	6.6 ± 1.4
2018 `TAU`	4.2 ± 0.4	5.4 ± 0.6	6.7 ± 1.2

year of observations. The table also shows the residual on the single 2020 validation observation for reference. `TAU` has smaller residual than `molecfit` on the validation observation for all three spectral regions. The mean autoencoder residual for each annual observing season is also less than the `molecfit` correction in each spectral region. `TAU` generally has a smaller residual on the 2020 validation observation than the mean residual across the previous observing seasons, this is however not the case for telluric lines in order 54 (spectral range 5945.1 Å - 5946.3 Å). Overall, the correction performance shows no clear temporal trend.

6.4.2.3 Endmember correlation

Fig. 6.7 shows telluric correction in a region of combined H_2O and O_2 tellurics in order 60. The top panel shows the residuals from the correction on the validation observation, while the bottom panel shows the individual spectra for H_2O and O_2 used by `TAU`. In the left side of the figure a small H_2O telluric is positioned on top of the wing of a stronger O_2 telluric. `TAU` has modeled the combined observed telluric transmission spectrum in this region by combining the lines. Weak H_2O tellurics are visible under each of the two stronger O_2 tellurics. The location of these weak tellurics could indicate that the autoencoder has not fully disentangled the H_2O and O_2 telluric spectra from each other. This is further discussed in Sect. 6.5.3

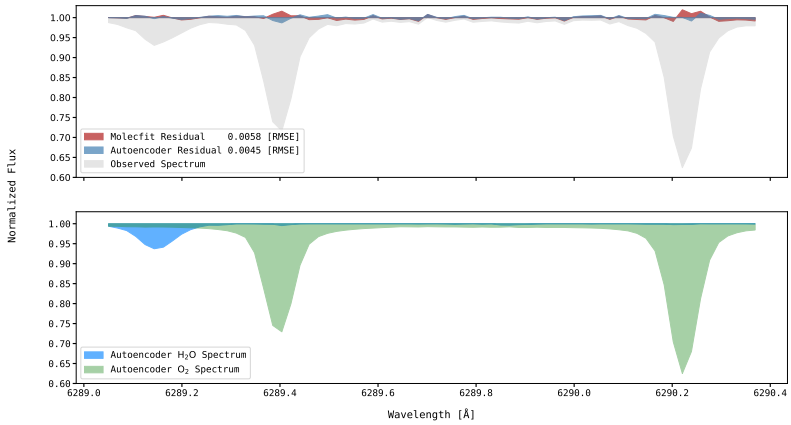


Figure 6.7: Correction of H_2O and O_2 tellurics. Top: Observed validation spectrum and residuals from telluric correction with `TAU` and `molecfit` in a region with H_2O and O_2 tellurics in order 60. Bottom: Autoencoder endmembers for H_2O and O_2 in the same spectral region.

6.4.2.4 Impact of training data

`TAU` was trained on solar observations, which are comparatively cheap to obtain. We also envision the possibility of training the autoencoder on nonsolar data in situations where a spectrograph is not fed by a solar telescope. Such data are bound to have larger BERVs, which could aid in disentangling the stellar and telluric components from each other. On the other hand, nonsolar data also has lower S/N, which would require a larger number of training observations to obtain high endmember fidelity. Obtaining nonsolar data is expensive, and it is important to know how the amount of training data impacts the generated results. For this reason, we retrained `TAU` on various amounts of training data and performed corrections with the resulting extracted tellurics.

Tab. 6.3 shows the impact on correction performance when reducing the size of the training data. We did this by randomly subsampling the original 838 observations to generate training sets of respectively 400 and 200 observations. We performed the sampling 5 times for both cases and report the mean RMSE and standard deviations for corrections on the 400 observations from 2015-2018. This results in 2000 corrections made from the tellurics extracted from networks trained on either 400 or 200 observations. The correction performance can be observed to decrease slightly with the size of the training data. This is expected,

Table 6.3: Correction performance dependence on training data size. Corrections are performed with tellurics extracted from a network trained either on the full 838 observations, or with tellurics extracted from networks trained on random samples of the training observations. We retrain the network for random samples of size either 400 or 200 observations. We perform the random sampling 5 times for both cases. Performance is shown in terms of the RMSE of the residual from corrections made on 400 random observations between 2015-2018 in the same spectral ranges shown in Tab. 6.2. The performance for "838" is the mean and standard deviation of the 400 corrections, while the performance for "400" and "200" is the mean and standard deviation for 2000 corrections each (5×400). Variance increases and performance decreases slightly for less training data.

N_{train}	Order 54 [10^{-3} RMSE]	Order 60 [10^{-3} RMSE]	Order 64 [10^{-3} RMSE]
838	4.0 ± 0.5	5.0 ± 0.6	6.7 ± 1.4
400	4.0 ± 0.5	5.0 ± 0.7	6.9 ± 1.5
200	4.0 ± 0.6	5.2 ± 0.9	7.0 ± 1.8

as feeding the autoencoder additional training data allows it to learn a more detailed telluric spectrum, leading to more accurate corrections. The correction performance is however quite stable across the training set sizes, indicating that the salient features of the data can be learned from a modest training set. Thus, the size of the training set is not directly critical in obtaining an accurate telluric spectrum. Rather, the critical part is to obtain sufficient abundance variation within the training data to learn the telluric spectrum across various atmospheric conditions. Generally, more training data carries more variation, but this is not necessarily the case. For instance numerous training observations recorded during the same night may not contain sufficient atmospheric variation to constitute a good training set.

6.4.2.5 Systematic correction performance

Natural noise artifacts like cosmic rays and photon noise impacts the residual. We demonstrate the systematic correction performance of `TAU` in Fig. 6.8, where telluric correction on a series of tellurics in order 54 is compared between a single residual from the 2020 validation observation and the combined residual from numerous corrections. The top panel shows the correction performance on the validation observation, where small correction artifacts are visible for `molecfit`

and autoencoder corrections. The bottom panel shows the combined residual from averaging residuals at each wavepoint over corrections performed by TAU on the 400 observations from 2015-2018 in the same spectral region.

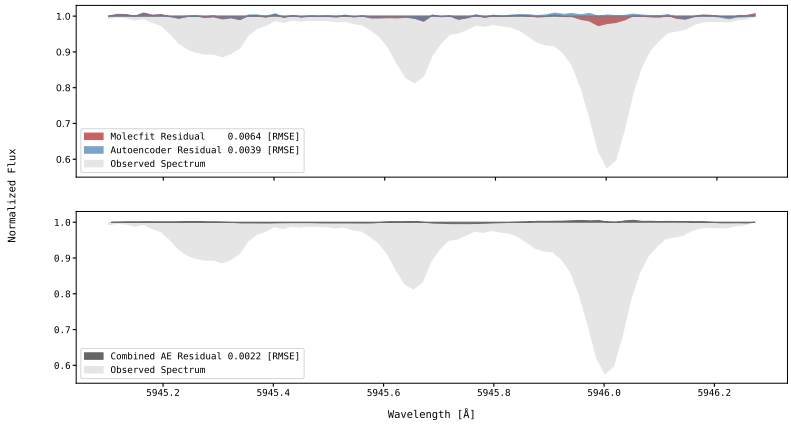


Figure 6.8: Correction of H_2O tellurics. Top: Residuals from telluric correction with TAU and molecfit on the validation observation for tellurics in order 54. Bottom: Combined residual in the same spectral region from correction of 400 observations between 2015 and 2018 performed with TAU. Combination is performed by averaging the residuals for each wave-point. The single observed spectrum from the top plot is shown for visual comparison.

Small systematic correction artifacts by TAU can be seen close to the 5946 Å line, where several telluric lines are superimposed on top of each other. In this region the autoencoder has a slight systematic over-correction effect. Overall the correction residual is within 1 per cent of the continuum level on all telluric lines shown in the figure. We found similar systematic correction performance within 1 per cent of the continuum level in telluric lines across the observed spectrum in the analyzed orders (53, 54, 60 and 64).

6.4.2.6 Telluric correction for exoplanetary atmosphere retrieval

TAU was trained on solar observations but is intended to be used for correction of night time observations. To demonstrate the ability to perform corrections on night time data with much lower S/N, we replicated the detection of Na in the atmosphere of HAT-P-70b [ZHB+19] by [BACM+22], where the authors

analyzed 40 spectra collected during a transit event of this ultra-hot Jupiter on 18 Dec 2020 between 21:28 and 02:10 UTC. The Na doublet is in a spectral region particularly sensitive to telluric absorption, so this is a good exercise to test the applicability of `TAU` to exoplanet atmospheric retrieval. The 40 spectra were first corrected for tellurics using either `molecfit` or `TAU`. We then followed the same transmission spectroscopy analysis described in [BACM$^+$22]. The result is shown in Fig. 6.9. The `molecfit` corrections were performed in the same manner as for the solar observations, with a total computational time of approximately 7 hours (excluding additional time for optimizing parameters manually). We inspected the `molecfit` input line list and removed the potential source of the overcorrection near 5886 Å, as already done in [BACM$^+$22]. `TAU` corrections were performed using the extracted tellurics without manual parameter optimization with a total computational time for the 40 corrections of approximately 10 seconds.

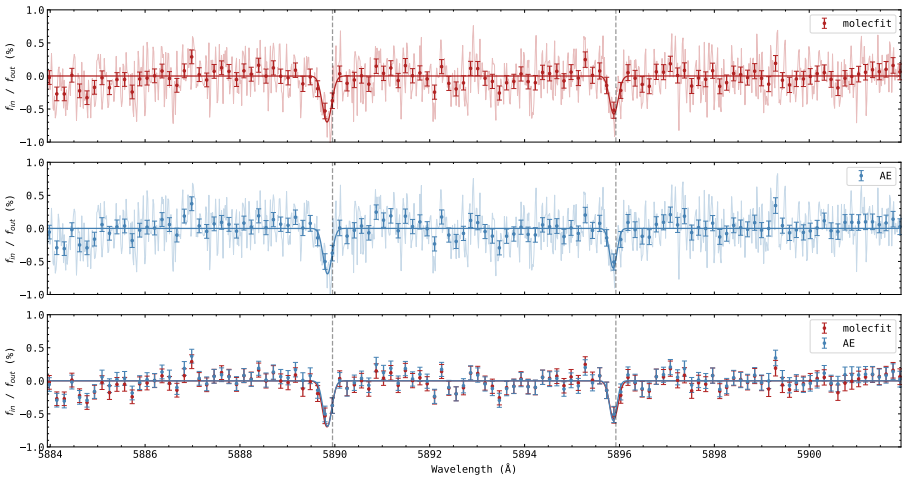


Figure 6.9: Transmission spectroscopy for the Na I doublet. The spectroscopy is performed using `molecfit` corrected spectra (top panel, red) and Autoencoder corrected spectra (middle panel, blue). The unbinned transmission spectra from either approach is shown in respectively light red and light blue while the binned master transmission spectra are shown as solid red and solid blue error bars. Each master transmission spectrum is fitted with a Gaussian model (solid red and solid blue lines). The vertical dotted gray lines show the theoretical location of each spectral line. The bottom panel shows a direct comparison of the binned master transmission spectra and the Gaussian models from either `molecfit` and `TAU`.

The retrieved Na D1 and Na D2 lines from the observations corrected with either

Table 6.4: Na D1 line results for the ultra-hot Jupiter HAT-P-70b using `molecfit` or `TAU` for telluric correction.

Type	`molecfit`	`TAU`
Depth [%]	0.58 ± 0.15	0.60 ± 0.16
Δv_{sys} [km s^{-1}]	-2.2 ± 1.4	-2.8 ± 1.2
FWHM [km s^{-1}]	10.9 ± 3.2	9.5 ± 2.9

Table 6.5: Na D2 line results for the ultra-hot Jupiter HAT-P-70b using `molecfit` or `TAU` for telluric correction.

Type	`molecfit`	`TAU`
Depth [%]	0.70 ± 0.14	0.69 ± 0.14
Δv_{sys} [km s^{-1}]	-5.9 ± 1.2	-5.4 ± 1.1
FWHM [km s^{-1}]	11.9 ± 2.8	11.1 ± 2.7

`molecfit` or `TAU` are shown in Tabs. 6.4 and 6.5. Due to the limited S/N of this data, we observe no significant differences in the results obtained with either `TAU` or `molecfit` corrections.

6.5 Discussion

In this section we discuss the results and touch on the advantages and limitations of the autoencoder model in addition to laying out the future work for our approach.

6.5.1 Comparison with `molecfit`

Our comparisons reveal how the autoencoder extracted spectra closely resemble the synthetic transmission spectra of H_2O and O_2 from `molecfit`. All telluric lines modelled by `molecfit` are also present in the autoencoder telluric endmembers. The two approaches however exhibit small differences regarding the depth, width and spectral location of lines.

The autoencoder telluric endmembers better match the actual width, depth and spectral position of observed tellurics leading to more accurate correction. Furthermore, corrections with `TAU` are performed approximately 10,000 times faster

than `molecfit`. This difference in compute time is significant and makes `TAU` much more feasible for correction of multiple spectra. Finally, the autoencoder approach can be used as a complementary data-driven validation tool to inspect the accuracy of synthetic approaches like `molecfit`, for instance by bringing attention to over-correction artifacts caused by line lists entry errors, not only in obvious cases like the 5886 Å telluric line in Fig. 6.4, but also in more subtle cases like the 5909 Å line in the same figure.

6.5.2 Advantages

The autoencoder approach for extracting telluric transmission spectra has certain advantages over other methods with the same aim. Firstly, since the approach is data-driven, it is possible to extract the combined telluric spectrum without relying on the precision of atmospheric measurements at the time of the observation, as well as external factors like molecular line lists. This is a major advantage over synthetic models, which are inherently limited by the precision of such factors. Furthermore, due to the high S/N and resolution of the training data from HARPS-N, it is possible to extract the spectral shape and location of telluric lines with very high accuracy. This type of accuracy on correction of telluric lines is increasingly important as we get closer to the 10 cm/s RV barrier for detecting Earth-like exoplanets.

Compared to other data-driven methods the autoencoder approach has the advantage that it can work and learn from immense datasets by training with mini-batch gradient descent. Traditionally neural networks have poor interpretability, but the highly constrained nature of the network architecture causes the learned representation to assume very specific properties, naturally leading to better interpretability of both the latent representation, as well as the extracted endmember spectra. The network constraints are highly customizable, allowing the framework to be adapted to other wavelength intervals or settings of spectral unmixing.

6.5.3 Limitations

For orders where the telluric components only exhibit minor variation, the encoder has difficulty learning the abundance variation. This is evident for order 60, where the H_2O component is much weaker than the O_2 component and thus constitutes less of the overall flux variation in the region. This caused the encoder to entangle the telluric abundances of H_2O and O_2, which lead to the decoder disentangling their spectra.

This correlation of features occurring in the encoding stage of an autoencoder trained with MSE is a known issue, which in the literature has been tackled by introducing either an orthogonality regularization term to the loss function [WYC+19] or by modeling the underlying low dimensional manifold as a product of submanifolds, each modeling a different factor of variation [FCMR21]. With the physical assumption that the H_2O and O_2 telluric components share at least a minor correlation through airmass, we circumvented the issue by stitching on a spectral region with strong tellurics, causing a disentanglement of the H_2O and O_2 endmembers while still allowing them to be correlated. The latent space correlation between the learned endmember abundances and consequently spectra, is likely not critical, as the correction of telluric lines in observed spectra is performed with the combined spectrum from all telluric components. From Fig. 6.7, we see how the combination of telluric components from H_2O and O_2 corrects for telluric lines in an unseen observed spectrum. As is evident, the correction is performed with high accuracy, suggesting the autoencoder can still accurately represent observed tellurics through a combination of moderately entangled telluric components. The question of whether the solar component has been fully disentangled (or to which degree) from the telluric components however remains.

Another potential limitation lies in the assumption that the components of the telluric spectrum can be seen as a fixed spectra scaled with an abundance weight. This assumption is based on the fact that temperature and pressure variations in the Earth atmosphere are likely small enough that the telluric spectrum can be represented like this. If the relative shape or spectral location of lines in the transmission spectrum of a molecular species undergo large temporal changes, then this assumption is unjustified. This could potentially be caused by large scale pressure and temperature variations in the Earth atmosphere. To retain high correction performance, this would mean that extracted endmembers used for correction would have to be learned from solar spectra observed relatively closely in time to the observation on which telluric correction is performed. However, our experiments on corrections across annual observing seasons show that no significant performance deterioration is apparent from temporal evolution. Finally, our physical model assumes an ideal spectrograph. If the instrumental profile changes significantly over time then our physical model is not valid. This makes `TAU` most applicable for well-stabilized spectrographs.

6.5.4 Future work

Future work includes extending `TAU` to other spectrographs and spectral ranges. This includes testing the approach on regions of saturated telluric lines in the redder part of the spectrum, which is currently a significant challenge for

ground-based astrophysics. Furthermore, the telluric weights used during telluric correction on new observations with the extracted endmembers are currently found using a least squares fit to known telluric lines in the spectrum. This could be improved by building `TAU` into a full forward model which includes stellar and telluric spectrum, such that these weights can be learned from all telluric lines simultaneously. This approach would be feasible due to the fast computation time of `TAU`.

6.6 Conclusion

We have introduced `TAU`, an open-source library that demonstrates a novel approach for extracting a compressed representation of observed solar spectra from HARPS-N using a highly constrained neural network autoencoder. We have shown how this representation can be used to extract endmembers that directly relate to the intrinsic solar spectrum as well as telluric transmission spectra of H_2O and O_2. After the autoencoder representation has been computed for a given spectrograph, the extracted components can be used to perform quick and accurate telluric correction on any observations from the same spectrograph. By comparing with `molecfit`, we have shown that the autoencoder extracted tellurics can be used as a master telluric spectrum with correction performance rivaling `molecfit` in large regions of the spectrum at a significantly reduced computational expense. `TAU` is data-driven and is thereby not affected by external factors like line list precision. This leads to improved correction accuracy over `molecfit` in various regions of the spectrum. `TAU` can be trained on observed solar spectra from any spectrograph and in this way extract a high-precision telluric profile for the given spectrograph. The extracted profile can be used to mitigate the telluric component in new observations, which could aid in the detection of faint radial velocity signals and atmospheric features of Earth analog exoplanets observed from ground-based telescopes.

Acknowledgements

We would like to thank HARPS-N for providing the public solar data used to train `TAU`. This data is based on observations made with the Italian Telescopio Nazionale Galileo (TNG) operated on the island of La Palma by the Fundación Galileo Galilei of the INAF (Istituto Nazionale di Astrofisica) at the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias. Without this vast quantity of high S/N data our approach would

not be feasible. Furthermore, we would like to thank `molecfit` for providing their open-source synthetic telluric correction framework, which we have not only used for initializing the autoencoder weights, but also for validating our extracted tellurics. Without a detailed synthetic telluric spectrum computed based on physical models, it would be significantly more complicated to validate our data-driven approach. Furthermore, we would like to acknowledge the DTU HPC cluster, which we have used for computing `molecfit` corrections. A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004)

6.7 Appendix A: Additional extracted endmembers

We provide additional results for extracted endmembers for order 53 and 54 in Figs. 6.10 and 6.11.

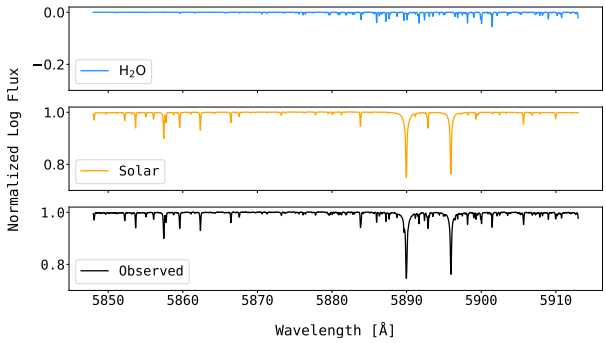


Figure 6.10: Extracted endmembers and an observed solar spectrum for order 53. The network has been trained using a two-dimensional latent space resulting in $R = 2$ endmembers. The extracted endmembers represent the H_2O (top, blue) and solar (middle, orange) components of the observed spectrum (bottom, black).

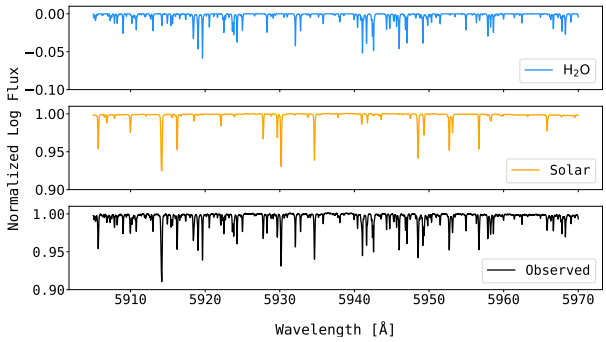


Figure 6.11: Extracted endmembers and an observed solar spectrum for order 54. The network has been trained using a two-dimensional latent space resulting in $R = 2$ endmembers. The extracted endmembers represent the H_2O (top, blue) and solar (middle, orange) components of the observed spectrum (bottom, black).

Chapter 7

Pantypes: Diverse Representatives for Self-Explainable Models

Rune D. Kjærsgaard[1], Ahcène Boubekki[2], Line K. H. Clemmensen[1]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Richard Petersens Plads 324, Kgs. Lyngby 2800, Denmark*

[2] *Department of Physics and Technology, The Arctic University of Norway (UiT), Tromsø*

Publication Status: Paper is submitted for review.

Submitted to: Association for the Advancement of Artificial Intelligence (AAAI 2024).

Abstract: Prototypical self-explainable classifiers have emerged to meet the growing demand for interpretable AI systems. These classifiers are designed to incorporate high transparency in their decisions by basing inference on similarity with learned prototypical objects. While these models are designed with diversity in mind, the learned prototypes often do not sufficiently represent all aspects of the input distribution, particularly those in low density regions. Such lack of sufficient data representation, known as representation bias, has been associated with various detrimental properties related to machine learning diversity and fairness. In light of this, we introduce *pantypes*, a new family of prototypical objects designed to capture the full diversity of the input distribution through a sparse set of objects. We show that pantypes can empower prototypical self-explainable models by occupying divergent regions of the latent space and thus fostering high diversity, interpretability and fairness.

7.1 Introduction

Machine learning (ML) systems are increasingly affecting individuals across various societal domains. This has put into question the black-box nature of these systems, and fostered the field of explainable AI (XAI), wherein model inference is corroborated with justifications and explanations in an effort to increase transparency and trustworthiness. In this line of research two approaches have arisen; that of ad-hoc black-box model explanations [SCD+17, YCN+15], and that of self-explainable models (SEMs) [CLT+19a, AMJ18]. A popular approach for SEMs substitutes traditional black-box networks with glass-box counterparts, where class representative prototypes are generated and used in the decision process [CLT+19a] leading to increased trustworthiness and interpretability.

The various initiatives emerging in the literature share the same overarching goals, but there is still a lack of consensus on the exact properties a SEMs should display [GHH+23]. We adopt three prerequisites properties of a SEM outlined in [GBH+22], namely *transparency*, *trustworthiness* and *diversity*.

Transparency may be defined by two properties; (i) the learned concepts are used in the decision making process without the use of a black-box model and (ii) the learned concepts can be visualized in the input space.

Trustworthiness may be defined by three properties; (i) the predictive performance of the model matches its closest black-box counterpart, (ii) explanations are robust and (iii) the explanations directly represent the contribution of the input features to the model predictions.

Diversity may be defined as the SEM using non-overlapping information in latent space to represent its concepts.

While significant work has been put forth in the literature to cement the transparency and trustworthiness axis of SEMs, only limited effort using qualitative measures exists for the diversity axis. Similarly, the relation between the diversity axis and appropriate inference remains largely unexplored. Diversity is typically ensured by introducing model regularization towards learning non-overlapping concepts [VL20]. However, this condition may not be strong enough, as non-overlapping concepts can still be learned in a small region of the input space, causing a lack of representativity for the full data distribution, known as representation bias [SLAJ22]. Representation bias can cause smaller sub-populations to remain hidden in low-density regions and ultimately cause biased inference [JXS+20]. To provide sufficient coverage and to mitigate the impact of data bias during model inference, it is critical to capture the full diversity of the data, and to have this diversity be represented in the prototypes learned by the SEM. To this end, we introduce pantypes, a new family of prototypical objects designed to empower SEMs by sufficiently covering the dataspace. Pantype generation is promoted using a novel volumetric loss inspired by a probability distribution known as a Determinantal Point Process (DPP) [KT+12a]. This loss induces higher prototype diversity, enables more fine-grained diversity control, and at the same time allows prototype pruning wherein the number of prototypes is determined dynamically dependent on the diversity expressed within each class. Prototype pruning enables the capacity to learn additional prototypes for complex classes and to grasp simple classes through a sparser set of objects, improving the interpretability of the class representatives.

Our contributions can be summarized as follows:

- Introduction of a volumetric loss, which promotes the generation of pantypes, a highly diverse set of prototypes.
- Quantitative measures for prototype representativity and diversity in SEMs.
- Dynamic class-specific prototype selection.

7.2 PanVAE

The modeling task at hand involves a classification setting on visual image data, where the SEM learns to classify $K > 0$ classes from a training set $X = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{N}$, where $\boldsymbol{x_i} \in \mathbb{R}^P$ is the i^{th} image and $\boldsymbol{y_i} \in \{0,1\}^K$ is a one-hot

label vector. We implement the pantypes on the foundation of a well-tested variational autoencoder based SEM, known as ProtoVAE [GBH+22]. This model uses an encoder function $f : \mathbb{R}^p \to \mathbb{R}^d \times \mathbb{R}^d$, to transform the input images into a posterior distribution $(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$. A latent representation $\boldsymbol{z_i}$ of the i^{th} image is then sampled from the distribution $\mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})$ and passed as input to a decoder function $g : \mathbb{R}^d \to \mathbb{R}^p$ to generate the reconstructed image $g(\boldsymbol{z_i}) = \hat{\boldsymbol{x}}_{\boldsymbol{i}}$. To enable transparent predictions, the model does not directly use the feature vector $\boldsymbol{z_i}$ during inference, but rather compares this vector to M prototypes per class $\boldsymbol{\Phi} = \{\phi_{kj}\}_{j=1,,M}^{k=1..K}$ via a similarity function sim : $R^d \to R^M$. The resulting similarity vector $\boldsymbol{s}_i \in \mathbb{R}^{K \times M}$ is then used in a glass-box linear classifier $h : \mathbb{R}^{\mathbb{M}} \to [0,1]^K$ to generate the class prediction $h(\boldsymbol{s_i}) = \hat{\boldsymbol{y}}_i$. The similarity function [CLT+19b] is given by:

$$\boldsymbol{s}_i(k,j) = \text{sim}(\boldsymbol{z}_i, \boldsymbol{\phi}_{kj}) = \log\left(\frac{||\boldsymbol{z}_i - \boldsymbol{\phi}_{kj}||^2 + 1}{||\boldsymbol{z}_i - \boldsymbol{\phi}_{kj}||^2 + \epsilon}\right), \tag{7.1}$$

where $0 < \epsilon < 1$. This construction allows the similarity vector to not only capture the distances to the prototypes, but to also reflect the influence of each prototype on the final prediction.

7.2.1 Loss Terms

To further enforce the properties of a SEM, we adopt the same prediction and VAE loss term structure as ProtoVAE:

$$\mathcal{L}_{\text{ProtoVAE}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{orth}}, \tag{7.2}$$

where

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{CE}(h(\boldsymbol{s}_i); \boldsymbol{y}_i) \tag{7.3}$$

is a cross-entropy (CE) prediction loss term ensuring inter-class diversity in the prototypes and

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = \frac{1}{N}\sum_{i=1}^{N} ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||^2 + \\ \sum_{k=1}^{K}\sum_{j=1}^{M} \boldsymbol{y}_i(k) \frac{\boldsymbol{s}_i(k,j)}{\sum_{l=1}^{M} \boldsymbol{s}_i(k,j)} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)||\mathcal{N}(\boldsymbol{\phi}_{kj}, \boldsymbol{I}_d))\end{aligned} \tag{7.4}$$

is the loss for a mixture of VAEs using the same network each with a Gaussian prior distribution centered on one of the prototypes [GBH+22]. Here $\boldsymbol{I}_d$ is a

$d \times d$ identity matrix. Finally, an orthonormality loss term is used:

$$\mathcal{L}_{\text{orth}} = \sum_{k=1}^{K} ||\bar{\boldsymbol{\Phi}}_k^T \bar{\boldsymbol{\Phi}}_k - \boldsymbol{I}_M||_F^2, \tag{7.5}$$

where $\bar{\boldsymbol{\Phi}}_k$ is the mean subtracted prototype vector for all prototypes of class k and $\boldsymbol{I}_M$ is an $M \times M$ identity matrix.

The orthonormality loss is included to foster intra-class prototype diversity and to uphold the diversity property of a SEM by inducing the learning of non-overlapping concepts in the latent space and thus avoiding prototype collapse [WLWJ21, JVLT21]. While this loss causes the prototypes to be orthogonal, it does not explicitly prevent the prototypes from occupying and representing a small region (volume) of the full data-space. Moreover, prototype orthonormality is typically achieved early during training, and further scaling of the orthonormality loss does not significantly alter the diversity of the prototypes (see results section).

Poor or skewed data representation, known as representation bias, has been associated with various detrimental properties related to ML fairness, where under-represented minority groups are negatively affected during inference [PJN+11a]. To mitigate these issues it is essential to achieve sufficient coverage of the full diversity represented in the data [SG19a]. We draw on this idea to empower the ProtoVAE model by exchanging its class-wise orthonormality diversity loss with a volumetric diversity loss, which causes the model to learn prototypical objects with various improved qualities, including an improved coverage of the embedding space. We call these learned objects *pantypes*. The loss term structure of our model is:

$$\mathcal{L}_{\text{PanVAE}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{vol}}, \tag{7.6}$$

where $\mathcal{L}_{\text{vol}}$ is the volumetric prototype loss, which not only prevents prevents prototype collapse, but causes higher prototype diversity, enables more fine-grained diversity control, and at the same time allows prototype pruning wherein the number of prototypes is determined dynamically dependent on the diversity expressed within each class.

7.2.2 Pantypes

Pantypes are prototypical objects learned in an end-to-end manner during model training. They are inspired by a probability distribution known as a Determinantal Point Process (DPP) [KT+12a], which can be used to sample from a population while ensuring high diversity. DPPs have recently garnered

attention in the ML community, and have been used to draw diverse sets in a range of ML applications including data from videos, images, documents, sensors and recommendation systems [GCGS14a, KT$^+$12a, LB12a, ZKL$^+$10a, KSG08a]. DPPs describe a distribution over subsets, such that the sampling probability of a subset is proportional to the determinant of an associated sub-matrix (a minor) of a positive semi-definite kernel matrix. The kernel matrix expresses similarity between feature vectors of observations through a kernel function $\boldsymbol{G}_{ij} = g(\boldsymbol{v}_i, \boldsymbol{v}_j)$. This global measure of similarity is then used to sample such that similar items are unlikely to co-occur. The kernel can be constructed in various ways including the radial basis function (RBF) kernel $\boldsymbol{G}_{ij} = e^{-\gamma||\boldsymbol{v}_i - \boldsymbol{v}_j||^2}$ or the linear kernel, leading to a similarity function of inner products known as the Gram matrix $\boldsymbol{G}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$. When using the Gram matrix, a DPP is equivalent to sampling with probability proportional to the volume of the paralellotope formed by the feature vectors of the sampled items. We utilize the linear kernel to construct a volumetric loss on the prototypes in the following way:

$$\mathcal{L}_{\text{vol}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\boldsymbol{G}_k|^{\frac{1}{2}}}, \tag{7.7}$$

where $\boldsymbol{G}_k \in \mathbb{R}^{M \times M}$ is the Gram matrix given by $\boldsymbol{G}_k = \boldsymbol{\Phi}_k^T \boldsymbol{\Phi}_k$ with Φ_{kj} as column vectors in $\boldsymbol{\Phi}_k$ and $|\boldsymbol{G}_k|$ is the Gramian (Gram determinant). $|\boldsymbol{G}_k|^{\frac{1}{2}}$ measures the M-dimensional volume of the parallelotope formed by the M columns of $\boldsymbol{\Phi}_k$ embedded in d-dimensional space. In other words, it expresses the diversity of the M prototypes of class k through the volume spanned by their feature vectors. This loss not only prevents prototype collapse by causing the feature vectors to diverge, but also directly encourages the pantypes to occupy different sectors of the data domain to express a large volume.

7.2.2.1 Prototype elimination

Increasing the scaling on the volume loss punishes pantypes that express a low volume and thus directly alters the diversity of the learned objects. With sufficient scaling, the volumetric loss forces pantypes out-of-distribution (OOD) if they are not necessary to represent the observed diversity of a class. This allows natural pruning, wherein the number of pantypes can be dynamically tuned by elimination of OOD pantypes. This is similar to the discipline of hyperspectral endmember unmixing, where a number of endmembers (prototypes) are disentangled from a hyperspectral image and linear combinations of the endmembers are used to reconstruct the input images. Following training, the learned endmembers can be associated with purity scores [BKL$^+$04], which express the quality of their explanations. These scores describe the maximal responsibility proportion of endmembers for reconstructing the original images. In other words, a high purity

score indicates that an endmember shares a high similarity with individual input images, while a low purity score indicates that an endmember is capturing noise and should be pruned. Such purity scores can be constructed from the similarity scores used in the linear classifier in our SEM. Thus, as proposed by [BKL+04], we can initiate the model with a sufficiently large number of pantypes, and use the similarity scores to prune individual OOD pantypes. We propose a heuristic for pruning, where a pantype can be pruned if it does not have the maximal similarity score for any of the training images (i.e. it does not individually represents any of the training images more than the other pantypes).

7.3 Results

We perform experiments across various real-world datasets to monitor the transparency, diversity and trustworthiness of PanVAE. These datasets are FashionMNIST (FMNIST) [XRV17], MNIST [LBBH98b], QuickDraw [HE17] and CelebA [LLWT15]. We demonstrate the trustworthiness of PanVAE by evaluating the predictive performance of the overall model and asses the diversity and transparency using qualitative assessments from visualizations of the input space, as well as quantitative measures of prototype quality and coverage. We compare PanVAE to the performance of ProtoVAE and ProtoPNet. The hyperparameters used for the experiments can be found in Appendix A.

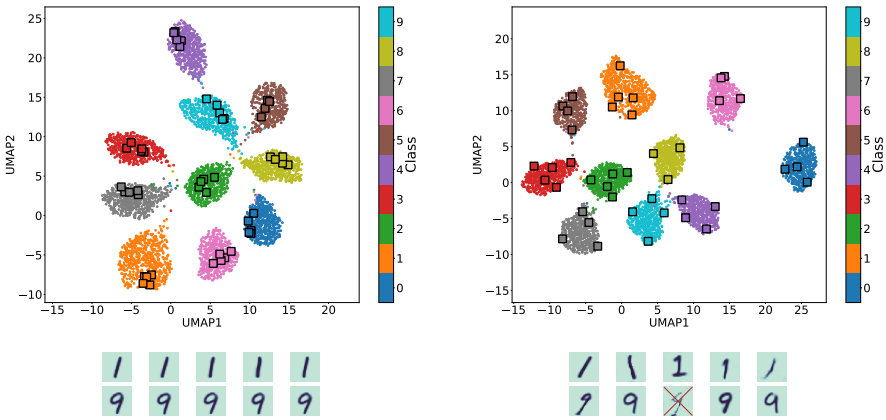
Table 7.1: Predictive performance (accuracy) of PanVAE ProtoVAE and ProtoPNet on MNIST, FMNIST, QuickDraw and CelebA. The values are the mean and standard deviation of three runs.

Dataset	ProtoPNet	ProtoVAE	PanVAE
MNIST	98.8 ± 0.1	**99.3 ± 0.1**	**99.4 ± 0.1**
FMNIST	89.9 ± 0.5	91.6 ± 0.1	**92.2 ± 0.1**
QuickDraw	58.7 ± 0.0	**85.6 ± 0.1**	**85.5 ± 0.1**
CelebA	98.2 ± 0.1	**98.6 ± 0.0**	**98.6 ± 0.0**

7.3.1 Predictive Performance

The results for the predictive performance are shown in Tab. 7.1, which demonstrates that PanVAE, like ProtoVAE, achieves higher predictive performance than ProtoPNET on the four datasets. There is no significant predictive perfor-

mance gap between PanVAE and ProtoVAE on the datasets. This underlines the trustworthiness of PanVAE.



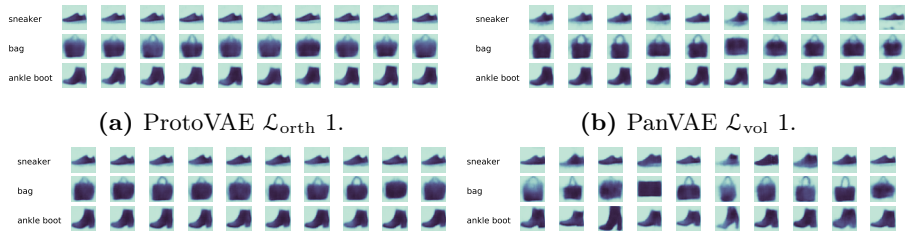
(a) ProtoVAE. **(b)** PanVAE.

Figure 7.1: ProtoVAE (a) and PanVAE (b) visualizations of the latent space and decoded prototypes learned on MNIST after 30 epochs of training. Top: UMAP representations of the latent space with learned prototypes overlaid as squares. Bottom: Decoded prototypes of class '1' and '9'. One of the prototypes from PanVAE does not have the maximal similarity for any training image, indicated by a red cross. PanVAE has captured variations in the digit '1' pertaining to right-handedness (first '1' from the left), left-handedness (second '1' from the left) and a traditional writing style (third '1' from the left).

7.3.2 Prototype Representation Quality

Firstly, we asses prototype representation quality using visual inspection of the learned prototypes and the associated latent space. This can be seen for the MNIST dataset on Fig. 7.1, where the prototypes for ProtoVAE and PanVAE are shown. The diversity of PanVAE is higher than ProtoVAE. The prototypes from ProtoVAE are mostly orthogonal in latent space, but only occupy a small region of the space. Contrarily, the volume loss in PanVAE has pushed the pantypes away from each other allowing them to occupy and represent diverse regions of the dataspace. This is reflected in the decoded prototypes, which show high diversity by representing various archetypical ways of drawing digits. For instance, the pantypes capture variations between left-handed and right-handed digits of "1" as well as the archetypical "1" with a horizontal base. Moreover,

PanVAE has found that the digits of "9" express less diversity and has thus pushed one of the pantypes OOD (indicated by a red cross in the figure). This form of prototype pruning by PanVAE allows the model to asses and represent the individual diversity expressed by each class.



(a) ProtoVAE $\mathcal{L}_{\text{orth}}$ 1. **(b)** PanVAE $\mathcal{L}_{\text{vol}}$ 1.

(c) ProtoVAE $\mathcal{L}_{\text{orth}}$ 100. **(d)** PanVAE $\mathcal{L}_{\text{vol}}$ 100.

Figure 7.2: Diversity control enabled by ProtoVAE and PanVAE. The figure shows the change in decoded prototype appearance as the respective diversity inducing losses are increased. The prototypes are shown for the FMNIST data of classes "sneaker", "bag" and "ankle boot" after 10 epochs of training. Figs. 7.2a and 7.2c show the difference between ProtoVAE prototypes with scale factor of 1 and 100 on the diversity loss $\mathcal{L}_{\text{orth}}$. Figs. 7.2b and 7.2d show the difference between PanVAE pantypes with scale factor of 1 and 100 on the diversity loss $\mathcal{L}_{\text{vol}}$.

Fig. 7.2 demonstrates the diversity control enabled by PanVAE by illustrating learned prototypes on the FMNIST datasets with different diversity loss scalings. The objective of the orthonormalization loss in ProtoVAE is to enforce intra-class diversity, and hence that the prototypes capture different concepts. While the loss ensures this, it only does so after sufficient training time (see Fig. 7.3 in Appendix 7.7). Fig. 7.2 shows that scaling the orthonormalization loss in ProtoVAE does not significantly alter the diversity of the representation. On the other hand, the volumetric loss in PanVAE allows direct control over the diversity of the representation.

Previous work in the literature on prototype based self-explainable classifiers often only qualitative asses the prototype diversity axis [GBH+22] (i.e. visual inspection of the diversity prerequisite of non-overlapping prototypes). We propose that self-explainable classifiers should not only be assessed with quantitative measures on the trustworthiness axis, but should also be evaluated by quantitative measures on the diversity axis. This includes thorough evaluations of how well the prototypes represent the dataspace. In order to do this we make use of measures of prototype quality and representativity by firstly measuring the prototype quality using the Davies-Bouldin (DB) index [DB79] and

secondly evaluating the diversity of the class representatives by assessing their data coverage.

7.3.2.1 Davies-Bouldin Index

The DB index is a measure of cluster quality defined by the average similarity between cluster C_i for $i = 1, ..., k$ and its most similar cluster C_j. The similarity measure R_{ij} quantifies a balance between inter- and intra-cluster distances. We adopt this measure and consider the prototypes in a SEM as cluster representatives and assign observations to their closest prototype in latent space according to maximal similarity scores. The intra-cluster size s_i is then measured as the average distance between prototype i and each data point belonging to the prototype, while the the inter-cluster distance d_{ij} is measured by the distance between prototypes i and j. From this the cluster similarity measure R_{ij} can be constructed such that it is non-negative and symmetric by:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}. \tag{7.8}$$

With these definitions in place the DB index may be defined by:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i=j} R_{ij}, \tag{7.9}$$

where a lower DB scores equates to a better representation of the underlying data. The DB scores for the different models can be seen in Tab. 7.2. PanVAE achieves the best DB scores in all cases, demonstrating the ability of the pantypes to represent the underlying dataspaces.

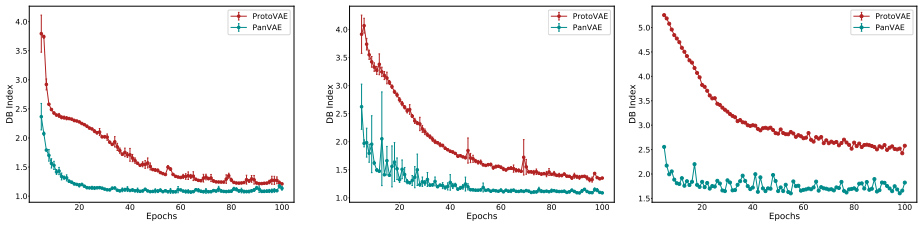
In addition to achieving higher final DB scores, PanVAE also does so using less training time. This is illustrated in Fig. 7.3, where the DB score evolution is shown for ProtoVAE and PanVAE over 100 epochs of training. PanVAE converges on a lower DB score much quicker than ProtoVAE. This diversity evolution behavior is also illustrated in latent space in Appendix B in Fig. 7.6.

7.3.2.2 Data Coverage

The DB index provides a measure of prototype quality in terms of prototype representation quality, but does not sufficiently asses how well the prototypes

Table 7.2: Davies-Bouldin scores of prototypes from the different models on the datasets used for our experiments. The values are the mean and standard deviation over three runs.

Dataset	ProtoPNet	ProtoVAE	PanVAE
MNIST	2.20 ± 0.18	1.21 ± 0.00	**1.13 ± 0.03**
FMNIST	3.43 ± 1.15	1.35 ± 0.01	**1.09 ± 0.01**
QuickDraw	2.52 ± 0.62	2.57 ± 0.01	**1.82 ± 0.01**
CelebA	27.09 ± 27.23	1.58 ± 0.15	**1.37 ± 0.01**

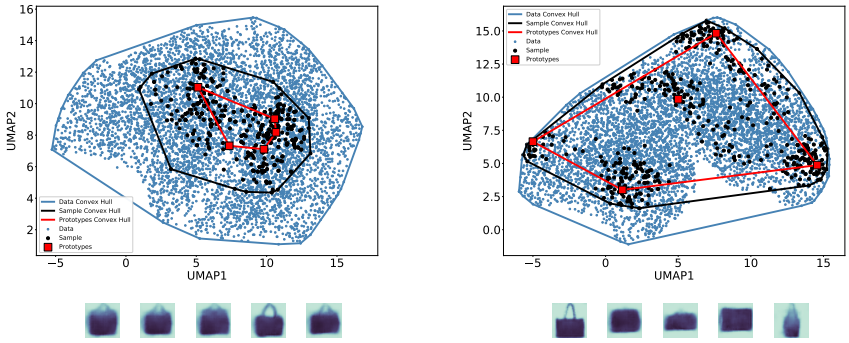


(a) MNIST. **(b)** FMNIST. **(c)** QuickDraw.

Figure 7.3: Evolution of prototype DB scores for PanVAE and ProtoVAE on MNIST, FMNIST and QuickDraw. Data points indicate mean values and associated standard deviations over three runs.

cover the diversity in the dataspace. Sufficient coverage of various aspects in the dataspace has been found critical in obtaining unbiased ML algorithms [JXS+20].

In order to asses prototype data coverage, we compare the volume spanned by observations represented by the prototypes to the volume of the full data distribution. Ideally, the prototypes are diverse enough, that they sufficiently cover a large volume of data they seek to represent. The coverage may be assessed through the volume of the convex hull of the data. We evaluate our pantypes on this premise by sampling the 100 nearest observations to each pantype. The proximity is measured in the full latent space in terms of the similarity score (Eq. 7.1). We then compute the volume spanned by the represented observations from their convex hull, and compare this to the volume of the original data. We illustrate the results of this procedure in Fig. 7.4 using a 2D UMAP projection of the 256 dimensional latent space for the "Bag" class in FMNIST. An additional coverage illustration for the "trouser" class is shown in Appendix B in Fig. 7.8. UMAP identifies a low dimensional manifold, where the data is uniformly distributed. The exact variation represented by this manifold can be hard to evaluate. Due to this we also demonstrate prototype coverage in PCA space for



(a) Bag prototypes from ProtoVAE. **(b)** Bag pantypes from PanVAE.

Figure 7.4: Prototype coverage in UMAP space from 20 epochs of training on FMNIST with 5 prototypes for the "bag" class for ProtoVAE (a) and PanVAE (b). Top: UMAP representations of the latent space with learned prototypes overlaid as red squares. The prototype convex hull in UMAP space is shown as a red outline around the prototypes and the full class dataspace convex hull is shown as a blue outline around the data. A sample of the 100 closest observations to each prototype is shown as black datapoints. The convex hull of the sampled observations is shown as a black outline. The PanVAE sample convex hull covers 77% of the volume of the full class convex hull, whereas the ProtoVAE sample convex hull covers 33%. Bottom: Decoded prototypes.

the FMNIST classes "dress" and "sneaker" in Appendix B in Figs. 7.9 and 7.10. We note that PanVAE has captured a severely underrepresented right-facing sneaker as a pantype in Fig. 7.10. Overall, the increased diversity of the pantypes allow them to occupy and represent a larger region of the dataspace.

7.3.2.3 Demographic Diversity

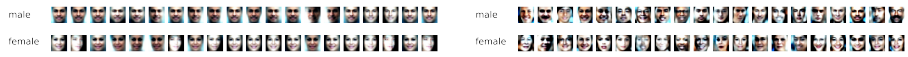
Sufficient representation of demographic groups has been found critical in ensuring ML fairness [JXS+20]. Image data used to train facial recognition algorithms have historically been biased towards White individuals, which are overrepresented in the training data, resulting in biased inference [BG18b]. The largest disparity is found between white skinned and dark skinned individuals.

Demographic diversity may be quantified using a measure of combinatorial

diversity, also known as diversity index [Sim49]. The combinatorial diversity is defined as the information entropy of the distribution [CDKV16a]:

$$H = -\sum_{i=1}^{k} p_i \log p_i, \tag{7.10}$$

where the combinatorial diversity measure H is the entropy, p_i is the probability of event i and $\sum$ is the sum over the possible outcomes k. This measure quantifies the information entropy of the demographic distribution over k demographic groups. A high entropy equates to a more diverse (fair) representation, which is not particularly biased towards any demographic group.



(a) ProtoVAE. **(b)** PanVAE.

Figure 7.5: Face prototypes learned on the UTK Face dataset. The learned prototypes are shown for ProtoVAE in (a) and for PanVAE in (b). PanVAE has captured variations in race as well as other unseen features such as facial hair in males. The ProtoVAE males all have somewhat neutral expressions with shut mouths while most of the females have slight smiles. The PanVAE males and females all exhibit large variations in expression from full smiles with visible teeth to neutral expressions without visible teeth.

We evaluate how the volumetric loss may aid in mitigating demographic data bias and enhance group level diversity. To do this we train PanVAE on the UTK Face dataset [ZSQ17], which contain images of about 20,000 individuals with associated sex and race labels. See Appendix A for the training details. The decoded facial prototypes from training on the UTK Face dataset can be seen in Fig. 7.5. To evaluate the demographic diversity, we asses the race of the nearest test image to each prototype and use this to compute the combinatorial diversity of the race distribution. The overall accuracy and diversity results are reported in Tab. 7.3. We also report the accuracy gap between White males and Black females. This accuracy gap has been identified as a ubiquitous problem in facial recognition algorithms. White males account for 23 percent of the individuals in the UTK Face data, while Black females account for 9 percent. PanVAE achieves a lower accuracy gap between these demographics due to a better accuracy on Black females. However, this comes at the expense of a lower accuracy on the majority sub-population of White males as compared to ProtoVAE.

Table 7.3: UTK results. The values are the mean and standard deviation of three runs. The overall accuracy is reported along with the individual accuracy and accuracy gap between White males and Black females. A positive gap value indicates that the mean accuracy is higher on White males compared to Black females. Demographic diversity is the information entropy of the distribution of races represented by the prototypes. The represented races are determined by the nearest test image to each prototype.

Metric	ProtoVAE	PanVAE
Accuracy All	**95.08 ± 0.11**	**95.42 ± 0.37**
Accuracy White Male	**96.35 ± 0.31**	95.21 ± 0.33
Accuracy Black Female	91.67 ± 0.53	**94.90 ± 0.39**
Accuracy Gap	4.69 ± 0.24	**0.32 ± 0.15**
Demographic Diversity	1.26 ± 0.06	**1.43 ± 0.07**

7.4 Discussion

The volumetric loss in PanVAE promotes the generation of diverse prototypes, which capture the underlying dataspace and represent distinct archetypical patterns in the data. This leads to increased representation quality and data coverage and can mitigate data bias. However, pantypes are most useful when the diversity expressed by the input data aligns with the diversity a study aims to enforce. This is closely related to the concepts of geometric and combinatorial diversity [CDKV16a], where geometric diversity expresses the volume spanned by a number of high-dimensional feature vectors and combinatorial diversity is related to information entropy of discrete variables. This means that geometric diversity is useful for ensuring what humans perceive as high *visual* diversity, while combinatorial diversity is useful for ensuring high *demographic* diversity (or fairness) of human understandable sensitive variables that take on a small number of discrete values (such as race). The volumetric loss in PanVAE exclusively ensures a large geometric diversity of the learned pantypes and as such only enforces visually diversity. This may not necessarily align with the diversity in unseen protected attributes such as race in facial image data. This misalignment can occur if features like background color and pose in the facial images exhibit larger visual variation than features related to demographic diversity such as skin tone. To enforce high demographic diversity, the images would either have to be pose aligned and background removed (or at at least background noise reduced) or the sensitive features would have to be incorporated directly into the model, if possible. We have trained PanVAE on the cropped and

aligned version of the UTK Face dataset to demonstrate that geometric and combinatorial diversity can be obtained simultaneously in noise reduced data with the volumetric loss. More balanced demographic representation can lead to better predictive performance for minority sub-populations in the data and consequently less disparate predictive performance between sub-populations. However, this usually comes at the expense of a reduction in performance for the majority group. Thus, the choice of representation should be carefully considered in coherence with the aim and target population of the trained model.

7.5 Conclusion

We have introduced pantypes, a new family of prototypical objects used in a SEM to capture the full diversity of the dataspace. Pantypes emerge by virtue of a volumetric loss and are easily integrated into existing prototypical self-explainable classifier frameworks. The volumetric loss causes the pantypes to diverge early in the training process and to capture various archetypical patterns through a sparse set of objects leading to increased interpretability and representation quality without sacrificing accuracy.

Acknowledgments

We would like to acknowledge the authors of the well-tested ProtoVAE. We have used the public code for this model as the foundation of PanVAE.

7.6 Appendix A

In this appendix we include the training details and the hyperparameters used for our experiments. The experiments for MNIST, FMNIST, QuickDraw and UTK Face have been performed on an Intel 6 core i7, UHD 630 CPU laptop. The experiments for CelebA have been performed on a GPU HPC cluster. For the MNIST, FMNIST and QuickDraw datasets we train the networks on images with the original dimensions from the published datasets. For CelebA we rescale the images from a dimension of 178 × 218 to a dimension of 224 x 224. For UTK Face we use the aligned and cropped version of the data and rescale the images from a dimension of 200 x 200 to 32 x 32. The UTK Face dataset contains images of all ages from 0-116. We filter the dataset to include any individuals over the age of 18. The UTK Face dataset is trained on 20 prototypes per class, which causes the initial volume of the randomly initialized prototypes to be excessively large, leading to computational precision issues. To resolve this, we downscale the volume loss kernel $\boldsymbol{G}_k$ with a multiplicative factor of $c = 0.1$ before computing the volume. This results in a volume loss of:

$$\mathcal{L}_{\text{vol}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|c \cdot \boldsymbol{G}_k|^{\frac{1}{2}}}. \tag{7.11}$$

For the MNIST, FMNIST and QuickDraw datasets we used the standard encoder and decoder structures reported for these datasets in the ProtoVAE paper [GBH+22]. For the UTK Face dataset we use the CIFAR-10 structure reported in the ProtoVAE paper. For CelebA we use a ResNet-34 encoder and the usual decoder designed to output 224 x 224 images.

The hyperparameters used for the experiments are reported in Tabs. 7.5 and 7.6. We tested a range of values in interval [0.01-1000] for the loss scaling parameters reported in Tab. 7.6. The final parameters were chosen to balance accuracy, DB scores and decoded prototype appearance.

Table 7.4: Overview of the datasets used for our experiments. K is the number of classes.

Dataset	N_{train} / N_{test}	Input Size	K
MNIST	60,000 / 10,000	28 x 28	10
FMNIST	60,000 / 10,000	28 x 28	10
QuickDraw	80,000 / 20,000	32 x 32	10
CelebA	162,770 / 39,829	224 x 224	2
UTK Face	16,000 / 3,210	32 x 32	2

Table 7.5: Overview of hyperparameters used for our experiments. LR indicates the learning rate and z Dim is the dimensionality of the latent space.

Dataset	LR	Epochs	Batch Size	z Dim
MNIST	$1e^{-3}$	100	128	256
FMNIST	$1e^{-3}$	100	128	256
QuickDraw	$1e^{-3}$	100	128	512
CelebA	$1e^{-3}$	50	128	512
UTK Face	$1e^{-3}$	50	128	512

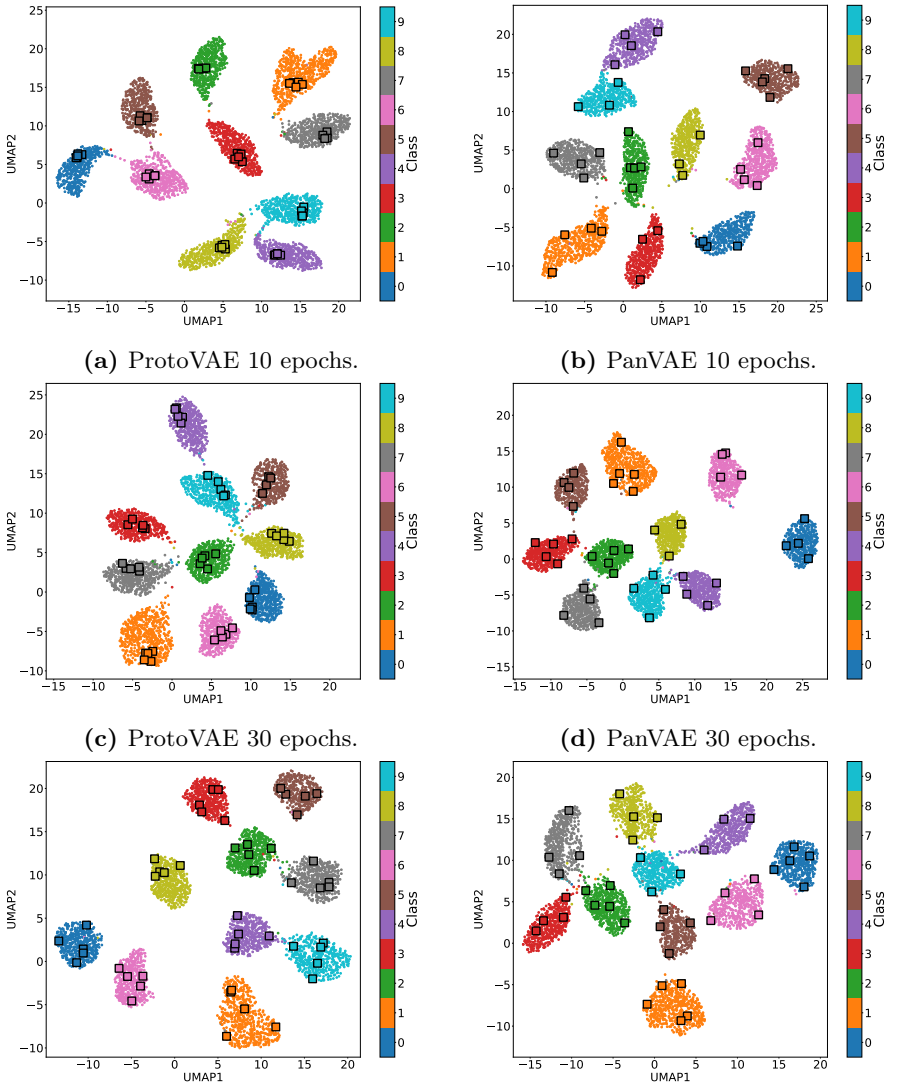
Table 7.6: Overview of number of prototypes pr. class (M) and loss term scalings used for our experiments. $\mathcal{L}_{\text{div}}$ indicates the scaling on the respective diversity inducing loss in ProtoVAE and PanVAE (orthonormalization or volumetric loss). $\mathcal{L}_{\text{rec}}$ is the reconstruction loss term in $\mathcal{L}_{\text{VAE}}$ and $\mathcal{L}_{\text{kl}}$ is the KL-divergence loss term in $\mathcal{L}_{\text{VAE}}$.

Dataset	M	$[\mathcal{L}_{\text{pred}}, \mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{kl}}, \mathcal{L}_{\text{div}}]$
MNIST	5	[1,1,1,1]
FMNIST	5	[1,1,1,1]
QuickDraw	10	[1,1,1,1]
CelebA (Proto)	10	[1,0.1,100,10]
CelebA (Pan)	10	[1,0.1,100,100]
UTK Face (Proto)	20	[1,1,1000,1]
UTK Face (Pan)	20	[1,1,1,0.1]

7.7 Appendix B

In this appendix we include additional illustrations of the concepts learned by PanVAE. All UMAP [MHM18] illustrations in this appendix and the paper throughout have been created using the following UMAP parameters: Minimum distance of 0.99, learning rate of 1.0, local connectivity of 1 and the number of neighbors at 25.

Fig. 7.6 shows the evolution of the latent space of MNIST during training. Here it is evident that PanVAE achieves high prototype diversity early in the training phase (as soon as 10 epochs), while the orthonormalization loss in ProtoVAE does not cause significant prototype diversity before 50 epochs of training. This finding is mirrored in Fig. 7.3, which shows that ProtoVAE uses significantly more training time to achieve good separation between the prototypes and that at separation convergence PanVAE achieves the best representation. Fig. 7.7 shows the final decoded prototypes learned on the QuickDraw dataset. Figs. 7.8, 7.9 and 7.10 show prototype data coverage on the FMNIST dataset in UMAP and PCA space.



(a) ProtoVAE 10 epochs.

(b) PanVAE 10 epochs.

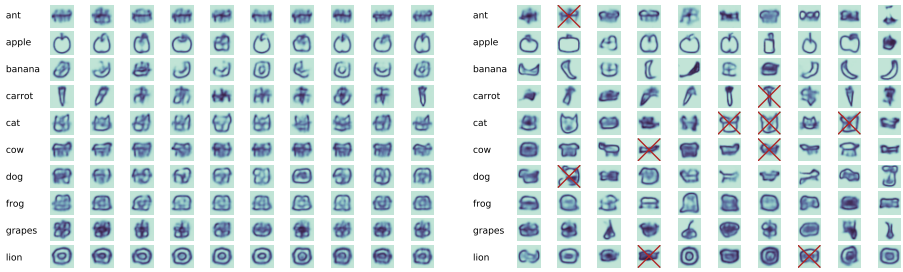
(c) ProtoVAE 30 epochs.

(d) PanVAE 30 epochs.

(e) ProtoVAE 50 epochs.

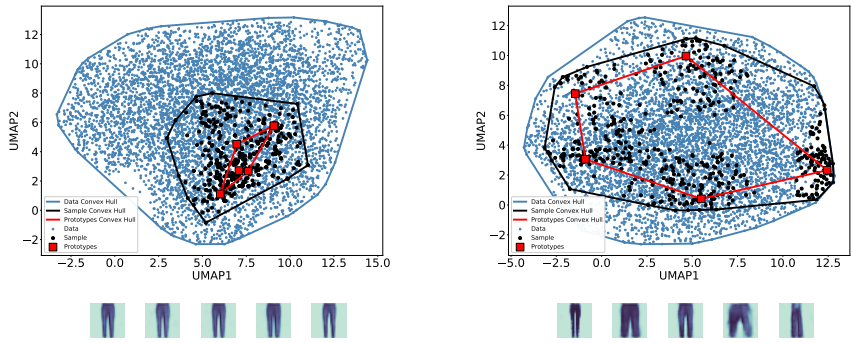
(f) PanVAE 50 epochs.

Figure 7.6: UMAP representations for the latent space of the MNIST data with overlaid latent representations of the prototypes (squares) for ProtoVAE and PanVAE respectively. The figures show the evolution of prototype latent location with training time. Both models are initiated with 50 total prototypes, but PanVAE is using prototype elimination and has eliminated 8 prototypes converging at 42 total prototypes.



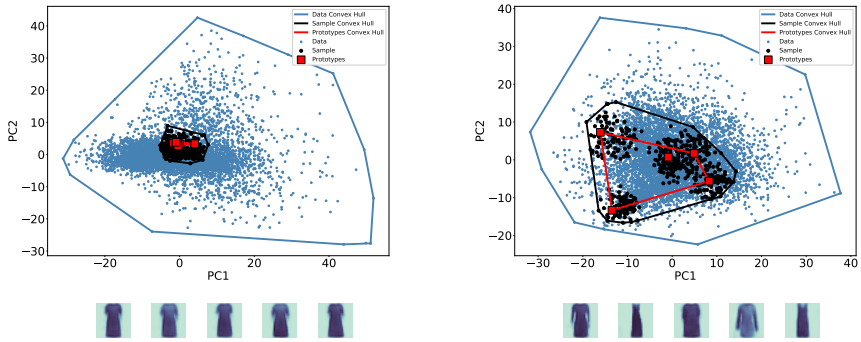
(a) ProtoVAE **(b)** PanVAE.

Figure 7.7: QuickDraw prototypes from ProtoVAE and PanVAE after 100 epochs of training with 10 prototypes per class. For PanVAE the pantypes that do not have the maximal similarity score with any training image have been marked with red crosses. In ProtoVAE all prototypes have maximal similarity score with at least one training image.



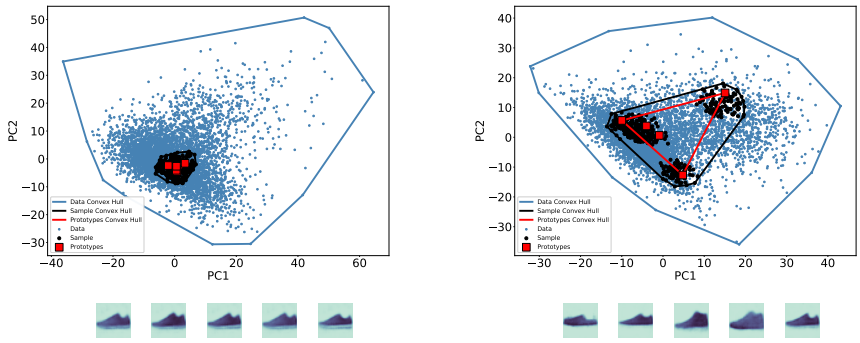
(a) Trouser ProtoVAE **(b)** Trouser PanVAE.

Figure 7.8: Prototype coverage from 20 epochs of training on FMNIST with 5 prototypes for the trouser class. The PanVAE sample convex hull covers 73% of the volume of the full class convex hull, whereas the ProtoVAE sample convex hull covers 25%.



(a) Dress ProtoVAE. **(b)** Dress PanVAE.

Figure 7.9: PCA coverage from 20 epochs of training on FMNIST with 5 prototypes for the dress class. The principal components account for 54% and 52 % of the variation in the Prototype and PanType networks respectively. The ProtoVAE sample convex hull covers 3% of the volume of the full class convex hull, whereas the PanVAE sample convex hull covers 24%.



(a) Sneaker ProtoVAE **(b)** Sneaker PanVAE.

Figure 7.10: PCA coverage from 20 epochs of training on FMNIST with 5 prototypes for the sneaker class. The principal components account for 65% and 68% of the variation in the ProtoVAE and PanVAE networks respectively. The ProtoVAE sample convex hull covers 2% of the volume of the full class convex hull, whereas the PanVAE sample convex hull covers 19%.

Chapter 8

Data Representativity for Machine Learning and AI Systems

Line K. H. Clemmensen[1,†], Rune D. Kjærsgaard[1,†]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Richard Petersens Plads 324, Kgs. Lyngby 2800, Denmark*

[†] These authors contributed equally to this work

Publication Status: Paper is under preparation.

To be submitted to: Association for Computing Machinery (ACM) Computing Surveys.

Abstract: Data representativity is crucial when drawing inference from data through machine learning models. Scholars have increased focus on unraveling the bias and fairness in models, also in relation to inherent biases in the input data. However, limited work exists on the representativity of samples (datasets) for appropriate inference in AI systems. This paper reviews definitions and notions of a *representative sample* and surveys their use in scientific AI literature. We introduce three measurable concepts to help focus the notions and evaluate different data samples. Furthermore, we demonstrate that the contrast between a representative sample in the sense of coverage of the input space, versus a representative sample mimicking the distribution of the target population is of particular relevance when building AI systems. Through empirical demonstrations on US Census data, we evaluate the opposing inherent qualities of these concepts. Finally, we propose a framework of questions for creating and documenting data with data representativity in mind, as an addition to existing dataset documentation templates.

8.1 Introduction

Machine learning and AI systems are increasingly governing important decisions affecting individuals at all levels of society. These automated decision frameworks have demonstrated various unwanted consequences as a result of biased data [PJN+11b, LI16, BG18a, RB19, MAR19, LGM88, ZOM19]. Oftentimes these systems are trained on samples (datasets) from a larger population. Biased results can arise if the sample does not accurately represent the target population, or if there is a lack of sufficient representation for subgroups within the data. While the literature of data bias in machine Learning and artificial intelligence (AI) systems is rich [SG19b], there exists only limited work on the connections between *data representativity* and AI systems. Terms like *representative sample* are used ubiquitously in the literature, often without further specification on the details or effects of this representativity. This paper analyzes and surveys data representativity in scientific literature relating to machine learning and AI systems by investigating how different notions of representativity are used and what effects adhering to different notions of data representativity has in relation to appropriate inference.

The term *representative sample* is an overloaded term and a generally accepted definition of what constitutes a *representative sample* (subset of observations) is hard to find in the literature. A few examples demonstrate that at least a couple of definitions of *representative sample* exist. The most general definition we found is from D'Excelle (2014) and states *"“Representative sampling” is a type of statistical sampling that allows us to use data from a sample to make conclusions*

that are representative for the population from which the sample is taken." [D4]. However, this definition leaves us with the important question of what we mean by representative. The following two examples of definitions clarify this point. 1) Meriam Websters' online dictionary says: *"Sampling in which the relative sizes of sub-population samples are chosen equal to the relative sizes of the sub-populations."* [Mer22]. 2) An online portal disseminating elementary statistics to graduate students writes *"A representative sample is where your sample matches some characteristic of your population, usually the characteristic you're targeting with your research."* [Sta22]. These examples illustrate that as we unfold the meaning of representative, questions arise, like what the target population is and which attributes/characteristics/sub-populations are relevant as well as how to measure a match between a sample and a population. OECD (Economic Co-operation and Development)'s definition of a representative sample acknowledges that several notions exist: *"In the widest sense, a sample which is representative of a population. Some confusion arises according to whether "representative" is regarded as meaning "selected by some process which gives all samples an equal chance of appearing to represent the population"; or, alternatively, whether it means "typical in respect of certain characteristics, however chosen"."* [OEC22]. Some of these ambiguities are linked: The definition of representative is linked to the target of the system/research/analysis and dictates which attributes, sub populations, and representative measures are of relevance. In this paper, we review various interpretations and notions of the term *representative sample* and link these to mathematical measures. Subsequently, we measure the match between the notions and the target of the analysis by looking at performance, diversity, and fairness metrics.

In 1979-1980 Kruskal and Mosteller wrote four papers on the term *representative sampling* with the motivation to unravel its ambiguities and imprecision [KM79a, KM79b, KM79c, KM80]. In addition, they called for caution as well as more specific expressions when referring to a representative sample. As they noted: "The reason for so much effort on one term is that the idea of representativeness is closely related to basic notions of statistical inference". In this paper, we take a closer look at data representativity for recent machine learning and artificial intelligence (AI) systems and before advancing, we will dwell on the nature of studies in AI and what this means for inference. AI systems are built both on observational data and on data from experiments gathered with the purpose of training the AI. Whereas randomized controlled experiments/trials are truly random samples, observational studies need to be carefully designed to tackle their inherent haphazardness [Ros10]. In observational studies, matching is performed to make treatment and control groups comparable, but unlike for experimentation, there is no basis for assuming that this extends to unmeasured factors [Ros10, Mon19]. Experimental studies are often used to make causal inferences, a basis which dates back to R.A. Fisher (1935) [Fis35]. However, causal relations can also be established through observational studies, like for

example the link between smoking and lung cancer [CHH+09]. We will therefore not further distinguish between the nature of the data or the AI systems.

As we draw conclusions from data or make predictions in artificial intelligence (AI) systems trained on data, it is important to understand what these data represent, and which inferences we can make. AI systems or machine learning (ML) models for decision making are widely used in industry and research, but care is not always put to the origin of the data, on which the systems are trained. This is for example seen in big data, where more data are considered better, and data often originate from a historical collection performed for e.g., control purposes or from scraping available internet sources rather than having been collected for the purpose, which it is later used for [BC11, KFK+20, Hua21, Ber17]. Other examples are more general for ML/AI and include representation bias stemming from the way we define and sample from a population, evaluation bias stemming from benchmark datasets with inherent biases, population bias when the distribution of attributes differ between dataset and target population, and sampling bias stemming from non-random sampling of subgroups [MMS+21, OCFK19, SG19c].

Amongst other, Kruskal and Mosteller found that *representative sample* was used as an assertive to underline a point without any scientific reasoning. Historically, the ImageNet competition has had a kind of implicit assertive, where scientists believed good results on the ImageNet dataset would mean good results for other image recognition tasks as well [MMS+21, DM20]. Torralba and Efros empirically illustrated in their paper 'Unbiased Look at Dataset Bias' (2011) that generalizations supporting this assertive were not necessarily a given, and described their findings as "if we add training data that does not match the biases of the test data this will result in a less effective classifier" [TE11].

Recently, focus has been put on the lack of transparency around dataset design and collection procedures as well as efforts to unbias existing datasets like e.g., the ImageNet [MMS+21, YQFF+20]. We will investigate these initiatives as well as the notions of a representative sample within the AI community. We have found sampling theories from the disciplines of analysis of physical material, design of experiments, as well as surveys in social sciences useful in terms of analyzing current practices and relating these to the ongoing work within ML and AI, where the historical emphasis on data representativity has been smaller.

To summarize, our contributions in this paper are:

- We provide an overview of the interdisciplinary topic of data representativity, organise the various notions of representativity, link mathematical measures to the notions when possible, and propose the use of three

measurable concepts.

- We describe and discuss relevant notions of representativity in literature and review their use in papers introducing datasets from the NeurIPS 2021 Track on Datasets and Benchmarks and ICCV 2021.
- We demonstrate contrasting perspectives on data representativity by empirically comparing two measurable concepts with opposing notions of representativity.
- We propose a framework of questions for creating and documenting data with representativity in mind.
- We provide new research directions on data representativity in ML and AI.

The rest of the paper is organized as follows. First, through literature about representative sampling, we will outline the general notions of a 'representative sample' (Section 8.2), give examples of their use in recent ML and AI literature, add mathematical measures for each notion, when possible, and propose to use three measurable concepts inn their place. In Section 8.3 we review the notions of representative sample used in the papers from the datasets and benchmarks track at NeurIPS 2021 and new benchmark datasets from ICCV 2021. Throughout these investigations, we find opposing opinions of sampling for coverage of the input space vs. probability sampling mimicking population distributions, which correspond to two of the measurable concepts. Consequently, we make empirical investigations demonstrating the qualities these opposing notions/concepts hold in Section 8.5. Finally, we suggest a framework for addressing data representativity in datasheets in Section 8.6 and round off with a discussion in Section 10.5.

8.2 Notions of a 'representative sample'

Since there is no specific, mathematical definition of a representative sample, and initial investigations identified at least a couple of different notions of what a representative sample is, we will review differing notions here.

Kruskal and Mosteller identified six notions/usages of a 'representative sample' in their first surveys from 1979 [KM79a, KM79b]: An assertive acclaim, absence of selective forces, a miniature of the population, an observation 'typical' or 'ideal' of the (sub)population, coverage of a population by the sample, and a reference to a sampling method later on specified in details. The sixth is a special notion in scientific writing, whereas the first five were found in both non-scientific as

well as scientific writing. We will use this framing here, and link more recent literature to these and add examples of their use in literature. We add existing mathematical or formal definitions belonging to the notions in the subsequent section.

Finally, we also add two novel notions we found in AI literature. We call them the copycat and no notion. Copycat refers to the creation of synthetic data representative of a target population. No notion refers to vague or no mentioning of representativity and likewise also no mentioning of non-representativity or limitations of the data representativity. The latter may seem harmless when presenting new datasets, but as the data is re-used, this can become harmful and an implicit notion of an assertive claim can grow in its place.

8.2.1 The assertive claim (the Emperor's new clothes)

The assertive claim, as described in the introduction, is used as an assertive to underline a point without any scientific reasoning and is dangerous both as a conscious acclaim and a subconscious notion when it comes without specification. It is recommended to avoid unjustified and unspecified use. We mentioned ImageNet as a historical example of an assertive claim of a representative sample [MMS+21, DM20]. Despite the broad acknowledgment of the ImageNet case as a cautionary tale on data representativity, the assertive notion continues to appear even in recent literature from acknowledged publication venues. One example is from the datasheet of a publication from the NeurIPS 2021 Track on Datasets and Benchmarks regarding time-sensitive questions [CWW21]: "It's sampled from large Wikipedia passages, it's representative of all the possible temporal-sensitive information."

8.2.2 The miniature (the model train set)

The miniature is best captured by Meriam-Webster's definition: "the relative sizes of sub-population samples are chosen equal to the relative sizes of the sub-populations" [Mer22]. This is a sample of the target population perfectly mimicking every (relevant) aspect (characteristic/distribution) of the population.

The miniature population has strong ties to the theory of sampling of physical material also related to chemical or biological analysis [Gy98, PME05a]. One of the guiding principles in the theory of sampling is to have as homogeneous a population (lot) as possible in order for a sample anywhere in the lot to mimic the lot best possible, which in turn minimizes sampling errors.

In other fields, it is common to subdivide the space into smaller groups, until each group exhibits homogeneity, and then randomly sample a miniature or a sample representative of that group with probability equal to the proportion of that group in the population [BJP13]. This is also referred to as strata sampling (simple random sampling within mutually exclusive groups of the target population/strata) or cluster sampling (random sampling of clusters/strata in the population and inclusion of all samples for the selected clusters) in fields like survey analysis [HJH12]. Ghojogh et al show that strata sampling always has lower variance than that of simple random sampling, in particular when strata have very different characteristics [GNG+20]. However, defining homogeneity in terms of subgroups may be delicate and constructing meaningful groups/strata is difficult if the population values/distributions are unknown.

Sampling from distributions is another way to construct miniature samples [Sha06, Mac05]. Sampling from distributions, and not least joint distributions, gives the possibility of matching distributions between sample and population rather than matching simpler characteristics, like e.g. averages. In high dimensions, these methods do suffer computationally, however. As a non-parametric alternative it is possible to sample from densities [KGC21b, RG16].

It is also possible to make a sample mimic certain characteristics of the population by sampling enough random samples to obtain a convergence in the measure of interest [BMZ21].

Recently, Yang et al (2020) [YQFF+20] proposed a framework to balance the demographics of ImageNet, but they also stated that this is only possible for one attribute at a time, as sub-categories will have too few samples if balancing across multiple attributes (e.g. race and gender). In consequence, the miniature analogy in itself breaks down, as we cannot account for all factors in the miniature, in particular not as the miniature decrease in sample size.

A concrete use of the miniature notion is seen in [dSMBC+21] where Machado et al predict suicide attempts in what they refer to as a representative sample of the US population. They write: "a representative sample of the adult population of the United States, oversampling black people, Hispanic individuals, and young adults aged 18– 24 years. ... Weighted data were adjusted to be representative of the civilian population ... data were weighted to reflect design characteristics of the NESARC and account for oversampling." The miniature notion is apparent in terms of reweighing characteristics to match the distribution of the population of interest. A certain notion of coverage and absence of selective forces can also be seen in terms of age and race, for which specific sampling strategies (oversampling) have been taken. This example illustrates that several notions are often used together, something we also note in our survey in Section 8.3.

In 'Understanding the Demographics of Twitter Users' by Mislove et al (2011) they conclude that Twitter users are not representative of the US population based on argumentation of non-matching demographic distributions for geography, gender, and race/ethnicity [MSLA+11]. This notion is related to that of a miniature, and we note that a dismissal of the representativity is in essence easier than proving it holds. However, even a dismissal of a sample as representative is limited to our understanding of the population. An understanding which for example is limited as explained by Taleb's Black Swan theory [Tal07, Tal20] about human's rationalization of rare and unpredictable events. Ruths and Pfeffer later on proposed eight steps to reduce biases and flaws in social media data [RP14], parts of these relate to the data collection and its documentation (similar to datasheets for datasets [GMV+21b]), and another part relates to correction for biases by population matching (miniature notion) or robustness testing across time and different samples.

8.2.3 Absence or presence of selective forces (justice balancing the scales)

Absence or presence of selective forces means that the sample is random as no forces are in play to select or de-select any specific types of observations in the target population; implying the purpose is to make inference about the target population, not the sample.

This notion ties to experimental modeling and coverage as follows. In the design of experiments literature, controllable factors and uncontrollable factors are distinguished [Mon19]. The controllable ones are indeed controlled to design as small an experiment as possible, yet with a suitable amount of observations and an appropriate coverage of the input space in order to make inference and optimize the response/output as a function of the controllable factors. Too many controlled factors make it hard to access all cross populations, and in addition there is no way of exhausting all possibilities.

Selective factors can also be uncontrollable or in the worst case go unnoticed. Examples of these are time-drifts in a production or non-response in surveys. These can pose problems to the statistical inference drawn from data. If observable, we can manage through our sampling design or sometimes even through post processing of data. However, unobserved or even unnoticed factors impose serious risks of bias and confounding.

In surveys, non-response is considered a substantial source of error caused by selection, one that is not directly related with the sampling. Selective forces can also influence survey responders through e.g., an interviewer effect. Errors

stemming from such selective forces can lead to potential biases, and several corrective efforts are usually applied to adjust for these [Gid12].

Selective sampling can also be performed on purpose, in survey sampling such examples are: quota sampling, purposive sampling, and referral sampling. These sampling designs are non-random and generalizations are therefore challenged, but sometimes samples of interest are so few, or participation recruitment so difficult, that convenience sampling designs can come in handy [Gid12].

In Kelly et al's 2019 opinion paper 'Key challenges for delivering clinical impact with artificial intelligence' [KKS+19], they mention representative sample as follows: "The curation of independent local test sets by each healthcare provider could be used to fairly compare the performance of the various available algorithms in a representative sample of their population." This notion of a representative sample speaks to some absence of selective forces in that it is believed each healthcare provider is best off providing its own sample, representative for their population, thus arguing for local models specific for a geographic area with specific demographics. Furthermore, distribution shifts are mentioned as a challenge for the AI models in healthcare, not only across healthcare providers, but also across time. This methodological discussion of whether a population should be seen as fixed or whether it itself is taken from an underlying stochastic process has ties all the way back to discussions from the 1903 ISI Berlin meeting (World Statistical Congress) [KM80].

8.2.4 Typical/ideal (Superman/Superwoman or the average man/woman)

Typical/ideal refers to typical or ideal exemplars which represent a population or subgroups of a population. This is not necessarily in a statistical sense, but may mean close to the average. An example is that in [LTG10], where cluster centers from Gaussian mixture models are sampled as representative observations of a larger dataset. In addition, a ML method like archetypal analysis [CB94b] carries some of this notion: Archetypes in the data are identified as linear combinations of the observations which describes a convex hull off the observations.

Another example of the notion of typical observations is from NeurIPS 2021, where typical names are sampled for construction of a dataset: "For each race and gender, we chose the top ten first names based on their overall frequency and representation within each group, excluding unisex names and names that differed by only one character." [LRD+21]. As the authors state: "The names we selected were derived using real-world data on demographic representations of first names, however demographic representation does not necessarily correlate

with implicit stereotypical associations".

We also found a use of a representative sample, meaning a sample representative of a specific target. In online tracking, this is used to help overcome occlusions when following a target in a video [OYC18]. This meaning is most related to that of typical exemplars, here typical of a specific target of interest.

8.2.5 Coverage (Noah's Ark)

Coverage seeks to include the heterogeneity of the population in the sample. A strong requirement for coverage would be that the sample should contain at least one observation from each relevant partition of the population. In contrast to the miniature, coverage does not require proportions within partitions to match those of the population. Harry V. Roberts suggested in 1971 sampling following the coverage notion in order to select a committee and avoid conscious and unconscious biases from appointing authorities [Rob71]. For the committee purpose, there is certain overlap with the typical/ideal notion. Along these lines, coverage is more about producing *representativeness* than about obtaining a likeness with the original population.

Density-based sampling approaches have proven useful under the coverage notion of representative sampling, where density estimates can be used to asses population imbalances and use this information for sampling to cover the heterogeneity of the population in the sample [KGC21b, RG16], or to reduce noise and improve performance in imbalanced classification settings [HLLL19].

An example where we meet the notion of coverage is in one of the recent proposals to address the lack of transparency around dataset collection and design in ML/AI, namely in *datasheets for datasets* by Gebru et al (2021) [GMV+21b]. One of the questions they propose concerns data representativity, and says: "Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable)." Apart from a clear notion of coverage, the description concerns some of the historical issues noted by the earliest endeavors of Anders Kiær (Director of Statistics Norway during 1877-1913) to go from full census to a representative sample, namely, how do we measure the representativeness? [KM80]. Coverage may or may not be what we go for, but if we go for it, how do we measure coverage, in particular considering joint distributions from several attributes? For

example if mean values or min/max of each attribute match between sample and population, this does not imply that the distributions of each attribute match between sample and target population. This only becomes more complex if we consider the joint distributions of the attributes. Second, we should note that if we strictly go for coverage, then distributions between sample and population most likely do not match, and e.g., variance or mean estimates based on the sample will differ. On the other hand, coverage has an intuitive attraction when it comes to inclusion and equality. We will demonstrate these aspects empirically in Section 8.5.

In a benchmark data publication with focus on real-world images [LGR+20], we additionally see a notion of coverage: "objectives for underwater image collection: ... a diversity of underwater scenes, different characteristics of quality degradation, and a broad range of image content should be covered."

Sampling with a notion of coverage in mind often means combining non-random and random sampling methods, whereas sampling with a miniature in mind often means using random probability sampling, for example strata sampling.

Coverage is also usually constructed purposefully to not mimic the underlying population, but rather to include the heterogeneity in the population, and this is often a preferred notion when fairness is part of the purpose of the modeling. In literature, some of the closest mathematical measures of coverage are those of diversity [CDKV16b].

8.2.6 Reference to sampling, later on specified

With this notion, the term 'representative sample' in itself becomes a 'vague term', and the exact meaning is specified in the context. Kruskal and Mosteller recommended this use of the term *representative sample*, bearing in mind that it needs always a specification. In their mind, the specification refers to the method of sampling, i.e., a description of how the data have been obtained. Apart from the sampling method/procedure we recommend also specifying the original population, the purpose of the sampling, and herein the notion (later refined to measurable concept) under which the sample is taken.

Another question Gebru et al propose to answer in a datasheet refers to the method of sampling [GMV+21b]: "If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?" Underlining the historical recommendations for a specification of sampling method when referring to a 'representative sample'. We will add, that any dataset is a sample of a larger set or population. In fact,

this question may also give some of the answers to the question of how the data is representative or not, as these answers heavily depend on the sampling strategy. Hopefully, answers are also well aligned with the first question in the motivation part of Gebru et al's datasheet, namely "For what purpose was the dataset created?" For some purposes, small sets of data, not generally representative of the entire population in question, can be good enough. Subsets of data may show that some characteristic thought to be absent or rare is in fact more frequent, or vice versa, that something thought of as universal is in fact missing to at least some degree. These subsets may be representative of only a part of the underlying population and thus form basis to dismiss one of the mentioned hypotheses, but not to draw any further inference about the entire population, see also [KM79c] for examples. With open source datasets, we should be careful, as the purpose or the hypothesis means we have collected specific data to enlighten us, and this data may not be useful to draw inference for other hypotheses or purposes.

As the method of sampling is specified, the notion of representativeness should be made clear and the reproduceability of the data/study possible. However, it is the notion and the purpose of the study that makes way for mathematical measures of the representativeness.

8.2.7 The copycat (synthetically generated)

This notion is used when real-world data (or parts of it) are copied or mimicked through synthetic data generation methods. In these settings the synthetic data are often claimed to represent the real-world data for instance by matching distributions. Alternatively, the synthetic data can be used to specifically target underrepresented regions of the original population distribution and thus be claimed more representative of uncommon instances than the original real-world data. Thus, synthetic data generation frameworks allow great flexibility and provide excellent test-beds for the study of data representativity. A recent example can be found in a paper publishing a novel text dataset [YIN+21]: "We took advantage of the synthesis pipeline to showcase how datasets can be constructed with properties that deliberately differ from real world distributions. Notably, we include samples of individuals with common (e.g., scientist) as well as uncommon occupations (e.g., spy)... and designed SynthBio to be more balanced with respect to gender and nationality compared to the original WikiBio dataset. ... Our paper takes the stance that in addition to evaluating on the world as it is, researchers benefit from having the option to evaluate their models on a more uniform distribution of the population. Synthesizing novel datasets is one technique that serves this goal. ... In addition, undesirable bias in real-world data, especially with respect to underrepresented groups, can be controlled in synthetic data, enabling evaluation of model performance on comparatively rare

language phenomena". We note that there are also notions of both a miniature and coverage in this example.

8.2.8 No notion

This notion, or rather lack thereof, indicates that it is simply not mentioned how or what the data may be representative of, or that it has no notion by not mentioning the limitations of the dataset. A newer benchmark dataset (from ICCV 2019) gives us an example of no notion of representativity [UPH+19]. They describe one of their contributions as: "A new object dataset from meshes of scanned real-world scene for training and testing point cloud classification". They indicate that the real-world scans of objects are more representative of problems expected to occur in vision tasks than computer generated object scans. This is undoubtedly true, but when it comes to the dataset as a benchmark of real-world scenes, the data representativity is more unclear. The real world examples consist of 15 categories of indoor objects; "we manually filter and select objects for 15 common categories". It is unclear how the categories were chosen or what they are representative of in terms of a larger population of common indoor objects. We recommend more explicit descriptions of data representativity, sampling and its purpose, see also [GMV+21b]. In some circumstances outlining the limitations of the data representativity may be more sensible than outlining the target population.

8.2.9 Measurable concepts

The notions do not provide clear definitions and sometimes several notions share the same underlying concept. Additionally, mathematical measures are not necessarily applicable to all notions. In this section we relate the notions to overarching concepts and connect these to mathematical measures that can be used to assess the concepts. In addition, we provide mathematical measures which can be used to asses the impact of data representativity on fairness.

We define three operational concepts for data representativity, see Table 8.1. 1) A sample as a *reflection* of the target population - mimicking the population distribution. The representativity can be measured by comparing the distributions of sample and target or by comparing specific measures (like averages) of interest. 2) A sample providing *coverage* of the population. The coverage of the sample can be measured through existing diversity measures of the sample (like geometric diversity or entropy). 3) Samples as *representatives* of subgroups in

the population, where the representativity e.g., can be measured through cluster metrics like the average distance to the representative within the subgroup.

The miniature and coverage notions naturally fit into the reflection and coverage concepts, respectively. Synthetic data (copycat) are often devised according to a reflection concept, but can in also be devised according to a coverage concept. The notion of selective forces likewise fits into either the reflection or coverage concept depending on the aim. The reference to sampling notion is also context dependent conditional on the specified sampling procedure. This notion can adhere to any of the three concepts depending on the specified sampling methodology and aim of the study.

Table 8.1: Overview of notions, concepts and related mathematical measures.

Concept	**Notion**	**Description**	**Examples of existing mathematical measures**
-	Assertive claim	Claiming representativeness without justification	None - Avoid
-	No notion	No indication of data representativity	None - Avoid
-	Reference to sampling	Special notion specified in context of sampling method	Context dependent
Reflection	Miniature	Sample mimics population distribution	Averages and average predictions as well as distributional comparisons between sample and population
	Selective forces	Truly random sample in observational studies like e.g. surveys	
	Copycat	Synthetic data created to mimic real-world data distribution	
Coverage	Coverage	Sample provides coverage by broadly representing the heterogeneity/diversity of the population	Diversity measures of the sample e.g. geometric coverage
	Selective forces	Truly random samples in experimental studies	
	Copycat	Synthetic data created for balanced coverage of real-world dataspace	
Representatives	Typical/ideal	Single observations are representatives of a group in the population	The representatives are e.g. approximated by the mean, median or mode of the group

The notion of a representative sample as an assertive claim, whether explicit or implicit should be avoided as it is not measurable. Not having a notion is likewise not recommended as it is not measurable and may lead to an implicit assertive use of the dataset. While Kruskal and Mosteller preferred to use a notion of a 'representative sample' as a vague term with the sampling procedure specified later, we recommend clearly stating a motivation for data representativity and to subsequently thoroughly document the sampling procedure and methods. We argue that a more explicit use of one of the measurable concepts of representativity will make the aim clearer; giving the sampling documentation a context in which it can be evaluated.

The reflection and coverage concepts often work from contrasting perspectives on data representativity and carry different inherent advantages and disadvantages. We will demonstrate these in Section 8.5.

8.2.9.1 Reflection

This concept may be assessed in various ways. As a first approach, statistical tests on averages and average predictions can give an indication of generalization between sample and population. Additional central tendency measures like median and mode and statistical dispersion measures like variance and interquartile range may also be analyzed. Furthermore, the notion may be examined by analyzing the distributions, for instance measuring the distributional departure of the sample from the population. This departure can be measured through the ℓ_∞ norm equivalent to the Kolmogorov-Smirnov (KS) statistic [LRC05]. Most distributional measures operate in one dimension, but some can be extended to compare multivariate distributions. For instance the generalization of the KS two sample statistic for 2D and 3D distributions due to Peacock [Pea83]. Other tests, like the maximum mean discrepancy (MMD) [GBR+12] are designed for comparing multidimensional distributions. Generally, the tests for comparing multidimensional distributions are computationally expensive for large, high-dimensional samples. Another popular distributional distance measure is the general Wasserstein distance [Vas69, Kan60] given by:

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Gamma(\mu,\nu)} \int_{M\times M} d(x,y)^p \mathrm{d}\pi(x,y) \right)^{1/p}, \tag{8.1}$$

where $p \geq 1$ and W_p is the p^{th} Wasserstein distance, $\Gamma(\mu,\nu)$ denote all joint distributions π that have marginals μ and ν, d is a metric (distance function) between points x and y that are being matched and M is a given metric space. When $p = 1$ the distance is also known as the Earth Mover Distance and carries a nice intuitive interpretation of visualizing the two distributions as piles of earth

(soil). The distributional departure is then measured by the work required to turn one pile into the other through an optimal transport problem.

8.2.9.2 Coverage

Coverage may be quantified through diversity measures. We bring attention to measures evaluating either combinatorial information, called *combinatorial diversity*, $C(\cdot)$ or geometric coverage, called *geometric diversity*, $G(\cdot)$. To define these metrics consider a set of observations X and a discrete categorical feature with k categories. This gives rise to a partition of the dataspace into k parts $X = X_1 \cup X_2 \cup .. \cup X_k$, leading to a combinatorial measure of diversity. The combinatorial diversity of a subset $S \subseteq X$ is defined as the Shannon entropy of the distribution [CDKV16b]:

$$C(S) = -\sum_{i=1}^{k} s_i \log s_i, \tag{8.2}$$

where the combinatorial diversity measure $C(S)$ is the Shannon entropy, $s_i = \frac{|S| \cap |X_i|}{|S|}$ is the probability of event i and $\sum$ is the sum over the possible outcomes. Thus combinatorial diversity (also known as diversity index [Sim49]) has roots in information theory and measures the degree of diversity through the Shannon entropy of the distribution. High entropy corresponds to high diversity. The combinatorial diversity measure is useful to quantify diversity in features with a set of discrete human-interpretable values [CDKV16b] such as race.

On the other hand, geometric diversity is motivated from a volumetric perspective [CDKV16b]. Each datapoint $x \in X$ is represented by a feature vector v_x. The geometric diversity of a subset $S \subseteq X$ is the n-volume of the parallelotope spanned by the n feature vectors $\{v_x : x \in S\}$, where $n = |S|$ is the size of the subset. Denoting the data matrix of the subset S as $\mathbf{D} \in \mathbb{R}^{p \times n}$, the (squared) n-volume of the n-parallelotope embedded in a p-dimensional space (where $p > n$) can be computed by means of the determinant of the Gramian matrix $\mathbf{G} = \mathbf{D}^T\mathbf{D}$ (with feature vectors as columns in $\mathbf{D}$). Thus the geometric diversity can be measured by:

$$G(S) = \sqrt{\mathrm{Det}(\mathbf{D}^T\mathbf{D})}, \tag{8.3}$$

where $G(S)$ is the geometric diversity of subset S, $\mathrm{Det}(\cdot)$ denotes the determinant and $\mathbf{D}$ is the data matrix of the subset S. Geometric diversity is motivated from a perspective of diverse feature vectors. Intuitively, diverse vectors can be interpreted as divergent and thus pointing in different directions. The

diversity of these can be measured by the volume of the parallelotope spanned by the vectors. Thus, the larger the volume, the higher the geometric coverage. Geometric diversity is closely related to a type of probability distribution known as *determinantal point process* (DPP) [KT$^+$12a], which can be used to draw samples proportional to their geometric diversity.

While geometric diversity can be a good measure of the coverage for a sample, or between different samples of the same size from the same population, it does not directly relate to the degree of coverage in the original population space. As the metric evaluates n-dimensional volumes, comparing geometric diversity between sample and population equates to comparing different dimensional volumes. On the other hand, comparing between different sized samples through combinatorial diversity is straightforward, as this measure operates intrinsically on normalized probabilities.

8.2.9.3 Representatives

A typical or ideal observation may be estimated as the mean, centroid or mode of the group it represents, and the representativeness may be measured by the variance in the group. Furthermore, in settings where representativeness of an underlying population is sought through data reconstruction from a combination of archetypes (ideal exemplars), the representativeness of these archetypes can be measured through a reconstruction loss between original and reconstructed data.

8.2.9.4 Fairness measures

Analogous to how the notions of representativity may be measured mathematically, various measures also exist to quantify the adverse effects of insufficient representation, known as representation bias. Representation bias occurs when parts of the input space are underrepresented [SG19b, SLAJ22], for instance a sampled population which underrepresents and fails to generalize well for parts of the population, which can manifest as disparate predictive accuracy for these groups [CJS18, AJJ19, JXS$^+$20]. Common cases include models trained on ImageNet [DDS$^+$09, SHB$^+$17] and commercial facial analysis algorithms [BG18c]. While these models are not intrinsically unfair, they may capture and increase biases present in the training data. This inherited bias can be measured through algorithmic fairness metrics.

Algorithmic fairness is often formulated in terms of independence relations

between model predictions $\hat{Y}$ and a protected attribute A (typically a binary feature $A \in \{0, 1\}$) denoting group membership under a protected category such as race or sex. A common notion of algorithmic fairness, known as demographic parity (or statistical parity), is defined to require independence between model predictions and a protected attribute:

$$\hat{Y} \perp\!\!\!\perp A \tag{8.4}$$

In regression settings $\hat{Y}$ is a real-valued random variable characterized by its cumulative distribution function (CDF). The departure of the CDF of $\hat{Y}$ from the CDF of $\hat{Y}$ conditional on the protected attribute A can be used as a measure of demographic parity [ADW19, RD21]. For a binary decision problem $\hat{Y} \in \{0, 1\}$ with a binary protected attribute $A \in \{0, 1\}$ the demographic parity constraint can be expressed by $P\{\hat{Y} = 1|A = 0\} = P\{\hat{Y} = 1|A = 1\}$ [HPS16], thus requiring equality of positive rates for subsets of the protected attribute.

The demographic parity criterion has been critiqued on various accounts [DHP+12, HPS16], which has lead to an alternative formulation of algorithmic fairness known as equalized odds. Equalized odds formulates the following conditional independence:

$$\hat{Y} \perp\!\!\!\perp A \mid Y, \tag{8.5}$$

The equalized odds constraint applies to targets and protected attributes in any space [HPS16]. For binary classification with a binary protected attribute, the constraint can be formulated as $P\{\hat{Y} = 1|A = 0, Y = y\} = P\{\hat{Y} = 1|A = 1, Y = y\}, y \in 0, 1$, where y is the outcome. In this setting $Y = 1$ is often considered the advantaged outcome, which leads to a popular relaxation of the equalized odds measure known as *equal opportunity* [HPS16]. This measure prohibits discrimination only within the advantaged outcome group and can be formulated as $P\{\hat{Y} = 1|A = 0, Y = 1\} = P\{\hat{Y} = 1|A = 1, Y = 1\}$. Equal opportunity thus requires equality of true positive rates.

8.3 Survey of use in AI literature

To provide insight into the use of data representativity in current literature, we conduct a survey of papers from two typical and highly recognized AI conferences; The Conference on Neural Information Processing Systems (NeurIPS) and The International Conference on Computer Vision (ICCV). We restrict our survey to papers contributing novel datasets at either the NeurIPS 2021 Track on Datasets and Benchmarks or at the main conference at ICCV. The NeurIPS track has 174 accepted papers contributing either high-quality datasets, new benchmarks or discussions on data related work; 108 of them contribute novel datasets. These

papers are required by NeurIPS guidelines to provide dataset documentation and intended uses. The organizers recommended using documentation such as datasheets for datasets, which encourages the authors to consider and document how their work relates to data representativity. The main conference at ICCV 2021 has 1612 accepted papers, of which we identify 32 contributing novel datasets. We conduct the survey by reviewing which notions each paper uses to describe the representativity of their dataset. A summary of the survey results can be found in Table 8.2.

Table 8.2: Summary of notions used in the 108 papers introducing novel datasets on the NeurIPS 2021 Track on Datasets and Benchmarks and 32 papers introducing novel datasets from ICCV 2021. For each paper we document which notion they use to describe the representativity of their dataset. A paper can use more than one notion of representativity.

	NeurIPS 2021		ICCV 2021	
Notion	**Number of papers**	**Percent**	**Number of papers**	**Percent**
No notion	2	1.9 %	1	3.2 %
Assertive	10	9.3 %	2	6.3 %
Miniature	15	13.9 %	10	31.3 %
Selective Forces	41	38.0 %	7	21.9 %
Typical / Ideal	14	13.0 %	4	12.5 %
Coverage	66	61.1 %	27	84.4 %
Reference to sampling	108	100.0 %	30	94 %
Copycat	18	16.7 %	5	15.6 %

8.3.1 Examples

We find that the various notions appear in a wide array of settings and range from implicit to explicit use. Here we provide noteworthy examples demonstrating how the authors use the notions to express the representativity of their datasets.

8.3.1.1 Assertive claim

The assertive claim appears in about 9% of surveyed NeurIPS publications and 6% of the surveyed ICCV papers. An example of the notion can be found in the datasheet of a publication regarding time-sensitive questions [CWW21]: "It's sampled from large Wikipedia passages, it's representative of all the possible temporal-sensitive information." Further examples include [MNJ+21]: "We select

16k most representative scenes and exhaustively annotate all the 3D bounding boxes of 5 categories..." and [MBC+21] "This data contains examples of slang, acronyms, lack of punctuation, poor orthography, concatenations, profanity, and poor grammar, among other forms of atypical language usage. This data is representative of the types of inputs that machine translation services find challenging."

8.3.1.2 Miniature

The miniature notion appears in roughly 14 % of the surveyed NeurIPS papers and 31% of the surveyed ICCV papers. It emerges in various settings including demographic population representativity of people [HWB+21]: "Tab. 1b shows a statistical summary of the eligible cohort. This cohort broadly reflects the Tufts student population in terms of age, racial and gender makeup." Likewise the notion is used in relation to population distribution of animals in a paper regarding animal pose estimation [YXZ+21]: "... the number of images in each family of AP-10K has a long-tail distribution, which reflects the true distribution of animals in the wild due to the commonness or rarity of the animals in some extent."

The notion also appears in more restricted forms, for instance claiming a miniature in terms of a specific geographic region [KTR+21]: "The class imbalance provides a challenge for machine learning algorithms but it is representative of the geographic region and an imbalance is generally common in real-world crop type mapping tasks." Furthermore, the notion also appears under the disguise of 'representative coverage' [RBM+21]: "By restricting a dataset to only those tweets matching a pre-defined vocabulary, a higher percentage of hateful content can be found. However, this sacrifices representative coverage for cost-savings, yielding a biased dataset whose distribution diverges from the real world we seek to model and to apply these models to in practice."

Finally, some publications state that their data are not representative in terms of the miniature notion [BD21]: "Two thirds of the dataset concentrate on as few as four countries: Germany, France, the UK, and Spain. This distribution is not representative of the actual distribution of church buildings across Europe but most likely correlated with the size and level of activity of the local Wikipedia communities and their propensity to enter information in Wikidata."

8.3.1.3 Selective forces

The selective forces notion appears in 38 % of the surveyed NeurIPS papers and 22% of the surveyed ICCV papers. The notion is mostly used to claim non-representativity due to the presence of selective forces in the sampling process. Examples include: [HWB+21]: "First, our dataset is limited in whom it represents. Because we draw from a convenience sample at our university, ages are skewed toward typical college students and the racial makeup reflects that our campus community is largely white and Asian." Another example is [Gil21] "Because our dataset comprises only named and published chaotic systems, it does not comprise a representative sample of the larger space of all low-dimensional chaotic systems." Yet another example includes a discussion on the difficulty of dealing with multiple selective forces [ARZV21]: "... sampling images randomly from an uncurated large collection removes specific biases such as search engine selection but not others, for example the geographic bias. Furthermore, we added one significant bias: there are no people in these pictures, despite the fact that a large fraction of all images in existence contain people"

On the other hand the absence of selective forces is used as an indication that no sampling biases exist, and hence that the data is representative of the population [ANM+21]: "To avoid introducing biases, all comments of the RP have been considered without further topical filtering. Furthermore, using a broad crowd to annotate the data should minimize the inclusion of person- specific biases."

8.3.1.4 Typical / Ideal

This notion is used in 13 % of the surveyed papers and also appears in relation to synthetic data generation, where representativity of the underlying population is sought modelled through variation on ideal / archetypical patterns or shapes. For example [KL21]: "The first part of the template specification describes a base sewing pattern that would then be parametrized and varied to produce new designs." "The training group of 12 templates aims to cover design spaces of typical simple garments, including skirts, dresses, tops, pants, jackets, hoodies, and jumpsuits, and reflect some topological variations among them."

The notion is also used to express representativity of a population through typical systems or methods [OGB+21]: "We propose four representative physical systems, as well as a collection of both widely used classical time integrators and representative data-driven methods (kernel-based, MLP, CNN, nearest neighbors)".

8.3.1.5 Coverage

The coverage notion is popular appearing in 61 % of the surveyed NeurIPS papers and 84% of the surveyed ICCV papers. The notion can be found in a wealth of settings and often appears as a claim of diversity in the data, for example [KL21]: "... the motivation was to resemble the variety of designs that exist within a garment type while covering this diversity uniformly." Another example is [KSC+21] "A diverse quantity of wild and lab culture mosquitoes is included in the database to capture the biodiversity of naturally occurring species." The notion is also use in terms of language coverage in [MCB+21]: "The dataset is our best effort to extract and represent as much diversity (in terms of various different languages) from Common Voice as possible." Additionally, the notion is commonly used in ICCV papers to support that the published image data is representative of the real world in terms of visual diversity [RRR+21]: "...we wanted scenes that are as photorealistic and visually diverse as possible."

8.3.1.6 Copycat

The copycat notion can be found in about 16 % of the surveyed papers and appears mostly in relation to synthetic data generators constructed to copy or mimic the distribution of real-world data. For instance [LKWN21]: "The synthetic datasets we release offer a wide variety of parameters that can be configured to simulate real-world data." Another example is [PSU21]: "Note that this data captures the behavior of real workers in the target domain modulo potential differences induced by the use of a synthetic speech generator."

The notion is however also used in tandem with the notion of selective forces to deliberately synthesize data that diverges from the real-world distribution. [YIN+21]: "We took advantage of the synthesis pipeline to showcase how datasets can be constructed with properties that deliberately differ from real world distributions. Notably, we include samples of individuals with common (e.g., scientist) as well as uncommon occupations (e.g., spy) (Table 3) and designed SynthBio to be more balanced with respect to gender and nationality compared to the original WikiBio dataset."

8.4 Survey Discussion

We find that the various notions of representative samples are still highly pertinent. Over 95% of the surveyed papers use at least one notion and all notions appear

in a wide range of settings. Overall, we observe similar occurrence rates for the notions across the two conferences, with the largest differences apparent in the use of miniature, selective forces, and coverage notions. For both conferences the coverage notion is especially prominent appearing in 61% of the surveyed NeurIPS paper and 84% of the surveyed ICCV papers. This might partially be attributed to the backdrop of the cautionary tale on lack of coverage in ImageNet, bu also partly due to questions in datasheets for datasets using a clear notion of coverage (e.g. geographic coverage) when inquiring about the representativity of the dataset. We also bring attention to the the assertive notion, which is rarely used but still has a somewhat high occurrence rate considering the recognition of the two conferences.

8.5 Demonstrations using data

To demonstrate contrasting perspectives on representativity, we empirically evaluate performance, fairness and diversity for samples created with either with the concept of coverage or reflection in mind. The samples are created from a US census data collection [DHMS21] through stratified random sampling to obtain a miniature and through either density based or determinantal point process (DPP) based sampling to achieve coverage.

The US Census data exhibit significant population skew between minority and majority groups of protected attributes as well as significant interstate geographical variation. For this reason the data provides a suitable testing ground to study the effects of data representativity in relation to representation bias. Models trained on biased data can result in learned mappings from input to output that are uncertain for underrepresented regions [SG19b], which may lead to disparate predictive accuracy for different groups [CJS18, AJJ19, JXS$^+$20], but can also cause adverse effects on overall performance under distributional shifts between training and target data.

For instance if models trained on specific states are applied to other states [DHMS21]. Representation bias can be mitigated by identifying and populating underrepresented parts of the data distribution [SG19b, JXS$^+$20]. Such mitigation efforts could be performed by obtaining additional data, by targeted data augmentation (eg. SMOTE [CBHK02]) or by probabilistic over-sampling of underrepresented data regions [KGC21b]. Representation bias can occur in real-world ML applications, where a systemic bias in the geographical distribution of US cohorts used to train models for clinical applications has been uncovered [KAL20]. This investigation found that 71% of the analyzed studies used cohorts from at least 1 of 3 states, namely California, Massachusetts or

New York, while 34 states did not contribute to any cohorts. California cohorts appeared in 39% of all analyzed studies. With this in mind, we also investigate the role of data representativity for drawing inference under distributional shifts, by comparing performance on in-distribution and out-of-distribution data for the different sampling strategies.

8.5.1 Data

The UCI Adult dataset from the 1994 Current Population Survey is organized by the US Census Bureau [KB96] and is a popular dataset in the machine learning community. This data has been used in hundreds of research papers, but its external validity has been questioned, and a collection of new datasets from US Census Bureau data have been proposed [DHMS21]. More specifically, these datasets are extracted from the American Community Survey Public Use Microdata Sample (ACS PUMS). They contain data on attributes like age, income, education, sex, ancestry and employment. The responses to the survey are controlled by privacy rules seeking to prevent re-identification of responders. Detailed documentation on the records can be found on the US Census Bureau websites. One of the proposed datasets is a replacement for the original UCI Adult dataset containing an income prediction task for a feature subset of the 2018 ACS PUMS data spanning all US states in addition to Puerto Rico.

To generate the dataset the ACS PUMS data are filtered to only include individuals over the age of 16 with at least one working hour per week and an income of at least 100 USD in the past year. This leaves a total of 1,664,500 individuals. Like the original UCI Adult dataset, this new dataset has a predefined income threshold of 50,000 USD used to binarize the targets into a classification setting. Fairness intervention tasks have been shown to be sensitive to the specific threshold value [DHMS21]. For this reason we create a modified version of the income dataset and omit the income threshold to form a regression task with the continuous income as target. We transform the income target using the natural logarithm to obtain homoscedasticity for the residuals in our regression model. An overview of the dataset can be seen in Table 8.3.

8.5.2 Methodology

We compare linear regression models fitted to the log transformed income using all features in Table 8.3 for the state of California (n=195,665). We evaluate model performances using 5-fold cross validation where for each iteration 20% (n=39,133) of the California data are used for testing and the remaining 80%

Table 8.3: Overview of the features in the modified US Census income data (n=1,664,500). Features COW, SCHL, MAR, POBP and RELP are modified from the original ACS PUMS data by binarizing into respectively government / non-government worker (COW), Bachelor's degree / no Bachelor's degree (SCHL), married / not married (MAR), US-born / non-US-born (POBP) and reference person / non-reference person (RELP). See the ACS PUMS dictionary documentation for full feature descriptions including original category codes.

Feature Type	Feature Name	Description	Data Type	Categories	Min/Max
Input	AGEP	Age	Continuous	-	17 - 96
Input	COW	Class of worker	Binary	2	-
Input	SCHL	Educational attainment	Binary	2	-
Input	MAR	Marital status	Binary	2	-
Input	POBP	Place of birth	Binary	2	-
Input	RELP	Relationship	Binary	2	-
Input	WKHP	Hours worked per week	Continuous	-	1 - 99
Input	SEX	Sex	Binary	2	-
Input	RAC1P	Race	Categorical	9	-
Target	PINCP	Total income	Continuous	-	104 - 1,423,000

(n=156,532) are used for training. We compare a model trained using the full training data (which we denote full census model) to models trained on samples of the training data following either the reflection or coverage concepts of representativity. For each iteration a miniature and coverage sample is drawn from the training data. The sample sizes are 20% (n=31,306) of the full census training data. We evaluate the concepts of representativity by comparing performances on a range of metrics including overall performance using the mean squared errors (MSE) as well as performance in terms of fairness and diversity criteria. We also evaluate performance on in-distribution and out-of-distribution data by comparing interstate and intrastate performance. For completeness we show additional results from logistic regression classification models on the original binarized income (50,000 USD threshold) in Appendix 8.8.

8.5.2.1 Generating Samples

We generate miniature samples using a population based probability sampling scheme known as proportional stratified random sampling. Based on various demographic features the data are subdivided into smaller groups (strata) that

exhibit homogeneity. Subsequently random samples are drawn from these strata. To ensure such sampling constitutes a true miniature of the underlying population requires either a relatively homogeneous population or an increasingly large sample the more sociodemographic features are considered. This is particularly the case with sociodemographic data containing minority groups, where strata can become too finely grained and be represented by statistically insufficient sample sizes [BJP13]. To sample rare ethnic groups disproportionate sampling (for instance oversampling of minority groups) can be used [Kal03, KBAW07, CSS21], but this can lead to adverse affects on overall population estimates. To avoid too finely grained strata we generate miniature samples by cross stratifying on three important protected sociodemographic features, namely age, sex and race. The sex feature contains 2 categories, while the race feature contains 9 categories. We bin the age feature into three bins containing age groups of [0-33],[33-66],[66-99]. This combines to a total of 54 strata. In section 9.4 we empirically demonstrate that our stratified random sampling mimics the population and that results on the miniature samples generalize to the population.

We generate coverage samples using two approaches. Firstly, a density based coverage approach using density weighted sampling proposed in [KGC21b]. The density around observations is measured by the mean distance to the nearest neighbors and the density measures are then used as sampling probabilities in a weighted random sampling scheme. This approach causes observations in low-density regions to be sampled with high probability and conversely observations from high-density regions to be sampled with low probability. In doing so, the density sampling equally covers the input space regardless of the demographic proportions in the population.

Secondly, we generate a diverse coverage sample using a determinantal point process (DPP) probability distribution [KT$^+$12a]. DPPs have been used to create diverse sets in a number of ML applications ranging from documents, sensors, videos, images and recommendations systems. [LB12b, KSG08b, GCGS14b, KT$^+$12a, ZKL$^+$10b]. The DPP is a distribution over subsets S such that the probability of a subset is proportional to the determinant of a positive semidefinite kernel matrix known as the L-ensemble $P(S) \propto \mathrm{Det}(L)$. The L-ensemble may be constructed as the Gramian of the data. Since inference through DPPs rely on inversion and eigendecomposition of the L-ensemble, this procedure is inefficient with large N, where typically the dual representation is used for efficient inference over large sets [KT$^+$12a]. DPPs model not only the content of the subsets, but also the size. To draw samples of a specific size k-DPPs, a conditional DPP modeling only subsets of cardinatliy k, was proposed [KT11]. We generate our DPP samples with the DPPy library [GPBV19] using k-DPPs through the dual representation.

8.5.2.2 Out-of-distribution performance

To investigate the role of data representativity for drawing inference under distributional shifts, we compare performance on in-distribution (the California test data) and out-of-distribution data (the remaining 49 US states and Puerto Rico) for the different sampling strategies.

8.5.2.3 Fairness metrics

We measure group level fairness between the overrepresented group of White individuals (accounting for 62.2% of the California data) and the underrepresented group of Native American individuals (accounting for 0.9% of the California data). In the ACS PUMS data Native Americans include both American Indian and Alaska Native individuals. We measure fairness based on *demographic parity* and *equalized odds* defined in Equations 8.4 and 8.5. We quantify demographic parity for our regression models by measuring the departure of the CDF of model predictions to the CDF of model predictions conditional on the protected attribute. We denote this departure the regression demographic disparity (RDD) and measure it using the ℓ_∞ norm. We measure the equalized odds disparity using an approach based on resampling of protected attributes [RBC20]. Here a synthetic resampled version of A is constructed, called fair dummies $\tilde{A}$, such that the triple $(\hat{Y}, \tilde{A}, Y)$ obeys equalized odds. The distribution of the fair triple $(\hat{Y}, \tilde{A}, Y)$ is then compared to that of the observed test data $(\hat{Y}, A, Y)$. We again measure the distributional departure using the ℓ_∞ norm and denote this the regression equalized odds disparity (REOD). For our classification models we measure fairness in terms of demographic parity and equal opportunity by the difference in positive rates and the difference in true positive rates between White and Native American individuals. We denote these measures the classification demographic disparity (CDD) and classification equal opportunity disparity (CEOD).

8.5.2.4 Coverage Metrics

We compare samples on combinatorial diversity $C(\cdot)$ and geometric coverage $G(\cdot)$ defined in Eqs. 10.2 and 8.3. Typically geometric coverage is computed from the determinant of the L-ensemble (Gramian), but for the US Census data $p \ll n$, which leads to a determinant and volume of zero. This necessitates an alternative formulation. We instead compute the diversity from the dual representation of the L kernel, which carries information about several important properties of the L-ensemble [KT$^+$12b].

8.5.2.5 Reflection Metrics

We evaluate the reflection concept of representativity for the samples both through statistical tests on average predictions between sample and population, as well as a measure of distance between overall sample and population distributions. For the distributional measure we report the first Wasserstein distance between samples and population for two features.

8.5.3 Results

The MSE on the in-distibution California test data can be seen in Table 8.4. The model trained on the full census training data has the lowest MSE followed by the miniature and DPP model, while the density model has the highest MSE. Table 8.4 also illustrates how the models score on fairness criteria for demographic parity and equalized odds between White and Native American individuals. The density model has the best performance in terms of demographic parity and equalized odds while the miniature and full census model have the worst performances. P-values from paired t-tests on sample results can be found in Appendix 8.8 in Table 8.9. Equivalent results for the classification setting can be found in Appendix 8.8 in Tables 8.7 and 8.8.

Table 8.4: Regression performance metrics evaluated with 5-fold cross validation on the California data. For each iteration 80% (n=156,532) of the California data is used for training a full census model and the remaining 20% (n=39,133) is used for testing. For each iteration a miniature and coverage sample is drawn from the full census training data and tested on the test data. The sample sizes are 20% (n=31,306) of the full census training data. The overall MSE across the 5 folds is shown in addition to regression demographic disparity (RDD) and equalized odds disparity (REOD) for White / Native American individuals in units of 10^{-2}.

Training Data	MSE	Parity (RDD)	Equality (REOD)
Full Census	**0.79 ± 0.01**	0.23 ± 0.02	0.16 ± 0.02
Miniature Sample	**0.79 ± 0.01**	0.24 ± 0.01	0.17 ± 0.02
Density Sample	0.83 ± 0.01	**0.16 ± 0.01**	**0.10± 0.03**
DPP Sample	0.80 ± 0.01	0.21 ± 0.01	0.15 ± 0.01

We report sample scores in terms of their combinatorial (Eq. 10.2) and geometric (Eq. 8.3) diversity in Table 8.5.

Table 8.5: Mean combinatorial diversity $C(\cdot)$ on the race feature and geometric diversity $G(\cdot)$ in units of $\cdot 10^{27}$ on all features for the five samples of each sample type. Standard deviations (SD) are also shown.

Sample Type	$C(\cdot)$	$G(\cdot)$
Miniature	1.18 ± 0.00	0.01 ± 0.00
Density	1.82 ± 0.00	3.58 ± 0.20
DPP	$\mathbf{1.94 \pm 0.00}$	$\mathbf{26.10 \pm 0.38}$

Table 8.6 reports distributional distances to asses the reflection concept of representativity for the different samples.

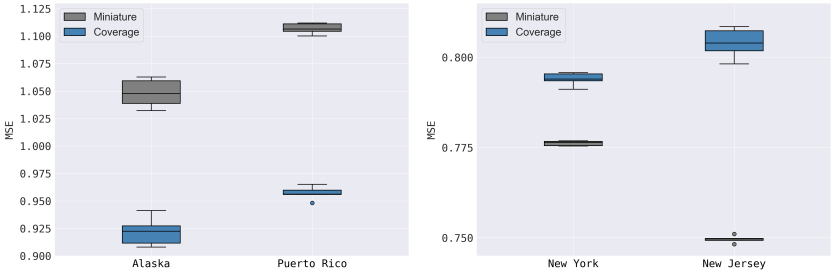
Table 8.6: Distributional comparisons to evaluate the reflection concept of representativity. Comparisons are reported as mean 1D Earth Mover Distances (EMD) between full census and the respective samples of each type. Standard deviations (SD) are also shown. EMD between population and sample is reported for the race and hours worked feature. The miniature samples where created by stratifying on the race feature, but not on the hours worked feature. Lower distance equates to higher evidence toward the reflection concept of representativity.

Sample Type	EMD Race	EMD Hours Worked
Miniature	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.07 \pm 0.01}$
Density	1.32 ± 0.02	3.71 ± 0.04
DPP	0.82 ± 0.01	1.15 ± 0.05

8.5.3.1 Out-of-distribution results

We demonstrate out-of-distribution performance by applying models trained on California to the remaining 49 states and Puerto Rico. Figure 8.1 compares MSE performance of miniature and density coverage models on two states similar and two states dissimilar to the California training data in terms of demographic distribution. See Fig. 8.2 in Appendix 8.8 for an out-of-distribution performance breakdown on the remaining states. MSE performance on in-distribution data is best for the model trained on miniature samples of the California training data, while MSE performance on out-of-distribution data is best for the model trained on coverage samples of the California training data. Overall the model trained on density coverage samples is on average better on 41 of the 50 states

and Puerto Rico with an average performance increase of 4% across all states. Similar results can be found for the classification case in Appendix 8.8, where the coverage model is better than the miniature model on 43 of the 50 states and Puerto Rico with an average accuracy increase of 1.5%. Fig. 8.3 in Appendix 8.9 shows results of models trained on a state with different demographic distribution than California. Here we use Massachusetts as training data and again find a model trained on miniature samples to achieve better predictive performance on in-distribution data, but worse performance on out-of-distribution data.



(a) MSE on states dissimilar to the training data (California).

(b) MSE on states similar to the training data (California).

Figure 8.1: MSE performance when applying models trained on the California data to states either demographically dissimilar (a) or similar (b) to the training data in terms of Pearson product-moment correlation between all features. The miniature model is trained on stratified random samples, while the coverage model is trained on density samples.

8.5.4 Summing up experiments on data

While the coverage sampling has merits such as robustness to distributional shifts and less disparate predictive performance between under- and overrepresented parts of the input space, the coverage sampling fails to accurately represent the distribution of the underlying population, and consequently incurs a loss in predictive power on the majority of said population, measured by the MSE. On the contrary, the miniature sampling accurately represents the underlying demographic distribution of the population allowing a similar interpretation of relations between sample and population. Consequently the miniature sampling is particularly appropriate for historical or in-distribution inference on the majority. This is evident for model performances on in-distribution data, where the miniature sampling achieves better predictive performance than the coverage

sample.

While we demonstrate improved race representation for our coverage sampling and consequently less disparate predictive accuracy between these groups, it should be noted that coverage procedures cannot be blindly applied to any dataset with the expectation of improved representation for marginalized groups. For instance [CDKV16b] shows that sampling for diverse image summaries with the notion of geometric coverage (DPP sampling) does not necessarily result in the desired improvement in gender representation in the generated summaries. This happens when instances of overrepresented and marginalized groups are not geometrically distinct (for instance with visually similar images of individuals of different race and gender). Likewise, the density based sampling approach relies on marginalized groups being positioned in low-density regions of the input space in order to achieve sufficient coverage of these. This underlines a point that achieving a representative sample under the concept of coverage should be seen in the context of the dataset and task at hand. Improved techniques for identifying and achieving optimal coverage of marginalized groups or regions in datasets provides an important future research direction.

8.6 Framework for data representativity

This section presents our proposed framework of questions for assessing data representativity when creating and documenting data. The framework naturally fits into both datasheets for datasets [GMV+21b] as well as shorter, more general data descriptions, and our aim here is to make it as concise and manageable as possible. With this in mind, and based on our literature study, proposed concepts, and empirical investigations, we propose answering and adhering to the following questions and guidelines:

8.6.1 Purpose:

What is the purpose of collecting/creating the data, and what/who is the target population? In addition, when building AI systems; what is the intended aim of the AI system along side its intended use?

8.6.2 Sampling methodology:

Which data representativity concept have you used to create your sample (reflection/coverage/representatives)? What is the sampling method and procedure used to create the data? The methodology should be specified to a degree that makes it reproducible. If a code base is used to create the data, we recommend making it open source.

8.6.3 Evaluation:

Are the collected data representative of the target population or 'good enough' for the aim? We recommend making this evaluation in accordance with the purpose and measurable data representativity concept, and not as a general statement of representativity. In addition, known limitations of the representativity, in terms of coverage as well as distributional match to target population, are always desirable to document for datasets to assess possible limitations, and not least because open source datasets may be used for purposes not originally anticipated. Finally, add measures of representativity in accordance with the sampling concept and to the extend possible.

8.7 Discussion

We found that the notions of what constitutes a 'representative sample' from the 1979 reviews by Mosteller and Kruskal are still pertinent. When building machine learning models and AI systems, particularly two contrasting views of representativity are of relevance: The concept of coverage vs. that of a reflection. We find that the two are useful for different purposes. Coverage is useful for robustness towards distribution shifts as well as mitigation of disparate predictive accuracy between overrepresented and marginalized groups.

The reflection concept is useful to mimic the target population allowing a similar interpretation of relations between sample and population as well as to obtain minimum average errors on the target population. However, we should keep in mind that average errors indicate that predictions are best for the majority, and not necessarily equal for population subgroups.

The notion of a 'representative sample' as an assertive acclaim without specification was mainly used in AI related literature as an implicit acclaim, without explicit mentioning of representativity, but with an indication of an inference

link (generalization from data) matching that of representativity. We call for attention on such implicit use, and recommend avoiding it, thus always specifying the sampling methodology as well as the purpose and target population of the collected data along with an evaluation of representativity and limits of same for the given sample. Such specification and evaluation is critical on the path towards fully transparent and trustworthy AI systems.

Through our investigations we found that we cannot talk about general representativeness of a sample, but need to consider data collection and representativeness in coherence with our purpose (and data analysis) whether this is a research hypothesis or an aim for our AI system.

As we reach limitations from our understanding of the target distributions and/or from a large number of attributes (and their interactions), it is practically impossible to make guarantees of representativeness. As a consequence, evaluations based on several datasets as well as 'in use' data (for deployed ML models or AI systems) are encouraged. Furthermore, accounting for all possible distribution shifts that may happen in the future (where our AI system will be in production), is also practically impossible. As an alternative, or rather addition, we suggest to perform continuous monitoring of AI systems and their performance while they are in production. An AI system may also at first be deployed in shadow mode if risks are too high to use predictions without further (live) testing.

Finally, we propose that further research into measurable concepts of data representativity is necessary. There is a need for measures that are computationally feasible for large high dimensional data and which can model joint distributions (parametric and non-parametric) as well as a need for further analysis into existing measures and their limitations.

Acknowledgements

The authors would like to acknowledge colleague Murat Kulahci for insightful feedback on the manuscript. Additionally, author Rune D. Kjærsgaard is funded by a university alliance scholarship between UiB (University of Bergen) and DTU (Technical University of Denmark).

8.8 Appendix A: Results for California

We show 5 fold cross validation (CV) results from applying models trained on California data to all other states. We show regression results from linear regression models using the continuous income target and classification results for logistic regression models using the binarized income (50,000 USD threshold). We compare miniature samples generated with stratified random sampling to coverage samples produced through density sampling.

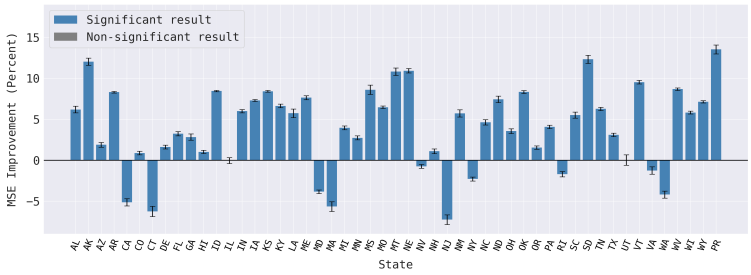
Table 8.7 shows in-distribution results for classification models trained using the binarized income with 50,000 USD threshold. Table 8.8 shows the associated p-values while Table 8.9 shows the p-values for the regression results in Table 8.4.

Table 8.7: Classification performance metrics evaluated with 5-fold cross validation of logistic regression models trained on the California data using the original binarized income (50,000 USD threshold) as target. The results are generated with the same procedure as in table 8.4. Accuracy is shown in addition to classification demographic disparity (CDD) and equal opportunity disparity (CEOD) for White / Native American individuals.

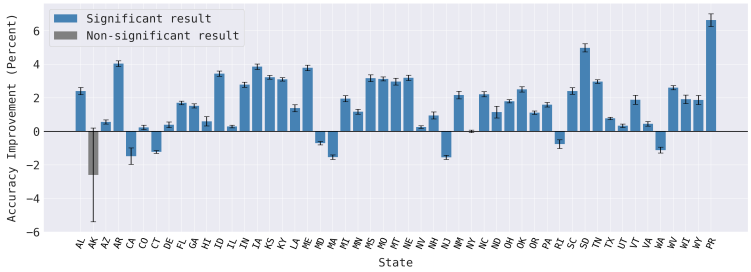
Training Data	Accuracy	Parity (CDD)	Equality (CEOD)
Full Census	**0.77 ± 0.00**	0.25 ± 0.02	0.30 ± 0.07
Miniature Sample	0.77 ± 0.00	0.27 ± 0.03	0.32 ± 0.08
Density Sample	0.76 ± 0.00	**0.21 ± 0.02**	**0.25 ± 0.05**
DPP Sample	0.76 ± 0.00	0.25 ± 0.02	0.30 ± 0.06

Table 8.8: Paired sample t-test on classification model results shown in Table 8.7. P-values are adjusted for multiple testing by controlling the FDR using the Benjamini-Hochberg procedure [BH95].

Sample Comparison	Accuracy p-value	CDD p-value	CEOD p-value
Full Census vs. Miniature	0.7200	0.3198	0.3715
Density vs. Miniature	0.0007	0.0096	0.0249
DPP vs. Miniature	0.0023	0.2622	0.3605
DPP vs. Density	0.0031	0.0249	0.0755



(a) California regression.



(b) California classification.

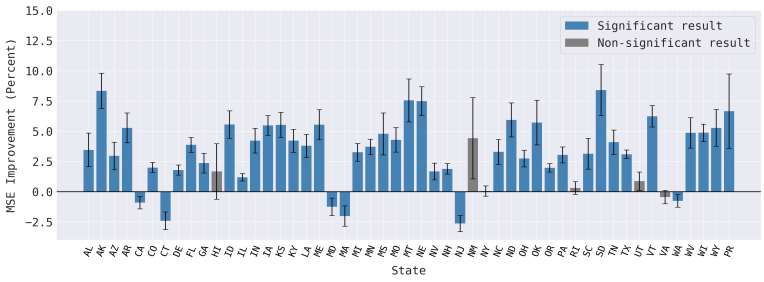
Figure 8.2: (a) Regression performance improvement on MSE when using coverage samples created with density sampling compared to using miniature samples created with proportional stratified sampling of the California training data. Error bars indicate the standard deviation. Positive values express that the MSE of the model trained on the coverage samples is improved over the MSE of the model trained on miniature samples. Negative values indicate that the coverage sampling deteriorates the performance compared to the miniature sampling. The mean MSE improvement across all states is 3.95%. False Discovery Rate (FDR) adjusted paired t-tests have been performed on all interstate differences. Grey results (Illinois and Utah) indicate states where model differences are not significant to an $\alpha = 0.05$ level. (b) Same as above for classification accuracy improvement on the binarized target value. The mean accuracy improvement across all states is 1.54%. All performance differences except for Alaska and New York are significant to an $\alpha = 0.05$ level when using a FDR adjusted paired sample t-test.

Table 8.9: Paired sample t-test on regression model results shown in Table 8.4. P-values are adjusted for multiple testing by controlling the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure [BH95].

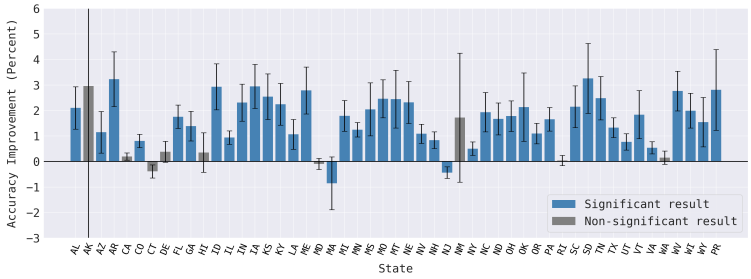
Sample Comparison	MSE p-value	RDD p-value	REOD p-value
Full Census vs. Miniature	0.0624	0.2962	0.1970
Density vs. Miniature	0.0002	0.0010	0.0102
DPP vs. Miniature	0.0192	0.0617	0.0102
DPP vs. Density	0.0010	0.0044	0.0332

8.9 Appendix B: Results for Massachusetts

We show 5 fold CV results from models trained on the Massachusetts data (n=40,114) applied to all other states. The Massachusetts data is significantly different from the California data both in terms of overall data size (n=40,114 vs. n=195,665), proportion of individuals with Bachelor's degree (49.1% vs. 38.7%) and in terms of proportion of White individuals (82.3% vs. 61.8%), but is similar on several other parameters like average weekly hours worked (37.4 vs. 37.9), proportion of government workers (12.5% vs. 14.9%), proportion of individuals aged 10-33 (33.3% vs. 32.4%) and proportion of married individuals (51.6% vs. 52.4%). Results compare performances for models trained on miniature samples of the training data to models trained on coverage samples created with the density sampling approach. Both sample types have size 20% of the training data. We find similar results as with the California data, where for both regression and classification using coverage samples over miniature samples deteriorates performance in terms of MSE and accuracy on in-distribution states similar to the Massachusetts training state (most notably on states CA, CT, MD, MA and NJ). However, model performance on states dissimilar to Massachusetts (out-of-distribution) is improved with an average performance increase across all states (including the training state) of 3.27% for regression and 1.54% for classification.



(a) MA regression.



(b) MA classification.

Figure 8.3: (a) Regression performance improvement on MSE when using coverage samples compared to miniature samples of the Massachusetts training data. Mean MSE performance improvement across all states is 3.27%. b) Same as above for classification. The mean accuracy improvement across all states is 1.54%. FDR adjusted paired t-tests have been performed on all interstate differences. Grey results indicate states where model differences are not significant to an $\alpha = 0.05$ level.

Chapter 9

Sampling To Improve Predictions For Underrepresented Observations In Imbalanced Data

Rune D. Kjærsgaard[1], Manja G. Grønberg[1], Line K. H. Clemmensen[1]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads 324, Kgs. Lyngby 2800, Denmark*

Publication Status: Paper is published.

Proceedings of Workshop on Data-Centric AI, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).
`https://datacentricai.org/neurips21/`

Abstract: Data imbalance is common in production data, where controlled production settings require data to fall within a narrow range of variation and data are collected with quality assessment in mind, rather than data analytic insights. This imbalance negatively impacts the predictive performance of models on underrepresented observations. We propose sampling to adjust for this imbalance with the goal of improving the performance of models trained on historical production data. We investigate the use of three sampling approaches to adjust for imbalance. The goal is to downsample the covariates in the training data and subsequently fit a regression model. We investigate how the predictive power of the model changes when using either the sampled or the original data for training. We apply our methods on a large biopharmaceutical manufacturing data set from an advanced simulation of penicillin production and find that fitting a model using the sampled data gives a small reduction in the overall predictive performance, but yields a systematically better performance on underrepresented observations. In addition, the results emphasize the need for alternative, fair, and balanced model evaluations.

9.1 Introduction

Production data is often gathered under very controlled settings, driven by a requirement of the data to fall within a specified range of variation, and experiments are often expensive leaving data insights to be derived from the available historical data. For this reason, production data commonly exhibit low variation expressed by most of the data lying in high-density areas with only few data points falling outside these areas. This is called imbalanced data and has been studied extensively for categorical targets ([HYS+17], [Kra16]), but only sparsely for continuous targets ([BTR17], [BTR19]). Previous works consider the imbalance to be caused by the target, where we on the other hand consider the imbalance mainly driven by the input variables. The premature ideas for this research were developed in [GSMC21].

Imbalance in the response variables is often handled through data-level approaches like over- or undersampling the classes, or through algorithm-level approaches like e.g. class priors, or by use of a hybrid of these ([Kra16], [JK19]). Here, we extend the data-level line of thought to consider sampling with respect to the input space. The assumption is that a balanced representation of the input space gives better inference for underrepresented parts of the input space. Thus, we propose (down)-sampling as a way to adjust for imbalance and demonstrate its use for production data, where we expect an imbalance due to the controlled settings.

In the following, we first discuss our three proposed sampling strategies to select a balanced training data set. Subsequently, we present our experimental setup and methods and the production data used for our experiments. Finally, we describe our results and discuss our findings and their perspectives.

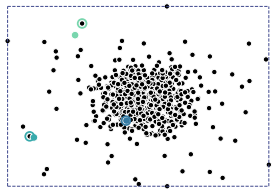
9.2 Sampling approaches

The main idea of this research is to obtain a more balanced data set than the original one by sampling a more balanced training data set. We will refer to the resulting data set as the new data set. We investigate three different sampling methods, where two of them, methods (a) and (b), are based on random sampling, and the last one, method (c), is density based.

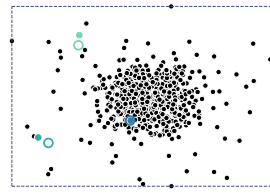
Our random sampling methods (a) and (b) combine a unique sampling approach (i) with random sampling of the observations from the training set (ii). The idea is that approach (i) mainly samples points on the edge of the data manifold (typically low-density areas) whereas approach (ii) mainly samples points in high-density areas of the manifold. Thus combining (i) and (ii), such that the new data set consists of a 50/50 combination of samples from (i) and (ii), the new data set has an almost equal amount of data from high- and low-density areas and is thus more balanced than the original.

Approach (i) samples points, z, uniformly within the hyper-rectangle spanned by the data. The sides of the hyper-rectangle are determined by the minimum and maximum value of each input variable, x_i, such that it has dimensions $\mathcal{Z} = [\min(x_1), \max(x_1)] \times ... \times [\min(x_p), \max(x_p)]$, where p is the number of variables. We then either use strategy (a), the nearest neighbour to the points z, denoted $1point$, or strategy (b), the mean of the 5 nearest neighbours to the points z, denoted $mean$, as samples in the new data set. The targets for the samples of $mean$ will be the mean of the targets for the 5 nearest neighbours. For some types of data, the mean is not necessarily meaningful, and a median approach would be a feasible alternative. Illustrations of the methods are found in Figure 9.1a and 9.1b. The filled coloured circles are the sampled points z, while the coloured rings are the (a) nearest neighbour to the sampled points or (b) the mean of the 5 nearest neighbours to the sampled points. The dotted lines represent the hyper-rectangle, within which we sample. Since the majority of data in imbalanced data sets are concentrated on a small part of the data manifold, the nearest neighbours to most points in the hyper-rectangle will lie on the edge of the data manifold. Thus, sampling methods (a-i) and (b-i) result in a lot of samples on the edge of the data manifold.

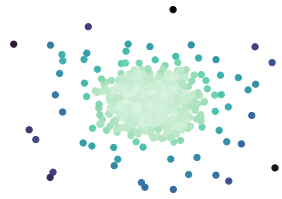
Approach (ii) samples points randomly with equal weight from the original data set. Due to the imbalance of the data, most of the points sampled by this approach lie in high-density areas. We sample points from both (i) and (ii) corresponding to 10% of the original (training) data. Thus, the size of the new data set for strategy (a) and (b) corresponds to 20% of the size of the original (training) data set.



(a) Random sampling a-i (1*point*).



(b) Random sampling b-i (*mean*).



(c) Density based sampling.

Figure 9.1: Illustration of the sampling methods. (a) and (b) illustrate approach (i) of random sampling. The coloured filled circles are the sampled points z, while the coloured rings are the (a) nearest neighbour to the sampled points or (b) the mean of the 5 nearest neighbours to the sampled points. The dotted lines represent the hyper-rectangle. (c) illustrates the density based sampling method. The colours reflect the sampling weights; the scaled mean distance to the 100 nearest neighbours.

The idea of the density based sampling method (c) is to obtain a more balanced data set by drawing a weighted random sample of the original data set with weights that reflect the inverse data density around each point. If a point is in a low-density area, the probability of drawing this point should be large, whereas if a point is in a high-density area the probability of drawing this point should be correspondingly low. We measure the data density around a point, x, as the mean distance to the 100 nearest neighbours of x. The sampling probabilities are then the mean distances scaled to sum to 1. The size of the new data set is 10% of the original data and the sample is drawn with replacement. Figure

9.1c illustrates how the density based sampling works. The colours reflect the sampling weights and thereby the measured data density around each point.

9.3 Method and data

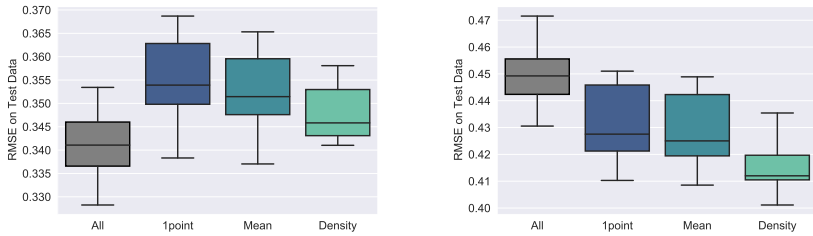
We investigate the three sampling approaches by applying them on a large biopharmaceutical data set from an advanced simulation of penicillin production in a 100,000 litre penicillin fermentation system known as industrial penicillin simulation (IndPenSim) ([GŞL+15], [GDVJ+19]). The data consist of 100 batches, where the first 90 are controlled with three different production control methods, and the last 10 batches contain faults resulting in process deviations. The latter batches are often few in historical data, but also those that give insights to the dynamics of the process away from the controlled settings.

The data set contains 113,935 observations of 2,238 variables. Of these variables, 39 are process variables of which one is the penicillin concentration. The remaining 2,199 are Raman spectroscopy measurements. We disregard the Raman spectra, 5 process variables containing missing values and two with no variation and analyse the rest (31 input variables) with the goal of predicting the penicillin concentration in the tank at each observation. We hold out 20% of the data for testing and consider the remaining 80% for training. We compare a linear regression model trained using all of the training data to models trained using only a sample of the training data.

9.4 Results

The root mean squared errors (RMSE) of the penicillin concentrations are shown in Figure 9.2. Of the sampling approaches, the density based approach gives the lowest RMSE on average. Figure 9.2a shows the RMSE on the full test set. Since the test data is imbalanced, none of the sampling approaches improve the RMSE over using all of the training data. However, the reduced performance has a low effect size; approximately a 3% decrease. Figure 9.2b shows the RMSE on the 10% most underrepresented observations from the test set measured by the mean distance to the 100 nearest neighbours. Here all sampling approaches improve the RMSE over using all training data.

Figure 9.3 shows the test set observations projected onto the first two principal components, which respectively explain 29.3% and 11.5% of the variance. This

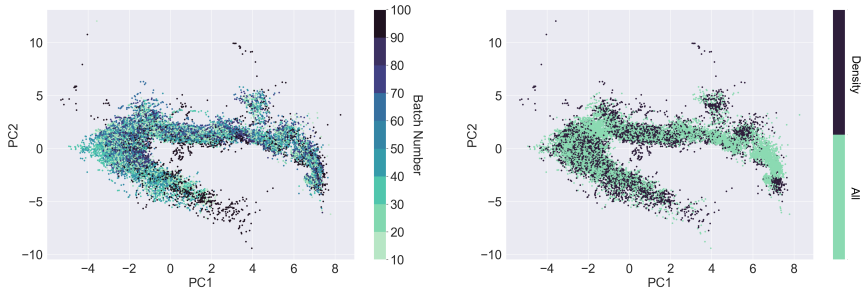


(a) Full imbalanced test set. **(b)** 10% most underrepresented.

Figure 9.2: Boxplots of the RMSE on the test data after 10 iterations of fitting the linear model using either the entire training set or samples from the three sampling approaches. (a) shows the performance on the full imbalanced test set, while (b) shows the performance on the 10% most underrepresented observations measured by the mean distance to the 100 nearest neighbours.

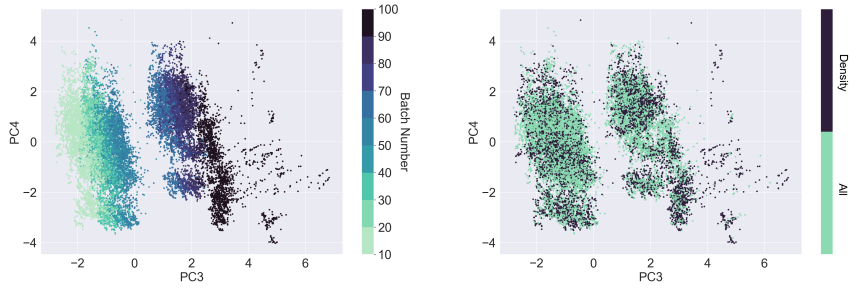
projection illustrates how the majority of the observations lie centralised on the data manifold in high-density areas, with only few observations lying on the edges of the manifold. Figure 9.3a displays the test set observations coloured according to batch number, which shows how the majority of observations in low-density regions originate from batches with process deviations (batches 91-100). Figure 9.3b illustrates the performance difference on the test set observations projected onto the first two principal components. Black data points indicate observations where the absolute residual from the density sampling approach is smaller than the absolute residual from using all data. The sampling has improved the performance for the majority of observations on the edge of the manifold (low-density, underrepresented areas). This is particularly the case in the upper part of the figure, where residuals for observations from the lowest density regions are all improved when using the density sample to train the model over using all data.

Figure 9.4 shows similar results with the test set observations projected onto the third and fourth principal components, which explain 9.1% and 6.9% of the variance. Figure 9.4a shows how the third principal component captures the variation across batches, with batches 91-100 again occupying the lowest density regions. Figure 9.4b illustrates the performance difference on the test set observations from using either all training data or the sample from the density approach. Again, the sampling has improved the performance for the underrepresented observations lying in low-density regions.



(a) Batch number **(b)** Residuals

Figure 9.3: The test data on the first two principal components. (a) shows the data coloured according to batch number. (b) shows the data coloured according to which approach between the density sampling method and using all data gives the lowest absolute residual.



(a) Batch number **(b)** Residuals

Figure 9.4: The test data on principal components three and four. (a) shows the data coloured according to batch number, while (b) is coloured according to the approach with the lowest absolute residual.

9.5 Discussion

The three strategies for sampling training data to adjust for imbalance all deteriorate the overall predictive performance compared to fitting a model on all the training samples, but only with a small effect size. However, residuals for underrepresented data have improved, illustrating that sampling can drive value for underrepresented data points/areas. In this context, we would like to raise the question of how to make a *balanced* and *fair* evaluation, as the RMSE on

imbalanced test data favours overrepresented inputs.

While we have shown our methods apply on production data, we expect them to also apply to other types of data, where balanced representative training data could be of particular importance. This could have a potential broader societal impact on domains with historical data containing underrepresented minorities.

Chapter 10

Fair Soft Clustering

Rune D. Kjærsgaard[1], Pekka Parviainen[2], Saket Saurabh[2,3], Madhumita Kundu[2], Line K. H. Clemmensen[1]

[1] *Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Richard Petersens Plads 324, Kgs. Lyngby 2800, Denmark*

[2] *Department of Informatics, University of Bergen (UiB), Norway*

[3] *Theoretical Computer Science Group, The Institute of Mathematical Sciences (IMSc), Chennai, India*

Publication Status: Paper is submitted for review.

Submitted to: Journal of Machine Learning Research (JMLR).

Abstract: Scholars in the machine learning community have recently focused on analyzing the fairness of learning models, including clustering algorithms. In this work we study fair clustering in a probabilistic (soft) setting, where observations may belong to several clusters determined by probabilities. We introduce new probabilistic fairness metrics, which generalize and extend existing non-probabilistic fairness frameworks and propose an algorithm for obtaining a fair probabilistic cluster solution from a data representation known as a fairlet decomposition. Finally, we demonstrate our proposed fairness metrics and algorithm by constructing a fair Gaussian mixture model on three real-world datasets. We achieve this by identifying balanced micro-clusters which minimize the distances induced by the model, and on which traditional clustering can be performed while ensuring the fairness of the solution.

10.1 Introduction

Decision making systems based on machine learning (ML) applications have demonstrated unwanted consequences as a result of biased data [PJN+11b, ZOM19]. This has fostered efforts towards artificial intelligence (AI) alignment, wherein ML systems are aligned with their intended objectives. This includes ensuring decisions are fair and do not show bias against or for certain population sub-groups. Many of these fairness interventions are based on the Disparate Impact (DI) doctrine [Rut87], which prohibits discrimination between different groups of protected attributes such as race or sex. For clustering, this type of non-discrimination is denoted group-level fairness [CMM21].

Clustering algorithms are an unsupervised ML approach used to partition a dataspace into clusters. These algorithms are widely used, particularly in settings where data labels are scarce. Here, clustering may be used as a feature engineering tool to supplement points with cluster assignments in an effort to increase expressive power of downstream models. If the underlying training data is unfair, this may propagate into the generated features and ultimately cause biased predictions. Fair clustering aims to prevent this.

The topic of fairness for clustering was initiated in a seminal work by [CKLV17], which considered group-level fairness obtained by modifying the input data for traditional hard clustering algorithms like k-center and k-median. The literature on fair clustering is largely focused on such non-probabilistic algorithms, where point assignments are deterministic [CMM21]. However, for a number of applications soft clustering is more appropriate. In our work, we consider group-level fair clustering in a probabilistic setting, where equal representation is ensured for protected groups in clusters found using soft clustering algorithms. As

an example, a bank might use a dataset containing information about educational attainment and wages of individuals to train a model with the goal of identifying potential customers and offering them loans or credit opportunities. The bank then trains a soft clustering algorithm to group customers into low or high risk candidates, with the soft assignments implying the probability (risk) that a given customer will default their loan. It should be pointed out, that a wage gap has been identified for women and people-of-color, who usually earn lower wages than White males [Pat16a], and that people-of-color often face additional adversities that lead to educational disparities as compared to White individuals [Sab16]. Thus, a clustering algorithm trained on this data would be prone to group White males as better prospective candidates and correspondingly deny people-of-color and women the potential for improvement, thus propagating the systemic bias from the training data to downstream decisions. Ensuring group-level fairness from a probabilistic cluster solution could prevent such decision-making systems from adversely affecting specific groups and thus ensure that the models adhere to the DI doctrine.

Fair probabilistic clustering has previously been studied by [EBTD20], where they considered probabilistic fairness in a setting of imperfect group membership knowledge. This considers the protected group membership in a probabilistic setting while considering the cluster assignments in a deterministic setting. We on the other hand, consider the case of deterministic protected group membership and probabilistic cluster assignments. In [ABDL20] they consider individual-level fairness in a probabilistic setting, where no protected groups exist, and fairness is achieved by ensuring similar individuals are treated similarly by the algorithm. Corresponding probabilistic assignments have been studied by [BCD$^+$20] and [BCD$^+$21]. In our work, we consider group-level fairness under the same conditions as [CKLV17], where protected groups exist and the goal is to construct a fairlet decomposition of this data, on which a traditional algorithm can be trained to obtain a balanced cluster solution. We generalize this to the probabilistic setting. No metrics for group-level fairness under probabilistic cluster assignments have been established [CMM21]. To this end, we propose probabilistic fairness metrics, which generalize current definitions for deterministic cluster assignment. Moreover, we demonstrate an algorithm for obtaining a fair cluster solution from a fairlet decomposition in the probabilistic setting. Finally, we demonstrate our metrics and algorithm by applying them on a fairlet decomposition constructed for a Gaussian mixture model [MB88].

Our contributions are:

- Probabilistic generalizations of metrics for group-level fairness.
- An algorithm for obtaining a fair probabilistic cluster solution from a fairlet decomposition.

- An approach for generating a fairlet decomposition for a GMM.

10.2 Cluster Fairness

In most work group-level fair clustering is defined in terms of balance or relaxations thereof, but may also be defined in terms of entropy [CMM21]. These metrics measure the representation equality of protected groups described by an attribute vector $\boldsymbol{A}$. In this work, we denote protected groups by colors $p \in P$.

10.2.1 Deterministic Assignment Fairness

Balance measures algorithmic fairness of a cluster solution by considering the degree of balance between protected groups within each cluster. This fairness definition complies with the the DI doctrine, and the goal is to obtain a balanced representation (similar fraction) of all groups within all clusters.

Consider a set of points D partitioned into a set of clusters C. Balance may be measured by comparing two fractions $r_{D,p}$ and $r_{c,p}$, where $r_{D,p}$ is the fraction of a color p in D and $r_{c,p} = \frac{|N_{c,p}|}{n_c}$ is the fraction of a color p in cluster c, where n_c is the number of observations in cluster c, and $N_{c,p}$ is the set of observations in the dataset belonging to both color p and cluster c. Now construct a fraction $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ and define the balance by:

$$B = \min_{c \in C, p \in P} \min \left(R_{c,p}, \frac{1}{R_{c,p}} \right), \tag{10.1}$$

where B is the balance and $R_{c,p} = \frac{r_{D,p}}{r_{c,p}}$ is a fraction for a given cluster c and color p [CMM21]. Balance is bounded in $B \in [0, 1]$ with higher balance being more fair. This metric measures the overall fairness of the cluster solution through the minimum balance across all clusters $c \in C$ and colors $p \in P$. Optimal balance ($B = 1$) is found when all clusters share the same color fraction $r_{c,p} = r_{D,p}\ \forall c, p$, while worst case balance ($B = 0$) is found when a cluster contains no members of a protected group $r_{c,p} = 0$.

Contrary to the balance metric, entropy does not measure the worst case fairness of all clusters, but rather quantifies the overall fairness through an information-theoretic perspective across all clusters simultaneously:

$$H = \min_{p \in P} \left(-\sum_{c=1}^{C} r_{c,p} \log r_{c,p} \right), \tag{10.2}$$

where H is the entropy [CMM21]. The entropy fairness is the level of information entropy across all clusters. Higher entropy equates to a more fair cluster solution. Optimal entropy fairness is found when all clusters share the same color fraction $r_{c,p}$ while worst case entropy fairness is found when all clusters are monochromatic.

10.2.2 Probabilistic Assignment Fairness

In soft clustering algorithms the point assignments are probabilistic and determined by a responsibility vector $\boldsymbol{\gamma}_c$ for each cluster c. The entries $\gamma_{i,c}$ in this vector describe the probability that the i^{th} data point is generated by component c. We use the responsibilities to construct a measure for weighted color contribution:

$$w_{c,p} = \frac{\sum_{i=1}^{N} \gamma_{i,c}\alpha_{i,p}}{\sum_{i=1}^{N} \gamma_{i,c}}, \tag{10.3}$$

where $w_{c,p}$ is the weighted contribution of color p to cluster c, $\gamma_{i,c}$ is the i^{th} entry in the responsibility vector $\boldsymbol{\gamma_c}$ and $\alpha_{i,p}$ is the i^{th} entry in a color vector $\boldsymbol{\alpha}_p$ constructed by setting $\alpha_{i,p} = 1$ if observation $\boldsymbol{x}_i \in p$ and 0 otherwise. The numerator represents the total color mass (weighted color contribution) for the given cluster, while the denominator represents the total mass of the cluster. Note that the weighted color contribution $w_{c,p}$ reduces to $r_{c,p}$ if $\boldsymbol{\gamma_c}$ dictates hard assignments (probabilities either 1 or 0). Thus $w_{c,p}$ generalizes the unweighted color contribution $r_{c,p}$ from the deterministic assignment setting.

We propose to substitute the weighted color contribution $w_{c,p}$ into the established fairness frameworks for deterministic assignment fairness in Eqs. 10.1 and 10.2:

$$B_{\text{soft}} = \min_{c \in C, p \in P} \min \left(W_{c,p}, \frac{1}{W_{c,p}} \right), \tag{10.4}$$

where B_{soft} is the soft assignment balance and $W_{c,p} = \frac{r_{D,p}}{w_{c,p}}$.

Equivalently we define the soft assignment entropy fairness by:

$$H_{\text{soft}} = \min_{p \in P} \left(-\sum_{c=1}^{C} w_{c,p} \log w_{c,p} \right), \tag{10.5}$$

10.2.3 Entropy Ratio

Unlike balance, entropy is not bounded in $H \in [0, 1]$, but we can normalize it by comparing the information entropy of the cluster solution to the optimal entropy of the cluster configuration:

$$H_{\text{ratio}} = H_{\text{soft}} / H_{\text{OPT}}, \tag{10.6}$$

where H_{OPT} is a cluster solution with optimal (largest) entropy under the given number of clusters. H_{OPT} is found when all clusters share the same color fraction. Thus $H_{\text{ratio}} \in [0, 1]$, where $H_{\text{ratio}} = 1$ when $w_{c,p} = r_{D,p}\ \forall c, p$ and $H_{\text{ratio}} = 0$ when all clusters are monochromatic with respect to color p.

10.3 Obtaining Fair Clusters

Standard cluster algorithms optimize an objective function and ignore the distribution of protected attributes. This may end up propagating inherent bias from the training data to the final model solution. To avoid this, the data can be modified by constructing a balanced representation. Fair cluster solutions can be found by generating a fair representation through a fairlet decomposition and subsequently performing clustering with a traditional color blind algorithm on the decomposition. The decomposition is constructed by identifying micro-clusters, called fairlets, which preserve balance.

For a binary protected attribute consisting of two colors, a decomposition can be specified as a (p_1, p_2)-fairlet decomposition (assuming $p_1 < p_2$) with balance parameters p_1 and p_2 indicating that all fairlets have a color fraction $r_{c,p} \geq \frac{p_1}{p_1+p_2}$. For a perfectly balanced dataset ($N_{p_1} = N_{p_2}$) it is possible to obtain a $(1, 1)$-fairlet decomposition, where each fairlet consists of exactly one point of each color. For this setting $r_{c,p} = r_{D,p}\, \forall c, p$ in the decomposition, which results in a balance of $B = 1$. To construct a fair clustering from the decomposition, centers are assigned for each fairlet and a traditional clustering is performed on the centers. Since the union of balanced micro-clusters is necessarily also balanced, this will ensure a fair clustering.

This procedure has been constructed for deterministic assignments in the literature. We demonstrate an algorithm for constructing a fair probabilistic clustering from any fairlet decomposition by modifying existing framework for deterministic clustering [BIO+19]. This is shown in Algorithm 1. The fair cluster solution is found by generating a fairlet decomposition, applying a traditional soft clustering

Algorithm 1 SOFTCLUSTERFAIRLET(Q)

Input: $Q = \{q_1, q_2, \cdots, q_\ell\}$ where every q_i is a fairlet with center c_i
Output: The algorithm returns a fair probabilistic clustering of D given a fairlet decomposition Q of D

multiset $\bar{D} \leftarrow \emptyset$ (initilization)
for all fairlets $q_i \in Q$ **do**
  $\bar{D} \leftarrow \bar{D} + \{|q_i|$ copies of $\mathrm{c}_i\}$ (sum of two multisets)
end for
$C \leftarrow$ Traditional probabilistic clustering of $\bar{D}$
$C^* \leftarrow \gamma_i$ (assign fairlet members the responsibility vector of their center in C)
return C^*

algorithm on the fairlet centers and subsequently assigning appropriate responsibilities to the fairlet members. Algorithm 1 provides the same theoretical fairness guarantees as previous works in the hard assignment setting (see Appendix 10.7 for details).

However, the algorithm does not ensure optimality of the decomposition and may result in a sub-optimal cost of the studied clustering objective depending on the spatial location of the points selected for each fairlet. To obtain a solution which maintains a fair representation of protected groups and simultaneously minimizes a clustering objective function, it is necessary to take the cost of the decomposition into account. Fairlet decompositions are tailored to specific objective functions like the k-median and k-means objective [BGK+18]:

$$\mathcal{L}_{\mathrm{k}}(D, Q) = \sum_{x \in D} d(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x})), \tag{10.7}$$

where $d(\cdot)$ is a metric (distance function) and $\beta_{\mathrm{Q}}(\boldsymbol{x})$ denotes the center location of the fairlet to which the data point $\boldsymbol{x}$ is mapped. For k-median clustering $\beta_{\mathrm{Q}}(\boldsymbol{x}) \in D$ and for k-means clustering $\beta_{\mathrm{Q}}(\boldsymbol{x}) \in \mathbb{R}^m$ for $D \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. The distance metric in k-means is $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^2$.

[CKLV17] define the total cost of the overall fair clustering assignment from D to C^* (Lemma 6) as:

$$\mathcal{L}_{\mathrm{k-tot}}(D, C^*) = \mathcal{L}_{\mathrm{k}}(D, Q) + \mathcal{L}_{\mathrm{k}}(\bar{D}, C^*), \tag{10.8}$$

where $\mathcal{L}_{\mathrm{k}}(D, Q)$ is the fairlet decomposition cost, and $\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*)$ is the cost on a transformed dataset $\bar{D}$, where for each fairlet q_i the fairlet center c_i appears $|q_i|$ times.

The cost on the transformed dataset $\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*)$ is the sum of distances of each

point in $\bar{D}$ to their assigned cluster center:

$$\mathcal{L}_{\mathrm{k}}(\bar{D}, C^*) = \sum_{\boldsymbol{x} \in \bar{D}} d(\boldsymbol{x}, \alpha_{C^*}(\boldsymbol{x})), \tag{10.9}$$

where $\alpha_{C^*}(\boldsymbol{x})$ is location of the center for which the data point $\boldsymbol{x}$ is mapped by the clustering C^*.

The goal of the fair clustering is to construct a fairlet decomposition which minimizes the cost in Eq. 10.8. [CKLV17] propose solving the problem by transforming it into a minimum cost flow (MCF) problem, where a directed graph is constructed. This graph may be modified to suit different cluster objective functions. To generate a decomposition the weights on the edges between nodes are represented by a distance function between points. The objective is to minimize the sum of distances from fairlet members to fairlet centers. The MCF approach has super-quadratic time in dataset size and becomes computationally expensive for large datasets. Alternative scalable approaches have been introduced, where the optimal fairlet decomposition is approximated and found in nearly linear time in dataset size [BIO+19]. In our results we illustrate that the scalable k-median fairlet decomposition introduced by [BIO+19] can be fed as input to Algorithm 1 to produce a fair probabilistic clustering, which can be assessed by our proposed metrics in Eqs. 10.4 and 10.5. This clustering may however have sub-optimal cost.

10.3.1 Probabilistic Model Fairlet Decomposition

To demonstrate our metrics and Algorithm, we directly translate the fair cost defined by [CKLV17] and construct a fairlet decomposition to minimize this cost for a probabilistic model known as a Gaussian mixture model (GMM).

A GMM describes the data distribution through a mixture of multivariate normal distributions $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$, covariance structure $\boldsymbol{\Sigma}$ and component weights $\boldsymbol{\pi}$. The distribution parameters can be inferred through the expectation maximization (EM) algorithm [Moo96], which iterates between updating the parameters (maximization step) and computing the responsibility $\gamma_{i,c}$ for all i, c (expectation step) until the likelihood converges. The responsibility can be computed by:

$$\gamma_{i,c} = \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\pi_c}{\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j} \tag{10.10}$$

where $\gamma_{i,c}$ is the probability that the data point $\boldsymbol{x}_i$ is generated by component c. Note that the total mass of a mixture component is $N_c = \sum_{i=1}^{N} \gamma_{i,c}$ and that the

sum of total component masses is the number of data points $N = \sum_{c=1}^{C} N_c$.

To construct a fairlet decomposition which is simultaneously fair and minimizes the distances between fairlet members in the space modelled by the GMM, we need a distance metric which takes into account the mixture model. The natural distance function for data modeled by a single multivariate Gaussian probability distribution $\mathcal{N}$ with covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$ is the Mahalanobis distance:

$$d_{\mathrm{M}}^2(\boldsymbol{x}, \mathcal{N}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{10.11}$$

where $d_{\mathrm{M}}^2(\boldsymbol{x}, \mathcal{N})$ is the squared Mahalanobis distance of a point $\boldsymbol{x}$ from the distribution $\mathcal{N}$.

The likelihood of a GMM is directly related to the Mahalanobis distance between observed points and presumed distributions. The log-likelihood of a data point belonging to a multivariate normal distribution is given by the logarithm of the probability density function of distribution $\mathcal{N}$:

$$\log\, L(\boldsymbol{x}) = -\frac{1}{2} \left[\log(|\boldsymbol{\Sigma}|) + \log(d_{\mathrm{M}}^2(\boldsymbol{x}, \mathcal{N})) + m \cdot \log(2\pi) \right], \tag{10.12}$$

where m is the multivariate dimension of $\mathcal{N}$.

When the data are modelled by a mixture of multiple Gaussians the covariance matrix $\boldsymbol{\Sigma}$ is not unique. To extend the notion of distance between points to this setting, the data space can be interpreted as a Riemannian manifold with metric $\boldsymbol{G}(\boldsymbol{x})$. This metric can be approximated leading to a model-weighted distance (MWD) [Tip99]:

$$d_{\mathrm{MWD}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{G} (\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{10.13}$$

where $d_{\mathrm{MWD}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the model-weighted distance between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and $\boldsymbol{G}$ is given by:

$$\boldsymbol{G} = \frac{\sum_{k=1}^{K} \boldsymbol{\Sigma}_k^{-1} \pi_k \int_{x_i}^{x_j} p(\boldsymbol{x}|k) d\boldsymbol{x}}{\sum_{k=1}^{K} \pi_k \int_{x_i}^{x_j} p(\boldsymbol{x}|k) d\boldsymbol{x}}, \tag{10.14}$$

where π_k is the mixing proportion of the k^{th} mixture component and $\int_{x_i}^{x_j} p(\boldsymbol{x}|k)\, d\boldsymbol{x}$ is the unidimensional integral of the probability density of the k^{th} component along the straight path between point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Computing the distance in this manner assumes a constant metric $\boldsymbol{G}$ along the path between the points. This metric can be interpreted as a probabilistically-weighted average of the inverse covariances of the different components in the mixture model. The integral is analytically tractable and is given by:

$$\int_{x_i}^{x_j} p(\boldsymbol{x}|k)\ d\boldsymbol{x} = \sqrt{\frac{\pi b^2}{2}} e^{-Z/2} \times \left[\operatorname{erf}\left(\frac{1-a}{\sqrt{2b^2}}\right) - \operatorname{erf}\left(\frac{-a}{\sqrt{2b^2}}\right)\right], \tag{10.15}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function and

$$b^2 = (\boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{v})^{-1}, \tag{10.16}$$

$$a = b^2 \boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u}, \tag{10.17}$$

$$Z = \boldsymbol{u}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u} - b^2 (\boldsymbol{v}^\intercal \boldsymbol{\Sigma}_k^{-1} \boldsymbol{u})^2, \tag{10.18}$$

with $\boldsymbol{u} = \boldsymbol{\mu}_k - \boldsymbol{x}_j$ and $\boldsymbol{v} = \boldsymbol{x}_i - \boldsymbol{x}_j$.

Equipped with a metric for computing distances between points we define the GMM fairlet decomposition cost by:

$$\mathcal{L}_{\mathrm{GMM}}(D, Q) = \sum_{\boldsymbol{x} \in D} d_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x})), \tag{10.19}$$

where the metric $\boldsymbol{G}$ describing data manifold is computed from a GMM on the original dataspace D. Similarly we define the GMM cost on the transformed dataset as:

$$\mathcal{L}_{\mathrm{GMM}}(\bar{D}, C^*) = \sum_{\boldsymbol{x} \in \bar{D}} d_{\mathrm{MWD}}(\boldsymbol{x}, \Gamma_{C^*}(\boldsymbol{x})), \tag{10.20}$$

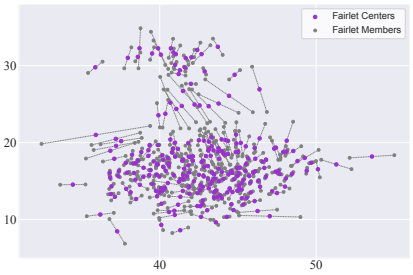
where $\Gamma_{C^*}(\boldsymbol{x})$ denotes the mean locations $\boldsymbol{\mu}$ of the components to which $\boldsymbol{x}$ is mapped. We restrict the distance of the k^{th} mixture component to be based on a $\boldsymbol{G}$ metric for the k^{th} component and it thus reduces to a weighted sum of Mahalanobis distances in the transformed dataspace $\bar{D}$, where the weights are dictated by the component responsibilities. This choice gives a more robust cost measure of the GMM fit. The direct translation of the total cost of the fair solution defined by [CKLV17] is then:

$$\mathcal{L}_{\mathrm{GMM-tot}}(D, C^*) = \mathcal{L}_{\mathrm{GMM}}(D, Q) + \mathcal{L}_{\mathrm{GMM}}(\bar{D}, C^*) \tag{10.21}$$

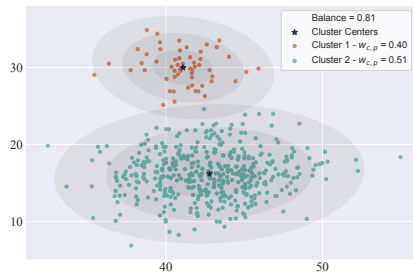
We generate a GMM fairlet decomposition by minimizing the GMM cost through a MCF algorithm[1]. We utilize the approach described in [CKLV17] for the k-median cost and change the weights on the edges of the graph to the MWD between points. Prior to running the algorithm the metric space is instantiated by fitting a traditional GMM with the desired number of components on the original data. The distribution parameters of these components are then used to

[1]Our code is publicly available at `https://github.com/RuneDK93/fair-soft-clustering`

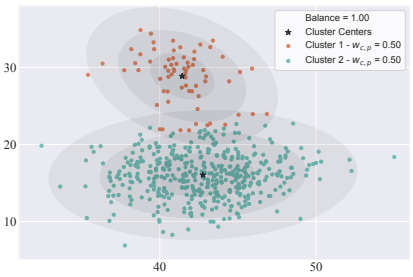
(a) Original data with colors.



(b) Fairlet decomposition.



(c) Traditional GMM.



(d) Fair GMM.

Figure 10.1: Illustration of our approach on simulated data of 500 points in $\mathbb{R}^2$. The original data points are shown colored according to their protected attribute in (a). A GMM fit on the original data is shown in (c). (b) shows a (1,1)-fairlet MWD decomposition of the data, while (d) shows the resulting fair solution from fitting a GMM on the fairlet centers in (b) and mapping the responsibilities $\gamma_{i,c}$ according to Algorithm 1. The points in (c) and (d) are colored according to the cluster index in $\gamma_{i,c}$ with highest probability. The weighted cluster color fractions $w_{c,p}$ in (c) and (d) are shown for the red class. Notice that the resulting balance is $B = 0.81$ for the traditional GMM fit in (c) and $B = 1.00$ for the fair GMM fit in (d).

generate the metric $\boldsymbol{G}$ and compute the model-weighted distances. The fairlet centers are then generated as the mean of the members in each fairlet.

Fig. 10.1 presents a visualisation of the approach on simulated data of 500 points in $\mathbb{R}^2$ with 250 red and 250 blue points. Fig 10.1a illustrates the original data with points colored according to their protected attribute. We apply a traditional GMM on the data to obtain a color blind solution shown in Fig. 10.1c. The balance of red and blue points allows us to construct a $(1, 1)$-fairlet

decomposition of the data through a perfect matching on the bichromatic graph. We construct the decomposition using the MCF approach by utilizing the distribution parameters of the colorblind solution to instantiate the $\boldsymbol{G}$ metric and use these distances on the edges of the graph. This results in a MWD fairlet decomposition Q of the data illustrated in Fig. 10.1b. The fairlet decomposition is then fed as input to Algorithm 1 to obtain the final fair clustering C^* shown in Fig. 10.1d. The fairness of both solutions is assessed with our proposed soft balance fairness metric (Eq. 10.4).

10.4 Results

We demonstrate our approach on real-world data by performing experiments on three widely used datasets in the fair clustering community. The datasets are Census[2], Bank[3] and Diabetes[4]. We select numerical features for the dimensions in the data point space and use 'sex' and 'marital status' as protected attributes. See Tab. 10.1 for an overview of the datasets.

Census: The dataset collects records of the 1994 US Census and presents an income prediction task based on various attributes of individuals. We select 'age', fnlwgt', 'education-num', 'capital-gain' and 'hours-per-week' as features representing the spacial dimensions of the data. We select 'sex' as the protected attribute.

Bank: The dataset [MCR14] is from a Portuguese phone call based bank marketing campaign. We select 'age', balance' and 'duration-of-account' as features representing the spacial dimensions of the data. We select 'marital-status' as the protected attribute.

Diabetes: The dataset [SDG+14] spans 10 years of information and outcomes of diabetes across 130 US hospitals. We select 'age' and 'time-in-hospital' as features representing the spacial dimensions of the data and select 'sex' as the protected attribute.

[2] https://archive.ics.uci.edu/ml/datasets/adult

[3] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[4] https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

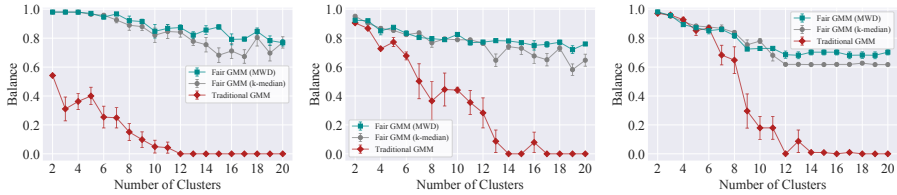
Table 10.1: Overview of the datasets used for our experiments. The table shows the number of spacial dimensions, the type of protected attribute and the color fraction for the three datasets.

Dataset	Dimension	Protected Att.	$r_{D,p}$
Census	5	sex	0.67
Bank	3	marital-status	0.62
Diabetes	2	sex	0.54

Similarly to [CKLV17] we sub-sample each dataset to 500 observations and preserve the protected attribute fraction from the original data. These fractions are $r_{D,p} = 0.67$ (Census), $r_{D,p} = 0.62$ (Bank) and $r_{D,p} = 0.54$ (Diabetes). For each dataset we apply a standard GMM on the original dataset and compare this to fair probabilistic clustering constructed by first finding a (1,2)-fairlet decomposition and then applying Algorithm 1 on the decomposition. We conduct the experiments for a fairlet decomposition generated by the method described in [BIO+19] (minimizing the k-median Euclidean distance cost) and for a fairlet decomposition generated by minimizing the MWD cost in a MCF setting. Algorithm 1 ensures that the fairness of both approaches is bounded, but the cost of the solutions are different.

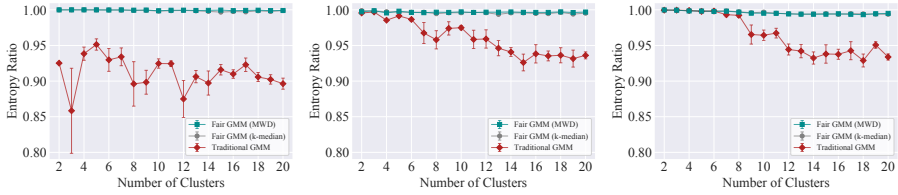
The final clustering outcome is dependent on the initialization of the GMM. We initialize the GMM using k-means clustering and repeat the overall clustering and fairlet decomposition 5 times with different random seeds for the initialization parameters to generate mean values and associated standard errors.

Fig. 10.2 shows the resulting soft balance (top row) according to Eq. 10.4 and soft entropy ratio (bottom row) according to Eq. 10.6. For all datasets the fairness disparity between the traditional and fair solutions increases sharply with the number of clusters. Observe that for a large number of mixture components, the colorblind model has a balance of zero for all datasets. The optimal GMM solution to the data thus requires monochromatic clusters. Additionally, the fair GMM shows less fairness variance and is thus more robust to the initialization. This is especially the case for entropy fairness, where the fair GMM is highly robust while the color blind GMM is much more sensitive to the initialization parameters. Note that the fair GMM generated from the Euclidean distance k-median decomposition and MWD decomposition achieve equivalent fairness scores.



(a) Census. **(b)** Bank. **(c)** Diabetes.

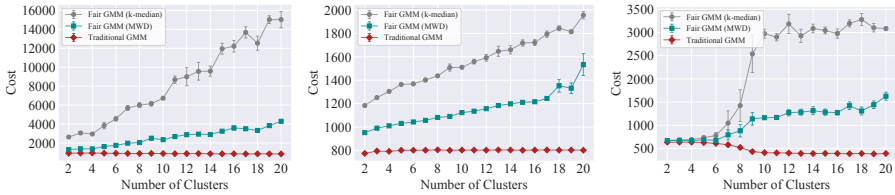
Figure 10.2: Fairness in terms of soft balance on the three datasets. Data points are the mean values from 5 iterations of different random seeds. Error bars indicate standard errors. Observe that the traditional GMM cluster approach obtains monochromatic cluster solutions ($B = 0$) for all datasets. Also note that due to Algorithm 1, the balance of the fair solutions are bounded. For instance the fair solutions of the Diabetes dataset are bounded at balance $B \geq 0.62$ (see Appendix 10.7 for details).



(a) Census. **(b)** Bank. **(c)** Diabetes.

Figure 10.3: Fairness in terms of soft entropy ratio on the three datasets.

The price of fairness is quantified by the negative impact on the cost of the solutions. This is illustrated in Fig. 10.4, where the costs of the different cluster solutions are shown for the three datasets. The costs for the fair solutions are computed by Eq. 10.21 while the cost of the traditional solution is computed by Eq. 10.20 on the original data as $\mathcal{L}_{\text{GMM}}(D, C)$, where C is a traditional GMM clustering on D. The traditional GMM solution has the lowest cost, while the fair MWD GMM has a larger cost, which increases with the number of cluster components. The fair solution from the Euclidean k-median fairlet decomposition has the highest cost, which is significantly higher than both the traditional GMM and fair MWD GMM for large k. Note that for the diabetes dataset, the cost of the fair and traditional cluster solutions are similar for low k. Likewise, the difference in cost for this data between the solutions from the Euclidean and MWD decompositions is small here.



(a) Census. **(b)** Bank. **(c)** Diabetes.

Figure 10.4: Total cost on the three datasets. The costs for the fair solutions are computed as $\mathcal{L}_{\mathrm{GMM-tot}}(D, C^*)$ according to Eq. 10.21 while the cost of the standard GMM is computed according to Eq. 10.20 on the original data as $\mathcal{L}_{\mathrm{GMM}}(D, C)$, where C is a traditional GMM clustering on D.

For our results we have used a direct translation of the fairlet decomposition cost from previous work based on minimization of distances induced by the clustering objective. A GMM cluster fit can also be evaluated in terms of likelihood. In Appendix 10.8 we present an approach for evaluating the likelihood of a GMM fairlet decomposition, which supports the results shown in Fig. 10.4.

10.5 Discussion

Our results demonstrate that the generalized fairness metrics can be used to assess fairness of probabilistic cluster solutions and that such fair solutions can be obtained through a fairlet decomposition of the data fed as input to Algorithm 1. We observe that the fair GMM ensures high fairness even for a large number of mixture components, whereas the fairness of the traditional GMM becomes progressively worse and ultimately dictates monochromatic clusters ($B = 0$). Additionally, we note that a fair GMM generated with the Euclidean k-median objective produces similarly fair solutions as the fair GMM found from a MWD decomposition. Generally, the fair solutions demonstrate much less variance on the entropy metric than on the balance metric across the different number of cluster components. This is because the balance measure is determined by the least balanced cluster, while the entropy measure takes the fairness of all clusters into account. However, the balance measure has an intuitive appeal, as it measures the worst case fairness among all clusters, and consequently it may be better suited for ensuring adherence to the DI doctrine.

From our experiments we also observe that the fair GMM solutions increase the cost over a traditional GMM. The cost increase is significantly lower for the MWD

decomposition than for the Euclidean k-median decomposition, especially for a large number of mixture components in higher-dimensional spaces. This is due to the fact that the k-median decomposition is designed to locate data points close in Euclidean space, which may be located far apart in the non-Euclidean space induced by the GMM. On the other hand, the MWD decomposition specifically connects points which lie close on the data manifold dictated by the GMM. Note that for a low number of mixture components and for data in a low-dimensional space (like the diabetes dataset), the Euclidean k-median decomposition does not increase the cost significantly over the MWD decomposition. This indicates than in such settings the Euclidean k-median decomposition approach introduced by [BIO+19] can be used as a highly scaleable alternative to the MWD decomposition for obtaining fair GMM solutions without significantly increasing the cost.

Our GMM fairlet decomposition is constructed from a direct translation of the fair cost introduced by [CKLV17], which involves generating a fairlet decomposition through minimization of distances induced by the clustering objective. While the Mahalanobis distance is directly connected to maximum likelihood estimation of a GMM, the connection between the likelihood and the MWD is less clear. Future research directions for GMM fairlet decompositions could involve linking the decomposition construction directly to the likelihood of the solution, or to construct a more scaleable GMM fairlet decomposition.

10.6 Conclusion

Previous work on fair clustering has focused on deterministic hard clustering algorithms like k-means and k-median, where data points belong to specific clusters in a binary sense. In this work we study fair soft clustering by proposing generalizations of group-level fairness metrics. These generalizations allow the fairness metrics to be used in the presence of soft clustering algorithms by reflecting the underlying probabilistic nature. Furthermore, we have demonstrated an approach for obtaining a fair probabilistic cluster solution from a fairlet decomposition of the data. This approach may be applied on decompositions tailored specifically to mixture models, but can also be used to modify fairlet decompositions from previous work on hard clustering algorithms. Ultimately, the resulting solutions are costlier than their traditional counterparts, but in turn provide guaranteed bounds on their fairness.

Acknowledgments and Disclosure of Funding

Author Rune D. Kjærsgaard is funded by a university alliance scholarship between UiB (University of Bergen) and DTU (Technical University of Denmark).

10.7 Appendix A: Algorithm 1 Fairness Bound

This section explains the theoretical bound on the fairness of the solution C^* provided by Algorithm 1.

Consider a dataset D with a binary protected attribute. A fairlet decomposition Q of this data can be specified as a (p_1, p_2)-fairlet decomposition with parameters p_1 and p_2 (where $p_1 < p_2$) indicating that all fairlets have a color fraction $r \geq \frac{p_1}{p_1+p_2}$. The color fraction obtained from the union of these fairlets is bounded according to Lemma 1 (analogous to Lemma 2 from [CKLV17]).

Lemma 1 (Combination):
Let $Y_1, Y_2 \subseteq D$ be disjoint. If C_1 is a clustering of Y_1 and C_2 is a clustering of Y_2, then $r(C_1 \cup C_2) \geq min(r(C_1), r(C_2))$.

Algorithm 1 combines the micro-clusters $q_1, q_2, \cdots, q_\ell$ (fairlets) into the probabilistic clustering C^* through a weighted combination dictated by the responsibilities $\boldsymbol{\gamma}$ of the fairlet centers. The weighted color fraction w of this combination is given by Eq. 10.3 and is bounded according to Lemma 2.

Lemma 2 (Weighted combination):
Let $Y_1, Y_2 \subseteq D$ be disjoint. If C_1 is a weighted clustering of Y_1 and C_2 is a weighted clustering of Y_2, then $w(C_1 \cup C_2) \geq min(w(C_1), w(C_2))$.

This means that the weighted color fractions w_c of the final mixture components in C^* are bounded by $w_c \geq \frac{p_1}{p_1+p_2}\ \forall\ c$. To take a concrete example consider the Diabetes dataset from our experiments in Sect. 10.4. We perform a $(1, 2)$-fairlet decomposition on the dataset and the weighted color fraction for any of the final mixture components is thus bounded by $w_c \geq \frac{1}{3}\ \forall\ c$. This dataset has an overall color fraction of $r_D = 0.54$. The soft balance of the final cluster solution is then bounded by $B \geq \frac{w_c}{r_D}$, i.e. $B \geq \frac{1/3}{0.54}$. This can be verified by inspecting Fig. 10.2, where the balance for the fair solution on the Diabetes dataset never drops below $\frac{1/3}{0.54} = 0.62$.

10.8 Appendix B: GMM Decomposition Likelihood

Our GMM fairlet decomposition is constructed by adapting the cost introduced by [CKLV17], which involves generating a fairlet decomposition through minimization of distances induced by the clustering objective. In our GMM fairlet decomposition we operate with the Mahalanobis and model-weighted distance. The cost of a GMM is typically not evaluated based on distances, but rather in terms of the log-likelihood. The log-likelihood of a data point belonging to a multivariate normal distribution is directly related to the Mahalanobis distance and is given by:

$$\log\, L(\boldsymbol{x}) = -\frac{1}{2}\left[\log(|\boldsymbol{\Sigma}|) + \log(d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})) + m \cdot \log(2\pi)\right], \tag{10.22}$$

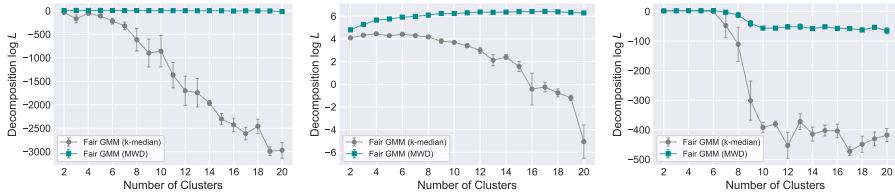
where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix, $d^2_{\mathrm{M}}(\boldsymbol{x}, \mathcal{N})$ is the squared Mahalanobis distance between data point $\boldsymbol{x}$ and distribution $\mathcal{N}$ and m is the multivariate dimension of $\mathcal{N}$.

The model weighted distance is a generalization of the Mahalanobis distance to the Gaussian mixture setting. The model-weighted distance reduces to the Mahalanobis distance in settings with a single Gaussian, or in regions of space where only a single component density $p(\boldsymbol{x}|k)$ is non-zero along the path between the points [Tip99].

While the Mahalanbis distance is directly related to the likelihood of a GMM solution, the connection between the model-weighted distance and the likelihood is less clear, and consequently the likelihood of the fairlet decomposition is harder to evaluate. However, we propose to estimate the likelihood by substituting the covariance matrix $\boldsymbol{\Sigma}$ in Eq. 10.22 with the model-weighted distance metric $\boldsymbol{G}$, and consequently the Mahalanobis distance with the model-weighted distance. The log-likelihood of a data point (fairlet member) belonging to a fairlet is then:

$$\log\, L_{\mathrm{Fairlet}}(\boldsymbol{x}) = -\frac{1}{2}[\log(|\boldsymbol{G}^{-1}|) + \log(d^2_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x}))) + m \cdot \log(2\pi)], \tag{10.23}$$

where $d^2_{\mathrm{MWD}}(\boldsymbol{x}, \beta_{\mathrm{Q}}(\boldsymbol{x}))$ is the model-weighted distance from fairlet member $\boldsymbol{x}$ to fairlet center $\beta_{\mathrm{Q}}(\boldsymbol{x})$ and $|\boldsymbol{G}^{-1}|$ is the determinant of the associated inverse model-weighted distance metric. Under this view Eq. 10.23 evaluates the likelihood that a fairlet member was generated by the fairlet it is assigned to. Fig. 10.5 shows the log-likelihood of the fairlet decompositions.



(a) Census. **(b)** Bank. **(c)** Diabetes.

Figure 10.5: Per observation average log-likelihood of the fairlet decompositions of the three datasets. The log-likelihood is evaluated with Eq. 10.23. Data points are mean values from 5 iterations of different random seeds. Error bars indicate standard errors.

Bibliography

[AADD+14] Étienne Artigau, Nicola Astudillo-Defru, Xavier Delfosse, François Bouchy, Xavier Bonfils, Christophe Lovis, Francesco Pepe, Claire Moutou, Jean-François Donati, René Doyon, et al. Telluric-line subtraction in high-accuracy velocimetry: a pca-based approach. In *Observatory Operations: Strategies, Processes, and Systems V*, volume 9149, page 914905. International Society for Optics and Photonics, 2014.

[ABDL20] Nihesh Anderson, Suman K Bera, Syamantak Das, and Yang Liu. Distributional individual fairness in clustering. *arXiv preprint arXiv:2006.12589*, 2020.

[ABM08] Peter Armitage, Geoffrey Berry, and John Nigel Scott Matthews. *Statistical methods in medical research*. John Wiley & Sons, 2008.

[ADW19] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

[AJJ19] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019.

[Alb16] Waleed Albattah. The role of sampling in big data analysis. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, BDAW '16, New York, NY, USA, 2016. Association for Computing Machinery.

[AMJ18] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

[ANM+21] Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. Rp-mod&rp-crowd: Moderator-and crowd-annotated german news comment datasets. In *NeurIPS Datasets and Benchmarks*, 2021.

[AP03] Pankaj K Agarwal and Cecilia M Procopiuc. Approximation algorithms for projective clustering. *Journal of Algorithms*, 46(2):115–139, 2003.

[APSN13] Anita S Acharya, Anupam Prakash, Pikee Saxena, and Aruna Nigam. Sampling: Why and how of it. *Indian Journal of Medical Specialties*, 4(2):330–333, 2013.

[ARK16] N. Armanfard, J. P. Reilly, and M. Komeili. Local feature selection for data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6):1217–1227, 2016.

[ARZV21] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *arXiv preprint arXiv:2109.13228*, 2021.

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[BACM+22] Aaron Bello-Arufe, Samuel HC Cabot, João M Mendonça, Lars A Buchhave, and Alexander D Rathcke. Mining the ultrahot skies of hat-p-70b: Detection of a profusion of neutral and ionized species. *The Astronomical Journal*, 163(2):96, 2022.

[BC11] Danah Boyd and Kate Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, September 2011.

[BCD+20] Brian Brubach, Darshan Chakrabarti, John Dickerson, Samir Khuller, Aravind Srinivasan, and Leonidas Tsepenekas. A pairwise fair and community-preserving approach to k-center clustering. In *International Conference on Machine Learning*, pages 1178–1189. PMLR, 2020.

[BCD+21] Brian Brubach, Darshan Chakrabarti, John P Dickerson, Aravind Srinivasan, and Leonidas Tsepenekas. Fairness, semi-supervised learning, and more: A general framework for clustering with stochastic pairwise constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6822–6830, 2021.

[BD21] Björn Barz and Joachim Denzler. Wikichurches: A fine-grained dataset of architectural styles with real-world challenges. *arXiv preprint arXiv:2108.06959*, 2021.

[Ber17] Maciej Bereswicz. A two-step procedure to measure representativeness of internet data sources. *International Statistical Review*, (85):473–493, 2017.

[BG98] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

[BG18a] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability, and Transparency; Proceedings of Machine Learning Research*, 81:1–15, 2018.

[BG18b] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[BG18c] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[BGK+18] Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*, 2018.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[BHFM+19] Megan Bedell, David W Hogg, Daniel Foreman-Mackey, Benjamin T Montet, and Rodrigo Luger. wobble: a data-driven analysis technique for time-series stellar spectra. *The Astronomical Journal*, 158(4):164, 2019.

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

[big19] Big data and representativity 2019 global market research report. `https://archive.researchworld.com/big-data-and-representativity-data-is-not-research/`, 2019.

[BIO+19] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

[BJP13] Marc H Bornstein, Justin Jager, and Diane L Putnick. Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental review*, 33(4):357–370, 2013.

[BK88] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

[BKH18] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin. Data sampling approaches with severely imbalanced big data for medicare fraud detection. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 137–142, 2018.

[BKH+21] Visar Berisha, Chelsea Krantsevich, P Richard Hahn, Shira Hahn, Gautam Dasarathy, Pavan Turaga, and Julie Liss. Digital medicine and the curse of dimensionality. *NPJ digital medicine*, 4(1):153, 2021.

[BKL+04] Mark Berman, Harri Kiiveri, Ryan Lagerstrom, Andreas Ernst, Rob Dunne, and Jonathan F Huntington. Ice: A statistical approach to identifying endmembers in hyperspectral images. *IEEE transactions on Geoscience and Remote Sensing*, 42(10):2085–2095, 2004.

[BLF+14] Jean-Loup Bertaux, Rosine Lallement, Stéphane Ferron, Cathy Boonne, and R Bodichon. Tapas, a web-based service of atmospheric transmission computation for astronomy. *Astronomy & Astrophysics*, 564:A46, 2014.

[BMD+12] Chad F Bender, Suvrath Mahadevan, Rohit Deshpande, Jason T Wright, Arpita Roy, Ryan C Terrien, Steinn Sigurdsson, Lawrence W Ramsey, Donald P Schneider, and Scott W Fleming. The sdss–het survey of kepler eclipsing binaries: Spectroscopic dynamical masses of the kepler-16 circumbinary planet hosts. *The Astrophysical Journal Letters*, 751(2):L31, 2012.

[BMZ21] Megan L. Blatchford, Chris M. Mannaerts, and Yijian Zeng. Determining representative sample size for validation of continuous, large continental remote sensing data. *International Journal of Applied Earth Observations and Geoinformation*, (94):102235, 2021.

[BN01] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

[Bro15] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[BSDG16] Adrian Barbu, Yiyuan She, Liangjing Ding, and Gary Gramajo. Feature selection with annealing for computer vision and big data learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):272–286, 2016.

[BSH+10] Jacob L. Bean, Andreas Seifahrt, Henrik Hartman, Hampus Nilsson, Günter Wiedemann, Ansgar Reiners, Stefan Dreizler, and Todd J. Henry. The CRIRES Search for Planets Around the Lowest-mass Stars. I. High-precision Near-infrared Radial Velocities with an Ammonia Gas Cell. , 713(1):410–422, April 2010.

[BTR17] Paula Branco, Luís Torgo, and Rita P Ribeiro. Smogn: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR, 2017.

[BTR19] Paula Branco, Luis Torgo, and Rita P Ribeiro. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99, 2019.

[CB94a] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[CB94b] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, (36):338–347, 1994.

[CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[CDKV16a] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[CDKV16b] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[CGS18] Sebastian Claici, Aude Genevay, and Justin Solomon. Wasserstein measure coresets. *arXiv preprint arXiv:1805.07412*, 2018.

[CHH+09] Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, , and Ernst L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *International Journal of Epidemiology*, (38):1175–1191, 2009.

[CHUK18] Yeounoh Chung, Peter J Haas, Eli Upfal, and Tim Kraska. Unknown examples & machine learning model generalization. *arXiv preprint arXiv:1808.08294*, 2018.

[Cic14] Andrzej Cichocki. Era of big data processing: A new approach via tensor networks and tensor decompositions. *arXiv preprint arXiv:1403.2048*, 2014.

[CJK04] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[CK23] Line H Clemmensen and Rune D Kjærsgaard. Data representativity for machine learning and ai systems. *arXiv preprint arXiv:2203.04706*, 2023.

[CKLV17] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

[CL21] Kyla Chasalow and Karen Levy. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–89, 2021.

[CLG01] Rich Caruana, Steve Lawrence, and Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, pages 402–408, 2001.

[CLP+12] Rosario Cosentino, Christophe Lovis, Francesco Pepe, Andrew Collier Cameron, David W. Latham, Emilio Molinari, Stephane Udry, Naidu Bezawada, Martin Black, Andy Born, Nicolas Buchschacher, Dave Charbonneau, Pedro Figueira, Michel Fleury, Alberto Galli, Angus Gallie, Xiaofeng Gao, Adriano Ghedina, Carlos Gonzalez, Manuel Gonzalez, Jose Guerra, David Henry, Keith Horne, Ian Hughes, Dennis Kelly, Marcello Lodi, David Lunney, Charles Maire, Michel Mayor, Giusi Micela, Mark P. Ordway, John Peacock, David Phillips, Giampaolo Piotto, Don Pollacco, Didier Queloz, Ken Rice, Carlos Riverol, Luis Riverol, Jose San Juan, Dimitar Sasselov, Damien Segransan, Alessandro Sozzetti, Danuta Sosnowska, Brian Stobie, Andrew Szentgyorgyi, Andy Vick, and Luc Weber. Harps-N: the new planet hunter at TNG. In Ian S. McLean, Suzanne K. Ramsay, and Hideki Takami, editors, *Ground-based and Airborne Instrumentation for Astronomy IV*, volume 8446 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 84461V, September 2012.

[CLT+19a] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[CLT+19b] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[CMM21] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.

[CMW+20] Samuel H. C. Cabot, Nikku Madhusudhan, Luis Welbanks, Anjali Piette, and Siddharth Gandhi. Detection of neutral atomic species in the ultra-hot Jupiter WASP-121b. , 494(1):363–377, May 2020.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[CSF+14] Cunha, D., Santos, N. C., Figueira, P., Santerne, A., Bertaux, J. L., and Lovis, C. Impact of micro-telluric lines on precise radial velocities and its correction. *A&A*, 568:A35, 2014.

[CSM+05] SA Clough, MW Shephard, EJ Mlawer, JS Delamere, MJ Iacono, K Cady-Pereira, S Boukabara, and PD Brown. Atmospheric radiative transfer modeling: A summary of the aer codes. *Journal*

of Quantitative Spectroscopy and Radiative Transfer, 91(2):233–244, 2005.

[CSS21] Sixia Chen, Alexander Stubblefield, and Julie A Stoner. Oversampling of minority populations through dual-frame surveys. *Journal of survey statistics and methodology*, 9(3):626–649, 2021.

[CWT+22] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022.

[CWW21] Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*, 2021.

[CWY+23] Yu-Neng Chuang, Guanchu Wang, Fan Yang, Zirui Liu, Xuanting Cai, Mengnan Du, and Xia Hu. Efficient xai techniques: A taxonomic survey. *arXiv preprint arXiv:2302.03225*, 2023.

[CY10] Danyang Cao and Bingru Yang. An improved k-medoids clustering algorithm. In *2010 the 2nd international conference on computer and automation engineering (ICCAE)*, volume 3, pages 132–135. IEEE, 2010.

[D+00] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

[DB79] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[DCS+21] X Dumusque, M Cretignier, D Sosnowska, N Buchschacher, C Lovis, DF Phillips, F Pepe, F Alesina, LA Buchhave, J Burnier, et al. Three years of harps-n high-resolution spectroscopy and precise radial velocity data for the sun. *Astronomy & Astrophysics*, 648:A103, 2021.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[DM20] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. *FAT* '20, January 27-30, 2020, Barcelona, Spain*, 2020.

[dSMBC+21] Cristiane dos Santos Machado, Pedro L. Ballester, Bo Cao, Benson Mwangi, Marco Antonio Caldieraro, Flávio Kapczinski, and Ives Cavalcante Passos. Prediction of suicide attempts in a prospective cohort study with a nationally representative sample of the us population. *Psychological Medicine*, pages 1–12, 2021.

[DTU23] DTU Computing Center. DTU Computing Center resources, 2023.

[D4] Ben D'Excelle. *Representative Sample*. Springer, 2014.

[EBTD20] Seyed Esmaeili, Brian Brubach, Leonidas Tsepenekas, and John Dickerson. Probabilistic fair clustering. *Advances in Neural Information Processing Systems*, 33:12743–12755, 2020.

[EMH19] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[FAA+16] Debra A. Fischer, Guillem Anglada-Escude, Pamela Arriagada, Roman V. Baluev, Jacob L. Bean, Francois Bouchy, Lars A. Buchhave, Thorsten Carroll, Abhijit Chakraborty, Justin R. Crepp, Rebekah I. Dawson, Scott A. Diddams, Xavier Dumusque, Jason D. Eastman, Michael Endl, Pedro Figueira, Eric B. Ford, Daniel Foreman-Mackey, Paul Fournier, Gabor Fűrész, B. Scott Gaudi, Philip C. Gregory, Frank Grundahl, Artie P. Hatzes, Guillaume Hébrard, Enrique Herrero, David W. Hogg, Andrew W. Howard, John A. Johnson, Paul Jorden, Colby A. Jurgenson, David W. Latham, Greg Laughlin, Thomas J. Loredo, Christophe Lovis, Suvrath Mahadevan, Tyler M. McCracken, Francesco Pepe, Mario Perez, David F. Phillips, Peter P. Plavchan, Lisa Prato, Andreas Quirrenbach, Ansgar Reiners, Paul Robertson, Nuno C. Santos, David Sawyer, Damien Segransan, Alessandro Sozzetti, Tilo Steinmetz, Andrew Szentgyorgyi, Stéphane Udry, Jeff A. Valenti, Sharon X. Wang, Robert A. Wittenmyer, and Jason T. Wright. State of the Field: Extreme Precision Radial Velocities. , 128(964):066001, June 2016.

[FCMR21] Marco Fumero, Luca Cosmo, Simone Melzi, and Emanuele Rodolà. Learning disentangled representations via product manifold projection. *arXiv preprint arXiv:2103.01638*, 2021.

[Fel20] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

[FHL14] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[Fis35] Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, 1935.

[FLTW20] Fenglei Fan, Mengzhou Li, Yueyang Teng, and Ge Wang. Soft autoencoder and its wavelet adaptation interpretation. *IEEE Transactions on Computational Imaging*, 6:1245–1257, 2020.

[FMI83] Kunihiko Fukushima, Sei Miyake, and Takayuki Ito. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5):826–834, 1983.

[FSS20] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[GBCB16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1 of *1*. MIT press Cambridge, 2016.

[GBH+22] Srishti Gautam, Ahcene Boubekki, Stine Hansen, Suaiba Salahuddin, Robert Jenssen, Marina Höhne, and Michael Kampffmeyer. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022.

[GBR+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[GC19] Benyamin Ghojogh and Mark Crowley. Instance ranking and numerosity reduction using matrix decomposition and subspace learning. In *Canadian Conference on Artificial Intelligence*, pages 160–172. Springer, 2019.

[GCGS14a] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014.

[GCGS14b] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014.

[GDRK14] Kevin Gullikson, Sarah Dodson-Robinson, and Adam Kraus. Correcting for telluric absorption: methods, case studies, and release of the telfit code. *The Astronomical Journal*, 148(3):53, 2014.

[GDVJ+19] Stephen Goldrick, Carlos A Duran-Villalobos, Karolis Jankauskas, David Lovett, Suzanne S Farid, and Barry Lennox. Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. *Computers & Chemical Engineering*, 130:106471, 2019.

[Gha21] Laurent El Ghaoui. Hyper-Textbook: Optimization Models and Applications gram matrix description, 2021.

[GHH+23] Srishti Gautam, Marina M-C Höhne, Stine Hansen, Robert Jenssen, and Michael Kampffmeyer. This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. *Pattern Recognition*, 136:109172, 2023.

[Gho21] Ghojogh, Benyamin. *Data Reduction Algorithms in Machine Learning and Data Science*. PhD thesis, 2021.

[Gid12] Lior Gideon. *Handbook of Survey Methodology for the Social Sciences*. Springer, 2012.

[Gil21] William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. *arXiv preprint arXiv:2110.05266*, 2021.

[GMT20] Alexander N Gorban, Valery A Makarov, and Ivan Y Tyukin. High-dimensional brain in a high-dimensional world: Blessing of dimensionality. *Entropy*, 22(1):82, 2020.

[GMV+21a] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[GMV+21b] Timnit Gebru, Jamie Morgenstern, Briana Vechhione, Jennifer Wrotmen Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. Datasheets for datasets. *arXiv:1803.09010v8*, 2021.

[GNG+20] Benyamin Ghojogh, Hadi Nekoei, Aydin Ghojogh, Fakhri Karray, and Mark Crowley. Sampling algorithms, from survey sampling to monte carlo methods: Tutorial and literature review. *arXiv:2011.00901v1*, 2020.

[GPBV19] Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. Dppy: Dpp sampling with python. *J. Mach. Learn. Res.*, 20:180–1, 2019.

[GŞL+15] Stephen Goldrick, Andrei Ştefan, David Lovett, Gary Montague, and Barry Lennox. The development of an industrial-scale fed-batch fermentation simulation. *Journal of biotechnology*, 193:70–82, 2015.

[GSMC21] M. Grønberg, K. Svendsen, I. Måge, and L. Clemmensen. Poster: Sampling to adjust for imbalance in production data. In *ENBIS 2021 Spring Meeting*, May 2021.

[GVL13] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

[Gy98] Pierre Gy. *Sampling for Analytical Purposes*. Wiley, 1998.

[H+70] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[HAR] HARPS-N Instrument Page. `http://www.tng.iac.es/instruments/harps/`.

[HE17] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[HG15] Zena Hira and Duncan Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015:1–13, 07 2015.

[HJH12] Mary Hibberts, R. Burke Johnson, and Kenneth Hudson. *Common Survey Sampling Techniques*. Springer, 2012.

[HKP12] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012.

[HLLL19] Yun Hou, Bailin Li, Li Li, and Jiajia Liu. A density-based undersampling algorithm for imbalance classification. In *Journal of Physics: Conference Series*, volume 1302, page 022064. IOP Publishing, 2019.

[HMG+16] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning, 2016.

[HPG14] Rob Heylen, Mario Parente, and Paul Gader. A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):1844–1868, 2014.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[HPT22] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[HS06] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[HSK+12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[Hua21] Jonathan Yinhao Huang. Representativeness is not representative - addressing major inferential threats in the uk biobank and other big data repositories. *Epidemiology*, (32):189–193, 2021.

[HWB+21] Zhe Huang, Liang Wang, Giles Blaney, Christopher Slaughter, Devon McKeon, Ziyu Zhou, Robert Jacob, and Michael C Hughes. The tufts fnirs mental workload dataset & benchmark for brain-computer interfaces that generalize. 2021.

[HYS+17] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[HZ94] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1994.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[JK19] J.M. Johnson and T.M. Khoshgoftaar. Survey on deep learning with class imbalance. *J Big Data*, 6(27), 2019.

[JMG22] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai (dcai). *arXiv preprint arXiv:2211.14611*, 2022.

[JPN+20] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3561–3562, 2020.

[JVLT21] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[JXS+20] Zhongjun Jin, Mengjing Xu, Chenkai Sun, Abolfazl Asudeh, and HV Jagadish. Mithracoverage: a system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2721–2724, 2020.

[Kal03] William D Kalsbeek. Sampling minority groups in health surveys. *Statistics in Medicine*, 22(9):1527–1549, 2003.

[KAL20] Amit Kaushal, Russ Altman, and Curt Langlotz. Geographic distribution of us cohorts used to train deep learning algorithms. *Jama*, 324(12):1212–1213, 2020.

[Kan60] Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.

[KB96] Ronny Kohavi and Barry Becker. Adult data set. *UCI machine learning repository*, 5:2093, 1996.

[KBAR+21] Rune D Kjærsgaard, Aaron Bello-Arufe, Alexander D Rathcke, Lars A Buchhave, and Line KH Clemmensen. Unsupervised spectral unmixing for telluric correction using a neural network autoencoder. *Conference: NeurIPS 2021 - Workshop on Machine Learning and the Physical Sciences.*, 2021.

[KBAR+23] Rune D Kjærsgaard, Aaron Bello-Arufe, Alexander D Rathcke, Lars A Buchhave, and Line KH Clemmensen. Tau: A neural network based telluric correction framework. *Astronomy and Astrophysics*, 2023.

[KBAW07] William D Kalsbeek, Walter R Boyle, Robert P Agans, and John E White. Disproportionate sampling for population subgroups in telephone surveys. *Statistics in medicine*, 26(8):1657–1674, 2007.

[KBC23] Rune D Kjærsgaard, Ahcene Boubekki, and Line KH Clemmensen. Pantypes: Diverse representatives for self-explainable models. *Manuscript submitted for publication*, 2023.

[KDS+22] Srikant Manas Kala, Kunal Dahiya, Vanlin Sathya, Teruo Higashino, and Hirozumi Yamaguchi. Lte-laa cell selection through operator data learning and numerosity reduction. *Pervasive and Mobile Computing*, 83:101586, 2022.

[KFK+20] Murat Kulahci, Flavia Dalia Frumosu, Abdul Rauf Khan, Georg Ørnskov Rønsch, and Max Peter Spooner. Experiences with big data: Accounts from a data scientist's perspective. *Quality Engineering*, (32):529–542, 2020.

[KGC21a] Rune D Kjærsgaard, Manja G Grønberg, and Line KH Clemmensen. Sampling to improve predictions for underrepresented observations in imbalanced data. *Proceedings of Workshop on Data-Centric AI, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).*, 2021.

[KGC21b] Rune D Kjærsgaard, Manja G Grønberg, and Line KH Clemmensen. Sampling to improve predictions for underrepresented observations in imbalanced data. *arXiv preprint arXiv:2111.09065*, 2021.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[KKS+19] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, (17), 2019.

[KŁ16] Jerzy Krawczuk and Tomasz Łukaszuk. The feature selection bias problem in relation to high-dimensional gene data. *Artificial intelligence in medicine*, 66:63–71, 2016.

[KL21] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3d garments with sewing patterns. *arXiv preprint arXiv:2109.05633*, 2021.

[KM79a] William Kruskal and Frederick Mosteller. Representative sampling, i: Non-scientific literature. *International Statistical Review*, (47):13–24, 1979.

[KM79b] William Kruskal and Frederick Mosteller. Representative sampling, ii: Scientific literature, excluding statistics. *International Statistical Review*, (47):111–127, 1979.

[KM79c] William Kruskal and Frederick Mosteller. Representative sampling, iii: the current statistical literature. *International Statistical Review*, (47):245–265, 1979.

[KM80] William Kruskal and Frederick Mosteller. Representative sampling, iv: the history of the concept in statics, 1895-1939. *International Statistical Review*, (48):169–195, 1980.

[KNS+15] W Kausch, S Noll, A Smette, S Kimeswenger, M Barden, C Szyszka, AM Jones, Hugues Sana, H Horst, and F Kerber. Molecfit: A general tool for telluric absorption correction-ii. quantitative evaluation on eso-vlt/x-shooterspectra. *Astronomy & Astrophysics*, 576:A78, 2015.

[KPR+17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[KPS+23] Rune D Kjærsgaard, Pekka Parviainen, Saket Saurabh, Madhumita Kundu, and Line KH Clemmensen. Fair soft clustering. *Manuscript submitted for publication*, 2023.

[Kra16] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[KS10] Betty R Kirkwood and Jonathan AC Sterne. *Essential medical statistics*. John Wiley & Sons, 2010.

[KS19] I. Måge K. Svendsen, L. Clemmensen. 5 points to consider before setting sails in data science projects in the food industry. *iProcess 2016-2019, Report no: 2019:01041*, 2019.

[KSC+21] Ivan Kiskin, Marianne Sinka, Adam D Cobb, Waqas Rafique, Lawrence Wang, Davide Zilli, Benjamin Gutteridge, Rinita Dam, Theodoros Marinos, Yunpeng Li, et al. Humbugdb: a large-scale acoustic mosquito dataset. *arXiv preprint arXiv:2110.07607*, 2021.

[KSG08a] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

[KSG08b] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

[KT11] Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011.

[KT+12a] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[KT+12b] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[KTR+21] Lukas Kondmann, Aysim Toker, Marc Rußwurm, Andres Camero Unzueta, Devis Peressuti, Grega Milcinski, Nicolas Longépé, Pierre-Philippe Mathieu, Timothy Davis, Giovanni Marchisio, et al. Denethor: The dynamicearthnet dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[KTS+12] Khamisi Kalegele, Hideyuki Takahashi, Johan Sveholm, Kazuto Sasai, Gen Kitagata, and Tetsuo Kinoshita. On-demand data numerosity reduction for learning artifacts. In *2012 IEEE 26th International Conference on Advanced Information Networking and Applications*, pages 152–159. IEEE, 2012.

[KTS+13] Khamisi Kalegele, Hideyuki Takahashi, Johan Sveholm, Kazuto Sasai, Gen Kitagata, and Tetsuo Kinoshita. Numerosity reduction for resource constrained learning. *Journal of information processing*, 21(2):329–341, 2013.

[KU12] Håkon Kile and Kjetil Uhlen. Data reduction via clustering and averaging for contingency and reliability analysis. *International Journal of Electrical Power & Energy Systems*, 43(1):1435–1442, 2012.

[KW13a] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[KW13b] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LB12a] Hui Lin and Jeff A Bilmes. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*, 2012.

[LB12b] Hui Lin and Jeff A Bilmes. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871*, 2012.

[LBBH98a] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBBH98b] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LBOM12] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[LFV19] Christopher Leet, Debra A. Fischer, and Jeff A. Valenti. Toward a Self-calibrating, Empirical, Light-weight Model for Tellurics in High-resolution Spectra. , 157(5):187, May 2019.

[LGM88] S. Lowry and "A blot on the profession G. Macpherson. *British Medical Journal*, 296(6623):657–658, 1988.

[LGR+20] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, (29):4376–4389, 2020.

[LI16] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[LKWN21] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv:2106.12543*, 2021.

[LLL+23] Yue Liu, Zitu Liu, Shuang Li, Zhenyao Yu, Yike Guo, Qun Liu, and Guoyin Wang. Cloud-vae: Variational autoencoder with concepts embedded. *Pattern Recognition*, 140:109530, 2023.

[LLWT15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[LMCH21] Adam B Langeveld, Nikku Madhusudhan, Samuel HC Cabot, and Simon T Hodgkin. Assessing telluric correction methods for na detections with high-resolution exoplanet transmission spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 502(3):4392–4404, 2021.

[LRC05] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

[LRD+21] Cécile Logé, Emily Ross, David Yaw Amoah Dadey, Saahil Jain, Adriel Saporta, Andrew Y Ng, and Pranav Rajpurkar. Q-pain: A question answering dataset to measure social bias in pain management. *arXiv preprint arXiv:2108.01764*, 2021.

[LS99] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LT23] Shiye Lei and Dacheng Tao. A comprehensive survey to dataset distillation. *arXiv preprint arXiv:2301.05603*, 2023.

[LTG10] Herbert K. H. Lee, Matthew Taddy, and Genetha A. Gray. Selection of a representative sample. *Journal of Classification*, (27):41–53, 2010.

[Mac75] Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

[Mac05] David J.C. MacKay. *Informaiton Theory, Inference, and Learning Algorithms*. Cambridge University Press, 7 edition, 2005.

[MAR19] M. Bogen A. Korlova A. Mislove M. Ali, P. Sapienzynski and A. Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *ACM on Human-Computer Interaction*, 2019.

[MB88] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

[MBC+21] Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

[MCB+21] Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[MCR14] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[Mer22] MeriamWebster.com. Definition of representative sampling, 2022.

[MG60] Madan Lal Mehta and Michel Gaudin. On the density of eigenvalues of a random matrix. *Nuclear Physics*, 18:420–427, 1960.

[MH10] Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 172–177. IEEE, 2010.

[MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[MK17] Claus Adolf Moser and Graham Kalton. *Survey methods in social investigation*. Routledge, 2017.

[MK20] Luke Melas-Kyriazi. The mathematical foundations of manifold learning. *arXiv preprint arXiv:2011.01307*, 2020.

[MMS+21] Ninahreh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, (54), 2021.

[MNJ+21] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021.

[Mon19] Douglas C. Montgomery. *Design and Analysis of Experiments*. Wiley, 10th edition, 2019.

[Moo96] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[MS18] Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *KI-Künstliche Intelligenz*, 32:37–53, 2018.

[MSLA+11] A. Mislove, S. S. Lehmann, Y.-Y. Ahn, J. p. Onnela, and J. Rosenquist. Understanding the demographics of twitter users. *Proceedings of: Fifth International AAAI Conference on Weblogs and Social Media*, (5):554–557, 2011.

[MSSW18] Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. *Advances in Neural Information Processing Systems*, 31, 2018.

[MZP21] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, 2021.

[NG-21] Data-Centric AI 2021 neurips workshop. `https://datacentricai.org/neurips21/`, 2021.

[OCFK19] Alexandra Olteanu, CarlosCastillo, FernandoDiaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data*, (2), 2019.

[OEC22] OECD. May 26, 2022, glossary of statistical terms, stats.oecd.org, 2022.

[OF96] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[OGB+21] Karl Otness, Arvi Gjoka, Joan Bruna, Daniele Panozzo, Benjamin Peherstorfer, Teseo Schneider, and Denis Zorin. An extensible benchmark suite for learning to simulate physical systems. *arXiv preprint arXiv:2108.07799*, 2021.

[OYC18] Weihua Ou, Di Yuuan, and Yongfeng Cao. Object tracking based on online representative sample selection via noon-negative least square. *Multimed Tools Appl*, (77):10569–10587, 2018.

[Pat16a] Eileen Patten. Racial, gender wage gaps persist in us despite some progress. 2016.

[Pat16b] Mildred L Patten. *Understanding research methods: An overview of the essentials*. Routledge, 2016.

[PE95] Tim J Peters and Jenny I Eachus. Achieving equal probability of selection under various random sampling strategies. *Paediatric and perinatal epidemiology*, 9(2):219–224, 1995.

[Pea83] John A Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.

[PJN+11a] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

[PJN+11b] P Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11, 2011.

[PKDN15] Joseph Prusa, Taghi M Khoshgoftaar, David J Dittman, and Amri Napolitano. Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE international conference on information reuse and integration*, pages 197–202. IEEE, 2015.

[PM78] GY Pierre and Lucien Marin. Unbiased sampling from a falling stream of particulate material. *International Journal of Mineral Processing*, 5(3):297–315, 1978.

[PME05a] Lars Petersen, Pentti Minkkinen, and Kim H. Esbensen. Representative sampling for reliability data analysis: Theory of sampling. *Chemometrics and Intelligent Laboratory Systems*, (77):261–277, 2005.

[PME05b] Lars Petersen, Pentti Minkkinen, and Kim H Esbensen. Representative sampling for reliable data analysis: theory of sampling. *Chemometrics and intelligent laboratory systems*, 77(1-2):261–277, 2005.

[PSSU18] Burkni Palsson, Jakob Sigurdsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Hyperspectral unmixing using a neural network autoencoder. *IEEE Access*, 6:25646–25656, 2018.

[PSU21] Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. *arXiv preprint arXiv:2107.01091*, 2021.

[RB19] I. Raji and J. Buolamwini. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial ai products. *AAAI/ACM Conf. AI, Ethics, and Society*, 2019.

[RBC20] Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems*, 33:361–371, 2020.

[RBM+21] Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. An information retrieval approach to building datasets for hate speech detection. *arXiv preprint arXiv:2106.09775*, 2021.

[RD21] Boris Ruf and Marcin Detyniecki. Implementing fair regression in the real world. *arXiv preprint arXiv:2104.04353*, 2021.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[RG16] Frédéric Ros and Serge Guillaume. Dendis: A new density-based sampling for clustering algorithm. *Expert Systems with Applications*, 56:349–359, 2016.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[RKRD17] Julian A Ramos Rojas, Mary Beth Kery, Stephanie Rosenthal, and Anind Dey. Sampling techniques to improve big data exploration. In *2017 IEEE 7th symposium on large data analysis and visualization (LDAV)*, pages 26–35. IEEE, 2017.

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[Rob71] Harry V. Roberts. Committee selection by statistical sampling. *The American Statistician*, (25):18–20, Feb 1971.

[Ros10] Paul R. Rosenbaum. *Design of Observational Studies*. Springer, 2010.

[RP14] Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, (346):1063–1064, 2014.

[RRR+21] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10912–10922, 2021.

[Rut87] George Rutherglen. Disparate impact under title vii: an objective theory of discrimination. *Va. L. Rev.*, 73:1297, 1987.

[Sab16] Liz Sablich. 7 findings that illustrate racial disparities in education. *Brookings. edu.(Brookings (https://www. brookings. edu/blog/brown-center-chalkboard/2016/06/06/7-findings-that-illustrate-racial-disparities-in-education/))*, 2016.

[SATC11] Ben Somers, Gregory P Asner, Laurent Tits, and Pol Coppin. Endmember variability in spectral mixture analysis: A review. *Remote Sensing of Environment*, 115(7):1603–1616, 2011.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[SDG+14] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[Sed13] Philip Sedgwick. Convenience sampling. *Bmj*, 347, 2013.

[SG19a] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

[SG19b] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2:8, 2019.

[SG19c] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv:1901.10002v1*, 2019.

[Sha06] William T. Shaw. Sampling student's t distribution-use of the inverse cumulative distribution function. *Journal of Computational Finance*, (9):37, 2006.

[Sha17a] Gaganpreet Sharma. Pros and cons of different sampling techniques. *International journal of applied research*, 3(7):749–752, 2017.

[Sha17b] Gaganpreet Sharma. Pros and cons of different sampling techniques. *International journal of applied research*, 3(7):749–752, 2017.

[SHB+17] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[Sim49] Edward H Simpson. Measurement of diversity. *nature*, 163(4148):688–688, 1949.

[SLAJ22] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. A survey on techniques for identifying and resolving representation bias in data. *arXiv preprint arXiv:2203.11852*, 2022.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[SS03] Sarjinder Singh and Sarjinder Singh. Simple random sampling. *Advanced Sampling Theory with Applications: How Michael 'selected'Amy Volume I*, pages 71–136, 2003.

[SSN+15] Alain Smette, Hugues Sana, S Noll, H Horst, W Kausch, S Kimeswenger, M Barden, C Szyszka, AM Jones, A Gallenne, et al. Molecfit: A general tool for telluric absorption correction-i. method and application to eso instruments. *Astronomy & Astrophysics*, 576:A77, 2015.

[ST02] Vin Silva and Joshua Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in neural information processing systems*, 15:721–728, 2002.

[Sta22] StatisticsHowto.Com. Representative sample: Simple definition, examples, 2022.

[Suh05] Diana D Suhr. Principal component analysis vs. exploratory factor analysis. *SUGI 30 proceedings*, 203:230, 2005.

[SWXS18] Jiayu Sun, Xinzhou Wang, Naixue Xiong, and Jie Shao. Learning sparse representation with variational auto-encoder for anomaly detection. *IEEE Access*, 6:33353–33361, 2018.

[Tal07] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

[Tal20] Nassim Nicholas Taleb. *Statistical Consequences of Fat Tails*. STEM Academic Press, 2020.

[TE11] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. *CVPR*, pages 1521–1528, 2011.

[Tho12] Steven K Thompson. *Sampling*, volume 755. John Wiley & Sons, 2012.

[Tip99] Michael E Tipping. Deriving cluster analytic distance functions from gaussian mixture models. 1999.

[TMZ05] O. Tamuz, T. Mazeh, and S. Zucker. Correcting systematic effects in a large set of photometric light curves. , 356(4):1466–1470, February 2005.

[TSL00] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[Tuc66] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[Unb22] Andrew Ng: Unbiggen AI ieee spectrum. `https://spectrum.ieee.org/andrew-ng-data-centric-ai`, 2022.

[UPH+19] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *Proceedings of: IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1588–1597, 2019.

[uRLA+16] Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U Khan. Big data reduction methods: a survey. *Data Science and Engineering*, 1(4):265–284, 2016.

[Vas69] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[VCR03] William D Vacca, Michael C Cushing, and John T Rayner. A method of correcting near-infrared spectra for telluric absorption. *Publications of the Astronomical Society of the Pacific*, 115(805):389, 2003.

[vDBA+19] David van Dijk, Daniel B Burkhardt, Matthew Amodio, Alexander Tong, Guy Wolf, and Smita Krishnaswamy. Finding archetypal spaces using neural networks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2634–2643. IEEE, 2019.

[VDKDG21] Leonard Elia Van Dyck, Roland Kwitt, Sebastian Jochen Denzler, and Walter Roland Gruber. Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. *Frontiers in Neuroscience*, 15:750639, 2021.

[VdMH08] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[VF05] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems: 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Proceedings 8*, pages 758–770. Springer, 2005.

[VL20] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[VLG96] Charles F Van Loan and G Golub. Matrix computations (johns hopkins studies in mathematical sciences). *Matrix Computations*, 5, 1996.

[VLL+10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[WBJ+22] Guanchu Wang, Zaid Pervaiz Bhat, Zhimeng Jiang, Yi-Wei Chen, Daochen Zha, Alfredo Costilla Reyes, Afshin Niktash, Gorkem Ulkar, Erman Okman, Xuanting Cai, et al. Bed: A real-time object detection system for edge devices. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4994–4998, 2022.

[WHWW14] Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[WLP+22] Sharon Xuesong Wang, Natasha Latouf, Peter Plavchan, Bryson Cale, Cullen Blake, Étienne Artigau, Carey M Lisse, Jonathan Gagné, Jonathan Crass, and Angelle Tanner. Characterizing and mitigating the impact of telluric absorption in precise radial velocities. *The Astronomical Journal*, 164(5):211, 2022.

[WLWJ21] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 895–904, 2021.

[WLZ+16] H Wang, ZeZXeZBePJ Lei, X Zhang, B Zhou, and J Peng. Machine learning basics. *Deep learning*, pages 98–164, 2016.

[WNC05] Rebecca Willett, Robert Nowak, and Rui Castro. Faster rates in regression via active learning. *Advances in Neural Information Processing Systems*, 18:179–186, 2005.

[WSGG12] Jin-Feng Wang, A Stein, Bin-Bo Gao, and Yong Ge. A review of spatial sampling. *Spatial Statistics*, 2:1–14, 2012.

[WYC+19] Wei Wang, Dan Yang, Feiyu Chen, Yunsheng Pang, Sheng Huang, and Yongxin Ge. Clustering with orthogonal autoencoder. *IEEE Access*, 7:62421–62432, 2019.

[XGF16] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[XLZ+19] Xinzheng Xu, Tianming Liang, Jiong Zhu, Dong Zheng, and Tongfeng Sun. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 328:5–15, 2019.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[XWCL15] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[YCN+15] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[YIN+21] Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. Synthbio: A case study in human-ai collaborative curation of text datasets. *arXiv preprint arXiv:2111.06467*, 2021.

[YLW23] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023.

[YQFF+20] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. *FAT* '20, january 27-30*, 2020.

[YXZ+21] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021.

[YZL+06] Jun Yan, Benyu Zhang, Ning Liu, Shuicheng Yan, Qiansheng Cheng, Weiguo Fan, Qiang Yang, Wensi Xi, and Zheng Chen. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE transactions on Knowledge and Data Engineering*, 18(3):320–333, 2006.

[ZAZ+20] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2):56–70, 2020.

[ZB23] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.

[ZBL+23a] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. Data-centric ai: Perspectives and challenges. *arXiv preprint arXiv:2301.04819*, 2023.

[ZBL+23b] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

[ZHB+19] G. Zhou, C. X. Huang, G. Á. Bakos, J. D. Hartman, David W. Latham, S. N. Quinn, K. A. Collins, J. N. Winn, I. Wong,

G. Kovács, Z. Csubry, W. Bhatti, K. Penev, A. Bieryla, G. A. Esquerdo, P. Berlind, M. L. Calkins, M. de Val-Borro, R. W. Noyes, J. Lázár, I. Papp, P. Sári, T. Kovács, Lars A. Buchhave, T. Szklenar, B. Béky, M. C. Johnson, W. D. Cochran, A. Y. Kniazev, K. G. Stassun, B. J. Fulton, A. Shporer, N. Espinoza, D. Bayliss, M. Everett, S. B. Howell, C. Hellier, D. R. Anderson, A. Collier Cameron, R. G. West, D. J. A. Brown, N. Schanche, K. Barkaoui, F. Pozuelos, M. Gillon, E. Jehin, Z. Benkhaldoun, A. Daassou, G. Ricker, R. Vanderspek, S. Seager, J. M. Jenkins, Jack J. Lissauer, J. D. Armstrong, K. I. Collins, T. Gan, R. Hart, K. Horne, J. F. Kielkopf, L. D. Nielsen, T. Nishiumi, N. Narita, E. Palle, H. M. Relles, R. Sefako, T. G. Tan, M. Davies, Robert F. Goeke, N. Guerrero, K. Haworth, and S. Villanueva. Two New HATNet Hot Jupiters around A Stars and the First Glimpse at the Occurrence Rate of Hot Jupiters from TESS. , 158(4):141, October 2019.

[ZKL+10a] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[ZKL+10b] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[ZNB22] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022.

[ZOM19] C. Vogeli Z. Obermeyer, B. Powers and S. Mullainan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

[ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[ZY18] Tonglin Zhang and Baijian Yang. Dimension reduction for big data. *Statistics and Its Interface*, 11(2):295–306, 2018.

[ZZ19] Hengqian Zhao and Xuesheng Zhao. Nonlinear unmixing of minerals based on the log and continuum removal model. *European Journal of Remote Sensing*, 52(1):277–293, 2019.

[ZZS13] Zeya Zhang, Zhiheng Zhou, and Dongkai Shen. Sample selection method in supervised learning based on adaptive estimated threshold. In *2013 International Conference on Machine Learning and Cybernetics*, volume 4, pages 1861–1864. IEEE, 2013.