Research Policy 50 (2021) 104144

Contents lists available at ScienceDirect

Research Policy

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

# Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures

Sam Arts [a,*], Jianan Hou [a], Juan Carlos Gomez [b]

[a] Department of management, strategy and innovation, Faculty of economics and business, KU Leuven, Korte Nieuwstraat 33, Antwerp 2000, Belgium
[b] Department of electronics engineering, University of Guanajuato campus irapuato-salamanca, Carretera Salamanca - Valle de Santiago, Salamanca, Mexico

ARTICLE INFO

*JEL codes:*
O320
O330
O340
D830

*Keywords:*
Natural language processing
Patent
Novelty
Impact
Breakthrough
Award

ABSTRACT

We develop natural language processing techniques to identify the creation and impact of new technologies in the population of U.S. patents. We validate the new techniques and their improvement over traditional metrics based on patent classification and citations in two case-control studies. First, we collect patents linked to awards such as the Nobel prize and the National Inventor Hall of Fame. These patents likely cover radically new technologies with a major impact on technological progress and patenting. Second, we identify patents granted by the United States Patent and Trademark Office but simultaneously rejected by both the European and Japanese patent office. Such patents arguably lack novelty or cover small incremental advances over prior art and should have little impact on technological progress. We provide open access to code, data, and new measures for all utility patents granted by the USPTO up to May 2018 (see https://zenodo.org/record/3515985, DOI: 10.5281/zenodo.3515985).

## 1. Introduction

The creation of new technologies is vital for firm productivity and economic growth (Romer, 1990). New technologies build on old knowledge and technologies and serve themselves as prior art for future generations (Weitzman, 1998). Despite the importance of technological progress, scholars face difficulties to identify and measure the initial creation of new technologies and the subsequent diffusion and impact of these new technologies. Prior work predominantly relies on patent statistics, which are broadly available across countries, industries, and time (Griliches, 1990; Hall et al., 2001; Nagaoka et al., 2010). Yet, there is a huge heterogeneity in the technical novelty and impact of patents. The large majority of patents arguably cover small incremental advances over old technologies and have little impact on technical progress (Griliches, 1990; Lemley and Shapiro, 2005). Only a small minority of patents introduce radically new technologies and extensively serve as prior art for future generations (Trajtenberg, 1990; Scherer and Harhoff, 2000). To measure the technical novelty and impact of patents, prior and current work has traditionally relied on patent classification and citations[1]. But, this approach has important limitations. Patent citations capture prior art but do not reflect the technical content of the patent itself. Thus, patent citations arguably cannot accurately measure the novelty of the technical content. Moreover, citations are sometimes an incomplete and biased representation of prior art (Alcacer and Gittelman, 2006; Lampe, 2012; Lei and Wright, 2017; Kuhn et al., 2020). In contrast to citations, patent classification does reflect the subject matter of the patent, but patent (sub)classes are usually too broad to capture the detailed technical content of the invention and measure technical novelty (Thompson and Fox-Kean, 2005; Arts et al., 2018; Righi and Simcoe, 2019). Therefore, patent classification and citations presumably cannot accurately identify the creation of new technologies – such as polymerase chain reaction, radio-frequency identification, or lithium-ion batteries – and the impact of these new technologies.

In this paper, we develop natural language processing (NLP) techniques to harness the technical content of patent documents. Inventors have to fully disclose their invention in exchange for legal protection. According to U.S. law, a patent must "contain a written description of the invention … in such full, clear, concise, and exact terms as to enable any person skilled in the art … to make and use the same, and shall set forth the best mode contemplated by the inventor or joint inventor of

* Corresponding author.
*E-mail addresses:* sam.arts@kuleuven.be (S. Arts), jianan.hou@kuleuven.be (J. Hou), jc.gomez@ugto.mx (J.C. Gomez).

[1] See for instance: Trajtenberg, 1990; Henderson et al., 1998; Dahlin and Behrens, 2005; Verhoeven et al., 2016; He and Luo, 2017 a,b; Funk and Owen-Smith, 2017; Arts and Fleming, 2018; Kneeland et al., 2020.

https://doi.org/10.1016/j.respol.2020.104144
Received 6 November 2019; Received in revised form 19 August 2020; Accepted 10 October 2020
Available online 3 November 2020

carrying out the invention."[2] Prior work started to explore the use of NLP to detect novelty in text documents[3], and to identify new technologies and trace the diffusion and impact of new technologies in patent text[4]. But, this stream of research currently has a number of shortcomings which limit the adoption of text-based metrics by the broader community. First, prior work provides no large scale validation of the proposed text-based measures. Second, it remains to be tested whether text-based metrics provide an improvement over the traditional metrics based on patent classification and citations. Third, prior work does not compare the performance of different text-based metrics so scholars do not know which measures to adopt. Fourth, prior work does not provide open access to the code, data, and new measures making replication and follow-on research difficult. The current paper contributes to this stream of research by addressing these shortcomings in turn.

To identify new technologies and measure patent novelty at the time of filing, we detect unigrams (keywords), bigrams (two consecutive words), trigrams (three consecutive words), and keyword combinations appearing for the first time in the title, abstract, or claims of a patent, and calculate the cosine similarity between the technical content of a focal patent and all prior patents. To measure the diffusion and impact of new technologies, we count the number of patents reusing the new keywords, bigrams, trigrams, or keyword combinations and calculate the cosine similarity between the content of a focal patent and all succeeding patents. To validate these alternative text-based metrics and their improvement over traditional measures based on patent classification and citations, we employ two case-control study designs. First, we manually collect patents linked to prestigious awards such as the Nobel Prize and the National Inventor Hall of Fame. These patents arguably cover radically new technologies with a major impact on subsequent technical progress and patenting. Second, we exploit the heterogeneity in the examination process at different patent offices, and the notion that the United States Patent and Trademark Office (USPTO) is perhaps granting too many weak or invalid patents (Jaffe and Lerner, 2004). Patent applications granted by the USPTO but simultaneously rejected by both the European (EPO) and Japanese (JPO) patent office presumably lack novelty or cover only small incremental advancement over prior art and should have little impact on technological progress.

Both case-control studies support the use of NLP to measure the technical novelty and impact of patents, and illustrate the improvement over traditional approaches based on patent classification and citations. Interestingly, the new text-based measures only weakly correlate with the traditional measures. We find that new keyword combinations and their reuse by later patents show the strongest discriminatory power to classify both patents linked to prestigious awards and rejected patents. This new measure outperforms the other text-based metrics as well as the traditional measures based on patent classification and citations. We provide open access to data and new measures via https://zenodo.org/record/3515985 (DOI: 10.5281/zenodo.3515985). Python code is available from https://github.com/sam-arts/respol_patents_code.

## 2. Identifying the creation and impact of new technologies

### 2.1. Data collection

We collect patent titles, abstracts and claims for all granted U.S. utility patents from the USPTO, the patent claims research dataset (Marco et al., 2016), and PATSTAT. We include all patents granted between March 1969 and May 2018 ($n$=6,252,916). We only have partial coverage of patents granted before 1976 (approximately 45%), and only information on their titles and abstracts, but not their claims. They are nonetheless included in order to establish a baseline dictionary.

For each patent, we concatenate title, abstract, and claims. Next, we lowercase the text and tokenize it to words using the following regular expression: [a-z0-9][a-z0-9-]*[a-z0-9]+|[a-z0-9]. We consider a word as a sequence of letters and numbers that could be separated by hyphens ("-"). Next, we remove words composed only by numbers, one-character words, stop words from the Natural Language Toolkit (NLTK) in the Python library[5], and words appearing in only one patent. In addition to natural stop words, we remove a manually compiled list of 32,255 very common keywords. First, we compile a list with the most frequently occurring keywords in patents. Next, we identify and exclude those keywords that are unrelated to the technical content of patents, but keep the frequently occurring technical keywords (e.g. internet, bluetooth, dna, rfid, glyphosate). The excluded keywords include both very common non-technical keywords (e.g. invention, discovery, claim, disclose, describe, include, patent) and very common mistakenly combined words which result from the tokenization process (e.g. comprisinga, combinefirst).

Next, we apply stemming to each word using the SnowBall method from the NLTK library. What remains is a collection of unique stemmed keywords which represent the technical content of the patent. The entire cleaned vocabulary contains 1,362,971 unique keywords and the average and median number of unique keywords per patent granted since 1976 is 61 and 56 respectively (stdev=29). As explained in Appendix A, we will provide open access to the list of cleaned and stemmed keywords for each U.S. patent granted up to May 2018. This dataset is an improvement over a prior open access dataset because it includes patent claims (and not just titles and abstracts), applies stemming to reduce words to their root, removes frequently occurring non-technical terms, and includes more recent patents (Arts et al., 2018).

### 2.2. Text-based measures

To measure technical novelty and impact, we develop different alternative measures based on the keywords extracted from the title, abstract, and claims of a patent. The analysis is restricted to patents filed between 1980 and February 2018 ($n$=5,645,845), but patents filed before 1980 are included to compile a baseline dictionary.

First, we calculate *new_word* as the number of unique keywords (unigrams) of a patent that appear for the first time in the U.S. patent database based on filing date. Thus, we identify the first patent using stemmed keywords such as "cryoablat" (US5147355), "stereolithographi" (US4575330), "finfet" (US6413802), and "crispr-ca" (US9951341). We also calculate *new_word_reuse* as the number of new keywords introduced by the focal patent weighted by the number of subsequent patents which reuse these particular keywords, which arguably reflects the impact of the new technology on technical progress. For instance, 644 patents reuse the keyword "stereolithographi"

[2] See 35 U.S. Code §112, available from https://www.law.cornell.edu/uscode/text/35/112

[3] See for example: Allan et al., 2003; Li and Croft, 2008.

[4] See for instance: Reitzig, 2004; Hasan, 2009; Gerken and Moehrle, 2012; Kaplan and Vakili, 2015; Nanda et al., 2015; Packalen and Bhattacharya, 2015; Bergeaud et al., 2017; Kelly et al., 2018; Krieger et al., 2018; Balsmeier et al., 2018; Ashtor, 2019; Watzinger and Schnitzer, 2019; Younge and Kuhn, 2019; Arts and Veugelers, 2020; Teodoridis et al., 2020; Motohashi and Zhu, 2020; Hain et al., 2020; de Rassenfosse et al., 2020; deGrazia et al., 2020.

[5] Examples of stop words from the NLTK library include: the, am, been, does, for, has.

**Table 1**
Sample award patents.

| Award | Period | #awards | #patents | #patents filed since 1980 |
|---|---|---|---|---|
| Nobel Prize | 1975-2018 | 305 | 80 | 59 |
| Lasker Award | 1975-2018 | 221 | 58 | 42 |
| A.M. Turing Award | 1975-2017 | 58 | 17 | 14 |
| National Inventor Hall of Fame | 1990-2018 | 130 | 112 | 60 |
| National Medal of Technology and Innovation | 1985-2014 | 226 | 76 | 42 |
| Benjamin Franklin Medal | 1998-2018 | 149 | 41 | 33 |
| Bower Award | 1990-2018 | 28 | 9 | 9 |
| | | 1,117 | 393 | 259 |

*Notes*: In case the same patent is linked to multiple awards, it is only included once and assigned to one award in the following order: Nobel Prize, Lasker Award, A.M. Turing Award, National Inventor Hall of Fame, National Medal of Technology and Innovation, Benjamin Franklin Medal, Bower Award. The analysis is restricted to patents filed since 1980 to have a sufficiently large baseline dictionary based on all patents filed before 1980.

and 3,238 patents reuse the keyword "finfet".[6] For patent $p$, $new\_word\_reuse_p = \sum_{i=1}^{n}(1+u_i)$ with $n$ equal to the number of new keywords introduced by patent $p$ and $u_i$ equal to the number of future patents which reuse the new keyword $i$. Hence, *new_word* measures the novelty of the patent at the time of filing and *new_word_reuse* measures impact, i.e. how the technical novelty introduced by the focal patent influences later patents. Overall, 1,039,112 new keywords are introduced since 1980, which are on average reused by eight patents (median=2, stdev=278). 10% of all patents introduce at least one new keyword. The average number of new keywords per patent is 0.18 (median=0, stdev=1.63). The average patent has 61 unique keywords so the average and median share of new keywords per patent are 0.3% and 0% respectively.

Second, we calculate *new_bigram* as the number of unique bigrams (two consecutive words) of a patent that appear for the first time in the U.S. patent database. For example, we identify the first patent using bigrams such as "wireless network" (US4789983), "flash memori" (US5053990), "web browser" (US5701451), or "sim card" (US5353328). In line with new keywords, we also calculate *new_bigram_reuse* as the number of new bigrams introduced by the focal patent weighted by the number of subsequent patents which reuse these particular bigrams, which arguably reflects the impact of the new technology on technical progress. For instance, 18,977 patents reuse bigram "flash memori" and 10,391 patents reuse bigram "web browser". Overall, 7,128,180 new bigrams are introduced since 1980, which are on average reused by six patents (median=2, stdev=82). 43% of all patents introduce at least one new bigram. The average number of new bigrams per patent is 1.26 (median=0, stdev=4.69).

Third, we calculate *new_trigram* as the number of unique trigrams (three consecutive words) of a patent that appear for the first time in the U.S. patent database. Thus, we identify the first patent using trigrams such as "graphic user interfac" (US4868785), "local area network" (US4366565), or "nucleic acid sequenc" (US4469863). We also calculate *new_trigram_reuse* as the number of new trigrams introduced by the focal patent weighted by the number of subsequent patents which reuse these particular trigrams. For instance, 27,660 patents reuse trigram "graphic user interfac" and 17,107 patents reuse trigram "nucleic acid sequenc". 11,119,812 new trigrams are introduced since 1980, which are on average reused by three patents (median=1, stdev=45). 54% of all patents introduce at least one new trigram. The average number of new trigrams per patent is 1.97 (median=1, stdev=6.08).

Fourth, we compute *new_word_comb* as the number of unique pairwise keyword combinations of a patent that appear for the first time. To do so, we calculate all possible pairs between any of the keywords of a patent. In contrast to bigrams and trigrams, it does not matter where the keywords appear in the patent, nor the order in which they appear. Patents introducing a new keyword are a subset of patents with new keyword pairs because new keywords by definition result in new combinations. For example, we identify the first patent using keyword combinations such as "vascular stent" (US4580568), "inkjet printhead" (US4677447), or "carbon nanotub" (US5346683). Similarly, we calculate *new_word_comb_reuse* as the number of new keyword combinations weighted by the number of later patents which reuse these particular keyword combinations, which presumably reflects the diffusion and impact of the new technology. For instance, 2,062 patents reuse "vascular stent" and 4,164 patents reuse "inkjet printhead".[7] Thus, *new_word_comb* measures patent novelty at the time of filing and *new_word_comb_reuse* measures impact, i.e. how the technical novelty introduced by the focal patent influences later patents. Overall, approximately 670 million new keyword combinations are introduced since 1980, which are on average reused by two later patents (median=0, stdev=44). 80% of the patents introduce at least one new keyword pair. The average and median number of new keyword pairs per patent is respectively 119 and 16 (stdev=6,705). The high mean and standard deviation are driven by very long patents. Some patent documents cover more than 100 pages, thousands of unique stemmed keywords, and millions of pairwise keyword combinations. Notice that the average patent has 61 unique keywords and 1,830 unique keyword pairs, so that the average and median share of new keyword pairs per patent are 6.5% and 0.9% respectively.

Finally, we calculate *backward_cosine* as the average cosine similarity between a focal patent and all patents filed in the five years before the focal patent.[8] To do so, each patent is represented as a vector of 1,362,971 dimensions where each dimension corresponds to one

[6] As a robustness check, we find that patents reusing a new keyword are indeed more likely to cite as prior art the patent which pioneered the particular keyword. We randomly sample 10,000 new keywords and identify 75,449 patents reusing these keywords. These reusing patents are approximately 4.5 times more likely (4.1% versus 0.9%) to cite the patent which introduced the keyword for the first time compared to a matched control group of similar patents not reusing the keyword. Each reusing patent is matched to one control patent based on technical content and approximate filing date. We calculate the similarity between a reusing patent and all other patents filed in the same year based on the overlap in unique stemmed keywords, and select the patent with the highest Jaccard index but not reusing the particular keyword as control (Arts et al., 2018).

[7] As a robustness check, we find that patents reusing a new keyword combination are indeed more likely to cite the patent which pioneered the particular keyword combination as prior art. We randomly sample 10,000 new keyword combinations which were reused by at least one future patent and identify 53,515 patents reusing these keyword combinations. These reusing patents are approximately 6.6 times more likely (9.2% versus 1.4%) to cite the patent which introduced the keyword combination for the first time compared to a matched control group of patents not reusing the particular keyword combination. Each reusing patent is matched to one control patent based on technical content and approximate filing date (Arts et al., 2018).

[8] We also calculated cosine similarity measures without using a moving time window of five years, but this gave extremely small values with very little variation across patents.

**Table 2**
Examples of awarded inventions.

| Invention | Award | Award year | Awardee | Patent | New words | New word combinations | New bigrams | New trigams |
|---|---|---|---|---|---|---|---|---|
| Discovery of HIV | Nobel Prize | 2008 | Luc Montagnier | US4839288 | hiv-2 (584) | virus hiv-2 (403) | | |
| Chimeric antibodies | Nobel Prize | 2018 | Gregory P. Winter | US5565332 | | antigen-bind cdr3 (618) | scfv fragment (278) | |
| BRCA1 gene locus | Lasker Award | 2014 | Mary-Claire King | US5622829 | brca1 (210) | cancer brca1 (153) | brca1 mutat (16) | |
| HIF-1 factor | Lasker Award | 2016 | Gregg L. Semenza | US5882914 | hif-1 (150) | acid hif-1 (88) | hypoxia induc (185) | hypoxia induc factor-1 (18) |
| Local area network | A.M. Turing Award | 2009 | Charles P. Thacker | US5088091 | host-to-host (28) | configur downlink (2557) | uplink port (130) | discard data packet (78) |
| Public-key encryption | A.M. Turing Award | 2015 | Martin Hellman | US4633036 | | authent public-key (284) | | |
| Polymerase chain reaction | National Inventor Hall of Fame | 1998 | Kary B. Mullis | US4683202 | helicas (533) | amplif primer (5249) | primer extens (1017) | |
| 3D printing | National Inventor Hall of Fame | 2014 | Charles Hull | US4575330 | stereolithographi (644) | surfac stereolithographi (407) | | |
| Metallocene catalysis | National Medal of Technology and Innovation | 2001 | John A. Ewen | US4530914 | zirconocen (156) | polyolefin metallocen (1103) | multimod molecular (115) | multimod molecular weight (111) |
| FinFET | National Medal of Technology and Innovation | 2014 | Chenming Hu | US6413802 | finfet (3238) | gate finfet (2632) | finfet transistor (205) | |
| Erbuim-doped fiber | Benjamin Franklin Medal | 1998 | Emmanuel Desurvire | US4963832 | | amplifi erbium-dop (678) | erbium-dop fiber (514) | erbium-dop fiber amplifi (330) |
| Lithium-ion battery | Benjamin Franklin Medal | 2018 | John Goodenough | US5910382 | lifepo4 (260) | batteri lifepo4 (210) | | |
| Transgenic animals | Bower Award | 1997 | Ralph L. Brinster | US4870009 | transgen (18054) | cell transgen (15357) | transgen anim (1426) | promot dna sequenc (56) |
| DNA sequencing | Bower Award | 2011 | George Church | US4942124 | | multiplex dna (625) | strand molecul (59) | singl strand molecul (23) |

*Notes:* The table includes for each award two examples of awarded inventions. For each awarded invention, the table displays the year of the award, the name of the awardee, the corresponding patent, one new word, new word combination, new bigram, and/or new trigram of the patent. The numbers between brackets show the total number of later patents which reuse these new words, new word combinations, new bigrams, and new trigrams.

keyword from the entire vocabulary and its value captures the frequency of this keyword in the particular patent document. In contrast to *new_word, new_bigram, new_trigram,* and *new_word_comb* which isolate n-grams or keyword pairs, cosine similarity relies on the entire combination of keywords of a patent and also takes into account the frequency of a keyword in a patent.[9] To calculate a measure of technical novelty, we calculate *1-backward_cosine*. More novel patents are arguably more dissimilar in content compared to prior patents. Next, we calculate *forward_cosine* as the average cosine similarity between the focal patent and all patents filed in the five years after the focal patent. To generate an impact measure, i.e. measuring how the new technology influences subsequent technical progress, we calculate *forward/backward_cosine* by dividing *forward_cosine* by *backward_cosine.* Thus, patents with the highest *forward/backward_cosine* are dissimilar to prior patents and similar to later patents. Patents which score among the highest (top 1%) on *forward/backward_cosine* include Amazon's one-click-buying patent (US5960411) and Google's patent for delivering, targeting, and measuring advertising over networks (US5948061). Because the average cosine similarity measures are very small, we standardize the measures for ease of interpretation.

### 2.3. Traditional measures

Prior and current research traditionally relies on patent classification and citation information to measure the technical novelty and impact of a patent. We calculate the most commonly used measures and compare their performance to our new text-based measures.

*New_subclass_comb* is the number of subclass pairs of a patent that appear for the first time in the U.S. patent database based on filing date (Fleming et al., 2007; Jung and Jeongsik, 2016; Verhoeven et al., 2016; Arts and Fleming, 2018). *New_subclass_comb_reuse* is the number of new subclass pairs of a patent weighted by the number of later patents which reuse the particular subclass combinations (Fleming et al., 2007). Similarly, *new_cit_comb* is the number of cited patent pairs which appear for the first time in the patent database, and *new_cit_comb_reuse* is calculated as the number of new cited patent pairs weighted by the number of future patents citing the same two patents (Uzzi et al., 2013; Arts and Fleming, 2018). *Originality* is calculated as one minus a Herfindahl index based on the share of cited patents from each primary patent class (Trajtenberg et al., 1997; Hsu and Lim, 2013; Kaplan and Vakili, 2015). Patents citing prior art from diverse fields are arguably more original or novel (Hirshleifer et al., 2017). *New_tech_origins* counts the number of new combinations between any of the patent classes of the focal patent and any of the classes linked to patents cited by the focal patent (Verhoeven et al., 2016)[10]. It reflects the extent to which a patent sources knowledge from technology fields that were previously never used. *Forward_cit* counts the number of patent citations received within ten years, which reflects the impact of the patent independent from its degree of novelty (Trajtenberg, 1990; Harhoff et al., 1999). Finally, *generality* is calculated as one minus a Herfindahl index based on the share of citing patents from each primary patent class (Jaffe et al., 1998; Henderson et al., 1998). Patents cited by patents from different classes arguably have a general purpose and broad impact across different fields.

[9] We find similar results if we also account for keyword frequency in the entire patent database by using tf–idf weights (term frequency–inverse document frequency).

[10] In line with Verhoeven et al. (2016), we treat all patents introducing a new combination in the same filing year as novel.

While *new_subclass_comb, new_cit_comb, originality*, and *new_tech_origins* reflect the novelty of a patent at the time of filing, *new_subclass_comb_reuse* and *new_cit_comb_reuse* measure the extent to which the technical novelty introduced by the focal patent influences technical progress. Finally, *forward_cit* and *generality* measure the diffusion and impact of a patent independent from its novelty. Because of the skewness of count variables, we use their logarithmic transformation after adding one for variables with zero values.

## 3. Validation award patents

To validate the text-based measures, we manually collect patents linked to prestigious awards (Carpenter et al., 1981). These patents arguably cover radically new technologies with a major impact on technical progress. Next, we use a case-control study design, matching each award patent to a control patent based on technical content and filing date. Finally, we test the ability of the different metrics to correctly classify award and control patents. Our assumption is that award patents are more likely to cover fundamentally new technologies with a significant impact on technical progress and patenting compared to control patents. Prior work indeed points out that the large majority of U.S. patents cover small incremental advances over old technologies and have a rather small – if any – impact on later patents (e.g. Griliches, 1990; Trajtenberg, 1990; Scherer and Harhoff, 2000; Lemley and Shapiro, 2005).

**Table 3**
Descriptive statistics award versus control patents.

| | Award patents | Text-matched control patents | | T test | |
|---|---|---|---|---|---|
| | Average | Average | Cohen's *d* | *t* | Pr(T < t) |
| New_word | 0.470 | 0.280 | -0.339 | -3.854 | 0.000 |
| | (0.596) | (0.525) | | | |
| New_word_reuse | 1.482 | 0.641 | -0.511 | -5.811 | 0.000 |
| | (1.980) | (1.228) | | | |
| New_bigram | 1.613 | 1.126 | -0.543 | -6.183 | 0.000 |
| | (0.886) | (0.908) | | | |
| New_bigram_reuse | 3.504 | 2.459 | -0.549 | -6.252 | 0.000 |
| | (1.894) | (1.909) | | | |
| New_trigram | 1.641 | 1.309 | -0.379 | -4.309 | 0.000 |
| | (0.888) | (0.869) | | | |
| New_trigram_reuse | 3.086 | 2.530 | -0.336 | -3.819 | 0.000 |
| | (1.649) | (1.667) | | | |
| New_word_comb | 4.837 | 3.932 | -0.534 | -6.072 | 0.000 |
| | (1.461) | (1.905) | | | |
| New_word_comb_reuse | 7.001 | 5.505 | -0.672 | -7.644 | 0.000 |
| | (1.956) | (2.469) | | | |
| 1-Backward_cosine | 0.113 | -0.113 | -0.227 | -2.587 | 0.005 |
| | (0.934) | (1.052) | | | |
| Forward/ backward_cosine | 0.018 | -0.018 | -0.035 | -0.401 | 0.344 |
| | (0.978) | (1.023) | | | |
| New_subclass_comb | 1.140 | 0.782 | -0.308 | -3.502 | 0.000 |
| | (1.231) | (1.095) | | | |
| New_subclass_comb_reuse | 2.084 | 1.304 | -0.389 | -4.429 | 0.000 |
| | (2.177) | (1.812) | | | |
| New_cit_comb | 2.173 | 1.927 | -0.130 | -1.485 | 0.069 |
| | (1.965) | (1.807) | | | |
| New_cit_comb_reuse | 2.928 | 2.426 | -0.206 | -2.345 | 0.010 |
| | (2.614) | (2.247) | | | |
| Originality | 0.363 | 0.311 | -0.172 | -1.962 | 0.025 |
| | (0.308) | (0.290) | | | |
| New_tech_origins | 0.039 | 0.034 | -0.026 | -0.295 | 0.384 |
| | (0.211) | (0.202) | | | |
| Forward_cit | 3.046 | 2.208 | -0.639 | -7.270 | 0.000 |
| | (1.374) | (1.244) | | | |
| Generality | 0.645 | 0.535 | -0.442 | -5.028 | 0.000 |
| | (0.232) | (0.266) | | | |

*Notes*: Only patents filed since 1980 are included (n=518 patents; 259 award patent, 259 text-matched control patents). All variables except *1-backward_cosine, forward/backward_cosine, originality*, and *generality* are log transformed after adding 1. Text-matched control patents are matched on stemmed keywords in the titles, abstracts and claims as well as on approximate filing date. Standard deviation between brackets. Cohen's *d* is the mean difference between award and control patents divided by the pooled standard deviation.

### 3.1. Data

*Award patents*. We collect information on the following awards: Nobel Prize (in physics, chemistry, and medicine), Lasker Award (often referred to as America's Nobel Prize), A.M. Turing Award (recognized as the Nobel Prize of Computing), National Inventor Hall of Fame, National Medal of Technology and Innovation, Benjamin Franklin Medal, and the Bower Award[11][12]. We manually match awards to patents except for the National Inventor Hall of Fame which directly provides the corresponding patent numbers. The websites of the award granting organizations and other websites such as Wikipedia provide detailed information to help with the matching. For each granted award, we identify whether it concerns a technological invention, and subsequently collect the names of the individual(s) receiving the award, their affiliation(s) at the time of discovery, information on the timing of the discovery, and the technical description of the awarded invention. We combine all this information to search for corresponding patents using Google patent search queries.

Our main goal is to select a representative sample of patents covering radically new technologies with a high impact on technical progress and patenting, and to avoid false positives rather than false negatives. Thus, we disregard awards and corresponding patents in case of doubt. We encountered a few difficulties. A person might receive an award for a collection of work or as a lifetime achievement rather than for a single invention. In this case, the award is not taken into account. We also exclude awards for which the description is too vague or missing key information to find corresponding patents. In a limited number of cases, the laureates have a large number of patents that closely relate to the description and timing of the awarded discovery. In case we find more than five corresponding patents, we exclude the award. For 65% of the awards linked to patents, we find one corresponding patent. For the remaining 35% of the awards, we find multiple but less than five corresponding patents which match on all criteria including the technical description and timing of discovery. In such cases, we select the patent with the earliest filing date. Nevertheless, all findings are robust for the subset of awards for which exactly one corresponding patent is found (65% of the sample)[13]. Finally, in case a single patent corresponds to multiple awards, we only include the patent once. For example, Kary Mullis won both the Nobel Prize in Chemistry and the National Inventor Hall of Fame for inventing polymerase chain reaction (PCR), a technique to multiply DNA segments. Table 1 provides an overview of the sample selection.

Table 2 provides examples of awarded inventions and corresponding patents. We briefly discuss a number of examples here. Luc Montagnier won the Nobel Prize in Physiology or Medicine in 2008 for the discovery of the human immunodeficiency virus (HIV) which causes AIDS. The corresponding patent (US4839288: "Retrovirus capable of causing AIDS, antigens obtained from this retrovirus and corresponding antibodies and their application for diagnostic purposes") was the first patent using the keyword "hiv-2" and keyword combination "virus hiv-2", reused by respectively 584 and 403 patents. HIV-2 is one of the two major types of the virus. The overall technical content of the patent is also very dissimilar from prior art. The patent is in the 100$^{th}$ percentile of all patents filed in the same year in terms of *1-backward_cosine*, i.e. one minus the

[11] Results are robust if we exclude any of the seven awards from the analysis. Including binary indicators for type and year of award in the regressions does not change our findings.

[12] The online appendix provides a more detailed description of the awards.

[13] Findings also remain robust if we include the first two patents for the subset of awards linked to multiple corresponding patents.

**Table 4**
Likelihood of award patent.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New_word | 0.800*** (0.258) | | | | | | | | | | | |
| New_word_reuse | | 0.469*** (0.082) | | | | | | | | | | |
| New_bigram | | | 1.007*** (0.179) | | | | | | | | | |
| New_bigram_reuse | | | | 0.507*** (0.088) | | | | | | | | |
| New_trigram | | | | | 0.518*** (0.161) | | | | | | | |
| New_trigram_reuse | | | | | | 0.254*** (0.081) | | | | | | |
| New_word_comb | | | | | | | 0.700*** (0.136) | | | | | |
| New_word_comb_reuse | | | | | | | | 0.738*** (0.128) | | | | |
| 1-Backward_cosine | | | | | | | | | 0.661*** (0.174) | | | |
| Forward/backward_cosine | | | | | | | | | | 0.030 (0.154) | | |
| New_subclass_comb | | | | | | | | | | | 0.572*** (0.156) | |
| New_subclass_comb_reuse | | | | | | | | | | | | 0.386*** (0.087) |
| New_cit_comb | | | | | | | | | | | | |
| New_cit_comb_reuse | | | | | | | | | | | | |
| Originality | | | | | | | | | | | | |
| New_tech_origins | | | | | | | | | | | | |
| Forward_cit | | | | | | | | | | | | |
| Generality | | | | | | | | | | | | |
| ll | -283.3 | -270.4 | -272 | -269.2 | -284.2 | -284.7 | -268.6 | -252 | -282.5 | -289.7 | -282.5 | -277.8 |
| Pseudo r2 | 0.09 | 0.13 | 0.12 | 0.13 | 0.08 | 0.08 | 0.13 | 0.19 | 0.09 | 0.07 | 0.09 | 0.11 |
| Precision (%) | 66 | 69 | 66 | 67 | 63 | 63 | 65 | 74 | 64 | 62 | 64 | 66 |
| Recall (%) | 64 | 65 | 68 | 68 | 61 | 61 | 73 | 76 | 65 | 61 | 64 | 64 |
| AUC | 0.70 | 0.73 | 0.73 | 0.73 | 0.68 | 0.69 | 0.74 | 0.79 | 0.68 | 0.66 | 0.70 | 0.71 |
| Marginal effect (%) | 10 | 17 | 20 | 21 | 10 | 9 | 24 | 32 | 14 | 1 | 15 | 17 |

average cosine similarity with all patents filed in the previous 5 years. It is also in the 97$^{th}$ percentile in terms of *forward/backward_cosine*, i.e. the average similarity with all patents filed in the following five years divided by the average similarity with all patents filed in the previous five years.

The 2014 Lasker award was granted to Mary-Claire King for her discovery of the BRCA1 gene locus that causes hereditary breast cancer. The corresponding patent (US5622829: "Genetic markers for breast, ovarian, and prostatic cancer") was the first to use keyword "brca1" and keyword combination "cancer brca1", reused by respectively 210 and 153 patents. The patent is also in the 99$^{th}$ percentile of all patents in terms of *1-backward_cosine* and in the 93$^{th}$ percentile of *forward/backward_cosine*.

Chenming Hu received the National Medal of Technology and Innovation in 2014 for inventing the FinFET transistor, which radically advanced semiconductor technology by taking up less surface than conventional two-dimensional transistors so that more transistors fit on a chip. The corresponding patent (US6413802: "Finfet transistor structures having a double gate channel extending vertically from a substrate and methods of manufacture") was the first patent with keyword "finfet", bigram "finfet transistor", and keyword combination "gate finfet", reused by respectively 3,238; 205; and 2,632 patents. But, the patent is only in the 29$^{th}$ percentile of *1-backward_cosine* and in the 80$^{th}$ percentile of *forward/backward_cosine*.

As a final example, the 2018 Benjamin Franklin Medal was awarded to John Goodenough for his invention of the rechargeable lithium-ion battery, which revolutionized portable electric power. The corresponding patent (US5910382: "Cathode materials for secondary (rechargeable) lithium batteries") was the first patent using keyword "lifepo4" (lithium iron phosphate) and keyword combination "batteri lifepo4", reused by respectively 260 and 210 patents. The patent is in the 94$^{th}$ percentile of *1-backward_cosine*, but only in the 43$^{th}$ percentile of *forward/backward_cosine*.

*Control patents*. Each award patent is matched to one control patent based on technical content and approximate filing date[14]. Most prior work relied on primary patent class and filing year to sample control patents (e.g. Jaffe et al., 1993). However, class-matched patents are often unrelated in content and therefore do not provide a good control group (Thompson and Fox-Kean, 2005; Arts et al., 2018).[15] To select control patents, we calculate the similarity between an award patent and all other patents filed in the same year based on the overlap in unique stemmed keywords, and select the patent with the highest Jaccard index as control. In case multiple patents have an identical Jaccard index, we

[14] By construction, control patents do not belong to the same patent family as an award patent. We rely on the patent family identifiers from PATSTAT, which groups patents sharing the same priority filing.

[15] Using 297 expert ratings, Arts et al. (2018) show that primary-class matched patents received an average similarity rating of 1.89 (median=2) on a scale from 1 to 7 while text-matched patents received an average similarity rating of 3.99 (median=4).

| (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) | (21) | (22) | (23) | (24) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | -0.146 | | -0.074 | | | |
| | | | | | | (0.304) | | (0.312) | | | |
| | | | | | | | | | 0.204** | | 0.255** |
| | | | | | | | | | (0.098) | | (0.119) |
| | | | | | | 0.591** | | 0.589** | | | |
| | | | | | | (0.241) | | (0.241) | | | |
| | | | | | | | | | 0.195 | | 0.171 |
| | | | | | | | | | (0.119) | | (0.126) |
| | | | | | | -0.048 | | -0.084 | | | |
| | | | | | | (0.212) | | (0.216) | | | |
| | | | | | | | | | -0.136 | | -0.202 |
| | | | | | | | | | (0.113) | | (0.124) |
| | | | | | | 0.500*** | | 0.433** | | | |
| | | | | | | (0.173) | | (0.168) | | | |
| | | | | | | | | | 0.568*** | | 0.489*** |
| | | | | | | | | | (0.158) | | (0.159) |
| | | | | | | 0.326* | | 0.431** | | | |
| | | | | | | (0.188) | | (0.193) | | | |
| | | | | | | | | | -0.085 | | -0.125 |
| | | | | | | | | | (0.177) | | (0.174) |
| | | | | | | | 0.534*** | 0.531*** | | | |
| | | | | | | | (0.160) | (0.180) | | | |
| | | | | | | | | | | 0.361*** | 0.322*** |
| | | | | | | | | | | (0.097) | (0.112) |
| 0.342** | | | | | | | 0.276* | 0.199 | | | |
| (0.143) | | | | | | | (0.147) | (0.170) | | | |
| | 0.392*** | | | | | | | | | 0.291** | 0.241* |
| | (0.114) | | | | | | | | | (0.119) | (0.138) |
| | | 1.353** | | | | | 1.262** | 1.368** | | | |
| | | (0.537) | | | | | (0.547) | (0.558) | | | |
| | | | 0.091 | | | | 0.037 | -0.238 | | | |
| | | | (0.589) | | | | (0.636) | (0.569) | | | |
| | | | | 0.670*** | | | | | | 0.421*** | 0.322** |
| | | | | (0.118) | | | | | | (0.133) | (0.142) |
| | | | | | 2.931*** | | | | | 1.790*** | 2.090*** |
| | | | | | (0.599) | | | | | (0.684) | (0.698) |
| -286.8 | -282.9 | -286.6 | -289.7 | -267.9 | -274.6 | -262.8 | -277.9 | -253.4 | -247.8 | -252.8 | -224.6 |
| 0.08 | 0.09 | 0.08 | 0.07 | 0.14 | 0.12 | 0.15 | 0.10 | 0.18 | 0.20 | 0.19 | 0.28 |
| 65 | 64 | 62 | 62 | 68 | 65 | 66 | 67 | 70 | 73 | 72 | 77 |
| 65 | 63 | 62 | 61 | 67 | 70 | 71 | 66 | 73 | 76 | 74 | 79 |
| 0.68 | 0.69 | 0.67 | 0.66 | 0.74 | 0.72 | 0.75 | 0.71 | 0.77 | 0.79 | 0.78 | 0.83 |
| 15 | 21 | 9 | 0 | 19 | 16 | | | | | | |

*Notes:* Logit regression, robust standard errors between brackets. Only patents filed since 1980 are included (n=518 patents; 259 award patents, 259 text-matched control patents). All variables except *1-backward_cosine, forward/backward_cosine, originality*, and *generality* are log transformed after adding 1. Control patents are one-to-one matched to award patents based on stemmed keywords in the titles, abstracts and claims, and approximate filing date. Control variables include number of *words, backward_citations, classes, subclasses*, and *primary_class*, and *filing_year* fixed effects. ll is log likelihood, precision represents the fraction of predicted award patents that are correctly classified, recall is the fraction of real award patents that are correctly classified, AUC is the area under the ROC-curve, marginal effect displays the average marginal effect (in %) of the particular measure increasing with one standard deviation.

select the patent with the closest filing date. The average Jaccard index between award and control patents is 0.22, which corresponds to two average patents of 61 unique keywords that have 22 keywords in common.

### 3.2. Results

Table 3 shows descriptive statistics for the award and control patents. Both the text-based and the traditional measures can distinguish award patents from control patents (i.e. *t*-test significant at 1%), except for *forward/backward_cosine* ($t$=-0.401; $p$=0.344), *new_cit_comb* ($t$=-1.485; $p$=0.069), *originality* ($t$=-1.962; $p$=0.025), and *new_tech_origins* ($t$=-0.295; $p$=0.384). All text-based novelty measures (*new_word, new_bigram, new_trigram,* and *new_word_comb*), measuring technical novelty at the time of filing, outperform the traditional novelty measures based on patent classification and citations (i.e. *new_subclass_comb, new_cit_comb, originality,* and *new_tech_origins*) in terms of both *t*-statistic and Cohen's *d* (i.e. mean difference between award and control patents divided by the pooled standard deviation). An exception is *1-backward_cosine. New_bigram* and *new_word_comb* perform best among all measures for novelty at the time of filing. Looking at all measures together, including the impact measures such as *forward_cit* and *generality*, we find that *new_word_comb_reuse* performs best in identifying new technologies with a major impact on later patents. Table A.2 in the online Appendix displays a correlation matrix. Interestingly, the new text-based measures are only weakly correlated with the traditional measures, and many of the pairwise correlations are even not significant.

Table 4 displays logit regressions with a binary indicator for award patent as outcome. We introduce each measure one by one in columns 1-18, and in columns 19-24 we jointly include all text-based novelty metrics, all traditional novelty metrics, all text-based and traditional novelty metrics combined, all text-based impact metrics, all traditional impact metrics, and finally all text-based and traditional impact metrics combined. In all regressions, we control for the number of unique stemmed *keywords* in the title, abstract and claims of a patent, the number of *backward_citations* to prior patents, the number of patent *classes*, and the number of patent *subclasses*.[16] We further include binary

[16] Findings are robust to the exclusion of these control variables.

**Table 5**
Descriptive statistics granted versus rejected patents.

| | Granted by EPO and JPO | Rejected by EPO and JPO | | T test | |
|---|---|---|---|---|---|
| | Average | Average | Cohen's *d* | *t* | Pr(T < t) |
| New_word | 0.167 | 0.099 | -0.184 | -110.000 | 0.000 |
| | (0.426) | (0.301) | | | |
| New_word_reuse | 0.356 | 0.221 | -0.168 | -95.733 | 0.000 |
| | (0.900) | (0.694) | | | |
| New_bigram | 0.677 | 0.504 | -0.241 | -140.000 | 0.000 |
| | (0.783) | (0.652) | | | |
| New_bigram_reuse | 1.299 | 1.011 | -0.208 | -120.000 | 0.000 |
| | (1.460) | (1.309) | | | |
| New_trigram | 0.880 | 0.747 | -0.163 | -92.782 | 0.000 |
| | (0.864) | (0.769) | | | |
| New_trigram_reuse | 1.522 | 1.332 | -0.138 | -78.780 | 0.000 |
| | (1.423) | (1.331) | | | |
| New_word_comb | 3.314 | 2.825 | -0.243 | -140.000 | 0.000 |
| | (2.054) | (1.974) | | | |
| New_word_comb_reuse | 4.135 | 3.559 | -0.241 | -140.000 | 0.000 |
| | (2.411) | (2.359) | | | |
| 1-Backward_cosine | 0.140 | -0.140 | -0.282 | -160.000 | 0.000 |
| | (0.959) | (1.020) | | | |
| Forward/backward_cosine | -0.003 | 0.003 | 0.006 | 3.375 | 1.000 |
| | (0.976) | (1.024) | | | |
| New_subclass_comb | 0.728 | 0.725 | -0.003 | -1.559 | 0.060 |
| | (1.038) | (1.023) | | | |
| New_subclass_comb_reuse | 1.012 | 0.986 | -0.019 | -10.798 | 0.000 |
| | (1.417) | (1.370) | | | |
| New_cit_comb | 2.449 | 2.464 | 0.008 | 4.615 | 1.000 |
| | (1.896) | (1.876) | | | |
| New_cit_comb_reuse | 2.971 | 2.828 | -0.063 | -36.031 | 0.000 |
| | (2.347) | (2.173) | | | |
| Originality | 0.384 | 0.373 | -0.038 | -21.405 | 0.000 |
| | (0.293) | (0.293) | | | |
| New_tech_origins | 0.008 | 0.006 | -0.019 | -10.639 | 0.000 |
| | (0.089) | (0.077) | | | |
| Forward_cit | 1.859 | 1.617 | -0.199 | -110.000 | 0.000 |
| | (1.249) | (1.178) | | | |
| Generality | 0.334 | 0.311 | -0.077 | -43.665 | 0.000 |
| | (0.305) | (0.305) | | | |

*Notes*: The sample includes all granted U.S. patents filed between 1980 and 2010 which are also filed at the EPO and the JPO. The sample is further restricted to U.S. patents which are also granted by both EPO and JPO (labelled *granted*), and U.S. patents which are rejected by both EPO and JPO (labelled *rejected*). Granted patents are one-to-one matched to rejected patents based on filing years at USPTO, EPO, and JPO, and based on stemmed keywords in the titles, abstracts and claims of the patent (n=1,302,956 patents; 651,478 *granted* patents, 651,478 text-matched *rejected* patents All variables except *1-backward_cosine, forward/backward_cosine, originality*, and *generality* are log transformed after adding 1. Standard deviation between brackets. Cohen's *d* is the mean difference between granted and rejected patents divided by the pooled standard deviation.

indicators for primary *patent_class* and *filing_year* to capture differences across fields and time. To assess the ability of the measures to correctly classify award patents, we calculate precision, recall, and area under the ROC-curve (AUC). Precision is the fraction of predicted award patents that are correctly classified. Recall is the fraction of real award patents that are correctly identified. AUC is a measure between 0.5 (no predictive power) and 1 (perfect classification). The average marginal effect of each measure is calculated as the increase in the likelihood of being an award patent associated with a one standard deviation increase of the particular measure.

The text-based novelty measures *new_word, new_bigram* and *new_word_comb*, measuring novelty at the time of filing, generally outperform the traditional novelty measures based on patent classification and citations (i.e. *new_subclass_comb, new_cit_comb, originality,* and *new_tech_origins*). The text-based novelty measures *1-backward_cosine* and *new_trigram* performs slightly worse compared to *new_word, new_bigram* and *new_word_comb*. *New_word_comb* has the strongest discriminatory power of all novelty measures, and particularly outperforms on recall. A one standard deviation increase in *new_word_comb* increases the likelihood of being an award patent with 24%. Exclusively relying on information available at the time of filing, it correctly identifies 73% of the award patents compared to at most 65% for the traditional novelty measures based on patent classification and backward citations. Thus, *new_word_comb* is able to identify a significant number of highly novel and impactful technologies at the time of filing which cannot be identified with the traditional measures. To give one example, Thomas Steitz received the Nobel Prize in Chemistry in 2009 for using x-ray crystallography to map the structure of ribosomes, which paved the way for more effective antibiotics. The corresponding patent (US6638908: "Crystals of the large ribosomal subunit") introduced many new word combinations such as "x-ray ribosome-bind", but cannot be identified with the traditional novelty measures because it has no new combinations of subclasses or prior art citations, and has a very low score for *originality* and *new_tech_origins*.

As expected, the impact metrics (measuring the impact of new technologies on technical progress and patenting) generally outperform the novelty metrics (measuring novelty at the time of filing). The latter finding is not a surprise given that award patents likely cover technological breakthroughs with a major impact on technological progress. Comparing all measures together, we find that *new_word_comb_reuse* has the strongest ability to correctly classify award patents. It has the highest precision (74%), recall (76%), and AUC (0.79) of all measures, including the traditional impact measures *forward_cit* and *generality*. *Forward_cit* performs best among the traditional measures based on patent classification and citations with a precision of 68%, a recall of 67%, and an AUC of 0.74. As such, *new_word_comb_reuse* has a 6% higher precision and a

9% higher recall in absolute terms compared to *forward_cit*[17]. Increasing *new_word_comb_reuse* with one standard deviation increases the likelihood of being an award patent with 32% while increasing *forward_cit* with one standard deviation increases the likelihood of being an award patent with 19%.

*New_word_comb_reuse* particularly outperforms in terms of recall, again illustrating that highly novel technologies with a major impact are sometimes missed by the traditional measures, including forward citations. To give another example, Peter Agre developed a technique to isolate, clone, and express a protein (aquaporin) which forms pores in the membrane of biological cells to facilitate the transportation of water. Agre's invention laid the foundation for the development of new drugs targeting a broad range of diseases, and rendered him the Nobel Prize in Chemistry in 2003. The corresponding patent (US5858702: "Isolation, cloning and expression of transmembrane water channel Aquaporin 5 (AQP5)") was the first to introduce amongst others keyword "aquaporin" (reused by 93 patents), and keyword combinations "express aquaporin" (reused by 46 patents) and "clone aquaporin" (reused by 13 patents). The patent cannot be identified with the traditional measures and for instance only received a handful of citations from later patents within twenty years. Notice that while *new_word_comb_reuse* has the highest recall of all measures, a significant share of award patents has relatively few reused new words or new word combinations, and as such cannot be identified with the new text-based measures either.

*New_word_comb_reuse* not only outperforms in terms of recall but also in terms of precision. Nevertheless, a significant share of the *predicted* award patents are in fact control patents, illustrating the measure is far from perfect. But, remember that our case-control design only relies on the assumption that the majority of patents cover small incremental advances over old technologies and have a small impact on technical progress. Some of the control patents might nonetheless also cover new technologies with a major impact, but simply did not receive one of the awards we study. For example, one control patent (US4648031: "Method and apparatus for restarting a computing system", assigned to IBM) covers a method for restarting a computer after system failure, introduced many new word combinations and new bigrams reused by thousands of later patents.

## 4. Validation rejected patents

As a second validation, we collect granted U.S. patents which arguably lack novelty – or only cover small incremental advances over prior art – and have little impact on technical progress and patenting. To be granted, a patent application has to demonstrate novelty, i.e. cover a new technical advance over all existing prior art. Yet, prior work suggests that the USPTO is perhaps granting too many weak or invalid patents that fail to meet the novelty requirement compared to the European (EPO) and Japanese (JPO) patent offices which follow the same patentability requirements but have a more careful and time-consuming examination process (Jaffe and Lerner, 2004; Lemley and Shapiro, 2005; Frakes and Wasserman, 2017). European and Japanese examiners devote approximately twice as much time to review a single patent application and reject a much larger share of all applications (Picard and Potterie, 2013; Lemley and Sampat, 2010). The time spend on patent examination is shown to correlate with the quality of the review (King, 2003). To exploit this heterogeneity in the examination process at different patent offices, we use the OECD Triadic Patent Family database to collect all patent applications that simultaneously sought protection at the USPTO, the EPO, and the JPO (OECD, 2019). Our main assumption is that patent applications granted by the USPTO but rejected by both the EPO and the JPO (labelled *rejected*) are more likely to lack novelty – or cover small incremental advances over prior art – and have less impact on subsequent technical progress, compared to patent applications granted by all three patent offices (labelled *granted*).

### 4.1. Data

Our sample only includes granted U.S. patents and not European or Japanese patent documents. But, we restrict the sample to granted U.S. patents that also sought protection at both the EPO and the JPO in order to exploit the heterogeneity in the examination process. Each *granted* patent (i.e. also granted by EPO and JPO) is matched to one *rejected* patent (i.e. rejected by EPO and JPO) based on technical content and filing years at USPTO, EPO, and JPO. Among all rejected patents which jointly match on filing years at the three patent offices, we select the patent with the highest Jaccard index based on the overlap in unique stemmed keywords. We restrict the analysis to patents filed between 1980 and 2010 to have a sufficiently large baseline dictionary and to have at least eight years of observation after filing. Notice that some of the patent filings which we label as rejected might actually be withdrawn by the applicants themselves and therefore not rejected as such by the examiners. Nevertheless, an applicant or attorney presumably withdraws a patent application in case they suspect it will be rejected. Further restricting the sample to patents filed up to 2000, to have at least 18 years of observation after filing, does not change our results. The final sample includes 1,302,956 U.S. patents, of which 651,478 *granted* patents and 651,478 text-matched *rejected* patents. The average Jaccard index between the granted and matched rejected patents is 0.14. For each patent, we calculate all text-based and traditional patent metrics as well as control variables.

### 4.2. Results

Overall, the results for *granted* versus *rejected* patents are in line with our previous findings for the prestigious awards. They support the use of NLP to identify the creation and impact of new technologies, and illustrate the improvement over traditional metrics based on patent classification and citations. Descriptive statistics are shown in Table 5 and logit regressions with a binary indicator for *granted* patent as outcome in Table 6. Table A.3 in the online Appendix displays the correlation matrix. Again, the correlations between the new text-based measures and the traditional measures based on patent citations and classification are low.

Both text-based and traditional measures can distinguish *granted* patents from *rejected* patents. All text-based novelty metrics, measuring novelty at the time of filing, again outperform the traditional novelty metrics based on patent classification and citations in terms of both *t*-statistic and Cohen's *d* (Table 5), and in terms of precision, recall, and AUC (Table 6). In line with the first validation exercise, *new_word_comb* has most discriminatory power to distinguish *granted* patents from *rejected* patents among all metrics for novelty at the time of filing. Increasing *new_word_comb* with one standard deviation increases the likelihood of being granted with 11% while increasing *new_subclass_comb* (i.e. the best performing traditional novelty measure) with one standard deviation increases the likelihood of being granted with 0.4%. Looking at all measures together, including the impact measures such as *forward_cit* and *generality*, we find that *new_word_comb_reuse* again performs best in terms of precision, recall, and AUC.

Not surprisingly, the overall predictive power of the measures to correctly classify *granted* and *rejected* patents is much lower compared to our first validation exercise based on a smaller sample of manually collected patents linked to awards that are likely to cover radically new technologies with a major impact on technical progress. In contrast to prestigious awards, the decision to grant a patent is predominantly based on the technical content and novelty of the patent and not on the (expected) importance or impact of the invention. Therefore, the impact measures do not significantly improve the prediction of which patent gets granted by the three patent offices. Moreover, patent applicants and

[17] Figure A.1 in online Appendix shows the ROC curves of *new_word_comb_reuse* and *forward_cit*, the two measures with the highest AUC

**Table 6**
Likelihood of grant by EPO and JPO.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New_word | 0.650*** | | | | | | | | | | | |
| | (0.006) | | | | | | | | | | | |
| New_word_reuse | | 0.263*** | | | | | | | | | | |
| | | (0.003) | | | | | | | | | | |
| New_bigram | | | 0.481*** | | | | | | | | | |
| | | | (0.003) | | | | | | | | | |
| New_bigram_reuse | | | | 0.219*** | | | | | | | | |
| | | | | (0.002) | | | | | | | | |
| New_trigram | | | | | 0.283*** | | | | | | | |
| | | | | | (0.003) | | | | | | | |
| New_trigram_reuse | | | | | | 0.148*** | | | | | | |
| | | | | | | (0.002) | | | | | | |
| New_word_comb | | | | | | | 0.232*** | | | | | |
| | | | | | | | (0.001) | | | | | |
| New_word_comb_reuse | | | | | | | | 0.199*** | | | | |
| | | | | | | | | (0.001) | | | | |
| 1-Backward_cosine | | | | | | | | | 0.418*** | | | |
| | | | | | | | | | (0.002) | | | |
| Forward/backward_cosine | | | | | | | | | | 0.047*** | | |
| | | | | | | | | | | (0.003) | | |
| New subclass comb | | | | | | | | | | | 0.017*** | |
| | | | | | | | | | | | (0.003) | |
| New_subclass_comb_reuse | | | | | | | | | | | | 0.041*** |
| | | | | | | | | | | | | (0.002) |
| New_cit_comb | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| New_cit_comb_reuse | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| Originality | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| New_tech_origins | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| Forward_cit | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| Generality | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| ll | -880974 | -882349 | -875304 | -878307 | -882068 | -883436 | -871304 | -871033 | -872456 | -887863 | -888012 | -887776 |
| Pseudo r2 | 0.025 | 0.023 | 0.031 | 0.028 | 0.023 | 0.022 | 0.035 | 0.036 | 0.034 | 0.017 | 0.017 | 0.017 |
| Precision (%) | 58 | 58 | 59 | 58 | 57 | 57 | 59 | 59 | 59 | 56 | 56 | 56 |
| Recall (%) | 54 | 54 | 56 | 56 | 55 | 56 | 60 | 61 | 59 | 53 | 53 | 54 |
| AUC | 0.602 | 0.601 | 0.616 | 0.610 | 0.600 | 0.597 | 0.625 | 0.626 | 0.622 | 0.585 | 0.585 | 0.586 |
| Marginal effect (%) | 6 | 5 | 8 | 7 | 6 | 5 | 11 | 11 | 10 | 1 | 0 | 1 |

attorneys are aware of the heterogeneity in the examination toughness at different patent offices, and presumably only file patents at the EPO and JPO above a certain threshold of technical novelty. Therefore, patents which jointly seek protection at the three offices are presumably more novel compared to patents exclusively filed in the U.S.. Given that we only sample patents jointly filed at the three offices makes it arguably more difficult to distinguish between *granted* and *rejected* patents, and lowers the predictive power of the different metrics.

Nevertheless, and despite the limited predictive power, *rejected* patents score significantly lower on almost all measures for novelty and impact. This finding supports prior work suggesting that compared to the EPO and JPO, the USPTO might be more likely to grant patents which lack novelty or cover small advances over prior art, and which have little impact on subsequent technical progress (Jaffe and Lerner, 2004). The social costs from the market power conferred by these patents might perhaps not be offset by the welfare gains from their disclosure and effect on increasing cumulative innovation (Lemley and Shapiro, 2005). Future work could more carefully study this research question with a better dataset and a stronger identification strategy.

## 5. Discussion and conclusion

New technology is a key determinant of firm productivity, economic growth, and welfare. To identify new technologies and measure technological progress, prior work has traditionally relied on patent statistics. A major drawback is that most patents cover small incremental advances over old technologies with little value while only a small minority of patents introduce radically new technologies with a major impact on technological progress and growth. To measure patent novelty and impact, prior work traditionally relied on patent classification and citations. But, citations capture prior art and not the technical content of the patent, and patent classification is typically too broad. Therefore, patent classification and citations arguably cannot accurately identify new technologies and the impact of these new technologies.

In this paper, we developed natural language processing techniques to harness the rich content of patent documents, identify new technologies and their impact on subsequent technological progress and patenting. We validated alternative text-based measures and their improvement over traditional metrics based on patent classification and citations by using patents linked to famous awards and patent rejections. Both validation studies support the use of text mining techniques to identify new technologies and measure patent novelty at the time of filing, and to measure the impact of these new technologies on subsequent innovation. They also illustrate the improvement over traditional measures based on patent classification and citations. In line with the classic view of invention as a cumulative and combinatorial search process, we find that new combinations of keywords appearing in the technical description of the patent has the strongest discriminatory power to identify new technologies and measure novelty at the time of filing. Moreover, weighting the new keyword combinations by the

| (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) | (21) | (22) | (23) | (24) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 0.193*** | | 0.186*** | | | |
| | | | | | | (0.006) | | (0.006) | | | |
| | | | | | | | | | 0.103*** | | 0.103*** |
| | | | | | | | | | (0.003) | | (0.003) |
| | | | | | | 0.187*** | | 0.190*** | | | |
| | | | | | | (0.004) | | (0.004) | | | |
| | | | | | | | | | 0.096*** | | 0.087*** |
| | | | | | | | | | (0.002) | | (0.002) |
| | | | | | | 0.090*** | | 0.100*** | | | |
| | | | | | | (0.003) | | (0.003) | | | |
| | | | | | | | | | 0.049*** | | 0.032*** |
| | | | | | | | | | (0.002) | | (0.002) |
| | | | | | | 0.149*** | | 0.153*** | | | |
| | | | | | | (0.001) | | (0.001) | | | |
| | | | | | | | | | 0.154*** | | 0.153*** |
| | | | | | | | | | (0.001) | | (0.001) |
| | | | | | | 0.322*** | | 0.319*** | | | |
| | | | | | | (0.003) | | (0.003) | | | |
| | | | | | | | | | -0.029*** | | -0.055*** |
| | | | | | | | | | (0.003) | | (0.003) |
| | | | | | | | 0.025*** | 0.007** | | | |
| | | | | | | | (0.003) | (0.003) | | | |
| | | | | | | | | | | 0.029*** | 0.006*** |
| | | | | | | | | | | (0.002) | (0.002) |
| -0.070*** | | | | | | | -0.072*** | -0.096*** | | | |
| (0.002) | | | | | | | (0.002) | (0.002) | | | |
| | 0.019*** | | | | | | | | | 0.005*** | -0.020*** |
| | (0.001) | | | | | | | | | (0.001) | (0.001) |
| | | 0.045*** | | | | | 0.066*** | 0.051*** | | | |
| | | (0.008) | | | | | (0.008) | (0.008) | | | |
| | | | 0.140*** | | | | 0.142*** | 0.108*** | | | |
| | | | (0.022) | | | | (0.022) | (0.023) | | | |
| | | | | 0.238*** | | | | | | 0.265*** | 0.233*** |
| | | | | (0.002) | | | | | | (0.002) | (0.002) |
| | | | | | 0.370*** | | | | | -0.212*** | -0.217*** |
| | | | | | (0.007) | | | | | (0.009) | (0.009) |
| -886971 | -887907 | -876312 | -888013 | -879109 | -886676 | -856752 | -875159 | -843801 | -867057 | -878664 | -860632 |
| 0.018 | 0.017 | 0.017 | 0.017 | 0.027 | 0.018 | 0.051 | 0.018 | 0.053 | 0.040 | 0.027 | 0.047 |
| 57 | 56 | 56 | 56 | 58 | 57 | 61 | 57 | 61 | 60 | 58 | 61 |
| 54 | 53 | 54 | 53 | 56 | 54 | 60 | 54 | 60 | 59 | 56 | 59 |
| 0.588 | 0.585 | 0.585 | 0.585 | 0.608 | 0.590 | 0.650 | 0.589 | 0.653 | 0.633 | 0.609 | 0.644 |
| -3 | 1 | 0 | 0 | 7 | 3 | | | | | | |

*Notes:* Logit regression, robust standard errors between brackets. The sample includes all granted U.S. patents filed between 1980 and 2010 which are also filed at the EPO and the JPO. The sample is further restricted to U.S. patents which are also granted by both EPO and JPO (labelled *granted*), and U.S. patents which are rejected by both EPO and JPO (labelled *rejected*). Granted patents are one-to-one matched to rejected patents based on filing years at USPTO, EPO, and JPO, and based on stemmed keywords in the titles, abstracts and claims of the patent ($n$=1,302,956 patents; 651,478 *granted* patents, 651,478 text-matched *rejected* patents). All variables except *1-backward_cosine, forward/backward_cosine, originality*, and *generality* are log transformed after adding 1. Control variables include number of *words, backward_citations, classes, subclasses*, and *primary_class, USPTO filing_year, EPO filing_year, and JPO filing_year* fixed effects. ll is log likelihood, precision represents the fraction of predicted *granted* patents that are correctly classified, recall is the fraction of real *granted* patents that are correctly classified, AUC is the area under the ROC-curve, marginal effect displays the average marginal effect (in %) of the particular measure increasing with one standard deviation.

number of later patents which reuse the same combinations provides an impact measure which outperforms all other measures, including the traditional impact measures based on forward citations.

Despite the benefits, the use of text also has several important limitations. Patents are written by inventors and attorneys in a way to increase the apparent novelty and chance of grant, and to maximize the scope of claims. Although inventors are required to disclose their invention in full, clear, concise, and exact terms, they might strategically use unclear writing and invent new jargon, which might bias the text-based measures.

Moreover, although stemming accounts for different spellings of the same word, it does not correct all spelling errors and does not account for synonyms (different words have the same meaning) and homonyms (same word has different meanings). Furthermore, a new technology might get a certain label after the pioneering patent is filed, or the patent might describe the technical content without using this label. In such cases, we might not identify the first patent introducing the technology. For example, Bluetooth – the short-range radio technology standard – got its name during a meeting of the three market leaders Intel, Ericsson, and Nokia in 1996. The first patent (US6590928: "Frequency hopping piconets in an uncoordinated wireless multi-user system"), invented by Jaap Haartsen who was inducted in the National Inventor Hall of Fame in 2015, does not mention the keyword Bluetooth. Nevertheless, because a patent contains on average not one but 61 unique stemmed keywords, the new technology and the corresponding patent might still be identified. The first Bluetooth patent for instance introduced amongst others the following new stemmed keyword combinations: "communic piconet" (reused by 375 patents) and "radio piconet" (reused by 109 patents). A piconet is an ad hoc network that connects wireless devices using Bluetooth technology.

To overcome the potential bias introduced by spelling errors, synonyms, homonyms, and the use of certain labels, one might consider the use of lemmatization, topic modelling, or word vectors (e.g. word2vec or Glove). The first patent(s) related to a topic or a word vector could be considered as introducing a new technology. But, to create topics or word vectors, it is necessary to train a model typically over all available

patents to learn relationships among words. We are reluctant to use dictionaries, topic modelling or word vectors because using information from later patents to assess novelty at the time of filing might cause spurious correlation between novelty and impact. To assess the novelty of a patent, one arguably should exclusively rely on information available at that point in time. Stemming is a more independent process that does not rely on learning word meaning from large contexts and has the advantage of being much faster to apply. Moreover, n-grams can also be interpreted as short topics, because they capture words within a limited and size-fixed context without exploring further word relations. A key benefit of the new measures proposed in this paper is that they can disentangle the two stages of inventive success, i.e. the creation of new ideas versus the eventual impact or diffusion of these ideas. This opens up opportunities to separately study the emergence of new technologies and the diffusion and exploitation of those technologies. What drives the creation of new technologies and which factors determine the exploitation and diffusion of new ideas?

The metrics based on cosine similarity have two additional limitations. First, they only provide a number between zero and one and do not identify the keywords related to new technology itself. Second, *forward_cosine* does not necessarily relate to the impact of the technical novelty introduced by the focal patent. Patents with a low novelty might have a high forward cosine because they are in a dense area with many similar patents.

Besides limitations related to the use of patent text, another restriction is that we only included granted U.S. patents and not patents granted by different patent offices nor scientific prior art. A certain discovery might first be disclosed in a scientific publication or in a non-U.S. patent, and only later (or never) appear in a U.S. patent document.

Finally, our text-based measures reflect the technical novelty and impact of a patent, and not its economic value in monetary terms. Other studies developed and validated indicators for the economic value of patents (Harhoff et al., 1999; Hall et al., 2005; Kogan et al., 2017; Moser et al., 2018).

We provide open access to code, data, and new measures for all U.S. utility patents granted up to May 2018 (see https://zenodo.org/record/3515985, DOI: 10.5281/zenodo.3515985). First, we provide for each patent a list of processed and cleaned keywords extracted from the title, abstract, and claims. This data can be used to measure and map the similarity between patents, inventors, firms, or geographical regions in technology space (e.g. Arts et al., 2018). The data can also be used to trace follow-on invention and diffusion of certain technologies or patents (e.g. de Rassenfosse et al., 2020), to measure knowledge spillovers from R&D (e.g. Myers and Lanahan, 2020), or spillovers between science and technology (e.g. Iaria et al., 2018). One fruitful opportunity is to match the dataset with other text documents such as scientific publications (e.g. Iaria et al., 2018), funding opportunity announcements (e.g. Myers and Lanahan, 2020), specifications of technical standards (Brachtendorf et al., 2020), or product descriptions (e.g. Argente et al., 2020).

Next, we disclose separate files for all new keywords, bigrams, trigrams, and new word combinations, together with the number of the patent introducing them for the first time (based on filing date), and the total number of patents using these new keywords, bigrams, trigrams, and new word combinations.

Finally, we calculate and disclose for every patent all the text-based measures discussed in the paper. Appendix A describes the different open access data files in greater detail. We hope our code, data, and new measures open up opportunities for future research.

## CRediT authorship contribution statement

**Sam Arts:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Resources, Software, Writing - original draft, Writing - review & editing. **Jianan Hou:** Data curation, Formal analysis, Investigation, Software, Validation, Writing - original draft, Writing - review & editing. **Juan Carlos Gomez:** Data curation, Methodology, Software, Resources.

## Declaration of Competing Interest

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

## Acknowledgements

The authors thank participants in presentations at the European Patent Office, Organisation for Economic Co-operation and Development (OECD), 2019 DRUID conference, 2019 PatentSemTech workshop, 2018 Summer school on data & algorithms for science, technology and innovation studies, 17th international conference on scientometrics & informetrics, 2019 European Meeting on Applied Evolutionary Economics, Friedrich Schiller University Jena, and KU Leuven. This work was supported by KU Leuven grant IMP/16/002.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.respol.2020.104144.

## References

Alcacer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. Rev. Econ. Stat. 88 (4), 774–779.
Allan, J., Wade, C., Bolivar, A., 2003. Retrieval and novelty detection at the sentence level. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 314–321.
Argente, D., Baslandze, S., Hanley, D., Moreira, S., 2020. Patents to products: product innovation and firm dynamics. In: FRB Atlanta Working Paper No. 2020-4. Available at SSRNhttps://ssrn.com/abstract=3587377 or. http://dx.doi.org/10.29338/wp2020-04.
Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manage. J. 39 (1), 62–84.
Arts, S., Fleming, L., 2018. Paradise of novelty-or loss of human capital? Exploring new fields and inventive output. Organ. Sci. 29 (6), 989–1236.
Arts, S., Veugelers, R., 2020. Taste for science, academic boundary spanning, and inventive performance of scientists and engineers in industry. Ind. Corp. Change 29 (4), 917–933.
Ashtor, J.H., 2019. Investigating cohort similarity as an ex ante alternative to patent forward citations. J. Empir. Legal Stud. 16 (4), 848–880.
Balsmeier, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., Li, G., Luck, S., O'Reagan, D., Yeh, B., Zang, G., Fleming, L., 2018. Machine learning and natural language processing on the patent corpus: data, tools, and new measures. J. Econ. Manag. Strateg. 27 (3), 535–553.
Bergeaud, A., Potiron, Y., Raimbault, J., 2017. Classifying patents based on their semantic content. PLoS ONE 12 (4), e0176310.
Brachtendorf, L., Gaessler, F., Harhoff, D., 2020. Truly standard-essential patents? A semantics-based analysis. In: CEPR discussion paper DP14726.
Carpenter, M.P., Narin, F., Woolf, P., 1981. Citation rates to technologically important patents. World Pat. Inf. 3 (4), 160–163.
Dahlin, K.B., Behrens, D.M., 2005. When is an invention really radical?: Defining and measuring technological radicalness. Res. Policy 34 (5), 717–737.
deGrazia, C.A., Frumkin, J.P., Pairolero, N.A., 2020. Embracing invention similarity for the measurement of vertically overlapping claims. Econ. Innov. New Technol. 29 (2), 113–146.
de Rassenfosse, G., Pellegrino, G., & Raiteri, E. (2020). Do patents enable disclosure? Evidence from the invention secrecy act. Available at SSRN: https://ssrn.com/abstract=3561896 or http://dx.doi.org/10.2139/ssrn.3561896.
Fleming, L., Mingo, S., Chen, D., 2007. Collaborative brokerage, generative creativity, and creative success. Adm. Sci. Q. 52 (3), 443–475.
Frakes, M.D., Wasserman, M.F., 2017. Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. Rev. Econ. Stat. 99 (3), 550–563.
Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. Manage. Sci. 63 (3), 791–817.
Gerken, J.M., Moehrle, M.G., 2012. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics 91 (3), 645–670.
Griliches, Z., 1990. Patent statistics as economic indicators: a survey. J. Econ. Lit. 28 (4), 1661–1707.

Hain, D., Jurowetzki, R., Buchmann, T., & Wolf, P. (2020). Text-based technological signatures and similarities: how to create them and what to do with them. arXiv preprint arXiv:2003.12303.
Hall, B.H., Jaffe, A.B., Trajtenberg, M., 2001. The NBER patent citations data file: lessons insights and methodological tools. In: NBER Working Paper.
Hall, H.B., Jaffe, A., Trajtenberg, M., 2005. Market value and patent citations. Rand J. Econ. 36 (1), 16–38.
Harhoff, D., Narin, F., Scherer, F.M., Vopel, K., 1999. Citation frequency and the value of patented inventions. Rev. Econ. Stat. 81 (3), 511–515.
Hasan, M.A., Spangler, W.S., Griffin, T., Alba, A., 2009. Coa: finding novel patents through text analysis. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1175–1184.
He, Y., Luo, J., 2017a. The novelty 'sweet spot' of invention. Des. Sci. 3, e21. https://doi.org/10.1017/dsj.2017.23.
He, Y., Luo, J., 2017b. Novelty, conventionality, and value of invention. In: Gero, J. (Ed.), Design Computing and Cognition '16. Springer, Cham. https://doi.org/10.1007/978-3-319-44989-0_2.
Henderson, R., Jaffe, A., Trajtenberg, M., 1998. Universities as a source of commercial technology. Rev. Econ. Stat. 80 (1), 119–127.
Hirshleifer, D., Hsu, P.H., Li, D., 2017. Innovative originality, profitability, and stock returns. Rev. Financ. Stud. 31 (7), 2553–2605.
Hsu, D.H., Lim, K., 2013. Knowledge brokering and organizational innovation: Founder imprinting effects. Organ. Sci. 25 (4), 1134–1153.
Iaria, A., Schwarz, C., Waldinger, F., 2018. Frontier knowledge and scientific production: evidence from the collapse of international science. Q. J. Econ. 133 (2), 927–991.
Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.
Jaffe, A.B., Fogarty, M.S., Banks, B.A., 1998. Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation. J. Ind. Econ. 46 (2), 183–205.
Jaffe, A.B., Lerner, J., 2004. Innovation and its Discontents: how our Broken Patent System is Endangering Innovation and Progress, and what to do about it. Princeton University Press, Princeton, NJ.
Jung, H.J., Lee, J.J., 2016. The quest for originality: a new typology of knowledge search and breakthrough inventions. Acad. Manage. J. 59 (5), 1725–1753.
Kaplan, S., Vakili, K., 2015. The double-edged sword of recombination in breakthrough innovation. Strateg. Manage. J. 36 (10), 1435–1457.
Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2018. Measuring Technological Innovation over the Long Run (No. w25266). National Bureau of Economic Research.
King, J.L., 2003. Patent examination procedures and patent quality. In: The National Research Council (Ed.), Patents in the knowledge-based economy. The National Academic Press.
Kneeland, M.K., Schilling, M.A., Aharonson, B.S., 2020. Exploring uncharted territory: knowledge search processes in the origination of outlier innovation. Organ. Sci. 31 (3), 535–557.
Kogan, L., Papanikolaou, D., Seru, A., Stoffman, N., 2017. Technological innovation, resource allocation, and growth. Q. J. Econ. 132 (2), 665–712.
Krieger, J.L., Li, D., Papanikolaou, D., 2018. Missing Novelty in Drug Development (No. w24595). National Bureau of Economic Research.
Kuhn, J., Younge, K., Marco, A., 2020. Patent citations reexamined. Rand J. Econ. 51 (1), 109–132.
Lampe, R., 2012. Strategic citation. Rev. Econ. Stat. 94 (1), 320–333.
Lei, Z., Wright, B.D., 2017. Why weak patents? Testing the examiner ignorance hypothesis. J. Public Econ. 148, 43–56.
Lemley, M.A., Shapiro, C., 2005. Probabilistic patents. J. Econ. Perspect. 19 (2), 75–98.
Lemley, M., Sampat, B., 2010. Examiner characteristics and the patent grant rate. Rev. Econ. Stat.
Li, X., Croft, W.B., 2008. An information-pattern-based approach to novelty detection. Inf. Process. Manage. 44 (3), 1159.
Marco, A.C., Sarnoff, J.D., deGrazia, C.A., 2016. Patent claims and patent scope. In: USPTO Economic Working Paper No. 2016-04.
Myers, K., & Lanahan, L. (2020). Research subsidy spillovers, two ways (June 9, 2020). Available at SSRN: https://ssrn.com/abstract=3550479 or http://dx.doi.org/10.2139/ssrn.3550479.
Moser, P., Ohmstedt, J., Rhode, P.W., 2018. Patent citations—an analysis of quality differences and citing practices in hybrid corn. Manage. Sci. 64 (4), 1926–1940.
Motohashi, K., Zhu, C., 2020. Technological competitiveness of China's internet platforms: comparison of Google and Baidu using patent text information. In: Research Institute of Economy, Trade and Industry (RIETI) Discussion Paper Series 20-E-045.
Nagaoka, S., Motohashi, K., Goto, A., 2010. Patent statistics as an innovation indicator. In: Handbook of the Economics of Innovation, Vol. 2, pp. 1083–1127. North-Holland.
Nanda, R., Younge, K., Fleming, L., 2015. Innovation and entrepreneurship in renewable energy. In: Jaffe, Adam B., Jones, Benjamin F. (Eds.), The Changing Frontier: Rethinking Science and Innovation Policy, pp. 199–232.
OECD. (2019). Triadic patent families database, version February 2019.
Packalen, M., Bhattacharya, J., 2015. New Ideas in Invention (No. w20922). National Bureau of Economic Research.
Picard, P.M., Potterie, B.V.P., 2013. Patent office governance and patent examination quality. J. Public Econ. 104, 14–25.
Reitzig, M., 2004. Improving patent valuations for management purposes—validating new indicators by analyzing application rationales. Res. policy 33 (6-7), 939–957.
Righi, C., Simcoe, T., 2019. Patent examiner specialization. Res. Policy 48 (1), 137–148.
Romer, P, 1990. Endogenous technological change. J. Polit. Econ. 98 (5), S71–S102.
Scherer, F.M., Harhoff, D., 2000. Technology policy for a world of skew-distributed outcomes. Res. Policy 29 (4-5), 559–566.
Teodoridis, F., Lu, J., & Furman, J. L. (2020). Measuring the direction of innovation: frontier tools in unassisted machine learning. Available at SSRN 3596233.
Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: a reassessment. Am. Econ. Rev. 450–460.
Trajtenberg, M., 1990. A penny for your quotes: Patent citations and the value of innovations. Rand J. Econ. 21 (1), 172–187.
Trajtenberg, M., Henderson, R., Jaffe, A., 1997. University versus corporate patents: a window on the basicness of invention. Econ. Innov. New Technol. 5 (1), 19–50.
Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342 (6157), 468. -47.
Verhoeven, D., Bakker, J., Veugelers, R., 2016. Measuring technological novelty with patent-based indicators. Res. Policy 45 (3), 707–723.
Watzinger, M., Schnitzer, M., 2019. Standing on the shoulders of science. In: CEPR Discussion Paper No. DP13766.
Weitzman, M.L., 1998. Recombinant growth. Q. J. Econ. 113 (2), 331–360.
Younge, K.A., & Kuhn, J.M. (2019). First movers and follow-on invention: evidence from a vector space model of invention. Available at SSRN 3354530.