International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V71I8P103

Volume 71 Issue 8, 16-27, August 2023

Original Article

An Analysis of Data Quality Requirements for Machine Learning Development Pipelines Frameworks

Sandeep Rangineni

Data Test Engineer, Information Technology, Pluto TV, California, USA.

Received: 10 June 2023 Revised: 21 July 2023 Accepted: 06 August 2023 Published: 26 August 2023

Abstract *- The importance of meeting data quality standards in the context of Machine Learning (ML) development pipelines is explored in this study. It delves deep into why good data is crucial to confidently deploying ML models. The primary goal of this research is to isolate and examine the most important aspects of data quality inside ML pipelines and how they affect model performance and generalizability. The study highlights the complex connection between data quality and ML model performance via an in-depth analysis of multiple phases within the ML pipeline, encompassing data collection, preprocessing, model training, and validation. The study highlights the importance of data quality in reducing bias, improving predicting accuracy, and making ML models more robust to outside influences. The study elaborates on the possible consequences of ignoring data quality issues by highlighting the difficulties given by data noise, incompleteness, and biases. Accuracy, consistency, completeness, relevance, and ethical issues are all part of the data quality criteria that are spelt forth. The study's relevance rests on providing a holistic perspective on the crucial importance of data quality within the landscape of ML development. The survey results provide ML professionals and businesses with a better appreciation for the importance of high-quality data in building trustworthy ML models. Trust in ML model outputs, adoption of ethical data practices, and effective dissemination of ML tools are all facilitated by their corresponding data quality needs being recognized and met.*

Keywords *- Data innovation, Data ecosystems, Machine learning, Data quality, Data management*.

1. Introduction

AI programs exploit the massive amounts of data generated by modern cultures. The second kind of ML is the topic of our study since it is gaining ground in applications such as systems that forecast outcomes based on inputs and propose the best options. These systems can process both structured and unstructured data (such as text, photos, and audio) to solve real-world problems in areas including healthcare, law enforcement, business, and transportation [52].

It is possible for ML systems to perpetuate prejudice against under-represented groups by using historical signals and incorrectly proxy measurements in settings such as employment hiring and criminal justice [46, 67]. It is estimated that between 10% and 30% of sales is spent on addressing data quality concerns [30], which may significantly impact a company's ability to run efficiently. Therefore, corporate and public stakeholders increasingly acknowledge the significance of data quality to decrease societal hazards, lower costs, and facilitate the efficient use of ML technologies. Due to the increasing prevalence of ML across sectors and the potentially life-altering nature of some of its earlier applications, the methods by which ML-based decision-support systems arrive at their conclusions are under increasing scrutiny [17, 44]. National and international organizations like the OECD1 and the Open Government Partnership are encouraging routines to ensure openness in ML datasets and development processes.

2. Background

There are several methods for gathering the training data needed for ML algorithms. Roh et al. [61] classify the many different ways that data may be collected for ML into three broad categories: Data collection consists of three stages: (1) discovery, (2) augmentation, and (3) creation and labelling of data using manual and semi-supervised methods. Use cases and the specifics of the data needed by an ML system determine the degree to which these techniques are implemented during data gathering.

Before reaching a practitioner of ML or the resulting product, data may be converted and handled by a number of different parties in bigger organizations and complicated innovation ecosystems.

Different Methods of Data Management Exist in the Middle: Academics and Business

It is important to note that how ML data is handled might vary greatly depending on whether the system is used in a research or commercial scenario [52]. Data management in academia is often delegated to individuals or small groups working on a specific project, who have complete autonomy over creating and maintaining their own data gathering, storage, and sharing infrastructures. To maintain consistency and collaboration among teams, however, researchers in the industry often use various independent platforms for data gathering, processing, and storage. The standard ISO/IEC 25024 addresses the latter issue by guiding organizations as they define data quality assurance standards and methods for monitoring them quantitatively.

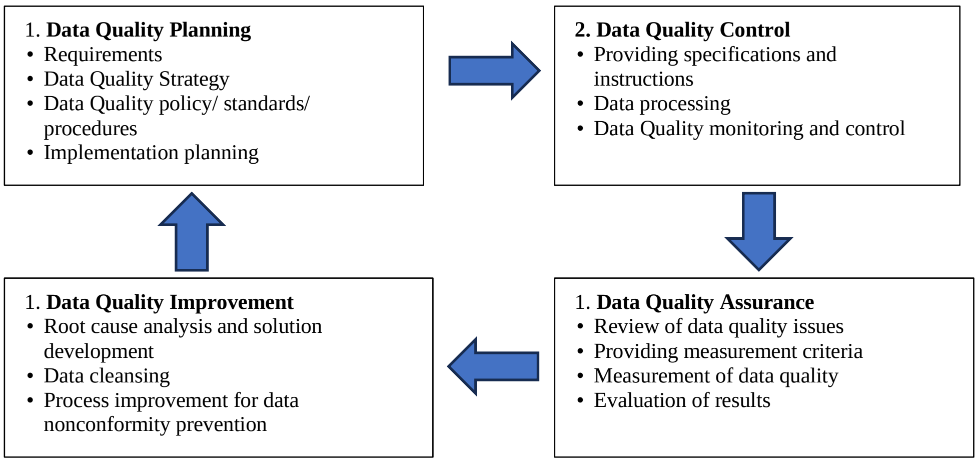

Fig. 1 Management of data quality

2.1. Implementation Follows Careful Planning for Data Quality

Our research was conducted with the intention of aiding ML professionals and data managers in the early stages of their quest to improve data quality. Practitioners may make more informed decisions about what measures to take for data quality control, assurance, and improvement if they have a firm grasp of the relevant standards. While the review's focus is not on the work involved in establishing certain data quality standards, assessment criteria, or methods for assessing data quality, we will provide examples when applicable. There are two aims with this piece. Our primary goal is to educate professionals on the relevant standards and best practices for data quality in the Machine Learning (ML) community. That is going to happen.

Involves compiling research over the last several years and classifying advice according to well-established data quality metrics in the discipline of data management. Our goal is to streamline the process by which businesses and individuals can get their data management systems ready for machine learning and plan for potential problems that may crop up throughout various phases of ML development.

3. Research and Methodology

Articles for this review were chosen based on the following research objectives, inclusion criteria, and search method, which are detailed below. Through thematic coding, we were able to expand our understanding of the growth of ML and the significance of data quality management within the field as a whole by analyzing the selected publications.

Articles Chosen Ahead of Time. Based on our experience working with ML models, we compiled a list of six papers [3, 23, 32, 34, 35, 58] on data quality planning and, more specifically, documentation.

It is an automatic search. To find relevant publications, we utilized Google Scholar to look for titles containing our study topics' keywords. Searching simply for article titles helps get rid of irrelevant items. Then, the results were narrowed down by reading the papers' abstracts and titles. Those deemed worthy of retention were the only ones who were kept on.

In the first step, we used the query "allintitle: "data quality" ("machine learning" OR "AI")" to search the whole of Google Scholar. The resulting number is 185.

We stopped after reviewing the first 30 results since so few fulfilled our inclusion criterion. Seven papers [12, 19, 21, 25, 27, 28, and 63] were kept after abstract review.

Snowballing: The process of reading and assessing the articles chosen using the aforementioned methods led us to discover other publications that addressed our study concerns. This method yielded eight articles [5, 9, 11, 36, 48, 53, 55, 57]. Our inclusion criteria were used to evaluate these papers after they were selected based on the descriptions supplied by the citing authors. Because we were interested in learning more about the research of the authors who mentioned this publication, we performed a forward search of papers that cited [64], which led us to one further item [53].

Table 1. Research type facets

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 articles	New ways of looking at existing fields through taxonomies or conceptual frameworks.
Opinion articles	Personal opinions on whether a technique is good or bad, or how it should be applied. Such articles do not rely on related work or research methods
Experience articles	Explanations of how a framework has been applied in practice, based on the experience of the author.

Table 2. The amount of results obtained from various google scholar

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

Coding of Thematic Separate column to call out any peculiar data quality concerns or needs that ML may impose.

We would want to define our findings' parameters before sharing them. Our primary focus was on theoretical frameworks that may be used to specify and design data quality requirements in ML; however, we made sure to take note of any applicable methods that were described in the literature. When it comes to preparing datasets for ML, several of the publications we looked at went above and beyond just "planning" data quality to provide guidance to data practitioners and managers.

Due to the fact that separate communities have traditionally tackled these areas, the connections between them are murky at best. Nonetheless, we make an effort to demonstrate the substantial overlap in Figure 2. According to Rising's [59] conception, justice concerns circumstances and outcomes, whereas ethics focuses on the choices that produce those outcomes. In this light, data ethics concerns how professionals use data to safeguard individuals' rights to secrecy and transparency and the safety and well-being of themselves and the environment [6]. However, data justice tackles disparities in how individuals are portrayed and dealt with based on the data they provide [69]. Data feminism identifies the power relations in society as the root cause of these inequalities and advocates for actions that address them [20].

Figure 2 shows how these works draw attention to how data-centric technology may either exacerbate or alleviate systemic problems in people's everyday lives. Interested readers are encouraged to pursue these issues independently since we did not actively seek out these perspectives and because space and time restrictions prevented us from discussing them in the depth they merit.

Our research also uncovered a second scoping difficulty associated with the nature of the data itself. For instance, we discovered that software tools (for data management or validating input or output data) may moderate training data quality.

Figure 2 shows how these works draw attention to the ways in which data-centric technology may either exacerbate or alleviate systemic problems in people's everyday lives.

Interested readers are encouraged to pursue these issues independently since we did not actively seek out these perspectives and because space and time restrictions prevented us from discussing them in the depth they merit.

Our research also uncovered a second scoping difficulty associated with the nature of the data itself. For instance, we discovered that software tools (for data management or validating input or output data) may moderate training data quality.

Figure 3: A depiction showing how the data processing structure in ISO/IEC 5259 (upper part) corresponds to our data quality pipeline (bottom part). Used a diagram from Chang [18].

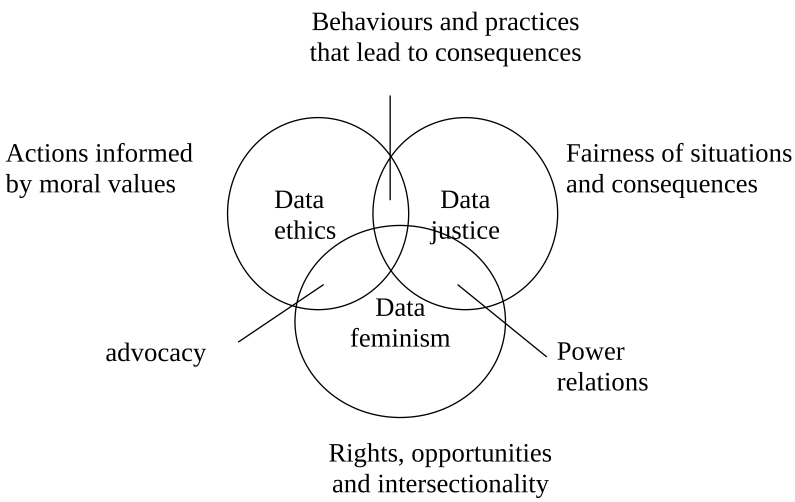

Fig. 2 Venn diagram of fields that complement data quality management

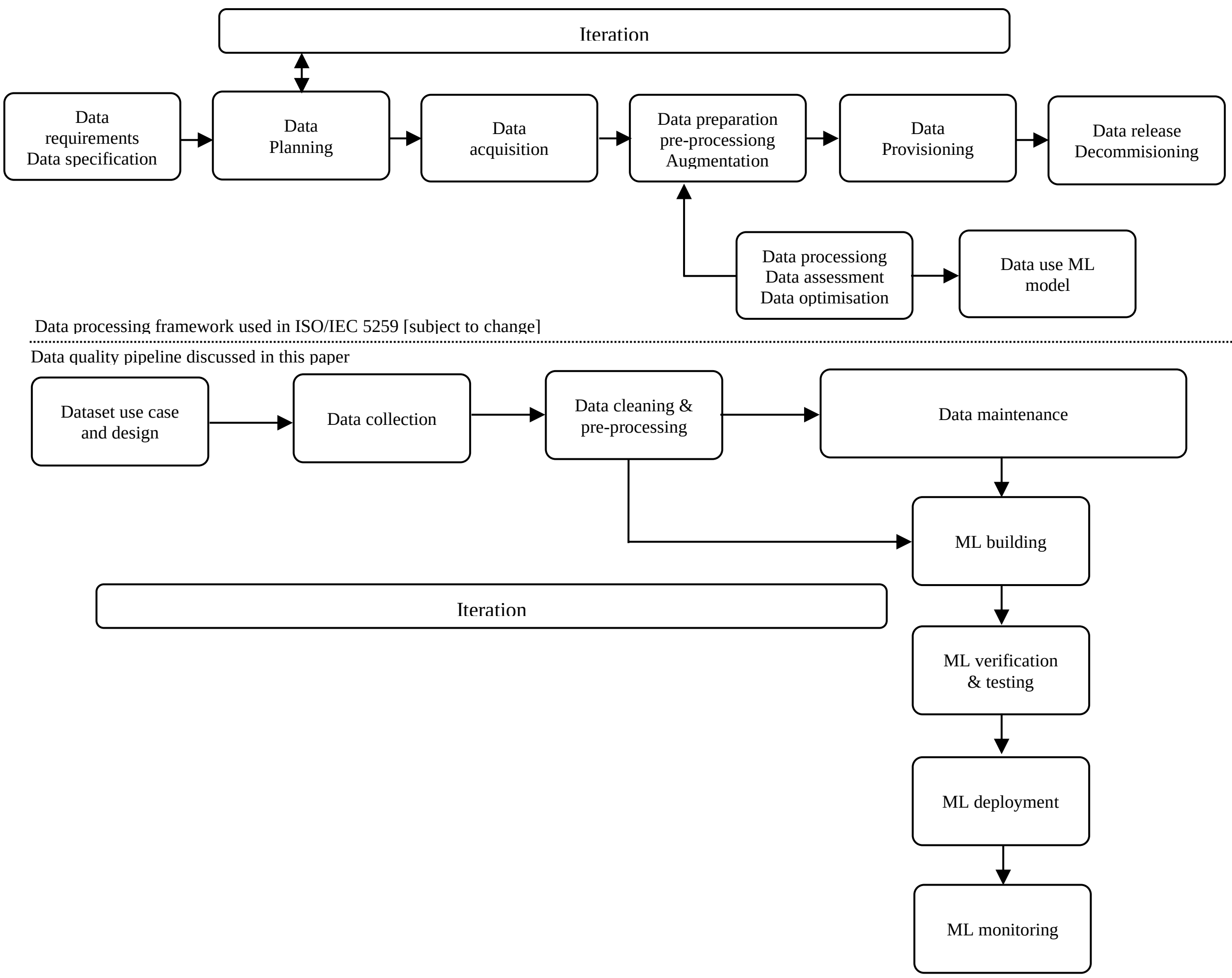

Fig. 3 Data processing structure in ISO/IEC 5259

3.1. Results

We organize our results by the major milestones in the history of machine learning. Because this is a cyclical process with several possible outcomes, no one set procedure can be used in every situation. Nonetheless, experts have uncovered several recurring patterns.

Fayyad et al. [22] presented a nine-step process for knowledge discovery in datasets as far back as 1996.8 According to the authors, the first step is to gain familiarity with the application domain and use case, then comes data collection, pre-processing, and reduction, then comes the identification and application of appropriate data mining methods, and finally comes the interpretation and implementation of the insights gained. Although the authors were aware of difficulties with data accessibility, HCI, and scaling models in knowledge discovery processes, they chose to concentrate on their pipeline's finer phases of data mining. Figure 3 (top portion) depicts the tentative data processing structure proposed by the future industry standard ISO/IEC 5259 [18].

More attention has been paid to recent scholarly analyses of the ML pipeline to dissect its many phases. In particular, they investigate the organizational and operational concerns unique to model creation, verification, deployment, and monitoring [5, 43].

For the sake of this paper, our results are organized into the steps shown at the bottom of Figure 3 and the first column of Table 3. Data pipelines may be challenging to consolidate across multiple operational settings since ML development seldom follows a pre-established order, as recognized by previous papers and standards. As a result, we cannot just assume that our phases would happen one after the other. Our simplified illustration probably will not reflect all the variations from the actual world. These charts are meant to illustrate how various steps in the ML development process correspond to various areas of data quality assurance. This is not a comprehensive analysis; therefore, we ask that you use your judgment to determine whether and how the following data quality standards apply to the non-linear cycles through which you create your datasets.

We would also want to stress that the outlined criteria for data quality are only recommended, not mandatory. Expecting them to be fully met is impractical, particularly in contexts where practitioners must balance conflicting demands for resources like time and money. Likewise, it is not uncommon for data management skills to evolve and improve as a project develops [7]. Therefore, the information presented here should be seen more aspirationally than prescriptively by the readers.

The data collection and labeling for the planned project [36] would need supervision, topic experience, and specialization.

Data collecting procedures characteristic of modern ML implementations differ from the previously advised rigorous study of requirements before data gathering [36]. The issue is that these methods seldom assess where the data came from, who was behind it, what technology was used, or what effect it may have. The questioning of assumptions about whether queries are answered with specific data properties is another challenge that may easily be disregarded when working with large data. Studies that sought to infer subjects' characteristics from their photographs are highlighted by Paullada et al. [53]. Human features have the mistaken impression that making such forecasts is feasible and useful.

In the context of real-time applications, when data is continually arriving, and models are continuously being trained, runtime verification approaches might be useful. To guarantee that the assumptions of the particular ML model are met, "online learning" involves constant monitoring to address any data quality concerns as they arise and bring them within acceptable boundaries [21]. Some use cases for ensuring data quality in online education may also need extra human resources for data labeling and the necessary technological infrastructure and tools.

3.2. Data Collection

Collecting data is the next step after establishing a data use case and operational needs. The above design considerations may be put into action in a variety of methods, including the use of software, annotator rules, and labeling platforms. How documentation, standards, and interfaces may aid in collecting high-quality data is discussed below. The Record of Facts Gathered. Many writers have released examples of good documentation structures. Data statements [9], data sheets [23, 35], and checklists [58] are all examples. These publications are meant to encourage dataset developers to pause for thought and consider their goals, assumptions, usage implications, and stakeholders' values before moving forward with data collecting.

Consumers may make more educated judgments about how to utilize a dataset and prevent unintended exploitation with the help of documentation regarding data-gathering techniques [23, 25]. This helps users determine whether the data are sufficient for their purposes [19]. This kind of paper has been actively promoted via sociocultural data-gathering systems like crowdsourcing, where data workers are recruited from around the globe to read texts, see photographs, and label data needed to train ML models. This includes keeping track of sample processes, converting experimental conditions into micro-tasks, and checking in with participants to ensure they contribute useful information [58].

The goal is to inspire requesters to establish standards of fair and courteous treatment of data workers in the workplace.

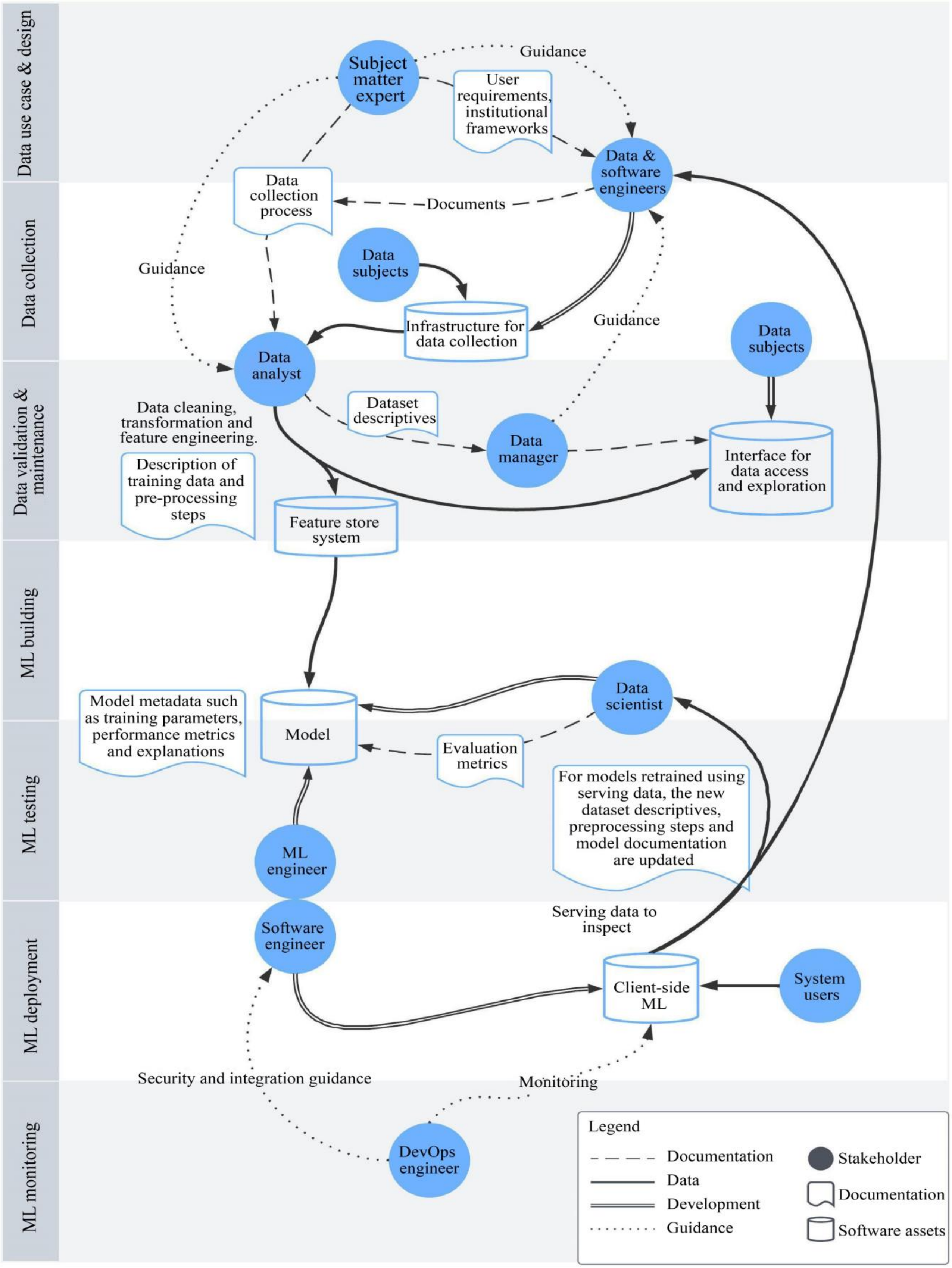

Fig. 4 Shows an example of an ML data quality process

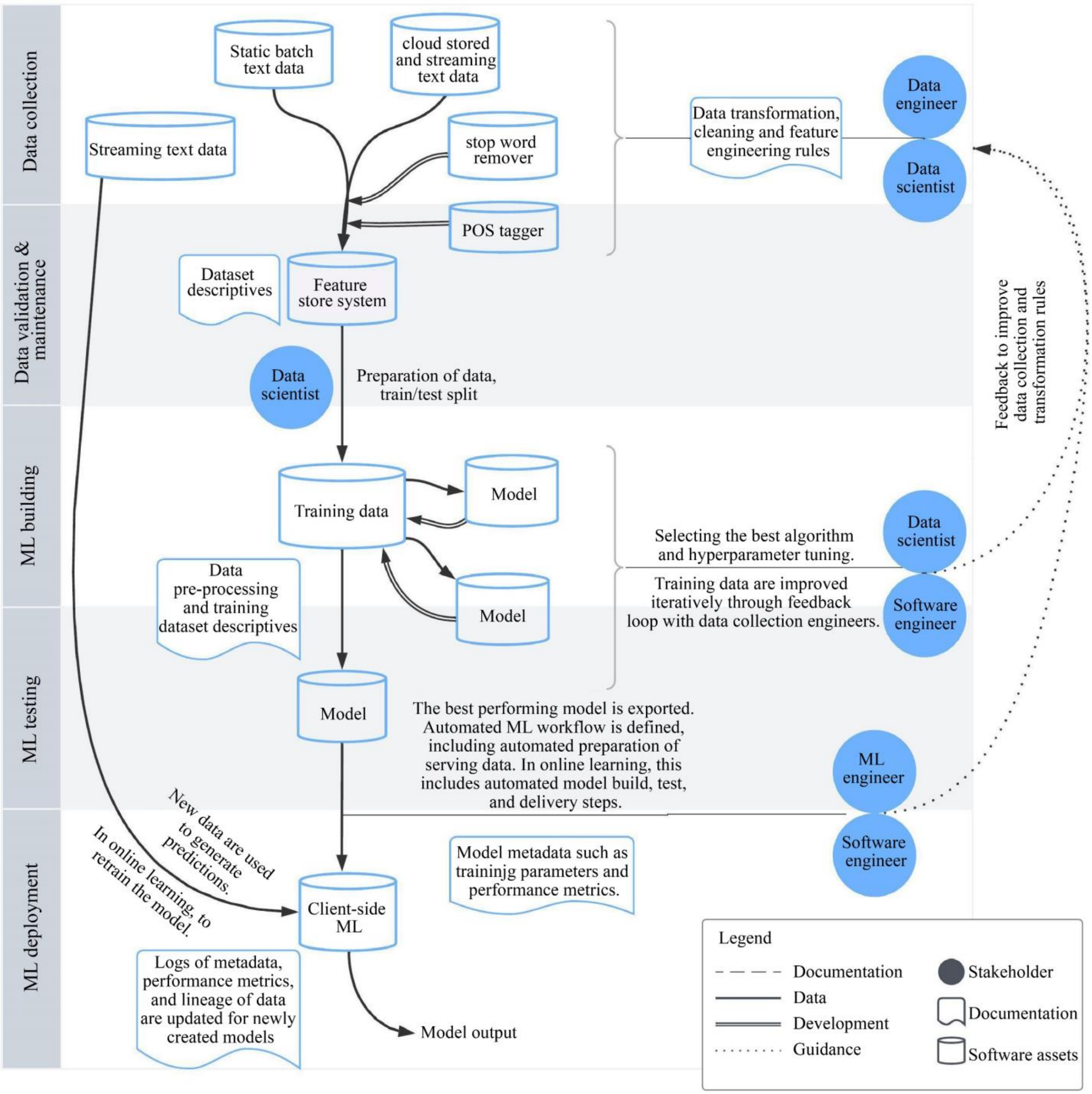

Fig. 5 A sample pipeline for a situation with multiple models and datasets

3.2.1. Data Collection Standards

Data heterogeneity may take the form of Methods of Information Input and Output. The continual data flow from sensors and online applications makes automated data collection a key feature of production ML. Software developers bear some of the burdens for guaranteeing high data quality in situations like these because they may create systems that send actionable warnings to users when problems are detected (such as when a feature is absent or has an unexpected value) [57].

3.2.2. Verifying and Updating Existing Data

In order for the data to be useful in an ML system, they must first be checked and cleaned once they have been acquired. Data quality assurance tasks are heavily weighted at this point in the machine learning development process.

Bertossi and Geerts [12] provide an example of how XAI approaches might be used to identify the causes of data inconsistencies and then recommend the most effective corrective measures.

However, data practitioners should still be mindful of recording their activities whenever feasible, even if formal data cleaning methods have not been utilized (by, for example, following pre-defined procedures or publishing in advance replicable code used to prepare the data).

Table 3. Additions to traditional data quality dimensions introduced by ML

Challenge	Data Quality Category			
	Intrinsic	**Contextual**	**Representational**	**Accessibility**
Legal and eithical	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 require 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. 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.

4. Conclusion

The study's results on important data quality criteria across Machine Learning (ML) development pipelines highlight the importance of high-quality data for successfully deploying ML models. This research set out to better understand how model performance and generalization are affected by the data quality utilized in machine learning workflows. The results highlighted the importance of high-quality data in reducing model bias, improving prediction accuracy, and bolstering ML models' overall resilience. The research highlighted the complex relationship between data quality and model performance by evaluating several phases of the ML pipeline, including data collection, preprocessing, model training, and validation. In addition, the study highlighted the difficulties brought on by noisy, incomplete, or biased data and outlined the possible consequences of ignoring data quality concerns. It outlined all the criteria for acceptable data quality, such as precision, consistency, completeness, relevance, and morality. The value of this research resides in the fact that it contributes to our knowledge development process, providing useful guidance to professionals and businesses as they work to create robust and efficient ML models. Developers of ML systems may do more to encourage the responsible and effective use of ML technologies if they acknowledge and solve data quality concerns.

References

[1] Amina Adadi, and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ariful Islam Anik, and Andrea Bunt, "Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Lora Aroyo et al., "Data Excellence for AI: Why Should You Care?," *Interactions*, vol. 29, no. 2, pp. 66–69, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Alejandro Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Rob Ashmore, Radu Calinescu, and Colin Paterson, "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–39, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jacqui Ayling, and Adriane Chapman, "Putting AI Ethics to Work: Are the Tools Fit for Purpose?," *AI and Ethics*, vol. 2, pp. 405–429, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yang Baolong, Wu Hong, and Zhang Haodong, "Research and Application of Data Management Based on Data Management Maturity Model (DMM)," *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. pp. 157–160, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Rachel K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development*, vol. 63, no. 4-5, pp. 4:1 - 4:15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Emily M. Bender, and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Emily M. Bender et al., "On the dangers of Stochastic Parrots: Can Language Models be Too Big?," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Laure Berti-Equille, "Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation," *WWW '19: The World Wide Web Conference*, pp. 2580–2586, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Leopoldo Bertossi, and Floris Geerts, "Data Quality and Explainable AI," *Journal of Data and Information Quality*, vol. 12, no. 2. pp. 1–9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Andrew Black, and Peter van Nederpelt, "Dimensions of Data Quality (DDQ) Research Paper," *DAMA NL Foundation*, pp. 1-113, 2020. [Google Scholar] [Publisher Link]

[14] Tolga Bolukbasi et al., "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4356–4364, 2016. [Google Scholar] [Publisher Link]

[15] Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models." *ArXiv*, pp. 1-214, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Paula Branco, Luís Torgo, and Rita P. Ribeiro, "A survey of Predictive Modeling on Imbalanced Domains," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–50, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[17] Samuel Budd, Emma C. Robinson, and Bernhard Kainz, "A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wo Chang, "ISO/IEC JTC 1/SC 42(AI)/WG 2(Data) Data Quality for Analytics and Machine Learning (ML)," *Information Technology Laboratory*, 2022. [Google Scholar] [Publisher Link]

[19] Haihua Chen, Jiangping Chen, and Junhua Ding, "Data Evaluation and Enhancement for Quality Improvement of Machine Learning," *IEEE Transactions on Reliability*, vol. 70, no. 2, pp. 831–847, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Catherine D'Ignazio, and Lauren F. Klein, *Data Feminism*, Cambridge: Massachusetts Institute of Technology, 2020. [Google Scholar] [Publisher Link]

[21] Lisa Ehrlinger et al., "A DaQL to Monitor Data Quality in Machine Learning Applications," *International Conference on Database and Expert Systems Applications*, pp. 227–237, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp 37–54, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[23] Timnit Gebru et al., "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Fernando Gualo et al., "Data Quality Certification using ISO/IEC 25012: Industrial Experiences," *Journal of Systems and Software*, vol. 176, p. 110938, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Venkat Gudivada, Amy Apon, and Junhua Ding, "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017. [Google Scholar] [Publisher Link]

[26] David Gundry, and Sebastian Deterding, "Trading Accuracy for Enjoyment? Data Quality and Player Experience in Data Collection Games," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, no. 156, pp. 1–14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[27] Nitin Gupta et al., "Data Quality for Machine Learning Tasks," *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4040–4041, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Thilo Hagendorff, "Linking Human and Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning," *Minds and Machines*, vol. 31, pp. 563–593, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Haibo He, and Edwardo A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[30] Deborah Henderson, and Susan Earley, *DAMA-DMBOK: Data Management Body of Knowledge*, 2nd ed., Technics Publications, p. 624, 2017. [Google Scholar] [Publisher Link]

[31] Fred Hohman et al., "Understanding and Visualizing Data Iteration in Machine Learning," *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[32] Sarah Holland et al., *The Dataset Nutrition Label*, Data Protection and Privacy, vol. 12, no. 12, 2020. [Google Scholar] [Publisher Link]

[33] Andreas Holzinger, "From Machine Learning to Explainable AI," *World Symposium on Digital Intelligence for Systems and Machines (DISA'18)*, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[34] Sara Hooker, "Moving Beyond "Algorithmic Bias is a Data Problem," *Patterns*, vol. 2, no. 4, p. 100241, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Ben Hutchinson et al., "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 560–575, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[36] Eun Seo Jo, and Timnit Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 306–316, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[37] Michael I. Jordan, and Tom M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[38] Ashish Juneja, and Nripendra Narayan Das, "Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application," *International Conference on Machine Learning, Big Data, Cloud, and Parallel Computing (COMITCon'19)*, pp. 559–563, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[39] Daniel S. Katz et al., "Software vs. Data in the Context of Citation," *PeerJ Preprints*, pp. 1-4, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[40] Guy Katz et al., "Towards Proving the Adversarial Robustness of Deep Neural Networks," *Arxiv*, pp. 19-26, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[41] Sunho Kim et al., "Organizational Process Maturity Model for IoT Data Quality Management," *Journal of Industrial Information Integration*, vol. 26, p. 100256, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[42] Laura Koesten et al., "Everything you Always Wanted to Know about a Dataset: Studies in Data Summarisation," *International Journal of Human-Computer Studies*, vol. 135, p. 102367, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[43] Dominik Kreuzberger, Niklas Kühl, and Sebastian Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," A*rXiv*, 2022. [CrossRef] [Publisher Link]

[44] Sampo Kuutti et al., "A Survey of Deep Learning Applications to Autonomous Vehicle Control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[45] Aleksander Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," *ArXiv*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[46] Ninareh Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[47] Merino Jorge et al., "A Data Quality in Use Model for Big Data," *Future Generation Computer Systems*, vol. 63, pp. 123–130, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[48] Margaret Mitchell et al., "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[49] Tanushree Mitra, Clayton J. Hutto, and Eric Gilbert, "Comparing Person-and Process-Centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1345–1354, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[50] Jose G. Moreno-Torres et al., “A Unifying View on Dataset Shift in Classification,” *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[51] Eirini Ntoutsi et al., “Bias in Data-Driven Artificial Intelligence Systems–An Introductory Survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[52] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–29, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[53] Amandalynne Paullada et al., “Data and its (dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research,” *Patterns*, vol. 2, no. 11, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[54] Kai Petersen et al., “Systematic Mapping Studies in Software Engineering,” *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, pp. 68–77, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[55] Joelle Pineau et al., “Improving Reproducibility in Machine Learning Research (a Report from the NeurIPS 2019 Reproducibility Program),” *Journal of Machine Learning Research*, vol. 22, no. 1, pp. 7459–7478, 2021. [Google Scholar] [Publisher Link]

[56] Claudio Santos Pinhanez et al., “Integrating Machine Learning Data with Symbolic Knowledge from Collaboration Practices of Curators to Improve Conversational Systems,” *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[57] Neoklis Polyzotis et al., “Data Lifecycle Challenges in Production Machine Learning: A survey,” *ACM SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[58] Jorge Ramírez et al., “On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. 2, pp. 1–34, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[59] Jimmy Rising, “*Justice and Ethics*,” Massachusetts Institute of Technology MIT, Cambridge, MA, Report., 2002. [Publisher Link]

[60] Anna Rogers, Tim Baldwin, and Kobi Leins, “Just What do You Think you’re Doing, Dave? A Checklist for Responsible Data Use in NLP,” *ArXiv*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[61] Yuji Roh, Geon Heo, and Steven Euijong Whang, “A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[62] Annabel Rothschild et al., “Towards Fair and Pro-Social Employment of Digital Pieceworkers for Sourcing Machine Learning Training Data,” *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[63] Tammo Rukat, Dustin Lange, Sebastian Schelter, and Felix Biessmann, “Towards Automated Data Quality Management for Machine Learning,” *Proceedings of the Workshop on MLOps Systems at the 3rd Conference on Machine Learning and Systems*, pp. 1–3, 2020. [Google Scholar] [Publisher Link]

[64] Nithya Sambasivan et al., “Everyone Wants to do the Model Work, Not the Data Work”: Data Cascades in High-Stakes AI,” *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[65] Sebastian Schelter et al., “Deequ-Data Quality Validation for Machine Learning Pipelines,” *Proceedings of the Machine Learning Systems Workshop at the Conference on Neural Information Processing Systems*, 2018. [Publisher Link]

[66] Shreya Shankar et al., “No Classification Without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World,” *ArXiv*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[67] Daniel Staegemann et al., “Determining Potential Failures and Challenges in Data-Driven Endeavors: A Real World Case Study Analysis,” *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*, pp. 453–460, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[68] Ikbal Taleb et al., “Big Data Quality Framework: A Holistic Approach to Continuous Quality Management,” *Journal of Big Data*, vol. 8, pp. 1–41, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[69] Linnet Taylor, “What is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally,” *Big Data and Society*, vol. 4, no. 2, pp. 1-14, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[70] Divy Thakkar et al., “When is Machine Learning Data Good?: Valuing in Public Health Datafication,” *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[71] Jennifer Wortman Vaughan, “Making Better Use of the Crowd: How Crowdsourcing can Advance Machine Learning Research,” *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1-46, 2017. [Google Scholar] [Publisher Link]

[72] April Yi Wang et al., “What Makes a Well-Documented Notebook? A Case Study of Data Scientists’ Documentation Practices in Kaggle,” *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[73] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan, “Whose AI dream? In Search of the Aspiration in Data Annotation,” *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[74] Richard Y. Wang, and Diane M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[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, no. 1, pp. 4–15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[76] Eric Wong, and Zico Kolter, “Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope,” *Proceedings of the International Conference on Machine Learning*, pp. 5286–5295, 2018. [Google Scholar] [Publisher Link]

[77] Amrapali Zaveri et al., “Quality Assessment for Linked Data: A Survey,” *Semantic Web*, vol. 7, no. 1, pp. 63–93, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[78] Sandeep Ranginenin, Arvind Kumar Bhardwaj, and Divya Marupaka, “An Overview and Critical Analysis of Recent Advances in Challenges Faced in Building Data Engineering Pipelines for Streaming Media,” *The Review of Contemporary Scientific and Academic Studies*, vol. 3, no. 6, pp. 1-5, 2023. [CrossRef] [Publisher Link]

[79] Divya Marupaka, Sandeep Rangineni, and Arvind Kumar Bhardwaj, “Data Pipeline Engineering in the Insurance Industry: A Critical Analysis of ETL Frameworks, Integration Strategies, and Scalability,” *International Journal of Creative Research Thoughts*, vol. 11, no. 6, pp. 530-539, 2023. [CrossRef] [Publisher Link]

[80] Sandeep Rangineni, Divya Marupaka, and Arvind Kumar Bhardwaj, “An Examination of Machine Learning in the Process of Data Integration,” *SSRG International Journal of Computer Trends and Technology*, vol. 71, no. 6, 2023. [CrossRef] [Publisher Link]