NBER WORKING PAPER SERIES

THE EFFECT OF PUBLIC SCIENCE ON CORPORATE R&D

Ashish Arora
Sharon Belenzon
Larisa C. Cioaca
Lia Sheer
Hansen Zhang

Working Paper 31899
http://www.nber.org/papers/w31899

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

We gratefully acknowledge support from The Henry Crown Institute of Business Research, the Fuqua School of Business, Qualcomm, and the Sloan Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w31899

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2023 by Ashish Arora, Sharon Belenzon, Larisa C. Cioaca, Lia Sheer, and Hansen Zhang. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Effect of Public Science on Corporate R&D
Ashish Arora, Sharon Belenzon, Larisa C. Cioaca, Lia Sheer, and Hansen Zhang
NBER Working Paper No. 31899
November 2023
JEL No. O3

ABSTRACT

We study the relationships between corporate R&D and three components of public science: knowledge, human capital, and invention. We identify the relationships through firm-specific exposure to changes in federal agency R\&D budgets that are driven by the political composition of congressional appropriations subcommittees. Our results indicate that R&D by established firms, which account for more than three-quarters of business R&D, is affected by scientific knowledge produced by universities only when the latter is embodied in inventions or PhD scientists. Human capital trained by universities fosters innovation in firms. However, inventions from universities and public research institutes substitute for corporate inventions and reduce the demand for internal research by corporations, perhaps reflecting downstream competition from startups that commercialize university inventions. Moreover, abstract knowledge advances per se elicit little or no response. Our findings question the belief that public science represents a non-rival public good that feeds into corporate R&D through knowledge spillovers.

Ashish Arora
Fuqua School of Business
Duke University
Box 90120
Durham, NC 27708-0120
and NBER
ashish.arora@duke.edu

Sharon Belenzon
Fuqua School of Business
Duke University
100 Fuqua Drive
Durham, NC 27708
and NBER
sharon.belenzon@duke.edu

Larisa C. Cioaca
Fuqua School of Business
Duke University
100 Fuqua Drive
Durham 27708
larisa.cioaca@duke.edu

Lia Sheer
Tel-Aviv University
Ramat-Aviv, P.O. Box 39040
Tel Aviv 6997801
Israel
liasheer@tauex.tau.ac.il

Hansen Zhang
Fuqua School of Business
Duke University
100 Fuqua Drive
Durham, NC 27708
hansen.zhang@duke.edu

The DISCERN dataset used in the paper is available at https://zenodo.org/records/4320782

1 Introduction

The American innovation ecosystem features a division of labor between universities that perform the bulk of basic research, startups that identify commercial applications for discoveries, and large firms that develop and scale up the applications. Fueled by federal support, university research has grown steadily since World War II. Since the passage of the Bayh-Dole Act in 1980, American universities have also increasingly turned to patenting and commercializing their discoveries (Aldridge & Audretsch, 2017). Figure 1 shows that university publications, university patents, and PhD dissertations have increased considerably since 1981. Whereas publications have grown by about 75% and PhD production by over 100%, university patents have increased twenty-five-fold, albeit from a small base.

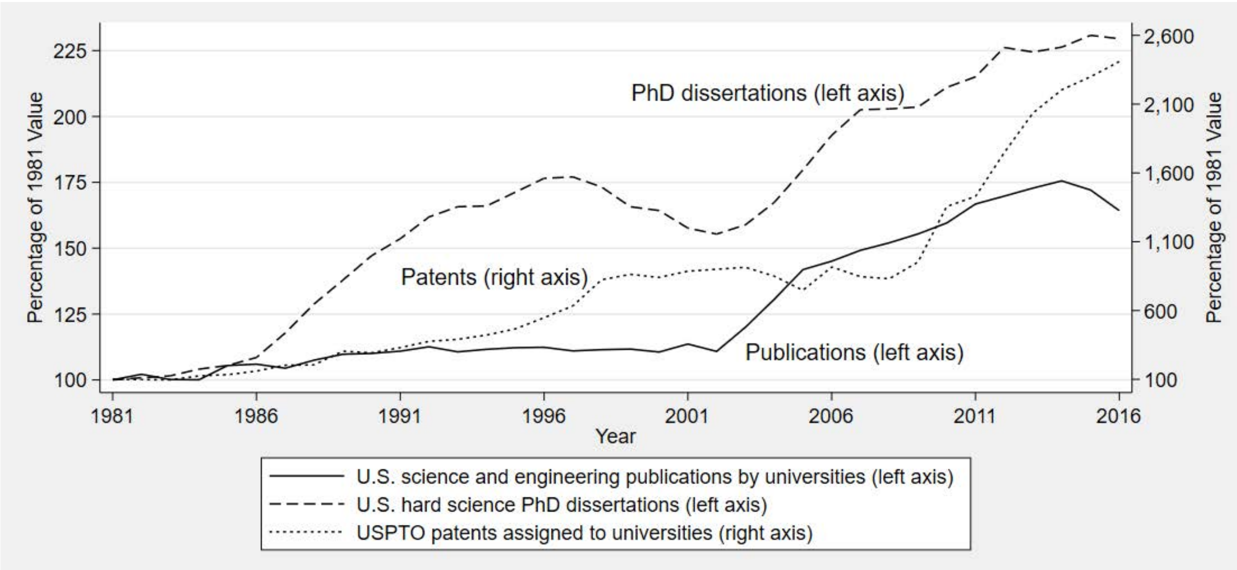
Over the same period, startups have also grown in importance as sources of new technology, while many large firms have withdrawn from upstream research (Arora, Belenzon, & Patacconi, 2018; Arora, Belenzon, & Sheer, 2021a; Mowery, 2009).[1] Though corporate laboratories such as Bell Labs, Xerox PARC, IBM, and DuPont are in decline, even today firms with more than 1,000 employees account for about 80% of business R&D investment (National Center for Science and Engineering Statistics, 2023b, table RD-12). Therefore, it is essential to understand how the growth of university research has affected innovation by established firms.

Doing so requires considering the different ways in which universities affect corporate R&D. In addition to producing scientific knowledge, universities also produce trained researchers (Schartinger, Rammer, Fischer, & Fröhlich, 2002), as well as inventions that can be used by startups or licensed to established firms. The impact of the different components of public science on corporate innovation can be complex (Cohen, Nelson, & Walsh, 2002). Moreover, corporate R&D itself has an upstream research component and a downstream development component, and these may respond differently to increases in public knowledge or increases in public invention. Our goal is to estimate how public science—scientific knowledge, human capital, and inventions from universities and other public research organizations—affects corporate R&D.

We develop a simple analytical framework to explore the relationships between corporate

[1] There are some notable exceptions to these broad trends. In several emerging technology fields, including artificial intelligence (AI) and quantum computing, leading companies such as IBM, Microsoft, and Google, continue to invest in upstream research. Some of the best-known AI researchers and quantum computing experts today work for corporations rather than universities (Hernandez & King, 2016). Corporate publications represented 10% of all publications at the International Conference on Machine Learning in 2004 and 30% in 2016 (Hartmann & Henkel, 2020). Based on data from Microsoft Academic Graph (Sinha et al., 2015; Wang et al., 2019), IBM and Microsoft produced more quantum computing publications than MIT during 2013-2020.

Figure 1: TRENDS IN UNIVERSITY SCIENCE, 1981-2016



Notes: This figure presents trends in university science over time, including U.S. science and engineering journal publications authored by university researchers (left axis), U.S. hard science PhD dissertations (left axis), and USPTO patents assigned to universities (right axis). All measures are normalized by their 1981 values. Publication data for 1981-1995 are from Appendix Table 5-44 of Science and Engineering Indicators 1998 (National Science Board, 1998). Publication data for 1995-2003 are from Appendix Table 5-42 of Science and Engineering Indicators 2010 (National Science Board, 2010). Publication data for 2003-2016 are from Appendix Table 5-41 of the Science and Engineering Indicators 2018 (National Science Board, 2018). Dissertation data are from ProQuest Dissertations & Theses Global, while patent data are from PatentsView.

R&D and three dimensions of public science: knowledge, human capital, and inventions. Corporate innovations can arise from *inventions* generated internally or acquired externally, particularly from universities. Scientific knowledge, both from *internal research* and *public knowledge* from universities, lowers the cost of internal invention.[2] *Human capital* is an input into both internal research and invention.

A firm's response to increased public science depends on three main factors. First, public knowledge can complement or substitute for internal research in reducing the marginal cost of internal inventions. Second, an increase in the supply of human capital reduces the cost of internal research and invention. Third, public inventions can substitute for internal inventions as inputs into the firm's innovations, thereby reducing the effective cost of innovation to the firm. Public inventions can also fuel market entry by startups, reducing the payoff to

[2]In a recent example of harnessing public knowledge to lower the cost of internal invention, Swiss pharmaceutical company Roche set up the Institute of Human Biology in May 2023 to enable its internal researchers to collaborate with academic researchers on exploratory research, bioengineering, and translational projects using organoids (Roche, 2023, May 4). This foundational research will not lead directly to the invention of new drugs, but will instead provide useful scientific knowledge that reduces the cost of invention (replacing animal models with organoids may better predict human responses to candidate drugs).

the focal firm's innovations. The effect of public science on the marginal returns to internal research and invention depends on the nature of these relationships, as noted in Section 3.

Our empirical analysis includes all publicly traded companies headquartered in the United States that had at least one year of reported R&D expenditures, at least one granted patent, and at least three years of consecutive financial records in Compustat between 1980 and 2015. We measure corporate R&D using company patents, scientific publications by corporate scientists, the employment of scientists profiled in the American Men & Women of Science (hereafter, "AMWS scientists"), and R&D expenditures. Measuring the relevance of public science to a focal firm's innovative activity is crucial to our analysis. We use a firm's previous publishing across OECD natural science subfields to identify relevant public knowledge. To identify relevant human capital, we use the SPECTER deep learning algorithm to measure the textual similarity between PhD dissertations and the focal firm's patents (Cohan, Feldman, Beltagy, Downey, & Weld, 2020).[3] We use a firm's previous patenting across technology subclasses to identify public inventions relevant to the firm.[4]

Estimating the effect of public science on corporate R&D suffers from a classical endogeneity problem: technological shocks that affect public science can also affect corporate R&D, leading to biased OLS estimates. Federal funding may offer a source of exogenous variation in the public science relevant to a firm. We exploit changes in federal funding that are driven by political rather than technological forces. Specifically, we use the federal agency R&D budgets that are predicted by the political composition of the relevant congressional appropriations subcommittees. Firms differ in the share of their publications published in various subfields. Subfields differ in the extent to which their publications are funded by different federal agencies. The combination reflects the extent to which firms are exposed to R&D funding shocks from different agencies. To arrive at a firm-specific instrumental variable for relevant public knowledge, we create a many-to-many crosswalk from OECD natural science subfields to publications, and from publications to R&D funding by federal agencies. We use a similar approach to develop firm-specific exogenous variation in human capital and public inventions.

We present three main results. First, we find that abstract public knowledge *per se*—publications in scientific journals—has little effect on the various components of corporate R&D. This means that corporate innovation is largely unresponsive to "pure" knowledge

[3]Recent work has used machine learning to establish connections between patents (e.g., Kelly, Papanikolaou, Seru, & Taddy, 2021), between patents and research grants (e.g., Myers & Lanahan, 2022), and to classify publications into fields (e.g., Angrist, Azoulay, Ellison, Hill, & Lu, 2020).

[4]We get similar results if dissertations are matched to firms using OECD natural science subfields, or if we use non-corporate publications cited by patents and a firm's previous patenting across technology subclasses to measure relevant public inventions. See subsection 6.8 for details.

spillovers.

Second, public invention reduces corporate R&D. An increase in relevant university patents of one standard deviation reduces corporate patents by about 51%, corporate publications by approximately 33%, and the employment of AMWS scientists by about 8%. Further, we find that an increase in public invention reduces the firm's profits, suggesting that, on balance, public inventions compete with corporate inventions more than they serve as inputs into corporate innovation.

Third, we find a positive effect of human capital on corporate R&D. An increase of one standard deviation in PhD dissertations that are textually similar to a focal firm's patents increases firm patents by approximately 53%, publications by approximately 22%, and the employment of AMWS scientists by approximately 9%. Higher human capital from universities also increases firm profits, consistent with a reduction in the cost of invention when relevant human capital becomes more abundant.

These effects vary across firms and industries. In particular, firms on the technology frontier appear to respond less to public invention as compared to followers and to benefit more from human capital. Similarly, public science appears to stimulate corporate research in life sciences to a greater extent than in other industries.

Taken together, our findings indicate that the public science that matters for corporate innovation—the science developed into patented inventions and embodied in the human capital of people—is both excludable and rivalrous. Thus, the expansion of public science may not lead to the sustained productivity growth that standard models of economic growth would predict. Our results also point to the importance of the growing technology commercialization activities of universities. Indeed, between 1980 and 2021, the share of basic research in the R&D performed by U.S. universities declined from 67% to 62%, while the share of applied research and development correspondingly grew from 33% to 38%, even as their R&D expenditures grew more than ten fold (in nominal terms) from around $6 billion to nearly $90 billion (National Center for Science and Engineering Statistics, 2023a).

We make two main contributions. First, we contribute to the literature that examines the effect of public science on corporate R&D, as briefly discussed in Section 2. We focus on established firms, rather than individual researchers, industries, regions, or national economies, the focus of prior studies. Our simple framework delineates how different components of public science, namely publications, patents, and people, affect upstream scientific research and downstream technology development in corporations. Our findings suggest that university research is most relevant for corporate innovation not as abstract, non-rivalrous ideas, but rather as embodied, market-supplied inputs. Incumbent corporations appear to have a limited ability to absorb and use abstract ideas produced by universities. It is only when those ideas are developed into inventions that they become relevant to firms, reducing the demand for internal invention by incumbent corporations and hence also reducing the demand for internal research. In clarifying the relationship between university research and corporate R&D, our findings also point to an important implication of university technology commercialization activities for R&D in incumbent firms. In particular, the expansion of university research, particularly more applied research, may spur additional competition from startups, with corresponding changes in corporate R&D.

Second, we make a data contribution by using funding acknowledgments and other bibliometric and textual linkages to connect federal agency funding to publications, PhD dissertations, and patents. We build on Babina, He, Howell, Perlman, and Staudt (2023), Myers and Lanahan (2022), and Azoulay, Ding, and Stuart (2009) by linking university publications, PhD dissertations, and patents with federal funding, and using exogenous changes in agency R&D funding to estimate their impact on corporate R&D. To our knowledge, we are the first to *indirectly* link federal funding to public knowledge, human capital, and public invention that is *relevant to* a given firm's R&D, even if not directly used by the firm. We exploit differences in the political composition of congressional appropriations subcommittees as a source of exogenous variation in agency R&D funding. This enables us to analyze the joint effect of the three components of public science on both upstream and downstream corporate R&D without the potential bias induced by how firms select the public science to use in innovation.

The paper proceeds as follows. Section 2 places this study in the related literature. Section 3 presents the conceptual framework that guides our empirical investigation. Section 4 discusses and summarizes the data, Section 5 outlines the econometric specifications, and Section 6 presents the results. Section 7 concludes and suggests directions for future work.

2 Related Literature

A voluminous literature has explored how public science affects corporate R&D through knowledge and training spillovers or the acquisition of university inventions. Early influential studies have surveyed industrial research managers on the perceived importance of public science to corporate innovation. These include the Yale survey on appropriability and technological opportunity (Klevorick, Levin, Nelson, & Winter, 1995; Nelson, 1986; Rosenberg & Nelson, 1994), the pioneering surveys by Mansfield (1991, 1995, 1998), the Carnegie Mellon survey on industrial R&D (Cohen et al., 2002), and the EU Community Innovation Survey (Beise & Stahl, 1999; Laursen & Salter, 2004; Tether & Tajar, 2008). These studies suggest that scientific research from universities is of limited *direct* value for corporate R&D. However, because these studies lack firm-specific measures of the stock of relevant public science, they do not directly address *how* public science affects corporate R&D.[5]

Other studies use citations to the non-patent literature (NPL) to measure the use of science in corporate invention (e.g., Fleming, Greene, Li, Marx, & Yao, 2019; McMillan, Narin, & Deeds, 2000; Narin, Hamilton, & Olivastro, 1997). These studies show that patent citations to scientific papers have increased over time, particularly for patents in the life-sciences, and for patents by startups. Most of the science cited is government-funded and produced by universities, federal laboratories, and other public research institutions, though AT&T, IBM, DuPont, and Merck also figure prominently. However, though these studies show that inventions have become closer to science, *how* public science affects corporate R&D remains unclear. We find that public science affects corporate R&D only when the knowledge is developed by universities into patents or embodied in people (PhD graduates).

Several recent studies estimate the effect of public funding for research on patented invention (Azoulay, Graff Zivin, Li, & Sampat, 2019; Myers & Lanahan, 2022), on the composition and intensity of corporate R&D (Mulligan, Lenihan, Doran, & Roper, 2022; Scandura, 2016), and on academic entrepreneurship (Babina et al., 2023). Myers and Lanahan (2022) exploit windfall grant funding resulting from non-competitive grant matching policies that vary across states and over time. They find that for every patent produced by grant recipients of the Department of Energy, three additional patents are produced by non-recipients. Babina et al. (2023) use windfall changes in agency funding to estimate the effect on university entrepreneurship, publishing, and patenting. We map agency R&D to public science relevant to a given firm to estimate how the different components of public science affect corporate R&D. We exploit differences in the political composition of congressional appropriations subcommittees as a source of exogenous variation in agency R&D funding, and in turn, as a source of exogenous variation in public science.

Our results also add to Azoulay et al. (2019), who analyze the effect of National Institutes of Health (NIH) grant funding for research and trace the impact on patenting by pharmaceutical and biotechnology firms during 1980-2012. They find that an increase of $10 million in NIH grant funding for a research area leads to 2.3 additional private patents, suggesting that public research encourages private innovation in the life sciences.[6] Our heterogeneity

[5]Over the past several decades, researchers have also investigated "additionality"—whether government spending crowds out or stimulates additional private R&D investments—at various levels of aggregation, including industries (e.g., Mamuneas & Nadiri, 1996), firms (e.g., Einiö, 2014; Lichtenberg, 1984; Moretti, Steinwender, & Van Reenen, 2021; Wallsten, 2000) and individuals (e.g., Goolsbee, 1998). Perhaps not surprisingly, given the diversity of approaches and levels of analysis, these studies have produced conflicting results (see reviews by David, Hall, & Toole, 2000; Dimos & Pugh, 2016). Previous studies have also documented substantial heterogeneity in response to government subsidies by firm size (González, Jaumandreu, & Pazó, 2005) and R&D intensity (Szücs, 2020).

[6]More than half of the patents resulting from NIH research grants are for diseases different from those

analysis similarly reveals that public knowledge provides some encouragement for corporate innovation in the life sciences, but that outside this unique setting, public knowledge appears to have little effect on patenting and publishing by incumbent firms. Our findings therefore caution against generalizing from the life sciences to other sectors.

Another strand of the literature focuses on the localization of spillovers from universities (e.g., Belenzon & Schankerman, 2013; Hausman, 2022; Tartari & Stern, 2021; Valero & Van Reenen, 2019). Tartari and Stern (2021) examine the effect of university funding on local startups at the zip code level. Consistent with our findings, they document a positive effect on local entrepreneurship from increases in funding for universities, but not for national laboratories. A possible explanation is that, unlike national laboratories, universities also embody knowledge in human capital used by new ventures. In other words, it is likely that human capital from universities is the source of new startups. Similarly, Hausman (2022) studies the effect of university innovation on local industrial agglomeration at the county-by-industry level. She documents higher growth in employment, wages, and corporate patenting after the passage of the Bayh-Dole Act in industries more closely related to the local university's technological strengths. Consistent with Tartari and Stern (2021), she finds that this growth is primarily driven by new ventures in university-linked industries. However, neither study analyzes the effect on incumbent firms. Indeed, incumbent R&D and profitability depend on whether startups commercializing university discoveries supply their innovations to incumbents or compete with them. Our results suggest that the competition effect is the dominant effect.

Overall, our paper differs from prior literature in a couple of important ways. First, we study the effects of three distinct components of public science—knowledge, human capital, and invention—on both upstream corporate R&D (scientific research or "R") and downstream corporate R&D (technology development or "D"). Second, we make progress on data and identification at the firm level rather than at the industry, zip code, or individual researcher level. For each firm, we measure the *potentially relevant* public knowledge, human capital, and public invention based on: (i) the textual similarity between publications, dissertations, and patents; (ii) the classification of patents and publications in various CPC subclasses and OECD subfields, respectively; and (iii) non-patent literature citations from patents to publications. We also match renowned scientists profiled in the American Men & Women of Science directories to thousands of R&D-performing, publicly traded, American firms and their subsidiaries over three-and-a-half decades. This allows us to measure corporate

initially funded, indicating the presence of knowledge spillovers. This highlights the importance of linking science to innovation without assuming that science affects innovation only in a narrowly defined intended area. We implement this approach when we measure the public science that is *potentially relevant* to the firm, and not just that which is *actually used* by the firm.

investment in research for firms that do not publish scientific publications.

3 Conceptual Framework

We adapt the framework from Arora et al. (2021a) to focus on the effect of public science on internal research and invention. Public science has at least three components: knowledge disclosed in scientific publications, trained human capital (Pavitt, 1991), and inventions based on public knowledge (Fabrizio & Di Minin, 2008). These potentially differ in how they affect internal research and invention by incumbent corporations. For instance, public knowledge may complement internal research or substitute for it. Inventions based on public knowledge substitute for internal inventions, and may even compete with the firm's innovations. Human capital, on the other hand, tends to increase internal research and invention.

3.1 Setup

A firm's product market profit, $\Pi(d)$, depends on its innovations—the number of inventions it introduces into the market—d. These inventions may be acquired from outside the firm or internally generated. Internal inventions are produced at a unit cost $w(k)\phi(r,u)$, where r is internal research and u is the stock of public knowledge that is relevant to the firm. The term $w(k)$ represents the wage of inventors and is assumed to fall as more human capital, k, is available to the firm. The term ϕ represents the inverse of invention productivity and is assumed to decrease with r at a diminishing rate. We also assume that ϕ decreases with u.[7]

The relationship between public knowledge and internal research in reducing the unit cost of internal invention is important for how the stock of public knowledge relates to investments in internal research.[8] Public knowledge may complement internal research because performing internal research provides the absorptive capacity to use the knowledge.[9]

We assume that the cost of internal research is given by $\gamma(k)\frac{1}{2}r^2$, which also depends on k, the supply of relevant human capital. In other words, increasing the number of trained PhD scientists produced by universities reduces the firm's cost of both internal research and internal invention.

[7]The cost function reflects a simple linear production function $d = \lambda(r,u)n$, where n is the number of inventors the firm employs and $\lambda(r,u)$ is the productivity of the inventors. Thus, the cost of internal inventions is simply $w(k)n$ so that $\phi = \frac{1}{\lambda(r,u)}$.

[8]Complementarity exists if $-\frac{\partial^2\phi}{\partial r \partial u} > 0$ and substitutability exists if $-\frac{\partial^2\phi}{\partial r \partial u} < 0$.

[9]There is a large literature on absorptive capacity that argues firms must invest in internal research to benefit from public knowledge (e.g., Cohen & Levinthal, 1990; Rosenberg, 1990). Baruffaldi and Poege (2020) show that firms are more likely to cite papers presented at conferences where the firm's scientists also participated.

Inventions by university researchers (henceforth, "public inventions") can either be inputs to the firm's own innovation or compete with the firm's innovations in the marketplace. For example, university spinoffs and startups could be acquired by the firm or instead compete with it in the marketplace, either directly or after being acquired by rivals (OECD, 2003).

To model public inventions as inputs to the firm's own innovation, we assume that the firm's innovation, d, is the sum of those derived from internal inventions, d_1, and those derived from public inventions, d_2. We assume that the firm can acquire public inventions at an increasing marginal cost represented by $a_0 d_2 + \frac{1}{2} a_1 {d_2}^2$.[10]

To model public inventions that compete with the firm's innovations, we allow the focal firm's product market profits to also depend on public inventions. Specifically, we assume $\Pi(d, \tilde{d}) = b_0 + b_1 d - \frac{1}{2} b_{11} d^2 - b_2 \tilde{d} - \frac{1}{2} b_{22} \tilde{d}^2 + b_{12} d \tilde{d}$, where $\tilde{d}$ stands for public inventions that compete with the firm's innovation. We assume that $\Pi(d, \tilde{d})$ increases with d, decreases with $\tilde{d}$, and is concave. Importantly, we assume that the firm takes the number of competing public inventions as given. Note that the marginal return to innovation (gross of the costs) is simply $b_1 - b_{11} d + b_{12} \tilde{d}$, which increases with $\tilde{d}$ if $b_{12} \geq 0$ and decreases with $\tilde{d}$ otherwise. We say that public inventions and internal inventions are strategic complements if $b_{12} \geq 0$ and strategic substitutes otherwise.

3.2 Implications for Firm Value and Innovation

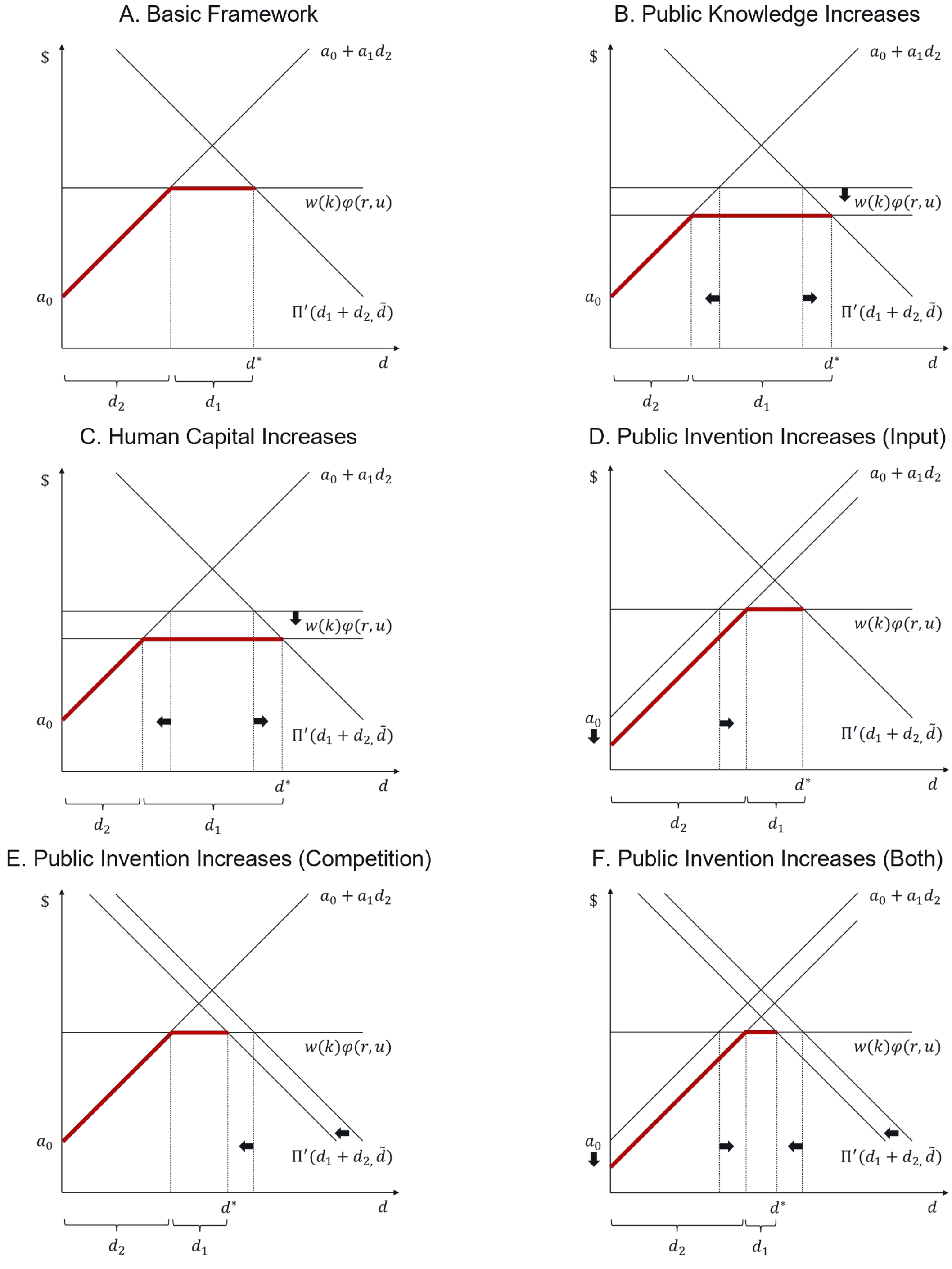
The value of the firm is $v(d_1, d_2) = \max_{d_1, d_2, r} \{\Pi(d_1 + d_2, \tilde{d}) - d_1 w(k)\phi - \gamma(k)\frac{1}{2} r^2 - a_0 d_2 - \frac{1}{2} a_1 {d_2}^2\}$. We assume that v is concave in its arguments. Panel A in Figure 2 summarizes the elements of our basic conceptual framework.

3.2.1 Public Knowledge

An increase in relevant public knowledge increases the value of the firm, v, by reducing the cost of internal invention. Formally, applying the envelope theorem, $\frac{\partial v}{\partial u} = -d_1 \frac{\partial \phi}{\partial u} > 0$. If internal research complements public knowledge (i.e., $-\frac{\partial^2 \phi}{\partial r \partial u} > 0$), then an increase in public knowledge will also increase internal research. If they are substitutes, then there are two opposing effects. Substitutability reduces the marginal return to internal research. However, a reduction in the cost of internal invention due to public knowledge increases the scale of internal invention, thereby increasing the marginal return on internal research.

[10] For simplicity, the total cost of public inventions acquired by the firm is assumed to be $a_0 d_2 + \frac{1}{2} a_1 {d_2}^2$. The assumption of a rising marginal cost of public invention implies that the firm has market power, perhaps due to its location or the specific inventions it can commercialize. The results are similar if the firm is a price taker and has an increasing cost of internal invention, except that an increase in demand for invention would leave internal invention and research unchanged but decrease invention sourced from the public sector.

Figure 2: Conceptual Framework



Notes: This figure presents our basic conceptual framework (Panel A). The firm's innovation, d, is the sum of internal inventions, d_1, and external inventions, d_2. The "demand" for innovation is represented by $\Pi'(d_1 + d_2, \tilde{d})$. The "supply" of public inventions is represented by $a_0 + a_1 d_2$, while the "supply" of internal inventions is represented by $w(k)\phi(r, u)$, where $w(k)$ is the wage of inventors, k is human capital, r is internal research, u is public knowledge, and $\gamma(k)\frac{1}{2}r^2$ is the cost of r. Comparative statics for increases in public knowledge (Panel B), human capital (Panel C), and public invention (Panels D, E, and F) are also included.

The effect on internal invention follows a similar logic. The direct effect of an increase in public knowledge is to reduce ϕ in the cost of internal invention, as shown in Panel B. As long as the marginal cost of internal invention decreases, overall innovation increases because the increase in internal invention is only partly at the expense of external invention.

3.2.2 Human Capital

As with public knowledge, an increase in human capital supply increases firm value. Formally, $\frac{\partial v}{\partial k} = -d_1 \phi \frac{\partial w}{\partial k} - \frac{1}{2} r^2 \frac{\partial \gamma}{\partial k} > 0$. An increase in the supply of human capital reduces the cost of internal invention and research, as shown in Panel C. Since external invention substitutes for internal invention, the former will fall.

3.2.3 Public Invention

Insofar as public inventions are inputs to the firm's own innovation, they increase firm value but decrease internal invention and research. An increase in public invention can be modeled as a reduction in a_0, as shown in Panel D, in which case $-\frac{\partial v}{\partial a_0} = d_2 > 0$. However, a reduction in the marginal cost of external invention will decrease internal invention, which will, in turn, decrease internal research. Intuitively, an increase in the supply of an input increases the firm's value. However, it will decrease the demand for substitute inputs.

Conversely, an increase in public sector inventions that compete with the firm's innovations, $\tilde{d}$, will decrease firm value. Formally, $\frac{\partial v}{\partial \tilde{d}} = -b_2 - b_{22}\tilde{d} + b_{12}d \leq 0$ because Π was assumed to fall with $\tilde{d}$, as shown in Panel E. Indeed, $b_{12} \leq 0$ is sufficient for this result (if b_2 and b_{22} are both positive). If $b_{12} < 0$, then an increase in $\tilde{d}$ will reduce d_1 and hence also will reduce r. Conversely, if $b_{12} > 0$, an increase in $\tilde{d}$ will increase d_1 and hence also will increase r. In other words, one has to examine the pattern of relationships with value as well as internal invention and research to assess how public inventions relate to corporate innovation. Table 1 summarizes the predictions of our basic conceptual framework.

3.2.4 Leaders and Followers

Even if the fruits of public science are available to all, they may not benefit all firms equally. It is plausible that for leading firms, which require "frontier" innovations, sourcing public inventions that match their needs is more difficult. By contrast, for follower firms trying to "catch up" to the technology frontier, public inventions may be more plentiful. If so, frontier firms would rely to a greater extent on internal inventions and also invest more

Table 1: The Predicted Effect of Public Science on Firm Value and Innovation

(1) Equation	(2) Comparative statics	(3) Effect on firm
A. Higher public knowledge		
Publications	$\partial r/\partial u$	$\uparrow$ if r complements u in lowering ϕ; $\downarrow$ or $\uparrow$ otherwise
Patents	$\partial d_1/\partial u$	$\uparrow$ if r complements u in lowering ϕ; $\downarrow$ or $\uparrow$ otherwise
Firm value	$\partial v/\partial u$	$\uparrow$
B. Higher human capital		
Publications	$\partial r/\partial k$	$\uparrow$
Patents	$\partial d_1/\partial k$	$\uparrow$
Firm value	$\partial v/\partial k$	$\uparrow$
C. Higher public invention (input)		
Publications	$-\partial r/\partial a_0$	$\downarrow$
Patents	$-\partial d_1/\partial a_0$	$\downarrow$
Firm value	$-\partial v/\partial a_0$	$\uparrow$
D. Higher public invention (competition)		
Publications	$\partial r/\partial \tilde{d}$	$\uparrow$ if d_1 and $\tilde{d}$ are strategic complements; $\downarrow$ otherwise
Patents	$\partial d_1/\partial \tilde{d}$	$\uparrow$ if d_1 and $\tilde{d}$ are strategic complements; $\downarrow$ otherwise
Firm value	$\partial v/\partial \tilde{d}$	$\downarrow$

Notes: This table summarizes the theoretical predictions regarding the effect of higher public knowledge, human capital, and public invention on the publications, patents, and value of the focal firm.

in internal research compared to follower firms.[11] This suggests that frontier firms may also respond differently to public science than followers. Public knowledge may substitute for internal research for followers but may complement internal research in frontier firms. Insofar as human capital reduces the cost of internal research, frontier firms would be more responsive to increases in human capital. On the other hand, followers may respond more to an expansion in the supply of public inventions.

4 Data

We combine data from several sources: (i) scientific publications by corporations, universities, federal laboratories, and other public research institutions, acknowledgments of federal grants by these publications, and citations by patents to publications from Dimensions (Digital Science, 2022); (ii) scientists profiled in the American Men & Women of Science

[11] Frontier firms may have a higher demand for inventions, may face a lower effective supply of public inventions, or internal research and public science may be strategic complements. These issues are explored in our empirical analysis.

directories; (iii) PhD dissertations from ProQuest Dissertations & Theses Global; and (iv) firm financial information from S&P's Compustat North America. We complement these data with scientific publication information from Clarivate's Web of Science, patent data from U.S. Patent and Trademark Office's PatentsView and the European Patent Office's PATSTAT, federal procurement contract data from the Federal Procurement Data System, and federal grant data from the Treasury DATA Act Broker (see Arora et al., 2021a; Arora, Belenzon, & Sheer, 2021b; Belenzon & Cioaca, 2021).

Corporate innovation and public science are multi-dimensional. Our measures capture both corporate innovation inputs (R&D expenditures and AMWS scientists) and outputs (publications and patents). Moreover, they capture upstream corporate science (publications and AMWS scientists) and downstream corporate invention (patents). As well, we measure three components of relevant public science: knowledge, human capital, and invention. The construction of the main variables used in our econometric analyses is summarized below and detailed in Online Appendix A.

4.1 Upstream Corporate Research: Publications and AMWS Scientists

We measure upstream corporate research using (i) the number of publications authored by scientists affiliated with the firm (from Arora et al., 2021a) and (ii) the number of scientists employed by the firm and profiled in the American Men & Women of Science (AMWS), a directory of accomplished North American scientists in science and engineering (similar to Kim & Moser, 2021). Using the digital editions of AMWS between 2005 and 2021, we identified 20,097 AMWS scientists who worked for 1,727 different firms in our panel between 1980 and 2015.

Both publications and AMWS scientists are noisy measures of corporate investment in research. In our estimation sample, the pairwise correlation between the annual flow of corporate publications and the number of AMWS scientists employed per firm is 0.68, suggesting that there is a strong shared component. Employing AMWS scientists is much more likely for firms that publish (54%) than firms that do not publish (11%). However, as Table 2 shows, 46% of the firms that publish do not employ AMWS scientists. Hence, we use both measures to capture upstream corporate R&D activity.

Table 2: Cross Tabulation of Measures of Upstream Corporate Research

	(1) Do not employ AMWS scientists		(2) Employ AMWS scientists		(3) Total	
	Count	%	Count	%	Count	%
Do not publish	1,046	89%	132	11%	1,178	100%
Publish	1,005	46%	1,189	54%	2,194	100%
Total	2,051	61%	1,321	39%	3,372	100%

Notes: This table provides a cross-tabulation of measures of upstream corporate research for the 3,372 firms included in our estimation sample. The unit of analysis is a firm.

4.2 Public Knowledge: Non-corporate Publications

We source scientific publications from Dimensions. This dataset provides information on which federal agencies (if any) provided the grants that funded each publication, enabling us to implement an identification strategy that uses exogenous variation in federal agency R&D funding.[12] The dataset also links university (and other non-corporate) publications to the patents that cite them, which we use to construct alternative measures of relevant public invention and human capital.

We use the OECD research classification system to determine the public knowledge that is *potentially relevant* to a firm's innovation.The 25 OECD natural science subfields (listed in Appendix Table A2) provide a standardized way of categorizing scientific publications into such scientific disciplines as mathematics, chemical sciences, and biological sciences. We assume that new publications in a particular subfield are most relevant to firms that have recently published in that subfield.

Our firm-year measure of relevant *Public knowledge* is the weighted sum of non-corporate publications. The weights are the focal firm's shares of publications across OECD subfields during the previous 5-year time cohort, as follows:

$$Public\ knowledge_{i,t} = \sum_{o \in O} Publications_{o,t} \times Precohort\ share\ of\ publications_{i,o} \tag{1}$$

The index o denotes OECD subfields. $Publications_{o,t}$ is the number of non-corporate publications published in year t in subfield o. $Precohort\ share\ of\ publications_{i,o}$ is firm i's share of publications in subfield o during the previous (lagged) 5-year time cohort, obtained by dividing the number of firm publications published in subfield o by the total number of firm publications in the time cohort. We generate a stock measure of *Public knowledge* using a perpetual inventory method with a 15% depreciation rate.

[12] As of 2022, the Dimensions dataset combined 131.5 million cited and citing publications, 6.3 million research grants with related funding organizations, as well as 149.7 million cited and citing patents.

4.3 Human Capital: PhD Dissertations

We measure human capital using PhD dissertations sourced from ProQuest Dissertations & Theses Global (hereafter, PQDT), recognized by the U.S. Library of Congress as the official repository for dissertations, and containing more than 5 million dissertations and theses from universities around the world between 1900 and 2021. We exclude "soft science" PhD dissertations from our data.[13] We also discard PhD dissertations from non-U.S. universities and all master's degree theses. We end up with 771,023 U.S. PhD dissertations awarded between 1985 and 2016 in 394 "hard science" research fields.

PhD dissertations are not typically cited by publications or patents. Therefore, we assess the relevance of trained human capital to corporate innovation based on the textual similarity between the abstracts of dissertations and the abstracts of company patents. We calculate that similarity using SPECTER, a deep learning algorithm that considers both the content and the context of scientific tests. In brief, SPECTER uses a transformer-based neural network to process natural language texts. Online Appendix A provides a detailed description of how we implement SPECTER in our variable construction.

Our firm-time cohort measure of relevant *Human capital* is the weighted sum of PhD dissertations, using the textual similarity to patents as weights:

$$Human\ capital_{i,t} = \sum_{d \in D} Maximum\ textual\ similarity_{d,i,t} \tag{2}$$

D is the set of PhD dissertations in the top 1,000 most similar dissertations for one or more of the patents granted to firm i during the 5-year time cohort t. $Maximum\ textual\ similarity_{d,i,t}$ is the maximum textual similarity score between the abstract of dissertation d and the abstracts of all patents granted to firm i during the 5-year time cohort t.[14]

A subset of PhD dissertations are published in scientific journals and (subsequently) cited by patents. We construct a complementary firm-year measure, *Human capital, cited*, as the

[13] Doing so is not straightforward because the variable that describes dissertations' research fields, "classterms," lists 308,862 different combinations of terms. We manually create a list of 1,027 disambiguated terms, then drop dissertations in such research fields as "literature," "history," and "social sciences."

[14] Our text-based measure captures the human capital that is *potentially relevant* to a firm's inventions without requiring "actual use" (e.g., NPL citations or employment history). For example, Arifur Rahman earned his PhD in Electrical Engineering from MIT in December 2000. His dissertation on interconnect technologies for integrated circuits was published in early 2001 in ProQuest Dissertations & Theses Global (document ID 304757014). SPECTER ranked Rahman's dissertation in the top 1,000 most similar dissertations for five of Lattice's patents granted in 2000, five granted in 2001, and another five granted in 2002. While none of these contemporaneous patents cited the dissertation, our measure nevertheless identified a link between Arifur and Lattice. Indeed, Rahman was subsequently hired by Lattice as a technical staff member in 2001. He went on to produce a number of semiconductor patents for Lattice (with filing dates starting in 2002) and subsequent corporate employers, including Intel, Altera, and Xilinx.

weighted sum of *published* PhD dissertations cited by patents in various patent subclasses, as detailed in Appendix A.[15] The weights are the focal firm's shares of patents across patent subclasses during the previous 5-year time cohort. We construct a third measure, *Human capital, OECD*, by first classifying PhD dissertations into OECD natural science subfields. We then use the focal firm's previous patenting across technology subclasses that rely on science from various OECD subfields to identify relevant human capital. We validate the logic behind our measures of firm-relevant human capital with three case examples included in Appendix C. We report results using the alternative measures in Section 6.8. Our findings are not sensitive to the specific approach used for measuring firm-relevant human capital.

4.4 Public Invention: University Patents

We measure public invention using patents granted to American universities. This measure reflects the extent to which universities *directly* develop inventions. We assume that university patents represent public inventions that firms can acquire (either by licensing or by acquiring the relevant startup) or have to compete against.

Our firm-year measure of *Public invention* is the weighted sum of university patents. The weights are the focal firm's shares of patents across patent subclasses during the previous 5-year time cohort, as follows:

$$Public\ invention_{i,t} = \sum_{s \in S} University\ patents_{s,t} \times Precohort\ share\ of\ patents_{i,s} \tag{3}$$

The index s denotes patent subclasses, identified using the first four digits of the current CPC classification from the U.S. Patent & Trademark Office (USPTO). $University\ patents_{s,t}$ is the count of patents granted to universities in subclass s in year t. $Precohort\ share\ of\ patents_{i,s}$ is firm i's share of patents in subclass s during the previous 5-year time cohort, obtained by dividing the number of firm patents granted in subclass s by the total number of firm patents in that time period.

In robustness checks, we use a broader measure of the supply of relevant public invention using publications that lead to inventions, as detailed in Appendix A. We construct *Public invention, broad* as the stock of non-corporate publications that are cited by patents in various patent subclasses, weighted by the share of the focal firm's patents across patent

[15] Continuing with the previous example, Arifur Rahman's dissertation was published under the title *Interconnect Limits on Gigascale Integration (GSI) in the 21st Century* (DOI 10.1109/5.915376) in 2001. This publication was subsequently cited by more than one hundred patents granted between 2004 and 2021, including patents assigned to IBM, Seagate Technologies, and Texas Instruments. Similar to our primary measure, our alternative measure captures the relevance of Rahman's human capital, at graduation, not only to his eventual employer, Lattice, but also to other firms that innovate in semiconductors.

subclasses during the previous 5-year time cohort. Because some publications are not directly cited by patents, yet still reflect external inventions that are potentially relevant to firms, we also construct another measure *Public invention, SPECTER* using the textual similarity between the abstracts of non-corporate publications and the abstracts of corporate patents. Textual similarity is assessed using the SPECTER algorithm. We report results using these alternative measures in Section 6.8. Our findings are not sensitive to the specific approach used for measuring firm-relevant public invention. Table 3 summarizes the main variables used in the econometric analyses.

4.5 Descriptive Statistics

Our estimation sample consists of an unbalanced panel of 3,372 U.S.-headquartered publicly traded firms over 1986-2015, totaling 41,698 firm-year observations.[16] Table 4 presents summary statistics for the main independent, dependent, and control variables.[17] Our sample contains a wide distribution of R&D expenditures, ranging from $0.6 million at the 10th percentile to $202.6 million at the 90th percentile, partly reflecting a wide distribution of firm sizes. On average, firms produce 28 patents and 16 publications per year and employ 5 AMWS scientists. Approximately 86% of firms have at least one patent and 65% have at least one publication between 1986 and 2015.

Firms vary substantially in their exposure to public science. The average stock of firm-relevant public knowledge (62,550 publications) represents a small fraction of the 2,714,527 publications added, on average, to Dimensions each year between 1986 and 2015. However, the average flow of firm-relevant human capital in a 5-year period (6,413 PhD dissertations) represents a larger fraction of the 28,537 PhD degrees in the hard sciences awarded by U.S. universities each year between 1986 and 2015.

Our measures of the three components of public science are strongly positively correlated, as shown in Appendix Table C15. In general, firms that face abundant relevant public knowledge also tend to face abundant relevant human capital (whether measured by PhD

[16] We begin with the sample of 4,520 firms over 1980-2015 from Arora et al. (2021b), totaling 60,885 firm-year observations. These are U.S.-headquartered publicly traded firms with at least one year of reported R&D expenditures, at least one granted patent, and at least three years of consecutive financial records from the first patent. We split this sample into 5-year time cohorts (e.g., 1980-1984, 1985-1989, etc.) to determine a firm's exposure to public science. Because observations from the first 5-year time cohort for each firm are used to calculate the firm's (i) lagged shares of patents across CPC subclasses and (ii) lagged shares of publications in each OECD subfield, they are subsequently excluded from the analysis sample. Similarly, because we lag independent and control variables by one year, additional observations are excluded, arriving at 41,698 firm-year observations.

[17] Summary statistics by main industry, for the instrumental variables, and for the alternative measures of public science are reported in Appendix Tables C16, A8, and C17, respectively.

Table 3: Main Variables

Variable name	Variable description
A. Dependent variables	
Patents	Patents granted by the USPTO to the focal firm
Publications	Scientific publications that have at least one author affiliated with the focal firm
AMWS scientists	Scientists profiled in AMWS that are employed by the focal firm
R&D expenditures	R&D expenditures reported by the focal firm
Tobin's Q	Market value divided by assets
B. Independent variables	
Public knowledge	Stock of non-corporate publications published in various OECD natural sciences subfields
Human capital	PhD dissertations, based on the textual similarity between abstracts of dissertations and abstracts of firm patents
Public invention	Stock of university patents granted by the USPTO in various CPC subclasses
C. Alternative independent variables	
Human capital, cited	Published PhD dissertations cited by patents in various CPC subclasses
Human capital, OECD	PhD dissertations mapped to various OECD subfields, based on the importance of the OECD subfields to patenting in various CPC subclasses
Public invention, broad	Stock of non-corporate publications cited by patents in various CPC subclasses
Public invention, SPECTER	Stock of non-corporate publications, based on the textual similarity between non-corporate publications and firm patents

Notes: This table summarizes the main variables used in the econometric analyses. Stock measures are constructed using a perpetual inventory method with a 15% depreciation rate. For example, $(Public\ knowledge,\ stock)_t = (Public\ knowledge)_t + (1-\delta)(Public\ knowledge,\ stock)_{t-1}$, where $\delta = 0.15$. We omit the term "stock" from variable names to simplify notation.

dissertations or the published versions of PhD dissertations) and public invention (whether measured by university patents or non-corporate publications cited by patents). Large firms, in particular, face more abundant relevant public science than small firms. Consistent with the idea that trained human capital and public invention are co-produced in universities, 62% of firms with above median human capital also have above median public invention, as shown in Appendix Table C18.

Table 4: Summary Statistics for Main Variables

	(1)	(2)	(3)	(4)	(5)	(6)
				Distribution		
	Observations	Mean	Standard deviation	10th	50th	90th
Public knowledge$_{t-1}$	41,698	62,550	85,302	0.0	0.0	178,684.6
Human capital$_{t-1}$	41,698	6,413	9,709	0.0	2,761.9	17,124.0
Public invention$_{t-1}$	41,698	266	512	0.0	56.0	780.7
Patents$_t$	41,698	28	157	0.0	1.0	44.0
Publications$_t$	41,698	16	94	0.0	0.0	16.5
AMWS scientists$_t$	41,698	5	32	0.0	0.0	5.0
R&D expenditures ($ mm)$_t$	36,712	142	656	0.6	11.8	202.6
Tobin's Q$_t$	36,800	34	688	0.4	1.7	16.2
R&D stock ($ mm)$_{t-1}$	41,698	603	3,134	1.0	38.6	773.3
Sales ($ mm)$_{t-1}$	41,439	3,101	14,336	4.0	192.5	5,420.4
R&D stock$_{t-1}$ / $Assets_t$	41,035	2	3	0.0	0.4	6.3

Notes: This table provides summary statistics for the main variables used in the econometric analyses. The analysis sample is at the firm-year level and includes an unbalanced panel of 3,372 U.S.-headquartered publicly traded firms from 1986 to 2015.

5 Econometric Framework

We turn to the empirical investigation of the theoretical predictions from Table 1.

5.1 Patents, Publications, AMWS Scientists, and R&D Expenditures Equations

We estimate the following specification for the relationship between corporate innovation and public science (bold indicates vector representation):

$$\ln(Y_{i,t}) = \alpha_0 + \alpha_1 \ln(X_{i,t-1}) + \boldsymbol{Z}'_{i,t-1}\boldsymbol{\omega} + \boldsymbol{\eta_i} + \boldsymbol{\tau_t} + \epsilon_{i,t} \tag{4}$$

We use multiple dependent and independent variables (see Appendix A for details on variable construction). $Y_{i,t}$ represents corporate innovation inputs (*R&D expenditures* and *AMWS scientists*) and outputs (*Publications* and *Patents*), for firm i in year t. $X_{i,t-1}$ represents the *Public knowledge* (stock), *Human capital*, and *Public invention* (stock) relevant to firm i's innovation in the lagged year or time cohort. The vector $\boldsymbol{Z}$ includes time-varying controls, such as $ln(Sales)_{t-1}$ for the R&D expenditures equation and $ln(R\&D\ stock)_{t-1}$ for the patents, publications, and AMWS scientists equations (where we also add an unreported indicator variable equal to 1 for firms without R&D expenditures prior to the focal year). In all specifications, we account for a possible direct federal funding effect by including $ln(Awards\ to\ focal\ firm)_{t-1}$, the lagged stock of federal grant and procurement dollars awarded to the focal firm and its subsidiaries. In the 2SLS specifications, we also include indicator variables equal to 1 for firms with zero-valued instruments in the prior year and a control for lagged *Agency exposure*.[18] The vectors $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ are firm and year fixed effects, respectively, and ϵ is an *iid* error term. When calculating natural logarithms, we add \$1 to variables measured in millions of dollars (e.g., *Sales, R&D stock*) and one unit to count variables (e.g., patents, publications, AMWS scientists). Standard errors are clustered at the firm level.

Our coefficient of interest is α_1. We expect the effect of public science on corporate innovation to vary by upstream and downstream R&D and by the specific component of public science. We also examine heterogeneity in effects by firm proximity to the technology frontier and by main industry.

One concern with our econometric framework pertains to our ln(1+x) transformation, which we implement to handle positively skewed count data with zeros (e.g., firms have zero publication flows in some years). We address this concern using the two-stage control function Poisson regression approach described in Lin and Wooldridge (2019) and implemented in Bellet, De Neve, and Ward (2023). We bootstrapped to estimate standard errors for the coefficient estimates. We obtain similar results to our main specifications.

5.2 Firm Value Equation

As noted in Section 3, public inventions may represent inputs to the firm's own innovation, in which case they would increase firm value. Public inventions may also compete with the firm's innovations, in which case they would decrease firm value. To assess how public inventions relate to corporate innovation *on average*, we estimate the following Tobin's Q specification:

$$\begin{aligned}\ln(Tobin's\ Q)_{i,t} =&\beta_0 \frac{R\&D\ stock_{i,t-1}}{Assets_{i,t-1}} + \beta_1 \ln\left(Public\ knowledge\right)_{i,t-1} \\ &+ \beta_2 \ln\left(Human\ capital\right)_{i,t-1} + \beta_3 \ln\left(Public\ invention\right)_{i,t-1} \\ &+ \boldsymbol{Z}'_{i,t-1}\boldsymbol{\omega} + \boldsymbol{\eta_i} + \boldsymbol{\tau_t} + \epsilon_{i,t}\end{aligned} \tag{5}$$

Tobin's Q is market value divided by book value of assets. The other elements of the specification are as previously described. Our coefficients of interest are β_1, β_2, and β_3 on the lagged firm-relevant public knowledge (stock), human capital, and public invention (stock), respectively.

[18] $Agency\ exposure_{i,t} = \sum_{s\in S}\sum_{a\in A} Reliance\ on\ public\ knowledge_{s,a} \times Precohort\ share\ of\ patents_{i,s}$ captures the weights used to calculate the instrument for public invention at the firm-year level.

5.3 Identification

A key econometric challenge is how to deal with the endogeneity of public science. We address it in an instrumental variable framework that uses the R&D budgets of federal agencies to predict firm-relevant public science. We construct a Bartik-style shift-share instrument for each component of public science. The "shift" represents federal financial support across OECD subfields (in the case of public knowledge), dissertation advisors (in the case of human capital), and patent subclasses (in the case of public invention). As multiple agencies provide such financial support, the shift for each subfield, advisor, and subclass is calculated as the weighted sum of financial support from each federal agency, where the weights capture how much of that agency's R&D budget is directed to that subfield, advisor, and subclass, respectively. The firm-specific "exposure share" is based on the firm's publishing across OECD subfields (public knowledge), the textual similarity of PhD dissertations to the firm's patents (human capital), and the distribution of the firm's patents across subclasses (public invention) in the pre-period.

A key identifying assumption is that federal agency funding for R&D is unrelated to technology and demand-side factors that also drive corporate innovation. To ensure that our results are not affected by potential violations of this assumption, we use two different approaches when building our instruments for public knowledge, human capital, and public invention. The first approach uses agency R&D budgets to construct the "shift." The second (and preferred) approach adds another step: we use two measures of the political composition of congressional appropriations subcommittees to predict agency R&D budgets, then use these *predicted* agency R&D budgets to construct the "shift." This second approach leverages the powerful and persistent roles of congressional appropriations subcommittees in federal budgeting (Davis, Dempster, & Wildavsky, 1966).

Another important identifying assumption is that firm "exposure shares" are unrelated to the same underlying factors that drive federal agency R&D budgets. For instance, if larger firms are more exposed to federal agencies that receive more R&D funding, instrumenting for public science with agency R&D funding may still lead to biased results. We examine the severity of this concern by estimating the relationship between firm size and federal R&D funding. We find a positive correlation between firm R&D stock and agency R&D funding, so we control for the lagged firm R&D stock or annual sales in all relevant specifications. Our results are qualitatively similar when we do not include the control for size.

5.3.1 Federal Funding for Public Science

The U.S. government is a substantial funder of public science. As shown in Appendix table A5, federal agencies' R&D budgets have increased from $104.6 billion per year in the 1980s to $156.1 billion per year in the 2010s (American Association for the Advancement of Science, 2021). The Dimensions dataset connects more than 4.6 million publications to their funding organizations, including federal agencies. These linkages are based on funding acknowledgments provided by the authors at publication and on administrative data collected from major funders, such as the National Science Foundation and the National Institutes of Health. We use the publications-to-grants and grants-to-federal agencies crosswalks from Dimensions, the hierarchical structure of federal agencies from the Global Research Identifier Database (GRID), and the PhD students-advisors crosswalk from PQDT to create instrumental variables for our various measures of public science.

In the simplest approach, we link federal funding for R&D with each of the three components of public science, then calculate Bartik-style shift-share instruments using firms' differential exposure to the common federal funding shocks. We report results using these instrumental variables in our robustness checks.

In our preferred approach, we address the concern that federal funding for R&D may reflect technological or demand shocks that also affect the R&D decisions of firms. Prior research suggests that political partisanship can influence federal budgets (Davis et al., 1966; Epp, Lovett, & Baumgartner, 2014). Because we need a source of agency-level variation in R&D funding, we focus on the political composition of congressional appropriations subcommittees. For each of the 12 main federal agencies (plus an "Other" category for smaller agencies), we identify which U.S. House and U.S. Senate subcommittees are responsible for reviewing their budget request to Congress, hearing testimony from government officials and other witnesses, and drafting the spending plan for each fiscal year. Appendix Table A6 summarizes the mapping between agencies and subcommittees.

For each subcommittee, we collect two pieces of information. The first measures how dominant the majority party is in the subcommittee. The variable *Majority party share* is the ratio of the number of members from the majority political party in the chamber over the total number of members in the subcommittee. The second measures the ideological orientation of the subcommittee. The variable *Democratness* is the ratio of the number of Democrats over the total number of members in the subcommittee. We use these variables to predict the R&D budget, then use the predicted R&D budget in constructing our Bartik-style shift-share instruments at the firm-year level.

The ideas behind this approach are as follows. When committees are more balanced, the majority party may have to engage in more give-and-take with the minority party. One way is to fund more of the minority party's priorities, which would result in bigger budgets. In addition, each member of the majority party may also have more bargaining power when the majority is small, leading to additional spending to benefit their constituents. In either case, we would expect agency R&D budgets to decrease when the majority party shares in the relevant subcommittees increase. Moreover, the ideological bent of the majority party may matter as well. Insofar as in the U.S. Republicans promote spending cuts while Democrats favor a larger federal government (Epp et al., 2014; Tavares, 2004), we would expect agency R&D budgets to increase when the share of subcommittee members who are Democrats increases. Appendix Table A7 shows that the political composition of congressional appropriations subcommittees predicts the R&D budgets of federal agencies in the anticipated directions. However, the political composition should be orthogonal to technological or demand shocks that also affect the R&D decisions of firms. If so, it is a source of exogenous variation in agency R&D budgets.

5.3.2 Instrument for Public Knowledge

Our preferred instrument for *Public knowledge* is the *predicted* federal funding for public knowledge published in each OECD subfield, weighted by the focal firm's shares of publications in each OECD subfield during the previous 5-year time cohort, as follows:

$$Predicted\ R\&D\ budget\ \text{-}\ public\ knowledge_{i,t} = \sum_{o \in O} Precohort\ share\ of\ publications_{i,o} \left(\sum_{a \in A} \widehat{R\&D\ budget}_{a,t} \times Reliance\ on\ agency_{o,a} \right) \tag{6}$$

O denotes OECD subfields. $Precohort\ share\ of\ publications_{i,o}$ is firm i's share of publications in subfield o during the previous 5-year time cohort, obtained by dividing the number of firm publications published in subfield o by the total number of firm publications. A is the set of 12 main federal agencies, plus an "Other" category for smaller agencies. $\widehat{R\&D\ budget}_{a,t}$ is the R&D budget predicted by *Majority party share* and *Democratness* for agency a in year t. $Reliance\ on\ agency_{o,a}$ is a share obtained by dividing the number of publications published in subfield o over 1980-2015 and funded by agency a by the total number of publications published in subfield o over 1980-2015.

5.3.3 Instrument for Public Invention

Our preferred instrument for *Public invention* is the *predicted* federal funding for publications that are relevant to university patents in each patent subclass, weighted by the focal firm's shares of patents across CPC subclasses during the previous 5-year time cohort, as follows:

$$\textit{Predicted R\&D budget - public invention}_{i,t} = \sum_{s \in S} \textit{Precohort share of patents}_{i,s} \left(\sum_{a \in A} \widehat{R\&D\ budget}_{a,t} \times \textit{Reliance on agency}_{s,a} \right) \tag{7}$$

S, *Precohort share of patents*$_{i,s}$, A, and $\widehat{R\&D\ budget}_{a,t}$ are as previously defined. *Reliance on agency*$_{s,a}$ is a share obtained by dividing the number of citations from university patents granted in subclass s over 1980-2020 to non-corporate publications published over 1980-2015 and funded by agency a by the total number of citations from university patents granted in subclass s over 1980-2020 to all non-corporate publications published over 1980-2015.

5.3.4 Instrument for Human Capital

We construct an analogous instrument for *Human capital*. Differently from the previous two instruments, we link each dissertation to a federal agency through the funding the PhD dissertation advisors received from each agency over the six-year period prior to the grant of the degree, and link each dissertation to a firm using the textual similarity to the firm's patents. Specifically, we match advisors to researchers in the Dimensions dataset using each dissertation advisor's name, school affiliation, and years of publishing activity and retrieve from Dimensions (i) the scientific publications authored by the advisors during the 6-year period preceding the PhD dissertation defense and (ii) the grant amounts and funding organizations for these publications. In our PhD dissertation dataset, 1,310,774 dissertations have advisor information, producing 1,472,326 dissertation-advisor pairs (some dissertations have more than one advisor). We assume that federal funding received by the advisor(s) of a PhD student during the 6-year duration of the PhD program affects the direction and content of the dissertation.

Our preferred instrument for *Human capital* is the predicted federal funding for each dissertation's advisors, weighted by the maximum textual similarity between the dissertation and a focal firm's patents granted in a 5-year time cohort, as follows:

$$\textit{Predicted R\&D budget - human capital}_{i,t} = \sum_{d \in D} \textit{Maximum textual similarity}_{d,i,t} \left(\sum_{a \in A} \widehat{R\&D\ budget}_{d,a} \times \textit{Share of agency}_{d,a} \right) \tag{8}$$

D is the set of PhD dissertations in the top 1,000 most similar dissertations for one or more of the patents granted to firm i during time cohort t. $Maximum\ textual\ similarity_{d,i,t}$ and A are as previously defined. $\widehat{R\&D\ budget}_{d,a}$ is the R&D budget predicted by *Majority party share* and *Democratness* for agency a at the beginning of the PhD program (i.e., five years prior to dissertation d's defense year). $Share\ of\ agency_{d,a}$ is obtained by dividing the funding amount (in \$) from agency a to the publications of the advisor(s) of dissertation d during the 6-year period ending in dissertation d's defense year by the total funding amount (in \$) from agency a to any publication published over the same period.

6 Estimation Results

6.1 Patents Equation

Table 5 presents the results using patents—our measure of corporate invention—as the dependent variable. Columns 1, 3, and 5 present OLS estimates for *Public invention*, *Human capital*, and *Public knowledge*, respectively. The coefficients are positive and statistically different from zero (p-values < 0.001). However, common shocks can affect both public science and corporate R&D, leading to biased OLS estimates. We address this concern in a 2SLS framework by instrumenting for *Public invention* using *Predicted R&D budget - public invention*, for *Human capital* using *Predicted R&D budget - human capital*, and for *Public knowledge* using *Predicted R&D budget - public knowledge.* The first stage results reported in Appendix Table A10 confirm that all components of public science are positively related to their respective instrumental variables (p-values < 0.001, F statistics > 104.7, see Lee, McCrary, Moreira, and Porter (2022)).

The 2SLS coefficient estimate on public invention becomes negative (Column 2, p-value < 0.001), while the estimate on human capital becomes even larger (Column 4, p-value < 0.001). Importantly, the negative effect of public invention and the positive effect of human capital persist when they are jointly estimated on the entire sample (Column 7) or a subsample of publishing firms (Column 8).[19] At the sample means, a one standard deviation increase in relevant public invention decreases company patents by 51%, while a one standard deviation increase in relevant human capital increases patents by 53% (Column 7).[20]

Conversely, the 2SLS estimate on public knowledge is small when estimated alone (Col-

[19] These results are robust to dropping the controls for $Agency\ exposure_t$ and $ln(Awards\ to\ focal\ firm)_{t-1}$. They are also robust to using an inverse hyperbolic sine transformation of the dependent variable.

[20] Average values for patent flow, public invention stock, and human capital are 28.30, 266.11, and 6,412.67, respectively. The standard deviation for public invention is 511.89 and for human capital is 9,708.77. The marginal effect of a one standard deviation increase in public invention is a decrease in firm patents of $511.89 \times 0.256(28.30 + 1)/(266.11 + 1) = 14.37$. The marginal effect of a one standard deviation increase in relevant human capital is an increase in firm patents of $9,708.77 \times 0.338(28.30 + 1)/(6,412.67 + 1) = 14.99$.

umn 6, p-value < 0.05) and becomes statistically indistinguishable from zero when estimated jointly (Columns 7 and 8). In light of our theoretical predictions from Table 1, finding no effect of public knowledge on corporate patents suggests that public knowledge does not lower the cost of invention. In turn, this also implies that public knowledge does not complement internal research in lowering the cost of invention.

To address concerns with our ln(1+x) transformation of the dependent variable, we implement the two-stage control function (CF) instrumental variable (IV) Poisson regression approach of Lin and Wooldridge (2019). We correct the standard errors in the second stage using panel bootstrapping with 100 replications and report results in Column 9. Our results are not sensitive to our preferred data transformation.

Table 5: Main Effect of Relevant Public Science on Company Patents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable:	$\ln(1+\text{Patents})_t$								Patents_t
	OLS	2SLS	OLS	2SLS	OLS	2SLS	2SLS	2SLS (Pub. firms)	2-Stage CF IV Poisson
$\ln(1+\text{Public invention})_{t-1}$	0.018***	-0.139***					-0.256***	-0.347***	-0.698***
	(0.005)	(0.029)					(0.035)	(0.053)	(0.187)
$\ln(1+\text{Human capital})_{t-1}$			0.024***	0.215***			0.338***	0.451***	1.062***
			(0.003)	(0.023)			(0.033)	(0.048)	(0.158)
$\ln(1+\text{Public knowledge})_{t-1}$					0.008***	0.033*	0.021	-0.005	-0.055
					(0.002)	(0.013)	(0.012)	(0.012)	(0.036)
$\ln(\$1+\text{R\&D stock})_{t-1}$	0.294***	0.323***	0.285***	0.237***	0.294***	0.288***	0.230***	0.270***	0.337**
	(0.020)	(0.023)	(0.020)	(0.018)	(0.020)	(0.020)	(0.019)	(0.025)	(0.101)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	28.30	28.30	28.30	28.30	28.30	28.30	37.83	31.02
Weak id. (Kleibergen-Paap)		454.39		743.52		440.16	116.09	65.87	
Firms	3,372	3,372	3,372	3,372	3,372	3,372	3,372	2,194	2,900
Observations	41,698	41,698	41,698	41,698	41,698	41,698	41,698	30,708	38,036
Adjusted R-squared	0.86	0.03	0.86	0.11	0.86	0.09	-0.04	-0.07	.

Notes: This table presents the estimation results using corporate patents as the dependent variable. The sample in Column 8 is restricted to publishing firms. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms. Column 9 reports estimates from a two-stage control function (CF) instrumental variable (IV) Poisson regression (Lin & Wooldridge, 2019). Standard errors are estimated by panel bootstrapping with 100 replications.

6.2 Publications Equation

Table 6 presents the results using corporate publications—our first measure of corporate internal research—as the dependent variable. Similar to the results for patents, after instrumenting, we estimate a negative and significant effect for public invention (Column 2, p-value < 0.001) and a positive and significant effect for human capital (Column 4, p-value $<$

0.001). These results persist when we jointly estimate them (Column 7), restrict the sample to publishing firms (Column 8), or use Poisson estimation (Column 9).[21] At the sample means, a one standard deviation increase in university invention decreases company publications by 33%, while a one standard deviation increase in relevant human capital increases publications by 22% (Column 7).[22]

Similar to the results for patents, the estimated effect of public knowledge on publications is not statistically different from zero (Columns 6-9), suggesting that knowledge that is not embodied in either people or inventions has little effect on corporate research as well.[23]

Table 6: Main Effect of Relevant Public Science on Company Publications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable:	$\ln(1+\text{Publications})_t$								Publications_t
	OLS	2SLS	OLS	2SLS	OLS	2SLS	2SLS	2SLS (Pub. firms)	2-Stage CF IV Poisson
$\ln(1+\text{Public invention})_{t-1}$	0.004	-0.108***					-0.162***	-0.252***	-0.705***
	(0.004)	(0.020)					(0.025)	(0.040)	(0.184)
$\ln(1+\text{Human capital})_{t-1}$			0.004	0.056***			0.139***	0.192***	0.561***
			(0.002)	(0.015)			(0.021)	(0.032)	(0.105)
$\ln(1+\text{Public knowledge})_{t-1}$					0.006***	-0.011	-0.008	-0.019	-0.087
					(0.002)	(0.010)	(0.010)	(0.010)	(0.062)
$\ln(\$1+\text{R\&D stock})_{t-1}$	0.179***	0.199***	0.178***	0.165***	0.176***	0.180***	0.163***	0.218***	0.608***
	(0.016)	(0.017)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.022)	(0.077)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	15.72	15.72	15.72	15.72	15.72	15.72	15.72	21.35	21.35
Weak id. (Kleibergen-Paap)		454.39		743.52		440.16	116.09	65.87	
Firms	3,372	3,372	3,372	3,372	3,372	3,372	3,372	2,194	2,194
Observations	41,698	41,698	41,698	41,698	41,698	41,698	41,698	30,708	30,708
Adjusted R-squared	0.88	0.01	0.88	0.06	0.88	0.05	-0.04	-0.09	.

Notes: This table presents estimation results for corporate publications. The sample in Column 8 is restricted to publishing firms. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms. Column 9 reports estimates from a two-stage control function (CF) instrumental variable (IV) Poisson regression (Lin & Wooldridge, 2019). Standard errors are estimated by panel bootstrapping with 100 replications.

[21] These results are also robust to dropping the controls for $Agency\ exposure_t$ and $ln(Awards\ to\ focal\ firm)_{t-1}$, and to using an inverse hyperbolic sine transformation of the dependent variable.

[22] Average values for publication flow, public invention stock, and human capital are 15.72, 266.11, and 6,412.67, respectively. The standard deviations for public invention and human capital are 511.89 and 9,708.77, respectively. The marginal effect of a one standard deviation increase in public invention is a decrease in firm publications of $511.89 \times 0.162(15.72 + 1)/(266.11 + 1) = 5.19$. The marginal effect of a one standard deviation increase in human capital is an increase in firm publications of $9,708.77 \times 0.139(15.72 + 1)/(6,412.67 + 1) = 3.52$.

[23] The consistency with the zero effect on patenting is gratifying. Even if public knowledge did not *directly* affect the marginal return to corporate research, if it increased patenting by the firm, it would *indirectly* increase the marginal return to research.

6.3 AMWS Scientists Equation

Table 7 presents the estimation results using firm employment of AMWS scientists—our second measure of corporate internal research—as the dependent variable. The patterns are very similar to those obtained using publications. Taken together, the results in Columns 5-9 indicate that relevant public knowledge has very little effect on company employment of renowned scientists. We find a negative effect for public invention (Column 7, p-value $<$ 0.05) and a positive effect for human capital (Column 7, p-value $<$ 0.001).

Evaluated at the sample means, the 2SLS estimates in Column 7 indicate that a one standard deviation increase in relevant public invention decreases employment of AMWS scientists by 8%, while a one standard deviation increase in relevant human capital increases employment of AMWS scientists by 9%.[24]

Table 7: Main Effect of Relevant Public Science on Company Employment of AMWS Scientists

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable:	ln(1+AMWS scientists)$_t$								AMWS scientists$_t$
	OLS	2SLS	OLS	2SLS	OLS	2SLS	2SLS	2SLS (Pub. firms)	2-Stage CF IV Poisson
ln(1+Public invention)$_{t-1}$	0.005*	-0.022					-0.033*	-0.069**	-0.384
	(0.003)	(0.012)					(0.015)	(0.024)	(0.214)
ln(1+Human capital)$_{t-1}$			0.006***	0.034***			0.047***	0.067***	0.292*
			(0.002)	(0.010)			(0.013)	(0.019)	(0.137)
ln(1+Public knowledge)$_{t-1}$					0.001	0.018**	0.015*	0.005	-0.020
					(0.001)	(0.007)	(0.006)	(0.006)	(0.039)
ln($1+R&D stock)$_{t-1}$	0.046***	0.051***	0.044***	0.036***	0.047***	0.044***	0.035***	0.044***	0.175**
	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.012)	(0.063)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	4.80	4.80	4.80	4.80	4.80	4.80	4.80	6.48	10.15
Weak id. (Kleibergen-Paap)		454.39		743.52		440.16	116.09	65.87	
Firms	3,372	3,372	3,372	3,372	3,372	3,372	3,372	2,194	1,321
Observations	41,698	41,698	41,698	41,698	41,698	41,698	41,698	30,708	19,710
Adjusted R-squared	0.93	0.01	0.93	0.02	0.93	0.01	0.00	-0.01	.

Notes: This table presents estimation results for corporate employment of AMWS scientists. The sample in Column 8 is restricted to publishing firms. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms. Column 9 reports estimates from a two-stage control function (CF) instrumental variable (IV) Poisson regression (Lin & Wooldridge, 2019). Standard errors are estimated by panel bootstrapping with 100 replications.

[24] Average values for the number of AMWS scientists employed, public invention stock, and human capital are 4.80, 266.11, and 6,412.67, respectively. The standard deviations for public invention and human capital are 511.89 and 9,708.77, respectively. The marginal effect of a one standard deviation increase in public invention is a decrease in AMWS scientists employed of $511.89 \times 0.033(4.80 + 1)/(266.11 + 1) = 0.37$. The marginal effect of a one standard deviation increase in human capital is an increase in AMWS scientists employed of $9,708.77 \times 0.047(4.80 + 1)/(6,412.67 + 1) = 0.41$.

6.4 R&D Expenditures Equation

Table 8 presents the estimation results for company R&D expenditures. Consistent with the previous three tables, the 2SLS estimates show a negative, though only marginally significant, effect of public invention (Column 7, p-value = 0.066), a positive and significant effect of human capital (Column 7, p-value < 0.01), and no effect of public knowledge (Columns 6-8). Evaluated at the sample means, the 2SLS estimates in Column 7 indicate that a one standard deviation increase in relevant public invention decreases company R&D expenditures by 38%, while a one standard deviation increase in relevant human capital increases them by 33%.[25]

Table 8: Main Effect of Relevant Public Science on Company R&D Expenditures

Dependent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln($1+R&D expenditures)$_t$							
	OLS	2SLS	OLS	2SLS	OLS	2SLS	2SLS	2SLS (Pub. firms)
ln(1+Public invention)$_{t-1}$	0.039	-0.078					-0.203	-0.141
	(0.020)	(0.085)					(0.111)	(0.117)
ln(1+Human capital)$_{t-1}$			0.042**	0.137**			0.225**	0.195*
			(0.014)	(0.052)			(0.078)	(0.080)
ln(1+Public knowledge)$_{t-1}$					0.015**	0.025	0.014	-0.017
					(0.005)	(0.031)	(0.031)	(0.031)
ln($1+Sales)$_{t-1}$	0.192***	0.193***	0.191***	0.189***	0.191***	0.191***	0.187***	0.169***
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.023)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	142.59	142.59	142.59	142.59	142.59	142.59	142.59	184.41
Weak id. (Kleibergen-Paap)		443.09		723.10		401.23	81.87	48.71
Firms	3,162	3,162	3,162	3,162	3,162	3,162	3,162	2,120
Observations	36,584	36,584	36,584	36,584	36,584	36,584	36,584	27,919
Adjusted R-squared	0.79	0.04	0.79	0.05	0.79	0.04	0.03	0.05

Notes: This table presents estimation results for corporate R&D expenditures. The sample in Column 8 is restricted to publishing firms. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

In summary, our key findings thus far are that (1) public invention, as measured by the stock of university patents, has a negative effect on corporate innovation, whereas (2) human capital, as measured by trained PhD scientists, has a positive effect, and (3) abstract public knowledge, not embodied in either people or inventions, has no effect.

[25] Average values for R&D expenditures, public invention stock, and human capital are 142.59, 287.74, and 6,931.84, respectively. The standard deviations for public invention and human capital are 527.80 and 10,118.23, respectively. The marginal effect of a one standard deviation increase in public invention is a decrease in R&D expenditures of $527.80 \times 0.203(142.59 + 0.000001)/(287.74 + 1) = 54.02$. The marginal effect of a one standard deviation increase in human capital is an increase in AMWS scientists employed of $10,118.23 \times 0.225(142.59 + 0.000001)/(6,931.84 + 1) = 46.82$.

6.5 Heterogeneous Effects: Frontier Firms vs. Follower Firms

Frontier firms may differ from followers in the type of inventions they produce, the value they derive from inventions, or both. To capture a firm's proximity to the technology frontier, we first count its annual flow of novel patents, where patent novelty is based on unique IPC combinations. Then, we create *Tech frontier* as an indicator variable equal to 1 for firm years with novel patents in the top decile compared to other sample firms in that year, and 0 otherwise.[26] We interact this indicator variable with our measures of *Public invention* and *Human capital*, respectively, and report second-stage 2SLS results in Table 9.[27]

The coefficient estimates on the interaction terms show substantial heterogeneity in the effect of public science on internal research and invention based on firm proximity to the technology frontier. While Tables 5, 6, and 7 show that, on average, firms respond to an increase in relevant public invention by withdrawing from patenting, publishing, and hiring of AMWS scientists, firms operating on the technology frontier do so to a lesser extent. Similarly, though both frontier firms and followers increase their patenting, publishing, and hiring in response to an increase in the supply of relevant human capital, frontier firms do so to a greater extent. We find similar results when we measure proximity to the technology frontier using patents that are first to be granted in a new CPC main group or subgroup (see Appendix Table B14).

To further explore these results, we capture a firm's ability to derive value from inventions using the average patent value from Kogan, Papanikolaou, Seru, and Stoffman (2017) normalized by market value. The indicator variable *High ability* equals 1 for firm years with average patent values in the top decile compared to other sample firms in that year and 0 otherwise. Table 10 reports the second stage of 2SLS estimation using the same instrumental variables as before. Unlike the results for firm proximity to the technology frontier, the coefficient estimates on the interaction terms are no longer significantly different from zero across specifications. A firm's ability to derive private value from inventions does not condition its response to relevant public science. In other words, the impact of public science on corporate innovation is more likely to be influenced by technological leadership than by an advantage in product markets.

Our results are consistent with the view that firms on the technology frontier may have more productive internal research or that these firms operate in technologies where public

[26] Appendix Table C19 shows the results of a mean comparison test of frontier firms versus followers. Frontier firms appear to have higher stocks of public knowledge and human capital than followers, but lower stocks of public invention.

[27] We use *Predicted R&D budget - public invention*, *Predicted R&D budget - human capital*, and their interactions with *Tech frontier* as instrumental variables for *Public invention*, *Human capital*, and their interactions with *Tech frontier*, respectively.

Table 9: Variation by Firm Proximity to the Technology Frontier: Unique IPC Combinations

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	$\ln(1{+}\text{Pat.})_t$	$\ln(1{+}\text{Pub.})_t$	$\ln(1{+}\text{AMWS sci.})_t$	$\ln(1{+}\text{Pat.})_t$	$\ln(1{+}\text{Pub.})_t$	$\ln(1{+}\text{AMWS sci.})_t$
$\ln(1{+}\text{Public invention})_{t-1} \times \textit{Tech frontier}_t$	0.236***	0.083***	0.043***			
	(0.010)	(0.009)	(0.008)			
$\ln(1{+}\text{Human capital})_{t-1} \times \textit{Tech frontier}_t$				0.125***	0.045***	0.026***
				(0.008)	(0.008)	(0.006)
$\ln(1{+}\text{Public invention})_{t-1}$	-0.216***	-0.141***	-0.033*	-0.221***	-0.143***	-0.033*
	(0.031)	(0.023)	(0.014)	(0.030)	(0.023)	(0.014)
$\ln(1{+}\text{Human capital})_{t-1}$	0.250***	0.100***	0.037**	0.264***	0.105***	0.038***
	(0.028)	(0.019)	(0.012)	(0.028)	(0.019)	(0.011)
$\ln(\$1{+}\text{R\&D stock})_{t-1}$	0.206***	0.153***	0.031***	0.201***	0.151***	0.029**
	(0.016)	(0.015)	(0.009)	(0.016)	(0.015)	(0.009)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	15.72	4.80	28.30	15.72	4.80
Weak id. (Kleibergen-Paap)	122.62	122.62	122.62	124.57	124.57	124.57
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents the second stage of 2SLS estimation for the effect of public invention and human capital on corporate patents, publications, and AMWS scientists when considering firm proximity to the technology frontier. To measure this proximity, we first count each firm's annual flow of novel patents, where patent novelty is based on unique IPC combinations. Then, we create the variable *Tech frontier* as an indicator equal to 1 for firm years with a flow of novel patents in the top decile compared to other sample firms in that year, and 0 otherwise. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

invention is less plentiful but with abundant supplies of human capital. In either case, it would result in frontier firms having a larger scale of internal research and invention. In turn, frontier firms would be more responsive to increases in human capital but less responsive to public invention.

6.6 Variation by Industry

Our sample includes firms from a diverse set of industries. Appendix Table C16 provides summary statistics by main industry, defined based on the firm's primary SIC4 code. Annual patent flows range from 14 for firms operating primarily in *Machinery, equipment, and systems* to 55 for firms operating primarily in *Computer, IT, and software*. The average number of AMWS scientists per firm ranges from 1 in *Machinery, equipment, and systems* to 11 in *Life sciences*. The most striking differences in terms of relevant public science appear in *Life sciences*, where firms have, on average, much higher stocks of relevant public knowledge and university patents.

We examine variation in the effect of public science on corporate patents and publications by main industry. Table 11 presents estimates from the second stage of 2SLS regressions using our preferred instrumental variables and their interactions with industry indicator variables.

Table 10: Variation by Firm Ability to Derive Value from Inventions: Patent Value

Dependent variable:	(1) $\ln(1{+}\text{Pat.})_t$	(2) $\ln(1{+}\text{Pub.})_t$	(3) $\ln(1{+}\text{AMWS sci.})_t$	(4) $\ln(1{+}\text{Pat.})_t$	(5) $\ln(1{+}\text{Pub.})_t$	(6) $\ln(1{+}\text{AMWS sci.})_t$
$\ln(\text{Public invention})_{t-1} \times \textit{High ability}_t$	0.040***	-0.003	-0.000			
	(0.003)	(0.002)	(0.001)			
$\ln(\text{Human capital})_{t-1} \times \textit{High ability}_t$				0.008	-0.007	-0.005**
				(0.005)	(0.004)	(0.002)
$\ln(1{+}\text{Public invention})_{t-1}$	-0.259***	-0.157***	-0.041**	-0.261***	-0.158***	-0.042**
	(0.035)	(0.025)	(0.015)	(0.035)	(0.025)	(0.015)
$\ln(1{+}\text{Human capital})_{t-1}$	0.349***	0.135***	0.055***	0.347***	0.135***	0.055***
	(0.033)	(0.021)	(0.013)	(0.033)	(0.021)	(0.013)
$\ln(\$1{+}\text{R\&D stock})_{t-1}$	0.230***	0.163***	0.036***	0.233***	0.163***	0.036***
	(0.019)	(0.016)	(0.009)	(0.019)	(0.016)	(0.009)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	15.72	4.80	28.30	15.72	4.80
Weak id. (Kleibergen-Paap)	123.47	123.47	123.47	123.39	123.39	123.39
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents the second stage of 2SLS estimation for the effect of public invention and human capital on corporate patents, publications, and AMWS scientists when considering firm ability to derive value from inventions. To measure this ability, we first calculate the average patent value from Kogan et al. (2017), normalized by market value, for each firm year. Then, we create the variable *High ability* as an indicator equal to 1 for firm years with an average patent value in the top decile compared to other sample firms in that year, and 0 otherwise. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

Our analysis reveals that the positive effect of human capital on firm patents and publications is robust across all industries. The negative effect of public invention is robust across all industries except *Life sciences*. In *Life sciences*, public knowledge complements internal research in reducing the cost of inventing, while external and internal inventions are strategic complements. This is consistent with incumbent firms collaborating with universities and investing in or acquiring startups to complete downstream development and commercialize the resulting products (Arora, Fosfuri, & Gambardella, 2001; Azoulay et al., 2019).

6.7 Firm Value Equation

Table 12 presents the estimation results for Tobin's Q, our measure of firm value. Columns 1-3 focus on the main effect of public science on firm value. Public invention has a negative and significant effect (p-values < 0.001), while human capital has a positive effect that is imprecisely estimated.[28] Interpreted in light of our theoretical predictions from Table 1, the negative effect of public invention suggests that university patents compete with firm inventions more than they serve as inputs into corporate innovation. When we consider

[28] We obtain a positive and significant estimate on human capital when using our alternative instrumental variables.

Table 11: Variation by Main Industry

Dependent variable:	(1) ln(1+Pat.)$_t$	(2) ln(1+Pub.)$_t$	(3) ln(1+Pat.)$_t$	(4) ln(1+Pub.)$_t$	(5) ln(1+Pat.)$_t$	(6) ln(1+Pub.)$_t$
ln(1+Public invention)$_{t-1}$	-0.165***	-0.145***				
	(0.042)	(0.031)				
$\times$ *Computer, IT, software*$_t$	0.032	0.010				
	(0.073)	(0.040)				
$\times$ *Electronics, semicond.*$_t$	0.152**	0.058				
	(0.054)	(0.040)				
$\times$ *Machinery, equipment, sys.*$_t$	-0.088*	0.050				
	(0.039)	(0.035)				
$\times$ *Life sciences*$_t$	0.015	0.150*				
	(0.059)	(0.061)				
$\times$ *Telecommunication*$_t$	0.123	-0.006				
	(0.136)	(0.048)				
$\times$ *Transportation*$_t$	0.114	-0.062				
	(0.096)	(0.044)				
ln(1+Human capital)$_{t-1}$			0.189***	0.042*		
			(0.024)	(0.017)		
$\times$ *Computer, IT, software*$_t$			0.030	0.005		
			(0.017)	(0.013)		
$\times$ *Electronics, semicond.*$_t$			0.067**	0.008		
			(0.023)	(0.016)		
$\times$ *Machinery, equipment, sys.*$_t$			0.017	0.044**		
			(0.020)	(0.017)		
$\times$ *Life sciences*$_t$			0.021	0.046**		
			(0.014)	(0.017)		
$\times$ *Telecommunication*$_t$			0.058	0.002		
			(0.034)	(0.018)		
$\times$ *Transportation*$_t$			0.028	-0.018		
			(0.036)	(0.034)		
ln(1+Public knowledge)$_{t-1}$					0.037*	-0.018
					(0.014)	(0.012)
$\times$ *Computer, IT, software*$_t$					0.025	0.015
					(0.018)	(0.012)
$\times$ *Electronics, semicond.*$_t$					0.030*	-0.002
					(0.013)	(0.008)
$\times$ *Machinery, equipment, sys.*$_t$					0.000	0.006
					(0.011)	(0.008)
$\times$ *Life sciences*$_t$					-0.030***	0.024*
					(0.009)	(0.011)
$\times$ *Telecommunication*$_t$					0.019	-0.000
					(0.020)	(0.011)
$\times$ *Transportation*$_t$					-0.013	-0.001
					(0.013)	(0.010)
ln($1+R&D stock)$_{t-1}$	0.314***	0.199***	0.239***	0.167***	0.286***	0.180***
	(0.023)	(0.017)	(0.018)	(0.016)	(0.020)	(0.016)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	15.72	28.30	15.72	28.30	15.72
Weak id. (Kleibergen-Paap)	19.82	19.82	107.54	107.54	62.85	62.85
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents the second stage of 2SLS estimation for the effect of relevant public science on corporate patents and publications by main industry. Industry classification is based on a firm's primary SIC4 code. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

heterogeneity in this effect by firm proximity to the technology frontier, we find that frontier firms are better positioned to compete with university-backed startups in the product market compared to followers (Columns 4 and 5).

Our results also indicate that increases in public knowledge reduce, not increase, value for incumbent firms. While we leave for future work a careful examination of this negative effect on market value, a potential direction would build on the idea that public knowledge is available for all firms to exploit. If the average incumbent firm is poorly positioned to exploit that knowledge relative to university-backed startups, the negative effect may be attributed to rent-dissipating competition between incumbents and startups in the technology market. That is, our results suggest that, insofar as public knowledge creates value, it is captured by startups and other private firms, at the expense of incumbent public firms.

Table 12: Firm Value Equation

	(1)	(2)	(3)	(4)	(5)
Dependent variable:			ln(Tobin's Q)$_t$		
	Baseline	Add Public knowledge	Add Human capital	Baseline with interaction	Add Public knowledge and Human capital
ln(1+Public invention)$_{t-1}$	-0.174***	-0.159***	-0.218***	-0.174***	-0.216***
	(0.030)	(0.030)	(0.052)	(0.030)	(0.051)
ln(1+Public knowledge)$_{t-1}$		-0.024	-0.037*		-0.041*
		(0.017)	(0.018)		(0.018)
ln(1+Human capital)$_{t-1}$			0.017		0.010
			(0.043)		(0.042)
ln(1+Public invention)$_{t-1} * Techfrontier_t$				0.108**	0.099**
				(0.038)	(0.037)
Tech frontier$_t$				-0.574**	-0.530**
				(0.177)	(0.172)
R&D stock$_{t-1}$ / $Assets_t$	0.218***	0.218***	0.220***	0.218***	0.220***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Year FE	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Mean DV	33.80	33.80	33.80	33.80	33.80
Weak id. (Kleibergen-Paap)	634.98	195.90	100.40	315.91	74.86
Firms	3,230	3,230	3,230	3,230	3,230
Observations	36,718	36,718	36,718	36,718	36,718

Notes: This table presents the second stage of 2SLS estimation for the effect of public science on firm value, measured using Tobin's Q. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

6.8 Robustness Checks

To probe the robustness of our findings, we perform several checks. First, we use alternative instrumental variables to estimate the effect of public science on corporate innovation. This allows us to assess the extent to which our findings are sensitive to potential violations of the key identifying assumptions. Second, we examine the impact of alternative measures of public invention and human capital on our main results. This allows us to assess the extent to which our findings are dependent on the specific measures used in our analysis. Third, in universities, PhD training happens as a part of the research process. We separate public invention from human capital to determine whether the two factors indeed have independent effects on corporate innovation outcomes. Fourth, we use measures of high-quality corporate innovation as dependent variables to test whether the effects of public invention and human capital are consistent across different measures of corporate innovation.

6.8.1 Alternative Instrumental Variables

We explore the sensitivity of our main results by using the R&D budgets of federal agencies, instead of the political composition-predicted R&D budgets, in constructing alternative instrumental variables. Appendix Tables A9 and B11 present the first and second stages of 2SLS estimation. We find that the negative (positive) effect of public invention (human capital) on corporate innovation persists.

6.8.2 Alternative Measures of Public Science

We check the robustness of our main results by using alternative measures of public invention and human capital. *Public invention, broad* is the stock of non-corporate publications that are cited by patents in various CPC subclasses, weighted by the firm's lagged patenting shares across CPC subclasses. *Human capital, cited* is a firm-year measure of published PhD dissertations cited by patents in various CPC subclasses, weighted by the firm's lagged patenting shares across CPC subclasses. *Human capital, OECD* is a firm-year measure of PhD dissertations in various OECD natural science subfields, weighted by the reliance of CPC subclasses on science published in various OECD subfields and by the firm's lagged patenting shares across CPC subclasses. Details about the construction of these measures are included in Appendix A.

Table 13 presents the second stage of 2SLS estimation. We find results consistent with Tables 5 and 6. As one might expect, the use of a broad measure of public invention reduces the power of the instrument, resulting in noisier but qualitatively similar results. But regardless of how we measure relevant public invention and human capital, the former has a negative and significant effect (statistically and economically) on company patents and publications (p-values < 0.001), while the latter has a positive and significant effect (p-values < 0.001).

Table 13: Alternative Measures of Relevant Public Invention and Human Capital

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	ln(1+Patents)$_t$			ln(1+Publications)$_t$		
	Broad	Cited	OECD	Broad	Cited	OECD
ln(1+Public invention, broad)$_{t-1}$	-0.172***			-0.109***		
	(0.023)			(0.018)		
ln(1+Public invention)$_{t-1}$		-0.196***	-0.517***		-0.150***	-0.366***
		(0.036)	(0.120)		(0.025)	(0.082)
ln(1+Human capital, cited)$_{t-1}$		0.330***			0.250***	
		(0.057)			(0.042)	
ln(1+Human capital, OECD)$_{t-1}$			0.433***			0.260***
			(0.099)			(0.068)
ln(1+Human capital)$_{t-1}$	0.304***			0.118***		
	(0.030)			(0.019)		
ln(1+Public knowledge)$_{t-1}$	0.012	0.049***	0.061***	-0.014	0.004	0.011
	(0.013)	(0.013)	(0.015)	(0.010)	(0.010)	(0.011)
ln($1+R&D stock)$_{t-1}$	0.234***	0.292***	0.310***	0.165***	0.185***	0.197***
	(0.019)	(0.021)	(0.026)	(0.017)	(0.016)	(0.019)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	28.30	28.30	15.72	15.72	15.72
Weak id. (Kleibergen-Paap)	115.51	156.30	21.58	115.51	156.30	21.58
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents estimates from the second stage of 2SLS regressions using alternative measures of firm-relevant public invention and human capital. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

6.8.3 Separating Public Invention From Human Capital

We explore the sensitivity of our main results to separating public invention from human capital in different ways. We construct alternative measures of *Public invention, broad* by including only patent-cited publications from federal laboratories, excluding any published PhD dissertations, excluding publications coauthored by PhD students, and excluding publications coauthored by the advisors of PhD students, respectively. We report results from the second stage of 2SLS estimation in Appendix Table B12. We find that the effects of public invention and human capital on corporate innovation are similar to those reported earlier. For instance, the elasticity of corporate patenting with respect to public invention ranges from -0.174 to -0.264, similar to -0.256, the elasticity reported in Column 7 of Table 5. The elasticity with respect to human capital ranges from 0.309 to 0.326, which is very close to the comparable estimate of 0.338 reported in Column 7 of Table 5. The patterns are similar for corporate publications and the employment of AMWS scientists.

We also account for differences in PhD production intensity across different scientific fields in a three-step process. First, for each of the 15 scientific fields (not including humanities and social sciences) included in the Dimensions Units of Assessment (UOA) classification system, we calculate the ratio between (i) the total funding amount received by publications of PhD students and their advisors published in the field and (ii) the total funding amount received by all publications published in the field. Second, we categorize fields with above (below) median ratios as having high (low) student-advisor funding. Third, we construct alternative measures of *Public invention, broad* by including only publications from fields with high or low student-advisor funding ratios, respectively. As shown in Appendix Table B12, the effects of public invention and human capital on corporate R&D are not sensitive to these permutations.

6.8.4 High-Quality Corporate Innovation

We report additional robustness checks using different measures of high-quality corporate innovation in Appendix Table B13. We use two criteria to define invention quality: "home-run" patents, which rank in the top 5% of the forward citations distribution per year, and "breakthrough" patents, which rank in the top 1% of the forward citations distribution up to five years after publication, relative to all patents filed the same year and in the same technology field. We also include analyses focusing on publications coauthored by AMWS scientists, publications cited by AMWS scientists, and employment of award-winning AMWS scientists. The results show that our findings are robust across all specifications.

7 Discussion and Conclusion

This paper shows that firms hire fewer scientists and produce fewer patents and publications in response to an increase in relevant public inventions. Conversely, when there is an increase in relevant human capital, firms tend to employ more scientists and produce more patents and publications. However, abstract public knowledge *per se* has very little effect on corporate patenting, publishing, or employment of scientists.

Our study highlights that the impact of public science on corporate innovation depends on its embodied components. While the public inventions represented by university patents appear to compete with corporate innovation, the PhD researchers produced alongside scientific knowledge enhance the payoffs to corporations from internal invention and research. These offsetting effects may result in a relatively small net effect of public science on corporate innovation.[29] The small net effect, however, conceals the diverse ways in which public and private R&D investments interact.

Indeed, firms' response to the increase in public science depends on their proximity to the technology frontier. Frontier firms tend to continue investing in internal research and invention, even in the presence of abundant public science. This is consistent with the observed surge in corporate scientific research in such emerging technology fields as artificial intelligence and quantum computing. The disparity in response may arise because frontier firms enjoy greater marginal returns from using internal research and invention than other firms, or because they operate in technologies where public invention is less abundant but human capital is nevertheless abundantly supplied. Consequently, frontier firms may benefit more from public knowledge and skilled PhDs to fuel their internal research and inventions than followers. On the other hand, firms operating in technologies with more abundant public invention would also tend to cut back on internal research and development. Such firms would naturally benefit less from expansion in the supply of human capital or public knowledge but would be very responsive to changes in public invention.

Our findings also relate to the growing literature on economic growth and productivity slowdown. The sluggish growth in productivity over the last three decades or more in the face of sustained growth in scientific output has puzzled observers. Our findings point to a possible reason. Romer (1990) and Jones (2022) stress that the non-rivalrous nature of ideas is a potent source of increasing returns and productivity growth. It should follow that the most powerful sources of increasing returns are ideas that are broadly usable, and whose production is publicly funded so that they can be placed in the public domain, available to all. This is the basic argument underlying the case for public support for scientific research in universities.

Yet the history of technical progress teaches us that abstract ideas are also difficult to use. Ideas have to be tailored for specific uses, and frequently, have to be embodied in people and artifacts before they can be absorbed by firms. However, such embodiment also makes

[29] Our estimates suggest that the rise in public science between 1986 and 2015 led to an average annual decrease in corporate patents of 1.5% and in corporate publications of 1.1%. Between 1986 and 2015, the stock of university patents relevant to our sample of firms increased by 660.85 (from 21.65 to 682.50), while human capital increased by 4,048.33 (from 5,896.36 to 9,944.69). Using the coefficients from Column 7 in Table 5, we estimate that the increase in university inventions decreased firm patents by $660.85 \times 0.256(28.30+1)/(266.11+1) = 18.56$, while the increase in human capital increased firm patents by $4,048.33 \times 0.338(28.30+1)/(6,412.67+1) = 6.25$. The net effect was a decrease in firm patents of 12.31, which represents a 1.5% decrease per year relative to the average annual patent flow of 28.30. Using the coefficients from Column 7 in Table 6, we estimate that the increase in university inventions decreased firm publications by $660.85 \times 0.162(15.72+1)/(266.11+1) = 6.70$, while the increase in human capital increased firm publications by $4,048.33 \times 0.139(15.72+1)/(6,412.67+1) = 1.47$. The net effect was a decrease in firm publications of 5.23, which represents a 1.1% decrease per year relative to the average annual publication flow of 15.72.

ideas less potent sources of increasing returns, turning non-rival ideas into rival inputs, whose use by rivals is easier to restrict. Our findings confirm that firms, especially those not on the technological frontier, appear to lack the absorptive capacity to use externally supplied ideas unless they are embodied in human capital or inventions. The limit on growth is not the creation of useful ideas but rather the rate at which those ideas can be embodied in human capital and inventions, and then allocated to firms to convert them into innovations. In other words, productivity growth may have slowed down because the potential users—private corporations—lack the absorptive capacity to understand and use those ideas.

The loss of absorptive capacity is partly related to the growing specialization and division of innovative labor in the U.S. economy. Not only do universities and public research institutes produce the bulk of scientific knowledge, but over the past three decades, publicly funded inventions and startups have grown in importance as sources of innovation. Concomitantly, many incumbent firms have substantially withdrawn from performing upstream scientific research. The withdrawal of many companies from upstream scientific research may have reduced their absorptive capacity—their ability to understand and use scientific advances produced by public science. If so, the division of innovative labor between universities and firms, wherein the former produce knowledge and the latter apply the knowledge to invent, appears to work much better for frontier firms. Non-frontier firms instead require universities or startups to convert ideas into inventions. The growing specialization involving universities, startups, and incumbents may therefore pose a challenge to maintaining a diverse and vibrant innovation ecosystem. The expansion of public science may widen the gap between frontier firms and followers, with ramifications for product market competition, as well as for the rate and direction of technical progress.

References

Aldridge, T. T., & Audretsch, D. (2017). The Bayh-Dole Act and scientist entrepreneurship. In *Universities and the entrepreneurial ecosystem* (pp. 57–66). Edward Elgar Publishing.

American Association for the Advancement of Science. (2021). *Historical Trends in Federal R&D.* (Available at https://www.aaas.org/programs/r-d-budget-and-policy/historical-trends-federal-rd. Accessed December 6, 2021.)

Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2020). Inside job or deep impact? extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, *58*(1), 3–52.

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate R&D. *Strategic Management Journal*, *39*(1), 3–32.

Arora, A., Belenzon, S., & Sheer, L. (2021a). Knowledge spillovers and corporate investment in scientific research. *American Economic Review*, *111*(3), 871–98.

Arora, A., Belenzon, S., & Sheer, L. (2021b). Matching patents to compustat firms, 1980-2015: Dynamic reassignment, name changes, and ownership structures. *Research Policy*, *50*(5), 104217.

Arora, A., Fosfuri, A., & Gambardella, A. (2001). Markets for technology and their implications for corporate strategy. *Industrial and Corporate Change*, *10*(2), 419–451.

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

Azoulay, P., Graff Zivin, J. S., Li, D., & 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.

Babina, T., He, A. X., Howell, S. T., Perlman, E. R., & Staudt, J. (2023). Cutting the innovation engine: How federal funding shocks affect university patenting, entrepreneurship, and publications. *The Quarterly Journal of Economics*.

Baruffaldi, S., & Poege, F. (2020). A firm scientific community: Industry participation and knowledge diffusion. *Max Planck Institute for Innovation & Competition Research Paper*(20-10).

Beise, M., & Stahl, H. (1999). Public research and industrial innovations in germany. *Research Policy*, *28*(4), 397–422.

Belenzon, S., & Cioaca, L. C. (2021). Guaranteed markets and corporate scientific research. *National Bureau of Economic Research Working Paper*(w28644).

Belenzon, S., & Schankerman, M. (2013). Spreading the word: Geography, policy, and knowledge spillovers. *Review of Economics and Statistics*, *95*(3), 884–903.

Bellet, C. S., De Neve, J.-E., & Ward, G. (2023). Does employee happiness have an impact on productivity? *Management Science*.

Bloomberg Government. (2023). *Your guide to navigating the federal budget process.* (Available at https://about.bgov.com/brief/your-guide-to-navigating-the-federal-budget-process/. Accessed September 30, 2023.)

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Pro-*

ceedings of the 58th annual meeting of the association for computational linguistics (pp. 2270–2282).

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

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: The influence of public research on industrial R&D. *Management Science*, *48*(1), 1–23.

David, P. A., Hall, B. H., & Toole, A. A. (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy*, *29*(4-5), 497–529.

Davis, O. A., Dempster, M. A. H., & Wildavsky, A. (1966). A theory of the budgetary process. *American Political Science Review*, *60*(3), 529–547.

Delron, J.-M., Guellec, D., Wu, C., & Liu, J. (2022). *Building a corpus of patents-articles siblings.* (Available at https://conference.nber.org/conf_papers/f176403.slides.pdf (Accessed March 2023))

Digital Science. (2022). *The data in Dimensions.* (Retrieved from https://www.dimensions.ai/dimensions-data/ on April 12, 2022.)

Dimos, C., & Pugh, G. (2016). The effectiveness of R&D subsidies: A meta-regression analysis of the evaluation literature. *Research Policy*, *45*(4), 797–815.

Einiö, E. (2014). R&D subsidies and company performance: Evidence from geographic variation in government funding based on the ERDF population-density rule. *Review of Economics and Statistics*, *96*(4), 710–728.

Epp, D. A., Lovett, J., & Baumgartner, F. R. (2014). Partisan priorities and public budgeting. *Political Research Quarterly*, *67*(4), 864–878.

Fabrizio, K. R., & Di Minin, A. (2008). Commercializing the laboratory: Faculty patenting and the open science environment. *Research Policy*, *37*(5), 914–931.

Fleming, L., Greene, H., Li, G., Marx, M., & Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, *364*(6446), 1139–1141.

González, X., Jaumandreu, J., & Pazó, C. (2005). Barriers to innovation and subsidy effectiveness. *RAND Journal of Economics*, 930–950.

Goolsbee, A. (1998). Does government R&D policy mainly benefit scientists and engineers? *American Economic Review*, *88*(2), 298–302.

Hartmann, P., & Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries*, *6*(3), 359–381.

Hausman, N. (2022). University innovation and local economic growth. *Review of Economics and Statistics*, *104*(4), 718–735.

Hernandez, D., & King, R. (2016). *Universities' AI talent poached by tech giants.* (Retrieved from https://www.wsj.com/articles/universities-ai-talent-poached-by-tech-giants-1479999601.)

Jones, C. I. (2022). The past and future of economic growth: A semi-endogenous perspective. *Annual Review of Economics*, *14*, 125–152.

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

Kim, S. D., & Moser, P. (2021). *Women in science. lessons from the baby boom* (Tech. Rep.). National Bureau of Economic Research.

Klevorick, A. K., Levin, R. C., Nelson, R. R., & Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. *Research Policy*, *24*(2), 185–205.

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

Laursen, K., & Salter, A. (2004). Searching high and low: What types of firms use universities as a source of innovation? *Research Policy*, *33*(8), 1201–1215.

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2022). Valid t-ratio inference for IV. *American Economic Review*, *112*(10), 3260–3290.

Lichtenberg, F. R. (1984). The relationship between federal contract R&D and company R&D. *American Economic Review*, *74*(2), 73–78.

Lin, W., & Wooldridge, J. M. (2019). Testing and correcting for endogeneity in nonlinear unobserved effects models. In *Panel data econometrics* (pp. 21–43). Elsevier.

Mamuneas, T. P., & Nadiri, M. I. (1996). Public R&D policies and cost behavior of the US manufacturing industries. *Journal of Public Economics*, *63*(1), 57–81.

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

Mansfield, E. (1995). Academic research underlying industrial innovations: Sources, characteristics, and financing. *The Review of Economics and Statistics*, 55–65.

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

McMillan, G. S., Narin, F., & Deeds, D. L. (2000). An analysis of the critical role of public science in innovation: The case of biotechnology. *Research Policy*, *29*(1), 1–8.

Moretti, E., Steinwender, C., & Van Reenen, J. (2021). The intellectual spoils of war? Defense R&D, productivity, and international spillovers. *National Bureau of Economic Research Working Paper*(w26483).

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

Mulligan, K., Lenihan, H., Doran, J., & Roper, S. (2022). Harnessing the science base: Results from a national programme using publicly-funded research centres to reshape firms' R&D. *Research Policy*, *51*(4), 104468.

Myers, K. R., & Lanahan, L. (2022). Estimating spillovers from publicly funded R&D: Evidence from the US Department of Energy. *American Economic Review*, *112*(7), 2393–2423.

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

National Center for Science and Engineering Statistics. (2023a). *National patterns of R&D resources: 2020–21 data update* (Tech. Rep. No. NSF 23-321). National Science Foundation. Retrieved from https://ncses.nsf.gov/pubs/nsf23321

National Center for Science and Engineering Statistics. (2023b). *Science and engineering indicators 2022* (Tech. Rep.). National Science Foundation. Retrieved from https://ncses.nsf.gov/pubs/nsb20225/data#

National Science Board. (1998). *Science and engineering indicators 1998* (Tech. Rep. No. NSB-1998-1). National Science Foundation. Retrieved from https://wayback.archive-it.org/5902/20150627201913/http://www.nsf.gov/

statistics/seind98/

National Science Board. (2010). *Science and engineering indicators 2010* (Tech. Rep. No. NSB-2010-1). National Science Foundation. Retrieved from https://wayback.archive-it.org/5902/20160210151754/http://www.nsf.gov/statistics/seind10/

National Science Board. (2018). *Science and engineering indicators 2018* (Tech. Rep. No. NSB-2018-1). National Science Foundation. Retrieved from https://www.nsf.gov/statistics/2018/nsb20181/

Nelson, R. R. (1986). Institutions supporting technical advance in industry. *The American Economic Review*, *76*(2), 186–189.

OECD. (2003). *Turning science into business: Patenting and licensing at public research organizations.* Retrieved from https://www.oecd-ilibrary.org/content/publication/9789264100244-en doi: https://doi.org/https://doi.org/10.1787/9789264100244-en

Pavitt, K. (1991). What makes basic research economically useful? *Research Policy*, *20*(2), 109–119.

Roche. (2023, May 4). Roche launches Institute of Human Biology to accelerate breakthroughs in R&D by unlocking the potential of human model systems. Retrieved from https://www.roche.com/media/releases/med-cor-2023-05-04

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, *98*(5, Part 2), S71–S102.

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

Rosenberg, N., & Nelson, R. R. (1994). American universities and technical advance in industry. *Research Policy*, *23*(3), 323–348.

Scandura, A. (2016). University-industry collaboration and firms' r&d effort. *Research Policy*, *45*(9), 1907–1922.

Schartinger, D., Rammer, C., Fischer, M. M., & Fröhlich, J. (2002). Knowledge interactions between universities and industry in austria: Sectoral patterns and determinants. *Research Policy*, *31*(3), 303–328.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web* (pp. 243–246).

Szücs, F. (2020). Do research subsidies crowd out private R&D of large firms? Evidence from European framework programmes. *Research Policy*, *49*(3), 103923.

Tartari, V., & Stern, S. (2021). *More than an ivory tower: The impact of research institutions on the quantity and quality of entrepreneurship* (Tech. Rep. No. 28846). National Bureau of Economic Research.

Tavares, J. (2004). Does right or left matter? cabinets, credibility and fiscal adjustments. *Journal of Public Economics*, *88*(12), 2447–2468.

Tether, B. S., & Tajar, A. (2008). Beyond industry–university links: Sourcing knowledge for innovation from consultants, private research organisations and the public science-base. *Research Policy*, *37*(6-7), 1079–1095.

Valero, A., & Van Reenen, J. (2019). The economic impact of universities: Evidence from across the globe. *Economics of Education Review*, *68*, 53–67.

Wallsten, S. J. (2000). The effects of government-industry R&D programs on private R&D: The case of the Small Business Innovation Research program. *The RAND Journal of Economics*, 82–100.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., ... Rogahn, R. (2019). A review of Microsoft Academic Services for science of science studies. *Frontiers in Big Data*, *2*, 45.

Appendix A Data and Variable Construction

A.1 Main Data Sources

We combined data from three main sources: Dimensions, American Men & Women of Science, and ProQuest Dissertations & Theses Global.

A.1.1 Dimensions

Digital Science's Dimensions project (July 31, 2021) provides data on scientific publications, grants, patents, and citations. The data include links between funding organizations, the grants they awarded, and the resulting publications, as well as links between patents and publications in the form of non-patent literature citations tracking the use of published research in invention.[30] The dataset is extensive, containing over 131.5 million publications from various sources, including 107,000 journals and 62 pre-print servers, 42.8 million patent-to-publication citations, and 6.3 million grants totaling $2.3 trillion from 656 funding agencies globally.

A.1.2 American Men & Women of Science

The American Men & Women of Science (AMWS) directory provides information on prominent scientists working in various scientific fields in the United States and Canada. The directory is published annually and includes information on individuals' professional affiliations, areas of research, and contact information. We acquired 17 electronic versions of the directory, covering editions published from 2005 through 2021.

The AMWS directory only profiles living scientists. We combined data from the 17 editions to create a comprehensive dataset with information on 200,706 living scientists from the 2021 edition as well as 17,657 deceased scientists from the 2005-2020 editions. Table A1 provides information on how many scientists were added to the 2021 edition from the historical versions.

We linked AMWS scientists to employers in a two-step procedure. First, we organized the unstructured, paragraph-based AMWS data into a structured, tabular format. We identified approximately 1.3 million professional positions corresponding to the 218,363 scientists. After discarding positions unique to academia (e.g., job titles containing the words professor, assistant professor, associate professor, editor, and lecturer), we identified approximately 459,830 positions in 210,224 unique organizations (e.g., Boeing) or sub-organizations (e.g., Applied Math Group). Second, we fuzzy-matched the 210,224 organization names with the 60,000 firm names in our panel (including both ultimate owners and their subsidiaries). After calculating the Levenshtein distance between the name strings using the Python package *TheFuzz*, we discarded potential matches with token set ratios below 90 (on a 0 to 100 scale), and manually checked the remaining potential matches. This process led to the identification of 12,817 matches between organization names and firm names. The 1,727 matched firms employed a total of 20,097 AMWS scientists during 1980-2015.

[30] Dimensions links scientific publications to supporting grants and funding organizations based on funding acknowledgments provided by authors at publication, as well as administrative data collected from major science funders.

Table A1: Construction of AMWS Dataset

AMWS edition	Scientists added to the final sample	Total scientists included in the edition
2005	2,212	132,812
2006	215	133,372
2007	800	132,312
2008	1,345	132,642
2009	1.240	131,880
2010	1,789	131,024
2011	997	135,790
2012	640	140,502
2013	408	145,439
2014	2,439	151,083
2015	139	158,884
2016	1,144	166,583
2017	2,845	169,048
2018	172	184,912
2019	209	180,225
2020	1,063	198,172
2021	200,706	203,272
Total scientists in the final sample	218,363	

Notes: This table underscores the importance of including historical versions of the AMWS directory. "Total scientists included in the edition" is the number of scientists featured in each edition. "Scientists added to the final sample" is the number of scientists not featured in subsequent editions due to death. We use the last (and most comprehensive) profile for these scientists in assembling our final dataset.

The AMWS directory provides full employment histories for the scientists profiled. Once we linked a scientist to a firm in our sample, we extracted the start and end year of the affiliation. We aggregated this information to the firm-year level by counting the number of AMWS scientists employed by a focal firm each year.

A.1.3 ProQuest Dissertations & Theses Global

ProQuest Dissertations & Theses Global (PQDT) is a comprehensive collection of over 5 million PhD dissertations and master's degree theses from thousands of universities worldwide. This dataset covers dissertations in various fields of study, including exact sciences, the humanities, and social sciences.

It is difficult to link PhD dissertations to firms because dissertations are not typically cited by other scientific publications or patents. To address this challenge, we used additional information from PQDT on the institutions where PhD candidates studied, the names of their advisors, the subject terms describing each dissertation's research fields, and dissertation abstracts.

We constructed our primary measure of firm-relevant human capital by using the textual similarity between the abstracts of dissertations and the abstracts of corporate patents, rather than relying on citation data (unlike publications, dissertations are not cited by patents).

We also matched each dissertation with its published version from Dimensions, if available. Since a dissertation often undergoes significant revisions before being published as a scientific publication, we compared the dissertation's abstract with the abstracts of all publications published by the same author within a decade from PhD graduation to identify the most similar publication to a focal dissertation. We then constructed a second measure of firm-relevant human capital, which allowed us to use patent citations to published dissertations to infer relevance to corporate R&D.

Moreover, we classified PhD dissertations into research fields. We then used the reliance of patenting subclasses on knowledge published across research fields to construct our third measure of firm-relevant human capital. PQDT provides a list of one or more non-standardized subject terms for each dissertation (e.g., "organic chemistry" or "health care; public health; and laboratories"). We manually created a list of 1,027 disambiguated subjects and discarded dissertations with a "soft science" subject, such as "literature," "history," and "social sciences." We also discarded PhD dissertations from non-U.S. universities as well as all master's degree theses. We ended up with a dataset of 771,023 U.S. PhD dissertations awarded between 1985 and 2016 in 394 "hard science" subjects. We manually assigned these subjects to the 25 OECD natural science subfields. Table A2 displays the resulting crosswalk for the most common subject terms. We then classified dissertations into one or more OECD subfields, which allowed us to capture the multidisciplinary nature of many dissertations.

We also faced the challenge of matching each PhD advisor to a researcher from Dimensions, which was necessary to construct instrumental variables for human capital. Instances of common names led to multiple ambiguous matches. To overcome this challenge, we restricted potential advisor matches using data on the PhD candidate's institutional affiliation and a 6-year time window that ended in the defense year. This allowed us to identify all the publications authored by each PhD advisor, along with the funding linkages between these publications and federal agencies. We used this information to construct instrumental variables for firm-relevant human capital.

A.2 Details on the Primary Independent Variables

A.2.1 Public Knowledge

Our *Public knowledge* measure captures the relevance of non-corporate publications to corporate R&D. Relevance is based on the firm's lagged publishing across the 25 OECD natural science subfields listed in Table A2.

To construct the measure, we first counted the number of non-corporate publications published in each OECD subfield each year. Then, we calculated each firm's shares of publications across OECD subfields by dividing (i) the number of firm publications in each subfield-time cohort by (ii) the total number of firm publications in the same time cohort. This allowed us to capture the importance of each subfield to a firm's research portfolio.

Our firm-year measure of relevant *Public knowledge* was constructed as the weighted sum of non-corporate publications, using the focal firm's shares of publications across OECD

Table A2: Crosswalk Between OECD Subfields and PhD Dissertation Subject Terms

OECD natural science subfield	Number of PhD dissertations	Most common subject term
1.01 Mathematics	41,106	Mathematics
1.02 Computer and information sciences	48,120	Computer science
1.03 Physical sciences and astronomy	41,254	Optics
1.04 Chemical sciences	83,023	Chemistry
1.05 Earth and related environmental sciences	24,932	Geology
1.06 Biological sciences	155,694	Molecular biology
1.07 Other natural sciences	1	Natural sciences
2.01 Civil engineering	16,205	Civil engineering
2.02 Electrical eng, electronic eng	46,092	Electrical engineering
2.03 Mechanical engineering	31,027	Mechanical engineering
2.04 Chemical engineering	19,101	Chemical engineering
2.05 Materials engineering	28,199	Materials science
2.06 Medical engineering	6,431	Biomedical engineering
2.07 Environmental engineering	31,526	Ecology
2.08 Environmental biotechnology	0	N/A
2.09 Industrial biotechnology	29	Tissue engineering
2.10 Nano-technology	2,053	Nanotechnology
2.11 Other engineering and technologies	17,615	Industrial engineering
3.01 Basic medical research	46,247	Pharmacology
3.02 Clinical medicine	41,111	Neurology
3.03 Health sciences	44,546	Public health
4.01 Agriculture, forestry, fisheries	21,334	Botany
4.02 Animal and dairy science	8,143	Animals
4.03 Veterinary science	3,670	Veterinary services
4.05 Other agricultural science	4,785	Food science

Notes: This table showcases the OECD natural science subfields that have been linked with ProQuest dissertations. It highlights the subject term most commonly used between 1980 and 2015 for each OECD subfield.

subfields during the previous 5-year time cohort as weights:

$$Public\ knowledge_{i,t} = \sum_{o \in O} Publications_{o,t} \times Precohort\ share\ of\ publications_{i,o} \tag{9}$$

The index o denotes OECD subfields. $Publications_{o,t}$ is the number of non-corporate publications published in year t in subfield o. $Precohort\ share\ of\ publications_{i,o}$ is firm i's lagged share of publications in subfield o. We calculated a stock measure of *Public knowledge* using a perpetual inventory method with a 15% depreciation rate.

A.2.2 Human Capital

Our primary *Human capital* measure captures the relevance of U.S. PhD dissertations to corporate R&D. Relevance is based on the textual similarity between the abstracts of dissertations and the abstracts of corporate patents. The following procedure details the construction of our measure.

1. We embedded our document corpus of abstracts from 771,023 U.S. PhD dissertations and 1.35 million corporate patents using the Scientific Paper Embeddings using

Table A3: OECD Subfields and Dissertation/Publication Counts

OECD natural science subfield	Number of PhD dissertations	Number of publications
1.01 Mathematics	41,106	476,621
1.02 Computer and information sciences	48,120	786,593
1.03 Physical sciences and astronomy	41,254	1,221,726
1.04 Chemical sciences	83,023	686,744
1.05 Earth and related environmental sciences	24,932	365,734
1.06 Biological sciences	155,694	967,170
1.07 Other natural sciences	1	128,23
2.01 Civil engineering	16,205	948,23
2.02 Electrical eng, electronic eng	46,092	656,393
2.03 Mechanical engineering	31,027	321,210
2.04 Chemical engineering	19,101	43,195
2.05 Materials engineering	28,199	278,763
2.06 Medical engineering	6,431	37,145
2.07 Environmental engineering	31,526	164,491
2.08 Environmental biotechnology	0	40,659
2.09 Industrial biotechnology	29	1,078
2.10 Nano-technology	2,053	5,870
2.11 Other engineering and technologies	17,615	156,777
3.01 Basic medical research	46,247	559,704
3.02 Clinical medicine	41,111	1,939,179
3.03 Health sciences	44,546	444,589
4.01 Agriculture, forestry, fisheries	21,334	105,698
4.02 Animal and dairy science	8,143	209,16
4.03 Veterinary science	3,670	13,224
4.05 Other agricultural science	4,785	9,732
All subfields	762,244	9,410,857

Notes: This table provides a breakdown of the number of ProQuest dissertations defended between 1980 and 2015 across 25 different OECD natural science subfields. As ProQuest does not categorize dissertations by OECD subfields, we manually linked the subject terms used in each dissertation with the appropriate subfield. In cases where a dissertation had multiple subject terms, we assigned fractional weights to each term. For instance, if a dissertation was labeled as "mathematics, electrical engineering," we attributed 0.5 dissertations to "mathematics" and 0.5 dissertations to "electrical engineering." This table also shows the count of Dimensions publications published between 1980 and 2015 in each of the 25 OECD natural science subfields. Because Dimensions does not automatically classify publications into OECD subfields, we extracted the OECD classification of publications that had digital object identifiers (representing approximately 72.12% of all publications in Dimensions) from Microsoft Academic Graph.

Citation-informed TransformERs (SPECTER) model (Cohan et al., 2020).[31] SPECTER is a specialized Bidirectional Encoder Representations from Transformers (BERT) model trained on 146,000 scientific papers (containing 26.7 million words) from Semantic Scholar and their forward citations. SPECTER has been compared to several other deep learning models specialized in technical documents, including BERT, PatentBERT, BERT for Patents, PatentSBERTa, SciBERT, RoBERTa, and ALBERT. As shown in Table A4, SPECTER has outperformed the alternative models in predict-

[31] Embedding is a way of representing text data as numerical vectors that capture the underlying meaning and context of the words.

ing forward citations. The out-performance is likely due to the fact that SPECTER is trained on scientific literature and patents, whereas other models are trained only on patents or general texts, such as the content of Wikipedia. Unlike vector representations used by term frequency-inverse document frequency (TF-IDF) algorithms, we embedded each dissertation abstract and each patent abstract into a densely bounded vector (dense meaning not having any missing values in the vector, and bounded meaning using values that can only fall between 0 and 1). Each word in the abstract was converted into a vector of 768 values between 0 and 1 (a 1 by 768 vector). Each abstract was converted into a matrix of size 768 by the number of words in the abstract. We condensed the matrix into a vector with 768 rows and one column using a mean pooling approach (averaging across rows).

2. Using the vectors from the previous step, we calculated the cosine similarity for each dissertation-patent pair (0.77 million PhD dissertations and 1.35 million patents). Due to the large number of abstract pairs, we used a high-performance computing (HPC) cluster to distribute the task over 40 NVIDIA A100-40GB GPUs, with each GPU running for more than six days continuously. The system processed and ranked over 1 trillion pairs of abstracts.[32]

3. For each corporate patent granted in year t, we identified the top 1,000 most similar dissertations granted in years $[t-1, t+1]$. Sample firms don't necessarily patent every year. To ensure we don't have zero relevant human capital, we focused on the 5-year time cohort as our relevant period. We identified all the dissertations that were similar (i.e., in the top 1,000) to the patents granted to a focal firm during each 5-year time cohort. Because a PhD dissertation could be similar to multiple corporate patents, we calculated the maximum textual similarity score between the dissertation and all the patents granted to the focal firm during the 5-year time cohort.

4. *Human capital* was constructed as the weighted sum of PhD dissertations, using the maximum similarity scores between dissertations and patents granted to the focal firm as weights:

$$PhD\ dissertations_{i,t} = \sum_{d \in D} Maximum\ textual\ similarity_{d,i,t} \tag{10}$$

D is the set of PhD dissertations in the top 1,000 most similar dissertations for one or more of the patents granted to firm i during the 5-year time cohort t. $Maximum\ textual\ similarity_{d,i,t}$ is the maximum textual similarity score between the abstract of dissertation d and the abstracts of all patents granted to firm i during the 5-year time cohort t.

[32]The computing needs included significant storage (more than 10 TBs) and memory resources (more than 4,800 GBs of RAM). The total computational time for the similarity task was approximately 60 days.

Table A4: Performance Comparison for Deep Learning Models

Percentile of similarity	Doc2Vec	PatentBERT	BERT for Patents	PatentSBERTa	SPECTER
Top 1%	61.28%	77.44%	43.60%	86.28%	87.81%
Top 3%	76.22%	88.11%	76.83%	94.51%	94.51%
Top 5%	81.10%	93.29%	86.59%	95.43%	97.56%

Notes: This table from Delron, Guellec, Wu, and Liu (2022) compares the performance of several deep learning models in identifying patent-paper pairs (i.e., the scientific publication that expresses the same technical content as a given patent). SPECTER was more accurate compared to other models. For a given set of patents, the Doc2Vec, PatentBERT, BERT for Patents, PatentSBERTa, and SPECTER models were used to identify the most similar publications. The outputs were then compared to the ground truth (correct pairs of publications) for each patent. SPECTER was able to rank the correct publication pair among the top 1% most similar publications for 87.81% of patents, which is a relatively high accuracy rate. Conversely, BERT for Patents was only able to achieve this for 43.60% of patents. This suggests that SPECTER is the more effective model for identifying the most similar publications to a given patent.

A.2.3 Public Invention

Our primary *Public invention* measure captures the relevance of university patents to corporate R&D. Relevance is based on the firm's lagged patenting across patent subclasses.

To construct the measure, we first counted the number of university patents granted in each subclass each year. University patents are those assigned to entities with a Global Research Identifier Data (GRID) organization type of "education."[33] During our sample period, there were 125,019 patents assigned to educational institutions. We identified patent subclasses using the first four digits of the current CPC classification from PatentsView. Then, we calculated each firm's shares of patents across subclasses by dividing (i) the number of firm parents in each subclass-time cohort by (ii) the total number of firm patents in the same time cohort. This allowed us to capture the importance of each subclass to a firm's invention portfolio.

Our firm-year measure of relevant *Public invention* was constructed as the weighted sum of university patents, using the focal firm's shares of patents across subclasses during the previous 5-year time cohort as weights:

$$\begin{aligned} Public\ invention_{i,t} = &\sum_{s \in S} University\ patents_{s,t} \\ &\times Precohort\ share\ of\ patents_{i,s} \end{aligned} \tag{11}$$

The index s denotes patent subclasses. $University\ patents_{s,t}$ is the count of patents granted to universities in subclass s in year t. $Precohort\ share\ of\ patents_{i,s}$ is firm i's lagged share of patents in subclass s.

[33] Other organization types are company, healthcare, nonprofit, facility, other, government, and archive. For more information on GRID, see https://www.grid.ac/.

A.3 Details on the Instrumental Variables

A.3.1 Data Sources on Agency R&D Budgets

We used data on federal R&D budgets from the "Total R&D by Agency, 1976-2020" series compiled by the American Association for the Advancement of Science (AAAS, 2021). Total R&D includes basic research, applied research, development, construction of R&D facilities, and major capital equipment for R&D. Each year, federal agencies are required to report their R&D budgets to the White House Office of Management and Budget (OMB). AAAS compiles these data, along with historical data published by OMB and survey data published by the National Science Foundation's National Center for Science and Engineering Statistics, into a data series of R&D budgets by agency, character, and discipline.

Table A5 summarizes the R&D budgets by agency and decade. It demonstrates the significant variation in R&D budgets between agencies and over time. Some agencies are significant funders of R&D (e.g., Defense, Health and Human Services), while others are not (e.g., Environmental Protection Agency, Department of Homeland Security). More importantly, the composition of federal R&D investments has changed over time. Defense-related R&D has dropped from 58% of all federal R&D budgets in the 1980s to only 49% in the 2010s. Conversely, human health-related R&D has increased from 11% of all federal R&D budgets in the 1980s to 23% in the 2010s. We exploit these differences and changes to "shock" the public science relevant to firms.

Table A5: R&D Budgets by Federal Agency and Decade

Federal agency	1980s	1990s	2000s	2010s
Dept. of Agriculture	20,701	24,101	29,876	27,674
Dept. of Commerce	8,179	13,841	15,394	17,162
Dept. of Defense	606,484	605,579	844,275	759,147
Dept. of Energy	126,045	112,475	114,065	148,106
Dept. of Health and Human Services	115,999	187,244	353,023	359,032
Natl. Institutes of Health	109,162	176,561	337,191	343,299
Other Subagencies	6,837	10,683	15,832	15,733
Dept. of Homeland Security	0	0	9,845	8,175
Dept. of Transportation	8,307	9,247	9,997	10,256
Dept. of Veterans Affairs	3,970	5,679	10,543	13,221
Dept. of the Interior	8,896	9,806	8,535	9,444
Environmental Protection Agency	7,099	8,721	7,833	5,899
Natl. Aeronautics and Space Admin.	95,993	145,265	139,002	120,671
Natl. Science Foundation	27,665	35,905	52,556	64,574
Others	16,407	16,390	13,792	16,419
Total	1,045,750	1,174,251	1,608,738	1,560,508

Notes: This table displays R&D budgets (in constant 2020 $ millions) by federal agency and decade. *Others* includes federal agencies that are not major funders of R&D (e.g., Department of Education, Department of Labor, etc.). Data are from the Total R&D by Agency, 1976-2020 series (American Association for the Advancement of Science, 2021).

A.3.2 Agency R&D Budgets

We construct a Bartik-style shift-share instrument for each component of public science. Our instrument *R&D budget - public knowledge* combines "shifts" to the federal funding for public knowledge published in each OECD natural science subfield with firm-specific "exposure shares" based on the firm's publishing across OECD subfields in the previous 5-year time cohort. The following procedure explains its construction:

1. We used data from AAAS to identify the value of the R&D budget appropriated by Congress to each of the 12 main federal agencies (plus an "Other" category for smaller agencies) in each year.[34]

2. We used the connections between federal agencies, grants, and publications from Dimensions to identify the federal agencies that funded each non-corporate publication.

3. For each OECD subfield, we calculated its reliance on funding from each federal agency by dividing (i) the number of publications published in the focal subfield over 1980-2015 and funded by a focal agency by (ii) the total number of publications published in the same subfield over 1980-2015.

4. We calculated each firm's shares of publications across OECD subfields by dividing (i) the number of firm publications in each subfield-time cohort by (ii) the total number of firm publications in the same time cohort.

5. We combined the shifts and exposure shares to calculate our first instrument:

$$\begin{aligned} \textit{R\&D budget - public knowledge}_{i,t} = & \sum_{o \in O} \textit{Precohort share of publications}_{i,o} \\ & \left(\sum_{a \in A} \textit{R\&D budget}_{a,t} \times \textit{Reliance on agency}_{o,a} \right) \end{aligned} \tag{12}$$

   O denotes OECD subfields. $\textit{Precohort share of publications}_{i,o}$ is firm i's share of publications in subfield o during the previous 5-year time cohort. A is the set of 12 main federal agencies, plus an "Other" category for smaller agencies. $\textit{R\&D budget}_{a,t}$ is the R&D budget of agency a in year t. $\textit{Reliance on agency}_{o,a}$ is a share obtained by dividing the number of publications published in subfield o over 1980-2015 and funded by agency a by the total number of publications published in subfield o over 1980-2015.

Our instrument *R&D budget - public invention* combines "shifts" to the federal funding for knowledge cited by university patents granted in each subclass with firm-specific "exposure shares" based on the firm's patenting across subclasses in the previous 5-year time cohort. Its construction broadly parallels that of the instrumental variable for *Public knowledge*, with two updates. First, to connect federal funding for science to public invention, we used

[34] The "Total R&D by Agency, 1976-2020" table includes "budget authority in millions of constant FY 2020 dollars." The constant-dollar conversions used OMB's chained price index, which can be found in historical table 10.1 available at https://www.whitehouse.gov/omb/historical-tables/.

the non-patent literature (NPL) citations and funding linkages from Dimensions to identify the federal agencies that funded each non-corporate publication cited by a university patent. Second, for each patent subclass, we calculated its reliance on public science funded by each federal agency. Our first instrument for *Public invention* was calculated as:

$$\begin{aligned} R\&D\ budget\ \text{-}\ public\ invention_{i,t} = & \sum_{s \in S} Precohort\ share\ of\ patents_{i,s} \\ & \left(\sum_{a \in A} R\&D\ budget_{a,t} \times Reliance\ on\ agency_{s,a} \right) \end{aligned} \tag{13}$$

The index s denotes patent subclasses. $Precohort\ share\ of\ patents_{i,s}$ is firm i's share of patents in subclass s during the previous 5-year time cohort, obtained by dividing the number of firm patents granted in subclass s by the total number of firm patents in that time period. A and $R\&D\ budget_{a,t}$ are as previously defined. $Reliance\ on\ agency_{s,a}$ is a share obtained by dividing the number of citations from university patents granted in subclass s over 1980-2020 to non-corporate publications published over 1980-2015 and funded by agency a by the total number of citations from university patents granted in subclass s over 1980-2020 to all non-corporate publications published over 1980-2015.

Our instrument *R&D budget - human capital* combines "shifts" to the federal funding for PhD dissertation advisors with the "exposure shares" of the similarity scores between the abstracts of dissertations and the abstracts of firm patents, as follows:

$$\begin{aligned} R\&D\ budget\ \text{-}\ human\ capital_{i,t} = & \sum_{d \in D} Maximum\ textual\ similarity_{d,i,t} \\ & \left(\sum_{a \in A} R\&D\ budget_{d,a} \times Share\ of\ agency_{d,a} \right) \end{aligned} \tag{14}$$

D is the set of PhD dissertations in the top 1,000 most similar dissertations for one or more of the patents granted to firm i during the time cohort t. $Maximum\ textual\ similarity_{d,i,t}$ is the maximum textual similarity score between the abstract of dissertation d and the abstracts of all patents granted to firm i during the 5-year time cohort t. A is as previously defined. $R\&D\ budget_{d,a}$ is the R&D budget for agency a at the beginning of the PhD program (that is, five years before the year of defense of dissertation d). $Share\ of\ agency_{d,a}$ is obtained by dividing the funding amount (in \$) from agency a to the publications of the advisor(s) of dissertation d during the 6-year period ending in dissertation d's defense year by the total funding amount (in \$) from agency a to any publication published during the 6-year period ending in the defense year of the dissertation d.

A.3.3 The Role of Subcommittees in the Congressional Appropriations Process

The U.S. federal budget includes two types of spending: discretionary spending and mandatory spending. Discretionary spending refers to the portion of the budget that is decided by Congress through the annual appropriations process, whereas mandatory spending includes expenditures that are mandated by law, such as Social Security and Medicare.

The U.S. House Appropriations Committee and its counterpart, the U.S. Senate Appropriations Committee, play a pivotal role in the legislative process, being responsible for passing appropriations bills that regulate the discretionary spending of federal agencies. Each committee is organized into subcommittees, and each subcommittee is charged with developing one regular annual appropriations bill that allocates funding for various agencies and activities that fall under its jurisdiction (Bloomberg Government, 2023). Importantly, the jurisdiction of each U.S. House appropriations subcommittee mirrors that of a corresponding U.S. Senate appropriations subcommittee. This pairing of subcommittees between the two chambers of Congress ensures symmetry and coordination in the appropriations process.

The composition and names of congressional appropriations subcommittees are not static over time, reflecting the evolving priorities and structure of the federal government. For example, the Homeland Security subcommittee was established in 2003 to oversee the newly created Department of Homeland Security, itself the result of combining all or part of 22 different federal departments and agencies. Since 2007, the U.S. House Appropriations Committee and the U.S. Senate Appropriations Committee have each included 12 subcommittees:

1. Agriculture, Rural Development, Food and Drug Administration, and Related Agencies;
2. Commerce, Justice, Science, and Related Agencies;
3. Defense;
4. Energy and Water Development;
5. Financial Services and General Government;
6. Homeland Security;
7. Interior, Environment, and Related Agencies;
8. Labor, Health and Human Services, Education, and Related Agencies;
9. Legislative Branch;
10. Military Construction, Veteran Affairs, and Related Agencies;
11. State, Foreign Operations, and Related Programs; and
12. Transportation, Housing and Urban Development, and Related Agencies

In this study, we leverage data from these 24 subcommittees on appropriations to construct our preferred instrumental variables. The central premise here is that while these subcommittees play a critical role in the allocation of governmental funds, their political composition is plausibly exogenous to the production of science. As such, variation in the political composition of the subcommittees offers a potentially powerful, and theoretically exogenous, source of variation in federal funding for public science.

A.3.4 Data Sources on the Political Composition of Subcommittees

Given the absence of a comprehensive data source about historical congressional appropriations subcommittees, we manually collected data from a variety of sources. We compiled information on the jurisdiction and membership roster of each subcommittee from the 95th Congress (1977-1978) through the 114th Congress (2015-2016).

We used the jurisdiction information to identify which subcommittees are responsible for which federal agencies. Table A6 summarizes the mapping between the 12 pairs of appropriations subcommittees and our 12 main federal agencies. The catch-all category of "Others" was mapped directly to the U.S. House and U.S. Senate Appropriations Committees.

Table A6: Crosswalk Between Appropriations Subcommittees and Federal Agencies

Subcommittees	USDA	DoC	DoD	DoE	HHS	DHS	DoT	VA	DoI	EPA	NASA	NSF
1. Agriculture, Rural Development, Food and Drug Administration	100%											
2. Commerce, Justice, Science		100%									75%	75%
3. Defense			100%								25%	
4. Energy and Water Development				100%					25%			
5. Financial Services and General Government												25%
6. Homeland Security						100%						
7. Interior, Environment									75%	100%		
8. Labor, Health and Human Services, Education					100%							
9. Legislative Branch												
10. Military Construction, Veterans Affairs								100%				
11. State, Foreign Operations												
12. Transportation, Housing and Urban Development							100%					
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Notes: This table maps the 12 appropriations subcommittees in the U.S. House and the U.S. Senate to 12 main federal agencies based on the jurisdictions of the subcommittees. The reported percentages represent the weights applied when calculating the two measures of political composition, *Majority party share* and *Democratness*, at the agency level.

We used the membership rosters to extract two pieces of information for each subcommittee (as well as the overall U.S. House and U.S. Senate Appropriations Committees):

1. **Majority party share**: To quantify how dominant the majority party was in the subcommittee, we calculated the ratio of (i) the number of members from the majority political party in the chamber and (ii) the total number of members in the subcommittee.

2. **Democratness**: To quantify the ideological orientation of the subcommittee, we calculated the ratio of (i) the number of Democrats and (ii) the total number of members in the subcommittee.

For agencies that fall under the jurisdiction of a single pair of appropriations subcommittees (e.g., Health and Human Services, which is overseen by the subcommittees on Labor, Health and Human Services, Education, and Related Agencies), we calculated a simple average *Majority party share* across the pair of subcommittees to arrive at an agency-year measure. We did the same for the *Democratness* measure.

For agencies that fall under the jurisdiction of two pairs of appropriations subcommittees (e.g., Department of the Interior, which is overseen by both the subcommittees on Energy and Water Development and the subcommittees on Interior, Environment, and Related Agencies), we calculated a weighted average *Majority party share* across the relevant subcommittees, using the percentages reported in Table A6 as weights.[35] We did the same for the *Democratness* measure.

Important for our identification strategy, the political composition of subcommittees predicts the R&D budgets of federal agencies. As shown in Table A7, the relationships between lagged *Majority party share* and *R&D budget* is negative and significant (Columns 2 and 4, p-values < 0.05). The relationship between lagged *Democratness* and *R&D budget* is positive, though imprecisely estimated (Columns 3 and 4).

Table A7: Political Composition Predicts Agency R&D Budgets

Dependent variable:	(1)	(2)	(3)	(4)
	R&D budget$_t$			
	Baseline	Add Majority party share	Add Democratness	Add both
Majority party share$_{t-1}$		-50.339*		-60.895*
		(22.256)		(24.018)
Democratness$_{t-1}$			5.502	12.082
			(8.760)	(9.304)
Time trend$_t$	0.259***	0.211**	0.279***	0.245**
	(0.067)	(0.065)	(0.081)	(0.077)
Mean DV	7.31	7.31	7.31	7.31
Observations	452	452	452	452
Adjusted R-squared	0.03	0.04	0.03	0.04

Notes: This table presents the OLS estimation results for the relationship of agency R&D budgets with two measures of the political composition of congressional appropriations subcommittees, *Majority party share* and *Democratness*. *Time trend* measures the number of years since 1979. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity.

A.3.5 Predicted Agency R&D Budgets

We use the two measures of the political composition of congressional appropriations subcommittees to construct a preferred Bartik-style shift-share instrument for each component

[35] We created these weights to attempt to capture the relative importance of different subcommittees in regulating the spending of federal agencies.

of public science. Instead of using the actual R&D budget of each agency in equations 12, 13, and 14, we first predict the agency R&D budget using the specification reported in Column 4 of Table A7, then use this predicted agency R&D budget to construct our instrumental variables. Table A8 provides summary statistics for all the instrumental variables used in the econometric analyses.

Table A8: Summary Statistics for Instrumental Variables

	(1)	(2)	(3)	(4)	(5)	(6)
				Distribution		
	Obs.	Mean	Std. dev.	10th	50th	90th
Predicted R&D budget - public invention ($ mm)$_{t-1}$	41,698	11,020	8,387	1,967	9,197	23,072
Predicted R&D budget - human capital ($ mm)$_{t-1}$	41,698	2,080	4,634	0	195	6,006
Predicted R&D budget - public knowledge ($ mm)$_{t-1}$	41,698	14,724	18,712	0	4,735	42,132
R&D budget - public invention ($ mm)$_{t-1}$	41,698	27,273	23,184	4,408	21,659	58,098
R&D budget - human capital ($ mm)$_{t-1}$	41,698	2,349	5,443	0	164	7,093
R&D budget - public knowledge ($ mm)$_{t-1}$	41,698	30,103	43,184	0	7,413	86,091
Predicted R&D budget - public invention, broad ($ mm)$_{t-1}$	41,698	4,691	4,069	820	3,567	10,362
Predicted R&D budget - human capital, cited ($ mm)$_{t-1}$	41,698	2	11	0	0	3
Predicted R&D budget - human capital, OECD ($ mm)$_{t-1}$	41,698	695	775	0	402	1,725

Notes: This table provides summary statistics for the instrumental variables used in the econometric analyses. The analysis sample is at the firm-year level and includes an unbalanced panel of 3,372 U.S.-headquartered publicly traded firms during 1986 to 2015.

A.3.6 First-Stage Estimation Results

Tables A9 and A10 present first-stage OLS regression results for the patents, publications, AMWS scientists, and R&D expenditures equations. They differ in that Table A9 uses the first set of instrumental variables, while A10 uses the second (and preferred) set of instrumental variables. Regardless of the approach used, all components of public science are positively related to their respective instrumental variables (p-values < 0.001, F statistics > 104.7, see Lee et al. (2022)).

A.4 Details on the Alternative Independent Variables

A.4.1 Human Capital Using PhD Dissertations Cited by Patents

We measured relevant human capital using an alternative approach that focused on published dissertations. We followed a four-step process to construct it:

1. To identify published dissertations, we matched ProQuest PhD students' dissertations with publications in Dimensions. We used both author names and the textual similarities between the abstracts of dissertations and the abstracts of publications to perform the matching. Specifically, we used the SPECTER model to measure the textual similarity between each ProQuest dissertation and all the publications authored

Table A9: Instrumental Variable Estimation (First Stage)

	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln(1+\text{Public invention})_{t-1}$		$\ln(1+\text{Human capital})_{t-1}$		$\ln(1+\text{Public knowledge})_{t-1}$	
	Baseline	For R&D expenditures	Baseline	For R&D expenditures	Baseline	For R&D expenditures
$\ln(\$1+\text{R\&D budget - public invention})_{t-1}$	0.563***	0.585***				
	(0.027)	(0.025)				
$\ln(\$1+\text{R\&D budget - human capital})_{t-1}$			0.345***	0.359***		
			(0.011)	(0.011)		
$\ln((\$1+\text{R\&D budget - public knowledge})_{t-1}$					1.481***	1.499***
					(0.066)	(0.067)
$\ln(\$1+\text{R\&D stock})_{t-1}$	0.069**		0.171***		0.093*	
	(0.022)		(0.017)		(0.046)	
$\ln(\$1+\text{Sales})_{t-1}$		0.006		0.012***		0.018
		(0.006)		(0.003)		(0.011)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	266.11	266.53	6,412.66	6,431.26	62,549.62	62,721.97
Firms	3,372	3,369	3,372	3,369	3,372	3,369
Observations	41,698	41,436	41,698	41,436	41,698	41,436
F statistic	271	295	5,410	6,081	2,010	2,367
R-squared	0.83	0.83	0.95	0.95	0.89	0.89

Notes: This table displays first-stage OLS regression results for instrumental variables based on the R&D budgets of federal agencies. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

by the same individual within a 10-year window after PhD graduation. Our assumption was that the top most similar publication corresponds to the published version of the student's dissertation.

2. For each patent subclass and year, we counted the number of PhD dissertations defended that year whose published versions were cited by patents granted in the focal subclass between 1980 and 2020.

3. We used the same firm patenting shares across CPC subclasses as previously described.

4. We calculated our firm-year measure of *PhD dissertations, cited* as follows:

$$PhD\ dissertations, cited_{i,t} = \sum_{s \in S} Cited\ PhD\ dissertations_{s,t} \times Precohort\ share\ of\ patents_{i,s} \qquad (15)$$

The index s denotes patent subclasses. $Cited\ PhD\ dissertations_{s,t}$ is the number of PhD dissertations defended in year t whose published version was cited by patents in subclass s during 1980-2020. $Precohort\ share\ of\ patents_{i,s}$ is firm i's share of patents in subclass s in the previous time cohort.

Table A10: Preferred Instrumental Variable Estimation (First Stage)

	(1)	(2)	(3)	(4)	(5)	(6)
	$\ln(1+\text{Public invention})_{t-1}$		$\ln(1+\text{Human capital})_{t-1}$		$\ln(1+\text{Public knowledge})_{t-1}$	
	Baseline	For R&D expenditures	Baseline	For R&D expenditures	Baseline	For R&D expenditures
$\ln(\$1+\text{Predicted R\&D budget - public invention})_{t-1}$	0.574***	0.599***				
	(0.027)	(0.025)				
$\ln(\$1+\text{Predicted R\&D budget - human capital})_{t-1}$			0.282***	0.298***		
			(0.010)	(0.011)		
$\ln((\$1+\text{Predicted R\&D budget - public knowledge})_{t-1}$					1.474***	1.500***
					(0.070)	(0.071)
$\ln(\$1+\text{R\&D stock})_{t-1}$	0.076***		0.189***		0.119*	
	(0.022)		(0.016)		(0.048)	
$\ln(\$1+\text{Sales})_{t-1}$		0.006		0.013***		0.021
		(0.006)		(0.003)		(0.011)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	266.11	266.53	6,412.66	6,431.26	62,549.62	62,721.97
Firms	3,372	3,369	3,372	3,369	3,372	3,369
Observations	41,698	41,436	41,698	41,436	41,698	41,436
F statistic	258	280	6,746	7,502	1,922	2,267
R-squared	0.83	0.83	0.96	0.96	0.89	0.89

Notes: This table displays first-stage OLS regression results for instrumental variables based on predicted R&D budgets of federal agencies, where two measures of the political composition of congressional appropriations subcommittees, *Majority party share* and *Democratness*, are first used to predict R&D budgets. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

A.4.2 Human Capital Using OECD Subfields

We also measured relevant human capital using an approach that focused on the classification of PhD dissertations into OECD natural science subfields. We followed a five-step process to construct it:

1. We classified PhD dissertations into one or more research fields using a manual cross-walk between dissertations' subject terms and the 25 OECD natural science subfields (see Table A2 for examples of the most commonly used subject term for each subfield).

2. We counted the number of PhD dissertations defended each year in each OECD natural science subfield.

3. For each patent subclass and time cohort, we calculated its reliance on knowledge published across OECD natural science subfields by dividing (i) the number of citations from patents granted in the focal subclass during the time cohort to publications published in the focal OECD subfield by (ii) the total number of citations from patents granted in the focal subclass during the time cohort to publications published in any subfield. For instance, if there were 100 NPL citations from subclass C01C (Ammonia; Cyanogen) in time-cohort t and 50 of those citations were to OECD subfield 2.04 Chemical engineering, then the $CPC-OECD\ share_{s,o,t}$ for subclass C01C and OECD subfield 2.04 in time cohort t was 0.5.

4. We used the same firm patenting shares across CPC subclasses as previously described.

5. We calculated our firm-year measure of *PhD dissertations, OECD* as follows:

$$PhD\ dissertations, OECD_{i,t} = \sum_{s \in S} Precohort\ share\ of\ patents_{i,s} \left(\sum_{o \in O} PhD\ dissertations_{o,t} \times CPC - OECD\ share_{o,s,t} \right) \tag{16}$$

The index s denotes patent subclasses. $Precohort\ share\ of\ patents_{i,s}$ is firm i's share of patents in subclass s in the previous time cohort. The index o denotes OECD natural science subfields. $PhD\ dissertations_{o,t}$ is the number of PhD dissertations defended in year t in OECD natural science subfield o. $CPC - OECD\ share_{o,s,t}$ is the relative importance of OECD subfield o to patent subclass s in time cohort t.

A.4.3 Public Invention Using Publications Cited by Patents

Our first alternative measure of relevant public invention is based on publications cited by patents. The following procedure explains its construction:

1. We counted the number of unique non-corporate publications cited by at least one patent (whether a corporate patent or not) from each patent subclass at the subclass and publishing year level.

2. We used the same firm patenting shares across CPC subclasses as previously described.

3. We calculated our firm-year measure of *Public invention, broad* as follows:

$$Public\ invention,\ broad_{i,t} = \sum_{s \in S} Publications\ cited\ by\ patents_{s,t} \times Precohort\ share\ of\ patents_{i,s} \tag{17}$$

The index s denotes patent subclasses. $Publications\ cited\ by\ patents_{s,t}$ is the number of non-corporate publications published in year t and cited by at least one patent (whether a corporate patent or a non-corporate patent) granted in subclass s during 1980-2020. $Precohort\ share\ of\ patents_{i,s}$ is firm i's share of patents in subclass s during the previous 5-year time cohort.

A.4.4 Public Invention Using Textual Similarity

Another approach for determining firm-relevant public invention uses the textual similarity between publication and corporate patents. We use the SPECTER algorithm again and implement the following procedure:

1. The *embedding step* generated a low-dimensional representation of words in a document corpus consisting of approximately 14.5 million scientific publications authored by researchers affiliated with U.S. entities and 1.35 million corporate patents.

2. Using the vectors obtained from the previous step, we computed the cosine similarity between each corporate patent and the publications published within a window of plus or minus one year from the year the patent was granted. We then ranked the publication abstracts in descending order of the cosine similarity score, keeping only the top 10,000 most similar publications for each patent abstract.

3. We determined the *pool of publications* relevant to each firm's patents granted within a specific time cohort. This list included all publications that were retained in the top 1,000 most similar publications for one or more of the firm's patents granted in the time cohort. We used the firm-cohort level rather than the firm-year level because not all firms in our analysis sample have granted patents every year, which would result in many instances where the pool of relevant public invention would be zero.

4. For each publication in the pool, we identified its *maximum similarity score* with the firm's patents granted in the time cohort.

5. We calculated our firm-year measure of *Public invention, SPECTER* as follows:

$$Public\ invention,\ SPECTER_{i,t} = \sum_{p \in P} Maximum\ similarity_{p,i,t} \tag{18}$$

   P is the set of publications in the top 1,000 most similar publications for one or more of the patents granted to firm i during time cohort t. $Maximum\ similarity_{p,i,t}$ is the maximum textual similarity score between the abstract of publication p and all the abstracts of the patents granted to firm i during time cohort t.

Appendix B Robustness Checks

We performed a variety of checks to test the robustness of the effect of public science on corporate patents, publications, and AMWS scientists.

B.1 Alternative Instrumental Variables

Table B11: Alternative Instrumental Variable Estimation (Second Stage)

Dependent variable:	(1) $\ln(1+\text{Patents})_t$	(2) $\ln(1+\text{Publications})_t$	(3) $\ln(1+\text{AMWS scientists})_t$	(4) $\ln(\$1+\text{R\&D expenditures})_t$
$\ln(1+\text{Public invention})_{t-1}$	-0.232***	-0.143***	-0.028*	-0.162
	(0.032)	(0.023)	(0.013)	(0.111)
$\ln(1+\text{Human capital})_{t-1}$	0.222***	0.098***	0.033**	0.162**
	(0.025)	(0.016)	(0.010)	(0.060)
$\ln(1+\text{Public knowledge})_{t-1}$	0.017	-0.001	0.020**	0.021
	(0.012)	(0.010)	(0.006)	(0.030)
$\ln(\$1+\text{R\&D stock})_{t-1}$	0.255***	0.169***	0.037***	
	(0.020)	(0.016)	(0.009)	
$\ln(\$1+\text{Sales})_{t-1}$				0.188***
				(0.021)
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Mean DV	28.30	15.72	4.80	142.59
Weak id. (Kleibergen-Paap)	121.24	121.24	121.24	86.52
Firms	3,372	3,372	3,372	3,162
Observations	41,698	41,698	41,698	36,584

Notes: This table presents the second stage of 2SLS estimation using alternative instrumental variables based on the R&D budgets of federal agencies. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

B.2 Separating Public Invention from Human Capital

We test the robustness of our main results to separating public invention from human capital in different ways. We report the second stage of 2SLS estimation in Table B12. In Column 1 we address the concern that knowledge and PhDs are jointly produced in universities. Specifically, we consider only patent-cited publications from federal laboratories (i.e., publications authored by scientists affiliated with laboratories owned by the federal government) in calculating our *Public invention, broad* measure and its corresponding instrumental variable. This approach leverages the fact that federal laboratories produce scientific knowledge but do not award PhD degrees. Their scientific output tends to be more relevant for invention as well. On average, publications coauthored by scientists affiliated with the federal laboratories receive more citations from patents compared to other publications, as shown in Table C20.

In Column 2, we exclude patent-cited publications that are based on PhD dissertations. To identify these publications, we examine the publication history of each PhD student in our dataset, searching for publications authored by the student within 10 years of their graduation that closely resemble the dissertation abstract, as outlined in Section **??**. We infer that these publications are published versions of the dissertation and remove them from the construction of our *Public invention, broad* measure and its corresponding instrumental variable. As a result of this procedure, 6,199 publications are excluded from the estimation sample.

Next, we exclude publications that have a PhD student as a coauthor (Column 3). We identify these publications by comparing the list of authors with our list of students. Publications with at least one coauthor who is a PhD student during the publication year are removed from the construction of our *Public invention, broad* measure and its corresponding instrumental variable. This procedure excluded 74,397 publications from the estimation sample.

In Column 4, we exclude publications that are authored by the advisors of PhD students. We identify these publications by comparing the list of authors with our list of advisors. Publications with at least one coauthor who is a PhD advisor (at any point in time) are removed from the construction of our *Public invention, broad* measure and its corresponding instrumental variable. This procedure excluded 391,007 publications from the estimation sample.

Next, we seek to account for differences in the intensity of human capital production across different scientific fields (Columns 5 and 6). We separate fields that have high (i.e., above median) ratios of funding received by PhD students and their advisors to total funding received from those that have low (i.e., below median) ratios. Publications from high ratio fields are dropped from the construction of our *Public invention, broad* measure in Column 5, while publications from low ratio fields are similarly dropped in Column 6.

The coefficient estimates on public invention and human capital remain consistent across all specifications.

Table B12: Separating Public Invention From Human Capital

	(1) Federal lab. pub.	(2) Without dissertation pub.	(3) Without PhD student pub.	(4) Without PhD advisor pub.	(5) Low student-advisor funding fields	(6) High student-advisor funding fields
A. Dependent variable:			$\ln(1{+}\text{Patents})_t$			
$\ln(1{+}\text{Public invention, broad})_{t-1}$	-0.264***	-0.174***	-0.183***	-0.178***	-0.179***	-0.213***
	(0.038)	(0.023)	(0.025)	(0.025)	(0.024)	(0.032)
$\ln(1{+}\text{Human capital})_{t-1}$	0.326***	0.309***	0.311***	0.309***	0.300***	0.338***
	(0.032)	(0.029)	(0.030)	(0.030)	(0.028)	(0.034)
$\ln(\$1{+}\text{R\&D stock})_{t-1}$	0.235***	0.236***	0.236***	0.236***	0.236***	0.235***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.020)
B. Dependent variable:			$\ln(1{+}\text{Publications})_t$			
$\ln(1{+}\text{Public invention, broad})_{t-1}$	-0.187***	-0.104***	-0.106***	-0.106***	-0.116***	-0.134***
	(0.029)	(0.017)	(0.018)	(0.018)	(0.018)	(0.024)
$\ln(1{+}\text{Human capital})_{t-1}$	0.135***	0.112***	0.111***	0.111***	0.110***	0.133***
	(0.022)	(0.019)	(0.019)	(0.019)	(0.018)	(0.022)
$\ln(\$1{+}\text{R\&D stock})_{t-1}$	0.163***	0.164***	0.164***	0.164***	0.164***	0.163***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
C. Dependent variable:			$\ln(1{+}\text{AMWS scientists})_t$			
$\ln(1{+}\text{Public invention, broad})_{t-1}$	-0.050**	-0.029**	-0.031**	-0.030**	-0.033**	-0.044**
	(0.016)	(0.010)	(0.011)	(0.011)	(0.011)	(0.014)
$\ln(1{+}\text{Human capital})_{t-1}$	0.055***	0.050***	0.050***	0.050***	0.050***	0.059***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.014)
$\ln(\$1{+}\text{R\&D stock})_{t-1}$	0.036***	0.037***	0.037***	0.037***	0.036***	0.036***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents the second stage of 2SLS estimation for the relationship between different (broad) measures of public invention and corporate patents (Panel A), publications (Panel B) and employment of AMWS scientists (Panel C). Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

B.3 High-Quality Corporate Innovation

We report robustness checks using different measures of high-quality corporate innovation in Table B13. In Columns 1 and 2, the dependent variable is "home-run patents" (i.e., patents in the top 5% of their cohort in terms of citations received) and "breakthrough patents" (patents in the top 1% of their cohort in terms of citations received), respectively. In Columns 3 and 4, the dependent variable is corporate publications coauthored with AMWS scientists (Column 3) and cited by AMWS scientists (Column 4), respectively. In Column 5, we use only firms' employment of AMWS scientists who have won major awards. Reassuringly, the coefficient estimates remain similar to those presented in Tables 5, 6, and 7.

Table B13: Measures of High-Quality Patents, Publications, and AMWS Scientists

Dependent variable:	(1) ln(1+Homerun patents)$_t$	(2) ln(1+Breakthrough patents)$_t$	(3) ln(1+Pub. with AMWS collab.)$_t$	(4) ln(1+Pub. cited by AMWS)$_t$	(5) ln(1+Award-winning AMWS sci.)$_t$
ln(1+Public invention)$_{t-1}$	-0.082***	-0.034**	-0.059***	-0.080**	-0.012
	(0.019)	(0.013)	(0.016)	(0.025)	(0.006)
ln(1+Human capital)$_{t-1}$	0.102***	0.047***	0.062***	0.105***	0.014
	(0.018)	(0.012)	(0.011)	(0.018)	(0.007)
ln($1+R&D stock)$_{t-1}$	0.082***	0.043***	0.048***	0.085***	0.007
	(0.011)	(0.007)	(0.010)	(0.014)	(0.004)
Year FE	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Mean DV	2.24	0.36	0.52	2.25	0.64
Weak id. (Kleibergen-Paap)	185.14	185.14	185.14	185.14	185.14
Firms	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698

Notes: This table presents estimates from the second stage of 2SLS regressions using measures of high-quality corporate innovation as the dependent variables. *Homerun patents* are in the top 5% of citations relative to all patents granted in the same year. *Breakthrough patents* are in the top 1% of citations up to five years after publication, relative to all patents filed the same year and in the same technology field. In column 3, we include only firms' publications that are *coauthored with AMWS scientists*. In column 4, we include only firms' publications that are *cited by AMWS scientists*. In column 5, we include only the employment of *award-winning AMWS scientists*. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

B.4 Alternative Measure of Frontier Firms

Table B14: Alternative Measure of Frontier Firms: First Patent in CPC

Dependent variable:	(1) ln(1+Pat.)$_t$	(2) ln(1+Pub.)$_t$	(3) ln(1+AMWS sci.)$_t$	(4) ln(1+Pat.)$_t$	(5) ln(1+Pub.)$_t$	(6) ln(1+AMWS sci.)$_t$
ln(1+Public invention)$_{t-1}$ × $Tech\ frontier_t$	0.058***	0.023***	0.007***			
	(0.003)	(0.003)	(0.002)			
ln(1+Human capital)$_{t-1}$ × $Tech\ frontier_t$				0.058***	0.023***	0.009*
				(0.005)	(0.005)	(0.004)
ln(1+Public invention)$_{t-1}$	-0.253***	-0.154***	-0.040**	-0.252***	-0.153***	-0.040**
	(0.034)	(0.024)	(0.015)	(0.034)	(0.024)	(0.015)
ln(1+Human capital)$_{t-1}$	0.333***	0.129***	0.053***	0.331***	0.128***	0.053***
	(0.032)	(0.020)	(0.013)	(0.032)	(0.020)	(0.013)
ln($1+R&D stock)$_{t-1}$	0.228***	0.161***	0.035***	0.226***	0.160***	0.035***
	(0.019)	(0.016)	(0.009)	(0.018)	(0.016)	(0.009)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	28.30	15.72	4.80	28.30	15.72	4.80
Weak id. (Kleibergen-Paap)	123.34	123.34	123.34	123.85	123.85	123.85
Firms	3,372	3,372	3,372	3,372	3,372	3,372
Observations	41,698	41,698	41,698	41,698	41,698	41,698

Notes: This table presents the second stage of 2SLS estimation for the effect of public invention and human capital on corporate patents, publications, and AMWS scientists when considering firm proximity to the technology frontier. To measure this proximity, we first count each firm's annual flow of novel patents, where patent novelty is based on patents that are first to be granted in a new CPC main group or subgroup. Then, we create the variable *Tech frontier* as an indicator equal to 1 for firm years with a flow of novel patents in the top decile compared to other sample firms in that year, and 0 otherwise. Standard errors (in parentheses) are robust to arbitrary heteroskedasticity and allow for serial correlation through clustering by firms.

Appendix C Additional Descriptive Statistics and Case Examples

Table C15: Correlations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) Public knowledge$_{t-1}$	1.00								
(2) Human capital$_{t-1}$	0.31	1.00							
(3) Human capital, cited$_{t-1}$	0.19	0.04	1.00						
(4) Human capital, OECD$_{t-1}$	0.42	0.19	0.30	1.00					
(5) Public invention$_{t-1}$	0.44	0.04	0.34	0.68	1.00				
(6) Public invention, broad$_{t-1}$	0.38	0.00	0.38	0.63	0.90	1.00			
(7) R&D stock$_{t-1}$	0.16	0.62	0.07	0.08	0.04	0.02	1.00		
(8) Sales$_{t-1}$	0.11	0.53	-0.01	0.05	-0.03	-0.04	0.63	1.00	
(9) Awards to focal firm$_{t-1}$	0.05	0.24	-0.01	0.00	-0.01	-0.02	0.14	0.16	1.00

Notes: This table displays pairwise Pearson correlations for the main and alternative independent variables, as well as the control variables, included in our econometric analyses.

Table C16: Summary Statistics by Main Industry

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Computer, IT, software		Electronics, semicond.		Machinery, equipment, sys.		Life sciences	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Public knowledge$_{t-1}$	55,891	79,005	57,437	70,224	63,374	101,788	139,621	104,963
Human capital$_{t-1}$	4,571	8,976	7,240	9,997	5,372	6,970	6,140	10,046
Public invention$_{t-1}$	190	271	144	208	196	309	1,072	872
Patents$_t$	55	401	50	188	14	50	15	55
Publications$_t$	25	181	12	52	3	14	44	163
AMWS scientists$_t$	6	48	3	20	1	8	11	66
R&D expenditures ($ mm)$_t$	239	991	162	605	46	168	214	882
Public knowledge$_t$/$Publications_t$	44,990	71,936	35,623	66,041	52,865	84,735	46,352	74,729
Public invention$_t$/$Patents_t$	90	218	39	101	88	208	492	690

Notes: This table provides summary statistics by main industry for our analysis sample. Industry classification is based on a firm's primary SIC4 code.

Table C16: Summary Statistics by Main Industry (Cont.)

	(1) Telecommunication	(2)	(3) Transportation	(4)	(5) Others	(6)
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Public knowledge$_{t-1}$	33,462	58,792	44,321	54,390	43,667	64,387
Human capital$_{t-1}$	5,006	9,119	10,964	13,245	6,689	9,990
Public invention$_{t-1}$	79	124	48	85	120	260
Patents$_t$	31	146	54	152	21	87
Publications$_t$	13	118	21	61	10	51
AMWS scientists$_t$	4	37	9	24	4	14
R&D expenditures ($ mm)$_t$	117	564	489	1,564	82	306
Public knowledge$_t$/$Publications_t$	35,066	59,613	23,768	42,498	29,188	51,975
Public invention$_t$/$Patents_t$	36	89	10	42	43	149

Notes: This table provides summary statistics by main industry for our analysis sample. Industry classification is based on a firm's primary SIC4 code.

Table C17: Summary Statistics for Alternative Measures of Public Science

	(1)	(2)	(3)	(4) Distribution	(5)	(6)
	Obs.	Mean	Std. dev.	10th	50th	90th
Public invention, broad$_{t-1}$	41,698	6,942	14,085	0.0	1,396.6	21,609.5
Human capital, cited$_{t-1}$	41,698	2	5	0.0	0.5	7.4
Human capital, OECD$_{t-1}$	41,698	1,509	833	0.0	1,458.9	2,550.1

Notes: This table provides summary statistics for the alternative measures of public science used in the robustness checks. The analysis sample is at the firm-year level and includes an unbalanced panel of 3,372 U.S.-headquartered publicly traded firms from 1986 to 2015.

Table C18: Cross Tabulation of Measures of Human Capital and Public Invention

	(1) High human capital		(2) Low human capital		(3) Total	
	Count	%	Count	%	Count	%
High public invention	691	62%	728	32%	1,419	42%
Low public invention	416	38%	1,537	68%	1,953	58%
Total	1,107	100%	2,265	100%	3,372	100%

Notes: This table provides a cross-tabulation of measures of relevant *Human capital* and *Public invention* stock for the 3,372 firms included in our estimation sample. High (low) means above (below) the median compared to other sample firms. The unit of analysis is a firm.

Table C19: Mean Comparison Tests: Frontier Firms Versus Follower Firms

	(1) Frontier firms - Follower firms	(2)	(3) Frontier firms	(4)	(5) Follower firms	(6)
	Difference in means	t	Mean	Std. dev.	Mean	Std. dev.
A. Unique IPC combination						
Public knowledge$_{t-1}$	57,217	56.5	114,182	58,103	56,965	85,900.1
Human capital$_{t-1}$	24,069	102.0	28,132	14,974	4,063	4,864.3
Public invention$_{t-1}$	-23	-3.8	245	349	268	526.4
Patents$_t$	236	33.5	241	449	5	12.3
Publications$_t$	126	29.8	130	270	3	16.3
AMWS scientists$_t$	37	24.7	39	96	1	4.3
R&D expenditures ($ mm)$_t$	972	35.2	1,009	1,740	37	113.8
R&D stock ($ mm)$_{t-1}$	4,758	34.3	4,897	8,856	138	440.8
Sales ($ mm)$_{t-1}$	19,743	32.0	20,912	39,247	1,169	4,753.0
B. Patent value						
Public knowledge$_{t-1}$	13,084	6.6	74,886	93,828	61,802	84,701
Human capital$_{t-1}$	-2,245	-22.5	4,296	4,210	6,541	9,930
Public invention$_{t-1}$	142	10.7	400	637	258	502
Patents$_t$	-24	-27.7	6	14	30	162
Publications$_t$	-13	-25.8	3	9	16	96
AMWS scientists$_t$	-4	-20.5	1	5	5	33
R&D expenditures ($ mm)$_t$	-122	-28.3	28	108	150	676
R&D stock ($ mm)$_{t-1}$	-502	-26.4	130	483	632	3,223
Sales ($ mm)$_{t-1}$	-2,589	-22.0	661	4,455	3,250	14,712
C. First patent in CPC						
Public knowledge$_{t-1}$	51,430	37.4	110,889	65,385	59,459	85,495.3
Human capital$_{t-1}$	21,806	59.3	26,908	18,311	5,102	7,090.8
Public invention$_{t-1}$	-20	-2.6	248	355	267	520.3
Patents$_t$	277	25.2	289	551	12	47.4
Publications$_t$	133	23.3	141	285	8	55.3
AMWS scientists$_t$	37	20.2	39	91	3	22.5
R&D expenditures ($ mm)$_t$	1,064	27.3	1,135	1,928	71	344.8
R&D stock ($ mm)$_{t-1}$	5,271	26.7	5,557	9,860	286	1,602.0
Sales ($ mm)$_{t-1}$	21,347	25.1	23,159	42,509	1,813	8,672.4

Notes: This table compares *Frontier firms* with *Follower firms* using various measures of relevant public science and corporate innovation. We identify *Frontier firms* using (A) a firm-year measure based on unique IPC combinations; (B) a firm-year measure based on patent values from Kogan et al. (2017); and (C) a firm-year measure based on first patenting in a CPC main group or subgroup. This table includes the 3,372 firms and 41,698 firm-years included in our estimation sample. The two-sample t-tests use unequal variances.

Table C20: Mean Comparison Tests: Federal Lab Publications Versus Other Publications

	(1)	(2)	(3)	(4)	(5)	(6)
	Federal lab pub. - Other pub.		Federal lab pub.		Other pub.	
	Difference in means	t	Mean	Std. dev.	Mean	Std. dev.
A. All subjects						
Funding per pub. ($)	45,837.1	4.7	60,669.3	62,099.9	14,832.3	11,278.0
Patent citations per pub.	0.1	2.7	0.5	0.3	0.3	0.2
Publication citations per pub.	7.4	6.2	19.8	7.0	12.4	3.1
Authors per pub.	5.1	5.1	9.1	6.4	4.0	0.8
B. Biology and medicine						
Funding per pub. ($)	75,537.5	4.8	96,557.2	99,607.0	21,019.6	15,397.4
Patent citations per pub.	0.3	4.2	0.7	0.4	0.4	0.2
Publication citations per pub.	9.2	6.9	24.0	7.7	14.8	3.6
Authors per pub.	1.3	3.4	5.9	2.2	4.6	1.1
C. Chemistry						
Funding per pub. ($)	22,186.6	4.0	33,904.5	34,321.2	11,717.9	8,909.5
Patent citations per pub.	0.1	1.8	0.6	0.4	0.4	0.3
Publication citations per pub.	7.8	4.1	21.1	10.6	13.3	5.9
Authors per pub.	0.7	2.7	4.6	1.4	4.0	0.7
D. Physics and engineering						
Funding per pub. ($)	20,534.0	3.9	29,317.9	32,537.8	8,783.9	8,222.8
Patent citations per pub.	0.0	0.1	0.3	0.2	0.3	0.2
Publication citations per pub.	6.0	5.5	15.6	6.4	9.5	2.8
Authors per pub.	9.3	5.4	12.7	11.0	3.4	0.6

Notes: This table compares annual averages over 1980-2020 for publications coauthored by scientists affiliated with the federal laboratories (*Federal lab pub.*) versus all other publications (*Other pub.*). *Funding per pub. ($)* is the average dollar amount of grant funding supporting a publication, deflated to constant 2012 dollars. *Patent citations per pub.* is the average number of citations received by a publication from patents. *Publication citations per pub.* is the average number of citations received by a publication, within five years, from other publications. *Authors per pub.* represents the average number of authors of a publication. The two-sample t-tests use unequal variances.

C.1 Examples of Relevant Human Capital

We validate the logic behind two of our measures of firm-relevant human capital with three case examples, as summarized in Tables C21 and C22 and detailed below.

Our primary measure of relevant human capital relies on the textual similarity between the abstracts of dissertations and the abstracts of firm patents. Specifically, a PhD graduate is relevant to a firm's R&D if his/her dissertation defended in year t is in the top 1,000 most similar dissertations to one or more of the firm's patents granted in years $[t-1, t+1]$.[36] For each PhD graduate from Column 1, we list the top 3 firms (Column 2) and up to 3 most similar patents per firm (Column 3) based on the textual similarity between abstracts.

Our complementary measure, *Human capital, cited*, relies on non-patent literature citations from patents in various CPC subclasses to the published version of the dissertation.[37] For each PhD graduate-firm pair, we list up to three patents that cite the published version of the dissertation (Column 4). We also list whether the PhD graduate worked for the firm, and during which years (Column 5).

Example 1: Dr. David Nichols

This example demonstrates that Dr. David Nichols' PhD dissertation was highly relevant to multiple firms, including Xerox, Microsoft, and Sun Microsystems, as evidenced by its citation in numerous patents assigned to these companies. Furthermore, Dr. Nichols' career trajectory included employment at Xerox and Microsoft, where he made substantial contributions to their patent portfolios over time.

Dr. Nichols earned his PhD in computer science from Carnegie Mellon University. In 1989, he defended his dissertation, titled "Multiprocessing in a network of workstations" (ProQuest document ID 303690418).

Dr. Nichols' research focused on enhancing the efficiency and reliability of computer network systems connecting distinct computers, known as "network file systems." These systems, which facilitated the collaboration of multiple computers operating as a loosely coupled collection of workstations, enabled them to work together without being tightly connected. Specifically, Dr. Nichols' research focused on the Andrew File System (AFS), a system widely used in UNIX applications for memory sharing among computers. To obtain a better understanding of AFS's performance, Dr. Nichols built a model utilizing a discrete-event simulation technique. The model allowed him to analyze the interactions between the AFS server and the linked computers. Through simulations, he gained valuable insights into the intricacies of AFS's functioning and developed strategies for its optimal performance.

Dr. Nichols's dissertation is closely linked to the prior work of his advisor, Dr. James H. Morris, at Xerox. From 1974 to 1983, Dr. Morris contributed to the creation of the Alto System at the Xerox Palo Alto Research Center, which was a groundbreaking innovation that formed the basis for modern personal computers. After leaving Xerox, Dr. Morris served as the director of Carnegie Mellon University's Information Technology Center from

[36] Exactly *how* relevant the graduate is depends on the maximum similarity between the dissertation and all the patents granted to the firm in a 5-year time cohort.

[37] We take into consideration the degree of relevance by using a focal firm's shares of patents across CPC subclasses during the previous 5-year time cohort as weights.

Table C21: Examples of Relevant Human Capital

(1) PhD graduate (graduation year)	(2) Firm / patent assignee	(3) Similar patents (grant year)	(4) Citing patents (application year)	(5) Worked for firm (years)
David Nichols (1989)	Xerox Corp.	4737931 (1988) 4843542 (1989) 4974173 (1990)	5469099 (1993)	Yes (1990-1996)
	Microsoft Corp.	4779187 (1988) 4825358 (1989) 4974159 (1990)	6981138 (2001) 7770023 (2005) 8112452 (2009)	Yes (since 2003)
	Sun Microsystems Inc.	4719569 (1988) 4884266 (1989) 4937734 (1990)	6134603 (1998) 6925644 (2003) 7660887 (2003)	No
Siddharth Ramachandran (1998)	Lucent Technologies Inc.	5596668 (1997) 5847690 (1998) 5858052 (1999)	6463088 (2000)	Yes (1998-2001)
	Micron Technology Inc.	5629246 (1997) 5837564 (1998) 5906771 (1999)		No
	Eastman Kodak	5629418 (1997) 5714301 (1998) 5916946 (1999)		No
Dirk Balfanz (2001)	Xerox Corp.	6016516 (2000) 6176425 (2001) 6340931 (2002)		Yes (2001-2007)
	Microsoft Corp.	6012052 (2000) 6172972 (2001) 6338079 (2002)	8782527 (2007) 8719847 (2010)	No
	Intel Corp.	6023509 (2000) 6173315 (2001) 6343067 (2002)		No

Notes: This table presents three case examples of relevant human capital. Columns 2 and 3 list three firms and up to three most similar patents per firm, respectively, based on the textual similarity between the abstract of the PhD dissertation and the abstracts of firm patents. Column 4 lists up to three patents per firm that cite the published version of the dissertation. Column 5 indicates whether the PhD graduate worked for the firm, as well as the years of employment, if applicable.

1983 to 1988. During this time, he collaborated with IBM to create a prototype university computing system called Andrew, on which Dr. Nichols's research was built. Dr. Nichols conducted an extensive comparison of his model's findings with real-world experiments on the AFS system. He delved into various factors that could impact the system's performance, such as network latency, processing speed, and hard drive access time. Through his research, he provided valuable insights into the functioning of the AFS system, demonstrating that its performance is primarily limited by the processing power of the involved computers. Furthermore, he found that AFS could handle a diverse range of tasks without becoming overwhelmed.

Table C22: Examples of Textually Similar Patents

(1) Patent number	(2) Patent title
4737931	Memory control device
4843542	Virtual memory cache for use in multi-processing systems
4974173	Small-scale workspace representations indicating activities by other users
4779187	Method and operating system for executing programs in a multi-mode microprocessor
4967378	Method and system for displaying a monochrome bitmap on a color display
4974159	Method of transferring control in a multitasking computer system
4719569	Arbitrator for allocating access to data processing resources
4884266	Variable speed local area network
4937734	High speed bus with virtual memory data transfer and rerun cycle capability
5596668	Single mode optical transmission fiber, and method of making the fiber
5847690	Integrated liquid crystal display and digitizer having a black matrix layer adapted for sensing screen touch location
5858052	Manufacture of fluoride glass fiber with phosphate coatings
5629246	Method for forming fluorine-doped glass having low concentrations of free fluorine
5837564	Method for optimal crystallization to obtain high electrical performance from chalcogenides
5906771	Manufacturing process for high-purity phosphors having utility in field emission displays
5629418	Preparation of titanyl fluorophthalocyanines
5714301	Spacing a donor and a receiver for color transfer
5916946	Organic/inorganic composite and photographic product containing such a composite
6016516	Remote procedure processing device used by at least two linked computer systems
6176425	Information management system supporting multiple electronics tags
6340931	Network printer document interface using electronics tags
6012052	Methods and apparatus for building resource transition probability models for use in pre-fetching resources, editing resource link topology, building resource link topology templates, and collaborative filtering
6172972	Multi-packet transport structure and method for sending network data over satellite network
6338079	Method and system for providing a group of parallel resources as a proxy for a single shared resource
6023509	Digital signature purpose encoding
6173315	Using shared data to automatically communicate conference status information within a computer conference
6343067	Method and apparatus for failure and recovery in a computer network

Notes: This table lists the titles of the textually similar patents from Table C21.

Using the SPECTER algorithm, we found that Dr. Nichols's dissertation is textually similar to several corporate patents granted in 1988-1990, as listed in Tables C21 and C22. For example, the 1989 patent titled "Virtual memory cache for use in multi-processing systems" (USPTO patent number 4843542) assigned to Xerox and Dr. Nichols' dissertation are both closely related to multi-processing systems, with a shared focus on improving performance and efficiency within such systems. The patent introduced a virtual memory cache that enhances performance and efficiency. Meanwhile, Dr. Nichols' dissertation explored the use of multiple processors in a network of workstations to optimize processing power and resource sharing. Overall, both the patent and the dissertation highlight the importance of performance and efficiency in multi-processing systems.

Dr. Nichols's dissertation was published in 1988 in *ACM Transactions on Computer Systems* under the title "Scale and performance in a distributed file system." This publication has been cited by 442 patents, including patents assigned to Sun Microsystems, Lucent Technologies, IBM, Xerox, EMC, Unisys Corporation, Microsoft, Oracle, NetApp, Hewlett Packard, Google, and AT&T, among others. For example, the 1993 patent titled "Method for delegating access rights through executable access control program without delegating access rights not in a specification to any intermediary nor comprising server security" (USPTO patent number 5649099) assigned to Xerox outlines a method that allows users to securely delegate specific access rights to others, even without complete trust. By utilizing rules called access control programs, the system maintains controlled and secure shared access. As a result, the system can decide whether to grant or deny a request.

Dr. Nichols was employed at Xerox PARC from 1990 to 1996.[38] Later, he joined Microsoft in 2003 (where he was still employed as of 2023). Although Dr. Nichols did not publish any scientific articles after completing his PhD degree, he contributed to numerous patents at both Xerox and Microsoft, many of which were related to his dissertation. While working at Xerox, Dr. Nichols played a significant role in designing and implementing the Tapestry system, which facilitates automatic filtering of electronic messages based on human feedback. He also co-led the Jupiter project, aimed at supporting collaboration through the concept of "network places." Some of his notable inventions at Xerox include "Method for controlling real-time presentation of audio/visual data on a computer system" (USPTO patent number 5692213), filed in 1995. Some of Dr. Nichols's notable inventions at Microsoft include "Method and system for resolving conflicts operations in a collaborative editing environment" (USPTO patent number 7792788), filed in 2005, and "Deployment, maintenance, and configuration of complex hardware and software systems" (USPTO patent number 7676806), also filed in 2005.

Example 2: Dr. Siddharth Ramachandran

This example demonstrates that Dr. Siddharth Ramachandran's published dissertation has significantly influenced the field of optical fiber and photonics devices, with multiple firms citing his work in their patents, including Lucent Technologies. Moreover, his expertise led him to work for renowned institutions like Bell Labs and OFS Labs, further contributing to advancements in the industry through his published research and patented inventions.

Dr. Ramachandran earned his PhD in electrical and computer engineering from the University of Illinois Urbana-Champaign in 1998. His dissertation, titled "Photoinduced optical integrated circuits and bulk photonic devices in chalcogenide glasses," delves into the unique properties of chalcogenide glasses that make them valuable for technological applications. Dr. Ramachandran explores how exposure to light can alter the structure of these glasses and enable energy transfer to rare earth elements, potentially benefiting lasers and communication devices. Additionally, he investigates methods to enhance the stability and longevity of the glass through heating processes and determines optimal operating conditions.

Using the SPECTER algorithm, we found that Dr. Ramachandran's dissertation is textually similar to several corporate patents granted between 1997 and 1999, as listed in Tables C21 and C22. For example, the 1997 patent titled "Single mode optical transmission

[38] https://www.linkedin.com/in/david-nichols-6829331/

fiber, and method of making the fiber" (USPTO patent number 5596668) assigned to Lucent Technologies has a strong link to Dr. Ramachandran's dissertation due to their shared focus on optical technologies, material properties, light interaction, and potential applications in telecommunications. Both the patent and dissertation concentrate on optical technologies, with Dr. Ramachandran's thesis examining chalcogenide glasses' properties for use in photonic, laser, and communication devices, while the patent is concerned with single-mode optical transmission fibers that are essential components of optical communication systems. Additionally, the patent and dissertation both involve the study of specific materials and their properties, with Dr. Ramachandran researching chalcogenide glasses' unique characteristics, and the patent concentrating on the manufacturing process of single-mode optical transmission fibers. Moreover, both the dissertation and the patent investigate materials and their interaction with light, with Dr. Ramachandran examining how exposure to light can alter the structure of chalcogenide glasses and enable energy transfer to rare earth elements, and the patent discussing an optical fiber that can efficiently transmit light signals over long distances. Finally, the patent and the dissertation have a connection to potential telecommunication applications. Optical fibers are a critical technology in modern telecommunication systems, while chalcogenide glasses have potential applications in communication devices and could be incorporated into future optical technologies.

Dr. Ramachandran's dissertation, entitled "Low-loss photoinduced waveguides in rapidly thermally annealed films of chalcogenide glasses," was published in Applied Physics Letters in 1999. This publication has been cited by eight patents, including one assigned to Lucent Technologies and titled "Mesa geometry semiconductor light emitter having chalcogenide dielectric coating" (USPTO patent number 6463088). Dr. Ramachandran's research has had a significant impact on the development of various inventions at Lucent Technologies, including light-emitting diodes (LEDs), laser diodes, and optoelectronic devices. LEDs are utilized in a broad range of applications, such as displays, indicator lights, and general lighting. The mesa geometry semiconductor light emitter with a chalcogenide dielectric coating has the potential to enhance the performance of LEDs, making them more energy-efficient and durable. Laser diodes are used in many applications, including data communications, optical storage, sensing, and medical equipment. The patented technology can improve the efficiency and output power of laser diodes, leading to better overall performance. Furthermore, the semiconductor light emitter outlined in the patent can be integrated into a variety of optoelectronic devices, such as photodetectors, optical modulators, or optical amplifiers, ultimately improving their performance.

Dr. Ramachandran's professional career spans over a decade of optical fiber and photonics device research. He began his work as a member of the technical staff at Bell Labs, a division of Lucent Technologies, in 1998 and continued until 2001. He then joined OFS Labs, a world-renowned institution in optical research and product development, where he worked from 2001 to 2009. Throughout his career, Dr. Ramachandran has authored numerous research articles on these subjects, including "Photoinduced index-tapered channel waveguides in chalcogenide glasses for guided mode-size conversion" published in 1998, "Spatially and spectrally resolved imaging of modal content in large-mode-area fibers" in 2008, and "Generation and propagation of radially polarized beams in optical fibers" in 2009. These publications reflect his expertise in optical fiber and photonics device research and demonstrate his contribution to the field.

Dr. Ramachandran has made significant contributions to multiple patented inventions during his tenure at Bell Labs and OFS Labs. For example, in 2007 he filed a patent titled "Visible continuum generation utilizing a hybrid optical source" on behalf of OFS Labs. This invention focuses on generating a visible light continuum by employing a hybrid optical source, which has potential applications in fields such as microscopy, imaging, and optical communications. In 2009, he filed another patent on behalf of OFS Labs called "Systems and techniques for generating Bessel beams" This patent involves the development of systems and methods for producing Bessel beams, a type of non-diffracting light beam that maintains its intensity profile over a long distance, making it highly useful in applications like optical trapping and laser machining. Both patented inventions are closely related to Dr. Ramachandran's doctoral studies. By building upon his doctoral research, Dr. Ramachandran has continued to contribute to the advancement of optical fiber and photonics devices.

Example 3: Dr. Dirk Balfanz

This example shows that Dr. Dirk Balfanz's dissertation on access control for ad-hoc collaboration has been highly relevant and influential in the field of information technology, as demonstrated by numerous citations in patents from such major tech companies as Microsoft and Xerox. Dr. Balfanz's expertise led him to work for both Xerox PARC and Google.

In 2001, Dr. Dirk Balfanz obtained his PhD in computer science from Princeton University. His dissertation titled "Access control for ad-hoc collaboration" explored various approaches to managing access control during interactions with unknown or untrusted parties. Dr. Balfanz demonstrated that it is possible to protect resources while allowing ad-hoc collaborations to take place, even though it may seem counterintuitive. His research has paved the way for refining access control logic, enhancing user-computer interaction models, and aiding programmers in securely dividing applications. The ultimate objective is to address the security challenges in our increasingly interconnected world, where ad-hoc collaborations are becoming more prevalent.

Dr. Balfanz acknowledged the substantial contributions of his collaborators at Microsoft Research and Xerox PARC. While at Microsoft Research, Dan Simon conceived the WindowBox idea, and Paul England offered guidance on Windows programming. At Xerox PARC, Drew Dean collaborated with Dr. Balfanz on the Placeless access control logic, while Doug Terry, Jim Thornton, and Mike Spreitzer provided further assistance. Ian Goldberg's expertise was critical in porting SSLeay to the PalmPilot, enhancing Copilot, and sharing programming tips for the Pilot. Bob Relyea from Netscape provided assistance with specific PKCS#11 details. Lastly, Andrew Appel's role in establishing the decidability proof for the logic in Chapter 2 of the dissertation was pivotal.

Using the specter algorithm, we found that Dr. Balfanz's dissertation is textually similar to various corporate patents granted between 2000 and 2002, as listed in Tables C21 and C22. For example, the patent "Remote procedure processing device used by at least two linked computer systems" granted to Xerox (USPTO patent number 6016516) and Dr. Balfanz's dissertation are connected within the broader context of information technology, with a focus on distributed systems and collaboration. Both address challenges in these areas, with the patent aimed at improving the ease and efficiency of combining and executing remote procedures, and the dissertation emphasizing the importance of security and access control in ad-hoc collaborations. Ultimately, both works contribute to enhancing the user experience in collaborative settings by facilitating seamless interactions with remote resources and ensuring secure access to shared information.

Dr. Balfanz's dissertation was published in the *2002 Proceedings of the ACM Conference on Computer-Supported Cooperative Work* under the title "Using speakeasy for ad hoc peer-to-peer collaboration." This publication has been cited by 23 patents, including those from Microsoft. Two such patents are 8719847, titled "Management and marketplace for distributed home devices," and 8782527, titled "Collaborative phone-based file exchange." Patent 8719847 deals with managing and integrating distributed home devices in a networked environment. The patent addresses challenges in securely connecting and managing devices, enabling users to share and access resources in a controlled manner. Dr. Balfanz's dissertation, which emphasizes the importance of secure access control in ad-hoc collaborations, is related to this patent in the sense that both works explore the need for secure and controlled sharing of resources in networked environments. Patent 8782527 focuses on facilitating secure and efficient file exchange between phones in a collaborative setting. The patent describes methods and systems for securely sharing files among collaborating parties using mobile devices. Dr. Balfanz's dissertation is related to this patent as both works share a focus on the importance of secure collaboration and access control when sharing resources, such as files, among multiple users.

Dr. Balfanz worked as a research staff member at Xerox from 2001 to 2007. After his tenure at Xerox, he joined Google in 2007 as a software engineer, where he focused on security, privacy, and abuse prevention.[39]

Although Dr. Balfanz has published papers like "Security Keys: Practical Cryptographic Second Factors for the Modern Web" in 2016 and "Origin-Bound Certificates: A Fresh Approach to Strong Client Authentication for the Web" in 2012, his primary contributions have come in the form of patent inventions. He has made numerous contributions to inventions during his time at Xerox PARC, including "Apparatus and methods for providing secured communication" (USPTO patent number 7392387) filed in 2007, and "Systems and methods for authenticating communications in a network" filed in 2004. Additionally, he has contributed to inventions at Google, including "System and method for authenticating to a participating website using locally stored credentials" filed in 2012, and "Methods and systems of adding a user account to a device" filed in 2014.

[39] https://www.linkedin.com/in/dirk-balfanz-7885852/