Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research

Bernard Koch
University of California, Los Angeles
bernardkoch@ucla.edu

Emily Denton
Google Research, New York
dentone@google.com

Alex Hanna
Google Research, Mountain View
alexhanna@google.com

Jacob Foster
University of California, Los Angeles
foster@soc.ucla.edu

Abstract

Benchmark datasets play a central role in the organization of machine learning research. They coordinate researchers around shared research problems and serve as a measure of progress towards shared goals. Despite the foundational role benchmarking practices play in the field, relatively little attention has been paid to the dynamics of benchmark dataset use and resuse within and across machine learning subcommunities. In this work we dig into these dynamics, by studying how dataset usage patterns differ across different machine learning subcommunities and across time from 2015-2020. We find increasing concentration on fewer and fewer datasets within task communities, significant adoption of datasets from other tasks, and concentration across the field on datasets that have been introduced by researchers situated within a small number of elite institutions. Our results have implications for scientific evaluation, AI ethics, and equity/access within the field.

1 Introduction

Datasets form the backbone of machine learning research (MLR). They are deeply integrated into work practices of machine learning researchers, operating as resources for training and testing machine learning models. Moreover, datasets serve a central role in the organization of MLR as a scientific field. Benchmark datasets establish stable points of comparison and coordinate scientists around shared research problems. Improved performance on these benchmarks is considered a key signal for collective progress; it is thus also an important form of scientific capital, sought after by individual researchers and used to evaluate and rank their contributions.

Datasets also exemplify machine learning tasks, typically through a collection of input and output pairs [1]. By institutionalizing benchmark datasets, task communities implicitly endorse these data as meaningful abstractions of a task or problem domain. The institutionalization of benchmarks influences the behavior of both researchers and end-users [2]. Because advancement on institutional benchmarks is viewed as an indicator of progress, researchers are encouraged to make design choices to maximize performance to gain credibility for their work. Institutionalization also signals to industry adopters that models can be expected to perform in the real world as they do on the benchmark datasets. The close alignment of datasets with "real world" tasks is thus critical not just to accurate measurement of collective scientific progress, but safe, ethical, and effective deployment of models in the wild.

Given their central role in the social and scientific organization of MLR, benchmark datasets have also become a central object of critical inquiry in recent years [3]. Dataset audits have revealed concerning biases that have direct implications for algorithmic bias and harms [4, 5, 6, 7]. Problematic

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

categorical schemas have been identified in popular image datasets, including poorly formulated categories and the inclusion of derogatory and offensive labels [8, 9]. Growing research into the disciplinary norms of dataset development have revealed concerning practices relating to dataset development and dissemination, such as unstandardized documentation and maintenance practices [10, 11, 12]. There is also growing concern regarding the limitations of existing datasets and standard metrics of evaluation for evaluating model behaviour in real-world settings and evaluating scientific progress in a problem domain [13, 14].

Despite the increase in critical attention to benchmark datasets, surprisingly little empirical attention has been paid to patterns of dataset use and reuse across the field as a whole. In this work we dig into these dynamics, by studying how dataset usage patterns differ across different machine learning subcommunities and across time from 2015-2020 in the Papers With Code (PWC) corpus.[1] More specifically, we study machine learning subcommunities that have formed around different machine learning tasks (e.g. *Sentiment Analysis* and *Face Recognition*) and examine: (i) the extent to which research within task communities is concentrated or distributed across different benchmark datasets, and (ii) patterns of dataset creation and movement between different task communites.

Overall, we find that the majority of papers within most tasks prefer datasets that were originally created for other tasks over ones created explicitly for their own task, even though most tasks have created more datasets than they have imported. Consistent with this finding, we see increasing concentration on fewer and fewer datasets within task communities. Lastly, we find that these dominant datasets have been introduced by researchers at just a handful of elite institutions.

The remainder of this paper is organized as follows. First, we motivate our research questions by underscoring the critical importance of benchmarks in coordinating machine learning research. Second, we describe our analyses on the PWC corpus, a catalog of datasets and their usage jointly curated manually by the machine learning community and algorithmically by Facebook AI Research. We then present our findings and discuss their implications for scientific validity, the ethical usage of MLR, and inequity within the field. We close by offering recommendations for possible reform efforts for the field.

2 The scientific, social, and ethical, importance of benchmark datasets

Following [1], we understand machine learning benchmarks as community resources against which models are evaluated and compared. Benchmarks typically formalize a particular task through a dataset and associated quantitative metric of evaluation. Benchmarking is the dominant paradigm for evaluation in MLR, and the field collectively views upward trends on benchmarks as noisy but meaningful indicators of scientific progress [2, 1, 15]. Over time, MLR has evolved strong norms to facilitate widespread benchmarking including the development of open-access datasets, formal competitions and challenges, and accompanying "black-box" software that allows researchers to test their algorithms on benchmark datasets with minimal effort.

The establishment of benchmark datasets as shared resources for evaluation across the MLR community has unique advantages for coordinating scientists around common goals. First, barriers to participation in MLR are reduced since well resourced institutions can shoulder the costs of dataset curation and annotation [2]. Second, by reducing otherwise complex comparisons to a single agreed upon measure, the scientific community can easily align on the value of research contributions and assess whether progress is being made on a particular task. Finally, a complete commitment to benchmarking has allowed MLR to relax reliance on slower institutions for evaluating progress like peer-review or theoretical integration. Together, these advantages have contributed to MLR's unprecedented transformation into a "rapid discovery science" in the past decade [16].

While there are clear advantages to benchmarking as a methodology of comparing algorithms and measuring progress in a problem domain, there are growing concerns regarding benchmarking cultures in MLR which tend to valorize state-of-the-art (SOTA) results on established benchmark datasets over other forms of quantitative or qualitative analysis. The necessity of SOTA results on well established benchmarks for publication acceptance has been identified as a barrier to the development of new ideas [17] and there have been growing calls for more rigorous and comprehensive empirical analysis

[1]paperswithcode.com

[2]However, machine learning model development still remains a resource intensive activity.

of models beyond standard top-line metrics, including reporting model size, energy consumption, fairness metrics, and more [18, 19, 20, 21]. The standard benchmarking paradigm also contributes to underspecification challenges in ML pipelines since a given level of performance on a held out benchmark test set doesn't guarantee a model has learned the appropriate causal structure of a problem [14]. In short, while community alignment on benchmarks and metrics can enable rapid algorithmic advancement, hyper focus on singular metrics at the expense of other more comprehensive forms of rigorous evaluation can lead the community astray and risk the development of models that poorly generalize to the real world.

The MLR community has begun to reflect on the utility of established benchmarks in the field and their appropriateness for evaluative purposes. For example, the Fashion-MNIST dataset was introduced because MNIST is perceived to be over-utilized and too easy [22], and the utility of ImageNet — one of the most influential ML benchmark in existence — as a meaningful measure of progress has been a focus of critical examination in the past couple years [23, 24]. SOTA chasing concerns are also compounded by the great capacity ML algorithms have to be "right for the wrong reason" [25], enabling SOTA results that rely on "shortcuts" rather than learning the causal structure dictated by the task [13]. [26] suggests the NLP community may have been "led down the garden path" by over-focusing on "beating" benchmark tasks with models that can easily manipulate linguistic form without any real capacity for language understanding. Recent dataset audits have also revealed that established benchmark datasets tend to reflect very narrow — typically white, male, Western — slices of the world [4, 5, 6, 7]. Thus, over-concentration of research on a small number of datasets and metrics can distort perceptions of progress within the field and have serious ethical implications for communities impacted by deployed models. Despite these discussions, little empirical work has considered whether over-concentration of research on a small number of datasets is a systemic issue. This prompts our first research question:

RQ1: How concentrated are machine learning task communities on specific datasets and has this changed over time?

There are also growing concerns regarding the gap between benchmark datasets and the problem domains that they are being used to evaluate progress in. For example, [12] found that computer vision datasets tend to be developed in a manner that is decontextualized from a particular task or application area. Supposedly "general purpose" benchmarks are often valued within the field, though the precise bounds of what makes a dataset suitable for general evaluative purposes remains unclear [15]. These observations prompt our second research question:

RQ2: How frequently do machine learning researchers borrow datasets from other tasks instead of using one created explicitly for that task?

Despite widespread recognition that datasets are critical to the advancement of the field, slow careful dataset development is often undervalued and disincentivized, especially relative to algorithmic contributions [27, 12]. Given the high value the MLR community places on SOTA performance on established benchmarks, researchers are also incentivized to reuse recognizable benchmarks to legitimize their contributions. Moreover, dataset development is time and labor intensive, making large scale dataset development potentially inaccessible to lower-resourced institutions. These observations prompt our final research question:

RQ3: What institutions are responsible for the major ML benchmarks in circulation?

3 Data

Our primary data source for this work is Papers With Code[3] (PWC), an open source repository for machine learning papers, datasets, and evaluation tables created by researchers at Facebook AI Research. PWC is largely community contributed — anyone can add a benchmarking result or a task, provided the benchmarking result is published in a paper as pre-print, in a conference or a journal. Once tasks and datasets are introduced by humans, PWC scrapes ArXiv using keyword searches to find other examples of the task or uses of the dataset.

We downloaded the complete PWC dataset on 06/16/2021 (licensed under CC BY-SA 4.0). In this study, we focus primarily on the "Datasets" archive, as well as papers utilizing those datasets. Each

[3] www.paperswithcode.com

dataset in the archive is associated with metadata such as the modality of the dataset (e.g., texts, images, video, graphs), the date the dataset was introduced, and the paper title that introduced the dataset (if relevant). At the time we found 4,384 datasets on the site and scraped 60,647 papers that PWC associates with those datasets using a PWC internal API.

In PWC papers, benchmarks, and (by transitivity) datasets are associated with tasks. For this analysis we were constrained to the 46,668 papers that use a dataset and are labeled with a task (see Figure 6 for a truncated histogram of usage across datasets). These papers collectively use 3,511 datasets. Studying the transfer of datasets between tasks imposes an additional constraint that we must know both the task of the paper that introduced the dataset ("the origin task") and the task of the paper that used the dataset later in time ("the destination tasks"). For example, ImageNet [28], was introduced as a benchmark for *Object Recognition* and *Object Localization* (origin tasks), but is now regularly utilized as a benchmark for *Image Generation* (destination task) among many others.

2,583 datasets on PWC were formally introduced in a paper affiliated with a task, utilized by 39,465 unique papers. An additional 640 datasets were introduced in a paper, but not labeled with tasks. Two authors manually labeled 50 of these dataset papers with tasks (see supplemental spreadsheet tab "Manually Tasked Datasets" for justifications) allowing us to include another 17,219 utilizing papers. We do not utilize the remaining 590 datasets and 2790 utilizing papers (14%).

PWC includes a taxonomy of tasks and subtasks but the graph is cyclic, making it hard to disentangle dataset transfer between broad tasks and finer-grained tasks (see data supplement tab "Task Relations"). For each transfer, we annotate both the transfer between the origin and destination, and the transfer between the origin's parents and the destination's parents. This approach allows us to accurately capture both dynamics between larger tasks (e.g., *Image Classification* and *Image Generation*), and between finer-grained tasks (e.g,. *Image-to-Image Translation* and *Image Inpainting* who are both children of *Image Generation*). Because we found dataset usages to be noisy (i.e., a paper would be associated with a dataset if the keyword appeared multiple times in the paper), we restricted each transfer to destination tasks that PWC had already associated with that dataset.

Datasets for Analyses 1 and 2 (RQ1, RQ2): Because our annotation system double counts transfers of a single dataset across different levels of organization, we chose to focus exclusively on high-level transfers between 313 parent tasks. Because the metrics we use in the analyses (particularly Gini and Creation Ratio) are biased in small samples, we chose to focus only on parent tasks with more than the median number of 31 papers. This resulted in a final sample of 133 tasks with 47,607 collective uses and 924 unique datasets (see supplemental spreadsheet tab "List of Tasks").

Dataset for Analysis 3 (RQ3): To study the distribution of successful datasets across institutions, we linked datset-introducing papers to the Microsoft Academic Graph (MAG) [29]. Affiliation concentration analyses were performed on the 2,461 datasets with papers that had the last author affiliation annotated in MAG.

4 Methods and Findings

4.1 Analysis 1 (RQ1): Concentration in Task Communities on Datasets

4.1.1 Methods

To measure how concentrated task communities are on certain datasets (RQ1), we calculated the Gini coefficient across the distribution of observed dataset usages within each task. Gini is a continuous measure of dispersion in frequency distributions. The metric is frequently used in social science to study inequality [30]. The Gini score varies between 0 and 1, with 0 indicating that the papers within a task use all datasets in equal proportions, and 1 indicating that only a single dataset is used across all dataset-using papers. Gini is calculated as the average absolute difference in the usage of all pairs of datasets used in the task, divided by the average usage of datasets. Formally, if x_i is the number of usages of dataset i out of all n datasets used in the task, then the Gini coefficient of dataset usage is,

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \tag{1}$$

Because Gini can be biased in small samples [31], we use the the sample corrected Gini, $G_s = \frac{n}{n-1}G$, and excluded tasks (or task-years when disaggregating by time) with fewer than 10 papers.

Regression Model 1: In addition to descriptive statistics, we built a regression model to assess the extent to which observed trends in Gini from year-to-year could be attributable to confounds like task size, task age, or other task-specific traits at that time. Our outcome is G_s in each task year from 2015-2020 (Figure 7 shows PWC coverage is limited for papers published before 2015). Our predictors of interest are:

1. **Year** (since we are interested in trends in concentration over time)
2. **CV, NLP, Methods** (three dummy variables indicating whether the task belongs to the Computer Vision, Natural Language Processing, or Methodology categories in PWC).

To absorb additional heterogeneity, we also included the following control covariates:

1. **Task size** in number of dataset-using/introducing papers for that task in that year
2. **Task age** (because younger tasks may have higher Gini coefficients)
3. **Random intercepts for each task** (because we have repeated observations over time)

Gini is bounded between 0 and 1 so we use beta regression [30], but apply the smoothing transformation in [32] to deal with the occasional task-year where the Gini is 0. We use a fully restricted model with the following interactions:

$$
\begin{aligned}
Beta(G_s) = \alpha + \beta_1 \text{Year} + \beta_2 \text{Task Size} + \beta_3 TaskAge\ + \quad & (2)\\
\beta_4 \text{CV} + \beta_5 \text{NLP}\ + \beta_6 \text{Methods} + \beta_7 \text{Full size}\ + \quad & (3)\\
\beta_8 \text{CV*Year} + \beta_9 \text{NLP*Year} + \beta_9 \text{Methods*Year} + \quad & \\
\beta_{10} \text{Year*Task Age *Task Size} \quad & (4)
\end{aligned}
$$

This model was favored over all simpler models on AICc.

4.1.2 Findings

Controlling for task age, task size, and task-specific effects, Model 1 finds significant evidence for increasing concentration in task communities for the full dataset over time, predicting a marginal increase in Gini of .065 from 2015-2020 (Figure 1 top green; Table 1). This trend is also visible in the overall distributions of Gini coefficients over this period (Figure 1 bottom). By 2020, the Gini coefficient for a task was .60. There are no statistically significant differences between Computer Vision and Methodology tasks compared to the full sample (Figure 1 top, Figure 5), but Model 1 suggests increases in concentration are attenuated in 2019-2020 for Natural Language Processing. During these two years, the model predicts NLP concentrations to decrease by .013 while the full sample increases by .012.

4.2 Analysis 2 (RQ2): Changes in Rates of Adoption and Creation of Datasets Over Time

4.2.1 Methods

We created two proportions to better understand patterns of dataset usage and creation within tasks as outcomes:

$$\text{Adoption Ratio} = \frac{\text{\# of Papers Using Datasets from Other Tasks}}{\text{\# of Papers Using Datasets from Other Tasks} + \text{\# of Papers Using Datasets from This Task}}$$

$$\text{Creation Ratio} = \frac{\text{\# of Datasets Created Within This Task}}{\text{\# of Datasets Created within this Task} + \text{\# of Datasets Imported from Other Tasks}}$$

Aggregated Descriptive Analyses: We first computed these proportions for each of the 133 parent tasks aggregated across all years, and subsetted these by the "Computer Vision," "Natural Language Processing," and "Methodology" categories.

Regression Models 2A & 2B: Because our outcomes are now ratios of "successful" counts out of "all" counts, they naturally follow a binomial distribution. We used a mixed effects logistic regression to model these outcomes with the same predictors as Model 1.

4.2.2 Findings

The top row of Figure 2 shows a wide variance in adoption ratios in both the full sample and the subcategories. Within the full sample, more than half of task communities use adopted datasets 57.1% of the time. However, this number varies dramatically across the three PWC subcategories. In more than half of Computer Vision communities, authors adopt 71.7% of their datasets from a different task, while half of Natural Language Processing communities adopt datasets less than 28.3% of the time. Methodology tasks adopt datasets from other tasks at very high rates as well (76.0%).

In the bottom row of Figure 2, we see a largely inverted trend. Of all unique datasets used in a task community, 66.7% are created specifically for that task in more than half of tasks. Within Computer Vision and Methods tasks, the median is lower at 58.9% and 63.3% with similar distributions across tasks. Most strikingly, 78% of datasets are created specifically for the task in more than half of NLP communities with a much tighter variance. We do note that there is a significant correlation between creation ratio and task size (Spearman's $\rho = .26\, p = 0$).

Regression Models 2A and 2B do not find any trends in adoption or creation ratios over time (data not shown).

Figure 1: **Top: Predicted concentration on datasets across task communities over time.** Gini predicted by Model 1 holding task size/age to means. Green plots show the estimated effects of the full dataset, other colors are fixed effects for categories. 95% confidence intervals shown. **Bottom: Distributions of concentrations over time.** Higher Gini indicates greater concentration on fewer datasets. We observe significant spread of Gini across different task communities, with the median trending upwards over time.

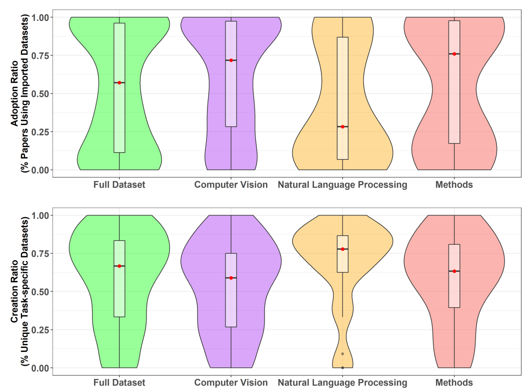


Figure 2: **Adoption (top) and Creation (bottom) Ratios for PWC parent Tasks** . Full dataset in green, tasks in the Computer Vision category in purple, Methods tasks in red, and Natural Language Processing tasks in orange. Red dot and line in boxplot indicate median. Width of violins indicates distribution of tasks.

4.3 Analysis 3 (RQ3): Concentration in Dataset-Introducing Institutions Over Time

4.3.1 Methods

To look at trends in Gini inequality across institutions and datasets over time for the larger set of dataset-using papers, we calculated the Gini coefficient G_s in each year for dataset usages by both dataset and by institution. We regressed this Gini on year, as well as the total number of papers used to estimate G_s, using a standard beta regression. We also mapped dataset-introducing institutions using the longitude and latitude coordinates provided for the first author's institution on Microsoft Academic.

4.3.2 Findings

Overall, we find that widely-used datasets are introduced by only a handful of elite institutions (Figure 3A). In fact, over 50% of dataset usages in PWC as of June 2021 can be attributed to just thirteen institutions. Moreover, this concentration on elite institutions as measured through Gini has increased to the mid .80s in recent years (Figure 3B red). This trend is also observed in Gini concentration on datasets in PWC more generally (Figure 3B blue).

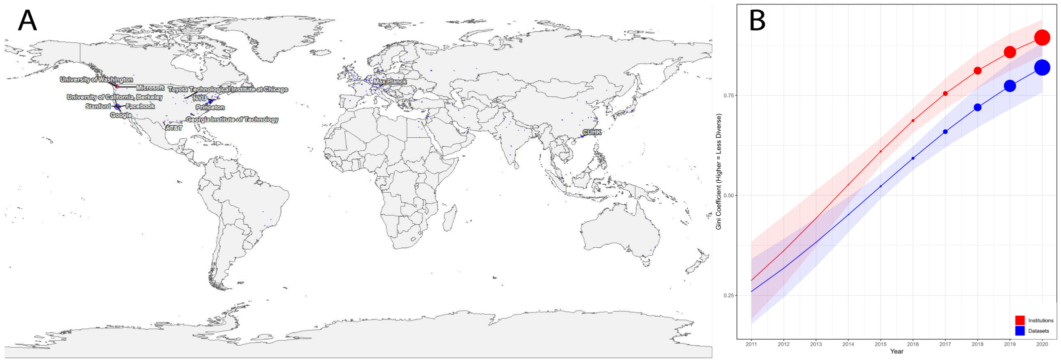


Figure 3: **Increases in concentration of dataset usages on institutions and datasets (non-task specific) over time. A:** Map of dataset usages per institution as of June 2021. Dot size indicates number of usages. Dot color indicates whether the institution is for-profit or not-for-profit. Institutions accounting for 50%+ usages labeled. **B:** Gini coefficient for concentration of dataset usages across the whole PWC dataset over time for both institutions and datasets. Ribbons indicate 95% confidence intervals.

5 Discussion

In this paper, we find that task communities are both heavily concentrated on a limited number of datasets, and that this concentration has been increasing over time (see Figure 1). Moreover, a significant portion of the datasets being used for benchmarking purposes within these communities were originally developed for a different task (see Figure 2). This result is striking given the fact that communities *are* creating new datasets — in most cases more than the unique number that have been imported from other tasks — but the newly introduced datasets are being used at lower rates. When examining PWC as a whole, we find that there is increasing inequality in dataset usage globally, and that more than 50% of all dataset usages in our sample of 28,749 were for datasets introduced by thirteen elite, primarily western, institutions.

While striking, there are valid reasons to expect widespread adoption and concentration on key datasets. First, a certain degree of research focus on a particular benchmark is both necessary and healthy to establish the validly and utility of the benchmark — or in some cases contest these properties — and gain community alignment around the benchmark as a meaningful measure of progress. Second, the curation of large-scale datasets is not just costly in terms of resources, but may require unique or privileged data (e.g., annonymized medical records, self-driving car logs) accessible to only a few elite academic and corporate institutions. Nevertheless, the extent of concentration we observe poses questions relating to the scientific rigor and ecological validity of machine learning research and underscores benchmarking as a vehicle for inequality in the field. In the remainder of this section we discuss our findings in relation to these two broad themes and outline recommendations that can be enacted at an individual and institutional level. We close by discussing limitations of this analysis and outlining directions for future work.

5.1 Scientific rigor and ecological validity of MLR

Heavy concentration of research on a small number of datasets for each task community is a fairly unsurprising result given the value placed on SOTA performance in established benchmark datasets — a valuation incentives individual researchers concentrate efforts on maximizing performance gains on well established benchmarks. However, as discussed in Section 2, over-concentrating research efforts on established benchmark datasets risks distorting measures of progress. Moreover, as the rate of technology transfer has accelerated, benchmarks have been increasingly used by industry practitioners to assess the suitability and robustness of different algorithms for live deployment. This transition has transformed epistemic concerns about overfitting datasets into ethical ones. For example, critical research on face recognition and generation datasets, has repeatedly highlighted the lack of diversity in standard benchmark datasets used to evaluate progress [4], even as the technologies are applied in law enforcement contexts that adversely affect those populations [37]. Figure 4c shows the top datasets in usage within the *Face Recognition* community. Here, we see a significant amount of high stakes reserch being concentrated on a small number of datasets, many of which contain significant racial and gender biases [4, 38]. An in depth examination of bias within the top benchmarks datasets in use within different task communities is outside the scope of this work. However, the systemic nature of bias concerns in ML datasets compounds the epistemic concerns of highly concentrated research.

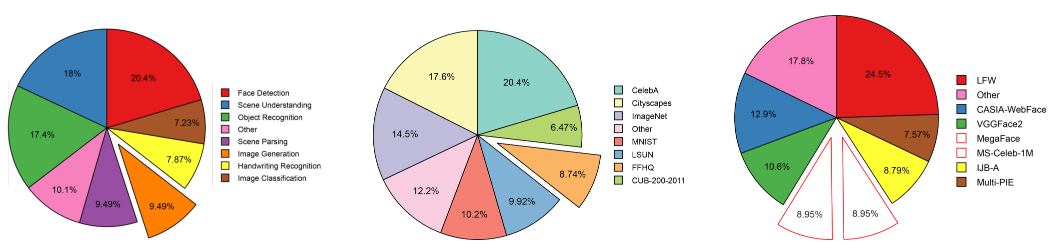
Our findings also indicate that datasets regularly transfer between different task communities. On the most extreme end, the majority of the benchmark datasets in circulation for some task communities were created for other tasks. For example, Figure 4 plots the dataset usages of *Image Generation* papers on PWC broken down by dataset name (Figure 4b) and origin task (Figure 4a). We observe only one of the datasets heavily used in the *Image Generation* community was designed specifically for this task. The widespread practice of adopting established datasets to train and evaluate models in new problem domains isn't inherently a problem. However, this practice does raise potential concerns regarding the extent to which datasets are appropriately aligned with a given problem space. Moreover, given the widespread prevalence of systematic biases in the most prominent ML datasets, adopting existing datasets, rather than investing in careful curation of new datasets, risks further entrenching existing biases.

Our findings relating to creation and adoption rates are quite nuanced, and the extent to which high adoption rates raise significant concerns to ecological validity are yet to be determined. Furthermore we believe it is worth distinguishing between at least two forms of dataset adoption that seem to be conflated in the PWC data. On the one hand, we observe datasets that have been developed for one task be adopted and *adapted* in some form for a new task through, for example, the addition of new annotations. On the other hand, we observe some datasets being adopted whole cloth from one task community to another. Each of these forms of dataset adoption potentially raises unique concerns regarding the validity of the benchmark in a given context. That said, our results add empirical support to the growing body of scholarship calling for dataset development and use to be rooted in context [3, 12], particularly important for application oriented tasks.

Our findings also compliment and support the growing calls to include forms of qualitative and quantitative evaluations beyond top-line benchmark metrics [18, 19, 20, 21]. Given the observed high concentration of research on a small number of benchmark datasets, we believe diversifying forms of evaluation is especially important to avoid overfitting to existing datasets and misrepresenting progress in the field. Reducing the near-singular emphasis on SOTA results on established benchmarks may also offer more voices the opportunity to shape the culture and trajectory of the field.

5.2 Social inequality in MLR

The extent of concentration we observe underscores that benchmarking is also a vehicle for inequality in science. The *prima facie* scientific validity granted by SOTA benchmarking is generically confounded with the social credibility researchers obtain by showing they can compete on a widely recognized dataset, even if a more context-specific benchmark might be more technically appropriate. We posit that this dynamics creates a "Matthew Efffect" where successful benchmarks, and the elite



(a) Origin tasks of datasets used by *Image Generation* community.

(b) Datasets used by *Image Generation* community.

(c) Datasets used by *Face Recognition* community.

Figure 4: **Top datasets used across *Image Generation* and *Face Recognition* task communities**: (a) Origin task communities of top *Image Generation*. Only 9.49% of *Image Generation* papers in PWC evaluate on datasets developed for *Image Generation*. (b) Names of top *Image Generation*. Only one of the top datasets, FFHQ [33], was developed for the task. (c) The small number of datasets in usage within the high stakes domain of *Face Recognition*. Two of the datasets, MegaFace [34] and MS Celeb-1M [35], have been recently retracted, the latter due to serious ethical violations [36].

institutions that introduce them, gain outsize stature within the field. To the extent that benchmarks shape the types of questions that get asked and algorithms that get produced, current benchmarking practices thus offer a mechanism through which a small number of elite institutions, both academic and for-profit, can shape the agenda of the field. Moreover, because research trends influence broader public discourse, opinions, and potentially even policy decisions, this influence extends into the broader social world as well.

The recently introduced NeurIPS Dataset and Benchmark Track is a clear example of an intervention that shifts incentive structures within the MLR community by rewarding dataset development and other forms of data work. We believe these sorts of interventions can play a critical role in incentivizing careful dataset development that is meaningfully aligned with problem domains. However, our finding that a small number of well-resourced institutions are responsible for most benchmarks in circulation today has implications for data oriented interventions in the field. Our research suggests that simply calling for ML researchers to develop more datasets, and shifting incentive structures so that dataset development is valued and rewarded may not be enough to diversify dataset usage and diversify the perspectives that are ultimately shaping and setting MLR research agendas. In addition to incentivizing dataset development, we advocate for equity oriented policy interventions that prioritize significant funding for people in less resourced institutions to create high-quality datasets. This would diversify — from a social and cultural perspective — the benchmark datasets in rotation.

5.3 Limitations and Future Work

In this paper, we provide the first field-scale analysis on dataset usage in MLR. Because our findings rely on a unique community-curated resource, our findings are contingent on the structure and coverage of PWC. The crowdsourced taxonomy of parent-child task relations in PWC is both noisy and open to interpretation. We have included the full list of parent tasks used in our analysis in the supplementary material, as well as the parent/child relations. We focused our adoption and creation rate analyses on parent-to-parent transfers in an effort to curtail any concerns regarding arbitrariness of task boundaries for fine grained tasks.

As with any dataset, PWC also reflects various forms of curatorial bias. To control for spurious labels of dataset usage, we conservatively only considered usages of a dataset valid if they shared a task label with the dataset. Our own curatorial decision influenced the final dataset as well. As noted in Section 3, there were also a large number of datasets that were not assigned an origin task. We manually assigned tasks to the top datasets (assignments and justifications for assignment included in Supplementary Material) from this set and dropped the remaining 14% of uses. Lastly, PWC is likely to reflect recency bias.

Finally, we emphasize that our findings are highly nuanced. We report trends that our analysis revealed, but refrain from imposing normative judgements on many of these trends. For example, the high rates of adoption raise potential concerns and points to an important future area of examination. The mere fact that datasets travel between task communities is not necessarily problematic, and indeed the widespread sharing of datasets has been central to methodological advancements in the field. We hope this work will offer a foundation for future empirical work examining the details of dataset transfer and the context specific implications of our findings.

6 Conclusion

Benchmark datasets play a powerful role in the social organization of the field of machine learning. In this work, we empirically examine patterns of creation, adoption, and usage within and across MLR task communities. We find that benchmarking practices are heavily concentrated on a small number of datasets for each task community and heavily concentrated on datasets originating from a small number of well resourced institutions across the field as a whole. We also find that many benchmark datasets flow between multiple task communities and are leveraged to evaluate progress on tasks for which the data was not explicitly designed. We hope this analysis will inform community-wide initiatives to shift patterns of dataset development and use so as to enable more rigorous, ethical, and socially informed research.

References

[1] David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. *ArXiv*, abs/2007.04792, 2020.

[2] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 294, New York, NY, USA, 2020. Association for Computing Machinery.

[3] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *NeurIPS Workshop on Machine Learning Retrospectives, Surveys, and Meta-analyses*, 2020.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of Machine Learning Research*, volume 81, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[5] S. Shankar, Yoni Halpern, Eric Breck, J. Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*, 2017.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[7] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

[8] Kate Crawford and Trevor Paglen. *Excavating AI: The Politics of Images in Machine Learning Training Sets*, 2019.

[9] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1537–1547, 2021.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[11] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Morgan Klaus Scheuerman, Emily Denton, and Alex Hanna. Do datasets have politics? disciplinary values in computer vision dataset development. *Computer Supported Cooperative Work (CSCW)*, 2021.

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[14] Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory

Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020.

[15] Amandalynne Paullada Emily Denton Alex Hanna Deborah I Raji, Emily M. Bender. Ai and the everything in the whole wide world benchmark. *NeurIPS Workshop on Machine Learning Retrospectives, Surveys, and Meta-analyses*, 2020.

[16] Randall Collins. Why the social sciences won't become high-consensus, rapid-discovery science. In *Sociological forum*, volume 9, pages 155–177. Springer, 1994.

[17] Tom Simonite. *Google's AI Guru Wants Computers to Think More Like Brains*, 2018.

[18] D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and A. Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *ICLR*, 2018.

[19] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019.

[20] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics.

[21] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards. In *arXiv:2009.13888*, 2020.

[22] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[23] L. Beyer, Olivier J. H'enaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *ArXiv*, abs/2006.07159, 2020.

[24] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. *International Conference on Machine Learning (ICML)*, 2020.

[25] Benjamin Heinzerling. NLP's Clever Hans moment has arrived. *The Gradient*, 2019.

[26] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of FAccT 2021*, 2021.

[27] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. 2021.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[29] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020.

[30] James B. McDonald and Michael Ransom. *The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation of Related Measures of Inequality*, pages 147–166. Springer New York, New York, NY, 2008.

[31] George Deltas. The small-sample bias of the gini coefficient: results and implications for empirical research. *Review of economics and statistics*, 85(1):226–234, 2003.

[32] Michael Smithson and Jay Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, 2006.

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[34] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[35] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. volume 9907, pages 87–102, 10 2016.

[36] Jules. Harvey, Adam. LaPlace. Exposing.ai, 2021.

[37] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The perpetual line-up: Unregulated police face recognition in america, 2016.

[38] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We describe limitations in Section 5.3.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] As discussed in Section5.3, our findings are highly nuanced. One potential negative societal impact of this work would be if the nuances of our analysis are lost. We have taken great care to articulate the extent to which our analysis is contingent on the particularities of the PWC repository and importance of future work examining in more depth the implications of this work for research validity in different task communities.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We have included our data in the supplementary material. Code will be released upon paper acceptance.
   (b) Did you specify all the training details [Yes] Parameters of regresion model described in Section 4.1.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] The PWC data is open source. We have included additional coding we added in the supplement.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Table 1: "Exponentiated coefficients for fixed effects in Regression Model 1"

	Term	Estimate	Std Error	Statistic	P-value
1	(Intercept)	1.21	1.22	0.96	0.34
2	Year	1.10	1.04	2.37	0.02
3	Task Size	2.73	1.16	6.63	0.00
4	Task Age	1.03	1.11	0.26	0.79
5	CV	0.96	1.21	-0.19	0.85
6	NLP	1.03	1.22	0.16	0.87
7	Methodology	0.68	1.21	-2.04	0.04
8	Year:Task Size	0.84	1.03	-5.95	0.00
9	Year:Task Age	0.98	1.02	-1.06	0.29
10	Task Size:Task Age	1.36	1.17	1.96	0.05
11	Year:CV	0.95	1.04	-1.33	0.18
12	Year:NLP	0.90	1.04	-2.57	0.01
13	Year:Methodology	1.06	1.04	1.48	0.14
14	Year:Task Size:Task Age	0.95	1.03	-1.80	0.07
15	SD(Task Random Intercepts)	1.66			

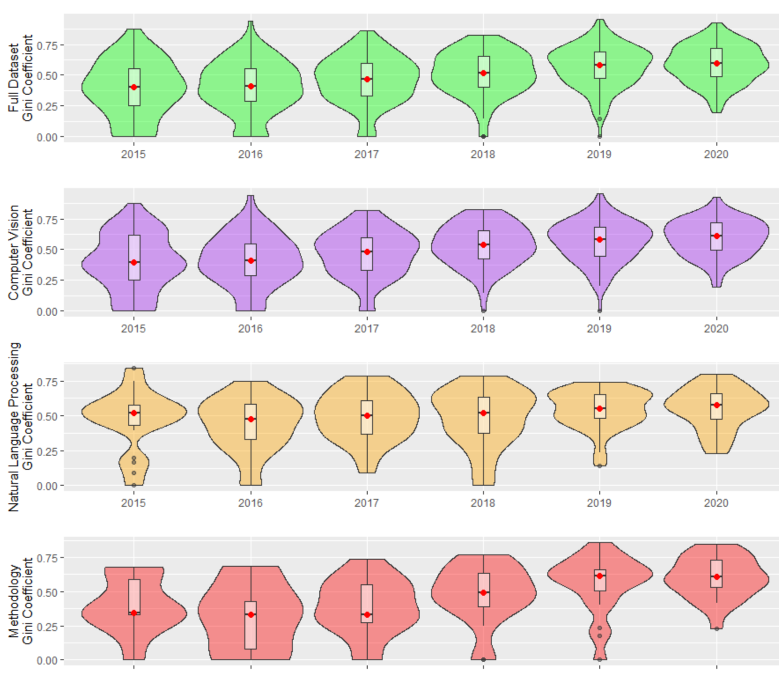


Figure 5: **Increases in concentration on datasets within task communities over time.** Higher Gini coefficient indicates greater concentration on fewer datasets. We observe significant spread of Gini across different task communities, with the median trending upwards over time for all modalities. Green is the full dataset, other colors indicate subsets of the data by PWC task category.

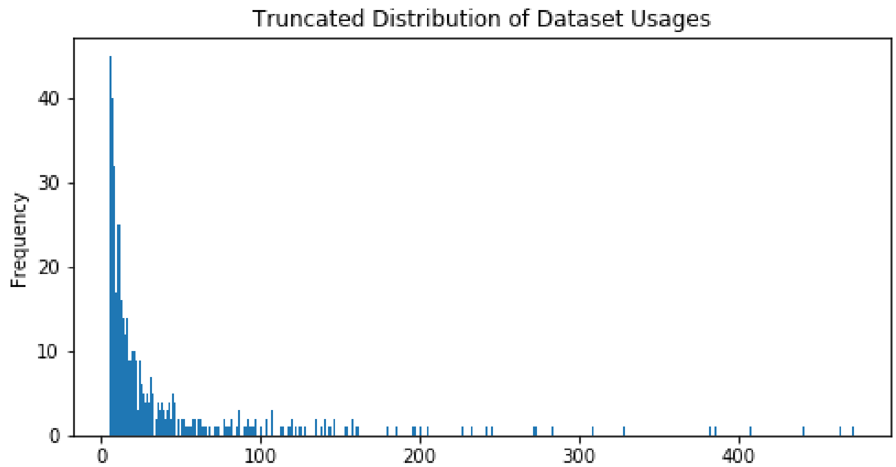


Figure 6: **Truncated distribution of usages per dataset in PWC.** Usages measured conservatively by only allowing usages from tasks the dataset was labeled for. 3760 datasets with less than 5 papers and 8 datasetwith over 500 uses dropped for clarity. 8 datasets are Penn Treebank,CelebA, SQuAD, KITTI, MNIST, Cityscapes, ImageNet, COCO.

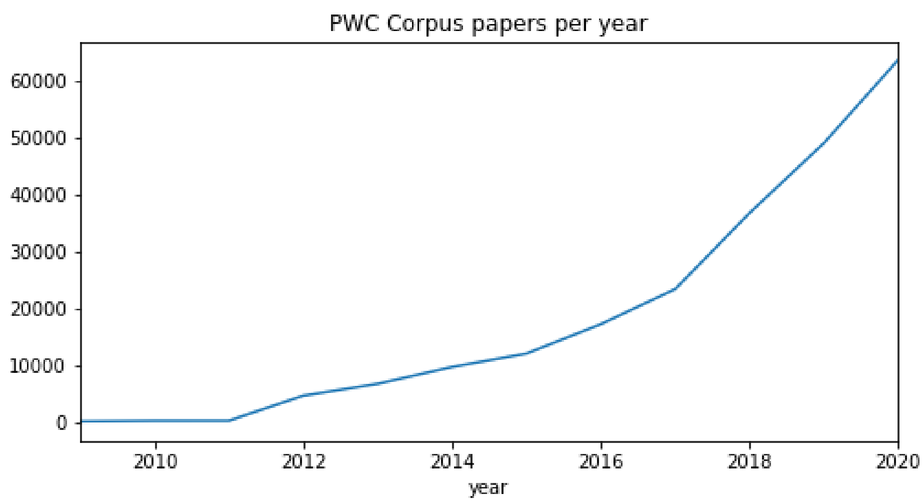


Figure 7: **Number of Papers in the Papers with Code Corpus**. Full set of "Papers with Abstracts" on Papers with Code as of June 2021. Total dataset size is 137,510 papers. Daily snapshots of this dataset are available at github.com/paperswithcode.