Growth Lab

Inventing Modern Invention: The Professionalization of Technological Progress in the US

Matte Hartog, Andres Gomez-Lievano, Ricardo Hausmann, and Frank Neffke

Growth Lab Working Paper Series
No. 229

April 2024

GROWTH LAB
HARVARD KENNEDY SCHOOL
79 JFK STREET
CAMBRIDGE, MA 02138

GROWTHLAB.HKS.HARVARD.EDU

This paper may be referenced as follows: Hartog, M., Gomez-Lievano, A., Hausmann, R., and Neffke, F. (2024). “Inventing Modern Invention: The Professionalization of Technological Progress in the US.” Growth Lab Working Paper, John F. Kennedy School of Government, Harvard University.

Inventing modern invention: the professionalization of technological progress in the US

Matte Hartog[1*], Andres Gomez-Lievano[1,3], Ricardo Hausmann[1], Frank Neffke[2,1]

[1]Growth Lab, Harvard University; 79 JFK Street, Cambridge, MA 02138, USA
[2]Complexity Science Hub; Josefstädter Str. 39, 1080 Wien, Austria
[3]Analysis Group; 111 Huntington Ave 14th floor, Boston, MA 02199, USA

[*]To whom correspondence should be addressed; E-mail: matte_hartog@hks.harvard.edu

Over the course of the mid-19th and early 20th century, the US transformed from an agricultural economy to the frontier in technology. To study this transition, we digitize half a million pages of patent yearbooks that describe inventors, organizations and technologies on over 1.6M patents. We combine this with demographic information from US census records and information on corporate research from large-scale repeated surveys of industrial research labs. Our data reveal that in the early 1920s a new system of innovation — based on teamwork and engineers — started to rapidly replace the existing craftsmanship-based invention that had dominated innovation in the 19th century. We argue that this new system relied on an organizational innovation: industrial research labs. These labs supported high-skill teamwork, replacing the collaborations within families with professional ties in firms and industrial research labs. The systemic shift in innovation had far-reaching consequences: it changed the division of labor in invention, led to an explosion of novelty and teamwork, and reshaped the geography of innovation in the US.

1 Introduction

In the canonical view of Schumpeterian innovation, inventors develop new technologies by recombining existing ones in new ways. Because the potential for new combinations increases exponentially with the stock of existing technologies (Arthur, 2009; Weitzman, 1998; Youn et al., 2015), innovation should in principle become easier as technological know-how accumulates. However, as the space of technological opportunity expands, searching it becomes harder. This reduces the productivity of research (Bloom et al., 2020) and forces inventors onto longer learning curves (Jones, 2009). However, it also puts a premium on teamwork (Wuchty et al., 2007): combining old technologies to produce new ones does not happen in the abstract, but often involves bringing various experts together who individually specialize, but collectively cover a broad knowledge base. Consistent with arguments put forward in studies of technological change by anthropologists (e.g., Henrich, 2004; Muthukrishna and Henrich, 2016; Shennan, 2001), we propose that social change does not necessarily always follow technological change, but rather, may also enable it. Accordingly, as a society's collective body of knowledge grows, it inevitably gets distributed across more and more people. The concomitant specialization increases the interdependencies among team members (Neffke, 2019), leading to rising coordination costs (Becker and Murphy, 1992). Organizational innovations that change these coordination costs by helping to aggregate the vast amounts of distributed knowledge that specialized learning generates can therefore have outsized effects on the nature and speed of collective learning. Here, we focus on a particularly prominent example of such an organizational innovation: the industrial research lab (Arora et al., 2020; Furman and MacGarvie, 2007; Gertner, 2012; Mowery, 1990; Mowery and Rosenberg, 1999; Reich, 1985). We show that the arrival of these labs in the US reverberated throughout the innovation system, from the micro-level of the characteristics of individual inventors and their inventions to the macro level of the US geography of innovation.

In their heydays, industrial research labs were a core component of the US innovation system. Corporate labs, such as AT&T's Bell Labs or Dupont's Experimental Station, employed hundreds of researchers with PhD degrees in engineering, physics, chemistry and mathematics (Gertner, 2012). They produced various Nobel prize winners and sometimes even entirely new academic subfields, such as information theory, pioneered by Bell Lab's Claude Shannon. Arora et al. (2020) even surmise that the accelerating growth of US labor productivity between 1920 and 1970, as well as its subsequent decline can in part be ascribed to the rise and fall of industrial research labs.

To study the role that industrial research labs played in US innovation, we digitize information on all but the universe of US patents granted between 1856 and 1945, including the inventors and firms to which intellectual property rights were assigned. Next, we link these inventors to census records and these firms to repeated economy-wide surveys of industrial research labs.

We then organize our analysis drawing on the analogy of knowledge production functions (Pakes and Griliches, 1980). In particular, we study how changes in the organizational mechanisms that coordinate invention affect the relation between (labor) inputs and characteristics of output of patented inventions. This yields a number of stylized facts about the shift in the US innovation system that unfolds in the 1920s:

- at the level of inventions, the 1920s mark a precipitous take-off of inventions that list radically new combinations of technologies;

- at the level of inventors, we observe lengthening learning curves, a rapid rise of engineers, a greater reliance an academic literature and the emergence of academic patenting;
- at the meso level, there is a sudden shift to teamwork that is not coordinated anymore through family ties but by organzational ties in firms and labs;
- at the macro level, invention reconcentrates in large cities, especially on the East Coast and in the Rustbelt and the new innovation system exhibits drastically lower participation rates of women and foreign-born inventors.

Together, these findings describe a broad systemic shift in the US innovation system that broadly resembles what historians of technological change have described as a move to "science-based" invention (Mowery and Rosenberg, 1999), which we will henceforth use as a shorthand reference for the new innovation system that emerges. We show that aspects of the emergence of science-based innovation in the US were intimately connected to the rise of the industrial research lab. First, almost all of the rise in teamwork between 1920 and 1945 can be attributed to research organized in firms and labs. Second, lab-based teams rarely leverage family ties, are more homogeneous in terms of skills and are more likely to engage in repeated and long-distance collaborations than teams outside of labs. Third, research labs produce disproportionately more inventions that rely on radically new combinations of technologies, especially when patents are filed by teams.

Our paper contributes to various debates in the literature on innovation and technological change. First and foremost, it highlights the importance of organizational innovation and how changes in the coordination of inventors created ripples throughout the economy, from the background and learning curves of individual inventors, to the nature and intensity of teamwork and ultimately the geography of innovation. Methodologically, it relates to large-scale efforts to study historical US patents (Esposito, 2023) and to link them to census data (Akcigit et al., 2017). Second, we contribute to the literature on economic history, corroborating several narratives about how the US transformed technologically and economically, such as the increasing dominance of large firms (Chandler Jr, 1993) and their industrial research labs (Gertner, 2012), as well as the rise of the Rustbelt as the technology hub of its era (Lamoreaux et al., 2007). Our contribution to this literature is that we quantify these phenomena and pinpoint them in time.[1] Third, we contribute to the literature in evolutionary economics and innovation, by focusing on the role of organizational change. In particular, we relate research inputs (inventors and inventor teams) and the organizational context in which they operate (research labs, firms, families) to the

[1]For instance, Lamoreaux and Sokoloff (2001) describe how corporate research evolves from an activity that was outsourced to contract-inventors in the 19th century, to in-house R&D departments in the 20th century. We corroborate, but also refine this: although already in the late 19th century, patents were often assigned to firms, the take-off of patenting by engineers and teams in large corporate labs was exceedingly rare before 1920.

technological content that is patented. Our findings offer a sharp contrast with Hanlon's (2022) account of the first industrial revolution in the UK, where progress was propelled by the emergence of engineers. Instead, we show that these engineers only started playing a role in the US 100 years later in the second industrial revolution and when they did, they did so in the context of the rise of industrial research labs. This also shows that, although associating radical innovation with the genius of lone inventor-entrepreneurs and more incremental innovation with organized corporate R&D – as in the literature on Schumpeterian regimes (Winter, 1984) – may have been warranted post World War II, it does not reflect the nature of invention in the early years of corporate R&D. Finally, we contribute a historical perspective to the contemporary literature on gender in innovation (Bell et al., 2019; Delgado and Murray, 2022; Ding et al., 2006; Ross et al., 2022), showing how the new science-based innovation regime affected the participation of women in US innovation.

We structure our exposition as follows. First, we provide some historical background on the rise of research labs in Germany and their subsequent adoption by US firms, setting this against broad changes in the US economy and US society. Next, we describe how we digitize data on patents and industrial research labs and merge them with historical census records. Finally, we use these data to describe the changing nature of US invention between 1856 and 1945. We first describe changes at the micro-level of inventors, their learning curves and the nature and novelty of their inventions. Next, we describe changes in teamwork and team coordination to then move to macro-level changes in participation of women and immigrants and the geography of invention. We end with a discussion of what our findings entail for our understanding of how societies learn as collectives and what lessons they hold for current policy debates.

2 Research labs

Research laboratories play an important role in the history of science. Scientist workplaces had existed for hundreds of years but the scientific research laboratory is a much more recent phenomenon. Although there were many precursors in, for instance, England and France, the scientific laboratory only truly came into being in early 19th century German chemistry (Rocke, 2021). The canonical example is the chemical laboratory founded by Justus von Liebig in the German town of Giessen in the 1820s (Michaelis, 2003), which acted as an inspiration for scientists in- and outside Germany, attracting scores of visitors who came to study its layout and operations (Schmidgen, 2021). Propelled by the so-called "laboratory revolution," such laboratories acted not just as places of research – where the experimental method was applied to methodically study and subjugate nature – but also as places of instruction – where this method was taught through demonstration of scientific experiments (e.g., Schmidgen, 2021).

Toward the end of the 19th century, the idea of the research lab had started spreading to the private sector. As in the academic laboratories, German chemistry played a leading role. By the 1870s and 1880s, large companies in the German dye industry, such as BASF and Hoechst had laboratories devoted to research that were integral parts of their corporate structures (Travis et al., 1992). These labs, in turn, were part of a larger system that had emerged in the German-speaking territories and that included university research programs, government and industry-sponsored research institutes, such as the Kaiser Wilhelm Society (forerunner of today's Max Planck institutes), and industrial R&D programs (Lenoir, 1998; Pithan, 2021). A key aspect of this system was the industrial sponsorship of research within universities, which aimed to benefit firms by collaborating with professors and their graduate students. Such collaborations were leveraged to establish in-house R&D organizations that set the standard for future science-based industries.[2]

These highly organized corporate laboratories stood in stark contrast to the laboratories of famous inventor-entrepreneurs elsewhere, such as Thomas Edison in the US or William H. Perkin in the UK (Travis et al., 1992). Although their laboratories were prolific producers of patents, they were not created as organizational units within large industrial firms to further the competitive position of these firms. On the contrary, in the US, many of the inventor-entrepreneurs did not aim to commercialize their inventions themselves, but rather sold them on a well-developed market for technology (Lamoreaux and Sokoloff, 2001). Consequently, in the 19th century, most US research labs resembled, and were extensions of, workshops of such individual inventors - they remained key to technological development (Nicholas, 2010). For instance, Edison's "Invention Factory" at Menlo Park (NJ) was controlled by Edison himself, not by his company, General Electric (GE). In fact, when GE set up its own research lab in 1900, it had no direct connection to Edison's lab (Travis et al., 1992).

This changed in the early 20th century with the rise of the kind of corporate research labs that had been pioneered in Germany. This shift from the entrepreneurial inventor to corporate science-based innovation was enabled by a number of changes in the institutional organization and economy of the US.

First, over the course of the 19th century, intellectual property rights of patents became increasingly assigned to firms. Although from its inception, the US patent office only granted inventions to individuals, not to firms, inventors were, and still are, allowed to transfer their intellectual property rights to others, including firms. Because commercializing ideas requires capital and organizational capacity, inventors often assigned their inventions either to individuals who provided financial backing or to firms. This led to a vibrant market for technology in the 19th century US economy, where firms devoted substantial resources not only to commercialize ideas of outside inventors, but also to

[2]For instance, Hounshell (1996) describes how Friedrich Bayer A.G. (later I.G. Farben) had already developed a comprehensive R&D structure by 1891. This consisted of a central research laboratory equipped with cutting-edge scientific instruments, a scientific and patent library, and a seminar room, complemented by more specialized application laboratories, staffed by scientists with doctorates from German research universities.

identify these ideas. For instance, Lamoreaux and Sokoloff (1999) document the emergence of a veritable industry around the provision of information about, and access to, new inventions, consisting of patent lawyers, technology intermediaries and specialized trade journals. Meanwhile, although firms were often granted free licenses to use their employees' inventions, it would take them until well into the 20th century to habitually draft contracts that forced workers to assign their patents to them (Lamoreaux and Sokoloff, 1999). In fact, in much of the 19th century, companies did not seem to see much value in hiring workers explicitly for their ability to invent,[3] an attitude that only started changing in the early decades of the 20th century.

Second, the late 19th century set the stage for an increasing scale in production (Pithan, 2021). Advances in transportation and communication had turned the US into a vast, integrated market. This scale was soon exploited by the rise of vertically integrated enterprises, organized as multi-divisional corporations. A wave of mergers sparked by the 1890 Sherman Antitrust Act then gave birth to giant conglomerates at the start of the 20th century (Bittlingmayer, 1985; Chandler Jr, 1990, 1993). A byproduct of the increased scale of operations of these firms was that technological improvements, which could be applied to the entire volume of production, became much more valuable, providing a strong rationale for investments in corporate R&D (e.g., Klepper, 1996).

Third, throughout the late 19th and early 20th century the US population had become increasingly educated. For instance, the high school movement (Goldin, 1998) led to an increase in secondary school graduation rates from 6% of the US population in 1900 to 30% in 1930 (Goldin and Katz, 2000). This coincided with the growth of the university system (e.g., Goldin and Katz, 1999), both in the number and size of universities, leading to a growth in higher-education enrollment rates from 3% of 18-year-olds in 1890, to 4% in 1900, 5% in 1910, 8% in 1920 and 16% in 1940 (Cohen, 2009). One key driver was the aforementioned rise of high school education. Another important shift was the Morrill Land Grant Acts of 1862 and 1890 that allowed setting up new land-grant colleges with the proceeds from sales of federally owned land. Although originally, the main focus of these new colleges lay on agricultural and mechanical studies, nowadays they include broad-based top ranking universities such as the Massachusetts Institute of Technology, Purdue University and Cornell University.

Finally, the early 20th century marks a shift in attitude toward science among management and in society as a whole. Pithan (2021) provides a detailed analysis of this, by studying the discourse that developed around industrial research. This work shows that in the first decade of the 20th century, an increasing societal emphasis on efficiency and conservation of natural resources elevated the prestige of science as a methodical way to rationalize production processes over traditional trial-and-error methods. In this context,

[3]This is vividly illustrated by the position that Lockwood, head of American Bell's patent department, expressed in a letter to general management in 1885: "I am fully convinced that it has never, is not now, and never will pay commercially, to keep an establishment of professional inventors, or of men whose chief business it is to invent [...]" (cited in Lamoreaux and Sokoloff, 1999, p. 42).

the idea of the laboratory – more than the university – as the "physical manifestation of science [and thus] the place of the scientific method," where professional, university-trained, scientists engaged in a "quest for truth leading to the superiority over nature and material welfare" (Pithan, 2021, p. 202) became corporations' vehicle of choice for embracing science.[4]

However, what likely impacted the public's perception of science, and therewith of the research lab, most was the outbreak of WWI. The war, and the British naval blockade of German products that accompanied it, dramatically exposed US companies' reliance on German labs and their chemicals, dyes, and other key materials (Carlson, 2013). Scientific research and its industrial applications became seen as the source of military power and industrial research therewith as a national duty to counter German supremacy (Pithan, 2021). After the war, the US government responded by initiatating a program to help firms compete with German companies after the war.[5] The result was a flurry of dedicated industrial research labs, physically separated from manufacturing sites and staffed by workers with expertise in science and advanced engineering (e.g., Carlson, 2013).

Although important, the institutional changes in patenting, education and industrial organization offer at best a partial explanation for the rise of industrial research labs that would take place in the 1920s. After all, many of these developments had already been underway for several decades and are therefore best understood as enabling circumstances that provided a fertile breeding ground for these labs. Meanwhile the actual take-off of US industrial research in the 1920s seems to have been triggered by the changes in attitudes towards science and industrial research described by Pithan (2021). However, because we are not primarily concerned with what causes the emergence of industrial research labs in the US, we won't take a strong stance on this. Instead, we focus on the marked traces these labs left in the patent record and how they coincided with major changes in the US innovation system.

3 Data

To study long-term changes in US invention, we focus on inventive activity between 1856 and 1945. We combine information from three different sources: patent yearbooks, the complete US Census from 1850 to 1940 (with the exception of the 1890 census, which was destroyed in a fire), and large-scale surveys of industrial research labs between 1920 and 1950. To extend our analysis until the year 2000, we supplement these data with records from EPO-PATSTAT and PatentsView. In this section, we give a high-level overview of how we collect, process and merge these data, providing further details in Appendix A.

[4]This shift was also reflected in the emergence of new terms such as "industrial chemist" in 1908 and "industrial research" in 1912 (Pithan, 2021).

[5]For instance, based on his study of DuPont and Kodak, Hounshell concluded that "without question, then, World War I led to a widespread quickening of interest in and enthusiasm for industrial R&D in the United States" (Hounshell, 1996, p. 21).

3.1 Patents

We start our data collection by digitizing scans of the *Annual Reports by the Commissioner of Patents*, henceforth referred to as *(patent) yearbooks*.[6] These yearbooks contain information on all patents granted by the USPTO in a given year, including the name and location of residence of inventors and the individuals or organizations to whom a patent's intellectual property was assigned ("assignees"). Using multiple copies of each yearbook, we collect about half a million scanned pages. We convert these images to text strings, using image processing and optical character recognition (OCR) algorithms. Next, we apply named-entity recognition algorithms to identify patent numbers, grant dates, names and places of residence of inventors, as well as the names and locations of assignees. The result is a structured dataset that describes 1,591,361 patents granted by the USPTO between 1856 and 1953.

The structure of the patent yearbooks changes in 1954, omitting grant dates, assignees and inventor locations. For the period 1954-1968, we therefore use information from the European Patent Office's EPO-PATSTAT database (European Patent Office, 2020). This dataset provides inventor and assignee names, but omits information on inventors' locations of residence. Moreover, from 1969 to 1975, EPO-PATSTAT no longer reliably reports countries of residence, which renders the data unusable for our purposes. For the period 1976-2000, we use the USPTO's own PatentsView database (USPTO, 2022), which provides names and locations of inventors, as well as of assignees.

To verify the accuracy of this process, we compared the data we digitized to records in EPO-PATSTAT. These data are generally less complete and only start in 1900. However, where available, there is high agreement with ours. Furthermore, we compare our data with Petralia et al.'s (2016) HistPat data. HistPat data do not provide information on inventor or assignee names, but do record the inventors' places of residence. In the period we cover, our data have substantially fewer missing values and higher accuracy (for earlier years, HistPat assesses its accuracy as 'high' in less than half of its records), mainly due to the fact that locations in the yearbooks are more clearly separated from other entities in a comparatively short text string, whereas HistPat parses a greater body of text – the entire patent documents – to find locations. Finally, we compare team sizes in our dataset to those in Van der Wouden (2020).[7] In a randomly drawn sample of patents where our and Van der Wouden's assessment of team size diverge, manual checks show that in 50 out of 50 cases our data are correct. Against state of the art datasets, our data can therefore be regarded as exceptionally complete and accurate.

[6]A sample page of these yearbooks is provided in Fig. A1A.

[7]The comparison was suggested by an anonymous reviewer. We are indebted to Frank van der Wouden for generously sharing information on patent-level team sizes in his data and a manually created ground truth dataset. The latter dataset suggests that our team size variable is 99.8% accurate.

3.2 Technology classes

Each patent application receives a set of technology codes from the United States Patent Classification (USPC classes). These codes are used to assign applications to so-called "art units" within the USPTO where they help patent examiners search the prior art (Righi and Simcoe, 2019). Importantly, a patent's USPC classes are determined by outside contractors, not by the inventors or patent examiners themselves.

At the highest level of aggregation, the USPC classification consists of 3-digit classes. In our datasets, we identify 474 such classes. Classes are further subdivided into about 150,000 unique subclasses.[8] We refer to these subclasses as "6-digit" codes, even though some subclasses may contain more or fewer digits. The USPTO regularly expands and modernizes the USPC system to reflect changes in technology. When it does so, it retroactively reclassifies all patents. This ensures that technology codes are harmonized across the entire period of analysis.

We use these technology codes for three purposes. First, following Hall et al. (2001) and Marco et al. (2015), we group patents into six broad technological sectors, based on a patent's primary technology code.[9] Second, we use the combination of primary and secondary classes (without making a distinction between the two) as a high-level description of the invention's content.

Third, we determine the "vintage" of technologies used in a patent, asking in which year each technology class reached a cumulative 1% of all patents having listed this class by 2015. Because the number of patents increases rapidly over time, we use a weighted cumulative count that weights each patent by the inverse of the total number of patent grants in a year, such that each year has equal weight in the determination of technological vintages. Tables 1 and 2 show the five most and least recent technology classes in terms of this vintage.

3.3 Demographic information of inventors

To learn more about the inventors on a patent, we merge them to US census records. To do so, we make use of the full-count non-anonymized census records between 1850 and 1940 provided by IPUMS (Ruggles et al., 2021).

[8]Classes tend to distinguish among technologies, whereas subclasses "delineate processes, structural features, and functional features of the subject matter encompassed within the scope of a class." `https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf`, p I.1.

[9]These categories are: *Mechanical*; *Chemical*; *Electrical & Electronic*; *Computers & Communication*; *Drugs & Medical*; and *Other* technologies. The primary class on which they are based "[...] is indicative of the invention as a whole or the main inventive concept using the claims as a guide." See `https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf`, p I.5.

Table 1: Vintage technology classes (3-digit)

code	**3-digit technology**	**vintage**
532	Organic compounds – part of the class 532-570 series	2001
726	Information security	1997
506	Combinatorial chemistry	1992
709	Multicomputer data transferring	1992
717	Data processing, software dev.	1992
295	Railway wheels and axles	1836
190	Trunks and hand-carried luggage	1836
142	Wood turning	1836
2	Apparel	1836
384	Bearings	1836

Table 2: Vintage technology classes (6-digit)

code	**6-digit technology**	**vintage**
359/200.3	Optical: systems and elements – Grooved shaft	2013
351/159.36	Optics – Means to limit movement	2012
185/41C	Motors: spring, weight, or animal powered – Centrifugal	2012
705/339	Data processing – Central recipient pick	2011
348/287	Television – Conductive grid at target	2011
295/4	Railway wheels and axles – Rack rail	1836
126/506	Stoves and furnaces – With food cooker	1836
144/36	Woodworking – Planing and matching	1836
408/71	Cutting by rotating axially moving tool – Rotary, work	1836
144/69	Woodworking – Auger cutter	1836

3.3.1 Census matching

We proceed in three steps. First, we find for each inventor a set of candidate matches in the US census, based on similarity in last names. Next, we generate a set of distances between the inventor and these candidate matches, including string distances for first names, initials and last names and kilometer distances between an inventor's place of residence as listed on the patent and the place of residence of each match candidate in the census.

Second, we create a ground truth dataset, by matching a small number of inventors by hand whose patents are listed on Wikidata. The additional information this yields on inventors' places of birth, dates of birth and names of family members typically makes it trivial to identify correct matches.

Third, we train two xgboost algorithms on this ground truth sample. The first produces match plausibility scores, based on name and place of residence information. The second adjusts these scores to differentiate between inventors with multiple plausible matches, and those that have a unique plausible match, downgrading the former relative to the latter. We repeat these steps for match candidates in the two census waves closest to the patent's grant date and then select the overall best match.[10] In the analyses below, we rely on a sample of high-confidence matches. The exact procedure and the out-of-sample performance of the xgboost models are described in Appendix A.3.

3.3.2 Gender, family ties and ethnicity

The census records allow us to study how socio-demographic characteristics of inventors change over time. Of particular interest are an inventor's occupation, age and place of birth. In principle, the census also records an individual's gender. However, we opt to infer the most likely gender from inventors' first names. This allows us to analyze gender dynamics also for inventors that could not be matched to census records, including inventors of patents granted after 1945.

Similarly, although census records contain family ties, we instead infer such ties from inventors' sharing the same last name. This may result in some false positives, especially for inventors with common last names. To investigate this, we construct null models in which last names are shuffled across patents.[11] This exercise suggests that the likelihood of two inventors' sharing the same last name by chance is negligible in our main period of interest.[12] At the same time, not all relatives share the same last name. Hereafter, "family

[10]For instance, for a patent granted in 1902, we try to find matches in the 1900 and 1910 censuses.

[11]That is, we estimate how often two inventors would share the same last name, had they been allocated to patents at random. To do so, we shuffle inventors' last names within a year and within groups of inventors whose last names have the same geographic origin. This ensures that each patent retains the same number of inventors, while allowing for higher copatenting rates among members of the same ethnic community (Almeida et al., 2015). This procedure is described in detail in Appendix B.2.

[12]The share of spurious family ties becomes significant only in the second half of the 20th century, with

ties" will therefore refer to family relations in which last names are typically shared (e.g., married couples, brothers, father and son, etc.).[13]

For the period after 1945, we have no country-of-birth information for inventors. Therefore, we proxy ethnic backgrounds using algorithms trained on Wikipedia and recent census data to infer the origins of last names (for details, see Appendix B.6). We focus on Hispanic and East-Asian surnames, because these are relatively easy to identify with such algorithms.

3.4 Organizational context

The patent yearbooks provide some information on the organizational context in which an invention was developed. In particular, we know whether the patent's intellectual property rights were transferred to an *assignee* or retained by the inventors at the moment the patent was issued. Using named-entity recognition, we distinguish between patents assigned to individuals and those assigned to organizations. It is not straightforward to differentiate between different types of organizations based on the information at hand. However, a closer inspection of assignee names suggests that before 1950 the vast majority of organizational assignees were firms. Therefore, in our analysis, we refer to all organizational assignees as firms. In robustness checks, we drop patents for which we could establish that they were assigned to non-firm organizations, such as governments, militaries or universities.

To gain additional information on how firms organize their inventive activity, we turn to a series of surveys conducted by the National Research Council: the *Industrial Research Laboratories of the United States* reports. To be included in these surveys, firms had to operate a dedicated "laboratory" with "separate and permanently established research staff and equipment", excluding "firms that indicated they only occasionally carry out research, using teams temporarily recruited for the purpose or assembled from their operating staffs" (p. 2 National Research Council, 1956, see also Furman and MacGarvie (2007)).

Separate reports of this survey were published in 1920, 1927, 1931, 1933, 1938, 1940, 1946, 1948, 1950 and 1956. Each report contains the name of each lab, a short description of its main activity, the lab's city or address, the (managing) directors and important researchers and, in some editions, the lab's major equipment and number of employees (Fig. A1C). Furthermore, the 1940 and 1946 editions also record founding years of labs. We will use these surveys to assess whether a patent assignee operated a research lab at the time the patent was granted.

---

the rise of, for instance, Indian and East-Asian surnames in the patent records (see Appendix B.2).

[13]Relying instead on family relations derived from the census is problematic in its own right. Census records only list family relations within a household. Therefore, identifying the full set of family relations not only requires matching multiple inventors on a patent to census records, but also matching different census waves to each other. Incomplete matches in either step reduce the number of identifiable family relations drastically and in possibly biased ways.

To do so, we obtain scanned versions of these reports from the *Hathi Trust Foundation*. Next, we use OCR to digitize their contents and subsequently apply named-entity recognition to extract the name of each lab. We then match labs to assignees on patents based on string similarity between lab and assignee names. This yields a total of 2,504 assignees for which we can establish that they operated an industrial research lab.

In the remainder, we will refer to patents as "firm-based" if they are assigned to organizations. Furthermore, if the organizational assignee operated a known industrial research lab, we refer to the patent as "lab-based". All other patents, i.e., patents that are unassigned or assigned to individuals, are called "standalone" patents.

It is important to note that the reason why inventors assigned their patents to firms changes throughout the 19th and 20th century. For instance, whereas 19th century inventors often had no employment relation with their assignees, selling intellectual property rights to partners that provided capital or who were better positioned to commercialize the technology, 20th century inventors were typically employed by the assignee, often in corporate research units (Lamoreaux and Sokoloff, 2001). Moreover, especially in this later period, the above definitions are likely to underestimate the number of patents in which firms or labs helped coordinate the innovation process. First, the definitions do not take into account that inventors may not always assign a patent to their employers. The opposite is in principle also possible: firms may commission (a team of) inventors, without employing them, a practice mostly associated with 19th century innovation. Second, to identify lab-based patents, we have to be able to match research labs to patent assignees. However, our surveys do not represent a complete census of research labs, but focus on the most prominent labs in the US. Moreover, we may fail to match assignees to their research labs whenever patents are assigned to entities with names that are sufficiently different from their firm's research lab.[14] To mitigate some of these problems, we manually checked whether we managed to identify correct matches for the largest assignees in our patent records, as well as for the most prominent research labs of this period.

3.5 Sample restrictions

To match patents to technology codes, we rely on patent numbers. Because these numbers are only listed in the patent yearbooks since 1856, our analysis starts in that year. Furthermore, we focus on the period 1856-1945, for which we can merge demographic information from the census, as well as information on labs from the industrial research lab surveys.

Because we can only match inventors to the US census, we drop patents for which not all inventors reside in the US and patents between 1969 and 1975, for which we lack country of residence information. Furthermore, we require that patent numbers and grant dates allow us to identify a patent's technology classes.

[14] Of a total of 7,990 research labs, 2,504 had matching assignees in the patent records.

When analyzing census-derived variables, we rely on a sample of well-matched inventors, aged 16 and older. For questions that involve inferred genders, we limit the sample to inventors whose gender can be inferred from their first names with high confidence (see Appendix B.5). For all other analyses, we impose no further restrictions.

Finally, we refrain from disambiguating inventors across patents and leave this as a task for future research. Instead, we create a sample of disambiguated *co-inventor dyads.* This is easier than disambiguating inventor names, because it is highly unlikely that we observe the same *pair* of last names on multiple patents by chance. An exception are pairs of inventors with common last names. Furthermore, if co-inventors are related, their last names are not independent. To minimize spurious matches, we drop all inventor dyads where one of the two last names is found in over 500,000 of the 650M census records,[15] or, for same-name dyads, where the shared last name is found in over 50,000 records. Details are provided in Appendix B.3.

3.6 Time windows

Because the number of granted patents grows at a roughly exponential rate, data are much sparser for earlier than for later periods. To strike a balance between temporal resolution and precision of estimates, we create windows, based, not on calendar time, but on temporal rank. To do so, we sort all patents by their grant dates and divide observations into groups of identical size, each containing N_w observations. Each window is then associated with the average grant year of the patents it spans. This allows us to plot timeseries of point estimates with standard errors that are roughly constant across windows. Note that we can choose N_w for each time series separately, which allows different precision/resolution trade-offs within a single graph.

4 Results

We organize our results from the micro to macro level. That is, we first describe changes at the patent level, then at the level of inventors, teams and organizations and end with a description of changes at the level of the innovation system as a whole.

4.1 Novelty

To analyze changes in the characteristics of patented inventions over the course of the 19th and 20th century, we focus on their novelty. The novelty of patents has been assessed in various ways, ranging from qualitative assessments by experts to analyses of forward and backward citations (Verhoeven et al., 2016). However, these approaches either scale poorly – as in the case of expert assessment – or require data that are not available for the

[15]Note that we establish how common a last name is using census, not patent records. To do so, we pool census records across all available waves.

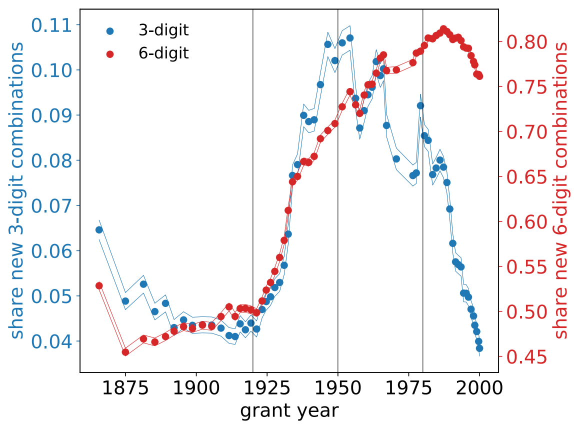


Figure 1: Novelty. Share of patents that list new combinations of 3-digit (blue) or 6-digit (red) technology classes.

period we analyze – as in the case of patent citations, which only start being used in the mid 1940s and truly widespread in the 1980s. Therefore, we infer how novel an invention is from a patent's technology classes. In particular, we ask whether a combination of technology classes had already been reported on an earlier patent, as originally proposed by Fleming (2001, see also, Clancy (2018); Pezzoni et al. (2022); Strumsky and Lobo (2015)).[16]

At the 6-digit level, new combinations of technology classes often represent incremental innovations, whereas new combinations of 3-digit classes should, on average, mark more radical departures from the existing state of the art. While this method provides an admittedly limited assessment of novelty, it is objective, because technology codes are determined by independent contractors, not by the inventors or other interested parties. Moreover, when we compare our measures with Kelly et al.'s (2021) measure of breakthrough patents, we find that breakthrough patents are 84% more likely to list novel combinations of technologies at the 3-digit level and 66% at 6-digit level. This shows that our novelty indicators important aspects of how much patents depart from the state of the art.[17]

Fig. 1 shows how the share of patents that we identify as novel changes over time. After falling in the late 19th and early 20th century, the share of patents that list new technological combinations increases rapidly between 1920 and 1950, a period thatAkcigit

[16]We consider a combination as new, even if it had already been used as a subset of technology classes on earlier patents. For instance, if patent B is the first to list codes 142 ("Wood turning") and 173 ("Tool driving or impacting"), we regard this as a new combination of codes, even if an earlier patent A already listed the codes 142 and 173, but also listed 147 ("Coopering").

[17]Kelly et al.'s breakthrough measure is based on an analysis of patent text. Breakthrough patents are defined as patents that are very dissimilar from prior patents but similar to future patents.

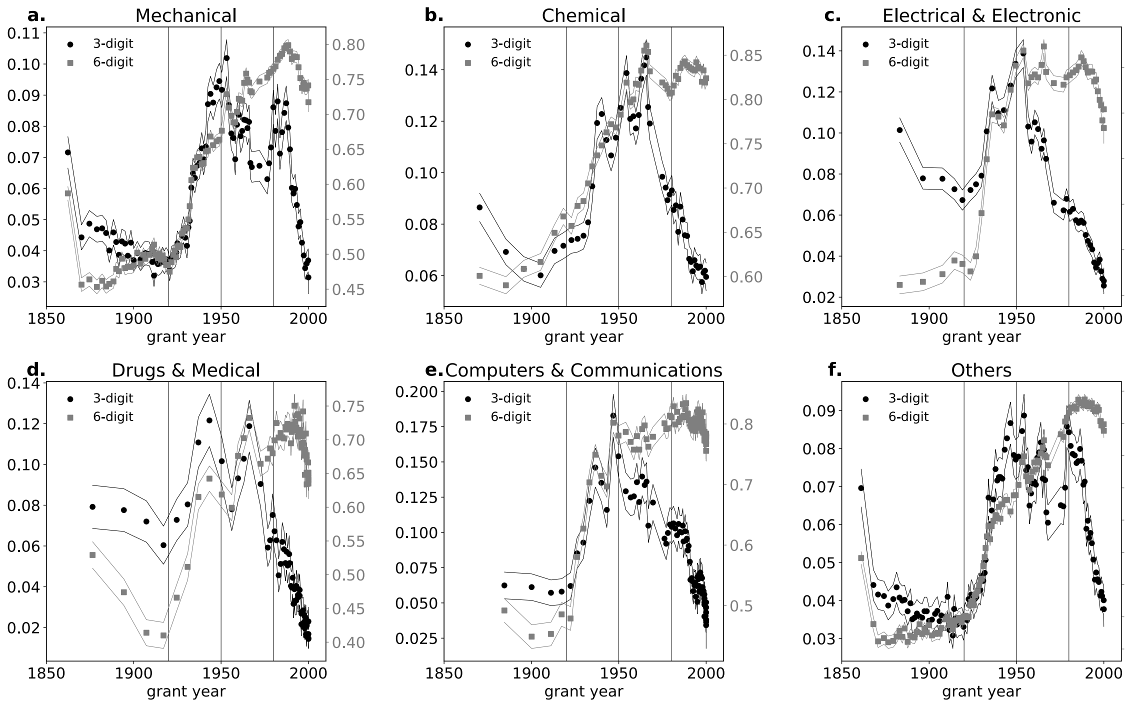


Figure 2: New combinations by broad technology class. Each graph shows the share of new 3-digit (black) and 6-digit (gray) technologies by the six broad primary classes defined in Hall et al. (2001).

et al. (2017) refer to as a "golden age of innovation" that witnesses the "rise of American ingenuity". Moreover, the rapid rise in novel combinations is not limited to a specific technological area, but visible in all six of Hall et al.'s (2001) broad technological sectors (Fig. 2).[18] However, this rise is more pronounced in newer, science-based technological areas – such as communication technologies, electronics and chemistry – than in more mature fields such as mechanical technologies and the category of "other" technologies, which includes agricultural, textiles and woodworking technologies.

Interestingly, the likelihood that patents list radically novel technological combinations starts falling again in the 1950s, whereas novelty that includes more incremental change keeps rising until at least the 1980s. We return to this observation in section 4.6.

[18]The patterns in Fig. 1 could potentially be driven by a small number of specific technologies. However, they do not change in a qualitative sense when we control for technology fixed effects.

4.2 Inventors

Between 1856 and 1945, the average age of inventors rises steadily (see Appendix B.4). However, their occupational backgrounds and the relation between an inventor's age and the nature of patented inventions changes abruptly in the 20th century.

4.2.1 Learning curves

We interpret the curves that track how inventions change with the age of their inventors as "learning curves". We estimate these learning curves for four different cohorts – inventors born between 1840 and 1859, 1860 and 1879, 1880 and 1899 and 1900 and 1919 – fitting the following regression equation:

$$y_{p(i)} = \alpha_{t(p(i))} + \sum_{c\in\mathcal{C}} \gamma_c C_{c(i)} + \sum_{c\in\mathcal{C}} \beta_c A_{it(p(i))} C_{c(i)} + \varepsilon_{p(i)}, \tag{1}$$

where $y_{p(i)}$ can be one of three variables: a dummy describing whether or not inventor i's patent $p(i)$ lists a novel combination of technologies, the number of distinct technology codes listed on patent $p(i)$, or the technological vintage of the "youngest" technology listed on the patent. As regressors, we include year fixed effects, α_t, cohort fixed effects, γ_c, and interactions of the age, $A_{it(p(i))}$, of inventor i in the year the patent was granted, $t(p(i))$, with cohort dummies, $C_{c(i)} \in \mathcal{C}$, where $\mathcal{C}$ is the set of cohorts.[19] To ensure that our sample covers the same age range for all cohorts, we restrict the analysis to inventors aged 15 to 35. This means that we study, for instance, patents of the 1880-1899 birth cohort between 1895 and 1935.

Table 3: Learning curves (3 digits)

Cohort	**Novelty**	**Vintage**	**Complexity**
1840	-0.0011 (0.0006)	-0.0078 (0.0274)	-0.0017 (0.0017)
1860	0.0001 (0.0002)	-0.0067 (0.0114)	0.0007 (0.0007)
1880	0.0003 (0.0002)	0.0293 (0.0125)*	0.0018 (0.0006)**
1900	0.0000 (0.0003)	0.2591 (0.0204)***	0.0024 (0.0009)**

Parameter estimates for age effects in equation (1). Cohorts are listed in rows, dependent variables in columns – *Novelty*: dummy variable for whether or not the patent lists a new combination of technologies; *Vintage*: vintage of most recent technology class listed on the patent; *Complexity*: number of distinct technology classes listed on the patent, where technology classes are recorded at the 3-digit level. Robust standard errors are in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

Results are shown in Table 3. Cohorts that enter the labor market in the 20th century display markedly different learning curves compared to earlier cohorts when it comes to

[19] Note that, because we do not disambiguate inventors, learning curves are estimated for entire cohorts, not individuals.

the vintage of the technologies they use and the number of technologies they combine. Whereas the first two cohorts' learning curves are flat, for the 1880-1899 and 1900-1919 cohorts, the number and vintage of technologies increases with age.[20] Moreover, these learning curves are substantially steeper for patents assigned to firms and industrial research labs (see Appendix D.1). In contrast, we find little evidence that novelty changes with inventors' age, regardless of their cohort.[21] This suggests that whereas the 20th century cohorts learn to use more and more recent technologies as they grow older, they do not become more "creative" with age in the sense of patenting novel technological combinations.[22]

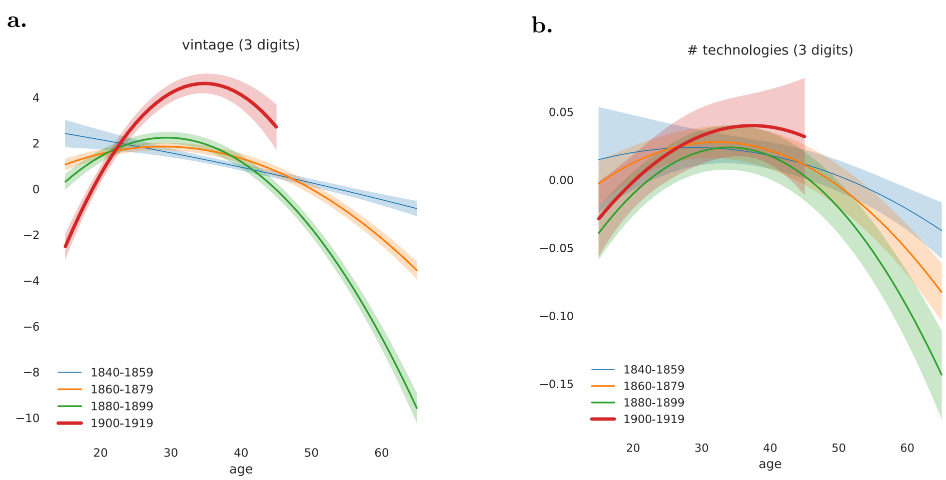


Figure 3: Learning curves. Within-sample fit of learning curves for four different cohorts: 1840-1959 (blue), 1860-1879 (orange), 1880-1899 (green), 1900-1919 (red). The left panel shows learning curves in terms of technological vintage, the right panel in terms of the number of technologies on a patent. The shaded areas represent 95% confidence intervals, using robust standard errors.

To further analyze the relations between novelty and complexity on the one hand and age on the other, Fig. 3 repeats the analysis for these variables, widening the age range to 15-65 and adding quadratic terms to allow for nonlinearities. This corroborates the findings in Table 3. In the earliest cohorts, at young ages, inventors tend to use more, and more modern, technologies than their older contemporaries. However, as they grow

[20]These statements should be taken relative to other contemporaneous cohorts. Overall, vintage and number of technologies increase steadily over time, regardless of the cohort. However, by adding year fixed effects, we remove such secular trends, comparing individuals across cohorts in the same year.

[21]Note that when we use 6-digit technology codes (see Appendix D.1, Table D1), for the two most recent cohorts, incremental novelty increases with age in a statistically significant way. However, these effects are small in an economic sense.

[22]Furthermore, in Appendix D.1, we show that these findings are robust to adding control variables for the occupational background of inventors and the organizational context in which the patent was created.

older, inventors in these cohorts fall behind younger inventors. In the 1880-1899, and even more so in the 1900-1919 cohort, this changes drastically. Entering the labor market in the 20th century, these inventors start their careers using fewer technologies and less recent vintages than the older cohorts they coincide with. However, as they grow older, they overtake older cohorts and peak in their thirties. This suggests that, after 1900, and even more so, in the 1920s, inventors need more time to become acquainted with modern technologies and to combine them in larger – presumably more complex – combinations.

4.2.2 Occupations

The 1920s also mark an abrupt change in the occupational backgrounds of inventors. Whereas 19th century invention was still dominated by blue collar workers and mechanics, in the 1920s, a new type of inventor emerges: the engineer. Engineers quickly come to dominate US invention. By the 1940s they have become responsible for 25% of patents, while representing only 0.7% of the US labor force (Fig. 4a and 4b).[23]

The rise of engineers coincides with the fledgling start of academic patenting. Fig. 4c shows the share of patents granted to inventors who are identified as professors in the census. Interestingly, these "academic patents" often represent early examples of university-industry collaboration. Of the 1,462 patents granted to professors between 1900 and 1945, 49% were assigned to organizations. However, in 91% of such cases, the organizations were firms. In contrast, only 9% of academic patents had been assigned to universities. In fact, universities only meaningfully start claiming ownership of patents in the 1940s (Fig. 4d, see also Arora et al., 2021). Moreover, these first university patents were not, as in later decades (e.g., Henderson et al., 1998), predominantly held by the largest research universities, but by universities in the American Rustbelt (see Table 4). Finally, the sudden shift in the 1920s to science-based innovation is also directly visible in inventors' reliance on academic literature, as using data from Marx and Fuegi (2020, 2022). Fig. 4e shows how the share of patents that cite academic papers rises from below 0.1% in the 19th century to over 2% in the 1940s.

4.3 Teamwork

With the shift to engineering occupations, we also witness the start of the steady rise in teamwork that would persist throughout the 20th century (see, for instance, Wuchty

[23] Because census occupations are self-reported, the observed rise of engineers may to some extent be artificial, merely reflecting a change in nomenclature or occupational identity, not a change in the actual occupational backgrounds of inventors. We explore this possibility in Appendix B.1, where we use census data to find potential precursors to engineering occupations. To do so, we look at how individuals change occupations across census waves. This yields a "skill-relatedness" network (Neffke and Henning, 2013; Neffke et al., 2017) that connects occupations between which exceptionally many individuals transition compared to a random benchmark. Adding occupations that are skill related to engineering occupations does not change the pattern observed in Fig. 4a (see Fig. B1).

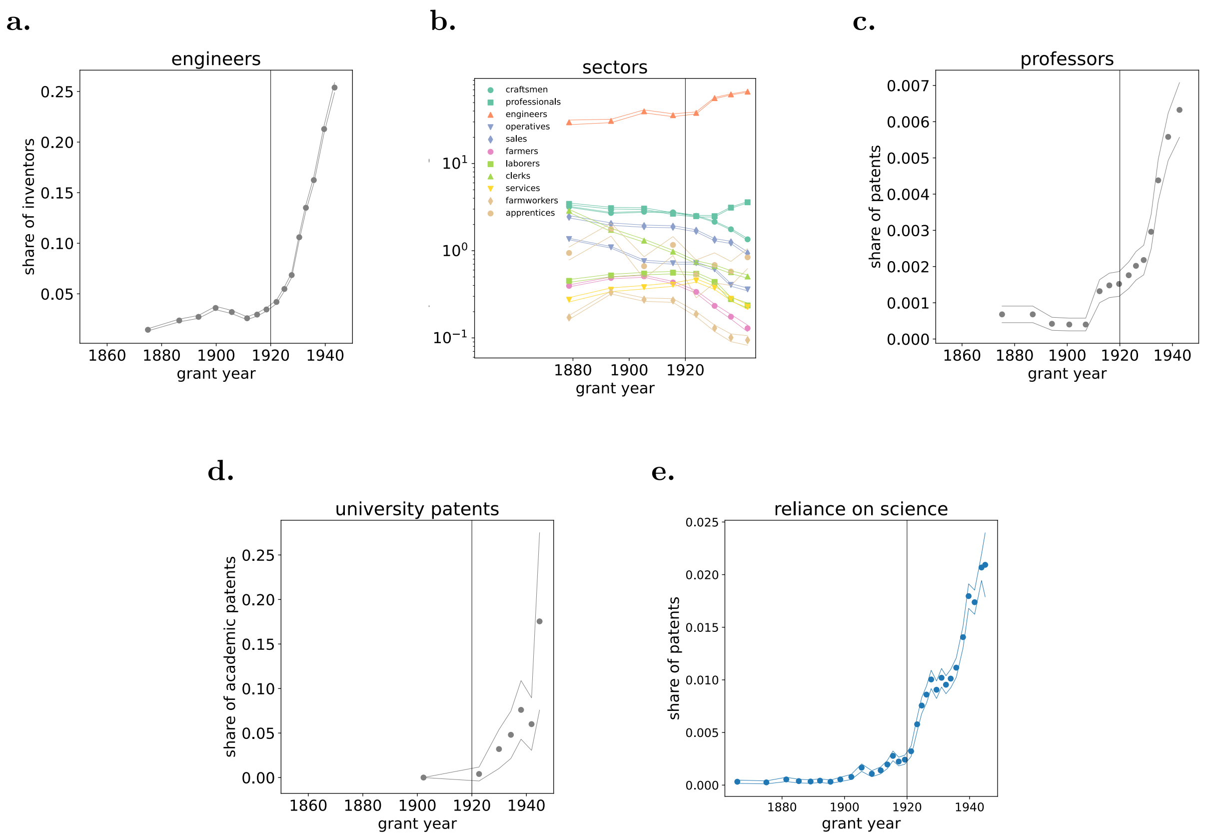


Figure 4: Labor inputs in invention. **a**: Share of engineers among inventors. **b**: Over-representation of occupational sectors among inventors vis-à-vis the US population: $\sigma^{p}_{ot}/\sigma^{pop}_{ot}$, where σ^{p}_{ot} is the share of patents held by inventors with occupation o in year t and σ^{pop}_{ot} the share of workers in the US population with occupation o in year t, interpolating between census waves in non-census years **c**: Share of inventors that list professor as their occupation ("academic patents"). **d**: Share of academic patents assigned to universities. **e**: Share of patents that include references to academic literature (Marx and Fuegi, 2020, 2022). Y-axis: share of patents. Time period: until 1945. All shares are calculated in samples where inventors could be matched with high confidence to census records. Lines display 95% confidence intervals.

Table 4: Share of academic patents by university (1900-1945)

University	**share**
Purdue University	0.206
Iowa State University	0.114
University of Wisconsin-Madison	0.063
Stanford University	0.057
Dartmouth College	0.048
University of Minnesota	0.044
University of Illinois	0.038
University of Tennessee	0.032
University of Michigan	0.025
Ohio State University	0.022

Universities' shares of patents that were (1) granted to inventors identified as professors in the US census and (2) assigned to a university between 1900 and 1945.

et al., 2007, for the period 1975-1995). Fig. 5 shows that this process was not gradual, but took off abruptly in the early 1920s. Until then, the share of patents granted to teams had, if anything, been slightly decreasing.

Not only the prevalence of teams changed, but also their nature. To show this, we create a dataset of coinventor dyads, i.e., of all pairwise combinations of inventors listed on a patent.[24] Table 5 shows the most frequent combinations of occupations in these inventor dyads.[25] Whereas 19th century collaborations tended to take place between skilled and less skilled roles – such as between skilled craftsmen and operators or between operators and common laborers – in the 1920s, collaboration shifts to teams of similarly, highly skilled professionals, yielding dyads of, for instance, two engineers, a chemist and a chemical engineer or a draftsman and a mechanical engineer.

Fig. 6 corroborates that teams become increasingly homogeneous in the early decades of the 20th century. First, the age difference between team members shrinks (Fig. 6a). Second, team members increasingly share similar expertise as evident in the share of dyads in which inventors list the same occupation (Fig. 6b) and in the increase in skill relatedness between a dyad's occupations (Fig. 6c).

[24]This analysis is restricted to inventor dyads where we can match both inventors with high confidence to a census record. Note that in the period we study, the vast majority of team patents, over 90%, are filed by teams of two inventors, with another 6% filed by teams of 3 inventors. Most teams are therefore represented by a single dyad.

[25]We exclude same-occupation pairs, as well as occupation pairs that involve farmers or managers, because such pairs are less informative about the expertise that is combined in teams: farmers represent a very large share of 19th century employment and management occupations tell us little about the technical skills of team members.

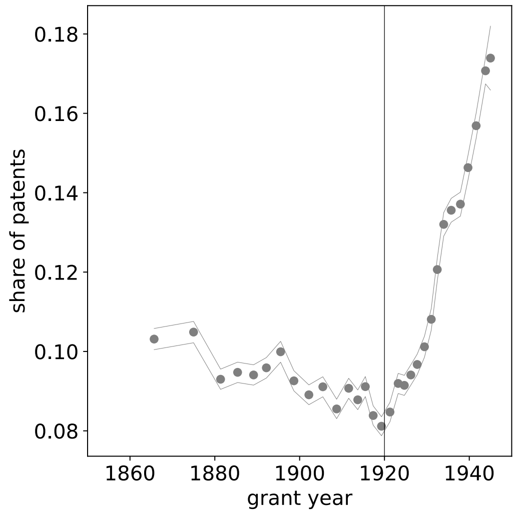


Figure 5: Teams. Share of patents granted to teams of inventors. Lines display 95% confidence intervals.

Table 5: Most common occupational pairs in coinventor dyads

occupation 1	**occupation 2**	**frequency**
1856-1900		
Machinists	Operative and kindred workers (n.e.c.)	65
Operative and kindred workers (n.e.c.)	Salesmen and sales clerks (n.e.c.)	30
Laborers (n.e.c.)	Operative and kindred workers (n.e.c.)	27
Operative and kindred workers (n.e.c.)	Carpenters	27
Laborers (n.e.c.)	Machinists	17
1920-1945		
Stationary engineers	Engineers, mechanical	110
Engineers, electrical	Engineers, mechanical	106
Draftsmen	Engineers, mechanical	98
Professional, technical and kindred workers (n.e.c.)	Chemists	91
Chemists	Engineers, chemical	66

Five most common occupation pairs in coinventor dyads, excluding same-occupation pairs (1856-1900: 57; 1920-1945: 74) and occupation pairs with missing occupations, managerial occupations (1856-1900: 126; 1920-1945: 160) or occupations in farming (1856-1900: 153; 1920-1945: 108). N.e.c.: not elsewhere classified.

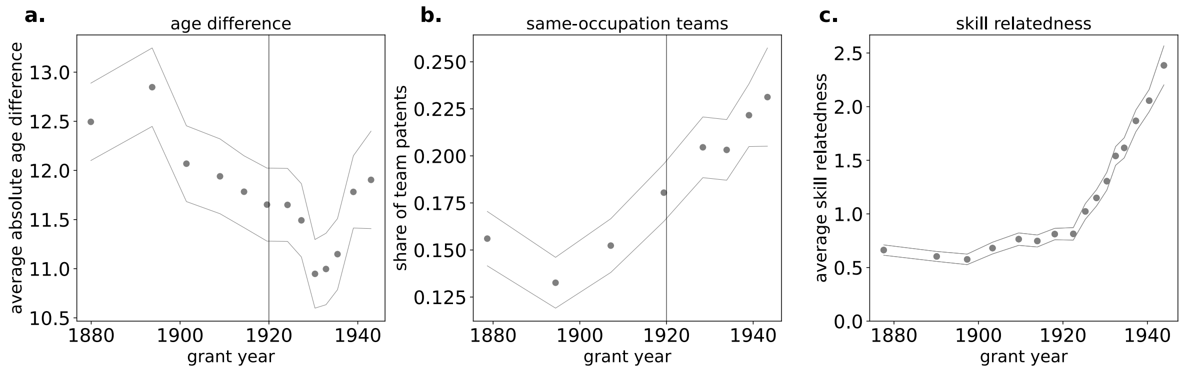


Figure 6: Team composition. Samples are restricted to inventor dyads for which both inventors can be matched to the census. **a**: Average age difference in coinventor dyads. **b**: Share of coinventor dyads in which both inventors record the same occupation in the census records. **c**: Average skill relatedness between the occupations in a dyad. Values larger than one signify that occupations are skill related: labor flows between the occupations exceed their random benchmark (see Appendix B.1). Values below one signify that occupations are unrelated: labor flows fall short of the random benchmark.

4.4 Corporate research and the rise of teamwork

What is driving these changes in US invention? We show that a key role is played by the emergence of organized corporate research, with the industrial research lab as its embodiment. These labs were responsible for much of the increased role of engineers: by 1940, 40% of the patents of engineers came out of industrial research labs (Fig. 7). More importantly, these labs were key facilitators of teamwork. To support this claim, we show that:

1. the rise in teamwork between 1920 and 1945 is almost completely due to teamwork in firms, especially in those with labs;
2. labs are associated with more frequent repeat-collaborations, fewer family ties, and higher levels of skill relatedness among team members;
3. labs allow teams to collaborate over longer distances; and
4. teams are more likely to patent radical innovations, but only if they work for firms, and especially for firms with research labs.

The rise of teamwork. In support of the first claim, Fig. 8a shows that today's widespread practice of team-based invention was originally supported by a remarkable

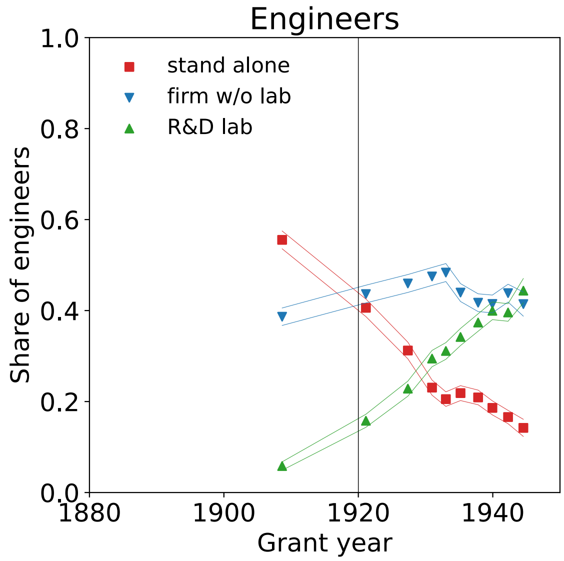


Figure 7: Workplaces of engineers. Share of inventors with engineering occupations who work as standalone inventors (red), for firms without research labs (blue) or for firms with research labs (green).

early-20th century organizational innovation: the industrial research lab. Whereas in the 19th century, teamwork often relied on family ties, firm-based teamwork gained prominence only in the 20th century. In the 1920s, team-based invention then shifted to industrial research labs. These labs quickly diffused and became dominant: in the 1940s over 40% of team patents came from firms that operated research labs.[26] What is more, lab-based inventors were 20% more likely to work in teams than inventors in firms without known research labs and over two times as likely as standalone inventors. In fact, Fig. 8b shows that the sudden rise in teamwork in the 1920s was wholly driven by increased teamwork in corporate research: whereas, after 1920, firm- and, especially, lab-based patents increasingly relied on teamwork, the prevalence of teamwork in standalone patents barely changed.

Repeat-collaborations. Firm- and lab-based teams were also more likely to collaborate repeatedly than standalone teams. To show this, Fig. 9a plots the likelihood that two inventors copatent more than once in the same decade, using the disambiguated inventor dyads described in section 3.5. Repeat-collaborations are almost twice as likely for team

[26]Note that this is likely an undercount, given that we will not have been able to identify all firms with research labs in our dataset. This is in part due to the incompleteness of the survey but also because firms without formal labs may have introduced research departments with lab-like characteristics.

a. **b.**

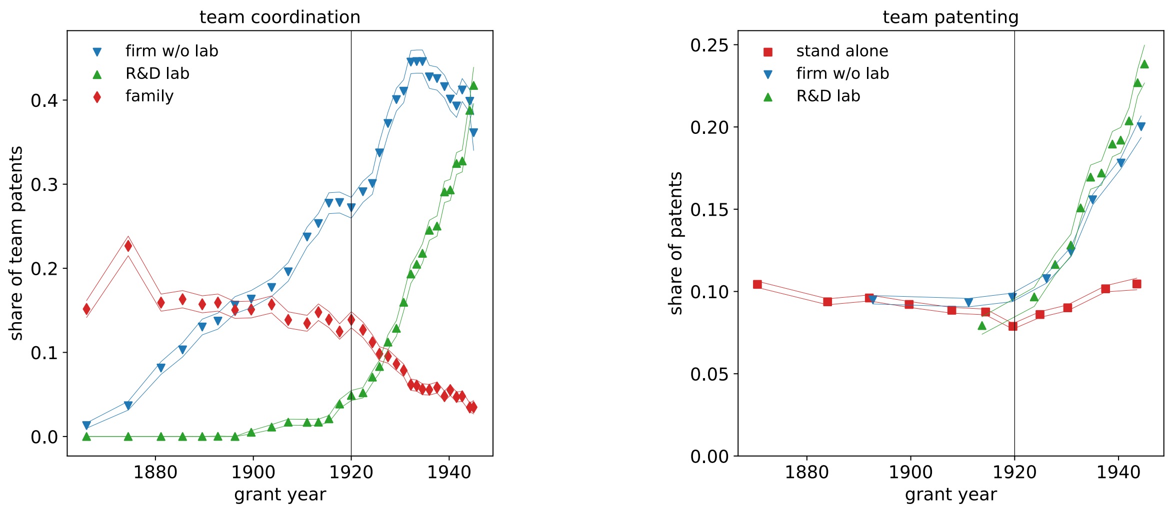


Figure 8: Team coordination. **a**: Share of team patents that are coordinated by firms without labs (blue), firms with labs (green) or family ties (red). **b**: Share of patents that are team patents.

patents assigned to firms or labs than for patents produced by standalone teams.

Family ties and skill relatedness. Fig. 9b shows that co-inventors in labs and firms were much less likely to be related (i.e., share the same last name) than members of standalone teams. This suggests that the shared organizational context in firms and labs was able to substitute for the trust embedded in family ties. Moreover, from 1920 on, skill relatedness between co-inventors grows rapidly, but much faster in firms and R&D labs than in standalone teams, suggesting that leveraging the new organizational ties resulted in better skill matches.

Long-distance collaboration. To show that firms and labs support long-distance collaboration, we count the number of collaborations between any given pair of US cities. Next, we analyze the spatial decay in collaboration, estimating gravity models with city of origin, o, and city of destination, d, fixed effects. To do so, we use Pseudo-Poisson Maximum Likelihood (PPML, Silva and Tenreyro, 2006) to estimate:

$$\mathbb{E}\left[C_{od}|d_{od}\right] = \exp\left[\gamma_o + \eta_d + \delta d_{od}\right], \tag{2}$$

where γ_o and η_d are origin and destination fixed effects and d_{od} is the logarithm of the Haversine distance between o and d. We estimate this model separately for standalone

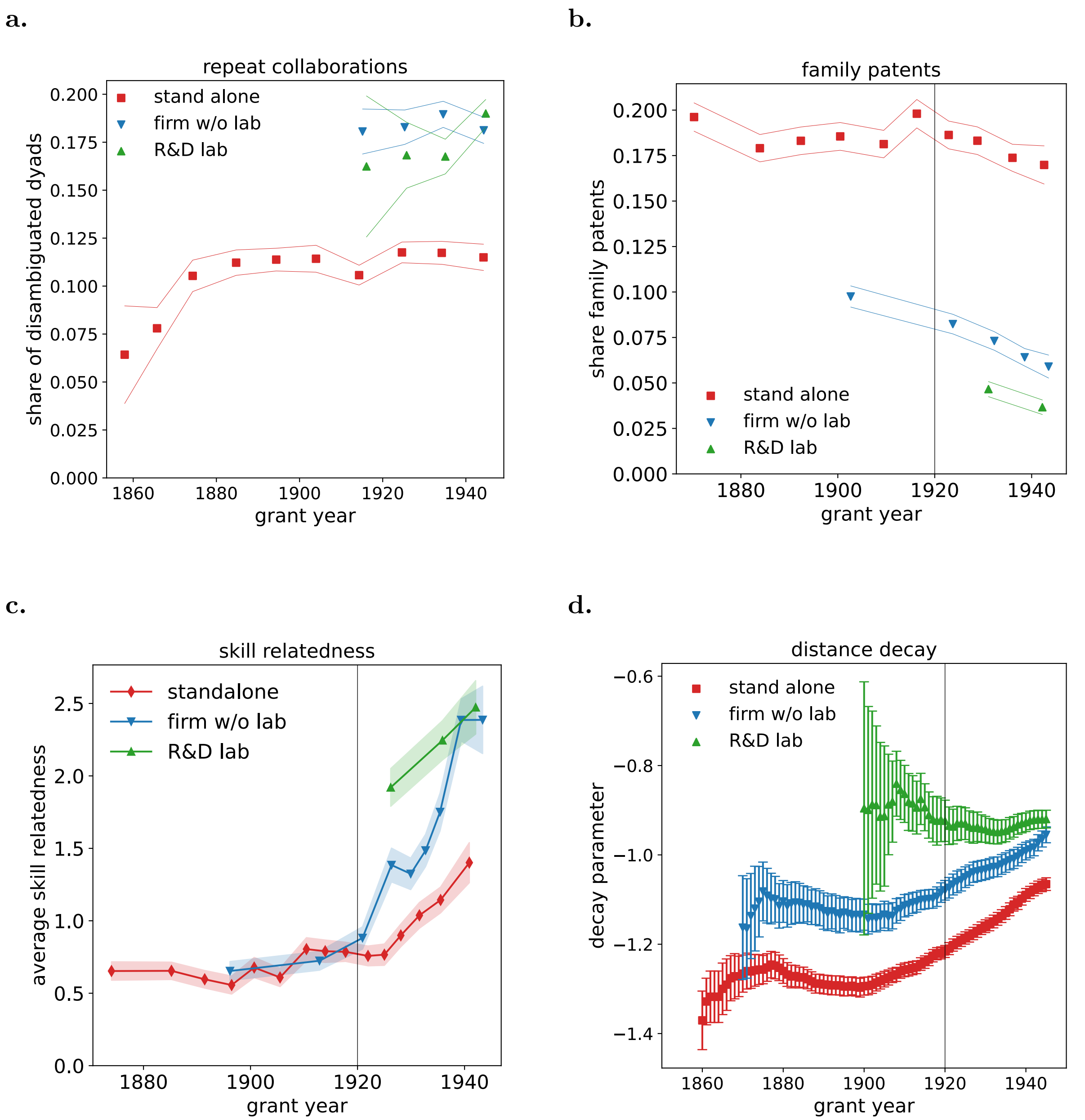


Figure 9: Team facilitation. **a**: Stability of inventor dyads: likelihood of repeated collaboration in a decade (i.e., that two inventors co-patent more than once in the same decade), using the disambiguated inventor dyads of section 3.5. **b**: Family patents: share of dyads where inventors share the same last name. **c**: Average skill relatedness between co-inventors (see Appendix B.1). **d**: Estimated distance decay parameters for collaborations between US cities. Vertical lines display 95% confidence intervals. Colors: team patents coordinated by firms without labs (blue), firms with labs (green) or family ties (red).

teams, teams in firms without research labs and teams in industrial research labs. We do so repeatedly over a moving time window that covers 10 years before and after the year reported on the horizontal axis. The estimated distance decay parameter is uniformly negative, showing that collaboration is constrained by distance. However, distance is least constraining for lab-based and most constraining for standalone teams.

Novelty. Do research labs also generate more novel inventions? We study this by analyzing what drives the changes in novelty of Fig. 1, loosely following a knowledge production function approach (Pakes and Griliches, 1980). In particular, we ask whether the new labor inputs that emerge in the 20th century are associated with more novel technological combinations and to what extent this depends on the organizational context in which these inputs are used.

To do so, we proxy labor inputs with dummies that capture whether or not one or more of the patent's inventors are engineers and whether the invention was granted to a single inventor or to a team of inventors. We then interact these inputs with dummies that code the patent's assignee type, distinguishing between standalone patents, firms, and firms with industrial research labs. Next, we fit the following linear probability models:

$$y_p^d = \alpha_{t(p)} + \beta_e E_p + \beta_t T_p + \sum_{o \in f,l} \beta_o O_{op} + \sum_{o \in f,l} \beta_{o\times E} O_{op} E_p + \sum_{o \in f,l} \beta_{o \times T} O_{op} T_p + \varepsilon_p, \qquad (3)$$

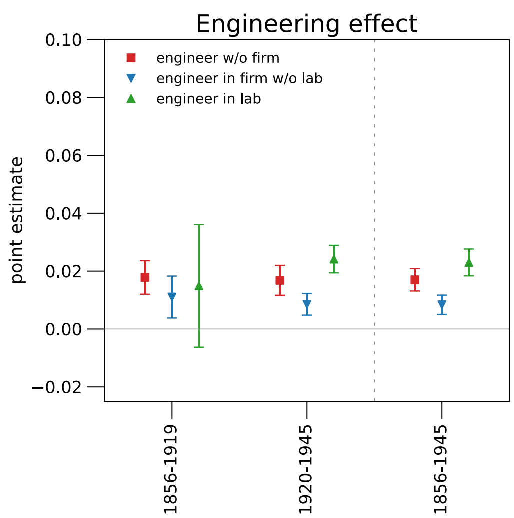
where y_p^d is a dummy variable for whether or not a patent lists a new combination of either 3- or 6-digit technology classes ($d \in \{3, 6\}$), $t(p)$ is the year in which patent p was granted, E_p a dummy for whether or not any of the inventors were engineers and T_p a dummy for team patents. Furthermore, O_p is a dummy group that describes the organizational context (assignee type) behind the patent, where f indicates firms and l industrial research labs. The omitted category refers to standalone patents.

We estimate this model in a sample of patents for which at least one inventor could be matched to census records. Full regression tables are reported in Appendix D.4. Here, we summarize outcomes by describing how novelty effects differ across organizational contexts. The left panel of Fig. 10 shows how engineers perform across organization types (i.e., $\hat{\beta}_e + \hat{\beta}_o + \hat{\beta}_{o\times E}$). The right panel does the same for teams (i.e., $\hat{\beta}_t + \hat{\beta}_o + \hat{\beta}_{o\times T}$). Effect estimates are always relative to a baseline of patents by standalone, solo inventors in non-engineering occupations.

Engineers and teams are both more likely to patent new technological combinations. However, whereas the engineering effect differs little across assignee types,[27] the team

[27]This means that engineering and firm or lab effects are not additive. In fact, the interaction effect of firm and engineering dummies is negative and its size exactly offsets the engineering effect. As a consequence, in firms, engineers do not generate more novelty than their other colleagues. For labs, the engineering and lab effects are less than additive, but don't fully cancel out. Engineers in labs do therefore patent slightly more novel inventions than other types of inventors.

a.



b.

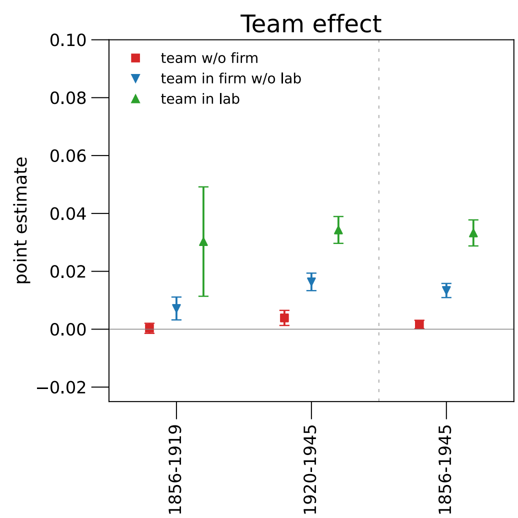


Figure 10: Novelty of patents and labor inputs. The left panel analyzes patents by engineers, the right panel team-based patents. Markers plot the difference in mean novelty at the 3-digit level between patents by engineers, respectively, teams and a baseline composed of standalone patents by solo inventors who are not engineers. Red: standalone patents, blue: patents of firms without industrial research labs, green: patents of industrial research labs. Estimates refer to one of three samples: patents granted between 1856 and 1920, patents granted between 1920 and 1945 and patents granted between 1856 and 1945. Vertical lines display 95% confidence intervals.

effect is contingent on the organizational context. In fact, standalone teams are no more likely to patent a new technological combination than the baseline of standalone, non-engineer, solo inventors. Once teams work for firms – and even more so once they work for firms that operate industrial research labs – the likelihood that a patented invention is novel goes up substantially. To put our estimates into context: in the period 1856-1945, on average 5.4% of inventions list novel technological combinations. Lab-based teams outperform the baseline by 3.3 percentage, or by 60% of this average.

This analysis shows that, although firms and labs lead the shift toward patenting by engineers and teams, when it comes to novelty generation, they only seem to have an impact on teams, not engineers. An important strength of the organizational innovation that the modern research lab represented lies therefore in its capacity to facilitate innovative teamwork.

These findings are highly robust. First, there is little difference between effects estimated in different time periods. This suggests that when it comes to the increase in novelty we observe in the 1920s, what changes is not that engineering- or team-based patents become more novel but that innovation is increasingly organized by firms and labs —which tends to produce more novel patents. Second, controlling for technology-time fixed effects that interact year dummies with the six broad technology classes defined by Hall et al. (2001) does not qualitatively change outcomes (see Appendix D.4).[28]

4.5 Macro level consequences for the US innovation system

4.5.1 System 1 and system 2 invention

Taken together, our findings suggest that the shift to engineers and lab-based teamwork played an important role in the rapid expansion of combinatorial exploration that started in the 1920s. To describe the spatial and societal consequences of this shift, we define two archetypal systems of invention. *System 1*, representing the craftsmanship-based system that dominates the 19th century, consists of patents by solo, standalone inventors without engineering backgrounds. *System 2*, representing the science-based system that takes off in the 1920s, consists of patents that are either invented by engineers, or created in research labs or by firm-based teams.[29]

These definitions somewhat mimic the innovation literature's distinction between Schumpeter Mark I and Mark II innovation patterns. However, whereas Schumpeter Mark I innovation is typically associated with creative destruction and radical change and Mark II

[28] Team effects are reduced by about one third when we add year-technology fixed effects, suggesting that firm and lab-based teams are more common in technological areas that witness more novel technological combinations.

[29] That is, to belong to system 1, a patent needs to fulfill all of the following three criteria: (1) list only one inventor, who (2) has no engineering occupation and (3) no organizational assignee. Instead, system 2 patents fulfill at least one of three criteria: the patent (1) lists an engineer among its inventors, or (2) is filed by a team and assigned to a firm, or (3) is assigned to a firm with a research lab.

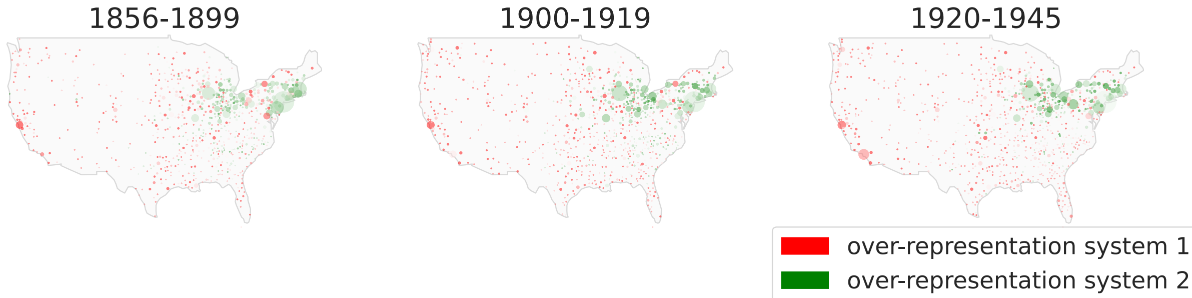


Figure 11: Overrepresentation of system 1 and 2 across US cities. Graphs show which system is overrepresented in each of 933 US cities, where overrepresentation is defined as the ratio of the city's share of patents in one system over the share of the city's patents in the other system. Cities in which system 2 is overrepresented are colored green, cities in which system 1 is overrepresented red. The less transparent the color, the greater the system's overrepresentation. Marker sizes represent the city's share of all US patents.

innovation with cumulative progress and incremental change (Breschi et al., 2000), our findings suggest the opposite: in the first half of the 20th century, it is the organized teams of system 2, working in firms and labs, that introduce radically new combinations of technologies, not the solo, standalone inventors of system 1.

4.5.2 Geography

Systems 1 and 2 are not only associated with different degrees of novelty, but also exhibit divergent spatial patterns. This is shown in Fig. 11, which plots the geography of system 1 and system 2. Cities in which system 1 patents are overrepresented (i.e., whose share of system 1 patents exceeds their share of system 2 patents) are colored red, cities in which system 2 is overrepresented are colored green. System 2 is overrepresented in an area that is nowadays known as the "American Rustbelt," but that, at the time, represented the cutting edge of the US economy (Lamoreaux et al., 2007). The rise of the innovation hubs of this region was therefore strongly intertwined with the rise of system 2.

In Appendix D.2, we show that the geography of innovation also changed in terms of the role that cities played. After innovation had become increasingly dispersed in the 19th century, in the early 20th century, innovation reconcentrated in a small number of technology hubs. This went hand in hand with the fall and rise of large cities. Around 1850, innovation had been heavily concentrated in the largest cities in the US. As innovation started becoming more dispersed over the course of the late 19th century, large cities lost their dominance. This trend halted and then reversed in the 20th century, when first system 2 concentrated in a few large cities, to be followed somewhat later by system 1.

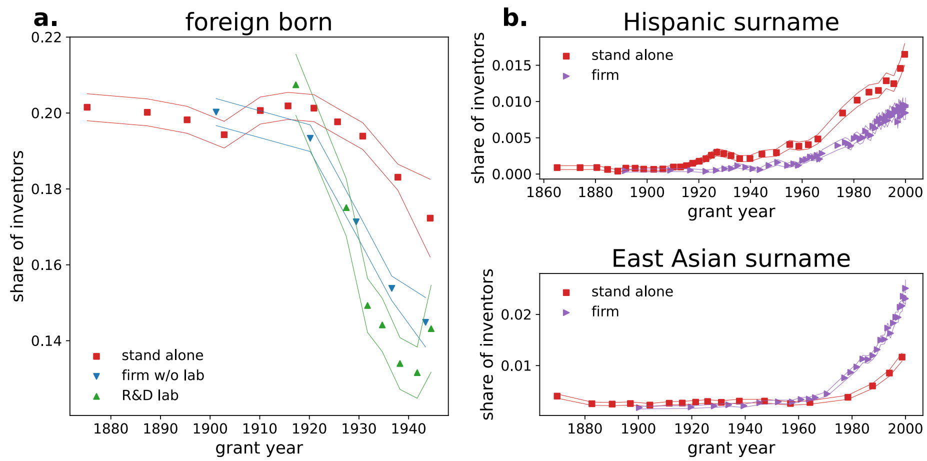


Figure 12: Immigrant participation rates **a**: share of foreign-born inventors by assignee type (1856-1945). **b**: upper panel: share of inventors with Hispanic surname, lower panel: share of inventors with East-Asian surnames (1856-2000). Colors: red: standalone patents, purple: firms; blue: firms without research labs; green: research labs.

4.5.3 Demographics and unequal participation

Immigrants play and have played an important role in US innovation (Akcigit et al., 2017; Kerr, 2013; Lissoni and Miguelez, 2024). Generally, immigrants are also overrepresented among inventors in the period we study (see Appendix D.3). However, after 1920, marked differences emerge between the firm- and lab-based patents of system 2 and the standalone patents of system 1 (see Fig. 12a). In these decades, the share of foreign-born inventors decreases across the board, mostly reflecting a relative decline of foreign-born individuals in the US population. However, foreign-born inventors' patenting shares drop much faster in firms and labs than on standalone patents.

To expand this analysis beyond the years for which we have census records, we analyze the patents of inventors with Hispanic and East Asian surnames, using the geographical origins of last names as proxies for inventors' ethnicity. Because we lack information about research labs after 1945, we only distinguish between standalone and firm-based patents. Comparing these two types of patents, we observe stark differences between inventors with Hispanic and East Asian surnames: whereas, relative to standalone patenting, inventors with Hispanic surnames are underrepresented in firm-based invention, inventors with East-Asian surnames are overrepresented (Fig. 12b).

In contrast to immigrants, women are notoriously underrepresented in patenting (e.g., Ding et al., 2006). Fig. B4 of Appendix B.5 shows that this was even more so in the past. Based on inventors' inferred genders, we find that, although the share of female inventors slowly increases over time, it remains below 2% throughout the 1856-1945 period.

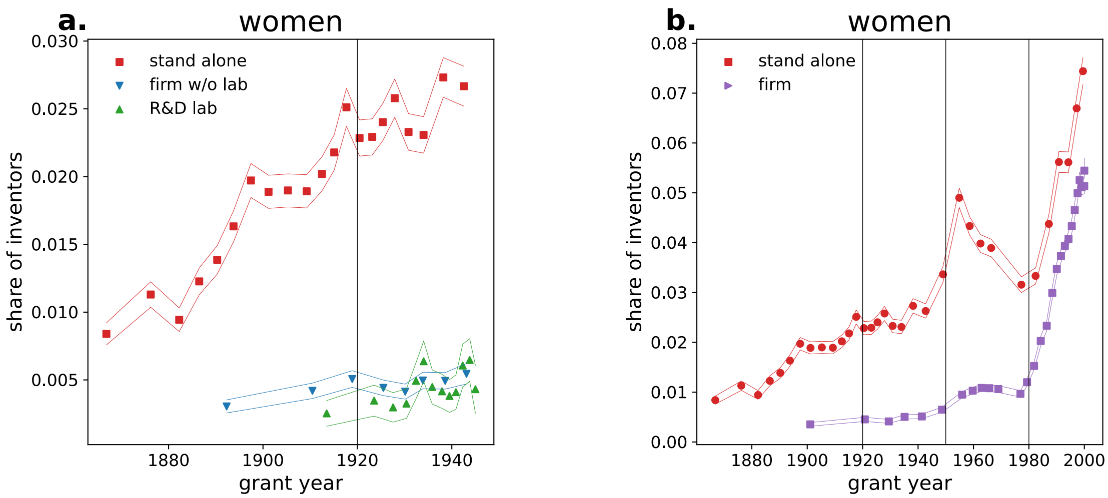


Figure 13: Participation of women in system 2. **a**: share of inventors with first names that are predominantly used by women in different organizational contexts (1856-1945) **b**: share of inventors with first names that are predominantly used by women in standalone versus firm-based patents (1856-2000).

Moreover, this rise comes to an abrupt halt at the end of WWI, after which female inventor shares start falling again.

However, when we distinguish between different organizational contexts, this drop turns out to be wholly attributable to firm- and lab-based patents (Fig. 13a). In contrast, the share of women on standalone patents keeps rising. In fact, women are drastically underrepresented in firm and lab-based patents compared to standalone patents. In the 1940s, female inventor shares are about a factor 5 lower among firm-based inventors than among standalone inventors. In labs, the female share is a factor 7 below the standalone share.[30] This gap in gender shares only starts closing in the late 1970s (Fig. 13b). However, in 2000, the share of female inventors is still about 25% lower in firms than among standalone inventors.

[30]In Appendix D.4, Table D9, we study female participation rates using the regression model of eq. (3) with as dependent variable a dummy that evaluates to one if we can identify at least one inventor with a first name that is likely to be female among the patent's inventors. Focusing on the fully interacted model in column (5), we find that the likelihood of finding female inventors on a patent is lower in firm- and, even more so, lab-based patents. Also patents that list engineers are less likely to list women. However, women were more likely to be listed on team-based patents. This suggests that system 2 included few women, both because there were very few female engineers and because women often did not participate in firm- and lab-based invention. Although teamwork counteracted this, it did not eliminate the resulting gender gap.

Scholars have identified a number of potential drivers of gender gaps in patenting. For instance, women may have fewer ties to industrial partners (Ding et al., 2006), lack successful mentors (Delgado and Murray, 2022) or role models (Bell et al., 2019). Other explanations refer to a lower patentability of inventions in fields where women are most active (Ding et al., 2006) or to outright biases that exclude women from being listed as coinventors (Ross et al., 2022). Although we cannot determine the deeper causes, our analysis suggests that gender gaps in patenting were exacerbated with the introduction of firm- and lab-based R&D.

4.6 Novelty in the 1945-2000 period

In the period 1856-1945, the rapid increase of patents that list new technological combinations seems to be mostly driven by a rise of patents by teams and engineers in firms and research labs. However, if we extend our knowledge-production function approach to the period 1856-2000, we find that the link between novelty and teamwork in corporate R&D gets severed after the 1950s.

For this longer period, we can no longer observe inventor demographics or industrial research labs, but we still see whether or not patents are filed by teams and whether or not they are assigned to firms. Fig. 14a shows results of a simplified version of the team analysis in Fig. 10b. After 1945, the greater propensity of firm-based teams to patent novel combinations of technologies shrinks. In the last quarter of the 20th century, firm-based teams even underperform standalone teams, suggesting that, after 1945, firms no longer enhance teams' capacity to generate novel technological combinations.

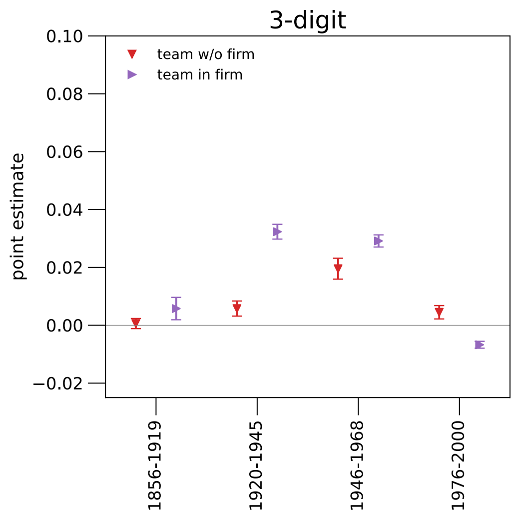
5 Discussion and conclusion

The 1920s are a pivotal point in the history of US invention. Supported by systemic changes in the way the US innovates, the country rapidly transitions from craftsmanship-based to science-based innovation, entering a 30-year period of increased exploration of radically new technological combinations.

We have argued that a key element in this transition was the diffusion of industrial research labs to the US. By offering a new and better way to tap into a growing collective body of technological knowledge, these labs represented an organizational innovation that became instrumental in coordinating teams of highly skilled inventors. This organizational innovation therewith likely was at the same time a response of firms to changes in the nature of technological know-how and a way to accelerate its advance. This put industrial research labs at the heart of a set of interlocking changes that would unfold throughout the US innovation system.

At the individual level, this is evident in the emergence of steep learning curves, with 20th century inventor cohorts inventors combining progressively more and more recent technologies as they age. Other individual-level changes reflect the turn to science-based

a.



b.

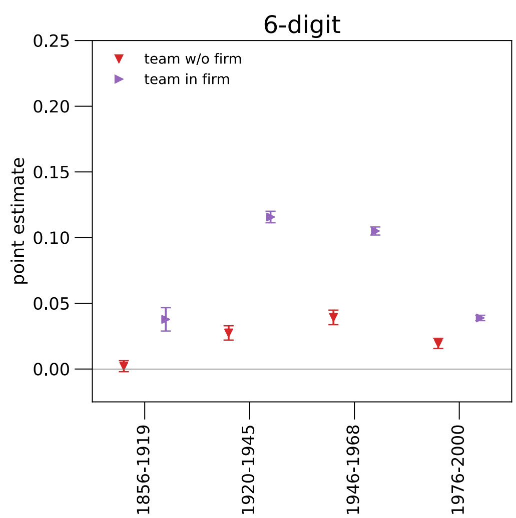


Figure 14: Novelty of team patents: 1856-2000. Markers plot the difference in mean novelty at the 3-digit level (left panel) or 6-digit level (right panel) between teams and a baseline composed of standalone patents by solo inventors. Red: standalone team patents, blue: firm-based team patents. Estimates refer to one of five time periods: 1856-1920, 1920-1945, 1946-1968, 1976-2000 or 1856-2000.

innovation, such as the rise of academic patenting, the increasing reliance on science as reflected in references to the academic literature and the arrival on the scene of a new type of innovation specialist: the engineer. The 1920s also mark the start of the long ascent of teamwork in invention that has persisted throughout the 20th century. These engineers and this teamwork were supported by the emergence at the meso level of a new organizational entity: the industrial research lab. Originating from Germany, these labs not only employed a disproportionate number of engineers, but were also responsible for most of the rise in teamwork. Moreover, they were particularly apt at coordinating teams in repeated collaborations and over long distances, replacing family ties as the dominant organizational mechanism in team invention. In fact, the research lab's effect on innovation seems to have operated more through facilitating teamwork than through the scores of engineers they hired, enhancing the capacity to generate novel combinations in teams, but not necessarily by engineers. Finally, with its concentration in the emerging technology hotspots of the Rustbelt, the new science-based innovation system also left deep traces in the geography of innovation, reconcentrating innovative activity in a small number of mostly large cities.

5.1 Policy implications

Although our analysis does not primarily focus on policy, the US history of industrial research labs we document — combined with arguments in the literature on this topic — holds an important lesson for policy making. In line with Arora et al.'s (2020) conjecture that research labs drove much of the productivity increases in the 20th century US economy, our findings suggest that organized industrial research in labs of large corporations played a key role in the acceleration of radical innovation in the 1920s. These labs were often run by companies that were *de facto* monopolies – as in the case of Bell Labs' parent company AT&T – or that at least had strong market power. Yet, in spite of the lack of competition, our study shows that these labs not only produced many patents, but also radical innovations based on ground-breaking fundamental research. In fact, it may have been precisely the monopoly position of firms like AT&T that allowed them to develop sprawling research portfolios. For instance, Bell Labs employed eleven Nobel prize winners and its research led to the invention or improvement of transistors, solar cells, satellite technology, deep-sea communication cables, information theory, and more (Gertner, 2012).

In fact, our analysis shows that the innovation activities of industrial research labs eclipsed the patenting output of academics and universities. These labs were not just better funded than most universities, offering state-of-the-art research facilities, but also connected to broad long-term goals, designed to serve the parent firms' markets. This allowed corporate labs to connect basic research — which is apparent in their innovations' reliance on science and engineers — to real-world problems. In this sense the research labs of the early 20th century developed an innovation strategy that resembles various elements of what we nowadays would call mission-oriented innovation policy (Mazzucato, 2021). For instance, Bell Labs did not have a narrow objective but aimed to improve communication far beyond its existing telephony business (Gertner, 2012). This provided a clear, yet broad mission to its workforce. AT&T's monopoly then allowed funding Bell Labs to undertake projects with very long time horizons. In fact, because monopolists are shielded from competitive pressures, they can focus on long-term profit growth, which must come either from cost-saving efficiency increases (e.g., replacing operators by electronic switches) or consumer-value creating improvements in service (e.g., satellite communication), instead of on the short-term defense of market shares (Schumpeter, 1942). Finally, the ideas that Bell Labs produced could be immediately tested and then implemented at scale by Western Electric, AT&T's manufacturing subsidiary. Together, these elements assured Bell Labs of long planning horizons, a broad, yet concrete mission and internal stakeholders that were invested in, and capable of, solving existing and identifying new engineering challenges for the labs.

Naturally, monopoly power also has downsides, such as keeping prices high for customers. However, in the case of Bell Labs, it was the monopoly that allowed AT&T to appropriate most of the benefits of the Labs' basic research, justifying the high investments in this research. At the same time, due to a peculiarity in AT&T's agreements with antitrust authorities, it also created benefits that spilled over to the wider economy: In order to maintain its telephony monopoly and avoid a break-up of the company, in 1956, AT&T agreed to license all of its existing patents royalty-free to American applicants, and future patents for small fees (Gertner, 2012). For AT&T, these arrangements had limited consequences: the company's core business was protected also without patents by its monopoly position. However, others could now freely use the know-how that Bell Labs had generated over the decades in businesses outside telephony. A vivid illustration of the value of this arrangement to US society as a whole is Bell Labs' transistor technology, which would become key to the rise of Silicon Valley.[31]

The decline of industrial research labs in the second half of the 20th century — signs of which we observe in the reduced novelty of firm-based patents — shifted the burden of basic research almost fully to the university system (Arora et al., 2020). However, arguably, this was not nearly as well equipped for this task, as it lacked many of the mission-orientation elements that R&D labs had leveraged.

[31] Although deals with antitrust authorities like AT&T's are uncommon, similar spillovers occur when corporations fail to appropriate all returns to their inventions and instead allow spin-offs or other companies to commercialize their intellectual property. A leading example is Xerox PARC, which invented many valuable technologies outside its main printer and copier business that eventually were successfully commercialized by other companies, such as Ethernet adapters by 3Com and PostScript by Adobe Systems (Chesbrough, 2002).

5.2 *Limitations and future research*

Our results likely understate the full importance of these labs. First, we will not have been able to match all patent assignees to their labs, especially when firms filed patents under names sufficiently distinct from those listed in the industrial research lab surveys. Second, firms may have operated research labs or similar organizational units, without being listed in the survey. Third, inventors will not have assigned the intellectual property rights for all of their inventions to their employers. In fact, employees often only had to assign such rights for inventions related to the firm's main line of business (see, for instance, Hertz, 1950, pp. 327-328).

Similar limitations apply to other aspects of our study. For instance, we were only able to match about 46% of inventors to census records and, given the decennial nature of the census, for the inventors we could match, these records may not always have offered up-to-date and accurate information.

Another class of limitations has to do with the fact that patents do not capture the full extent of innovation nor of the efforts that individuals and firms put in. On the one hand, it is well-known that much innovation goes unpatented (Archibugi, 1992). On the other hand, in spite of legal obligations to do so, patents may not list all contributors. This would mean that a number of solo-inventor patents were in reality the result of teamwork. Such miss-measurement likely biases estimated coefficients towards zero, implying that most of our estimates are likely conservative.

Third, our study ignores inventions that were the result of international collaborations. Although the majority of US inventor teams were entirely based in the US, international collaborations can already be observed in the 19th century. While interesting, we believe that this aspect of the evolution of US invention deserves a study of its own.

Fourth, our analysis sketches trends and correlations and can only hint at causal relations. For instance, we don't know whether the turn to science-based invention was a result of technological change, increasing levels of education or driven by corporate strategy. Literature in economic history suggests that these factors all played a role and reinforced one another. For instance, the division of labor in the growing teams that industrial research labs coordinated relied on the availability of specialized inventors, who themselves were the product of expanded education in the engineering schools of the Morrill Land Grant Acts. In fact, seven of the ten universities listed in Table 4 are land-grant colleges. Similarly, organized corporate research became more profitable after a wave of mergers had created very large firms in various sectors of the economy.

Our study also suggests a number of questions that could be studied in more detail in future research. One question relates to whether organized corporate research is conducive to radical innovation. Our study tentatively contradicts the canonical view that the highly organized nature of Schumpeter Mark II innovation is best suited for incremental, not radical innovation — at least in the period in which this type of innovation first emerged. Other questions pertain to the consequences of the shift in the innovation system. These include how the exclusion of women and immigrants from science-based innovation materialized and why the system first emerged in the American Rustbelt.

Another question relates to the role played by engineers. Engineers became a dominant force in the 1920s, when they were hired in large numbers by the emerging R&D labs that characterized the golden age of American ingenuity of the second industrial revolution. Interestingly, according to recent work by Hanlon (2022), engineers also played an outsized role in innovation in the UK. However, Hanlon's work describes innovation in the first, not the second, industrial revolution. So whereas in the UK, engineers revolutionized invention in the 1820s, American engineers only rose to prominence a full century later, in the 1920s. Moreover, they did so not as individual inventors, but as part of team-based innovation in large corporations.

Finally, from 1950 on, corporate R&D shifts toward incremental innovation, with firms — and in particular firm-based teams — becoming less likely to patent radically new combinations of technologies. This resonates with the finding in Wu et al. (2019) that today's very large research teams struggle to generate radically new ideas in science and technology. One challenge of running large teams is that the number of bilateral links increases with the square of the number of team members. A potential explanation for the reduced capacity for radical innovation by corporate inventor teams is that existing organizational forms fail to accommodate this rising complexity of coordination. This hypothesis draws our attention to new collaboration technologies, such as Slack, Zoom and other online platforms. These platforms may help coordinate teamwork in new ways, just like labs had done 100 years before. If so, they could set in motion developments that ripple throughout the economy, from the division of labor to a reshuffling of global production structures. Studying team collaboration in modern-day industrial research could therefore yield important insights, shifting attention from how technological innovations change the future of work to how organizational innovations change the future of teamwork.

Acknowledgements

Acknowledgements: The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. We are grateful to Sergio Petralia and Riccardo Crescenzi for bringing the existence of historical books on patents and R&D labs to our attention and early discussions of this research and to Frank van der Wouden for sharing his data with us. Furthermore, we thank Ana Grisanti for helping us create ground truths for optical character recognition purposes, and Mali Akmanalp for programming and data engineering advice. We gratefully acknowledge the feedback we received from Llewellyn Thomas, Stefania Scrofani, James McNerney and Lingfei Wu, as well as from participants at the 2022 DRUID conference, Growth Lab Seminar at Harvard University, UNU-Merit seminar at Maastricht University, the Structure of Technology workshop at the Santa Fe Institute, the Complexity Economics Seminar Series at Oxford University, the MORS seminar at Berkeley Haas School of Business and the IPL seminar at the Toronto Munk School of Global Affairs and Public Policy.

Funding: Frank Neffke gratefully acknowledges financial support from the Austrian Research Promotion Agency (FFG), project #873927 (ESSENCSE).

References

Akcigit, U., Grigsby, J., and Nicholas, T. (2017). Immigration and the rise of american ingenuity. *American Economic Review*, 107(5):327–331.

Almeida, P., Phene, A., and Li, S. (2015). The influence of ethnic community knowledge on indian inventor innovativeness. *Organization Science*, 26(1):198–217.

Ambekar, A., Ward, C. B., Mohammed, J., Male, S., and Skiena, S. (2009). Name-ethnicity classification from open sources. In *Knowledge Discovery and Data Mining*.

Archibugi, D. (1992). Patenting as an indicator of technological innovation: a review. *Science and public policy*, 19(6):357–368.

Arora, A., Belenzon, S., Kosenko, K., Suh, J., and Yafeh, Y. (2021). The rise of scientific research in corporate america. Technical report, National Bureau of Economic Research.

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

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. Simon and Schuster.

Becker, G. S. and Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly journal of economics*, 107(4):1137–1160.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., and Van Reenen, J. (2019). Who becomes an inventor in america? the importance of exposure to innovation. *The Quarterly Journal of Economics*, 134(2):647–713.

Bittlingmayer, G. (1985). Did antitrust policy cause the great merger wave? *The Journal of Law and Economics*, 28(1):77–118.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Breschi, S., Malerba, F., and Orsenigo, L. (2000). Technological regimes and schumpeterian patterns of innovation. *The economic journal*, 110(463):388–410.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Carlson, W. B. (2013). Innovation and the modern corporation: From heroic invention to industrial science. In *Science in the twentieth century*, pages 203–226. Routledge.

Chandler Jr, A. D. (1990). Scale and scope: The dynamics of industrial capitalism.

Chandler Jr, A. D. (1993). *The visible hand*. Harvard university press.

Chesbrough, H. (2002). Graceful exits and missed opportunities: Xerox's management of its technology spin-off organizations. *Business History Review*, 76(4):803–837.

Clancy, M. S. (2018). Inventing by combining pre-existing technologies: Patent evidence on learning and fishing out. *Research Policy*, 47(1):252–265.

Cohen, A. M. (2009). *The Shaping of American Higher Education: Emergence and Growth of the Contemporary System*. Jossey-Bass Publishers.

Delgado, M. and Murray, F. (2022). Faculty as catalysts for new inventors: Differential outcomes for male and female phd students. In *Academy of Management Proceedings*, number 1, page 17297. Academy of Management Briarcliff Manor, NY 10510.

Ding, W. W., Murray, F., and Stuart, T. E. (2006). Gender differences in patenting in the academic life sciences. *science*, 313(5787):665–667.

Esposito, C. R. (2023). The geography of breakthrough invention in the united states over the 20th century. *Research Policy*, 52(7):104810.

European Patent Office (2020). Patstat - worldwide patent statistical database - spring 2020.

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

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

Gertner, J. (2012). *The idea factory: Bell Labs and the great age of American innovation*. Penguin Press, New York.

Goldin, C. (1998). America's graduation from high school: The evolution and spread of secondary schooling in the twentieth century. *The Journal of Economic History*, 58(2):345–374.

Goldin, C. and Katz, L. F. (1999). The shaping of higher education: The formative years in the united states, 1890 to 1940. *Journal of Economic Perspectives*, 13(1):37–62.

Goldin, C. and Katz, L. F. (2000). Education and income in the early twentieth century: Evidence from the prairies. *The Journal of Economic History*, 60(3):782–818.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. Working Paper 8498, National Bureau of Economic Research.

Hanlon, W. W. (2022). The Rise of the Engineer: Inventing the Professional Inventor During the Industrial Revolution. Technical Report w29751, National Bureau of Economic Research, Cambridge, MA.

Henderson, R., Jaffe, A. B., and Trajtenberg, M. (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and statistics*, 80(1):119–127.

Henrich, J. (2004). Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the tasmanian case. *American Antiquity*, 69(2):197–214.

Hertz, D. B. (1950). The theory and practice of industrial research.

Hounshell, D. A. (1996). The evolution of industrial research in the united states. In Rosenbloom, R. S. and Spencer, W. J., editors, *Engines of Innovation*. Harvard Business School Press, Boston, MA.

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

Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2):363–375.

Jung, T. and Ejermo, O. (2014). Demographic patterns and trends in patenting: Gender, age, and education of inventors. *Technological Forecasting and Social Change*, 86:110–124.

Kaltenberg, M., Jaffe, A., and Lachman, M. E. (2021). Matched inventor ages from patents, based on web scraped sources.

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

Kerr, W. R. (2013). Us high-skilled immigration, innovation, and entrepreneurship: Empirical approaches and evidence. Technical report, National Bureau of Economic Research.

Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *The American economic review*, pages 562–583.

Lamoreaux, N. R., Levenstein, M., and Sokoloff, K. L. (2007). Financing invention during the second industrial revolution: Cleveland, ohio, 1870–1920.

Lamoreaux, N. R. and Sokoloff, K. L. (1999). Inventors, firms, and the market for technology in the late nineteenth and early twentieth centuries. In *Learning by doing in markets, firms, and countries*, pages 19–60. University of Chicago Press.

Lamoreaux, N. R. and Sokoloff, K. L. (2001). Market trade in patents and the rise of a class of specialized inventors in the 19th century united states. *American Economic Review*, 91(2):39–44.

Lenoir, T. (1998). Revolution from above: the role of the state in creating the german research system, 1810-1910. *The American Economic Review*, 88(2):22–27.

Li, Y. and Neffke, F. (2023). Evaluating the principle of relatedness: Estimation, drivers and implications for policy. Technical report, Center for International Development at Harvard University.

Lissoni, F. and Miguelez, E. (2024). Migration and innovation: Learning from patent and inventor data. *Journal of Economic Perspectives*, 38(1):27–54.

Marco, A. C., Carley, M., Jackson, S., and Myers, A. F. (2015). The uspto historical patent data files: Two centuries of innovation. *SSRN working paper*.

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

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

Mazzucato, M. (2021). *Mission economy: A moonshot guide to changing capitalism*. Penguin UK.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Michaelis, A. R. (2003). Justus von liebig, frs: creator of the world's first scientific research laboratory. *Interdisciplinary Science Reviews*, 28(4):280–286.

Mowery, D. C. (1990). The Development of Industrial Research in U.S. Manufacturing. *American Economic Review*, 80(2):345–349.

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

Muneepeerakul, R., Lobo, J., Shutters, S. T., Goméz-Liévano, A., and Qubbaj, M. R. (2013). Urban economies and occupation space: Can they get "there" from "here"? *PloS one*, 8(9):e73676.

Muthukrishna, M. and Henrich, J. (2016). Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690):20150192.

National Research Council (1956). *Industrial Research Laboratories of the United States, Including Consulting Research Laboratories.* National Research Council, Washington, D.C.

Neffke, F. and Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3):297–316.

Neffke, F. M. (2019). The value of complementary co-workers. *Science advances*, 5(12):eaax3370.

Neffke, F. M., Otto, A., and Weyh, A. (2017). Inter-industry labor flows. *Journal of Economic Behavior & Organization*, 142:275–292.

Nicholas, T. (2010). The role of independent invention in u.s. technological development, 1880–1930. *The Journal of Economic History*, 70(1):57–82.

Pakes, A. and Griliches, Z. (1980). Patents and r&d at the firm level: A first report. *Economics letters*, 5(4):377–381.

Petralia, S., Balland, P.-A., and Rigby, D. L. (2016). Unveiling the geography of historical patents in the United States from 1836 to 1975. *Scientific Data*, 3(1):160074.

Pezzoni, M., Veugelers, R., and Visentin, F. (2022). How fast is this novel technology going to be a hit? antecedents predicting follow-on inventions. *Research Policy*, 51(3):104454.

Pithan, D. (2021). *Corporate research laboratories and the history of innovation.* Routledge.

Reich, L. S. (1985). *The making of American industrial research: science and business at GE and Bell, 1876-1926*. Studies in economic history and policy. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

Righi, C. and Simcoe, T. (2019). Patent examiner specialization. *Research Policy*, 48(1):137–148.

Rocke, A. (2021). 41The Rise of Academic Laboratory Science: Chemistry and the 'German Model' in the Nineteenth Century. In *History of Universities: Volume XXXIV/1: A Global History of Research Education: Disciplines, Institutions, and Nations, 1840-1950*. Oxford University Press.

Ross, M. B., Glennon, B. M., Murciano-Goroff, R., Berkes, E. G., Weinberg, B. A., and Lane, J. I. (2022). Women are credited less in science than men. *Nature*, 608(7921):135–145.

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., and Sobek, M. (2021). *IPUMS USA: Version 11.0 [Dataset]. Minneapolis.* IPUMS Minnsesota: MN.

Schmidgen, H. (2021). Laboratory. *Encyclopedia of the History of Science*, 4(1).

Schumpeter, J. (1942). *Socialism, Capitalism and Democracy.* Harper and Brothers, NY.

Shennan, S. (2001). Demography and cultural innovation: a model and its implications for the emergence of modern human culture. *Cambridge archaeological journal*, 11(1):5.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4):641–658.

Sood, G. and Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name.

Strumsky, D. and Lobo, J. (2015). Identifying the sources of technological novelty in the process of invention. *Research Policy*, 44(8):1445–1461.

Travis, A. S., Hornix, W. J., Bud, R., and Homburg, E. (1992). The emergence of research laboratories in the dyestuffs industry, 1870–1900. *The British journal for the history of science*, 25(1):91–111.

USPTO (2022). PatentsView. `https://www.patentsview.org/download/`.

van Dam, A., Gomez-Lievano, A., Neffke, F., and Frenken, K. (2023). An information-theoretic approach to the analysis of location and colocation patterns. *Journal of Regional Science*, 63(1):173–213.

Van der Wouden, F. (2020). A history of collaboration in us invention: changing patterns of co-invention, complexity and geography. *Industrial and Corporate Change*, 29(3):599–619.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research policy*, 45(3):707–723.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics*, 113(2):331–360.

Winter, S. G. (1984). Schumpeterian competition in alternative technological regimes. *Journal of Economic Behavior & Organization*, 5(3-4):287–320.

Wu, L., Wang, D., and Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827):1036–1039.

Youn, H., Strumsky, D., Bettencourt, L. M., and Lobo, J. (2015). Invention as a combinatorial process: evidence from us patents. *Journal of the Royal Society interface*, 12(106):20150272.

A Data sources

Our analyses combine three different types of data. The main data set contains information on patents issued by the United States Patent and Trademark Office (USPTO). We add to this existing data from the European Patent Office's (EPO) PATSTAT and the USPTO's PatentsView datasets. For US-based inventors, we combine these patent data with demographic information from the US population censuses between 1850 and 1940. Finally, for US assignees, we add information on industrial research labs. Fig. A1 provides a schematic overview of all data processing and merges that were carried out.

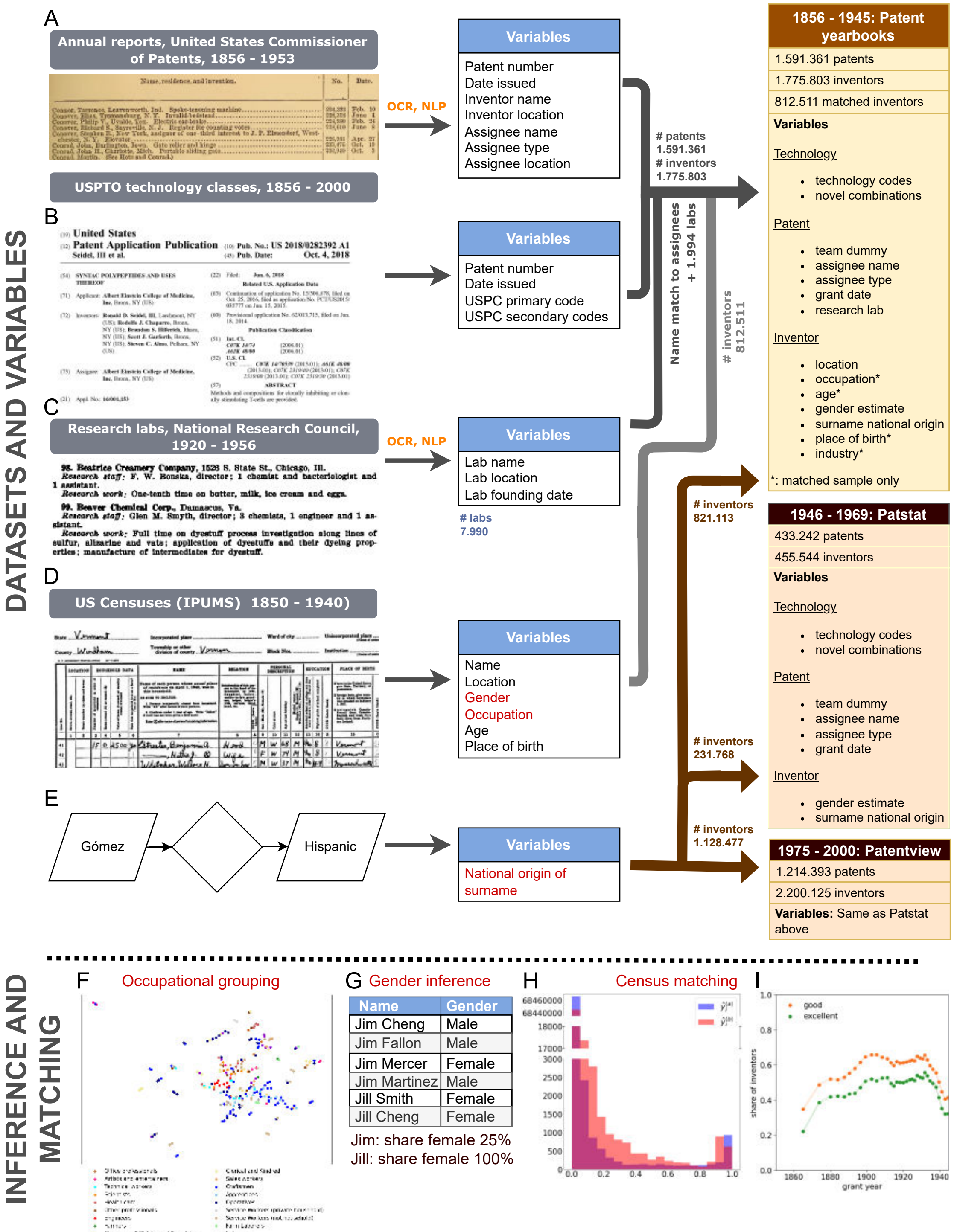


Figure A1: **A**: Sample page of the 1880 Annual report of the Commissioner of Patents. **B** First page of patent number 2018/0282392 A1. **C** Sample page of 1931 edition of "Industrial Research Laboratories of the United States". **D** Sample page of US Census form. **E** Identification of most likely national origin of surnames. Hispanic surnames are identified using models trained on the 2000 US Census (Sood and Laohaprapanon, 2018), East-Asian surnames using models trained on Wikipedia (Ambekar et al., 2009). **F**: Projection of cross-occupation labor flow network on 2-dimensional UMAP embedding. **G** First names are associated with genders by majority vote in about 650M census records. We focus on a sample where first names display little ambiguity in gender, requiring that at least 90% of census records list the same the gender for the first name. **H**: Histogram of $\hat{y}_i^{(a)}$ (blue) and $\hat{y}_i^{(b)}$ (red) in a random sample of about 68M potential matches. Some potential matches of seemingly high quality when using only name- and geographical distances are downgraded in the second-stage xgboost model, which also accounts for the quality of alternative matches. **I**: Match rate: share of inventors that can be matched with moderate accuracy ($\hat{y}_i^{(b)} \geq .95$, orange) and high accuracy ($\hat{y}_i^{(b)} \geq .99$, green) to one of the two census waves closest in time to the patent's grant date.

A.1 Patent yearbooks

Patent yearbooks have been scanned by different universities.[32] For most years, we can therefore access multiple digital copies of the same yearbook. All image files for these scanned copies were accessed through the *Hathi Trust Foundation*, except for scans obtained from the *Smithsonian Foundation*, which were obtained directly from this foundation. Together, this amounts to about half a million scanned pages.

To convert scans to text, we apply image preprocessing and optical character recognition (OCR) algorithms. Next, we correct errors in the extracted text by cross-validating the output across different scans of the same yearbook, using a majority vote to decide on the correct string.

We further process these strings, using natural language processing (NLP) algorithms to separate inventor names and places of residence, assignee names and assignee locations, short descriptions of the invention, patent numbers and grant dates. To do so, we first manually create a ground truth that separates the aforementioned entities, which can be found in the online code appendices. Some reports only list first names in full for the first inventor, providing initials and last names for any additional inventors. However, the yearbooks typically contain additional lines that provide full first and last names for the other inventors as well. These lines can be identified by the string "(See <name of first inventor>)". Whenever possible, we supplement first-name information from these lines. However, in some cases and years only initials are provided for second and subsequent inventors, which complicates the merge to census records.

Patents can be assigned (i.e., transfer intellectual property rights) to individuals or organizations other than the original inventors. If this happened before or at the time the patent was granted, the yearbooks contain this information, following the string "assigned to". To distinguish between patents assigned to organizations (e.g., "assigned to Wright Metal Incorporated") and patents assigned to individuals (e.g. "assigned to Benjamin Reece and Neary Claflin"), we once again rely on NLP, manually creating a ground truth training set for named entity extraction. The vast majority of organizations in this period, over 99%, are firms.

Because the same organization may be listed in different yearbooks under different names (e.g. "General Electrics" versus "General Electric Inc." versus "General Electric Incorporated") we manually align common terms (e.g., replacing "Mfg." by "Manufacturing") and then apply a fuzzy name-matching algorithm that calculates string similarities across all assignees. We use this to disambiguate organization names, merging names likely to refer to the same organizations.

This process allows us to extract detailed information on patents, their inventors and assignees for the period 1856-1953. In the analysis, we exclude the years 1873, 1874, 1878,

[32] These are the following universities: Harvard University, Princeton University, University of California, University of Chicago, University of Illinois at Urbana-Champaign, University of Michigan and University of Wisconsin.

1908, 1909, 1951 and 1952. In 1874, yearbooks are missing altogether. In the other years, first names of inventors are not reported, which complicates the match to census data. Furthermore, we restrict the sample of patents to utility patents, excluding reissues or translations of existing patents, provisional patents, design patents and (organic) plant patents. Finally, we drop patents where one or more inventors are located outside the US.

A.2 Technology classes

Technology classes and subclasses in the USPC classification system are provided by the USPTO for bulk download in the CASSIS Patents Assignments File and the *Bulk Data Products* repository (`https://bulkdata.uspto.gov/`). These data provide grant numbers and dates for all patents, as well as the list of primary and secondary technology codes.

We match the patents in the yearbooks to these data using the patent number. However, OCR errors may lead to ambiguous and/or imperfect matches. In these cases, we add the patent's grant date to improve the match.

Finally, we aggregate technology classes into broader categories using the classification of (Hall et al., 2001). Because this classification is not available for all patents, we infer a correspondence between NBER subcategories and USPC main classes, using the primary technology classes for patents that are classified in both systems. We use this correspondence to add the aggregated NBER classes to as many patents as possible.

A.3 Census data

We obtain US census records from the *Integrated Public Use Microdata Series*, or *IPUMS* (Ruggles et al., 2021). These records, about 650 million in total, contain the answers to census questions for all US residents for the years 1850, 1860, 1870, 1880, 1900, 1910, 1920, 1930 and 1940. 1890 is unavailable because a fire destroyed most of the records of that year. In our analysis, we use information on first and last names, years of birth, places of residence, industries and occupations.

Matching inventors to the US census To match inventors to census records, we proceed in five steps:

1. For each inventor, we find a set of candidate matches based on the string distance between the inventor's last name and all last names in a US census wave. On average there are 326,697 candidate matches for each patent-inventor combination.

2. We create a ground truth for a subset of inventor-individual matches, relying on inventors whose patents are listed on Wikidata, which adds detailed information on date and place of birth, places of residence, as well as spouses, children and parents.

3. We train a first-round xgboost model on this ground truth that predicts correct matches from information on distances between an inventor and all candidate matches in the census, using as predictors string distances for first names, last names and initials, as well as the geographical distance between the place of residence in IPUMS and on the patent. This model provides for each inventor-candidate pair a score that describes the quality of each match candidate: $\hat{y}_a$.

4. We train a second-round xgboost model on a different set of ground truth observations that uses as predictors $\hat{y}_{ij}^{(a)}$, as well as the top k $\hat{y}_{ij'}^{(a)}$ scores across all match candidates associated with inventor i. This yields for each inventor-candidate pair a second match score $\hat{y}_{ij}^{(b)}$.

5. We link inventors to individuals by choosing the individual in the US census with the highest $\hat{y}_{ij}^{(b)}$ across all candidates j.

The first-round model effectively decides how different distances between inventor and census individuals should be weighted when choosing among multiple match candidates. The second-round model helps determine how much confidence we should have in a match, taking into consideration that confidence should be low if either $\hat{y}_{ij}^{(a)}$ is low, or if there are multiple candidates j with more or less equal values of $\hat{y}_{ij}^{(a)}$.

Fig. A2 shows how well our models perform within our ground truth dataset. To do so, we split inventors into a 75% train and a 25% test dataset and then add all potential match candidates from the census to the corresponding inventors. We fit our models using the train dataset and use the fitted model to predict out-of-sample for each inventor in the test dataset which is the best match among all available candidates. We denote this match candidate by the subscript $j^*(i)$. Next, we calculate the deciles of $\hat{y}_{ij^*(i)}^{(b)}$ in the test sample, $\hat{y}_d^{(b)}$ for $d \in \{0.1, 0.2, ..., 1.0\}$. Finally, we plot the share of correctly matched inventors on the y-axis against the share of the sample that we match if we choose $\hat{y}_{ij^*(i)}^{(b)} \geq \hat{y}_p^b$ in descending order of $\hat{y}_d^{(b)}$.

The graph shows that up to 60% of inventors can be matched with very high accuracy to census records. After this, the match rate starts falling. Our decision to match inventors to census individuals whenever $\hat{y}_{ij^*(i)}^{(b)} \geq 0.99$ amounts to a true positive rate of 0.92 in our ground truth data. in contrast, choosing a cut-off such that $\hat{y}_{ij^*(i)}^{(b)} \geq 0.95$ implies a true positive rate of 0.89.

Fig. A1H shows the histograms of $\hat{y}_i^{(a)}$ and $\hat{y}_i^{(b)}$ in a random sample of about 68M potential matches taken from the entire population of inventors and their match candidates (i.e., not limited to our ground truth data). In this paper, we only use matches for which $\hat{y}_i^{(b)} \geq .99$. Although the second stage xgboost estimates, $\hat{y}_i^{(b)}$, do not improve match quality, they do downgrade a number of seemingly high-quality matches. This typically happens when inventors with common names have multiple close matches in the

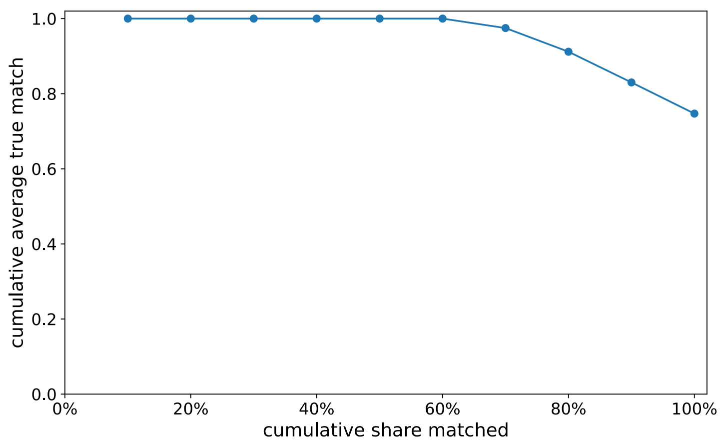


Figure A2: Out-of-sample performance of matching algorithm Vertical axis shows the true positive rate in out-of-sample matches between inventors and candidate matches in the US census for deciles of the top second-stage match scores across all of an inventor's match candidates, $\hat{y}^{(b)}_{ij^*(i)}$, in descending order of match quality. The horizontal axis shows the share of all inventors that were matched at the corresponding levels of $\hat{y}^{(b)}_{ij^*(i)}$ or higher.

census data. Because a multiplicity of good matches actually complicates selecting the right match, the second stage correctly lowers the estimated quality of such matches.

We repeat this analysis, matching each inventor to the two nearest census waves, except for patents after 1940, where inventors are matched only to the 1940 wave. Next, we select the match candidate with the highest $\hat{y}^{(b)}_i$ of all candidates in either wave, conditionally on the associated individual being at least 16 years old. In case one of the two matches is younger than 16 years, we select the other match, provided that $\hat{y}^{(b)}_i \geq .95$. If both matches involve individuals younger than 16 years, we do not match the inventor to the census. The match rate at both levels of accuracy over time is given in Fig. A1I.

A.4 Industrial research labs

Data about industrial research labs are extracted from the National Research Council's *Industrial Research Laboratories of the United States* surveys (Fig. A1C). Waves of this survey were conducted in 1920, 1927, 1931, 1933, 1938, 1940, 1946, 1948, 1950 and 1956. These surveys provide information on the name of each lab, its main activity, location, (managing) directors and important researchers.

We digitize these surveys using OCR algorithms. Next, we clean the names by removing common words such as 'company' and 'limited' and we calculate the Jaro-Winkler string similarity between every lab and patent assignee. Finally, we match records if the similarity is greater than 0.95 or if the similarity is greater than 0.90 and at least one of the words in the names matches perfectly.[33]

To determine when a given assignee started operating a research lab, we use information on the founding years from the 1940 and 1946 editions. For labs that are not reported in these editions, we set the founding date to the year of the first survey that mentions these labs. Finally, when labs are mentioned in editions that predate the founding year, we override the founding years by the year of the earliest edition that mentions the lab.

B Variable construction

B.1 Occupations

The census data contain harmonized occupation codes in the *occ1950* classification created by IPUMS from transcribed text originally noted down by census enumerators. With over 250 different job titles, the occ1950 classification is too detailed to use in the descriptive analyses of this study. Therefore, we aggregate these job titles into broader classes. One concern is that the prevalence of some occupational titles changes, even when jobs do not. In particular, the occupations that are reported in the census are ultimately based on self-reported jobs. As a consequence, the emergence of engineers in the census may not only be due to an actual expansion of engineers in the population, but also because more individuals start describing themselves as engineers. To address this concern, we group occupations into classes, based on an analysis of labor flows between the detailed occupations of the occ1950 classification.

To create labor-flow-based occupation groupings, we first create a matrix F with elements F_{ij} that contain the total number of individuals who move from occupation i to occupation j between two census waves. To do so, we start by selecting all individuals with non-missing occupation codes. Next, the level of detail at which occupations are described can vary across census waves. This is the case for *Professors* (codes 10-29) and *Scientists* (codes 61-69), which in some census waves are subdivided by field and in others only reported as aggregates. We combine these two sets of occupations into the two aggregate classes of professors and scientists.

Next, we construct for each pair of sequential decades the total number of individuals who listed occupation i in the first decade and occupation j in the next decade. This provides us with decadal cross-occupational labor flows. Note that due to the loss of the 1890 census in a fire, labor flows that start in 1880 end in 1900. Because the population grows from 23M in 1850 to 132M in 1940, these counts will be dominated by job switches in later decades. To remedy this, we normalize the flows in each decade by the sum total of all flows in that decade to express them as shares that add up to 1 in each decade. Next, we multiply these shares with the grand total of all flows across all decades divided

[33] Similar procedures using other similarity measures, e.g. using the Levenshtein distance, with different thresholds, yield similar results.

by eight, the number of decade pairs. This ensures that flows from each year are weighted equally.

To turn the flow matrix, F, into a matrix of flow intensities, we calculate *skill relatedness* (Neffke and Henning, 2013). Skill relatedness quantifies whether an observed flow between two occupations surpasses a random benchmark. Here, we follow van Dam et al. (2023), who develop an information-theoretic framework to generate Bayesian estimates of the amount of surprise involved in observing a flow F_{ij}, given the total inflows into occupation j and the total outflows from occupation i. To be precise, we will estimate the point-wise mutual information $pmi(i, j) = \log \frac{q_{ij}}{q_i q_j}$, where q_{ij} is the probability of observing an individual moving from occupation i to j, q_i the (marginal) probability that an individual leaves occupation i and q_j the marginal probability that an individual moves to occupation j. We collect these estimates in matrix PMI.

We convert this measure of relatedness into a measure of distance by subtracting matrix PMI from the maximum value across all its elements. Following recommendations of Muneepeerakul et al. (2013) and Li and Neffke (2023), we drop elements of PMI that are not significantly ($p = 0.01$) larger than 0, taking such occupations to be unrelated.[34] In distance matrix, D, these unrelated entries are set to a value of ten times the maximum observed distance.

Using this distance matrix, we estimate a 10-dimensional *Uniform Manifold Approximation and Projection* (UMAP, McInnes et al., 2018) embedding on which we project all occupations (see Fig. A1F for a projection on a 2-dimensional embedding). Finally, we use the Python implementation of Campello et al.'s (2013) *HDBSCAN* algorithm to cluster occupations at two different, nested hierarchical levels. The resulting two-level clusters are provided in Table B1. In the analyses, we only use the first, highest-level clusters.

Table B1: Occupational classification - flow-based grouping

MAILMEN

Mailmen: 84: Express messengers and railway mail clerks; 85: Mail carriers; 91: Telegraph operators; 93: Ticket, station, and express agents

TRANSPORT SERVICES

Continued on next page

[34] Note that we retain the direction of the flows and do not symmetrize this relatedness matrix. Moreover, diagonal elements are treated the same as any other element.

Table B1: Occupational classification - flow-based grouping (Continued)

Transport Services: 64: Conductors, railroad; 78: Baggagemen, transportation; 83: Dispatchers and starters, vehicle; 179: Brakemen, railroad; 180: Bus drivers; 182: Conductors, bus and street railway; 195: Motormen, street, subway, and elevated railway; 204: Switchmen, railroad; 225: Guards, watchmen, and doorkeepers; 230: Policemen and detectives; 236: Watchmen (crossing) and bridge tenders

PRINTING

Printing: 106: Bookbinders; 112: Compositors and typesetters; 116: Electrotypers and stereotypers; 117: Engravers, except photoengravers; 147: Photoengravers and lithographers; 151: Pressmen and plate printers, printing; 172: Apprentices, printing trades

LOGGING

Logging: 29: Foresters and conservationists; 52: Surveyors; 124: Inspectors, scalers, and graders, log and lumber; 201: Sawyers; 246: Lumbermen, raftsmen, and woodchoppers

WHITE COLLAR

Government: 67: Inspectors, public administration; 70: Officials and administrators (n.e.c.), public administration; 96: Auctioneers; 228: Marshals and constables; 233: Sheriffs and bailiffs

Financial: 1: Accountants and auditors; 79: Bank tellers; 80: Bookkeepers; 81: Cashiers; 87: Office machine operators; 94: Clerical and kindred workers (n.e.c.)

Other White Collar: 7: Authors; 10: Clergymen; 11: Professors; 17: Editors and reporters; 28: Farm and home management advisors; 32: Librarians; 44: Recreation and group workers; 45: Religious workers; 46: Social and welfare workers, except group; 48: Psychologists; 50: Miscellaneous social scientists; 53: Teachers (n.e.c.); 62: Buyers and department heads, store; 66: Floormen and floor managers, store; 71: Officials, lodge, society, union, etc.; 72: Postmasters; 74: Managers, officials, and proprietors (n.e.c.); 75: Agents (n.e.c.); 76: Attendants and assistants, library; 82: Collectors, bill and account; 89: Stenographers, typists, and secretaries; 95: Advertising agents and salesmen; 99: Insurance agents and brokers; 101: Real estate agents and brokers; 102: Stock and bond salesmen; 103: Salesmen and sales clerks (n.e.c.)

N.E.C.: 31: Lawyers and judges; 39: Personnel and labor relations workers; 47: Economists; 49: Statisticians and actuaries; 65: Credit men; 92: Telephone operators

FARMING

Farming: 60: Farmers (owners and tenants); 61: Farm managers; 238: Farm foremen; 239: Farm laborers, wage workers; 240: Farm laborers, unpaid family workers

Continued on next page

Table B1: Occupational classification - flow-based grouping (Continued)

Engineers

Engineers: 4: Architects; 16: Draftsmen; 18: Engineers, aeronautical; 19: Engineers, chemical; 20: Engineers, civil; 21: Engineers, electrical; 22: Engineers, industrial; 23: Engineers, mechanical; 24: Engineers, metallurgical, metallurgists; 25: Engineers, mining; 26: Engineers (n.e.c.); 181: Chainmen, rodmen, and axmen, surveying

Service Work

Health Care: 9: Chiropractors; 13: Dentists; 37: Optometrists; 38: Osteopaths; 40: Pharmacists; 42: Physicians and surgeons; 54: Technicians, medical and dental; 55: Technicians, testing; 57: Therapists and healers (n.e.c.)

Low Skill Service: 15: Dietitians and nutritionists; 34: Nurses, professional; 35: Nurses, student professional; 77: Attendants, physicians and dentists office; 97: Demonstrators; 104: Bakers; 184: Dressmakers and seamstresses, except factory; 187: Fruit, nut, and vegetable graders, and packers, except factory; 190: Laundry and dry cleaning operatives; 192: Milliners; 210: Housekeepers, private household; 211: Laundresses, private household; 212: Private household workers (n.e.c.); 213: Attendants, hospital and other institution; 214: Attendants, professional and personal service (n.e.c.); 216: Barbers, beauticians, and manicurists; 217: Bartenders; 218: Bootblacks; 219: Boarding and lodging house keepers; 220: Charwomen and cleaners; 221: Cooks, except private household; 222: Counter and fountain workers; 223: Elevator operators; 226: Housekeepers and stewards, except private household; 227: Janitors and sextons; 229: Midwives; 231: Porters; 232: Practical nurses; 235: Waiters and waitresses; 237: Service workers, except private household (n.e.c.)

N.E.C.: 143: Opticians and lens grinders and polishers; 244: Gardeners, except farm, and groundskeepers

Arts And Design

Arts And Design: 5: Artists and art teachers; 14: Designers; 41: Photographers; 98: Hucksters and peddlers; 121: Furriers; 122: Glaziers; 158: Tailors and tailoresses; 198: Photographic process workers

Entertainment

Artists: 2: Actors and actresses; 12: Dancers and dancing teachers; 33: Musicians and music teachers

Sports And Farm Work: 6: Athletes; 27: Entertainers (n.e.c.); 51: Sports instructors and officials; 58: Veterinarians; 63: Buyers and shippers, farm products; 191: Meat cutters, except slaughter and packing house; 241: Farm service laborers, self-employed

N.E.C.: 215: Attendants, recreation and amusement

Continued on next page

Table B1: Occupational classification - flow-based grouping (Continued)

Shipyard Work

Nautical: 69: Officers, pilots, pursers and engineers, ship; 178: Boatmen, canalmen, and lock keepers; 200: Sailors and deck hands; 242: Fishermen and oystermen

Laborers: 245: Longshoremen and stevedores; 248: Laborers (n.e.c.)

N.E.C.: 111: Cement and concrete finishers

Textile Workers

Textile Workers: 131: Loom fixers; 185: Dyers; 202: Spinners, textile; 207: Weavers, textile; 209: Operative and kindred workers (n.e.c.)

Blue Collar

Locomotive Workers: 118: Excavating, grading, and road machinery operators; 129: Locomotive engineers; 130: Locomotive firemen; 155: Stationary engineers; 203: Stationary firemen; 224: Firemen, fire protection

Oilers And Hoistmen: 113: Cranemen, derrickmen, and hoistmen; 196: Oilers and greaser, except auto

Mining And Railroads: 125: Inspectors (n.e.c.); 137: Mechanics and repairmen, railroad and car shop; 177: Blasters and powdermen; 193: Mine operatives and laborers; 194: Motormen, mine, factory, logging camp, etc.

Carpentry: 110: Carpenters; 140: Millwrights; 146: Pattern and model makers, except paper; 165: Apprentice carpenters

Metal Workers: 105: Blacksmiths; 120: Forgemen and hammermen; 123: Heat treaters, annealers, temperers; 141: Molders, metal; 152: Rollers and roll hands, metal; 188: Furnacemen, smeltermen and pourers; 189: Heaters, metal

Machine Workers: 127: Job setters, metal; 132: Machinists; 160: Tool makers, and die makers and setters; 167: Apprentice machinists and toolmakers

Repair Office Machinery: 43: Radio operators; 135: Mechanics and repairmen, office machine; 136: Mechanics and repairmen, radio and television

Electrical: 115: Electricians; 128: Linemen and servicemen, telegraph, telephone, and power; 142: Motion picture projectionists

Motorized Transportation: 3: Airplane pilots and navigators; 133: Mechanics and repairmen, airplane; 134: Mechanics and repairmen, automobile; 163: Apprentice auto mechanics

Continued on next page

Table B1: Occupational classification - flow-based grouping (Continued)

Motorized Transportation: 183: Deliverymen and routemen; 205: Taxicab drivers and chauffers; 206: Truck and tractor drivers; 243: Garage laborers and car washers and greasers; 247: Teamsters

Masonry: 108: Brickmasons, stonemasons, and tile setters; 149: Plasterers; 156: Stone cutters and stone carvers; 164: Apprentice bricklayers and masons

Other Construction: 144: Painters, construction and maintenance; 145: Paperhangers; 150: Plumbers and pipe fitters; 159: Tinsmiths, coppersmiths, and sheet metal workers; 161: Upholsterers; 170: Apprentices, building trades (n.e.c.); 173: Apprentices, other specified trades; 174: Apprentices, trade not specified; 197: Painters, except construction or maintenance

N.E.C.: 90: Telegraph messengers; 100: Newsboys; 107: Boilermakers; 109: Cabinetmakers; 114: Decorators and window dressers; 138: Mechanics and repairmen (n.e.c.); 148: Piano and organ tuners and repairmen; 154: Shoemakers and repairers, except factory; 162: Craftsmen and kindred workers (n.e.c.); 166: Apprentice electricians; 168: Apprentice mechanics, except auto; 169: Apprentice plumbers and pipe fitters; 171: Apprentices, metalworking trades (n.e.c.); 175: Asbestos and insulation workers; 176: Attendants, auto service and parking; 186: Filers, grinders, and polishers, metal; 199: Power station operators; 208: Welders and flame cutters; 234: Ushers, recreation and amusement

N.E.C.

N.E.C.: 8: Chemists; 30: Funeral directors and embalmers; 36: Scientists; 56: Technicians (n.e.c.); 59: Professional, technical and kindred workers (n.e.c.); 68: Managers and superintendents, building; 73: Purchasing agents and buyers (n.e.c.); 86: Messengers and office boys; 88: Shipping and receiving clerks; 119: Foremen (n.e.c.); 126: Jewelers, watchmakers, goldsmiths, and silversmiths; 139: Millers, grain, flour, feed, etc.; 153: Roofers and slaters; 157: Structural metal workers

The flow-based clustering of occupations yields one group of occupations that contains all engineering occupations. However, this group also includes three further occupations that have strong labor-flow connections to engineering jobs: *Architects*, *Draftsmen* and *Chainmen, rodmen, and axmen, surveying* (i.e., occupations related to land surveying).

To check the robustness of our results, we also consider a different way of grouping occupations into broad sectors, using the hierarchical structure of the occ1950 codes. Because of our interest in the rise of engineers among inventors, we subdivide the sector "Professional, technical," which contains the engineering occupations, into a number of smaller subclasses. Furthermore, we separate apprentices as a subclass of the sector "Operatives". Table B2 summarizes these groupings.[35]

[35]The missing category consists of the following occupation codes: *Members of the armed services*; *Not yet classified*; *Keeps house/housekeeping at home/housewife*; *Imputed keeping house* (1856-1900);

Table B2: Occupational classification - hierarchical grouping

Codes	**Class name**
0 - 99	Professional, technical
41 - 49	Engineers
7, 10, 12-19, 23-29, 61-69, 81-84	Scientists
1, 4, 6, 31, 51, 57, 74, 77, 91	Artists
8, 32, 34, 58-59, 70-71, 73, 75, 97-98	Health care
0, 9, 55, 72	Office professionals
2-3, 33, 92, 94-96	Technical workers
36, 52-54, 56, 76, 78-79, 93, 99	Other professionals
100 - 123	Farmers
200 - 290	Managers, Officials, and Proprietors
300 - 390	Clerical and Kindred
400 - 490	Sales workers
500 - 594	Craftsmen
600 - 690	Operatives
600 - 621	Apprentices
622 - 690	Operatives
700 - 720	Service Workers (private household)
730 - 790	Service Workers (not household)
810 - 840	Farm Laborers
910 - 970	Laborers
595, 979 - 999	Missing

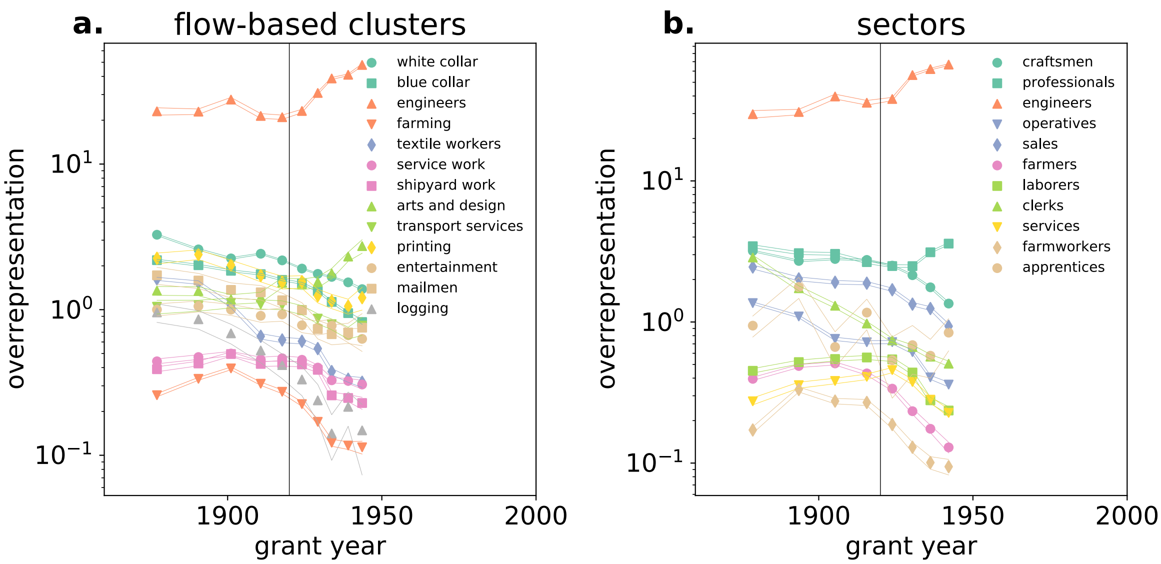


Figure B1: Overrepresentation of broad occupation classes in patent records. **a**: Overrepresentation of flow-based occupational clusters. **b**: Overrepresentation of occupational sectors.

We can now determine which occupations are overrepresented in the patent records, by comparing their shares among inventors to their shares in the working-age population. To do so, we calculate the following quantity:

$$\rho_{ot} = \frac{N_{ot}/\sum_{o'} N_{o't}}{POP_{ot}/\sum_{o''} POP_{o''t}} \tag{4}$$

where N_{ot} is the number of patents by inventors with occupation o in period t and POP_{ot} the number of working-age individuals in the census records with occupation o in period t. To determine POP_{ot} outside census years, we interpolate linearly between two census waves.

Fig. B1 plots the overrepresentation of flow-based clusters (B1a) and of broad occupational sectors (B1b) over time. It shows that engineers had always been heavily overrepresented among inventors, but that this overrepresentation increases even more in the 1920s. Another group of occupations that emerges in this period can be identified in the flow-based clusters: arts and design occupations. In contrast, blue collar workers and craftsmen become increasingly less overrepresented over time.

Helping at home/helps parents/housework; *At school/student*; *Retired*; *Unemployed/without occupation*; *Invalid/disabled w/ no occupation reported*; *Inmate*; *New Worker*; *Gentleman/lady/at leisure*; *Other non-occupational response*; *Occupation missing/unknown*; and *N/A (blank)*.

B.2 Distinguishing family teams

There are two ways in which we can assess whether co-inventors are related in our dataset. First, for patents until 1953, the matched census records allow us to construct family ties. This approach has the advantage that it allows identifying a large variety of family ties, such as siblings, cousins, uncles and nephews, grandparents and grandchildren, etc.. However, the construction of these family ties relies on matching inventors to the census, and for more distant family ties, linking census individuals across multiple census waves. However, both of these types of linkages are imperfect, complicating the use of census linkages to determine family ties.

Instead, therefore, we rely on a second approach and identify family ties based on shared last names. That is, we assume that two inventors are related if they share the same last name. Because not all family ties are necessarily associated with shared last names, this approach results in an under-count of family-based teamwork. Moreover, even when two inventors share the same last name, they are not necessarily related. Such spurious family ties will be particularly common for inventors with common last names. The nature of this problem changes over time. In the 1940s, the most common last name is *Smith* and inventors with that last name hold 0.78% of all patents. This is followed by *Johnson* (0.51%), *Miller* (0.47%), *Brown* (0.37%) and *Anderson* (0.35%). From the 1980s on inventors with last names that are common in certain East-Asian countries start dominating this list. For instance, the top 5 surnames on patents in 2010 is composed of *Kim* (1.17%), *Lee* (1.15%), *Chen* (.86%), *Wang* (0.77%) and *Park* (0.61%). This reflects the high frequency of certain surnames in specific countries (especially in Asia). For instance, we find that 19.9% of inventors residing in Korea record the surname *Kim* and the five most common Korean surnames account for 49.3% of all inventors in Korea. Similarly high percentages are found for inventors residing in Taiwan (top 5 surnames: 33.1% of total), China (30.5%), Malaysia (19.7%), Hong Kong (18.8%) and Denmark (14.4%).

To correct our estimates of family-based team-patents, we create a random benchmark in which we shuffle inventor surnames across patents. However, we do this in such a way that inventors whose name suggests a certain "ethnicity" can only swap patents with inventors with the same inferred ethnicity. This acknowledges that inventors may be more likely to form teams within an ethnic community.

To construct the ethnicity variable, we drop all inventors residing in the US, Canada or Australia, because the populations of these countries have always included many migrants from all over the world. Furthermore, because China, Taiwan, Hong Kong, Singapore and Malaysia share many of the same last names, we group the inventors residing in these countries into a single category. Similarly, we group all inventors in Spain and Spanish-speaking Latin American countries, as well as Germany, Austria and Liechtenstein. Finally, countries with fewer than 100,000 patents are grouped in a residual category, *RoW*. For every surname, we now determine the most likely geographical origin as the country

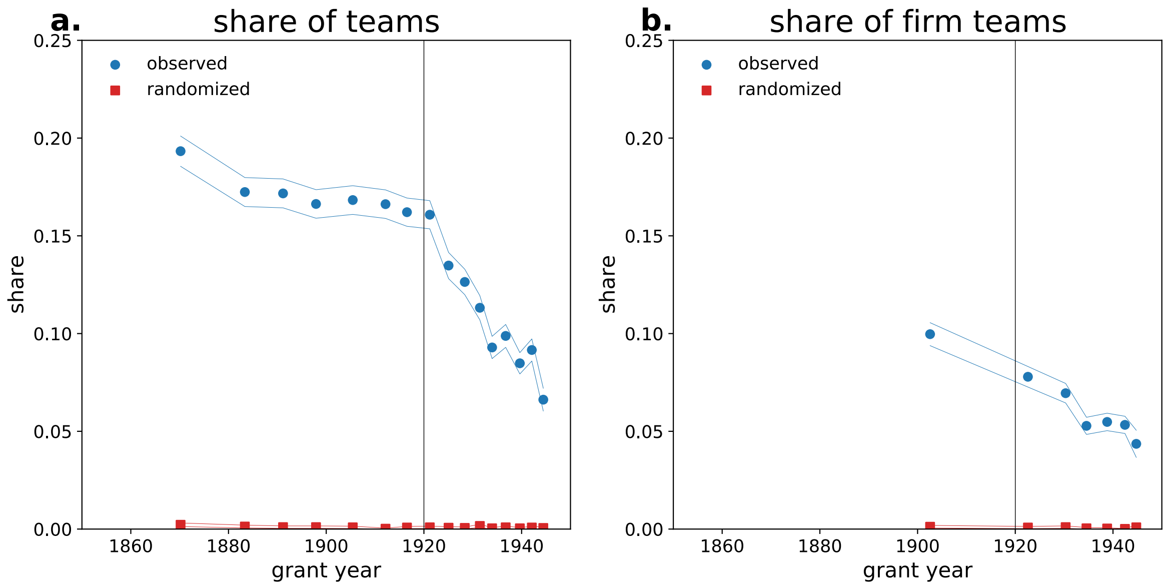


Figure B2: Null model comparison family teams. **a**: Share of teams that list inventors with the same last name. **b**: Share of firm patents that list inventors with the same last name. Blue: shares in observed data, red: shares in reshuffled data. Thin lines display 95% confidence intervals.

that accounts for the greatest share of inventors with that surname.[36]

Finally, we randomize surnames across inventor-patent combinations. To do so, we shuffle the surname column within (grant-year, name-origin) pairs. Fig. B2 shows that for the period reported in Fig. 8, these null-model shares are negligible.

B.3 Disambiguated inventor dyads

Disambiguating inventor names is complicated, due to the size of the US population and the limited information that exists about each inventor. As long as first and last name combinations are rare, inventor disambiguation across patents would in principle be possible. However, the same individual may report slightly different names on different patents and in the census, a problem compounded by spelling mistakes and OCR errors. Because of this, we consider disambiguating inventors across patents to be beyond the scope of this paper.

However, disambiguating *pairs* of co-inventors is much easier. To see this, let p_n denote the share of individuals in the US census with surname n. Consequently, if we randomly

[36]This procedure leads to some problems for Indian surnames, which are also very common in the UK. For these names, we make a number of manual adjustments that can be found in the code in the accompanying Supporting Material.

pick two individuals from the census, the probability of observing a pair with surnames a and b equals $p_a p_b$. Furthermore, we drop inventors whose last name is very common, that is, we drop inventors with last names that are found over 500k times across all 9 US censuses. Because there are roughly 650M records in these censuses, if we were to pick two random individuals, the probability of drawing any given combination of two last names is at most $\left(\frac{600k}{650M}\right)^2 = \frac{1}{1.69M}$.

However, in family-based teams, inventors often share the same last name. That is, for family-based pairs, surname draws are not uncorrelated. To disambiguate these inventor pairs, we drop all same-surname pairs if the surname is repeated over 50k times across all censuses. Finally, we drop inventors with East-Asian or Indian surnames. This yields a total of 133k unique inventor dyads across all patents granted between 1856 and 1949.

To explore the validity of this approach, we calculate the lifespan of each inventor dyad. That is, for dyads that are associated with more than one patent, we calculate the time between their first and last patent. We find that 99.9% of dyads that are listed on more than one patent have a lifespan shorter than 24 years, which we deem plausible.

In Fig. 8c, we use this disambiguated dyads sample, S_{disamb}, to analyze repeat collaborations. To do so we calculate for each decade the share of patents that resulted from repeated collaborations. That is, we sum all patents by the disambiguated dyads in S_{disamb} that patent more than once in a given decade and divide this sum by the total number of patents that were granted to any disambiguated dyads in this decade.

B.4 Inventor age

Linking inventors to census records allows us to calculate the age of an inventor at the time that a patent was granted. This reveals an increase in the average age of inventors in the late 19th and early 20th century that accelerates in the 1920s (Fig. B3).

Because we don't disambiguate inventor names, our data refer to the average age of inventors on any patent, not the age when they start patenting as reported by, for instance, Jones (2009). For comparison, we therefore use data collected by Kaltenberg et al. (2021) that allows us to calculate the equivalent number for the period 1975-2018. In this period, average inventor age also rises, from about 40 to 50. Interestingly, however, the average age of inventors in 1975 (about 40) is well below the age we record for the pre-WWII period. Inventor age must therefore have fallen some time after WWII. Jung and Ejermo (2014) document a similar fall in inventor age in the late 1990s and early 2000s for Sweden, attributing this to a change in technological paradigm.

B.5 Gender

To assess the gender of inventors, we could rely on the matched census records. This approach has two disadvantages. First, we would only be able to determine a gender for the inventors that we can accurately match to the census records. Second, and more importantly, our census records only allow us to match inventors until 1945. Instead,

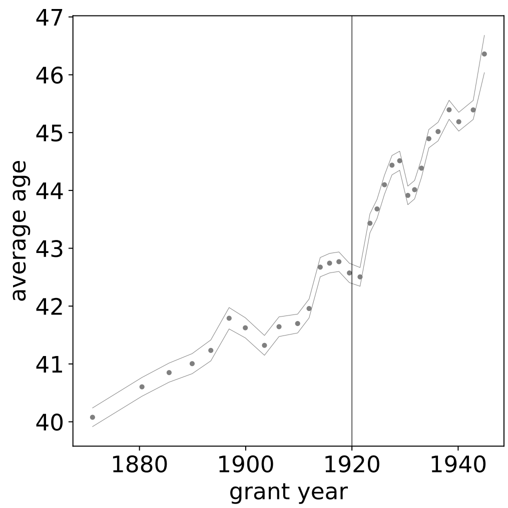


Figure B3: Average age of inventors.

therefore, we infer the most likely gender of an inventor based on their first names (see Fig. A1G).

To do so, we calculate which share of individuals with a given first name listed their gender as female or male in the US censuses between 1850 and 1940. Whenever this percentage (in the US population not among inventors) exceeds 90%, we associate inventors with the most common gender for their first names in the census records. Otherwise, we set gender to missing. We also set gender to missing if the first name is shorter than 3 characters, to avoid inferring genders from initials misidentified as first names. Fig. B4 shows how the share of patents granted to women in this sample changes over time.

To assess the robustness of our analysis, we repeat the analysis of Fig. 13 in the main text for inventors whose first names were often used by both men and women (i.e., where fewer than 90% of individuals listed the same gender). For these cases, we now decide on the gender by majority vote, i.e., inventors are classified as female if at least 50% of census records with these first names list female as a gender, and otherwise as male. Fig. B5a shows the evolution of female inventor shares using this more ambiguous sample, corroborating the patterns described in the main text.

B.6 National origins of surnames

To infer the national origins of surnames, we rely on predictions from models that have been trained on datasets that contain surnames and their associated nationality by Sood and Laohaprapanon (2018), using their *ethnicolr* Python package. We focus on Hispanic and East-Asian surnames, because these are relatively easy to identify. Hispanic surnames are identified from models trained on data from the 2000 US Census, and East-Asian

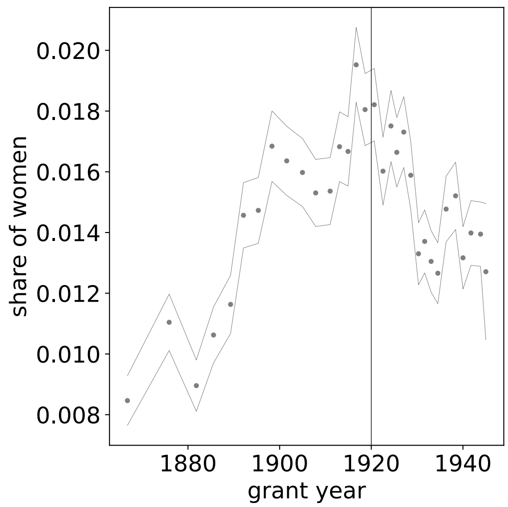


Figure B4: Share of women. Share of women among all inventors whose first name allows to infer gender with high confidence.

surnames from models trained on Wikipedia data (Ambekar et al., 2009).

Both algorithms yield an accuracy score that can be interpreted as the estimated probability that a given surname is of Hispanic or East-Asian origin respectively. Whenever this score doesn't surpass 0.9, we refrain from labeling the surname as Hispanic or East-Asian. As a consequence, the estimated inventor shares are undercounts. However, our main interest is in how these shares differ between firm-based and standalone patents and how this difference evolves over time. For this purpose, we only need to assume that the degree of undercounting does not vary systematically over time or with whether or not an inventor patents on behalf of a firm.

To assess how robust this analysis is, we repeat it for individuals with surnames that are *probably* of Hispanic or East-Asian origin (Fig. B5b and B5c). That is, we calculate the shares of inventors whose surname suggest one of these two origins with a probability that is nonzero, yet below 90%. The findings corroborate those reported in the main text in Fig. 12b.

C Graphs

Many graphs in the main text display how average characteristics of patents and their inventors or assignees change over time. When choosing time windows over which to average observations, we need to balance a sufficiently high temporal resolution with reasonably precise point estimates of these means. This precision depends on the number of observations in a given time interval. Because the number of observations, i.e., the

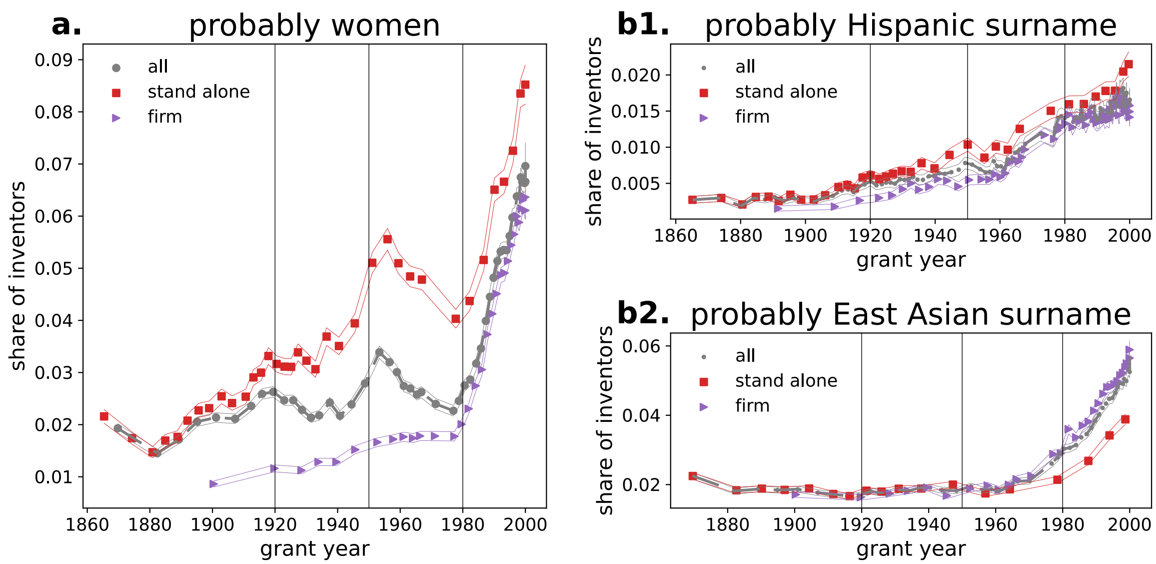


Figure B5: Robustness of female participation analysis. **a**: Share of inventors with a first name that is used in at least 50% of census records by individuals that list their gender as female, dropping all individuals whose first name coincides with male or female genders in at least 90% of records. **b1**: Share of surnames that are probably of Hispanic origin (i.e., with positive but below 90% probability). **b2**: Share of surnames that are probably of East-Asian origin (i.e., with positive but below 90% probability). Gray: all patents, red: standalone patents, purple: firm patents.

number of patented inventions, grows roughly exponentially, the width of the ideal time window shrinks over time.

To resolve this, we create windows, not based on time, but on temporal rank. That is, we sort all patents by their exact grant date and then divide the data into groups of identical size, each containing N observations. Next, we calculate the average grant year associated with each group and use this average year as the horizontal coordinate in the graph. Meanwhile, estimated quantities are are plotted along the vertical axis, with confidence intervals calculated as $\pm 1.96 \times$ (standard error of the mean). Note that the group size differs across, and sometimes even within graphs. This is due to the fact that some types of observations are more numerous than others. For instance, patents by research labs are rather rare, especially in earlier years, whereas firm-based patents are relatively common in most decades. Therefore, striking a useful balance between precision and time resolution leads to different group sizes for these two types of patents.

D Additional results

D.1 Learning curves

Section 4.2.1 of the main text shows how the novelty, vintage and complexity of patented technologies changes with age, based on an analysis of the 3-digit technologies listed on a patent. Here, we repeat this analysis using 6-digit technology codes. The findings for vintage and complexity are very similar to those based on 3-digit codes (Fig. D1). However, as shown in the first column of Table D1, we now find that age is statistically significantly associated with increases in the likelihood that a patent lists a novel combination of 6-digit codes. This suggests that creativity (in the sense of patenting new technological combinations) increases with work experience, but only in more incrementally novel combinations, not the radically new combinations of broad 3-digit classes. However, we should note that the parameter estimate implies that between the age of 15 and 35, the likelihood of patenting a novel combination of 6-digit technology codes rises by about 3pp. Compared to the baseline probability of patenting novel combinations of between 50% and 75%, this increase is modest at best.

Table D2 analyzes whether the 20th century learning curves differ between inventors in firms and labs than for standalone inventors. These findings suggest that the learning curves we observe in 20th century cohorts (in terms of learning how to combine more and more recent technologies) are mostly driven by corporate research.

Table D3 adds control variables for having an engineer on the patent or for organizational context to the analysis of Table 3 of the main text. Adding these control variables does not change the substance of our conclusions, although the addition of the engineer dummy somewhat reduces point estimates and renders two of them insignificant.

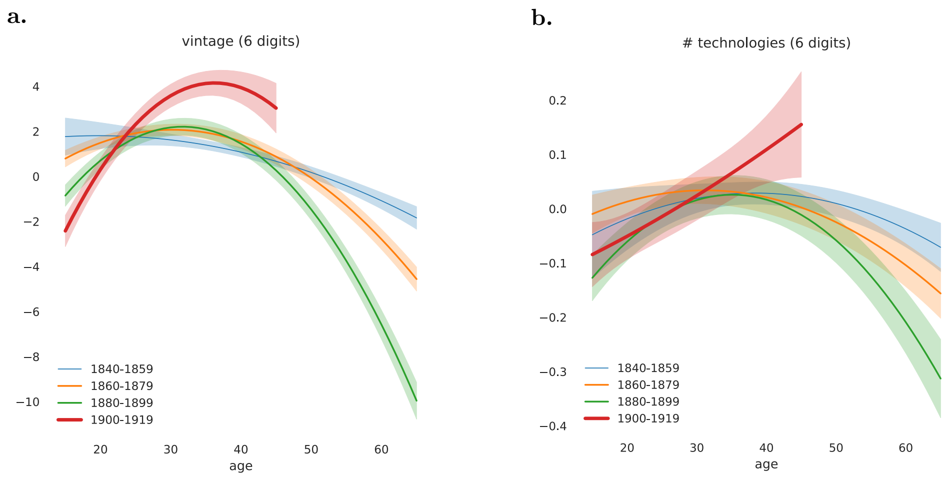


Figure D1: Learning curves for 6-digit technology classes. Within-sample fit of learning curves for four different cohorts: 1840-1959 (blue), 1860-1879 (orange), 1880-1899 (green), 1900-1919 (red). The left panel shows learning curves in terms of technological vintage, the right panel in terms of the number of technologies on a patent. The shaded areas represent 95% confidence intervals, using robust standard errors.

D.2 Geography

In this section, we analyze changes in the geography of innovation. In Fig. D2a, we show changes in the geographical concentration of patenting. To do so, we calculate the effective number of cities in US patenting. That is, we calculate the exponentiated entropy of the distribution of patents across cities:[37]

$$H_t^p = e^{-\sum_{r\in R} \sigma_{rt}^s \log \sigma_{rt}^s}, \tag{5}$$

where σ_{rt}^s is the share of patents of system $s \in \{1, 2\}$ filed by inventors residing in city r in period t. We normalize H_t^p by the analogous quantity for the distribution of population across cities to account for the large population shifts taking place in this period.

Relative to how population spreads out across the US, patenting becomes more spatially concentrated over the course of the 19th century. However, in the 20th century, this process reverses and invention starts diffusing to more and more cities, first in system 2 and then somewhat later also in system 1.

This pattern is mimicked by the role that large cities play in innovation. Fig. D2b shows the share of patents filed by inventors in the 25 largest cities in the US. Whereas in the 19th century, these cities slowly become less dominant in the patent record, at the turn of the century, this trend reverses. By 1940, large cities have regained their mid-19th

[37] The "effective number of cities", H_t^p, quantifies the number of cities that would yield the same observed entropy, had patenting been distributed equally across them (Jost, 2006).

Table D1: Learning curves (6 digits)

Cohort	Novelty	Vintage	Complexity
1840	-0.0012 (0.0014)	-0.0078 (0.0340)	-0.0000 (0.0034)
1860	-0.0005 (0.0005)	-0.0004 (0.0156)	-0.0011 (0.0015)
1880	0.0016 (0.0004)***	0.0495 (0.0169)**	0.0040 (0.0014)**
1900	0.0013 (0.0005)**	0.2417 (0.0229)***	0.0037 (0.0020)

Parameter estimates for age effects in eq. (1). Cohorts are listed in rows, dependent variables in columns – *Novelty*: dummy variable for whether or not the patent lists a new combination of technologies; *Vintage*: vintage of most recent technology class listed on the patent; *Complexity*: number of distinct technology classes listed on the patent, where technology classes are recorded at the 6-digit level. Robust standard errors are in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

century primacy. Again, the reversal is led by system 2, with system 1 following some years later.

Fig. D2c further illustrates how system 2 leads a geographical shift in invention, that is later followed by system 1. To construct this figure, on each time window, we first limit the set of cities to those that account for at least 0.1% of the overall US population. Next, we divide the share of patents that a city holds by the share of the US population it hosts. This yields a locational vector that captures the overrepresentation of a city in inventive activity:

$$\rho_{cst} = \frac{N_{cst}/\sum_{c'} N_{c'st}}{POP_{ct}/\sum_{c''} POP_{c''t}}, \tag{6}$$

where N_{cst} is the number of patents in city c, system s and period t, and POP_{ct} the population of city c and period t. To determine POP_{ct} between census years, we interpolate linearly between two census waves.

We repeat this procedure once for the patents of system 1 and once for the patents of system 2. Next, we calculate the correlation between the locational vectors of system 1 and 2. Plotting this correlation in Fig. D2c shows that the locational vectors of system 1 and 2 diverge until the mid-1920s, but then start converging again.

D.3 Foreign-born inventors

Fig. D3a shows that immigrants are overrepresented in the population of inventors for most of the period we study. However, this is not uniformly the case. In fact, in the late 19th century, immigrant inventors' overrepresentation diminishes and for a short period, foreign-born inventors are *underrepresented* in the patent records. From the 1920s on, foreign-born inventors once again become strongly overrepresented. Moreover, as shown in Fig. D3b, which plots the overrepresentation of inventors from the six most important countries of birth, this aggregate dynamic hides much variation. For instance, unlike

Table D2: Learning curves by organizational context)

A: Standalone inventors			
Cohort	**Novelty**	**Vintage**	**Complexity**
1840	-0.0011 (0.0006)	-0.0087 (0.0272)	-0.0015 (0.0018)
1860	0.0000 (0.0002)	-0.0069 (0.0114)	0.0003 (0.0007)
1880	0.0003 (0.0002)	-0.0149 (0.0129)	0.0012 (0.0007)
1900	-0.0005 (0.0004)	0.0954 (0.0244)***	-0.0003 (0.0012)
B: Firms without labs			
Cohort	**Novelty**	**Vintage**	**Complexity**
1840	-0.0023 (0.0022)	-0.1737 (0.1240)	-0.0061 (0.0061)
1860	0.0003 (0.0006)	-0.0246 (0.0382)	0.0026 (0.0017)
1880	0.0003 (0.0004)	-0.0161 (0.0287)	0.0024 (0.0013)
1900	0.0005 (0.0004)	0.3573 (0.0307)***	0.0042 (0.0013)**
C: Labs			
Cohort	**Novelty**	**Vintage**	**Complexity**
1860	0.0039 (0.0044)	-0.5315 (0.3774)	-0.0073 (0.0154)
1880	-0.0001 (0.0012)	-0.0998 (0.0108)	0.0026 (0.0042)
1900	0.0001 (0.0007)	0.4502 (0.0623)***	0.0051 (0.0025)*

Parameter estimates for age effects in equation (1). Cohorts are listed in rows, dependent variables in columns – *Novelty*: dummy variable for whether or not the patent lists a new combination of technologies; *Vintage*: vintage of most recent technology class listed on the patent; *Complexity*: number of distinct technology classes listed on the patent, where technology classes are recorded at the 3-digit level. Robust standard errors are in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

UK-born inventors, inventors born in Sweden or Germany, both of which industrialized relatively late, rose to prominence only in the early 20th century. Similarly, we see an inflection point in the time-series of Russian-born inventors after the Russian revolution (Fig. D3b).

D.4 Regression results

This section presents full tables and robustness checks for the regression analyses in the main text. The regression models aim to associate the likelihood that a patent introduces a new technological combination to the labor inputs (engineers and teams) and coordination modes (firms and industrial research labs) of system 2. To proxy labor inputs, we ask whether or not at least one inventor on the patent is an engineer. Furthermore, we ask whether the patent was produced by a solo inventor or by a team of inventors. To assess the mode of coordination behind a patent, we distinguish between patents that were assigned to firms, patents that were assigned to firms with research labs and patents that were assigned to individuals (either to the inventors themselves or other individuals). Our

Table D3: Learning curves with control variables

A: With engineer dummy			
Cohort	**Novelty**	**Vintage**	**Complexity**
1840	-0.0011 (0.0006)	0.0085 (0.0269)	-0.0015 (0.0017)
1860	0.0001 (0.0002)	-0.0108 (0.0114)	0.0007 (0.007)
1880	0.0003 (0.0002)	0.0101 (0.0125)	0.0016 (0.0006)**
1900	-0.0002 (0.0003)	0.1828 (0.0201)***	0.0017 (0.0009)
B: With standalone / firm / lab dummies			
Cohort	**Novelty**	**Vintage**	**Complexity**
1840	-0.0012 (0.0006)*	-0.0512 (0.232)*	-0.0020 (0.0017)
1860	0.0001 (0.0002)	0.0153 (0.0096)	0.0009 (0.0007)
1880	0.0003 (0.0002)	0.0284 (0.0099)**	0.0017 (0.0006)**
1900	-0.0000 (0.0003)	0.1894 (0.0167)***	0.0023 (0.0009)*

Parameter estimates for age effects in equation (1). Cohorts are listed in rows, dependent variables in columns – *Novelty*: dummy variable for whether or not the patent lists a new combination of technologies; *Vintage*: vintage of most recent technology class listed on the patent; *Complexity*: number of distinct technology classes listed on the patent, where technology classes are recorded at the 3-digit level. Robust standard errors are in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

primary goal is to assess whether certain types of organization enhance the capacity of innovation inputs to produce novel combinations. That is, we are interested in interaction effects between inputs and organizational contexts.

Furthermore, we distinguish between effects on radical and on incremental novelty. To do so, we run all analyses twice, once using novelty at the level of 3-digit technology codes and once at the level of 6-digit technology codes. Finally, to determine whether estimated parameters change over time, we split our data into four periods: the period before the take-off of research labs (1856-1919), the period in which research labs start dominating invention (1920-1945) and the periods 1946-1968 and 1976-2000. Results are summarized in Tables D4-D8.

D.5 1856-1945

To assess the robustness of these findings, we run models with year fixed effects (reported in the main text) and year-NBER-sector fixed effects, where the latter interact year dummies with the 6 technological sector dummies in Hall et al. (2001), using the following specification:

1. *engineers*: dummy for whether or not one of the inventors is an engineer;
2. *team*: dummy for whether or not the patent lists a team of inventors;

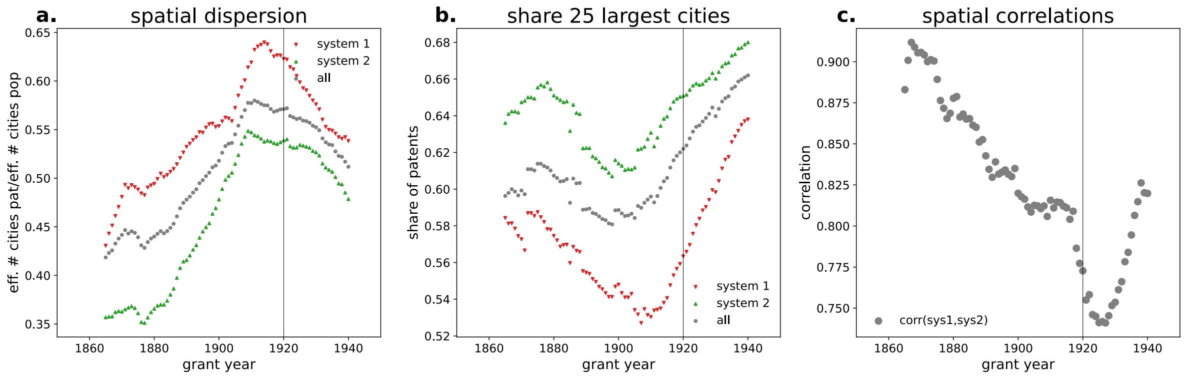


Figure D2: Geography. **a**: Concentration of inventive activity: effective number of cities – defined as e^H, where H is the entropy of the distribution of patents across cities – divided by the effective number of cities in terms of population. **b**: Share of patents in the 25 most populous cities in the US. **c**: Correlation between vectors that reflect the overrepresentation of patenting across cities vis-à-vis their population. Gray: all patents, red: system 1 patents, green: system 2 patents.

3. *eng*×*team*: interaction of 1 and 2;
4. *firm*: dummy for whether or not the patent was assigned to a firm;
5. *RnDlab*: dummy for whether or not the patent was assigned to a firm with a research lab;
6. *eng*×*firm*: interaction of 1 and 4;
7. *eng*×*RnDlab*: interaction of 1 and 5;
8. *eng*×*team*×*firm*: interaction of 1, 2 and 4;
9. *eng*×*RnDlab*: interaction of 1 and 5;
10. *team*×*RnDlab*: interaction of 2 and 5;
11. *eng*×*team*×*RnDlab*: interaction of 1, 2 and 5;

In these models we restrict the sample when fitting these models to patents where we can match at least one inventor to census records.

The main focus of our analysis is on the role of engineers and teams as labor inputs into the invention process and how these labor inputs are enhanced by the organizational context in which they are deployed. Gauging such interaction effects directly from the

a. **b.**

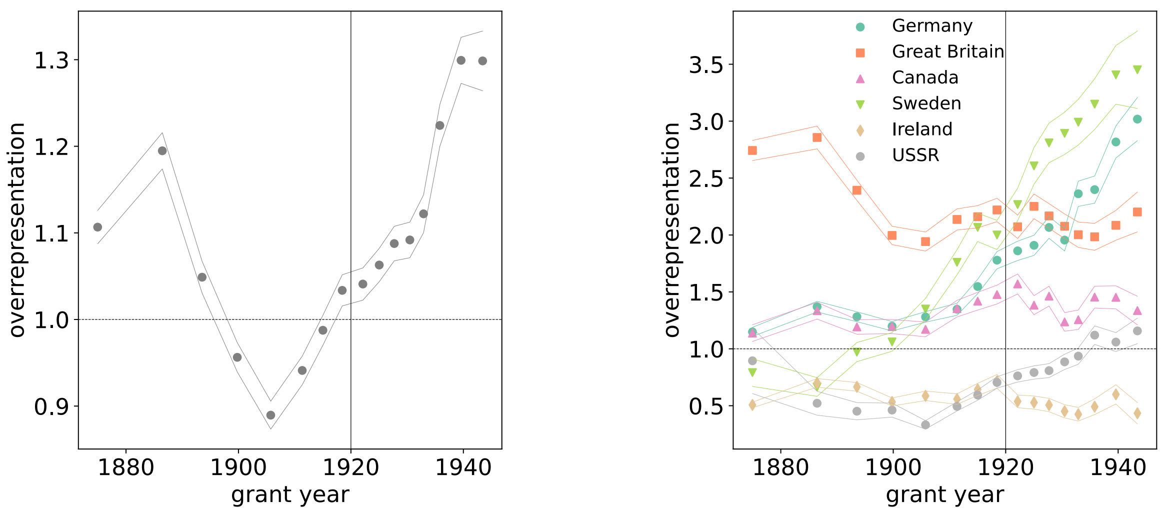


Figure D3: Demographics. **a**: Overrepresentation of foreign-born individuals in patent records: $\sigma^p_{ft}/\sigma^{pop}_{ft}$, where σ^p_{ft} is the share of patents held by foreign-born inventors in year t and σ^{pop}_{ft} the share of foreign-born individuals in the US population in year t, interpolating between census waves in non-census years. **b**: Overrepresentation of foreign-born individuals in patent records for the six largest countries of origins among US inventors: $\sigma^p_{bt}/\sigma^{pop}_{bt}$, where σ^p_{bt} is the share of patents in year t held by inventors born in country b and σ^{pop}_{bt} their share in the US population in year t, interpolating between census waves in non-census years.

Table D4: Novelty regression - Model A (year fixed effects)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.010***		0.010***	0.017***	0.017***	0.017***	0.015***		0.016***	0.021***	0.021***	0.018***	0.009***		0.008***	0.015***	0.015***	0.017***
	(0.001)		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.003)	(0.003)	(0.001)		(0.001)	(0.003)	(0.003)	(0.003)
team		0.005***	0.005***	0.002*	0.002*	0.002*		0.001	0.001	0.001	0.001	0.000		0.008***	0.008***	0.003*	0.003*	0.004**
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
eng×team			0.002	-0.001	-0.001				-0.014*	-0.025***	-0.025***				0.003	0.011	0.011	
			(0.003)	(0.006)	(0.006)				(0.006)	(0.007)	(0.007)				(0.003)	(0.008)	(0.008)	
firm				0.010***	0.006***	0.006***				0.005***	0.004***	0.004***				0.013***	0.007***	0.007***
				(0.000)	(0.000)	(0.000)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
eng×firm				-0.014***	-0.014***	-0.014***				-0.015**	-0.014**	-0.010*				-0.013***	-0.014***	-0.016***
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.003)	(0.004)	(0.003)
team×firm				0.006***	0.006***	0.006***				0.003	0.003	0.003				0.005**	0.006**	0.005*
				(0.001)	(0.002)	(0.001)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)
eng×team×firm				0.003	0.001					0.027*	0.027*					-0.010	-0.012	
				(0.007)	(0.007)					(0.013)	(0.013)					(0.009)	(0.009)	
RnDlab					0.021***	0.020***					0.020***	0.020***					0.020***	0.020***
					(0.001)	(0.001)					(0.003)	(0.003)					(0.001)	(0.001)
eng×RnDlab					-0.007*	-0.006*					-0.015	-0.016					-0.006	-0.004
					(0.003)	(0.003)					(0.013)	(0.012)					(0.003)	(0.003)
team×RnDlab					-0.002	-0.001					0.004	0.003					-0.003	-0.002
					(0.003)	(0.003)					(0.011)	(0.010)					(0.003)	(0.003)
eng×team×RnDlab					0.007						-0.011						0.008	
					(0.008)						(0.033)						(0.008)	
Constant	0.054***	0.054***	0.054***	0.051***	0.050***	0.050***	0.044***	0.045***	0.044***	0.044***	0.044***	0.044***	0.064***	0.064***	0.063***	0.058***	0.058***	0.058***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.046***		0.045***	0.070***	0.072***	0.069***	0.063***		0.065***	0.086***	0.086***	0.081***	0.041***		0.041***	0.064***	0.065***	0.063***
	(0.002)		(0.002)	(0.004)	(0.004)	(0.004)	(0.005)		(0.005)	(0.007)	(0.007)	(0.006)	(0.002)		(0.003)	(0.005)	(0.005)	(0.005)
team		0.019***	0.018***	0.009***	0.010***	0.009***		0.006**	0.006**	0.003	0.003	0.003		0.031***	0.030***	0.021***	0.021***	0.020***
		(0.001)	(0.001)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
eng×team			-0.004	-0.017	-0.017				-0.012	-0.036*	-0.036*				-0.010	-0.013	-0.013	
			(0.005)	(0.010)	(0.010)				(0.014)	(0.018)	(0.018)				(0.006)	(0.013)	(0.013)	
firm				0.034***	0.021***	0.021***				0.017***	0.012***	0.012***				0.046***	0.028***	0.028***
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.001)	(0.001)	(0.001)
eng×firm				-0.048***	-0.053***	-0.051***				-0.063***	-0.069***	-0.061***				-0.046***	-0.048***	-0.048***
				(0.005)	(0.005)	(0.005)				(0.011)	(0.011)	(0.010)				(0.006)	(0.006)	(0.006)
team×firm				0.018***	0.020***	0.019***				0.013*	0.016**	0.017**				0.008*	0.011**	0.010**
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.004)	(0.004)	(0.004)
eng×team×firm				0.014	0.011					0.054	0.057					0.008	0.001	
				(0.012)	(0.013)					(0.029)	(0.030)					(0.014)	(0.016)	
RnDlab					0.071***	0.070***					0.075***	0.075***					0.067***	0.067***
					(0.002)	(0.002)					(0.006)	(0.006)					(0.002)	(0.002)
eng×RnDlab					-0.020***	-0.016**					0.023	0.019					-0.020**	-0.014**
					(0.006)	(0.005)					(0.027)	(0.025)					(0.006)	(0.005)
team×RnDlab					-0.016***	-0.013**					-0.023	-0.025					-0.018***	-0.014**
					(0.005)	(0.004)					(0.021)	(0.020)					(0.005)	(0.005)
eng×team×RnDlab					0.022						-0.028						0.028*	
					(0.013)						(0.071)						(0.013)	
Constant	0.527***	0.527***	0.525***	0.515***	0.514***	0.514***	0.476***	0.476***	0.475***	0.472***	0.472***	0.472***	0.583***	0.582***	0.579***	0.559***	0.558***	0.558***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table D5: Novelty regression - Model B (year×technology- sector fixed effects)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.006***		0.006***	0.015***	0.015***	0.015***	0.012***		0.014***	0.020***	0.020***	0.016***	0.005***		0.004**	0.012***	0.013***	0.014***
	(0.001)		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.003)	(0.003)	(0.001)		(0.001)	(0.003)	(0.003)	(0.002)
team		0.003***	0.003***	0.001	0.001	0.001		0.001	0.001	0.001	0.001	0.000		0.006***	0.005***	0.002	0.002	0.003*
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
eng×team			0.003	-0.001	-0.001				-0.013*	-0.024**	-0.024**				0.005	0.011	0.011	
			(0.003)	(0.005)	(0.005)				(0.006)	(0.007)	(0.007)				(0.003)	(0.007)	(0.007)	
firm				0.006***	0.003***	0.003***				0.002**	0.001	0.001				0.008***	0.005***	0.005***
				(0.000)	(0.000)	(0.000)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
eng×firm				-0.015***	-0.014***	-0.013***				-0.016***	-0.015**	-0.011*				-0.014***	-0.012***	-0.014***
				(0.002)	(0.003)	(0.002)				(0.004)	(0.005)	(0.004)				(0.003)	(0.003)	(0.003)
team×firm				0.005***	0.005**	0.005***				0.003	0.003	0.003				0.004*	0.004*	0.004
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.002)	(0.002)	(0.002)
eng×team×firm				0.004	0.002					0.027*	0.029*					-0.008	-0.011	
				(0.006)	(0.006)					(0.012)	(0.012)					(0.007)	(0.008)	
RnDlab					0.013***	0.013***					0.009***	0.009***					0.013***	0.012***
					(0.001)	(0.001)					(0.002)	(0.002)					(0.001)	(0.001)
eng×RnDlab					-0.009**	-0.007**					-0.018	-0.020					-0.007*	-0.006*
					(0.003)	(0.002)					(0.011)	(0.010)					(0.003)	(0.003)
team×RnDlab					-0.001	0.000					0.006	0.005					-0.003	-0.001
					(0.002)	(0.002)					(0.009)	(0.008)					(0.003)	(0.002)
eng×team×RnDlab					0.008						-0.016						0.010	
					(0.006)						(0.029)						(0.007)	
Constant	0.054***	0.054***	0.054***	0.052***	0.052***	0.052***	0.044***	0.045***	0.044***	0.044***	0.044***	0.044***	0.065***	0.064***	0.064***	0.060***	0.060***	0.060***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	0.033***		0.032***	0.059***	0.061***	0.059***	0.050***		0.051***	0.075***	0.075***	0.071***	0.029***		0.028***	0.052***	0.054***	0.052***
	(0.002)		(0.002)	(0.004)	(0.004)	(0.004)	(0.005)		(0.005)	(0.007)	(0.007)	(0.006)	(0.002)		(0.003)	(0.005)	(0.005)	(0.005)
team		0.012***	0.011***	0.006***	0.006***	0.006***		0.004	0.003	0.001	0.001	0.001		0.020***	0.019***	0.016***	0.016***	0.015***
		(0.001)	(0.001)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
eng×team			0.001	-0.013	-0.013				-0.007	-0.031	-0.031				-0.003	-0.008	-0.008	
			(0.005)	(0.011)	(0.011)				(0.014)	(0.018)	(0.018)				(0.006)	(0.013)	(0.013)	
firm				0.020***	0.011***	0.011***				0.005**	0.002	0.002				0.030***	0.018***	0.018***
				(0.001)	(0.001)	(0.001)				(0.002)	(0.002)	(0.002)				(0.001)	(0.001)	(0.001)
eng×firm				-0.046***	-0.046***	-0.045***				-0.064***	-0.068***	-0.059***				-0.042***	-0.040***	-0.040***
				(0.005)	(0.005)	(0.005)				(0.011)	(0.011)	(0.010)				(0.006)	(0.007)	(0.006)
team×firm				0.011***	0.015***	0.015***				0.013*	0.015**	0.017**				0.001	0.006	0.006
				(0.003)	(0.003)	(0.003)				(0.005)	(0.005)	(0.005)				(0.004)	(0.004)	(0.004)
eng×team×firm				0.018	0.012					0.055	0.060*					0.012	0.002	
				(0.012)	(0.013)					(0.029)	(0.030)					(0.015)	(0.016)	
RnDlab					0.049***	0.049***					0.044***	0.044***					0.047***	0.046***
					(0.002)	(0.002)					(0.006)	(0.006)					(0.002)	(0.002)
eng×RnDlab					-0.019**	-0.014**					0.018	0.012					-0.020**	-0.013*
					(0.006)	(0.005)					(0.027)	(0.025)					(0.006)	(0.006)
team×RnDlab					-0.019***	-0.015**					-0.021	-0.024					-0.020***	-0.015**
					(0.005)	(0.005)					(0.021)	(0.020)					(0.005)	(0.005)
eng×team×RnDlab					0.027*						-0.043						0.033*	
					(0.014)						(0.071)						(0.014)	
Constant	0.528***	0.528***	0.527***	0.521***	0.520***	0.520***	0.476***	0.476***	0.475***	0.475***	0.475***	0.475***	0.583***	0.583***	0.581***	0.568***	0.567***	0.567***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	1508504	1508504	1508504	1508504	1508504	1508504	779970	779970	779970	779970	779970	779970	728534	728534	728534	728534	728534	728534

Robust standard errors in parentheses. : *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

regression tables is complicated.[38] To show the role of interaction effects more clearly, we plot some explicit comparisons derived from the regression results in Figs D4-D7. These figures show how the association of engineers or teams with the novelty of an invention depends on whether these engineers or teams work for firms or research labs. The omitted category against which effects are compared consists of solo inventors, who are not engineers and who patent outside firms and labs.

Our analysis corroborates that engineers tend to patent novel combinations more often, regardless of the organizational context in which they operate (Figs. D4 and D5). This shows that the results in the main text are robust when controlling for technology-year fixed effects. However, it should be noted that relatively few engineers file patents in a standalone capacity: in the period 1856-1945, firms account for 72% of engineers and 76% of teams that include engineers. Both shares rise over time, reaching over 80% by the 1920s.

Also the results on team patenting hold when controlling for technology-year fixed effects (Figs. D6 and D7). However, here we observe a slight reduction in point estimates for firm-based and lab-based teams in model B.

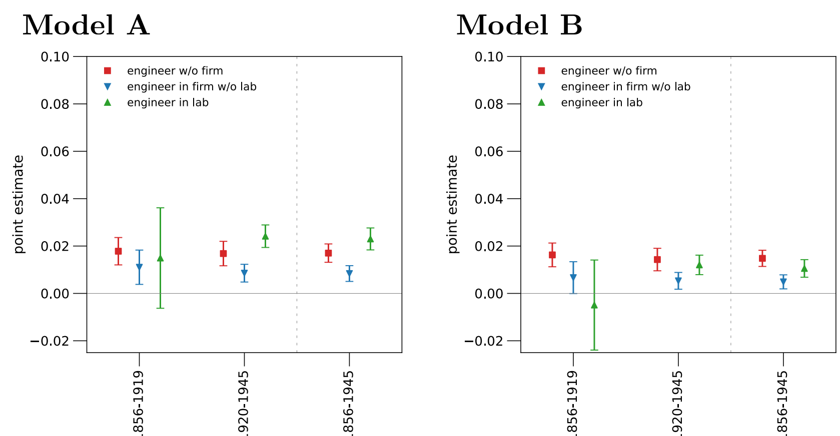


Figure D4: Summary of regression analyses: Engineer effects (3-digit novelty). Omitted category is patents by solo inventors outside firms. Vertical spikes display 95% confidence intervals. **Model A** includes year fixed effects and is reported in the main text. **Model B** includes sector × year fixed effects. Vertical spikes display 95% confidence intervals.

[38]For instance, to calculate the novelty effect of a lab-based team of non-engineers in model 6 of Table D4, we would need to add up the 2nd, 4th, 6th, 8th and 10th coefficients and calculate the standard error of this sum.

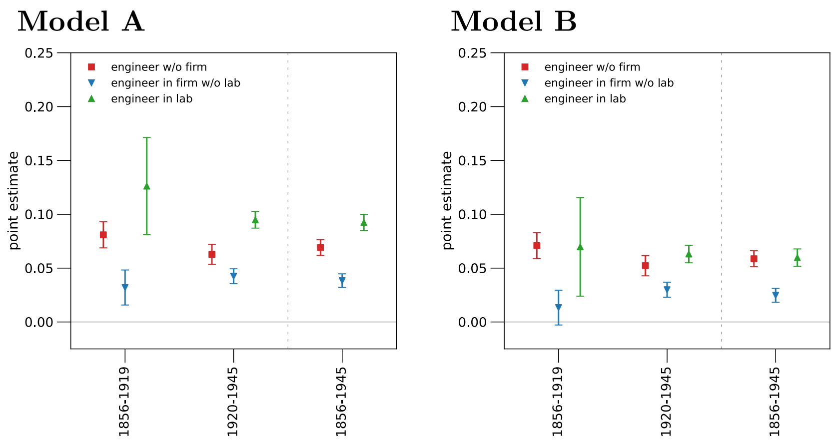


Figure D5: Summary of regression analyses: Engineer effects (6-digit novelty). Omitted category is patents by solo inventors outside firms. **Model A** includes year fixed effects and is reported in the main text. **Model B** includes sector × year fixed effects. Vertical spikes display 95% confidence intervals.

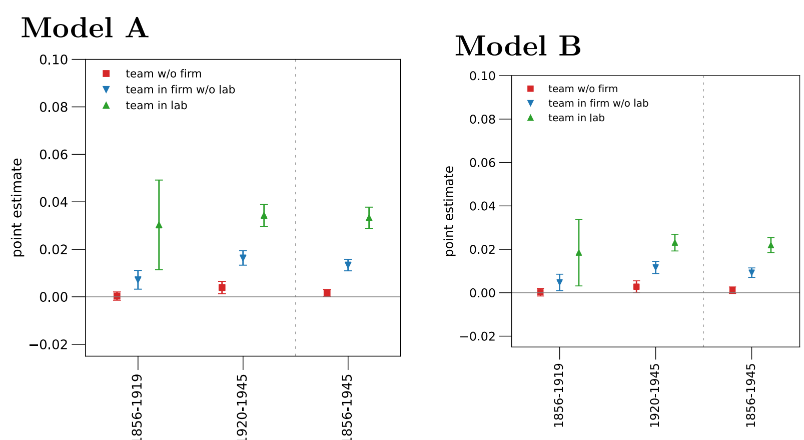


Figure D6: Summary of regression analyses: Team effects (3-digit novelty). Omitted category is patents by solo inventors outside firms. **Model A** includes year fixed effects and is reported in the main text. **Model B** includes sector × year fixed effects. Vertical spikes display 95% confidence intervals.

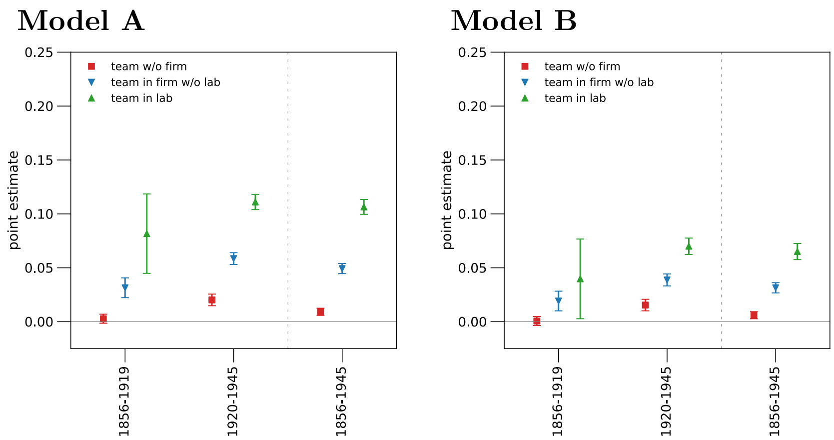


Figure D7: Summary of regression analyses: Team effects (6-digit novelty). Omitted category is patents by solo inventors outside firms. **Model A** includes year fixed effects and is reported in the main text. **Model B** includes sector $\times$ year fixed effects. Vertical spikes display 95% confidence intervals.

D.6 1856-2000

Next, we analyze a smaller model to study how the effects on a patent's novelty change over the entire time period from 1856 to 2000. Because we don't observe demographic information or research labs in this period, we limit the analysis to estimating effects of teams and firms. Furthermore, so far, we have grouped all assignees other than individuals into one category, which we labeled "firms". However, some of these assignees are actually better classified as other types of organizations, such as universities and government agencies. This is only an issue after 1945: before 1945, fewer than 1% of organizational patents are assigned to other organizations than firms. Therefore, when expanding the sample from 1856 to 2000, we test the robustness of our results, dropping all patents that were assigned to non-firm organizations and rerun our regression analyses, using the following specification:

1. *team*: dummy for whether or not the patent lists a team of inventors;
2. *firm*: dummy for whether or not the patent was assigned to a firm;
3. *team*$\times$*firm*: interaction of 1 and 2;

Tables D6-D8 describe the outcomes. We summarize the results in figures that show how the team effect changes between firm-based and standalone patents (Fig. D8-D9). These figures illustrate the robustness of the findings reported in Fig. 14 of the main text. Results do not change much, neither when we add technology-year fixed effects, nor when we drop patents assigned to non-firm organizations.

Table D6: Novelty regression - Model A* (year fixed effects)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.007***		0.008***	0.001		0.000	0.008***		0.004**	0.016***		0.020***	0.004***		0.008***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.007***	0.007***		0.005***	0.005***		0.013***	0.012***		0.021***	0.021***		-0.006***	-0.008***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			-0.004***			0.003			0.005*			-0.010***			-0.003*
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.063***	0.061***	0.060***	0.045***	0.044***	0.044***	0.064***	0.059***	0.058***	0.091***	0.080***	0.078***	0.059***	0.066***	0.064***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.025***		0.019***	0.006**		0.003	0.031***		0.021***	0.040***		0.037***	0.023***		0.022***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.044***	0.042***		0.018***	0.016***		0.048***	0.046***		0.078***	0.076***		0.031***	0.028***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			-0.002			0.015**			0.007*			-0.014***			-0.007***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.653***	0.635***	0.632***	0.476***	0.473***	0.473***	0.582***	0.563***	0.561***	0.727***	0.686***	0.681***	0.768***	0.755***	0.750***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

Robust standard errors in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table D7: Novelty regression - Model B* (year×technological sector fixed effects)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	-0.000		0.007***	0.001		0.000	0.010***		0.004**	0.010***		0.015***	-0.001**		0.005***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.007***	0.010***		-0.001*	-0.002*		0.014***	0.013***		0.009***	0.010***		-0.010***	-0.009***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			-0.013***			0.003			0.006***			-0.009***			-0.005***
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.065***	0.061***	0.060***	0.045***	0.045***	0.045***	0.064***	0.058***	0.058***	0.092***	0.088***	0.086***	0.061***	0.068***	0.067***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.082***		0.048***	0.003		0.000	0.033***		0.022***	0.029***		0.028***	0.017***		0.019***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.114***	0.101***		0.008***	0.007***		0.053***	0.051***		0.054***	0.053***		0.020***	0.018***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			0.009***			0.016**			0.009*			-0.010**			-0.007***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.638***	0.597***	0.591***	0.476***	0.475***	0.475***	0.581***	0.560***	0.558***	0.730***	0.701***	0.697***	0.770***	0.763***	0.759***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3398183	3398183	3398183	779970	779970	779970	728534	728534	728534	675001	675001	675001	1214678	1214678	1214678

Robust standard errors in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

Table D8: Model C* (year fixed effects, dropping non-firm organizations)

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.005***		0.009***	0.001		0.001	0.013***		0.006***	0.015***		0.020***	-0.001***		0.004***
	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.002)	(0.000)		(0.001)
firm		0.014***	0.016***		0.002***	0.002**		0.020***	0.019***		0.020***	0.020***		-0.008***	-0.007***
		(0.000)	(0.000)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			-0.012***			0.003			0.008***			-0.010***			-0.005***
			(0.001)			(0.002)			(0.002)			(0.002)			(0.001)
Constant	0.063***	0.057***	0.056***	0.045***	0.044***	0.044***	0.063***	0.055***	0.055***	0.091***	0.081***	0.079***	0.061***	0.066***	0.066***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)
Observations	3368109	3368109	3368109	779970	779970	779970	728534	728534	728534	671112	671112	671112	1188493	1188493	1188493

	1856 - 2000			1856 - 1919			1920 - 1945			1946 - 1968			1976 - 2000		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
team	0.114***		0.056***	0.005*		0.002	0.048***		0.027***	0.044***		0.039***	0.019***		0.019***
	(0.001)		(0.001)	(0.002)		(0.002)	(0.002)		(0.003)	(0.001)		(0.003)	(0.001)		(0.002)
firm		0.151***	0.130***		0.021***	0.020***		0.074***	0.070***		0.082***	0.080***		0.030***	0.028***
		(0.001)	(0.001)		(0.002)	(0.002)		(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)
team×firm			0.017***			0.016**			0.018***			-0.014***			-0.009***
			(0.001)			(0.005)			(0.004)			(0.003)			(0.002)
Constant	0.629***	0.577***	0.570***	0.476***	0.473***	0.473***	0.580***	0.550***	0.548***	0.726***	0.683***	0.678***	0.769***	0.756***	0.752***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3368109	3368109	3368109	779970	779970	779970	728534	728534	728534	671112	671112	671112	1188493	1188493	1188493

Robust standard errors in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Patent lists a new combination of 3-digit technology codes. **Bottom panel**: Dependent variable: Patent lists a new combination of 6-digit technology codes.

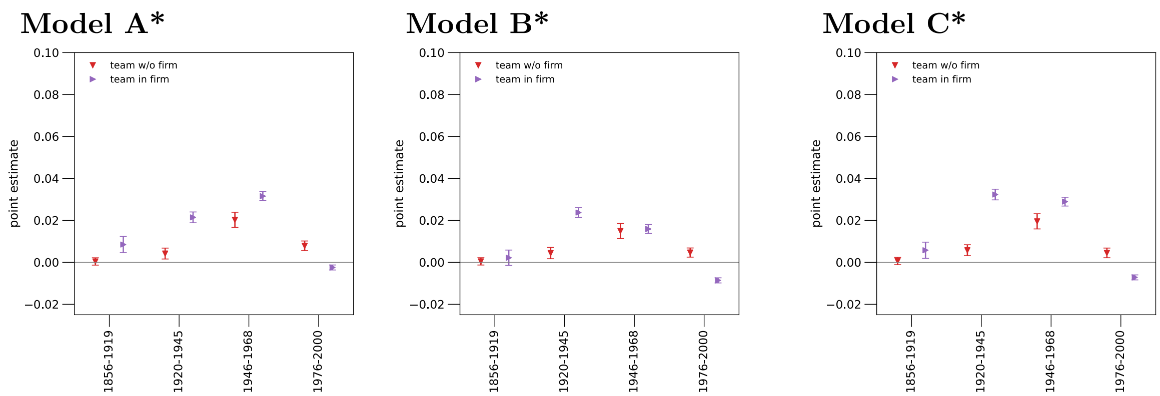


Figure D8: Team effects 1856-2000 (3-digit novelty). Omitted category is patents by solo inventors outside firms. **Model A*** includes year fixed effects and is reported in the main text; **Model B***: includes sector × year fixed effects. **Model C***: includes year fixed effects, but drops patents assigned to non-firm organizations. Vertical spikes display 95% confidence intervals.

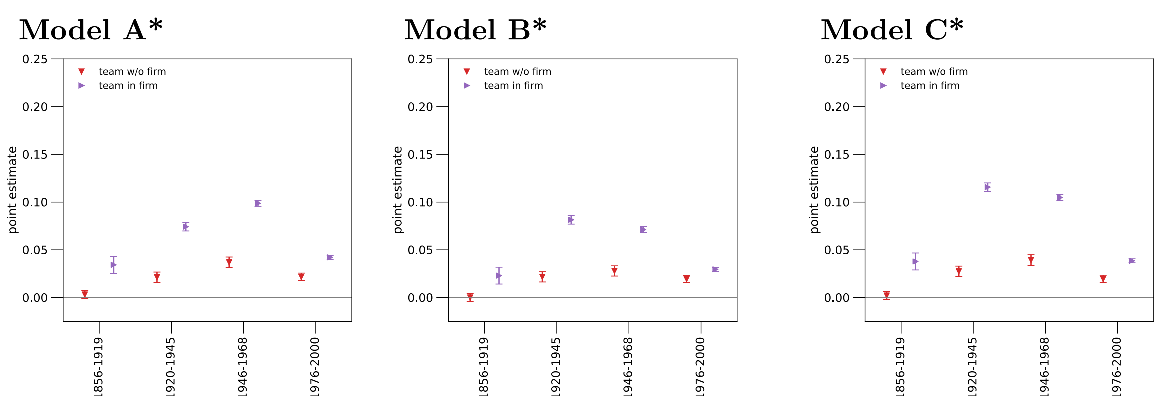


Figure D9: Team effects 1856-2000 (6-digit novelty). Omitted category is patents by solo inventors outside firms. **Model A*** includes year fixed effects and is reported in the main text; **Model B***: includes sector × year fixed effects. **Model C***: includes year fixed effects, but drops patents assigned to non-firm organizations. Vertical spikes display 95% confidence intervals.

D.7 Gender

Finally, we use the regression model of eq. (3) to analyze the likelihood, not that a patent lists a new combination of technologies, but that the patent lists an inventor that we identify as female based on their first name. Table D9 shows the results. The results described in footnote 30 in the main text refer to model (5) for the period 1856-1945.

Table D9: Gender regression (Model A, year fixed effects)

	1856 - 1945						1856 - 1919						1920 - 1945					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
engineers	-0.014***		-0.014***	-0.020***	-0.020***	-0.021***	-0.014***		-0.014***	-0.016***	-0.016***	-0.016***	-0.013***		-0.013***	-0.023***	-0.023***	-0.024***
	(0.000)		(0.000)	(0.000)	(0.000)	(0.001)	(0.001)		(0.000)	(0.001)	(0.001)	(0.001)	(0.000)		(0.000)	(0.001)	(0.001)	(0.001)
team		0.013***	0.014***	0.015***	0.015***	0.015***		0.010***	0.010***	0.010***	0.010***	0.010***		0.015***	0.017***	0.023***	0.023***	0.023***
		(0.000)	(0.000)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
eng×team			-0.003**	-0.001	-0.001				-0.002	-0.002	-0.002				-0.006***	-0.006	-0.006	
			(0.001)	(0.003)	(0.003)				(0.003)	(0.003)	(0.003)				(0.001)	(0.004)	(0.004)	
firm				-0.019***	-0.019***	-0.019***				-0.016***	-0.016***	-0.016***				-0.021***	-0.021***	-0.021***
				(0.000)	(0.000)	(0.000)				(0.000)	(0.000)	(0.000)				(0.000)	(0.000)	(0.000)
eng×firm				0.017***	0.016***	0.015***				0.013***	0.013***	0.013***				0.020***	0.020***	0.019***
				(0.000)	(0.000)	(0.001)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
team×firm				-0.001	-0.000	-0.001				0.002	0.001	0.001				-0.008***	-0.007***	-0.008***
				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)				(0.001)	(0.001)	(0.001)
eng×team×firm				-0.004	-0.005					-0.003	-0.003					0.000	-0.001	
				(0.003)	(0.003)					(0.005)	(0.005)					(0.004)	(0.004)	
RnDlab					-0.002***	-0.002***					-0.004***	-0.004***					-0.000	-0.001
					(0.000)	(0.000)					(0.001)	(0.001)					(0.000)	(0.000)
eng×RnDlab					0.002***	0.003***					0.002	0.001					0.002**	0.002***
					(0.001)	(0.001)					(0.001)	(0.002)					(0.001)	(0.001)
team×RnDlab					-0.001	-0.001					0.009	0.009					-0.003*	-0.003*
					(0.001)	(0.001)					(0.006)	(0.006)					(0.001)	(0.001)
eng×team×RnDlab					0.003						-0.003						0.005	
					(0.003)						(0.016)						(0.003)	
Constant	0.017***	0.015***	0.015***	0.021***	0.021***	0.021***	0.016***	0.014***	0.015***	0.017***	0.017***	0.017***	0.018***	0.015***	0.016***	0.025***	0.025***	0.025***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	1423314	1423314	1423314	1423314	1423314	1423314	735254	735254	735254	735254	735254	735254	688060	688060	688060	688060	688060	688060

Robust standard errors in parentheses. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. **Top panel**: Dependent variable: Dummy for at least one female inventor among the patent's inventors.