JANUARY 2021

Comparing Corporate and University Publication Activity in AI/ML

CSET Data Brief

AUTHORS

Simon Rodriguez

Tim Hwang

Rebecca Gelles

Introduction

Corporate labs dominate the public consciousness around artificial intelligence (AI) and machine learning (ML). The widely discussed defeat of world Go champion Lee Sedol by DeepMind's AlphaGo system in 2016 arguably triggered the contemporary wave of popular coverage and mainstream excitement around the technology. Apple's Siri and Amazon's Alexa—both products manufactured by private companies—are two of the most recognizable examples of AI in use today.

This popular focus on corporate activity in ML—particularly of U.S. companies—might lead to the assumption that private labs entirely dominate the ML research relative to the work being conducted in universities. But is this actually the case?

This data brief explores this question by comparing aggregate publication and citation activity among leading U.S. companies and universities publishing research literature on ML.

We find the following:

- Nearly all major corporate labs associated with the AI "revolution" in recent years publish significantly fewer ML papers than universities.
- Despite this, papers produced by corporate labs attract significantly more citations to their research work. But, even here, significant differences exist among the various corporate labs considered to be leading in the technology in terms of publication rates and citations received.
- On a year-to-year basis, the most cited papers in ML do not originate from corporate labs acting alone. Instead, cross-collaborations between companies and university authors tend to be the more frequent source of leading papers in the field for industrial labs. More generally, papers published solely by university authors have constituted a stable majority of the top 100 leading papers in ML by citations since 2010.

Our findings have two important implications for policymakers and strategists thinking about U.S. national competitiveness in AI. For one, narrowly focusing on the widely reported corporate activity in the AI industry may miss important trends taking place elsewhere in the field of ML. Second, policymakers need to consider the distinctions among various leading technology companies as they differ significantly in their level of involvement and impact in the field of ML.

Methodology

The dataset used in this analysis is drawn from Dimensions, a joint database created by Digital Science that tracks more than 128 million scholarly publications, grants, data, and metrics. Using the methodology discussed in an earlier CSET publication, we first extracted a set of 1,269,033 papers related to the topic of AI.[1]

Our use of Dimensions in this analysis is predicated on its extensive integration of GRID, an open source system for linking affiliated organizations. This allows for a straightforward aggregation of the total number of publications being produced by a given organization and its subsidiaries. For example, GRID links a subsidiary lab like "Google (Canada)" as a child of the parent organization "Google." We leveraged this feature to create a dataset of the top 100 organizations publishing in ML as ranked by the total number of publications since 2010.

While providing a more accurate matching of a given paper with the organizational affiliations of its authors, our use of the Dimensions dataset and the CSET classifier does appear to undercount the overall number of papers published by a given organization. For instance, a broader search for Facebook affiliated papers across research databases suggests that the company has published 239 AI papers in the last decade, as opposed to the 171 papers that appear in Dimensions. Similarly, 4,736 AI papers associated with Carnegie Mellon University appear in Dimensions, while a broader search suggests a higher publication count of 10,054. Sampling across organizations in our dataset suggests that these errors are distributed roughly evenly, and that our overall conclusions below about the relative publication activity hold.

In order to enable a comparison between these top 100 organizations and leading private labs, we added information from Dimensions on Amazon, Apple, Facebook, and Google to our dataset. Two other major U.S. technology companies—IBM and Microsoft—were already included in the dataset since they were among the top 100 most active organizations by publication. This specific selection of companies is based on their representation on the board of the Partnership on AI, a nonprofit coalition founded in 2016 that is committed to the responsible use of artificial intelligence, that we interpret as a signal of public prominence in the industry and in the research field.[2]

This dataset of 104 organizations and their 334,833 papers was then queried to generate a table that linked organization names with their respective publication count and citations to their papers from within the field of AI. These organizations were also labeled as either being a university or company, allowing a comparison across these groups.

Separately, all of the AI papers drawn from Dimensions—even those published by those outside the 104 organizations examined earlier—were segmented by the year of publication. This allowed an examination of the papers published in a given year that have received the highest number of citations from other papers within the field of AI to date. These papers were sorted into three categories describing the type of organizational collaboration that produced them. Papers were categorized as being produced by authors solely affiliated with universities, solely affiliated with companies, or a mix of the two.

Discussion

The most immediate pattern that emerges from our analysis is the differential between the number of ML publications from private labs and the number published by universities. Only two companies—IBM and Microsoft—appear in the list of top 100 organizations ranked by publication volume in Dimensions over the last decade. The difference between the number of papers published by the top 100 organizations and companies like Amazon, Apple, and Facebook is stark.[3]

Figure 1. Bar chart showing number of ML publications by organization since 2010. For a full list of organization names, reference counts, and publication counts, please refer to https://github.com/georgetown-cset/Comparing-Corporate-and-University-Publication-Activity.

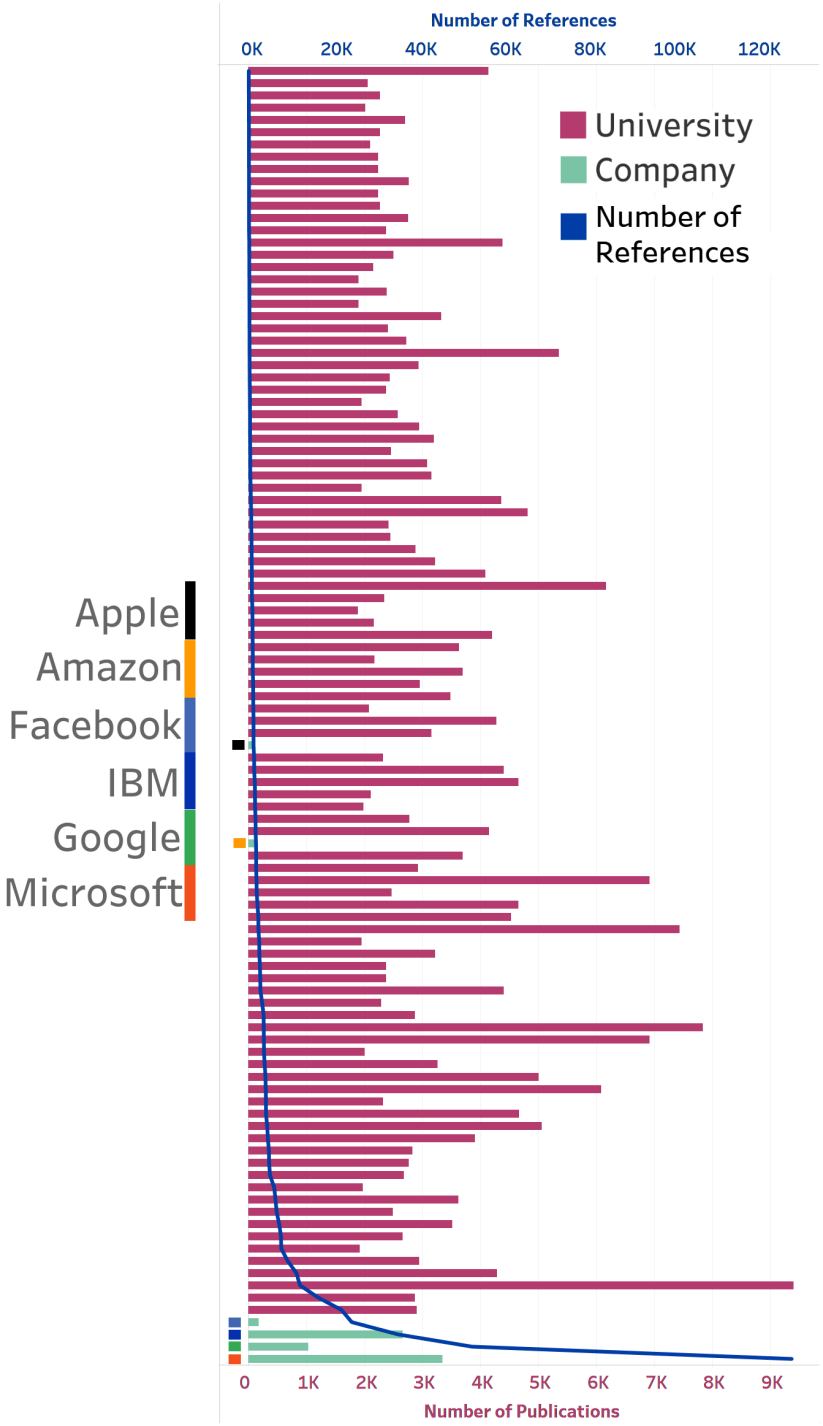
Source: Dimensions

It should be emphasized that publication volume is not the same as research activity. Private companies vary significantly in their willingness to publish research results openly. So, while companies like Apple and Amazon may appear as comparatively small from a research publications perspective, it is erroneous to equate this metric with these organizations lagging in terms of investment or interest in ML overall.

Aggregate publication counts are one way of viewing the relative research activity among universities and U.S. private labs, but they may give a misleading impression of the relative importance of various organizations to the research field.

Another relevant metric is citation count: how many other papers within the field of AI cite the research being published by each organization? While companies may publish relatively less than universities, the research they publish may be relatively more impactful in this sense.

Figure 2. Bar chart showing citation and publication activity across organizations since 2010. For a full list of organization names, reference counts, and publication counts, please refer to https://github.com/georgetown-cset/Comparing-Corporate-and-University-Publication-Activity.



Source: Dimensions

Our analysis indicates that the ML research published by private labs appearing in Dimensions generates significantly more citation activity than the papers from universities. Figure 2 plots citation activity against the publication data depicted in Figure 1.

There is a dramatic differential between companies and universities in citations. Google, despite publishing only half as many papers as the hundredth ranked organization by publication volume in Dimensions, garnered over 50 thousand citations. This is 22.7 times more than the average number of citations received by universities in the dataset, 2,331 citations. The ML research produced by Microsoft has on aggregate generated the most citations among all organizations in our dataset, nearly 130 thousand citations. Even among companies that publish significantly fewer papers such as Apple and Amazon, citation counts are still at least 10 times higher than their publication count. The most significant disparity between publication and citation counts in our dataset is Facebook. The company has published only 176 times in the past decade but has amassed over 24 thousand citations.

As is the case in many other fields, citations on papers within the field of ML follow a power law.[4] We find that the median citation count among the papers produced by these 104 organizations is zero. At the same time, the most highly cited papers receive thousands of citations.

Among organizations, we find that there is a low correlation between the number of publications any one university produces to the number of citations it generates ($R^2 = 0.0397$). In contrast, the private entities in our datasets show a much stronger correlation between publication counts and citation counts ($R^2 = 0.6901$). However, even among corporations, major differences exist. Citations generated by companies producing a comparable number of publications, such as Microsoft and IBM, differ greatly. In Dimensions, Microsoft has published only 26 percent (687) more papers than IBM in ML over the past decade but has garnered 264 percent (93,146) more citations.

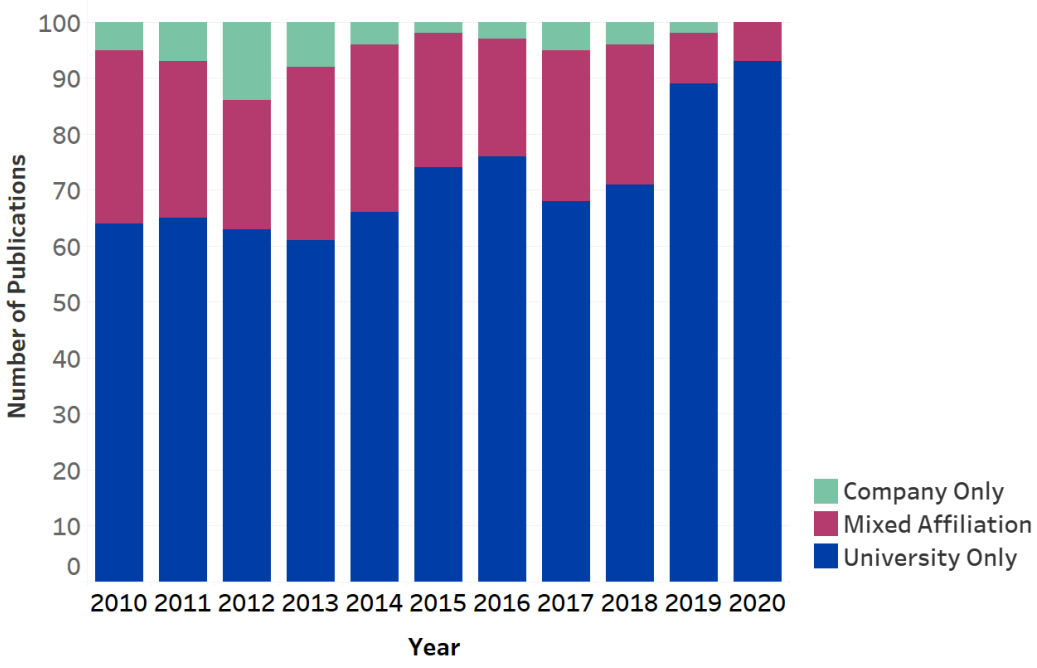
While the Dimensions dataset that this analysis is based on covers a large number of publications in ML, its limitations should be recognized. One major issue is that the Dimensions dataset mostly covers English language publications. We believe that this may distort some of the results, particularly around major Chinese research institutions. Tsinghua University, Zhejiang University, and Shanghai Jiao Tong University are some of the most active organizations in terms of publication volume, yet their overall citation counts are quite low. We hypothesize that this may result, in part, from unseen citations that are not counted in our analysis since Chinese language ML

papers citing these publications are not largely included in Dimensions. This may influence the relative ranking of institutions in our analysis. Future analyses will work to expand our understanding of this "dark matter" of citation activity by more fully incorporating scholarly research in the Chinese language.

We conducted an analysis to understand how patterns in publication and citation activity may have changed across the past decade of ML research. To do this, we used the Dimensions dataset to draw out the list of the hundred most cited papers within the field of ML from each year of publication since 2010. These papers were then labeled based on whether they were written by authors solely affiliated with universities, solely affiliated with companies, or a mix of the two.

Figure 3 is a stacked bar chart showing the organizational affiliation of the top 100 papers by number of citations from within the field of ML for each year of publication since 2010. Each record represents a single publication. Records labeled as "mixed affiliation" represent papers with at least one company and one university author.

Figure 3. Stacked bar chart of top 100 papers by citations for each year of publication since 2010.



Source: Dimensions

While Figure 2 illustrates that papers produced by corporate labs generate significantly more citation activity within the ML field, Figure 3 shows the degree to which this leadership depends on collaborations between private companies and universities. "Company only" papers constitute only a small portion of the leading papers by citations on a year-to-year basis: co-authorship with a university author is by far the more common path for corporate research to appear among the top papers.

Universities also appear to account for a stable majority of leading papers over time. Since 2010, leading papers produced solely by university authors have accounted for more than half of the most cited papers being released each year. While Figure 3 suggests that the portion of universities may be increasing over time, we believe that this is likely an artifact of the recency of the papers released in 2019 and 2020. The papers published in these years will continue to receive citations, shifting the relative balance of organizational affiliations over time.

The importance of university research and cross-collaborations between university and industrial labs is further emphasized when examining the distribution of citations across these top 100 papers in each year. The following table shows the proportion of citations associated with papers in this top 100 listing emerging from different kinds of authorship in our dataset. Universities and cross-collaborations have consistently accounted for papers that capture the majority of citations to the leading papers over the last decade.

Table 1. Annual distribution of citations to top 100 papers across companies and universities.

Year	Total Citations Among Top 100 Papers	Citations to Company Papers	Citations to Mixed Affiliation Papers	Citations to University Papers
2010	111,525	4,153 (4%)	35,668 (32%)	71,704 (64%)
2011	125,284	8,709 (7%)	32,334 (26%)	84,241 (67%)
2012	84,367	12,841 (15%)	22,734 (27%)	48,792 (58%)
2013	71,164	5,286 (7%)	24,966 (35%)	40,912 (57%)
2014	84,126	2,808 (3%)	28,304 (34%)	53,014 (63%)
2015	102,387	1,263 (1%)	41,007 (40%)	60,117 (59%)
2016	57,554	1,066 (2%)	17,839 (31%)	38,649 (67%)
2017	52,629	1,954 (4%)	22,063 (42%)	28,622 (54%)
2018	28,995	2,487 (9%)	7,045 (24%)	19,363 (67%)
2019	9,926	205 (2%)	1,088 (11%)	8,633 (87%)
2020	3,061	0 (0%)	166 (5%)	2,875 (94%)

Source: Dimensions

Conclusion: Nuancing the AI Policy Debate

This analysis introduces two important nuances for those developing national strategy around AI to keep in mind. Consistent with the popular coverage around ML, our data bolsters the argument that private companies do wield a significant influence in the research field. Overall, corporate labs dominate in the creation of research that attracts significant attention from the research field as measured by citations. However, their research output is far from being a representative sample of the entire universe of exploration in the field of ML. Corporate labs account for only a small portion of the overall papers being released and many of the leading papers come entirely from university labs. Policymakers should be careful not to conflate private research as representing the entire field.

Second, analysts should also avoid assuming that the "big tech" private labs conducting research in ML are a monolithic category. This analysis reveals substantial differences between leading companies as to their relative publication activity and impact on the field as measured by citations. For instance, Microsoft and IBM, despite both leading in publication rates, appear to differ wildly in the citations generated by their work. In turn, these companies differ significantly from the limited publication and citation activity seen with companies like Amazon and Apple. Forthcoming analysis of the dataset used in this brief will also show important differences in the substantive areas of research various companies are publishing in, as well. Any policy shaping the incentives or regulations around these companies will need to take this significant heterogeneity into account.

Acknowledgments

The authors would like to thank Avonelle Davis, Melissa Deng, Shelton Fitch, Andrew Kim, Matt Mahoney, Igor Mikolic-Torreira, Dewey Murdick, and Lynne Weil for their feedback and assistance on this project.

Document Identifier: doi: 10.51593/20200067

Endnotes

[1] Dewey Murdick, James Dunham, and Jennifer Melot, "AI Definitions Affect Policymaking" (Center for Security and Emerging Technology, June 2, 2020), https://cset.georgetown.edu/research/ai-definitions-affect-policymaking/.

[2] "Meet the Team," Partnership on AI, accessed August 12, 2020, https://www.partnershiponai.org/team/.

[3] A brief analysis of the per capita output of universities and companies on our list confirmed that these results hold even taking the overall size of these research organizations into account.

[4] It should be highlighted that the comparison of citations to publication activity in Figure 2 does duplicate citation counts where co-authorships between institutions exist. For example, a citation to a paper co-authored by Google and Stanford University would be counted on both institutions in our plot. This may impact the relative ranking of different institutions, particularly given the lopsided distribution of citations across papers. Co-authorship on a particularly highly cited paper may cause the co-authoring organizations to appear significantly "ahead" of their peers in citation counts. However, our overall conclusions comparing universities to corporate labs hold even taking into account this effect due to the high citation counts associated with ML papers published by corporate labs in general.