NBER WORKING PAPER SERIES

NEW IDEAS IN INVENTION

Mikko Packalen
Jay Bhattacharya

Working Paper 20922
http://www.nber.org/papers/w20922

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2015

A previous version of this paper was circulated under the title "Words in Patents: Research Inputs and the Value of Innovativeness in Invention." We thank Darius Lakdawalla, Dana Goldman, Alan Garber, Richard Freeman, John Ham, Josh Graff Zivin, David Blau, Joel Blit, Subhra Saha, Tom Philipson, Neeraj Sood, Pierre Azoulay, Grant Miller, Jeremy Goldhaber-Fiebert, and Gerald Marschke for their comments on early drafts of this paper and for their encouragement. We also thank seminar participants at the Harvard Business School, Stanford School of Medicine, University of Guelph, and especially Bruce Weinberg's working group on innovation and science at the NBER for excellent feedback. Despite all this help, the authors are responsible for all errors in the paper. We acknowledge financial support from the National Institute on Aging grant P01-AG039347. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

New Ideas in Invention
Mikko Packalen and Jay Bhattacharya
NBER Working Paper No. 20922
January 2015
JEL No. I1,O31,O32,O33

ABSTRACT

A key decision in research is whether to try out new ideas or build on more established ideas. In this paper, we evaluate which type of work is more likely to spur further invention. When recent advances create superior opportunities for invention, their adoption as research inputs in the invention process promotes technological progress. The gains from pursuing such innovative research paths may, however, be very limited as new ideas are often initially raw and poorly understood. We determine idea inputs in invention based on the text of nearly every US patent granted during 1836–2010. We find that inventions that build on new ideas early are more likely to spur subsequent invention than inventions that rely on ideas of older vintage. Our results are important because they suggest a benefit from encouraging and supporting innovative research that tries out new ideas — avoiding stagnation in technological advance.

Mikko Packalen
University of Waterloo
Department of Economics
200 University Avenue West
Waterloo, ON N2L 3G1
Canada
packalen@uwaterloo.ca

Jay Bhattacharya
117 Encina Commons
CHP/PCOR
Stanford University
Stanford, CA 94305-6019
and NBER
jay@stanford.edu

1 Introduction

Anyone involved in research must choose whether to build their work on recent advances or rely on more established knowledge. This is a choice faced by scientists and inventors as well as private and public financiers of research, such as pharmaceutical firms, the National Science Foundation, and the National Institutes of Health. When recent advances create superior opportunities for invention, innovative research that pursues those new opportunities promotes technological progress. Many of the potential benefits of such innovative research may, however, never be realized as risk aversion, the principal-agent problem, limited rationality, and entrenched interests may bias researchers, firms, and research agencies against innovative research paths (e.g. Kuhn 1962; March 1991; Ahuja and Lambert 2001).

Yet, favoring less innovative research directions is not necessarily foolish, as the private and social benefits of work that tries out new ideas may be quite limited. Important new scientific and technological advances are often initially raw and poorly understood (Marshall, 1920; Usher, 1929; Kuhn, 1962). Organizations also often lack the necessary complementary capabilities to build on new ideas (Nerkar, 2003). Work that builds on fresh ideas thus need not result in much useful invention. Knowledge about the properties of recent advances may also initially progress so fast that inventions building on it soon become obsolete. Moreover, when the knowledge base in recombinant invention is expansive enough, advances that add to it have little impact on what can be achieved with invention (Weitzman, 1998).

That the benefits of trying out new ideas may be small is not a mere theoretical possibility. There exists both anecdotal and quantitative evidence (Utterback, 1996; Fleming, 2001) suggesting that knowledge needs to mature and deepen before it becomes most useful in spurring subsequent inventions.

The benefits of trying out new ideas in invention may thus be large, small, non-existent, or even negative. In this paper, we offer a quantitative comparison of invention that builds on new ideas and invention that builds on more established ideas: we examine which type of work spurs more subsequent invention.

We find that inventions that build on new ideas are more likely to spur subsequent invention than inventions that rely on ideas of older vintage. Our results are important because they suggest a benefit from encouraging and supporting innovative research that tries out new ideas — avoiding stagnation in technological advance. The results add to the heretofore sparse systematic evidence on the benefits of pursuing innovative research directions (Fleming, 2001; Ahuja and Lampert 2001; Schoenmakers and Duysters 2010).[1]

We examine patent texts to determine which inventions build on newer ideas. The textual approach reveals organically which technologies and scientific discoveries have been popular idea inputs in invention.[2,3] We rely on patent citations to determine how much subsequent innovation each invention generated. Because patent citations may reflect similarity rather than cumulative invention, we develop a novel citation measure that reflects only inventions that build upon the cited invention. We also contribute by organizing and examining patent-level data for 1836–2010. To our knowledge, existing large-scale patent-level analyses have focused on the post–1975 time period.

2 Methods

We first propose a new way to identify idea inputs in technological innovation. We then explain how we use this information on idea inputs to measure of the vintage of idea inputs in each in-

[1] Ahuja and Lampert (2001) and Schoenmakers and Duysters (2010) find that, in chemical and pharmaceutical industries, highly cited patents cite more recent patents than do other patents, suggesting that the use of emerging knowledge facilitates invention. These analyses are, however, subject to an important caveat: a citation may reflect mere similarity rather than cumulative invention (e.g. Jaffe et al., 2002). Hence, differences in the age of cited patents may reflect mere differences in the extent of similar inventions rather than differences in idea inputs, and differences in forward citations may reflect mere differences in the extent of similar inventions rather than differences in subsequent advance. We address this issue by measuring the age of idea inputs from text instead, and by identifying breakthrough inventions based on an approach that excludes forward citations that may reflect mere similarity.

[2] Existing textual approaches to measuring innovativeness from text (Evans, 2011; Grodal and Thoma, 2009; Azoulay et al., 2011; Bhattacharya and Packalen, 2011) have relied on predefined word lists.

[3] Our analysis also complements the recombination theory of invention (e.g. Usher, 1922, Schumpeter, 1939, Weitzman, 1998) by providing systematic evidence on what new ideas and matter are recombined in technological innovation and on how important new knowledge is as an idea input.

vention. Next, we present our approach to measuring the extent of subsequent advance spurred by each patent. Finally, we discuss how we link these constructed variables to estimate whether work that builds on newer ideas spurs more further invention.

2.1 Identifying Idea Inputs from Patent Texts

Existing analyses of invention have typically captured idea inputs from patent citations (e.g. Caballero and Jaffe, 1993; Popp, 2002).[4] Two well-known drawbacks of this approach are that (1) in any given domain citations can reveal only a very limited set of idea inputs (Rosenberg, 1982) and that (2) because the main purpose of patent citations is to delineate the boundaries of a patent rather than disclose prior art the patent built upon, at least half of patent citations do not reflect idea inputs (Jaffe et al., 2000).[5] For these reasons, we sought to develop a new approach.

In our approach, we measure idea inputs from patent text. By design, patent texts distribute information about advancements in knowledge: a patent text describes the invention and its components. Consequently, we expected that by indexing *words* and *word sequences* in patents we would uncover at least a subset of the knowledge and matter that were recombined in the invention process that led to the invention. Below we show that this indeed turned out indeed to be the case — the indexing revealed important prior inventions and scientific discoveries that have served as idea inputs in the invention process.[6]

The patent data we index spans 175 years of US patents (1836–2010). We first index all words

[4] An alternative existing approach measures idea inputs from subclass designations in patents (Fleming, 2001). A drawback of that approach is that subclasses capture a very limited set of idea inputs. In related work, Alexopoulos (2011) uses classifications of technical books to measure rates of invention.

[5] Addressing the latter concern by excluding backward citations for which the cited and citing patents or their components lie in same technology categories (see section 2.4) would leave one with idea inputs that are even more limited in their number and in what the inputs cover.

[6] Because the purpose of patent text is to describe the invention, there does not appear to exist much reason for inventors to include words that do not reflect the components of the invention. One source of noise is that some inventions are re-named or named first only after proven valuable (e.g. drugs). It is also possible that a new word or word sequence represents the output of the patent, as opposed to an input. However, this property does not drive our results: for a given idea, the number of such patents is at most one, whereas the number of patents we consider innovative because of the mention of the new idea is generally orders of magnitudes higher. Moreover, our results are robust to excluding for each new concept the patent that received the most citations.

and 2- and 3-word sequences that appear in each patent. By *word* we mean a character sequence that is separated from other character sequences with whitespace. For each patent, the indexed text includes the title, abstract, body, and claims.

We refer to the indexed words and word sequences as *concepts*. To determine when the idea represented by a concept was a new input to the inventive process, we determine for each concept the year in which first it appears in the patent data. We refer to this year of arrival as the idea's *cohort*.[7,8] For all post–1870 cohorts, we examine the initial lists of words and word sequences that appear often in patents and exclude words and word sequences that appear to reflect changes in spelling or presentation of patents rather than changes in the nature of inventive activity. Please see the Data Appendix for details.

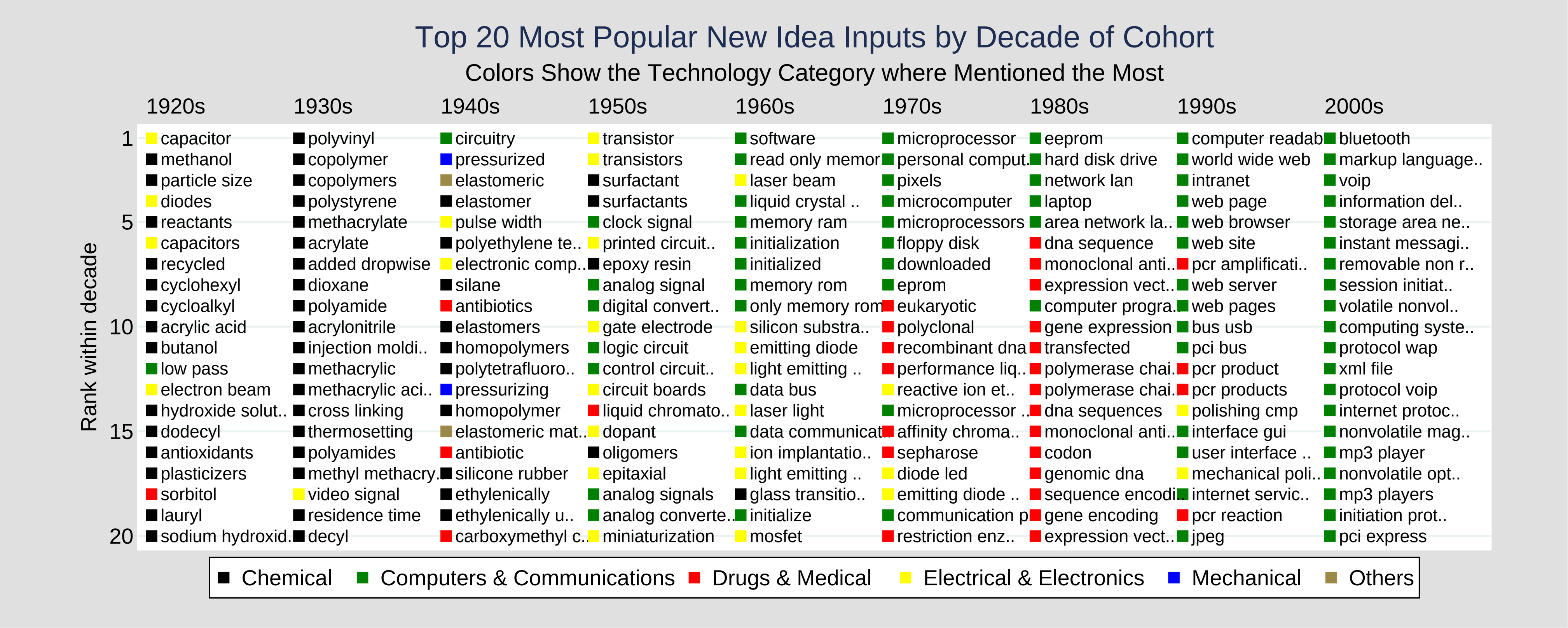
After indexing concept mentions in each patent and the cohort of each concept, we rank concepts in each cohort based on the number US patents that mention each concept. This ranking enables us to focus attention on the best idea inputs in each cohort.

Table 1 lists the 20 most popular new concepts in each decade from 1920s to 2000s (part 1) and from 1840s to 1910s (part 2). For this summary table, concepts are grouped by the decade of their cohort. The colored squares affixed to each concept name in the table indicate the technology category with the most patents that mention the concept.

[7] Cohort years and the timing of invention are measured from grant years of patents because the application year is often ambiguous and not readily available. Newer patents have multiple application years for patents that are based on a continuation application. For older patents it must be extracted from OCR text.

[8] Due to OCR errors and typos, we ignore the initial mention of concepts that are mentioned less than 5 times during the subsequent 25 years. For such concepts the cohort is set as the earliest year in which (1) the concept is mentioned, and (2) the concept is mentioned at least 5 times during the 25 years that follow.

Table 1, Part 1: Top 20 Most Popular New Idea Inputs by Decade of of Cohort, 1920s-2000s.



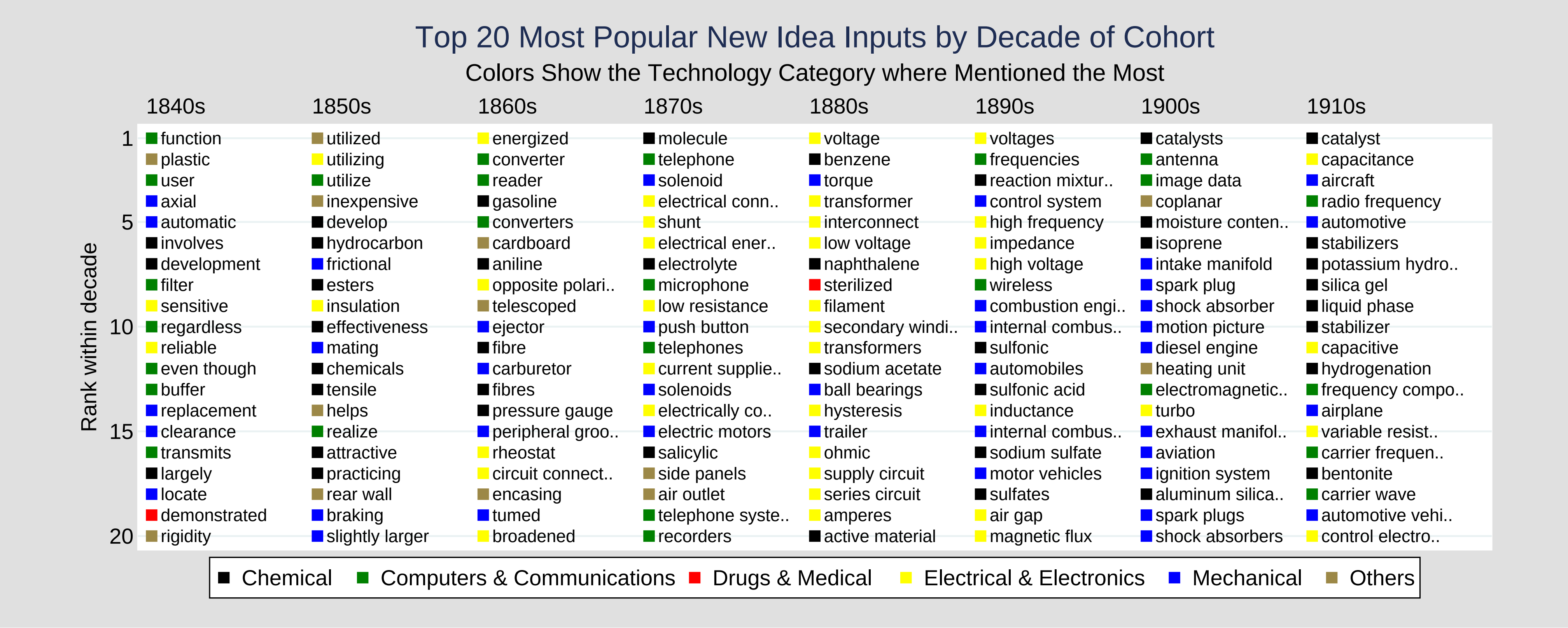
Top 20 Most Popular New Idea Inputs by Decade of Cohort

Colors Show the Technology Category where Mentioned the Most

Rank within decade	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
1	🟨 capacitor	⬛ polyvinyl	🟩 circuitry	🟨 transistor	🟩 software	🟩 microprocessor	🟩 eeprom	🟩 computer readab..	🟩 bluetooth
	⬛ methanol	⬛ copolymer	🟦 pressurized	🟨 transistors	🟩 read only memor..	🟩 personal comput..	🟩 hard disk drive	🟩 world wide web	🟩 markup language..
	⬛ particle size	⬛ copolymers	🟫 elastomeric	⬛ surfactant	🟨 laser beam	🟩 pixels	🟩 network lan	🟩 intranet	🟩 voip
	🟨 diodes	⬛ polystyrene	⬛ elastomer	⬛ surfactants	🟩 liquid crystal ..	🟩 microcomputer	🟩 laptop	🟩 web page	🟩 information del..
5	⬛ reactants	⬛ methacrylate	🟨 pulse width	🟩 clock signal	🟩 memory ram	🟩 microprocessors	🟩 area network la..	🟩 web browser	🟩 storage area ne..
	🟨 capacitors	⬛ acrylate	⬛ polyethylene te..	🟨 printed circuit..	🟩 initialization	🟩 floppy disk	🟥 dna sequence	🟩 web site	🟩 instant messagi..
	⬛ recycled	⬛ added dropwise	🟨 electronic comp..	⬛ epoxy resin	🟩 initialized	🟩 downloaded	🟥 monoclonal anti..	🟥 pcr amplificati..	🟩 removable non r..
	⬛ cyclohexyl	⬛ dioxane	⬛ silane	🟩 analog signal	🟩 memory rom	🟩 eprom	🟥 expression vect..	🟩 web server	🟩 session initiat..
	⬛ cycloalkyl	⬛ polyamide	🟥 antibiotics	🟩 digital convert..	🟩 only memory rom	🟥 eukaryotic	🟩 computer progra..	🟩 web pages	🟩 volatile nonvol..
10	⬛ acrylic acid	⬛ acrylonitrile	⬛ elastomers	🟨 gate electrode	🟨 silicon substra..	🟥 polyclonal	🟥 gene expression	🟩 bus usb	🟩 computing syste..
	⬛ butanol	⬛ injection moldi..	⬛ homopolymers	🟩 logic circuit	🟨 emitting diode	🟥 recombinant dna	🟥 transfected	🟩 pci bus	🟩 protocol wap
	🟩 low pass	⬛ methacrylic	⬛ polytetrafluoro..	🟩 control circuit..	🟨 light emitting ..	🟥 performance liq..	🟥 polymerase chai..	🟥 pcr product	🟩 xml file
	🟨 electron beam	⬛ methacrylic aci..	🟦 pressurizing	🟨 circuit boards	🟩 data bus	🟨 reactive ion et..	🟥 polymerase chai..	🟥 pcr products	🟩 protocol voip
	⬛ hydroxide solut..	⬛ cross linking	⬛ homopolymer	🟥 liquid chromato..	🟨 laser light	🟩 microprocessor ..	🟥 dna sequences	🟨 polishing cmp	🟩 internet protoc..
15	⬛ dodecyl	⬛ thermosetting	🟫 elastomeric mat..	🟨 dopant	🟩 data communicat..	🟥 affinity chroma..	🟥 monoclonal anti..	🟩 interface gui	🟩 nonvolatile mag..
	⬛ antioxidants	⬛ polyamides	🟥 antibiotic	⬛ oligomers	🟨 ion implantatio..	🟥 sepharose	🟥 codon	🟩 user interface ..	🟩 mp3 player
	⬛ plasticizers	⬛ methyl methacry..	⬛ silicone rubber	🟨 epitaxial	🟨 light emitting ..	🟨 diode led	🟥 genomic dna	🟨 mechanical poli..	🟩 nonvolatile opt..
	🟥 sorbitol	🟨 video signal	⬛ ethylenically	🟩 analog signals	⬛ glass transitio..	🟨 emitting diode ..	🟥 sequence encodi..	🟩 internet servic..	🟩 mp3 players
	⬛ lauryl	⬛ residence time	⬛ ethylenically u..	🟩 analog converte..	🟩 initialize	🟩 communication p..	🟥 gene encoding	🟥 pcr reaction	🟩 initiation prot..
20	⬛ sodium hydroxid..	⬛ decyl	🟥 carboxymethyl c..	🟨 miniaturization	🟨 mosfet	🟥 restriction enz..	🟥 expression vect..	🟩 jpeg	🟩 pci express

⬛ Chemical 🟩 Computers & Communications 🟥 Drugs & Medical 🟨 Electrical & Electronics 🟦 Mechanical 🟫 Others

Table 1, Part 2: Top 20 Most Popular New Idea Inputs by Decade of of Cohort, 1840s-1910s.



Top 20 Most Popular New Idea Inputs by Decade of Cohort

Colors Show the Technology Category where Mentioned the Most

Rank within decade

	1840s	1850s	1860s	1870s	1880s	1890s	1900s	1910s
1	🟩 function	🟫 utilized	🟨 energized	■ molecule	🟨 voltage	🟨 voltages	■ catalysts	■ catalyst
	🟫 plastic	🟨 utilizing	🟩 converter	🟩 telephone	■ benzene	🟩 frequencies	🟩 antenna	🟨 capacitance
	🟩 user	🟩 utilize	🟩 reader	🟦 solenoid	🟦 torque	■ reaction mixtur..	🟩 image data	🟦 aircraft
	🟦 axial	🟫 inexpensive	■ gasoline	🟨 electrical conn..	🟨 transformer	🟦 control system	🟫 coplanar	🟩 radio frequency
5	🟦 automatic	■ develop	🟩 converters	🟨 shunt	🟨 interconnect	🟨 high frequency	■ moisture conten..	🟦 automotive
	■ involves	■ hydrocarbon	🟫 cardboard	🟨 electrical ener..	🟨 low voltage	🟨 impedance	■ isoprene	■ stabilizers
	■ development	🟦 frictional	■ aniline	■ electrolyte	■ naphthalene	🟨 high voltage	🟦 intake manifold	■ potassium hydro..
	🟩 filter	■ esters	🟨 opposite polari..	🟩 microphone	🟥 sterilized	🟩 wireless	🟦 spark plug	■ silica gel
	🟨 sensitive	🟨 insulation	🟫 telescoped	🟨 low resistance	🟨 filament	🟦 combustion engi..	🟦 shock absorber	■ liquid phase
10	🟩 regardless	■ effectiveness	🟦 ejector	🟦 push button	🟨 secondary windi..	🟦 internal combus..	🟦 motion picture	■ stabilizer
	🟨 reliable	🟦 mating	■ fibre	🟩 telephones	🟨 transformers	■ sulfonic	🟦 diesel engine	🟨 capacitive
	🟩 even though	■ chemicals	🟦 carburetor	🟨 current supplie..	■ sodium acetate	🟦 automobiles	🟫 heating unit	■ hydrogenation
	🟩 buffer	■ tensile	■ fibres	🟦 solenoids	🟦 ball bearings	■ sulfonic acid	🟩 electromagnetic..	🟩 frequency compo..
	🟦 replacement	🟫 helps	■ pressure gauge	🟨 electrically co..	🟨 hysteresis	🟨 inductance	🟨 turbo	🟦 airplane
15	🟦 clearance	🟩 realize	🟦 peripheral groo..	🟦 electric motors	🟦 trailer	🟦 internal combus..	🟦 exhaust manifol..	🟨 variable resist..
	🟩 transmits	■ attractive	🟨 rheostat	■ salicylic	🟨 ohmic	■ sodium sulfate	🟦 aviation	🟩 carrier frequen..
	■ largely	■ practicing	🟨 circuit connect..	🟫 side panels	🟨 supply circuit	🟦 motor vehicles	🟦 ignition system	■ bentonite
	🟦 locate	🟫 rear wall	🟫 encasing	🟫 air outlet	🟨 series circuit	■ sulfates	■ aluminum silica..	🟩 carrier wave
	🟥 demonstrated	🟦 braking	🟦 tumed	🟩 telephone syste..	🟨 amperes	🟨 air gap	🟦 spark plugs	🟦 automotive vehi..
20	🟫 rigidity	🟦 slightly larger	🟨 broadened	🟩 recorders	■ active material	🟨 magnetic flux	🟦 shock absorbers	🟨 control electro..

■ Chemical 🟩 Computers & Communications 🟥 Drugs & Medical 🟨 Electrical & Electronics 🟦 Mechanical 🟫 Others

Table 1 shows that starting from the late 19th century the most popular idea inputs captured by our approach generally represent important prior inventions and scientific discoveries. The approach put forward here thus works as intended: it captures important idea inputs in invention — pieces of knowledge that were recombined in the invention process that led to the patented invention. In Table A1 in the appendix we provide further supporting evidence, this time from the perspective of which idea inputs render patents innovative in each decade, using the measure of innovativeness defined below in section 2.4.

The textual approach complements the citation and subclass based approaches to measuring idea inputs in invention. Besides capturing many more types of idea inputs than either existing approach, an advantage of the textual approach is that a non-expert will often find it much easier to understand the meaning of ideas uncovered from text than the meaning of ideas uncovered from citations or subclasses; subclass names and patent and scientific article titles are often narrow and very technical. In our application, for example, this is beneficial because the lists of idea inputs enable the reader to verify that the ideas uncovered from text are indeed ideas that drove technological change and were relatively new inputs in the years that followed the idea's cohort year, and that, consequently, the patents that built on these ideas early represented innovative work that tried out a new idea.[9]

2.2 Measuring the Vintage of Idea Inputs

We calculate the vintage of idea inputs in each patent based on the appearances of idea inputs with the most potential. To identify these idea inputs, we apply the principle of revealed preference: within each idea input cohort, we consider those new ideas that are mentioned the most often by the end of the sample period to be the most potent new ideas. For this purpose, we first identify the top 100 new concepts in each cohort based on the number of patents that mention them by the

[9] Other rationales for pursuing text analysis include: noise in citations (Jaffe et al. 2000; Alcacer et al., 2009; Lampe, 2012), the fact that citations can reflect similarity rather cumulative invention (section 2.3), sparsity of patent-to-science citations in pre–1980s data, and sparsity of any citations in pre–1947 patents.

end of the sample. For each patent, we then determine the age of the newest top 100 concept that appears in it. We refer to this measure as *Age of Idea Inputs*; this measure is our main measure of the vintage of idea inputs in each patent. We later calculate a similar measure based on appearances of the top $10,000$ concepts in each cohort.

A rationale for focusing on ideas with the most potential is that inventors, firms and funding agencies who are considering pursuing or funding research that builds on a new idea are likely to have private information (or beliefs) about the input's long-term potential relative to other new ideas. Hence, rather than contemplating whether early work on a random new idea is worth pursuing, their decision is more likely to be between pursuing work on a potent new idea now and postponing that work until the idea has matured. For this decision, the key uncertainty revolves around whether work on the new ideas with the most potential is worthwhile soon after their arrival or only later; Fleming (2001) suggests that work should be directed elsewhere until new ideas have matured sufficiently.

2.3 Measuring Advances Spurred by Each Invention

We measure the advances spurred by each invention from the citations the patent has received from other patents. Citations are a commonly used measure of patent value (e.g. Harhoff et al., 1999; Hall et al., 2005) and knowledge flows (e.g. Jaffe et al., 1993). Fewer than half of citations, however, actually reflect cumulative invention (Jaffe et al., 2002). This is in part because the main purpose of patent citations is not the disclosure of prior art upon which the invention built, but rather to delimit the scope of the patent by indicating which parts of the citing invention are not novel and therefore not covered by the patent (e.g. Jaffe et al., 1993; Strumsky et al., 2010). A citation may thus merely indicate that two inventions, or some of their components, are similar in the sense that the inventions or some of their components are near one another in the technology space, even if they are built upon different ideas or principles (e.g. Jaffe et al., 2002).

The concern that citations may reflect the mere similarity of inventions weakens the case for using citations to measure cumulative invention. To address this issue, we construct a new measure: the count of citations for which the novel parts of the citing invention are not anywhere near the novel parts of the cited invention in the technology space. Among citations, such citing and cited patent pairs seem the most likely to reflect cumulative invention.

To construct this measure, we first determine how close the citing and cited patents' novel parts are in the technology space. The novel parts are specified by the claims of each patent. The primary and multiple secondary technology classification codes assigned to a patent in turn delineate what types of technologies are covered by the claims (Strumsky et al., 2010; U.S. Patent and Trademark Office, 2005). We infer whether the novel parts of the citing invention are near the novel parts of the cited invention in the technology space from the technology codes of each citing and cited patent pair. The technology space is specified by patent examiners who assign the technology codes to patents and also maintain the classification.

At the 3-digit level, the classification system used in patents has over 400 technology classes. Because different 3-digit codes may cover closely related technologies, a citing invention may be near a cited invention even when the two do not share a 3-digit technology code. Fortunately, Hall et al. (2011) mapped the 3-digit technology classes into 6 broad technology categories. This mapping allows us to extract those citing and cited patent pairs that are unlikely to cover similar technologies: patent pairs for which the technology categories spanned by their 3-digit technology class codes do not overlap.

In our approach, we first determine for each patent the primary and all secondary 3-digit technology codes assigned to the invention. Next, we determine which technology categories are spanned by these technology classes.[10] We then calculate the number of citations received from patents which technology categories do not overlap with any of the technology categories of the cited patent. We refer to the count of such received citations as *No-Overlap Citations*. From this

[10] The approach is distinct from counting citations for which the category of the primary technology class is different for the citing and cited patents because, for instance, 26% of patents granted during 1920–2010 have technology codes in multiple categories.

count, we also construct an indicator variable that captures whether a patent is among the top 5% most cited by this citation measure among patents granted in the same technology class in the same year. We refer to this measure as *Top 5% by No-Overlap Citations*. These are our preferred measures of the extent of subsequent advance spurred by each patent. We also report results based on two traditional measures of cumulative invention: the count of total received citations and the corresponding top 5% most cited status of each patent (variables *Total Citations* and *Top 5% by Total Citations*, respectively).

2.4 Estimating Link Between the Age of Idea Inputs and Subsequent Advances

In non-parametric analyses, we examine how the number of received patent citations varies by the age of idea inputs, as measured by the constructed *Age of Idea Inputs* variable. For these analyses we first normalize each citation measure so that its mean is the same across all technology class and year pairs. We then group patents based on the age of idea inputs and calculate the mean of received patent citations for each group of patents.

Because some technology areas may adopt new ideas faster than others, the vintage of ideas that can be considered relatively new may vary across technology areas. To address this issue, we also compare citations to patents that are among the first to use a potent new idea within a given research area against citations to other, less innovative, patents in that same area. For these analyses, we construct a dummy variable that captures whether a patent is among the top 5% most recent based on the *Age of Idea Inputs* variable. The comparison group for each patent is other patents granted in the same technology class in the same year. We refer to this indicator variable as *Top 5% by Age of Idea Inputs*.[11]

In parametric regression analyses, we employ as outcome variables the four measures that capture received citations; we eschew from using the *Age of Idea Inputs* variable to avoid the use

[11] We previously employed such a measure in Packalen and Bhattacharya (2015a).

of a non-linear fixed effects specification. The main explanatory variable is the variable *Top 5% by Age of Idea Inputs*. We include patent length as a control variable – measured by the number of characters – because longer patents are more likely to include any given concept. We include a separate fixed effect parameter for each technology class and grant year pair (within estimation). As the citation measures that serve as dependent variables are all either binary or count variables, we employ conditional logit and Poisson models.

3 Data and Descriptive Statistics

Our data consist of nearly all US patent documents granted during 1836–2010.[12] Figure A1 in the appendix shows the number of patents granted in each year. For 1976–2010, the patent data are a machine-readable transfer from the original patents. In these data, fields such as title, abstract, claims, and references are clearly indicated. For 1836–1975, the patent data are an Optical Character Recognition ("OCR") transfer from the original patent images, which we performed on 4+ million patents. In these data, only patent number and grant year are separately indicated; elements such as title, application year, claims, and references must be determined by searching the ASCII scan of each patent for the relevant markers. Please see the Data Appendix for details on our data organization, extraction and disambiguation efforts.

The key descriptive statistic is the extent of variation in the age of idea inputs in patents. Figure A2 in the appendix shows the distribution of the *Age of Idea Inputs* variable by time period, when the age of idea inputs is determined based on mentions of the top 100 concepts in each cohort. We limit the analysis to patents granted since 1880 and mentions of idea inputs from the post-1870 cohorts because we have not inspected the lists of new words and word sequences for the pre–1870 cohorts; so many words and word sequences that do not reflect new idea inputs first appear in the data during those early years that we deemed a manual elimination of such concepts

[12] The data cover over 99.8% of patents granted in any given year, over 99.93% of patents granted before 1976, and over 99.99% of patents granted after 1976.

for the pre–1870s cohorts to be too resource-intensive. Figure A2 shows that within each time period there is variation in whether a patent builds on the most potent ideas early. Comparison of the distributions in Figure A1 across time periods suggests that since the 1970s there has been a considerable increase in the pace at which new idea inputs are adopted in invention.

Figure A3 in the appendix in turn depicts the mean for the outcome variable and *No-Overlap Citations* by technology category for each year. We note that only patents granted since 1947 include a references section; we do not index in-text citations. The Figure A3 shows that patents receive *No-Overlap Citations* in all technology categories.

4 Results

Figure 1 shows the mean of received citations by the age of the newest idea input. The main panel and the upper right panel depict the results for the two preferred citation measures, *Top 5% by No-Overlap Citations* and *No-Overlap Citations*, respectively. The bottom right panel depicts the results for the measure *Total Citations*.

In each panel the relationship is downward-sloping. Inventions that build on the most potent new ideas early thus appear to spur more subsequent invention than do inventions that build on these ideas only later. This result suggests that it is wise to pursue early work that tries out the new ideas that hold the most long-term potential rather than postpone work on those ideas until the ideas have matured. This result runs counter to an influential earlier analysis according to which knowledge may need to mature first (Fleming, 2001).[13]

Figure 2 presents the following comparison for each year: citations to patents that are among

[13] Fleming (2001) concluded that that "Organizations that seek technological breakthroughs should experiment with new combinations, possibly with old components." Our findings suggest that organizations should experiment with relatively new components too. The approach of Fleming (2001) focuses on the average component, whereas we focus on the age of the most recent idea input. Moreover, we infer idea components from words in patent texts whereas Fleming (2001) infers them from technology subclasses. A caveat to both sets of results arises because not all inventive efforts are successful. However, such truncation may not be that significant because the bar to patent is so low (Fleming, 2001). This caveat notwithstanding, our analysis shows that the use of new ideas in invention matters as it changes the distribution of outcomes and that words in patents are an important predictor of patent citations.

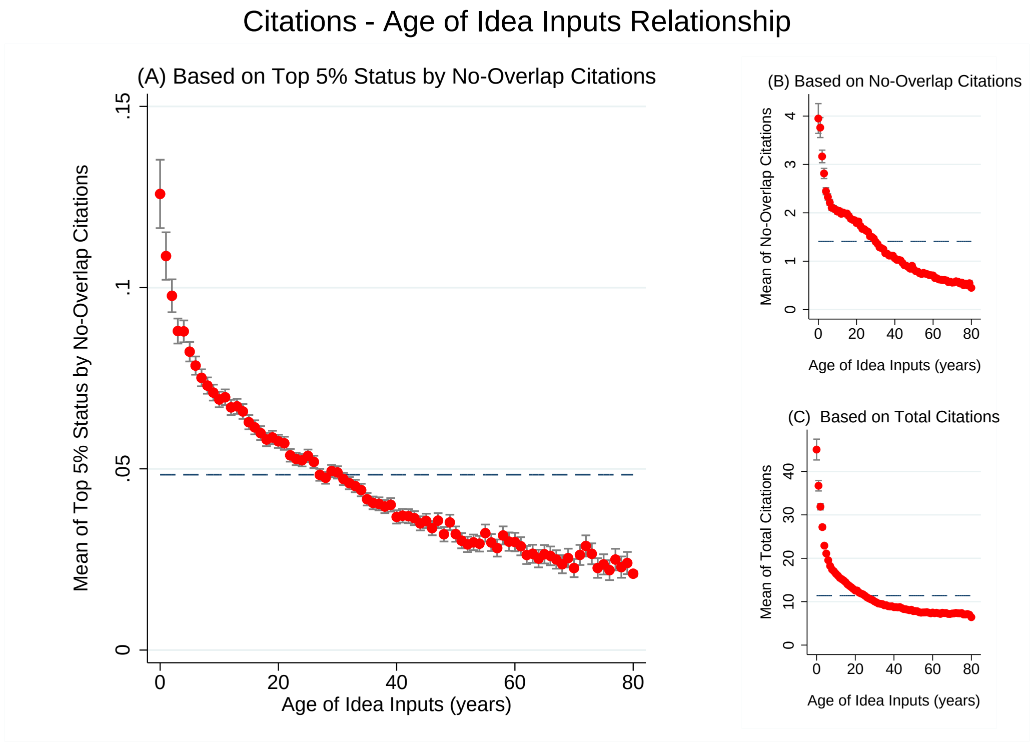


Figure 1: Estimates of the citations–age of idea inputs relationship. In the main panel (A) the outcome variable is the indicator variable *Top 5% by No-Overlap Citations*; in the side panels (B) and (C) the outcome variables are *No-Overlap Citations* and *Total Citations*, respectively. Age of idea inputs is determined based on mentions of the top 100 concepts in each cohort. The reported values are averages for patents granted during 1976–2005; observations are weighted so that observations from each year receive the same total weight as observations from any other year. Capped lines indicate 95% confidence intervals.

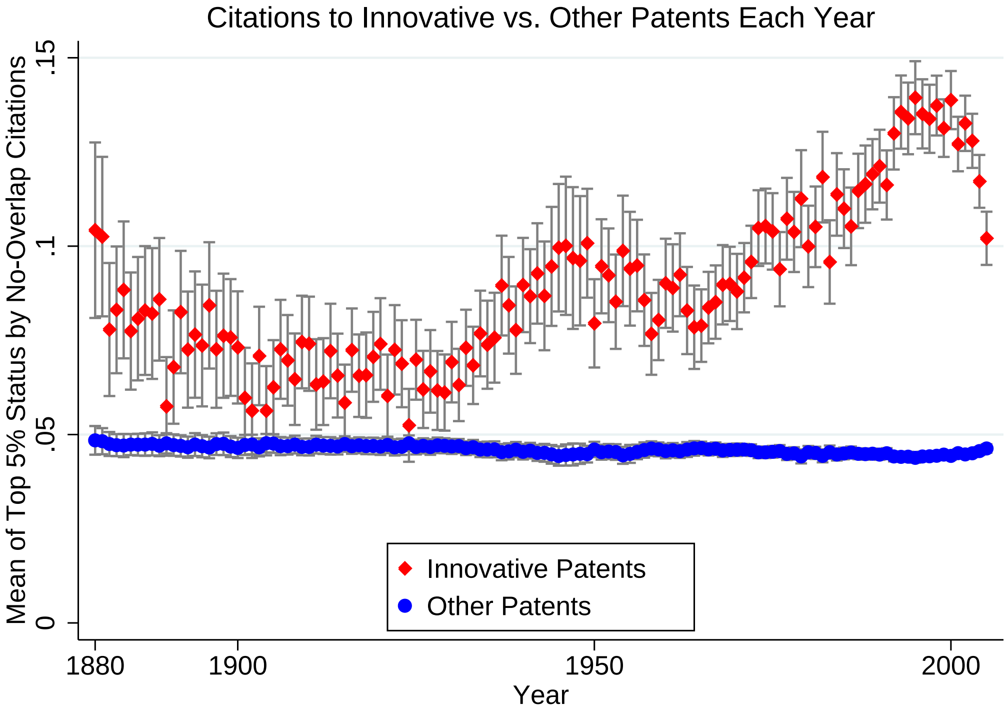


Figure 2: Citations to innovative patents vs. citations to other patents. Horizontal axis depicts the grant year of patents. Vertical axis depicts the mean of the indicator variable *Top 5% by No-Overlap Citations*. Patents granted each year are divided to two groups based on the variable *Top 5% by Age of Idea Inputs*: innovative patents (patents with top 5% status by age of idea inputs) and other patents. Age of idea inputs is calculated based on mentions of the top 100 concepts in each cohort. Capped lines indicate 95% confidence intervals.

the top 5% most recent by the age of the newest idea input vs. citations to other patents. To determine which patents belong in the former, innovative, group of patents, each patent is compared to other patents granted in the technology class in the same year. The results again support the conclusion that inventions that use new ideas early spur more subsequent advance than do inventions that are less innovative in terms of the ideas that they build upon.

Table 2 on the next page and Tables A2, A3 and A4 in the appendix show the results from parametric regression analyses. In Table 2, the four columns show the age of idea inputs–received citations relationship for each of the four citation measures. In Table A2, each of the seven columns shows one robustness check. Column 1 shows the results when when the most cited patent associated with each top 100 concept is reassigned from the innovative group of patents to the control group of patents.[14] Column 2 shows the results when the analysis is extended to the top $10,000$ concepts in each cohort. Column 3 shows the results when the comparison group for each patent is patents granted in the same *sub*class in the same year. Columns 4 and 5 show results when the analysis is restricted to either patents which first inventor is located in the US or patents which first inventor is in a foreign country. Columns 6 and 7 show results when the analysis is restricted to either patents that cite scientific literature or patents that do not cite scientific literature.[15] Tables A3 and A4, respectively, show the results by time period and by technology category.

Across the citation measures, robustness checks, time periods, and technology categories, the odds ratios shown in Table 2 and Tables A2, A3 and A4 indicate that the pattern already found in

[14] In this analysis, for each top 100 concept one of the innovative patents is reassigned as non-innovative, specifically the innovative patent that has received the most *No-Overlap Citations* relative to patents in its control group (i.e. for that patent the indicator variable measuring whether the patent is innovative is reassigned from 1 to 0). This analysis addresses the concern that the estimates are driven by citations received by patents for which a new concept is an output as opposed to an idea input. This approach establishes a lower bound for use of a new idea-citations link; the approach is not meant to suggest that the reassigned patent — or any patent for that matter — necessarily covers the concept in question.

[15] The domestic/foreign status indicator variable and cites/does not cite science indicator variable are additional predictors of both breakthrough invention status and the age of idea inputs. Patents with a domestic first inventor are two times more likely to be breakthrough inventions than patents with a foreign first inventor (Packalen and Bhattacharya, 2015b). Patents that cite science are 50% more likely to be breakthrough inventions than patents that don't cite science. We discuss the link between the use of new ideas and science citation status in the earlier version of this paper (Packalen and Bhattacharya, 2012).

Table 2: Estimates of the Age of Idea Inputs–Received Citations Relationship for Each Citation Measure. Sample time period is 1976–2005.

	(1)	(2)	(3)	(4)
Dep. Var.:	Top 5% by No-Overlap Citations	Top 5% by Total Citations	No-Overlap Citations	Total Citations
Model:	Conditional Logit	Conditional Logit	Poisson	Poisson
Top 5% by Age of Idea Inputs	2.488^{***}	2.589^{***}	1.878^{***}	1.598^{***}
	(.036)	(.036)	(.019)	(.011)
Patent Length	1.452^{***}	1.678^{***}	1.304^{***}	1.278^{***}
	(.009)	(.007)	(.004)	(.002)
Fixed Effects	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs
Number of Groups (Fixed Effects)	10717	10717	10534	10709
Observations	2784582	2784582	2783485	2784559

Reported coefficients are odds ratios (columns 1-2) and incidence rate ratios (columns 3-4). The odds ratio on patent length measures the effect of a one standard deviation increase in the variable. The model includes a separate fixed effect for each year-technology class pair. Observations are weighted so that observations from a given year received the same total weight as observations from any other year. Standard errors in parentheses; standard errors are clustered by the groups that corresponding to the fixed effects. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the non-parametric analyses is robust. Patents that are innovative in that they build on the most potent new ideas early receive much more citations than do other patents. Supporting the pursuit of such innovative research directions has measurable benefits, as such inventions spur much more subsequent advance than does the average invention.

5 Conclusion

Our quantitative analysis suggests that work that tries out a new idea early spurs much more subsequent advance than does other work. Encouraging and supporting innovative research that tries out new ideas thus appears to have an important benefit — avoiding stagnation in technological advance. The findings offer quantitative support for programs like the National Science Foundation's Small Business Innovation Research program (NSF SBIR), which supports innovative, risky, R&D projects that build on new ideas and have the potential to be transformative.[16]

Two additional contributions of the analysis are the development of a new approach for measuring idea inputs – we measure idea inputs from text – and the development of new patent data – we extend the span of available data from several decades to nearly two centuries. Given the central role of research inputs as drivers of technological and scientific progress and the centrality of patent data in analyses of invention, we expect these additional contributions to serve as fruitful research inputs in subsequent research.

One intriguing direction for future research is to examine to what extent scientific work that tries out new ideas spurs subsequent scientific advances. Findings obtained here need not extend to science because researcher incentives are different in science from what they are in invention (e.g. Aghion et al., 2008). The direction of invention is largely disciplined by the for-profit motive of firms, whereas scientists generally do not risk failure if they shun innovative ideas to protect

[16] According to NSF, '[t]ransformative research often results from a novel approach or new methodology' (National Science Foundation 2014a). The NSF SBIR program is designed to enable the pursuit of R&D projects that are 'based on innovative, transformational technology with potential for great commercial and/or societal benefits' (National Science Foundation 2014b).

the value of their own human capital and past ideas. Because entrenched interests can exclude innovative ideas with relative ease in science, the private and social benefits of trying out new ideas need not be aligned in science to the same degree that we expect them to be aligned in invention.

References

Aghion, P., Dewatripont, M. and J. C. Stein, 2008, "Academic Freedom, Private-Sector Focus, and the Process of Innovation," *RAND Journal of Economics*, vol. 39, pp. 617-35.

Ahuja, G. and C. M. Lampert, 2001, "Entrepreneurship in a Large Corporation: A Longitudinal Study of How Established Firms Create Breakthrough Invention," *Strategic Management Journal*, vol. 22, pp. 521-43.

Alcacer J., Gittelman M. and B. N. Sampat, 2009, "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis," *Research Policy*, vol. 38, pp. 415-27.

Alexopolos, M., 2011, "Read All about It!! What Happens Following a Technology Shock?" *American Economic Review*, vol. 101, pp. 1144-79.

Azoulay, P., Manso, G. and J. Graff Zivin, 2011, "Incentives and Creativity: Evidence from the Academic Life Sciences," *RAND Journal of Economics*, vol. 42, pp. 527-54.

Bhattacharya, J. and M. Packalen, 2011, "Opportunities and Benefits as Determinants of the Direction of Scientific Research," *Journal of Health Economics*, vol. 30, pp. 603-15.

Caballero, R. J. and A. Jaffe, 1993, "How High are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth," *NBER Macroeconomics Annual*, vol. 8, pp. 15-83.

Evans, J. A., 2010, "Industry Induces Academic Science to Know Less about More," *American Journal of Sociology*, vol. 116, pp. 389-452.

Fleming, L., 2001, "Recombinant Uncertainty in Technological Search," *Management Science*, vol. 47, pp. 117-32.

Grodal, S. and G. Thoma, 2009, "Cross-Pollination in Science and Technology: Concept Mobility in the Nanobiotechnology Field," *Annals of Economics and Statistics*, vol. 93.

Hall, B., Jaffe, A. and M. Trajtenberg, 2001, "The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools," NBER Working Paper No. 8485.

Hall, B. H., Jaffe, A. and M. Trajtenberg, 2005, "Market Value and Patent Citations," *RAND Journal of Economics*, vol. 36, pp. 16-38.

Harhoff, D., Narin, F., Scherer, F. M. and K. Vopel, 1999, “Citation Frequency and the Value of Patented Innovations,” *Review of Economics and Statistics*, vol. 81, pp. 511-5.

Jaffe A., Trajtenberg, M. and M. S. Fogarty, 2002, “The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Study of Patentees,” in Jaffe, A. and M. Trajtenberg (eds.) *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, MIT Press.

Jaffe, A., Trajtenberg, M. and R. Henderson, 1993, “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations,” *Quarterly Journal of Economics*, vol. 108, pp. 577-98.

Kuhn, D., 1962, *The Structure of Scientific Revolutions*. University of Chigaco Press.

Lai, R., D’Amour, A., Yu, A., Sun, Y. and L. Fleming, 2013, “Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010),” Mimeo.

Lampe, R., 2012, “Strategic Citation,” *Review of Economics and Statistics*, vol. 94, pp. 320-33.

March, J. G., 1991, “Exploration and Exploitation in Organizational Learning,” *Organizational Science*, vol. 2, pp. 71-87.

Marshall, A., 1920, *Principles of Economics*, 8th ed., London: Macmillan and Co.

National Science Foundation, 2014a, “Characteristics of Potentially Transformative Research,” web page (http://www.nsf.gov/about/transformative_research/characteristics .jsp; last accessed 8/20/2014).

National Science Foundation, 2014b, “Small Business Innovation Research Program Phase I Solicitation (SBIR) June 2014 Submission,” web page (http://www.nsf.gov/pubs/ 2014/nsf14539/nsf14539.htm; last accessed 8/20/2014).

Nerkar, A., 2003, “Old Is Gold? The Value of Temporal Exploration in the Creation of New Knowledge,” *Management Science*, vol. 49, pp. 211-29.

Packalen, M. and J. Bhattacharya, 2015a, “Age and the Trying Out of New Ideas,” Mimeo.

Packalen, M. and J. Bhattacharya, 2015b, “Cities and Ideas,” Mimeo.

Schoenmakers, W. and G. Duysters, 2010, “The Technological Origins of Radical Inventions,” *Research Policy*, vol. 39, pp. 1051-9.

Schumpeter, J., 1939, *Business Cycles*. McGraw-Hill: New York.

Strumsky, D., Lobo, J. and S. van der Leeuw, 2010, “Using Patent Technology Codes to Study Technological Change,” Santa Fe Institute Working Paper 10-11-028.

Utterback, J. M., 1996, *Mastering the Dynamics of Innovation*, Harvard Business School Press.

Usher, A. P., 1922, *History of Mechanical Inventions*, McGraw-Hill Book Company, New York.

U.S. Patent and Trademark Office, 2005, *Handbook of Classification*.

Weitzman, M., 1998, “Recombinant Growth,” *Quarterly Journal of Economics*, vol. 113, pp. 331-60.

Appendix: Additional Tables and Figures

Table A1: Lists of which idea inputs render patents innovative in each decade. Each embedded list below contains a decade-specific list of the idea inputs that render one or more patents in that decade innovative. Innovativeness of each patent is measured by the dummy variable *Top 5% by Age of Idea Inputs* (defined in section 2.4). A concept renders a patent innovative if the concept is the newest concept in a patent and the patent has the *Top 5% by Age of Idea Inputs* status. The six columns of each table contain the following items:

1. Decade.
2. Concept name.
3. Cohort of the concept (the year the concept first appeared in patents).
4. Number of times a mention of the concept in a patent during the decade in question renders the patent innovative, relative to the patent's comparison group. The comparison group for each patent is other patents granted in the same technology class in the same year.
5. Cumulative share out of all innovative patents during the decade in question (calculated based on column 4).
6. The total number of patents that mention the concept (during 1836-2010).

Within each decade, the concepts are ordered by column 4. To open an embedded list (a PDF file), click on one of the decades listed below (the links do not access the internet).

Top 100 New Idea Inputs by Cohort for 1880s

Top 100 New Idea Inputs by Cohort for 1890s

Top 100 New Idea Inputs by Cohort for 1900s

Top 100 New Idea Inputs by Cohort for 1910s

Top 100 New Idea Inputs by Cohort for 1920s

Top 100 New Idea Inputs by Cohort for 1930s

Top 100 New Idea Inputs by Cohort for 1940s

Top 100 New Idea Inputs by Cohort for 1950s

Top 100 New Idea Inputs by Cohort for 1960s

Top 100 New Idea Inputs by Cohort for 1970s

Top 100 New Idea Inputs by Cohort for 1980s

Top 100 New Idea Inputs by Cohort for 1990s

Top 100 New Idea Inputs by Cohort for 2000s

Table A2: Estimates of the Age of Idea Inputs–Received Citations Relationship: Robustness Checks. Sample time period is 1976–2005, the dependent variable is *Top 5% by No-Overlap Citations*, and the model is Conditional Logit.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reassignment Approach	Top 10,000 Concepts	Subclass Comparisons	Domestic Patentees	Foreign Patentees	Cites Science	Does Not Cite Science
Top 5% by Age of Idea Inputs	2.189***	2.271***	2.064***	2.250***	2.307***	2.175***	2.578***
	(.034)	(.030)	(.022)	(.036)	(.053)	(.050)	(.037)
Patent Length	1.463***	1.458***	1.427***	1.492***	1.387***	1.399***	1.441***
	(.009)	(.009)	(.004)	(.010)	(.011)	(.014)	(.007)
Fixed Effects	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Subclass Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs
Number of Groups (Fixed Effects)	10717	10717	125035	9864	8070	6734	10131
Observations	2784582	2784582	2706574	1511912	1214288	691570	2058878

Reported coefficients are odds ratios. The model includes a separate fixed effect for each year-technology class pair, except in column 3 where the model includes a separate fixed effect for each year-technology subclass pair. For further notes, please see notes to Table 2.

Table A3: Estimates of the Age of Idea Inputs–Received Citations Relationship by Time Period. The dependent variable is *Top 5% by No-Overlap Citations*, and the model is Conditional Logit.

	(1)	(2)	(3)	(4)
	1880s-1910s	1920s-1960s	1970s-1980s	1990s-2000s
Top 5% by Age of Idea Inputs	1.195***	1.580***	2.037***	2.534***
	(.023)	(.021)	(.031)	(.042)
Patent Length	.966***	1.080***	1.325***	1.471***
	(.006)	(.004)	(.007)	(.010)
Fixed Effects	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs
Observations	1101316	2158399	1398471	2076232
Number of Fixed Effects	13844	19314	8215	6608

Reported estimates are odds ratios. The model includes a separate fixed effect for each year-technology class pair. For further notes, see notes to Table 2.

Table A4: Estimates of the Age of Idea Inputs–Received Citations Relationship by Technology Category. Sample time period is 1976–2005, the dependent variable is *Top 5% by No-Overlap Citations*, and the model is Conditional Logit.

	(1)	(2)	(3)	(4)	(5)	(6)
	Chemical	Computers & Comm.	Drugs & Medical	Electronics	Mechanical	Other
Top 5% by Age of Idea Inputs	2.534***	2.120***	1.668***	2.696***	2.800***	2.680***
	(.071)	(.071)	(.130)	(.081)	(.069)	(.066)
Patent Length	1.445***	1.463***	1.256***	1.537***	1.447***	1.534***
	(.015)	(.014)	(.036)	(.012)	(.011)	(.011)
Fixed Effects	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs	Year-Tech Class Pairs
Number of Groups (Fixed Effects)	1973	1120	363	1385	2855	3021
Observations	468650	456738	268325	543265	518117	529487

Reported estimates are odds ratios. The model includes a separate fixed effect for each year-technology class pair. For further notes, see notes to Table 2.

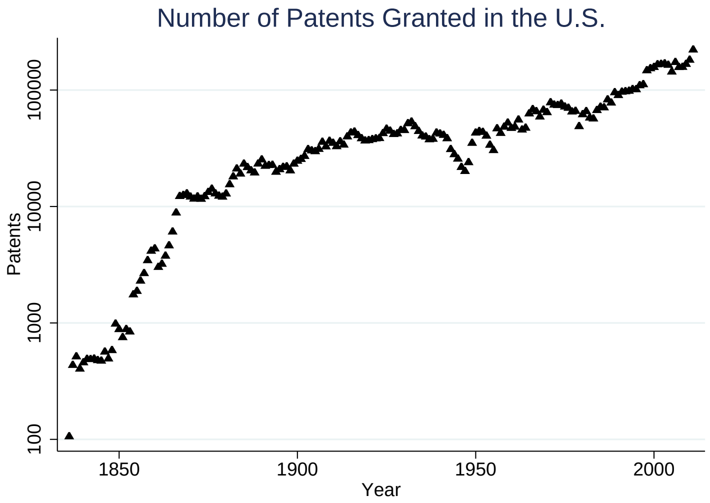


Figure A1: Number of US patents granted each year during 1836–2010.

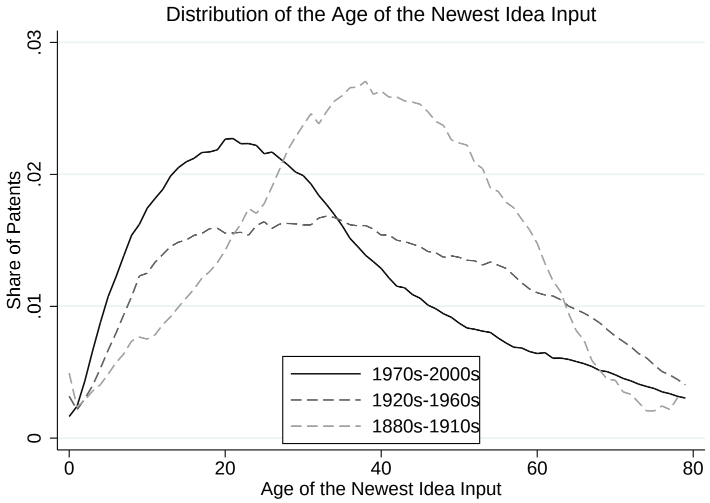


Figure A2: Distribution of the 5th percentile of the age of idea inputs for three time periods.

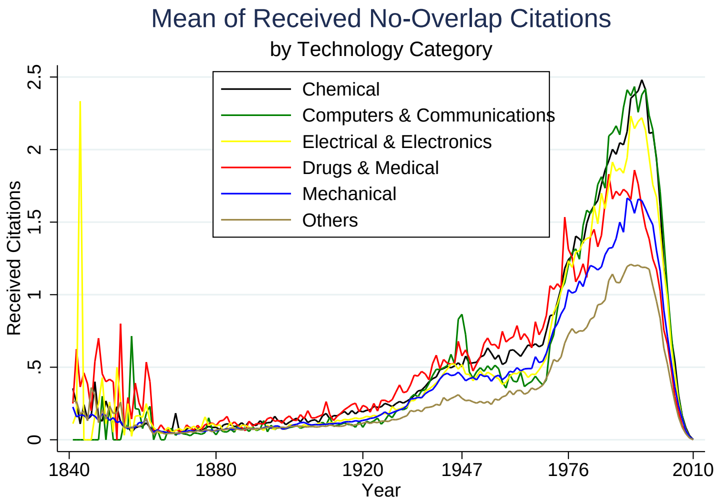


Figure A3: Mean of received *No-Overlap Citations* by technology category each year 1840–2010.

Data Appendix

Overview

Our main data source is US patents granted during 1836–2010 (“Patent Document Data”). For patents granted during 1976–2010 the patent documents are readily available in ASCII form (available here). For patents granted during 1836–1975 the patent documents are available as images of the original patents (available here). To transform these images to ASCII text, we apply an optical character recognition (“OCR”) program to these images of the 4 million patents granted during 1836–1975. From each patent we extract the following elements: patent text (title, abstract, brief summary, description, claims), citations to prior patents, and citations to scientific literature and other non-patent references. We limit the analysis to utility patents, which form the vast majority of patents.

Our other data sources are (1) Table of Issue Years and Selected Document Types Issued Since 1836 (“Grant Year File”, available here), which indicates the grant year for each patent number; (2) U.S. Patent Grant Master Classification File (“Master File”, available here), which indicates patents granted each year as well as patent classes assigned to each patent; (3) Mapping of Technology Classes to Technology Categories and Technology Category Labels (“Technology Category File”, available here) as developed by Hall et al. (2001), which links each 3-digit patent technology class to one of six broad technology categories; and (4) cleaned inventor location data for patents granted during 1975–2010 as developed by Lai et al. (2013), which we use perform a robustness check that limits the analysis to patents granted to inventors located in the U.S.

Differences to previous version in terms of data construction. (1) In the previous version of this paper (Packalen and Bhattacharya, 2012) we used patent document data for 1920–2010. Here, we extend the analysis to patent documents granted since 1836. (2) In the previous version we relied on an existing ASCII transformation of the images for patents granted during 1920–1975, whereas here we apply an OCR program to transform the images ourselves. This results in less OCR errors, enables us to extend the analysis to also to patents granted during 1836–1919 as well as patents granted during 1971–1975 comprehensively (the existing ASCII transformation had information on less than 50% of patents granted during this 5-year period). (3) We now cull through the list of top concepts in each cohort more carefully to exclude concepts that likely do not reflect idea inputs (see item 7.11 below); previously we excluded concepts that include any one of less than 100 words and character sequences, we now exclude concepts that include any one of more than 1,000 words or character sequences.

Details

1. Obtain and organize the Grant Year File. These data specify the number of the first utility patent granted each year. We use these data to determine the grant year of each patent in the Master File.

2. Obtain and organize the Master File, and link the Master File to the Grant Year File. The Master File specifies the patent number, a primary technology classification, and multiple secondary technology classifications for granted patents. The assigned classifications are updated when the classification system changes as a result of the introduction of new patent classes. Only the current version is available. We use version 1110 (mcfcls1110.txt, November 2011). We combine these data with the Grant Year File to obtain a list of patents granted each year, which enables us to examine the completeness of the patent document data. We use the assigned primary technology class for each patent in the Master File in determining the comparison group for each patent in the analyses (patents with the same primary technology class and grant year are compared to one another). Approximately 1000 patents have no assigned primary technology class. We assign each such patent to the technology class that appears most often among the secondary technology classes of the patent. The primary technology class is also combined with the Technology Category File to assign each patent to one of 6 technology category (see below). We use the primary and all secondary assigned technology classes of citing and cited patents in determining which citations are Non-Overlap Citations.

3. Obtain the Technology Category File. These data specify 6 broad technology categories and map each technology class to one category. The technology class 850 is not mapped in these data; we map it to technology category 4 (and subcategory 43 though we do not use subcategory information). We use the mapping to assign each patent to a technology category based on the patent's primary technology class, enabling us to (1) determine the technology category in which each concept appears most often and (2) obtain technology category specific estimates of the age of idea inputs–received citations link. We also use the technology category mapping to determine the technology categories spanned by the primary and all secondary technology classes of patent; this information is used in determining which citations are Non-Overlap Citations.

4. Obtain and organize the Patent Document Data.

4.1 Download and unzip the data. We use these data to determine the patent text, citations to patents, and citations to the scientific literature and other non-patent references.

4.2 Apply OCR algorithms to older (pre–1976) patent document data. This transforms images of patent document to ASCII files.

4.3 Store information on newer patents (post–1975) in patent-specific files. The data have multiple patents in each file but individual patents within each file are clearly indicated. Information on some patents appears in multiple places.

5. Determine which patents in the Patent Document Data appear also in the Master File. We only consider information in those patents in the Patent Document Data that appear also in the Master File.

6. Determine patent text. In the newer data, the fields we consider as patent text (title, abstract, brief summary, description, claims) are clearly indicated. In the older data, the fields are only indicated among the scanned text. For these data, we determine where patent text ends and citations begin by searching for indications of the presence of phrases such as "CITED REFERENCES" and "following references are of". Any text considered as being part of a bibliography is not included in the textual analysis.

7. Index words and 2- and 3-word sequences (Concepts) in each patent.

7.1 Construct a list of all text in a patent. We add a space character between text from different text fields and paragraphs.

7.2 Replace special characters with the space character. Exceptions: parentheses, brackets, and braces are deleted; period, comma, colon and semi-colon are replaced with “ X ” when followed by whitespace (so that indexed word sequences do not contain words from different sentences or words from different independent clauses of a sentence); period, comma, colon and semi-colon are replaced with the space character when not followed by whitespace (as in those cases the characters may reflect something other than punctuation that separates two sentences or two independent clauses within a sentence).

7.3 Change all alphabetic characters to lowercase. In principle, analysing to what extent mentions of a given concept begin with an upper case letter could be used to exclude concepts such as “Microsoft” that do not represent innovation inputs in the traditional sense. However, such an approach would also exclude important inventions such as “Teflon”.

7.4 Eliminate possessive case. Character sequence “ ’s ” is replaced with the space character. Character sequence “ s’ ” is replaced with the character “ s ”.

7.5 Replace certain control sequences with whitespace. We exclude character sequences such as “ldquo”, “apos”, “centerdot”, etc. and character sequences that begin with certain characters such as the character “ x ” followed by a number, which reflect changes in how patents are recorded.

7.6 Replace character sequences that have two or more consecutive numeric characters with whitespace. Concepts with multiple consecutive numeric characters are often page numbers, publication years and typographical control sequences.

7.7 Eliminate excess whitespace. All whitespace longer than the space character is replaced with the space character.

7.8 Extract all such words and 2-, and 3-word sequences from the list that satisfy character length limits on concept length and on length of individual words within concepts. We only extract 1-grams with 3-29 characters, 2-grams with 7-59 characters and 3-grams with 11-89 characters. We only extract 2- and 3-grams for which each word is at least 3 characters long.

7.9 Exclude concepts with DNA or RNA sequence information. We exclude all concepts that include one or more words that consist only of characters in the set “a,c,g,t” or the set “a,c,g,u.

7.10 Exclude concepts that include certain common words. Concepts for which any of the words is “the” are excluded. Concepts for which either the first or last word is a common word such as “than”, “and”, and “have” are excluded.

7.11 Exclude concepts which appearance as new concepts likely does not reflect new idea inputs We cull through the list of top 100 concepts for each cohort and exclude concepts to manually exclude concepts that likely do not reflect idea inputs. We delete, for example, concepts with words such as“filed”, “valid”, “jan”, “novel”, and concepts that include character sequences such as “natl acad”, “priority”, “application”, “provisional”, “federally” , “sponsored” and “envisi”. The complete list of words and character sequences that we use to eliminate concepts is included as an embedded file here (click to open an internal PDF file; does not access the internet). We exclude all concepts that include a word sequence in this embedded list.

7.13 Save the list of concepts that were not excluded.

8. Index concept cohorts and total mentions, and rank concepts in each cohort.

8.1 Combine Patent Document Data with the Technology Category File. We assign each patent to a technology category based on the technology class-technology category mapping in the Technology Category File and on the primary technology class of each patent.

8.2 Determine the cohort of each concept. Generally, the cohort of a concept is the year in which the concept first appeared in any patent. However, as indicted in the main text, we ignore the initial mention of concepts that are mentioned less than 5 times during the subsequent 25 years. For such concepts the cohort is set as the earliest year in which (1) the concept is mentioned, and (2) the concept is mentioned at least 5 times during the 25 years that follow.

8.3 For each concept, calculate the number of patents that mention the concept by the end of the sample. We also calculate how many times each concept appears in a given technology category.

8.4 Rank concepts in each cohort. We rank concepts based on the number of patents that mention each concept by the end of the sample.

9. Construct patent-specific variables measuring the age of the newest idea input. We construct this variable both based on mentions of the top 100 concepts in each cohort and based on mentions of the top 10,000 concepts in each cohort.

10. Index citations to patents and citations scientific and other non-patent references

10.1 Index Patent Citations. In the newer Document Data, patent citations are indicated in a separate field. In the older Document Data, patent citations are among the scanned text. We extract these patent citations by first searching for indications of the presence of phrases such as “CITED REFERENCES” and “following references are of” and then analyzing the text that follows. We extract the patent number, grant year and inventor name in each reference. We use the grant year and inventor name information in these citations to compensate for OCR errors in the following way: we only include citations for which either the cited grant year is within 10 years of the actual grant year of the cited patent or the first letter of the cited inventor name matches the first letter of the actual inventor name of the cited patent (the actual grant year is determined from the Master File and Grant Year File; the actual inventor name is determined from citations in the newer patent data to pre–1976 patents).

10.2 Index Citations to Scientific and Other Non-Patent References. In the newer Document Data, non-patent references are indicated in a separate field (there are additional non-patent references in the patent text but we do not consider them to limit the scope of the analysis). To distinguish scientific references from other non-patent references, we first search the non-patent references for terms that would indicate that the reference is to a patent reference, technical publication, marketing material, or web page (the searched terms include terms such as “ser. no.”, “patent”, “pat. appl”, “derwent”, “’database wpi”, “’search report”, “office action”,“advertisement”, “ibm technical bulletin”, “disclosure”, “language abstract”, “withdrawn”, “JP”, “EP”, “english translation”, “www.”, “website”, etc.). We designate references for which such terms are not found as potential scientific citations. Among the potential scientific citations we then search for an indication of a publication year (we first search inside parentheses for a 4-digit number between 1500 and 2015, and — when such sequence is not found — we then search for 2-digit numbers that follow either the character “ ’ “ or the character“ / ”.) Those citations for which a publication year is found are considered scientific references. In the older Document Data, non-patent references are among the scanned text. For these older data, we extract non-patent references by searching for indications of the presence of the phrase“OTHER REFERENCES” within the“CITED REFERENCES” section (see Index Patent Citations step above) and then search for publication years (a 4-digit number) Older patents for which such publication year is found among other references are assigned as having a non-patent reference. The search is stopped when an indication is found for the presence of phrases as“CERTIFICATE”, “FOREIGN PATENTS” or “CORRECTION”.

10.3 Disambiguate Scientific References To disambiguate the scientific references, we find citations that have the same publication year and are similar to one another. After indexing citations by publication year, we seek citations that have two double quotations (which typically surround a title) and examine which of such citations are similar to one another. We then extend the similarity comparison to all references among citations with to references with a given publication year. In these comparisons, we exclude certain character sequences such as “pages” that can be expected to be present in some citations to a scientific reference but not in other citations to the same reference. We also exclude character sequences that include non-alphabetic characters and character sequences that are shorter than 3 characters. We consider two citations to be to the same scientific reference when the SequenceMatcher comparison in Python returns a value above 0.9. We do not disambiguate Chemical Abstracts and references to GenBank accession numbers. We use the disambiguate scientific references to construct the table that lists the Top 20 most cited scientific references in each cohort (shown in the earlier version Packalen and Bhattacharya (2012) of this paper).

11. Index patent titles. In the newer Document Data, patent titles are indicated in a separate field. For the older Document data, we extract patent title based on the appearance of capital letters near the beginning of the text. We use these data to display patent titles in the table that lists the Top 20 most cited patents in each cohort (shown in the earlier version Packalen and Bhattacharya (2012) of this paper).