*DOI:*
10.1145/3592616

# A survey of data quality requirements that matter in ML development pipelines

MARIA PRIESTLEY, King's College London, United Kingdom
FIONNTÁN O'DONNELL, Open Data Institute, United Kingdom
ELENA SIMPERL, King's College London, United Kingdom

The fitness of the systems in which Machine Learning (ML) is used depends greatly on good quality data. Specifications on what makes a good quality dataset have traditionally been defined by the needs of the data users - typically analysts and engineers. Our article critically examines the extent to which established data quality frameworks are applicable to contemporary use cases in ML. Using a review of recent literature at the intersection of ML, data management, and Human Computer Interaction (HCI), we find that the classical "fitness-for-use" view of data quality can benefit from a more stage-specific approach that is sensitive to where in the ML lifecycle the data are encountered. This helps practitioners to plan their data quality tasks in a manner that meets the needs of the stakeholders who will encounter the dataset, whether it be data subjects, software developers or organisations. We therefore propose a new treatment of traditional data quality criteria by structuring them according to two dimensions: 1) the stage of the ML lifecycle where the use case occurs vs. 2) the main categories of data quality that can be pursued (intrinsic, contextual, representational and accessibility). To illustrate how this works in practice, we contribute a temporal mapping of the various data quality requirements that are important at different stages of the ML data pipeline. We also share some implications for data practitioners and organisations that wish to enhance their data management routines in preparation for ML.

CCS Concepts: • **Information systems** → **Database performance evaluation**; *Data cleaning*; • **Social and professional topics** → **Quality assurance**.

Additional Key Words and Phrases: data quality, machine learning, data ecosystems, data management, data innovation

**ACM Reference Format:**
Maria Priestley, Fionntán O'Donnell, and Elena Simperl. 2023. A survey of data quality requirements that matter in ML development pipelines. *ACM J. Data Inform. Quality* xx, x, Article xxx (x 2023), 40 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

Today's societies are producing immense volumes of data that get used by Artificial Intelligence (AI) systems. At the core of AI applications is the field of Machine Learning (ML), which relies on the use of data to classify or detect patterns in existing information (unsupervised ML), as well as using past data to "train" algorithms to solve new tasks (supervised ML) [37]. Our paper focuses on the latter subset of ML, which is growing in popularity in systems created to predict probable

Authors' addresses: Maria Priestley, maria.1.priestley@gmail.com, King's College London, United Kingdom; Fionntán O'Donnell, research@theodi.org, Open Data Institute, United Kingdom; Elena Simperl, elena.simperl@kcl.ac.uk, King's College London, United Kingdom.


1936-1955/2023/x-ARTxxx $15.00
https://doi.org/XXXXXXX.XXXXXXX

outcomes based on certain inputs, or to make recommendations about which decisions would be optimal in a given scenario.

These systems work with structured as well as unstructured data (e.g. text, images, audio) to address practical use cases in fields such as clinical diagnosis, criminal justice, financial lending, manufacture and autonomous vehicles, among others [52]. In the remainder of this article, we will refer to ML systems as software systems in which ML models or algorithms are deployed, typically for the purposes of solving a problem in the real world.

Poor quality datasets and data science pipelines can compromise ML systems in a number of ways. This includes historical signals and inappropriately proxied measures that make ML systems vulnerable to reproducing past discrimination against under-represented groups (e.g. in contexts such as job hiring and criminal justice), or propagating abusive content [46, 67]. Messy or inaccurate data can also disturb the operational efficiency of businesses, with estimates of 10% to 30% of revenue being spent on resolving data quality issues [30]. The importance of data quality is therefore increasingly being recognised by private and public stakeholders who want to mitigate social risks, reduce costs and support the effective assimilation of ML technologies in society.

The growing use of ML across industries, and the high-stakes nature of some of the above uses [17, 44], is being accompanied by greater scrutiny of the processes that determine the output of ML-based decision-support systems [51]. Routines for ensuring transparency in ML datasets and ML development pipelines are being encouraged by national and international organisations such as the OECD[1] and the Open Government Partnership[2]. The UK government has recently published an Algorithmic Transparency Standard[3] alongside templates designed to help public sector organisations to document the datasets that underlie their ML tools. Similar trends are happening in industry, where new standards are currently being developed to guide businesses on how to define, implement and measure data quality throughout the ML development lifecycle [18]. Standards of this kind crystallise an ever growing corpus of academic literature that has explored ML data quality challenges and ways to mitigate them [25, 57, 64, 70].

The growing range of ML data management guidelines, frameworks and standards presents practitioners with a vast range of possible criteria to aspire to, on top of the traditional data management practices that were established in previous decades. This raises a twofold challenge: 1) how to navigate the ML literature and select only those data quality requirements that are meaningful to the practitioner's use case, and 2) how to address the new requirements using frameworks and practices that are already familiar to the data management community.

Our paper aims to help data practitioners to navigate these challenges by distilling some of the key concepts from recent literature in the fields of ML, data management and Human Computer Interaction (HCI). Our contributions include:

- An overview of some of the key data quality requirements that matter in ML systems.
- An illustration of how these requirements map onto traditional data quality criteria.
- A structure for identifying the most salient data quality requirements depending on the stage of the ML lifecycle where the data use case occurs.

The remainder of this paper is structured as follows. In Section 2, we present the background literature that motivates our work. We then present our methodology for conducting a literature review in Section 3, followed by a summary of results in Section 4 and discussion of the findings in Section 5.

[1]https://oecd.ai/en/dashboards/ai-principles/P7 [accessed 24/01/22]
[2]https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ [accessed 24/01/22]
[3]https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub [accessed 14/01/23]

## 2 BACKGROUND

Training data for ML algorithms can be collected in a variety of ways. In their comprehensive survey of data collection methods for ML, Roh et al. [61] group these into three categories: 1) data acquisition (including discovery, augmentation and generation), 2) data labelling (using manual or semi-supervised approaches) and 3) improvement (cleaning the data itself or improving the model built upon it). The extent to which these data collection methods are used varies depending on the use case and the type of data upon which an ML system relies.

In larger organisations and complex innovation ecosystems, the data may pass through multiple stakeholders and be transformed in various ways before it reaches an ML practitioner or their resultant product. Because of this, the topic of data quality is beginning to transcend beyond the field of data management and into the realm of Human Computer Interaction (HCI), which accommodates holistic considerations such as how people search for relevant datasets [42], how developers perceive data work [64] and the best ways of using crowdsourcing to generate, evaluate or label data [71]. While the role of these dynamic processes and multi-stakeholder configurations is increasingly being recognised by data practitioners, it is less clear how traditional data quality frameworks and notions of data accountability are adapting to ML development pipelines [35].

### 2.1 Data management practices differ between academia and industry

Longstanding definitions of data quality have viewed good quality data as “data that are fit for use by data consumers” [74]. This has been accompanied by granular specifications of what makes a good quality dataset, with essential dimensions such as accuracy, completeness, consistency and validity being just some of the 60 dimensions identified in the wider data management literature [13]. Practical applications of these dimensions typically focus on smaller subsets of the most relevant qualities, and can be found in the UK government’s data strategy[4], professional associations for information management such as AHIMA[5], and the requirements of open data and open science initiatives, where datasets should ideally be linked by the peer-reviewed code or publication that uses them.

It is worth noting that ML data tend to be managed differently depending on whether the system is within an academic or industry setting [52]. In academia, data management is typically contained within the projects of individuals or small teams, who are able to design and amend the data collection, storage and sharing systems at their discretion. Industry researchers, however, often rely on separate data collection, processing and storage systems that sit across multiple company functions, requiring formal data management guidelines to ensure consistency and coordination across teams.

Formal practices of this kind are sometimes established with the help of industry standards. For example, the International Organization for Standardization (ISO) standard ISO/IEC 25012 provides guidance on how to define the data quality characteristics that matter to an organisation. Defining these characteristics is a pre-requisite to deciding how data quality can be evaluated in a practical sense. This latter task is addressed by the standard ISO/IEC 25024, which guides organisations in defining the data quality assurance criteria and ways of measuring them quantitatively. These data quality models are complemented by standards which recognise that organisations differ in their preparedness to define and execute data quality assurance. The ISO 8000-61 standard specifies the pure activities of enhancing data quality processes, while ISO 8000-62 defines ways to assess the maturity, or readiness, of organisations to implement these data quality tasks.

[4]https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework [Accessed 06/10/21]
[5]https://library.ahima.org/PB/DataQualityModel [Accessed 13/11/21]

More recently, ISO has begun to develop the ISO/IEC 5259 standard which focuses on data quality for the fields of analytics and ML, as well as ISO/IEC DIS 8183 that addresses the AI data life cycle framework. These newer standards address processes that can be employed by various stakeholders at different stages of the AI life cycle, which differs from earlier data quality guidelines that tended to view quality as a uniform outcome that fulfils a pre-defined list of desired criteria. These standards are still under development, so there is value in publications that inform practitioners of how data quality applies to ML tasks.

## 2.2 Data quality means meeting the needs of different users

Traditionally, data quality compliance has meant meeting the needs of the immediate data users (e.g. analysts or engineers who value clean machine-readable data). However, this singular focus can flatten the variety of uses and data quality requirements that are encountered at various stages of the ML development pipeline over the longer-term [57, 70]. For instance, data quality aspects that are important to ML developers are likely to be different from what was important to upstream data subjects, who may have valued mechanisms for expressing consent and data usage preferences. Similarly, downstream users of trained ML algorithms, such as software developers and organisations, may have their own preferences for specific data qualities when procuring the system, including aspects such as security, provenance, legal compliance, and the capacity to meet business goals in real-world contexts. It is therefore useful to consider data quality processes in ML as being less about obtaining a finished outcome and more about creating a dynamic artefact that is imbued with the potential to be improved and shaped by different stakeholders to meet their own requirements.

Many of the data quality issues that could reasonably concern the above mentioned stakeholders can already be accommodated by the granular data quality specifications produced in the field of data management. For example, the list of 60 dimensions created by Black and van Nederpelt [13] includes qualities related to data accuracy, lineage, currency, coverage, legal compliance and usability. These dimensions are subsumed by higher-order characterisations that capture the intrinsic, contextual, accessibility, and representational aspects of datasets [74].

While the advent of data-centric technologies has been accompanied by a proliferation of updated data quality definitions and metrics tailored to fields such as big data [68] and linked data [77], contemporary authors continue to find value in existing data quality characterisations and conceptual structures. For example, in their "Data Quality in Use" model for Big Data, Merino et al. [47] draw on Wang and Strong [74]'s canonical distinction between the intrinsic, contextual, accessibility, and representational aspects of datasets when using the above mentioned industry standards ISO/IEC 25012 and ISO/IEC 25024. Other efforts have been made to adapt traditional data quality management practices to specific fields. This includes the work of Kim et al. [41], who developed new frameworks for assessing and improving the maturity of IoT data quality processes based on the standards ISO 8000-61 and ISO 8000-62.

## 2.3 The four dimensions of data quality

Below we will draw on Wang & Strong's [74] categorisation of the intrinsic, contextual, accessibility, and representational aspects of datasets to illustrate some of the ways in which previously established data quality categories already apply to ML problems.

*Intrinsic* data quality has traditionally been understood to reflect the extent to which data values conform to the actual or true values [74]; this includes specific requirements such as accuracy, provenance, and cleanliness, the latter of which covers practices such as the addressing missing values and redundant cases. Besides the usual data qualities needed for statistical analysis (e.g. addressing missing data, anomalies), an intrinsic quality that is increasingly valued by ML practitioners and regulators relates to data lineage and traceability. For data that require multiple pre-processing steps or transactions between organisations, the origins of their features becomes important. Traceability makes it possible to interpret and audit the history that precedes the output of ML algorithms [33], but despite recent regulations on explainable AI (XAI)[6], traceability is not yet shortlisted in the data quality framework used by the UK government[7], suggesting that this data quality characteristic may need to be promoted in the context of ML.

*Contextual* data quality relates to the extent to which data are pertinent to the task of the data user [74]; this includes dimensions such as relevance, timeliness, completeness, and appropriateness. An essential question that is considered here is the extent to which the sample of cases contained in the dataset diverges from the true distribution of cases that are likely to be encountered when the ML model is deployed. Possible sources of divergence may include historical time or geographic representation. For example, temporality has been flagged as a potential source of difficulty in textual data, where models trained on historical text corpora, such as Google News articles, have been found to reproduce past social stereotypes (e.g. the word "man" being associated with "computer programmer" and "woman" with "homemaker") [14]. If left untreated, the use of such data in downstream applications (e.g. web search rankings, question retrieval) can perpetuate or amplify the biases that were and continue to be present in broader society. Other contextual biases have been detected in image data, with publicly available image corpora such as ImageNet and Open Images coming predominantly from amerocentric and eurocentric contexts [66]. Insufficient representation of some geographic regions, such as Asia or Africa, has meant that ML algorithms have less information to learn about these contexts. This results in solutions that perform poorly for under-represented groups (e.g. passport photo software that does not recognise the facial expressions of ethnic minorities, or electronic soap dispensers that do not respond to darker skin tones). These cases urge ML data practitioners to think critically about the context captured by their dataset and the degree to which it reflects the use case and lived experience of the end users.

*Representational* data quality refers to the extent to which data are presented in an intelligible and clear manner, including requirements such as being interpretable, easy to understand, represented concisely and consistently [74]. In practical terms, these qualities can be implemented through practices such as standardisation and documentation. Standardisation refers to conventions for capturing information in a consistent manner, including machine-readable data structures and formats for capturing specific attributes (e.g. date, location, measurement error). This helps engineers to ingest datasets from multiple sources and build interoperable solutions. Documentation about the dataset provides an additional layer of descriptive information to support the creation of ML applications. For example, it can help engineers to understand where the dataset sits in relation to the physical world (e.g. the calibration of equipment, seasonality of data collection, contextual limitations) [64], so that the training data or model output can be transformed accordingly. It is worth highlighting that when the limitations of a dataset are made explicit in the documentation, this helps subsequent users to take the steps needed to improve the quality of the dataset for their specific use case. Some solutions even allow for the dataset to remain unchanged while the ML algorithm is tuned to produce more robust or socially equitable outcomes [14, 29].

The *accessibility* category refers to the extent to which data are available, obtainable and secure. The rise of big data and ML applications in recent decades has been accompanied by calls for publishing datasets in an open manner, as well as secure access mechanisms for restricted datasets, so that their value can be realised [75]. For ML stakeholders who work with personal or commercially

[6]https://ico.org.uk/media/about-the-ico/consultations/2616434/explaining-ai-decisions-part-1.pdf [accessed 18/11/21]
[7]https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework [accessed 26/01/22]

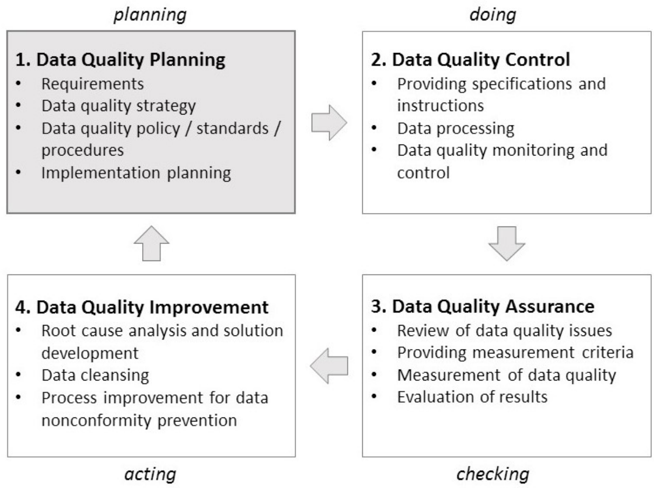


Fig. 1. Data quality management process. Adapted from Kim et al. [41] and ISO 8000-61. The focus of our paper is shaded in grey.

sensitive data, advances in the accessibility of data have been tempered by security and legal precautions (e.g. compliance with GDPR and intellectual property rights).

The data quality concerns exemplified above are already within the scope of the concepts and frameworks that have been established in data management literature, suggesting that this field already has a good grounding for defining the data quality dimensions that will continue to remain important to ML. What is new, however, is that ML development is characterised by complex configurations of datasets, data services and data handlers, which makes individuals more vulnerable to abstain from taking action due to the belief that data quality is somebody else's problem [34]. This diffusion of responsibility can be addressed by providing clearer indicators about which data quality aspects are and are not out of scope of particular ML roles.

## 2.4 Why knowledge of desirable data quality practices is important

The struggle of clarifying which data quality requirements are important is not exclusive to ML. Even where detailed data quality standards and practices exist, organisations and/or practitioners have to specify which data quality characteristics are relevant to their use case and how to define them.

In a study of organisations that applied the ISO/IEC 25012 data quality standard, Gualo et al. [24] found that practitioners struggled to identify and describe the data quality rules that applied to their use case. The authors found that providing examples of what the requirements can look like helps to guide practitioners in clarifying their own rules.

Another challenge relates to information overload. Long lists of requirements have been found to deter practitioners from applying traditional standards, with Kim et al. [41] showing that there is value in simplified frameworks that are tailored to a specific use case or technology.

Both of the above challenges are encountered during the initial stage of planning and defining which data qualities to evaluate. In other words, they occur at the beginning of the data quality management process defined by the standard ISO 8000-61, as illustrated in Figure 1. Without the planning stage, it becomes harder for a practitioner to develop the right data quality rules and select the tools to enforce them.

### 2.5 Data quality planning precedes implementation

Our paper aims to support ML practitioners and data managers at the planning stage of their data quality journey. We identify a series of considerations that can help them to define their own requirements and data quality strategy. By understanding the requirements that exist, practitioners can be better positioned to select the most meaningful data quality control, assurance and improvement steps for their use case. Although the tasks of implementing specific data quality measures, evaluation criteria and tools for checking data quality are outside the scope of this review, we will mention examples where relevant.

Our goal in this paper is twofold. Firstly, we want to inform practitioners of the data quality requirements and practices that exist and are meaningful in the field of ML. This will be done by synthesising recent academic literature and grouping the recommendations according to the dimensions of data quality that are already familiar to the field of data management. Secondly, to assist readers in selecting a smaller set of data quality practices that may apply to their use case, we map the recommendations onto specific stages and stakeholders in the ML development pipeline. In doing so, we hope to make it easier for organisations and individuals to prepare their data management routines for ML and to anticipate some of the scenarios that may arise at each stage of the ML development pipeline.

## 3 METHODOLOGY

Our literature review was conducted using a systematic mapping protocol [54] in order to select a small set of relevant articles from the much larger collection of literature emerging at the intersection of data quality and ML. Below we present the research questions, inclusion criteria and search strategy that were used to select articles for review. We analysed the selected articles using thematic coding, which revealed additional themes related to the development stages of ML and the scope occupied by data quality management in the wider ML literature.

### 3.1 Research Questions

Our review aimed to identify and discuss the data quality requirements that are important to ML development, and how they differ from more established data management practices. For this purpose, we defined the following research questions:

- Where do the data quality requirements of ML sit in relation to traditional data quality frameworks from data and information management?
- Does ML present any new challenges that are not yet accommodated by traditional data quality frameworks?

The above questions deal with data quality management planning, as opposed to implementation. This is a distinction that has previously been recognised in industry standards such as ISO 8000-61, as depicted in Figure 1. The planning stage (1) deals with the identification of data quality requirements and strategies for implementing them, while the implementation stages (2-4) are about translating these plans into practical rules and techniques for data quality control, assurance and improvement. This distinction between data quality planning and implementation informed the selection criteria of our review.

### 3.2 Selection Criteria

Our interest in data quality planning (as distinct from implementation) helped to limit the scope of our literature review and make the topic small enough to be discussed in a single paper. Specifically, our targeted papers dealt with philosophical or experiential perspectives on data quality frameworks, as opposed to papers that evaluated specific data management techniques or proposed new solutions

Table 1. Research type facets. Adapted from Petersen et al. [54]. We have shaded in grey the research categories that were targeted by our study.

| Category | Description |
|---|---|
| Validation research | Techniques that are novel and have not yet been implemented in practice. (e.g. experiments) |
| Evaluation research | Practical implementation and evaluation of techniques. (e.g. to identify benefits and drawbacks when applied in industry) |
| Solution proposal | Proposed solution to a problem. This includes new techniques or extensions of an existing technique. |
| Philosophical papers | New ways of looking at existing fields through taxonomies or conceptual frameworks. |
| Opinion papers | Personal opinions on whether a technique is good or bad, or how it should be applied. Such papers do not rely on related work or research methods. |
| Experience papers | Explanations of how a framework has been applied in practice, based on the experience of the author. |

for managing data quality. Our choice of research categories is highlighted in Table 1 alongside the other possible types of research as defined in the systematic mapping protocol of Petersen et al. [54].

Our inclusion criteria were as follows:

- The abstract of the paper must discuss conceptual frameworks for defining data quality requirements in relation to ML, or experiences of how these requirements have been defined in practice.
- The paper was published between 2015 and 2022, in order to provide a contemporary overview.
- The paper is peer-reviewed and published in a journal, conference, or workshop.
- The paper may come in the form of a full-length article, extended abstract, or workshop description.

Our exclusion criteria were as follows:

- The abstract of the paper focuses only on techniques for data quality processing, assurance or improvement, rather than conceptual frameworks for defining the data quality requirements.
- The abstract of the paper only considers the data quality requirements of a specific industry that uses ML (e.g. healthcare, finance, materials science).
- The paper does not contain information about the publisher.
- The paper is an early iteration of a later work (e.g. if a similar workshop was delivered by the same authors multiple times, we selected only the latest version).

There was some overlap between our inclusion and exclusion criteria. For example, many abstracts discussed conceptual frameworks in addition to validating specific techniques, developing new prototypes or sector-specific solutions. We included these papers as long as the the main part of the abstract was generalisable (i.e. discussing data quality concepts that apply to general ML applications, and not focusing only on a specific industry or solution).

### 3.3 Search Strategy

Our literature search strategy consisted of three stages: 1) pre-selected articles that were already known to us, 2) automatic search on Google Scholar and selected conference proceedings, and 3) forward and backward snowballing to identify further papers.

*Pre-selected articles*

We began with a list of six articles [3, 23, 32, 34, 35, 58] related to data quality planning, and in particular documentation, that were already known to us based on our previous work with ML models.

*Automatic search*

We used Google Scholar to search for articles whose title included keywords related to our research questions. Limiting the search only to titles helped to eliminate marginally relevant papers from the results. The results were then filtered by examining the titles and abstracts of the papers. Only those that met the selection criteria were retained.

We began by searching the entire Google Scholar corpus using the query "`allintitle: “data quality" ("machine learning" OR "AI")`". This returned 185 results. We truncated our analysis after examining the first 30 results, as many of them did not meet our inclusion criteria. After examining the abstracts, seven articles were retained [12, 19, 21, 25, 27, 28, 63].

We then conducted searches inside the proceedings of two leading academic conferences in machine learning and human-computer interaction: International Conference on Machine Learning (ICML) and Conference on Human Factors in Computing Systems (CHI). This was done using Advanced searches in Google Scholar, where the "published in" box was filled with the name of each conference. We adapted the search query to each venue's area of specialisation. For example, when searching through CHI proceedings, we used used a slightly more lenient query due to the smaller size of the search space: "`allintitle: data (quality OR "machine learning" OR AI)`". This returned 19 results, nine of which met our inclusion criteria [2, 26, 31, 49, 56, 62, 64, 70, 73]. We also adapted the query for ICML, as the conference already specialises in ML. A search for "`allintitle: "data quality" OR "data management"`" returned 16 results, one of which was identified as relevant [38]. Table 2 summarises each of our search queries, the number of results returned by them, and the number of papers that were subsequently selected for our discussion.

We are aware that there may be other venues with relevant contributions that were not included in our selection.

*Snowballing*

After we started reading and reviewing the papers selected using the above techniques, we came across references to other papers that were relevant to our research questions. Eight papers were identified in this way [5, 9, 11, 36, 48, 53, 55, 57]. These papers were initially chosen based on the descriptions provided by authors who cited them, and then assessed using our inclusion criteria. One further article [53] was identified using a forward search of articles that cited [64], as we were curious about the work of other authors who cited this paper. Our general approach to snowballing was informal. Due to time constraints, we did not conduct a systematic review of all possible forward and backward citations.

While we sought to gather a representative sample of papers, it is important to acknowledge that

Table 2. Number of papers identified in each Google Scholar search.

| Articles published in | Search query | Results | Reviewed | Selected |
|---|---|---|---|---|
| [any venue] | allintitle: “data quality" ("machine learning" OR "AI") | 185 | The first 30 results. | 7 |
| International Conference on Machine Learning | allintitle: "data quality" OR "data management" | 16 | 16 | 1 |
| Conference on Human Factors in Computing Systems | allintitle: data (quality OR "machine learning" OR AI) | 19 | 19 | 9 |

the 32 papers reviewed here are only a small part of the growing number of articles related to data quality in ML that exist in reality.

### 3.4 Thematic coding

After selecting the papers, we read them and extracted information that helped to answer our research questions. Relevant information was recorded for each paper using a spreadsheet with the following groups of columns:

- Basic information about the paper - 5 columns: title, authors, publication venue, year, how the paper was found (e.g. automated search, snowballing or existing knowledge).
- Comments raised by the paper in relation to each of the four traditional data quality dimensions - 4 columns: intrinsic, contextual, representational, accessibility (as described by Wang and Strong [74]).
- 1 column to highlight any unusual data quality issues or requirements presented by ML.

Once we started reading the papers, we found that some of the authors' comments and data quality requirements were targeted to specific stages in the ML development pipeline. For this reason, we added the following set of columns to organise our notes:

- Stages of the ML development lifecycle - 8 columns: dataset use case and design, data collection, data cleaning and pre-processing, data maintenance, ML building, ML verification and testing. ML deployment, ML monitoring (as described in Section 4).

Information about each paper was coded using the 18 columns described above. We reviewed this spreadsheet to synthesise common themes at the intersection of two dimensions: each stage of ML development vs. the four traditional categories of data quality. This is the structure we use to present our results.

### 3.5 Scope of the findings

Before presenting our results, we want to clarify their scope. Although the initial goal of our paper was concerned with theoretical frameworks that can help to define and plan data quality requirements in ML, we also noted down any practical techniques mentioned by the authors. Many of our reviewed papers went beyond data quality "planning" to make recommendations on how data practitioners and managers should prepare their datasets for ML. We did not review these techniques in a systematic manner, as this would merit a separate review of its own. However, we

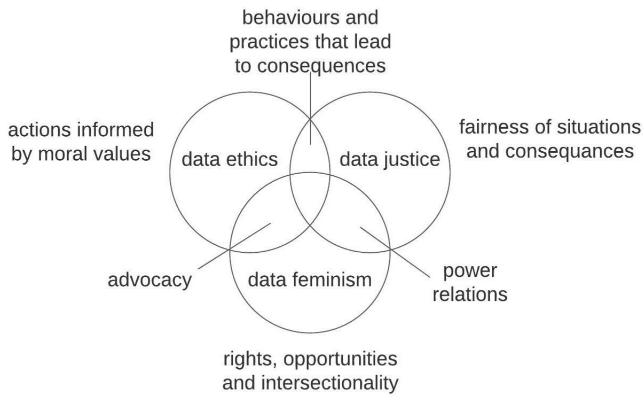


Fig. 2. Venn diagram of fields that complement data quality management.

included some of the techniques in our findings in order to illustrate how data quality plans can be translated into practical specifications, assurance techniques and solutions that apply during stages 2 - 4 of the process depicted in Figure 1.

Besides extending into practical techniques, many of our selected papers discussed topics that went beyond our original focus on data quality in the technical sense. Specifically, they overlapped with other related communities of research and practice in data management, such as data ethics, data justice and data feminism. These fields have historically been addressed by different communities, so the relations between them are not neatly delineated. Nonetheless there is significant overlap which we attempt to illustrate in Figure 2. Rising [59] presented an understanding where justice is about situations and consequences, while ethics is about the actions that lead to consequences. In line with this, data ethics deals with the way that practitioners manage data to ensure privacy, fairness, accountability, security and environmental sustainability [6]. On the other hand, data justice addresses inequalities in the way people are represented and treated as a result of the data that they emit [69]. Data feminism traces the cause of such inequalities to the power relations present in society, and advocates for actions that support the political, social, and economic equality of the sexes, including intersections across other social dimensions such as race and class, sexuality, ability, age, religion, and geography [20].

As illustrated in Figure 2, each of these literatures highlights how systemic challenges in the lived experience of ordinary people are embedded in data, and their potential to be reinforced or mitigated through data-centric technologies. While we did not explicitly search for these perspectives, and time and space constraints prevent us from covering them in the detail that they deserve, we encourage interested readers to investigate these topics separately.

Another scoping challenge which emerged during our review was related to the definition of data. Our initial intention was to focus on observational data (for training, testing and serving models), but this was later expanded due to the substantial attention that our reviewed papers dedicated to the quality of software systems, ML models and their accompanying documentation. Although there is some ambiguity among academics as to whether it is constructive to view software as data [39], we have included the aspects of ML model and documentation quality that emerged during our review. For instance, we found that training data quality can be mediated by software systems (e.g. for data maintenance, or for checking input or output data). Moreover, the inclusion of model and documentation quality helped to highlight the areas where ML model quality is dependent on good quality training, testing or serving data, as well as metadata in the form of documentation. For these reasons, our discussion of data quality grew to include model, software and documentation quality.

## 4 RESULTS

We structure our findings according to the main stages of ML development. Because this is an iterative process that involves numerous decision pathways, there is no single agreed-upon workflow that is universally applicable to every scenario. Nonetheless, a number of commonalities have been identified by researchers.

As early as 1996, Fayyad et al. [22] proposed a sequence of nine stages that constitute the task of knowledge discovery in datasets[8]. The authors suggested that the process typically begins with developing an understanding of the application domain and use case, followed by data collection, preprocessing, and reduction, before moving on to identifying and applying relevant data mining methods, as well as interpreting and acting on their insights. While the authors recognised that knowledge discovery workflows also include challenges related to data accessibility, human-computer interaction, and model scaling, their pipeline focused on the granular steps contained within data mining. A similar focus on data is adopted by the upcoming industry standard ISO/IEC 5259, whose provisional data processing framework in illustrated in Figure 3 (upper) [18].

Recent academic discussions of the ML pipeline have been more detailed in separating out the different stages undergone by ML data. Specifically, they explore model development, verification, deployment, and monitoring, which pose different requirements in terms of organisational and operational considerations [5, 43].

For the purposes of this paper, we organise our findings into a series of stages listed in the first column of Table 3 and illustrated in Figure 3 (lower). Our first five stages (from dataset design to ML building) are adapted from the foundational work of Fayyad et al. [22], and the last three stages (ML verification to deployment and monitoring) are additions derived from more recent literature. We use Figure 3 to anticipate how our terminology maps onto the framework of the forthcoming ISO/IEC 5259 standard.

Earlier publications and standards acknowledge that ML development rarely follows a pre-defined sequence, meaning that data pipelines are difficult to consolidate across different operational contexts. Our stages must therefore not be assumed to occur in a linear sequence. There are a number of ways in which reality may diverge from the stylised view presented in our diagram. The first of these relates to data iteration, where the steps of model building and testing are frequently followed by the need to collect new data, or enriching the existing dataset [5, 11, 31, 34]. Other scenarios that are becoming increasingly common are multi-dataset-multi-model pipelines, where existing ML models are used for pre-processing data or training new ML models [5]. We will flag these scenarios when we discuss our findings in the subsections below.

While time and space constraints prevent us from anticipating every possible workflow that may occur in reality, we illustrate a simple example of an ML data quality pipeline in Figure 4. Additionally, we use Figure 5 to illustrate a more specific scenario of an ML application trained on text data that might involve multiple data sources and multiple models. The purpose of these diagrams is to show how different aspects of data quality assurance can map onto different stages of the ML development process. This is not an exhaustive view and we encourage readers to be critical in evaluating how the data quality requirements discussed below would apply to their own non-linear cycles of dataset development.

[8]Within the scope of knowledge discovery, the specific role of ML is to provide the data mining methods that help to discover new knowledge in the form of approximations, predictions or observable patterns.

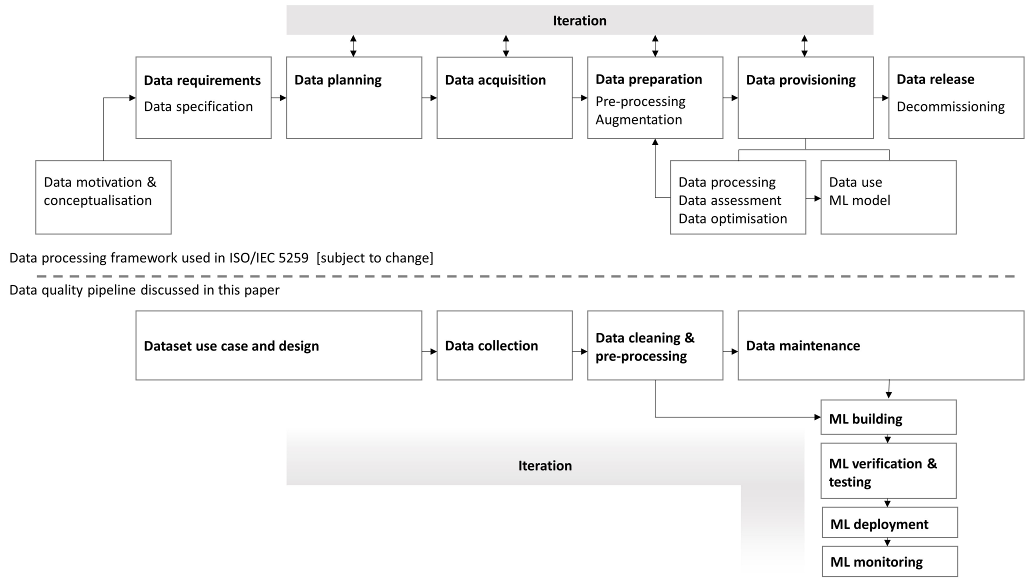


Fig. 3. An illustration of how our data quality pipeline (lower) maps to the data processing framework used in ISO/IEC 5259 (upper). Diagram adapted from Chang [18].

We would also like to highlight that the data quality requirements described here should be viewed as desirable rather than essential. It is unrealistic to expect them to be achieved in their entirety, especially where practitioners have competing priorities such as time and cost. It is also common for data management capabilities to change and mature throughout the progression of a project [7]. So readers should treat the information reported here as aspirational rather than prescriptive.

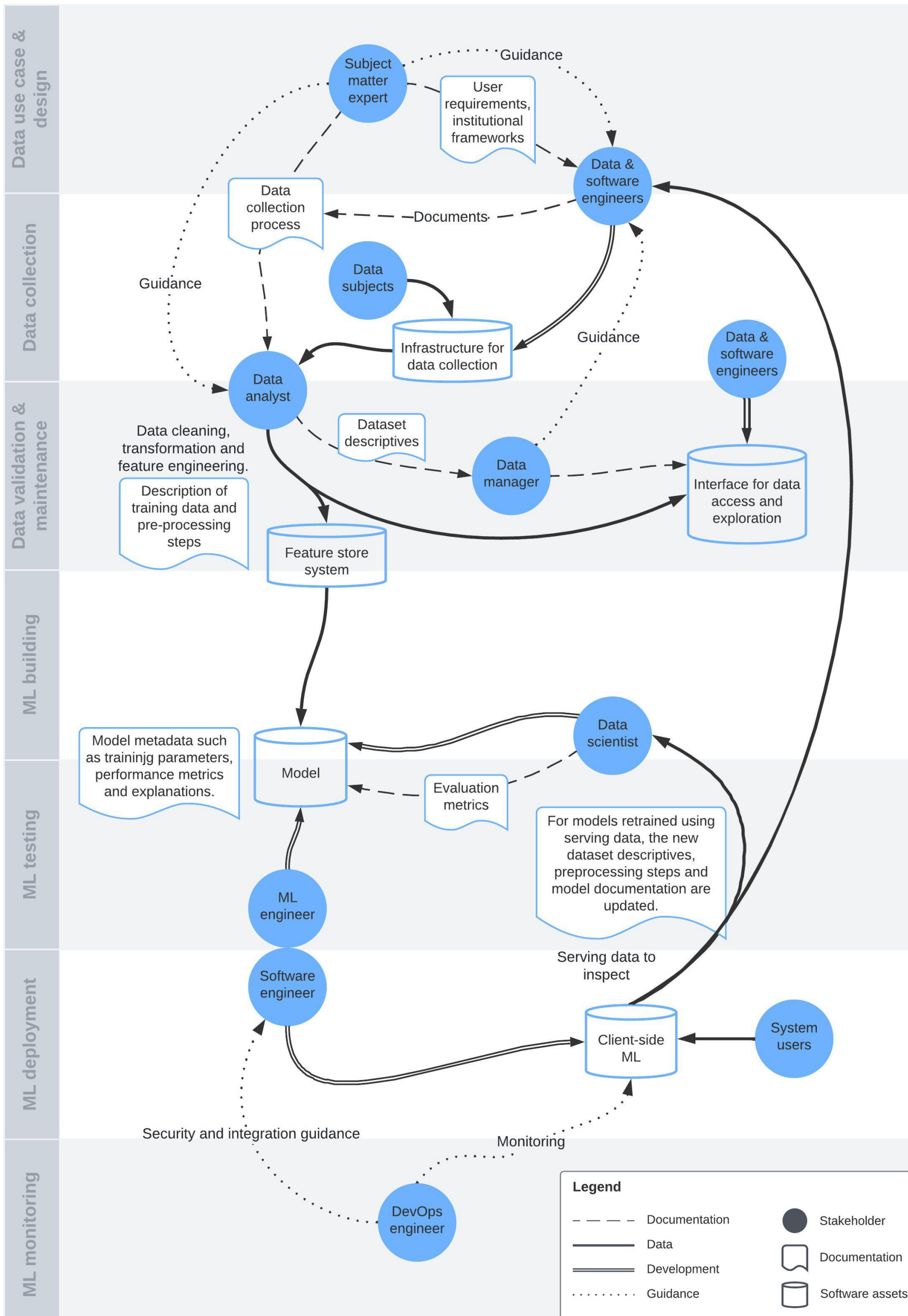


Fig. 4. ML data quality pipeline example.

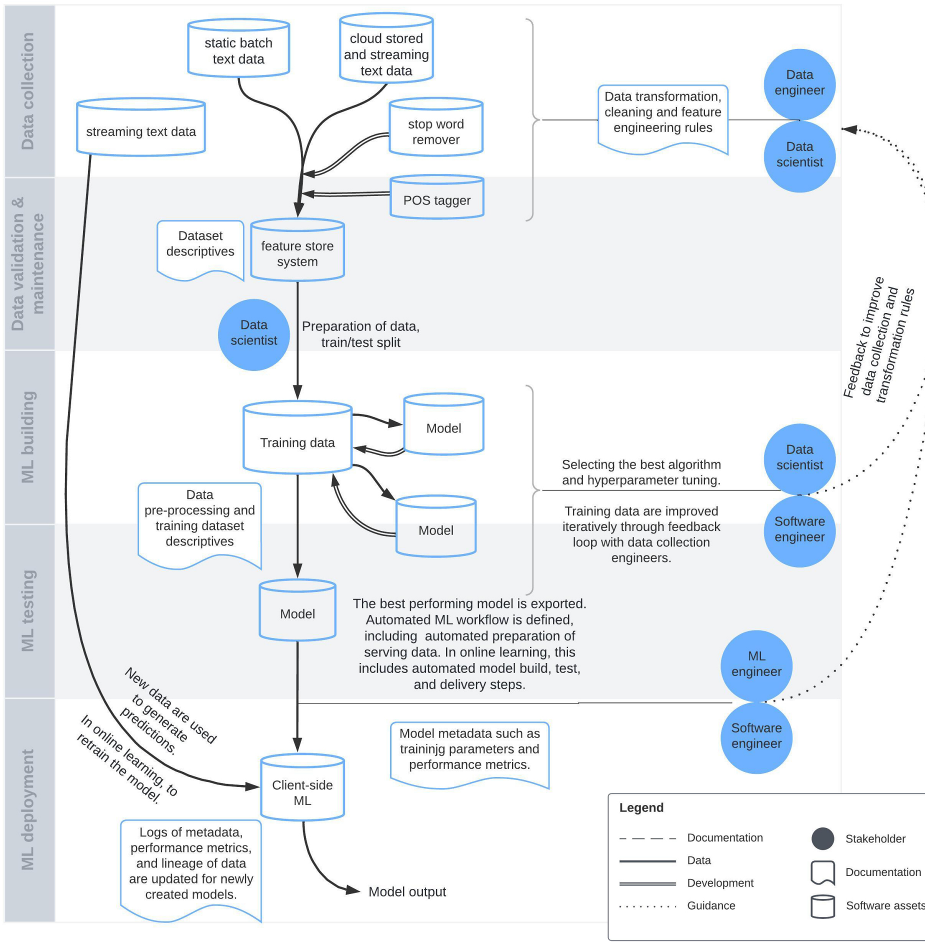


Fig. 5. Example pipeline for multi-model-multi-dataset scenario.

Table 3. ML data quality considerations classified according to different categories of quality (horizontal) and stages of the ML development pipeline (vertical).

| Development stage | Data quality category | | | |
|---|---|---|---|---|
| | Intrinsic | Contextual | Representational | Accessibility |
| Dataset use case and design | Accuracy of data can be supported by hiring human annotators and field experts in advance. [49, 52, 53] | Relevance of data can be ensured by determining what features are required in advance. [9, 36, 53] | Clarity and credibility of the metadata can be improved by including documentation on user requirements and dataset design. [35] | Availability of data can be supported by infrastructure for data collection and management (particularly in large organisations). [25, 52, 57]<br><br>Validity of data for online learning can be assured by putting in place runtime verification tools. [21, 50] |
| Data collection | Accuracy can be improved by:<br>• Human-in-the-loop approaches for data labelling and augmentation. [49, 73]<br>• Data collection tools that raise actionable alerts to warn users of unexpected values in advance.[38, 57]<br>• Screening and training of data workers. [49, 70, 73] | Context coverage can be supported by institutional guidelines on potential power imbalances, ethics and inclusivity.[9, 36, 62, 70] | Clarity of the metadata can be supported by documenting the data collection process (e.g. using datasheets, checklists). [9, 23, 35, 48, 58]<br><br>Consistency of data can be improved using standardisation. [25, 33] | Regulatory compliance can be supported by institutional frameworks and procedures for consent, transparency, ethics and privacy. [9, 36, 70] |

Table 3. ML data quality considerations classified according to different categories of quality (horizontal) and stages of the ML development pipeline (vertical).

| Development stage | Data quality category | | | |
|---|---|---|---|---|
| | Intrinsic | Contextual | Representational | Accessibility |
| Data cleaning and preprocessing | Uniqueness of data entries and features can be improved by removing redundant cases and reducing the complexity of the features. [5, 38, 57]<br><br>Completeness can be supported by automated pre-processing and ML aids for augmentation and annotation. [5, 27, 38, 70] | Contextual bias can be detected using ground-truth correlations. [32, 52, 53]<br><br>Contextual validity can be improved by balancing the classes and measuring how well the dataset fits the real-world problem. [3, 5, 8, 9, 25, 38, 70] | Clarity of the data pre-processing sequence can be improved using documentation and publication of code. [52, 72]<br><br>Consistency of data sourced from heterogeneous sources can be supported by reformatting standards, normalising and aggregation. [38, 70]<br><br>Precision can be improved by using representational standards that allow for uncertainty. [3, 5] | Security of sensitive data supported by anonymisation [33, 70] |
| Data maintenance | | Contextually biased data can be improved using curation, including infrastructure, tools, and practices for maintaining nonstatic datasets that grow over time. [3] | Maintainability at scale is supported by standards. [3]<br><br>Clarity of the dataset can be supported by user interfaces for dataset exploration. [25, 32, 33, 52, 57]<br><br>Clarity of the metadata can be supported by documentation on:<br>• data content (e.g. nutrition labels) [25, 32]<br>• maintenance plan [36]<br>• mission statement[36] | Availability of data can be facilitated by infrastructure for differential access and sharing (e.g. via data trusts). [32, 35]<br><br>Identifiability of the correct dataset (out of multiple versions) can be guided by version control and DOIs. [25, 32, 35, 57] |

Table 3. ML data quality considerations classified according to different categories of quality (horizontal) and stages of the ML development pipeline (vertical).

| Development stage | Data quality category | | | |
|---|---|---|---|---|
| | Intrinsic | Contextual | Representational | Accessibility |
| ML Building | Uniqueness of features supported by dimensionality reduction. [5, 57]<br><br>Completeness of data improved by enrichment. [33, 34, 57] | Contextual validity supported by selecting the right features. Contextually biased data can be improved by re-sampling or re-weighting the training distribution. [5, 16, 25, 57] | Clarity of the ML building process can be elucidated using model reproducibility checklists [35, 55] and by embedding structured meta-knowledge into the documentation [56].<br><br>Clarity of model performance can be supported by documentation on evaluation metrics and statistics. [31, 48] | Availability of code and model data can be supported by publication, in addition to the above steps. [32, 35, 55] |
| ML verification and testing | | Contextual fit of the model can be assessed using benchmarked evaluation in different conditions/scenarios. [5, 48] These evaluation metrics must be verified by checking the overlap between training and test datasets [19]. | Clarity of model performance results can be improved by model cards, including contextual evaluation results. [48]<br><br>Transparency of model can be supported by sensitivity testing and explanations. [1, 2, 35] | Availability of test data (real or synthetic) made possible by sharing. [48]<br><br>Security of restricted training data can be assured by adversarial testing for data poisoning, model stealing and inversion. [5, 48] |
| ML deployment | Validity of serving data can be ensured by following the data preparation rules of the original model, and by checking for representational drift. [57, 65] | Contextually sensitive ML options include client-side ML (federated learning). [33] | Interpretability of the model output can be supported by explanations. [1] | |

Table 3. ML data quality considerations classified according to different categories of quality (horizontal) and stages of the ML development pipeline (vertical).

| Development stage | Data quality category | | | |
|---|---|---|---|---|
| | Intrinsic | Contextual | Representational | Accessibility |
| ML monitoring | | Fidelity of the model in evolving contexts can be monitored by checking the distribution and features of data fed into the model. [3, 57] | | Security of restricted training data can be assured by monitoring for adversarial attacks. [5, 48] |

### 4.1 Dataset use case and design

The initial steps to ensuring data quality begin before data are collected. These steps include clarifying the use case for which the data are sought and investigating the operational and/or infrastructural requirements of gathering the data. These preparatory steps must be recorded in the dataset's documentation, in order to inform current and/or future colleagues about the requirements of the use case.

It is common that the precise data quality requirements will not be known upfront, and new use cases may emerge as the model matures. This means that practitioners will likely need to return iteratively to the dataset design and data collection process [31]. In cases where additional data are required but cannot be collected iteratively, other methods are available to enhance the dataset, as we will discuss later. With this in mind, the preparatory steps described below should be viewed as a desirable rather than essential part of the data quality pipeline.

#### *4.1.1 Use case documentation.*

The definition of ML data requirements must begin by consulting with relevant stakeholders [35]. Those who are commissioning the system should be consulted to understand not only the problem that the ML needs to address, but also the anticipated characteristics of the end users (e.g. demographics, cultural and environmental context). This information can support the acquisition of training data that are representative of the population of interest, thus increasing the likelihood that the output of the ML system will match their needs [9, 36, 70]. Some questions that ML researchers may want to consider include asking how much supervision, domain expertise, and specialisation would be needed to collect and label data for the scoped project [36].

The careful analysis of requirements prior to data collection, as recommended above, is different to the data collection practices that are typical of contemporary ML implementations [36]. Our recommendation reflects an "interventionist" approach, which contrasts with minimally supervised data collection methods such as Web crawling and crowdwork that have traditionally been used to generate large volumes of data. The problem is that these approaches do not typically evaluate the origin, motivation, platform, or potential impact of the gathered data. This has been flagged as one of the causes of historical and representational bias in ML systems that use those data, with numerous authors urging for slower and more methodical approaches to data collection [36, 53]. This includes the recruitment and training of data workers, as they are an integral part of how ML data come into being [73].

Another issue that can get overlooked with big data is the interrogation of assumptions about which questions are answerable with certain data attributes in the first place. For example, Paullada et al. [53] draw attention to studies that attempted to predict personal attributes from photos of human faces, under the false assumption that these predictions are possible and worthwhile to make. Careful documentation of the use case and underlying assumptions about relevant data attributes can help practitioners and organisations to avoid collecting data signals that may subsequently get discarded.

#### *4.1.2 Data availability and coherence.*

Once the use case and requirements for a dataset are known, it is important to conduct further checks into the availability of the required data. Whereas discussions in traditional data management tended to focus on static datasets that were already accessible to practitioners, common ML use cases include big data and real-time analytics where data reside in multiple storage systems characterised by streaming, heterogeneous and cloud-based data [52, 57]. Data that are dispersed across multiple sources tend to have different schemas and approaches to storage and access [33, 52]. This can lead to difficulties in discovering what data are available, their structure and how to parse, query or store them, which complicates the task of integrating information into a single dataset suitable for ML. Several authors have therefore noted that traditional data quality approaches designed for relational and static datasets may not be sufficient when dealing with the kinds of large-scale decentralised ML pipelines that are increasingly being used for operational and organisation-wide decision making [25, 57].

As a result of the above, managing data quality in industrial use cases may require new infrastructure that ingests data and converts them into a form that is more compatible with the ML trainer [57]. This may involve the creation of data warehouses to extract, clean, transform, and integrate data. For instance, Paleyes et al. [52] discuss how Data Oriented Architectures (DOA) can help to make data flowing between elements of business logic more explicit and accessible, simplifying the tasks of data discovery, collection and labelling.

For real-time applications, runtime verification techniques can help to deal with data that arrive continuously and where models are trained continuously. This form of "online learning" requires continuous monitoring to correct data quality issues on-the-fly and ensure that they are within acceptable bounds to match the assumptions of the respective ML model [21]. This may include checking that the operational input distribution is similar to that represented by the original model, to avoid issues of distribution shift [50].

Besides technical infrastructure and tools for data quality assurance in online learning, some use cases may also require additional human resources for data labelling. Access to human annotators and field experts may be a particularly significant bottleneck in data labelling tasks, such as those in medical fields [52, 70]. Here, the framing of tasks, labour conditions, and legal issues pertaining to data collection and distribution will need to be investigated as part of the technical and institutional infrastructure that precedes data collection [53]. For example, Mitra et al. [49] and Thakkar et al. [70] discussed the importance of preparatory measures in the form of screening and training of data workers, with Mitra et al. [49] finding that this preemptive approach produced better quality data than what would typically be achieved through automated post-processing of noisy data.

## 4.2 Data collection

Once the data use case and operational requirements are in place, the process of data collection can start. The design decisions made in the previous step may be implemented in a number of ways, such as through software systems, annotator guidelines, and labelling platforms. Below we discuss the ways in which documentation, standards and interfaces can support the acquisition of data that are high in quality.

### *4.2.1 Data collection documentation.*

The data collection process should be documented as early as possible during task design [53]. Numerous authors have shared templates on how to structure the documentation. This includes datasheets [23, 35], data statements [9], and checklists [58]. These documents are intended to help dataset creators to become more intentional and reflective about their data collection objectives, underlying assumptions, implications of use, and stakeholder values as they work. Benefits to this include an improved understanding of the dataset's contextual validity, by asking questions about how the dataset instances or sampling approach can be made more reflective of the larger population (e.g. in terms of geographic or demographic coverage) [23, 70], or application context [48].

For consumers, documentation about data collection methods provides the information needed to make informed decisions about using a dataset and to avoid unintentional misuse [23, 25]. It supports users in deciding whether the data are comprehensive enough for their use case [19]. In some cases, the documentation may reveal assumptions that would not be readily apparent from basic metadata or dataset content [35]; for instance, a recent crawl of old news articles would benefit from a statement to explain that the time of data collection is different from the original time of creation of the data values.

Besides understanding the dataset, some documentation frameworks are designed to equip downstream practitioners with the transparency needed to repeat the data collection process (e.g. for the purposes of gathering alternative datasets with similar characteristics, auditing or repeating an experiment in different contexts) [23, 53, 58]. Documentation methods of this kind have been particularly encouraged in sociocultural data collection mechanisms, such as crowdsourcing, where data workers are hired worldwide to read texts, view images and video, and label the data that are used to develop ML models. This means recording operations related to sampling, mapping experimental conditions to micro-tasks, and ensuring quality contributions from participants [58].

In this vein, data users are beginning to assess quality not only in terms of the characteristics of the data (e.g. accuracy), but also the working conditions, skills and aspirations of the individuals who annotated those data [73]. Authors in the field of HCI envision that as ML practitioners respond to the push for better documentation, this creates an opportunity for data labour practices to also be documented and reviewed. To this end, Rothschild et al. [62] propose that crowdsourced ML datasets can be accompanied by a cover sheet that describes the precise hiring and employment practices. The intention is to encourage requesters to create institutional norms around just and respectful employment for data workers.

#### *4.2.2 Data collection standards.*

As noted in our earlier discussion of data use cases, a major challenge in ML data collection in industrial applications relates to data heterogeneity, which can be manifested as unstructured, semi-structured, and structured data of disparate types [25]. During the data collection process, the user requirements established in the previous step (dataset use case and design) need to be translated into common standards that allow datasets to be linked and that capture the necessary information. For example, streaming data from the web may need to be filtered and converted to a more structured format, while data from IoT sensors may require standardised semantics to capture the types of equipment used, as well as accommodating uncertainty around measurements.

#### *4.2.3 Data collection interfaces.*

One of the novel aspects of production ML is that data collection is automated rather than manual (e.g. data arrives continuously from sensors or web applications). In cases like this, part of the responsibility for ensuring good data quality lies with software engineers, who can design systems that generate actionable alerts to inform users of potential data quality issues (e.g. if a feature is missing or has an unexpected value) [57]. Other examples of data collection interfaces can take a more creative format, such as data collection games. However, Gundry and Deterding [26] found that such interfaces can present a trade-off between participant enjoyment and data quality, where games elicited more enjoyment but led to less accurate data compared to an equivalent control.

### 4.3 Data validation and maintenance

Once the data have been collected, they typically undergo a process of checking and cleaning before being usable for an ML system. This stage of the ML development pipeline bears a large bulk of the activities related to data quality assurance. Below we discuss these tasks, which include pre-processing, validating the contextual coverage of the data, data quality metrics, user interfaces for inspecting data, dataset accessibility and maintenance over the longer term.

#### *4.3.1 Pre-processing.*

Data collection is often followed by pre-processing tasks such as feature selection, deduplication, removal of outliers, consistency checking, anonymisation and imputation of missing values [52, 70]. As was done during the data collection step, information about the preprocessing steps should be recorded in the dataset documentation. This helps subsequent data consumers to determine whether the data are readily compatible with their chosen task or if they need to undertake further transformations (e.g. dimensionality reduction, bucketing, tokenization, removal of instances, normalisation etc.) [23, 57]. Another aspect of data composition that can be useful to inspect and report in some ML use cases relates to potential dependencies that may exist between features, where information leakages between variables could later cause the trained ML models to produce unrealistically accurate predictions during testing [57].

While the nature of the above work is not unusual in relation to longstanding data management literature that has dealt with validity, consistency and integrity concerns, literature from the field of ML has highlighted constraints in the order in which the data preprocessing tasks should be executed. Differences in the sequence of data pre-processing steps have been found to produce radically different ML results (e.g. correcting the data for missing values using imputation can affect outliers in the dataset)[27]. Given that the search space of all possible sequences of data pre-processing tasks is combinatorially large, some authors have proposed algorithmic solutions for establishing the optimal pre-processing pipeline[11]. Others have drawn attention to formal ways of establishing and treating the reasons behind problematic data. For example, Bertossi and Geerts [12] suggest that explainable AI techniques can be applied to identify the features that cause inconsistencies in data and use this information to predict the best repair actions.

But even where formal data cleaning techniques have not been used, data practitioners can still take care to document their actions where possible (e.g. using pre-defined protocols or ex-ante publication of reproducible code that was used to prepare the data). One possible way of doing this is through the use of interactive notebooks to weave together code and documentation [72]. Data validation routines and publication of pre-processing code is particularly valuable in contexts where data preparation is decoupled from the ML pipeline, providing more transparency and opportunities to detect bugs, feedback loops, or changes in data dependencies [52].

#### *4.3.2 Data context and coverage.*

The period after data collection is a good time to re-evaluate contextual characteristics of the dataset and the degree to which they align with the intended use case. In sociocultural data, important factors to explore could include cultural biases related to gender, race, ethnicity, or religion [9, 25]. Guidance on which protected characteristics to look out for can be found within practical toolkits such as “AI Fairness 360” [8], and checklists can be used to document such information to ensure legal and ethical compliance [60].

Additionally, practitioners should consider the possibility that some variables captured in a dataset may not explicitly refer to demographic groups, but still contain stereotype-aligned correlations [32, 53]. For example, variables such as wages or location may be strongly correlated with specific populations in a given region. To surface these kinds of relationships, practitioners may need to compute comparisons to variables from other datasets considered to be "ground truth", such as Census Data [32]. In use cases that do not capture human data, it may also be useful to evaluate the variance of data in capturing different environmental contexts, such as the environment in which autonomous vehicles are trained in the lab and how it may differ from situations in the real world [52].

While some of the contextual biases described above may be detectable in the existing data through effort, others may become clear only once the dataset is deployed through ML in production. This is especially true of unstructured data (e.g. text, images) where the features are opaque and difficult to inspect. In cases like this, it is important to document the populations from whom the data originate. Numerous authors have observed that ML systems perform better for users whose demographic characteristics match those represented in the training data [9, 36]. The contextual origins of datasets must therefore be recorded in the documentation as a means to preempt scientific and ethical issues that may result from the use of data from certain populations to develop ML technology for other populations. Bender and Friedman [9] provide examples of data statements for NLP datasets, which can be used to provide the context needed by developers and users to better understand how the subsequent ML results might generalise, how best to deploy the software, and what biases might be embedded in it. For datasets that originate from crowdworkers, it is important to additionally report any potential sampling and selection biases, as well as response bias, design bias and ethical integrity aspects (e.g. informed consent, minimum wage), that will allow the experimental setup to be traced or reproduced where necessary [48, 58].

#### 4.3.3 *Data quality metrics.*

In addition to the qualitative descriptions of dataset use cases, collection and pre-processing steps discussed earlier, during the data maintenance step it is beneficial to include quantitative metrics about the dataset. Several generalised and context-specific frameworks have been proposed for this in the literature.

Holland et al. [32] developed a web-based "dataset nutrition label" that comprises of seven modules to display general aspects such as metadata, provenance, variables, statistics, pair plots, probabilistic models, and ground truth correlations. In contrast to this standardised approach, Gudivada et al. [25] recommend metrics that are more task-specific. For example, the data quality metrics that matter most in classification tasks are proposed to include class overlap, outliers, boundary complexity, label noise, and class imbalance. Regression tasks, on the other hand, benefit from data quality metrics regarding outliers and missing values. This suggests that data practitioners who are responsible for maintaining the dataset may need to refer back to the anticipated ML use case in order to decide which metrics would be most meaningful to consider and report.

#### 4.3.4 *User interfaces.*

Besides quantitative metrics, the above proposals for data quality metrics have also advocated for the use of dashboards and visual aids for data inspection and sanity checks (e.g. min max values in continuous data, distribution of categorical values) [25, 32, 52, 57]. Holzinger [33] highlights that "at the end of the pipeline there is a human, who is limited to perceive information in dimensions. It is a hard task to map the results, gained in arbitrarily high dimensional spaces, down to the lower dimensions." To this end, interactive software tools can help users to explore the data through pair plots, distributions, correlations, histograms or heatmaps, and evaluate their suitability for certain demographics or other criteria.

#### 4.3.5 *Accessibility.*

Maintaining a dataset after its creation can present a number of accessibility questions, especially for personal or commercially sensitive datasets whose disclosure could pose risks to privacy, security, or intellectual property [32]. Before publishing, data managers will need to determine the usage affordances of the dataset, its policies and designated owners [35]. Specific mechanisms may need to be identified for achieving good data availability while simultaneously protecting them from unauthorised access (e.g. by defining user entitlements to data access, including metadata containing licence type and DOI) [25, 32, 35]. One possibility here is the use of specialised infrastructure (e.g. data trusts) that allow for secure data storage, retrieval and purging mechanisms between trusted parties. In cases where direct access to data is not possible, proxy metrics such as the data "nutrition label" described earlier may provide sufficient information for auditing and accountability purposes [32].

#### 4.3.6 Maintenance.

Datasets will require governance standards and specifications to support their maintenance, especially in larger organisations that handle multiple datasets [25]. This documentation should include information about the conventions used for naming and organising the data, their meaning, source and version history [32, 35, 57], as well as specifying the complex relationships that may exist between multiple data sources.

For datasets that deal with contextually significant data (e.g. from specific geographic regions, populations or industries), data managers may have an interest in maintaining them in ways that help to address data coverage issues over the longer-term. This can involve the establishment of open repositories and data trusts with the goal of gathering more representative data [36]. As part of this, data managers can develop "mission statements" to communicate their curation goals and encourage external contributions that can make the collection more contextually representative in future.

### 4.4 ML building

In many contexts, the previous data collection and preparation steps are likely to have been carried out by a person different to the one who builds the ML model. For this reason, the ML practitioner would ideally go back and check the dataset's documentation to make sure that it meets their use case requirements. This can help them to avoid using the data for a purpose that may be morally or ethically objectionable to the original curators [53].

Once the dataset is confirmed to be suitable, the process of building ML can begin. Some of the initial data work may be similar to the data pre-processing stage mentioned earlier, but here the requirements will depend to a greater extent on the selected ML techniques and use case. Examples of possible tasks include feature selection, enrichment and sampling. We summarise these requirements below, followed by a discussion on data accessibility issues that accompany ML models.

#### 4.4.1 Feature selection.

During the initial development of a model, an important part of data preparation involves selecting or engineering a set of features that are most predictive of the outcome [57]. This includes removing redundancies (e.g. correlated variables) or using dimensionality reduction methods (e.g. PCA) before using the data as model input. However, preparations of this kind are not always feasible with unstructured data such as images, language, and video, where high dimensionality and large sizes make it hard to identify relevant features from the outset [34]. Some work on feature selection may therefore be put on hold until ML models are more mature, where the focus shifts from preparatory steps on the incoming dataset towards ex-post feature selection as a way to optimise resources and reduce latency while still retaining the same accuracy in the model.

#### 4.4.2 ML-informed data pre-processing and enrichment.

Once relevant features are selected, ML developers may need to re-examine data quality challenges related to contextual coverage and cleanliness, where the limitations of the dataset may need to be mitigated through enrichment and/or sampling approaches before feeding them to the model. Below we discuss each of these processes.

Exploration into data coverage that was initiated at the data collection and pre-processing stages should continue during the ML building process. In particular, ML practitioners should be mindful that it is not always possible for the preceding data handlers to obtain a priori knowledge of potentially sensitive features (e.g. gender, race), especially in high dimensional data such as images, language, and video. In cases like this, ML in itself can become a tool for detecting smaller subsets of data that would most benefit from enrichment or using modeling choices to mitigate bias [33, 34]. In the case of enrichment, the first step is to contextualise the available data, and then augment the existing features with new signals from other datasets or acquire new labels [57]. Solutions of this kind have been applicable in contexts such as gender biased text data, where authors have proposed the use of further data collection and improvement steps, such as crowdwork and debiasing algorithms, to identify and remove discriminative word mappings from training data [14].

Besides enriching the available data, another solution for creating contextually relevant datasets involves sampling. Such practices target the dataset's representativeness, rather than size, as the quality that will influence the performance of an ML model. Several authors have noted that a small number of representative observations can be more effective than using an extremely large but biased dataset [25]. Indeed, using all available data to train models can sometimes have a detrimental effect [28].

Examples of this have been especially prominent in research that deals with imbalanced datasets, where the outcome of interest is under-represented in the observation space (e.g. fraud detection, clinical diagnosis). Here, techniques such as under-sampling and synthetic data have been found to enhance model performance [16]. Others have proposed that training datasets should be filtered in other contexts that deal with human behaviour. For example, Hagendorff [28] propose to single out data from certain subpopulations that are deemed more competent, eligible, or morally versed for a specific task.

One of the downsides of re-sampling approaches is that they can be costly to implement and require the practitioner to know in advance which features are responsible for the undesirable bias [34]. To this end, some authors have proposed algorithmic approaches for identifying subsamples of training data that are most effective at meeting the desired model metrics (e.g. log loss, AUC, and calibration) [57].

In addition to mitigating bias, ML tools can also be used to enhance the cleanliness of datasets for specific models. As mentioned during the pre-processing stage, automated techniques can be used to select the optimal sequence of data preprocessing tasks that maximise the performance of the ML model [11].

#### 4.4.3 *Multi-dataset-multi-model scenarios.*

Another common scenario involves practitioners reusing existing ML models as part of their data pre-processing steps, or relying on an existing ML model as a starting point to train a second model for a new domain. These scenarios have implications for data quality because they determine part of the context to which the data quality needs to be tailored.

For example, in the NLP domain, it is common to reuse tools such as part-of-speech (POS) taggers, dependency parsers, and pre-defined stop word lists to prepare the data for subsequent use in a model. To do this, the practitioner will typically need to prepare their text data by removing special characters and tokenising the string into a list of words that can be read by the pre-processing tool.

In other cases, model reuse forms a more substantial part of the ML development process. This is common with complex models that could take weeks of computation on multiple machines, where using existing models as a starting point can save valuable time and resources when training a second model. For example, a convolutional neural network (CNN) trained on human faces that already has the capacity to extract the main features (e.g. eyes, noses, etc.) can prove more efficient than training a new CNN from scratch [5]. This is termed "transfer learning" in the literature, and it typically means using one of a few "foundation models" created by large organisations that had access to huge data and computational power [15].

An important data quality challenge here relates to knowledge about the data on which the model was trained, and the data used to evaluate the model. For example, duplicate entries in a dataset can produce an overlap between the datasets used to train and evaluate a model, which can cause the performance metrics to be exaggerated [19].

Another issue to consider is the extent to which the original model's intended usage matches that of the new application. Foundational models that are built to be generalisable can come at the expense of specificity. For example, their training data may not sufficiently capture an operational context that is characterised by specific demographic or cultural traits. In cases like this, reusing and tuning a trained model helps to improve model performance only if the tuning is done using a dataset that contains task-specific data entries [19]. Some authos have called for smaller reusable models that are trained on contextually-relevant, rather than large, datasets [10].

#### 4.4.4 Documentation.

Where possible, the ML building process should be accompanied by documentation that has all the necessary information to reproduce or verify the model [35, 55].

This includes defining the metrics and statistics used to evaluate the model, as well as reporting the measures of central tendency (e.g. mode, median and mean) and uncertainty around observed effects (e.g. range, quartiles, absolute deviation, variance and standard deviation) [48]. Documentation practices at this stage can also provide an opportunity to examine and reflect on the data properties that significantly affect the model accuracy, and whether there are any dependencies to other data and infrastructure that may affect the outcome [57]. Besides the model results, the documentation at should also report the provenance of the model (e.g. who developed it, potential conflicts of interest, when it was developed, versioning etc.) [48].

The above information can come in the form of separate documents, or as comments and variable identifiers embedded in the code. Pinhanez et al. [56] preset an example from the field of conversational systems where practitioners have tended to structure their documentation in a manner that is readable by machines. The authors discussed how documentation of this kind can have its own computational value when building new tools to assist the developers.

Besides assisting collaboration between ML developers, the documentation also provides an opportunity to disclose decisions and facts that can be used by the broader community to better understand what the model does [48]. As with dataset maintenance, the model documentation should also be accompanied by versioning information and DOIs, which could be done through institutional repositories or other open platforms where the model itself or its metadata are housed [35]. In commercially sensitive settings, the level of disclosure may be tempered by the requirement to protect intellectual property rights.

#### 4.4.5 Accessibility.

In contexts where openness is possible, a growing number of research venues are encouraging ML practitioners to publish their models for the purposes of review and verification (e.g. checking experimental conditions, hyperparameters, proper use of statistics, robustness), as well as supporting the replication of existing models in subsequent innovation and research. Structured guidelines for sharing ML models can be found in reproducibility checklists, such as the one proposed by Pineau et al. [55]. Such checklists cover both the accessibility of model code as well as training data.

In the publication of code, practitioners in industry may first need to ensure that their applications do not contain software that is protected by intellectual property, or is built on top of proprietary libraries. Although this is an important consideration, prior research has observed that many authors from industry were indeed able to submit code [55]. In cases where the model cannot be shared at all and practitioners still want to provide access for model verification and review, they can share minimal information on model performance across various factors [48]. One way of doing this is to use "model cards", which are short documents that describe model evaluation procedures and results across different settings that are relevant to the intended application domain [48]. We will elaborate on these procedures in the next section. Additionally, models that use decision thresholds can include a threshold slider in the digital documentation that accompanies a model [48], allowing users to view performance parameters across different decision thresholds.

With regard to the publication of data, ML practitioners are typically encouraged to share the training and test data that underpin their model. However, this presents a challenge to ML models that rely on commercially sensitive or personal data (e.g. in healthcare or finance). For cases like this, synthetic training and test data can be generated using distribution hypotheses from the original data [32], or complementary empirical results can be provided using open-source benchmark datasets in addition to results based on the confidential data [55]. ML practitioners should also be mindful of using and distributing training data that come from unknown sources; this includes benchmark datasets scraped from the Web, whose licensing and copyright restrictions are unclear, or datasets that may have become deprecated [53].

## 4.5 ML testing

In many documented cases of adverse ML outcomes, the issues with training data became apparent only after the solution was deployed in real-world contexts. In order to avoid this, ML practitioners and auditors can test the system for contextual bias and security issues before releasing the system. We discuss these considerations below.

### *4.5.1 Performance metrics and explainability.*

The evaluation metrics for ML models have traditionally focused on generic cues such as information loss, false positive and false negative rates. However, more recently researchers have started encouraging practitioners to develop context-specific criteria that rely on specific types of test data. For example, to assess the contextual coverage of an ML model, its performance can be tested in different demographic and intersectional groups (e.g. by age, race, gender, geography) [48]. This is particularly important in cases where protected attributes may be underrepresented in the training dataset, prompting fairness concerns [34]. When deciding which factors to present in the intersectional analyses, practitioners must be cautious to preserve the privacy of individuals; this can be done through collaboration with policy, privacy, and legal experts to decide which groups may be responsibly inferred, and how this information can be stored and accessed [48]. For practitioners who are struggling to find test data for populations outside of the initial domain used in training, possible solutions include using synthetic datasets to represent use cases that may otherwise go unevaluated [48].

Besides testing performance on different demographic groups, different business contexts may also be relevant to consider (e.g. plant recognition worldwide or in the Pacific Northwest, vehicular crash tests with one or another phenotype in dummies) [48]. This allows stakeholders (policymakers, developers and individuals) to compare models not only based on generic evaluation metrics, but also on social and economic dimensions such as ethics, inclusivity and fairness, making it possible to take remedial action where necessary.

In addition to representativeness, other meaningful metrics might include reflections on model performance in real business settings, for instance by estimating customer conversion rates [52], model size and energy consumption incurred by the model [53]. Additionally, sensitivity studies of dataset parameters can give insight into the features that have an impact on the model's prediction [35]. This does not only help to support transparency and explainability fror data users, but it can also help practitioners to understand the effect that errors in specific features can have on a model's output and performance. This understanding is vital to applying data quality assurance and correction tools [63].

One of the themes emerging from the above authors is that performance metrics need to be tailored to the specific use case of the model. Often this involves trade-offs between traditional evaluation metrics such as precision and recall [25], as well as contextually sensitive issues such as test-set accuracy, robustness and fairness, compactness and privacy, where maximising one performance metric may come at the expense of another [34]. Because of the subjective nature of the model evaluation process, and the various different metrics that practitioners can choose to prioritise, these decisions can be communicated to users using "model cards" that contextualise the results according to different benchmarks that matter in the intended application [48]. As was the case with dataset documentation, the use of visualisations can help to demonstrate cross-sectional analyses of model performance according to different metrics.

Besides performance, model inspection and visualisation methods can also support the interpretability of the model, which can in turn influence its perceived quality [4]. This falls within the field of explainable AI (XAI), which aims to help practitioners and operators to analyse the output of ML models and the reasons behind automated decisions. Possible ways of doing this include providing natural language explanations based on decision trees, using model visualisations to support understanding, and explaining the outcome by example [1]. Whereas many XAI approaches have focused on model-based explanations, Anik and Bunt [2] proposed that data-centric explanations can be equally meaningful to evaluating the trustworthiness of ML models, both by engineers as well as end-users.

#### 4.5.2 *Access Security.*

The steps taken to test the security of ML models will depend on whether they are open or closed, and whether their data are subject to privacy restrictions. There may be ambiguous cases where the training and evaluation data may need different levels of disclosure. For example, the training data may be proprietary or require a non-disclosure agreement, while the evaluation datasets are shared publicly for third-party use [48]. Open datasets that have been anonymised will require a careful review to mitigate the risk of de-anonymisation; ideally this would be done by someone who has a good background knowledge of the hypothetical enemy [33].

Another evaluation that is important to conduct during the testing stage relates to weighing the benefits of detailed reporting practises outlined earlier against the potential risks of exposing confidential data. Adversarial testing should be conducted to make sure that the public-facing model output cannot be used to recreate the original data [48], especially in cases that provide confidence intervals and interactive interfaces (e.g. sliders) in digitised model documentation. Besides test-based approaches, practitioners can also opt for using theoretical models for proving that their models are safe against adversarial attacks [40, 45, 76].

### 4.6 ML deployment

Once ML models are trained and ready for deployment, the focus of data quality work shifts from internal operations on training and test data, and instead looks at assuring the quality of serving data that enter the system from the outside.

Mechanisms are needed to ensure that the serving data undergo the same preparation steps as the steps that were applied to the raw training data [57]. This can be especially challenging in settings where new data arrive continuously, and where they are used to retrain and deploy updated models. The latter case will require additional measures for preventing adversarial attacks such as data poisoning [52] or spam [57].

Other precautions also apply to models that do not ingest new training data. For example, proprietary models can be stolen by repeatedly querying the system (e.g. via a public prediction API) and monitoring the outputs to reverse engineer a substitute model [52]. Another similar risk relates to model inversion, where querying can be used to recover parts of a private training dataset, thereby breaking its confidentiality [52]. These risks are especially likely in models that report confidence values alongside their predictions.

To mitigate the above risks, ML developers should work closely with software engineers in order to ensure that public-facing systems built on top of the ML are robust against malicious attacks. Recent trends in ML have discussed the development of new engineering approaches such as federated learning to foster privacy, data protection and security [33]. Federated learning works by allowing devices to learn a shared prediction model collaboratively while keeping the training data securely on the user's own computer.

### 4.7 ML monitoring

Once a model is deployed, the focus on serving data should continue. At this stage, the work shifts to monitoring the properties of incoming data and ensuring that they are contextually similar to the data that the model was trained on. Polyzotis et al. [57] and Schelter et al. [65] propose analyses that can be used to detect training-serving skew in pre-defined variables. However, others note the difficulty in trying to establish which columns must be inspected, and what the required thresholds should be [63]. Chen et al. [19] suggest that the thresholds can be based on the expected distribution of a targeted population for relevant features (e.g. the usage frequency of a phrase, or the number of individuals with a particular skin tone).

Some monitoring activities can be automated and communicated to the users of ML systems via alerts. This may include data integrity checks, anomaly detection, and performance metrics [52]. Additionally, the system can be designed to gather additional data about misuses or outliers while the model operates in the real world, providing DevOps engineers and ML developers with more information for mitigating security and performance issues in subsequent versions of the model.

### 4.8 Challenges for stakeholders

In this final section of our results, we summarise the data quality requirements that matter within specific stakeholder roles. Knowledge of relevant responsibilities can help practitioners to understand and resolve the data quality issues that are within their capacity, and to articulate their own requirements to relevant colleagues.

- **Subject matter experts** are typically consulted during the early stages of defining the dataset use case and design. These experts can advise on which data features are relevant in their domain of expertise, and the anticipated characteristics of the end users (e.g. in terms of demographic, cultural or environmental traits). Discussions of this kind should help the data collector to assess how much supervision and domain expertise would be needed to collect, label and document the dataset.
- **Data engineers and software engineers** may be involved in different stages of the ML development pipeline. During the initial stages of dataset design and collection, they may be asked to select or build systems for data storage, access, transformation and linking. During the stage of ML deployment, their role may shift to building a user-facing ML system that is secure against attacks or unauthorised access, while at the same time being transparent and user-friendly. Other responsibilities may include building systems for monitoring incoming data and generating alerts if they do not meet a pre-defined set of criteria.

- **Data managers** work with the data validation and maintenance stage of the ML pipeline. Their role is to collaborate with other stakeholders to ensure that a dataset is clean, contextually relevant, well-documented, and accessible in the right way. Important responsibilities include determining the policies and designated owners of the dataset, and ensuring that it is protected from unauthorised access where necessary. Data managers must also take responsibility for putting together relevant documentation about how the dataset was collected, its naming conventions, purpose, and version history.
- **Data analysts and data scientists** are involved during the stages of data validation and ML building. After data are collected, they are likely to carry out pre-processing tasks such as feature selection, deduplication, removal of outliers, consistency checking, anonymisation and imputation of missing values. Data analysts may also be required to inspect the dataset to identify potential biases, protected characteristics, or stereotype-aligned correlations. When it comes to building ML models, data scientists may be tasked with selecting or creating new features, enriching and/or resampling the dataset. In every task, it is important that the practitioner records the sequence of actions they perform on the data, and the properties and limitations they may discover about the dataset.
- **ML engineers** are mostly involved with the ML building and testing stages, and they are likely to collaborate closely with data scientists whose role is to validate and prepare the dataset. ML engineers will make decisions about which data features to use in the model, how to split the training and evaluation data, and whether to build a new model or re-use an existing one. They may need to consult with subject matter experts in order to establish which performance criteria should be prioritised and the different contexts in which the model needs to be tested.
- **DevOps engineers** work with ML engineers and software developers to oversee the ML system once it has been deployed. Their responsibilities include monitoring the properties of incoming and outgoing data to make sure that the system is operating reliably. These responsibilities may be subsumed by ML engineers in the absence of DevOps staff.

An important caveat we would like to restate is that our findings are not exhaustive, and capture only a small selection of recurring themes that came up during our review. We therefore encourage readers to remain open to other data quality requirements that may matter to them and their colleagues, bearing in mind that these may not have been covered here.

## 5 DISCUSSION

Our paper provided a literature review of data quality requirements that matter during ML development. We found that these requirements can be broadly accommodated within the data quality frameworks traditionally endorsed by data management research, including routines for data collection, processing and documentation. What is unique about the experience of ML practitioners is that their data quality requirements and corresponding tasks are disaggregated across different stages of the ML development pipeline.

Each stage of ML development embodies a new purpose with its own data uses and quality requirements, meaning that the traditionally accepted definition of data quality as “fitness for use” should not be viewed as a singular outcome. Instead, data quality must be defined using stage-specific approaches that are sensitive to where in the ML lifecycle the data are encountered and who encounters them [25, 57]. Because of this, the four traditionally used categories of data quality (intrinsic, contextual, representational and accessibility) must be addressed differently at different stages of the ML development pipeline, as we will discuss below.

Requirements around intrinsic data quality may initially be targeted to the data collection stage, where careful monitoring and human-in-the-loop methods can support the acquisition of data that are accurate, reliably sourced and clean from the outset. Once the data have been collected, the requirements may shift to removing any remaining inconsistencies and redundancies as part of general data maintenance. When it is time for the dataset to be used to train an ML model, the intrinsic requirements will include determining an appropriate level of dimensionality and ensuring the completeness of relevant features. During the later stages of ML deployment, the tasks of intrinsic data quality shift from working with training data to the preparation of serving data received from the outside world.

With contextual data quality, the authors in our review highlighted the importance of understanding the anticipated ML use case and characteristics of the end users before data collection. This understanding is needed to design the data collection process to gather data that adequately reflect their purpose and the environment in which the trained ML will be deployed. Other contextual requirements during data collection, especially in sociocultural contexts, relate to compliance with ethical and inclusivity guidelines. After the data are collected, their contextual integrity must be evaluated and, where necessary, improved through the curation of additional data. When approaching the early stages of ML development, the contextual fit of the dataset may be improved through steps such as feature selection and re-sampling of the training distribution. Once the ML model is built, requirements around performance can be assessed using benchmarked evaluation in different contexts. After the model is deployed, the data requirements shift to monitoring the quality of serving data in terms of their distribution and features, to ensure that they align with data characteristics upon which the model was trained.

A large part of the representational aspect of data quality involves documenting how the above requirements were met. In the earlier stages of dataset development, documentation should focus on the user requirements and dataset design, followed by summaries of the dataset collection process, cleaning, maintenance and evaluation steps. Other representational requirements that may arise during the data collection stage relate to the standards used to capture data, as well as the quality of user interfaces for data collection and exploration.

Lastly, requirements around accessibility include the quality and security of infrastructure used for data storage, access and maintenance, which must be in place before the data are available to develop ML models. This can be supported by institutional frameworks and guidelines on consent, transparency and privacy of datasets. When it comes to data security, the later stages of model development require thorough testing and monitoring processes to mitigate against adversarial attacks that could poison training data or expose private datasets.

From the above summary of the intrinsic, contextual, representational and accessibility requirements of ML datasets, we see that the responsibility for managing data quality is distributed across various stakeholders. This includes subject matter experts, data analysts, software engineers, ML engineers and DevOps specialists (or site reliability engineers). Distinguishing between these different classes of users is necessary if we are to understand the radically different backgrounds and tasks that are needed to keep ML data quality pipelines running smoothly [57]. In Figures 4 and 5 of the Results section, we presented illustrated examples of this complex web of relationships and the nature of their interactions with datasets. At the intersection of dataset development and ML pipelines, we came across a number of synergies and tensions that have implications on data quality but have been less explicit in previous data quality frameworks. These are as follows:

- **Ethical and legal requirements** - numerous articles in our review commented on ethical issues such as working with sensitive data, the impact of data-driven decisions on human life, and potential security risks. Rather than being a distinct and temporally-constrained task, we observed that these requirements transcend different stages of the ML lifecycle. This is in line with the observation made by Gebru et al. [23] that the best way to elicit information about ethical and legal compliance is by requiring practitioners to document specific stages of the dataset development process.

- **Amount of data** - early advances in ML were motivated and, in some cases, enabled by the availability of big datasets, and big data remain necessary in many ML applications such as autonomous vehicles and clinical diagnosis [33, 75]. However, numerous researchers in our review highlighted that bigger datasets are not always better. Earlier trends of opportunistic data collection and post hoc justifications of large datasets are gradually moving towards a requirement for more deliberative data collection methods [35, 36], sampling techniques[16] and minimal data architectures [57] to deliver better performance without reducing model accuracy.
- **Representational standards** - adherence to common standards and metadata already has a long history in traditional data management literature. However, ML applications that are built on social and cultural data require practitioners to reconcile different vocabularies and unique ways of perceiving the world with the need for standardised and homogeneous datasets to be fed into ML systems [36]. This requirement for contextual sensitivity is being met by the growing use of semantic standards that use ontologies and annotate data with graph-like properties [33].
- **Software requirements** - software quality can impact data quality in a number of ways. Software infrastructures may determine how data are structured and collected, how access to datasets is granted, and how the dataset is presented for exploration by prospective users (e.g. via visualisations or dashboards). When ML models are integrated into client-side applications, software developers need to ensure that model training and serving data are protected against adversarial attacks, and that they do not inadvertently expose any personal or commercially sensitive data [33].
- **Documentation** - rather than being a post-hoc activity that accompanies completed datasets, the authors in our review viewed documentation as a pre-emptive activity that should span the entire ML development lifecycle. The stages of dataset design, collection, ML training and testing should each yield documents that can support communication and decision-making between successive stakeholders [35, 70]. This is especially valuable in larger organisations where the data and ML activities are separated across teams, or where they are vulnerable to information loss due to staff handover.

In order to show how the above implications map onto the four traditional data quality dimensions that were discussed earlier, we summarise them in Table 4.

Table 4. Additions to traditional data quality dimensions introduced by ML.

| Challenge | Data quality category | | | |
|---|---|---|---|---|
| | Intrinsic | Contextual | Representational | Accessibility |
| Legal and ethical | Some intrinsic aspects of datasets, particularly in personal or sociocultural data, now require greater pre-processing to identify and anonymise or remove sensitive and/or protected characteristics (e.g. gender, race, age). | The relevance of sociocultural data to specific use cases requires an assessment of the presence and distribution of legally protected characteristics. | Documentation of the dataset and its development process can help to anticipate and prevent ethical or legal risks. | Compliance with ethical and legal requirements requires controlled access mechanisms that preserve the security of personal and proprietary data (e.g. data trusts). |
| Bias | | Small contextually relevant datasets can lead to better and fairer performance than large data. | Documenting the environment in which data were collected helps practitioners to assess contextual relevance and to mitigate bias. | |
| Software | Data collection and management software can be used to improve the intrinsic quality of data (e.g. through runtime verification and alerts). | Runtime verification tools can be used to detect contextual drift. | Visualisations and dashboards can make it easier to inspect the quality of a dataset.<br><br>Documentation facilitates the handover of information across different stages of ML development. This is especially useful in scenarios where datasets and ML are developed by multiple teams. | Software built on top of ML models needs to be tested to ensure that model training and serving data are protected against adversarial attacks. |

Many of the processes described above span across multiple stakeholders, whose ability to self-organise into a robust data quality workflow will require the support of higher-level institutional structures. Part of that is about providing incentives to individuals and organisations [23, 53]. At present, the field of ML suffers from the devaluation of data work, with model development tasks being held in higher esteem than data quality processes [35, 64]. In response to this, authors have advocated for the professionalisation of data work as a means to promote best practices in data management and accountability. Practical approaches to this include establishing membership organisations and review panels with standardised codes of conduct [36]. Participation in these schemes will impose greater costs which may be felt particularly strongly by smaller stakeholders such as startups and SMEs. For this reason, policy makers could explore solutions for achieving economies of scale through consortia and trusts[9] that pool the resources needed by practitioners to produce good quality data.

Besides institutional change as a long-term strategy for improving data quality, it is equally important to consider actionable steps that can be taken in the shorter term by individuals and teams that wish to improve their practices. Reviewing the complete range of data quality enhancement tools and protocols goes beyond the scope of this article, but several examples of such tools were encountered during our review. In the sphere of documentation, there exist various checklists, such as those for reporting crowdsourcing experiments [58] and model reproducibility [55], as well as datasheets [23], cover sheets on employment practices [62], data nutrition labels [32], model cards [48], notebooks [72] and explainability toolkits[10]. When it comes to mitigating the risks of ML models through data, readers may be interested in ethics assurance tools such as AI Fairness 360 [11], legal and ethical checklists for NLP [60], and verification tools for streaming and serving data [21, 65]. Lastly, for readers who are interested in sharing ML datasets and the models built upon them, repositories hosted by CodeOcean, GitHub, Zenodo and HuggingFace[12] can serve as good candidates. We encourage interested readers to investigate the relevance, advantages and drawbacks of these tools in their specific use case.

### 5.1 Relevance to use cases

Earlier in this paper we noted that data quality frameworks and standards present practitioners with dozens of possible criteria to comply with. These are accompanied by a growing range of tools for data pre-processing, documentation and assurance. It is impossible for all of these requirements to be met, nor is it necessary. Previous studies that explored the application of data quality standards found that practitioners benefit from seeing examples of data quality requirements, as it helps to clarify their own needs [24]. It was also found that there is value in simplified data quality frameworks that are tailored to specific use cases or technologies [41].

Our review sought to assist ML practitioners who are trying to define their data quality requirements. Firstly, we synthesised previous literature to illustrate the common data quality requirements that can exist in ML. Secondly, by mapping these requirements to different stages of the ML pipeline, we provide a way for readers to see the requirements that are likely to precede and follow their specific task, and to discern which data quality outcomes to focus on in their role. This type of clarity is needed to prevent the diffusion of responsibility and to ensure that every stakeholder is proactive at mitigating data quality issues that are within their capacity.

Besides individuals who work directly with data, we anticipate that our review will be useful to coordinators of data innovation projects that involve multiple stakeholders. Our own experience of

[9]https://theodi.org/article/data-trusts-in-2020/ [accessed 1/03/22]
[10]ELI5 python package: https://github.com/TeamHG-Memex/eli5 [accessed 19/08/22]
[11]https://aif360.mybluemix.net/ [accessed 19/08/22
[12]https://huggingface.co/docs [accessed 18/08/22]

this includes a series of projects that emerged from a Public Private Partnership (PPP) between the European Commission and the Big Data Value Association (BDVA). These projects included the European Data Incubator (EDI), EuRopEAn incubator for trusted and secure data value Chains (REACH) and EUHubs4Data. Their goal was to facilitate data-driven innovation in startups and SMEs through collaboration between data providers, data users, business coaches and legal experts assembled from different geographic regions. The review provided in this paper can help managers of similar initiatives to understand the data quality requirements of colleagues who are responsible for different parts of the data value chain, and to signpost participants to resources that will support their data quality practice and documentation.

# 6 CONCLUSION

Shifting data practices from current priorities driven by availability or convenience towards high quality data will require the effort of decision makers and practitioners at every level of organisations and policy. We hope to have contributed a useful vocabulary for perceiving and articulating some of the nuanced data quality requirements that can be resolved by practitioners in different parts of the ML pipeline.

## 6.1 Limitations

Our review is limited by the relatively small sample of articles, which represent only a minor portion of the growing number of literature that is emerging at the intersection of data management, ML and HCI. It is possible that the keywords and sources that were used during our search for articles, as well as the insights drawn from them, were influenced by the authors' fields of expertise. For example, we did not elaborate greatly on concepts such as data ethics, data feminism and data justice, which relate to data quality but lie outside the technical focus adopted in our review.

Another limitation relates to the simplification of our findings for the purpose of this review. For example the ML pipelines illustrated in Figures 4 and 5 (along with the sequence of stages in Table 3) use a linear sequence of data quality assurance steps that is unlikely to be structured like this in reality. Specifically, multi-dataset-multi-model and agile data iteration scenarios are more common than the waterfall-ish view used to report our findings. Moreover, much of our discussion focused on desirable or ideal scenarios rather than what is feasible. So we did not do justice to the important trade-offs and negotiations that occur when some parts of data quality may need to be adapted or sacrificed in favour of practical requirements and business goals.

## 6.2 Future work

Our review focused mainly on defining the requirements of ML data as part of the "planning" stage of the data quality management process illustrated in Figure 1. We did not systematically review the literature on how these plans can be implemented through tools for data quality control, assurance and improvement. As the fields of ML and data quality research continue to grow, we envision a demand for reviews that are able to compile a list of the available data quality tools, and compare their overlaps, differences, and blind spots.

We also encourage organisations to look beyond the traditional view of data quality as a fixed outcome that meets a list of pre-defined criteria, and to consider a more dynamic perspective that specifies the particular data quality requirements that are valued within their part of the ML lifecycle. Movement in this direction is already underway in the standardisation community, with standards such as ISO/IEC 5259 and ISO/IEC DIS 8183 beginning to incorporate the ML life cycle into data quality recommendations. In Figure 3 we illustrated how the findings of our review may be reconciled with the pipeline of the forthcoming ISO/IEC 5259 standard. We encourage future researchers to investigate the practical application of this standard by individuals and organisations, and to share their experiences with the wider community.


## ACKNOWLEDGMENTS

We would like to acknowledge the support of the Open Data Institute (ODI) Fellowship Programme, as well as colleagues from ODI who shared valuable resources and insights that led to the creation of this paper. We are also grateful to three anonymous reviewers whose thorough comments significantly helped to improve the quality of this article. Any errors or omissions are our own.

This work was part-funded by the European Union's Horizon 2020 research and innovation programme under the projects EUHubs4Data (grant number 951771) and MediaFutures (grant number 951962). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.


## REFERENCES


[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: explaining training data of machine learning systems to promote transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[3] Lora Aroyo, Matthew Lease, Praveen Paritosh, and Mike Schaekermann. 2022. Data excellence for AI: why should you care? *Interactions* 29, 2 (2022), 66–69.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[5] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–39.

[6] Jacqui Ayling and Adriane Chapman. 2021. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* (2021), 1–25.

[7] Yang Baolong, Wu Hong, and Zhang Haodong. 2018. Research and application of data management based on Data Management Maturity Model (DMM). In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. 157–160.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[9] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[11] Laure Berti-Equille. 2019. Learn2clean: Optimizing the sequence of tasks for web data preparation. In *The World Wide Web Conference*. 2580–2586.

[12] Leopoldo Bertossi and Floris Geerts. 2020. Data quality and explainable AI. *Journal of Data and Information Quality (JDIQ)* 12, 2 (2020), 1–9.

[13] A Black and P van Nederpelt. 2020. Dimensions of Data Quality (DDQ) Research Paper. http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf. *DAMA NL Foundation* (2020). [Online; accessed 6/10/21].

[14] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[15] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[16] Paula Branco, Luís Torgo, and Rita P Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)* 49, 2 (2016), 1–50.

[17] Samuel Budd, Emma C Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* 71 (2021), 102062.
[18] Wo Chang. 2022. ISO/IEC JTC 1/SC 42(AI)/WG 2(Data) Data Quality for Analytics and Machine Learning (ML). https://jtc1info.org/wp-content/uploads/2022/06/01_06_Wo_2022_05_24_ISO-IEC-JTC1-SC42-WG2-Data-Quality-for-Analytics-and-Machine-Learning-Wo-Chang-NIST-final.pdf. [Online; accessed 01/12/2022].
[19] Haihua Chen, Jiangping Chen, and Junhua Ding. 2021. Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability* 70, 2 (2021), 831–847.
[20] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.
[21] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. 2019. A DaQL to monitor data quality in machine learning applications. In *International Conference on Database and Expert Systems Applications*. Springer, 227–237.
[22] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37–37.
[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[24] Fernando Gualo, Moisés Rodríguez, Javier Verdugo, Ismael Caballero, and Mario Piattini. 2021. Data quality certification using ISO/IEC 25012: Industrial experiences. *Journal of Systems and Software* 176 (2021), 110938.
[25] Venkat Gudivada, Amy Apon, and Junhua Ding. 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software* 10, 1 (2017), 1–20.
[26] David Gundry and Sebastian Deterding. 2022. Trading Accuracy for Enjoyment? Data Quality and Player Experience in Data Collection Games. In *CHI Conference on Human Factors in Computing Systems*. 1–14.
[27] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. 2021. Data quality for machine learning tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4040–4041.
[28] Thilo Hagendorff. 2021. Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning. *Minds and Machines* 31, 4 (2021), 563–593.
[29] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
[30] D Henderson and S Earley. 2017. DAMA-DMBOK: data management body of knowledge. *Technics Publications* (2017).
[31] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[32] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy* 12, 12 (2020), 1.
[33] Andreas Holzinger. 2018. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 55–66.
[34] Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns* 2, 4 (2021), 100241.
[35] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.
[36] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.
[37] Michael I Jordan and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260.
[38] Ashish Juneja and Nripendra Narayan Das. 2019. Big data quality framework: Pre-processing data in weather monitoring application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE, 559–563.
[39] Daniel S Katz, Kyle E Niemeyer, Arfon M Smith, William L Anderson, Carl Boettiger, Konrad Hinsen, Rob Hooft, Michael Hucka, Allen Lee, Frank Löffler, et al. 2016. Software vs. data in the context of citation. *PeerJ Preprints* 4 (2016), e2630v1.
[40] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. 2017. Towards proving the adversarial robustness of deep neural networks. *arXiv preprint arXiv:1709.02802* (2017).
[41] Sunho Kim, Ricardo Pérez-Castillo, Ismael Caballero, and Downgwoo Lee. 2022. Organizational process maturity model for IoT data quality management. *Journal of Industrial Information Integration* 26 (2022), 100256.
[42] Laura Koesten, Elena Simperl, Tom Blount, Emilia Kacprzak, and Jeni Tennison. 2020. Everything you always wanted to know about a dataset: Studies in data summarisation. *International Journal of Human-Computer Studies* 135 (2020), 102367.

[43] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl. 2022. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *arXiv preprint arXiv:2205.02302* (2022).

[44] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. 2020. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems* 22, 2 (2020), 712–733.

[45] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[47] Jorge Merino, Ismael Caballero, Bibiano Rivas, Manuel Serrano, and Mario Piattini. 2016. A data quality in use model for big data. *Future Generation Computer Systems* 63 (2016), 123–130.

[48] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[49] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.

[50] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45, 1 (2012), 521–530.

[51] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[52] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. 2022. Challenges in deploying machine learning: a survey of case studies. *Comput. Surveys* 55, 6 (2022), 1–29.

[53] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021), 100336.

[54] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*. 1–10.

[55] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research* 22 (2021).

[56] Claudio Santos Pinhanez, Heloisa Candello, Paulo Cavalin, Mauro Carlos Pichiliani, Ana Paula Appel, Victor Henrique Alves Ribeiro, Julio Nogima, Maira De Bayser, Melina Guerra, Henrique Ferreira, et al. 2021. Integrating machine learning data with symbolic knowledge from collaboration practices of curators to improve conversational systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[57] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[58] Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

[59] Jimmy Rising. 2002. Justice and Ethics. (2002).

[60] Anna Rogers, Tim Baldwin, and Kobi Leins. 2021. Just What do You Think You're Doing, Dave?'A Checklist for Responsible Data Use in NLP. *arXiv preprint arXiv:2109.06598* (2021).

[61] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1328–1347.

[62] Annabel Rothschild, Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl DiSalvo, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Towards fair and pro-social employment of digital pieceworkers for sourcing machine learning training data. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–9.

[63] Tammo Rukat, Dustin Lange, Sebastian Schelter, and Felix Biessmann. 2020. Towards automated data quality management for machine learning. In *ML Ops Work. Conf. Mach. Learn. Syst.* 1–3.

[64] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[65] Sebastian Schelter, Stefan Grafberger, Philipp Schmidt, Tammo Rukat, Mario Kiessling, Andrey Taptunov, Felix Biessmann, and Dustin Lange. 2018. Deequ-data quality validation for machine learning pipelines. In *Machine Learning Systems Workshop at the Conference on Neural Information Processing Systems (NeurIPS)*.

[66] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[67] Daniel Staegemann, Matthias Volk, Tuan Vu, Sascha Bosse, Robert Häusler, Abdulrahman Nahhas, Matthias Pohl, and Klaus Turowski. 2020. Determining Potential Failures and Challenges in Data Driven Endeavors: A Real World Case Study Analysis.. In *IoTBDS*. 453–460.

[68] Ikbal Taleb, Mohamed Adel Serhani, Chafik Bouhaddioui, and Rachida Dssouli. 2021. Big data quality framework: a holistic approach to continuous quality management. *Journal of Big Data* 8, 1 (2021), 1–41.

[69] Linnet Taylor. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society* 4, 2 (2017), 2053951717736335.

[70] Divy Thakkar, Azra Ismail, Pratyush Kumar, Alex Hanna, Nithya Sambasivan, and Neha Kumar. 2022. When is Machine Learning Data Good?: Valuing in Public Health Datafication. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[71] Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18, 1 (2017), 7026–7071.

[72] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. 2021. What makes a well-documented notebook? a case study of data scientists' documentation practices in kaggle. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[73] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[74] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[75] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. 2020. Preparing medical imaging data for machine learning. *Radiology* 295, 1 (2020), 4–15.

[76] Eric Wong and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*. PMLR, 5286–5295.

[77] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. Quality assessment for linked data: A survey. *Semantic Web* 7, 1 (2016), 63–93.