The State of Data Curation at NeurIPS: Appendix

A. Rubric

	CURATORIAL ELEMENT	DESCRIPTION	DOCUMENTATION LEVEL	
			Criteria to meet minimum standard	Criteria to meet standard of excellence
	SCOPE			
1	Context, purpose, motivation	This information explains the purpose of dataset creation for the specified domain.	Documentation discusses the problem domain, what problems the new dataset addresses, the relevance of those problems, and the need for a new dataset in comparison to existing datasets.	Documentation explains how the context of the dataset affects possible reuse and includes reflection on the dataset creators' awareness of social, political, and historical context.
2	Requirements	The translation process from a "real-world" problem to a "ML problem" for which the dataset is created [98, 105] consists of numerous decisions, expertise, and worldviews that should be documented in order to understand the context in which the problem situation was framed.	Documentation states how the problem was formulated and how the dataset creation plan was generated.	Documentation includes reflection on how the problem formulation introduces intrinsic biases.
	ETHICALITY AND REFLEXIVITY			
3	Ethicality	Ethical considerations are critical to the fair and accountable creation and (re)use of datasets.	Documentation discusses how the benefits of creating the dataset outweigh any harms of creating it (see proportionality principle), and it discusses informed consent if the dataset is about humans.	Documentation goes beyond requirements listed in ethics framings like guidelines/policies/checklists. For example, documentation discusses alternate methods of dataset creation that were not used because of potential ethical harm.
4	Domain knowledge & data practices	Creating a dataset involves, often tacit, expertise about one or more domains as well as data practices. Articulating both types of nuance required in dataset development makes data work more transparent [42, 56, 98, 109, 135].	Documentation states the domain-specific expertise and data skills required in developing the dataset.	Documentation discusses the required expertise needed to understand the intended purpose of the dataset and to reuse it.
5	Context awareness	Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.	Documentation includes a positionality statement.	Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.

6	Environmental footprint	This element is for dataset creators to reflect and quantify the footprint of their dataset creation process [6].	Documentation contains a quantitative assessment of environmental footprint and clearly defined scope of what was measured.	Documentation includes a lifecycle assessment and the corresponding environmental footprint, and an assessment of design choices and rationale for the choices.
DATA PIPELINE				
7	Data collection	Disclosing data sources is essential in the data collection process. Further reflection on the process of selecting those sources can reveal important interpretive assumptions [98] and historical and representational biases [56].	If data was collected, documentation states how and why data and metadata were collected from the data source(s). If data was synthesized, documentation discusses: 1) how and why the data was synthesized and 2) whether the data was synthesized to match labels, if used.	If data was collected, documentation discusses the process of defining criteria for selecting data source(s), specifies the criteria, explains why those criteria were chosen, and how the selected data sources are evaluated against these criteria. If data was synthesized, documentation includes a reflection on potential intrinsic biases of the synthesis process, how the synthesis process shaped the features of the data, the limitations of the synthesis process, and how the synthesized data relates to the real-world distribution of the data it represents.
8	Data processing	Data processing involves cleaning, transforming, and wrangling data. Data processing decisions have impacts on the ultimate "cleaned" data that is used [77, 98]. Detailed documentation of this process enables outcomes of the model to be traced back to processing decisions.	Documentation discusses the process of cleaning, transforming, or wrangling data.	Documentation goes beyond what is done to discuss how the decisions about data processing were made and why, and potential impacts of the processing decisions.
9	Data annotation	Data annotation or labelling, regardless of the guidelines provided to reduce worker bias, can lead to disagreements on how data should be annotated (either between annotators or between dataset creators and annotators).The inclusion of this documentation highlights what is considered the "ground truth" [22, 98, 99]	Documentation discusses the process of annotation. If any labels are used, the documentation includes the following: If labels are derived from the data: documentation discusses how data was interpreted to generate labels.	Documentation discusses the process of annotation with depth and reflexivity by including a reflection on how annotations (including labels, if used) represent differing worldviews and social backgrounds.

		by the dataset creators which impacts how annotation is performed [57].	If the labels were created first and the data was derived from the labels: documentation discusses how the relationship of the data to the labels was verified. If labels are obtained from elsewhere: documentation discusses where they were obtained from, how they were reused, and how the collected annotations and labels are combined with existing ones.	Additionally, if labels are derived from the data: documentation discusses how the labels are robust, i.e., not sensitive to variability and how disagreements on annotation were reconciled.
DATA QUALITY				
10	Suitability	Suitability is a measure of a dataset's quality with regards to the purpose defined.	Documentation discusses how the dataset is appropriate for the defined purpose.	Documentation discusses how dimensions such as accuracy, completeness, timeliness, and consistency contribute to the quality of the dataset in being used for the defined purpose. For example, timeliness (i.e., age) of data should be appropriate for the defined purpose.
11	Representativeness	Representativeness is a measure of how well a sample set of data represents the entire population. Sampling procedures and decisions about data sources can introduce extrinsic bias [98]. For example, choosing Reddit or Twitter as a data source can perpetuate dominant social biases rather than being a representative sample of the target population [6].	Documentation defines the population and discusses the extent to which the sampling procedure is representative of the population.	Documentation includes reflection on how the dataset creation process overall, and the sampling procedures specifically, affect extrinsic bias.
12	Authenticity	Authenticity of a dataset is about whether the dataset "is what it purports to be" [23, 25, 26, 46, 110], which is a responsibility of dataset creators [74]. Authenticity can be established by assessing the identity and the integrity of the record [23, 24, 55, 69, 81, 84]. Integrity of a dataset is about whether "the material is complete and unaltered" [7, 13, 27, 46, 95].	Documentation discusses how authenticity has been established and maintained, i.e., • Has the identity and origin of all data been verified? • For data that is obtained, it is clear how the dataset creators have verified the identity of the dataset they reuse.	Documentation states how others can establish the authenticity of this dataset, i.e., • Documentation provides a persistent identifier and provenance information for the dataset in order for reusers to establish identity.

			• For data that is generated, it is clear how they have been created and by whom. • Has the integrity of all data been verified? • For data that is processed in any way, it is clear how processing steps may have impacted integrity.	• Documentation provides mechanisms for reusers to verify the integrity of their dataset.
13	Reliability	Reliability is about how well the dataset is "capable of standing for the facts to which it attests" [23], i.e., how certain we can be that its data points reflect what they represent.	Documentation discusses how the reliability of the dataset has been established and maintained, including the verification steps taken to ensure reliability, where necessary, i.e., • It is clear for each data element what synthetic or real-world phenomenon it represents.	Documentation states how others can establish the reliability of the dataset, i.e., • Documentation provides mechanisms to enable verification of what synthetic or real-world phenomenon each data element represents.
14	Structured documentation	Context documents in standardized structures provide information on the content of the dataset which is critical in establishing its usage in a well defined format.	Documentation includes a standardized context document. Acceptable formats include context documents that follow an established structure such as datasheets, data statements, and nutrition labels.	The context document addresses all mandatory items.
DATA MANAGEMENT				
15	Findability	Ensuring findability is about enabling the dataset to be discovered for reuse after its development [138].	Documentation discusses how the dataset is findable by providing a globally unique and persistent identifier (URLs are not persistent).	Documentation includes metadata and both the metadata and data are stored in a searchable repository.
16	Accessibility	Accessibility is about enabling the dataset to be obtained after its development [138].	Documentation states all information and tools required to access the content of the data, and the identifier navigates to the metadata and data.	Documentation includes a communications protocol, an authentication and authorization procedure, and provides metadata that will be available even if data access is removed.
17	Interoperability	Interoperability ensures that the dataset can be integrated with other applications and workflows [138].	Documentation discusses how the dataset integrates with other data, workflows, applications, etc. (i.e., that both the metadata and data are readable by humans and machines).	Documentation has metadata and data that both use controlled vocabularies and link to other resources using qualified references.

18	Reusability	Ensuring reusability requires providing information such as relevant provenance and usage [138].	For both metadata and data, provenance information includes at least all of the following: 1) where the data came from, 2) who collected it, and 3) when it was collected.	Documentation has metadata and data that are both described using domain-relevant standards, state license and usage information, and provide additional provenance documentation as described by FAIR best practices.

B. Rubric worksheet

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard		Criteria to meet standard of excellence	
		Pass/Fail	Comments	Full/Partial/None	Comments
	SCOPE				
1	Context, purpose, motivation				
2	Requirements				
	ETHICALITY AND REFLEXIVITY				
3	Ethicality				
4	Domain knowledge & data practices				
5	Context awareness				
6	Environmental footprint				
	DATA PIPELINE				
7	Data collection				
8	Data processing				
9	Data annotation				
	DATA QUALITY				
10	Suitability				
11	Representativeness				
12	Authenticity				
13	Reliability				
14	Structured documentation				
	DATA MANAGEMENT				
15	Findability				
16	Accessibility				
17	Interoperability				
18	Reusability				

C. Toolkit

C.1 Overview of research

Background and Motivation: The usage of artificial intelligence has increased exponentially with applications in predicting outcomes related to education, employment, housing, and many more social, economic, and financial aspects of our lives. Archival studies have long dealt with large amounts of data and concerns of representativeness, ethics, integrity, and more with the use of data curation methods, theories, and frameworks. Machine learning research (MLR) has pinpointed the data underlying predictive models to be the largest contributor in introducing bias [107, 117, 121]. Emerging studies have advocated for the prioritization of rigorous data curation practices often referred to as "data work" or "dataset development" in MLR [6, 42, 60]. Introducing data curation concepts and principles can therefore improve the transparency and accountability of the dataset creation process within MLR.

Objectives: We assess ML dataset development processes using principles and methods from archival studies and digital curation. We perform a synthesis and organization of existing work to enable the coherent usage of data curation frameworks, a taxonomy of data curation terms used within machine learning research, and a review of gaps and opportunities for data curation in machine learning.

Method: Our research design for this study consists of the following:

1. Synthesizing literature on data curation concepts and principles central to ML data work.
2. Exploring the relevance of data curation concepts and principles through an illustration of how they can be adapted, translated, and operationalized for ML data work.
3. Demonstrating the gaps and overlaps in how ML data practices already perform data curation, how data curation is discussed in MLR, and how data curation can be further adopted.

Goals and contributions: This project deepens the scholarly and practical connections between the data curation and machine learning research communities and initiate directions for improvement within MLR's data practices. The outcomes present a novel perspective on improving documentation practices in machine learning through data curation. Through this project, we aim to further establish the connections between the data curation and machine learning research communities.

C.2 Application guidance

Scope of application: The rubric is intended for two types of users.

1. Firstly, dataset creators can use the rubric as a resource to prompt and facilitate critical engagement and reflection throughout their dataset creation process.
2. Secondly, existing datasets can be evaluated prior to publishing or reuse by applying the rubric to determine gaps that require further documentation and areas where bias can be introduced. In both cases, we aim for the rubric to be a practical and useful resource for researchers to engage with the dataset creation process using a data curation lens. The rubric was developed for the evaluation of ML datasets and has elements specific to the domain, including: requirements, data annotation, environmental footprint, and structured documentation.

C.2.1 Applying the rubric to your own dataset

The overall process for using the rubric is as follows:

1. Read the rubric to get familiarized with the elements and details that will be needed.
2. Review each element in the rubric individually.
    a. For each element, first assess whether the minimum standard of documentation has been fulfilled. To do this, provide a pass/fail evaluation, where a pass is granted if all aspects specified under the minimum standard were discussed and a fail if they were only partially discussed or not discussed at all.
    b. Next, assess whether the documentation meets a standard of excellence, only if the minimum criteria received a pass. The standard of excellence is a full/partial/none evaluation. A full is

granted if all aspects specified in the standard of excellence column were discussed, a partial is granted if one or more (but not all) were discussed, and a fail if none were discussed.
    c. It is important to note both for points 2a and 2b that the quality of the responses/documentation is not being assessed but rather if the element was considered and reflected on in any capacity. The purpose of the rubric is to demonstrate the dataset creators' thought process and provide transparency so that its reuse is based on a complete understanding of the dataset.
3. For each element, along with the grade, a comment on what specific information was used to determine that grade must be provided. Other comments and questions can also be included.

The evaluation of each dataset can take 30-60 minutes.

C.2.2 Applying the rubric to existing datasets through publications

The overall process for using the rubric is as follows:

1. Read the rubric to get familiarized with the elements and details that will be needed.
2. Gather and review all pertinent information that can be found about the dataset. This will include the research paper, appendices, the linked dataset, and any documentation associated with the externally linked dataset (e.g., README on github).
3. Review each element in the rubric individually by looking for it across all the information gathered in step 1. Some of the elements will be easier to locate than others because they will be titled specifically, whereas others may be discussed at any point.
    a. For each element, first assess whether the minimum standard of documentation has been fulfilled. To do this, provide a pass/fail evaluation, grant a pass if all aspects specified under the minimum standard were discussed and a fail if they were only partially discussed or not discussed at all.
    b. Next, assess whether the documentation meets a standard of excellence, only if the minimum criteria received a pass. The standard of excellence is a full/partial/none evaluation. A full is granted if all aspects specified in the standard of excellence column were discussed, a partial is granted if one or more (but not all) were discussed, and a fail if none were discussed.
    c. It is important to note both for points 2a and 2b that the quality of the responses/documentation is not being assessed nor the correctness of the technicalities but rather if the element was considered and reflected on in any capacity. The purpose of the rubric is to demonstrate the dataset creators' thought process and provide transparency so that its reuse is based on a complete understanding of the dataset and how it was developed.
4. For each element, along with the grade, a comment on what specific information was used to determine that grade must be provided. Other comments and questions can also be included.
5. For each dataset, evaluators must provide a reflection on their overall assessment of the documentation and rigour demonstrated in the dataset creation process.
6. For each dataset, evaluators must provide a confidence rating for their evaluation.

We estimate the evaluation of each dataset will take about 30-60 minutes once you are familiar with the framework.

C.2.3 How to interpret authenticity, reliability, and representativeness

It may be worth noting that the archival and digital curation perspectives that inform the evaluation framework are particularly important to interpreting the meaning of certain dimensions. Above all, the cluster of authenticity, integrity and reliability needs to be understood from this angle. They are closely related aspects, often treated or addressed by similar mechanisms, but they can be seen as analytically separate concepts. Here is an example.

When you download a data set of weather observations from a platform, you may want to verify if the file you have downloaded in fact is the data set you wanted to get, i.e., is it an authentic copy? You may be able to verify this with various checksums, both on the level of the file (e.g. a hashcode of the file, as commonly provided for downloads) and on the level of observations in some cases. In this case, you are concerned with **authenticity** - you want to verify that the data set is what it purports to be.

Authenticity does not guarantee you, however, that the observations in the data set are any good. A good observation of weather data is one that you can rely on to accurately represent how the weather actually was at the temporal and spatial locations covered by the data. Other aspects of goodness are reflected in the many quality standards for data, but when you want the data set to be able to stand in for the facts it represents, you are concerned with **reliability.** In other words, reliability is very much about the relationship of the data to whatever it represents. If the data set is a compilation of social media posts, then reliability will relate to the question whether these contributions were really posted, etc.

Integrity on the other hand refers to questions of tampering, errors, etc. For example, a dataset that lacks integrity is one for which we cannot assert that it contains *all* the items it originally contained, or that none of the items have been altered, falsified, or faked.

Consider a textbook case for records and archives for the difference between the three. A *passport* is a document that comes with very special features to prove that it can *stand in for the fact* that you are a citizen of the issuing country. Its **integrity** refers to the question whether it has been tampered with - has the photo been peeled off, have pages been removed or added? etc. The passport comes with features to prevent and check integrity. Its **authenticity** refers to the fact that it is indeed a passport of that country and that it indeed asserts the facts it states. Most of its special features are designed to make it easy to verify that (cf. banknotes). But imagine: a government could issue a perfectly authentic passport for a person who doesn't exist. That would be authentic, but it would not be reliable. The **reliability** rests on the relationship to the person it represents. We trust an authentic passport to be reliable because we trust the processes that governments have instituted and honed over the centuries to ensure that passports are only *issued to* authenticated citizens. But border control will use a machine-readable passport to look up and compare the information shown with the information stored in a database. When they do that, they verify reliability. For a deep dive into the archival perspective on what makes records authentic and reliable, see [5, 9].

Consider next a digital photograph taken during sunlight with a pro-grade digital lens reflex camera of a pantone color set of *whites* with standardized, specified colors, where the white balance is erroneously set at 'fluorescent light'. White balance relates to the color temperature of light: our eyes automatically adjust to different color temperatures, but a digital sensor does not. How an image looks on a screen is the result of computing it. In this case, the colors will not look very white on the photo without corrections to where the 'white point' should be located. The photo as taken is an **authentic** photo providing an **unreliable** representation of its subject. If you transfer the photograph yourself out of the camera you can also put in place mechanisms to verify integrity (including fixity checks and integrity checks using hash sums and the like on the file).

If you notice the error in color and then manually edit the binary code of the RAW file to set the white balance to the correct 'sunlight' setting, the photograph would in fact *lose* the property of 'integrity' since it has been tampered with (the hashcodes won't match), and it would *lose* the property of 'authenticity' since that was not the original setting, but it would gain in 'reliability' since the resulting color rendering would be a more accurate representation of how the colors should look. In this particular case, the fact that the subject of the photograph is standardized provides a *ground truth* that aids in verifying and assessing the photograph. Professional image processing software will be able to document both the 'as-taken' setting and 'to-use' setting of the photograph. Most photos, of course, are of subjects where this is much harder, and if the photograph is directly processed into a JPEG file, correcting white balance is much more difficult.

Finally, **representativeness** is related to reliability, but its perspective is much more narrowly focused on the question whether a data set accurately *represents* the overall set of observations or entities that it claims to be a sample of. For instance, for a data set of social media posts, the question will arise if it's representative of all platforms, all users, all topics, all media types, or various combinations of dimensions. All the statistical concepts around sampling apply as usual. Other data sets are not sampled out of an identified population but claim to stand for a general category so that representativeness is evaluated analytically, and so on.

C.2.4 How to interpret findability, accessibility, interoperability, and reusability (FAIR)

Note that this group of criteria are a direct representation of the widely used FAIR principles [138] for research data sets, adopted and adapted for machine learning. We provide a simple checklist to assess whether the documentation of the dataset discusses the application of FAIR principles. This checklist is derived from the following tools and resources:

- Minglu Wang and Dany Savard. 2023. The FAIR Principles and Research Data Management. (September 2023). https://doi.org/10.5206/EXFO3999
- FAIR data maturity model
- https://zenodo.org/records/5111307#.Yj3Vi5rMI-Q
- https://ardc.edu.au/resource/fair-data-self-assessment-tool/
- https://fairaware.dans.knaw.nl/

1. Findable
    a. A globally unique (cannot be reused by someone else) and persistent (valid over time) ID (like DOI) is assigned to the data.
    b. The dataset is described by metadata (PID, license, description, provenance, etc.). Further guidelines and definitions of provenance can be found from the DCMI and our glossary.
    c. The metadata specifies the identifier.
    d. The metadata and data is stored in a searchable repository.
2. Accessible
    a. The identifier navigates to the metadata and data.
    b. Retrieval of the data is specified by a standard communications protocol (i.e., all information and tools that are required are communicated to access the content of the dataset) which is open and free to access.
    c. The communications protocol specifies the authentication and authorization procedure, if needed (i.e., if the dataset is not open and free-to-access, the protocol specifies how access would be granted).
    d. The metadata record is available even if the data is not.
3. Interoperable
    a. Metadata and data are *in principle* readable by humans and machines (i.e., has a structured format, open standard).
    b. Metadata and data use controlled vocabularies (standardized and universal terms for indexing and information retrieval). Metadata standards can be found in the RDA Metadata Standards Catalog (https://rdamsc.bath.ac.uk/).
    c. Metadata and data is linked to other metadata and data using qualified references (i.e., relationship to the resource is specified).
4. Reusable
    a. Metadata and data are well-described as per domain-relevant standards, have detailed provenance (where did the data come from, who collected it, when, etc.), and clear and accessible license and usage information.

C.2.5 Guiding principles

We specify the following principles as “rules of thumb” to guide the evaluation of datasets:

1. Evaluate explicit documentation.

Evaluations should be made on the basis of documentation provided by the dataset creators, rather than performing evaluations ourselves.

2. Provide traceable comments.

The comments provided in the rubric to support the grade for each element should make recoverable the basis for the evaluation.

3. Minimum is easy, excellence is hard.

The evaluations for the minimum standard are meant to be *generous*. On the other hand, the standard of excellence criteria advocates for a high level of criticality, which is significantly harder to attain (compared to the minimum standard). The evaluations should only grant a ‘Full’ if all criteria are satisfied.

4. Don’t make excuses.

If there is no documentation provided to evaluate an element, then don't make excuses for the dataset creators and evaluate it yourself or think of it as unnecessary. If you truly feel the element does not apply for that dataset, then that means it's feedback for the rubric and that the element needs further work so it applies to all types of datasets.

C.2.6 Reflections & recommendations

In addition to the instructions on the process of using the rubric to evaluate datasets, the following recommendations are provided based on common reflections, challenges, and questions:

1. Completing an evaluation using the rubric requires iteration. A single pass through the rubric is often insufficient, especially for datasets that include various sources of documentation. The first iteration should be a step-by-step completion of each element in the rubric by looking for relevant information, keywords in the research paper or other dataset documentation. However, in doing so, sections of the documentation may be missed. It is therefore suggested to first evaluate the dataset by applying the rubric sequentially and then reviewing all the dataset documentation sequentially. The final step should be iterating as needed and zooming out.
2. The evaluation of elements will be interconnected, there can be notes to refer to the comment for another element.
3. If a context document is provided, it must be used to evaluate the elements. Although, the document will only provide information to fill in gaps rather than be sufficient to completely evaluate any element.
4. None of the elements should receive an N/A comment or grade.
5. The standard of excellence criteria should only be evaluated if the minimum standard criteria passes.
6. A failure for any element should not be provided based on the quality of the dataset but rather the documentation and reflection on the process of developing the dataset. For example, if the documentation acknowledges that the sample is not representative and can therefore introduce a bias- this is not considered a 'Fail'.
7. It is important to not evaluate the technical details provided but only evaluate the documentation. This means that evaluators should refrain from inferring the thought process or intention of the dataset creators based on their technical understanding of why the creators would develop their dataset in one way versus another. It is key to rely on explicit documentation only. This is important because the rubric assesses critical reflection around the dataset process not the quality of the dataset developed.

C.2.7 FAQ

1. Is there a difference between labeling and annotation?

Please refer to the glossary for definitions differentiating the two terms. The rubric doesn't require evaluation of the "labeling" process if the dataset does not have labels.

2. How to evaluate consistency and timeliness for suitability?

Data quality is often defined as fitness for purpose and is multi-dimensional, meaning that it's measured through more than one data quality dimension such as accuracy, completeness, etc. Suitability, in the rubric, evaluates whether dataset creators ensure that their dataset's quality meets the purpose defined. For example, a dataset of math problems may not require timely data but may require consistent data (i.e., data presented in the same format). For standard of excellence, multiple data quality dimensions will apply for evaluation but potentially not all.

3. Is representativeness applicable to synthetic data?

Representativeness is still applicable to synthetic datasets because synthetic data is still representative of reality. However, this is a *conceptual* representativeness rather than a *statistical* one.

4. Why does the evaluation criteria for authenticity discuss data processing specifically?

Data processing alters the authenticity of a digital object. Authenticity is dependent on the bits of information in a file. For example, if you download a dataset with a hash code and make copies of it, all copies will have the same hash code. However, if you perform data processing (which changes the bits), the hash code will no longer be the same. In the rubric, for the minimum standard, you evaluate whether the dataset creators validate and verify the authenticity of the data they are collecting. Whereas for standard of excellence, you evaluate whether they have

processes to ensure people that reuse their dataset are able to claim authenticity (i.e., maintaining the chain of authenticity).

5. For the data quality elements, are we evaluating that the dataset is suitable, authentic, has integrity, is representative, and is reliable OR that the dataset creators discuss their processes for ensuring these? If there is no mention of these qualities specifically, how do we evaluate them?

For data quality elements, you are evaluating whether the dataset creators discussed their processes for ensuring that their dataset is suitable, authentic, reliable, has integrity, and the extent to which it is representative (and why if it is not). Remember the guiding principle- "evaluate explicit documentation". We have added another guiding principle- "don't make excuses". If no documentation is provided for these data quality elements, then don't make excuses for the dataset creators and evaluate it yourself or think of it as unnecessary. If you truly feel the element does not apply for that dataset, then that means it's feedback for the rubric and that the element needs further work so it applies to all types of datasets.

6. Does hosting a dataset on huggingface make it 'findable'?

It depends, if it's hosted on huggingface but does not have a persistent identifier like a DOI, then it is not findable. See next question.

7. Why are URLs not acceptable for findability?

URLs are not considered "findable" because of the high likelihood of link rot (that the link over time will no longer be available). There are studies that show that academic papers are highly perceptible to link rot, eg: see [62]. Instead, we want persistent identifiers like DOIs to make sure the dataset is findable in the future.

8. What is the difference between findability and accessibility?

Findability is about a dataset being easily located. For example, if a publication provides a zenodo link to a dataset, that would make it findable (zenodo assigns a DOI to everything it publishes). So here we're looking for a dataset being easily located, indexed, catalogued, etc.

Accessibility is about whether a dataset can be opened and used and read. For example, is it in a format you can read, can you download it (i.e., is it retrievable), is the access blocked off via password-protection, are there access and authorization protocols?

A dataset would then be findable if there was a link pointing to it but not accessible if you couldn't open it because you didn't have the password for it and there was no documentation of an access protocol. On the other hand, if a dataset was open-access (eg, through github) but didn't have a persistent identifier (eg DOI) and wasn't indexed in a repository like zenodo then it would be accessible but not findable. Since accessibility rests on *accessing* the content, a URL alone is not enough to make it accessible either. So even if the dataset is available through github there must be other documentation that provides any further information needed to access the content and metadata.

9. Can you provide further clarification for evaluating interoperability (especially standard of excellence)?

For the minimum standard, the documentation must explain how the dataset can be integrated with other data and workflows. An example of that is that the data can be exported to popular, standard formats. For the standard of excellence, the data and metadata must use controlled vocabularies and link to other resources with qualified references. For example, metadata can be created using controlled vocabularies like the W3C's Data Catalog Vocabulaire (DCAT) model which defines terms like dataset vs data service, catalog (as a subclass of dataset), and so on. Please see this blurb from FAIR about qualified references:

> "A qualified reference is a cross-reference that explains its intent. For example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*, or *X see also Y*. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good data model. To be more concrete, you should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers)." [33].

Zenodo also has a webpage that describes how it fulfills the FAIR principles for its datasets [143].

C.3 Rubric

See Section A of the Appendix.

C.4 Rubric worksheet

See Section B of the Appendix.

C.5 Sample evaluations

Please note that the sample evaluations were performed using the version of the rubric at the time of evaluating datasets from round 3. Note also that the description column and cited references are deleted below for space, see full rubric with references.

The State of Data Curation at NeurIPS: Appendix
Example 1

Paper: FS-Mol: A Few-Shot Learning Dataset of Molecules [129]

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard	PASS/ FAIL	Criteria to meet standard of excellence	Full/ Partial/ None
			SCOPE		
1	Context, purpose, motivation	Pass	Paper introduction discusses the problem domain and why a new dataset is needed; see 'related work' in paper and appendix B in supplementary material ('related work details') for comparison to existing datasets.	Full	Section 7 of paper discusses how dataset can be used outside of its original context ("it is now possible to evaluate… we note that transfer of results to realistic projects is not guaranteed to be successful…")
2	Requirements	Pass	Section 2 of paper (especially " 2.2 Desired Attributes of a QSAR Few-Shot Dataset and Benchmark") explicitly derives design requirements to create the dataset.	Partial	No explicit discussion of intrinsic biases introduced by problem formulation; other approaches to formulating the problem are discussed in 'related work' section of paper (discussing other datasets and their features)
			ETHICALITY AND REFLEXIVITY		
3	Ethicality	Pass	No discussion of consent (no human data); pg 9 'societal impacts' section discusses benefits of creating the dataset.	Fail	No additional discussion of ethical consideration throughout the paper or supplementary documentation.
4	Domain knowledge & data practices	Pass	On pg 2 of papers, authors state aim to "demonstrate the utility of few-shot learning methods in an important domain, namely QSAR, which does not provide an obvious generic pretraining corpus (such as in NLP or computer vision). The proposed dataset is specifically designed to replicate the challenges of machine learning in the very low data regime of drug-discovery projects" (focus on drug-discovery domain)	Partial	README in GitHub repo discusses activities to be undertaken to re-use the dataset "Hence, in order to be able to run MAT, one has to clone our repository via…" – not directly discussing any domain knowledge needed.
5	Context awareness	Fail	Research goals are described but not positioned relative to researchers' intellectual/political believes; researcher positions not disclosed/no positionality statement included.	None	Failed minimum criteria.

6	Environmental footprint	Fail	No assessment of environmental footprint	None	Failed minimum criteria
DATA PIPELINE					
7	Data collection	Pass	ExtractDataset.ipynb from GitHub repo describes how data were gathered by querying ChEMBL; section 3 of paper explains data acquisition process in detail ("the reason why we remove large assays is…")	Partial	Section B of supplementary material describes other few-shot learning and molecular property datasets (e.g. why they used ChEMBL instead of other sources); no explicit discussion of criteria for source selection, why criteria were chosen, or how other sources were validated against criteria.
8	Data processing	Pass	ExtractDataset.ipynb from GitHub repo describes how data were cleaned and split into test vs validation assays.	Full	Section 3 of paper describes decisions behind data processing (e.g. "In this way, our proposed meta-testing tasks closely mimic the new-lead optimization problem, where a completely unseen task is presented for adaptation.")
9	Data annotation	Pass	"Binary Classification Task" section of paper discusses some annotation activity	None	No discussion of robustness of annotations.
DATA QUALITY					
10	Suitability	Pass	Section 6 and first paragraph of section 7 describe and demonstrate dataset appropriateness for purpose.	Partial	Documentation does not explicitly discuss accuracy/completeness/timeliness of the chosen dataset, but Section 6 of the paper demonstrates the utility of the dataset for its intended purpose by providing "a set of results for all three categories of few-shot learning, with representative methods of the use of this dataset in each".
11	Representativeness	Pass	Section 3 on pg 3 of main paper describes how the 'sample' of the dataset is taken from the overall population (the ChEMBL database); also on pg 9 "the few-shot baselines we provide checkpoints and results for are only a representative set, rather than a complete survey of the current state of the field"	None	No explicit discussion of biases.
12	Authenticity	Pass	No explicit discussion of authenticity but extractdataset.ipynb does discuss how initial raw data were obtained (e.g. describes process by which database was queried)	Partial	No explicit discussion of future authenticity/preservation processes, but does discuss in section A of supp material how dataset documentation facilitates re-use more generally.
13	Reliability	Pass	Section 5 of paper discussing benchmarking procedures (i.e. making sure that the dataset is useful for what it's supposed to be useful for)	Partial	No explicit discussion of reliability management in the context of future re-use; section A of supplementary material discusses how the dataset documentation facilitates re-use.
14	Integrity	Fail	No discussion of dataset integrity or preservation processes (section H of	None	Failed minimum criteria.

			supplementary document does not actually discuss a maintenance plan or means of maintaining accuracy/consistency over time).		
15	Structured documentation	Fail	No standardized context document	None	Failed minimum criteria
DATA MANAGEMENT					
16	Findability	Fail	No persistent identifier provided.	None	Failed minimum criteria
17	Accessibility	Pass	Section F of supplementary material describes computational resources used; GitHub README states the tools and steps required to access data content.	Partial	GitHub repo includes a code of conduct document, as well as protocols for contributing and for security reporting.
18	Interoperability	Pass	README in GitHub repo describes how to use the dataset with “three key few-shot learning methods”; dataset.ipynb describes the machine/human readable metadata.	Full	Dataset.ipynb describes the controlled vocabularies for specific dataclasses (e.g. task_name as a string describing the task each point is taken from)
19	Reusability	Fail	From data contents of GitHub repo it does not appear that data or metadata contain provenance information about where the dataset came from/when/who collected it; license is included in the GitHub repo.	None	Failed minimum criteria.

Paper: American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers [18]

	CURATORIAL ELEMENT	DOCUMENTATION LEVEL			
		Criteria to meet minimum standard	PASS/ FAIL	Criteria to meet standard of excellence	Full/ Partial/ None
			SCOPE		
1	Context, purpose, motivation	Pass	Paper Introduction and section 6 (Applications) discusses the problems and relevance, and 'Related Literature' (section 2) discusses other similar datasets.	Full	'Applications' section on pg 6 of supplementary material discusses "multiple applications that can be facilitated by the American Stories dataset"
2	Requirements	Pass	Paper Introduction (pg 2, "To address these limitations, we develop…") introduces certain requirements.	Partial	On pg 3 of paper ,documentation reflects on the bias potentially introduced by scanning illegible newspapers; other approaches are discussed in Section 2 on Related Literature (but not specifically other approaches the authors considered)
			ETHICALITY AND REFLEXIVITY		
3	Ethicality	Pass	Some harms (e.g. offensive language) are discussed in Section 7: Conclusion. Consent is discussed in datasheet (pg 14 of supplementary material)	Partial	Some additional discussion of copyrights/accessibility on pg 3 of paper
4	Domain knowledge & data practices	Pass	Pg. 23 of paper (the datasheet) addresses the professors, research assistants, and students involved in data collection	Partial	Datasheet states "There are a large number of potential uses in the social sciences, digital humanities, and deep learning research"
5	Context awareness	Pass	No positionality statement but several mentions throughout the datasheet showing awareness of social context ("This dataset contains unfiltered content composed by newspaper editors, columnists, and other sources. It reflects their biases and any factual errors that they made."), and section 7 of the paper reflects on the historicity of dataset contents	Partial	Section 3 of paper touches on assumptions going into methodological choices (e.g. on pg 3, "We do not OCR ads because…")
6	Environmental footprint	Fail	No environmental assessment.	None	Failed minimum criteria
			DATA PIPELINE		
7	Data collection	Pass	Described in 'Composition' (pg 11) and 'Collection Process' (pg 13) sections of datasheet in supplementary material	Partial	We have a lot of information about how the data were collected, but I still don't see where in the documentation it specifies the criteria they used to select data sources or how data sources were validated against these criteria (e.g. why the library of congress dataset?).
8	Data processing	Pass	Pre-processing section of datasheet (pg 14 of supplementary material) describes process of cleaning and wrangling data	Full	Sections 3, 4, and 5 of main paper discuss the implications of processing decisions (e.g. on computing cost and efficiency)

9	Data annotation	Pass	Student annotation is discussed ins Section 5 'Pipeline Evaluation' of main paper	Full	Student annotations were used as 'ground truth' for model training; see pg 5 of supplementary material
DATA QUALITY					
10	Suitability	Pass	Section 5 of paper evaluates the pipeline for accuracy, legibility, and comparison to other OCR engines	Full	See explanation for minimum criteria
11	Representativeness	Pass	Sampling approach discussed in datasheet (pg 13 of supplementary material) – it includes everything in the Chronicling American scan collection.	Full	Section 3 of paper discusses how illegible papers and their inclusion/exclusion in the dataset could bias results.
12	Authenticity	Pass	Pipeline for generating data is included in the Github repo (https://github.com/dell-research-harvard/AmericanStories?tab=readme-ov-file); no explicit discussion of authenticity	None	No explicit discussion of authenticity in future re-use.
13	Reliability	Pass	Section 5 of paper (Pipeline Evaluation) describes verification and validation processes used to ensure reliability.	Full	Maintenance section of datasheet discusses how errors will be corrected in future (and uploaded to HuggingFace)
14	Integrity	Pass	Documentation does not explicitly discuss integrity but datasheet does emphasize that "material is complete and unaltered"	Full	Maintenance section of datasheet describes preservation processes in place (e.g. old versions still accessible via HuggingFace)
15	Structured documentation	Pass	Paper and supplementary material include a datasheet (Gebru et al)	Full	All mandatory components of datasheet are answered.
DATA MANAGEMENT					
16	Findability	Pass	DOI available on HuggingFace page (10.57967/hf/0757)	Full	Data and metadata stored in searchable repo (HuggingFace)
17	Accessibility	Pass	Steps for accessing data listed on HuggingFace page data card and described in 'Distribution' section of datasheet (pg 15 of supplementary material	Full	Communications protocol described in 'Maintenance' section of datasheet (supp material pg 16)
18	Interoperability	Pass	Pg 4 of paper describes readable formats of metadata and data ("The raw files are in a json format, and the Hugging Face repo comes with a setup script that easily allows people to download both raw and parsed data to facilitate language modeling and computational social science applications."; lots of metadata info included on HuggingFace page	Full	See HuggingFace page for controlled metadata vocabularies
19	Reusability	Pass	Some provenance information included in metadata (e.g. where it came from, associated newspaper, but not who collected it/when)	Partial	Pg 16 of supplementary material (datasheet) states "The dataset is distributed under a Creative Commons CC-BY license. The terms of this license can be viewed at https://creativecommons.org/licenses/by/2.0/"

C.6 Glossary

Table 1: Glossary Terms

Term	Definition	Discussion/Example	Sources
Context document	"Interventions designed to accompany a dataset or ML model, allowing builders to communicate with users".	Context documents are standardized documentation formats that convey information about the dataset, types of context documents include datasheets, nutrition labels, etc.	[11]
Data annotation	Although data annotation and labelling are often used interchangeably in ML, labelling is a subset of annotation. (See labelling) Data annotation refers to the process of adding information to a dataset to provide more context. For example, adding metadata.	Annotation can include metadata about the units of measurement.	
Data practices	"What and how data are collected, managed, used, interpreted, released, reused, deposited, curated, and so on…"	Data practices are the decisions made in the collecting, interpretation, etc. of data.	[10]
Extrinsic bias	Extrinsic bias refers to bias that exists within the dataset which are reflections of social, historical biases.	"Extrinsic bias is concerned with a view of a biased dataset "from the outside." The argument is that an already-biased dataset can cause even innocent software to produce a biased outcome - and may look like people saying things such as "the data made me do it." … If we fail to remember that a dataset is biased, then we may treat it as "fair" or "representative," harming people who have been excluded from it."	[98]
Informed consent	Informed consent is a standard ethical principle of research with human subjects that rests on the commitment that participants • Are fully informed • Decide voluntarily • Before research is conducted. Its application in online environments is complicated by the shift in technology and methods (see [28]), but the principle remains important.	In conventional human subject studies such as interviews, an IRB reviews ethics protocols and evaluates if the research is compliant with principles such as the proportionality principle. In social media research, things get complicated. In some situations, implied consent (see [53]) can be present but must be justified. In the case of LLMS, widespread data collection without consent has prompted massive ethical and legal concerns.	A discussion of Twitter research ethics [28] is a good start.
Intrinsic bias	"The ways in which we change the data "from the inside" of data science work-processes while we are preparing the data for modeling." Intrinsic bias is the bias data workers introduce to the dataset.	"Through practices of data wrangling, curation, and feature-engineering, humans make a series of decisions about how to treat their data, and those decisions may inadvertently introduce bias into the data."	[98]

Labelling	Labelling is a specific type of annotation that involves assigning a predefined category to a data item.	Labelling tweets on Twitter as 'human-generated' or 'bot-generated'.	
PID: persistent identifier	"Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. [IDs] must be persistent. It takes time and money to keep web links active, so links tend to become invalid over time. Registry services guarantee resolvability of that link into the future, at least to some degree."	ORCID iDs are persistent identifiers for people. DOIs are persistent identifiers for journal articles, datasets, etc.	[34]
Population	Mathematical term used to describe a group of units sharing a common trait.		
Positionality statement	"Researcher/Practitioner Self-Disclosure: Practice should involve a disclosure of the researcher's position in the world, her or his goals, as well as the researcher's position in her or his intellectual and, to an appropriate extent, political beliefs"	See [72] for examples.	[3]
Problem formulation	The process with which a problem is formulated and the methods we use to define and measure it define the lens with which we see the problem. "...the problems we solve with data science are never insulated from the larger process of getting data science to return actionable results…, these ends are very much an artifact of a contingent process of arriving at a successful formulation of the problem, and they cannot be easily decoupled from the process at arriving at these ends."[105:9].	For example, O'Neil describes how proxies are used to quantify university excellence through indicators that are easily collected such as SAT scores, student-teacher ratios, and acceptance rates rather than through students' learning experience, happiness, productivity, personal fulfillment, etc. [101]. See [105] for additional examples.	[105]
Proportionality principle	In ethics, it is understood that actions have positive and negative effects simultaneously. This is called *double effect*. "Applications of double effect always presuppose that some kind of proportionality condition has been satisfied. Traditional formulations of the proportionality condition require that the value of promoting the good end outweigh the disvalue of the harmful side effect."	In medicine, a surgeon may cause harm to a patient's skin (negative) in order to save their heart (positive). It would not be permissible for a surgeon to open up someone's chest just to get a better look or take a selfie with it because that would violate the proportionality principle.	[92]
Provenance	Provenance information provides a trail of history about how the data originated, how it's changed, who was involved, and more.	See the following blurb from the FAIR principles… "For others to reuse your data, they should know where the data came from … who to cite and/or how you	[35]

		wish to be acknowledged. Include a description of the workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else that you may have transformed or completed?"	
Reflexivity	"Questions of reflexivity ask us to consider who we should listen to and why, how to place actors' ideas in a larger field of power, questions about our own relationship to actors' theories of the world. Reflexivity asks us to approach our work with epistemological unease because we are always at risk of reproducing categories that reify power."		[93, 132]

C.7 Further readings

The following readings 1) showcase how data curation is discussed in data science and machine learning studies, 2) contain context for relevant data curation terms, concepts, and frameworks, and 3) provide important terminology for ML benchmarks.

C.7.1 Data curation in data science

A vast amount of literature points to the datasets used for training machine learning models to be the source for introducing bias in model results leading to a call for increased documentation of datasets used in ML. Emerging research has proposed context documents – "interventions designed to accompany a dataset or ML model, allowing builders to communicate with users" [11]. The following are types of relevant context documents.

1. 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 (November 2021), 86–92. https://doi.org/10.1145/3458723

Datasheets are one of the most popular methods of documenting the process of developing datasets as well as providing a dataset description. This paper is a good introduction to how dataset documentation is evaluated [32].

2. 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*, April 21, 2020, Honolulu HI USA. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445

Madiao et al. developed a resource - checklist for AI fairness - based on findings of current practitioners processes, needs, and requirements for developing fair AI models [85].

3. 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

Bender and Friedman develop 'data statements'- a resource for NLP training datasets to be documented in order to mitigate bias and exclusion [5].

Topics like dataset documentation in ML are often discussed as a part of data practices, data work, or dataset development. The following studies talk about stages of dataset development processes, how data scientists or data workers approach their data work, and the importance and impact of decisions made during the dataset development.

1. Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–37. https://doi.org/10.1145/3476058

This paper discusses how documentation captures underlying values of data practices in machine learning (specifically computer vision tasks) [121]. Specifically, publications are analyzed to understand the documentation and communication of datasets. The findings showcase the practices that are silenced (such as data work, context, positionality, and care) over those that are (wrongly) embraced such as model work, universality, and so on. This reading helps reflect on and understand how intrinsic bias can be introduced within datasets.

2. 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*, May 06, 2021, Yokohama Japan. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

Through interviews with AI practitioners, Sambasivan et al. find that poor data practices in high-stakes AI domains (i.e., practices that do not prioritize data quality) lead to data cascades which are negative impacts of data issues [117].

3. Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP (January 2022), 1–14. https://doi.org/10.1145/3492853

Miceli et al. discuss that while we often recognize that there is bias in the datasets and their processes used for ML models, it is often ignored that this bias is a result of power inequities [93]. The authors analyze data bias, data work, and data documentation from a "power-aware" framing as compared to a "bias-oriented" one. This paper provides an interesting shift in perspective which further illuminates the importance of reflexivity in data work.

4. Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022, New Orleans LA USA. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644

This paper studies how data processing leads to different types of forgetting and where and how each type of forgetting occurs in the machine learning stack [98]. Forgetting is conceptualized as the practice that occurs when choices are made about what data is kept, what it represents and so forth (therefore by designing a dataset in a given way, we *remember* only its current state, and *forget* the decisions, the erased data, etc.). This is a great paper for a deep dive into the various types of design decisions that impact the eventual dataset.

The previous studies discuss aspects of data curation as dataset development. However, some ML studies have started discussing the importance of data curation by referencing archival studies and digital curation directly. These are included below.

1. 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*, July 21, 2021, Virtual Event USA. ACM, Virtual Event USA, 695–703. Retrieved November 11, 2022 from https://dl.acm.org/doi/10.1145/3461702.3462598

This study discusses principles for ethical data curation based on race critical race theory and data feminism to improve the reflection of power, bias, and values in data processes and thereby improve transparency and accountability of AI systems [68].

2. 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* (*FAccT '21*), March 01, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 610–623. . https://doi.org/10.1145/3442188.3445922

This paper discusses the potential risks of language models (and by extension other ML/AI systems) [6]. The authors recommend a shift towards careful, reflective practices around datasets and model development along with a greater focus towards documentation.

3. 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*, January 27, 2020, Barcelona Spain. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

This paper highlights that practices from archival studies have experience dealing with consent, power dynamics, transparency, and ethics and that these practices should be adopted into data collection and annotation practices in machine learning [56].

C.7.2 Data curation

Data curation involves "maintaining and adding value to digital research data for current and future use" [20]. The following studies introduce data/digital curation terminology and the data curation lifecycle model (parallel to ML model pipelines) with the aim to familiarize how the data curation field approaches data work.

1. Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3, 1 (August 2008), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

The paper introduces the curation lifecycle model by emphasizing it as a lifecycle (as opposed to a linear process). Each stage of the model is briefly introduced [45].

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

This paper discusses each stage in depth including the tasks performed, how each stage leads to the next, and the expected outcomes [47].

3. Digital Curation Centre. Glossary. *Digital Curation Centre*. Retrieved January 21, 2024 from https://www.dcc.ac.uk/about/digital-curation/glossary

This is a glossary of common digital curation terms - to be returned to as a resource, as needed [21].

4. Carole L Palmer, Nicholas M Weber, Trevor Muñoz, and Allen H Renear. Foundations of Data Curation: The Pedagogy and Practice of "Purposeful Work" with Research Data. 16.

This is an introductory paper to the field of data curation and its place within archival studies, library studies, and computer science [103].

C.7.3 Benchmarking in ML

Benchmarking is often not a well discussed topic in machine learning papers. The below list is compiled to introduce commonly used terms including: benchmark dataset, benchmark tasks, simulator, synthetic dataset, baseline method, benchmark suite, etc.

1. Matthew Stewart. 2023. The Olympics of AI: Benchmarking Machine Learning Systems. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b

This paper explains terms benchmark, benchmark dataset, benchmark tasks, baseline method, and benchmark suite [89].

2. Ramona Leenings, Nils R. Winter, Udo Dannlowski, and Tim Hahn. 2022. Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage* 257, (August 2022), 119298. https://doi.org/10.1016/j.neuroimage.2022.119298

This paper explains benchmark term and concept [71].

3. Kim Martineau. 2021. What is synthetic data? *IBM Research Blog*. Retrieved January 21, 2024 from https://research.ibm.com/blog/what-is-synthetic-data

This paper explains the term *synthetic data* [61].

4. Nataniel Ruiz. 2019. Learning to Simulate. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/learning-to-simulate-c53d8b393a56

This paper explains the term *simulator* [115].

D. Additional information about rubric evaluations

D.1 List of datasets

Table 2: List of datasets evaluated from the NeurIPS Datasets & Benchmarks track

Dataset Number	Round	Paper Title	Dataset Abbreviation	Publication Year	Reference
1	Training	Programming Puzzles	program_puzzles	2021	[123]
2	Training	Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation	open_bandit	2021	[116]
3	Training	SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables	scigen	2021	[96]
4	Training	MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing	moma-lrg	2022	[83]
5	Training	CEDe: A collection of expert-curated datasets with atom-level entity annotations for Optical Chemical Structure Recognition	cede	2022	[49]
6	Round 1	LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation	loveda	2021	[137]
7	Round 1	RELLISUR: A Real Low-Light Image Super-Resolution Dataset	rellisur	2021	[1]
8	Round 1	Measuring Mathematical Problem Solving With the MATH Dataset	math	2021	[44]
9	Round 1	DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection	dgraph	2022	[142]
10	Round 1	Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery	change_event	2022	[87]
11	Round 1	CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets	caesar	2022	[54]
12	Round 1	GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization	globem	2022	[141]
13	Round 1	ClimateSet: A Large-Scale Climate Model Dataset for Machine Learning	climateset	2023	[58]
14	Round 1	BubbleML: A Multiphase Multiphysics Dataset and Benchmarks for Machine Learning	bubbleML	2023	[41]
15	Round 1	DataComp: In search of the next generation of multimodal datasets	datacomp	2023	[30]
16	Round 2	The CPD Data Set: Personnel, Use of Force, and Complaints in the Chicago Police Department	cpd	2021	[48]
17	Round 2	The Tufts fNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces that Generalize	tufts	2021	[51]
18	Round 2	How Would The Viewer Feel? Estimating Wellbeing From Video Scenarios	viewer	2022	[90]
19	Round 2	The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only	refinedweb	2023	[108]

20	Round 2	Stanford-ORB: A Real-World 3D Object Inverse Rendering Benchmark	stanfordorb	2023	[63]
21	Round 3	FS-Mol: A Few-Shot Learning Dataset of Molecules	fs-mol	2021	[129]
22	Round 3	Evaluating Out-of-Distribution Performance on Document Image Classifiers	eval_ood	2022	[67]
23	Round 3	Dungeons and Data: A Large-Scale NetHack Dataset	dungeons	2022	[37]
24	Round 3	VisAlign: Dataset for Measuring the Alignment between AI and Humans in Visual Perception	visalign	2023	[70]
25	Round 3	American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers	amerstories	2023	[18]
26	Round 4	KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects	kespeech	2021	[133]
27	Round 4	Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI	habitat	2021	[112]
28	Round 4	FLAIR: a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery	flair	2023	[31]
29	Round 4	MedSat: A Public Health Dataset for England Featuring Medical Prescriptions and Satellite Imagery	medsat	2023	[120]
30	Round 4	PUG: Photorealistic and Semantically Controllable Synthetic Data for Representation Learning	pug	2023	[9]
31	Round 5	Constructing a Visual Dataset to Study the Effects of Spatial Apartheid in South Africa	spatial_apart	2021	[125]
32	Round 5	A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning	spoken_lang	2021	[59]
33	Round 5	Ambiguous Images With Human Judgments for Robust Visual Event Classification	ambiguous	2022	[118]
34	Round 5	SCAMPS: Synthetics for Camera Measurement of Physiological Signals	scamps	2022	[91]
35	Round 5	Objaverse-XL: A Universe of 10M+ 3D Objects	objaverse	2023	[17]
36	Round 5	Timers and Such: A Practical Benchmark for Spoken Language Understanding with Numbers	timers	2021	[82]
37	Round 5	CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge	creak	2021	[102]
38	Round 5	CLEVRER-Humans: Describing Physical and Causal Events the Human Way	clevrer	2022	[88]
39	Round 5	OpenProteinSet: Training data for structural biology at scale	openprotein	2023	[2]
40	Round 5	SSL4EO-L: Datasets and Foundation Models for Landsat Imagery	ssl	2023	[130]
41	Round 5	OpenFilter: A Framework to Democratize Research Access to Social Media AR Filters	openfilter	2022	[114]

42	Round 5	PROSPECT: Labeled Tandem Mass Spectrometry Dataset for Machine Learning in Proteomics	prospect	2022	[126]
43	Round 5	MoCapAct: A Multi-Task Dataset for Simulated Humanoid Control	mocapact	2022	[136]
44	Round 5	MADLAD-400: A Multilingual And Document-Level Large Audited Dataset	madlad	2023	[64]
45	Round 5	Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images	seeing	2023	[76]
46	Round 5	STAR: A Benchmark for Situated Reasoning in Real-World Videos	star	2021	[139]
47	Round 5	BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling	bitod	2021	[75]
48	Round 5	ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment	actionsense	2022	[19]
49	Round 5	RenderMe-360: A Large Digital Asset Library and Benchmarks Towards High-fidelity Head Avatars	renderme	2023	[104]
50	Round 5	PTADisc: A Cross-Course Dataset Supporting Personalized Learning in Cold-Start Scenarios	ptadisc	2023	[50]
51	Round 5	ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild	conflab	2022	[113]
52	Round 5	CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer	csaw	2021	[127]
53	Round 5	WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges	wikichurches	2021	[4]
54	Round 5	Mathematical Capabilities of ChatGPT	mathgpt	2023	[29]
55	Round 5	SubseasonalClimateUSA: A Dataset for Subseasonal Forecasting and Benchmarking	subseas	2023	[97]
56	Round 5	COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening	covid	2021	[140]
57	Round 5	Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks	shifts	2021	[86]
58	Round 5	Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark	wukong	2022	[36]
59	Round 5	Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets	sign_lang	2022	[100]
60	Round 5	SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis	synmob	2023	[144]

D.2 Method for selecting datasets

In order to select datasets to evaluate, we first filtered all publications from the NeurIPS Datasets and Benchmarks track based on the following inclusion and exclusion criteria:

Inclusion:

1. Submitted paper creates a dataset.

Exclusion:

1. The paper discusses supplementary material and documentation, but it is not available.
2. The paper contributes a dataset but there are other contributions also made, which are discussed to a greater length, and the discussion of the dataset creation is less than one section of the paper.

Edge cases were discussed among the authors and categorized as included or excluded. From the remaining datasets, we randomly selected datasets to evaluate for each round.

D.3 Evaluation consistency

Additional note on methods: In the final round, four reviewers performed double coding for 30 datasets, each reviewing on average 15 datasets. Accordingly, we measured IRR with a one-way mixed, consistent, average-measures intra-class coefficient (ICC). The final round, as with Round 3 and 4, also consisted of a disagreement review. After the disagreement review, additional corrections were made for consistency. This included:

1. If a reviewer changed their score from 'pass' to 'fail' for the minimum standard, the standard of excellence score was automatically changed to 'none'.
2. If a reviewer changed their score from 'fail' to `pass' for the minimum standard, the standard of excellence score was automatically changed to match the 2nd reviewer's score.

Additional results (IRR): We measured IRR per datasets and rounds as well as rubric categories. Specifically, the lowest ICC value for a given dataset in the initial rounds (training to round 4) was 0.45, indicating fair agreement, while the highest was 0.94, signifying excellent agreement. Subsequently, in the final round, the median ICC value for the 30 datasets evaluated was 0.90, with the highest ICC value for a given dataset as 1 indicating perfect agreement (Figure 4).

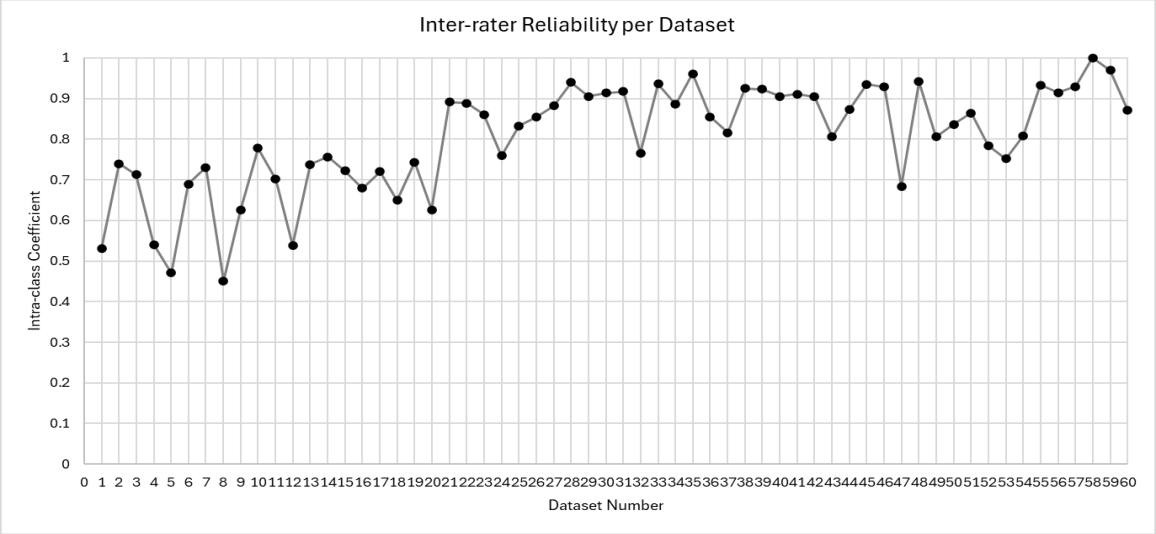

Figure 4. IRR for datasets. Datasets 1-30 measured with two-way ICC and 31-60 with one-way ICC.

Despite the greater level of interpretation present when evaluating IRR across elements as compared to datasets, the final round had a median ICC value >0.82 across all rubric categories (Figures 3a and 3b). Greater degree of variability can also be seen for the earlier rounds as compared to round 5, with round 5 ICC values having a median of 0.83-0.98 across rubric categories. Furthermore, many elements had perfect agreement especially for the minimum standard criteria ('context, purpose, motivation', 'context awareness', 'environmental footprint', 'data collection', 'data processing', 'data annotation', 'suitability', 'reliability', 'structured documentation', 'findability', and 'reusability'). The IRR for rubric categories from rounds 1-4 is shown in Figure 5.

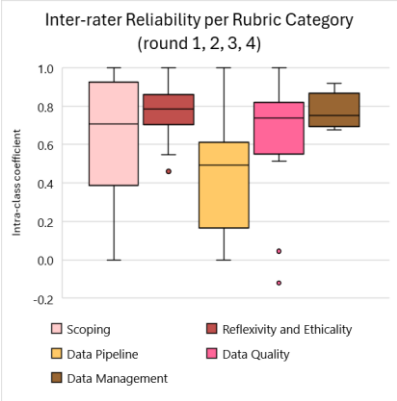

Figure 5. IRR for rubric categories calculated using two-way ICC for rounds 1-4.

Additional results (consistency): In addition to IRR, we measured inconsistency across evaluations by calculating the number of disagreements between reviewers. Disagreements were categorized in the following manner for training and rounds 1-4:

- Minor disagreement, standard of excellence (Minor, exc): instances where 1 of 3 reviewers disagree, e.g., Full, None, Full
- Major disagreement, minimum standard (Major, min): instances where 1 of 3 reviewers disagree, e.g., Pass, Pass, Fail
- Major disagreement, standard of excellence (Major, exc): instances all reviewers disagree, e.g., Full, Partial, None
- No disagreement, standard of excellence: instances where 1 of 3 reviewers gives a Partial evaluation and the other 2 reviewers agree, e.g., Partial, None, None

This was further simplified in round 5:

- Major disagreement, minimum standard (Major, min): instances where 1 of 2 reviewers disagree, e.g., Pass, Fail
- Major disagreement, standard of excellence (Major, exc): instances 1 of 2 reviewers disagree, e.g., Full, None
- No disagreement, standard of excellence: instances where 1 of 2 reviewers gives a Partial evaluation, e.g., Partial, None, None

We observed that the inconsistencies across datasets had markedly decreased by the final round. Figure 6 illustrates the trend in inconsistencies across datasets throughout multiple rounds of evaluations. The trend in major disagreements showed an even more pronounced reduction: for major inconsistencies under the minimum standard (Major, min), the inconsistency rates decreased significantly from initially high levels in the training and early rounds to much lower levels by Round 5. Similarly, major inconsistencies under the standard of excellence (Major, exc) see a sharp reduction, highlighting the impact of targeted improvements in rubric clarity and rater understanding. When considering all types of inconsistencies combined, the graph shows a substantial decrease from over 30% in the earliest phases to around 2% by the end of the final round, demonstrating nearly complete alignment among evaluators. This uniformity is indicative of the rubric's maturity as a tool for assessing dataset documentation quality.

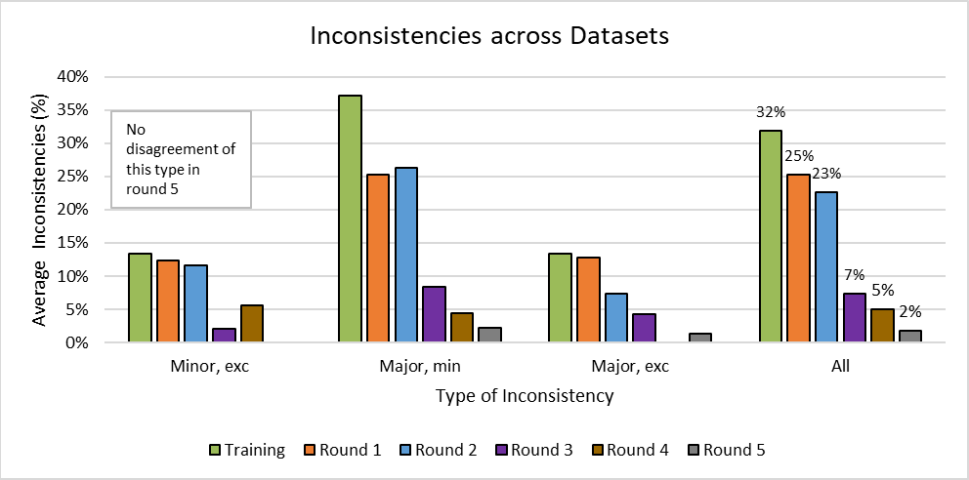

Figure 6. Inconsistencies in the rubric application across datasets.

In analyzing the inconsistencies among elements across several rounds, we found a clear trend of decreasing inconsistencies across almost all rubric elements. Progressing through rounds, the inconsistencies in even the most challenging elements had been markedly reduced (Figure 7a). By round 5, the percentage of inconsistencies reduced to 10% and under and only persisted for 7 of the 18 rubric elements (Figure 7b).

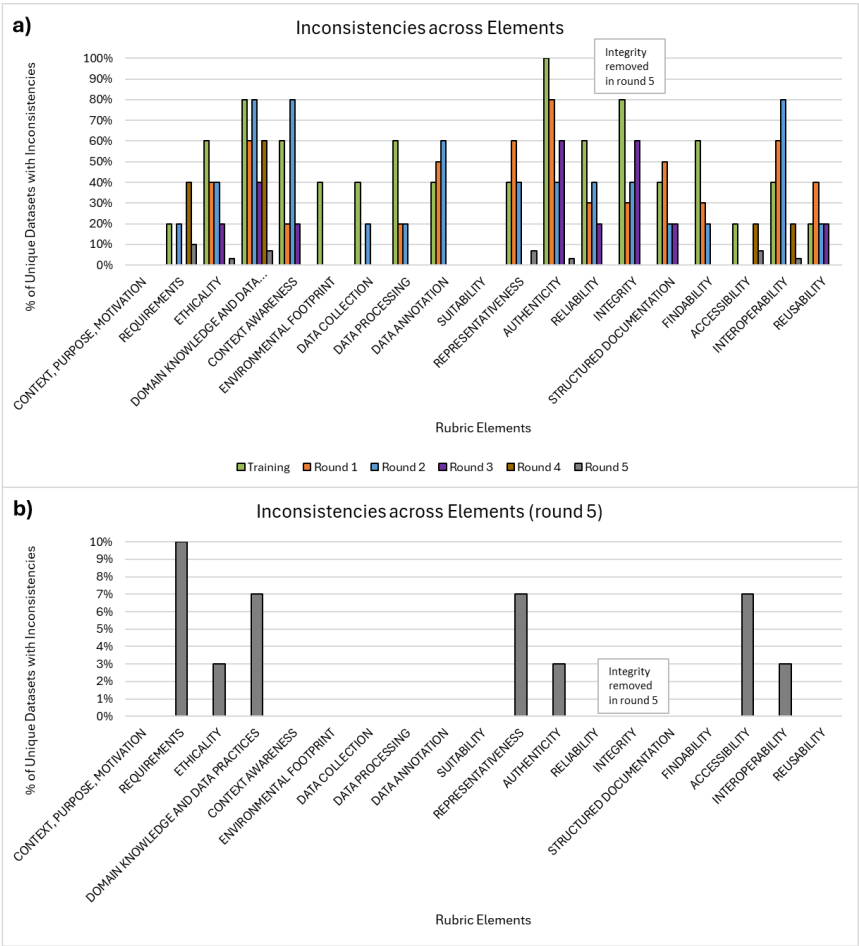

Figure 7. Inconsistencies in the rubric application. (a) Shows inconsistencies across elements, and particularly in the final round (b).

D.4 Positionality, reflection, and contributions

Positionality begins by recognizing that knowledge is always produced from a certain perspective and that this perspective is always dependent on where we stand. Feminist standpoint theory emphasizes that there is no knowledge without a subject who knows. By identifying the location and situatedness of the knowing subject, we gain a better understanding of the knowledge in question. This makes the knowledge more robust, not less - it is a *stronger* version of objectivity than the flawed idea that knowledge can be fully disassociated from a subject and would embody a perspective of 'nowhere' [38–40]. By clarifying from where we speak, we can be more objective [94]. This is true for individuals as much as for groups. A key emphasis is often on recognized dimensions of intersectionality including but not limited to gender, age, sex, race, class, religion, dis/abilities, or geographies. In addition, the knowledges each team member brings to a group are also highly relevant.

Positionality statements can take many different shapes dependent on context and purpose. Beyond simply declaring aspects of identity, there is most value in *reflecting* how specifically these aspects *intersect* and shape the work being presented. Below, we will include positionality statements from each person disclosing what they are comfortable to disclose, followed by joint reflections on how the interaction of these perspectives have influenced this project.

Tegan: My primary academic training is in machine learning, specifically deep learning. Much of my master's and doctoral work examined generalization and learning behaviour of deep networks on real-world data (e.g. for climate and video data) using empirical methodology, giving me a strong and grounded appreciation for the importance of data in machine learning. As a professor, I am increasingly engaged in the interdisciplinary research I believe is necessary to make the field of machine learning/AI more responsible and empirically rigorous. All of my training and much of my perspective remain influenced by the dominant paradigms of the field, which generally have strong western, white, colonial/extractive, heteronormative, and patriarchal biases; my identity as a mixed-race, female-presenting, first-generation scholar has likely sensitized me to these influences. My work seeking change in the field (e.g. to our notions of novelty, representation, intelligence, rigor, or appropriate scientific practice) has been mostly internally reform-oriented.

Christoph: I am a white central European immigrant settler in Toronto, Canada, with an invisible disability. This identity has afforded me an odd juxtaposition of experiencing how privilege and oppression can intersect in our societies and has sensitized me over the years to theories and frameworks that help us make sense of these issues, including the intersectional feminist theories that shape this positionality statement. My education was in computer science and business informatics, but I am now a professor of information and am interacting closely with fields outside of computer science. My first published paper applied self-organizing maps to software cost estimation, but my doctoral and postdoctoral work developed computing and decision analysis frameworks for large-scale digital curation problems in libraries and archives. I became interested in long-term perspectives of environmental sustainability and the broader implications of technology on social justice partially because of the longer-term perspectives offered by ideals of stewardship and archiving. In the past decade I have grappled with the myths and illusions that are common in computing and that I too inherited via my education - including the flawed idea that technology is or can be neutral and that people's minds are information processors - and have built 'critical friendships' with other fields that can help computing researchers, educators and practitioners see beyond these conventional horizons and make better sense of the role of computing in our societies in order to help reorient computing for environmental sustainability and social justice. This negotiation between different fields and disciplines can be uncomfortable especially when it implies critique. In this project I aim to be pragmatic in building bridges between the conceptual spaces of these fields and to translate valuable insights from the stewardship perspective of digital curation into practical guidelines that ML can adopt and use.

Eshta: As a brown, female PhD student, I feel a lot of privilege in the opportunity to be a researcher where so many like me have not had the same opportunities, spaces, and circumstances. While I am a minority in these social identities, I am not in others and that is likely what has led to a position of privilege that I acknowledge. I am therefore also in a position where I can discuss and reflect on the impacts of my identities and worldviews on how and what I research. My academic background is rooted in information technology and data science, but my doctoral work is within a sociotechnical space. My perspective has thus shifted from seeing technology as a neutral entity, a tool that must be used and advanced further, to recognition of the need for a critical approach to technology that takes an ethics and justice based lens. The exposure to different worldviews that I have gained during my doctoral research has altered how I perceive the use of technology. For this project, I have pursued balancing a technical and

critical perspective so that the adoption of data curation that we recommend has practical benefits and uptake within the ML community.

Harshit: HG brings a perspective informed by his background in data science and information technology, shaped by his studies in South Asia. He is attuned to both the potential and the challenges of scaling technology in contexts vulnerable to climate change. In his work at the Department of Computer Science, HG evaluates how technological advancements can be balanced against their environmental impacts, striving to produce evidence that supports stringent environmental standards through a public health lens. His position is situated at the intersections between machine learning, environmental policy, and health outcomes.

Reyna: I am an Asian female PhD student. My primary academic training is in computer science and statistics, with a focus on climate informatics. Much of my academic research uses empirical methodologies to build machine learning and statistical models on real-world datasets, such as climate and scRNA-seq data. Throughout my PhD journey, I have recognized the importance of adopting a critical approach that incorporates ethics, responsibility, and sustainability in technology. This understanding is driven by my passion for studying climate change and the potential and limitations of machine learning to address it. I am particularly committed to promoting explainable AI, ethical AI, and responsible AI for climate change. Through my doctoral research, I have embraced a multifaceted and inclusive approach, striving to balance technical excellence with ethical considerations. This includes advocating for the responsible use of machine learning to tackle climate change emphasizing the importance of sustainability and social justice in my work.

Ciara: I am a PhD candidate at a faculty of information, where my primary research focus is on how people work with cross-domain data. I also work as a data scientist with the federal government, where I develop AI/ML governance and support the implementation of AI/ML projects. In my academic research, I draw on data and information practice scholarship, intersectional feminist and queer data studies, and interdisciplinary studies. My commitments to interdisciplinarity and feminist perspectives are influenced by my positions as a mixed-race woman, first generation university student, and settler in Canada. These commitments mean that I approached this project centering the situatedness of the rubric (that is: we are proposing a set of data curation best practices for ML rather than implying a sole universal 'right' way to do data work) and paying attention to the ways in which subject matter domain might have impacted the documentation of the datasets we evaluated.

Note: The fact that these statements above are all slightly different is an expression of their *authenticity* - we refrained from homogenizing them to fit a common structure, choosing instead to leave our individual voices here and present additional context.

Our perspectives interacted most directly in the making of the rubric. In our previous work, we discussed how disagreements in evaluations occurred because of differences in perspectives and areas of knowledge, including overlapping technologies, negotiating the depth of data curation expertise needed to apply the rubric, and challenges in scoping the extent of documentation dataset creators are responsible for [8]. These disagreements were deliberated to reach a balance between points of view. Particularly, it is the different points of view that enabled the rubric to take a shape in which both data curation and ML concepts were harmonized. When assembling the team, the senior researchers intentionally selected candidates with diverse perspectives. The composition of this team proved well-suited for doing the translation and operationalizing process of data curation for ML. Simply having two varying perspectives on the same dataset and the same criterion helped create a more nuanced view, and this triangulation of perspectives often enriched the debate and reflection on data practices. Other teams with diverse compositions with an interest in ML data practices as well as interdisciplinarity complementarity may produce different assessment results, but a combination of perspectives will be valuable. We thus recommend having diversity and interdisciplinary complementarity in the application of the rubric.

We did not aim to neutralize the specifics of each individual viewpoint in a supposedly 'neutral' rubric - something that is never feasible - but instead sought *consistency* in the evaluation process. It is important to overcome misleading ideals of curation as neutral. As with other technical work, the neutral stance is illusory. Fields concerned with curation have also grappled with the realization that they cannot be neutral at all [111]. As a consequence, "current archival thought now recognizes and explores the implications of the subjective and inherently political nature of archival processes" such as appraisal: the decision what to keep and what to discard [122:162].

Our work in developing this framework extends a 'critical friendship' [14] from data curation towards machine learning. While machine learning already performs curation, it does so without adopting the field's standards which can aid in advancing the state of the art. Our aims were to clearly communicate knowledge from the data curation literature and communicate with machine learning researchers, in order to normatively encourage better practice.

Table 3: Author contributions

Author	Contributions
Eshta Bhardwaj (she/her)	Her contributions for this project include the conceptualization of the framework, aiding with project administration, developing the methodology of conducting the evaluations, conducting evaluations, analyzing and visualizing the results, and writing, editing, and reviewing all drafts of the published work.
Harshit Gujral (he/him)	His contributions to this paper include the development and iteration of the evaluations, conducting evaluations, its writing, particularly methods, and results, and a discussion about reporting the environmental footprint of machine learning.
Siyi Wu (she/her)	Her contributions to this paper include development and iteration of evaluations, conducting evaluations, writing, review and editing.
Ciara Zogheib (she/her)	Her contributions include conducting iterative evaluations of dataset documentation, and writing (results section, positionality) and edits of the final manuscript.
Tegan Maharaj (they/them)	Their contributions to this paper include conceptualization, funding, methodology, comments on iterations of the rubric, writing, review and editing.
Christoph Becker (he/him)	His contributions include conceptualization, resources, funding, methodology, supervision, validation, writing, reviewing and editing.

D.5 How to report environmental footprint

We recommend the following strategies to quantify the environmental footprint of dataset development:

1. Carbon Footprint Estimation Tools: Tools like LLMCarbon, which has been designed to provide end-to-end carbon footprint estimations for large language models, offer a valuable resource for dataset creators in ML [78]. These tools allow for the prediction of carbon outputs based on diverse parameters such as hardware use, model architecture, and operational practices before the actual computational tasks begin.
2. Efficient Data Management Strategies: Research shows that end-to-end utilization of each GB stored in a data center is associated with approximately 5.12 kWh of energy consumption [16]. Reducing data redundancy and implementing data pruning techniques can decrease the volume of data that needs to be stored and processed, subsequently reducing the energy consumed during these stages. Several researchers have called for using end-to-end efficiency as an evaluation criterion for publishing ML research on computationally intensive models besides accuracy and related measures [43, 106, 124, 131].
3. Standardize Carbon Reporting: Developing a standardized protocol for reporting the carbon emissions of ML projects, as suggested by existing research [65, 78, 80, 106], would facilitate greater transparency and accountability within the industry. For the very least, researchers can gather a rough estimate of the electricity consumption of their ML projects and can get an estimate of corresponding carbon dioxide equivalent emissions using tools like the ML Emissions Calculator [65] and Green Algorithms tool [66]. This could also involve detailed reporting of energy sources, hardware specifications, and operational efficiencies.
4. Large but Sparsely Activated Networks: The previous research discusses the energy efficiency of using large but sparsely activated deep neural networks (DNNs), which can consume less than one-tenth the energy of large, densely activated DNNs without sacrificing accuracy [106]. For data processing, training, and deployment, the paper also emphasizes the impact of choosing energy-efficient data centers and hardware, along with strategically choosing data centers at a geographic location with a high renewable energy mix in the electricity grid [106].
5. Life cycle assessment (LCA) approach: For large language models conducting a Life Cycle Assessment (LCA) is recommended to quantify the carbon footprint across all stages of a language model's life cycle, from

equipment manufacturing to operational use and beyond [79]. This methodology, as discussed in the BLOOM model analysis [79], includes the energy consumed during model training and the emissions during model deployment and inference.

We hope these recommendations provide a much-needed starting point for the ML community to begin quantifying the environmental footprint of their projects. However, reporting alone will not mitigate the environmental impacts of data curation in ML. Existing research encourages the ML community to engage in efficient data management strategies, optimize model architectures, and adopt green computing practices like selecting data centers that use renewable energy [43, 106, 124, 131]. Research suggests that while efficiency improvements are crucial, they are not always sufficient on their own to reduce overall carbon emissions due to these rebound effects [73, 128].

Rebound effects manifest when improvements in computational efficiency lead to an increased use of ML technologies, as lower operational costs and enhanced capabilities encourage more frequent training and deployment of larger models, potentially increasing total energy consumption [12, 15, 128]. Historical data in sectors like automotive and residential energy use demonstrate that efficiency gains often lead to increased consumption, as savings are reinvested into more or expanded use of the technology, rather than resulting in a net decrease in energy use [12, 52, 134]. This can be exacerbated by indirect rebound effects that occur when the application of energy-efficient ML technologies in various sectors leads to broader and more intensive use of these technologies, subsequently increasing energy demand across those sectors, despite individual efficiency improvements.

In addition to energy efficiency-based measures, we call for the inclusion of digital sufficiency-based measures to mitigate potential rebound effects [73, 119]. These measures include limiting the growth of computational demands by setting strict computational budgets that reflect actual needs rather than maximum capacities. Additionally, it involves the creation of algorithms designed to perform effectively with minimal energy use, emphasizing necessity over excess. Regulatory measures are also critical, aimed at enforcing sustainable practices across the digital and computing sectors to ensure that efficiency improvements result in genuine reductions in carbon emissions. These mitigation strategies integrate sufficiency with technological innovation, ensuring that the advancement of ML contributes positively to environmental sustainability.

E. Changes to Rubric and Toolkit

Table 4: List of Changes to Rubric and Toolkit

Location of Change	Description and Rationale
Toolkit: Application Guidance	The evaluation of the minimum standard of documentation was updated. It previously stated that a pass is granted for *any* amount or type of discussion around the element and a fail is granted *only* if there is no discussion around the element at all. In the new version of the toolkit, it states that a pass is granted if all aspects specified under the minimum standard were discussed and a fail if they were only partially discussed or not discussed at all. Based on the changes made to criteria of the rubric elements, the minimum standard could only be achieved if all criteria were fulfilled (not just one or few).
Toolkit: How to interpret authenticity, reliability, and representativeness	An additional example of how to interpret authenticity and reliability was added to aid in better understanding of the criteria of the elements.
Toolkit: Sample evaluations	Both sample evaluations were updated. Sample evaluations were updated to reflect the completion of a more recent version of the rubric.
Rubric: Context, purpose, motivation	The criteria for the standard of excellence were made more explicit so that evaluators would have similar interpretations. Previously it stated, “documentation explains how dataset can be reused beyond its original context”. The current version expects documentation to discuss whether and how reuse is possible.
Rubric: Requirements	The criteria for the standard of excellence were simplified, and the requirement to “state different approaches in formulating the problem apart from the final presented plan” was removed.

	The criterion was simplified because we recognized that documenting alternative approaches was not an essential characteristic of excellence: there are good reasons not to document paths not chosen.
Rubric: Context awareness	The criterion for the minimum standard was simplified by moving the requirement for a "reflection on the dataset creators' awareness of social, political, and historical context" to the standard of excellence criteria for 'context, purpose, motivation'. The criterion was moved for clarity: the reflection on context ties in with purpose and motivation, while the 'context awareness' dimension focuses on positionality and reflexivity.
Rubric: Domain knowledge and data practices	The phrasing for both criteria were updated. The phrasing conveys similar criteria as the original, but is more explicitly stated to enable more consistent evaluation.
Rubric: Data collection	The criteria for both minimum standard and standard of excellence were updated to include multiple types of data. The criteria were augmented to address both collected data and synthetic data because it originally did not ask for documentation regarding how data was synthesized and whether that introduces intrinsic biases.
Rubric: Data annotation	The criteria for both minimum standard and standard of excellence were updated to include the assessment of labels obtained from multiple sources. The previous criteria did not have differing criteria based on how the labels were obtained/derived.
Rubric: Authenticity	We combined the criteria for authenticity to include integrity based on [55] to simplify the rubric.
Rubric: Reliability	Based on feedback from the reviewers, we added a clarifying statement to make the criteria for the minimum standard clearer. We also changed the criteria for the standard of excellence to clarify how it can be evaluated.
Rubric: Structured documentation	We slightly updated the phrasing to only allow context documents to have established structures rather than formats that are not established and therefore may be difficult to evaluate consistently.
Rubric: Reusability	Minor changes to the phrasing were made to the minimum standard to improve clarity.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B. Moeslund. 2021. RELLISUR: A Real Low-Light Image Super-Resolution Dataset. August 29, 2021. Advances in Neural Information Processing Systems.

[2] Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Lukas Jarosch, Dan Berenberg, Ian Fisk, Andrew Martin Watkins, Stephen Ra, Richard Bonneau, and Mohammed AlQuraishi. 2023. OpenProteinSet: Training data for structural biology at scale. 2023. Advances in Neural Information Processing Systems.

[3] 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*), 2011. Association for Computing Machinery, New York, NY, USA, 675–684. https://doi.org/10.1145/1978942.1979041

[4] Björn Barz and Joachim Denzler. 2021. WikiChurches: A Fine-Grained Dataset of Architectural Styles with Real-World Challenges. 2021. Advances in Neural Information Processing Systems.

[5] 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

[6] 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* (*FAccT '21*), March 01, 2021. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[7] June M. Besek and Philippa S. Loengard. 2007. Maintaining the Integrity of Digital Archives. *Colum. J.L. & Arts* 31, (2007), 267.
[8] 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 *2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. Association for Computing Machinery, New York, NY, USA, 1055–1067. https://doi.org/10.1145/3630106.3658955
[9] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. 2023. PUG: Photorealistic and Semantically Controllable Synthetic Data for Representation Learning. 2023. Advances in Neural Information Processing Systems.
[10] Christine L. Borgman. 2017. *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.
[11] Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–27. https://doi.org/10.1145/3479582
[12] Christina Bremer, Harshit Gujral, Michelle Lin, Lily Hinkers, Christoph Becker, and Vlad C Coroamă. 2023. How Viable are Energy Savings in Smart Homes? A Call to Embrace Rebound Effects in Sustainable HCI. *ACM Journal on Computing and Sustainable Societies* 1, 1 (2023), 1–24.
[13] 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
[14] Christoph Becker. 2023. Computing's Critical Friends. In *Insolvent: How to Reorient Computing for Just Sustainability*. MIT Press, 399. Retrieved June 8, 2024 from https://direct.mit.edu/books/oa-monograph/5594/chapter/4218561/Computing-s-Critical-Friends
[15] Vlad C Coroamă and Friedemann Mattern. 2019. Digital rebound–why digitalization will not redeem us our environmental sins. In *Proceedings 6th international conference on ICT for sustainability. Lappeenranta. http://ceur-ws. org*, 2019. .
[16] David Costenaro and Anthony Duer. 2012. The megawatts behind your megabytes: going from data-center to desktop. *Proceedings of the 2012 ACEEE Summer Study on Energy Efficiency in Buildings, ACEEE, Washington* (2012), 13–65.
[17] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. 2023. Advances in Neural Information Processing Systems.
[18] 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. 2023. Advances in Neural Information Processing Systems.
[19] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment. 2022. Advances in Neural Information Processing Systems.
[20] Digital Curation Centre. What is digital curation? Retrieved from https://www.dcc.ac.uk/about/digital-curation
[21] Digital Curation Centre. Glossary. *Digital Curation Centre*. Retrieved January 21, 2024 from https://www.dcc.ac.uk/about/digital-curation/glossary
[22] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data Feminism*. MIT Press.
[23] Luciana Duranti. 1995. Reliability and Authenticity: The Concepts and Their Implications. *Archivaria* (May 1995), 5–10.
[24] Luciana Duranti. 1998. *Diplomatics: New Uses for an Old Science*. Scarecrow Press.
[25] 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
[26] Luciana Duranti. 2007. The InterPARES 2 Project (2002-2007): An Overview. *Archivaria* (2007), 113–121.
[27] Luciana Duranti and Heather MacNeil. 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project. *Archivaria* (October 1996), 46–67.
[28] Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 (January 2018), 2056305118763366. https://doi.org/10.1177/2056305118763366
[29] Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical Capabilities of ChatGPT. 2023. Advances in Neural Information Processing Systems.
[30] 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. November 02, 2023. Advances in Neural Information Processing Systems.

[31] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sebastien Giordano, and Boris Wattrelos. 2023. FLAIR : a Country-Scale Land Cover Semantic Segmentation Dataset From Multi-Source Optical Imagery. 2023. Advances in Neural Information Processing Systems.

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

[33] GO FAIR. I3: (Meta)data include qualified references to other (meta)data. *FAIR Principles*. Retrieved January 18, 2024 from https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/

[34] GO FAIR. F1: (Meta) data are assigned globally unique and persistent identifiers. *FAIR Principles*. Retrieved January 20, 2024 from https://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/

[35] GO FAIR. R1.2: (Meta)data are associated with detailed provenance. *FAIR Principles*. Retrieved January 20, 2024 from https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/

[36] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 Million Large-scale Chinese Cross-modal Pre-training Benchmark. 2022. Advances in Neural Information Processing Systems.

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

[38] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. https://doi.org/10.2307/3178066

[39] Sandra Harding. 1992. Rethinking Standpoint Epistemology: What Is "Strong Objectivity"? In *Feminist Epistemologies*. Routledge.

[40] Sandra G. Harding. 1986. *The Science Question in Feminism*. Cornell University Press.

[41] 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. November 02, 2023. Advances in Neural Information Processing Systems.

[42] 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (2022), 1–29. https://doi.org/10.1145/3555760

[43] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43.

[44] 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. August 29, 2021. Advances in Neural Information Processing Systems.

[45] Sarah Higgins. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3, 1 (2008), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

[46] Sarah Higgins. 2009. DCC DIFFUSE Standards Frameworks: A Standards Path through the Curation Lifecycle. *International Journal of Digital Curation* 4, 2 (October 2009), 60–67. https://doi.org/10.2218/ijdc.v4i2.93

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

[48] 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. August 29, 2021. Advances in Neural Information Processing Systems.

[49] 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. June 29, 2022. Advances in Neural Information Processing Systems.

[50] Liya Hu, Zhiang Dong, Jingyuan Chen, Guifeng Wang, Zhihua Wang, Zhou Zhao, and Fei Wu. 2023. PTADisc: A Cross-Course Dataset Supporting Personalized Learning in Cold-Start Scenarios. 2023. Advances in Neural Information Processing Systems.
[51] 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. August 29, 2021. Advances in Neural Information Processing Systems.
[52] Kent M Hymel and Kenneth A Small. 2015. The rebound effect for automobile travel: asymmetric response to price changes and novel features of the 2000s. *Energy Economics* 49, (2015), 93–103.
[53] Information and Privacy Commissioner of Ontario. Consent may be implied in some cases. *Information and Privacy Commissioner of Ontario*. Retrieved January 20, 2024 from https://www.ipc.on.ca/part-x-cyfsa/consent-and-capacity/elements-of-consent/consent-may-be-implied-in-some-cases/
[54] 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. June 29, 2022. Advances in Neural Information Processing Systems.
[55] Asen O. Ivanov. 2019. The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice. Ph.D. University of Toronto (Canada). Retrieved May 26, 2024 from https://www.proquest.com/docview/2281198823/abstract/11E3C0C910F94374PQ/1
[56] 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*, January 27, 2020. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829
[57] Julian Posada. 2023. *Platform Authority and Data Quality*. Retrieved from https://www.berggruen.org/ideas/articles/decoding-digital-authoritarianism/
[58] Julia Kaltenborn, Charlotte 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* 36, (2023), 21757–21792.
[59] Gaoussou Youssouf Kebe, Padraig Higgins, Patrick Jenkins, Kasra Darvish, Rishabh Sachdeva, Ryan Barron, John Winder, Donald Engel, Edward Raff, Francis Ferraro, and Cynthia Matuszek. 2021. A Spoken Language Dataset of Descriptions for Speech-Based Grounded Language Learning. 2021. Advances in Neural Information Processing Systems.
[60] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. (September 2022). https://doi.org/10.2139/ssrn.4217148
[61] Kim Martineau. 2021. What is synthetic data? *IBM Research Blog*. Retrieved January 21, 2024 from https://research.ibm.com/blog/what-is-synthetic-data
[62] 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 (December 2014), e115253. https://doi.org/10.1371/journal.pone.0115253
[63] 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. November 02, 2023. Advances in Neural Information Processing Systems.
[64] Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. 2023. Advances in Neural Information Processing Systems.
[65] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700* (2019).
[66] Loic Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science* 8, 12 (2021), 2100707.
[67] Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating Out-of-Distribution Performance on Document Image Classifiers. 2022. Advances in Neural Information Processing Systems.
[68] 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*, July 21, 2021. ACM, Virtual Event USA, 695–703. Retrieved November 11, 2022 from https://dl.acm.org/doi/10.1145/3461702.3462598
[69] Brent Lee. 2005. Authenticity, Accuracy and Reliability: Reconciling Arts-related and Archival Literature. 2005. . Retrieved May 26, 2024 from https://www.semanticscholar.org/paper/Authenticity%2C-Accuracy-and-Reliability%3A-Reconciling-Lee/b96f00340d39bb0b971f1e6261358ee50abd1565

[70] 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. 2023. Advances in Neural Information Processing Systems.
[71] Ramona Leenings, Nils R. Winter, Udo Dannlowski, and Tim Hahn. 2022. Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage* 257, (August 2022), 119298. https://doi.org/10.1016/j.neuroimage.2022.119298
[72] Calvin Liang. 2021. Reflexivity, positionality, and disclosure in HCI. *Medium*. Retrieved January 20, 2024 from https://medium.com/@caliang/reflexivity-positionality-and-disclosure-in-hci-3d95007e9916
[73] Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, and Jacques Combaz. 2022. Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability* 14, 9 (2022), 5172.
[74] 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. *Sci Data* 7, 1 (May 2020), 144. https://doi.org/10.1038/s41597-020-0486-7
[75] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. BiToD: A Bilingual Multi-Domain Dataset For Task-Oriented Dialogue Modeling. 2021. Advances in Neural Information Processing Systems.
[76] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. 2023. Advances in Neural Information Processing Systems.
[77] 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*, June 21, 2022. ACM, Seoul Republic of Korea, 1136–1153. https://doi.org/10.1145/3531146.3533175
[78] Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476* (2023).
[79] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research* 24, 253 (2023), 1–15.
[80] Sasha Luccioni, Bruna Trevelin, and Margaret Mitchell. 2024. The Environmental Impacts of AI - Primer. *Hugging Face Blog*. Retrieved from https://huggingface.co/blog/sasha/ai-environment-primer
[81] Luciana Duranti and Randy Preston (Eds.). 2009. International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records. *Records Management Journal* 19, 1 (2009). https://doi.org/10.1108/rmj.2009.28119aae.003
[82] Loren Lugosch, Piyush Papreja, Mirco Ravanelli, Abdelwahab Heba, and Titouan Parcollet. 2021. Timers and Such: A Practical Benchmark for Spoken Language Understanding with Numbers. 2021. Advances in Neural Information Processing Systems.
[83] 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. June 29, 2022. Advances in Neural Information Processing Systems.
[84] H. MacNeil. 2013. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. Springer Science & Business Media.
[85] 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*, 2020. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376445
[86] Andrey Malinin, Neil Band, Yarin Gal, Mark Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. 2021. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. 2021. Advances in Neural Information Processing Systems.
[87] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. 2022. Change Event Dataset for Discovery from Spatio-temporal Remote Sensing Imagery. June 29, 2022. Advances in Neural Information Processing Systems.

[88] Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. 2022. CLEVRER-Humans: Describing Physical and Causal Events the Human Way. 2022. Advances in Neural Information Processing Systems.
[89] Matthew Stewart. 2023. The Olympics of AI: Benchmarking Machine Learning Systems. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b
[90] 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. June 29, 2022. Advances in Neural Information Processing Systems.
[91] Daniel McDuff, Miah Wander, Xin Liu, Brian L. Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. 2022. SCAMPS: Synthetics for Camera Measurement of Physiological Signals. 2022. Advances in Neural Information Processing Systems.
[92] Alison McIntyre. 2023. Doctrine of Double Effect. In *The Stanford Encyclopedia of Philosophy* (Winter 2023), Edward N. Zalta and Uri Nodelman (eds.). Metaphysics Research Lab, Stanford University. Retrieved January 20, 2024 from https://plato.stanford.edu/archives/win2023/entries/double-effect/
[93] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proc. ACM Hum.-Comput. Interact.* 6, GROUP (2022), 1–14. https://doi.org/10.1145/3492853
[94] Gerald Midgley. 2000. *Systemic intervention: philosophy, methodology, and practice*. Springer, New York.
[95] Reagan Moore. 2008. Towards a Theory of Digital Preservation. *International Journal of Digital Curation* 3, 1 (August 2008), 63–75. https://doi.org/10.2218/ijdc.v3i1.42
[96] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. August 29, 2021. Advances in Neural Information Processing Systems.
[97] Soukayna Mouatadid, Paulo Orenstein, Genevieve Elaine Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Edward Knight, Maria Geogdzhayeva, Samuel James Levang, Ernest Fraenkel, and Lester Mackey. 2023. SubseasonalClimateUSA: A Dataset for Subseasonal Forecasting and Benchmarking. 2023. Advances in Neural Information Processing Systems.
[98] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *CHI Conference on Human Factors in Computing Systems*, 2022. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517644
[99] 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*, May 06, 2021. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445402
[100] Gokul Nc, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh M. Khapra. 2022. Addressing Resource Scarcity across Sign Languages with Multilingual Pretraining and Unified-Vocabulary Datasets. 2022. Advances in Neural Information Processing Systems.
[101] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
[102] Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. 2021. Advances in Neural Information Processing Systems.
[103] 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), 16.
[104] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. 2023. RenderMe-360: A Large Digital Asset Library and Benchmarks Towards High-fidelity Head Avatars. 2023. Advances in Neural Information Processing Systems.
[105] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19*), 2019. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567
[106] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).

[107] 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 (November 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[108] 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. November 02, 2023. Advances in Neural Information Processing Systems.

[109] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. 2021. Advances in Neural Information Processing Systems.

[110] Alex H. Poole. 2015. How has your science data grown? Digital curation and the human factor: a critical literature review. *Arch Sci* 15, 2 (June 2015), 101–139. https://doi.org/10.1007/s10502-014-9236-y

[111] Ricardo L. Punzalan and Michelle Caswell. 2016. Critical Directions for Archival Approaches to Social Justice. *The Library Quarterly* 86, 1 (January 2016), 25–42. https://doi.org/10.1086/684145

[112] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. 2021. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. 2021. Advances in Neural Information Processing Systems.

[113] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. 2022. Advances in Neural Information Processing Systems.

[114] Piera Riccio, Bill Psomas, Francesco Galati, Francisco Escolano, Thomas Hofmann, and Nuria M. Oliver. 2022. OpenFilter: A Framework to Democratize Research Access to Social Media AR Filters. 2022. Advances in Neural Information Processing Systems.

[115] Nataniel Ruiz. 2019. Learning to Simulate. *Medium*. Retrieved January 21, 2024 from https://towardsdatascience.com/learning-to-simulate-c53d8b393a56

[116] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. August 29, 2021. Advances in Neural Information Processing Systems.

[117] 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*, 2021. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445518

[118] Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. 2022. Ambiguous Images With Human Judgments for Robust Visual Event Classification. 2022. Advances in Neural Information Processing Systems.

[119] Tilman Santarius, Jan CT Bieser, Vivian Frick, Mattias Höjer, Maike Gossen, Lorenz M Hilty, Eva Kern, Johanna Pohl, Friederike Rohde, and Steffen Lange. 2023. Digital sufficiency: conceptual considerations for ICTs on a finite planet. *Annals of Telecommunications* 78, 5 (2023), 277–295.

[120] Sanja Scepanovic, Ivica Obadic, Sagar Joglekar, Laura Giustarini, Cristiano Nattero, Daniele Quercia, and Xiao Xiang Zhu. 2023. MedSat: A Public Health Dataset for England Featuring Medical Prescriptions and Satellite Imagery. 2023. Advances in Neural Information Processing Systems.

[121] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (October 2021), 1–37. https://doi.org/10.1145/3476058

[122] Sarita Schoenebeck and Paul Conway. 2020. Data and Power: Archival Appraisal Theory as a Framework for Data Preservation. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (October 2020), 162:1-162:18. https://doi.org/10.1145/3415233

[123] Tal Schuster, Ashwin Kalyan, Alex Polozov, and Adam Tauman Kalai. 2021. Programming Puzzles. June 08, 2021. Advances in Neural Information Processing Systems.

[124] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM* (2020). Retrieved from https://cacm.acm.org/research/green-ai

[125] Raesetje Sefala, Timnit Gebru, Luzango Mfupe, Nyalleng Moorosi, and Richard Klein. 2021. Constructing a Visual Dataset to Study the Effects of Spatial Apartheid in South Africa. 2021. Advances in Neural Information Processing Systems.

[126] Omar Shouman, Wassim Gabriel, Victor-George Giurcoiu, Vitor Sternlicht, and Mathias Wilhelm. 2022. PROSPECT: Labeled Tandem Mass Spectrometry Dataset for Machine Learning in Proteomics. 2022. Advances in Neural Information Processing Systems.

[127] Moein Sorkhei, Yue Liu, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, and Kevin Smith. 2021. CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer. 2021. Advances in Neural Information Processing Systems.

[128] Steve Sorrell and John Dimitropoulos. 2008. The rebound effect: Microeconomic definitions, limitations and extensions. *Ecological Economics* 65, 3 (2008), 636–649.

[129] 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. 2021. Advances in Neural Information Processing Systems.

[130] Adam J. Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. 2023. SSL4EO-L: Datasets and Foundation Models for Landsat Imagery. 2023. Advances in Neural Information Processing Systems.

[131] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[132] Paige L. Sweet. 2020. Who Knows? Reflexivity in Feminist Standpoint Theory and Bourdieu. *Gender & Society* 34, 6 (December 2020), 922–950. https://doi.org/10.1177/0891243220966600

[133] Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. 2021. KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects. 2021. Advances in Neural Information Processing Systems.

[134] Brinda A Thomas, Zeke Hausfather, and Inês L Azevedo. 2014. Comparing the magnitude of simulated residential rebound effects from electric end-use efficiency across the US. *Environmental Research Letters* 9, 7 (2014), 074010.

[135] 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (2022), 414:1-414:29. https://doi.org/10.1145/3555139

[136] Nolan Wagener, Andrey Kolobov, Felipe Vieira Frujeri, Ricky Loynd, Ching-An Cheng, and Matthew Hausknecht. 2022. MoCapAct: A Multi-Task Dataset for Simulated Humanoid Control. 2022. Advances in Neural Information Processing Systems.

[137] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. August 29, 2021. Advances in Neural Information Processing Systems.

[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. *Sci Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[139] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. 2021. Advances in Neural Information Processing Systems.

[140] Tong Xia, Dimitris Spathis, Chlo{\"e} Brown, J. Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, and Cecilia Mascolo. 2021. COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening. 2021. Advances in Neural Information Processing Systems.

[141] 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. June 29, 2022. Advances in Neural Information Processing Systems.

[142] Huang Xuanwen, Yang Yang, Wang Yang, Wang Chunping, Zhang Zhisheng, Xu Jiarong, Chen Lei, and Vazirgiannis Michalis. 2022. DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. .

[143] Zenodo - Research. Shared. FAIR Principles. Retrieved January 18, 2024 from https://about.zenodo.org/principles/

[144] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. 2023. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. 2023. Advances in Neural Information Processing Systems.