Tracing the Endless Scientific Frontier in Publication Texts

Sam Arts, Nicola Melluso, Reinhilde Veugelers

Presenting author: **Nicola Melluso**

Department of Management, Strategy and Innovation
KU Leuven

July 2023

Motivation

- New scientific ideas drive scientific and technological progress, economic growth, and social welfare (Bush, 1945; Price, 1963; Jones & Summers, 2020)
- Opportunities for highly novel ideas are shrinking over time
  - Slowdown of canonical science progress and decreasing disruption (Chu & Evans, 2019; Park et al., 2023)
  - Burden of knowledge (Jones 2009; Bloom et al. 2020)
- Difficulties in identifying and evaluating the value of new scientific ideas (e.g. bias against novelty/risk) with breakthrough that can be missed (Boudreau et al., 2016; Wang et al., 2017; Veugelers et al., 2023)
- Heterogeneity in novelty and impact of new scientific ideas (Seglen, 1992)
  - General patterns of progress in science: most of the time incremental improvements in existing paradigms and occasional big shocks in paradigms (Uzzi et al., 2013)

Background

- Approaches to measure novelty and impact of science by using patterns of citations to prior art (references) (Uzzi et al., 2013; Wang et al., 2017)
- Limitations:
  - Bibliographic data only capture prior art but not the scientific content, or, on the contrary, reflect little-to-no intellectual influence on the authors citing them (Teplitskiy et al., 2022)
  - Overlap of indicators: novelty overlaps with interdisciplinarity when looking at citations (Fontana et al., 2020)
  - No established validation: do new combinations of citations and keywords capture novel scientific ideas?
  - Difficult to interpret the meaning of indicators based on prior art (citation-based):
    what does the paper actually introduce?

What do we do: Natural Language Processing

- We use Natural Language Processing (NLP) techniques to develop metrics of **novelty** and **impact** based on text, circumventing the use of citations
- Scientific ideas are embedded in the text of scientific articles and new scientific ideas can only be identified through shifts in language (Kuhn 1970)



"*Language shifts are the fingerprint of paradigm shifts*"
Thomas Kuhn

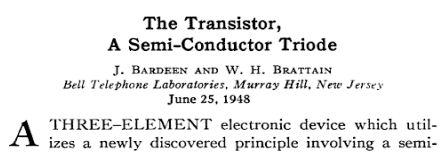
Data

- Microsoft Academic Graph (MAG) (now OpenAlex)
  - Coverage from 1780 until 2020 (about 100M scientific works)
  - Open access and broader coverage than the Web of Science or Scopus
- We develop text-based metrics for **72,245,396** journal and conference papers from 1901 to 2020 in english language
- Concatenate title, abstract, eliminate stop words, words < 2 characters, numbers, words which appear only once in all papers, lemmatization

New word

- **new_word**
  first time appearance of a new word in a paper
- **new_word_reuse**
  number of times reused in future papers

"transistor"
(reused 152,036)

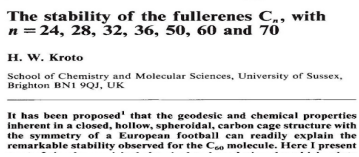


The Transistor,
A Semi-Conductor Triode

J. Bardeen and W. H. Brattain
Bell Telephone Laboratories, Murray Hill, New Jersey
June 25, 1948

A THREE–ELEMENT electronic device which utilizes a newly discovered principle involving a semi-

"fullerene"
(reused 94,149)



The stability of the fullerenes C_n, with $n = 24, 28, 32, 36, 50, 60$ and 70

H. W. Kroto

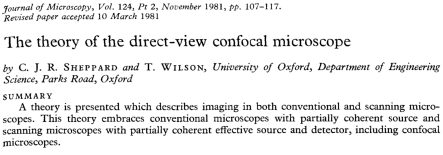
School of Chemistry and Molecular Sciences, University of Sussex, Brighton BN1 9QJ, UK

It has been proposed[1] that the geodesic and chemical properties inherent in a closed, hollow, spheroidal, carbon cage structure with the symmetry of a European football can readily explain the remarkable stability observed for the C_{60} molecule. Here I present

New bigram

- **new_bigram**
  first time appearance of two consecutive words in a paper
- **new_bigram_reuse**
  number of times reused in future papers

"confocal microscope"
(reused 43,401)



Journal of Microscopy, *Vol.* 124, *Pt* 2, *November* 1981, *pp.* 107–117.
Revised paper accepted 10 *March* 1981

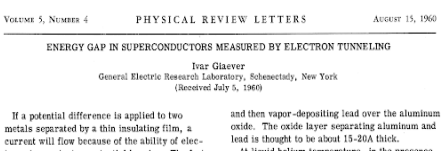
The theory of the direct-view confocal microscope

by C. J. R. SHEPPARD *and* T. WILSON, *University of Oxford, Department of Engineering Science, Parks Road, Oxford*

SUMMARY

A theory is presented which describes imaging in both conventional and scanning microscopes. This theory embraces conventional microscopes with partially coherent source and scanning microscopes with partially coherent effective source and detector, including confocal microscopes.

"electron tunneling"
(reused 4,759)



VOLUME 5, NUMBER 4 PHYSICAL REVIEW LETTERS AUGUST 15, 1960

ENERGY GAP IN SUPERCONDUCTORS MEASURED BY ELECTRON TUNNELING

Ivar Giaever
General Electric Research Laboratory, Schenectady, New York
(Received July 5, 1960)

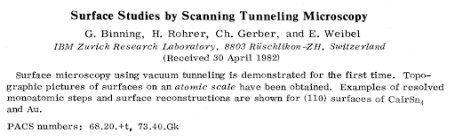
If a potential difference is applied to two metals separated by a thin insulating film, a current will flow because of the ability of elec-

and then vapor-depositing lead over the aluminum oxide. The oxide layer separating aluminum and lead is thought to be about 15-20Å thick.

New trigram

- **new_trigram**
  first time appearance of three consecutive words in a paper
- **new_trigram_reuse**
  number of times reused in future papers

"scanning tunneling microscopy"
(reused 17,011)



Surface Studies by Scanning Tunneling Microscopy

G. Binning, H. Rohrer, Ch. Gerber, and E. Weibel
IBM Zurich Research Laboratory, 8803 Rüschlikon-ZH, Switzerland
(Received 30 April 1982)

Surface microscopy using vacuum tunneling is demonstrated for the first time. Topographic pictures of surfaces on an *atomic scale* have been obtained. Examples of resolved monoatomic steps and surface reconstructions are shown for (110) surfaces of $CaIrSn_4$ and Au.

PACS numbers: 68.20.+t, 73.40.Gk

"polymerase chain reaction"
(reused 61,626)



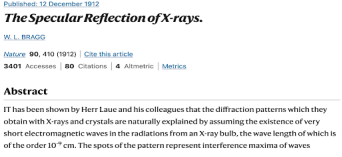
Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction

K. MULLIS, F. FALOONA, S. SCHARF, R. SAIKI, G. HORN, AND H. ERLICH
Cetus Corporation, Department of Human Genetics, Emeryville, California 94608

New word combinations

- **new_word_comb**
  first time of appearance of a word combination (no matter the order in the text)
- **new_word_comb_reuse**
  number of times reused in future papers

"x-ray diffraction"
(reused 528,592)



Published: 12 December 1912

The Specular Reflection of X-rays.

W. L. BRAGG

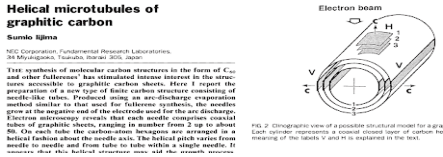
Nature **90**, 410 (1912) | Cite this article

3401 Accesses | **80** Citations | **4** Altmetric | Metrics

Abstract

IT has been shown by Herr Laue and his colleagues that the diffraction patterns which they obtain with X-rays and crystals are naturally explained by assuming the existence of very short electromagnetic waves in the radiations from an X-ray bulb, the wave length of which is of the order 10^{-9} cm. The spots of the pattern represent interference maxima of waves

"carbon nanotube"
(reused 12,245)



Helical microtubules of graphitic carbon

Sumio Iijima

NEC Corporation, Fundamental Research Laboratories, 34 Miyukigaoka, Tsukuba, Ibaraki 305, Japan

Tiny synthesis of molecular carbon structures in the form of C_{60} and other fullerenes' has stimulated intense interest in the structures accessible to graphitic carbon sheets. Here I report the preparation of a new type of finite carbon structure consisting of needle-like tubes. Produced using an arc-discharge evaporation method similar to that used for fullerene synthesis, the needles grow at the negative end of the electrode used for the arc discharge. Electron microscopy reveals that each needle comprises coaxial tubes of graphitic sheets, ranging in number from 2 up to about 50. On each tube the carbon-atom hexagons are arranged in a helical fashion about the needle axis. The helical pitch varies from needle to needle and from tube to tube within a single needle. It appears that this helical structure may aid the growth process

FIG. 2 Clinographic view of a possible structural model for a graphitic tubule. Each cylinder represents a coaxial closed layer of carbon hexagons. The meaning of the labels V and H is explained in the text.

Descriptives: Textual Units

		Reuse						
	n. papers	**Mean**	**Std**	**Min**	**Q1**	**Q2**	**Q3**	**Skewness**
new_word	7,305,080	75.86	2,943.04	1.00	1.00	3.00	7.00	243.49
new_bigram	32,104,336	26.48	569.07	1.00	1.00	2.00	8.00	359.33
new_trigram	75,573,271	8.00	141.93	1.00	1.00	2.00	4.00	677.10
new_word_comb	1,129,014,727	48.68	1,101.71	1.00	1.00	2.00	9.00	240.32

Notes: Descriptive table of the textual units identified in 72,245,396 journal and conference papers from 1901 to 2020 in english language. *Q1* is the 25% percentile. *Q2* is the median (50% percentile). *Q3* is the 75% percentile. The *skewness* 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).

Descriptives: Textual Units

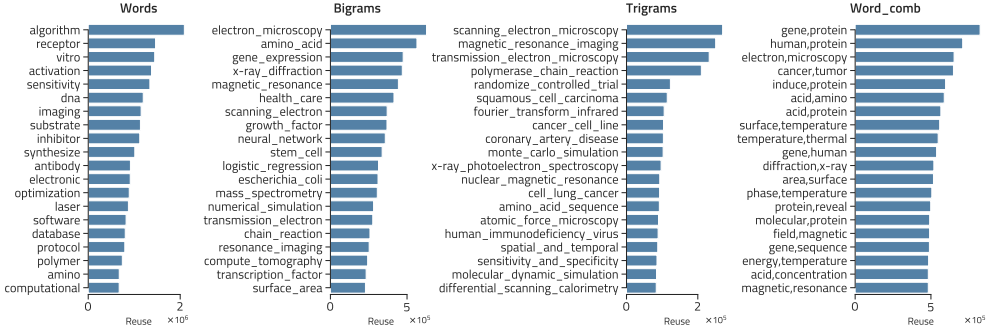


Figure: Top 20 reused words, bigrams, trigrams and word combinations.

Paper Level

- **new_word**
  n. of unique new words introduced by the focal paper
- **new_bigram**
  n. of new bigrams that do not contain one of the new words introduced by the paper
- **new_trigram**
  n. of new trigrams that do not contain one of the new words or new bigrams introduced by the paper
- **new_gram**
  n. of new words + n. of new bigrams that do not contain one of the new words introduced by the paper + n. of new trigrams that do not contain one of the new words or new bigrams introduced by the paper
- **new_word_comb**
  n. of new word combinations (word pairs) that do not contain one of the new words introduced by the paper, do not correspond to a new bigram introduced by the paper, and do not contain one of the three words contained in the new trigrams introduced by the paper

Descriptives: Paper Level

	Mean	Std	Q1	Q2	Q3	Skewness
new_word_bin	0.079	0.270	0	0	0	3.123
new_word	0.101	0.409	0	0	0	8.227
new_word_reuse	7.670	970.971	0	0	0	751.848
new_bigram_bin	0.267	0.442	0	0	1	1.053
new_bigram	0.426	0.923	0	0	1	4.332
new_bigram_reuse	11.483	403.185	0	0	1	536.651
new_trigram_bin	0.469	0.499	0	0	1	0.126
new_trigram	1.046	1.651	0	0	2	2.885
new_trigram_reuse	8.371	155.778	0	0	3	667.781
new_gram_bin (NG)	0.545	0.498	0	1	1	-0.182
new_gram	1.573	2.424	0	1	2	3.425
new_gram_reuse	27.524	1,101.109	0	1	5	575.007
new_word_comb_bin (NC)	0.513	0.500	0	1	1	-0.052
new_word_comb	14.598	59.141	0	1	11	33.636
new_word_comb_reuse	747.275	45,935.323	0	1	39	683.762
NG = 1 & NC = 1	0.404	0.491	0	0	1	0.391
NG = 1 or NC = 1	0.654	0.476	0	1	1	-0.649

Notes: Descriptive table of textual metrics for 72,245,396 journal and conference papers from 1901 to 2020 in English language. All the metrics are exclusive (e.g. new_bigram is the n. of new bigrams that do not contain the new words introduced by the focal paper). *Q1* is the 25% percentile. *Q2* is the median (50% percentile). *Q3* is the 75% percentile. The *skewness* 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).

Descriptives: Paper level

	# papers	Share of total papers	Share of new_word_bin	Share of new_bigram_bin	Share of new_trigram_bin	Share of NG	Share of NC
new_word_bin	5,702,598	0.079	1.000	0.528	0.750	1.000	0.857
new_bigram_bin	19,299,593	0.267	0.156	1.000	0.767	1.000	0.830
new_trigram_bin	33,848,424	0.469	0.126	0.437	1.000	1.000	0.759
new_gram_bin (NG)	39,390,166	0.545	0.145	0.490	0.859	1.000	0.741
new_word_comb_bin (NC)	37,069,460	0.513	0.132	0.432	0.693	0.787	1.000

Notes: Descriptive of the binary exclusive metrics and their overlaps for n= 72,245,396 papers published between 1901 and 2020. Each row represents the number of papers that score positively for the binary metrics. Each column (except for "papers") represents the subset of papers (share of papers) that score positively for the particular binary metrics.

Condition	Metric	Mean	Std Dev.	Q1	Q2	Q3	Skewness
new_word_bin = 1	new_word	1.281	0.777	1.000	1.000	1.000	7.129
new_bigram_bin = 1	new_bigram	1.594	1.151	1.000	1.000	2.000	4.405
new_trigram_bin = 1	new_trigram	2.232	1.781	1.000	2.000	3.000	2.844
new_gram_bin = 1	new_gram	2.884	2.645	1.000	2.000	4.000	3.403
new_word_comb_bin = 1	new_word_comb	28.450	80.141	3.000	10.000	29.000	25.827

Notes: Summary statistics for exclusive metrics conditional on scoring positively in their binary form. n= 72,245,396 papers published between 1901 and 2020. Q1 is the 25th percentile, Q2 is the median (50th percentile) and Q3 is the 75th 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.

Benchmark

- Distance metrics
  - **cosine_avg**
    1 - average cosine similarity with 5-years prior papers
  - **cosine_max**
    1 - maximum cosine similarity with 5-years prior papers
- Citation-based metrics
  - **wang**
    sum of the distance of novel combinations of cited journals (of cited papers) (Wang et al., 2017)
  - **uzzi**
    atypical combination (10-th percentile of z-score) of prior knowledge as measured by the proximity between pairs of journals (of cited papers) (Uzzi et al., 2013)
  - **cd** (*disruptiveness)*
    cd (disruption score at 5 years level) quantifies the extent to which a paper disrupts or develops the existing literature (Funk & Owen-Smith, 2017)

SPECTER document embedding (Cohan, 2020)

- transformer-based (BERT-like) language model tailored for scientific language (best performing model at the state-of-the-art)
- 768-dimensions vector representation of the title and abstract of a focal paper
- accounting for the context in which words appear (synonymity, polysemy, change in language...)

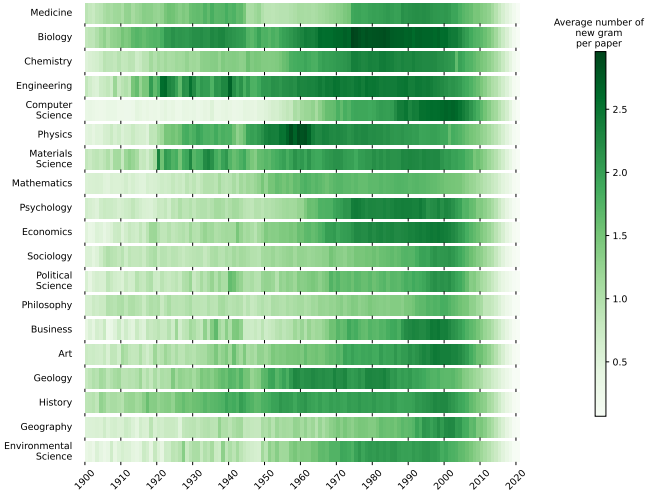
Descriptives: Paper Level

	cosine_max	cosine_avg	wang >0	uzzi <0	cd
NG_excl = 1 & NC_excl = 1	0.4291	0.9933	0.1246	0.3201	0.0049
NG_excl = 1 or NC_excl = 1	0.4231	0.9933	0.1142	0.2913	0.0043
NG_excl = 1 & NC_excl = 0	0.4073	0.9932	0.0987	0.2384	0.0036
NG_excl = 0 & NC_excl = 1	0.4210	0.9933	0.0958	0.2534	0.0035
NG_excl = 0 & NC_excl = 0	0.3889	0.9931	0.0515	0.1762	0.0019

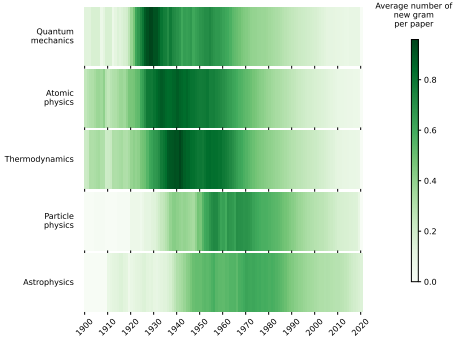
Notes: Average values of *cosine_max*, *cosine_avg*, *wang >0*, *uzzi <0* and *cd* conditional on scoring positively on one of the novelty metrics. *wang >0* indicates that a paper is novel according to *wang*. *uzzi <0* indicates that a paper is novel according to *uzzi*

.

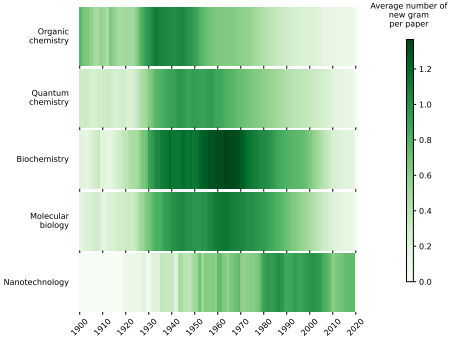
Novelty over time by fields



Novelty over time by subfields



(a) Subfields of Physics



(b) Subfields of Chemistry

Validation

- **Nobel prize papers**
  - Papers likely introduced fundamentally new scientific ideas with a major impact on scientific progress
  - Nobel prize award specifies the novel idea that motivates the prize
- Case-control design
  - Match one-to-one Nobel prize papers with control papers (database of Nobel prize papers from Li et al. (2019))
    - Control papers: same year, journal, volume, issue
  - Ability to distinguish treatment from control papers
    - T-test, Cohen's d
    - Precision, recall, AUC (area under the ROC-curve)

Validation: Nobel Prizes

- Gerd Binnig and Heinrich Rohrer introduce for the first time the “scanning tunnel microscopy” for which they won the Nobel prize in Physics in 1986
  - The paper titled “Surface studies by scanning tunneling microscopy”, published in *Physical review letters* in 1982 introduce for the first time the gram “scanning tunneling miscropy” (reused 17,011 times by future papers)

Surface Studies by Scanning Tunneling Microscopy

G. Binning, H. Rohrer, Ch. Gerber, and E. Weibel
IBM Zurich Research Laboratory, 8803 Rüschlikon-ZH, Switzerland
(Received 30 April 1982)

Surface microscopy using vacuum tunneling is demonstrated for the first time. Topographic pictures of surfaces on an *atomic scale* have been obtained. Examples of resolved monoatomic steps and surface reconstructions are shown for (110) surfaces of $CaIrSn_4$ and Au.

PACS numbers: 68.20.+t, 73.40.Gk



Validation: Nobel Prizes

ccc

Nobel Prize	Prize Motivation	Paper	new_gram	new_word_comb
1936 Physics	Discovery of the positron	Anderson C. D. (1933)	positron (94,149)	
1937 Physics	Experimental discovery of the diffraction of electrons by crystals	Davisson & Germer (1927)	diffraction of electron (347)	diffraction factor (31,543)
1946 Chemistry	Discovery that enzymes can be crystallized	Sumner (1926)	crystallization of enzyme (3,371)	enzyme isolation (23,532)
1948 Chemistry	Research on electrophoresis and absorption analysis	Tiselius (1937)		electrophoresis immune (3,975)
1950 Physics	Development of the photographic method	Lattes et al. (1947)	pion production (5,960)	
1956 Physics	Researches on semiconductors and discovery of the transistor effect	Bardeen & Brattain (1948)	transistor (152,037)	
1959 Physics	Discovery of the antiproton	Chamberlain et al. (1955)	antiproton electron (720)	
1962 Chemistry	Studies of the structures of globular proteins	Kendrew et al. (1958)	muscle three-dimensional (6,427)	
1962 Medicine	Discoveries concerning the molecular structure of nucleic acids	Crick (1953)	coiled-coil (8,111)	
1965 Medicine	Discoveries concerning genetic control of enzyme and virus synthesis	Monod et al. (1963)	allosteric transition (678)	
1967 Medicine	Discoveries concerning the primary physiological and chemical visual processes in the eye	Eldred et al. (1953)	supraspinal (5,016)	
1973 Physics	Experimental discoveries regarding tunneling phenomena in semiconductors and superconductors	Giaever (1960)	electron tunneling (4,692)	superconductor tunneling (2,102)
1977 Physics	Fundamental theoretical investigations of the electronic structure of magnetic and disordered systems	Mott (1949)	exciton transition (11,120)	
1981 Chemistry	Theories concerning the course of chemical reactions	Woodward & Hoffmann (1965)	sigmatropic (2,101)	
1997 Medicine	Discovery of prions, a new biological principle of infection	Prusiner (1982)	prion (1,282)	
2001 Chemistry	Work on chirally catalysed hydrogenation reactions	Knowles & Sabacky (1968)	catalytic asymmetric hydrogenation (247)	
2014 Chemistry	Development of super-resolved fluorescence microscopy	Hell & Wichmann (1994)	emission depletion (956)	

Notes: For each Nobel prize paper, the table displays the year and field of the award, the prize motivation, the short reference of the Nobel prize publication, one example new word or new bigram or new trigram, and/or new word combination introduced by the paper. The numbers between brackets show the total number of future papers which reuse these new grams and/or word combinations.

Validation: Nobel Prizes

ccc

Nobel Prize	Prize Motivation	new_gram	new_word_comb	new_gram_reuse	new_word_comb_reuse	wang	uzzi
1936 Physics	Discovery of the positron	9	17	102,119	16,576	0.000	504.802
1937 Physics	Experimental discovery of the diffraction of electrons by crystals	24	498	87,407	619,220	0.000	723.095
1946 Chemistry	Discovery that enzymes can be crystallized	1	4	3,371	23,544	0.000	121.250
1948 Chemistry	Research on electrophoresis and absorption analysis	1	2	7	5,189	0.000	98.344
1950 Physics	Development of the photographic method	5	274	7,012	66,931	0.000	114.837
1956 Physics	Researches on semiconductors and discovery of the transistor effect	2	2	152,048	18	0.000	150.199
1959 Physics	Discovery of the antiproton	1	2	5,023	17	0.000	511.484
1962 Chemistry	Studies of the structures of globular proteins	12	166	9,236	40,838	0.000	-4.675
1962 Medicine	Discoveries concerning the molecular structure of nucleic acids	3	55	8,118	4,456	0.000	190.530
1965 Medicine	Discoveries concerning genetic control of enzyme and virus synthesis	3	27	1,196	95,798	0.201	1.079
1973 Physics	Experimental discoveries regarding tunneling phenomena in semiconductors and superconductors	1	2	4,759	6,086	0.000	306.496
1977 Physics	Fundamental theoretical investigations of the electronic structure of magnetic systems	4	60	11,234	74,309	0.000	2.431
1981 Chemistry	Theories concerning the course of chemical reactions	2	0	2,103	0	0.000	624.344
1997 Medicine	Discovery of prions, a new biological principle of infection	8	52	1,282	54,670	2.445	-19.373
2001 Chemistry	Work on chirally catalysed hydrogenation reactions	11	310	578	208,922	0.000	32.703
2014 Chemistry	Development of super-resolved fluorescence microscopy	9	6	1,907	171	0.000	-32.684

Notes: For each Nobel prize paper, the table displays the year and field of the award, the prize motivation, the number of new grams and new word combinations introuced for the first time by the paper, and the number of times they are reused. Additionally, citation-based metrics are reported.

Validation: Nobel Prizes

	Nobel prize papers (n=501)				Control papers (n=501)				Cohen's d	t	Pr(|T| >|t|)
	Mean	Stdev	Min	Max	Mean	Stdev	Min	Max			
new word_binary	0.289	0.454	0	1	0.158	0.365	0	1	-0.32	-5.063	0.0000***
new_word	0.253	0.425	0	2,197	0.137	0.334	0	1,609	-0.302	-4.78	0.0000***
new_word_reuse	1.478	2.735	0	12,342	0.572	1.548	0	9,128	-0.408	-6.457	0.0000***
new_bigram_binary	0.599	0.491	0	1	0.375	0.485	0	1	-0.458	-7.256	0.0000***
new_bigram	0.754	0.749	0	3,332	0.425	0.618	0	2,398	-0.48	-7.591	0.0000***
new_bigram_reuse	2.904	2.921	0	12,423	1.469	2.22	0	9,454	-0.553	-8.756	0.0000***
new_trigram_binary	0.649	0.478	0	1	0.463	0.499	0	1	-0.38	-6.013	0.0000***
new_trigram	0.995	0.887	0	3,497	0.592	0.721	0	2,708	-0.499	-7.892	0.0000***
new_trigram_reuse	2.773	2.474	0	12,382	1.606	2.007	0	8,515	-0.518	-8.204	0.0000***
new_word_comb_binary (NC)	0.800	0.412	0	1	0.511	0.5	0	1	-0.639	-10.112	0.0000***
new_word_comb	2.771	1.991	0	7,644	1.61	1.925	0	6,463	-0.593	-9.389	0.0000***
new_word_comb_reuse	6.642	3.998	0	13,922	3.7	4.007	0	13,821	-0.735	-11.634	0.0000***
new_gram_bin (NG)	0.828	0.377	0	1.000	0.675	0.469	0.000	1.000	-0.361	-5.714	0. 0000**
new_gram	1.454	0.933	0	4.025	1.001	0.876	0.000	3.296	-0.500	-7.913	0. 0000**
new_gram_reuse	4.530	2.913	0	12.485	2.588	2.535	0.000	10.319	-0.711	-11.256	0.0000**
cosine_max	0.407	0.124	0.028	0.725	0.397	0.145	0.024	0.961	-0.076	-1.198	0.2312
cosine_avg	0.993	0.005	0.976	1	0.994	0.005	0.973	1	0.077	1.226	0.2205
wang	0.036	0.167	0	1,389	0.019	0.138	0	1.53	-0.109	-1.73	0.0840*
uzzi	6.949	0.751	6.207	8,505	7.157	0.771	6.307	9,748	0.274	4.329	0.0000***
cd	0.097	0.209	-0.522	0.974	0.042	0.145	-0.5	0.966	-0.307	-4.858	0.0000***

Notes: n=1,002 papers of which 501 Nobel prize papers and 501 matched control papers. Each Nobel prize paper is matched to one randomly selected control paper published in the same year, journal, volume and issue. All measures except binary indicators, the metrics based on cosine similarity and cd are log transformed after adding 1 for measures with 0 values. Cohen's d is the mean difference between award and control patents divided by the pooled standard deviation. *** p<0.01, ** p<0.05, * p<0.10

Validation: Nobel Prizes

- Likelihood of Nobel prizes: **Novelty**

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) new_word		0.607**									0.155		
		(0.251)									(0.228)		
(2) new_bigram			1.059***								0.491***		
			(0.179)								(0.190)		
(3) new_trigram				1.300***							0.801***		
				(0.190)							(0.199)		
(4) new_word_comb					0.578***						0.305***	0.291***	0.322***
					(0.111)						(0.113)	(112)	(0.114)
(5) new_gram						1.380***						1.067***	1.095***
						(0.183)						(0.185)	(0.184)
(6) cosine_max							-0.107						
							(0.702)						
(7) cosine_avg								-29.730*					
								(16.617)					
(8) wang									0.317				0.940
									(0.532)				(0.843)
(9) uzzi										-0.325***			-0.279***
										(0.071)			(0.075)
Pseudo R-squared	0.081	0.082	0.108	0.110	0.100	0.124	0.093	0.096	0.093	0.098	0.127	0.130	0.146
Precision (%)	62.29	62.67	66.06	65.21	65.11	66.40	64.41	64.12	64.06	64.45	67.15	66.53	69.38
Recall (%)	66.35	67.30	68.99	69.20	67.72	68.78	66.96	67.59	68.91	69.22	67.72	68.35	70.25
AUC	0.6691	0.6794	0.7136	0.7118	0.7038	0.7278	0.6995	0.6923	0.6963	0.6998	0.7307	0.7339	0.7469
Marginal Effect (%)		5.12	14.40	18.57	21.48	23.75	0	-3.20	1.10	-5.38			

Notes: Logit regression, robust standard errors between brackets. All measures, except metrics based on cosine similarity and cd are log transformed after adding 1 for measures with 0 values. All models include publication year and field of study fixed effects, and additionally control for whether the paper has an abstract available, text length, and the number of unique papers and journals cited by the focal paper. AUC is area under the ROC curve. Marginal effects are calculated as the % increase in the likelihood of being a Nobel prize paper associated with an increase in the metric with one standard deviation. *** p<0.01, ** p<0.05, * p<0.10

Validation: Nobel Prizes

- Likelihood of Nobel prizes: **Reuse**

	(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) new_word_reuse		0.178***						0.098**		
		(0.043)						(0.047)		
(2) new_bigram_reuse			0.305** *					0.128***		
			(0.041)					(0.049)		
(3) new_trigram_reuse				0.393***				0.252***		
				(0.054)				(0.060)		
(4) new_word_comb_reuse					0.280***			0.153***	0.197***	0.129***
					(0.036)			(0.040)	(0.047)	(0.039)
(5) new_gram_reuse						0.409***			0.277***	0.328***
						(0.047)			(0.047)	(0.050)
(7) cd								2.237***		2.249***
								(0.546)		(0.548)
Pseudo R-squared	0.081	0.106	0.116	0.132	0.154	0.125	0.107	0.102	0.107	0.109
Precision (%)	62.29	66.54	67.93	70.22	70.18	71.57	67.13	69.53	70.39	72.73
Recall (%)	66.35	72.01	71.64	73.32	74.16	73.53	70.38	71.73	71.73	72.57
AUC (%)	0.6691	0.7217	0.7560	0.7558	0.7621	0.7802	0.7231	0.7699	0.7729	0.7866
Marginal Effect (%)		8.75	16.93	18.21	20.38	22.61	9.67			

Notes: Logit regression, robust standard errors between brackets. All measures, except *cd* are log transformed after adding 1 for measures with 0 values. All models include publication year and field of study fixed effects, and additionally control for whether the paper has an abstract available, text length, and the number of unique papers and journals cited by the focal paper. Auc is area under the ROC curve. Marginal effects are calculated as the % increase in the likelihood of being a Nobel prize paper associated with an increase in the metric with one standard deviation. *** p<0.01, ** p<0.05, * p<0.10

Findings

- Text based measures can better identify novel ideas compared to citation based indicators
- Pool of different textual-metrics at paper level
- Validation of text-based metrics that better identify novel ideas compared to citation based indicators
- Interpretable novelty and impact indicators

Conclusions

- Limitations
  - we cannot use the full text due to the open access data restrictions
  - we do not have a true set of novel papers, which is why any validation exercise remains imperfect
- Robustness check
  - *Review papers* vs control papers
  - Analysis of only the title of publications (similar results)
- Future work
  - How new scientific ideas diffuse in patents?
  - How does the context (surrounding words) in which new scientific ideas appear and diffuse change?
  - *Novelty* vs *Interdisciplinarity*

Conclusions

We will provide open access to
all text metrics for all papers in MAG published until 2020 (n=72,529,264).

We hope our data will open up opportunities for future research.

Thank you for your attention and for your questions

nicola.melluso@kuleuven.be
sam.arts@kuleuven.be
reinhilde.veugelers@kuleuven.be