September 2021

Small Data's Big AI Potential

CSET Issue Brief

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Husanjot Chahal
Helen Toner
Ilya Rahkovsky

Executive Summary

This issue brief provides an introduction to and overview of "small data" artificial intelligence approaches—that is, approaches that help with situations where little or no labeled data is available and that reduce our dependency on massive datasets collected from the real world. According to the conventional understanding of AI, data is an essential strategic resource and any meaningful progress in cutting-edge AI techniques requires large volumes of data. This overemphasis on "big data" ignores the existence and overshadows the potential of the approaches we describe in this brief, which do not require massive datasets for training.

We present our analysis in two sections. The first introduces and classifies the main small data approaches, which we conceptualize in terms of five rough categories—transfer learning, data labeling, artificial data, Bayesian methods, and reinforcement learning—and lays out reasons for why they matter. In doing so, we aim not only to point out the potential benefits of using small data approaches, but also to deepen nontechnical readers' understanding of when, and how, data is useful for AI. Drawing from original CSET datasets, the second section presents some exploratory findings evaluating the current and projected progress in scientific research across small data approaches, outlining which country leads, and the major sources of funding for this research. We conclude the following four key takeaways based on our findings:

a) Artificial intelligence is not synonymous with big data, and there are several alternative approaches that can be used in different small data settings.

b) Research into transfer learning is growing especially rapidly (even faster than the larger and better-known field of reinforcement learning) making this approach likely to work better and be more widely used in the future than it is today.

c) The United States and China are competing closely in small data approaches, with the United States leading in the two largest categories of reinforcement learning and Bayesian

methods, and China holding a small but growing lead in the fastest-growing category of transfer learning.

d) Tentatively, transfer learning may be a promising target for greater U.S. government funding, given its smaller share of investments in small data approaches relative to investment patterns across AI as a field.

Table of Contents

Introduction

Conventional wisdom says that cutting-edge artificial intelligence is dependent on large volumes of data. According to this conception of AI, data is therefore an essential strategic resource, and how much data a country (or company) has access to is seen as a key indicator of AI progress. This understanding of data's role in AI is not completely inaccurate—many current AI systems do make use of large amounts of data. But policymakers will be led astray if they believe that this is an enduring truth about all AI systems. An overemphasis on data ignores the existence—and underestimates the potential—of several AI approaches that do not require massive labeled datasets or data collected from real-world interactions. In this brief, we name these as "small data" approaches.

What we are calling "small data" is not a clean-cut category, and therefore does not have a single, formal, agreed upon definition. Academic writings discuss small data in relation to the application area under consideration, often tying it to the size of the sample, for instance kilobytes or megabytes versus terabytes of data.[1] Popular media articles attempt to describe small data in relation to varied factors like its usability and human comprehension, or as data that comes in a volume and format that makes it accessible, informative, and actionable, especially for business decisions.[2] Many references to data often end up treating it as an all-purpose resource. However, data is not fungible, and AI systems in different domains call for distinct types of data and distinct types of approaches depending upon the problem at hand.[3]

This study describes small data with a policymaker's perspective in mind. Government actors are often considered potentially strong players in AI because of the nature of real-world interactions they have access to and their ability to collect massive amounts of data on it—examples include climate monitoring data, geological surveys, border control, social security, voter registration, vehicle and driver records, among others. Most comparisons of countries' AI competitiveness label China as having a unique advantage based on its access to more data, citing its large population, data collection capabilities, and lack of privacy protections.[4] Part of our motivation in writing this paper is to shed light on a set of technologies that makes this less true than is often assumed.[5]

Finally, it is sometimes suggested that government organizations will only be able to benefit from the AI revolution if they can digitize, clean, and label large amounts of data. While this suggestion holds merit, it is inaccurate to think of all progress in AI as contingent upon such conditions. This belief overshadows the idea that the future of AI may not only be about big data, as well as that AI innovation in the government sector (and beyond) can still happen without massive investments in big data infrastructures.

In what follows, we aim not only to point out the potential benefits of using small data approaches, but also to deepen nontechnical readers' understanding of when and how data is useful. This brief can be viewed as a primer on small data approaches or approaches that could minimize reliance on "big data."[6] This analysis is split into two sections. The first section is a technical explainer on what "small data" approaches are, which categories form a part of these approaches, and why they matter. It lays out the conceptual foundation for the data analysis drawn in section two. The second section draws from original CSET datasets, specifically our merged corpus of scholarly literature capturing over 90 percent of the world's scholarly output, to present our findings on small data approaches across three pillars—research progress, national competitiveness, and funding. We seek to examine the current and projected progress in scientific research across these approaches, as well as to identify which country leads, and the major sources of funding for the research studied. We conclude this brief with four key takeaways based on our findings.

Classifying "small data" approaches

The research in this paper is broken down in terms of five rough categories of "small data" approaches: a) transfer learning, b) data labeling, c) artificial data generation, d) Bayesian methods, and e) reinforcement learning. These categories, which we describe in more detail below, are imperfect. AI and machine learning research incorporates a wide array of different methods, approaches, and paradigms used to solve many different types of problems, and therefore defies easy categorization. Our aim in delineating the categories below is to give the reader a sense of some of the rough conceptual approaches that make it possible to train AI systems without large, pre-labeled datasets. The categories we use are not cleanly separable in practice, and they are neither mutually exclusive nor collectively exhaustive.

Transfer learning works by first learning how to perform a task in a setting where data is abundant, then "transferring" what it has learned there to a task where much less data is available. This is useful in settings where only a small amount of labeled data is available for the problem of interest, but a large amount of labeled data is available for a *related* problem.

For example, someone building an app to identify rare bird species might only have a handful of photos of each bird, each labeled with its species. To use transfer learning, they could first train a basic image classifier using a much larger, more generic image database such as ImageNet, which has millions of images labeled according to thousands of categories. Once that classifier could distinguish dogs from cats, flowers from fruit, and sparrows from swallows, they could feed it the much smaller dataset of rare birds. The model could then "transfer" what it already knows about how to put images into categories, using that knowledge to learn the new task (identifying rare bird species) from much less data.

Data labeling is a category of approaches that starts with limited *labeled* data but abundant *unlabeled* data. Approaches in this category use a range of methods to make sense of the unlabeled data available, such as automatically generating labels (automated labeling) or identifying datapoints for which labels would be especially useful (active learning).

For example, active learning has been used in research on skin cancer diagnoses. One image classification model was initially trained on 100 photos labeled according to whether they depicted skin cancer or healthy skin.[7] The model then had access to a larger set of potential training images, from which it could choose 100 additional photos to be labeled and added to its training data. In order to learn as much as possible from the available data, the model was designed to choose the additional photos to be labeled based on which images would be most informative in learning to distinguish photos of healthy skin from photos of skin cancer.

Artificial data generation is a category of approaches that seeks to maximize how much information can be extracted from a small amount of data by creating new datapoints or other related techniques. This can range from simply making small changes to existing data (e.g., cropping or rotating images in an image classification dataset) to more complex methods that aim to infer the underlying structure of the available data and extrapolate from there.

A simple example is that computer vision researchers have been able to use computer-aided design (CAD) software—a widely available tool used in industries from shipbuilding to advertising—to generate realistic, 3D images of everyday objects, then use those images to augment an existing image dataset.[8] Approaches like this are more feasible when a separate source of information about the data of interest exists—in this case, crowdsourced CAD models. In other cases, more complex methods may be required. In general, data generation requires strong assumptions of one kind or another to be made about the data in question, and how useful the generated data will be depends on how valid those assumptions are.

The ability to generate additional data is not only useful when working with small datasets. In cases where details of any individual piece of data might be sensitive (for example, an individual's health record) but the overall distribution of data is of interest to researchers, synthetic data can be used to obscure private information by making random changes to the data to make it less identifiable.[9]

Bayesian methods are a large class of approaches to machine learning and statistics that have two features in common**.** First, they try to explicitly incorporate information about the structure of the problem—so-called "prior" information—into their approach to solving it.[10] This contrasts with most other approaches to machine learning, which tend to make minimal assumptions about the problem in question. By incorporating this “prior” information before improving further based on the available data, Bayesian methods are better suited to some contexts where data is limited, but it is possible to write out information about the problem in a useful mathematical form. Second, Bayesian approaches focus on producing well-calibrated estimates of the uncertainty of their predictions. This is helpful in settings where there is limited data availability because Bayesian approaches to estimating uncertainty make it easier to identify datapoints that, if collected, would be most valuable in reducing uncertainty.

As an example of Bayesian work using small data, Bayesian approaches have been used to monitor global seismic activity, which is relevant both for detecting earthquakes and for verifying nuclear treaties. By developing a model that starts with incorporating prior knowledge from seismology, researchers can make the most of available data to improve the model from there.[11]

The family of Bayesian methods is a large one, and does not consist solely of approaches that are especially good at working with small datasets. For simplicity, we have erred on the side of inclusiveness for this study, though that likely means that some of the research included in this category uses large datasets.

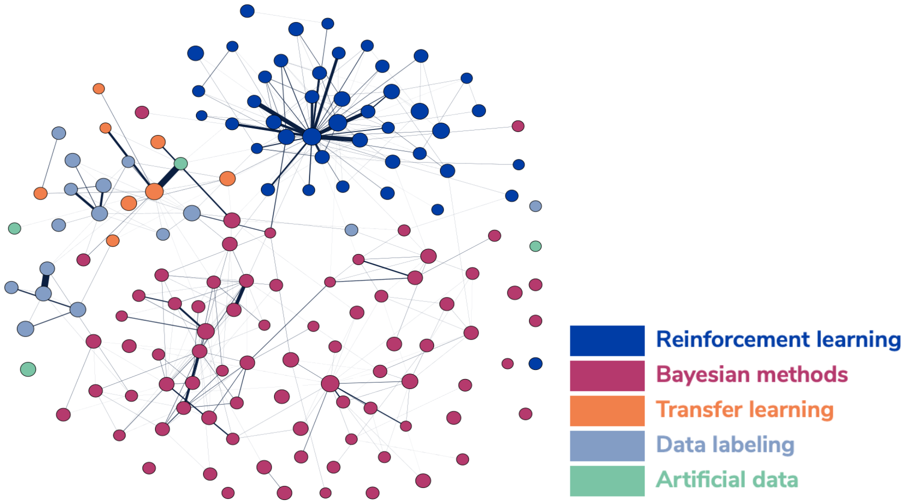
Reinforcement learning is a broad term that refers to machine learning approaches in which an agent (the computer system) learns how to interact with its environment via trial and error. Reinforcement learning is often used to train game playing systems, robots, and autonomous vehicles.

For example, reinforcement learning has been used to train AI systems that learned to play video games—from simple arcade games such as Pong, to strategy games such as Starcraft. In each case, the system starts out knowing very little (or nothing) about how to play the game, but gradually learns by trying things out and seeing what creates a positive reward signal. (In the example of video games, the reward signal is often in the form of points scored by the player.)[12]

Reinforcement learning systems often end up learning from large amounts of data and requiring immense computing resources, so they may seem like an unintuitive category to include here. We include them nonetheless because the data they use is generally produced while the system is training—often in a simulated environment—rather than being collected and labeled in advance. In reinforcement learning problems, the agent's ability to interact with its environment is critical.

Figure 1 is a representation of how these different areas connect to each other. Each dot represents a research cluster (i.e., a group of papers) that we identified as belonging to one of the categories above (see Appendix for methodological details). The thickness of the lines connecting one research cluster to another represents the strength of citation linkages between the two research clusters.[13] Absence of a line indicates no citation linkage. We can see that while clusters do tend to be most connected with other clusters in the same category, there are also plenty of connections between clusters of different categories. The figure also shows that the clusters we identified under "reinforcement learning" form an especially coherent grouping, whereas the "artificial data" clusters are much more scattered.

Figure 1. Network graph of small data research clusters



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Significance of small data approaches

Approaches to AI that do not rely on large, pre-collected, labeled datasets offer a number of advantages over more data-intensive approaches. Among other factors, these approaches can:

Reduce capability differentials between large and small entities

The growing value of large datasets for many AI applications has created concern about disparity in different organizations' ability to collect, store, and process the required data. This dynamic has the potential to create a gap between the AI "haves," such as big tech companies, and "have-nots," depending upon who can afford to meet these demands. If approaches such as transfer learning, automated labeling, Bayesian methods, etc. make it possible to apply AI with less data, the barriers for entry for smaller organizations will be reduced on the data front.[14] This can contribute to reducing the capability differential between larger and smaller entities.[15]

Reduce the incentive to collect large amounts of personal data

Several surveys have indicated that a majority of Americans feel that AI will significantly reduce personal privacy.[16] Such concerns emerge from the idea that big tech companies continue to collect more and more consumer data linked with individual identities to train their AI algorithms. Certain small data approaches have the potential to lessen such concerns by reducing the need to collect real-world data for training machine learning models. In particular, approaches that enable the generation of new data artificially (like synthetic data generation), or that use simulations for training algorithms, either do not rely on data generated by individuals or have the potential to synthesize the data to remove sensitive personally identifiable attributes.[17] Although this does not mean that all privacy concerns would be solved, by reducing the need to collect large amounts of real-world data, such approaches could enable the use of machine learning in a way that makes concerns regarding large-scale collection, use, or disclosure of consumer data less acute.[18]

Bolster progress in areas with access to few data points

Much recent progress in AI has been enabled by the explosion of available data. For many important problems, however, little or no data may exist that can be fed into an AI system. For instance, imagine building an algorithm predicting disease risk for a set of people that do not have electronic health records, or forecasting the likelihood of eruptions in volcanoes with long eruptive recurrence.[19] Small data approaches can provide us with a principled way to deal with this lack or absence of data. It can do so by transferring knowledge from a related problem, making use of both labeled and unlabeled data. Small data can also help by using the few data points we have in hand to create more, leveraging prior knowledge about the domain in question, or venturing into a new domain altogether by building simulations or encoding structural assumptions.

Circumvent dirty data problems

Certain small data approaches can benefit big organizations where, although data may exist, it is a long way from being clean, neatly structured, and ready for analysis. The U.S. Department of Defense, for instance, has a large amount of "dirty data" as a result of siloed data infrastructures and legacy systems, which demands time-consuming and labor-intensive processes of data cleaning, labeling, and organizing.[20] Approaches in the data labeling category can make it easier to work with large amounts of unlabeled data by automatically generating labels, for example. Transfer learning, Bayesian methods, or artificial data approaches can significantly reduce the scale of dirty data problems by shrinking the amount of data that needs to be cleaned, relying instead on related datasets, structured models, and synthetic data, respectively.[21]

More generally, we also believe that it is important for policymakers whose work relates to AI to have a clear understanding of the role that data plays—and does not play—in AI development. The abovementioned factors will not apply to all of the approaches we describe. For example, reinforcement learning does generally require large amounts of data, but that data is generated during the training process (e.g., as the AI system moves a robot arm or navigates around a virtual environment) rather than being gathered in advance.

Findings

To explore how research into small data approaches is progressing, we used CSET's research cluster dataset to identify research related to the five categories described above (transfer learning, data labeling, artificial data generation, Bayesian methods, and reinforcement learning). Research clusters are groups of scientific research articles connected by citation links—instances where a researcher is communicating that they are using ideas, methods, results, or in any way building upon the work of other researchers.[22]

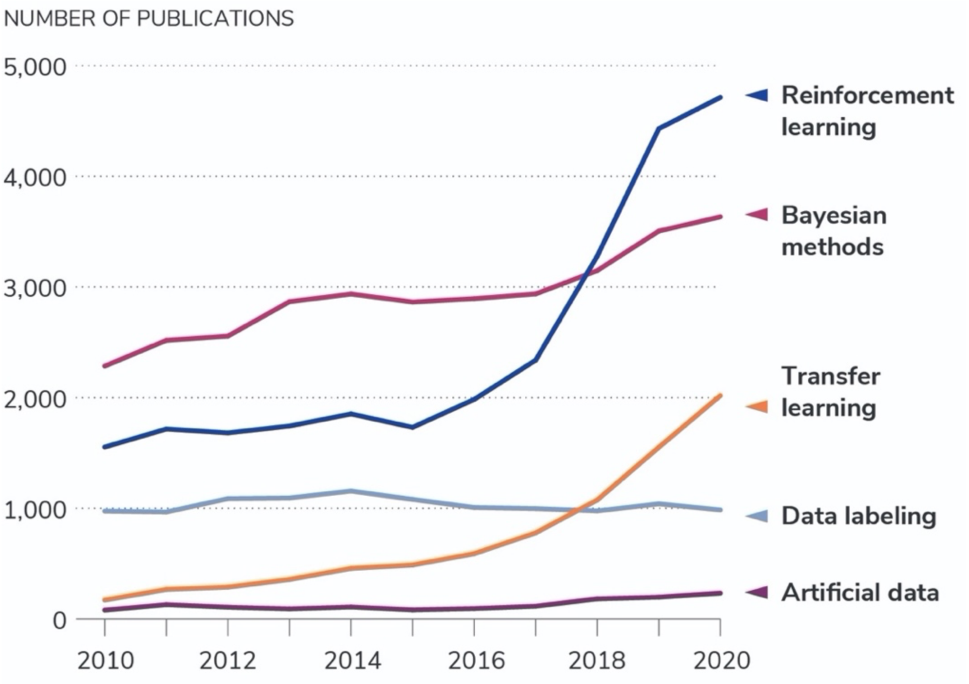
For our analysis, we identified 150 research clusters as belonging to one of our five categories. For comparison, the dataset includes 735 AI clusters.[23] The 150 identified clusters comprised about 80,324 papers, which were drawn from CSET's merged corpus of scholarly literature, containing over 90 percent of the world's scholarly output. To determine which papers fall into our "small data" categories, we first collaborated with technical experts to define a set of keywords that related to each of our five categories. Next, we searched for clusters where any of those keywords were among the top phrases extracted from papers in the cluster. Lastly, we manually excluded clusters that were clearly unrelated to small data. Once we had identified the 150 clusters we wanted to use, each of which was associated with one of our five categories, we treated all papers within those research clusters as belonging to the corresponding category.[24] In following this approach, we have attempted to balance accuracy and inclusiveness, but it is very possible that we missed including relevant papers that do not cite authors in their research community as much, or that some research papers we included may be connected to a cluster due to citations but may not directly address the topic under consideration. Therefore, we encourage readers to consider the analysis in the sections below as exploratory rather than definitive. See Appendix A for more details on our methodology.

In the subsections below, we present our findings regarding all papers identified in relevant research clusters across three pillars—research progress, national competitiveness, and funding. Through this analysis we hope to examine the current and projected progress in scientific research in developing these approaches, which country leads, and the major sources of funding for this research.

Research progress

In terms of research volume, our five categories of "small data" approaches have had very different trajectories over the last decade. As shown in Figure 2, reinforcement learning and Bayesian methods are the two largest categories in terms of number of papers. While the number of papers in Bayesian clusters grew steadily over the decade, reinforcement learning clusters only grew starting in 2015, then saw especially rapid growth between 2017 and 2019. This is likely due to revolutionary advances in deep reinforcement learning that had suffered from technical challenges until 2015.[25] By contrast, the number of papers published annually in clusters for artificial data generation and data labeling research stayed fairly low across the decade. Finally, the transfer learning category started out small in 2010, but by 2020 had grown substantially.

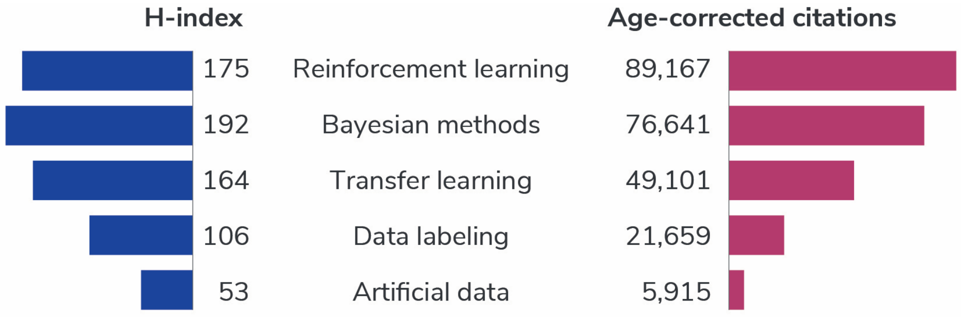
Figure 2. Trends in small data publications, 2010–2020



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Of course, the sheer number of publications does not account for paper quality. We consider two metrics to assess the quality of papers in each category's clusters: h-index and age-corrected citations. H-index is a commonly-used metric that captures the publication activity and total citation impact of a collection of papers—in our case, the papers in clusters attributed to each category.[26] One limitation of the h-index, however, is that it does not account for paper age (i.e., the fact that older papers have had more time to accumulate citations). Therefore, h-index undervalues groups of papers where the most influential papers are newer and have not yet gathered citations.[27] To adjust for this, figure 3 also depicts age-corrected citations. As can be seen in the figure, on h-index alone, reinforcement learning and Bayesian methods are roughly even, but after accounting for the age of papers, reinforcement learning comes out on top. This means that for the research clusters that we identified, the cumulative impact of Bayesian methods appears higher, but reinforcement learning stands out for its relatively recent surge in paper production and citation impact.

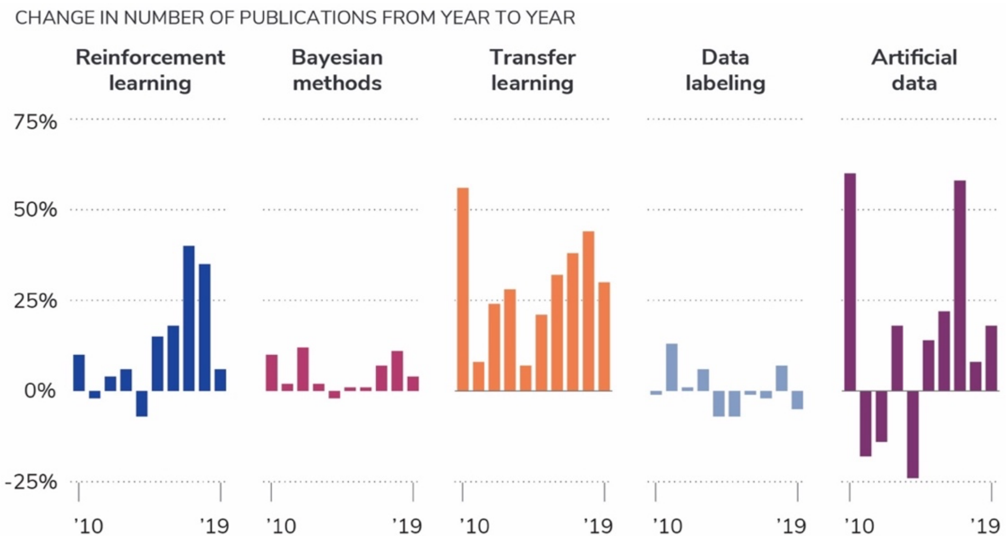
Figure 3. H-index and age-corrected citations by category, 2010–2020



Source: CSET merged corpus of scholarly literature, as of February 1, 2021.

However, it would be wrong to assume that reinforcement learning has grown the fastest in the past decade. Looking more closely at the growth of each category over time, Figure 4 makes clear that transfer learning has seen the most consistent growth between 2011 and 2020, with the highest growth in all but two years. The chart also shows the growth that artificial data generation has seen over the last five years, which is less apparent in Figure 3 due to the low number of total papers in this category. However, it has also seen the largest dips in growth figures between 2012 and 2015 making it hard to draw specific conclusions for this category's growth trajectory.

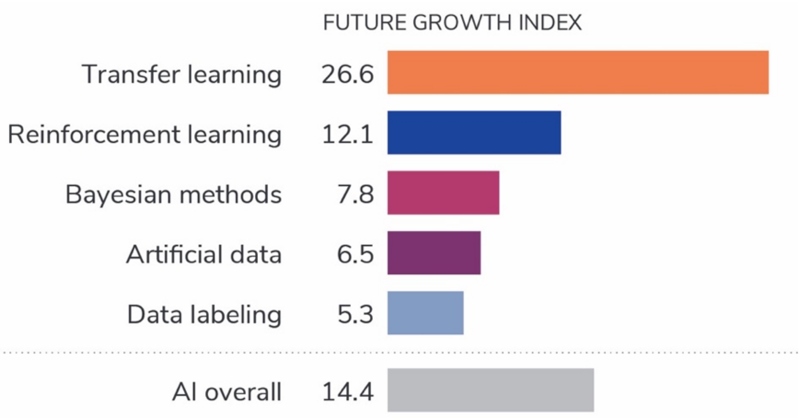
Figure 4. Year-on-year growth by category, 2011–2020



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Figure 5 compares the forecast three-year growth for each category according to a CSET-developed forecasting model,[28] with an additional category of "AI overall" papers included as a benchmark to compare against.[29] As shown in the figure, transfer learning is the only category forecasted to grow faster than AI research as a whole, far outstripping all the other categories and consistent with its ongoing growth in previous years.

Figure 5. Growth forecast for 2023 by category



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

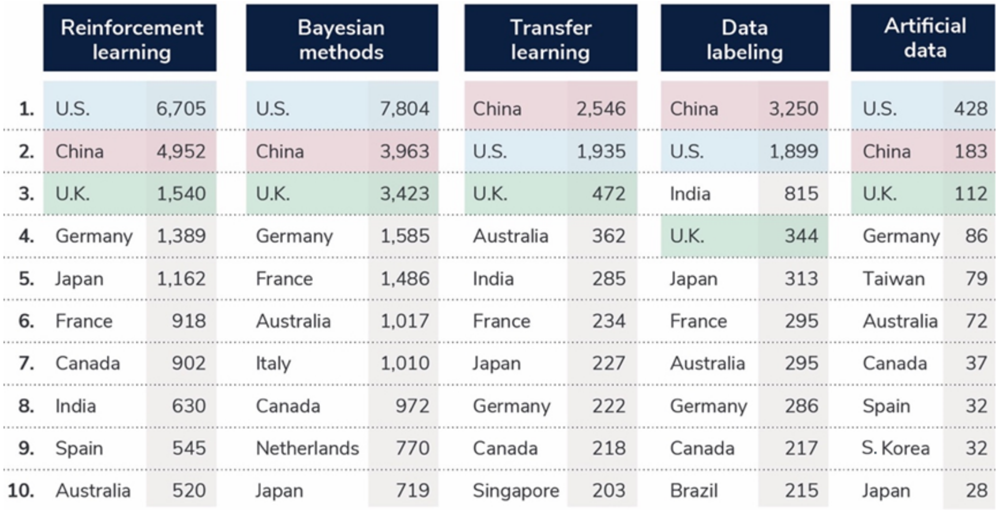
Note: Future growth index is calculated based on CSET forecasts of research cluster growth. See Appendix A for more details on methodology.

National competitiveness

In this section we explore national competitiveness in small data approaches by looking at research progress made by the top 10 countries globally in each of these approaches.[30] We use simple measures such as number of papers published and number of age-adjusted citations to gain an initial picture of countries' relative standing in each category, but we encourage readers to explore other indicators to fully understand a country's potential in small data approaches.

Table 1 displays the total number of papers produced by category for the top 10 countries in terms of small data publications. Consistent with the results for AI research overall, China and the United States are the top two producers of papers in clusters we identified as containing research related to small data, closely followed by the United Kingdom. China leads in the total number of scholarly publications in data labeling and transfer learning approaches, whereas the United States is a leader in Bayesian methods, reinforcement learning, and artificial data generation.

Table 1. Number of publications by category for top 10 countries globally



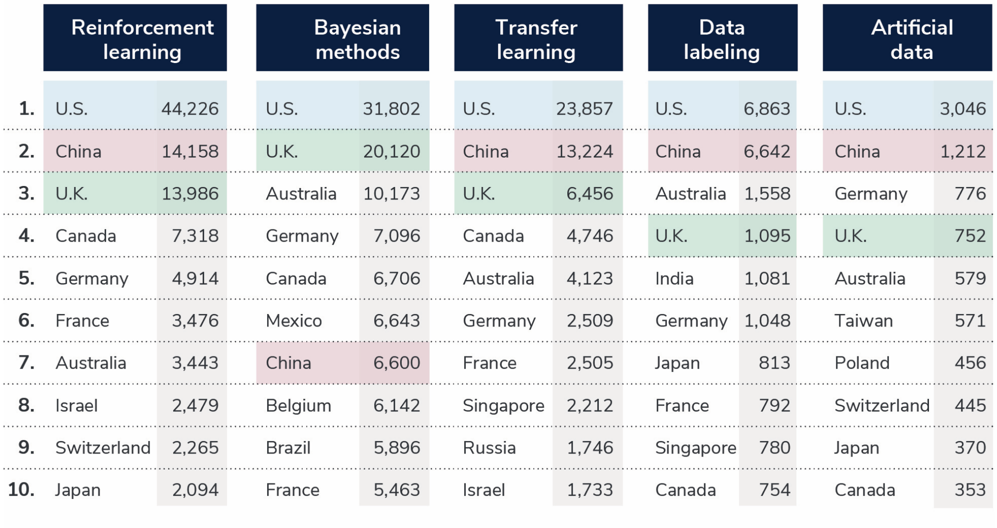
	Reinforcement learning		Bayesian methods		Transfer learning		Data labeling		Artificial data	
1.	U.S.	6,705	U.S.	7,804	China	2,546	China	3,250	U.S.	428
2.	China	4,952	China	3,963	U.S.	1,935	U.S.	1,899	China	183
3.	U.K.	1,540	U.K.	3,423	U.K.	472	India	815	U.K.	112
4.	Germany	1,389	Germany	1,585	Australia	362	U.K.	344	Germany	86
5.	Japan	1,162	France	1,486	India	285	Japan	313	Taiwan	79
6.	France	918	Australia	1,017	France	234	France	295	Australia	72
7.	Canada	902	Italy	1,010	Japan	227	Australia	295	Canada	37
8.	India	630	Canada	972	Germany	222	Germany	286	Spain	32
9.	Spain	545	Netherlands	770	Canada	218	Canada	217	S. Korea	32
10.	Australia	520	Japan	719	Singapore	203	Brazil	215	Japan	28

Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

It is interesting to note that, other than the United States and China, all countries in the top 10 ranking for all of small data research are U.S. allies or partners, with countries like Russia notably absent from the list.[31] However, this trend in data here could also be due to the fact that we are counting papers with multiple authors belonging to different countries multiple times, and papers where researchers in the United States and allied countries collaborated are reflecting a higher count individually due to double counting. Our analysis of coauthorships in these papers supports this assessment.[32]

Paper citations are often used as a measurement of research quality and impact,[33] and our findings indicate that China's high volume of research may not be high-quality research across all small data categories. As shown in Table 2, China ranks below the United States across all approaches when looking at age-corrected citations (which can be roughly interpreted as the number of citations per year). China's rank is second on age-corrected citations across all small data categories except for Bayesian methods, where its position falls further down to seventh. This implies that while China may be producing a lot of papers on Bayesian methods, the quality and impact of its research in this category suffers the most in comparison to other approaches. The United States leads on age-corrected citations globally across all approaches.[34]

Table 2. Number of age-corrected citations by category for top 10 countries globally

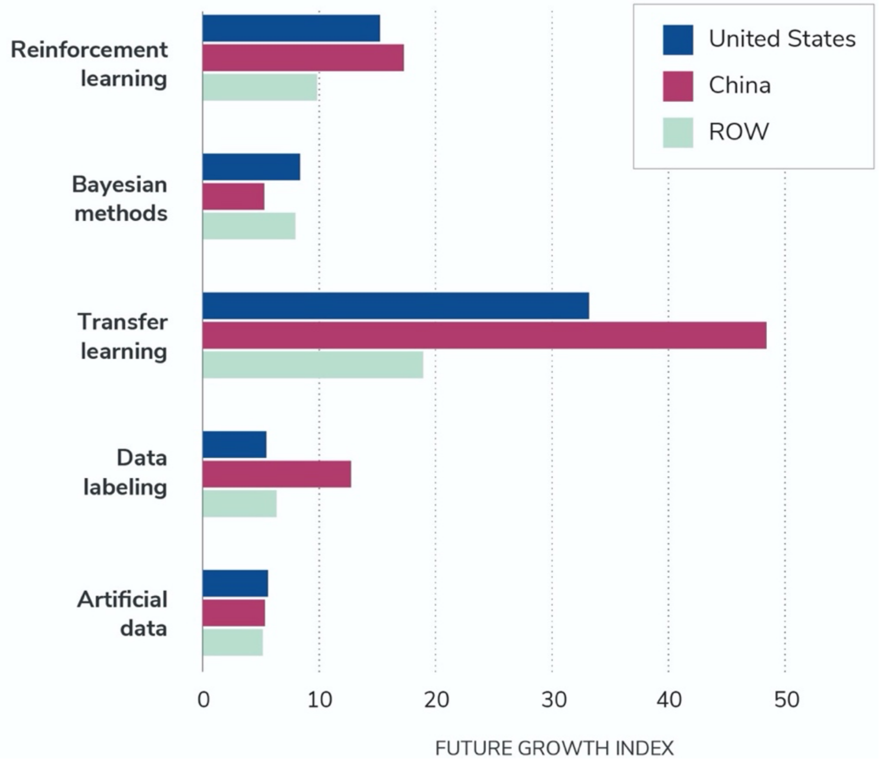


	Reinforcement learning		Bayesian methods		Transfer learning		Data labeling		Artificial data	
1.	U.S.	44,226	U.S.	31,802	U.S.	23,857	U.S.	6,863	U.S.	3,046
2.	China	14,158	U.K.	20,120	China	13,224	China	6,642	China	1,212
3.	U.K.	13,986	Australia	10,173	U.K.	6,456	Australia	1,558	Germany	776
4.	Canada	7,318	Germany	7,096	Canada	4,746	U.K.	1,095	U.K.	752
5.	Germany	4,914	Canada	6,706	Australia	4,123	India	1,081	Australia	579
6.	France	3,476	Mexico	6,643	Germany	2,509	Germany	1,048	Taiwan	571
7.	Australia	3,443	China	6,600	France	2,505	Japan	813	Poland	456
8.	Israel	2,479	Belgium	6,142	Singapore	2,212	France	792	Switzerland	445
9.	Switzerland	2,265	Brazil	5,896	Russia	1,746	Singapore	780	Japan	370
10.	Japan	2,094	France	5,463	Israel	1,733	Canada	754	Canada	353

Source: CSET merged corpus of scholarly literature, as of February 1, 2021.

Figure 6 looks at three-year growth forecasts broken down by country. The most notable finding here is how much higher Chinese growth in transfer learning approaches is projected to be, relative to the United States and the rest of the world. If accurate, this forecast would imply that China is likely to pull further ahead in transfer learning, at least in terms of the number of papers published.

Figure 6. Growth forecast for 2023 by category for the United States, China, and the rest of the world (ROW)



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

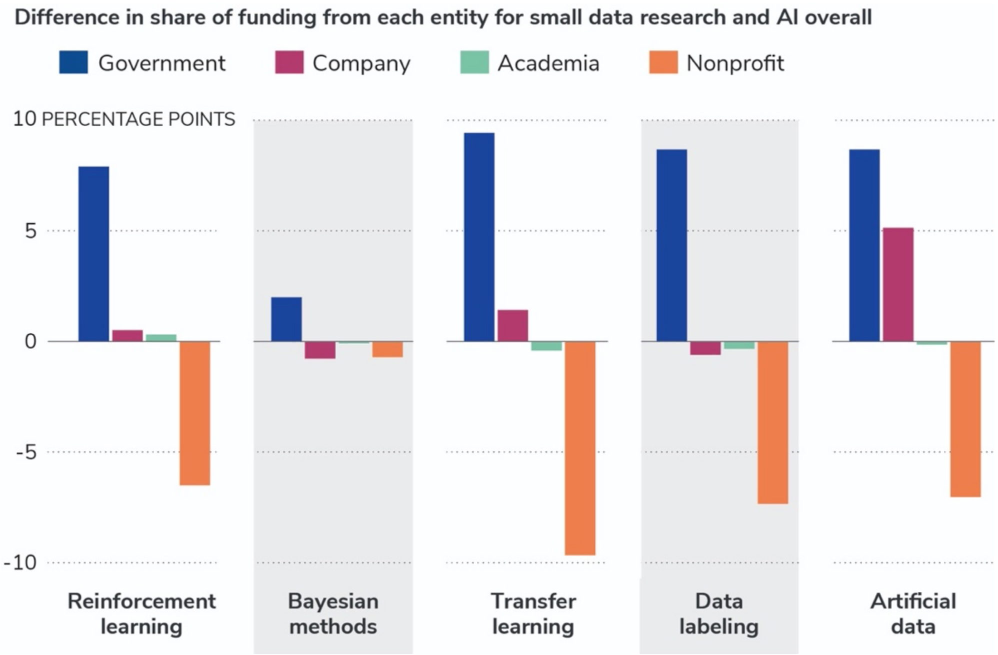
Note: Future growth index is calculated based on CSET forecasts of cluster growth. See Appendix A for more details on methodology.

Funding

We analyzed funding data available for small data approaches to obtain estimates of the type of entities that fund papers in the research clusters that we identified as belonging to these approaches. An important caveat to the findings presented here is that we only had funding information for around 20-30 percent of the papers under consideration, though we do not have reason to believe that there are systematic differences between papers with and without funding data available.[35]

Across disciplines, and among government, companies, academia, and nonprofit organizations, government actors tend to be the biggest funders of research, while authors are most commonly affiliated with academia. Keeping this in mind, we compared the results of small data research with AI research overall to see how much it differed.[36] It was interesting to note that globally, government funding makes up a larger share of funding for the clusters we identified as relating to small data approaches than in AI as a whole. As shown in Figure 7, across all five categories, the share of government funding is disproportionately high when compared with the funding breakdown for AI research overall. We also observe that nonprofits represent a smaller proportion of funding for small data research than they normally do for the rest of AI. The funding patterns for Bayesian methods are most similar to those for AI overall.

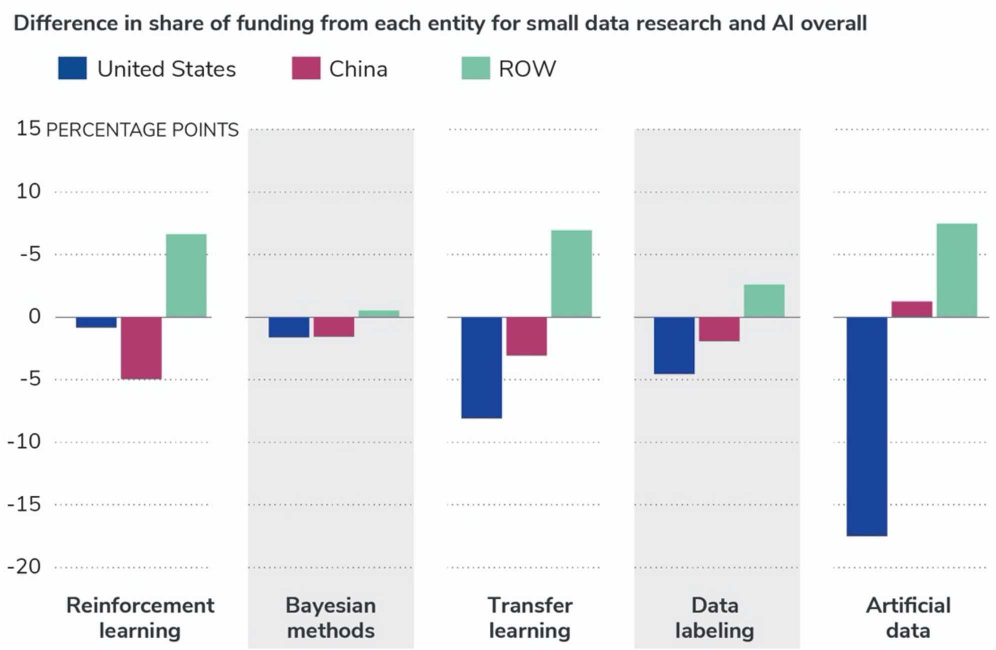
Figure 7. Funding sources for small data approaches relative to AI overall



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Figure 8 further breaks down the funding information associated with government entities by country. Our results indicate that despite the overall trend for government funding to be overrepresented in small data, the U.S. government's share of funding for small data research is lower than its share of AI research in general. On the other hand, private sector companies tend to fund a greater share of small data research in the United States than of AI research overall (see Figure 9 in Appendix B for details).

Figure 8. Government funding for small data approaches relative to AI overall, by China, the United States, and the rest of the world (ROW)



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

This trend is almost opposite when we look at figures for the rest of the world, where government actors fund a much higher share of small data research, especially when compared to the private sector. Interestingly, nonprofit organizations, like research trusts and foundations, in the rest of the world have a lesser tendency to finance small data papers in comparison to their support for all of AI (see Appendix B, Figure 10 for details).

In China, except for artificial data generation, the share of government funding for small data approaches is smaller than AI overall, though the discrepancy is not as large as in the United States.

Key Takeaways

This paper provides an introduction to and overview of a range of "small data" approaches to AI. To conclude, we offer the following key points based on our findings:

Artificial intelligence is not synonymous with big data, and especially not with large, pre-labeled datasets. The role big data has played in the AI boom of the last decade is undeniable, but thinking of large-scale data-gathering and labeling as a prerequisite for AI progress will lead policymakers astray. Alternative approaches are diverse and can be used differently in different settings: if data on the problem at hand is scarce but data on a related problem is abundant, perhaps transfer learning can be useful; if the problem can be dealt by accessing a simulated or real environment where the agent can learn by trial and error rather than pre-collected data, reinforcement learning is likely needed; and so on.

Research into transfer learning is growing especially rapidly—even more quickly than the larger and better-known field of reinforcement learning. By implication, this approach is likely to work better and be more widely used in the future than it is today. Policymakers facing a lack of data for a problem of interest would therefore be well-served by seeking to identify related datasets that could perhaps serve as a starting point for a transfer learning-based approach.

The United States and China are competing closely in small data approaches as the top two countries (by number of research papers) in each of the five categories we considered, according to our research cluster-based methodology. While the United States has a large lead in the two largest approaches (reinforcement learning and Bayesian methods), China holds a small but growing lead in transfer learning, the fastest-growing category.

Tentatively, transfer learning may be a promising target for greater U.S. government funding. Relative to investment patterns across AI as a field, U.S. government funding occupies a smaller share of funding for small data approaches. This could either be because research in these fields are not prioritized by U.S government actors, or because private sector players in the country tend to allocate a proportionately higher share of funds in researching these approaches. Either way, given transfer learning's position as a rapidly emerging field, it may represent a promising opportunity for increased funding from U.S. government sources.

Authors

Husanjot Chahal is a research analyst with CSET, where Helen Toner is the director of strategy, and Ilya Rahkovsky is a data scientist.

Acknowledgements

The authors are grateful to Michael Osborne at University of Oxford and Matt Sheehan at the Paulson Institute for reviewing the report and providing feedback and suggestions. We thank Cody Marie Wild at the Center for Human-Compatible Artificial Intelligence at the University of California, Berkeley, and Andreas Stuhlmüller at Ought for their consultations and thoughtful reviews on the technical write-up of this project. Thanks also to Tim Rudner at CSET, Chris Painter at the Joint Artificial Intelligence Center, and Tien Pham and Raghuveer M. Rao at the Army Research Laboratory for their support during the early phases of this research.

We also thank Igor Mikolic-Torreira, Catherine Aiken, and Autumn Toney at CSET for their continued guidance on the research and writing process. Many thanks to Farhana Hossain for help with data visualization, and Shelton Fitch, Lynne Weil, Matt Mahoney, and Melissa Deng for their editorial support.

Document Identifier: doi: 10.51593/20200075

Appendix A: Methodology

Identifying research clusters

This paper utilizes an existing dataset of research clusters created by CSET’s data science team.[37] Research clusters (RCs) are groups of scientific research articles linked by citations. The articles used to identify clusters are drawn from CSET’s merged corpus of scientific literature, which draws from five datasets that together account for roughly 90 percent of the world's scholarly output: Dimensions, Microsoft Academic Graph, China National Knowledge Infrastructure, arXiv, and Web of Science. The breadth and international coverage of this corpus, containing 109.8 million unique research papers in Chinese, English, French, German, Japanese, Portuguese, Spanish, and other languages, makes it possible to do the types of cross-national comparisons we include in this paper.[38] For this paper, we use the version of CSET’s RC data from February 12, 2021.

For each RC, CSET’s data science team has extracted phrases (a combination of one or more adjectives with a noun) that occur most commonly in paper titles and abstracts. For our analysis in this paper, we made use of these extracted phrases to identify RCs likely to contain research in our five categories of interest (transfer learning, data labeling, artificial data generation, Bayesian methods, and reinforcement learning). After collaborating with technical experts to come up with phrases that would be likely to identify research in these categories, we identified a short list of RCs in each category as follows:

- *Transfer learning:* RCs in which at least one of the following extracted phrases appeared in 10 percent or more of papers: “transfer learning,” “zero-shot learning,” “one-shot learning,” “few-shot learning.”

- *Data labeling:* RCs in which at least one of the following extracted phrases appeared in 10 percent or more of papers: “automatic image annotation,” “semi-supervised learning,” “active learning.”

- *Artificial data generation*: RCs in which at least one of the following extracted phrases appeared in 10 percent or more of papers: "synthetic data," "virtual sample generation."[39]

- *Bayesian methods*: RCs in which 10 percent or more of papers used an extracted phrase that included "Bayesian" (for example, "Bayesian inference," "Bayesian network," etc.) and in which more than 10 percent of papers were AI papers.[40]

- *Reinforcement learning*: RCs in which 10 percent or more of papers included the extracted phrase "reinforcement learning."

The underlying method used here, science mapping using bibliometric-based clustering, is analytically useful because it allows us to discover communities in a network, and by basing our extracted phrase search within citation-linked clusters we sought to use a relatively small set of extracted phrases to discover relevant small data communities. However, interlinkages resulting from a citation-based approach also present the possibility that some resulting papers may not directly address the topic under consideration but are connected to a cluster due to citations, or some relevant papers with unconnected bibliographies are missed. Additionally, clustering solutions that work on many thousands of publications can be challenging to verify given the sheer volume of data involved, and we end up having to choose between accuracy versus inclusiveness. We aimed to balance the two by taking two steps.

Firstly, we manually sense-checked the results of our extracted phrase searches to identify a threshold below which the vast majority of clusters were not relevant. We found that across all RCs, any RC in which our extracted phrase figured in less than 10 percent of the papers was unlikely to be focused on that topic, hence we disregarded those clusters. Secondly, for every RC above the 10 percent threshold, we screened the top 10 core papers, top 10 cited papers, and top 10 phrases within the RC to check if they relate to the topic under consideration. RCs that were clearly unrelated to the topic of interest were dropped, totaling 91 overall.

These dropped clusters numbered three in transfer learning, 17 in data labeling, 10 in artificial data generation, 15 in Bayesian methods, and 46 in reinforcement learning categories. This helped us in eliminating clusters where the phrase of interest was being used in a different sense, e.g. RCs on "active learning" as a pedagogical technique for students, RCs that were primarily psychology clusters under "reinforcement learning," or RCs that primarily focused on mathematics or statistics rather than on machine learning or AI, per se.

For reference, here we present basic information on the RCs included in our analysis:

Table 3. Basic information on RCs in each "small data" category

Category	# of RCs Identified	# of Papers Within RCs Identified (2010–2020)	% AI Papers in RCs Identified (Mean) (Min, Max)[41]
Transfer learning	9	8,102	78% (52%, 89%)
Data labeling	16	11,405	74% (31%, 88%)
Artificial data generation	4	1,433	45% (6%, 85%)
Bayesian methods	78	32,247	41% (10%, 85%)
Reinforcement learning	43	27,137	63% (7%, 94%)
Total	150	80,324	

Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Growth forecasts

Our findings refer to the "future growth index," a metric calculated from CSET's forecasts of research cluster growth. CSET has ranked research clusters according to the probability that they will experience extreme growth—i.e., more than 8 percent growth per year in the number of papers in the cluster—between 2020 and 2023.[42] To present these ranks in a more intuitive form, the future growth index used in this paper takes the inverse of the average forecast rank for clusters in a category, then multiplies it by a scaling factor of 100,000.

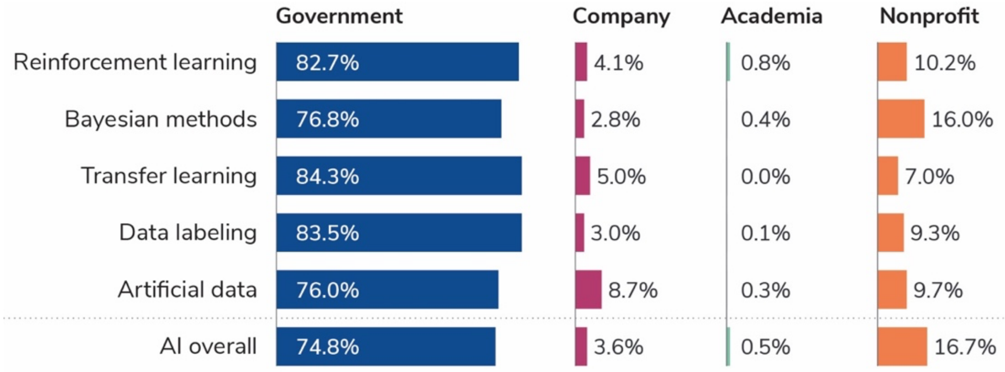
For example, clusters in the category of transfer learning were ranked, on average, 3,761st among the 55,000 ranked clusters in terms of their forecast probability of experiencing extreme growth. Transfer learning's growth index was therefore equal to 1/3761 * 100,000 = 26.6.

For growth forecasts broken down by country, we selected all small data papers published by each country along with research clusters where these papers belong. Then we assigned a growth ranking to each paper based on their research cluster. Finally, we created a weighted average of growth rankings weighted by the number of papers in each cluster for every country.

Observing funding patterns

For all of the small data research captured in CSET's database, we had funding data available for about 30 percent of the papers across all categories. This data provided us information on funding entities and their country of origin. For many of these papers (roughly 20-30 percent of all papers), we had information available on the type of funding entity—government, companies, academia, and nonprofits. We used the available information to look at percentage shares of each sector's funding for these papers. The following table lists the percentage share of papers funded by the five small data categories, and all of AI overall.

Table 4. Percentage share of papers funded by entity type



	Government	Company	Academia	Nonprofit
Reinforcement learning	82.7%	4.1%	0.8%	10.2%
Bayesian methods	76.8%	2.8%	0.4%	16.0%
Transfer learning	84.3%	5.0%	0.0%	7.0%
Data labeling	83.5%	3.0%	0.1%	9.3%
Artificial data	76.0%	8.7%	0.3%	9.7%
AI overall	74.8%	3.6%	0.5%	16.7%

Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

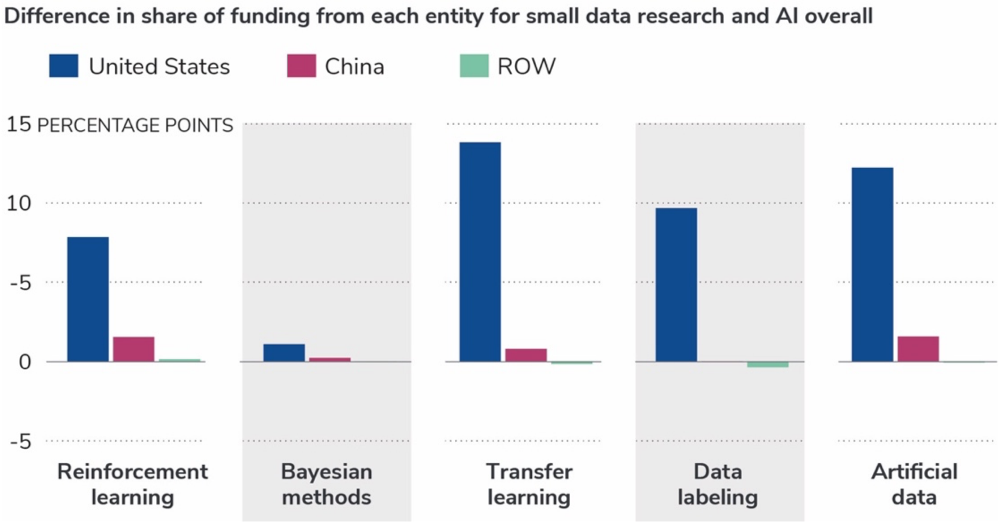
To observe differences in funding patterns between small data approaches and AI overall, we subtracted the percentage share of all AI papers from each category's share to observe the difference in percentage points and presented the results in the findings section.

Appendix B: Additional figures

The three figures in this appendix present data on funding information for small data research categories by country, broken out for companies, nonprofits, and academic organizations. Coupled with the data presented for government entities in Figure 8, the four charts together represent all funding information as available across the four entity types we studied in this paper.

Figure 9 presents data on the share of funding for small data research by companies in the private sector. It appears that, in comparison to government entities, the private sector in the United States tends to fund a greater share of small data research versus all of AI research in general.

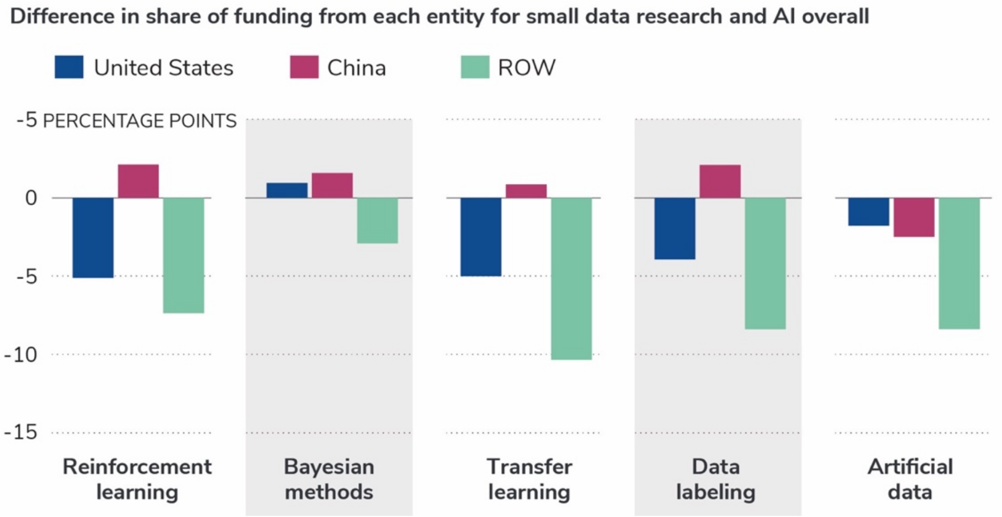
Figure 9. Company funding for small data approaches relative to AI overall, by China, the United States, and the rest of the world (ROW)



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Figure 10 presents trends in small data funding by nonprofit organizations like research trusts and foundations, by country. It appears that across most categories, nonprofits in the United States and the rest of the world tend to underfund small data research versus AI overall. In comparison, nonprofit organizations in China tend to emphasize funding all small data categories by a little, except for artificial data generating approaches, where Chinese nonprofit funding shares tend to dip.

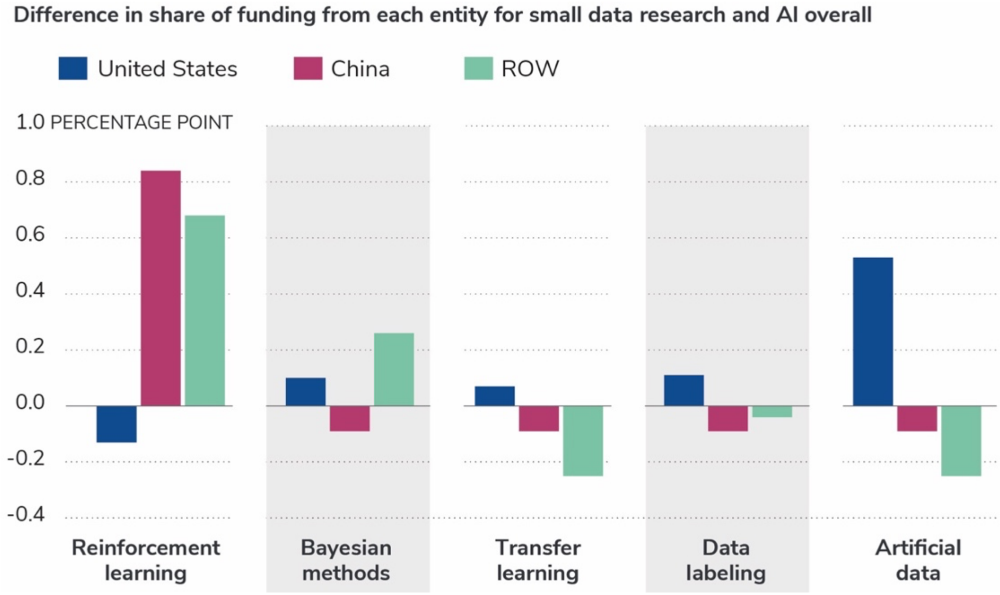
Figure 10: Nonprofit funding for small data approaches relative to all of AI, by China, the United States, and the rest of the world (ROW)



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Finally, Figure 11 presents funding patterns across academic organizations for small data research categories. Please note that in comparison to the two other entity types we discussed above, the percentage point difference between funding for small data research versus all of AI by academic organizations is very little—less than 1 percent overall. It can therefore be concluded that academia's funding patterns for small data mirrors its funding pattern for all of AI research across all approaches for all countries.

Figure 11: Academic funding for small data approaches relative to all of AI, by China, the United States, and the rest of the world (ROW)



Source: CSET merged corpus of scholarly literature, as of February 12, 2021.

Endnotes

[1] Chi Chen and Shyue Ping Ong, “AtomSets – A Hierarchical Transfer Learning Framework for Small and Large Materials Datasets,” arXiv preprint arXiv: 2102.02401 (2021), https://arxiv.org/pdf/2102.02401.pdf; H. James Wilson and Paul R. Daugherty, “Small Data Can Play a Big Role in AI,” *Harvard Business Review*, February 17, 2020, https://hbr.org/2020/02/small-data-can-play-a-big-role-in-ai; Rafael S. Pereira, Alexis Joly, Patrick Valduriez, and Fabio Porto, “Hyperspherical embedding for novel class classification,” arXiv preprint arXiv: 2102.03243 (2021), https://arxiv.org/pdf/2102.03243.pdf.

[2] Ahmed Banafa, “Small Data vs. Big Data: Back to the Basics,” BBVA OpenMind, July 25, 2016, https://www.bbvaopenmind.com/en/technology/digital-world/small-data-vs-big-data-back-to-the-basics/; “What is small data (in just 4 minutes),” *Wonderflow Blog*, April 1, 2019, https://www.wonderflow.ai/blog/what-is-small-data; Ben Clark, “Big Data vs. Small Data – What’s the Difference?,” iDashboards, December 19, 2018, https://www.idashboards.com/blog/2018/12/19/big-data-vs-small-data-whats-the-difference/; Priya Pedamkar, “Small Data vs Big Data,” Educba, accessed August 2021, https://www.educba.com/small-data-vs-big-data/.

[3] Husanjot Chahal, Ryan Fedasiuk, and Carrick Flynn, “Messier than Oil: Assessing Data Advantage in Military AI” (Center for Security and Emerging Technology, July 2020), https://cset.georgetown.edu/publication/messier-than-oil-assessing-data-advantage-in-military-ai/.

[4] Daniel Castro, Michael McLaughlin, and Eline Chivot, “Who Is Winning the AI Race: China, the EU or the United States?” (Center for Data Innovation, August 19, 2019), https://datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/.

[5] Access to large amounts of raw data is not the only factor that matters in measuring data competitiveness given the presence of several alternative AI approaches that do not require such voluminous data. Even if big data is viewed as a key factor underlining China’s advantage in this domain, it makes it even more important for democratic countries to leverage small data techniques to reconfigure the competitive landscape. Tim Hwang, “Shaping the Terrain of AI Competition” (Center for Security and Emerging Technology, June 2020), https://cset.georgetown.edu/publication/shaping-the-terrain-of-ai-competition/.

[6] Here big data is conceptualized as massive datasets collected from real-world interactions that are generally required to be stored, cleaned, transformed, labeled, and optimized before being deployed to train AI algorithms.

[7] Yarin Gal, Riashat Islam, and Zoubin Ghahramani, "Deep Bayesian Active Learning with Image Data," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, 1183-1192.

[8] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. "Learning Deep Object Detectors from 3D Models," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*, 2015, 1278-1286.

[9] Defence Science and Technology Laboratory, "Synthetic Data," UK Government, August 12, 2020, https://www.gov.uk/government/publications/synthetic-data.

[10] Please note that non-Bayesian methods may also incorporate information about the structure of the problem. However, Bayes has another advantage, well-calibrated uncertainty, relevant to small data.

[11] Nimar Arora, Stuart Russell, Paul Kidwell, and Erik Sudderth, "Global Seismic Monitoring: A Bayesian Approach," *Proceedings of the AAAI Conference on Artificial Intelligence* 25, no. 1 (2011).

[12] David Silver et al., "Mastering the game of Go without Human Knowledge," DeepMind, October 19, 2017, https://deepmind.com/research/publications/mastering-game-go-without-human-knowledge.

[13] Proximity of research clusters cannot be interpreted quantitatively here, and is just designed to position clusters with a lot of connections closer together.

[14] Even though small data approaches could enable certain entities to overcome barriers pertaining to data, other factors like dependence on compute and rare expertise may continue to prevent under-resourced smaller actors from reducing capability differentials in comparison to bigger players.

[15] Small data approaches may not reduce the capability differential between big and small entities on all fronts. It may be the case that having a lot of relevant data for specific tasks may likely be useful in a lot of circumstances. However, in some areas, small data approaches may democratize access to modern machine learning for those entities that do not have the capacity to collect and clean big data.

[16] Darrell M. West, "Brookings survey finds worries over AI impact on jobs and personal privacy, concern U.S. will fall behind China," *Brookings Institution*, May 21, 2018, https://www.brookings.edu/blog/techtank/2018/05/21/brookings-survey-finds-worries-over-ai-impact-on-jobs-and-personal-privacy-concern-u-s-will-fall-behind-china/.

[17] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni, "The Synthetic data vault," MIT Data to AI, October 2020, https://dai.lids.mit.edu/wp-

content/uploads/2018/03/SDV.pdf; Jack Goetz and Ambuj Tewari, "Federated Learning via Synthetic Data," arXiv preprint arXiv: 2008.04489 (2020), https://arxiv.org/abs/2008.04489.

[18] "Eliminate Privacy Concerns with Synthetic Data," Gretel, accessed August 2021, https://gretel.ai/synthetics; U.S. Government Accountability Office, "Internet Privacy: Additional Federal Authority Could Enhance Consumer Protection and Provide Flexibility," Government Accountability Office, January 2019, https://www.gao.gov/assets/700/696446.pdf.

[19] Rosa Sobradelo and Joan Martí, "Using Statistics to Quantify and Communicate Uncertainty During Volcanic Crises," in *Observing the Volcano World* (Berlin: Springer, 2018), 571-583, https://link.springer.com/chapter/10.1007/11157_2017_15; Milena Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine* 178, no. 11 (August 2018), https://www.researchgate.net/publication/327138260_Potential_Biases_in_Machine_Learning_Algorithms_Using_Electronic_Health_Record_Data; Sema K. Sgaier, Vincent Huang, and Grace Charles, "The Case for Causal AI," *Stanford Social Innovation Review*, Summer 2020, https://ssir.org/articles/entry/the_case_for_causal_ai.

[20] Sydney J. Freedberg, Jr., "Exclusive: Pentagon's AI Problem is 'Dirty' Data: Lt. Gen. Shanahan," *Breaking* Defense, November 13, 2019, https://breakingdefense.com/2019/11/exclusive-pentagons-ai-problem-is-dirty-data-lt-gen-shanahan/.

[21] Transfer learning can potentially circumvent "dirty data" problems only if the related dataset which we are transferring knowledge from is clean and organized. If not, we will not be able to circumvent the issue as a whole, however, the bigger advantage can still be amortizing the time and cost of data cleaning across problems. For example, the same dataset of bird species in general can be used for two (or more) purposes—for training algorithms to identify birds in general, and for transferring that knowledge to an algorithm to identify rare bird species, and more.

[22] Ilya Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth," *Fronters in Research Metrics and Analytics*, April 6, 2021, https://www.frontiersin.org/articles/10.3389/frma.2021.630124/full.

[23] AI clusters are defined as clusters where more than 50 percent of papers are AI papers. For methodological details and more information on how an AI-relevant paper is defined, please see James Dunham, Jennifer Melot, and Dewey Murdick, "Identifying the Development and Application of Artificial Intelligence in Scientific Text," arXiv preprint arXiv:2002.07143 (2020), https://arxiv.org/pdf/2002.07143.pdf. Note that not all of the clusters we

identified as belonging to “small data” approaches meet this definition of “AI clusters”; see Appendix A for more details.

[24] For instance, for a research cluster we identified using extracted phrases related to transfer learning, our analysis treats all papers in that cluster as transfer learning papers.

[25] Matthew Botvinick et al., “Reinforcement Learning, Fast and Slow,” *Trends in Cognitive Sciences* 23, no. 5 (May 2019): 408-422, https://www.sciencedirect.com/science/article/pii/S1364661319300610#bib0065; Volodymyr Mnih et al., “Human-level control through deep reinforcement learning,” *Nature* 518 (2015): 529-533, https://www.nature.com/articles/nature14236.

[26] Wolfgang Glänzel, “On the Opportunities and Limitations of the H-index,” *Science* Focus, January 2006, https://www.researchgate.net/publication/28806524_On_the_Opportunities_and_Limitations_of_the_H-index.

[27] The indicator is based more on long-term observations, so it does not show a decay in productivity as the number of citations continue to increase even if no new papers are published. Glänzel, “On the Opportunities and Limitations of the H-index.”

[28] The methodology used to forecast growth is described in Richard Klavans, Kevin W. Boyack, and Dewey A. Murdick, “A novel approach to predicting exceptional growth in research,” *PLoS One* 15, no. 9 (2020), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0239177. See Appendix A for more details on methodology.

[29] The growth forecast for “AI overall” refers to the average growth ranking for research clusters where at least 50 percent of papers are predicted to be “AI papers,” according to the methodology described in Dunham, Melot, and Murdick, “Identifying the Development and Application of Artificial Intelligence in Scientific Text.”

[30] Research papers are attributed to a country based on the authors’ organizational (universities, companies, government, etc.) affiliation as listed on the paper published. This means that a paper with multiple authors based in different countries will figure in the counts of multiple countries. Therefore, all of the research papers linked with one country should not be seen as exclusively belonging to that country.

[31] This finding aligns with global trends for all of AI. In contrast, when looking at top 10 countries by scholarly publications in general (i.e., not limited to AI), Russia figures in the list across most databases.

[32] For all small data papers, we estimated the intensity of international collaboration by looking at the average number of countries coauthoring a publication. If a set of papers does not have any international collaboration, then the average number of countries authoring those papers is 1. If all papers involve the collaboration of the home country and one more country, then the average number of countries authoring those papers is 2, etc. Small data papers published by U.S. authors have coauthors from an average of 1.5 other countries, while small data papers published by Chinese authors have coauthors from an average of 1.4 other countries—only a small difference. On the other hand, some U.S. allies have more intense international collaboration. For example, the United Kingdom's papers have coauthors from an average of 1.7 other countries. Australian and Canadian papers have coauthors from an average of 1.8 other countries, indicating that a higher share of international coauthored papers may have boosted their paper counts.

[33] K.A. Khor and L.-G. Yu, "Influence of international co-authorship on the research citation impact of young universities," *Scientometrics* 107 (March 2016): 1095-1110, https://link.springer.com/article/10.1007/s11192-016-1905-6.

[34] When we look at the same results for h-index, China's ranking for transfer learning slips to second, but it continues to lead in data labeling approaches. Its h-index ranking for Bayesian methods falls further down to the fifth position after France and Germany.

[35] Apart from the 20-30 percent papers with funding information, the rest of the papers were either not funded or we did not observe funding activity. It is not clear how much funding information we miss, but based on the analysis of high quality academic datasets, such as Web of Science and Scopus, we expect to have observed the vast majority of acknowledged funding events.

[36] AI research overall had 30 percent of funding information available.

[37] Rahkovsky et al., "AI Research Funding Portfolios and Extreme Growth."

[38] Because we identify small data clusters using extracted phrases in English, it is worth noting that our methodology depends on the assumption that research published in other languages either has an English-language abstract available, or cites sufficient English-language research to be included in a cluster with research that our methodology could identify.

[39] Please note that we also initially incorporated the keyword "simulation" (and variants), "self play", "sim2real," for this category, but we did not obtain any relevant results so we dropped the keyword.

[40] Papers were classified as “AI papers” using the methods described in Dunham, Melot, and Murdick, “Identifying the Development and Application of Artificial Intelligence in Scientific Text.”

[41] Note that the mean (or average) percentage of AI papers in the identified RCs is calculated by averaging the proportion of AI in each cluster belonging to that category. Because clusters vary widely in how many papers they contain, the average percentage of AI papers across clusters cannot be used to calculate the overall percentage of AI papers in the category. We use the per-cluster mean, min, and max here to give a sense of the range of values across clusters. Papers were classified as “AI papers” using the methods described in Dunham, Melot, and Murdick, “Identifying the Development and Application of Artificial Intelligence in Scientific Text.”

[42] Research cluster growth was forecasted using the methodology described in Klavans, Boyack, and Murdick, “A novel approach to predicting exceptional growth in research.”