Cinus et al. *EPJ Data Science* (2025) 14:51
https://doi.org/10.1140/epjds/s13688-025-00568-4

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Measuring how computer science research translates into innovation and development

Federico Cinus[1], Ali Septiandri[2], Marios Constantinides[2,3*] and Daniele Quercia[2,4]

Handling Editor: Johannes Wachs

*Correspondence:
marios.constantinides@cyens.org.cy
[2]Nokia Bell Labs, Cambridge, UK
[3]CYENS Centre of Excellence, Nicosia, Cyprus
Full list of author information is available at the end of the article

Abstract

What factors determine the impact of a scientific paper? Does its impact extend to patents and software development, or is it primarily confined to academic circles? While current literature predominantly adopts a descriptive approach, emphasizing patent citations as indicators of industry impact, the role of research in driving software development is overlooked. To address this gap, we quantitatively assessed the impact of research papers on both patents and software repositories. With a computational social science approach, we collected, curated, and analyzed a large-scale dataset of 200K papers published between 1980 and 2022 across the research areas of AI, Computer Vision, Data Mining, Databases, HCI, and NLP, including conferences like NeurIPS, ICML, ACL, CVPR, CHI, KDD, and The Web Conference. We found that, on average, 7.1% of papers from these venues became patents and 11.6% went into repositories—significantly higher than top general science journals (3.8% for patents and 0.02% for repositories). Despite being a minority, these papers have received a disproportionate number of citations—4% of AI papers became patents, and 18% went into repositories, yet they have received 29% and 42% of the area's academic citations, respectively. However, after correcting for papers published at different times with survival analysis, we found that there is a significant time lag between patents or repositories and papers (10-15 years for patents, 5 years for repositories in Computer Vision and NLP, and even longer for top general science journals at 30 years). As for consistent trends, Deep Learning has become exponentially popular, and "papers with code" are becoming the norm. Finally, we showed that a paper's publication venue and the extent to which a paper builds upon (un)conventional knowledge determine the impact on patents and repositories, with greater conventionality predicting impact on patents, and lesser conventionality predicting impact on repositories.

Keywords: Research impact; Innovation; Development; Patents; Repositories

1 Introduction

The impact of research on innovation is often measured by the amount of scientific knowledge that is 'transferred' to produce novel products and services [1–4]. Typically this impact is measured through bibliometrics and network science, both of which require a combination of heterogeneous data to understand the dynamics and impact of scientific research [4]. In practice, Semantic Scholar [5] and Microsoft Academic Graph (now Ope-

Springer

nAlex [6]) have been used [2] to model citations among millions of research papers using tools such as Grobid [7] to generate bibliographic data and Crossref to cross-reference them [8]. By analyzing 70 thousand patent citations to premier Human-Computer Interaction (HCI) research venues, Cao et al. [2] found that HCI research has a significant impact on innovation, with systems-oriented research being the most highly cited by patents. More broadly, by analyzing 4.8 million U.S. patents and 32 million research papers, Ahmadpoor et al. [9] found that nanotechnology and computer science were among the fields closest to patents. It has also been observed that high-impact science is primarily built upon a combination of well-established ideas and some new, and unconventional ideas, emphasizing the importance of striking a balance between the two [10].

However, this body of literature may overlook another crucial aspect of research impact, that is, its influence on software development. A relatively small body of literature has explored the characteristics of influential code repositories, similar to the study of influential accounts on social media [11]. This literature found that influential repositories on GitHub are, unsurprisingly, those with high numbers of followers [12, 13], which, in turn, tend to guide others towards new software projects [14, 15]. In the contemporary research landscape, the interplay between academia and development is becoming increasingly important, with the potential to drive innovation and spur technological advancements. Hence, it is important to quantify the full extent of research papers' impact on both innovation and development while considering a broader spectrum of influential factors.

To partially address these gaps, we explore the factors that contribute to the impact of scientific papers, aiming at gaining a comprehensive understanding of the two pathways through which research can influence innovation and development. Drawing upon previous literature [2, 9, 10, 16], we ask questions about factors related to temporal aspects, the distribution of topics, authors' affiliations, and the extent to which papers draw upon (un)conventional knowledge. These questions are not typical research questions with hypotheses but rather focus on the factors that make research papers impactful. The simplicity hides substantial work on cross-referencing a variety of datasets (which have never been cross-referenced at this scale). In answering these questions, we made two main contributions:

1. We collected and curated a large-scale dataset of 200,000 research papers published between 1980 and 2022 across the research areas[1] of AI, Computer Vision, Data Mining, HCI, and NLP (Sect. 3). This dataset contains 85,000 citations between these papers and United States Patent and Trademark Office (USPTO) patents, and 350 million stars and forks of the GitHub repositories associated with these papers. Using this dataset, we developed a set of five metrics that capture the amount of research impact on patents and repositories, its time evolution, and the main factors that shape it, including scientific topics, authors' institutions, and papers' conventionality.

[1] These areas were selected among the fifty-six subcategories within Google Scholar's "Engineering and Computer Science" category, and they are considered foundational to modern Computer Science with their papers impacting many disciplines (e.g., drug discovery and protein folding [17]). However, this classification puts NLP, Computer Vision, and AI separately. Although there is an overlap, for example, in multimodal AI systems combining language and vision, Google Scholar's classification considered these three areas to maintain distinct identities due to unique challenges and data types they handle. At the same time, our selection criteria and the focus on specific venues may introduce biases, particularly in terms of the overrepresentation of highly cited papers and the underrepresentation of niche subdomains. To partly tackle that, we included a comparative sample of average (papers from venues with lower h-index) computer science papers.

2. We found that a small number of highly-cited papers went into (were cited by) patents or into (were linked by) repositories (Sect. 4). Overall, 9% of them went into patents and 11% into repositories, but with a significant time lag for this translational impact. This lag goes from 2 to 30 years for patents, and for repositories, it is either 1 year or the paper does not go into a repository at all. Also, the extent to which a paper builds upon (un)conventional knowledge determines its impact on patents and repositories, with greater conventionality being associated with patents, and lesser conventionality with repositories.

In light of these results, we discuss the implications of our work for future quantitative analyses and modeling of the impact of research papers on society (Sect. 5).

2 Related work

The computing field continually advances through specialized tasks such as language generation and pattern recognition. Conferences like NeurIPS, ICML, ACL, CVPR, KDD, CHI, and The Web Conference (Table 4, Appendix) stand as vital hubs for knowledge exchange across various computational domains, with their papers significantly influencing the modern technological advancements.

The interplay between these research outputs and their subsequent impact on generation of patents and software repositories has been explored, but these two have been mainly studied in isolation.

Impact of research on innovation It is often measured by the amount of scientific knowledge that is 'transferred' to produce novel products and services [2, 3]. Two typical ways of measuring this impact are through bibliometrics and network science, both of which require a combination of heterogeneous data.

Semantic Scholar [5] and Microsoft Academic Graph (now OpenAlex [6]) have been used [2] to model the citations among millions of research papers. That was achieved by using: Grobid [7] to generate bibliographic data; Crossref to cross-reference such data [8]; and bespoke tools to ultimately match such data with innovation metrics [18].

More recently, by analyzing 70K patent citations to premier HCI research venues, Cao et al.'s [2] study found that HCI research has a substantial impact on innovation, particularly with systems-oriented research being the most frequently cited by patents. Tijssen [19] examined the impact of Dutch-authored research papers on international inventions, emphasizing the complex dynamics of knowledge flows, including indirect links between research and innovation. Despite similarities with our work, their method relied primarily on simple bibliometric counting methods, which may not fully capture the extent of these indirect impacts.

The field of scientometrics, leveraging tools from network science, provides a quantitative approach to understanding the dynamics and impact of scientific research [4]. By analyzing 4.8 million U.S. patents and 32 million research papers, Ahmadpoor and Jones [9] found that mathematics was the field most distant from patents, while nanotechnology and computer science were among the fields closest to them. Mariani et al. [20] analyzed the U.S. patent citation network from 1926–2010, and introduced the rescaled PageRank metric, which proved more effective in early identification of historically significant patents than traditional citation counts. Similarly, Park et al. [21] examined six decades

of data across 45 million papers and 3.9 million U.S. patents, and found that recent papers and patents are more closely following established trends in science and technology, making it less likely to break from a field's traditions and forge new research directions.

Impact of research on development The impact of research on development remains relatively unexplored. However, a body of literature has explored the characteristics of influential code repositories, similar to the study of influential accounts on social media [11]. This literature found that influential repositories on GitHub are, unsurprisingly, those with high numbers of followers [12, 13], which, in turn, tend to guide others towards new software projects [14, 15].

Research questions In summary, our study seeks to address existing research gaps by conducting a comprehensive analysis to assess the impact of research on both innovation and development. For the purposes of this study, we define innovation as the process of translating research findings into new technologies, products, or methods that create significant advancements in a field. Development, on the other hand, refers to the subsequent stage where these innovations are refined, scaled, and integrated into practical applications, including commercial products, industrial processes, or societal implementations. Drawing upon previous literature [2, 9, 10, 16], we set out to understand the factors that make research papers impactful, including temporal aspects, the distribution of topics, authors' affiliations, and the extent to which papers draw upon (un)conventional knowledge. More specifically, we formulated five key questions:

- Q_1: *Impact.* What is the impact of papers on patents and repositories?
- Q_2: *Time.* How long does it take for papers to have an impact?
- Q_3: *Topics.* Which topics determine the impact?
- Q_4: *Institutions.* Which institutions produce impactful papers?
- Q_5: *Conventionality.* How does combining conventional and unconventional knowledge in papers affect their impact?

3 Methods

3.1 Data collection

Papers We collected 200 million papers from Semantic Scholar's [5] May 2023 release, and filtered them based on three criteria. Papers were retained if they: *a)* had comprehensive metadata fields, including 'venue', 'title', 'year', 'citation count', 'citations', 'DOI', 'ArXiv id', and 'authors'; *b)* were published between 1980 and 2022; and *c)* were written in the English language. This filtering yielded a dataset of 50 million papers.

Patents We collected 15 million USPTO patents from the Google Patents Public Dataset, filed between 1980 and 2022. For each patent, we extracted relevant attributes using BigQuery [22], including publication number, country code, publication date, inventor data, abstract, title, and non-patent citations.

Repositories We collected 1 million papers from "Papers With Code" [23] that had at least one link to a GitHub repository.

To allow for reproducibility, we made our data and code publicly available: https://social-dynamics.net/impact and discuss the data sources in detail in Appendix B.

3.2 Data preprocessing

We first extracted citations by matching papers referenced in patents with the Semantic Scholar metadata. We then identified and matched the titles of the venues under study (Table 4, Appendix), and filtered the dataset based on this selection.

Retrieve citations in patents To link Semantic Scholar's papers metadata and USPTO patents, we faced a challenge: patents cite documents using varied plain text formats without consistent document identifiers, apart from other cited patents. To systematically extract the title and author data from these references, we used the deep-learning framework Grobid [7]. For data retrieval efficiency, we indexed both Semantic Scholar papers and references within USPTO patents by the first letter of the author's last name and the initial letter of the paper's title. Titles from both sources were then embedded using the Sentence-T5 model [24].

After embedding, we executed pairwise comparisons within each indexed group, characterized by the initial letter of the author's last name and the title's first letter. Cosine similarity was used to identify potential title matches by applying the elbow rule on the similarity curve to set a distance threshold at 0.06 (Appendix, Fig. 7a). To ensure precision, we cross-referenced author names from Semantic Scholar with those in patent references using the Levenshtein distance metric as suggested by Marx and Fuegi [25]. We set a similarity threshold above 0.8 for this comparison using an elbow strategy (Appendix, Fig. 7b). Our method overcomes the shortcomings of two alternative methods: the rule-based scoring technique [25] and a pure machine learning-based approach [8]. The former comes with reproducibility issues and adaptability to new datasets [26], while the latter is not scalable, given the extensive pairwise comparisons.

Select research areas To ensure comprehensive coverage, we selected venues from six core Computer Science areas using Google Scholar's classifications: *Artificial Intelligence (AI)*; *Human-computer interaction (HCI)*; *Natural Language Processing (NLP)*; *Database & Information Systems*; *Data Mining & Analytics*; and *Computer Vision*.

For comparative analysis, for each area, we included ten highly-ranked venues based on Google Scholar's h5 index.[2] Papers from these venues account for over 88% of all citations within each area (Fig. 6, Appendix). These areas were benchmarked against two reference areas: average *Computer Science* and top *General Science*. For the average Computer Science, we randomly selected ten venues from the top 10 listings in other Computer Science venues, and for the top General Science, we selected the ten most impactful venues, including Nature, Science, and Lancet. The full list of venues and statistics is provided in Table 4, Appendix. Finally, we manually matched the titles of these venues with any alternative titles found on Semantic Scholar.

Final datasets statistics After excluding papers from the comparative groups (i.e., average Computer Science and top General Science), we were left with 200,000 papers. We excluded benchmarks, tutorials, surveys, and other non-research-oriented publications, which are unlikely to end up in patents or repositories. Of these, 28,000 went into patents

[2] Google Scholar defines the h5 index as the h-index for publications from the past five full years. It represents the highest value of h for which there are h papers released between 2018 and 2022, each receiving no fewer than h citations.

Table 1 Variables used in the analysis

Variable	Description
P	Set of patents in the corpus (1980-2022, USPTO).
R	Set of repositories in the corpus (2008-2022, GitHub).
X	Generic set of citing entities; could be P or R.
U	Set of papers in the corpus (1980-2022, written in the English language).
H	Set of papers from the venues under study.
H'	Set of papers not from the venues under study.
$c_{\text{in}}(y)$	Set of papers associated with a document y (another paper or a patent).
$c_{\text{out}}(u)$	Set of documents (among papers and patents) citing paper u.
$H \mid X$	Set of papers in H that went into patents or into repositories.
$(H \mid X)'$	Set of papers in H that did not become patents or go into repositories.
$H' \mid X$	Set of papers in H' that went into patents or into repositories.
$X \mid H$	Set of patents or repositories that cite papers in H.
$U \mid H$	Set of papers that cite papers in H.
$r(H)$	Proportion of papers in H that went into patents.
$r(H')$	Proportion of papers in H' that went into patents.

(AI: 21.8%, Computer Vision: 30.5%, Data Mining: 9.3%, Database: 20.4%, HCI: 5.9%, NLP: 12.2%) and 34,000 went into repositories (AI: 38.8%, Computer Vision: 28.7%, Data Mining: 5.2%, Database: 5.4%, HCI: 0.4%, NLP: 21.6%).

3.3 Metrics

Our metrics target either *patents* or *repositories*. For simplicity, when discussing either, we use the variable X to indicate one of the two sets. A list of the symbols used is provided in Table 1.

Impact We use citations from patents to measure a paper's impact on innovation, and the creation of a GitHub repository to measure the impact on development. We used a Z-test to evaluate the difference between $r(H \mid X)$ and $r(H' \mid X)$, defined as follows: $r(H) = \frac{|(H|X)|}{|H|}$, and $r(H') = \frac{|(H'|X)|}{|H'|}$, where H is the set of papers from the venues under study; $H|X$ is the set of papers in H that went into patents (or repositories) X; H' is the complement set of papers (i.e., all papers in the venues other than those considered in H); and $H'|X$ is the set of papers in H' that went into patents (or repositories). A straightforward extension to count the number of total citations (denoted as $c(\cdot)$) in the H set is $r(H) = \frac{c(H|X)}{c(H)}$.

To assess the difference in academic impact, we conducted an unpaired t-test between the average citation counts of papers that went into patents (or into repositories) $\mu_{H|X}$, and those that did not $\mu_{(H|X)'}$, where $\mu_{H|P} = \frac{1}{|(H|X)|} \sum_{h \in (H|X)} c_{\text{out}}(h)$, and $\mu_{(H|X)'} = \frac{1}{|(H|X)'|} \sum_{h \in (H|X)'} c_{\text{out}}(h)$.

Time We investigated the time difference between the publication of a paper and its first citation in a patent (or link to a repository). This analysis posed a time-to-event problem involving *right-censored* papers, meaning that the event of interest (first association with a patent) may not occur at the time we wish to draw inferences. To tackle this challenge, we used the survival analysis model Kaplan-Meier estimator [27].

The non-parametric estimator of the survival function $S(t)$, which represents the probability that a paper's "life" (i.e., the time from its publication year until it receives its first citation by a patent) extends beyond time t, is given by $\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$, where t_i denotes a time when at least one event (citations by a patent) occurred, d_i represents the

number of events (i.e., citations by a patent) that happened at time t_i, and n_i refers to the papers known to have "survived" (i.e., not yet cited by a patent or censored) up to time t_i.

Topics To identify the areas where the papers had an impact on patents (or repositories), we conducted a comparative analysis of the topics covered by two sets of papers: those that went into patents (or into repositories) and those that did not. To identify the topics, we employed a Latent Dirichlet Allocation (LDA) model [28] on all the abstracts. We then extracted the topics from both sets (i.e., those that went into patents or into repositories, and those that did not) to conduct further analysis and comparison (Fig. 8, Appendix). To determine the number of topics, we used a coherence-based measure [29] (Fig. 9a, Appendix) as it typically improves topics' interpretability compared to a perplexity measure [28].

Institutions We collected all papers going into patents or repositories and linked these papers to the unique institutions of their co-authors. We then aggregated the total number of such papers for each institution, enabling us to identify the institutions with the highest impact. In our analysis, we focused on the top 50 institutions by the total number of papers and ranked them according to the average proportion of papers going into patents or repositories across all research areas.

Conventionality We studied how conventional and unconventional knowledge affects the impact of papers on patents and repositories. Building upon Uzzi et al. [10]'s methodology, we used the pairwise combinations of references within a paper's bibliography as our proxy for conventionality. For each paper, we defined a conventionality score based on all possible pairs of venues referenced in the paper. For each pair of venues, we computed the empirical co-occurrence observed in the whole dataset and the co-occurrence that would be randomly observed in a null model. More specifically, we used a randomization technique as per [10]. For each paper and venue, we preserved two quantities: the distribution of citations and the chronological sequence of those citations (i.e., the year of publication). To accommodate these criteria, we augmented the pair of venues (venue_i, venue_j) with the publication year of the paper in which the two venues are referenced. Hence, we denote the list of co-occurrences as: $\{(\text{venue}_i, \text{venue}_j, \text{year}) \cdots\}$. To randomize these co-occurrences, we shuffled all entries related to venue_j, pairing them with entries from the same years. This process was repeated 1000 times. As a result, the co-occurrence count remains constant, preserving both citation patterns and chronological order through the selection step in the shuffle process.

3.4 Limitations of data processing

We acknowledge that our analysis might not comprehensively cover all published papers, given its reliance on Semantic Scholar. Our dataset includes papers from the top 10 venues in six core areas of computer science as classified by Google Scholar, with a focus on highly impactful and cited papers. Consequently, our results may represent an upper bound, as less-cited papers are less likely to appear in patents or repositories. At the same time, our dataset covers 200,000 papers written in the English language, which, while extensive, represents a fraction of the total literature. This may introduce language and regional bias, or lead to an overemphasis on trends observed in highly active research areas and institutions, potentially understating the impact in less active regions or subfields. Despite

such limitations, Semantic Scholar is extensively used in previous studies [2] and its scope is comparable to Google Scholar's [30]. Future research could benefit from incorporating alternative databases such as Crossref [31] and OpenAlex [6]. Moreover, while our method of associating Semantic Scholar articles with patent citations through title matching proved to be effective, it is possible that not all relevant papers were identified. Future efforts could investigate alternative title matching strategies such as those provided by Biblio Glutton [8] or the approach by Marx and Fuegi [25]. Another limitation concerns the specific intentions behind citations, and our method's ability to accurately discern them. For example, when we attempted to classify citation intentions through terms suggested by previous work [32] (e.g., extension, future, and use), we obtained suboptimal results due to the difficulty in establishing a reliable benchmark for these intentions. Other potential confounding variables may include the time of paper publication, or the number of authors [33]. While we adjusted for the time of publication through survival analysis, our analysis did not account take into account the number of authors in each paper as it may not be directly linked with research impact. Additionally, our metric for capturing research impact is limited to data exclusively derived from the US Patent and Trademark Office (for patents) and GitHub (for code repositories). Future studies could extend the scope of data sources to achieve a more comprehensive evaluation of research impact. Finally, the use of repositories to measure the impact of computer science research has limitations, especially as many papers include code or datasets at submission time. This may skew our metrics as the initial repository creation may not reflect subsequent impact. However, to partly fix that, we also performed our analysis on top general science papers, which served as a baseline to adjust for any potential bias. Alternative measures include tracking software adoption and usage over time, which can provide a more accurate picture of research impact.

4 Results

Throughout the results, we refer to two sets of papers: *(a) those that go into patents*, and *(b) those that go into repositories*.

4.1 What is the impact of papers on patents and repositories?

A small number of highly cited papers went into patents and repositories. Despite a mere 5.9% of AI papers (out of 105K) went into patents (Table 2, and Table 5, Appendix), these papers disproportionately received 42.8% of academic citations in AI. 15.1% of Computer Vision (out of 57K) and 9.1% of Data Mining (out of 29K) papers went into patents but received as much as 59.8% and 44.8% of academic citations in Computer Vision and Data Mining (Fig. 1a), respectively.

Similarly, 16.9% of Computer Vision and 18.7% of NLP (out of 57K) papers went into repositories, and received 30.6% and 32.4% of academic citations in Computer Vision and NLP. 12.5% of AI papers went into repositories, and received 35.5% of academic citations in AI (Fig. 1b).

The impact of the six research areas is, on average, greater than top General Science and the average Computer Science. The research area of Computer Vision has the highest percentage of papers that went into patents (Table 2, and Table 5 in Appendix), with three times more than top General Science and almost two times more than the average Computer Science papers. Conversely, NLP has the highest percentage of papers that went into repositories, with 18 times more often than the average Computer Science papers.

Table 2 For each research area (in the rows), the numbers in parentheses denote percentages of papers going into patents or repositories, and percentages of academic citations received by papers going into patents or repositories. The percentages of papers going into patents or repositories were calculated by dividing the number of papers going into patents (or repositories) by the total number of papers. Similarly, the percentages of academic citations received by papers going into patents or repositories were calculated by dividing the number of academic citations of papers going into patents or repositories by the total number of academic citations in all papers. The main research areas considered (last six rows) are benchmarked against two reference areas: average *Computer Science* and *top General Science* (first two rows)

Research area	Papers going into patents	Papers going into repositories	Papers	Number of academic citations of papers going into patents	Number of academic citations of papers going into repositories	Number of academic citations
Top General Science	32,550 (5.5%)	244 (0.0%)	586,864	18,343,830 (31.6%)	141,183 (0.2%)	58,141,570
Computer Science	4806 (8.0%)	408 (0.7%)	60,232	734,263 (32.3%)	13,062 (0.6%)	2,270,612
AI	6254 (5.9%)	13,237 (12.5%)	105,762	2,852,156 (42.8%)	2,368,236 (35.5%)	6,667,624
Database	5838 (12.4%)	1831 (3.9%)	47,113	1,138,226 (44.2%)	110,701 (4.3%)	2,576,666
Data Mining	2666 (9.1%)	1770 (6.0%)	29,369	631,109 (44.8%)	148,549 (10.6%)	1,407,968
HCI	1698 (6.4%)	136 (0.5%)	26,733	254,767 (22.3%)	5066 (0.4%)	1,141,780
NLP	3494 (8.9%)	7394 (18.7%)	39,475	883,689 (46.3%)	617,508 (32.4%)	1,908,013
Computer Vision	8742 (15.1%)	9795 (16.9%)	57,864	3,769,944 (59.8%)	1,924,702 (30.6%)	6,299,765

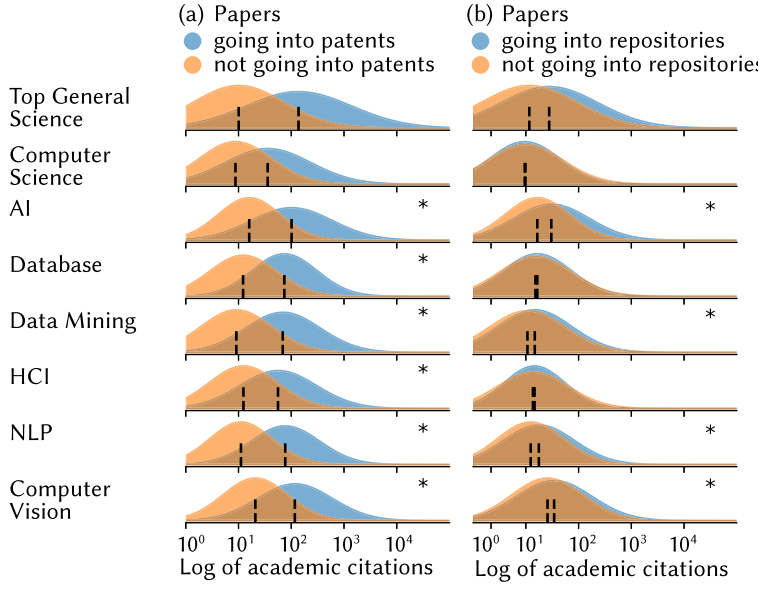

Figure 1 For each research area (in the rows), the number of academic citations is presented on a log scale. (a) Distributions of papers going into patents (blue) and papers not going into patents (orange). (b) Distributions of papers going into repositories (blue) and papers not going into repositories (orange). A wider gap between the distributions in each plot indicates a greater ability of papers going into patents and repositories to attract more academic citations compared to those that did not. The asterisk symbol (*) indicates a statistically significant difference between the two averages of the distributions at the 0.001 level

4.2 How long does it take for papers to have an impact?

The probability of papers going into patents within the first two years after publication is low. This probability significantly increases afterwards and continues to rise, even beyond 30 years (Fig. 2a). This means, for example, that in a given set of 100 Computer Vision

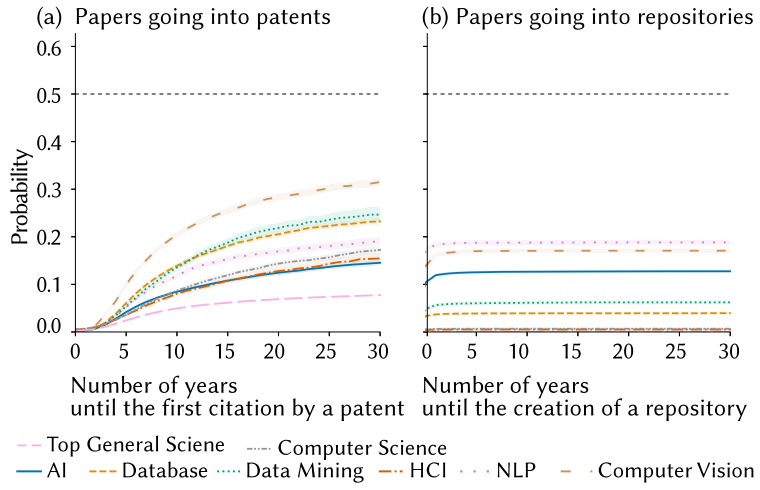

Figure 2 For each research area, the probability of its papers went into: (a) patents for the first time, and (b) repositories for the first time

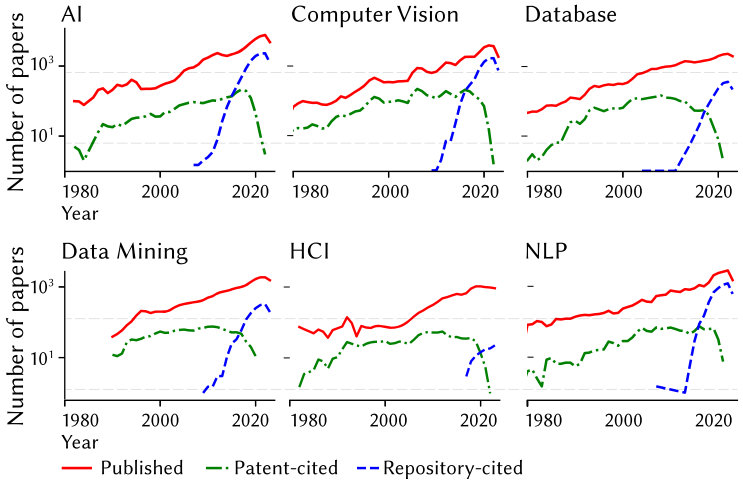

Figure 3 For each research area, we show the total number of published papers (red), papers going into patents (green), and papers going into repositories (blue). Total numbers are log-transformed

papers, none will go into patents within the first two years after publication. After 5 years, 10 papers may go into patents, and this number can increase to as many as 30 papers after 30 years. By contrast, most of the papers (>80%) did not go into repositories, and the ones that did would only do so within the first year of publication (Fig. 2b). This means, for example, in the same set of 100 Computer Vision papers, 15 of them will go in a repository within the first year of publication, and this number will only increase to 16 afterwards, without further increase even after a much longer period.

Across all research areas, we observed a fast publication cycle but a slow patent production. There is an upward trend of published papers, with annual numbers reaching several thousands (Fig. 3). As for papers that went into patents, there was a continuous rise from 1980 to 2018, followed by a decline.

Two explanations could be possible. One possible explanation is the extended publication process for patents compared to papers. As a result, papers can take up to a decade to start gathering patent citations [2]. Therefore, the number of times a paper goes into a patent depends on how many years have passed since the paper was published. To account for this time dependency, we factored in any delays with a survival analysis detailed in Sect. 3. As observed in Fig. 2, this explanation holds true given that the number of patent citations is increasing over time. However, this might not be the sole explanation. A second explanation could be attributed to the prevalence of a research area (i.e., overproduction of papers) that results in an excess of papers, surpassing the rate of patent production. Put differently, should a particular research area become excessively productive, it could create a bottleneck on the long patenting process for which the average total tendency (the time from filing to patent grant) is around 24-30 months [34].

4.3 Which topics determine the impact?

We observed an increasing popularity of State-of-the art NLP models, Graph-based learning, and Data-centric AI. There are three main areas of considerable growth (Fig. 4 - top). First, deep learning topics have seen growth since 2012, with papers that went into patents and repositories making up 30% of the total. This increase may have been influenced by new neural network architectures at that time such as ImageNet [35] and, later, Attention-based models [36]. Second, Graph-based Learning started to have a considerable impact on repositories (10% of the total papers). Third, Data-centric AI papers have consistently influenced both patents (15%) and repositories (10%).

Interest in topics such as Algorithmic optimization in learning has diminished in both patents and repositories (Fig. 4 - middle), despite being a consistently popular topic in academia (Fig. 9b). Similarly, papers associated with Information retrieval & search that went into patents have shown a declining trend starting from 2012, coinciding with the rise of deep learning-based NLP models.

We found an enduring popularity of Computer vision, Computational efficiency algorithms, and Human-centered information systems. Computer vision consistently ranks high in terms of papers going into patents and repositories. In 2012, as much as 15% of papers in Human-centered information systems went into patents. However, papers

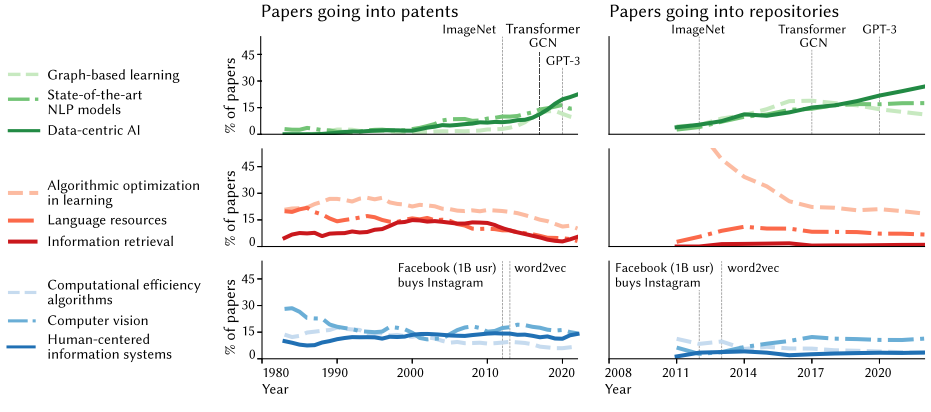

Figure 4 Percentage of papers about a given topic over the years. The percentage values are smoothed using a rolling window of four years for papers going into patents, and two years for papers going into repositories

that went into repositories dipped the very same year, only to gradually recover thereafter (Fig. 4 - bottom).

4.4 Which institutions produce impactful papers?

Among the top 25 institutions, more than half are in the U.S., and one-third are in China. China is narrowing the gap with the U.S. in research output [37], including papers that go into patents, even if we have been considering the U.S. patent system only.

Microsoft (U.S., UK, and China, takes a particularly prominent position when considering papers going into patents (Table 3). This holds true across various research areas, with leadership in Computer Vision (31.5%), Database (19.8%), NLP (22.1%), HCI (19.6%). Stanford University leads in AI (8.7%), while IBM leads in Data Mining (17.8%).

Tencent has 15.7% of papers going into repositories in AI, 23.8% in Computer Vision, and 18.5% in NLP. Fudan leads in NLP (26.1%), followed closely by Google (25.1%) which also leads in Computer Vision (36.7%), and Alibaba leads in Data Mining (19.3%).

4.5 How does combining conventional and unconventional knowledge in papers affect their impact?

A paper's conventionality scores are computed on all possible pairs of its citations, and reflect whether its citations are commonly or uncommonly found in other papers. That is, each citation pair comes with a frequency observed in the dataset, and a frequency that would be randomly observed in our null model detailed in Sect. 3.3. We took the z-scores of the conventionality scores to center them at zero such that values below 0 reflect unconventionality, while those above 0 reflect conventionality. Hence, each paper is associated with a distribution of z-scores (one for each pair of venues in the bibliography). In a conservative fashion, following the methodology in [10], we took the 10th percentile value from each paper, which is the lowest conventionality score that only 10% of papers have. This measure's core concept is that when a Computer Vision paper cites CVPR, ECCV, and ICCV (common Computer Vision conferences), it is considered more conventional (with a score above 0), while citing NeurIPS, ACL, and CHI is less common and, as such, seen as less conventional, resulting in a score below 0. Figure 5 shows the cumulative distribution of such percentile values in a research area.

Papers that went into patents use conventional combinations of prior work, whereas papers that went into repositories blend conventional and unconventional combinations. Papers that went into patents are based on conventional combinations of prior work (Fig. 5a), while those that went into repositories tend to be more unconventional and, as such, more novel (Fig. 5b). To illustrate this difference with a specific example, we examined the two most unconventional papers: *"Happy Dance, Slow Clap: Using Reaction GIFs to Predict Induced Affect on Twitter"* by Shmueli et al. [38] and *"DeepLens: Interactive Out-of-distribution Data Detection in NLP Models" by Song et al. [39]*. These papers draw from uncommon pairs of venues such as CHI and COLING, and CHI and ICLR and went into several repositories.

5 Discussion

By analyzing 200,000 papers from top computer science venues, we found that a small number of highly cited papers went into patents and into repositories. Despite being a minority, these papers receive a disproportionate number of citations. However, the time it

Table 3 The top 25 institutions with the most papers that went into patents and into repositories (rows), divided into six research areas (columns)

Institution	Avg. Impact	Papers going into patents						Papers going into repositories					
		AI	CV	DM	DB	HCI	NLP	AI	CV	DM	DB	HCI	NLP
Google (USA)	12.6%	5.7%	15.7%	11.9%	14.9%	8.6%	9.5%	12.3%	**36.7%**	6.8%	2.9%	0.9%	25.1%
Microsoft (USA)	11.8%	7.2%	24.9%	15.7%	18.5%	**19.6%**	17.8%	2.3%	12.6%	5.2%	2.3%	0.3%	15.4%
Stanford University (USA)	11.4%	**8.7%**	23.9%	14.6%	16.6%	7.8%	15.5%	12.3%	13.4%	11.9%	3.5%	0.7%	8.1%
Microsoft Research Asia (China)	10.9%	8.3%	22.8%	12.3%	14.3%	9.4%	**22.1%**	12.0%	14.6%	7.1%	4.6%	**2.4%**	0.5%
Massachusetts Institute of Technology (USA)	10.8%	7.0%	29.1%	7.8%	9.1%	10.8%	15.5%	9.1%	12.3%	9.8%	6.1%	2.1%	11.0%
Carnegie Mellon University (USA)	10.7%	5.8%	24.4%	14.6%	15.8%	9.0%	14.1%	8.1%	14.2%	4.3%	3.5%	0.3%	14.1%
University of California, Berkeley (USA)	10.1%	3.6%	23.5%	16.0%	18.4%	9.1%	15.2%	8.3%	15.7%	3.9%	1.9%	0.5%	5.5%
Cornell University (USA)	9.0%	5.8%	17.5%	10.7%	13.8%	3.0%	12.9%	5.8%	13.3%	13.2%	4.3%	0.2%	7.5%
ETH Zurich (Switzerland)	9.0%	5.0%	13.7%	2.5%	14.3%	4.9%	0.8%	14.2%	22.9%	4.2%	3.6%	0.7%	20.9%
Microsoft Research (UK)	8.8%	6.6%	**31.5%**	4.9%	**19.8%**	7.0%	11.7%	11.3%	6.2%	1.2%	2.0%	0.2%	3.3%
University of Washington (USA)	8.7%	8.4%	13.6%	11.1%	12.3%	7.6%	12.1%	5.9%	12.3%	3.1%	3.5%	0.5%	14.5%
University of Illinois Urbana-Champaign (USA)	8.5%	3.7%	22.4%	11.8%	11.0%	4.8%	8.3%	9.2%	8.1%	7.7%	6.1%	0.2%	8.3%
Georgia Institute of Technology (USA)	8.3%	4.0%	14.1%	11.7%	8.7%	6.1%	4.5%	7.8%	12.1%	5.9%	5.1%	0.1%	19.9%
Tencent (China)	7.9%	1.7%	1.1%	0.3%	2.4%	1.0%	1.4%	**15.7%**	23.8%	13.2%	14.4%	1.0%	18.5%
IBM (USA)	7.9%	5.5%	4.8%	**17.8%**	13.5%	8.9%	4.7%	8.2%	8.1%	4.8%	1.5%	0.5%	16.0%
University of Massachusetts Amherst (USA)	7.6%	7.2%	18.6%	12.4%	14.5%	1.5%	13.8%	4.1%	4.2%	6.1%	4.2%	1.5%	3.5%
Chinese University of Hong Kong (China)	7.6%	7.0%	12.5%	6.5%	5.6%	1.6%	3.0%	8.9%	21.0%	4.0%	3.2%	0.7%	17.1%
University of Michigan-Ann Arbor (USA)	7.2%	2.2%	13.7%	13.6%	12.9%	2.2%	5.4%	5.9%	10.1%	5.1%	4.2%	0.2%	11.6%
Tsinghua University (China)	7.2%	4.0%	7.4%	4.7%	4.8%	2.2%	4.6%	8.1%	15.0%	8.0%	8.0%	0.3%	19.6%
Alibaba Group (China)	6.8%	0.9%	1.1%	3.6%	2.0%	0.9%	1.4%	5.0%	12.0%	**19.3%**	**15.3%**	0.9%	19.6%
University of Maryland, College Park (USA)	6.8%	3.4%	12.4%	12.3%	12.4%	8.7%	15.7%	7.8%	4.1%	0.8%	0.3%	0.2%	4.1%
Fudan University (China)	6.7%	2.1%	3.4%	2.7%	4.3%	0.6%	6.2%	9.6%	12.1%	7.3%	4.3%	1.4%	**26.1%**
Shanghai Jiao Tong University (China)	6.5%	4.8%	3.5%	4.2%	5.3%	1.9%	4.3%	7.1%	14.7%	11.1%	9.1%	1.1%	11.6%
Univ. of Science and Technology of China (China)	6.5%	2.5%	5.7%	4.9%	4.3%	0.7%	5.8%	8.0%	20.4%	9.0%	11.8%	0.7%	4.0%
Zhejiang University (China)	6.4%	2.7%	3.9%	4.7%	4.8%	1.6%	0.8%	9.1%	14.3%	8.4%	8.1%	0.7%	18.4%

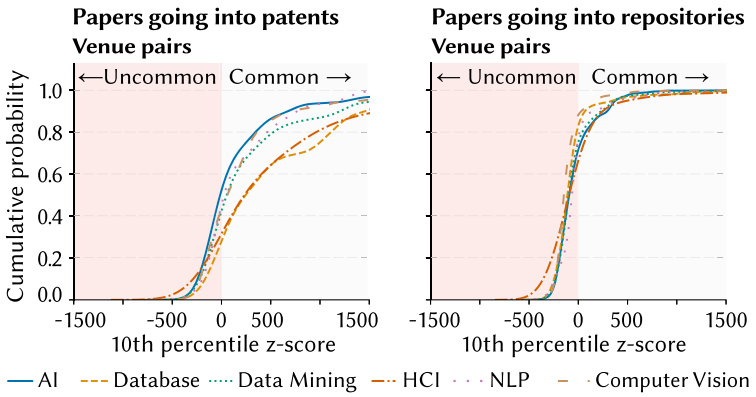

Figure 5 A research area's conventionality based on whether its papers—those that went into patents (left plot) and those that went into repositories (right plot)—have (un)usual combinations of citations. The more the area in the red region, the more unconventional (e.g., AI)

takes for them to go into patents is long, ranging from 2 to 30 years after their publication, with an increasing probability thereafter. Conversely, papers that to go into repositories are either picked up within the first year of their publication (i.e., the time is less than a year) or not at all. Interestingly, the most intriguing aspect of our analysis lies in how the measure of conventionality separates the impact of scientific publications. Papers that build on more conventional knowledge tend to have a greater impact on patents, while those that incorporate more unconventional ideas are more influential in repositories. While patents seem to favor stability and safety, often drawing from well-established knowledge, repositories thrive on novelty, embracing innovative and sometimes risky ideas. This distinction mirrors broader industry trends, such as in aviation, where safety and reliability are paramount [40], leading to a preference for incremental improvements over radical changes. In contrast, fields such AI, where innovation is key, tend to push the boundaries, even at the risk of safety concerns [41].

While our analysis provides valuable insights into the impact of computer science research, readers should interpret the results with caution. The selective nature of our sample may overstate the influence of highly active subdomains such as AI and computer vision while understating emerging or interdisciplinary fields such as Responsible AI. Additionally, the use of patent citations to measure a paper's impact on innovation introduces another layer of complexity. Not all patents are equally innovative, especially in areas such as device manufacturing and software development, where some patents may be less relevant to true technological advancement. This variability could lead to an overestimation of the impact of papers associated with patents that are not highly innovative. Despite that shortcoming, patents have been shown to be a good proxy for innovation [42]. Generalizing our findings to other scientific disciplines should consider differences in publication practices and citation behaviors.

Drawing from our findings, we contextualized them with those documented in prior literature and propose four recommendations for enhancing the translational landscape in patents and repositories of Computer Science papers.

The first recommendation is about encouraging translational work on patents and repositories. Translational research aims to transform research findings into real-world products and services that, in turn, drive progress. However, our study revealed that only

a subset of papers went into patents and into repositories, emphasizing the potential to bridge the gap between theory and practical applications. To enhance collaboration between academia and industry, dedicated translational research events within conferences could facilitate knowledge exchange and collaboration between researchers and industry experts [43]. Additionally, conferences could establish awards for most impact translational research or provide grants for repository development. Similarly, initiatives such as sponsorships for applied research through events like hackathons [44] or competitions [45] at universities can help translate research into tangible products.

The second recommendation is about striking a balance between scientific focus and practical progress. In line with prior research [2], we confirmed a significant time lag between papers and patents, even after accounting for publication delays and the time since a paper's publication. This can be attributed to AI recently experiencing an overwhelming surge in paper production [46], paradoxically creating a bottleneck in translating research into patents and practical applications. Addressing this paradox requires exploring ways to expedite the transition from academic research into patentable solutions.

The third recommendation is about making space for distinct modes of production. It may be that certain topics are best suited for a specific mode of production, whether it is patent creation or academic paper publication. Some subjects may require ample time to evolve, and therefore, they need to remain within academic circles for an extended period. A prime example is neural networks 40 years ago: had this topic not been nurtured over decades, it may not have reached the level of maturity it enjoys today.

The fourth recommendation is about exploring unconventional combinations of prior research. More conventionality predicted patent impact, while less conventionality predicted repository impact. This difference can be explained by the distinct characteristics of these two types of translational impact. Patents are problem-focused, addressing specific practical issues with precision, while repositories have a broader scope, offering adaptable resources for various purposes. Hence, patents and repositories require different levels of conventionality. This offers an opportunity for papers to strike the right balance of (un)conventional knowledge.

6 Conclusion

We present empirical evidence indicating that academic papers exert influence beyond academic spheres. We tracked two types of impact: one on innovation (through patents), and the other on development (through repositories). Despite this substantial impact, our findings reveal an opportunity for greater translational impact, where academia and industry need to bridge the gap. Simultaneously, they highlight that certain academic areas tend to overly concentrate on specific topics, while other areas are underrepresented but hold the potential for significant real-world impact.

Abbreviations

AI, Artificial Intelligence; USPTO, United States Patent and Trademark Office; HCI, Human-Computer Interaction; NLP, Natural Language Processing; CVPR, Conference on Computer Vision and Pattern Recognition; ICML, International Conference on Machine Learning; ACL, Association for Computational Linguistics; CHI, Conference on Human Factors in Computing Systems; KDD, Knowledge Discovery and Data Mining; GCN, Graph Convolutional Network; LDA, Latent Dirichlet Allocation; DOI, Digital Object Identifier; API, Application Programming Interface; ICLR, International Conference on Learning Representations; GROBID, GeneRation Of BIbliographic Data.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00568-4.

Additional file 1. Appendix (PDF 745 kB)

Acknowledgements
This work was done at Nokia Bell Labs. MC was supported by Nokia Bell Labs and the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No. 739578).

Author contributions
FC collected the data and conducted the analysis. FC, AS, MC, and DQ conceived the experiments and wrote the manuscript.

Funding information
Nokia Bell Labs provided support in the form of salaries for authors [FC, AS, MC, DQ], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Data availability
Semantic Scholar full-corpus datasets can be obtained by using the Semantic Scholar Datasets API, https://api.semanticscholar.org/api-docs/datasets. To use the API, users need to request an API key via the request form, https://www.semanticscholar.org/product/api#api-key-form. The patent data are available in Google Patents Public Data, https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-public-data. Links between papers and repositories are available in Papers With Code, https://production-media.paperswithcode.com/about/links-between-papers-and-code.json.gz. To allow for reproducibility, we made our data and code publicly available: https://social-dynamics.net/impact.

Declarations

Ethics approval and consent to participate
This study was approved by Nokia Bell Labs.

Consent for publication
All authors have reviewed and approved the manuscript for publication.

Competing interests
The authors declare no competing interests.

Author details
[1]Sapienza University of Rome, Rome, Italy. [2]Nokia Bell Labs, Cambridge, UK. [3]CYENS Centre of Excellence, Nicosia, Cyprus. [4]Politecnico di Torino, Turin, Italy.

Received: 7 April 2025 Accepted: 1 July 2025 Published online: 16 July 2025

References

1. Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–132
2. Cao H, Lu Y, Deng Y, Mcfarland D, Bernstein MS (2023) Breaking out of the ivory tower: a large-scale analysis of patent citations to HCI research. In: Proceedings of the ACM conference on human factors in computing systems (CHI), pp 1–24
3. Mendoza XPL, Sanchez DSM (2018) A systematic literature review on technology transfer from university to industry. Int J Bus Syst Res 12(2):197–225
4. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási A-L (2018) Science of science. Science 359(6379):eaao0185. https://doi.org/10.1126/science.aao0185
5. Semantic Scholar. https://www.semanticscholar.org/product/api
6. OpenAlex. https://openalex.org/
7. Grobid. https://github.com/kermitt2/grobid
8. Biblio Glutton. https://github.com/kermitt2/biblio-glutton
9. Ahmadpoor M, Jones BF (2017) The dual frontier: patented inventions and prior scientific advance. Science 357(6351):583–587
10. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472
11. Cha M, Haddadi H, Benevenuto F, Gummadi K (2010) Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the international AAAI conference on web and social media (ICWSM), vol 4, pp 10–17
12. Badashian AS, Stroulia E (2016) Measuring user influence in github: the million follower fallacy. In: Proceedings of the international workshop on CrowdSourcing in software engineering, pp 15–21
13. Badashian AS, Esteki A, Gholipour A, Hindle A, Stroulia E (2014) Involvement, contribution and influence in github and stack overflow. In: CASCON, vol 14, pp 19–33
14. Blincoe K, Sheoran J, Goggins S, Petakovic E, Damian D (2016) Understanding the popular users: following, affiliation influence and leadership on Github. Inf Softw Technol 70:30–39
15. Blincoe K, Harrison F, Damian D (2015) Ecosystems in github and a method for ecosystem identification using reference coupling. In: IEEE/ACM working conference on mining software repositories. IEEE, pp 202–211

16. Manjunath A, Li H, Song S, Zhang Z, Liu S, Kahrobai N, Gowda A, Seffens A, Zou J, Kumar I (2021) Comprehensive analysis of 2.4 million patent-to-research citations maps the biomedical innovation and translation landscape. Nat Biotechnol 39(6):678–683
17. Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7):878
18. Jefferson OA, Jaffe A, Ashton D, Warren B, Koellhofer D, Dulleck U, Ballagh A, Moe J, DiCuccio M, Ward K, et al (2018) Mapping the global influence of published research on industry and innovation. Nat Biotechnol 36(1):31–39
19. Tijssen RJ (2001) Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows. Res Policy 30(1):35–54
20. Mariani MS, Medo M, Lafond F (2019) Early identification of important patents: design and validation of citation network metrics. Technol Forecast Soc Change 146:644–654
21. Park M, Leahey E, Funk RJ (2023) Papers and patents are becoming less disruptive over time. Nature 613(7942):138–144
22. Google Patents Public Data. https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-public-data
23. PapersWithCode. https://paperswithcode.com/api/v1/docs/
24. Ni J, Ábrego GH, Constant N, Ma J, Hall KB, Cer D, Yang Y (2021) Sentence-t5: scalable sentence encoders from pre-trained text-to-text models. arXiv preprint. arXiv:2108.08877
25. Marx M, Fuegi A (2022) Reliance on science by inventors: hybrid extraction of in-text patent-to-article citations. J Econ Manag Strategy 31(2):369–392
26. Reliance on Science. https://github.com/mattmarx/reliance_on_science
27. Dudley WN, Wickham R, Coombs N (2016) An introduction to survival statistics: Kaplan-Meier analysis. J Adv Pract Oncol 7(1):91
28. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
29. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D (2009) Reading tea leaves: how humans interpret topic models. Adv Neural Inf Process Syst 22
30. Hannousse A (2021) Searching relevant papers for software engineering secondary studies: semantic scholar coverage and identification role. IET Softw 15(1):126–146
31. Crossref. https://www.crossref.org/
32. Cohan A, Ammar W, Zuylen M, Cady F (2019) Structural scaffolds for citation intent classification in scientific publications. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers). Association for Computational Linguistics, Minneapolis, pp 3586–3596. https://doi.org/10.18653/v1/N19-1361. https://aclanthology.org/N19-1361
33. Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. Science 316(5827):1036–1039
34. The USPTO patent examination research dataset: a window on the process of patent examination. https://www.uspto.gov/sites/default/files/documents/PatEx
35. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25
36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
37. Li D, Tong TW, Xiao Y Is China Emerging as the Global Leader in AI? Harvard Business Review. https://hbr.org/2021/02/is-china-emerging-as-the-global-leader-in-ai
38. Shmueli B, Ray S, Ku L-W (2021) Happy dance, slow clap: using reaction gifs to predict induced affect on Twitter. arXiv preprint. arXiv:2105.09967
39. Song D, Wang Z, Huang Y, Ma L, Zhang T (2023) Deeplens: interactive out-of-distribution data detection in nlp models. In: Proceedings of the ACM conference on human factors in computing systems (CHI), pp 1–17
40. Downer J (2024) Rational accidents: reckoning with catastrophic technologies. MIT Press, Cambridge
41. Burki T (2024) Crossing the frontier: the first global ai safety summit. Lancet Digit Health 6(2):91–92
42. Nagaoka S, Motohashi K, Goto A (2010) Patent statistics as an innovation indicator. In: Handbook of the economics of innovation, vol 2. Elsevier, Amsterdam, pp 1083–1127
43. Buxton B (2008) The long nose of innovation. Insight 11:27
44. Briscoe G (2014) Digital innovation: the hackathon phenomenon
45. Kaggle: Kaggle competitions (2010). https://www.kaggle.com/competitions
46. McKinsey & Company: the state of AI in 2022—and a half decade in review (2022). https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

Publisher's note