# How Does Industry Affiliation of Academic Scientists Change the Rate and Direction of Research?*

Amir Sariri†

September 2024

**Abstract**

The implications of academic collaborations with industry have long been the nexus of contentious debate. The prevailing concern is that industry causes research to lose its fundamental depth and become commercially driven. This paper presents evidence that these concerns are misplaced for areas of research that have commercial value–that is, the specific domains for which these concerns are raised. The empirical analysis uses large-scale, manually improved bibliometric data from artificial intelligence research. For identification, I use the unexpected and significant success of the neural network techniques revealed at the ImageNet 2012 benchmark competition, which sharply increased the industry's demand for AI scientists, but more so for scientists with higher expertise in the breakthrough field. While industry affiliation significantly increases the usefulness of research, it does not diminish its novelty. In addition, industry affiliation raises both the publication and the quality of science produced by academic scientists. Results are consistent with the explanation that, at least in the medium term, academics in short supply can negotiate higher academic freedom while utilizing commercial resources for their research.

*Keywords*: Industry Affiliation, University-Industry Interface, Artificial Intelligence
*JEL Classifications*: O31, O32, O33

*For their insightful feedback, I thank Ajay Agrawal, Ashish Arora, Pierre Azoulay, Kevin Bryan, Avi Goldfarb, Joshua Gans, Ryan Hill, Matt Marx, Henry Sauermann, Jungkyu Suh, and Dror Shvadron. Financial support from the Royal Bank of Canada's Borealis AI Foundation is acknowledged.

†Purdue University Daniels School of Business

# 1 Introduction

With the increasing proximity of commercial technologies to cutting edge science, firms have risen to become producers of knowledge (Sohn, 2021; Zucker & Darby, 2014). Several studies have seen this as a diffusion of academic science to industry, an important impetus for the growth of technological innovation (Zucker *et al.*, 1998; Jaffe, 1989). Others, however, have cautioned against a deviation of research towards topics of commercial importance and delays in the dissemination of knowledge (Blumenthal *et al.*, 1996; Cohen *et al.*, 1998; Henderson *et al.*, 1998). Several studies have explored potential implications of university-industry interface for academic research in the context of academic entrepreneurship (Stuart & Ding, 2006), patenting of university inventions (Azoulay *et al.*, 2009), university-industry research centers (Cohen *et al.*, 1994), and collaborative research (Evans, 2010). The effect of industry affiliation of academics scientists with firms on the rate and direction of scientific production, however, has received barely any attention. This paper focuses on this issue using detailed hand-collected data on the industry affiliation of scientists in the field of artificial intelligence research.

The field of AI research is ideal for evaluating the effect of industry affiliation on research. Apart from the dramatic rise in the macro-level economic and social significance of AI, this field offers a unique opportunity for causal inference in the context of industry affiliation. In 2012, a team of researchers from the University of Toronto (hereafter AlexNet) demonstrated a step-function increase in the predictive power of the artificial neural networks (hereafter NN) approach to machine learning, causing a sharp increase in the demand for scientific labor from industry. However, this increase was higher for scientists whose line of research was closer to the successful technique and lower for those in distant fields.

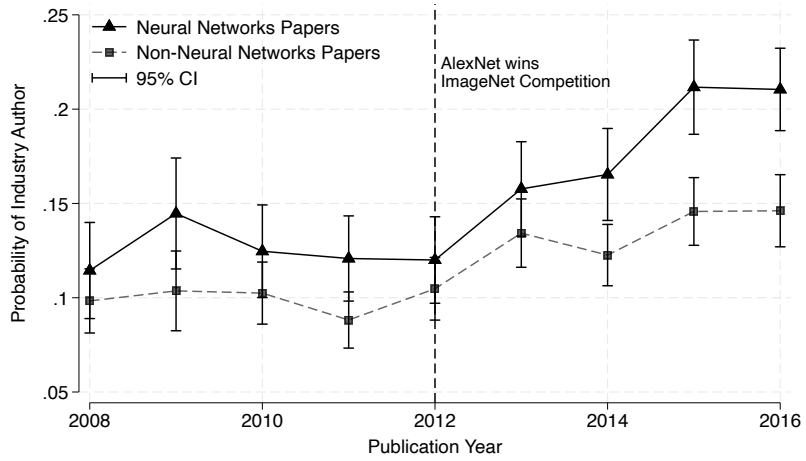
The figures in Figure 1 offer a visual summary of the impact of ImageNet 2012 on the university-industry interface based on publications in the top-5 conference proceedings in the field of AI. The top figure shows the fraction of authors affiliated with industry by whether the paper was NN or not. While industry is increasingly active in both domains, the rate of increase is significantly larger for NN research. The bottom graph plots the two-way fixed effect estimates of within-author changes

in industry affiliation. Specifically, it shows the change in the probability of scientists' industry affiliation before versus after 2012 and across scientists who had NN expertise before 2012 and those who did not. Compared to the base year of 2012, NN experts affiliated with industry 8 percentage points more, a 67% increase compared to the yearly average of scientists' industry affiliation, and nearly 80% compared to this yearly average before 2012. On the one hand, these steep increases may be viewed as promising signals of a rapid diffusion of knowledge to industry. On the other hand, policy makers, university administrators and the popular press express concerns about a shift in the production of scientific knowledge from basic to applied.

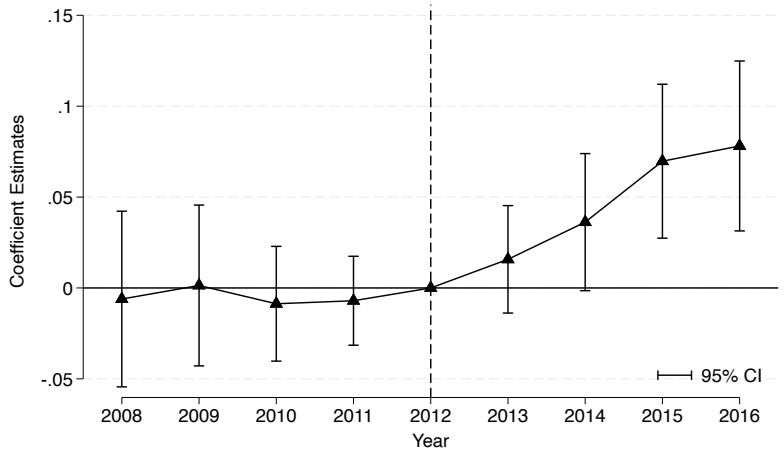
In the first set of results, the instrumental variable results show that industry affiliation increases scientists' paper out by 70%, however, the results appear to be driven mostly by authors who affiliate after the shock, with an average effect of only 11% boost in publications overall. In terms of the direction of research, I find that industry affiliation does not significantly decrease the novelty of research, though it dramatically boosts the commercial potential of the research. In terms of the latter, I find that industry affiliation causes a five-fold increases in the average annual patent cites to the affiliated scientists' publications. This suggests that the scarcity of NN experts in light of the sharp increase in industry's demand for their talent may have increased their ability to negotiate non-pecuniary benefits. In negotiating their compensation, researchers bargain with firms not only over the disclosure of their knowledge through publication (Stern, 2004), but also over their freedom in choosing the direction of their research. Therefore, increased demand for a fairly fixed supply of scientific experts may have paved the path to the affiliates' maintaining their academic freedom while utilizing superior commercial resources for research.

The main contribution of this paper is to add important nuance to the theoretical prediction that firms hire scientists in later-stage domains to direct research towards higher-payoff outcomes Aghion *et al.* (2008). My findings suggest that control-rights considerations is also a function of the labor market characteristics, and so industry affiliation may not always result in a push away from curiosity-driven work done in universities. In particular, merely attributing disclosure regimes to organizational structure as, say, university or industry (Nelson, 1959; Levin *et al.*, 1987) overlooks considerable differences in market conditions as well as the state of scien-

**Figure 1:** Disproportionate Change in NN Industry Research and Industry-Affiliation of Scientists with NN Expertise after ImageNet 2012




**(a)** Change Industry Publications of NN versus non-NN research




**(b)** Change in Industry Affiliation by pre-2012 NN Expertise

Notes: Figure 1a shows yearly average industry affiliation of authors on papers published in the proceedings of the top-5 AI conferences NeurIPS, AAAI, CVPR, ICLR, and ICML. Figure 1b shows the coefficient estimates $\boldsymbol{\beta}$ from the linear regression
$\text{Industry Affiliation}_{it} = \text{NN Expert}_i \times \mathbf{Year}_t\boldsymbol{\beta} + \gamma_i + \delta_t + \epsilon_{it}$, where standard errors are clustered by scientist. The coefficient estimates show within-scientist changes in the probability of affiliating with industry before versus after the shock and by whether the scientist was in the top quartile of NN expertise based on their publications between 2000 and 2012.

tific research being explored. Gans *et al.* (2008), for example, suggest this by noting industrial research may be amenable to the norms of open science. Therefore, my empirical findings support the idea that the quality of research by public and private sector can converge due to the underlying goals of investigators or research sponsors (Stokes, 1997; Geison, 1995), but also due to broader market-level factors and competitive forces.

The remainder of this paper proceeds as follows. In Section 2, I provide a brief overview of the literature. In Section Section 3, I provide a brief history of AI research on neural networks, the focus of my empirical setting. In this section, I also describe the ImageNet 2012 event that serves as an exogenous shock for my identification strategy. I describe data in Section 4 and my empirical strategy in Section 5. The main results are reported in Section Section 6.

## 2 University-Industry Collaboration and Research

An upswing in universities' commercial efforts has driven researchers to study the influence of opportunities to commercialize scientific research on the production of academic science. Most notably, academic patenting and its implications for the rate and direction of research have received significant attention (Agrawal & Henderson, 2002; Henderson *et al.*, 1998). In addition, scholars' pursuit of collaborative research with industry has gone from sporadic and frowned-upon to pervasive and encouraged. Such collaborations commonly transpire in four forms: academic entrepreneurship, patenting of university inventions, university-industry research centers, and collaboration on journal publications. This paper shares with this fourth track of research a focus on the individual scientists as the level of study.

Much of the policy debate on the university-industry interface has focused on the impacts of the two sectors' collaboration on the rate and direction of scientific inquiry. A small but growing number of studies have taken on the more specific question of how opportunities for commercialization or collaboration with industry may influence research itself. For example, Azoulay *et al.* (2009) report a moderate shift in the content of scientists' research towards questions of commercial interest and Evans (2010) reports similar results by showing that industry nudges academic

research towards more speculative and exploratory research. On the other hand, Bikard *et al.* (2018) suggest that in commercially important research areas, industry collaboration results in more follow-on papers and fewer follow-on patents than does purely academic research does. My findings concur with this conclusion but focus on the formal affiliation of academic scientists with industry, as opposed to collaborative research.

The classical view of industry's role in science is that hiring scientists enables controling the direction of scientific research (Aghion *et al.*, 2008). Scientists, on the other hand, bargain over norms of open science, such as the ability to publish their results and the type of research they are able to pursue (Stern, 2004; Gans *et al.*, 2008). A key input to the realized outcomes is the career motivations of scientists. Roach & Sauermann (2014) show that PhDs who have a strong preference for peer recognition and contribution to society have a particularly high 'price of publishing.' More broadly, however, the bargaining power of scientists hinges on the expected value of their knowledge to the firm as well as the market competition to capture this value. A highly sought-after PhD whose expertise is rare relative to the market demand may have no reason to accept a wage discount for the right to publish – it may be part of their reservation wage. Furthermore, scientific talent who are pursuing research careers in industry but presumably receive competing offers need to evaluate their options. Thus, firms may indeed need to publish high-quality research that appeals to the academic talent in order to attract and retain the best and brightest from universities (Rosenberg, 1990; Simeth & Raffo, 2013).

These arguments imply that science experts who are in short supply may be particularly well positioned to benefit from collaborating with industry. The desired natural experiment to establish this proposition is a scientific breakthrough that signals significant commercial value for a technique that is not expected to be successful, suddenly raising commercial sector's demand for some talent but not others. Although the supply will eventually adjust and the scientists' bargaining power will dissipate, the equilibrium condition is likely to last in the medium term given the lengthy process of training new PhDs. Switching fields is also highly expensive and often requires learning by collaboration with those trained in the target specialty (Zucker *et al.*, 1998). The next section describes how features of research in artificial

intelligence provides a natural experiment comparable to the one described.

# 3 Neural Networks: from the 1940s to date

In 1943, Warren McCullock and Walter Pitts first proposed a neural network model for mathematical computations at the University of Chicago. Their creation could model the basic OR/AND/NOT functions but lacked a mechanism for learning. In 1949, Donald Hebb, a professor at McGill University, theorized in his influential book "The Organization of Behavior" that learning occurs through changes in the synaptic connectivity of the human brain. In 1957, Frank Rosenblatt, a psychologist at Cornell University inspired by the Hebbian theory of learning, advanced McCulloch-Pitts' neural network model and conceived the idea of Perceptrons: he proposed a linear regression-like model for enabling artificial neurons to *learn* (Rosenblatt, 1957). This led to the construction of *Mark I Perceptron* in 1960 (Figure 2), the first computer that could learn and classify simple shapes. The impression that this model of neural network could simply be expanded to solve difficult problems of AI led to an absurd hype that came to an end in the early 1970s with major cuts to and the freezing of research funding, a period widely known in the popular culture as the *First AI Winter*. This downturn is widely ascribed to the 1969 book "Perceptrons: An Introduction to Computational Geometry" by Marvin Minsky and Seymour Papert of MIT (Minsky & Papert, 1988), who showed that Rosenblatt's approach to AI could not learn the simple XOR function, meaning that it was a dead end and should be abandoned.

Among others, Paul Werbos, a PhD candidate at Harvard, dealt with the issue raised by Minsky and Papert in his 1974 PhD thesis. He suggested that the operation termed *backpropagation* could train multilayer neural network (Werbos, 1974) - note that "neurons" in artificial neural network is a generalization of Rosenblatt's idea of Perceptrons. Approximately a decade later, David Rumelhart and Ronald Williams from UC Berkeley and Geoffrey Hinton from Carnegie Mellon University published two papers that popularized backpropagation and brought neural network back to life (Rumelhart *et al.*, 1986b,a). A year after this paper was published, Hinton moved to Canada, and later convinced the Canadian Institute for Advanced Research (CIFAR) to fund his research. Two years later, an influential paper by researchers at Technische

**Figure 2:** Mark I Perceptron at the Cornell Aeronautical Laboratory

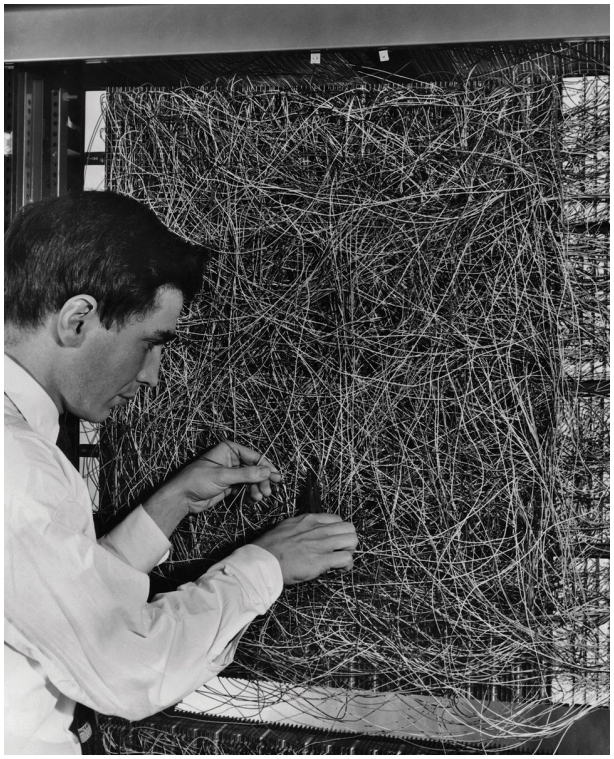
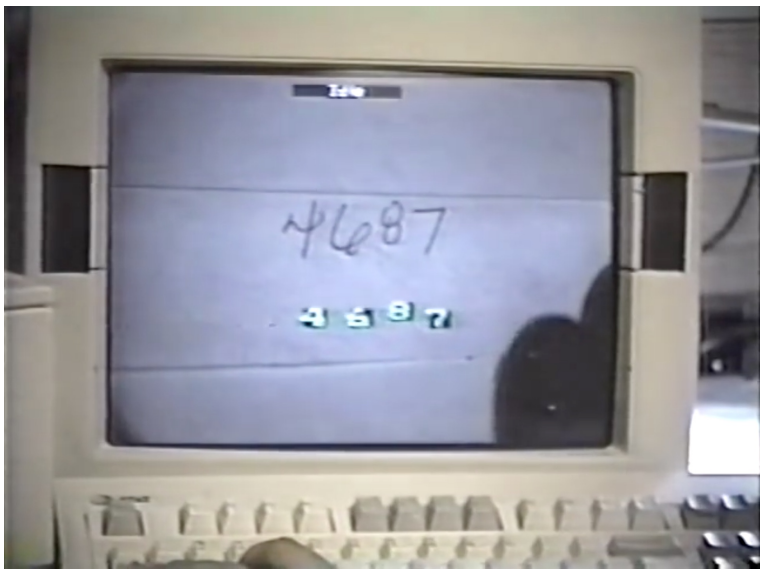


**Figure 3:** Yann LeCun demonstrating the first convolutional network that could recognize handwritten digits in 1993.



Universität Wien and UC San Diego showed that multilayer neural network could be used to approximate any mathematical function (Hornik *et al.*, 1989) - that is, neural network are a sort of universal approximator. In the same year, Yann LeCun and his colleagues at Bell Labs successfully applied backpropagation with a modified architecture for neural network called *Convolutional neural network (CNN)* to the recognition of handwritten zip code digits provided by the U.S. Postal Service (LeCun *et al.*, 1989) (Figure 3). This algorithm later became the core engine of the nationally deployed cheque-reading systems.

Regardless, by the late 1980s it was clear that backpropagation could not train a different model of neural network called *Recurrent neural network (RNN)*, which was useful for machine learning applications with variable inputs and variable outputs, such as speech recognition. Although in 1997 two professors introduced a method that made training RNNs possible (Hochreiter & Schmidhuber, 1997), lack of sufficient compute power to train these models fast enough and an overall loss of the scientific community's faith in neural network brought about what is known as the *Second AI Winter*. One of the highlights of this period is that almost all papers with the subject of neural network were routinely rejected by top-tier journals and for conference

proceedings.

Hinton was, however, able to convince CIFAR to fund his new program on neural computation during this time when the "mainstream machine learning community could not have been less interested in neural nets."[1] Soon after, Hinton and two of his students, Simon Osindero and Yee-Whye The, published a breakthrough paper that popularized what became known as deep learning (Hinton *et al.*, 2006). Although this and subsequent empirical papers on deep learning showed impressively low error rates in image recognition on a standard image dataset called MNIST, simpler *shallow* techniques from the late 90s had already shown similarly low error rates of approximately 1%. Thus, it was difficult to market the importance of deep learning with high performance on the MNIST dataset only. Scientific benchmarks that are widely used by researchers in this field is itself an under-explored but important characteristic of this field.

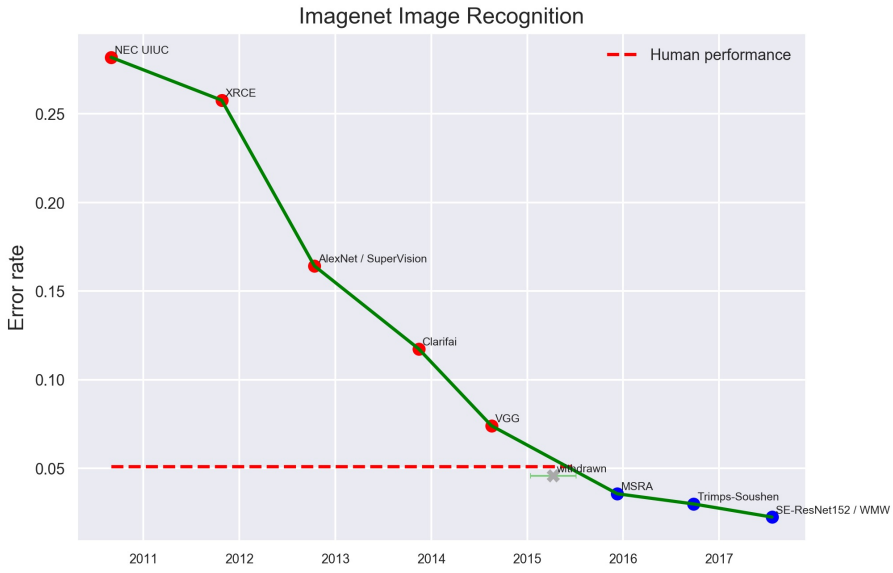
A few years later, Andrew Ing of Stanford University showed that Graphical Processing Units (GPUs) are 70 times faster for training deep neural network than are conventional CPUs (Raina *et al.*, 2009). With the creation of very large neural network trained with GPUs, better performance results followed. Shortly after, graduate students who worked with Hinton on speech recognition using deep learning and GPUs went on to internships with Microsoft (2009) and Google (2011), where they helped to replace both companies' existing speech recognition technology infrastructures with deep neural network.[2] This collaborative work between Hinton's group at the University of Toronto and research groups at Microsoft, Google, and IBM was summarized in a paper by 11 authors from all four groups (Hinton *et al.*, 2012). As well, in 2011, Andrew Ng, along with Google's computer scientists, formed the Google Brain Project, which remains one of the largest and most productive industry research labs in the field of machine learning.

Finally, in late 2012 Hinton and two of his graduate students Alex Krizhevsky and Ilya Sutskever, applied deep convolutional neural network with GPUs to a benchmark image recognition and classification competition called Large Scale Visual Recog-

[1]Kate Allen. How a Toronto professor's research revolutionized artificial intelligence, Science and Technology reporter, Apr 17 2015. http://bit.ly/1cEyL8R

[2]Leading breakthroughs in speech recognition software at Microsoft, Google, IBM. University of Toronto News. November 26, 2012. http://bit.ly/2hL9K40

**Figure 4:** Geoffrey Hinton, Alex Krizhevsky and Ilya Sutskever won the ImageNet Competition with AlexNet algorithm in 2012.




nition Challenge (widely known as the ImageNet Competition) and won the competition with a 15.3% error rate, over 70% better than the second closest error rate that year (Figure 4). My interviews with machine learning scientists from both academia and industry suggest that the work by Hinton et al at the ImageNet Competition was the beginning of the influx of industry's interest in machine learning research that ensued.

# 4 Data

Data is a corpus of academic publications in the large umbrella field of artificial intelligence research. Specifically, I collect all papers published in the proceedings of five top AI conferences from 2000 to 2016. I use academic conference proceedings instead of academic journals because, as a norm due to rapid progress int he field, the bulk of research in AI appears in proceedings. These conferences are International Conference on Machine Learning (ICML), the International Conference on Learning Representations (ICLR), the Conference on Neural Information Processing Systems (NeurIPS), the main conference of the Association for the Advancement of Artificial

Intelligence (AAAI), and the Conference on Computer Vision and Pattern Recognition (CVPR). These conferences are identified based on interviews with 15 leading AI scientists, who broadly viewed them as *high impact* and *academically oriented*. Roughly half of the consulted scientists are faculty members at universities and half are from industry.

To capture changes in the direction of research, I use a measure of novelty described in Section 5. I also a more established and accurate measure of commercial potential. This measure is based on the number of patent-to-paper citations, which I obtain by linking each paper to the Reliance of Science database (Marx & Fuegi, 2020, 2022).

To ensure accuracy, I hand-collect the titles of the papers directly from conference proceedings published online. This process yields 18,636 papers. Then, I search each title in the Microsoft Academic Graph (MAG) to collect additional bibliometric data (Sinha *et al.*, 2015).[3] At the time of data collection, MAG contained over 200 million publications, offering a well-documented application programming interface (API) for data collection, and an extensive coverage of computer science conference proceedings, which significantly facilitates data collection. However, MAG has several serious limitations in terms of coverage and accuracy that I addressed by hand-collecting data from other sources such as Google Scholar, CiteSeerX, arXiv, DBLP Computer Science Bibliography, Semantic Scholar, and personal webpages of the individual scientists, faculty, and corporate labs.

Another major sources of my bibliometric data is OpenAlex (Priem *et al.*, 2022) (hereafter, "OA"). In May 2021, Microsoft announced that MAG will be retired on December 31, 2021. At this time, a nonprofit organization called OurResearch launched the OA project to continue and improve on the services previously provided by MAG. While MAG serves as an important source of the OA's research catalogue,

[3]I use Microsoft Academic as my primary source for collecting basic information on conference proceedings of NeurIPS, ICML, ICLR, AAAI, and CVPR (Sinha *et al.*, 2015). To do this, I use Microsoft Academic Knowledge Exploration Service (MAKES) Semantic Interpretation API and Entity API. These two APIs provide access to a variety of schemas from which I collect the following attributes for each paper: title, paper ID, conference series name, authors' names, authors' affiliations, fields of study, citation count, publication year, references, and the link to the PDF file of the publication. For more information, visit https://aka.ms/msracad.

OA has greatly increased the quality and coverage of its data by expanding its sources to other major data providers such as Crossref, a DOI registration with data on 150 million scientific articles. In addition, the OA data has made important improvements in disambiguating author names and using advanced natural language processing algorithms to assign topics to papers. To utilize these enhanced features, I also collect the bibliometric data by OA on each of the conference papers, and when appropriate, reconcile this information with the MAG data to maximize quality. In the remainder of this section, I will note if any merging or data reconciling between the two sources have taken place.

## 4.1 Identifying authors and affiliations

Unlike the ISI Web of Science that lists first initial, a middle initial, and the last name of authors (Agrawal *et al.*, 2017), MAG automatically indexes the full name of authors. Nonetheless, it is reasonable to assume there are still errors in such automatic operation, particularly in assigning unique identifiers to authors due to name ambiguity. As a first step to disambiguate author names, I cluster author names whose first and last two name characters are identical. In each cluster, I manually check each pair of names and cross check names with the papers from which they have been collected.[4] I also conduct a battery of additional tests to identify unusual author attributes that may point to name ambiguity. For example, a name ambiguity was detected and rectified by a test that tagged each author with more than four affiliation changes during the study period. The final step in name disambiguation was to identify and resolve conflicts between these author names and those provided by OA. Before any sample attrition, this process yields 22,065 authors.

For institutional affiliation of scientists, I again started with MAG. MAG indexes the institution affiliation of authors by automatically scanning the front page of papers. This automatic system, however, is prone to error. For example, a common

[4]For example, a variation of the name Yann LeCun is Yann Le Cun, which appears in the data as two distinct names. The first two and last two characters of the names, *Ya* and *un*, match. Thus, the clustering algorithm puts them together for a manual correction. Increasing the threshold to a three-character match from each end may result in a false negative error that is too high, and lowering the threshold to a one-character match from each end will result in clusters that are too large and infeasible to check manually.

error identified in the manual inspection of data is that MAG appears to commonly captures the affiliation of the author whose name appears spatially adjacent to that of the focal author on the front page. This is particularly the case when there are several authors or when author names are printed relatively close to each other. While OA provides significantly enhanced institutional affiliation data, here I also detected systematic errors during my manual inspections. A particularly salient error occurs when the affiliation of authors is incorrectly specified in open-source repositories, or the paper is simply not included in such a repository. This is because OA indexes author names by crawling digitized information on available databases, thus necessarily inheriting errors and missing information.

Given industry affiliation is the main objective of interest in this paper, and the low confidence in available data, I decided to gather this information manually by viewing the front page of each paper and recording each author's institutional affiliation, or affiliations if they listed more than one institution under their name. While quite costly, it appeared to be the only viable path to gather reliable information. Next, in order to collect a unique list of institutions, I manually map all research centers, schools, and departments within larger entities to their corresponding institutions. For example, I map Brookes Vision Group to the University of Oxford and Wyss Institute for Biological Inspired Engineering to Harvard University. Finally, I label each institution as industry if the institution is for-profit entity of any kind. Most institutions can be easily assigned a label since they have clear identifiers of type in their names such as "University," "School of," or "Inc." For all other cases that are not obvious, I manually search and label the institution. Before any sample attrition, this process yields 1,484 unique institutions, with 15% labeled as industry.

## 4.2 Field of Study

The conference proceeding papers in my data do not list keywords, which are an essential step towards differentiating between NN and and non-NN papers, as well as developing secondary measures of scientists' expertise in NN based on their publications. MAG provides automatically generated topics using NLP predictive algorithms called hierarchical topic modeling. This model generates a list of keywrods associated with each paper, but my manual inspections revealed high variability in the quality

and detailedness of the keywords generated. Here lies another major improvement made possible by OA. Specifically, OpenA builds on a more accurate clustering algorithm that assigns topics by taking into account a multitude of factors beyond the title and abstract of the paper, but also including the paper's citation network and journal. More importantly, the model is trained on a labeled dataset with micro-level field assignment to papers. A key feature of OA topics is that, unlike MAG, each paper is assigned topics at a comparable level of detail, and then hierarchically mapped to subfields, fields, and the overarching scientific domains. For each paper, I gather the lowest-level topic assigned to it by OA.

While a significant step forward, the topics alone do not suggest to what extent each paper is relevant to the NN specialty. While some topics (e.g., "Neural Network Fundamentals and Applications") can be clearly identified as NN, other topics (e.g., "Advances in Transfer Learning and Domain Adaptation") may not be immediately obvious as strongly related to NN. A manual inspection is not feasible. With 543 unique topics in my data, a significant breadth of knowledge is required to objectively assign a relevance score to each topic. Fortunately, assigning a relevance score is the sort of task that is quite amenable to the novel large language models. To do this, I use the most advanced commercial GPTs' available in 2024, OpenAI's ChatGPT4o and Anthropic's Claude Opus 3. In both cases, I prompt the language model with the list of unique topics in my data and the command to assign a score from 0 to 1 to each topic based on the extent to which the topic builds on or contributes to deep learning. The two scores are highly correlated, though my manual checks show that Claude Opus performed slightly better.

Given only a minority of topics are scored as highly related to NN–69 out of 543 topics, I manually checked for type I error and found no errors. To assess type II error, I used a fuzzy match algorithm to detect a set of 14 scientific keywords that strongly indicate an NN paper in the title of each paper. I then compared the Claude Opus probabilities with this highly mechanical approach and found that in only 40 papers the Claude Opus assigned a score less than 0.9 to the topic of a paper with keyword match probability of more than 0.9. An inspection of these 40 papers revealed that, in most cases, the paper's topic was correctly assigned by OA (e.g., research related to meteorology), but a low NN-relevance probability was assigned

because the paper was an application of NN to problems in a disparate field (e.g., "Advanced Techniques in Reservoir Management"). Replacing the Claude score with the keyword match score in these 40 cases result in the measure NNRelevance. Each paper is labeled as NN, if it has an NNRelevance score of 0.9 or higher.

# 5 Empirical Strategy

This section describes measures and the empirical strategy used to estimate the main results. There are two sets of outcomes analyzed, pertaining to the rate and direction of research. For explanatory variables, the two main types of variable used identify industry affiliation and neural network expertise of the scientists. The strategy for causal identification of the effect of industry affiliation is an instrumental variable approach that exploits the sudden increase in private sector's demand for scientific labor after 2012 which varied based on scientists' existing expertise in NN methods. The remainder of this section describes the methodology used for measures and identification in more detail.

## 5.1 Outcomes: Rate & Direction

**Rate of Research:** For changes in the rate of research I use two established measures of academic productivity, published papers and citation-weighted publications. Since the analysis is done at the author-year level, the measure of publications calculates the yearly average papers produced by each scientist. An issue with using publication counts is that because scientific fields differ in their baseline productivity, effect sizes would be difficult to interpret as large or small. Therefore, I also normalize the yearly average papers by each scientist. A key consideration in normalizing publications is that temporal shocks such as changes in public funding for research can shift baseline productivity across years. To account for this fact, I normalize the publication counts by year.

The primary shortcoming of using raw publication counts is that it lacks information about the quality of the knowledge produced. The de facto approach to account for quality is to use citation-weighted publications. However, this measure

comes with its own problems. In particular, the number of academic cites received as a measure of quality is inherently biased due being right-truncated. When citation information is gathered at a specific point in time, the econometrician does not observe the citations that a later-published paper would have received, had she waited just as long after publication as an earlier-published paper. One remedy is to only count cites received by a given paper within a fixed period of time after the paper is published. This approach requires ensuring each of the backward cites contain reliable bibliometric information, which can become infeasible in fields such as AI that receive hundreds of thousands of references from across disciplines. This remedy also fails to account for the fact that field-level discovery shocks that raise references can happen at a time outside of the fixed window selected.

Due to these challenges, I use the alternative measure of normalizing citations received by year. Specifically, I calculate each scientist's total yearly cites, then normalize this number across all scientists within the same year.

**Direction of Research:** The two measures used to assess the direction of research are commercial potential and novelty. As described in Section 4, the measure of commercial potential is based on the number of citations made by patents to academic papers (Marx & Fuegi, 2022, 2020). This measure is similar in nature to academic cites, thus bearing the same right-truncation issues described earlier. Therefore, I tackle these issues similarly by normalizing patent references to authors' papers by year. In particular, I calculate each scientist's total patent cites received in a given year, then normalize across all scientists' patent cites within that year.

Measuring novelty in research is challenging and critically depends on how it is interpreted. For example, one can view the number cites, the measure I use for capturing the rate of research, as an indication of novelty. Such a definition presumes novelty as research that is both unique and useful. In this paper, I use a more constrained definition of novelty as research that is distinct in recombining existing ideas. Specifically, I index each paper's references as proxy for the underlying knowledge that institute the focal paper's ideas, then calculate the extent to which the recombination used is atypical compared to that of other papers.

Using backward cites to capture the nature of ideas encoded in scientific papers has a long tradition in social science studies of science and innovation. For example. For example, Azoulay *et al.* (2007) show that it is possible to use bibliometrics data as a "proxy for inter-person differences in the level of commercial orientation of scientific research." Papers with similar references not only draw from the same composite of prior knowledge, but also are highly likely to contribute to the same literature. More closely related to the present paper, I build my measure of novelty on Uzzi *et al.* (2013) who use papers' mix of references as a proxy for novelty and conventionality of the content of research and show that both novelty and conventionality are associated with a higher probability of a "hit paper."

Let $T$ be the set of years in the dataset, $J_t$ be the set of papers published in year $t$, and $R_j$ be the set of references in paper $j$. The frequency $f_{r,t}$ of a reference $r$ in year $t$ is calculated as:

$$f_{r,t} = \sum_{j \in J_t} \mathbb{1}(r \in R_j)$$

where $\mathbb{1}(r \in R_j)$ is an indicator function that is 1 if $r$ is in $R_j$, and 0 otherwise.

Next, the reference co-occurrence frequency $c_t(r_1, r_2)$ of a pair of references $(r_1, r_2)$ in year $t$ is:

$$c_t(r_1, r_2) = \sum_{j \in J_t} \mathbb{1}(r_1 \in R_j \wedge r_2 \in R_j)$$

where $\mathbb{1}(r_1 \in R_j \wedge r_2 \in R_j)$ is an indicator function that is 1 if both $r_1$ and $r_2$ are in $R_j$, and 0 otherwise.

Then, the novelty score for a paper $j$ published in year $t$ is calculated based on the average frequency and co-occurrence frequency of its references. The steps are as follows:

1. Calculate the average reference frequency $\bar{f}_{j,t}$:

$$\bar{f}_{j,t} = \frac{1}{|R_j|} \sum_{r \in R_j} f_{r,t}$$

2. Calculate the average co-occurrence frequency $\bar{c}_{j,t}$:

$$\bar{c}_{j,t} = \frac{1}{\binom{|R_j|}{2}} \sum_{(r_1,r_2)\in R_j} c_t(r_1, r_2)$$

where $\binom{|R_j|}{2}$ is denotes the number of possible pairs of references in paper $j$.

3. Calculate the novelty score $N_{j,t}$:

$$Novelty_{j,t} = \frac{1}{\bar{f}_{j,t} \cdot \bar{c}_{j,t}}$$

where, by design, the novelty score is standardized to the range of 0 to 1.

## 5.2 Estimation Strategy

The baseline regressions are:

$$y_{it} = \beta \text{Industry Affiliation}_{it} + \gamma_i + \delta_t + \epsilon_{it}$$

where $y_{it}$ is one of the four main measures of the rate and direction of research, $\gamma_i$ and $\delta_t$ are individual scientist and year fixed effects, and Industry Affiliation is an indicator that equals one for the first year in which the scientist affiliates with industry and all subsequent years. In my sample, this measure is equal to an indicator that equals to 1 if the author has at least one industry affiliation in each year. This is because no author stops affiliating with industry once they do it for the first time. Put differently, no author who moved to industry ever publishes again in subsequent years without indicating industry affiliation in at least one of their publications in that year.

**Instrumental Variable:** The instrumental variable strategy relies on the assumption that the industry demand shock to the scientific labor was higher for scientists with higher expertise in the breakthrough technique, neural networks. To construct a measure of expertise, I use information about the proximity of each scientist's pre-2012 publications to neural network research. Specifically, I transform the NNRelevance variable described in Section 4 in three steps to produce a scientist-level measure of NN expertise.

$$\text{NNExpertise}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left( \frac{\text{NNRelevance}_{ij}}{a_j} \times (t_j - 1999) \right)$$

where $n_i$ is the number of pre-2012 publications by scientist $i$, $a_j$ is the number of authors on paper $j$, and $t_j$ is the publication year of paper $j$.

The first transformation is based on the consideration that to the extent publications signal corporate labs about a given scientist's expertise in a given field, those signals are attenuated by the number of authors on the paper. In the absence of an objective measure of the level of each author's contribution, I divide the NNRelevance of each paper by its count of authors. Second, it is also reasonable to assume that the more recent publications to the shock provide more precise information about one's real research interests. A scientist may have published a paper with high NNRelevance, but they may have long switched topics. This is especially likely given before the 2012 shock NN was viewed by the majority of the scientific community as unpromising path to AI. To account for this recency effect, I further linearly weight each paper's NNRelevance by how close the published year of the paper is to the shock year of 2012. To maximize available information, I use the full history of papers in my data between 2000 and 2012. Lastly, I calculate the continuous measure of NNExpertise as the within-scientist average of the resulting weighted NNRelevance metric described.

Industry Affiliation, is then instrumented with:

$$\text{IV}_{it} = \mathbb{1}(t > 2012) \times \text{Expertise}_{it}.$$

The IV strategy in this paper assumes that the authors' expertise in neural networks

# 6 Results

This section shows results in two parts. The first set of results show the effect of industry affiliation on the rate of research based on publications and citation-weighted publications. The second set of results show the effect of industry affiliation on the direction of research using the measure of novelty and commercial potential.

**Table 1:** Industry Affiliation and Publications

| DepVar: | Pubs | | | Normalized Pubs | | |
|---|---|---|---|---|---|---|
| | (1-1) | (1-2) | (1-3) | (1-4) | (1-5) | (1-6) |
| Industry Affiliation | 0.097* | 0.305*** | 0.187*** | 0.054 | 0.153*** | 0.150*** |
| | (0.051) | (0.063) | (0.064) | (0.041) | (0.050) | (0.050) |
| Mean of DV | 1.692 | 1.692 | 1.692 | −0.000 | −0.000 | −0.000 |
| Scientist FE | | X | X | | X | X |
| Year FE | | | X | | | X |
| *N* | 11,371 | 11,371 | 11,371 | 11,371 | 11,371 | 11,371 |

Notes: This table shows the relationship between the scientist's move to industry and their yearly average publications. Standard errors clustered by scientist in parentheses. Statistical significance is *(10%), **(5%), or ***(1%).

## 6.1 Rate of Research

Table 1 shows that the change in industry affiliation is correlated with an increase in the number of publications. For brevity, I will focus on the results in Columns 3 and 6, the specification that includes both scientist and year fixed effects. Column 3 shows that affiliating with industry is associated with 0.187 more publications per year or 11% increase on a base yearly average publication of 1.69. Alternatively, Column 6 shows that this increase is equivalent to $0.15\sigma$.

While these effects are large and significant, they do not account for the quality of publications. To account for quality, results in Table 2 shows estimates using the same specification, except for the dependent variable being the average yearly citation-weighted publications. Column 3 shows that after accounting for time-invariant author qualities and transitory shocks, industry affiliation is associated with 143 more cites per paper, an 83% increase over the base yearly average. The estimate in Column 6 shows that this effect translates to $0.26\sigma$, which together with the results in Table 1 suggests that those who move to industry not only publish more papers, but also publish higher quality ones.

While suggestive, these results do not account for scientists selection into collaborating with industry. The key concern here is that the unobserved qualities of

**Table 2:** Industry Affiliation and Citations

| DepVar: | Cites | | | Normalized Cites | | |
|---|---|---|---|---|---|---|
| | (2-1) | (2-2) | (2-3) | (2-4) | (2-5) | (2-6) |
| Industry Affiliation | 143.438*** | 121.133** | 143.523*** | 0.307*** | 0.247*** | 0.259*** |
| | (27.147) | (55.014) | (53.926) | (0.051) | (0.081) | (0.080) |
| Mean of DV | 171.579 | 171.579 | 171.579 | 0.000 | 0.000 | 0.000 |
| Scientist FE | | X | X | | X | X |
| Year FE | | | X | | | X |
| *N* | 11,371 | 11,371 | 11,371 | 11,371 | 11,371 | 11,371 |

Notes: This table shows the relationship between industry affiliation and their yearly average citation-weighted publications. Standard errors clustered by scientist in parentheses. Statistical significance is *(10%), **(5%), or ***(1%).

the scientists drive both their productivity as well as higher demand from industry to collaborate with them. The instrumental variable estimates shown in Table 3 aim to tackle this issue. As discussed in Section 5 the instrumental variable strategy works by exploiting the unexpected success of the neural networks approach to machine learning, which differentially increased industry's demand for scientists with higher neural networks expertise. All estimates include the full set of fixed effects.

Column 1 in Table 3 shows the first-stage estimates along with the weak identification test of the instrumental variable. Given I do not assume IID errors, the usual Cragg-Donald Wald test is invalid ($F$-statistic = 70), so instead I show the $F$-statistic of the cluster-robust Kleibergen-Paap Walk test, which is larger than the rule-of-thumb value of 10. In addition, for each of the second-stage results, I show the $p$-value of the Anderson-Rubin Wald test, which in the case of an exactly identified model, provides weak-instrument robust test of the significance of the endogenous regressor in the structural equation.

The second-stage results in Columns 2 and 3 show that yearly average publications and annually normalized publications increase by 72% and $1.04\sigma$, respectively. Similarly, Columns 4 and 5 show that yearly average citation-weighted publications and its normalized analogue increase by nearly a factor of 3 and $1.04\sigma$, respectively. In all cases, the IV results are statistically significant, though also meaningfully larger

**Table 3:** The Effect of Industry Affiliation on the Rate of Research

| | 1st Stage | 2nd Stage | | | |
|---|---|---|---|---|---|
| DepVar: | (3-1) Industry Affiliation | (3-2) Pubs | (3-3) Normalized Pubs | (3-4) Cites | (3-5) Normalized Cites |
| Post2012 × NN Expert | 0.030*** | | | | |
| | (0.006) | | | | |
| Industry Affiliation | | 1.216** | 1.037** | 503.449** | 1.040** |
| | | (0.566) | (0.443) | (231.503) | (0.457) |
| Mean of DV | | 1.692 | −0.000 | 171.579 | 0.000 |
| Scientist FE | X | X | X | X | X |
| Year FE | X | X | X | X | X |
| Kleibergen-Paap Wald F | 29 | | | | |
| $p$-value Anderson-Rubin Wald $\chi 2$ | | 0.026 | 0.013 | 0.030 | 0.021 |
| $N$ | 11,371 | 11,371 | 11,371 | 11,371 | 11,371 |

Notes: This table shows 2SLS IV estimates of the effect of affiliating with industry on yearly average publications and citation-weighted publications. Standard errors clustered by scientist in parentheses. Statistical significance is *(10%), **(5%), or ***(1%).

than the OLS estimates. The effect size differences suggest a positive correlation between Industry Affiliation and the error term, reinforcing the intuition that higher-quality scientists are more likely to select into collaborating with industry labs. One explanation is that these scientists may face higher labor demand from companies.

## 6.2 Direction of Research

As described in Section 5, I use two measures of direction, novelty and commercial potential, where the latter is proxied by the number of patent-to-paper citations. Before turning to numerical estimates, it is demonstrative to assess how the direction of research overall changed after the ImageNet 2012 event, irrespective of the activity at the university-industry interface. Figure 5 offers two graphical summaries of field-level changes in direction. In each, I plot two-way fixed effects estimates of changes in the measure of research direction before versus after the 2012 shock, and across neural network and non-neural network papers. Specifically, for each, I report estimates in $\boldsymbol{\beta}$ from the linear regressions of the form:

$$y_j = \alpha \text{Neural Network}_j + \text{Neural Network}_i \times \mathbf{Year}_j \boldsymbol{\beta} + \mathbf{Year}_j \boldsymbol{\gamma} + \epsilon_j.$$

where $\text{Neural Network}_j$ is an indicator for whether paper $j$ is a neural network paper–

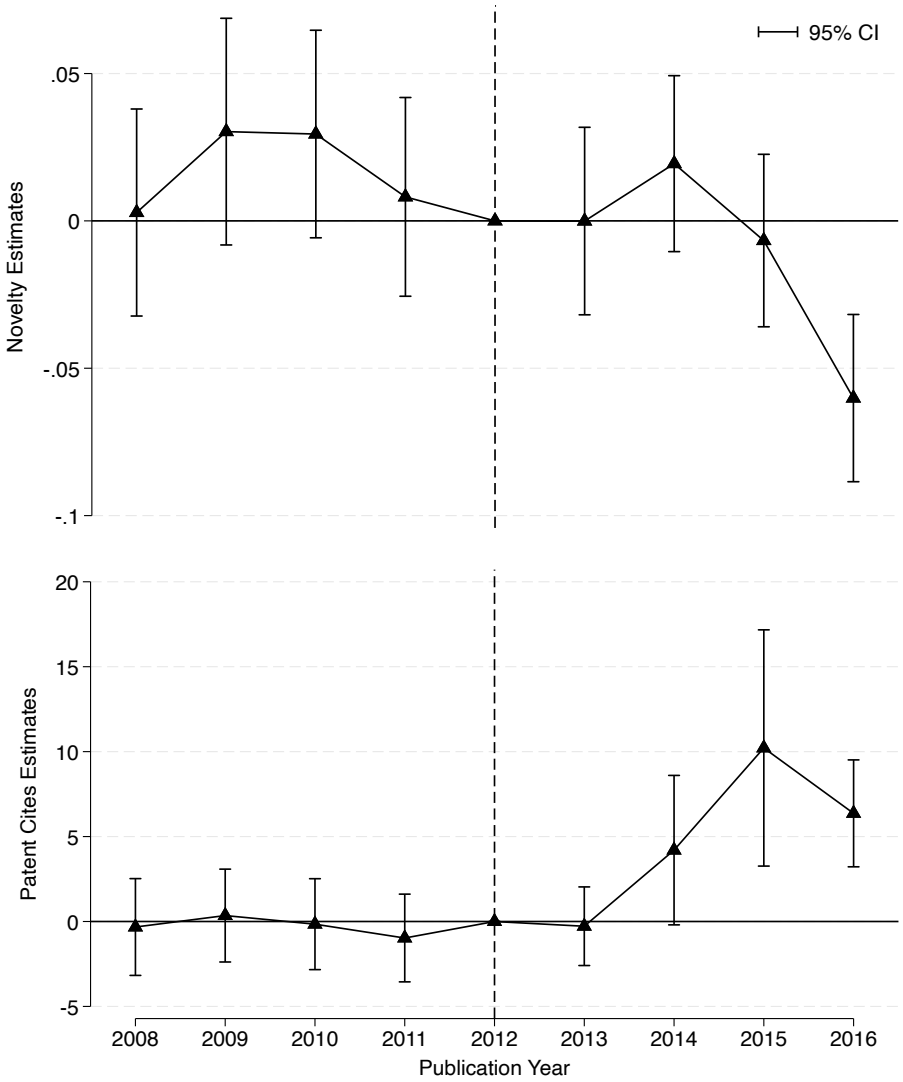
that is, the NNRelevance measure is 0.9 or higher, and $\mathbf{Year}_j$ is a vector of indicators for its year of publication.

The top plot of Figure 5 shows a sharp drop in the novelty of neural network papers compared to publication on other research topics after the ImageNet 2012 event. Conversely, the bottom graph shows a comparably large increase in the commercial potential of neural network publications, as measured by the number of in-patent cites to the papers. At the surface, these figures suggest that neural network research has become less novel and more applied, which is aligned with the type of popular and policy concerns noted earlier. However, it is important to note that while these results show field-level changes in the direction of research, they do not suggest that these changes are driven by industry collaboration of academic scientists.

Table 4 shows the OLS and IV estimates of the effect of industry affiliation on the measures of direction of research. Comparing the first two columns show that while the OLS estimates are significant and negative, the 2SLS IV estimate are insignificant, corroborated with the Anderson-Rubin Wald test of the significance of the endogenous regressor. The next two columns show the same results with dependent variable equal to novelty score normalized by year. While the IV results are insignificant, the correlation in Column 3 shows that those who affiliate with industry demonstrate a $0.1\sigma$ reduction in the novelty of their research. Overall, these results do not provide substantive evidence of a change in the novelty of research as a result of industry affiliation, though it shows that those who affiliate produce, on average, less novel research.

The remaining results focus on the effect of industry affiliation on the commercial potential of research, as measured by the yearly average patent cites received by scientists' papers. For brevity, I will focus on the IV results shown in Columns 6 and 8. Column 6 shows that industry affiliation increases the scientists' patent cites by 22. Compared the baseline average of 4.5 patent cites per year, this increase is equivalent to a staggering 500% rise in patent citations. Alternatively, results using normalized patent cites as dependent variable show that industry affiliation increase patent citations by $1.06\sigma$. Note also that the IV estimates are meaningfully larger than the OLS results. This is in part expected because the IV estimates show the local treatment effect for scientists who affiliate with industry, whereas OLS provides the average

**Figure 5:** Disproportionate Change in the Direction of Neural Network Research after the ImageNet 2012 Competition




Notes: These figures show the TWFE estimates of changes in the novelty (top graph) and commercial potential (bottom graph) of research before and after the ImageNet 2012 shock, and across NN and non-NN publications.

**Table 4:** The Effect of Industry Affiliation on the Direction of Research

| DepVar: | Novelty Score | | Normalized Novelty Score | | Patent Cites | | Normalized Patent Cites | |
|---|---|---|---|---|---|---|---|---|
| | (4-1) OLS | (4-2) IV | (4-3) OLS | (4-4) IV | (4-5) OLS | (4-6) IV | (4-7) OLS | (4-8) IV |
| Industry Affiliation | −0.021*** | −0.093 | −0.101*** | −0.539 | 5.658*** | 22.931*** | 0.235*** | 1.058** |
| | (0.008) | (0.114) | (0.038) | (0.543) | (1.396) | (7.668) | (0.081) | (0.478) |
| Mean of DV | | 0.294 | | −0.000 | | 4.495 | | 0.000 |
| Scientist FE | X | X | X | X | X | X | X | X |
| Year FE | X | X | X | X | X | X | X | X |
| $p$-value Anderson-Rubin Wald $\chi 2$ | | 0.416 | | 0.320 | | 0.003 | | 0.025 |
| $N$ | | 11,371 | | 11,371 | | 11,371 | | 11,371 |

Notes: This table shows the OLS and 2SLS IV estimates of the effect of industry affiliation on the direction of research. The measures of the direction of research used as dependent variable are in the column super headers Standard errors clustered by scientist in parentheses. Statistical significance is *(10%), **(5%), or ***(1%).

effect of industry affiliation. Secondly, the large difference between the estimates may also be driven by bias in the OLS estimates. Together these results suggest that while industry affiliation increases the commercial potential of the research produced by those who affiliate with industry, it does not appear to significantly reduce the novelty of research.

# References


Aghion, Philippe, Dewatripont, Mathias, & Stein, Jeremy C. 2008. Academic freedom, private-sector focus, and the process of innovation. *The RAND Journal of Economics*, **39**(3), 617–635.

Agrawal, Ajay, & Henderson, Rebecca. 2002. Putting patents in context: Exploring knowledge transfer from MIT. *Management science*, **48**(1), 44–60.

Agrawal, Ajay, McHale, John, & Oettl, Alexander. 2017. How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy*, **46**(4), 853–867.

Azoulay, Pierre, Ding, Waverly, & Stuart, Toby. 2007. The determinants of faculty patenting behavior: Demographics or opportunities? *Journal of economic behavior & organization*, **63**(4), 599–623.

Azoulay, Pierre, Ding, Waverly, & Stuart, Toby. 2009. The impact of academic patenting on the rate, quality and direction of (public) research output. *The Journal of Industrial Economics*, **57**(4), 637–676.

Bikard, Michaël, Vakili, Keyvan, & Teodoridis, Florenta. 2018. When collaboration bridges institutions: The impact of university–industry collaboration on academic productivity. *Organization Science*, **30**(2), 426–445.

Blumenthal, David, Causino, Nancyanne, Campbell, Eric, & Louis, Karen Seashore. 1996. Relationships between academic institutions and industry in the life sciences?an industry survey. *New England Journal of Medicine*, **334**(6), 368–374.

Cohen, Wesley, Florida, Richard L, & Goe, W Richard. 1994. *University-industry research centers in the United States*. Center for Economic Development, H. John Heinz III School of Public Policy and Management, Carnegie Mellon University.

Cohen, Wesley M, Florida, Richard, Randazzese, Lucien, & Walsh, John. 1998. Industry and the academy: uneasy partners in the cause of technological advance. *Challenges to research universities*, **171**(200), 59.

Evans, James A. 2010. Industry Induces Academic Science to Know Less about More 1. *American Journal of Sociology*, **116**(2), 389–452.

Gans, Joshua S, Murray, Fiona E, & Stern, Scott. 2008. Patents, papers, pairs and secrets: Contracting over the disclosure of scientific knowledge. *Manuscript, Sloan School of Management, Massachusetts Institute of Technology*.

Geison, Gerald L. 1995. *The private science of Louis Pasteur*. Vol. 306. Princeton University Press.

Henderson, Rebecca, Jaffe, Adam B, & Trajtenberg, Manuel. 1998. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *The review of economics and statistics*, **80**(1), 119–127.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen,

Patrick, Sainath, Tara N, *et al.* 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, **29**(6), 82–97.

Hinton, Geoffrey E, Osindero, Simon, & Teh, Yee-Whye. 2006. A fast learning algorithm for deep belief nets. *Neural computation*, **18**(7), 1527–1554.

Hochreiter, Sepp, & Schmidhuber, Jürgen. 1997. Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

Hornik, Kurt, Stinchcombe, Maxwell, & White, Halbert. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**(5), 359 – 366.

Jaffe, Adam B. 1989. Real effects of academic research. *The American Economic Review*, 957–970.

LeCun, Yann, Boser, Bernhard, Denker, John S, Henderson, Donnie, Howard, Richard E, Hubbard, Wayne, & Jackel, Lawrence D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, **1**(4), 541–551.

Levin, R, Klevorick, NR, & Winter, S. 1987. Appropriating returns from industrial research and development in Baily, M. and Winston, C. redactors ?Brooking papers on Economic activity. *Wachington, the brooking institution.*

Marx, Matt, & Fuegi, Aaron. 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, **41**(9), 1572–1594.

Marx, Matt, & Fuegi, Aaron. 2022. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, **31**(2), 369–392.

Minsky, Marvin Lee, & Papert, Seymour. 1988. Perceptrons; an introduction to computational geometry. *258 p. : ill.*

Nelson, Richard. 1959. The simple economics of basic scientific research. *Journal of political economy*, **67**(3), 297–306.

Priem, Jason, Piwowar, Heather, & Orr, Richard. 2022. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts.*

Raina, Rajat, Madhavan, Anand, & Ng, Andrew Y. 2009. Large-scale deep unsupervised learning using graphics processors. *Pages 873–880 of: Proceedings of the 26th annual international conference on machine learning.* ACM.

Roach, Michael, & Sauermann, Henry. 2014. Not all scientists pay to be scientists: PhDs' preferences for publishing in industrial employment. *Research Policy*, **43**(1), 32 – 47.

Rosenberg, Nathan. 1990. Why do firms do basic research (with their own money)? *Research Policy*, **19**(2), 165–174.

Rosenblatt, Frank. 1957. *The perceptron, a perceiving and recognizing automaton Project Para.* Cornell Aeronautical Laboratory.

Rumelhart, David E., Hinton, Geoffrey E., & Williams, Ronald J. 1986a. Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, **1**, 318–362.

Rumelhart, David E., Hinton, Geoffrey E., & Williams, Ronald J. 1986b. Learning representations by back-propagating errors. *Nature*, **323**(6088), 533–536.

Simeth, Markus, & Raffo, Julio D. 2013. What makes companies pursue an open science strategy? *Research Policy*, **42**(9), 1531–1543.

Sinha, Arnab, Shen, Zhihong, Song, Yang, Ma, Hao, Eide, Darrin, Hsu, Bojune Paul, & Wang, Kuansan. 2015. An overview of microsoft academic service (mas) and applications. *Pages 243–246 of: Proceedings of the 24th international conference on world wide web*. ACM.

Sohn, Eunhee. 2021. How local industry R&D shapes academic research: Evidence from the agricultural biotechnology revolution. *Organization Science*, **32**(3), 675–707.

Stern, Scott. 2004. Do scientists pay to be scientists? *Management science*, **50**(6), 835–853.

Stokes, Donald E. 1997. *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Stuart, Toby E, & Ding, Waverly W. 2006. When do scientists become entrepreneurs? The social structural antecedents of commercial activity in the academic life sciences. *American Journal of Sociology*, **112**(1), 97–144.

Uzzi, Brian, Mukherjee, Satyam, Stringer, Michael, & Jones, Ben. 2013. Atypical combinations and scientific impact. *Science*, **342**(6157), 468–472.

Werbos, Paul John. 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*.

Zucker, Lynne G., & Darby, Michael R. 2014. Defacto and Deeded Intellectual Property: Knowledge-Driven Co-Evolution of Firm Collaboration Boundaries and IPR Stragtegy. *Annals of Economics and Statistics*, 221–251.

Zucker, Lynne G., Darby, Michael R., & Brewer, Marilyn B. 1998. Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises. *American Economic Review*, **88**(1), 290 – 306.

# Online Appendix: "How Does Industry Affiliation of Academic Scientists Change the Rate and Direction of Research?"

Amir Sariri