

# Global AI Talent Report 2020



jfgagne

The report is best read [online](#) to navigate the interactive data visualizations. If you are on mobile, check out the [top highlights here](#), and an explanation of the [critical technical AI roles here](#).

## Organization of this report

As these are new metrics and a much wider scope from previous years, we have focused on presenting the data with highlights of trends and outliers to help you get started in exploring the information. Analysis requires a lot of ground-level understanding that we will look to add in subsequent articles and collaborations, along with further points of quantitative analysis that we outline in the conclusion for us and others to take up.

This report is divided into four primary sections. The first section is an introduction and executive summary with the rankings of the top countries by talent pool totals. The second explains the different specialized technical roles along the value chain of productizing AI, from research to deployment. The third section explores arXiv data, a pre-publishing platform that is the closest thing to a census of AI research, and looks at location, movement, and gender. Finally, the fourth section is where we look at our early estimates of supply and demand in industry for all of the different roles covered in the first section. We touch on methodology throughout and have included an appendix that summarizes methodologies, as well as another appendix summarizing areas for further exploration to add more granularity to these new metrics.

## INTRO

A recent [Technology Quarterly](#) by The Economist was subtitled: “After years of hype, many people feel AI has failed to deliver....” It considered two core problems: access to the required data, and that algorithms weren’t all that smart, yet. This is unsurprising since so much of the communication and media about artificial intelligence (AI) and machine learning (ML) promised magic algorithms if given the right data.

However, this perspective is only partly true, as there is much unrealized potential of AI that can be achieved with currently available data and algorithms—given enough of another scarce resource to apply it: talent. Our previous estimates of the talent pool have focused on AI researchers—those who are making the algorithms smarter and working on ever larger data sets—and this focus perhaps added to the illusion that that was all that was needed.

Taking AI from research to real-world impact is a long value chain that depends on a range of skill sets and experience. It is common to see people who can, and do, fill multiple roles as some of the rarer skills are in need across the value chain. Even so, we think it is useful to categorize and explore these roles separately to better understand what it takes to build and run AI solutions<sup>1</sup>, and what accessibility to that talent looks like.

## EXECUTIVE SUMMARY

### New metrics, and a broader view, by which to measure talent

This year, we've added estimates of several other critical specialized technical roles in the value chain of developing an AI product: ML engineering, technical implementation, and data architecture. We measured the size of the available talent pool for industry through self-reported data on social media, and demand via the monthly total job postings for the same roles.

We also expanded our view of the AI talent pool to look at not just fundamental researchers, but those doing the applied work. To do so, we have shifted from using conferences as a proxy, whose limited seats don't capture the full growth of the ecosystem, to the AI/ML (cs.AI, cs.LG and stat.ML) repositories on arXiv. ArXiv is where researchers pre-publish their papers (i.e. prior to peer review for acceptance to a publication or conference<sup>2</sup>) and is perhaps the closest thing there is to a census of AI research. It also gives a much broader view of AI growth by including papers on applied methods.

### Here are the brief highlights:

The total number of authors publishing each year on arXiv has grown an average of **52.69% per year since 2007**. Last year saw **nearly 58,000 authors** and we estimate it to be **86,000 by year-end**. Among the four industry roles we looked at for the talent pool we estimate there to be about **478,000 people** (see section "The Specialized Technical Roles of the AI Value Chain").

The national rankings are in the two graphs below, and should be viewed with an important note on methodology: we have assigned location based on the location of the headquarters of the organization the author is affiliated with, in order to emphasize where the intellectual property (IP) is owned. This choice gives significant weight to countries like the United States whose Big Tech companies have AI research labs all over the world.

As we have in the past, we continue with several quantitative analyses of the pool of research talent on arXiv (for full detail, see section "The People Making AI Smarter").

### **Talent remains global and mobile (at least pre-pandemic)**

Our analysis shows that at least for the talent contributing to research, the pool truly is global. Collaborations continue across the world, except for the Global South, whose ecosystems are much

---

<sup>1</sup> Our categories emphasize the aspects of roles in *building* an AI solution, as opposed to *running* it, though the needed skill sets could cover both.

<sup>2</sup> There is a strong acceptance for pre-publishing in AI as it is easy to test out methods and see if they are replicable and useful. Granted, we did not sort for popularity as a signifier of impact but only to see where the volume of research is.

earlier in their development. Ireland outranks everyone with over 15 average collaborations per author, while most countries sit around 4-6.

Migration may be a dated metric with the pandemic and remote work, but it does help to show which countries (or companies) have more pull versus those who seem to supply others, as well as those who seem to be more insular with relatively few people coming and going. Of course, the US has by far the biggest pull of anyone. This may be partly undone by new visa policies, creating a big opportunity for their neighbors, but the many international labs of its Big Tech companies may mitigate some of this.

### **The gender balance has made little gains**

Gender in arXiv is even less balanced than we had seen when looking at conferences, and has barely moved since 2007, only going from 12% to 15% today.<sup>3</sup> The ratios do vary significantly country to country, but the pandemic has [more greatly affected the output of female researchers](#) and so this year could be a step backward overall. Another factor is in who contributes to publishing, which our self-representation data sets help shed light on.

### **Few of those in industry appear to do full-time fundamental research**

On job-related social media we see that there are only about 4,100 people presenting themselves as professional researchers for industry, perhaps indicating a number of publishers working primarily as engineers or another role and doing research part-time. We have also found this to be anecdotally true among our peers. While this may be out of a preference for applied work, we have found that the proportion of demand for pure researchers in industry is equally small (about 1.7% of job postings) and the pure research work in industry isn't widely available.<sup>4</sup> Adding data on graduation rates, employment records in academia and private research labs, and understanding the average publishing lifespan and output of a researcher will help with answering more clearly how industry impacts the output of research.

This phenomenon of people splitting time between applied work and research could further affect the gender balance in research contributions. [Some reporting has shown](#) that women are less likely to get jobs that allow them to continue contributing to research. With the challenge of getting jobs doing pure research after graduation, this puts even more pressure on incoming cohorts to include high numbers of women to affect the overall ratio of women's contributions to research.

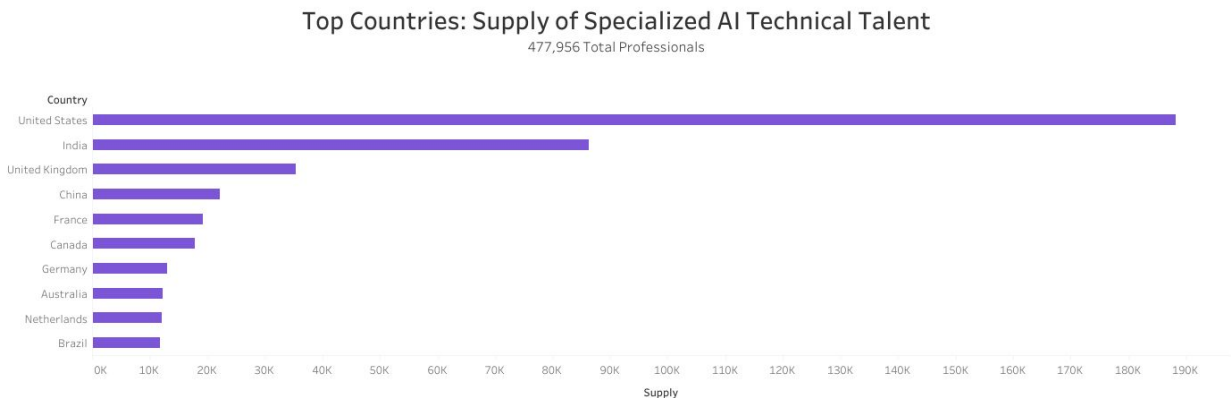
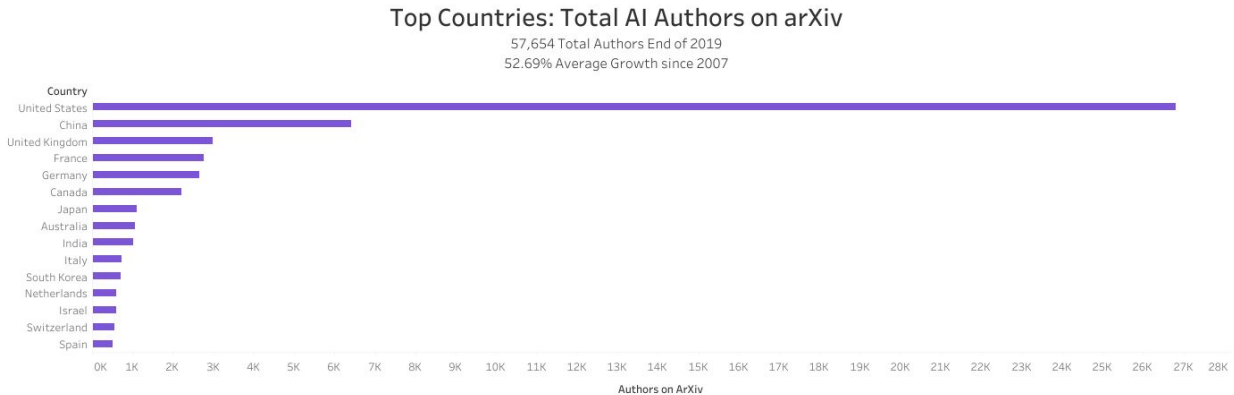
### **Demand for new roles has been steady, but saw a big dip in 2020**

Unfortunately, data on demand is limited at this stage, but we can get some insight from tracking monthly job postings. What we have found is that the proportionate mix of specialized engineering, technical implementation, and research roles in demand is closely matched with the mix of supply: about 61% for implementation roles building the software around AI capabilities, 38% for AI engineering roles building the core AI capabilities, 1% for researchers. We do not know the aggregate amounts, but the monthly flows were growing steadily in 2019 around 2-6% for different job titles. Unsurprisingly, we also saw 20-30% drops in demand for the relevant job titles during 2020, but both 2019 and 2020 show significant outliers for those entering the scene and persisting through the pandemic.

---

<sup>3</sup> At conferences it was only slightly better, last year's report showed the proportion of women published at the top conferences was 18%.

<sup>4</sup> There is also a possibility that most people stay in research in academia and simply do not create job profiles for themselves, and that the professional degrees and online training in AI are making up a greater proportion of the supply thanks in part to being significantly shorter than research oriented PhD tracks.



## The Specialized Technical Roles of the AI Value Chain

The methods for building AI products have evolved, and the challenges have grown as we've taken aim at new problems, but we are clear on the standard types of expertise needed to bring AI from theoretical ideas to concrete products.

At Element AI, we have the benefit of the founders' 20 years of experience of developing AI products, and being connected with the broader ecosystem to check and validate these categories. While there are differences and nuance in the details of roles across companies (and geographies), the following series of job titles and job descriptions cover the specialized technical roles that make up the value chain of developing an AI product:

- Research
- AI/ML Engineering
- Data Engineering/Architecture
- AI/ML Productization

Below we expand on their respective roles and qualifications.

### Research

Researchers continue to rapidly expand AI capabilities. The race to tipping points of efficiency or effectiveness in the underlying technology is so important that many professional researchers are either full-time in industry, or are at least affiliated—even though the research is open as a common standard. This openness is important as it allows for collaboration between organizations (academic or industrial) to better define research agendas and develop novel methods.

Research advances quickly in part because of pre-publishing on arXiv, which allows for new methods to be quickly assessed and built upon. However, the best papers are accepted by publications and conferences that offer their discoveries greater reach and scrutiny for improvement, making both pre-publishing and official publishing important parts of the researcher role.

Even as this report and the market add focus to the other roles below, it is important that the role of researcher not lose emphasis. Organizations that benefit from cutting edge techniques cannot expect to maintain their lead for long in this rapidly evolving field without reinvesting in good research. Besides, researchers often have excellent skill sets that can be transitioned to more applied roles, and so the role remains a great starting point for top talent.<sup>5</sup>

**Qualifications:** The qualifications for this role include a PhD in machine learning, computer science, artificial intelligence or a related quantitative field; experience and mastery of scientific programming and relevant libraries; experience contributing to research communities important for pushing the quality of research.

## AI/ML Engineering

These engineers are the key bridge between the fundamental research in AI and its application in the real world. The significant difference with other software engineers is that they are “coding with data”. Much of their work is to take the latest techniques for types of learning and work with data sets to get the techniques to learn something specific and applicable to a problem.

Their titles most commonly are “data scientist”, an example of Tesler’s theorem of how AI techniques stop being called AI as they become standard. They may also be called AI/ML Engineer or Applied Research Scientist. The top engineers apply their deep technical expertise, scientific rigor and creativity, developed through years of experience, to solve practical problems. This requires them to understand user needs while being able to conduct the needed research to get the relevant available techniques workable at scale.

They, too, often pre-publish their findings on applied methods on arXiv, contributing to a very collaborative and science-driven field. Due to this, we’ve captured both researchers and applied researchers in our assessment of arXiv (see following section). The applied work is growing in importance because of the challenge of getting AI to work at scale in the real world, though it is not clear what the precise split is.<sup>6</sup> There are also far more people reporting to be working in the Engineering role (150,500) than are publishing on arXiv (projected to be 86,000 by end of 2020).

The self-reported data and demand data from social media is a closer reflection of the job market. It shows how the demand is there for data scientists and AI developers and not researchers, which helps explain why one can find researcher-level expertise across the value chain because of the need to fill in (see section “the talent to build it out”).

---

<sup>5</sup> There is an open question of the average publishing lifespan of researchers, and whether they publish less when joining industry as they split more of their time with applied work. It also appears in the data below that very few authors are full-time professional researchers (4.6%), and we see many people in the engineer role contributing to research on a part-time basis. Better understanding these numbers precisely should help determine a balanced investment in research output.

<sup>6</sup> This makes arXiv publishing categorized by country useful for people looking to see if there is enough talent to open a lab, as well as to see who is closest to the latest advancements to stay ahead of the curve. Future work should differentiate between applied and fundamental work on arXiv to determine the true split.

**Qualifications:** A PhD or Masters in a quantitative discipline (e.g., Computer Science, Mathematics, Operations Research, Physics, Electrical Engineering)<sup>7</sup>; significant understanding of the underlying theory of deep learning and related AI fields; experience with statistical software (e.g., R, Python, MATLAB, pandas) and with scientific programming; ability to run models on often noisy data that exist in a commercial context; experience in modifying models/techniques to adapt to data limitations; skills in selecting the right statistical tools given a data analysis problem.

## Data Engineering/Architecture

Given the nature of the data scientist's work to code with data, making a flow of data available to train an AI model and then to work at scale is a huge part of the job. Data Engineers or Architects build modern database structures that store terabytes of industrial data used to develop AI models. They construct, test and maintain optimal data pipeline architectures, and ensure that the architecture will support the requirements of the business. Further, they are responsible for best practices in terms of data organization, standards and versioning, and ensuring compliance with internal and external regulations.

What is unfortunately hidden, but is a majority of people's hours, is the data wrangling that happens to get the data itself in shape. Data cleansing, labeling and augmentation can take up to [65% of the hours](#) of a machine learning project [according to Cognilytica](#). This data preparation work varies widely as data engineers can also develop automated methods for these tasks, but many data wrangling tasks require a degree of human expertise especially when working on a novel field of application.

Our report will only look at the skill set of a Data Engineer/Architect as a specialized technical role needed in the value chain, but we strongly recommend this area to not be discounted and to be valued labor because AI operates on a "garbage in—garbage out" principle.<sup>8</sup>

**Qualifications:** A university degree in computer science or other IT field is often a necessary base along with a technology skill set of Spark, Kafka, Pulsar, Cassandra, Hadoop, as well as NoSQL and Relational databases. Most important is experience with different components of a data engineer's role: implementing data transformation methods and supporting robust pipelines in production; architecting data and optimizing it for various software design patterns; and expertise with methods that ensure large-scale quality on top of major cloud platforms.

## AI/ML Productization

Building a product around an AI capability requires significant traditional software expertise, familiarity with AI methods, and a deep understanding of the end user's context. We've subdivided this category into two distinct roles, one that is more focused on assessing what is a feasible solution to build out and another more technical role that does the building. As AI tooling becomes more democratized, the qualifications to be an AI dev will converge towards that of a data analyst and more easily supplemented by widely available online training courses.

**AI developers** are the key people for building out the software environment around an AI capability that makes it into a fully-functioning product. They participate in the elaboration, architecture, design, development, testing, deployment, operation, maintenance, and enhancement of AI models. In productizing AI models, they also need to be able to help in the evaluation and selection of the appropriate technology platform, frameworks and deployment architecture for each given problem to solve, as well as help maintain AI models deployed in production. This role extends to designing, implementing and operating friendly and scalable

---

<sup>7</sup> It has been impressive to see how those in tangential fields have been able to adapt themselves with available online resources.

<sup>8</sup> Check out the company [Samasource](#) for an approach for profitably valuing data labeling work.

APIs, and even working on the UX and interface if the team is small.

**Qualifications:** Bachelor's degree in computer science or equivalent work experience that will have given them knowledge of Web GUI frameworks and mastery of a few programming languages. As these are novel development projects, it will be good to have at least 5 years of experience on large scale projects (preferably microservice-based solutions), and relevant experience with container-based deployment and automation tools.<sup>9</sup>

**Data analysts** are technically adept solutions developers. Their role is to analyze and understand a problem and relevant datasets to assess the best approach for designing a solution. While they are focused on the end user, they need the technical background to be able to effectively assess the feasibility of a solution (verify data distributions, identify data gaps, evaluate accuracy of labels, reproducibility and traceability around collection, transformation, and associated analytical tasks) and communicate the full context to a technical team to build it out or have the business address the gaps.

**Qualifications:** While a specific degree is not required, they will need experience working with programming, data visualization, statistical and data wrangling tools relevant to data analysis and machine learning (e.g., scikit-learn, pandas DataFrame, NumPy, R, awk, RapidMiner, Tableau, D3.js), experience in wrangling unstructured data (e.g., images, pdf) and structured data (e.g., csv), and a broad knowledge of statistical and machine learning techniques with the ability to run and benchmark existing models.

### Supporting Staff

Consultants, solution designers, and AI-aware industry operations (e.g. accounting, legal, etc.) are all important roles in AI adoption, and require an understanding of the dynamics of this new technology. We did not include a deep look at this group as they do not need to have technical AI backgrounds to fill their role, and can be brought up to speed quickly. We did think it important enough to mention as a way of pointing to the value of AI awareness in order to better understand and adapt to the new dynamics of AI software.

While this group is relatively small and focused on getting organizational leadership to see the value of AI and commit to adoption, it will eventually mean major portions of the workforce who are upskilling to successfully adopt and collaborate with this new technology.

## Limitation of Typologies

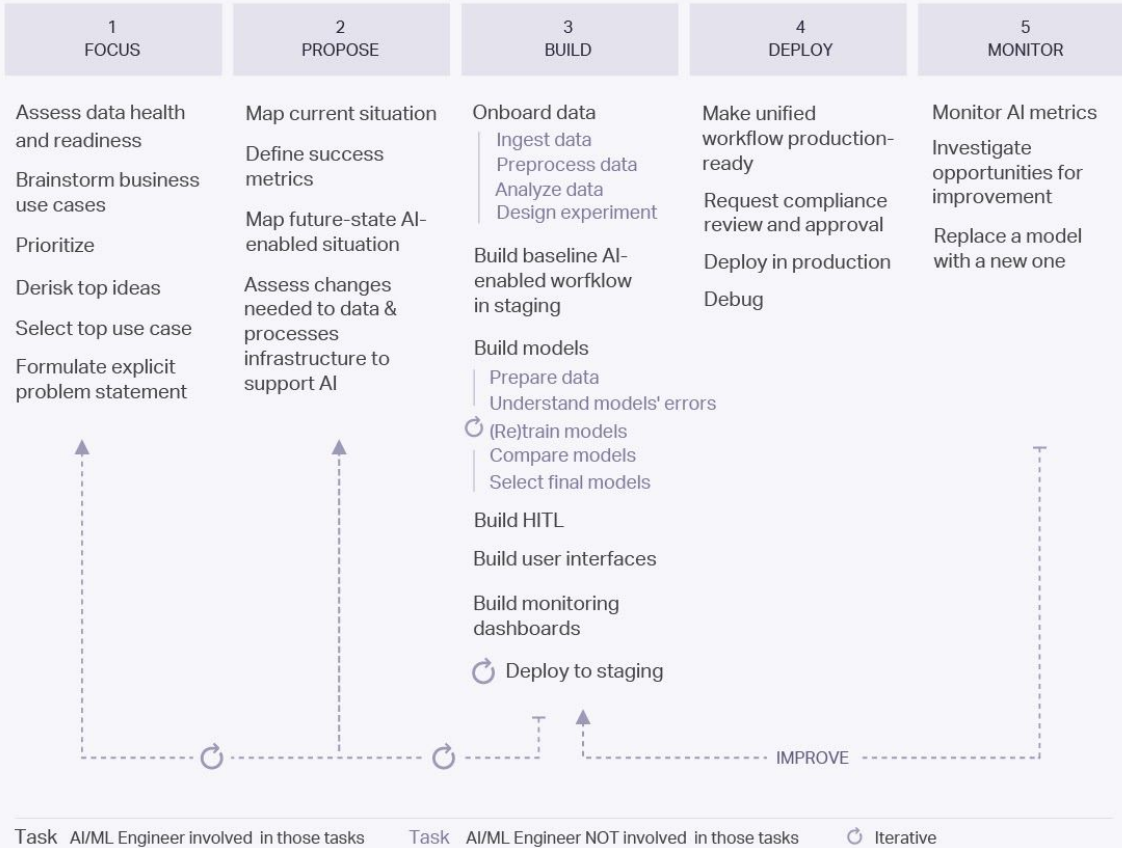
The roles are imperfect categories, and the associated job titles can often cover jobs with little-to-no AI work.<sup>10</sup> However, as AI becomes the standard approach to software, all of these titles will benefit from AI skills—and the market has lots of appetite to reward it. Further, this new standard will mean even more emphasis on data wrangling, where data work will become the primary input for building software.

---

<sup>9</sup> This has generally meant that the AI dev roles are filled with those with data scientist qualifications, or quickly build them in the role. See discussion in “Supply and Demand Along the Talent Value Chain” below for how this has affected talent trends.

<sup>10</sup> Even if it was a perfect typology, there would be an issue of international comparability between other existing typologies in use.

## GETTING AN AI MODEL INTO PRODUCTION



## The People Making AI Smarter

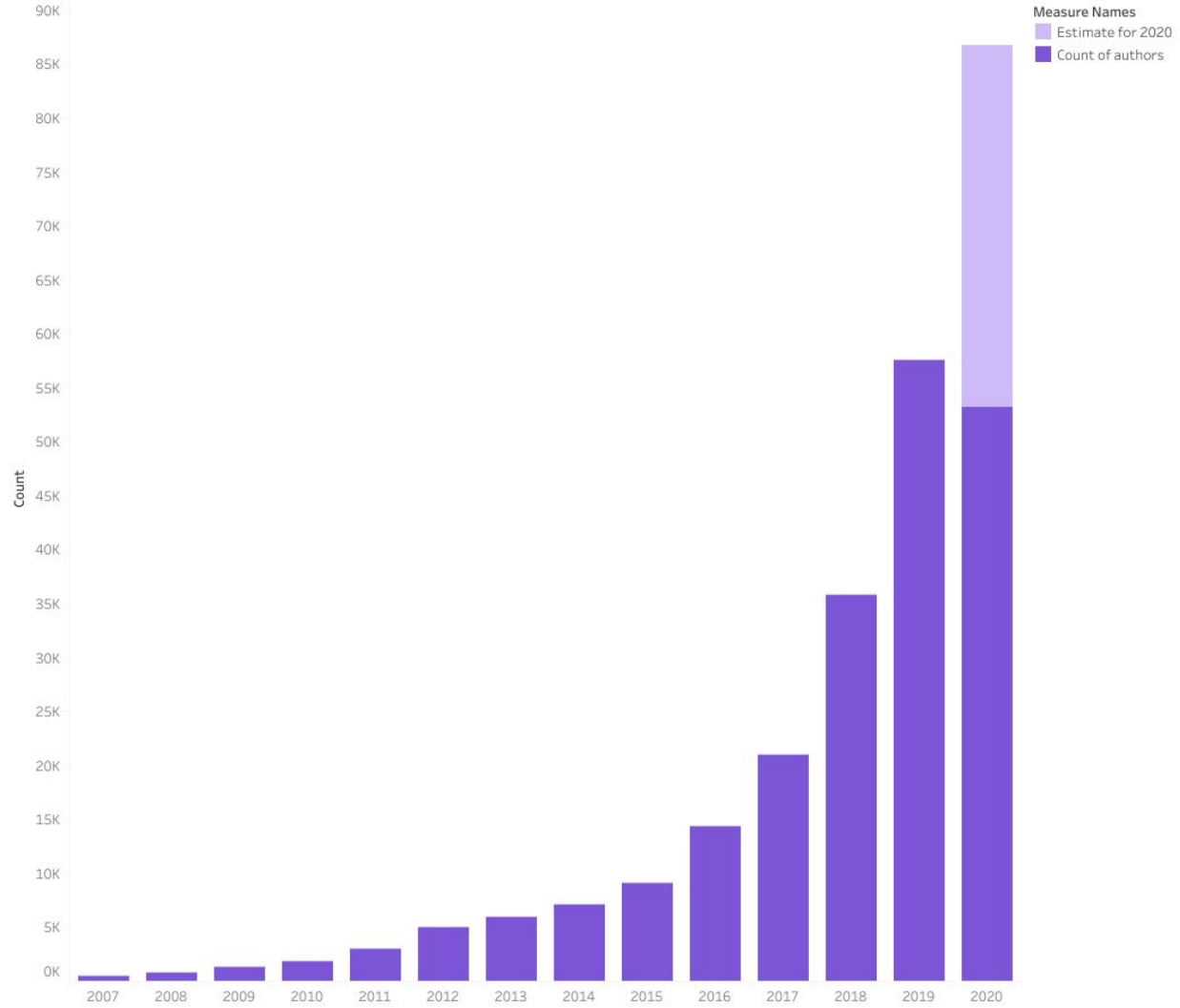
### ArXiv data covering applied and fundamental researchers

Last year we concentrated on the most prestigious scientific conferences around AI, with the goal of understanding the most impactful research. That yielded a count of 22,400 authors going into 2019. This year we decided to broaden our look into research by trying to have the closest thing possible to a "census of research", the pre-publishing repository arXiv. This approach has given us a better understanding of how AI research is unfolding into more than what we had seen in the past: from a small circle of researchers working on making AI real, into a constellation of experts finding ways in which AI can be implemented in different fields.

In other words, this change should enable us to be less focused on "fundamental AI research" and add more "applied AI research" into our scope. This had a number of impacts that we will point to throughout the report, as well as in the detailed section on methodology.

In the below graph, we have applied the methodology to previous years up to 2007 to show the global growth curve from this new perspective.

## Annual Count of AI Authors on ArXiv



Total number of authors publishing on arXiv each year. Author count based on arXiv metadata at time of writing. Estimate based on average growth for past 5 years.

A significant change from the conference methodology is in the growth trends. While growth was up 36% from 2015 and 19% in 2016 at the conferences, we're seeing 47.69% and 53.88% respectively in author growth on arXiv. This shows how much of the activity is outside the traditional, peer-reviewed channels, as well as the limit of conference seats to show growth. A key reason behind the acceptability of eschewing the peer-reviewed world is that it is quick and easy to determine whether something works or not. Conferences do still provide an important filter for the "best ideas" and influence the application of ethical standards (see [ICLR](#) and [NeurIPS](#) conferences that now require ethical considerations in submitted papers), but the volume on arXiv is indicative of the number of people getting engaged with AI tools and can thus feed the talent pool.

These papers (about 78,000 in total, or [a little under 3% of all science publications](#) according to Stanford HAI's 2019 AI Index) are also likely to cover applied approaches rather than fundamentally new discoveries, again making the field all the more accessible for those with less training.

Though, this should not be a reason to let up in the training. This growth has been due to the sustained investment in higher-education, building momentum as these multi-year PhDs take time to manifest.

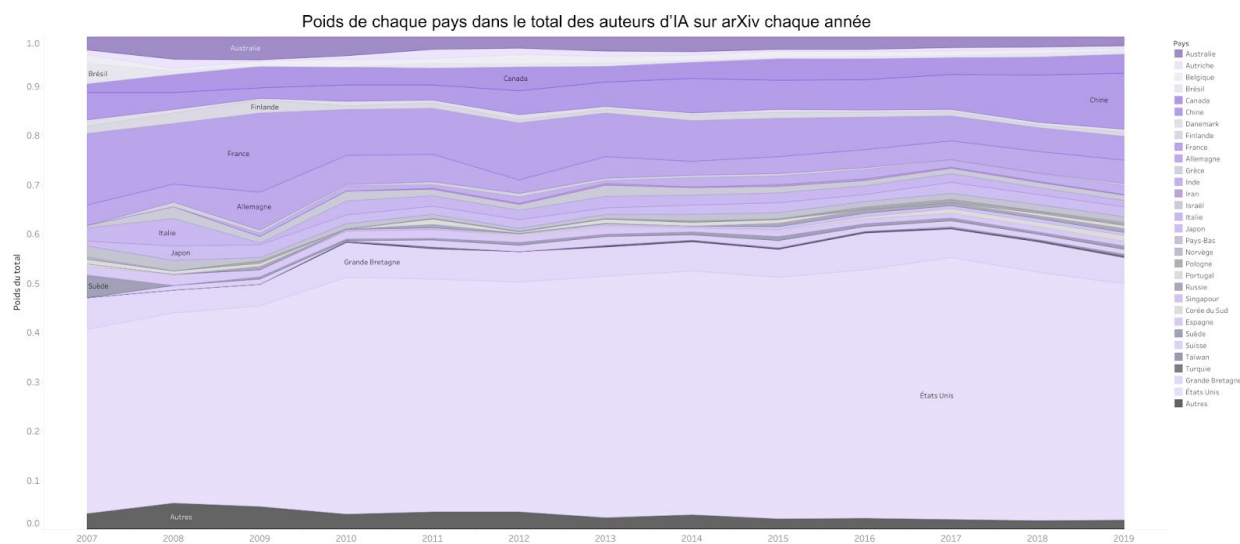
Keeping that momentum will be important as some of the growth in authorship has actually been from experts from other fields who use ML methods choosing to simply add ML as a category of publishing in order to indicate their use of cutting-edge techniques.<sup>11</sup> Further, AI PhDs are not likely to all stay in academia, and will provide high-calibre talent in industry along the entire value chain (or perhaps take their skills to another field of research and not consider it significant to include ML as a publishing category).

All in all, the growth is remarkable since the 478 authors in 2007, virtually no one at all, to just under **58,000** at the beginning of this year and a projected **86,000 by year-end 2020**.

## National Rankings

The US remains dominant, and China is in its own field

The national rankings hold, with the US still on a level of its own as the biggest player with 47.89% of the total number of authors. This must be heavily caveated by our methodology's focus on a western-centric arXiv and only taking on latin-alphabet papers. There is also additional weight given to the US due to the dominance of its Big Tech players, as we've attributed the location of industry authors to the headquarters of their affiliation regardless of which local office they work in. While this discounts the value of geographic proximity stimulating local collaborations and training, it prioritizes where the intellectual property (IP) that is generated is in fact owned.



As we only have a breakdown by country for this year (these were derived from a subset of the population data), we cannot say directly the changes in rankings from previous years. However, compared to the previous years of conference publishing, the top players are the usual suspects: China is in its own ranking at around 11.4%; UK (5.3%), France (4.9%), Germany (4.7%), and Canada (3.9%) in their own grouping; then dropping off to Japan (2%), Australia (1.9%), India (1.8%), Italy (1.3%) and South Korea (1.3%). These “next 10” make up 38.6% of the total number of authors, and with the US, these top 11 make up 86.5%; not very surprising profiles to see as all of these are “Global North” countries with significant tech sectors. Others on this list could catch up fairly quickly with some targeted investment, thanks to the help of relatively small numbers as a whole.

<sup>11</sup> Publishers on arXiv can select multiple categories to publish on.

## Average Number of International Collaborations Per Author

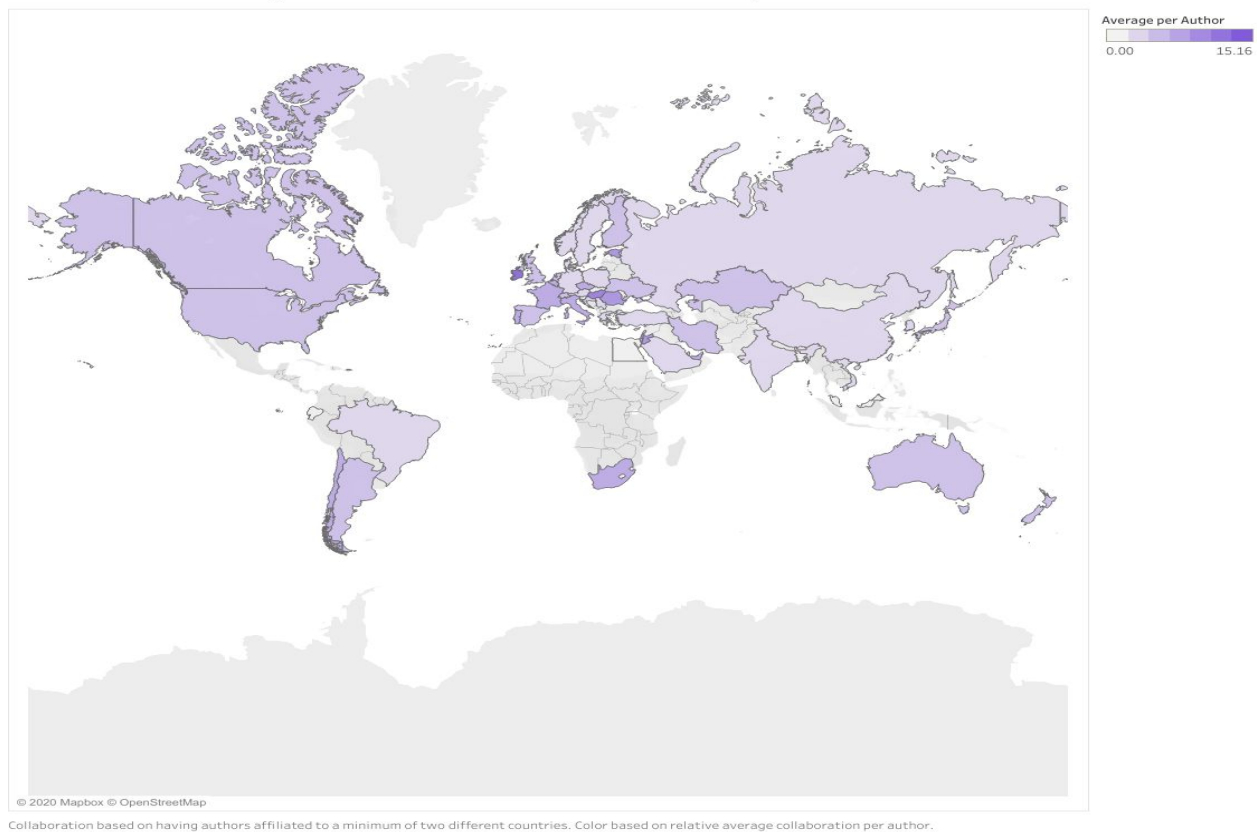
Collaborations continue globally, with smaller countries standing out

This isn't to say that the work is not shared globally. By identifying the country of each author on a paper, we found that collaborations continue everywhere with everyone, with an unfortunate exception of the Global South who are largely absent from the map of collaborations. This year, we've switched focus from which countries are collaborating with each other, to the average number of collaborations per author in a given country.

The big winners are mostly small countries, and an intuitive explanation is that it is out of necessity due to fewer options locally. While the US is not high on the list of average collaborations per author with a middling (4-6) average per author similar to other AI-mature countries, it still has by far the most number of total international collaborations. Notably China and South Korea are on the low end with an average of under 3 per author.

On the other end is an extreme outlier, Ireland, with 15 per author. Runners-up are Singapore and France with around 8 and 7 per author respectively, and both are AI-mature in their own right. There are other countries that scored between Singapore and Ireland (e.g. Hungary (11.5) or Estonia (9.5) or Belgium (8.5)), but these countries should not be considered as much because of their significantly small number of researchers.

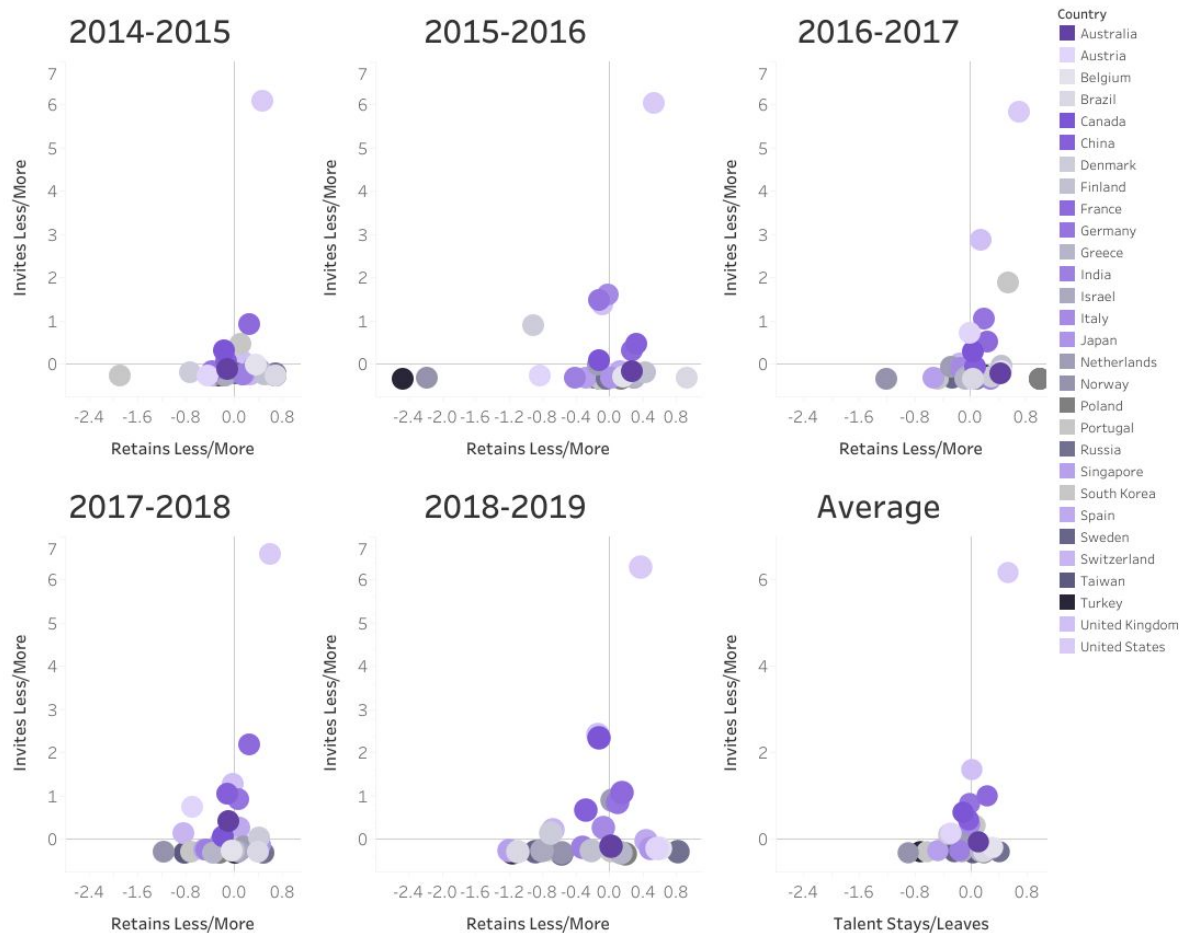
Average International Collaborations per Author



## Talent Attraction and Retention

More countries are able to resist the pull of the US than they were 5 years ago, and 2020 may accelerate that

Plotting countries based on their inflow and outflow of AI talent, we see that countries fall into one of four types: Inviting Countries, Producer Countries, Anchored Countries and Platform Countries.



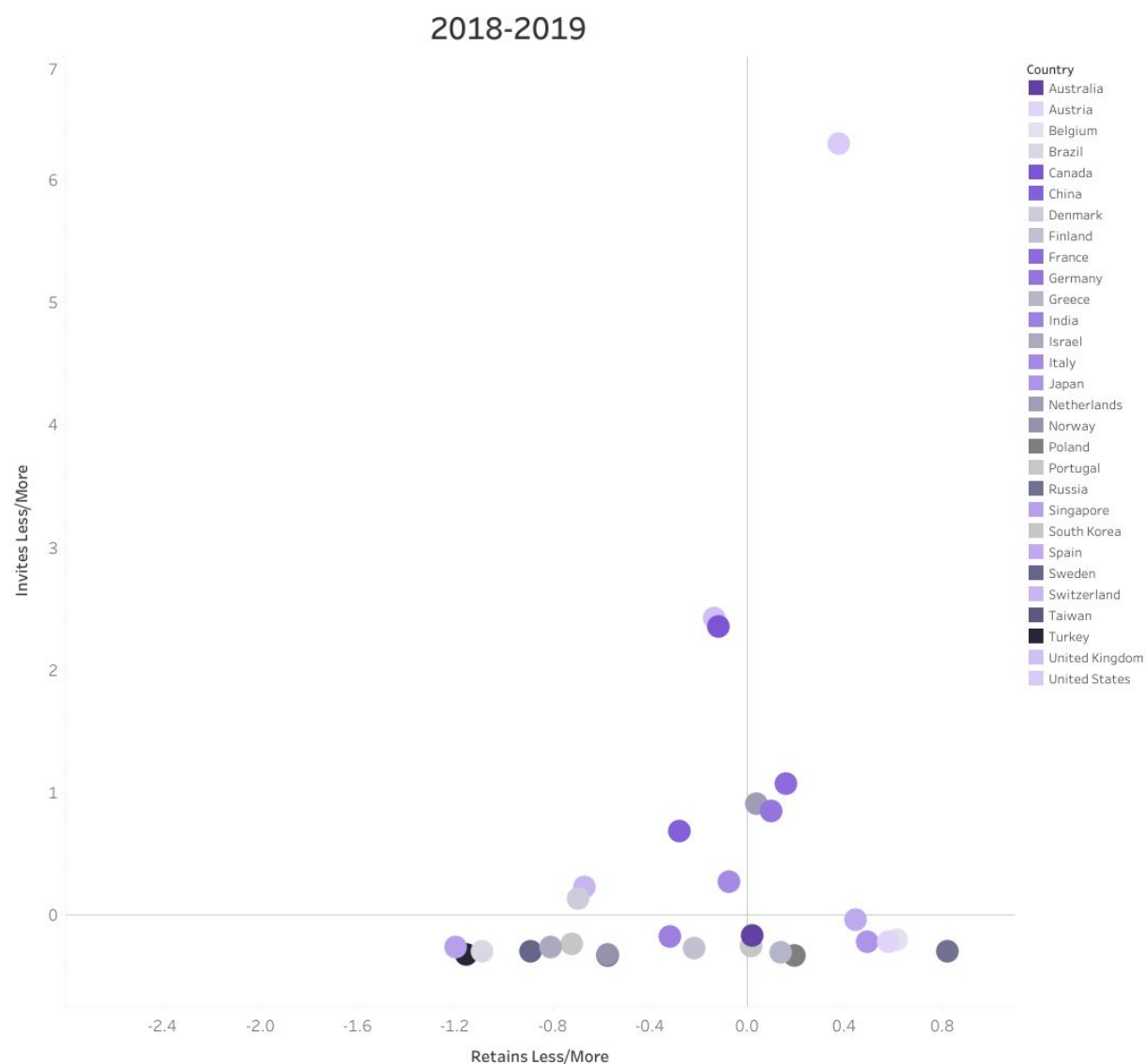
The graphs above show a comparison of countries' capacity to retain and attract talent. The X axis represents capacity to attract talent compared by measuring the chance that authors start in another country and end up in a target country. The Y axis represents capacity to retain talent by measuring the chance that an author that starts in a country stays there. Both axes are measures of standard deviation from the average of all countries.

By looking at the location of author affiliations year to year, we were able to compare the talent inflow and outflow in each country as a percentage of the country's overall talent pool of authors. Talent flow was measured by comparing the chance of an author moving to and from countries compared to the average chance for all countries. We calculated the average inflow and average outflow of all the countries and then looked at each individual country's distance in standard deviations from the average inflow and distance from the average outflow to give a normalized score.

The reasoning for a normalized score here is that countries don't exist in a vacuum. One country's capacity to pull talent is another's incapacity to retain it. We are comparing to an average so that we can see how the efforts of each country affect their weight in the global market for research talent.

The “inviting” measure shows the chance that authors end up in a country, representing the pull an ecosystem is able to exert on talent, whereas “retaining” is the measure of the chance that an author stays during a year.

To see how countries compared in this push-and-pull dynamic we plotted the “inviting” scores on the x-axis and “retaining” scores on the y-axis. These values allowed us to categorize countries into the four distinct groups outlined below based on the quadrant of the graph they fell into.



Comparison of countries' capacity to retain and attract talent. X axis represents capacity to attract talent as measured by the chance that authors start in another country and end up in target country. Y axis represents capacity to retain talent as measured by the chance that an author that starts in a country stays there. Both axis are measures of standard deviation from the average of all countries.

### Producer Countries - Bottom Left Quadrant

We think of India, Singapore, and Israel as producer countries, because they saw less inflow and more outflow, as a proportion of the country's talent pool, than average. Israel has seen some setbacks going from 0.3 on its best year to -0.8 in 2018-2019 for retaining talent, as has Singapore, which went from 0.16 in its best year to -1.19 in 2018-2019 for retaining.

Many of those publishing on arXiv are developing a skill set that is desired along the value chain, and so these countries that have a higher outflow are seeing their talent investment leaving to create value in other countries. Singapore is an important example of the danger, as this outflow contributed to their zero growth in their supply data.

### **Anchored Countries - Bottom Right Quadrant**

Japan, Belgium and Russia are what we call anchored countries. They had less talent inflow and less talent outflow, as a proportion of the country's overall talent pool, than average. It signals the relative stability of their talent pools, but perhaps also risky insularity.

Japan may rank 7th for researchers, but has some of the least movement in and out of the country, and is below average for international collaborations (28th overall). As AI moves quickly, it is important to be well connected with the global community to keep up with the latest advancements.

Serbia does not figure in the graph because of its number of authors is too small, but it is worth noting that it made a big improvement from a -3.7 to a +.98 "staying" score, which can also make this a section for countries that are making the important first step of retaining the talent they create, if not attracting it from abroad. Belgium and the Netherlands have also shown that working on affecting talent flows have paid off in sustained growth. They have gone from being developing ecosystems that struggled to hold on to talent, to steady growth relative to their population, despite being smaller than their neighboring countries (from -0.2 to 0.9 inviting for the Netherlands, and Belgium went from 0.3 "staying" to 0.6, doubling their comparative capacity to retain people). The Netherlands has actually done well enough to qualify as an "inviting" country.

### **Platform Countries - Top Left Quadrant**

Several countries saw both more inflow and more outflow, as a proportion of the country's talent pool, than average. These countries are succeeding at attracting workers who were trained abroad while also seeing significantly more movement out than average. These ecosystems, which we term platform countries are best exemplified by the U.K, China and Canada. These countries are known for their competitive AI hotspots and present attractive options for talent in terms of high-calibre peers and institutions. However, these successes can generate competitive pressure on their ability to retain talent, especially through attracting the interest of international AI labs to set up shop locally.

The UK and Canada have had large gains. The UK went from an historically low 0.05 "invite" in 2014/2015 to 2.4 this year, and Canada went from 0.3 "invite" in 2015/2016 to 2.3 (with most of the gain happening in the last year). Switzerland's strategy has shown success for inviting talent -0.2 in 2014/2015 to 0.2 in 2018/2019, perhaps due to its big income difference with Italy just south of the border, but has also cost them in retaining power (0.26 to -0.66).

### **Inviting countries - Top Right Quadrant**

The US, France and Portugal all saw more inflow and less outflow, as a proportion of the country's talent pool, than average. This means that these countries are relatively more successful at both retaining the talent they started with and attracting more talent from other ecosystems. We call these ecosystems inviting countries.

The US dominates in attracting talent in research and academia, and has plenty of jobs to keep people there after graduation.<sup>12</sup>

According to a [survey](#) by The Center for Security and Emerging Technology (CSET), 38% of US AI PhDs did their undergraduate degree abroad and 48% were born outside of the US. This is in line with their findings

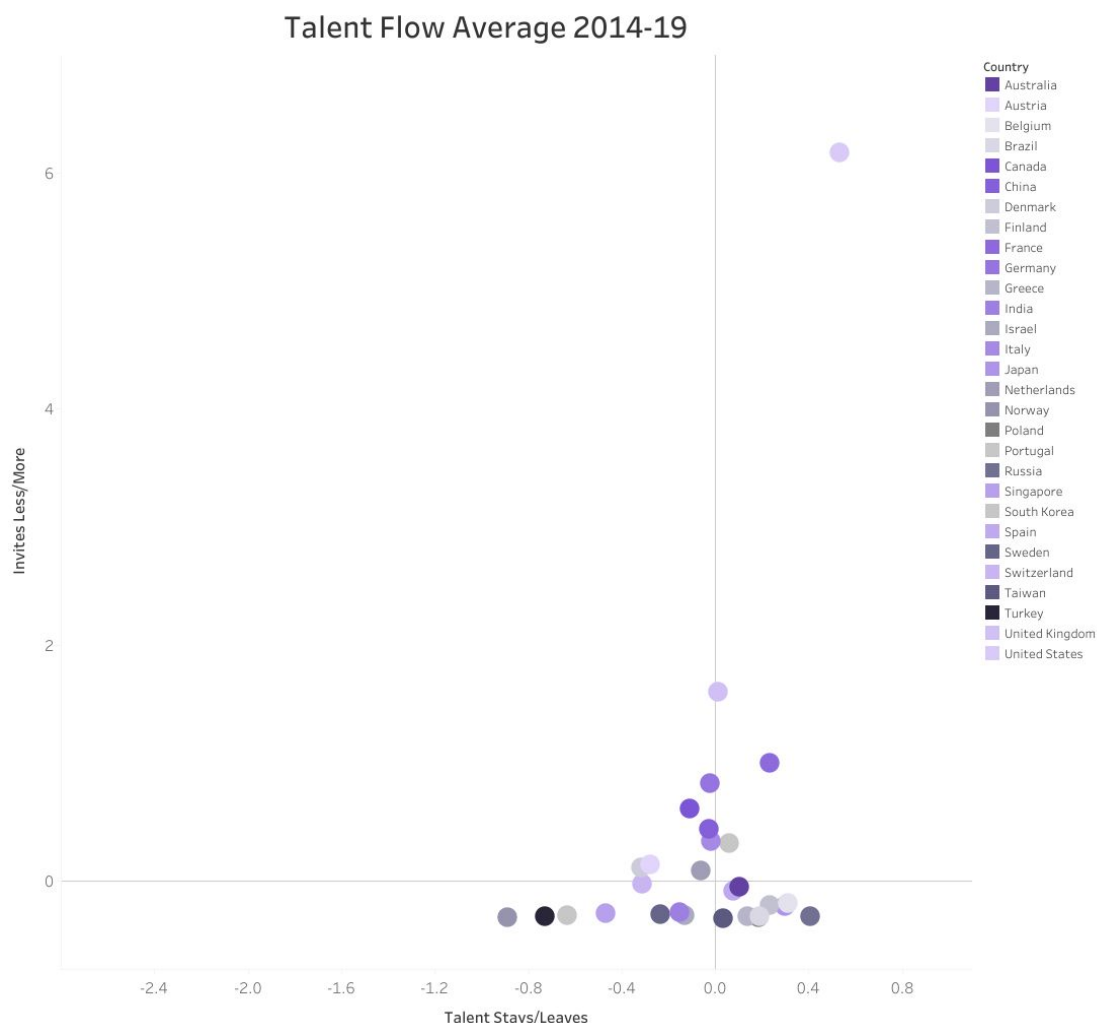
---

<sup>12</sup> Since our study defines a researcher's location based on the headquarters of the company where the researcher works, these results could be seen to inflate US numbers. This is because many of the companies establishing labs around the world are headquartered in the United States.

that 55% of the US STEM PhDs are US citizens, while Chinese and Indian nationals make up 16% and 6% respectively. At some universities, the percentages of foreign graduate students is notably higher: at New York University's Tandon School of Engineering, for example, [80% of graduate students are reported to have come from abroad](#). The CSET also found that between 82–92% of the US AI PhDs stay in the United States to work in the first five years after degree completion. While the primary motivator for coming was the quality of education (82%), only 42% said job opportunities were a factor, so this situation may be mostly a factor of convenience.

However, the stricter rules proposed by the new H-1B Visa policy in the US is making it more difficult for students to stay after graduation, or even enter as an international student with some of the more strict visa policies proposed. This represents an opportunity for countries with ample job markets to attract and retain talent, and will be important to follow in the coming years. Many countries seem to be able to attract and/or retain talent much better than they were 5 years ago, and so could be well positioned to take advantage. Canada in particular has shown it can. In 2017-2018, Canada had a 0.062 score on invites, for 2018-2019 it has a 2.347 score, winning back a lot of the talent it had lost in previous years to Silicon Valley.

Still, the US is starting with a big lead. On average, our data show there is a 4% chance of those doing research outside of the US to end up publishing there. By comparison, there's a 1% chance of someone ending up in the UK.

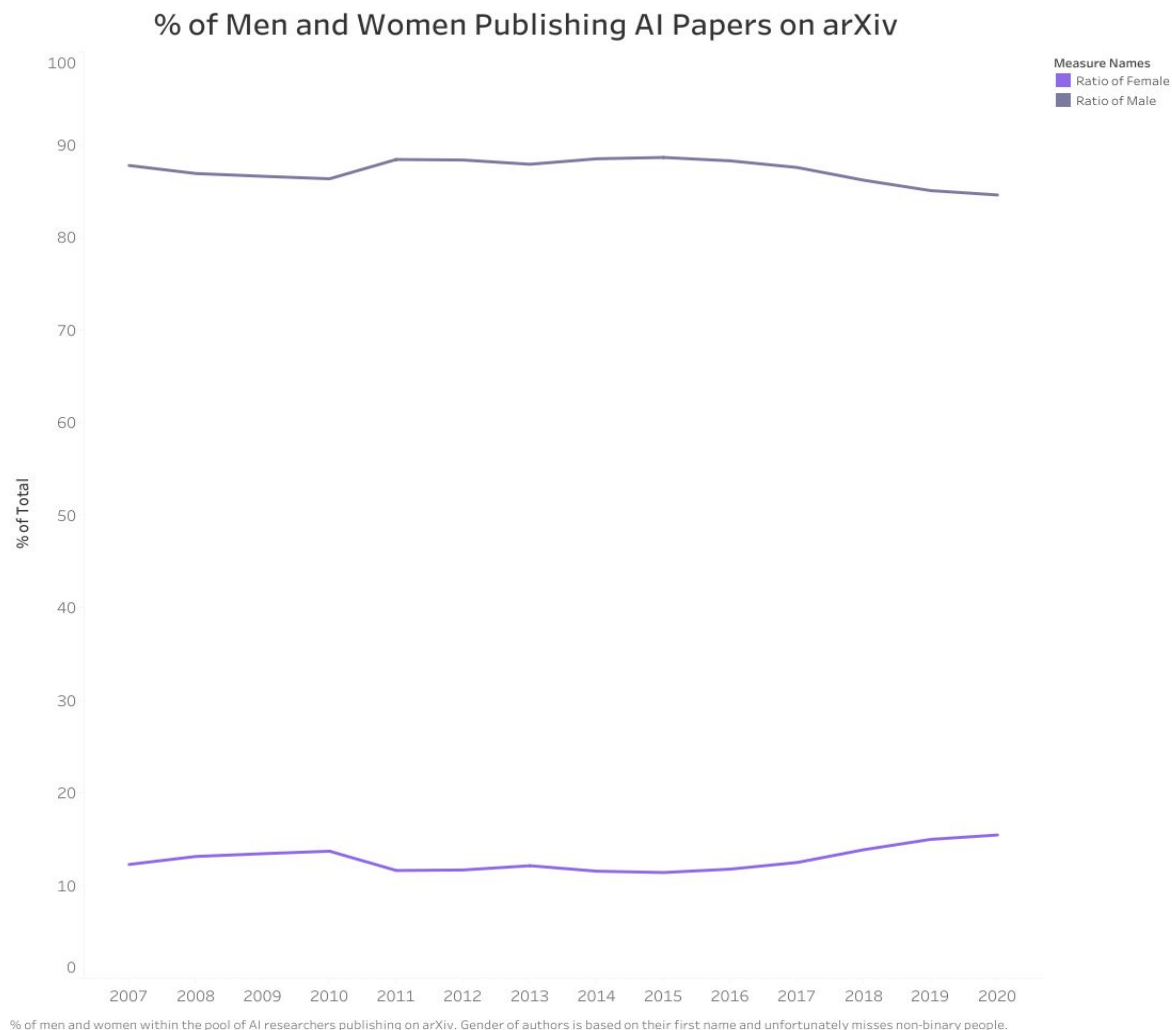


These talent flows should be watched carefully in the next year to see how the pandemic has affected geographic mobility. AI workers are sophisticated and able to express their talent via internet connections, which theoretically makes them 100% “mobile”. This mobility may undermine the natural advantages of the countries with the top schools and jobs—and increase the competition by allowing talent to better pick their preferred mix of compensation (and tax), locations (for life and work), field of application, and colleagues. However, it will be a testament to the importance (or lack thereof) of proximity to other researchers, versus simply contact and access. Studies indicate this proximity produces better results, but the question remains if researchers see it that way.

## Gender Balance

The ratio budes only slowly over 13 years, but the stories vary widely between countries

The gender measure was based on the names of the authors. Using the US census data, we created a list of probabilities for each name and kept those with a higher than 95% probability of being of one gender rather than the other.<sup>13</sup> We recognize this as a crude measure due to the ambiguity of many names, the US-centricity, and of course how people may not identify with either gender.

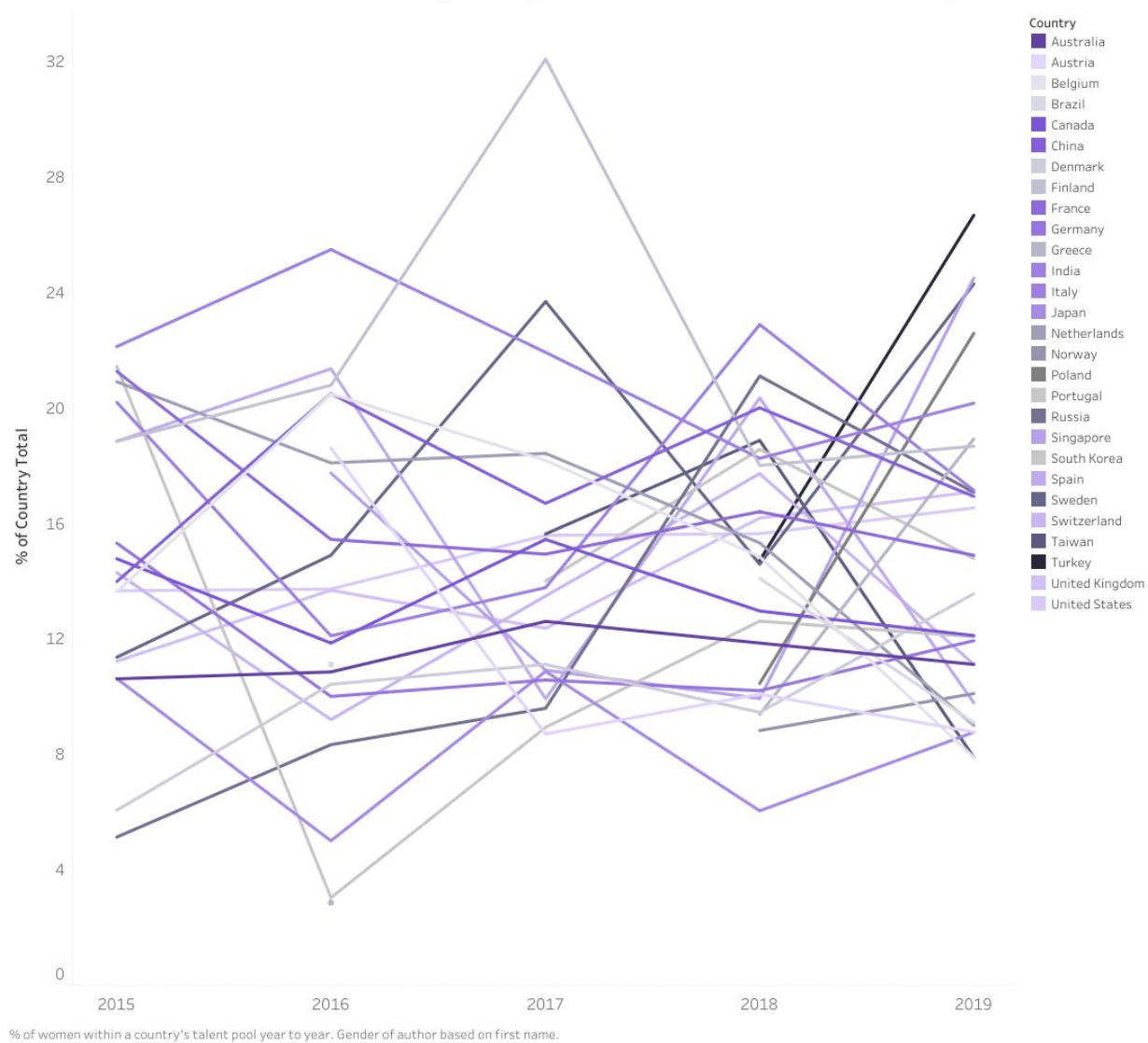


<sup>13</sup> Based on the logic from this paper: <https://journal.r-project.org/archive/2016-1/wais.pdf>

When looking at the aggregated ratio, there is slight convergence. The percentage of women has grown from 12.26% in 2007 to 15.44% in 2020. Steady, but slow. It is difficult to assess country performance here because we can see wide fluctuations of the overall ratio year over year. These fluctuations are likely due to many publishers being students, who then move out of research and/or move to other countries after graduation. This movement away from publishing is enhanced for women, who tend to get jobs with fewer publishing opportunities than men. Therefore, a given year's ratio for a country depends a lot on the cohorts of incoming students, which makes a more volatile result.

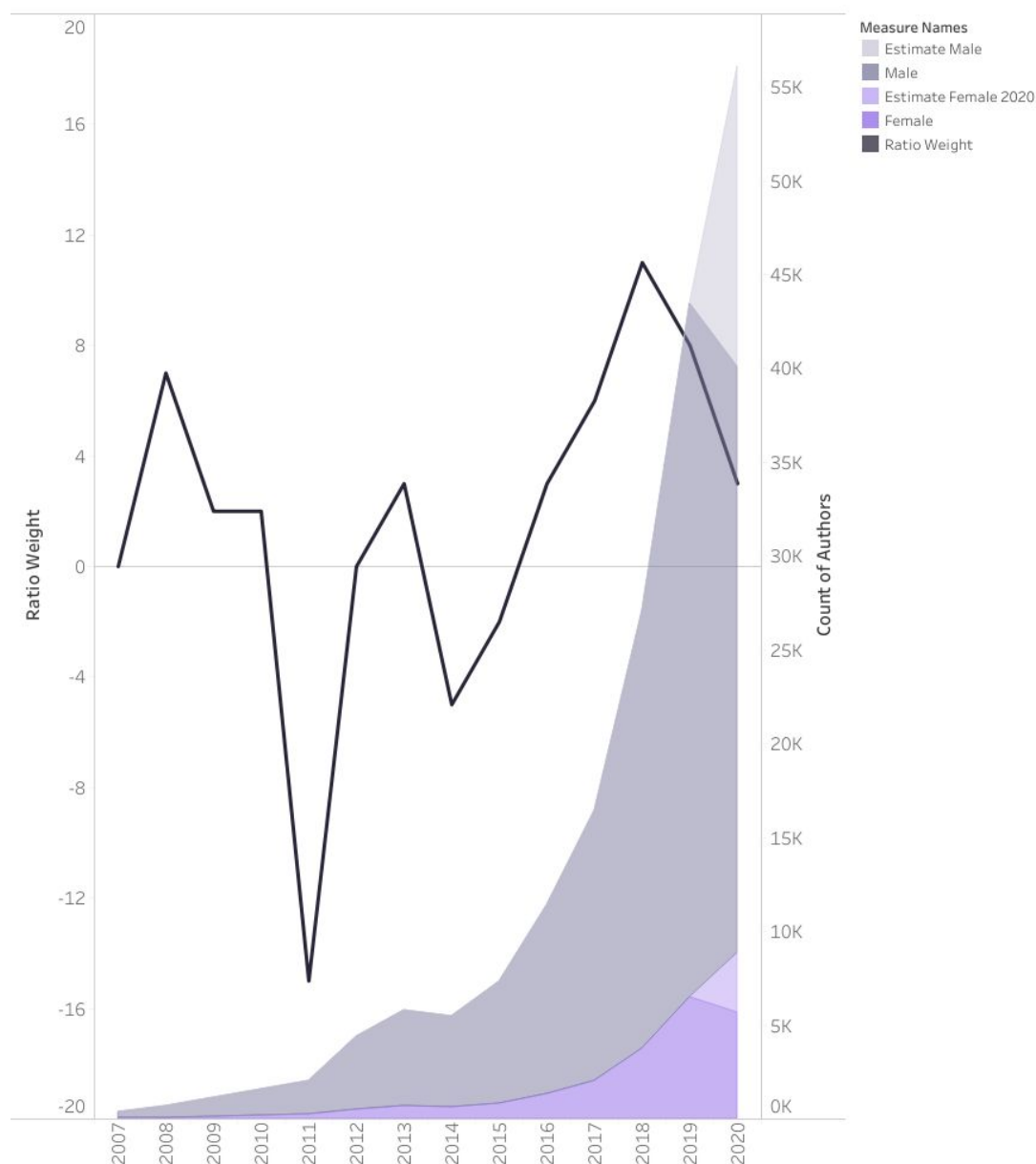
Some countries have shown significant and persistent growth in the data: Turkey (26.67%), Singapore (24.49%), Sweden (24.29%), Poland (22.58%), Greece (18.92%), Russia (17.07%), and Denmark (13.56%). The primary influencers of the overall ratio however remain the US (16.53%) and China (16.93%) due to their size. Given the recognition in the field of the value of diverse leadership, women will be in high demand, and retaining them will likely have a positive impact on attracting more women to the field in that country.

Ratio of Women Publishing AI Papers on arXiv for Each Country



Observing the overall growth rate, we can also see that there are years where the growth of male authors far outpaced female authors, causing a drop in the ratio. Since 2007, the average growth rate of women has been 55.44%, where for men it has been 51.21%. This should be enough to close the gap over time, but the years of missed relative growth stall this. If the average rates of growth continue for the rest of the year, the ratio will continue to balance. But, the impact of the pandemic has been shown to [disproportionately impact women](#). Given these uncertain times and the challenges of the return to school, we can't disregard the possibility of a lower number of female authors in 2020 compared to 2019, which would be an unfortunate first.

## Number of Men and Women Publishing AI Papers on arXiv



Number of men and women publishing, with gender estimated from their first name, and trend line showing annual percentage change in ratio of women. 2020 data is as of Aug 2020. Estimates for end of 2020 based on average of weighted growth of last 5 years. Excludes names that the algorithm had less than 95% confidence in classifying.

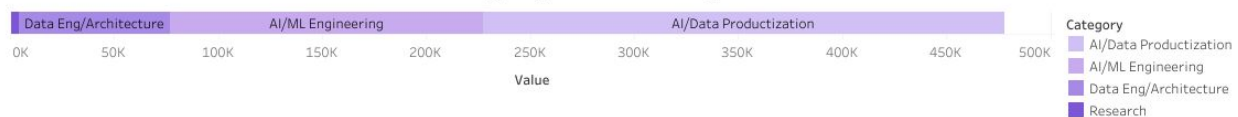
# The People to Build It Out

## Supply available to industry — how AI talent shape themselves to the commercial market

Using the standard roles and titles we categorized, we collected self-representation data from social media. The collection was based on keyword usage. We searched both by title (e.g. “data scientist”) and keywords for their skills lists (e.g. “machine learning” + “tensorflow” + “PhD”). Growth is a measure of those with AI jobs for the first time, covering roughly the last 1-3 years.

In total, we counted **477,956 people worldwide in the below roles**. As typologies of skills vary from country to country, it is difficult to find previous estimates of the total talent pool to compare to. One data point could be Tencent’s 2017 estimate of 300,000 people, which accounted for *all* employees working in AI companies or departments (including support and administrative staff). With a focus on technical talent, our measure certainly shows a significant growth in response to the increased demand over the last few years.

Total of Each Role Category in Industry Talent Pool



The proportionate mixes are roughly equal between supply and demand globally (when excluding data engineering/architecture), generally 61% AI/Data Productization, 38% Engineering, and around 1% Research.<sup>14</sup> There are no countries where the share of researchers gets close to that of the other roles, the closest is Canada with 2.36% of its talent pool being researchers. However, some countries have more engineers than implementers: France (61.20% engineers) and Germany (63.21% engineers). On the other end of the extremes, there is China with 86.82% implementers.

The total number of self-reporting researchers is 4,149. This is much lower than the total publishing at conferences last year (22,000) and the total on arXiv (projected to be 86,000 by end of year). This tells us that few fundamental researchers list themselves on social media oriented towards industry job markets, or if they do are perhaps marketing themselves as engineers to get the applied jobs that are available. While we cannot say precisely from our view how many researchers continue doing fundamental work, the demand data below shows that the supply is responding to demand.

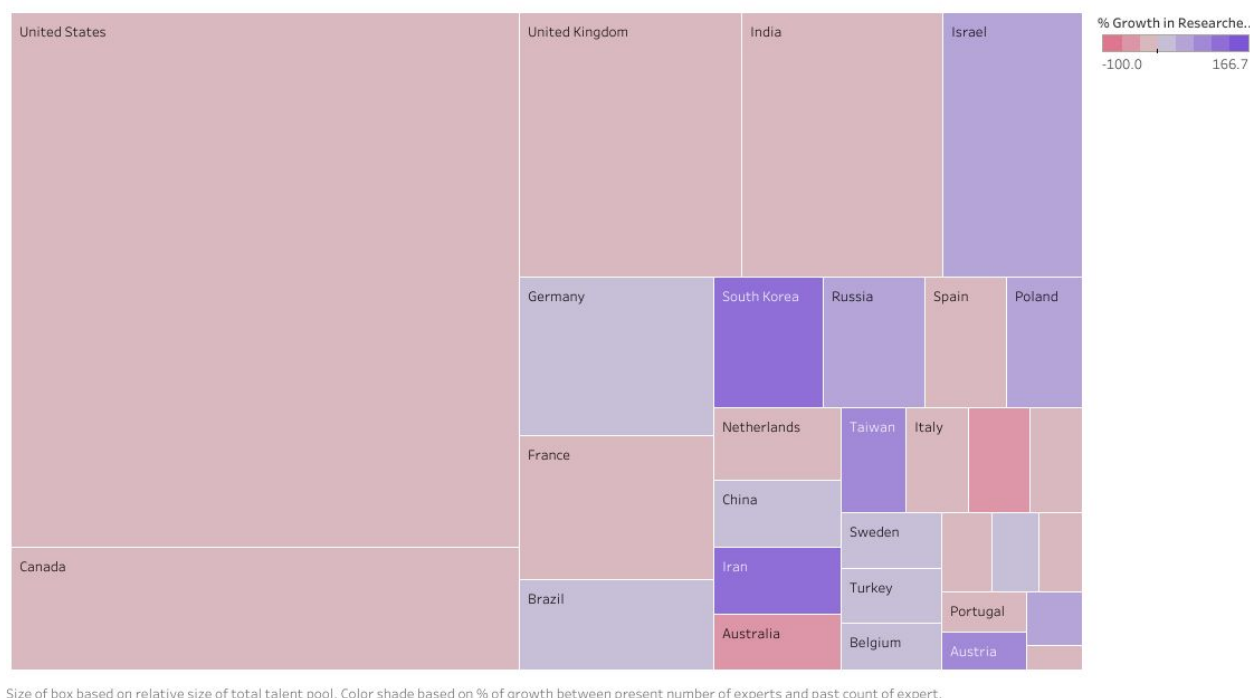
In other words, this data represents those researchers who continue to research professionally in the private sector, showing the difficulty in continuing to publish novel fundamental work.<sup>15</sup> Canada punches

<sup>14</sup> While the relative scale to the other roles is similar for researchers, the difference between supply and demand of researchers is much more stark. Supply here is about 1.02% of the total (again, excluding data engineering/architecture), whereas demand is 1.77% of the total (see the next section on demand).

<sup>15</sup> The Center for Security and Emerging Technology’s survey of AI PhDs showed that 73% did applied research as part of their work, as opposed to 53% doing basic research and 37% doing engineering work. Though, they admit this may be a skewed sample with 54% of respondents coming from Academia: “Recent CSET research analyzing the career paths of US AI PhD graduates from top-ranked programs between 2014–2018 based on CV coding found 34 percent work in academia and 60 percent work in the private sector.”

above its weight here, which is perhaps explained by the number of international firms coming to set up AI labs specifically aimed at attracting research talent.

## Research - 4,149 total

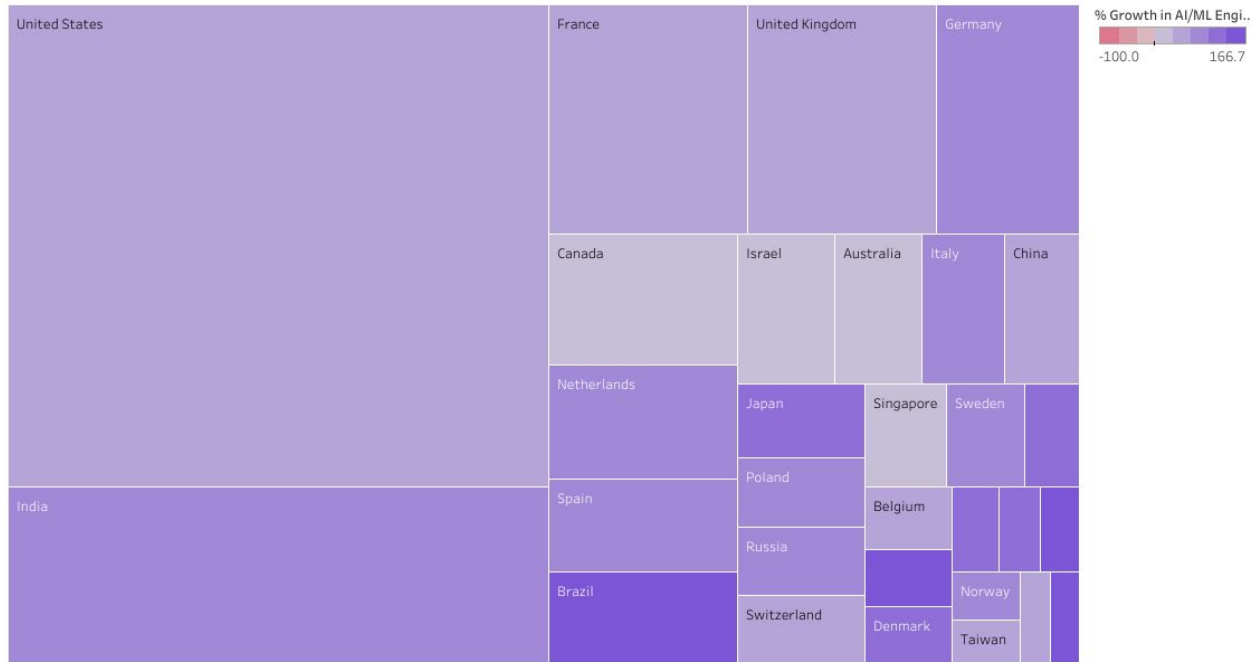


Some countries of note in terms of researcher growth are South Korea with 133% growth, Taiwan, Iran and Austria with growth rates of 90-110%. Israel, Poland, Russia, and Greece are also showing significant growth, between 35%-60%. On the other end, Australia and Switzerland have gone negative at -43%. Shrinking numbers in the top countries may show a trend towards more researchers moving into engineering and implementation, but the AI industry should be careful to not underinvest in research.

We've found that the growth rate of specifically AI devs (-68.8%) is inverse to that of data scientists (102.7%). However, AI devs and data scientists often have the same skill set today and AI devs are able to simply change their title for higher pay of the data scientist playing an engineering role (about \$30,000/year more on average according to Glassdoor). This ability to change title may explain why these two roles are moving in such drastically opposite directions. This isn't true for a data analyst (-28% growth), also in the productization category, who is much less likely to have or need the same depth of technical expertise for their role, but may also be seeing negative growth as people skill themselves up into data scientists.

One reason the skill sets have been similar is because AI devs often need a good understanding of the underlying engineering in order to fit novel AI tools to software. The challenging problems to solve in AI software development is the same reason AI has been largely inaccessible to smaller teams who can't afford both ML engineers and specialized AI devs. The aim to democratize AI with standard, out-of-the-box tools could allow the average developer with some online AI training to implement AI in their software. This is good for smaller teams as it allows them to put more of their budget towards implementation rather than engineering, and still be able to tweak the out-of-the-box methods with some customization.

## AI/ML Engineering - 150,559 total



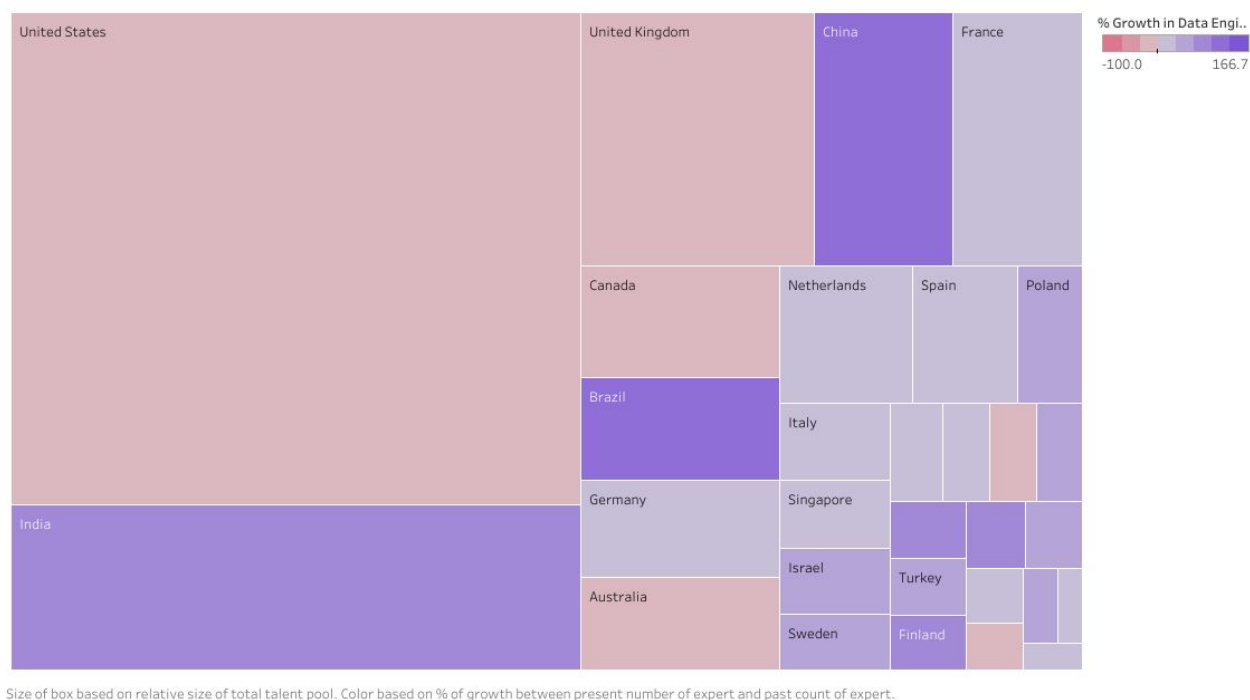
## AI/Data Productization - 250,500 total



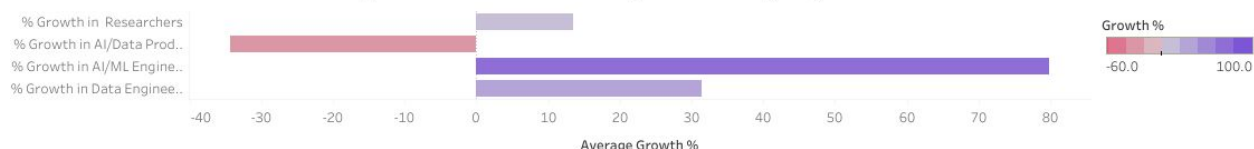
One trend to watch is emerging ecosystems with a growing number of data engineers. This can be a sign of organizations who have built out AI solutions, but have then encountered the challenge of getting it to work at scale on live data flows. This isn't a fault, as we have seen it as a signal of maturing towards truly

integrated AI systems. These countries are Brazil (110%), Finland (87%), India (83%), and South Korea (94%). The large ecosystems that have shrinking numbers of data engineers amidst an overall growth of the category, may be a sign of offshoring or migration of those in the role, as well as a consolidation (and even automation) of the work in companies [who are building the tools to supplement the work](#).

## Data Engineering/Architecture - 72,748 total



## Average Global Growth per Category

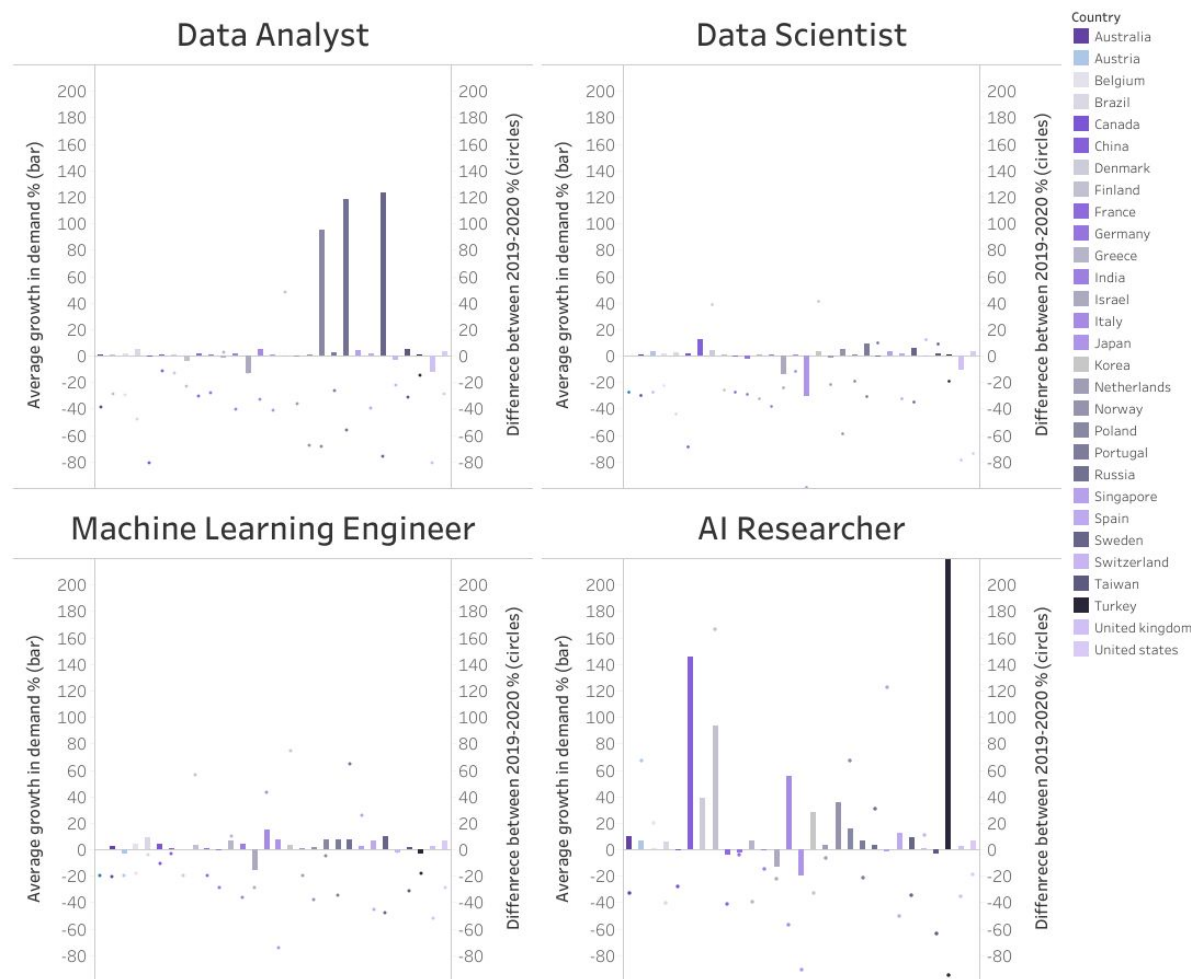


## Industry Demand for Talent

Demand for new roles was stable pre-2020, then fell 20-30% during the pandemic—but some outliers stood out

To measure an indicator of demand, we searched job aggregator sites for relevant job titles (e.g. “data scientist”, “machine learning engineer”, etc.) on a monthly basis to collect the number of job postings. We then compared the variation from month to month to measure the growth in demand for each job title.

Unfortunately, we only know the net growth, and so do not know how many job postings have been closed versus re-opened, or if jobs left unclosed have in fact been filled but not taken off the list.<sup>16</sup>



Average growth in demand for job title, based on averaging variance in number of job posting with related keywords from month to month.  
Difference between 2019-2020 is difference between 2019 monthly average and 2020 monthly average.

The overall trends show that the median growth rates for data analyst, data scientist and ML engineer roles have all attained a certain stability in 2019 between 1.24% and 3.28%. ML researcher roles meanwhile have a much higher median growth rate of 6.28%. Add to this that researchers make up 1.77% of demand measured here, while they only make up comparatively 1.02% of the talent supply available to industry, we can expect that competition will get even tighter for research talent and more researchers may be drawn out of academia.

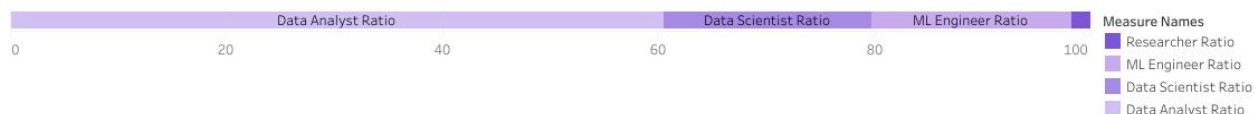
Notable outliers are Poland, Russia and Sweden who are booming (between 95% and 125% growth) when it comes to the growth of data analysts. We can also see important outliers in the demand for ML researchers with Turkey (225%), China (145%), Finland (166%) and a handful of ecosystems still actively

<sup>16</sup> One should note that the demand data are not separated into the categories we laid out in the section “Specialized Technical Roles”, and are only for the specific job title mentioned in each graph as we did not have 2019 numbers for all of the titles to assess growth rates. It is also important to note that “machine learning engineer” and “data scientist” would both fit the AI/ML Engineering role category, and that our data for demand does not cover the Data Engineering/Architecture role category.

looking for researchers: Italy, Denmark, Norway, South Korea, Poland, and Spain all between 10% and 55% growth.

These numbers can signify local shifts in strategy as well as new entrants on the scene. The boom in demand for researchers is in part seems like it is largely driven by China looking to extend its success in inviting more academic researchers (seen earlier in the section “Talent Attraction and Retention”) to other industrial researchers, taking advantage of the many Chinese nationals in the US suddenly running into visa challenges.

### Average Ratio in Demand for Each Job Title

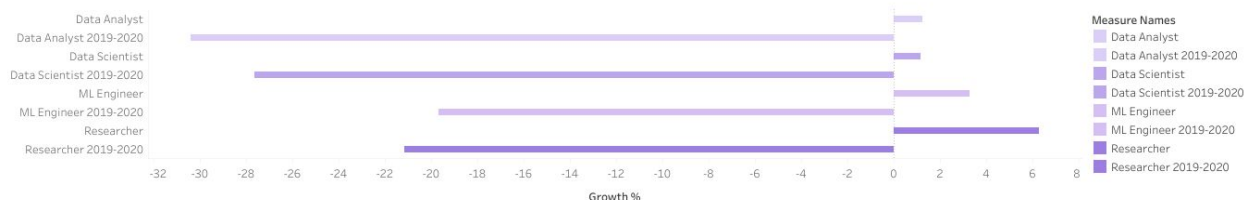


The difference we see in demand between “data scientist” and “ML engineer” show that they are virtually the same proportionately in terms of job offerings. We see them as requiring the same skill set as the top data science work is indeed ML work, thus both parts of our role category of AI/ML Engineering. However, monthly job postings for “ML engineer” grew at twice the rate (3.28%) than those for “data scientist” (1.17%), indicating that many countries are intentionally ramping up for ML innovation and the AI skill set has perhaps not yet normalized.

Looking at the 2020 data so far, we can see the impact of Covid-19. Almost every ecosystem saw a drastic slow down in demand for AI talent. Comparing the annual averages for 2019 and 2020 up until August, we can see that job postings for data analysts slowed down by 30%, data scientists by 27%, ML engineers by 20%, and researchers by 21%.

Certain ecosystems have taken the slow down as an opportunity to grow, with countries persisting in their new growth, or even jumping in for the first time, during the pandemic. South Korea raised its hiring across the board except for researchers in 2020. Singapore did the same and inched up its hiring for all roles and even doubled down on researchers (122% growth in 2020), perhaps realizing that the supply was not growing on the island.

### Median Growth of Demand



Median growth in demand for each job title, based on averaging variance in number of job posting with related keywords from month to month. Columns with no year are median for 2019. Difference between 2019-2020 is difference between 2019 monthly average and 2020 monthly average.

Highly established ecosystems have slowed down demand more significantly compared to developing ecosystems who are playing catch-up. Early results from McKinsey’s latest Global AI survey, publishing in November, shows that respondents from a majority of organizations are more likely to increase AI investments over the next three years than to decrease them. We will have to see whether that will revamp the demand for talent, or if more of the investment will go towards infrastructure in the recovery.

For larger players who want control over their tech stack and data to drive novel applications and solutions, it will still be important to hire and develop data science talent. But as they standardize their

own tools, they are increasing demand for more implementation roles. Further, as more out-of-the-box capabilities become available, the people further down the value chain working on implementations of working models will be key and should expect to see high demand for their skills.

## Conclusion: Bridging the Talent Gap

Whether the full potential of AI is over-hyped is [for another discussion](#), but we can say that practical AI success is not just a formula of high-level experts and access to the right data. The AI industry initially focused on the very high-level experts because only they could administer the new techniques emerging with the right knowhow for applying them to novel areas. Now there is a recognition that the dynamics of this new technology require more than just engineers and people who can build nice models in order to deploy it effectively.

AI is a new generation of software that is adaptive to the data fed to it; it is coded with data rather than logical rules. Traditional software is static by comparison, and AI needs a new ecosystem of support and infrastructure to not only be built but also governed once it is deployed. For AI to work at scale, lots of new talent is needed for engineering, building infrastructure, developing new business models, and monitoring objectives.

[A survey by The European Commission](#) recently found businesses identified access to the right skill sets as being the number one impediment to adopting AI. Our Global AI Talent Report is just a first look at these professional roles and how mismatches in demand and supply may continue. A lot remains to be done to paint a full picture. We suggest a number of ideas in “Appendix: For Future Research” to add depth and help create a clearer view on who has the needed talent to deliver AI, and where they can be found.

As AI matures it will become more pervasive. We will see new specialized roles emerge for managing the new dynamics of AI, but eventually everyone will need to update their digital skills to collaborate with this new technology. Already, we have seen that most people can grasp the concept of an AI-powered recommendation algorithm, and adjust their behavior to affect the output of the algorithm. However, people have very limited choice and only blunt tools for manipulating an algorithm to their needs. When the different tooling and skill sets standardize along the value chain, it will vastly increase the choice and access to AI technology and engender far more innovation than we have yet seen with AI software.

To get there, we have a challenge of bridging the gap between proof-of-concept in the lab and real world deployment. Researchers and engineers play an important role right now in helping close that gap, but they cannot do it alone. They, and the institutions that train them, need to focus on standardizing their tools and processes so that others can more easily collaborate with others down the value chain.

## Appendix: For Future Research

### **Searching for skill sets rather than titles**

A key update to better understand supply and demand will be to take a skills-based approach rather than title-based. Titles are likely to shift as roles standardize and new ones emerge. Tracking combinations of skill sets will more quickly identify how those roles take shape. Further, sourcing the needed data for this will likely add more granularity that will allow for the kind of analysis we do on the research data, such as measuring geographic movement and gender.

### **Better understanding the relationship between industry and academia**

It is difficult to assess precisely what the impact of industry is on research. We know that it can over-hire, leaving institutions dry and unable to train the next generation. But it can also generate a lot of productive and useful work by giving researchers access to real-world data and problems. To better answer this question, some straightforward additions would be:

- Average publishing lifespan of a researcher
- Whether affiliations are corporate or academic
- Splitting research into applied and fundamental, and subcategories of each where possible
- Better understanding of movement by gender, visiting faculty, and researchers in private labs

### **On-the-ground understandings of different regions**

Integrating survey data from different locals would be an important supplement to understand the stories behind the data, or where the data has a blind spot. China is likely very underrepresented in this report, as are other countries who publish in non-latin alphabets or are simply at an earlier stage of maturity. Even in developed ecosystems, regional data will help with understanding distribution of private labs and the weight of different hubs.

## **Appendix: Methodology & Caveats**

### **Typology and Social Media**

We consulted with industry experts from within Element AI and outside about their experience with AI related projects. From those conversations, we broke down the types of expertise needed to bring AI from theoretical ideas to concrete products, which led us to a series of job titles and descriptions that form our working hypothesis about AI expertise. While the reality on the ground will tend to have more blurred boundaries, it is our belief that these categories well represent the different skill sets needed. Using the job titles, we then collected self representation data from social media. The collection was based on keywords found in job titles (e.g. “data scientist” as a job title), in skill lists (e.g. “machine learning”) and in technology know-how (e.g. “tensorflow”). The growth rate is a measure of those with AI jobs for the first time over those who have had previous jobs with the same role, covering roughly the last 1-3 years given average job length in the tech industry. The nature of this measure does count those in the same first job for a long time towards growth.

### **Job demand**

To measure an indicator of demand, we searched job aggregator sites for different job titles (e.g. “data scientist”, “machine learning engineer”, etc.) on a monthly basis to collect the number of job postings. We then compared the variation from month to month to measure the growth in demand for each job title.

### **arXiv and Demographic Data**

Last year we concentrated the research on the most prestigious scientific conferences around AI, with the goal of understanding the most impactful research. This year we broadened our look into research by trying to have the closest thing possible to a “census of research” that would give us a better understanding of how AI research is unfolding: from a small circle of researchers working on making AI real, into a constellation of experts finding ways in which AI can be implemented in different fields. In other words, this new methodology should enable us to focus not just on “fundamental AI research” and broaden the scope to encompass more applied AI research.

Comparing last year’s report with this year’s highlights the volume of academic research done around AI and its applications that doesn’t make its way into conferences. In our new approach, industry has less influence on the overall numbers.

We downloaded all of the papers from the cs.AI, cs.LG and stat.ML repositories, as they were the ones with the most direct relation with AI, and the least amount of chance of containing papers not related to AI. Within these papers, we kept all those that had either a Latex, PDF, Tex, or other text readable file, leaving out others who only had either HTML files or other coding language files. From there, due to technical limitations, we only kept papers using latin script; leaving out some papers such as those written in Chinese characters.

With this sample, we created a heuristic to extract the author names and affiliations from each paper. The heuristic was created by going over 3,500 papers and creating a list of all possible affiliations found in the paper. From this list of potential affiliations, we extracted the affiliations from the papers with a regex script. This left us with 35,418 papers that contained affiliations recognized by our script. From the author/affiliation list created, we then defined the most statistically probable affiliation for each author for each year based on the affiliations recurring in each paper the author had. This left us with the probable affiliations for each author for each year they published. This methodology, tested against a sample of 200 papers, came back with a 98.7% success rate.

The localization of each affiliation was derived by hand and was based on the same logic as past reports: university systems (e.g. the University of California (UC) system or UParis system) were simplified (i.e. all UC universities are counted as simply being in California), and all companies are counted as located at their headquarters (e.g. all Google labs are counted as being in California). This logic is based on two considerations: the first being pragmatic in nature in the sense that, for example, Google has labs around the world, but authors rarely specified which one they work in; and the second one is more to address an underlying point of the report where we postulate that most labs don't keep the value of their research in the lab's host country but bring it back to the headquarters of the organization.

The gender measure was based on the names of the authors. Using the US census data, we created a list of probabilities for each name and kept those with a higher than 95% probability of being of one gender rather than the other. (based on the logic from this paper: <https://cran.r-project.org/web/packages/gender/gender.pdf>)

Talent inflow was measured by comparing the chance of an author moving to and from countries, compared to the average chance. Inviting is the measure of the chance that authors end up in a country, while staying/leaving is the measure of the chance that another move/not move during a year. The average was calculated by averaging chances and comparing to the average of all averaged chances.

## Thank You's

**Written with** Simon Hudson and Yoan Mantha

**Data research and visualizations by** Yoan Mantha

**Special thanks to those at Element AI and elsewhere who provided invaluable commentary and support:**

Alexandra Mousavizadeh of [Tortoise Media](#)  
Bruno Lanvin of the [Portulans Institute](#)  
Helen Mullings, Paul Walsh and Andrew Fyffe of [Quantum Black](#)  
Michael Chui and David DeLallo of [McKinsey Global Institute](#)  
Time Davies and Nicolás Grossman of the [Global Data Barometer](#)

and

Frédérique Bouchard, Valérie Bécaert, Simon Bélanger, Annabelle Martin, David Bédard, Catherine Lefebvre, Jean-Philippe Reid, Adam Salvail, Lara O'Donnell, Pierre-Luc Beaubien, Benoit Hamelin, Julien-Pier Boisvert, Christian Jauvin

**Thanks to the team supporting the distribution of this report:** Kayla Gillis, Morgan Guegan, Guillaume Gagnon, Kevin Clark, Julien Desrosiers, Marie-Calude Savard, and Robyn Crump

**Translated into French by** Melissa Guay and Guillaume Gagnon