Standing on the shoulders of science

Martin Watzinger Monika Schnitzer*

June 7, 2019

The goal of science is to advance knowledge, yet little is known about its value for marketplace inventions. Analyzing U.S. patents, we establish three new facts about the relationship between science and the value of inventions. First, we show that a patent directly building on science is on average 2.9 million U.S. dollars more valuable than a patent in the same technology but unrelated to science. Based on the analysis of the patent text, we show second that the novelty of patents predicts their value, and third that science-intensive patents are more novel. This documents that science introduces new concepts that are valuable for marketplace inventions. Our study informs the debate on the merits of science for corporate innovation and the origins of breakthrough inventions.

JEL Codes: O30, O34, O33, O31

Key words: Value of innovation, science, patent novelty

*Martin Watzinger and Monika Schnitzer: Ludwig Maximilian University Munich, Department of Economics, Akademiestrasse 1, 80799 Munich, Germany; Martin Watzinger : martin.watzinger@econ.lmu.de, Monika Schnitzer (corresponding author): schnitzer@econ.lmu.de. We thank Jeff Furman, Fabian Gaessler and Fabian Waldinger for helpful comments and discussions. We thank Mohammad Ahmadpoor, Ben Jones, and Bill Kerr for sharing their data. We gratefully acknowledge financial support of the Deutsche Forschungsgemeinschaft through SFB-TR 190.

"Science, by itself, provides no panacea [...] But without scientific progress no amount of achievement in other directions can insure our health, prosperity, and security as a nation"

Vannevar Bush

Introduction

According to philosopher Francis Bacon "The true and legitimate goal of the sciences is to endow human life with new discoveries and resources" (Mokyr, 2016). Yet surprisingly little is known about how much scientific knowledge contributes to the development of marketplace innovations and their commercial success. Some well-known examples document that science can play an important role for the development of technological breakthroughs. Ferdinand Braun and Guglielmo Marconi could not have developed the wireless telegraph before Heinrich Hertz showed the existence of electromagnetic waves. Similarly, the development of the transistor at the Bell Laboratories would have been difficult to imagine without the scientific understanding of the physics of semiconductors. Skeptics have argued that these cases are the exception rather than the rule and that ideas for inventions usually come from other sources than science (Kline and Rosenberg, 1986; von Hippel, 1988). They are concerned that science is trapped in an ivory tower and has little direct application in private companies. Given the public good character of scientific knowledge and the reward structure for scientists (Stephan, 1996), it is indeed non-obvious that they engage in basic research that matters for private sector applications. Moreover, skeptics argue that the knowledge transfer from university to industry does not work effectively or doubt the reliability of the knowledge produced at universities (Goozner, 2005; Butler, 2008; Freedman *et al.*, 2015; Bikard, 2018).

While these are potentially important concerns, to date there is no empirical evidence that would allow us to quantify the overall impact of science on marketplace innovations. In this study, we provide such a quantification by measuring the contribution of science to the value of patented inventions in the private sector. 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 science and how much of it is technology-specific. We solve this challenge with the help of a metric for how science-intensive a patent is. 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.

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 it makes use of. 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). Kogan *et al.* (2017) 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 the average patent value for a given distance to science for 1.2 million U.S. patents filed between 1980 and 2010.

We find that patents directly based on science have on average a private value that is 2.9 million U.S. dollars larger than patents filed in the same technology class and year but only losely related to science. Patents with a distance of two (three) have an implied value of science of $ 2.2 million ($ 0.9 million) U.S. dollars. 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 technology innovation" throughout the economy (Stokes, 2011, p.84). 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 alternative measures for patent value based on citations and patent scope from Kuhn and Thompson (2017).

What makes science-based patents particularly valuable? We identify the novelty of scientific ideas as the link between scientific research and private sector value. To establish this link, we develop a new measure of patent novelty based on the novelty of words in the text of the patent. For this purpose, we calculate for each patent the probability that a given combination of keywords has been used before. We call a patent "novel" if it contains keyword combinations with low probability. We document that patent novelty predicts the value of patents in a very similar way as the science-intensity of patents does. Finally, 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. More specifically, we document that the novelty of science-based patents is driven by novel concepts, not by novel combinations of existing concepts.

Our paper contributes to the literature in two main ways. First, it highlights that science creates value in the private sector, not only directly, but also indirectly, and it quantifies the respective value contributions in dollars. In 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 is primarily focussed 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.[1] 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. We add to these findings by showing that based on stock-market returns the direct effect of science accounts — in dollar terms — for only

[1]Narin *et al.* (1997) reports that over time patents in the private sector cite more and more scientific articles.

around one third of the overall private value. By estimating the monetary private value of science-based patents our results allow to gauge the private incentives of patenting inventors to use science for innovation purposes.

Second, our paper takes an important step towards understanding the role of science for the value of patents by showing that science and the novelty of the innovations protected by the patent 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. Yet, the mechanisms for how science creates value have been little explored. Fleming and Sorenson (2004) have argued that science alters inventors' search processes and leads them to useful new knowledge combinations. We find direct evidence for this mechanism studying the use of novel keyword combinations. Our measure is inspired by Uzzi *et al.* (2013) who use uncommon combinations of citations to journals as a measure for the novelty of an academic article.[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.[3] By linking patent novelty to patent value and science to patent novelty we provide a rationale for why science matters for private sector innovation.

Data

Our starting point is a dataset which contains information on the monetary value of 1.8 million 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

[2] Uzzi *et al.* (2013) show that more novel articles are more likely - under some conditions - to end up in the top 5% of the citation distribution.

[3] 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.

of the filing company within a narrow window around the grant date of the patent.

We calculate for each of these patents 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 where practical inventions and scientific advances are directly linked (Marx and Fuegi, 2019). 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. 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. 19% of all patents directly cite a patent (D=1), 55% are indirectly based on science (D=2 and D=3) and 26% are not based on science (D=4 or larger). The average value of a patent is 12.9 million in constant 1982 U.S. dollars. Appendix A gives a detailed description of the data construction and sources.

Results

In the following, 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.

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. We start by presenting average dollar values of patents along with 95% confidence bounds for different distances to science. As shown in Figure 1, Panel (a), a science-based patent that directly cites an academic article (D=1) has an average value of 17.7 million dollars. This value decreases as the distance to science increases. Patents with a distance of two have on average a value of 14.1 million dollars while patents with a distance larger than three have a value between 8 and 10 million dollars. The last set of patents ("unconnected") do not contain any reference to science or science-based patents. These patents have on average a value of 8.9 million dollars.

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 as the value distribution of less science-intensive patents, the share of patents at each percentile should be 1%. Figure 1, 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.

[Figure 1 about here.]

We examine next 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. It has been noted that "to a considerable extent the scientific enterprise evolves in disciplines that from their beginnings have been closely tied to fields of technology" (Stephan, 1996). This is why we ask how much of the patent value is technology specific and how much can be attributed to the value of science.

To be able 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 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 by looking at the average value of patents that are not science-related. The value of non-science-related patents is the sum of the technological component and the idiosyncratic component whereas by definition the 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. In the following, we define patents that have a distance to science of four as patents that derive little or no value from science. This choice is arbitrary but informed by the data. As shown in Figure 1, Panel (a), the average value of patents is falling as distance to science increases up to a distance of 4. For larger distances, the average patent value remains constant. This suggests that science adds little additional value if a patent has a distance to science of four or larger. If patents with a distance of four still derive value from science, our estimates for the value of science represent a lower bound.

Figure 1, Panel (c) presents the average science value within technology classes by distance to science. To derive these values for patents of different scientific intensity, we calculate the difference between the value of a patent of a given distance to science and the value of a (D-4)-patent of the same technology and year. Then we take averages over all patents of a given distance. Patents that are directly based on science (D=1) have an average science-value of \$ 2.9 million dollars. This means a (D=1)-patent is on average \$ 2.9 million dollars more valuable than a (D=4)-patent of the same technology class, filed in the same year. Patents that are indirectly based on science (D=2 and D=3) have an implied science value of \$ 2.2 and \$ 0.9 million dollars, respectively. This is how much value they add to patents over and above the technology component. These values are significantly lower than the raw values presented in Panel (a) of Figure 1. This indicates that science-intensive patents are indeed more prevalent in high-value technologies than in low-value technologies.

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.[4] Less science-intensive patents tend to have values close to the median of the value distribution. Science-intensive patents instead are more likely to have a value that is in the tails of the science value distribution. Relative to a (D=4)-patent in the same technology and year, more science-intensive patents are more likely to be in the upper and the lower tail of the value distribution. 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 for this could be the high rate of irreproducible research results which has been said to be as 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. In Appendices B and C we show that these results are robust to alternative method choices and aross technologies.

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

Patent novelty and patent value

[Figure 2 about here.]

As argued above, the goal of science is to advance knowledge by making new discoveries. This why we study next whether the value of a patent is related to the novelty of its content. For this purpose, we construct a new measure of patent novelty. Using this new measure, we establish as the second fact that patent novelty predicts patent values.

In the history of technology and innovation, inventions are often conceptualized as the outcome of successfully combining ideas, either by combining new ideas or resources or by combining 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). Economist Joseph Schumpeter defines the essence of enterprise and entrepreneurship to be "the carrying out of new combinations." In *The Theory of Economic Development,* he writes that: "As a rule, the new combinations must draw the necessary means of production from some old combinations . . . development consists primarily in employing existing resources in a different way, in doing new things with them" (Schumpeter, 1934).

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. In contrast, the word "mouse" was combined with the word "display" for the first time in 1981 in the pioneering patents of Xerox.

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 every patent are taken from the dataset of Arts et al. (Arts *et al.*, 2018). 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 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 are the pairwise word combinations used in the particular patent. We call patents with a smaller average probability more novel.

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. Panel (b) indicates that a higher patent novelty is associated with an upward shift in the patent value distribution. For this purpose, 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. Figure 2, Panel (b) shows that more novel patents (i.e., patents with a low probability of 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.

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 to control for technology-specific effects. Again, there is a clear positive relationship between novelty and science value. In Panel (d), we show the distribution of the sum of the science and the idosyncratic value component across the percentiles of the value distribution for patents with higher or lower novelty; i.e., patents with below average 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 the 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 science, novelty is associated with a value premium over and above the technology-related value component, but also with higher risk. "Newness is not, by itself, an economic advantage" (Kline and Rosenberg, 1986).

[Figure 3 about here.]

A comparison of Figures 1 and 2 reveals that the patterns displayed are strikingly similar. While not being conclusive, this suggests that the science content of a patent and the novelty of the patent content are related. Consistent with this intuition, we establish as a third fact that 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 is the patent. Panel (a) 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 more science-intensive patents contain more novel word combinations; i.e., they are more novel. In Appendix C.5, we show that these findings are robust to using the emergence of new words and the average age of words as alternative novelty indicators.

Patent novelty and distance to science

As argued above, there are two complementary ways in which science can increase patent novelty. First, science might 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).

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 Poincaré, "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 tell which combinations not to pursue by providing an understanding of why a 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 to synthesize 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).

To operationalize these two views of how science might help the development of new technologies, we distinguish how novel the word combinations in a patent are from how novel the single words in a given patent are. To classify words in a patent as more or less novel, we calculate for each word the ratio of how often a particular word has been used before relative to the total number of words used. We call a word more novel the smaller this ratio; i.e., the smaller the probability a particular word has been used before. Then we take the 10th percentile of this probability distribution of word usage within a patent to arrive at a single measure of word novelty. If we find that science-intensive patents use new words to create new combinations, this will support the view that science provides new building blocks for the invention. In contrast, if science-intensive patents use novel combinations of common words, this will be in line with science guiding the inventor to valuable combinations of known elements.

Panel (c) of Figure 3 shows for each patent the novelty of the pairwise word combinations on the horizontal axis and the measure for word novelty on the vertical axis. We distinguish areas with a large share of science-intensive patents (red - dark shaded), an intermediate share of science-intensive patents (green - medium shaded) and a low share of science-intensive patents (grey - light shaded). Without adjusting for differences in technologies, science-intensive patents do not tend to use more novel words than non-science-intensive patents. If we adjust for technology and year, we find that a larger share of patents with novel words are science-intensive, as shown in Figure 3, Panel (d). This is in line with the intuition that science introduces new concepts that are combined successfully into novel combinations to develop new inventions.

Conclusion

Our study shows that science adds value to the private sector on a broad scale. This is far from obvious, given the public good nature of scientific knowledge and the skepticism expressed in several studies about the effectiveness of the knowledge transfer from university to industry and about the reliability of academic science (Goozner, 2005; Butler, 2008; Freedman *et al.*, 2015; Bikard, 2018). Understanding how much value science creates for society is fundamental for the case of public science funding. By illuminating the commercial value of science, our study provides a lower bound for its total value for society. Extrapolated to the U.S. economy, we find this lower bound for the additional value created by science for marketplace inventions to be 720 U.S. dollars per capita and year. This is about 25% of the total value of patented inventions in the U.S. (Appendix D). 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.

References

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

Arora, A., Belenzon, S. and Sheer, L. (2017). *Back to Basics: Why do Firms Invest in Research?* Tech. rep., National Bureau of Economic Research.

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

Azoulay, P., Zivin, J. S. G., Li, D. and Sampat, B. N. (2015). *Public R&D investments and private-sector patenting: evidence from NIH funding rules*. Tech. rep., National Bureau of Economic Research.

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

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.

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.

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.

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*.

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.

KELLY, B., PAPANIKOLAOU, D., SERU, A. and TADDY, M. (2018). *Measuring technological innovation over the long run*. Tech. rep., Working Paper.

KERR, W. R. (2008). Ethnic scientific communities and international technology diffusion. *The Review of Economics and Statistics*, **90** (3), 518–537.

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.

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

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yug, A. Z. and Fleming, L. (2014). Disambiguation and co-authorship networks of the u.s. patent inventor database (1975-2010). *Research Policy*, **43** (6), 941–955.

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.

Marx, M. and Fuegi, A. (2019). Reliance on science in patenting, mimeo.

Mokyr, J. (2016). *A culture of growth: the origins of the modern economy*. Princeton University Press.

Narin, F., Hamilton, K. S. and Olivastro, D. (1997). The increasing linkage between us technology and public science. *Research policy*, **26** (3), 317–330.

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

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.

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

Schumpeter, J. A. (1934). *Theory of economic development*. Harvard University Press.

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.

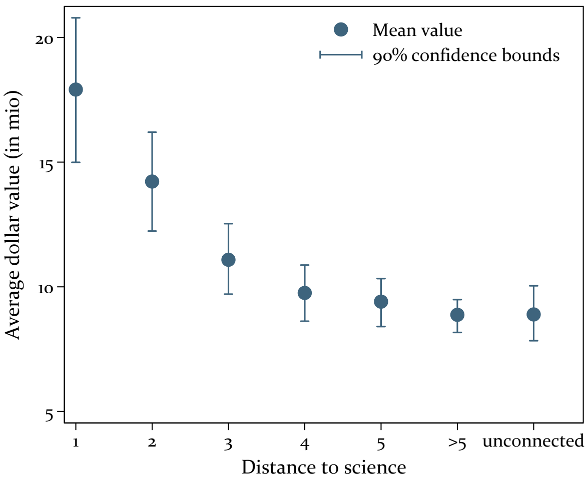
Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

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.

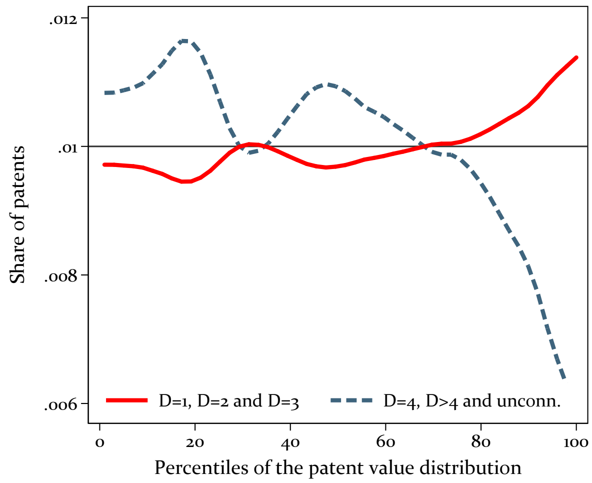
Uzzi, B., Mukherjee, S., Stringer, M. and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, **342** (6157), 468–472.

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

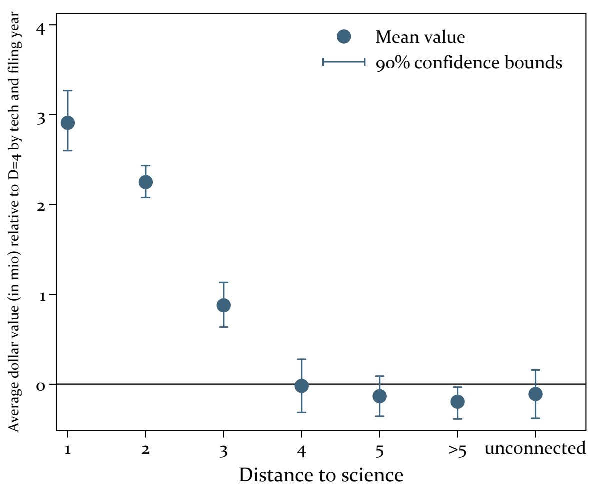
Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, pp. 331–360.



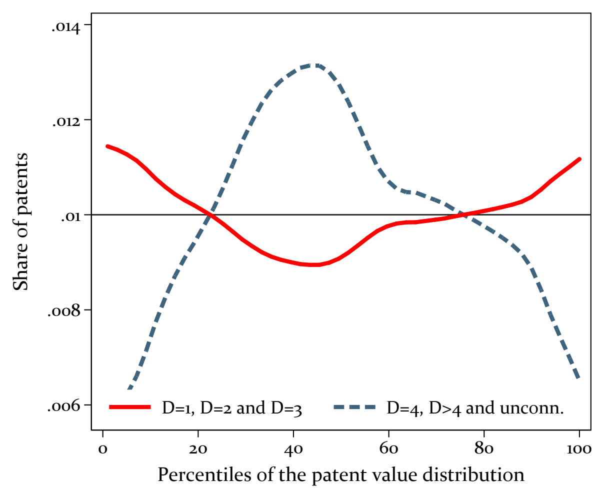
(a) Averages



(b) Distribution



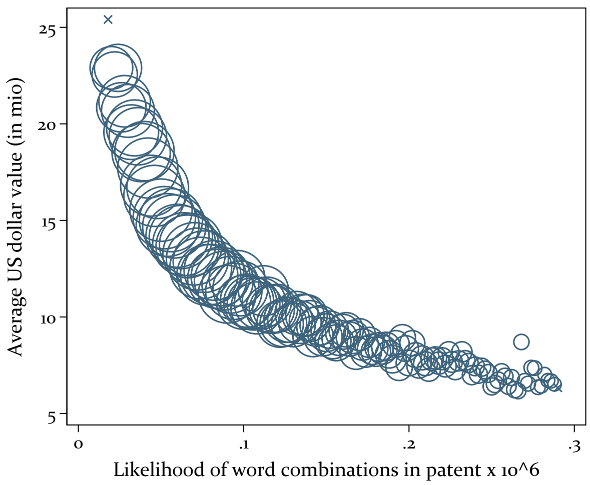
(c) Averages accounting for technology and year



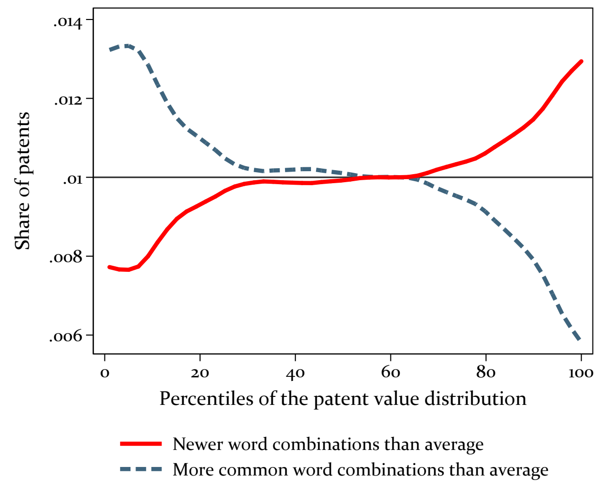
(d) Distribution accounting for technology and year

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

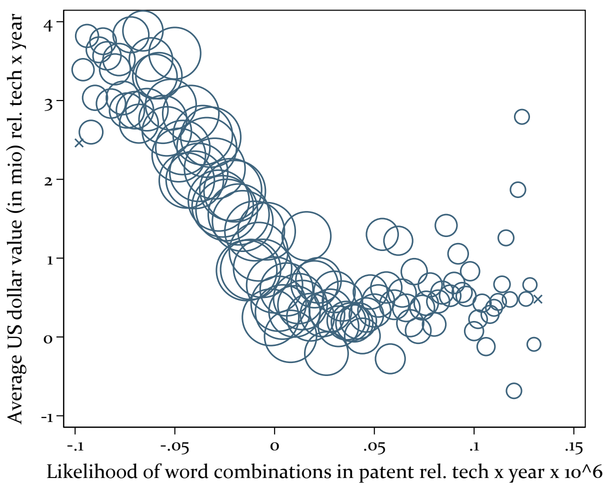
Panel (a) shows the average patent value for all distances to science. The values of U.S. patents are from Kogan et al. (Kogan *et al.*, 2017). The distance to science of U.S. patents is calculated with Microsoft Academic and Patstat using the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). The distance to science is defined by citation links. The 95% confidence bounds are based on standard errors bootstrapped by CPC technology class. 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 0.01 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 CPC technology class and filing year and a distance of four. In Panel (d), we show the distribution of normalized patent values.



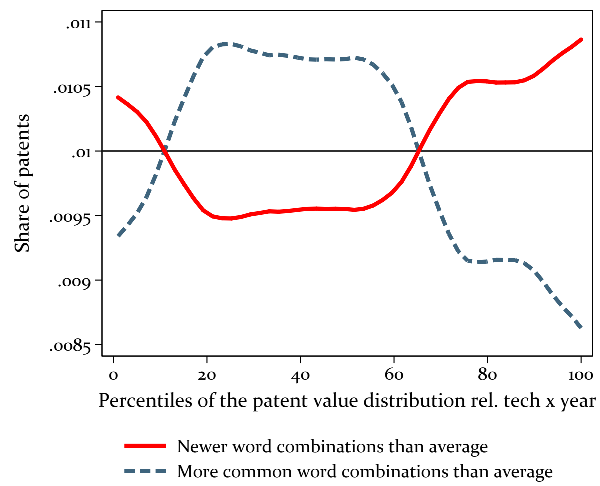
(a) Raw data



(b) Distribution



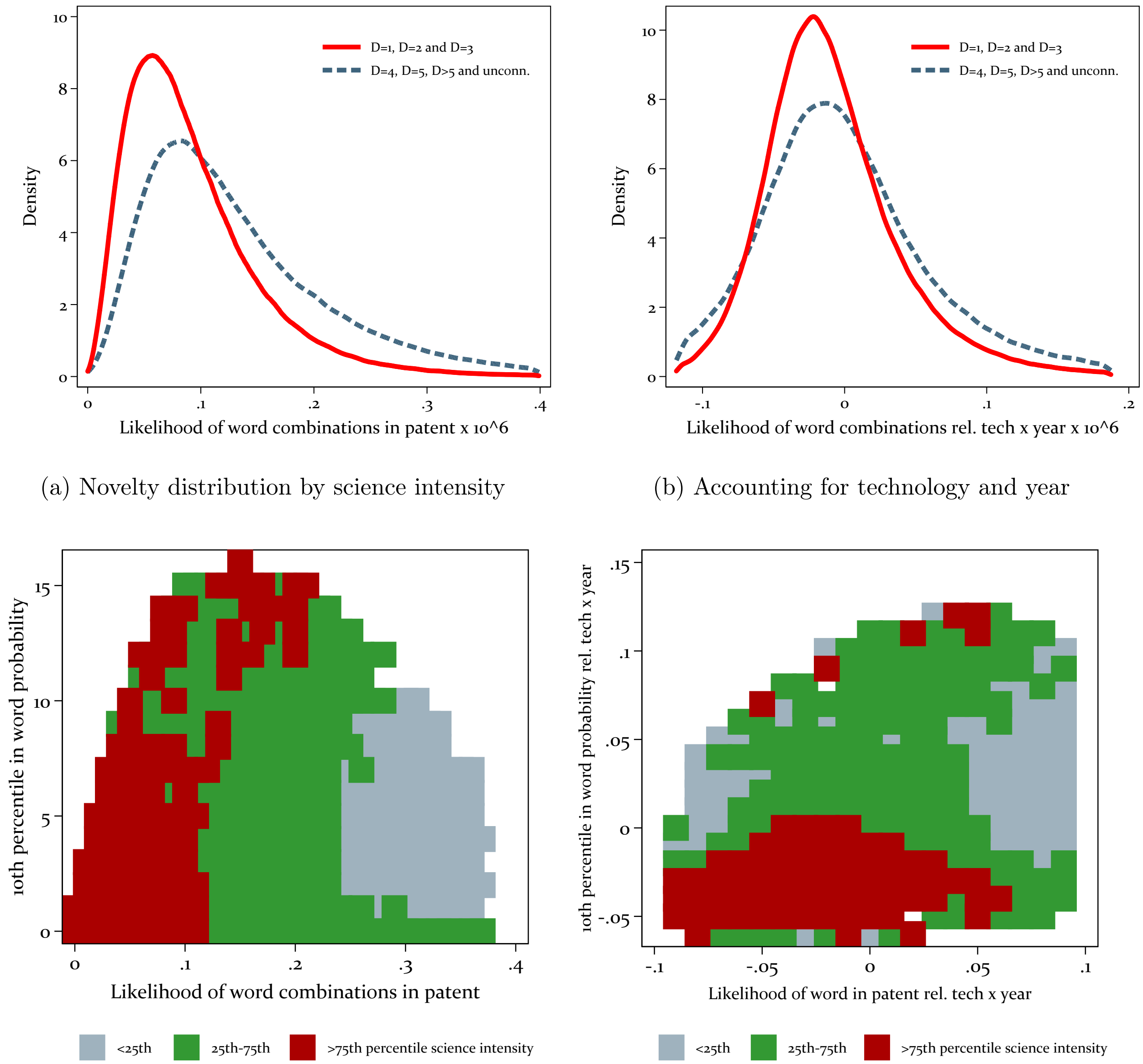
(c) Accounting for technology and year



(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 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

(c) Novelty and random matching benchmark

(d) 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. In Panel (c), we plot the average likelihood of word combinations (x-axis) along with the 10th percentile of the word novelty in a patent. We use different colors to indicate which share of patents is science-intensive. We distinguish areas with a large share of science-intensive patents (>75th percentile, red - dark shaded), an intermediate share of science-intensive patents (25th-50th percentile, green - medium shaded) and a low share of science-intensive patents (<25th percentile, grey - light shaded). We only keep data points with at least 100 patents. In Panel (d), we adjust the likelihood of word combinations and the word probability by technology class and filing year.

Supplementary material (online only)

A Data

For our analysis, we calculate distance to science for each patent following the method of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). We then match this data with patent values calculated by Kogan et al. (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.

Distance to science:

[Figure 4 about here.]

Ahmadpoor and Jones (Ahmadpoor and Jones, 2017) define a patent's distance to science using citation links.[5] 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. (Arora *et al.*, 2017) and Azoulay et al. (Azoulay *et al.*, 2015).[6] A patent that cites a (D=1)-patent but no scientific article has a distance of two (D=2), and so on (Figure 4). Citing another patent that is 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 (2019), 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 Marx

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

[6] 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.

and Fuegi (2019) with the values calculated by Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). In cases where Ahmadpoor and Jones (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. (Kogan *et al.*, 2017). Kogan et al. (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. (Arts *et al.*, 2018). Arts et al. (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

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/

- **All other patent characteristics** are from Patstat

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

B Results in Regression Form

We investigate the relationship between patent values and distance to science using regression methods. Results are presented in Table 1. We find that a patent that directly cites an academic article (D=1) has an average value of 17.8 million dollars (column 1). This value decreases with distance to science. Patents with a distance of two have on average a value of 14.1 million dollars while patents with a distance larger than three have a value between 8 and 11 million dollars. Patents that are completely unconnected to science have a value of 8.0 million dollars. The higher average value of science-intensive patents reflects an upward shift in the value distribution of patents with higher science intensity. This is shown in regression form in Table 2. In columns (1) and (2), we use the probability of a patent having a value in the top and in the bottom 5% of all patent values as outcomes. Patents that are closer to science have a higher likelihood to be in the top 5% and a lower likelihood to be in the bottom 5%.

In column 2 of Table 1, we present the average science component of the patent value by distance to science. In the regression, we use technology class × filing year fixed effects and leave out the dummy for (D=4)-patents. Patents that are directly based on science (D=1) have an average science-value component of $ 2.9 million dollars. Patents that are indirectly based on science (D=2 and D=3) have a science-value component of $ 2.2 and $ 0.9 million dollars, respectively. Columns 3 and 4 in Table 1 show that the pattern is robust to using forward citations or patent scope as alternative measures of patent value.

Additive separability of the technological and the science components is the simplest one among many potential assumptions about the relationship of science and non-science contributions to patent value. One alternative could be to assume multiplicative separability instead. In column 5 of Table 1, we present estimates using the logarithm of value as an outcome variable. The estimates show the same pattern, with (D=1)-patents more valuable than (D=2) and (D=3)-patents, which in turn are more valuable than a patent of distance of four.

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 Table 2, we use 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.

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.[7]

The results presented in column 6 of Table 1 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 7 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.

Why are patents based on scientific articles more valuable? Innovation is often thought of as the new combination of ideas. To judge the novelty of an innovation protected by a patent we look at pairwise combinations of words that occur in a particular patent. The words for every patent are taken from Arts et al (Arts *et al.*, 2018). We first count how

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

often a particular pairwise word combination was used in previous patents up to the filing year. We normalize the count of each word combination with the total number of word combinations up to the filing year of the patent. We denote this number as the probability of a particular word combination used in a particular patent. In a second step, we take the average over the probabilities of all word combinations within a patent to calculate the average probability per patent.

Column 8 of Table 1 shows the correlation between our patent novelty indicator and the average dollar value. In column 9, we control for technology $\times$ year fixed effect.[8] In both specifications, higher novelty measured by a lower likelihood of word combinations increases the patent value. If we do not control for technology and year, the effect is again driven by an upward shift in the whole distribution. Columns 3 and 4 of Table 2 show that more novel patents are more likely to be in the tails of the value distribution. Yet, once we look at the science value component we find that closeness to science increases both, the likelihood of a patent being in the top 5% and the bottom 5% of the distribution (Columns 5 and 6).

Columns 10 and 11 of Table 1 show that patents that are closer to science are also more novel. In Column 10, we show the raw correlation. Patents with a distance of one have a lower average likelihood of word combinations than patents with a higher distance. In Column 11, we control for average novelty in technology and year of patents with a distance of four. Again, patents closer to science are more novel. In Columns 7 and 8 of Table 2, we use the likelihood that a patent is in the top 5% or bottom 5% of the distribution of the science-value component as an outcome. Patents that are more novel have a higher likelihood to be in the tails of this distribution. So, relative to the technological component, more novel patents show a larger variance in values.

In Table 3, we show the share of science patents for different levels of average likelihood of word combinations and the 10th percentile of the word novelty in a patent in order to replicate the results in Panels C and D of Figure 3. To simplify the exposition we use dummy

[8]Note that 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.

variables to indicate above and below median values of the two independent variables. The highest share of patents with a distance of science smaller than four ("science-intensive") are among patents with a below median likelihood of word combinations and a below median 10th percentile of the word novelty in a patent (Column 1). This finding is robust if we look at the share of patents with a distance of one (Column 2) and if we control for technology and year (Columns 3 and 4).

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

C Alternative data and method choices

In this section, we discuss the results using alternative data and methodological choices. In the next subsection we use the alternative distance to science measures of Ahmadpoor and Jones (Ahmadpoor and Jones, 2017), and in subsection 2 we use different normalizations to account for year and technology effects. In Subsection 3, we show the effect over the whole value distribution. In subsection 4 we split the results by technology, and in the last subsection we discuss word age as an alternative measure for novelty.

C.1 Using Ahmadpoor and Jones distance values

In Figures 5a and 5b, we use the distance-to-science measure based on the data of Ahmadpoor and Jones (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 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 with distance to science in absolute and normalized values.

[Figure 5 about here.]

C.2 Different normalizations and assignee-fixed effects

In the main part of the paper, we normalize the patent values by the value of patents with a distance of 4 with the same technology class and filing year. One might ask whether patents with a distance of D=4 are the right comparison group. As a robustness check, we use unconnected patents as a control group and present results in Figure 6a. The quantitative magnitudes of the effects are similar to our main specification. As a further alternative, we normalize using USPC instead of CPC technology classes and present results in Figure 6b. The results are the same.

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. If this were the case, our result might be driven by assignee-fixed effects. Comparing patent values within assignee is difficult to do with the Kogan et al. (Kogan *et al.*, 2017) data. The reason is that all patents published by the same assignee at the same date have the same value as their evaluation is based on the same abnormal stock market returns. This is why for this exercise we use forward citations as an outcome variable. In Figure 6c, we normalize the number of forward citations by each combination of assignee, filing year and technology class. This implies that we look only at citation differences within the same assignee. We find the same pattern as in our main result.

[Figure 6 about here.]

C.3 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 CPC technology classification with a distance of four. This pattern is already visible in the raw data in Figure 7a. In Figure 7b 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.[9] The patent values are falling with distance to science over all percentiles. This confirms that our results are not driven by outliers.

[Figure 7 about here.]

C.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 8, 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.

[Figure 8 about here.]

[9] 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.

C.5 Alternative measures for novelty

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 9a and 9b, 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 9a 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 9b). 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 9c compares the patent value for patents with and without new words for each year. Patents with new words are more valuable throughout the whole sample. Figure 9d 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.

[Figure 9 about here.]

D Macroeconomic extrapolation

[Figure 10 about here.]

To understand the overall impact of science on private sector innovation we need to gauge how much science contributes directly or indirectly to the patent value in the economy. In Figure 10, we show the direct and indirect impact of science for all of the U.S. and for each state separately. For this exercise, we use in the first row the science value component for the filing years 2000 to 2005, sum them up and scale them by the U.S. population. The direct value of science is $ 203 dollars per person and year; i.e., 7% of the total patent value of $ 2727 dollars per person and year. The indirect value of science - i.e., the science value component of all patents with a distance to science larger than one - is $ 519 dollars per person and year. This means that the indirect effect is more than twice the size of the direct effect and accounts for around 70% of the overall effect of science, which is $ 722. In total, science contributes 25% of the overall U.S. patent value per person.

This extrapolation is based on some strong assumptions. First, we assume that without science, the science-intensive patents would have the average value of a patent in the same technology and with the same filing year with a distance of four. Second, our estimates in the first row neglect the value of patents of companies that are not publicly traded. In row two and the following rows, we include an extrapolation for values of patents of companies that are not publicly traded based on technology class, filing year and distance to science. However, these patents might be more or less valuable than patents of public companies. Third, these estimates only include the private value to companies, but not the additional value of innovation to consumers that is not reflected in higher profits for the innovating companies. For these reasons, our estimate is most likely an approximation of the lower bound of the value of science in the economy.

The benefits of science and innovation are not equally distributed across states (Figure 10). We use the geolocated patent data of Li et al (Li *et al.*, 2014) to assign patents to states along with the imputed patent value data. The states are ordered by the total science value per capita. California, New Jersey, and Massachusetts stand out in terms of overall value generated by patents. In terms of value created directly and indirectly by science, it is California and New Jersey that come out top, but science also plays a large role in Massachusetts and Texas. In contrast, North Carolina and Florida derive little value from science.

In Figure 11a, we explore further the heterogeneity of the science value component over time and across industries. In Panel (a), we show the total science value component of (D=1), (D=2) and (D=3)-patents ("direct & indirect science value") along with the patent value not related to science per person and year of all U.S. patents filed in the U.S. between 1990 and 2005. To assess the economy-wide impact we extrapolate patent values by distance to science, technology class and filing year to all patents assigned to companies in the economy.

In each year, the science component is around 25% of the total patent value per person. The share of the science component in the overall patent value is largest in the years of the dotcom boom when tech companies like Google developed their first patents. All dollars are deflated to 1982 in accordance with the deflation used by Kogan et al. (Kogan *et al.*, 2017). To put these numbers in perspective, Panel (b) shows the total funding for R&D in the U.S. and the share funded by universities per person. This data is from the National Science Foundation. Total R&D funding was \$ 686 dollars per person in 2005, while university funding was \$ 95 dollars per person. In 2005, the science value component was \$ 1060 dollars per person and the total patent value \$ 4083 dollars per person and year.

The benefits of science and innovation, in general, are not equally distributed across industries. To assign a patent to an SIC code, we use the data of Kerr (Kerr, 2008).[10] Panel 11c shows the average total value of patents by two-digit SIC code over the years 2000 to 2005. Both in terms of total patent values and value derived from science industrial machinery is leading. On a per patent basis, chemicals derive the largest value from science (Figure 11d).

[Figure 11 about here.]

[10] We thank Bill Kerr for sharing his data.

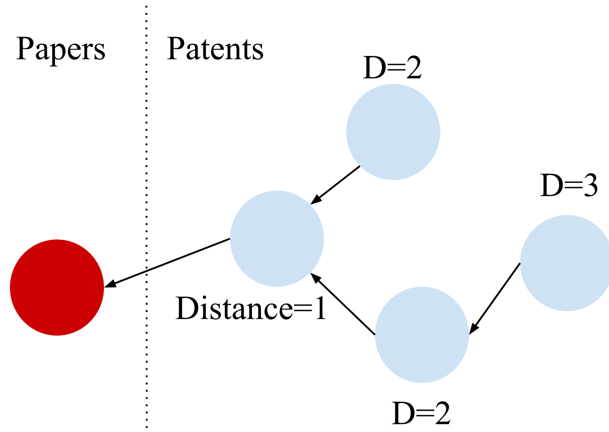
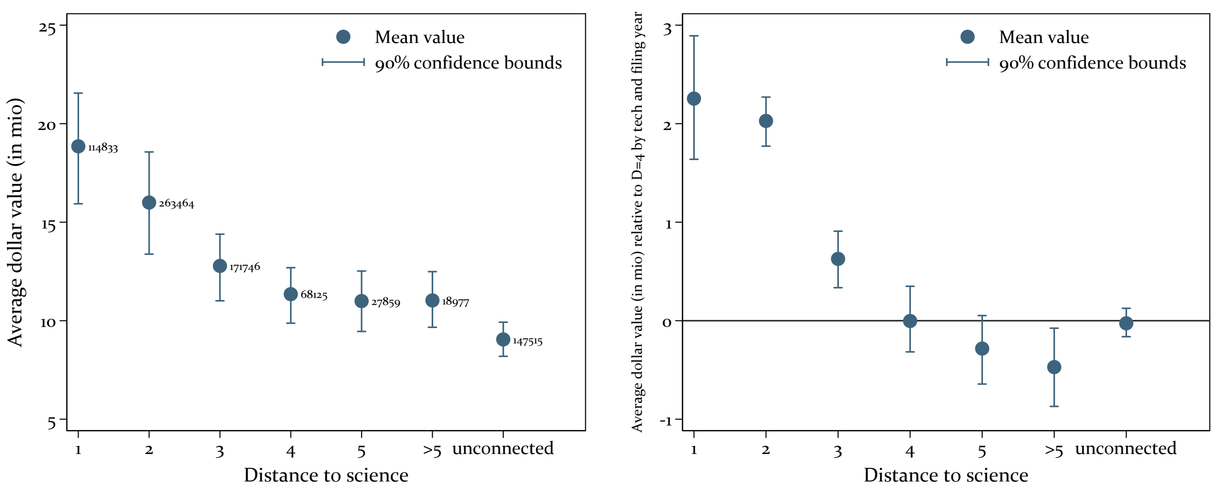


Figure 4: Distance to science

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

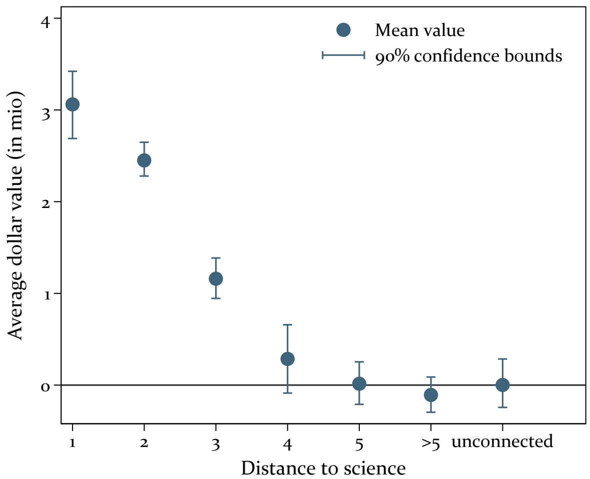


(a) Ahmadpoor and Jones (Ahmadpoor and Jones, 2017) distance data

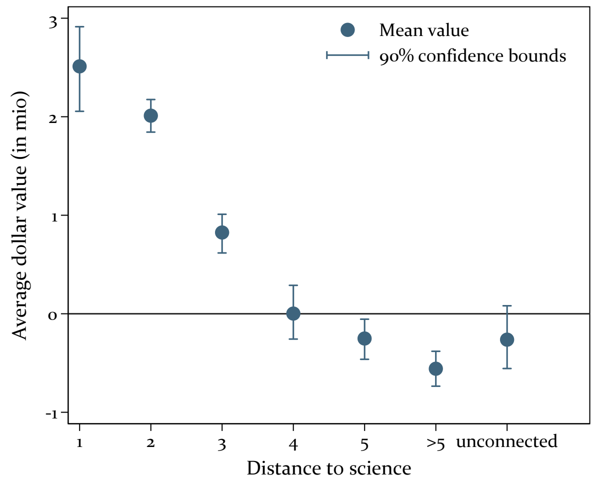
(b) Ahmadpoor and Jones (Ahmadpoor and Jones, 2017) distance data normalized

Figure 5: Value of patents by distance to science

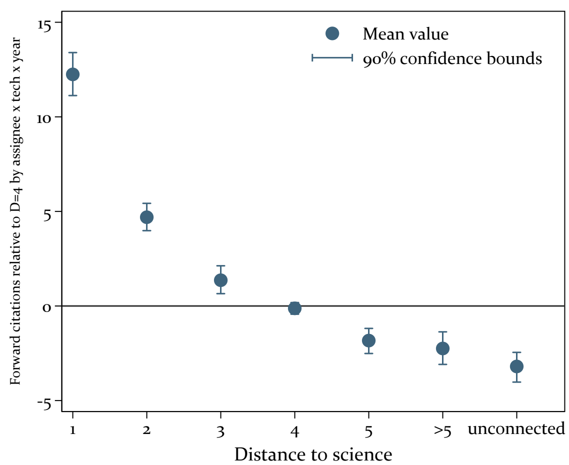
Panel (a) shows the raw data for patent values by distance to science data for a 10% sample of patents. The values of U.S. patents are from Kogan et al. (Kogan *et al.*, 2017) and the distance-to-science data are from Ahmadpoor and Jones (Ahmadpoor and Jones, 2017). In Panel (b), we normalize the patent value by the average value of a patent with the same CPC technology class and filing year and a distance of four.



(a) Relative to unconnected



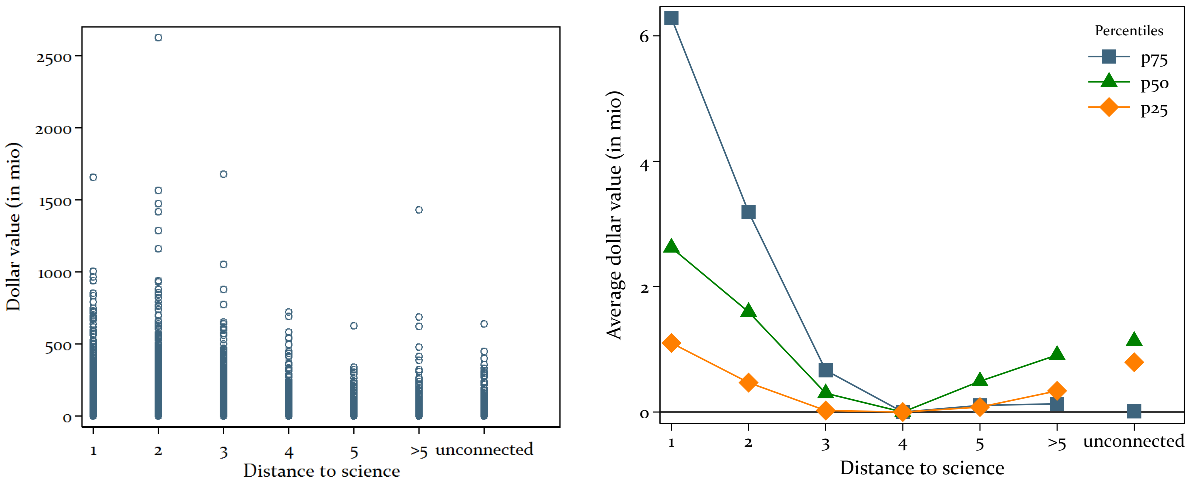
(b) Based on USPC instead of CPC-technology class



(c) Forward citations normalized by assignee

Figure 6: Different normalizations and assignee-fixed effects

In Panel (a), we show the average value of patents normalized by the value of patents that were filed in the same year but are unconnected to science. In Panel (b), instead of the value of patents in the same filing year and CPC patent classes we use patents in the same filing year and the same USPC patent classes. In Panel (c), we use forward citations instead of patent values as the outcome and adjust for assignee-fixed effects.

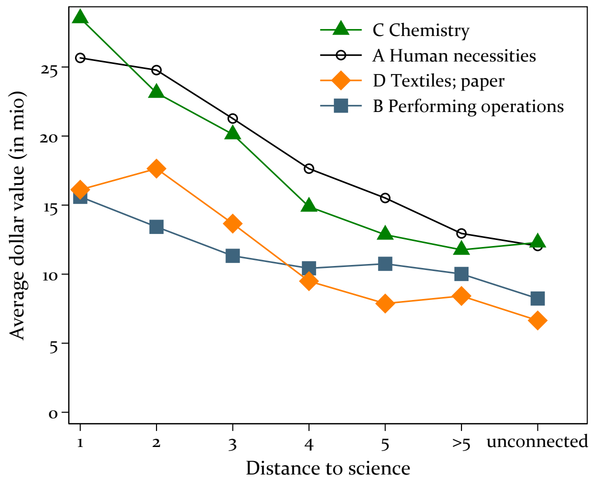


(a) Raw data

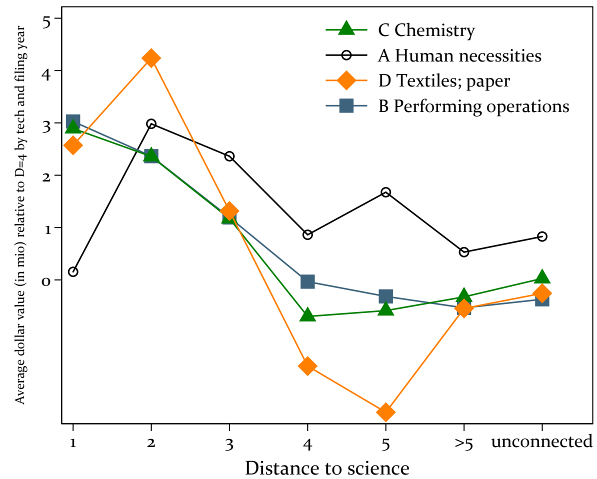
(b) Value percentiles residualized with same percentile in D=4

Figure 7: Effects over the value distribution

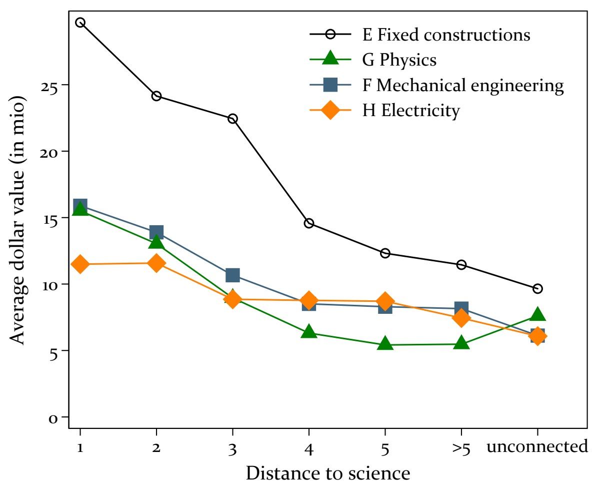
Panel (a) shows the raw data for patent values by distance to science data for a 10% sample of patents. The values of U.S. patents are from Kogan et al (Kogan *et al.*, 2017). The distance to science of U.S. patents is calculated with Marx and Fuegi Marx and Fuegi (2019) and Patstat using the method 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 4.



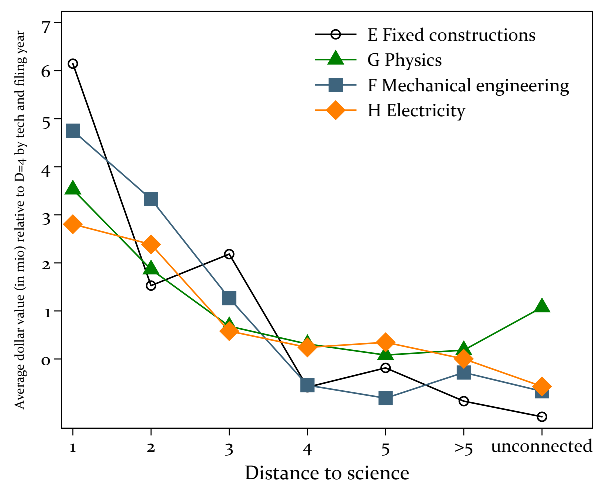
(a) Value by broad patent class A-D



(b) Accounting for technology× year



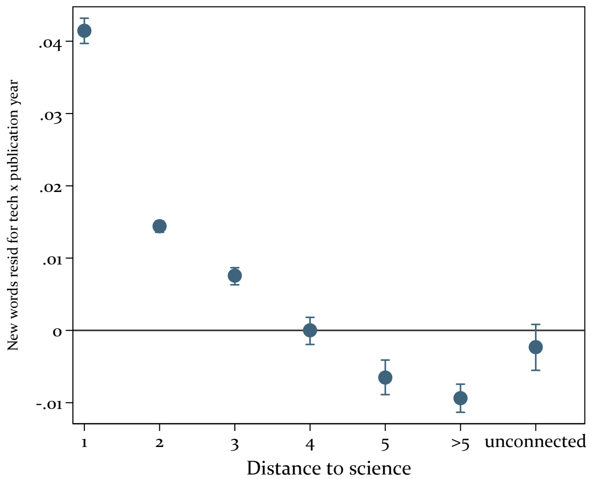
(c) Value by broad patent class E-G



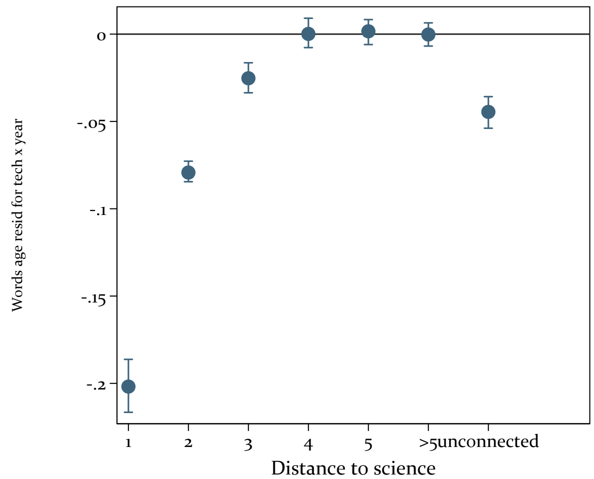
(d) Accounting for technology × year

Figure 8: Sample splits by technology

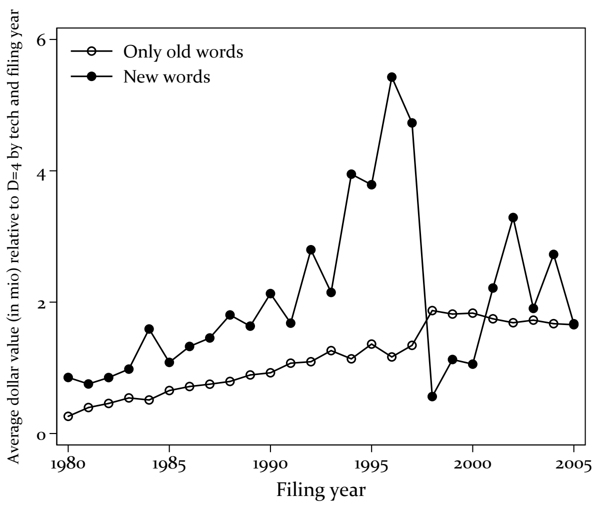
In this figure, we split patents by broad technology fields measured by one-digit CPC classes. In Panels A and C we show the raw data. In Panels B and D we normalize by the average patent value of a patent in the same four-digit CPC class and filing year with a distance of 4.



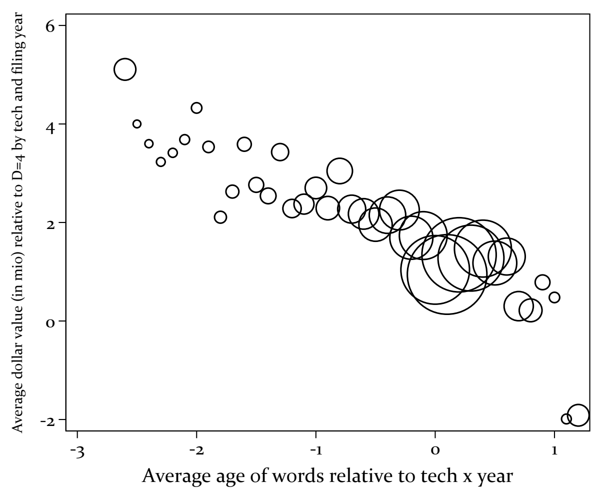
(a) Novelty: New words



(b) Novelty: Average word age



(c) New words and value



(d) Word age and value

Figure 9: 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 dollar value of a patent relative to a patent with the same filing year and CPC technology class separately for patents with and without a new word 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 CPC technology class.

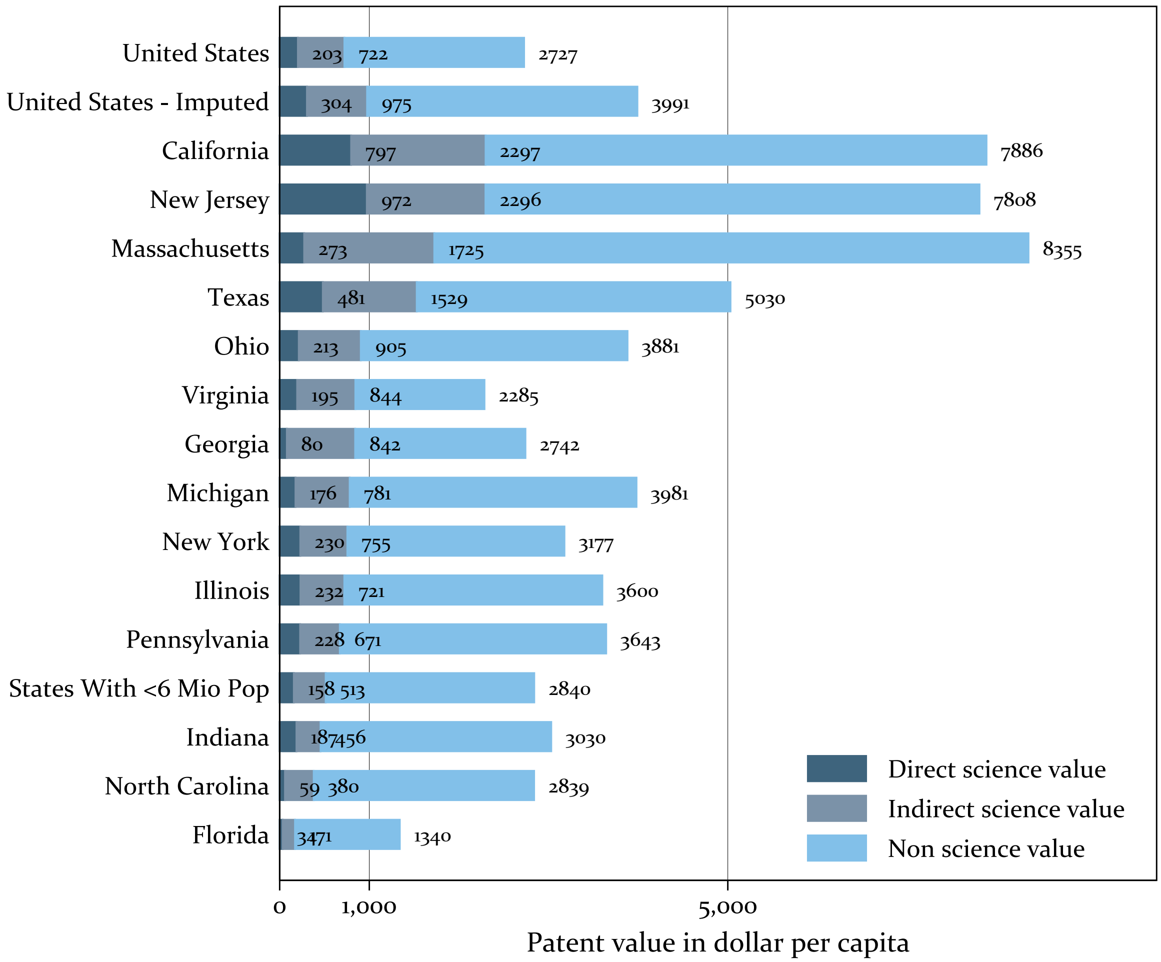
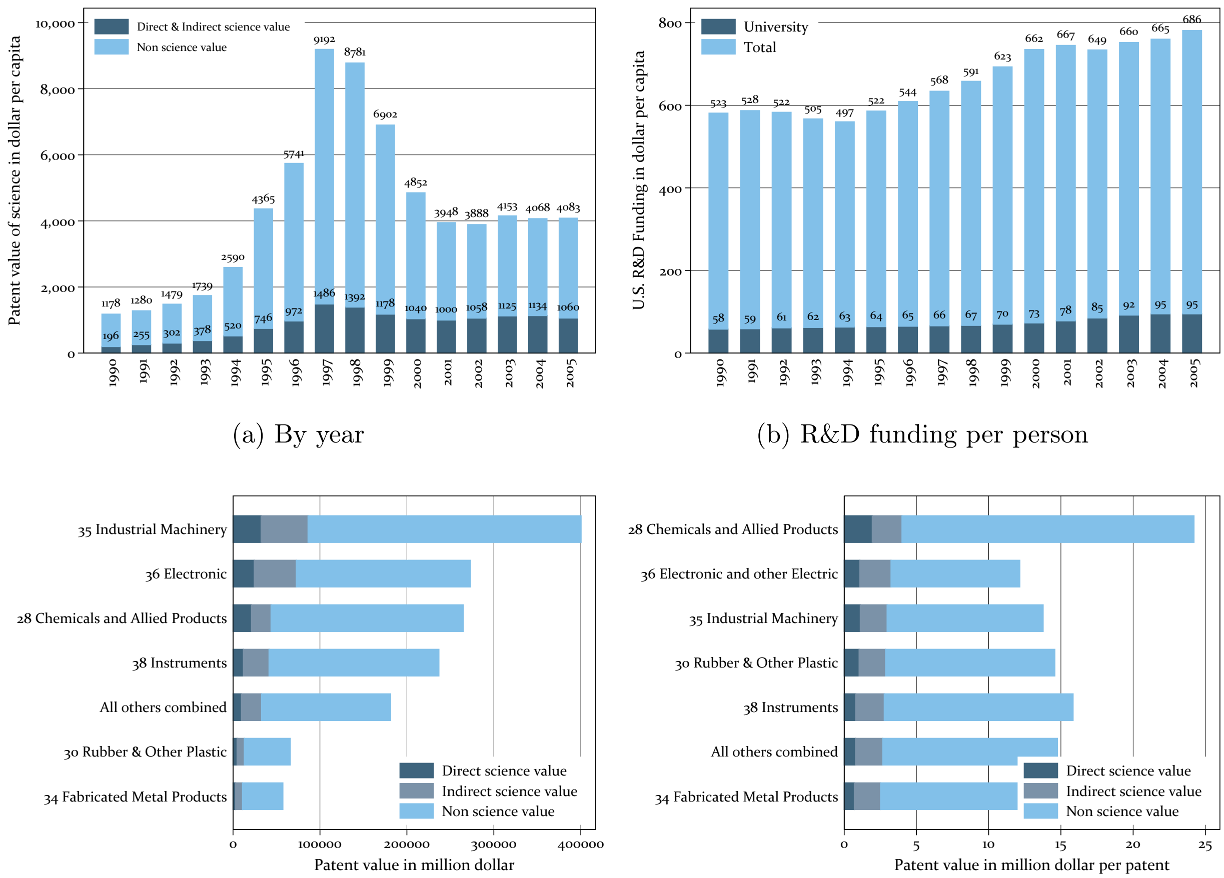


Figure 10: The science value component across states - 2000 to 2005 average

This figure shows the value of patents per person and year in the U.S. (rows 1 and 2) and across states. We plot separately the implied direct and the implied indirect value of science, as well as the value that is not related to science. For each patent, we derive the value attributable to science by subtracting the average value of a patent with the same filing year and the same CPC technology class with a distance equal to four from the patent value. If the patent under consideration has a distance to science of one, we call the difference "Direct science value component." If the distance to science is larger than one we call the difference "Indirect science value component." The "nonscience value component" is the remainder of the patent value after subtracting the direct and the indirect science value. We divide the aggregated value by the number of persons in the U.S. and the different states in 2000 to calculate per capita values.
In the first row, we use only the patent values of Kogan et al. (Kogan *et al.*, 2017) and aggregate to the whole United States. In the second row, we use the data of Kogan et al. (Kogan *et al.*, 2017) and add imputed values for those patents for which no Kogan et al. (Kogan *et al.*, 2017) data is available. We impute values for patents assigned to companies by CPC technology class and filing year. In the following rows, we split patents by the state of the inventor. The state of the inventor is from Li et al. (Li *et al.*, 2014). If the inventors on a patent are located in multiple states we assign a share of the value to each state. To ease presentation, we aggregate the 36 states with a population of less than 6 million in 2000 to one ("States with $<$6 Mio Pop").



(a) By year

(b) R&D funding per person

(c) 2-digit SIC industry- total

(d) 2-digit SIC industry- per patent

Figure 11: The science value component over time

Notes: Panel (a) shows the value of U.S. patents per capita over time. For each patent, we derive the value attributable to science by subtracting the average value of a patent with the same filing year and the same CPC technology class with a distance of four from the patent value. If the patent under consideration has a distance to science of one, we call the difference “Direct science value component” while if the patent has a distance larger than one we call the difference “Indirect science value component.” “Nonscience value component” is the remainder of the patent value after subtracting the direct and the indirect science value components. We divide the aggregated value by the number of persons in the U.S. and the different states in 2000 to calculate per capita values. We use imputed values for patents for which no Kogan et al. (Kogan *et al.*, 2017) data is available. We impute values for patents assigned to companies by CPC technology class and filing year. Panel (b) shows the average university and total R&D funding per person and year. The data is from the NSF and deflated to constant 1982 dollars, as in Kogan et al. (Kogan *et al.*, 2017). In Panel (c), we split patents by industry and aggregate the values up for the year. The figure gives the total value per industry averaged over the years 2000 to 2005. To split the value of each patent by industry, we use the data of Kerr (Kerr, 2008), which gives a probabilistic transition table from USPC technology classes to SIC codes. In Panel (d), we plot the average value per patent and per industry.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Patent Value					Text similarity		Patent novelty			
Outcome:	Dollar	Dollar	Citations	Patent Scope	ln(Dollar)	Text sim.	Dollar	Dollar	Dollar	Prob. of word combination	Prob. of word combination
Distance 1	17.8***	2.9***	19.5***	12.3***	0.8***	7.6***				8.0***	-1.7***
	(1.9)	(0.4)	(1.3)	(3.0)	(0.1)	(0.3)				(0.4)	(0.2)
Distance 2	14.1***	2.2***	7.2***	10.4***	0.6***	5.1***				9.3***	-0.8***
	(1.3)	(0.3)	(0.5)	(2.0)	(0.0)	(0.2)				(0.3)	(0.1)
Distance 3	11.0***	0.9***	2.0***	7.3***	0.2***	2.6***				10.4***	-0.2***
	(0.9)	(0.2)	(0.4)	(1.2)	(0.0)	(0.2)				(0.2)	(0.1)
Distance 4	9.7***									11.5***	
	(0.7)									(0.2)	
Distance 5	9.4***	-0.1	-1.1***	-3.7**	-0.0	0.2**				12.9***	0.1**
	(0.6)	(0.2)	(0.3)	(1.6)	(0.0)	(0.1)				(0.2)	(0.1)
Distance >5	8.9***	-0.2	-2.3***	-5.4***	0.0	-0.2				15.1***	0.6***
	(0.4)	(0.2)	(0.3)	(2.1)	(0.0)	(0.1)				(0.3)	(0.1)
Unconnected	8.9***	-0.1	-3.1***	-3.9	-0.0					15.8***	0.5***
	(0.7)	(0.2)	(0.6)	(8.0)	(0.0)					(0.5)	(0.1)
Probability of word combinations								-49.1***	-11.9***		
								(4.5)	(1.7)		
Text similarity							3.4***				
							(1.0)				
Tech x Year FE	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Mean Dep.	13.0	13.0	25.9	11.8	0.8	10.0	13.0	13.0	13.0	10.4	10.4
Obs.	1085180	1085180	1085180	226066	1085180	1084282	1084282	1085180	1085180	1085180	1085180

Table 1: Patent value and distance to science

This table shows the results from OLS regressions. 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 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 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 patent scope z-score within art unit provided by Kuhn and Thompson (Kuhn and Thompson, 2017) as outcome variable. In column 5, we take the natural logarithm of the patent value of Kogan et al. (Kogan *et al.*, 2017) as dependent variable. In column 6, 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 7, we use text similarity as independent variable. In column 8, we use a novelty indicator - the average probability of word 2-tuples of each patent - to explain the value of a patent. In column 9, we add technology x filing year fixed effects. In column 10, we use the novelty indicator as outcome variable and correlate it with the distance to science. In column 11, we add technology class x filing year fixed effects The standard errors are clustered on the CPC technology class level. *, ** and *** indicate that the coefficient is significantly different from zero on the 10%, 5%, and 1% level.

	(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***	3.2***			7.4***	8.7***		
	(1.2)	(0.5)			(0.8)	(2.0)		
Distance 2	5.5***	4.8***			5.6***	5.3***		
	(0.6)	(0.5)			(0.5)	(1.2)		
Distance 3	4.1***	6.5***			4.3***	3.8***		
	(0.4)	(0.6)			(0.3)	(0.7)		
Distance 4	3.4***	6.3***			3.4***	3.3***		
	(0.3)	(0.7)			(0.3)	(0.5)		
Distance 5	3.0***	5.6***			3.2***	2.9***		
	(0.3)	(0.9)			(0.2)	(0.4)		
Distance >5	2.4***	5.3***			2.6***	2.3***		
	(0.2)	(0.8)			(0.2)	(0.4)		
Unconnected	2.3***	3.7***			2.5***	2.4***		
	(0.3)	(0.5)			(0.3)	(0.5)		
Probability of word combinations			-26.8***	27.3***				
			(2.8)	(4.7)				
Probability of word combinations rel. tech x year							-7.2***	-3.0**
							(1.0)	(1.4)
Mean Dep.	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
Obs.	1085180	1085180	1085180	1085180	1085180	1085180	1085180	1085180

Table 2: 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. *, ** and *** indicate that the coefficient is significantly different from zero on the 10%, 5%, and 1% level.

	(1)	(2)	(3)	(4)
	Raw		Controlling Tech × Year	
Outcome:	Share D<4	Share D=1	Share D<4	Share D=1
Below median likelihood of word combinations and word probability	85	29	81	25
Below median likelihood of word combinations / above median word probability	79	17	72	20
Above median likelihood of word combinations / below median word probability	70	19	78	19
Above median likelihood of word combinations and word probability	65	11	69	14
Obs.	1085180	1085180	1085128	1085128

Table 3: Share of science-intensive patents by novelty

This table shows the average share of patents with a distance of science smaller than four (Columns 1 and 3) and with a distance of science of one (Columns 2 and 4) for different levels of average likelihood of word combinations and the 10th percentile of the word novelty in a patent. To construct dummies for the different levels we split the average likelihood of word combinations and the 10th percentile of the word novelty in a patent at their respective medians. In Columns 1 and 2 we show the raw data and in Column 3 and 4 we adjust the average likelihood of word combinations and the 10th percentile of the word novelty in a patent by technology and filing year.