MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] The dataset was created to better understand

the dynamics of dataset usage

in machine learning research, both

within task communities and within Machine Learning Research (MLR)

overall. The paper it was created for

addresses three research questions:

1. How concentrated is benchmark

usage?

2. Where do benchmarks come

from?

3. Which institutions are creating

widely-used benchmarks?

The dataset addresses a gap in being

a field-level survey of benchmark

dataset usage across MLR. Other work

at the time of publication has largely

used ethnographic or audit methods.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] The dataset was created purely for academic

research. It was created by researchers

at the UCLA Department

of Sociology (Koch, Foster) and the

Google Responsible & Ethical AI team

(Denton, Hanna).

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] At the time of research, Bernard Koch

was supported by an US NSF Graduate

Research Fellowship and Doctoral

Dissertation Research Improvement

Grant. The DDRIG grant was

given explicitly to work on this project.

All authors were supported by their

respective employers, and Foster was

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

[answer5] There are multiple types of instances that are represented in the dataset:

1. Papers: The raw data consists of papers in MLR that have been annotated with datasets and/or specific tasks. We consider all papers that have been annotated with a dataset in our raw data source, paperswithcode. com (PWC).

2. Datasets: We considered all datasets catalogued in PWC at the time of retrieval.

3. Tasks: We considered all tasks in MLR catalogued in PWC's benchmark archive that annotated a dataset or datasetintroducing paper.

4. Usages: Dataset usages is a derivative instance of datasetusing papers because many papers use multiple datasets.

5. Institutions: Lastly we consider the affiliation of the last author of each dataset-introducing

paper (if the paper and institution
are in the Microsoft Academic
Graph).

[question6] How many instances are there in total (of each type, if appropriate)?

[answer6] This answer is complicated by
the significant cleaning steps idiosyncratic
to each analysis (described
below). The raw data consist
of 4,384 datasets and datasetusing
60,647 papers downloaded from
PWC. Below we produce the table
in the paper of the actual counts for
each of the analyses after cleaning. Additionally, there were 442 institutions
in Analysis 3.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] The dataset should be considered a
sample of all datasets used in MLR
because PWC is primarily curated by
crowd-sourcing and scraping ArXiv.
To estimate the extent of coverage
bias and recency bias in PWC, we
searched PWC for all papers published
in ten top ML conferences (NeurIPS,
ICML, ICLR, ACL, AAAI, AISTATS,
KDD, CVPR, SIGIR, IJCAI) between
2015 and 2020 (46,774 papers according
to Microsoft Academic). We found
that 58.9% of these papers appeared
in PWC. However these 58.9% of papers
accounted for 89.3% of collective
citations received by the 46,774 papers,
suggesting that the missing papers
are primarily lower impact papers.

When we disaggregated this analysis by year, we find that coverage ranges from 38.9% of 2015 papers to 68.8% of 2020 papers, but the proportion of citations covered in each year from 2015 on never drops below 86%. While these numbers suggest that omitted papers may be less cited and thus unlikely to propose widely used datasets, we note that they do not address possible under-annotation of dataset creations or usages within included PWC papers. In addition, we also introduce several additional sampling biases by design due to resource constraints and to ensure the quality of our data. These details are discussed extensively in the paper's appendix.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] The raw dataset is available at https://github.com/paperswithcode/paperswithcode-data. The fields vary extensively with the different instance types. Fields for PWC papers include the paper URL, ArXiv ID, title, abstract, PDF url, journal/proceedings, authors, tasks, date, and methods. Dataset instances contain the URL, short name, full name, homepage, a description, affiliated paper, introduction date, any ethic warnings, the modality of the dataset (e.g., text, images, video), associated tasks, language, variants, and number of citing papers. All other instances are essentially derivative of these types except for links for dataset-introducing

papers to Microsoft Academic Graph.
With these links, extensive data about
the paper and it's authors are available
online or through the API. Details
of the graph schema are available
at https://docs.microsoft.com/
en-us/academic-services/graph/
reference-data-schema.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Not applicable. This is a social science dataset, not a supervised learning dataset.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] The PWC data is messy in that there
are datasets and/or papers that are not
associated with papers or tasks (see below
for more details). Whether this
forces us to drop this data is dependent
on the constraints of the analysis.
That being said our most significant
omission was dropping a number
of lightly used datasets because it
was too time-intensive to label them
with tasks. Starting with the raw data,
we scraped 92,874 dataset usages from
46,697 dataset-using PWC papers labeled
with tasks. Only 49,589 (53%) of
those usages were to datasets already
labeled with tasks in PWC.We first labeled
the 45 highest-used datasets with
their original tasks, skipping datasets
that did not seem designed for MLR
or where origin tasks were unidentifiable
from language used in the paper
or website. By manually labeling
the 90 largest datasets with tasks, we

recovered 33,739 usages, leaving just
10.2% of total dataset usages unlabeled
with tasks across 550 datasets. It was
too time-intensive for us to label these
last 550 datasets. However, our results
do not change when we include only
45 manually annotated datasets versus
90.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] All relationships between datasets, usages,
papers, and institutions are made
explicit through columns in the data.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] Not applicable. This is not a supervised
learning dataset.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] Yes, we discuss these limitation in
full in the Appendix of the paper, so
I won't reproduce them here.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] The dataset links to PWC and the
Microsoft Academic Graph. Unfortunately,
neither resource is archived but
we provide the raw data used from each. The PWC data is regenerated
daily and Microsoft Academic Graph
was/has been taken offline as of December

31, 2021. There are no restrictions
on using either data source for
academic research

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
[answer15] No.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16] No.

If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] YES

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19]

[question20] Any other comments?
[answer20]

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] We emailed the PWC maintainers
at the beginning of this project and
learned that PWC papers are annotated
with tasks and datasets through
a combination of community crowdsourcing
and keyword searching within
ArXiV papers. We do not believe this
annotation process is made explicit on
the website, but interested parties may
be able to contact PWC for more information.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] Most of the PWC data is
made publicly available at https:
//github.com/paperswithcode/
paperswithcode-data. Dataset usages
were scraped using an internal
API (see README and code). MAG
data was retrieved using the MAG API
(not online as of December 31, 2021).

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] Sampling is described above. Additional
sampling biases induced for each
analysis are described in more detail
within the "Data" section of the paper
and the appendix.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] Only authors

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] June to July 2021.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer26] No. All of the data used here is publicly

available and does not involve or identify human subjects.

If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipB] YES
[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27]
[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer28]
[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32]

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so,

please provide a description. If not, you may skip the remaining questions in this section.

[answer33] Please read the "Data" section and
appendix of the paper first since this
description is supplementary of that
text. Extensive preprocessing was
done on the dataset. This was largely
described in the paper but we describe
data curation in more detail. To learn
more, please see the code in Main-
Data.ipynb. First we use the PWCprovided
file "evaluation-tables.json"
to reconstruct a task ontology of parent
and child tasks. We note that
PWC has since made the ontology editable
directly, so looking at the benchmarking
tables may not be the most
up to date. We then go through the
"papers-with-abstracts.json" and label
all papers with tasks. We also incorporate
90-dataset introducing papers
into these papers that were manually
labeled with tasks. Next we merge
PWC's "datasets.json" file with these
task-labeled papers, maintaining only
papers that use a dataset in PWC.
To better understand what tasks the
paper was originally introduced for,
we only keep 2,673 datasets out of 4,384 that were actually introduced in
a paper. Finally we merge in the file
"dataset$_c$iting$_p$apers.txt"thatwescrapedusingPWC$'$ sinternalAPI, tellinguswhichofpaper$'$ Analysis 2
Cleaning In the next
stage, we look at dataset transfers between
"origin" tasks (i.e., tasks associated
with the dataset-introducing paper)
and "destination" tasks (i.e., tasks
associated with the dataset using paper.
This process is somewhat complicated

because both origin and destination papers will be labeled with multiple tasks, and we wanted to focus on higher-level tasks that were parent to another task (see paper). To this end, we impose the following restrictions:

1. A transfer cannot be from one origin tasks to another.

2. An origin does not transfer to a destination's parents, only origin parents can transfer to destination parents.

3. A parent task cannot transfer to it's own children (even if those children are themselves parent tasks).

To weed out noisy algorithmic annotations, we also discard any usages by papers that do not share at least one PWC task in common with the dataset task annotations. We further filtered the data by only considering the 50% largest tasks (133) with more than 34 dataset usages. When dissagregating over time to study time trends, we also dropped any task year with fewer than 10 usages to reduce noise in our analyses. Analysis 1 Cleaning Analysis 1 differs from Analysis 2 in that we do not constrain ourselves to datasets introduced in papers (4,384 datasets instead of 2,673). We do however discard any usages by papers that do not share at least one PWC task in common with the dataset task annotations. For comparability, we further restrict ourselves to the same 133 tasks used in Analysis 1. When dissagregating over time to study time trends, we again dropped any task year with fewer than 10 usages.

Analysis 3 Cleaning Analysis 3
takes the final dataset from Analysis
1. However it is further constrained
to usages of datasets introduced in papers
that appear in Microsoft Academic
Graph. Moreover, the last author
of the dataset-introducing paper
must have an affiliation annotation in
MAG.
Figure 4 This is not a cleaning
decision per se, but we note that the
"Other" category varies in size for the
three pie charts for legibility.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] Yes, the raw data should also be available
in the GitHub. See the README
for details.

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35]

[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37]

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
[answer38]

[question39] What (other) tasks could the dataset be used for?
[answer39] The dataset could be used for other
science of science analyses relating
to dataset usage, benchmarking practices,
or inequality within MLR. That
being said, we recommend that interested

readers consider collecting more

up to date data given how quickly the

field changes.

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] Yes, there are two issues. First, because

MLR is an extremely fast moving

field and could look very different

in two years, we would urge interested

parties to consider collecting

more timely data. Second, annotations

in PWC are likely to change and improve

over time. Third, MAG is no

longer online so those metadata annotations

of papers are no longer available.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] Given the low ethical risks of the data,

I cannot think of any. That being said

any research with policy implications

should consider the timeliness of the

data.

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] No.

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] The data will be available on GitHub.

[question45] When will the dataset be distributed?

[answer45] The data will be online by the end of

2021 (two weeks beyond the publication

deadline).

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] The dataset will be licensed under CC

BY-NC-SA. From Creative Commons,

"This license allows reusers to distribute,

remix, adapt, and build upon

the material in any medium or format

for noncommercial purposes only, and only so long as attribution is given

to the creator. If you remix, adapt,

or build upon the material, you must

license the modified material under

identical terms."

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] No.

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] No.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The dataset will be hosted at github.

com/kochbj/https://github.com/

kochbj/Reduced_Reused_Recycled.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] Raising a GitHub issue is preferred.

Emailing the first author also works.

[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52] Not as of now.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
[answer53] The dataset is unlikely to be updated

unless errors are discovered by interested

users. These updates will be

made visible on the front page of the

Github.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
[answer54] Not applicable.

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
[answer55] We do not anticipate maintaining multiple

versions of the dataset beyond the

original release. This will be stated on

the Github frontpage.

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
[answer56] Because of the CC BY-NC-SA, others

are free to use and adapt the dataset

as they see fit. Forking the Github

would probably be the best way to do

so. If other parties contribute significant

updates, we can discuss merging

them into the original GitHub.

[question57] Any other comments?
[answer57] No.