Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology

J. Klinger, J. Mateos-Garcia, and K. Stathoulopoulos

Nesta, 58 Victoria Embankment, London, EC4Y 0DS, United Kingdom

Abstract

General Purpose Technologies (GPTs) that can be applied in many industries are an important driver of economic growth and national and regional competitiveness. In spite of this, the geography of their development and diffusion has not received significant attention in the literature. We address this with an analysis of Deep Learning (DL), a core technique in Artificial Intelligence (AI) increasingly being recognized as the latest GPT. We identify DL papers in a novel dataset from ArXiv, a popular preprints website, and use CrunchBase, a technology business directory to measure industrial capabilities related to it. After showing that DL conforms with the definition of a GPT, having experienced rapid growth and diffusion into new fields where it has generated an impact, we describe changes in its geography. Our analysis shows China's rise in AI rankings and relative decline in several European countries. We also find that initial volatility in the geography of DL has been followed by consolidation, suggesting that the window of opportunity for new entrants might be closing down as new DL research hubs become dominant. Finally, we study the regional drivers of DL clustering. We find that competitive DL clusters tend to be based in regions combining research and industrial activities related to it. This could be because GPT developers and adopters located close to each other can collaborate and share knowledge more easily, thus overcoming coordination failures in GPT deployment. Our analysis also reveals a Chinese comparative advantage in DL after we control for other explanatory factors, perhaps underscoring the importance of access to data and supportive policies for the successful development of this complex, 'omni-use' technology.

Contents

1 Introduction

What do the steam engine, the electric motor and the microprocessor have in common? They are all powerful General Purpose Technologies (GPTs) that can be applied in multiple sectors creating waves of change that ripple across the economy [1]. It is not a coincidence that economic eras are often named after their 'core' GPTs: the Steam Age, the Age of Electricity, the Information Revolution and today, a 'Second Machine Age' driven by advances in Artificial Intelligence (AI) [2, 3].

The emergence of a GPT can also change the economic fortunes of nations and regions: it is hard to disentangle Britain's ascendancy from the steam engine, or the USA's from electrification and the combustion engine. The arrival of microelectronics and the Internet shifted economic power from the East Coast of the US to Silicon Valley in the West. Today, the rhetoric of an AI 'global race' implies that those countries that develop strong AI industries will be able to dominate more markets and industries. Governments across the world are responding with national strategies to grow their AI sectors [4] .

But where do GPTs such as AI appear and why, and how do they transform geographies of innovation and production? We still lack good answers to these questions. Although economic geographers and regional scientists have studied disruptive GPT-like innovations that create new opportunities for countries and regions, they rarely consider their links with the rest of the economy [5, 6]. Yet it is precisely this connectivity that defines GPTs, and could also explain where they emerge, and their geographical impact [7].

In this paper we seek to address this gap in the literature with an analysis of the geography of Deep Learning (DL) research, one of the technologies driving recent advances in AI systems that are increasingly being recognized as the latest GPT [8, 9, 10]. We consider how the geography of DL has evolved since its emergence in the early 2010s, and study its link with local research and industrial capabilities. Our analysis draws on the literature on technological discontinuities and a recent body of research on economic complexity and related diversity that looks at how the industrial and knowledge composition of regions and countries drive their diversification into new products and technologies [11, 12]. In doing this, we provide new evidence about the geography of AI research, a question of great interest for policymakers.

Our analysis draws on a novel combination of data sources and methods: we obtain our principal dataset from `arXiv`, a preprints site widely used by scientists and engineers, and identify DL papers in its computer science section with `CorEX`, an information theory algorithm that can detect clusters of related words in corpora of text. We also use data from `CrunchBase`, a technology business directory, to identify and map industries that are related to DL and might spur its development. Our data analysis pipeline illustrates the opportunities that novel data science methods create for the analysis of emerging technologies such as DL.[1]

The rest of this section reviews relevant literatures in economics, economic geography and AI. Section 2 describes how we collected and enriched the data and classified papers from `arXiv` computer science corpus into the DL category. Section 3 presents our findings in three steps. First, we consider whether DL displays three defining features of a GPT (*rapid growth, rapid diffusion into new fields, and impact in new fields*). Second, we study the geographical aspects of its diffusion. Third, we model the link between regional specialization in DL research and activity

[1]The code we have used in our analysis is available for review in `https://github.com/nestauk/arxiv_ai`.

in related knowledge and industrial bases. Section 4 discusses the findings and its limitations, and outlines issues for further research.

1.1 General Purpose Technologies as engines of growth

GPTs are technologies or clusters of related technologies *'characterized by the potential for pervasive use in a wide range of sectors and by their technological dynamism'* [1, 13]. They enable productivity improvements in multiple industries by automating or greatly improving the efficiency of key production tasks such as the use of energy for work, or the transfer and processing of information. The steam engine replaced human, animal and natural motor power in mining, textiles and transport [3]. Electricity cheaply illuminated homes and workplaces, and the combustion engine detached energy from a fixed grid, making production and transport more flexible [14]. Micro-electronics transformed the speed and scale of computation across the economy.

If we imagine the technology system as a network of ideas being constantly recombined, then we will find GPTs sitting near its center [15]. GPTs induce cascades of complementary innovations in the sectors that deploy them, some of which may also be widely applicable. For example, Information and Communication Technologies (ICTs) based on cheap microchips gave birth to the video-games industry, which subsequently spurred the development of Graphical Processing Units (GPUs) now used for parallel processing of information in other sectors. This exploration of new GPT opportunities requires trial-and-error and can take time. For example, US factories did not start to realize the benefits of electric power until they reorganized their layout to harness the flexibility of small electric motors, decades after the introduction of electricity [14].

The networked nature of GPTs creates the risk of *coordination failures* in its deployment: rapid change makes their evolution hard to predict, and might encourage a 'wait-and-see' strategy among potential adopters and providers of complementary skills, infrastructures and standards. This can hinder the exploration of the new opportunities the GPT offers, and delay or halt follow-on innovations [16].

1.2 Towards a geography of GPTs

When they arrive, GPTs transform the economic conditions and production processes of many industries. Consider for example the changes brought about by the advent of steam to textiles and transport in Britain, or more recently, the impact of the Internet in media or retail. Since industries tend to cluster in specific locations to access dense talent pools, reduce transaction costs and learn from each other, the impact of GPTs will also be unequally distributed in space [17]. If a GPT is 'competence-destroying' for an industry (that is, if it eliminates previous sources of comparative advantage like the Internet did with control over physical distribution channels in the music industry), then those locations where the industry concentrates will experience a negative shock. At the same time, a GPT can create windows of opportunity to enter a sector, like the Internet did with new media clusters.

Economic geographers have studied similar discontinuities through the lens of the product life-cycle. The idea is that the technologies used by an industry follow a trajectory with distinct phases, and that each of these phases has a different geography. Early in the life-cycle, when a new market or technological opportunity is revealed, there is a phase of experimentation when entrepreneurs explore different designs to harness this opportunity [18, 19][2]. In this phase of the product life-cycle, there is uncertainty about the technologies and capabilities required to succeed in the market, lowering barriers to entry for new entrepreneurs and regions [21]. Eventually, this experimentation yields a standard or dominant design and the industry moves from product to process innovation. Economies of scale become more important, leading to industrial and geographical consolidation.[3] We would expect something similar to happen when a GPT arrives,

[2](the beginning of the automobile industry is a paradigmatic example of this phase, with inventors and entrepreneurs exploring in parallel various energy sources for the automobile, from the combustion engine to electrical and steam-powered motors, [20])

[3]At the same time, there might be some dislocation of activity as standardized parts of the production process are outsourced or off-shored to other locations with cheaper costs.

with an initial phase of geographical volatility when new entrants come into the market, followed by a shake-out and increasing concentration once a dominant design is established.

What factors determine whether a region is able to enter and successfully compete in the development of the GPT in the first place? A growing body of literature on *Economic Complexity* and *Economic Relatedness* suggests that a region's ability to enter a new market or technology depends on the presence of related capabilities that can be re-purposed or recombined to explore new opportunities. This is referred to as the *Principle of Relatedness* [11, 12, 20, 22]. Building on this idea, GPTs that can be applied in multiple industries could benefit from the co-location of R&D sectors that develop the technology and industrial sectors where it can be applied. Proximity between developers, adopters and suppliers of skills and infrastructure facilitates communication and reduces the risk of coordination failures, improving the prospects for GPT deployment and helping the location gain a comparative advantage in the technology [23].

1.3 Empirical setting: Artificial Intelligence and Deep Learning

Having discussed the concept of GPTs, we now turn our attention to the empirical setting for our analysis: Artificial Intelligence, and more specifically the Deep Learning techniques underpinning it.

Artificial Intelligence (AI) systems have been defined as *'self-training structures of Machine Learning predictors that automate and accelerate human tasks'* [24]. In turn Machine Learning (ML) is *'the field that thinks about how to automatically build robust predictions from complex data'* [24]. ML emerged in the 1970s in response to the failure of rule-based approaches where human experts hard-coded knowledge in Artificial Intelligence systems [25]. ML's approach is to instead develop algorithms that can recognize patterns in labeled data with less need for human intervention, and use the resulting models to make predictions about new observations. Economic analyses of AI focus on its ability to reduce the costs of prediction, an important task in many industries [26].

Deep Learning (DL) is a new ML technique that processes large and complex datasets it through networks of synthetic neurons where subsequent layers learn increasingly abstract representations of the data that eventually become an input into prediction [27]. Although the neural network literature goes back to the 1950s, this approach only became feasible in recent years thanks to the availability of large, labeled datasets from the web, and powerful GPUs. Since the early 2010s, Deep Learning has been proven to be *'unreasonably effective'* in many applications, from image and video recognition to translation and gaming, fueling a surge of interest and investment in AI [28].

Ultimately, AI researchers strive for generality: developing algorithms that can transfer their predictive prowess across domains, and respond effectively to new situations. Sustained progress towards that goal has led a growing number of economists to declare DL-driven AI a new GPT that will revolutionize the economy [2]. DL also represents an *'invention in the methods of invention'* that could transform how new ideas are discovered, improving productivity of R&D in fields such as drug discovery, genomics or material sciences [9, 29]. Publication, patenting and venture capital trends support this view, with rapid growth in DL activity and diffusion into other disciplines and industries [9].

The GPT nature of AI would also explain stagnant productivity growth despite rapid technological progress: businesses still need to reorganize their operations [2], the education system needs to address skills shortages, and suitable digital and regulatory infrastructures have to be developed to create value from AI-driven growth.

What about the geography if AI? A recent review of its international trade aspects argues that the localized nature of AI knowledge spillovers (the fact that organizations need to be based in the locations where investments on R&D take place to benefit from them) could justify national policies to support its development [30]. Governments across the world appear to share this view, and many have announced national strategies to compete in the 'AI global race'. There is a growing belief that China, with its large STEM workforce, powerful Internet platforms and vast amounts of data is 'winning' this race [31]. Meanwhile, European researchers and policymakers fear that the EU falling behind for lack of talent and leading AI-driven businesses [32]. These perceptions imply

that AI GPT is disrupting the geography of digital production and innovation. As AI researcher Andrew Ng points out *'Since AI changes the foundation of many technology systems - everything ranging from web search to autonomous driving to customer service chatbots - it also gives many countries the opportunity to 'leapfrog' the incumbents in some application areas'* [33]. In the rest of this paper, we monitor these geographical changes and study their drivers using a novel preprints dataset and state-of-the-art Natural Language Processing (NLP) methods.

2 Data collection and classification

Our analysis relies on several data sources and preprocessing activities:

1. We combine data from `arXiv`, `GRID` (Global Research Identifier) and `MAG` (Microsoft Academic Graph) to create a geocoded dataset of research activity in computer science disciplines where we identify DL papers with `CorEx`, a topic modeling algorithm. We also measure the relatedness between computer science subjects based on their co-occurrence in `arXiv` papers.
2. We use `CrunchBase`, a business directory, to map industrial activities that might be relevant for the development of DL clusters. We measure relatedness between those industries and DL using a machine learning model that predicts industrial sectors with company descriptions.

We go through these two streams of data collection and classification in turn.

2.1 Identifying and mapping DL papers in `arXiv` data

We generate the DL dataset for our analysis by matching three non-proprietary open data sources; `arXiv`, Microsoft Academic Graph (`MAG`), and the Global Research Identifier Database (`GRID`). The data sources are matched in the following order, according to the procedure described in Sections 3.1.1- 2.1.3:

$$\{\texttt{arXiv} \xrightarrow{\text{matched to}} \texttt{MAG}\} \xrightarrow{\text{matched to}} \texttt{GRID}$$

By following this pipeline of data collection, we create a dataset with the features described in Table 1 for further processing as described in Section 2.1.4.

Feature	**Data source**	**Comments**
Article title	`arXiv`	Assured to be consistent with `MAG` title after matching procedure.
Article abstract text	`arXiv`	To be used for topic modeling (Section 2.1.4).
Subject classification	`arXiv`	Assigned by the author.
Is article published?	`MAG`	Always true, as implicitly assured by match to `MAG`.
Publication date	`MAG`	Publication date in MAG, rather than `arXiv` submission date.
Citation count	`MAG`	Used for cross-check by selecting 'high quality' publications (Section 3).
Institute affiliation (all authors)	`MAG`	This replaces the potentially incomplete set of authors from `arXiv`.
Institute location	`GRID`	

TABLE 1: *Features extracted in the data collection procedure.*

2.1.1 arXiv

arXiv is a 'real-time' open archive of academic preprints widely used by researchers in quantitative, physical and computational science fields. Data from each of over 1.3 million papers can be accessed programmatically via the arXiv API. As arXiv papers are self-registered, we ensure that papers are not simply 'junk' articles by requiring that all papers are matched to a journal publication or conference proceeding, as presented in Section 2.1.2. We also have anecdotal evidence that the archive contains many high quality papers, since a short study of conference proceeding from the prestigious AI Conference on Neural Information Processing Systems in 2017 reveals that over 55% of these were published on arXiv.

Is arXiv a suitable data source for the analysis of industrial R&D? We believe that this is the case. The AI research community has a strong culture of openness in its publication of research findings, software and benchmark datasets, which are perceived as a way to attract scientific talent [34]. Some of the most active DL institutions in our corpus include corporations such as Google, Microsoft, IBM, Baidu or Huawei.

From the initial set of over 1.3 million papers, approximately 134,000 have been selected for analysis as they fall under the broad category of 'Computer Science' (cs) or the specific category of 'Statistics - Machine Learning' (stat.ML).

2.1.2 Microsoft Academic Graph (MAG)

Microsoft Academic Graph (MAG) is an open API offering access to 140 million academic papers and documents compiled by Microsoft and available as part of its 'Cognitive Services'. For the purpose of this paper, MAG helps to ensure that article retrieved from arXiv have been published in a journal or conference proceeding, as well as providing citation counts, publication date and author affiliations. The matching of the arXiv dataset described in Section 3.1.1 is performed in two steps.

We begin by matching the publication title from arXiv to the MAG database. The database can be queried by paper title, although fuzzy-matching[4] or near-matches are not possible with this service. Furthermore, since paper titles in MAG have been preprocessed, one is required to apply a similar preprocessing prior to querying the MAG database. There is no public formula for achieving this, so we explicitly describe the following steps to emulate the MAG preprocessing:

1. Identify any 'foreign' characters (for example, Greek or accented letters) as non-symbolic;
2. Replace all symbolic characters with spaces; and
3. Ensure no more than one space separates characters.

This procedure leads to a match rate of 90%, for the set of arXiv articles used in this paper. We speculate that papers could be missing for several reasons: the titles on arXiv could significantly different from those on MAG; the latter procedure may be insufficient for some titles; the arXiv paper may not be published in a journal; and MAG may not otherwise contain the publication. It may be possible to recuperate some of these papers, however this is currently not a limiting factor in our analysis.

2.1.3 Global Research Identifier Database (GRID)

We use the Global Research Identifier Database (GRID) to enrich the dataset with geographical information, specifically a latitude and longitude coordinate for each affiliation that we can then geocode into countries and regions.[5]. The GRID data is particularly useful since it provides institute names and aliases (for example, the institute name in foreign languages). Each institute name from MAG is matched to the comprehensive list from GRID as follows:

[4] 'Fuzzy-matching' refers to the process of finding a likely match for a set of text (such as a word or sentence) amongst a choice of texts. A naive example would be comparing the ratio of the number of characters between texts, and identifying the texts with the highest ratio as a match.

[5] We do this with a point-in-polygon approach using boundary (shapefile) data from the Natural Earth public map dataset.

1. If there is an exact match amongst the institute names or aliases, then extract the coordinates of this match. Assign a 'score' of 1 to this match (see step 3. for the definition of 'score').
2. Otherwise, check whether a match has previously been found. If so, extract the coordinates and score of this previous match.
3. Otherwise, find the `GRID` institute name with the highest matching score, by convoluting the scores from various fuzzy-matching algorithms in the following manner:

$$\frac{1}{\sqrt{N}}\sqrt{\sum_{n=0}^{N} F_n(m_{\texttt{MAG}}, M_{\texttt{GRID}})^2} \tag{1}$$

where N is the number of fuzzy-matching algorithms to use, F_n returns a fuzzy-matching score (in the range $0 \rightarrow 1$) from the n^{th} algorithm, $m_{\texttt{MAG}}$ is the name from `MAG` to be matched and $M_{\texttt{GRID}}$ is the comprehensive list of institutes in the `GRID` data.

The form of Equation 1 ensures that effect of a single poor fuzzy-matching score is to vastly reduce the preference for a given match. Therefore, good matches are defined according to Equation 1 as having multiple good fuzzy-matching scores, as measured according to different algorithms. We use a prepackaged set of fuzzy-matching algorithms implementing the Levenshtein Distance metric [35], and specifically, two algorithms applying a token-sort-ratio and a partial-ratio respectively.

After this stage of data matching, we are left with approximately 240,000 unique institute-publication matches with at least one computer science subject in their `arXiv` categories.

2.1.4 Topic modeling

We analyze the abstracts in our corpus using Natural Language Processing to identify papers related to DL. This involves tokenizing the text of the abstracts and removing common stop-words, very rare words and punctuation. We lemmatize the tokens based on their part-of-speech tag, and create bi-grams and tri-grams. Documents with less than twenty tokens are removed from the sample. After these steps, there are over 168,000 features (unique 'words') in the dataset.

There are different approaches to identify DL papers in this preprocessed corpus. Previous work has used a keyword-search approach based on a predefined vocabulary of terms [9]. Here, we follow an alternative topic modeling strategy which identifies clusters of words in the data without an initial vocabulary, and provides a score for each topic in a document, simplifying the labeling process.

More specifically, we use the Correlation Explanation (`CorEx` [36]) algorithm, which takes an information-theoretic approach to generate n combinations of features in the data which maximally describe correlations in the dataset. Using a one-hot bag-of-words representation, we optimally find $n = 28$ topics by tuning n with respect to the 'total correlation' variable, as advised by the `CorEx` authors. The generated topics contain words which are sorted in terms of their contribution of each feature to total correlation. We assign a score S^j for each topic j (containing N^j words w_i with topic weights T_i^j) to each document W such that:

$$S^j = \sum_{i=0}^{N^j} T_i^j \delta(w_i, W) \tag{2}$$

where:

$$\delta(w_i, W) = \begin{cases} 1 & \text{if } w_i \in W \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Topics are then assigned to each document only if the following condition is satisfied:

$$S^j \geq \gamma \, T^j_{\text{max}} \tag{4}$$

where γ is a threshold parameter that we assign below, and $T^j_{\max}$ is the maximum topic weight. The form of the above asserts that documents must contain a sufficient number of components of topics to be assigned to the topic. Clearly, a larger choice of γ leads to a lower frequency of documents assigned to the topic whilst improving the overall recall.

After inspecting the model outputs, we identify two topics related to DL, containing keywords such as `neural_network`, `deep_learning`, or `convolutional_neural_networks`. We label as 'Deep Learning' those papers where either of these topics is present with a γ above 0.5, giving us a set of 15,062 DL papers (11% of the total unique papers).[6].

2.1.5 Research relatedness

In Section 3.5, we study the link between DL specialization in a region and the presence of related research and industrial activities. We proxy research relevance using the relatedness between research subjects based on their co-occurrence in `arXiv` papers.[7] To measure this relatedness, we calculate the cosine similarity between vectors representing the subjects that appear in different papers in the corpus. Sub-section 3.1 presents the results.

2.2 Building the industrial dataset

2.2.1 `CrunchBase`

We use `CrunchBase`, a commercial directory of technology companies, to measure industrial activity in a region. The version of `CrunchBase` we use contains information about 257,000 organizations, including a short description of their activities, the sectors where they operate, the year when they were founded and their geographical coordinates.[8]

Recent analyses of technology clusters in `CrunchBase` suggest that it correlates well with other measures of regional technological activity, and it is increasingly being used in economics and management research [37, 38]. `CrunchBase` presents two important advantages for our analysis: first, it has global coverage (like our `arXiv` corpus) and individual organization locations, so it is easy to merge with our `arXiv` data at a suitable geographical level. Second, it contains text descriptions of company activities and labels for the sectors where they operate, which we can use to generate measures of similarity between these sectors and DL papers in the `arXiv` data using the strategy we describe below.

2.2.2 Research-industry relatedness

We estimate the relatedness between industrial activities in `CrunchBase` and research in `arXiv` by training a supervised machine learning model that predicts the sector where a company in `CrunchBase` operates based on its description[9].

This model is then used for out-of-sample prediction of the `CrunchBase` categories of `arXiv` papers, based on the text in their abstract. Specifically, we assign categories where the prediction probability is at least 0.99.[10] We then calculate the share of papers by `arXiv` subject predicted to be in a `CrunchBase` category to measure their relatedness. Subsection 3.1 presents the results.

[6] We also create a more restrictive DL category containing only those papers either topic is present with a γ above 0.5, resulting in a total of 1,604 papers. A visual inspection of a random sample of papers in both groups suggests that their outputs are similarly relevant so we opt to focus on the larger set. This is further motivated by our interest in understanding the diffusion of DL methods in various computer subjects.

[7] Researchers who submit their papers to `arXiv` label them with a set of relevant research categories. We focus our analysis in Computer Science (`cs`) subjects as well as those in the `stat.ML` subject.

[8] As before, we geocode `CrunchBase` companies using a point-in-polygon approach with boundaries from Natural Earth.

[9] We focus on those observations with the longest and more informative descriptions, comprising around 115,000 companies. We perform a grid-search to select the best performing model, a logistic regression classifier with L1 regularization.

[10] By setting a high threshold for classification of `arXiv` papers into `CrunchBase` categories we seek to remove noise in the transference of the model across corpora with potential differences in their languages.

3 Analysis

3.1 Descriptives

3.1.1 arXiv

Table 2 presents some descriptive statistics for papers classified as DL and the rest of the corpus. DL papers have, on average, been published more recently, they tend to contain fewer arXiv subjects, and involve collaborations with a somewhat higher number of institutions. They also tend to receive more citations , specially after we control for the number of years since publication. This suggests that DL is a relatively recent topic, and that DL papers are, on average, more influential than the rest.

dl_cat	dl	non_dl
total	15602	115587
year_average	2015.937	2013.572
field_average	2.798	2.930
institute_average	1.754	1.705
citation_average	25.627	18.003
citation_p_year_average	6.182	2.505

TABLE 2: *Descriptive statistics for DL / non-DL papers in arXiv dataset*

Figures 1, 2 and 3 present the distribution of DL and non-DL activity over arXiv computer science subjects, countries and regions for the top categories in each variable.

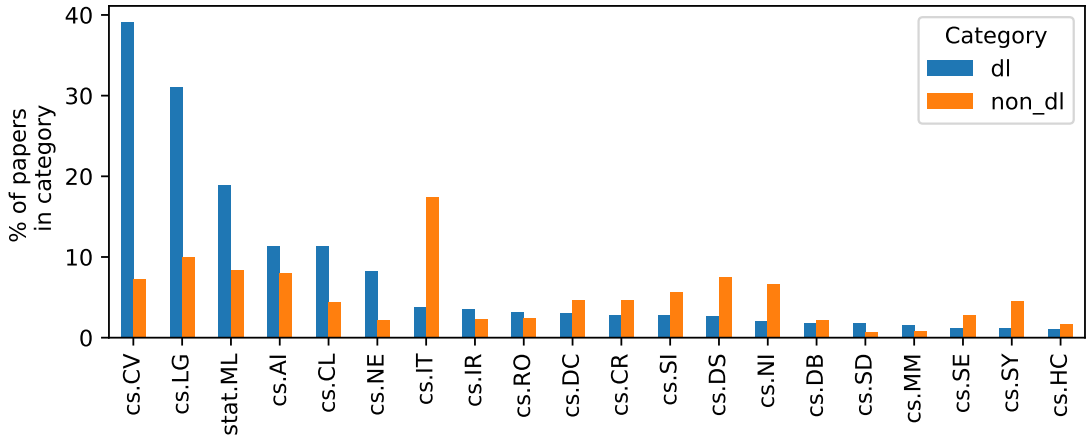

FIGURE 1: *Distribution of DL/non DL papers by arXiv category*

Some observations:

1. DL papers are highly concentrated in a small number of arXiv subjects: Computer Vision (cs.CV), Computer Learning (cs.LG), Machine Learning (stat.ML), Artificial Intelligence (cs.AI) and Neural Networks (cs.NE). The set of DL-intensive subjects includes some that rely on unstructured datasets where DL has achieved important breakthroughs, and in fields that specialize in the development of ML and AI methods.

2. The US has the biggest share of DL and non-DL papers, with around a third of all publications in both categories. China is overrepresented in DL: its share of DL papers is more than double

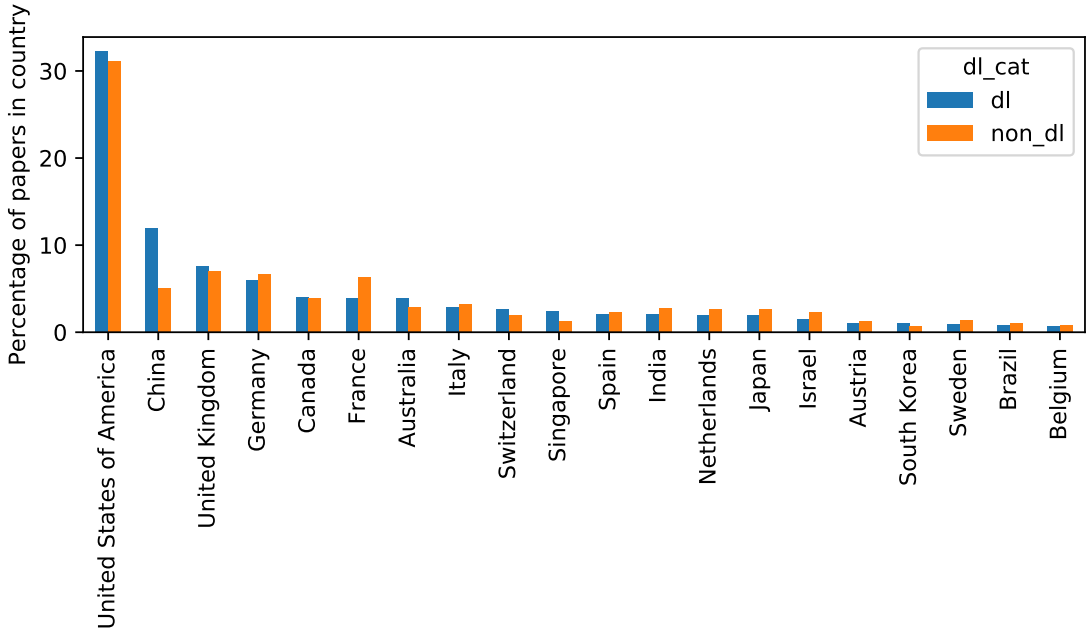

FIGURE 2: *Distribution of DL/non DL papers by country (top 20 countries)*

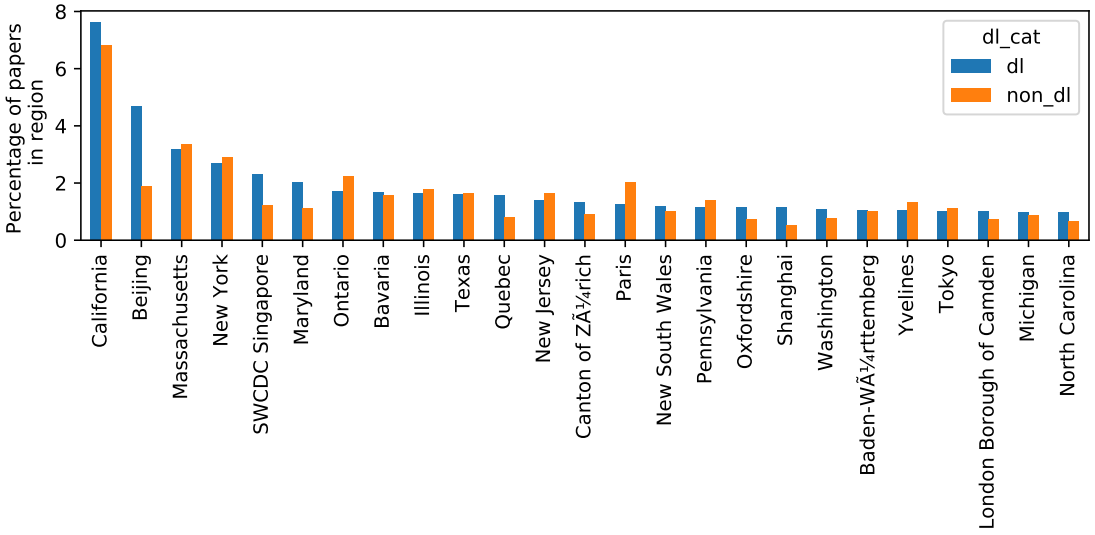

FIGURE 3: *Distribution of DL/non DL papers by region (top 35 regions)*

its share of non-DL papers. By contrast, France is underrepresented in DL.

3. North American regions dominate the global rankings of DL activity. California, Massachusetts, New York, Maryland, Illinois and Texas rank highly by volume of DL activity. Ontario and Quebec in Canada also have high levels of activity, consistent with Canada's strong research base on AI. Beijing, the South West Development Corporation in Singapore, Maryland and Quebec are over-represented in DL, with substantially higher shares of activity in DL than in the rest of the corpus). Notably, only one EU region (Bavaria) appears in the top ten of global DL research in `arXiv`.

Figure 4 displays a heatmap of the proximities between different `arXiv` subjects (as well as the DL category) based on their co-occurrence on papers, sorted by their proximity to the DL category.

Consistent with Figure 1, DL papers are closer to computer science subjects involving unstructured

data and subjects that research ML, AI and neural networks. These subjects also tend to to co-occur with each other, forming a 'cluster' of data analytics research in `arXiv`. Our analysis also reveals intuitive connections between other `arXiv` subjects such as Computers and Society (`cs.CY`) and Human Computer Interaction (`cs.HC`) or between Logic (`cs.LO`) and Programming Languages (`cs.PL`), supporting the idea that these proximities are a meaningful measure of relatedness between computer science subjects in `arXiv`.

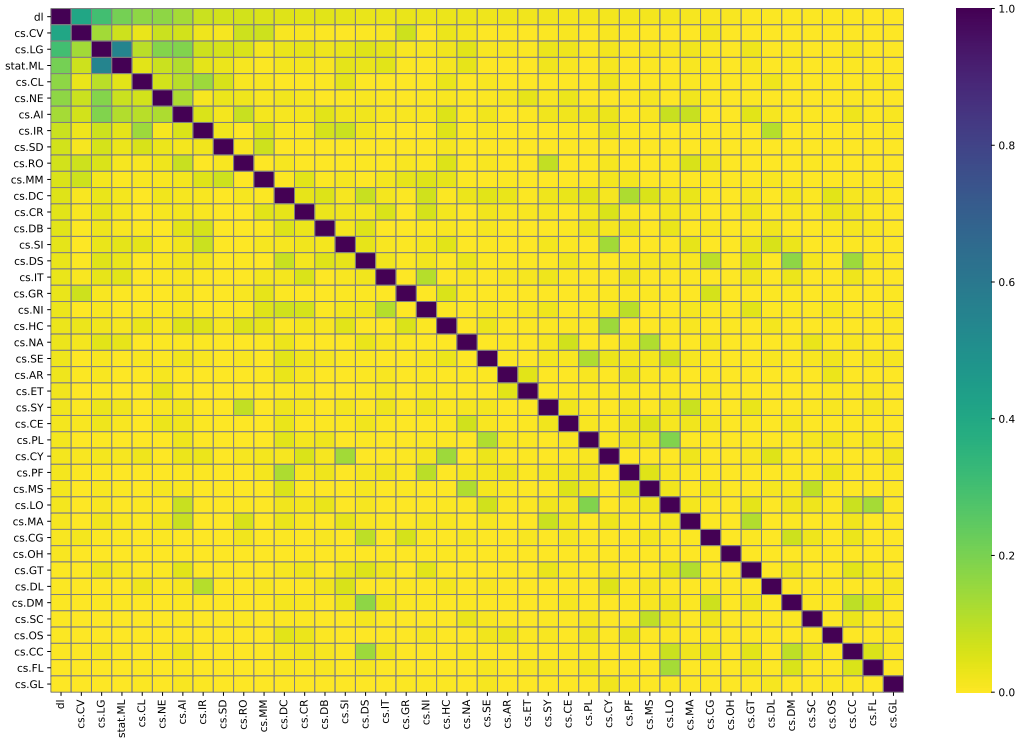

FIGURE 4: *Proximities between `arXiv` subjects based on co-occurrence in papers*

3.1.2 `CrunchBase` data

Figure 5 presents the regional distribution of activity in `CrunchBase`. California is again the top region by number of organizations. Technology company activity in `CrunchBase` is more concentrated than research in `arXiv` (California accounted for 15% of all activity in `CrunchBase`, while it only captured 7% of the activity in `arXiv`). US States and Indian regions have a stronger presence here than they did in `arXiv`. Chinese provinces are, by contrast, less visible.

Figure 6 compares levels of activity in `arXiv` and `CrunchBase`. Although there is a strong correlation between both datasets (ρ=0.67), we note some divergences. For example, there are several UK counties around London with a strong presence in `CrunchBase` but low activity in `arXiv`. Conversely, some Japanese prefectures display high levels of `arXiv` activity but few organizations in `CrunchBase`[11].

We end our descriptive analysis by considering the proximity between `arXiv` categories (including DL) and `CrunchBase` sectors based on the machine learning analysis outlined in 2.2.1. The heatmap in 7 presents the share of all papers in an `arXiv` subject (and DL) that were labeled in a `CrunchBase` category. It shows that DL papers were classified more often in Data Analytics, Artificial Intelligences and Software `CrunchBase` sectors. We also detect intuitive relations between other `arXiv` categories and `CrunchBase` sectors: for example, Robotics (`cs.RO`) is related to Science and Engineering, Sound (`cs.SO`) is related to Music and Audio, and Cryptography (`cs.CR`) is related to Privacy and Security. It is however worth noting that some of the similarities we

[11] These results underscore the importance of triangulating our results against other data sources in future research.

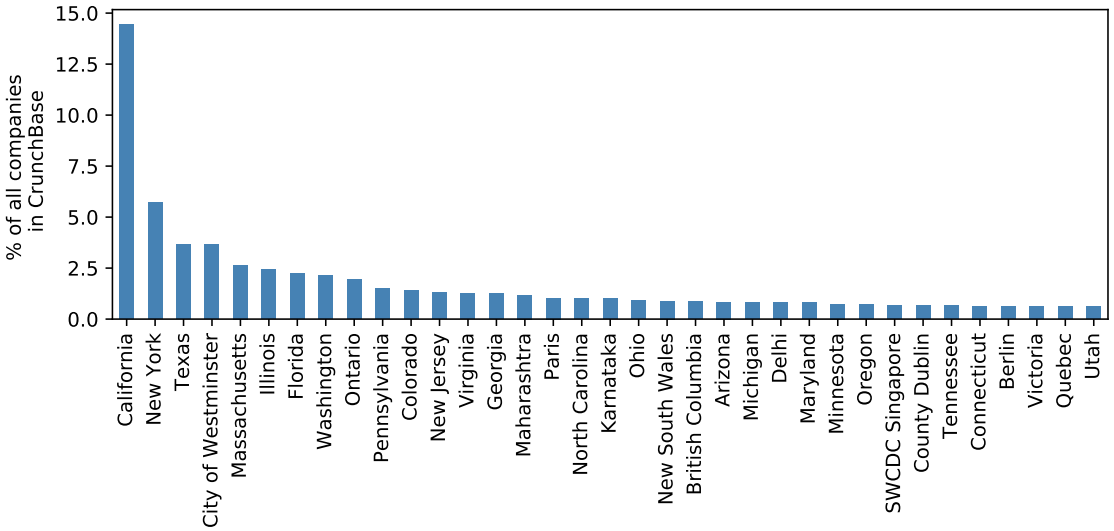

FIGURE 5: *Share of* `CrunchBase` *activity by region*

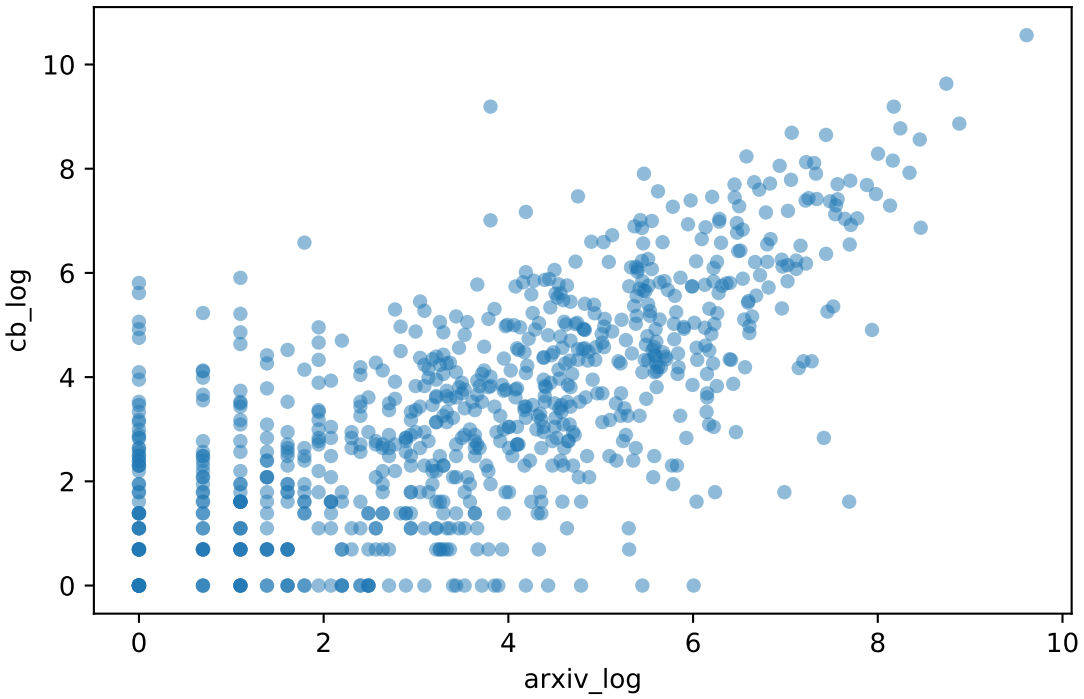

FIGURE 6: *Relationship between* `arXiv` *and* `CrunchBase` *activity (logged)*

identify could be linguistic rather than semantic (for example, our model detects a strong similarity between Game Theory - `cs.GT` and Gaming, which could be partly explained by their use of similar language rather than a shared knowledge base).

3.2 GPT aspects of DL research in `arXiv`

We now move into our first question: Is DL a GPT? In answering this, we seek to ensure that our interpretation of further results is valid, and contribute to the literature on the GPT nature of AI using a new dataset and classification method [9]. Previous analyses of patent data in [13] have looked for GPTs using patent class growth and citations, while more recently, [9] measure growth in DL publishing and patenting with a keyword-based approach. They also consider levels of publishing in application fields outside of Computer Science to measure the generality of DL[12]. Our analysis builds on all this work.

Inspired by the original definition of a GPT, we have devised the following three GPT tests for DL:

3.2.1 Rapid growth

The first component of the definition of a GPT is 'technological dynamism', which we measure, like [9], by looking at growth in activity. If DL is a GPT with broad applicability, we should see an increase in the number of DL papers in `arXiv` as more researchers explore its potential.

Figure 8 presents the evolution of DL and non-DL publishing in `arXiv`. It shows that `arXiv` is becoming an increasingly popular venue for computer science research, and that DL is gaining relative importance in it. The share of DL papers in the total has grown fivefold, from 3% before 2012, to 15% afterwards [13].

3.2.2 Generality

The second GPT test for a technology is *rapid diffusion in new fields*: is DL being adopted in multiple domains or restricted to a small number of areas? To assess this, we measure the number of DL papers in different `arXiv` subjects [14].

Figure 9 presents the results. The top panel displays yearly changes in the shares of DL by `arXiv` subject (based on 3 year moving averages), and the bottom panel compares shares of DL activity in a category before and after 2012, focusing on the top 35 computer science subjects in `arXiv` by total levels of activity.

DL also fulfills the second GPT test, with a visible upward trend in the relative importance of DL in many computer science subjects, specially since 2012, the year of publication for [39], a landmark paper in the use of DL in computer vision. Further, the bottom panel of 9, shows that virtually all computer science subjects in our corpus have experienced an increase in the relative importance of DL research since 2012. As before, this is particularly visible in subjects that use unstructured data (e.g. Computer Vision) or specialize in the development of AI and ML methods (Neural Networks, Computer Learning etc.) [27].

3.3 Impact in other fields

The third GPT test is *impact in new fields*: does DL generate follow-on innovations in the fields that adopt it? Following convention, we use citations as a proxy for that impact.

[12] It is interesting to note that they classify computer vision papers and patents outside of DL. This contrasts with our finding that Computer Vision is one of the main application areas for DL, underscoring the value of unsupervised approaches for the analysis of fast moving technology fields.

[13] The results are similar if we focus on the most highly cited papers every year.

[14] As mentioned, most papers are labeled with multiple `arXiv` subjects. We allocate a paper to a subject if it appears in it at least once.

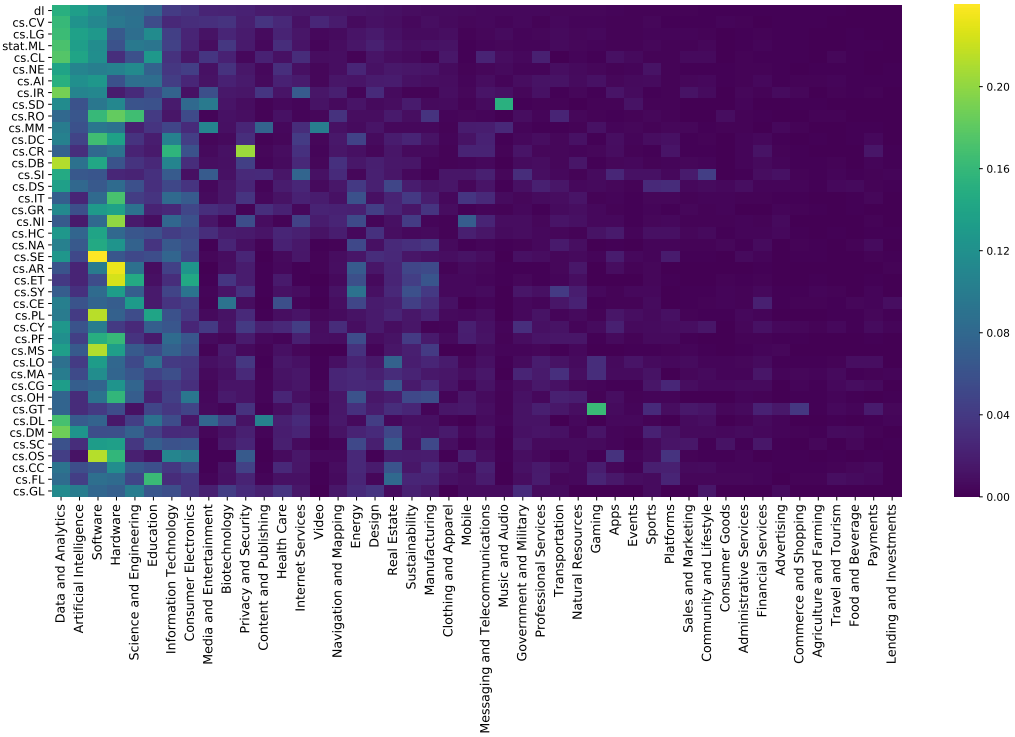

FIGURE 7: *Proximity between* ***arXiv*** *disciplines and* ***CrunchBase*** *sectors*

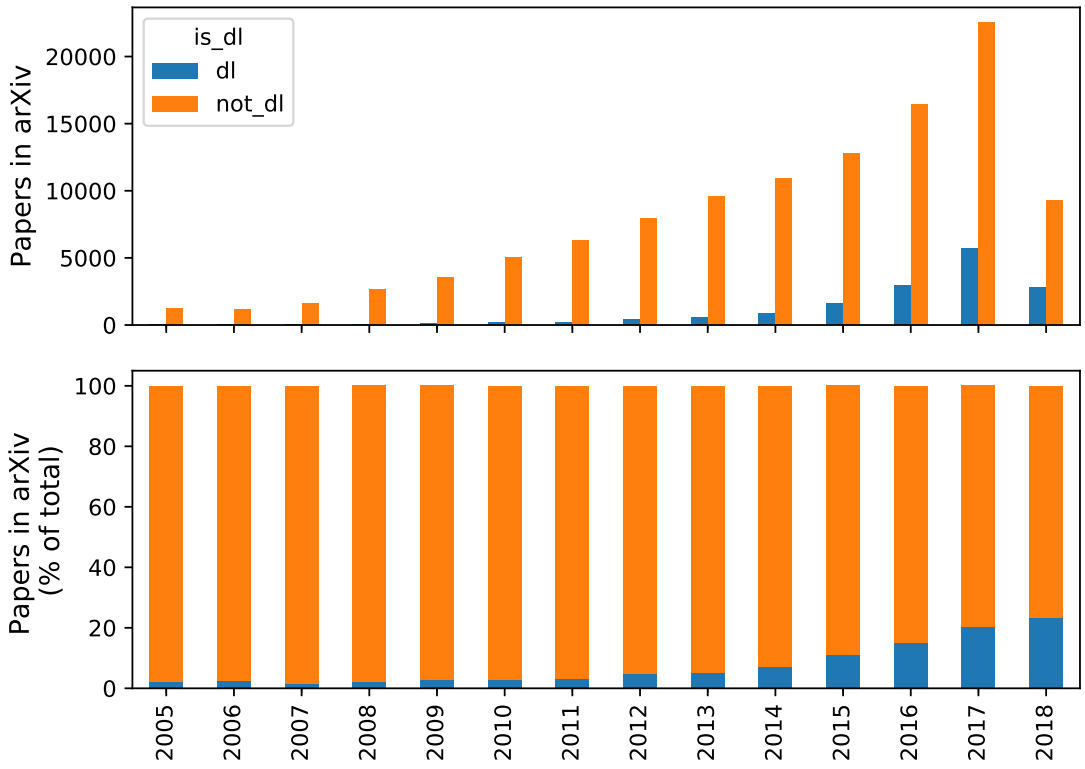

FIGURE 8: *Publication activity in* ***arXiv*** *(2008-2012)*

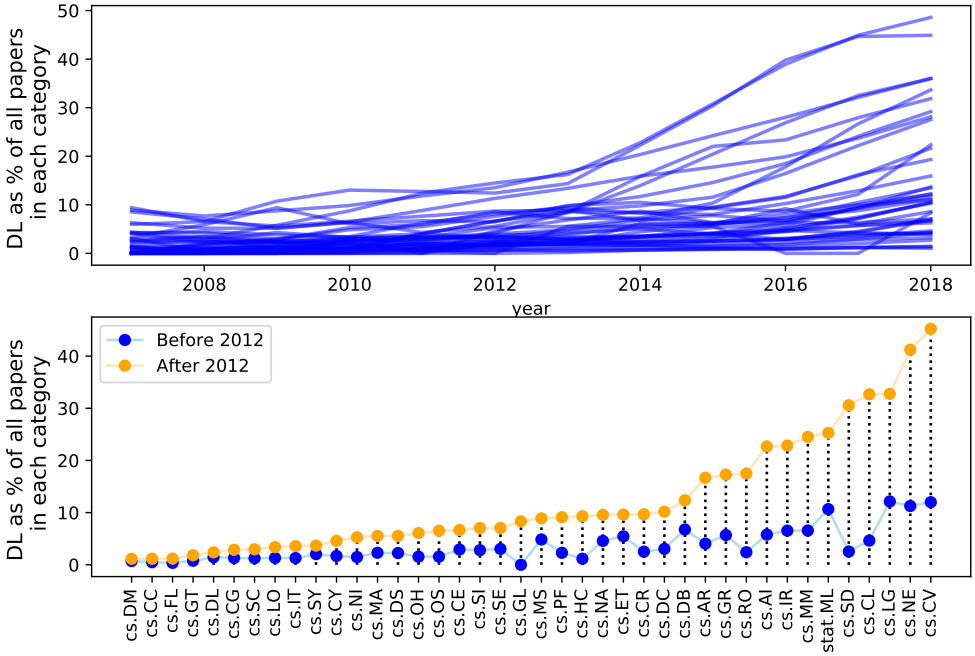

FIGURE 9: *DL as a share of activity in different* ***arXiv*** *subjects. Top panel shows yearly trends for all subjects in the* ***arXiv*** *data. Bottom panel compares shares of DL activity in a subject before and after 2012.*

Figure 10 compares the shares of DL papers in a `arXiv` subjects with their share of *highly cited papers* in that same subject over different periods [15]. In all cases, most `arXiv` subjects are above the diagonal (this is, DL papers are overrepresented among the highly cited ones in the subject). This pattern becomes more apparent over time, supporting the idea that DL is becoming more influential in the fields where it is applied.

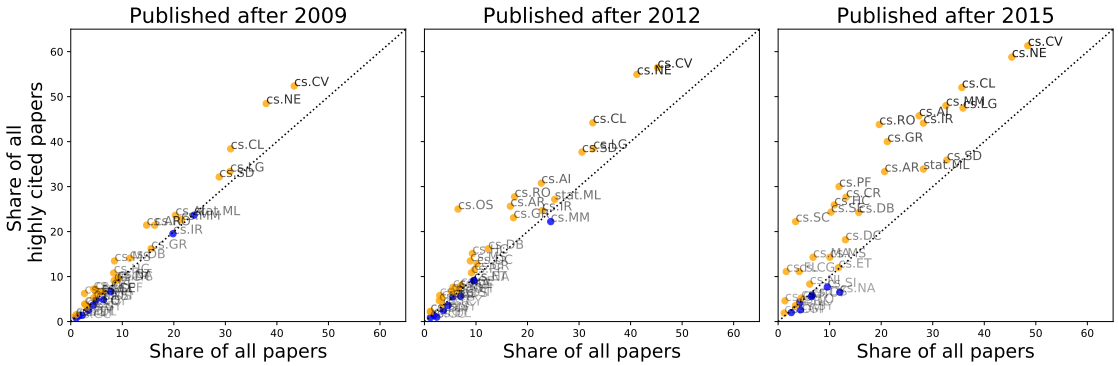

FIGURE 10: *DL papers as a share of all papers in an* ***arXiv*** *category, and as a share of all highly cited papers for papers published after 2009 (left panel), 2012 (center panel) and 2015 (right panel).*

Together, these results support the idea that DL is a GPT: its levels of activity are growing rapidly, it is spreading into more fields, and it is generating an impact (or at least attracting attention, in terms of the number of citations it receives) in the fields where it is applied.

[15] Highly cited papers are those in the top citation quartile for each year.

3.4 Evolution in the Geography of DL research

We now turn to the analysis of the geography of DL research, considering whether its evolution follows the cycle of volatility and consolidation we would expect based on the literatures reviewed in Section 1.2. To do this, we analyze changes in national and regional DL specialization using relative comparative advantage (RCA) indices. We define the RCA_{dl} of a location (country or region) i as:

$$\text{RCA}_{dl,i} = \frac{(\frac{A_{dl,i}}{A_{c,i}})}{(\frac{A_{dl,n}}{A_{c,n}})} \tag{5}$$

Where $A_{dl,i}$ and $A_{c,i}$ are the research activity of the location in DL and in all `arXiv` categories, and $A_{dl,n}$ and $A_{c,n}$ are the totals of DL activity and activity in all `arXiv` categories in all locations. A $\text{RCA}_{dl,i}$ above 1 implies that the country is relatively specialized in DL, while the opposite is true if the $\text{RCA}_{dl,i}$ is below 1. RCAs allow us to measure changes in DL research while controlling for rapid growth in computer science activity, and for differences in size between locations. Since RCAs tend to lose robustness in observations with low levels of activity, we focus our analysis on the larger countries and regions. We also remove low quality papers from the data by focusing on those above the median of citations for the year when they were published.

Figure 11 presents DL specialization by country after 2012 (map in the right panel) and changes in DL specialization since 2012 for the most active countries. It shows that China has the strongest comparative advantage in DL R&D. Interestingly, this has not changed significantly since 2012, suggesting that the development of advanced AI capabilities in China predate the recent explosion of interest in DL. We also see rapid growth in the specialization of other Asian countries such as Singapore and Korea. By contrast, all European countries in the chart with the exception of the United Kingdom and France have become less competitive in DL research (and France has, in any case, low levels of specialization in the DL). Canada and the US have also increased their DL specialization since 2012.

These changes are consistent with the idea of volatility in the early stages of GPT development, with some countries climbing up in the research rankings rapidly while others fall behind. It is also interesting to note, qualitatively, that the trends we observe echo popular narratives about the current state of the 'AI race', with China in the ascendant while European countries fall behind in relative terms. After an initial slow response to the emergence of DL, the US is catching up [9].

Figure 12 presents similar figures but this time focusing on regions. The map shows high levels of activity in a small number of regions in the East and West coast of the US, Canada, China and East Asia, Central Europe, France, Britain and Adelaide in Australia (which hosts the Australian Institute for Machine Learning Research). The right-hand panel shows US states such as Maryland, California and New York becoming more specialized in DL since 2012. Perhaps the most notable change is in Oxfordshire in the UK, which has multiplied its RCA_{dl} more than seven-fold since 2012. Interestingly, we see that most of the largest regions in DL activity have also gained specialization in DL, suggesting potential advantages to scale in developing a DL research cluster. One potential explanation we explore in 3.5 is that these larger regions have sufficient scale to host the combination of research and industrial capabilities required to develop the DL GPT.

We conclude by considering changes in the dispersion and concentration of DL activity since 2009. Does the geography of DL research follow the cycle of volatility and consolidation we expect from the product life-cycle literature?

Figure 13 shows the recent evolution in volatility and concentration of DL activity in the largest nations and regions, focusing again on highly cited papers.

The patterns in the violin-plots in the top panel are consistent with the idea that DL experienced an initial phase of volatility (high dispersion and a flatter distribution in RCAs) followed by growing stability (less dispersion and a normal distribution with fewer locations displaying high RCAs). Also in line with what we expected, the bottom panels show a sudden decline in the shares of activity accounted for by the top countries / regions around 2012, followed by an increase in

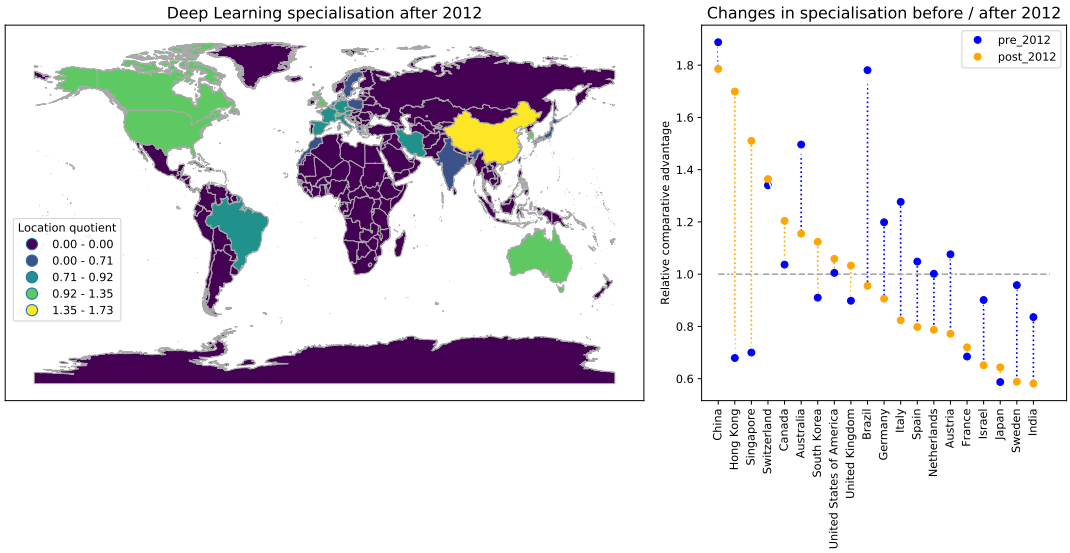

FIGURE 11: *The map in the left panel shows RCA_dl by nation for papers published after 2012, focusing on papers above the median of citations in their publication year, and countries in the top 90 percentile for total level of activity. The figure in the right panel compares changes in RCA_{dl} between the period before 2012 and afterwards, focusing on the top 20 countries by total level of DL activity.*

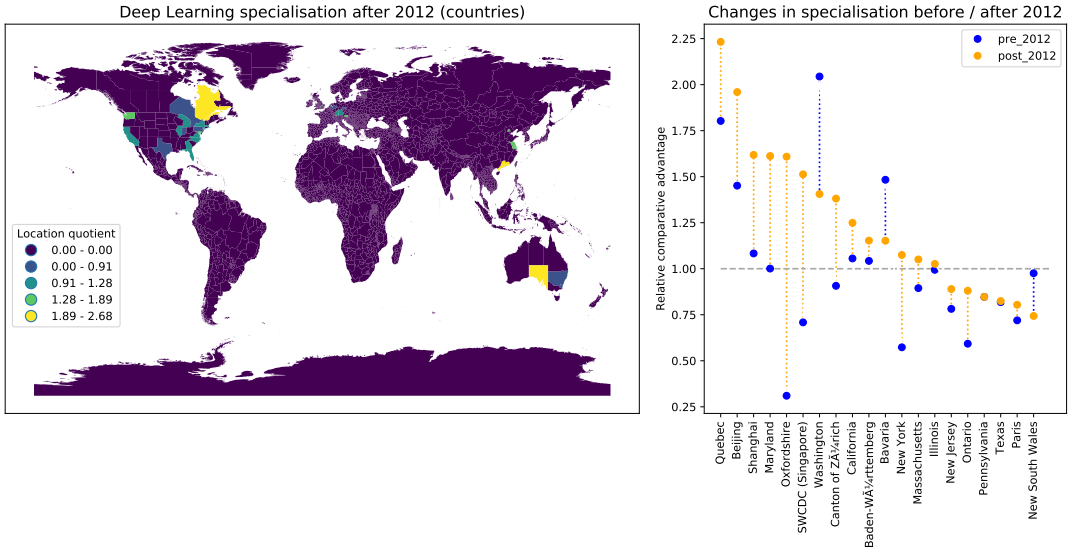

FIGURE 12: *The map in the left panel shows RCA_{dl} by region for papers published after 2012, focusing on papers above the median of citations for papers in their publication year, and region in the top 99 percentile for total level of DL activity. The figure in the right panel compares changes in RCA_{dl} between the period before 2012 and afterwards, focusing on the top 20 regions by total level of DL activity.*

concentration afterwards. Having said this, the pattern of dis-location in DL is also present in the broader `arXiv` corpus, suggesting that changes in concentration could be influenced by other factors, such as growing use of `arXiv` or lower barriers to entry at the beginning of the period, with open resources such as `arXiv` allowing more locations to participate in computer science research. These are all interesting questions to explore in further work.

Our analysis also shows that DL research is more geographically concentrated than computer

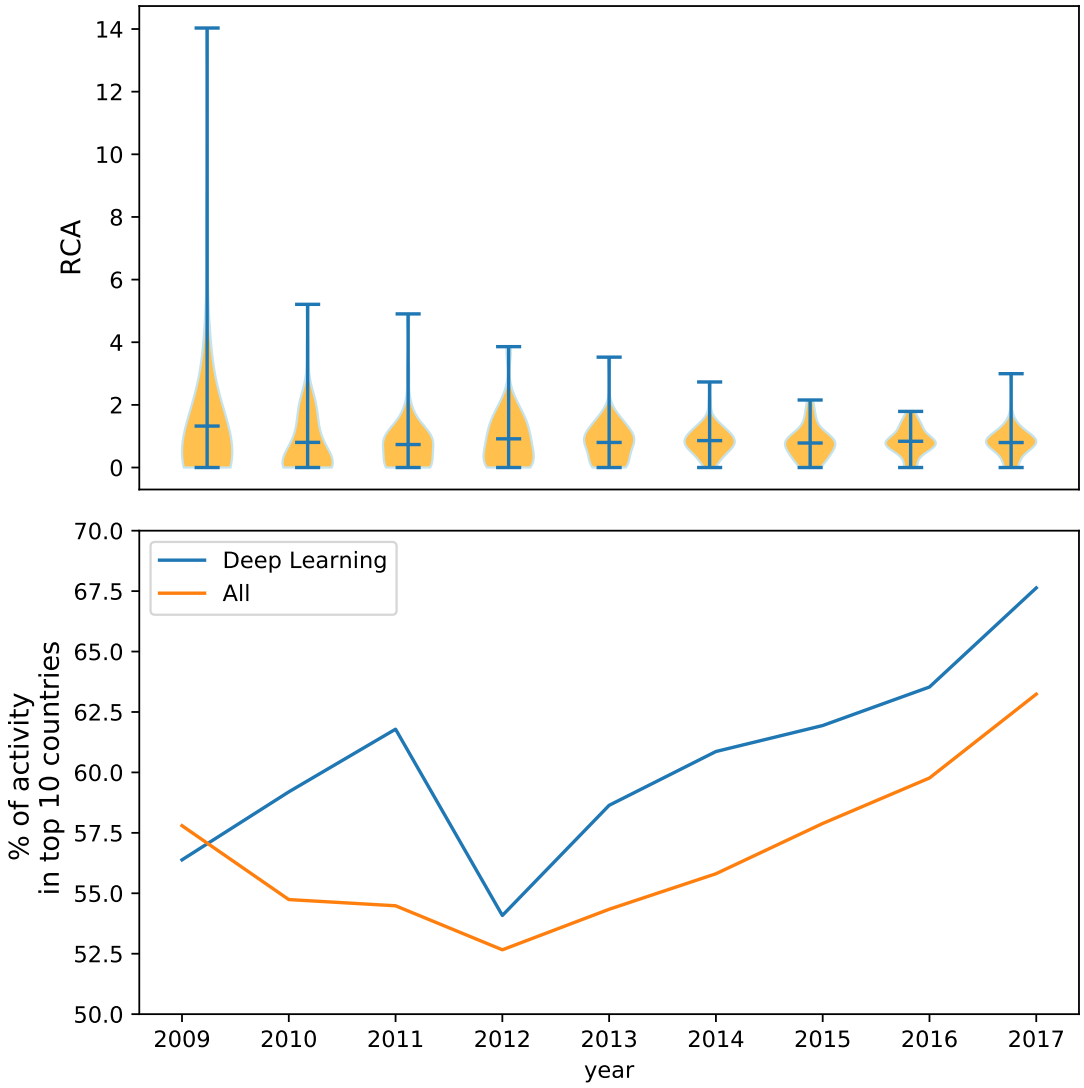

FIGURE 13: *The top panel shows the evolution in the dispersion of RCA_{dl} by country between 2009 and 2017 only considering papers above the citation median for the year, and the top 50 countries by level of activity in* `arXiv`*. The bottom panel shows the percentage of highly cited papers that concentrate in the top 10 countries*

science overall. One potential explanation is that DL research is more complex, requiring proximity for successful collaboration ([40] find something similar in their analysis of complex technologies using patent data). This is what we would expect in a GPT that relies on coordination between developers and adopters. We focus on that interaction for the remainder of this section.

3.5 Drivers of DL cluster emergence

After showing that DL behaves like a GPT in its growth, diffusion, impact and geography, we turn to the analysis of the local drivers associated with its development. As we said, GPTs benefit from coordination between developers and adopters: developers aware of market needs can customize and promote their technologies to new industries. Adopters aware of GPT opportunities can find new ways to apply these technologies to their own situation. We would expect this mutual awareness to be higher when developers and adopters are close to each other, making it easier to collaborate, network and share knowledge. This means that regions where developer and adopter

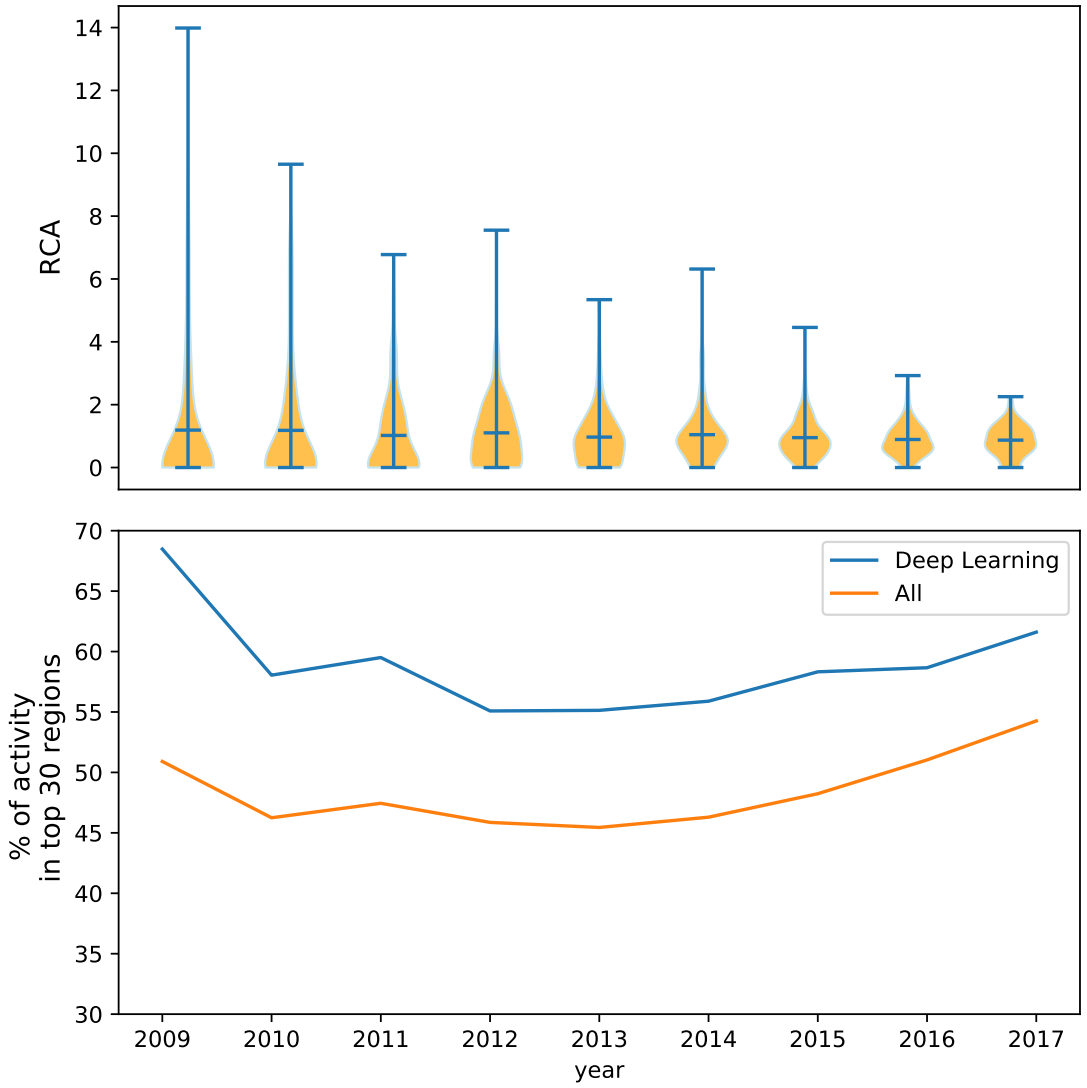

FIGURE 14: *The top panel shows the evolution in the dispersion of RCA_{dl} by region between 2009 and 2017 only considering papers above the citation median for the year, and the top 150 regions by level of activity in* ***arXiv****. The bottom panel shows the percentage of highly cited papers that concentrate in the top 30 regions*

sectors co-locate should be more competitive in the development of a GPT.

We test this hypothesis with the following model specification:

$$\begin{aligned} \text{RCA}_{dl,t1} = &\beta_0 + \beta_1 \text{RCA}_{dl,t0} + \beta_2 \text{arXiv}_{sp} + \beta_3 \text{CrunchBase}_{sp} + \\ &\beta_4 \text{arXiv}_{sp} \text{CrunchBase}_{sp} + \beta_5 \text{arxiv}_{sp} * \text{CrunchBase}_{tot} + \\ &\beta_6 \text{arXiv}_{tot} + \beta_7 \times \text{is\_China} + \epsilon \end{aligned} \tag{6}$$

In it, we estimate the link between DL specialization after 2012 ($\text{RCA}_{dl,t1}$) and the presence of related research and industrial capabilities (arXiv_{sp} and CrunchBase_{sp}) and their interaction ($\text{arXiv}_{sp} \times \text{CrunchBase}_{sp}$) before 2012, capturing the idea of GPT complementarities between research and industry.[16] We also include an interaction between relevant research capabilities and

[16]The measures of related activity weight levels of regional specialization in research subjects and industrial activities by the DL similarity vectors described in 2 and 3.1.

total `CrunchBase` activity ($\text{arXiv}_{sp} \times \text{CrunchBase}_{tot}$) to capture the benefits from deploying a GPT in industries less directly related to it.

We control for the levels of specialization in DL before 2012, total `arXiv` activity, and a dummy for whether a region is Chinese or not (is_China). We take the logarithm of all totals, calculate z-scores for all variables and focus our analysis on regions in the highest level of `arXiv` activity (i.e. the top quartile) to reduce noise in the RCAs and remove a long tail of regions with little or no DL activity.[17]

The correlation matrix in Figure 15 shows an association between DL_{t1} and several independent variables and controls, including China. The correlation between between arxiv_{sp} and CrunchBase_{sp} is low, suggesting that locations with high specialization in research subjects relevant for DL do not always specialize in relevant industries. Strong correlations between some independent variables suggest the presence of multicolinearity [18].

Table 3 presents the results of our regression analysis with different specifications. Model 4 is the specification in 6. We review some key results:

1. There is a robust link between a region's specialization in DL before 2012 and afterwards. This suggests that the volatility in the geography of DL we described above is not absolute, with some DL specialization persisting over time.

2. There is a significant link between the interactions of arXiv_{sp} with CrunchBase_{sp} and with $\text{CrunchBase}_{\text{tot}}$, and DL_{t1}. This supports the hypothesis that GPT development benefits from the co-location of developers and adopters. Interestingly, once we consider this complementarity, the link between related research activity and a region's comparative advantage in DL loses significance. This suggests that the presence of relevant industries is an important ingredient in the development of a DL cluster.

3. The link between the is_China dummy and the development of a DL cluster after 2012 is strong and significant after we control for other explanatory factors such as regional research and industrial levels of activity. Together with the low R^2 of our models, this suggests that we our model is missing important national and regional factors that play a role in the development of DL clusters such as access to skills and data, infrastructure, regulation and supportive policies [41]. We plan to bring them into the analysis in future work.

We conclude by comparing model outputs for DL with other computer science subjects using the same specification (while focusing in the relevant research and industrial activities for each subject). One could think of these other subjects as quasi-controls allowing us to explore whether the patterns we detect in DL are also present in other fields, or DL is unique in some way. Through this, we also attempt to control for other trends which could be driving our results, such as secular changes in the usage of `arXiv`.

The results in Figure 16 shows that, in general, the interactions between `arXiv` and `CrunchBase` activity that we have detected in DL are not pervasive amongst other DL subjects. Interestingly, complementarities between research and industry are more important for data-related subjects such as Computer Vision, Computer Learning, Machine Learning or Computer Learning. Other subjects such as Data Structures (`cs.DS`), Network architecture (`cs.NI`), Social and Information Networks or Logic (`cs.LO`) seem less reliant on these complementarities, perhaps because they are more mature (reducing the need for coordination between developers and adopters).[19] It is also worth noting that DL and Computer Vision are the main subjects with a strong and positive association between the is_China dummy and subject specialization, suggesting that China has specific endowments that facilitate the development of these subjects, such as large unstructured datasets and targeted policies .[20]

[17] Our results are robust to changes in these thresholds

[18] During our robustness tests we have removed some of these interaction terms without significant changes in the results.

[19] The exception to this, Information Theory (`cs.IT`) appears to be a catch-all subject present in almost 20% of the computer science `arXiv` corpus

[20] This result will also be driven by the overlaps between DL and Computer Vision outlined in 3.2

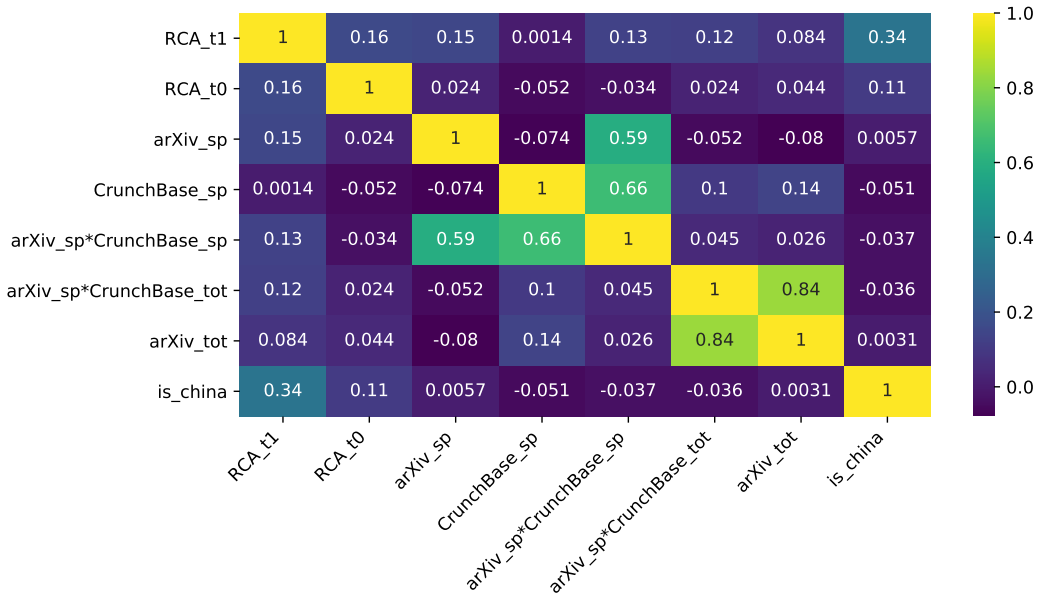

FIGURE 15: *Correlation matrix between key variables in our model. All variables have been normalized.*

	Model 1	Model 2	Model 3	Model 4
y	RCA_t1	RCA_t1	RCA_t1	RCA_t1
RCA_{t0}	0.12***	0.121***	0.124***	0.126***
	(0.044)	(0.044)	(0.044)	(0.043)
$arXiv_{sp}$	0.155***	0.156***	-0.012	0.006
	(0.044)	(0.044)	(0.084)	(0.084)
$CrunchBase_{sp}$		0.023	-0.162*	-0.135
		(0.044)	(0.09)	(0.09)
$arXiv_{sp} \times CrunchBase_{sp}$			0.261**	0.229**
			(0.111)	(0.111)
$arXiv_{sp} \times CrunchBase_{tot}$				0.207**
				(0.08)
$arXiv_{tot}$	0.09**	0.086*	0.092**	-0.083
	(0.044)	(0.044)	(0.044)	(0.081)
is_China	1.54***	1.545***	1.549***	1.586***
	(0.213)	(0.213)	(0.212)	(0.212)
R^2	0.147	0.146	0.154	0.165
n	451	451	451	451

TABLE 3: *Dependent variable is $RCA_{dl,t1}$. Standard errors in brackets are clustered by country. *** p<0.01, ** p<0.05, * p<0.10.*

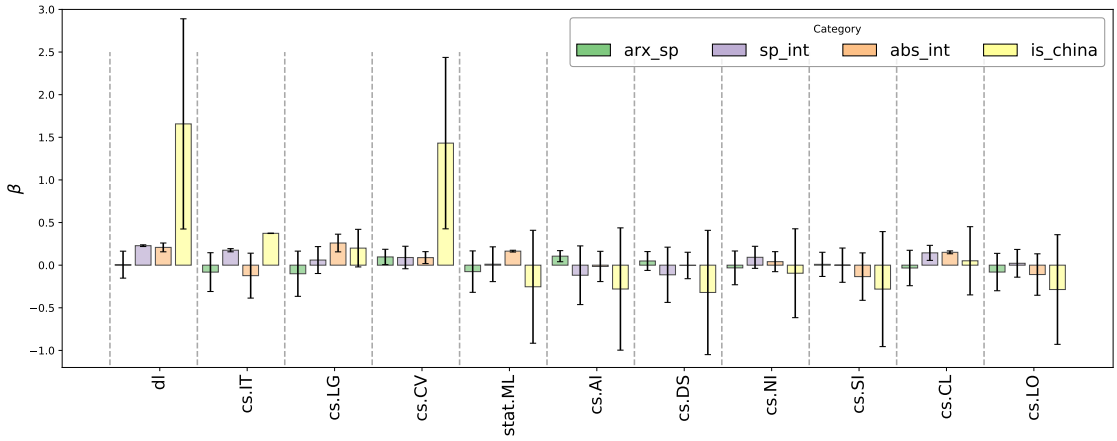

FIGURE 16: *Regression coefficients and confidence intervals for models using the specification in 6 in* `arXiv` *subjects with the highest levels of activity*

4 Conclusion

4.1 Discussion and implications

We have studied the geography DL, a new paradigm for AI. Our analysis of `arXiv`, a popular preprints website used by researchers in academia and industry supports the idea that DL has the features of a GPT technology: it has experienced rapid growth and is being applied in an increasing number of computer science subjects where it generates high-impact work (which we proxy with citations). This confirms the conclusions of previous studies such as [9], and also suggests that in spite of recent criticisms of the DL paradigm, and in particular the lack of robustness stemming from its reliance on large datasets for training [42]), researchers in multiple domains of computer science who are perhaps less likely to be swayed by hype than policymakers and entrepreneurs, are applying it in ways that their peers find interesting and useful.

If DL is a GPT, what are the geographical dimensions of its development? Our review of the literature suggested that the emergence of a GPT might involve an initial shift in the geography of research as new 'entrants' come into the scene, followed by consolidation as central hubs of activity emerge. Our analysis at the national and regional level support this idea: we see international shifts in activity since 2012, when DL started to gain visibility, followed by growing geographical concentration. We also note that DL is at all points more geographically concentrated than computer science research, lending support to the hypothesis in [30] that knowledge spillovers in AI research are localized, justifying national and sub-national policies to support its development.

This higher geographical concentration also suggests that DL researchers benefit from co-location. We have further studied this idea with a model that estimates the link between co-location of relevant research and industrial capabilities and DL development. The results of the analysis, considering DL on its own and comparing it with other computer science subjects, supports the idea that the co-location of researchers able to develop a GPT and adopters who can explore its application favors the development of stronger DL research clusters. This result is also present in other data analytics subjects, highlighting the link between AI and broader trends towards 'datafication' in the economy and society [43].

In terms of policy, our findings suggest that the attention that DL is attracting from national and regional policymakers is warranted by its GPT nature and evidence of localized knowledge spillovewrs. What is less clear is the extent to which the 'window of opportunity' to enter the field remains open given the growing concentration of DL research activity that we identify. Our findings also echo the public narrative about the emergence of China as a global AI leader (together with the USA, Canada and Asian countries such as Singapore and Korea and, perhaps to a lesser extent, the UK), while EU countries lag behind.

Our analysis of the drivers of DL cluster development support the idea that co-location and collaboration in dense ecosystems of research and industrial activity offers a fertile ground for the development of GPTs that rely on new combinations of ideas from various fields and applicable in multiple sectors. Proximity between researchers and businesses could address some of the coordination failures between GPT developers and adopters identified in the literature [1]. One important challenge for policymakers is how to enhance these complementarities without exacerbating regional inequalities. While a geographical diversity of needs could justify dispersing research geographically so as to explore DL opportunities in a wider set of industrial and social contexts, this might weaken agglomeration economies and knowledge spillovers derived from clustering. New, detailed and timely sources of data such as those we use in this analysis can help understand and balance these trade-offs.

4.2 Limitations and issues for further research

Our use of `arXiv` data raises some concerns. To begin with, this is a platform with low barriers to entry, so many of the papers there might be of low quality. We have tried to address this problem by matching the `arXiv` data with `MAG`, and focusing key parts of our analysis on highly cited, hopefully higher quality papers. Future work should expand and further validate our conclusions in other data sources such as patents or open source projects.

Second, to which extent does our research data capture changes in technology development and business diffusion? Throughout our analysis we have assumed that the clustering of DL research is a good proxy for DL R&D development activities with an industrial application. Although anecdotal evidence suggests high level of industry participation in `arXiv`, and we find a strong correlation between the levels of activity in `arXiv` and `CrunchBase`, there is risk of biases if different research communities, sectors or countries display variation in their propensity to publish their work in `arXiv`. Further triangulation of `arXiv` data with other sources, including peer-reviewed research in comparable disciplines, as well as industry patenting and the financial performance of companies in DL-related sectors, would help to address these concerns.

Third, there is the issue of causality. While our analysis has a longitudinal dimension, and qualitatively controls for unobservables by comparing DL model estimates with other computer science subjects, we cannot rule out that other local factors such as access to skills and finance or a supportive policy environment might be underpinning the links between research and industrial activity and DL research clustering that we have detected. Going forward, we would like to incorporate in our analysis shocks to industrial activity with an exogenous element, such as regulatory changes, or industrial policy interventions so as to identify more precisely the causal effects of research/industry co-location in DL cluster development.

There are many interesting directions to extend our work:

First, our analysis says little about the mechanisms behind the link between research / industry co-location and DL cluster development: are these links driven by knowledge spillovers, the formation of a technical talent pool that researchers and industry both tap on, or access to finance (e.g. adopters fund development activities in regional research institutions)? A better understanding of those mechanisms would help to address the issues of causality above, and yield policy-relevant implications about what programs to put in place to strengthen DL clusters.

Second (and relatedly), our analysis takes a siloed view of DL research clusters, only considering geographical proximity to other DL researchers and technology businesses as a source of valuable knowledge about new techniques and business applications. In reality, researchers access this knowledge through many other channels and further afield, including via popular international collaborations such as NIPS. Going forward, we will address this by studying the network of co-authorships and citations in our data, and trying to understand the role of international conferences in the dissemination of knowledge in DL. This analysis could reveal cross-country flows of ideas and collaborations going against the narrative of a zero-sum global AI race dominating popular debates.

Third and last, we have not considered in detail the technological characteristics of the DL 'dominant design': what are its features and components, and how stable are they? What are the

parallel paths for DL that have been explored and set aside? Should some of them be maintained to avoid a premature lock-in to suboptimal standards for the large-scale deployment of the AI GPT [44]? As we mentioned before, some researchers have expressed concerns about the lack of robustness and interpretability in DL systems, calling for their combination with older paradigms for AI development. New techniques and methods are being developed in response to this. Identifying what they are, and overlaying their geography with the geography of DL explored in this paper could yield a richer understanding of the diversity of evolutionary paths for emerging technologies, and their spatial dimensions. Rich text data from papers could be marshaled for this, using the same NLP approach we followed in this paper.

All these ideas highlight the analytical and policy opportunities for using new data sources for the analysis of emerging technologies, and turning AI-related methods and tools towards the analysis of AI itself.

References

[1] Timothy F Bresnahan and Manuel Trajtenberg. General purpose technologies 'engines of growth'? *Journal of econometrics*, 65(1):83–108, 1995.

[2] Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In *Economics of Artificial Intelligence*. University of Chicago Press, 2017.

[3] Joel Mokyr et al. *The gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press, 2002.

[4] Stephen Cave and Seán SOhÉigeartaigh. An ai race for strategic advantage: Rhetoric and risks. In *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*, 2018.

[5] David B Audretsch and Maryann P Feldman. Innovative clusters and the industry life cycle. *Review of industrial organization*, 11(2):253–273, 1996.

[6] Allen Scott and Michael Storper. Regions, globalization, development. *Regional studies*, 37(6-7):579–593, 2003.

[7] Timothy Bresnahan and Pai-Ling Yin. Reallocating innovative resources around growth bottlenecks. *Industrial and Corporate Change*, 19(5):1589–1627, 2010.

[8] Ajay K Agrawal, Joshua S Gans, and Avi Goldfarb. Economic Policy for Artificial Intelligence. Working Paper 24690, National Bureau of Economic Research, June 2018.

[9] Iain M Cockburn, Rebecca Henderson, and Scott Stern. The impact of artificial intelligence on innovation. Technical report, National Bureau of Economic Research, 2018.

[10] Jason Furman and Robert Seamans. Ai and the economy. Technical report, National Bureau of Economic Research, 2018.

[11] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[12] Koen Frenken, Frank Van Oort, and Thijs Verburg. Related variety, unrelated variety and regional economic growth. *Regional studies*, 41(5):685–697, 2007.

[13] Bronwyn H Hall and Manuel Trajtenberg. Uncovering gpts with patent data. Technical report, National Bureau of Economic Research, 2004.

[14] Paul A David. The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review*, 80(2):355–361, 1990.

[15] W Brian Arthur. *The nature of technology: What it is and how it evolves*. Simon and Schuster, 2009.

[16] Elhanan Helpman and Manuel Trajtenberg. A time to sow and a time to reap: Growth based on general purpose technologies. Working Paper 4854, National Bureau of Economic Research, September 1994.

[17] Michael E Porter. *Clusters and the new economics of competition*, volume 76. Harvard Business Review Boston, 1998.

[18] P. Anderson and M. L. Tushman. Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative science quarterly*, pages 604–633, 1990.

[19] W. J. Abernathy and J. M. Utterback. Patterns of industrial innovation. *Technology review*, 80(7):40–47, 1978.

[20] Steven Klepper. Entry, exit, growth, and innovation over the product life cycle. *The American economic review*, pages 562–583, 1996.

[21] A. Scott and M. Storper. Regions, globalization, development. *Regional studies*, 37(6–7):579–593, 2003.

[22] Mercedes Delgado, Maryann Feldman, Koen Frenken, Edward Glaeser, Canfei He, Dieter F Kogler10, Andrea Morrison, Frank Neffke11, David Rigby12, Scott Stern, et al. The principle of relatedness. In *Unifying Themes in Complex Systems IX: Proceedings of the Ninth International Conference on Complex Systems*, page 451. Springer, 2018.

[23] Ron Boschma. Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74, 2005.

[24] Matt Taddy. The technological elements of artificial intelligence. Technical report, National Bureau of Economic Research, 2018.

[25] J. Markoff. Machines of loving grace: The quest for common ground between humans and robots. *HarperCollins Publishers*, 2016.

[26] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The simple economics of artificial intelligence*. Harvard Business Press, 2018.

[27] I. Goodfellow et al. Deep learning. *MIT press*, 2016.

[28] A. Karpathy. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*, 2015.

[29] Ajay Agrawal, John McHale, and Alex Oettl. Finding needles in haystacks: Artificial intelligence and recombinant growth. Technical report, National Bureau of Economic Research, 2018.

[30] Avi Goldfarb and Daniel Trefler. Ai and international trade. Technical report, National Bureau of Economic Research, 2018.

[31] Greg Williams. Why China will win the global race for complete AI dominance. *Wired UK*, April 2018.

[32] Ian Sample Science editor. Scientists plan huge European AI hub to compete with US. *The Guardian*, April 2018.

[33] AI Index. The Artificial Intelligence Index: 2017 Annual Report. Technical report, 2017.

[34] Nick Bostrom. Strategic implications of openness in ai development. *Global Policy*, 8(2):135–148, 2017.

[35] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 1966.

[36] G. Ver Steeg and A. Galstyan. Discovering structure in high-dimensional data through correlation explanation. *Advances in Neural Information Processing Systems (NIPS)*, 27, 2014.

[37] Jean-Michel Dalle, Matthijs den Besten, and Carlo Menon. Using crunchbase for economic and managerial research. 2017.

[38] Stefano Breschi, Julie Lassébie, and Carlo Menon. A portrait of innovative start-ups across countries. 2018.

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[40] Pierre-Alexandre Balland and David Rigby. The geography of complex knowledge. *Economic Geography*, 93(1):1–23, 2017.

[41] Miles Brundage. Modeling progress in ai. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

[42] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[43] Mayer-Schönberger Viktor and Cukier Kenneth. Big data: A revolution that will transform how we live, work, and think. *Houghton Mifflin Harcourt*, 2013.

[44] Philippe Aghion, Paul A David, and Dominique Foray. Science, technology and innovation for economic growth: linking policy research and practice in 'stig systems'. *Research policy*, 38(4):681–693, 2009.