Machine Learning Data Practices through a Data Curation Lens: An Evaluation Framework

Eshta Bhardwaj*
eshta.bhardwaj@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Harshit Gujral
harshit.gujral@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Siyi Wu
reyna.wu@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Ciara Zogheib
ciara.zogheib@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Tegan Maharaj
tegan.maharaj@utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Christoph Becker
christoph.becker@utoronto.ca
University of Toronto
Toronto, Ontario, Canada

ABSTRACT

Studies of dataset development in machine learning call for greater attention to the data practices that make model development possible and shape its outcomes. Many argue that the adoption of theory and practices from archives and data curation fields can support greater fairness, accountability, transparency, and more ethical machine learning. In response, this paper examines data practices in machine learning dataset development through the lens of data curation. We evaluate data practices in machine learning *as* data curation practices. To do so, we develop a framework for evaluating machine learning datasets using data curation concepts and principles through a rubric. Through a mixed-methods analysis of evaluation results for 25 ML datasets, we study the feasibility of data curation principles to be adopted for machine learning data work in practice and explore how data curation is currently performed. We find that researchers in machine learning, which often emphasizes model development, struggle to apply standard data curation principles. Our findings illustrate difficulties at the intersection of these fields, such as evaluating dimensions that have shared terms in both fields but non-shared meanings, a high degree of interpretative flexibility in adapting concepts without prescriptive restrictions, obstacles in limiting the depth of data curation expertise needed to apply the rubric, and challenges in scoping the extent of documentation dataset creators are responsible for. We propose ways to address these challenges and develop an overall framework for evaluation that outlines how data curation concepts and methods can inform machine learning data practices.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Machine learning*; • **General and reference** → Evaluation.

KEYWORDS

data practices, datasets, dataset creation, datasheets, documentation, evaluation, machine learning, rubric

ACM Reference Format:
Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine Learning Data Practices through a Data Curation Lens: An Evaluation Framework. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3658955

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658955

1 INTRODUCTION

The pervasive usage of predictive machine learning (ML) models has not dwindled in the face of ever-growing research discussing cases of biased results [2, 6, 9, 15, 26, 30, 31, 41, 54, 70, 71, 78, 89, 100, 114, 121, 141]. Bias in ML models often causes discriminatory, unfair, or unethical judgements towards specific populations. Past research has shown that algorithms can generate gendered biases such as image captioning models that produce gender-specific predictions based on image context [52], analogy generators that associate genders with stereotypical activities [17], and neural machine translation systems that generate gendered outputs [135]. Algorithms can also produce racial biases in facial recognition [20], inaccurate classifications of racial minorities as "hateful" in online hate detection [3], and prioritization of referrals for complex medical care for white people over black people on average [81]. The biases found in these cases and others are widely attributed to the choices made about datasets used for training ML models [43, 58, 84].

Reused datasets are not always fit for a new model's intended purpose. Koch et al. show that benchmark datasets created in one task community are used in other communities, which raises the risk of inappropriate usage [75]. Paullada et al. discuss similar concerns on the implications of dataset benchmarks that are reused across tasks and the creation of data derivatives that reuse datasets outside their original context [113]. Appropriate data use is also hindered by the hidden, tacit, and undervalued nature of the practices underlying data collection, processing, and implementation. As Hutchinson et al. point out, "How can AI systems be trusted when the processes that generate their development data are so poorly understood?" [63, p. 560]. In addition, knowledge related to using and forming data is often obfuscated because of the tacit skills and expertise involved [104, 133] but also because data work is undervalued and taken for granted in the face of performance metrics related to models [13, 18, 63]. These factors contribute greatly to the lack of transparency and accountability in ML models.

Attempts to address these issues look towards the study of data practices in ML. Data practices in this context are defined as "…what and how data are collected, managed, used, interpreted, reused, deposited, curated, and so on…" [18, p. 55], and are also referred to as data work [124] and dataset development [72, 113, 125]. Many studies have highlighted that the overall lifecycle for dataset development should get greater recognition for its impact on predictive models and as a result requires a more intentional strategy [13, 51, 63, 72, 111, 113, 116, 125]. This has led to a greater focus on the development of context documents – "interventions designed to accompany a dataset or ML model, allowing builders to communicate with users" [19, p. 2]. Other research on dataset development has explored the needs of practitioners in performing documentation [51, 58, 76], the challenges and opportunities in reducing bias and increasing fairness and accountability of data used in ML [4, 86, 98, 99, 131], the impacts of data preprocessing on ML models [16, 46, 90], aspects of fairness in dataset annotation [76], and many more. Particularly, this study adds to emerging research that discusses the adoption of principles from archival studies and digital curation into dataset development processes for machine learning research (MLR) [13, 25, 67, 80, 134].

Digital curation is defined as "the active involvement of information professionals in the management, including the preservation, of digital data for future use" [140, p. 335]. The broader domain of digital curation includes all digital objects. Data curation is a subset of this domain that focuses solely on data objects. Data curation involves "maintaining and adding value to digital research data for current and future use" [22, p. 1]. Studies call for ethical data curation [80] and methods from archival studies as these fields have long dealt with large amounts of data and concerns of representativeness, ethics, and integrity [25, 67, 134]. While these studies propose principles and practices that can be adopted from data curation in theory, there is a gap in applying the concepts within ML to demonstrate their feasibility and usefulness in practice.

In this work, we present an application of a data curation lens within dataset development in ML to obtain a practical understanding of data practices. We review and consolidate the literature on ML data work documentation and data curation frameworks and leverage these theoretical foundations to study whether data curation can feasibly provide frameworks for improved fairness, accountability, and transparency in ML dataset development. Our overall **research question** is: *How should data curation concepts and methods inform ML data practices?* Our aim is to explore, at the intersection of these fields, how ML data practices currently perform data curation and how data curation can be enacted more effectively and rigorously. Our **working hypothesis** is that data curation frameworks can be effectively used to guide and evaluate data practices in ML. We therefore use data curation frameworks to conceptualize and evaluate existing ML practices *as data curation*. Our goal is that in the near future, data curation is routinely recognized and rigorously performed as a key part of ML research, including its norms and peer review standards. We present a summary of literature from data curation to establish its importance in ML and use it as a lens for ML. By examining data practices in MLR through the lens of data curation, we aim to contribute to effective dataset development in ML that supports transparent, fair, and accountable ML practices and outcomes.

To connect the two fields, we designed a toolkit to identify gaps and overlaps. It includes a rubric to evaluate the documentation of the contents of datasets as well as the design decisions made in the process of developing datasets based on criteria adapted from the fields of digital and data curation, library, and archival studies. We applied the rubric on sample datasets from NeurIPS, the Conference on Neural Information Processing Systems, a leading global venue for AI/ML research. The design of the framework therefore moves towards the adoption of data curation principles and concepts by influencing evaluation standards. We analyzed the rubric evaluations to understand the entanglement of data practices in the disciplines and determine the feasibility and relevance of assessing ML data work using data curation perspectives. The process of designing and applying this rubric revealed strengths and weaknesses of current dataset development but also challenges in adapting principles from a data-focussed field like data curation for the model-focussed field of ML. We present our findings in four themes and discuss the limitations of adapting nuanced, practice-based processes from data curation into ML given their differing field epistemologies. We also present pathways to address the four challenges and make recommendations to further progress interdisciplinarity between the fields.

2 BACKGROUND

Below, we first review current practices of data work in machine learning research (Section 2.1) and briefly describe foundational data curation concepts (Section 2.2). We then discuss ML studies that start to bridge the fields of ML and data curation and archival studies (Section 2.3). Finally, we discuss why and how machine learning can adopt data curation to improve current data practices (Section 2.4) by extending current studies' use of data curation concepts.

2.1 Data Work in Machine Learning Research

In response to the call for accountability and transparency, the development of context documents became the prevalent method of demonstrating the data work involved in ML research. Datasheets, for example, are now a commonly used documentation framework for describing the contents of datasets and select data design decisions made by the dataset creators [42]. There are also specific structures of context documents for different types of datasets. For example, data statements for natural language processing (NLP) datasets contain specifications on demographic information about the dataset annotator, quality of the dataset, provenance, etc. [12]. Similarly, AI fairness checklists were developed to aid practitioners by providing a structured framework to identify and address issues within their projects [92]. Model cards aim to "standardize ethical practice and reporting" within ML models [101, p. 221]. Model cards include details about the models, their intended use, impacts of the model on the real-world, evaluation data, details on the training data, and ethical considerations [101]. Explainability fact sheets are used for similar documentation but are specifically geared towards the method applied in a predictive model. The fact sheet contains an evaluation of the method's functional and operational requirements, the criteria used for the evaluation, any security, privacy or other vulnerabilities that may be introduced by the method, and the results of this evaluation [129].

Simultaneously, dataset development research, sometimes referred to as data science work in ML, became a focal subject of study. Prominently many of these works unearthed how extrinsic and intrinsic biases impact the outcomes of ML models. For example, data cascades - "compounding events causing negative, downstream effects from data issues, that result in technical debt over time" [124, p. 5] - result from data practices being undervalued, lack of preparedness in handling data quality in high-stakes domains, data being reused out of context, and data scarcity causing potential downstream risks to groups. Documentation of computer vision datasets have also been analyzed to unearth the values that are prioritized by dataset creators and the field in general [125]. "The kinds of data collected, how it is collected, and how it is analyzed all reflect disciplinary and researcher values" [125, p. 4]. The results showcase that current practices of dataset development in ML prioritize model development over dataset development, efficiency over reflexive and critical curation, the collection of large, diverse datasets over emphasis on the context and circumstances of the data included in the dataset, and advocate for neutrality and impartiality in their data development process as compared to disclosing their positionality and worldviews [125]. Types of intrinsic biases that occur in ML projects have also been organized by building a "forgettance stack" with types of forgetting that occur throughout the ML pipeline [105]. "...forgetting in data science can also be harmful or cause violence, not least because our choice of what we deem unimportant enough to forget to improve our memory, impacts on our understanding of histories, data, exploitation, harm, and so on" [105, p. 3]. On the other hand, focussing on intrinsic biases is also seen as failing to acknowledge the power dynamics at play in situations [99]. By placing the focus on a bias-oriented framing rather than a power-oriented one, there is a loss of awareness of how labour conditions, social processes, and relationships between dataset creators and consumers impact the data bias present within ML models [99]. Instead, it is proposed that research must "...interrogate the set of power relations that inscribe specific forms of knowledge in machine learning datasets" [99, p. 9].

While many of these studies of dataset development discuss "data curation", the term is often used generally to discuss data collection [58, 84, 92]. Contrarily, data curation as a field takes an encompassing lifecycle view and considers many data work processes beyond data collection. The relevance of broader data curation studies to ML is rarely recognized, but several studies identify the opportunities in adopting practices from data curation into MLR.

2.2 Theoretical Framework of Data Curation

The information fields of archives, records management, and digital curation share principles, practices, challenges, and knowledge frameworks, but also diverge in areas. Data curation has been defined by institutions in varying ways, on occasion coupled with digital curation [107]. An important synthesis is made between perspectives that see data curation as digital curation, as value-added infrastructure service, and as an object of archival interest [107]. Data curation can be defined as, "...the activity of managing data throughout its life cycle; appropriately maintaining its integrity and authenticity; ensuring that it is properly appraised, selected, securely stored, and made accessible; and supporting its usability in subsequent technology environments." [107, p. 203].

The Digital Curation Center's lifecycle model consists of stages of curation that projects undergo and helps in identifying roles and responsibilities, processes and best practices, standards and policies, and their documentation [55]. The sequential stages of the DCC curation lifecycle model are 'conceptualize', 'create or receive', 'appraise and select', 'ingest', 'preservation action', 'store', 'access, use, and reuse', and 'transform' [55]. Data curation emphasizes that each stage of curation must be purposeful and attend to stewardship and future use [110]. The focus lends itself towards "improvement of data products" and ensuring data is valuable now and in the future [110]. For each stage of curation, technical, legal, ethical, and operational considerations are made.

2.3 Data Curation in Machine Learning Research

The existing body of knowledge in archival studies, data management, and data curation provide opportunities for adoption within dataset development in ML. Some ML studies have recognized this. For example, Jo and Gebru urge, "By showing the rigor applied to various aspects of the data collection and annotation process in archives, an industry of its own, we hope to convince the ML community that an interdisciplinary subfield should be formed…" [67, p. 307].

Archival science offers sophisticated methods of evaluating, filtering, and curating data that require a high degree of supervision and intervention. While this poses a challenge in some subfields of ML, lessons can be learned from archives around current key issues in ML including consent, inclusivity, power, transparency, and ethics [67]. For example, archives have codes of conduct and ethics to ensure violations do not occur and data curators consider and document whether data should be collected at all based on potential risks and benefits ensuring transparency and supervision of collected data. Leavy et al. further emphasize the importance of "... [enabling] critical reflection and responsibility for the potential effects of the use of data" [80, p. 695]. Their proposed framework for ethical curation consists of 4 principles detailing how to examine the power dynamics of whose voices, labour, and perspectives are included in data curation, how to consider the context and situatedness of data, how to recognize that data curation is a continuous and reflexive process, and how to question the forms of knowledge that are considered legitimate and are included in the data curation process as compared to those that are not.

Similar to Jo and Gebru, who point out the need for interventions in ML data, Bender et al. describe the risk of documentation debt due to large amounts of uncurated and undocumented data that is used to train large language models [13]. The lack of accountability and transparency lead to encoded bias in the datasets used for training. In turn, Bender et al. recommend "making time … for doing careful data curation and documentation, for engaging with stakeholders early in the design process…" [13, p. 619].

Research at the intersection of archives and ML often focuses on how algorithms can automate archival processes such as extraction, indexing and retrieval, appraisal, and redaction [25], but some emphasize "...the opportunity for recordkeeping contributions to the advancement and appropriate use of AI by bringing expertise on provenance, appraisal, contextualisation, transparency, and accountability to the world of data" [25, p. 11]. A critical archival approach is required towards datasets in AI to enable reflection on ethical issues such as access, consent, traceability, and accountability [134].

2.4 How Can Data Curation Benefit ML Data Work?

The ML model development pipeline consists of data collection, data processing, model building, training, model evaluation, and model deployment [49, 109]. Data curation has similar stages in its lifecycle model. For example, 'create or receive', 'appraise and select', and 'ingest' relate to data collection in ML, while 'transform' can involve data cleaning, data augmentation, and data wrangling in ML. However, data curation prioritizes two key aspects within the lifecycle that make it distinct from how dataset development is performed in ML.

First, data curation has defined inputs, outputs, outcomes, tasks, and reasons for performing each stage in the lifecycle. Importantly, all of these elements are defined and implemented through policies that hold curators and involved stakeholders accountable while also **enabling transparency**. The 'appraise and select' stage evaluates which data should be retained versus discarded for long-term curation. This process is interventionist and requires curators to make judgements on the benefits and risks of storing or discarding the data. In contrast, this is currently missing in ML dataset development where many subfields are driven by collecting the largest amount of data possible. In fact, ML publications introducing new datasets consider the size of the data collected an important contribution when discussing their work. On the other hand, the 'appraise and select' stage is performed for 5 reasons: to reduce the amount of data to be curated, to enable efficient retrieval, to enable timely preservation activities, to limit cost of data storage, and to capture legalities of data storage and access [57]. The tasks performed in this stage are documented through an appraisal policy which structures the process of making appraisal decisions among other agreed upon requirements for accessibility, retention, etc. The appraisal policy also supports the collection development policy which is an outcome of the prior stage, 'receive'. In the next stage, 'ingest', in which data is submitted for curation, the appraisal schedule is determined to ensure that there is timely reappraisal of the data being curated to determine needs for further retention and long-term value. These defined guidelines and expectations from each stage of the curation lifecycle enable reuse due to comprehensive documentation, the establishment of clear context and purpose for data curation, and high level of intervention that decreases the risk of introducing intrinsic bias and increases the likelihood of removing or addressing extrinsic bias. Similar standards and processes can be adopted into ML dataset development.

Secondly, data curation takes a **lifecycle approach** focusing on adding and maintaining long-term value across each stage, which is reflected in the norms, standards, and practices of data curation communities. The inclusion of stages like 'preservation' and 'access, use, and reuse' centralizes these reuse-oriented concerns in data curation. These concepts are considered not solely within their specified stages but throughout the dataset lifecycle. For example, considerations around long-term access inform the 'conceptualize' stage and data management methods throughout the lifecycle, the 'receive' stage identifies access and reuse rights, and the 'ingest' stage considers legal ownership issues. The data curation lens therefore not only provides standards and practices but also highlights the value of a cyclical view.

Pennock outlines the benefits of a lifecycle approach for digital curation, stating that digital materials change throughout their curation process and adopting a lifecycle model facilitates its continuous management [117]. This continuity lends itself to the ability to retain authenticity and integrity. A study of data curation at the ICPSR find that data work is often thought of as sequential and is represented through a pipeline but in actuality "data curation … is a highly collaborative process occurring across a distributed system over time" [133, p. 20].

Data curation supports greater reflexivity on the importance of each stage of data work. It highlights that data reuse now and in the future is dependent on a holistic approach for creating more transparent and accountable datasets which is only possible through meaningful dataset development. In the next section, we discuss the development of a resource that is aimed towards enabling critical dataset development in ML through a data curation lens.

3 METHODS

Below, we demonstrate how data curation concepts can be adapted, translated, and operationalized for ML data work.

3.1 Development Process

Our framework for evaluating ML datasets centers on a rubric developed in a multi-stage design pictured in Fig. 1. We started by identifying aspects of data curation currently used in ML dataset creation and those that can be further informed by data curation frameworks. Based on concept mapping between the two disciplines supported through literature reviews, we organized dimensions of data curation concepts and principles relevant to ML. We developed the rubric iteratively based on existing literature from digital curation lifecycle models [55], FAIR data principles [138], considerations of environmental sustainability and justice [10, 120], prior work on digital curation assessment frameworks [11], and current ML documentation frameworks. The framework builds on the significant impact of datasheets [42] and takes the logical next step. Datasheets [42] are focused largely on the content of datasets. Our rubric prompts dataset creators to adopt a reflexive stance about their curation decisions. The earliest drafts of the rubric went through an internal review process in which the authors iteratively discussed and improved the descriptions and evaluation criteria. This included adding, removing, splitting and grouping elements, narrowing down the data quality dimensions most apt for ML datasets, exploring qualitative and quantitative evaluation

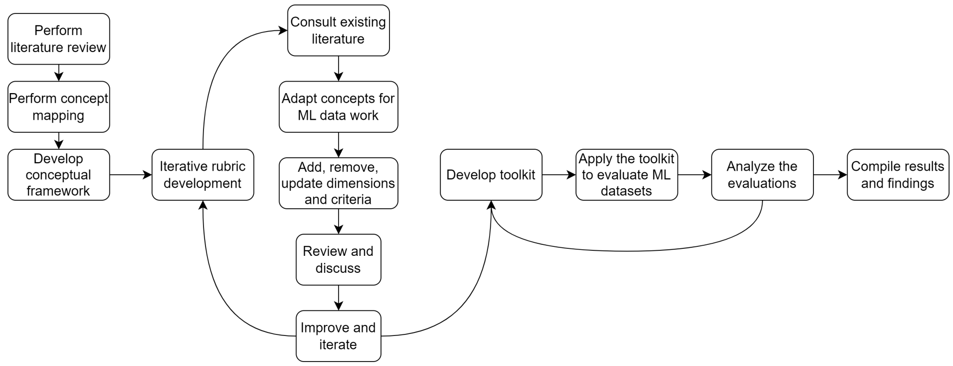

Figure 1: Multi-stage development and evaluation process of the rubric and toolkit

metrics, and arriving at two levels of evaluation, namely a minimum standard and a standard of excellence. After several iterations of the conceptual framework, we developed additional resources to support the use of the rubric, packaged together as a toolkit.

We used the toolkit to evaluate select datasets published in the NeurIPS benchmarks and datasets track [142]. We collected quantitative and qualitative results on the ratings and comments to understand how data curation is performed in ML, whether data curation principles were effectively adapted for ML datasets to enable feasible evaluations, whether there were elements that emerged as being irrelevant to evaluating ML datasets, and to study the reviewers' experience, feedback and reflections from applying the rubric. The ratings and results contributed to iterative revisions of the toolkit. In addition, in the final set of evaluations, the reviewers re-examined the evaluations performed for each dataset, and asynchronously resolved disagreements in the ratings by providing comments on whether they agreed or disagreed with another review and accordingly updating their evaluation rating. The reviewers collaboratively discussed the remaining disagreements which helped in further refinement of the rubric and toolkit. In the following sections, we outline the contents of the rubric and toolkit. In Section 4, we present our observations and findings from using the rubric to evaluate ML datasets.

3.2 Rubric

The rubric elements assess the documentation of data composition and data design decisions (i.e., data work) in 19 dimensions across five groups. The full rubric is provided in Appendix A. Below, we briefly discuss a few sample elements within each group.

Scope contains the elements '*context, purpose, motivation*' and '*requirements*'. The latter element's criteria expect 1) a dataset creation plan and 2) considering how problem formulations can introduce intrinsic biases. This echoes data curation's emphasis on data management plans that are established at the beginning to guide the entire curation process. The rubric contains these elements because establishing the scope is "...a translation task from a problem in-the-world, into a problem in-the-business, and then into a data science formulation…Each translation step requires additional interpretation into data sources and data formulations, imposing further decisions upon the humans who carry out the work" [105, p. 9]. Capturing these decisions through documentation helps unveil the politics and values involved in setting scope [105, 112, 118].

There is an emphasis on reflexivity throughout the rubric, such as being intentional and accountable *while* deciding on the purpose for creating a dataset, but a group of elements are centrally concerned with **ethicality and reflexivity**: '*ethicality*', '*domain knowledge and data practices*', '*context awareness*', and '*environmental footprint*'. The criteria for evaluating 'ethicality' includes a discussion of informed consent and weighing benefits and harms of the dataset. The criteria expects dataset creators to demonstrate 'context awareness' by looking inwards and considering how their dataset is a non-neutral representation of the real-world impacted by their perspectives, field epistemologies in which their research is situated, and social, political, and historical context [127]. Dataset creators are also asked to document how their 'domain knowledge' expertise and 'data practices' shape the dataset. Curatorial work requires craft and unstandardized methods: "...curators organize their work by first developing a gestalt, abstract mental representation of the data to envision what the final released dataset will entail; they then use their judgement and expertise to interpret standards, [and] creatively come up with solutions…" [133, p. 13]. Documentation of this tacit knowledge makes it explicit which supports informed choices about reuse. This is supported by Heger et al.'s findings which discuss that ML practitioners "...noted that information that is implicit or tacit is at risk of being lost if it is not documented" [51, p. 13].

Elements that document **key stages of the ML pipeline** are included in the rubric because they demonstrate the foundation of how the dataset was developed, namely '*data collection*', '*data processing*', and '*data annotation*'. While these elements are familiar to dataset creators, the rubric offers the opportunity to approach these elements from a different perspective. For example, aside from disclosing the data sources from which data was collected, the rubric urges reflection on how choices in 'data collection' have embedded interpretative assumptions because the act of selecting data or "discovering" data, especially one source over another, is a human, subjective act that involves interpretation [105]. The rubric also prompts for reflexivity *in the process* of 'data collection' rather than at its end. It suggests that criteria for selecting data sources should be discussed and decided prior to its collection in an active process of assessing whether data sources fit the criteria.

Ultimately this process must be documented so that data reuse is more transparent, similar to collection development policies in data curation.

The rubric underscores the application of the data curation lens through the elements about **data quality dimensions**, including '*suitability*', '*representativeness*', '*authenticity*', '*reliability*' and '*integrity*', along with '*structured documentation*'. 'Suitability' prompts dataset creators to reflect on whether their dataset aligns with the purpose they established at the start of the dataset development process and whether the quality of the dataset enables the fulfillment of that purpose. 'Representativeness' is included to promote awareness of introducing extrinsic biases through data collection. Dataset creators are asked to define the population represented in their dataset and comment on whether a representative sample is included. 'Authenticity', 'reliability', and 'integrity' are inter-related elements but are analytically separate concepts and specifically defined in archival and digital curation fields. An authentic dataset is one that "is what it purports to be" [32–34, 56, 119]. This means that the development of the dataset should include discussion of how 'authenticity' was established i.e., how the dataset creators verified the origin of the data they collected. Additionally, it should discuss how authenticity is impacted once the collected data is preprocessed and how the now derived dataset will continue to maintain authenticity. Establishing this chain of authenticity ensures that the dataset that is created is based on verified data and the future reuse of the new dataset can also have a claim of authenticity. A reliable dataset is one that is "capable of standing for the facts to which it attests" i.e., that the data points reflect what they represent [32]. The rubric prompts the assessment of the maintenance of 'reliability' while creating the dataset and how reliability can be maintained once the dataset is reused. A dataset with 'integrity' is one where "the material is complete and unaltered" [14, 21, 35, 56, 102]. The rubric prompts evaluators to check whether dataset creators discuss how integrity has been maintained during dataset creation and future management of integrity. Lastly, we include the 'structured documentation' element within this category as the rubric prompts evaluators to assess whether a context document was included to provide documentation about the quality of the dataset's contents.

To increase transparency and *appropriate* reuse of datasets in ML, the rubric adopts and adapts the widely used **FAIR principles for data management** [138]. The FAIR (*findability, accessibility, interoperability, reusability*) principles were first produced to improve the stewardship and management of research datasets but since then have been adopted into numerous disciplines, including AI/ML [7, 39, 66, 88, 108]. In the rubric, documentation for each of 'findability', 'accessibility', 'interoperability', and 'reusability' is prompted as individual elements with the principles split into minimum standard and standard of excellence based on their importance and relevance for ML datasets [138]. Inclusion of the principles in the rubric enables increased transparency and reusability while fostering improved collaboration.

3.3 Toolkit

The conceptual framework of the rubric is complemented with 1) instructions detailing how dataset creators can use the rubric to evaluate their own processes and how dataset re-users (or reviewers) can evaluate existing datasets, 2) guiding principles, recommendations, and FAQ to help in evaluating datasets using the rubric, 3) guidance on interpreting the FAIR principles and authenticity, reliability, integrity, and representativeness, 4) a glossary, 5) and sample evaluations. The toolkit is provided in Appendix B.

The rubric is used to evaluate a minimum standard and a standard of excellence. The former is evaluated on a pass/fail basis, the latter using none/partial/full. The minimum standard criteria relay the expected level of documentation from all ML datasets while the standard of excellence criteria advocates for a high level of criticality and the documentation only receives "full" when all sub-criteria are satisfied. The guiding principles, recommendations, and FAQ sections provide overarching suggestions such as how to approach the evaluation of a dataset that has multiple sources of documentation such as the publication, appendix, website, GitHub page, etc.

Additional guidance is provided for the data quality dimensions 'representativeness', 'authenticity', 'reliability', and 'integrity' as these elements must distinctly be evaluated from an archival and digital curation perspective. For example, 'representativeness' is related to 'reliability' but more closely focussed on whether the dataset accurately represents the overall set of observations or entities that it claims to be a sample of. Similar guidance is provided around the FAIR principles with simplified explanations and links to self assessment tools and checklists based on the FAIR principles.

4 FINDINGS

4.1 Application

In order to study whether data curation concepts were feasible for ML dataset development in practice, a set of authors with varied exposure to both ML and digital curation fields conducted a sample set of evaluations using datasets published in NeurIPS. Further information about the authors' expertise is discussed in Appendix C.1. The evaluations were conducted in four rounds (training, round 1, round 2, and round 3).

We started with a training round so reviewers could become adept with applying the rubric, become familiarized with new concepts and terminology using the toolkit as supplementary material, and ask questions to improve their understanding. The training round consisted of 5 randomly selected datasets published in the NeurIPS benchmarks and datasets track from 2021-2023. Next, three rounds of evaluations were performed on (20) randomly selected datasets; the first round consisted of 10 datasets and the remaining of 5 each. The datasets are listed in Appendix C.2. Following each round, we worked on resolving any disagreements, questions and feedback by improving the rubric and toolkit, and addressing any concerns raised by the reviewers.

We analyzed the ratings and comments for all the evaluations by measuring inter-rater reliability (IRR). We calculated IRR by using two-way mixed, consistency, average-measures intra-class coefficient (ICC) given our fully crossed design to assess the consistency of the raters' evaluations of rubric elements across subjects [96]. Since the ratings for the variables (i.e., rubric elements) were measured on an ordinal scale (i.e., full, partial, none and pass, fail), the ICC was the best suited statistic to assess IRR [47]. ICC values of 1 indicate perfect or complete agreement, 0 indicates random

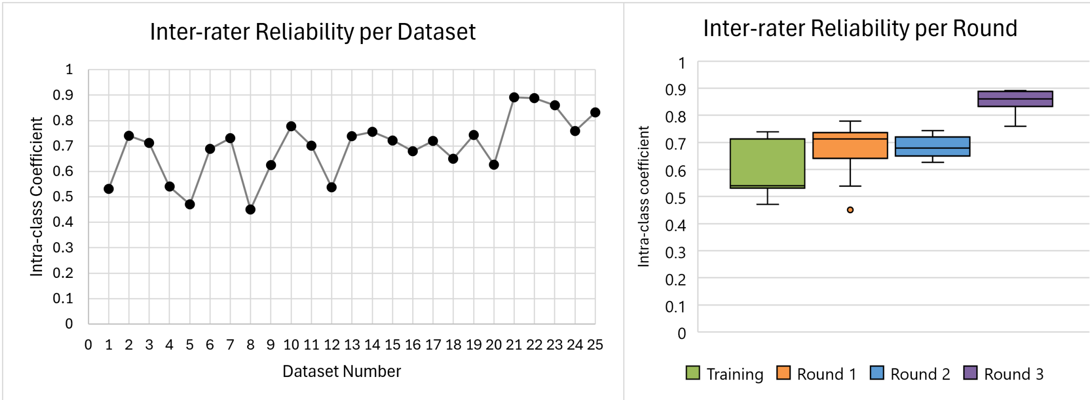

Figure 2: IRR across datasets and rounds

agreement, and negative values indicate systematic disagreement. ICC values of less than 0.40 indicate poor IRR, values between 0.40 and 0.59 indicate fair IRR, values between 0.60 and 0.74 indicate good IRR, and values between 0.75 and 1.0 indicate excellent IRR [24].

Fig. 2 shows the progression of IRR from training (datasets 1-5), round 1 (datasets 6-15), round 2 (datasets 16-20), and round 3 (datasets 21-25). The lowest ICC value is 0.45 for dataset 8 which indicates fair agreement while the highest is 0.89 for dataset 21 which indicates excellent agreement. Across datasets, the IRR values span from fair to excellent which indicates the difficulty in obtaining truly consistent evaluations. Nonetheless, the distribution of IRR per round shows lower variability in the consistency as the rounds progress indicating gradual improvement through iterations. Furthermore, round 3 has all 5 ICC values indicating excellent agreement (ICCs between 0.76 and 0.89). We also compared reliability across elements, which had mixed results, as discussed further in Appendix C.3.

To assess the extent to which the rubric and toolkit improvements between the evaluation rounds were impacting the consistency of the evaluations, we analyzed the disagreements by datasets and elements. A summary of the average number of inconsistencies across datasets can be found in Appendix C.4. Most importantly, the metric demonstrates that the overall percentage of all disagreements decreased from training (32%),round 1 (25%), round 2 (23%), to round 3 (7%) indicating that the iterative development of the rubric was improving the consistency of the evaluations.

We reviewed the inconsistencies across each element of the rubric to determine whether any specific elements were standing out as infeasible to adapt from data curation to ML or required further improvements to adapt. To measure this, we calculated the percentage of datasets with inconsistencies for each rubric element as shown in Appendix C.5.

We analyzed responses after each round and consequently introduced changes to the toolkit that would reduce inconsistencies iteratively. As a result, we were able to identify improvements in some rubric elements and recurrent challenges in others. Difficulties with curation-specific terminology such as the difference between 'findability' and 'accessibility' was addressed by clearer definitions and examples. Other difficulties, such as the applicability of data quality evaluation for datasets that were synthetic (not collected) were addressed with better guidance. Lastly, 'reliability', 'authenticity', and 'integrity' were difficult to evaluate because while the documentation provided by dataset creators acknowledged the limitations of the dataset, it was not in terms that addressed these elements specifically.

To get a better understanding of the reason for the inconsistencies after round 2, we analyzed the evaluations by comparing the evaluation comments against a "reference" comment. Based on this analysis, specific patterns emerged on the reasons for disparate evaluations between raters. These reasons, in turn, revealed four types of challenges in applying data curation concepts to ML contexts. A sample set of analyzed evaluations is presented in Table 1, and discussed in further detail in the following section.

In response to the range of evaluation outcomes for the 'structured documentation' dimension, we reviewed the current data practices reported by the dataset creators in more detail by analyzing the context documents provided with publications. Our review, detailed in Appendix C.6, indicates that out of 25 datasets assessed, 6 lacked an accompanying context document. Of the 19 datasets with context documentation, we identified limitations that undermine their completeness and utility. This review highlights instances where modifications to standard datasheets or checklists by dataset creators lead to the omission of essential curation details. We document cases where the provided information was ambiguous or could not be independently verified, emphasizing the need for improved documentation standards to uphold the integrity of data curation processes.

4.2 Challenges

Table 1 introduces four challenges we identified through the evaluation results. They illustrate the difficulties of designing an evaluation framework that assesses ML concepts using data curation principles. These challenges are not comprehensive but serve as a demonstration of salient issues in this interdisciplinary space.

4.2.1 False Friends. Some elements refer to terms shared between data curation and ML (or computing broadly) that have non-shared meanings *(false friends)*. For example, 'reliability' in engineering and computing disciplines refers to expected consistency in performance (i.e., that a system will perform as expected in a given time period and environment). For datasets, this is often interpreted as the trustworthiness of data in terms of accuracy and consistency [137]. However in data curation, reliability is defined as whether data is "capable of standing for the facts to which it attests" [32]. For the example provided in Table 1, raters evaluated the standard of excellence for 'reliability' for dataset 19, which has criteria stating that the documentation discusses the management of reliability for appropriate reuse in the future, i.e., how the dataset structure and documentation enable reliable re-purposing and reuse. Interpreting this criteria as "dataset reliability" leads to consideration towards whether the dataset would be accurate over time for reuse and consistently available. Accordingly, Rater 3's evaluation points to a discussion on maintenance and findability of the dataset rather than an evaluation of processes in place to ensure that the dataset will continue to represent the information it is about even when it is reused and repurposed.

4.2.2 Interpretative Flexibility. The rubric's more open ended criteria lead to *interpretative flexibility*, which can result in divergent ratings. For example, the evaluation of ethicality in dataset 20 surfaced how different standards and expectations can collide, resulting in a full range of evaluations. While one rater was fully satisfied by a discussion of potential negative impacts (full), another recognized these statements as typical but expected more (partial), and the third considered them insufficient (none).

4.2.3 Depth of Analysis. The third challenge arose as a result of reviewers bringing differing expertise and different technical know-how to evaluating an element, but the important question is how deep an evaluation can and should go beyond surface documentation. Table 1 points to an example of this for evaluating 'interoperability' for dataset 17. The criteria direct reviewers to assess whether the metadata and data are readable by humans and machines. This can be interpreted by evaluating whether the dataset is made available in a standardized and documented format. Data format standardization however has multiple levels. For example, even a structurally simple standardized 'container' format such as CSV must be complemented with clear definitions of each column. For more complex data, the recursive analysis and exhaustive models of dependency networks can become effort intensive [44].

4.2.4 Scoping. The last challenge in designing the rubric was *scoping* the expected standard of documentation from dataset creators. For example, in evaluating the maintenance of integrity while developing the dataset (minimum standard) and management of integrity for appropriate reuse in the future (standard of excellence), it is challenging to scope which points in the data pipeline the dataset creators are responsible for documenting processes around integrity. In the example of dataset 16, raters reported confusion around whether the integrity should be evaluated based on the integrity of the collected data, or the maintenance of integrity in the data pipeline, or the integrity of the final produced dataset. Similarly, in the example shown in Table 1 of dataset 19, 'domain knowledge and data practices' was challenging to evaluate because it was unclear whether expertise in collecting the data, the problem domain overall, or developing the dataset needed to be documented.

5 DISCUSSION

5.1 Limitations

We identify two key limitations in the application of the rubric. First, using the rubric to evaluate ML datasets requires training, practice, and familiarity with data curation concepts. Performing evaluations iteratively and taking part in workshops and discussions help improve the required data curation knowledge. This also creates a potential scenario in which ML experts may be expected to acquire an unreasonable amount of expertise in data curation prior to applying the rubric. Uptake of such a rubric requiring specialized knowledge and skills that are improved over time is presently a limitation on the immediate resolution of using data curation to improve fairness, accountability, and transparency in ML dataset development.

Second, our current evaluation framework is used to explore the connections between data curation and ML dataset development through its application on a select set of datasets where evaluations are performed by a select set of reviewers. In addition, the reviewers were trained in using the rubric. Furthermore, the results from the application of the rubric are on the basis of randomly selected datasets that *aim* to represent ML datasets at large. This means that our findings are predicated on these factors. This further implies that the improvements made to the toolkit are on the basis of difficulties faced in evaluating a sample set of datasets. We report IRR metrics that showcased improved consistency in responses between each round. However, we cannot distinguish to what degree the ICC values improve because the toolkit was updated and improved after each round, or because the reviewers became more consistent at interpreting and evaluating the rubric criteria.

5.2 Pathways Forward

We outline some recommendations to address the challenges based on the lessons learned from applying a data curation lens to examine ML data practices. These challenges combine problems that can be fixed with tensions that will remain present and need navigation, thus they present opportunities for growth between the disciplines through continued exploration of the intersections in data practices, including further toolkit development.

The presence of *false friends* across fields suggests that evolving documentation can aid with defining, understanding, and navigating the differences in shared terms. Toolkit components like the glossary and FAQ can provide evolving required context to ensure that shared terms between ML and data curation are evaluated as intended.

The challenge of *interpretative flexibility* presents an opportunity to engage in generative discussion and collaboration that broadens the association between data curation and ML dataset development. As with any form of descriptive evaluation, the rubric necessitates interpretation. The recent emergence of research intersecting these fields means that evaluation across the disciplines is complex. One consideration for generative discussion is to what extent (and if it all) the evaluators need to agree in their assessments. It can be argued that a better approach would be to embrace the flexibility of the evaluations within the format of the rubric and create an evaluation framework that doesn't result in ratings and comments but questions and recommendations to foster collaboration instead

Table 1: Sample set of round 2 evaluations and challenges

Dataset	Element	Evaluation Comments (paraphrased)	Reference Comment	Reason for Inconsistency	Challenge
19	Reliability, standard of excellence	Rater 2 (none) mentioned that there was no specific discussion of reliability as it pertains to reuse. Rater 1 (partial) pointed to the maintenance section of the datasheet. Rater 3 (full) pointed to a DOI and maintenance plan as assurance for long-term reliability.	I would rate this as none. Despite the maintenance section in the datasheet, the response does not discuss maintenance as it pertains to maintaining reliability when the dataset is repurposed and reused.	Reliability is interpreted from a software or computing perspective which considers consistent performance rather than a data curation perspective which considers how data will remain true to the facts it represents through reuse.	False friends
20	Ethicality, standard of excellence	Rater 1 (none) stated that the documentation doesn't go beyond standard ethics statements. Rater 2 (partial) stated that documentation on potential negative impacts is identified. Rater 3 (none) states that there is no identifiable risk in this dataset.	I would rate this as a none because there is no further discussion of ethics beyond typical negative impacts statements.	Rater 3 interprets this dataset as being as low-risk for ethicality and doesn't believe there is a need to "go beyond requirements listed in ethics framings".	Interpretative flexibility
17	Interoperability, both levels	Rater 1 gave a fail (minimum) and none (excellent) and mentioned there was no explicit documentation of how the dataset integrates with other workflows. Rater 2 (also fail/none) mentioned that machine and human readability is discussed implicitly because data is in a CSV format but fails for lack of discussion on integration. Rater 3 (pass/full) mentioned all relevant info was given on GitHub.	I would rate this as a pass for minimum standard because data is in a popular, standard format. I would rate this as none for excellence because controlled vocabularies and qualified references for linking were not used/discussed.	It is difficult to decide to which extent human and machine readability should be evaluated. The reference comment indicates that a popular, standard format is sufficient. However, while CSV is a popular format, it can only be processed if all columns are fully defined. The reviewers would need expertise about multiple data formats and their structures to fully assess this.	Depth of analysis
19	Domain knowledge and data practices, minimum standard	Raters 1 and 2 gave a fail because there was no explicit documentation about this element. Rater 3 gave a pass, and mentioned that expertise is required in curation, web crawling, and natural language processing.	I would rate this as a fail, because there is no explicit discussion on how the process of developing this dataset required special skills/expertise.	The documentation describes the curation of LLMs as intensive and specialized (presented as a description of the problem domain). This is however not a description of the knowledges required to develop this dataset. The challenge for the raters is to interpret the extent of documentation the dataset creators are responsible for.	Scoping

of dissonance. In fact, one of the identified challenges of enforcing triangulation is that it acts as a barrier to collaboration [5]. Instead, approaching interpretative disagreements as a way to understand evaluators' perspectives can prompt deeper reflexivity [5, 68]. This can be especially helpful in progressing the interdisciplinarity between the fields at this early stage of intersection.

The challenge of *depth of analysis* is linked to *interpretative flexibility* because requiring agreement in evaluations means requiring identical levels and types of expertise from the reviewers. In other words, the optimal depth of analysis will vary because depth of expertise varies and disagreements happen on different levels. However, as we discussed above, if the evaluation framework does not require agreement among reviewers, the disagreements arising due to *depth of analysis* can become prompts for deeper levels of assessment. Disagreements would then become generative and would be used as a starting point for discussion.

The related challenge of *scoping* occurs due to inevitable entanglement of data curation and ML. The processes of curation and dataset development are inseparable in practice, yet conceptually separable even when occurring contemporaneously. Setting clear boundaries on the expectations from data creators can aid in scoping the documentation they are responsible for. But as datasets regularly reuse prior datasets, it is not easy to determine the appropriate boundary of responsibility for the quality of data curation. A guiding principle for this boundary, adopted from data curation, can be to maintain the chain of custody, i.e., dataset creators should be expected to provide all possible documentation relating to their processes and pointing to others' documentation for processes outside their control.

6 CONCLUSION

Jo and Gebru "hope[d] to convince the ML community that an interdisciplinary subfield should be formed…" [67, p. 307]. In order to make sense of the intersecting terminologies and concepts in this interdisciplinary space, we must develop the right tools. Here, we explore what form and content these tools might take. The paper explored the intertwined relationship of data practices in data curation and ML and presented a method for how data curation concepts can be adapted for ML dataset development. The process of exploring this intersection of fields yielded a high-level framework of dimensions and criteria as well as insights into the challenges of merging these fields' perspectives. We adopted standards for transparency and accountability built into data curation processes to evaluate the documentation of dataset development in ML.

Based on our data, we claim that the evaluation enabled by the framework identifies strengths and weaknesses in order to prioritize targeted improvements by incorporating data curation methods where they are most needed. As a diagnostic aid, the formative evaluation helps ML practitioners decide how to improve their dataset's documentation and develop staged objectives to improve their practices. Aggregate evaluation results highlight priorities, such as environmental footprint disclosures. By incorporating data curation norms, evaluation criteria, and terminology into evaluation guidelines for ML, the framework contributes to normalizing the idea that data curation is part of ML and guides the community in systematically addressing and evaluating it.

This work answers calls for data curation in AI/ML [25, 67], supports the examination of intrinsic and extrinsic biases in the dataset development process, and facilitates greater reflexivity [80]. Our results demonstrate the potential of collaboration between data curation and ML data work, with the toolkit as a resource for bridging the gap in practice.

ACKNOWLEDGMENTS

This research was partially supported by NSERC through RGPIN-2016-06640 and the Canada Foundation for Innovation.

REFERENCES

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B. Moeslund. 2021. RELLISUR: A Real Low-Light Image Super-Resolution Dataset. *Advances in Neural Information Processing Systems*.

[2] Daron Acemoglu, David Autor, Jonathon Hazell, and Pascual Restrepo. 2022. Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics* 40, S1 (April 2022), S293–S340. https://doi.org/10.1086/718327

[3] Zo Ahmed, Bertie Vidgen, and Scott A. Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science* 11, 1 (Feb. 2022), 8. https://doi.org/10.1140/epjds/s13688-022-00319-9

[4] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D'Ambra, and K.N. Shen. 2021. Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management* 60 (Oct. 2021), 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

[5] Mandy M. Archibald. 2016. Investigator Triangulation: A Collaborative Strategy With Potential for Mixed Methods Research. *Journal of Mixed Methods Research* 10, 3 (July 2016), 228–250. https://doi.org/10.1177/1558689815570092

[6] Sapna Arora, Ruchi Kawatra, and Manisha Agarwal. 2021. An Empirical Study - The Cardinal Factors towards Recruitment of Faculty in Higher Educational Institutions using Machine Learning. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*. 491–497. https://doi.org/10.1109/SPIN52536.2021.9566057 ISSN: 2688-769X.

[7] Nongnuch Artrith, Keith T. Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. 2021. Best practices in machine learning for chemistry. *Nature Chemistry* 13, 6 (June 2021), 505–508. https://doi.org/10.1038/s41557-021-00716-z

[8] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 675–684. https://doi.org/10.1145/1978942.1979041

[9] Ransome Epie Bawack, Samuel Fosso Wamba, Kevin Daniel André Carillo, and Shahriar Akter. 2022. Artificial intelligence in E-Commerce: a bibliometric study and literature review. *Electronic Markets* 32, 1 (March 2022), 297–338. https://doi.org/10.1007/s12525-022-00537-z

[10] Christoph Becker. 2023. *Insolvent: How to Reorient Computing for Just Sustainability*. MIT Press.

[11] Christoph Becker, Emily Maemura, and Nathan Moles. 2020. The Design and Use of Assessment Frameworks in Digital Curation. *Journal of the Association for Information Science and Technology* 71, 1 (2020), 55–68. https://doi.org/10.1002/asi.24209

[12] 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. https://doi.org/10.1162/tacl_a_00041

[13] 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* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[14] June M. Besek and Philippa S. Loengard. 2007. Maintaining the Integrity of Digital Archives. *Columbia Journal of Law & the Arts* 31 (2007), 267.

[15] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 619, 7970 (July 2023), 533–538. https://doi.org/10.1038/s41586-023-06185-3

[16] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, Athens Greece, 981–993. https://doi.org/10.1145/3468264.3468536

[17] 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. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[18] Christine L. Borgman. 2017. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

[19] Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–27. https://doi.org/10.1145/3479582

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[21] Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14 (May 2015), 2–2. https://doi.org/10.5334/dsj-2015-002

[22] Digital Curation Centre. [n. d.]. What is digital curation? https://www.dcc.ac.uk/about/digital-curation.

[23] Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A Dataset for Answering Time-Sensitive Questions. *Advances in Neural Information Processing Systems* (2021).

[24] Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6, 4 (Dec. 1994), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

[25] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2022. Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage* 15, 1 (Feb. 2022), 1–15. https://doi.org/10.1145/3479010

[26] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future Healthcare Journal* 6, 2 (June 2019), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

[27] Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. 2023. American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers. Advances in Neural Information Processing Systems.

[28] Digital Curation Centre. [n. d.]. Glossary. https://www.dcc.ac.uk/about/digital-curation/glossary

[29] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data Feminism*. MIT Press.

[30] Yanqing Duan, John S. Edwards, and Yogesh K Dwivedi. 2019. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management* 48 (Oct. 2019), 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

[31] Laurine Duchesne, Efthymios Karangelos, and Louis Wehenkel. 2020. Recent Developments in Machine Learning for Energy Systems Reliability Management. *Proc. IEEE* 108, 9 (Sept. 2020), 1656–1676. https://doi.org/10.1109/JPROC.2020.2988715

[32] Luciana Duranti. 1995. Reliability and Authenticity: The Concepts and Their Implications. *Archivaria* (May 1995), 5–10. https://archivaria.ca/index.php/archivaria/article/view/12063

[33] Luciana Duranti. 2005. The long-term preservation of accurate and authentic digital data: the INTERPARES project. *Data Science Journal* 4 (2005), 106–118. https://doi.org/10.2481/dsj.4.106

[34] Luciana Duranti. 2007. The InterPARES 2 Project (2002-2007): An Overview. *Archivaria* (2007), 113–121. https://www.archivaria.ca/index.php/archivaria/article/view/13155

[35] Luciana Duranti and Heather MacNeil. 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project. *Archivaria* (Oct. 1996), 46–67. https://archivaria.ca/index.php/archivaria/article/view/12153

[36] GO FAIR. 2017. F1: (Meta) data are assigned globally unique and persistent identifiers. https://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/.

[37] GO FAIR. 2017. I3: (Meta)data include qualified references to other (meta)data. https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/.

[38] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (Jan. 2018). https://doi.org/10.1177/2056305118763366

[39] Sakinat Folorunso, Ezekiel Ogundepo, Mariam Basajja, Joseph Awotunde, Abdullahi Kawu, Francisca Oladipo, and Abdullahi Ibrahim. 2022. FAIR Machine Learning Model Pipeline Implementation of COVID-19 Data. *Data Intelligence* 4, 4 (Oct. 2022), 971–990. https://doi.org/10.1162/dint_a_00182

[40] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems.*

[41] Swati Garg, Shuchi Sinha, Arpan Kumar Kar, and Mauricio Mani. 2021. A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management* 71, 5 (Jan. 2021), 1590–1610. https://doi.org/10.1108/IJPPM-08-2020-0427

[42] 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 (Nov. 2021), 86–92. https://doi.org/10.1145/3458723

[43] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out?: do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of FAT* '20.* ACM, Barcelona Spain, 325–336. https://doi.org/10.1145/3351095.3372862

[44] David Giaretta. 2011. *Advanced Digital Preservation.* Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16809-3

[45] GO FAIR. [n. d.]. R1.2: (Meta)data are associated with detailed provenance. https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/

[46] Carlos Vladimiro Gonzalez Zelaya. 2019. Towards Explaining the Effects of Data Preprocessing on Machine Learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE).* 2086–2090. https://doi.org/10.1109/ICDE.2019.00245

[47] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23–34. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

[48] Eric Hambro, Roberta Raileanu, Danielle Rothermel, Vegard Mella, Tim Rocktäschel, Heinrich Kuttler, and Naila Murray. 2022. Dungeons and Data: A Large-Scale NetHack Dataset. Advances in Neural Information Processing Systems.

[49] Hannes Hapke and Catherine Nelson. 2020. *Building Machine Learning Pipelines.* "O'Reilly Media, Inc.".

[50] Sheikh Md Shakeel Hassan, Arthur Feeney, Akash Dhruv, Jihoon Kim, Youngjoon Suh, Jaiyoung Ryu, Yoonjin Won, and Aparna Chandramowlishwaran. 2023. BubbleML: A Multiphase Multiphysics Dataset and Benchmarks for Machine Learning. *Advances in Neural Information Processing Systems.*

[51] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–29. https://doi.org/10.1145/3555760

[52] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 793–811. https://doi.org/10.1007/978-3-030-01219-9_47

[53] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *Advances in Neural Information Processing Systems.*

[54] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura. 2019. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications* 124 (June 2019), 226–251. https://doi.org/10.1016/j.eswa.2019.01.012

[55] Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3 (Aug. 2008), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

[56] Sarah Higgins. 2009. DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle. *International Journal of Digital Curation* 4, 22 (Oct. 2009), 60–67. https://doi.org/10.2218/ijdc.v4i2.93

[57] Sarah Higgins. 2012. The lifecycle of data management. In *Managing Research Data* (1 ed.), Graham Pryor (Ed.). Facet, 17–46. https://doi.org/10.29085/9781856048910.003

[58] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proc. CHI'2019.* ACM, Glasgow Scotland Uk, 1–16. https://doi.org/10.1145/3290605.3300830

[59] Thibaut Horel, Lorenzo Masoero, Raj Agrawal, Daria Roithmayr, and Trevor Campbell. 2021. The CPD Data Set: Personnel, Use of Force, and Complaints in the Chicago Police Department. *Advances in Neural Information Processing Systems.*

[60] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, and Honglak Lee. 2022. CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition. *Advances in Neural Information Processing Systems.*

[61] Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. 2022. DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection. *Advances in Neural Information Processing Systems.*

[62] Zhe Huang, Liang Wang, Giles Blaney, Christopher Slaughter, Devon McKeon, Ziyu Zhou, Robert Jacob, and Michael C. Hughes. 2021. The Tufts fNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces that Generalize. *Advances in Neural Information Processing Systems.*

[63] 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 (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[64] Information and Privacy Commissioner of Ontario. [n. d.]. Consent may be implied in some cases. https://www.ipc.on.ca/part-x-cyfsa/consent-and-capacity/elements-of-consent/consent-may-be-implied-in-some-cases/

[65] Md Mofijul Islam, Reza Manuel Mirzaiee, Alexi Gladstone, Haley N. Green, and Tariq Iqbal. 2022. CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets. *Advances in Neural Information Processing Systems.*

[66] Ashish Kumar Jha, Sneha Mithun, Umesh B. Sherkhane, Vinay Jaiswar, Zhenwei Shi, Petros Kalendralis, Chaitanya Kulkarni, M. S. Dinesh, R. Rajamenakshi, Gaur Sunder, Nilendu Purandare, Leonard Wee, V. Rangarajan, Johan van Soest, and Andre Dekker. 2022. Implementation of Big Imaging Data Pipeline Adhering to FAIR Principles for Federated Machine Learning in Oncology. *IEEE Transactions on Radiation and Plasma Medical Sciences* 6, 2 (Feb. 2022), 207–213. https://doi.org/10.1109/TRPMS.2021.3113860

[67] 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.* ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[68] P. Lynne Johnstone. 2007. Weighing up triangulating and contradictory evidence in mixed methods organisational research. *International Journal of Multiple Research Approaches* 1, 1 (Oct. 2007), 27–38. https://doi.org/10.5172/mra.455.1.1.27

[69] Julian Posada. 2023. *Platform Authority and Data Quality.* Technical Report. https://www.berggruen.org/ideas/articles/decoding-digital-authoritarianism/

[70] Julia Kaltenborn, Charlotte Emilie Elektra Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. 2023. ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning. *Advances in Neural Information Processing Systems.*

[71] Bryan Kelly and Dacheng Xiu. 2023. Financial Machine Learning. *Foundations and Trends® in Finance* 13, 3-4 (Nov. 2023), 205–363. https://doi.org/10.1561/0500000064

[72] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. (Sept. 2022). https://doi.org/10.2139/ssrn.4217148

[73] Kim Martineau. 2021. What is synthetic data? https://research.ibm.com/blog/what-is-synthetic-data

[74] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9, 12 (Dec. 2014), e115253. https://doi.org/10.1371/journal.pone.0115253

[75] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research.

Advances in Neural Information Processing Systems (2021).
[76] Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300330
[77] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Shangzhe Wu, and Jiajun Wu. 2023. Stanford-ORB: A Real-World 3D Object Inverse Rendering Benchmark. *Advances in Neural Information Processing Systems*.
[78] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. 2023. Learning skillful medium-range global weather forecasting. *Science* 382, 6677 (Dec. 2023), 1416–1421. https://doi.org/10.1126/science.adi2336
[79] Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating Out-of-Distribution Performance on Document Image Classifiers. Advances in Neural Information Processing Systems.
[80] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. 2021. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In *Proc. of 2021 AAAI/ACM Conf. on AI, Ethics, and Society*. ACM, Virtual Event USA, 695–703. https://dl.acm.org/doi/10.1145/3461702.3462598
[81] Heidi Ledford. 2019. Millions of black people affected by racial bias in health-care algorithms. *Nature* 574, 7780 (Oct. 2019), 608–609. https://doi.org/10.1038/d41586-019-03228-6
[82] Jiyoung Lee, Seungho Kim, Seunghyun Won, Joonseok Lee, Marzyeh Ghassemi, James Thorne, Jaeseok Choi, O.-Kil Kwon, and Edward Choi. 2023. VisAlign: Dataset for Measuring the Alignment between AI and Humans in Visual Perception. Advances in Neural Information Processing Systems.
[83] Ramona Leenings, Nils R. Winter, Udo Dannlowski, and Tim Hahn. 2022. Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage* 257 (Aug. 2022), 119298. https://doi.org/10.1016/j.neuroimage.2022.119298
[84] Nianyun Li, Naman Goel, and Elliott Ash. 2022. Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 396–410. https://doi.org/10.1145/3514094.3534147
[85] Calvin Liang. 2021. Reflexivity, positionality, and disclosure in HCI. https://medium.com/@caliang/reflexivity-positionality-and-disclosure-in-hci-3d95007e9916
[86] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (Aug. 2022), 669–677. https://doi.org/10.1038/s42256-022-00516-1
[87] Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann E. Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sierman, Dina V. Sokolova, Martina Stockhause, and John Westbrook. 2020. The TRUST Principles for digital repositories. *Scientific Data* 7 (2020). https://doi.org/10.1038/s41597-020-0486-7
[88] Joe Logan, Paul J. Kennedy, and Daniel Catchpoole. 2023. A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Scientific Data* 10, 1 (Sept. 2023), 595. https://doi.org/10.1038/s41597-023-02430-6
[89] A. L. D. Loureiro, V. L. Miguéis, and Lucas F. M. da Silva. 2018. Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems* 114 (Oct. 2018), 81–93. https://doi.org/10.1016/j.dss.2018.08.010
[90] Lydia R. Lucchesi, Petra M. Kuhnert, Jenny L. Davis, and Lexing Xie. 2022. Smallset Timelines: A Visual Representation of Data Preprocessing Decisions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1136–1153. https://doi.org/10.1145/3531146.3533175
[91] Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, Ehsan Adeli, and Li Fei-Fei. 2022. MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing. *Advances in Neural Information Processing Systems*.
[92] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445
[93] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. 2022. Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery. *Advances in Neural Information Processing Systems*.
[94] Matthew Stewart. 2023. The Olympics of AI: Benchmarking Machine Learning Systems. https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b
[95] Mantas Mazeika, Eric Tang, Andy Zou, Steven Basart, Jun Shern Chan, Dawn Song, David Forsyth, Jacob Steinhardt, and Dan Hendrycks. 2022. How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios. *Advances in Neural Information Processing Systems*.
[96] Kenneth O. McGraw and S. P. Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1, 1 (1996), 30–46. https://doi.org/10.1037/1082-989X.1.1.30
[97] Alison McIntyre. 2023. Doctrine of Double Effect. In *The Stanford Encyclopedia of Philosophy* (winter 2023 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2023/entries/double-effect/
[98] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 1–35. https://doi.org/10.1145/3457607
[99] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (Jan. 2022), 1–14. https://doi.org/10.1145/3492853
[100] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics* 19, 6 (Nov. 2018), 1236–1246. https://doi.org/10.1093/bib/bbx044
[101] 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*. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596
[102] Reagan Moore. 2008. Towards a Theory of Digital Preservation. *International Journal of Digital Curation* 3, 1 (Aug. 2008), 63–75. https://doi.org/10.2218/ijdc.v3i1.42
[103] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. *Advances in Neural Information Processing Systems*.
[104] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356
[105] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644
[106] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445402
[107] Daniel Noonan and Tamar Chute. 2014. Data Curation and the University Archives. *The American Archivist* 77, 1 (2014), 201–240. https://www.jstor.org/stable/43489591
[108] Natasha Noy and Carole Goble. 2023. Are We Cobblers without Shoes?: Making Computer Science Data FAIR. *Commun. ACM* 66, 1 (Jan. 2023), 36–38. https://doi.org/10.1145/3528574
[109] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. 2023. Challenges in Deploying Machine Learning: A Survey of Case Studies. *Comput. Surveys* 55, 6 (July 2023), 1–29. https://doi.org/10.1145/3533378
[110] Carole L Palmer, Nicholas M Weber, Trevor Muñoz, and Allen H Renear. 2013. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. (2013).
[111] Praveen Paritosh. 2018. The Missing Science of Knowledge Curation: Improving Incentives for Large-scale Knowledge Curation. In *Companion of WWW '18*. ACM Press, Lyon, France, 1105–1106. https://doi.org/10.1145/3184558.3191551
[112] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567
[113] 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 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336
[114] Evan D. Peet, Brian G. Vegetabile, Matthew Cefalu, Joseph D. Pane, and Cheryl L. Damberg. 2022. *Machine Learning in Public Policy: The Perils and the Promise of Interpretability*. Technical Report. RAND Corporation. https://www.jstor.org/stable/resrep44898
[115] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. *Advances in Neural Information Processing Systems*.
[116] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. *Advances in Neural Information Processing Systems* (2021).

[117] Maureen Pennock. 2007. Digital curation: a life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal* 1 (2007).
[118] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3147–3156. https://doi.org/10.1145/2702123.2702298
[119] Alex H. Poole. 2015. How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science* 15, 2 (June 2015), 101–139. https://doi.org/10.1007/s10502-014-9236-y
[120] Bogdana Rakova and Roel Dobbe. 2023. Algorithms as Social-Ecological-Technological Systems: an Environmental Justice Lens on Algorithmic Audits. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. 491–491. https://doi.org/10.1145/3593013.3594014
[121] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. 2022. Tackling Climate Change with Machine Learning. *Comput. Surveys* 55, 2 (Feb. 2022), 42:1–42:96. https://doi.org/10.1145/3485128
[122] Nataniel Ruiz. 2019. Learning to Simulate. https://towardsdatascience.com/learning-to-simulate-c53d8b393a56
[123] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *Advances in Neural Information Processing Systems*.
[124] 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*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518
[125] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–37. https://doi.org/10.1145/3476058
[126] Tal Schuster, Ashwin Kalyan, Alex Polozov, and Adam Tauman Kalai. 2021. Programming Puzzles. *Advances in Neural Information Processing Systems*.
[127] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598
[128] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 13158–13173. https://proceedings.neurips.cc/paper_files/paper/2022/hash/552ef803bef9368c29e53c167de34b55-Abstract-Datasets_and_Benchmarks.html
[129] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 56–67. https://doi.org/10.1145/3351095.3372870
[130] Megan Stanley, John Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. 2021. FS-Mol: A Few-Shot Learning Dataset of Molecules. Advances in Neural Information Processing Systems.
[131] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3465416.3483305
[132] Paige L. Sweet. 2020. Who Knows? Reflexivity in Feminist Standpoint Theory and Bourdieu. *Gender & Society* 34, 6 (Dec. 2020), 922–950. https://doi.org/10.1177/0891243220966600
[133] Andrea K. Thomer, Dharma Akmon, Jeremy J. York, Allison R. B. Tyler, Faye Polasek, Sara Lafia, Libby Hemphill, and Elizabeth Yakel. 2022. The Craft and Coordination of Data Curation: Complicating Workflow Views of Data Science. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 414:1–414:29. https://doi.org/10.1145/3555139
[134] Nanna Bonde Thylstrup. 2022. The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains. *Media, Culture & Society* 44, 4 (May 2022), 655–671. https://doi.org/10.1177/01634437211060226 Publisher: SAGE Publications Ltd.
[135] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. 2021. The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. *Ethics and Information Technology* 23, 3 (Sept. 2021), 419–433. https://doi.org/10.1007/s10676-021-09583-1
[136] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *Advances in Neural Information Processing Systems*.
[137] 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. http://www.jstor.org/stable/40398176
[138] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18
[139] Xuhai Xu, Han Zhang, Yasaman S. Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Scott Kuehn, Mike A. Merrill, Paula S. Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve A. Riskin, Jennifer Mankoff, and Anind Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. *Advances in Neural Information Processing Systems*.
[140] Elizabeth Yakel. 2007. Digital curation. *OCLC Systems & Services: International digital library perspectives* 23, 4 (Jan. 2007), 335–340. https://doi.org/10.1108/10650750710831466
[141] Zhenpeng Yao, Yanwei Lum, Andrew Johnston, Luis Martin Mejia-Mendoza, Xin Zhou, Yonggang Wen, Alán Aspuru-Guzik, Edward H. Sargent, and Zhi Wei Seh. 2023. Machine learning for a sustainable energy future. *Nature Reviews Materials* 8, 3 (March 2023), 202–215. https://doi.org/10.1038/s41578-022-00490-5
[142] Serena Yeung and Joaquin Vanschoren. 2021. Announcing the NeurIPS 2021 Datasets and Benchmarks Track. https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c
[143] Zenodo - Research. Shared. [n. d.]. FAIR Principles. https://about.zenodo.org/principles/