Tracing the Endless Scientific Frontier in Publication Text

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

- Current approaches to measure novelty and impact in science
  - Patterns of citations to prior art (references) (Uzzi et al., 2013; Wang et al., 2017)
  - Combinations of MeSH keywords (Azoulay et al., 2011; Boudreau et al., 2016; Carayol et al., 2016)
  - Expert assessment (Bornmann et al., 2019)
- 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?
  - Lack of external validity beyond the subjective assessments of typically small samples

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

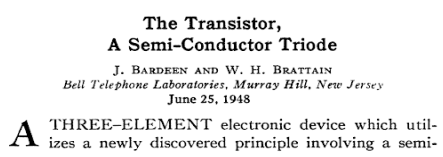
Data

- Microsoft Academic Graph (MAG) (now OpenAlex)
  - Coverage from 1780 until 2020 (about 270M publication records)
  - 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)

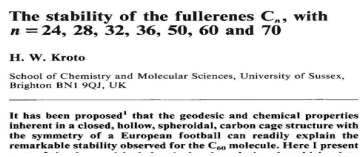


The Transistor,
A Semi-Conductor Triode

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 32,305)



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

H. W. Kroto

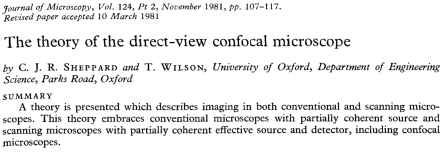
School of Chemistry and Molecular Sciences, University of Sussex, Brighton BN1 9QJ, UK

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

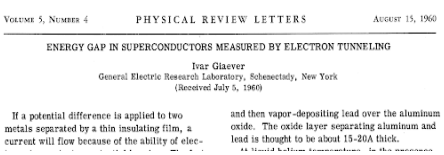
The theory of the direct-view confocal microscope

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)

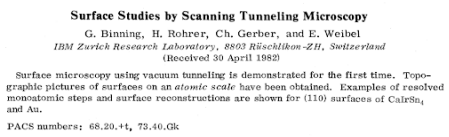
If a potential difference is applied to two metals separated by a thin insulating film, a current will flow because of the ability of elec-

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)



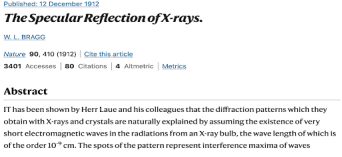
Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction

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

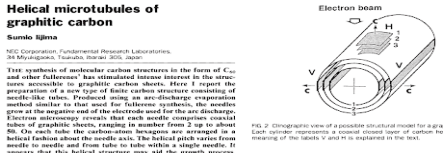
Nature **90**, 410 (1912) | Cite this article

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[1] 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 M is explained in the text.

Descriptives: Textual Units

		Reuse						
	#	**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. All textual units are reused at least once. *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

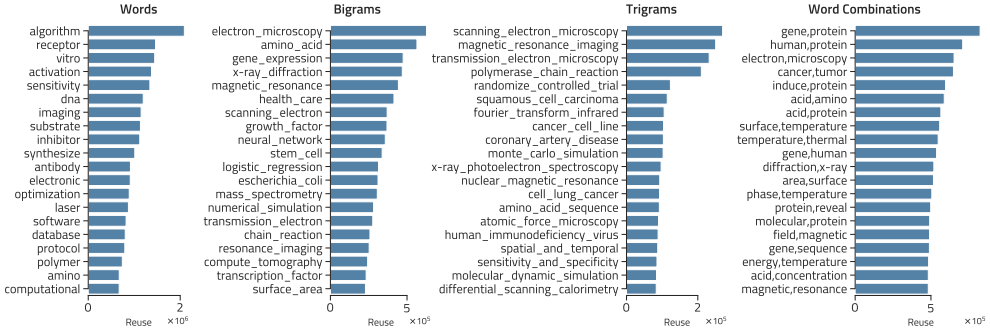


Figure: Top 20 reused words, bigrams, trigrams and word combinations.

Paper Level Textual Metrics

- **new_word | new_word_reuse**
  n. of unique new words introduced by the focal paper
- **new_bigram | new_bigram_reuse**
  n. of new bigrams that do not contain one of the new words introduced by the paper
- **new_trigram | new_trigram_reuse**
  n. of new trigrams that do not contain one of the new words or new bigrams introduced by the paper
- **new_gram | new_gram_reuse**
  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 | new_word_comb_reuse**
  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 Textual Metrics

	Mean	Std	Q1	Q2	Q3	Skewness
new_word_bin	0.079	0.27	0	0	0	3.123
new_word	0.101	0.409	0	0	0	8.227
new_word_reuse	7.67	970.971	0	0	0	751.848
new_bigram_bin	0.265	0.441	0	0	1	1.067
new_bigram	0.414	0.893	0	0	1	4.126
new_bigram_reuse	10.676	375.354	0	0	1	576.908
new_trigram_bin	0.451	0.498	0	0	1	0.198
new_trigram	0.954	1.52	0	0	1	2.762
new_trigram_reuse	7.548	133.301	0	0	3	671.247
new_gram_bin (NG)	0.545	0.498	0	1	1	-0.182
new_gram	1.469	2.224	0	1	2	3.256
new_gram_reuse	25.894	1076.944	0	1	5	601.745
new_word_comb_bin (NC)	0.511	0.5	0	1	1	-0.045
new_word_comb	14.523	58.996	0	1	10	33.718
new_word_comb_reuse	734.479	45453.872	0	1	38	688.475
NG = 1 & NC = 1	0.407	0.491	0	0	1	0.377
NG = 1 or NC = 1	0.654	0.476	0	1	1	-0.649

Notes: Descriptive table of textual metrics at paper level 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	0.523	0.711	1.000	0.855
new_bigram_bin	19,122,209	0.265	0.156	1.000	0.702	1.000	0.826
new_trigram_bin	32,571,551	0.451	0.124	0.412	1.000	1.000	0.76
new_gram_bin (NG)	39,390,166	0.545	0.145	0.485	0.827	1.000	0.738
new_word_comb_bin (NC)	36,942,016	0.511	0.132	0.428	0.670	0.787	1.000

Notes: Descriptive of the binary textual 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 textual metrics. Each column (except for "papers") represents the subset of papers (share of papers) that score positively for the particular binary metrics.

Condition	Measure	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	1.563	1.103	1.000	1.000	2.000	4.163
new_trigram_bin = 1	new_trigram	2.117	1.633	1.000	2.000	3.000	2.693
new_gram_bin = 1	new_gram	2.695	2.403	1.000	2.000	3.000	3.267
new_word_comb_bin = 1	new_word_comb	28.401	80.079	3.000	10.000	28.000	25.848

Notes: Summary statistics for textual 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

- Citation-based metrics
  - **wang**
    sum of the distance of novel combinations of cited journals (of cited papers) (Wang et al., 2017)
  - **uzzi**
    atypicality of all cited journal pairs (paper's 10-th percentile of z-score) (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)
- Textual Distance metrics (document embeddings of title and abstract)
  - **cosine_avg**
    1 - average cosine similarity with 5-years prior papers
  - **cosine_max**
    1 - maximum cosine similarity with 5-years prior papers

SPECTER document embedding (Cohan, 2020)

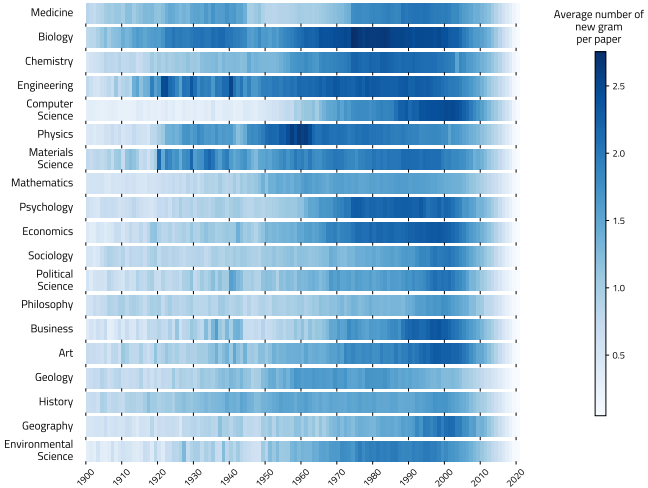
- transformer-based (BERT-like) language model tailored for scientific language (best performing model at the state-of-the-art)
- 768-dimensions vector representation of the title and abstract of a focal paper
- accounting for the context in which words appear (synonymity, polysemy, change in language...)

Benchmark: Paper Level Textual-based vs Citation-based Metrics

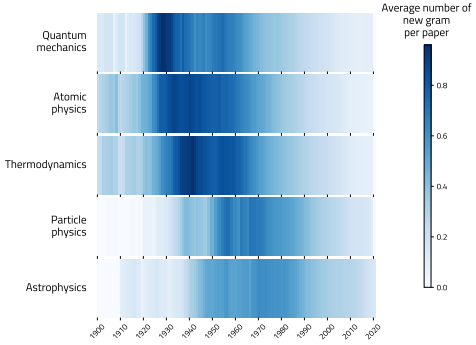
	cosine_max	cosine_avg	wang >0	uzzi <0	cd
NG = 1 & NC = 1	0.4291	0.9933	0.1246	0.3201	0.0049
NG = 1 or NC = 1	0.4231	0.9933	0.1142	0.2913	0.0043
NG = 1 & NC = 0	0.4073	0.9932	0.0987	0.2384	0.0036
NG = 0 & NC = 1	0.4210	0.9933	0.0958	0.2534	0.0035
NG = 0 & NC = 0	0.3889	0.9931	0.0515	0.1762	0.0019

Notes: Average values of *cosine_max, cosine_avg, wang >0, uzzi <0* and *cd* conditional on scoring positively on the textual novelty metrics. *NG = 1* indicates that the paper introduces at least one new gram (new word or new bigram or new trigram). *NC = 1* indicates that the paper introduces at least one new word combination (excluding the ones belonging to the new bigrams or trigrams). *wang >0* indicates that a paper is novel according to *wang. uzzi <0* indicates that a paper is novel according to *uzzi.*

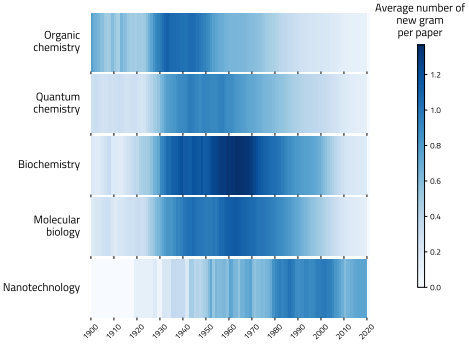
Novelty over time by fields



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 Nobel prize papers from control papers
    - T-test, Cohen's d
    - Precision, recall, AUC (area under the ROC-curve)

Validation: Nobel Prizes

- Binnig and Rohrer introduce for the first time the "scanning tunnel microscopy" for which they won the Nobel prize in Physics in 1986
  - The paper "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

The Nobel Prize in Physics 1986



Photo from the Nobel Foundation archive.
Ernst Ruska
Prize share: 1/2



Photo from the Nobel Foundation archive.
Gerd Binnig
Prize share: 1/4



Photo from the Nobel Foundation archive.
Heinrich Rohrer
Prize share: 1/4

The Nobel Prize in Physics 1986 was divided, one half awarded to Ernst Ruska "for his fundamental work in electron optics, and for the design of the first electron microscope", the other half jointly to Gerd Binnig and Heinrich Rohrer "for their design of the scanning tunneling microscope"

Validation: Nobel Prizes

89% of the Nobel papers introduce at least one new gram or new word combination that has been also found in their prize motivation summary page
(e.g. `https://www.nobelprize.org/prizes/physics/1986/summary/`)

Prize	Prize motivation	Paper Reference	new gram (reuse)	new word comb (reuse)
Physics 1923	work on the elementary charge of electricity and on the photoelectric effect	Millikan (1910)		charge,elementary (4107)
Physics 1936	Discovery of the positron	Anderson (1933)	positron (94,146)	
Physics 1937	discovery of the diffraction of electrons by crystals	Davisson & Germer (1927)	electron_diffraction (46307)	
Medicine 1952	discovery of streptomycin, the first antibiotic effective against tuberculosis	Schatz et al. (1944)	streptomycin (23334)	
Physics 1956	researches on semiconductors and their discovery of the transistor effect	Bardeen & Brattain (1948)	transistor (152047)	
Physics 1959	discovery of the antiproton	Chamberlain et al. (1955)	antiproton (5023)	
Medicine 1969	discoveries concerning the replication mechanism and the genetic structure of viruses	Hershey & Chase (1952)		nucleic,viral (8846)
Medicine 1971	discoveries concerning the mechanisms of the action of hormones	Butcher & Sutherland (1962)	nucleotide_phosphodiesterase (2863)	
Physics 1973	discoveries regarding tunneling phenomena in semiconductors and superconductors	Giaever (1960)	electron_tunneling (4759)	
Medicine 1979	development of computer assisted tomography	Hounsfield (1973)		sensitive,tomography (11498)
Chemistry 1981	theories concerning the course of chemical reactions	Fukui et al. (1952)		frontier,orbital (7301)
Medicine 1982	discoveries concerning prostaglandins and related biologically active substances	Hamberg et al. (1974)	prostaglandin_endoperoxide (1339)	
Medicine 1983	discovery of mobile genetic elements	McClintock (1950)		chromosome,instability (8896)
Physics 1986	work in electron optics, and for the design of the first electron microscope	Binning et al. (1982)	scanning_tunnel_microscopy (17914)	
Chemistry 1999	studies of the transition states of chemical reactions using femtosecond spectroscopy	Rose et al. (1988)		dissociation,femtosecond (907)
Chemistry 2001	work on chirally catalysed hydrogenation reactions	Nozaki et al. (1968)		chiral,synthesis (34820)
Physics 2005	contribution to the quantum theory of optical coherence	Glauber (1963)	optical_coherence (50175)	
Medicine 2016	discoveries of mechanisms for autophagy	Takeshige et al. (1992)		autophagy,genetic (3396)

Notes: New grams (new words or new bigram or new trigram) and new word combinations introduced for the first time by papers linked to Nobel prize winners and also found in the Nobel prize motivation summary page.

Validation: Nobel Prizes

Prize	Prize motivation	new_gram	new_word_comb	new_gram_reuse	new_word_comb_reuse	wang	uzzi
Physics 1923	work on the elementary charge of electricity and on the photoelectric effect	1	23	46	25,680	0.000	12.233
Physics 1936	Discovery of the positron	9	17	102,119	16,576	0.000	1.023
Physics 1937	discovery of the diffraction of electrons by crystals	24	495	87,407	612,942	0.000	32.241
Medicine 1952	discovery of streptomycin, the first antibiotic effective against tuberculosis	3	19	23,339	60,314	0.000	7.342
Physics 1956	researches on semiconductors and their discovery of the transistor effect	2	1	152,048	3	0.000	8.222
Physics 1959	discovery of the antiproton	1	2	5,023	17	0.000	11.342
Medicine 1969	discoveries concerning the replication mechanism and the genetic structure of viruses	50	1,066	20,540	640,243	0.000	-0.234
Medicine 1971	discoveries concerning the mechanisms of the action of hormones	3	2	2,918	5,237	0.342	-1.680
Physics 1973	discoveries regarding tunneling phenomena in semiconductors and superconductors	1	2	4,759	6,086	0.000	21.234
Medicine 1979	development of computer assisted tomography	2	12	9	36,560	0.000	75.557
Chemistry 1981	theories concerning the course of chemical reactions	7	59	1,368	50,970	0.000	-3.761
Medicine 1982	discoveries concerning prostaglandins and related biologically active substances	10	73	2,299	7,100	1.130	-3.184
Medicine 1983	discovery of mobile genetic elements	11	94	486	24,001	0.000	35.646
Physics 1986	work in electron optics, and for the design of the first electron microscope	2	7	17,927	1,387	0.000	6.994
Chemistry 1999	studies of the transition states of chemical reactions using femtosecond spectroscopy	0	1	0	907	2.442	-5.019
Chemistry 2001	work on chirally catalysed hydrogenation reactions	10	307	577	205,048	0.000	32.276
Physics 2005	contribution to the quantum theory of optical coherence	9	55	50,624	8,334	0.000	1.234
Medicine 2016	discoveries of mechanisms for autophagy	8	133	144	7,230	0.000	1.651

Notes: Novelty metrics scores (textual and citation based) for a sample of Nobel prize papers.

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	0.297	0.458	0.000	1.000	0.184	0.388	0.000	1.000	-0.268	-4.247	0.0000***
new_word	0.260	0.430	0.000	2.197	0.154	0.341	0.000	1.609	-0.273	-4.326	0.0000***
new_word_reuse	1.485	2.736	0.000	12.343	0.590	1.545	0.000	9.128	-0.403	-6.375	0.0000***
new_bigram_binary	0.599	0.491	0.000	1.000	0.425	0.495	0.000	1.000	-0.352	-5.578	0.0000***
new_bigram	0.698	0.695	0.000	3.045	0.430	0.562	0.000	2.303	-0.423	-6.693	0.0000***
new_bigram_reuse	2.909	2.911	0.000	11.284	1.441	2.148	0.000	9.495	-0.574	-9.084	0.0000***
new_trigram_binary	0.727	0.446	0.000	1.000	0.623	0.485	0.000	1.000	-0.223	-3.525	0.0000***
new_trigram	1.012	0.795	0.000	3.434	0.701	0.661	0.000	2.708	-0.425	-6.728	0.0000***
new_trigram_reuse	2.866	2.335	0.000	9.793	1.740	1.871	0.000	8.732	-0.532	-8.423	0.0000***
new_word_comb_binary (NC)	0.804	0.397	0.000	1.000	0.663	0.473	0.000	1.000	-0.324	-5.135	0.0000***
new_word_comb	2.718	1.946	0.000	7.424	1.866	1.822	0.000	6.438	-0.452	-7.157	0.0000***
new_word_comb_reuse	6.654	4.055	0.000	15.230	4.145	3.800	0.000	13.849	-0.639	-10.107	0.0000***
new_gram_bin (NG)	0.828	0.377	0.000	1.000	0.721	0.449	0.000	1.000	-0.260	-4.112	0. 0000**
new_gram	1.388	0.895	0.000	3.932	0.970	0.793	0.000	2.944	-0.494	-7.826	0. 0000**
new_gram_reuse	4.462	2.878	0.000	12.485	2.554	2.358	0.000	9.512	-0.725	-11.478	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	1.000	0.994	0.005	0.973	1.000	0.077	1.226	0.2205
wang	0.036	0.167	0.000	1,389	0.019	0.138	0.000	1.53	-0.109	-1.730	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		0.456**									0.074		
		(0.226)									(0.238)		
(2) new_bigram			0.775***								0.394**		
			(0.164)								(0.186)		
(3) new_trigram				0.698***							0.407**		
				(0.161)							(0.177)		
(4) new_word_comb					0.450***						0.302***	0.287***	0.287***
					(0.082)						(0.092)	(0.091)	(0.091)
(5) new_gram						0.940***						0.689***	0.694***
						(0.165)						(0.180)	(0.179)
(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	0.093	0.095	0.112	0.109	0.120	0.124	0.093	0.094	0.094	0.098	0.129	0.132	0.138
Precision (%)	64.41	64.26	65.82	64.31	65.94	65.22	64.60	64.66	64.41	64.60	67.20	67.26	67.70
Recall (%)	69.96	69.12	70.38	68.91	69.96	69.33	70.17	69.96	69.96	72.06	71.01	71.22	73.11
AUC	0.6966	0.6982	0.7176	0.7110	0.7234	0.7295	0.6968	0.6979	0.6967	0.6996	0.7328	0.7369	0.7401
Marginal Effect (%)		2.72	10.74	11.26	18.27	17.24	-0.60	-1.84	1.12	-5.13			

Notes: Logit regression, robust standard errors between brackets. All measures, except metrics based on cosine similarity 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		0.157***						0.108**		
		(0.036)						(0.038)		
(2) new_bigram_reuse			0.266** *					0.126***		
			(0.037)					(0.042)		
(3) new_trigram_reuse				0.284***				0.131***		
				(0.046)				(0.050)		
(4) new_word_comb_reuse					0.278***			0.197***	0.193***	0.195***
					(0.032)			(0.036)	(0.035)	(0.036)
(5) new_gram_reuse						0.346***			0.244***	0.235***
						(0.039)			(0.042)	(0.041)
(7) cd							2.517***			2.329***
							(0.579)			(0.616)
Pseudo R-squared	0.093	0.107	0.138	0.126	0.165	0.167	0.116	0.186	0.193	0.210
Precision (%)	64.41	64.44	68.14	67.54	70.06	69.98	67.27	71.02	71.49	71.89
Recall (%)	69.96	67.02	71.43	70.38	73.74	72.48	70.80	73.11	73.74	75.21
AUC (%)	0.6966	0.7086	0.7418	0.7315	0.7647	0.7651	0.7234	0.7792	0.7844	0.7961
Marginal Effect (%)		7.71	14.69	13.16	22.56	19.10	9.83			

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 seem to better identify novel ideas compared to citation based indicators
- Impact and diffusion of novelty with text-based measures can be studied through their reuse rather than only citations
- Pool of different textual-based measures at paper level
- 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