Standing on the shoulders of science

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

September 8, 2022

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 and establish three new facts about the relationship between science and the value of inventions. First, we show that patent value is decreasing in distance-to-science. Patents building directly on scientific publications are on average 26% more valuable than patents in the same technology which are disconnected from science. Patents closer to science are also more likely to be in the tails of the value distribution (i.e., greater risk and greater reward). Next, we use patent text analysis to show that patent value increases with patent novelty. Third, we find that science-intensive patents are more novel. We discuss firm heterogeneity and the causes behind the science premium. Overall, firms that consistently use science for invention produce higher value patents generally, and especially when they "build on the shoulders" of their own scientific work.

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 are the bedrock of industrial R&D, only some of those activities build directly on science—translating discoveries from laboratories and scientific publications into novel inventions and commercial products. Other corporate innovation efforts rely only indirectly on science—experimenting, tinkering, optimizing, and inventing without the aid (or constraints) of "the republic of science," but still using tools and technologies enabled through centuries of scientific advance. Firms' level of engagement in science is an important component of their R&D strategy and a potential source of talent and competitive advantage (Henderson and Cockburn, 1994; Cockburn *et al.*, 2000; Stern, 2004). Yet, surprisingly little is known about how building on scientific knowledge affects the value of a firm's inventions.

In this paper, we investigate how firms generate value by inventing "on the shoulders of science." This question is critical for R&D strategy, as investing in the expertise required to absorb scientific knowledge (e.g., in-house scientists or cooperative projects with universities) is both capital- and time-intensive. Such costs are only justified if managers expect a significant science premium, i.e., that inventions based on science are more valuable than inventions that do not use science as an input.

While the social value of science has been immense, formal science is not an obvious source of competitive advantage. Scientific advances are usually published, so that everyone can access them. Thus, basic theory would suggest that any science premium should be competed away by entrants, and access to science should not confer a sustainable competitive advantage. Furthermore, popular discourse in the technology sector and prominent strands of the management literature are highly skeptical of science-based inventing, instead favoring the view that valuable inventions result from applied industrial engineering and user innovation (see Section 2.1). By quantifying the relative value of inventions by their proximity to scientific publications, we aim to inform R&D strategy and adjudicate between opposing narratives on the (private) value of science. After empirically establishing a proximity-to-science premium, we further probe why science-led R&D produces a different patent value distribution. Our analysis provides evidence that science is primarily a tool for exploring and articulating novel idea space—propelling inventive activity towards both higher risk and higher reward areas. Finally, we evaluate how firms differ in their ability to capitalize on the benefits of science. We discuss frictions that inhibit firms' application of science and create persistent performance differences in their invention quality.

To quantify the value of science in corporate R&D, we measure how different degrees of building on science contribute to the private value of patents. Any attempt to measure this contribution is complicated by the fact that science plays a larger role in some technologies than in others (Stephan, 1996). This makes it difficult to distinguish how much of the value of an invention is due to its proximity to science and how much of it is technology-specific. We deal with this challenge with the help of a metric for a patent's level of science-intensity. By comparing the values of more and less science-intensive patents within different technology classes, we can isolate the science component and the technology component of the value of each invention.[1]

To classify patents with respect to their distance-to-science, we build on Ahmadpoor and Jones (2017). When a company files for a patent, it has to list all prior art on which the patents build, including scientific articles. This provides a direct link between the patent and the scientific knowledge of which it makes use. 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 of (D=2), and so on. We match this data with the patent values from Kogan *et al.* (2017), which we also refer to as KPSS (as shorthand). KPSS derive patent values from excess stock returns of the filing company around the date of the patent publication. Combining these two data sets, we can calculate

[1] Conceptually, we view our analysis as a conservative estimate of the private value of science. Beyond the direct value captured through patents, scientific knowledge may generate private value through several additional channels. For example, firms benefit from the productivity-enhancing features of working with science-enabled communications technologies and computing systems breakthroughs. Both patented and non-patented inventions rely on knowledge generated by the scientific community, even when the practitioners involved do not formally cite scientific journal articles. For example, firms routinely hire Ph.D. scientists and engineers whose knowledge and techniques are products of their frontier scientific training, even when their subsequent R&D output does not link directly to published research.

the average patent value for a given distance-to-science for 1.1 million U.S. patents filed between 1980 and 2009.

Our main analyses describe three key correlations. First, we find that patents directly based on science (D=1) have an average private value that is 26% greater than patents filed in the same technology class and year, but only loosely related to science (D=4). Patents with a distance of two (D=2) or three (D=3) from science have private values of 18% and 7% greater than the (D=4) group, after controlling for technology $\times$ year. This propagation of value generated by science to patents that are not directly science-based suggests that scientific progress can be the "remote dynamo of technological innovation" throughout the economy (Stokes, 2011, p.84), with effects beyond the immediately useful applications. Yet, we also show that more science-intensive patents are more risky; i.e., more likely to end up in the tails of the value distribution. In auxiliary results, we show that our main findings are stable when using alternative measures for distance-to-science, based on text similarity and when using measures for patent value based on citations, patent scope, and patent litigation.

As our second finding, we identify a link between patent novelty and private sector value. To establish this link, 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.

Third, 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. In addition to the correlation between novelty and distance-to-science, these characteristics appear to contribute independently to patent value. Both patents that are below and above median novelty, have average valuations increasing in proximity to science, but novel patents enjoy slightly larger average values at each distance from 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"? In the last section of this paper, we evaluate whether imperfect information, risk aversion, a value cliff in the supply of ideas, or R&D cost heterogeneity can account for this phenomenon. While all of these reasons might account to some degree for differing company R&D strategies vis-a-vis science, we show supporting evidence of R&D cost heterogeneity being a key explanation. Thus, our analysis implies that the overall payoff from building on science depends on both the science premiums reported in this paper and the cost of the firm's (cumulative) investment in absorptive capacity.

Our paper contributes to the literature in three main 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 and dollar terms. Regarding intent, this is close to the early surveys of Edwing 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).

By estimating the private value of science-based patents, 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 is a source of competitive advantage— one that is not equally available to established firms that lack experience with science-based invention (Cohen and Levinthal, 1990; Henderson and Cockburn, 1994). 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).

Second, our results demonstrate a key risk-vs-reward tradeoff in science-based innovation. Our findings show that proximity to science is not only associated with greater average private patent values, but also with more risk—i.e., the likelihood of extreme outcomes on both ends of the value distribution. Such tail-risk suggests that science-based R&D is akin to the exploration arm of classic "explore vs. exploit" models (March, 1991; Manso, 2011; Azoulay *et al.*, 2011; Akcigit and Kerr, 2018). Even if the expected value of a science-based patent is larger, risk-aversion or path dependence in R&D decision-making (Cyert *et al.*, 1963; Argote and Greve, 2007; Hall and Lerner, 2010; Eggers, 2012; Chan *et al.*, 2007; Krieger *et al.*, 2021) might lead firms away from science-driven R&D.

Third, 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. 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) argue that science alters inventors' search processes and leads them to useful new knowledge combinations. By linking patent 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]

[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.* (2018) have shown 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 both by citation distance and by text similarity between articles and patents. In a new working paper Arora *et al.* (2021a) use similar data to evaluate how participation in science and first-mover advantage (in building on science) affects the private value of patents. Different from our analysis, the authors focus on within-firm variation.

2 Context: Science in Invention

In this section, we first outline competing narratives on whether science makes a valuable contribution to private sector inventions. In a second step, we explore the question of how science can help companies invent.

2.1 Competing narratives

The canonical anecdotes of technology history are filled with famous private sector inventions that used modern science as a springboard for breakthroughs. Ferdinand Braun and Guglielmo Marconi could not have developed the wireless telegraph before Heinrich Hertz demonstrated the existence of electromagnetic waves. The development of the transistor at the Bell Laboratories would have been difficult to imagine without the scientific understanding of the physics of semiconductors. Similarly, the biotechnology industry was born out of the pioneering scientific work of academic scientists-turned-entrepreneurs like Genentech founder Herbert Boyer.[3] However, technology and management scholars have argued that these cases, while indeed powerful examples, are the exception rather than the rule.[4] This alternative view places applied industrial engineering and user innovation at the center of the innovation process. Scholars have frequently noted how novel science gets "trapped in the ivory tower," or stalls in commercialization due to frictions in knowledge flows, intellectual property, contracting, and the reliability of academic science (Goozner, 2005; Butler, 2008; Harris, 2011; Osherovich, 2011; Freedman *et al.*, 2015; Bikard, 2018).

Recent trends in the corporate landscape also cast doubt on the relative value of science on private sector innovation. Large firms have retreated from internal scientific research (Arora *et al.*, 2018, 2020), while venture capital investments have moved towards faster experimentation and less capital-intensive software-based business models (Ewens *et al.*, 2018). In the

[3] Boyer published some of the seminal papers on recombinant DNA as a professor at the University California, San Francisco prior to cofounding Genentech alongside venture capitalist Robert Swanson.

[4] These skeptics assert that inventions are born from other sources outside formal scientific study (Kline and Rosenberg, 1986; von Hippel, 1988).

cultural and business zeitgeist, slow-moving research endeavors like PhDs, postdocs and peer review do not conform to a technology era dominated by the famous Mark Zuckerberg motto "move fast and break things."[5] This shift might reflect a gradual realization that the rewards of translating frontier science into products are not worth their challenges—or at least in comparison to more narrowly focused applied engineering and software development activities. Alternatively, science-based innovation might indeed provide superior expected private value relative to other sources of innovation, in which case the corporate withdrawal from in-house science is merely a symptom of R&D risk aversion, changes in the cost structure of science-intensive inventing, or broader trends in the organization of corporate innovation.[6]

2.2 Science as tool for navigation and novelty

Why might building on science lead to a higher expected value of corporate patents? For the purpose of invention, science may serve as a map for technological search (Fleming and Sorenson, 2004). With this perspective, science enables more efficient invention by documenting both promising paths (strong "shoulders" to stand on) and dead-ends to avoid.[7] However, the "republic of science" often prizes speed, novelty and individual credit over search efficiency (Partha and David, 1994; Stephan, 2012). Thus, even though science offers tools that push the technology frontier to new heights, science also generates (more than) its fair share of irreproducible findings and "shaky shoulders" for both scientists and inventors to

[5]While product cycles have sped up and R&D teams have adopted "lean" methodologies at technology firms, evidence shows that the organization of academic science has moved in a different direction. Scientific productivity increasingly requires more specialization, larger teams, and more resources in order to overcome the "burden of knowledge" and navigate complex problems (Jones, 2009; Wuchty *et al.*, 2007; Bloom *et al.*, 2020).

[6]For example, the move towards "open innovation" and accessing new innovations through the markets for technology (Chesbrough *et al.*, 2006; Arora *et al.*, 2004; Gans and Stern, 2003; Mowery, 2009; Bhaskarabhatla and Hegde, 2014) has allowed downstream firms in some industries to forgo scientific research under their own roof, while still accessing the technological offspring of those research activities via markets for technology.

[7]At the extreme, scientific articles not only provide a "map" or "foundation" for new inventions, they also directly produce the invention itself. Patented inventions and scientific projects are sometimes co-produced and co-disclosed as patent-paper pairs (Murray, 2002). These pairs describe the same (or highly similar) discoveries and may be best identified using their overlapping language (Magerman *et al.*, 2015). In the words of a patent attorney, such patent applications usually start by "slapping a patent coversheet" on the text of a draft scientific article. On average, patent-paper pairs are associated with more forward citations to the scientific article (Murray and Stern, 2007).

build on (Osherovich, 2011; Begley and Ellis, 2012; Begley and Ioannidis, 2015; Freedman *et al.*, 2015; Azoulay *et al.*, 2015). Thus, science's incentive system pushes the methods, ideas, and language of science towards the exploration end of the explore-exploit spectrum (March, 1991; Azoulay *et al.*, 2011; Akcigit and Kerr, 2018). More novelty is associated with both more risk and more (potential) reward.

In summary, while there is reason to believe that scientific input can contribute to the inventive process, there is no consensus on the value of that input. In the next section we outline the data that we use to adjudicate this open question.

3 Data

Our starting point is a dataset which contains information on the monetary value of 1.8 million public firm patents from 1926 to 2009 (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 2.5 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 of two (D=2), and so on (see Figure A.1 in Appendix A.1). 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 1.1 million U.S. patents filed between 1980 and 2009. 21% of all patents directly cite a scientific article (D=1), 55% are indirectly based on science (D=2 and D=3) and 24% are not based on science (D=4 or larger). The (unadjusted) average value of a patent is $12.9 million in constant 1982 U.S. dollars.

Our primary measure of novelty is built from data on words used patents Arts *et al.* (2018). We calculate how often a given pairwise word combination occurs relative to other patent word combinations, up to a given year. In additional analyses we also use text-based measures of patent similarity to scientific articles cited in the patent, word age, and structural novelty of patented chemicals.

Appendix A.1 gives a detailed description of all the data construction and sources.

Examples: Distance-to-Science Specific examples are useful for understanding the range of inventions represented in the data. Even among well-cited patents, we observe large 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 of which are detailed technical drawings of the cooling system and its 80 different enumerated components (e.g., control valves, linear actuator, 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 different 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, as well as on 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 technologies such as medical devices, telecommunications, transportation, machine tools, and food processing. Therefore, many of our regression specifications use technology class $\times$ year fixed effects to restrict comparisons to more comparable subgroups.

Appendix A.2 provides additional examples using patents from CPC class A61L ("Methods or apparatus for sterilising 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

In the following section, we first map the relation between private sector value and distance-to-science. We show that patents closer to science have a higher value than patents further away from science. In a second step we show that patents that are more novel are also more valuable. Lastly, we show that patents closer to science are also more novel, highlighting a potential reason why science-based patents are more valuable.

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

Our ***first fact*** documents the relationship between patent values and distance-to-science. We show that more science-intensive patents are on average more valuable, but also riskier; i.e., they are more likely to be in the tails of the value distribution.

[Insert Figure 1 Here]

We start by presenting how relative dollar values of patents differ by distance from science. To estimate relative percentage (%) differences in KPSS values across distance-to-science groups, we run ordinary least squares (OLS) regressions with (D=4) as the baseline (omitted) distance-to-science group and KPSS values as the outcome.[8]

We then convert the regression coefficients to percentage differences relative to (D=4) average values. As shown in Panel (a) of Figure 1, a science-based patent that directly cites an academic article (D=1) has an average value that is 82% 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 or three have average values 44% and 14% higher than those in (D=4), respectively. When patents have a distance from science larger than four (D=5, D>5 or "unconnected"), their average values are between 6% and 16% less valuable

[8] Distance-to-science of four as the baseline is a basically arbitrary choice, but one informed by the data, since the relative value premium begins to flatten after (D=4). Intuitively, the connection to scientific publications is quite tenuous at (D=4), and using this baseline group also ensures that our estimates are somewhat conservative, as oppossed to using higher degrees as the baseline.

than those with distance of four. Column (1) of Table 1 presents the same results in regression form.

[Insert Table 1 Here]

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 (Table A.4).[9] We interpret the dollar magnitudes with caution, adjusting the KPSS values downwards using the most conservative estimate of ex-ante probability of patent grant. Doing so deflates the values by 12%.[10] Thus, we find that science-based patents that directly cite an academic article (D=1) have an average value of $15.82 million dollars, which is $8.27 million more than the average value of patents with (D=4) and $8.69 million more than those unconnected to science.[11] The average patent values decline to $12.49, $9.90, and $8.69 million for distances of two, three and four, respectively.

The higher average value of science-based patents reflects an upward shift in the value distribution of patents with higher science intensity. Panel (b) of Figure 1 plots the share of science-intensive patents (D=1, D=2 and D=3) and the share of less science-intensive patents (D=4, D=5, D>5 and unconnected) over the percentiles of the value distribution of all patents. If the value distribution of more science-intensive patents were the same that of less science-intensive patents, the share of patents at each percentile should be 1%. Figure 1,

[9] To arrive at dollar values for individual patents, (Kogan *et al.*, 2017) makes a number of assumptions about the distribution of firm returns and the (ex-ante) probability of patent grants. While the (Kogan *et al.*, 2017) results appear fairly robust to alternative distributional assumptions (see Footnote 11 in (Kogan *et al.*, 2017)), we focus on percentage differences as our primary results since our interest is in the relative value differences patents of different characteristics. Doing so allows us to apply the consistent quantitative valuation method of (Kogan *et al.*, 2017), without relying too heavily on any assumptions that move the magnitude in any direction.

[10] Carley *et al.* (2015) find that acceptance rates vary between 50% and 60% in the 1991–2001 period. We use the low end of that range (50%) instead of the average (56%), which is used in Kogan *et al.* (2017). This adjustment deflates all the patent dollar values by 12%. Conceptually, this adjustment provides a more conservative set of estimates by increasing the amount of market information "surprise" associated with the patent grant. That increased surprise could be a result of more conservative expectations about the likelihood of patent issuance, or imperfect information regarding the existence of pending patent applications. Table A.5 in Appendix B.2 shows how different assumptions on patent grant probabilities affect patent valuations.

[11] 7.13/8.69 = 0.82%, the percentage difference in value of a science-based patent relative to (D=4) patent mentioned above.

Panel (b) shows that there are fewer science-intensive patents at the lower end of the value distribution, while there are more at the upper end. The pattern for less-science-intensive patents is (mechanically) reversed. They are overrepresented at the lower end of the value distribution, while they are significantly underrepresented at the upper end.

We next examine whether this regularity between distance-to-science and patent value simply reflects differences across technologies, perhaps because science is used predominantly in technologies that are on average more valuable. 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." This is why we consider how much of the patent value is technology-specific and how much can be attributed to the value of science.

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 a technological component, a proximity to science component, and an idiosyncratic component, and that these components are additively separable. The technological component is assumed to be the same for all patents with the same technology class and the same filing year, independent of their distance to science. The science component is present in patents closely based on science, while it is absent in patents unrelated to science. The idiosyncratic component captures the patent value residual after accounting for the science and technological components. We assume that the idiosyncratic component has an expected value of zero.

Under these assumptions, we can isolate the technological component through the value of patents that are distant from science. The value of non-science-related patents is the sum of the technological component and the idiosyncratic component, where by definition, the proximity to science component is zero. As the technological component is assumed to be the same for all patents in the same technology class and year, we can filter out the idiosyncratic component by taking averages.

$$
\begin{aligned}
Y_{ijt} &= \alpha_0 + \alpha_1 \times (D=1)_i + \alpha_2 \times (D=2)_i + \alpha_3 \times (D=3)_i \\
&+ \alpha_4 \times (D=5)_i + \alpha_5 \times (D>7)_i + \alpha_1 \times (D=unconnected)_i + \delta_{jt} + \varepsilon_{ijt}
\end{aligned}
\tag{1}
$$

where Y_{ijt} is the patent level outcome (e.g., KPSS value, citations, text novelty) for patent i that is associated with (CPC 4-digit) technology class j and filing year t. The variables (D=1,2,3,5,>5, unconnected) are indicator variables for distance-to-science, where (D=4) is the reference category. δ_{jt} is the technology class $\times$ filing year group fixed effect (i.e., the technology component).

Figure 1, Panel (c) and Table 1, Column (2) present the average patent value % premium—within technology class and filing year—relative to (D=4), by distance-to-science. Patents that are directly based on science (D=1) have an average science-value 26% greater than that of the average (D=4)-patent of the same technology class-year. Patents indirectly based on science with distances of (D=2) and (D=3) have implied values of 18% and 7% greater than (D=4) patents, respectively. These values are lower than the raw values presented in Panel (a) of Figure 1—indicating that science-intensive patents are more prevalent in high-value technology classes than in low-value classes. However, the within technology-year estimates mean that proximity to science has a meaningful relationship with patents' private values.

Our main patent value regression specifications do not include firm fixed effects because our primary goal is to understand how the *choice* to build on science is associated with patent value. While those choices vary project by project, they may also reflect firm-level decisions on how to organize R&D. As such, some of the average effects in our distance-to-science coefficients may be due to differences in firm quality and capabilities. Firms that tend to invent closer to the scientific frontier may simply produce better (or worse) patents on average—regardless of the given patent's distance-to-science. Furthermore, "better firms" may invent closer to the scientific frontier because those firms, as a function of their competitive position, have both the talent and resources to engage in such costly exploration.

However, the firm-sorting explanation does not fully explain the results. Once we account for firm-specific factors, the overall patterns from Table 1 still hold. Appendix B.3 and Table A.6 describe our approach and the results with firm fixed effects. Overall, we find that that average patent values are decreasing in distance from science when we estimate the same same coefficients *within* firms. Controlling for firm fixed effects (along with technology $\times$ year), we find that (D=1) patents still have an average value between 9%–25% greater than patents at the same firm with distances of (D=4), although the dollar amount of the premium is smaller. Overall, the firm fixed effects results are consistent with some firm-quality sorting into science-based innovation, yet the proximity to science premium still persists within firms.

In Figure 1, Panel (d), we show the distribution of the sum of the science value component and the idiosyncratic component; i.e., the residual in value that is not due to the technology and year, across the percentiles of the value distribution.[12] Less science-intensive patents tend to have values close to the median of the value distribution. Science-intensive patents, by contrast, are more likely to have a value that is in the extremes of the science value distribution. Relative to a (D$\geq$4)-patent in the same technology and year, more science-intensive patents are more likely to be in the upper and the lower tails of the value distribution.[13] This suggests that the value premium of science, over and above the value of the technology, comes at the price of an increasing risk of tail outcomes. One potential explanation could be the high rate of irreproducible research results which has been estimated as high as 50 percent (Osherovich, 2011; Freedman *et al.*, 2015). Thus, the science value premium may to some extent be the compensation for the risk that investors associate with science-intensive patents.

[12] We cannot separately identify the science-value from the idiosyncratic value component for a particular patent.

[13] Table A.9 in Appendix B.8 shows that the relationship between distance-to-science and the probability of a patent being in the top or bottom 5% of the value distribution. After controlling for technology and filing year, we see that patents closer to science are more likely to be in both extreme ends of the value distribution.

Alternative Measures of Patent Value: Citations, patent scope and litigation

Patent private value 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 3-6 in Table 1 explore three other proxies for patent value: forward citations, patent scope and propensity to be involved in litigation.

The results are broadly consistent with the KPSS value regressions. Column 3 shows the patent forward citation results. Controlling for tech class $\times$ year fixed effects, we find that patents with (D=1) average 99% more forward citations than patents with (D=4). The relative difference is even stronger than the stock market valuation premium—indicating that social returns through knowledge flows may be above and beyond what the inventor firms capture.[14] The citation premium is is 36% and 9% for (D=2) and (D=3), while patents with (D$\geq$5) have fewer citations than the (D=4) group.

Next, we evaluate the relationship between distance from science and patent scope. Intuitively, patents with greater scope are more valuable by claiming a broader swath of intellectual property space for their assignees, who can then more easily exclude competitors from their technology's domain.[15] We merge our set of patents to the patent scope index from Kuhn and Thompson (2017), and run an ordinary least squares specification, including tech class $\times$ year fixed effects. Column 4 shows that patent scope is increasing in proximity to science. The difference in scope between a (D=1) and (D=4) patent within the same tech-year is equivalent to a 100% increase over the sample mean.[16]

Columns 5 and 6 repeat the exercise with the binary litigation outcome, both with and without tech class $\times$ year fixed effects. While we can only measure litigation as a binary

[14] 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), and unlike KPSS, they are an ex-post measure of quality, so we cannot easily quantify the gap between private and social value capture.

[15] Kuhn and Thompson (2017) show that a patent's scope may be reliably measured using the number of words in its first claim.

[16] The sample mean is 0.12 ("Mean Dep." in Column 4 of Table1) and the regression coefficient of (D=1) is 0.12.

event, it is a useful signal of patent value.[17] Patent litigation is expensive, so firms will (for the most part) only fight in court for patents that are believed to be valuable.[18] With and without class $\times$ year fixed effects, we find that the likelihood of litigation is increasing in proximity to science. Column 5 indicates that patents with (D=1) have a more than double increase in likelihood of litigation relative to D=4 patents (10.36 vs. 4.39 per 1,000 patents). Column 6 shows that even within tech class and vintage, patents that build more directly on science are more likely to end up in court. Taken together with the KPSS, citation and scope results, this finding shows that patents which 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.

Additional Results and Robustness

Appendix B presents a number of additional results and robustness checks. Appendix B.4 shows the results by broad (1-digit) CPC classes. Appendix B.5 shows how the distance-to-science measure effectively captures which patents draw more directly on the language and ideas of their associated scientific journal articles. Appendixes B.6 and B.7 show robustness to alternative distance-to-science measures. Appendix B.8 describes the regressions examining the tails of the value distribution. We show that the main distance-to-science results hold up across different percentiles of the value distribution in Appendix B.9.

[17] 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.

[18] 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)

4.2 Patent novelty and patent value

Next, we study whether the value of a patent is related to the novelty of its content. If the goal of science is to advance knowledge by making new discoveries, then inventions relying directly on science have the potential to introduce more novel ideas. For these analyses, we construct a new measure of patent novelty. Using this measure, we establish the ***second fact*** of the paper: that patent novelty predicts 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 that is based on the content of the patent. 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.[19]

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) dataset. We then divide this count with 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

[19]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/`.

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. We call patents with a smaller average probability more novel. Appendix B.10 provides similar analyses using alternative definitions of patent novelty such as new words and word age (Figure A.7), as well as chemical novelty (Table A.8). The correlations between these alternative measures and KPSS patent value and distance-to-science are quite similar to our main measure of novelty.

Novelty and patent value

Figure 2 shows that the novelty of a patent – measured by the average probability of word combinations – predicts the patent value and the likelihood that the value of a patent is in the tails of the distribution. Panel (a) shows that there is a positive relationship between novelty and patent value. The pattern suggests increasing returns to novelty, as the marginal gains from novelty increase as word combinations become more rare. Panel (b) indicates that a higher patent novelty is associated with an upward shift in the patent value distribution. We split all patents into those that have a below average probability of word combinations (i.e., higher novelty) and those that have an above average probability. Panel (b) shows that more novel patents (i.e., with rare word combinations) are less likely to be at the lower end of the value distribution and more likely to be at the upper end. The picture is reversed for patents that are less novel.

[Insert Figure 2 Here]

In Panel (c), we plot the relationship between patent novelty and patent value relative to (D=4)-patent values of the same technology and the same year. We also residualize the novelty measure (x-axis), such that x-axis values below (above) zero represent combinations of words that are less (more) common than the average word combinations within technology-year. Again, there is a clear positive relationship between novelty and science value.[20] In

[20]The regression version of this analysis is reported in Columns 1 and 2 of Table 2 (without and with

Panel (d), we show the distribution of patent values for patents with below and above average probability of word combinations relative to the probability of a (D=4)-patent in the same technology and year. Highly novel patents are again more likely to be in both tails of the value distribution, while patents with a lower novelty are in the middle of the value distribution, relative to its technology and year. Thus, as in the case of distance-to-science, novelty is associated with a value premium over and above the technology-related value component, but also with higher risk. Kline and Rosenberg (1986) captured the spirit of this relationship in writing, "newness is not, by itself, an economic advantage."

[Insert Table 2 Here]

4.3 Patent novelty and distance-to-science

As argued above, there are two complementary ways in which science can increase patent novelty. First, science can provide new insights that can be combined with older ideas. This view is akin to how Vannevar Bush described the relation between science and invention in his influential 1945 report *Science: The endless frontier*:

> *Basic science (...) creates the fund from which the practical applications of knowledge must be drawn. New products and new processes do not appear full-grown. They are founded on new principles and new conceptions, which in turn are painstakingly developed by research in the purest realms of science. Today, it is truer than ever that basic research is the pacemaker of technological progress.* (Bush, 1945).

This description is thought to reflect the realities in the large science-intensive corporate laboratories of the post-war period (Smith and Hounshell, 1985; Godin, 2006).

technology $\times$ year fixed effects). This specification differs from the one presented in Figure 2, because here we take the averages over all patents in a technology and year combination and not only over patents with a distance of D=4. The negative correlations in both specifications indicate that that patent value is decreasing in the likelihood of word combinations.

Second, science can guide the inventor to more fruitful combinations of known elements (Rosenberg *et al.*, 1990; Fleming and Sorenson, 2004). According to mathematician Henri Poincare, "the true work of the inventor consists in choosing among (...) combinations so as to eliminate the useless ones or rather to avoid the trouble of making them" (Weitzman, 1998).

Science can help in determining which combinations not to pursue by providing an understanding of why a particular combination might or might not work. For example, enormous amounts of energy and ingenuity were wasted by alchemists on attempts to transform lead into gold before science demonstrated that nothing short of an atomic reaction could achieve this end. Scientific knowledge also guided the development of the Haber-Bosch method of synthesizing ammonia. During the first trial runs, Carl Bosch struggled with the problem that the hydrogen proved to be corrosive for the high-pressure reactor chamber made of steel. Using basic chemistry, he deduced that the problem was due to the carbon contained in the steel walls of the chamber. His solution was to build a double wall reactor chamber with iron on the inside, which contains no carbon, and steel on the outside (Jeffreys, 2008).

In our final set of analyses, we explore the relationship between novel combinations of ideas and proximity to science. The strikingly similar patterns displayed in Figures 1 and 2 suggest that the novelty of patents and their distance-to-science are related. Consistent with this intuition, we establish as a ***third fact***: patents that are more science-intensive exhibit a higher patent novelty on average.

In Panels (a) and (b) of Figure 3, we show the novelty distributions for relatively more science-intensive patents (D=1, D=2, D=3) and for relatively less science-intensive patents (D=4, D=5, D>5, unconnected). As defined above, the lower the likelihood of a pairwise word combination in a patent is, the more novel the patent. Panel shows the novelty distribution for the raw data. In Panel (b) of Figure 3, we adjust for differences in technology and year. The novelty distribution for more science-intensive patents has its peak to the left and at a higher density than the novelty distribution for less science-intensive patents, both in the raw data and when controlling for technologies. This confirms that patents closer to science contain more novel word combinations; i.e., they are more novel (on average). Columns 3 and 4 of Table 2 display this relationship in regression form. In the averages (Column 3) and within technology-year (Column 4), we see that novelty is decreasing with distance-to-science—i.e., word combinations become more common, on average, as patents move further away from connections to science.

[Insert Figure 3 Here]

However, we also observe in Panel b) of Figure 3 that the relationship is non-linear and asymmetric. While patents more proximate to science are less likely to be in the "less novel" half of the distribution (right tail), they are also less likely to be in the far left tail of the novelty distribution. This asymmetry suggests that connection to science is associated with mid-range and above average novelty, while the extreme novel patents are more likely to be distant from science. Perhaps, those unconnected to science are less constrained in language (or imagination) than inventions tethered by the norms and formalities of science. In Appendix B.10, we show that these findings are robust to using the emergence of new words, the average age of words, of patent chemical novelty as alternative novelty indicators.

Finally, we combine our measures of distance-to-science and novelty to assess whether they contribute separately to patent value. The results are presented in Appendix B.11. We interact distance-to-science with indicators for below and above median novelty. Specifically, patents are below (above) median novelty if they have new word combinations that are below (above) median for their CPC (4 digit) class and year. We then generate graphs equivalent to Figure 1 for the below median and above median novelty groups. The results in Figure A.8 in Appendix B.11 reveal very similar patterns for both groups. Both with and without adjusting for tech class × year, we see that relative patent value is increasing in proximity to science. The same is true regardless of whether we look at % differences or average patent dollar values.

While the relationship between distance-to-science and relative patent value is quite similar for both novelty groups, Appendix Figure A.8 in Appendix B.11 shows that patent values are consistently higher (shifted upward) for the high novelty sub-groups. Thus, novelty and science both appear to contribute to private value. Since the two characteristics are at least partially co-determined,[21] we cannot quantify their relative influence on patent values. However, since both novelty subgroups have patent values increasing in their proximity to science, novelty is clearly not the sole mechanism behind our main results. Rather, we interpret these patterns as evidence that using science as a tool to explore and express new ideas is a path associated with both greater novelty and value capture.

5 Discussion: Heterogeneity in the Science Premium

Our main results suggest that firms reap greater benefits from more novel and science-driven invention. So why don't all firms simply choose to invent "on the shoulders" of science? In this section we discuss possible explanations as to why the science premium persists given that firms *choose* their preferred mode of R&D. Overall, we find that the greater share of value from inventing with (and without) science is captured by the subset of firms that consistently use science in patenting. However, we still find a science premium after accounting for firm fixed effects or scientific-intensity—implying that risk aversion and/or convex marginal costs of inventing with science limit that mode of innovation.

Imperfect information. We first acknowledge the possibility that managers might not be aware of the (average) additional value of science-proximate patenting. Given the narratives around patenting in the information technology industry (see Section 2.1), we would not be surprised if executives often underestimate the value of inventing based on recent science. That said, firms might be well aware of the average value of relying on science in inventing,

[21] In addition to the patterns found in Figure 3, Figure A.8 in Appendix B.11 demonstrates the strong correlation between novelty and proximity to science, as the two panels that adjust for technology $\times$ year fixed effects (Panels (b) and (d)) exhibit much larger differences between the low vs. high novelty subgroups than we see in the unadjusted raw differences (Panels (a) and (b)).

yet still choose other strategies for performing R&D.

Risk Aversion. As emphasized in the results sections, the average premium for building on science also comes with increased tail-risk (i.e., more worthless patents). The additional expected value of science-proximate patenting might not appeal to firms that are risk averse due to a variety of factors (Cyert *et al.*, 1963; Argote and Greve, 2007; Hall and Lerner, 2010; Eggers, 2012; Krieger *et al.*, 2021).[22] Based on our firm fixed effects regressions (Appendix Figure A.6) and analyses described below, we believe that risk-aversion plays a role, since the science premium persists within patenting firms that choose different R&D strategies.

Value cliff in the supply of ideas. Marginal benefits to building on the shoulders of science may be non-linear and drop off dramatically beyond a certain level. In this scenario, the returns to exploring with science "fall off a cliff" after the initial (obvious) connections from science to invention are "picked over." Observed average returns to science-based invention might appear higher than pure "tinkering" invention, even though marginal projects have similar ex-ante values. Without a series of natural experiments that exogenously increase different types of R&D (e.g., D=1, D=2, D=3, etc.), we cannot rule this story out entirely. However, this story would run counter to much observed behavior around commercializing science. U.S.-based universities have displayed growing enthusiasm for commercialization science, as evidenced by their secular increase in patenting and licensing of science-driven intellectual property,[23] and are doing more to encourage commercialization through setting up their own accelerators and entrepreneurship competitions. Similarly, venture capitalists have exhibited a renewed interest in companies that aim to commercialize "tough tech," which includes areas like energy systems, life sciences, and defense technologies that are

[22] The prior literature suggests that firms might vary greatly in their level of risk-aversion. Related, a recent set of papers has explored how firms differ in their level of short-termism (Sampson and Shi, 2020; Gormsen and Huber, 2022), suggesting that firms may have different tolerance for uncertain R&D explorations. While this study does not involve the data required to effectively separate firms by their implied discount rates or level of risk aversion, we believe that future research may be able to test how changes in these firm-level factors shifts the direction of innovative output.

[23] See the Association of University Technology Managers (AUTM) reports on trends in university patenting, spinoff company formation and licensing (e.g., `https://autm.net/AUTM/media/SurveyReportsPDF/FY20-US-Licensing-Survey-FNL.pdf`).

more likely to rely on academic discoveries for their core intellectual property.[24]

Thus, despite the contracting frictions in commercializing frontier science (e.g., license negotiations, asymmetric information) and the cost disadvantages relative to developing information technologies (Ewens *et al.*, 2018), industry trends suggest that both scientific institutions and investors believe that the marginal academic idea is worth the effort of expanding their scope of activity. In short, we do not expect the "value cliff" story to drive the science premium.

R&D Cost Heterogeneity and Path Dependency. 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 (absorptive capacity without in-house research) is costly and may not yield positive net present value for all firms. Reasons for R&D cost heterogeneity in reliance on science 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), imprinting (Levinthal, 2003), and absorptive capacity (Cohen and Levinthal, 1990).

We investigate this possibility by looking at how the value of science in patenting differs across firms. Heterogeneity in the costs of using or access to science for invention may lead firms to sort into different R&D strategies. As discussed in Section 4, firm-level sorting based on overall invention quality/capabilities could show up in the aggregate data as a science premium. However, we find that adding in firm fixed effects, in addition to technology × year fixed effects, does not eradicate the science premium (see Appendix Table A.6 in Appendix B.3). The magnitude of the science premium (in dollar terms) is smaller due to firm fixed effects, which implies that firm differences in inventing quality are indeed associated with

[24] Trends in "tough tech" financing are detailed in a report by The Engine, an independent venture capital firm seeded by MIT and Harvard University: `https://www.engine.xyz/wp-content/themes/the-engine-wp-theme/templates/pitchbook-assets/The-Engine-Pitchbook.pdf`.

their propensity to use science.

The logical next question is what drives this variation in firm quality. Difficult-to-observe factors like general managerial skill and the extraordinary creativity of individual contributors might drive persistent firm differences. Other sources of heterogeneity include hard (though not impossible) to change firm characteristics like firm location, investor pressures, and the existence or timing of competitor breakthroughs. Where patenting firms have significantly more agency is in their choice of R&D strategy. A firm's level of reliance on science has important (high commitment) implications for how it explores and develops new technology—i.e., sourcing ideas, modes of experimentation, types of expertise, project selection criteria, novelty, etc. Thus, we expect that a firm's science intensity would affect the overall benefits of patenting, as well as the premium from relying on science in a given invention.

In order to reasonably characterize firm differences in reliance on science, we limit the sample to firms with at least 10 patents. Next, we split the sample by firms above and below the median in the percentage of their (cumulative) patents to-date that directly cite scientific articles (D=1).[25] Panel (a) of Figure A.9 in Appendix B.12 shows the distribution of distance-to-science patents for the two groups. The sample split indeed appears to identify different level of reliance on science on the firm level: The above median firms have a larger share of (D=1) and (D=2) patents than the below median firms. The below median firms have a larger share for (D>3) patents than the above median firms.

To see whether above and below median firms differ in patent values, we then run our main patent value regression specifications with interaction terms for distance-to-science $\times$ above/below median, accounting for technology class and time fixed effects. We plot the results in Panel (b).

Three findings stick out. First, firms above the median produce higher value patents

[25] Intuitively, both groups are firms with a non-trivial inventive footprint, but representing different "modes" of invention.

generally, across every level of distance-to-science (see Panel (b) of Appendix Figure A.9).[26] The gap in average patent value is consistently around \$4 million. This large value difference suggests that that firms exploring with the aid of science produce more valuable inventions generally—including the majority of their inventions that only use science indirectly (D>1). Though we cannot infer the direction of causality in that relationship, we suspect that firms producing highly valuable patents without science would not have a strong incentive to adopt more science intensive modes of invention.[27] Second, we see that the marginal value of building on science is greater for firms above the median in their typical reliance on science. Panel B shows that firms above the median gain a roughly \$2 million of average incremental value by producing a (D=1) patent instead of a (D=4) patent, while that difference is only about \$1.6 million (with a wider confidence interval, nearly straddling \$0) for firms below the median. Third, we still see an overall downward slope for both groups—suggesting that other factors like risk aversion and/or the marginal cost of science-based invention (even for science-intensive firms) also contribute to the science premium.

Another dimension of a firm's "science intensity" is whether the firm performs in-house research. 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). In Table A.10 in Appendix B.12, we investigate how building on a firm's own (*in-house*) publications correlates with patent dollar value. We first show that (D=1) patents that cite at least one in-house publication (with authors linked to the same firm) are significantly more valuable than (D=1) patents that only cite external research (Columns 1 and 2). On average, citing an in-house publication is associated with a patent value more than 8 times larger than other (D=1) patents. We further show that firms that are in the above median group for reliance on science generate higher value from patents that

[26] We use the terminology "produce higher value patents," though we cannot distinguish between value creation and value capture in this sample split. It is possible that above the median firms either produce higher quality inventions or are simply better at extracting value from patents.

[27] Indeed, in supplemental analyses (not shown) we find that firms—both above and below the median in reliance on science—tend to have a smaller percentage of (D=1) patents as they mature and patent more over time.

cite in-house research than firms that less frequently rely on science (Columns 3–5). Thus, building on in-house science appears highly valuable, especially for high reliance-on-science firms.

In summary, less science-intensive firms tend to have lower value patents, and appear to gain a smaller benefit (if any at all) from building on science—both for external and in-house science. Patenting firms without many prior 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 and risky. While science-based inventing may seem attractive in the context of patent value, the long-term investments required to consistently build on science are considerable in many industries. Furthermore, the persistence of a science premium suggests that risk aversion and/or convex marginal cost in using science directs firms to lower payoff modes of invention.

6 Conclusion

Our study shows that building more directly on science is associated with more novel inventions and capturing greater private value from those inventions. 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, 1990; Henderson and Cockburn, 1994; Stern, 2004).

Along with the increased expected rewards from science, our results show that building on science is a relatively risky approach to corporate innovation. We find that patents closer to science and relatively novel patents are both more likely to end up in the tails of the patent value distribution.

In combination, our results suggest that science helps firms push the technological frontier by building on more disparate ideas to introduce and combine more novel technologies—many of which fall flat commercially, while others propel their firm's growth. 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.

Akcigit, U. and Kerr, W. R. (2018). Growth through heterogeneous innovations. *Journal of Political Economy*, **126** (4), 1374–1443.

Argote, L. and Greve, H. R. (2007). A behavioral theory of the firm 40 years and counting: Introduction and impact. *Organization Science*, **18** (3), 337–349.

Arora, A., Belenzon, S. and Dionisi, B. (2021a). *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. (2017). *Back to basics: why do firms invest in research?* Tech. rep., National Bureau of Economic Research.

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

—, — and — (2021b). Knowledge spillovers and corporate investment in scientific research. *American Economic Review*, **111** (3), 871–98.

—, 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., Furman, J. L., Krieger, J. L. and Murray, F. (2015). Retractions. *Review of Economics and Statistics*, **97** (5), 1118–1136.

—, 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.

—, — and Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, **42** (3), 527–554.

Backman, T. W., Cao, Y. and Girke, T. (2011). Chemmine tools: an online service for analyzing and clustering small molecules. *Nucleic acids research*, **39** (suppl_2), W486–W491.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, **483** (7391), 531–533.

— and Ioannidis, J. P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, **116** (1), 116–126.

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

Bikard, M. (2018). Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science*, **29** (5), 818–836.

— 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., Jones, C. I., Van Reenen, J. and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, **110** (4), 1104–44.

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

Brunnermeier, M. and Nagel, S. (2004). Hedge funds and the technology bubble. *The journal of Finance*, **59** (5), 2013–2040.

Bush, V. (1945). *Science, the endless frontier: A report to the President*. US Govt. print. off.

Butler, D. (2008). Translational research: crossing the valley of death. *Nature News*, **453** (7197), 840–842.

Carley, M., Hedge, D. and Marco, A. (2015). What is the probability of receiving a us patent. *Yale JL & Tech.*, **17**, 203.

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., Henderson, R. M. 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. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, **35** (1), 128–152.

Cyert, R. M., March, J. G. *et al.* (1963). *A behavioral theory of the firm*, vol. 2. Englewood Cliffs, NJ.

Eggers, J. P. (2012). Falling flat: Failed technologies and investment under uncertainty. *Administrative Science Quarterly*, **57** (1), 47–80.

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.

Freedman, L. P., Cockburn, I. M. and Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS biology*, **13** (6), e1002165.

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.

Godin, B. (2006). The linear model of innovation: The historical construction of an analytical framework. *Science, Technology, & Human Values*, **31** (6), 639–667.

Goozner, M. (2005). *The $800 million pill: The truth behind the cost of new drugs*. Univ of California press.

Gormsen, N. J. and Huber, K. (2022). Discount rates: Measurement and implications for investment.

Greenwood, R. and Nagel, S. (2009). Inexperienced investors and bubbles. *Journal of Financial Economics*, **93** (2), 239–258.

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools. *NBER Working Paper No. 8498*.

— and Lerner, J. (2010). The financing of r&d and innovation. In B. H. Hall and N. Rosenberg (eds.), *Handbook of The Economics of Innovation, Vol. 1*, *Handbook of the Economics of Innovation*, vol. 1, North-Holland, pp. 609–639.

Harris, G. (2011). Federal research center will help develop medicines. *New York Times*, (January 22).

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.

Jeffreys, D. (2008). *Hell's cartel: IG Farben and the making of Hitler's war machine*. Macmillan.

Jones, B. F. (2009). The burden of knowledge and the death of the renaissance man: Is innovation getting harder? *The Review of Economic Studies*, **76** (1), 283–317.

Kelly, B., Papanikolaou, D., Seru, A. and Taddy, M. (2018). *Measuring technological innovation over the long run*. Tech. rep., Working Paper.

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., Li, D. and Papanikolaou, D. (2021). Missing Novelty in Drug Development*. *The Review of Financial Studies*, hhab024.

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.

Kuhn, J. M. and Thompson, N. (2017). How to measure and draw causal inferences with patent scope.

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.

MANSO, G. (2011). Motivating innovation. *The Journal of Finance*, **66** (5), 1823–1860.

MARCH, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, **2** (1), 71–87.

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.

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.

OSHEROVICH, L. (2011). Hedging against academic risk. *Science-Business eXchange*, **4** (15), 416–416.

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*.

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.

ROSENBERG, N. *et al.* (1990). Why do firms do basic research (with their own money)? *Research Policy*, **19** (2), 165–174.

SAMPSON, R. C. and SHI, Y. (2020). Are us firms becoming more short-term oriented? evidence of shifting firm time horizons from implied discount rates, 1980–2013. *Strategic Management Journal*.

SCHMOCH, U. (2008). Concept of a technology classification for country comparisons. *Final report to the world intellectual property organisation (wipo), WIPO*.

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.

SMITH, J. K. and HOUNSHELL, D. A. (1985). Wallace h. carothers and fundamental research at du pont. *Science*, **229** (4712), 436–442.

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

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.

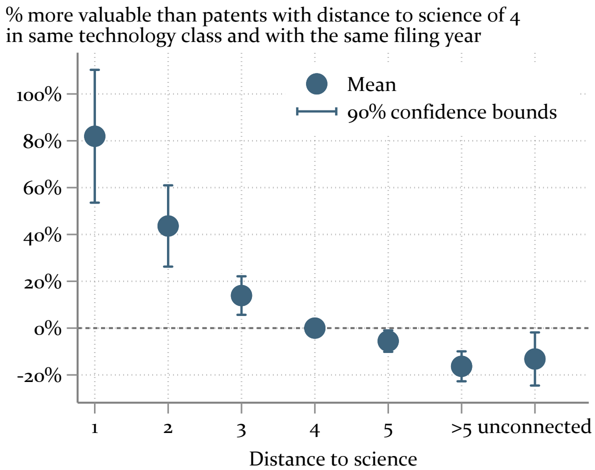
Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L. and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 990–998.

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

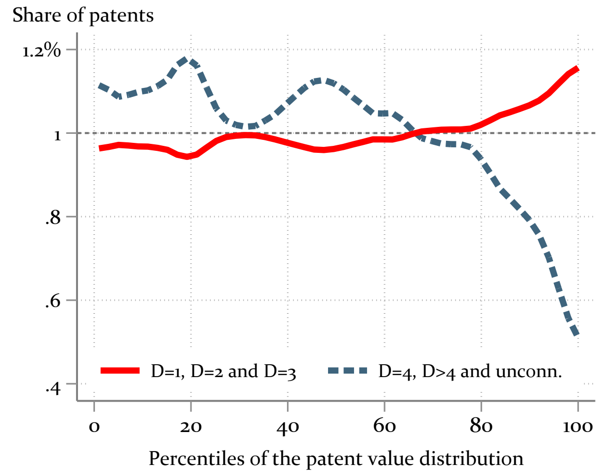
Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, pp. 331–360.

Wuchty, S., Jones, B. F. and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, **316** (5827), 1036–1039.

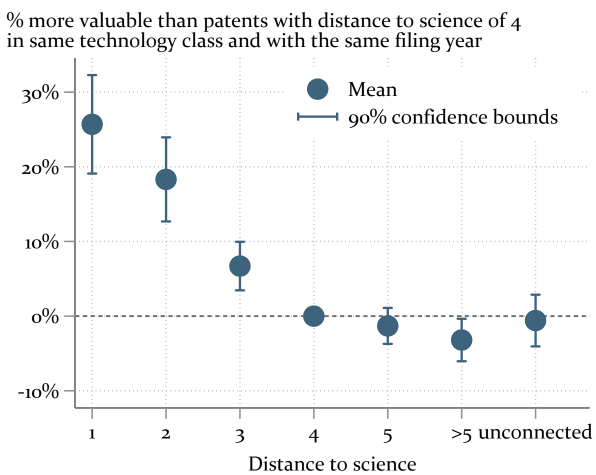
Tables & Figures



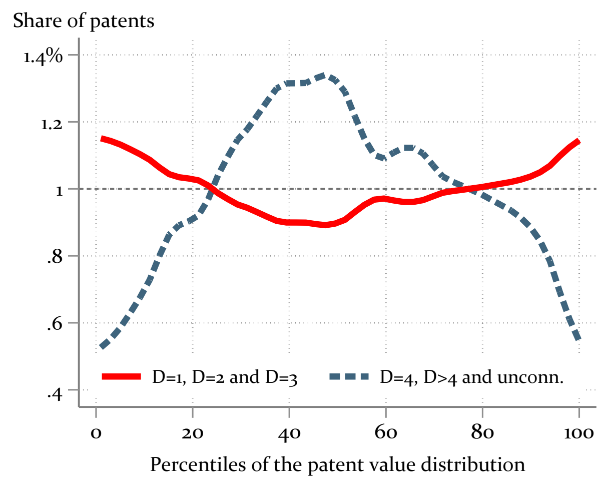
(a) Averages



(b) Distribution



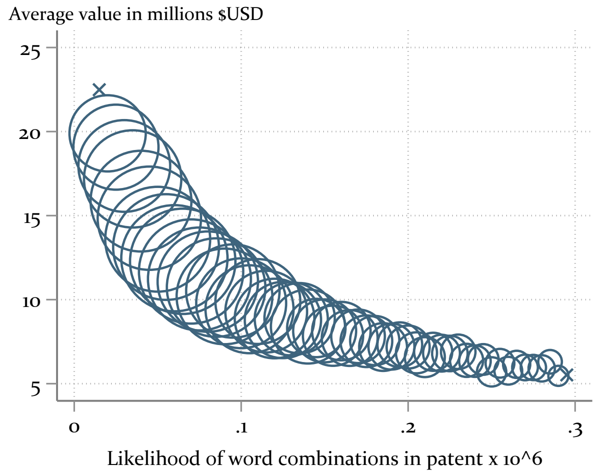
(c) Averages accounting for technology and year



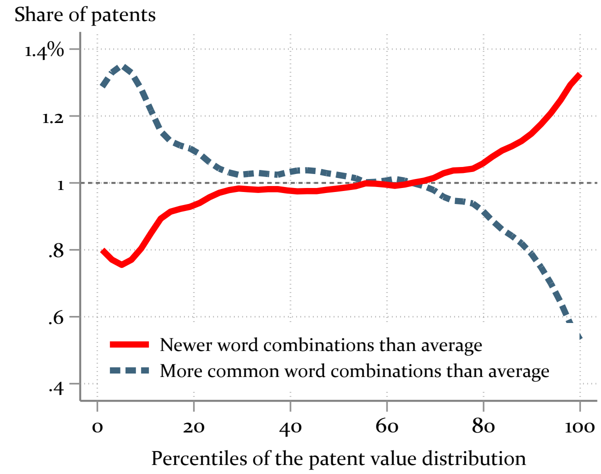
(d) Distribution accounting for technology and year

Figure 1: Distance-to-science, patent value and risk.

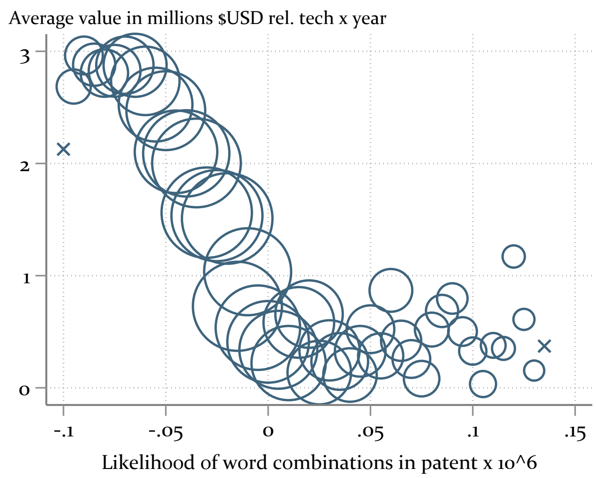
Panel (a) shows the average increase in patent value for all distances to science relative to patents with a distance of four (D=4) in percent (%). The values of U.S. patents are from Kogan *et al.* (2017). The distance-to-science of U.S. patents is calculated using data from Marx and Fuegi (2020) and 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 1. Panel (b) shows the distribution of patent values across the percentiles of the value distribution of all patents for more science-intensive patents (D=1, D=2 or D=3; solid red line) and less science-intensive patents (D>3 or unconnected; dashed blue line). The horizontal line at 1% shows the distribution of all patents across the percentiles of the value distribution. In Panel (c), we residualize the patent value by the average value of a patent with the same (four-digit) CPC technology class and filing year and a distance of four, then display relative (%) values indexed to D=4. The values correspond to the coefficients in Table 1, Column 2. In Panel (d), we show the distribution of the patent values normalized by technology and year.



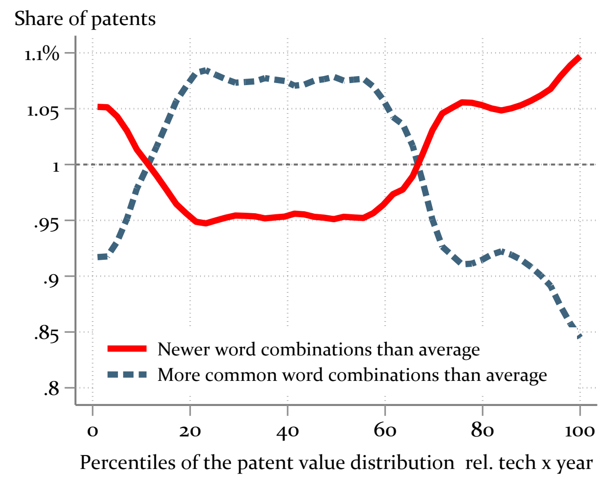
(a) Raw data



(b) Distribution



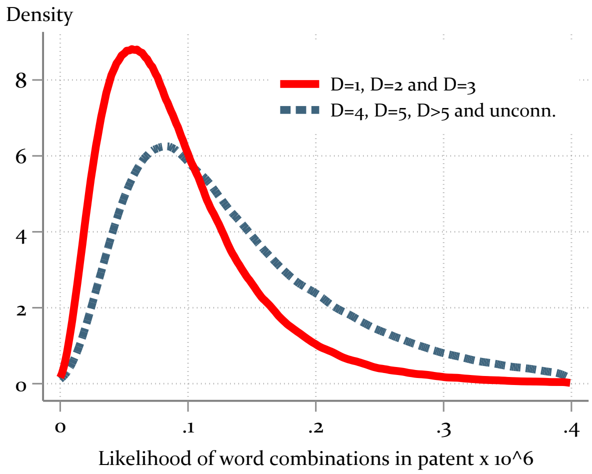
(c) Accounting for technology and year



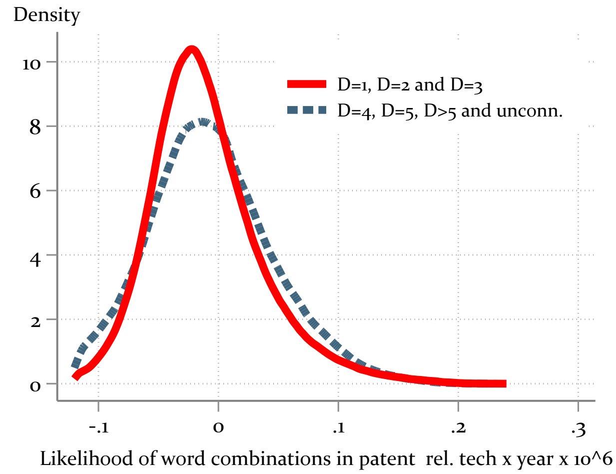
(d) Distribution accounting for technology and year

Figure 2: Patent novelty, patent value and risk.

Panel (a) shows the average patent value for every likelihood of pairwise combinations of words that occur in a particular patent as an indicator for patent novelty. Smaller probabilities are interpreted as higher novelty. The winsorized values are marked with "X." The size of the bubbles represents the number of patents underlying each point. Panel (b) shows the distribution of patent values across the percentiles of the value distribution of all patents for patents with below average pairwise word combination probability (solid red line) and for above average pairwise word combination probability (dashed blue line). In Panel (c), we plot the average residualized patent value by residualized pairwise word combination probability. We residualize the value and the word combination probability for the interaction of (four-digit) CPC technology class and filing year. Panel (d) shows the distribution of residualized patent values by distance of patents to science across the percentiles of the value distribution of all patents for patents with below average pairwise word combination probability in a technology and year (solid red line) and for above average pairwise word combination probability in a technology and year (dashed blue line).



(a) Novelty distribution by science intensity



(b) Accounting for technology and year

Figure 3: Patent novelty and distance-to-science.

Panel (a) shows the kernel density plot of the average likelihood of pairwise combinations of words that occur in a particular patent for more science-intensive patents (D=1, D=2 or D=3; red line) and for less science-intensive patents (D>3 or unconnected; dashed blue line). Smaller probabilities are interpreted as a higher novelty. In Panel (b), we residualize the patent value and the likelihood of word combinations by the average value of a patent with the same technology class and filing year and a distance of four.

	(1)	(2)	(3)	(4)	(5)	(6)
	Patent Value					
Outcome:	Dollar Percent	Dollar Percent	Citations Percent	Patent Scope	Prob(Litigation) x 1000	
Distance 1	0.82	0.26	0.99	0.12	10.36	4.40
	(0.17)	(0.04)	(0.09)	(0.03)	(1.44)	(0.62)
Distance 2	0.44	0.18	0.36	0.10	6.77	1.50
	(0.11)	(0.03)	(0.04)	(0.02)	(0.60)	(0.42)
Distance 3	0.14	0.07	0.09	0.07	4.93	0.16
	(0.05)	(0.02)	(0.02)	(0.01)	(0.38)	(0.34)
Distance 4					4.39	
					(0.41)	
Distance 5	-0.06	-0.01	-0.06	-0.03	3.88	-0.52
	(0.03)	(0.01)	(0.01)	(0.02)	(0.40)	(0.43)
Distance >5	-0.16	-0.03	-0.14	-0.05	3.43	-0.73
	(0.04)	(0.02)	(0.02)	(0.02)	(0.30)	(0.40)
Unconnected	-0.13	-0.01	-0.13	-0.04	2.49	-0.57
	(0.07)	(0.02)	(0.03)	(0.08)	(0.26)	(0.41)
Tech x Year FE	No	Yes	Yes	Yes	Yes	
Mean Dep.	11.56	11.56	29.88	0.12	6.40	6.40
Obs.	1135757	1135757	1135757	231474	1135757	1135757

Table 1: Patent Value Measures

Table 1 reports regression results on how various patent value outcomes with the independent variable of distance-to-science. The calculation of distance-to-science is based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. Columns 1–3 are generated by ordinary least squares (OLS) specifications, and we report the coefficients in terms of their percentage increases relative to D=4. For example, the coefficient for the first value in Column 1 may be interpreted as an 82% increase relative to (D=4). In Column 1, the outcome is (adjusted) patent values from Kogan *et al.* (2017). In Column 2, we control for (four-digit) CPC technology class × filing year fixed effects such that coefficients represent relative differences within tech-year. In Column 3, the outcome variable is the count of citing families. In Column 4, we use the within art unit patent scope index provided by Kuhn and Thompson (2017) as the outcome variable in an OLS regression. Columns 5 and 6 are both OLS specifications where the outcome variable is 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). For all models, the standard errors are clustered on the CPC technology class level.

	(1)	(2)	(3)	(4)
	Patent value		Patent novelty	
Outcome:	Dollar	Dollar	Probability of word combination	Probability of word combination
Distance 1			7.91	-1.69
			(0.36)	(0.18)
Distance 2			9.37	-0.76
			(0.28)	(0.11)
Distance 3			10.45	-0.23
			(0.17)	(0.07)
Distance 4			11.90	
			(0.16)	
Distance 5			13.95	0.30
			(0.18)	(0.07)
Distance >5			16.58	0.79
			(0.20)	(0.13)
Unconnected			16.25	0.39
			(0.50)	(0.11)
Probability	-43.39	-10.20		
of word combinations	(3.84)	(1.47)		
Tech x Year FE	No	Yes	No	Yes
Mean Dep.	11.56	11.56	10.45	10.45
Obs.	1135669	1135669	1135669	1135669

Table 2: Patent value and text novelty

This table shows OLS regression results on the relation of patent value, patent novelty and distance-to-science. In Columns 1 and 2, the outcome variable is the adjusted Kogan *et al.* (2017) patent values.The independent variable is the focal patent's probability of word combinations based on the data from Arts *et al.* (2018). In Columns 3 and 4, the outcome variable is the patent's probability of word combinations, and the indepenent variables are the distance-to-science measure based on Ahmadpoor and Jones (2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Columns 2, 3, 4 we additionally control for filing year × (four-digit) CPC technology class fixed effects. The standard errors are clustered on the CPC technology class level.

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.

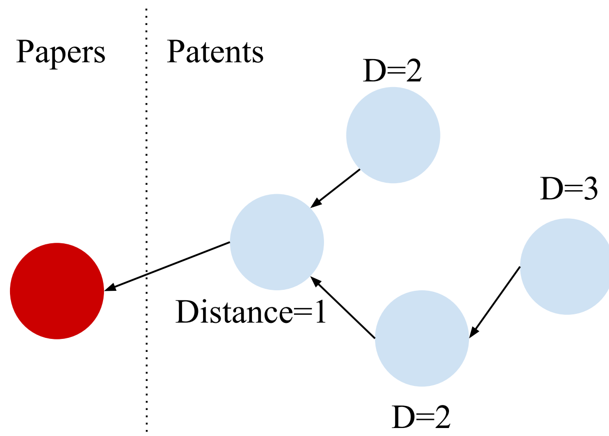


Figure A.1: Distance-to-science

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.[28] 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).[29] 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

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

[29] 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.

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

As an alternative measure of patent novelty, we take the subset of patents which describe chemical compounds and calculate their compound novelty. For any given patent, we calculate the pairwise similarity between each one of its described chemical structures' similarity to all compounds represented in prior patents in the same 3-digit CPC class. One minus the maximum of those pairwise similiarities is a patent's "chemical patent novelty score." To do so, we use the crosswalk of patents-to-compounds from SureChEMBL, which extracts the standardized (SMILES code) chemical structures represented in patents. To calculate the pairwise similarity scores, we use the ChemmineR package (Backman *et al.*, 2011). We describe these analyses in Appendix B.10.

Sources:

http://chembl.blogspot.com/2015/03/the-surechembl-map-file-is-out.html

https://chemminetools.ucr.edu/

https://www.bioconductor.org/packages/devel/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html

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 scope** is from Kuhn and Thompson (Kuhn and Thompson, 2017). Specifically, we use the z-score within art unit for our results.

  Source: *http://jeffreymkuhn.com/index.php/data/*

- **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 Patent Characteristics and Distance-to-Science

The following tables describe characteristics of patents in our data set. All the variables relate to characteristics at the time of patent issuance, rather than subsequent outcomes like stock market reaction, forward citations, and litigation.

20.7% of patents in our data directly cite scientific publications. Among those (D=1) patents, they average 6.1 citations to scientific journal articles, and the average age of those journal articles is 8.04 years prior to the patent's publication. Appendix Tables A.1, A.2 and A.3 present descriptive statistics for patents in our sample, by distance-to-science. We find that more recent patents are more likely to cite science. In our analysis data, the average patent issuance year of 2001.8 for (D=1) patents, and steadily descends with distance-to-science all the way to 1989.2 for unconnected patents. Patents closer to science also tend to have larger teams, more patents in their patent "family,"[30] take longer for the patent office to process (from application to publication), and have fewer words in their claims. Appendix Table A.3 shows that these relationships hold even after controlling for patent publication year and CPC (four digit) technology class.

Appendix Table A.2 and Appendix Table A.3 further describe the profile of backward citations, by the focal patent's distance-to-science. We see that patents closer to science have more backwards citations—indicating that proximity to science involves building off a larger base of prior art. We find that science patents also build off a wider foundation of extant inventions, as the share of same-technology (CPC class) citations is increasing in distance-to-science. Patents that are most distant to science also cite older prior art. The average age of a cited patent is between 7.38–8.33 years old for (D=1,2,3) patents, while the range grows to 10.9–22.0 year old for (D>3). Finally, we use the patent-to-patent text similarity measure developed by Kuhn and Thompson (2017) to assess how distance-to-science correlates with similarity between focal patents and their (backwards) cited patents. In both the

[30] A patent family is the set of patents, applied for across different countries, that aim to protect the same invention by the same inventor(s).

raw averages and the year- and technology-adjusted regressions, we see that the maximum cited patent similarity is decreasing in distance from science while the average cited patent similarity is increasing in distance-to-science. In other words, inventions closer to science cite patents that are more varied in their word similarity to the focal patent. These patterns fit the view that science, as an exploration tool, helps navigate broader patent space. The common language of shared scientific methods enables specific comparisons (i.e., clear performance improvements) to highly similar innovations, while also uncovering new connections and sparking recombination across inventions with disparate product applications.

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.[31] 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.

Just as the science-based McKesson restocking system patent shared the same technology class as the scientifically distant Coca Cola vending machine patent (see Section 2 in the main text of the paper), 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

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

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	234946	2001.8	6.23	2.89	1100.2	110.6	166.3
Distance 2	386061	2001.9	4.13	2.64	1040.9	113.4	164.3
Distance 3	240145	2000.9	3.93	2.54	923.2	122.8	167.1
Distance 4	105705	1997.4	3.77	2.40	795.8	137.6	175.5
Distance 5	60579	1994.0	3.54	2.23	728.4	141.4	181.6
Distance >5	67355	1991.5	3.39	2.07	686.5	146.0	181.8
Unconbected	40966	1989.2	3.51	2.05	701.5	144.0	171.1
Total	1135757	1999.7	4.39	2.57	955.5	118.9	167.2

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

NOTES: Table A.1 presents 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	8.00	5.27	0.55	0.31
Distance 2	11.8	0.14	0.56	7.38	5.18	0.54	0.32
Distance 3	9.77	0.17	0.56	8.33	6.10	0.52	0.34
Distance 4	9.93	0.18	0.56	10.9	8.11	0.52	0.35
Distance 5	10.7	0.16	0.57	12.9	9.44	0.50	0.35
Distance >5	10.2	0.14	0.60	15.0	10.9	0.48	0.36
Unconnected	8.58	0.13	0.59	22.0	12.7	0.44	0.37
Total	11.2	0.15	0.56	9.18	6.38	0.53	0.33

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

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

Outcome:	(1) Nb. Inventors	(2) Processing Days	(3) Claim Words	(4) Nb. Backwards Cites	(5) Share Same-Tech	(6) Avg. Age Cites	(7) Max. Sim. Cited	(8) Avg. Sim. Cited
Distance 1	0.329	101.1	-13.54	6.47	-0.055	-0.59	0.054	-0.037
	(0.00706)	(10.32)	(0.878)	(0.085)	(0.0013)	(0.024)	(0.00078)	(0.00055)
Distance 2	0.184	58.03	-12.82	3.43	-0.042	-1.22	0.037	-0.025
	(0.00647)	(9.445)	(0.800)	(0.078)	(0.0012)	(0.022)	(0.00071)	(0.00050)
Distance 3	0.0796	-4.608	-9.634	1.62	-0.024	-1.12	0.0076	-0.010
	(0.00657)	(9.598)	(0.787)	(0.079)	(0.0012)	(0.022)	(0.00072)	(0.00051)
Distance 5	-0.0220	10.47	5.677	-1.51	0.014	1.38	-0.011	1.8e-06
	(0.00890)	(13.00)	(1.343)	(0.11)	(0.0016)	(0.030)	(0.00097)	(0.00069)
Distance >5	-0.0494	16.15	6.206	-3.74	0.050	2.87	-0.023	0.0099
	(0.00887)	(12.96)	(1.746)	(0.11)	(0.0016)	(0.030)	(0.00097)	(0.00069)
Unconnected	-0.0617	33.51	1.343	-4.17	0.055	9.82	-0.052	0.020
	(0.0106)	(15.47)	(2.352)	(0.13)	(0.0022)	(0.040)	(0.0014)	(0.00097)
Tech x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,135,747	1,135,725	232,049	1,135,747	1,114,025	1,114,025	1,108,568	1,108,568

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). 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 patent issuance year 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

Columns 1–3 of Appendix Table A.4 are analogous to those Table 1, but with coefficients that correspond to level differences rather than percentage differences. This table generates the patent value magnitudes that we report in Section 4.1. We report the same estimates as Table 1 for Columns 4–6, which show the patent scope and litigation outcomes as alternative measures of value. Appendix Table A.4 also has two additional Columns (7 and 8) which report the text similarity between the focal patent and it's most similar scientific journal articles as described above in Appendix A.1 and B.5.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Patent Value						Text similarity	
Outcome:	Dollar	Dollar	Cita-tions	Patent Scope	Probability of Litigation x 1000		Text sim.	Dollar
Distance 1	15.82	2.64	22.17	0.12	10.36	4.40	7.13	
	(1.56)	(0.36)	(1.47)	(0.03)	(1.44)	(0.62)	(0.31)	
Distance 2	12.49	1.88	8.07	0.10	6.77	1.50	4.76	
	(1.08)	(0.31)	(0.63)	(0.02)	(0.60)	(0.41)	(0.26)	
Distance 3	9.90	0.69	2.02	0.07	4.93	0.16	2.46	
	(0.77)	(0.19)	(0.38)	(0.01)	(0.38)	(0.34)	(0.20)	
Distance 4	8.69				4.39			
	(0.62)				(0.41)			
Distance 5	8.21	-0.13	-1.41	-0.03	3.88	-0.52	0.10	
	(0.46)	(0.15)	(0.27)	(0.02)	(0.40)	(0.43)	(0.09)	
Distance >5	7.27	-0.33	-3.10	-0.05	3.43	-0.73	-0.46	
	(0.33)	(0.18)	(0.36)	(0.02)	(0.30)	(0.40)	(0.09)	
Unconnected	7.55	-0.06	-2.94	-0.04	2.49	-0.57		
	(0.64)	(0.22)	(0.69)	(0.08)	(0.26)	(0.41)		
Text similarity								3.04
								(0.87)
Tech x Year FE	No	Yes	Yes	Yes	No	Yes		
Mean Dep.	11.56	11.56	29.88	0.12	6.40	6.40	9.94	11.56
Obs.	1135757	1134139	1134139	231474	1135757	1134139	1134626	1134626

Table A.4: Patent value (in levels) and text similarity, by distance-to-science

This table shows OLS regression results. In Column 1, we use the patent values of Kogan et al. (Kogan *et al.*, 2017) as outcome variable. The independent variable is the distance-to-science measured by citation links. The calculation of distance-to-science is based on the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). Unconnected patents are patents for which we could not find a citation link to any scientific article. In Column 2, we control for filing year x (four-digit) CPC technology class fixed effects. We use the distance-to-science of four as a baseline. In Column 3, we use the number of citing patent families as outcome variable. This data is from Patstat. In Column 4, we use the within art unit patent scope index provided by Kuhn and Thompson (2017) as the outcome variable in an OLS regression. Columns 5 and 6 are both OLS specifications where the outcome variable is 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). In Column 7, we use the text similarity between the abstract of the scientific article cited and the patent abstract as outcome variable. If there is more than one cited article, we take the maximum of the similarity per patent. In Column 8, we use text similarity as independent variable and KPSS dollar value as the outcome. The standard errors are clustered on the CPC technology class level.

B.2 Different assumptions for KPSS value estimates

This section shows the sensitivity of our patent value measures to different assumptions about probability of patent grant. Our primary measure of patents' private value is from Kogan *et al.* (2017) [KPSS]. While we primarily use the KPSS values as a measure of *relative* value across distance from science and novelty groups (e.g., see Figure 1 and Table 1), we also report dollar magnitudes adjusted for conservative estimates of patent grant rates (see Section 4.1). An important assumption in the KPSS patent valuation method is choosing the value for a patent's ex-ante probability of success, π. Using a higher probability of patent success mechanically increases all patent values since the method scales all patent values by $(1-\pi)^{-1}$. The intuition is that a big positive stock market response is even more indicative of a valuable patent is the market was already expecting that the patent had a decent chance of approval.

Appendix Table A.5 shows how various adjustments to the KPSS values change the dollar magnitudes of average patent values by distance-to-science. Columns 1 and 2 of Appendix Table A.5 report the average patent values, by distance-to-science, based on the adjusted values used throughout the main body of the paper, without and with technology $\times$ year fixed effects. As described in Section 4.1, these "new baseline" estimates deflate all KPSS values by 12% to match the most conservative end of the spectrum for probability of patent grant. In other words, conditional on the market being aware of a patent application, we assume the greatest average market "surprise" for patent publication events. For contrast, Columns 4 and 5 show the same regressiosn using the undadjusted ("old baseline") KPSS values.

Columns 5 and 6 use a more extreme adjustment: assuming that the market had no information about a patent application prior to patent publication. This alternative shows a 40%-50% drop in average patent values with this approach [Column (1) vs. Column (5)], but the general relationship between distance from science and relative value holds. We think this additional analysis is useful for offering a sort of lower-bound, however reality probably falls in between the "no information benchmark" and the original KPSS patent values. Prior to the American Inventor's Protection Act (AIPA), which was enacted in November 2002, patent applications were not necessarily publicly available prior to grants. Nor was the market totally ignorant of firms' inventions, since firms might still publicly disclose inventions that were "patent pending" through a variety of communication channels, or have publicly disclosed patent documents at non-US patent agencies. Thus, the "no information benchmark" values serve as an extreme lower bound.

	(1)	(2)	(3)	(4)	(5)	(6)
	New baseline Conservative acceptance rate		Old baseline Original KPSS		No information benchmark	
Outcome:	Dollar	Dollar	Dollar	Dollar	Dollar	Dollar
Distance 1	15.82	2.64	17.72	2.95	9.73	1.90
	(1.56)	(0.36)	(1.75)	(0.40)	(0.97)	(0.21)
Distance 2	12.49	1.88	13.98	2.10	7.47	1.31
	(1.08)	(0.31)	(1.21)	(0.35)	(0.68)	(0.18)
Distance 3	9.90	0.69	11.09	0.77	5.78	0.48
	(0.77)	(0.19)	(0.87)	(0.21)	(0.47)	(0.11)
Distance 4	8.69		9.74		4.67	
	(0.62)		(0.70)		(0.33)	
Distance 5	8.21	-0.13	9.20	-0.15	4.25	-0.04
	(0.46)	(0.15)	(0.51)	(0.17)	(0.24)	(0.08)
Distance >5	7.27	-0.33	8.15	-0.37	3.69	-0.12
	(0.33)	(0.18)	(0.37)	(0.20)	(0.17)	(0.09)
Unconnected	7.55	-0.06	8.45	-0.07	3.85	0.04
	(0.64)	(0.22)	(0.71)	(0.24)	(0.34)	(0.12)
Tech x Year FE	No	Yes	No	Yes	No	Yes
Mean Dep.	11.56	11.56	12.95	12.95	6.79	6.79
Obs.	1135757	1134139	1135757	1134139	1135757	1134139

Table A.5: Different assumptions for KPSS calculation

This table shows the results from OLS regressions of distance from science on Kogan *et al.* (2017) patent value estimates. In Columns 1 and 2, we use the same adjusted KPSS values used in the main body of the paper (deflated 12%). Columns 3 and 4 used the original unadjusted KPSS patent value estimates. Columns 5 and 6 use the extreme "no information benchmark" which assumes the stock market had no knowledge of a patent application prior to the patent's ultimate publication. In Columns 2, 4 and 6 we add technology class $\times$ filing year fixed effects. The standard errors are clustered on the CPC technology class level.

B.3 Firm fixed effects

Another concern might be that by comparing patents with a different distance-to-science we are comparing different company types. Some companies might be closer to science and at the same time produce more valuable patents. To see if our results are driven by firm-specific factors, we control in the below analyses for firm fixed effects. We base the firm fixed effect on the firm identifier provided by the DISCERN database, which matches patents to Compustat firms (Arora *et al.*, 2021b). If a firm identifier is missing we use the assignee identifier from PATSTAT that is based on the disambiguated assignee name.[32]

Controlling for firm fixed effects is difficult to do with Kogan *et al.* (2017) data. The reason is that all patents published by the same firm on the same date have the same value because their evaluation is based on the same abnormal stock market returns. If we calculate the firm fixed effects using dummies for each firm in an OLS regression, dates with multiple patents then receive disproportional weight in the estimate of the firm fixed effects without adding new information. We address this problems in two ways which we describe in turn.

Calculating firm fixed effects using a weighted regression. As first approach, we use a weighted regression with the number of patents per firm and issue date as weights to calculate the fixed effects. Then we use KPSS values net of fixed effects as the outcome for our main regression specification (which is unweighted).

Table A.6 shows our patent value results with firm fixed effects. Column (3) corresponds to our main results (in percentage terms), but with firm fixed effects in addition to the 4-digit CPC $\times$ year fixed effects. The percentage results with firm fixed effects are very similar to our baseline results, which we reproduce in Column (1). In column (4) we use the KPSS levels as outcomes and control for firm fixed effects. Here, the science premium is smaller: \$0.64 million for (D=1) patents relative to (D=4) patents, compared to a premium of \$2.64 million without firm fixed effects in our baseline regression (Column 2). These results suggest

[32] As KPSS values are only available for firms and we are interested in the firm at the time of patent assignment, assignees might be a reasonable approximation of firm identifier

that part of the science premium is driven by firm specific characteristics, i.e. that firms that create more value select into invention more proximate to science. However, even *within-firm* comparisons of patent value still point to a substantial science premium that is decreasing in distance-to-science.

Only use the first patent per firm and issue date. To address the multiple patents problem, a second approach is to include only one patent per firm and issue date. This sample restriction also ensures that each issue date receives the same weight in the calculation of the firm fixed effect. In column (5) and (6) of Table A.6, we show the results including only the "first" patent per firm and issue date, the patent with the lowest publication number per firm and issue date, in the sample. The results in percentage terms are smaller in magnitude, but show the same overall pattern as columns (1) and (3). In levels the results are larger, with premiums in between the baseline specification (column 2) and the full sample with firm fixed effects (column 4). In column (7), we use forward citations as outcome and estimate firm fixed effects directly in the main regression. The results yield the same pattern as our main results. In column (8) we use the probability that a patent is involved in litigation as the outcome and find that patents that are more proximate to science are more likely to be involved in patent litigation, also controlling for firm fixed effects.

Reconciling our results with Arora *et al.* (2021a)

The concurrent Arora *et al.* (2021a) paper analyzes a similar setup as we do and also includes firm fixed effects in their regression. They find only a small science premium within firms of around 1% for patents citing scientific articles, as opposed to patents that do not cite scientific articles. This premium is significantly smaller than the science premium of about 25% that we find in our firm fixed results in B.3. In Table A.7. In this section, we aim to reconcile our findings with theirs.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Baseline		With firm fixed effects					
	Dollar		Dollar		Dollar		Cita-tions	Prob(Litigation)
	%	Level	%	Level	%	Level	%	x 1000
Distance 1	0.26	2.64	0.25	0.64	0.09	1.26	0.69	2.77
	(0.04)	(0.36)	(0.08)	(0.18)	(0.03)	(0.37)	(0.06)	(0.41)
Distance 2	0.18	1.88	0.19	0.48	0.06	0.86	0.25	0.87
	(0.03)	(0.31)	(0.06)	(0.15)	(0.02)	(0.27)	(0.02)	(0.35)
Distance 3	0.07	0.69	0.06	0.16	0.02	0.34	0.07	0.02
	(0.02)	(0.19)	(0.05)	(0.12)	(0.01)	(0.19)	(0.01)	(0.30)
Distance 4								
Distance 5	-0.01	-0.13	-0.00	-0.01	-0.01	-0.08	-0.05	-0.53
	(0.01)	(0.15)	(0.05)	(0.12)	(0.01)	(0.22)	(0.01)	(0.40)
Distance >5	-0.03	-0.33	0.03	0.09	0.02	0.28	-0.11	-0.62
	(0.02)	(0.18)	(0.05)	(0.12)	(0.02)	(0.24)	(0.01)	(0.37)
Unconnected	-0.01	-0.06	0.08	0.20	0.02	0.30	-0.11	-0.57
	(0.02)	(0.21)	(0.06)	(0.15)	(0.02)	(0.25)	(0.02)	(0.36)
Firm FE:	No	No	Yes	Yes	First	First	Yes	Yes
Tech x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean Dep.	11.56	11.56	11.56	11.56	15.10	15.10	29.88	6.39
Obs.	1136496	1136496	1132296	1132296	370230	370230	1132278	1132278

Table A.6: Firm fixed effects

This table reports regression results using various patent value as outcomes and distance-to-science as independent variable. In Columns 1 and 2 we repeat our baselines results in percent terms and levels with (4-digit) CPC $\times$ filing year fixed effects but without firm fixed effects. In Columns 3 and 4, we add firm fixed effects to our baseline specification. Columns 5 and 6 repeat the specifications of column 3 and 4 but we only use the patent with the lowest publication number ("First") per firm and issue date. In Column 7 we use forward citations as outcomes. Columns 8 is an OLS specifications where the outcome variable is an indicator variable for whether or not the focal patent was ever involved in litigation (multiplied by 1000). For all models, the standard errors are clustered on the CPC technology class level.

In column 1 of Table A.7, we repeat our results with firm fixed effects from column (3) in Table A.6, which shows that a D=1 patents is 25% more valuable relative to a patent in the same class and filing year with a distance of 4. The first main difference of our empirical setup to the setup of Arora *et al.* (2021a) is that they compare patents with citation to science to patents without. Thus, they compare D=1 to all other patents and not to D=4 patents. In column 2, we mimic their approach and find a science premium of 11%. This is plausible as many companies have at the same time D=1 and D=2 patents that have a higher value than D=4 patents. In the following, we use 11% as a benchmark to compare the results of Arora *et al.* (2021a) with ours.

The second main difference is the sample. We use the full sample of Kogan *et al.* (2017), while Arora *et al.* (2021a) use only companies those headquarters is in the US. We believe that their data is either equivalent or closely related to the DISCERN database compiled and published by a subset of the authors Arora *et al.* (2017). If we restrict our sample to patents covered in DISCERN, we find in column 3 an imprecisely estimated science premium of zero (95% CI [-10.2%, 10.1%]). While the headline number of Arora *et al.* (2021a) is 1%, this number also drops close to 0% with many controls.

The key difference between the DISCERN sample and the latter is that the full sample also includes companies whose headquarters are outside the US, but which are still listed on the US stock market. Figure A.2 shows the KPSS values for the DISCERN sample and for all other companies over time. The patents in the DISCERN sample are more valuable on average. They also show a much larger increase in the Dot-Com Bubble years.

This strong, and arguably exogenous, increase in patent value in the years 1996 to 2002 make it difficult to estimate the (time-invariant) firm fixed effect. For example, if we estimate the firm FE using dummy variables and OLS, OLS will give disproportional weight to the very high values in the Dot-Com period.

Consistent with the idea that the Dot-Com period poses a statistical problem, we drop companies affected by the bubble in column 4 and find a science premium of 8%. To identify companies affected by the bubble, we use the price-to-sales ratio in 1999 and sort companies into terciles based on the price-to-sales ratio. Companies in the highest tercile we call "bubble companies.” Using data from Compustat, we classify 39,018 patents as belonging to bubble companies and 645,398 patents as belonging to non-bubble companies. 62,222 patents we cannot classify as either price or sales data are missing. This approach is similar to the method of Brunnermeier and Nagel (2004) and Greenwood and Nagel (2009), which identify stocks affected by the Dot-Com bubble.

To be able to keep all data of all years, we next address the problem using a statistical adjustment. In Column 5, we scale the KPSS values with the the mean of the KPSS value in a issue year and 4-digit CPC technology class. This strategy is inspired by Hall *et al.* (2001), which suggests scaling forward citations by the average of forward citations in a technology class and year. This rescaling purges any common systematic movement of KPSS value over time and technology, i.e., also the effect of the Dot-Com Bubble. It also purges differences in value by patent cohort. Yet, as we are interested in the science-premium within a year and technology, we think this is acceptable. If we use scaled dollars as outcome, we find a science-premium of 14% in terms of scaled patent values in column 5.

In column 6, we use an indicator of whether the KPSS value of a patent is in the Top 10% of patents in an issue year and a 4-digit technology class. This transformation again levels differences between patent cohorts. Using this outcome we find that science-based patents are 19% more likely to be in the Top 10% of KPSS values in a given issue, year and technology.

In columns 1 to 6, we used fixed effects based on a weighted regression as discussed in Section B.3 to address the problem of multiple patents of a firm per issue date. In columns 7 and 8, we only keep the first patent per firm and issue date in our sample. Then we estimate the firm fixed effect directly in the main regression. In column 7, we find a science premium in terms of Top 10% patents of 6%. In column 8, we use all distances to science as the independent variable and find a science premium of 10% relative to D=4 patents.

To summarize, we can closely approximate the results of Arora *et al.* (2021a) with our data. The reasons that we obtain different results are two-fold. First, we compare D=1 patents relative to D=4, while Arora *et al.* (2021a) compare D=1 patents to all other patents. Second, Arora *et al.* (2021a) use a US only sample, while we use the full sample of Kogan *et al.* (2017). It is difficult to estimate firm fixed effects in the US-only sample due to the relatively larger importance of the Dot-Com Bubble. If we account for the Dot-Com bubble, we also find in the US sample a sizable science premium.

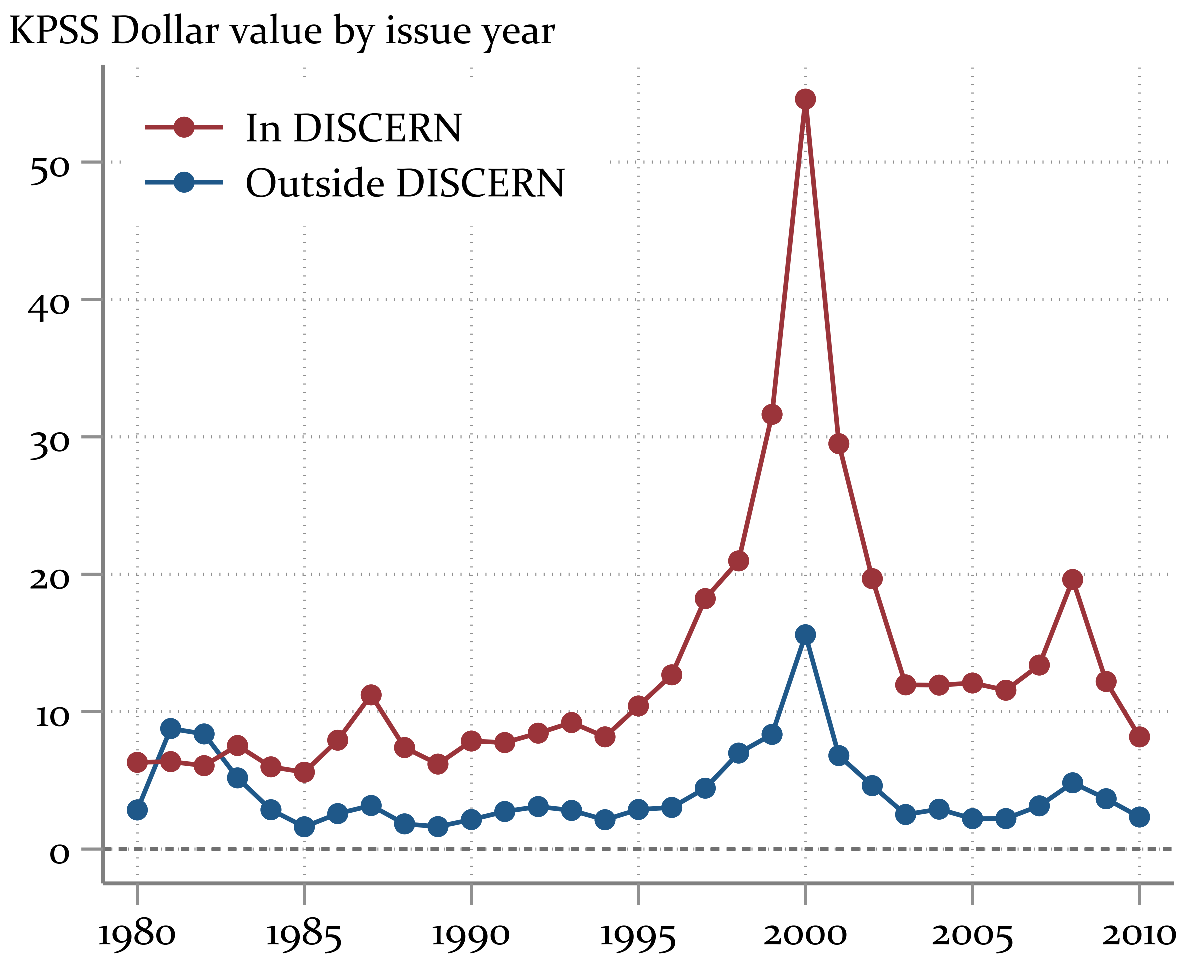


Figure A.2: KPSS values

This figure shows the KPSS value for the DISCERN sample and for all other companies.

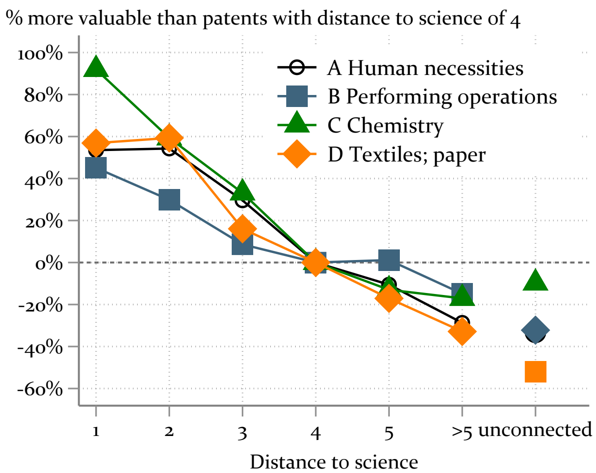
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sample:	All	All	DISCERN	DISCERN w/o bubble companies	DISCERN		DISCERN	
Outcome:	Dollar %	Dollar %	Dollar %	Dollar %	Scaled Dollar %	I[Top 10% dollar] %	I[Top 10% dollar] %	%
Distance 1	0.25	0.11	-0.00	0.08	0.14	0.19	0.06	0.10
	(0.08)	(0.05)	(0.05)	(0.04)	(0.04)	(0.04)	(0.02)	(0.03)
Distance 2	0.19							0.07
	(0.06)							(0.02)
Distance 3	0.06							0.02
	(0.05)							(0.02)
Distance 4								
Distance 5	-0.00							0.00
	(0.05)							(0.02)
Distance >5	0.03							-0.00
	(0.05)							(0.03)
Unconnected	0.08							-0.00
	(0.06)							(0.03)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	First	First
Tech x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	1132338	1132338	740323	643379	740323	740323	253329	253329

Table A.7: Firm FE with different sample

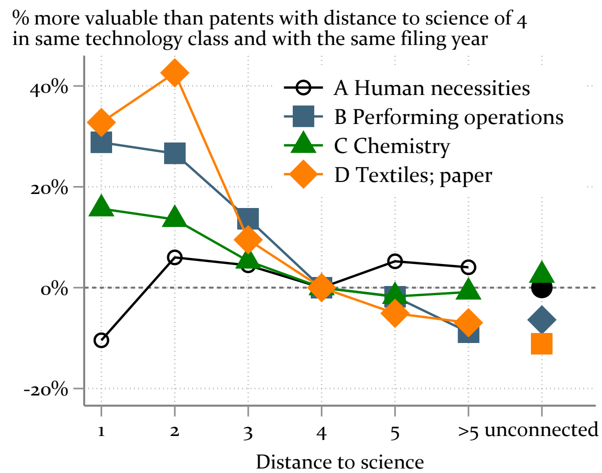
In Columns 1 and 2 we use the full sample and KPSS values as outcomes. We report the results in percent terms and levels with (4-digit) CPC $\times$ filing year fixed effects and firm fixed effects. In Columns 3 to 8 we restrict the sample to the DISCERN sample. In column 3 we the DISCERN sample and use KPSS value as outcome. In column 4 we drop patents of bubble companies. Bubble companies are defined as being in the highest tercile of the price to sales ratio in 1999 and included in Compustat. In column 5, we standardize the KPSS value by the average of the KPSS value in the issue year and 4-digit CPC technology class. In columns 6 to 8 we use as outcome an indicator whether the patent is in the Top 10% of KPSS values in a publication year and 4-digit CPC technology class. In column 7 and 8 we only use as sample only the first patent on the issue date per firm.

B.4 Split by technology

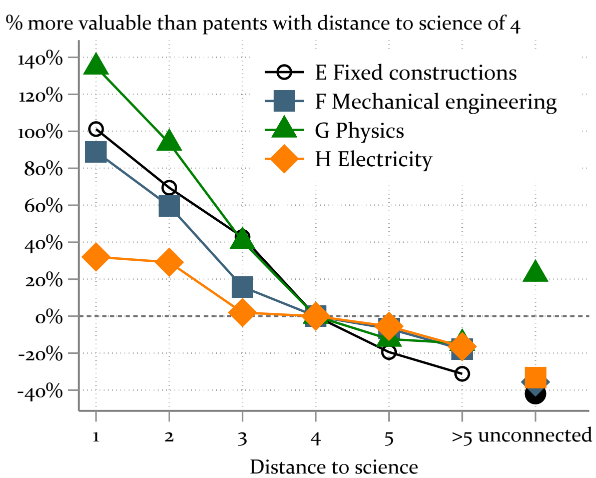
One concern might be that the observed effects are driven by a single technology that benefits particularly from science. This is not the case. In Figure A.3, we show the main graph separately for broad technology categories measured by one-digit CPC classes. Panels (a) and (b) show the raw data. In Panels (c) and (d) we normalize by the average values of patents in the same four-digit CPC technology classification and filing year. For all technologies, there is a decrease in value by distance-to-science, most pronounced for drugs and chemicals. Results by other technology classifications such as the classification in Hall et al. (Hall *et al.*, 2001) and Schmoch (Schmoch, 2008) are available from the authors on request.



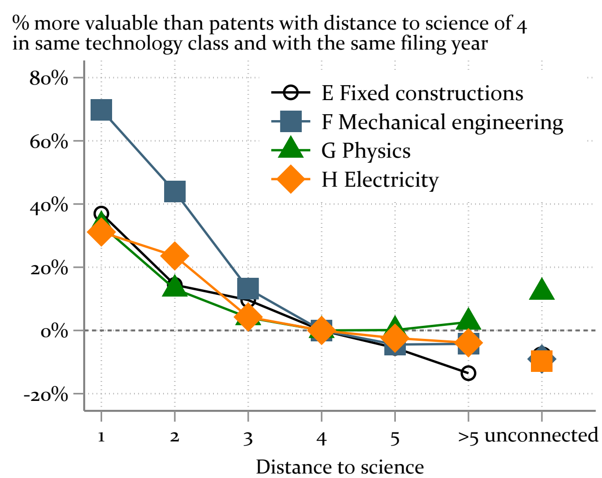
(a) Value by broad patent class A-D



(b) Accounting for technology× year



(c) Value by broad patent class E-G



(d) Accounting for technology × year

Figure A.3: Sample splits by technology

In this figure, we split patents by broad technology fields measured by one-digit CPC classes. In Panels (a) and (c) we show the raw data. In Panels (b) and (d) we normalize by the average patent value of a patent in the same four-digit CPC class and filing year with a distance of four.

B.5 Comparing Patent and Scientific Article Text

One potential concern one might have 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 therefore include more citations, but without actually using science.

To see whether patents close to science make use of its content, we compare the texts of scientific articles and the text 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 of the patent we use the "term frequency-inverse document frequency" (tf-idf) method.[33]

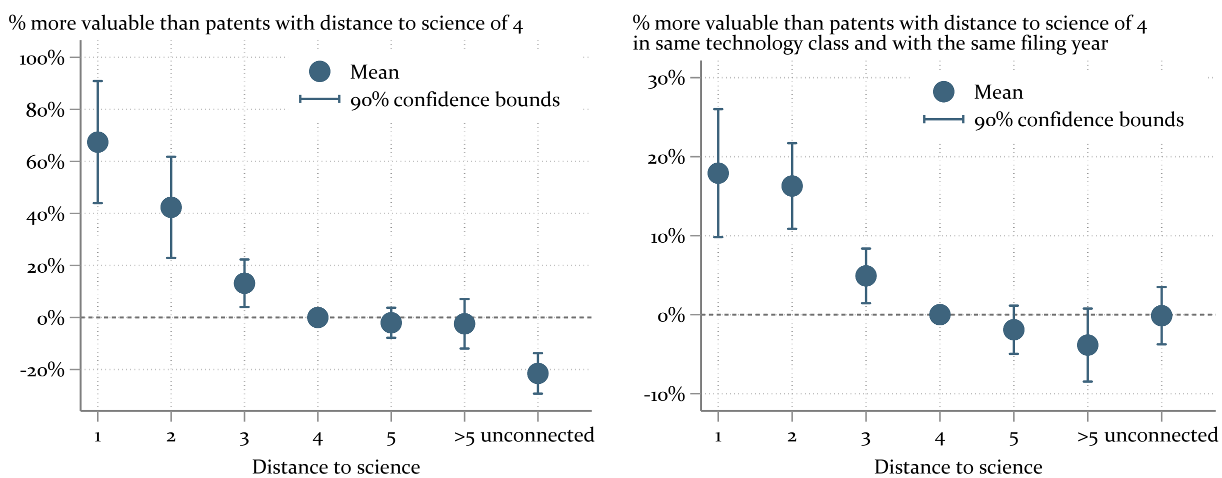
The results presented in Column 7 of Table A.4 show that patents with a citation distance of D=1 have a higher text similarity to scientific articles than patents more distant to science. This suggests that citation distance reflects indeed how much a patent is related to science. Consistent with the idea that patents with more scientific content have a higher value, Column 8 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 that are unrelated to the scientific content of the patent.

B.6 Using Ahmadpoor and Jones distance-to-science values

In Figures A.4a and A.4b, we use the distance-to-science measure based on the data of Ahmadpoor and Jones (2017). The data of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017) is based on Web of Science while our measure is based on the data of Microsoft

[33] We use the "gensim" implementation in Python for our calculations (see: `https://radimrehurek.com/gensim/`).

Academic. There are two main differences. First, Ahmadpoor and Jones (Ahmadpoor and Jones, 2017) have many more unconnected patents (165 thousand unconnected patents out of 0.8 million overall patents) than we do (54 thousand unconnected patents out of 1.1 million overall patents). Second, we aggregate the patents with a distance larger than 5 as the number of patents goes down dramatically for larger distances. We find the same overall pattern; i.e., the patent values decrease in percent relative to D=4 with distance-to-science.



(a) Ahmadpoor and Jones (2017) distance

(b) Controlling for technology and year

Figure A.4: Value of patents by distance-to-science

Panel (a) shows the average patent value for all distances to science relative to patents with a distance of 4 (D=4). The values of U.S. patents are from Kogan et al. (2017). The distance-to-science of U.S. patents is from Ahmadpoor and Jones (2017). In Panel (b), we residualize the patent value by the average value of a patent with the same (four-digit) CPC technology class and filing year and a distance of four, then display relative (%) values indexed to D=4.

B.7 Number of citations to science and value

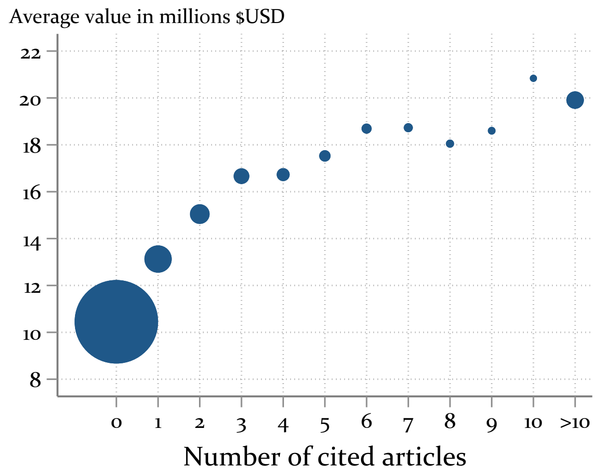
Our primary measure of distance-to-science does not account for the intensity of citations to science. A patent that cites one single scientific article and a patent citing dozens of journal articles both end up in the (D=1) group. To investigate whether the number of citations to scientific articles is correlated with patents' private value, we use number of cited articles as an alternative distance-to-science measure.

Appendix Figure A.5 shows the results of this analysis. Both panels graph Kogan *et al.* (2017) values for different numbers of citations to scientific articles. The size of each observation circle is proportional to the number of observations for each value of "number of cited articles." The modal patent has no citations to science. Panel (a) displays the raw averages and shoes a fairly steady increase in average value as the number of cited articles increase.

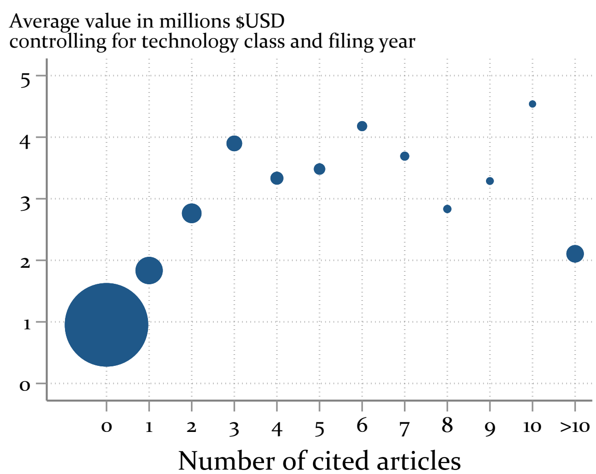
In Panel (b), we residualize values using the average value of patents in the same technology class and filing year. The residualized values show a similar pattern up until seven cited articles, after which the values move more erratically. Notably, patents with more than 10 cited articles have an average value closest to patents with merely one scientific citation.

We cannot pinpoint the exact reason for the drop off in residualized values beyond 10 citations, but we can speculate. Within some technology classes, a large number of scientific citations may be indicative of a patent being a more incremental innovation. Relying extensively on a series of past discoveries could mean that the patent is situated in a more mature and crowded technological space. While reliance on science is generally valuable in those technology classes, excessive citations might signal a narrowness or lack of novelty that reduces the patent's private value. Such a pattern is in line with Krieger *et al.* (2021), who find that drug candidates that are highly similar in structure to prior drugs have lower KPSS patent values.

More generally, the number of scientific cites among the (D=1) is a noisy measure of how reliant a paper is on science. Some subfields simply have a norm of more citations Marx and Fuegi (2020). Additional citations in those fields are more likely to be ceremonial in nature. In such high citation scientific areas, we do not expect more citations to correlate with higher private value, nor relying any more on science than in areas where citing patterns are more selective.



(a) Averages



(b) Averages accounting for technology and year

Figure A.5: Number of articles cited and patent values

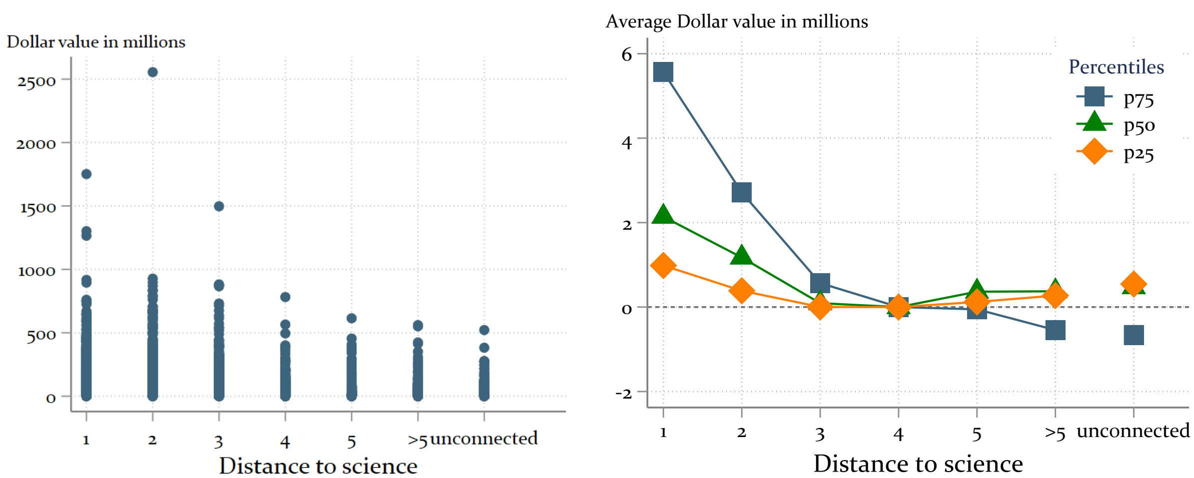
Panel (a) shows the average patent value and patents are split by the number of scientific articles cited in the patent. The data on non-patent citations is from Marx and Fuegi (2020). In Panel (b), we residualize the patent value by the average value of a patent with the same (four-digit) CPC technology class and filing year and a distance of four.

B.8 Patent Tail Outcomes

Our primary regression results look at how average patent values vary with distance from science (Table 1). However, as we see in Figure 1 (Panels (b) and (d)) The increase in value due to science also comes with an increased likelihood of tail outcomes accounting for technology and year. In Columns 5 and 6 of Appendix Table A.9, we investigate the likelihood that a patent is in the top 5% or bottom 5% of the distribution of the science component as an outcome. The distribution is taken over all patents. Patents that are closer to science have a higher likelihood to be in the tails of this distribution. So, accounting for the technological component, science-based patents show a larger variance in values.

B.9 Effects over the entire value distribution

The main paper shows that average patent values decrease with distance to science relative to the average value of a patent with the same filing year and the same (four-digit) CPC technology classification with a distance of four. This pattern is already visible in the raw data in Figure A.6a. In Figure A.6b we show the value by distance of science over the 25th, 50th and 75th percentile of the value distribution. We residualize each percentile with the same percentile of patents with D=4.[34] The patent values are falling with distance-to-science over all percentiles. This confirms that our results are not driven by outliers.



(a) Raw data

(b) Value percentiles residualized with same percentile in D=4

Figure A.6: Effects over the value distribution

Panel (a) shows the raw data for patent values by distance-to-science data for a 10% sample of patents. The values of U.S. patents are from Kogan et al (Kogan *et al.*, 2017). The distance to science of U.S. patents is calculated with Marx and Fuegi Marx and Fuegi (2020) and Patstat using the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). The distance-to-science is defined by citation links. A patent that directly cites an academic article has a distance of D=1. A patent that cites a (D=1)-patent but not an academic article has a distance of D=2. Patents are defined as "Unconnected" if there is no citation link to an academic article. In Panel (b), we show the average patent value for all distances to science along with the number of patents in each distance. Panel (b) shows the 25th, 50th and 75th percentiles of the patent value distribution by distance-to-science normalized by its percentile at a distance of four.

[34] Here, we do not account for technology and year, as for many technology-filing year combinations there are not enough patents to obtain a distribution for every distance-to-science.

B.10 Alternative measures for novelty

New words and word age

In our main specification, we measure how likely or unlikely the word combinations used in a patent are to determine the novelty of a patent. In Figures A.7a and A.7b, we use the data of Arts et al (Arts *et al.*, 2018) to calculate two alternatives measures for novelty. The first measure indicates whether a patent has a new word. A word is new if it was not used in any patent before. Figure A.7a shows that the share of patents with a new word decreases monotonically with its distance-to-science. The second measure is the average age of words used in a patent. We calculate the age of a word by calculating the difference between the filing year and the filing year of the patent in which it was first used. The average word age is systematically lower for patents that are closer to science (Figure A.7b). If word age is indicative for the age of the ideas they encode, patents closer to science contain more novel ideas.

Both of the alternative novelty measures are positively related to patent value. Figure A.7c compares the patent value for patents with and without new words for each year, controlling for technology $\times$ year fixed effects. From 1980–1998, we see that patents with new words have a positive (average) value premium over those without new words. Then, from 1998–2000, patents with new words are actually less valuable on average. Presumably this period featured a lot of highly valued, but (relatively) low novelty software patents without new words. Firms may have been able to capture value from "jumping on the bandwagon" with software patents in that brief period, but the "bursting" of the dot com bubble in 2000 suggests that many of those software patents were indeed overvalued in the longer run.

Figure A.7d shows the relation between the average age of words of a patent relative to the year and technology class and the residualized patent value. We see a negative relation between patent value and the average age of words.

Chemical Patents and Compound Novelty

A potential shortcoming of measuring novelty with text analysis is that inventors and patent lawyers have many degrees of freedom in choosing words. Word or word combination novelty might reflect the author's preferences or popular tech buzzwords, moreso than innate qualities of the invention. Endogenous language choices might erroneously hide or suggest true novelty (e.g., consider the use and abuse of word combinations involving terms like "crypto" and "artificial intelligence" over time). To avoid the pitfalls of using language to measure novelty, ideally innovation scholars would have standardized and quantifiable measures of technological similarity for all technology classes over time.

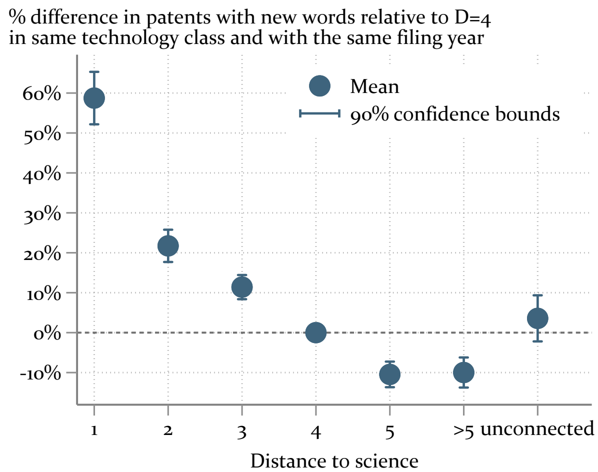
While such standardized novelty measures do not exist for most technologies, chemical patents allow for such structured measures of invention qualities. We use the crosswalk of patents-to-compounds from SureChEMBL, which extracts the standardized (SMILES code) chemical structures represented in patents.[35] Patents often have more than one chemical described in their claims, and these chemicals need not be the central component of the the new invention.

One way that chemists quantify the similarity of two given molecules is to calculate their share of overlapping fragments as a "Tanimoto" (or Jaccard) chemical similarity score from zero (no overlap) to one (total overlap). Using the general approach described in Krieger *et al.* (2021) and with the aid of the ChemmineR package (Backman *et al.*, 2011), we calculate all the maximum pairwise (backwards) similarity between each of the compounds in a given patent and all previously patented chemicals from the same 3-digit CPC class. The average maximum similarity to prior patent chemicals is 0.86 with a median of 1. This general lack of novelty is unsurprising since patents contain multiple compounds and inventive novelty might come from recombination of existing compounds or complementary technologies claimed by the patent. Thus, our chemical similarity measure is a conservative measure of novelty.

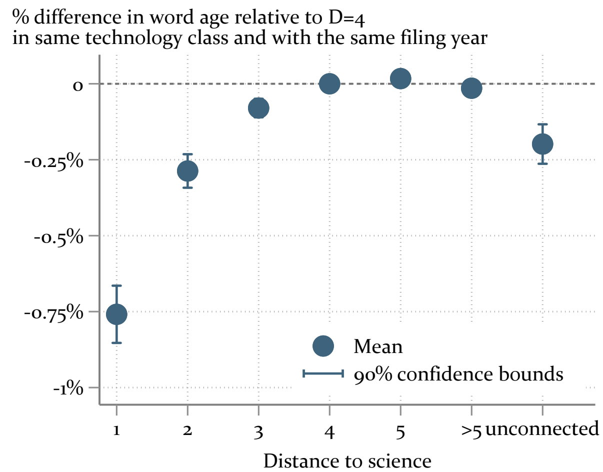
[35] The data includes 187,958,584 patent-compound pairs from 1960–2014. For more information on the data see: `http://chembl.blogspot.com/2015/03/the-surechembl-map-file-is-out.html`.

As a robustness check on our main results, we merge the chemical patent similarity measures to our main analysis data set and evaluate the 58,668 public firm chemical patents in our data. Table A.8 presents the results. Controlling for technology (4-digit CPC) $\times$ year fixed effects, we find that chemical similarity to past patents is negatively associated with KPSS dollar values (Column 1). In Column 2, we show that the most novel compounds (those with backwards similarity < 0.25) have a \$5.35 million higher average KPSS valuation than less novel chemical patents.

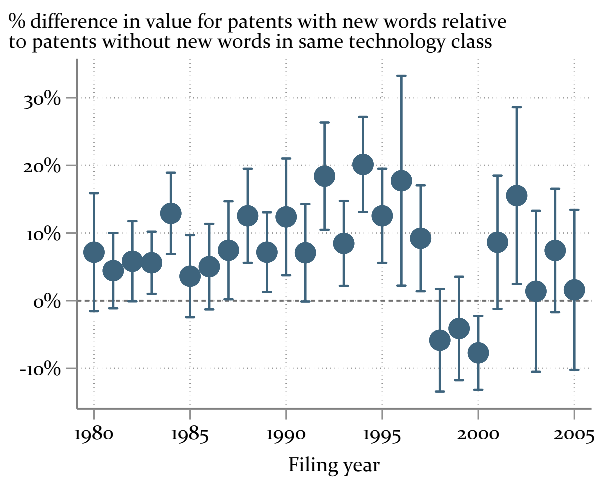
Finally, in Column 3 we evaluate the association between distance-to-science and chemical patent novelty. Compared to (D=4) patents and controlling for technology (4-digit CPC) $\times$ year fixed effects, (D=1) and (D=2) patents have lower similarity (more novelty) on average. Interestingly, patents totally unconnected to science also tend to have less similarity to past patented chemicals, while the other distance groups show no difference from (D=4). This pattern mirrors the relationships we find using the new words and word age measures (Figures A.7a and A.7b), where unconnected to science patents also exhibited greater novelty than the (D>3) groups.



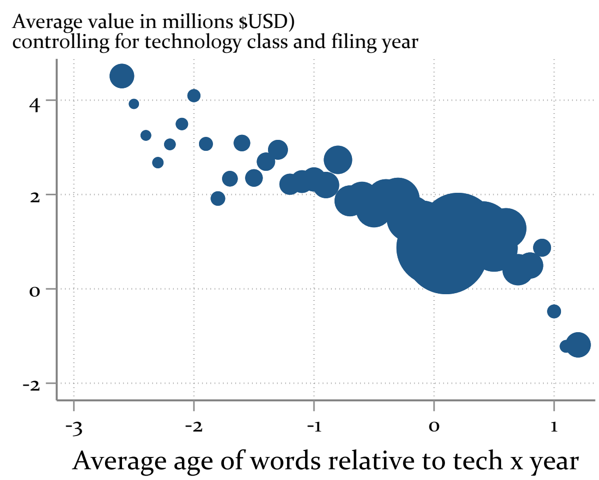
(a) Novelty: New words



(b) Novelty: Average word age



(c) New words and value



(d) Word age and value

Figure A.7: Value and novelty

Panel (a) shows the share of patents that have a new word by distance-to-science. A new word is a word that has not been mentioned before in a patent according to the data of Arts et al (Arts *et al.*, 2018). In Panel (b), we plot the average word age by distance-to-science. The age of a word in a patent is the difference between the filing year of the patent and the year the word was first used in a patent. In Panel (c), we plot the average percentage difference in dollar value of a patent with a new word relative to a patent without a new word within the same filing year and (four-digit) CPC technology class over time. In Panel (d), we plot the average word age and the average patent value. The word age is relative to the mean of patents in the same filing year and the same (four-digit) CPC technology class.

Outcome:	(1) Dollars	(2) Dollars	(3) Chemical Similarity
Chemical Similarity	-4.85		
	(2.36)		
Most Novel Compounds		5.35	
		(1.99)	
Zero Novelty Compounds		-0.48	
		(0.74)	
Distance 1			-0.037
			(0.014)
Distance 2			-0.012
			(0.0069)
Distance 3			-0.00081
			(0.0044)
Distance 5			0.0052
			(0.0058)
Distance >5			-0.0012
			(0.0060)
Unconnected			-0.060
			(0.020)
Tech x Year FE	Yes	Yes	Yes
Observations	60,009	60,009	60,009

Table A.8: Chemical novelty, patent value and distance-to-science

NOTES: Table A.8 presents OLS regressions for the subset of patents associated with one or more chemical structure. The patent to chemical crosswalk comes from SureChEMBL (`www.surechembl.org`), and chemical similarity is calculated from zero (no similarity) to one (total structural overlap) based on Tanimoto scores. Most novel compound patents are those with at least one compound with a maximum chemical similarity to previously granted patents of zero or less than 0.25. Zero novelty compounds patents are those which have no compounds with a backwards similarity less than 1. The omitted category is for patents with between 0.25 and 0.99 chemical similarity to prior patents. The dependent variable in Columns 1 and 2 is the (adjusted) KPSS patent dollar value. Column 3 presents the correlations between distance-to-science and patent maximum chemical similarity to prior patented chemicals. All models have technology (4-digit CPC) $\times$ year fixed effects.

B.11 Patent value by distance-to-science and novelty

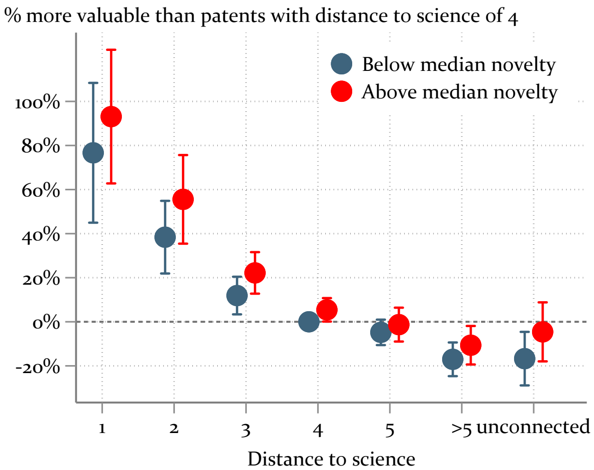
Distance-to-science and novelty are correlated, however within distance-to-science group, we still observe variation in novelty of word combinations. This variation allows us to explore how patent value differs if we turn novelty "on" or "off." In Appendix Figure A.8 we compare patent values by both distance-to-science and novelty. Specifically, we regress patent value on interactions between distance-to-science and an indicator for whether patents are below or above median novelty. Novelty is measured by the likelihood of pairwise combinations of words that occur in a particular patents after account for technology class and filing year (i.e., the same novelty measure as Figure 2, Panels (c) and (d)). In Appendix Figure A.8, Panels (a) and (b) graph patent values as percentage differences relative to the (D=4), low novelty group, with and without technology $\times$ year fixed effects. Panels (c) and (d) show the equivalent regression results, but reporting coefficients in levels (millions of \$USD), rather than percentage differences.

We discuss the results at the end of Section 4.3.

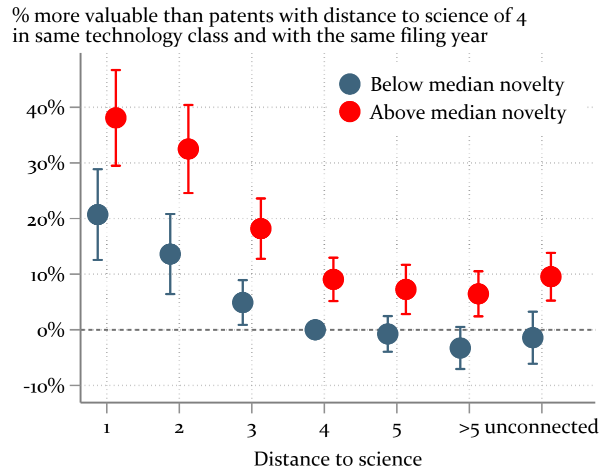
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Patent Value				Patent Value relative to D=4 by tech and filing year			
Outcome: Probability of	Top 5%	Bottom 5%	Top 5%	Bottom 5%	Top 5%	Bottom 5%	Top 5%	Bottom 5%
Distance 1	7.8	3.1			1.3	-0.6		
	(1.1)	(0.5)			(0.2)	(0.2)		
Distance 2	5.5	4.8			0.9	-0.5		
	(0.6)	(0.5)			(0.2)	(0.2)		
Distance 3	4.1	6.6			0.3	-0.3		
	(0.4)	(0.6)			(0.1)	(0.1)		
Distance 4	3.3	6.2						
	(0.3)	(0.7)						
Distance 5	2.7	5.6			-0.0	0.1		
	(0.2)	(0.8)			(0.1)	(0.1)		
Distance >5	2.0	5.3			-0.2	0.2		
	(0.2)	(0.8)			(0.1)	(0.1)		
Unconnected	2.2	4.0			0.0	0.3		
	(0.3)	(0.5)			(0.1)	(0.1)		
Probability of word combinations			-26.7	25.1			-4.8	1.0
			(2.7)	(4.5)			(0.8)	(0.5)
Mean Dep.	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Obs.	1135757	1135757	1135669	1135669	1135757	1135757	1135669	1135669

Table A.9: Distribution of patent value

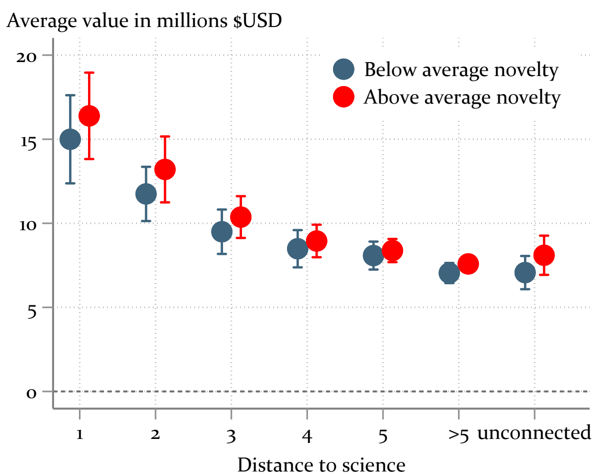
This table shows the distribution of patent values by distance-to-science and by patent novelty. In columns 1 and 3, we use the patent values of Kogan et al. (Kogan *et al.*, 2017) to assign to each patent an indicator equal to one if its value is in the top 5% of all patent values. In columns 2 and 4, we assign an indicator equal to one if the patent is in the bottom 5% of the patent value distribution. In columns 1 and 2, the independent variable is the distance-to-science measured by citation links. Unconnected patents are patents for which we could not find a citation link to any scientific article. In columns 3 and 4, we use a novelty indicator - the average probability of 2-tuple word combination of each patent - as an outcome variable. In columns 5 to 8, we use instead of the distribution of all patent values the distribution of the science value component to calculate the top and bottom 5% of the distribution. The science value component is derived by calculating the residual of the patent value and the average patent value of a patent in the same technology and year with a distance of four. The standard errors are clustered on the CPC technology class level.



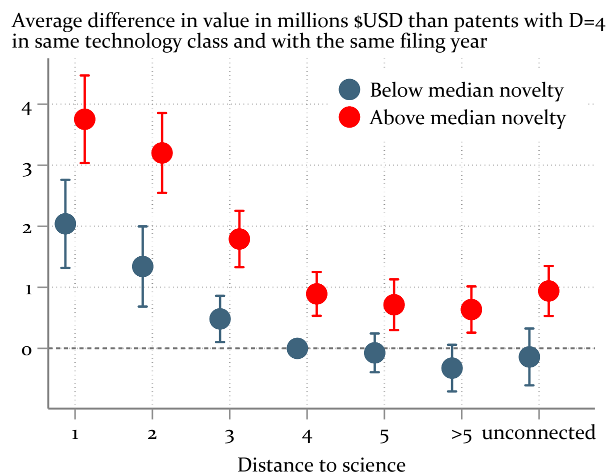
(a) % differences



(b) % difference, accounting for tech and year



(c) Averages ($ levels)



(d) $ level differences, accounting for tech and year

Figure A.8: Patent value by distance from science and novelty

Panel (a) shows the average patent value for all distances to science relative to patents with a distance of (D=4), and separates by patents with above and below median novelty. Novelty is measured by the likelihood of pairwise combinations of words that occur in a particular patents and we account for technology class and filing year. The values of U.S. patents are from Kogan et al. (Kogan *et al.*, 2017). The distance-to-science of U.S. patents is calculated with Microsoft Academic and Patstat using the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). The distance-to-science is defined by citation links. The 90% confidence bounds are based on standard errors clustered by (four-digit) CPC technology class. In Panel (b), we residualize the patent value by the average value of a patent with the same (four-digit) CPC technology class and filing year and a distance of four. We again plot seperate values for patents with above and below average novelty. Panel (c) shows the average dollar value of patents in levels for all distances to science for both below and above average novelty patents. Panel (d) residualizes dollar patent values by the average value of a patent with the same (four-digit) CPC technology class and filing year and a distance of four.

B.12 Firm Heterogeneity: Science Intensity

We explore firm heterogeneity in firm's "science intensity." We focus on two dimensions of science intensity. First we consider firms' propensity to cite scientific publications (reliance on science), and second we examine in-house vs. external citations within the (D=1) group.

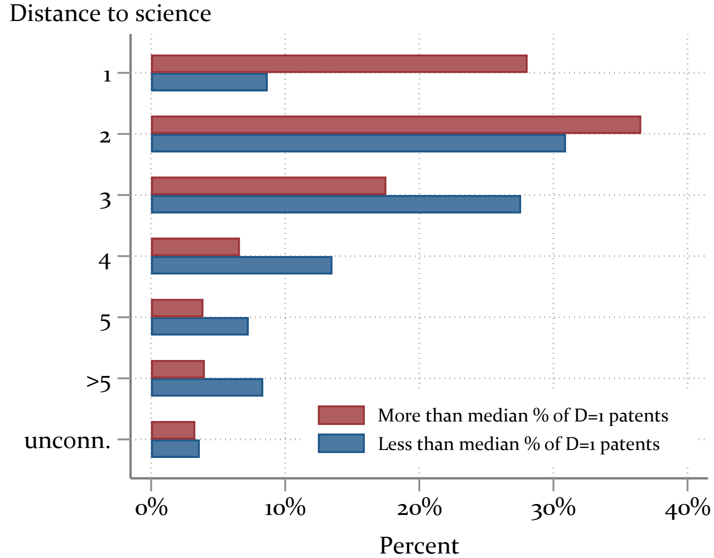
In Appendix Figure A.9 Panel (a) we show the distribution of patents across distance-to-science, split by above and below median levels of firms' "reliance on science."[36] Notably, firms above the median rely directly on scientific articles in just under 30% of patents. The above median firms have an interquartile range between 17%–53% for (D=1) patents in any given year, as opposed to 0%–7% for the below median group. The modal patent in both groups is (D=2), but the below median group is twice as likely to have patents with a distance (D$\geq$4).

Panel (b) graphs the coefficients comparable to the specification in Panel (c) of Figure 1. We regress KPSS patent values on distance-to-science and control for technology (CPC 4-digit) $\times$ year fixed effects, reporting differences relative to the (D=4) group. Here, we additionally split the sample by whether the producing firm was above vs. below median in reliance on science intensity, and use the below median (D=4) group as the reference group. Another difference from is that we report coefficients in levels (dollars) rather than percentages, because the average values are generally lower for the below median group, and thus make percentage premia difficult to interpret across the two groups. Instead, we focus on the average marginal value (dollar) differences across groups. We find that (within technology $\times$ year groups), above median reliance on science firms gain an average of \$1.98 million, while below median firms gain only \$1.22 million. Under the assumption that the marginal cost of a (D=1) project is lower for the above median reliance on science firms—more experience and expertise in using science for invention—the difference in profitability of a (D=1) project across the two types of firms is likely even greater. Thus, these patterns would reinforce

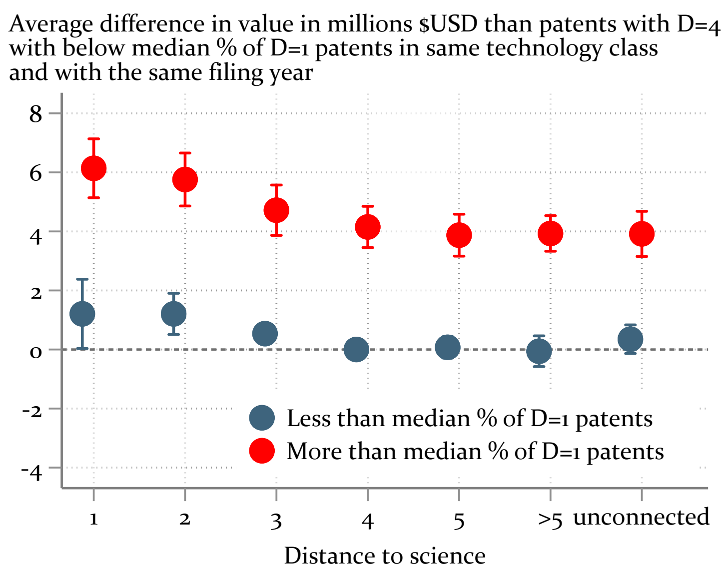
[36] We define reliance on science intensity as the firm's % of (D=1) cumulative patents, normalized by year (the median percentage changes over time).

differences in reliance on science rates, unless the long-term benefits of investing in science were inspires low reliance firms towards more (D=1) invention as a longer term investment in absorptive capacity.

Next, in Appendix Table A.10 we show the additional value associated with in-house (D=1) citations to science. Each specification suggests that when firms invent by building off scientific publications authored by their own scientists, those patents are associated with substantially more KPSS dollar values, after controlling for technology $\times$ year fixed effects. Columns 3 and 4 show that both high (above median) reliance on science and in-house (D=1) publications are associated with greater patent value. Our most saturated regression in the table (Column 5) interacts distance-to-science groups with whether the firm is below/above the median in reliance on science, and then further interacts the (D=1) groups by whether they cite in-house publications. Those coefficients show that high reliance on science firms generate more average value when they invent based on their own scientific publications than do firms building on in-house publications, but who less frequently use science in invention.



(a) Distribution of Patenting, by Distance-to-Science



(b) KPSS Values, Controlling for technology and year

Figure A.9: Patenting Rates and Values, by Distance-to-Science and Firm (D=1) Intensity

Panel (a) shows the percentage of patents for each distance-to-science, by whether the assignee firm was in the above vs. below median of percentage of (cumulative) patents that build directly on science (D=1). Panel (b) shows the average KPSS value for the above vs. below median groups and by distance-to-science. The baseline group is (D=4) for below median reliance on science firms, such that the y-axis should be interpretted as incremental value relative to that group's average value. The underlying regression controls for technology and year fixed effects.

	(1) Dollars	(2) Dollars	(3) Dollars	(4) Dollars	(5) Dollars
Distance 1	0.92	2.17		1.39	
	(0.35)	(0.42)		(0.36)	
Distance 2		1.89		1.45	
		(0.31)		(0.32)	
Distance 3		0.70		0.56	
		(0.19)		(0.20)	
Distance 4 ,					
Distance 5		-0.13		-0.10	
		(0.15)		(0.17)	
Distance >5		-0.33		-0.14	
		(0.18)		(0.18)	
Unconnected		-0.088		-0.05	
		(0.22)		(0.22)	
In-House Publication (D=1)	8.41	8.45			
	(3.15)	(3.15)			
High Reliance (Above Median)			4.44	4.29	
			(0.52)	(0.51)	
High Reliance, In-House (D=1)			7.99	7.83	7.81
			(3.02)	(3.14)	(3.18)
Low Reliance, In-House (D=1)				2.19	2.33
				(0.78)	(0.87)
Tech-Year FE	Yes	Yes	Yes	Yes	Yes
Distance X High/Low Reliance	No	No	No	No	Yes
Obs.	1147779	1147779	1098452	1098452	1098452

Table A.10: Firm Heterogeneity: Science Intensity

This table shows the distribution of patent values by distance-to-science and firm scientific intensity. All outcome variables are in dollar terms based on KPSS patent values. Column 1 reports the (D=1) premium and the additional value for (D=1) patents that cite publications authored by scientists employed at the same firm as the patent assignee ("in-house"). Column 2 adds in additional coefficients for distances to science D>1. Column 3 shows the average premium for firms with "high reliance" on science (above the median in firm's percentage of patents with D=1 in the given publication year), as well as the interaction between high reliance and (D=1) in-house patents. Column 4 and 5 further include the interaction between (D=1) in house patents and and indicator for firms that have"low reliance" on science (below the median in firm's percentage of D=1 patents in given year). The standard errors are clustered at the CPC (4-digit) technology class level.