Research Policy 49 (2020) 103946

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

In-text patent citations: A user's guide

Kevin A. Bryan[a,*,1], Yasin Ozcan[b,1], Bhaven Sampat[c,1]

[a] University of Toronto, Rotman School of Management, 105 St. George Ave, Toronto, Ontario M5S3E6, Canada
[b] NBER, United States
[c] Columbia University, United States

A B S T R A C T

We introduce, validate, and make publicly available a new data source for innovation research: scientific references in patent specifications. These references are common and algorithmically extractable. Critically, they are very different from the "front page" prior art commonly used to proxy for inventor knowledge. Only 24% of front page citations to academic articles are in the patent text, and 31% of in-text citations are on the front page. We explain these differences by describing the legal rules and practice governing citation. Empirical validations suggest that in-text citations are qualitatively different. Further, there is suggestive evidence that they more accurately proxy for knowledge flows, consistent with their legal role.

1. Introduction

What prior knowledge do inventors and firms use as inputs to their research? These spillovers are at the core of urban economics, growth, and the economics of innovation. The empirical challenge is that "knowledge flows...are invisible; they leave no paper trail by which they may be measured and tracked" (Krugman, 1991). Trying to make these flows visible, researchers have surveyed firms (Levin et al., 1987; Cohen et al., 2000), written qualitative industry histories (Hippel, 1988), and investigated inventor biographies (Khan et al., 2014; Moser, 2005). However, by far the most commonly-used measure has been prior art cited on the "front page" of granted patents (Jaffe et al., 1993; Narin, 1994; Trajtenberg, 1990).

Several pioneering studies, beginning in the 1990s, investigated the impact of science on industrial innovation by looking at front page citations from private sector patents to academic patents (Jaffe et al., 1993; Henderson et al., 1998). But patents also cite non-patent references, including scientific literature which for two reasons may represent an even more promising way to evaluate the effects of public science. First, non-patent citations are less likely to come from examiners than patent-patent citations (Lemley and Sampat, 2012). Moreover, since most academic research is disseminated through publications (rather than patents) patent citations to publications capture a broader range of potential impact than patent-patent citations (Agrawal and Henderson, 2002).

That said, the interpretation of front page citations as knowledge flows is tenuous. These citations are not simply a list of earlier patents, academic articles, and other documents which were relevant to the invention. Rather, they derive from a legal "duty of disclosure" requiring inventors to list documents material to the patentability of their claims. Material to patentability means there is no need to cite tools that help build the invention, information used to avoid unpromising paths, basic facts that focus effort, or research suggesting a technological hole or a market need. That is, front page citations in patents are not analogous to references in academic papers.[2]

Consider an alternative measure of knowledge flows: citations in the specification text itself. Patent specifications by law must include enough detail that someone "skilled in the art" could replicate the invention. As Fig. 1 shows, inventors fulfill this requirement in part by citing literature while describing the background of their invention, why it is novel, and how it is built. These descriptive parts of patents are often largely written by the inventors themselves, rather than by patent attorneys who focus on the more legally consequential claims and prior art disclosure. Both their legal purpose and their practical construction suggest that in-text citations should better measure the real knowledge inventors use to motivate and construct their inventions.

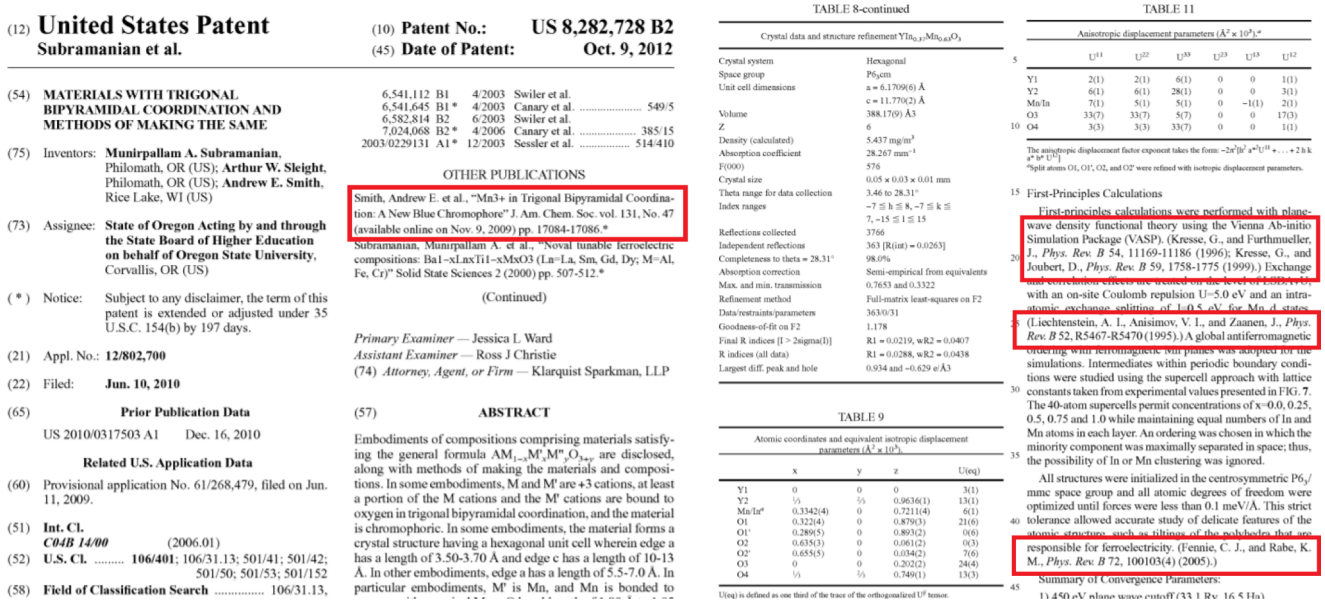
We thank the Alfred P. Sloan Foundation (grant G-2019-12272) for and the National Institute on Aging (award number R24 AG048059 to the National Bureau of Economic Research) for financial support.

* Corresponding author.
E-mail address: kevin.bryan@rotman.utoronto.ca (K.A. Bryan).

[1] The three coauthors contributed equally to all portions of this manuscript.

[2] See Meyer (2000) on the differences between academic and front page patent citation patterns.

https://doi.org/10.1016/j.respol.2020.103946
Received 2 April 2019; Received in revised form 9 December 2019; Accepted 14 February 2020
Available online 24 February 2020
0048-7333/



(12) **United States Patent**
Subramanian et al.

(10) **Patent No.: US 8,282,728 B2**
(45) **Date of Patent: Oct. 9, 2012**

(54) **MATERIALS WITH TRIGONAL BIPYRAMIDAL COORDINATION AND METHODS OF MAKING THE SAME**

(75) Inventors: **Munirpallam A. Subramanian**, Philomath, OR (US); **Arthur W. Sleight**, Philomath, OR (US); **Andrew E. Smith**, Rice Lake, WI (US)

(73) Assignee: **State of Oregon Acting by and through the State Board of Higher Education on behalf of Oregon State University**, Corvallis, OR (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 197 days.

(21) Appl. No.: **12/802,700**

(22) Filed: **Jun. 10, 2010**

(65) **Prior Publication Data**

US 2010/0317503 A1 Dec. 16, 2010

Related U.S. Application Data

(60) Provisional application No. 61/268,479, filed on Jun. 11, 2009.

(51) **Int. Cl.**
C04B 14/00 (2006.01)

(52) **U.S. Cl.** **106/401**; 106/31.13; 501/41; 501/42; 501/50; 501/53; 501/152

(58) **Field of Classification Search** 106/31.13,

6,541,112 B1 4/2003 Swiler et al.
6,541,645 B1 * 4/2003 Canary et al. 549/5
6,582,814 B2 6/2003 Swiler et al.
7,024,068 B2 * 4/2006 Canary et al. 385/15
2003/0229131 A1 * 12/2003 Sessler et al. 514/410

OTHER PUBLICATIONS

Smith, Andrew E. et al., "Mn3+ in Trigonal Bipyramidal Coordination: A New Blue Chromophore" J. Am. Chem. Soc. vol. 131, No. 47 (available online on Nov. 9, 2009) pp. 17084-17086.*

Subramanian, Munirpallam A. et al., "Novel tunable ferroelectric compositions: Ba1−xLnxTi1−xMxO3 (Ln=La, Sm, Gd, Dy; M=Al, Fe, Cr)" Solid State Sciences 2 (2000) pp. 507-512.*

(Continued)

Primary Examiner — Jessica L Ward
Assistant Examiner — Ross J Christie
(74) *Attorney, Agent, or Firm* — Klarquist Sparkman, LLP

(57) **ABSTRACT**

Embodiments of compositions comprising materials satisfying the general formula $AM_{1-x}M'_xM''_yO_{3+\gamma}$ are disclosed, along with methods of making the materials and compositions. In some embodiments, M and M' are +3 cations, at least a portion of the M cations and the M' cations are bound to oxygen in trigonal bipyramidal coordination, and the material is chromophoric. In some embodiments, the material forms a crystal structure having a hexagonal unit cell wherein edge a has a length of 3.50-3.70 Å and edge c has a length of 10-13 Å. In other embodiments, edge a has a length of 5.5-7.0 Å. In particular embodiments, M' is Mn, and Mn is bonded to

TABLE 8-continued

Crystal data and structure refinement $YIn_{0.37}Mn_{0.63}O_3$

Crystal system	Hexagonal
Space group	$P6_3cm$
Unit cell dimensions	a = 6.1709(6) Å c = 11.770(2) Å
Volume	388.17(9) Å3
Z	6
Density (calculated)	5.437 mg/m³
Absorption coefficient	28.267 mm⁻¹
F(000)	576
Crystal size	0.05 × 0.03 × 0.01 mm
Theta range for data collection	3.46 to 28.31°
Index ranges	−7 ≤ h ≤ 8, −7 ≤ k ≤ 7, −15 ≤ l ≤ 15
Reflections collected	3766
Independent reflections	363 [R(int) = 0.0263]
Completeness to theta = 28.31°	98.0%
Absorption correction	Semi-empirical from equivalents
Max. and min. transmission	0.7653 and 0.3322
Refinement method	Full-matrix least-squares on F2
Data/restraints/parameters	363/0/31
Goodness-of-fit on F2	1.178
Final R indices [I > 2sigma(I)]	R1 = 0.0219, wR2 = 0.0407
R indices (all data)	R1 = 0.0288, wR2 = 0.0438
Largest diff. peak and hole	0.934 and −0.629 e·Å3

TABLE 9

Atomic coordinates and equivalent isotropic displacement parameters (Å² × 10³).

	x	y	z	U(eq)
Y1	0	0	0	3(1)
Y2	⅓	⅔	0.9636(1)	13(1)
Mn/In*	0.3342(4)	0	0.7211(4)	6(1)
O1	0.322(4)	0	0.879(3)	21(6)
O1'	0.289(5)	0	0.893(2)	0(6)
O2	0.635(3)	0	0.061(2)	0(3)
O2'	0.655(5)	0	0.034(2)	7(6)
O3	0	0	0.202(2)	24(4)
O4	⅓	⅔	0.749(1)	13(3)

U(eq) is defined as one third of the trace of the orthogonalized U^{ij} tensor.

TABLE 11

Anisotropic displacement parameters (Å² × 10³).*

	U^{11}	U^{22}	U^{33}	U^{23}	U^{13}	U^{12}
Y1	2(1)	2(1)	6(1)	0	0	1(1)
Y2	6(1)	6(1)	28(1)	0	0	3(1)
Mn/In	7(1)	5(1)	5(1)	0	−1(1)	2(1)
O3	33(7)	33(7)	5(7)	0	0	17(3)
O4	3(3)	3(3)	33(7)	0	0	1(1)

The anisotropic displacement factor exponent takes the form: $-2\pi^2[h^2 a^{*2}U^{11} + \ldots + 2\,h\,k\,a^*b^*U^{12}]$

*Split atoms O1, O1', O2, and O2' were refined with isotropic displacement parameters.

First-Principles Calculations

First-principles calculations were performed with plane-wave density functional theory using the Vienna Ab-initio Simulation Package (VASP). (Kresse, G., and Furthmueller, J., *Phys. Rev. B* 54, 11169-11186 (1996); Kresse, G., and Joubert, D., *Phys. Rev. B* 59, 1758-1775 (1999).) Exchange and correlation effects are treated on the level of LSDA+U with an on-site Coulomb repulsion U=5.0 eV and an intra-atomic exchange splitting of J=0.5 eV for Mn d states (Liechtenstein, A. I., Anisimov, V. I., and Zaanen, J., *Phys. Rev. B* 52, R5467-R5470 (1995).) A global antiferromagnetic ordering with ferromagnetic Mn planes was adopted for the simulations. Intermediates within periodic boundary conditions were studied using the supercell approach with lattice constants taken from experimental values presented in FIG. 7. The 40-atom supercells permit concentrations of x=0.0, 0.25, 0.5, 0.75 and 1.0 while maintaining equal numbers of In and Mn atoms in each layer. An ordering was chosen in which the minority component was maximally separated in space; thus, the possibility of In or Mn clustering was ignored.

All structures were initialized in the centrosymmetric $P6_3/mmc$ space group and all atomic degrees of freedom were optimized until forces were less than 0.1 meV/Å. This strict tolerance allowed accurate study of delicate features of the atomic structure, such as tiltings of the polyhedra that are responsible for ferroelectricity. (Fennie, C. J., and Rabe, K. M., *Phys. Rev. B* 72, 100103(4) (2005).)

Summary of Convergence Parameters:

1) 450 eV plane wave cutoff (33.1 Ry, 16.5 Ha)

Fig. 1. Example of front page citations (left) versus in-text citations (right)

Despite this theoretical advantage in tracking knowledge flows, innovation researchers have not used in-text citations for two reasons.[3] First, there is a folk belief, which we will show is untrue, that citations in the text almost always appear on the front page. Second, since patents do not have bibliographies and in-text citations have no standardized format, they are difficult to extract.

Our primary contribution in this paper is to develop a practical algorithm for extracting these citations. Using this algorithm, we produce and describe a new public database of all 2,786,041 front page and in-text citations to every article since 1984 in 248 journals across 20 fields, available at http://doi.org/10.7910/DVN/ZEZWBX. The difference between the two types of citations is dramatic. Only 31% of in-text citations appear on the front page of a given patent, and only 24% of front page citations appear in the patent text.[4]

After describing the formal and practical differences in the legal origins of in-text versus front page citations, our secondary contribution is an early validation of in-text citations through four empirical investigations. First, we investigate the type of academic research applicants cite. In-text and front page citations draw on different academic journals: over 22% of citations would have to be to a different journal in order to equalize the frequency distributions. Further, for medical citations where we can map to NIH MeSH codes capturing the nature of the citation, the frequency distributions of in-text versus front page citations strongly differ. That is, on the questions of which types of biomedical research are used by inventors, or which journals have research spillovers to inventors, in-text citations are qualitatively different from front page citations, not merely a less noisy proxy for knowledge transfer.

Second, we investigate whether in text citations better proxy for ground truth about knowledge flows to inventors using "patent-paper pairs" where a biotechnology patent derives from the same research as an academic article (Murray and Stern, 2007). References in the underlying academic article bibliography are cited 68–138% more often in the text of corresponding patents than on the front page. Third, we confirm that result by showing that the number of in-text citations to academic research in a firm's patents is more strongly correlated with the firm R&D manager's stated reliance on public sector research and open science, using data from the survey in Roach and Cohen (2013).

Finally, we discuss a variable used in citation studies for which in-text citations are a worse proxy: measuring patent value. Patents with more front page scientific citations are more valuable according to four separate measures, and the link between in-text citations and value is generally either nonexistent or less strong. This ought not be surprising: we might expect inventors and their attorneys to perform more comprehensive prior art searches for patents they expect will be more valuable (Sampat, 2010).

The in-text scientific references in this paper and its associated database complement both the recent broad expansion in techniques exploiting patent text, and recent analyses of front page scientific citations. The first generation of patent-based studies largely used count measures like patent classes, the number of forward and backward front page citations, and so on. A series of recent papers construct metrics from raw patent text using modern computing power and natural language processing. For example, Kelly et al. (2017) identify impactful patents by examining text that is unlike existing patents but similar to future patents, Kuhn and Thompson (2017) use the length of claims to measure scope, and Kaplan and Vakili (2015) use topic modeling to identify breakthrough patents. In addition to using patent text to study innovation, several research teams have recently taken advantage of computational advances to extract front page citations to academic research and to link them to scientific literature databases (e.g., Marx and Fuegi, 2019). These data have been used for analyzing the impact of NIH research on private sector patenting (Azoulay et al., 2015; Li et al., 2017), the impact of publicly funded energy research on applied technology (Popp, 2017), the economic impact of universities (Jefferson et al., 2018), and the distance between science and technology in different fields (Ahmadpoor and Jones, 2017).

[3] The only previous paper using individual in-text references, by two of the present authors, investigated citations to articles in 44 medical journals while studying the effect of the NIH open access mandate (Bryan and Ozcan (2019)). The only large-scale institutional attempt we know of to extract in-text references is an EPO trial beginning in 2006 to include a "summary of references for the reader's convenience" with in-text patent and academic references attached to patent pdfs (EPO (2007)). We are unaware of any research in economics or innovation that has used this measure. At an aggregate level, Tamada et al. (2006) attempts to find "regular expressions" in the text that look like references to academic science, then counts these across various classes in Japanese patents. Recent work in progress by Matt Marx and coauthors is attempting to scale up a database of these references to a broader set of journals.

[4] Our publication data contains all articles in selected journals from 1984 to 2016, and our patent dataset includes all public US patents as of May 2018.

2. The empirics of in-text citations

The major practical difficulty with using in-text references is that they have no standard format. Patents do not have bibliographies, and references to publications are made in the flow of text. For instance, US6551784 writes that "methods for aligning sequences using the CLUSTAL program are well described by Higgins and Sharp in Gene, 73: 237–244 (1988) and in CABIOS 5: 151–153 (1989))." Extracting that reference requires knowing both that there are two references to academic publications in this sentence, and that the CABIOS article refers to one written by Higgins and Sharp. Some in-text references are incredibly vague, such as "Genomic DNA was obtained from leaf tissue according to Doyle and Doyle (1987)" in US6483012.

The most natural approach to follow is a coarsened match between article metadata and text in a patent. Coarse matches permit mis-spellings, various combinations of metadata, and so on.[5] The problem in our setting is that identifying a chunk of text as containing a scientific reference is itself a challenging problem in the absence of a bibliography or the one-line-represents-one-citation format of front page citations. We would therefore need to apply the coarsened algorithm directly to over a terabyte of text.

To avoid false positives and limit computational burden, our algorithm relies on limiting the text searched to only the blocks likely to contain a reference to science. We begin by dropping any paragraph which does not contain a four-digit year, then index every word in these paragraphs, search for one of thousands of potential journal name abbreviations, and attempt to match partial article metadata like title keywords, author names, and page numbers to the text surrounding the year and journal abbreviation. This algorithm can even capture citations where metadata are not located near each other in the patent text. For example, in US6605754, "Comai et al have previously described a chimeric plant promoter combining elements of the CaMV35S and the mannopine synthase (mas) promoters (1990, Plant Mol Biol, 15:373-381)" is correctly identified despite the author name being nowhere near the other metadata, and the Higgins and Sharp paper in CABIOS discussed above is also found. We discuss the algorithm in detail in Appendix B. Manual investigation suggests a false positive rate well below 1%.

We use this procedure to extract all in-text and front page references to every research article published between 1984 and 2016 in 248 prominent journals drawn from fields as diverse as biology, medicine, engineering, computer science, and social science. These 3,389,853 articles have been cited collectively 2,786,041 times in patents granted since 1984, with 1,573,143 citations of the front page and 1,212,898 in-text. 9.7% of the articles are cited at least once, with the probability of being cited highest for biomedical articles. 6.0% are cited at least once in-text, and 7.9% are cited at least once on the front page. Very recent articles are, of course, unlikely to be have been cited yet, so these figures understate the general citation propensity. For articles published in 2000, nearly 16% have been cited at least once in the full-text and/or on the front-page of issued patents.

The critical fact about in-text citations is their lack of overlap with front page citations: only 24% of the front page citations are cited in-text in the same patent, and only 31% of the in-text citations are cited on the front page.[6] That is, patentees use the two types of citations in very different ways.

Table 1 shows summary statistics on front-page and full-text citations. The lack of overlap between in-text and front page citations occurs across patentee types, geographies, time, and industry. 24% of front page citations by American inventors appear in-text, and 31% of in-text citations appear on the front page; for foreign inventors, the percentages are 24% and 32%. Patents assigned to academic research institutions have overlap of 31% and 36%, while those assigned to organizations outside academia have overlap of 21% and 29%. So-called "triadic" patents, which are filed in the US, Europe, and Japan-(this is often used as in indicator for more important patents) have overlap of 23% and 32%, while nontriadic patents have overlap of 26% and 30%.[7] Patents in medical or biotechnology classes have overlap of 27% and 31%, while those in other fields have overlap of 19% and 32%.[8] To ensure our results are not being driven by patentees who "flood" examiners with hundreds of references, we can restrict to patents with fewer than 20 front-page and 20 in-text references; the overlaps in this restricted set are 26% and 30%.[9]

In Appendix B, we provide complete details on our algorithm and the sample of journals we attempt to match to patents, show that the relative distributions of in-text and front-page citations are similar and highly skewed, give examples of how our algorithm treats various types of citations, and describe more precisely how we classify "university", "biomedical" and other groups mentioned above.

3. Why front page and in-text citations differ

It has long been known that front page citations may miss references to prior scientific work used by inventors. Indeed, one of the first papers to empirically examine front page citations notes that in-text references "may be more related to the history, usefulness, and development of the invention." Nonetheless, they used front page citations since they are "far easier to extract" (Narin and Noma, 1985).

Summary statistics on these front page citations, especially for citations to academic research, have been reported in OECD documents and in the NSF Science and Engineering Indicators since the 1990s. The number of front page citations is one measure used by funders, including the NIH, to investigate the impact of their grants. Front page citations to patents are widely used as measures of knowledge flows between inventors, and as proxies for the importance, value, or quality of the cited inventions (Jaffe and Rassenfosse, 2017). Recent research suggests front-page citations to academic patents are not strongly correlated with survey-based indicators of the extent to which firms rely on public research; front-page citation to non-patent literature is more strongly correlated (Roach and Cohen, 2013). However qualitative and historical studies suggest that even front page citations to science appear to map poorly into the directly measured prior knowledge of inventors (Tijssen, 2002; Meyer, 2000).[10]

To understand more clearly why patents cite such a radically

[5] E.g., US6190856 has an article by Erkki Koivunen cited as "Korvunen"; these mis-spellings are not uncommon. US6130090 cites a Bradley and Liu paper giving the year as 1996 instead of the correct 1997. US630824 cites a 1989 paper as being published in "Genetics" when it actually appeared in the journal "Genomics"; ironically, this article was written by one of the inventors!

[6] This difference is not a result of misclassifications by the matching algorithm. In Section 4, we perform two robustness checks where in-text and front page citations are classified by hand; within these sets the overlap is similarly small.

[7] Triadic patents are often considered a proxy for high value patents. It is therefore interesting to note that while triadic and non-triadic patents have an almost identical number of in-text citations conditional on having at least one (5.79 and 5.59, respectively), triadics have far more front page citations (6.65 and 4.68). We return to this point in the third empirical exercise in Section 4.

[8] Medical patents make up 59% of patents citing at least one article in-text, and 51% of those citing at least one on the front page.

[9] Dropping any *article* which is cited more than 20 times on the front page or in-text, to remove potential skew due to highly-cited articles which are ritualistically cited by examiners or copied across patents by lawyers at a given company, gives overlap of 23% and 36%.

[10] Nearly half of front page citations to patents are added by examiners (Alcácer, Gittelman, and Sampat (2009)). Only around 5% of front page citations to academic articles are added by the examiner (M. A. Lemley and Sampat (2012)), so the examiner problem is less severe for these types of citations, though of course it still unclear whether lawyers, inventors, or someone else inside the inventing firm added a given reference.

Table 1
Summary statistics on in-text versus front-page citations.

	All	US	Non-US	Univ	Non-Uni	Triadic	Non-Tri	Biomed	Non-BM
N of Patents	342667	227473	115194	78142	264525	178191	163511	168472	173230
Avg. # of In-Text	3.54	4.05	2.54	5.27	3.03	3.83	3.23	5.42	1.72
Avg. # of Front Page	4.60	5.22	3.35	6.21	4.11	5.41	3.70	6.24	2.98
Share of In-Text on FP	31.1	30.9	32.0	35.9	28.7	32.0	30.0	30.7	32.4
Share of FP In-Text	24.0	23.9	24.3	30.1	21.1	22.7	26.2	26.7	18.6

different set of prior knowledge in their specification text than on the front page, and what this difference means for scholars of innovation, let us examine how patentees legally ought to use these citations, and how they do so in practice.

Consider first front page citations, whose legal origin lies in patentees' "duty of disclosure". Everyone involved in filing a patent "has a duty…to disclose to the Office all information known to that individual to be material to patentability" (37 CFR 1.97). That is, *any* individual involved in the filing of the patent, whether the initial inventors, a drafter, a patent agent, or a patent attorney, must disclose on an Information Disclosure Statement (IDS) any prior publications they know which are relevant to the novelty and/or non-obviousness of the invention.[11] These individuals must cite prior art even if they learn of it years long after the invention is complete.[12] If known prior art is not disclosed, there is a risk the patent office will find "inequitable conduct," which is grounds for unenforceability of a patent (Cotropia et al., 2013). The documents on the IDS, alongside relevant prior art found by a patent examiner, make up the front page citations in a granted patent. Front page citations therefore represent a list of prior documents, known by any person involved in preparing the patent, which might be relevant to the novelty or non-obviousness of the patent's claims.

In-text citations have a completely different legal origin. The specification must "enable" the invention by describing its background, showing how it solves a useful problem, and showing how a person "skilled in the art" can make and use it without excessive experimentation. The applicant is not required to cite anything formally in the patent specification text, since she can simply describe the invention's background and method of construction using text and graphics. Often, it is easier to "incorporate by reference" aspects of the background and method.[13] For instance, a patent application for a new cancer drug could describe the method of discovery by writing "we inject our mice with cancer using the technique developed by Smith (2017)." The reference to Smith replaces lengthy details on how exactly that injection method works.[14] These in-text citations, therefore, serve a role more like academic citations than front page citations. Knowledge flows like basic motivating facts, open scientific puzzles, and tools used to construct the invention, are often part of an invention's method and background but not material to patentability.

3.1. Patent strategy and practice

In addition to laws on the books discussed above, patent practice and legal strategy will shape what we see cited on the front-page vs. in the full-text practice.

Based on our understanding from legal scholars and practitioners, inventors are typically more involved in drafting patent specifications than they are in prior art searches. And once drafted, the patent specification typically doesn't change much (technically it cannot without filing a continuation-in-part, which may or may not benefit from the priority date of the original application; MPEP 201.07). So in-text citations generally are generated around the time the original application is drafted, sometimes based on citations in provisional applications or scientific articles accompanying the patent.

As noted above, front page citations come from applicant information disclosure statements (IDS, PTO form 1449) and the examiner search that typically follows (reported on the PTO Form 892). In contrast to full-text citations, these front page citations come in over the course of prosecution process, and include inputs from attorneys who prepare the IDS based on information from inventors and their own prior art searches, and examiners' own searches at the USPTO.[15]

Importantly while the "duty of candor" applies to the entire patent process, failure to enable an invention would not consititute a violation of it (though may result in an enablement rejection) whereas failure to disclose known prior art on an IDS would. (MPEP 37 CFR 1.56). Violation of the duty of candor involves severe penalties for applicants and their attorneys, and could render the subject patent unenforceable. In this context, applicants (and attorneys) many err on the side of caution in their front page citations, since there is no penalty for citing too much. There is also some argument that attorneys may "flood the patent office" to hide actual relevant references (Taylor, 2012), though there is no strong evidence on this. On the other hand, it is also known that in some fields a substantial share of patents include little or no applicant provided art, possibly because inventors don't read competitors' patents for fear of willful infringement liability or don't care much about the validity of any given patent, but instead about accumulating large portfolios.[16]

These strategic aspects of front-page citation practice, and the legal meaning of front-page citations above would seem to call into question some of the early assumptions in citation analysis, such as Trajtenberg's assumption that the front-page citation process "apparently does generate the right incentives to have all relevant [prior art] cited, and only those" (Trajtenberg, 1990).

[11] Brasseler, U.S.A. I, L.P. v. Stryker Sales Corp., Fed. Cir. 2001: "Once an attorney, or an applicant has notice that information exists that appears material and questionable, that person cannot ignore that notice in an effort to avoid his or her duty to disclose."

[12] Another complicating factor is that there is a strategic aspect to whether applicants search for or cite prior art (Lampe (2012), Sampat (2010)) and heterogeneity among patent examiners in the extent to which they do so (M. A. Lemley and Sampat (2012)).

[13] See 37 CFR 1.71 and 37 CFR 1.57.

[14] The situation is slightly different outside the United States. European patents, for instance, do not have a duty of disclosure, and therefore tend to have fewer front page citations; they also operate under the requirement that the patent specification is interpretable by a more specialized reader than a U.S. patent, and hence may also see differences with in-text references. That said, the nature of in-text citations in European patents may not be much different than in the United States. European specifications must "indicate the background art which, as far as is known to the applicant, can be regarded as useful to understand the invention" (EPC Rule 42.1(b)). We have not validated the empirical properties of in-text citations for non-U.S. patents, but at least by the letter of the law, the qualitative distinction between front page and in-text citations is the U.S. and Europe is not large.

[15] As an empirical matter front-page citations to non-patent literature are less likely to be listed as "added by examiner" than front-page citations to patents (M. A. Lemley and Sampat 2012). Based on published PTO data about 4 percent of non-patent references are from examiners, compared to 40 percent of patent references.

[16] The duty of candor does not require affirmative search, only disclosure of known prior art. See Sampat (2010) for more discussion.

But there are potential strategic aspects to in-text citation as well. Inventors and attorneys may want to provide enough information to satisfy the disclosure requirement, but not enough to "actually" enable competitors to practice the invention.[17] Both legal scholarship and the economics literature provide mixed evidence on how much patents actually disclose, and how seriously applicants take the disclosure requirement (Ouellette, 2011; Devlin, 2009; Fromer, 2008; Cohen et al., 2000) though none of the literature focuses on in-text citation practices per se.

Finally, some attorneys and law firms consider it best legal practice to reference everything in the specification on the front-page because there is little cost to doing so. However, there is no legal requirement to do so, since, as emphasized in the previous section, prior art citations and citations in the specification have different purposes. Only the in-text citations that also bear on novelty and/or non-obviousness need to be actually need to be cited on the front-page, though actual citing practices may vary by firms, attorneys, and even the importance of specific inventions.

We should clarify that much of the existing empirical literature on strategic citation involves citations to patents, not citations to non-patent literature (NPL) like academic research. The legal role of NPL prior art is not different from patent-based prior art: any prior publication can limit claims, and omission of any prior publication violates the duty of disclosure. Nonetheless, direct research on how inventors and patent attorneys treat NPL as a strategic decision relative to patent citations is limited.

3.2. Timing of references

As we have seen, patents are not submitted all at once. Rather, applicants submit different forms and information to the patent office over time, update and revise their patent draft, and update the Information Disclosure Statement when they become aware of new relevant prior art. We can use the dynamic nature of the application process to directly investigate which citations applicants were aware of at the time of the initial disclosure to the patent office.

In the case of in-text citations, we investigate the initial application of all granted patents who application was first filed between 2001 and April 25, 2017.[18] We run our algorithm on both the application text and the text of the final granted patent. Only 7.3% of the in-text citations in granted patents were not in the initial application.[19] This is consistent with the legal argument above: changing the text enabling the invention risks changing the priority date of the patent.

On the other hand, applicants *must* update their Information Disclosure Statement when they become aware of a publication which is potentially material to the patent's claims. In the most comprehensive analysis of this topic, Kuhn et al. (2019) note that “many applicant citations are submitted long after the application is filed, in successive rounds of submitting forms to the USPTO, where citations become less and less similar.” Using data on front-page patent-patent citations in all patents issued between 2005 and 2014, they find that 72% of these references not submitted at initial filing, including examiner citations (28%) and late arriving applicant art (44%).

To our knowledge this figure has not been directly computed for front-page patent-NPL citations.[20] To examine this, we traced by hand a random sample of 100 frontpage non-patent references in patents from the Cotropia et al. (2013) dataset. (The Cotropia dataset starts with a 1% random sample of patents issued in 2007). Specifically, we downloaded and reviewed all information disclosure statements and other forms for these patents, and determined when a front-page NPL reference cited in a patent first appeared during prosecution. Of these 100 front page NPL references, we found that 3 were from the examiner, and additional 31 were not on the original information disclosure statement. That is, at least 34% were not in the original disclosure.

Since this is an admittedly small sample, we also used simple text matching algorithms to locate over nearly 8000 non-patent references from the Cotropia dataset across digitized information disclosure statements. While this is less precise (because of differences in formatting in the IDS and final patent, and OCR errors, and errors in matching), the results were similar to the manual match. Specifically for the 7000 references to non-patent literature we were were able to locate on any IDS, 36% were not on the earliest IDS.

To summarize, while 93% of in-text citations to science are in the original published application, only about two-thirds of front-page patent-NPL references are in the initial information disclosure statement. To the extent that later citations are prime facie less useful as indicators of knowledge flow (Kuhn et al., 2019) it would seem in-text references perform better.[21]

4. Validating in-text citations

We have shown that in-text citations are algorithmically extractable, have little overlap with front page citations, and legally ought play a different role than in-text citations. Before investigating how in-text and front page citations perform empirically in a variety of settings, consider how a researcher should evaluate proxy variables more generally.

Assume that a parameter of interest has ground truth σ. For instance, σ may be a binary variable representing that an inventor knew of and used the information in a given academic article in her invention, or a continuous variable representing the degree of similarity between that article and the invention. Let $\hat{\sigma}$ represent a proxy for σ, where $\hat{\sigma} = f(\sigma) + \epsilon$. If f is the identity function and $\hat{\epsilon}$ is always equal to zero, then the proxy variable perfectly represents the ground truth. Otherwise, the proxy may be biased, noisy, or both.

Suppose a researcher wishes to estimate a function $y = f(\sigma, X) + \mu$, where X are other covariates and μ is noise. For instance, y may represent the value of invention as a function of its similarity to pre-existing inventions conditional on field, or the probability of using a piece of knowledge in a new invention as a function of geographic distance from the originator of that knowledge.

The question of what makes a “good” proxy variable is a well-studied instance of the problem of how to use latent variables. It may be surprising to some readers that even proxies which are positively correlated with the true σ can generate coefficients of the wrong sign in a linear regression (e.g., Krasker and Pratt, 1986). More broadly, properly accounting for errors-in-variables reduces “true” statistical power, and hence poorly correlated proxy variables may overstate confidence in a particular relationship. We therefore wish to answer three questions about our “new” proxy variable. First, are there economically interesting questions where in-text and front page citations measure qualitatively different things? Second, are there economically interesting questions where in-text citations are better correlated with a ground truth σ of interest? Third, are there economically interesting questions

[17] As a practical matter, only 35 percent of applications get a rejection on Section 112 grounds - failure to meet disclosure - while 72 percent do for non-obviousness reasons, i.e. in light of prior art on the IDS or found by examiners. See Frakes and Wasserman (2017).

[18] Application text is not available prior to that date.

[19] And *every* citation in the final grant was in the initial application.

[20] Given that these are less likely to come from examiners (Christopher A Cotropia, Lemley, and Sampat 2013; Lemley and Sampat 2012) we would intuitively expect fewer “late” citations for front-page NPL references than front-page patent references.

[21] There are two caveats here. First, patent applications are published 18 months after issue, but to our understanding the specification does not change much between initial filing and publication for reasons already discussed. Second, our figures above are for all non-patent art, not “science” references per se.

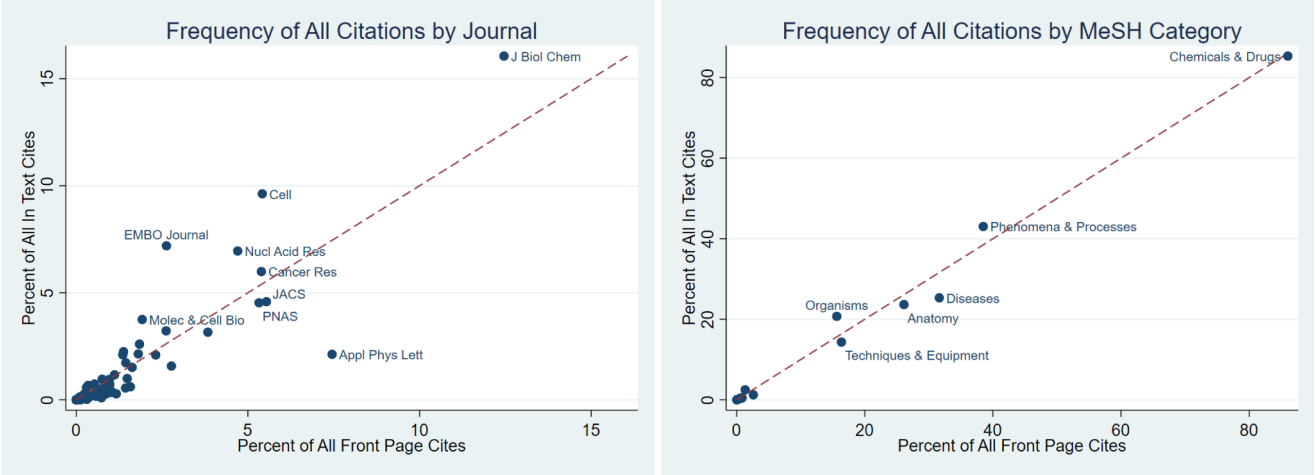


Fig. 2. Relative frequency of citations to a given journal (left) or to articles with a given non-exclusive Pubmed MESH code (right)

for which front page citations have been used, where front page citations are *better* proxies? The answer to all three questions is yes. We caution the reader, however, not to misinterpret how much is known about the proper use of in-text citations: much more work is needed to fully delineate when a researcher would prefer them to traditional front page measures.

4.1. Do in-text and front page citations measure the same things, noisily?

Let σ represent some feature of knowledge used by an inventor. Concretely, let σ represent the probability an inventor draws on the knowledge in a particular scholarly journal, or draws on research studying a particular method or field. Assume our two proxies $\hat{\sigma}_{it}$ and $\hat{\sigma}_{fp}$ represent noisy observations of σ; that is, $\hat{\sigma}_{it} = \sigma + \epsilon_1$ and $\hat{\sigma}_{fp} = \sigma + \epsilon_2$ where ϵ_1, ϵ_2 are equally-biased noise with potentially different variance.

Note that we can directly test that the two proxies are both equally biased but noisy versions of the ground truth. If, for instance, 6% of the scientific knowledge inventors use comes from the New England Journal of Medicine, or 20% of the knowledge biomedical inventors use has to do with properties of organisms, then as the number of citations of each type we observe increases, the probability they differ *from each other* will converge to zero.

Empirically, we perform two tests in line with this intuition. First, we construct the distribution of all 2,786,041 in-text and front page citations across our 248 journals. 22% of all citations would need to be switched to a different journal is order to create an equivalent distribution between in-text and front-page citations. A χ^2 test of equality of distributions across these categories rejects equality ($p < .00001$). The differences in citation frequency to even prominent journals is extreme. In-text citations would have to be directed toward the journal Applied Physics Letters at 3.5 times their actual rate to equal the relative frequency of citation that journal receives on the front page. Likewise, front page citations to the EMBO Journal would need to be 2.74 times as frequent to match their relative frequency in the patent text. These are not small sample issues: both jounals are cited more than 100,000 times in our data. We plot relative frequencies for all journals in the left panel of Fig. 2.

The same incongruence occurs if we examine what types of knowledge is being cited. For the 39% of citations that can be linked to Pubmed, cited articles can be categorized by non-exclusive Medical Subject Heading (MeSH) codes. The MeSH system categorizes articles by 14 high-level topics such as "Anatomy", "Psychiatry and Psychology", "Chemicals and Drugs", and so on. We used Pubmed to map each article to its major MeSH code(s). Even in common categories, there are substantial differences between the types of articles cited in-text and on the front page. An in-text citation is 33% more likely to be coded as about "Organisms" than a front page cite, and a front page cite is 25% more likely to be coded as about "Diseases". Again, these are not small sample issues: over 200,000 total citations are made to each of those categories.[22]

While Fig. 2 shows the distribution of articles across broad 1-digit MeSH codes, we can also look at specific subcategories. A particularly interesting subcategory in the MeSH scheme is "Investigative Techniques" (MeSH Tree Number E05), which includes topics like assays and epidemiological methods. A given in-text citation is 12.3% more likely (one-tailed t-test: $p < .00001$) to cite an "Investigative Technique" article than a front page citation, consistent with our discussion in Section 3. The right panel of Fig. 2 shows relative frequencies of in-text and front page citations by MeSH code.

A researcher investigating which journals are the primary vectors for the transmission of academic knowledge to inventors, or which types of biomedical knowledge is most used by inventors, will therefore obtain different answers depending on whether they proxy for knowledge transfer using in-text or front-page citations.

4.2. Do in-text citations track inventor knowledge with less noise?

In the previous subsection, we have shown that in-text and front-page citations proxy for *something* in a qualitatively different way. In Section 3.2, we showed that in-text citations are much more likely to have appeared on the first relevant document submitted to the patent office than front page citations, hinting that inventors may have more likely to know the cited knowledge at the time of invention.. We now examine, with two datasets, whether in-text citations are a better proxy measure for directly measured knowledge flows.

Let σ represent whether an inventor used the information in an academic article to motivate or construct their invention. Though this variable is in general non-observable, manual investigation or surveys of the knowledge used by inventors does exist for two small datasets: patent-paper pairs, and the Carnegie Mellon survey of R&D managers. We will refer to $\hat{\sigma}_{it}$ as a better proxy variable if it predicts the relevant σ with either less bias or less noise than $\hat{\sigma}_{fp}$, a distinction we return to at the end of the subsection.

Consider first patent-paper pairs, defined as patents with a simultaneous academic article describing the same invention. Do in-text or front page citations more closely match the knowledge flows cited in

[22] Since categories are non-exclusive, we cannot use a χ^2 test of equality of categorical distributions. However, each of the individual 14 categories has a different relative propensity to be cited in-text versus on the front page at $p < .00001$.

the article? Murray and Stern (2007) collects 171 articles in the journal Nature Biotechnology with a related patent filed at least partially on the basis of that article, as judged by a reader with subject-matter expertise. If in-text citations are a superior measure of the type of knowledge flows represented by academic citations, the research cited in the academic article should appear more frequently in the patent text than the front page.

Recall that our algorithmic method for identifying in-text citations requires starting with a fixed list of academic articles. Therefore, for this comparison we instead read each patent manually, counting the total number of in-text and front page references to academic articles in *any* journal, and the number of references in the original article's bibliography that are also cited in-text or on the front page.[23]

The biotechnology patents in this sample cite academic work much more heavily than the modal patent. The 171 patents have a mean of 26.9 (median: 15) front page references to academic articles, and a mean of 41.5 (median: 29) in-text references. The majority of references of each type are unrelated to the research cited in the corresponding Nature Biotechnology article; this is both because the patent text is generally very long and detailed compared to the academic writeup, and because the patent often covers an invention broader than the particular result in the underlying article.

On average, each Nature Biotechnology article in the Stern-Murray sample has about 30 referenced articles. Of these, an average of 6.9 articles are also mentioned in the text of the corresponding patent, while only 4.1 are cited on the patent's front page. That is, a citation in the underlying Nature Biotechnology article is 68% more likely to be found in the corresponding patent's specification text than in the front page citations. For the median patent-paper pair, the difference is even more stark. The front page of the median patent contains only 6.3% of the corresponding article's academic references, while the text contains 13.0% of these references, a 108% increase.

More than 25% of patents in the Murray-Stern sample have *zero* front page references which match a reference in the article bibliography, and 43.3% have no more than a single such reference. Patents with at most zero or one in-text citation matching the article bibliography are far less common, at 10.5% and 25.7% respectively. That is, a researcher who relied on front page citations rather than in-text citations to investigate knowledge flows would be almost 2.5 times more likely to incorrectly conclude that patent did not rely on *any* of the knowledge contained in the corresponding academic paper's references. The correlation of the total number of in-text and front page references matching the article bibliography for a given patent-paper pair is only 0.48, though this correlation overstates the overlap; even when there are, for instance, 3 in-text and 2 front page citations that match the article's bibliography, those 5 citations are often entirely distinct. Indeed, the in-text academic references and front page academic citations are identical in only 3 of the 171 patents.

Do in-text citations contain more of the corresponding article's academic citations simply because the inventor has lazily copy-and-pasted parts of the background from the article into the patent? In our experience manually reading both the article and the patent, it was very rare to find identical language. Only 5 of the 171 patents contained every academic reference from the corresponding article in the patent text, and in only 12.9% of the patents were even half of the underlying article's citations included in the patent text. A reader may ask why patents that make up a patent-paper pair do not cite *all* of the references in the original article. There are two reasons. First, the patent in general does not describe precisely the same invention and claims as the result described in the original article; rather, the original article often describes a single claim of the invention. Second, when manually examining these patent-paper pairs, similar basic science is often cited with different yet scientifically-equivalent references in the article and the patent.[24]

Our second test of the relative noisiness of in-text versus front page citations uses private sector R&D survey responses rather than academic citations as ground truth. In particular, are front page or in-text citations of academic research better correlates of firms' stated reliance on public sector research in a large survey? The Carnegie Mellon Survey (Cohen et al., 2002) of industrial R&D managers asked how much their firm relies on public sector spillovers for their inventions, as well as a series of questions about their reliance on "open science" like conferences, books and articles, versus "closed science" like contract work with academics. A follow-up study counted front page citations to public sector patents and non-patent literature in surveyed firm's patents (Roach and Cohen, 2013). The former doesn't correlate at all with the R&D manager's stated response on the percent of a firm's research using public sector knwledge, and the latter correlates relatively weakly.

To check whether in-text citations to academic research may better predict firm's actual stated use of public sector knowledge, we manually count all in-text and front page references to journal publications in all 6148 patents filed by 614 surveyed firms between 1991 and 1993. There are 8307 total front page citation of academic journal articles, and 9296 in-text cites. The raw correlation between the two count measures, at the individual patent level, is .52.

Fig. 3 plots the correlation between the number of in-text citations or front page citations to academic research and the R&D manager's estimate of whether less than 10%, 10–40%, 40–60%, 60–90%, or greater than 90% of their research projects rely on public sector knowledge. This plot is monotonically and strongly increasing for the in-text measure, and increasing though less precise for the front page measure. In Table 2, we show that, in line with Fig. 3, the in-text measure explains more of the survey response variance across a variety of specifications.

How much less noisy is the in-text measure? Consider a Bayesian exercise where an analyst is initially uncertain whether in-text or front page citation propensity for a given firm is a better predictor of the firm's stated reliance on public sector knowledge. The two potential "models" are not nested, so there is no traditional frequentist test of which model is more precise. However, under that uniform prior about which equally parsimonious model is a better predictor, the more precise model produces a greater relative likelihood of seeing the observed firm-level data.

In particular, the Bayesian model selection literature shows that the difference in BIC of non-nested models is related to that problem (Kass and Raftery, 1995; Raftery, 1998). A difference in the BIC of 6, by a standard rule of thumb (Raftery, 1995), is "strong" evidence for one model over another. In particular, when the difference is BIC is more than 6, one model is at least 20 times more likely to explain the data observed than another under a uniform prior about which model is better. Note the final row in Table 2: in-text citations are more strongly predictive for *every* model, whether without controls, or after controlling for covariates like industry and the number of scientists at each firm.[25]

Bayesian model selection cannot distinguish between the setting

[23] Note that extracting these references by hand means we are able to handle the non-trivial number of patents with typos on references, such as misspelled author names, misstated years, or obscure journal abbreviations. Reassuringly, the overwhelming majority of references we find by hand are ones that our algorithm would match if given the proposed academic article.

[24] The Murray-Stern patent-paper pair sample is a small, biomedical focused sample. Identifying pairs requires manually reading the academic article and proposed related patent, with expert knowledge about the field in question. We are unaware of any alternative or automated method to identify patent-paper pairs that does not use front page citation data to assist with the match.

[25] Appendix Table A.1 shows that in-text citations are also a better proxy for the "open science" factor in Roach and Cohen (2013), measuring the reliance of a firm's R&D on publicly available science.

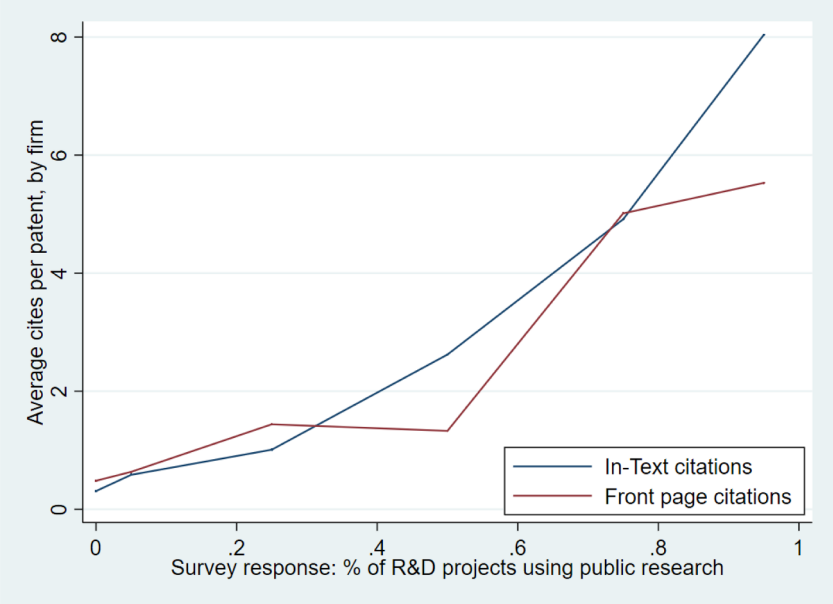


Fig. 3. Survey response is per firm to "what fraction of your unit's R&D projectors rely on public sector knowledge", on a five point scale: 0–10%, 10–40%, 40–60%, 6090%, 90–100%.

Table 2
Ordered logit models relating percent of firms' R&D projects using public research to science references.

In-Text Cites per Patent	0.0882^***		0.0619^***		0.0582^***	
	(0.0122)		(0.0171)		(0.0161)	
Front-Page Cites per Patent		0.0892^***		0.0470^*		0.0404
		(0.0140)		(0.0223)		(0.0228)
Total Firm Patents					0.133	0.143
					(0.0878)	(0.0887)
Fraction Scientists					1.589^***	1.555^***
					(0.326)	(0.334)
Observations	615	615	615	615	614	614
Pseudo R^2	0.020	0.015	0.046	0.043	0.053	0.050
BIC	1680.7	1687.8	1629.6	1635.7	1628.4	1634.4
Industry Controls	No	No	Yes	Yes	Yes	Yes

where $\hat{\sigma}_{it}$ is a less biased predictor than $\hat{\sigma}_{fp}$ from one where it is a less noisy predictor. Nonetheless, in-text citations are better correlated with two different measures of ground truth for "knowledge flows to inventors", and in a way that cannot be explained by small sample size of the empirical exercises.

4.3. When do in text citations perform poorly?

The previous exercises do not mean that front page citations are inferior proxies for all variables of interest to innovation scholars. Prior research suggests that front page scientific references can serve as a proxy for high-value patents, as measured by forward citations or other metrics (Fleming and Sorenson, 2004; Sorenson and Fleming, 2004).[26] While the interpretation of this result is unclear (for example, Sampat, 2010 suggests that firms have incentives to search for prior art more diligently for more important inventions), comparing how front page references and in-text references to science respectively correlate with patent value can help us better understand the information contained in each measure.

To do so, we collect data on all front page and in-text citations from 489,346 patents issued between 2006 and 2008 to articles published in the 248 journals described in Section 2. Of these patents, most (93%) cite no scientific articles from our set. Of patents with front page references to a scientific article, 57% also cite at least one scientific article in text. And of the patents with a full text reference, 69% cite at least one article on the front page.

We also collected data on four different measures of invention value: (1) forward citations in later patents; (2) the stock market reaction to patent issuance (Kogan et al., 2017); (3) whether the patents were renewed to at least year 8; and (4) whether the patents are part of triadic patent families. Prior research has used each of these measures as an indicator of patent value.

Table 3 shows summary statistics on each of the variables in this model. Tables 4 and 5, and Appendix Tables 2 and 3 show results from OLS regressions relating the value measures to the number of front page and in-text references (Models 1–3) and to indicators for whether there were any front page or in-text references (Models 4–6). We find front page backward citations to science are positively correlated with forward citations (Table 4) and with whether the patent is part of a triadic patent family (Table 5), consistent with prior research (Sorenson and Fleming, 2004; Fleming and Sorenson, 2004). To our knowledge the relationship between front page science references and stock market reaction to patent issue nor maintenance decisions has been examined before, and here the relationships are less robust across specifications

[26] Other research shows a weak or even negative relationship between extent of science citation in firms patents and forward citations (Trajtenberg, Henderson, and Jaffe (1992), Gittelman and Kogut (2003), Cassiman, Veugelers, and Zuniga (2008)). Patent-to-patent citations by inventors/attorneys tend to be focused on canonical inventions in an area and high-quality prior patents, while examiner-added citations focus on similarity (Moser, Ohmstedt, and Rhode (2018)).

Table 3
Summary statistics for value vs. science reference analyses.

Variables	(1) N	(2) mean	(3) sd	(4) min	(5) max
Number of front page science refs	489,346	0.332	2.533	0	195
Number of full text science refs	489,346	0.331	2.568	0	224
Number of overlapping science refs	489,346	0.178	1.772	0	170
Any full text science refs?	489,346	0.0513	0.221	0	1
Any front page science refs?	489,346	0.0596	0.237	0	1
Forward citations	489,346	9.577	24.36	0	2,120
Maintained at 8?	489,346	0.572	0.495	0	1
Stock Reaction	199,986	10.94	30.40	0.000237	1,457
Triadic Patent?	489,346	0.285	0.451	0	1

(Appendix Tables 2 and 3). That said, for all four measures of value, in most specifications, front page citations are more strongly related to value than in-text citations.[27]

The precise mechanisms for this are unclear. It may be that a patent's similarity or proximity to science, captured by front page science citations cited as prior art material to patentability, is more predictive of the private value of a patent to a firms than is whether the patent is based on science. This could be true, for example, if scientific inputs cited in text were in the public domain and available to competitors as well. Alternatively, it may be that applicants submit more front-page prior art or search more intensively for their more important inventions, to "bulletproof" these patents against validity challenges or guard against duty of candor violations, whereas for reasons discussed above this is not necessary to do for in-text citations. Whatever the reasons, this final validation emphasizes that front-page and full-text citations are fundamentally different, and each potentially useful for measuring

Table 4
OLS models relating forward citations to science references.

Variables	(1) Forward	(2) Forward	(3) Forward	(4) Forward	(5) Forward	(6) Forward
Number of front page science refs	0.3***		0.3***			
	(0.02)		(0.02)			
Number of full text science refs		0.1***	0.007			
		(0.01)	(0.01)			
Any front page science refs?				6.4***		6.1***
				(0.3)		(0.3)
Any full text science refs?					3.9***	0.7*
					(0.4)	(0.4)
issyear = 2007	−1.9***	−1.9***	−1.9***	−1.9***	−1.9***	−1.9***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
issyear = 2008	−3.9***	−3.9***	−3.9***	−3.9***	−3.9***	−3.9***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Constant	11***	11***	11***	11***	11***	11***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Observations	489,346	489,346	489,346	489,346	489,346	489,346
R-squared	0.091	0.090	0.091	0.092	0.090	0.092
Patent class FE	Yes	Yes	Yes	Yes	Yes	Yes
BIC	4467047.51	4467491.42	4467060.39	4466099.53	4467184.33	4466101.66

Table 5
OLS models relating whether a patent is in a triadic family to science references.

Variables	(1) Triad?	(2) Triad?	(3) Triad?	(4) Triad?	(5) Triad?	(6) Triad?
Number of front page science refs	0.006***		0.007***			
	(0.0003)		(0.0003)			
Number of full text science refs		−0.0002	−0.003***			
		(0.0003)	(0.0003)			
Any front page science refs?				0.1***		0.1***
				(0.003)		(0.004)
Any full text science refs?					0.04***	−0.01***
					(0.004)	(0.005)
issyear = 2007	−0.007***	−0.006***	−0.007***	−0.007***	−0.006***	−0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
issyear = 2008	−0.01***	−0.01***	−0.01***	−0.01***	−0.01***	−0.01***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.3***	0.3***	0.3***	0.3***	0.3***	0.3***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	489,346	489,346	489,346	489,346	489,346	489,346
R-squared	0.114	0.113	0.114	0.115	0.113	0.115
Patent class FE	Yes	Yes	Yes	Yes	Yes	Yes
BIC	550821.01	551418.98	550739.87	550319.26	551301.72	550318.84

[27] Further, a formal Bayesian model selection procedure as in the previous subsection strongly prefers models using front page citations to proxy for forward citations and triadic patents.

different concepts: whether a proxy $\hat{\sigma}$ is "good" depends, of course, on what it is meant to proxy for.

5. Concluding remarks

We have shown that in-text citations can be accurately and comprehensively extracted from patents, and that these citations have little overlap in a given patent with front page citations. There are important contexts where in-text citations are used in a qualitatively different way from front page citations. A public database covering 248 journals for over three decades is available alongside this paper. Given the frequency with which front page citations have been used to proxy for otherwise hard-to-observe aspects of the invention process, a more complete understanding of precisely how the two different types of citations are generated is needed.

In addition to having completely different legal uses, in-text citations possess two practical benefits compared to front page citations. First, non-granted patent applications do not have any front page citations listed.[28] For studies that require the use of contemporaneous data, it is often infeasible to wait five or more years for patents to be granted. In-text citations appear in applications, alloing that contemporaneous data to be examined. Second, pre-1947 U.S. patents do not have front page citations, while in-text citations can, in theory, be extracted for patents going back to the 1800s. For example, U.S. patent 2,295,481 A, applied for in 1939 by a scientist at Merck, contains no front page at all, but cites in the specification text just like modern patents: "Thus, Domagk (Deutsche Med. Wochsch., 61, 250, 1935) claimed that Prontosil, a derivative of diazotized sulphanilamide, was moderately effective against pneumococci, especially of Type III."

As for future research, the actual text of patents remains an incredibly underutilized resource. Rather than relying on count measures or features like a patent's class, machine learning methods (e.g., Mullainathan and Speiss, 2017, Gentzkow et al., 2017) can "read" the text of the patent and hence uncover information on precisely what knowledge a patent recombines, the exact way certain types of knowledge were used in the invention, and so on. In-text citations should prove value not just in better capturing actual knowledge flows, but in the ability to use the words around those citations to understand exactly how, when, and why inventors build on the past.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.respol.2020.103946.

Appendix A. : Additional results

Bayesian Model of Patent Citations Predicting Firms' Use of Open Science in Research

In-Text Cites per Patent	0.0338^***		0.0255^**		0.0216^***	
	(0.00583)		(0.00721)		(0.00535)	
Front-Page Cites per Patent		0.0304^***		0.0163^**		0.0117^*
		(0.00504)		(0.00518)		(0.00538)
Total Firm Patents					0.102^**	0.107^**
					(0.0298)	(0.0294)
Fraction Scientists					0.625^**	0.643^**
					(0.188)	(0.198)
Constant	−0.0295	−0.0287	−0.0454^***	−0.0433^***	−0.167^***	−0.170^***
	(0.0505)	(0.0530)	(0.00245)	(0.00211)	(0.0245)	(0.0256)
Observations	615	615	615	615	614	614
R^2	0.037	0.025	0.122	0.113	0.145	0.138
BIC	1534.0	1541.6	1470.4	1476.8	1465.6	1470.9
Industry Controls	No	No	Yes	Yes	Yes	Yes

OLS models relating stock market reaction to science references

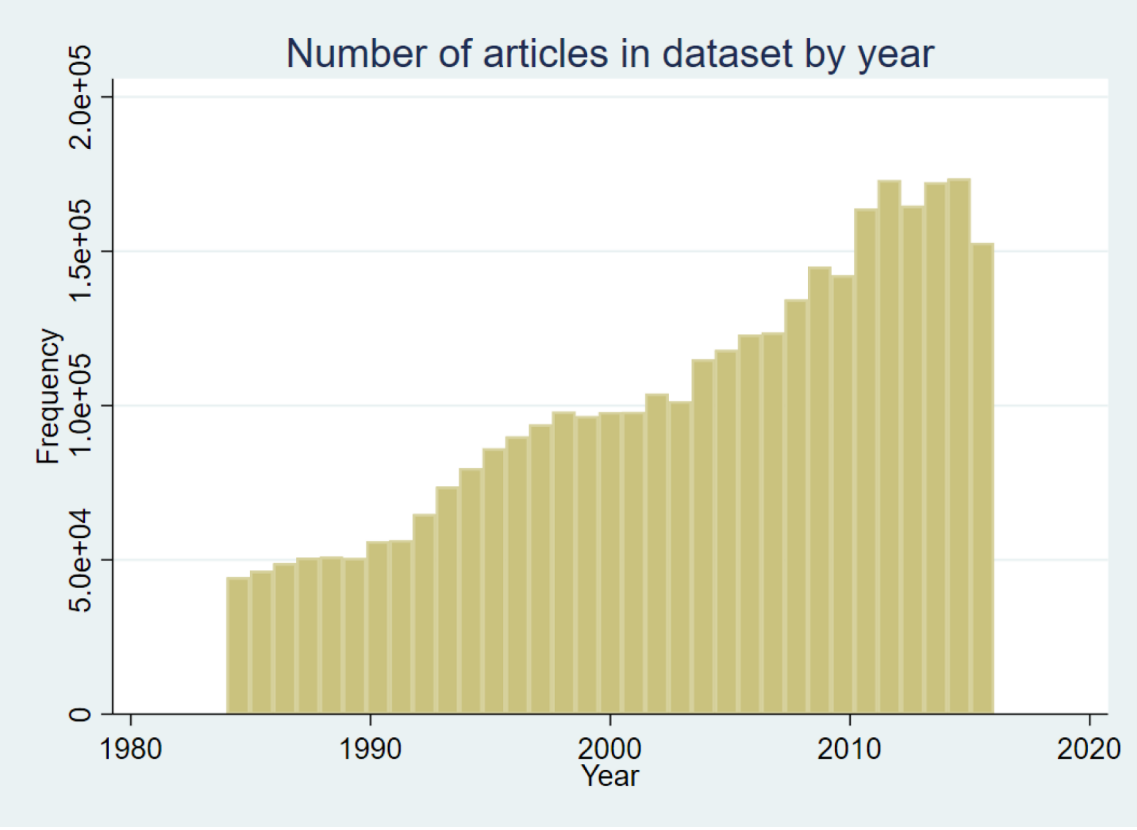
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Stock	Stock	Stock	Stock	Stock	Stock
Number of front page science refs	0.10**		0.10**			
	(0.04)		(0.04)			
Number of full text science refs		0.02	-0.004			
		(0.04)	(0.04)			
Any front page science refs?				0.3		-0.7
				(0.5)		(0.5)
Any full text science refs?					2.8***	3.2***
					(0.7)	(0.8)
issyear = 2007	1.5***	1.5***	1.5***	1.5***	1.5***	1.5***
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
issyear = 2008	6.7***	6.7***	6.7***	6.7***	6.7***	6.7***
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
Constant	8.4***	8.4***	8.4***	8.4***	8.3***	8.3***
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)

[28] The information disclosure statements with applicant prior art citations can be filed throughout the application process, and examiner searches are conducted after the application is filed. More practically, neither of these is readily available in machine readable form.

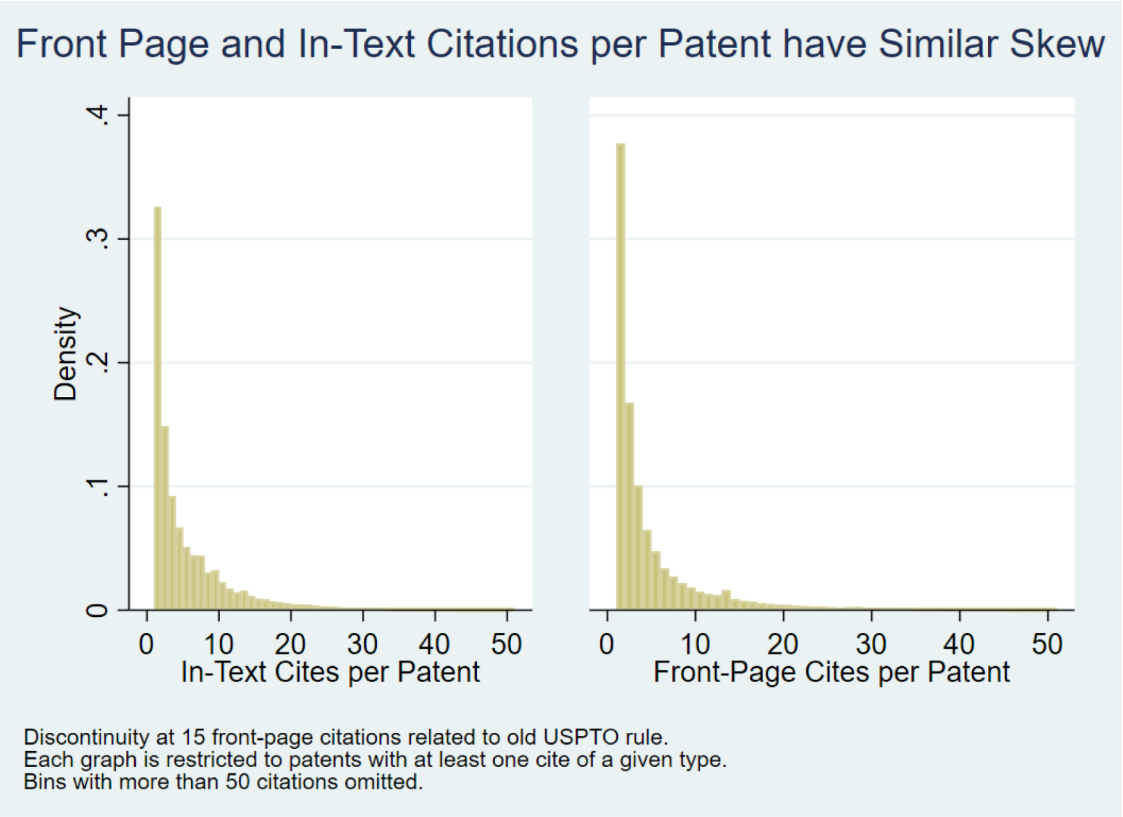
Observations	199,986	199,986	199,986	199,986	199,986	199,986
R-squared	0.126	0.126	0.126	0.126	0.126	0.126
Patent class FE	Yes	Yes	Yes	Yes	Yes	Yes
BIC	1906330.95	1906339.77	1906343.14	1906339.67	1906298.57	1906307.13

OLS models relating whether maintained to year 8 to science references

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	Maintained	Maintained	Maintained	Maintained	Maintained	Maintained
Number of front page science refs	0.0006*		0.001***			
	(0.0003)		(0.0003)			
Number of full text science refs		−0.002***	−0.002***			
		(0.0003)	(0.0003)			
Any front page science refs?				−0.04***		−0.03***
				(0.004)		(0.004)
Any full text science refs?					−0.06***	−0.04***
					(0.004)	(0.005)
issyear = 2007	0.0003	0.0003	0.0003	0.0005	0.0004	0.0004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
issyear = 2008	−0.007***	−0.007***	−0.007***	−0.007***	−0.007***	−0.007***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Constant	0.6***	0.6***	0.6***	0.6***	0.6***	0.6***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	489,346	489,346	489,346	489,346	489,346	489,346
R-squared	0.076	0.076	0.076	0.076	0.076	0.076
Patent class FE	Yes	Yes	Yes	Yes	Yes	Yes
BIC	661556.99	661525.07	661520.23	661408.7	661365.09	661328.82



Articles in Web of Science dataset across 248 journals by year.



The distribution of front page and in-text citations per patent are both highly skewed.

Appendix B. Data construction

We begin with a list of academic articles and patents. The patent data comes from the publicly available USPTO Patent Application and Grant Publication Full Text files. These files include the full text of the patent applications and grants, as well as bibliographic information, including the application and publication dates, and the inventor names and locations.

For articles, we begin with a list of the 20 most-cited journals in 16 fields, gathered from Google Scholar. These fields are Chemicals and Materials, Nanotechnology, Biochemistry, Oil Petroleum and Natural Gas, Engineering and Computer Science, Artificial Intelligence, Robotics, Mechanical Engineering, Operations Research, Structural Engineering, Sustainability Technology, Health and Medical Science, Biomedical Technology, Oncology, Physics and Mathematics, and Probability and Statistics. 26 of the journals on these lists are duplicates, and 30 are not catalogued in Web of Science (largely ArXiv working series and conference proceedings), leaving 264 journals. Sixteen further journals had very few articles on Web of Science. Removing these leaves 248 unique journals, listed in Appendix C.[29]

From this set, we gather metadata for list of articles from a predefined set of journals from Thomson Reuters' Web of Science, for all articles published between 1984 and 2016. Second, we standardize special characters and drop ambiguous words from titles (e.g., the chemical formula for Carbon-14, which can appear as ^{14}C, C-14, *C*14, and so on).[30] Third, we algorithmically construct a list of potential journal abbreviations using the full journal title, the standard Web of Science abbreviation, and alternative abbreviations for common words (algorithm available on request). For instance, the WOS abbreviation for the American Economic Review is Am Econ Rev, but patents also contain many references to Amer Econ Rev, Amer Econ Review, and AER. Fourth, we take all paragraphs in the patent text which contain a four-digit number from 1900 to 2017, and search indexed versions of those paragraphs for a combination of metadata in any article in our sample as described below.

The most natural way to match is with a coarsened matching procedure. For instance, in Marx and Fuegi (2019) or Ahmadpoor and Jones (2017), front page patent citations are matched to scientific articles in this manner. The important difference is that front page citations appear in one known line of text. Therefore, an algorithm can operate on exactly that one line. In our case, citations appear in arbitrary formats throughout the full specification text of patents, without a bibliography. It is often ex-ante non-obvious where these citations might appear. We therefore need an algorithm that limits the space of text to be searched, does not require the analyst to know ex-ante where a citation might appear, and nonetheless makes few errors.

Patents are provided as xml files. A single line in this xml file may contain an entire paragraph in the patent application text, hence a line may contain thousands of characters. Investigating a subset of the files, we have identified that there are very few lines longer than 7000 characters in length; therefore, we have kept only the first 7000 characters of each line in the xml file. Patents from earlier years are provided as text files rather than XML, with a single paragraph divided into multiple lines; in this case we appended up to 10 consecutive lines to each other and treated them as a single paragraph. In the minority of cases where a paragraph was divided into more than 10 lines, we split the paragraph into subparagraphs each made of 10 lines.

[29] These are generally new journals, or conference proceedings only partially cataloged by the Web of Science.

[30] Indeed, the exact format of titles of the same article in PubMed, Web of Science, and Google Scholar is often completely different, largely due to how special characters are handled.

Each line in the xml file starts with an xml tag identifying the information in that line. Through investigation of a subset of the files, we have found that references are nearly always included in lines with certain tags such as li and p. Therefore, we dropped the remainder of the files, and kept only lines with these tags. The remaining portions of the files also contain a minimal amount of citations, but investigation by hand suggests that these are mostly repetitions of citations also made elsewhere within the same document.

Using the article list and the queryable patent text dataset, we now run the following algorithm:

1) Eliminate any lines not containing the four digits of at least one year from 1986 to 2017. The underlying assumption is that when a journal article is cited, then its year must exist in the citing paragraph.
2) From these remaining lines, to speed up the process of journal name lookup in this still large search space, we create an index of the words contained in each line in each file. In other words, we create word - line number - file number tuples, representing every instance of where each word appears in which line of which file. This allows us to query the index rather than perform operations on the direct xml and text of patents, which are collectively over a terabyte of data.[31]
3) Identify lines that contain any of the journal names in our list of 248 journals. For a given journal name, we take the words from the journal name that have two or more characters, identify the word-line number-file number tuple set for each word, and take the intersection of those sets. For example, for New England Journal of Medicine, we intersect the tuple sets for the four words New, England, Journal, and Medicine.

For the purposes of this search, the journal list is augmented by various common abbreviations of the same journal name, using the algorithm mentioned previously. For example, to capture New England Journal of Medicine, many different abbreviations were searched for, including the following: “NEJM”, “N.E.J.M", “N. Engl. J. Med", and “New England J. Medicine". We now move from the journal level to the article level search.

1) Consider the lines that contain a given journal-year pair, recalling that “line” in the patent files generally refers to paragraph of text. We search for the article first author last name close to the journal name. Article title lengths vary across articles, and the full title may be present in the citation. Therefore, we identify proximity to the journal name as being within 150 characters plus the length of the article title before and after the journal name location. We eliminate lines that do not contain the first author's last name or the year within this proximity. In this step, we are only identifying the first citation to a single journal within a single line. In other words, if two different articles from the same journal are cited within a single line of the patent application, then we may or may not capture the second one depending on how far apart it is located from the first citation. We have no reason to believe that missing such citations would bias our results.
2) Among the matches identified so far, we keep matches that include either the article page numbers *or* the first four words of the article in the proximity (i.e. within 150 + article title length characters of the journal title). Although the described algorithm can be applied to majority of the articles in our set, we have applied some modifications due to foreign characters, punctuations, or data quality issues.
3) Special Characters in the title: If the article title contains a special character such as a punctuation, then we modify step 4. Instead of using the first four words of the title, we use the first four words without a punctuation *and* require at least five out of the first six words to be present individually (i.e. instead of searching six consecutive words, we search for each word separately, and check if at least five out of six exists within the proximity). Note that if article page numbers match, then there is no need to resort to this modification as the page number match will capture the citation in the original algorithm.[32]
4) Authors with multi-word last names: Some authors have last names consisting of more than a single word. In these cases, we run the algorithm using only the last word in the author's last name.
5) Data Reporting Issue 1: For some articles from the WOS data, the article author appears to be a study group or an institution. In these observations, unlike the rest of the data, the author name is not a quoted string. For these observations we are not able to use the author last name to conduct the match. So, instead of requiring the author last name match in step 5, we change the condition in step 4 from an OR to an AND condition: both title first four words AND the article page no have to exist within the patent line to be considered a match.
6) Data Reporting Issue 2: Occasionally, for corrections and other errata, the WOS data includes the original article's “pageno” at the end of the title field. When both the original article and errata page numbers are included in the patent, we count this as a citation to both articles. Any algorithm of this type needs to balance between Type I and Type II errors. In this context, a Type I error is erroneously claiming the existence of a citation. Investigation by hand suggests that the matches identified by the algorithm contain less than one percent Type I errors. A Type II error happens if the algorithm fails to identify an existing citation. For instance, “In 1989 Stephan J. Weiss in the New England Journal of Medicine conducted bacterial sensitivity studies on E. Coli and toxicity on tissue in guinea-pigs” in patent application 12/101,775 is too vague, lacking both an article title and a journal issue number, for our algorithm to match it with a specific article. The extent of Type II errors of this kind is difficult to quantify. We investigated a number of less restrictive algorithms, but generally they resulted in many more Type I errors with very few additional legitimate matches. We discuss this further, with examples, in the following subsection.

Finally, when searching for front page citations, we use the exact same algorithm as above. It goes without saying that a coarsened matching approach would generate fewer false negatives on front page citations; however, for comparability of in-text and front-page citations, it makes sense to use the most similar possible matching algorithm. Note that in our first empirical exercise in the main text, we count both types of citations by hand, and find a similar pattern of non-overlap.

It is difficult therefore to directly compare our algorithm to other matches of front page citations. However, in our data, 274,822 patents have at least one front page cite. In Ahmadpoor and Jones (2017), using a coarsened match of all 32 million post-1945 articles in Web of Science to all post-

[31] We also considered dropping all paragraphs which do not include [, {, <li, “et al”, “et. al.” or <i. Doing so modestly reduces the size of the file to be queried, and captures 99.9% of the citations we otherwise recover in our primary algorithm. The speed improvement was minor, but researchers considering expanding our algorithm to a larger journal list may wish to consider the speed-up.

[32] Umlauts and special characters in Author name: Some author names include foreign characters that may be spelled in more than one way in the English alphabet. Therefore, we attempted the following three changes in first author last names: ae into a, oe into o, and ue into u. We then repeated the algorithm with the updated author last names. The rate of false positives this induces is high, but could conceivably be manually handles: e.g., changing the name Xu to Xue, the rerunning algorithm, picks up a false positive. This can perhaps be fixed in future versions with a dictionary of common names including special characters.

1976 articles, they find roughly 759,000 total patents with at least one front-page cite. That is, though we only use post 1976 pubs and 248 journals, we nonetheless are capturing 36% of the front page cites in Ahmadpoor and Jones (2017). Our match also finds 267,576 articles with at least one citation, versus 1.41 million in the longer and much larger dataset matched by Ahmadpoor and Jones (2017). This suggests that the journals we focus on include many of those which are most likely to be cited by an inventor.

Some notes on matches our algorithm misses and catches

All matching algorithms balance between type I and type II references. Below are a sample of in-text references which our current algorithm is unable to find.

1) Wrong author name: US6,190,856 has Erkki Koivunen cited as ”Kolvunen” in-text (“Kolvunen, E. et al., J. Cell Biol. 124:373 (1994)”); US6,423,693 has a paper by Dominic Wells as “Wells” correctly in front page but as “Walls” in the specification text (“Walls, 1993, FEBS Lett. 332:170-182”)
2) Wrong year: US6,130,090 has a Bradley and Liu paper as 1997 not 1996, when 1996 is correct; US6,143,551 has a Hauf paper with the year ”in press” but of course our algorithm needs the actual year
3) Too vague: US6,008,016 just cites “Li et al” with no context or reference in other part of the patent text (“THPs were purified from the resulting filtrate using chromatographic procedures that are standard for the isolation of fish AFPs from serum (Fourney et al., 1984; Li et al., 1985; Ng et al., 1986)”); US6,483,012 cites “Genomic DNA was obtained from leaf tissue according to Doyle and Doyle (1987)” with no further detail.
4) Outright typos: US6,265,535 has year of one in-text citation as 19994 (“cells were grown and differentiated into adipocytes as described previously (Garcia de Herreros et al., 1989, J. Biol. Chem. 264:19994)”).
5) Wrong journal: US6,309,824 has a 1989 paper listed as being published in ”Genetics” when the actual journal is called Genomics; ironically, this particular article was written by the inventor himself!

In our experience, typos are more frequent in-text than in prior art, partially justifying the assumption that lawyers are not too careful about checking in-text cites (e.g., a citation to a paper by Paradkar in US5,914,367 is misspelt in-text but not in the front page NPL).

On the other hand, our algorithm does correctly match a number of challenging in-text citations. For example, in US6,605,754 (“Comai et al have previously described a chimeric plant promoter combining elements of the CaMV35S and the mannopine synthase (mas) promoters (1990, Plant Mol Biol, 15:373-381)”, the reference is found despite the author name Comai being very far from the reference, and no title being included.

Categorizing University, Medical and other patentees

We refer to an article as “medical” if it appears in a journal categorized as “Oncology”, “Health and Medical Science”, “Biomedical Technology” or “Biochemistry” in the Google Scholar journal rankings.

We refer to a patent assignee as “university-based” if the assignee contains any of a series of references to universities or university-affiliated teaching hospitals. This list, also used in Bryan and Ozcan (2019), is based on manual examination of patents in a large number of languages.

“University” was a designation given to patents with any of the following in one of their patent assignee strings: “university”, “alumni”, “ univ”, “national cancer”, “brigham”, “jackson lab”, “research center”, “akademie”, “vib ”, “RIKEN”, “Eye & Ear”, “medical school”, “national jewish health”, “eth zurich”, “Center for”, “univeristy”, “higher education”, “cold spring harbor”, “akadamie”, “centre for”, “fundacio”, “Université”, “centre”, “planck”, “universuty”, “Universitât”, “fundacion”, “UNIVERSITÀ”, “agence nationale”, “insitute”, “UNIVERSITÉ”, “eye and ear infirmary”, “Society for”, “Unversity”, “cancer centre”, “universite”, “institue”, “istituto”, “cancer center”, “fondation”, “universiteit”, “universitet”, “universitaet”, “city of hope”, “educational fund”, “zentrum”, “consejo”, “ecole”, “universtiy”, “centro”, “kettering”, “mayo”, “schule”, “institucio”, “centrum”, “hospital for sick”, “children's hospital”, “academisch”, “universita”, “universit´at”, “unviersity”, “georgia tech”, “school of”, “consiglio nazionale”, “intellectual properties”, “fondazione”, “national centre”, “centro nacional”, “centre national”, “foundation”, “regents”, “council”, “fred hutchinson”, “general hospital corporation”, “universidade”, “research hospital”, “medical center”, “foundation”, “universitat”, “universidad”, “colegio”, “univerisite”, “institut”, “institute”, “instituto”, “trustees”, “academia”, “academy”, or “college”. These strings were picked following manual investigation in order to limit type I and type II errors, and attempt to capture academic research hospitals as well as universities themselves.

Appendix C. List of Covered Journals

ACCOUNTS OF CHEMICAL RESEARCH
ACI STRUCTURAL JOURNAL
ACS APPLIED MATERIALS & INTERFACES
ACS NANO
ACTA BIOMATERIALIA
ACTA MECHANICA
ADVANCED ENERGY MATERIALS
ADVANCED FUNCTIONAL MATERIALS
ADVANCED MATERIALS
ANGEWANDTE CHEMIE-INTERNATIONAL EDITION
ANNALS OF APPLIED PROBABILITY
ANNALS OF APPLIED STATISTICS
ANNALS OF BIOMEDICAL ENGINEERING
ANNALS OF ONCOLOGY
ANNALS OF OPERATIONS RESEARCH
ANNALS OF PROBABILITY
ANNALS OF STATISTICS

ANNALS OF SURGICAL ONCOLOGY
ANNUAL REVIEW OF BIOCHEMISTRY
ANTIOXIDANTS & REDOX SIGNALING
APPLIED ENERGY
APPLIED PHYSICS LETTERS
APPLIED SOFT COMPUTING
ARTIFICIAL INTELLIGENCE
ASTRONOMY & ASTROPHYSICS
ASTROPHYSICAL JOURNAL
AUTONOMOUS ROBOTS
BERNOULLI
BIOCHEMICAL JOURNAL
BIOCHIMICA ET BIOPHYSICA ACTA-BIOENERGETICS
BIOCHIMICA ET BIOPHYSICA ACTA-MOLECULAR CELL RESEARCH
BIOENERGY RESEARCH
BIOFABRICATION
BIOFUELS BIOPRODUCTS & BIOREFINING-BIOFPR
BIOINSPIRATION & BIOMIMETICS
BIOMASS & BIOENERGY
BIOMATERIALS
BIOMECHANICS AND MODELING IN MECHANOBIOLOGY
BIOMEDICAL MATERIALS
BIOMEDICAL MICRODEVICES
BIOMETRIKA
BIORESOURCE TECHNOLOGY
BLOOD
BRITISH JOURNAL OF CANCER
BRITISH MEDICAL JOURNAL
CANCER
CANCER CELL
CANCER DISCOVERY
CANCER LETTERS
CANCER RESEARCH
CELL
CHEMICAL COMMUNICATIONS
CHEMICAL REVIEWS
CHEMICAL SOCIETY REVIEWS
CHEMISTRY OF MATERIALS
CIRCULATION
CLINICAL CANCER RESEARCH
COCHRANE DATABASE OF SYSTEMATIC REVIEWS
COMPOSITE STRUCTURES
COMPUTATIONAL MECHANICS
COMPUTATIONAL STATISTICS & DATA ANALYSIS
COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING
COMPUTERS & INDUSTRIAL ENGINEERING
COMPUTERS & OPERATIONS RESEARCH
COMPUTERS & STRUCTURES
CURRENT OPINION IN STRUCTURAL BIOLOGY
EARTHQUAKE ENGINEERING & STRUCTURAL DYNAMICS
EMBO JOURNAL
ENERGIES
ENERGY & ENVIRONMENTAL SCIENCE
ENERGY
ENERGY AND BUILDINGS
ENERGY CONVERSION AND MANAGEMENT
ENERGY FOR SUSTAINABLE DEVELOPMENT
ENGINEERING ANALYSIS WITH BOUNDARY ELEMENTS
ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE
ENGINEERING FAILURE ANALYSIS
ENGINEERING STRUCTURES
EUROPEAN CELLS & MATERIALS
EUROPEAN JOURNAL OF CANCER
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
EXPERT SYSTEMS WITH APPLICATIONS
FEBS JOURNAL

FEBS LETTERS
FINITE ELEMENTS IN ANALYSIS AND DESIGN
FREE RADICAL BIOLOGY AND MEDICINE
GASTROENTEROLOGY
GLOBAL CHANGE BIOLOGY BIOENERGY
IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)
IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION
IEEE ROBOTICS & AUTOMATION MAGAZINE
IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS
IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING
IEEE TRANSACTIONS ON CYBERNETICS
IEEE TRANSACTIONS ON FUZZY SYSTEMS
IEEE TRANSACTIONS ON HAPTICS
IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS
IEEE TRANSACTIONS ON NEURAL NETWORKS
IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE
IEEE TRANSACTIONS ON POWER ELECTRONICS
IEEE TRANSACTIONS ON ROBOTICS
IEEE TRANSACTIONS ON SUSTAINABLE ENERGY
IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART B-CYBERNETICS
IEEE-RAS INTERNATIONAL CONFERENCE ON HUMANOID ROBOTS
IMMUNITY
INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
INTERNATIONAL JOURNAL OF BIOCHEMISTRY & CELL BIOLOGY
INTERNATIONAL JOURNAL OF CANCER
INTERNATIONAL JOURNAL OF ENERGY RESEARCH
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE
INTERNATIONAL JOURNAL OF HYDROGEN ENERGY
INTERNATIONAL JOURNAL OF MECHANICAL SCIENCES
INTERNATIONAL JOURNAL OF NON-LINEAR MECHANICS
INTERNATIONAL JOURNAL OF OPERATIONS & PRODUCTION MANAGEMENT
INTERNATIONAL JOURNAL OF PRODUCTION ECONOMICS
INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
INTERNATIONAL JOURNAL OF RADIATION ONCOLOGY BIOLOGY PHYSICS
INTERNATIONAL JOURNAL OF ROBOTICS RESEARCH
INTERNATIONAL JOURNAL OF SOCIAL ROBOTICS
INTERNATIONAL JOURNAL OF SOLIDS AND STRUCTURES
JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION
JOURNAL OF APPLIED MECHANICS-TRANSACTIONS OF THE ASME
JOURNAL OF BIOLOGICAL CHEMISTRY
JOURNAL OF BIOMEDICAL MATERIALS RESEARCH PART A
JOURNAL OF BIOMEDICAL MATERIALS RESEARCH PART B-APPLIED BIOMATERIALS
JOURNAL OF BIOMEDICAL NANOTECHNOLOGY
JOURNAL OF BUSINESS & ECONOMIC STATISTICS
JOURNAL OF CANADIAN PETROLEUM TECHNOLOGY
JOURNAL OF CLINICAL INVESTIGATION
JOURNAL OF CLINICAL ONCOLOGY
JOURNAL OF COMPOSITES FOR CONSTRUCTION
JOURNAL OF CONSTRUCTIONAL STEEL RESEARCH
JOURNAL OF ECONOMETRICS
JOURNAL OF ENGINEERING FOR GAS TURBINES AND POWER-TRANSACTIONS OF THE ASME
JOURNAL OF FIELD ROBOTICS
JOURNAL OF HIGH ENERGY PHYSICS
JOURNAL OF INTELLIGENT & ROBOTIC SYSTEMS
JOURNAL OF MACHINE LEARNING RESEARCH
JOURNAL OF MATERIALS CHEMISTRY
JOURNAL OF MATERIALS CHEMISTRY B
JOURNAL OF MECHANICAL DESIGN
JOURNAL OF NANOMATERIALS
JOURNAL OF NANOPARTICLE RESEARCH
JOURNAL OF NANOSCIENCE AND NANOTECHNOLOGY
JOURNAL OF NATURAL GAS CHEMISTRY
JOURNAL OF NEURAL ENGINEERING
JOURNAL OF NUCLEAR MATERIALS
JOURNAL OF OPERATIONS MANAGEMENT

JOURNAL OF PETROLEUM GEOLOGY
JOURNAL OF PETROLEUM SCIENCE AND ENGINEERING
JOURNAL OF PHYSICAL CHEMISTRY C
JOURNAL OF POWER SOURCES
JOURNAL OF STATISTICAL SOFTWARE
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
JOURNAL OF THE MECHANICAL BEHAVIOR OF BIOMEDICAL MATERIALS
JOURNAL OF THE MECHANICS AND PHYSICS OF SOLIDS
JOURNAL OF THE NATIONAL CANCER INSTITUTE
JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY
JOURNAL OF THORACIC ONCOLOGY
JOURNAL OF TISSUE ENGINEERING AND REGENERATIVE MEDICINE
JOURNAL OF TURBOMACHINERY-TRANSACTIONS OF THE ASME
JOURNAL OF VIBRATION AND ACOUSTICS-TRANSACTIONS OF THE ASME
KNOWLEDGE-BASED SYSTEMS
LANCET
LANCET ONCOLOGY
LEUKEMIA
MARINE AND PETROLEUM GEOLOGY
MATHEMATICAL FINANCE
MATHEMATICAL PROGRAMMING
MATHEMATICS OF OPERATIONS RESEARCH
MECCANICA
MECHANISM AND MACHINE THEORY
MECHATRONICS
MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING
MEDICAL ENGINEERING & PHYSICS
MOLECULAR AND CELLULAR BIOLOGY
MOLECULAR BIOLOGY OF THE CELL
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY
NANO ENERGY
NANO LETTERS
NANO RESEARCH
NANO TODAY
NANOMEDICINE
NANOSCALE
NANOSCALE RESEARCH LETTERS
NANOTECHNOLOGY
NANOTOXICOLOGY
NATURE CHEMICAL BIOLOGY
NATURE CHEMISTRY
NATURE GENETICS
NATURE MATERIALS
NATURE MEDICINE
NATURE NANOTECHNOLOGY
NATURE PHOTONICS
NATURE PHYSICS
NATURE REVIEWS CANCER
NATURE REVIEWS CLINICAL ONCOLOGY
NATURE STRUCTURAL & MOLECULAR BIOLOGY
NEURAL NETWORKS
NEUROCOMPUTING
NEURON
NEW ENGLAND JOURNAL OF MEDICINE
NPG ASIA MATERIALS
NUCLEIC ACIDS RESEARCH
OIL & GAS JOURNAL
OIL & GAS SCIENCE AND TECHNOLOGY-REVUE D IFP ENERGIES NOUVELLES
ONCOGENE
OPERATIONS RESEARCH
PETROLEUM EXPLORATION AND DEVELOPMENT
PETROLEUM GEOSCIENCE
PETROLEUM SCIENCE AND TECHNOLOGY

PHYSICAL REVIEW B
PHYSICAL REVIEW D
PHYSICAL REVIEW LETTERS
PHYSICS LETTERS B
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA
PRODUCTION AND OPERATIONS MANAGEMENT
PROGRESS IN PHOTOVOLTAICS
RENEWABLE & SUSTAINABLE ENERGY REVIEWS
RENEWABLE ENERGY
ROBOTICS AND AUTONOMOUS SYSTEMS
ROBOTICS AND COMPUTER-INTEGRATED MANUFACTURING
SMALL
SOLAR ENERGY
SPE DRILLING & COMPLETION
SPE JOURNAL
SPE PRODUCTION & OPERATIONS
SPE RESERVOIR EVALUATION & ENGINEERING
STATISTICS AND COMPUTING
STATISTICS IN MEDICINE
STRUCTURAL AND MULTIDISCIPLINARY OPTIMIZATION
STRUCTURAL CONTROL & HEALTH MONITORING
STRUCTURAL SAFETY
THIN-WALLED STRUCTURES
TISSUE ENGINEERING PART A
TISSUE ENGINEERING PART B-REVIEWS
TISSUE ENGINEERING PART C-METHODS
TRANSPORTATION SCIENCE
TRENDS IN BIOCHEMICAL SCIENCES
TRIBOLOGY INTERNATIONAL
TRIBOLOGY LETTERS
VEHICLE SYSTEM DYNAMICS
WEAR
WIND ENERGY

References

Agrawal, A., Henderson, R., 2002. Putting patents in context: exploring knowledge transfer from Mit. Manag. Sci. 48 (1), 44–60.

Ahmadpoor, M., Jones, B.F., 2017. The dual frontier: patented inventions and prior scientific advance. Science.

Alcácer, J., Gittelman, M., Sampat, B., 2009. Applicant and examiner citations in U.S. patents: an overview and analysis. Res. Policy 38 (2), 415–427. https://doi.org/10.1016/j.respol.2008.12.001.

Azoulay, P., Zivin, J.S.G., Li, D., Sampat, B.N., 2015. Public R&D investments and private-sector patenting: evidence from NIH funding rules. Rev. Econ. Stud.

Bryan, KA., and Ozcan, Y. 2019. "The impact of open access mandates on invention".

Cassiman, B., Veugelers, R., Zuniga, P., 2008. In search of performance effects of (in) direct industry science links. Ind. Corp. Change 17 (4), 611–646.

Cohen, W.M., Nelson, R.R., Walsh, J.P., 2000. Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). 7552. pp. 50. https://doi.org/10.1093/dnares/dsr014. NBER Working Paper.

Cohen, W., Nelson, R., Walsh, J., 2002. Links and impacts: the influence of public research on industrial R&d. Manag. Sci.

Cotropia, CA., Lemley, MA., Sampat, B., 2013. Do applicant patent citations matter? Res. Policy 42 (4), 844–854. https://doi.org/10.1016/j.respol.2013.01.003.

Devlin, A., 2009. The misunderstood function of disclosure in patent law. Harv. JL Tech. 23, 401.

EPO. 2007. "Cited references in european patent documents." http://www.wipo.int/export/sites/www/cws/en/taskforce/citation{_}practices/docs/epo{_}citation{_}practice{_}summary.pdf.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strateg. Manag. J. 25 (8–9), 909–928.

Frakes, MD., Wasserman, MF., 2017. Is the Time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. Review of Economics and Statistics 99 (3), 550–563.

Fromer, JC., 2008. Patent disclosure. Iowa Law Rev. 94, 539.

Gentzkow, M., Kelly, B., Taddy, M., 2017. Text as Data. NBER Working Paper.

Gittelman, M., Kogut, B., 2003. Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. Manag. Sci. 49 (4), 366–382.

Henderson, R., Jaffe, A.B., Trajtenberg, M., 1998. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. Rev. Econ. Stat. 80 (1), 119–127.

Hippel, E., 1988. The Sources of Innovation. Oxford University Press.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.

Jaffe, AB., de Rassenfosse, G., 2017. Patent citation data in social science research: overview and best practices. J. Assoc. Inf. Sci. Technol.

Jaffe, A., Trajtenberg, M., Handerson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations Author (s): Adam B . Jaffe, Manuel Trajtenberg and Rebecca Henderson. Q. J. Econ. 108 (3), 577–598. https://doi.org/10.2307/2118401.

Jefferson, O.A., Jaffe, A., Ashton, D., Warren, B., Koellhofer, D., Dulleck, U., Ballagh, A., et al., 2018. Mapping the global influence of published research on industry and innovation. Nat. Biotechnol. 36 (1), 31.

Kaplan, S., Vakili, K., 2015. The double-edged sword of recombination in breakthrough innovation. Strateg. Manag. J. 36 (10), 1435–1457. https://doi.org/10.1002/smj.2294.

Kass, RE., Raftery, AE., 1995. Bayes factors. J. Am. Stat. Assoc.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2017. Measuring Technological Innovation over the Long Run. pp. 1–57 Working Paper.

Khan, B.Z., Branstetter, L., Chien, C., Diebolt, C., Dreyfuss, R., Lamoreaux, N., Moser, P., et al., 2014. Inventing in the Shadow of the Patent System: Evidence from 19th-centrury Patents and Prizes for Technological Innovations. https://doi.org/10.3386/w20731. NBER Working Paper #20731.

Kogan, L., Papanikolaou, D., Seru, A., Stoffman, N., 2017. Technological innovation, resource allocation, and growth. Q. J. Econ. 132 (2), 665–712.

Krasker, WS., Pratt, JW., 1986. Bounding the effects of proxy variables on regression coefficients. Econometrica 54 (3).

Krugman, PR., 1991. Geography and Trade. MIT Press, Cambridge, MA.

Kuhn, J., and Thompson, N.. 2017. "The ways we've been measuring patent scope are wrong: how to measure and draw causal inferences with patent scope".

Kuhn, J., Younge, K., Marco, A., 2019. Patent citations reexamined. Forthcoming. RAND J. Econ.

Lampe, R., 2012. Strategic citation. Rev. Econ. Stat.

Lemley, M.A., Sampat, B., 2012. Examiner characteristics and patent office outcomes. Rev. Econ. Stat. 94 (3), 817–827.

Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., 1987. Appropriating the returns from industrial research and development; comments and discussion. Brook. Pap. Econ. Act. 3, 783. https://doi.org/10.2307/2534454.

Li, D., Azoulay, P., Sampat, B.N., 2017. The applied value of public investments in

biomedical research. Science 356 (6333), 78–81.
Marx, M., Fuegi, A., 2019. Reliance on Science in Patenting. Working Paper.
Meyer, M., 2000. What is special about patent citations? Differences between patent and scientific citations. Scientometrics 49 (1), 93–123. https://doi.org/10.1023/A:1005613325648.
Moser, P., Ohmstedt, J., Rhode, P., 2018. Patent citations – an analysis of quality differences and citing practices in hybrid corn. Manag. Sci.
Moser, P.. 2005. "How do patent laws influence innovation?Evidence from nineteenth-century world's fairs." 10.1257/0002828054825501.
Mullainathan, S., Speiss, J., 2017. Machine learning: an applied econometric approach. J. Econ. Persepect.
Murray, F., Stern, S., 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge?. An empirical test of the anti-commons hypothesis. J. Econ. Behav. Organ. 63 (4), 648–687. https://doi.org/10.1016/j.jebo.2006.05.017.
Narin, F., 1994. Patent bibliometrics. Scientometrics 30 (1), 147–155. https://doi.org/10.1007/BF02017219.
Narin, F., Noma, E., 1985. Is technology becoming science? Scientometrics 7 (3-6), 369–381. https://doi.org/10.1007/BF02017155.
Ouellette, L.L., 2011. Do patents disclose useful information. Harv. JL Tech. 25, 545.
Popp, D., 2017. From science to technology: the value of knowledge from different energy research institutions. Res. Policy 46 (9), 1580–1594.
Raftery, AE., 1995. Bayesian model selection in social research. Sociol. Methodol.
Raftery, AE., 1998. Bayes Factors and Bic: Commets on Weakliem. University of Washington Technical Report.
Roach, M., Cohen, W.M., 2013. Lens or prism? Patent citations as a measure of knowledge flows from public research. Manag. Sci. 59 (2), 504–525. https://doi.org/10.1287/mnsc.1120.1644.
Sampat, B., 2010. When do patent applicants search for prior art? J. Law Econ.
Sorenson, O., Fleming, L., 2004. Science and the diffusion of knowledge. Res. Policy 33 (10), 1615–1634.
Tamada, S., Naito, Y., Kodama, F., Gemba, K., Suzuki, J., 2006. Significant difference of dependence upon scientific knowledge among different technologies. Scientometrics 68 (2), 289–302. https://doi.org/10.1007/s11192-006-0112-2.
Taylor, R.B., 2012. Burying. Mich. Telecomm. Tech. L. Rev. 19, 99.
Tijssen, RJ.W., 2002. Science dependence of technologies: evidence from inventions and their inventors. Res. Policy 31 (4), 509–526. https://doi.org/10.1016/S0048-7333(01)00124-X.
Trajtenberg, M., Henderson, R., Jaffe, A., 1992. Ivory Tower Versus Corporate Lab: an Empirical Study of Basic Research and Appropriability. National Bureau of Economic Research.
Trajtenberg, M., 1990. A penny for your quotes: patent citations and the value of innovations. RAND J. Econ. 21 (1), 172. https://doi.org/10.2307/2555502.