NBER WORKING PAPER SERIES

RELIANCE ON SCIENCE BY INVENTORS: HYBRID EXTRACTION OF IN-TEXT PATENT-TO-ARTICLE CITATIONS

Matt Marx
Aaron Fuegi

Working Paper 27987
http://www.nber.org/papers/w27987

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

We thank Kyu Shim, Dmitrii Shelekhov, Elizabeth Terilli, Olivia Bible, and Anand Patel for constructing the list of known-good citations and scoring random samples of algorithm output. Chris Ackerman and Murtadha AlBahrani automated the downloading of patent data. We thank David Orange, Elisabeth Perlman, Felix Poege, and Joseph Staudt for insightful feedback. All computation was performed on the Boston University Shared Computing Cluster, which provided more than half a million CPU-hours to this project. This work was funded by a grant from the Alfred P. Sloan Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-Article Citations
Matt Marx and Aaron Fuegi
NBER Working Paper No. 27987
October 2020
JEL No. O31,O32,O33,O34

ABSTRACT

We curate and characterize a complete set of citations from patents to scientific articles, including nearly 16 million from the full text of USPTO and EPO patents. Combining heuristics and machine learning, we achieve 25% higher performance than machine learning alone. At 99.4% accuracy, coverage of 87.6% is achieved, and coverage above 90% with accuracy above 93%. Performance is evaluated with a set of 5,939 randomly-sampled, cross-verified "known good" citations, which the authors have never seen. We compare these "in-text" citations with the "official" citations on the front page of patents. In-text citations are more diverse temporally, geographically, and topically. They are less self-referential and less likely to be recycled from one patent to the next. That said, in-text citations have been overshadowed by front-page in the past few decades, dropping from 80% of all paper-to-patent citations to less than 40%. In replicating two published articles that use only citations on the front page of patents, we show that failing to capture those in the body text leads to understating the relationship between academic science and commercial invention. All patent-to-article citations, as well as the known-good test set, are available at http://relianceonscience.org.

Matt Marx
Questrom School of Business
Boston University
595 Commonwealth Avenue
Boston, MA 02215
and NBER
mattmarx@bu.edu

Aaron Fuegi
Boston University
aarondf@bu.edu

All data and documentation is available at http://relianceonscience.org

1 Introduction

Innovation is fundamental to economic growth, and many of the most valuable innovations rely on scientific discoveries (Poege et al. 2019). Indeed, science can serve as a "map" for commercial inventors who which to exploit raw technologies, or to inspire their own research & development efforts (Fleming and Sorenson 2004). But tracing the scientific heritage of innovation is challenging for lack of a "paper trail" documenting the inputs used by scientists and engineers in the process of invention. In a seminal paper, Jaffe, Tratjenberg, and Henderson (1993) proposed using patent citations for this purpose, as patent applicants (in the U.S., at least) are obligated to cite all relevant prior art. Their approach of tracking citations from a focal patent to prior patents has been widely adopted, though it is limited by the fact that only a small percentage of scientific articles are ever patented (Belenzon and Schankerman 2013) and thus can at best offer a partial view into how inventors rely on scientific input.

Spurred by the recent availability of open-source databases of patent-to-article citations (Marx and Fuegi 2020), scholars have begun to use citations not only to other patents but also to scientific articles in order to paint a more complete picture of the reliance on science by patent assignees and inventors (e.g. Ahmadpoor and Jones 2017; Arora, Belenzon, and Sheer 2017; Bikard and Marx 2019; Fleming et al. 2019; Marx and Hsu 2019), Roach and Cohen (2013, page 505) conduct a survey of R&D managers, finding that "citations to nonpatent references, such as scientific journal articles, correspond more closely to managers' reports of the use of public research than do the more commonly employed citations to patent references." However, these datasets and analyses are limited in a fundamental way as they consider only the "official" citations explicitly listed on the front-page of patents. However, citations to scientific articles also appear in the body text of the patents (a.k.a "in-text").[1]

The distinction between front-page and in-text citations is not immaterial. As Bryan, Ozcan, and Sampat (2020) observe, citations on the front page of patents serve a *legal* purpose by delimiting the validity of the patent; in other words, a granted patent is presumed to be valid even when considering the citations on the front-page of the patent. For this reason, such citations tend to be carefully reviewed by patent attorneys. By contrast, citations in the full body text are not legally binding and tend to be supplied by the inventors themselves. As such, they have the potential to more accurately represent the sources of scientific inspiration upon which the inventors actually drew in the invention process. In a sample of in-text citations gleaned from articles published in 248 journals since 1984, Bryan et al. report only a 15% overlap between front-page and in-text citations[2] and moreover suggest that in-text citations do a better job of capturing the scientific articles upon which the scientists truly relied upon for inspiration.

1. Citations in the body text of patents go by many names, including "embedded" citations and "fulltext" citations. Following Bryan et al. (Bryan, Ozcan, and Sampat 2020) and deferring to their terminology given that they pioneered the extraction of such citations, we adopt their "in-text" moniker.

2. We find even less overlap, barely 10 percent, when considering all articles in all journals since 1800.

Researchers have been aware of "in-text" citations for decades (Narin and Noma 1985), but the complication of extracting citations embedded in sentences and paragraphs has discouraged the construction of a complete set of in-text patent-to-article citations.[3] Using a hybrid approach that leverages both heuristic-based and machine-learning methods, we retrieve nearly 16 million in-text citations from worldwide patents[4] since 1836 to scientific articles as captured by the Microsoft Academic Graph, PubMed, and Digital Object Identifiers (Sinha et al. 2015) since 1800.

Each citation is assigned a confidence score as well as its origin from applicants vs. examiners. We characterize recall (i.e., coverage, or 1 - false negatives) using a set of 5,939 "known-good" citations, which were cross-verified by multiple research assistants and which the authors have never seen. We calculate precision (i.e., accuracy, or 1 - false positives) with a random sample of 1,000 citations stratified by confidence level. Depending on the confidence score selected, recall of 87.6% can be achieved at 99.4% precision. Of course, recall and precision vary inversely, so we provide ROC-style curves at each confidence level so that users can select the optimal confidence score for their application. Both the patent-to-article citations and the known-good test set can be downloaded at relianceonscience.org. We then characterize in-text vs. front-page citations:

- **Prevalence.** In-text citations have become outnumbered by front-page citations, falling from 80 percent in the late 1970s to less than 40 percent of all scientific citations as of 2019.
- **Recency.** Front page citations point to more recent prior art, whereas in-text citations reach further back in time.
- **Localization.** In-text scientific citations are less localized than either front-page scientific citations or patent-to-patent citations.
- **Self-citation.** In-text citations are less likely to cite articles authored by the inventors on the patent than are front-page citations.
- **Interdisciplinarity.** In-text citations are more interdisciplinary than front-text citations.
- **"Recycling".** Front-page citations are "recycled" in future patents almost 40 percent more often than in-text citations.

We also replicate published findings, adding in-text citations to the front-page citations used for those articles. Including in-text citations in our replication of Ahmadpoor and Jones (2017) reveals that patents are about 40 percent closer to the academic/industry interface than they report. Similarly, we find that Li et al. (2017) understate the percentage of NIH grants that

3. The 248 journals analyzed by Bryan, et al. (Bryan, Ozcan, and Sampat 2020) represent about 1.2% of the more than 20,000 journals and about 5.7% of all articles. The Clarivate Web of Science reports 3,136,867 articles published in those 248 journals from 1984-2017, compared with 55,405,317 total articles.

4. The full text of USPTO patents is examined since 1836. The full text of EPO patents is available since 1978.

give rise to commercial application—as captured by citations from industry patents—by about 20 percent because they did not have access to in-text citations.

In the remainder of the paper, we first describe our "hybrid" algorithm for extracting in-text citations from the full body text of worldwide patents and then linking them to scientific articles, which leverages both heuristic-based and machine-learning approaches. Second, we describe our performance-characterization methods, including the construction of a test set of more than 5,939 "known-good" citations from more than 9,000 randomly-sampled patents. Third, we provide stylized facts regarding front-page vs. in-text citations. Fourth, we replicate prior work using front-page citations only, which underscores the importance of using in-text citations for analysis. All patent-to-article matches, plus the known-good test set, are available for download at http://relianceonscience.org; file contents are described in Appendix C.

2 Challenges of Extracting and Linking In-text Citations to Scientific Articles

Even front-page citations to scientific articles, each of which appears on a separate line, are not simple to extract as they lack a consistent structure and may contain misspellings, errors, and incomplete information. Authors are often but not always at the start of the citation. Years are often incorrect and occasionally missing. Titles are routinely misspelled, truncated, and sometimes absent entirely. In-text citations add at least three layers of complexity. First, whereas each front-page citation appears on its own line, in-text citations only rarely appear as part of a structured bibliography; in the vast majority of cases, they are embedded within longer sentences and paragraphs. Second, precisely because they are embedded, the citations are often embedded in "commentary" about the citation, which can be difficult to separate from identifying information such as author, title, or page numbers. Third, some citations are extremely vague, including only an author name and a year or containing a backreference such as "ibid" with a possibly-unclear antecedent. Figure 1 shows an example of a single patent, which has ten citations in a single paragraph, most of them back-to-back within a single sentence.

Figure 1: Scientific citations embedded in the body-text of a patent

> The recombinant moiety is inserted into a microorganism by transformation and transformants are isolated and cloned, with the object of obtaining large populations capable of expressing the new genetic information. Methods and means of forming recombinant cloning vehicles and transforming organisms with them have been widely reported in the literature. See, e.g., H. L. Heynecker et al, Nature 263, 748-752 (1976) Cohen et al, Proc. Nat. Acad. Sci. U.S.A 69, 2110 (1972); ibid., 70, 1293 (1973); ibid., 70 3240 (1973); ibid., 71, 1030 (1974) Morrow et al, Proc. Nat. Acad. Sci. U.S.A. 71, 1743 (1974) Novick, Bacteriological Rev., 33, 210 (1969); Hershfel et al, Proc. Soc. Nat'l. Acad. Sci. U.S.A. 71, 3455 (1974) and Jackson et al, ibid. 69, 2904 (1972). A generalize discussion of the subject appears in S. Cohen, Scientific American 233, 24 (1975). These and other publications alluded to herein are incorporated by reference.

Notes: Figure shows an example paragraph from patent 5583013. There are ten citations to scientific articles in this paragraph, several of which use "ibid" to refer to previous citations in the same sentence.

Bryan et al. (2020) pioneered the automatic extraction of citations from the body text of patents, focusing on articles published in 248 top journals since 1984. They used rule-based heuristics, relying on both the journal name and year of publication to filter the full text of patents for possible citations. In order to minimize computational requirements, they did not attempt fuzzy matching but required exact matches on year, author, journal (including abbreviations), and either page numbers or the first four words of the title.

An advantage of rule-based heuristics is that they can be fine-tuned for performance, including exception handling of unusually-structured citations. At the same time, customizing such systems can be cumbersome, and executing rules at scale can be computationally intensive—especially if fuzzy matching is employed. At least partially motivated by such concerns, others have turned to machine-learning techniques for extracting scientific citations from the body text of patents. Verberne, Chios, and Wang (2019) combined Conditional Random Fields and Flair approaches on a subset of 33,338 biotechnology patents. (Rassenfosse and Verluise 2020) also apply a machine learning approach, using the open-source GROBID library (Lopez 2009), but applied to the full-text of all USPTO patents. Moreover, they use GROBID to extract not just citations to other patents, product brochures, and legal proceedings.

We adopt a hybrid of rule-based and machine-learning (ML) approaches. The two approaches complement each other both in methodology and advantages. ML, when it works, yields structure of the embedded citation that can contribute to high confidence. The rules-based approach is more computationally expensive but finds about 25 percent more matches than machine learning alone, particularly those with unconventional structure or formatting.

2.1 Rule-based extraction

The full text of USPTO patents since 1836 and EPO patents since 1978 is downloaded from Google Patents and the USPTO website. Prior to 1976, natively digital patent data is no longer available from the USPTO. We initially used the OCRed files from Fleming, et al. (2019) but found

Google's OCR to be more reliable (though far from perfect). We account for errors in optical character recognition (OCR) in two ways. First, we replace '@' with 'a' when it is embedded within a word, such as "self-driving c@r" or (erroneously) "electr@chemical reaction." Second, words containing a possibly-spurious hyphen are split in two if neither of the hyphen-separated words is in the dictionary (e.g., bioinf-ormatics, but not super-conductivity. We removed HTML and other formatting characters and then transliterated accented characters for matching with the Microsoft Academic Graph.

The rule-based method applies to all patents from all years, regardless of whether the underlying text was OCRed. Our method is distinguished from Bryan et al. (2020) in two ways (in addition to covering all articles from all journals). First, whereas they scan the body text for both a year and a journal name/abbreviation to identify portions of the body text that might include a citation, we can find citations without a journal name and even without a year. Second, we make extensive use of fuzzy matching for author names, article titles, bibliographic information, and journal names—any of which might be misspecified. These techniques prove critical in extracting and matching citations, especially from OCRed text. In what may be the first patent citation to a scientific article, Raymond Seilliere and Louis Riot of Paris refer to "[t]he theory of the mechanical equivalent of heat, first established by Jule [sic] in 1843," as shown in Figure2.

Figure 2: Perhaps the first scientific citation in a patent

UNITED STATES PATENT OFFICE.

RAYMOND SEILLIÈRE AND LOUIS M. T. RIOT, OF PARIS, FRANCE.

IMPROVEMENT IN APPLICATION OF SUPERHEATED STEAM TO STEAM-ENGINES.

Specification forming part of Letters Patent No. **202,591**, dated April 16, 1878; application filed February 7, 1878.

To all whom it may concern:

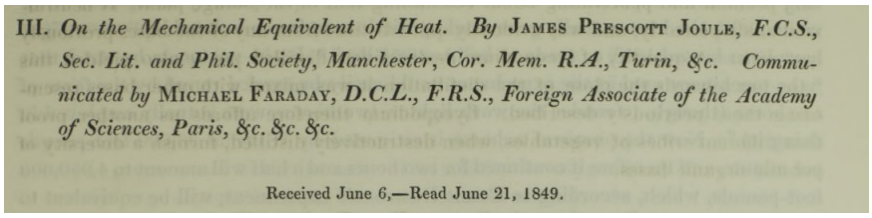
Be it known that we, RAYMOND SEILLIÈRE and LOUIS M. T. RIOT, of Paris, France, have invented a new and Improved Application of Superheated Steam to Steam-Engines, of which the following is a specification:

Most of the steam-engines at present employed use saturated steam, although many attempts have been made and apparatus devised to substitute superheated steam for saturated steam as the dynamic agent, with a view to the more perfect utilization of the fuel. The theory of the mechanical equivalent of heat, first established by Jule in 1843, and in practice, and the many attempts which have been made to employ superheated steam have failed to give the looked-for results, owing to the phenomena not having been sufficiently studied. Thus, it has been hitherto assumed as a general principle in physics that there is an equality of temperature between steam and its origin, so long as these two remain in contact or in communication, and hence it has been concluded that, in order to superheat steam, it must be isolated or separated from its origin, and all arrangements hitherto devised have been based on these data.

Notes: Figure shows the first known scientific citation, from patent 202,591, which was granted in 1878. Note that the author's name is misspelled, and the title is rendered differently than in the original. The year is correct, but the journal is omitted.

The citation is made to the article "On the Mechanical Equivalent of Heat" in the *Philosophical Transactions of the Royal Society*. Note that the author's name is misspelled in the patent (Jule vs. Joule, not corrected in the OCR data), hence flexible matching was employed. Moreover, the first few words of the title are different, but the algorithm can still pick up the match.

Figure 3: Perhaps the first scientific article cited by a patent



III. *On the Mechanical Equivalent of Heat. By* James Prescott Joule, *F.C.S., Sec. Lit. and Phil. Society, Manchester, Cor. Mem. R.A., Turin, &c. Communicated by* Michael Faraday, *D.C.L., F.R.S., Foreign Associate of the Academy of Sciences, Paris, &c. &c. &c.*

Received June 6,—Read June 21, 1849.

2.1.1 Year-based citation extraction

Most but not all scientific citations in patents contain the year of the article. Thus we begin by searching the full text of all patents for four-digit strings that could plausibly be years (i.e., 1800-2020). Numbers in this range appear approximately 70 million times in the full text of U.S. patents. We extract a symmetric, 500-character window surrounding each of these (possible) years.

Of course, not every four-digit string 1800-2020 represents a year, let alone the year of a citation. In the interest of minimizing false positives, as well as some computational savings, we discard windows containing years that are either preceded or followed by a digit or dash (as might be in a page range), as well as those preceded by page/pg/pp/p/vol/v as these are probably page or volume numbers. We also eliminate spurious citation years by checking for disconfirming words either before and after the year. For example, "windows 2000" and "1900 angstroms" both contain what appear to be years but are not scientific citations; The full list of pre- and post-year filter words appears in Appendix A. We also remove a small set of context-spanning pre-and-post-year phrases such as "from <year> to", "of <year> to."

Of particular concern is the risk of mistaking citations to *patents* as citations to scientific articles. When mapping scientific citations from the front pages of patents, where each citation is on a separate line without additional information, one can simply screen out citations that point to patent-related materials. Doing so is less straightforward when extracting citations from the body text of patents because a scientific citation may be mentioned in the same paragraph—or even in the same sentence—as a patent citation. For example, the following excerpt mentions a patent and a scientific article:

> *By way of background, attention is called to the following patents: U.S. Pat. nos. 3,676,337 (Kolm) and 3,902,994 (Maxwell et al). Further background material may be obtained from a report dated October 1977, but not yet published so far as the inventor is aware; the report is entitled "Magnetite recovery in coal washing by high gradient magnetic separation."*

The scientific citation is to an unpublished 1977 article by Maxwell entitled "Magnetite Recovery in Coast Washing by High Gradient Magnetic Separation: Final Report." The referent is not immediately obvious because the patent refers to "the inventor". Yet we are able to link this scientific citation with the article.

To avoid false positives, we filter out years preceding filing dates as is common for applicants to write when referencing patents (which applied for, or granted). Specifically, we screen for the phrase "filing date" or the word "filed" (followed optionally by qualifiers such as "electronically at the uspto" or "in uspto"), then followed by a date including month and day (in either order) and finally the year detected in the earlier step. As an illustration, the following extracted window contains the year 2007:

> *the PWWFA is explained in more detail in a pending U.S. patent application entitled, Parallel Wrapped Wave Front Arbiter, by inventors Wladyslaw Olesinski, Hans Eberle, and Nils Gura, having Ser. No. 11/731,590, and filing date 29 Mar. 2007, which is hereby incorporated by reference*

This citation to a patent might have been incorrectly mapped to an article by Olesinski, Eberle, and Gura in the 2007 proceedings of *High Performance Switching and Routing* titled "PWWFA; the Parallel Wrapped Wave Front Arbiter for Large Switches." The mistake is understandable, given that the pending patent application was titled "Parallel Wrapped Wave Front Arbiter" and the inventors on the patent are identical to the authors on the paper. (Indeed, one would reasonably describe this as a patent-paper pair (Murray 2002), but the focal patent cited the *patent* in the pair but not the paper.) By noticing that the year is referred to in the context of a patent filing, we can avoid this erroenous mapping.

Once spurious years have been removed, it is possible that multiple citation years appear in the same extract, as is visible in Figure 1. Thus the next step is to separate multiple citations within an extract. Semicolons are often used to separate citations, though some inventors use semicolons to delimit authors; thus, we treat semicolons as a citation separator only if there are five or fewer within a 500-character window. We also segment citations within an extract using cue phrases that indicate a citation is about to start or has just ended. Cue phrases include "as described in the article of", "In the article", and "described by" (the full list is found in Appendix A). The word "by" is particularly useful as a cue as it immediately precedes an author name in most cases. These cue phrases are also used to shrink the extracted window. For example, the paragraph in Figure 1 is split into the following ten windows (note that none of these citations contains the title of the article):

1. *See, e.g., H.L. Heynecker et al, Nature 263, 748=752 (1976)*
2. *Cohen et al, Proc. Nat. Acad. Sci U.S.A. 69, 2110 (1972)*
3. ***ibid**., 70, 1293 (1973)*
4. ***ibid**, 70, 3240 (1973)*
5. ***ibid**, 71, 1030 (1974)*
6. *Morrow et al, Proc. Nat. Acad. Sci. U.S.A. 71, 1743 (1974)*
7. *Novick, Bacteriological Rev., 33, 210 (1969)*

8. *Hershfelt et al, Proc. Soc. Nat'l. Acad. Sci. U.S.A. 71, 3455 (1974)*
9. *Jackson et al.,* ***ibid*** *69, 2904 (1972)*
10. *S. Cohen, Scientific American 233, 24 (1975)*

With the potential citations captured and segmented, we also address dependencies. In the paragraph from Figure 1, extracts 3, 4, 5, and 9 contain "ibid" as a backreference, where the word indicates that information is to be borrowed from the prior citation unless it is overriden by new information. In these cases, we remove the word "ibid" from the potential citation and append the entire previous citation except for the year. Although there is the potential for conflicting data to exist, as the ibid citation may include not only a new year but also different volume and/or page numbers (even journal names), our bet is that the correct citation will be found given the updated year. Extracted window 3 is resolved as follows:

"ibid., 70, 1293 (1973)"

does not contain an author, journal, or title - only three numbers, the third in parentheses probably a year. The first two numbers might be volume and first page but are unclear. The notion of *ibid* is to refer back to the previous citation, in this case extracted window 2:

"Cohen et al, Proc. Nat. Acad. Sci. U.S.A. 69,2110 (1972)"

We need from the prior citation Cohen and Proc. Nat. Acad. Sci. U.S.A. (PNAS) (no title is supplied), but it is not immediately obvious which is the author vs. journal vs. (again, missing) title. In practice, we replace the "ibid" with the contents of the prior citation aside from the year, which we take from the ibid. The resulting snippet we then attempt to link is:

"*Cohen et al, Proc. Nat. Acad. Sci. U.S.A. 69,2110* **70, 1293 (1973)**"

The boldfaced portions are from the *ibid* citation, and the italicized portions are from the immediately-previous citation. Although the resulting snippet has two volumes and two page numbers, our bet is that the bibliometric details 70 and 1293 will match an article by Cohen in PNAS with higher confidence in 1973 than in 1972.

Having extracted windows with possible years, dropped extracts containing spurious years, and replaced backreferences, each extract containing a potential scientific citation is then written out on a separate line and fed through a linking and scoring process described in Section 2.3.

2.1.2 Extracting citations when year is missing

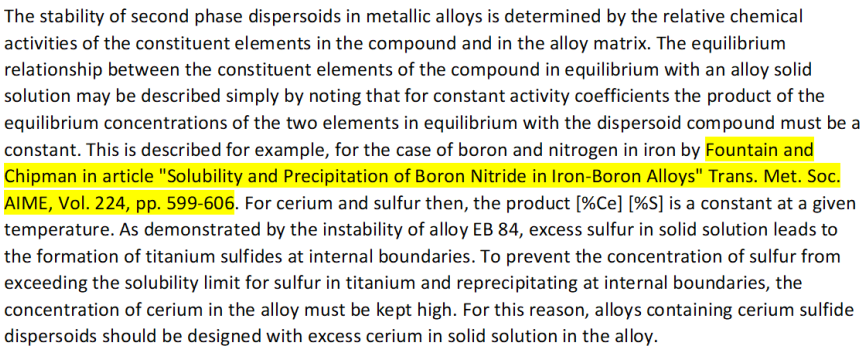
Not every scientific citation contains the year of the article, so we search for citations *without* years in two ways. First, we scan the full text of all USPTO patents for any of 40,000 journal names and more than 150,000 synonyms.[5] This approach differs from Bryan et. al. (2020) both in that we look

5. Journal abbreviations are collected from https://images.webofknowledge.com/images/help/WOS/A_abrvjt.html and supplemented considerably by our own manual review.

for *all* known journal names, and that we search for these in the entire full-text of patents—not just in extracts that contain years. To avoid false positives, we insist on a case-sensitive match to avoid false positives with single-word journals such as Science or Nature.

Second, we look for potential citations even when we cannot match a journal *or* a year, based solely on article titles. Searching the full text of patents for 200,000 journal names and synonyms is itself complicated, but scanning for the titles of more than 179 million unique titles would be computationally prohibitive. As a shortcut, we scan the full text of all patents for phrases surrounded in quotes where a) at least two-thirds of the words were capitalized, and b) the quoted phrase was not already captured in the already-extracted windows containing what appears to be a four-digit year. We then fuzzy-match these quoted strings, which were not otherwise captured in windows either with years or journal names, against the titles of all 179 million articles for potential citations. Figure 4 shows an example.

Figure 4: Finding citations without years via quoted titles



Notes: An example of locating a citation despite not having a year in the text to anchor on. This is done by scanning for quoted phrases where two-thirds of the words are capitalized. Any such phrases that were not already extracted as one of the 70 million windows with a possible date is captured and fuzzy-matched against all 179 million unique titles in the Microsoft Academic Graph.

As with potential citations extracted via year, we write each of these out on a single line and feed it to the linking and scoring process described in Section 2.3. The linking and scoring process can find the article cited, even when lacking a year, as long as the author is present along with other information possibly including journal, title, and other bibliographic info.[6]

2.2 Machine learning

In addition to our rule-based approach, we employ machine learning techniques to extract citations from the body text of patents. Like de Rassenfosse and Verluise (2020), we adopt the open-source

6. We do however require that the first author be specified in the citation. Moreover, it is not sufficient to have only an author and a year without any journal, title, or bibliographic information (e.g., "Smith, 2005") as such citations are too vague to link with confidence. Perhaps in the case of extraordinarily uncommon names there might be only a single article in a given year by that author, but we do not pursue this avenue. In rare cases, an author-year-only citation is shorthand for a citation appearing on the front page of the patent.

GROBID library which has been trained to extract citations from text and tag fields including author, title, journal, and page numbers. We employ the GROBID library in two ways. First, we pass to GROBID entire paragraphs of text from the body of patents. Second, we pass to GROBID the extracted 500-character windows surrounding years, as described in Section 2.1.

Whereas our rule-based approach returns a snippet of text in which we think a citation may appear, GROBID attempts to identify the author, journal, title, etc. and then returns these as structured fields. It thus eliminates "commentary" and other extraneous information from a potential citation and makes the linking process more reliable. As with our rule-based methods, we write out the fields captured by GROBID on a single line and feed that to the linking and scoring algorithm below. Where a citation is found both by rules and ML, we adopt the higher confidence score of the two.

GROBID is useful in handling an unusual type of citation shorthand. Above we discussed handling the *ibid* shorthand where the inventor borrows from the immediately-prior citation in the body text. Another type of shorthand involves listing only the author and year in the body for a citation that appears on the front page. GROBID extracts what it thinks are components of a citation—often incorrectly—but sometimes it extracts an author and year (only), without a journal name, title, etc. In some cases this shorthand is supplemented by a bibliography in the body text of the patent, in which case we should have found the citation anyway, but when such a bibliography is missing we can make the connection by comparing the author and name against those citations already found on the front page. By checking incomplete citations in the text, such as "Smith (2005)", against the front-page citations. Failure to do so would lead the measure of front-page/in-text overlap to be understated.

2.2.1 Comparing heuristic-based methods with machine learning

Which method performs better, machine learning or heuristics? We can examine this question from two perspectives. First, let us consider which algorithm finds more matches. Of the nearly 16 million matches, 71 percent were found by both methods. Of the remaining 29 percent, 23 percent were found *only* by our heuristics. By comparison, six percent of matches were exclusive to GROBID. Hence, a machine-learning system alone substantially underperforms our hybrid approach. (This boost in performance of approximately one-fourth is found not only in the raw number of matches produced but also in the coverage or "recall" rate (i.e., 1 - false negatives). As described in Section 3.1, recall rates for our hybrid approach are about 25 percent higher than when using machine learning alone. Thus it is not the case that the heuristics merely add noise by generating millions of low-confidence, incorrect matches. Rather, machine learning fails to capture a large chunk of citations that should be found.)

That only six percent of matches are exclusive to GROBID may call into question the wisdom of combining the two approaches, given the additional computation required. However, a second

perspective considers the accuracy of the matches found. To evaluate this, we consider the 71 percent of matches that are common to heuristics and machine learning. Which approach has higher confidence scores? We had anticipated that GROBID would achieve higher confidence scores because it extracts structure, but this proved to be less of an advantage than we expected. About half of the time, the two method produce equal confidence scores for the same match. 17 percent of the time, heuristics produce higher confidence scores, and 23 percent of the time, machine learning produces higher confidence scores. We conclude that machine learning is somewhat helpful in finding a small number of matches that would be missed by heuristics alone, and occasionally obtains higher confidence scores. But a machine-learning approach on its own would substantially underperform our hybrid approach, missing nearly a quarter of actual matches.[7]

2.3 Linking and scoring

The foregoing steps reduce the number of potential citations from more than 70 million to about 40 million. The next major step in the process is to determine which of those 40 million actually represent citations from patents to scientific articles. Directly comparing 40 million citations with more than 179 million articles would require quadrillions of pairwise comparisons, so we pre-segment the database of unstructured citations by the year of the article in the supposed citation. (Citations without a year are processed separately.) We also segment the database of unstructured citations based on words ("biochemical"), numbers ("99"), and journal names ("Applied Physics Letters") into a set of 22 million files. This allows us, when searching for a particular paper, to only look at the relatively small set of citations that might match it. Marx and Fuegi (2020) describes this file-based hash table in greater detail.

Although the file-based hash table reduces the search space from quintillions of comparisons to merely trillions, we perform the task in two steps designed to keep it computationally tractable. We first execute computationally-inexpensive "loose" matching to reduce the set of potential matches from trillions to a few billion. This is done by matching only the author and year as well as the longest (or second longest) word in the title of the article, the starting page number (or volume number, if page number is missing)[8], or the journal name/abbreviation. "Loose" matching generates approximately 1.7 billion possible citations, of which only a tiny percentage are correct. We then apply the following, more expensive scoring techniques to determine which of those are likely correct and assign a confidence score to each. (Applying the following scoring techniques to the trillions of potential matches would be computationally prohibitive.)

7. It is possible that ML would improve with additional training, which we have not yet undertaken.

8. We prefer page numbers because they tend to be larger (e.g., 1178-1183), thus avoiding spurious matches. By comparison, many volume numbers are merely one or two digits.

2.3.1 Scoring author name

Author names are sometimes misspelled, so the "loose" matching above employs fuzzy techniques to avoid false negatives. Except for the first character of the name, which must match exactly, a one-character addition, subtraction, or transposition is allowed for a name with between 5 and 9 characters. Longer names may have two such changes. Names of only two or three letters must match exactly. We allow fuzzy matching to account for input errors on the part of the applicant, but the risk is that we will erroneously match an author's name to a random word. For example, an author with surname Grant could be matched to Grand. We partially penalize the fuzzy match (but do not fully rule it out) if it targets a word in the dictionary that is not a common name.

We downweight names that are abbreviations for months (Jun), frequent given names (Morgan), scientific terms (Diamond), or that consist of just two letters. Further, if we are able to determine the author's first initial from the citation and it does not match the database, we sharply downweight the match.[9]

2.3.2 Scoring title

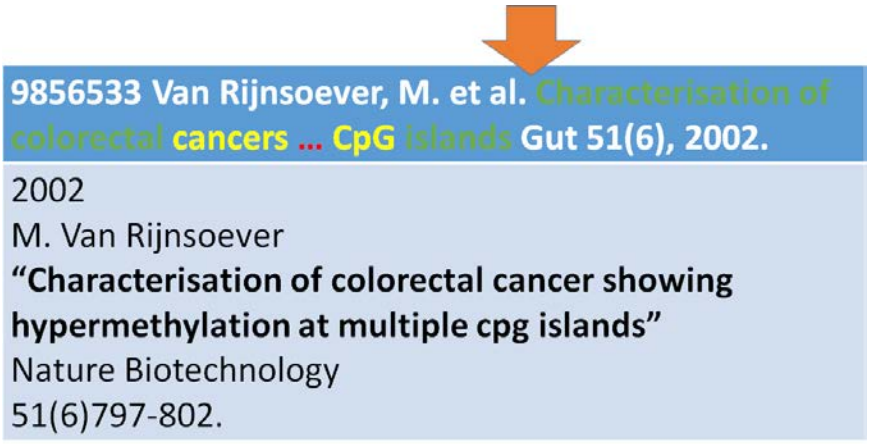
Titles are problematic to match exactly due to truncations and misspellings. Acknowledging that titles can be truncated, Bryan et al. (2020) match exactly the first four words of the title. We undertake a multi-step approach to scoring titles in order to increase coverage while preventing false positives.

One challenge is that it is not known where in a citation the title starts—if it is even present at all. (Recall that none of the in-text citations in Figure 1 included a title.) We proceed by looking for occurrences of each word in the title in the extracted window containing the year of the citation (or, if the window was extracted via journal-name match). For each word, we record its "offset", i.e., the number of words that precede it in the extract. This offset is then decremented by the order of the word in the title, e.g., if *luminescence* were the third word in the title, and it appeared as the fifteenth word in the extract, we would say its offset is twelve. If we are able to establish a modal offset among the title words, we then presume that the title starts at that word of the extract.

We then score the title based on how many of the words of the actual title can be found in sequence, starting at the offset. Partial credit is given for either 1) a slightly misspelled word in the expected sequence, or 2) a correctly-spelled word slightly out of sequence. Stopwords receive only partial credit, even if in the expected sequence and spelled correctly. Longer titles naturally have the potential for higher scores. We calculate both an overall title score and a number-of-initial-words-correct score and use both in the final scoring computation. the process is illustrated with an annotated example in Figure 5.

9. We do not currently take advantage of full first names, or authors other than the first author.

Figure 5: Flexible matching and scoring of titles



Notes: The extracted window appears in the first line of the figure, and the fields from the Microsoft Academic Graph in the remaining lines. We begin by computing the offset from the start of the extracted window (minus its position in the title) for every word of the title (shown in bold on lines 4-5). The offset for "Characterisation" is 7, for "of" is 8-1=7, for "colorectal" is 9-2=7, and so on. Note that the offset for CpG is 11-9=2. There is no offset for "hypermethylation" as this word does not appear in the citation. We set the overall title offset to be 7 as the preponderance of title words have that offset. We then score each word as to whether it matches exactly vs. closely (e.g., "cancer" vs. "cancers") and whether it appears at the expected offset (as with "colorectal") as opposed to in a different position ("CpG" and "islands") or not at all ("hypermethylation"). Partial credit is given for words in the expected position but with different spelling, or words spelled correctly but in a different positions. These scores are summed and then normalized by the length of the title, though giving credit to longer titles that match more words. Common words like "the" receives less credit.

2.3.3 Scoring bibliographic details

Bibliographic elements include the year, volume, issue, starting page, ending page, and journal. Citations almost always include the year, but occasionally the year is incorrectly specified. We allow for years to be off by one, so if the article is listed in the database as 1993 but the citation indicates 1992, a match can still happen but with a slight penalty. If however the citation listed the year as 2004, an article from 1997 could in no case match. If the citation does not contain any year, we perform the match with only author, longest-title-word, and first-page number, with a penalty for not having year (stronger than for year being off by one).[10] We also downweight supposed citations where the article was published more than five years after the patent was granted, with a severe penalty for articles published more than ten years following the patent grant.

Bibliographic scoring depends also on the number of elements (volume, issue, starting page, ending page) matched, whether these appear in set patterns, and the length of individual elements. Confidence increases dramatically if bibliographic data are found in sequence, such as <volume>-<page> and especially <volume>-<issue>-<first page>-<last page>, and especially if these sequences are preceded by "Vol." or when "p." or "pp." . More credit is given for volume, issue, and pages with more digits (e.g., page 8 vs. page 4322). By the same token, if Vol., p., etc is followed by a number that does not match the structured data, we severely downweight the match.

10. Our logic for penalizing a missing year less than a year off by more than one is that an incorrect year serves as disconfirming evidence, whereas a missing year serves only as lack of evidence.

This process is illustrated in Figure 6.

Figure 6: Flexible matching and scoring of bibliographic information

score	NPL
+++++	Smith et al, 2000, Nature, **407(77):312-9**.
++++	Smith et al, 2000, Nature, **407:312-9**.
+++	Smith et al, 2000, Nature, **407:312.**
++++	Smith et al, 2000, Nature, **vol. 407, p. 312**.
+	Abcewigz and Sutton 1977, J. Immunology **3:20-28**
-----	Smith et al, 2000, Nature, 407(74):307-8.

Notes: We scan for the presence of volume, issue, first and last page in the extracted window. Higher scores are achieved by matching more pieces of information, especially in sequence and when identified via due words like "vol" and "pp". Longer numbers score higher; single digits and especially 1 are penalized. The figure shows six scenarios. In the first, we match volume followed by issue and then then first and last page (note that the last page often contains only the distinctive digits). These appear all in sequence; this receives the maximum score. Slightly less credit is given if the issue is not found in sequence, though this is common. A bit less still if only the first page number is available as in the third scenario. The fourth scenario offers a bit more credit, even though it has only volume and first page, because both of these are delineated with the cue words "vol." and "p." which indicate these are truly a volume and page. Absent such markers, as in scenario 3, less credit is given. Even less credit is given in scenario 5 when the volume is a single digit and the page sequence might be mistaken for a date because both are less than 32. Finally, in the sixth scenario we sharply downweight the match if all the bibliometric information is in sequence but does *not* match the original citation, as indicated in red.

Credit is given for a journal-name match if the full or abbreviated journal title is found in the potential citation. The journal score, combined with title, bibliographic, and author score are combined to yield an overall score for the patent-article citation.

2.3.4 Downweighting unlikely cross-disciplinary matches

Finally, we downweight the citation if the top-level CPC category of the patent and the top-level OECD categorization of the article represent a highly unusual pairing. We compute the percentage of citations with confidence = 10 (i.e., highest) from a given CPC class to a given OECD category. If the OECD category is matched by fewer than 10, 5, or 1 percent, then the confidence score is decremented by increasing margins (e.g., if a patent is mapped to an article in a category for which its CPC maps less than 1% of the time, we downweight that most severely). There are 37,942,078 total citations to science from patents. Of those, 15,987,190 are found in the body text, 13,249,991 for USPTO patents and 2,737,149 for EPO patents.

3 Performance characterization

We characterize the performance of our extraction, linking, and scoring procedure above in terms of false negatives and false positives. Counting false positives gives a sense of accuracy (a.k.a.

"precision"), i.e., if the algorithm reports a match, what are the chances that it is correct? Counting false negatives gives a sense of coverage (a.k.a. "recall"), i.e., what percentage of true matches did the algorithm find? Because one can increase precision by decreasing recall and vice-versa, it is essential to calculate both and understand the tradeoffs. As noted above, we anticipate that researchers will have heterogenous preferences for precision vs. recall.[11]

3.1 Construction of known-good validation set (false negatives)

Recall is defined as 1 minus false negatives. To determine recall, we need a gold standard of "known good" citations against which the output of the algorithm can be compared. Once this known-good test set is created, it can be run repeatedly, but the construction itself is time-consuming. Moreover, it is essential that the developers *not* be involved in its curation, lest the algorithm be overfitted to the test set.

We hired several research assistants to create the known-good test set. Each was given a set of randomly-selected patents from which to extract citations from the body text. 9,000 patents were randomly sampled from 1836-2019, with oversampling on the pre-1976 (i.e., OCR) era. Research assistants found more than 6,200 citations, but we kept only those citations found independently from the same set of randomly-selected patents by at least two of the research assistants (5,939 in total). We further segmented these citations into two categories:

1. **Correct:** citation has the correct year and first author's surname (i.e., without misspellings) and some other identifying information including one or more of journal, title, volume, issue, or page. A **Correct** citation may or may not have the journal specified, but if specified the journal cannot be wrong. 5,086 of 5,939 were labeled **Correct** by a research assistant.
2. **Reasonable:** similar to the **Correct** category, except that the citation may have the year off by one and may have misspelled a letter of the first author's surname. 5,629 of 5,939 are **Reasonable** according to a research assistant.

The remaining 310 known-good citations contained substantial error or omissions but could still be matched by a human. We report recall for the **Correct** citations alone, for the set also including **Reasonable** citations, and finally for **All**. Again, the authors have not seen the text of these known-good citations. We make public the known-good test set, consisting of a patent number and a Microsoft Academic Graph ID, but *not* the extracted snippet from the body text of the patent.

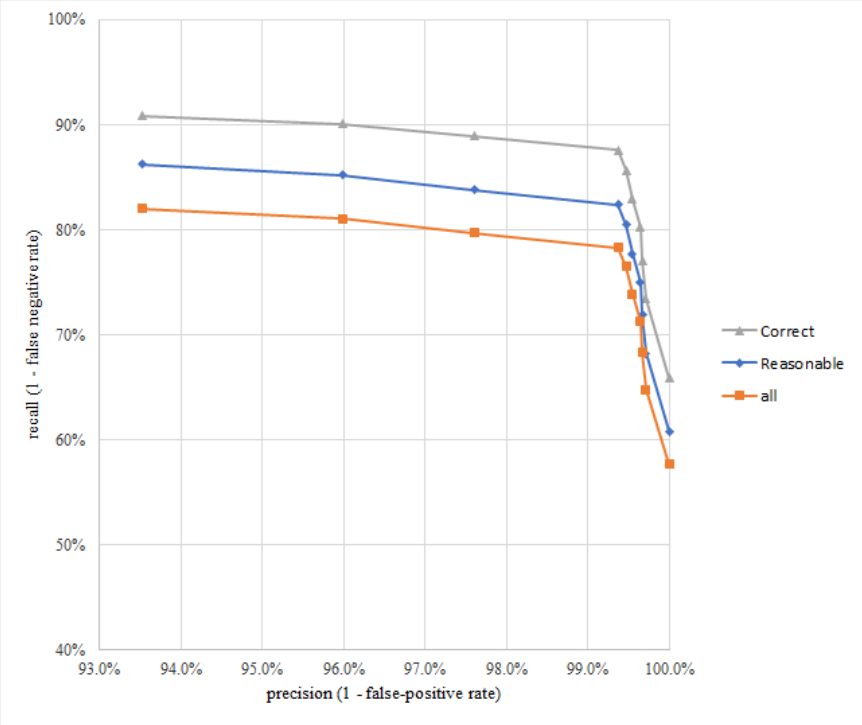
Our algorithm found 87.6% of the **Correct** known-good citations and 82.3% of the **Reasonable** known-good citations, focusing on matches with confidence of at least 4, which corresponds to greater than 99 percent precision. 78.3% of **All** known-good were found. Permitting higher levels

11. Although it has become commonplace to report a single F-measure of performance, this metric weights false positives and false negatives equally whereas researchers may care more about one or the other.

of false positives, recall of over 90 percent can be achieved. We plot recall vs. precision at each confidence level for **Correct**, **Reasonable**, and **All** known-good citations in Figure 7 above.

Returning to the advantages of the hybrid approach as compared to machine learning alone, we note that recall rates using GROBID without the heuristic-based rules results in recall rates about 25 percent lower. This is consistent with our finding that the hybrid approach produces 23% more matches than machine-learning alone. Thus the heuristic-based rules are not simply adding noise but are critical to finding nearly one-quarter of the citations to scientific articles.

Figure 7: Recall vs. Precision



Notes: N = 1,000, 100 matches per confidence level. Confidence levels drop with precision; for example, at 93.5% precision and 90.5% recall, confidence >= 1 whereas at 100% recall and 57% recall, confidence = 10. **Correct:** citation has the correct year and first author's surname (i.e., without misspellings) and some other identifying information including one or more of journal, title, volume, issue, or page (i.e., year and author alone is not Correct). A **Correct** citation may not have the journal specified but cannot have the *wrong* journal. 5,086 of 5,939 were labeled **Correct** by a research assistant. **Reasonable:** citation may have the year off by one and may have a letter misspelled of the first author's surname, and has some other identifying information. It may not have the journal specified, or it may have the wrong journal. 5,629 of 5,939 are **Reasonable** according to a research assistant.

3.2 Precision

Calculating precision can be done straightforwardly by counting false positives in a random sample of generated output. An overall random sample will yield overall precision rates; to understand precision for individual confidence scores, one needs to stratify the sample accordingly. Importantly, a random sample can be used only once to score precision; if the algorithm is changed, that same set of citations may no longer be generable. Moreover, if the algorithm is tweaked to improve performance on the random sample, it must be discarded.

We use a stratified random sample of 1,000 results, with 100 at each confidence level (1-10). We show precision levels at each confidence threshold in Figure 7. Precision is indicated on the x-axis. Working left to right, each point represents precision at confidence levels 1-10. At confidence level 10, the rightmost plotted point, we found no mistakes for precision of 100%. Precision remains above 99% when considering matches at or above confidence level 4. Considering all matches–i.e., at or above confidence level 1, precision of 93.5% is obtained.

Among the mistakes found in scoring were several "near misses." for example, we linked the text of a paragraph

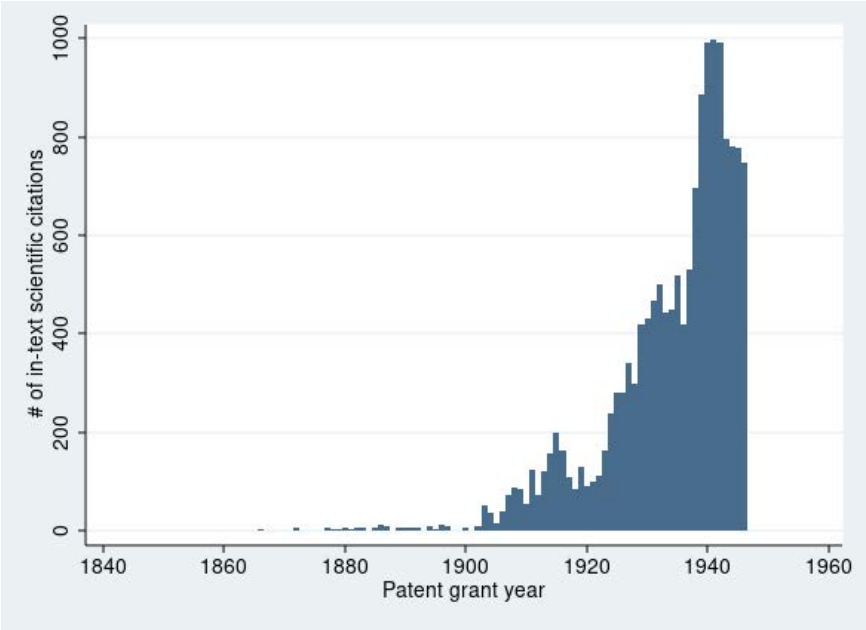
> berry (2002) "assessing probability of ancestry using simple sequence repeat profiles: applications to maize hybrids and inbreds" genetics vol. 161 pp. 813-824

to a 2003 article also by Berry, also in *Genetics*, and with a nearly-identical title: "assessing probability of ancestry using simple sequence repeat profiles applications to maize inbred lines **and soybean varieties**." The crucial difference is bolded. Note however that this match was assigned a confidence score of only 3. As shown in Figure 7, insisting on a confidence score of 4 or higher ensures fewer than one percent false positives.Confidence scores are provided for all citations in the distribution files. In our experience with front-page citations, many scholars choose only to use citations with confidence score = 10 in order to avoid any false positives. As is apparent from Figure 7, however, there is a substantial dropoff in recall when moving from confidence score 4 to 10 but with only a slight increase in precision.

4 Comparing in-text citations with front-page citations to scientific articles

Having assembled a complete set of scientific citations in patents, we can fully characterize in-text citations and compare them with those that appear on the front page. In this section, we focus on USPTO patents. As shown in Figure 8, in-text scientific citations are rare until the early 1900s. In the two decades prior to the introduction of front-page citations in 1947, the number grows to several hundred per year.

Figure 8: Count of in-text scientific citations, 1836-1946

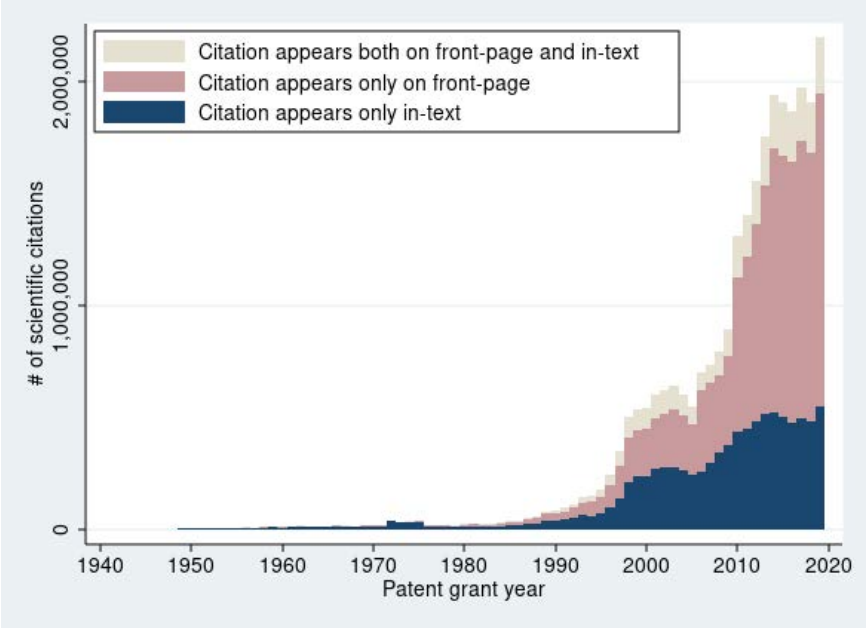


Notes: Each bar indicates the number of in-text citations per year.

4.1 Overlap with Front-page Citations

Starting in 1947, the USPTO began including citations on the front pages of patents. In Figure 9 we plot annual counts of three types of citations: a) those only on the front page, b) those only in the body text, and c) those that appear in both places. As is visible from Figure 9, only a small portion of scientific citations appear in both locations.

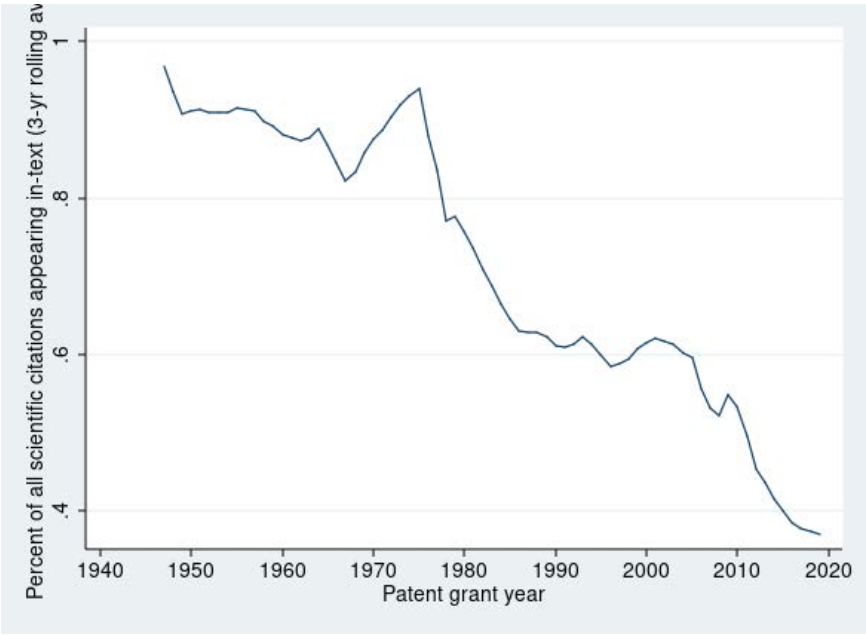
Figure 9: Front-page vs. in-text scientific citations, 1947-2019



Bryan et al. (2020) observed that among the 248 journals they study, only about a third of in-text citations appear on the front page and only about a quarter of front-page citations appear in the text. When analyzing the universe of scientific citations, the picture is even a bit more extreme when considering citations from all patents to all articles from all journals: only 10.5% of scientific citations appear both on the front page and in the body text (57.8% are only on the front page, and 31.6% are only in the text).[12]

Moreover, the two types of scientific citations exhibit different, nearly opposing trends over time. Those only on the front-page appear to have grown more rapidly than those on the front page. Figure 10 makes this latter comparison more explicit by graphing the percentage of scientific citations each year that appear in the body text (whether or not they are on the front page). Most striking is the sharp decrease in the percentage of scientific citations that are embedded in the body text. Prior to the 1980s, 80-90% of all scientific citations were found in the body text. Since then, the percentage of scientific citations found in the body text of patents has decreased gradually, to the point where less than 40% of scientific citations appear in the body text.

Figure 10: Percentage of scientific citations found in the body text of patents, 1947-2019



Notes: Prior to 1947, all scientific citations appeared in the body text.

What explains the sharp decrease in the relative use of in-text scientific citations? Although Figure 9 shows an explosion in the number of front-page scientific citations in the past ten years, Figure 10 indicates that this trend, although recently exacerbated, has been ongoing for decades. Interviews with patent attorneys suggest that there may be a generational shift afoot. One attorney

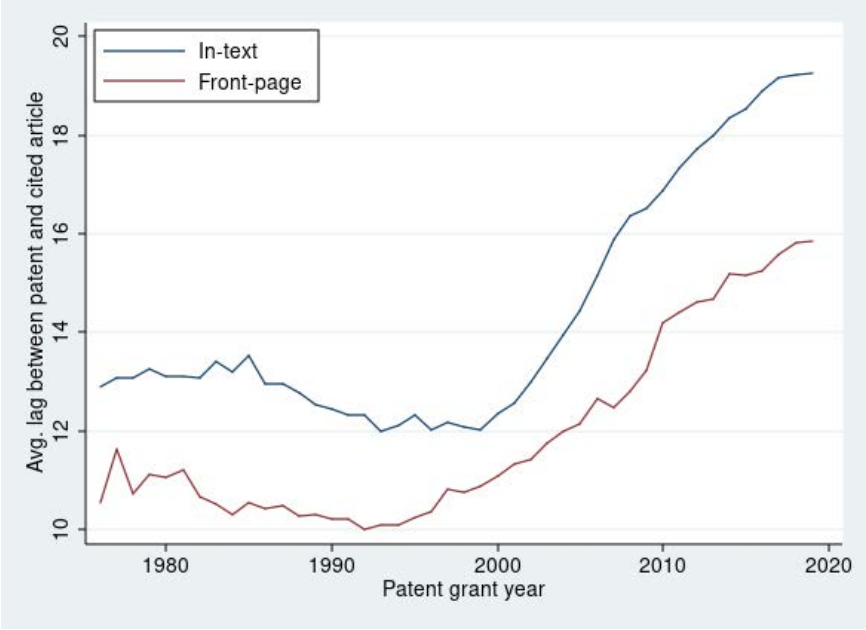
12. In establishing overlap rates it was necessary to manually extract front-page references from OCRed patents for the years 1947-1975 as Google Patents does not report many front-page citations during this period although they appear in the section *Other Citations*. Failing to correct the data from Google Patents would result in understating the overlap.

we interviewed suggested that *all* scientific citations should be in the Invention Disclosure Summary (IDS) and thus appear on the front page of the patent, as in-text citations do not affect the legal scope of the patent. Moreover, he advises clients not to include any citations in the body text of the patent. A USPTO patent examiner we interviewed expressed confusion that in-text citations exist independent of the patent's front page: *"I cannot think of any reason not to include a body-text citation in the IDS."* The same attorney speculated that newly-minted attorneys are averse to including citations in the body text, and that in-text citations should have dropped in the past decade.

4.2 Recency

Front-page and in-text citations also differ in the temporal lag between the citing patent and the cited article. On average, front-page citations reference scientific articles from 14.4 years earlier compared with 16.6 years for in-text citations. The approximate 2-year difference in the age of citations between front-page and in-text is (in unreported results) robust to controlling for year, technology class, and the patent itself. As seen in Figure 11, these differences are rather consistent over time. The average lag lengthens over time as the body of scientific articles continues to grow; if anything, the gap between in-text and front-page citation lags increases.

Figure 11: Lag (in Years) Between Citing Patent and Cited Article



Notes: Figure graphs patents since 1976.

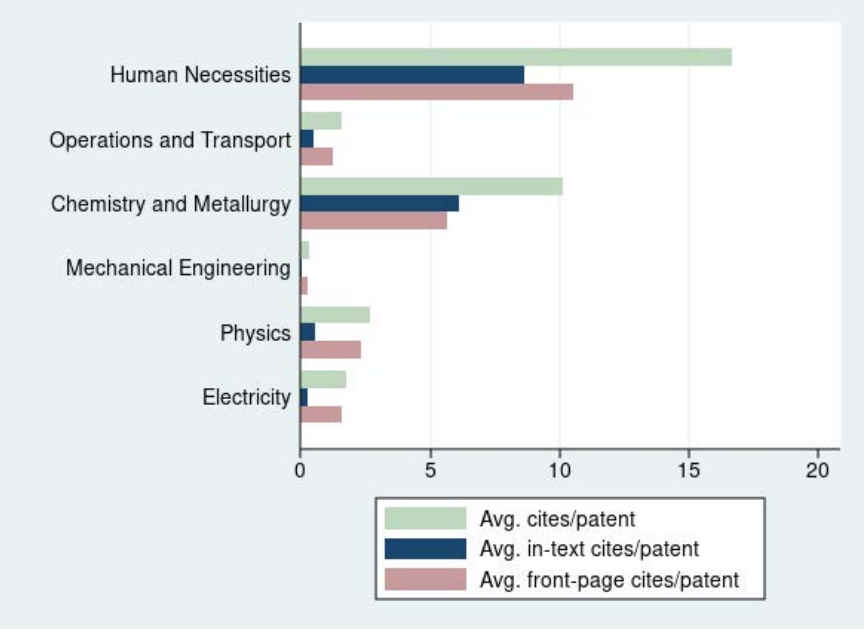
4.3 Technological Fields and Interdisciplinarity

We next explore heterogeneity by technological fields, defined as top-level CPC technical classifications. Figure 12 shows the average number of scientific citations per patent in each CPC field,

by in-text, front-page, or both combined. Human Necessities as well as Chemistry and Metallurgy have the most scientific citations per patent. (Half of patents in both of these categories have at least one scientific citation somewhere in the patent, but the Human Necessities patents with citations have more of them).

Consistent with Section 4.1, front-page citations are more prevalent in nearly every category. Patent Mechanical Engineering has virtually no in-text citations, and they are likewise rare in Electricity and Physics. However, in-text and front-page citations are virtually neck-and-neck in Human Necessities; in Chemistry and Metallurgy, in-text are slightly more common.

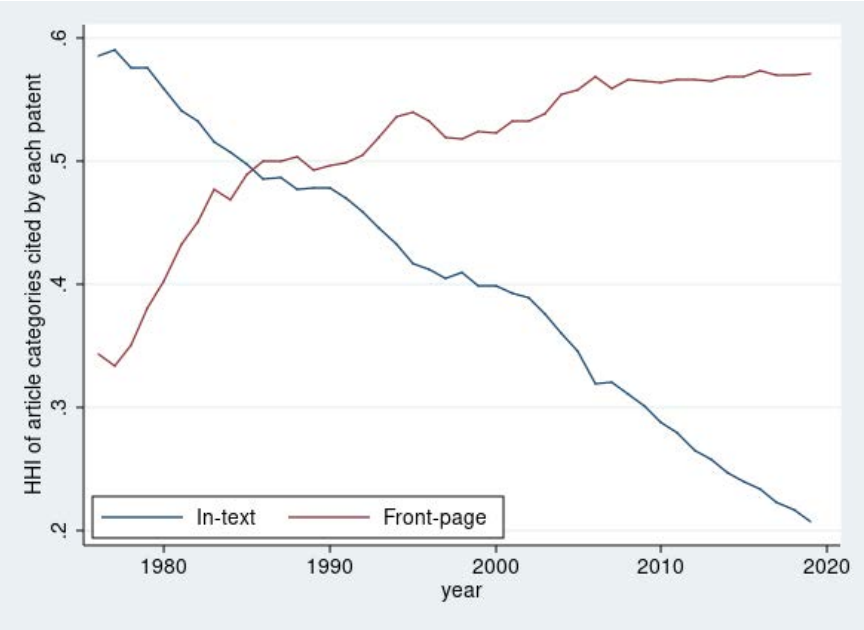
Figure 12: Average number of scientific citations per patent, by CPC



Notes: Figure graphs patents since 1976. Counts for patents in CPC classes Textiles and Fixed Constructions not shown.

Front-page and in-text citations also differ in their degree of interdisciplinarity. We compute interdisciplinarity as the Hersch-Herfindahl Index of the 251 Web of Science categories for the scientific articles cited by each patent (Appendix B details the process of mapping Microsoft Academic Graph articles to Web of Science categories.) Looking at patents with multiple front-page and in-text citations—otherwise HHI would be 1 by construction—in-text citations are consistently more interdisciplinary than front-page citations. This may reflect wider knowledge on the part of the scientists who construct the in-text citations, or perhaps focus by the patent attorneys who assemble the list of legally-binding citations for the front page of the patent. This contrast also holds in unreported results when accounting for the year, CPC class, assignee, and number of scientific citations in the patent. Interestingly, the contrast between in-text and front-page citations has reversed over time, as shown in Figure 13. In-text citations originally had a lower HHI, crossing over with that of front-page citations in the mid-1980s. In unreported results, these trends hold when controlling for various factors including the number of citations in the patent.

Figure 13: Interdisciplinarity of front-page vs. in-text cites

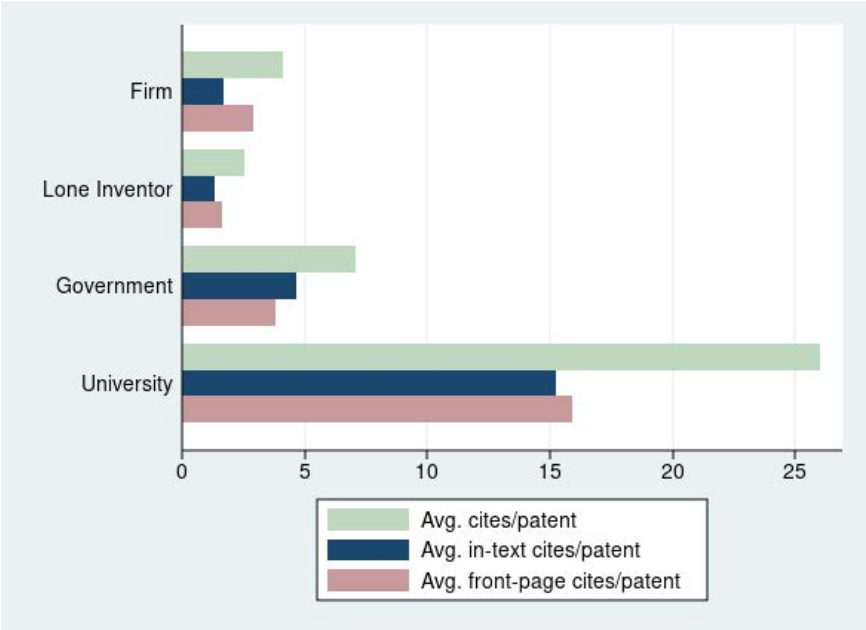


Notes: Includes only patents since 1976.

4.4 Assignee Type

Different types of assignees use citations in distinct ways, as shown in Figure 14. University-affiliated inventors use far more scientific citations than any other type of assignee. Scientific citations are next most prevalent in government patents. Firms are the only assignee type that uses front-page citations substantially more often than in-text citations, possibly reflecting their concern with the validity of the patent as opposed to its scientific heritage.

Figure 14: Average number of scientific citations per patent, by assignee type



Notes: Patents since 1976 are included.

4.5 Self-citation

We compare rates of self-citation across front-page and in-text scientific citations. Lacking a robust identifier across patents and articles, we determine self-citation as any inventor on the patent with the following series of steps:

1. Determine whether the surnames match exactly or nearly, where "nearly" indicates that both surnames are more than five characters long and fewer than of the characters must be changed to convert one to the other (i.e., Levenshtein distance). Moreover, the surnames must start with the same letter (e.g., "Rogers" and "Bogers" are not matched). Two names are treated as a preliminary match if the surname meets these criteria and the first initials also match.
2. To avoid the situation where the author "J Smith" is assumed to be the same as the inventor "Jesse Smith", we score surnames according to their inverse frequency of appearance in the Microsoft Academic Graph. For instance, surname "Smith" would be downscaled to near-zero as it is among the most common author names. Surnames that comprise less than 0.007% of all authors (i.e., 2nd percentile) are not downscaled. If only two authors match between the paper and patent, and both of them represent more than 0.005% of all authors, we conclude that there is no match.
3. Regardless of surname, matches are considered exact if both first and second initials are present for both names and both match.

Unconditionally, 6 percent of in-text citations are self-citations compared with 10 percent of

front-page citations. This finding holds in unreported results when applying fixed effects for year, CPC class, and assignee, or for the patent itself. There could be multiple explanations for this difference. One might have expected that inventors adding citations to the body text of the patent would favor including their own citations into the text, but it appears instead that they are less provincial than patent attorneys. This is broadly consistent with our earlier finding that in-text citations point to older prior art and may reflect broader awareness of relevant literature on the part of inventors than the patent attorneys who tend to be responsible for front-page citations.

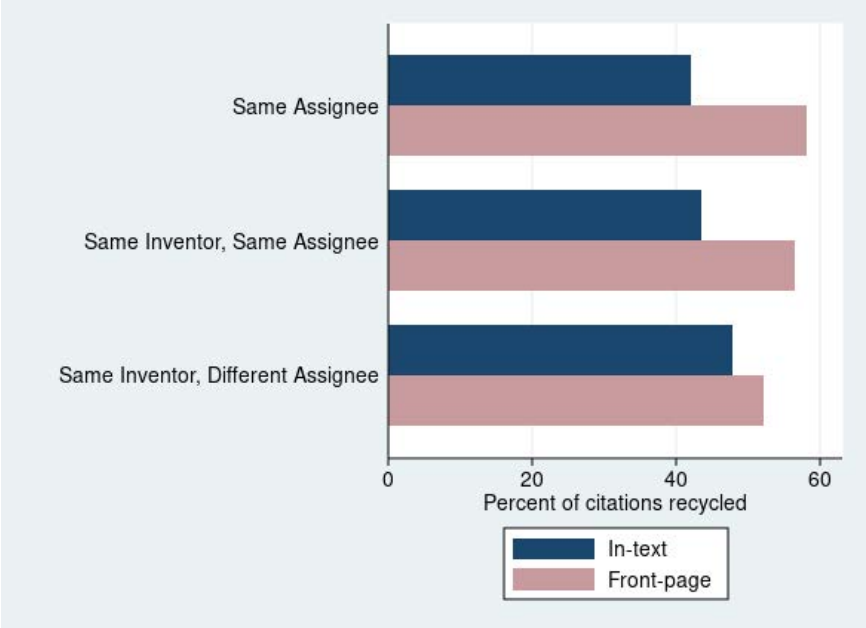
4.6 Citation "Recycling"

In this section we investigate the degree to which the two types of citations are re-used in subsequent patents. For each patent citation to a scientific article, we check whether that article has been cited in the previous patent-filing year. We refer to this re-use of citations as "recycling." All comparisons are made within top-level CPC classes.

Figure 15 shows that front-page citations are recycled at a considerably higher rate than in-text citations. For patents within the same assignee and CPC class, front-page citations are recycled 40 percent more often than in-text citations.

Next, we consider the recycling of citations by inventors. As above, we find that subsequent patents with the same inventor *within the assignee* re-use front-page citations more often. But when the inventor's next patent is at another assignee, s/he is only somewhat more likely to re-use in-text citations than those on the front page.

Figure 15: "Recycling" of Scientific Citations



Notes: Patents since 1976 are included. All comparisons are within the same CPC class. A "recycled" scientific citation is one used again in the subsequent year. (Extending to a five-year window produces similar results.)

4.7 Localization

Belenzon and Schankerman (2013) were the first to show evidence of localization among scientific citations in patents, analyzing citations to *academic* patents using the Jaffe, Tratjenberg, and Henderson (1993) case-control methodology. In unreported results, we replicate their results using in-text citations, either , or instead of, the front-page citations they used. Mindful however of the limitations of the case-control approach (Thompson and Fox-Kean 2005), in Table 1 we implement Thompson's (2006) methodology of distinguishing between citations added by examiners vs. the applicants themselves (see also Singh and Marx (2013) for a large-scale implementation).

Table 1 estimates the geographic proximity between a citing patent and the cited item, with fixed effects for the patent. We begin in column (1) by establishing that front-page citations are localized, consistent with Belenzon and Schankerman (2013). We do this by comparing front-page citations added by examiners against those added by applicants. The average distance between the patent and a front-page scientific citation added by an applicant is about 25 miles lower than for one added by an examiner.

In column (2) we replace front-page citations added by applicants with in-text citations added by applicants comparing these with examiner-added front-page citations from the same patent. (Few if any in-text citations are added by examiners, so such a comparison would be moot.) We find that in-text citations, unlike front-page citations, are *not* localized. Rather, they cite scientific articles on average 19 miles further away than front-page citations added by examiners. The contrast holds when comparing in-text scientific citations against applicant front-page scientific citations in column (3).

Finally, in column (4) we compare the localization of citations to the scientific literature vs. citations to other patents. For this comparison, we exclude citations to patents that were added by examiners (a test we can perform only since 2001). An even greater disparity (more than 100 miles more on average) exists here.

Table 1: Localization of Patent Citations

	Are front-page scientific citations localized?	Are in-text citations more localized than front-page?		Are scientific citations more localized than patent-to-patent citations?
	(1)	(2)	(3)	(4)
Applicant scientific citation (front)	-26.883***			
	(3.625)			
Applicant scientific citation (in-text)		18.666***	53.346***	
		(3.486)	(0.859)	
Applicant scientific citation (front/in-text)				109.094***
				(0.586)
Observations	4095145	6030608	10002384	22756061
R^2	0.298	0.252	0.252	0.191
Patent fixed effects	yes	yes	yes	yes
Scientific front-page (examiner)	yes	yes	no	no
Scientific front-page (applicant)	yes	no	yes	yes
Scientific in-text (applicant)	no	yes	yes	yes
Patent-to-patent (applicant)	no	no	no	yes

Notes: In columns (1-2) we follow Thompson (2006) in establishing the localization of scientific citations by comparing scientific citations added by patent examiners with those added by the applicant and applying patent fixed effects. The remaining columns dispense with examiner citations but continue to use patent fixed effects in comparing front-page vs. in-text citations (3) and scientific vs. patent-to-patent citations (4).

5 Replication

Our descriptive statistics above show that front-page and in-text citations are distinct. But does it matter, and what do we learn from analyzing in-text citations? In this section we attempt to replicate results from the literature.

5.1 Distance from the academic/industry interface

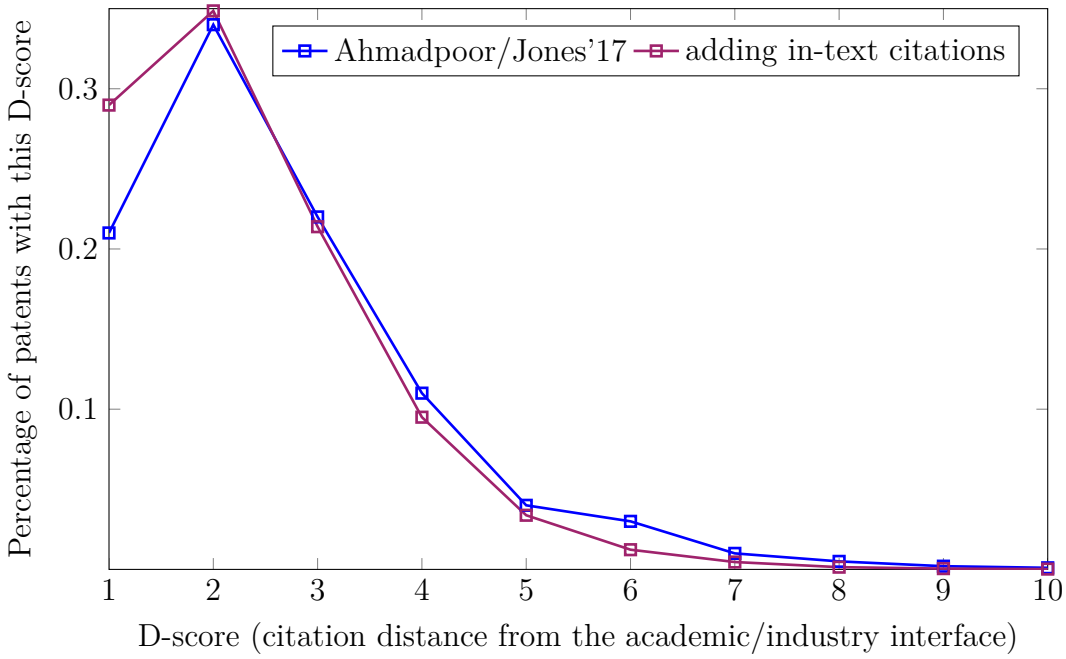
Ahmadpoor and Jones (2017) use front-page scientific citations to measure the distance from the academic/industry interface. They compute a D-score, categorizing articles and patents by the number of citations away from the interface.

We replicate their analysis for patents, establishing the number of patents 1976-2015 for each degree of separation from 1-10. Their distribution is shown in Figure 16. Most striking is the higher percentage of patents with a D-score of 1, which jumps eight percentage points (21% to 29%) when including in-text citations. This represents an increase of 38 percent.

The average D-score shrinks for nearly every NBER category. The largest difference is in Resins, whose mean D-score shrinks from 1.98 to 0.36. Several NBER categories even adjust their mode

D-score from 4 to 3: Furniture and House Fixtures; Motors and Engines and Parts, Receptacles, and Transportation. Agriculture, Food, and Textiles drops from a modal D-score of 2 to 1, showing that the category is closer to the academic/industry interface than previously thought.

Figure 16: Distance of patents from academic/industry interface



Notes: Patents since 1976 are included. All comparisons are within the same CPC class. A "recycled" scientific citation is one used again in the subsequent year. (Extending to a five-year window produces similar results.)

5.2 The applied value of public investments in biomedical research

Li, Azoulay, and Sampat (2017) link patents to grant-supported articles captured by PubMed to show that 31% are cited on the front pages of patents.

We replicate their methodology as closely as possible, though including in-text citations as well. This means considering scientific citations from patents granted until 2012 that are in the life sciences (NBER categories 1 and 3, see Jaffe et al) and that are assigned to firms. They report that 31 percent of all grants (112,408 of 365,380) during the period had commercial impact as measured by front-page citations from patents to articles that acknowledged those grants. Our matching algorithm finds that 161,674 grants have had commercial impact, raising the total to 44.2 percent of all grants. Of those, 26,864 or 16.6 percent were found only in the text of patents.[13]

Taken together with the replication results from Ahmadpoor and Jones (2017), it is apparent that failing to consider in-text citations can result in an substantial understatement of the

13. This magnitude is likely smaller than the difference found in the Ahmadpoor and Jones (2017) replication because one grant may have given rise to several articles, any of which could have been cited on the front-page of a patent. Only in the case where none of the articles associated with a grant was cited on the front-page, but one or more were cited in-text, would there be a difference in this statistic.

commercial impact of academic science. This is consistent with the distinct role of, and minimal overlap between, front-page and in-text citations.

6 Conclusion

Scholars have long sought to trace the heritage of innovation via patent citations. Recent advances in computation have enabled the curation of datasets containing scientific citations from the front page of patents, but as others before us have noted, many citations appear in the body text of patents. Moreover, these citations are not identical to those on the front page. The difficulty of extracting scientific citations from the full text of patents has until now prevented scholars from including these in their analyses. Our contribution here is fourfold.

First, we make available a complete set of nearly 16 million citations from the full-text of patents (USPTO since 1836, and EPO since 1978) to scientific articles. These patent-to-article citations are publicly available for download under an ODC-By license (allowing academic, personal, and commercial use, with attribution). We characterized the coverage and accuracy and provide ROC-like curves so that users of the data can assess the confidence-score cutoff they wish to use. Instead of relying entirely on machine learning, we invest in developing heuristic-based rules that boost coverage by about 25 percent.

Second, we curated the most extensive set of known-good citations from the full-text of patents to scientific articles, 5,939 in all. These citations were harvested from a random sample of 9,000; only citations found by multiple people were retained. The authors have never seen this test set, so the algorithm is not overfitted to it. This test set is also available for download and use by other researchers.

Third, we provide the first descriptive characterization of in-text scientific citations at scale. We compare front-page and in-text citations with regard to recency (i.e., citation lag), localization, self-citation, interdisciplinarity, variation by types of patent assignees, "recycling" of citations, and the percentage of front-page vs. in-text citations in patents over time. From these descriptive statistics we conclude that in-text citations are less provincial than those that appear on the front page, overcoming boundaries of time, proximity, authorship, and disciplines.

Fourth, we replicate the findings of prior articles to show how ignoring in-text citations can lead to understating the connection between academic science and commercial invention. We find that Li. et al. (2017) underestimate the fraction of NIH grants used in downstream commercial development by 20% and that Ahmadpoor and Jones (2017) overestimate the distance of patents from the academic/industrial interface (by nearly 40%) because they were limited to front-page citations. Both of these findings underscore the importance of including in-text citations in analyses.

This work is not without limitations. Although we are able to retrieve 87.6% of scientific citations from the full text of patents when they are correctly specified, and can handle misspellings and even missing data, we do not quite approach the 93% recall rate achieved with front-page

citations. Of course the task is much more difficult with in-text citations, but our algorithm depends on author names which are not always listed (just journal/volume/page in a few cases). Sometimes only the author and year are listed in the text of the patent, but the full citation is written out on the front-page. We might be able to increase recall (as well as overlap) by cross-correlating the two sources.

Our hope is that this dataset will spur previously-infeasible research, including the heritage of innovation, technology transfer from academia to industry, and a deeper understanding of the career paths of scientists and inventors.

References

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The dual frontier: Patented inventions and prior scientific advance." *Science* 357 (6351): 583–587.

Arora, Ashish, Sharon Belenzon, and Lia Sheer. 2017. *Back to basics: why do firms invest in research?* Technical report. National Bureau of Economic Research.

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

Bikard, Michaël, and Matt Marx. 2019. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science.*

Bryan, Kevin A, Yasin Ozcan, and Bhaven Sampat. 2020. "In-text patent citations: A user's guide." *Research Policy* 49 (4): 103946.

Fleming, Lee, Hillary Greene, G Li, Matt Marx, and Dennis Yao. 2019. "Government-funded research increasingly fuels innovation." *Science* 364 (6446): 1139–1141.

Fleming, Lee, and Olav Sorenson. 2004. "Science as a map in technological search." *Strategic management journal* 25 (8-9): 909–928.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic localization of knowledge spillovers as evidenced by patent citations." *the Quarterly journal of Economics* 108 (3): 577–598.

Li, Danielle, Pierre Azoulay, and Bhaven N Sampat. 2017. "The applied value of public investments in biomedical research." *Science* 356 (6333): 78–81.

Lopez, Patrice. 2009. "GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications." In *International conference on theory and practice of digital libraries,* 473–474. Springer.

Marx, Matt, and Aaron Fuegi. 2020. "Reliance on science in patenting." *Strategic Management Journal.*

Marx, Matt, and David H Hsu. 2019. "The Entrepreneurial Commercialization of Science: Evidence From'Twin'Discoveries." *Boston University Questrom School of Business Research Paper Forthcoming.*

Murray, Fiona. 2002. "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering." *Research policy* 31 (8-9): 1389–1403.

Narin, Francis, and Elliot Noma. 1985. "Is technology becoming science?" *Scientometrics* 7 (3-6): 369–381.

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi. 2019. "Science quality and the value of inventions." *Science Advances* 5 (12): eaay7323.

Rassenfosse, Gaétan de, and Cyril Verluise. 2020. "PatCit: A Comprehensive Dataset of Patent Citations."

Roach, Michael, and Wesley M Cohen. 2013. "Lens or prism? Patent citations as a measure of knowledge flows from public research." *Management Science* 59 (2): 504–525.

Singh, Jasjit, and Matt Marx. 2013. "Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity." *Management Science* 59 (9): 2056–2078.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. "An overview of microsoft academic service (mas) and applications." In *Proceedings of the 24th international conference on world wide web,* 243–246.

Thompson, Peter. 2006. "Patent citations and the geography of knowledge spillovers: evidence from inventor-and examiner-added citations." *The Review of Economics and Statistics* 88 (2): 383–388.

Thompson, Peter, and Melanie Fox-Kean. 2005. "Patent citations and the geography of knowledge spillovers: A reassessment." *American Economic Review* 95 (1): 450–460.

Verberne, Suzan, Ioannis Chios, and Jian Wang. 2019. "Extracting and matching patent in-text references to scientific publications."

Appendices

A Filter and chunking words

A.1 Pre-year filters

after / about / above / access / additional / addresses / and / apparatus / appliance / approached / approaching / approximately / assembly / at / becomes / before / between / beyond / cdma / chart / circuit / comprising / computer / contained contains / controller / counted / cpu / device / arly / every / example / exceed / exceeding / exceeds / excel / exchange / fectamine / fig / gate / generator / glycol / having / interface / into / invention / irradiated / jpeg / late / layer / least / letters / line / lipofectamine / machine / magnification / magnified / medium / mediums / memory / meters / method / methods / microprocessor / module / near / nearly / network / networks / number / numbered / numeral / over / plate / portion / powerpoint / process / processor / processors / program / provided / provides / psi / reach / reached / reaches / register / rotation / section / segment / server / since / standard / step / subroutine / substrate / supplied / system / systems / table / temperature / terminal / typically / window / windows / with / within / word

A.2 Post-year filters

amp / ampere / amperes / amps / angstroms / atmospheres / barrels / bits / bp / btu / btus / bytes / calories / cc / cells / celsius / centigrade / centipoise / centipoises / chi / cm / cms / cpm / cubic / cultures / cycles / daltons / degree / degrees / degress differences / dpi / examples / fahrenheit / fectamine / feet / filaments / footpounds / foot-pounds / fpm / ft / gallons / gals / gm / gms / grams / hertz / horsepower / hours / hr / hrs / instrument / inventors / kcal / kilograms / kilowatts / kg / kgs / km / kms / kw / kws / lb / lbs / letters / lines / lipofectamine / liter / liters / meter / meters / micrometers / microns / microseconds / mile / miles / milligrams / milliliters / milliseconds / minutes / ml / mol / movements / mpas / msec / nanometers / nanoseconds / nucleotides / numbered / oersteds / ohm / ohms / ounces / parts / pound / pounds / ppmv psi / psig / pulses / radians / rads / results / revoutions / rpm / samples / sec / seconds / square / standard /states / step / sec / stepsec / steps / times / tons / units / variations / vol / volt / volts / watts / weight / wppm / yards / years

A.3 Pre-citation delimiters

Note: characters appearing within brackets are optionally matched. An underscore indicates a required leading space.

_In addition, / _Preferably, / _Recent work done / _Theoretically it is assumed / _Theoretically, it is assumed / _There is, however, / _While it is stated / _as described in the article of / _as discussed by / _commonly referred to as / _most recently by / _prepared by / _prepared by the method of / _suggested by / _suggested, for example, by / _was used to / _which reports / _which were said to / , for example, / , hereby / , however, / , reports the / , such a / , which are specifically / , which may / According to / After such / Appears in / As pointed out by / As reported therein / As taught in papers such as / Briefly, / Discussed by / For example / Hereafter / Hereinafter / In accordance with the invention / In addition to / In such / In the article / In the preferred embodiment / Many such / Most recently, / One such / Procedure of (but NOT proceedings of) / Reported in / See also / See also, e[.]g[.], / See e[.]g[.] / See for example / See generally / See generally, / See in this regard / See the article / See the article entitled / See, for example, / See, for example, / See, for example, the articles, / The application of the technique / The organisms are / Therefore, / To our knowledge / Work of / [.] Other applicable / a recent study / a report by / according to the method of / according to the procedure described by / according to the procedure of / adapted from / as described by / as described in detail elsewhere / as described in detail previously / as described in the article of / as described, for example, in / as disclosed by / as noted by / as reported by / as set forth by / authored by / by reference herein; / by the method of / cf[.] for example / confirmed by / defined by / demonstrated by / demonstrated first by / demonstrated originally by / described by / described in an article by / described in the following publications: / described in, for example, / escribed, e[.]g[.], in / described, for example, in / described, for instance, in / developed by / disclosed by, for example, / disclosed, for example, in / e[.]g[.] / e[.]g[.] the article, / e[.]g[.], as disclosed by / e[.]g[.], see / e[.]g[.], the articles of / edited by / evidenced, for example, by / explained, for example, in / found, for example, in / illustrated by / improved by / in an article by / in other contexts by / in the article by / in the article of / in the preferred embodiment / in the terminology of / incorporated by reference / incorporated herein by reference / introduced by / invented by / known in the art / observed by / outlined by / performed by a modification of / prepared by / prepared by the method of / presented by / proposed by / provided by / reported by / reported in the literature and patented by / reviewed by / see, for example, the articles of / shown by / stipulated by / study by / suggested by / suggested, for example, by / summarized by / taught by / the analysis of / the articles, / the interested reader / the method described by / the reader is referred / the work of / therefore, / to our knowledge / using the procedure of / work by

A.4 Post-citation delimiters

Note: characters appearing within brackets are optionally matched.

_Preferably, / _and / , for example, / , hereby / , however, / , reports the / , such a / , which are specifically / , which may / . See / According to / After such / Appears in / As pointed out by / As reported therein / As taught in papers such as / Briefly, / Discussed by / For example / Hereafter / Hereinafter / In a paper entitled / In accordance with the invention / In addition to / In addition, / In such / In the article / In the preferred embodiment / Many such / Most recently, / One such / Procedure of (but NOT proceedings of) / Recent work done / Reported in / See also / See also, e[.]g[.], / See e[.]g[.] / See for example / See generally / See generally, / See in this regard / See the article entitled / See, for example, / See, for example, / See, for example, the articles, / The application of the technique / The organisms are / Theoretically it is assumed / Theoretically, it is assumed / There is, however, / Therefore, / To our knowledge / While it is stated / Work of /])] and /])[]) and /])[,] and / [,] and / [.] In / [.] Other applicable / a recent study / a report by / a report in / according to the method of / according to the procedure described by / according to the procedure of / adapted from / as described / as described by / as described in / as described in detail elsewhere / as described in detail previously / as described in the article of / as described, for example, in / as disclosed by / as discussed by / as identified in / as noted by / as reported by / as reported in / as set forth by / as suggested in / authored by / by reference herein; / by the method of / cf[.] for example / commonly referred to as / confirmed by / defined by / defined in / demonstrated by / demonstrated first by / demonstrated originally by / described at page / described by / described in / described in an article by / described in the following publications: / described in, for example, / described, e[.]g[.], in / described, for example, in / described, for instance, in / developed by / disclosed by, for example, / disclosed in / disclosed, for example, in / discussed in / e[.]g[.] / e[.]g[.] the article, / e[.]g[.], as disclosed by / e[.]g[.], see / e[.]g[.], the articles of / edited by / evidenced, for example, by / explained, for example, in / found in / found, for example, in / has indicated that / identified in / illustrated by / improved by / in a letter entitled / in an article appearing in the / in an article by / in an article entitled / in article / in other contexts by / in the art / in the article by / in the article of / in the literature / in the preferred embodiment / in the terminology of / in their article entitled / incorporated by reference / incorporated herein by reference / introduced by / invented by / known in the art / may be found in / most recently by / observed by /outlined by / outlined in / performed by a modification of / pointed out in the paper entitled / prepared by / prepared by the method of / presented by / presented in / proposed by / provided by / provided in / published by / published in / reported by / reported in the literature and patented by / reports in / reviewed by / reviewed in / see, for example, the articles of / set forth in / shown by / stipulated by / study by / suggested by / suggested, for example, by / summarized by / taught by / the analysis of / the articles, / the interested reader / the method described by / the method of / the paper entitled / the reader is referred / the work of / therefore, / to our knowledge / under the title / using the procedure of / was reported in / were reported in / which discusses / which reports / which were said to / work by

B Mapping Microsoft Academic Subjects to Web of Science Categories

The Clarivate web of science categorizes articles into one of 251 subjects. Microsoft Academic Graph does not provide a similar taxonomy; instead, it automatically extracts more than 200,000 fields from the abstracts and titles of the articles themselves. Thus a high-level categorization of MAG articles is not readily available.

We used a DOI-based crosswalk to map approximately 15 million WoS articles to their MAG equivalents. Each of the automatically-extracted keywords in MAG was then given a frequency-weighted distribution of WoS subjects by collapsing the relevant keywords for articles crosswalked from WoS each of the 251 subjects. We then merged this frequency table to each MAG paper's list of keywords to compute the most likely WoS subject.

C File description appendix

Aside from a redistribution of the Microsoft Academic Graph, relianceonscience.org contains two original files. The first, *_pcs_mag_doi_pmid.tsv*, contains the patent-to-article matches described above (including both front-page and in-text matches) and is detailed in Table C1. The second, *bodytextknowngood.tsv*, contains the known-good test set described in Section 3.1 and is detailed in Table C2.

Table C1: Variable description for *_pcs_mag_doi_pmid.tsv* at relianceonscience.org. Both front-page and in-text citations are included in the file; the field "wherefound" distinguishes between the two types.

Variable	Type	Notes
patent	string	Only patents for which our algorithm found a citation are included.
reftype	string	**App** = from applicant. **Exm**=from examiner (Note: all citations from non-USPTO patents are labeled as examiner unless otherwise indicated in the unstructured citation.) **Unk** = if unspecified in the unstructured citation
wherefound	string	**frontonly**, **bodyonly**, or **both** (i.e., both on the front page of the patent, and also in the body text)
confscore	numeric	Assigned confidence score to the match. Only matches with confidence 3 or above are included.
uspto	binary	Indicates whether the patent is from the USPTO. International patents start with a two-letter code for the granting office.
magid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
doi	string	Digital Object Identifier, if available
pmid	numeric	PubMed ID, if available

Notes: File includes only matches found by the algorithm in this paper. No descriptive data for the articles or patents is included, only the unique identifiers. Information about the articles can be linked from relianceonscience.org.

Table C2: Variable description for the curated known-good test set, *bodytextknowngood.tsv*, at relianceonscience.org.

Variable	Type	Notes
patent	string	Patent number
magid	numeric	Unique identifier for each paper in the Microsoft Academic Graph
doi	string	Digital Object Identifier, if available
pmid	numeric	PubMed ID, if available

Notes: File includes known-good matches from 9,000 randomly-sampled patents from 1836-2019, with oversampling on the pre-1976 (i.e., OCR) era. Research assistants found more than 6,200 citations, but we kept only those citations found independently from the same set of randomly-selected patents by multiple research assistants (5,939 in total). .