Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text

Sam Arts*
Department of management, strategy and innovation
Faculty of economics and business
KU Leuven
sam.arts@kuleuven.be

Nicola Melluso
Department of management, strategy and innovation
Faculty of economics and business
KU Leuven
nicola.melluso@kuleuven.be

Reinhilde Veugelers
Department of management, strategy and innovation
Faculty of economics and business
KU Leuven
reinhilde.veugelers@kuleuven.be

ABSTRACT

New scientific ideas drive progress, yet measuring scientific novelty remains challenging. We use natural language processing to detect the origin and impact of new ideas in scientific publications. To validate our methods, we analyze Nobel Prize-winning papers, which likely pioneered impactful new ideas, and literature review papers, which typically consolidate existing knowledge. We also show that novel papers have more intellectual neighbors published after them, indicating they are ahead of their intellectual peers. Finally, papers introducing new ideas, particularly those with greater follow-on reuse, attract more citations.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Keywords: natural language processing; science; novelty; impact; breakthrough; Nobel

JEL codes: O30, O31, O32, O33, I23

* Corresponding author

The authors gratefully acknowledge Pierre Azoulay (the editor) and the anonymous reviewers for their insightful and constructive feedback, which greatly enhanced the quality of this work. We also thank participants at the 2023 NBER-SI-SSF workshop, the WOEPSR23 Conference, DRUID23, REGIS Summer School 2023, and the 2024 OpenAlex Virtual User Conference for their valuable input, especially Stefano Baruffaldi, Chiara Franzoni, Dietmar Harhoff, Vincent Thorge Holst, Jianan Hou, Erin Leahey, Rodrigo Ito, and Fabio Montobbio for their thoughtful comments. Financial support from KU Leuven Grant 3H200208 is acknowledged.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

1. Introduction

New scientific ideas are fundamental to driving progress in science, technology, and economic prosperity (Bush, 1945; Mokyr, 2002). The significance of new scientific discoveries such as the transistor, magnetic resonance imaging, polymerase chain reaction, or carbon nanotubes cannot be overstated. However, despite their critical role, identifying and measuring novel scientific ideas—and tracing their diffusion and impact—remains challenging.

To measure new scientific ideas and the novelty of scientific papers, prior work has traditionally relied on their patterns of citations to prior work (Uzzi et al., 2013; Wang, Veugelers, & Stephan, 2017; Wu, Wang, & Evans, 2019). Yet, this approach has important limitations. Citations capture prior art but not the scientific content and contribution of the paper itself. Thus, citation-based methods may fail to accurately identify new scientific ideas and measure the novelty in the content of a paper (Fontana et al., 2020). Moreover, citations may not always reflect intellectual influence or impact of novel ideas. A survey of 9,000 academics across 15 fields found that more than half of cited papers had only minor or very minor intellectual influence (Teplitskiy et al., 2022).

In this paper, we move beyond citations and build on Kuhn's (1962) argument that scientific ideas are embedded in the text of scientific literature, making shifts in language critical for identifying new scientific ideas. We apply Natural Language Processing (NLP) techniques to harness the text of scientific publications, identifying both the origin and impact of new scientific ideas. By now, a substantial body of research, scattered across various disciplines from physics to sociology, has begun exploring NLP for detecting novel scientific ideas and their impact.[1] Yet, several shortcomings remain. First, existing work has yet to develop text metrics that can simultaneously identify new scientific ideas and measure a paper's novelty at the time of publication, as well as trace the diffusion and impact of these ideas over time.[2] Differentiating between novelty at publication and later adoption allows us to separate the emergence of

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

new ideas from their eventual success, facilitating an analysis of the factors influencing the success or failure of novel ideas. To accurately assess a paper's novelty, it is important to consider only prior work available at the time of publication, excluding any subsequent publications or data, to avoid success bias when identifying novel ideas. Second, existing studies lack comprehensive large-scale validation of text metrics. Without evidence that these metrics effectively capture new scientific ideas and their influence on future research, their utility remains questionable. Third, there is little consensus on which text metrics to adopt and whether they offer an improvement over traditional citation-based measures. The proliferation of proposed metrics complicates comparisons across studies, hindering scientific progress. Lastly, prior research has not provided open access to the underlying code, data, and metrics, which limits broader adoption. Given the scale of processing required for the entire corpus of scientific work, open access to these resources could facilitate the broader use of text metrics and support community efforts to further improve these measures. This paper aims to address these gaps in the literature.

We use the titles and abstracts of all scientific publications covered in the January 2024 snapshot of OpenAlex, which covers the largest corpus of scientific work published in the entire history from 1666 until 2023, and has the advantage of being open access (Priem, Piwowar, & Orr, 2022; Lin et al., 2023). To detect new scientific ideas and measure a paper's novelty at publication, we identify words, noun phrases, and novel combinations of words or noun phrases that appear for the first time. For example, we track the first papers introducing the word "*perceptron*," the noun phrase "*atomic force microscope*," or the word combination "*dna*" and "*microarray*." Alternatively, instead of distilling individual keywords, noun phrases, or combinations of these, we measure a paper's novelty based on the similarity of its entire text to all prior papers. To do this, we use SPECTER, a pre-trained document-level embedding of scientific papers, which has the advantage of accounting for synonyms, polysemy,

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

and the context of words (Cohan et al., 2020).[3] To measure the impact or influence of new scientific ideas, we count the number of subsequent papers reusing new words, noun phrases, or combinations of either words or noun phrases. For instance, 19,495 papers reuse "*perceptron*," 17,322 reuse "*atomic force microscope*," and 25,916 reuse the combination of "*dna*" and "*microarray*."

To validate our text metrics and their improvement over traditional citation-based measures, we analyze Nobel Prize papers, which likely introduced impactful new ideas. The Nobel Prize also provides a means to validate whether the novel ideas identified by our text metrics across the entire corpus of scientific literature align with the contributions highlighted in the official Nobel Prize documentation. We also examine literature review papers, which typically consolidate existing knowledge rather than introduce novel ideas. Additionally, we show that more novel papers tend to have a higher proportion of their closest intellectual neighbors published after them, indicating they are ahead of their peers, and are more likely to use novelty-signaling language such as "discover" or "innovative." Finally, we demonstrate that papers introducing novel ideas—particularly those with substantial follow-on reuse—attract more citations and have a higher likelihood of becoming highly cited. These studies support the use of NLP in identifying new scientific ideas, measuring the novelty of papers at the time of publication, and assessing the impact of these ideas on subsequent scientific work. Furthermore, the results illustrate the improvement of text-based metrics over traditional citation-based measures. We provide open access to all data (https://zenodo.org/records/13869486) and code (https://github.com/nicolamelluso/science-novelty).

2. Identifying the Origin and Impact of New Scientific Ideas

2.1 Data collection

We collect all scientific publications from OpenAlex, concatenate the title and abstract of each paper, and process the text using standard NLP techniques: including tokenization, POS-

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

tagging, dependency parsing, chunking, lemmatization, cleaning, baseline removal, and vectorization (see Appendix C for details).[4] We use title, abstract and full-text of publications published between 1666 and 1900 to construct a baseline dictionary and restrict the analysis to papers published between 1901 and 2023 (n=75,295,921).

2.2 Text metrics for novelty

First, we calculate *New Word* as the number of unique unigrams of a paper that appear for the first time in history. Hence, we identify the first paper introducing words such as "*apoptosis*" or "*photon*." Second, we calculate *New Phrase* as the number of unique noun phrases of a paper that appear for the first time in history.[5] Hence, we identify the first paper introducing noun phrases such as "*optical coherence tomography*" or "*vascular endothelial growth factor*." Third, we compute *New Word Combination* as the number of unique pairwise combinations of words that appear for the first time, regardless of the location or order of the words. For instance, we identify the first paper using the combination of words such as "*angiogenesis"* and "*therapeutic*" or "*carbon"* and *"nanotube*." Fourth, we compute *New Phrase Combination* as the number of unique pairwise combinations of noun phrases that appear for the first time, regardless of the location or order of the phrases. For instance, we identify the first paper using the combination of phrases such as "*enzyme*" and "*superoxide dismutase*" or "*atom transfer*" and "*radical polymerization*." We exclude words, phrases, and combinations that appear only once in the entire corpus. Note that papers introducing new words or phrases introduce new combinations of words or phrases by construction. Finally, we generate a SPECTER document-embedding vector for each paper and calculate *Semantic Distance* as one minus the maximum cosine similarity between the focal paper and all prior papers from the past 5 years. This metric captures how distinct the focal paper is from the most similar prior work. For example, the seminal paper by Kirkpatrick, Gelatt, and Vecchi (1983) on simulated annealing scores in the top 1% for Semantic Distance.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

2.3 Text metrics for impact of new scientific ideas

To assess the influence of new scientific ideas on subsequent research, we analyze how often these ideas appear in later publications. First, we calculate *New Word Reuse* as the number of new unigrams introduced by the focal paper, weighted by the number of subsequent papers that reuse these unigrams. For instance, 495,985 publications reuse "*apoptosis*" and 226,688 reuse "*photon*." For paper p, $New\ Word\ Reuse_p = \sum_{i=1}^{n}(1 + u_i)$ with n representing the number of new unigrams introduced by paper p and u_i equal to the number of future papers which reuse the new unigram i. While *New Word* captures a paper's novelty at publication (ex ante), *New Word Reuse* measures the influence of the new scientific ideas on later research (ex post). Second, we compute *New Phrase Reuse* as the number of new noun phrases introduced by the focal paper, weighted by the number of later papers reusing those noun phrases. For example, "*optical coherence tomography*" is reused in 42,324 papers, and "*vascular endothelial growth factor*" in 55,897. Third, we compute *New Word Combination Reuse*, which measures the number of new word combinations weighted by the number of later papers incorporating those combinations. For instance, 28,907 papers reuse the combination "*angiogenesis*" and "*therapeutic*" and 149,120 reuse the combination "*carbon*" and "*nanotube*." Finally, we compute *New Phrase Combination Reuse* as the number of new noun phrase combinations weighted by the number of subsequent papers reusing those combinations. For instance, 6,518 papers reuse the combination of "*enzyme*" and "*superoxide dismutase*" and 6,144 papers reuse the combination of "*atom transfer*" and "*radical polymerization*."

As shown in Appendix D, papers that reuse new words or noun phrases are approximately 53 and 45 times more likely, respectively, to cite the pioneering paper that introduced these words or phrases, compared to matched control papers from the same journal, year, and subfield that do not reuse these words or phrases. A similar pattern holds for the reuse of new word or phrase combinations. These findings suggest that the reuse of new ideas correlates

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

with citations to the paper pioneering these ideas, confirming their intellectual influence. However, the data also reveal that only a small minority of reusing papers cite the pioneering paper, and as illustrated in Figure D.1, this likelihood declines significantly as the time between the pioneering and reusing paper increases. This suggests that researchers tend to cite more recent, closely related work rather than the foundational studies that originally introduced the insights upon which they build. These results highlight the complementary value of text-based metrics in capturing the diffusion and influence of new ideas, particularly in light of prior research indicating that citations alone provide a noisy and incomplete measure of intellectual influence (Cozzens, 1989; Teplitskiy et al., 2022).

2.4 Traditional metrics for novelty

Prior research traditionally relies on citations to measure the novelty of a paper. First, Uzzi et al. (2013) define novelty as an atypical combination of prior knowledge, where the observed frequency of any pair of cited journals is compared to the frequency of that pair occurring by chance. This comparison results in a normalized z-score that measures the (a)typicality of each pair of cited journals. The *Uzzi* score for a focal paper is determined as the 10^{th} percentile of the z-scores for all pairs of journals cited by the focal paper, i.e. focusing on the most atypical combinations introduced by the focal paper. Lower values of *Uzzi* indicate more atypical or novel papers.

Second, Wang, Veugelers and Stephan (2017) define a focal paper's novelty as the sum of the distances between first-time combinations of cited journals. Following their method, we first identify all pairs of journals cited together for the first time. We then calculate the distance for each new pair based on their co-citation frequencies and, finally, compute the *Wang* metric as the sum of these distances for all new combinations.

Finally, Funk and Owen-Smith (2017) introduce a metric (*CD*) that characterizes whether a paper is disruptive or consolidating. A paper is considered disruptive if it is cited

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

without its predecessors being cited—those papers that it references. Conversely, a paper is seen as consolidating if it is frequently cited alongside its predecessors. *CD* cannot measure a paper's novelty at the time of publication (ex-ante) since it relies on subsequent citations (ex-post). Unlike *Uzzi* and *Wang*, *CD* is not a direct measure of a paper's novelty but rather a measure of the nature of its impact. Nevertheless, we include this measure in our analysis due to its growing use in the science of science community (e.g., Wu, Wang, & Evans, 2019; Park, Leahey, & Funk, 2023). In our analysis, we use the *Uzzi* atypicality and *CD* scores provided by SciSciNet (Lin et al., 2023).[6]

In this paper, we primarily use the novelty metrics (both citation-based and text-based) as continuous or count variables. Given the skewness of these measures, we also performed robustness checks by transforming them into binary indicators, classifying papers as novel if they rank in the top 5% (or top 1%) of a given metric within the same scientific subfield and year of publication. Although not reported, these checks confirmed that our results are robust across all validation tests, regardless of whether we use the raw measures or their binary transformations.

3. Descriptive Statistics

Table A.1 presents summary statistics at the level of new words, noun phrases, and their pairwise combinations. Between 1901 and 2023, approximately 6 million new words, 27 million new noun phrases, 1 billion new word combinations, and about 0.8 billion new noun phrase combinations were introduced. As expected, most new words, phrases, and combinations are reused by a small number of papers, while a select few are widely adopted. Among the most influential are the word "*positron*," first introduced in 1933 and reused by 96,988 publications; the noun phrase "*x-ray diffraction*," introduced in 1914 and reused by 344,357 papers; the combination of words "*electron*" and "*microscopy*," appearing in 1934 and reused by 674,072 papers; and the combination of noun phrases "*catalase*" and "*superoxide dismutase*,"

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

originating in 1970 and reused in 35,423 papers. Table A.2 presents the top 10 most frequently reused new noun phrases introduced in each decade.

Table 1 presents summary statistics at the paper level, with Panel A showing ex ante metrics, which measure novelty at the time of publication, and Panel B showing ex post metrics, which assess impact after publication. Only a subset of papers introduce new scientific ideas. Approximately 7% of papers introduce a new word, 25% introduce a new noun phrase, 61% introduce a new word combination, and 65% introduce a new noun phrase combination. The average (median) paper introduces 0.08 (0) new words, 0.36 (0) new noun phrases, 13.22 (2) new word combinations, and 10.98 (2) new noun phrase combinations. Although the number of new word and noun phrase combinations may appear substantial, it is important to note that the average paper contains 816 unique word pairs and 157 unique noun phrase pairs. Consequently, the average (median) share of new word combinations per paper is only 1.9% (0.4%), and the average share of new noun phrase combinations is 9.0% (4.5%). As expected, the summary statistics reveal a significant skew across all text-based novelty metrics, and underscore the need to control for text length when assessing a paper's novelty.

'Table 1'

Figure 1 illustrates the average number of new phrases introduced by papers across all fields of study from 1901 to 2023, revealing significant variation in scientific novelty both across fields and over time.[7] For example, fields like Physics and Astronomy and Material Science exhibit higher rates of scientific novelty compared to others. The time trends also reflect the historical evolution of scientific fields, such as the surge in groundbreaking discoveries in Biochemistry, Genetics, and Molecular Biology after 1950. As shown in Figure A.1 , this substantial heterogeneity in scientific novelty persists even at the more granular subfield level.

'Figure 1'

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

The novelty of scientific papers varies both across subfields and within the same subfield over time and among papers published in the same subfield and year. A variance decomposition model using *New Phrase* as a novelty measure reveals that differences between subfields explain approximately 2% of the total variance, while variation within subfields over time accounts for 5%. The remaining 93% is attributed to differences among papers within the same subfield and year. Certain subfields and time periods, such as Molecular Biology from 1974 to 1976, stand out for particularly high rates of novel scientific ideas.

The novelty of papers also varies significantly across journals. Differences between journals account for about 10% of the variance, variation within journals over time contributes 7%, and differences among papers within the same journal and year explain the remaining 83%. Journals like Cell and the Journal of Experimental Medicine consistently publish a notable share of papers introducing new scientific ideas. Interestingly, the relatively modest contribution of temporal variation is somewhat surprising given previous findings on the declining novelty (or disruptiveness) of papers over time (Bloom et al., 2020; Park, Leahey, & Funk, 2023).

As shown in Table A.3 of the Appendix, there is a strong positive correlation among text-based novelty metrics, indicating that novel ideas in a paper are often captured by multiple metrics simultaneously. However, average correlations between text-based and traditional citation-based novelty metrics are low, suggesting that text metrics assess novelty in a fundamentally different way than citation-based measures.

Finally, we examine the paper-level summary statistics for metrics that capture the impact of new scientific ideas. As shown in Panel B of Table 1, the average (median) value for *New Word Reuse* is 3.4 (0), for *New Phrase Reuse* 6.6 (0), for *New Word Combination Reuse* 197.8 (6), and for *New Phrase Combination Reuse* 67.8 (6). These descriptive statistics again

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

highlight the significant skew in the reuse of new scientific ideas, with only a small minority of novel papers having a substantial impact.

4. Nobel Prize Papers

Nobel Prize papers are recognized for pioneering new scientific ideas, particularly those with a profound impact on scientific progress. As such, they provide a benchmark for validating the effectiveness of our text metrics in identifying novel scientific ideas and their influence on subsequent scientific work. Using official documentation of laureates' contributions, we evaluate whether the novel scientific ideas identified by our metrics in Nobel Prize papers align with those described in the official documentation. Additionally, we perform a case-control study to evaluate the effectiveness of text-based and citation-based metrics in distinguishing Nobel Prize papers from matched control papers, based on the assumption that Nobel Prize papers are more likely to introduce new scientific ideas, particularly those with a significant impact on subsequent research.

4.1 Data

We collect papers linked to Nobel prizes in Chemistry, Physics and Physiology or Medicine from Li et al. (2019). Each Nobel prize paper is randomly matched to one control paper not linked to any Nobel prize but published in the same journal, year and subfield.[8] Our sample includes 584 Nobel Prize papers published between 1902 and 2007, linked to 234 Nobel Prizes awarded between 1908 and 2016. Among the 234 Nobel Prizes, at least one corresponding paper introduces a new word for 43%, a new phrase for 75%, a new word combination for 79%, and a new phrase combination for 85%.

For each Nobel Prize, we collected official online documentation detailing the recognized contributions, including the summary page, press release, and Nobel Lecture(s). Nobel Lectures are authored by the laureates on topics relevant to the work for which the prize was

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

awarded (Nobel Foundation, 2024).[9] The texts were combined, preprocessed similarly to the papers, and references to the titles of Nobel Prize papers were removed (see Appendix C). Reassuringly, 88% of new words, 92% of new noun phrases, 90% of new word combinations, and 91% of new noun phrase combinations introduced by Nobel Prize papers are also mentioned in the corresponding official documentation. This demonstrates that our text metrics are effective, though not perfect, in identifying new scientific ideas across the entire body of scientific literature.

'Table 2'

Table 2 presents examples of Nobel Prizes, highlighting the prize motivation, a corresponding paper, and a new scientific idea identified by our text metrics that aligns with the Nobel Prize motivation. For instance, William Shockley, John Bardeen, and Walter Brattain shared the 1956 Nobel Prize in Physics "for their research on semiconductors and the discovery of the transistor effect." One of their corresponding papers, published in Physical Review in 1948, was the first to introduce the term "*transistor*," which has since been reused by 128,821 subsequent papers. In 1997, Stanley Prusiner was awarded the Nobel Prize in Medicine "for his discovery of prions–a new biological principle of infection." His corresponding paper, published in Science in 1982, introduced the combination of the words "*prion*" and "*protein*," which has been reused by 13,779 papers. The 1993 Nobel Prize in Chemistry was awarded to Kary Mullis "for his invention of the polymerase chain reaction method." His corresponding paper, published in 1986 in Cold Spring Harbor Symposia on Quantitative Biology, pioneered the noun phrase "*polymerase chain reaction*," which has been reused by 116,146 papers. Lastly, in 2011, Ralph M. Steinman received the Nobel Prize in Medicine "for his discovery of the dendritic cell and its role in adaptive immunity." His corresponding paper, published in 1973 in the Journal of Experimental Medicine, was the first to introduce

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

the combination of the phrases "*dendritic cell*" and "*macrophage*," which has since been re-used by 8,299 papers.

4.2 Descriptive results

Table A.4 presents descriptive statistics for Nobel Prize papers and matched control papers. Both text-based and traditional citation-based metrics effectively distinguish Nobel Prize papers from control papers, as indicated by the significant results of the Mann-Whitney test (p=0.000). The only exceptions are *Semantic Distance* and *Uzzi*. All text metrics that measure novelty at the time of publication (*New Word*, *New Phrase*, *New Word Combination*, *New Phrase Combination*) outperform traditional citation-based novelty measures (*Wang* and *Uzzi*). These findings provide evidence that text metrics are more effective at identifying new scientific ideas and measuring the novelty of papers at the time of publication. Among the text metrics, *New Phrase* performs best in identifying Nobel Prize papers.

As expected, and illustrated in Panel B, text metrics that measure the reuse of new scientific ideas after publication (*New Word Reuse*, *New Phrase Reuse*, *New Word Combination Reuse*, *New Phrase Combination Reuse*) also distinguish Nobel Prize papers from control papers, outperforming all metrics that capture novelty at the time of publication. Each of these impact-based text metrics also outperforms *CD*, the disruptiveness index. Of all metrics, *New Phrase Reuse* performs the best in identifying Nobel Prize papers. As Nobel Prize papers are renowned for pioneering impactful ideas, these findings demonstrate that text metrics tracking idea reuse effectively capture the influence and lasting impact of their novel contributions.

4.3 Regressions

Table 3 presents results from logit regressions with a binary indicator for Nobel Prize paper as the outcome variable. The models control for the number of unique words and phrases in the title and abstract of the paper, as well as whether the paper has an abstract available

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

(Letchford, Moat & Preis, 2015; Milojević, 2017). Additionally, we account for the number of cited papers and journals and include fixed effects for publication year and subfield. To evaluate the performance of different metrics in correctly classifying Nobel Prize and control papers, we calculate precision (proportion of correctly classified Nobel Prize papers), recall (proportion of actual Nobel Prize papers correctly identified), and the area under the curve (AUC), which ranges from 0.5 (no predictive power) to 1 (perfect classification). Average marginal effects quantify the increase in the likelihood of a paper being a Nobel Prize paper per one standard deviation increase in each metric.

‘Table 3’

The regression results in Table 3 align closely with the descriptive statistics in Table A.4. Panel A shows that all text metrics measuring novelty at the time of publication are significant at the 1% level and outperform traditional citation-based metrics, except *for Semantic Distance*, which is not significant at conventional levels. Binary versions of text metrics, such as *New Word (Binary)*, which identifies whether a paper introduced at least one new word, are statistically significant but generally perform worse and are omitted from the table for brevity. This indicates that Nobel Prize papers typically introduce multiple new words, noun phrases, or combinations, making count-based measures more effective at capturing novelty than binary indicators. Consistent with prior literature, traditional citation-based metrics (*Wang* and *Uzzi*) fail to distinguish Nobel Prize papers from controls (Fontana et al., 2020). Of all novelty metrics, *New Phrase* has the strongest discriminatory power.

In Model 8, which includes all text metrics, *New Word* becomes statistically insignificant, but this model achieves higher precision, recall, and predictive power compared to using *New Phrase* alone, the best-performing single metric. In Model 9, which incorporates both text and citation-based metrics, predictive power remains unchanged. These results highlight

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

the effectiveness of text metrics, particularly *New Phrase*, as indicators of a paper's novelty at the time of publication.

Panel B shows that text metrics capturing the reuse of new scientific ideas generally outperform those measuring novelty at the time of publication in predicting Nobel Prizes. This finding is reassuring, as it suggests that text metrics effectively capture the broader influence and enduring impact of novel contributions introduced by Nobel Prize papers. Among all metrics, *New Phrase Reuse* emerges as the strongest predictor.

While the text metrics demonstrate reasonable accuracy in predicting Nobel Prize papers, they are obviously far from perfect. On one hand, a notable portion of predicted Nobel Prize papers are actually control papers, i.e., false positives. However, these control papers may also introduce new scientific ideas but simply did not receive a Nobel Prize, making them not necessarily false positives for novelty. For instance, a paper by Federico Capasso and colleagues published in Science in 2000 (in the same year, journal, and subfield as John Hall's Nobel-winning work) introduced the phrase "*midinfrared quantum cascade laser*." On the other hand, some Nobel Prize papers are missed by the text metrics, i.e., false negatives. An example is the set of papers associated with the 2004 Nobel Prize in Chemistry awarded to Aaron Ciechanover, Avram Hershko, and Irwin Rose for their discovery of ubiquitin-mediated protein degradation. Additionally, while Nobel Prize-winning papers provide a useful validation benchmark, they have inherent limitations. The sample size is small, the focus is limited to certain fields, and novelty alone is not the sole criterion for awarding a Nobel Prize. These papers primarily reflect highly impactful new ideas rather than a complete representation of novel contributions in science.

5. Literature Review Papers

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

As a second validation, we collect a large sample of literature review papers (n=34,428) along with a matched control sample of original, non-review papers. Our assumption is that review papers primarily summarize existing scientific knowledge rather than introducing new scientific ideas (Wu, Wang, & Evans, 2019). In other words, we expect review papers to be less likely to pioneer new insights compared to control papers. As detailed in Appendix E, the data and results support this assumption and align closely with the Nobel Prize validation findings. Given that the vast majority of scientific papers are perhaps not very novel, it is unsurprising that the text metrics are less effective at distinguishing review papers from control papers than at differentiating Nobel Prize papers from controls.

6. Publication Timing of Intellectual Neighbors

As an additional validation, we check whether more novel papers have a greater proportion of their closest intellectual neighbors—defined as the most similar papers—published after them, compared to less novel papers. In other words, when a paper introduces a new scientific idea, related research is expected to be published after it, indicating that the paper is ahead of its intellectual peers. To test this, we identify all OpenAlex papers indexed in PubMed and published between 1901 and 2010 (n=11,542,812) and use the PubMed Related Citations Algorithm to find the five most similar articles for each paper based on title, abstract, and MeSH terms (Lin & Wilbur, 2007). As outlined in Appendix F, we find support for this using both text-based and citation-based novelty metrics, though the text-based metrics generally show a stronger effect on the proportion of intellectual neighbors published after the focal paper.

7. Language Denoting Novelty

Papers introducing new scientific ideas are more likely to use language explicitly signaling novelty, such as terms like “discover,” “introduce,” “novel,” or “innovative.” Following the approach of Leahey et al. (2023), we construct a binary indicator for papers that include any

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

of these novelty-signaling terms in their title or abstract and test the ability of all metrics to correctly classify such papers. This analysis is conducted on the full sample of papers published between 1901 and 2023 (n=75,295,921). As illustrated in Appendix G, while all metrics demonstrate some predictive power, text-based metrics generally outperform others in identifying papers that explicitly highlight their novelty through this type of language.

8. New Scientific Ideas Fuel Scientific Progress

The motivation to identify papers that pioneer novel scientific ideas stems from their potential to significantly advance scientific progress, traditionally measured by the citations these papers receive (e.g., Uzzi et al., 2013; Wang, Veugelers, & Stephan, 2017). There are persistent concerns about the science funding system's ability to adequately support novel research, potentially resulting in missed opportunities for scientific progress and breakthroughs (Azoulay, Graff Zivin, & Manso, 2011; Alberts et al., 2014; Franzoni, Stephan, & Veugelers, 2022). In Appendix H, we analyze the full sample of papers published between 1901 and 2010 (n=37,154,406) and show that papers introducing new scientific ideas at the time of publication tend to attract more citations over time and are more likely to become highly cited. Figure 2 illustrates how the likelihood of a paper being among the top 1% most-cited (above the 99th percentile in citations within the same subfield and year) varies across percentile ranges of each novelty metric. Notably, text-based metrics capturing a paper's novelty at publication generally predict top-cited papers more effectively than traditional novelty metrics (*Uzzi* and *Wang*). Reassuringly, text-based metrics capturing the reuse of new scientific ideas generally outperform those measuring novelty at the time of publication in predicting citation outcomes. This indicates that, despite our earlier finding that papers reusing new scientific ideas do not consistently cite the pioneering papers introducing these ideas, these metrics effectively capture the broader influence and impact of new scientific ideas on subsequent literature, as reflected in citation counts.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

'Figure 2'

9. Discussion and Conclusion

New scientific insights drive progress in science, technology, and the economy. However, identifying and measuring novel scientific ideas—and tracing their diffusion and impact—remains a persistent challenge. Citation-based metrics have traditionally served as proxies for novelty—or, relatedly, atypicality or disruptiveness—in scientific papers. Yet, citations primarily reflect connections to prior work rather than the intrinsic scientific content or contributions of the papers themselves. Moreover, while citations are widely accepted as a measure of impact, they often serve rhetorical or strategic purposes, such as lending authority or adhering to conventions, rather than signaling substantive intellectual influence (Cozzens, 1989; Teplitskiy et al., 2022). As a result, citation-based metrics often fall short in accurately identifying novel scientific ideas at the time of publication and in capturing their true intellectual impact on scientific progress (Fontana et al., 2020).

In this paper, we employ natural language processing techniques to analyze and vectorize scientific content, enabling the identification of novel ideas and the measurement of scientific novelty at the time of publication, as well as the assessment of their subsequent reuse in the literature. Our findings confirm the effectiveness of text-based metrics in identifying novelty and measuring its impact, demonstrating that these metrics outperform traditional citation-based approaches.

Interestingly, we find that papers reusing new scientific ideas do not consistently cite the original work that introduced the idea, particularly when the idea was established long ago. This highlights how text-based metrics offer a novel perspective on the diffusion and influence of new ideas, extending beyond citation-based measures.

However, text-based approaches are not without limitations (particularly those stemming from the quality of underlying publication data) which future researchers must carefully

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

consider. Appendix I provides a detailed overview of these limitations and offers suggestions for advancing this line of research. Despite these challenges, a key contribution of text-based metrics is their ability to capture the two main stages of scientific progress: the discovery of new ideas and their subsequent diffusion and use. This capability opens new avenues for studying not only the emergence of scientific ideas but also their influence and spread across the scientific community.

References

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. "Rescuing US Biomedical Research from Its Systemic Flaws." *Proceedings of the National Academy of Sciences* 111:16 (2014), 5773-5777. 10.1073/pnas.1404402111

Arts, Sam, Jianan Hou, and Juan Carlos Gomez. "Natural Language Processing to Identify the Creation and Impact of New Technologies in Patent Text: Code, Data, and New Measures." *Research Policy* 50:2 (2021) 104144. 10.1016/j.respol.2020.104144

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso, "Incentives and Creativity: Evidence from the Academic Life Sciences," *The RAND Journal of Economics* 42:3 (2011), 527–554. 10.1111/j.1756-2171.2011.00140.x

Bhattacharya, Jay, and Mikko Packalen. "Stagnation and Scientific Incentives," NBER working paper 26752 (2020).

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl, "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science," *Management Science* 62:10 (2016), 2765–2783. 10.1287/mnsc.2015.2285

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?" *American Economic Review* 110:4 (2020) 1104–1144. 10.1257/aer.20180338

Bush, Vannevar, *Science, the Endless Frontier: A Report to the President* (Washington: United States Government Printing Office, 1945).

Chai, Sen, and Anoop Menon, "Breakthrough Recognition: Bias Against Novelty and Competition for Attention," *Research Policy* 48:3 (2019) 733–747. 10.1016/j.respol.2018.11.006

Cheng, Mengjie, Daniel Scott Smith, Xiang Ren, Hancheng Cao, Sanne Smith, and Daniel A. McFarland, "How New Ideas Diffuse in Science," *American Sociological Review* 88:3 (2023) 522–561. 10.1177/00031224231166955

Cohan, Arman, Sergey Feldman, Iz Beltagy, Doug Downey and Daniel Weld, "SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 2270–2282. 10.18653/v1/2020.acl-main.207

Cozzens, Susan. "What do citations count? The rhetoric-first model" *Scientometrics* 15, (1989) 437-447.

Evans, James A., "Electronic Publication and the Narrowing of Science and Scholarship," *Science* 321:5887 (2008) 395–399. 10.1126/science.1150473

Evans, James A., "Industry Collaboration, Scientific Sharing, and the Dissemination of Knowledge" *Social Studies of Science* 40:5 (2010) 757–791. 10.1177/0306312710379931

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Fontana, Magda, Martina Iori, Fabio Montobbio, and Roberta Sinatra, "New and Atypical Combinations: An Assessment of Novelty and Interdisciplinarity," *Research Policy* 49: 7 (2020), 104063. 10.1016/j.respol.2020.104063

Fortunato, Santo, Carl T. Bergstrom, Katy Borner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang and Albert-László Barabási, "Science of science," *Science* 359:6379 (2018) 10.1126/science.aao0185

Foster, Jacob G., Andrey Rzhetsky, and James A. Evans, "Tradition and Innovation in Scientists' Research Strategies," *American Sociological Review* 80:5 (2015) 875–908. 10.1177/0003122415601618

Franzoni, Chiara, Paula Stephan, and Reinhilde Veugelers, "Funding Risky Research," *Entrepreneurship and Innovation Policy and the Economy* 1 (2022) 103–133. 10.1086/719252

Funk, Russell J., and Jason Owen-Smith, "A Dynamic Network Measure of Technological Change," *Management Science* 63:3 (2017) 791–817. 10.1287/mnsc.2015.2366

Gerow, Aaron, Yuening Hu, Jordan Boyd-Graber, David M. Blei, and James A. Evans, "Measuring Discursive Influence Across Scholarship," *Proceedings of the National Academy of Sciences* 115:13 (2018) 3308–3313. 10.1073/pnas.1719792115

Hofstra, Bas, Vivek V. Kulkarni, Sebastian Muñoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A. McFarland, "The Diversity–Innovation Paradox in Science" *Proceedings of the National Academy of Sciences* 117:17 (2020) 9284–9291. 10.1073/pnas.1915378117

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger, "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science" *The Quarterly Journal of Economics* 133:2 (2018) 927–991. 10.1093/qje/qjx046

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation Over the Long Run," *American Economic Review: Insights,* 3:3 (2021), 303–320. 10.1257/aeri.20190499

King, Gary, Michael Tomz, and Jason Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science* 44 (2000), 347-361. 10.2307/2669316

Kirkpatrick, Scott, C. Daniel Gelatt Jr., and Mario P. Vecchi, "Optimization by Simulated Annealing," *Science* 220:4598, (1983), 671–680. 10.1126/science.220.4598.671

Kojima, Masayasu, Hiroshi Hosoda, Yukari Date, Masamitsu Nakazato, Hisayuki Matsuo and Kenji Kangawa, "Ghrelin Is a Growth-Hormone-Releasing Acylated Peptide from Stomach," *Nature* 402 (1999), 656–660. 10.1038/45230

Kuhn, Thomas S., *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).

Kuhn, Tobias, Matjaž Perc, and Dirk Helbing, "Inheritance Patterns in Citation Networks Reveal Scientific Memes," *Physical Review X* 4:4 (2014), 041036. 10.1103/PhysRevX.4.041036

Leahey, Erin, Jina Lee, and Russell J. Funk, "What Types of Novelty Are Most Disruptive?" *American Sociological Review* 88:3 (2023), 562–597. 10.1177/00031224231168074

Leibel, Christian, and Lutz Bornmann. "What do we know about the disruption index in scientometrics? An overview of the literature," *Scientometrics* 129:1 (2024), 601–639. 10.1007/s11192-023-04873-5

Letchford, Adrian, Helen Susannah Moat, and Tobias Preis, "The advantage of short paper titles," Royal Society Open Science 2:8 (2015) 150266. 10.1098/rsos.150266

Li, Jichao, Yian Yin, Santo Fortunato, and Dashun Wang, "A Dataset of Publication Records for Nobel Laureates," *Scientific Data* 6 (2019), 33. 10.1038/s41597-019-0033-6

Lin, Zihang, Yian Yin, Lu Liu, and Dashun Wang, "SciSciNet: A Large-Scale Open Data Lake for the Science of Science Research," *Scientific Data* 10 (2023), 315. 10.1038/s41597-023-02198-9

Lin, Jimmy, and W. John Wilbur, "PubMed Related Articles: A Probabilistic Topic-Based Model for Content Similarity," *BMC Bioinformatics* 8 (2007), 423. 10.1186/1471-2105-8-423

Milojević, Staša, "Quantifying the Cognitive Extent of Science," *Journal of Informetrics* 9:4 (2015), 962–973. 10.1016/j.joi.2015.10.005

Milojević, Staša, "The Length and Semantic Structure of Article Titles–Evolving Disciplinary Practices and Correlations with Impact," *Frontiers in Research Metrics and Analytics* 2 (2017). 10.3389/frma.2017.00002

Mokyr, Joel, *The Gifts of Athena: Historical Origins of the Knowledge Economy* (Princeton and Oxford: Princeton University Press, 2002).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Nobel Foundation, "Statutes of the Nobel Foundation," *NobelPrize.org* (2024). Online: https://www.nobelprize.org/organization/statutes-of-the-nobel-foundation/, last accessed 2024/12/02.

Packalen, Mikko, and Jay Bhattacharya, "New Ideas in Invention," NBER working paper 20922 (2015).

Packalen, Mikko, and Jay Bhattacharya, "Age and the Trying Out of New Ideas," *Journal of Human Capital* 13:2 (2019), 341–373. 10.1086/703160

Packalen, Mikko, and Jay Bhattacharya, "NIH Funding and the Pursuit of Edge Science," *Proceedings of the National Academy of Sciences* 117:22 (2020), 12011–12016. 10.1073/pnas.1910160117

Park, Michael, Erin Leahey, and Russell J. Funk, "Papers and Patents Are Becoming Less Disruptive Over Time," *Nature* 613:7942 (2023), 138–144. 10.1038/s41586-022-05543-x

Priem, Jason, Heather Piwowar, and Richard Orr, "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts," *arXiv preprint arXiv:2205.01833* (2022). 10.48550/arXiv.2205.01833

Shi, Feng, and James A. Evans, "Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines," *Nature Communications* 14 (2023), 1641. 10.1038/s41467-023-36741-4

Shibayama, Sotaro, Deyun Yin, and Kuniko Matsumoto, "Measuring novelty in science with word embedding," *PloS One* 16:7 (2021), e0254034. 10.1371/journal.pone.0254034

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Teplitskiy, Misha, Eamon Duede, Michael Menietti, and Karim R. Lakhani, "How Status of Research Papers Affects the Way They Are Read and Cited," *Research Policy* 51:4 (2022), 104484. 10.1016/j.respol.2022.104484

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Benjamin Jones, "Atypical Combinations and Scientific Impact," *Science* 342:6157 (2013), 468–472. 10.1126/science.1240474

Wang, Jian, Reinhilde Veugelers, and Paula Stephan, "Bias Against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators," *Research Policy* 46:8 (2017), 1416–1436. 10.1016/j.respol.2017.06.006

Wu, Lingfei, Dashun Wang, and James A. Evans, "Large Teams Develop and Small Teams Disrupt Science and Technology," *Nature* 566:7744 (2019), 378–382. 10.1038/s41586-019-0941-9

[1] For a comprehensive overview of papers utilizing text metrics, see Arts, Hou, and Gomez (2021), Azoulay, Graff Zivin, and Manso (2011), Bhattacharya and Packalen (2020), Boudreau et al. (2016), Chai and Menon (2019), Cheng et al. (2023), Evans (2008, 2010), Fortunato et al. (2018), Foster, Rzhetsky, and Evans (2015), Gerow et al. (2018), Hofstra et al. (2020), Iaria, Schwarz, and Waldinger (2018), Kelly et al. (2021), Kuhn, Perc, and Helbing (2014), Milojević (2015), Packalen and Bhattacharya (2015, 2019, 2020), Park, Leahey, and Funk (2023), Shi and Evans (2023), Shibayama, Yin and Matsumoto (2021).

[2] The two papers most closely aligned with our approach are Cheng et al. (2023) and Hofstra et al. (2020). Cheng et al. (2023) track the diffusion of new scientific ideas using AutoPhrase, a machine learning tool trained on a broader corpus and sources such as Wikipedia, including

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

post-publication data, to extract key phrases based on their prominence and repetition. This approach may inadvertently select scientific ideas that have already gained traction and proven successful. Furthermore, the authors only consider phrases that appear at least 120 times across scientific publications as novel ideas. As we illustrate later, only the top 1.6% most impactful new noun phrases are reused at least 120 times. Hofstra et al. (2020) applies topic models to US doctoral dissertation titles and abstracts to assess novelty and how later dissertations incorporate these ideas. However, we believe that methods such as topic models and AutoPhrase may introduce bias by disproportionately identifying ideas that have already become significantly successful.

[3] Note that, like other machine learning tools (e.g., topic modeling and AutoPhrase), SPECTER is trained on the full corpus, including post-publication data, potentially biasing it toward identifying successful ideas.

[4] As 20% of the publications have only a title without an abstract, this could introduce potential bias. To address this, we include a binary indicator for abstract availability and control for text length in our regression analyses. Additionally, as a robustness check, we recalculated all text metrics using only titles across the full sample (see Appendix B). The main findings remain consistent, suggesting that titles alone can effectively identify new scientific ideas and their impact. However, the predictive power of title-based metrics is lower, highlighting the complementary value of abstracts.

[5] A noun phrase consists of a central noun or pronoun (the head) and accompanying words (dependents) that refine or specify its meaning.

[6] *Uzzi*, *Wang*, and *CD* metrics are undefined for papers that either lack citations to prior work or do not reference papers from at least two different journals (Leibel & Bornmann, 2024). To avoid selection bias, we assign a value of zero to these papers and include them in our

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

analysis. Importantly, our results remain consistent even when these papers are excluded from the analysis.

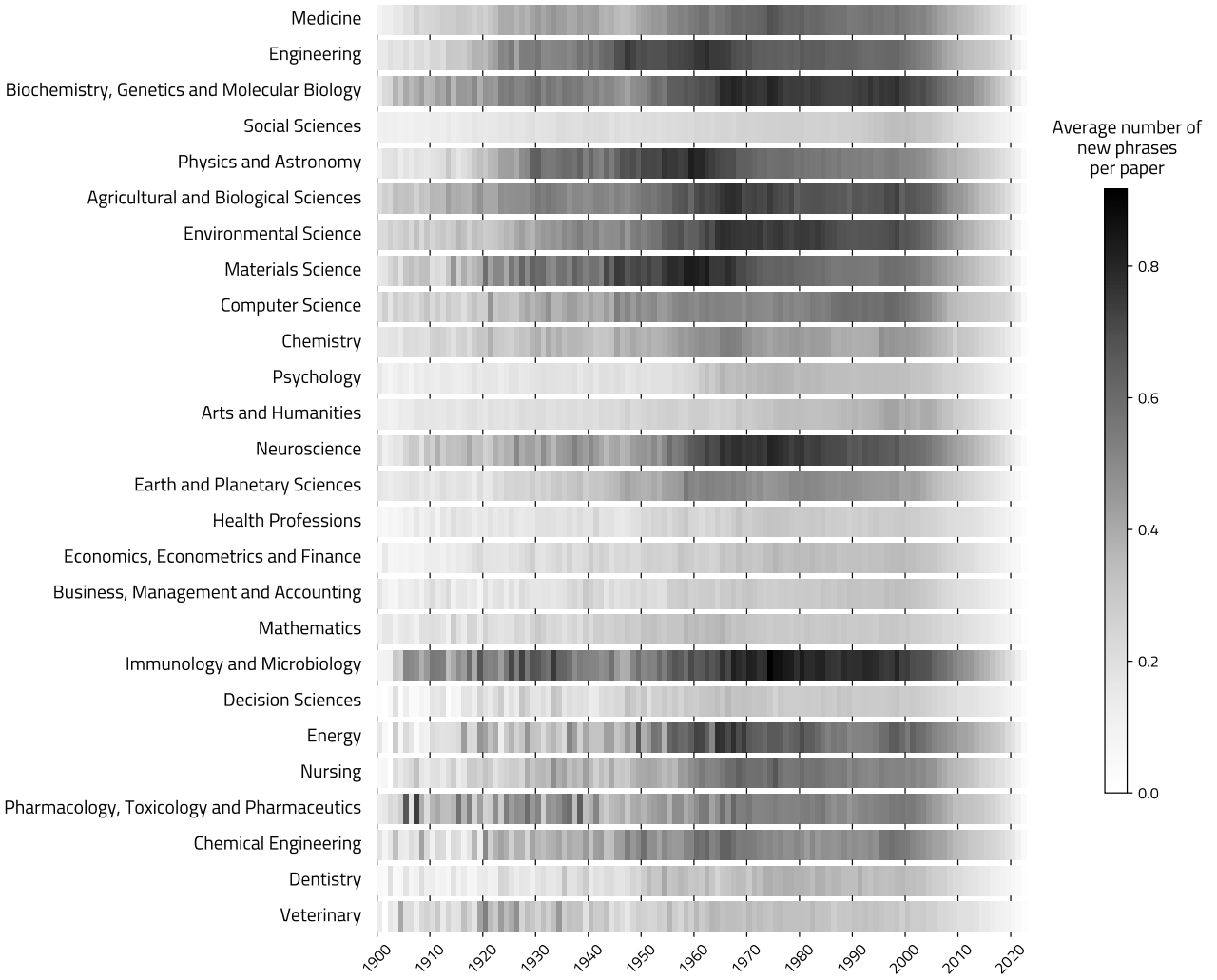
[7] OpenAlex assigns each paper to one of 252 subfields within a hierarchy of 26 fields, using subfield labels derived from Scopus. A machine learning model determines the appropriate subfield for each paper based on the text of its title and abstract, as well as co-citation patterns. More information can be found here: https://docs.openalex.org/api-entities/topics and a list of subfields can be found here https://api.openalex.org/topics.

[8] We use 252 subfield classifications from OpenAlex. If no subfield match is available, we use the coarser 26 field classifications.

[9] Papers in our sample are explicitly referenced in these lectures (Lin et al., 2019). We collected a total of 481 lectures, with each laureate delivering one and each prize averaging two lectures. Of these, 91% were available as PDFs, while the remaining were plain text sourced directly from the laureates' Nobel Prize pages.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

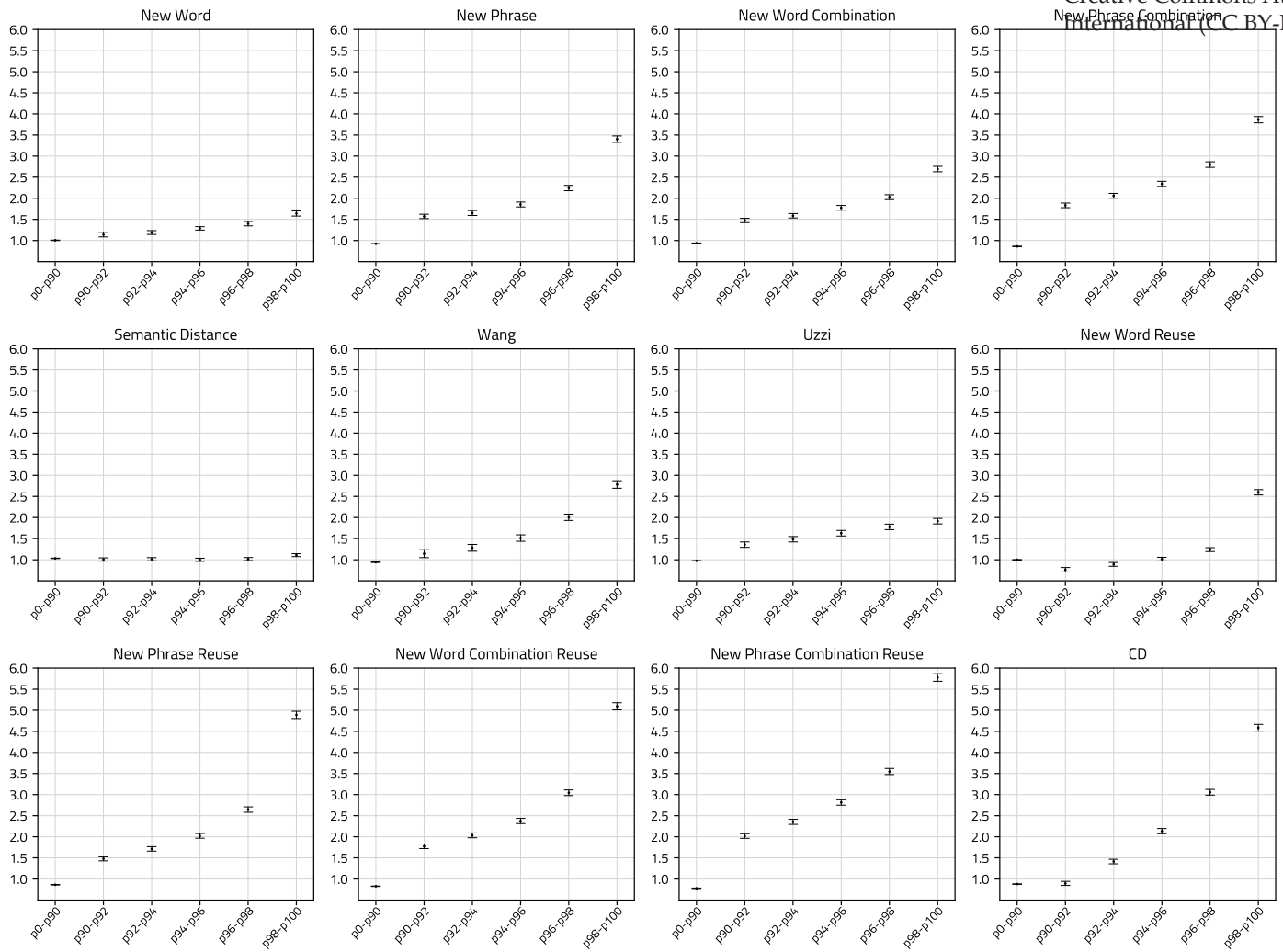
Figure 1: Scientific Novelty by Field of Study and Year



Notes: Paper-level average number of new noun phrases by fields of study (OpenAlex classification) and year (n= 75,295,921 papers published between 1901 and 2023). Fields are top down ordered by the overall number of publications.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Figure 2: Predicted Probability of Paper Being Top Cited



Notes: Figure 2 plots the predicted probability (in %) of a paper being among the top 1% most-cited (above the 99th percentile of received citations within its subfield and year), estimated using 12 separate linear probability models. Each model includes indicator variables for whether a paper's metric value, within its subfield and year, falls into the 0th–90th percentile (reference category) or one of five 2-percentile ranges from the 90th to 100th percentile. For consistency in comparisons, *Uzzi* is inverted (e.g., p98–p100 corresponds to p0–p2), ensuring higher novelty aligns with higher percentile values. The predicted probabilities are shown with 99.999% confidence intervals. Control variables in the models include whether the paper has an abstract, the number of unique words and phrases in title and abstract, the number of unique cited papers and journals, and fixed effects for subfield and year of publication. The figure is organized into 12 panels, each representing a different metric. The analysis includes all papers from the OpenAlex database published between 1901 and 2010 (n=37,154,406).

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Table 1: Summary Statistics Paper Level

	Mean	St.Dev.	Min	p25	p50	p75	p95	p99	Max	Skew
Panel A: Ex ante										
New Word (Binary)	0.068	0.252	0	0	0	0	1	1	1	3.432
New Word	0.083	0.361	0	0	0	0	1	2	281	17.244
New Phrase (Binary)	0.245	0.430	0	0	0	0	1	1	1	1.184
New Phrase	0.360	0.779	0	0	0	0	2	3	223	4.128
New Word Combination (Binary)	0.614	0.487	0	0	1	1	1	1	1	-0.469
New Word Combination	13.222	97.659	0	0	2	11	55	147	256,995	980.874
New Phrase Combination (Binary)	0.649	0.477	0	0	1	1	1	1	1	-0.624
New Phrase Combination	10.948	45.742	0	0	2	11	47	104	44,871	139.806
Semantic Distance	0.136	0.046	0.010	0.105	0.131	0.162	0.220	0.271	0.740	0.700
Wang	0.192	2.088	0	0	0	0	0.844	4.208	6,052.686	503.870
Uzzi	14.911	147.781	-445.270	0	0	0.034	51.951	363.955	67,782	43.508
Panel B: Ex post										
New Word Reuse	3.368	367.690	0	0	0	0	2	20	769,278	812.412
New Phrase Reuse	6.555	268.222	0	0	0	0	12	71	416,735	522.944
New Word Combination Reuse	197.827	2,322.053	0	0	6	52	584	3,061	1,359,901	119.531
New Phrase Combination Reuse	67.818	402.917	0	0	6	40	271	932	275,329	79.233
CD	0.002	0.039	-1	0	0	0	0.005	0.094	1	9.578
Panel C: Controls										
Abstract (Binary)	0.801	0.400	0	1	1	1	1	1	1	-1.504
N. of Words	31.767	25.579	1	9	30	47	71	103	2,489	3.089
N. of Phrases	13.373	12.215	0	3	11	20	34	48	1,728	3.249
N. of cited Papers	18.203	26.856	0	0	9	28	62	114	6,315	6.304
N. of cited Journals	9.611	12.774	0	0	5	15	33	54	1,132	2.735

Notes: n= 75,295,921 papers published between 1901 and 2023. p25, p50, p75, p95 and p99 are respectively the 25th, 50th, 75th, 95th and the 99th percentile. The skewness (skew) of the distributions is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. A positive skewness indicates that the distribution has a long right tail, while a negative skewness indicates a long left tail.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Table 2: Examples of Nobel Prizes

Prize	Short Prize Motivation	Paper	New word (phrase) or new combination of words (phrases)
Chemistry 1934	Discovery of heavy hydrogren	Urey, Harold C., Ferdinand G. Brickwedde, and George M. Murphy, "A Hydrogen Isotope of Mass 2," *Physical Review* 39:1 (1932), 164.	hydrogen_isotope (4,074)
Physics 1936	Discovery of the positron	Anderson, Carl D., "The Positive Electron," *Physical Review* 43:6 (1933), 491.	positron (94,146)
Medicine 1952	Discovery of streptomycin, the first antibiotic effective against tuberculosis	Schatz, Albert, Elizabeth Bugle, and Selman A. Waksman, "Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria," *Proceedings of the Society for Experimental Biology and Medicine* 55:1 (1944), 66–69.	streptomycin (23,334)
Physics 1956	Researches on semiconductors and discovery of the transistor effect	Bardeen, John, and Walter H. Brattain, "The Transistor, a Semi-Conductor Triode," *Physical Review* 74:2 (1948), 230.	transistor (152,047)
Physics 1959	Discovery of the antiproton	Chamberlain, Owen, Emilio Segrè, Clyde Wiegand, and Thomas Ypsilantis, "Observation of Antiprotons," *Physical Review* 100:3 (1955), 947.	antiproton (5,023)
Medicine 1962	Discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material	Watson, James D., and Francis H. C. Crick, "The Structure of DNA," *Cold Spring Harbor Symposia on Quantitative Biology* 18:1 (1953), 123–131.	dna,genetic_material (1,985)
Medicine 1963	Discoveries concerning the ionic mechanisms involved in excitation and inhibition in the peripheral and central portions of the nerve cell membrane	Eccles, John C., Peter Fatt, and Koichi Koketsu, "Cholinergic and Inhibitory Synapses in a Pathway from Motor-Axon Collaterals to Motoneurones," *The Journal of Physiology* 126:3 (1954), 524.	inhibitory_synapsis (2,275)
Medicine 1969	Discoveries concerning the replication mechanism and the genetic structure of viruses	Hershey, Alfred D., and Martha Chase, "Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage," *The Journal of General Physiology* 36:1 (1952), 39–56.	dna,viral (58,844)
Physics 1971	Discovery for the invention and development of the holographic method	Gabor, Dennis, "A New Microscopic Principle," *Nature* 161:4098 (1948), 777–778.	holography (6,829)
Medicine 1979	Development of computer assisted tomography	Hounsfield, Godfrey N., "Computerized Transverse Axial Scanning (Tomography): Part 1. Description of System," *The British Journal of Radiology* 46:552 (1973), 1016–1022.	computerize,tomography (21,835)
Physics 1986	Fundamental work in electron optics, and for the design of the first electron microscope	Binnig, Gerd, Heinrich Rohrer, Christoph Gerber, and Ed Weibel, "Surface Studies by Scanning Tunneling Microscopy," *Physical Review Letters* 49:1 (1982), 57.	scanning_tunnel_microscopy (17,914)
Chemistry 1993	Invention of the polymerase chain reaction (PCR) method	Mullis, Kary, Fred Faloona, Sherry Scharf, Randall Saiki, Glenn Horn, and Henry Erlich, "Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction," *Cold Spring Harbor Symposia on Quantitative Biology* 51:1 (1986), 263–273.	polymerase_chain_reaction (209,968)
Medicine 1997	Discovery of Prions - a new biological principle of infection	Prusiner, Stanley B., "Novel Proteinaceous Infectious Particles Cause Scrapie," *Science* 216:4542 (1982), 136–144.	prion,protein (13,779)
Medicine 1998	Discoveries concerning nitric oxide as a signalling molecule in the cardiovascular system	Katsuki, Shiro, Walter Arnold, Charanjit Mittal, and Ferid Murad, "Stimulation of Guanylate Cyclase by Sodium Nitroprusside, Nitroglycerin, and Nitric Oxide in Various Tissue Preparations and Comparison to the Effects of Sodium Azide and Hydroxylamine," *Journal of Cyclic Nucleotide Research* 3:1 (1977), 23–35.	nitric_oxide,stimulation (4,641)
Chemistry 2001	Work on chirally catalysed hydrogenation reactions	Nozaki, Hiroshi, Hiroshi Takaya, Seiichi Moriuti, and Ryoji Noyori, "Homogeneous Catalysis in the Decomposition of Diazo Compounds by Copper Chelates: Asymmetric Carbenoid Reactions," *Tetrahedron* 24:9 (1968), 3655–3669.	chiral,synthesis (45,954)
Physics 2005	Contribution to the quantum theory of optical coherence	Glauber, Roy J., "The Quantum Theory of Optical Coherence," *Physical Review* 130:6 (1963), 2529.	optical_coherence (2,982)
Physics 2007	Discovery of Giant Magnetoresistance	Baibich, Mário N., Jacques M. Broto, Albert Fert, Francis N. Van Dau, Frédéric Petroff, Paul Etienne, and J. Chazelas, "Giant Magnetoresistance of (001) Fe/(001) Cr Magnetic Superlattices," *Physical Review Letters* 61:21 (1988), 2472.	giant_magnetoresistance (3,229)
Medicine 2007	Discoveries of principles for introducing specific gene modifications in mice by the use of embryonic stem cells	Thomas, Kirk R., and Mario R. Capecchi, "Site-Directed Mutagenesis by Gene Targeting in Mouse Embryo-Derived Stem Cells," *Cell* 51:3 (1987), 503–512.	gene_target (4,543)
Medicine 2009	Discovery of how chromosomes are protected by telomeres and the enzyme telomerase	Szostak, Jack W., and Elizabeth H. Blackburn, "Cloning Yeast Telomeres on Linear Plasmid Vectors," *Cell* 29:1 (1982), 245–255.	polymerase,telomere (1,753)
Medicine 2011	Discovery of the dendritic cell and its role in adaptive immunity	Steinman, Ralph M., and Zanvil A. Cohn, "Identification of a Novel Cell Type in Peripheral Lymphoid Organs of Mice. I. Morphology, Quantitation, Tissue Distribution," *The Journal of Experimental Medicine* 137:5 (1973), 1142–1162.	macrophage,dendritic_cell (8,299)

Notes: 20 illustrative examples of Nobel prizes, a corresponding paper, and a new word (phrase) or new combination of words (phrases) introduced by the paper and found in the corresponding Nobel prize motivation page. Words (lemmatized) in phrases are separated by the underscore (e.g. 'optical_coherence' is a phrase) and new word (phrase) combinations are separated by the comma (e.g.'chiral,synthesis' is a new word combination). The reuse by later papers is shown in parentheses.

Review of Economics and Statistics Just Accepted MS.
https://doi.org/10.1162/rest_a_01561

Table 3: Predicting Nobel Prize Papers

	Text Metrics					Traditional Metrics		Text Metrics Combined	All Metrics Combined
Panel A: Ex ante									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
New Word	1.336***							0.243	0.257
	(0.292)							(0.324)	(0.324)
New Phrase		1.906***						1.498***	1.493***
		(0.193)						(0.213)	(0.214)
New Word Combination			0.675***					0.275**	0.288**
			(0.090)					(0.112)	(0.112)
New Phrase Combination				0.984***				0.596***	0.605***
				(0.139)				(0.153)	(0.154)
Semantic Distance					-0.696			-1.069	-1.145
					(1.352)			(1.460)	(1.459)
Wang						0.226			-0.030
						(0.656)			(0.677)
Uzzi							-0.217		-0.369**
							(0.177)		(0.182)
Log-Likelihood	-709.1	-665.6	-694.2	-695.0	-721.8	-721.9	-721.3	-649.8	-648.3
Pseudo-R^2	0.113	0.167	0.131	0.130	0.097	0.097	0.098	0.187	0.189
Precision (%)	67.32	69.87	68.88	68.61	66.31	66.55	66.06	72.20	72.17
Recall (%)	65.45	65.63	64.93	65.28	63.89	63.54	63.19	67.19	67.53
AUC	0.7198	0.7635	0.7361	0.7362	0.7086	0.7083	0.7084	0.7745	0.7761
Marginal Effects (%)	9.09	20.93	23.05	32.15	-0.81	0.64	-1.78		
Panel B: Ex post									
	(10)	(11)	(12)	(13)		(14)		(15)	(16)
New Word Reuse	0.338***							0.111*	0.108*
	(0.052)							(0.065)	(0.064)
New Phrase Reuse		0.582***						0.482***	0.471***
		(0.050)						(0.055)	(0.055)
New Word Combination Reuse			0.314***					0.169***	0.179***
			(0.034)					(0.039)	(0.039)
New Phrase Combination Reuse				0.386***				0.215***	0.199***
				(0.051)				(0.058)	(0.058)
CD						2.060***			1.723***
						(0.451)			(0.473)
Log-Likelihood	-696.6	-625.6	-671.6	-688.0		-708.9		-598.3	-591.3
Pseudo-R^2	0.128	0.217	0.160	0.139		0.113		0.251	0.260
Precision (%)	68.32	74.42	69.80	70.20		66.67		75.82	76.92
Recall (%)	64.41	67.19	67.01	67.88		63.89		68.58	69.44
AUC	0.7296	0.7938	0.7575	0.7470		0.7208		0.8139	0.8189
Marginal Effects (%)	13.23	25.66	23.02	23.31		7.85			

Notes: Logit regressions with a binary outcome for Nobel Prize paper, robust standard errors in parentheses. Sample includes 1,168 papers (584 Nobel Prize papers matched with 584 control papers from the same year, journal, and subfield). All measures, except for *Semantic Distance* and *CD*, are log-transformed (after adding 1 for zero values). Models control for publication year and subfield fixed effects, abstract availability, text length (unique words and phrases in title and abstract), and number of unique papers and journals cited. AUC represents the area under the ROC curve. Marginal effects show the percentage increase in the likelihood of being a Nobel Prize paper associated with a one-standard-deviation increase in the metric. Baseline model (controls only): Precision = 66.25, Recall = 64.06, AUC = 0.7083. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$