Small Data Explainer - The impact of small data methods in everyday life

Maren Hackenberg[1,2], Sophia G Connor[1], Fabian Kabus[1], June Brawner[3], Ella Markham[3,4], Mahi Hardalupas[3], Areeq Chowdhury[3], Rolf Backofen[5,6], Anna Köttgen[7], Angelika Rohde[8], Nadine Binder[2,9], Harald Binder[1,2,6], and the Collaborative Research Center 1597 Small Data

[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany
[2]Freiburg Center for Data Analysis, Modelling and AI, University of Freiburg, Freiburg, Germany
[3]The Royal Society, London, United Kingdom
[4]University of Edinburgh, Edinburgh, United Kingdom
[5]Department of Computer Science, Faculty of Engineering, University of Freiburg, Germany
[6]Centre for Integrative Biological Signaling Studies (CIBSS), University of Freiburg, Freiburg, Germany
[7]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg, Germany
[8]Department of Mathematical Stochastics, Faculty of Mathematics and Physics, University of Freiburg, Germany
[9]Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

Corresponding author: maren.hackenberg@uniklinik-freiburg.de

Abstract

The emergence of breakthrough artificial intelligence (AI) techniques has led to a renewed focus on how small data settings, i.e., settings with limited information, can benefit from such developments. This includes societal issues such as how best to include under-represented groups in data-driven policy and decision making, or the health benefits of assistive technologies such as wearables. We provide a conceptual overview, in particular contrasting small data with big data, and identify common themes from exemplary case studies and application areas. Potential solutions are described in a more detailed technical overview of current data analysis and modelling techniques, highlighting contributions from different disciplines, such as knowledge-driven modelling from statistics and data-driven modelling from computer science. By linking application settings, conceptual contributions and specific techniques, we highlight what is already feasible and suggest what an agenda for fully leveraging small data might look like.

1 Introduction

In recent years, there has been an increasing recognition of the potential of small data in technology development and decision-making processes, complementing and potentially even outpacing the trajectory seen for big data [1]–[5]. While there is no consensus on the definition of small data, small data approaches are typically needed in scenarios where there is a limited amount of information, such as the analysis of a small number of observations in a dataset. Whether a dataset is considered small also depends on the amount of heterogeneity between observations. For example, a clinical patient dataset with six patients might be considered small, while an experiment with six mice in biomedical research would not be considered small, due to large diversity in human beings as compared to animal models in biomedical research. A further important factor is the complexity of the questions to be explored with a dataset. For example, training large language models (LLMs) with machine learning techniques requires vast training datasets of millions or billions of documents. The training of a specialised LLM with only thousands of documents then is a small data challenge, while the use of simpler statistical modelling techniques with clinical patient data might only face a small data challenge when the number of observations is below a few dozen patients.

Insights from small data are particularly valuable in scenarios where large datasets are not available or where unique and specific data points can provide critical understanding [1], [6]. Unlike big data approaches, which rely on vast datasets to identify overall trends and patterns, small data derives insights from smaller and often more homogenous datasets, which might put the focus on specific, sometimes overlooked subgroups of a population and thus address diversity in a targeted way [1], [7]. Even when there are large datasets available, they may exclude under-represented groups, or small groups might not be represented in patterns extracted by big data techniques, reinforcing biases [8], [9]. Therefore, small data is increasingly seen as a fruitful and necessary alternative remedying the limitations of big data approaches [5].

Small data problems occur in a range of fields and hence, small data methodologies have also been developed across many different research areas (e.g., computer science, mathematics, statistics) [10]–[12]. While this shows the relevance of small data research in many domains, it also means that research may be impeded due to a lack of interdisciplinary communication [2], [13]. In particular, the lack of a shared language for small data approaches across disciplines is a major hurdle, in addition to technical challenges when using small data methods, such as overfitting and validation [14], [15]. To foster such a shared language, this explainer jointly addresses audiences from many different communities developing small data methods.

Specifically, we aim to raise awareness of, and illustrate, the unique challenges posed by small data, which originate from limited sample sizes and limited information in the face of complex data analysis tasks. This is relevant to a range of audiences, including: 1) policymakers and policy professionals who are interested in understanding how advancements in data science can inform policy decisions; 2) data scientists familiar with statistical methods, modelling techniques, and machine learning algorithms, who are interested in understanding small data techniques, challenges, and potential solutions; and 3) computer scientists, mathematicians, and statisticians developing algorithms, tools, and techniques for processing and analysing data.

This explainer was guided by three primary questions, answered through extensive desk review and evidence synthesis:

1. What are the fundamental principles underlying the concept of small data, and how do they differ from those of big data, particularly in the context of technology development and decision-making?
2. Why has there been a growing interest in leveraging small data as an alternative approach to big data, and what are the overarching implications for inclusive technology development and decision-making processes?
3. How can various methods and techniques for working with small data be adapted and optimised to address the unique challenges and opportunities present in the context of AI development and personalised applications?

To approach these questions, we look at small data from multiple perspectives. Section 2 focuses on the relevance of both big and small data for policy and decision making, and the core components of small data applications. We present an overview of big data versus small data and introduce some limitations of big data that may be addressed by using small data. In particular, we focus on how big data struggles to represent and deal with extreme values in a data distribution or subgroups that might need to be described by their own distribution. This means that under-represented and minority groups can often become invisible or ignored when the outputs from big data methods are used to inform decisions, which is especially concerning for high-stakes policy decisions. This section also includes case study examples of small data use (e.g., rare disease treatment and wearable assistive technology), and introduces the themes of similarity, transfer, and uncertainty for structuring small data applications and subsequent discussion of specific methods. It is targeted towards a broad audience with a specific focus on policymakers.

In Section 3, also targeted towards a broad audience, we highlight some exemplary application fields of small data. For some of these, some tailored small data approaches have already been developed, such as for rare disease settings or precision medicine, which could serve as templates for other application fields, where small data approaches are still lacking.

In Section 4, we describe small data methodologies in more detail (e.g., approaches for incorporating similarity, foundation models, few-shot learning, meta learning, and the combination of knowledge-driven and data-driven modelling). Many of these methodologies were developed by different disciplines and fields, which may potentially hinder awareness and adoption of methods by different research groups. To address this, we highlight the contributions of different disciplines and explore potential links. We also present some common challenges in using small data, such as overfitting and validation. This section therefore takes a more technical approach and is tailored for an interdisciplinary data science audience.

We conclude with some potential opportunities and propose an agenda for increasing the use of small data in in Section 5.

2 Small data and policy

2.1 The significance of big data for research and decision making

The ability to process vast amounts of information quickly and efficiently has made possible the identification of broad trends, prediction of future behaviours, and optimisation of complex systems. These discoveries and innovations are fuelled by so-called big data. The term big data initially referred to datasets so vast they required specialised computing solutions for analysis [16], yet such advanced computing power nowadays is easily available. Despite the name, sheer size is not what defines big data. In simple terms, big data is more accurately characterised by the ability to search, aggregate, and cross-reference extensive datasets, with potential challenges due to a large variety of data types and data quality issues [8]. However, from a sociotechnical lens, big data rests at the intersection of 1) technology (maximising the power of computation and algorithms to gather and analyse large datasets); 2) analysis (identifying patterns within large datasets to make claims about processes); and 3) mythology (the widespread trust in large datasets to offer the possibility of truth, objectivity or accuracy previously not possible) [17].

The contemporary emphasis on big data began in the early 2000s, coinciding with advancements in computing power and the large varied quantities of data generated and collected by digital technologies, including those connected through the internet. Global technology companies began to harness big data to transform their business models, creating highly personalised user experiences [18], [19]. The success of these enterprises underscored the value of big data, leading to the widespread adoption of big data practices across various sectors, including healthcare, finance and public policy [20].

The corresponding increased awareness of the opportunities of data and focus on gathering data for better understanding structures and processes has led to significant advancements, regardless of whether specific applications are big data or not. For example, data availability in healthcare has enabled breakthroughs in precision medicine [21] and predictive maintenance in manufacturing [22]. More than this, big data created “a radical shift in how we think about research […] Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality”[17].

A clinician specialising in Gaucher disease, a rare genetic disorder that leads to the buildup of glucocerebroside in various organs [23] due to the deficiency of a specific enzyme, wants to assess whether a patient is responding well to enzyme replacement therapy (ERT). It is often challenging for the clinician to find the correct dose of ERT, as the disease presents with diverse manifestations and many different factors need to be considered [24]. In order to address these issues, the clinician might have access to a clinical registry of patients with this medical condition, which contains information as to their ERT doses. Despite the likely small number of observations, which is to be expected for a rare disease registry, a machine learning approach could be used to deliver dose predictions, ideally leveraging all available data, leading to a small data challenge.

When the clinician has a new patient, for example a 6-year-old child trialling ERT again after a pause, the machine learning approach should take into account how similar the new patient is to the patients already present within the clinical registry. Often, there will be considerable heterogeneity across patients in a registry, which means that the patients are dissimilar to each other. Therefore, the new patient might need to be matched to the most similar subgroup to allow for modelling and prediction based on this subgroup, while still taking some information from the other subgroups into account. In this case, for example, there might only be a limited subgroup of patients under the age of 10. For these reasons, it can be difficult to match the current patient to previous patients in order to make accurate predictions. When the new patient cannot be appropriately matched, transfer techniques are required to still enable prediction modelling by at least using some amount of information, e.g., by extrapolating across age groups.

Box 1: Hypothetical exemplary scenario from the field of rare disease small data applications.

2.2 Limitations of big data and the problem of the 'average man'

One key limitation of big data approaches is data availability: in many contexts, large datasets are simply not available, especially in specialised fields like rare diseases (see Box 1 for an exemplary scenario, and Section 3.1 for more details) or niche markets. Small data approaches, on the other hand, thrive on detailed, context-specific information from smaller datasets, making them invaluable where data is scarce.

A second key limitation of big data approaches is the mythology surrounding them, in particular, the assumption that bigger datasets lead to more reliable conclusions. The belief that 'bigger is better', where traditional approaches in business, governance, and technology prioritise large numbers, goes hand in hand with advancements in computing and AI, which excel at processing vast amounts of information but often falter when confronted with extreme values (Figure 1, Panel A) and unique scenarios (Figure 1, Panel B). For instance, AI models for self-driving cars have shown critical failures when confronted with unexpected situations, such as encountering a backward-propelled wheelchair, even when trained extensively on wheelchair-representative datasets [25].

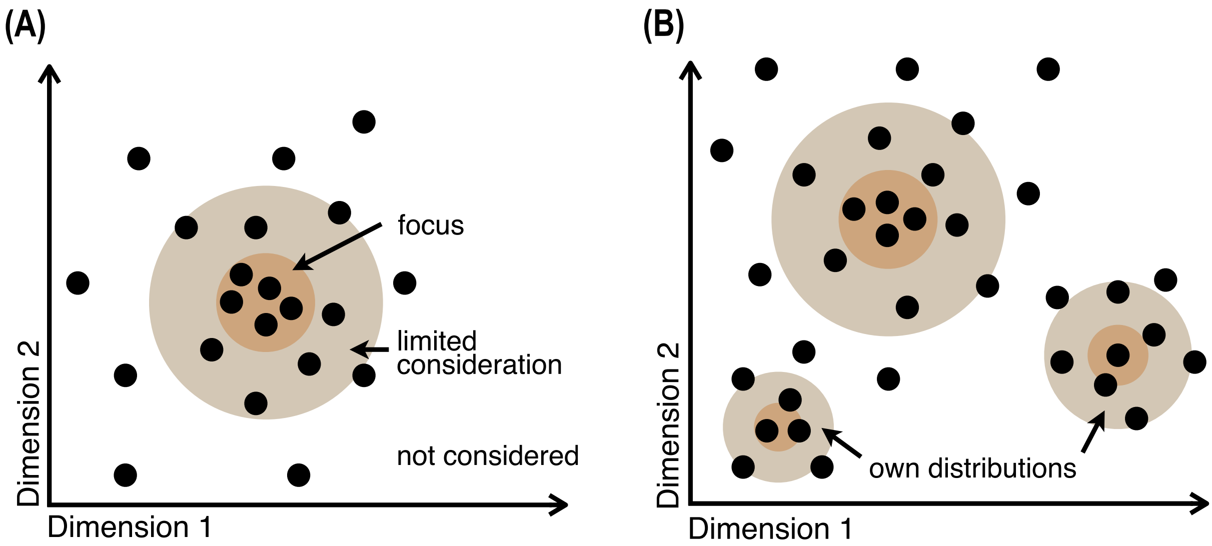

Figure 1: Conceptual illustration of small data challenges with hypothetical data where individuals are described by two dimensions, such as when considering factors “age” and “years of experience with digital technologies” when designing a chatbot system based on large language models (LLMs). (A) Small data conceptualised as the extreme values of a distribution, e.g., a Gaussian distribution, corresponding to the concept of Quetelet [26]. In this figure, the individuals in the “Focus” area are within the average, whereas individuals that deviate from this average are considered less often or not at all. (B) Conceptualisation of small data via subgroups of individuals which might be described by their own distribution, e.g., not treating children as small adults, or analysing the specific needs of elderly users.

This highlights how big data approaches, which can be useful for identifying general trends within large datasets, might obscure the nuances and specificity present in smaller datasets, which might cover a more diverse range of scenarios. Consequently, individuals and scenarios that do not fit the majority pattern are marginalised, leading to significant knowledge gaps and inadequately addressed needs. When these approaches are used for policy decisions, this can lead to ineffective, inefficient and inequitable decision-making.

The necessity of considering alternatives to big data approaches becomes even more apparent when knowledge of the historical context of statistical methodologies is applied to human populations and social phenomena. Specifically, the concept of the ‘average man’ was introduced in the 19th century by Belgian astronomer and statistician Adolphe Quetelet FRS. Quetelet’s work laid the foundation for modern statistics and the field of social physics, which applies statistical methods to social phenomena. In his pivotal work, Quetelet proposed the idea of the ‘average man’ (‘l’homme moyen’), a hypothetical individual whose characteristics represent the average values of a population’s various traits. He applied the Gaussian distribution, which had been used primarily in astronomy and other physical sciences, to human traits. In doing so, he suggested that most people cluster around a central, average value, with fewer individuals exhibiting extreme values at the margins.

Quetelet's concept had profound implications for modern statistics and social science. First, it introduced the idea that human characteristics, such as height, weight, and even behaviours, could be quantified and analysed statistically. Second, Quetelet asserted that the average value within a distribution (the l'homme moyen) should be the primary focus when studying human attributes [26]. This idea laid the groundwork for the development of fields like demography, public health, and social statistics. However, it also led to an overemphasis on the average, sometimes at the expense of understanding the diversity and complexity of human population, neglecting those who fall outside the norm. This is illustrated in Figure 1, Panel A.

As an alternative conceptualisation, small groups might not be considered as the extreme margins of a Gaussian distribution, but as having their own, separate distribution (1, Panel B). This implies the need to reason about these distributions, which is made difficult by limited availability of data. A solution is to assess to what extent information from the majority distribution or other subgroups can be adapted.

The implications of Quetelet's work are still evident today with the reliance on averages to use big data in decision making and policy making. However, it is important to remember that "…people are individual people and not an average" [27]. While averages can provide useful summaries, they also mask important variations and lead to decisions that do not adequately address the necessities of all individuals, particularly those with less common traits or needs. This has led to a growing recognition of the importance of small data and personalised approaches that can capture the unique and diverse experiences of individuals, thereby fostering more inclusive and effective political and technological solutions, ultimately benefiting the entire population.

An example of this is closed captioning [28], which was initially developed for the deaf and hard of hearing, however is now widely used in noisy environments like gyms and bars, and by individuals learning new languages or trying to follow along without sound. Similarly, automatic doors [29], Text-to-speech (TTS) and Speech-to-text (STT) systems [30], voice assistants [31] like Amazon's Alexa and Apple's Siri, as well as smartphone accessibility features [32] were also originally designed to assist disabled people but have developed a much wider scope given their convenience and ease of use in various situations.

2.3 The rise of small data

In an era where big data dominates decision-making, small data, which focuses on detailed, context-specific information from smaller datasets, offers unique and critical insights.

Small data is sometimes referred to as the 'original data': early-modern scientific discoveries, from those of Darwin's evolutionary theory to Einstein's fundamental laws, predate big data approaches and contemporary computational abilities. Humans are also naturally small data learners: given a single image of a car, children can generalise the concept and recognise similar objects. This is an impressive example of one-shot learning, i.e., learning a class from a

Glossary: ***Foundation models*** are large-scale machine learning models trained on vast amounts of diverse data, allowing them to serve as a general-purpose base for a wide range of downstream tasks. These models, such as large language models (LLMs) and multimodal AI systems, acquire broad representations that can be adapted to more specific applications with limited available data through fine-tuning or in-context learning. These capabilites make them a powerful tool for both big and small data settings.

single labelled example, which is not straightforward to obtain with classical machine learning approaches [33]. Therefore, it might be attractive to utilise small data ideas also in big data contexts. Certain methodologies already have integrated small data with big data to transfer information between the big dataset and the small dataset. For example, foundation models now allow the transfer of information from a huge body of text, images, or other data to settings with limited data [34] (see Box 2 for an exemplary scenario).

Consequently, small data is (re-)emerging as a crucial complement to big data, in part because of its relevance to advances in technologies like AI. This approach is particularly important in areas where context-specific insights are critical, such as in rare disease research (see Box 1), and personalised marketing strategies (see Box 3).

Suppose a policymaker in a large tech company aims to improve the security and robustness of the company's software products against cybersecurity threats, e.g., in accordance with a report recently published by the White House Office of the National Cyber Director [35]. To achieve this, they focus on addressing known classes of vulnerabilities through use of secure software development practices. The most commonly used programming language for the company's products is C++, which is also one of the languages that lacks memory safety [36]. In addition, the policymaker knows that it is common practice among developers in their company to use LLM-based products for assistance when generating code. While recent versions of C++ allow for memory-safe programming, the C++ code used to train LLMs is likely to span several decades, i.e., represent old non-safe practices with potential memory safety bugs [37]. Thus, the challenge would be to use limited amounts of memory-safe examples of code available to generate via the LLM further memory-safe code for the requested task. Ideally, there would be a similarity metric to assess different LLMs and determine which can best provide this code, in order to select the best solution. Alternatively, LLMs could be fine-tuned to generate safer code. Users should at least be provided with a warning if the generated code deviates from the safety standards of the company, or the model might even be able to more precisely indicate how confident it is that its output is indeed memory-safe.

Box 2: Hypothetical exemplary scenario from the field of generative AI and foundation models.

Suppose a finance tech start up ("FinTech") is developing an app that helps its users make informed financial decisions by providing them personalised recommendations. As consumer trust is an important factor for the adoption of the product [38], [39], they focus on a strong adherence to user privacy, in particular, by storing information only on the device of the user, e.g., a smartphone. In addition, the recommendation algorithm is focused on data minimisation, by asking the user only for absolutely necessary information and storing only aggregate data where possible, e.g., only the average proportion of income spent on rent or eating out. This limits the privacy impact, e.g., in case of a data breach. To enable precise predictions as a basis for recommendations, missing information that is still required might be extrapolated from similar users, e.g., from a database of users who agreed to such use of their data. For example, the prediction algorithm might determine how the individual fits different archetypes, i.e., a student who likes to travel or a married individual with a stable income and children, using similarity measures. As banks continue to open up their data to third parties [40], the app should give specific recommendations on retirement planning by transferring the generic recommendation model onto the specific spending habits of the user provided by the bank. As financial advice, especially concerning investments, involves uncertainty, which might be increased by data minimisation, the app should include uncertainty evaluations to help the individuals make informed decisions about the risks involved, e.g., when deciding for a saving strategy. The app might even indicate what additional information would be needed to reduce uncertainty.

Box 3: Hypothetical exemplary scenario from the field of finances and trusted personalised recommendations.

The growing interest in small data is also driven by advancements in data collection and analysis technologies that make it easier to gather and interpret more detailed information from smaller samples. For example, wearable devices, mobile sensors, and specialised software enable the continuous collection of granular data, data that is very detailed and specific, providing a rich source of insights that were previously inaccessible [41] (see Box 4).

As another example, mixed-methods research combines quantitative data from large samples with qualitative insights from smaller groups to provide a more comprehensive understanding of complex phenomena. This integrative approach is particularly valuable in fields like social sciences, where understanding human behaviour requires analysis of both broad patterns and individual experiences. For a primer on these themes see [42]. Small data techniques are also relevant where limited, granular datasets are collected by individual researchers during day-to-day experiments, as these datasets may be richly diverse, covering a range of heterogeneous scenarios, and relatively small, so-called long-tail data [6]. Long-tail data is valued for its diversity and specificity, which can provide critical insights that large, aggregated datasets might overlook.

Furthermore, small data approaches can help foster innovation and resilience in technologies,

Suppose a healthcare provider aims to decrease the number of hospital admissions resulting from recurrent falls among patients. To do this, they give wearable devices with accelerometers to individuals [43]. The detection of falls needs to be performed on the device with the limited data available there. After a period of time, there is enough accelerometer data available for the individual, such as how they normally move, so that personalisation and thus improved fall detection might be feasible. If the provider has a database of accelerometer data generated from use of this wearable device, the data from the individual could be mapped to subgroups of other individuals, or alternatively, a prediction model could be used that is aware of subgroups and has already been trained on the database. Ideally, this analysis in the context of similar data should be performed without transferring the data of individuals to a central server, to protect privacy.
However, fall detection using accelerometer data is challenging due to significant variability in how movements are recorded, as well as differences in individual anatomy [44], [45]. This variability might introduce considerable uncertainty in similarity assessment and predictions. Therefore, approaches for quantifying uncertainty in all steps are needed. In addition, some of the patients might not like the wearable device, but want to use the accelerometer on their phones instead [46]. This means that information transfer is required to link this new data type to the original datasets and models.

Box 4: Hypothetical exemplary scenario from the field of assistive technologies.

as they specifically acknowledge subgroups, potentially in a wide range of scenarios, making them more robust in the face of dynamic environments. Thus, personalised and context-specific technology solutions will rely on the utilisation and inclusion of small data.

2.4 Common themes: similarity, transfer, and uncertainty

For making small data challenges manageable, it can be useful to identify sub-tasks that appear in many small data settings. In the examples given above (in particular Boxes 1-4), there were three themes that repeatedly emerged: similarity, transfer, and uncertainty. These are illustrated in Figure 2, and are briefly introduced in the following. A detailed description of the corresponding technical approaches is provided in Section 4.

2.4.1 Similarity

Similarity is relevant when comparing different datasets for potential subsequent integration. Integrating datasets increases the number of observations available for modelling, and involves assessing the similarity of individuals to other individuals or groups of individuals, such as in rare disease applications, to leverage evidence from similar individuals. In other settings, similarity of datasets that include several observational units needs to be assessed, e.g., when deciding whether to combine datasets for modelling. As similarity will typically be assessed quantitatively, modelling techniques that are subsequently used also need to incorporate such quantitative information on similarity.

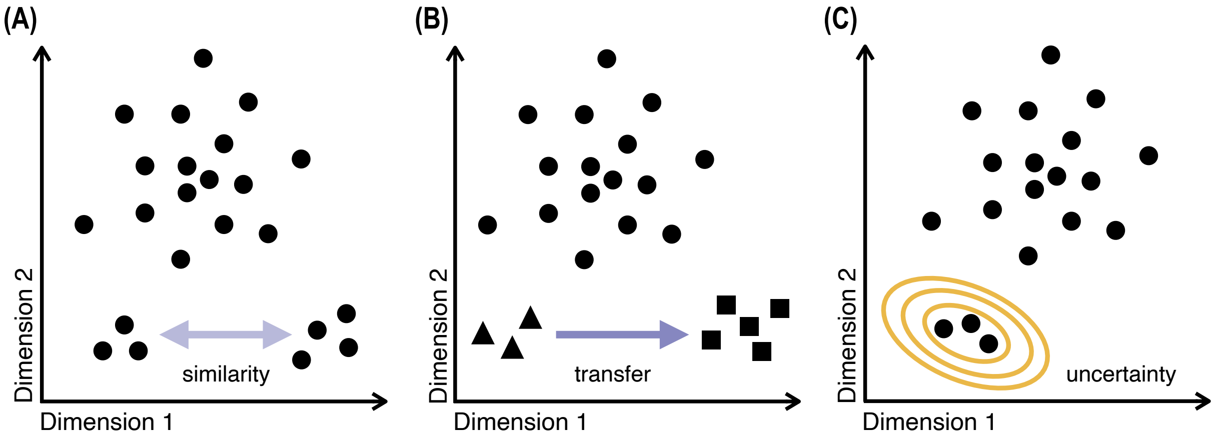

Figure 2: Illustration of common themes, namely similarity, transfer, and uncertainty, when addressing small data challenges. The example uses hypothetical data where individuals are described by two dimensions, such as when considering factors "age" and "years of experience with digital technologies" when designing a chatbot system based on LLMs. (A) Assessing the similarity of groups for potentially incorporating information, e.g., deciding whether to jointly analyse data from children and older adults when considering inclusive chatbot design. (B) More general information transfer, when data/information cannot be directly combined, e.g., leveraging child-centered user experience described in the scientific literature when designing a chatbot for older adults. (C) Quantifying uncertainty, to assess whether there is enough data for obtaining reliable evidence, e.g., deciding whether there is enough information to reasonably tailor a chatbot for children.

When there is already a big dataset or (pre-)trained model, e.g., a foundation model, approaches for similarity can be useful for assessing whether a dataset at hand is close enough to the typical settings represented by the big dataset or model to warrant use. Subsequent adaptation, when there is sufficient similarity, might then include personalisation of the model, e.g., by identifying subgroups represented by the large model that are the most similar to the dataset at hand. This corresponds to acknowledging the heterogeneity of big datasets and potentially reducing them into interlinked small datasets.

2.4.2 Transfer

The combination of similar datasets already implicitly entails a transfer of information from one dataset to the other. Yet, we understand transfer in a broader sense, relating not only to the transfer of information between similar datasets, but also to a more general information transfer from external sources to enrich a small dataset. In particular, information from other sources to be incorporated can include, e.g., other types of data (other data forms, databases, etc., see Figure 2, Panel B) or pre-trained models (e.g., LLMs), reflecting implicit knowledge, to facilitate modelling with small data by additional prior knowledge. This requires methods for explicitly transferring information. Use of methods for transfer ideally should include some quantification of similarity between the different types of information sources to assess whether information transfer indeed is beneficial, or whether it instead is detrimental, e.g., when LLMs override the information in the dataset at hand that is used as context.

In some settings, knowledge of the domain at hand might not be available in formats ready for information transfer, such as mechanistic knowledge or heuristic knowledge of experts. A formalisation step is then required, before information transfer can be performed. For example, knowledge on relations between quantities of interest could be expressed in the form of mathematical equations, even if it is not possible to fully and precisely specify these equations. Methods for information transfer then need to incorporate the uncertainty in specification.

2.4.3 Uncertainty

Due to limited information, uncertainty is a particularly pressing issue to be considered in small data settings. While uncertainty has been conventionally considered in the context of variability in parameter estimation in statistical models, such as when estimating treatment effects, the term can also be understood in a broader sense. In particular, there are many other kinds of uncertainties, including general sources of uncertainty in modelling, such as the choice of appropriate models [47]. Some sources of uncertainty are more specific to small data, particularly uncertainty arising from quantification of similarity and transfer of information, e.g., in the formalisation of expert knowledge for inclusion into modelling.

Where uncertainty cannot be avoided, comprehensive quantification of the different sources of uncertainty is needed. This enables informed use of the resulting models and predictions. Such quantification can also enable deliberate acceptance of uncertainty, e.g., when making informed decisions on data minimisation for a trade-off between uncertainty and privacy (see Box 3 for an example).

2.5 Implications of small data for evidence-informed policy-making

The policy implications of small data approaches are significant. During times of financial constraint, decision-makers often rely on data-supported strategies to set spending priorities and justify investments in public funds. However, this reliance can marginalise individuals who do not fit within the majority, such as those with disabilities or exceptional access requirements [48].

Likewise, dominant big data research methods prioritise large groups and population averages, potentially disregarding extreme values and reinforcing the ‘invisibility’ of marginalised groups when it comes to decision making. For example, research into assistive technologies often lacks generalisability due to its small, varied sample sizes [49]. Moreover, current big data tools are not designed to highlight patterns or trends specific to underrepresented groups, instead amplifying the averages and leaving out critical insights from the margins.

To address these issues, it is essential to recognise that including the margins in research and design processes can benefit everyone. Moving away from big data approaches towards personalised learning and informed, self-aware learners can help create a more equitable and responsive educational system. Embracing small data and dynamic instability in policymaking can lead to better outcomes for both individuals and society as a whole.

3 Exemplary application fields of small data

3.1 Rare diseases and N-of-1 studies

Small data methods are crucial for studying rare diseases, i.e., diseases afflicting less than one in 2,000 persons [50], as this is a prime example for medical conditions with limited data (see Box 1 for a hypothetical example). Correspondingly, this area has produced numerous methodologies specifically designed for small datasets. Still, it is a challenge to adapt machine learning, in particular deep learning techniques, which typically require more data [51]. Nevertheless, such methodological developments can be used as a blueprint for other small data settings. It is important to note that many of these techniques are not explicitly labelled as small data methods, even though they apply to these scenarios, reflecting the challenge of the lack of shared terminology around small data methods.

N-of-1 studies are a particular type of small data study which produce data limited by design, as they focus on collecting and analysing data from a single participant [52]. This is often performed in order to determine how effectively the participant responds to a specific intervention, and may involve the need to combine multiple different types of data from the individual, which requires transfer methodologies. N-of-1 studies are better suited for chronic conditions rather than rapidly progressive diseases, and commonly deployed to assess neurological or musculoskeletal disease [53]. This is because chronic conditions are often relatively stable or slowly progressive, and require long term management, which enables multiple crossover periods to assess efficacy of various treatments on the same patient.

Small data methods can assess medication efficacy, as well as the efficacy from surgical procedures, behavioural modifications, or medical device assessments [54]. These methodologies can then be applied to similar study designs where only a single participant is assessed. When aggregating many N-of-1 studies, the combined data could, e.g., be used for future predictions [55]. Yet, to do this, one must assess how similar the datasets are and develop methods to combine datasets when they are not in exactly the same format, i.e., tailored approaches are needed for information transfer.

3.2 Precision medicine and case-based reasoning

Personalisation of models, predictions and decisions in small data settings aims to tailor particular outcomes to the unique characteristics of a single observational unit or a small group. In precision or personalised medicine, this involves adapting specific therapies to suit each individual, accurately diagnosing diseases, or predicting individual health outcomes by taking into account unique genetic, lifestyle or environmental factors [48]. Effectively, this means that even a rather large, potentially multi-modal dataset might be reduced to many small subsets when looking for homogeneous groups of patients that could receive a personalised prediction. Therefore, much like in rare disease settings, many methodologies developed in the field of precision medicine are useful for small data settings, though they are often not explicitly identified as such. Often, methodologies initially developed for specific diseases are also applicable to other small data scenarios once researchers recognise similar underlying features.

Oncology has consistently led to advancements in precision medicine, largely because of the genetic basis of many cancers and the rapid progression of genomic technologies, e.g., allowing for more reliable and affordable cancer diagnosis [49], [56]. These advances enable the comprehensive screening of an individual's genome in just days, facilitating the development of tailored diagnostic, prognostic and therapeutic strategies. Big data analysis is effective at identifying patterns across large datasets, but small data methods are much more adept at tasks such as detecting rare mutations. Small data methodologies can provide assistance, for example, in analysing data from an individual's own cells to predict how they will respond to treatment, through personalised in vivo and in vitro cancer models [57]. However, this raises many small data challenges, such as looking for similar data to increase the dataset size, transferring information from other sources, and assessing the uncertainty, i.e., quantifying the reliability of predictions from such cancer models.

The paradigm of case-based reasoning (CBR) considers an extremely personalised setting where the data from the patient at hand, i.e., a single observational unit, is used as a starting point for reasoning. CBR then offers a modelling methodology for using previous similar examples of a problem and its solutions, each called a "case", to offer recommendations for a current similar issue [58]. This recommendation is then stored for use with future cases, allowing for iterative learning [59]. As for other settings in precision medicine, similarity assessment is therefore crucial in CBR because the accuracy of the solution depends on the accuracy of the evaluation of similarity between the new case and stored cases [60]. Furthermore, a machine learning approach must be able to search for similar cases effectively and time efficiently [58]. Uncertainty needs to be taken into account, as the available information from previous cases may be incomplete, imprecise, or subjective [61], [62]. It is also beneficial to develop transfer learning techniques that allow recognition of useful knowledge from other systems, and support their integration into the CBR model [63].

3.3 Assistive technologies, quantified self, and wearable health devices

The robust pace of advancements in energy-efficient computing and sensor miniaturisation [64] has enabled wearable devices to become increasingly more capable as assistive technologies. Continuous use of wearable health devices throughout the day can facilitate novel data-driven paradigms in healthcare and personal health management by providing real-time, comprehensive data from sensors to enable personalised monitoring, e.g., fall detection (see Box 4), early detection of health issues, and tailored interventions [65]. In particular, this technology has led to a rapidly expanding "quantified self" movement, where individuals use self-tracking to improve their health and well-being [66]. The corresponding data, although potentially longitudinally extensive, is similar to rare disease situations, as it pertains to a single individual. In particular, the devices often do not transfer all recorded data to centralised servers for combined analysis with data from other individuals. Instead, the recorded data primarily stays on-device and correspondingly analysis is performed there, e.g., in an edge computing solution [67]. If the device at least intermittently has access to external servers, these can be leveraged for additional information, e.g., additional data and/or trained models hosted by the manufacturer. However, it is still desirable to analyse the small data on-device, and use this external information as additional context to improve predictions made from the small dataset [68], as these predictions should still ideally be personalised to the user. This means that some matching to subgroups represented by the external data is needed, including assessment of similarity and potential approaches for information transfer. Much like above, this setting is similar to rare disease applications, yet with the additional complication of having the different sources of information in different locations. On the device itself, the model based on the larger amount of information can then be adapted to the user, to reduce the prediction uncertainty [69].

3.4 Data minimisation

Data minimisation, i.e., the collecting and processing of only the data required for a specific purpose, is a crucial concept that must be taken into account whenever dealing with data from individuals. Data minimisation is governed by regulations such as the EU General Data Protection Regulation (GDPR) [70], which also underscores the necessity of informed consent for collecting data and mandates that models be designed to provide explanations for their decision-making processes when required [71]. Small data techniques support data minimisation as they can potentially still achieve acceptable performance with limited data. This allows for deliberate collection of less data, reducing unnecessary privacy intrusions (see Box 3 for a hypothetical example). The data of individuals also might not need to be copied to central servers, e.g., as indicated for applications with wearable devices. Together, this lowers the risk of unintended data leaks [72], [73]. This is helpful in a wide range of scenarios, from medicine to finance, where data minimisation should be encouraged. Minimisation, as enabled by small data methods, might even be beneficial for other reasons, as collection and analysis of smaller data is often cheaper and can be performed with higher quality [1]. Yet, deliberate minimisation requires careful assessment of the resulting uncertainty in order to perform an informed trade-off. Also, filling in the gaps left by deliberate minimisation by incorporating similar data that is already available might require development of targeted techniques for information transfer, if there is only limited similarity between the data to be incorporated and the minimised data.

3.5 Generative AI

LLMs, such as that behind ChatGPT, are an example of foundation models, the latest generation of generative AI approaches which provide general representations learned from large-scale

Glossary: ***In-context learning*** allows large language models (LLMs), and more generally foundation models, to perform new tasks without retraining by providing task-specific context information, e.g., prompt text that includes examples, as part of the input during inference. Unlike traditional learning methods that require extensive fine-tuning, in-context learning enables LLMs to adapt dynamically to new problems by interpreting the provided context. For example, when tasked with translating a sentence or generating a summary, the model can infer an appropriate approach based on the information contained in the query itself. This capability has significant implications for rapid prototyping and adaptability in real-world applications.

datasets [34]. Correspondingly, they are adaptable for a wide range of creative tasks, such as generating artwork [74], [75] or music [76], [77], as well as practical tasks such as data augmentation, text generation, and image synthesis [78]–[80]. They can also be applied to fields such as biology or medicine [81]–[83]. Once tailored to a specific scenario, such foundation models can be used to simulate and explore hypothetical situations, even with limited real data available. This ability to generate synthetic data based on a learned structure, combined with expert insights or established knowledge from a specific field, is particularly attractive in small data settings, as it allows for the conversion of additional knowledge into data.

Yet small data problems still arise, as even large-scale datasets have areas that are hardly covered, e.g., as illustrated in 2. When an LLM is used to generate text pertaining to a part of the underlying statistical distribution that is only supported by limited data, it might also have a tendency to hallucinate, which means it fabricates a plausible but incorrect answer [84]. This setting, where the text used to prompt the LLM corresponds to the small data at hand, while the LLM represents implicit knowledge from a large-scale dataset, is an example of a more general class of small data settings where a small dataset is used as context information for in-context learning with foundation models [85]. Therefore, the challenge is to assess how close the small data at hand is to the implicit knowledge represented by the foundation model to detect potential issues with information transfer, such as hallucination in LLMs. Alternatively, foundation models can be fine-tuned to the small data at hand by adapting the model parameters. However, this might be challenging with limited data. Even if such an adaptation is not feasible or straightforward, a model should at least be adapted to warn the user when the provided response is based on an underrepresented knowledge area. The importance of acknowledging and communicating underlying uncertainties in models should be considered when formulating policies that rely on predictions or insights from foundation models, ensuring there is clear communication from developers on uncertainties and limitations in the model predictions and that policies account for potential biases, gaps in knowledge and model limitations.

Glossary: ***Few-shot learning*** refers to a method in AI where models are trained to perform tasks using only a small number of labeled examples. This technique is especially useful in scenarios where gathering extensive labeled datasets is impractical, such as rare medical conditions or highly specific tasks in industrial applications. By leveraging knowledge from previously seen tasks or datasets, few-shot learning enables models to generalize effectively, bridging the gap between data-rich and data-scarce environments.

4 Methods for working with small data

4.1 Contributions of different disciplines

Small data problems occur in a broad range of diverse application fields, as illustrated by the examples of small data presented in Box 1-4 and in Section 3. This diversity in applications indicates an increasing recognition of the need to develop specific small data approaches that can be used across different disciplines. Currently, since different small data application scenarios are typically addressed by different disciplines, such as computer science, mathematics, or statistics, there is high diversity in methodology developed to address small data challenges. This means that small data research is currently scattered across disciplines, impeding the realisation of its full potential.

A particular challenge is the limited exchange of methods between different data science disciplines. For example, mathematics and statistics have been traditionally seen as the disciplines for small data challenges, as research there has historically been performed in the context of applications with strong assumptions, which are used to gain mechanistic insight and to compensate for typically small datasets [1]. Take, for instance, infectious disease modelling, where assumptions regarding disease transmission and recovery rates are employed to predict the spread of diseases like influenza or COVID-19 when real-time data is scarce [86]. Similarly, in tumour growth modelling, assumptions about the growth rates and limitations of resources within the tumour are used to estimate the progression of cancer and optimise treatments with limited data [87].

Yet, the computer science community has also recently developed important small data techniques. These approaches are typically based on AI, and known under labels such as transfer learning, few-shot learning or meta-learning, but are only slowly finding their way into other data science communities [88]. In particular, many of the recent innovations in generative AI have considerable potential for small data challenges, but are currently scarcely discussed within this context. Conversely, even though the benefits of incorporating assumptions and domain knowledge into AI approaches are increasingly recognised (e.g., [89]–[92]), the modelling expertise in the statistics and mathematics community is only slowly being incorporated in current AI approaches developed in computer science.

Glossary: ***Meta-learning***, often referred to as "learning to learn", trains models to adapt to new tasks using prior knowledge from multiple datasets. Instead of building a single model optimized for a specific task, meta learning equips the model with a generalized capability to handle diverse problems. This approach is particularly relevant in small data contexts, where the ability to leverage knowledge from related domains can significantly enhance performance.

Small data applications typically also require strong domain knowledge, but domain experts often do not have access to state-of-the-art small data methods. Even though such connections between disciplines are increasingly recognised and needed, it also becomes apparent that they are often obscured by the lack of a common language [14]. This is particularly relevant with respect to small data methods, as none of the disciplines have developed a comprehensive small data framework. Clarification and mutual understanding of relevant terminology could remove hurdles and be a first step towards such a shared framework and language across disciplines.

To achieve this, it is essential to reflect on the contributions of the different disciplines to current small data methods. Firstly, we will present relevant small data approaches used in different communities, including corresponding terminology. We begin by giving a brief overview of the contributions of disciplines and methods developed there to the small data themes similarity, transfer and uncertainty outlined in Section 2.4, before going into more detail on selected specific classes of methods and approaches.

4.2 Similarity

Quantifying and incorporating similarity into modelling has traditionally been dealt with extensively in the statistics community. For example, approaches for matching [93] and re-weighting individuals [94], [95], allow researchers to make different groups more comparable, or to integrate information from other groups when focusing on a specific subgroup [96], [97]. The concept of localised regression [98] allows researchers to make models, such as regression models, specific to individuals, and can be extended to more flexible nonparametric approaches [99]. Further, approaches for quantifying the similarity of fitted models [100], [101] can be leveraged for assessing the similarity of different groups, e.g., in models for clinical time-to-event data of rather small patient groups [102], [103]. Dissimilarity of groups of individuals is also considered in meta-analysis techniques, developed in clinical biostatistics, which allow for the estimation of overall treatment effects across several clinical trials. Differences at the dataset level can also be taken into account by re-calibrating existing models to local conditions [104]. Methods for taking into account heterogeneity due to dissimilarity have also been developed for prediction tasks [105], and there have been suggestions for quantifying similarity in such settings [106], [107].

Glossary: ***Representation learning*** focuses on transforming raw input data into compact, meaningful representations that capture the essential features and patterns within the data. These representations make it easier for downstream tasks, such as classification or clustering, to identify underlying relationships. Techniques like autoencoder deep neural networks are commonly used to uncover these representations, which are particularly valuable in small data settings. For example, in image recognition, representation learning can distil critical features like edges and textures into an efficient format, enabling accurate predictions with limited training samples.

Glossary: ***Neuro-symbolic AI*** combines the pattern recognition capabilities of neural networks with the structured reasoning of symbolic AI. This hybrid approach aims to address the limitations of purely neural or symbolic systems by integrating data-driven learning with explicit knowledge representation and reasoning. Neuro-symbolic AI enables systems to reason logically about their predictions while handling complex, unstructured data like images or text.

From a computer science perspective, similarity can also be learned from data, called metric learning [108]–[111]. Neural networks can also be used for learning a potentially complex mapping into a latent space where similarity can be quantified (e.g., [112], [113]). For incorporating similarity at the dataset level, so-called dataset meta features can be leveraged in meta-learning approaches [114].

4.3 Transfer

Approaches for information transfer have been primarily driven by the computer science community and will be presented in more detail in the following subsections. Such approaches incorporate techniques for transferring information from trained models via pre-training and fine-tuning strategies, or specifically in the context of LLMs, via techniques for retrieval-augmented generation and in-context learning. More generally, transferring knowledge based on a representation learned from data is enabled by approaches in the field of representation learning. This includes techniques for few-shot learning, where a model is trained to perform tasks with very limited training examples.

Another strategy to transfer information is to incorporate explicit knowledge. Corresponding approaches are developed in the field of neuro-symbolic AI for incorporating knowledge, e.g., represented by mathematical equations or logical systems. More generally, approaches for combining knowledge-driven and data-driven modelling enable AI-based models to leverage such explicit knowledge, e.g., encoded in differential equations.

4.4 Uncertainty

The topic of uncertainty is at the core of mathematical and applied statistics, e.g., when providing uncertainty of estimates in statistical models via confidence intervals. Quantifying uncertainty is also a crucial aspect of (network) meta-analysis [115]. The need to quantify uncertainty of model parameters can also be extended to modelling techniques such as differential equations that might be used for incorporating knowledge [116]. Beyond parameter estimation, the need to quantify uncertainty more broadly encompasses uncertainty in model prediction, as well as in optimisation strategy and hyperparameter tuning [117]–[119]. The latter is of particular relevance in AI-based models, which heavily rely on such hyperparameters to optimise performance. An important class of approaches for reducing uncertainty by learning across many datasets is provided by ensemble methods and specifically meta-learning.

4.5 Foundation models

As introduced briefly in Section 2, foundation models and in particular LLMs have opened a new era of generative AI, as they provide powerful representations adaptable to a wide range of tasks and applications [34]. Typically, they have many parameters (billions or even hundreds of billions) and are first trained on large datasets to learn a general representation. This is referred to as pre-training, while subsequently adjusting the model to a task-specific, potentially a smaller target dataset of interest, is called fine-tuning. Such a pre-training and fine-tuning strategy is a prototypical example of information transfer that may be helpful in small data settings, e.g., if a pre-trained model can be fine-tuned on the small target dataset at hand.

To avoid the need for large amounts of manually annotated training data, foundation models are often trained in a self-supervised or weakly supervised fashion [120]. In self-supervised training, labelled samples are derived from the input data itself, e.g., by predicting a part of the input from other parts or by reconstructing the original input after random perturbation [121], [122]. Weak supervision means training on noisy or only partially labelled samples that can be generated in an automated way [122].

Foundation models are typically based on a transformer neural network architecture [123]. Transformers have originally been developed for natural language processing (NLP) tasks, such as sentiment classification, text summarization, question answering or translation. They leverage the concept of self-attention to learn the patterns that underlie given input sequences, such as sentences. Even though their core architecture is not inherently generative, transformers can be adapted to generative tasks by combining them with additional components such as autoregressive decoding [124], where the model generates one word at a time conditional on previously generated words. Alternatively, they can be used within a generative adversarial setup for text generation tasks with transformers as generator and discriminator components (e.g., [125]–[128]). This technique is leveraged, for example, in the GPT models (generative pre-trained transformer), a series of autoregressive language models that form the backbone of ChatGPT [129], [130]. Another prominent example of a language foundation model based on transformers is BERT (bidirectional encoder representations from transformers) [131], which has also been adapted for other applications, e.g., for protein structure prediction [132].

The data that can be generated with such models is not limited to text. For example, foundation models and transformers have found their way into a wide range of data types and applications, including image [133], [134] and audio data [76], [77]. They have even been extended across multiple modalities [78]–[80]. A well-known example are image generators such as DALL-E [74], [75], where images are generated based on a text prompt entered by the user. Notably, transformer-based foundation models even enable the generation of tabular data, such as data collected from individuals, e.g., customers, study participants, or patients. For example, prior-data fitted networks [135] leverage a transformer architecture for training on a large number of synthetic tabular datasets and subsequently perform inference on a potentially small real-world target dataset without re-training [135]–[137]. Such foundation models for synthetic data are particularly interesting from a small data perspective, as they leverage synthetic data as prior knowledge, which allows the minimization of training data, in line with data minimisation principles.

4.6 Representation learning, in-context learning, and few-shot learning

Many generative AI approaches can be considered a form of representation learning. This means that they learn to embed the input data into a meaningful representation, which captures its underlying patterns and structure and allows for generating new data based on the representation. Such representation learning is closely linked to transfer learning, as the learned representation allows researchers to generalise different information sources, as well as datasets, and transfer information between them. For example, such a representation enables information transfer from the dataset used for pre-training to a target dataset of interest, or even between different data modalities (e.g., in multi-modal foundation models).

A powerful representation is also a key feature of in-context learning. In-context learning performs a new task based on a small set of examples presented within the context (the prompt) at inference time [85]. This property is found in particular in LLMs, where demonstrations of the task are provided to the model as part of the prompt (in natural language). This enables application on novel tasks without the need for fine-tuning the LLM. Further, it may enable use of LLMs in small data settings, where in-context learning can be used to tailor the pre-trained LLM to the small data task at hand.

Another strategy to tailor LLMs to a specific task of interest based on a learned representation is retrieval-augmented generation. Here, an LLM is linked to an external data source or knowledge base to include additional information [138]–[140]. This enables information transfer between the external knowledge base and the LLM representation, and thus can also be a strategy for leveraging the power of LLMs for small data settings. The idea of including knowledge into a data-driven model such as an LLM is tightly linked to principles of neuro-symbolic AI and more generally combining data-driven and knowledge-driven modelling, as explained in more detail below.

Few-shot learning techniques can be employed when a small dataset at hand, comprising only few observations, has a similar format as the data used for training a large-scale model. Few-shot learning mimics humans' ability to learn tasks after having seen only a few samples. The aim is to leverage the representation obtained from a larger dataset to aid in making predictions for a smaller target set [108]. Similarly to the setting of in-context learning, a few-shot approach is able to make predictions for a new task not seen earlier during training by reasoning only from a few examples (analogous to the context) provided during inference. In this case, the context is given by the observations of the few-shot task.

Few-shot learning is an attractive tool for tackling small data challenges, as it aims to optimise performance when training data is scarce. For example, research in few-shot learning focuses on how to best extract features from training data, such that the model can learn a feature space that also covers the subsequent few-shot tasks, i.e., aiming to find an optimal representation for generalising to the few-shot tasks. The representation should further be structured such that training on the subsequent task should not lead to forgetting critical structure from previous trainings [141]. In small data situations, few-shot methods also allow for increasing the set of examples used for training (with data augmentation or generative approaches) or transferring knowledge from prior learning [142], also across domains. For example, in cross-modal few-shot learning, large-scale public image databases such as ImageNet [143] are leveraged for the few-shot task of medical image diagnosis challenges [144]–[147].

4.7 Meta-learning

Meta-learning refers to a set of approaches that allow for sharing information between several (potentially small) datasets. Such information transfer can, e.g., occur via hyperparameters, neural network architectures, or other meta-parameters, such as initial values of model parameters. Meta-learning approaches then aim to train a model jointly on several datasets so that it can be fine-tuned quickly for new tasks with few examples [148], [149]. Specifically, prior knowledge is extracted from a meta-dataset of several different datasets and/or tasks and subsequently leveraged to fine-tune on a separate target dataset. This strategy has been particularly successful for joint meta-learning of prediction models [148], [150]. As several datasets are incorporated at once, typically less fine-tuning is required on the target dataset [151], which is particularly attractive when the target dataset is small.

However, a necessary condition for meta-learning is that source and target datasets are not too dissimilar. To address this, meta-learning can be combined with similarity techniques for quantifying and incorporating the similarity of datasets, e.g., via dataset meta-features, i.e., characteristics of the target dataset. For example, auxiliary neural networks have been used for generating such dataset meta-features for tabular datasets [114].

4.8 Knowledge-driven modelling

In many real-world applications, a substantial body of knowledge about the subject at hand may already be available, e.g., the specialist knowledge of domain experts or knowledge encoded in rules, systems, or mathematical equations. Knowledge-driven modelling or the incorporation of such existing prior knowledge into modelling may allow for the reduction of the amount of extra data needed and is thus particularly attractive for small data scenarios.

This idea has been at the core of much of the research in statistics and mathematical statistics, meaning that these disciplines are a key contributor to small data methodologies [1]. There has been a recent surge in research in these disciplines, for infusing this broad small data expertise into settings where big data techniques need to be adapted to small data settings. This is complemented by developments in computer science, where ideas of knowledge driven-modelling are, e.g., reflected in neuro-symbolic AI. This approach combines data-driven modelling by neural networks with symbolic AI techniques, which rely on a symbolic representation of explicit knowledge and logical reasoning, e.g., leveraging expert systems, rule-based systems designed to mimic the decision-making capabilities of a human domain expert, logic programming, semantic web techniques, or planning and scheduling systems [152]–[154].

In contrast to neural networks, which are powerful for learning complex patterns from data but are inherently opaque and difficult to interpret, such symbolic AI techniques can provide greater transparency, explainability, and verifiability and therefore increase trustworthiness. Neuro-symbolic AI aims to integrate these two paradigms and combine their respective strengths to provide hybrid tools that integrate reasoning and learning from data in an efficient and transparent way [154], [155]. Thus, they can provide more explainable AI approaches and allow for incorporating prior knowledge into learning [155]. These properties may be particularly desirable in sensitive applications where explainability is especially important, such as in precision medicine, or for policymaking in general.

As indicated, knowledge-driven modelling is a dominant paradigm in statistics and many areas of mathematics and physics. Here, research often builds on a strong body of prior knowledge, which is typically reflected in strong modelling assumptions, e.g., in statistics, when prespecifying the structure of a regression model for assessing treatment effects in a clinical trial. Similarly, in physics or systems biology, temporal dynamics are typically modelled by differential equations that are defined based on domain knowledge, e.g., about the relevant interactions and regulatory processes to describe a small set of molecular or cellular quantities [156]. For building on such knowledge-driven modelling tools, while allowing for additional flexibility, they can be combined with flexible, general-purpose data-driven modelling techniques that require less specific domain knowledge [157], promising tools that are more data-efficient, and at the same time more interpretable. For example, neural differential equations [158], [159] integrate mechanistic knowledge into neural networks by combining them with differential equations, thus directly combining data-driven and knowledge-driven modelling. More generally, neural networks can be used as components to replace parts of a mechanistic model, such as a system of differential equation systems that cannot be fully specified [160], [161]. From a technical perspective, such approaches are often based on the paradigm of differentiable programming, which generally allows the integration of modelling components from different communities, e.g., machine learning and scientific computing [162].

When knowledge on a subject of interest is specific enough that it can be phrased in terms of a prior distribution for a parameter in a model, statistical modelling provides a rich set of Bayesian techniques for model building and evaluation [163]. Yet, when such a precise specification is not feasible, but knowledge still is available in the form of models, there are other ways for incorporating such knowledge, e.g., by augmenting the original data with data from the model, post-filtering [164], domain adaptation [165], or regularization [166]. Pre-existing knowledge can also be incorporated based on expert input in the form of probabilistic relationships of variables [167]. Approaches for incorporating knowledge into machine learning have been more generally investigated and categorised under the label "informed machine learning" [168], also with specific developments for other application domains, such as physics [169].

4.9 General challenges

4.9.1 Overfitting

Training models on limited data incurs the risk of overfitting, i.e., when the model has learned characteristics unique to the limited training data and is aligned too closely to it, but does not generalise to other datasets. While overfitting can also occur in large datasets, small data settings are particularly prone to it due to the limited information available. With fewer data points, there may not be enough diversity to cover a wide range of situations, making it more likely that the model will pick up noise or specific patterns that do not generalise [170, p. 108–114].

A related problem can occur when the observations fall into several categories, and some categories are represented more often than others. For example, in a longitudinal clinical dataset on a severe chronic disease, there might be more patients with comorbidities than without, i.e., the presence of comorbidities is overrepresented. The model is then biased towards predicting the more prevalent class, leading to overrepresentation, or may fail to predict the minority class, leading to underrepresentation [171]. In small data settings, there may simply not be enough data points within a singular class in order to properly assess it.

When models are based on historical data, that is, data that was collected in the past, there is no guarantee that this data will remain in this form over time, or is the same in other datasets [172]. Potentially necessary adaptation could again be based on approaches for similarity and transfer.

4.9.2 Validation

When developing a model on a specific dataset of interest, the data is typically split into a dataset for training the model and a second dataset for testing the model [173]. Sometimes, when training the model, data from the test dataset is accidentally used during the training process. This phenomenon is known as data leakage. While this is an issue for both big data and small data, it tends to have more pronounced and unpredictable effects on small datasets compared to larger ones [174]. Because the datasets are so small, this can lead to an overestimation of model accuracy, a decrease in its ability to generalise to new data (known as generalisability), and the introduction of bias.

After internal validation, models need to go through external validation to see whether the results are generalisable to other, similar datasets [175]. In a small data scenario, this might be rather challenging to realise if there is no sufficient external data available, or if only dissimilar data is available. Assessment and incorporation of similarity is therefore crucial. To remove hurdles, policies should facilitate streamlined data exchange and recommend the creation of data in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles [176] to enable joint use of similar datasets for validation, or already for model development.

5 Conclusions and future perspectives

In an increasingly data-centric world, it is important to make full use of new technologies, such as AI techniques. As shown, the potential benefits are not limited to big data settings, but in particular small data settings can benefit and also need more attention, e.g., to adequately serve underrepresented groups. Novel data analysis and modelling techniques for small data settings can then complement technical developments, such as in assistive technologies, to realise their full potential.

We have provided a variety of case studies and application settings not only to highlight the broad applicability of small data solutions, but also to show the potential from areas where small data techniques are more established, e.g., for rare diseases, to those where the potential is apparent but techniques still need to be established.

However, a wide range of applications, each with contributions from many stakeholders and academic disciplines, runs the risk of confusing newcomers and reinventing the wheel. To enable sufficient communication across disciplines, a shared language is needed as a first step towards a joint small data framework. This motivated us to jointly present ideas from different disciplines, such as the 'average man' from the social sciences, transfer learning from computer science, and the quantification of similarity and uncertainty from statistics and mathematics.

As the common themes of similarity, transfer, and uncertainty may be more generally useful for structuring the field of small data methods, we have provided an overview of how these themes are approached by the different disciplines, including links between them. Hopefully, this will stimulate the fusion of knowledge-driven approaches, as traditionally used in statistics and mathematics, and data-driven approaches, as now prominently established with deep learning techniques in computer science. Such a fusion could be the key to unlocking AI techniques for small data settings. As method development in different disciplines is currently rather siloed, further investment may be warranted to gain commitment from different disciplines and allow for the development of common methods, as well as analysis of barriers to interdisciplinary small data research.

As an example of a specific technique that could be considered a starting point for such endeavours, foundation models, such as LLMs, appear to be a particularly promising approach for bringing additional information to small data sets. As described, there is currently a vibrant community developing foundation models for different fields, and training/fine-tuning approaches that can incorporate many different types of data and models. This flexibility could also offer an avenue for contributions from different disciplines, where disciplines focused on knowledge-driven approaches could, for example, provide knowledge in the form of synthetic data that can then be ingested by data-driven approaches. Of course, this will also depend on establishing a shared language and a forum where different disciplines can develop a common understanding in order to make the best use of different expertise.

Developments in the methods community must be accompanied by a strong push to expand its application across more fields. So far, many assume that fields which cannot offer big data are cut off from novel methods, such as AI. Therefore, it is of prime importance to raise awareness for the opportunities of small data. One way this could be achieved is through dedicated initiatives that bring together all stakeholders. This would also be critical for ensuring feedback from the specific application fields into the broader methodological community, for steering the small data field in the right direction, and for realising the full potential of small data in everyday life.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB 1597.

References

[1] J. J. Faraway and N. H. Augustin, “When small data beats big data,” *Statistics & Probability Letters*, vol. 136, pp. 142–145, 2018. doi: 10.1016/j.spl.2018.02.031.

[2] H. Chahal, H. Toner, and I. Rahkovsky, “Small data’s big AI potential,” Center for Security and Emerging Technology, 2021. doi: 10.51593/20200075.

[3] S. Ehrenberg-Silies, “Beyond big data,” Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB), 2019. doi: 10.5445/IR/1000133938.

[4] E. Strickland, “Andrew Ng, AI minimalist: The machine-learning pioneer says small is the new big,” *IEEE Spectrum*, vol. 59, no. 4, pp. 22–50, 2022. doi: 10.1109/MSPEC.2022.9754503.

[5] I. H. Rather, S. Kumar, and A. H. Gandomi, “Breaking the data barrier: A review of deep learning techniques for democratizing AI with small datasets,” *Artificial Intelligence Review*, vol. 57, no. 9, p. 226, 2024. doi: 10.1007/s10462-024-10859-3.

[6] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone, “Big data from small data: Data-sharing in the ‘long tail’ of neuroscience,” *Nature Neuroscience*, vol. 17, no. 11, pp. 1442–1447, 2014. doi: 10.1038/nn.3838.

[7] E. B. Hekler, P. Klasnja, G. Chevance, N. M. Golaszewski, D. Lewis, and I. Sim, “Why we need a small data paradigm,” *BMC Medicine*, vol. 17, pp. 1–9, 2019. doi: 10.1186/s12916-019-1366-x.

[8] S. Shilo, H. Rossman, and E. Segal, “Axes of a revolution: Challenges and promises of big data in healthcare,” *Nature Medicine*, vol. 26, no. 1, pp. 29–38, 2020. doi: 10.1038/s41591-019-0727-5.

[9] S. Leonelli, “Data—from objects to assets,” *Nature*, vol. 574, no. 7778, pp. 317–320, 2019. doi: 10.1038/d41586-019-03062-w.

[10] Z. Li, H. Yao, and F. Ma, “Learning with small data,” in *International Conference on Web Search and Data Mining*, 2020, pp. 884–887. doi: 10.1145/3336191.3371874.

[11] T. Shaikhina and N. A. Khovanova, “Handling limited datasets with neural networks in medical applications: A small-data approach,” *Artificial Intelligence in Medicine*, vol. 75, pp. 51–63, 2017. doi: 10.1016/j.artmed.2016.12.003.

[12] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, *et al.*, “Machine learning methods for small data challenges in molecular science,” *Chemical Reviews*, vol. 123, no. 13, pp. 8736–8780, 2023. doi: 10.1021/acs.chemrev.3c00189.

[13] P. Kokol, M. Kokol, and S. Zagoranski, “Machine learning on small size samples: A synthetic knowledge synthesis,” *Science Progress*, vol. 105, no. 1, 2022. doi: 10.1177/00368504211029777.

[14] L. Faes, D. A. Sim, M. van Smeden, U. Held, P. M. Bossuyt, and L. M. Bachmann, “Artificial intelligence and statistics: Just the old wine in new wineskins?” *Frontiers in Digital Health*, vol. 4, p. 833 912, 2022. doi: 10.3389/fdgth.2022.833912.

[15] J. Bornschein, F. Visin, and S. Osindero, “Small data, big decisions: Model selection in the small-data regime,” 2020. arXiv: 2009.12583, pre-published.

[16] L. Manovich, “Trending: The promises and the challenges of big social data,” in *Debates in the Digital Humanities*, University of Minnesota Press, 2012, pp. 460–475. doi: 10.5749/minnesota/9780816677948.003.0047.

[17] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Information, Communication and Society*, vol. 15, no. 5, pp. 662–679, 2012. doi: 10.1080/1369118X.2012.678878.

[18] V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: Murray, 2013, 242 pp.

[19] T. H. Davenport and J. Dyché, “Big data in big companies,” *International Institute for Analytics*, 2013. [Online]. Available: https://www.iqpc.com/media/7863/11710.pdf.

[20] M. Hillyer. “Here’s how technology has changed the world since 2000,” World Economic Forum. (2020), [Online]. Available: https://www.weforum.org/stories/2020/11/heres-how-technology-has-changed-and-changed-us-over-the-past-20-years/.

[21] E. A. Ashley, “Towards precision medicine,” *Nature Reviews Genetics*, vol. 17, no. 9, pp. 507–522, 2016. doi: 10.1038/nrg.2016.86.

[22] J. Lee, E. Lapira, B. Bagheri, and H.-a. Kao, “Recent advances and trends in predictive manufacturing systems in big data environment,” *Manufacturing Letters*, vol. 1, no. 1, pp. 38–41, 2013. doi: 10.1016/j.mfglet.2013.09.005.

[23] J. Stirnemann, N. Belmatoug, F. Camou, C. Serratrice, R. Froissart, C. Caillaud, *et al.*, “A review of Gaucher disease pathophysiology, clinical presentation and treatments,” *International Journal of Molecular Sciences*, vol. 18, no. 2, p. 441, 2017. doi: 10.3390/ijms18020441.

[24] H. C. Andersson, J. Charrow, P. Kaplan, P. Mistry, G. M. Pastores, A. Prakesh-Cheng, *et al.*, “Individualization of long-term enzyme replacement therapy for Gaucher disease,” *Genetics in Medicine*, vol. 7, no. 2, pp. 105–110, 2005. doi: 10.1097/01.GIM.0000153660.88672.3C.

[25] J. Treviranus, “The value of being different,” in *International Web for All Conference*, 2019, pp. 1–7. doi: 10.1145/3315002.3332429.

[26] D. Tafreshi, “Adolphe Quetelet and the legacy of the “average man” in psychology.,” *History of Psychology*, vol. 25, no. 1, p. 34, 2022. doi: 10.1037/hop0000202.

[27] G. Marsden, A. Maunder, and M. Parker, “People are people, but technology is not technology,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1881, pp. 3795–3804, 2008. doi: 10.1098/rsta.2008.0119.

[28] G. J. Downey, *Closed Captioning: Subtitling, Stenography, and the Digital Convergence of Text with Television*. Baltimore: Johns Hopkins University Press, 2008, 387 pp.

[29] E. Steinfeld and J. Maisel, *Universal Design: Creating Inclusive Environments*. Hoboken, NJ: John Wiley & Sons, Inc, 2012, 382 pp.

[30] *Web Accessibility: A Foundation for Research*. Springer London, 2019.

[31] A. Pradhan, K. Mehta, and L. Findlater, ""Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities," in *Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2018, pp. 1–13. doi: 10.1145/3173574.3174033.

[32] K. Ellis and G. Goggin, "Disability, locative media, and complex ubiquity," in *Ubiquitous Computing, Complexity and Culture*, Routledge, 2015, pp. 272–287.

[33] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

[34] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, *et al.*, "On the opportunities and risks of foundation models," 2022. arXiv: 2108.07258, pre-published.

[35] "Back to the building blocks: A path toward secure and measurable software," White House Office of the National Cyber Director, 2024. [Online]. Available: https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/Final-ONCD-Technical-Report.pdf.

[36] Department of Homeland Security, "The case for memory safety roadmaps," Cybersecurity Infrastructure and Security Agency, 2023. [Online]. Available: https://www.cisa.gov/sites/default/files/2023-12/The-Case-for-Memory-Safe-Roadmaps-508c.pdf.

[37] R. Khoury, A. R. Avila, J. Brunelle, and B. M. Camara, "How secure is code generated by ChatGPT?," 2023. arXiv: 2304.09655, pre-published.

[38] H. Stewart and J. Jürjens, "Data security and consumer trust in FinTech innovation in germany," *Information & Computer Security*, vol. 26, no. 1, pp. 109–128, 2018. doi: 10.1108/ICS-06-2017-0039.

[39] A. Torno, D. R. Metzler, and V. Torno, "Robo-what?, robo-why?, robo-how? – a systematic literature review of robo-advice," *Pacific Asia Conference on Information Systems*, 2021. [Online]. Available: https://aisel.aisnet.org/pacis2021/92.

[40] L. Wewege, J. Lee, and M. C. Thomsett, "Disruptions and digital banking trends," *Journal of Applied Finance & Banking*, vol. 10, no. 6, pp. 1–2, 2020. [Online]. Available: https://ideas.repec.org//a/spt/apfiba/v10y2020i6f10_6_2.html.

[41] L. Piwek, D. A. Ellis, S. Andrews, and A. Joinson, "The rise of consumer health wearables: Promises and barriers," *PLoS medicine*, vol. 13, no. 2, 2016. doi: 10.1371/journal.pmed.1001953.

[42] *Handbook of Mixed Methods in Social and Behavioral Research*. Thousand Oaks, California: SAGE Publications, 2003.

[43] A. Ramachandran and A. Karuppiah, "A survey on recent advances in wearable fall detection systems," *BioMed Research International*, vol. 2020, pp. 1–17, 2020. doi: 10.1155/2020/2167160.

[44] R. Igual, C. Medrano, and I. Plaza, “Challenges, issues and trends in fall detection systems,” *BioMedical Engineering OnLine*, vol. 12, no. 1, p. 66, 2013. doi: 10.1186/1475-925X-12-66.

[45] C. Medrano, I. Plaza, R. Igual, Á. Sánchez, and M. Castro, “The effect of personalization on smartphone-based fall detectors,” *Sensors*, vol. 16, no. 1, p. 117, 2016. doi: 10.3390/s16010117.

[46] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, “UP-Fall detection dataset: A multimodal approach,” *Sensors*, vol. 19, no. 9, p. 1988, 2019. doi: 10.3390/s19091988.

[47] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021. doi: 10.1007/s10994-021-05946-3.

[48] National Research Council (US), *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington (DC): National Academies Press (US), 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK91503/.

[49] F. Ciardiello, D. Arnold, P. G. Casali, A. Cervantes, J.-Y. Douillard, A. Eggermont, *et al.*, “Delivering precision medicine in oncology today and in future — the promise and challenges of personalised cancer medicine: A position paper by the european society for medical oncology (ESMO),” *Annals of Oncology*, vol. 25, no. 9, pp. 1673–1678, 2014. doi: 10.1093/annonc/mdu217.

[50] I. Villalón-García, M. Álvarez-Córdoba, J. M. Suárez-Rivero, S. Povea-Cabello, M. Talaverón-Rey, A. Suárez-Carrillo, *et al.*, “Precision medicine in rare diseases,” *Diseases*, vol. 8, no. 4, p. 42, 2020. doi: 10.3390/diseases8040042.

[51] J. Banerjee, J. N. Taroni, R. J. Allaway, D. V. Prasad, J. Guinney, and C. Greene, “Machine learning in rare disease,” *Nature Methods*, vol. 20, no. 6, pp. 803–814, 2023. doi: 10.1038/s41592-023-01886-z.

[52] S. McDonald, F. Quinn, R. Vieira, N. O’Brien, M. White, D. W. Johnston, *et al.*, “The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: A systematic literature overview,” *Health Psychology Review*, vol. 11, no. 4, pp. 307–323, 2017. doi: 10.1080/17437199.2017.1316672.

[53] R. Mirza, S. Punja, S. Vohra, and G. Guyatt, “The history and development of n-of-1 trials,” *Journal of the Royal Society of Medicine*, vol. 110, no. 8, pp. 330–340, 2017. doi: 10.1177/0141076817721131.

[54] N. B. Gabler, N. Duan, S. Vohra, and R. L. Kravitz, “N-of-1 trials in the medical literature: A systematic review,” *Medical Care*, vol. 49, no. 8, pp. 761–768, 2011. doi: 10.1097/MLR.0b013e318215d90d.

[55] D. R. Zucker, R. Ruthazer, and C. H. Schmid, "Individual (n-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations," *Journal of Clinical Epidemiology*, vol. 63, no. 12, pp. 1312–1323, 2010. doi: 10.1016/j.jclinepi.2010.04.020.

[56] K. T. Schmidt, C. H. Chau, D. K. Price, and W. D. Figg, "Precision oncology medicine: The clinical relevance of patient-specific biomarkers used to optimize cancer treatment," *The Journal of Clinical Pharmacology*, vol. 56, no. 12, pp. 1484–1499, 2016. doi: 10.1002/jcph.765.

[57] M. J. Walker, J. Bourke, and K. Hutchison, "Evidence for personalised medicine: Mechanisms, correlation, and new kinds of black box," *Theoretical Medicine and Bioethics*, vol. 40, pp. 103–121, 2019. doi: 10.1007/s11017-019-09482-z.

[58] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994. doi: 10.3233/AIC-1994-7104.

[59] C. Dussart, P. Pommier, V. Siranyan, G. Grelaud, and S. Dussart, "Optimizing clinical practice with case-based reasoning approach," *Journal of Evaluation in Clinical Practice*, vol. 14, no. 5, pp. 718–720, 2008. doi: 10.1111/j.1365-2753.2008.01071.x.

[60] T. W. Liao, Z. Zhang, and C. R. Mount, "Similarity measures for retrieval in case-based reasoning systems," *Applied Artificial Intelligence*, vol. 12, no. 4, pp. 267–288, 1998. doi: 10.1080/088395198117730.

[61] R. Weber, "Fuzzy set theory and uncertainty in case-based reasoning," *Engineering Intelligent Systems for Electrical Engineering and Communications*, vol. 14, no. 3, p. 121, 2007.

[62] N. Löw, J. Hesser, and M. Blessing, "Multiple retrieval case-based reasoning for incomplete datasets," *Journal of Biomedical Informatics*, vol. 92, p. 103 127, 2019. doi: 10.1016/j.jbi.2019.103127.

[63] D. W. Aha, M. Molineaux, and G. Sukthankar, "Case-based reasoning in transfer learning," in *Case-Based Reasoning Research and Development*, Springer, 2009, pp. 29–44. doi: 10.1007/978-3-642-02998-1_4.

[64] A. Tricoli, N. Nasiri, and S. De, "Wearable and miniaturized sensor technologies for personalized and preventive medicine," *Advanced Functional Materials*, vol. 27, no. 15, p. 1 605 271, 2017. doi: 10.1002/adfm.201605271.

[65] L. Lu, J. Zhang, Y. Xie, F. Gao, S. Xu, X. Wu, *et al.*, "Wearable health devices in health care: Narrative systematic review," *JMIR mHealth and uHealth*, vol. 8, no. 11, e18907, 2020. doi: 10.2196/18907.

[66] S. Feng, M. ki, A. Dhir, and H. Salmela, "How self-tracking and the quantified self promote health and well-being: Systematic review," *Journal of Medical Internet Research*, vol. 23, no. 9, e25171, 2021. doi: 10.2196/25171.

[67] E. Covi, E. Donati, X. Liang, D. Kappel, H. Heidari, M. Payvand, *et al.*, “Adaptive extreme edge computing for wearable devices,” *Frontiers in Neuroscience*, vol. 15, 2021. doi: 10.3389/fnins.2021.611300.

[68] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, “Convergence of edge computing and deep learning: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020. doi: 10.1109/COMST.2020.2970550.

[69] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, “Deep learning and model personalization in sensor-based human activity recognition,” *Journal of Reliable Intelligent Environments*, vol. 9, no. 1, pp. 27–39, 2023. doi: 10.1007/s40860-021-00167-w.

[70] European Parliament, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

[71] J. Dexe, U. Franke, K. Söderlund, N. van Berkel, R. H. Jensen, N. Lepinkäinen, *et al.*, “Explaining automated decision-making: A multinational study of the GDPR right to meaningful information,” *The Geneva Papers on Risk and Insurance - Issues and Practice*, vol. 47, no. 3, pp. 669–697, 2022. doi: 10.1057/s41288-022-00271-9.

[72] I. Sanchez-Rola, M. Dell’Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, *et al.*, “Can I opt out yet?: GDPR and the global illusion of cookie control,” in *ACM Asia Conference on Computer and Communications Security*, 2019, pp. 340–351. doi: 10.1145/3321705.3329806.

[73] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013. doi: 10.1073/pnas.1218772110.

[74] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, *et al.*, “Zero-shot text-to-image generation,” in *International Conference on Machine Learning*, 2021, pp. 8821–8831. [Online]. Available: https://proceedings.mlr.press/v139/ramesh21a.html.

[75] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with CLIP latents,” 2022. arXiv: 2204.06125, pre-published.

[76] Y.-S. Huang and Y.-H. Yang, “Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,” in *ACM International Conference on Multimedia*, 2020, pp. 1180–1188. doi: 10.1145/3394171.3413671.

[77] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, *et al.*, “UniAudio: An audio foundation model toward universal audio generation,” 2024. arXiv: 2310.00704, pre-published.

[78] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, “Visual ChatGPT: Talking, drawing and editing with visual foundation models,” 2023. arXiv: 2303.04671, pre-published.

[79] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, *et al.*, “mPLUG-2: A modularized multi-modal foundation model across text, image and video,” in *International Conference on Machine Learning*, 2023, pp. 38 728–38 748. [Online]. Available: https://proceedings.mlr.press/v202/xu23s.html.

[80] H.-K. Ko, G. Park, H. Jeon, J. Jo, J. Kim, and J. Seo, “Large-scale text-to-image generation models for visual artists' creative works,” in *International Conference on Intelligent User Interfaces*, 2023, pp. 919–933. doi: 10.1145/3581641.3584078.

[81] W. Hou and Z. Ji, “Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis,” *Nature Methods*, pp. 1–4, 2024. doi: 10.1038/s41592-024-02235-4.

[82] D. Peng, L. Zheng, D. Liu, C. Han, X. Wang, Y. Yang, *et al.*, “Large-language models facilitate discovery of the molecular signatures regulating sleep and activity,” *Nature Communications*, vol. 15, no. 1, p. 3685, 2024. doi: 10.1038/s41467-024-48005-w.

[83] J. Chang and J. C. Ye, “Bidirectional generation of structure and properties through a single molecular foundation model,” *Nature Communications*, vol. 15, no. 1, p. 2323, 2024. doi: 10.1038/s41467-024-46440-3.

[84] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, “Sources of hallucination by large language models on inference tasks,” 2023. arXiv: 2305.14552, pre-published.

[85] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, *et al.*, “Language models are few-shot learners,” 2020. arXiv: 2005.14165, pre-published.

[86] C. J. E. Metcalf and J. Lessler, “Opportunities and challenges in modeling emerging infectious diseases,” *Science*, vol. 357, no. 6347, pp. 149–152, 2017. doi: 10.1126/science.aam8335.

[87] T. Alarcón, H. M. Byrne, and P. K. Maini, “A cellular automaton model for tumour growth in inhomogeneous environment,” *Journal of Theoretical Biology*, vol. 225, no. 2, pp. 257–274, 2003. doi: 10.1016/S0022-5193(03)00244-3.

[88] S. Friedrich, G. Antes, S. Behr, H. Binder, W. Brannath, F. Dumpert, *et al.*, “Is there a role for statistics in artificial intelligence?” *Advances in Data Analysis and Classification*, vol. 16, pp. 823–846, 2022. doi: 10.1007/s11634-021-00455-6.

[89] J. N. Kather, N. Ghaffari Laleh, S. Foersch, and D. Truhn, “Medical domain knowledge in domain-agnostic generative AI,” *npj Digital Medicine*, vol. 5, no. 1, pp. 1–5, 2022. doi: 10.1038/s41746-022-00634-5.

[90] T. Dash, S. Chitlangia, A. Ahuja, and A. Srinivasan, “A review of some techniques for inclusion of domain-knowledge into deep neural networks,” *Scientific Reports*, vol. 12, no. 1, p. 1040, 2022. doi: 10.1038/s41598-021-04590-0.

[91] J. Nie, J. Jiang, Y. Li, H. Wang, S. Ercisli, and L. Lv, “Data and domain knowledge dual-driven artificial intelligence: Survey, applications, and challenges,” *Expert Systems*, vol. 42, no. 1, e13425, 2025. doi: 10.1111/exsy.13425.

[92] J. C. M. Darwin, G. Mendel, R. Franklin, J. Watson, *et al.*, "Pooling the strengths of data and models," *Nature Biomedical Engineering*, vol. 5, no. 4, pp. 291–292, 2021. doi: 10.1038/s41551-021-00721-0.

[93] P. C. Austin, "The performance of different propensity score methods for estimating marginal odds ratios," *Statistics in Medicine*, vol. 26, no. 16, pp. 3078–3094, 2007. doi: 10.1002/sim.2781.

[94] P. C. Austin and E. A. Stuart, "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies," *Statistics in Medicine*, vol. 34, no. 28, pp. 3661–3679, 2015. doi: 10.1002/sim.6607.

[95] O. Borgan, L. Goldstein, and B. Langholz, "Methods for the analysis of sampled cohort data in the cox proportional hazards model," *Annals of Statistics*, vol. 23, no. 5, pp. 1749–1778, 1995. doi: 10.1214/aos/1176324322.

[96] V. Weyer and H. Binder, "A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures," *BMC Bioinformatics*, vol. 16, p. 294, 2015. doi: 10.1186/s12859-015-0716-8.

[97] J. Richter, K. Madjar, and J. Rahnenführer, "Model-based optimization of subgroup weights for survival analysis," *Bioinformatics*, vol. 35, no. 14, pp. i484–i491, 2019. doi: 10.1093/bioinformatics/btz361.

[98] C. Loader, *Local Regression and Likelihood*. New York, NY: Springer New York, 1999, 290 pp.

[99] R. J. Samworth, "Optimal weighted nearest neighbour classifiers," *Annals of Statistics*, vol. 40, no. 5, pp. 2733–2763, 2012. doi: 10.1214/12-AOS1049.

[100] W. Liu, F. Bretz, A. J. Hayter, and H. P. Wynn, "Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region," *Biometrics*, vol. 65, no. 4, pp. 1279–1287, 2009. doi: 10.1111/j.1541-0420.2008.01192.x.

[101] H. Dette, K. Möllenhoff, S. Volgushev, and F. Bretz, "Equivalence of regression curves," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 711–729, 2018. doi: 10.1080/01621459.2017.1281813.

[102] N. Binder, K. Möllenhoff, A. Sigle, and H. Dette, "Similarity of competing risks models with constant intensities in an application to clinical healthcare pathways involving prostate cancer surgery," *Statistics in Medicine*, vol. 41, no. 19, pp. 3804–3819, 2022. doi: 10.1002/sim.9481.

[103] N. Binder, H. Dette, J. Franz, D. Zöller, R. Suarez-Ibarrola, C. Gratzke, *et al.*, "Data mining in urology: Understanding real-world treatment pathways for lower urinary tract systems via exploration of big data," *European Urology Focus*, vol. 8, no. 2, pp. 391–393, 2022. doi: 10.1016/j.euf.2022.03.019.

[104] K. G. M. Moons, A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, *et al.*, “Risk prediction models: II. external validation, model updating, and impact assessment,” *Heart*, vol. 98, no. 9, pp. 691–698, 2012. doi: 10.1136/heartjnl-2011-301247.

[105] V. M. T. de Jong, K. G. M. Moons, M. J. C. Eijkemans, R. D. Riley, and T. P. A. Debray, “Developing more generalizable prediction models from pooled studies and large clustered data sets,” *Statistics in Medicine*, vol. 40, no. 15, pp. 3533–3559, 2021. doi: 10.1002/sim.8981.

[106] G. R. Klein, J. B. Greenhouse, B. D. Stein, and H. J. Seltman, “Characterizing patterns of care using administrative claims data: ADHD treatment in children,” *Health Services and Outcomes Research Methodology*, vol. 11, no. 3, pp. 115–133, 2011. doi: 10.1007/s10742-011-0076-4.

[107] Z. Huang, W. Dong, H. Duan, and H. Li, “Similarity measure between patient traces for clinical pathway analysis: Problem, method, and applications,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 4–14, 2014. doi: 10.1109/JBHI.2013.2274281.

[108] W. Ren, Y. Tang, Q. Sun, C. Zhao, and Q.-L. Han, “Visual semantic segmentation based on few/zero-shot learning: An overview,” *IEEE/CAA Journal of Automatica Sinica*, pp. 1–21, 2023. doi: 10.1109/JAS.2023.123207.

[109] B. M. Mathisen, A. Aamodt, K. Bach, and H. Langseth, “Learning similarity measures from data,” *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 129–143, 2020. doi: 10.1007/s13748-019-00201-2.

[110] J. Wu, H. Shen, Y.-D. Li, Z.-B. Xiao, M.-Y. Lu, and C.-L. Wang, “Learning a hybrid similarity measure for image retrieval,” *Pattern Recognition*, vol. 46, no. 11, pp. 2927–2939, 2013. doi: 10.1016/j.patcog.2013.04.008.

[111] K. Liu, A. Bellet, and F. Sha, “Similarity learning for high-dimensional sparse data,” in *International Conference on Artificial Intelligence and Statistics*, 2015, pp. 653–662. [Online]. Available: https://proceedings.mlr.press/v38/liu15.html.

[112] C. H. Grønbech, M. F. Vording, P. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther, “scVAE: Variational auto-encoders for single-cell gene expression data,” *Bioinformatics*, vol. 36, no. 16, pp. 4415–4422, 2020. doi: 10.1093/bioinformatics/btaa293.

[113] M. Hackenberg, P. Harms, M. Pfaffenlehner, A. Pechmann, J. Kirschner, T. Schmidt, *et al.*, “Deep dynamic modeling with just two time points: Can we still allow for individual trajectories?” *Biometrical Journal*, 2022. doi: 10.1002/bimj.202000366.

[114] H. S. Jomaa, L. Schmidt-Thieme, and J. Grabocka, “Dataset2Vec: Learning dataset meta-features,” *Data Mining and Knowledge Discovery*, vol. 35, no. 3, pp. 964–985, 2021. doi: 10.1007/s10618-021-00737-9.

[115] A. Nikolakopoulou, D. Mavridis, and G. Salanti, “Planning future studies based on the precision of network meta-analysis results,” *Statistics in Medicine*, vol. 35, no. 7, pp. 978–1000, 2016. doi: 10.1002/sim.6608.

[116] C. Kreutz, "An easy and efficient approach for testing identifiability," *Bioinformatics*, vol. 34, no. 11, pp. 1913–1921, 2018. doi: 10.1093/bioinformatics/bty035.

[117] J. K. H. Franke, G. Köhler, N. Awad, and F. Hutter, "Neural architecture evolution in deep reinforcement learning for continuous control," 2020. arXiv: 1910.12824, pre-published.

[118] F. Wenzel, J. Snoek, D. Tran, and R. Jenatton, "Hyperparameter ensembles for robustness and uncertainty quantification," in *Advances in Neural Information Processing Systems*, 2020, pp. 6514–6527. [Online]. Available: https://papers.neurips.cc/paper_files/paper/2020/hash/481fbfa59da2581098e841b7afc122f1-Abstract.html.

[119] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, e1484, 2023. doi: 10.1002/widm.1484.

[120] Y. Yuan, "On the power of foundation models," in *International Conference on Machine Learning*, 2023, pp. 40 519–40 530. [Online]. Available: https://proceedings.mlr.press/v202/yuan23b.html.

[121] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems*, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/a2b15837edac15df90721968986f7f8e-Paper.pdf.

[122] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, 2023. doi: 10.1038/s41586-023-06221-2.

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[124] M. Stern, N. Shazeer, and J. Uszkoreit, "Blockwise parallel decoding for deep autoregressive models," in *Advances in Neural Information Processing Systems*, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/c4127b9194fe8562c64dc0f5bf2c93bc-Abstract.html.

[125] L. Zhao, Z. Zhang, T. Chen, D. Metaxas, and H. Zhang, "Improved transformer for high-resolution gans," in *Advances in Neural Information Processing Systems*, 2021, pp. 18 367–18 380. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/98dce83da57b0395e163467c9dae521b-Abstract.html.

[126] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," in *Advances in Neural Information Processing Systems*, 2021, pp. 14 745–14 758. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/7c220a2091c26a7f5e9f1cfb099511e3-Abstract.html.

[127] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "ST-GAN: Spatial transformer generative adversarial networks for image compositing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9455–9464. doi: 10.1109/CVPR.2018.00985.

[128] X. Li, V. Metsis, H. Wang, and A. H. H. Ngu, "TTS-GAN: A transformer-based time-series generative adversarial network," in *Artificial Intelligence in Medicine*, Springer International Publishing, 2022, pp. 133–143. doi: 10.1007/978-3-031-09342-5_13.

[129] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[130] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, *et al.*, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023. doi: 10.1109/JAS.2023.123618.

[131] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[132] R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, *et al.*, "Single-sequence protein structure prediction using a language model and deep learning," *Nature Biotechnology*, vol. 40, no. 11, pp. 1617–1623, 2022. doi: 10.1038/s41587-022-01432-w.

[133] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022. doi: 10.1109/TPAMI.2022.3152247.

[134] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–41, 10s 2022. doi: 10.1145/3505244.

[135] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter, "Transformers can do Bayesian inference," in *International Conference on Learning Representation*, 2022. [Online]. Available: https://openreview.net/pdf?id=KSugKcbNf9.

[136] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=cp5PvcI6w8_.

[137] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025. doi: 10.1038/s41586-024-08328-6.

[138] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[139] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, *et al.*, "Retrieval-augmented generation for large language models: A survey," 2024. arXiv: 2312.10997, pre-published.

[140] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," *Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17 754–17 762, 2024. doi: 10.1609/aaai.v38i16.29728.

[141] L. Wang, Q. Li, Y. Zhong, and J. Zhu, "Few-shot continual learning: A brain-inspired approach," 2021. arXiv: 2104.09034, pre-published.

[142] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2387–2402, 2021. doi: 10.1109/JSTARS.2021.3052869.

[143] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[144] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2017. doi: 10.1109/cvpr.2017.369.

[145] F. Shakeri, M. Boudiaf, S. Mohammadi, I. Sheth, M. Havaei, I. B. Ayed, *et al.*, "FHIST: A benchmark for few-shot classification of histological images," 2022. arXiv: 2206.00092, pre-published.

[146] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *IEEE International Symposium on Biomedical Imaging*, 2018, pp. 168–172. doi: 10.1109/ISBI.2018.8363547.

[147] M. Samiei, T. Würfl, T. Deleu, M. Weiss, F. Dutil, T. Fevens, *et al.*, "The TCGA meta-dataset clinical benchmark," 2019. arXiv: 1910.08636, pre-published.

[148] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017, pp. 1126–1135. [Online]. Available: http://proceedings.mlr.press/v70/finn17a.html.

[149] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," presented at the International Conference on Learning Representations, 2018. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[150] J. Vanschoren, “Meta-learning,” in *Automated Machine Learning: Methods, Systems, Challenges*, Cham: Springer International Publishing, 2019, pp. 35–61. doi: 10.1007/978-3-030-05318-5_2.

[151] S. Jadon and A. Jadon, “An overview of deep learning architectures in few-shot learning domain,” 2023. arXiv: 2008.06365, pre-published.

[152] P. Hitzler and M. K. Sarker, *Neuro-Symbolic Artificial Intelligence: The State of the Art*. Amsterdam Berlin Washington, DC: IOS Press, 2022, 408 pp.

[153] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, and L. Zhou, “Neuro-symbolic approaches in artificial intelligence,” *National Science Review*, vol. 9, no. 6, nwac035, 2022. doi: 10.1093/nsr/nwac035.

[154] A. d’Avila Garcez and L. C. Lamb, “Neurosymbolic AI: The 3rd wave,” *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 387–12 406, 2023. doi: 10.1007/s10462-023-10448-w.

[155] A. Sheth, K. Roy, and M. Gaur, “Neurosymbolic artificial intelligence (why, what, and how),” *IEEE Intelligent Systems*, vol. 38, no. 03, pp. 56–62, 2023. doi: 10.1109/MIS.2023.3268724.

[156] F. J. Bruggeman and H. V. Westerhoff, “The nature of systems biology,” *Trends in Microbiology*, vol. 15, no. 1, pp. 45–50, 2007. doi: 10.1016/j.tim.2006.11.003.

[157] R. Li, L. Li, Y. Xu, and J. Yang, “Machine learning meets omics: Applications and perspectives,” *Briefings in Bioinformatics*, vol. 23, no. 1, pp. 1–22, 2022. doi: 10.1093/bib/bbab460.

[158] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, “Neural ordinary differential equations,” 2019. arXiv: 1806.07366, pre-published.

[159] G. Köber, S. Pooseh, H. Engen, A. Chmitorz, M. Kampa, A. Schick, *et al.*, “Individualizing deep dynamic models for psychological resilience data,” *Scientific Reports*, vol. 12, no. 1, p. 8061, 2022. doi: 10.1038/s41598-022-11650-6.

[160] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, *et al.*, “Universal differential equations for scientific machine learning,” 2021. arXiv: 2001.04385, pre-published.

[161] J. H. Lagergren, J. T. Nardini, R. E. Baker, M. J. Simpson, and K. B. Flores, “Biologically-informed neural networks guide mechanistic modeling from sparse experimental data,” *PLOS Computational Biology*, vol. 16, no. 12, e1008462, 2020. doi: 10.1371/journal.pcbi.1008462.

[162] M. Innes, A. Edelman, K. Fischer, C. Rackauckas, E. Saba, V. B. Shah, *et al.*, “A differentiable programming system to bridge machine learning and scientific computing,” 2019. arXiv: 1907.07587, pre-published.

[163] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, *et al.*, “Bayesian statistics and modelling,” *Nature Reviews Methods Primers*, vol. 1, no. 1, pp. 1–26, 2021. doi: 10.1038/s43586-020-00001-2.

[164] Y. Ba, G. Zhao, and A. Kadambi, “Blending diverse physical priors with neural networks,” 2019. arXiv: 1910.00201, pre-published.

[165] J. A. Gongora, H. P. Marshall, C. Hase, T. Bradberry, and S. Ballerini, “Applications of deep learning and domain adaptation to the study of mountain snow,” *American Geophysical Union*, vol. 2019, 2019. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2019AGUFM.H33L2110G.

[166] M. Raissi, “Deep hidden physics models: Deep learning of nonlinear partial differential equations,” *Journal of Machine Learning Research*, vol. 19, p. 25, 2018. [Online]. Available: https://www.jmlr.org/papers/volume19/18-046/18-046.pdf.

[167] L. von Rueden, S. Houben, K. Cvejoski, C. Bauckhage, and N. Piatkowski, “Informed pre-training on prior knowledge,” 2022. arXiv: 2205.11433, pre-published.

[168] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, *et al.*, “Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 614–633, 2023. doi: 10.1109/TKDE.2021.3079836.

[169] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, “Physics-informed machine learning,” *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021. doi: 10.1038/s42254-021-00314-5.

[170] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org.

[171] M. M. Rahman and D. N. Davis, “Addressing the class imbalance problem in medical datasets,” *International Journal of Machine Learning and Computing*, pp. 224–228, 2013. doi: 10.7763/IJMLC.2013.V3.307.

[172] S. Vollmer, B. A. Mateen, G. Bohner, F. J. Király, R. Ghani, P. Jonsson, *et al.*, “Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness,” *BMJ*, vol. 368, 2020. doi: 10.1136/bmj.q2390.

[173] M. K. Uçar, M. Nour, H. Sindi, K. Polat, *et al.*, “The effect of training and testing process on machine learning in biomedical datasets,” *Mathematical Problems in Engineering*, vol. 2020, 2020. doi: 10.1155/2020/2836236.

[174] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, “Data leakage inflates prediction performance in connectome-based machine learning models,” *Nature Communications*, vol. 15, no. 1, p. 1829, 2024. doi: 10.1038/s41467-024-46150-w.

[175] Q.-Y. Li, B.-H. Tang, Y.-E. Wu, B.-F. Yao, W. Zhang, Y. Zheng, *et al.*, "Machine learning: A new approach for dose individualization," *Clinical Pharmacology & Therapeutics*, vol. 115, no. 4, pp. 727–744, 2024. doi: `10.1002/cpt.3049`.

[176] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, pp. 1–9, 2016. doi: `10.1038/sdata.2016.18`.