Industrial and Corporate Change, Volume 00, Number 00, pp. 1–44
doi:10.1093/icc/dtaf044
Original Article

OXFORD

How small is big enough? Open labeled datasets and the development of deep learning

Daniel Souza[1,*] Aldo Geuna[2,3,4] and Jeff Rodríguez[5]

[1]Department of Management, Economics and Industrial Engineering, Polytechnic University of Milan, Via Raffaele Lambruschini, 4/B, 20156 Milan, Italy. e-mail: danielfernando.desouza@polimi.it, [2]Department of Cultures, Politics and Society, University of Turin, Lungo Dora Siena, 100A, 10153 Turin, Italy. e-mail: aldo.geuna@unito.it, [3]Collegio Carlo Alberto, Piazza Vincenzo Arbarello, 2, 10122 Turin, Italy. e-mail: aldo.geuna@unito.it, [4]Innovation, Equity & The Future of Prosperity Programme, Canadian Institute for Advance Research (CIFAR), MaRS Centre, West Tower 661 University Ave., Suite 505, Toronto, ON M5G 1M1, Canada. e-mail: aldo.geuna@unito.it and [5]OECD, Council and Executive Committee Secretariat, 2 rue André Pascal, 75775 Paris, France. e-mail: jeff.rodriguez@oecd.org

*Main author for correspondence.

We investigate the emergence of Deep Learning as a technoscientific field, emphasizing the role of open labeled datasets. Through qualitative and quantitative analyses, we evaluate the role of datasets like Canadian Institute of Advanced Research - 10 classes (CIFAR-10) in advancing computer vision and object recognition, which are central to the Deep Learning revolution. Our findings highlight CIFAR-10's crucial role and enduring influence on the field, as well as its importance in teaching ML techniques. Results also indicate that dataset characteristics such as size, number of instances, and number of categories, were key factors. Econometric analysis confirms that CIFAR-10, a small-but-sufficiently large open dataset, played a significant and lasting role in technological advancements and had a major function in the development of the early scientific literature as shown by citation metrics.

JEL Classification: O31, O35, H5

1. Introduction

Artificial intelligence (AI) technologies promise to revolutionize the knowledge production process. At the core of one of the most important approaches to the AI revolution are machine learning (ML) algorithms: computer programs that improve performance as they are exposed to an increasing amount of data. An example of disruptive technology based on ML is AlphaFold—an AI algorithm developed by Google's offshoot DeepMind first released in 2018, which solved one of the most challenging problems in the field of biology: the prediction of protein's structures based on amino-acid sequences (Callaway, 2020; Jumper *et al.*, 2021). A more recent example is ChatGPT, a Large Language Model (LLM) developed by OpenAI. It is based on the GPT (Generative Pre-training Transformer) architecture and is trained to generate human-like text. ChatGPT and other LLMs available in the early 2020s have been identified as having impact in diverse areas that go from medicine (Jeblick *et al.*, 2023) to journalism (Pavlik, 2023) and their impact on science is heavily discussed (Stokel-Walker and Van Noorden, 2023).

These breakthroughs, and many others are underpinned by developments in Deep Learning (DL), a subset of ML models that relies on neural networks and requires vast amounts of data to be trained (LeCun *et al.*, 2015). Due to the extremely promising results in wide areas of

application, DL has been regarded as a new method of invention and potentially a general-purpose technology on which the next industrial revolution may be based (Crafts, 2021). Although a growing literature has studied the impact of DL on the knowledge production process (Klinger *et al.*, 2021; Bianchini *et al.*, 2022), little attention has been given to its inception and to the specific role played by Open Labeled Datasets (OLDs).

In this paper, we analyze the emergence of DL as a technoscientific field, that is, a domain in the middle of scientific enquiry and technical problem-solving (Kastenhofer and Molyneux-Hodgson, 2021). More specifically, we examine how OLDs have contributed to the growth and consolidation of DL, focusing on their distinct characteristics. Within this perspective, we regard OLDs as technological artifacts that allow the development of the field. We draw on the literature discussing the emergence of new scientific disciplines to provide a picture of the development of DL as the dominant approach in ML & AI, and the role of OLDs in that process. We perform an analysis of the technological and scientific use of OLDs that includes both qualitative and quantitative elements. We devote particular attention to the role played by Canadian Institute of Advanced Research - 10 classes (CIFAR-10), the most used dataset in the ML literature indexed at the *Papers with Code* website.[1] We carried out a set of semi-structured interviews with relevant actors and we implemented a survey of academics and ML practitioners who have used CIFAR-10 in their work; on the basis of the qualitative evidence we modeled the use OLDs in technological and scientific development proxied by patent (technology) and scholarly (science) citations in the period 2000–2022.

Compute, data, and algorithmic advances are the needed ingredients of the DL revolution (Sevilla *et al.*, 2022; Koch and Peterson, 2024). In early 2010s, increased computing power availability (see the arrival of 2D and 3D GPUs) was in line with the doubling approximately every 6 months of computing requirements by new DL algorithms running on OLDs (Sevilla *et al.*, 2022). The main tenet of this paper is that once the bottleneck of computer power was no longer a major problem, the potential of neural network approaches to AI—theoretically developed over the last fifty years of the 20th century—could be realized and further advanced through the use of OLDs. Given that AI as a field shifted towards an evaluation system based on *benchmarking*—quantification of progress based on predictive accuracy on example datasets (Koch and Peterson, 2024), OLDs became fundamental to develop better algorithms/architectures. Models (algorithms and architectures) were developed to solve specific tasks using specific OLDs; they would not exist without the dataset, as the specific OLD allowed for the development of more refined and accurate models. OLDs that required less computing power, such as CIFAR-10, a small-but-sufficiently-large dataset, enabled the testing and refinement of new model architectures like AlexNet, which succeeded in solving tasks using huge and complex datasets that were previously unattainable with the same computational resources. OLDs should be considered as the necessary testing tools that had to be developed to allow progress in the DL modelling.

The qualitative evidence we put together supports the view that OLDs, and CIFAR-10 in particular, were fundamental for the technological and scientific developments that lead to the DL revolution and still shape the trajectory of the field. We trace the creation of CIFAR-10 to the CIFAR NCAP Summer School in 2008, where the labeling of the dataset was conducted mostly by graduate students over the supervision of Geoffrey Hinton, a prominent scholar in the field, and two of his students, Alex Krizhevsky and Vinod Nair. We also learned through our interviews that CIFAR-10 became a benchmark due to its technical specifications, namely the nature of the images, their size, the number of samples, and categories. The survey confirms the insights of the interviews and highlights that CIFAR-10 is used extensively in the training of computer scientists working with ML. Many researchers not only teach courses using CIFAR-10, but also were themselves exposed to the dataset while following graduate programs. This finding highlights teaching as an important channel through which CIFAR-10 impacted the field of DL.

By examining data from 28,393 conference proceedings and journal publications in the ML literature that utilized OLDs to train models between 2010 and 2022, we assess the technological and scientific relevance of these papers based on their citations in patents and academic literature. Our econometric analysis confirms the significant role of CIFAR-10 in the technological and scientific development of DL. Specifically, we find that papers using CIFAR-10—a small but

[1] See Section 4.2 for details.

sufficiently large dataset—had a substantial early impact on the scientific literature, as evidenced by high academic citation counts, and continue to be relevant today, as shown by their higher patent citation counts. This indicates that the technical characteristics that initially contributed to the dataset's success continue to drive research and technological advancements in DL, particularly in computer vision and image recognition. We compared the CIFAR-10 and ImageNet datasets, demonstrating that CIFAR-10 has been and continues to be significant for technological developments, while ImageNet keeps on playing a prominent role in scientific developments within the DL literature.

The rest of the paper proceeds as follows. In the next section, we present the conceptual framework used followed by the historical and institutional background of DL research and OLDs in Section 3. Section 4 describes the empirical methodology, data collection, the construction of the sample and presents descriptive statistics. Section 5 reports and discusses the results of the analysis. Section 6 concludes the paper.

2. Conceptual framework

Since Kunh's *The structure of scientific revolutions* (Kuhn, 1970), the sociology of science—and more recently the economics of science—has been interested in studying the conditions of emergence of new disciplines or subdisciplines within the scientific endeavor. The most important idea presented by Kuhn is how scientific knowledge does not always grow in a stable and incremental fashion, but it can also go through short periods of big changes, in which new paradigms emerge and consolidate.

In this paper, we explore in particular the question of how OLDs contributed to the process of making DL the dominant paradigm within AI (Schmidhuber, 2015; Kersting, 2018; Chah, 2019), after being dismissed for a long time in favor of symbolic AI (Waldrop, 2019; Weber and Prietl, 2021). To do so, we mostly rely on the theoretical contributions of Frickel and Gross (2005). They argued that there are parallelisms between social movements and what they call Scientific/Intellectual Movements (SIMs). Just as social movements, SIMs involve the pursue of common projects and objectives by a group of people that must rely upon repertoires of collective action to face resistance from others in the scientific or intellectual community. Since SIMs resemble social movements that emerge to challenge some previous paradigm and therefore inevitably face some level of resistance, they also must deal with the problems of collective action: "The emergence of new social forms in science and academe invariably requires some level of spatial, temporal, and social coordination" (Frickel and Gross, 2005).

Following Kuhn (1970), we also consider that SIMs emerge at times of scientific crisis, when research anomalies linked to old paradigms have accumulated beyond a tolerable threshold. However, the contempt towards the dominant paradigm is only a prerequisite and never enough to generate a SIM. For an intellectual movement of that sort to be successful, the leaders must articulate a distinctive research program. Doing so requires certain structural conditions, especially the access to resources, such as employment for the members of the SIM, access to laboratories, academic positions that allow to publish their results, and organizational resources that allow the members of the SIM to come together and create *epistemic cultures*, and discuss repertoires of thought and action that allow them to advance their intellectual agenda.

After the initial conditions are given, SIMs also have the need (like social movements) to recruit new members, to do so a locus of exchange and discussion where novel research is presented to old members and potential new recruits becomes a major condition for the success of the movement. These scenarios of micromobilization can take the form of seminars, conferences, PhD positions, or summer schools. SIMs must find ways to validate itself both internally, building a narrative of its history and identity, and externally, against opponents (Frickel and Gross, 2005).

In the case in hand, in the 1980s and 1990s, there was a group of scientists, in different universities around the United States, Europe, and Canada who were not satisfied with the direction of the research programs in AI, based mostly on symbolic systems. Among them, Geoffrey Hinton, a University of Toronto professor, who was convinced that DL "had to be the future of AI" (Goldman, 2022). He and some of his colleagues—particularly Yan LeCun and Yoshua Bengio—were at the forefront of the DL revolution. Waldrop (2019) describes what happens during this contentious period of the 1980s and early 1990s:

> Today's deep-learning revolution has its roots in the "brain wars" of the 1980s when advocates of two different approaches to AI were talking right past each other. On one side was an approach—now called "good old-fashioned AI"—that had dominated the field since the 1950s. Also known as symbolic AI, it used mathematical symbols to represent objects and the relationship between objects. [. . .] But by the 1980s, it was also becoming clear that symbolic AI was impressively bad at dealing with the fluidity of symbols, concepts, and reasoning in real life. In response to these shortcomings, rebel researchers began advocating for artificial neural networks, or connectionist AI, the precursors of today's deep-learning systems (Waldrop, 2019, p. 1075).

The Canadian Institute for Advanced Research (CIFAR), a research funding organization based in Toronto that finances basic research with a high-risk, high-reward philosophy was, since its foundation in the 1980s, consistently interested in the advancement of AI and was at the forefront of the upsurge of ML technologies (Chah, 2019). CIFAR became the institutional setting on which those "rebel researchers" were able to join forces and form their own epistemic culture. CIFAR provided access to symbolic resources in the form of positions—like fellowships—for some of them, but also material resources in the form of funding (not in a significant amount) for conferences, meetings, and summer schools, all of them part of the micromobilization scenarios needed to recruit new members, discuss novel ideas, and in general advance their agenda.

2.1 Open science

In a series of works in the early 2000s, Paul A. David elaborated on the concept of *open science* contrasting it with the increasing reliance on Intellectual Property Rights (IPRs) in the production of science (David, 2000 and 2003). Open science in its original conception, takes a descriptive sense, referring to a new paradigm born:

> with Renaissance mathematics, the cultural ethos and social organization of western European scientific activities during the late sixteenth and seventeenth centuries [. . .] –departing from the previously dominant regime of secrecy in the pursuit of 'Nature's secrets' (David, 1998, p. 15).

This new paradigm shaped the organization of the scientific endeavor in the West, including the imperatives of public disclosure of discoveries, and the methods that lead to those discoveries. This openness was supported by a public (open) system of Universities and research communities, and a series of norms, including communalism, universalism, desinterestedness, originality, and skepticism (Merton, 1973), which created a reward system based on collegiate reputation that was achieved by validated claims to priority in discovery or invention (David, 2003).

Since its inception, the concept has positioned itself as opposed to a "closed" science based on IPRs like patents and copyrights, which jeopardize the traditional ethos of open science. Scholars like Dasgupta and David (1994) have warned about the social and economic problems that might arise from the enclosure of scientific knowledge within the framework of IPRs. Among other potential hazards, they mention a suboptimal level of production of basic science, which has the greatest spillovers; and scientists getting more and more engaged in duplicitous work, unable to access a big part of the stock of codified knowledge in the form of patents created by a culture of "intellectual capitalism" (David, 2004).

More recently, the advent of digital technologies, and in particular the internet, has given rise to a slightly different conceptualization of open science that lies "between the age-old tradition of openness in science and the tools of information and communications technologies (ICTs) that have reshaped the scientific enterprise" (OECD, 2015). In this conception, open science is (loosely) defined as "efforts by researchers, governments, research funding agencies, or the scientific community itself to make the primary outputs of publicly funded research results—publications and the research data—publicly accessible in digital format with no or minimal restriction as a means for accelerating research; these efforts are in the interest of enhancing transparency and collaboration, and fostering innovation" (OECD, 2015). In this conception, open science is part of an "open ecosystem" that encompasses open access journals, open data, open software, open collaboration, open peer review, among others.

One element of the open science ecosystem is particularly relevant for this work is open data. The European Commission provides a definition, stating that "Open Data is data that is made

available by (public) organisations, businesses, and individuals for anyone to access, use, and share" (European Commission, 2018). Access to data can have many advantages or purposes. Data (e.g. from public records) can be used for original research; for reproducing and validating (or not) existing knowledge; or to explore new research avenues.

Certainly data have become relevant for many areas of scientific inquiry; but for DL in particular, data are a *conditio sine qua non* for its very existence, since the neural networks on which it is built rely on the availability of large amounts of data. Open data, built collaboratively, clearly labeled, and free to access on the Internet was key to the emergence and eventual dominance of DL within AI (Martens, 2018).

2.2 The GPU revolution

To achieve the status of a dominant paradigm within the ML literature, DL had to overcome a series of systemic bottlenecks that impeded its development. Although the theoretical basis for AI, based on ML algorithms and convolutional neural networks, was established in the 1980s, the first significant bottleneck from the 1990s onwards was the availability of large amounts of training data necessary to "feed" the DL models.

Besides the availability of training data, the development of DL depended also on the increase of computing power. Because of the enormous amounts of data to be processed and the increasing complexity of the algorithms used to analyze that data, computing capacity became a bottleneck for the development of DL until the second decade of the twenty-first century. Despite the excitement with neural networks in the 1980s and 1990s "computers were not powerful enough to allow this approach to work on anything but small, almost toy-sized problems" (Dean, 2020).

The paradigm of general-purpose computing on GPU cards, which were originally used for gaming, "because of GPU cards' high-floating point performance relative to CPUs, started to allow neural networks to show interesting results on difficult problems of real consequence" (Dean, 2020). In particular, from the mid-1990s the performance of GPUs increased significantly with 2D and 3D acceleration on the same unit. The coming on the market of Nvidia GeForce 256 in 1999 is usually considered the turning point of the industry. The consequence of these technological advances was that "computers finally started to become powerful enough to train large neural networks on realistic, real-world problems" (Dean, 2020).

By 2009 when CIFAR-10 was launched, the technological conditions for its use and exploitation were mature. CIFAR-10 became a dataset that could be manipulated on personal computers (see Table 1 below for a comparison of computing requirements of mostly used OLDs), and used as a toy-dataset to train and improve algorithms that could later be used on more complex datasets such as ImageNet.

2.3 AI as a technoscience

Different from other intellectual movements, AI in general and DL in particular can be better understood as a technoscience, located in the intersection between traditional scientific research and technological applications (Raimbault and Joly, 2021). Different from pure sciences, the quest of technoscience is not only motivated by a search for new knowledge in an abstract way, but also to the solution of practical problems. In fact, "technoscience is 'face to face' with the things. It is less interested in what they are or what regular behaviors they are naturally disposed to exhibit, and more interested in what they can become or what they might offer" (Bensaude-Vincent *et al.*, 2011). Usually, in the policy arena, technosciences are often referred to as Pasteur Sciences (Stokes, 2011).

In the case of DL, the practical applications go from medical image analysis, language translation, object detection for autonomous vehicles, content filter, and many others (Cf. Bengio *et al.*, 2021). It is no surprise, then, that technosciences develop strong links with industry, as shown by the fact that most of the academics that ignited the DL revolution ended up joining the industry (Geoffrey Hinton in Google; Yan LeCun in Facebook; and Yoshua Bengio in his own venture, Element AI).

The theoretical implications of considering AI as a technoscience and DL as a paradigm within it, mean a deviation from a traditional analysis of a scientific discipline. For example, even if traditional criteria, like the priority of discovery (Merton, 1957) still apply, it does in a different way. More than publications in academic journals, breakthroughs are shown

Table 1. Computational requirements

Dataset	Description	Instances	Primary task	Year Introduced	Creator Affiliation	Current Best Model Performance	Hardware Burden	Estimated Time on Super-computer	Estimated Time on Laptop
COCO	Complex everyday scenes of common objects in their natural context	2,500,000	Object recognition	2015	Microsoft	Models today only reach error rates up to **38.7%**	Target error rate of **10%** requires estimated 10^{31} flops	3.17×10^{5} years	6.34×10^{9} years
ImageNet	Labeled object image database	14,197,122	Object recognition, classification	2009	Princeton University	Best model *OmniVec* reached error rate of **8%**	Reaching human-level error rate of **5%** requires 10^{26} flops	3.17 years	6.34×10^{4} years
CIFAR-10 Dataset	Many small, low-resolution, images of 10 classes of objects	60,000	Classification	2009	University of Toronto	Most models can reach **99%+** accuracy	Reaching human-level error rate of **6%** requires 10^{21} flops	$\sim$ 16 minutes	230 days
MNIST database	Database of handwritten digits	70,000	Classification	1994	AT&T Bell Labs	Most models can reach **99%+** accuracy	To train a similar model 10^{12} flops	10^{-6} seconds	$<$ 10 seconds

This table shows the computational requirements to train State-Of-The-Art (SOTA) ML models using the 4 most common dataset in our sample. See methodological appendix for the construction of the table. Source: Shermatov (2024).

through *competitions*, in which the new techniques (in this case, the algorithms) are tested against a certain benchmark to validate the real-world performance of the *discovery*. In other words, the understanding of causal mechanisms in the aim of proving or disproving a certain theoretical perspective becomes secondary, while practical (technological) results are of the utmost importance.

This is very clear in the case of DL. Bengio *et al.* (2021) highlight that "DL scored a dramatic victory in the 2012 ImageNet competition, almost halving the error rate for recognizing a thousand different classes of object in natural images". Other authors like LeCun *et al.* (2015) and Schmidhuber (2015) also underscore the performance of the algorithms in those competitions as the most important milestones in the paradigm shift. 2012 became known as the year of the "DL revolution". The practical performance then becomes more relevant than the actual understanding of the mechanism that drives those results. In fact, "once a DL system has been trained, it's not always clear how it's making its decisions" (Waldrop, 2019). Articles in scientific journals play a role in the development of this new technoscience but other forms of knowledge diffusion and creation of reputation such as conference presentations, conference proceedings, and patents are of similar or higher importance (Fortnow, 2009; Meyer *et al.*, 2009; Franceschet, 2010). For example, in the full sample of 37,242 articles identified in this paper as composing the relevant literature in DL for computer vision and image recognition, around 55% were conference proceedings. Science and technology are interlinked, and publications and proceedings are cited more frequently and faster in patents protecting downstream technological development. To try to capture developments in the field we must therefore use both patents and publications because the latter would only provide a limited representation of the evolution of the science.

3. Institutional background

3.1 Winning the *brain wars*: The emergence of DL as a dominant paradigm within AI

DL is a subfield of ML that is inspired by the structure and function of the brain's neural networks. It involves training artificial neural networks, which are composed of layers of interconnected nodes or "neurons" to learn from large amounts of data. These networks can be used to perform a wide variety of tasks, such as image and speech recognition, natural language processing, and decision making. DL is often used in combination with other techniques, such as reinforcement learning, to solve complex problems (LeCun *et al.*, 2015).

DL is a subset of ML, which in turn is defined as "concerned with the question of how to construct computer programs that automatically improve with experience" (Mitchell, 1997). ML, on the other hand, is one of the most important approaches of AI. AI, ML, and DL are interrelated but differ from each other. Chah (2019) makes the following distinction:

> Although the three terms—AI, ML and DL—are intricately linked, nuanced differences in their specific definitions can make the difference between whether the term is used precisely or whether the actual operations on the ground are obfuscated. The dynamic definition of AI affects what state-of-the-art advances are considered as AI for a particular time and place. ML is primarily concerned with training machines to learn from data, following closely the original definition by Arthur Samuel in 1959. To implement the ever-changing state-of-the-art techniques that exhibit AI capabilities, DL is one of the most popular sets of ML techniques in use today (Chah, 2019, p. 3).

The concept of DL was coined in 2006 by Geoffrey Hinton and his colleagues (LeCun *et al.*, 2015; Chah, 2019). However, the concepts on which this technology is based, started to develop long before with the work on artificial neural networks in the 1940s (Schmidhuber, 2015; Chah, 2019) that evolved in the 1980s into the convolutional neural network (Fukushima, 1980).

Despite being a growing field, DL was at the peripheries of AI for many decades. According to Yan LeCun, one of the main proponents and intellectual architects of the DL revolution "In the late 1990s, neural nets and backpropagation were largely forsaken by the machine-learning community and ignored by the computer-vision and speech-recognition communities"

(LeCun *et al.*, 2015). By 2006, a paper by Hinton *et al.* (2006) reignited the interest, by showing the possibility of using DL to achieve state-of-the-art results (1.25 percent error rate) in recognizing handwritten digits. In 2012, a DL algorithm developed by Krizhevsky *et al.* (2017) won the ImageNet classification competition, one of the most challenging image recognition databases at the time. AlexNet, the winning algorithm, became a milestone that positioned DL as the dominant paradigm within ML and consequently within AI.

AlexNet was developed by Alex Krizhevsky in collaboration with Ilya Sutskever and Geoffrey Hinton. Both Krizhevsky and Hinton were also behind the creation of CIFAR-10, which became the basis for the development of the AlexNet algorithm (Fergus, Interviewee 2; Bengio, Interviewee 5).

3.2 The development of Open Labeled Datasets and CIFAR-10

CIFAR's rationale for support was to fund high-risk basic research by providing resources designed to foster networking and to create institutional space for alternative research approaches. In 2004, CIFAR supported a diverse group of unorthodox scientists led by Geoffrey Hinton into pursuing an ambitious program in AI called Neural Computation and Adaptive Perception Program (NCAP) (Silverman, Interviewee 7; Brownell, 2016).

On his account of the beginning of the NCAP program, Prof. Silverman highlights how this group of people were full of new ideas, but did not have the institutional spaces to present and discuss them:

> But when I'm trying to create a scenario, where, as they spoke, [...] that they sort of, had come together as a group, informally because they didn't have anybody to talk to in their own departments. There they were, they had their own disciplines, they made it into, they had faculty appointments, they were achieving in their own departments, but basically, their interests had taken them in a much broader, different way [to] understand how the brain processes information, not really staying in a single lane, if you know what I mean. And so that, that, that was that resonated with me." He continues saying that they thought "We're smart, and nobody wants us. Because we're trying to work on this really tough problem. (Silverman, Interviewee 7).

CIFAR became then the institutional space that provided with basic resources for that group of people to come together and start exploring their common interests. Geoffrey Hinton was joined in his research effort by Yoshua Bengio, and Yann LeCun. Their work became a seminal piece in the paradigm shift that saw DL become the dominant approach in AI. "Their work together led to a number of advances, including a breakthrough AI technique called DL, which is now integral to computer vision, speech recognition, natural language processing, and robotics" (Farrow, 2019). Because of this work, they received the A.M. Turing Award, considered as the "Nobel Prize of Computing".

Open Labeled Datasets. Before 2009, the two main datasets used for computer vision and object recognition tasks were CALTECH-101 and MNIST (Modified National Institute of Standards and Technology database). CALTECH-101 is a dataset that contains pictures of objects belonging to 101 categories. It contains about 40–800 images per category, with most categories containing about 50 images. It was collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc'Aurelio Ranzato (Li *et al.*, 2022). MNIST was one of the first annotated datasets used in ML models and consists of large collection of images of handwritten digits taken from the CENSUS bureau, it contains 60,000 black and white training images and 10,000 testing images, it was released by AT&T Bell Labs in 1998 (Goltsev *et al.*, 2004). The CIFAR team worked mostly with MNIST.

In 2006, Rob Fergus, Antonio Torralba, and William T. Freeman released the "80 million tiny images", a new dataset that could overcome some of the limitations of the existing ones. They automatically collected low-resolution images from different search engines (Altavista, Ask, Flickr, Cydral, Google, Picsearch, and Webshot) and loosely labeled with one of the 53,464 non-abstract nouns in English, as listed in the Wordnet lexical database (Torralba *et al.*, 2008). However, in the original paper in which he introduces the CIFAR databases, Krizhevsky (2009) mentions that he is trying to solve

> A [. . .] problematic aspect of the tiny images dataset is that there are no reliable class labels which makes it hard to use for object recognition experiments. We created two sets of reliable labels. [. . .]. Using these labels, we show that object recognition is significantly improved by pre-training a layer of features on a large set of unlabeled tiny images. (Krizhevsky, 2009, p. 1)

In 2008, Geoffrey Hinton, along with two of his students, Vinod Nair and Alex Krizhevsky, had the idea to manually label a subset of the "80 million tiny images" (Fergus, Interviewee 2), to address one of the problems they encountered when using this large dataset for unsupervised training. To label the images, they took advantage of the NCAP Summer School that took place in August 2008. The students that participated occupied some of the time labeling the images according to a protocol written by Alex Krizhevsky and Rob Fergus (Fergus, Interviewee 2). Rob Fergus account of the process is that

> the data [from the 80 million tiny images] need[ed] to be manually cleaned in order to make a sort of good supervised training dataset. And Geoff wants you to do this. And so he organized the CIFAR summer school, he got all the summer school students sitting down. So how did it work? So I think Alex Krizhevsky and I wrote a labelling routine to actually, you know, have labelling interface where all the students would sit down, and we will go through the images, cleaning them up, and we decided that, Geoff decided he was going to pick, you know, these 10 super categories, and then each one of which had subcategories that form the CIFAR-100. (Fergus, Interviewee 2).

In that way, the CIFAR team created the CIFAR-10 dataset, which consists of 60,000 32x32 color images in 10 classes, with 6000 images per class; and the CIFAR-100 is just like the CIFAR-10, except it has 100 classes containing 600 images each. The datasets overcame some of the problems encountered in older open datasets, while keeping an architecture similar to that of MNIST. These two datasets were subsequently used to train computer vision algorithms through a procedure called *supervised learning*. LeCun *et al.* (2015) explain supervised learning as follows:

> Imagine that we want to build a system that can classify images as containing, say, a house, a car, a person or a pet. We first collect a large data set of images of houses, cars, people and pets, each labeled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the output scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters, often called weights, are real numbers that can be seen as 'knobs' that define the input–output function of the machine. In a typical deep-learning system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labeled examples with which to train the machine (LeCun *et al.*, 2015, p. 436).

It is clear from this definition that supervised learning requires vast amounts of data and very well labeled, so that the machine can be trained with this clean set of images and thus learn how to recognize objects of the same classes. By 2009, there were no datasets that combined a large number of images, a rigorous labelling process, and well-constructed categories that made it easy to manipulate. CIFAR-10 and CIFAR-100 became quickly benchmarks for new algorithms of computer vision using DL. However, it was CIFAR-10 that had the most impact. According to experts' accounts (Fergus, Interviewee 2) beyond the reliability, the size of this dataset and its simplicity were key for its success. In fact, it was light enough so that it was easy to manipulate and work with, specially to train algorithms that require a lot of computer power, but had enough data to properly train a neural network. As Fergus noted, the fact that "every student can run CIFAR-10 on their laptop" contributed to its widespread adoption.

CIFAR databases were made available for free on the web from the University of Toronto, very much in line with the Open Data paradigm mentioned above. Along with the other characteristics, the easy availability became one the distinctive characteristics and main advantages of the CIFAR datasets.

During the same period, other research teams were developing similar image databases. One notable example is ImageNet, created by Fei-Fei Li (currently at Stanford University, formerly at the University of Illinois Urbana-Champaign) and Christiane Fellbaum (Princeton University), and introduced in 2009. ImageNet includes a large collection of labeled object images and rapidly became a benchmark for state-of-the-art computer vision algorithms. The dataset was annotated using Amazon Mechanical Turk (MTurk).[2] For a more detailed description of ImageNet and other similar datasets, see Section 4.2 and Table C1.

In the following sections, we will evaluate whether OLDs, particularly CIFAR-10, have influenced the development and evolution of DL as a technoscience. Specifically, we will explore whether the unique characteristics of these open datasets contributed to the development of the foundational convolutional neural network (CNN) architectures that sparked the DL revolution.

4. Methods and data

4.1 Method

For the empirical analysis, we use a mixed-methods approach (see Appendix A for details). The initial step was to conduct semi-structured interviews with relevant actors, including prominent academics working on the field of AI and DL, as well as CIFAR personnel linked directly or indirectly to the creation of CIFAR-10. Two kinds of interviews were conducted: general interviews with academics working on AI, not necessarily related to CIFAR datasets, with the aim of getting an understanding of the field and some general features that practitioners might look for in a training dataset; and more specific interviews with strategic individuals who were directly or indirectly related to the development of the CIFAR datasets. In total, we conducted seven interviews, out of which two were with field experts not linked to CIFAR; and five with persons linked to CIFAR.

Second, we surveyed researchers and practitioners who referred to CIFAR datasets in their articles. The survey aimed to validate the information obtained from the interviews and develop a broader assessment of the impact of CIFAR-10 on DL and Computer Vision. The sample population included corresponding authors from the subset of papers referencing CIFAR datasets out of 6,060 paper, we were able to retrieve 3,033 valid emails (see Qualitative Methodology in Appendix A for the response analysis and survey questions). We were able to collect 295 answers to the survey, which corresponds to a response rate of 9.4%. The response analysis indicates that our sample is representative for most variables available for the population (total and valid emails).

Finally, we concentrated on publications in the ML literature that utilized OLDs for model training between 2010 and 2022, following the release of the CIFAR and ImageNet datasets. We conducted an econometric analysis aimed at examining the relationship between the use of specific OLDs and receiving citations from patent and scientific publications. Our method involved comparing publications that referenced CIFAR-10 and ImageNet with those that did not but used one of the other similar labeled datasets, while controlling for various confounding factors. Specifically, we estimate regressions of the following models:

$$\begin{aligned}\mathbb{E}[Citations_{jst}] = \exp(&\beta_1 \text{CIFAR-10 (only)}_{jst} + \beta_2 \text{CIFAR-10 (others)}_{jst} \\ &+ \beta_3 X_{jst} + \alpha_j + \delta_s + \gamma_t + \varepsilon_{jst})\end{aligned} \tag{1}$$

$$\begin{aligned}\mathbb{E}[Citations_{jst}] = \exp(&\beta_1 \text{CIFAR-10 (only)}_{jst} + \beta_2 \text{CIFAR-10 (others)}_{jst} + \beta_3 \text{ImageNet}_{jst} \\ &+ \beta_4 X_{jst} + \alpha_j + \delta_s + \gamma_t + \varepsilon_{jst})\end{aligned} \tag{2}$$

For each focal paper published of type j (scholarly journal or conference proceeding), in a scientific area s and year t, we measure its outcome using different metrics that capture their technological and scientific citation impact. Our main explanatory variables are binary indicators

[2] MTurk is a crowdsourcing platform provided by Amazon Web Services that connects businesses and researchers with a global pool of remote workers. It is designed to handle tasks that are difficult for machines but relatively easy for humans, such as labeling datasets.

that assume value 1 if the paper mentioned only CIFAR-10, CIFAR-10 and other datasets or ImageNet. Our main dependent variables which we use as measure of technological and scientific relevancy are the total number of patents citing the articles and the total number of scientific citations.

To ensure a fair comparison of our articles, we develop an empirical design that allows us to compare ML/DL articles that are similar in observable characteristics but differ in the datasets used for model training. Thus, we incorporate a set of control variables, X_{jst}, which describe various characteristics of the focal papers and are related to citation impact. These controls include the number of authors, the number of references, the presence of international collaboration, and the share of authors affiliated with companies, as these factors may influence both the use of CIFAR-10/ImageNet and citation impact.

Additionally, we include as control variables observable characteristics of the labeled datasets used in the papers, such as the number of OLDs mentioned, the number of modalities (i.e. different types of data beyond images, such as text and audio), and the number of ML prediction tasks performed with these datasets.[3] These dataset characteristics serve as proxies for the types of DL models being developed and refined, helping us to partially control for factors that could simultaneously affect both citation counts and the use of the primary OLDs under investigation.

We then estimate a fixed-effects Poisson model, including the independent and control variables discussed above and a set of fixed effects, which includes type of publication α_j, scientific fields δ_s, and calendar year γ_t to control for time-invariant features that may also explain citation impact. In the robustness checks we use Negative Binomial model and different variable operationalizations and sample definitions to test the consistence with. We use the same setting for both technology and scientific citation impacts.

We performed a split sample analysis focusing on two periods 2010–2014 and 2015–2022. We identified 2014 a year of structural change in the motivations to use CIFAR-10 because it was the year where state-of-the-art DL models consistently surpassed human-level accuracy in image classification tasks.[4] Thus, from 2015 onward CIFAR-10 was essentially a "solved problem": prediction accuracy and error rates of trained models were comparable to humans-levels. We perform the split sample analysis for both models 1 and 2: while the former allows us to estimate how many citations papers using CIFAR-10 are expected to receive compared to papers that do not (conditional on observable paper characteristics), the latter model enables us to assess the expected citations for papers using CIFAR-10 or ImageNet in comparison to others. By comparing the coefficients of the CIFAR-10 and ImageNet variables, we can evaluate how papers utilizing datasets of different dimensions and complexity perform in terms of citations.

4.2 Data

We constructed a novel and unique dataset that includes both detailed bibliometric information on publications and patents in the ML literature on image recognition and object classification and the OLDs used by them to train DL models. Our data collection process involves identifying OLDs similar to CIFAR-10, the most used dataset in the ML literature indexed at the *Papers with Code* website.[5] Papers with Code is a platform launched by ML practitioners in July 2018 to share open resources associated with AI development, with a focus on ML (Martínez-Plumed *et al.*, 2021). Currently, Papers with Code comprises approximately 135,000 research papers, encompassing over 11,000 benchmarks that address 5000 distinct tasks.

First, we identified all the tasks mentioned in Papers with Code that involved models trained with CIFAR-10. These tasks refer to different types of predictions or inferences made using models trained on specific data. For example, image datasets can be used to train ML models to solve tasks such as image classification, object detection, and anomaly detection. See Table B1 in Appendix B

[3] Further considerations regarding tasks are provided in Section 4.2. As the field advances, datasets are increasingly applied to new tasks. Nonetheless, we use the number of unique tasks identified for each dataset up to July 2023 as a proxy for the breadth of application of a specific OLD in these fields.

[4] The human error rate on CIFAR-10 is estimated to be around 6%. Working papers originally published at the end of 2014 achieved an error rate of around 3–4% (Graham, 2015).

[5] The Papers with Code platform offers access to all its contents under the CC BY-SA licence, which can be downloaded from the website paperswithcode.com/about. In the context of this study, we obtained the data to perform our analysis on July 17, 2023.

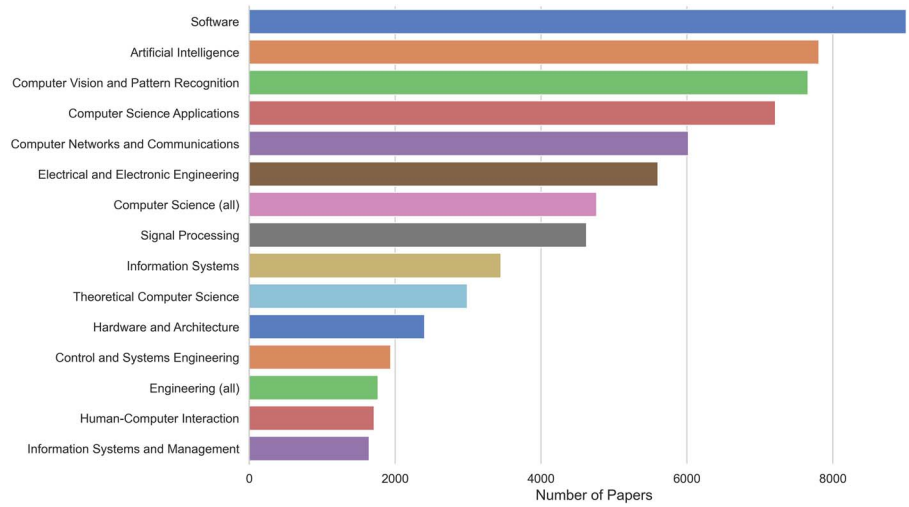


Figure 1. Distribution of Publications by Subject Area. *Note*: This figure displays the distribution of ASJC codes assigned by Scopus to each paper based on its journal, conference, or other publication venue. Note that papers may be assigned multiple ASJC codes.

for a detailed list of these tasks related to CIFAR-10. Our analysis identified 46 tasks related to CIFAR-10 that have been utilized in developing ML models. We then used these tasks to identify other datasets used to train models that handle at least one task similar to those involving CIFAR-10. For information on the most frequently used datasets, see Table C1 in Appendix B.

The Papers with Code platform aggregates ML research papers that are openly accessible and accompanied by source code, mostly sourced from the open access online repositories like *arXiv*. To obtain better bibliographic and citation coverage, we collected scientific publications using a list of annotated datasets of interest on *Scopus*, Elsevier's citation database. Specifically, we used the Scopus database[6] to find any publication that mention a datasets in our list in their titles, abstracts, or keywords. As annotated datasets frequently have variations or subsets tailored to specific tasks, we searched for the full names, shortened names, and variants of each dataset.This approach allowed us to identify 37,242 Scopus indexed scientific publications that mention a total of 264 unique labeled datasets.[7]

Using Scopus' All Science Journal Classification (ASJC) codes, we identified the scientific fields of the publications that frequently use the datasets in our study, as shown in Figure 1. Unsurprisingly, most of the papers in our sample are published in journals from the fields "Software," "Artificial Intelligence," "Computer Vision and Pattern Recognition," and "Computer Science Applications". Interestingly, there is also significant representation from "Electrical and Electronic Engineering," "Hardware and Architecture," and "Control and Systems Engineering," suggesting that these datasets have technological applications beyond strictly computer science disciplines.

To understand the technological developments linked to papers using OLDs, we complement the Papers With Code and Scopus data with patent-publication citation links from the *Reliance on Science* dataset (Marx and Fuegi, 2020, 2022). We gather all the front-page and in-text citations of patents granted worldwide that reference scientific papers in our sample. Since our focus is on technological developments rather than intellectual property concerns, we aggregate these patents into patent families using data from the EPO-PATSTAT database (Autumn 2023 version). We

[6] We utilized pybliometrics, a Python package for accessing the Scopus API. See Rose and Kitchin (2019) for further details on the package.

[7] In the remainder of this work, we will consider publications that *mention a dataset in the title, abstract, or keywords* and *use a dataset to train ML models* as equivalent. We acknowledge that this approach has limitations, as authors might not always mention the dataset used to train their models in these sections, or may mention only some of them. We explored using backward citations to introductory papers, but this method proved less precise and biased. Despite these limitations, we believe that referencing datasets in the title, abstract, or keywords provides a strong indication of the dataset's importance in the publication and is the most reliable way to identify relevant papers in this literature.

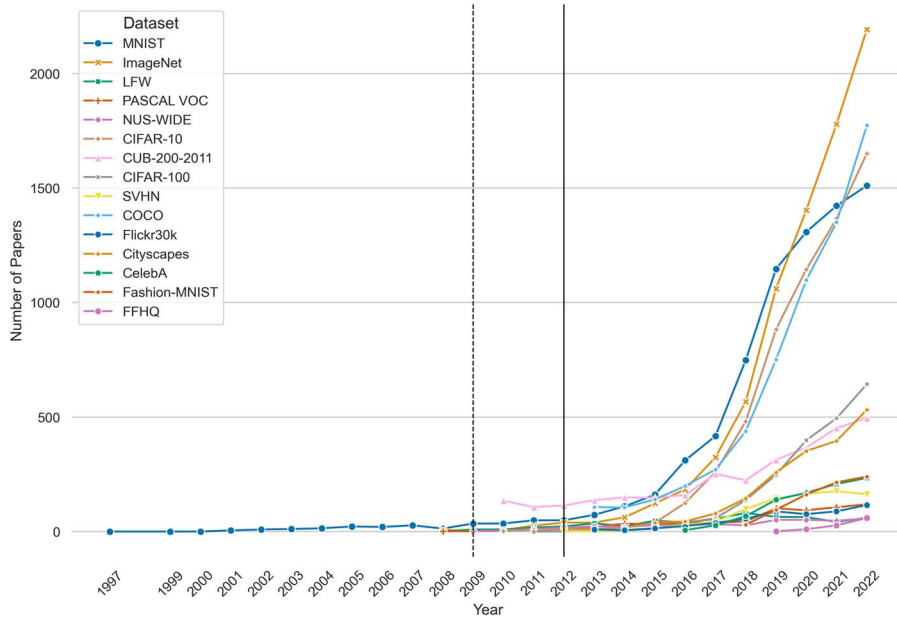


Figure 2. The Rise of Annotated Image Datasets. *Note*: This figure illustrates the annual growth in the number of publications referencing the 15 most commonly used annotated image datasets. The vertical dashed line denotes 2009, the year ImageNet and CIFAR-10 were introduced, while the solid horizontal line marks 2012, the year of the DL revolution.

identified 31,170 patents families citing 14,435 papers from our sample of journal articles and conference proceedings, either on the front-page, in-text, or both.[8]

Considering the relative distribution of the datasets used by publications in our sample, Figure 2 illustrates the number of yearly publications citing the 15 most common datasets in computer vision and image recognition literature using ML, accounting for 81.67% of papers. The significant increase, particularly after 2012—the year of the DL revolution—was primarily driven by papers citing ImageNet, MNIST, COCO, and CIFAR-10,[9] which account for 62.71% of the total sample. CIFAR-10 and ImageNet represent 33.23% of the publications in our sample and show similar trends: both were introduced in 2009 by young scholars who believed in the potential of labeled datasets to advance DL. These datasets consist of natural images and have been used for various tasks such as image classification, object recognition, and image generation.[10] They are the first two most used databases in Papers with Code. The main difference between them is that CIFAR-10 is much smaller, with 60,000 images and 10 categories, compared to ImageNet's 14,197,122 images and over 20,000 categories. Additionally, ImageNet was central to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) from 2010 to 2017, which incentivized the development of ML models using this dataset.

Regarding the other two datasets, MNIST is a dataset of handwritten digits introduced in 1998. It comprises 60,000 training examples and 10,000 test examples, with digits that have been size-normalized and centered in fixed-size images. COCO, introduced in 2014 by a Microsoft group, contains images of complex everyday scenes with common objects in their natural context. It includes 91 object categories, 82 of which have more than 5000 labeled instances, totaling 2,500,000 labeled instances in 328,000 images. Unlike ImageNet, COCO has fewer categories but more instances per category, and it is used for tasks such as detection, segmentation, and captioning.

[8] We have decided to include in-text citations as well, because we believe that non-patent literature cited only on the front page would not adequately cover less conventional scientific publications, such as conference proceedings, data introductory papers, and other sources likely to be found throughout the full patent text.

[9] For more information on these datasets, visit their official sites: ImageNet, MNIST, COCO, and CIFAR-10.

[10] According to our analysis using the Papers with Code platform, CIFAR-10 has been utilized in 46 unique tasks and ImageNet in 64 tasks, with 24 tasks overlapping between the two (52.17% of CIFAR-10 tasks). See Table B1 for details.

Table 1 provides an overview of the computing capacity required for the four most commonly used datasets. Using the best supercomputer in 2024 and a typical research laptop as benchmarks, and referencing state-of-the-art algorithms that outperform humans on CIFAR-10 and MNIST, it is clear that MNIST is now too simple for complex tasks, while COCO remains too challenging to "solve" fully. Therefore, we believe ImageNet is the best candidate for comparison with CIFAR-10. It is important to note that CIFAR-10 is both sufficiently complex and manageable in size. For instance, training a model to achieve human-level accuracy of 94% on CIFAR-10 would take an average research laptop about 10 seconds (Jordan, 2024).

To analyze the citation patterns described in Section 4.1, we consider only publications between 2010 and 2022, reducing our sample to 36,859 publications.[11] We chose 2010 as the starting year because it marks the introduction of the two most popular OLDs to the ML community: CIFAR-10 and ImageNet. We further restrict our sample to conference proceedings and journal articles, as these are the types of publications where we expect to see ML models trained using OLDs, excluding review papers and data introduction papers. This restriction leaves us with a sample of 35,705.

Since we want to compare similar articles, we removed those lacking fundamental bibliometric information used as control variables.[12] Additionally, not all publications indexed by Scopus can be found in Reliance on Science and vice versa, due to duplicated IDs and other issues. Thus, to ensure reliable information about patent citations, we drop from our sample 4944 conference proceedings and journal articles for which we cannot confirm the number of patent citations received. This leaves us with 28,416 papers and proceedings.[13]

Among these, we identified 23 papers (top 0.1% in the citation distribution of the full sample and 0.08% in the restricted sample) as outliers, which received citations orders of magnitude higher than the others.[14] These publications represent breakthroughs so significant that they are not comparable with the average paper in the field. To ensure the comparability, robustness, and stability of our estimations, we excluded these outliers from our sample.

Our final data sample for the econometric analysis includes 28,393 journal articles and conference proceedings, as well as 252 labeled datasets. Table 2 provides descriptive statistics for our regression sample. Approximately 15.4% of the sample mentioned the CIFAR-10 dataset in the 10 years following its release, while 19.9% cited the ImageNet dataset. However, only 4.1% mentioned CIFAR-10 exclusively, indicating that CIFAR-10 is often used alongside other datasets. On average, each paper mentions only 1.3 datasets, with the third quartile being 1 dataset. The average number of authors per paper is 4.32, and 24.5% of the papers include at least one international collaboration (i.e. authors from multiple countries). Private companies are also represented as well, with 4% of authors affiliated with companies (9.4% of the papers have one company affiliated author). Focal papers have an average of 36 backward citations and 16 forward citations, with considerable variation. The average number of different modalities used is 1.25, indicating that most papers focus solely on images. Finally, the number of unique tasks overlapping with CIFAR-10 tasks across all datasets used in the focal papers is approximately 16, representing 34.5% of similar tasks. In summary, this sample consists of papers using the most common labeled datasets in computer vision, with tasks closely related to those central to the DL revolution.

5. Findings

In this section, we explore the role played by OLDs and in particular of CIFAR-10 on the development of DL using our three sources of information: the qualitative interviews, the survey,

[11] Few papers used large labeled datasets before the DL revolution in 2012; thus, we lose very few papers in this step.

[12] We have 28 publications missing author information, 1761 missing references, 1074 missing affiliation information, and four missing subject area information. We also exclude two papers that came from dataset that are described in Papers With Code, but do not have any paper indexed to it in the platform.

[13] We perform robustness checks using the sample without excluding those publications in Appendix D.

[14] The most cited paper in our sample is "Deep Residual Learning for Image Recognition," published in the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition by a Microsoft group. This paper introduced the Residual Networks (ResNet) architecture, a key component in modern DL models (e.g. Transformers, AlphaGo Zero), and has received 95,139 citations. It was the most cited paper globally for five consecutive years, according to Google Scholar (source: NatureIndex).

Table 2. Summary statistics

Statistic	*N*	Mean	St. dev.	Min	Pctl(25)	Median	Pctl(75)	Max
CIFAR-10	28,393	0.154	0.361	0	0	0	0	1
CIFAR-10 (only)	28,393	0.041	0.198	0	0	0	0	1
ImageNet	28,393	0.199	0.399	0	0	0	0	1
Nb. authors	28,393	4.320	2.581	1	3	4	5	100
Nb. references	28,393	36.212	20.505	1	22	33	47	811
International collaboration	28,393	0.245	0.430	0	0	0	0	1
Share company affiliation	28,393	0.040	0.152	0	0	0	0	1
Nb. patent citations	28,393	0.157	1.049	0	0	0	0	48
Nb. scientific citations	28,393	16.365	73.377	0	0	2	10	2279
Nb. dataset	28,393	1.300	0.657	1	1	1	1	7
Nb. modalities	28,393	1.246	0.465	1	1	1	1	5
Nb. tasks	28,393	46.647	29.430	1	25	54	67	183
Nb. tasks similar CIFAR-10	28,393	15.876	16.082	1	2	5	24	46
Share tasks similar CIFAR-10	28,393	0.345	0.350	0.022	0.043	0.109	0.522	1

Summary statistics for regression sample on publications mentioning annotated image datasets.

and the bibliometric data. We use different approaches to triangulate and pinpoint how CIFAR-10 contributed to making DL a dominant paradigm within AI, as well as the factors that explain the widespread use of the CIFAR-10 in industrial and academic settings.

5.1 Interviews analysis

5.1.1 Bridging the gap – Dataset characteristics

From the interviews that we conducted, the first element that prominent practitioners mentioned, was how CIFAR-10 went a step ahead of the MNIST but was more manageable than ImageNet, creating a sort of bridge between those two moments of the development of DL. Yoshua Bengio (Interviewee 5) mentions how the team at CIFAR had achieved some success with MNIST but "we didn't have datasets of comparable size for natural images". This was confirmed by Rob Fergus (Interviewee 2), and Yan LeCun (Interviewee 4) both of whom mentioned that there was a "gap" that CIFAR-10 helped to fill.

An important element that made CIFAR-10 a bridge is that it used a small number of categories, like MNIST, but also natural images, like ImageNet.

> It was much harder than the 10 digits [of MNIST], it was much, much harder. So it was useful, but the size was the same 60,000 training examples. So that mean, we could use the same kind of architectures. (Bengio, Interviewee 5).

That also helps explain why it was CIFAR-10 (and not CIFAR-100) the one that had the most impact:

> Yes, CIFAR-10 was the one that really had a big impact. For one it was exactly the same format that MNST, 10 categories. When people started working with CIFAR-100, it was much harder. So there are 100 categories, yeah, but you have the same amount of data so that the accuracy is much worse. So CIFAR-100 has been used, but as far as I know, not nearly as much as CIFAR-10. (Bengio, Interviewee 5).

5.1.2 Testing architectures and scaling up

The second element that emerges from the interviews it that CIFAR-10 was simple enough to test and iterate different algorithms and architectures, without requiring prohibitive amounts of computer power. Those architectures could then be used in more challenging datasets. Yoshua Bengio, Yan LeCun, and Rob Fergus insisted, in very similar terms, on the potential of CIFAR-10 for trying different architectures and iterate experiments:

> So in a way, what I'm trying to say is working with CIFAR-10 we discovered architecture tricks, if you want a methodology, for training deeper networks that Alex was then able to apply to ImageNet. So, yeah, CIFAR-10 was kind of instrumental on the path to the modern revolution of computer vision with DL. (Bengio, Interviewee 5).

> So I think, when you're trying to develop a new method, you've got to be able to iterate experiments quickly. And ImageNet is still, [. . .] a bit too big to do that with. And it turns out that the performance on CIFAR-10 generalizes quite well to other data sets like ImageNet. So, you can prototype on CIFAR-10. And then, you know, get some promising stuff, and then move over to something a bit bigger and harder. (Fergus, Interviewee 2).

This characteristic became instrumental in the development of AlexNet, which marked the turning point in the DL revolution:

> The AlexNet paper, I'm not sure would have happened, had it not been for CIFAR-10. Because otherwise, it would have been very difficult for them [Alex and Ilya] to go directly to the ImageNet dataset, which was quite new at the time, and definitely quite big at the time, too, and challenging to use. (Fergus, Interviewee 2)

5.1.3 Pedagogical potential

The third element that according to the interviewees help explain the success and persistence of CIFAR-10 as relevant tool for DL, is its pedagogic value. Since working on it does not require onerous computational capabilities, it can be easily used for teaching purposes. Bengio notes that:

> My students started to use it pretty soon, like, we were hungry for that. And we were aware of it even before it was released, because Geoff [Hinton] was talking about it. And you know, we were in close communication with Geoff [Hinton]. (Bengio Interviewee 5)

In the same line of reasoning, Fergus stated:

> Once you've got a lot more people interested in DL, it was a great sort of introductory data set. I mean, small enough, you can do it in, you know, if you're teaching a class, you can use it, because every student can run CIFAR-10 on their laptop, more or less. (Fergus Interviewee 2)

5.2 Survey analysis

5.2.1 Survey and respondents description

The survey has received 295 complete responses, with a total response rate of 9.4% at the time of closure. The vast majority of respondents (228) hold a doctorate degree (PhD), and most of the respondents are employed in academia. 20% of respondents work in industry or in a combination of industry and academia, when we look at the article affiliation we find a much lower share, about 11%, confirming the importance of mobility of researchers from academia to industry.

5.2.2 The importance of the datasets

Figure 3 reports responses on the importance of CIFAR-10 for the development of DL and Computer Vision were overwhelmingly positive. 76% believe CIFAR-10 was very or extremely important for the development of DL and 73% for the development of Computer Vision. 44% considered CIFAR-10 as extremely important for the development of DL in general, not only for Computer Vision. Though CIFAR-10 included labeled images, it is considered important for the development of the general field of DL.

5.2.3 Use compare to other OLDs

We asked the respondents to rate the reasons why they chose CIFAR-10 compared to similar datasets in the public domain. Based on our interviews, we included the quality of labeling,

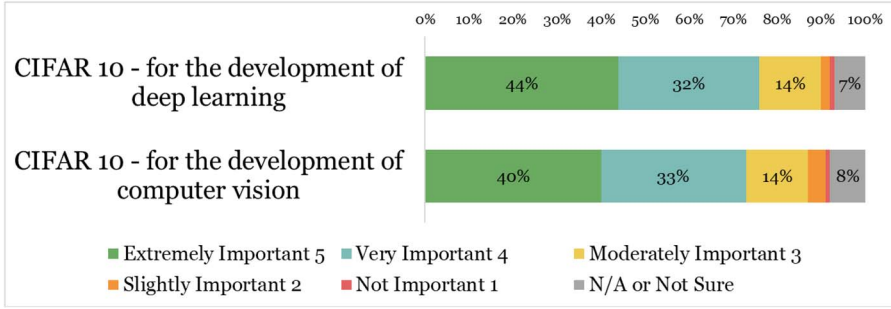


Figure 3. Survey results—CIFAR-10 Datasets Impact on DL & Computer Vision. *Note*: This figure shows the distribution of answers for the Impact question.

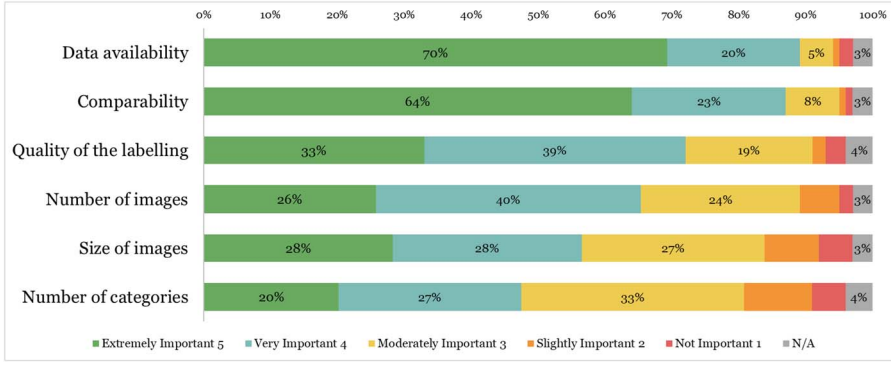


Figure 4. Survey results—comparing CIFAR-10 with Similar Datasets. *Note*: This figure shows the overall distribution of response for the question comparing CIFAR-10 to other datasets.

comparability as a benchmark, number of categories and images, image size, and data availability. Respondents rated each section on a Likert scale ranging from 1 (not important) to 5 (extremely important). Figure 4 presents the results of this question. Around 90% of respondents rated availability and comparability as very or extremely important. Quality of labeling and number of images were also considered important in explaining the choice of CIFAR-10 by 72% and 66% of the respondents.

5.2.4 Pedagogical use of the datasets

Figure 5 reports an interesting dimension of the survey: the pedagogical use of CIFAR-10. A significant number of respondents—193 (65%)—reported that they were introduced to the dataset during their studies (at the Bachelor, Master, or PhD level), and most respondents in academia routinely use it in their teaching programs. Furthermore, the responses to the open-ended question highlight the importance of CIFAR-10 as a pedagogical tool.

The last question of the questionnaire was open, we asked to describe why they thought that CIFAR-10 was important for the development of DL or CV. Out of the 295 complete questionnaires analysed, we have got 182 quite detailed answers with a lot of interesting insights. To analyse them, we have used the premium version of ChatGPT asking the algorithm to "identify the 5 main themes in the list of answers".

- **1. Benchmarking and comparison**: CIFAR-10 is frequently cited as a standard benchmark for evaluating and comparing the performance of various algorithms and models. It provides a common platform for fair comparisons and validation, which is essential for developing and testing new methods in DL and computer vision.
- **2. Accessibility and ease of use**: The dataset is noted for its accessibility and ease of use. It is readily available, simple to download, and manageable in terms of size and computational

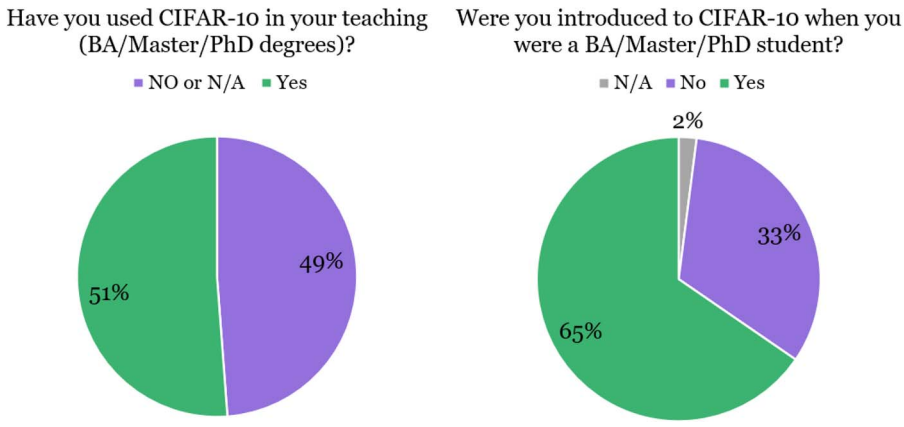


Figure 5. Survey results—integration of CIFAR-10 in teaching environment. *Note*: The graphs illustrate the responses of participants regarding their usage of CIFAR-10 in teaching, as well as their introduction to CIFAR-10 based on their academic background.

requirements. This makes it an ideal choice for both beginners and researchers without access to extensive computational resources.

- **3. Educational value and prototyping:** CIFAR-10 serves as an excellent educational tool for new learners and students. Its simplicity and comprehensibility make it a good starting point for understanding and experimenting with DL concepts. Additionally, it is suitable for rapid prototyping and initial testing of new ideas before scaling up to more complex datasets.
- **4. Quality and characteristics of the dataset:** The dataset is appreciated for its well-labeled, high-quality images. It offers a balanced number of categories and samples, which are sufficiently challenging for various image classification tasks. Its small image size and the diversity of the data allow for efficient experimentation and training.
- **5. Historical and continued relevance:** CIFAR-10 has historical significance in the field of computer vision and DL, having been used in many foundational studies and developments. Despite advancements in technology and the availability of larger datasets, it remains relevant due to its widespread use and the wealth of existing research that has utilized it as a benchmark.

We also created a word cloud of the most common terms used in the answers to the open question.[15] We excluded some frequently used terms like "CIFAR" and "database" to get a more accurate idea of the reasons respondents assign importance to CIFAR-10. Figure 6 shows that "benchmarking" and "learning" are the most used terms, with 49 and 41 times, respectively. These results are consistent with the analysis made through ChatGPT.

The evidence from both interviews and survey is consistent in highlighting that the specific characteristics (size, complexity, generalization) of CIFAR-10 made it the technological tool needed to develop and test convolutional neural network algorithms that gave rise to the DL revolution. We also find consistent evidence that the accessibility, versatility, use as a benchmark and pedagogical use of CIFAR-10 supported its continuous use and relevance even if much more complex and targeted OLDs became available in the 10 years after its release.

5.3 Econometric analysis

5.3.1 Results

In Table 3, we present the results of estimation equations 1 and 2 using as outcome variable the number of patent citations received by a paper using OLDs. Column 1 shows that papers mentioning only CIFAR-10 in the title, abstract, or keywords received, on average, $e^{0.418} - 1 = 51.89\%$ more patent citations than papers that do not mention it. Considering only the first

[15] The word cloud was generated using Voyant Tools, an online open-source text analysis software available at https://voyant-tools.org/.

Figure 6. Word cloud of main terms used in the open-ended question in the survey. *Note*: This figure shows the most common terms used by respondents in the open-ended question on why CIFAR-10 was important for the development of DL or CV.

half of the decade after the creation of CIFAR-10 (2010–2014), as shown in column 2, papers mentioning CIFAR-10 accrued, on average, nearly double the citations ($e^{0.692} - 1 = 99.77\%$) compared to those using other datasets. In the later period (2015–2022), as shown in column 3, papers using only CIFAR-10 continued to receive on average a higher number of citations than other papers, though the effect is less significant in terms of magnitude ($e^{0.366} - 1 = 44.20\%$) and statistical significance. Papers using CIFAR-10 and other datasets receive, on average, the same number of citations as those not using CIFAR-10 across all periods. In columns 4–6, we estimate regressions using specification 2 and find similar results for papers using only CIFAR-10 and those using CIFAR-10 along with other datasets. Papers using ImageNet received, on average, 24.86% more patent citations over the entire period ($e^{0.222} - 1 = 24.86\%$), primarily driven by papers published before 2015 (column 5, $e^{0.959} - 1 = 160.91\%$). CIFAR-10-only papers are expected to receive, on average, more citations ($e^{1.090} - 1 = 197.43\%$ in column 5) than papers using ImageNet, although the effect magnitude is comparable. Interestingly, more recent papers using ImageNet receive (not statistically significant for those published between 2015 and 2022, column 6), on average, a similar number of patent citations as papers using datasets other than CIFAR-10.

Table 4 shows the results of Poisson regressions for scientific citations. Columns 1–3 present results for specification 1. As we can observe in column 1, on average papers that mention only CIFAR-10 or CIFAR-10 and other datasets between 2010 and 2022 receive less citations than paper that do not cite it, but this difference is not statistically significant at conventional confidence levels. However, if we take into consideration only the first half of the 2010 decade (2010–2014) as in column 2, we find that papers using only CIFAR-10 or CIFAR-10 and others accrued on average 64.38% and 181.51% more citations than articles using other datasets. When considering the period post-2014 (2015–2022) in column 3, we see that papers using CIFAR-10, both alone and in combination with other datasets, received fewer citations compared to the average citations received by other papers. The results in column 3 are significant in both magnitude ($e^{-0.346} - 1 = -29.25\%$) and statistical terms for papers using CIFAR-10 along with other datasets, suggesting that the scientific citation impact of CIFAR-10 was primarily concentrated in its early years.

Table 3. Labeled datasets and patent citations

	Patents citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.418*	0.692**	0.366†	0.486**	1.090***	0.385†
	(0.169)	(0.211)	(0.199)	(0.179)	(0.211)	(0.207)
CIFAR-10 (others)	–0.055	0.339	–0.017	0.030	0.943	0.008
	(0.187)	(0.423)	(0.171)	(0.202)	(0.621)	(0.181)
ImageNet				0.222*	0.959**	0.067
				(0.105)	(0.355)	(0.105)
Log (Nb. authors)	0.480***	–0.103	0.681***	0.474***	–0.091	0.680***
	(0.099)	(0.157)	(0.109)	(0.100)	(0.153)	(0.109)
Log (Nb. references)	0.674***	1.751***	0.472**	0.654***	1.621***	0.466**
	(0.149)	(0.160)	(0.147)	(0.144)	(0.150)	(0.146)
International collab.	0.087	0.191	0.049	0.084	0.113	0.048
	(0.079)	(0.241)	(0.093)	(0.079)	(0.274)	(0.093)
Share company affil.	1.147***	2.518***	0.964***	1.117***	2.199***	0.955***
	(0.198)	(0.422)	(0.145)	(0.194)	(0.433)	(0.146)
Nb. datasets	0.014	0.162	0.044	0.007	0.182	0.042
	(0.090)	(0.121)	(0.081)	(0.090)	(0.136)	(0.081)
Nb. tasks	0.009***	0.010**	0.007***	0.007***	0.002	0.006**
	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)
Nb. modalities	0.093	–0.597	0.183*	0.145†	–0.477	0.198*
	(0.070)	(0.392)	(0.077)	(0.078)	(0.405)	(0.085)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	27,905	1620	26,220	27,905	1620	26,220
Dependent variable mean	0.15951	0.54691	0.13596	0.15951	0.54691	0.13596
Pseudo R^2	0.26575	0.25381	0.26234	0.26654	0.26795	0.26241

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number of patent families that cited the focal paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Columns (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2) and (3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

In columns 4–6 of Table 4, we consider model specification 2, where we compare papers citing either CIFAR-10 or ImageNet with those using other similar datasets. The results in column 4 indicate that, overall, papers mentioning CIFAR-10 do not differ significantly in terms of scientific citations compared to those mentioning other datasets. In contrast, papers mentioning ImageNet receive significantly more citations on average ($e^{0.386} - 1 = 47.11\%$). However, the difference in citations between papers mentioning CIFAR-10 and those not mentioning it is positive and significant in the first period (column 5) but becomes statistically insignificant in the last period (column 6). Meanwhile, papers using ImageNet consistently receive more citations on average in every period, though they received fewer citations than papers using CIFAR-10 in the first period. Specifically, during 2010–2014, papers mentioning CIFAR-10, either alone or with other datasets, received 101.58% and 295.11% more citations, respectively, compared to those using other datasets but not ImageNet. In comparison, papers using ImageNet received 81.14% more citations than those not using CIFAR-10.

5.3.2 Robustness checks

In Appendix D, we perform a series of robustness checks and sensitivity analysis using various estimation methods, different sample definitions, and alternative model specifications.

Table 4. Labeled datasets and scientific citations

	Scientific citations					
Model	Full (1)	2010-2014 (2)	2015-2022 (3)	Full (4)	2010-2014 (5)	2015-2022 (6)
CIFAR-10 (only)	–0.026	0.497**	–0.051	0.108	0.701***	0.082
	(0.090)	(0.172)	(0.101)	(0.095)	(0.186)	(0.108)
CIFAR-10 (others)	–0.294	1.035*	–0.346*	–0.155	1.374*	–0.207
	(0.191)	(0.489)	(0.156)	(0.216)	(0.584)	(0.179)
ImageNet				0.386***	0.594*	0.384***
				(0.092)	(0.284)	(0.092)
Log(Nb. authors)	0.351***	–0.130	0.438***	0.341***	–0.127	0.427***
	(0.072)	(0.178)	(0.076)	(0.072)	(0.178)	(0.076)
Log(Nb. references)	1.202***	1.367***	1.180***	1.180***	1.324***	1.157***
	(0.135)	(0.265)	(0.137)	(0.138)	(0.257)	(0.140)
International collab.	0.324***	0.335†	0.322***	0.321***	0.294	0.322***
	(0.041)	(0.173)	(0.038)	(0.042)	(0.179)	(0.039)
Share company affil.	1.271***	1.268**	1.284***	1.224***	1.079*	1.240***
	(0.151)	(0.471)	(0.155)	(0.145)	(0.454)	(0.151)
Nb. datasets	0.036	0.337*	0.041	0.029	0.320†	0.034
	(0.060)	(0.139)	(0.047)	(0.061)	(0.165)	(0.046)
Nb. tasks	0.008***	0.007**	0.007***	0.004***	0.003	0.004**
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Nb. modalities	0.189***	0.213	0.194***	0.290***	0.279	0.298***
	(0.045)	(0.187)	(0.047)	(0.047)	(0.192)	(0.048)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	28,393	1734	26,659	28,393	1734	26,659
Dependent variable mean	16.365	39.354	14.870	16.365	39.354	14.870
Pseudo R^2	0.41097	0.27955	0.42151	0.41527	0.28836	0.42589

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number scientific citations received by a paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2) and (3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Alternative statistical models. The Poisson model is not the only method for handling highly skewed count data.[16] Table D1 presents estimates from a negative binomial regression[17] based on specification 1. When scientific citation count is used as the dependent variable, the direction of the CIFAR-10 indicator variable coefficients remains consistent, although the significance levels differ. This alternative model does not alter our main finding: the influence of CIFAR-10 in scientific literature is predominantly concentrated in the earlier period. Results for patent citations are qualitatively similar.

Further restricted sample. The publications in our initial sample are quite diverse, including both journal articles and conference proceedings. To improve comparability, we further refine our sample to include only conference proceedings. These proceedings are more likely to represent recent advancements in ML models using labeled datasets. We restrict this further to papers that utilize datasets covering at least 10% of the tasks addressed by CIFAR-10 and are indexed

[16] Another approach involves using OLS to estimate models with log-transformed dependent variables or employing the inverse hyperbolic sine transformation. We chose not to use these estimation methods because our dependent variables include many zeros, and results can be sensitive to the arbitrary addition of a constant to handle these zero observations.

[17] Negative binomial regressions lack a fixed-effects estimator that is as consistent as the one in the Poisson fixed-effects model. To address this, we substitute fixed effects with categorical variables that control for the same factors as the fixed effects in the Poisson model.

in Papers With Code, aiming to minimize noise. Table D2 presents results for patent citations, while Table D3 shows results for scientific citations. The findings are qualitatively similar to our main results and exhibit greater significance and magnitude, providing additional evidence of the influence of CIFAR-10 and ImageNet on technological and scientific advancements in DL.

Enlarged sample. To ensure consistency in comparing patent and scientific citations, we initially removed a substantial number of observations where patent citations could not be accurately measured (13.86% of papers with missing patent citation values), as well as various publication types such as reviews, book chapters, and data papers. We then re-estimated our main specifications using an enlarged sample that includes papers with missing patent citation counts and all publication types for both patent citations (Table D4 and scientific citations and Table D5). The results are qualitatively consistent with our original findings.

Alternative dataset indicator variable. Since papers often benchmark new ML models against multiple datasets, isolating the citation impact of dataset size and complexity is challenging. To address this issue, we refined our indicator variable to distinguish between papers using only CIFAR-10 and those using CIFAR-10 in combination with other datasets. The variable for papers using only CIFAR-10 is more likely to reflect the effect of a small, yet sufficiently large, dataset, while the variable for papers using CIFAR-10 alongside other datasets also captures the influence of combining multiple datasets. To test the sensitivity of our analysis to this variable definition, we estimated models with an alternative specification where the independent variable is set to 1 for any publication mentioning CIFAR-10, regardless of whether it is used alone or with other datasets, and 0 otherwise. Table D6 shows that while the results for patent citations are consistent in sign with the full sample and the 2015–2022 period, they are no longer statistically significant. This suggests that papers more significant in technological development predominantly use CIFAR-10 alone, supporting the idea that small-but-large-enough datasets play a unique role in advancing DL models in computer vision. Results in Table D7 are qualitatively similar and confirm our main findings.

Citation lags. DL has experienced rapid growth in recent years, especially following the 2012 revolution and the release of the ChatGPT chatbot at the end of 2022. This surge has arguably heightened interest in older publications in the field and altered citation patterns in ways that publication year fixed effects may not fully capture. To assess the sensitivity of our results to different citation specifications, we compare citation counts within a fixed number of years after publication. This approach helps account for the dynamic nature of citation trends. Given the recent nature of our sample, we focus on a 3-year citation window to avoid losing too many observations from more recent years. Tables D8 and D9 present the results for patent and scientific citations within this 3-year window, respectively, demonstrating that the findings remain qualitatively consistent.[18]

Overall, our main results remain consistent, confirming the robustness of our findings across various control variables, fixed effects, sample definitions, and model specifications.

6. Discussion

Through our interviews, we learned that CIFAR-10 became a benchmark due to its technical specifications, including the nature of the images, their size, and the number of samples and categories. The timing of its release was also crucial to its popularity, as no other similar OLD was available at the time. ImageNet, released in 2009 by a team of university researchers and associated with the ImageNet Large Scale Visual Recognition Challenge, was also significant but proved too large and complex. Even today, in 2024, "solving" ImageNet with the best model and the largest supercomputer would take more than three years.

The survey confirms the insights from the interviews and reveals an additional role that CIFAR-10 played in the diffusion of DL methods. We present evidence that CIFAR-10 is extensively used in training computer scientists working with ML. Many researchers not only teach courses using

[18] We also conducted the analysis using the citation count of patent families sourced from Elsevier's PlumX Analytics. This metric includes only front-page citations from the European Patent Office (EPO), World Intellectual Property Organization (WIPO), Intellectual Property Office of the United Kingdom (IPO), United States Patent and Trademark Office (USPTO), and Japan Patent Office (JPO). The results were qualitatively similar.

CIFAR-10 but were also exposed to the datasets during their own graduate programs. This finding highlights teaching as a significant channel through which CIFAR-10 influences the field of DL.

The econometric analysis of the technological and scientific roles played by OLDs confirms that CIFAR-10 has had a significant influence on the development of DL compared to other OLDs, including its closest competitor, ImageNet. For science, we find that CIFAR-10's contribution was particularly important in the early years of DL development, with patent citations to CIFAR-10 remaining frequent in recent years. The role of ImageNet for the development of DL has been more prominent and continuous, likely due to its complexity, which allows for the testing and development of more advanced models. However, CIFAR-10 continued to outperform ImageNet (and all other OLDs) in technological citations even in recent years.

In terms of scientific complexity, CIFAR-10 was effectively "solved" by 2014, when state-of-the-art DL models achieved an error rate of around 3–4%, surpassing human-level accuracy in image classification tasks. Its sufficient complexity and status as a benchmark make it particularly useful in applied industrial research, where the speed of research and cost controls are more important than new scientific achievements. This continued use and technological relevance can explain the frequency of patent citations in recent years.

Based on the qualitative and quantitative evidence collected, it can be argued that CIFAR-10's lower computational requirements, ease of use, and the availability of a trained workforce make it more suitable for technology-oriented developments, as reflected in patent activity. These developments, which focus less on pushing the scientific frontier, are likely to rely more on CIFAR-10 compared to ImageNet and other more recent, complex datasets. The latter's increased complexity and higher computational demands make it less accessible for such practical applications.

This study has some limitations. First, while we have tried to interview active researchers in computer vision during the DL revolution, we were unable to interview the creators of CIFAR-10, Geoffrey Hinton, Vinod Nair, and Alex Krizhevsky. Gaining further insight into their motivations could illuminate the choice of dataset characteristics and how these are related to the development of DL models they were working on. Another limitation is the difficulty in identifying the specific OLDs used in each paper. Despite experimenting with different approaches, pinpointing the datasets in ML papers remains challenging. Future studies could employ more precise extraction algorithms to identify the datasets used, leveraging the full text of papers. Additionally, this study is primarily descriptive, making it challenging to establish causal effects of dataset usage. We do not observe the full process of building and refining ML model architectures or which datasets were effectively used prior to publication. Forthcoming investigations could exploit exogenous shocks in the availability of OLDs to understand their impact on the development of the field.

7. Conclusion

This paper aims to shed light on the role played by OLDs in the development of DL. Understanding the fundamental building blocks of this emerging technoscience is crucial, as these foundational elements will likely impact socioeconomic development in the coming years. Current advancements continue to be influenced by early events.

We find that CIFAR-10, a small yet sufficiently complex, well-labeled, and easily accessible database, was fundamental for the developments leading to the DL revolution and continues to shape the field's trajectory. We identify CIFAR-10 as one of the most important technological artifacts used to develop DL algorithms and architectures. We trace the creation of this dataset to the CIFAR NCAP Summer School in 2008, where graduate students, supervised by Geoffrey Hinton, a prominent scholar in the field, carried out the labeling of the datasets.

The evolution of AI in the early 2020s has been marked by significant investments by private companies in data collection and computing capacity to develop advanced LLMs expected to profoundly impact society. A few large companies, which have been recruiting top DL scientists (similar to the career trajectory of the lead scientists behind CIFAR-10) and attracted a substantial share of new graduate and postgraduate DL researchers (as evidenced by the current debate on universities' challenges in retaining DL scientists and our own data on the share of researchers working for companies), have the capacity to shape both the scientific and technological trajectory of DL.

Previously, the field developed with an open science approach, where public and private actors adhered to the ethos of open science by sharing data and methods. However, this approach has changed significantly. We may be entering a new phase in DL development characterized by a more traditional separation between science and technology, consistent with Dasgupta and David (1994) characterization of traditional science. If this is the case, there is an urgent need for substantial investment in public science conducted at universities. The "small is beautiful" model exemplified by the CIFAR-10 database may no longer be viable. Nonetheless, the widespread diffusion of CIFAR-10 and its origins reflect a human capital imprint of "open science" ethics that could be leveraged to maintain competitive dynamics in the DL field.

Acknowledgments

We extend our gratitude to the interview participants for sharing their insights on the development of Deep Learning. Special thanks go to Stefano Bianchini, Stijn Kelchtermans, Michele Pezzoni, and attendees of the 6th BETA Workshop, OIS Research Conference 2023, Munich Summer Institute Workshop 2023, REGIS Summer School 2023, Atlanta Conference on Science and Innovation Policy 2023, DRUID 2023, Eu-SPRI Conference 2023, 2nd International Conference on Science of Science and Innovation, and the International Joseph A. Schumpeter Society Conference 2024 for their valuable suggestions. We also appreciate the comments from seminar participants at the University of Manchester, Université Côte d'Azur, and Università Cattolica del Sacro Cuore. Special thanks to Fazliddin Shermatov for his outstanding research assistance. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Aldo Geuna, a fellow at the Canadian Institute for Advanced Research (CIFAR) in the Innovation, Equity, and the Future of Prosperity Program, reports financial support was provided by CIFAR's Catalyst Grant. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

We gratefully acknowledge financial support from the Canadian Institute for Advanced Research (CIFAR) through the Innovation Equity and the Future of Prosperity Program Catalyst Grant, as well as from the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS—Growing Resilient, INclusive, and Sustainable project (GRINS PE00000018 – CUP D13C22002160001).

References

Bengio, Y., Y. Lecun and G. Hinton (2021), 'Deep learning for AI,' *Communications of the ACM*, **64**(7), 58–65.

Bensaude-Vincent, B., S. Loeve, A. Nordmann and A. Schwarz (2011), 'Matters of interest: the objects of research in science and Technoscience,' *Journal for General Philosophy of Science*, **42**(2), 365–383.

Bianchini, S., M. Müller and P. Pelletier (2022), 'Artificial intelligence in science: an emerging general method of invention,' *Research Policy*, **51**(10), 104604.

Brownell, C. (2016), *How the artificial intelligence revolution was born in a Vancouver hotel*, CA. Financial Post. Toronto, ON, Canada. https://financialpost.com/technology/how-the-artificial-intelligence-revolution-was-born-in-a-vancouver-hotel (visited on January 13, 2023).

Callaway, E. (2020), "It will change everything': DeepMind's AI makes gigantic leap in solving protein structures,' *Nature*, **588**(7837), 203–204.

Chah, N. (2019), 'Down the deep rabbit hole: untangling deep learning from machine learning and artificial intelligence,' *First Monday*, **24**(2).

Crafts, N. (2021), 'Artificial intelligence as a general-purpose technology: an historical perspective,' *Oxford Review of Economic Policy*, **37**(3), 521–536.

David, P. A. (1998), 'Common agency contracting and the emergence of "Open Science" institutions,' *The American Economic Review*, **88**(2), 15–21.
David, P. A. (2000), 'The digital technology boomerang: new intellectual property rights threaten global "Open Science",' In: J. Bas and J. E. Stiglitz (eds), *Proceedings of the World Bank Annual Conference on Development Economics – Europe*. Washington, DC: World Bank. (Originally presented at the World Bank ABCDE Conference, Paris, June 26–28, 2000; Department of Economics Working Paper 00-16, Stanford University, October 2000).
David, P. A. (2003), 'The economic logic of open science and the balance between private property rights and the public domain in scientific data and information: A primer,' in J. M. Esanu and P. F. Uhlir (eds), *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*. Google-Books-ID: OULvaEq8YFoC. Washington, DC, USA: National Academies Press.
David, P. A. (2004), 'Can "Open Science" be protected from the evolving regime of IPR protections?' *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft*, **160**(1), 9–34.
Dasgupta, P. and P. A. David (1994), 'Toward a new economics of science,' *Research Policy*, **23**(5), 487–521.
Dean, J. (2020), '1.1 the deep learning revolution and its implications for computer architecture and Chip Design,' in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. San Francisco, CA, USA: IEEE, pp. 8–14.
European Commission (2018) *AI and Open Data: a crucial combination*. Publications Office of the European Union, https://data.europa.eu/en/publications/datastories/ai-and-open-data-crucial-combination (visited on April 1, 2023).
Farrow, J. (2019), *Turing Award honours CIFAR's 'pioneers of AI'*. Toronto, ON, Canada: CIFAR. https://cifar.ca/cifarnews/2019/03/27/turing-award-honours-cifar-s-pioneers-of-ai/ (visited on January 13, 2023).
Fortnow, L. (2009), 'ViewpointTime for computer science to grow up,' *Communications of the ACM*, **52**(8), 33–35.
Franceschet, M. (2010), 'The role of conference publications in CS,' *Communications of the ACM*, **53**(12), 129–132.
Frickel, S. and N. Gross (2005), 'A general theory of scientific/intellectual movements,' *American Sociological Review*, **70**(2), 204–232.
Fukushima, K. (1980), 'Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,' *Biological Cybernetics*, **36**(4), 193–202.
Goldman, S. (2022), *10 years later, deep learning 'revolution' rages on, say AI pioneers Hinton*, LeCun and Li. https://venturebeat.com/ai/10-years-on-ai-pioneers-hinton-lecun-li-say-deep-learning-revolution-will-continue/ (visited on September 20, 2022).
Goltsev, A., E. Kussul and T. Baidyk (2004), 'A process of differentiation in the assembly neural network,' In N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal and S. K. Parui (eds), *Neural Information Processing*. Springer, 452–457.
Graham, B. (2015), *Fractional Max-Pooling* (arXiv:1412.6071), arXiv.
Hinton, G. E., S. Osindero and Y.-W. Teh (2006), 'A fast learning algorithm for deep belief nets,' *Neural Computation*, **18**(7), 1527–1554.
Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J. and Ingrisch, M. (2023), 'ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports,' *European Radiology*, **34**(5), 2817–2825.
Jordan, K. (2024), *94% on CIFAR-10 in 3.29 Seconds on a Single GPU*. (arXiv:2404.00498), arXiv.
Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool et al. (2021), 'Highly accurate protein structure prediction with AlphaFold,' *Nature*, **596**(7873), 583–589.
Kastenhofer, K. and S. Molyneux-Hodgson (eds) (2021), *Community and Identity in Contemporary Technosciences*. Vol. **31**. Springer International Publishing: Cham.
Kersting, K. (2018), 'Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines,' *Frontiers in Big Data*, **1**, 6.
Klinger, J., J. Mateos-Garcia and K. Stathoulopoulos (2021), 'Deep learning, deep change? Mapping the evolution and geography of a general purpose technology,' *Scientometrics*, **126**(7), 5589–5621.
Koch, B. J. and D. Peterson (2024), *From Protoscience to Epistemic Monoculture: How Benchmarking Set the Stage for the Deep Learning Revolution*. (arXiv:2404.06647), arXiv.
Krizhevsky, A. (2009), *Learning multiple layers of features from tiny images*.
Krizhevsky, A., I. Sutskever and G. E. Hinton (2017), 'ImageNet classification with deep convolutional neural networks,' *Communications of the ACM*, **60**(6), 84–90.
Kuhn, T. S. (1970), *The structure of scientific revolutions*, 2nd edn. University of Chicago Press: Chicago,
LeCun, Y., Y. Bengio and G. Hinton (2015), 'Deep learning,' *Nature*, **521**(7553), 436–444.
Li, F.-F., M. Andreeto, M. Ranzato and P. Perona (2022), Caltech 101 (1.0)[Data set], *CaltechDATA*, **3**(086), 10–22002.

Martens, B. (2018), 'The impact of data access regimes on artificial intelligence and machine learning,' *European Commission, Joint Research Centre (JRC)*, **2018**(09).
Martínez-Plumed, F., E. Gómez and J. Hernández-Orallo (2021), 'Futures of artificial intelligence through technology readiness levels,' *Telematics and Informatics*, **58**, 101525.
Marx, M. and A. Fuegi (2020), 'Reliance on science: worldwide front-page patent citations to scientific articles,' *Strategic Management Journal*, **41**(9), 1572–1594.
Marx, M. and A. Fuegi (2022), 'Reliance on science by inventors: hybrid extraction of in-text patent-to-article citations,' *Journal of Economics & Management Strategy*, **31**(2), 369–392.
Merton, R. K. (1957), 'Priorities in scientific discovery: a chapter in the sociology of science,' *American Sociological Review*, **22**(6), 635–659.
Merton, R. K. (1973), *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago and London: University of Chicago Press.
Meyer, B., C. Choppy, J. Staunstrup and J. Van Leeuwen (2009), 'ViewpointResearch evaluation for computer science,' *Communications of the ACM*, **52**(4), 31–34.
Mitchell, T. M. (1997), *Machine Learning. en*. New York, NY, United States: McGraw-Hill.
OECD (2015), *Making Open Science a Reality, OECD Science, Technology and Industry Policy Papers, No. 25*. Paris: OECD Publishing.
Pavlik, J. V. (2023), 'Collaborating with ChatGPT: considering the implications of generative artificial intelligence for journalism and media education,' *Journalism & Mass Communication Educator*, **78**(1), 84–93.
Raimbault, B. and P.-B. Joly (2021), 'The emergence of Technoscientific fields and the new political sociology of science,' in K. Kastenhofer and S. Molyneux-Hodgson (eds), *Community and Identity in Contemporary Technosciences. Sociology of the Sciences Yearbook*. Springer International Publishing: Cham, pp. 85–106.
Rose, M. E. and J. R. Kitchin (2019), 'Pybliometrics: scriptable bibliometrics using a python interface to Scopus,' *SoftwareX*, **10**, 100263.
Schmidhuber, J. (2015), 'Deep learning in neural networks: an overview,' *Neural Networks*, **61**, 85–117.
Sevilla, J. et al. (2022), 'Compute trends across three eras of machine learning,' *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy*. Piscataway, NJ, USA: IEEE.
Shermatov, F. (2024), *National Computing Capacities for Frontier AI/DL Research*, MA thesis. Sorbonne University Association & University of Turin: Paris, France.
Stokel-Walker, C. and R. Van Noorden (2023), 'What ChatGPT and generative AI mean for science,' *Nature*, **614**(7947), 214–216.
Stokes, D. E. (2011), *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, DC, USA: Brookings Institution Press.
Torralba, A., R. Fergus and W. T. Freeman (2008), '80 million tiny images: a large dataset for non-parametric object and scene recognition,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 1958–1970.
Waldrop, M. M. (2019), 'What are the limits of deep learning?' *Proceedings of the National Academy of Sciences*, **116**(4), 1074–1077.
Weber, J. and B. Prietl (2021), 'On the rise of data-driven AI and its epistemontological foundations,' in A. Elliott (ed), *The Routledge Social Science Handbook of AI*. New York, London: Routledge, 58–73.

Appendix A. Qualitative methodology

The qualitative empirical material for this article is derived from a series of interviews with DL experts involved in the DL revolution, managers, and administrative staff at CIFAR associated with the DL funding program, and computer science PhD students. The survey was conducted among computer scientists and ML practitioners with scientific publications that mention CIFAR-10 in the title, abstract, and keywords indexed by Scopus.

First, we outline our study design. Next, we detail the mechanics of our interviews, including how we used the interview guide and conducted the sessions. Then, we present the survey text and response analysis.

Interviews

For the qualitative section, we conducted a series of semi-structured interviews of two kinds: shorter conversations were held with academics working on AI, regardless of their direct involvement with CIFAR-10, to gain a broad understanding of the field and identify general features that practitioners might seek in a training dataset; and in-depth interviews with key individuals who were directly or indirectly involved in the development of the CIFAR-10/CIFAR-100 datasets. Table A1 provides a comprehensive list of all the interviews conducted. Some of these interviews contributed to refining

Table A1. List of interviews

Interview number	Interviewee	Affiliation	Position	Interviewer	Date
1	Bruno Casella	University of Turin	PhD student	Daniel Souza	12/07/2022
2	Rob Fergus	NYU/ DeepMind	Professor/ Researcher Scientist	Daniel Souza/ Aldo Geuna/ Jeff Rodriguez	21/07/2022
3	Gianluca Mittone	University of Turin	PhD student	Daniel Souza	26/07/2022
4	Yann LeCun	NYU/Meta AI	Professor/VP & Chief AI Scientist	Daniel Souza/ Aldo Geuna	28/07/2022
5	Yoshua Bengio	Université de Montréal	Full Professor	Daniel Souza/ Aldo Geuna/ Jeff Rodriguez	17/10/2022
6	Rachel Parker	CIFAR	Sr Director, Research	Daniel Souza/ Aldo Geuna/ Jeff Rodriguez	18/11/2022
7	Melvin Silverman	CIFAR	Former VP of Research	Daniel Souza/ Aldo Geuna/ Jeff Rodriguez	08/12/2022

the research question, others provided empirical material for our conclusions, and some served both purposes.

The selection of interviewees was opportunistic, leveraging existing contacts. From these initial contacts, we employed a snowball sampling method to reach individuals outside our direct network, focusing on those recommended by interviewees and those with experience directly related to the creation of CIFAR-10 and CIFAR-100. Additionally, we conducted shorter interviews with individuals peripherally related to the topic, selected for their direct, personal knowledge of specific facts. This approach resulted in seven in-depth interviews that were transcribed, along with many off-the-record conversations.

We framed the interviews as conversations, with most conducted online and a few by phone, typically lasting between 15 minutes to an hour. Whenever permission was granted, we recorded the conversations, though participants could designate specific comments as off the record at any time. They were also given the option to review sections of the article in which they were mentioned before publication to ensure accuracy and agreement with how their comments were used. Interviewees had the choice to determine whether they wished to be identified by name or remain anonymous.

Interview guide

In-depth interviews were based on a guide reproduced below, which aws adapted in minor ways for each interview to reflect the fact that not all interviewees would have the same information to impart.

Final interview guide for in-depth interviews

Research question: What is the impact of CIFAR-10/CIFAR-100 on the development of Deep Learning?

Interview goal: Understand the role of CIFAR-10 and CIFAR-100 in the Deep Learning Revolution. We are also trying to better understand the chain of events that led to the development of these two datasets around 2008/2009, particularly the Summer School of August 2008.

Questions:

1. What was the impact of CIFAR-10/CIFAR-100 databases and how would you measure it?
   (a) Did it help the development of neural network algorithms?
   (b) Did it help the development of computer vision?
   (c) Did it help the development of other research topics in artificial intelligence?

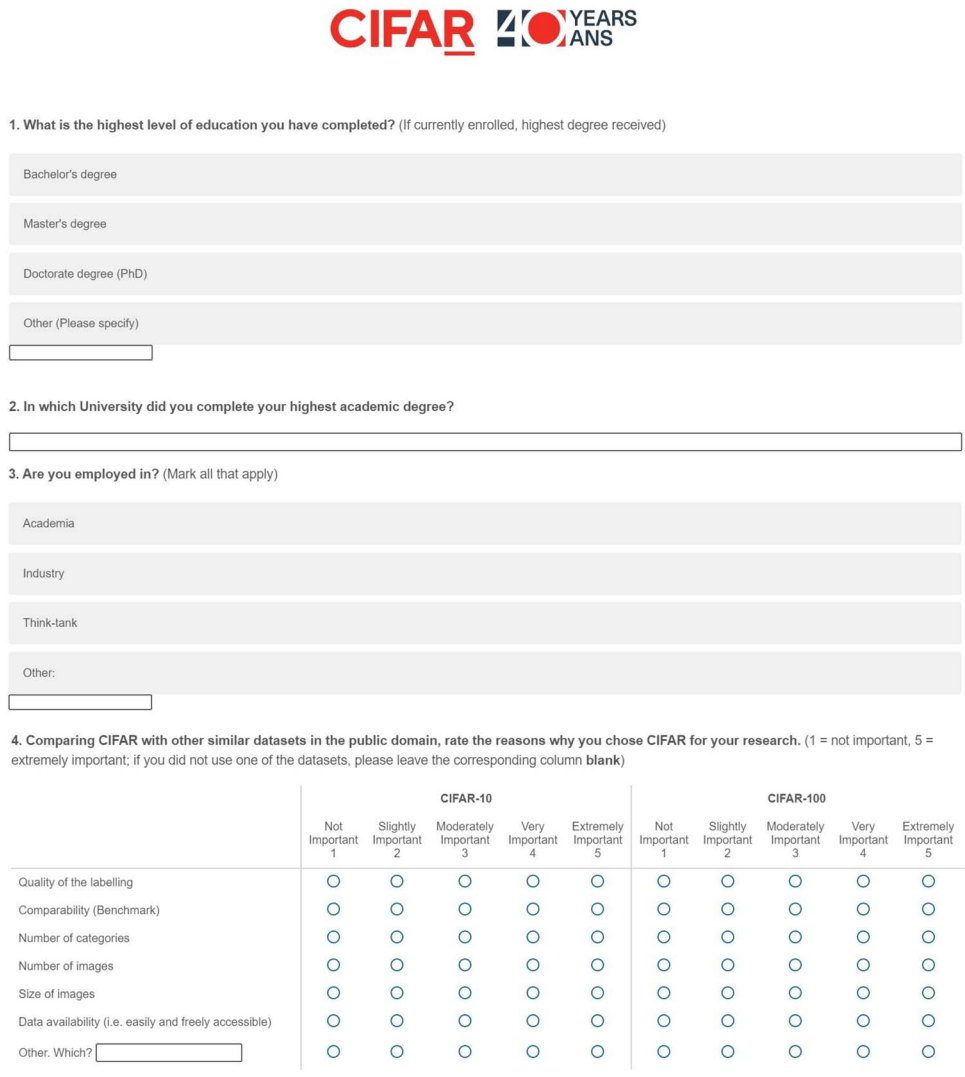


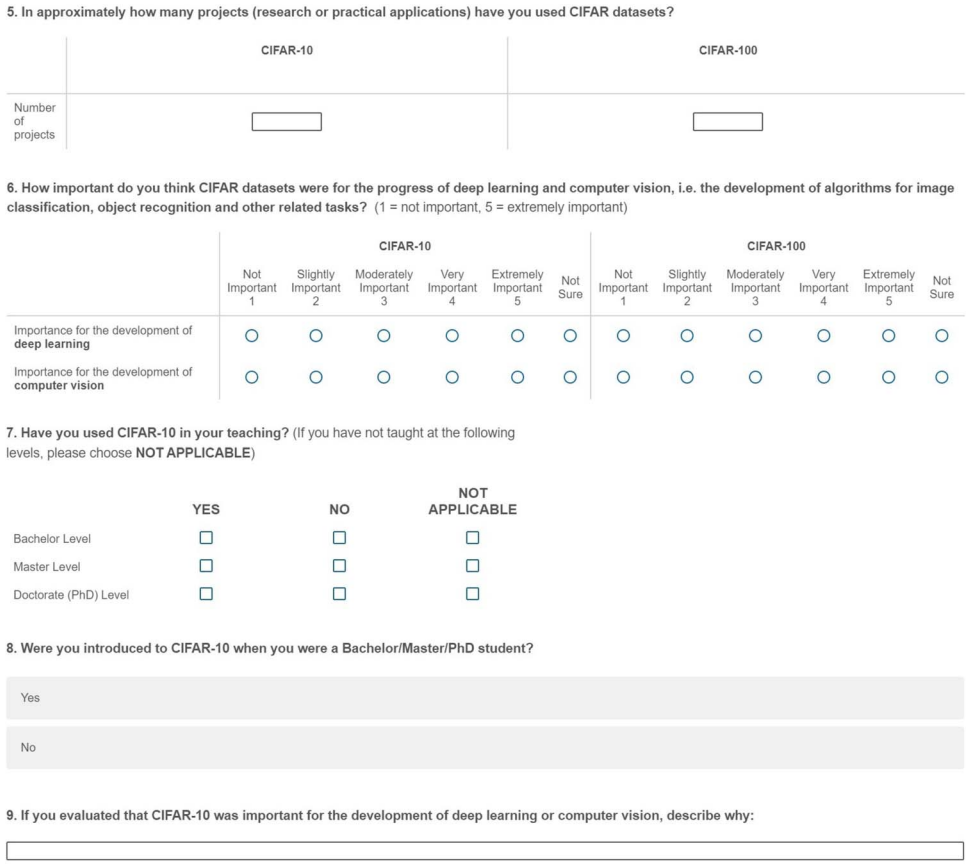
Figure A1. Survey text.

(d) Should we measure the impact on publications?
(e) Should we measure the impact on patents?
(f) Should we measure the impact on working papers?
(g) Should we measure the impact on conference proceedings papers?
(h) Should we measure the impact on media?

2. Can you tell us about the history of the AI projects at CIFAR?
3. How was the process of creating CIFAR-10/CIFAR-100?
4. Do you remember the NCAP summer school of 2008? Was that the moment in which CIFAR-10/100 were born? Was the whole process of labelling finished during the summer school or did it require additional work?
5. Who decided to give the name of CIFAR in CIFAR-10/100? Was it related to the funding of the project?

Wrap-up

- Who else do you think I should engage on this in relation to their work with CIFAR-10/CIFAR-100?


5. In approximately how many projects (research or practical applications) have you used CIFAR datasets?

	CIFAR-10	CIFAR-100
Number of projects		

6. How important do you think CIFAR datasets were for the progress of deep learning and computer vision, i.e. the development of algorithms for image classification, object recognition and other related tasks? (1 = not important, 5 = extremely important)

	CIFAR-10						CIFAR-100					
	Not Important 1	Slightly Important 2	Moderately Important 3	Very Important 4	Extremely Important 5	Not Sure	Not Important 1	Slightly Important 2	Moderately Important 3	Very Important 4	Extremely Important 5	Not Sure
Importance for the development of **deep learning**	○	○	○	○	○	○	○	○	○	○	○	○
Importance for the development of **computer vision**	○	○	○	○	○	○	○	○	○	○	○	○

7. Have you used CIFAR-10 in your teaching? (If you have not taught at the following levels, please choose **NOT APPLICABLE**)

	YES	NO	NOT APPLICABLE
Bachelor Level	☐	☐	☐
Master Level	☐	☐	☐
Doctorate (PhD) Level	☐	☐	☐

8. Were you introduced to CIFAR-10 when you were a Bachelor/Master/PhD student?

Yes

No

9. If you evaluated that CIFAR-10 was important for the development of deep learning or computer vision, describe why:

Figure A1. Continued

- Are you interested in seeing the results of this research?
- Thank you, very grateful for your time and thoughts.

Transcription. Most interviews were recorded and transcribed by the authors. When interviewees declined to be recorded, or when recording was impractical, shorthand notes were taken during the interview and subsequently expanded into detailed notes as soon as possible afterward.

Survey. The inputs from the interviews were used to produce a survey that was distributed to ML practitioners and academics.

The questionnaire consisted of 9 questions; 3 of the questions were related to the informant (education, place of work), and 4 directly to the evaluation of the CIFAR datasets. Figure A1 shows the full battery of questions.

To select the universe of possible respondents, we used the contact details of authors of papers extracted from Scopus that had used CIFAR-10 in their research. Out of the total of 6060 papers extracted, we were able to recover a valid email address of a corresponding author for 3033 papers. We sent a total of four requests to answer to the questionnaire to those authors in the period from September 2022 to February 2023.

The survey had a response rate of 9.7%, with 392 authors starting the survey (13%) and 295 completing it. The authors were from different geographical locations, with most affiliations in China and the United States.

Table A2 presents the summary statistics of our response analysis. The table includes the Kolmogorov–Smirnov test (addressing the variance in the distribution) to compare the three sample considered: Total population, Population Survey Sent, and Population Survey Answered. We included in our response analysis the following variables: Year of publication, Number of authors, Citations count, Type publication (Journal versus others), International collaborations, Number of OLDs used, Use of ImageNet, and Authors affiliated to a company.

Table A2. Summary statistics for response analysis

Descriptive statistics for total population of papers								
Statistic	**Year**	**Number of authors**	**Citation count**	**Type publications**	**Int. collaboration**	**OLDs**	**ImageNet**	**Company Affil.**
N	6060	6056	6060	6060	6060	6013	6060	5874
Ndist	14	22	266	2	2	8	2	2
Mean	2020.09	4.06	41.44	0.37	0.23	2.19	0.25	0.11
St. dev.	1.78	1.95	1184.43	0.48	0.42	1.03	0.43	0.31
Min	2010	1	0	0	0	1	0	0
Pctl(25)	2019	3	0	0	0	1	0	0
Pctl(75)	2021	5	8	1	0	3	0	0
Max	2023	36	90038	1	1	8	1	1
Kolmogorov-Smirnov test for total population ∗ Corresponding Email								
D	0.11061	0.051373	0.0635	0.19911	0.023343	0.028154	0.00090898	0.0033341
P-value	<2.2e-16	4.661e-05	1.666e-07	<2.2e-16	0.2207	0.08453	1	1
Descriptive Statistics for Papers with Corresponding Email Addresses								
N	3033	3033	3033	3033	3033	2987	3033	2935
Ndist	14	21	131	2	2	8	2	2
Mean	2020.57	4.26	10.84	0.57	0.26	2.26	0.25	0.1
St. dev.	1.59	2.12	54.59	0.5	0.44	1.05	0.43	0.31
Min	2010	1	0	0	0	1	0	0
Pctl(25)	2020	3	0	0	0	2	0	0
Pctl(75)	2022	5	6	1	1	3	0	0
Max	2023	36	1508	1	1	8	1	1
Kolmogorov-Smirnov test for corresponding email ∗ responded to survey								
D	0.15676	0.053131	0.056518	0.047216	0.021125	0.033868	0.045209	0.0074897
P-value	3.681e-06	0.4337	0.3569	0.5866	0.9998	0.9282	0.6419	1
Descriptive statistics for papers whose authors responded to survey								
N	295	295	295	295	295	283	295	280
Ndist	8	12	43	2	2	7	2	2
Mean	2020.99	3.97	9.86	0.62	0.28	2.21	0.2	0.1
St. dev.	1.42	1.86	51.95	0.49	0.45	1.08	0.4	0.3
Min	2016	1	0	0	0	1	0	0
Pctl(25)	2020	3	0	0	0	1	0	0
Pctl(75)	2022	5	6	1	1	3	0	0
Max	2023	12	816	1	1	7	1	1
Kolmogorov-Smirnov test for total population ∗ responded to survey								
D	0.25667	0.029133	0.11405	0.24632	0.044468	0.0085604	0.0443	0.010824
P-value	< 2.2e-16	0.9708	0.001327	2.998e-15	0.6342	1	0.6389	1

The tables above present descriptive statistics of variables for the population of papers analyzed in our study. These variables include publication year, number of authors, citation count, journal publications (using journal publications from aggregation type as a benchmark of comparison), international collaboration, number of datasets used, use of the Imagenet dataset, and affiliation of authors. The table shows the number of observations (*N*), the number of unique values (Ndist), the mean, standard deviation, minimum, maximum, and 25th and 75th percentiles for each variable. Table 1 provides statistics for the entire population, while Tables 2 and 3 present statistics for subsets of papers based on whether they had corresponding email addresses and whether their authors responded to our survey. At the bottom of each table, we report the results of the Kolmogorov-Smirnov test, which assesses the distributional differences between the variables in different tables. Specifically, we report the results of the KS test between Tables 1 and 3, Tables 1 and 2, and Tables 2 and 3, to identify any significant differences in the distribution of variables between the tables. These tables offer valuable insights into the characteristics of the papers in our sample and provide a foundation for further analysis.

The year of publication was the only factor hypothesis rejected by all of the three tests. Respondents are associated to papers published more recently; however, the difference is only in term of months, and is mainly due to no response from a few old papers. There are no significant differences between respondents and the population for which we had the email for the other seven variables we have considered.

When we compare respondents to the total population we see that journal articles are more frequent compare to other outlets, this is due to the fact that email addresses are difficult to be found in proceedings thus the email population was already biased in favour of journals. The

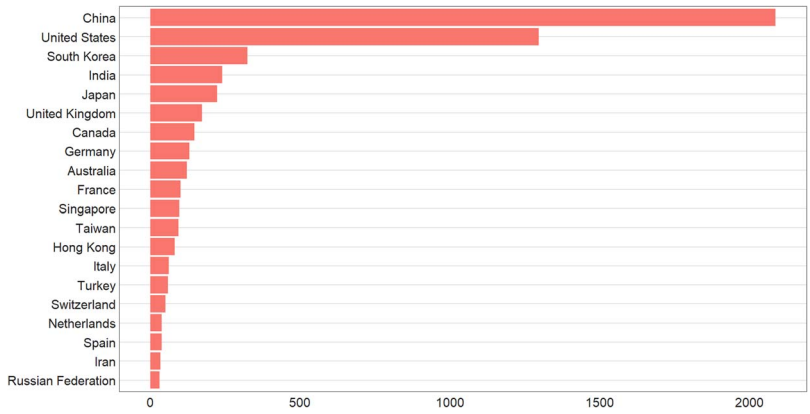


Figure A2. Distribution of CIFAR Papers among top 20 affiliations. *Note*: The graph illustrates the fractional count of papers based on affiliations. The top 20 affiliations listed in the graph collectively account for 90% of the CIFAR papers.

respondent sample includes also papers with fewer citations compared to the total population; the bias was already present in the email population as the most highly cited articles are in category of proceedings.

Appendix B. Dataset construction

In this appendix, we report the procedure we followed to construct the dataset used in the econometric analysis.

We obtained data on labeled datasets' names, introduction dates, and associated tasks from Papers With Code on July 17, 2023. We identified 358 datasets, including their names, full names, and variants, which had at least one task overlapping with CIFAR-10 tasks. A list of CIFAR-10 and ImageNet tasks can be found in Table B1. Missing introduction dates were filled automatically by querying for the name of the introductory paper on Scopus or manually when the introductory paper was unavailable. Additionally, we collected data on the number of papers indexed in Papers With Code connected to each dataset.

We then queried the Scopus API using Pybliometrics with the following query structure for each dataset name:

TITLE-ABS-KEY("dataset") AND PUBYEAR AFT intro-year.

This query identified papers published after the year of introduction, allowing for a 2-year margin to account for discrepancies between the first online appearance and official publication dates. When the number of papers identified on Scopus using this query significantly exceeded[19] the number of papers indexed on Papers With Code, we discarded the results. To refine the search for datasets with short or general names like "BSD," "Flowers," or "APRICOT," we appended "dataset" and "database" to the dataset names, ensuring the results were specific to machine learning papers. The query structure described above was executed on August 9, 2023. The following list of dataset names was queried using the outlined steps and yielded at least one publication indexed on Scopus:

102 Category Flower Dataset, A Visible-infrared Paired Dataset for Low-light Vision, AFHQ, AFHQ Cat, AFHQV2, AI2 Diagrams, AI2D, APRICOT dataset, ARC-100, ARID dataset, ASIRRA, Abnormal Event Detection Dataset, AbstractReasoning, AdvNet, AmsterTime, AmsterTime: A Visual Place Recognition Benchmark Dataset for Severe Domain Shift, Animal Faces-HQ, Animal Species Image Recognition for Restricting Access, ArtDL, BAM!, BCI database, BCI dataset, BCN 20000, BCNB, BSD database, BSD dataset, BSDS300, BTAD, Bamboo dataset, BarkNet 1.0, Behance Artistic Media, Bentham dataset, Bentham project, Berkeley Segmentation Dataset, BigEarthNet,

[19] We considered double the number of papers indexed as the maximum threshold after preliminary tests.

Table B1. Tasks performed using CIFAR-10 and ImageNet

CIFAR-10	ImageNet
Image classification	Image classification
Image generation	Image generation
Semi-supervised image classification	Semi-supervised image classification
Image clustering	Image clustering
Long-tail learning	Long-tail learning
Neural architecture search	Neural architecture search
Density estimation	Density estimation
Binarization	Binarization
Stochastic optimization	Stochastic optimization
Quantization	Quantization
Small data image classification	Small data image classification
Image compression	Image compression
Conditional image generation	Conditional image generation
Adversarial defense	Adversarial defense
Object recognition	Object recognition
Unsupervised image classification	Unsupervised image classification
Adversarial robustness	Adversarial robustness
Network pruning	Network pruning
Classification with binary weight network	Classification with binary weight network
Data augmentation	Data augmentation
Robust classification	Robust classification
Classification with binary neural network	Classification with binary neural network
Open-world semi-supervised learning	Open-world semi-supervised learning
Neural network compression	Neural network compression
Anomaly detection	Biologically plausible training
Graph classification	CW attack detection
Image retrieval	Classification
Out-of-distribution detection	Classification consistency
Learning with noisy labels	Color image denoising
Image classification with label noise	Continual learning
Semi-supervised image classification (cold start)	Contrastive learning
Personalized federated learning	Data free quantization
Unsupervised anomaly detection with specified settings – 30% anomaly	Domain generalization
Unsupervised anomaly detection with specified settings – 20% anomaly	Few-shot image classification
Unsupervised anomaly detection with specified settings – 1% anomaly	Few-shot learning
Unsupervised anomaly detection with specified settings – 0.1% anomaly	Generalized zero-shot learning
Unsupervised anomaly detection with specified settings – 10% anomaly	Image classification with differential privacy
Adversarial attack	Image colorization
Sequential image classification	Image compressed sensing
Model poisoning	Image deblurring
Sparse learning and binarization	Image inpainting
Novel class discovery	Image recognition
Hard-label attack	Image super-resolution
Clean-label backdoor attack (0.05%)	Incremental learning
Nature-inspired optimization algorithm	JPEG decompression
Long-tail learning on CIFAR-10-LT (ρ=100)	Knowledge distillation
	Linear-probe classification
	Model compression
	Object detection
	Parameter prediction
	Partial domain adaptation
	Prompt engineering
	Self-supervised image classification
	Sparse learning
	Unconditional image generation
	Unsupervised domain adaptation
	Variational inference
	Video matting
	Video visual relation detection
	Weakly supervised object detection
	Weakly supervised object localization
	Zero-shot learning
	Zero-shot object detection
	Zero-shot transfer image classification

This table lists all the tasks associated with CIFAR-10 and ImageNet, the two most commonly used labeled datasets in Papers With Code. Data collected on July 17, 2023, and compiled by the authors.

Boombox, BraTS 2016, BreakHis database, BreakHis dataset, Breast Cancer Histopathological Database, Breast Cancer Immunohistochemical Image Generation, CASIA-FASD, CCPD, CIFAR-10, CIFAR-10 Image Classification, CIFAR-10 image generation, CIFAR-100, CIFAR-100 vs CIFAR-10, CINIC-10, CIRCO, CIRR, CLEVR, CLEVR-Dialog, COCO, COCO 2014, COCO 2015, COCO 2017, COCO minival, COCO panoptic, COCO test-challenge, COCO test-dev, COCO+, COCO-Animals, COCO-CN, CORe50, COVID-19 Image Data Collection, COWC, CUB, CUB Birds, CUB-200-2011, CUB-LT, CURE-OR, CalTech 101 Silhouettes, Caltech-101, Caltech-256, Caltech-UCSD Birds-200-2011, Cars Overhead With Context, Cats and Dogs dataset, CelebA, CelebA-HQ, CelebA-Test, CelebAMask-HQ, CelebFaces Attributes Dataset, Challenging Unreal and Real Environments for Object Recognition, Chaoyang dataset, ChestX-ray8, Chinese City Parking Dataset, ChineseFoodNet, Cityscapes, Cityscapes test, Cityscapes val, Clothing1M, Cluttered Omniglot, Compose Image Retrieval on Real-life images, Compositional Language and Elementary Visual Reasoning, DF20, DF20 - Mini, DFUC2021, DTD dataset, Danish Fungi 2020, Deep PCB, DeepFashion, DeepFashion, DeepFashion2, DeepFish, DeepScores, DeepWeeds, Deepfashion2 validation, DensePose, DensePose-COCO, Describable Textures Dataset, DiagSet, Digits database, Digits dataset, Dry Bean Dataset, ELEVATER, EMNIST, EMNIST-Balanced, EMNIST-Digits, EMNIST-Letters, EgoHOS, EuroSAT, European Flood 2013 Dataset, Extended MNIST, Extended Yale B database, Extended Yale B dataset, Extended Yale-B, FER2013 database, FER2013 dataset, FFHQ, FGVC Aircraft, FGVC-Aircraft, FRGC database, FRGC dataset, Face Recognition Grand Challenge database, Face Recognition Grand Challenge dataset, FaceForensics++, Facial Expression Recognition 2013 Dataset, Fashion-Gen, Fashion-MNIST, Fishyscapes, Flickr database, Flickr dataset, Flickr-Faces-HQ, Flickr30k, FlickrLogos-32, Flowers database, Flowers dataset, Flowers-102, Food-101, Food-101N, Freiburg Groceries, Functional Map of the World, GOZ, GPR1200, Galaxy Zoo DECaLS, GasHisSDB, George Washington database, George Washington dataset, Google Landmarks, Google Landmarks Dataset v2, Grocery Store dataset, HR-ShanghaiTech, Hotels-50K, Hyper-Kvasir Dataset, IAM Handwriting, IAM database, IAM dataset, IAPR TC-12, IAPR TC-12 Benchmark, IARPA Janus Benchmark-B, IARPA Janus Benchmark-C, ICFG-PEDES, ICubWorld, IJB-B, IJB-C, ILSVRC 2015, ILSVRC 2016, INSTRE, IRMA database, IRMA dataset, ISBNet, Image Retrieval from Contextual Descriptions, ImageCoDe, ImageNet, ImageNet Detection, ImageNet-10, ImageNet-100, ImageNet-32, ImageNet-9, ImageNet-A, ImageNet-C, ImageNet-Caltech, ImageNet-LT, ImageNet-O, ImageNet-R, ImageNet-Sketch, ImageNet32, ImageNet64x64, Imagenette, In-Shop, InLoc, Incidents database, Incidents dataset, InstaCities1M, JFT-300M, JFT-3B, JHU CoSTAR Block Stacking Dataset, JHU-CROWD, JHU-CROWD++, Kannada-MNIST, Kitchen Scenes, Konzil, Kuzushiji-49, Kuzushiji-MNIST, Kvasir-Capsule, LFW database, LFW dataset, LHQ, LIDC-IDRI database, LIDC-IDRI dataset, LLVIP, LSUN, LSUN Bedroom, LaSCo, LabelMe, Labeled Faces in the Wild, Large-scale Scene UNderstanding Challenge, Lemons quality control dataset, Letter Recognition Data Set, Letter database, Letter dataset, Localized Narratives, Logo-2K+, MAMe, MIAD, MINC dataset, MLRSNet, MNIST, MNIST Large Scale dataset, MNIST-8M, MNIST-full, MNIST-test, MS-COCO, MSCOCO, MSRA Hand, MUAD, MVTEC ANOMALY DETECTION DATASET, MVTec AD, MVTec D2S, MVTecAD, Materials in Context Database, Melodic Design, Meta-Dataset, Microsoft Common Objects in Context, Million-AID, Moving MNIST, MuMiN, Multi-Modal CelebA-HQ, MultiMNIST, N-Caltech 101, NAS-Bench-201, NCT-CRC-HE-100K, NUS-WIDE, New Plant Diseases Dataset, Notre-Dame Cathedral Fire, NumtaDB, OFDIW, OMNIGLOT, ObjectNet, OmniBenchmark, Omniglot, OnFocus Detection In the Wild, Open Images V4, Open MIC, Open Museum Identification Challenge, Optical Recognition of Handwritten Digits, Oxford 102 Flower, Oxford 102 Flowers, Oxford Buildings, Oxford-IIIT Pet Dataset, Oxford105k, Oxford5k, PASCAL VOC 2007 database, PASCAL VOC 2007 dataset, PASCAL VOC 2011, PASCAL VOC 2011 test, PASCAL VOC 2012 database, PASCAL VOC 2012 dataset, PASCAL VOC 2012 test, PASCAL VOC 2012 val, PASCAL VOC database, PASCAL VOC dataset, PASCAL Visual Object Classes Challenge, PCam, PGM dataset, PKU-Reid, PROMISE12, Pano3D, PatchCamelyon, Patzig, Perceptual Similarity, PhotoChat, Places database, Places dataset, Places-LT, Places2, Places205, Places365, Places365-Standard, PlantVillage database, Procedurally Generated Matrices (PGM), Processed Twitter, QMNIST, Quick, Draw! Dataset, QuickDraw-Extended, RESISC45 database, RESISC45 dataset, RF100, RIT-18, RPC database, RPC dataset, RVL-CDIP, Recipe1M, Recipe1M+, Replica dataset, Retail Product Checkout, Ricordi, Riseholme-2021, Road Anomaly, Rotated MNIST, Rotating MNIST, SI-Score, STAIR Captions, STL-10,

STN PLAD, STN Power Line Assets Dataset, SUN Attribute, SUN397, SVHN, SVLD, Saint Gall, Schiller dataset, Schwerin, Self-Taught Learning 10, Semi-Supervised iNaturalist, Semi-iNat, Sequential MNIST, Sewer-ML, ShanghaiTech, ShanghaiTech A, ShanghaiTech B, Shiller, ShoeV2, Silhouettes database, Silhouettes dataset, SketchHairSalon, SketchyScene, So2Sat LCZ42, Spot-the-diff, Stanford Cars, Stanford Dogs, Stanford Online Products, Street View House Numbers, StreetStyle, Structured3D, StyleGAN-Human, Stylized ImageNet, TMED, TUM-GAID, Tencent ML-Images, Thyroid Disease database, Thyroid Disease dataset, Thyroid database, Thyroid dataset, Tiny ImageNet, Tiny Images, Tiny-ImageNet, TransNAS-Bench-101, Tsinghua Dogs, Twitter100k, UBI-Fights, UCF-CC-50, UCSD Anomaly Detection Dataset, UCSD Ped2, UCSD-MIT Human Motion, UFPR-AMR, UMIST, UMist, UPIQ, USPS database, USPS dataset, Unified Photometric Image Quality, VOC12, VegFru, VehicleX, Verse dataset, Visual Madlibs, Visual Wake Words, VocalFolds, WHU-Hi, WIT dataset, Washington RGB-D, WebVision, WebVision-1000, Wikipedia-based Image Text, Wine Data Set, Wine database, Wine dataset, Wuhan UAV-borne hyperspectral image, YFCC100M, beanTech Anomaly Detection, cats vs dogs, ciFAIR-10, cifar10, cifar100, fMoW, fashion mnist, food101, iCartoonFace, iNat2021, iNaturalist, iNaturalist 2018, iNaturalist 2019, iSUN, imagenet-1k, mini-ImageNet-LT, smallNORB, tieredImageNet, xBD.

From the original list of 358 unique labeled datasets, we managed to identify on Scopus 37,242 papers citing 264 unique labeled datasets. The discrepancy is due to some datasets not being identified precisely enough using the described steps, i.e. having too many results even after adding the words "dataset" and "database" in the query, or having no results at all. The labeled datasets we could not find were either not indexed by Scopus or did not mention the datasets in the title, abstract, or keywords. We then merged this information with Papers With Code's annotated datasets information to obtain the complete sample with all the necessary data.

Appendix C. Additional descriptives

Table C1 reports the main characteristics of 15 selected labeled datasets in our sample, including their names, supporting institutions, introduction years, number of categories, and instance counts.

Additionally, we constructed Table 1 using estimates from Shermatov (2024) to provide computational requirements for running state-of-the-art (SOTA) models on the four most commonly used open labeled datasets (OLDs) in the literature: ImageNet, COCO, MNIST, and CIFAR-10. These estimates were derived from Epoch AI's methods for assessing the training compute of deep learning systems, including operation counts, GPU time, and performing calculations based on SOTA data compiled by the PaperswithCode platform. More details can be found at https://epochai.org/blog/estimating-training-compute.

These calculations are intended for illustrative and comparative purposes only. Computing power requirements can differ significantly across datasets and architectures. We provide rough estimates by comparing the compute demands of leading models with the compute capabilities of different hardware. We present estimates for hardwares with two levels of performance: supercomputers and research laptops. For supercomputers, we use the Frontier exascale machine, which delivers 1194 PFlop/s, as a benchmark. For research laptops, we reference average devices with NVIDIAGeForceRTX4080 or AMDRadeonRX7900XTXGPUs, which provide roughly 5x less flops/s. Actual flops allocated for deep learning tasks can vary greatly depending on the specific model and its configuration.

For example, the top CIFAR-10 model by Google Research Brain Team, ViT-H/14, requires a substantial amount of flops to achieve 99.5% accuracy. A simpler model, "airbench," requires 3.6 times fewer flops to achieve human-level accuracy of 94% (Jordan, 2024). On an average researcher laptop, training this model on CIFAR-10 would take approximately *10 seconds to achieve 94% accuracy*.

Table C1. Open labeled datasets characteristics

Dataset	Full name	Created by	Introduced Year	Categories	Instances	Num. Papers
ImageNet	ImageNet Large Scale Visual Recognition Challenge	Princeton University	2009	21,841	14,197,122	5,658
MNIST	Modified National Institute of Standards and Technology	AT&T Bell Laboratories	1998	10	70,000	5,647
COCO	Common Objects in Context	Microsoft	2014	80	2,500,000	4,894
CIFAR-10	Canadian Institute for Advanced Research 10	University of Toronto	2009	10	60,000	4,366
CUB-200-2011	Caltech-UCSD Birds-200-2011	California Institute of Technology	2011	200	11,788	2,419
CIFAR-100	Canadian Institute for Advanced Research 100	University of Toronto	2009	100	60,000	1,515
Cityscapes	Dataset for urban scene understanding and autonomous driving	Daimler AG and University of Tübingen	2016	30	5,000	1,443
CelebA	Celebrities Attributes Dataset	Chinese University of Hong Kong	2014	10,177	202,599	639
Fashion-MNIST	Dataset for benchmarking machine learning algorithms	Zalando Research	2017	10	70,000	624
SVHN	Street View House Numbers	Stanford University	2011	10	604,388	606
PASCAL VOC	Pattern Analysis, Statistical Modelling and Computational Learning – Visual Object Classes Challenge	University of Oxford	2005	20	27,450	519
Flickr30k	Flickr 30k Dataset	University of Illinois	2014	1	31,783	388
BSD	Berkeley Segmentation Dataset	Berkeley Vision and Learning Center	2003	1	500	207
Extended Yale B	Extended Yale Face Database B	Yale University	2001	38	2,414	144
FRGC	Facial Recognition Grand Challenge	National Institute of Standards and Technology	2006	1	50,000	45

This table provides information on 15 datasets from our sample, including their names, supporting institutions, introduction years, number of categories, instance counts, and number of research papers in our sample using them. Elaborated by the authors.

Appendix D. Robustness checks and sensitivity analysis

In this Appendix, we report the robustness checks and sensitivity analysis we run and discussed in the Section 5.3 of this article.

Table D1. Robustness check: negative binomial

	Patent citations			Scientific citations		
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.357*	1.491*	0.292†	0.013	0.606†	0.013
	(0.153)	(0.596)	(0.161)	(0.069)	(0.352)	(0.074)
CIFAR-10 (others)	–0.048	0.459	–0.039	–0.168†	1.353*	–0.175*
	(0.142)	(0.550)	(0.138)	(0.097)	(0.672)	(0.084)
ImageNet	0.168*	0.598†	0.084	0.426***	0.403	0.440***
	(0.083)	(0.316)	(0.089)	(0.052)	(0.313)	(0.054)
Log (Nb. authors)	0.536***	0.031	0.639***	0.344***	–0.205*	0.395***
	(0.079)	(0.167)	(0.084)	(0.040)	(0.100)	(0.041)
Log (Nb. references)	0.603***	1.601***	0.502***	0.947***	1.023***	0.958***
	(0.095)	(0.235)	(0.099)	(0.110)	(0.128)	(0.114)
International collab.	0.046	0.333†	0.002	0.416***	0.288**	0.416***
	(0.068)	(0.192)	(0.079)	(0.047)	(0.101)	(0.047)
Share company affil.	0.920***	2.329***	0.783***	1.226***	0.931*	1.224***
	(0.156)	(0.565)	(0.158)	(0.148)	(0.454)	(0.143)
Nb. datasets	0.031	–0.109	0.047	–0.003	0.102	0.009
	(0.067)	(0.149)	(0.068)	(0.038)	(0.199)	(0.036)
Nb. tasks	0.006***	0.005	0.006***	0.003***	0.005*	0.002***
	(0.001)	(0.004)	(0.001)	(0.001)	(0.003)	(0.001)
Nb. modalities	0.159*	–0.207	0.198**	0.218***	0.258*	0.226***
	(0.069)	(0.318)	(0.070)	(0.040)	(0.124)	(0.042)
Observations	28,393	1734	26,659	28,393	1734	26,659
Dependent variable mean	0.15676	0.51096	0.13373	16.365	39.354	14.870
Pseudo R^2	0.15171	0.10799	0.15478	0.10241	0.04174	0.10559
Over-dispersion	0.17058	0.18649	0.18141	0.50494	0.56484	0.50624

This table reports estimates of regressions of the model described in equation 2. The dependent variable for columns (1–3) is the total number of patent families citing the focal papers, while for columns (4–6) it is the total number of scientific citations received by the focal papers. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract or keywords. Columns (1 and 4) report our baseline results of the estimates stemming from a Poisson regression. Columns (2–3) and (5–6) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P< 0.1$; $^{*}P< 0.05$; $^{**}P< 0.01$; $^{***}P< 0.001$.

Table D2. Robustness check: restricted sample – patent citations

	Patents citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.769***	1.305**	0.621**	0.876***	1.883***	0.668**
	(0.181)	(0.421)	(0.209)	(0.199)	(0.500)	(0.215)
CIFAR-10 (others)	0.049	0.184	0.073	0.189	1.029†	0.138
	(0.216)	(0.368)	(0.192)	(0.227)	(0.543)	(0.197)
ImageNet				0.338*	1.170**	0.167
				(0.140)	(0.411)	(0.122)
Log (Nb. authors)	0.500***	0.437*	0.608***	0.492***	0.421*	0.605***
	(0.101)	(0.186)	(0.118)	(0.101)	(0.200)	(0.118)
Log(Nb. references)	0.509**	1.536***	0.347*	0.453**	1.267***	0.321†
	(0.173)	(0.310)	(0.173)	(0.158)	(0.199)	(0.167)
International collab.	–0.015	0.050	–0.032	–0.030	–0.111	–0.037
	(0.102)	(0.419)	(0.105)	(0.103)	(0.460)	(0.104)
Share company affil.	1.425***	3.533***	1.111***	1.360***	3.017***	1.082***
	(0.244)	(0.288)	(0.184)	(0.235)	(0.308)	(0.181)
Nb. datasets	–0.170	–0.124	–0.077	–0.170	–0.297	–0.074
	(0.106)	(0.193)	(0.099)	(0.106)	(0.205)	(0.099)
Nb. tasks	0.021***	0.036***	0.015***	0.018***	0.030***	0.014***
	(0.003)	(0.007)	(0.004)	(0.003)	(0.009)	(0.004)
Nb. modalities	–0.080		0.190	–0.038		0.205
	(0.179)		(0.164)	(0.173)		(0.164)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	15,407	799	14,504	15,407	799	14,504
Dependent variable mean	0.17005	0.61452	0.14679	0.17005	0.61452	0.14679
Pseudo R^2	0.29363	0.36343	0.28420	0.29624	0.39371	0.28484

This table reports estimates of regressions of the models described in equations 1 and 2 using a sample that includes only conference proceedings and datasets with at least 100 papers indexed by Papers With Code and 5 or more (10%) tasks overlapping with CIFAR-10. The dependent variable is the total number of patent families that cited the focal paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D3. Robustness check: restricted sample – scientific citations

	Scientific citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.217†	0.773*	0.157	0.418***	0.997***	0.368**
	(0.128)	(0.305)	(0.148)	(0.119)	(0.298)	(0.134)
CIFAR-10 (others)	–0.313	0.735	–0.357†	–0.099	1.182†	–0.142
	(0.231)	(0.454)	(0.185)	(0.236)	(0.603)	(0.185)
ImageNet				0.552***	0.624†	0.577***
				(0.073)	(0.340)	(0.074)
Log (Nb. authors)	0.336***	0.193	0.372***	0.319***	0.197	0.353***
	(0.087)	(0.216)	(0.093)	(0.087)	(0.217)	(0.093)
Log (Nb. references)	1.069***	0.994*	1.080***	0.992***	0.918*	1.001***
	(0.166)	(0.408)	(0.165)	(0.157)	(0.378)	(0.158)
International collab.	0.221***	0.096	0.244***	0.199***	0.009	0.229***
	(0.058)	(0.278)	(0.051)	(0.058)	(0.283)	(0.052)
Share company affil.	1.271***	2.000***	1.268***	1.179***	1.754***	1.182***
	(0.128)	(0.471)	(0.117)	(0.117)	(0.471)	(0.108)
Nb. datasets	–0.010	0.411	0.017	0.007	0.318	0.040
	(0.126)	(0.277)	(0.110)	(0.131)	(0.318)	(0.113)
Nb. tasks	0.015***	0.021***	0.013**	0.009†	0.016*	0.007
	(0.004)	(0.005)	(0.005)	(0.005)	(0.007)	(0.005)
Nb. modalities	0.235		0.310†	0.346*		0.425**
	(0.178)		(0.165)	(0.168)		(0.158)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	15,600	895	14,703	15,600	895	14,703
Dependent variable mean	16.558	41.165	15.062	16.558	41.165	15.062
Pseudo R^2	0.41044	0.30952	0.42224	0.42279	0.32472	0.43591

This table reports estimates of regressions of the models described in equations 1 and 2 using a sample that includes only conference proceedings and datasets with at least 100 papers indexed by Papers With Code and 5 or more (10%) tasks overlapping with CIFAR-10. The dependent variable is the total number scientific citations received by a paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: †$P < 0.1$; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Table D4. Robustness check: enlarged sample – patent citations

	Patents citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.412*	0.692**	0.359†	0.481**	1.089***	0.379†
	(0.169)	(0.214)	(0.199)	(0.179)	(0.214)	(0.206)
CIFAR-10 (others)	–0.052	0.349	–0.012	0.034	0.952	0.014
	(0.187)	(0.426)	(0.171)	(0.202)	(0.624)	(0.181)
ImageNet				0.225*	0.958**	0.072
				(0.104)	(0.353)	(0.104)
Log (Nb. authors)	0.514***	–0.130	0.723***	0.508***	–0.118	0.721***
	(0.105)	(0.157)	(0.113)	(0.105)	(0.154)	(0.114)
Log (Nb. references)	0.639***	1.745***	0.434**	0.620***	1.620***	0.428**
	(0.152)	(0.157)	(0.150)	(0.148)	(0.146)	(0.149)
International collab.	0.097	0.191	0.061	0.094	0.113	0.061
	(0.079)	(0.239)	(0.092)	(0.079)	(0.272)	(0.092)
Share company affil.	1.155***	2.509***	0.971***	1.125***	2.188***	0.962***
	(0.198)	(0.417)	(0.147)	(0.194)	(0.430)	(0.147)
Nb. datasets	0.019	0.160	0.051	0.012	0.180	0.049
	(0.091)	(0.121)	(0.081)	(0.090)	(0.135)	(0.081)
Nb. tasks	0.008***	0.010**	0.006***	0.006***	0.002	0.006**
	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)
Nb. modalities	0.089	–0.646†	0.181*	0.142†	–0.507	0.197*
	(0.070)	(0.381)	(0.077)	(0.078)	(0.396)	(0.085)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	28,433	1658	26,705	28,433	1658	26,705
Dependent variable mean	0.15855	0.54403	0.13503	0.15855	0.54403	0.13503
Pseudo R^2	0.26884	0.25634	0.26651	0.26965	0.27031	0.26660

This table reports estimates of regressions of the models described in equations 1 and 2 using an enlarged sample that encompasses all kinds of publication outlets. The dependent variable is the total number of patent families that cited the focal paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D5. Robustness check: enlarged sample – scientific citations

	Scientific citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.192	1.250**	0.082	0.363*	1.416***	0.251†
	(0.145)	(0.397)	(0.133)	(0.150)	(0.388)	(0.136)
CIFAR-10 (others)	–0.102	0.313	–0.107	0.045	0.550	0.038
	(0.167)	(0.425)	(0.166)	(0.188)	(0.474)	(0.183)
ImageNet				0.437***	0.530*	0.429***
				(0.100)	(0.208)	(0.109)
Log (Nb. authors)	0.419***	0.188	0.468***	0.401***	0.181	0.452***
	(0.057)	(0.166)	(0.062)	(0.057)	(0.160)	(0.062)
Log (Nb. references)	1.186***	1.337***	1.178***	1.164***	1.290***	1.157***
	(0.101)	(0.180)	(0.098)	(0.102)	(0.178)	(0.100)
International collab.	0.263***	0.367*	0.254***	0.256***	0.332†	0.250***
	(0.044)	(0.166)	(0.049)	(0.046)	(0.170)	(0.052)
Share company affil.	1.313***	1.950**	1.297***	1.254***	1.711**	1.244***
	(0.137)	(0.595)	(0.146)	(0.130)	(0.607)	(0.139)
Nb. datasets	0.049	0.476**	0.037	0.048	0.442**	0.038
	(0.048)	(0.148)	(0.040)	(0.047)	(0.160)	(0.039)
Nb. tasks	0.007***	0.009***	0.006***	0.003**	0.005	0.002†
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Nb. modalities	0.136*	–0.022	0.154**	0.254***	0.039	0.272***
	(0.059)	(0.182)	(0.059)	(0.059)	(0.179)	(0.058)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication ear fixed effect	YES	YES	YES	YES	YES	YES
Observations	33,693	2062	31,630	33,693	2062	31,630
Dependent variable mean	18.511	45.833	16.731	18.511	45.833	16.731
Pseudo R^2	0.43621	0.33513	0.44532	0.44171	0.34236	0.45083

This table reports estimates of regressions of the models described in equations 1 and 2 using an enlarged sample that encompasses all types of publication outlets and papers missing patent citation information. The dependent variable is the total number of scientific publications that cited the focal paper. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D6. Robustness check: alternative datasets indicator variables – patent citations

	Patents citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10	0.126	0.491*	0.131	0.207	1.011**	0.156
	(0.143)	(0.226)	(0.146)	(0.158)	(0.336)	(0.158)
ImageNet				0.228*	0.963**	0.073
				(0.103)	(0.342)	(0.105)
Log (Nb. authors)	0.483***	–0.105	0.686***	0.477***	–0.092	0.684***
	(0.099)	(0.156)	(0.109)	(0.100)	(0.152)	(0.109)
Log (Nb. references)	0.674***	1.756***	0.472**	0.655***	1.623***	0.465**
	(0.149)	(0.159)	(0.147)	(0.145)	(0.151)	(0.146)
International collab.	0.084	0.189	0.046	0.081	0.111	0.046
	(0.079)	(0.236)	(0.093)	(0.079)	(0.268)	(0.093)
Share company affil.	1.154***	2.516***	0.971***	1.123***	2.196***	0.961***
	(0.197)	(0.418)	(0.144)	(0.193)	(0.426)	(0.145)
Nb. datasets	–0.046	0.132	–0.005	–0.051	0.170	–0.006
	(0.082)	(0.098)	(0.076)	(0.082)	(0.108)	(0.076)
Nb. tasks	0.008***	0.010*	0.007***	0.006***	0.002	0.006**
	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)
Nb. modalities	0.102	–0.600	0.191*	0.155*	–0.478	0.208*
	(0.070)	(0.392)	(0.076)	(0.078)	(0.406)	(0.084)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	27,905	1620	26,220	27,905	1620	26,220
Dependent variable mean	0.15951	0.54691	0.13596	0.15951	0.54691	0.13596
Pseudo R^2	0.26509	0.25357	0.26186	0.26593	0.26791	0.26195

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number of patent families that cited the focal paper. The response variables are indicator variables that are equal to one if a paper mentions CIFAR-10 or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D7. Robustness check: alternative datasets indicator variables – scientific citations

	Scientific citations					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10	–0.212	0.891*	–0.255†	–0.074	1.177*	–0.117
	(0.153)	(0.392)	(0.133)	(0.174)	(0.468)	(0.154)
ImageNet				0.386***	0.575*	0.384***
				(0.092)	(0.274)	(0.093)
Log (Nb. authors)	0.352***	–0.125	0.440***	0.342***	–0.122	0.429***
	(0.073)	(0.181)	(0.077)	(0.072)	(0.181)	(0.076)
Log (Nb. references)	1.202***	1.362***	1.180***	1.180***	1.320***	1.157***
	(0.135)	(0.261)	(0.136)	(0.138)	(0.253)	(0.140)
International collab.	0.322***	0.339*	0.320***	0.319***	0.301†	0.320***
	(0.040)	(0.172)	(0.038)	(0.042)	(0.178)	(0.039)
Share company affil.	1.275***	1.262**	1.289***	1.227***	1.082*	1.244***
	(0.150)	(0.480)	(0.154)	(0.144)	(0.464)	(0.150)
Nb. datasets	0.012	0.377**	0.015	0.005	0.371**	0.008
	(0.056)	(0.119)	(0.047)	(0.056)	(0.140)	(0.047)
Nb. tasks	0.007**	0.007**	0.007***	0.004***	0.003	0.004**
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Nb. modalities	0.192***	0.216	0.199***	0.294***	0.282	0.303***
	(0.045)	(0.189)	(0.047)	(0.047)	(0.194)	(0.048)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	28,393	1734	26,659	28,393	1734	26,659
Dependent variable mean	16.365	39.354	14.870	16.365	39.354	14.870
Pseudo R^2	0.41067	0.27869	0.42113	0.41498	0.28699	0.42552

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number scientific citations received by a paper. The response variables are indicator variables that are equal to one if a paper mentions CIFAR-10 or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2) and (3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2022, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D8. Robustness check: labeled datasets and patent citations – 3-years window

	Patents citations – 3 years window					
Model	Full (1)	2010–2014 (2)	2015–2022 (3)	Full (4)	2010–2014 (5)	2015–2022 (6)
CIFAR-10 (only)	0.431*	0.530*	0.409†	0.494*	0.894***	0.432†
	(0.188)	(0.232)	(0.215)	(0.196)	(0.218)	(0.224)
CIFAR-10 (others)	–0.078	–1.519†	0.053	0.005	–0.982	0.084
	(0.209)	(0.652)	(0.200)	(0.222)	(0.761)	(0.210)
ImageNet				0.216*	0.867**	0.081
				(0.107)	(0.326)	(0.119)
Log (Nb. authors)	0.530***	0.082	0.684***	0.523***	0.091	0.681***
	(0.103)	(0.246)	(0.121)	(0.104)	(0.243)	(0.121)
Log (Nb. references)	0.684***	1.831***	0.513**	0.665***	1.716***	0.506**
	(0.181)	(0.188)	(0.186)	(0.177)	(0.180)	(0.185)
International collab.	0.119	0.262	0.075	0.117	0.204	0.075
	(0.090)	(0.234)	(0.108)	(0.090)	(0.262)	(0.108)
Share company affil.	1.294***	2.944***	1.101***	1.264***	2.658***	1.091***
	(0.191)	(0.510)	(0.171)	(0.188)	(0.535)	(0.171)
Nb. datasets	0.034	0.368*	0.027	0.029	0.389*	0.025
	(0.105)	(0.126)	(0.103)	(0.104)	(0.133)	(0.103)
Nb. tasks	0.008**	0.010***	0.007***	0.006**	0.002	0.006**
	(0.002)	(0.003)	(0.002)	(0.002)	(0.004)	(0.002)
Nb. modalities	0.064	–0.592	0.160†	0.113	–0.483	0.178†
	(0.083)	(0.393)	(0.091)	(0.089)	(0.396)	(0.098)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	10,114	1579	8429	10,114	1579	8429
Dependent variable mean	0.31016	0.32552	0.31119	0.31016	0.32552	0.31119
Pseudo R^2	0.13808	0.26681	0.13424	0.13902	0.27749	0.13437

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number of patent citations received by a paper within 3 years of the publication year. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2019, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Table D9. Robustness check: labeled datasets and scientific citations – 3-years window

	Scientific citations – 3-years window					
Model	**Full (1)**	**2010–2014 (2)**	**2015–2022 (3)**	**Full (4)**	**2010–2014 (5)**	**2015–2022 (6)**
CIFAR-10 (only)	0.084	0.800**	0.063	0.217*	0.960***	0.196*
	(0.096)	(0.235)	(0.093)	(0.088)	(0.230)	(0.090)
CIFAR-10 (others)	–0.388*	–0.284	–0.367*	–0.249	–0.034	–0.227
	(0.174)	(0.452)	(0.170)	(0.211)	(0.483)	(0.205)
ImageNet				0.393***	0.473**	0.391***
				(0.102)	(0.180)	(0.102)
Log (Nb. authors)	0.465***	–0.039	0.517***	0.452***	–0.038	0.503***
	(0.097)	(0.108)	(0.103)	(0.097)	(0.107)	(0.104)
Log (Nb. references)	1.315***	1.514***	1.297***	1.292***	1.483***	1.274***
	(0.132)	(0.184)	(0.134)	(0.136)	(0.185)	(0.138)
International collab.	0.308***	0.269**	0.311***	0.310***	0.245*	0.315****
	(0.041)	(0.098)	(0.047)	(0.039)	(0.103)	(0.045)
Share company affil.	1.200***	1.365***	1.202***	1.149***	1.233***	1.152***
	(0.199)	(0.373)	(0.200)	(0.190)	(0.350)	(0.191)
Nb. datasets	0.089*	0.533***	0.072†	0.083†	0.532***	0.067
	(0.044)	(0.092)	(0.043)	(0.043)	(0.099)	(0.043)
Nb. tasks	0.006***	0.002	0.006***	0.002*	–0.002	0.003*
	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Nb. modalities	0.194*	0.118	0.205***	0.294***	0.165	0.307***
	(0.060)	(0.119)	(0.061)	(0.059)	(0.120)	(0.060)
Pub. venue type fixed effect	YES	YES	YES	YES	YES	YES
Subject area fixed effect	YES	YES	YES	YES	YES	YES
Publication year fixed effect	YES	YES	YES	YES	YES	YES
Observations	10,346	1734	8612	10,346	1734	8612
Dependent variable mean	21.480	11.685	23.452	21.480	11.685	23.452
Pseudo R^2	0.33006	0.32315	0.32350	0.33576	0.32905	0.32935

This table reports estimates of regressions of the models described in equations 1 and 2. The dependent variable is the total number of scientific citations received by a paper within 3 years of the publication year. The response variables are indicator variables that are equal to one if a paper mentions only CIFAR-10, CIFAR-10 among other datasets or ImageNet in the title, abstract, or keywords. Column (1) reports our baseline results of the estimates stemming from a Poisson regression. Columns (2 and 3) report estimates of the same equation in a subset of the sample comprised of papers published from 2010 to 2014 and those published from 2015 to 2019, respectively. Columns (4–5) report estimates when adding a dataset indicator variable also for papers using ImageNet. Exponentiating the coefficients and differencing from one yields numbers interpretable as elasticities. All the specifications include publication venue type, publication year, and scientific field fixed effects. Standard errors are clustered at the journal/conference level. Significance levels: $^{\dagger}P < 0.1$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.