The Rise of American Corporate Science

Ashish Arora* Sharon Belenzon† Konstantin Kosenko‡ Jungkyu Suh§
Yishay Yafeh¶‖

January 16, 2021

[PRELIMINARY DRAFT. PLEASE DO NOT CIRCULATE]

Abstract

Corporate science in America emerged in the interwar period, as some companies set up state-of-the-art corporate laboratories, employed highly-skilled scientists, and conducted basic research of the kind we would associate today with academic institutions. Using a newly-assembled data set on both publicly-traded and private US companies between 1926 and 1940 with information on corporate ownership, corporate organization (diversification and business group affiliation) and corporate research (publications, patents and their citations and market values), we attempt to explain the historical rise of corporate quasi-academic research. We argue that this phenomenon was driven by companies trying to make up for institutional voids, in this case, the weakness of academic research, which made access to science in certain fields costly for US firms. Measuring institutional voids, or field-specific scientific backwardness, in two different ways, we find that large and diversified corporations, especially those relatively close to the technological frontier, were the ones prone to internalize the provision of basic research, a classic public good. We also find that corporate research was positively correlated with innovative and valuable patents (inventions) and with stock market valuation, suggesting that investment in corporate research paid off. The results shed light on the link between corporate organization and corporate science and, in particular, on the possible role of large firms and diversified business groups in advancing science in less-than-fully-developed economies.

*Duke University, Fuqua School of Business and NBER (ashish.arora@duke.edu)
†Duke University, Fuqua School of Business and NBER (sharon.belenzon@duke.edu)
‡Bank of Israel (konstantin.kosenko@boi.org.il)
§Duke University, Fuqua School of Business (jungkyu.suh@duke.edu)
¶Hebrew University of Jerusalem, Jerusalem School of Business Administration, CEPR and ECGI (yishay.yafeh@huji.ac.il)
‖We thank seminar participants at Boston University and the ASSA 2021 annual meeting, as well Tim Simcoe and Bhaven Sampat for helpful comments and suggestions. All remaining errors are our own.

1 Introduction

Today, we typically think of universities as the locus where scientific knowledge is produced, whereas private corporations apply the scientific knowledge generated in academia. This division of innovative labor, however, is historically recent. In this paper we show that in the interwar period, some U.S. corporations engaged in the pursuit of university-like basic science, whereas others carried out applied research as companies today tend to do.

We argue that the initiation of corporate science is related to "institutional voids" – a phenomenon whereby, in less-than-fully-developed economies as the U.S. may have been, at least in certain respects before WW-II, certain companies can make up for the missing (market or other) institutions and provide quasi-public goods themselves (Belenzon & Berkovitz, 2010; Belenzon, Berkovitz, & Rios, 2013; Belenzon & Tsolmon, 2016; Khanna, 2000; Khanna & Yafeh, 2007). For example, in the early stages of Korea's industrialization, the Hyundai group provided facilities for training labor and carrying out (applied) research, overcoming the weakness of Korea's academic institutions at the time. There are also many examples of conglomerates and business groups making up for the deficiencies of financial markets in emerging (and occasionally developed) economies by operating internal capital markets that can overcome the information and agency problems plaguing public markets. In addition, Khanna and Yafeh (2007) describe cases of diversified business groups making up for missing institutions related to entrepreneurship (in India, for example), contract enforcement and more. By combining data on patent and ownership chains among European firms, Belenzon and Berkovitz (2010) find that business groups may foster innovation via internal capital markets.

In the context of corporate research in the United States of the interwar period, the major institutional void we focus on is the weakness of American academia in certain scientific fields (e.g., organic chemistry), which were dominated by European, especially German, universities. Although hard to imagine given the standing of American universities today, the relative backwardness of U.S. academic institutions in certain scientific disciplines before World War II made access to science in these fields costly for U.S. companies, especially in the 1930s when international trade became restricted. In response, we show that some (though not all) U.S. corporations, primarily those belonging to large and diversified corporate entities and those close to the technological frontier, set up research laboratories and employed skilled scientists to carry out basic science. These companies, we argue, were large and diversified enough to make the provision of basic research worth their while.

We study these issues using a newly-assembled, historical data, which builds on, and extends, the data set constructed by Kandel, Kosenko, Morck, and Yafeh (2019) (hereafter KKMY) to study regulation-induced changes in U.S. corporate ownership in the 1930s and 1940s. The data set includes both publicly-traded and private firms which were active in corporate science (academic publications) or applied research (patenting) during the period 1926 to 1940.[1] We emphasize the inclusion of private firms, as some of them played an important role in corporate research and omitting them might create an incomplete picture of corporate science in interwar America. As in KKMY, we use information on corporate ownership and control. In particular, we identify firms that are controlled by other (parent) firms, or control their own subsidiaries, and firms which are part of diversified business groups with multiple affiliates, often operating across industries. Such inter-firm ties matter because, prior to 1940, U.S. corporate ownership was quite different from what it is today; multi-firm entities—business groups and conglomerates of various types—dominated the U.S. economy. It turns out that these diversified corporate entities played an especially important role in corporate science. We supplement the information on corporate ownership and organization with data on firm size (assets) and profitability, collected from Moody's Manuals and from the Center for Research in Securities Prices (CRSP).

We then combine the KKMY data set with extensive information on corporate science derived from numerous sources. We use Microsoft Academic Graph (MAG) to assemble data on scientific publications authored by scholars employed by corporations, their scientific discipline, citations etc. We also match firms to the Industrial Research Laboratories of the United States (hereafter the IRL directory) on the existence of corporate science laboratories and on the personnel employed in them. We complement the information on corporate science by using information on patenting, the output of applied corporate research and development. Our data set includes not only the number of patents by company from the United States Patent and Trademark Office (USPTO, provided through Google patents and IFI claims services), but also other measures on patents such as forward prior art citations received and stock market values from Kogan, Papanikolaou, Seru, and Stoffman (2017).

Finally, for each firm, we construct measures of relative American backwardness by scientific field and measures of proximity to the technological frontier. Scientific backwardness is measured using two alternative methods. In the first method, we calculate the ratio of prominent American scientists trained in Europe divided by American scientists without experience in Europe. In the second method we calculate, for each scientific field, the number of academic publications by European authors in the field

[1] To be included in the analysis, companies must have had at least one publication or one patent during our sample period.

most relevant to the firm's operations, divided by all publications by U.S.-based scientists in the same field. Proximity to the technological frontier is measured through patent citations and a count of identical patent class combinations granted before a patent's grant date (Fleming, 2001). The combined dataset comprises of 469 firms and 7,035 firm-year observations.

Our historical sample period, ranging from 1926 to 1940, is particularly suitable for the issue at hand.[2] First and foremost, this is the period in which the phenomenon of corporate science emerged in the United States. Second, this is a period when U.S. academic institutions were still relatively weak, at least in certain disciplines, and therefore an ideal period to test the institutional voids (scientific backwardness) theory. Finally, our sample period ends before the onset of World War II, when the U.S. government became deeply involved in scientific development both directly and through military procurement contracts. During our sample period, the U.S. government had little influence on corporate research and development activities.

We attempt to make three contributions to the literature. First, we provide empirical evidence on, and explanations for, the rise of corporate research in America and the emergence of an innovation ecosystem which played an important role in promoting American science and the U.S. economy to world prominence. We show that in the interwar era, corporate basic research and academic science were substitutes (rather than complements). In this respect, the present study is closely related to Arora, Belenzon, and Patacconi (2019). The internalization of R&D functions of U.S. firms in the early twentieth century is closely studied by Lamoreaux and Sokoloff (1997) from the perspective of the independent inventors and their gradual conversion into salaried R&D employees in larger firms. Nicholas (2010, 2014) attributes one of the causes of this transition to the increasing complexity of the technology underlying the chemicals and electric industries. Our findings add a different perspective to the rise of corporate science. From a managerial perspective, in-depth accounts of specific firms and industries, as in Jenkins and Chandler (1975), Maclaurin and Harman (1949), Reich (1985) and Hounshell and Smith (1988) document the various motives behind the establishment of large R&D laboratories. Our findings highlight the role of scientific gaps (and the higher cost of accessing domestic university knowledge) as a motivator of R&D internalization. Moreover, our systematic evidence allows us to show that the response to this scientific gap was heterogeneous across firms, allowing us to ask why some firms responded while others did not.

Second, we provide evidence in support of the missing institutions (or institutional voids) frame-

[2]Our sample begins in 1926 as this is the first year in the KKMY sample and the year in which the CRSP data on stock market data becomes available.

work in a context hitherto not documented in the literature. Khanna (2000), and Khanna and Palepu (2000) apply this idea to emerging markets; the present paper finds that this concept may be useful in understanding phenomena related to the U.S. economy in the first half of the twentieth century. Quasi-academic research in America was driven by the attempts of large and diversified corporate entities to make up for the weakness of American academia at the time. The finding that large companies, and companies operating in multiple fields, were the ones most prone to invest in basic scientific knowledge is consistent with Nelson (1959b). It also provides a historical perspective on the current debate regarding the role of large U.S. corporations in advancing science and promoting innovation. Whereas Autor, Dorn, Katz, Patterson, and Van Reenen (2020) describe large U.S. corporations as highly innovative "superstar firms", others, such as Gutiérrez and Philippon (2020), view them as primarily inefficient entities shielded from competitive pressures.

Third, the data set we construct, combining information on U.S. corporate ownership and U.S. corporate science in the interwar period is the most comprehensive of its kind and, in our view, a contribution in and of itself. This data set will open the way to future research on the possible links between corporate characteristics, research and development, government policy and institutional context.

The rest of the paper is organized as follows. In the next section we summarize the rise of American corporate science in the early twentieth century. Section 3 describes our data. Sections 4 and 5 present our econometric specifications and estimation results. Section 6 concludes.

2 The rise of corporate science in America

Prior work has established that American firms in the early twentieth century internalized various managerial functions, one of which was research and development (Chandler Jr, 1993). Lamoreaux and Sokoloff (1997), in their study of American inventors between 1870 and 1911, show that independent inventors that had previously provided patents and contract R&D work from outside firm boundaries began to be directly employed by firms.[3] Between 1921 and 1940, the number of firms operating R&D labs increased more than sevenfold, from 297 to 2264, according to a survey produced by the National Research Council (Mowery & Rosenberg, 1998). However, prior work has not established whether the rise was even between the "R" and the "D": did American firms invest in scientific research ("R") or was

[3]The archetypal example of this evolution is Thomas Edison, who had started his career with contract research work for telegraph companies, eventually establishing his own firm with one of America's first internalized R&D laboratory at Menlo Park.

the internalization of R&D functions principally for testing and quality control ("D")?

Anecdotal evidence suggests that the early years of corporate research were limited to development activity, but that this gradually changed for certain firms circa the late 1920s and early 1930s. The first R&D laboratories had narrow and well defined goals: Charles Dudley's tenure at the Pennsylvania Railroad company that began in 1875 was focused on examining the metallurgical properties of the rail tracks that were supplied to the firm by steel companies. Thomas Edison's Menlo Park facility concerned itself primarily with product invention and was known for shunning fundamental investigations that Edison considered "purely aesthetic" (Wise, 1985).[4]

Robert Duncan, the founder of the Mellon Institute (the premier contract R&D organization of its time) and an advocate of corporate research, lamented that the factories of American firms were dominated by foremen that stuck to traditional practices and that managers were too myopic to wait the "two, three, or even five years" for scientific projects to reach their potential (Servos, 1994, pp. 223–225). Corporate researchers (and chemists in particular) were mainly tasked with routine analysis and troubleshooting, which motivated Duncan's establishment of a specialized contract research organization in the Mellon Institute in 1913. The institute grew steadily in contract dollars ($300,000 to $800,000) and staff size (83 to 145 industrial fellows funded by firms) between 1919 and 1929. Fellows such as George Curme made crucial contributions to replacing coal tar with petroleum for certain fine chemicals, while Union Carbide's contract with the institute yielded the glycol antifreeze, which became a key product for the firm (Servos, 1994, p. 223). This, and changes instigated by like-minded research managers such as Willis Whitney at GE Research Labs, Frank Jewett at AT&T Bell Labs, and Charles Stine at DuPont's Purity Hall signalled a deeper commitment to scientific research at American firms. As we show in section 3, American firms published three times as many scientific papers in 1940 as they did in 1926. The quality of research also improved dramatically: Irving Langmuir (GE) was awarded his Nobel Prize in chemistry (1932) and Clinton Davisson (AT&T) in physics (1937).

Why did firms invest in science? There are four explanations.[5] First, German chemical firms such as Bayer, Hoechst, and Agfa fared well in the international organic synthetic dyes market by building on their corporate research (Reich, 1985, p.41). This set a precedent to emulate for American firms in similar sectors. Second, American inventions were being challenged by European competition that brought a more fundamental understanding of natural phenomena. GE's electric lighting business that was started

[4] According to a foreign-trained lab employee, Edison argued that "We can't be like those old German professors who, as long as they can get their black bread and beer, are content to spend their whole lives studying the fuzz on a bee." (Wise, 1985)

[5] This section draws on (Arora, Belenzon, Patacconi, & Suh, 2020a).

by Edison in 1879, for instance, was based on carbon-filament high-vacuum incandescents. However, Walther Nernst (the 1920 Nobel Laureate in Chemistry) at Göttingen invented a glower that required no vacuum to operate and was more efficient. Westinghouse eventually acquired the patent rights to the Nernst glower in a bid to compete with GE (Wise, 1985). This was one of the motivations behind the establishment of GE Research Laboratory (GERL) in 1900. Langmuir, one of Nernst's American doctoral students, was also recruited by Whitney to further his work on surface chemistry.

Third, American firms had often reached the technological "frontier" and the costs of accessing external inventions were thereby increasing. When firms attempted to acquire technology from abroad, a lack of scientific sophistication prevented them from fully exploiting such imports. In hopes of starting rayon businesses, DuPont purchased the Bevan, Cross and Topham patents from England in the 1910s. However, the patents required scientific know-how that American firms did not yet possess, and it was only after British firms opened factories in the United States that viscose rayon production could begin in earnest (Hounshell, 1988). Moreover, the reversal of globalization after the Great Depression, manifested in various forms of capital controls and the erection of tariff walls, may have raised the costs of accessing knowledge outside of North America (Mauro, Sussman, & Yafeh, 2002, 2006; Rajan & Zingales, 2003). For instance, restrictions on trade would have made it costlier to reverse engineer imported products from Germany, while capital restrictions may have discouraged foreign investment necessary for the smooth transfer of technology. This may have further incentivized American firms to "make" the scientific knowledge they could not "buy" cheaply from European competitors.[6]

In other instances, the science required for improvement was still underdeveloped, which led to researching firms contributing to new scientific disciplines. U.S. petrochemical firms and chemical engineering illustrate this point. The U.S. chemical industry changed from using coal feedstocks to petroleum. This transition was partly facilitated by the abundance of this resource in Pennsylvania (first discovered in 1859) and later in Texas (1901). U.S. market demand for "lighter" petroleum products such as gasoline increased as innovations in internal combustion technology increased automobile sales.[7] With this transition, American petrochemical firms by the 1920s found themselves at the technological frontier when it came to refining hydrocarbons. There was an acute need for improved methods to refine crude oil into end products efficiently. Enos (1962) counts at least seven major cracking methods that increased thermal cracking capacity from under 100 barrels a day in 1913 with the Burton process to over 10,000 barrels

[6] Incidentally, the movement of talented emigrés fleeing Nazi persecution during the 1930s may have also aided with this change in strategy (Moser, Voena, & Waldinger, 2014).

[7] Before the advent of mass automobile transportation, the principal end purpose for petroleum products was for heating oil ("heavier" Kerosene).

a day in 1940 from the Houdry process. Firms often invented new processes (such as Esso's Tube and Tank method or Fluid Catalytic Cracking), and absorbed and supported new inventors where possible (e.g. SOCONY's support of the Houdry process from France).

In developing these techniques, American petrochemical firms also contributed to the establishment of chemical engineering in universities. According to Rosenberg and Landau (1992) "the development of the discipline of chemical engineering was associated, to a striking degree, with a single institution: MIT". The rise of chemical engineering at MIT centered around William Walker's Research Laboratory of Applied Chemistry (RELAC), founded in 1908. Walker supported the formation of the department of chemical engineering at MIT in 1920 and co-wrote the first American textbook on the subject in 1924. However, these developments would probably not have taken place without industry involvement. For one, Walker himself was a former industrial scientist and had worked with Arthur D. Little, who later formed his own research firm in 1909. The RELAC was established with the purpose of soliciting research contracts from industry (Geiger, 1986). Many of the best ideas and data originated from industry. The critical concept of reducing chemical engineering into "unit operations", for instance, was theorized by Arthur Little (Hounshell & Smith, 1988, Ch.14).

[Insert Figure 1 Here]

Finally, the research university system in America was still catching up to the European gold standard at the time. Figure 1 contrasts the budgets (from Geiger (1986, pp.273-4)) and scientific output (papers) of American research universities before and after the first World War and shows that there was a sustained increase in both. Conversely, the number of publications by the nation's top 16 universities at the end of World War I was less than a quarter of its level at the eve of World War II. In 1919, the budgets of these universities were less than half of their level in 1937. Five out of the 42 Nobel Laureates in physiology (12%), three out of the 39 in chemistry (8%) and six out of the 46 in physics (13%) awarded between 1901 (the first awards) and 1939 went to American scientists. This is in stark contrast to the post-1940 period (1940-2020), where the American share jumps to 55% for physiology, 68% for chemistry, and 51% for physics. The gap was wider in select fields of chemistry and physics: Only two Americans (Irving Langmuir and Karl Compton) were invited to the famous 1927 (5th) Solvay Conference on Electrons on Photons. This is indicative of the types of transatlantic intellectual interactions that occurred before the 1930s: American physicists would receive postdoctoral training in Germany, while German physicists would come on lecture tours in the United States (Fleming & Bailyn, 1968; Holton, 1981).

U.S. universities before the end of the Second World War relied heavily on private funding, a significant portion of which came from contract research. Corporate managers in the electrical and chemical firms (the "high-tech" industries of the time) may therefore have found basic science from universities wanting. A comparison of the research expenditures between the best corporate laboratories and universities of the era underlines this point. In its "Research: A National Resource" report published in 1938, the National Resources Planning Board under the NRC surveyed 1,450 American colleges and universities and found that the top 150 spent a total of $50 million on research (by contrast, the bottom 1,300 were estimated to spend a total of $1 million) between 1935 and 1936 (Council, 1938). For the top 150 schools, this boils down to around $333,333 per school (in historical dollars). The University of Chicago ($2,557,803 in 1929-30), and the University of California ($2,350,000 in 1928-29) were the top research spenders.

If we compare these figures to the top corporate laboratories, it would not seem surprising if universities struggled to retain their brightest faculty: according to Hounshell and Smith (1988, p.612), DuPont's 1925, 1930, and 1935 budgets were $1.99, $5.5, and $6.6 million. This implies that DuPont's R&D budget in 1930 was as large as that of the University Chicago and California put together. AT&T's R&D data from Maclaurin and Harman (1949, p.158) shows that the 1925, 1930 and 1935 budgets were $11,666,281, $23,241,590, $15,372,678 respectively. While accounting standards may differ between corporate and university settings, these comparisons shed some light on the magnitudes of corporate and university research programs at the time.[8]

To be clear, there were areas of American scientific and technological excellence. Agricultural sciences and mechanical and civil engineering, for instance, were nurtured since the Morril Act of 1862 that established the land grant colleges. In select sectors such as machine making, American firms were technologically advanced to the extent that German competitors were playing catch-up by reverse engineering imported American machines (Richter & Streb, 2011). In addition, the seeds of excellence for American universities may have been sown before the onset of World War II, due to various institutional circumstances relating to the incentive structure for faculty hiring and promotion (MacLeod & Urquiola, 2020; Urquiola, 2020).

The rest of the paper explores these issues by using large scale data on corporate science. To test

[8]Irving Langmuir, for instance, was disenchanted with the lack of research support at Stevens Institute of Technology, where he joined as a faculty member in 1906. "To his chagrin, he found few students with an interest in science. His attempts to upgrade laboratory facilities and the quality of student work met with hostility from students and indifference from his colleagues. To make matters worse, he had little time for research. When he left in 1909 to join the GE Research Laboratory, Langmuir found a position that met his needs far better" (Reich, 1985, p.111).

the idea that firms invested in scientific areas where domestic academic science was lagging, we generate measures of comparative advantage among scientific disciplines between the United States and Europe. We then test whether American corporate involvement in science was stronger in fields where American universities were comparatively weaker. We also test whether such involvement was more likely for firms at the technological frontier by examining the patent characteristics of "science-active" firms. Finally, we ask whether such strategies helped firms generate new and useful products that contributed to the value of the company.

3 Data

Our unbalanced panel of firms is constructed by matching several data sets: the corporate ownership and financial statements data set assembled by KKMY, augmented by market value data on other listed companies from CRSP; the USPTO data from Google Patents; the Microsoft Academic Graph (MAG) data set, which provides information on scientific publications. The combined data set covers the period 1926-1940.

We begin with 234 firms from KKMY that patent at least once within our sample period in an IPC that cites at least five scientific articles between 1947 and 1957. This restricts our sample to firms that are "at risk" of beginning scientific research.[9] We select patenting firms to focus our analysis on firms which engage in innovation. Since KKMY comprises a subset of large, industrial firms active during the 1920s, this sample may omit small firms or firms that became active in the 1930s. We address this coverage issue by including 235 listed firms from CRSP that patent (Kogan, Papanikolaou, Seru, & Stoffman, 2017). Therefore, our basic sample consists of 469 private and public American firms (7,035 firm-years) that patent at least once in our sample period. Of these, there are 358 firms (and 5,370 firm-years) for which we have financial statement (assets and earnings) and ownership data between 1926 and 1930. We look at the first five years of our sample period for firm size data to mitigate the concern that firm size is endogenous to the R&D activities of our firms.

3.1 Corporate Ownership and Financial Statements

We collect corporate ownership data in order to link firms under common ultimate ownership and construct a measure of the diversification of the entity they belong to. The sample of firms for which

[9] Examples of excluded patent classes include B27M (woodworking), B60P (loading transportation vehicles), and E03D (Water Closets or Flushing Valves thereof). Around 26% of patenting firms in our sample are lost to this restriction.

ownership data is from KKMY, who collect data on the control of U.S. non-financial corporations for the years of 1926, 1929, 1932, 1937, 1940 (and for 1950, after the end of our sample period). Using Moody's Manuals on non-financial sectors (Railroads, Public Utilities and Industrials Manuals, which are available from Mergent Online (http://webreports.mergent.com/), the authors start with the largest 200 non-financial corporations, ranked by total assets, as reported in Berle and Means (1932), and construct ownership trees for these firms, their parents and subsidiaries. The chains of control, from the ultimate owner to all the subsidiaries, is based on Moody's definition of control, which uses both equity links and other considerations. Ultimate owners of control chains (individuals, families or, in some cases, a widely held apex company) are identified using a variety of archival data sources.[10]

Financial statement variables — Balance sheet data on earnings and assets are not available before 1950 from conventional sources such as S&P Compustat. Therefore, we build on KKMY, who collect data on firm assets and earnings for the sample firms for the years 1926, 29, 32, 37, 40 (and 50), using Moody's Manuals.[11] We expand this data set for the intervening years from the same source. To classify the industries in which these firms operate, we use descriptions of firm "occupations" in Moody's Manuals. We then manually connect each industry name to one of the 85 3-digit industry codes in the revised 1947 SIC tables (reported by the BEA in 1958).[12] We augment the data set by collecting all available end-of-the-year stock market value data for listed firms, both firms included in the KKMY data set and other listed firms, using the CRSP Monthly Stock File for North American firms. For listed firms which appear on CRSP but not in the KKMY sample, we obtain data on their financials from Graham, Leary, and Roberts (2015), though not on their ownership.[13]

Measures of Business Group Ownership and Diversification — Following KKMY, we define a business groups as three listed firms under common ownership. For years in our sample where ownership data are not collected (i.e., years other than 1926,29,32,37,40), we impute business group affiliation years if they do not change between consecutive collection years.[14] In addition to using business group data, we measure levels of diversification by calculating Herfindahl (HHI) indices of sales distributions across 3-digit SIC industries at the ultimate owner level. For each ultimate owner-year, we calculate the share of each 3-digit industry out of total sales and sum the squared shares across industries. A Herfindahl index of 1 implies that the group of firms owned by the ultimate owner derives all of its sales from a

[10]See appendix A.2 for details on the construction of ownership chains and the identification of ultimate owners.

[11]Figure 6 in Appendix A.2 reproduces the 1949 entry for the Porto Rico Telephone Company.

[12]Source: http://www.bea.gov/industry/io_histsic.htm

[13]This dataset is also manually collected from Moody's Industrial Manuals.

[14]That is, if a firm is owned by General Electric in 1926 and 1929, years 1927 and 1928 are imputed for the firm as GE affiliate years.

single industry.[15]

3.2 Patents

Our patent data are derived from Google's public patent data set. This dataset digitizes a majority of patent characteristics hitherto unavailable for patents granted before the 1980s. There are 637,190 patents granted between 1926 and 1940 by the USPTO. We collect information on grant date, assignee and inventor names, Cooperative Patent Classification (CPC) codes, as well prior art citations made to patents. Based on this, we calculate normalized forward patent citations by dividing the total number of prior art citations received by a focal patent by the average number of citations received by the patents granted in the focal patents' issue year.[16] We also measure the "novelty" of a patent by counting the number of times the same CPC combination of a patent has been granted since 1790 (Fleming, 2001).[17] A Combination Familiarity score of zero implies that the technical combination embedded in the patent is brand new (has never appeared before). For instance, Wallace Carother's Nylon patent for DuPont (US2130948A) combines eight different CPC subclasses, some for polyamides and others for fibers, which was an unprecedented combination at its time (hence, the resulting familiarity score is zero).[18]

Patent assignees are matched to firms using a fuzzy string matching algorithm. This builds on Kogan, Papanikolaou, Seru, and Stoffman (2017), who match listed firms in CRSP from 1926 to patent assignee names. Our string matching algorithm takes into account abbreviations frequent in the era (e.g., firms in the railroad sector may be abbreviated as RR (railroad), RW (railway), RC (rail company)), and name variants for certain companies (e.g. AT&T's Bell Labs, SOCONY for the Standard Oil Company of New York). We match the assignee names of 89,328 patents to the 468 firms in our panel between 1926 and 1940. Of these, 233 firms are found in the KKMY sample, 349 are found in CRSP, and 114 firms are found in both.

[15] It is important to note that we calculate HHI for any multi-firm entity regardless of the number of listed affiliates. Therefore, the calculation of HHI is not restricted to business groups only.

[16] This cohort-based normalization is important because a procedural change at the USPTO starting from 1947 substantially increased citations afterwards. Using 5- or 10-year lagged forward citations for our sample of patents would have introduced biases because of this reform.

[17] "Combination Familiarity" of patent i, R_i is $k \times exp(\frac{publication\ date\ as\ patent\ i - publication\ date\ as\ patent\ k}{time\ constant\ of\ knowledge\ loss})$ where k = number of patents granted before patent i and *time constant of knowledge loss* is set to 5 years.

[18] These are D01F6/60,D01F6/58, D01F6/605, C08G69/26, C08G69/28, D01D5/06, Y10S8/21, Y10T428/2904, Y10T428/2976, and Y10T428/2978

3.3 Scientific Publications

We use Microsoft Academic Graph (MAG) to source 283,992 peer-reviewed scientific publications between 1926 and 1940. MAG coverage of publication data from the early twentieth century is superior to other data sources that provide bibliographic information for scientific articles such as Clarivate Web of Science. We exclude papers in the humanities and the social sciences based on their OECD subfields.[19] Similar to patents, we calculate normalized forward publication citations by dividing raw forward citations received by publications up until 2019 by the average number of forward citations received by the focal cohort (papers published in the same year). Using Marx and Fuegi (2020), we also identify which publications are cited by a U.S. patent in the future.[20]

To compare the rise of American corporate science against university science, we match 140,766 author affiliations from 283,992 papers to each sector. For universities, we filter publications that contain affiliations that indicate academic authorship such as "University", "College", "Institute of Technology", "School" etc. We ensure that government laboratories and nonprofit institutes (e.g., the National Institutes of Health or the Carnegie Institutions), are not classified as university-related affiliations. This process yields 60,305 affiliations (123,657 papers) that are related to universities for 1926-1940. From this list of global university publications, we limit our sample to affiliations in the United States only. For corporations, we follow the same fuzzy string matching algorithm used for matching patents. In addition to name variants and abbreviations already identified, we ensure that eponymous charitable foundations and hospitals (e.g., by DuPont, Carnegie, and Rockefeller) are not erroneously classified as corporate publications. We match 3,194 corporate publications to 201 sample firms. Of these, 110 are found in KKMY sample, 162 firms are found in CRSP, and 71 are found in both samples.

3.4 Corporate Labs

We also obtain data on the size of R&D labs operated by firms in our sample. We collect information from a national survey by the National Research Council (NRC) conducted since 1920. The "Industrial Research Laboratories of the United States, including Consulting Research Laboratories" is a comprehensive survey whose respondents consisted of "directors of [corporate] laboratories", who were contacted through a "thorough search of periodicals and membership lists, and through a very good response to

[19]These fields of science (FOS) have been defined by 2002 revision (6th edition) of the Frascati Manual, which is a manual on the collection of R&D statistics that national experts from OECD member countries agreed to in 1963. See https://www.oecd.org/science/inno/38235147.pdf for full list of classifications.

[20]We use version 17 of this dataset, available from https://zenodo.org/record/3236625#.XlS0KS2ZNTY

a notice of the forthcoming [survey] in the technical journals." (Service, 1931, p.5). Data from these surveys have been used in Mowery and Rosenberg (1999), Nicholas (2011), Field (2003), and Furman and MacGarvie (2007). We manually match our firm to firms in the 1927 (999 firms), 1931 (1,620 firms), 1933 (1,562 firms) and 1938 (1,769 firms) surveys. We collect the number of total personnel that were employed in labs, as well as details on the type of research conducted at the matched firms.

3.5 Gap in university science

We measure the "void" or "gap" in university science in America compared to Europe by i) the number of scientists trained at or affiliated with a European university and ii) the forward citation-weighted scientific publications authored by scientists in each region.

American Men of Science — We assume that scientific areas where the United States is ahead in will exhibit more homegrown scientific talent, while areas where the U.S. lags behind Europe will feature more European-educated scientists. For instance, the founder of the American Journal of Chemistry (Ira Remsen) studied at Götingen, while the alma mater of the founder of the American Journal of Mathematics (James Sylvester) was the University of Cambridge (Kevles, 1979). On the other hand, areas such as agriculture and civil engineering, where the United States did not lag as far behind, did not require a similar import of overseas talent. We collect information on European education/affiliation by American scientists from the 1921 (3rd) edition of the Cattel Directory of American Men of Science.

Published by James McKeen Cattel since 1906 and running its 38th edition in 2020, the American Men of Science directory (hereafter AMS directory) is one of the oldest and most comprehensive listing of scientists active in the United States (Moser & San, 2020). James Cattel was the inaugural editor of *Science* and relied on his correspondence with scientific societies, universities, and corporate laboratories to compile his list of scientists in each discipline. The directory lists the full name, title, current employment, field of study, birthplace and birthdate, education, employment. as well as professional society affiliations. To measure relative scientific strength before our sample period (which starts from 1926), we focus on the first three editions (1906, 1910, 1921), which are freely available from Hathitrust.[21] We choose the 1921 edition because the Optical Character Recognition (OCR) quality is highest and it provides the most comprehensive listing of scientists.[22] From the raw PDF files downloaded from Hathitrust, we extract 8,232 author entries. Upon further cleaning, we obtain 7,245 scientists that have "non-missing" subjects

[21] https://catalog.hathitrust.org/Record/003255132 for details.

[22] The number of listed scientists increases from around 4,000 in 1906 to 5,500 in 1910 and 9,500 in 1921. And Cattel, in his 1921 preface, concedes that this increase "may to some extent be due to a more complete collection of the material".

(24% information loss in total), of which 1,649 have training or experience in European institutions. We take care to extract only the most representative discipline for each scientist.[23] The number of scientists for each region are broken down by field in the left subcolumn of table 1. Although we apply extensive manual checks, it is possible that scientists exclusively trained in the United States up to the doctoral level are recruited by European institutions and show up as "European" due to their affiliations. While random checks suggest that this is quite rare, this motivates our alternative measure of national scientific "strength", which is based on scientific publication output.

Web of Science — Since Microsoft Academic Graph does not contain an address field,[24] we rely on Clarivate's Web of Science (WoS), which provides country data wherever addresses are recorded. We collect the address information on 307,847 publications published between 1900 and 1920 from WoS and classify all author addresses into US, Europe, and "Rest of World" regions based on their country names. Because 85% of addresses are missing in WoS for this period, we leverage the *American Men of Science* directory to identify American scientists by author name. We then determine whether the rest of the names are European using a naive-Bayes nationality classifier from Ye, Han, Hu, Coskun, Liu, Qin, and Skiena (2017).[25] This process yields 162,442 publications by Europeans and 65,316 publications by Americans between 1900 and 1920. The gap at the scientific field level is then defined for 39 OECD subfields after a crosswalk from WoS field (from Marx and Fuegi, 2020). After excluding the social sciences and humanities, we are left with 15 sub-fields for which at least one "European" or "American" published between 1900 and 1920. To adjust for quality differences, we next weigh these publication counts by the number of forward paper citations received until 2020. These numbers are broken down by field in the right sub-column of table 1.

Scientific field gap crosswalk to firms — Since our regional scientific activity data (from AMS and WoS) are encoded at the scientific (OECD) sub-field level, we crosswalk them to the firm level based on how much each firm patents in a patent class, and on how much a patent class relies on a scientific field. To illustrate for the WoS-based measure, we first define the number of European papers published between 1900 and 1920 relevant for each 4-digit IPC based on the share of patents in that IPC that cite

[23]See Appendix B.1 for details on cleaning the AMS data

[24]Affiliations, which are available in MAG, rarely contain full addresses, and by implication, country names during our desired period between 1900 and 1920.

[25]Offered through the "Nameprism API": https://name-prism.com/about. For more detail on which nationalities are classified as European, please see appendix B.2.

the OECD sub-field in their front page Non-Patent Literature section:[26]

$$European\ Papers_{IPC,subfield} := \frac{NPL\ Citations_{IPC,subfield}}{NPL\ Citations_{IPC}} \times European\ Papers_{subfield} \quad (1)$$

We sum $European\ Papers_{IPC,subfield}$ over all OECD sub-fields for each IPC to obtain the total number of papers "relevant" to it: $European\ Papers_{IPC}$. We then crosswalk this IPC level value to a firm using the share of firm patents in each IPC:[27]

$$European\ Papers_{firm,IPC} := \frac{Patents_{firm,IPC}}{Patents_{firm}} \times European\ Papers_{IPC} \quad (2)$$

Summing $European\ Papers_{firm,IPC}$ over the 4-digit IPCs for each firm, we obtain the number of European publications relevant to each firm. We repeat the same procedure for American publications published between 1900 and 1920 in WoS. We then divide the crosswalked number of European publications by the crosswalked American publications at the firm level, which we define as our historical scientific gap measure for WoS.[28]

To calculate the gap measure based on the number of European-affiliated vs non-affiliated scientists in the 1921 edition of the American Men of Science, we replace $European\ Papers_{subfield}$ in equation 1 with the number of scientists in that OECD sub-field which have been trained, at least in part, in Europe (while $American\ Papers_{subfield}$ are replaced with the scientists without European experience). After crosswalking the number of European-affiliated scientists and non-affiliated scientists relevant to a firm with the same procedure as above, we divide the former by the latter to generate an additional AMS-based historical (1921) scientific gap measure.

3.5.1 Comparison between gap measures

[Insert Table 1 Here]

[26]We use data for patents granted for the first ten year period since NPL citations were formalized in U.S. patent documents (between 1947 and 1957).

[27]The shares are calculated for firm patents granted during our entire sample period between 1926 and 1940.

[28]For example, AT&T, which is in the 90th percentile in this score, patents most often in the IPC: H01J (ELECTRIC DISCHARGE TUBES OR DISCHARGE LAMPS), where 15% of its patents between 1926 and 1940 are classified under. Patents in this IPC, in turn, cite the Chemical Sciences most often (26%), followed by Electrical Engineering (23%) and Physical Sciences (21%) between 1947 and 1957. As we see in table 1, Chemical Sciences and Physical Sciences have European-to-American ratios that are higher than the average, which contributes to the higher firm-level gap score AT&T receives. In contrast, General Ice Cream Corp, which is below the 10th percentile in this score, patents most often in A23G (COCOA; COCOA PRODUCTS) during our sample period, where the highest number of NPL citations between 1947 and 1957 are made to Biological Sciences, among the fields in table 1. Biological sciences, in turn, has a European-to-American ratio below the average, which contributes to the firm receiving a low gap score.

Table 1 compares the two measures of scientific "strength" (relative backwardness) from WoS and AMS. The "Ratio" columns for each measure presents the number of European-affiliated scientists (for the AMS) and European-authored papers (for the WoS) divided by the number of scientists never affiliated with a European institution and the number of American-authored papers, respectively. Intuitively, this can be thought of as the "gap" or "lag" that exists between European and American institutions (fields with relatively large values are those where the scientific gap between Europe and the U.S. is large). The two measures do not yield identical results. For instance, the AMS ratio is smaller in mean and variance compared to the WoS one. This may partly be due to the name-based nationality imputations we conduct for authors without affiliations.[29] Moreover, AMS covers only scientists residing in America, whereas WoS covers papers published by scientists working outside America as well. Notwithstanding this, when we compare the two measures by their rankings, ten out of the twelve fields where both measures have non-missing values differ by no more than three ranks (e.g. Mathematics is ranked 2nd in the AMS-based ratios, and is also ranked 2nd in the WoS-based ranking). The two outliers are basic medical research, and agriculture, forestry and fisheries. For medicine, we suspect that there may be an over-representation of practitioners (i.e., practicing physicians). In agriculture, forestry, and fisheries, there may be a measurement error given that it encompasses a wide variety of fields. At the firm level, i.e., when the observations are weighted by the industries or scientific sub-fields of firms in the sample, the correlation between the two measures (r=0.527) is slightly higher than the correlation at the scientific field level (r=0.508), suggesting that fields with the highest mismatches between AMS and WoS are not very important in the patent classes used by our sample firms (figure 2).

[Insert Figure 2 Here]

3.6 Descriptive Statistics

[Insert Table 2 Here]

Table 2 presents descriptive statistics at the firm-year level. The most number of observations is 7,035 (469 firms observed for 15 years). "Lab Size" counts the number of lab personnel reported in the IRL directory. There are only around a third of the total observations here because the IRL was collected for only five years by the NRC (1927,31,33,37,40). As observed from the difference in median and mean values (and as we also show in table 4), scientific publications and lab personnel are skewed to the right.

[29] In unreported robustness checks, we replicate our results without imputing author nationality.

The average gross income and assets are 879 million and 1.4 billion respectively. These are slightly larger than the values for the "pre-period" between 1926 and 1930, reflecting the growth of the U.S. economy. The median level of concentration (measured by HHIs) among firms in our sample is 1, while the average is 0.82. This implies that, among firms linked in ownership chains, those that are diversified in distinct industries are a minority.

[Insert Figure 3 Here]

Figure 3 presents trends in corporate investment in science during the interwar period. Average corporate publications exhibit an upward trend, reflecting an aggregate increase from just under 200 corporate-authored papers in 1926 to around 700 papers in 1940. In addition, about 12.5% of all firms published in 1926, increasing to to 17.6% of all firms in 1940. The staff employed in corporate laboratories mirrors the same upward trend. These figures reflect an expansion of corporate science both on the extensive margin (measured by total publications) and intensive margin (measured by share of firms that publish). This rise in corporate science, however, is not monotonous, peaking in the early 1930s. The increase in patenting is less steep but more monotonic, with an aggregate rise from around 3,000 to 6,500 corporate patents per year.

[Insert Figure 4, Table 3 Here]

Figure 4 shows that these trends mask substantial heterogeneity. For instance, 158 firms out of our sample of 469 firms never operate an R&D lab, while more than half (268) of the firms never publish a scientific article. Perazich and Field (1940) estimate that less than 1% of all firms accounted for a third of all industrial research employment, for 1921, 1927 and 1938 respectively (see table 3 for details).[30]

[Insert Figure 5 Here]

Figure 5 presents the correlation between the scientific gap and corporate science across industries. Corporate investments in science seem to occur in industries where the U.S. lags behind European science: For instance, construction, which relies on civil engineering where the gap is small, exhibits less corporate science investment than communications, which relies partly on chemistry, where the gap is large. This pattern is consistent with our conjecture in section 2; it also mandates the use of industry fixed effects throughout the empirical analysis.

[30]Moreover, a mere 45 firms in 1938 employed half of total research personnel.

4 Econometric Framework

4.1 Corporate science and gaps in university science (OLS)

We estimate the following specification for firm i in year t:

$$Corpsci_{it} = \beta_0 + \beta_1 Gap_i + \boldsymbol{Z_i'}\boldsymbol{\gamma} + \tau_t + \phi_c + \epsilon_{it} \tag{3}$$

where $Corpsci_{it}$ is defined as the investment in corporate science by our sample firms, measured by number of lab employees or the scientific publication stock; Gap_i is defined as the ratio of European science to American science relevant to each firm; $\boldsymbol{Z_i'}$ is a vector of controls. We include level values for each science measure since our gap measure is a ratio. We also control for the size of the firm (assets) and include year (τ_t) and 2-digit industry (ϕ_c) fixed effects. We expect $\hat{\beta}_1 > 0$.

4.2 Interaction with firm characteristics (OLS)

$$Ccorpsci_{it} = \beta_0 + \beta_1 Gap_i \ + \beta_2 Gap_i \times TechFrontier_i + \beta_3 TechFrontier_i + \boldsymbol{Z_i'}\boldsymbol{\gamma} + \tau_t + \phi_c + \epsilon_{it} \tag{4}$$

We hypothesized in section 2 that firms needed to internalize research as they neared the technological frontier. We test this hypothesis by interacting scientific gaps with how technologically advanced a firm is. We measure this by the average forward citations the firm's patents receive. We expect $\hat{\beta}_2 > 0$. In addition, we also test whether firm investment in science is related to other firm attributes. Prior research argues that larger firms have the ability to conduct basic research (Schumpeter, 1939) and that both large and diversified firms have the incentive to do so (Arrow, 1962; Nelson, 1959b).[31] We therefore replace $TechFrontier$ with dummies for whether the firm's size (assets) are above the sample median, whether the firm is part of a multi-firm business group, and whether the firm is controlled by an ultimate owner whose holdings are more diversified than the sample median, which is based on the owner's HHI for sales across industries.

[31] Assuming imperfect markets for technologies (where knowledge can otherwise be traded), more diversified firms are better able to internalize the fruits of basic research, which is more uncertain in terms of its eventual application and therefore wider in potential scope.

4.3 Performance consequences of corporate science

4.3.1 "Home-run" patent equation (IVE)

We estimate the "technological returns" to corporate science, as measured by the production of high quality patents. Specifically, we ask whether firms that responded to the gap by investing in internal research produced superior inventions in terms of novelty, stock market value, and forward patent citations. As a benchmark, we first estimate non-causal correlations using the following OLS specification, which includes firm publications as a proxy for corporate science and controls for the current (contemporaneous) WoS gap measure:

$$HomeRun_{it} = \beta_1 Corpsci_{it} + \boldsymbol{Z_i'}\boldsymbol{\gamma} + \tau_t + \phi_c + \epsilon_{it} \tag{5}$$

$HomeRun_{it}$ is measured using the number of patents that are in the top 5% of stock market value (Kogan, Papanikolaou, Seru, & Stoffman, 2017) and novelty scores (Fleming, 2001).

Given that corporate science is endogenous, we obtain causal estimates by instrumenting for it using the historical scientific gap. We estimate through two stage least squares, where in the first stage, we regress $Corpsci_{it}$ on the pre-sample university science gap measure (from equation 3) and other controls, and in the second stage we regress $HomeRun_{it}$ on the fitted values of investment in corporate science ($\widehat{Corpsci}_{it}$) obtained in the first stage.

4.3.2 Tobin's Q equation

As an alternative to estimating "technological" (patent-based) returns, we also estimate the contribution of corporate research to firm value:

$$ln(Q)_{it} = \beta_1 ln(Pubstock_{it-1}) + \beta_2 ln(Patstock_{it-1}) + \boldsymbol{Z_i'}\boldsymbol{\gamma} + \tau_t + \phi_c + \epsilon_{it} \tag{6}$$

where Q is the market-to-book ratio (or Tobin's Q). $ln(Pubstock_{it-1})$ and $ln(Patstock_{it-1})$ refer to the natural log of one plus lagged publication and patent stock respectively. Because publication and patent stocks are strongly serially correlated, lagged values of these variables cannot alleviate endogeneity concerns. Again, we obtain causal estimates of β_1 by instrumenting for $ln(Pubstock_{it-1})$ using the historical university science gap measure.

5 Results

[Insert Table 4 Here]

Table 4 presents univariate comparisons between firms with large (above median) and small (below median) WoS-based scientific gaps. Evidently, high-gap firms are more likely to invest in corporate research. High-gap firms are also larger in terms of total assets than lower-gap companies and we therefore control for firm size in all regression specifications.

5.1 Corporate science and gaps in university science (OLS)

[Insert Table 5 Here]

Table 5 presents results from OLS estimations of equation 3. As expected, we find that gaps in university science, measured either on the basis of academic publications (WoS) or on the basis of scientist training (AMS), are positively correlated with both the corporate publication stock and the size of the corporate lab personnel. The coefficients, except that in column 8, are positive and statistically significant at the 5% level. In column 8, the coefficient of interest, *Gap in university science, 1921 (AMS)* is positive, but the t-statistic is below conventional significance levels (1.36). To get a sense of the magnitude of the impact of scientific backwardness on corporate science, we calculate the effect of a one standard deviation increase in either gap measure on the publication stock.[32] In both cases, one standard deviation increase in the gap measure is associated with an increase of about 1.8 (1.84 for WoS measure, 1.80 for AMS measure) in the publication stock, a 66% increase relative to the sample mean of 2.7. A similar increase in the gap is associated with 5.71 (WoS measure) to 6.77 (AMS measure) larger lab personnel, around 13% and 15% of the sample average. The subsequent analyses primarily focus on the number of corporate publications as the measure of corporate investment in science, since the IRL directory data on labs is only available for five years (1927,31,33,38,40) out of the sample period.

5.2 Interaction with firm characteristics (OLS)

The results thus far suggest that firms in sectors where American scientific knowledge lagged behind Europe were more likely to invest in corporate research. We now turn to examine which firm characteristics affect investment in corporate science. We focus on proximity to the technological frontier, size,

[32]In general, the WoS standard deviations are an order of magnitude larger than the standard deviations of the AMS measure. Because publications in WoS are a rarer event than the presence of scientists in AMS, we include the number of publications (for columns 1-4) and scientists (for columns 5-8) as controls.

and business group affiliation (or diversification). To proxy for proximity to the technological frontier we use information on the quality of patents. Table 6 column 1 shows that firms whose patents receive more forward citations were more likely to respond to a gap in university science. Specifically, a firm with patents in the 75th percentile of forward citations respond around 15% more in terms of new scientific publications compared to firms whose patent citations are in the 25th percentile. In column 2, we similarly find that firms with higher average forward patent citations are more likely to respond to gaps.

In table 7, we replicate these results by using the AMS measure of the gap. The directions of the interaction term coefficient for forward patent citations is identical, though we fail the replicate statistical significance. These results are consistent with our conjecture in section 2 that, on average, technologically more advanced firms experienced a more acute need for internal research, when faced with weaker domestic universities.

[Insert Table 6 Here]

We also test for the effect of size. Column 3 of table 6 shows that firms with assets above the sample median responded to a gap in university science around seven times more than those with below-median assets, suggesting that larger firms were more likely to internalize the effects of corporate research. Similarly, firms affiliated with business groups published about four times more in response to gaps in university science (column 4). We also find that firms controlled by diversified ultimate owners[33] are more likely to respond to a university science gap, albeit this effect is not with statistical significance at the 5% level (column 4). We also replicate these results using the AMS-based gap measure in table 7. We conclude that corporate size and organization and, in particular, affiliation with business groups and other multi-firm entities, affect the likelihood of engaging in research, presumably because of the ability of such corporate structures to internalize the benefits generated by the creation of basic knowledge.

5.3 Performance consequences of corporate science

Nelson (1959a, p. 119) noted that "Research laboratories may be created and maintained by firms for many purposes, including (...) quality control, (...) improvement of manufacturing methods, improvement of existing products and development of new uses for them, development of new products and processes, and *scientific research to acquire knowledge enabling more effective work to be done to achieve the above purposes*" (emphases added). If so, firms investing in scientific research ought also to have better and more

[33]Inverse of diversification is measured by the Herfindahl index of sales distribution over 3-digit industries

valuable new inventions. We use two measures: patents deemed valuable by investors, whose issuance is associated with increases in the firm's stock price; and novel patents, which combine patent classes which are rarely combined before. As detailed below, investments in internal research do indeed produce more valuable inventions.

5.3.1 "Home-run" patent equation (IVE)

[Insert Table 8 Here]

OLS estimates from table 8 column 1 show that firm publication stock is positively correlated with number of highly novel patents (those that are within 5% of novelty scores measured by the count of subclass combinations (Fleming, 2001)). To obtain plausibly causal estimates of this relationship, we instrument firm publication stock with gaps in university science. The gaps in university science are calculated before the beginning of the sample period and stem from historical circumstances on both sides of the Atlantic that are exogenous to technological opportunity. The first stage regression of publication stock against gaps in university science (column 2 for WOS, column 4 for AMS gap measures) are significant, with F-stats at 30 and 24 respectively.[34] Because American universities were catching up to European standards during the interwar period, we also control for the "contemporary" or "current" gaps calculated for each firm and year.[35]

As one would expect, we find that the "current" gap measure is negatively correlated with corporate publication stock. In the second stage regressions, we find that the increase in publication stock caused by gaps formed before our sample period increases the number of highly novel patents. The estimate in column 3 of Table 8 is about 1.75 times larger than the OLS estimate in column 1. This suggests that the OLS estimates, treating the publication stock as an exogenous factor rather than as an outcome of the scientific gap, are biased downwards.

[Insert Table 9 Here]

We replicate this result for the number of patents that are in the top 5% of stock market value (table 9). We find in column 3 of table 9 that a one standard deviation larger publication stock due to larger gaps in university science (measured through WoS) leads to around 3.9 more patents in the top 5% of

[34] Recent research shows that these to be appropriate critical values when combined with the observed second stage t-stats (Lee, McCrary, Moreira, & Porter, 2020).

[35] The $European\ Papers_{subfield}$ in the context of section 3.5 are now calculated for each year, instead of summed up for paper publication years 1900 and 1920. The crosswalks are identical, with the exception that the gap measure is calculated at the firm-year level, not the firm level.

stock market value (or around 4.5 times the sample mean). A similar magnitude is observed in column 5, which instruments for the publication stock using the historical AMS gap measure.

5.3.2 Tobin's Q equation

[Insert Table 10 Here]

Finally, we estimate the market-to-book ratio equation (6). Column 1 of table 10 shows that publication stock is positively related to the market-to-book ratio of firms. A one standard deviation larger publication stock is associated with an increase of Tobin's Q by around 0.01 (2% of the sample mean). We also split the sample by gap measures to probe whether "responding firms" benefited more from science. Comparing columns 2 and 3, we find that firms whose gaps in university science are smaller than the median show a negative and statistically insignificant effect of their publication on the market-to-book ratio. In contrast, those with above-median gaps have positive and significant effects. We replicate the same results in columns 4 and 5 by splitting the sample by median AMS gaps. We find that the correlation between Q and gaps are 1.7 times larger for those above median gap scores compared to those below the median. These results suggest that investments in science are positively related to market value, and that this relationship is stronger for firms with larger scientific gaps.

We also instrument publication stock by the gap in university science to examine whether this result is causal. As expected, we find in column 6 that the first stage regression coefficient of publication stock against gaps in university science is positive and statistically significant. However, we do not find significant results for the second stage. This may partly be due to the fact that corporate R&D was a relatively new concept and early capital markets in the United States may not yet have developed the sophistication to fully appreciate the value of corporate research. Conversely, it is possible that markets were concerned with spillovers to rivals from in-house research and penalized firms for what they perceived as wasteful activity. It is also possible that the market-to-book ratio is affected by many firm-level characteristics and is imprecisely measured, making it difficult to estimate precisely the effect of corporate research on it. When we look at a more narrow definition of financial returns, we find that the stock market value of patents is positively related in the second stage to publication stock predicted by both of our scientific gap measures (columns 3 and 5 of table 9).

6 Discussion and Conclusion

We argue that the rise of corporate research in America in the interwar period can be attributed to the weakness of American academia in certain scientific fields, and to the ability of large, diversified and relatively technologically advanced corporations to fill up this institutional void by internalizing the provision of scientific knowledge, a classic public good. The rise in corporate research in America took place against the background of increasing limits to international trade and technology transfers; by the 1930s, licensed technology from competitors abroad was no longer readily available to support the newly developing U.S. industries. Corporations were required to acquire a deep mastery of the underlying science.

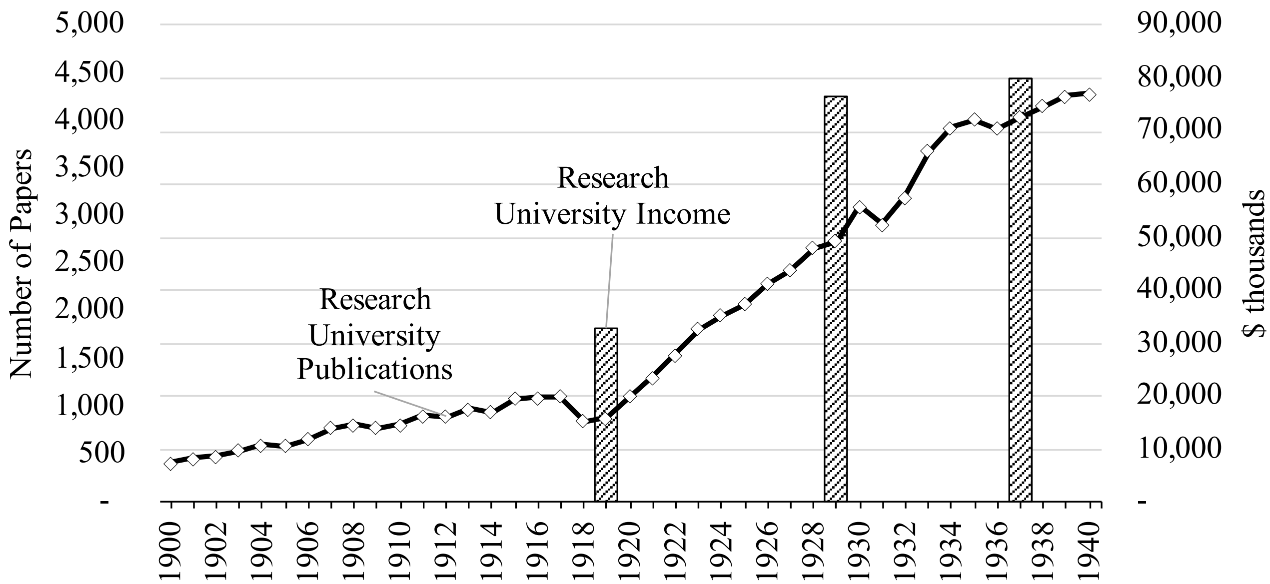
American universities have increased the quality and quantity of their scientific output since the interwar period. Eventually, this process led to the division of labor between academia and industry we observe today. Arora, Belenzon, and Patacconi (2018) show that the scientific publication stock of American public firms declined in the late twentieth century, 40 or 50 years after our sample period. Besides the growing scientific might of American universities, it is possible that corporate research in the 1930s was easier protect from rivals (few could effectively use it), whereas by the 1980s knowledge spillovers may have become increasingly costly. This conjecture is consistent with Arora, Belenzon, and Sheer (2020b) who show that companies cut back on research when spillovers to rivals increased relative to the value from internal use.

Much about the rise and apparent decline of industrial research remains poorly understood. How did investment in research by a firm affect its competitors? Did rivals benefit from some spillovers, or were they forced to innovate themselves? What was the role of "patient" shareholders in allowing management to pursue investment projects with distant and uncertain payoffs?

The historical evidence we provide is related also to contemporary policy issues and debates in both developed and in emerging economies. For example, to what extent should governments in emerging economies attempt to fill institutional voids, and to what extent can these voids be overcome by private corporations? Do private corporations obtain competitive advantage by filling up institutional voids? Is this worse, in terms of social welfare, than leaving the voids unaddressed because of the limited capabilities of many emerging market governments? In developed economies, to what extent is the historical evidence on research carried out by private entities of relevance to the present-day debate about the costs and benefits of large technologically advanced firms?

In addition to providing new evidence on corporate research in America in the interwar period, we assemble the most extensive historical sample of American firms that were involved in innovation during that period, including information on the scientific output of these firms and on the relative gap between American and European universities. We hope that these newly developed data will contribute to future research on the open questions we raised.

Figure 1: The Rise of U.S. Research Universities (1900-1940)



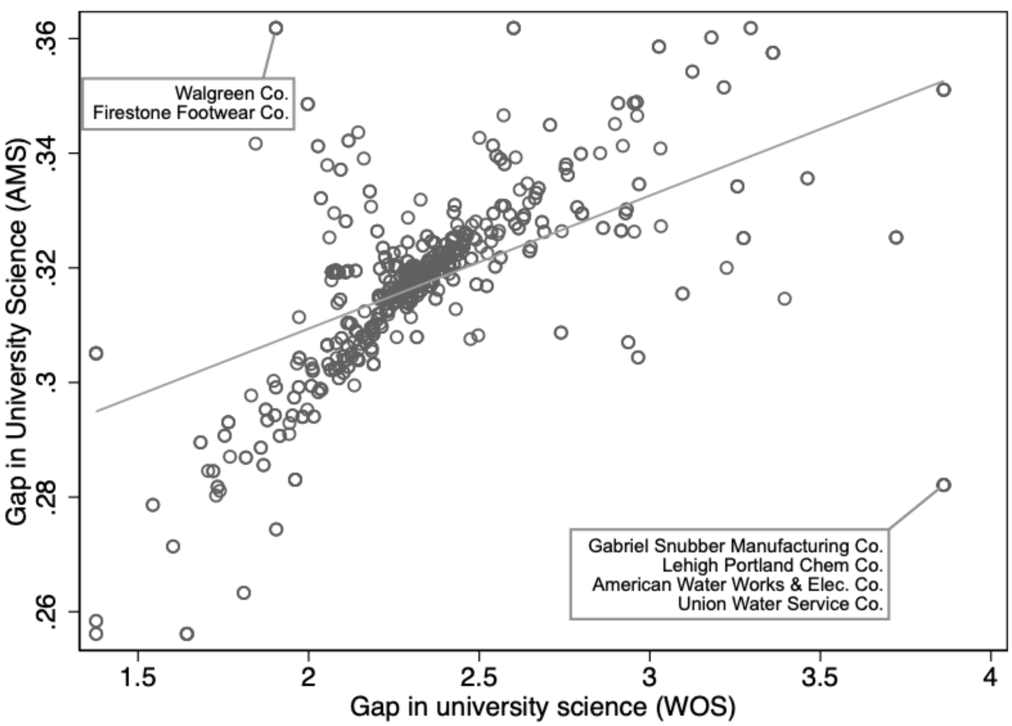
Notes: The bar graph (right axis) plots the budgets data of prominent research universities from Geiger (1986, pp.273-4). The universities represented are University of California, University of Illinois, University of Michigan, University of Minnesota, University of Wisconsin, California Institute of Technology, University of Chicago, University of Pennsylvania, Columbia University, Cornell University, Harvard University, Johns Hopkins University, Massachusetts Institute of Technology, Princeton University, Stanford University, and Yale University. The line graph (left axis) counts the total number of scientific publications published by these universities from the Science Citation Index-Expanded (SCI-Expanded) from Clarivate Web of Science.

Table 1: Number of Scientists and Papers, Europe vs America

	American Men of Science (No. of Scientists)			Web of Science (No. of Citation Weighted Papers)		
OECD Subfield Equivalent	American	European	Ratio	American	European	Ratio
2.05 Materials engineering	82	41	0.50	3,806	15,456	4.06
1.01 Mathematics	525	229	0.44	5,334	19,556	3.67
2.07 Environmental engineering	68	27	0.40	-	-	-
3.02 Clinical medicine	178	65	0.37	43,007	81,883	1.90
1.03 Physical sciences and astronomy	605	219	0.36	12,802	42,719	3.34
3.03 Health sciences	63	22	0.35	5,121	9,373	1.83
3.01 Basic medical research	928	324	0.35	32,556	34,614	1.06
1.04 Chemical sciences	1,189	383	0.32	31,330	75,596	2.41
2.02 Electrical eng, electronic eng	177	54	0.31	125	143	1.14
1.05 Earth and related environmental sciences	561	167	0.30	7,996	1,189	0.15
1.06 Biological sciences	1,482	360	0.24	39,262	44,261	1.13
2.03 Mechanical engineering	134	32	0.24	-	-	-
2.01 Civil engineering	120	19	0.16	1,010	636	0.63
4.01 Agriculture, forestry, fisheries	385	60	0.16	2,112	3,594	1.70

Notes: This table presents the number of scientists (from AMS) and citation-weighted articles (from WoS) that have non-missing subject and affiliation fields. The “Ratio” column for the AMS sub-columns divides the number of European-affiliated American scientists by those unaffiliated with Europe. The rows are downward-sorted by this value. The “Ratio” column for WoS sub-columns divides the number of European-affiliated papers (published globally) divided by American-affiliated papers.

Figure 2: Comparison of Gaps in University Science



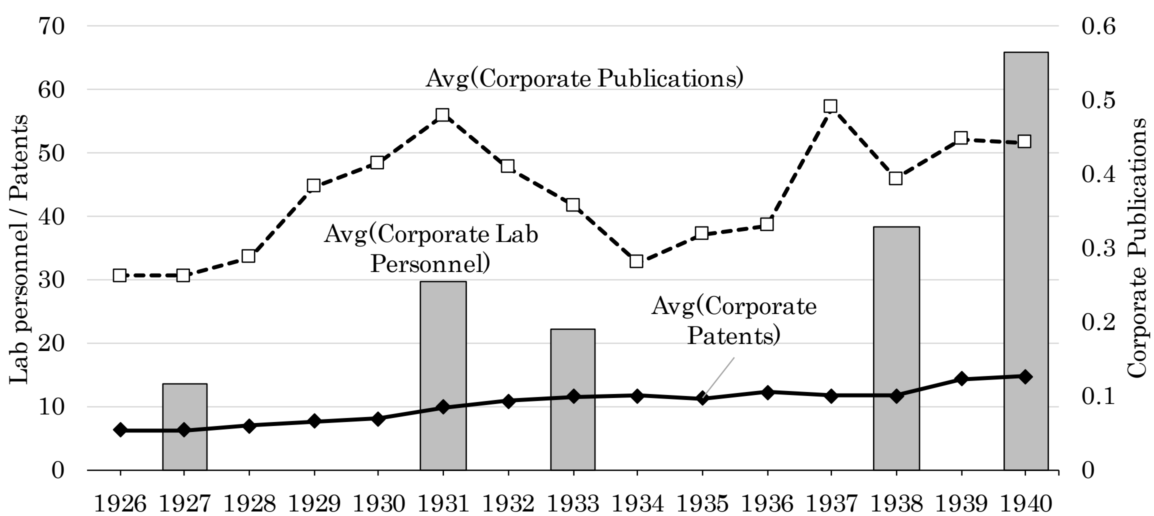
Notes: This figure compares the two scientific gap measures at the firm level. Higher values represent a larger gap between Europe and the United States. The AMS gap measure (on the vertical axis) tends to be positively correlated with the WoS gap measure (on the horizontal axis).

Table 2: Summary Statistics of Main Variables

	Observations	Mean	Median	Std. Dev.	Min	Max
Lab Size	2320	43.49	0	214.86	0	4669
Patents Granted Per Year	7035	13.13	1	54.88	0	838
Publications Authored Per Year	7035	0.46	0	3.72	0	88
Patent Stock	7035	71.18	7.54	316.85	0	4441.06
Publication Stock	7035	2.7	0	22.16	0	440.33
Forward Patent Citations	4035	0.77	0.62	0.8	0	18.42
KPSS Patent Value	2629	2.58	1.03	4.37	0.03	56.65
Gross Income ($1M)	2789	879.27	270.66	1864.59	-1.98	20655.93
Total Assets ($1M)	4305	1369.71	418.87	3272.52	7.43	60114.66
Market Capitalization ($1M)	3856	1103.23	248.91	2903.02	0.69	37352.08
Business Group Affiliated = 1	3104	0.41	0	0.49	0	1
Gross Income ($1M) (1926-1930)	728	833.95	235.17	1865.36	-1.98	14366.3
Total Assets ($1M) (1926-1930)	1330	1258.17	445.75	2723.14	12.39	36047.36
Ultimate Owner Earnings HHI (1926-1930)	431	0.82	1	0.23	0.27	1

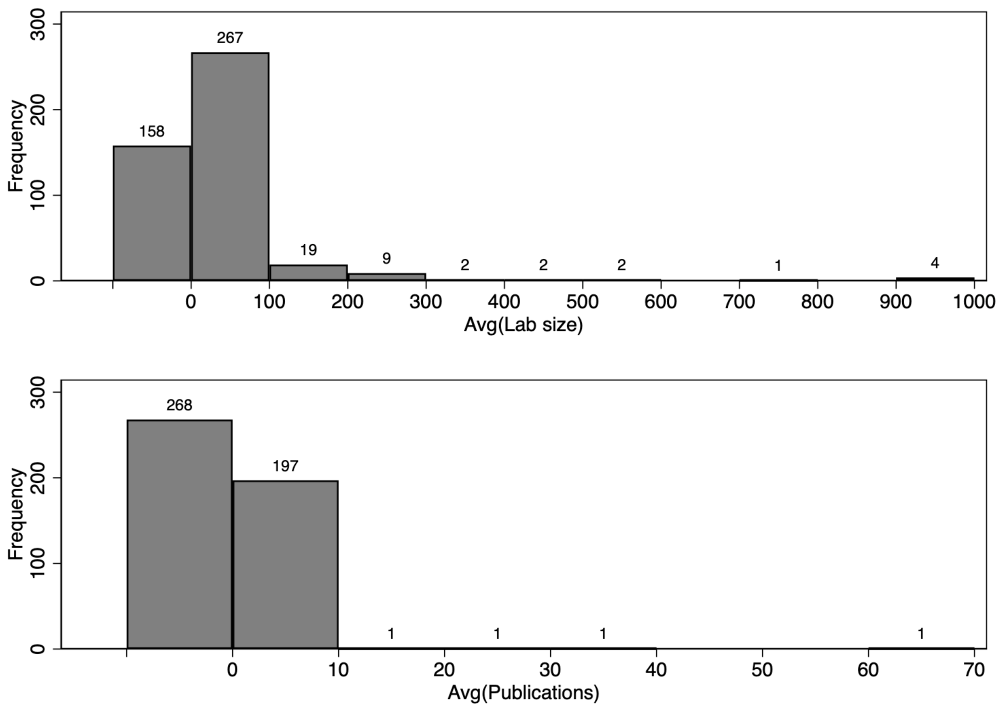
Notes: Observations are at the firm-year level and the sample period is 1926-1940. Forward Patent Citations is first defined at the patent level as the number of forward prior art citations received normalized by the average number of forward citations for the patent's grant year cohort. This value is averaged at the focal firm-year level to produce the Forward Patent Citations measure in the table. Patent and publication stock are calculated using a perpetual inventory method with a 15% rate of depreciation. KPSS Patent Value is the value of a patent (in million dollars) based on the cumulative abnormal returns in the firm's market value at the issuance event of the patent (Kogan, Papanikolaou, Seru, & Stoffman, 2017). The Ultimate Owner Earnings Herfindahl-Hirschman Index (HHI) is calculated based on the distribution of earnings across the industries where each ultimate owner is active in. All current dollar amounts are deflated to 2005 dollars using https://www.measuringworth.com/datasets/usgdp12/result.php.

Figure 3: Emergence of Corporate Science, 1926-1940



Notes: The bar graph indicates the average number of personnel employed at corporate laboratories for firms in our sample matched to the Industrial Research Laboratories Directory. The broken line indicates the average number of publications by firms in our sample matched to Microsoft Academic Graph. The solid line indicates the number of patents by firms in our sample matched to USPTO utility patents.

Figure 4: Heterogeneity of Corporate Science



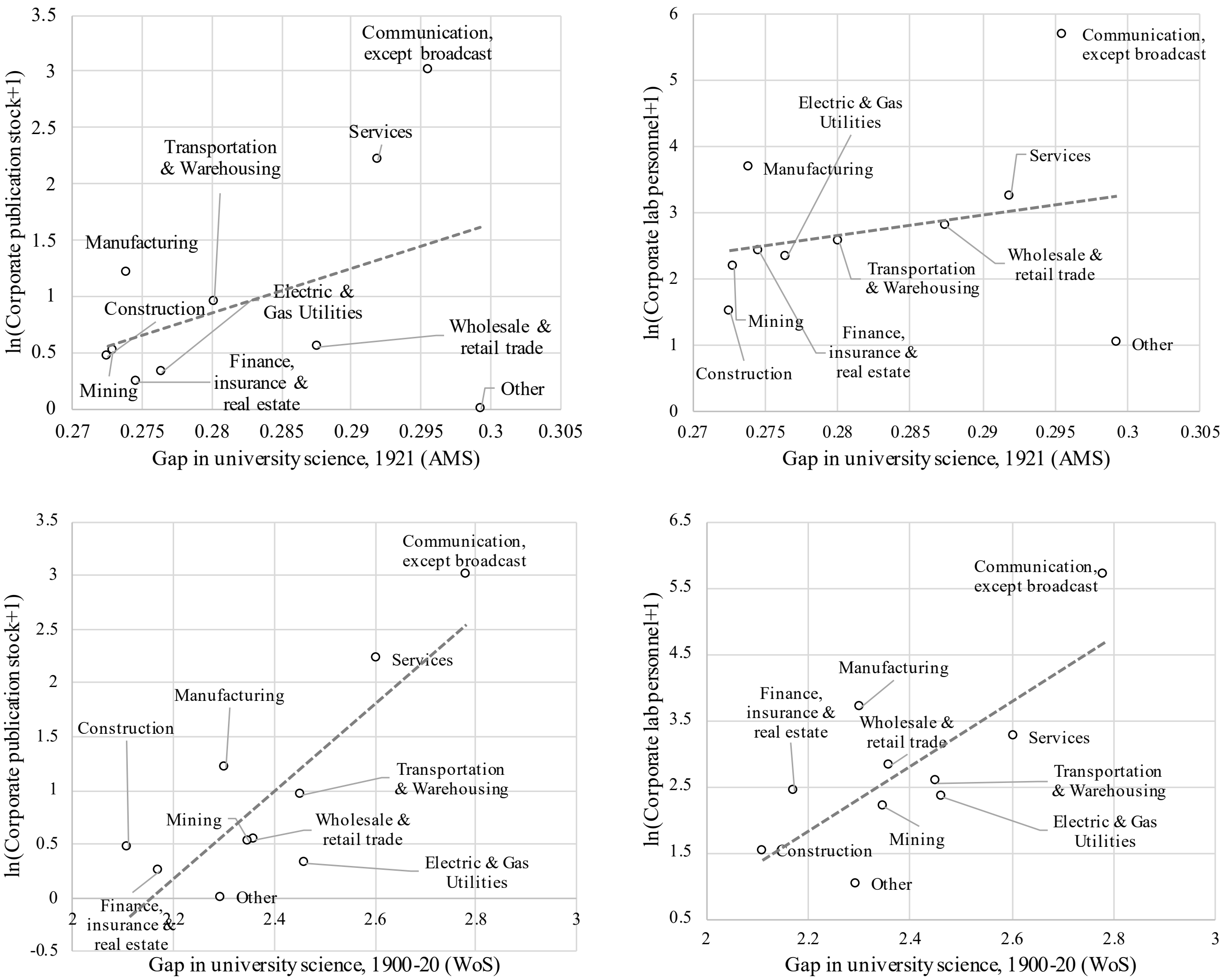
Notes: The upper histogram bins the number of personnel employed at corporate laboratories for firms in our sample. 158 firms (the leftmost bar) report no employed lab personnel in our sample period. The lower histogram bins the number of publications authored by firms in our sample. 238 firms (the leftmost bar) do not author any scientific publications in our sample period.

Table 3: Number of Companies Accounting for Each Third of Industrial-Research Employment, By Relative Size of Staff, 1921, 1927, and 1938

	1921		1927		1938	
	Companies	Employment	Companies	Employment	Companies	Employment
All Companies	462	9,350	926	18,982	1,722	44,292
Companies with:						
Largest staffs	4	3,117	4	6,327	13	14,764
Intermediate staffs	25	3,117	79	6,327	127	14,764
Smallest staffs	436	3,117	843	6,327	1,583	14,764

Notes: This is a reproduction of table 1 in Perazich and Field (1940).

Figure 5: Corporate Science vs Gaps in University Science, by Industry



Notes: Industry-level scatter plots of firm investment in science and the gaps in the relevant academic discipline. The upper left panel plots the natural log of one plus the average publication stock against the average AMS gaps in university science measures . The upper right panel replaces the publication stock with the number of personnel at R&D labs, from the IRL directory. The bottom panels are similar except that the scientific gaps are based on WoS.

Table 4: Summary Statistics of Main Variables, by Gaps

	T-Test: Large-Small Gap (WoS)		Large Gap (WoS)			Small Gap (WoS)		
	Diff.	Std. Error	Count	Mean	sd	Count	Mean	sd
Corporate Lab Personnel	23.133	17.436	233	55.009	243.552	231	31.876	104.907
Publications Authored Per Year	0.569*	0.326	234	0.749	4.944	235	0.180	0.683
Patents Granted Per Year	7.691	4.737	234	16.980	68.194	235	9.289	24.915
Total Assets ($1M) (1926-1930)	588.371**	269.349	185	1405.617	3378.142	173	817.246	1102.805
	T-Test: Large-Small Gap (AMS)		Large Gap (AMS)			Small Gap (AMS)		
	Diff.	Std. Error	Count	Mean	sd	Count	Mean	sd
Corporate Lab Personnel	22.296	17.439	232	54.640	243.471	232	32.344	106.175
Publications Authored Per Year	0.553*	0.326	234	0.741	4.947	235	0.188	0.671
Patents Granted Per Year	8.411*	4.735	234	17.341	68.929	235	8.930	22.683
Total Assets ($1M) (1926-1930)	341.780	270.407	177	1294.102	3207.940	181	952.302	1696.830

Notes: Mean comparison tests run at the firm level. "Large Gap" firms are those above the median WoS-based gap measure. "Small Gap" firms are those below the median WoS-based gap measure. Gaps are defined as the number of European publications divided by American publications in the natural sciences from Clarivate Web of Science between 1900 and 1920.

Table 5: Corporate Science and Gaps in University Science

Dependent Variable	Publication stock				Lab personnel			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gap in university science (WOS), 1900-20	6.265		7.152		43.763		33.408	
	(0.889)		(1.401)		(15.532)		(13.189)	
ln(American university science), 1900-20 (WOS)	3.183		2.990		-15.638		-7.198	
	(0.840)		(1.020)		(15.032)		(14.542)	
ln(European university science), 1900-20 (WOS)	-2.805		0.092		17.960		20.645	
	(0.823)		(0.868)		(14.454)		(15.295)	
Gap in university science, 1921 (AMS)		180.454		147.763		1002.905		531.961
		(28.743)		(39.387)		(506.640)		(468.285)
ln(American university science), 1921 (AMS)		19.661		17.165		20.104		-83.027
		(4.226)		(7.293)		(81.303)		(87.981)
ln(European university science), 1921 (AMS)		-20.225		-14.419		-20.084		98.189
		(4.448)		(7.217)		(86.546)		(91.429)
ln(Assets)	3.912	3.745	4.307	4.157	56.449	55.378	67.974	67.564
	(0.549)	(0.532)	(0.601)	(0.582)	(10.396)	(10.156)	(12.742)	(12.736)
Average of Dependent Variable	3.632	3.632	3.642	3.642	56.594	56.594	56.643	56.643
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	No	No	Yes	Yes	No	No	Yes	Yes
R^2	0.050	0.047	0.149	0.148	0.136	0.135	0.225	0.224
Number of Firms	425	425	422	422	392	392	389	389
Number of Observations	4,305	4,305	4,293	4,293	1,391	1,391	1,386	1,386

Notes: Unit of analysis is at the firm-year level. *ln(American University Science), 1900-20 (WOS)* and *ln(European University Science), 1900-20 (WOS)* refer to the number of citation weighted scientific articles published in scientific fields germane to a focal firm's industry from the United States and Europe, respectively (one is added before the natural log transformation). *Gap in university science, 1900-20 (WOS)* is defined at the firm level as the number of European scientific publications relevant to a firm divided by American publications relevant to a firm. *Gap in university science, 1921 (AMS)* is defined as the number of American scientists with a European affiliation divided by those without any such experience. Both "Gap" measures are winsorized at the 1% and 99% level. Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

Table 6: Corporate Publications and Gaps in University Science (WoS), by Firm Characteristics

	DV: Publication stock			
	(1)	(2)	(3)	(4)
Gap in university science, 1900-20 (WOS)	8.553	2.417	4.723	17.912
	(2.139)	(1.072)	(4.229)	(4.678)
Gap in university science, 1900-20 (WOS) $\times$ Forward Patent Citations	2.550			
	(1.157)			
Forward Patent Citations	-4.309			
	(2.521)			
Gap in university science, 1900-20 (WOS) $\times$ Assets above median (pre 1930)		16.199		
		(2.765)		
Assets above median (pre 1930)		-43.670		
		(7.155)		
Gap in university science, 1900-20 (WOS) $\times$ Business Group Dummy (Imputed)			17.833	
			(5.779)	
Business Group Dummy (Imputed)			-43.104	
			(13.247)	
Gap in university science, 1900-20 (WOS) $\times$ HHI above median (pre 1930)				-2.988
				(5.071)
HHI above median (pre 1930)				4.575
				(11.233)
ln(American university science), 1900-20 (WOS)	3.915	3.348	6.501	10.767
	(1.698)	(1.177)	(3.246)	(3.562)
ln(European university science), 1900-20 (WOS)	-0.254	-0.058	0.234	-1.929
	(1.489)	(1.011)	(2.819)	(2.915)
ln(Assets)	5.404	5.947	6.575	7.473
	(0.773)	(0.916)	(1.082)	(1.089)
Average of Dependent Variable	4.967	3.843	6.625	7.886
Year Fixed Effects	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes
R^2	0.191	0.165	0.179	0.206
Number of Firms	396	358	199	196
Number of Observations	2,948	4,028	1,702	1,855

Notes: Unit of analysis is at the firm-year level. *Business Group Dummy (Imputed)* refers to firms that are Business Group affiliates (defined as ownership chains with at least three firms). *Assets above median (pre 1930)* equals one if firm average assets between 1926-1930 are above median. *HHI above median (pre 1930)* equals one if average Herfindahl indices calculated at the Ultimate Owner level (based on industry sales distribution between 1926-1930) are above median. The rest of the variable definitions are identical to those in table 2 and 5. Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

Table 7: Corporate Publications and Gaps in University Science (AMS), by Firm Characteristics

	DV: Publication stock			
	(1)	(2)	(3)	(4)
Gap in university science, 1921 (AMS)	203.558	18.171	137.554	392.287
	(61.042)	(30.035)	(98.290)	(118.616)
Gap in university science, 1921 (AMS) × Forward Patent Citations	16.673			
	(16.826)			
Forward Patent Citations	-3.827			
	(5.143)			
Gap in university science, 1921 (AMS) × Assets above median (pre 1930)		293.919		
		(52.198)		
Assets above median (pre 1930)		-98.450		
		(17.173)		
Gap in university science, 1921 (AMS) × Business Group Dummy (Imputed)			350.856	
			(139.300)	
Business Group Dummy (Imputed)			-113.331	
			(44.114)	
Gap in university science, 1921 (AMS) × HHI above median (pre 1930)				-5.488
				(102.303)
HHI above median (pre 1930)				0.095
				(31.851)
ln(American university science), 1921 (AMS)	10.655	16.264	18.979	51.327
	(8.024)	(7.531)	(18.062)	(21.759)
ln(European university science), 1921 (AMS)	-6.762	-13.407	-10.890	-42.070
	(8.094)	(7.462)	(17.962)	(21.704)
ln(Assets)	5.251	5.798	6.486	7.442
	(0.759)	(0.898)	(1.053)	(1.061)
Average of Dependent Variable	4.967	3.843	6.625	7.886
Year Fixed Effects	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes
R^2	0.191	0.161	0.184	0.213
Number of Firms	396	358	199	196
Number of Observations	2,948	4,028	1,702	1,855

Notes: Unit of analysis is at the firm-year level. The variable definitions are identical to those in table 2, 5 and 6. Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

Table 8: Corporate Science and "Home-Run" Patents (Novelty)

	DV: Top 5% Novelty				
	(1) OLS	(2) 1st Stage IV	(3) 2nd Stage IV	(4) 1st Stage IV	(5) 2nd Stage IV
Publication Stock (100s)	0.273		0.470		0.548
	(0.020)		(0.050)		(0.071)
Gap in university science, 1900-20 (WOS)		6.040			
		(1.101)			
Gap in university science, 1921 (AMS)				123.111	
				(25.082)	
Gap in university science, current (WOS)	0.433	-0.742	0.457	1.317	0.467
	(0.417)	(1.233)	(0.410)	(1.272)	(0.447)
ln(Assets)	0.762	2.436	0.310	2.277	0.131
	(0.160)	(0.404)	(0.143)	(0.384)	(0.200)
ln(Patent stock)	1.868	4.229	1.049	4.181	0.724
	(0.153)	(0.480)	(0.160)	(0.474)	(0.228)
Average of Dependent Variable	2.895		2.895		2.895
Kleibergen-Paap rk Wald F statistic		30.099		24.091	
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes
R^2	0.635				
Number of Firms	425	425	425	425	425
Number of Observations	4,293	4,293	4,293	4,293	4,293

Notes: Analysis is at the firm-year level. Industry Fixed Effects are applied at 2 digit SIC codes. The instrument is the firm-level *Void in university science*. The dependent variable in columns 3-4 is the number of firm patents in the bottom 5% measured by IPC Combination Familiarity (Fleming, 2001). Columns 2-3 present first and second stage estimation results where the instrument is the ratio of European papers to American ones (*Gap in university science, 1900-20 (WOS)*) for papers published between 1900 and 1920. Columns 4-5 present first and second stage results where the instrument is the ratio of European-affiliated scientists to those without European affiliations (*Gap in university science, 1921 (AMS)*). *Gap in university science, current (WOS)* calculates the ratio of European to American papers for the focal year. Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

Table 9: CORPORATE SCIENCE AND "HOME-RUN" PATENTS (STOCK MARKET VALUE)

	DV: Top 5% Xi				
	(1) OLS	(2) 1st Stage IV	(3) 2nd Stage IV	(4) 1st Stage IV	(5) 2nd Stage IV
Publication Stock (100s)	0.040		0.145		0.183
	(0.012)		(0.027)		(0.040)
Gap in university science, 1900-20 (WOS)		7.996			
		(1.320)			
Gap in university science, 1921 (AMS)				154.720	
				(29.599)	
Gap in university science, current (WOS)	-0.702	-2.036	-0.574	0.921	-0.528
	(0.309)	(1.631)	(0.277)	(1.703)	(0.297)
ln(Assets)	1.418	2.565	1.169	2.384	1.079
	(0.218)	(0.475)	(0.200)	(0.458)	(0.185)
ln(Patent stock)	0.587	5.259	0.050	5.189	-0.143
	(0.081)	(0.568)	(0.088)	(0.561)	(0.136)
Average of Dependent Variable	0.861		0.861		0.861
Kleibergen-Paap rk Wald F statistic		36.708		27.324	
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes
R^2	0.247				
Number of Firms	327	327	327	327	327
Number of Observations	3,569	3,569	3,569	3,569	3,569

Notes: Analysis is at the firm-year level. Industry Fixed Effects are applied at 2 digit SIC codes. The instrument is the firm-level *Void in university science*. The dependent variable in columns 1-2 is the number of focal firms patents in the top 5% of U.S. patents issued in the focal year, measured by KPSS patent values (Kogan, Papanikolaou, Seru, & Stoffman, 2017). Columns 2-3 present first and second stage estimation results where the instrument is the ratio of European papers to American ones (*Gap in university science, 1900-20 (WOS)*) for papers published between 1900 and 1920. Columns 4-5 present first and second stage results where the instrument is the ratio of European-affiliated scientists to those without European affiliations (*Gap in university science, 1921 (AMS)*). *Gap in university science, current (WOS)* calculates the ratio of European to American papers for the focal year. Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

Table 10: CORPORATE SCIENCE AND MARKET-TO-BOOK RATIOS

	DV: ln(Tobin's Q)						
	Baseline (OLS)	WoS Gap (OLS)		AMS Gap (OLS)		WoS Gap (IVE)	
	(1)	(2) Small	(3) Large	(4) Small	(5) Large	(6) 1st Stage	(7) 2nd Stage
$\ln(\text{Pubstock}_{t-1})$	0.032	-0.001	0.045	0.022	0.039		-0.027
	(0.008)	(0.019)	(0.009)	(0.016)	(0.010)		(0.082)
Gap in university science, 1900-20 (WOS)						0.221	
						(0.041)	
$\ln(\text{Patstock}_{t-1})$	0.023	0.042	0.010	0.032	0.010	0.223	0.036
	(0.004)	(0.007)	(0.006)	(0.007)	(0.006)	(0.013)	(0.018)
Average of Dependent Variable	0.591	0.607	0.575	0.604	0.577		0.591
Kleibergen-Paap rk Wald F statistic						26.941	
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes		Yes
Industry Fixed Effects	Yes	Yes	Yes	Yes	Yes		Yes
R^2	0.345	0.400	0.395	0.391	0.422		0.010
Number of Firms	325	170	155	171	154		325
Number of Observations	3,399	1,740	1,659	1,789	1,610	3,399	3,399

Notes: Unit of analysis is at the firm-year level. Columns 1-5 present results from estimating the Tobin's Q equation with logged publication stock. Columns 6 and 7 present the first and second stage IV estimates where the instrument is the ratio of European to American publications published between 1900 and 1920 (*Gap in university science, 1900-20 (WOS)*). Columns 2 and 3 are split by median values of the *Gap in university science, 1900-20 (WOS)* measure (2 being below median and 3 above). Columns 4 and 5 are split by median values of the *Gap in university science, 1921 (AMS)* measures (4 being below median and 5 being above). Industry Fixed Effects are applied at 2 digit SIC codes. Standard errors are robust to arbitrary heteroscedasticity.

References

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., & Patacconi, A. (2019). A theory of the US innovation ecosystem: Evolution and the social value of diversity [Publisher: Oxford University Press]. *Industrial and Corporate Change*, *28*(2), 289–307.

Arora, A., Belenzon, S., Patacconi, A., & Suh, J. (2020a). The changing structure of American innovation: Some cautionary remarks for economic growth. *Innovation Policy and the Economy*, *20*(1), 39–93.

Arora, A., Belenzon, S., & Sheer, L. (2020b). Back to Basics: Why do Firms Invest in Research? *American Economic Review*.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention, In *The rate and direction of inventive activity: Economic and social factors*. Princeton University Press.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & Van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics*, *135*(2), 645–709. https://doi.org/10.1093/qje/qjaa004

Belenzon, S., & Berkovitz, T. (2010). Innovation in business groups. *Management Science*, *56*(3), 519–535.

Belenzon, S., Berkovitz, T., & Rios, L. A. (2013). Capital markets and firm organization: How financial development shapes European corporate groups. *Management Science*, *59*(6), 1326–1343.

Belenzon, S., & Tsolmon, U. (2016). Market frictions and the competitive advantage of internal labor markets. *Strategic Management Journal*, *37*(7), 1280–1303.

Berle, A., & Means, G. (1932). Private property and the modern corporation. *New York: Mac-millan.*

Bonbright, J. C., & Means, G. C. (1932). *The holding company: Its public significance and its regulation.* McGraw-Hill Book Company.

Buchanan, N. S. (1936). The public utility holding company problem [Publisher: HeinOnline]. *Calif. L. Rev.*, *25*, 517.

Chandler Jr, A. D. (1993). *The visible hand*. Harvard University Press.

Council, N. R. (1938). *Research: A National Resource : II. Industrial Research*. Washington, DC, The National Academies Press. https://doi.org/10.17226/20233

Enos, J. L. (1962). Invention and innovation in the petroleum refining industry, In *The rate and direction of inventive activity: Economic and social factors*. Princeton University Press.

Field, A. J. (2003). The most technologically progressive decade of the century. *American Economic Review*, *93*(4), 1399–1413.

Fleming, D., & Bailyn, B. (1968). *The intellectual migration: Europe and America, 1930-1960* (Vol. 2). Charles Warren Center for Studies in American History, Harvard University.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management science*, *47*(1), 117–132.

Furman, J. L., & MacGarvie, M. J. (2007). Academic science and the birth of industrial research laboratories in the US pharmaceutical industry [Publisher: Elsevier]. *Journal of Economic Behavior & Organization*, *63*(4), 756–776.

Geiger, R. L. (1986). *To advance knowledge : The growth of American research universities, 1900-1940.* New York : Oxford University Press, 1986.
Bibliography: p. 279-320.

Graham, J. R., Leary, M. T., & Roberts, M. R. (2015). A century of capital structure: The leveraging of corporate America [Publisher: Elsevier]. *Journal of Financial Economics*, *118*(3), 658–683.

Gutiérrez, G., & Philippon, T. (2020). *Some Facts about Dominant Firms* (tech. rep.). National Bureau of Economic Research, Inc.

Holton, G. (1981). The formation of the American physics Community in the 1920s and the coming of Albert Einstein [Publisher: Springer]. *Minerva*, *19*(4), 569–581.

Hounshell, D., & Smith, J. K. J. (1988). *Science and corporate strategy : Du Pont R&D, 1902-1980*. Cambridge [Cambridgeshire] ; New York : Cambridge University Press, 1988. https://find.library.duke.edu/catalog/DUKE000776227
Bibliography: p. 613-728.

Hounshell, D. A. (1988). The Making of the Synthetic Fiber Industry in the United States.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science [Publisher: Oxford University Press]. *The Quarterly Journal of Economics*, *133*(2), 927–991.

Jenkins, R., & Chandler, A. D. (1975). *Images and enterprise: Technology and the American photographic industry, 1839 to 1925*. Johns Hopkins University Press Baltimore.

Kandel, E., Kosenko, K., Morck, R., & Yafeh, Y. (2019). The great pyramids of America: A revised history of US business groups, corporate ownership, and regulation, 1926–1950. *Strategic Management Journal*, *40*(5), 781–808.

Kevles, D. (1979). The physics, mathematics, and chemistry communities: A comparative analysis (A. Oleson & J. Voss, Eds.). In A. Oleson & J. Voss (Eds.), *The Organization of Knowledge in Modern America, 1860-1920*. Baltimore, Johns Hopkins University Press. Retrieved January 28, 2019, from http://resolver.caltech.edu/CaltechAUTHORS:20171127-134808791

Khanna, T. (2000). Business groups and social welfare in emerging markets: Existing evidence and unanswered questions [Publisher: Elsevier]. *European Economic Review*, *44*(4-6), 748–761.

Khanna, T., & Palepu, K. (2000). Is group affiliation profitable in emerging markets? An analysis of diversified Indian business groups [Publisher: Wiley Online Library]. *The journal of finance*, *55*(2), 867–891.

Khanna, T., & Yafeh, Y. (2007). Business groups in emerging markets: Paragons or parasites? *Journal of Economic literature*, *45*(2), 331–372.

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

Lamoreaux, N. R., & Sokoloff, K. L. (1997). *Inventors, Firms, and the Market for Technology in the Late Nineteenth and Early Twentieth Centuries*. National Bureau of Economic Research Cambridge, Mass., USA.

Lee, D. L., McCrary, J., Moreira, M. J., & Porter, J. (2020). Valid t-ratio Inference for IV. *arXiv preprint arXiv:2010.05058*.

Lundberg, F. (1937). *America's 60 families*. Vanguard Press New York.

Maclaurin, W. R., & Harman, R. J. (1949). *Invention & innovation in the radio industry*. Macmillian.

MacLeod, W. B., & Urquiola, M. S. (2020). Why Does the US Have the Best Research Universities? Incentives, Resources, and Virtuous Circles. *NBER Working Paper*.

Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles [Publisher: Wiley Online Library]. *Strategic Management Journal*.

Mauro, P., Sussman, N., & Yafeh, Y. (2002). Emerging Market Spreads: Then versus Now*. *The Quarterly Journal of Economics*, *117*(2), 695–733. https://doi.org/10.1162/003355302753650364

Mauro, P., Sussman, N., & Yafeh, Y. (2006). *Emerging markets and financial globalization: Sovereign bond spreads in 1870-1913 and today*. Oxford University Press.

Moser, P., & San, S. (2020). Immigration, Science, and Invention. Lessons from the Quota Acts. *Lessons from the Quota Acts (March 21, 2020)*.

Moser, P., Voena, A., & Waldinger, F. (2014). German Jewish émigrés and US invention. *American Economic Review*, *104*(10), 3222–55.

Mowery, D., & Rosenberg, N. (1998). *Paths of innovation : Technological change in 20th-century America*. Cambridge ; New York : Cambridge University Press, 1998.
Includes bibliographical references (p. 181-199) and index.

Mowery, D. C., & Rosenberg, N. (1999). *Paths of innovation: Technological change in 20th-century America*. Cambridge University Press.

Nelson, R. R. (1959a). The economics of invention: A survey of the literature. *The Journal of Business*, *32*(2), 101–127.
Nelson, R. R. (1959b). The simple economics of basic scientific research. *Journal of political economy*, *67*(3), 297–306.
Nicholas, T. (2010). The role of independent invention in US technological development, 1880–1930 [Publisher: JSTOR]. *The Journal of Economic History*, 57–82.
Nicholas, T. (2011). Did R&D firms used to patent? Evidence from the first innovation surveys [Publisher: Cambridge University Press]. *The Journal of Economic History*, *71*(4), 1032–1059.
Nicholas, T. (2014). *Scale and innovation during two US breakthrough eras*. Harvard Business School.
Perazich, G., & Field, P. M. (1940). *Industrial research and changing technology*. WPA, Nat. Research Project.
Pinchot, G. (1928). The power monopoly; its make-up and its menace.
Rajan, R. G., & Zingales, L. (2003). The great reversals: The politics of financial development in the twentieth century. *Tuck Symposium on Corporate Governance*, *69*(1), 5–50. https://doi.org/10.1016/S0304-405X(03)00125-9
Reich, L. S. (1985). *The making of American industrial research : Science and business at GE and Bell, 1876-1926*. Cambridge Cambridgeshire ; New York, Cambridge University Press
Leonard S. Reich. ill. ; 24 cm. Includes index. Bibliography: p. 258-301. Studies in economic history and policy.
Richter, R., & Streb, J. (2011). Catching-up and falling behind: Knowledge spillover from American to German machine toolmakers [Publisher: Cambridge University Press]. *The Journal of Economic History*, *71*(4), 1006–1031.
Rosenberg, N., & Landau, R. (1992). Successful Commercialization in the Chemical Process Industries, In *Technology and the Wealth of Nations*. Stanford University Press.
Rüegg, W. (2004). *A history of the university in Europe: Volume 3, universities in the nineteenth and early twentieth centuries (1800–1945)* (Vol. 3). Cambridge University Press.
Schumpeter, J. A. (1939). *Business cycles* (Vol. 1). McGraw-Hill New York.
Service, N. R. C. U. R. I. (1931). *Industrial Research Laboratories of the United States, Including Consulting Research Laboratories*. RR Bowker Company.
Servos, J. W. (1994). Changing partners: The Mellon Institute, private industry, and the federal patron [Publisher: JSTOR]. *Technology and Culture*, *35*(2), 221–257.
Urquiola, M. (2020). *Markets, Minds, and Money: Why America Leads the World in University Research*. Harvard University Press.
Wise, G. (1985). *Willis R. Whitney, General Electric, and the origins of US industrial research*. Columbia University Press.
Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S. (2017). Nationality classification using name embeddings, In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM.

A Details on Firm Ownership Data Construction

A.1 Corporate Historical Documents and Data Sources

- Bureau of Economic Analysis (BEA, 1958), U.S. Department of Commerce, Benchmark Federal Trade Commission (FTC) Annual Reports: www.ftc.gov/os/annualreports/index.shtm
- Input-Output Data: Historical SIC Data, www.bea.gov/industry/io_histsic.htm
- Interstate Commerce Commission (ICC) Reports
- Moody's Manuals, 1926-1940: http://webreports.mergent.com/
- Statistics of Income: http://www.irs.gov/pub/irs-soi/
- National Association of Railroad and Utility Commissioners
- National Resources Committee (NRC) (1939), The Structure of the American Economy (Washington DC: US Government Print Office).
- Regulation of Stock Ownership in Railroads, 71st Congress, 3d Session, House Report No. 2789, Vol.2, February 1931
- Securities and Exchange Commission (SEC) Annual Reports: www.sec.gov/about/annrep.shtml
- Survey of American Listed Corporations: Reported Information on Registrants with the SEC under the Securities Exchange Act of 1934, 1939-40.
- Temporary National Economic Committee (TNEC), (1940), The Distribution of Ownership in the 200 Largest Nonfinancial Corporations, monograph 29 (1-2) (Washington DC: U.S. Government Printing Office): http://www.bpl.org/govinfo/online-collections/federal-executive-branch/temporary-national-e
- Twentieth Century Fund, Committee on Taxation (1937), Facing the Tax Problem (New York: Twentieth Century Fund).

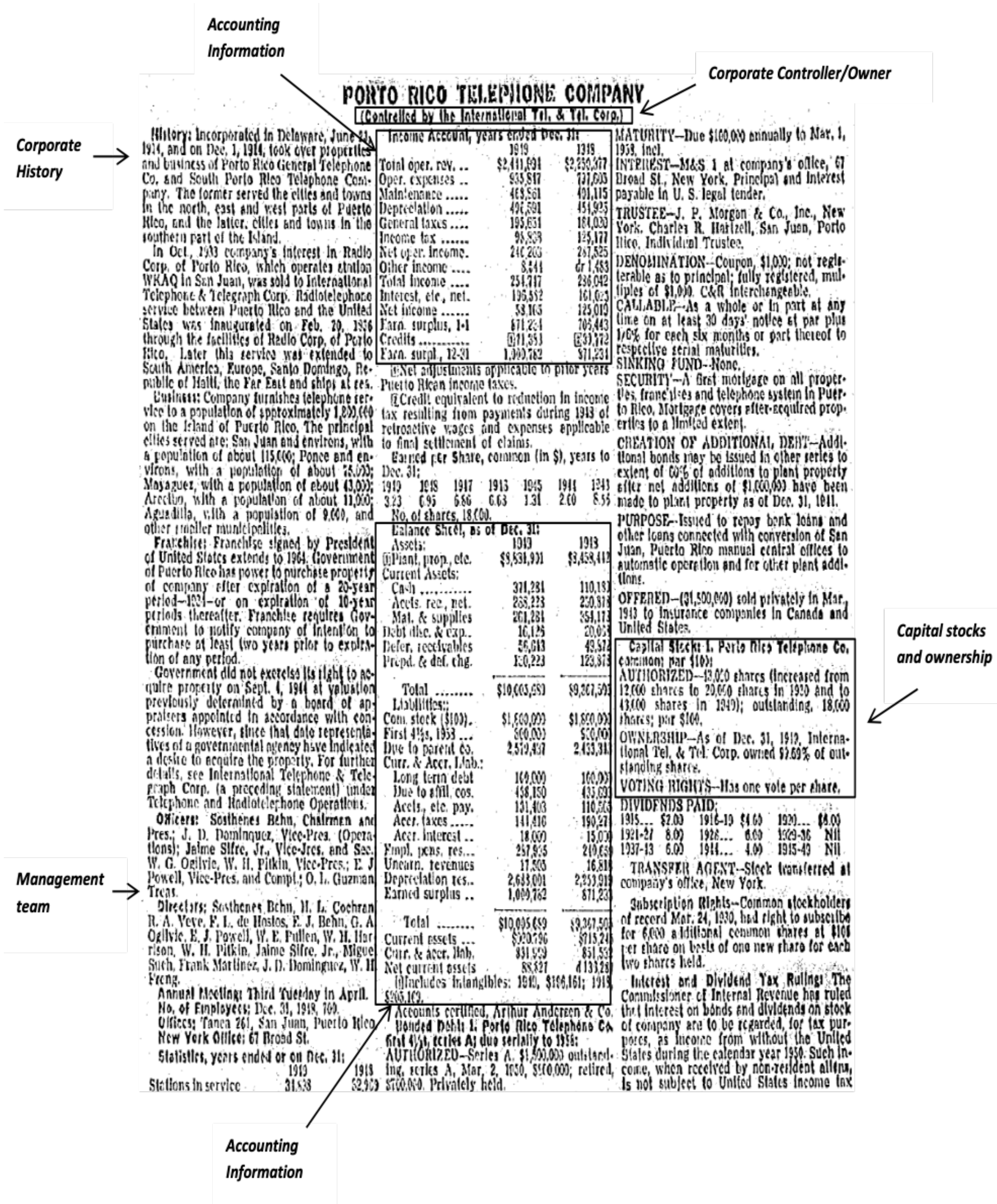
A.2 Corporate Histories

- http://www.Archive.org
- Encyclopedia of American Business History (Facts on File, 2005): http://www.Fundinguniverse.com
- The New York Times Archives: http://www.nytimes.com/ref/membercenter/nytarchive.html
- The Wall Street Journal Archives: http://pqasb.pqarchiver.com/wsj/search.html

A.3 Control Chains

We use Moody's Manuals to track companies controlling, or controlled by, the 200 companies on the B&M list. In each volume, a company report is followed by reports on its controlled subsidiaries (which are identified without an explicitly specified control threshold held by the controlling company). For example, if company A controls company B and company B, in turn, controls company C and all three firms belong to the railroad sector, the A-B-C control chain will appear in Moody's Railroads Manual in the same sequence with the identity of the corporate controller usually reported next to the company name. We examine if one or more companies are controlled by another corporation included in the original list and,

Figure 6: A Moody's Manuals entry: The Porto Rico Telephone Company, 1949



PORTO RICO TELEPHONE COMPANY

(Controlled by the International Tel. & Tel. Corp.)

History: Incorporated in Delaware, June 21, 1914, and on Dec. 1, 1914, took over properties and business of Porto Rico General Telephone Co. and South Porto Rico Telephone Company. The former served the cities and towns in the north, east and west parts of Puerto Rico, and the latter, cities and towns in the southern part of the Island.

In Oct., 1933 company's interest in Radio Corp. of Porto Rico, which operates station WKAQ in San Juan, was sold to International Telephone & Telegraph Corp. Radiotelephone service between Puerto Rico and the United States was inaugurated on Feb. 20, 1936 through the facilities of Radio Corp. of Porto Rico. Later this service was extended to South America, Europe, Santo Domingo, Republic of Haiti, the Far East and ships at sea.

Business: Company furnishes telephone service to a population of approximately 1,800,000 on the Island of Puerto Rico. The principal cities served are: San Juan and environs, with a population of about 115,000; Ponce and environs, with a population of about 75,000; Mayaguez, with a population of about 43,000; Arecibo, with a population of about 11,000; Aguadilla, with a population of 9,000, and other smaller municipalities.

Franchise: Franchise signed by President of United States extends to 1964. Government of Puerto Rico has power to purchase property of company after expiration of a 20-year period—1934—or on expiration of 10-year periods thereafter. Franchise requires Government to notify company of intention to purchase at least two years prior to expiration of any period.

Government did not exercise its right to acquire property on Sept. 4, 1944 at valuation previously determined by a board of appraisers appointed in accordance with concession. However, since that date representatives of a governmental agency have indicated a desire to acquire the property. For further details, see International Telephone & Telegraph Corp. (a preceding statement) under Telephone and Radiotelephone Operations.

Officers: Sosthenes Behn, Chairman and Pres.; J. D. Dominguez, Vice-Pres. (Operations); Jaime Sifre, Jr., Vice-Pres. and Sec.; W. G. Ogilvie, W. H. Pitkin, Vice-Pres.; E. J. Powell, Vice-Pres. and Compt.; O. L. Guzman, Treas.

Directors: Sosthenes Behn, H. L. Cochran, R. A. Veve, F. L. de Hostos, E. J. Behn, G. A. Ogilvie, E. J. Powell, W. E. Pullen, W. H. Harrison, W. H. Pitkin, Jaime Sifre, Jr., Miguel Such, Frank Martinez, J. D. Dominguez, W. H. Feeng.

Annual Meeting: Third Tuesday in April.

No. of Employees: Dec. 31, 1948, 700.

Offices: Tanca 261, San Juan, Puerto Rico.

New York Office: 67 Broad St.

Statistics, years ended or on Dec. 31:

	1949	1948
Stations in service	31,838	32,909

Income Account, years ended Dec. 31:

	1949	1948
Total oper. rev. ..	$2,441,891	$2,250,377
Oper. expenses ..	935,817	737,603
Maintenance	463,561	401,115
Depreciation	496,591	454,925
General taxes	195,651	184,030
Income tax	98,808	125,177
Net oper. income.	240,203	267,525
Other income	8,441	dr 1,483
Total income	254,717	236,042
Interest, etc., net.	196,552	161,023
Net income	58,165	125,019
Earn. surplus, 1-1	871,231	705,443
Credits	(j)71,353	(k)30,372
Earn. surpl., 12-31	1,000,762	871,231

(j)Net adjustments applicable to prior years Puerto Rican income taxes.

(k)Credit equivalent to reduction in income tax resulting from payments during 1948 of retroactive wages and expenses applicable to final settlement of claims.

Earned per Share, common (in $), years to Dec. 31:

1949	1948	1947	1946	1945	1944	1943
3.23	6.95	6.86	6.63	1.31	2.00	8.55

No. of shares, 18,000.

Balance Sheet, as of Dec. 31:

Assets:	1949	1948
(i)Plant, prop., etc.	$8,831,901	$8,458,412
Current Assets:		
Cash	371,281	110,181
Accts. rec., net.	288,223	250,878
Mat. & supplies	261,281	354,173
Debt disc. & exp..	16,125	20,063
Defer. receivables	56,613	49,574
Prepd. & def. chg.	130,223	123,873
Total	$10,005,989	$9,367,503
Liabilities:		
Com. stock ($100).	$1,800,000	$1,800,000
First 4½s, 1958 ...	800,000	500,000
Due to parent co.	2,579,437	2,433,311
Curr. & Accr. Liab.:		
Long term debt	100,000	100,000
Due to affil. cos.	458,150	435,693
Accts., etc. pay.	131,403	110,563
Accr. taxes	141,416	190,271
Accr. interest ..	18,000	15,000
Empl. pens. res...	257,935	210,659
Unearn. revenues	17,505	16,818
Depreciation res..	2,683,001	2,253,919
Earned surplus ..	1,000,762	871,231
Total	$10,005,989	$9,367,503
Current assets ...	$920,796	$715,246
Curr. & accr. liab.	831,929	851,533
Net current assets	88,821	d 136,287

(i)Includes intangibles: 1949, $196,161; 1948, $205,169.

Accounts certified, Arthur Andersen & Co.

Bonded Debt: 1. Porto Rico Telephone Co. first 4½s, series A; due serially to 1958:

AUTHORIZED—Series A, $1,500,000 outstanding, series A, Mar. 2, 1950, $800,000; retired, $700,000. Privately held.

MATURITY—Due $100,000 annually to Mar. 1, 1958, incl.

INTEREST—M&S 1 at company's office, 67 Broad St., New York. Principal and interest payable in U. S. legal tender.

TRUSTEE—J. P. Morgan & Co., Inc., New York. Charles R. Hartzell, San Juan, Porto Rico, Individual Trustee.

DENOMINATION—Coupon, $1,000; not registerable as to principal; fully registered, multiples of $1,000. C&R interchangeable.

CALLABLE—As a whole or in part at any time on at least 30 days' notice at par plus 1/6% for each six months or part thereof to respective serial maturities.

SINKING FUND—None.

SECURITY—A first mortgage on all properties, franchises and telephone system in Puerto Rico. Mortgage covers after-acquired properties to a limited extent.

CREATION OF ADDITIONAL DEBT—Additional bonds may be issued in other series to extent of 60% of additions to plant property after net additions of $1,000,000 have been made to plant property as of Dec. 31, 1941.

PURPOSE—Issued to repay bank loans and other loans connected with conversion of San Juan, Puerto Rico manual central offices to automatic operation and for other plant additions.

OFFERED—($1,500,000) sold privately in Mar., 1943 to insurance companies in Canada and United States.

Capital Stock: 1. Porto Rico Telephone Co. common; par $100:

AUTHORIZED—43,000 shares (increased from 12,000 shares to 20,000 shares in 1930 and to 43,000 shares in 1949); outstanding, 18,000 shares; par $100.

OWNERSHIP—As of Dec. 31, 1949, International Tel. & Tel. Corp. owned 99.69% of outstanding shares.

VOTING RIGHTS—Has one vote per share.

DIVIDENDS PAID:

1915...	$2.00	1916-19	$4.00	1920...	$8.00
1921-27	8.00	1928...	6.00	1929-36	Nil
1937-43	6.00	1944...	4.00	1945-49	Nil

TRANSFER AGENT—Stock transferred at company's office, New York.

Subscription Rights—Common stockholders of record Mar. 24, 1930, had right to subscribe for 6,000 additional common shares at $105 per share on basis of one new share for each two shares held.

Interest and Dividend Tax Ruling: The Commissioner of Internal Revenue has ruled that interest on bonds and dividends on stock of company are to be regarded, for tax purposes, as income from without the United States during the calendar year 1950. Such income, when received by non-resident aliens, is not subject to United States income tax

Source: http://webreports.mergent.com

if this is the case, combine their control chains. Therefore, each control chain in our sample is a long sequence of firms consisting of an apex corporation and its subsidiaries, each of which has control over the next one. In most cases, control chains include firms belonging to the same industrial category (e.g., railroads), but there are occasionally multiple control chains in different categories with the same ultimate owner as well as (e.g., a few cases of public utility apex companies controlling industrial companies).

A.4 Ultimate Controlling Shareholders

Moody's Manuals do not provide any information on the identity of the controllers of apex firms. To identify the owners of apex corporations that are not controlled by any other entity, we use the following sources:

1. For the 1926-1929 period: Pinchot (1928), the Wall Street Journal (WSJ) and the New York Times (NYT) archives as well as additional sources such as internet searches, historical documents, corporate files, www.archives.org and www.fundinguniverse.com.

2. For the 1929-1932 period: Table XII, Berle and Means (1932), Bonbright and Means (1932), Buchanan (1936), Lundberg (1937), the Encyclopedia of American Business History (2006), the WSJ and NYT archives and www.fundinguniverse.com.

3. For the 1937-1940 period: National Resources Committee (1939, Chapter IX and Appendix 13) and TNEC (1940).

B Details on scientific gap calculations

B.1 American Men of Science Directory

The AMS directory lists information on each scientist in a consistent manner: the last name is followed by the title, first name, current employment and residence and main discipline. Information on date and place of birth, alma mater, past employment and membership in professional societies follow. The final item in each entry lists a detailed list of keywords that describe the focal scientist's research interests. We wish to extract i) the main discipline in which each scientist works in and ii) any European degrees conferred.

The general data challenge is that the OCR on the image files, while relatively high quality, still have high error rates when classifying punctuation marks (commas, periods, and semicolons) that are essential for separating out the entries into their constituent parts. Therefore, rather than splitting the text into its constituents, we directly search for the information we need. For main disciplines, we collect 131 scientific fields from a list of deceased scientists listed at the end of the 1906 and 1921 editions of AMS.[36] We conduct regular expressions (regex) on each AMS entry to determine which disciplines correspond to each scientist.[37] We further clean this data by determining the location of the regex match: if the matched discipline occurs after the birth date (Oct. 23, 75 for Gilbert Lewis in figure 7), we remove the match. This prevents descriptions for research interests that occur later ("Thermodynamic theory and its application to chemistry; ... ; entropy of elements; third law of thermodynamics") from matching as the main discipline in which the scientist works. In the case of Gilbert Lewis, we prevent terms such as "electro-magnetic theory" or "non-Newtonian mechanics" to match with stemmed tokens for "Electrical Engineering" and "Mechanical Engineering". Afterwards, we manually crosswalk disciplines found in AMS into their equivalents in OECD subfields.

[36] These are more feasible to collect manually as the entries are structured as names, discipline, years of birth and death.
[37] "Technology", "General Science", and "Engineering" without specifying a field (mechanical, civil, mechanical, chemical) are excluded as they are too general.

Figure 7: American Men of Science Entry for Gilbert Lewis (1921)

Lewis, Dr. G(ilbert) N(ewton), University of California, Berkeley, Calif. ***Chemistry.** Weymouth, Mass, Oct. 23, 75. Nebraska, 90–93; A.B, Harvard, 96, A.M, 98, Ph.D, 99; Leipzig and Göttingen, 00–01. Teacher, Phillips Acad, 96–97; instr. chem, Harvard, 99–00, 01–06, on leave in charge weights and measures, Bur. Govt. Laboratories, P. I, 04–05; asst. prof. physico-chem. research, Mass. Inst. Tech, 07–08, assoc. prof, 08–11, prof, 11–12, acting director, research lab, 07–09; *prof. chem. and dean col. chem, California, 12–* Major, lieut. col, chief of defense div, gas service, A.E.F, and chief of training div, C.W.S. Chevalier Légion d'honneur. Nat. Acad; Physical Soc; Chem. Soc; Philos. Soc; Am. Acad. Thermodynamic theory and its application to chemistry; free energy tables; equilibrium in numerous reactions; electric potentials of the common elements; properties of solutions and the activity of ions; distribution of thermal energy; specific heat of electrons; the principle of relativity and non-Newtonian mechanics; application of four-dimensional vector analysis to electro-magnetic theory; the geometry of the space time manifold of relativity; ultimate rational units; calculation of Stefan's constant; the structure of the atom and the molecule and the theory of valence; entropy of elements; third law of thermodynamics.

Source: Entry on Gilbert Lewis from the 1921 edition of the American Men of Science Directory.

For alma maters and professional experience, we collect the list of all universities that were active in Europe between 1801 and 1945 from Wikipedia,[38] which in turn is heavily based on Rüegg (2004). Similar to main disciplines, we use regular expressions to determine whether each entry contains a match to at least one of these universities.[39] We further clean this data by removing matches for migrants that were born in Europe but trained exclusively in America: any match that occurs before the birth date of the scientist is excluded. Even after this cleaning, there will remain cases where an American is trained (until his doctoral degree) in the United States, only to be recognized by foreign institutions. We therefore complement this with scientific publication output data from Clarivate Web of Science.

B.2 Web of Science Affiliations Coding

For the period between 1900 to 1920, the Microsoft Academic Graph data does not record the country of publication. Also, we find that the affiliations sections rarely list the full address of the author for this period, which leads MAG to omit country data from affiliation data. We therefore rely on Clarivate Web of Science, which has previously been used for research on the impacts of World War I on scientific production (Iaria et al., 2018). Of 307,847 publications listed in Web of Science, 15% (44,356) have country data.

Therefore, we classify the remaining 85% of publications by a combination of the Cattell directory

[38]https://en.wikipedia.org/wiki/List_of_modern_universities_in_Europe_(1801\OT1\textendash1945). We also collect data on early modern universities (established between 1501-1800) https://en.wikipedia.org/wiki/List_of_early_modern_universities_in_Europe and medieval universities (established before 1500) https://en.wikipedia.org/wiki/List_of_medieval_universities that were likely active in the early 20th century.

[39]Schools with very short names such as the University of Pau (France, 1722) and Literary University of Vic (Spain, 1599) are excluded because of high false positive match rates.

and a name-classification algorithm. We first classify any name that we are able to match to the 1906 and 1921 versions of the directory as American. For names not matched to the Cattell directory, we use a naive-bayes classifier (Ye et al., 2017) to identify European names.[40] The classifier outputs a probability that a name belongs to one of 37 nationalities.[41] For each name, we identify as "European" those names which are classified as Eastern European, Southern Slavic, Baltic, Romanian, Russian, Spanish, Portuguese, German, Italian, French, Danish, Finnish, Swedish, Greek, Norwegian, or "Celtic/English".

Table 11: WoS Countries and Regions (1/1)

country	region	country	region	country	region	country	region
Africa	ROW	London	EUR	Switzerland	EUR	Uganda	ROW
Argentina	ROW	Malta	EUR	Syria	ROW	Rwanda	ROW
Australia	ROW	Mexico	ROW	Thailand	ROW	Ruanda Urundi	ROW
Austria	EUR	Mozambique	ROW	The Netherlands	EUR	Nigeria	ROW
Bahamas	ROW	N WALES	EUR	Turkey	ROW	Manchuria	ROW
Barbados	ROW	N Z	EUR	UK	EUR	Esthonie	EUR
Belgium	EUR	Netherlands	EUR	Ukraine	EUR	Ecudor	ROW
Belize	ROW	New Zealand	ROW	Uruguay	ROW	BURMA	ROW
Bermuda	ROW	Nicaragua	ROW	USA	USA	West Africa	ROW
Brazil	ROW	North Ireland	EUR	USSR	EUR	Ukriane	EUR
British	EUR	North Wales	EUR	Venezuela	ROW	Southern India	ROW
British East Af	EUR	Norway	EUR	W Indies	ROW	Palestine	ROW
British Hondurs	EUR	NS Wales	EUR	Wales	EUR	NY	USA
Bulgaria	EUR	Nyasaland	ROW	Western Austral	ROW	Kenya Colony	ROW
BWI	EUR	NZ	ROW	WIA	ROW	ISA	?
Canada	ROW	P I	ROW	Yemen	ROW	Iraq	ROW
CEYLON	ROW	Panama	ROW	Yugoslavia	EUR	Great Britain	EUR
Chile	ROW	Peoples R China	ROW	SUISSE	EUR	East Africa	ROW
CHINA	ROW	Persia	ROW	Finnland	EUR	Yugoslavie	EUR
Colombia	ROW	Peru	ROW	BW1	EUR	Western Samoa	ROW
Costa Rica	ROW	Philippine Isl	ROW	West Indies	ROW	Saskatchewan	ROW
Croatia	EUR	Philippine Isla	ROW	Russland	EUR	Russian Turkest	EUR
Cuba	ROW	Philippines	ROW	Prague	EUR	No Ireland	EUR
Czech Republic	EUR	Philippines Isl	ROW	Pakistan	ROW	Jugoslavia	EUR
CZECHOSLOVAKIA	EUR	Phillipine Isla	ROW	Malaysia	ROW	Johannesburg	ROW
Denmark	EUR	PI	ROW	Argentine	ROW	Inida	ROW
Egypt	ROW	Poland	EUR	Taiwan	ROW	Indien	ROW
England	EUR	Portugal	EUR	Kenya	ROW	Estonie	EUR
Federated Malay	ROW	Prussia	EUR	Bengal	ROW	Cook Islands	EUR
Fiji	ROW	Romania	EUR	Fed Malay State	ROW	BRITISH W INDIES	EUR
Finland	EUR	Russia	EUR	South America	ROW	Sri Lanka	ROW
FMS	?	S AFRICA	ROW	Philippline Isl	ROW	Siberia	EUR
France	EUR	S Australia	ROW	Morocco	ROW	Lithuania	EUR
Germany	EUR	S India	ROW	Korea	ROW	Isle Wright	EUR
Greece	EUR	S Wales	EUR	Isle Of Man	EUR	Byelarus	EUR
Guatemala	ROW	Schweden	EUR	Engalnd	EUR	British West In	EUR
Guyana	ROW	Scotland	EUR	Ecuador	ROW	Philippine	ROW
HOLLAND	EUR	Senegal	ROW	Czechoslovakio	EUR	Belgian Congo	EUR
Honduras	ROW	Siam	ROW	Czechoslovak Re	EUR	Turkestan	ROW
Hong Kong	ROW	Sierra Leone	ROW	Columbia	ROW	Tunisia	ROW
Hungary	EUR	Singapore	ROW	Trinidad	ROW	Paris	EUR
India	ROW	South Africa	ROW	Tasmania	ROW	Maroc	ROW
Ireland	EUR	South Australia	ROW	Mauritius	ROW	Hongrie	EUR
Italien	EUR	South India	ROW	Estonia	EUR	Chili	ROW
Italy	EUR	South Korea	ROW	Esthonia	EUR	Tchecoslovaquie	EUR
Jamaica	ROW	Spain	EUR	Dutch E Indies	EUR	Haiti	ROW
Japan	ROW	Sudan	ROW	Dominican Repub	ROW	Berlin	EUR
Latvia	EUR	Sweden	EUR	Union Of South	ROW	Belguim	EUR

Notes: The table lists the country affiliations of publications found in Clarivate Web of Science's Science Citation Index-Expanded between 1900 and 1920. "Region" has been imputed by the authors.

[40] Please see https://www.name-prism.com/api for access to the API

[41] Full list of nationalities in Nameprism are: Eastern European, Southern Slavic, Baltics, Romanian, Russian, Celtic English, Spanish, Indonesian, Jewish, Portuguese, Chinese, German, Italian ,South Asian, French, Central Asian, Filipino, Bangladeshi, South African, Nubian, Thai, Vietnamese, Cambodian, Danish, Burmese, Finnish, Persian, Swedish, Maghreb, Greek, Pakistani ,Arabian Peninsula, West African, Japan, Norway, Turkish, Malay, South Korean, East African

Table 12: WOS Countries and Regions (2/2)

country	region	country	region	country	region
Union S Africa	ROW	Cihina	ROW	Paraguay	ROW
Ukraina	EUR	Britsh India	EUR	New Mexico	USA
Serbia	EUR	British India	EUR	Lebanaon	ROW
ROUMANIA	EUR	USRR	EUR	Istanbul	ROW
Polen	EUR	URSS	EUR	Isreal	ROW
Haut Congo Belg	EUR	UA	EUR	Estland	EUR
Gr Britain	EUR	Ruanda	ROW	E Indies	ROW
Chilli	EUR	R De P	ROW	E Africa	ROW
Cananda	ROW	N Ireland	EUR	Breslau	EUR
UdSSR	EUR	Dutch East Indi	EUR	Azerbaidjan	ROW
Schweiz	EUR	Czecho Solvakia	EUR	Venezuella	ROW
New Zeland	ROW	Cyprus	EUR	UL	?
Lebanon	ROW	Chechoslovakia	EUR	Slovenia	EUR
Israel	ROW	Cairo	ROW	Republic Chili	ROW
Iceland	EUR	Ukrainia	EUR	Phillipine Isl	ROW
Hawaii	USA	Scothland	EUR	Nothern Ireland	EUR
Czechoslvakia	EUR	Puerto Rico	ROW	Netherlands Ind	EUR
Abyssinia	ROW	Oslo	EUR	Lebanan	ROW
W Africa	ROW	Irlande	EUR	Georgian SSR	EUR
Sud Mandschurei	ROW	Guadeloupe	EUR	Denamrk	EUR
Porto Rico	ROW	Ethiopie	ROW	CSR	EUR
North Africa	ROW	C I	?	Britain	EUR
Netherland	EUR	BRASIL	ROW	Anglo Egyptian	EUR
Luxembourg	EUR	BELGIQUE	EUR	Yugoslavija	EUR
Irish Free Stat	EUR	Begium	EUR	Union South Afr	ROW
Iran	ROW	Bangladesh	ROW	UKx	EUR
Dominican Rep	ROW	W Germany	EUR	Sumatra	ROW
Central India	ROW	USRS	EUR	Slovakia	EUR
BWA	EUR	TH	?	RHODESIA	ROW
USAa	USA	S Africia	ROW	Northern Ireland	EUR
Ungarn	ROW	Republic Panama	ROW		
RUMANIA	EUR	Madras	ROW		
R Argentina	ROW	LURSS	EUR		
N Nigeria	ROW	Lettonia	EUR		
Mailand	ROW	Jugoslawien	EUR		
Jerusalem	ROW	Cameroon	ROW		
Czlchoslovakia	EUR	ARSSR	EUR		
Czechoslavakia	EUR	Zwitzerland	EUR		
United Kingdom	EUR	USS	USA		
Republ Libanaise	ROW	Rumanien	EUR		
Norwegen	EUR	N Rhodesia	ROW		
Northern Irelan	EUR	Czechoslovak	EUR		
Java	ROW	Cent India	ROW		
Hungry	EUR	Bolivia	ROW		
GSSR	EUR	Belgien	EUR		
Ethiopia	ROW	Armenia	EUR		
Egpyt	ROW	Uzbekistan	EUR		
Czecho Slovakia	EUR	Rep of Georgia	EUR		

Notes: The table lists the country affiliations of publications found in Clarivate Web of Science's Science Citation Index-Expanded between 1900 and 1920. "Region" has been imputed by the authors.