Benchmark Data Repositories for Better Benchmarking

Rachel Longjohn[1*], **Markelle Kelly**[2*], **Sameer Singh**[2], **Padhraic Smyth**[2]
Department of Statistics[1], Department of Computer Science[2]
University of California, Irvine
Irvine, CA 92697
{rlongjoh,kmarke,sameer,pjsmyth}@uci.edu

Abstract

In machine learning research, it is common to evaluate algorithms via their performance on standard benchmark datasets. While a growing body of work establishes guidelines for—and levies criticisms at—data and benchmarking practices in machine learning, comparatively less attention has been paid to the data repositories where these datasets are stored, documented, and shared. In this paper, we analyze the landscape of these *benchmark data repositories* and the role they can play in improving benchmarking. This role includes addressing issues with both datasets themselves (e.g., representational harms, construct validity) and the manner in which evaluation is carried out using such datasets (e.g., overemphasis on a few datasets and metrics, lack of reproducibility). To this end, we identify and discuss a set of considerations surrounding the design and use of benchmark data repositories, with a focus on improving benchmarking practices in machine learning.

1 Introduction

Evaluating machine learning (ML) algorithms on benchmark datasets is a central pillar of ML research. This performance benchmarking facilitates direct comparison across different techniques, which is important, for example, in the publication of research that introduces a novel method or for selecting the most appropriate approach for a particular application [1–5]. Ideally, these benchmark datasets serve as proxies for real-world tasks, so that performing well on the task represents meaningful advancement toward some desired real-world ML capability [6–10]. Benchmarking can help quantify progress on these tasks over time, and the availability of a well-studied, standard task evaluation environment can be a critical first step before moving to real-world applications, especially in high-stakes or expensive domains. In addition, evaluating with a benchmark dataset can be useful as a sanity check when developing a new methodology, as well as for ML education and training [11, 12].

Early data repositories, such as the UCI ML Repository, arose to address the data needs that come with ML benchmarking [13]. These repositories started as relatively small-scale efforts, but as the field of ML has rapidly grown, they have become more sophisticated, supporting additional features such as leaderboards comparing the benchmarked performance of ML models on a given dataset [14–20]. ML data repositories are fundamentally different from traditional domain-specific data repositories. For example, they tend to contain datasets from a wide variety of domains, and the process of selecting a dataset is often less about a scientific or engineering application and more about the compositional characteristics of the data and its associated tasks, for which a particular class of methods is applicable, e.g., multivariate spatiotemporal or network datasets.

[*]Denotes equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-2744

Today, these ML data repositories—in particular, HuggingFace Datasets, Kaggle, OpenML, Papers with Code Datasets, TensorFlow Datasets, and the UCI ML Repository—are widely used. However, relatively little work has been devoted to understanding them and the specific factors involved in their design. In this paper, we introduce the term *benchmark data repository* to describe repositories that support the discovery and use of datasets for evaluating ML models (for brevity, we will use *benchmark repository* in the remainder of the paper). Our focus is the role of benchmark repositories in ML research, particularly the relationship between benchmark repositories and criticisms of current data and benchmarking practices in ML.[2] Each of Sections 2-6 reviews one of these issues, progressing through the dataset lifecycle—from creation and development to documentation and sharing to use and reuse for model evaluation [21–23]. For each criticism, we identify ways in which benchmark repositories can be part of the solution, motivated by existing standards, observed trends in the field, and examples from our experience as repository curators. We believe that these recommendations will be useful both for the owners of benchmark repositories in designing and improving their repositories, and more broadly to the creators and users of benchmark datasets in determining how to store, document, and find data. To the best of our knowledge, this paper is the first to define and establish best practices specifically for benchmark repositories, as well as to connect the practices of benchmark repositories to ML and data repository practices in general.

2 Valuing Datasets as Research Contributions

Data work maintains a legacy of being under-valued and under-incentivized by the ML community [6, 24–32], often regarded as an "engineering exercise" [33] or "operational" [26]. Several recent initiatives, such as the NeurIPS Datasets and Benchmarks track[3] and the Journal of Data-centric Machine Learning Research [34], have sought to change this pattern by providing peer-reviewed venues for publishing papers on data contributions. In this section, we posit that benchmark repositories can also help recognize datasets as intellectual contributions to the scholarly ecosystem by providing 1) dataset citations, 2) "connection metadata," and 3) dataset licenses. Reinforcing the value of data work incentivizes dataset creators to pay greater care and attention during dataset development and documentation—effects that propagate throughout the dataset lifecycle [24, 26, 27].

2.1 Dataset Citations and Metrics

For a dataset to operate in the ML ecosystem as a first-class research contribution, researchers must be able to locate it and its metadata via a persistent stable URL (as is the norm with published papers). In particular, the assignment of a *persistent identifier* (PID), such as a DOI, that can reliably be used to access a dataset has been widely recommended by experts [32, 35–39]. However, ML datasets and their documentation frequently lack PIDs and are often only available via GitHub or personal/research group websites [32, 40, 41]. Repositories can help address this by minting DOIs for submitted datasets (e.g., as Kaggle[4] and HuggingFace Datasets[5] do).

PIDs are the foundation of *dataset citations*, which give proper attribution to dataset creators, rather than solely citing associated publications [36, 42–49]. In ML, however, datasets are often referred to using combinations of names, descriptions, and associated papers, which can be challenging to disambiguate [40]. In contrast, many data repositories already provide standardized dataset citations that can be easily copied in a desired format (e.g., BibTeX) and include the minted DOI (Figure 1). Beyond giving credit, citing a dataset enables researchers to track its usage throughout the literature, which is particularly relevant in ML, e.g., for performance comparisons.

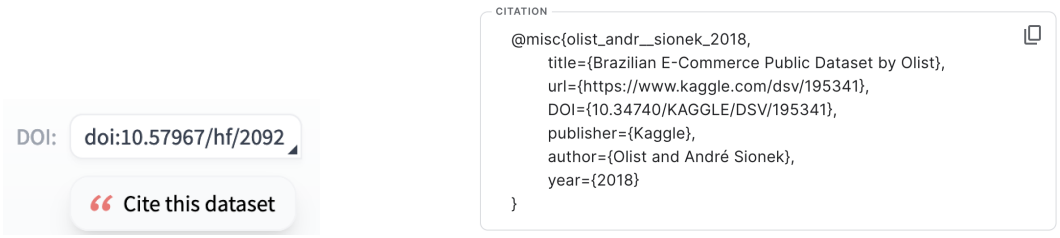
Furthermore, metrics such as the number of citations, number of views, or number of downloads can help quantify data impact, highlighting the value of the dataset in terms of its contribution to the ML community and potentially benefiting a variety of stakeholders (e.g., the researchers whose work is being cited or funders assessing a return on investment [50–53]). Repositories can provide the infrastructure for tracking these metrics of interest; for example, OpenML counts the number of

[2]While in this paper we focus on data used for model evaluation, we note that many of our points are also relevant to pretraining data.

[3]https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c

[4]https://www.kaggle.com/discussions/product-feedback/108594

[5]https://huggingface.co/blog/introducing-doi



(a) HuggingFace [57] (b) Kaggle [58]

Figure 1: Examples of DOIs and citations in repositories.

(a) Papers with Code[6] (b) UCI ML Repository [60]

Figure 2: Examples of connecting datasets to papers in repositories.

times a dataset has been used in experiment runs and the number of times it has been downloaded [14]. Data organizations, such as Scholix, the Data Usage Metric Working Group and Project Counter [54–56], are working towards more sophisticated frameworks for the provision of data metrics, and benchmark repositories are in a prime position to foster collaborations with these efforts.

2.2 Connection Metadata

Repositories can also support the treatment of datasets as research contributions via *connection metadata*, which connects a dataset to associated research entities (such as the dataset's creators or maintainers, publications, code, or other datasets) [59].

The dataset construction process is rife with consequential, "value-laden" decisions [8, 28, 29, 33]. The rationale behind these decisions may be described in an *introductory paper*: a publication that combines the narrative style of an article with the technical description of a dataset and its design process (also referred to as "data articles" [61] or "dataset descriptors"[7]). Introductory papers can give the data a story beyond standardized documentation, providing useful context about the problem, background on data collection procedures, and guidance about tasks for which the data have already been used. When these papers are peer-reviewed, it can lend additional credibility to the dataset for those considering it for re-use. Introductory papers can be included in benchmark repositories as a standardized metadata field (Figure 2), raising the visibility of these important documents.

Repositories can also identify an individual who agrees to serve as a dataset's *point of contact*: someone responsible for answering questions about the dataset and addressing any issues. Ideally, this person is also one of the dataset's creators, as they are best equipped to answer questions about the data and facilitate re-use [41, 62]. Although long-term data maintenance, and determining different stakeholders' responsibilities in that maintenance, remain challenging tasks [28, 63, 64], establishing a point of contact can help prevent the development of a disconnect between a dataset and its creators, which is not uncommon in ML [65–67]. By requiring that contact information for a responsible individual be specified in metadata, repositories can encourage an ongoing connection between the dataset, those who created it, and those who want to re-use it.

2.3 Dataset Licenses

It is widely recommended that datasets come with clear use guidance, often via a *license* [35, 39, 68]. These safeguards can help prevent the unintended use of data, an important part of respecting datasets as intellectual contributions. Data repositories can include licenses as part of a dataset's metadata. They can also make selecting a license easier on dataset donors, e.g., by showing which licenses

[6]https://paperswithcode.com/dataset/imagenet
[7]https://www.nature.com/sdata/journal-information

https://doi.org/10.52202/079017-2744

are popular, providing help text and links to the license language, or comparing the salient parts of different popular licenses. For ML datasets in particular, licensing can be complicated (e.g., it is ambiguous if models trained on a dataset count as "derivative work" [69]), and dataset licenses commonly used in other domains may not effectively restrict data use in ML (e.g., training commercial ML pipelines) [40, 70]. In addition, current licensing practices for ML datasets are often irregular, conflicting, poorly documented, or over-permissive given the dataset content [71]; in one survey of over 1800 text datasets, 69% had "unspecified" licenses on HuggingFace [72]. To help mitigate these issues, benchmark repositories can encourage the use of licenses that were constructed with ML use cases in mind, such as the Montreal Data License [69], or others, as more work is done in this area.

3 Addressing Issues with Dataset Content

Over the past decade or two, numerous issues with common benchmark datasets have been discovered, including technical flaws such as labeling errors and annotation artifacts [5, 73–75], privacy and copyright violations [40, 76–78], inclusions of hate speech or other harmful content [79, 80], representational biases [77, 81–83], and miscellaneous ethical issues [40, 66, 84]. Without clear documentation or careful data auditing, it is easy for these problems to go undiscovered well after a dataset's initial release and propagate harmful effects to downstream results [38, 78]. Further, even once an issue is discovered, updating or deprecating a dataset can be ineffective [40, 75]. Benchmark repositories can help detect and address dataset issues by collecting contextual metadata, performing quality reviews, and supporting the revision and deprecation of datasets.

3.1 Contextual Metadata

Benchmark datasets are often disseminated without detailed information about their broader context [9, 27, 28, 76, 85, 86]. By collecting *contextual metadata*, including information about a dataset's source, funding, collection, annotation, and preprocessing, benchmark repositories can illuminate the assumptions and motivations of dataset creators and flag potential dataset issues [6, 7, 30, 87–89]. To this end, several standards and schemata that include contextual metadata have been established, including Data Cards [90], datasheets for datasets [68], the Dataset Nutrition Label [91, 92], and the FAIR principles [35]. Such metadata can help ML practitioners detect issues earlier [93]; several "retrospective" datasheets for well-known datasets have demonstrated how contextual information raises red flags and could have contributed to earlier detection of data issues [77, 94].

In particular, information about the source of a dataset can alert data users to privacy or consent issues, representation biases, the potential for harmful content, or a mismatch with their target domain [95, 96]. For example, multiple facial image datasets include mugshots or surveillance camera footage [97–99]—raising red flags about the consent and privacy of the photographed individuals. Another example is the NIST Face Recognition Vendor Test dataset, which was funded by the U.S. Department of Homeland Security and contains data from the U.S. Mexican visa archive [100, 101]. In the use of this dataset for general facial recognition evaluation (e.g., [102]), its source and original intent are cause for concern about its transferability [76]. Generally, understanding the origins of a dataset can help ML researchers determine if it is appropriate for their use case, discouraging the use of benchmarks that are poor proxies for the task they are supposedly evaluating [6, 7, 23, 87, 95].

Data selection, filtering, and annotation processes are important design decisions that can significantly impact downstream performance [7, 38, 78, 103, 104]. One example is the systematic exclusion of text authored by or about marginalized groups in large, web-scraped text datasets due to curation and data filtering processes [29, 33, 38, 77]. Another pervasive issue is biased annotations, which are often crowd-sourced [27, 38, 105–108], e.g., as have been documented in the ImageNet dataset [7, 104]. However, these processes tend to be under-documented [29]; for instance, in a survey of over 100 papers introducing computer vision datasets, 36.6% did not provide any description of the human annotators; only 7.8% reported annotator demographics [32].

Clearly documenting data source, intent, collection, and processing procedures sheds light on these dataset issues early on in the data lifecycle. However, dataset creators do not necessarily prioritize metadata on their own [32, 41], and documentation is often scattered and unstandardized [76, 85, 93]. Benchmark repositories can work against this pattern by requiring dataset creators to provide detailed documentation of contextual information (and making it easily accessible), e.g., via an accompanying datasheet [68] and/or published introductory paper, which thoroughly describe a dataset's context.

Housing type	Correct value	Value in repository
Own	3	2
Rent	2	1
Free	1	3

Figure 3: The Statlog (German Credit Data) dataset [109], hosted by the UCI ML Repository, is a sample of customer records from a German bank, with the task of classifying each individual as a good or bad credit risk. In the repository documentation, 8 categorical variables have their levels mixed up or incorrectly described (e.g., see attribute 15, the type of housing the debtor lives in, above). Groemping [67] tracked down papers which describe the dataset's origins [110–113] to construct a proper code table. She donated the corrected dataset as the South German Credit dataset in 2019 [67] but the original dataset from 1994 has nonetheless been widely used in ML research.

3.2 Quality Review

Ideally, quality issues with benchmark datasets and their metadata are detected early and corrected; otherwise, these concerns should either be documented or used as a rationale to withdraw the dataset. For example, datasets containing personally identifiable information should not be released [27] and documentation errors (as in Figure 3) should be quickly amended. Benchmark repositories can help identify these problems throughout the data lifecycle by (1) performing a pre-release *quality review* [95] to catch issues before a dataset is shared, and (2) by serving as a centralized location to collect users' reports and concerns [75] to flag issues throughout a dataset's use and reuse. With stringent quality assurance, ML researchers can reliably look to a repository for high-quality datasets [38], making it easier to avoid using unvalidated, problematic datasets for benchmarking.

Quality reviews can help counter the current lack of incentive for ML dataset creators to consider ethical issues, which has been pointed to as a major contributor to the numerous ethical problems with benchmark datasets [27]. For example, benchmark data collection often does not undergo institutional ethical review [78]; in one survey [32], only 5 out of 100 papers introducing datasets with human subjects mentioned an institutional review board (IRB) or equivalent ethical review. As a result there has been a call for more intervention in data curation, involving curators who can focus on developing conduct codes and ethical review processes rather than relying on dataset creators [27, 30, 33, 88]. It is an open question to what extent repositories should be involved in these decisions; several popular repositories (e.g., Zenodo, Mendeley) view their role as only providing infrastructure and not conducting any kind of data review. However, we posit that benchmark repositories are well-positioned in the data pipeline to perform at least basic ethical checks and initiate a movement towards interventionism. We point to the growing body of literature on ethical data curation for ML [30, 78, 89, 95, 106, 114] as a starting point for the development of ethical review processes.

Conducting thorough quality assurance can be particularly difficult for benchmark repositories because they typically host data from a variety of domains. In contrast, disciplinary repositories, which specialize in a particular domain, often have a community of experts with the knowledge to conduct quality reviews. We point to requiring peer-reviewed introductory papers as a potential step in this direction, as the publishing venue may be able to perform more targeted reviews, and an increasing number of venues also incorporate ethical reviews.[8] Repositories could also outsource reviews for datasets via a network of experts such as the Data Curation Network.[9]

3.3 Dataset Revision and Deprecation

Although quality review can help catch serious issues before the release of a dataset, inevitably, some datasets will need to be updated, corrected, or deprecated. As a centralized data source, benchmark repositories can help support the revision and deprecation of datasets.

Benchmark repositories can support *dataset revision* by documenting data versions and connecting each dataset to a responsible point of contact. When different versions of a dataset are not clearly associated with unique version numbers, differing versions may be used interchangeably [117, 118] (e.g., as in Figure 4). Repositories can enforce versioning by assigning a new version number whenever a data file is changed. Documentation of the revision, including what was changed or

[8] e.g., https://medium.com/@icml2024pc/ethics-review-at-icml-e3b4ce1afd54,https://neurips.cc/public/EthicsGuidelines
[9] https://datacurationnetwork.org

 https://doi.org/10.52202/079017-2744

Index	Sepal length	Sepal width	Petal length	Petal width
35	4.9	3.1	1.5	0.2
38	4.9	3.6	1.4	0.1
90	5.5	2.5	4	1.3

(a) The original data for the 35th, 38th, and 90th iris flowers.

Index	Sepal length	Sepal width	Petal length	Petal width
35	4.9	3.1	1.5	0.2
38	4.9	3.6	1.4	0.1
90	5.5	2.5	5	1.3

(b) An alternative version of the data with the petal length of the 90th flower incorrect.

Index	Sepal length	Sepal width	Petal length	Petal width
35	4.9	3.1	1.5	0.1
38	4.9	3.1	1.5	0.1
90	5.5	2.5	4	1.3

(c) An alternative version of the data with erroneous entries for the 35th and 38th flowers.

Figure 4: The Iris dataset from the UCI ML Repository is widely used for evaluating clustering and classification algorithms [115]. Each observation corresponds to an iris flower, including sepal and petal measurements and its specific species (out of three classes). After years of use, it was discovered that there were multiple different widely-publicized versions of this dataset, with differing measurements for certain observations. Consequently, the reported performances of classification models on Iris (across a large number of published papers) are not necessarily comparable [116].

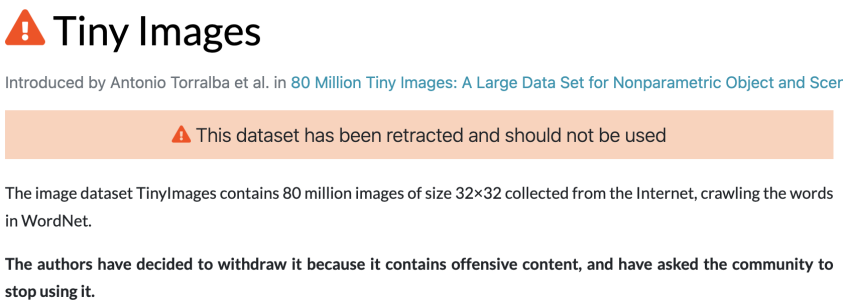


Figure 5: The Papers with Code dataset page for the deprecated Tiny Images dataset.

removed and a rationale for the changes, should also be provided [32, 95]. In addition, by associating datasets with a responsible point of contact (see Section 2.2), repositories can help streamline the resolution of questions or issues regarding a dataset.

Repositories can also support the *deprecation* of datasets. Currently, there is no standardized process for dataset deprecation: creators often withdraw their dataset without an explanation of why it was withdrawn or explicit instructions not to use the dataset (or other post-deprecation protocols). For example, in [75]'s case study of six high-profile dataset retractions, three (MS-Celeb-1M, Duke MTMC, and HRT Transgender) did not provide any reason for the dataset's removal. Moreover, deprecation reports are posted in a scattered, decentralized manner via news articles, conference papers, or researcher or lab websites [75]. Ultimately, it can be unclear to researchers if a dataset is acceptable to use; it is not uncommon for datasets to remain in use after their deprecation, including in published, peer-reviewed papers [7, 40, 75]. To mitigate this, benchmark repositories can (1) establish a process for deprecating a dataset, in which its creators submit a standardized report, detailing the reasons for deprecation and post-deprecation protocols, and (2) maintain a page connected to the dataset DOI (see Section 2.1) with the deprecation report and original metadata [119]. If a deprecation report is clearly displayed in the same place where a dataset was available, it clarifies to researchers (and reviewers) that the dataset should not be used (e.g., see Figure 5).

4 Promoting Data Usability and Reproducibility

Recent work has pointed to a need for improved (re)usability of data [28, 122] and reproducibility of benchmarked results [123–127] in ML. When a dataset lacks clear metadata, it can lead to critical misunderstandings in the reuse phase of its lifecycle (as in Figure 6). Unambiguous metadata are also a necessary foundation for benchmark reproducibility, ensuring that data are used in the same way across evaluations. Benchmark reproducibility is critical for ML research: it enables the verification of published results, provides a starting point for experimentation and follow-up work, and makes contributions easier for others to use, potentially increasing research impact [5, 128, 129]. Benchmark repositories can support usability and reproducibility with the metadata they require and provide to

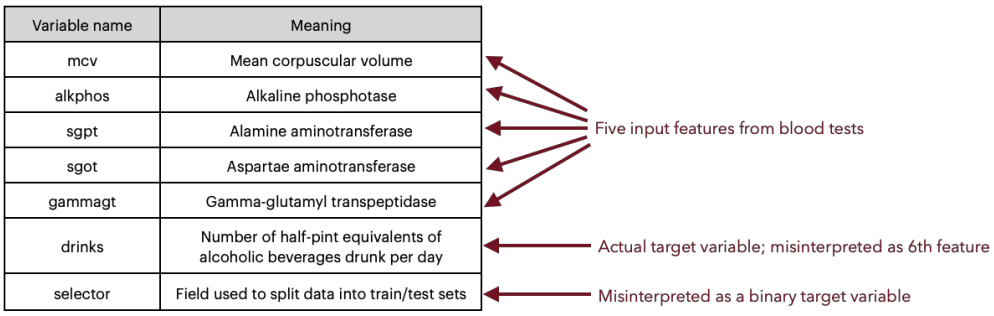


Figure 6: The BUPA Liver Disorders dataset is a popular classification benchmark from the UCI ML Repository [120]. Each row contains information on an individual's consumption of alcoholic drinks and their results on several blood tests targeting alcohol-related liver issues; the intended task is to predict alcohol consumption based on these test results. The last column of the dataset is an indicator, added by the dataset creators, intended to split the rows into training and test sets; however, the data documentation did not clearly explain the meaning of each column. It was subsequently found that many highly-cited papers using this dataset had mistakenly treated this last column as the class label, producing "meaningless results" [121].

users [130, 131], particularly compositional and task-specific metadata (in addition to responsible points of contact and dataset versions—see Sections 2.2 and 3.3 [123]).

If repositories include benchmarked results alongside datasets (discussed further in Section 5), they can further support reproducibility by collecting and providing metadata about the benchmarked results themselves [123].

4.1 Compositional and Task Metadata

Datasets are more usable for ML researchers when accompanied by metadata describing the dataset composition and relevant tasks [35, 39, 132–134]. These metadata help expedite onboarding for a new dataset [41, 135], prevent misunderstandings, and promote data reproducibility [7, 136–139].

Compositional metadata describe the makeup of a dataset, e.g., for tabular datasets, what each instance represents, descriptions of each feature, the total number of rows and columns, the dataset label or target, the presence of missing data, and recommended data splits [68, 140]. When this information is clear, datasets are more interoperable, meaning they can be more easily processed and incorporated into different workflows [35, 135]. For example, a researcher might want to evaluate an existing model on a new dataset; if this dataset has detailed compositional metadata, the researcher can quickly and easily determine which columns they need to use and what preprocessing is required. If this metadata follows a standardized schema (e.g., Croissant [122]), model evaluation may even be done automatically or semi-automatically.

Task metadata include the intended or appropriate ML tasks for a dataset (e.g., image classification or time-series prediction) and specialized metadata relevant to those tasks (e.g., for human [78] or medical [136] image, NLP [141], or ecological [139] data). As an example, HuggingFace Datasets [17], which specializes in NLP data, collects metadata on language and multilinguality, text creation, and fine-grained NLP tasks (e.g., sentiment classification, multiple-choice question answering, or word sense disambiguation). Such specialized metadata make it easy for ML practitioners to find and use data that fit a specific application.

Thus, to improve data usability and benchmark reproducibility, repositories can require that data donors provide high-quality compositional and task metadata, including specialized task metadata, where appropriate [132]. Repositories can also help streamline metadata creation processes, which donors can find overwhelming or time-consuming [41, 142], e.g., with adaptive metadata collection, auto-filling or updating basic fields [41], or supplementary training and tools [143]. We note that to some extent, the responsibility to provide accurate and complete metadata ultimately falls on the data donor. However, repositories can reduce the risk of incorrect, incomplete, or manipulated documentation by enforcing metadata schemata, using quality review processes, and providing user-friendly metadata creation tools—taking some of this burden from the donors themselves.

https://doi.org/10.52202/079017-2744

4.2 Benchmark Metadata

In ML benchmarking, implementation details such as software dependencies, random seeds, and hyperparameter values can have a significant impact on results [5, 124, 144], as can the details of metric computation, including data splits, metric definitions, and aggregation of results [5, 15]. Thus, clearly documenting this *benchmark metadata* is critical for reproducibility [123, 145]. Beyond validating results, the ability to replicate an analysis facilitates "hands-on" experimentation with benchmarked models on a given dataset, enabling researchers to test potential modifications, perform additional evaluations, or debug other models [128, 146]. To this end, if a repository displays a particular benchmarked result for a dataset, they should ensure that specific details on all settings and hyperparameter values used to obtain that result are available [123]. The software environments, dependencies, code, and data files necessary to re-run analyses should also be documented and accessible [41, 117, 128, 147–149]. By ensuring that detailed metadata on a dataset's content, composition, task, and benchmarked results are available, repositories can provide ML practitioners with a holistic understanding of the benchmark [41, 150].

5 Encouraging Holistic Evaluation

It has become common practice to tabulate benchmarked metrics for different ML methods with a dataset leaderboard; several benchmark repositories offer leaderboard features, including Kaggle [18] and Papers with Code.[10] Leaderboarding has become predominant in ML evaluation, and state-of-the-art performance is a key factor in peer review processes [5, 151]. However, this leaderboard culture has been criticized for a "near singular" focus on the incremental improvement of a narrow set of metrics (e.g., classification accuracy) [7, 152]. Such fixation on a specific metric is unlikely to yield broadly applicable results and can stifle the growth of new, diverse ideas [5–7, 151]. Further, as measures of uncertainty are seldom incorporated, seemingly record-breaking performance improvements are not always statistically significant [151, 153, 154] (e.g., a review of the MS MARCO leaderboard found no significant difference in performance between the top three models [155]). In this paper, we refrain from taking a stance on whether benchmark repositories should include leaderboards. However, as several repositories currently act as centralized purveyors of leaderboards, we briefly discuss how they can promote more comprehensive evaluation, helping address problems with benchmarking that manifest later in the dataset lifecycle.

5.1 Analysis Beyond Single Metrics

Recent work on best practices for ML benchmarking recommends evaluating performance more holistically [6, 7, 156, 157]. This could include a variety of metrics, capturing model size and complexity, energy consumption, inference latency, and the amount of data used [1, 3, 5, 152, 158, 159]. Additional in-depth assessment—such as error analysis or disaggregated evaluations—can provide a more nuanced portrait of model behavior, capturing bias, fairness, or robustness [6–8, 106, 160]. Thus, to enhance their leaderboards, benchmark repositories can include this sort of comprehensive information on model performance. In addition to incentivizing progress in a number of dimensions, this approach reflects that the ideal model is context-dependent, enabling practitioners to choose a model based on the criteria most relevant to their use case [5, 7, 8, 158].

5.2 Metric Uncertainty

Metrics shown without any measure of uncertainty can prompt fallacious conclusions, e.g., that one model performs definitively better than another. Instead, including uncertainty makes these benchmarked results more informative [1, 161], and a growing body of methodologies has developed for estimating uncertainty, computing confidence intervals, and performing statistical significance testing in the context of model comparison [5, 144, 155, 162–164]. In light of this, several guidelines for ML evaluation call for the inclusion of variance, uncertainty, and statistical significance in model analysis [87, 95, 106, 151, 165]. Repositories with leaderboards can support this movement by enforcing the reporting of uncertainty alongside point estimates of metrics.

[10] e.g., https://paperswithcode.com/sota/image-classification-on-cifar-10

6 Diversifying Benchmark Datasets

Benchmarking for a particular type of ML model is often concentrated on a limited set of datasets and tasks [6, 8, 23]. This lack of diversity can encourage overfitting to a specific benchmark dataset (e.g., via random seed or hyperparameter fishing) [165, 166]. Over-adaptation also happens at a macro level over time, as new models leverage tricks and strategies from earlier work [5, 87]. Moreover, these benchmarks are often not directly relevant to the real-world behavior they are evaluating—for example, the GLUE benchmark [167] has been commonly used to evaluate natural language understanding, but it is mainly comprised of sequence matching tasks [5]. As a result, there is often a disconnect between benchmarked performance and real-world model behavior [10, 87, 168]. Thus, the overrepresentation of specific tasks, data types, and test datasets can ultimately bias long-term research directions [5, 6, 8] and limit the generalizability of model evaluations [9, 95, 106, 169]. To fight these patterns of overfitting and overuse, repositories can support the discovery and use of diverse, relevant, and continuously evolving datasets.

6.1 Living Datasets

To hinder the overfitting of models to a specific test set, leaderboards can evaluate submitted models on a private, hitherto unused test set [170–172] (e.g., as done by Kaggle) or on out-of-distribution data [151, 173]. Extending this principle, leaderboards can also support "living" or evolving datasets, to which dataset creators continuously add new examples or tasks and remove outdated or erroneous examples. While this means that the benchmarked performances of two models evaluated at two different points in time may not be directly comparable (and data versioning, as discussed in Section 3.3, is critical), robust models will generally outperform those using a specific trick or artifact as evaluations are repeated over time [5]. These living datasets also track the real-world evolution of data—for instance, in the context of autonomous driving, new types of vehicles appear on the roads [38]—which static benchmark datasets fail to capture [75]. By evaluating models on living datasets, repositories can help shift focus away from a specific static set of examples, de-incentivizing overfitting and helping bridge the gap between benchmarked and real-world performance.

6.2 Dataset Discoverability

When choosing benchmark datasets for evaluating an algorithm, it has become the default to select the same datasets already used in the literature [9, 76]. Often, however, there also exists a plethora of other high-quality datasets that could have been used but did not win the "benchmark lottery" [5] and were left undiscovered. To support *dataset discoverability*, existing standards emphasize the importance of standardized, rich metadata [39, 132, 133], which enable searching for datasets via keywords, filtering, and controlled vocabularies [174].

For benchmark repositories, this search is often task-driven: ML practitioners need to find datasets for which a certain type of model is applicable, based on compositional properties and relevant tasks (see Section 4.1). Thus, to improve benchmark dataset discovery, repositories can support search based on compositional and task metadata [33, 41, 130]—for example, the UCI ML Repository's search functionality includes a filter for classification, regression, clustering, or other datasets. Overall, by promoting the discovery and use of a more diverse set of evaluation datasets, repositories can build a barrier to the over-representation of specific benchmark tasks or datasets and encourage more generalizable model evaluation.

7 Discussion and Conclusion

7.1 Key Takeaways

A common thread throughout the criticisms of ML data and benchmarking practices we discuss in this paper is a need for the intervention of a third party—separate from dataset creators and users—in addressing these issues [5–7, 26–28, 31–33, 41, 75, 78, 95, 132]. While improving the state of ML evaluation will be a community effort, involving the efforts of conferences and journals, policymakers, nonprofit organizations, and individual practitioners [27], in this paper we posit that benchmark repositories can play a major role in this effort, instigating far-reaching changes to the culture surrounding datasets and benchmarking in ML. We summarize our key takeaways as follows.

 https://doi.org/10.52202/079017-2744

- Repositories can highlight the status of datasets as valuable scholarly contributions.
- Repositories are well-positioned in the data pipeline to address issues with dataset content.
- Repositories can facilitate data reuse and benchmark reproducibility by ensuring salient metadata is provided for datasets (and, if applicable, benchmark evaluations).
- Repositories with leaderboard features can enforce best practices for model evaluation.
- Repositories can provide a platform for discovering new, relevant, high-quality datasets, counteracting the overuse of a small set of standard benchmark datasets.

7.2 Limitations

The long-term feasibility and impact of our suggestions are predicated upon larger shifts in community norms and attitudes about data-centric work, which will rely upon proper incentivization, an open challenge in the ML community [24, 27, 28]. We discuss here potential incentives for both individual researchers and repositories—although incentives for other actors (e.g., universities, companies, publishers) are also worth exploring.

A key incentive for researchers to become involved in repository efforts is funding; this is becoming more available as agencies such as the U.S. National Institutes of Health (NIH) and National Science Foundation (NSF) pay increasing attention to the data-sharing ecosystem.[11] For example, the NSF's program for Community Infrastructure for Research in Computer and Information Science and Engineering explicitly calls out funding support for data repositories,[12] and the U.S. National Artificial Intelligence Research Resource Task Force identifies repositories as important to their goal of "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem" [135].

To incentivize dataset reviewers, repositories could follow the model of "volunteer journals" such as the *Journal of Machine Learning Research*. These public journals demonstrate how high-quality shared resources can be developed through dedicated volunteer efforts, offering inspiration for a parallel system of oversight, reviewing, and maintenance for repositories. For example, similar to the role of Action or Associate Editors (AEs) in these journals, repositories could have a set of curators who are responsible for identifying relevant experts to review a dataset. By framing data curation and review as an academic service in the same vein as more traditional editorial roles, repositories could help incentivize participation in the review process.

Further work is also needed to determine how to incentivize repositories themselves to enforce best practices (e.g., requiring data donors to select a license or provide task metadata). One potential avenue is to establish standards for benchmark repositories, building upon standards for data repositories in general, such as CoreTrustSeal.[13] Establishing standards or repository certification processes will be most effective if the ML community cultivates an expectation that such requirements are met (e.g., as in archival settings [27]).

Though we may draw useful inspiration from these ideas, it remains unclear what incentivization strategies will work best to spur large-scale buy-in from repositories, dataset creators, and dataset users in implementing and maintaining best practices.

7.3 Looking Ahead

Going forward, we hope that a growing appreciation of data work will permeate the ML community, serving as a catalyst for investment into data infrastructure in ML and broader researcher involvement in data repositories. In light of this, we believe the ideas in this paper lay a foundation for further discussion and research about how benchmark repositories can be utilized, and improved, for better benchmarking in ML.

[11] https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

[12] https://new.nsf.gov/funding/opportunities/circ-community-infrastructure-research-computer-information/nsf23-589/solicitation

[13] https://www.coretrustseal.org/

Acknowledgements

This research was supported in part by the National Science Foundation under award CCRI-1925741 and in part by the Hasso Plattner Institute (HPI) Research Center in Machine Learning and Data Science at the University of California, Irvine.

References

[1] Jeyan Thiyagalingam, Mallikarjun Shankar, Geoffrey Fox, and Tony Hey. Scientific machine learning benchmarks. *Nature Reviews Physics*, 4(6):413–420, April 2022.

[2] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967, 2021.

[3] Frank Hoffmann, Torsten Bertram, Ralf Mikut, Markus Reischl, and Oliver Nelles. Benchmarking in classification and regression. *WIREs Data Mining and Knowledge Discovery*, 9(5): e1318, September 2019.

[4] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 2006 International Conference on Machine Learning*, pages 161–168, 2006.

[5] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

[6] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021. URL `https://openreview.net/pdf?id=j6NxpQbREA1`.

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

[8] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 294, New York, NY, USA, 2020. Association for Computing Machinery. URL `https://doi.org/10.1145/3351095.3373157`.

[9] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*, 2020.

[10] Kiri L. Wagstaff. Machine learning that matters. In *Proceedings of the 2012 International Conference on Machine Learning*, page 1851–1856, Madison, WI, USA, 2012. Omnipress.

[11] Stephanie C. Hicks and Rafael A. Irizarry. A guide to teaching data science. *The American Statistician*, 72(4):382–391, 2018. URL `https://doi.org/10.1080/00031305.2017.1356747`.

[12] Emilio Serrano, Martin Molina, Daniel Manrique, and Luis Baumela. Experiential learning in data science: From the dataset repository to the platform of experiences. *Intelligent Environments 2017*, 22:122–130, 2017.

[13] Pat Langley. The changing science of machine learning. *Machine Learning*, 82(3):275–279, 2011.

[14] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

 https://doi.org/10.52202/079017-2744

[15] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G Mantovani, Jan N van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021): Track on Datasets and Benchmarks*, 2021.

[16] Pavel Brazdil, Jan N van Rijn, Carlos Soares, and Joaquin Vanschoren. Metadata repositories. In *Metalearning*, chapter 16, pages 297–310. Springer, 2022.

[17] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

[18] Konrad Banachewicz and Luca Massaron. *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd, 2022.

[19] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

[20] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10:1–13, 2017.

[21] Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys*, 54(5), May 2021. URL https://doi.org/10.1145/3453444.

[22] Jaihyun Park and Ryan Cordell. The ripple effect of dataset reuse: Contextualising the data lifecycle for machine learning data sets and social impact. *Journal of Information Science*, 2023. URL https://doi.org/10.1177/01655515231212977.

[23] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021): Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=zNQBIBKJRkd.

[24] Katy Ilonka Gero, Payel Das, Pierre Dognin, Inkit Padhi, Prasanna Sattigeri, and Kush R Varshney. The incentive gap in data work in the era of large models. *Nature Machine Intelligence*, 5(6):565–567, 2023.

[25] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 173–184, 2022.

[26] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "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*, pages 1–15, 2021.

[27] Eun Seo Jo and Timnit Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 306–316, New York, NY, USA, January 2020. Association for Computing Machinery. URL https://dl.acm.org/doi/10.1145/3351095.3372829.

[28] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 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)*, page 560–575, New York, NY, USA, 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3442188.3445918.

[29] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.179. URL https://aclanthology.org/2024.naacl-long.179.

[30] Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, FAccT '24, New York, NY, USA, 2024. Association for Computing Machinery.

[31] Benjamin Heinzerling. NLP's clever hans moment has arrived. *Journal of Cognitive Science*, 21(1), 2020.

[32] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? Disciplinary values in computer vision dataset development. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5, pages 1–37. ACM Press, New York, NY, USA, 2021.

[33] Meera A. Desai, Irene V. Pasquetto, Abigail Z. Jacobs, and Dallas Card. An archival perspective on pretraining data. *Patterns*, 5(4):100966, 2024. URL https://doi.org/10.1016/j.patter.2024.100966.

[34] Luis Oala, Manil Maskey, Lilith Bat-Leah, Alicia Parrish, Nezihe Merve Gürel, Tzu-Sheng Kuo, Yang Liu, Rotem Dror, Danilo Brajovic, Xiaozhe Yao, Max Bartolo, William A Gaviria Rojas, Ryan Hileman, Rainier Aliment, Michael W. Mahoney, Meg Risdal, Matthew Lease, Wojciech Samek, Debojyoti Dutta, Curtis G Northcutt, Cody Coleman, Braden Hancock, Bernard Koch, Girmaw Abebe Tadesse, Bojan Karlaš, Ahmed Alaa, Adji Bousso Dieng, Natasha Noy, Vijay Janapa Reddi, James Zou, Praveen Paritosh, Mihaela van der Schaar, Kurt Bollacker, Lora Aroyo, Ce Zhang, Joaquin Vanschoren, Isabelle Guyon, and Peter Mattson. DMLR: Data-centric machine learning research - past, present and future. *Journal of Data-centric Machine Learning Research*, 2024. URL https://openreview.net/forum?id=2kpu78QdeE.

[35] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 03 2016.

[36] Victoria Stodden, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John PA Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016.

[37] Alyssa Goodman, Alberto Pepe, Alexander W Blocker, Christine L Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, et al. Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, 10(4): e1003542, 2014.

[38] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8):669–677, Aug 2022.

[39] National Science and Technology Council. Desirable characteristics of data repositories for federally funded research. https://doi.org/10.5479/10088/113528, 2022.

[40] Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021.

[41] Amy K. Heger, Liz B. Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), Nov 2022. URL https://doi.org/10.1145/3555760.

[42] Task Group on Data Citation Standards and CODATA-ICSTI Practices. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, 12(0):CIDCR1–CIDCR75, 2013.

[43] Data Citation Synthesis Group. Joint declaration of data citation principles, 2014. URL https://doi.org/10.25490/a97f-egyk.

[44] Micah Altman, Christine Borgman, Mercè Crosas, and Maryann Matone. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, 41(3):43–45, 2015.

[45] Gianmaria Silvello. Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1):6–20, 2018.

[46] Paul Groth, Helena Cousijn, Tim Clark, and Carole Goble. Fair data reuse–the path through data citation. *Data Intelligence*, 2(1-2):78–86, 2020.

[47] Martin Fenner, Mercè Crosas, Jeffrey S Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone, et al. A data citation roadmap for scholarly data repositories. *Scientific Data*, 6(1):28, 2019.

[48] Joan Starr, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R Downs, Ruth Duerr, Laurel L Haak, Melissa Haendel, Ivan Herman, Simon Hodson, et al. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1:e1, 2015.

[49] Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, et al. A data citation roadmap for scientific publishers. *Scientific Data*, 5(1):1–11, 2018.

[50] Daniella Lowenberg, John Chodacki, Martin Fenner, Jennifer Kemp, and Matthew B. Jones. *Open Data Metrics: Lighting the Fire*. Zenodo, November 2019. URL https://doi.org/10.5281/zenodo.3525349.

[51] Daniella Lowenberg. Recognizing our collective responsibility in the prioritization of open data metrics. *Harvard Data Science Review*, 4(3), 7 2022. URL https://hdsr.mitpress.mit.edu/pub/8vfnfvag.

[52] John E Kratz and Carly Strasser. Making data count. *Scientific Data*, 2(1):1–5, 2015.

[53] Martin Fenner, Daniella Lowenberg, Matt Jones, Paul Needham, Dave Vieglais, Stephen Abrams, Patricia Cruse, and John Chodacki. Code of practice for research data usage metrics release 1. Technical report, PeerJ Preprints, 2018.

[54] Helena Cousijn, Patricia Feeney, Daniella Lowenberg, Eleonora Presani, and Natasha Simons. Bringing citations and usage metrics together to make data count. *Data Science Journal*, 18: 9–9, 2019.

[55] Project Counter. Code of Practice for Research Data, 2018. URL https://www.projectcounter.org/code-of-practice-rd-sections/foreword/.

[56] Adrian Burton, Martin Fenner, Wouter Haak, and Paolo Manghi. Scholix Metadata Schema for Exchange of Scholarly Communication Links, November 2017. URL https://doi.org/10.5281/zenodo.1120265.

[57] HuggingFaceFW. fineweb (revision af075be), 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb.

[58] Olist and André Sionek. Brazilian e-commerce public dataset by olist, 2018. URL https://www.kaggle.com/dsv/195341.

[59] DataCite Metadata Working Group. Making and Using Connection Metadata, 2023. URL https://support.datacite.org/docs/making-and-using-connection-metadata.

[60] Aleksej Logacjov, Atle Kongsvold, Kerstin Bach, Hilde Bremseth Bårdstu, and Paul Jarle Mork. HARTH. UCI Machine Learning Repository. URL https://doi.org/10.24432/C5NC90.

[61] Claire Austin, Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte. Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, 18(2):77–92, 2017. URL https://rdcu.be/cTICG.

[62] Irene V Pasquetto, Christine L Borgman, and Morgan F Wofford. Uses and reuses of scientific data: The data creators' advantage. *Harvard Data Science Review*, 2019.

[63] Christine L. Borgman and Philip Bourne. Why it takes a village to manage and share data. *Harvard Data Science Review*, 4(3), Jul 2022. URL https://hdsr.mitpress.mit.edu/pub/wyxni26q.

[64] Christine L. Borgman. The lives and after lives of data. *Harvard Data Science Review*, 1(1), Jul 2019. URL https://hdsr.mitpress.mit.edu/pub/4giycvvj.

[65] Chhavi Yadav and Léon Bottou. Cold case: The lost MNIST digits. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Joanna Radin. "Digital natives": How medical and indigenous histories matter for big data. *Osiris*, 32(1):43–64, 2017.

[67] Ulrike Groemping. South German credit data: correcting a widely used data set. Reports in mathematics, physics, and chemistry, Beuth Hochschule für Technik Berlin, 2019. URL http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf.

[68] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 11 2021. URL https://doi.org/10.1145/3458723.

[69] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. Towards standardization of data licenses: The Montreal data license. *arXiv preprint arXiv:1903.12262*, 2019.

[70] Gopi Krishnan Rajbahadur, Erika Tuck, Li Zi, Dayi Lin, Boyuan Chen, Zhen Ming, Daniel M German, et al. Can I use this publicly available dataset to build commercial AI software?–A case study on publicly available image datasets. *arXiv preprint arXiv:2111.02374*, 2021.

[71] Moming Duan, Qinbin Li, and Bingsheng He. Modelgo: A practical tool for machine learning license analysis. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1158–1169, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645520. URL https://doi.org/10.1145/3589334.3645520.

[72] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint arXiv:2310.16787*, 2023.

[73] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021): Track on Datasets and Benchmarks*, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

[74] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-2017.

[75] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM Press, New York, NY, USA, 2021.

[76] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. *Proceedings of the AAAI 2020 Workshop on AI Evaluation*, 2020.

[77] John Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning: A retrospective datasheet for BookCorpus. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round1.pdf.

[78] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems*, 36, 2024.

[79] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

[80] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA, 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3351095.3375709.

[81] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[82] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[83] Joon Sung Park, Michael S. Bernstein, Robin N. Brewer, Ece Kamar, and Meredith Ringel Morris. Understanding the representation and representativeness of age in AI data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 834–842, New York, NY, USA, 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3461702.3462590.

[84] M Carlisle. Racist data destruction?, 2019. URL https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8.

[85] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336, 2020.

[86] Xinyu Yang, Weixin Liang, and James Zou. Navigating dataset documentations in AI: A large-scale analysis of dataset cards on Hugging Face. *ICLR*, 2024.

[87] David J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1 – 14, 2006. URL https://doi.org/10.1214/088342306000000060.

[88] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3442188.3445922.

[89] Susan Leavy, Eugenia Siapera, and Barry O'Sullivan. Ethical data curation for AI: An approach based on feminist epistemology and critical theories of race. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 695–703, New York, NY, USA, 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3461702.3462598.

[90] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAcct)*, page 1776–1826, New York, NY, USA, 2022. Association for Computing Machinery. URL https://doi.org/10.1145/3531146.3533231.

[91] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. *Data Protection and Privacy*, 12(12):1, 2020.

[92] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954*, 2022.

[93] Karen L. Boyd. 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. URL https://doi.org/10.1145/3479582.

[94] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

[95] Stella Biderman and Walter J. Scheirer. Pitfalls in machine learning research: Reexamining the development cycle. In Jessica Zosa Forde, Francisco Ruiz, Melanie F. Pradier, and Aaron Schein, editors, *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pages 106–117. PMLR, 12 Dec 2020. URL https://proceedings.mlr.press/v137/biderman20a.html.

[96] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Jul 2021. URL https://doi.org/10.1145/3457607.

[97] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2325–2333, 2016.

[98] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing.

[99] Andrew Founds, Nick Orlans, Whiddon Genevieve, and Craig Watson. NIST special database 32 - multiple encounter dataset II (MEDS-II), Jul 2011. URL https://doi.org/10.6028/NIST.IR.7807.

[100] P Phillips, Patrick Grother, Ross Micheals, D Blackburn, Elham Tabassi, and M Bone. Face recognition vendor test 2002: Evaluation report, 2003-03-01 2003. URL https://doi.org/10.6028/NIST.IR.6965.

[101] Samuel Dooley, George Z Wei, Tom Goldstein, and John Dickerson. Robustness disparities in face detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 38245–38259. Curran Associates, Inc., 2022.

[102] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Dalong Du, Jiwen Lu, and Jie Zhou. WebFace260M: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (2):2627–2644, 2023. URL https://doi.org/10.1109/TPAMI.2022.3169734.

[103] Sandeep Rangineni. An analysis of data quality requirements for machine learning development pipelines frameworks. *International Journal of Computer Trends and Technology*, 71(9):16–27, 2023.

[104] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From ImageNet to image classification: Contextualizing progress on benchmarks. In *Proceedings of the 2020 International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[105] Shilad Sen, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. Turkers, scholars, "arafat" and "peace": Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 826–838, New York, NY, USA, 2015. Association for Computing Machinery. URL https://doi.org/10.1145/2675133.2675285.

[106] Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/2021.naacl-main.385.

[107] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8 (32), 2022. URL https://doi.org/10.1126/sciadv.abq6147.

[108] Patrik Sörqvist, Linda Langeborg, and Mårten Eriksson. Women assimilate across gender, men don't: The role of gender to the own-anchor effect in age, height, and weight estimates. *Journal of Applied Social Psychology*, 41(7):1733–1748, 2011. URL https://doi.org/10.1111/j.1559-1816.2011.00774.x.

[109] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. URL https://doi.org/10.24432/C5NC77.

[110] WM Häußler. Empirische ergebnisse zu diskriminationsverfahren bei kreditscoringsystemen. *Zeitschrift für Operations Research*, 23:B191–B210, 1979.

[111] WM Häußler. Methoden der punktebewertung für kreditscoringsysteme. *Zeitschrift für Operations Research*, 25:B79–B94, 1981.

[112] Ludwig Fahrmeir and Alfred Hamerle. Kategoriale regression in der betrieblichen planung. *Zeitschrift für Operations Research*, 25:B63–B78, 1981.

[113] Ludwig Fahrmeir and Alfred Hamerle. *Multivariate Statistische Verfahren*. de Gruyter, 1984.

[114] Madhulika Srikumar, Rebecca Finlay, Grace Abuhamad, Carolyn Ashurst, Rosie Campbell, Emily Campbell-Ratcliffe, Hudson Hongo, Sara R Jordan, Joseph Lindley, Aviv Ovadya, et al. Advancing ethics review practices in AI research. *Nature Machine Intelligence*, 4(12): 1061–1064, 2022.

[115] R.A. Fisher. Iris. UCI Machine Learning Repository, 1988. URL https://doi.org/10.24432/C56C76.

[116] James C Bezdek, James M Keller, Raghu Krishnapuram, Ludmila I Kuncheva, and Nikhil R Pal. Will the real iris data please stand up? *IEEE Transactions on Fuzzy Systems*, 7(3): 368–369, 1999.

[117] Sandra L Sawchuk and Shahira Khair. Computational reproducibility: A practical framework for data curators. *Journal of eScience Librarianship*, 10(3):7, 2021.

[118] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[119] Helenmary Sheridan, Anthony J Dellureficio, Melissa A Ratajeski, Sara Mannheimer, and Terrie R Wheeler. Data curation through catalogs: A repository-independent model for data discovery. *Journal of eScience Librarianship*, 10(3):4, 2021.

[120] Richard Forsyth. Liver Disorders. UCI Machine Learning Repository, 1990. URL https://doi.org/10.24432/C54G67.

[121] James McDermott and Richard S. Forsyth. Diagnosing a disorder in a classification benchmark. *Pattern Recognition Letters*, 73:41–43, 2016. URL https://doi.org/10.1016/j.patrec.2016.01.004.

[122] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, et al. Croissant: A metadata format for ML-ready datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, pages 1–6, 2024.

[123] Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c429429bf1f2af051f2021dc92a8ebea-Paper.pdf.

[124] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. AAAI Press, Menlo Park, 2018. URL https://doi.org/10.1609/aaai.v32i1.11694.

[125] Gerald Friedland. *Repeatability and Reproducibility*, pages 201–208. Springer International Publishing, Cham, 2024. URL https://doi.org/10.1007/978-3-031-39477-5_15.

[126] Antonio Maffia, Helmar Burkhart, Florina M. Ciorba, and Michael Resch. *On Benchmarking of Deep Learning Systems: Software Engineering Issues and Reproducibility Challenges*. Philosophisch-Naturwissenschaftliche Fakultät der Universität Basel, 2023. URL https://books.google.com/books?id=rRw50AEACAAJ.

[127] Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*, 2022.

[128] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9), Jan 2023. URL https://doi.org/10.1145/3555803.

[129] Rachael Tatman, Jake Vanderplas, and Sohier Dane. A practical taxonomy of reproducibility for machine learning research. In *2nd Reproducibility in Machine Learning Workshop*, 2018.

[130] Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya G. Parameswaran. "It took longer than I was expecting:" Why is dataset search still so hard? Preprint at https://www.madelonhulsebos.com/assets/dataset_search_survey.pdf., 2024.

[131] Joan Giner-Miguelez, Abel Gómez, and Jordi Cabot. A domain-specific language for describing machine learning datasets. *Journal of Computer Languages*, 76:101209, 2023. URL https://doi.org/10.1016/j.cola.2023.101209.

[132] Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, et al. The TRUST principles for digital repositories. *Scientific Data*, 7(1):144, 2020.

[133] NIH. Selecting a data repository, 2022. URL https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository.

[134] CCSDS. Reference model for an open archival information system (OAIS), 2012. URL http://www.oais.info/.

[135] National Artificial Intelligence Research Resource Task Force. Strengthening and democratizing the U.S. artificial intelligence innovation ecosystem: An implementation plan for a national artificial intelligence research resource, 2023. URL https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

[136] Katherine Elfer, Emma Gardecki, Victor Garcia, Amy Ly, and Hytopoulous et al. Reproducible reporting of the collection and evaluation of annotations for artificial intelligence models. *Modern Pathology*, 37(4):100439, 2024. URL https://doi.org/10.1016/j.modpat.2024.100439.

[137] Mateusz Pawlik, Thomas Hütter, Daniel Kocher, Willi Mann, and Nikolaus Augsten. A link is not enough–reproducibility of data. *Datenbank-Spektrum*, 19:107–115, 2019.

[138] Chris Welty, Praveen Paritosh, and Lora Aroyo. Metrology for AI: From benchmarks to instruments. *arXiv preprint arXiv:1911.01875*, 2019.

[139] Ayelet Shavit and Aaron M Ellison. *Stepping in the same river twice: Replication in biological research*. Yale University Press, 2017.

[140] Julia Stoyanovich and Bill Howe. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3), 2019.

[141] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. URL https://doi.org/10.1162/tacl_a_00041.

[142] Zhihong Xu, John Watts, Sarah Bankston, and Laura Sare. Depositing data: A usability study of the Texas data repository. *Journal of eScience Librarianship*, 11(1):6, 2022.

[143] U.S. Bureau of Economic Analysis. Advisory committee on data for evidence building: Year 2 report. https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf, 2022.

[144] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, 2021.

[145] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, Jan 2021. URL https://dl.acm.org/doi/10.5555/3546258.3546422.

[146] Donghyun Kang, TaeYoung Kang, and Junkyu Jang. Papers with code or without code? Impact of GitHub repository usability on the diffusion of machine learning research. *Information Processing & Management*, 60(6):103477, 2023. URL https://doi.org/10.1016/j.ipm.2023.103477.

[147] Odd Erik Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, 379(2197):20200210, 2021.

[148] Anna Laurinavichyute, Himanshu Yadav, and Shravan Vasishth. Share the code, not just the data: A case study of the reproducibility of articles published in the journal of memory and language under the open data policy. *Journal of Memory and Language*, 125:104332, 2022.

[149] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. What makes a well-documented notebook? A case study of data scientists' documentation practices in Kaggle. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. URL https://doi.org/10.1145/3411763.3451617.

[150] Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Karthik Ram, and Jane Greenberg. The role of metadata in reproducible computational research. *Patterns*, 2(9):100322, 2021.

[151] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of Goodhart's law. *Advances in Neural Information Processing Systems*, 33:407–417, 2020.

[152] Rachel L Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5), 2022.

[153] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2015.

[154] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014. URL http://jmlr.org/papers/v15/delgado14a.html.

[155] Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? A case study of the ms marco document ranking leaderboard. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2283–2287, 2021.

[156] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023. URL https://www.science.org/doi/abs/10.1126/science.adf6369.

[157] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[158] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.393.

[159] Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, Feb 2019. URL https://doi.org/10.1145/3317287.3328534.

[160] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1073.

[161] Niklas Tötsch and Daniel Hoffmann. Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer Science*, 7:e398, 2021. URL https://doi.org/10.7717/peerj-cs.398.

[162] Yvette Graham, Nitika Mathur, and Timothy Baldwin. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL https://aclanthology.org/W14-3333.

[163] Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017. URL https://aclanthology.org/Q17-1033.

[164] Shira Wein, Christopher Homan, Lora Aroyo, and Chris Welty. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3138–3161, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.196.

[165] Kajsa Møllersen and Einar Holsbø. What is the state of the art? Accounting for multiplicity in machine learning benchmark performance. *arXiv preprint arXiv:2303.07272*, 2023.

[166] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 2019 International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

[167] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[168] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1334.

[169] David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.85.

[170] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pages 117–126, 2015.

[171] Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 2015 International Conference on Machine Learning*, pages 1006–1014. PMLR, 2015.

[172] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ee39e503b6bedf0c98c388b7e8589aca-Paper.pdf.

[173] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 2021 International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

[174] DataCite Metadata Working Group. Datacite metadata schema documentation for the publication and citation of research data and other research outputs, 2021. URL https://doi.org/10.14454/3w3z-sa82.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/079017-2744