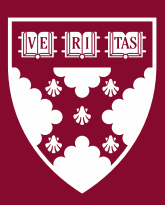
Working Paper 21-128

Standing on the Shoulders of Science

Joshua L. Krieger
Monika Schnitzer
Martin Watzinger



Harvard Business School

Standing on the Shoulders of Science

Joshua L. Krieger
Harvard Business School

Monika Schnitzer
Ludwig Maximilian University Munich

Martin Watzinger
University of Münster

Working Paper 21-128

We thank Jeff Furman, Fabian Gaessler, Rem Koning, Kyle Myers, and Fabian Waldinger for helpful comments and discussions. We thank Mohammad Ahmadpoor, Ben Jones, and Bill Kerr for sharing their data. Watzinger and Schnitzer gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through SFB-TR 190. Watzinger thanks the REACH – EUREGIO Start-Up Center for their kind support.

Funding for this research was provided in part by Harvard Business School.

Standing on the shoulders of science

Joshua L. Krieger* Monika Schnitzer† Martin Watzinger‡§

September 20, 2023

Forthcoming at The *Strategic Management Journal*

Abstract

Today's innovations rely on scientific discoveries of the past, yet only some corporate R&D builds directly on scientific output. In this paper, we analyze U.S. patents to investigate how firms generate value by building on prior art "closer" to science. We show that patent value is decreasing in distance-to-science. Overall, we find a science premium within firms ranging from 5.0% and 18.3%. If we allow for firm sorting into different modes of R&D based on their relative advantage, i.e., when we do not control for firm fixed effects, we find an even larger science premium: patents building directly on scientific publications are 4.0%–42.3% more valuable than patents in the same technology that are not directly based on science.

JEL Codes: O30, O34, O33, O31

*Joshua L. Krieger: Harvard Business School, 60 N. Harvard St., Boston, MA 02163; jkrieger@hbs.edu.

†Monika Schnitzer: Ludwig Maximilian University Munich, Akademiestrasse 1, 80799 Munich, Germany; schnitzer@econ.lmu.de.

‡Martin Watzinger: University of Münster, Am Stadtgraben 9, 48143 Münster, Germany; martin.watzinger@wiwi.uni-muenster.de.

§We thank Jeff Furman, Fabian Gaessler, Rem Koning, Kyle Myers, and Fabian Waldinger for helpful comments and discussions. We thank Mohammad Ahmadpoor, Ben Jones, and Bill Kerr for sharing their data. Watzinger and Schnitzer gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through SFB-TR 190. Watzinger thanks the REACH – EUREGIO Start-Up Center for their kind support.

1 Introduction

While scientific advances form the bedrock of industrial R&D, only some of these activities directly build on science—translating discoveries from laboratories and scientific publications into novel inventions and commercial products. Other corporate innovation efforts indirectly rely on science—experimenting, tinkering, optimizing, and inventing without the direct aid (or constraints) of "the republic of science," yet still utilizing tools and technologies enabled by centuries of scientific progress. The level of a firm's engagement with science is a crucial component of its R&D strategy and can be a potential source of competitive advantage (Henderson and Cockburn, 1994; Cockburn *et al.*, 2000; Stern, 2004). However, surprisingly little is known about how leveraging scientific knowledge influences the value of a firm's inventions.

In this paper, we explore whether and how firms derive value by inventing "on the shoulders of science." This exploration is central to R&D strategy, given the significant resources required to assimilate scientific knowledge. (e.g., in-house scientists or collaborative projects with universities). To justify such expenditures, managers must anticipate a significant science premium, i.e., that inventions rooted in science are more valuable than those that do not directly incorporate scientific insights.

Although the societal value of science is immense, formal science is not an obvious source of competitive advantage. Scientific breakthroughs are typically published, making them accessible to all. Consequently, any science premium should be eroded by new entrants, implying that access to science does not provide a lasting competitive edge. Moreover, popular discourse in the tech sector and dominant threads of management literature express skepticism towards science-based inventing. That dissenting perspective contends that valuable inventions stem from applied industrial engineering and user innovation. By gauging the relative value of inventions based on their closeness to scientific publications, we aim to guide R&D strategy and mediate between contrasting views on the private value of science.

To determine the "science premium" of a patent, we assess how varying degrees of reliance

on science influence the private value of patents. To categorize patents by their distance-to-science, we draw from Ahmadpoor and Jones (2017). When a company files a patent, it must list all preceding art, including scientific articles, upon which the patent is based. This establishes a direct link between the patent and the scientific knowledge it utilizes. A patent directly citing a scientific paper is assigned a distance of one (D=1) to science. A patent that references a (D=1)-patent but doesn't cite a scientific article itself has a distance of (D=2), and so forth. We align this data with patent values from Kogan *et al.* (2017), which we also abbreviate as KPSS. Merging these datasets allows us to correlate patent value with distance-to-science for 1.2 million U.S. patents issued between 1980 and 2010.

KPSS calculate patent values using excess stock returns of the filing company around the patent publication date. If a firm receives multiple patents on a particular date, KPSS distribute the total computed value equally among all patents. This means every patent filed on the same date by the same company holds identical value in the KPSS dataset. To tackle this "multiple patents problem," we employ various estimation methods, yielding a spectrum of plausible estimates for the science premium.

Initially, we contrast the value of patents directly based on science (D=1) with those not directly rooted in science (D>1). Controlling for firm fixed effects, we discern a science premium within firms ranging between 5.0% and 18.3%, contingent on the assumptions we adopt to address the multiple patent issue. When we account for firm sorting across R&D strategies and innovation quality (i.e., cross-firm analysis without firm fixed effects), the science premium spans from 4.0% to 42.3%.

However, these estimates might undervalue the impact of science on patent worth, as patents in the control group might still be indirectly influenced by science. To account for this indirect impact of proximity to science, we next compare the value of patents directly based on science (D=1) against those with a tenuous link to science (D=4). The derived estimates are roughly double those of our primary specification. The science premium, when accounting for firm fixed effects, lies between 9.0% and 33.5%. Without controlling for firm

fixed effects, the science premium varies from 17.2% to 95.6%.

We also observe that patents with a distance of two (D=2) or three (D=3) from science possess private values exceeding the (D=4) group but are less than patents directly based on science (D=1). This cascading value from science to patents not directly rooted in science suggests that scientific advancements can act as the "remote dynamo of technological innovation" across the economy (Stokes, 2011, p.84), influencing beyond immediate practical applications.

In the auxiliary results, we demonstrate that our primary findings remain consistent when using alternative metrics for distance-to-science or when incorporating measures for patent value. By retaining our main specification and showcasing results via alternative methods to address the multiple patent dilemma, as well as different estimation techniques, we identify a median science premium of 10.7% across 24 plausible specifications when considering firm fixed effects. 22 out of these 24 estimates significantly deviate from zero at the 5% level. Without accounting for firm fixed effects, the median science premium stands at 22.0%, with 17 out of the 24 estimates being significantly different from zero at the 5% level.

Having established an overall science premium, we then investigate possible sources. To understand how distance to science relates to the innovative content of inventions, we develop a new measure of patent novelty based on the novelty of word combinations in the text of the patent. For this purpose, we calculate for each patent the probability that a given combination of words has been used before. We call a patent "novel" if it contains low-probability word combinations. We document that patent novelty predicts the value of patents in a very similar way to a patent's science-intensity. Additionally, we establish that the content of more science-intensive patents is more novel and that the novelty of the content decreases with distance-to-science.

These results pose a puzzle. If there is a significant science premium, why do some *but not all* companies invent "on the shoulders of science"? It turns out that while the benefits of using science are clear on average, not every company is equipped to tap into

these benefits. We find that companies that invest in absorptive capacity for science, either by employing scientist-inventors or by conducting their own in-house research, have a higher science premium for their patents. This also implies that for many smaller companies, without the scale of innovation to spread the fixed costs of investing in absorptive capacity, they might find the process of science-based inventing as too long, costly, or unpredictable, even if it could lead to more valuable patents. In short, while the appeal of using science is clear, economic tradeoffs and path-dependencies hold some companies back.

Our paper contributes to the literature in three ways. First, it highlights that science-driven R&D is associated with substantial value in the private sector, not only directly, but also indirectly (i.e., D>1), and it quantifies the respective value contributions in percentage terms. Regarding intent, this is close to the early surveys of Edwin Mansfield, which showed that in the 1980s and the 1990s, around 20 percent of all newly introduced products benefited substantially from recent academic science (Mansfield, 1991, 1995, 1998). The recent literature primarily focuses on patents that are directly science-based and on value measures such as forward citations and patent renewal payments, which reflect only indirectly and partially the private value of the patents for the owner. Sorenson and Fleming (2004) show that science-based patents have more follow-on citations. Poege *et al.* (2019) find that the quality of cited scientific articles is positively related to various monetary and non-monetary measures of patent value. Ahmadpoor and Jones (2017) document that forward citations decrease with distance-to-science, and that patents close to science are more likely to be renewed. The benefits of academic science and industry seem to flow both ways, as academic-industry collaboration and citation boosts quality and productivity for both sides (Bikard *et al.*, 2019; Bikard and Marx, 2020).

Second, our paper takes an important step towards understanding the role of science in patent value by showing that science and the novelty of patents go hand in hand. Basic science is frequently credited with stimulating technological innovations.[1] By linking patent

[1]In the context of World War I, Iaria *et al.* (2018) have recently shown that scientists produce more patent-relevant scientific articles if they have access to frontier knowledge. Fleming and Sorenson (2004)

novelty to patent value and science to patent novelty, we provide a rationale for why science matters for private sector innovation—enabling a different, and more fat-tailed, type of technology search.[2] Furthermore, our findings support the view that the prevailing decline in corporate in-house R&D is more likely due to shifts in organizational boundaries or in the costs of performing in-house R&D, than in the private gains of translating science into inventions (Arora *et al.*, 2018).

Third, by estimating how the private value of science-based patents differs by investment in absorptive capacity, our results shed light on firms' incentives to use science in the innovation process. The findings suggest that investments in the firm's ability to build on science are a source of competitive advantage—one that is not equally available to all firms, especially those that lack scale (Cohen and Levinthal, 1990a; Henderson and Cockburn, 1994). These results complement the findings of the concurrent paper of Arora *et al.* (2021) that show that there is a significant first-mover advantage (in building on science) in terms of patent value using similar data to ours. Different from our analysis, the authors focus on within-firm variation.

2 Context: Competing Narratives

Historical narratives often highlight how the private sector has harnessed modern science for significant inventions. The development of the wireless telegraph by Ferdinand Braun and Guglielmo Marconi was based on Heinrich Hertz's foundational work on electromagnetic waves. Similarly, the invention of the transistor at Bell Laboratories was grounded in a deep understanding of semiconductor physics. The emergence of the biotechnology sector can

argue that science alters inventors' search processes and leads them to useful new knowledge combinations.

[2]Thus, our study complements the indirect evidence in Fleming and Sorenson (2004), which shows that science increases forward citations in fields in which it is hard to innovate. Recently, Kelly *et al.* (2021) have demonstrated that the value of patents as measured by Kogan *et al.* (2017) is negatively correlated with their text similarity to earlier patents. We add to these findings by demonstrating that patent novelty systematically correlates with the scientific content of a patent, measured by citation distance or by text similarity between articles and patents.

trace its roots to academic figures like Genentech founder Herbert Boyer.[3]

In the realm of invention, Fleming and Sorenson (2004) posits that science might serve as a guiding compass for technological exploration. The scientific process identifies promising trajectories and potential pitfalls. In certain instances, scientific articles can also directly lead to inventions. These inventions and scientific projects often manifest as patent-paper pairs (Murray, 2002), highlighting similar or identical discoveries.[4]

The "republic of science" ethos prioritizes rapid advancements and original contributions. While this might lead to valuable private sector inventions, it also underscores that scientific results should be published and available to the public (Partha and David, 1994; Stephan, 2012). Once results are published, they transition to the public domain, permitting any firm to use these insights. Thus, any competitive edge gained from employing scientific insights may be transient. As competitors can tap into the same scientific knowledge base, they might rapidly replicate innovations, potentially diluting any unique science-based advantage. In reality, not all companies can capitalize on scientific benefits because they may lack the subject matter expertise and access to physical tools necessary for absorbing and using new scientific knowledge (Cohen and Levinthal, 1990b). This absence of absorptive capacity could serve as an entry barrier, such that companies possessing absorptive capacity might enjoy a premium for harnessing science (Gambardella, 1992; Cockburn and Henderson, 1998).

However, many scholars in technology and management contend that profitable science-based inventions are exceptions, rather than the norm. These critics suggest that many inventions primarily emerge from areas outside formal scientific research (Kline and Rosenberg, 1986; von Hippel, 1988). They underscore the importance of applied industrial engineering and user-driven innovation. This idea is in line with recent trends in the corporate world, indicating a decline in internal scientific research, which might lead to a loss in absorptive capacity for science (Arora *et al.*, 2018, 2020). At the same time, venture capital has favored

[3]Boyer, during his time at the University of California, San Francisco, contributed to recombinant DNA research before co-establishing Genentech with venture capitalist Robert Swanson.

[4]These pairs can be identified through their shared linguistic elements (Magerman *et al.*, 2015; Murray and Stern, 2007).

rapid experimentation and software-driven business models. (Ewens *et al.*, 2018). This shift, epitomized by Mark Zuckerberg's famous "move fast and break things" mantra seems at odds with the deliberate nature of science-based invention. These broader trends could suggest that the challenges of translating frontier science into marketable products outweigh the potential benefits. [5]

In summary, while there are reasons to believe that scientific input can enhance the inventive process, there remains no consensus on the value of that input. The following section outlines the data we use to address this unresolved question.

3 Data

Our starting point is a dataset which contains information on the expected monetary value of 1.8 million public firm patents issued from 1926 to 2010 (Kogan *et al.*, 2017). The private value of the patent is estimated by studying movements in stock prices following the days that patents were issued to the firm. Specifically, the value is approximated using the abnormal stock market return of the filing company within a narrow window around the grant date of the patent.

For each of these patents, we calculate its "distance" to prior scientific advances using the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). We use information on 3.6 million patents issued by the U.S. Patent and Trademark Office (USPTO) from 1980 to 2010, and information on journal articles indexed by Microsoft Academic (Sinha *et al.*, 2015). We then locate patents that directly cite journal articles; i.e., patents to which practical inventions and scientific advances are directly linked (Marx and Fuegi, 2020). A patent that directly cites a scientific paper is assigned a distance of one (D=1) to science. A patent that cites a (D=1)-patent, but does not cite a scientific article itself, has a distance

[5] Alternatively, it might indicate that science-based innovation retains its value, and the corporate shift away from science is a reflection of changing R&D strategies or broader innovation trends. The trend towards "open innovation" and sourcing innovations from technology markets (Chesbrough *et al.*, 2006; Arora *et al.*, 2004; Gans and Stern, 2003; Mowery, 2009; Bhaskarabhatla and Hegde, 2014) allows firms to benefit from external research without direct in-house scientific endeavors.

of two (D=2), and so on. The distance for each patent to science is thus defined by the minimum citation distance to the boundary where there is a direct citation link between patent and scientific article.

Combining the information on patent values and citation links, we construct a dataset that contains patent values for 1.2 million U.S. patents issued between 1980 and 2010. 20.2% of all patents directly cite a scientific article (D=1), 53.4% are indirectly based on science (D=2 and D=3) and 26.3% are not based on science (D=4 or larger). The (unadjusted) average value of a patent is $12.8 million in constant 1982 U.S. dollars. Appendix A.1 gives a detailed description of all the data construction and sources.

Specific examples help to understand the range of inventions represented in the data. Even among well-cited patents, we observe significant qualitative differences between patents by their distance from science. Take Coca-Cola's 1997 patent titled "Apparatus for icing a package" (5671604). The solo-inventor patent describes a vending machine refrigeration system featuring a spray nozzle that cools the stored items with a water mist. It has no citations to science and a distance from science of five (D=5). The patent contains seven figures, all detailed technical drawings of the cooling system and its 80 different enumerated components (e.g., control valves, linear actuators, vortex cooling devices).

In the same technology class as the Coca-Cola vending machine patent (G07F, "Coin-freed or like apparatus"), one can also find McKesson Automation's patent number 7010389, "Restocking system using a carousel," intended to aid in dispensing of medical supplies. The patent's figures bear considerable similarity to the Coca-Cola patent. Both include detailed drawings of a motorized storage system that brings items to a stationary user. Unlike the soda vending machine, the McKesson patent includes a computer, printer, and hand-held wireless device system. Furthermore, the patent cites 15 scientific articles, including some from journals such as the *International Journal of Bio-Medical Computing* and the *American Journal of Hospital Pharmacy*. These articles present evidence on how the implementation of similar database systems improved operations at hospitals and the health benefits of

deploying monitoring systems to prevent toxic multi-drug interactions.

The actual variation in how patents describe their inventions and build on prior art is impossible to capture meaningfully in a small set of illustrative examples, and the corpus of patents is so large that a counter-example is always just a simple Google search away. Thus, in our regression analyses that follow, we emphasize that our methods are useful for describing *average* correlations between groups. Of course, there are obvious differences between medical devices, telecommunications, transportation, machine tools, and food processing technologies. Therefore, all of our regression specifications control for technology class and time fixed differences.

Appendix A.2 provides additional examples using patents from CPC class A61L ("Methods or apparatus for sterilizing materials"). Across examples, the distinguishing features of science-based patents have more to do with their *process* or R&D than their complexity or sophistication. Patents that are more proximate to formal science use the tools and language of science to search for a technology solution, identify its novelty, and communicate its value.

Appendix A.2 also describes the patent sample, with summary statistics broken down by distance-to-science (Tables A.1, A.2 and A.3). Patents closer to science are different across a number of interesting dimensions. Most notably, patents closer to science tend to have more inventors, shorter claims and take longer for the USPTO to process. Their prior art also looks different, as science-based patents tend to build on a larger and broader scope set of backwards citations.

4 Results

The next subsection documents the relationship between patent values and distance-to-science. We show that more science-intensive patents are on average more valuable. The following subsection demonstrate that patents closer to science are more novel and more novel patents are more valuable offering a potential explanation for the science premium. Finally,

we highlight that the benefits derived from science-based inventions vary significantly among companies, explaining why not every firm pursues innovation rooted in scientific research.

4.1 The private value of patents by their distance-to-science

We start by presenting how relative dollar values of patents differ by whether a patent cites or does not cite science (across all firms). To separate science-related from non-science-related patent value, we need to make assumptions about the data-generating process. We assume that the value of a patent is generated by an additive separable technological component, a proximity to science component, a time component, and an idiosyncratic component. The technological component is assumed to be the same for all patents with the same technology class, independent of their distance to science.[6]

To estimate relative percentage (%) differences in KPSS values by whether or not a patent cites science, we first estimate the absolute difference in patent value by distance-to-science using the following specification

$$Y_{ijkt} = \alpha_0 + \alpha_1 \times (D = 1)_i + \delta_j + \gamma_t + \varepsilon_{ijt} \tag{1}$$

where Y_{ijt} is the KPSS patent value for patent i that is associated with the (CPC 4-digit) technology class j, filing time period t and firm k. The variables (D=1) is an indicator variable whether or not a patent cites an academic article. δ_j is a fixed effect for each technology class and γ_t is a fixed effect for the issue week.[7] In a second step we convert the regression coefficients to percentage differences relative to all patents that are not (D=1) average values.

[Insert Table 1 Here]

[6]Stephan (1996) captures the ties between science and particular industries, writing that "to a considerable extent the scientific enterprise evolves in disciplines that from their beginnings have been closely tied to fields of technology."

[7]The USPTO issues patents only once per week on Tuesday, so the issue week identifies the issue date.

Column 1 of Table 1 shows the coefficients α_1 of Equation 1 without firm fixed effects. In this specification, a science-based patent that directly cites an academic article (D=1) has a science premium of 19.9% (95% CI [9%, 31%]), i.e., patents that cite science have an 19.9% greater value than patents that do not cite science controlling for technology and issue week.[8] Panel (a) of Figure 1 corresponds to Column 2 of Table 1 and shows how patent dollar value decreases in distance to science.

Columns 2 through 9 show the results of plausible alternative specifications for estimating the science premium, which we discuss in turn.

Relative to all other patents vs. relative to D=4 patents

[Insert Figure 1 Here]

In our main specification, we compare the value of patents that cite science with those that do not. However, patents that do not cite science might use science indirectly because they might be based on prior science-based innovation. To operationalize the notion that patents are indirectly based on science, Ahmadpoor and Jones (2017) developed the concept of distance to science. A patent that directly cites a scientific article has a distance of 1 (D=1), a patent that does not cite a scientific article but cites a D=1 patent has a distance of 2, and so on. If a company has no scientific capabilities, it might be difficult to invent, even at a distance of 2 to science, because some understanding of the scientific basis might be required.

To estimate relative percentage (%) differences in KPSS values across distance-to-science groups, we add to the specification in Equation 1 indicator variables (D=1,2,3,5,>5, unconnected)

[8] Our preferred specifications report patent value differences in relative (%) terms, but to acquire a sense of the order of magnitude, we present the same results in levels ($USD) in Appendix B.1.

for the distance-to-science of patent i, where (D=4) is the reference category.[9]:

$$\begin{aligned} Y_{ijkt} = \alpha_0 + \alpha_1 \times (D=1)_i + \alpha_2 \times (D=2)_i + \alpha_3 \times (D=3)_i \\ +\alpha_4 \times (D=5)_i + \alpha_5 \times (D>5)_i + \alpha_6 \times (D=unconnected)_i + \delta_j + \gamma_t + \varepsilon_{ijt} \end{aligned} \tag{2}$$

In this specification, a science-based patent that directly cites an academic article (D=1) has an average value that is 36.9% (95% CI [18%, 56%]) greater than a patent four degrees removed from science (D=4) (Column 2 in Table 1). This value decreases as the distance to science increases. Patents with a distance of two (three) have average values 22.8% (7.4%) higher than those in (D=4), respectively.[10] Figure 1 corresponds to Column 1 of Table 1 and shows how patent dollar value decreases in distance to science.

Whether all other or D=4 patents are the right counterfactual depends on the question. If we are interested in the overall contribution of science to corporate R&D, than we should compare the value of a patent based on science (D=1) with a patent that is neither directly nor indirectly based on science (D=4). If we are interested in the average increase in patent value if a random company (relative to its current patent portfolio) would start to build directly on science, then the specification in Equation 1 and the science premium of 19.9% is the correct number.

The multiple patents problem In Column 3, we address an idiosyncrasy of the KPSS data. KPSS estimates the patent value from the assignee's abnormal stock market returns around the date of the patent issuance. If there are multiple patents N_k are issued to firm k on the same day, Kogan *et al.* (2017) assigns each patent a fraction $\frac{1}{N_k}$ of the estimated total value based on the same abnormal stock market returns. Therefore—if multiple patents are

[9] Distance-to-science of four as the baseline is an arbitrary choice. The results are very similar if we use other baselines.

[10] Our preferred specifications report patent value differences in relative (%) terms, but to acquire a sense of the order of magnitude, we present the same results in levels ($USD) in Appendix B.1.

present—we only observe an average by firm and date of the actual value of a patent, i.e.

$$\overline{Y_{kt}} = \frac{1}{N_{kt}} \sum_i Y_{ijkt}$$

where $\overline{Y_{kt}}$ is the average patent value of patents of firm k at date t.

The multiple patents per day issue mechanically biases the science premium estimate towards zero if a company publishes patents with different distances to science on the same date. This multiple patent problem is (potentially) not a minor issue: In our data, 237,564 of all 1,176,990 patents cite science (D=1). 82.0% or 194,763 of patents citing science (D=1) are issued on a date on which the same assignee gets also issued other patents that do not cite science (D>1). This means that for 82.0% of all patents citing science, we only observe an average value that includes the value of patents that do not cite science.[11] Therefore, the actual patent value Y_{ijkt} of patent i is unknown if there are multiple patents per firm-date, and we need to make additional assumptions to be able to estimate Equation 1.

The first potential assumption is that the reported KPSS average is the actual patent value, and all patents issued on the same date have truly the same value. Then, the specifications in Columns 1 or 2 are correct. A second potential assumption is that patents issued on the same date to the same assignee do not have the same value. Then, we need to model the data generating process of the average patent value $\overline{Y_{kt}}$. To do this, we aggregate Equation 1 on the firm and issue week level by taking averages of the dependent and the independent variables:

$$\overline{Y_{kt}} = \alpha_0 + \alpha_1 \cdot \overline{(D=1)_i} + \overline{\delta_j} + \gamma_t + \varepsilon_{kt} \tag{3}$$

[11] For example, two patents in CPC subclass B05D with publication numbers US5289639 and US5290586 were issued to IBM in 1994, on the same date. The first patent cites science and has 35 forward citations; the second patent does not cite science and has three forward citations. Many studies show that forward citations are correlated with patent value (e.g., gathered in surveys). Therefore, we would expect that the first patent is more valuable than the second (Trajtenberg, 1990; **?**, among others). However, in our data, both patents have the same value of 4.53 million Dollars.

where $\overline{(D=1)_i}$ is the average over the indicator whether or not a patent cites science (e.g., if the firm recieves two patents in one week and only one cites science, $\overline{(D=1)_i} = 0.5$). Since we aggregate on the firm-week level and each firm potentially has patents in multiple technologies, we also take the average over the full set of indicator variables identifying the technology classes, $\overline{\delta_j}$.

Using the aggregated specification in Column 3 of Table 1, the science premium of a patent that directly cites an academic article (D=1) is 4.0% (95% CI [-9%, 17%]). If we control for all distances-to-science in Column 4, a patent that directly cites an academic article (D=1) has an average value that is 17.2% (95% CI [-3%, 38%]) greater than a patent four degrees removed from science (D=4).

These mean estimates are smaller than those in Columns 1 and 2 because they are based on a differently weighted sample. In Columns 1 and 2, we give each patent the same weight in the estimate. In Columns 3 and 4, we give each firm-issue week observation the same weight independent of the underlying number of patents. This is potentially important because the number of patents per firm and issue week range from 1 to 258, with a median of 8 and a mean of 16.5.

In Columns 5 and 6, we address this issue by using a weighted regression with the number of patents per firm and issue week as weights. In this weighted regression, the science premium of a patent that directly cites an academic article (D=1) is 42.3% (95% CI [9%, 75%]). In Column 6, we again change the control group and compare the value of the patent that directly cites science (D=1) with a D=4 patent and find an increase in patent value of 95.8% (95% CI [18%, 174%]). The value premium decreases as the distance to science increases. Patents with a distance of two (D=2) or three (D=3) have average values 67.2% and 19.4% higher than those in (D=4), respectively.

The specification in Column 5 (relative to D>1, weighted regression, and using KPSS averages as outcomes) is our preferred specification for three reasons: First, comparing D=1 patents to D>1 patents gives us a conservative estimate of the science premium. Second, we

are interested in a patent-level measure of a science premium; therefore, each patent should get the same weight in the regression. This means the estimation should be weighted. Third, we consider it unlikely that all patents in a firm $\times$ date combination have precisely the same value. To strengthen the latter argument that patents issued on the same date to the same firm do not have the same patent value, Table A.6 in Appendix B.2 shows summary statistics for different patent quality indicators for patents citing science and not science. The analysis sample in this table is the subset of patents where a given firm had multiple patents issued on a single date and where at least one paper did and did not cite science. For each of these indicators, studies suggest a relationship between the indicator and the patent's private or social economic value.[12] Of the 15 indicators shown, all but two point to a higher value for patents citing science than patents not citing science. The two exceptions are "Radicalness" proposed by Shane (2001), where there is no difference, and Grant Lag, where it is thought that more valuable patents have a shorter time lag between filing and issuing (Harhoff and Wagner, 2009; Régibeau and Rockett, 2010). In our data, patents citing science have a longer grant lag.

Alternative Measures of Patent Value and Distance-to-science The private value of a patent is the outcome of interest for profit-maximizing firms; however, other proxies for value are informative as robustness checks and links to other types of social value (e.g., knowledge flows, spillovers). Columns 7 and 8 in Table 1 show the results of two other proxies for patent value: forward citations and propensity to be involved in litigation.

The results are broadly consistent with the science premium regressions. Column 7 shows the patent forward citation results. Controlling for technology class and issue date fixed effects, we find that patents with (D=1) average 67.5% (95% CI [60%, 75%]) more forward citations than patents that do not cite academic articles. The relative difference is even larger than the stock market valuation premium—indicating that social returns through

[12]Squicciarini *et al.* (2013) discusses the relevant studies for each indicator.

knowledge flows may be above and beyond what the inventor firms capture.[13]

Columns 8 repeats the exercise with a binary litigation indicator as an outcome. While we can only measure litigation as a binary event, it is a useful signal of patent value.[14] Patent litigation is expensive, so firms will (for the most part) only fight in court for patents that are believed to be valuable.[15] Our results shows that the likelihood of litigation is higher for patents that directly cite science. Column 8 indicates that patents with (D=1) have 70.1% (95% CI [52%, 88%]) higher likelihood to be involved in litigation than patents that do not cite science. Taken together with the KPSS and the forward citations results, this finding shows that patents that build more directly on science are not only valued higher, but firms judge them more worthy of expensive courtroom battles in the years post-grant.

Another potential concern about our estimation of the science premium of patents is that the distance-to-science calculated by citations might measure not only how much a patent uses science but also the quality of the inventor. A high-quality inventor might be more aware of scientific research and include more citations without using science.

To see whether patents close to science use its content, we compare the texts of scientific articles and those of patents. We calculate the pairwise text similarity between a patent and the articles cited in the patent. Then, we take the maximum over all the similarities of a patent to its cited articles to determine the distance to the closest article. To calculate the similarity between the abstracts of the article and the patent, we use the "term frequency-inverse document frequency" (tf-idf) method.[16]

[13]That said, we recognize that patent-to-patent citations are an imperfect measure of knowledge flows (Roach and Cohen, 2013; Marx and Fuegi, 2020; Kuhn *et al.*, 2020). Unlike KPSS, they are an ex-post quality measure, so we cannot easily quantify the gap between private and social value capture.

[14]The dependent variable comes from merging our data to the USPTO's Patent Number and Case Code File dataset, a comprehensive link between patent litigation cases in U.S. district courts and patents between 2003 and 2016. 94% of the court cases are coded as patent infringement suits, with the remaining cases involving disputes around ownership/inventorship, patent validity, royalties, false markings, and other procedural issues involving patents.

[15]The American Intellectual Property Lawyer's Association (AIPLA) estimated litigation costs of $250,000 – $950,000 for cases with less than $1 million at risk, and between $2.4 million and $4 million for cases with more at stake (`https://apnews.com/press-release/news-direct-corporation/a5dd5a7d415e7bae6878c87656e90112`)

[16]The number of observations is lower in the similarity regression as text data is limited for articles.

Consistent with the idea that patents with more scientific content have a higher value, Column 9 shows that the value of a patent increases with its text similarity to scientific articles. This suggests that the relation between citation distance and patent value presented as our main result above is not a result of a spurious correlation driven by third factors unrelated to the patent's scientific content.

Estimating the *within-firm* science premium The empirical facts documented above describe the average value of patents across the full sample of firms after controlling for the technology class and time-invariant characteristics. This *overall* science premium could have two broad (and not mutually exclusive) explanations: First, firms that are "better" at inventing (have a higher patent value on average) are more likely to use science (sorting), or second scientific proximity is associated with higher value inventions after controlling for sorting (within-firm science premium).[17]

To see whether our results are driven by sorting or if there is a within-firm science premium, we add a firm fixed effect dummy variable based on the firm identifier in the KPSS data to all specifications. The resulting estimates in Table 2 give a within-firm science premium that controls for (time-invariant) sorting across firms. Column 1 shows the result from estimating Equation 1 on the patent level and focusing on D=1 vs. all other patents. This specification yields a science premium of 5.0% (95% CI [-0%, 10%]). In Column 2, we find that a science-based patent that directly cites an academic article (D=1) has an average value that is 9.0% (95% CI [2%, 16%]) greater than a patent four degrees removed from science (D=4). This value decreases as the distance to science increases. Patents with a distance of two (three) have average values 5.3% (2.8%) higher than those in (D=4), respectively. If not for the multiple patents problem, Columns 1 and 2 would be sufficient for estimating the within-firm science premium. However, if the assumption that all patents filed by the same firm on the same date have the same patent value (as reported in the KPSS

[17]Sorting could be a result of firm's strategic and path-dependent choices, such as investment in scientific expertise and locating near academic clusters. Alternatively, sorting might be a reflection of the firm's (less changeable) initial endowments in resources or management quality.

data) does not hold, we need to take additional steps to estimate the within-firm premium.

In Columns 3–6, we aggregate the data on the firm × issue week level and then add the firm fixed effect. In Columns 3 and 4, we treat every firm week as an equal weight observation and find a science premium of 14.0% (95% CI [5%, 23%]) in Column 3 and an increase in patent value of 21.2% (95% CI [9%, 34%]) relative to D=4 in Column 4. In Columns 5 and 6, we weight each observation by the number of patents granted in a firm-issue week cell. The weighted regression produces a larger but noisier overall science premium of 18.3% (95% CI [-1%, 38%]) in Column 5 and an increase in patent value of 33.5% (95% CI [5%, 62%]) relative to D=4 in Column 6, and again with the value decreasing in distance-to-science. The weighted regression aggregated on firm-week and estimated relative to D>1 (Columns 5) is our preferred specifications for the within-firm science premium. It gives a conservative estimate of the science premium, addresses the KPSS multiple patents issue, and does not ignore the additional information in instances where prolific patenting firms have multiple grants.

As in Table 1, Columns 7 and 8 of Table 2 show the premium for patent forward citations and the likelihood of litigation. In both cases, we find a statistically significant science premium for D=1 patents: 49.1% (95% CI [43%, 55%]) and 43.2% (95% CI [28%, 59%]), respectively. In Column 9, we find a positive correlation between text similarity to scientific articles and patent value but with a smaller magnitude than the across-firm version of the analysis (Column 9 of Table 1).

Together, these results indicate that part of the overall across-firm premium results from firms with different inventing capabilities sorting into different modes of invention. However, even after controlling for sorting, the market believes that firms capture more value from patents closer to science than those that are more distant to science. Further, the results appear robust to adjusting for the multiple patents problem.

Specification curve Since our analyses have many researcher degrees of freedom, we present specification curves in Figure 2. These plots display the results from many plausible alternative empirical strategies (Simonsohn *et al.*, 2020). Each specification is described in Table 5. The preferred specification estimates the science premium relative to D>1, averages KPSS values on the firm and issue week level, and uses a weighted regression. The median estimate for the overall science premium is 22.0%, and 17 of 24 estimates are significantly different from zero on the 5% level (Panel (a)). The median estimate for the within-firm science premium is 10.7%, and 22 of 24 estimates are significantly different from zero on the 5% level (Panel (b)). When we exclude specifications that discard substantial portions of the data (i.e., all specifications excluding "Excluding multiple patents" and "Only science or not"), we get a range of plausible estimates for the overall science premium between 4.0% and 181.1% and for the within-firm science premium estimates between 2.6% and 33.4%.

4.2 Interpreting the science premium

The main finding of our results section is that patents closer to science have a higher value. That fact leads to two natural interpretation questions: 1) Why is there a science premium? 2) Given that science premium, why do not all firms leverage scientific knowledge in their patenting strategies? In this section, we address each of these questions through additional analysis.

4.2.1 Why is there a science premium?

If the goal of science is to advance knowledge by making new discoveries, then inventions relying directly on science might introduce more novel ideas into private-sector patenting. If novel patents are also, on average, more valuable, the introduction of novelty might be one reason for the science premium. To see if this is the case, we construct a new measure of patent novelty. Using this measure, we establish that novelty is associated with patent proximity to science and is correlated with patent values.

Measuring patent novelty In the history of technology and innovation, inventions are often conceptualized as the outcome of successfully combining ideas, either by combining new ideas or existing ones in a novel way. In *A History of Mechanical Inventions,* Abbott Payson Usher writes: "Invention finds its distinctive feature in the constructive assimilation of preexisting elements into new syntheses, new patterns, or new configurations of behavior" (Weitzman, 1998). Following this concept of invention as a novel combination of ideas or resources, we develop a new measure for patent novelty based on the patent's content. More specifically, we measure how novel the combinations of words are that are used in a patent. For example, the word "mouse" combined with the word "trap" was used in patents since at least 1870. By contrast, the word "mouse" was combined with the word "display" for the first time in 1981 in Xerox's pioneering patents.[18]

Our measure of patent novelty is constructed as follows. In a first step, we count how often a particular pairwise combination of different words was used in the abstracts of previous patents up to the filing year. The sets of words for each patent are from the Arts et al. (Arts *et al.*, 2018) data set. We then divide this count by the total number of pairwise word combinations up to the filing year of the patent. We denote this ratio as the probability of a word combination. In a second step, we take the average over the respective probabilities of all pairwise word combinations within a patent to determine the average probability per patent. The smaller the average probability of pairwise word combinations, the more novel the pairwise word combinations used in the particular patent. In a last step, to arrive at our final patent novelty measure, we multiply the novelty measure by minus one, such that higher values of the novelty measure indicate more novel patents; we winsorize it to account for outliers and add the minimum to make the resulting values positive. [19]

Patent novelty and distance-to-science

[18]Though notably, the term "computer mouse" was in colloquial use starting as early as 1968, when Doug Engelbart first demonstrated the device after filing the first computer mouse patent in 1967 under the title "X-Y Position Indicator for a Display System." See `https://dougengelbart.org/`.

[19]The number of observations is lower in the novelty regression than in our main specification because we do not have abstracts for some patents.

[Insert Table 2 Here]

[Insert Figure 3 Here]

Table 3 explores the relationship between novel combinations of ideas and proximity to science. In Columns 1–6, the dependent variable is the patent novelty measures. Columns 1 and 2 show the results controlling for technology and issue week, and in Columns 3 and 4, we add firm fixed effects. Without controlling for firm fixed effects, patents that directly cite academic articles (D=1) have a 7.6% (95% CI [6%, 9%])) a higher novelty than patents that do cite academic articles. Relative to D=4 patents, D=1 patents have a 9.4% (95% CI [7%, 11%])) a higher novelty. We find very similar results if we control for firm fixed effects in Columns 3 and 4 or if we use text similarity to scientific articles as an alternative measure for distance to science in Columns 5 and 6: patents that are based more on science have more novel word combinations in their text.

Novelty and patent value Figure 3 shows that the novelty of a patent – measured by the average probability of word combinations – predicts the patent value. In Panel (a), we show a binned scatter plot by sorting all patents into 5% bins according to their novelty. Then, we plot the average patent value and novelty within these bins. Without controlling for anything, there is a positive relationship between novelty and patent value. The pattern suggests increasing returns to novelty. In Panel (b), we plot the relationship between patent novelty and patent value, controlling for technology class and issue week. Again, there is a clear positive relationship between novelty and patent value.

The regression version of Panel (b) is reported in Column 7 of Table 3. To make results comparable across specifications, we show the elasticity of patent value with respect to our novelty measure. We find a positive correlation indicating that patent value increases with increasing novelty. In column 8, we add firm fixed effects and account for multiple patents by aggregating on the firm $\times$ issue week level. Here, the correlation between the patent novelty

and patent value is negative. This change indicates that novelty is a firm characteristic—e.g., a firm with very novel patents on average has relatively less value from its most novel patents than from its less novel patents. However, closer examination reveals that the 15% least novel observations drive this negative (within-firm) relationship between novelty and patent value. That pattern is consistent with Panel (b) of Figure 3, which also shows that the positive relation between value and novelty breaks down if patents have below-average novelty.

To address this problem, Column 9 examines the (within-firm) correlation between KPSS dollar value and the share of a firm-week's patents that are among the 25% most novel patents as an explanatory variable. Again, we find a positive relation between novelty and patent value even if we condition on firm fixed effects.

This evidence suggests that science is associated with novelty, and novelty is associated with a science premium. While correlations are not transitive, it seems plausible that the increase in novelty is at least one path in the causal chain from science to patent value. That said, we do not claim that novelty is the only reason for the observed science premium. For example, scientific insights might make it possible for a patent to make more general claims about which technology is protected. That story is consistent with Table A.6 in Appendix B.2 that shows that patents citing science have, on average, a larger scope than patents that do not cite science.

4.2.2 Why do not all firms invent "on the shoulders" of science?

Our main results suggest that firms reap greater benefits from more novel and science-driven inventions. So why do not all firms choose (or sort) to invent "on the shoulders" of science?

Imperfect information. We first acknowledge the possibility that managers might need to be made aware of the (average) additional value of science-proximate patenting. Given the narratives around patenting in the information technology industry (see Section 2),

executives may underestimate the value of inventing based on recent science. In the remainder of this section, we focus on why perfectly informed firms might still differ in their choice to engage in or reap the benefits of science-based inventing.

Heterogeneity in absorptive capacity. Even though the rewards to science-proximate innovation are higher on average, the fixed costs of accessing that value might differ across firms. Setting up in-house R&D or gathering the expertise needed to build on scientific insights (both named "absorptive capacity" in the following) is costly and may not yield positive net present value for all firms (Cohen and Levinthal, 1990a). Reasons for heterogeneity in absorptive capacity include (but are not limited to) scarcity of talent with expertise in the scientific literature or methods, access to financing, location relative to scientific clusters and talent, as well as path dependencies like "trapped factors" (Bloom *et al.*, 2013), adjustment costs (Chan *et al.*, 2007; Krieger *et al.*, 2022), and imprinting (Levinthal, 2003).

Heterogeneity in absorptive capacity may lead firms to sort into different R&D strategies. As discussed above with regards to firm fixed effects, firm-level sorting based on overall invention capabilities could show up in the aggregate data as a science premium. Accounting for firm-specific factors does suggest a fair amount of company sorting into different types of invention (see Table 1 vs. Table 2). Thus, understanding the choice to use science-based invention requires investigating how the science premium differs across firms.

In a first step, we examine if firms with more patents also have a higher science premium. This is based on the idea that larger inventors can spread the fixed costs of building absorptive capacity over more patents. To do this, we count the number of patents per firm each issue year and define the 100 firms with the most patents as "large."[20] The result is shown in Column 2 of Table 4. The Top 100 patent holders per issue year have a larger science premium than companies outside the Top 100. Companies outside the top 100 have even a smalll negative science premium on average. However, we are cautious to over-interpret those

[20]On average, there are 864 firms per year in the sample. As patenting is concentrated in a few firms, around 77% of all patents are in large firms.

correlations since these coefficients are imprecisely estimated with wide confidence bounds.

In Column 3, we add an indicator of whether the firm is in the Top 100 in terms of D=1 patents in an issue year. By focusing specifically on the number of D=1 patents, rather than overall patenting volume, helps isolate whether the company's R&D strategy is science-based or not. Around 51% of all firms that are in the Top 100 by total patents are also in the Top 100 in terms of D=1 patents, i.e., 49% are not. Firms that are in the Top 100 of D=1 patents have a significantly higher science premium than companies that are not in the Top 100 of D=1 patents. The science premium of being in the Top 100 by total patents is imprecisely estimated, small or even negative once we control for a firm being in the Top 100 of D=1 patents. These results show that the science premium differs systematically across companies. Companies that often build on science also benefit more from building on science, while companies that do not have the scale in building science have a low or negative science premium.

In Columns 4 to 6, we directly look at absorptive capacity. Engaging directly in science via academic publishing is the most direct path by which firms build absorptive capacity for new science (Henderson and Cockburn, 1994; Cockburn *et al.*, 2000; Stern, 2004). We use the patent-paper pair data developed in Marx and Fuegi (2020) and Marx and Fuegi (2022) to measure absorptive capacity. A patent is part of a patent-paper pair if at least one of the inventors is also an author in one of the cited academic articles. We combine this patent-paper pair data with data on whether a patent cites an academic article where at least one of the authors is affiliated with the assignee. In total, we find that 17,192 or 7.2% of all D=1 patents have one or both of these markers of absorptive capacity. We call these patents in the following patent "based on in-house science." In Column 4, we show that patents based on in-house science have a significantly higher science premium than patents that do not (54.0% 95% CI: [14%, 94%]).

If absorptive capacity is a company characteristic, we might still observe a science premium for inventions that build on external science. A company with access to scientific

expertise should be better able to evaluate and use science, even if that work is outside the narrow scientific experiments or expertise of its internal scientists. To evaluate this type of absorptive capacity, we define a firm as having an "In-house lab" if it has 25 or more patents based on in-house science. We restrict to 25 or more patents to observe whether a company consistently invests in absorptive capacity, filtering out one-off collaborations with academic scientists. By this definition, 107 companies have an "in-house lab." In Column 5, we look at all patents and find that companies with an in-house lab have a significant science premium. In Column 6, we drop all patents that directly rely on in-house science, and again run the same regression, interacting indicators for whether the firm has an in-house lab and whether the focal patent cites science. We find virtually the same result, consistent with the notion that absorptive capacity extends outside the immediate expertise of the scientists engaged by the company.

In summary, less science-intensive firms tend to have lower-value patents and gain a smaller benefit (if any) from building on science—both for external and in-house science. Patenting firms with few science-based inventions may be fully informed about the (average) benefits of building on science yet regard the path to developing absorptive capacity as too long, expensive, or risky. While science-based inventing may seem attractive in the context of patent value, the long-term investments required to build on science consistently are considerable in many industries.

5 Conclusion

Our study shows that building more directly on science is associated with more valuable and novel patents. Thus, while scientists since Isaac Newton have been known to see further "by standing on the shoulders of giants," our study suggests that many inventors in the private sector see further by standing on the shoulders of science.

By their very nature, our estimates provide an incomplete picture of the private value

derived from science. Beyond patented inventions, R&D organizations benefit from applying the tools and training originating in the scientific community. These indirect benefits are possible because firms hire scientists and engage with the research frontier (Cohen and Levinthal, 1990a; Henderson and Cockburn, 1994; Stern, 2004).

Our results suggest that science helps firms push the technological frontier by building on more disparate ideas to introduce and combine more novel technologies. While its value is seemingly available for science-driven firms to capture, science's potential in corporate innovation remains an important area for study. How best to access, engage with and build upon the ever-expanding base of scientific knowledge and methods is an exciting challenge for both R&D managers and scholars.

References

Ahmadpoor, M. and Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, **357** (6351), 583–587.

Arora, A., Belenzon, S. and Dionisi, B. (2021). *First-mover Advantage and the Private Value of Public Science*. Tech. rep., National Bureau of Economic Research.

—, — and Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, **39** (1), 3–32.

—, — and Sheer, L. (2020). Knowledge spillovers and corporate investment in scientific research. *The American Economic Review, forthcoming*.

—, Fosfuri, A. and Gambardella, A. (2004). *Markets for Technology: The Economics of Innovation and Corporate Strategy*. The MIT Press, MIT Press.

Arts, S., Cassiman, B. and Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, **39** (1), 62–84.

Azoulay, P., Graff Zivin, J. S., Li, D. and Sampat, B. N. (2019). Public r&d investments and private-sector patenting: evidence from nih funding rules. *The Review of economic studies*, **86** (1), 117–152.

Bhaskarabhatla, A. and Hegde, D. (2014). An organizational perspective on patenting and open innovation. *Organization Science*, **25** (6), 1744–1763.

Bikard, M. and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, **66** (8), 3425–3443.

—, Vakili, K. and Teodoridis, F. (2019). When collaboration bridges institutions: The impact of university industry collaboration on academic productivity. *Organization Science*, **30** (2), 426–445.

Bloom, N., Romer, P. M., Terry, S. J. and Van Reenen, J. (2013). A trapped-factors model of innovation. *American Economic Review*, **103** (3), 208–13.

Chan, T., Nickerson, J. A. and Owan, H. (2007). Strategic management of r&d pipelines with cospecialized investments and technology markets. *Management Science*, **53** (4), 667–682.

Chesbrough, H., Vanhaverbeke, W. and West, J. (2006). *Open Innovation: Researching a New Paradigm*. OUP Oxford.

Cockburn, I. M. and Henderson, R. M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *The Journal of Industrial Economics*, **46** (2), 157–182.

—, — and Stern, S. (2000). Untangling the origins of competitive advantage. *Strategic Management Journal*, **21** (10/11), 1123–1145.

Cohen, W. M. and Levinthal, D. A. (1990a). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, **35** (1), 128–152.

— and — (1990b). Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, pp. 128–152.

Ewens, M., Nanda, R. and Rhodes-Kropf, M. (2018). Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics*, **128** (3), 422 – 442.

Fleming, L. and Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, **25** (8-9), 909–928.

Gambardella, A. (1992). Competitive advantages from in-house scientific research: The us pharmaceutical industry in the 1980s. *Research policy*, **21** (5), 391–407.

Gans, J. S. and Stern, S. (2003). The product market and the market for ideas commercialization strategies for technology entrepreneurs. *Research Policy*, **32** (2), 333 – 350, special Issue on Technology Entrepreneurship and Contact Information for corresponding authors.

Harhoff, D. and Wagner, S. (2009). The duration of patent examination at the european patent office. *Management Science*, **55** (12), 1969–1984.

Henderson, R. and Cockburn, I. (1994). Measuring competence? exploring firm effects in pharmaceutical research. *Strategic Management Journal*, **15** (S1), 63–84.

Iaria, A., Schwarz, C. and Waldinger, F. (2018). Frontier knowledge and scientific production: evidence from the collapse of international science. *The Quarterly Journal of Economics*, **133** (2), 927–991.

Kelly, B., Papanikolaou, D., Seru, A. and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, **3** (3), 303–320.

Kline, S. J. and Rosenberg, N. (1986). An overview of innovation. the positive sum strategy: Harnessing technology for economic growth. *The National Academy of Science, USA*.

Kogan, L., Papanikolaou, D., Seru, A. and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, **132** (2), 665–712.

Krieger, J. L., Li, X. and Thakor, R. T. (2022). Find and replace: R&d investment following the erosion of existing products. *Management Science*.

Kuhn, J., Younge, K. and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, **51** (1), 109–132.

Levinthal, D. A. (2003). Imprinting and the evolution of firm capabilities. *The SMS Blackwell handbook of organizational capabilities: Emergence, development, and change*, pp. 100–103.

Magerman, T., Van Looy, B. and Debackere, K. (2015). Does involvement in patenting jeopardize one's academic footprint? an analysis of patent-paper pairs in biotechnology. *Research Policy*, **44** (9), 1702–1713.

Mansfield, E. (1991). Academic research and industrial innovation. *Research policy*, **20** (1), 1–12.

— (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *Review of Economics and Statistics*, **77** (1), 55–65.

— (1998). Academic research and industrial innovation: An update of empirical findings. *Research policy*, **26** (7-8), 773–776.

Marco, A. C., Tesfayesus, A. and Toole, A. A. (2017). Patent litigation data from us district court electronic records (1963-2015).

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

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

Mowery, D. C. (2009). Plus ca change, industrial rd in the third industrial revolution. *Industrial and Corporate Change*, **18** (1), 1–50.

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: exploring tissue engineering. *Research policy*, **31** (8-9), 1389–1403.

— and Stern, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge? an empirical test of the anti-commons hypothesis. *Journal of Economic Behavior and Organization*, **63** (4), 648–687.

Partha, D. and David, P. A. (1994). Toward a new economics of science. *Research policy*, **23** (5), 487–521.

Poege, F., Harhoff, D., Gaessler, F. and Baruffaldi, S. (2019). Science quality and the value of inventions. *arXiv preprint arXiv:1903.05020*.

Régibeau, P. and Rockett, K. (2010). Innovation cycles and learning at the patent office: does the early patent get the delay? *The Journal of Industrial Economics*, **58** (2), 222–246.

Roach, M. and Cohen, W. M. (2013). Lens or prism? patent citations as a measure of knowledge flows from public research. *Management Science*, **59** (2), 504–525.

Shane, S. (2001). Technological opportunities and new firm creation. *Management science*, **47** (2), 205–220.

Simonsohn, U., Simmons, J. P. and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, **4** (11), 1208–1214.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-j. P. and Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, ACM, pp. 243–246.

Sorenson, O. and Fleming, L. (2004). Science and the diffusion of knowledge. *Research policy*, **33** (10), 1615–1634.

Squicciarini, M., Dernis, H. and Criscuolo, C. (2013). *Measuring patent quality: Indicators of technological and economic value*. Tech. rep.

Stephan, P. E. (1996). The economics of science. *Journal of Economic literature*, **34** (3), 1199–1235.

— (2012). *How economics shapes science*, vol. 1. Harvard University Press Cambridge, MA.

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

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 990–998.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand journal of economics*, pp. 172–187.

von Hippel, E. (1988). *The Sources of Innovation*. Oxford University Press.

Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, pp. 331–360.

Tables & Figures

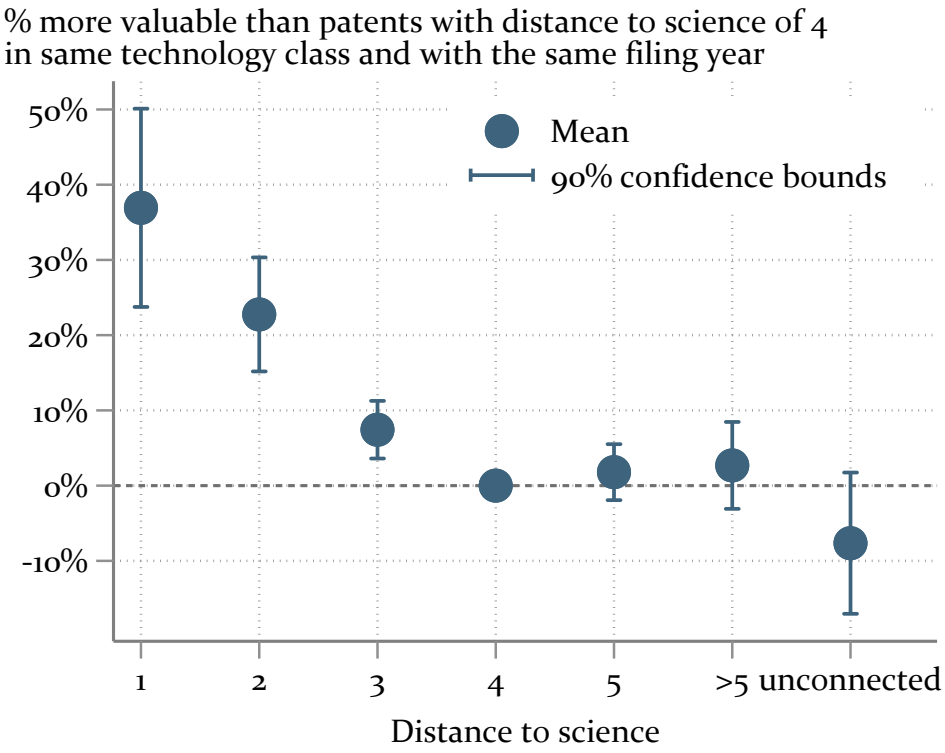
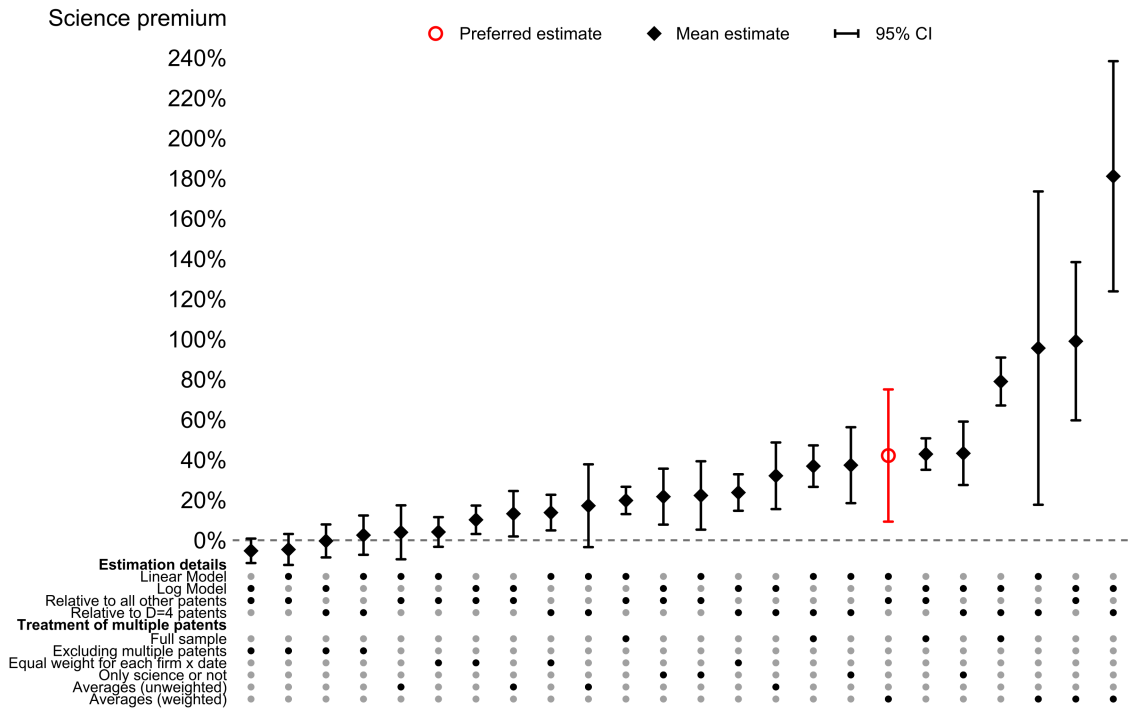


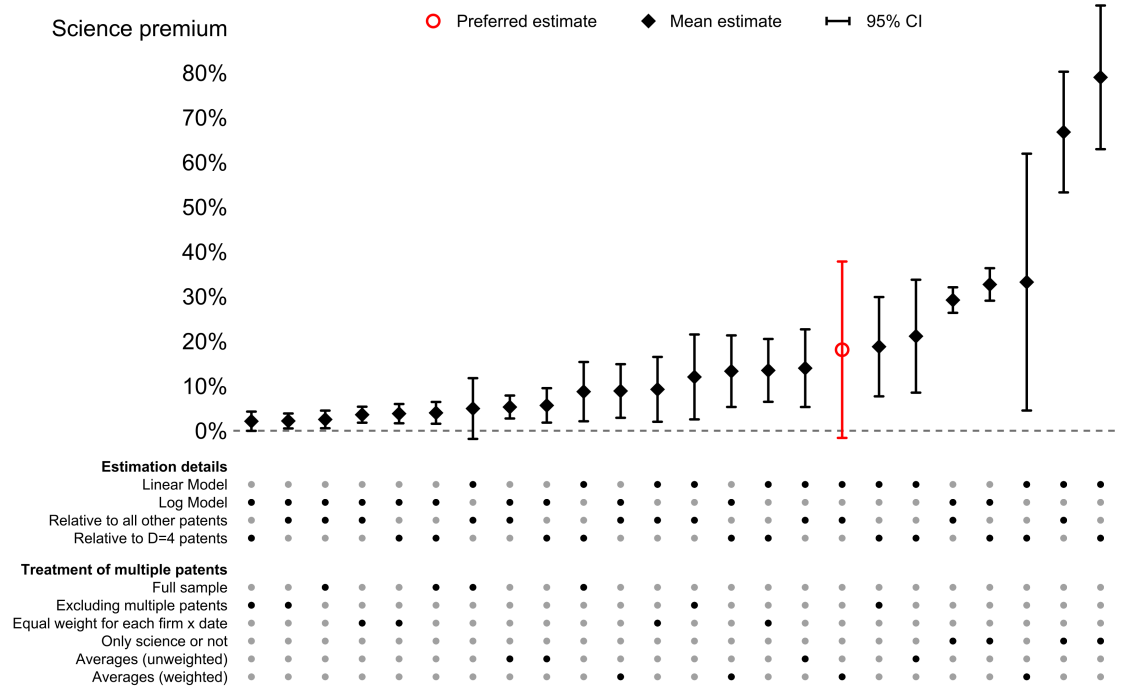
Figure 1: Distance-to-science and patent value

This Figure shows the average increase in patent value for all distances to science relative to patents with a distance of four (D=4) in percent (%). The values of U.S. patents are from Kogan *et al.* (2017). The distance-to-science of U.S. patents is calculated using data from Marx and Fuegi (2020) and the method of Ahmadpoor and Jones (2017). The distance-to-science is defined by citation links. The values correspond to the coefficients in Table 1, Column 2.

,



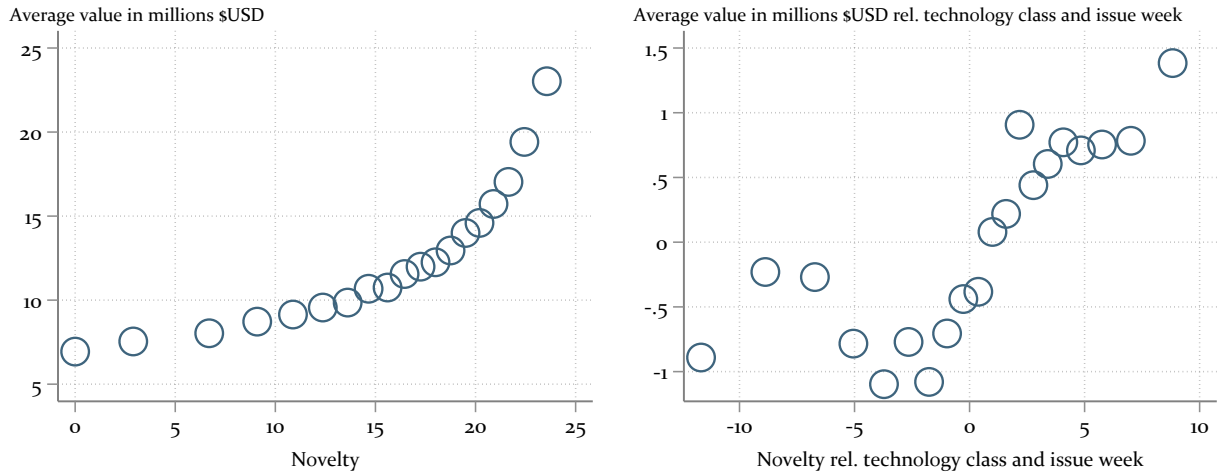
(a) Without firm fixed effects



(b) With firm fixed effects

Figure 2: Specification curve

Note: The specification curves shows the mean science premium estimate and the 95% confidence intervals for all specifications described in Table 5. Each specification includes controls for the patent technology on the CPC4 level and issue week fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level. Panel (a) shows all specifications without firm fixed effects and Panel (b) displays the science premium results with firm fixed effects.



(a) Raw data

(b) Accounting for technology and issue week

Figure 3: Patent novelty and patent value.

Figure 3 shows the relationship between patent text novelty and (KPSS) values. Panel (a) shows the average patent value for each ventile of our novelty indicator, that is based on the likelihood of pairwise combinations of words that occur in a particular patent. Each bubble represents 5% of all patents in our sample. In Panel (b), we plot the average patent value residualized by (four-digit) CPC technology class fixed effects and the issue week fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcome:	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Citations Percent	Prob(Litigation) Percent	Dollar Percent
Distance 1	19.9	36.9	4.0	17.2	42.3	95.8	67.5	70.1	
	[9,31]	[18,56]	[-9,17]	[-3,38]	[9,75]	[18,174]	[60,75]	[52,88]	
Distance 2		22.8		21.4		67.2			
		[11,34]		[8,35]		[15,120]			
Distance 3		7.4		6.1		19.4			
		[2,13]		[-1,13]		[-4,42]			
Distance 5		1.8		-0.4		8.5			
		[-3,6]		[-7,7]		[-10,27]			
Distance >5		2.7		-1.5		10.7			
		[-6,11]		[-11,8]		[-20,41]			
Unconnected		-7.7		-5.3		-22.8			
		[-20,5]		[-18,8]		[-61,16]			
Text similarity									3.0
									[1,5]
Firm FE	No	No	No	No	No	No	No	No	No
Aggregating on firm x issue week	No	No	Yes	Yes	Weighted	Weighted	No	No	No
Mean Dep.	12.8	12.8	18.3	18.3	18.3	18.3	29.5	6.2	12.8
Obs.	1176990	1176990	348179	348179	348179	348179	1176990	1176990	1134632

Table 1: Patent Value and Distance to Science - without Firm Fixed Effects

Table 1 reports regression results relating patent value with distance-to-science. In Columns 1 to 6 and Column 9, the outcome variable is (adjusted) patent values from Kogan *et al.* (2017). In Column 7, the outcome variable is the count of citing patent families. In Columns 8 is the outcome variable an indicator variable for whether or not the focal patent was ever involved in litigation. The litigation outcome variable is based on the data from Marco *et al.* (2017). The calculation of distance-to-science used in Columns 1 to 8 as independent variables is based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Column 9, the independent variable is the similarity of the patent text and the text of the (directly and indirectly) cited academic articles. We report the coefficients in terms of their percentage increases relative to D=4 patents. For example, the coefficient for the first value (Distance 1) in Column 2 may be interpreted as an 36.9% increase in patent value of D=1 patent relative to a D=4 patent. All specifications include issue week and technology fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcome:	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Citations Percent	Prob(Litigation) Percent	Dollar Percent
Distance 1	5.0	9.0	14.0	21.2	18.3	33.5	49.1	43.2	
	[-0,10]	[2,16]	[5,23]	[9,34]	[-1,38]	[5,62]	[43,55]	[28,59]	
Distance 2		5.3		9.4		19.1			
		[-0,11]		[1,18]		[-3,41]			
Distance 3		2.8		1.6		8.8			
		[-1,7]		[-4,7]		[-6,24]			
Distance 5		1.5		3.4		5.5			
		[-2,5]		[-3,9]		[-6,17]			
Distance >5		4.9		8.8		14.4			
		[-3,12]		[-1,18]		[-8,37]			
Unconnected		-5.2		4.2		-10.2			
		[-16,6]		[-9,17]		[-39,19]			
Text similarity									1.0
									[-0,2]
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Aggregating on firm x issue week	No	No	Yes	Yes	Weighted	Weighted	No	No	No
Mean Dep.	12.8	12.8	18.3	18.3	18.3	18.3	29.5	6.2	12.8
Obs.	1175935	1175935	347067	347067	348179	348179	1175935	1175935	1133603

Table 2: Patent Value and Distance to Science - Firm Fixed Effects

Table 2 reports regression results relating patent value with distance-to-science, including firm fixed effects. In Columns 1 to 6 and Column 9, the outcome variable is (adjusted) patent values from Kogan *et al.* (2017). In Column 7, the outcome variable is the count of citing patent families. In Columns 8 is the outcome variable an indicator variable for whether or not the focal patent was ever involved in litigation. The litigation outcome variable is based on the data from Marco *et al.* (2017). The calculation of distance-to-science used in Columns 1 to 8 as independent variables is based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Column 9, the independent variable is the similarity of the patent text and the text of the (directly and indirectly) cited academic articles. We report the coefficients in terms of their percentage increases relative to D=4 patents. For example, the coefficient for the first value (Distance 1) in Column 2 may be interpreted as an 9.0% increase in patent value of D=1 patent relative to a D=4 patent. In addition to filing firm fixed effects, all specifications include issue week and technology fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcome:	Probability of word combinations						Dollar	Dollar	Dollar
	Percent	Percent	Percent	Percent	Percent	Percent	Percent	Percent	Percent
Distance 1	7.6	9.4	5.4	6.5					
	[6,9]	[7,11]	[5,6]	[6,8]					
Distance 2		3.8		2.4					
		[3,5]		[2,3]					
Distance 3		0.6		0.6					
		[-0,1]		[0,1]					
Distance 5		-2.4		-2.4					
		[-3,-2]		[-3,-2]					
Distance >5		-7.6		-6.8					
		[-9,-7]		[-8,-6]					
Unconnected		-9.0		-7.9					
		[-11,-7]		[-9,-7]					
Text similarity					1.7	1.3			
					[2,2]	[1,1]			
Novelty							13.1	-15.2	
							[2,24]	[-28,-2]	
Top 25% novelty									5.2
									[2,8]
Firm FE	No	No	Yes	Yes	No	Yes	No	Yes	Yes
Aggregating on firm x issue week	No	No	No	No	No	No	No	Weighted	Weighted
Mean Dep.	15.4	15.4	15.4	15.4	15.4	15.4	12.8	17.3	17.3
Obs.	1176954	1176954	1175897	1175897	1134629	1133599	1176954	416459	416459

Table 3: Patent value, patent novelty and distance-to-science

Table 3 shows OLS regression results on the relation of patent value, patent novelty and distance-to-science. In Columns 1 to 6, the outcome variable is the probabilty of the focal patent's word combinations (i.e., 1-novelty) based on the data from Arts *et al.* (2018). In Columns 1 to 4 the independent variables are the distance-to-science measure based on Ahmadpoor and Jones (2017) and in columns 5 and 6 the text similarity of the patent text and the text of the (directly and indirectly) cited academic articles. In Columns 7 to 9 the outcome variable is the adjusted Kogan *et al.* (2017) patent values, and the focal patent's novelty is the independent variable. We control in all specifications for (four-digit) CPC technology class level and issue week fixed effects. In Columns 3, 4, 6, 8 and 9 we add a firm fixed effects. The confidence intervals are based on standard errors that are clustered on the firm level.

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome:	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent	Dollar Percent
Distance=1	19.9	-1.2	-22.5	16.4	-11.6	-10.0
	[9,31]	[-24,21]	[-41,-4]	[6,27]	[-27,4]	[-25,5]
I(Top 100)		2.4	5.8			
		[-21,26]	[-27,38]			
I(Top 100) × (D=1)		26.7	-39.2			
		[-3,56]	[-103,24]			
I(Top 100 D=1)			-4.2			
			[-42,34]			
I(Top 100 D=1) × (D=1)			90.2			
			[29,152]			
Based on in-house science × (D=1)				54.0		
				[14,94]		
In-house lab					19.0	19.2
					[-17,55]	[-17,55]
In-house lab × (D=1)					56.3	53.1
					[26,87]	[23,83]
Mean Dep.	12.8	12.8	12.8	12.8	12.8	12.8
Obs.	1176990	1176990	1176990	1176990	1176990	1159798

Table 4: Heterogeneity of the science premium by firm characteristics

Table 4 reports regression results on the heterogeneity of the science premium by firm characteristics. In all columns, the outcome is patent value from Kogan *et al.* (2017) and regressions all include an indicator variable for whether the patent directly cites academic articles or not ("Distance 1" or "(D=1)"). Indicator variables for "(Top 100)" and "(Top 100 D=1)" refer to whether the patent filing firm was one of that year's top 100 patenting firms in terms of overall patenting volume or number of patents citing science (D=1), respectively. The "Based on in-house science" indicates whether at least one of the authors of the patent is also and author in one of the focal patent's cited academic articles. The "in-house lab" extends the same concept to the firm-level, by indicating whether the firm has 25 or more patents based on in-house science. Column 6 excludes all patents that directly rely on in-house science.. We report the coefficients in terms of their percentage increases relative to all patents that do not cite science. All regression include issue week and technology class fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level.

Table 5: Specification curve details

Specification	Description
Estimation details	
Linear model	These specifications estimate the model in Equation 1 or Equation 3 in case of the averages model.
Log Model	In these specifications, we use the natural logarithm of the dependent variable as outcome.
Relative to all other patents	These specification include as explanatory a dummy for whether a patent cites a scientific article or not instead of dummies for all distances to science. This implies that the science premium is measured relative to all other patents that do not cite science.
Relative to D=4 patents	In this specification we classify each patent by its distance to science as in Equation 1 or Equation 3. A patent that directly cites a scientific article has a distance of 1 (D=1), a patent that does not cite a scientific article but cites a D=1 patent has a distance of 2 and so on. Patents that are D=2 are likely to use science input indirectly and therefore might not be a suitable baseline to calculate the science premium. In this specifications, we choose as baseline for the regression patents with a distance of 4, i.e. patents that are only indirectly based on science.
Treatment of multiple patents	
Full Sample	We do adjust for multiple patents and use all patent values as reported by Kogan *et al.* (2017) as outcome.
Excl. multiple patents	Specification drops all firm $\times$ issue week observations with more than one patent. Therefore only patents for which the value of patent is uniquely determined are included.

Equal weight for each firm × date	These specifications give each firm × date observation the same weight by reweighting each patent with the inverse of the number of patents per firm × date combination. This specification is intermediate to the "Full Sample" specification which gives full weight to multiple patents and the "Excluding multiple patents" specification which gives zero weight to observations with multiple patents.
Only science or not	Specification keeps only firm × issue week observations where all patents are citing science or do not cite science. As we are interested in the difference in patent value of patents citing science with patents that do not cite science, the observed patent value averages are informative either about the former (if all cite science) and the latter (if no patents cite science).
Averages (unweighted)	In these specifications we estimate Equation (3) and average the dependent variables and in the independent variable on the firm × issue week level. Without weighing each firm × issue week observations has equal weight in the estimate.
Averages (weighted)	These specifications are the same as "Averages (unweighted)" but we weigh the estimate with the number of patents per firm × issue week. Thus firms with more patents get a larger weight than companies with less patents.

Appendices

A Data

A.1 Data Sources

For our analysis, we calculate distance-to-science for each patent following the method of Ahmadpoor and Jones (2017). We then match this data with patent values calculated by Kogan *et al.* (2017) and with patent characteristics from a variety of sources. We use all patents that have a non-missing patent value and in whose technology class and filing year there is at least one patent with a distance-to-science of four.

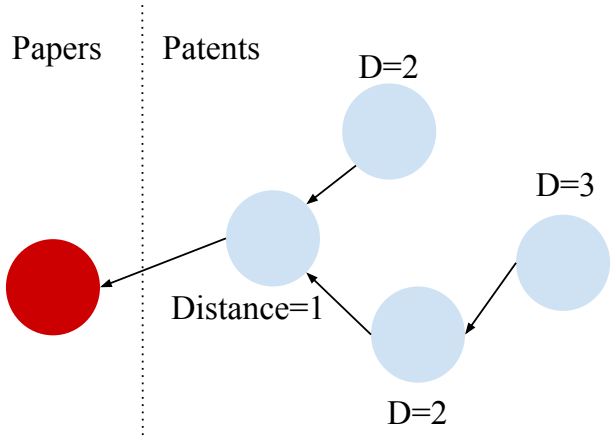


Figure A.1: Distance-to-science

This figure is adapted from (Ahmadpoor and Jones, 2017). It shows the distance to science for patents based on citation proximity to scientific articles.

Distance-to-science: Ahmadpoor and Jones (2017) define a patent's distance to science using citation links.[21] A patent that directly cites a scientific paper has a distance to science of one (D=1). Patents cite academic articles or other patents to give credit to prior art on which the technology disclosed in the patent is based. Patent-to-article citations are used in many recent papers to capture the link between science and innovation, e.g. Arora *et al.* (2020) and (Azoulay *et al.*, 2019).[22] A patent that cites a (D=1)-patent but no scientific article has a distance of two (D=2), and so on (Figure A.1). Citing another patent that is

[21] We thank Mohammad Ahmadpoor and Ben Jones for sharing their data.

[22] Roach and Cohen (Roach and Cohen, 2013) suggest that patent-to-article citations reflect knowledge flows from academia to the private sector better than the commonly used patent-to-patent citation.

based on a scientific article provides evidence that the citing patent is also based to some degree on science, but less directly so.

To determine the distance-to-science of individual patents we use data from Marx and Fuegi Marx and Fuegi (2020), which provides a link from academic articles in Microsoft Academic to patents. Then we use data in PATSTAT to obtain patent-to-patent citations. We cross-check the values of our distance-to-science measure based on Marx and Fuegi (2020) with the values calculated by Ahmadpoor and Jones (2017). In cases where Ahmadpoor and Jones (2017) arrive at a smaller distance-to-science, we substitute their values.

Sources:

https://www.openicpsr.org/openicpsr/project/108362/version/V12/view

https://www.microsoft.com/en-us/research/project/academic/

https://www.epo.org/searching-for-patents/business/patstat.html\#tab-1

Patent value: We match the distance-to-science information with the data on patent values of Kogan *et al.* (2017). Kogan *et al.* (2017) use abnormal stock market returns around the publication date of the patent to infer the value of a patent. Therefore, the data measures the ex-ante expected net present value of the patent for the filing company. This dataset contains patent values for 1.8 million U.S. patents filed between 1926 and 2009.

Source: *https://iu.app.box.com/v/patents*

Patent novelty: For our novelty measure, we use information on words in patents from Arts *et al.* (2018). Arts *et al.* (2018) tokenize the titles and abstract texts of patents, clean and alphabetically sort the resulting words. The resulting word vector contains on average 37 words per patent and in sum 526,561 words. For the novelty measure, we count how

often a particular pairwise word combination occurs in a patent abstract and standardize it with the total number of pairwise word combinations up to this year. We also calculate for each word how common it is. To do this, we count for each word how often it was used in the past and standardize it with the total number of words used. In a last step, to arrive at our final patent novelty measure, we multiply the novelty measure by minus one, such that higher values of the novelty measure indicate more novel patents; we winsorize it at the 5th and 95th percentile to account for outliers and add the minimum to make the resulting values positive.

Source: *https://dataverse.harvard.edu/dataverse/patenttext*

Other patent characteristics:

- **Text similarity**: We calculate the pairwise text similarity between a patent and the articles cited in the patent. Then we take the maximum over all the similarities of a patent to its cited articles to determine the distance to the closest article. To calculate the similarity between the abstracts of the article and of the patent we use the "term frequency-inverse document frequency" (tf-idf) method. We use the "gensim" implementation in Python for our calculations (https://radimrehurek.com/ gensim/). Article abstracts are from the OpenAcademic Graph (Tang *et al.*, 2008; Sinha *et al.*, 2015) and patent abstracts are from Patstat. For each term used in the abstracts of the patent and the article, tf-idf measures how often this word appears in the abstract and then standardizes this value with the probability that this term appears in general. Using the tf-idf value for each term, we can build a word vector for each of the abstracts. Then we determine the similarity between the abstracts of the patent and the article abstract by calculating the correlation between the two-word vectors. If a patent cites several articles, we take the maximum in similarity.

  Source: *https://www.openacademic.ai/oag/* 1

- **Patent Litigation** is from the USPTO's Patent Litigation Docket Reports Data (Marco *et al.*, 2017).

  Source: *https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-litigation-docket-reports-data*

- **All other patent characteristics** are from Patstat–including application dates and (four digit patent classes).

  Source: *https://www.epo.org/searching-for-patents/business/patstat.html*

A.2 Additional Examples:

U.S. Patent number 6120536 ("Medical devices with long term non-thrombogenic coatings") was published in July 2000 and assigned to Schneider USA Inc. (later purchased by Boston Scientific). It describes a drug-eluting coating applied to a metallic stent in order to prevent blood clots.[23] The patent builds directly on science (D=1), with 8 citations to scientific publications and 35 additional non-patent citations (mostly conference presentations and technical reports). These publications include articles from *The Journal of Biomedical Materials Research*, *The Society of Thoracic Surgeons* and the *American Society for Artificial Internal Organs*. One of the two inventors, Michael Helmus, is an author of four of those non-patent citations, two of which are his own grant applications. The patent's word similarity to its average cited publication and most similar scientific publication are both in the in the top quartile of patents within the CPC technology class (A61L)—meaning that the language employed in the patent is highly similar to its most proximate scientific articles. The application itself is rich in data, presenting eight different figures plotting drug release over time, using different coating conditions and concentrations.

[23] As of May 2021, the patent had 566 forward citations.

Just as the science-based McKesson restocking system patent shared the same technology class as the scientifically distant Coca Cola vending machine patent, the drug-eluting stent patent above (6120536) shares the same CPC class as many patents that are mostly or totally disconnected from science (D>5). These include materials that prevent oxidation of medical implants (5543471, D=6), a film that shrinks upon contact with excess water (patent number 5641562, D=5), and air filters which contain tea extracts that might deactivate viruses (5747053, D=5). Clearly, all of the above inventions benefit indirectly from the scientific advances of the modern era from physics, chemistry and microbiology. However, as applied engineering efforts, their search process and method for communicating the invention's distinctive features and value is different since they do not relate their work to formal scientific findings.

	Nb. Patents	Year	Family Size	Nb. Inventors	Days Processed	Claim Words App.	Claim Words Filed
Distance 1	237569	2001.6	6.20	2.88	1098.2	110.6	166.2
Distance 2	388638	2001.7	4.13	2.63	1039.9	113.4	164.2
Distance 3	241155	2000.7	3.93	2.53	923.7	122.8	167.1
Distance 4	107934	1996.6	3.78	2.38	803.7	137.1	174.9
Distance 5	62590	1993.1	3.57	2.22	745.2	141.5	182.4
Distance >5	74778	1991.1	3.43	2.07	702.4	146.1	183.6
Unconnected	64290	1987.3	3.46	1.95	696.7	143.5	171.6
Total	1176954	1999.1	4.36	2.54	950.3	118.9	167.2

Table A.1: Patent Characteristics, by Distance-to-Science

NOTES: Table A.1 presents average patent characteristics for patents in the analysis sample, by degree of distance from science.

	Nb. Cites	Share Self-Cites	Share Same-Tech	Age	StdDev. Age	Max. Sim. Cited	Avg. Sim. Cited
Distance 1	13.3	0.14	0.54	8.01	5.27	0.55	0.31
Distance 2	11.8	0.14	0.56	7.36	5.17	0.54	0.32
Distance 3	9.87	0.17	0.56	8.27	6.05	0.52	0.34
Distance 4	10.2	0.18	0.56	10.6	7.85	0.52	0.35
Distance 5	11.1	0.16	0.56	12.3	8.96	0.50	0.35
Distance >5	10.4	0.14	0.59	14.3	10.6	0.49	0.35
Unconnected	8.87	0.13	0.59	20.3	12.4	0.45	0.37
Total	11.3	0.15	0.56	9.27	6.44	0.53	0.33

Table A.2: Backward Citations, by Distance-to-Science

Table A.2 describes average backwards citation characteristics for patents in the analysis sample, by degree of distance from science.

Outcome:	(1) Nb. Inventors Percent	(2) Processing Days Percent	(3) Claim Words Percent	(4) Nb. Backwards Cites Percent	(5) Share Same-Tech Percent	(6) Avg. Age Cites Percent	(7) Max. Sim. Cited Percent	(8) Avg. Sim. Cited Percent
Distance 1	13.1	10.6	-7.5	68.3	-9.0	-6.1	10.3	-10.5
	[12,14]	[8,14]	[-8,-7]	[66,71]	[-9,-9]	[-7,-6]	[10,11]	[-11,-10]
Distance 2	7.1	5.9	-7.1	35.7	-6.8	-13.0	7.2	-6.9
	[7,8]	[3,9]	[-8,-6]	[34,38]	[-7,-6]	[-13,-13]	[7,7]	[-7,-7]
Distance 3	2.9	-0.7	-5.3	16.2	-3.8	-11.7	1.4	-2.7
	[2,3]	[-3,2]	[-6,-5]	[14,18]	[-4,-3]	[-12,-11]	[1,2]	[-3,-2]
Distance 5	-0.5	1.0	3.8	-14.2	1.9	13.3	-1.7	-0.0
	[-1,0]	[-2,5]	[2,5]	[-16,-12]	[1,2]	[13,14]	[-2,-1]	[-0,0]
Distance >5	-1.8	0.7	4.4	-39.8	8.5	28.2	-3.5	2.5
	[-2,-1]	[-3,4]	[3,6]	[-42,-38]	[8,9]	[27,29]	[-4,-3]	[2,3]
Unconnected	-4.0	0.6	0.9	-52.4	10.6	89.1	-9.2	6.3
	[-5,-3]	[-3,4]	[-2,3]	[-55,-50]	[10,11]	[88,90]	[-10,-9]	[6,7]
Mean Dep.	12.8		12.8	12.8	12.8	12.8	12.8	12.8
Obs.	1176954	1176931	232069	1176954	1154374	1154374	1144852	1144852

Table A.3: Patent Characteristics and Backwards Citations, by Distance-to-Science (Regressions)

Table A.3 presents OLS regression results of distance-to-science on a a variety of patent characteristics. In each regression, the omitted group is patents where degree is equal to four (D=4) and the results are reported in percent differences. The outcome variable in Column 1 is the number of inventors listed on the patent. In Column 2, the dependent variable is number of days between a patent's first application and issuance. Columns 3 show results for number of words in patents' first claim in the issued patent. Column 4's outcome is the total number of backwards citations to other patents. Column 5 uses the share of each patent's backward citations that go to same (4-digit) CPC technology class as the focal patent. Column 6 reports results for the average age of backwards citations. Column 7 and 8's outcomes are the maximum and average similarity of the focal patent's text to the text of it's cited patents. All models include issue week and technology class (4-digit CPC code) fixed effects.

B Additional Regressions and Robustness Results

B.1 Patent Value Outcomes (in Levels), by Distance-to-Science and Text Similarity

Tables A.4 and A.5 are the same as the Tables 1 and 2 but here we report the the coefficients in levels rather than percent changes relative to not citing science or D=4.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcome:	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Citations	Prob(Litigation)	Dollar Million
Distance 1	2.4	4.1	0.7	2.9	5.0	8.4	17.5	3.8	
	[1,4]	[2,6]	[-2,3]	[-0,6]	[1,9]	[3,13]	[15,20]	[3,5]	
Distance 2		2.5		3.6		5.9			
		[1,4]		[1,6]		[2,9]			
Distance 3		0.8		1.0		1.7			
		[0,1]		[-0,2]		[-0,3]			
Distance 5		0.2		-0.1		0.8			
		[-0,1]		[-1,1]		[-1,2]			
Distance >5		0.3		-0.3		0.9			
		[-1,1]		[-2,1]		[-2,4]			
Unconnected		-0.8		-0.9		-2.0			
		[-2,1]		[-3,1]		[-6,2]			
Text similarity									4.1
									[1,7]
Firm FE	No	No	No	No	No	No	No	No	No
Aggregating on firm x issue week	No	No	Yes	Yes	Weighted	Weighted	No	No	No
Mean Dep.	12.8	12.8	18.3	18.3	18.3	18.3	29.5	6.2	12.8
Obs.	1176990	1176990	348179	348179	348179	348179	1176990	1176990	1134632

Table A.4: Patent Value and Distance to Science in Levels - without Firm Fixed Effects

This table reports regression results about the relation of patent value with distance-to-science controlling across all patents in the sample. In Columns 1 to 6 and Column 9, the outcome variable is (adjusted) patent values from Kogan *et al.* (2017). In Column 7, the outcome variable is the count of citing families. In Columns 8 is the outcome variable an indicator variable for whether or not the focal patent was ever involved in litigation (multiplied by 1000). The litigation outcome variable is based on the data from Marco *et al.* (2017). The calculation of distance-to-science used in Columns 1 to 8 as independent variables is based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Column 9, the independent variable is the similarity of the patent text with the text of the (directly and indirectly) cited academic articles. We report the coefficients in terms of their percentage increases relative to D=4 patents. For example, the coefficient for the first value (Distance 1) in Column 1 may be interpreted as an 37% increase in patent value of D=1 patent relative to a D=4 patent. All regression include issue week and technology fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Outcome:	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Dollar Million	Citations	Prob(Litigation)	Dollar Million
Distance 1	0.6	1.1	2.5	3.6	2.3	3.7	13.2	2.5	
	[-0,1]	[0,2]	[1,4]	[2,5]	[-0,5]	[1,6]	[12,15]	[2,3]	
Distance 2		0.7		1.6		2.1			
		[-0,1]		[0,3]		[-0,4]			
Distance 3		0.3		0.3		1.0			
		[-0,1]		[-1,1]		[-1,3]			
Distance 5		0.2		0.6		0.6			
		[-0,1]		[-0,2]		[-1,2]			
Distance >5		0.6		1.5		1.6			
		[-0,2]		[-0,3]		[-1,4]			
Unconnected		-0.6		0.7		-1.1			
		[-2,1]		[-1,3]		[-4,2]			
Text similarity									1.3
									[-0,3]
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Aggregating on firm x issue week	No	No	Yes	Yes	Weighted	Weighted	No	No	No
Mean Dep.	12.8	12.8	18.3	18.3	18.3	18.3	29.5	6.2	12.8
Obs.	1175935	1175935	347067	347067	348179	348179	1175935	1175935	1133603

Table A.5: Patent Value and Distance to Science in Levels - Firm Fixed Effects

This table reports regression results about the relation of patent value with distance-to-science controlling for firm fixed effects. In Columns 1 to 6 and Column 9, the outcome variable is (adjusted) patent values from Kogan *et al.* (2017). In Column 7, the outcome variable is the count of citing families. In Columns 8 is the outcome variable an indicator variable for whether or not the focal patent was ever involved in litigation (multiplied by 1000). The litigation outcome variable is based on the data from Marco *et al.* (2017). The calculation of distance-to-science used in Columns 1 to 8 as independent variables is based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Column 9, the independent variable is the similarity of the patent text with the text of the (directly and indirectly) cited academic articles. We report the coefficients in terms of their percentage increases relative to all other patents or D=4 patents. All regression include issue week and technology fixed effects. The 95% confidence intervals are based on standard errors that are clustered on the firm level.

B.2 Comparing quality indicators for patents citing science and not citing science among multiple patents issued to same firm in same week

Table A.6 shows that when we evaluate patents issued to the same firm on the same day, we still see that patents citing science are different on a number of important patent value, scope, and content measures.

Table A.6: Comparing quality indicators for patents citing science and not citing science among multiple patents issued to same firm in same weeks

	Not citing science	Citing science	Diff	P-Value
Kogan et al. (2017)				
KPSS patent value	18.4	18.4	0.0	1.00
USPTO Patent Number and Case Code File				
Litigated x 1000	5.6	8.1	2.5	0.00
OECD patent quality data (Squicciarini et al., 2013)				
Patent renewal	11.8	12.0	0.2	0.00
Patent family size	4.4	5.0	0.6	0.00
# of claims	17.1	19.2	2.1	0.00
Patent Scope	1.8	2.0	0.2	0.00
Grant lag / 100	9.4	10.1	0.7	0.00
Forward citations - 5 years	9.5	14.6	5.1	0.00
Forward citations - 7 years	13.6	20.6	7.0	0.00
I[Top 1% of cited patents] x 100	0.9	2.0	1.1	0.00
Generality index x 100	48.3	50.8	2.5	0.00
Originality index x 100	72.8	75.1	2.4	0.00
Radicalness index x 100	36.6	36.5	-0.0	0.91
Kelly et al. (2021) patent quality data				
Patent quality - 10 yr	1.1	1.1	0.1	0.00
... net of grant year x 100	3.2	9.4	6.2	0.00
I[Top 10% in patent quality] x 100	11.5	16.9	5.3	0.00

Note: This table shows average values for different quality indicators for patent citing science and patents not citing science that were issued on the same day and to the same firm. To arrive at these averages we first take averages by firm and issue week and the average of those averages across our analysis same. The table then compares patents that cite science and do not cite science holding firm and issue week constant. The first row, KPSS patent value, is mechanically the same for both groups because KPSS assigns the same value to all patents from the same firm-week. The p-values result from a t-test with unequal variances.