JULY 2020

Identifying AI-Related Companies: A Conceptual Outline and Proof of Concept

CSET Data Brief

CSET
CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Zachary Arnold
Rebecca Gelles
Ilya Rahkovsky

Executive Summary

Identifying what constitutes an "artificial intelligence company" is useful for many policy and analytic tasks. However, companies relate to AI in different ways. Some dimensions of AI relevance will be more important than others for any given task, and many are difficult to capture in data. With these challenges in mind, CSET is developing a flexible, multi-source, data-driven framework for identifying AI-related companies. In this data brief, we explain our efforts and present a test version. We show that the metrics and data sources chosen to define AI relevance strongly affect which companies qualify as "AI."

Many high-profile tech companies count as top "AI companies" according to some plausible metrics, but not others:

Rank	By recent PhD hires from top U.S. AI programs	By AI publications, worldwide	By top AI conference papers, worldwide	By AI patent families since 2010, worldwide
1	Google*	Microsoft	Microsoft	IBM
2	Facebook	Google*	Google*	Microsoft
3	Microsoft	Siemens	Siemens	Google*
4	Amazon	Samsung	Yahoo	Samsung
5	Apple	Philips	Alibaba	Siemens
6	Alphabet†	Yahoo	NEC	Intel
7	IBM	Mitsubishi	Facebook	Sony
8	Uber	Intel	Amazon	Philips
9	Intel	Bosch	LinkedIn	NEC
10	DeepMind	Huawei	Mitsubishi	Baidu

* Excludes DeepMind
† Excludes Google and DeepMind

At the same time, many somewhat lower-profile companies, such as Siemens, Mitsubishi, Bosch, and NEC, qualify as top "AI companies" according to these metrics. The metrics are examined in greater detail in this report, and the full results of the test framework are available online.[1]

Identifying AI companies in theory

Many questions in artificial intelligence policy involve identifying a pool of AI-related companies. For example, to map technology transfer risks, researchers might examine investment transactions involving foreign investors and domestic AI companies. To gauge the overall health of America's AI sector, they might measure venture capital flows into AI companies. And to predict future AI development and deployment trends, they might study the research activities of AI companies.

This situation poses two problems. First, companies can be active in AI in many different ways, creating different "AI company" populations of interest for policymakers and researchers. For example, a researcher motivated by technology transfer risks might focus on companies pursuing sensitive or strategic AI applications. In contrast, researchers gauging the general health of the AI sector would probably cast a wider net, drawing in companies applying well-known AI techniques to consumer needs. And research on future AI development trends would likely spotlight companies focusing on R&D. In short, different research questions demand different ways of identifying "AI companies."

Second, once a relevant dimension of AI involvement is identified, finding real-world data capturing that dimension can prove difficult. Any single metric will have flaws; for example, patenting statistics are often used to assess companies' and nations' research productivity, but experts caution that these statistics can be misleading on their own.[2] In many cases, combining and comparing multiple metrics allows a richer understanding.

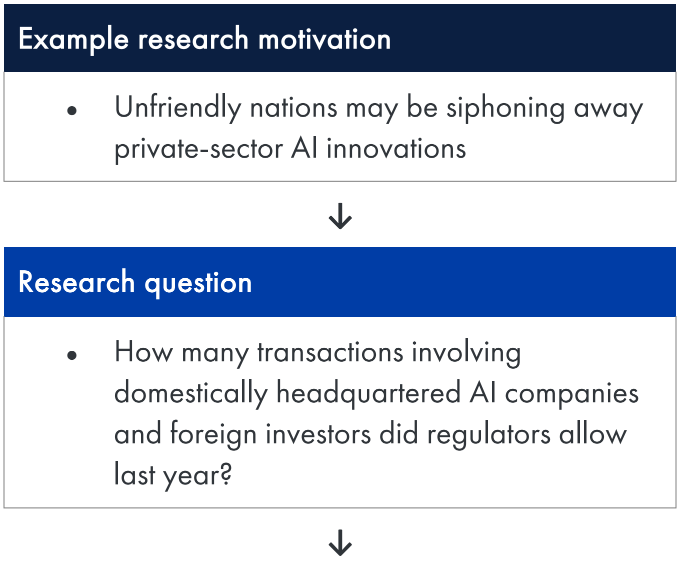
With both of these issues in mind, CSET is developing a flexible framework for identifying AI companies, coupled with the datasets and computing infrastructure to implement that framework. With these resources, researchers and policymakers will be able to identify "AI companies" that are relevant to their particular interests and analytic needs.

CSET's framework consists of a set of dimensions of AI relevance—for example, conducting AI research, developing products that involve or relate to AI, employing people with AI skills, applying AI to significant societal needs, and being recognized by experts as having AI capacity. The underlying infrastructure consists of datasets related to those dimensions of relevance. Along with these datasets, CSET is compiling a list of potential AI companies, and cleaning and structuring this list in order to reliably identify companies and their affiliates wherever they occur in the datasets.

Table 1: Example dimensions of AI relevance and associated datasets

Example dimension of AI relevance	Example datasets
Conducting AI-related research and development	Publication data, patent filings, researcher job postings
Developing products that involve or relate to AI	Business descriptions, corporate press releases, patent filings
Employing people with AI-related skills	Personnel databases, job postings
Applying AI to significant societal needs	Government procurement, grant and prize data
Being recognized by experts as having AI capacity	Expert-compiled lists, venture capital data, news mentions, survey data, grant and prize data

With these resources in place, analysts will be able to turn their conceptual definitions into concrete, specific parameters that can be applied to data and incorporated into policy, as in the following example:

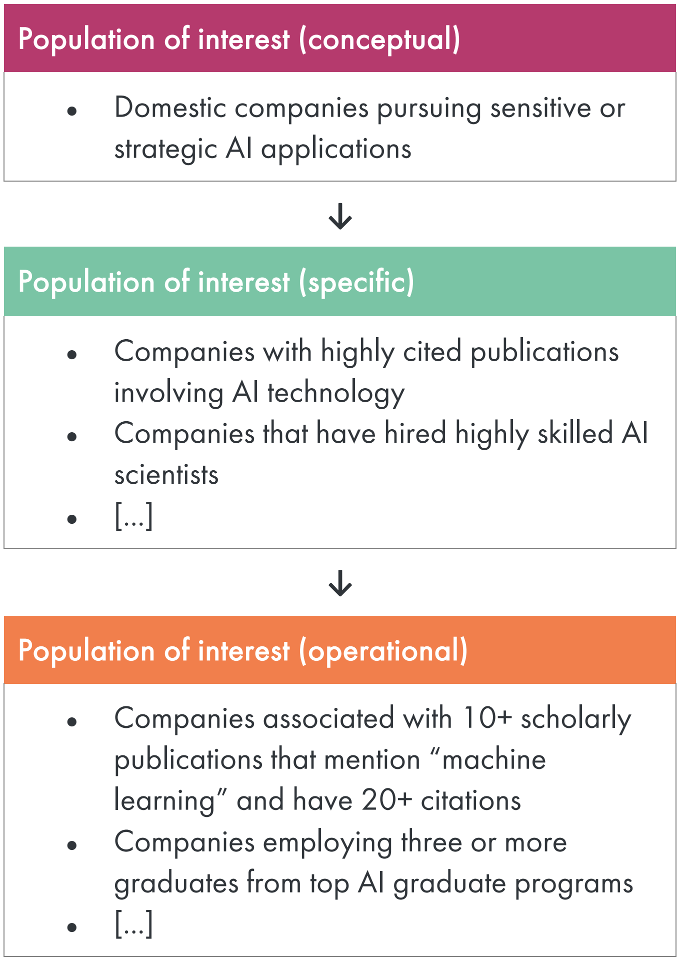


Example research motivation

- Unfriendly nations may be siphoning away private-sector AI innovations

Research question

- How many transactions involving domestically headquartered AI companies and foreign investors did regulators allow last year?



Population of interest (conceptual)

- Domestic companies pursuing sensitive or strategic AI applications

Population of interest (specific)

- Companies with highly cited publications involving AI technology
- Companies that have hired highly skilled AI scientists
- [...]

Population of interest (operational)

- Companies associated with 10+ scholarly publications that mention "machine learning" and have 20+ citations
- Companies employing three or more graduates from top AI graduate programs
- [...]

Identifying AI companies in practice: A proof of concept

To show how this framework functions, we built a test implementation using CSET data holdings related to AI publications, patents, and personnel. (The appendix provides source information.) We count companies' AI-related scholarly publications, their publications in the proceedings of top AI conferences, their AI-related patent families[3] with both applied and granted filings,[4] and the number of graduates each company is known to have hired recently from top U.S. AI PhD programs.[5] The test population is defined, somewhat arbitrarily, as all companies with at least two such hires—164 companies in total.[6] The full list of companies, their respective counts, and related metadata is available online.[7] The top 10 companies according to each count are as follows:

Table 2: Top 10 companies by different AI-related metrics—rank and count

Rank	By recent PhD hires from top U.S. AI programs		By AI publications, worldwide		By top AI conference papers, worldwide		By AI patent families since 2010, worldwide	
		Count		*Count*		*Count*		*Count*
1	Google[*]	220	Microsoft	1,651	Microsoft	313	IBM	9,047
2	Facebook	115	Google[*]	978	Google[*]	138	Microsoft	6,115
3	Microsoft	106	Siemens	977	Siemens	101	Google[*]	4,632
4	Amazon	86	Samsung	730	Yahoo	88	Samsung	3,358
5	Apple	57	Philips	547	Alibaba	81	Siemens	3,280
6	Alphabet[†]	46	Yahoo	441	NEC	50	Intel	2,603
7	IBM	38	Mitsubishi	330	Facebook	33	Sony	2,525
8	Uber	32	Intel	321	Amazon	31	Philips	2,337
9	Intel	28	Bosch	303	LinkedIn	30	NEC	2,334
10	DeepMind	24	Huawei	233	Mitsubishi	28	Baidu	2,100

Table 2 shows how different metrics can yield significantly different groups of "AI companies." Some companies, notably Google and Microsoft, rate highly on all three metrics. Others, such as Facebook, Apple, and Amazon, hire large numbers of elite AI PhD students but have less patenting and publication activity. The full dataset available online also shows how many companies outside "Big Tech" have a significant AI footprint according to one or more metrics. For example, companies with multiple PhD hires from elite U.S. AI programs include Ford Motor (11 hires), Walmart (8), Northrop Grumman (5), and 3M (5), among others, and blue-chip firms like Mitsubishi, Siemens, and Philips rank highly in AI publications.

To further demonstrate how different metrics can capture different dimensions of AI relevance, in Tables 3 and 4, we compare two metrics in greater detail: patent families and top conference papers. A company that publishes AI-related papers in the proceedings of highly competitive, theoretically oriented

[*] Excludes DeepMind

[†] Excludes Google and DeepMind

academic conferences may be more likely to be involved in pre-commercial research. This might especially interest science and technology policymakers. In contrast, metrics such as patent counts may more directly reflect companies' capacity to apply and commercialize AI technology, and may be more relevant to trade and economic policymakers.

Table 3 shows how the top-patenting companies rank according to publication count, and Table 4 shows how the top-publishing companies rank according to patent count. Some companies, such as IBM, Philips, and Sony, are much further ahead in patents than top papers; the reverse is true for others, such as Alibaba, Amazon, Facebook, and LinkedIn. All of these could plausibly be considered "AI companies," but these results suggest they may interact with the AI R&D process in different ways.

As this brief example shows, by combining different datasets and using them to develop multiple metrics of AI involvement, analysts and policymakers can examine them and their peers holistically, then choose the metrics and companies most relevant to the specific issues at hand.

Table 3: Top 10 companies by count of AI patent families, with corresponding rank by count of top AI conference papers

Company	Rank (AI patent families since 2010, worldwide)	Rank (top AI conference papers, worldwide)
IBM	1/164	30/164
Microsoft	2	1
Google (excluding DeepMind)	3	2
Samsung	4	12
Siemens	5	3
Intel	6	17
Sony	7	36
Philips	8	33
NEC	9	6
Baidu	10	13

Table 4: Top 10 companies by count of top AI conference papers, with corresponding rank by count of AI patent families

Company	Rank (top AI conference papers, worldwide)	Rank (AI patent families since 2010, worldwide)
Microsoft	1/164	2/164
Google (excluding DeepMind)	2	3
Siemens	3	5
Yahoo	4	[no AI patent families found]
Alibaba	5	23
NEC	6	9
Facebook	7	21
Amazon	8	30
LinkedIn	9	[no AI patent families found]
Mitsubishi	10	17

Next steps

As we work toward a comprehensive, multidimensional framework for identifying AI companies, CSET will incorporate new metrics, datasets, and companies into our research. We welcome potential collaborators and feedback on our research agenda. Please contact zachary.arnold@georgetown.edu to discuss.

Acknowledgments

Thanks to James Dunham, Alexandre Georgieff, Jennifer Melot, Igor Mikolic-Torreira, Anna Milanez, Dewey Murdick, Karine Perset, and Alexandra Vreeman for helpful comments and technical support.

Document Identifier: doi: 10.51593/20200018

Appendix: Data sources

Publication data is derived from the Dimensions (Digital Science) dataset. We searched for AI publications, as identified by CSET's publication classifier, and identified companies according to their GRID identifiers (when available) or regular expression-based string queries (when not). "Top publications" counts are based on publications in the proceedings of nine top AI conferences,[8] as designated by CSRankings. (Four additional conferences[9] are included in CSRankings but lack data in Dimensions, and were excluded from our analysis.) Our analysis covers all publications in Dimensions, regardless of date or location of publication.

Patent data is provided by 1790 Analytics and covers patents applied for worldwide from 2010 to the present. We count patent families identified by 1790 as related to AI, and identified companies using regular expression-based string queries. We count families including at least one granted patent as well as families including patents that were applied for but not granted, given the large number of recent patent applications in the AI field not yet examined. Our calculations may differ from other patent analyses due to different underlying datasets, different means of identifying AI-related patents, and different treatment of patents not yet granted.[10]

Hiring data was assembled as part of CSET's *Keeping Top AI Talent in the United States* study. We analyze hiring data for graduates with PhD dissertations dated between 2014 and 2019, inclusive. Further details on this dataset can be found in Section A of the study's appendix.

Endnotes

[1] "Center for Security and Emerging Technology: AI Companies Brief," GitHub, https://github.com/georgetown-cset/ai-companies-brief.

[2] See, e.g., Cheryl Xiaoning Long and Jun Wan, "China's patent promotion policies and its quality implications," *Science and Public Policy*, vol. 46, no. 1 (February 2019): 91-104, https://academic.oup.com/spp/article-abstract/46/1/91/5004405.

[3] "A patent family is a collection of patent applications covering the same or similar technical content." "Patent Families," European Patent Office, accessed June 10, 2020, https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families.html. See generally Patrick Thomas and Dewey Murdick, "Patents and Artificial Intelligence – A Data Primer," Center for Security and Emerging Technology (forthcoming 2020).

[4] In other words, we count families including at least one granted patent as well as families including patents that were applied for but not granted, given the large number of recent patent applications in the AI field which have not yet been examined.

[5] Specifically, we analyze hiring data for graduates with PhD dissertations dated between 2014 and 2019, inclusive.

[6] When a company and its subsidiary both meet this criterion, we analyze the subsidiary as a separate company if it originated independently and was acquired, but not if it originated as a division of the parent. For example, we treat LinkedIn as separate from its parent Microsoft, but not Microsoft Research Asia, and we treat DeepMind as separate from its parent Google, but not Google Brain. Alphabet, a holding company, is an unusual case; we aggregate it with its non-Google subsidiaries only.

[7] "Center for Security and Emerging Technology: AI Companies Brief," GitHub, https://github.com/georgetown-cset/ai-companies-brief.

[8] The International Joint Conference on Artificial Intelligence (IJCAI), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV), International Conference on Knowledge Discovery and Data Mining (SIGKDD), Annual Meeting of the Association for Computational Linguistics (ACL), North American Chapter of the Association for Computational Linguistics (NAACL), Conference on Empirical Methods in Natural Language Processing (EMNLP), Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), and International World Wide Web Conference (WWW).

[9] The AAAI Conference on Artificial Intelligence (AAAI), IEEE International Conference on Computer Vision (ICCV), International Conference on Machine Learning (ICML), and Annual Conference on Neural Information Processing Systems (NeurIPS).

[10] Compare, e.g., "Live data from partners," *OECD AI Policy Observatory*, Organisation for Economic Co-operation and Development, accessed June 26, 2020, https://oecd.ai/data-

from-partners (identifying AI patents using the Microsoft Academic Graph dataset and classification algorithm).