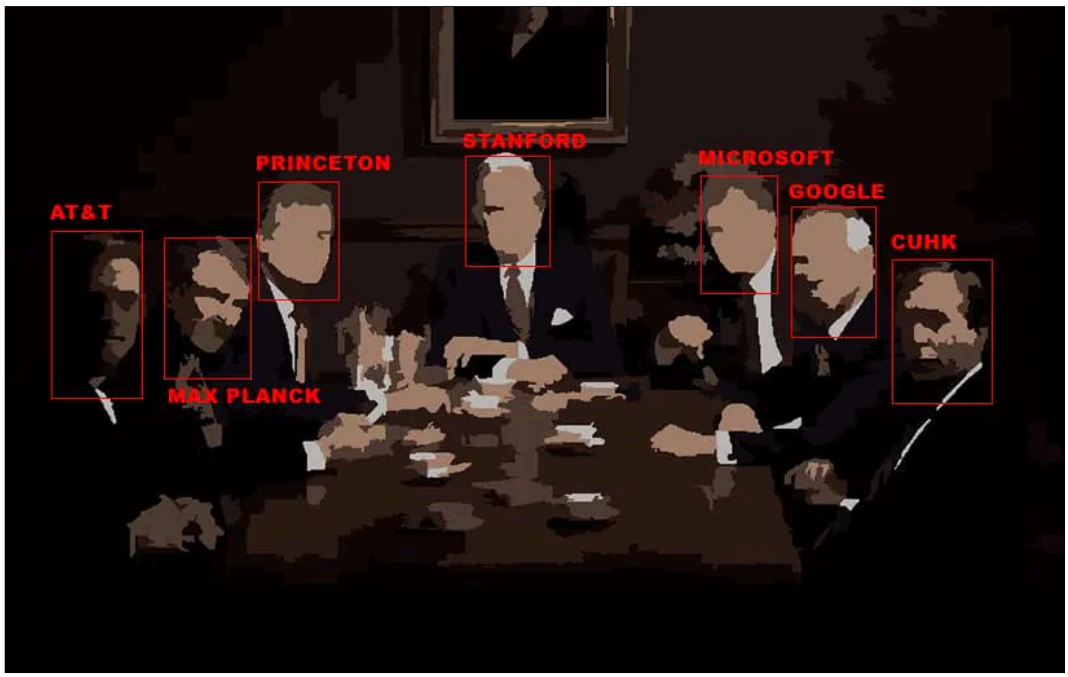
A Cartel of Influential Datasets Is Dominating Machine Learning Research, New Study Suggests

Originally published at https://www.unite.ai/a-cartel-of-influential-datasets-are-dominating-machine-learning-research-new-study-suggests/

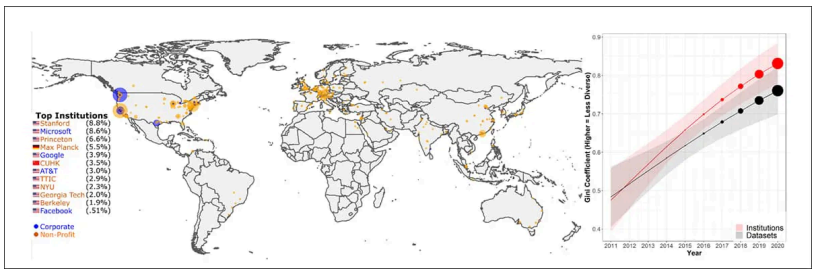


A new paper from the University of California and Google Research has found that a small number of 'benchmark' machine learning datasets, largely from influential western institutions, and frequently from government organizations, are increasingly dominating the AI research sector.

The researchers conclude that this tendency to 'default' to highly popular open source datasets, such as ImageNet, brings up a number of practical, ethical and even political causes for concern.

Among their findings – based on core data from the Facebook -led community project Papers With Code (PWC) – the authors contend that *'widely-used datasets are introduced by only a handful of elite institutions'*, and that this 'consolidation' has increased to 80% in recent years.

> *'[We] find that there is increasing inequality in dataset usage globally, and that more than 50% of all dataset usages in our sample of 43,140 corresponded to datasets introduced by twelve elite, primarily Western, institutions.'*



A map of non-task specific dataset usages over the last ten years. Criteria for inclusion is where the institution or company accounts for more than 50% of known u coefficient for concentration of datasets over time for both institutions and datasets. Source: https://arxiv.org/pdf/2112.01716.pdf

The dominant institutions include Stanford University, Microsoft , Princeton, Facebook, Google, the Max Planck Institute and AT&T . Four out of the top ten dataset sources are corporate institutions.

The paper also characterizes the growing use of these elite datasets as *'a vehicle for inequality in science'*. This is because research teams seeking community approbation are more motivated to achieve state-of-the-art (SOTA) results against a consistent dataset than they are to generate original datasets that have no such standing, and which would require peers to adapt to novel metrics instead of standard indices.

In any case, as the paper acknowledges, creating one's own dataset is a prohibitively expensive pursuit for less well-resourced institutions and teams.

> *'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 these dynamics creates a "Matthew Effect" (i.e. "the rich get richer and the poor get poorer") where successful benchmarks, and the elite institutions that introduce them, gain outsized stature within the field.*

The paper is titled *Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research*, and comes from Bernard Koch and Jacob G. Foster at UCLA, and Emily Denton and Alex Hanna at Google Research.

The work raises a number of issues with the growing trend towards consolidation that it documents, and has been met with general approbation at Open Review. One reviewer from NeurIPS 2021 commented that the work is *'extremely relevant to anybody involved in machine learning research.'* and foresaw its inclusion as assigned reading at university courses.

From Necessity to Corruption

The authors note that the current culture of 'beat-the-benchmark' emerged as a remedy for the lack of objective evaluation tools that caused interest and investment in AI to collapse a second time over thirty years ago, after the decline of business enthusiasm towards new research in 'Expert Systems':

> *'Benchmarks typically formalize a particular task through a dataset and an associated quantitative metric of evaluation. The practice was originally introduced to [machine learning research] after the "AI Winter" of the 1980s by government funders, who sought to more accurately assess the value received on grants.'*

The paper argues that the initial advantages of this informal culture of standardization (reducing barriers to participation, consistent metrics and more agile development opportunities) are beginning to be outweighed by the disadvantages that naturally occur when a body of data becomes powerful enough to effectively define its 'terms of use' and scope of influence.

The authors suggest, in line with much recent industry and academic thought on the matter, that the research community no longer poses novel problems if these can't be addressed through existing benchmark datasets.

They additionally note that blind adherence to this small number of 'gold' datasets encourages researchers to achieve results that are overfitted (i.e. that are dataset-specific and not likely to perform anywhere near as well on real-world data, on new academic or original datasets, or even necessarily on different datasets in the 'gold standard').

> *'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.'*

Government Influence in Computer Vision Research

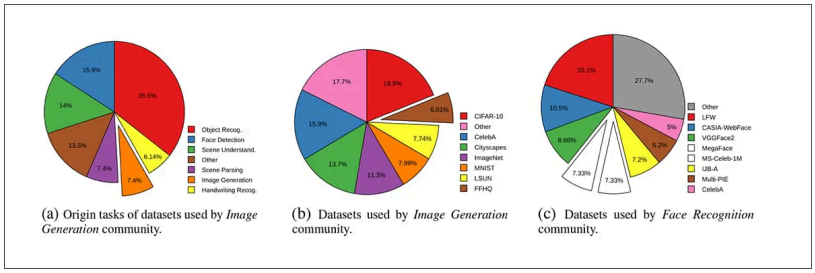
According to the paper, Computer Vision research is notably more affected by the syndrome it outlines than other sectors, with the authors noting that Natural Language Processing (NLP) research is far less affected. The authors suggest that this could be because NLP communities are *'more coherent'* and larger in size, and because NLP datasets are more accessible and easier to curate, as well as being smaller and less resource-intensive in terms of data-gathering.

In Computer Vision, and particularly regarding Facial Recognition (FR) datasets, the authors contend that corporate, state and private interests often collide:

> *'Corporate and government institutions have objectives that may come into conflict with privacy (e.g., surveillance), and their weighting of these priorities is likely to be different from those held by academics or AI's broader societal stakeholders.'*

For facial recognition tasks, the researchers found that the incidence of purely academic datasets drops dramatically against the average:

> *'[Four] of the eight datasets (33.69% of total usages) were exclusively funded by corporations, the US military, or the Chinese government (MS-Celeb-1M, CASIA-Webface, IJB-A, VggFace2). MS-Celeb-1M was ultimately withdrawn because of controversy surrounding the value of privacy for different stakeholders.'*



The top datasets used in Image Generation and Face Recognition research communities.

In the above graph, as the authors note, we also see that the relatively recent field of Image Generation (or Image Synthesis) is heavily reliant on existing, far older datasets that were not intended for this use.

In fact, the paper observes a growing trend for the 'migration' of datasets away from their intended purpose, bringing into question their fitness for the needs of new or outlying research sectors, and the extent to which budgetary constraints may be 'genericizing' the scope of researchers' ambitions into the narrower frame provided both by the available materials and by a culture so obsessed with year-on-year benchmark ratings that novel datasets have difficulty gaining traction.

> *'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.'*

Regarding the machine learning luminaries (including Andrew Ng) who have increasingly called for more diversity and curation of datasets in recent years, the authors support the sentiment, but believe that this kind of effort, even if successful, could potentially be undermined by the current culture's dependence on SOTA-results and established datasets:

> *'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 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 being used to evaluate modern ML methods.'*

6th December 2021, 4:49pm GMT+2 – Corrected possessive in headline. – MA

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More