



ARL-SR-0493 • FEB 2024



From Scientific Institutions to Research Concepts to Global Collaboration and Beyond: Uncovering Patterns in Scientific Networks

by Mark A Tschopp

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



From Scientific Institutions to Research Concepts to Global Collaboration and Beyond: Uncovering Patterns in Scientific Networks

by Mark A Tschopp
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE		2. REPORT TYPE		3. DATES COVERED	
February 2024		Special Report		START DATE 01/01/2023	END DATE 12/31/2023
4. TITLE AND SUBTITLE From Scientific Institutions to Research Concepts to Global Collaboration and Beyond: Uncovering Patterns in Scientific Networks					
5a. CONTRACT NUMBER		5b. GRANT NUMBER		5c. PROGRAM ELEMENT NUMBER	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER	
6. AUTHOR(S) Mark A Tschopp					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLA-B Aberdeen Proving Ground, MD 21005-5066				8. PERFORMING ORGANIZATION REPORT NUMBER ARL-SR-0493	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)		11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.					
13. SUPPLEMENTARY NOTES ORCID IDs: 0000-0001-8471-5035; email: <mark.a.tschopp.civ@army.mil>.					
14. ABSTRACT This final report highlights work associated with the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) sabbatical program, awarded to Dr Mark Tschopp for the 2023 calendar year. The motivation behind this sabbatical is to further develop skills in data science, natural language processing, machine learning, and artificial intelligence, in addition to exploring the interdisciplinary field of the science of science. This report aims to encapsulate the essence of the sabbatical. The three independent studies included herein are related to (1) many domains of global collaboration (research concepts, institutions, temporal evolution, etc.), (2) hype words and promotional language in NSF title/abstracts, and (3) funding agency-to-article linkage (study on data quality in open-source databases and in articles). Additional activities conducted during the sabbatical included efforts to disseminate knowledge in machine learning and artificial intelligence within the broader DOD workforce.					
15. SUBJECT TERMS sabbatical, science of science, data science, machine learning, artificial intelligence					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 213	
a. REPORT UNCLASSIFIED	b. ABSTRACT UNCLASSIFIED	c. THIS PAGE UNCLASSIFIED			
19a. NAME OF RESPONSIBLE PERSON Mark A Tschopp				19b. PHONE NUMBER (Include area code) 301-938-5389	

STANDARD FORM 298 (REV. 5/2020)

Prescribed by ANSI Std. Z39.18

Contents

List of Figures	vi
List of Tables	x
Acknowledgments	xi
Executive Summary	xii
1. Introduction	1
2. Methods	2
2.1 A Brief Introduction to OpenAlex	2
2.2 Navigating the JSON Format	4
2.3 API Calls within OpenAlex	5
3. Research Study: Global Collaboration in Science	7
3.1 Introduction	7
3.2 Methods	9
3.3 Results and Discussion	10
3.3.1 Temporal Evolution of Global Science	10
3.3.2 Temporal Evolution of Scientific Productivity for Top Countries	15
3.3.3 Frequency of Countries, Institutions, Authors, and Concepts	22
3.3.4 Global Collaboration between Leading Countries	25
3.3.5 Global Collaboration between Leading Countries over Time	31
3.3.6 US Institution Collaboration with Leading Countries	34
3.3.7 Global Collaboration for Research Concepts	39
3.3.8 Interconnectedness of Research Concepts	48
3.3.9 Global Collaboration between Leading Countries within Specific Research Areas	60
3.4 Concluding Remarks: Insights and Future Research Directions	65
4. Research Study: Revolutionary Study on the Unprecedented Use of Novel Hype and Promotional Language in Funded Agreements	67

4.1	Exceptional Introduction	67
4.2	Pioneering Methods	70
4.3	Paradigm-shifting Results and Discussion	74
4.3.1	Identifying Hype in Titles and Abstracts	74
4.3.2	NSF Title and Abstract Distributions	76
4.3.3	Hype Maps within NSF Titles and Abstracts	76
4.3.4	Top Hype Words within NSF Titles and Abstracts	80
4.3.5	Temporal Evolution of Hype Words through Time	82
4.3.6	Hype Words in Books – Google Books Ngrams	90
4.3.7	Hype Word Comparison with Google Books Ngrams	92
4.3.8	Hype Word Temporal Evolution over Time	95
4.3.9	Using Large Language Models to Focus on Content	101
4.4	Concluding Remarks: Insights and Future Research Directions	102
5.	Research Study: Understanding Funding of Science from Open Scientific Databases	104
5.1	Introduction	104
5.2	Methodology	107
5.3	Results and Discussion	109
5.3.1	Data Quality for Funder Award IDs in Open Databases	109
5.3.2	Data Cleaning of Award IDs, Part I	110
5.3.3	Data Cleaning of Award IDs, Part II	112
5.3.4	Data Cleaning of Award IDs, Part III	114
5.3.5	Data Cleaning of Award IDs, Part IV	116
5.3.6	Frequency of Articles per Award and Awards per Article	116
5.3.7	Distribution of Awards, Articles and Citations by Award FY	118
5.3.8	Distribution of Awards, Articles and Citations by Award Funding Amount	122
5.3.9	Distribution of Cost per Article and Cost per Citation by Award Funding Amount	123
5.3.10	Discussion	127
5.4	Concluding Remarks: Insights and Future Research Directions	129

6. Courses: Machine Learning and Artificial Intelligence 101	132
6.1 Machine Learning 101 Short Course	132
6.2 One-Minute Machine Learning: The Book	140
6.3 Artificial Intelligence 101 Short Course	143
6.4 Machine Learning Operations, MLOps	145
7. Measurable Deliverables	147
7.1 Invited Seminars and Talks	147
7.2 Short Courses	149
7.3 Conferences, Workshops, Events	149
7.4 VIP Engagement	150
8. Conclusion	151
9. References	155
Appendix A. Interconnectedness of Level 0 and Level 1 Concepts	158
Appendix B. Title Hype Word Comparisons by Hype Category	178
Appendix C. Abstract Hype Word Comparisons by Hype Category	186
List of Symbols, Abbreviations, and Acronyms	194

List of Figures

Fig. 1	Interconnected entities in OpenAlex	3
Fig. 2	JSON format for OpenAlex: meta	5
Fig. 3	JSON format for OpenAlex: results	6
Fig. 4	Evolution of scientific articles since 1900	11
Fig. 5	Evolution of book chapters since 1900	12
Fig. 6	Evolution of science books since 1900	13
Fig. 7	Global scientific article contributions by country	15
Fig. 8	Evolution of global articles with US affiliations	16
Fig. 9	Evolution of global articles with China affiliations	17
Fig. 10	Evolution of global articles with UK affiliations	18
Fig. 11	Evolution of global articles with Japan affiliations	19
Fig. 12	Evolution of global articles with German affiliations	20
Fig. 13	Temporal evolution of global collaboration	21
Fig. 14	Distribution of number of distinct countries for articles	22
Fig. 15	Distribution of number of distinct institutions for articles	23
Fig. 16	Distribution of number of authors for articles	24
Fig. 17	Distribution of number of concepts for articles	25
Fig. 18	Global collaboration heat map for all articles	26
Fig. 19	Global collaboration heat map of articles in English	29
Fig. 20	Venn diagram for multi-country collaborations	30
Fig. 21	Global collaboration heat map (1990s)	31
Fig. 22	Global collaboration heat map (2000s)	32
Fig. 23	Global collaboration heat map (2010s)	33
Fig. 24	US federal institutions collaboration heat map (2018–2023)	35
Fig. 25	US top academic institutions collaboration heat map	36
Fig. 26	US collaborative academic institutions collaboration heat map	37
Fig. 27	US agency funding heat map (2018–2023)	38
Fig. 28	Material science concept connections (Sankey plot)	41
Fig. 29	Computer science concept connections (Sankey plot)	42
Fig. 30	Global collaborations in level 0 concept areas	44

Fig. 31	US collaborations in level 0 concept areas.....	45
Fig. 32	China collaborations in level 0 concept areas	46
Fig. 33	UK collaborations in level 0 concept areas	47
Fig. 34	Interconnectedness of business concepts (Sankey plot)	49
Fig. 35	Interconnectedness of computer science concepts (Sankey plot).....	50
Fig. 36	Level 0 concept co-occurrence heat map (2018–present)	52
Fig. 37	Level 0 concept co-occurrence heat map (1998-2007)	53
Fig. 38	Level 0 concept co-occurrence heat map (2008-2017)	54
Fig. 39	Schematic of Venn diagrams for indices	56
Fig. 40	Jaccard index for concept collaboration heat map.....	57
Fig. 41	Sørensen-Dice coefficient for concept collaboration heat map	58
Fig. 42	Overlap coefficient for concept collaboration heat map.....	59
Fig. 43	Global collaboration heat map for AI and ML	62
Fig. 44	Global collaboration heat map for nanotechnology	63
Fig. 45	Global collaboration heat map for quantum	64
Fig. 46	Word cloud of 139 hype and promotional adjectives	69
Fig. 47	XML format for NSF awards.....	71
Fig. 48	Example of award abstract with hype words	75
Fig. 49	Histograms of word count distributions in NSF awards.....	77
Fig. 50	Tree map of hype words in NSF award titles (2018–2022).....	78
Fig. 51	Tree map of hype words in NSF award abstracts (2018–2022).....	79
Fig. 52	Tree map of hype words in NSF award abstracts (2018–2022).....	80
Fig. 53	Stacked bar chart of temporal evolution of hype	87
Fig. 54	Stacked bar chart of temporal evolution of hype (normalized)	88
Fig. 55	Dashboard for hype language within NSF dataset	89
Fig. 56	Sunburst diagram for Google Books Ngrams of hype words	91
Fig. 57	Yearly frequencies for hype word: scalable.....	95
Fig. 58	Yearly frequencies for “scale” hype words	97
Fig. 59	Yearly frequencies for hype words	99
Fig. 60	Yearly frequencies for hype words	100
Fig. 61	Rewriting the most-hyped NSF award abstract of 2022.....	102

Fig. 62	Grants, works, citations for Army funders	106
Fig. 63	Grants, works, citations pie chart for Army funders	106
Fig. 64	Funding section within OpenAlex and CrossRef	108
Fig. 65	Initial pie chart for award IDs	111
Fig. 66	Pie chart for award IDs after cleaning #1	112
Fig. 67	Pie chart for award IDs after cleaning #2	114
Fig. 68	Pie chart for award IDs after cleaning #3	115
Fig. 69	Final pie chart for award IDs	117
Fig. 70	Probability distribution of articles and grants for ARO funding	118
Fig. 71	Distribution of awards, articles and citations by fiscal year	121
Fig. 72	Familiarity with ML prior to course.....	132
Fig. 73	Word map of course participant position titles	133
Fig. 74	Word map of course participant organizations	134
Fig. 75	ML 101 course outline and concepts map	135
Fig. 76	ML 101 NLP section added.....	137
Fig. 77	Anonymous course ratings for ML for Everyone 2023	138
Fig. 78	Machine learning book based on ML101 course	141
Fig. 79	Machine learning book with Edward Tufte template	142
Fig. 80	AI course outline	143
Fig. 81	AI books	144
Fig. A-1	Interconnectedness of art concepts (Sankey plot)	159
Fig. A-2	Interconnectedness of biology concepts (Sankey plot)	160
Fig. A-3	Interconnectedness of business concepts (Sankey plot)	161
Fig. A-4	Interconnectedness of chemistry concepts (Sankey plot)	162
Fig. A-5	Interconnectedness of computer science concepts (Sankey plot).....	163
Fig. A-6	Interconnectedness of economics concepts (Sankey plot)	164
Fig. A-7	Interconnectedness of engineering concepts (Sankey plot).....	165
Fig. A-8	Interconnectedness of environmental science concepts	166
Fig. A-9	Interconnectedness of geography concepts (Sankey plot)	167
Fig. A-10	Interconnectedness of geology concepts (Sankey plot)	168
Fig. A-11	Interconnectedness of history concepts (Sankey plot).....	169

Fig. A-12 Interconnectedness of materials science concepts (Sankey plot)	170
Fig. A-13 Interconnectedness of mathematics concepts (Sankey plot).....	171
Fig. A-14 Interconnectedness of medicine concepts (Sankey plot)	172
Fig. A-15 Interconnectedness of philosophy concepts (Sankey plot).....	173
Fig. A-16 Interconnectedness of physics concepts (Sankey plot)	174
Fig. A-17 Interconnectedness of political science concepts (Sankey plot)	175
Fig. A-18 Interconnectedness of psychology concepts (Sankey plot)	176
Fig. A-19 Interconnectedness of sociology concepts (Sankey plot)	177

List of Tables

Table 1	Entities and URLs within OpenAlex	4
Table 2	Accessing OpenAlex API example	7
Table 3	Country breakdown for total/collaborative articles	14
Table 4	139 hype words and categories	72
Table 5	40 most frequent hype words and categories (2018–2022)	81
Table 6	Temporal change in top 40 hype words	84
Table 7	Temporal change in bottom 40 hype words	85
Table 8	Difference in hype words from 1988–1992 to 2018–2022	94
Table 9	Articles funded by ARO grant with incorrect award ID	110
Table 10	Error variant examples on W911NF from award ID	112
Table 11	ARO awards, articles, and citations by award funding.....	122
Table 12	Article and citation cost by award funding (average)	124
Table 13	Article and citation cost by award funding (median).....	125
Table B-1	Attitude category: hype word comparison in NSF titles	179
Table B-2	Importance category: hype word comparison in NSF titles	180
Table B-3	Novelty category: hype word comparison in NSF titles	181
Table B-4	Problem category: hype word comparison in NSF titles	181
Table B-5	Quality category: hype word comparison in NSF titles	182
Table B-6	Rigor category: hype word comparison in NSF titles	183
Table B-7	Scale category: hype word comparison in NSF titles	184
Table B-8	Utility category: hype word comparison in NSF titles	185
Table C-1	Attitude category: hype word comparison in NSF abstracts	187
Table C-2	Importance category: hype word comparison in NSF abstracts	188
Table C-3	Novelty category: hype word comparison in NSF abstracts	189
Table C-4	Problem category: hype word comparison in NSF abstracts	189
Table C-5	Quality category: hype word comparison in NSF abstracts	190
Table C-6	Rigor category: hype word comparison in NSF abstracts	191
Table C-7	Scale category: hype word comparison in NSF abstracts	192
Table C-8	Utility category: hype word comparison in NSF abstracts	193

Acknowledgments

I extend my heartfelt gratitude to the US Army Combat Capabilities Development Command Army Research Laboratory (ARL) for providing an enriching and transformative sabbatical experience. The opportunity to embark on this journey, immerse myself in a new environment, acquire new skills, and engage in creative thinking has been truly exceptional. It was not uncommon to encounter surprised expressions from both colleagues within the ARL community and those outside of it, with remarks like, “Wait, what?! ARL has a sabbatical program?” Frequently, the response internally centered on the perceived lack of time for a sabbatical, weighed down by existing responsibilities and ongoing programs. However, I firmly believe that there is rarely a perfect time for career-enhancing endeavors; sometimes, one must seize the moment and take the leap to grow and evolve.

I also extend my sincere thanks to the students, postdoctoral researchers, faculty, and staff at Northwestern University and the University of Chicago for their invaluable contributions to the discussions that have greatly enriched my sabbatical experience. Their collective insights have played a pivotal role in shaping my understanding and fostering a newfound appreciation for diverse research realms, including computational social science, network science, machine learning, and generative artificial intelligence. I am particularly grateful to Professor Brian Uzzi at Northwestern University and Professors James Evans and Rebecca Willett at the University of Chicago for their warm welcome into their respective groups, labs, and centers. Engaging in discussions with them and their group members has significantly advanced my comprehension of the current state of the art in the various domains discussed herein.

Executive Summary

In an era marked by the rapid transformation of science and technology through data science, analytics, machine learning, and artificial intelligence, organizations are compelled to adopt data-driven decision-making practices to stay competitive. This involves not only interrogating data through descriptive analytics but also diagnosing data using diagnostic analytics, making predictions through predictive analytics, and implementing changes through prescriptive analytics. Each domain safeguards its datasets, holding valuable insights and knowledge ready for exploration. The motivation behind this sabbatical is to further develop skills in data science, natural language processing, machine learning, and artificial intelligence, extending their applications across the interdisciplinary realm of the science of science. This executive summary aims to encapsulate the essence of a comprehensive report derived from a sabbatical dedicated to multiple independent studies (three research studies are detailed within). It also outlines additional activities conducted during the sabbatical, including efforts to disseminate knowledge in machine learning and artificial intelligence within the broader DOD workforce.

Science of Science: Central to this sabbatical is an examination of the science of science, an innovative field aiming to unravel complex relationships across scientific domains. It delves into the extraction of valuable knowledge and insights by exploring expansive networks (e.g., over 200 million articles, over 1 billion citations) within each scientific domain. This exploration encompasses scientists, institutions, articles, grants, citations, and patents, revealing the interconnectedness of data across diverse domains.

Unlocking the Value of Open-Source Data: The three research studies detailed herein serve as exemplars, highlighting the intrinsic value of open-source data and the current ease of accessing information. While these studies offer a glimpse into the potential, it is evident that research and development organizations can harness open-source datasets to drive innovation, promote collaboration, and refine decision-making processes. This is achieved through understanding critical entities such as scientists, research output, collaborations, funding sources, uniqueness, and scientific impact. The pursuit of knowledge and insight within these comprehensive datasets extends beyond academic exercise; it is imperative for organizations seeking to thrive in the dynamic and swiftly evolving data-rich landscape.

Key Findings: The forthcoming bullets of this executive summary unpack specific insights derived from the three research studies, shedding light on discoveries during the course of these investigations.

- *Global Collaboration.* The exponential growth of scientific articles, doubling approximately every 14 years, is mirrored by a noteworthy shift in collaborative efforts. Presently, around 20% of all scientific articles involve collaborations across multiple countries. While the US has traditionally been a common collaborative partner, this trend is evolving, with China emerging as a collaborative force not only for top countries but also with US federal institutions, academic entities, and projects funded by US agencies. This study systematically quantifies these evolving global collaboration dynamics, along with other focused examinations such as exploring critical research domains of artificial intelligence, quantum science, and nanotechnology.
- *Hype Language in National Science Foundation (NSF) Awards.* The dynamics of the award process and science communication have witnessed a substantial influence from hype language. A comprehensive analysis of 139 hype words across over 480,000 NSF awards spanning four decades reveals a remarkable surge in the frequency of hype language in titles and abstracts over the last 30 years—some cases experiencing an increase exceeding an order of magnitude (10 words). This escalates to 39 hype words surpassing an order of magnitude when considering the base rate of change in digitized books. Many of these fall into the utility and scale hype category, emphasizing the perceived usefulness and broad scale of the research. The study concludes with a perspective on how large language models may play a future role in mitigating this growing trend in scientific discourse.
- *Funding Agencies: from Awards to Articles.* The nexus between funding agencies and their return on investment lies in the intricate relationship with the scientific knowledge generated, primarily disseminated through scientific articles, and the impact of that science, often measured through ensuing citations. This study delves into extracting award ID information from open-source databases, elucidating the linkage from funding to articles to citations for the Army Research Office. A notable aspect is the agency award ID in the acknowledgment section of articles, where only 74% conform to correct

formats, necessitating cleaning operations for the 15% of incorrectly formatted award IDs (with 10% entirely lacking an award ID). With refinements pushing the compliance rate to 85%, this linkage provides insights into the number of articles and citations per award, the associated cost for each metric, and further downstream analysis.

Educating a Broad Workforce in Machine Learning and Artificial Intelligence:

Empowering the DOD workforce with advanced data analytics methods like machine learning and artificial intelligence is crucial for navigating the evolving landscape. Despite the availability of numerous courses, often attended by only a fraction of the total workforce, the challenge persists in ensuring a comprehensive understanding of the concepts of machine learning and artificial intelligence across diverse staff members. This report provides a brief overview of a machine learning 101 and artificial intelligence 101 short course developed by Dr. Mark Tschopp. Complementary to other efforts within the Army, this course plays a pivotal role in the essential task of educating and preparing the Army and DOD workforce for the challenges and opportunities presented by the data-driven future.

Key Takeaways: In the current landscape, open-source data is a valuable resource, but organizations often hold an equally valuable wealth of internal data. In some instances, integrating both external and internal knowledge sources can unlock deeper insight and transformative outcomes for organizations. The combination of these data sets can be crucial for maximizing additive value. This transformative potential only lies through carefully curating data, strategically aggregating, fostering a holistic understanding through data science, and predicting responses and outcomes via machine learning and artificial intelligence. Forward-thinking R&D organizations are progressively incorporating data scientists with this expertise, unlocking profound insights into their core business functions.

This report emphasizes how R&D organizations can strategically leverage the growing field of science of science, thereby fostering a deeper comprehension of organizational uniqueness, their roles in the broader scientific landscape, and the impact of their work in the open scientific domain. The evolving science operational environment, fueled by rapid advancements in machine learning and artificial intelligence, underscores the need for acquiring talent and organizational skills that bridge the gap between data and science.

For those examining this report through a DOD lens, questions like “why focus on hype language?” or “why not use DOD datasets?” or “how can these findings be applied?” may naturally arise. First, the studies chosen here are driven by curiosity, inspired by recently-published scientific studies, conversations within research groups, or personal interest. They represent only a subset of the many studies pursued during this sabbatical. For instance, the hype language study was selected to further develop skills in aggregating large amounts of text data, utilizing natural language processing techniques for data cleaning, and analyzing the results—while the topic itself was both interesting and timely.

The primary goal of the sabbatical was to immerse in a new environment, acquire new skills, and engage with external researchers. While DOD datasets were explored in many instances, they may not be included in this report for various reasons. Regarding the application of these findings, the analyses presented are straightforward and can be easily adapted to address diverse research questions. With the increasing use of APIs and other tools for sharing curated data, such analyses can be tailored to explore various organizational aspects. For example, one could investigate global collaboration partners, track the evolution of these collaborations, assess their impact in terms of citations, identify partners contributing to greater impact, examine the most collaborative research areas, and make comparisons with other similar organizations. In essence, this report aims to showcase the potential of open-source datasets available today while inspiring new questions for future research endeavors to answer.

1. Introduction

Scientific curiosity, the fundamental pursuit of “why,” has propelled the technological landscape witnessed today. Across diverse domains, each new scientific discovery does not halt progress; instead, it unfurls deeper questions and prompts contemplation on how to translate these findings into broader impact. This perpetual cycle fuels a continuous influx of ideas, driven by the exponential growth of rich data, innovative techniques, advanced tools, and the enhanced skillsets of countless scientists. Collectively, they shape the shared understanding of the world. In the current era marked by unprecedented global collaboration and technological strides, science assumes a pivotal role, prompting researchers to scrutinize the intricate workings of the complex scientific ecosystem.

The inquiry of how the scientific enterprise works has given rise to a thriving trans-disciplinary field of study known as the Science of Science.^{1,2} This interdisciplinary domain employs computational methods to dissect and comprehend the network structure, dynamics, and impact of scientific research. Rooted in the analysis of patterns governing the production, dissemination, and consumption of scientific knowledge, the Science of Science integrates insights from diverse disciplines, including sociology, information science, network science, data analytics, and, more recently, advances in machine learning (ML) and artificial intelligence (AI). Researchers within this field aim to unveil the underlying mechanisms steering scientific discovery and innovation. From investigating scientists’ careers and collaboration dynamics to dissecting citation patterns and investigating the roles of funding agencies and policy, the Science of Science offers a holistic lens with which to view science from a 30,000-ft view. It provides researchers with a comprehensive view of the scientific enterprise, shedding light on what works and what does not work. As this discipline grows, it not only enriches the comprehension of knowledge generation and dissemination but also provides a high-altitude perspective for policymakers, funding agencies, and scientists alike, fostering a more informed and collaborative scientific landscape.

Some of the biggest advances in driving this new field forward is the development of large fully-open datasets that allow researchers to finally answer questions about the science ecosystem at scale through easy-to-use interfaces. Microsoft Academic Graph was perhaps the first open heterogeneous graph containing scientific records,

citation relationships, authors, institutions, journals, conferences and field,^{3,4} and this database has been carried on with OpenAlex.⁵ Needless to say, with a dataset of millions of grants, articles, authors, and billions of citations, geographical boundaries no longer constrain the pursuit of knowledge into the scientific enterprise. This report explores a few studies that encounter both the complexities with using these databases as well as the extraordinary usefulness of them: (1) investigating the roles of funding agencies with OpenAlex, (2) studying the hype language surrounding prestigious awards with National Science Foundation (NSF) database,^{6,7} and (3) diving into the transformative impact of open databases on connecting awards to scientific articles.

The objective of this report is to understand the ability of open scientific databases for quantifying the impact of scientists, teams, institutions, funding agencies, countries, and articles. In this report, a brief introduction to extracting information from these open-access scientific databases will first be discussed, including how to use the application programming interfaces (APIs) to communicate directly with these open databases. Then, multiple research studies are pursued to show the utility and quality of these datasets: global collaboration for top countries, hype language in NSF awards, and the usefulness of funding agency award info. A section describing several short introductory courses developed on the topics and concepts of ML and AI are presented. Last, some measurable deliverables and outcomes are listed: invited seminars and talks, conferences and workshops attended, invited panels participated in, short courses taught, etc. This final report details the studies, the methods, the results, and the deliverables from this sabbatical.

2. Methods

2.1 A Brief Introduction to OpenAlex

OpenAlex is an openly accessible catalog of the global research system. At its core, the OpenAlex dataset describes various scholarly entities and their interconnected relationships, forming an extensive network or, more technically, a heterogeneous directed graph. This dataset includes entities such as works, authors, sources, institutions, concepts, publishers, and funders, creating a data representation of the global research landscape, as illustrated with Fig. 1.

As of the latest update on January 30, 2024, OpenAlex boasts impressive cover-

Interconnected Entities in OpenAlex

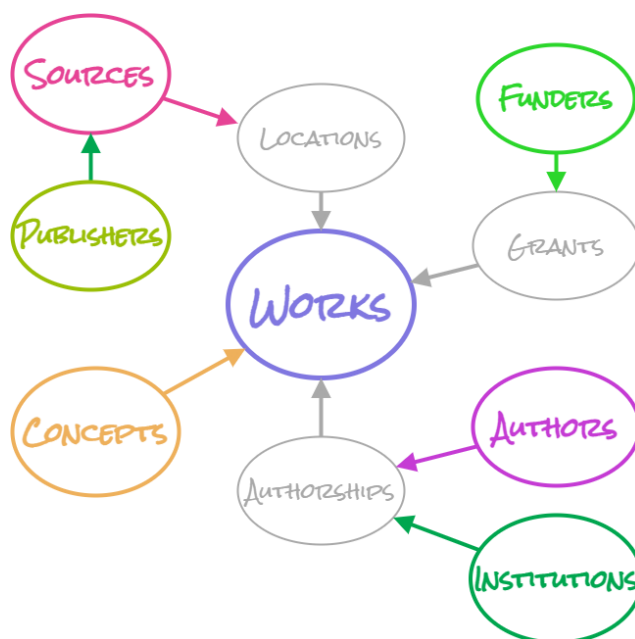


Fig. 1 This visual illustrates the intricate web of connections among various entities within the OpenAlex dataset. The entities include works (scholarly documents), authors (creators of works), sources (hosting venues for works), institutions (affiliated universities and organizations), concepts (topics assigned to works), publishers (organizations distributing works), and funders (organizations funding research).

age (see <https://openalex.org/stats>), encompassing 248 million works, 53 million open-access works, 28 million works from the Global South,^{*} 3 million datasets, 90 million authors (5 million with ORCIDs), 12 million authors from the Global South, 251,000 sources, 46,000 of which are open access, 10,000 publishers, 32,000 funders, 107,000 institutions, and 65,000 concepts.

Table 1 lists the various entities along with URL web links to both an example and the application programming interface (API). As shown in the example, each entity in OpenAlex is assigned a unique OpenAlex ID, allowing for stability and identification of specific resources within the database. The OpenAlex ID comprises two parts: the base, which is always <https://openalex.org/>, and the Key, a unique primary identifier (first letter) corresponding to the entity's type (e.g., Work, Author, Source, Institution, Concept, Publisher, Funder). Canonical External IDs, such as DOI for works, ORCID for authors, ISSN-L for sources, ROR ID for institutions and funders, and Wikidata ID for concepts and publishers, provide standardized

^{*}The Global South is a term used to identify regions within Latin America, Asia, Africa, and Oceania.

identifiers across various systems. For example, for works, this ID can also be tied to IDs within other databases: DOI (“doi”), Microsoft Academic Graph (“mag”), PubMed (“pmid”), etc.

Table 1 The various accessible entities within OpenAlex along with a URL to an example record and to the API. The examples chosen: “The Limitations of Deep Learning in Adversarial Settings”⁸ (works), Mark A. Tschoopp (authors), DEVCOM Army Research Laboratory (institutions), Army Research Office (funders), Science (sources), American Association for the Advancement of Science (publisher), and Computer Science (concept).

Entity	URL for Example Entity	URL Code to Access API
Works	https://openalex.org/W2180612164	https://api.openalex.org/works
Authors	https://openalex.org/A5073812740	https://api.openalex.org/authors
Institutions	https://openalex.org/I166416128	https://api.openalex.org/institutions
Funders	https://openalex.org/F4320338281	https://api.openalex.org/funders
Sources	https://openalex.org/S3880285	https://api.openalex.org/sources
Publishers	https://openalex.org/P4310315823	https://api.openalex.org/publishers
Concepts	https://openalex.org/C41008148	https://api.openalex.org/concepts

OpenAlex not only offers a wealth of data but now also presents a user-friendly web interface for exploring and analyzing the dataset, both through a user interface/dashboard (try it through <https://openalex.org/>) and an API for algorithmic analysis (<https://api.openalex.org/>). This resource proves a valuable database for science of science to institutions, researchers, governments, publishers, funders, and anyone interested in global research and scholarly communication. Its open nature positions OpenAlex as an alternative to paywalled services like Elsevier’s Scopus and Clarivate’s Web of Science, emphasizing inclusivity, affordability, and availability. The complete dataset is freely available under the CC0 license.

2.2 Navigating the JSON Format

Utilizing the OpenAlex API enables research and information retrieval from the dataset. It is essential to understand the nature of information provided by the API. To illustrate, focus on the entity “works”, encompassing articles, proceedings, books, datasets, and more. In Fig. 2, an excerpt from the JSON response of an API call is presented, specifically highlighting the “meta” key and its associated contents. Nested within this key is a dictionary comprising various keys and values. Noteworthy components include the total count of entities returned from the query (247,597,644 records), the response time in milliseconds (54 ms), the page number (pertinent for multiple pages), the count of records returned per page (25), and the count of groups returned (applicable when utilizing a group_by function).

The `"page"` and `"per_page"` keys play a pivotal role in navigating through multiple JSON files, particularly when aggregating extensive records surpassing the 200-per-page limit.

JSON format for OpenAlex: `"meta"`

```
"meta": {  
  "count": 247597644,  
  "db_response_time_ms": 54,  
  "page": 1,  
  "per_page": 25,  
  "groups_count": null  
},  
"results": [  
  {
```

Fig. 2 The JSON format for OpenAlex within the key `"meta"` that describes the total number of entities returned by the API (`"count"`), the response time (`"db_response_time_ms"`), and other information

Direct your attention to the subsequent key presented in the JSON response—namely, `"results"`. In Fig. 2, the `"results"` key, positioned at the bottom, houses a list of entries (25 in this instance, not shown). Each entry follows a dictionary format, featuring various keys that can be leveraged to extract pertinent information associated with each entry. Figure 3 provides an initial glimpse into the first entry within `"results"`. The default is to sort the entries by citation count, so this first entry is the Lowry et al.⁹ 1951 journal article entitled “PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT,” which was cited 304,009 times. This entry furnishes details such as the DOI ([10.1016/s0021-9258\(19\)52451-6](https://doi.org/10.1016/s0021-9258(19)52451-6)), OpenAlex ID ([W1775749144](https://openalex.org/works/W1775749144)), title, publication year, language, publication location, work type, and open-access status (`true`). It’s imperative to recognize that this overview only captures a fraction of the comprehensive information encapsulated within an entry. Additional data encompasses authorship details, countries involved, affiliated institutions, publishing charges, citation counts and rankings, associated concepts, work locations, optimal open-access repositories, alignment with sustainable development goals, references, abstract content, links to the list of citations, and citation counts per year (since 2012). Accessing this wealth of information programmatically enables researchers to swiftly address queries, aggregate relevant data, and conduct diverse analyses, thereby enhancing their understanding of the available information.

2.3 API Calls within OpenAlex

Navigating from a dataset of 247 million works to a refined understanding of a

JSON format for OpenAlex: "results"

```
"id": "https://openalex.org/W1775749144",
"doi": "https://doi.org/10.1016/s0021-9258(19)52451-6",
"title": "PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT",
"display_name": "PROTEIN MEASUREMENT WITH THE FOLIN PHENOL REAGENT",
"publication_year": 1951,
"publication_date": "1951-11-01",
"ids": {
  "openalex": "https://openalex.org/W1775749144",
  "doi": "https://doi.org/10.1016/s0021-9258(19)52451-6",
  "mag": "1775749144",
  "pmid": "https://pubmed.ncbi.nlm.nih.gov/14907713"
},
"language": "en",
"primary_location": {
  "is_oa": true,
  "landing_page_url": "https://doi.org/10.1016/s0021-9258(19)52451-6",
  "pdf_url": null,
  "source": {
    "id": "https://openalex.org/S140251998",
    "display_name": "Journal of Biological Chemistry",
    "issn_l": "0021-9258",
    "issn": [
      "1083-351X",
      "0021-9258",
      "1067-8816"
    ],
    "is_oa": true,
    "is_in_doaj": true,
    "host_organization": "https://openalex.org/P4310320990",
    "host_organization_name": "Elsevier BV",
    "host_organization_lineage": [
      "https://openalex.org/P4310320990"
    ],
    "host_organization_lineage_names": [
      "Elsevier BV"
    ],
    "type": "journal"
  },
  "license": "cc-by",
  "version": "publishedVersion",
  "is_accepted": true,
  "is_published": true
},
"type": "article",
"type_crossref": "journal-article",
"indexed_in": [
  "crossref"
],
"open_access": {
  "is_oa": true,
  "oa_status": "hybrid",
  "oa_url": "https://doi.org/10.1016/s0021-9258(19)52451-6",
  "any_repository_has_fulltext": true
},
```

Fig. 3 The JSON format for OpenAlex within the key "results" that contains information about OpenAlex ID, DOI, title, publication year, location of the published work, whether it is open access, etc.

specific area of interest requires leveraging the flexibility offered by additional parameters in the API URL. The OpenAlex API allows users to tailor their queries by filtering, sorting, grouping, and conducting keyword searches within the dataset—explained within the technical documentation of OpenAlex.¹⁰ To illustrate this, Ta-

ble 2 provides a step-by-step example of constructing a URL for API access. In the initial step, initiate with the API address and specify the entity type that is to be accessed (works), bearing in mind that filters may vary depending on the entity type. Subsequent steps showcase how to progressively filter the dataset. For instance, the second step demonstrates filtering for a specific type of work, such as articles (inclusive of both journal articles and conference proceedings). The third step narrows down the dataset to records with publication years from 2018 to the present. Further refinements follow, such as filtering for works associated with institutions in the US, limiting the dataset to publications featuring global collaboration (> 1 institution), specifying a minimum citation threshold (> 100), and finally, focusing on publications associated with the level 1 concept of AI (using the OpenAlex ID C154945302). This curated query results in 4,746 publications, offering a focused dataset for more in-depth exploration across various dimensions within the JSON information.

Table 2 An example of accessing the OpenAlex API in successive steps: (1) initial URL, (2) only journal articles and conference proceedings, (3) for publication years from 2018 to present, (4) only when US is one of the author countries, (5) when there are multiple countries involved, (6) when citations are above 100, and (7) when one of the concepts is AI

Step #	URL Code to Access API	# of Records
1	https://api.openalex.org/works	247,597,668
2	https://api.openalex.org/works?filter=type:article	203,698,955
3	https://api.openalex.org/works?filter=type:article,publiation_year:2018-2024	43,751,516
4	https://api.openalex.org/works?filter=type:article,publiation_year:2018-2024,institutions.country_code:US	5,237,793
5	1">https://api.openalex.org/works?filter=type:article,publiation_year:2018-2024,institutions.country_code:US,countries_distinct_count:>1	1,935,598
6	1,cited_by_count:>100">https://api.openalex.org/works?filter=type:article,publiation_year:2018-2024,institutions.country_code:US,countries_distinct_count:>1,cited_by_count:>100	39,292
7	1,cited_by_count:>100,concepts.id:C154945302">https://api.openalex.org/works?filter=type:article,publiation_year:2018-2024,institutions.country_code:US,countries_distinct_count:>1,cited_by_count:>100,concepts.id:C154945302	4,746

3. Research Study: Global Collaboration in Science

3.1 Introduction

Is collaboration a catalyst for scientific progress? The diminishing prevalence of solo-authored papers over time suggests a shifting trend towards collaborative efforts in scientific research.¹¹ This surge in collaboration is driven by the recognition

that team-based endeavors often yield more impactful outcomes than solo-authored works. Wuchty, Jones, and Uzzi's extensive analysis of nearly 20 million papers and over 2 million patents spanning five decades demonstrates the ascendancy of teams in knowledge production across diverse research domains.¹¹ Moreover, their findings reveal that articles produced collaboratively garner more frequent citations, signifying greater impact within the research community. The mounting evidence underscores the positive correlation between collaboration, knowledge production, and the heightened impact of scientific contributions.

Yet, as collaboration becomes more ingrained in the fabric of scientific inquiry, the question arises: what about collaboration that transcends the geographical boundaries of countries? The advent of the Internet has revolutionized the exchange of information globally, epitomized by the phrase "death of distance" in scientific discourse. The COVID-19 pandemic has only accelerated this trend, pushing scientific interactions into the virtual realm. The so-called death of distance prompts inquiries into the frequency of papers featuring collaborations between multiple countries. Existing studies indicate a steady increase in the percentage of papers involving international collaborations.¹² However, delving deeper, questions emerge: with whom do the top countries collaborate? How does this trend manifest for institutions and funding agencies? In which research areas do countries forge collaborations, and how has this landscape evolved over time? The growing impact of science's globalization, coupled with concerns about the openness of scientific endeavors, fuels curiosity within the scientific community, as well as among policymakers and funding agencies, making this area of research a focal point for exploration and understanding.

Building on the foundations laid by team science and recognizing the collaborative shift in the scientific landscape, this study aligns itself within the domain of the Science of Science. It explores the evolutionary trajectory of scientific articles, concurrently delving into the dynamics of global scientific collaboration. The impetus for this inquiry stems from persistent questions surrounding collaborative endeavors among countries, the nuanced structures of emerging open databases such as OpenAlex, global institutional collaborations, and the intersecting realms of interdisciplinary interactions on a global scale.

In a world where intellectual property effortlessly crosses national borders and

knowledge operates on a global stage, understanding the nuances of collaboration is vital. This study illuminates global collaboration patterns, specifically examining interactions between top (elite) countries, and further delving into the landscape of collaboration within national laboratories, universities, and potentially extending to individual researcher collaborations on a global scale. The utilization of OpenAlex serves as a lens to extract insights into the collaborative landscape and assess data quality within this extensive repository.

The implications of this study extend across policy, practice, and future research, providing valuable insights into intellectual discoveries that transcend geographic boundaries. Unraveling the intricate pathways through which scientific knowledge circulates globally, these findings aim to guide decisions pertaining to international partnerships in advancing science. Designed for those captivated by global collaboration across diverse domains, this study serves as a resource and starting point for understanding its impact on scientific progress and potentially safeguarding discoveries within the realm of intellectual property. Rooted in intellectual curiosity, this pursuit not only satiates personal interest but also establishes a groundwork for subsequent investigations, benefitting various stakeholders and communities.

3.2 Methods

The analysis of global collaboration in this study relies on the OpenAlex database, a resource that captures the interconnected landscape of global research entities through including country information with institution affiliations for authors. Utilizing the OpenAlex API in a similar manner to what was shown in Table 2, users can probe this database through a hyperlink, employing various filters to refine searches based on parameters such as the year, involved institutions, authors and their affiliations, countries, number of distinct countries, and research concepts, among others.

The API returns information in JSON format as shown in Figs. 2 and 3, providing a structured output that can be easily accessed and processed through Python scripts. These scripts facilitate the transformation of the retrieved data into variables, enabling further analysis within the Python programming environment. In addition to the flexibility of filtering the database, the API allows results to be sorted by citation counts or grouped by different categories, such as work type (e.g., article, book chapter, book, paratext) or country. This versatility in sorting and grouping en-

hances the depth of analysis and provides nuanced insights into global collaboration patterns.

Given the focus on global collaboration in many of the analyses, the API calls often start with a primary country, such as the US. The results are then grouped by secondary collaborating countries, allowing for the measurement of collaboration intensity between the primary country and its global counterparts. This approach facilitates the examination of collaborative networks and highlights the extent of international research cooperation.

3.3 Results and Discussion

3.3.1 Temporal Evolution of Global Science

Before addressing questions about global collaboration, it is crucial to understand the foundational aspects of the OpenAlex database, particularly the evolution of the number of articles over time. Figure 4 illustrates the annual production of articles since 1900, encompassing both peer-reviewed journal articles and conference proceedings. The plot reveals a nearly monotonic increase in the number of articles per year, with notable interruptions during World War I and World War II, as indicated in the plot. Intriguingly, the peak in the maximum number of articles per year occurs around 2020, followed by a decline, notably influenced by the global COVID-19 pandemic. The red line serves as a visual guide, signaling that the number of articles doubles approximately every 14 years. Notably, the logarithmic scale on the y axis captures the nonlinear growth in publications over time. The OpenAlex database currently hosts an extensive collection of over 200 million articles. As an illustration of its dynamic nature, the database witnesses a continuous influx of new content; to put this in perspective, if 7.2 million articles were produced annually, this would result in the addition of approximately 20,000 articles each week. Notably, the curation of the database in terms of these articles is performed through scripting rather than manual curation. The QR code provides direct access to the most recent data used for this plot.

Figure 5 delves into the production of book chapters over time, mirroring the format of Fig. 4 with key distinctions. The OpenAlex dataset contains over 20 million book chapters, approximately 10% of the total articles. Unlike the decreasing trend in articles post-2020, the number of book chapters steadily rises, with the maximum observed in 2023. Noteworthy is the historical fluctuation in book chapter produc-

Evolution of Scientific Articles Since 1900

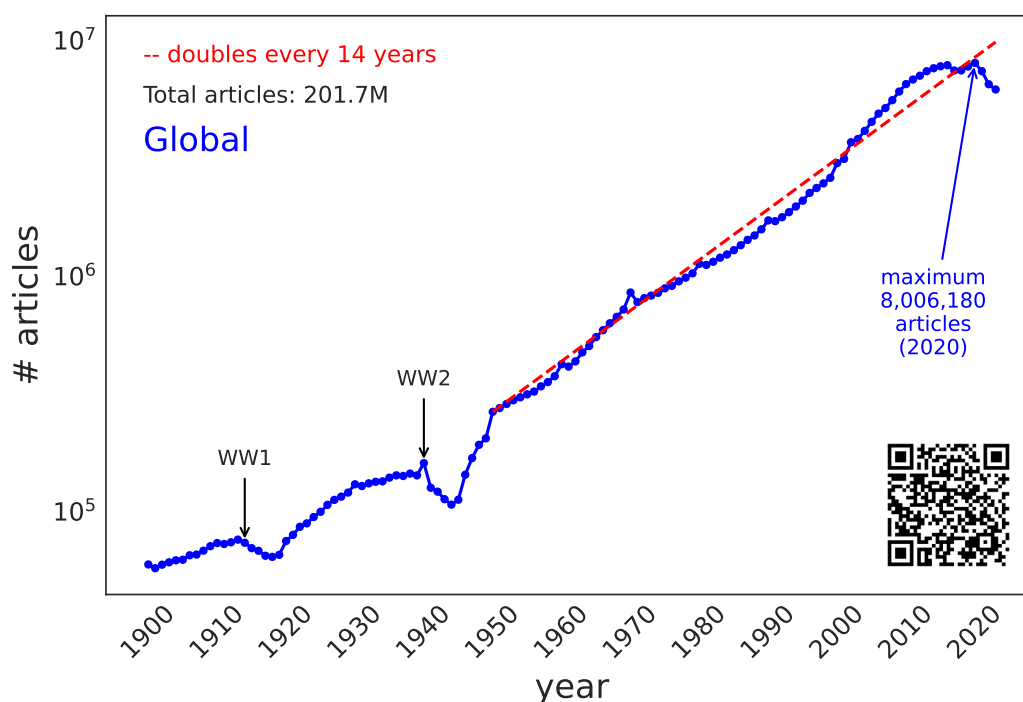


Fig. 4 The evolution of the number of scientific articles produced globally per year since 1900. The “article” type refers to peer-reviewed journal articles and conference proceedings. The beginning of the first two world wars are identified as is the maximum in the produced articles per year (peaks in 2020). The red line shows a doubling of articles every 14 years as a guide to the eye. In this figure and subsequent figures, the QR code in the lower right-hand corner is a link to the OpenAlex data source.

tion, especially during World Wars, giving way to a post-World War II era marked by a consistent increase. This evolution sees the number of book chapters doubling approximately every ten years.

Turning attention to another medium for disseminating scientific knowledge, Fig. 6 explores the annual production of books over time. The format aligns with Figs. 4 and 5, revealing interesting trends. The maximum number of books occurs in 2011, with a substantial decline in subsequent years, particularly after 2020. Despite a historical trend of monotonic increase since around 1950, this pattern shifts in the last decade. While multiple factors may contribute to this decline, the popularization of the h-index as a metric for assessing scientific productivity and impact could play a significant role in the incentive system for scientists, i.e., one book takes significantly more effort to produce than one journal article, but each can increase the h-index by one. The logarithmic y axis accentuates the magnitude of this decrease

Evolution of book chapters since 1900

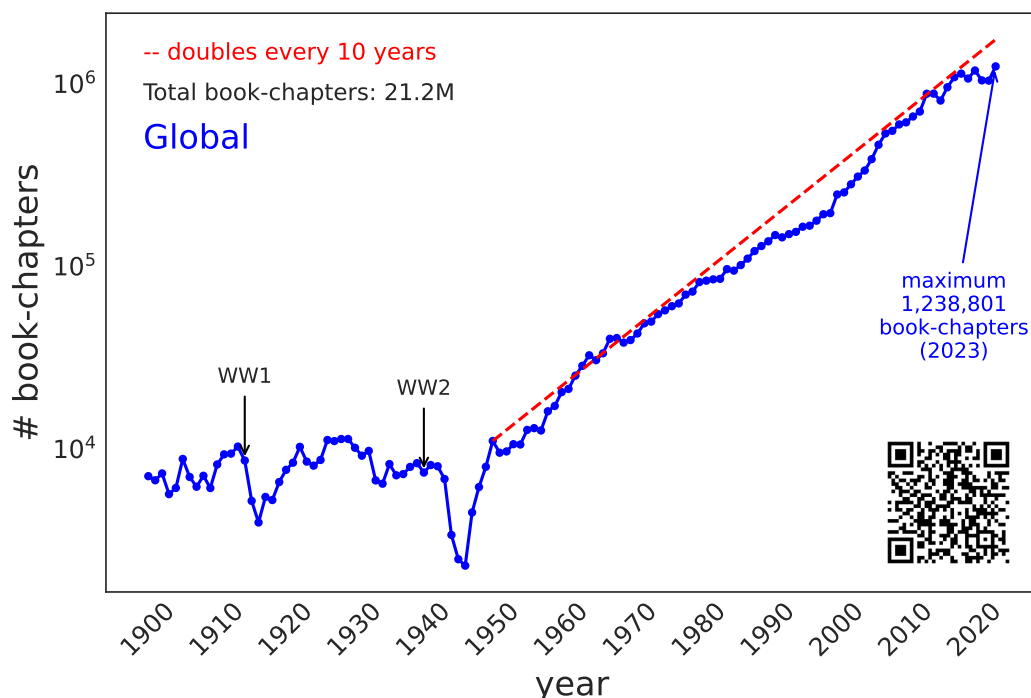


Fig. 5 The evolution of the number of book chapters produced globally per year since 1900. The beginning of the first two world wars are identified as is the maximum in the produced book chapters per year (peak is in 2023). The red line shows a doubling of book chapters every 10 years as a guide to the eye. The total number of book chapters is approximately 10% of the total scientific articles.

over the past decade.

The global distribution of articles can be elucidated by examining the countries listed in the authors' affiliations. Table 3 provides a comprehensive breakdown of countries listed within the author affiliations for articles in the OpenAlex database. The initial two rows distinguish whether the country is mentioned in the article. Notably, approximately 58% of articles lack a specified country. Further investigation into a subset of these articles reveals instances where institutional information is also absent. Various factors contribute to this, ranging from articles featuring only author names without affiliations to algorithmic limitations in extracting such details. Despite the dataset's imperfections, more than 84 million articles include author affiliations with country information, leading to over 100 million country-paper pairs.* It is assumed that trends within this subset are representative of the full sample population.

*101,771,796 country-paper pairs representing 225 countries

Evolution of science books since 1900

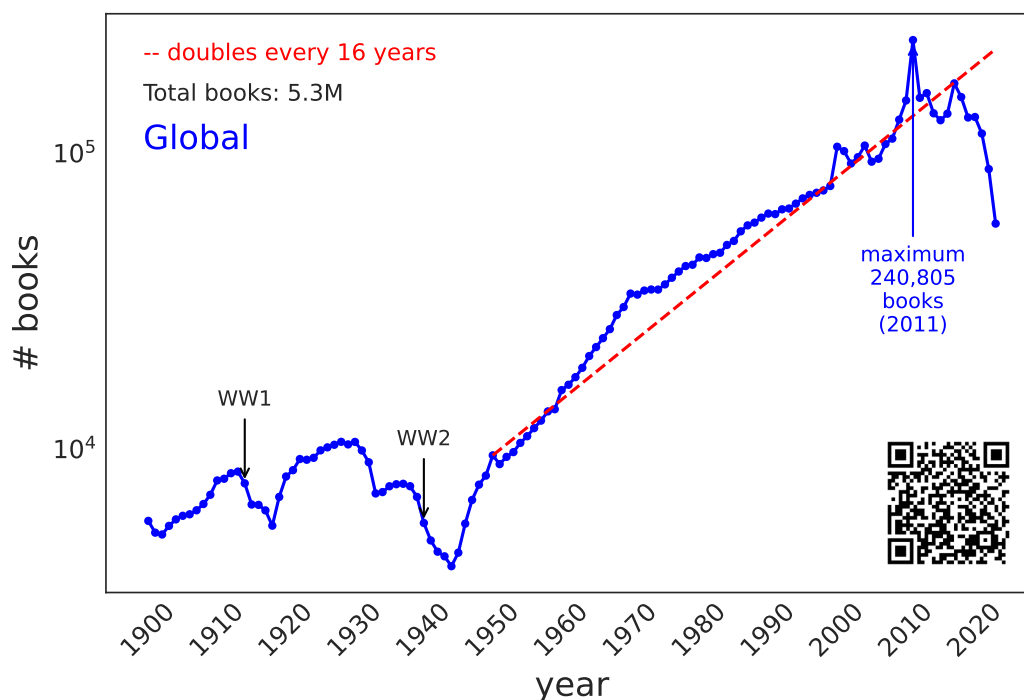


Fig. 6 The evolution of the number of science books produced globally per year since 1900. The beginning of the first two world wars are identified as is the maximum in the produced books per year (peaks in 2011). The red line shows a doubling of books every 16 years as a guide to the eye. The total number of books is approximately 2.6% of the total scientific articles.

Figure 7 presents a bubble plot illustrating the contribution of various countries to articles in the OpenAlex database. Each bubble's size corresponds to the frequency with which a particular country is mentioned in the affiliations of articles. This dynamic visualization provides a snapshot of the cumulative contributions of countries across all articles, emphasizing the prevalence of each country in scholarly publications. The bubble plot aligns with the insights presented in Table 3 but adds a visual dimension to showcase the relative impact of different countries on a global scale.

It's crucial not to interpret the percentages in Table 3 and the bubble sizes in Fig. 7 as the proportion of authors from the listed countries. For example, a single publication may enumerate multiple countries in the affiliation section, counting once for each distinct country. Additionally, a publication with 10 authors from one country contributes only once to that country. As a result, the percentages exceed 100%. The “% collaboration” column represents the proportion of articles listing the given country minus the percentage of those articles exclusively affiliated with that coun-

Table 3 The breakdown of countries within author affiliations for all articles in OpenAlex. The first two rows highlight the number of unknown and known country affiliations within OpenAlex and their corresponding percentages. For the country information, “% of total” refers to the percent of articles that include at least one author with an affiliation from that country. The column “% collaboration” is the percentage of articles with that country listed that have other countries as well.

Country Code	Country	# Publications	% of total	% collaboration
unknown	unknown	119,236,553	58.7	N/A
known	known	84,051,614	41.3	16.2
US	United States of America	23,167,541	27.6	23.4
CN	China	12,240,083	14.6	16.9
GB	United Kingdom	6,188,074	7.4	39.6
JP	Japan	4,805,921	5.7	20.3
DE	Germany	4,796,137	5.7	38.9
FR	France	4,104,957	4.9	36.2
CA	Canada	3,066,151	3.6	39.7
IN	India	2,891,882	3.4	20.2
IT	Italy	2,691,939	3.2	39.9
BR	Brazil	2,573,234	3.1	20.3
AU	Australia	2,318,166	2.8	43.4
ES	Spain	2,280,556	2.7	38.0
RU	Russia	1,888,689	2.2	25.1
ID	Indonesia	1,731,123	2.1	6.8
KR	Korea	1,678,691	2.0	27.0
NL	Netherlands	1,590,320	1.9	49.2
PL	Poland	1,365,976	1.6	23.2
CH	Switzerland	1,137,710	1.4	57.0
SE	Sweden	1,083,114	1.3	48.4
TR	Turkey	1,033,422	1.2	19.5
IR	Iran	1,011,450	1.2	22.1
BE	Belgium	983,932	1.2	48.7
TW	Taiwan	938,596	1.1	27.5
MX	Mexico	736,539	0.9	30.8
DK	Denmark	689,193	0.8	48.9
IL	Israel	637,533	0.8	39.9
NO	Norway	606,732	0.7	44.3
AT	Austria	606,066	0.7	52.4
ZA	South Africa	525,813	0.6	40.3
PT	Portugal	518,406	0.6	45.6
FI	Finland	513,298	0.6	46.5
CZ	Czechia	509,722	0.6	49.6
EG	Egypt	503,009	0.6	39.5
MY	Malaysia	485,870	0.6	39.2
GR	Greece	453,387	0.5	41.8
SG	Singapore	441,438	0.5	53.4
AR	Argentina	423,881	0.5	32.9
SA	Saudi Arabia	407,084	0.5	67.5
NZ	New Zealand	374,528	0.4	46.3
CO	Colombia	366,627	0.4	27.6

Global Article Contributions to Open Science

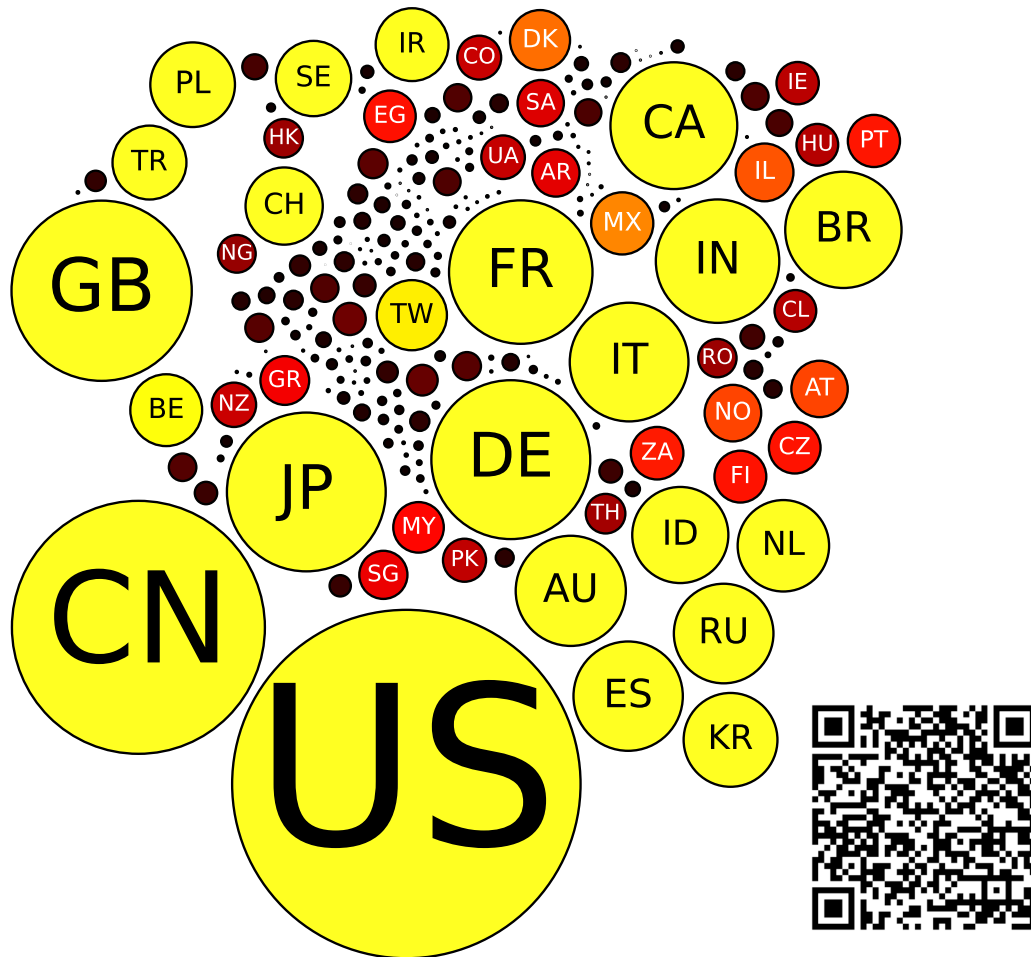


Fig. 7 A bubble chart illustrating the number of scientific articles contributed by each country. The size of each circle is proportional to the total number of articles listing affiliations from that specific country. For the full names of countries, refer to Table 3.

try. This reveals how often other countries are listed for those publications. While this offers a snapshot of global science and publications, it does not capture the temporal evolution of articles, changes in research concepts over time, or the evolution of collaboration between countries. Subsequent analyses will delve into these aspects.

3.3.2 Temporal Evolution of Scientific Productivity for Top Countries

Exploring the evolution of scientific articles over time provides intriguing insights into the scientific output of different countries. In Fig. 8, the continuous growth in the number of articles per year with affiliations from the US is evident, with occasional dips during the first and second world wars. Remarkably, the peak in

articles authored by the US occurred in 2020, coinciding with the global impact of the COVID-19 pandemic.

Evolution of Global Articles with US Affiliations

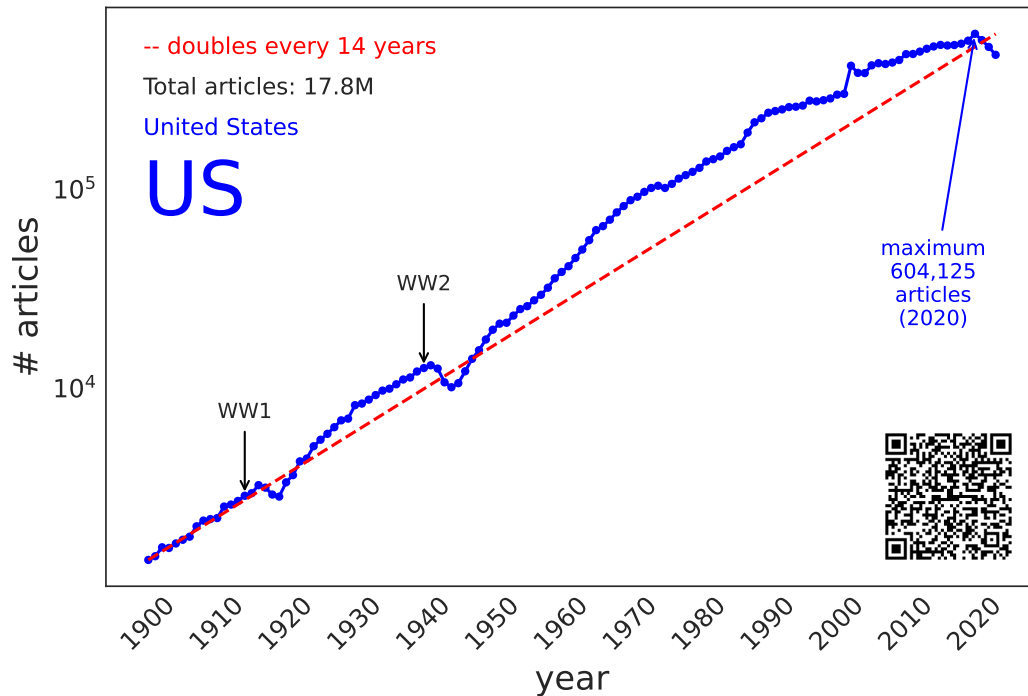


Fig. 8 The evolution of the number of articles produced globally per year since 1900 that list an author affiliation located within the US

Figure 9 shifts the focus to China, revealing a distinct narrative. The rapid pace of article production, highlighted by the steep trajectory on the y axis, sets China apart from other top-contributing countries. Examining the period stretching from the 1900s to the 1970s, China's output remained below 1000 articles per year, reaching the milestone of 1000 articles annually with affiliations from China around 1980. Notably, China's scientific productivity during the world wars appears less affected, possibly due to lower article numbers or the country's relative neutrality, which could have shielded it from the direct impact experienced by nations deeply involved in the conflicts. Also of note is the rapid ascent in articles published around the 2000s, which experienced a sizeable dip in the 2010s, and is now on a path where the rate of scientific output doubles every 5 years.

Figures 10–12 extend this analysis to the UK, Japan, and Germany, respectively. Following a trajectory parallel to the US trend, these countries exhibit overall growth, punctuated by dips during wartime. Japan, for instance, experiences a notable pro-

Evolution of Global Articles with China Affiliations

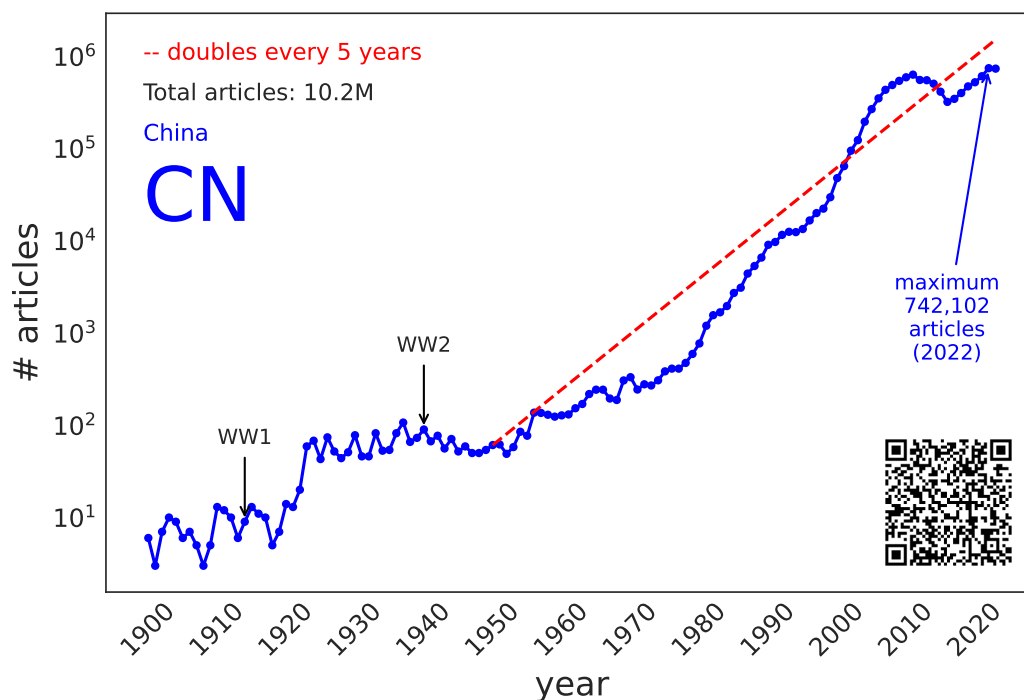


Fig. 9 The evolution of the number of articles produced globally per year since 1900 that list an author affiliation located within China

ductivity dip during World War II. Subsequently, it resumes the upward trend typical of scientific article output until the mid to late 20th century. However, Japan's output has shown signs of stabilization over the last few decades. This persistent but plateauing productivity prompts intriguing questions about the dynamics within a world where scientific output is on a continuous ascent. Unlike other nations that demonstrate sustained growth, Japan's consistent but unchanging productivity over recent decades may suggest a unique position or challenges within the evolving landscape of global scientific contributions.

Germany, on the other hand, faces the most substantial loss in scientific output during both world wars. However, Germany's recovery following World War II is remarkable, with a resurgence along the original steady growth trend similar to other countries. Germany's ability to recover and rejoin the scientific growth trend underscores its resilience and adaptability, contributing to the complex interplay of factors shaping the scientific output of different nations.

This exploration underscores the global dynamics of scientific productivity, captur-

Evolution of Global Articles with UK Affiliations

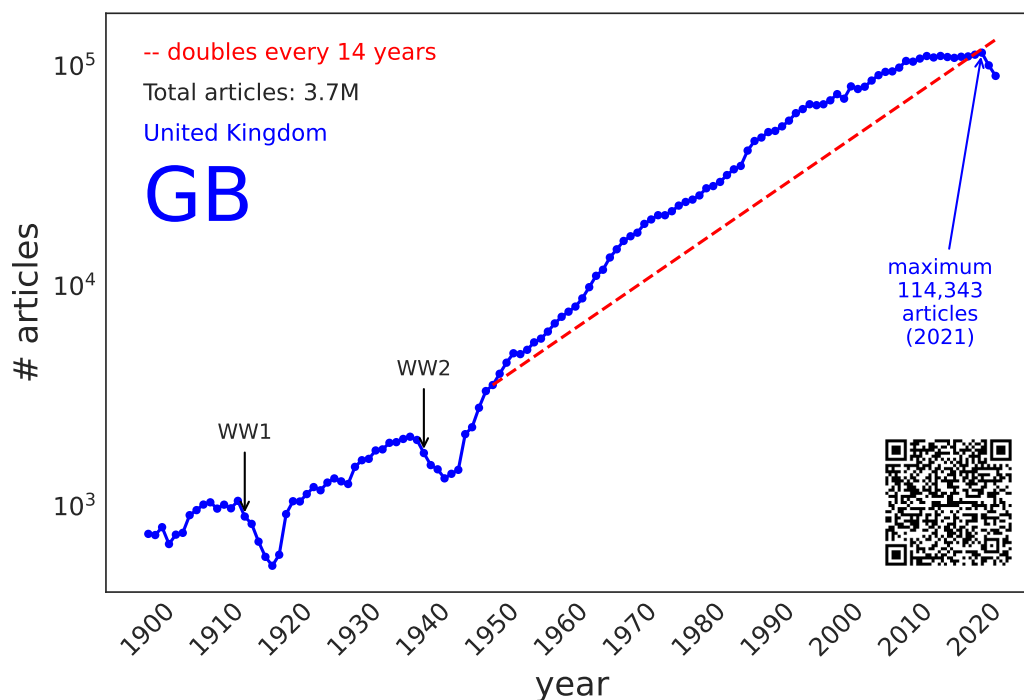


Fig. 10 The evolution of the number of articles produced globally per year since 1900 that list an author affiliation located within the United Kingdom of Great Britain and Northern Ireland

ing unique trajectories for each country and revealing the impact of historical events on their scientific output. Further examination in subsequent figures delves into specific aspects, providing a comprehensive understanding of the evolving landscape of scientific contributions.

As the evolution of scientific collaboration is examined, the question arises: what proportion of articles, over time, involves contributions from multiple countries? Figure 13 elucidates this trend, showcasing the percentage of collaborative articles since 1900. Leveraging the capabilities of OpenAlex, articles were filtered with distinct affiliations from more than one country and then grouped them by publication year. Intriguingly, the percentage of collaborative articles remained relatively constant in the first half of the 20th century but exhibited a steady increase thereafter, surpassing 20% in recent years. This trend prompts contemplation about the factors driving such collaboration. Is it fueled by incentives for scientists, such as access to added capabilities, instrumentation, expertise, and impact? Or does it reflect the increasing trend of joint affiliations by scientists, whereby they hold appointments

Evolution of Global Articles with Japan Affiliations

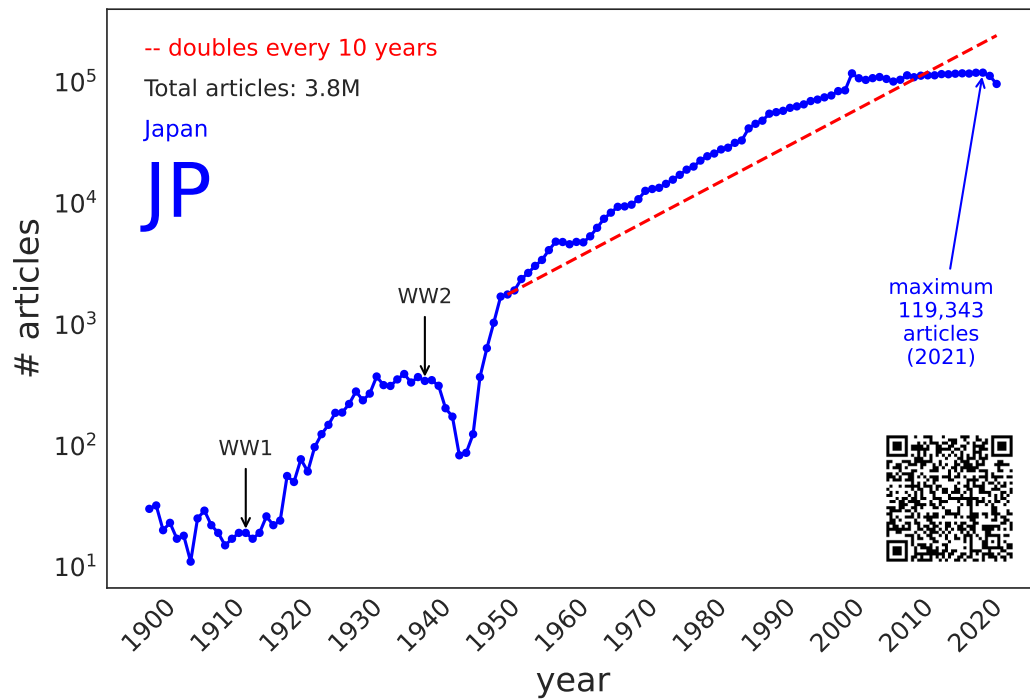


Fig. 11 The evolution of the number of articles produced globally per year since 1900 that list an author affiliation located within Japan

with institutions in multiple countries, enabling them to draw additional resources towards research? While not explored in this study, some of these questions can also be answered via large databases like OpenAlex.

Evolution of Global Articles with Germany Affiliations

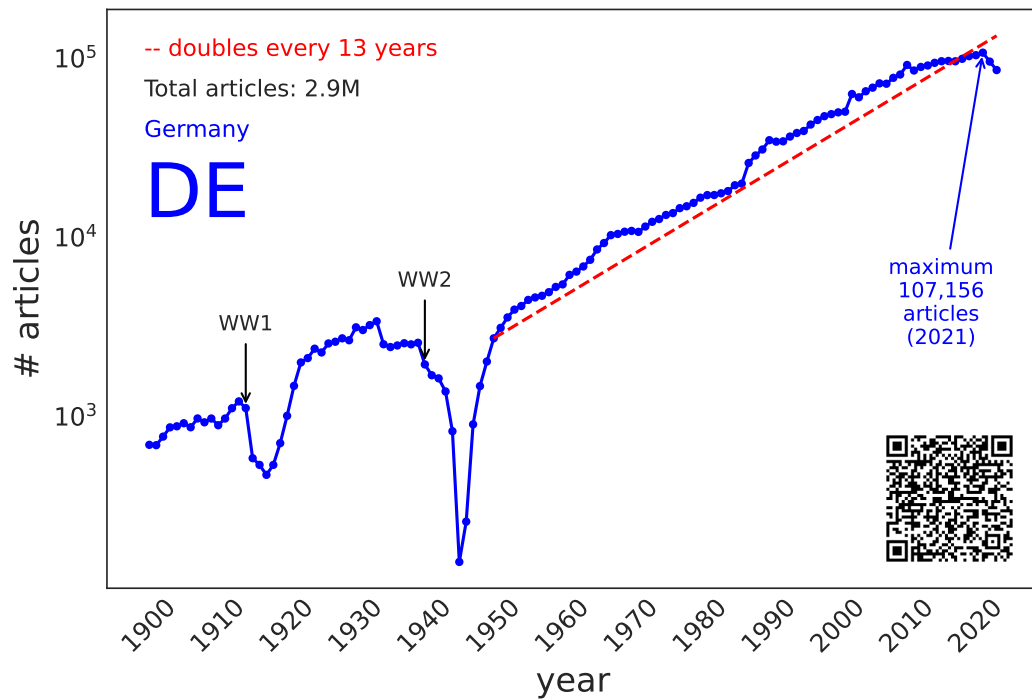


Fig. 12 The evolution of the number of articles produced globally per year since 1900 that list an author affiliation located within Germany

Temporal Evolution of Global Collaboration

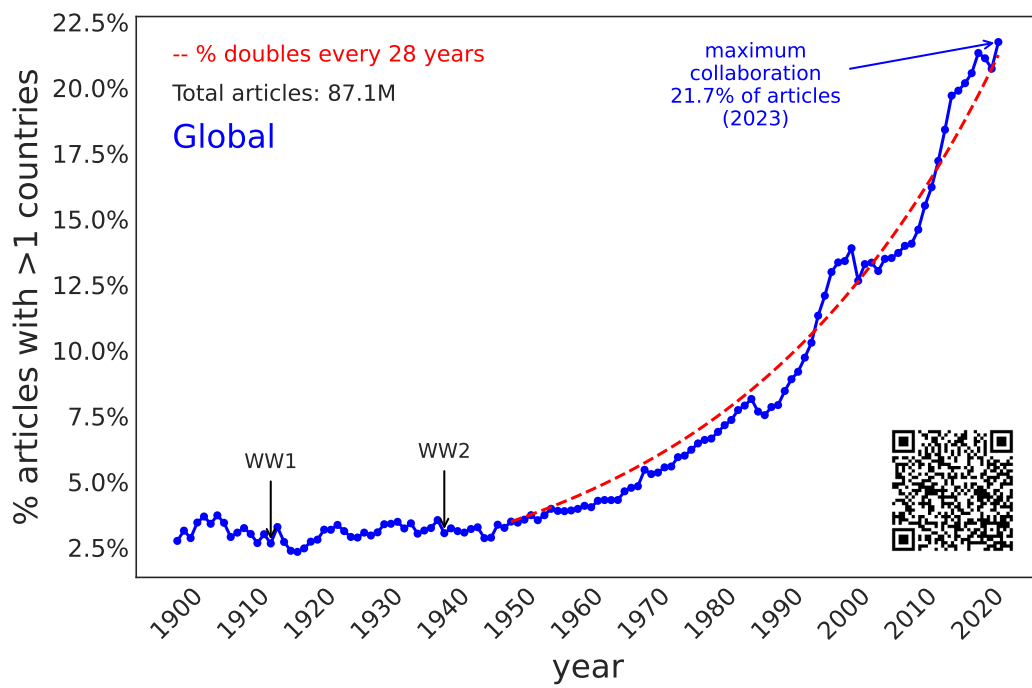


Fig. 13 The trend in the number of articles with authors from multiple countries since 1900

3.3.3 Frequency of Countries, Institutions, Authors, and Concepts

Given the established base rate of collaboration exceeding 20%, the exploration now delves into the distribution of distinct countries involved in these collaborative publications. Figure 14 delineates the percentage of articles against the count of distinct countries within author affiliations. In the plot's lower left corner, the substantial presence of articles without any specified country—totaling 116 million out of the 203 million articles in the dataset—becomes evident, constituting approximately 57% of the dataset. Beyond this, 36% of articles feature one distinct country, with a subsequent monotonic decrease in percentages as the count of distinct countries increases. To accentuate these percentages, the plot incorporates annotations listing them beside the respective data points, while the count of distinct countries is highlighted below each point in red. The logarithmic scale on the y axis effectively reveals the continual decrease, even for small percentages that may span several orders of magnitude, providing a nuanced perspective on articles with varying counts of distinct countries.

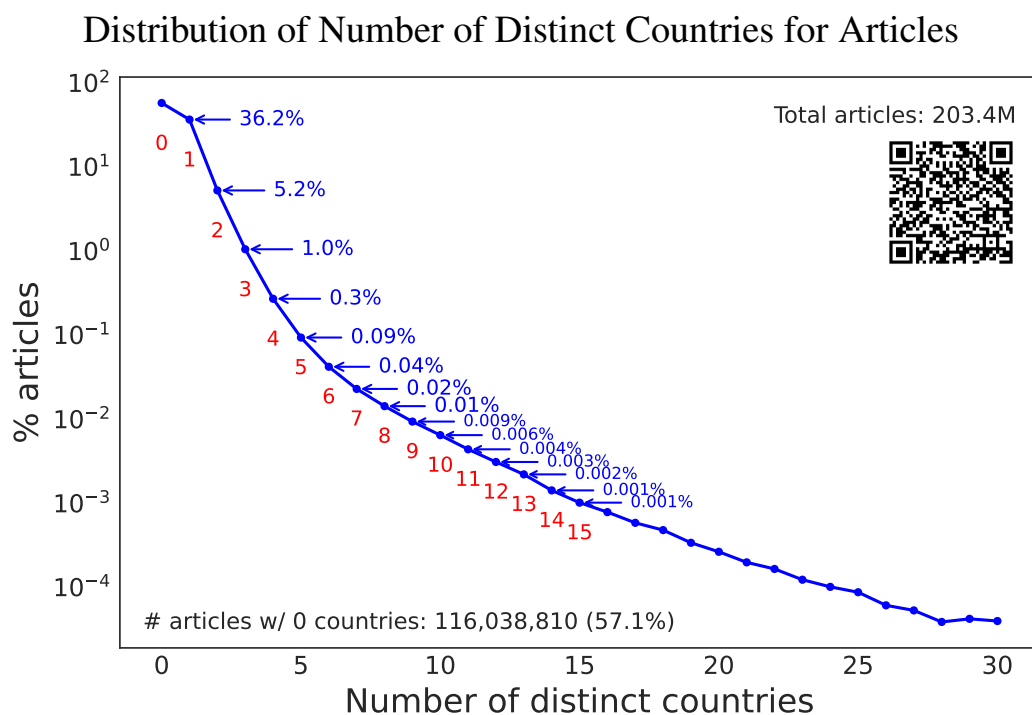


Fig. 14 The distribution of the percentage of articles involving various numbers of distinct countries within the author affiliations. The percentages represent the proportion of articles relative to the total number of articles in the database. The y axis employs a logarithmic scale, emphasizing the continual monotonic decrease in percentage with an increasing number of distinct countries collaborating on a publication.

Given the prevalence of articles lacking country information, the inquiry shifts to the presence of institutions listed for each article. Figure 15 illustrates the percentage of articles across various counts of distinct institutions. Notably, a substantial fraction of the database comprises articles with zero institutions listed—exceeding 119 million articles or 58.7%. This considerable proportion closely aligns with the absence of information about distinct countries; specifically, the 116,171,340 entries with zero countries also exhibit zero institutions. Furthermore, a breakdown reveals that 3,878 entries possess institution information but lack country details, while 3,069,481 entries include country information but lack institution specifics. Within these entries, the number of authors emerges as a pivotal factor; single-author articles constitute over 50%, with two- and three-author papers comprising a significant portion of the remaining entries. This intricate interplay underscores the multifaceted nature of authorship, institutions, and country affiliations within the dataset.

Distribution of Number of Distinct Institutions for Articles

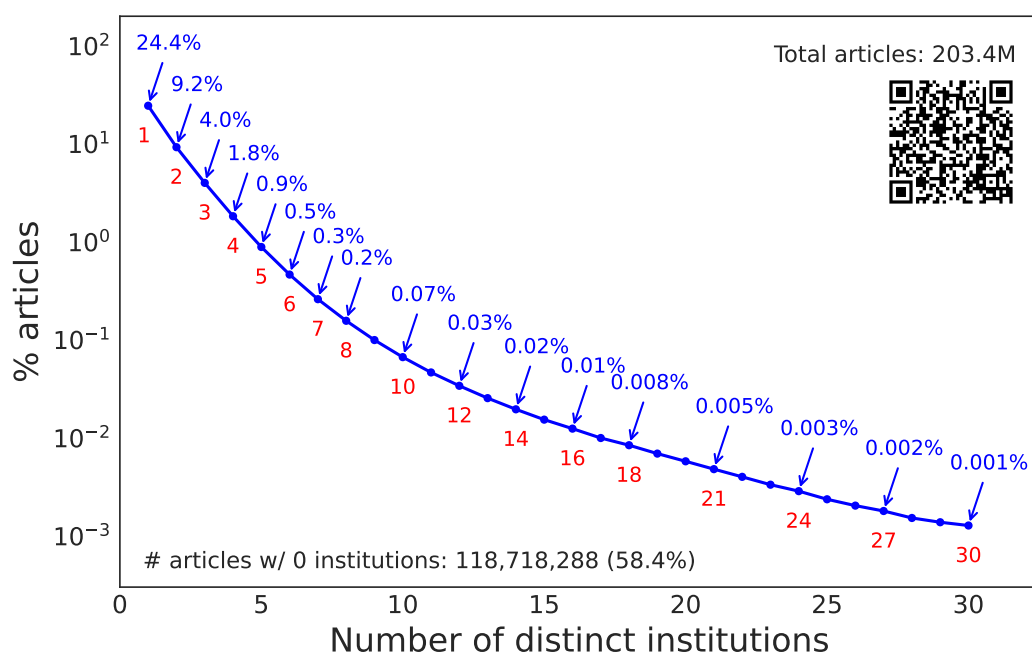


Fig. 15 The distribution of the percentages of articles involving different numbers of distinct institutions within the author affiliations. The percentages represent the proportion of articles relative to the total number of articles in the database. The y axis employs a logarithmic scale, emphasizing the continual monotonic decrease in percentage with an increasing number of distinct institutions collaborating on a publication.

Exploring the composition of articles within the database, an examination of au-

thorship reveals the distribution of articles across varying author counts. Figure 16 depicts the percentage of articles relative to the number of distinct authors per article. Notably, articles with zero authors constitute just under 10 million entries, roughly 5% of the total 200 million articles. Among the remaining articles, the majority feature a single author, followed by two-author publications as the next highest percentage. Subsequently, three-author articles, four-author articles, and so forth demonstrate a diminishing percentage as the number of authors increases. A noteworthy trend emerges, where the percentage of articles consistently decreases monotonically with each incremental increase in the number of authors, extending up to 30 authors, as illustrated in Fig. 16. This decreasing relationship is accentuated by employing a logarithmic scale for the y axis, emphasizing the nuanced dynamics of authorship distribution within the dataset.

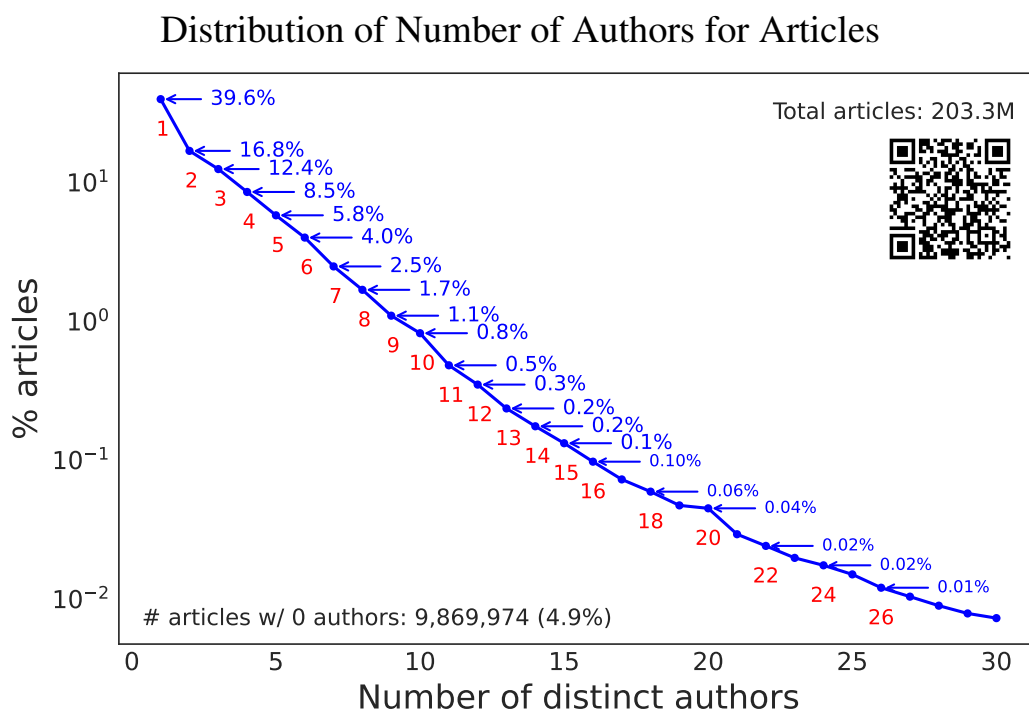


Fig. 16 The distribution of percentages of articles involving different numbers of authors. The percentages represent the proportion of articles relative to the total number of articles in the database. The y axis employs a logarithmic scale, emphasizing the continual monotonic decrease in percentage with an increasing number of distinct authors collaborating on a publication.

Turning the attention to the diversity of research concepts across articles, Fig. 17 presents the percentage of articles corresponding to the number of distinct concepts. It's crucial to note that these concepts result from the ML classification of article

titles, abstracts, and journal titles, thereby organizing articles into a taxonomy of research topics. A striking observation emerges: a mere fraction, approximately 27,000 articles, out of the vast 200 million articles in the database, lack any assigned concepts. For the majority of articles, one research concept is predominant, followed by articles with three concepts constituting the second-highest percentage. The distribution continues, encompassing two research concepts and beyond. Intriguingly, the percentage of articles with two to 20 concepts maintains a consistent pattern before embarking on a monotonic decrease for counts surpassing 20 concepts. This trend, outlined in Fig. 17, unveils the distribution of research concepts across the dataset, showcasing a nuanced interplay between the breadth of conceptual coverage and the declining frequency of occurrence as concepts diversify.

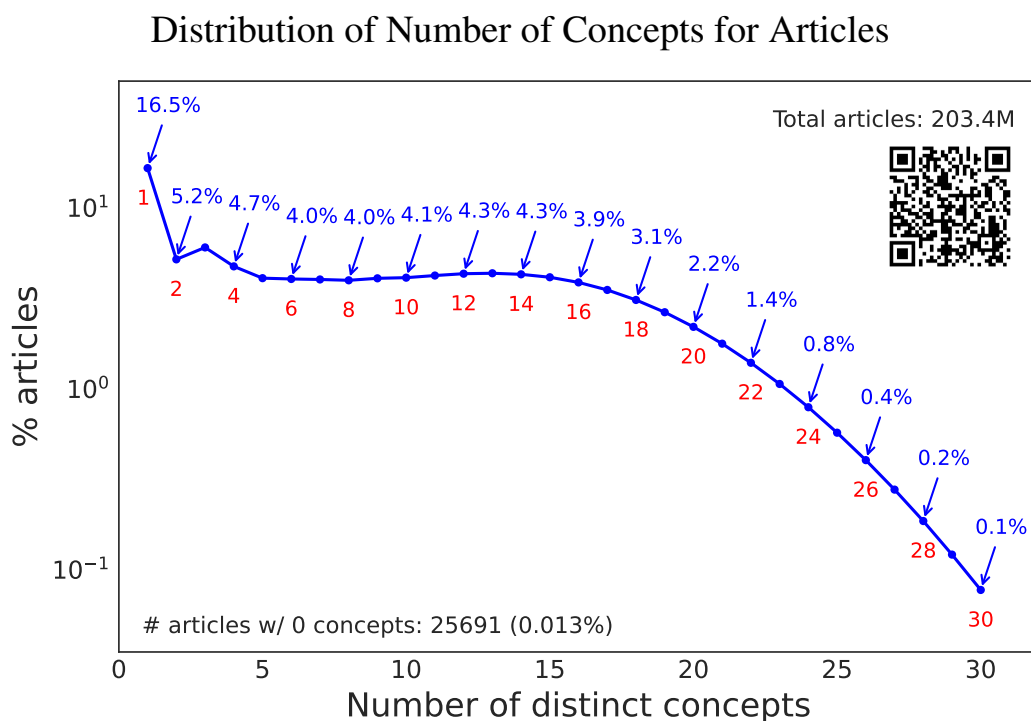


Fig. 17 The distribution of percentage of articles involving different numbers of research concepts tagged within OpenAlex. The percentages represent the proportion of articles relative to the total number of articles in the database. The y axis employs a logarithmic scale to highlight the wide range of percentages for different numbers of concepts.

3.3.4 Global Collaboration between Leading Countries

Having delved into the distinct countries and institutions represented in the database, the exploration now turns to the frequency of collaboration between countries. Fig-

ure 18 presents a heat map illustrating global collaboration among the top 12 countries in terms of scientific article production since 2018.

Global Collaboration Heatmap for All Articles (2019–present)

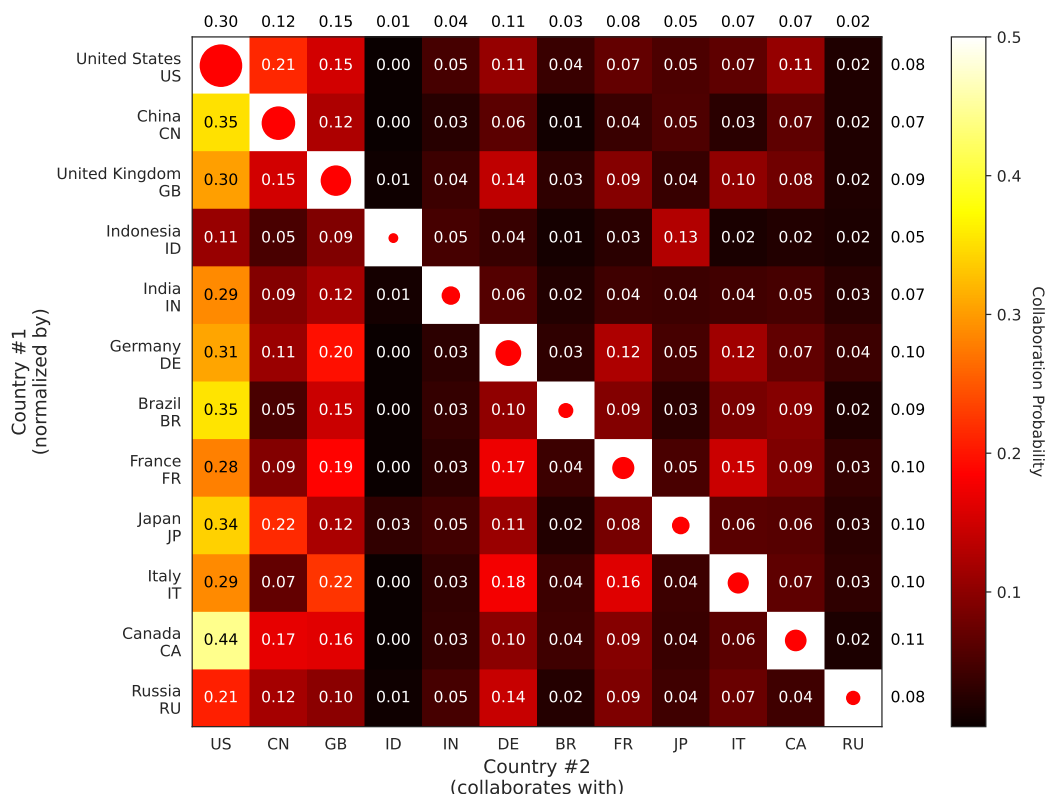


Fig. 18 A heat map illustrating global collaboration among the top 12 most prolific countries since 2018. Countries are arranged based on their scientific article productivity, with the y axis representing the top-down order and the x axis depicting the left-to-right order. The circular regions (highlighted in red) along the diagonal are proportional to the number of articles featuring multiple countries for Country #1. Each row is normalized by this count, converting each element within the heat map into a probability indicating how often a multi-country article for Country #1 (left) also contains Country #2 (below). The egocentric collaboration index is displayed to the right and the collaborative role index is displayed above the heat map.

First, the order of the countries on the left reflects their total article production over the specified time span, encompassing both articles with exclusive affiliation to that country and articles featuring multiple countries in their affiliations. This informative heat map format will be progressively unraveled, aiding in the comprehension of collaborative patterns.

Second, the focus shifts within the heat map to articles where the affiliations involve more than one distinct country. The red circles along the diagonal represent the number of articles featuring multiple countries that include Country #1. Specif-

ically, the size of the red circle at each country's position signifies the number of collaborative articles involving that country. For instance, observing this plot reveals that the US not only holds the top position for the overall number of articles but also boasts the highest number of collaborative multi-country articles, evident from the size of its corresponding red circle. Intriguingly, while Indonesia ranks fourth in total article production globally, its collaborative output with other countries is relatively modest.

Third, attention turns to the values and corresponding colors within individual elements of the heat map. These values denote the probability that Country #2 appears on a multi-country article that also includes Country #1. Importantly, this probability has been normalized by the total number of collaborative articles that Country #1 produces, introducing an asymmetry of colors within the heat map. To illustrate, consider the comparison between the US and China. While both countries have a similar frequency of occurrence in each other's collaborative articles, the normalization by the total output of articles with multiple countries reveals nuanced patterns. For instance, China is involved in 21% of collaborative articles from the US, whereas the US is present in 35% of China's collaborative articles. This asymmetry underscores that countries with lower total output often collaborate with those with higher article output, prominently observed in partnerships involving the US, China, and the UK. However, regional collaborative partnerships are also evident; European countries, such as the UK, Germany, France, and Italy, tend to engage in more collaborative endeavors within their region. Further analysis, clustering countries based on continents or geographical proximity, could likely unveil additional collaborative patterns.

The heat map introduces two indices aimed at quantifying collaborative patterns among top-producing countries. First, the Egocentric Collaboration Index (ECI) is displayed on the right, measuring the extent of collaboration from the perspective of each nation with other leading countries. This index is derived by averaging the probabilities within the corresponding row. Notably, the ECI exhibits less variation, offering a glimpse into countries that frequently engage in collaborations with other top nations, such as France, Japan, Italy, and Canada, while pointing out those with lower collaboration rates, like Indonesia (more details on this will follow). Second, the Collaborative Role Index (CRI) is displayed above, providing insights into the collaborative role of each country for other leading nations. Similarly, the CRI is

derived by averaging the probabilities within the corresponding column. The index above reveals that the US, for instance, is a prominent collaborative partner for other leading countries, contrasting with Indonesia, which engages less frequently in collaborations with other top nations. Instead, Indonesia fosters partnerships with regional counterparts like Malaysia, Japan, Australia, Taiwan, China, Thailand, India, Korea, and others (not shown). The CRI is an average probability that that country is on multi-country publications from other leading countries, i.e., when other leading countries publish a multi-country scientific article, the US is on that paper 30% of the time for the time frame given. Despite being averages of probabilities within the heat map, caution must be exercised when interpreting the nuanced meanings of these indices, as they offer valuable insights into discernible trends. For instance, while the CRI can be interpreted as an average probability, the ECI is distorted by articles that contain more than one other leading country, i.e., the sum of the probabilities can exceed 100%. If the probabilities from all other countries were used for the ECI, it would be proportional to the average number of distinct countries on its multi-country articles; herein, the ECI includes a subset of only the leading countries shown in the heat map.

Returning to Indonesia, despite ranking fourth in total publications over the last five years, it surprisingly has the fewest collaborations with other countries among the top 12 producing nations. One potential factor contributing to this trend is the substantial number of scientific articles written in Indonesian. Notably, since 2018, Indonesia has produced the highest count of non-English scientific articles (495,700 articles, predominantly in Indonesian). Brazil follows as the second-highest contributor to non-English publications (421,800 articles, mainly in Portuguese), with a significant drop in non-English articles for other nations (France ranks third with 182,000 articles in French). Both Indonesia and Brazil exhibit low CRI index values (columns), akin to Russia and India, even though the latter two primarily publish articles in English. Supporting this observation, when Indonesia and Brazil engage in collaborations with other countries, 85% of the time, the article is written in English. In contrast, when the article features only these countries in the affiliations (i.e., only one distinct country), the language switches to English only 52% of the time. Figure 19 presents the heat map with a focus on articles written exclusively in English. Consequently, Indonesia shifts from the fourth to the eighth position in total articles produced, while Brazil is replaced in the top 12 by Australia. For the remainder of this study, the analysis concentrates on articles written in English.

Global Collaboration Heatmap for All English Articles (2019–present)

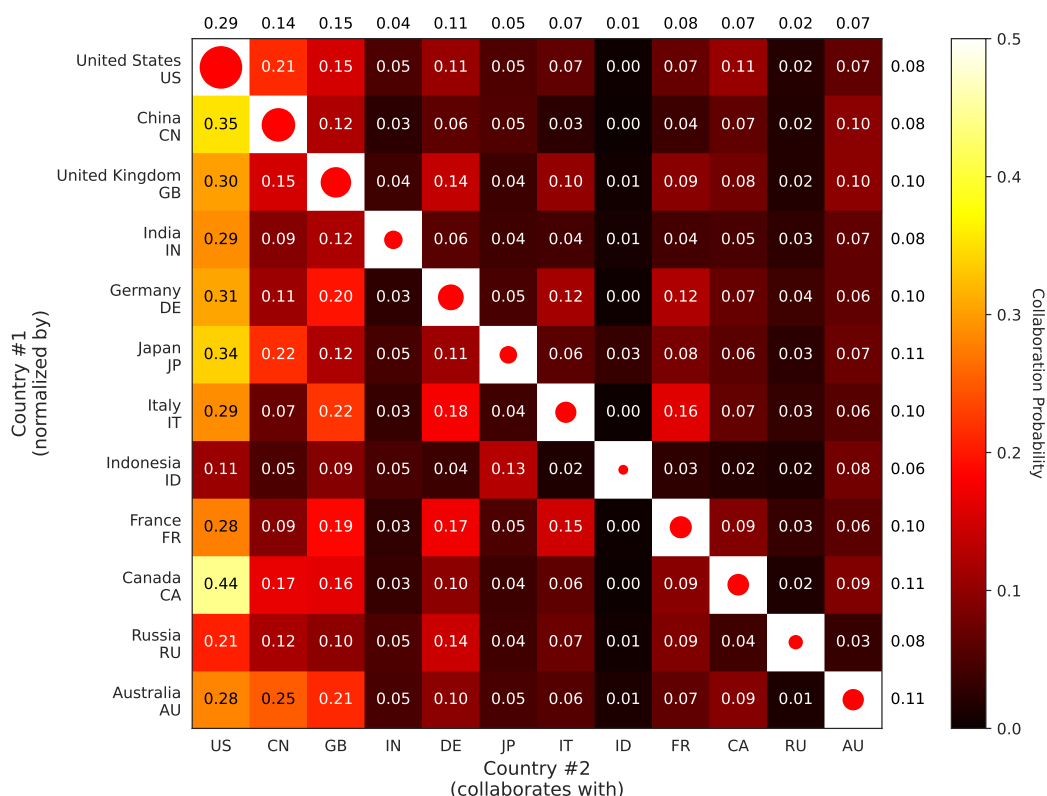


Fig. 19 A heat map illustrating global collaboration among the top 12 most prolific countries since 2018. The formatting is similar to Fig. 18 with the only difference being that only articles written in English are used to calculate the top 12 countries and the global collaborations.

To delve deeper into the nuances of the ECI and the CRI, Fig. 20 illustrates Venn diagrams representing the collaborative relationships of the US with China and the UK, and with China and Russia. Each colored region in the Venn diagram signifies a specific subset of multi-country publications, with the number indicating the total count for that subset. For instance, the large red section in the US circle denotes the number of multi-country publications since 2018 that lack China or the UK as an institutional affiliation of a co-author (in Fig. 20a). Similarly, shaded regions indicating overlap between two countries exclude the third country, while the intersection of all three countries represents multi-country publications where affiliations from each country exist for the co-authors.

Regarding the ECI for the US, it entails the sum of the overlap subsets between the US and every other country, divided by the total number of multi-country publications produced by the US. Notably, in the example Venn diagram, the region

Venn Diagram for Multi-Country Collaborations

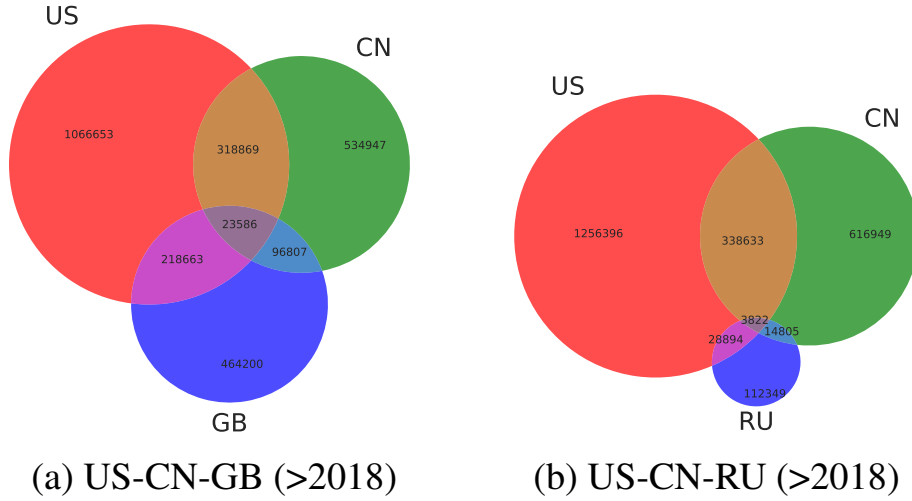


Fig. 20 A Venn diagram showing the numbers of multi-country publications since 2018 for (a) US-CN-GB and (b) US-CN-RU. The numbers represent the total number of multi-country articles in each colored region, e.g., the US has 1,066,653 multi-country publications without CN or GB as a co-author, the US has 218,663 multi-country publications with GB and without CN, and the US has 3,882 multi-country publications with both CN and RU.

where all three countries are listed on a publication is counted twice for the ECI exhibited for these three countries. However, when all leading countries are included, the complexity of numerous overlapping regions would render a 2D Venn diagram impractical (but the overlap subsets between more than 2 countries exists and is counted multiple times for the ECI).

On the other hand, the CRI for the US is computed from the overlap subsets between the US and every other leading country, with each subset divided by the total number of multi-country publications for the other countries. These two indices can differ significantly due to variations in the total number of publications for each country. For instance, in Fig. 20b, collaborative publications between the US and Russia are approximately an order of magnitude lower than those with the US and China. From the US's (and China's) egocentric perspective, collaboration with Russia appears minimal, while from Russia's egocentric viewpoint, collaboration with the US (and China) represents a more substantial fraction of their total multi-country publications. Thus, the average of this probability from the perspective of all collaborating countries, as represented by the CRI, offers insights into how often these other countries collaborate with the country at hand. With a comprehensive grasp of the ECI and CRI, depicted in the accompanying Venn diagrams, the focus

now returns to the examination of various heat maps. These visualizations continue to unravel and analyze global collaboration patterns, offering a nuanced perspective on collaborative relationships among leading nations.

3.3.5 Global Collaboration between Leading Countries over Time

Figures 21–23 present collaboration heat maps spanning the last three decades: the 1990s, the 2000s, and the 2010s, respectively. Throughout this period, collaboration among the top countries has shown a modest increase. Notably, the US and the UK consistently maintained leading positions in scientific article production. A pivotal shift occurred with China’s ascent, transitioning from 9th globally in the 1990s to the 2nd position in the 2010s. Beyond the sheer increase in article production, China’s engagement in articles involving multiple nations saw a substantial rise, evident in the CRI value above the heat map (shifting from 0.03 in the 1990s to 0.11 in the 2010s).

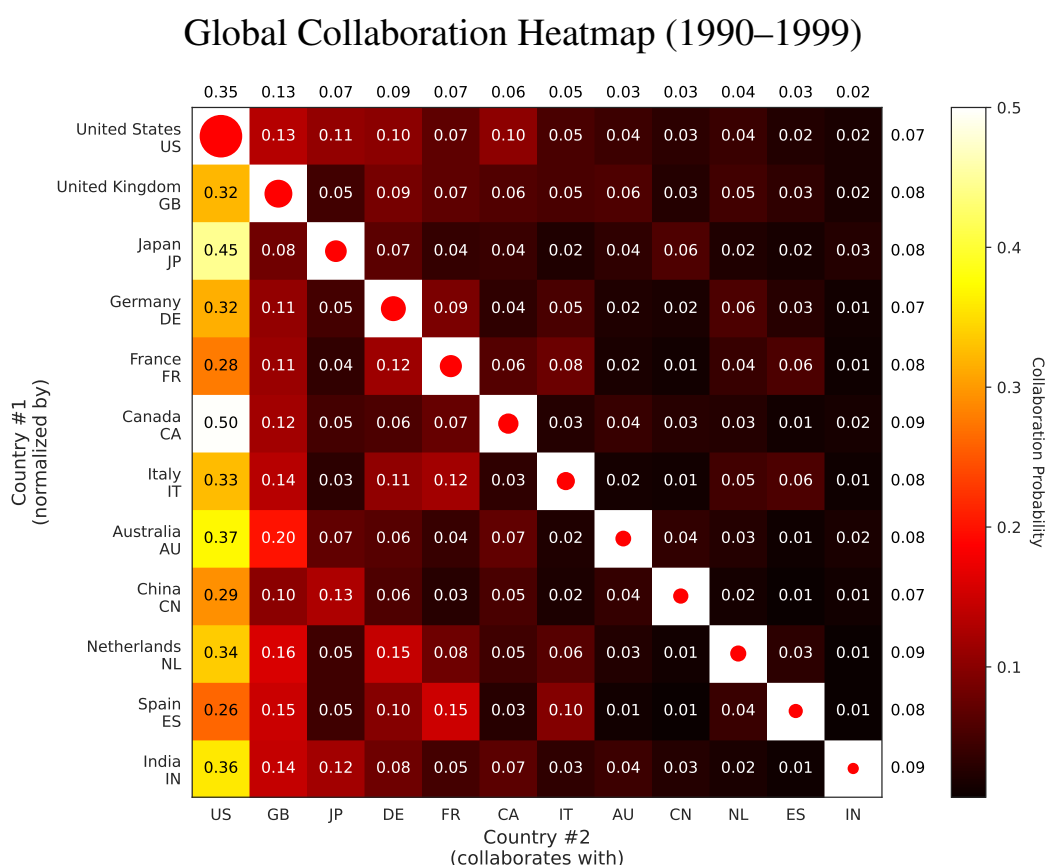


Fig. 21 A heat map illustrating global collaboration among the top 12 most prolific countries in the 1990s. The formatting is similar to Figs. 18 and 19.

In the 1990s, several countries in the bottom half of the top 12 had relatively fewer

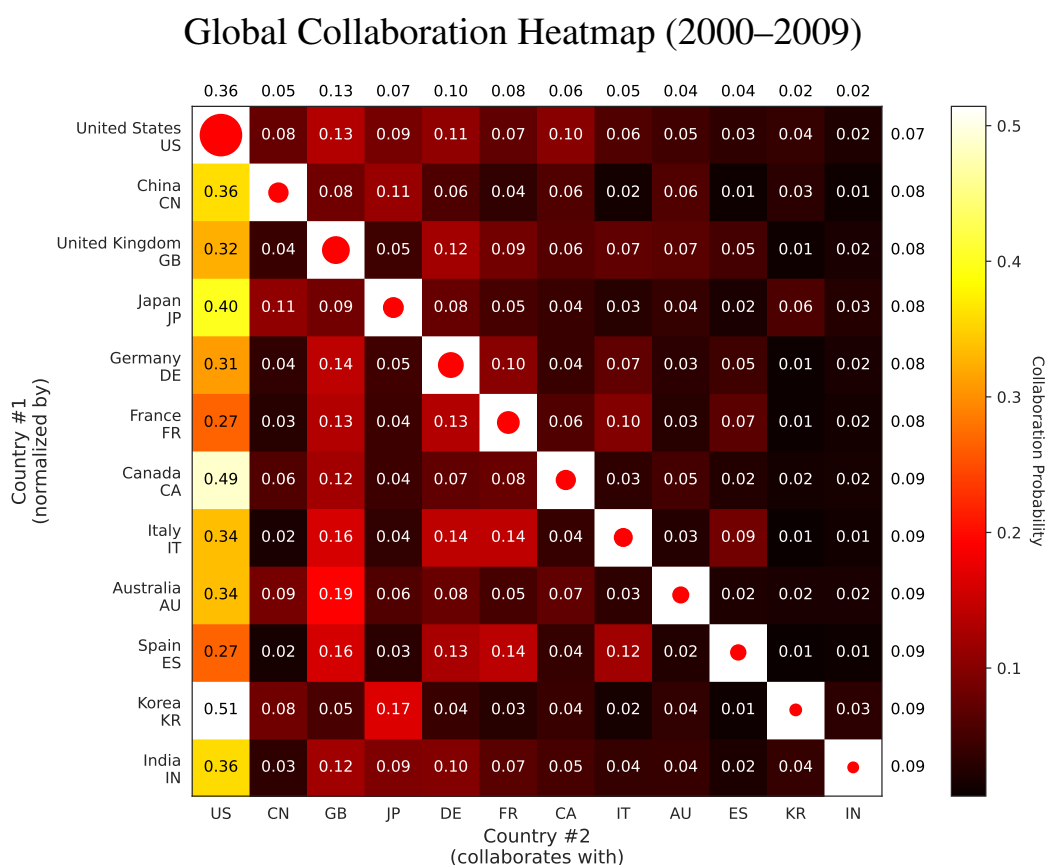


Fig. 22 A heat map illustrating global collaboration among the top 12 most prolific countries in the 2000s. The formatting is similar to Figs. 18 and 19.

collaborations within this group. However, scientific collaboration has witnessed a general uptrend over the decades, with the exception of India, which experienced a stagnation in collaborations despite an increase in scientific articles. The dynamics of the Top 12 have also seen countries moving in and out, influenced by closely matched scientific outputs. It's important to note that the chosen years are somewhat arbitrary, and countries at the border of the Top 12 might have shifted under different year ranges.

Across the decades, certain countries consistently rank among the world's top contributors to scientific output: the US, China, UK, Germany, Japan, France, Canada, Italy, Australia, and India. The recent emergence of Indonesia, as discussed earlier, adds an intriguing dimension, suggesting a rapid increase in the scientific presence of Southeast Asian countries.

This heat map serves as a dynamic lens into the evolving collaboration between

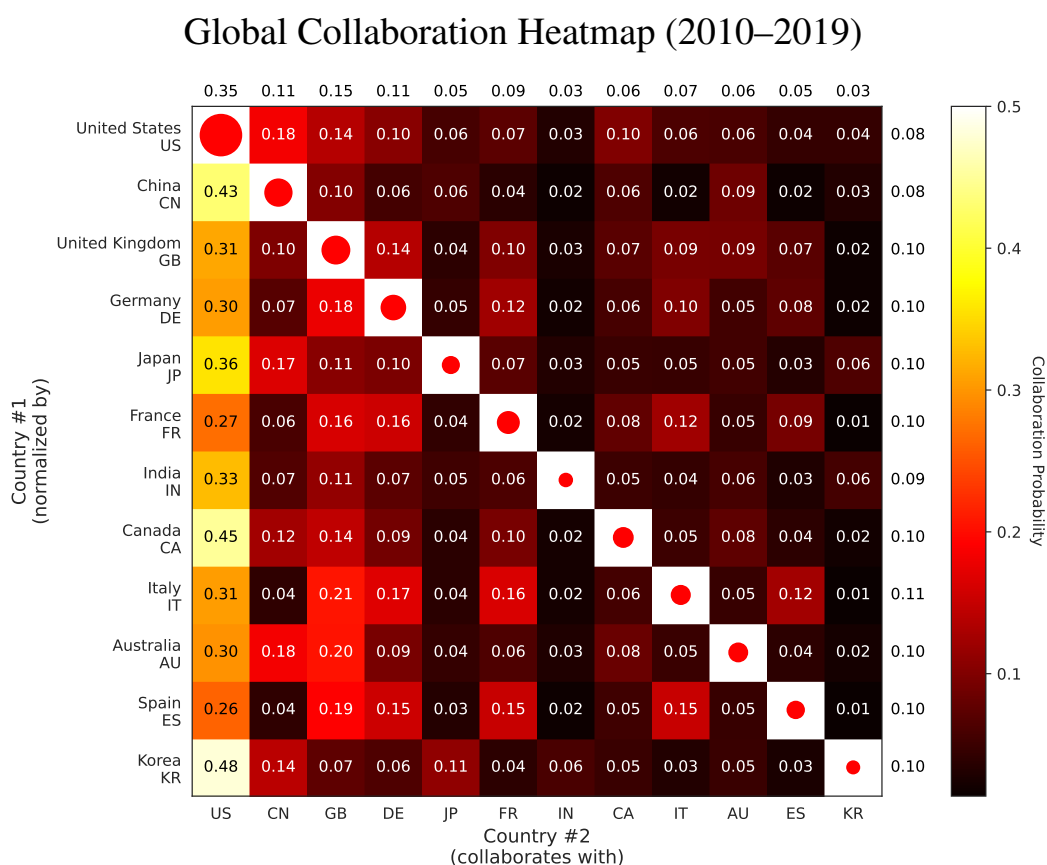


Fig. 23 A heat map illustrating global collaboration among the top 12 most prolific countries in the 2010s. The formatting is similar to Figs. 18 and 19.

country pairs. An intriguing example lies in the asymmetry within the relationship between the US and China. In the 1990s, collaborative articles involving the US and China accounted for 3% and 29%, respectively. This asymmetry persisted into the 2000s, with percentages at 7% (US articles with China) and 36% (China articles with the US). The trend continued into the 2010s, with percentages growing to 18% (US articles with China) and 43% (China articles with the US). Notably, in Fig. 19 (2018 to the present), these percentages further increased to 21% and 35%, respectively.

While the focus here was on the US-China relationship and its asymmetry, a broader pattern emerges across various examples: the greater the discrepancy in scientific productivity between two countries, the more pronounced the asymmetry in the percentages of collaborative articles. It’s essential to acknowledge that higher percentages could also be influenced by the collaborative output of countries with higher overall article production. The Matthew effect, also known as “the rich get richer,”

seems to apply to the total scientific productivity of countries in its global collaboration with other countries. This does not necessarily mean that high-producing countries have more collaborative scientists (by percentage of total), just more of them. This greater number of scientists leads to a higher probability that scientists from other countries will interact with them at conferences or the like, leading to more multi-country collaborations and publications. The global shift to virtual conferences, meetings and seminars following COVID-19 enables even more opportunities for scientists to engage with scientists around the world without the requirement of physical presence.

3.3.6 US Institution Collaboration with Leading Countries

Turning the focus to a more granular scale, the exploration delves into global collaboration at the institutional level, as illustrated in Fig. 24. The y axis of this heat map features a myriad of institutions, predominantly federal laboratories and federally funded R&D centers within the US, maintaining similarity with previous heat maps. Each element within the heat map represents a probability, with the denominator being the total number of publications by institutions whose authors list more than one distinct country in their affiliations. The numerator signifies the count of publications listing the country corresponding to the bottom of the heat map within at least one author's affiliation.

For instance, 41% of the National Energy Technology Laboratory's multi-country papers involve China, while 31% of JHU Applied Physics Laboratory's such papers are in collaboration with the UK and France. The CRI index (above) the heat map maintains a similar interpretation as earlier, gauging the extent of collaborative papers with the country at the bottom for all institutions listed. It reveals that a substantial portion of collaboration involves countries such as China, the UK, Germany, Japan, Italy, France, Canada, Australia, Spain, and Korea. The ECI (right) has a similar interpretation as well, except now it is for the institutions at the left rather than for countries.

In terms of scale, this heat map, expressed in probability values rather than frequencies, prompts a question about the total volume of collaborative articles involving more than one country and these US institutions over the given timeframe. The answer: 55,150 articles. Among these, NETL has the smallest number of collaborative articles (300). Notably, out of the 55,150 articles, only 3,268 are published

US Government Institutions Collaboration Heatmap

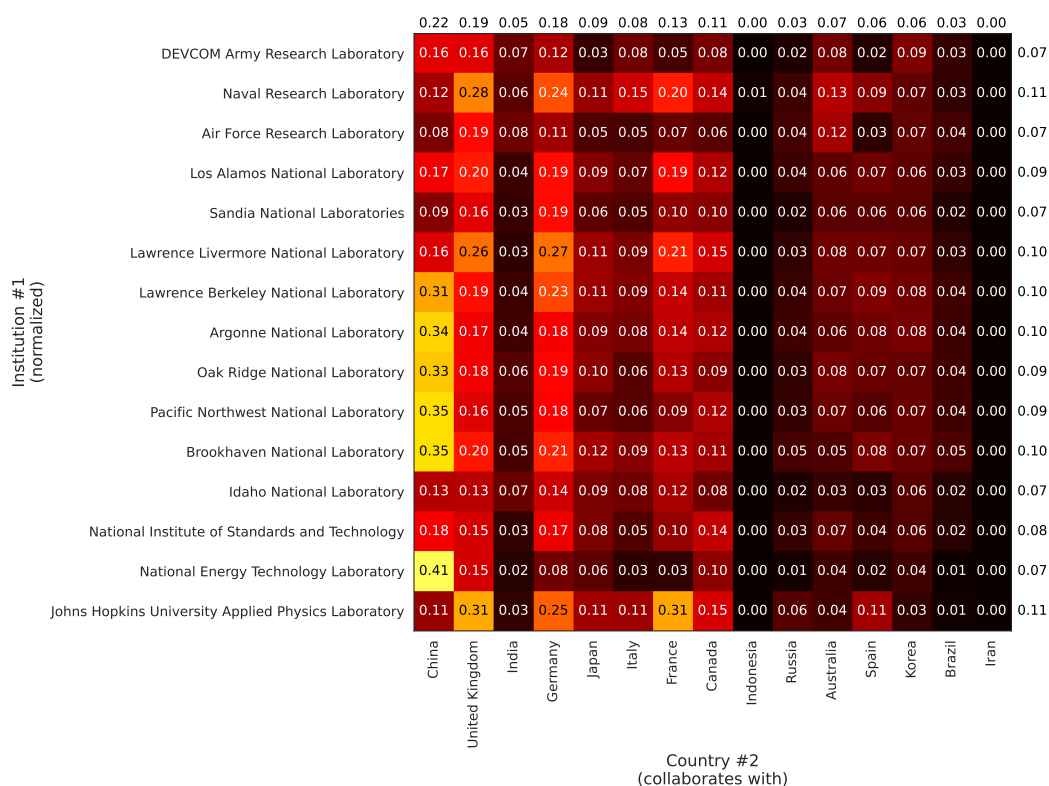


Fig. 24 A heat map illustrating US institutional collaboration with the top 15 most prolific countries (2018–2023). The formatting is akin to Figs. 18 and 19, with the y axis now featuring various government and federally-funded research and development centers. Probabilities in the heat map signify the percentage of papers with more than 1 distinct country listing an author affiliated with an institution from the respective country. Color bar (removed herein) is similar to past heat maps, ranging from 0 to 0.50.

without an academic institution as a collaborator. This highlights that most publications within these institutions, involving collaborations with other countries, also feature an academic collaborator, either in the US or abroad.

This prompts an inquiry into how these collaborations compare to those involving academic institutions. In Fig. 25, a heat map depicting global collaboration is presented with the y axis featuring the top-producing academic institutions from 2018 to 2023. The range of articles published, listing more than one distinct country during this period, spans from 27,310 articles (UC-San Francisco) to 95,330 articles (Harvard University), aggregating to 458,330 articles across all these institutions.

Comparing this to Fig. 24, the collaboration targets for academia closely align with those in the national laboratory ecosystem. These academic institutions, operat-

US Top Academic Institutions Collaboration Heatmap

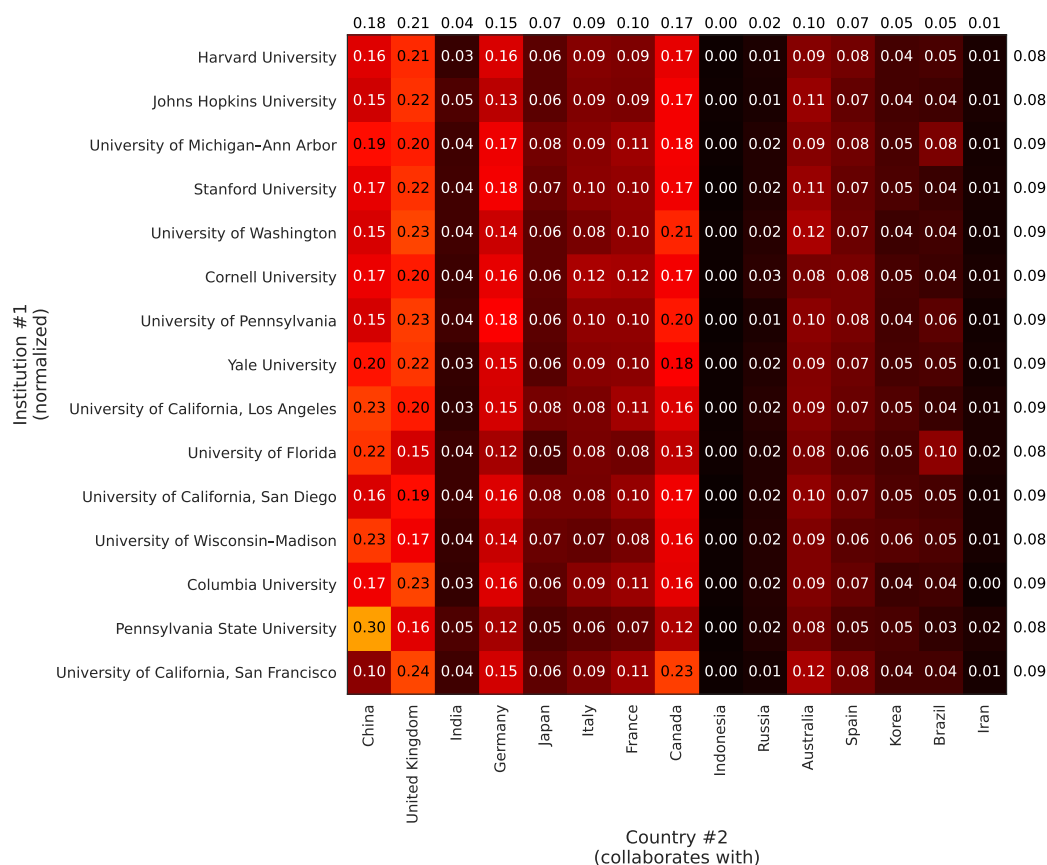


Fig. 25 A heat map illustrating US academic institutional collaboration with the top 15 most prolific countries between 2018–2023. The formatting is similar to Fig. 24, except now the y axis features the top producing academic institutions over this time period (ordered top to bottom). Color bar (removed herein) is similar to past heat maps, ranging from 0 to 0.50.

ing across diverse fields, consistently engage in global collaborations, notably with China, the UK, Germany, Japan, Italy, France, Canada, Australia, Spain, Korea, and Brazil. It’s essential to note that these institutions represent some of the top academic producers.

However, considering the broader academic landscape in the US, Fig. 26 unveils a heat map of top institutions as defined by the ECI values on the right of the plot. Here, these academic institutions demonstrate high probabilities of collaboration with multiple countries, in particular China, the UK, Germany, and Canada. It’s crucial to clarify that the institutions listed in Fig. 26 are not ranked based on the frequency of publications with more than one distinct country. For instance, the University of Chicago (23,870 articles) and California Institute of Technology

(17,560 articles) publish over an order of magnitude more articles from 2018–2023 than Northern Illinois University (1,367 articles) or Brandeis University (1,568 articles).

US Collaborative Academic Institutions Collaboration Heatmap

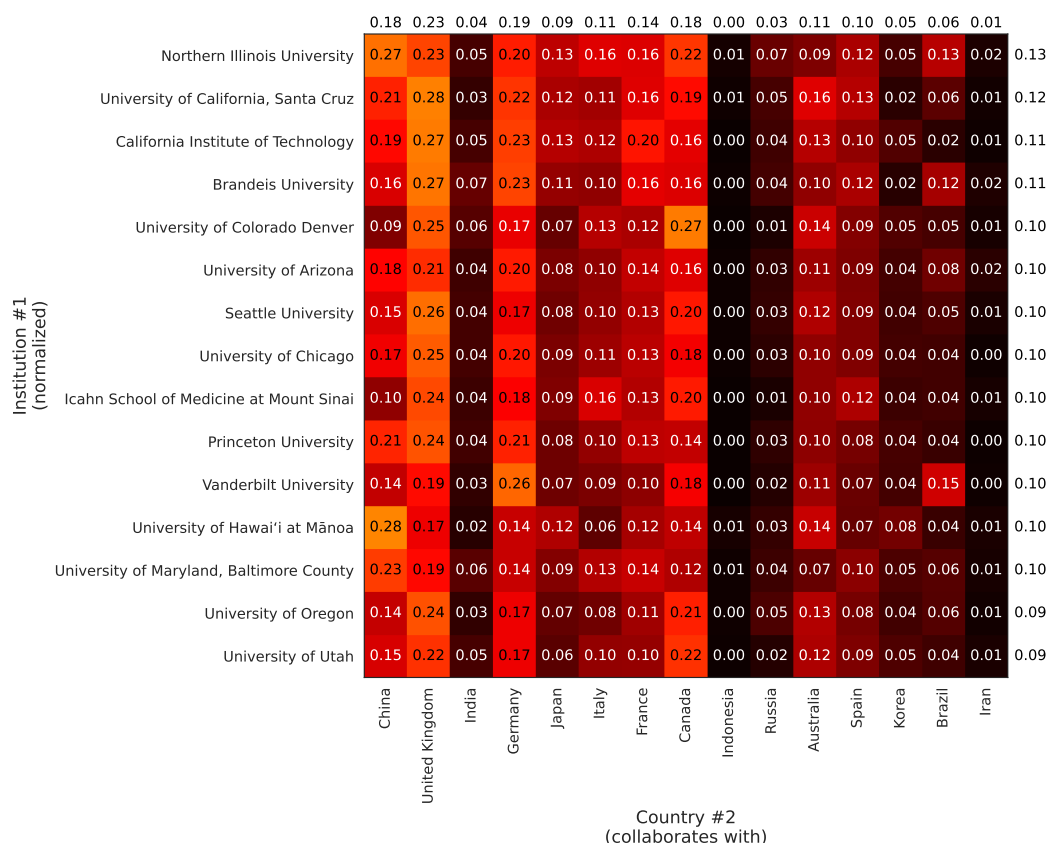


Fig. 26 A heat map illustrating US top collaborative academic institution collaboration with the top 15 most prolific countries between 2018–2023. The formatting is similar to Figs. 24 and 25, except now the y axis features the academic institutions with the highest collaboration index (right-hand side) over this time period (ordered top to bottom).

An intriguing aspect to explore within this dataset is the role of funding agencies that underpin the research leading to global collaborations. Figure 27 presents a heat map showcasing funding agencies acknowledged in publications with affiliations across different countries. This encompasses notable entities such as the NSF, the National Institutes of Health, the Department of Energy (DOE), various defense agencies (Army Research Office [ARO], Office of Naval Research [ONR], Air Force Office of Scientific Research [AFOSR]), and various government laboratories supporting both basic and applied research and development.

Distinguishing itself from previous heat maps, this visualization considers the pro-

US Funding Agency Heatmap

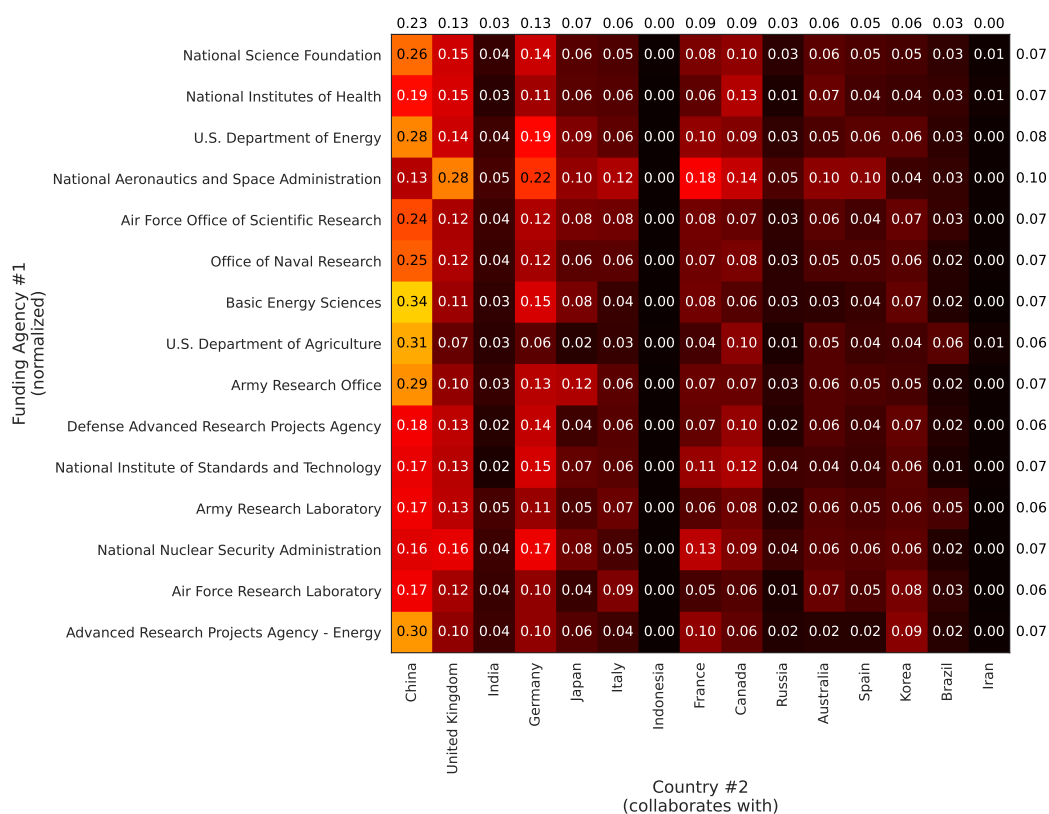


Fig. 27 A heat map illustrating US agency funding of collaborative articles with the top 15 most prolific countries between 2018–2023. The formatting is similar to Fig. 18, except now the y axis features multiple US government agencies who are listed within the acknowledgments section as partially funding the research. The probabilities within the heat map represent the percentage of papers (funded by the agency) with more than 1 distinct country that list an author with an institution from the country on the x axis.

cess where Principal Investigators (PIs) submit proposals to federal agencies, which are then reviewed and funded (or declined). Following approval, the PI carries out the research outlined in the proposal, often engaging with researchers from other countries to enhance the project. Consequently, the funded work contributes to publications, with authors frequently acknowledging the supporting funding agencies upon publication.

Remarkably, funding agencies primarily devoted to basic research, such as NSF, DOE and DOE Basic Energy Science, AFOSR, ONR, and ARO, exhibit substantial percentages of globally collaborative publications, particularly with countries like China, the UK, Germany, France, or Canada. The first column of the heat map illustrates collaborations with China. Notably, China is less frequently listed as a

collaborator for federal agencies like National Aeronautics and Space Administration (NASA), Defense Advanced Research Projects Agency (DARPA), National Institute of Standards and Technology (NIST), the National Nuclear Security Administration (NNSA), and agreements originating directly from federal defense laboratories.

NASA stands out with the highest collaboration rates between its PIs and researchers from other countries (highest ECI value), especially with significant percentages involving the UK, Germany, and France. This heightened collaboration may be attributed, in part, to the nature of research conducted, particularly in fields like astronomy and astrophysics where hyper-authorship has become more prevalent, resulting in author lists comprising hundreds or even thousands of contributors from various countries.

Exploring global collaboration at the institutional level reveals nuanced complexities. For instance, national laboratories frequently collaborate with academic institutions, and both entities demonstrate substantial partnerships with other countries. Consequently, collaborations between institutions within the US emerge as a confounding factor when attempting to pinpoint the primary driving force behind global collaborations. This complexity echoes the challenges of credit allocation for individual authors within multi-author publications, extending the difficulty to attributing the predominant factor for collaborations with other countries within multi-institution publications.

The intricacies persist when considering funding agencies, where attributing contribution to global collaborations becomes intricate. Many publications involve numerous authors from various institutions, each acknowledging funding from different agencies. Adding to this complexity, different funding agencies allocate resources across a portfolio of research areas, some of which naturally involve more global collaboration. This intricate web of interactions raises a follow-up question, which will be explored in the next section: How do research areas contribute to multi-country collaborative publications?

3.3.7 Global Collaboration for Research Concepts

In this section, the impact of research areas, or research concepts, on global collaboration is explored. The research area can affect various factors such as the total number of published papers, citation rates, authorship distribution, and even the

collaborative dynamics within a specific research field. Before exploring this effect, first look into the process of mapping papers to research concepts within OpenAlex.

Concepts, in this context, represent abstract ideas tied to publications. These were first called “fields of study” in Microsoft Academic Graph to represent the semantic contents of the document.³ They are organized in a hierarchical structure across five levels. Level 0 encompasses broad concepts like materials science, engineering, computer science, and sociology (totaling 19 concepts), while level 5 delves into much more specific areas such as pandemic, coronavirus, black hole (networking), chirality (physics), and rectifier (neural networks). Employing an automated classifier, OpenAlex assigns concepts to works—over 65,000 concepts in total—based on titles, abstracts, and the interconnections among concepts.

This ML algorithm predicts a similarity score between the article’s text and each concept. After obtaining scores for all concepts, a threshold is applied to finalize concept assignments, as depicted in Fig. 17. Remarkably, over 99.9% of articles have at least one concept assigned. Each assigned concept is presented as a list accompanied by details such as a concept score, name, identifier, ancestors, related concepts, and a link to the corresponding Wikidata item.

This section primarily centers on level 0 concepts, yet its applicability extends seamlessly to a diverse array of concepts within the expansive hierarchical network. Before probing the details, it’s essential to underscore the hierarchical structure of the concept graph. To illustrate this, examine a selection of level 0 concepts and their interconnections with concepts at the immediately lower level, level 1.

Figures 28 and 29 present Sankey plots that visually capture the relationships between a level 0 concept and its level 1 counterparts. For this illustration, two exemplars were chosen from the top level 0 concepts: material science and computer science. A striking observation is the substantial number of level 1 concepts associated with each level 0 concept. This pattern holds true for all 19 level 0 concepts, with a slightly reduced count for some of the social science and humanities concepts at level 1.

The width of the lines in these Sankey plots reflects the predicted score derived from the title and abstract, signifying the similarity between different concepts. It’s noteworthy that a high score for a lower-level concept can influence the scores and

Materials Science Level 0 Concept Sankey Plot (by score)

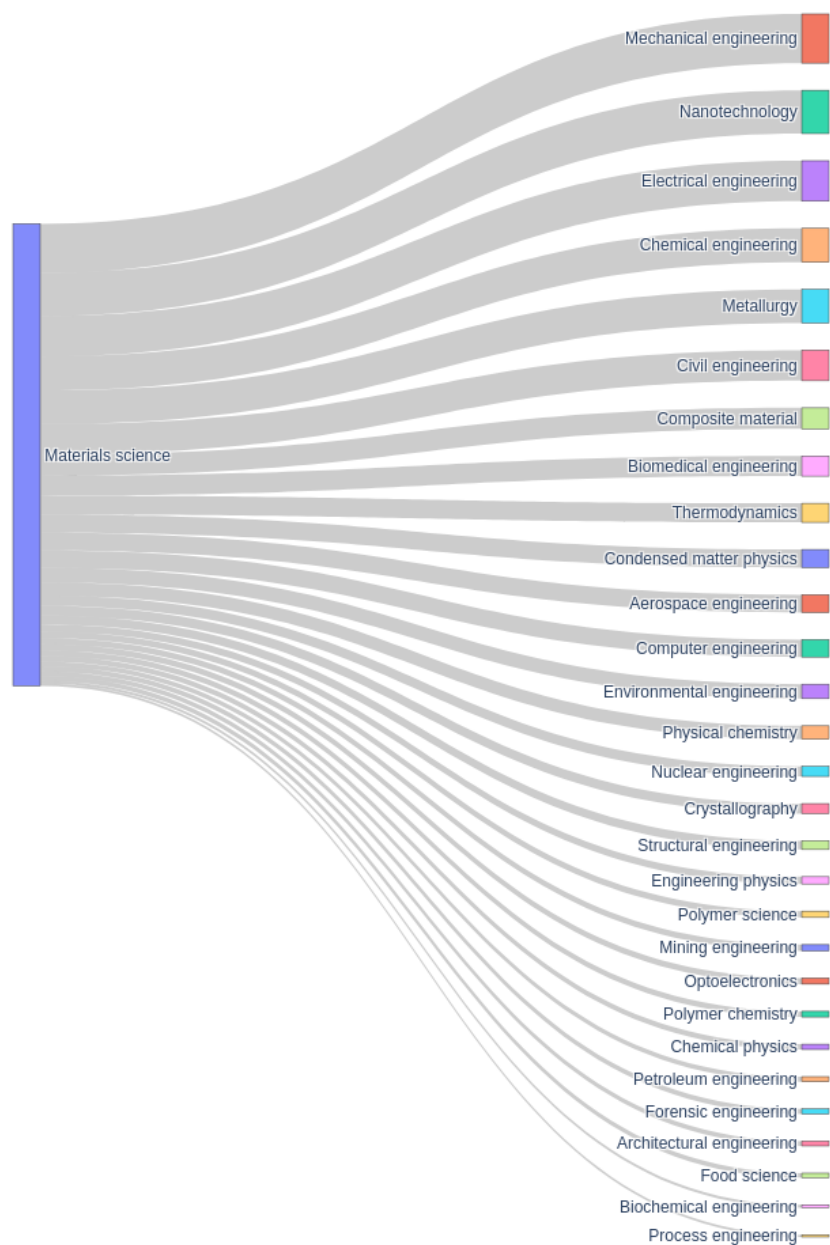


Fig. 28 A Sankey plot illustrating the connections from the level 0 concept “material science” (on the left) to its diverse level 1 counterpart concepts (on the right). The thickness of the connecting lines in this network is determined by the predicted score within OpenAlex, serving as a metric that indicates the similarity between concepts.

inclusion of the concepts above it. Consequently, these upper-level concepts are often listed, even if their individual scores fall below the inclusion threshold.

Now equipped with insights into the realm of research concepts within the database

Computer Science Level 0 Concept Sankey Plot (by score)

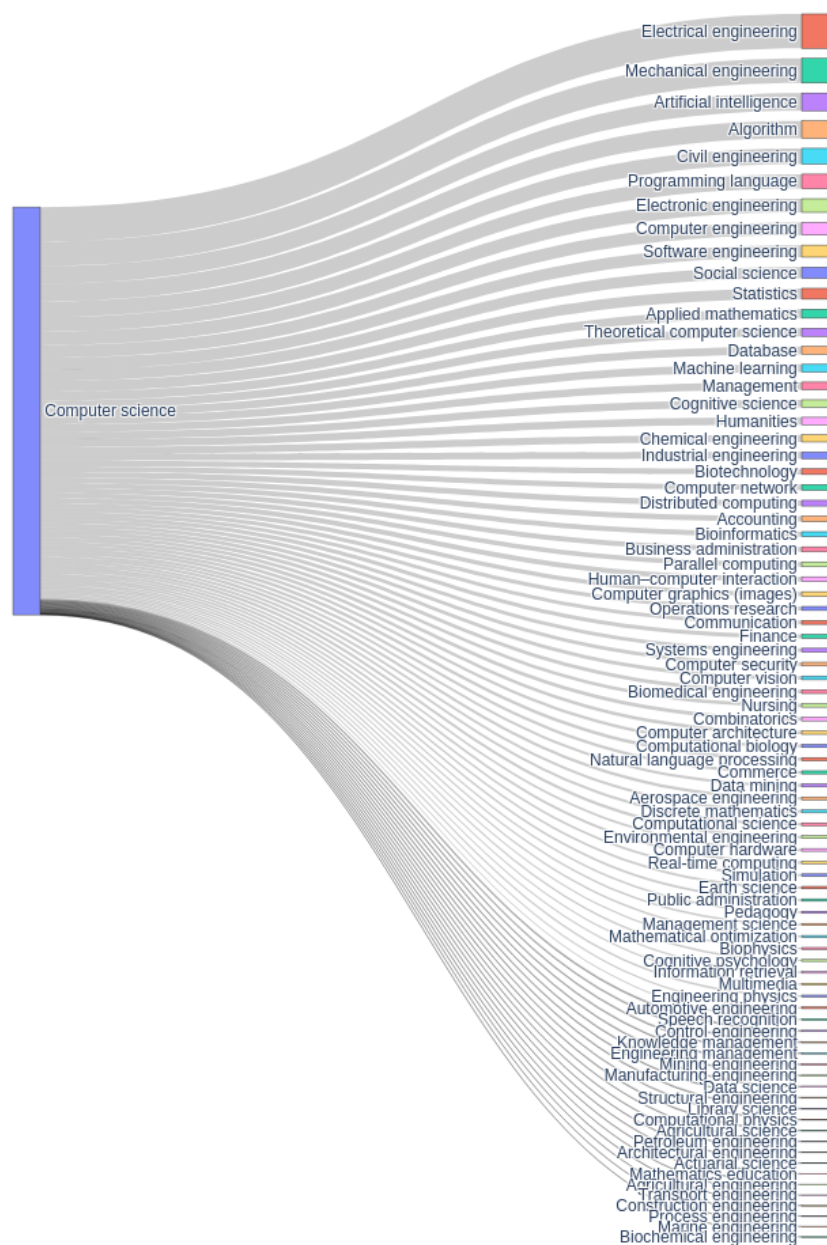


Fig. 29 A Sankey plot depicting the connections originating from the level 0 concept “computer science” (positioned on the left) to its various level 1 counterpart concepts (displayed on the right). The line widths in this network are determined by the predicted scores within OpenAlex, reflecting the similarity between the associated concepts.

and the mapping of articles to these concepts, the intricate interplay between these concepts and global collaboration can be dive into. In Fig. 30, a heat map unveils the landscape of collaboration across articles associated with level 0 concepts. The con-

cepts are arranged alphabetically, while the countries along the bottom are strategically ordered based on the production of multi-country articles spanning the 2018-to-2023 timeframe.

The values within the heat map maintain a familiar format, aligning with prior encounters with such visualizations. To illustrate, consider the level 0 concept “engineering.” Within this domain, 28% (or 0.28) of all articles with multiple countries affiliated are attributed to the US, while China claims a comparable share of 29% (or 0.29). Shifting focus to the domain of medicine, the US emerges as a dominant force, contributing to 43% of all globally collaborative articles, surpassing other nations significantly (with UK second at 21%).

The CRI value above the heat map is similar to past measures, except it is now quantifying the propensity for global collaboration across all concept areas, with the US boasting the highest index (0.33). Meanwhile, the ECI adorning the right side of the heat map sheds light on the frequency with which top countries engage in multi-country publications for each concept. This index exhibits minor fluctuations, indicating that the relationship between concepts and countries is similar from the self-centered concept perspective.

In Fig. 31, the US’s global collaborations within distinct level 0 concept areas are explored. The heat map unveils that the US engages in substantial collaboration, particularly with China, followed by robust associations with the UK, Germany, and Canada. Notably, STEM fields, including material science, physics, engineering, chemistry, mathematics, and biology, exhibit the highest collaboration probabilities in US-China partnerships. Intriguingly, US-UK collaborations showcase elevated probabilities in non-STEM areas like history, political science, sociology, economics, and business.

Figure 32 delves into China’s collaborative landscape across level 0 concept areas, revealing distinctive patterns compared to the US. China demonstrates a predilection for strong collaborations with fewer countries across diverse areas. Foremost among its collaborators is the US, with heightened probabilities in fields such as medicine and biology. While the UK and Australia also emerge as significant collaborators, their contribution remains a fraction of the collaborative output seen with the US.

Global Collaborations in Level 0 Concept Areas (2018–2023)

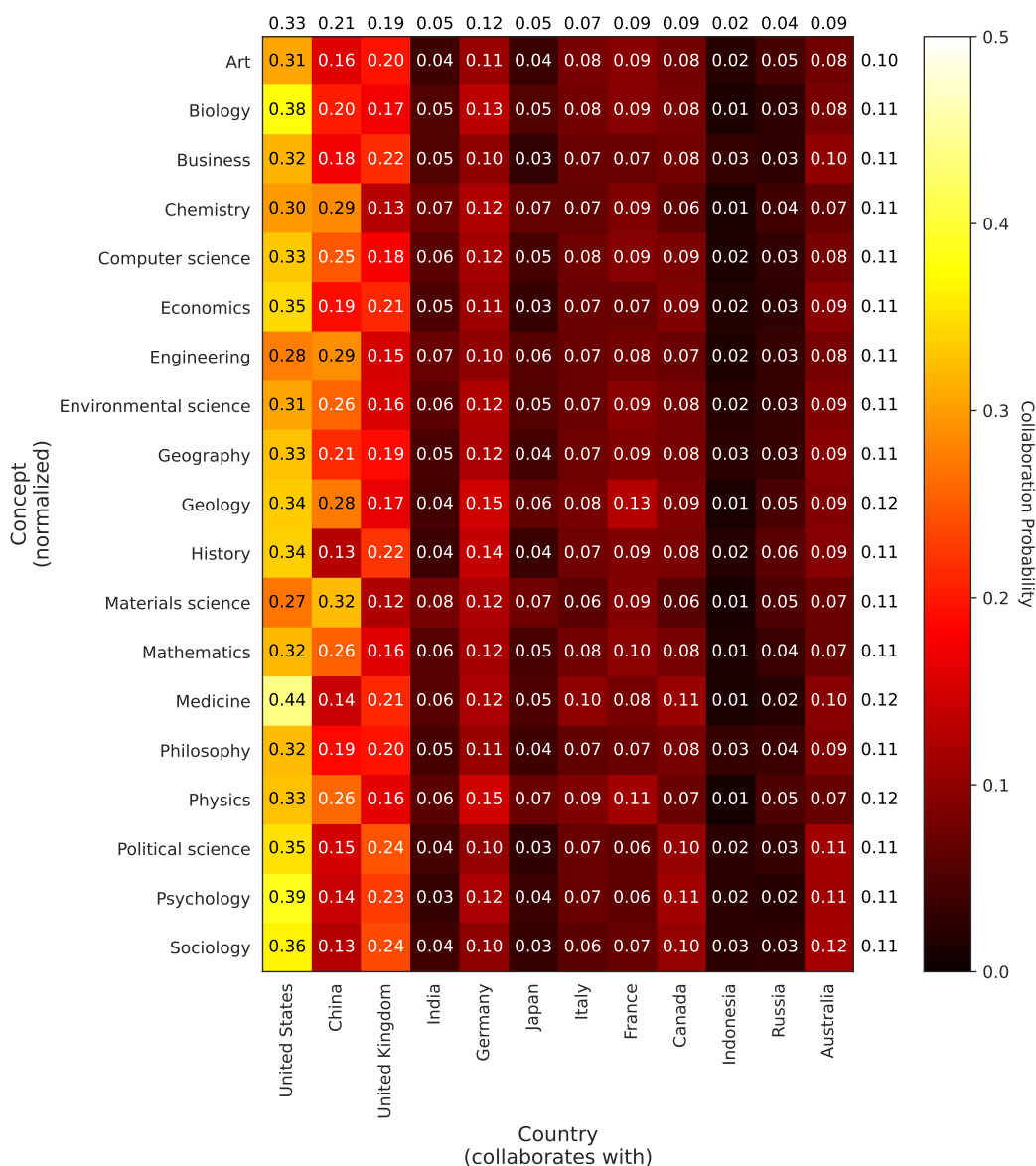


Fig. 30 A heat map unfolding the intricacies of global collaboration from 2018–2023, anchored in the level 0 concepts associated with articles within OpenAlex. The y axis showcases concept names arranged alphabetically, while the x axis reveals the leading countries in producing multi-country articles. Values within the heat map denote the probability that the country at the bottom is affiliated with a multi-country publication linked to the left-listed concept. The summed index above each column elucidates the degree of each country’s entwinement with publications in these concept domains, while the right-side index compiles row probabilities, delineating the top countries’ association with publications tagged with these concept areas.

The collaborative picture of the UK, illustrated in Fig. 33, mirrors some aspects of both the US and China. Similar to the US, the UK showcases elevated probabilities

US Collaborations in Level 0 Concept Areas (2018–2023)

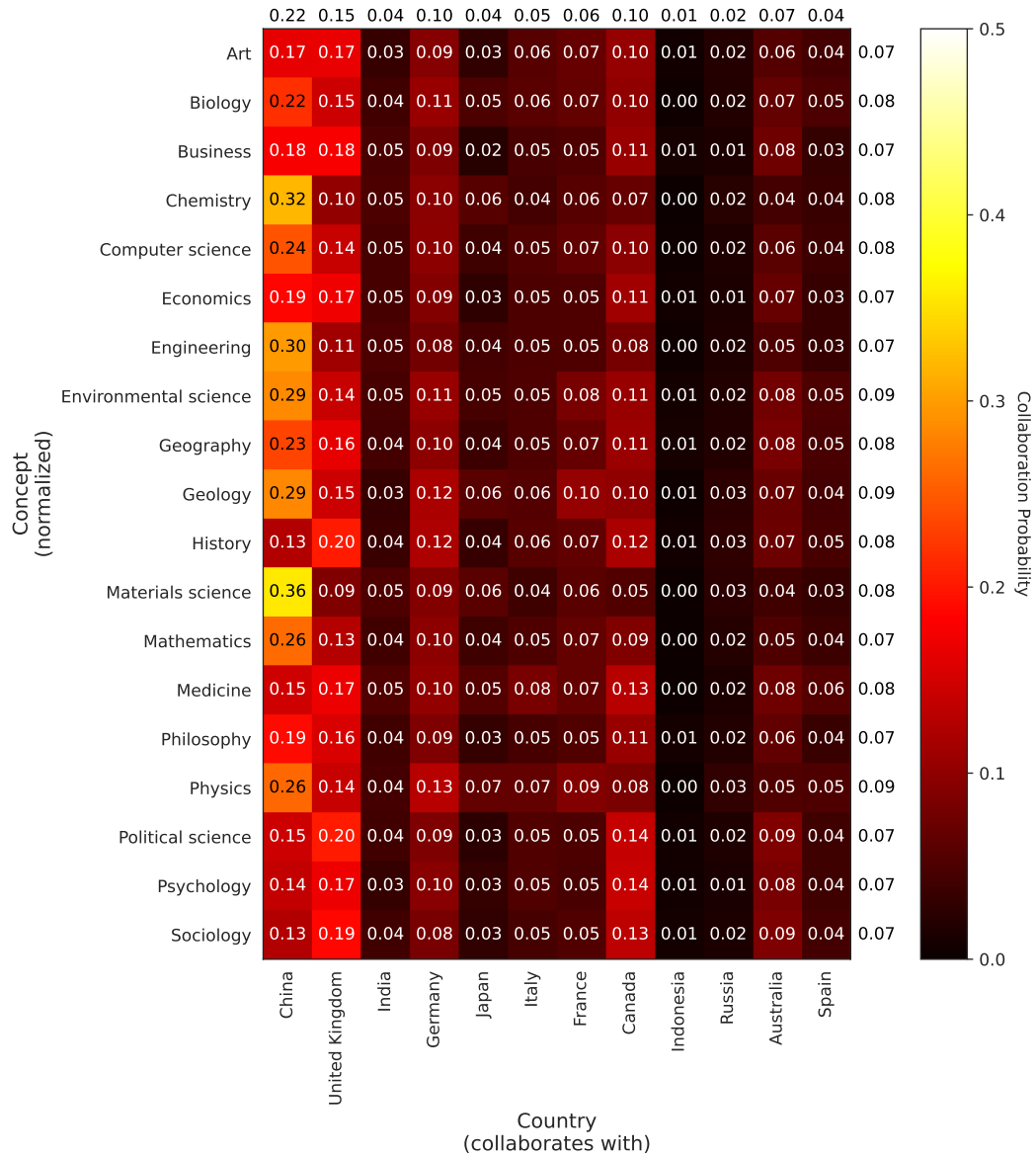


Fig. 31 A heat map uncovering the dynamics of global collaboration from 2018–2023 for articles associated with the US. The y axis presents concept names alphabetically, while the x axis reveals the leading countries in producing multi-country articles. Values within the heat map represent the probability of the US being affiliated with a multi-country publication related to the left-listed concept. The summed index above each column illuminates the extent of the US’s involvement in publications across these concept domains.

in collaborations with the US and China, but the heat map also highlights robust partnerships with European counterparts, Canada, and Australia. Intriguingly, UK-US collaborations lean towards medicine and biology, aligning with China's preferences, while UK-China collaborations predominantly gravitate towards material science and engineering, surpassing collaborations even with the US.

China Collaborations in Level 0 Concept Areas (2018–2023)

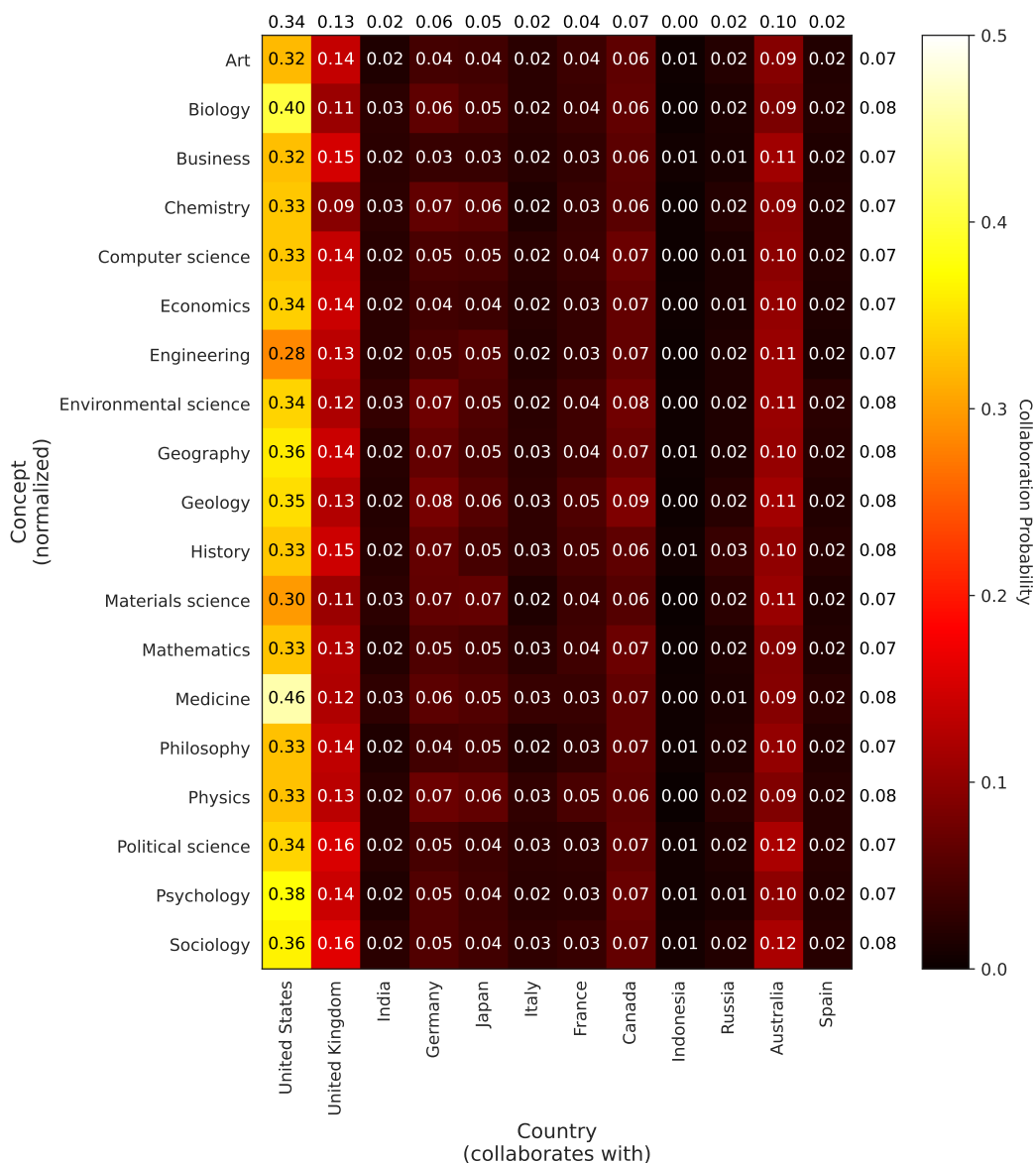


Fig. 32 A heat map delving into global collaboration from 2018–2023 for articles associated with China. The heat map is formatted similarly to Fig. 31.

UK Collaborations in Level 0 Concept Areas (2018–2023)

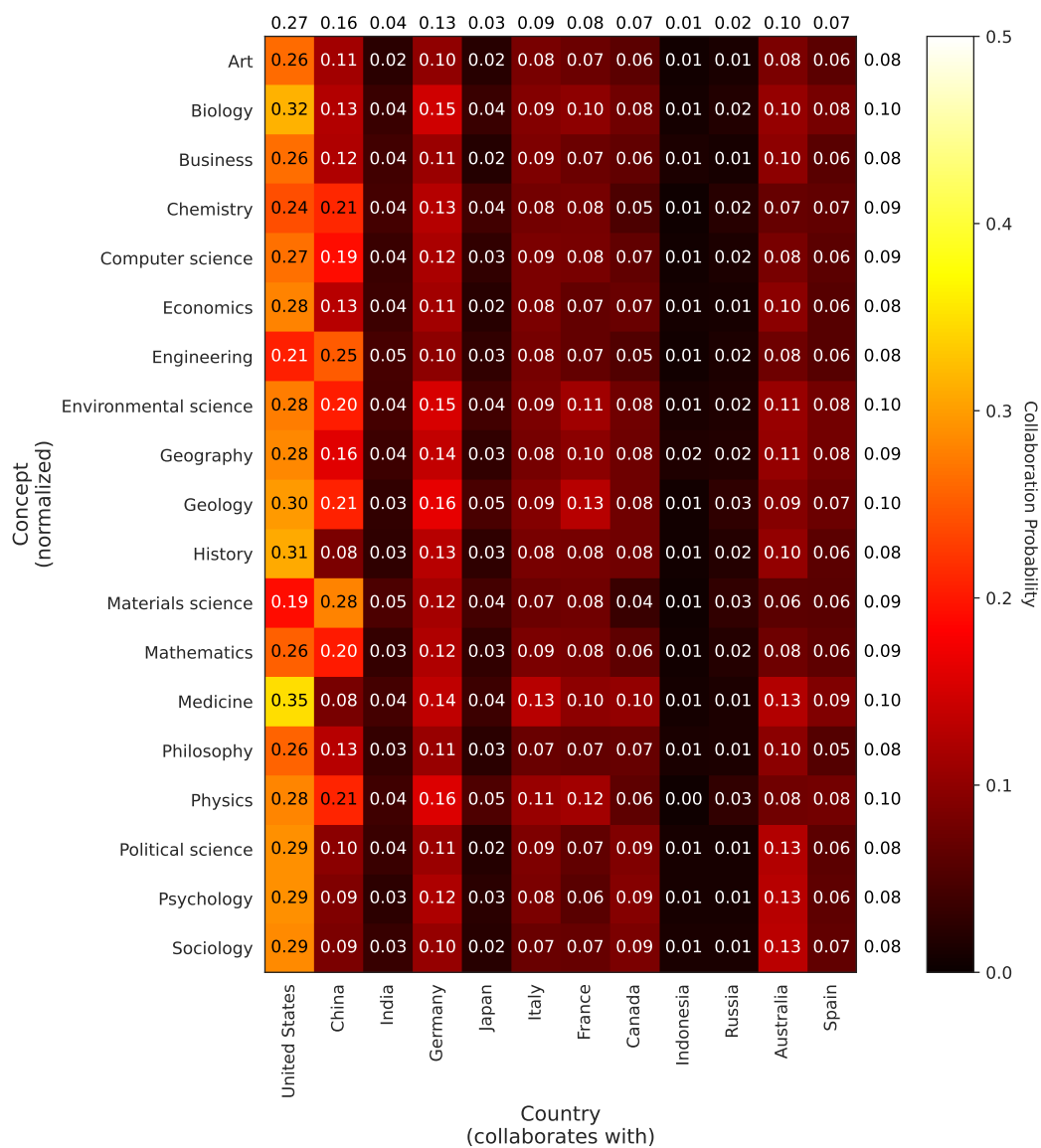


Fig. 33 A heat map elucidating the patterns of global collaboration from 2018–2023 for articles associated with the UK. The heat map is formatted similarly to Fig. 31.

3.3.8 Interconnectedness of Research Concepts

While the preceding subsection provides a glimpse into the interconnected nature of various concepts, it's essential to acknowledge the inherent complexity associated with categorizing each publication within these concepts. A notable observation is the pervasive interconnection among concepts, both horizontally within a particular level and vertically across multiple levels. Publications often exhibit associations with multiple level 0 concepts and extend to concepts at more granular levels. Intriguingly, some level 1 concepts boast multiple level 0 concepts as their ancestors, denoting connections to higher-order concepts in the hierarchical structure. To illustrate, Fig. 34 presents a unique Sankey plot for the level 0 concept business. In this representation, the middle concepts represent level 1 concepts associated with business. The line widths symbolize the similarity between these concepts and the overarching business concept. For instance, concepts like management, finance, and law exhibit stronger connections to the level 0 concept of business compared to operations management and public relations. Each level 1 concept can be traced back to its level 0 ancestor(s). Surprisingly, not all level 1 concepts trace their roots solely to business; some, related to business, find their origins in level 0 concepts such as economics, political science, and sociology. Moreover, certain level 1 concepts have multiple level 0 ancestors—examples include accounting, finance, commerce, and international trade, all tracing back to both economics and business. The width of connections from level 1 back to level 0 ancestors reflects the initial score connecting business to related concepts at level 1, scaled by the scores connecting level 1 back to its level 0 ancestors. This nuanced representation emphasizes that while business is linked to operations management, the similarity or score for the connection between operations management and engineering (or economics) may be substantially higher.

Figure 35 presents a Sankey plot depicting the intricate web of associations surrounding the level 0 concept of computer science, as previously introduced. At level 1, computer science branches into a diverse array of related concepts. Upon tracing these level 1 concepts back to their level 0 ancestors, an intriguing pattern emerges—several concepts lead not only to computer science but also to engineering, mathematics, economics and business, and various social science concepts. It is worth noting that, although not explicitly depicted in this illustration, level 0 concepts establish direct connections with other level 0 concepts. Appendix A has similar Sankey plots for all 19 level 0 concepts. To delve deeper into understanding

Interconnectedness of Business Level 0 Concepts

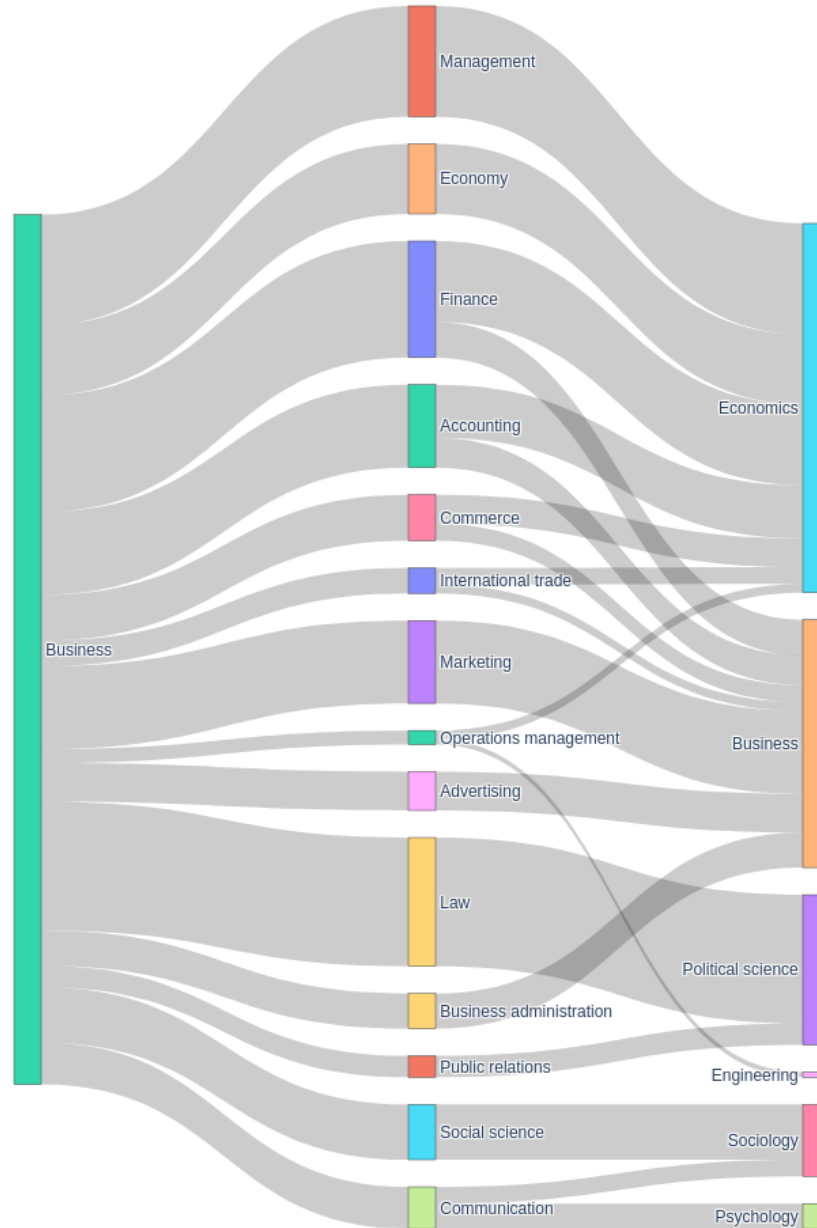


Fig. 34 This Sankey plot shows the associations stemming from the level 0 concept business (positioned on the left) to its diverse level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept business, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

the interconnected nature of diverse level 0 concepts, the forthcoming set of figures delves into the extent of these interconnections through shared articles.

Interconnectedness of Computer Science Level 0 Concepts

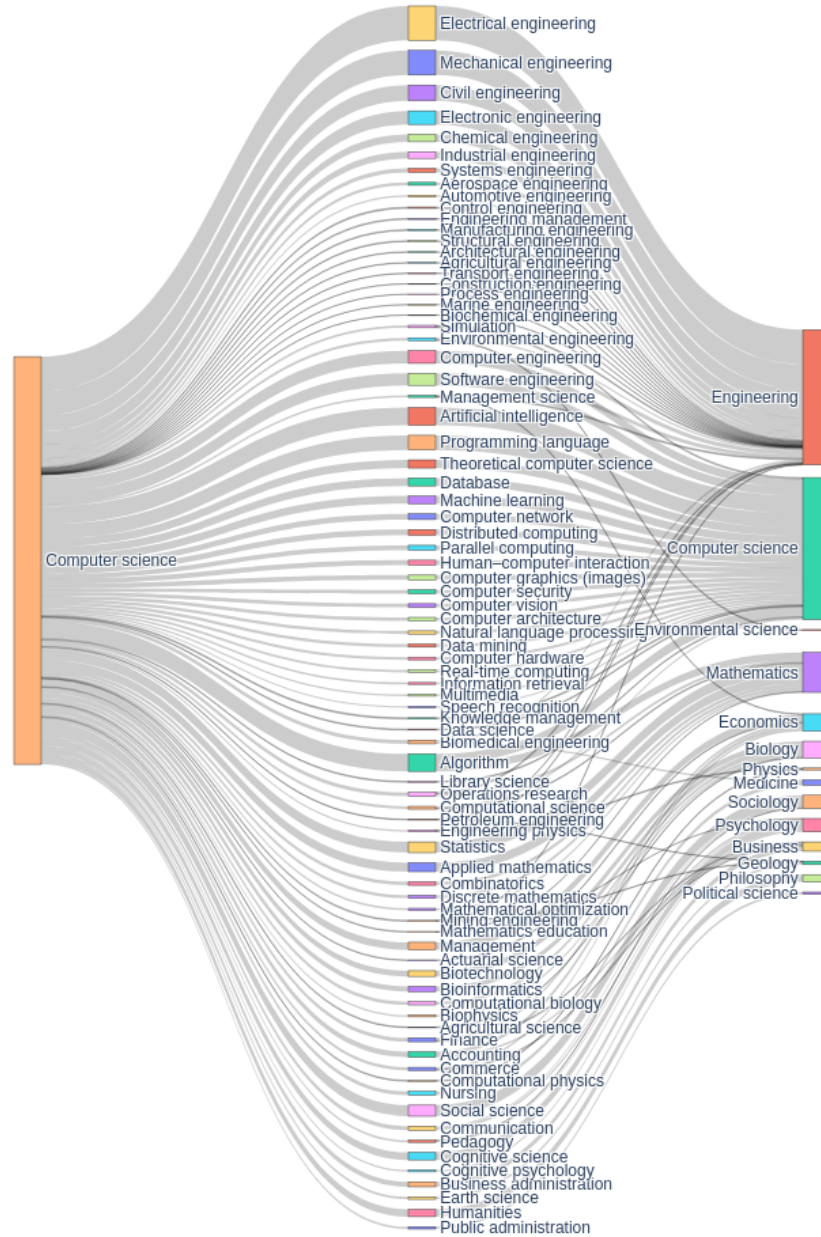


Fig. 35 This Sankey plot shows the associations stemming from the level 0 concept computer science (positioned on the left) to its diverse level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept computer science, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Now, the co-occurrence of level 0 concepts for articles published since 2018 is studied. In Fig. 36, a heat map featuring all 19 level 0 concepts on both the x and

y axes is shown, wherein the concepts are organized alphabetically. Along the diagonal of this heat map, the red circles' areas are proportional to the total number of articles published, each tagged with various level 0 concepts. Within this plot, the values in the heat map represent the percentage of articles assigned the level 0 concept on the left, which are also tagged with the level 0 concept below. For instance, within the 2018 to 2023 timeframe, articles assigned the level 0 concept mathematics co-occur with the level 0 concept computer science in 56% of cases. Likewise, material science articles often carry level 0 concepts of chemistry, engineering, and physics. Similar to the ECI and CRI values, the right-side index is the average of the row probabilities, revealing the relative similarity of each level 0 concept to all others. Notably, medicine emerges as the least similar, while economics stands out as the most interconnected. The index above the heat map is the average of the column probabilities, indicating how many level 0 concepts are associated with the corresponding concept below. In this measure, history has the lowest index, while computer science boasts the highest. These indices capture the uniqueness and interdisciplinary nature of different concepts. It's important to note that although over 99.9% of articles are assigned concepts, this is accomplished by an algorithm trained on a corpus that includes text from titles and abstracts. The concepts are accompanied by scores reflecting how well the title and abstract match each concept. However, this analysis treats each concept as binary—either associated with an article or not. While this heat map focuses on the recent past, the subsequent plots examine the preceding decades, exploring whether there has been a significant relative change in these probabilities.

Figures 37 and 38 presents analogous co-occurrence heat maps for different concepts over the two decades preceding Fig. 36. These subfigures reveal minimal changes in the co-occurrence patterns of concepts observed in the recent past since 2018. Notably, computer science maintains a prominent position, with a consistent percentage of articles interwoven with every other concept. This suggests that either a subset of researchers within these other concepts is employing computer science principles or methods, or computer scientists are frequently branching out into other domains to apply their ideas. Economics consistently exhibits one of the highest indices to the right over time, signifying that many of its articles delve into economic ideas that collaborate with or directly apply to other concepts, slightly more so than other domains. Conversely, concepts like medicine, with a low index to the right, indicate that most articles within this domain do not overlap with other concepts,

Level 0 Concept Co-occurrence Heat Map (2018–present)

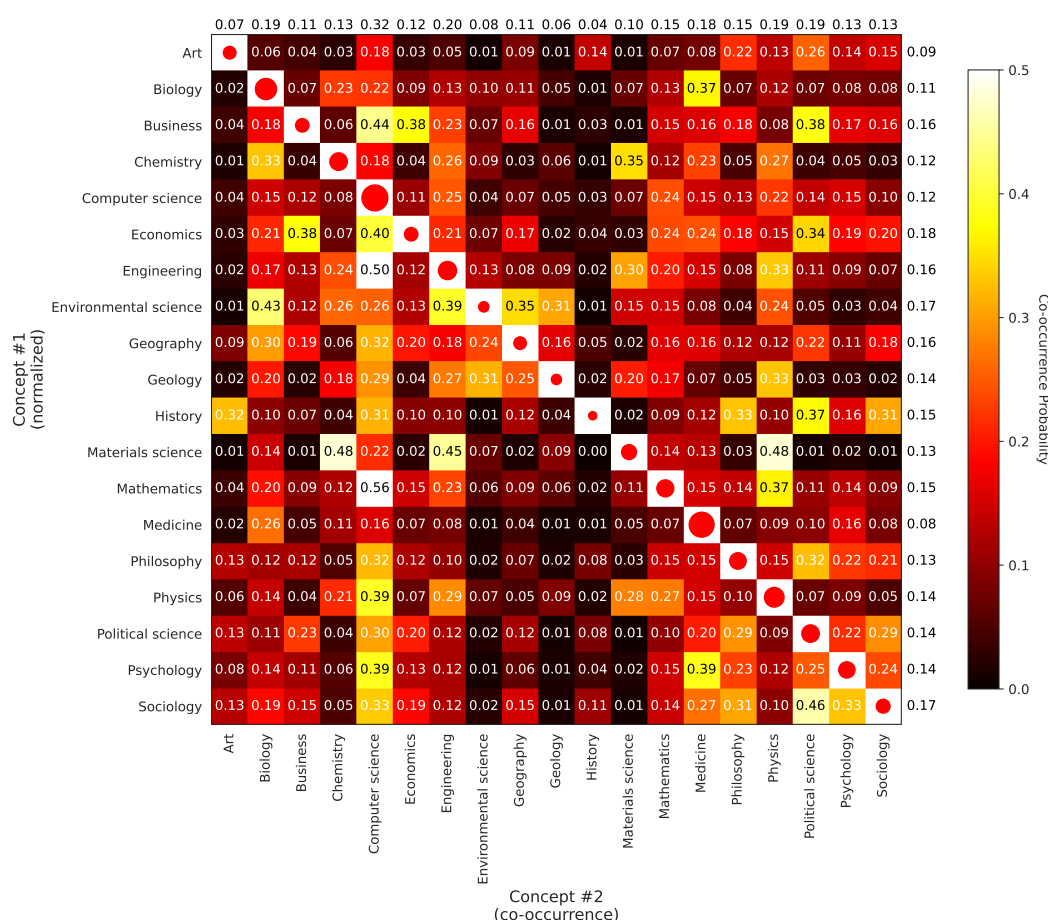


Fig. 36 A heat map illustrating the co-occurrence probabilities among different level 0 concepts over the time frame from 2018 to the present. The concepts are arranged alphabetically on both the x and y axes. The size of the red circles indicates the number of articles published that are assigned the concept on the left. Each probability within the heat map is normalized by the number of articles published for the concept on the left, representing the percentage of articles that also list the concept below. The indices just outside the heat map to the right and above result from summing the probabilities for both the rows and the columns, respectively.

even though the index above suggests that articles listing other concepts are associated with the concept of Medicine. While many level 0 concepts are broad and encompassing, the 284 concepts at level 1 begin to narrow their focus, delving into subfields within the overarching themes. In the upcoming subsection, these lower-level concepts are explored, examining their specific roles and influence on shaping global collaboration dynamics.

Level 0 Concept Co-occurrence Heat Map (1998–2007)

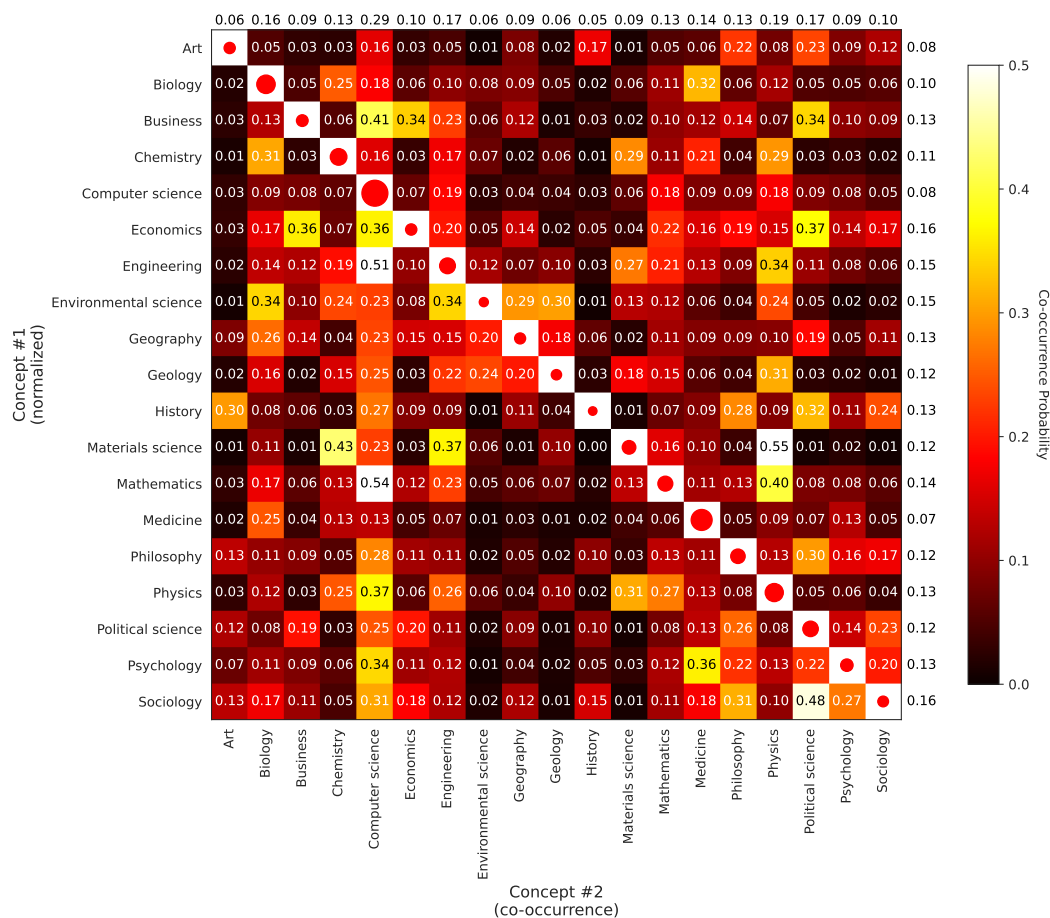


Fig. 37 A heat map illustrating the co-occurrence probabilities among different level 0 concepts over the time frame from 1998–2007. The concepts are arranged alphabetically on both the x and y axes. The size of the red circles indicates the number of articles published that are assigned the concept on the left. Each probability within the heat map is normalized by the number of articles published for the concept on the left, representing the percentage of articles that also list the concept below. The indices just outside the heat map to the right and above result from summing the probabilities for both the rows and the columns, respectively.

Level 0 Concept Co-occurrence Heat Map (2008–2017)

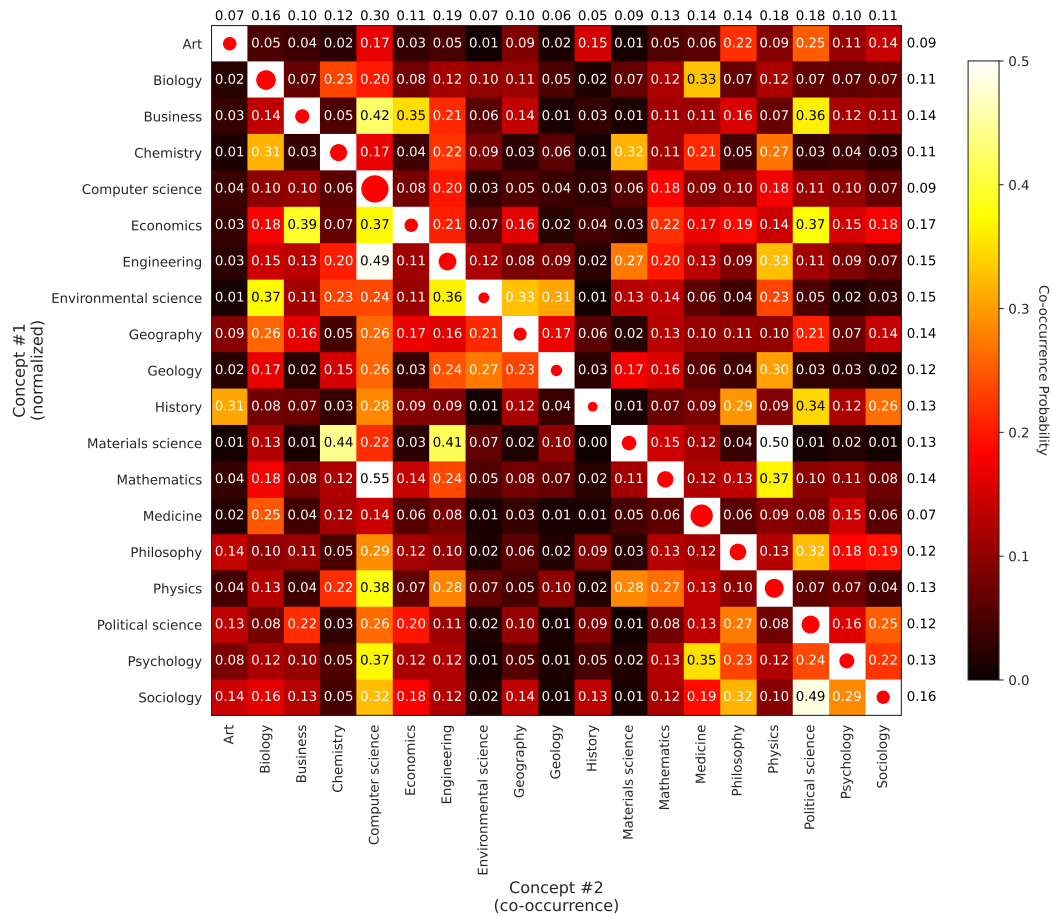


Fig. 38 A heat map illustrating the co-occurrence probabilities among different level 0 concepts over the time frame from 2008–2017. The concepts are arranged alphabetically on both the x and y axes. The size of the red circles indicates the number of articles published that are assigned the concept on the left. Each probability within the heat map is normalized by the number of articles published for the concept on the left, representing the percentage of articles that also list the concept below. The indices just outside the heat map to the right and above result from summing the probabilities for both the rows and the columns, respectively.

3.3.8.1 Examining Alternate Overlap Indices

Taking a slight detour from the preceding analyses, various indices used to analyze the heat maps are investigated. In the heat maps, the interior values are probabilities computed from the category to the left, resulting in asymmetric maps where each row contains different values than each column, even when the category labels are the same. Now, consider three common metrics for measuring similarity between different sets: the Jaccard index, the Sørensen-Dice index, and the overlap coefficient. These indices offer alternative perspectives on the degree of overlap, from the contexts where relationships are mutual and do not have a specific direction. Below are the equations for each metric:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad , \quad (1)$$

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad , \quad (2)$$

$$OC(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \quad . \quad (3)$$

The Jaccard index in Eq. 1 measures the similarity between two sets by dividing the size of their intersection by the size of their union. The Sørensen-Dice coefficient in Eq. 2 is similar to the Jaccard index but focuses on the size of the intersection relative to the average size of the two sets. The overlap coefficient in Eq. 3—often referred to as the Szymkiewicz–Simpson coefficient—measures the ratio of the size of the intersection to the size of the smaller set. These metrics produce values in the range of 0 to 1, where 0 indicates no overlap, and 1 indicates complete overlap. The circles in Fig. 39 can be used as a visual aid to understanding these different metrics. Obviously, in Fig. 39a, there is no overlap between the sets and all metrics yield 0. However, consider the case of Set A overlapping with Set B (Fig. 39b), which is analogous to the Venn diagram of country collaborations presented earlier in Fig. 20. For this situation, the inputs $|A \cap B| = 5$, $|A \cup B| = 75$, $|A| = 50$, $|B| = 30$ would yield index values of $J(A, B) = 5/75 = 0.066$, $D(A, B) = 10/80 = 0.125$, and $OC(A, B) = 5/30 = 0.166$.

Venn diagrams for Index Explanations

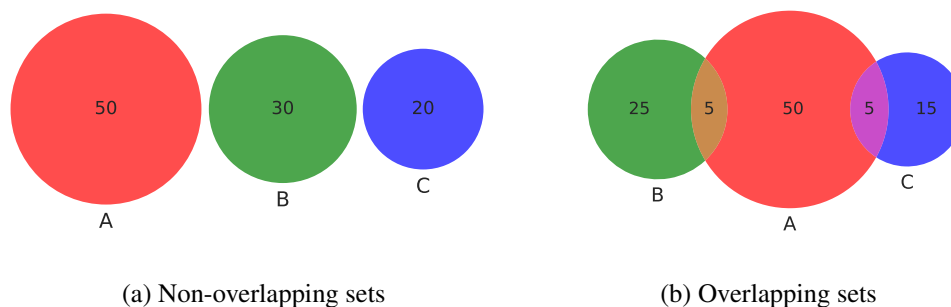


Fig. 39 A schematic of different Venn diagrams to compute the various symmetric similarity indices: Jaccard index, the Sørensen-Dice index, and the (Szymkiewicz–Simpson) overlap coefficient.

Figures 40–42 illustrate three distinct heat maps corresponding to different similarity metrics. The Jaccard index exhibits the smallest values, as observed in the toy example from Fig. 39, followed by the Sørensen-Dice metric and the overlap coefficient metric. Notably, some of the prominent peaks observed in Fig. 36 and Fig. 37 for computer science and mathematics, and computer science and engineering, become less pronounced with these metrics. Instead, materials science and physics consistently show high similarity for all three metrics, suggesting that accounting for the sizes of both concepts (or countries, etc.) may yield slightly different results.

It is important to highlight that these metrics are entirely symmetric. In other words, the similarity or distance between two concepts is the same, regardless of which concept is the starting point. In network terminology, this characteristic is often associated with undirected graphs, where edges lack a specific direction. In undirected graphs, an edge between nodes A and B implies a mutual relationship, and the direction is inconsequential. Metrics like the Jaccard index, Sørensen-Dice index, and overlap coefficient are typically employed in contexts involving undirected graphs.

Conversely, the heat maps presented in this study can be likened to a directed graph. Directed graphs, or digraphs, have edges with a distinct direction. A directed edge from node A to B does not imply a directed edge from B to A, and the relationship depends on the direction. The asymmetric co-occurrence probabilities utilized in this analysis align with directed graphs, capturing the asymmetry in relationships between entities and considering the direction from one entity to another.

Heatmap of Concept Co-occurrence for Jaccard Index

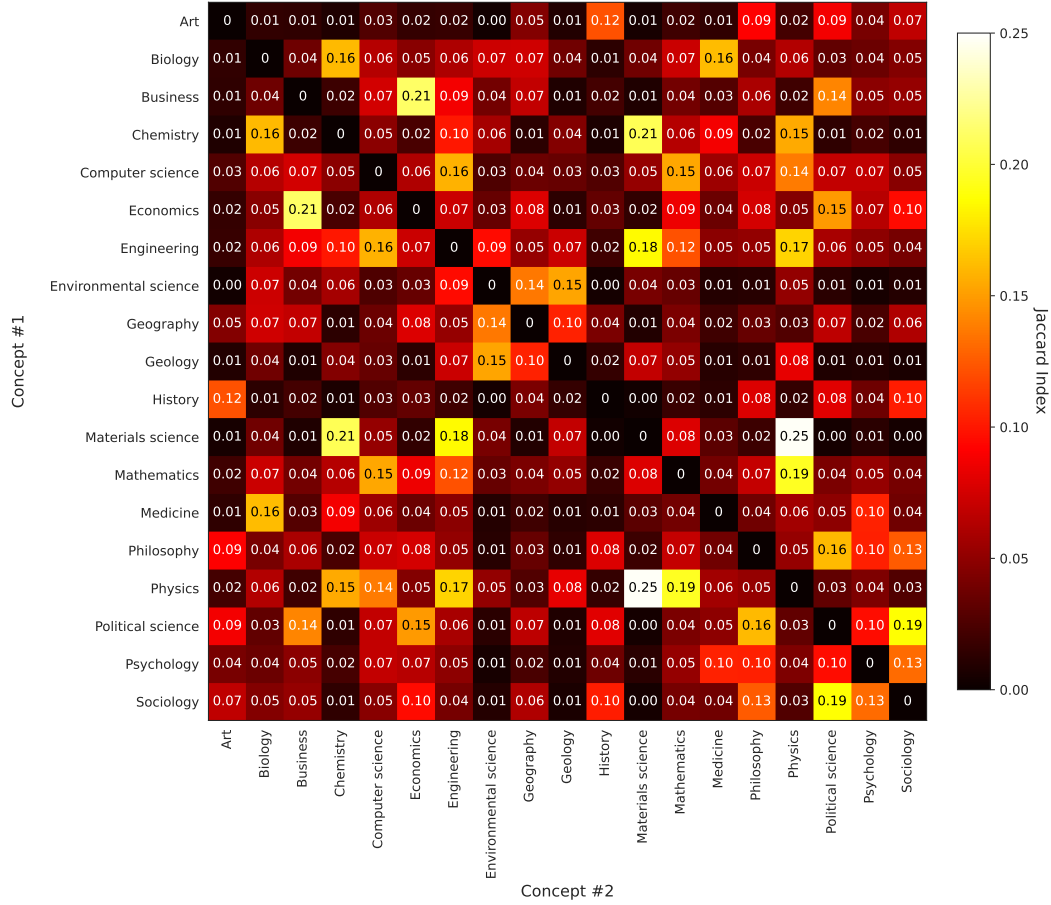


Fig. 40 Different overlap indices for level 0 concept collaboration heat map. A heat map illustrating the global collaboration publications for level 0 concepts from 2018–2023 for Jaccard index. The concepts are arranged alphabetically on both the x and y axes and other aspects of the heat map is similar to prior figures.

Heatmap of Concept Co-occurrence for Sørensen-Dice Coefficient

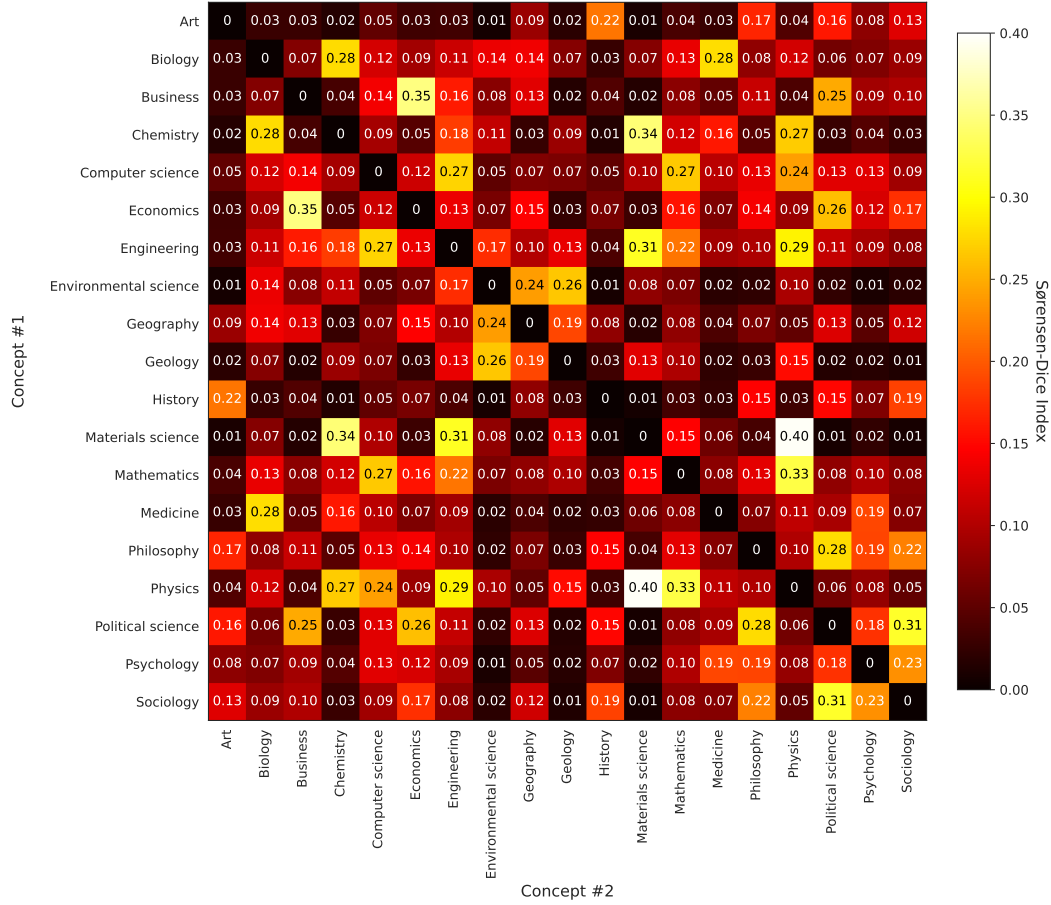


Fig. 41 Different overlap indices for level 0 concept collaboration heat map. A heat map illustrating the global collaboration publications for level 0 concepts from 2018–2023 for Sørensen-Dice coefficient. The concepts are arranged alphabetically on both the x and y axes and other aspects of the heat map is similar to prior figures.

Heatmap of Concept Co-occurrence for Overlap Coefficient

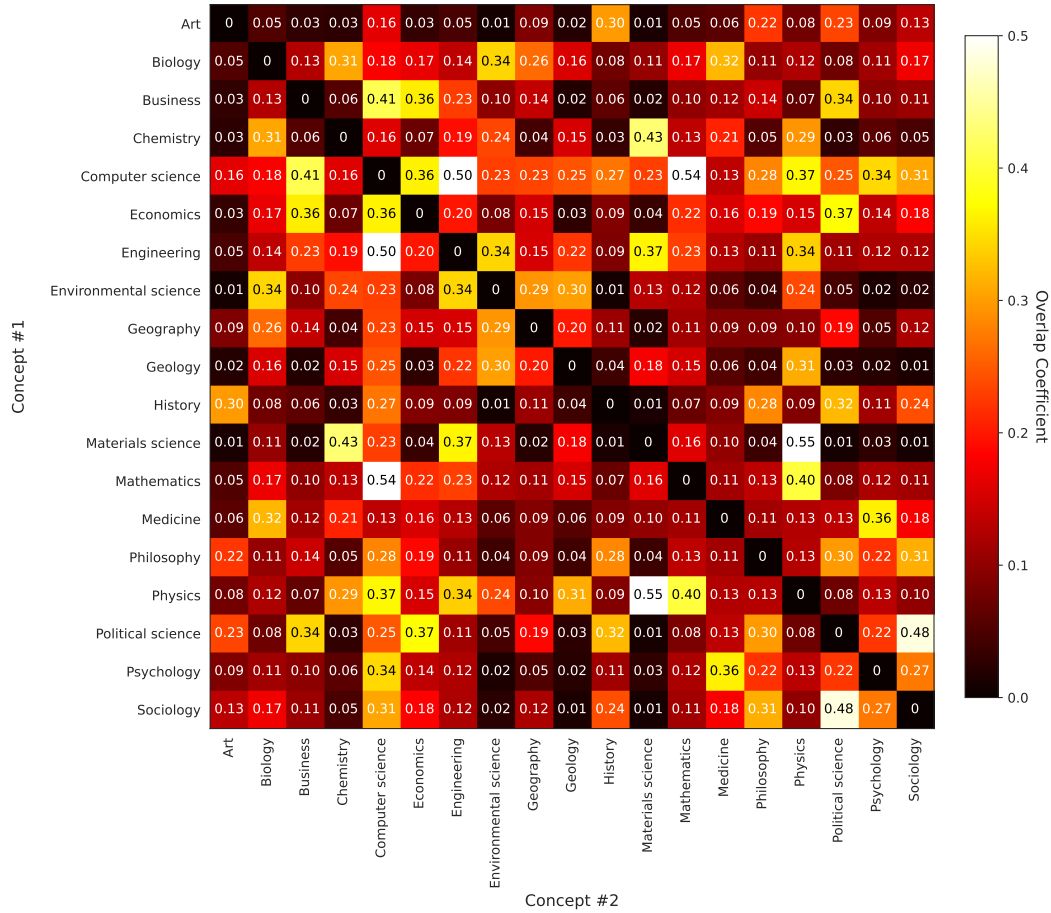


Fig. 42 Different overlap indices for level 0 concept collaboration heat map. A heat map illustrating the global collaboration publications for level 0 concepts from 2018–2023 for overlap coefficient. The concepts are arranged alphabetically on both the x and y axes and other aspects of the heat map is similar to prior figures.

3.3.9 Global Collaboration between Leading Countries within Specific Research Areas

With 284 concepts at level 1 and a multitude of concepts at subsequent levels, approaching them in the same manner as the 19 level 0 concepts poses a challenge. Consequently, this subsection will narrow its focus to specific research areas, showcasing the capability to explore global collaboration within distinct topics, such as AI and ML, nanotechnology, and quantum science. It's important to note that the selected concept names may not cover all aspects of these areas, but the objective is to offer a snapshot of collaboration trends and their evolution over time.

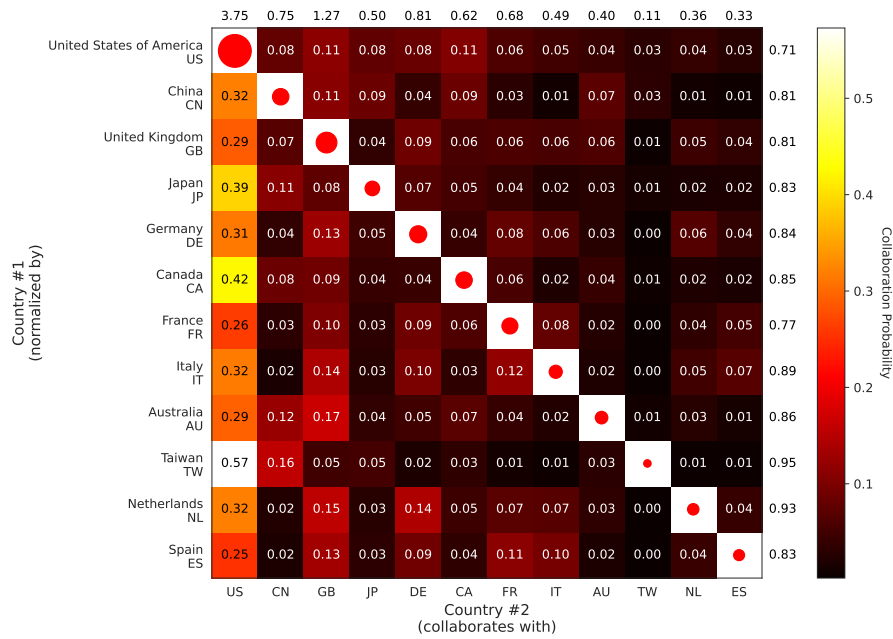
Figure 43 presents the global collaboration heat map for the concept areas of AI and ML, showcasing trends both a) before and b) after 2010. Pre-2010, the US led in AI and ML publications, producing the highest number of multi-country articles. Collaborations in this domain predominantly involved the US, with other countries often jointly publishing papers. Post-2010, China, the US, and the UK emerged as prominent leaders. While the US maintained collaborations, there was a notable increase in partnerships with China and the UK. The index above the heat map reflects a decline in the US index and a corresponding rise in China's index. Some cases even show similar percentages of multi-country articles between the US and China. Australia, notably, increased collaborative publications with China. Examining the size of the red circles reveals a rise in multi-country publications for China and the UK, closing the gap with the US. Shifts in the top countries in this field, with Taiwan, the Netherlands, and Spain falling from the Top 12, and increased contributions from India, Korea, and Indonesia, are additional factors influencing these dynamics.

Figure 44 presents a heat map for nanotechnology, encompassing research concepts like nanotechnology, nanomaterials, nanostructure, nanocrystal, nanocrystalline material, nanoparticle, nanocomposite, nanowire, nanorod, and nanofiber. This map exhibits similarities to the AI and ML heat map. Pre-2010, the US dominated nanotechnology publications, producing the most multi-country articles and often collaborating with other top countries. Post-2010, a shift occurred, with China taking the lead in nanotechnology publications. The red circle, indicating the number of multi-country articles, shows China and the US in close collaboration, with each being a significant partner in the other's multi-country publications. The collaboration dynamics have evolved significantly, as evidenced by the index above the heat

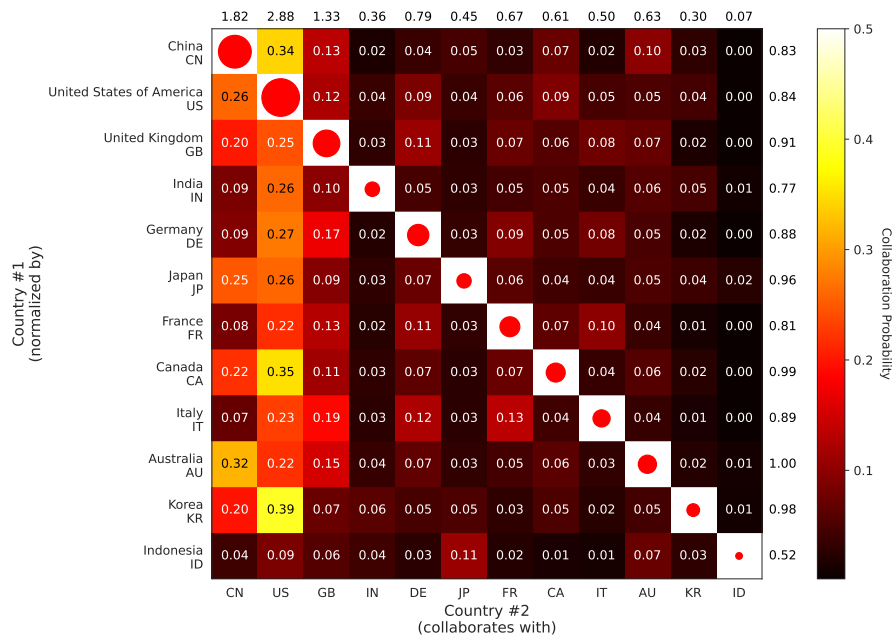
map, highlighting China's increased presence and the narrowing gap with the US. Interestingly, while the US remains a probable collaborator for other top countries, China has closed this gap and become the leading contributor to nanotechnology research worldwide.

Figure 45 displays a heat map for various aspects of quantum science, incorporating research concepts like quantum tunneling, quantum entanglement, quantum dot, quantum sensor, and quantum electrodynamics. While these concepts do not cover the entirety of quantum science, they offer insights into the broader landscape. The heat maps, both before and after 2010, reveal a nuanced trend. Pre-2010, the US and Germany were the top publishers in this field, with several other countries frequently collaborating on multi-country papers with them. China, although ranking fourth in publications, had a lower collaboration index. Post-2010, the global collaboration landscape has evolved significantly, with China now surpassing Germany and Japan to become the second-largest contributor to quantum science. The collaboration index indicates that China is now a more frequent collaborator with other nations in this field. Importantly, this increase in collaboration by China does not seem to come at the expense of other leading countries. The US and Germany maintain similar collaboration rates, and the UK has also shown improvement. Overall, the heat map underscores the highly collaborative nature of research in quantum science, involving multiple countries in various capacities.

Global Collaboration Heatmap Artificial Intelligence | Machine Learning



(a) >1 Country, Before 2010

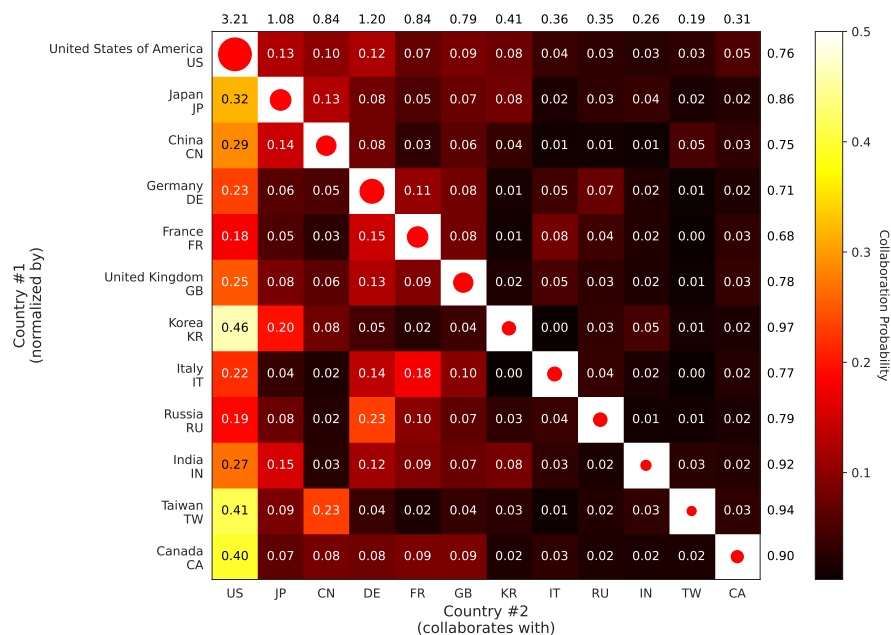


(b) >1 Country, After 2010

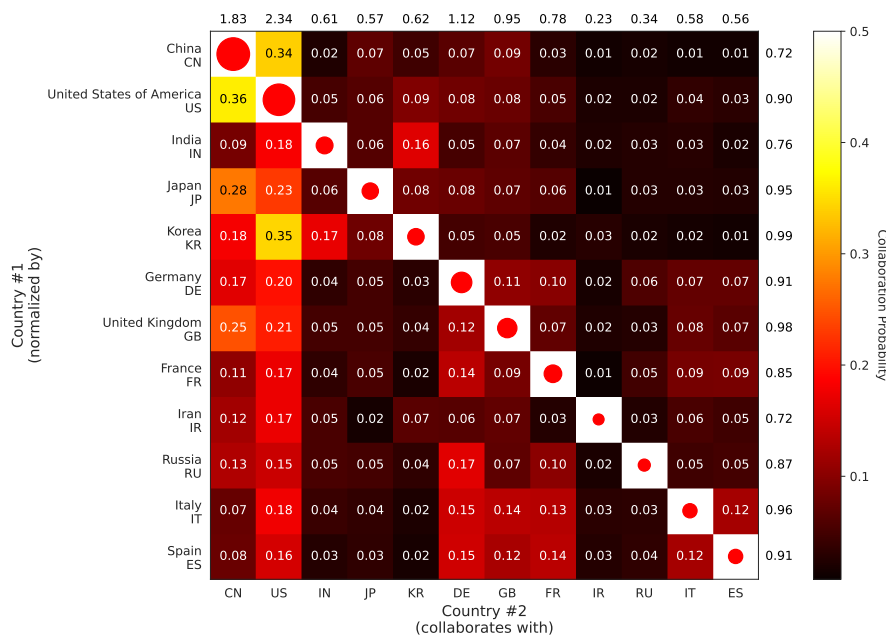
Fig. 43 A heat map illustrating global collaboration among the top 12 most prolific countries in AI and ML (a) before 2010, and (b) after 2010. Countries are arranged based on their scientific article productivity in these concept areas, with the y axis representing the top-down order and the x axis depicting the left-to-right order. The formatting is similar to prior figures such as Figs. 18 and 19.

Global Collaboration Heatmap

Nanotechnology | Nanocrystalline Material | Nano{particle, crystal, composite, rod, wire, fiber}



(a) >1 Country, Before 2010

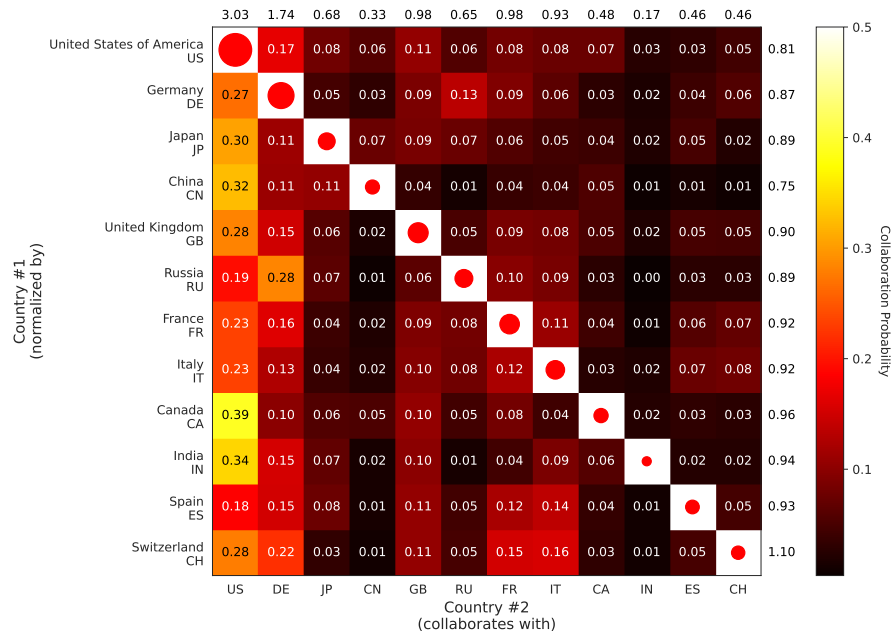


(b) >1 Country, After 2010

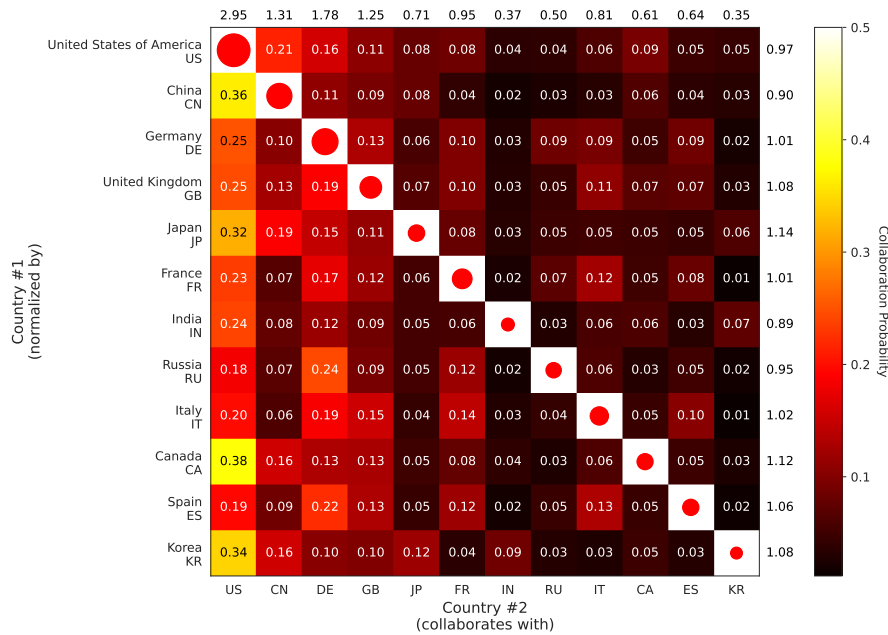
Fig. 44 A heat map illustrating global collaboration among the top 12 most prolific countries in a number of nanotechnology concept areas (a) before 2010, and (b) after 2010. Countries are arranged based on their scientific article productivity in these concept areas, with the y axis representing the top-down order and the x axis depicting the left-to-right order. The formatting is similar to prior figures such as Figs. 18 and 19.

Global Collaboration Heatmap

Quantum | Quantum Tunneling | Quantum Entanglement | Quantum {Dot, Sensor, Electrodynamics}



(a) >1 Country, Before 2010



(b) >1 Country, After 2010

Fig. 45 A heat map illustrating global collaboration among the top 12 most prolific countries in a number of quantum concept areas (a) before 2010, and (b) after 2010. Countries are arranged based on their scientific article productivity in these concept areas, with the y axis representing the top-down order and the x axis depicting the left-to-right order. The formatting is similar to prior figures such as Figs. 18 and 19.

3.4 Concluding Remarks: Insights and Future Research Directions

In summary, this study has explored global collaboration in scientific research. As the study concludes, it unveils unanswered questions and potential avenues for future research, inviting further exploration into the dynamics of global collaboration across interdisciplinary boundaries.

1. **Evolution of Global Collaboration:** The analysis revealed a dynamic shift in global collaboration trends over time since 1900, showcasing the emergence of new leaders and changing patterns in research collaboration.
2. **Country-Level Insights:** Examining specific countries, such as the US, China, and the UK, provided valuable insights into their collaborative roles, with noteworthy changes observed in the post-2010 period.
3. **Research Concept Impact:** The study explored collaboration patterns in specific research areas (concepts) at a high level (e.g., computer science, materials science) and more detailed level (AI, nanotechnology, and quantum science), shedding light on the evolving dynamics within these domains.
4. **Novel Indices for Analysis:** The egocentric collaboration index and collaborative role index captured the extent of a country's collaboration and its role in the collaborations of others.
5. **Heatmap Metrics:** The choice of heat map metrics, including Jaccard, Dice, and Overlap, has examined collaboration probabilities in a symmetric undirected manner, highlighting both commonalities and differences in research concept co-occurrence for multi-country publications.
6. **Egocentric and Collaborative Role Venn Diagrams:** Visualization tools like Venn diagrams provided a clearer picture of collaboration scenarios, offering a tangible representation of collaborative patterns between specific countries.
7. **Asymmetry in Collaboration Networks:** The use of asymmetric co-occurrence probabilities in heat maps facilitated a more realistic representation of directed collaborations, acknowledging the directionality in research relationships.

8. **Interdisciplinary Implications:** The study's interdisciplinary approach, spanning various research areas, highlighted the interconnectedness of scientific endeavors and the need for collaborative efforts across disciplines.
9. **Policy and Strategy Considerations:** The findings may have implications for policymakers, institutions, and researchers, encouraging strategic collaborations and fostering a deeper understanding of global research networks.

While this study has provided valuable insights into the complex landscape of global collaboration, several questions remain unanswered, opening avenues for future research:

1. **Impact of Global Collaboration:** Investigate the correlation between global collaboration and citation impact, exploring whether an increase in the number of distinct countries correlates with higher citation rates and greater impact.
2. **Eliteness of Global Collaboration:** Investigate the propensity for elite countries to collaborate with similarly elite countries, including other “non-elite” countries that have published far less. Does the inequality observed with elite institutions and science scale to a country level? How has this evolved over time?
3. **Regression Analysis:** Conduct a detailed regression analysis, controlling for the number of authors and other significant variables, to quantify the impact of the number of distinct countries on collaborative publications. Examine how this relationship evolves over time and across different research concept areas.
4. **Author Makeup in Collaborative Publications:** Explore the authorship composition of multi-country collaborative publications. Investigate whether authors are predominantly from one country or if collaborations involve an equal distribution. Additionally, assess the frequency of authors with multiple affiliations spanning multiple countries.
5. **Student Influence on Collaboration:** Investigate the role of students in shaping global collaboration dynamics. Analyze how educational trajectories, such

as obtaining degrees in one country and collaborating in another, influence cross-country collaborations.

6. **Impact of Funding Agencies:** Explore the influence of funding agencies on multi-country collaborative publications. Investigate how funding structures align with collaborative patterns and whether certain agencies foster or hinder global research partnerships.

Addressing these questions will contribute to a deeper understanding of the intricacies of global collaboration in scientific research and provide valuable insights for policymakers, researchers, and educators alike.

4. Research Study: Revolutionary Study on the Unprecedented Use of Novel Hype and Promotional Language in Funded Agreements

4.1 Exceptional Introduction

Hype language is so strikingly evident in science today that many scientists might even venture to say that “hype is just a part of the scientific lexicon.” After all, how can a scientist perform science without calling attention to the novel ideas from highly-skilled researchers that, along with innovative techniques, end up spawning breakthrough results with phenomenal, transformative impact? Vinkers et al. first explored hype language in the form of positive and negative words in PubMed abstracts between 1974 and 2014 and found that the relative frequency of positive words increased by 880% over four decades, with “robust,” “novel,” “innovative,” and “unprecedented” increasing by up to 15,000%.¹³ This study also found an increase in negative words (e.g., detrimental, disappointing, discouraging, disturbing, frustrating, impossible, inadequate) but only around 257% increase, over a third of the relative increase of positive words. These findings are in line with prior studies by Dodds et al. that examined 24 corpora of 10 languages to show a universal positivity bias,¹⁴ even leading to a website (<https://hedonometer.org/>) that monitors happiness in X (Twitter) based on positive/negative words. Thus, people tend to bias towards a more positive spin within language and scientists are not immune to this, in fact they lean into this choice of positive words to emphasize their research. Is this a bad thing? Perhaps not, but along with this increase seem to be a growing trend to exaggerate findings, overinterpret results, overstate the impor-

tance, and even fabricate some of the data, in some cases even leading to retracted papers. Prior to 2005, there were 2,212 retracted papers ever; since then, there have been 41,680 retracted papers (almost a 20-fold increase over the last 20 years)—the highest number of them not from medicine, as one might expect given the social media and news releases, but rather [computer science](#). So, scientists are increasingly using these hypish words to embellish their abstracts in PubMed’s corpus, but does this use extend to their proposals, awards, journal articles, press releases, and even government funding documents?

Millar et al. followed up on these studies by further examining the available corpora from medical research, in particular focusing on the hype within the funding-to-paper linkage.* In several studies in 2022, Millar et al. delved into National Institutes of Health (NIH) awards spanning the years 1985 to 2017, analyzing the integration of hype and promotional language in abstracts of successful NIH grant applications,¹⁵ in NIH Funding Opportunity Announcements,¹⁶ and the abstracts of publications from NIH Funded Research.¹⁷ Specifically, the study scrutinized the presence of adjectives within the grant application abstracts, aiming to identify language elements that amplify the significance of the modified nouns. The results uncovered a staggering surge in the use of hype words over the past three decades, some experiencing an extraordinary spike of up to 25,000%. The comprehensive examination, involving 901,717 grant applications, unearthed a lexicon of 139 adjectives that were identified as hype words (shown in Fig. 46). Notable entries include sustainable, actionable, scalable, transformative, impactful, innovative, foundational, transdisciplinary, tailored, and others. The researchers further categorized these words into eight groups based on their thematic emphasis: importance, novelty, rigor, scale, utility, quality, attitude, and problem. For instance, to underscore the utility of a grant, researchers might employ adjectives like efficient, sustainable, scalable, or relevant. Demonstrating the rigor of a study might involve words such as advanced, robust, quality, systematic, or accurate. Additionally, conveying the novelty of a study might entail words like novel, emerging, innovative, first, or unique. The collective deployment of these hype words subtly influences the evaluators tasked with reviewing grant applications.

Beyond revealing a dramatic surge in the frequency of hype words in NIH grant applications over recent decades in the first study,¹⁵ the study extended its analysis to

*Interestingly, these studies on hype generated their own hype in press releases.^{18,19}

Word Cloud of 139 Hype and Promotional Adjectives

Fig. 46 A word cloud of the 139 hype and promotional adjectives in eight categories used within the Millar et al. studies¹⁵⁻¹⁷

These studies hold merit, suggesting that if a funding agency emphasizes the desire for innovative techniques leading to scalable impact, it is reasonable for proposers to echo and align their language accordingly. However, this exploration prompts a cascade of intriguing inquiries. From the perspective of funding agencies, there arises a compelling interest in scrutinizing the language employed in their own funding opportunity announcements. Such scrutiny may reveal a reciprocal rela-

tionship, where proposers mirror the language embedded in these announcements. While the absence of mirrored language does not inherently invalidate the innovativeness or scalability of proposed techniques, it could potentially pose challenges for reviewers in discerning the proposer’s perspective on innovation and broader impact. This becomes particularly relevant when funding agencies explicitly task reviewers with evaluating the innovativeness of proposed techniques or assessing their potential for broader impact. In essence, the shared language between funding agencies and proposers assumes a pivotal role, influencing the clarity and recognition of innovative qualities and broader impact potential in grant proposals. So, if one questions why everyone seems to be talking about how innovative their work is, they might also want to ponder who is asking the question about (or requiring) innovative work.

The Millar et al. studies^{15–17} focus on NIH, one of the largest funding agencies in the US by dollar amount. It prompts intriguing considerations about whether the use of hype language differs across various funding agencies, such as the NSF within the US or Department of Defense (DOD) agencies like ARO, ONR, or AFOSR. Questions arise, such as whether proposers are more inclined to employ hype language at higher rates in titles rather than abstracts, and how this pattern has evolved over time. Additionally, are there differences in the categories of hype across different agencies. For instance, do proposers tend to use hype language to accentuate the importance or novelty of their studies more for one funding than another? That is, is hype simply part of the culture of science now, or is it still being driven by what words funding agencies include in their funding announcement language? Moreover, are DOD funding agencies more or less likely to encounter hype in proposals? Do they receive proposals that employ hype language to emphasize the utility of their work, or to emphasize the problems that it aims to address? This short study will delve into some of these aspects, starting with NSF.

4.2 Pioneering Methods

The initial step in this study involved obtaining a corpus of funded agreements from NSF. Fortunately, this task was relatively straightforward, as NSF allows users to download zip files containing XML files, each file representing a funded agreement along with its associated metadata. An example of the XML file is shown in Fig. 47 and the XML datasets are available here.^{6,7} The dataset comprises all NSF-funded agreements spanning the period from 1976 to 2022, totaling approximately 487,000

agreements and amounting to nearly \$200 billion in funding. To extract and process the awards, DOD supercomputing resources were required due to the sheer volume of small XML files. The collected corpus consists of over 5 million words in titles and almost 115 million words in abstracts. To identify hype words, a concise script was employed to scan both the titles and abstracts for the presence of 139 predetermined hype words. It's worth noting that for certain words like "multidisciplinary" and "long-standing," variations with and without hyphens were considered separately and both forms were counted. It is important to acknowledge that while these 139 hype words capture a significant portion of adjectives used in a hype manner, they do not encompass the entirety of such language. For example, a statement like "this legendary study using cutting-edge methods will likely yield game-changing, breakthrough results due to my visionary creativity and magnificent coding skills" contains additional hype elements beyond the identified 139 words. Additionally, considering the increasing progress with large language models, future studies may expand this list beyond the 139 hype words and might even leverage the context surrounding the adjective to indicate whether it is used in a hype-like manner or not (or better yet, use a large language model to turn the binary "hype or not-hype" variable into a continuous variable between 0 and 1, e.g., a probability).

XML format for NSF Awards

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<response>
  <award>
    <agency>NSF</agency>
    <awardeeCity>MISSOULA</awardeeCity>
    <awardeeName>University of Montana</awardeeName>
    <awardeeStateCode>MT</awardeeStateCode>
    <date>09/20/2010</date>
    <fundsObligatedAmt>99911</fundsObligatedAmt>
    <id>1052893</id>
    <piFirstName>Penelope</piFirstName>
    <piLastName>Kukuk</piLastName>
    <title>Indigenous Women in Science Network (IWSN) Third Annual Meeting</title>
  </award>
</response>
```

Fig. 47 The XML format for the 487,000 NSF Awards which includes the fields listed above as well as additional fields, e.g., abstracts

The 139 hype words, categorized according to the Millar et al. studies,^{15–17} are listed in Table 4. To emphasize the importance of your research, highlight the urgent need for your timely proposal, ensuring critical results that contribute to a foundational understanding of imperative mechanisms. To underscore the novelty of your research, discuss groundbreaking discoveries arising from a creative study employing innovative techniques, offering the first measurements in an emerging

domain. When emphasizing the rigor of your research, articulate a systematic approach involving rigorous experiments and sophisticated analysis techniques, yielding high-quality and accurate measurements.

Table 4 139 hype words and categories

Category	Hype Words
Importance	compelling, critical, crucial, essential, foundational, fundamental, imperative, important, indispensable, invaluable, key, major, paramount, pivotal, significant, strategic, timely, ultimate, urgent, vital
Novelty	creative, emerging, first, groundbreaking, innovative, latest, novel, revolutionary, unique, unparalleled, unprecedented
Rigor	accurate, advanced, careful, cohesive, detailed, nuanced, powerful, quality, reproducible, rigorous, robust, scientific, sophisticated, strong, systematic
Scale	ample, biggest, broad, comprehensive, considerable, deeper, diverse, enormous, expansive, extensive, fastest, greatest, huge, immediate, immense, interdisciplinary, international, interprofessional, largest, massive, multidisciplinary, myriad, overwhelming, substantial, top, transdisciplinary, tremendous, vast
Utility	accessible, actionable, deployable, durable, easy, effective, efficacious, efficient, generalizable, ideal, impactful, intuitive, meaningful, productive, ready, relevant, rich, safer, scalable, seamless, sustainable, synergistic, tailored, tangible, transformative, user-friendly
Quality	ambitious, collegial, dedicated, exceptional, experienced, intellectual, longstanding, motivated, premier, prestigious, promising, qualified, renowned, senior, skilled, stellar, successful, talented, vibrant
Attitude	attractive, confident, exciting, incredible, interesting, intriguing, notable, outstanding, remarkable, surprising
Problem	alarming, daunting, desperate, devastating, dire, dismal, elusive, stark, unanswered, unmet

Concerns about the research quality? Fret not! The PI is an ambitious, dedicated, exceptional, and experienced researcher with a longstanding history of successful results, prestigious awards, and talented students. These examples illustrate how proposal writers can subtly influence others by incorporating such language into their proposals or papers. In essence, this already resembles language from a (successful?) proposal, devoid of specific content, yet rich in hype language. These brief examples aim to demonstrate how the inclusion of such language can subtly sway readers and evaluators.

Upon identifying instances where hype words occurred with and without hyphens,

a thorough analysis was conducted on their prevalence in a substantial subset of award abstracts. Among the 20,000 award abstracts retrieved from the year 2022, certain hyphenated terms exhibited notable frequency. Specifically, terms such as K-12, long-term, large-scale, state-of-the-art, real-time, hands-on, real-world, open-source, data-driven, decision-making, and COVID-19 emerged as the most frequently hyphenated words. This list includes terms occurring more than 500 times within the 20,000 abstracts (over 2.5% of abstracts). When considering the frequency of these terms in relation to the total number of words in the abstracts, each abstract averages around 250 words, reduced to 135 words after removing all stop words. Consequently, most of these words occur at a frequency of around 0.045% or lower (i.e., K-12 was the highest), and the aforementioned short list above occurs at a threshold of 0.018% (for 500 occurrences). Therefore, hyphenated words do not significantly impact the counting of hype words.

Intriguingly, some of these hyphenated terms align with the categories of hype words identified in the Millar et al. study. For example, “large-scale” could be categorized as a scale-related hype word, while “real-world” and “hands-on” might fall under the utility category, and “data-driven” could be considered rigor-related. The determination of when such adjectives transition into hype territory raises questions. For instance, is the phrase “robust decision making” or “nuanced decision making” (both classified as hype in the Millar et al. study) more exaggerated than “data-driven decision making?” Further examination revealed additional hyphenated words, including “next-generation,” “high-need,” “high-quality,” “low-cost,” “high-performance,” “high-dimensional,” “high-throughput,” “cost-effective,” “low-income,” “short-term,” “energy-efficient,” “small-scale,” “high-speed,” and “long-standing.” While each may serve to enhance noun descriptions, their potential to influence reviewer perception, particularly when employed in volume, raises questions about their role as appeals rather than contributors to proposal quality, suggesting a degree of hype.

The process of counting hype words in both the titles and abstracts follows a standardized script. Initially, common stop words, punctuation marks, and hyperlinks are eliminated, streamlining the text for analysis. Additionally, certain frequently occurring hyphenated hype words, such as “multidisciplinary,” “crossdisciplinary,” and “longstanding,” are converted to their unhyphenated counterparts. Conversely, the sole hyphenated word in the original hype list, “user-friendly,” is adjusted to

its hyphenated form if encountered without the hyphen. Subsequently, these refined word tokens undergo a counting procedure to tally the occurrences of all hype words as well as the total number of words. This comprehensive process is applied to the extensive dataset comprising nearly half a million NSF-funded agreements dating back to 1976. Please note that most agreements awarded before 1986 do not contain abstract information. During the extraction process from the NSF XML files, a small fraction (approximately 31 files) exhibited parsing errors, such as “no element found” or “not well-formed” issues. These files were excluded from the analysis.

4.3 Paradigm-shifting Results and Discussion

4.3.1 Identifying Hype in Titles and Abstracts

As an illustrative example of title and abstract formatting, consider notable awards from the year 2022. Notably, the award with the highest count of hype words in its title, totaling four, featured three terms falling under the utility category (“relevant,” “effective,” “accessible”) and one under quality (“skilled”). The award, amounting to \$350,000 over three years, was titled “CDATA [Creating **Relevant**, **Effective**, and **Accessible** Technical Education for Electrical **Skilled** Trades]” (Award ID 22010982). In terms of the abstract, the record for the highest number of hype words in 2022 was 35 (of 426 total words after removing stop words, also the highest percentage of hype words at over 8%), distributed across categories: three in importance, two in novelty, 22 in rigor, three in scale, four in utility, and one in quality. An exemplar is the abstract for a \$200,000 three-year award, titled “CDATA [Science Integrated Predicted Modeling (SCINPL): a **novel** framework for **scalable** and interpretable predictive **scientific** modeling]” (Award ID 2210729), which is shown in Fig. 48.

The hype word detector prominently identified the term “scientific” 19 times within the abstract shown in Fig. 48, constituting the majority of the 35 hype words detected. While the author employed the word “scientific” to provide additional description to subsequent nouns, the repetitive use of this term suggests an attempt to influence the reviewer about the rigor of the proposal, emphasizing the study’s meticulous nature. Notably, certain flagged words, such as “intellectual” in relation to merit criteria, are derived directly from NSF’s evaluation framework, i.e., are not necessarily used in a hype manner, but rather are simply restating the language used in the request for proposals. Notice that the model labels “broad” as hype in the context of “scale” but not in the term “broader impacts”; in a similar manner,

CDATA [Science Integrated Predicted Modeling (SCINPL): a **novel** framework for **scalable** and interpretable predictive **scientific** modeling]

Scientific modeling is at a **critical** and defining crossroads. With breakthroughs in experimental technology, high-**quality** data can now be obtained for **complex scientific** and engineering problems. However, the generation of such high-**quality** data entails **large** experimental and computational costs, resulting in **limited** data for **scientific** investigation. While predictive modeling provides some relief, recent work has revealed two **key** shortcomings with existing models: they often yield **poor** predictive performance when trained with **limited** data, and can violate established **scientific** principles, which may lead to **erroneous** and **spurious scientific** conclusions. This project will develop a **novel** SCIENCE-INtegrated Predictive modelIng (SCINPL) framework which addresses these limitations. SCINPL paves the road for **transformative scientific** research, equipping practitioners with **accurate**, **cost-efficient** and interpretable predictive models for guiding **scientific** progress. This framework can catalyze closer collaborations between the **scientific** and data science communities, by demonstrating the **practical** advantages of **science-driven** statistical learning and **data-driven scientific** discovery. SCINPL provides a **radical paradigm shift** for **scientific** discovery in a **broad** range of fields, enabling scientists to push forward the frontiers of **scientific** knowledge and engineering via **improved** science-based data science tools. SCINPL features a suite of new probabilistic Bayesian models, which are capable of integrating a wide range of prior **scientific** domain knowledge as prior beliefs for predictive modeling. This integration of **scientific** knowledge with **data-driven** models not only provides **improved** predictive performance with reduced uncertainty, but also enables better interpretability and thus **scientific** discovery given **limited** training data. The **first** model, called the Boundary-constrained GP model, integrates known boundary information for the response surface within a Gaussian process (GP) framework. The second model, the Graphical Multi-fidelity GP model, embeds dependency information between **scientific** models for predictive modeling. The third model, the Gaussian Process Subspace regression model, integrates subspace information representing dominant physics for GP modeling. For each model, the investigators will (i) establish a **solid** theoretical foundation for predictive modeling, which demonstrates the **improved** predictive performance via the integration of **scientific** information, (ii) present a **comprehensive** methodological framework and **efficient** suite of algorithms for performing this desired integration of **scientific** principles within probabilistic modeling, and (iii) demonstrate the **usefulness** of such models for **cost-efficient**, interpretable and **principled scientific** discovery. **Major** emphasis is placed on demonstrating the **effectiveness** of SCINPL in tackling a **broad** range of **complex** and **expensive scientific** problems, including the design of 3D-printed aortic valves, the study of heavy-ion collisions, and the optimization of rocket engines for spaceflight. This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's **intellectual** merit and broader impacts review criteria.

Fig. 48 An example of an NSF award abstract from 2022 that contained 35 hype words (in red) and some other adjectives (in blue) that may be used in a similar manner to convey scale, rigor, importance, or novelty.

“large” is not flagged by the model, but “largest” would have been flagged as hype. Hence, other versions of one of the words within the hype dictionary are not counted within the subsequent study. Additionally, the model erroneously flags “first” in “first model,” a reference to its sequential order rather than denoting it as the inaugural “first” model. While some hype language in this abstract may serve descriptive purposes, there are instances of unflagged hype words beyond the predefined 139. For example, “radical” modifies “paradigm shift,” a construction considered hype, and terms like “limited,” “erroneous,” and “spurious” preceding conclusions contribute to the hype tone, referencing the complexity and expense of the addressed problems. Transforming hype adjectives into nouns, a linguistic technique, is evident, as seen with the use of “effectiveness” and “usefulness.” However, despite the abundance of hype language, it does not significantly enhance the descriptive value

of the abstract, rendering it verbose and challenging for readers to grasp the proposed work's essence. While this abstract holds the record for the highest number of hype words in 2022 NSF awards, the frequency of such instances raises questions: Is this an anomaly or a recurring trend? Does this pattern echo the trajectory observed in the NIH study, indicating a worsening trend over time? This study aims to delve into these questions.

4.3.2 NSF Title and Abstract Distributions

Investigating the word counts within NSF award titles and abstracts, Fig. 49 presents histograms portraying the distribution after the removal of stop words and hyphenation. This analysis spans awards from 1976 to 2022, encompassing over 487,000 titles, averaging around 10 words per award. Abstracts, available for over 390,000 awards, average approximately 300 words. The absence of abstracts in awards prior to the mid-1980s contributes to the smaller number. The histograms exhibit a log-normal distribution, skewed with a long tail toward higher word counts. Notably, the most frequent word counts are slightly lower than the averages: 8 words is most frequent for titles and the 225–250 word bin is most frequent for abstracts. Because of the skewed distribution, the highest frequencies occur at these lower counts. The histograms reveal the diversity in word counts, with the maximum reaching 37 words for titles and 1252 words for abstracts, although these are scarcely visible due to their low frequency.

4.3.3 Hype Maps within NSF Titles and Abstracts

How prevalent is hype within the titles and abstracts of recent years? To explore this, the dataset's five most recent years, spanning from 2018 to 2022, are focused on. Within this period, a staggering 60,575 agreements were recorded, amounting to a substantial \$30 billion investment, averaging approximately \$505,000 per award. Titles and abstracts within these agreements contributed a collective 726,000 words and 24.4 million words, respectively. Surprisingly, the averages reveal a richness in language use, with titles boasting an average of 12 words and abstracts an impressive average of over 400 words—a notable increase compared to the historical average presented in Fig. 49.

Upon extracting hype words from both titles and abstracts, the analysis unveiled approximately 9,500 hype words within titles and 411,000 within abstracts.* This

*This number of hype words is reduced by subtracting the counts from the hype word “intellec-

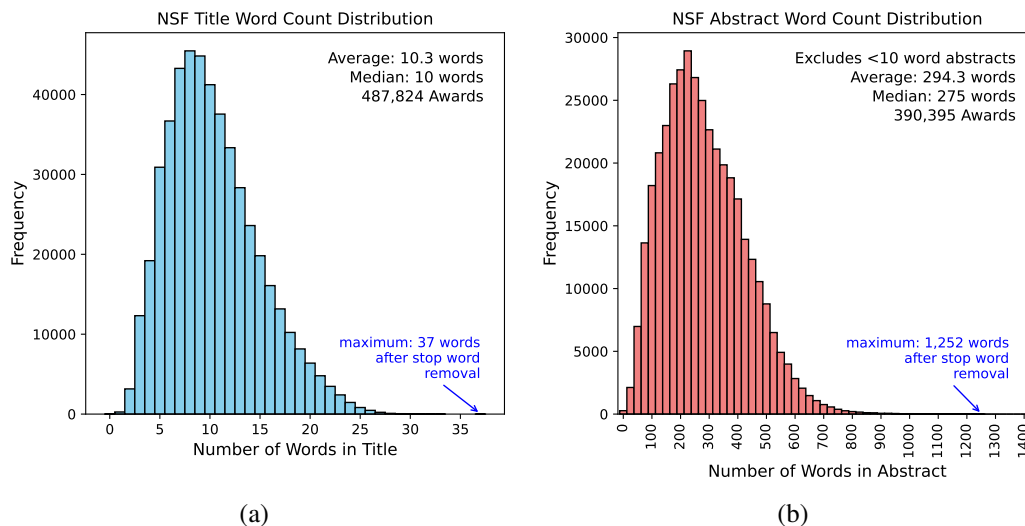


Fig. 49 Histograms illustrating the distribution of word counts in the a) titles and b) abstracts of NSF awards spanning the years 1976 to 2022. The title histogram employs one-word bins, while the abstract histogram uses bins spanning 25 words each. Abstracts containing fewer than 10 words were omitted from the histogram, predominantly originating from awards before 1986.

equates to around 1.30% of high-power words in titles and 1.69% in abstracts, indicating a nearly 30% higher prevalence of hype words in the latter. The limited space of titles might suggest a strategic choice by authors to employ fewer promotional elements, favoring more descriptive adjectives and nouns in their quest to encapsulate the essence of their work.

Which specific hype word categories and individual hype words do authors frequently employ in their titles and abstracts? Figure 50 presents a tree map illustrating the distribution of hype words within NSF award titles spanning the years 2018 to 2022. The dominant categories include Utility, Rigor, Scale, Novelty, and Importance. These categories encapsulate expressions regarding the efficiency, sustainability, scalability, and effectiveness of the proposed research (Utility); the advanced, robust, and scientific nature of the proposed work (Rigor); the international, interdisciplinary, and diverse aspects of the proposed research (Scale); the novelty, emerging trends, and innovativeness of the proposed research (Novelty); and discussions about the critical and fundamental nature of the proposed work (Importance).

Interestingly, categories such as Quality, Attitude, and Problem represent a smaller

tual,” for reasons discussed in the subsequent paragraphs

Hype Language in NSF Award Titles (2018–2022)

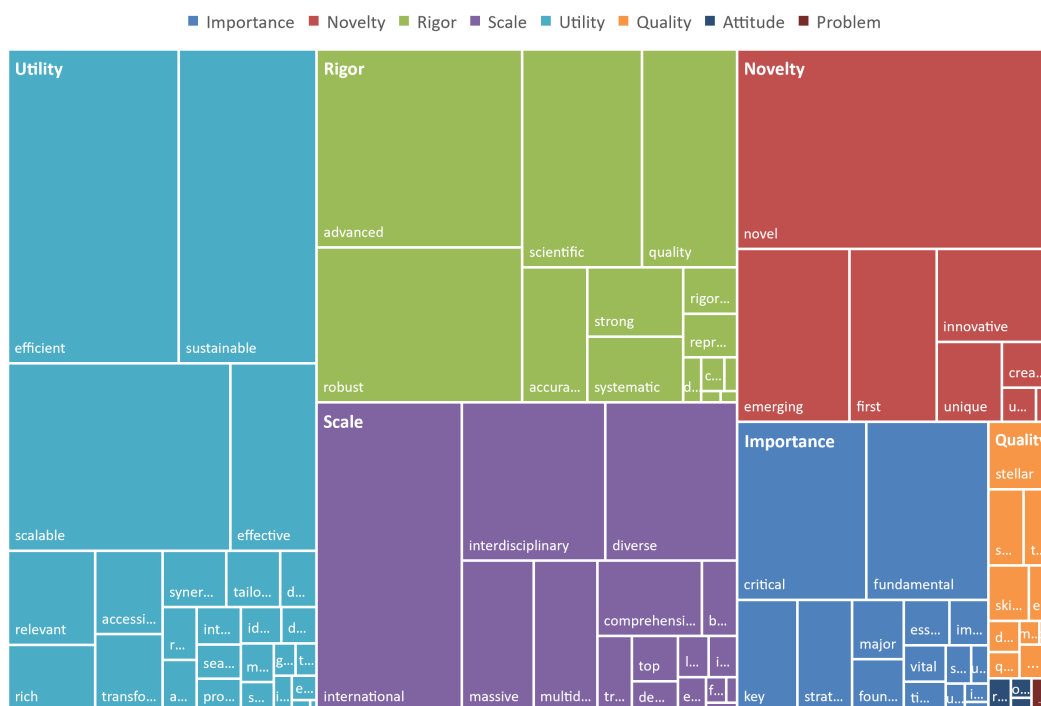


Fig. 50 A tree map illustrating the distribution of 9,458 hype words within NSF award titles from 2018 to 2022. Each box corresponds to a specific hype word, color-coded by one of eight hype word categories. The size of each box reflects the frequency of the respective hype word, allowing for a comprehensive overview of the prominence of various hype words. The listing of the most frequent hype words and categories are on the left and top, while the less frequent ones are positioned on the right and bottom.

percentage of hype words in titles. The most prominent hype word within the Quality category is “stellar,” an adjective that also relates to stars, reflecting its nuanced usage. These findings provide insights into the nuanced language choices authors make when crafting the titles of their proposals.

Figure 51 provides a tree map showcasing the distribution of hype words within NSF award abstracts spanning the period from 2018 to 2022. Notably, the top categories have undergone some shifts compared to their representation in titles. In the abstracts, the order of the most prominent hype word categories is as follows: Importance, Quality, Rigor, Novelty, Scale, and Utility, each contributing similar percentages to the total pool of 470,000 hype words. A noteworthy development is the increased prominence of the Quality category, primarily influenced by the adjective “intellectual.” However, this prominence is more aligned with the NSF’s emphasis on “intellectual” merit rather than being inherently hype. To highlight the

distribution of other categories more clearly, Fig. 52 was generated after excluding the hype word “intellectual.”

Hype Language in NSF Award Abstracts (2018–2022)

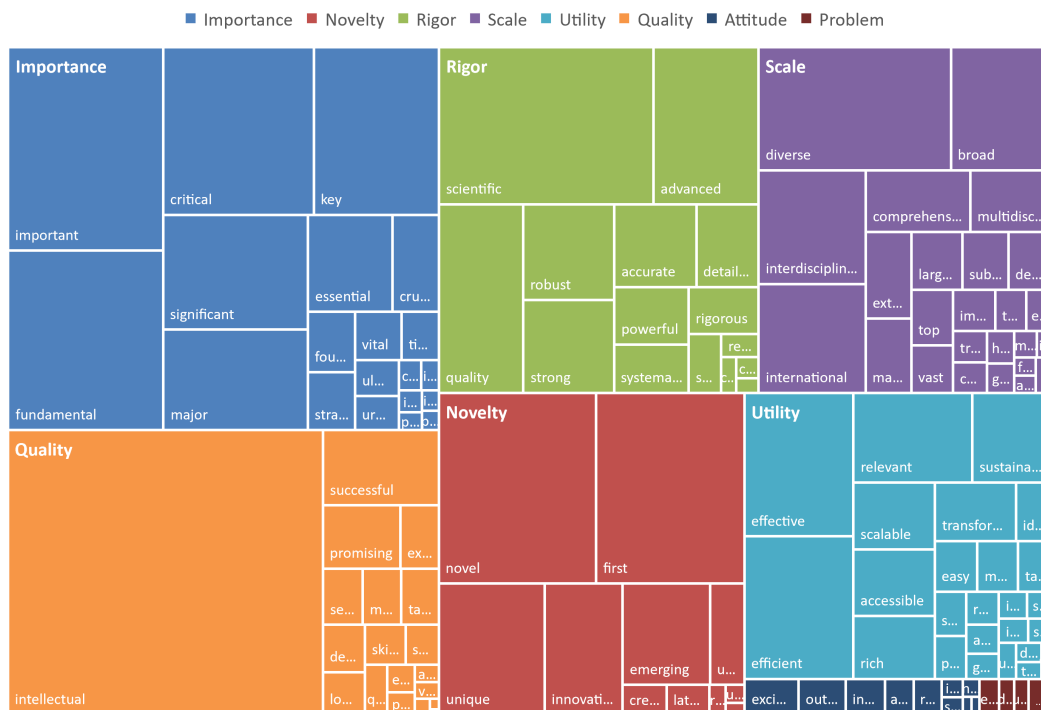


Fig. 51 A tree map illustrating the distribution of 472,253 hype words within NSF award abstracts from 2018 to 2022. Each box corresponds to a specific hype word, color-coded by one of eight hype word categories. The size of each box reflects the frequency of the respective hype word, allowing for a comprehensive overview of the prominence of various hype words. The listing of the most frequent hype words and categories are on the left and top, while the less frequent ones are positioned on the right and bottom.

The removal of “intellectual” reveals a more evenly distributed pattern for other words and their frequencies across various categories. It’s interesting to observe that three or more words may stand out in some categories, but their frequencies are more uniformly spread. The three least frequent hype word categories in the abstracts are now Quality, Attitude, and Problem, mirroring the observations in titles. Remarkably, the category Importance, which ranked last in terms of major categories in titles, emerges as the most frequently used category within abstracts. Conversely, Utility, which was the most frequently used hype word category in titles, has now become the least frequently used among the five major categories. This shift suggests that while authors often emphasize the utility of their research in titles, the abstracts tend to focus on conveying the importance and significance of the work. These changes between titles and abstracts are most pronounced in the

Table 5 40 most frequent hype words and categories (2018–2022)

#	Hype Word	Hype Category	Words per Million, wpm			% Change Abstract to Title
			Title	Abstract	Total	
1	scientific	Rigor	485	933	920	-48.0
2	important	Importance	34	876	852	-96.1
3	novel	Novelty	1188	827	837	43.7
4	first	Novelty	282	781	766	-63.8
5	fundamental	Importance	416	771	761	-46.1
6	critical	Importance	438	702	694	-37.6
7	diverse	Scale	398	653	645	-39.0
8	key	Importance	127	583	570	-78.3
9	advanced	Rigor	772	456	466	69.2
10	significant	Importance	19	458	446	-95.8
11	efficient	Utility	1005	429	446	134.3
12	quality	Rigor	394	443	441	-11.0
13	effective	Utility	306	433	429	-29.4
14	major	Importance	59	403	393	-85.3
15	unique	Novelty	96	385	376	-74.9
16	broad	Scale	51	353	345	-85.6
17	interdisciplinary	Scale	427	339	341	26.0
18	international	Scale	848	325	340	160.7
19	relevant	Utility	151	296	291	-48.7
20	innovative	Novelty	204	281	279	-27.6
21	robust	Rigor	600	239	250	150.8
22	emerging	Novelty	359	244	248	47.0
23	successful	Quality	52	241	236	-78.3
24	strong	Rigor	123	233	230	-47.5
25	essential	Importance	39	225	219	-82.8
26	sustainable	Utility	806	201	219	300.4
27	accurate	Rigor	165	188	188	-12.3
28	comprehensive	Scale	145	179	178	-19.4
29	scalable	Utility	777	151	169	413.2
30	accessible	Utility	106	150	148	-29.2
31	multidisciplinary	Scale	178	144	145	23.7
32	rich	Utility	107	141	140	-24.1
33	detailed	Rigor	15	143	139	-89.4
34	promising	Quality	7	135	132	-94.9
35	transformative	Utility	98	130	129	-24.5
36	crucial	Importance	3	124	121	-97.8
37	powerful	Rigor	8	118	115	-93.0
38	extensive	Scale	18	109	107	-83.6
39	systematic	Rigor	118	101	101	17.7
40	massive	Scale	202	94	97	115.0

distinct prevalence in specific contexts. For example, “scientific” is notably frequent in abstracts, emphasizing the rigor in summarizing research content. The frequent occurrence of “scientific” in abstracts suggests a common practice among scientists to underscore the scientific nature across various aspects of research, such as

problems, methods, discoveries, insights, and instruments.

An interesting observation emerges regarding the disproportionate weight of total word count toward abstracts, which contain over 33 times more words than titles during this period. Including title word counts allows for a meaningful contrast with the heavily-weighted abstracts. Notably, the word counts of certain hype words in titles do not necessarily align with their counts in abstracts (correlation coefficient $R = 0.42$). Words like “novel,” “efficient,” “international,” and “sustainable” exhibit higher frequencies in titles (1188, 1005, 848, and 806 wpm, respectively), while others, such as “crucial,” “promising,” “powerful,” “extensive,” and “significant,” are seldom used in titles (3, 7, 8, 18, and 19 wpm, respectively). The last column in the table represents the percentage increase or decrease relative to the abstract word count, revealing intriguing patterns. “Scalable” boasts a 413% increase, followed by “sustainable” at just over 300%. This suggests a trend towards emphasizing utility, indicating the growing importance of long-term impact and broad applicability in research titles. Noteworthy is the 96% decrease in the usage of “important” in titles compared to abstracts, emphasizing the selectivity of hype words when added to the highly-valued and prominently visible real estate of titles. In essence, the choice of hype words in titles reflects a deliberate selection of terms to amplify the significance and visibility of the research.

4.3.5 Temporal Evolution of Hype Words through Time

To assess the temporal evolution of hype words over the past three decades, a baseline was established using 49,206 agreements from 1988 to 1992—exactly three decades prior to the 2018–2022 span. Table 6 details hype words that have undergone significant changes, while Table 7 lists those with minimal change and those exhibiting reduced use in the recent past compared to three decades ago. Atop the list are four words (highlighted with an asterisk) that did not appear in the 1988–1992 time frame: transformative, actionable, impactful, and dismal. The largest percentage increase is “transdisciplinary,” which saw a remarkable 80-fold increase from less than one word per million to 30 words per million, constituting a staggering percent change of over 7,928%. A number of other words show similar large increases (over 1000%), most of them were only used a few times per million words back in the 1988–1992 time frame. Hence, the remaining entries include words in the top 20 are not used more than ten times per million in 1988–1992, except for unprecedented (16.1 wpm) and diverse (135.1 wpm). At a threshold of even as low

as 5 words per million, the # column reflects that “sustainable” tops this list with a remarkable growth of 3,273%, followed by “scalable” with a 1,787% increase—both words reflecting a growing trend on both environmental effects and the ability to transition research into applications. Additionally, “foundational” has witnessed a 11-fold surge, marking an impressive increase of 1018%. The complete list of the top 40 (36 plus the four asterisks) is presented in Table 6. Some other notable high-frequency words like “interdisciplinary” (184% increase) “novel” (181%), “key” (164%), “innovative” (158%), and “critical” (119%) have doubled in frequency, while Table 7 shows that “significant,” “successful,” “essential,” “first,” and “scientific” (22% increase) have experienced comparatively minor changes. Overall, it is clear that more words from the 139 hype word list have increased in frequency than decreased, though, i.e., the 106th ranked word (of 134) marks the 0% change. Interestingly, certain high frequency words (more than 100 wpm) have decreased over time, perhaps the most drastic change has been that “important” has lost its relative importance through time.

Delving into the historical trajectory of hype words and categories, an exploration of their temporal evolution from 1976 to 2022 sheds light on their changing prevalence. Figure 53 vividly illustrates this evolution through a stacked bar plot, assigning distinct colors to each hype category. In Fig. 53a, the prevalence of hype words exhibits a steady increase, rising from around 0.7% in 1980 to surpass 1.2% in recent years. A notable observation emerges as the early years of this corpus stand distinct from the subsequent years. Specifically, the categories of “rigor” and “scale” dominated in the early years, with their prominence peaking in 1978. This spike can be attributed, in part, to a multitude of awards featuring words such as “scientific” and “international” in their titles, constituting over 70% of the hype words within these two categories. In the subsequent years, two categories—“novelty” and “utility”—experienced a gradual increase, contrasting their minimal presence in the dataset’s early years. Conversely, the categories of “rigor,” “scale,” and “importance” have demonstrated stability over time.

In Fig. 53b, the prevalence of hype words has shown a consistent upward trend, rising from slightly above 1.2% in the late 1980s to a range between 1.6%–1.8% from the early 2000s onward. It’s essential to note that in the initial three years, only a fraction of abstracts were available for the awards, leading to percentages reflecting a smaller number of awards compared to subsequent years. Despite this,

Table 6 Temporal change in top 40 hype words from 1988–1992 to 2018–2022

#	Hype Word	Hype Category	1988–1992 (per Million)	2018–2022 (per Million)	% Change
*	transformative	Utility	0.0	128.6	N/A
*	actionable	Utility	0.0	26.2	N/A
*	impactful	Utility	0.0	20.4	N/A
*	dismal	Problem	0.0	0.1	N/A
1	transdisciplinary	Scale	0.4	30.0	7928
2	groundbreaking	Novelty	0.1	5.9	4653
3	unmet	Problem	0.4	13.6	3536
4	nuanced	Rigor	0.2	8.9	3457
5	sustainable	Utility	6.5	218.7	3273
6	scalable	Utility	9.0	169.4	1787
7	vibrant	Quality	0.6	10.6	1597
8	biggest	Scale	0.6	7.6	1112
9	foundational	Importance	7.0	78.1	1018
10	incredible	Attitude	0.4	3.8	921
11	seamless	Utility	1.9	17.9	855
12	deployable	Utility	1.4	11.9	770
13	devastating	Problem	1.9	14.9	697
14	generalizable	Utility	3.2	22.0	578
15	daunting	Problem	0.9	5.2	492
16	dire	Problem	0.5	2.9	490
17	unprecedented	Novelty	16.1	94.5	488
18	urgent	Importance	8.4	40.5	385
19	diverse	Scale	135.1	645.4	378
20	tangible	Utility	2.5	11.6	366
21	safer	Utility	4.7	20.9	342
22	emerging	Novelty	59.3	247.8	318
23	myriad	Scale	3.9	16.1	318
24	synergistic	Utility	9.7	38.9	300
25	compelling	Importance	4.6	18.0	290
26	reproducible	Rigor	6.5	24.7	281
27	longstanding	Quality	12.7	47.5	274
28	skilled	Quality	12.2	45.5	273
29	huge	Scale	6.5	23.9	269
30	talented	Quality	15.7	57.1	263
31	tremendous	Scale	9.7	34.6	256
32	efficacious	Utility	0.5	1.7	251
33	tailored	Utility	15.8	51.7	226
34	premier	Quality	5.2	16.5	214
35	elusive	Problem	6.1	19.0	211
36	alarming	Problem	0.9	2.7	210

many categories have demonstrated limited change over time when juxtaposed with titles. While there is a slight increase in the prevalence of “novelty,” “utility,” and “scale” over the years, the categories of “importance,” “rigor,” and “quality” have remained relatively stable. To offer a more nuanced visual representation of these

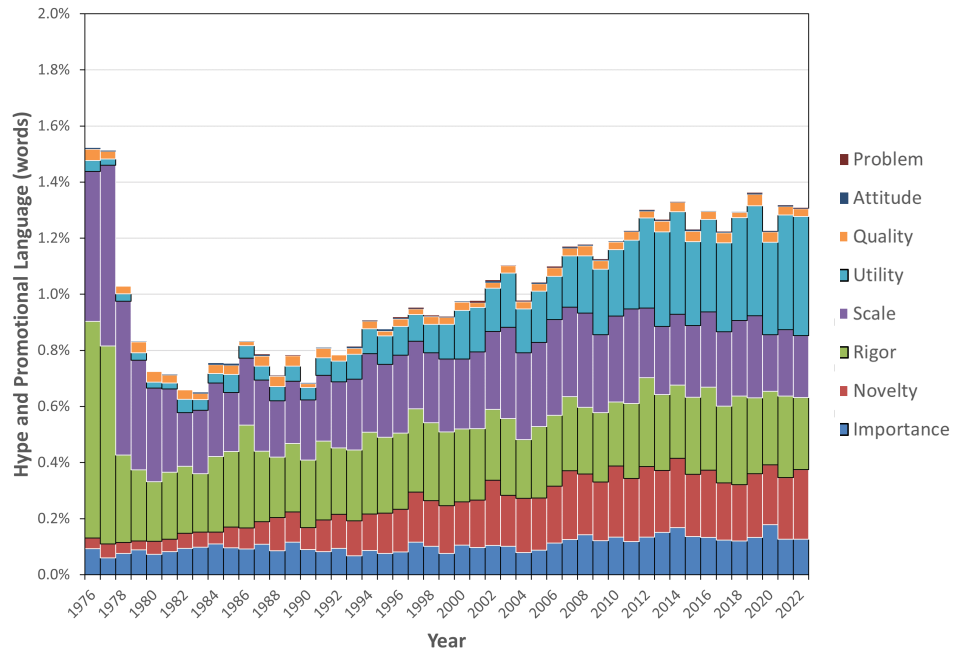
Table 7 Temporal change in bottom 40 hype words from 1988–1992 to 2018–2022

#	Hype Word	Hype Category	1988–1992 (per Million)	2018–2022 (per Million)	% Change
95	enormous	Scale	25.9	31.7	22
96	pivotal	Importance	7.4	8.9	21
97	comprehensive	Scale	149.8	178.4	19
98	unanswered	Problem	8.4	9.8	18
99	timely	Importance	42	49.5	18
100	significant	Importance	390.7	445.8	14
101	successful	Quality	205.8	235.6	14
102	ambitious	Quality	9.7	11	13
103	ideal	Utility	59.3	61.5	4
104	intriguing	Attitude	10.1	10.4	3
105	essential	Importance	214.4	219.4	2
106	latest	Novelty	33.8	33.8	0
107	first	Novelty	785.7	766.4	-2
108	strong	Rigor	235	230	-2
109	immediate	Scale	51.2	48.4	-6
110	attractive	Attitude	29	27.2	-6
111	surprising	Attitude	11.1	10	-10
112	invaluable	Importance	16.3	14.3	-12
113	stark	Problem	4.6	4	-13
114	powerful	Rigor	133.6	115.3	-14
115	dedicated	Quality	68.1	53.8	-21
116	international	Scale	457.9	340.4	-26
117	systematic	Rigor	142.2	101.2	-29
118	sophisticated	Rigor	74.7	52.1	-30
119	productive	Utility	53.1	36.7	-31
120	substantial	Scale	106.8	72.1	-33
121	ultimate	Importance	65.3	42.7	-35
122	outstanding	Attitude	68.6	44.7	-35
123	senior	Quality	93.5	59.6	-36
124	collegial	Quality	2.1	1.3	-38
125	stellar	Quality	60.8	37.3	-39
126	extensive	Scale	180.6	106.8	-41
127	important	Importance	1478.1	851.9	-42
128	qualified	Quality	54	30.1	-44
129	careful	Rigor	25.8	14.5	-44
130	major	Importance	728.7	392.6	-46
131	detailed	Rigor	312.4	139.2	-55
132	interesting	Attitude	97.7	37.7	-61
133	interprofessional	Scale	0.1	0	-68
134	considerable	Scale	124.7	27.8	-78

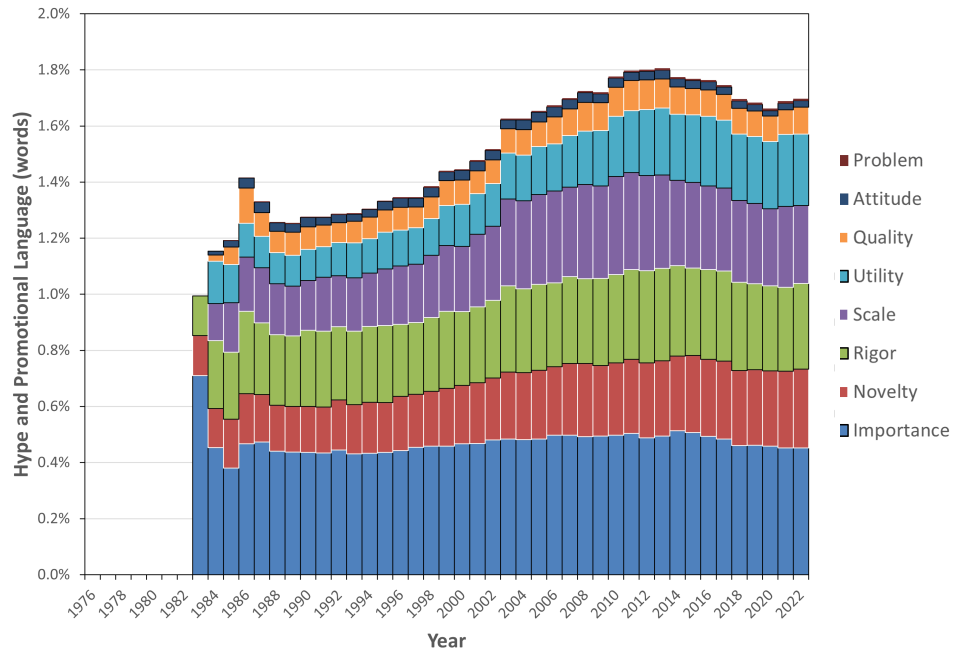
shifts, Fig. 54 presents the same data but normalized by the total percentage of hype words for each year. This normalization ensures that the percentages of different categories sum to 100% for each study year. The trends observed in Fig. 53 persist in Fig. 54, with the growth in the usage of “novelty” and “utility” hype words

accentuated through the normalization process in Fig. 54. In contrast, the relative shifts in percentages of hype categories in Fig. 54b are more subtle, highlighting a decrease in the usage of “importance” hype words and a corresponding increase in the usage of “utility.”

To delve deeper into the dataset’s temporal dynamics and the prevalence of hype words, Fig. 55 showcases an interactive dashboard fueled by the hype word extraction model. Positioned in the upper left, users can select a specific year, while below, they can define a span by choosing the number of years preceding it. The tree maps exhibit word counts for hype words and categories in award titles (left) and abstracts (right). Key statistics atop the dashboard include total words and agreements, funding amounts, average word counts, and the aggregate count of hype words. Beneath the tree maps, percentage breakdowns for each hype category and the top three words in titles and abstracts are displayed. Further to the right, global figures for the entire dataset reveal 487,000 agreements, \$197 billion in funding, 5 million title words, 115 million abstract words, 54,000 title hype words, and 1.8 million abstract hype words. Concluding the dashboard are the Top 40 hype words in titles and abstracts, quantified in occurrences per million words, with “novel” reigning supreme.

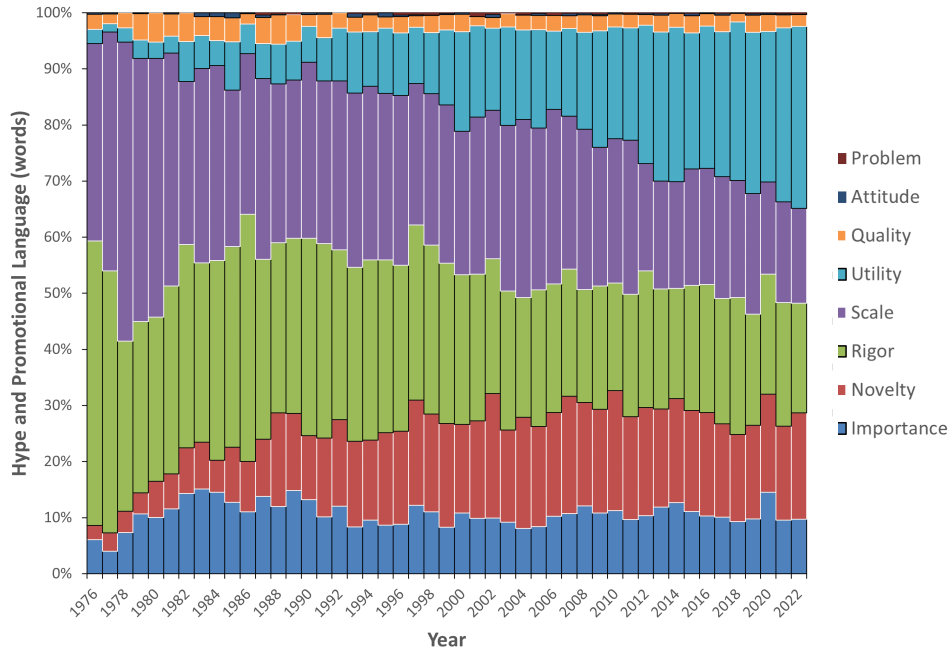


(a) Hype categories for titles

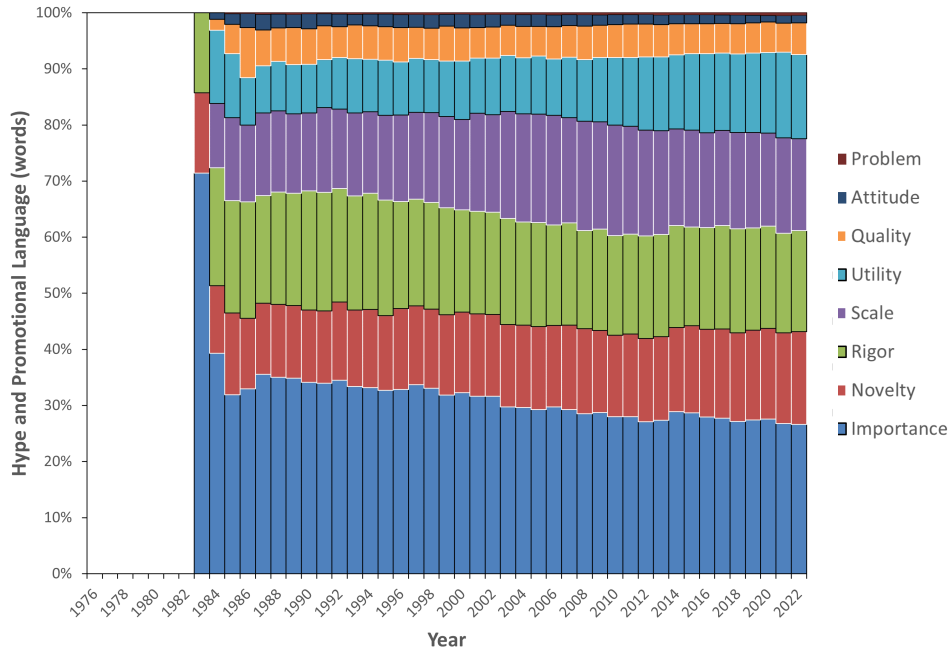


(b) Hype categories for abstracts

Fig. 53 A stacked bar chart offering insights into the temporal evolution of percentages associated with various hype word categories within a) titles and b) abstracts from 1976 onward. Each unique color corresponds to a distinct hype word category, and each bar symbolizes the stacked percentage of hype words (excluding stop words) in relation to the total words present within the title for a specific year. This visualization examines the shifting prevalence of different hype word categories over the years in NSF awards.



(a) Hype categories for titles



(b) Hype categories for abstracts

Fig. 54 The temporal evolution of percentages associated with various hype word categories within a) titles and b) abstracts from 1976 onward. Notably, starting in the mid-1980s, the NSF dataset began to include abstracts alongside titles. Each distinct color in the chart corresponds to a specific hype word category, and each bar visually represents the stacked percentage of hype words (excluding stop words) relative to the total words within the abstract for a given year. This visual representation provides valuable insights into the changing prevalence of different hype word categories over time within the abstracts of NSF awards.

Dashboard for Hype Language within NSF dataset

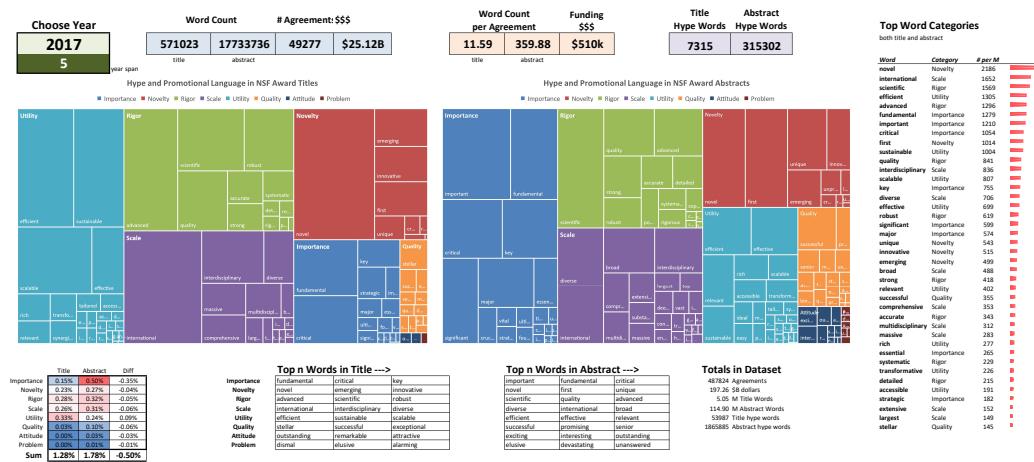


Fig. 55 A dashboard view for hype language within the NSF awards dataset. The upper left features two dropdown lists: one (above) for the year and one (below) for the span (going backwards). Once selected, the tree maps and associated statistics update with statistics for those years (or year). Shown is the dashboard for the years from 2013–2017, encompassing 49,277 awards for a total of \$25 billion in funding; the average title length is approximately 11.5 words, the average abstract length is 360 words; there are 7,315 hype words in the titles, and 315,302 hype words in the abstracts. The top hype words are novel, international, scientific, efficient, advanced, and fundamental.

4.3.6 Hype Words in Books – Google Books Ngrams

While the preceding analysis has delved into the highest-frequency hype words and those that have seen drastic increases over the decades, such trends may be reflective of broader shifts in language usage over time. For example, the increased frequency of the word “novel,” exceeding 2000 occurrences per million words in Table 6, might be a general linguistic trend. This prompts the question: “What is the baseline rate of these words” usage over time?” To address this, the Google Books Ngram Viewer^{20,21} was employed to track the frequency of various hype words from 1976 to 2019, representing the culmination of the English 2019 corpus in the Ngram Viewer. The English 2019 corpus, encompassing books predominantly in the English language from around the globe, comprises a vast dataset. Version 2, released in 2012, contains at least four and a half million books with close to half a trillion words, and the subsequent version 3, utilized in this analysis and released in 2020, likely includes an even more extensive collection of digitized books and words. Consequently, this dataset stands as the largest digitized compilation of books available, serving as a robust baseline for comparing the use of hype language.

Figure 56 depicts a sunburst plot showcasing different hype categories and hype words within the Google Books Ngram dataset for English 2019 corpus from 2015 to 2019. This corpus, rich in scientific literature, is often criticized for its potential bias away from common usage. However, for the purposes of this study, the prevalence of scientific literature in the corpus may enhance its suitability as a reflection of the use of these hype words in NSF awards. Generally, the frequencies of the hype words exhibit minimal year-to-year variations, as indicated by a positive correlation coefficient above 0.99 for all years from 2006 to 2019, and a correlation coefficient of 0.96 when compared to frequencies from 1976. The inner circle presents statistics for the hype categories, derived from the frequencies of the hype words within each category. Similar to observations in the NSF dataset, the most prevalent hype categories include importance, novelty, scale, utility, and rigor, with additional contributions from quality, attitude, and problem (not labeled). The outer circle displays the frequency of different hype words within these categories. Notably, some of the top hype words in terms of frequency (labeled in figure) align with those observed in the NSF dataset: novel (ranked as 1), scientific (3), international (5), critical (7), first (8), and important (11). Of these, “first” emerges as the most frequent hype word in English 2019 corpus by a significant margin. It’s

worth mentioning that, like some other hype words, "first" serves various grammatical functions, being used as an adjective, adverb, noun, and number (in this case, functioning as an adjective approximately 80% of the time).

Sunburst Diagram for Google Books Ngrams of Hype Words

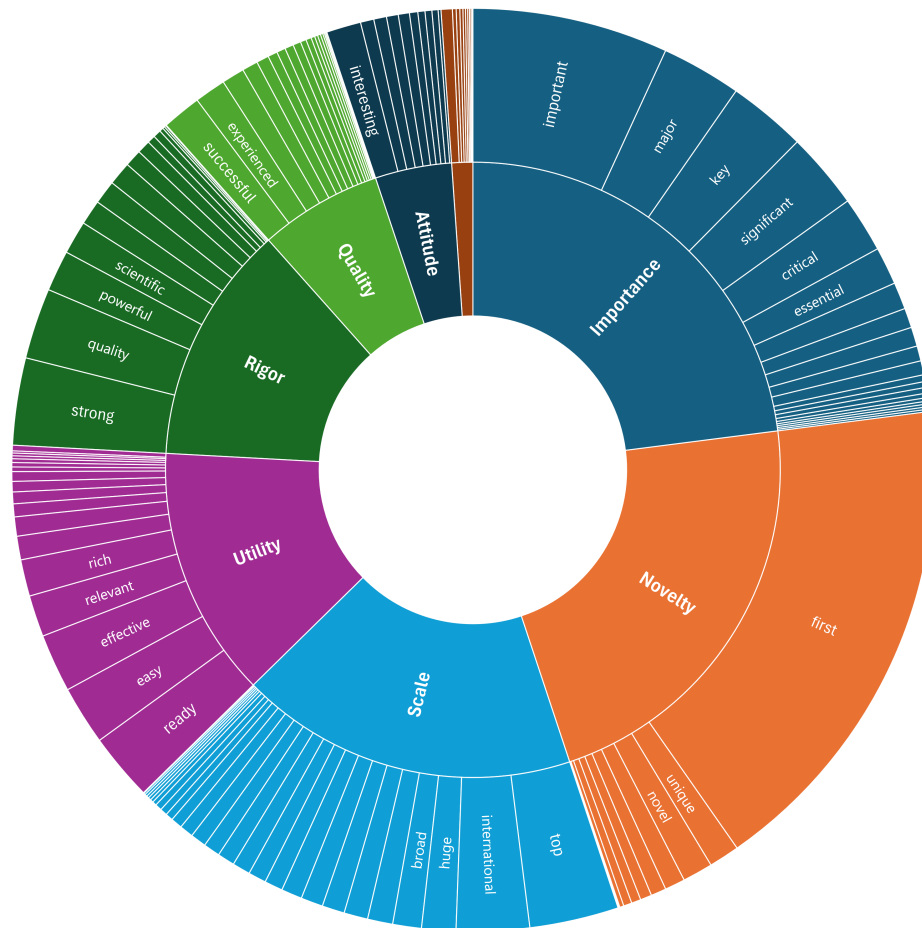


Fig. 56 A sunburst diagram for Google Books Ngrams dataset generated for hype word statistics from the English 2019 corpus for the years 2015–2019. The inner circle represents the different hype categories, obtained from the sum of the percentages for the different hype words within those categories and sorted clockwise in decreasing order. The outer circle represents the various hype words within those categories, sorted clockwise for each category in decreasing order. Not all hype words are labeled due to space and the only category not labeled is “Problem.”

The examination of correlation coefficients between the frequencies of the 139 identified hype words provides insights into whether the usage of these words is merely a reflection of broader language trends. Surprisingly, the correlation coefficients for the frequencies in the same year (2019) are found to be low: 0.17 for NSF titles and the English 2019 corpus, and 0.39 for NSF abstracts and the English 2019 corpus. Notably, this low positive correlation is not consistent over time. In the period from

1985 to 2002, the correlation coefficients (for NSF abstracts) consistently exceeded 0.70, indicating a strong similarity between the frequencies of hype words used in books and those in NSF abstracts. However, after 2002, there has been a gradual decline in the correlation of these frequencies, suggesting a divergence between the use of hype words in books and their use in NSF abstracts (as well as titles).

4.3.7 Hype Word Comparison with Google Books Ngrams

For a direct comparison with the NSF hype word analysis, it is imperative to adjust the frequencies obtained from the Google Books Ngrams dataset to account for both stop words and hyphenation. To achieve this, stop words and hyphenation information from Natural Language Toolkit (NLTK) in Python were employed to refine the frequencies within the English 2019 corpus, categorized by year. The collective percentage of stop words and hyphenation has consistently ranged from around 35% to 40% since 1976. With these adjustments, the frequencies derived from the English 2019 corpus can now be effectively compared with the frequencies obtained from the NSF awards.

Table 8 presents the top 40 hype words and categories based on their percentage increase in frequency (% increase columns) over the five-year period from 2018 to 2022, comparing data from NSF titles and abstracts against the Google Books Ngram dataset. The relative % difference is shown and the last column displays the rank order of the three preceding columns, i.e., for deployable, it was the 9th highest percentage in 1988–1992, the 2nd highest in 2018–2022, and the second highest relative difference. The average ranking or rank.avg used is the rank of a number in a list of numbers, i.e., its size relative to other values in the list; if more than one value has the same rank, the average rank is returned.

Notably, when sorted by their percentage increase over Google Books English 2019 corpus in 2018–2022,^{*} five words emerge in the top 10: scalable, deployable, interdisciplinary, transdisciplinary, and multidisciplinary. In the percent increase columns, a 100% increase signifies a frequency within NSF titles/abstracts double that of the English 2019 corpus; a 200% increase implies a tripled frequency, and so forth.[†] For instance, a 32,592% increase means that “scalable” is observed 326.92

^{*}Recall that this corpus ends in 2019, so the missing values are padded with those from 2019.

[†]A percent increase of -100% means that those words did not appear over the span of years selected, i.e., this happened 61 times in 1998–1992 and 29 times in 2018–2022 for the 278 hype words (rank.avg of 248 and 264, respectively).

times more in NSF titles than in the English 2019 corpus. The top 39 words in Table 8 were all used over an order of magnitude more than in the English 2019 corpus (up from 11 in 1988–1992). The “utility” category stands out with 18 occurrences in the top 40, followed by “scale” with 7, and “rigor” with 6, among others. This observation suggests a notable trend toward emphasizing the practical impact of research, possibly in response to the growing emphasis on assessing the overall impact of basic research during the evaluation process by funding agencies and reviewers. Including such terminology in abstracts and titles may serve as a strategic approach to address these evolving expectations.

Examining the rank order shows that there are a number of words that did not appear in 1988–1992 but appeared in high frequency in 2018–2022 time span: transdisciplinary, transformative, impactful, actionable, generalizable. There were 25 hype words that did not appear in either time range (absent in all the titles): attractive, collegial, compelling, considerable, daunting, desperate, devastating, dismal, enormous, groundbreaking, immense, indispensable, interprofessional, intriguing, invaluable, myriad, notable, nuanced, overwhelming, paramount, prestigious, renowned, substantial, tremendous, and unanswered. A number of these are in the “scale” hype category (8 of them: considerable, enormous, immense, interprofessional, myriad, overwhelming, substantial, tremendous) as well as the “problem” hype category (5 of them: daunting, desperate, devastating, dismal, unanswered). For “scale,” the frequency of these words indicates that convincing reviewers of the {inter, trans, multi}disciplinary nature of the proposed research is far more important than convincing them of the immense, overwhelming, and substantial nature of the research. For “problem,” it appears that it is not effective to use real estate to convince the reviewers of the dismal, dire, and unanswered problems, even though these sorts of hype words may reappear within the proposal itself. In Appendices B and C (NSF titles and abstracts, respectively), all of the hype words are broken out for each hype category with corresponding rankings based on only titles or abstracts.

Table 8 Difference in hype words from 1988–1992 to 2018–2022

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase)	2018–2022	Relative % Difference	Rank.avg Order
1	scalable	utility	title	5,191	32,592	27,401	[1, 1, 1]
2	deployable	utility	title	1,116	8,681	7,565	[9, 2, 2]
3	scalable	utility	abstract	1,154	6,193	5,039	[8, 3, 3]
4	interdisciplinary	scale	title	2,374	5,591	3,216	[3, 4, 6]
5	transdisciplinary	scale	title	-100	4,711	4,811	[248, 5, 4]
6	interdisciplinary	scale	abstract	1,734	4,427	2,693	[6, 6, 7]
7	deployable	utility	abstract	96	4,153	4,057	[73, 7, 5]
8	multidisciplinary	scale	title	2,027	3,882	1,855	[5, 8, 13]
9	multidisciplinary	scale	abstract	1,556	3,062	1,506	[7, 9, 18]
10	transdisciplinary	scale	abstract	396	2,867	2,471	[28, 10, 8]
11	robust	rigor	title	2,225	2,744	519	[4, 11, 38]
12	synergistic	utility	title	698	2,673	1,975	[15, 12, 11]
13	stellar	quality	title	3,121	2,551	-570	[2, 13, 278]
14	generalizable	utility	abstract	614	2,477	1,863	[21, 14, 12]
15	reproducible	rigor	title	-23	2,261	2,284	[127, 15, 9]
16	sustainable	utility	title	-15	2,119	2,135	[121, 16, 10]
17	efficient	utility	title	248	1,883	1,635	[39, 17, 15]
18	transformative	utility	abstract	-100	1,673	1,773	[248, 18, 14]
19	user-friendly	utility	abstract	658	1,565	907	[17, 19, 27]
20	longstanding	quality	abstract	307	1,518	1,211	[35, 20, 22]
21	synergistic	utility	abstract	313	1,492	1,179	[34, 21, 23]
22	novel	novelty	title	1,040	1,481	441	[11, 22, 42]
23	impactful	utility	abstract	-100	1,468	1,568	[248, 23, 16]
24	actionable	utility	title	-100	1,467	1,567	[248, 24, 17]
25	generalizable	utility	title	-100	1,398	1,498	[248, 25, 19]
26	diverse	scale	abstract	390	1,382	992	[29, 26, 26]
27	transformative	utility	title	-100	1,233	1,333	[248, 27, 20]
28	innovative	novelty	abstract	422	1,192	771	[25, 28, 32]
29	reproducible	rigor	abstract	132	1,164	1,032	[58, 29, 24]
30	actionable	utility	abstract	-100	1,164	1,264	[248, 30, 21]
31	stellar	quality	abstract	1,078	1,075	-4	[10, 31, 236]
32	advanced	rigor	title	879	1,063	184	[12, 32, 69]
33	robust	rigor	abstract	722	1,038	316	[14, 33, 54]
34	fundamental	importance	abstract	622	1,027	406	[19, 34, 44]
35	novel	novelty	abstract	370	1,000	630	[30, 35, 35]
36	seamless	utility	title	128	999	871	[60, 36, 29]
37	scientific	rigor	abstract	680	981	301	[16, 37, 56]
38	foundational	importance	abstract	617	947	329	[20, 38, 53]
39	impactful	utility	title	-100	902	1,002	[248, 39, 25]
40	emerging	novelty	title	82	853	771	[76, 40, 31]

4.3.8 Hype Word Temporal Evolution over Time

Probing the temporal evolution of the identified hype words may also be of interest. Figure 57 illustrates the yearly frequencies of the hype word “scalable” in words per million, comparing the frequencies based on NSF titles and abstracts within a given year with those based on the English 2019 corpus. The primary axis corresponds to the NSF title-abstract combination, while the secondary axis on the right represents the English 2019 corpus. Notably, the secondary axis is often scaled to a smaller range to highlight variations in the English 2019 corpus. Instances of yearly variation are observed within the NSF award corpus, prompting the adoption of five-year ranges for statistical analysis, such as 1988 to 1992 or 2018 to 2022. The plot begins at 1985, signifying the inclusion of abstracts with awards in the NSF dataset. In the initial years post-1985, a substantial increase in abstract word count occurred until 1988, indicating the progressive inclusion of more abstracts in the award XML.

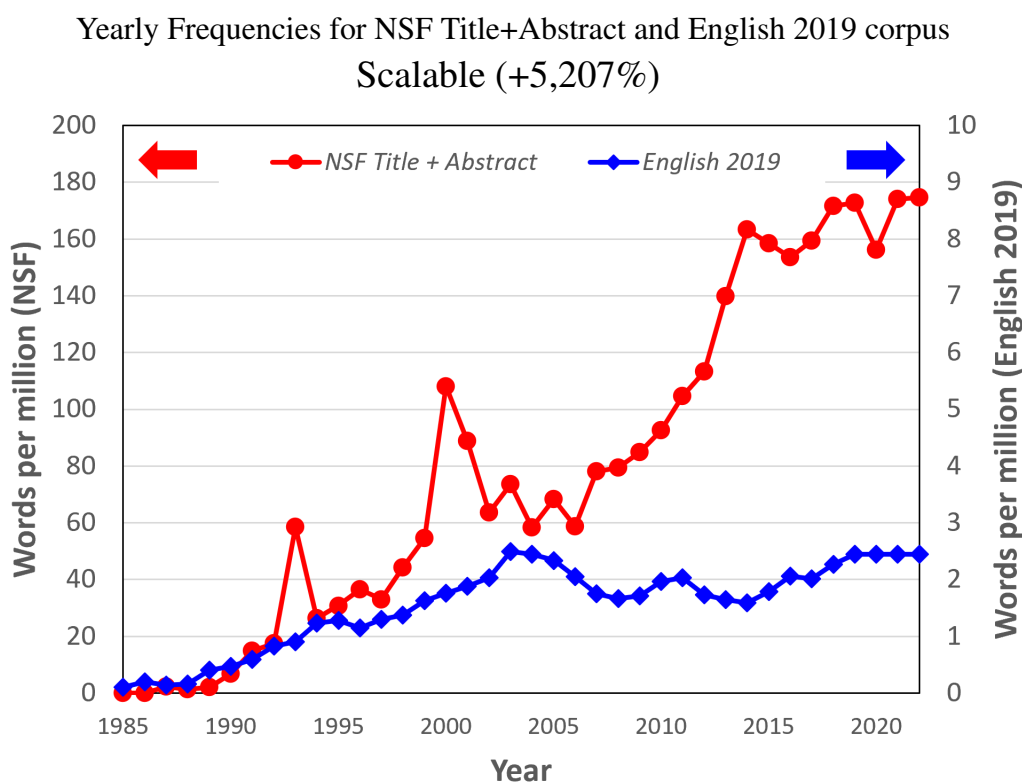


Fig. 57 The evolution of yearly frequencies for the hype word “scalable” (wpm) for: NSF title+abstract (red, primary y axis) and the Google Books English 2019 corpus (blue, secondary y axis). The scales on the primary and secondary y-axes are selected to highlight the evolution over different wpm ranges. The % increase of NSF title+abstract from English 2019 corpus for 1988–1992 and 2018–2022 is calculated; then the % difference over that 30-year span is shown in the plot title.

The percentage value displayed after the word “scalable” in the title indicates the calculated percentage difference and its significance. The words per million were computed for both datasets within two distinct 30-year periods: 1988 to 1992 and 2018 to 2022. The percentage difference between NSF values and English 2019 corpus values was calculated for each period, where 0% implies identical word frequencies, and 100% indicates that the NSF word frequency is precisely double that of the English 2019 corpus. The difference between these percentages offers a comprehensive measure of the increased (or decreased) use of the hype word, considering both its temporal evolution and its comparison with a baseline frequency in the English 2019 corpus. For example, the 5207% “scalable” percentage indicates that its use is 53 times more frequent in 2018 to 2022 compared to 30 years prior, accounting for the nominal base rate within the Google Books digitized dataset. It’s important to note that a 0% increase suggests no change over time relative to the base rate, without specifying whether the word frequency is above or below the base rate. An NSF word frequency double the base rate in both time ranges yields a 0% increase overall.

Figure 58 provides a visual representation of the evolution of four words within the scale category: international, interdisciplinary, multidisciplinary, and transdisciplinary. Interestingly, the base rate does not exhibit significant growth over time for all words, but the word frequency for the selected four words undergoes considerable changes across the specified range. The arrangement of these words is based on their decreasing frequency in NSF titles and abstracts in 1985, and each word follows a distinct trajectory up to 2022. The hype word “international” (-14%) experiences a net decrease over the 30-year span, while the remaining words demonstrate marked increases.

The term “interdisciplinary” (+2,667%) gradually intensifies in frequency until 2005-2010, surpassing 400 words per million (wpm). “Multidisciplinary” (+1,489%) undergoes a significant increase around 1995, maintaining high frequencies until 2010, exceeding 150 wpm. On the other hand, the hype word “transdisciplinary” (+2,668%) emerges around 2010 from low word frequencies (5 wpm or lower) to reaching frequencies above 30 wpm. The varied trajectories of these words highlight the dynamic shifts in their usage patterns over the three decades.

Beyond their temporal evolution, understanding the nuanced distinctions among

Yearly Frequencies for NSF Title+Abstract and English 2019 corpus

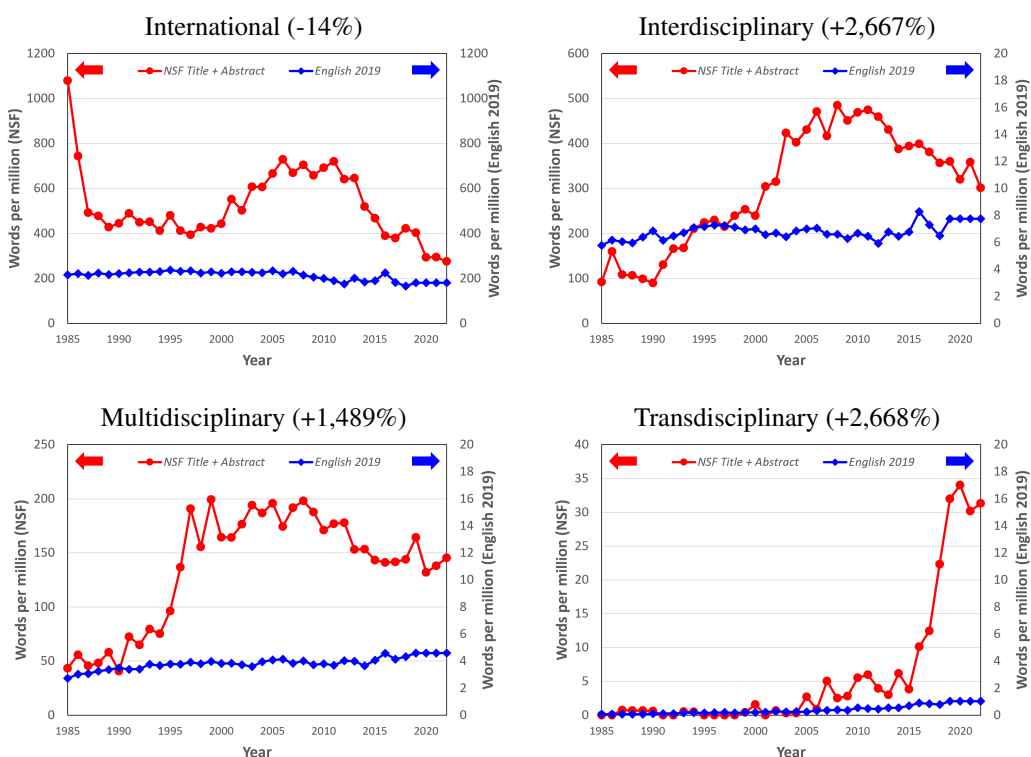


Fig. 58 The evolution of yearly frequencies for international and {inter, multi, trans}disciplinary hype words (wpm) for: NSF title+abstract (red, primary y axis) and the Google Books English 2019 corpus (blue, secondary y axis). The scales on the primary and secondary y-axes are selected to highlight the evolution over different wpm ranges. The % increase of NSF title+abstract from English 2019 corpus for 1988–1992 and 2018–2022 is calculated; then the % difference over that 30-year span is shown in the plot title. This plot is organized from left to right and top to bottom by frequency of use in 1985.

the words within the scale category adds depth to the analysis. “Interdisciplinary,” a term prevalent in NSF titles and abstracts in the mid-1980s, reflects an approach that integrates insights and methodologies from various disciplines, addressing complex issues by combining disciplinary perspectives. This popularity 30 years ago might be attributed to the academic landscape’s emphasis on collaborative efforts across traditional disciplines during that period.

On the other hand, “multidisciplinary” experiences a surge in frequency around 1995, signifying a heightened recognition of the value in engaging multiple disciplines, each maintaining its independence. This approach allows for a comprehensive exploration of a topic from distinct disciplinary angles. The notable increase in usage might be linked to a growing acknowledgment of the importance of diverse

disciplinary contributions.

In contrast, “transdisciplinary,” emerging around 2010, represents a more holistic and integrative paradigm that transcends disciplinary boundaries, fostering collaboration beyond traditional academic silos. The significant rise in frequency suggests a contemporary shift towards more interconnected and holistic research approaches. This change may reflect an evolving scholarly landscape that increasingly values collaborative, boundary-spanning research endeavors.

These interpretations provide a qualitative backdrop to the quantitative trends observed in the frequency patterns of “interdisciplinary,” “multidisciplinary,” and “transdisciplinary” over the 30-year span. They showcase not only shifts in academic preferences but also the evolving nature of collaborative research paradigms.

Figures 59 and 60 provide an insightful look into the evolution of 12 selected words that exhibit a positive percent increase over the 30-year span. Figure 59, inspired by the evolution of hype word “transdisciplinary,” highlights several words with extremely low frequencies before 2000. Notably, these words experience a sharp increase in frequency from an instantiation year of around 2005 (e.g., “transformative”) to 2010 (“impactful”). With the exception of “reproducible” (Rigor category) and “foundational” (Importance category), these words predominantly fall within the Utility category, elucidating how the research will be beneficial.

On the other hand, Fig. 60 showcases another 12 words that had higher frequencies in 1985. While many of these words might not exhibit as high a percent increase over the 30-year span, their initial higher frequencies make a sevenfold increase for “novel” or an eightfold increase in “innovative” or an order of magnitude increase in “diverse” substantial in words per million. Intriguingly, words like “diverse” were used less frequently in NSF titles and abstracts before 2005 but have been gaining increased usage relative to their occurrence in books over time. Some words, like “novel,” maintained a similar rate of use in the 1988–1992 timeframe but are now used seven times more frequently in NSF titles and abstracts than in the digitized book corpus. Conversely, words like “scientific” and “key” were used more frequently in NSF titles and abstracts 30 years ago and are now employed even more often.

While the majority of these words exhibit an increase over time, it’s essential to

Yearly Frequencies for NSF Title+Abstract and English 2019 corpus

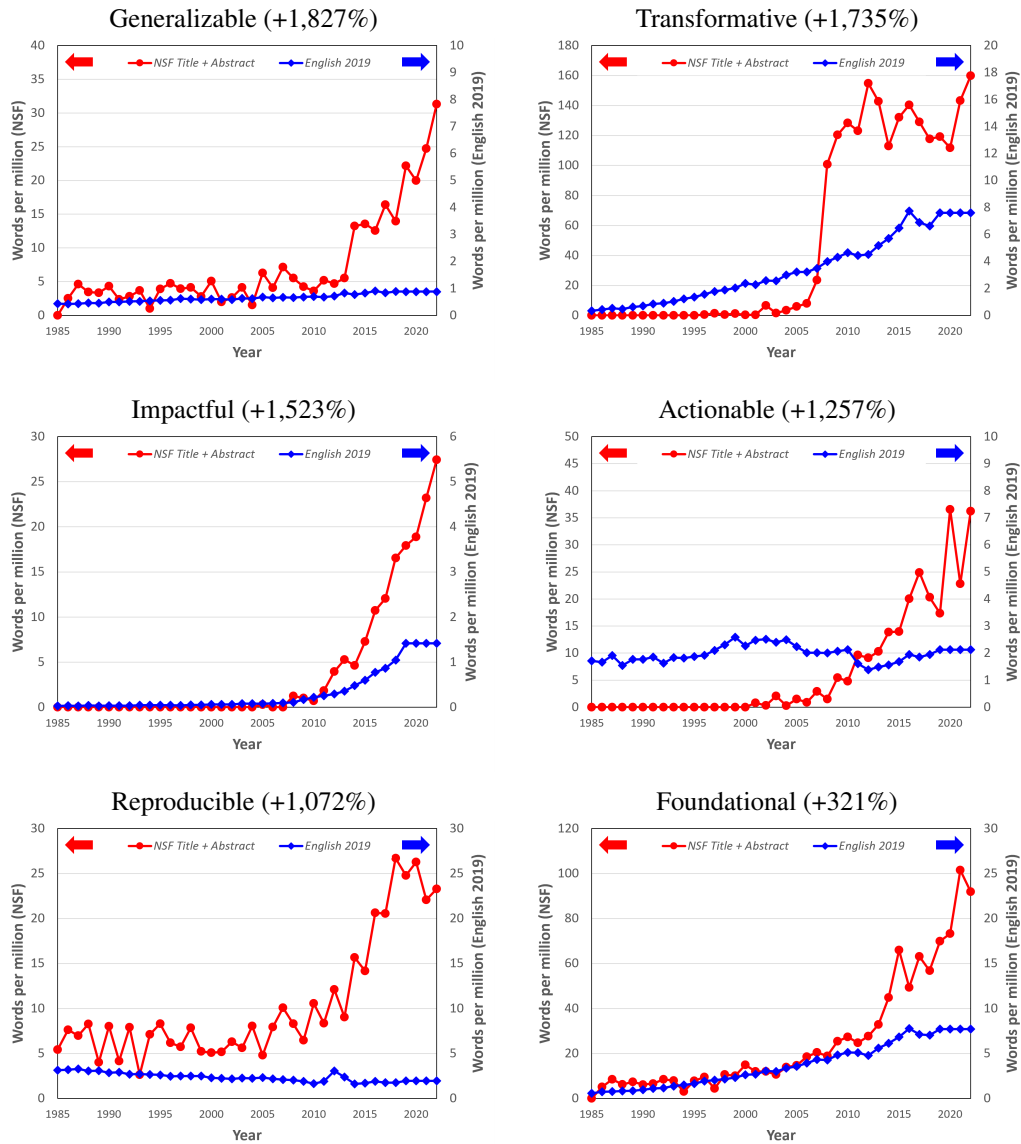


Fig. 59 The evolution of yearly frequencies for hype words (wpm) of generalizable, transformative, impactful, actionable, reproducible, and foundational for: NSF title+abstract (red, primary y axis) and the Google Books English 2019 corpus (blue, secondary y axis). The scales on the primary and secondary y-axes are selected to highlight the evolution over different wpm ranges. The % increase of NSF title+abstract from English 2019 corpus for 1988–1992 and 2018–2022 is calculated; then the % difference over that 30-year span is shown in the plot title. This plot is organized from left to right and top to bottom by this percentage.

note that certain hype words have decreased in frequency relative to the English 2019 corpus over time. Out of the 139 hype adjectives, 109 words saw an increase in frequency over time (averaging a 56 wpm increase, 88% higher), while 30 hype adjectives experienced a decrease in frequency (averaging a -17 wpm, 33% lower).

Yearly Frequencies for NSF Title+Abstract and English 2019 corpus

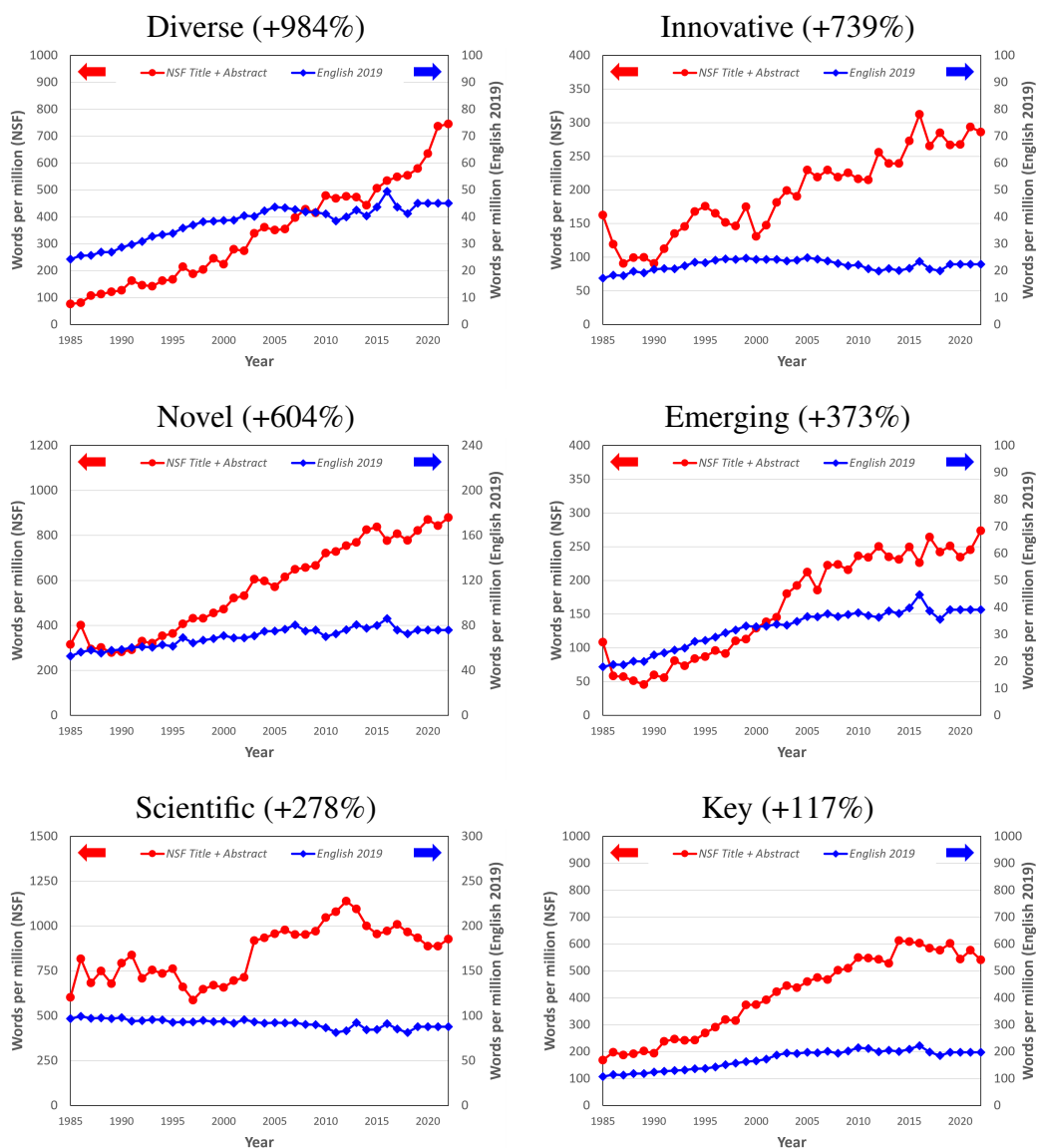


Fig. 60 The evolution of yearly frequencies for hype words (wpm) of diverse, innovative, novel, emerging, scientific, and key for: NSF title+abstract (red, primary y axis) and the Google Books English 2019 corpus (blue, secondary y axis). The scales on the primary and secondary y-axes are selected to highlight the evolution over different wpm ranges. The % increase of NSF title+abstract from English 2019 corpus for 1988–1992 and 2018–2022 is calculated; then the % difference over that 30-year span is shown in the plot title.

This nuanced analysis sheds light on the dynamic trends in the usage of hype words over the past three decades.

4.3.9 Using Large Language Models to Focus on Content

In the grand scheme, what is the utility of comprehending hype words and their usage in titles, abstracts, and even proposals? Fundamentally, as a funding agency, the aim is to support the most commendable proposals based on the merit of their ideas and approaches rather than the abundance of flowery adjectives in their expression. As a funding agency, sound data-driven decisions are sought—ones firmly grounded in the substance of the research. The study of hype language opens avenues for intriguing investigations. For example, does the removal of such language impact reviewers' assessments, even with identical content? Is there a discernible threshold for hype language, beyond which proposals face penalties? What are the most objectionable words that may influence or deter reviewers? Has the threshold for hype language evolved over time, with present-day reviewers more tolerant than those from three decades ago? This last question may involve examining both failed proposals as well as the successful awarded ones. In many ways, the natural experiments are much easier to grasp than an A versus B test for hype language.

Consider the NSF award with the highest number of hype words in the abstract for 2022. Through an iterative process of guidance and prompts to ChatGPT (GPT-3.5) to reduce hype words while retaining content, Fig. 61 presents a revised version of the most hyped abstract in 2022. The total number of hype words now stands at 13, and the abstract length is just over 200 words—both reductions exceeding 50% from the original 35 hype words and 428-word length. While further reductions in hype words are possible, the objective is not to eliminate them entirely but to illustrate how minimizing such language can enhance readability. Whether this contributes to better or worse reviews depends on the reviewer. Do they critique more rigorously when the content is clearer? Can hype words and verbosity confound the reviewer, shifting their focus from content to hype? Alternatively, do hype words assist in directing the reviewer's attention to the innovative or scientific aspects of the research? If any of these scenarios hold true, it incentivizes PIs to include hype language. This could be one of the factors contributing to the observed increase in promotional language in award titles and abstracts over time.

The reworking of this exaggerated abstract sparks an intriguing consideration. Could large language models serve not only to enhance grammar and clarity in presenting ideas to reviewers but also as a means to mitigate the potential impact of hype language on their assessments? The observed rise in hype and promotional language

CDATA [Science Integrated Predicted Modeling (SCINPL): a novel framework for scalable and interpretable predictive scientific modeling]

Modeling in science is at a critical juncture, driven by advances in experimental technology that provide high-quality data for intricate scientific problems. However, the cost of generating such data presents challenges, limiting scientific investigation. Existing predictive models have shortcomings, struggling with poor performance on limited data and occasionally violating established principles, leading to potential inaccuracies. This project introduces the Science-INtegrated Predictive modelIng (SCINPL) framework to address these limitations. SCINPL aims to develop predictive models to enhance science, fostering collaboration between scientific and data science communities. It represents a paradigm shift, empowering scientists across diverse fields with improved tools for data-driven scientific discovery. SCINPL includes innovative probabilistic Bayesian models capable of integrating prior knowledge for predictive modeling. The models, namely the Boundary-constrained GP, Graphical Multi-fidelity GP, and Gaussian Process Subspace regression, enhance predictive performance, reduce uncertainty, and enable better interpretability. The project emphasizes a solid theoretical foundation, a comprehensive methodological framework, and efficient algorithms for integrating scientific principles into probabilistic modeling. The effectiveness of SCINPL is demonstrated across complex problems, including 3D-printed aortic valve design, heavy-ion collision studies, and rocket engine optimization for spaceflight. This NSF-supported project aligns with the Foundation's mission, emphasizing intellectual merit and broader impacts in advancing the discovery of new science.

Fig. 61 This abstract version was iteratively edited in a series of prompts with ChatGPT (GPT 3.5) for the NSF award abstract from 2022 that contained the most hype words (in red). The new version contains 13 hype words (from 35) and 204 words (from 428), reducing the total numbers by about 50%.

over the decades suggests that reviewers and funding agencies might be swayed by the phrasing of proposals in addition to their content. Large language models could standardize the language in a way that tempers excessive hype, thus ensuring that reviewers can evaluate proposals on a fair and level playing field. Nevertheless, challenges persist, as these models may struggle to grasp the nuanced meaning and context inherent in scientific work. As of today, this option may not be fully realized, but it presents an intriguing vision for the future.

4.4 Concluding Remarks: Insights and Future Research Directions

This study employed a multi-faceted approach to investigate the evolution and impact of hype language in open science proposals. The primary dataset consisted of NSF award titles and abstracts spanning from 1976 to 2022, providing a comprehensive view of language trends over several decades. Complementing this, Google Books ngrams were utilized to analyze English language usage from 1976 to 2019, offering a broader perspective on the prevalence of hype words in a vast corpus of literature. NLTK tools were employed for stop word removal and addressing hyphenation in the Google Books dataset. The analysis involved the identification of hype words and categorization into specific thematic groups (Table 4), enabling a nuanced exploration of language trends. Additionally, iterative modifications were made to abstracts using prompts and ChatGPT to reduce hype language while re-

taining content, providing insights into potential readability improvements. This multi-pronged methodology facilitated a comprehensive understanding of the temporal dynamics and potential implications of hype language in the context of open science proposals.

1. **Hype in Awards:** The analysis shows the distribution of hype language within both titles and abstracts (Figs. 50 and 52, respectively). The hype categories most prominent are (in decreasing order): Utility (how useful is the research), Rigor (how rigorous are the methods), Scale (how large are the network interactions), Novelty (how innovative is the research), and Importance (how significant is the research), with a few categories that are not observed all that much (Quality, Attitude, Problem). The most frequently used words are (Table 5, in decreasing order): novel (1982 words per million), efficient, scientific, advanced, international, fundamental, critical, first, diverse, and sustainable (967).
2. **Hype Evolution over Time:** The study revealed the evolution of hype language in open science proposals from 1976 to 2022, showcasing shifts in emphasis and frequency of specific hype words and categories. Utility and Novelty categories have shown the largest increases in NSF titles and abstracts over time, at the expense of Scale, Rigor, and Importance (Figs. 53 and 54). The largest increases in use of hype words over 3 decades are (Table 6, in decreasing order): transdisciplinary (7626%), sustainable, scalable, foundational, reproducible, safer, diverse, emerging, skilled, unprecedented (491%).
3. **Hype in Awards vs Hype in Books:** Understanding the difference between hype language's utility in NSF awards versus Google Books Ngram's English 2019 corpus provides a better understanding of what hype words and categories are merely a reflection of changes in usage and what are actually unique to awards. The hype categories appear to be similar in their rank order, with the top 5 and bottom 3 categories the same (Fig. 56), but the correlation between the award hype words and that in the largest collection of digitized books is low (correlation coefficients below 0.40). Also, this correlation has been shifting from high correlation (above 0.70) decades ago to much lower in the recent past, a sign that the emerging growth of hype language in awards

is not in line with the English 2019 corpus. The largest differences for 2018–2022 are: scalable (32,592% increase), deployable, interdisciplinary, transdisciplinary, and multidisciplinary (all over 2800% increase).

4. **Future Prospects of Language Models:** As hype words have become more prevalent, having a balanced approach that effectively conveys the novelty and significance of research without compromising clarity is a good procedure for PIs to follow. Iterative adjustments to one example demonstrated how reducing hype words may enhance readability, raising questions about how both hype and clarity impacts reviewer assessments. Unanswered questions about reviewers' sensitivity to hype language prompt considerations about whether clearer language could lead to fairer evaluations. One example is given for the most hype word abstract of 2022 (compare Figs. 48 and 61). Considering the potential influence of language models, their role in providing standardized yet context-aware language can lead to a future vision wherein they contribute to fair and unbiased reviews.

The extent to which hype matters has been shown to be significant. It does influence reviewers, potentially affecting their responses to proposals. Adjective choices can sway reviewers' perceptions, sometimes overshadowing the merit of the work itself. Additionally, there likely exists a threshold for acceptable hype for each reviewer, beyond which it becomes detrimental. When hype crosses this boundary, it risks eliciting a negative response from reviewers, resulting in rejected proposals, papers, or diminished recognition within the research community due to a lack of citation.

5. Research Study: Understanding Funding of Science from Open Scientific Databases

5.1 Introduction

Understanding the intricacies of scientific funding is important to unraveling the web of knowledge creation. Scientific publications routinely conclude with acknowledgment paragraphs, expressing gratitude to colleagues for insightful discussions, institutions for providing essential facilities and equipment, and funding agencies whose financial support made the research possible. Yet, what if researchers delved deeper into this wealth of information, scrutinizing the symbiotic relationship between scientific publications and the agencies that fuel their

progress? This study explores the funding dynamics within scientific literature, focusing on the vital link between research outcomes and the entities that invest in their realization. By analyzing the funding landscape, one aims to uncover patterns, trends, and insights that inform about the funding process—how science transverses from proposal to funding agency to award to its ultimate execution and scientific discovery.

Within the confines of this concise study, the focus turns to the information embedded in the acknowledgment sections of scientific publications, specifically honing in on details related to funding agencies and their corresponding agreement numbers. Among the myriad of open sources housing such information, this investigation zeroes in on OpenAlex, a comprehensive repository for mining funding details within academic literature. While the scope of this analysis could extend to various funding agencies, the attention is directed towards the ARO for an in-depth exploration of funding information within publication records.

The ARO stands out as a pivotal supporter of basic research within the Department of the Army, playing a central role in disbursing a significant portion of grants to PIs and researchers in academia, as shown in Figs. 62 and 63. As a funding agency for scholarly research, the ARO constitutes a majority of awards acknowledged in papers (54.2%), a majority of scientific articles published (61.3%), and the majority of citations accumulated from those articles (70%).* With a commitment to advancing foundational knowledge leading to future capabilities, the ARO constitutes a substantial contributor to the scientific community. By choosing the ARO as the focal point, a nuanced understanding is provided for the interplay between funding agencies and the scholarly outputs they nurture. This deliberate selection facilitates a granular analysis of the funding landscape, revealing insights that may have broader implications for the funding and execution of basic research across diverse domains.

In reality, conducting a thorough examination of the connection between the ARO and the impact of publications presents a complex challenge, especially given the time constraints of this study. Preliminary data analyses from databases like Cross-

*The second highest funder of scientific research for the Army is DEVCOM Army Research Laboratory, i.e., the parent organization of ARO. In many cases, when DEVCOM Army Research Laboratory is listed, this is often referring to cooperative agreement awards, a mechanism for collaboration between internal ARL scientists and external organizations (mainly academia and industry).

Percentage of Grants, Articles, and Citations for Army Funders

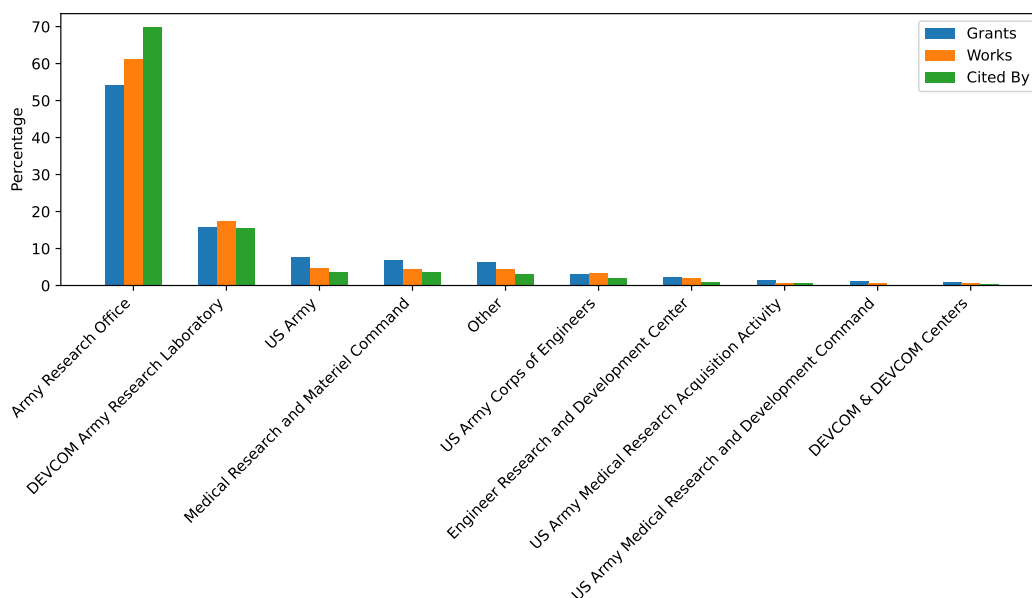


Fig. 62 A bar chart showing the percentage of grants, works, and citations for all Army funding agencies. Percentages are calculated from articles with an Army funding agency in the acknowledgments section and in OpenAlex’s “funder” category.

Percentage of Grants, Articles, and Citations for Army Funders

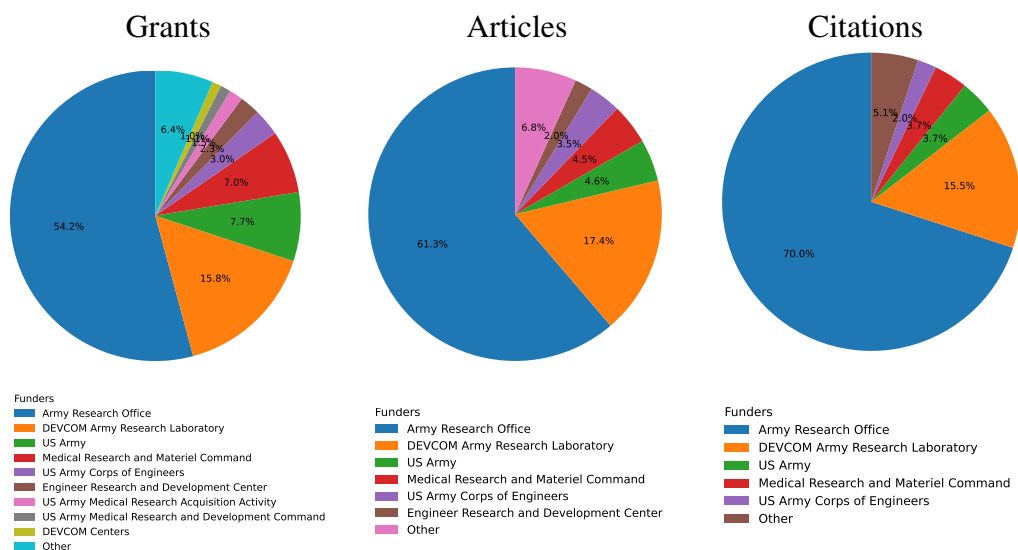


Fig. 63 A pie chart showing the percentage of grants, works, and citations for all Army funding agencies. Percentages are calculated from articles with an Army funding agency in the acknowledgments section and in OpenAlex’s “funder” category. The Other category includes those funders with less than 1% of the pie chart.

Ref and OpenAlex have revealed nuances in the quality of the information, emphasizing the need for careful data cleaning. It's worth noting that, due to limitations in available information, some datasets may remain less refined. This study highlights the pragmatic constraints of working with real-world data, where the intricacies of the information provided by PIs can vary. Despite these challenges, the goal remains to extract meaningful insights from the available data. The subsequent sections elaborate on the methodologies employed, ensuring transparency about the encountered constraints during the analysis process. By managing expectations regarding the data quality, this study aims to contribute a balanced perspective on the relationship between funding agencies, particularly the ARO, and the scholarly impact of publications. The results section will provide a more detailed exploration of the findings, including insights gained and challenges faced during the analysis.

5.2 Methodology

In the evaluation of Funding Agency information, scientific records from OpenAlex are leveraged, encompassing publication details such as the journal, DOI, title, publication year, and more. These records also include Authorships data, detailing authors' positions, affiliations, and countries, along with citation details, research concepts, sustainable development goals, abstracts, references, citations, and grants supporting the work, including the funder and award ID.

To illustrate the transformation from acknowledgment text within a publication to the grants section within the OpenAlex API, Fig. 64 exemplifies the original text containing funding agencies, the converted "grants" section in OpenAlex, and the "funder" section in CrossRef. The selected example is the most cited publication funded by ARO in 2022 at the time of this report, a review article with 162 citations and 764 references. This figure aims to demonstrate the general appearance of acknowledgment sections and their translation into structured grants data.

However, this process is not without complexities, as the example reveals. The funding text often references multiple organizations, raising questions about the distinction between entities like Novo Nordisk and Lundbeck foundations or Adelson and Simon foundations. The ambiguity extends to specific grants, such as the one from the ARO attributed to author M.N., prompting considerations about the broader implications for which authors are covered by which funding sources—there are only two authors on the paper in Fig. 64 and at least 7 different funding agencies ac-

Fluid transport in the brain

In original publication:

This work is supported by the Novo Nordisk and Lundbeck Foundations, the European Research Council (ERC) under the European Union's (EU) Horizon 2020 Research and Innovation Program Grant Agreement 742112, EU Horizon 2020 Research and Innovation Program Grant 666881 (SVDs@target), National Institute of Neurological Disorders and Stroke Grant R01-NS-100366, National Institute on Aging Grant RF1-AG-057575, US Army Research Office Grant Multidisciplinary University Research Initiative W911NF1910280 (to M.N.), Fondation Leducq Transatlantic Networks of Excellence Program, and the Adelson and the Simon Foundations.

In OpenAlex JSON file:

```
"grants": [
  {
    "funder": "https://openalex.org/F4320320890",
    "funder_display_name": "Fondation Leducq",
    "award_id": null
  },
  ...
  {
    "funder": "https://openalex.org/F4320338281",
    "funder_display_name": "Army Research Office",
    "award_id": "MURI W911NF1910280"
  }
],
```

In CrossRef JSON file:

```
"funder": [
  {
    "DOI": "10.13039/501100001674",
    "name": "Fondation Leducq",
    "doi-asserted-by": "publisher"
  },
  ...
  {
    "DOI": "10.13039/100000183",
    "name": "DOD | United States Army | RDECOM | Army Research Office",
    "doi-asserted-by": "publisher",
    "award": [
      "MURI W911NF1910280"
    ]
  }
],
```

Fig. 64 The acknowledgments or funding section within scientific publications are converted into lists of funding agencies and award IDs within OpenAlex and CrossRef. Shown above is the original text and how it is mapped onto funding agencies ("funder_display_name" in OpenAlex) with unique identifiers ("funder") and award IDs ("award_id").

knowledge. Notably, the challenge lies not only in accurately listing the funding agencies and award IDs but also in ensuring consistent and comprehensive coverage across all authors involved. With an average of over 125,000 scientific articles published weekly in 2022, the necessity of automation becomes apparent, prompting critical inquiry into the ability of automated models to accurately extract and compile this intricate information. The ensuing sections delve into the analysis of the funding conversion for these open datasets, providing insights into the potential limitations and solutions encountered during the extraction and database compilation process.

Define the standard format for ARO award IDs is necessary for understanding the cleaning operations. A correct award ID follows the pattern: W911NF-XX-X-XXXX. All ARO award IDs commence with “W911NF,” succeeded by two digits indicating the year of the award (e.g., 09 for 2009, 23 for 2023, etc.). The subsequent single character designates the award type: grant (1), cooperative agreement (2), contract (3), or an alternative specifier (9, C, D, P). The final four digits, assigned during acquisition processing, represent the order of processing, albeit potentially with gaps. For uniformity, hyphens separating characters are omitted.

5.3 Results and Discussion

5.3.1 Data Quality for Funder Award IDs in Open Databases

For an in-depth analysis of the link between awards and publications, the author downloaded data from the OpenAlex database encompassing all scientific articles since 2013-10-01 (19,582 articles) that list the ARO as one of the Funding Agencies. The focus initially narrows down to the top 100 most cited articles acknowledging ARO, prompting the investigation: How accurate are the award IDs associated with these highly cited publications?

Table 9 provides a detailed breakdown, including the rank order, citation counts, excerpts from the acknowledgments mentioning ARO, and an evaluation of whether the listed award IDs are correct. Several noteworthy observations emerge. Firstly, certain publications merely acknowledge the Funding Agency without providing an associated award ID, or they reference the program without specifying the ID (#16, #89). Secondly, a common issue arises where publications inadvertently omit key letters or numbers from the correct award ID (#32, #34, #96, #100). Thirdly, some publications refer to IDs that are not award IDs but rather proposal IDs used for tracking purposes before the award is granted (e.g., 73315-PH [#66], 55036-CH [#68]). Lastly, in certain cases, although the text accurately mentions the award ID, extraction algorithms face challenges, as seen in example #82, possibly due to specifics like a program manager’s mention causing confusion.

Crucially, over 90% of award ID discrepancies stem from inaccuracies within the publications themselves. This finding suggests that regardless of whether the data is sourced from open datasets or paid subscriptions (e.g., Web of Science), reliance on information extracted directly from papers is likely to yield inaccuracies in award ID representation.

Table 9 Publications funded by ARO grant with incorrect award ID listed in top 100 cited articles (FY18–FY23)

#	# Citations	Acknowledgments Reference to Funding Agency	OK?
16	519	ARO’s atomtronics MURI	No
32	373	Grant W911NF-09-0001 from the US Army Research Office (ICB)	No
34	363	Army Research Office (award no. W911NF121054), ARO-LPS	No
66	247	Army Research Office (ARO) (Grant No. 73315PH)	No
68	240	Army Research Office under the Single Investigator Grant 55036-CH	No
82	205	Army Research Office (Program Manager: Dr. Robert Mantz) under Award No. W911NF1510187	Yes
89	194	Army Research Office–MURI	No
96	184	ARO award W911NF-17-0512	No
100	177	ARO MURI Award No. W911NF-14-0247	No

This analysis extended to the top 1000 most cited articles acknowledging ARO as a Funding Agency, encompassing 1135 awards. Notably, some papers acknowledged ARO as the funding agency multiple times. Among these awards, 12% lacked a listed award ID, 18% were in an incorrect format (as determined by a search for W911NF and correct length), while 70% possessed the correct format. The absence of award IDs could stem from statements like “the authors acknowledge ARO for funding” or challenges in extracting the award ID from the text.

For the entire dataset, ARO appeared as the Funding Agency in 22,203 instances. Figure 65 shows a breakdown of the award IDs from the entire data set examined herein. Within this set, 10.9% of the time, award IDs were absent, 15.1% were in an incorrect format, while 74% adhered to the correct format. It’s important to note that, at this stage, “correct format” is defined by the presence of W911NF and adherence to the proper length for the award ID.

5.3.2 Data Cleaning of Award IDs, Part I

To refine the award IDs, a multi-step process aimed at reducing nonstandard formats and rectifying any potential misallocations to the wrong agency was employed. Manual rule insertion became necessary at times, aiding in the identification and classification of various issues encountered during award ID cleaning, as illustrated in Fig. 65.

In the subsequent step, certain award IDs exhibited the presence of two or more IDs joined by “and” or concatenated with a comma. These instances were straightforwardly split into distinct award IDs, resulting in the addition of 173 IDs to the

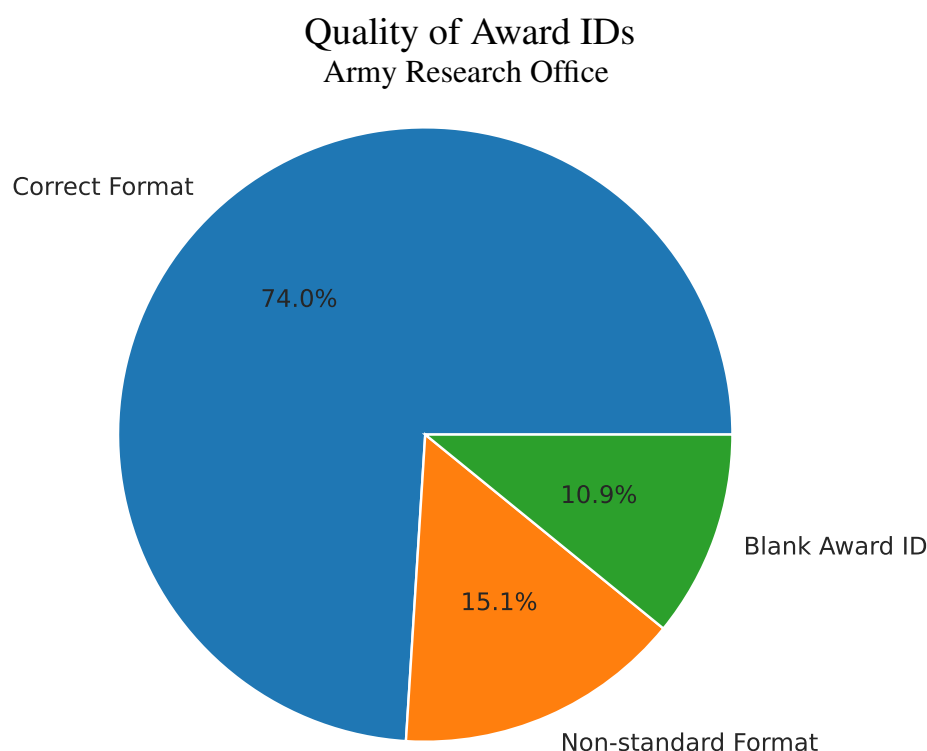


Fig. 65 A pie chart for ARO award IDs within OpenAlex “grants” section from scientific articles published within the FY2014–FY2023 time frame. This comprises 19,582 articles that acknowledge ARO as one of the funding agencies, of which 22,203 award records are listed.

19,788 non-blank award IDs. Each award then underwent a cleaning procedure that eliminated any characters preceding W911NF and removed all punctuation within the award ID, including hyphens, spaces, periods, colons, commas, ampersands, question marks, and more. This splitting and character removal led to a 3.75% increase in awards conforming to the correct format.

The subsequent cleaning phase addressed the diverse variations of W911NF within the first six characters of the award ID. Table 10 catalogs these variants, ranging from too few to too many characters or containing one, two, or three erroneous substitutions. At this point, whether the award ID is incorrectly entered in the article or whether the conversion is incorrect is hard to understand at scale without recreating the process. It is suspected that some of these—like “VV” instead of “W” or “I” instead of “1”—are due to the use of optical character recognition. By identifying and substituting these variations with the correct W911NF, a further improvement was achieved, with a few percentage points more of award IDs aligning with the correct format. A comprehensive breakdown of award IDs from the entire dataset

post-cleaning is presented in Fig. 66.

Table 10 Error variant examples on W911NF from award ID used within published articles

Error	Examples
Too few characters	NF, W911, W99NF, W9HNF, W91NF, W11NF, W19NF, W911N, W911F, W81NF, W81WH
Right number, not 911	W011NF, WN11NF, W91INF, W991NF, W811NF, WG11NF, W9IINF, W311NF, W922NF, W900KK
Right number, other	N911NF, W911NR, W911BF, W911NE, W911NG, @911NF, W911SR, W911NS, W911WF, 2911NF, W911TA, W911W6, W911NH, W911VF
Too many characters	WN911NF, W1911NF, W911NSF, W911INF, W911NNE, W911NFF, W911NFD, WF911NF, WG911NF, VV911NF, W911INF, W1911NF, W911NFNF, W911NF1N

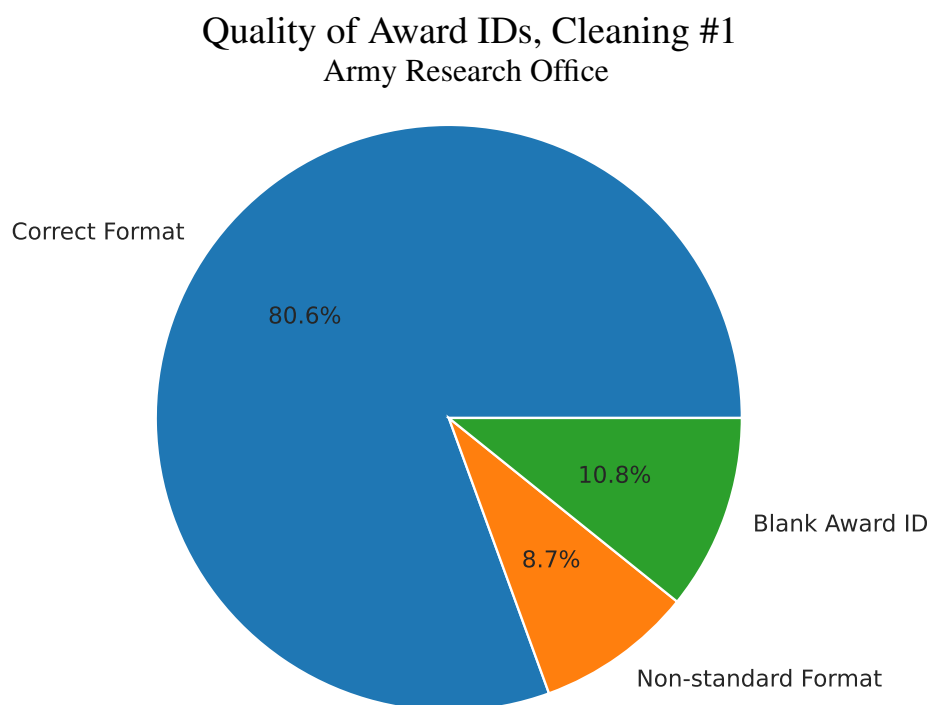


Fig. 66 A pie chart for ARO award IDs within OpenAlex “grants” section from scientific articles published within the 2014–2023 time frame. This snapshot is after applying a cleaning operation detailed within the text and Table 10.

5.3.3 Data Cleaning of Award IDs, Part II

The subsequent phase of cleaning focused on identifying award IDs in proposal format rather than the correct award format. This approach facilitates mapping by Funding Agencies with a proposal-to-award ID correspondence, allowing for the

correction of these entries within the databases. Proposal IDs adhere to the format XXXXX-XX, with the initial five characters being numbers, followed by two letters denoting the program. In certain cases, an additional three letters may signify additional information, such as affiliation with the Young Investigator Program (YIP), participation in a Multidisciplinary University Research Initiative (MURI), involvement in a Research Instrumentation Program, or engagement in an Other Transaction Agreement.

In the course of this process, 434 proposal IDs were effectively mapped to the correct award IDs. Interestingly, 13 proposal IDs were in the correct format but faced rejection. In several instances, PIs made references to the proposal ID in published papers before any decision on the award was finalized and ultimately they were rejected. This phenomenon could be filed under "tactics that PIs think may influence program managers" (though, in practice, they do not). Furthermore, researchers occasionally cited internal designations for progress reports or fund transfers within the government.

Notably, among the nonstandard formats, many were readily identifiable as originating from other government funding agencies, such as the NSF with formats like CMMI/CNS/DMR-XXXXXXX, the AFOSR with formats beginning with FA9550, or even extremely old ARO agreements starting with DAAD or DAAG.

This observation prompts an inquiry into whether ARO award IDs are erroneously mapped to other agencies within the article's information. While this might initially suggest suboptimal performance of the algorithm extracting this data, consider a prevalent scenario illustrated by the following example:

"Supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Agency (IARPA), via U.S. Army Research Office Contract No. W911NF-17-C-0050."

In such instances, the challenge arises in deciding whether to allocate the award ID to ODNI, IARPA, ARO, or a combination of them. In the instance mentioned, the award ID was associated with IARPA, with blank award IDs assigned to the other two agencies. Another instance involves text stating, "This work was also supported by a MURI grant," which was a few lines removed from any reference to ARO, the Funding Agency tied to this grant. Subsequent cleaning addressed

the misallocation of ARO award IDs to other agencies. This involved replacing the ARO award ID with the correct one and eliminating the initially assigned ID, which, in many cases, belonged to entities like NSF, AFOSR, ONR, NASA, DTRA, or was left blank. Figure 67 provides a visual representation of the percentages of award IDs in the correct format, those that are blank, and those that are incorrect. Following this cleaning operation, 84% of award IDs are now in the correct format. It's noteworthy that some blank award IDs have been corrected, as indicated by the decrease in the blank award ID percentage. Incorrect award IDs are further categorized into "Other," representing cases where no ARO award ID was formatted within the article, and cases where the award ID contained W911NF but lacked the correct format concerning the rest of the ID (incorrect length).

Quality of Award IDs, Cleaning #2 Army Research Office

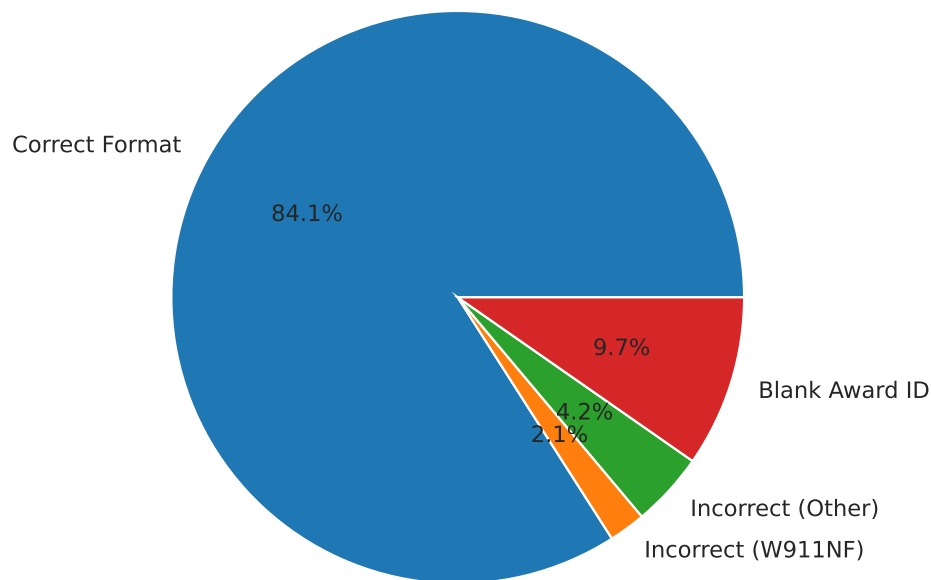


Fig. 67 A pie chart for ARO award IDs within OpenAlex “grants” section from scientific articles published within the 2014–2023 time frame. This snapshot is after applying a second cleaning operation detailed within the text.

5.3.4 Data Cleaning of Award IDs, Part III

Now the award IDs that start with the correct ARO identifier but deviate from the standard award ID format are explored. Upon closer inspection, a recurrent pattern emerges—often, the one-character designation for the type of agreement is absent (e.g., W911NF-XX-0001 instead of W911NF-XX-1-0001), or a zero is missing from the final four digits (e.g., W911NF-XX-X-001 instead of W911NF-XX-X-

0001). In the former case, the absence of the agreement type introduces variability, although grants (1) and cooperative agreements (2) prevail as the most common types. In the latter case, the absence of a zero in, for example, 001, could be erroneously mapped into 0001 or 001X. Though various errors exist, it appears that the most prevalent one is the absence of a zero. To rectify this, a placeholder character was strategically introduced in the correct position, aligning the award ID with the correct format for future reconciliation with lists of actual award IDs. Figure 68 illustrates that a majority of these incorrect award IDs are now categorized within the correct format, increasing the percentage to nearly 86% of ARO awards listed in articles.

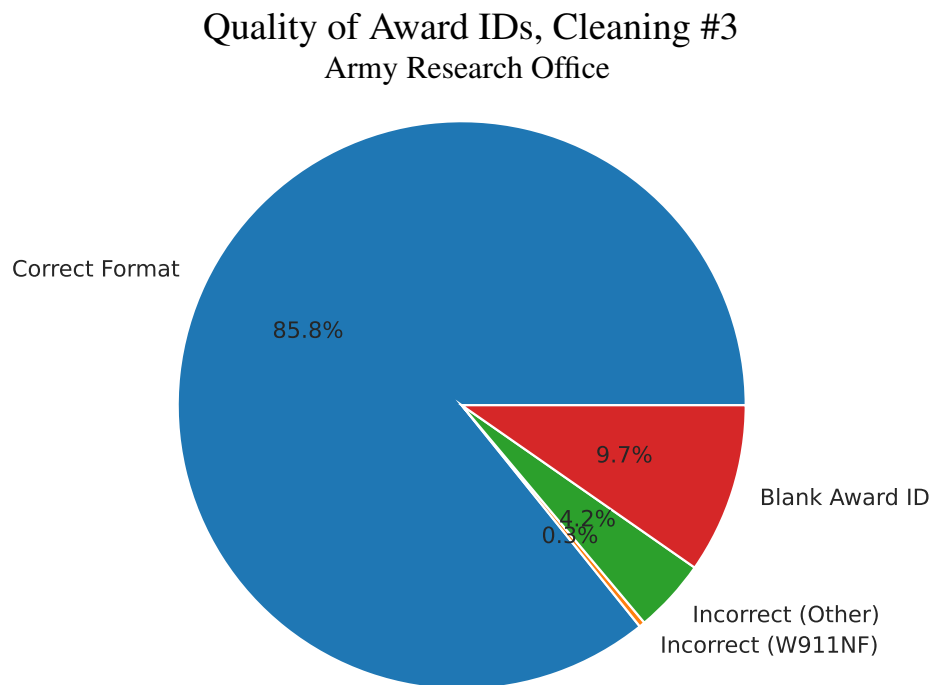


Fig. 68 A pie chart for ARO award IDs within OpenAlex “grants” section from scientific articles published within the 2014–2023 time frame. This snapshot is after applying a third cleaning operation detailed within the text.

Reconciling ARO award IDs in articles becomes challenging at this juncture. The “Other” category, primarily populated with award IDs from other agencies, does not include any ARO-formatted award IDs listed under different agencies. This suggests that while the ARO is likely mentioned as a Funding Agency in the text, no award ID from any agency matches the ARO format. Similarly, the presence of blank award IDs in the dataset suggests that the ARO is mentioned as a Funding Agency in the text, but no award ID from any agency conforms to the ARO format.

This presents a limitation in further aligning ARO award IDs with corresponding agencies or IDs for these entries.

5.3.5 Data Cleaning of Award IDs, Part IV

Continuing the exploration, the award IDs with the correct format can be investigated. Initially, scrutiny was given to the characters representing the year and agreement type to identify any abnormalities. The year parameter exhibited no issues, predominantly falling within the range of 2005 to 2023. However, the agreement type revealed an interesting observation—many award IDs contained the characters “R” or “S,” signifying the broad agency announcement (BAA) or program solicitation, neither of which provides pertinent information about the award. Out of the 19,623 award IDs in the correct format, 74 referenced the BAA, and 34 referenced the solicitation; these were reassigned to a category indicating they are in the correct format but not valid.

Subsequently, the awards underwent comparison with lists of actual award IDs, and a Python dictionary was crafted to rectify award IDs that could be accurately mapped. Additionally, wherever a placeholder character was inserted, it was replaced with the number or letter most commonly occurring award. Following this step, 19,202 award IDs out of the total 19,249 (99.76%) were accurately associated with actual ARO awards. For the 47 award IDs not found, a thorough examination of the most accessible articles was conducted, revealing instances where these IDs matched exactly as listed in the papers.*

Figure 69 presents the final breakdown of award IDs for all articles since 2013 acknowledging the ARO as a Funding Agency that supported their work. Approximately 85% of all award IDs not only adhere to the correct format but have been successfully validated against actual awards.

5.3.6 Frequency of Articles per Award and Awards per Article

With the verification of ARO award IDs as actual awards, the dataset comprises close to 20,000 papers and their corresponding awards. The probability distribution functions for the number of articles produced per award and the number of grants acknowledged per article are illustrated in Fig. 70.

*Remarkably, in one case, it was discovered that the PI consistently used the same incorrect award ID in every paper published until the final report.

Quality of Award IDs, Final Cleaning Army Research Office

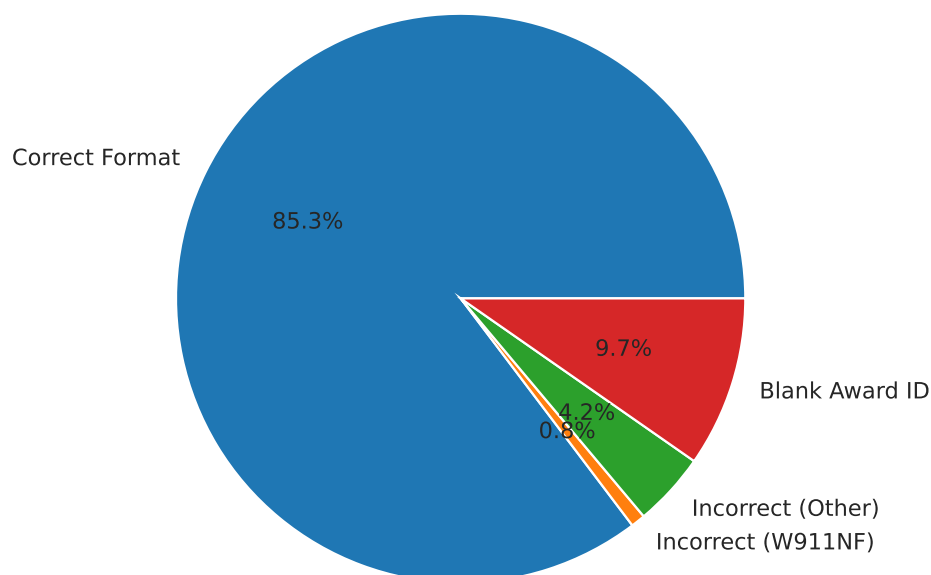
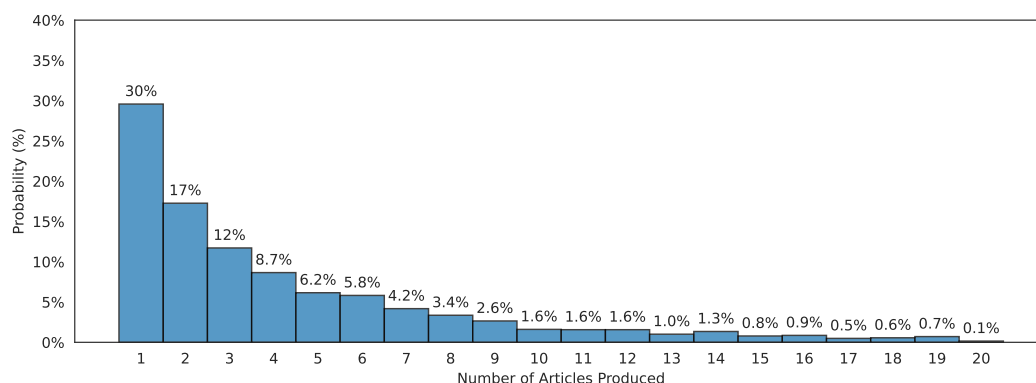


Fig. 69 A pie chart for ARO award IDs within OpenAlex “grants” section from scientific articles published within the 2014–2023 time frame. This snapshot is after the final cleaning and validation stage.

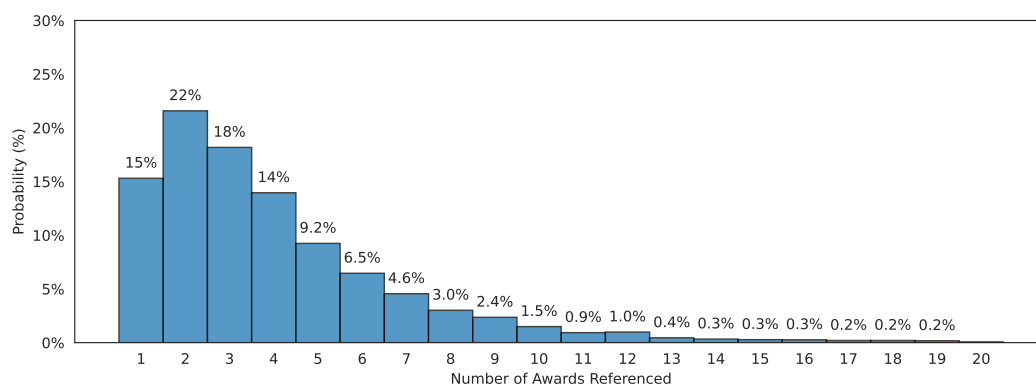
In Fig. 70a, the x axis represents the number of articles acknowledging ARO, each accompanied by a verified award ID. The y axis signifies the percentage of grants producing a specific number of articles; probability distribution function means that the sum of all percentages equals 100%. The highest percentage is attributed to each award producing a single article (30%), followed by each award yielding two articles (17%), and a decreasing trend continues with each additional work per grant.

Similarly, Fig. 70b illustrates the distribution of how many grants are referenced within a single article. The highest percentage occurs when two awards are referenced, accounting for 22% of articles. Only 15% of articles reference funding from a single award, specifically an ARO grant. After reaching a peak at two grants per article, three grants per article is slightly lower (18%) and the percentage monotonically decreases as the number of grants per article increases. The other agencies most highly associated with articles funded by ARO: NSF (37%), AFOSR (13%), ONR (9.8%), DOE (8.8%), Defense Advanced Research Projects Agency (5.9%), National Natural Science Foundation of China (4.9%) and Army Research Laboratory (4.7%). This probability distribution plot underscores that funding from a

Probability Distribution Functions of Article–Award Relationship



(a) Probability distribution function of articles produced (for a single award)



(b) Probability distribution function of awards referenced (in a single article)

Fig. 70 A probability distribution function plot for (a) number of articles produced for each grant, and (b) number of grants referenced for each article

single agency is often interwoven with funding from other agencies in the generation of scientific articles and research discoveries.

5.3.7 Distribution of Awards, Articles and Citations by Award FY

Exploring the dynamics between awards, their respective fiscal years, the resulting number of articles, and the cumulative citations garnered by those articles provides valuable insights. Figure 71 portrays distribution plots that illuminate this relationship over fiscal years in which the awards were funded. Three distinct distributions are presented:

- a) Number of Awards by Fiscal Year: This distribution illustrates the frequency of awards allocated in each fiscal year.

- b) Number of Articles Produced per Award by Award Fiscal Year: Here, the focus shifts to the quantity of articles generated by awards in each fiscal year, shedding light on the productivity of individual awards.
- c) Number of Citations Accumulated by Articles from Awards by Fiscal Year: Examining the citations amassed by articles produced from awards in different fiscal years provides a measure of the scholarly impact over time.

For ease of comparison, the award fiscal year labels are standardized across all sub-figures, facilitating a comprehensive visual assessment of the distributions. Figure 71 serves as a comprehensive visual tool to understand the intricate interplay between award years, article productivity, and citation impact.

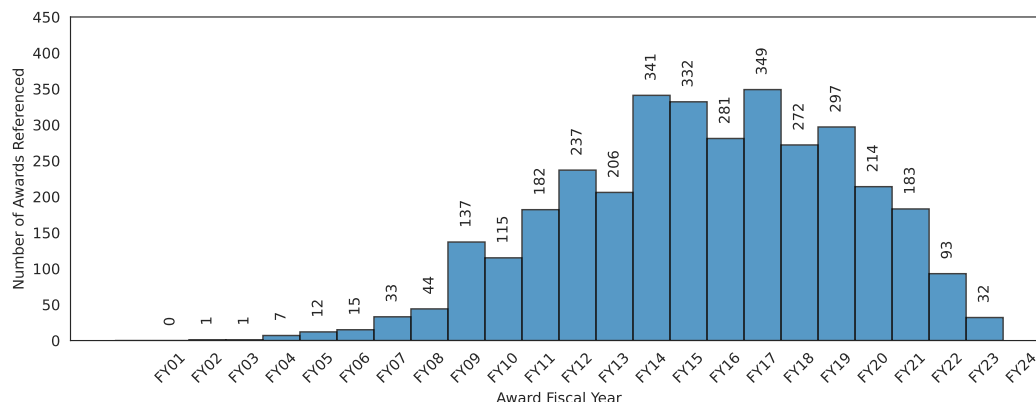
In Fig. 71a, the depiction of the number of awards reveals a gradual increase post-fiscal year 2005, surging above 200 awards from fiscal year 2012 to 2020, followed by a decline in fiscal years 2022 and 2023. It is imperative to interpret this distribution with the caveat that it relies solely on Funding Agency references found in the acknowledgments section of publications. Consequently, awards may not appear in this plot if there were no associated publications or if the articles did not explicitly attribute funding to the ARO. Notably, instances where ARO is acknowledged without a specified award (blank, 10%), or with incorrect award information (4%), merit consideration. In such cases, the responsibility could lie with the authors or may stem from potential limitations in the extraction algorithm. The intricacies of these discrepancies were not explored further to establish causality. The absence of awards in the early to mid-2000s might be attributed to a transition from an older award ID format to the current W911NF format. Similarly, the reduced presence of awards in FY22 and FY23 likely stems from the absence of publications resulting from those awards.

In Fig. 71b, the visualization portrays the number of publications associated with the awards. Notably, the initial x axis labels include entries for articles with blank award IDs ("Blank") and those with an incorrect format for the award ID ("N/A"). Focusing on the period from FY 2011 to FY 2019, there is a consistent production of approximately 1000 to 2500 articles from awards linked to those years. The peak in the number of papers occurs with awards funded in 2014, reaching a maximum of 2,472 articles. Similar to the trends observed in Fig. 71a, recent fiscal years have seen a lower output of articles from awards, indicating that articles stemming from

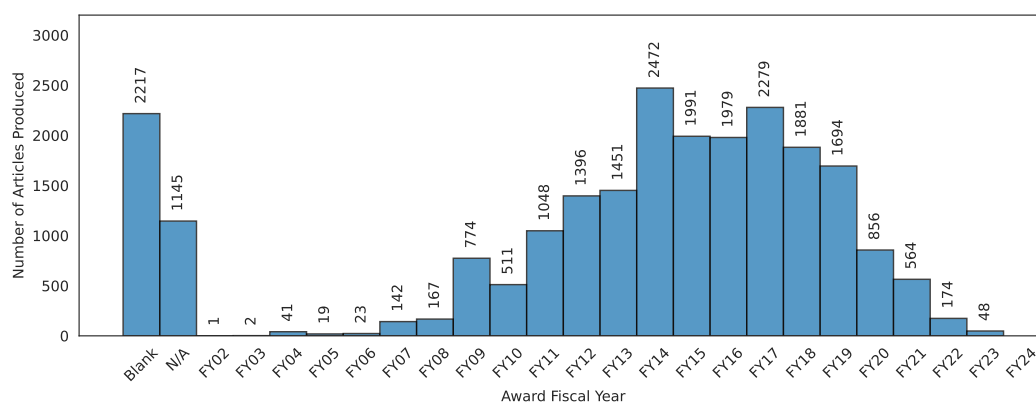
awards are a lagging metric. Analysis of articles with blank award IDs or incorrect formatting reveals their significant contribution to the overall body of work attributing to ARO, as previously depicted in pie charts.

Moving to Fig. 71c, the visualization displays the number of citations linked to papers generated under the awards. The highest number of citations, exceeding 100,000, is associated with awards funded in fiscal year 2014. This metric serves as an additional lagging indicator, reflecting the impact of these awards over time. A notable example is the comparison between Figs. 71b and 71c, revealing that awards funded from fiscal year 2020 have initiated scientific discoveries resulting in over 856 articles. However, these 856 articles have accumulated only 5,060 citations within the research communities. This underscores the temporal aspect of awards, wherein they take time to yield scientific articles, and subsequently, these articles require time to gain recognition within the scientific community and garner citations in the evolving scientific literature.

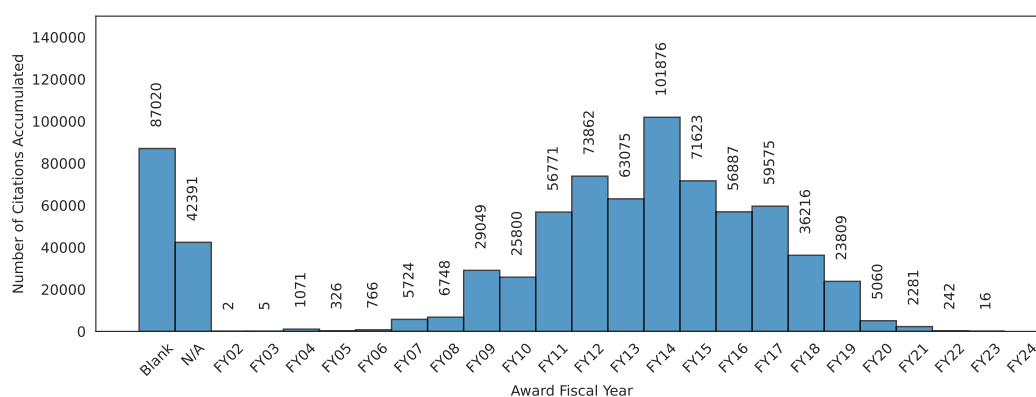
Distribution of Awards, Articles and Citations by Award Fiscal Year



(a) Distribution of number of awards



(b) Distribution of number of articles produced



(c) Distribution of number of article citations

Fig. 71 A distribution plot for (a) the number of awards made per fiscal year, (b) the number of publications grouped by award fiscal year, and (c) the number of citations of those publications grouped by award fiscal year. For all three subfigures, the x axis is the award fiscal year, even if the publications referencing the award or the citations to the publications occurred in later years. Blank award IDs and incorrect format award IDs are also shown to the left.

5.3.8 Distribution of Awards, Articles and Citations by Award Funding Amount

Redirecting the focus to the funding amounts associated with these awards provides insight into the award-article relationship. For this analysis, since articles were restricted to only those after 2013-10-01, only awards that started after this date and that ended before 2023-10-01 were included. Hence, this encompasses a 10-year span of articles and awards. The process for analyzing the data is as follows. Initially, all awards featuring award IDs are correlated with their respective funding amounts and then categorized into intervals outlined in the first column of Table 11. This step provides an overview of the correlation between award amounts and the outcomes of these awards. Each bin is chosen to encompass more than 50 awards. Table 11 encapsulates the average funding amount, average duration (in years), the count of awards, the number of articles, and the number of citations for each bin. In most cases, the average funding amount aligns closely with the midpoint between the two bracketing amounts of the bin.

Table 11 Number of ARO awards, articles, and citations after binning awards by their funding amounts, including the average funding amount, length of time, and funding per year

Funding Bin	Average Funding	Average Length (yr)	Funding per Year	Number of Awards	Number of Articles	Number of Citations
\$4k–\$50k	\$34,865	0.89	\$39,165	51	103	1,840
\$50k–\$100k	\$61,507	0.94	\$65,135	167	434	12,779
\$100k–\$200k	\$151,314	1.72	\$88,100	137	521	13,995
\$200k–\$300k	\$245,380	2.06	\$119,248	123	465	7,997
\$300k–\$400k	\$359,921	3.11	\$115,648	336	1,828	47,548
\$400k–\$500k	\$448,223	3.25	\$138,077	275	1,653	45,724
\$500k–\$600k	\$556,741	3.17	\$175,693	154	1,042	23,560
\$600k–\$750k	\$655,628	3.42	\$191,981	93	586	16,019
\$750k–\$1M	\$871,694	3.46	\$252,078	81	480	11,515
\$1M–\$1.5M	\$1,162,692	3.41	\$340,593	51	402	7,503
\$1.5M–\$5M	\$2,571,590	3.89	\$660,613	55	429	12,219
\$5M+	\$10,049,380	4.90	\$2,049,253	51	1,449	48,967

As the funding amount increases across the bins, there is a corresponding rise in the average award duration. For awards less than \$100K, the average duration is approximately 1 year; within the \$300k to \$600k range, the average duration extends to about 3 years; and for awards surpassing this, the duration starts to increase to four years or more. Unsurprisingly, the funding per year exhibits a monotonic increase with the size of the award. Notably, the highest concentration of awards lies within the \$300k to \$500k range, totaling 611 awards, akin to the scale of a single

PI grant. This range aligns with the highest count of articles and citations among awards under \$5 million, based on the selected bins. In the bin above \$5 million, the average funding amount experiences a significant surge, surpassing \$10M on average, and spanning almost five years. This category yields over 1,400 articles that have accumulated approximately 49,000 citations. Awards in this range are likely associated with MURIs or university-affiliated research centers (UARCs). The number of articles and citations for these are large compared to other categories except those within the \$300k to \$500k, but the funding amounts are drastically different between these ranges, making a normalized-by-funding-amount comparison a more fair comparison.

5.3.9 Distribution of Cost per Article and Cost per Citation by Award Funding Amount

Delving into the funding amounts tied to these awards naturally raises questions about the average cost per article and citation. Table 12 utilizes the same bins to present averages for the number of articles per award, citations per award, citations per article, cost per article, and cost per citation. As expected, with the escalation of funding and project duration, the number of articles produced per award witnesses an increase. Interestingly, the number of citations per award exhibits a growth trend with escalating award amounts, reaching a peak in the \$400k to \$600k funding range, then decreasing some until over \$5 million funding range. The citations per article displays a similar trend, except the first peak comes around the \$300k to \$500k funding range.

While spending on awards over \$500k increases, the number of articles and citations does not undergo a significant boost beyond the \$300k to \$600k award range. However, a notable increase occurs in the cost per article and cost per citation, as evident in the last two columns of Table 12. Notably, these costs continue to rise with the size of the award. This escalation could stem from various factors such as a shift from single PI to multiple PIs, heightened administrative overhead for larger sums, a transition from scientific discovery to science development, an expanded organizational structure, a need for more senior scientific staff, or increased oversight. The changing nature of the scientific goals as the dollar amount increases, coupled with heightened scrutiny on larger funding amounts (increased oversight, program reviews, etc.), may contribute to these trends. The type of program associated with the award also undergoes a transformation as funding amounts increase,

Table 12 Cost of articles and citations after binning awards by their funding amounts, including number of articles per award, number of citations per award, and the citations per article (averages)

Funding Bin	Articles per Award	Citations per Award	Citations per Article	Cost per Article	Cost per Citation
\$4k–\$50k	2.0	36.1	17.9	\$17,263	\$966
\$50k–\$100k	2.6	76.5	29.4	\$23,667	\$803
\$100k–\$200k	3.8	102.2	26.9	\$39,789	\$1,481
\$200k–\$300k	3.8	65.0	17.2	\$64,907	\$3,774
\$300k–\$400k	5.4	141.5	26.0	\$66,156	\$2,543
\$400k–\$500k	6.0	166.3	27.7	\$74,568	\$2,695
\$500k–\$600k	6.8	153.0	22.6	\$82,282	\$3,639
\$600k–\$750k	6.3	172.2	27.3	\$104,050	\$3,806
\$750k–\$1M	5.9	142.2	24.0	\$147,098	\$6,131
\$1M–\$1.5M	7.9	147.1	18.7	\$147,505	\$7,903
\$1.5M–\$5M	7.8	222.2	28.5	\$329,691	\$11,575
\$5M+	28.4	960.1	33.8	\$353,704	\$10,466

with smaller amounts typically tied to instrumentation awards, young investigator awards, and single PI awards, in contrast to multiple PI awards, MURIs, centers of excellence, or UARCs for larger amounts.

While the findings in Tables 11 and 12 suggest that lower funding amounts yield more articles and citations on a cost basis, several considerations temper this interpretation. First, smaller funding amounts for shorter durations often represent incremental, near-term efforts compared to larger single PI grants. Faculty members might utilize a single PI grant to support a PhD student, whereas smaller funding amounts might sustain a postdoctoral researcher—someone more experienced but with a more limited time frame in their position.

Second, articles frequently attribute funding to multiple awards, implying that the listed funding may be a minority share within a more extensive effort funded by various sources. This scenario might create an illusion of funds going further when, in reality, they are combined with other funding sources.

Last, and perhaps most crucially, the presented funding amounts are specific to awards resulting in scientific articles. Some award amounts may never lead to a publication, and this is more likely for smaller funding amounts. It is highly improbable that higher funding amount awards never result in a single article (unless the work is more sensitive). Therefore, what is presented may only be a fraction of

the total funded projects. Further exploration is essential to comprehensively understand these effects and their implications on the observed trends.

The focus of the analysis so far has primarily been on average statistics in understanding the relationship between awards and articles. However, delving into the distribution within funding range bins revealed the presence of outliers—instances of an unusually high number of articles for a small funding amount or a low number of articles for a large funding amount. These outliers have the potential to distort the averages presented in Table 12. To address this, Table 13 employs median values for articles, citations, and funding amounts to provide a more robust calculation. The resulting data indicate a decrease in both articles per award and citations per award across all funding ranges, suggesting the influence of certain awards skewing the averages. Notably, the cost per article within the single PI funding range is now in the range of \$90k–\$120k, translating to approximately 3–4 articles for a \$360k grant. This cost per article experiences a substantial increase—by a factor of 3 or more (>\$370,000/article)—for much larger funding amounts (\$1M or greater). As discussed earlier, various factors could contribute to this escalation, encompassing administrative overhead, a shift towards more applied research, or a move towards team science methodologies.

Table 13 Cost of articles and citations after binning awards by their funding amounts, including number of articles per award, number of citations per award, and the citations per article (medians)

Funding Bin	Articles per Award	Citations per Award	Citations per Article	Cost per Article	Cost per Citation
\$4k–\$50k	1.0	20.0	20.0	\$38,819	\$1,940
\$50k–\$100k	2.0	18.0	9.0	\$30,000	\$3,333
\$100k–\$200k	2.0	27.0	13.5	\$75,000	\$5,555
\$200k–\$300k	2.0	30.0	15.0	\$120,022	\$8,001
\$300k–\$400k	4.0	47.0	11.8	\$90,000	\$7,659
\$400k–\$500k	5.0	57.0	11.4	\$90,000	\$7,894
\$500k–\$600k	4.0	45.0	11.2	\$139,421	\$12,392
\$600k–\$750k	4.0	73.0	18.2	\$162,500	\$8,904
\$750k–\$1M	4.0	53.0	13.2	\$217,000	\$16,377
\$1M–\$1.5M	3.0	42.0	14.0	\$373,567	\$26,683
\$1.5M–\$5M	3.0	39.0	13.0	\$841,358	\$64,719
\$5M+	19.0	378.0	19.9	\$375,928	\$18,895

The awards within the funding range band exceeding \$5 million was further examined to unravel the dynamics at the upper echelons of the funding spectrum. Notably, some awards surpassing \$20 million were administered by ARO, although

the funding originated from other agencies such as DARPA and IARPA. Despite their substantial funding, a subset of these awards resulted in minimal open-access publications, often fewer than 5. It's crucial to recognize that the objectives of such high-value awards may not invariably align with scientific discovery and publication within the open community. It's imperative to bear in mind that the values presented in Tables 12 and 13 are contingent on the data incorporated in the analysis, and variations in included information will impact these outcomes.

Strategic acknowledgment. Briefly revisiting the outliers identified in the analysis helps to understand the anomalies in the award-article relationships. In the quest to pinpoint outliers within the dataset, a linear model was applied to the funding-to-article relationship (x-to-y). The largest mean squared error between the predicted and actual numbers of articles was then used to flag abnormalities. Notably, one particularly egregious outlier featured a PI who secured \$585,000 in funding in 2018, producing a staggering 174 articles at a cost of just over \$3,000 per article. A closer examination reveals that around 60% of these articles, i.e., 105 publications were generated during the project's performance period of three years, with an additional 69 articles emerging after the award concluded in 2021. Another PI had received an award five years prior, amounting to \$372,000, yielding 72 articles—of which 23 were published within the period of performance and 49 articles after the program ended (in some cases almost 6 years later and still acknowledging in articles).

These cases prompt the question of whether such behavior actually incentivizes follow-on awards or whether the PIs just think that it does. If indeed there is an incentive, one might anticipate more instances of this nature within the scientific community. While these particular cases stand out prominently in the ARO dataset, it's prudent to acknowledge that similar behaviors could occur on a smaller scale among other researchers. As the scientific community grapples with evaluating scientists in the era of hyper-authorship, where articles boast hundreds of authors, concerns also emerge regarding the awards funding this work. How should funding agencies navigate prolific senior authors who find themselves on hundreds of papers, acknowledging funding agencies each time? From the authors' perspective, this strategy appears strategic, showcasing high productivity from a scientific standpoint and a notable return on investment for program managers. In fact, multiple awards held by a more senior PI combined with this strategy could be used to show that all programs yield a high return on investment (as illustrated by the case

above, where ARO and NSF were credited for funding each publication).

Where does this leave early career scientists, though? The current publish-or-perish system of rewarding prolific output may inadvertently exclude another vital group of scientists: early career researchers. These individuals might engage in less of the observed behavior, not due to a lack of incentive, but rather as they are in the process of establishing their standing within the scientific community. While they might not churn out 174 articles on a single PI grant, their approach provides program managers with a clearer understanding of the specific work that the research funding supports. This distinction sheds light on the challenges faced by early career scientists in navigating the current landscape of research funding and recognition.

5.3.10 Discussion

In the early stages of funding, the correlation between funding amounts and the number of publications presented a straightforward analysis, resulting in robust linear relationships. However, the contemporary funding environment is notably more intricate. Major funding agencies now operate through diverse funding mechanisms, each designed with distinct objectives, selection criteria, and funding conditions. Analyzing these intricacies at the level of the funding agency proves challenging due to the diversity of stated objectives. If the sole objective were to generate scientific articles, the landscape might present a more level playing field. However, this is far from the only goal. Untangling the factors that contribute to variations in publication patterns within identical funding mechanisms demands a nuanced and detailed analysis, surpassing the scope of this study. The essential exploration of such nuances relies on comprehensive data linking grants and publications, a crucial dimension yet to be fully integrated into the current analyses.

It's crucial to note that establishing the link between funding and articles can be approached through various methods, including examining the acknowledgments section (as undertaken in this study), reporting by PIs to funding agencies, exploring open science repositories (e.g., OpenAlex, tailored specifically for funding), or directly communicating with the PI. Each method has its distinct advantages and drawbacks. For acknowledgments, a potential concern arises in the form of strategic acknowledgment, where authors may align acknowledgments with their personal agendas, potentially acknowledging high-status funding bodies more frequently than warranted by the actual contribution of that funding or strategically

leaving low-status funding bodies off of the acknowledgments. Take, for instance, the strategic acknowledgment of NIH and NSF, which predominantly fund basic science—authors may strategically acknowledge them to bolster their reputation concerning these agencies, which results in more funding awards, producing more articles and bolstering reputation in a repeating cycle (i.e., the so-called Matthew “rich get richer” effect). Notably, since October 2018, over 50% of articles acknowledging ARO funding also acknowledge NSF funding. This statistic implies multiple valid reasons for acknowledging multiple awards, beyond strategic considerations. Authors might utilize funds from different agencies to complement their research, collaborate with others funded by different agencies, or grapple with attributing their intellectual contributions to a single award. In essence, the assumption of a straightforward, unique link from a single award to a single article is often a fallacy, given the complexity of research collaborations and resource utilization.

In handling incorrect award IDs and linking them to the correct ones, the process can be refined and automated further with various similarity metrics. While expert-driven processes were predominantly employed in the current analysis, automating the procedure could greatly benefit from quantifying the similarity between incorrect and correct IDs using similarity/distance metrics. Several similarity metrics are applicable to text, such as award IDs. A straightforward metric is the Hamming distance, which assesses differences in corresponding characters for strings of equal length. However, due to varying character lengths in award IDs, Hamming distance may not be suitable in many instances. An alternative is the Jaro distance, offering a quantitative measure of similarity that accounts for matching characters, transpositions, and their positions within the strings. Transpositions involve the swapping of adjacent characters, enhancing the accuracy of similarity assessments in string matching algorithms. The Levenshtein distance stands out as one of the most effective metrics, incorporating character insertion, deletion, and substitution. An extension, the Damerau-Levenshtein distance, includes transpositions. All these metrics generate a score between zero and one, with zero indicating no similarity and one indicating an exact match. Leveraging these scores can aid in automated processes to identify the correct award ID or narrow down the possibilities in a systematic manner.

5.4 Concluding Remarks: Insights and Future Research Directions

This study delves into the intricate relationship between scientific publications and the funding agencies supporting them, with a particular focus on the ARO. Acknowledgment sections in scientific papers, expressing gratitude to funding entities, provide a rich source of information. The analysis aims to uncover patterns and trends that inform the broader scientific community. Despite the challenges posed by real-world data constraints, particularly in data cleaning and variability in PIs' information quality, this study navigates through complexities to extract various insights. The ARO, as a pivotal supporter of basic research, becomes a focal point for an in-depth exploration, shedding light on the nuanced interplay between funding agencies and the scientific outputs of their awards. While recognizing the pragmatic constraints, this study details its methodology, aiming to contribute a balanced perspective on this relationship. Notably, this study provides detailed insights into some of the challenges encountered during the analysis, with a focus on the methodology employed to overcome those, drawing attention to the transformation from acknowledgment text to structured grants data within OpenAlex (see Fig. 64).

1. **Challenges in Award IDs:** The examination of funding agencies and their awards is instrumental in gaining a deeper understanding of how these awards contribute to scientific articles and subsequent citations, exemplifying their impact on the scientific community. However, within the top 100 cited articles funded by ARO, nine encountered various issues with their award IDs, indicative of broader challenges within the 19,582 articles acknowledging ARO's support. These challenges encompassed instances of "no award ID listed," "listing proposal ID instead," "incorrect formatting for the award ID" (involving missing or multiple characters), and "listed correctly but encountering extraction algorithm failures." During the initial analysis of the complete dataset, it was found that 74% of award IDs adhered to the correct format, 10.9% lacked any award ID, and 15.1% presented a nonstandard ARO award ID format.
2. **Data Refinement Process:** This study engaged in a data cleaning endeavor, documented to address and rectify issues present in the dataset, with a particular focus on mitigating the 15% of incorrect formats. Noteworthy discrepancies emerged during this process, including numerous instances of incor-

rect starting characters in award IDs that commence with "W911NF," such as insufficient characters, the absence of "911," having the right number of characters but the wrong letters, and an excess of characters. Notably, certain errors were attributable to factors beyond the articles themselves, where optical character recognition occasionally misinterpreted "1" as "T" or "W" as "VV." While these discrepancies are manageable on a manual level, devising more efficient methods becomes imperative for larger datasets. Another facet of the cleaning process unearthed situations where ARO award IDs were erroneously linked with other funding agencies. Conversely, there were instances where the ARO award ID itself incorporated references to NSF, AFOSR, or ONR award IDs. Additionally, even when the award ID was in the correct format, PIs would make references to the BAA or the funding solicitation announcement. A potential enhancement lies in cross-referencing IDs with comprehensive lists of known awards and their corresponding IDs, facilitating the mapping of awards with missing characters to the accurate award ID.

3. **In-Depth Analysis:** A substantial portion of the effort centered on the meticulous refinement of award IDs within the open dataset. This refined dataset was subsequently integrated with funding amounts and the period of performance length to delve into comprehensive award outcomes, including the number of articles per award and the number of citations per award. The analytical exploration unveiled intriguing patterns. The probability distribution for the number of articles per award showcased a notable peak at one, with a monotonic decrease in frequency as the count of articles per award increased (Fig. 70a). Additionally, investigating the number of awards referenced by a single article revealed a distinct pattern, with the highest probability for two awards, followed by three, and then one—underscoring the prevalence of articles supported by multiple awards (Fig. 70b). Further insights were derived from subsequent plots (Fig. 71), unraveling the intricate relationships among awards, articles, and citations concerning the fiscal year of the award. These plots illuminate a lag in the performance of articles and citations compared to awards. Lastly, the awards were binned based on funding amount, offering a nuanced perspective on how the number of articles and citations varies with award funding. This analysis even delves into the breakdown of the cost per article for the dataset under examination (Table 11 and Table 12).

In light of the challenges posed by the identified errors, the question arises: what insights can be gleaned to enhance the extraction of funding information from articles?

One insight is that scaling these analyses to broader applicability demands a concerted effort in automating the cleaning processes employed herein. This is crucial to minimize noise in analyses of funding agencies. Another insight is that a robust solution may lie in reimagining the current system. What if funding agencies could share their lists of award IDs and metadata through an accessible API? Currently, agencies rely on varied dashboards or spreadsheets, which can hinder accessibility. However, by offering easily-accessible information, funding agencies pave the way for publishers to act as a crucial filter. Publishers can verify the presence of both the funding agency and award ID, cross-referencing against the agency's API. This approach would efficiently screen acknowledgments lacking award IDs or containing errors like misspellings, missing characters, or references to internal documents. Standardized acknowledgment statements would further simplify natural language processing extraction from acknowledgment statements, facilitating seamless linkages between funding agencies, articles, authors, and their impact. In some ways, this is merely an extension of CrossRef's Open Funder Registry (OFR, <http://crossref.org/fundref/>) which matches funding agencies with unique IDs. As this study shows, agencies is only part of the award-to-article link and some ability to incorporate valid award IDs would be a substantial improvement.

The benefits of this proposed system extend to funding agencies, publishers, and the scientific community at large. For funding agencies, instant access to award-related articles, regardless of PI submissions, becomes possible. The system incentivizes PIs to accurately include information, fostering accountability for both authors and funding agencies. The alternative, attempting to identify every author associated with an award and match incorrectly formatted award IDs, proves intricate. Many multiple PI awards lack authors beyond the proposal documentation, demanding extensive efforts to uncover and piece together award ID information. Therefore, implementing this recommended system emerges as a forward-looking solution for streamlining and enhancing the integrity of funding information across scientific publications.

6. Courses: Machine Learning and Artificial Intelligence 101

6.1 Machine Learning 101 Short Course

What started in 2020 as a series of seminars on the basic concepts of ML has evolved over the past few years into a 2-day course (about 15 hours in total with over 500 slides) covering a range of topics on ML. In 2023, the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) hosted the third “Machine Learning for Everyone: May The Fourth Be With You” course virtually. While the course incorporated pop culture references to Star Wars, Marvel, Spaceballs, Star Trek, Austin Powers, National Lampoon’s Christmas Vacation, Futurama, and others to maintain attention spans, it also imparted a substantial amount of information on ML topics. The concepts ranged from linear regression to traditional neural networks to more complex architectures like convolutional neural networks for images and sequence models used within large language models (i.e., the concepts behind generative pre-trained transformers [GPT]).

Over the last several years, the short course has attracted over 2500 participants registering from more than 300 distinct institutions worldwide, including academia, industry, and government agencies. The pre-course surveys asked registrants about their familiarity with ML; a majority of the participants classified their familiarity with ML as slightly familiar, somewhat familiar, and not at all familiar, as shown in the tree map in Fig. 72. The captured audience represents the envisioned target audience. The target audience was not experts in ML but rather individuals beginning their journey. Unlike other introductory courses, this course demonstrated the algorithms and mathematics behind ML using a platform familiar to virtually everyone: Microsoft Excel.

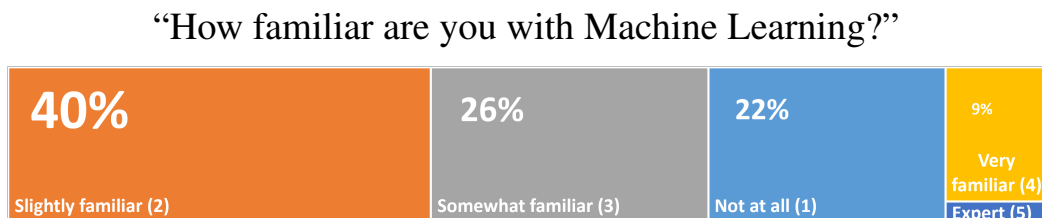


Fig. 72 A tree map showing the breakdown of participant pre-course responses for “How familiar are you with machine learning coming into the course?” The options were on a scale of 1 to 5 with text explanation given as well, i.e., not at all (1), slightly familiar (2), somewhat familiar (3), very familiar (4), and expert familiarity (5). The size of the different regions within the tree map is proportional to the number of participants that answered as such.

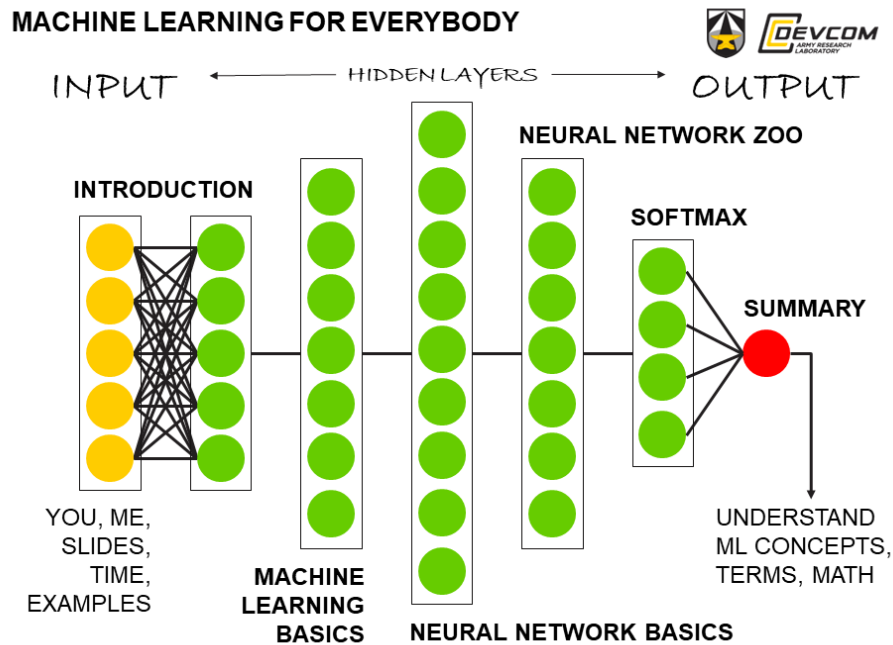
[illegible]

Fig. 74 A word map showing the words within organizations for course participants

others (autoencoders and generative adversarial networks [GANs]), etc. While there is a tremendous amount of information, it has been streamlined to cover only those basic concepts essential to understanding neural networks and how they operate—linear algebra, matrix multiplication, linear regression, logistic regression, gradient descent, etc.—at the expense of other ML models like gradient-boosted machines, decision trees, support vector machines, and other statistical learning models. Over the years, the more complex neural networks section has evolved into a full day, discussing topics to include AI successes and failures (ML201), feature engineering (ML201), autoencoders and GANs (ML301), convolutional neural networks (ML301), sequence models (ML301), and added this year, natural language processing (ML301). The word map shown in Fig. 75 was the initial word map from year 1 of the course which serves as a roadmap of concepts, which are crossed off as the course evolves.

Natural language processing ML models have been publicized greatly since the emergence of OpenAI’s ChatGPT in November 2022. In prior versions of the ML course, the concepts behind large language models were presented in the Sequence Models section, which discussed how recurrent neural networks, gated recurrent unit, and long-/short-term memory models worked. In this manner, this section was agnostic to whether the sequence was a sequence of words or a sequence of signal readings from an internet-of-things device. This was purposely done because describing how large language models take words and map them into an embedding

Machine Learning for Everyone: May The Fourth Be With You 2023



(a) ML 101 course outline

ESSENTIAL MACHINE LEARNING TOPICS

I started with a list of what I thought were the most important topics, like...

DOES $y = mx + b$ OR DOES $y = f(Wx + b)$?

Machine Learning & AI Definition
 Steepest/Gradient Descent
Classification vs Regression
 Linear Regression
 Logistic Regression
 Loss and Cost Functions
Variance vs Bias Tradeoff
 Training, Dev, Test Sets
 LOOCV, K-fold Cross Validation
 Optimization
 Gaussian Process Regression
 Kernel (Ridge) Regression
 Radial Basis Function
Support Vector Machines
 L1/L2 Regularization
 Ridge, LASSO, ElasticNet Regression

Sigmoid
ReLU
RMSProp
MaxPool
VGG-16, VGG-19
Inception
RNN
GRU
LSTM
 One-to-Many
 Many-to-One

W
 layers
 nodes
 One-hot
 Softmax
 Adam Optimization
 Convolution filter
 LeNet
 Vanishing/Exploding Gradients
ResNet
 Convolutional Neural Networks
 Initialization
 Autoencoder
 Natural Language Processing
Sequence Models

b
Neural Networks
 Deep Learning
Activation Functions
 Feed Forward
 Dropout
 Regularization
 Momentum
Hyperparameters
 Cross-Entropy, KL Divergence
 ML Platforms
 Backpropagation
 Python Toolkits
 Transfer Learning

(b) Word map of ML concepts

Fig. 75 Two slides from the ML 101 course are shown: a) the outline of the short course, and b) a word map of essential ML concepts taught within the course

space whereby the embedding vectors are consumed by these different sequence models would have taken longer to explain and added considerable complexity. In May 2023, the natural language processing section was added. A few slides from this addition are shown in Fig. 76. For instance, this section elaborates on the concept of an embedding space and embedding vectors and what they mean. The often-cited GloVe (Global Vectors for Word Representation) embedding vector dictionary was added into Excel (the 50-dimensional version) to show the relationships between words and their embedding vectors. Figure 76a shows a colored version* of several words like skywalker, jedi, vader, and darth. The GloVe embeddings are produced using a probabilistic co-occurrence matrix, which has some disadvantages.† The Word2Vec embedding model as a neural network-produced embedding space is then discussed (as shown in Fig. 76b), along with how the input and output has to be structured to be used in a neural network architecture (i.e., the skipgram technique and the continuous bag of words [CBOW] technique). This introduction to natural language processing stops just short of teaching transformers for both time limitations and the increased complexity of these architectures.

The course participants were asked to give anonymous course feedback, both quantitative (scale of 1 to 5, strongly disagree to strongly agree) on several statements and qualitative (“What did you think of the course? Anything else that you wish to add?”). The May The Fourth 2023 feedback is very similar to feedback from past courses and even an improvement over the May The Fourth 2022. This year’s course improved in every evaluation criterion; the eight course evaluation statements are shown in Fig. 77. Highest improvement was in “my overall evaluation of the course is excellent,” which was a huge jump from prior courses (over 10%). The changes since last version are the improved content (for understanding), the additional of natural language processing and discussion of current AI topics, and the overall flow/organization of the course. The lowest score was still “the course gave me confidence to do more advanced learning,” which can be expected—many participants come in knowing very little about ML, so this can even be an encouraging score given that the majority still strongly agree (i.e., score of 5) with the statement.

*i.e., conditional formatting in Excel was used to color the values of the 50-D embedding vector

†For example, the antonyms “good” and “bad” can often be used interchangeably within a sentence, e.g., “the restaurant has good ratings” or “the restaurant has bad ratings.” Hence the embedding vectors will be similar. Another disadvantage is that the embedding vectors for GloVe lack context surrounding the words. The musical artist “Prince” has the same embedding vector as “prince,” the son of a king, even though the word has different meanings in different contexts.

Machine Learning for Everyone: May The Fourth Be With You 2023

NATURAL LANGUAGE PROCESSING GLOBAL VECTOR ... GLOVE



"skywalker"

[0.81037, -0.79427, -0.30319, 0.27545, 0.51568, -0.63509, 0.054371, 0.025869, -0.45726, -0.067687, 0.21436, 1.1255, 0.038657, 0.080528, -0.49525, 1.1644, 0.27232, -0.015318, -0.19989, 1.375, -1.0267, 0.87457, -0.89811, 0.31633, 0.86519, -0.043719, -0.28484, -0.64548, 0.075151, -0.75939, -0.19859, -0.49455, 0.33003, 0.98297, 1.4413, 0.42309, -0.5735, -1.3914, -0.47725, -0.86097, -0.33381, 0.85012, -0.52006, -0.77692, -0.16556, -0.0049513, 2.1455, -1.4426, -1.6192, 0.54356]

"skywalker"



"jedi"



"vader"



"darth"

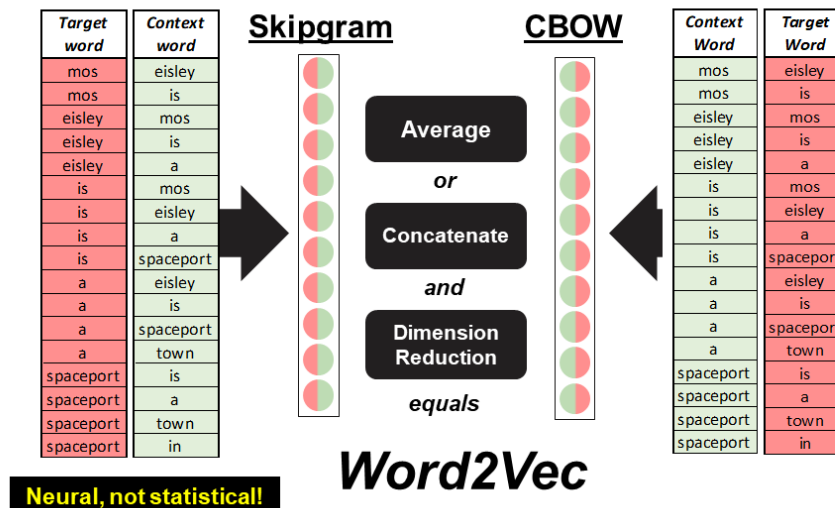


(a) GloVe and Embedding Vectors

NATURAL LANGUAGE PROCESSING HOW ARE THEY TRAINED? CBOW & SKIPGRAM



What if we approached this like a neural network?



(b) Word2Vec and Embedding Vectors

Fig. 76 Two slides from the ML 101 course's section on natural language processing showing a) the GloVe and embedding vector representations for various words, and b) the use of skipgram and continuous bag of words for training a Word2Vec model

Anonymous Course Ratings Machine Learning for Everyone 2023

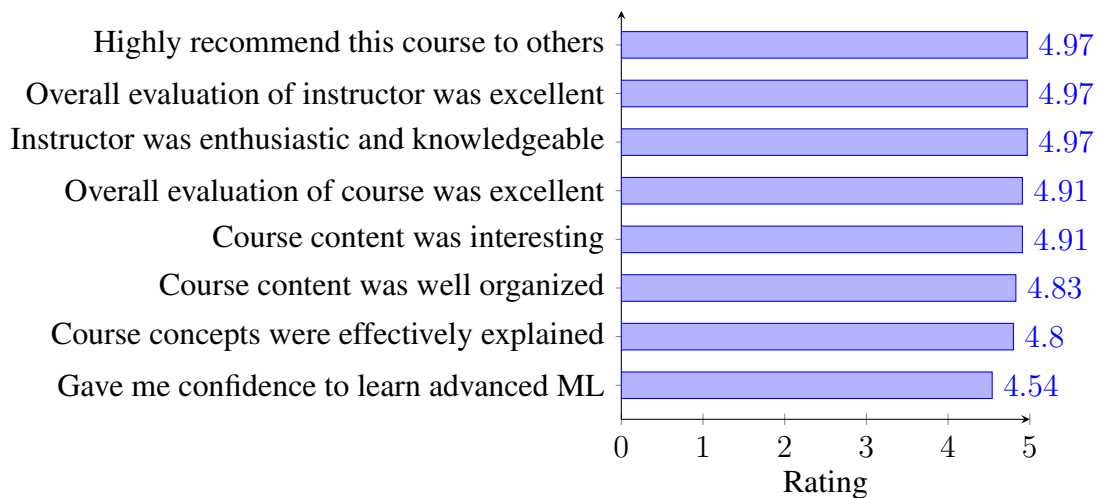


Fig. 77 Anonymous course ratings for the Machine Learning for Everyone 2023 short course for different statements from 1 (strongly disagree) to 5 (strongly agree). The paraphrased statements that the participants were asked to evaluate are listed in the horizontal bar chart along with the corresponding average value (out of 5).

For “What did you think of the course? Anything else that you wish to add?” section, there were quite a few comments. In addition to the ratings, the comments reflect that the course participants both learned about machine learning and its concepts in addition to enjoying the instructor and course structure. Some comments reflect opportunities to improve the course, which has resulted in increased ratings over time. Here are just some of the (hundreds of) comments:

1. Quite frankly, it was the best and most entertaining presentation on this subject I have ever experienced. Terrific job.
2. This course was the best ML/AI I have ever taken. I do not know anything about Star Wars, but the funny captions, comments and pictures made the training even more enjoyable.
3. The content was intriguing and very eye catchy. The instructor was excellent teacher. The concept was well-presented, very easy to understand at the surface level. Started with literally no background or training in ML, I have learned a lot more than I would ever had thought I can be. Thanks again for the great content and thanks for making me (generalized as someone who doesn’t know anything about ML) understand and stay focused throughout

the whole course. Thanks!

4. The course tackled an astronomical amount of information, collectively known as machine learning, with well-articulated simplicity and clarity. This one-day course seamlessly presented an overview of ML topics with stimulating visuals and graphics that would normally take hundreds of textbook pages, weeks-long online courses, and entire college semesters to cover!
5. This was so great! If I had to give it a score, I would give it a 10 out of 9. Thank you for putting this material out there and making it approachable for people who don't have backgrounds in linear algebra or other advanced math. The pacing of the material was great and sending out the excel spreadsheets in advance was highly conducive to helping me follow along during the lesson. [...] Thanks for making something I once thought was unattainable actually seem much more accessible!
6. Excellent training! Kept me engaged. Mark has renewed my long-forgotten love for math. I can't wait to learn more about ML!
7. As a self-proclaimed Star Wars nerd, I was really looking forward to the Star Wars-themed machine learning course. I absolutely loved the way the memes/quotes/references were incorporated in a way that both held my attention throughout the course and made learning the concepts infinitely more enjoyable. Additionally, the instructor's enthusiasm was very contagious! As for the material itself, I really appreciated the way it was presented—especially the Excel-based examples. Often times, machine learning is presented as something that requires tons of specialized programming knowledge which makes learning the basics a somewhat intimidating undertaking. It was refreshing to see excel being used for the smaller data sets with an explanation of how python, MATLAB, etc. is needed to larger data sets. To me, that makes the content a lot more relatable. [...] It was a really clever way to put the content in real-world context.
8. There was a lot of math for someone who never took linear algebra, but the very thoughtful and fun Star Wars theme MORE than made up for it. As a new hire in a DOD research lab, this course should be used as a model for how to teach non-experts about your field – I would attend these all the time if they were all half this good!

9. Truly excellent session! You found the perfect balance between not enough math/theory and too much. Goldilocks level of perfection. I really enjoyed the Star Wars theme too - it made for great levity additions.
10. From the perspective of someone who is just starting to dive into the world of machine learning, I thought this course was great! Mark was knowledgeable, well-organized, and clearly enthusiastic about the material. I appreciate the use of Star Wars to help keep things fun and engaging! I also liked that Mark chose to demonstrate neural network building in Excel rather than Python or MATLAB. It was an intuitive way to visualize the process without getting bogged down by syntax and esoteric functions, even if Excel would not be the most practical way to develop neural networks for a real application. [...] Overall, though, this was a great experience, and I would definitely recommend this course to others!
11. Absolutely awesome. The instructor is a “natural teacher,” someone who is just born to teach. [...] Truly awesome talent, presentation, pedagogy, everything.
12. Wow! Course was fantastic. I was pretty intimidated by the whole idea of machine learning beforehand, but this primer really helped take the edge off. I’m the kind of person who needs to be able to look under the hood of tools before I’m comfortable using them and the way Mark unpacked the equations and approaches used in the various methods was great!
13. Overall, a fantastic job at this course, but I feel like it needed more Star Wars references (joking... maybe). I now think that Excel is probably the best/most informative way to represent these algorithms on the smaller scale. I think it improved how easily I understood the concepts.

6.2 One-Minute Machine Learning: The Book

In parallel with the development and delivery of the “Machine Learning for Everyone: May The Fourth Be With You” course, a book on the subject has been crafted. This book, an extension of the 2-day course, encapsulates the essence of the ML journey. What sets this course—and by extension, the accompanying book—apart is its commitment to demystifying ML for beginners in a fun manner. Figure 78 shows the cover of the book “One Minute Machine Learning: For the “Somewhat

Familiar” with ML Crowd.” The target audience, deliberately not confined to experts in the field, includes those embarking on the initial stages of their ML journey. To capture the whimsy of the virtual course, the book uses the Edward Tufte template, synonymous with the visually appealing incorporation of data, plots, and charts into a book form. Figure 79 shows an example of the Tufte L^AT_EX template with a wide margin where additional information (or witty pop culture references) can be placed; full length figures or text boxes can be also used as shown the the beginning of the subsection. The book mirrors the course’s trajectory, covering the spectrum of ML topics, from foundational concepts like linear regression to the intricacies of traditional neural networks.

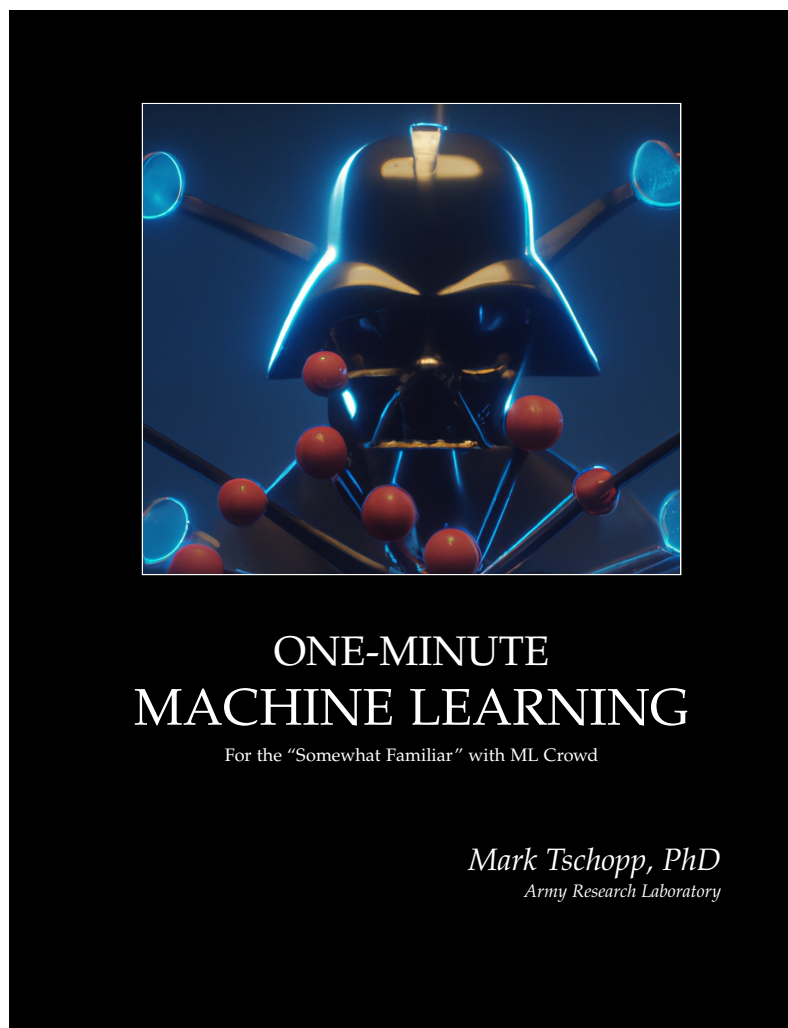


Fig. 78 The cover for the ML book based on the “Machine Learning for Everyone: May The Course Be With You” short course

In summary, the book encapsulates the spirit of the “Machine Learning for Every-

one” course, offering a guide that seamlessly combines accessibility with depth, making the intricate world of ML accessible to a broad and eager audience.

8.2 Optimization of Neural Networks

How is optimization different for neural networks?

Training neural networks involves finding the optimal values for the model parameters that minimize the loss function. There are a few different concepts for efficiently and effectively optimizing these parameters:

DIFFERENT GRADIENT DESCENT ALGORITHMS can impact the training process by accelerating finding an optimal solution and by accessing different optimal solutions. There are multiple different variants including momentum, adagrad, adadelata, and adam optimization. These algorithms can help decrease the number of iterations required to find the optimal solution.

DIFFERENT DATA SAMPLING METHODS can also impact the training process, accelerating the optimization process. Three common data sampling methods used in training neural networks are stochastic gradient descent, batch gradient descent, and mini-batch gradient descent. These methods can help decrease the time spent in each iteration to find the optimal solution faster.

GRADIENT DESCENT, while a powerful optimization algorithm, is not without its challenges. One common issue arises when the step size, also known as the learning rate, is set too large or too small. If the learning rate is too large, the algorithm may overshoot the optimal solution and fail to converge. On the other hand, if the learning rate is too small, the convergence may be slow, requiring many iterations to reach the optimal solution.

Another challenge in optimization arises when dealing with high-dimensional data. In multiple dimensions, the optimization landscape becomes more complex, with potentially many local minima and flat regions. This makes it difficult to find the global minimum and can lead to suboptimal solutions.

Additionally, gradient descent may exhibit oscillatory behavior in some cases. This occurs when the optimization path oscillates around the optimal solution, unable to converge. This oscillation can be caused by a high learning rate, conflicting gradients, or complex optimization surfaces.

These challenges highlight the importance of carefully tuning the learning rate and considering more advanced optimization algorithms to address issues like local minima and oscillations in high-dimensional spaces.

BUT FIRST, A FEW IMPORTANT CONCEPTS...

“Patience you must have, my young padawan, for with **GRADIENT DESCENT**, slowly but surely you will reach the minimum.”



Fig. 79 The Edward Tufte template for the ML book whereby a wide margin is used to allow for additional information accompanying the main text

6.3 Artificial Intelligence 101 Short Course

In the fall of 2022, a 3.5-hour module on AI was conducted for one of the Army Management Staff College's courses targeting senior leaders. On August 30, 2023, the course expanded into a full-day program and was delivered to approximately 100 participants at the Army Materiel Command's (AMC) Joint Munitions Command (JMC) headquarters (HQ) at Rock Island Arsenal. Figure 80 presents the outline slide for the course, organized into three primary sections: AI Foundations, AI and ML, and AI in Application.



Fig. 80 The outline slide for AI 101 one-day short course

The AI Foundations section covered broad topics such as defining AI and distinguishing what AI is not, delving into the history of AI from the 1940s and 1950s, and exploring the reasons behind the emergence of AI at this particular juncture. The AI and ML section delved into various ML algorithms, including unsupervised and supervised algorithms. It provided insights into how neural networks are trained, the concepts of deep learning, the functioning of convolutional neural networks, the mechanics of sequence models and large language models, and an introduction to reinforcement learning. The final section explored AI applications, addressing successes and failures, discussing the risks and ethics associated with

AI, analyzing the global landscape of AI with a focus on China, and examining the future trajectory of AI.

The AI course was infused with a blend of wit, drawing parallels to pop culture and weaving in references to a diverse array of sci-fi movies, books, and series. Figure 81 provides a glimpse of the book covers that enriched the course material,^{22–31} spanning topics from navigating AI’s goal-setting challenges (Human Compatible²²) to envisioning AI’s societal role in the future (Homo Deus,²³ Age of AI,²⁴ The War on Normal People,³⁰ Competing in the Age of AI²⁸). Delving into the course included explorations into the historical narratives of AI (Genius Makers²⁶), the evolution of AI hardware (Chip War²⁹), and the complex interplay of AI in social media and information warfare (LikeWar²⁷), as well as its role in autonomous weapons systems (Army of None³¹). Balancing the factual with the imaginative, the course embraced sci-fi fiction with works like Machines Like Me. Television series, prominently featuring AI episodes in Black Mirror, along with a plethora of documentary series, further enriched the content. These references not only brought the history and concepts of AI to life but also engaged and held the audience’s attention throughout the learning experience.



Fig. 81 An ensemble of book covers that complemented the exploration of AI concepts in the course, spanning topics from the challenges of AI goal-setting to its profound impact on society, history, hardware, and imaginative realms of sci-fi fiction

Overall, feedback for the course mirrored that of the original ML course. In that course, substantial improvements were made over the past few years based on re-

ceived feedback, iteratively improving the course.

6.4 Machine Learning Operations, MLOps

This subsection offers a concise overview on Machine Learning Operations (MLOps), learned from several courses taken during the sabbatical on the topic. Unlike traditional ML courses that often focus solely on algorithms, MLOps delves into the broader infrastructure surrounding ML systems. Recognizing that the algorithm constitutes only a fraction of the complete ML pipeline, understanding the multifaceted process is necessary for introducing ML into a production environment. The following ten key steps encapsulate the MLOps journey:

1. **Define the Problem:** Clearly articulate the problem the ML model aims to solve, setting precise performance benchmarks.
2. **Data Verification:** Rigorously validate data correctness, completeness, and consistency to ensure the model trains on reliable information.
3. **Feature Extraction:** Identify pertinent features from available data to shape the model's training process.
4. **Model Selection:** Choose the most effective ML model, grounded in comprehensive development and testing phases.
5. **Model Training and Testing:** Engage in the training of the chosen model on preprocessed data, evaluating its performance using relevant metrics.
6. **Deployment:** Seamlessly integrate the selected ML model into the production environment, often demanding software engineering skills.
7. **Monitoring and Maintenance:** Implement continuous monitoring, fine-tuning, and periodic retraining to sustain the model's accuracy and performance.
8. **Data Management:** Establish a robust data pipeline ensuring the model has constant access to essential data inputs, updating training data regularly.
9. **Error Analysis:** Conduct in-depth analysis of model errors during deployment, identifying areas for improvement and fine-tuning.
10. **Security and Privacy:** Safeguard both data and model, ensuring robust security measures and upholding data privacy standards.

Exploring the data lifecycle is also essential for ML operations, which requires emphasizing the tools essential for data transformation, anomaly detection, schema and metadata development, and data provenance. Delving into the ML data lifecycle in a production environment requires numerous stages for ensuring model accuracy, reliability, and security, such as:

1. **Data Collection:** Aggregate data from diverse sources, encompassing internal databases, log files, and external datasets. Emphasize fairness in data collection, accounting for diversity, representativeness, and privacy considerations.
2. **Data Transformation:** Shape raw data into a format suitable for training ML models. This involves processes such as data normalization, encoding categorical variables, and aligning data with the algorithm's requirements.
3. **Data Schema and Provenance:** Define a robust data schema and trace the origin and evolution of data to ensure a reliable and traceable data pipeline.
4. **Data Validation:** Scrutinize data for correctness, completeness, and consistency, including checks for missing values, duplicates, and errors.
5. **Data Statistics:** Generate descriptive statistics to comprehend data distribution and identify potential outliers.
6. **Data Labeling:** For supervised ML, accurately label data considering factors such as manual labeling, crowdsourcing, and active learning. Tools like Snorkel.ai can facilitate this process.
7. **Data Skewness:** Address data distribution skewness to prevent bias toward specific classes. Techniques may include oversampling, undersampling, generating synthetic samples, or using weighting techniques.
8. **Feature Selection:** Identify the most relevant features for training ML models through methods like feature importance scores, feature correlation analysis, or dimensionality reduction.
9. **Metadata Management:** Store metadata about the data pipeline and the model for enhanced understanding, debugging, and pipeline improvement. This includes data version, source, statistics, model parameters, and performance information.

Thinking beyond ML algorithms requires knowledge of how these algorithms integrate with the larger MLOps picture. The ML data pipeline in a production environment involves a sequential process of collecting, transforming, defining schema and provenance, validating, generating statistics, labeling data, addressing skewness, selecting features, and managing metadata. This intricate process requires considerations of fairness in data collection, the trade-offs in data labeling, the impact of skewness and feature selection on model performance, and the utilization of tools (e.g., TensorFlow Extended) to streamline pipeline management. Understanding the broader MLOps framework ensures a comprehensive approach to deploying and maintaining ML models, aligning them seamlessly with organizational objectives and end-user needs.

7. Measurable Deliverables

During this sabbatical, the researcher dedicated efforts to exploring the intersection of AI, ML, network science, computational social science, and business topics. The subsequent section delineates tangible outcomes and contributions, encompassing invited seminars and talks, the development of short courses on AI and ML, participation in conferences, workshops, and events, and active involvement in organizing and participating in VIP visits to ARL Central. Simultaneously, a commitment to personal growth and skill development was undertaken through relevant online courses, aimed at enhancing expertise in key areas. While primarily pursued for individual enrichment, these courses significantly contributed to deepening the researcher's knowledge, augmenting the overall impact of this dedicated period of research and learning, as detailed in this report.

7.1 Invited Seminars and Talks

1. **Tschopp, M.A.**, “An Arsenal of Experiences: Maneuvering the Frontlines of a Modern PhD Career... to ARL,” University of Chicago, November 16, 2023.
2. **Tschopp, M.A.**, “Star Wars and Machine Learning: Lessons Learned from the “somewhat familiar with ML” crowd,” Carnegie Mellon University, November 3, 2023.
3. **Tschopp, M.A.**, “From Hallowed Halls to Army Strong: Navigating the World of Army-Academia Collaboration,” Carnegie Mellon University, November 3, 2023.

4. **Tschopp, M.A.**, “Adventures in Career-Shifting: Navigating the Unpredictable Roads of a Modern PhD Career... and... National Laboratories!,” Case Western Reserve University, Career experiences seminar series, September 19, 2023.
5. **Tschopp, M.A.**, “Science of Science: Future Challenges and Opportunities,” Columbia University, ARO Workshop on Decentralized Science, August 17, 2023.
6. **Tschopp, M.A.**, “Standing on the Shoulders of Giants (Without Falling Off): Navigating the World of Scientific Knowledge,” 2023 National Defense Science and Engineering Graduate (NDSEG) Conference, August 1, 2023.
7. **Tschopp, M.A.**, “Adventures in Career Shifting: Navigating the Unpredictable Roads of a Modern PhD Career,” Northwestern University, ME 513 Career experiences seminar series, May 17, 2023.
8. **Tschopp, M.A.**, “From Padawan to Jedi Master: Machine Learning with a Star Wars Twist,” Northwestern University, Northwestern Institute on Complex Systems (NICO) Seminar Series, May 10, 2023.
9. **Tschopp, M.A.**, “The Sociology of Science: How to Stand on the Shoulders of Giants (Without Falling Off),” Georgia Institute of Technology, “Advances in Multiscale Mechanics and Materials Design” symposium in honor of David McDowell’s retirement, April 14, 2023.
10. **Tschopp, M.A.**, “The Science of Science: How to Stand on the Shoulders of Giants (Without Falling Off),” University of North Carolina–Chapel Hill, April 4, 2023.
11. **Tschopp, M.A.**, “From Hallowed Halls to Army Strong: Navigating the World of Army-Academia Collaboration,” University of North Carolina–Chapel Hill, April 4, 2023.
12. **Tschopp, M.A.**, “Career Experiences: Working at National Laboratories,” University of North Carolina–Chapel Hill, April 4, 2023.
13. **Tschopp, M.A.**, “DATA is the New Bacon: Why Data is Sizzlin” Right Now,” Army Management Staff College, Kickoff Invited Lecture for Data Founda-

tions course, February 24, 2023; March 31, 2023; April 28, 2023; June 23, 2023.

7.2 Short Courses

1. **Tschopp, M.A.**, “Artificial Intelligence 101,” Rock Island Arsenal, AMC Joint Munitions Command HQ, September 30, 2023 (8-h total).
 - ★ >100 registered participants.
 - ★ Post-course survey: “Instructor was able to engage the audience” (4.5/5.0), “Instructor was knowledgeable and communicated effectively” (4.6/5.0), and “Instructor presented material in interesting and engaging manner” (4.7/5.0)
 - ★ Comments: “I really loved this course! I have already recommended it to just about everyone I have spoken to, haha!,” “Great class. Mark is also a very good instructor,” “Mark did a great job!,” “S-tier break screens, Mark”, “I have to pick a favorite session??? I was intrigued with the whole thing!”
2. **Tschopp, M.A.**, “Machine Learning for Everyone: May The Fourth Be With You,” Microsoft Teams, May 4–5, 2023 (15-h total).
 - ★ >350 registered participants, 3000 person-hours of ML instruction, 2 guest speakers.
 - ★ Post-course survey: “Overall evaluation of course is excellent” (4.91/5.0), “Instructor is excellent” (4.97/5.0), and “Highly recommend to others” (4.97/5.0)
 - ★ “Dr. Tschopp’s course was one of the rare courses that teaches an incredibly complex subject effectively to a diverse audience from PhDs to old engineers like me 22 years out of college (25 out of linear algebra!). Outstanding job!”

7.3 Conferences, Workshops, Events

1. 2023 National Cybersecurity Education Colloquium (NCEC), Chicago, IL, September 18–22, 2023 (invited panelist, invitation special topic).
 - ★ “How to do business with ARL,” invited panelist for “Federal funding opportunities” workshop, September 22, 2023, and

- ★ Invited member of invitation-only committee on “Generative AI Tools and Cybersecurity,” September 21, 2023
- 2. ARO “Decentralized Science” workshop, Columbia University, New York, NY, August 17–18, 2023 (invited speaker).
- 3. 2023 National Defense Science and Engineering Graduate (NDSEG) Conference, July 29, 2023 – August 2, 2023.
- 4. 2023 International Conference on Science of Science Innovation, Northwestern Kellogg School, June 26–29, 2023.
 - ★ Award for Hackathon Team “ScholarConnect”
- 5. 2023 Midwest Machine Learning Symposium, 16–17 MAY 2023.
- 6. “The University of Chicago and Caltech Conference on AI + Science,” workshop, March 28–30, 2023.
- 7. “Imagining the Futures of Collaboration in Research,” Northwestern University workshop, Buffett School of Public Policy, March 1, 2023.
- 8. “Internet of Battlefield Things” Collaborative Research Alliance boot camp, 2-day IOBT Bootcamp, Chicago, IL, January 12–13, 2023.

7.4 VIP Engagement

- 1. SEN Duckworth (D-IL) Military Legislative Assistant and Staff Delegation Visit (ARL Central), November 2023
 - ★ **Tschopp, M.A.**, “University and Defense Partnerships in ARL Central”
- 2. BG Trybula Visit (ARL Central), April 2023
 - ★ Led and hosted visit by BG Trybula to ARL CENTRAL
 - ★ **Tschopp, M.A.**, “Welcome to ARL Central!”

8. Conclusion

In an era marked by the rapid transformation of science and technology through data science, analytics, machine learning, and artificial intelligence, organizations find themselves compelled to embrace data-driven decision-making practices. Staying abreast of advancements and remaining competitive requires not only interrogating data through descriptive analytics but also diagnosing data using diagnostic analytics, making predictions through predictive analytics, and effecting constructive changes through prescriptive analytics. Each domain safeguards its datasets, holding valuable insights and knowledge ready for exploration. The motivation behind this sabbatical is to further develop skills in data science, natural language processing, machine learning, and artificial intelligence, extending their applications across the interdisciplinary realm of the science of science.

In this sabbatical report, diverse topics were delved into which were situated at the intersection of artificial intelligence, machine learning, network science, computational social science, and business. This multifaceted exploration unfolded across several studies, collectively contributing to the comprehension of the expansive landscape encompassing scientific research, collaboration dynamics, language trends, impacts of funding agencies, and the dissemination of knowledge through educational initiatives.

The three research studies detailed herein serve as exemplars, highlighting the intrinsic value of open-source data and the current ease of accessing information. While these studies offer a glimpse into the potential, it is evident that research and development organizations can harness open-source datasets for driving innovation, promoting collaboration, and refining decision-making processes. This is achieved through understanding of critical entities such as scientists, research output, collaborations, funding sources, uniqueness, and scientific impact. The pursuit of knowledge and insight within these comprehensive datasets extends beyond academic exercise; it is imperative for organizations seeking to thrive in the dynamic and swiftly evolving data-rich landscape.

Key Findings: The following bullets unpack specific insights derived from the three research studies, shedding light on discoveries during the course of these investigations.

- *Global Collaboration.* The exponential growth of scientific articles, doubling approximately every 14 years, is mirrored by a noteworthy shift in collaborative efforts. Presently, around 20% of all scientific articles involve collaborations across multiple countries. While the US has traditionally been a common collaborative partner, this trend is evolving, with China emerging as a collaborative force not only for top countries but also with US federal institutions, academic entities, and projects funded by US agencies. This study systematically quantifies these evolving global collaboration dynamics, along with other focused examinations such as exploring critical research domains of artificial intelligence, quantum science, and nanotechnology.
- *Hype Language in NSF Awards.* The dynamics of the award process and science communication have witnessed a substantial influence from hype language. A comprehensive analysis of 139 hype words across over 480,000 NSF awards spanning four decades reveals a remarkable surge in the frequency of hype language in titles and abstracts over the last 30 years—some cases experiencing an increase exceeding an order of magnitude (10 words). This escalates to 39 hype words surpassing an order of magnitude when considering the base rate of change in digitized books. Many of these fall into the utility and scale hype category, emphasizing the perceived usefulness and broad scale of the research. The study concludes with a perspective on how large language models may play a future role in mitigating this growing trend in scientific discourse.
- *Funding Agencies: from Awards to Articles.* The nexus between funding agencies and their return on investment lies in the intricate relationship with the scientific knowledge generated, primarily disseminated through scientific articles, and the impact of that science, often measured through ensuing citations. This study delves into extracting award ID information from open-source databases, elucidating the linkage from funding to articles to citations for the ARO. A notable aspect is the agency award ID in the acknowledgment section of articles, where only 74% conform to correct formats, necessitating cleaning operations for the 15% of incorrectly formatted award IDs (with 10% entirely lacking an award ID). With refinements pushing the compliance rate to 85%, this linkage provides insights into the number of articles and citations per award, the associated cost for each metric, and further downstream analysis.

Machine Learning 101 and Artificial Intelligence 101: Empowering the DOD workforce with advanced data analytics methods like machine learning and artificial intelligence is crucial for navigating the evolving landscape. Despite the availability of numerous courses, often attended by only a fraction of the total workforce, the challenge persists in ensuring a comprehensive understanding of the concepts of machine learning and artificial intelligence across diverse staff members. This report provides a brief overview of a machine learning 101 and artificial intelligence 101 short course that was developed by Dr. Mark Tschopp. Complementary to other efforts within the Army, this course plays a pivotal role in educating and preparing the Army and DOD workforce for the challenges and opportunities presented by the data-driven future.

Participation in conferences, workshops, and events showcased an active engagement with the scientific community. Invited seminars and talks covered a range of topics, from career experiences to the societal impact of science, fostering collaborations and knowledge exchange across institutions. The sabbatical also led to impactful VIP engagements, where interactions with military legislative assistants, staff delegations, and senior leaders demonstrated the value of science within defense organizations as well as the importance of partnership and collaboration in the pursuit of science. These deliverables show that sabbatical can be both about focused learning of new skills and knowledge while still remaining engaged and active within the broader DEVCOM ARL organization.

Key Takeaways. In the current landscape, open-source data is a valuable resource, but organizations often hold an equally valuable wealth of internal data. In some instances, integrating both external and internal knowledge sources can unlock deeper insight and transformative outcomes for organizations. The combination of these data sets can be crucial for maximizing additive value. This transformative potential only lies through carefully curating data, strategically aggregating, fostering a holistic understanding through data science, and predicting responses and outcomes via machine learning and artificial intelligence. Forward-thinking R&D organizations are progressively incorporating data scientists with this expertise, unlocking profound insights into their core business functions.

This report emphasizes how R&D organizations can strategically leverage the growing field of science of science, thereby fostering a deeper comprehension of organi-

zational uniqueness, their roles in the broader scientific landscape, and the impact of their work in the open scientific domain. The evolving science operational environment, fueled by rapid advancements in machine learning and artificial intelligence, underscores the need for acquiring talent and organizational skills that bridge the gap between data and science.

The primary goal of the sabbatical was to immerse in a new environment, acquire new skills, and engage with external researchers. For applying these findings, the analyses presented are straightforward and can be easily adapted to address diverse research questions. With the increasing use of APIs and other tools for sharing curated data, such analyses can be tailored to explore various organizational aspects. For example, one could investigate global collaboration partners, track the evolution of these collaborations, assess their impact in terms of citations, identify partners contributing to greater impact, examine the most collaborative research areas, and make comparisons with other similar organizations. In essence, this report aims to showcase the potential of open-source datasets available today while inspiring new questions for future research endeavors to answer. In summary, this sabbatical allowed for the opportunity to explore, collaborate, and disseminate knowledge at the intersection of these diverse fields. The studies conducted, courses developed, and engagements undertaken collectively contribute to the broader scientific landscape, fostering understanding, collaboration, and innovation.

9. References

1. Wang D, Barabási AL. The science of science. Cambridge University Press; 2021. 1
2. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási AL. Science of science. *Science*. 2018;359(6379). 1
3. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu BJP, Wang K. An overview of Microsoft Academic Service (MAS) and applications. In: Proceedings of the 24th International Conference on World Wide Web; ACM; 2015. 2, 40
4. Wang K, Shen Z, Huang C, Wu CH, Eide D, Dong Y, Qian J, Kanakia A, Chen A, Rogahn R. A review of microsoft academic services for science of science studies. *Front Big Data*. 2019;2. 2
5. Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. 2022;2205.01833. 2
6. NSF Awards Web API. [accessed 2023 Mar 28]. <https://resources.research.gov/common/webapi/awardapisearch-v1.htm>. 2, 70
7. NSF Download Awards by Year. [accessed 2023 Mar 28]. <https://www.nsf.gov/awardsearch/download.jsp>. 2, 70
8. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P); 2016; Saarbrücken, Germany. 372–387. 4
9. Lowry OH, Rosebrough NJ, Farr AL, Randall RJ. Protein measurement with the folin phenol reagent. *J Biol Chem*. 1951;193(1):265–275. 5
10. OpenAlex API Technical Documentation. [accessed 2024 Jan 30]. <https://docs.openalex.org>. 6
11. Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. *Science*. 2007;316(5827):1036–1039. 7, 8
12. Adams J. The fourth age of research. *Nature*. 2013;497:557–560. 8

13. Vinkers CH, Tijdink JK, Otte WM. Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: retrospective analysis. *BMJ*. 2015;351:h6467. [67](#)
14. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, Mitchell L, Harris KD, Kloumann IM, Bagrow JP, Megerdooian K, McMahon MT, Tivnan BF, Danforth CM. Human language reveals a universal positivity bias. *Proc Natl Acad Sci*. 2015;112(8):2389–2394. [67](#)
15. Millar N, Batalo B, Budgell B. Trends in the use of promotional language (hype) in abstracts of successful national institutes of health grant applications, 1985–2020. *JAMA Netw Open*. 2022;5(8):e2228676. [68](#), [69](#), [70](#), [71](#)
16. Millar N, Batalo B, Budgell B. Trends in the use of promotional language (hype) in national institutes of health funding opportunity announcements, 1992–2020. *JAMA Netw Open*. 2022;5(11):e2243221. [68](#), [69](#), [70](#), [71](#)
17. Millar N, Batalo B, Budgell B. Promotional language (hype) in abstracts of publications of national institutes of health–funded research, 1985–2020. *JAMA Netw Open*. 2023;6(12):e2348706. [68](#), [69](#), [70](#), [71](#)
18. Lowe D. A novel approach to new types of novel novelty. *Science*; 2022 Aug 30 [accessed 2024 Jan 1]. <https://www.science.org/content/blog-post/novel-approach-new-types-novel-novelty>. [68](#)
19. Gaffney T. Unprecedented! incredible! scientific grant applications are getting heavier on hype. *STAT News*; 2022 Aug 30 [accessed 2024 Jan 1]. <https://www.statnews.com/2022/08/30/science-grant-nih-hype-application/>. [68](#)
20. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL. Quantitative analysis of culture using millions of digitized books. *Science*. 2011;331(6014):176–182. [90](#)
21. Google Books Ngram Viewer. [accessed 2024 Jan 1]. <https://books.google.com/ngrams/>. [90](#)
22. Russell S. *Human compatible*. Penguin Books; 2019. [144](#)

23. Harari YN. Homo deus. Harvill Secker; 2016. [144](#)
24. Kissinger HA, Schmidt E, Huttenlocher D. The age of AI. Little, Brown and Company; 2021. [144](#)
25. McEwan I. Machines like me. Anchor; 2020. [144](#)
26. Metz C. Genius makers. Dutton; 2021. [144](#)
27. Singer PW, Brooking ET. LikeWar. Eamon Dolan/Houghton Mifflin Harcourt; 2018. [144](#)
28. Iansiti M, Lakhani KR. Competing in the age of AI. Cambridge University Press; 2021. [144](#)
29. Miller C. Chip war. Scribner; 2022. [144](#)
30. Yang A. The war on normal people. Hachette Books; 2018. [144](#)
31. Scharre P. Army of none. WW Norton & Company; 2018. [144](#)

Appendix A. Interconnectedness of Level 0 and Level 1 Concepts

Art Level 0 Concept Sankey Plot

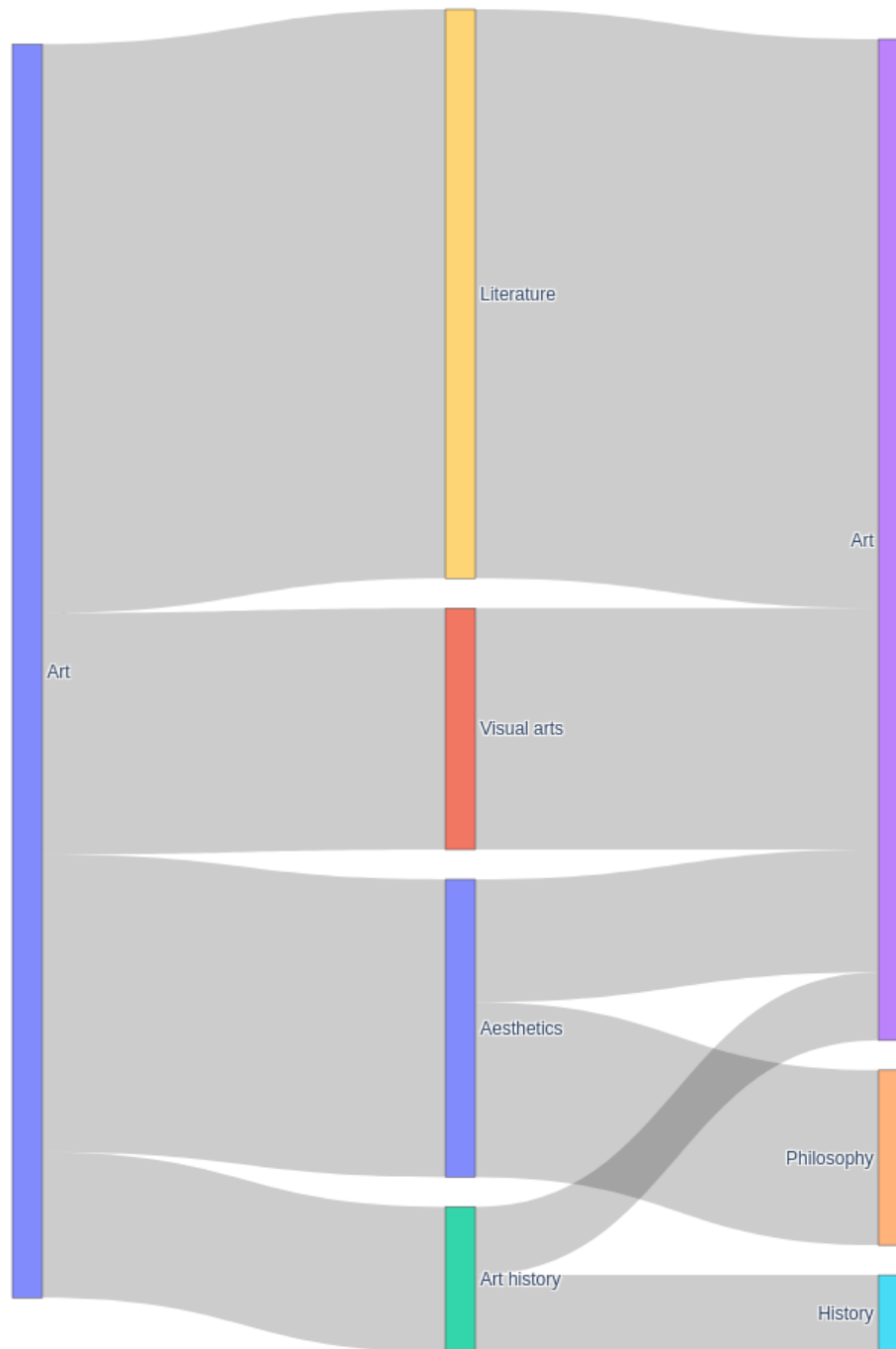


Fig. A-1 This Sankey plot visually shows the associations stemming from the level 0 concept Art (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Art, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Biology Level 0 Concept Sankey Plot

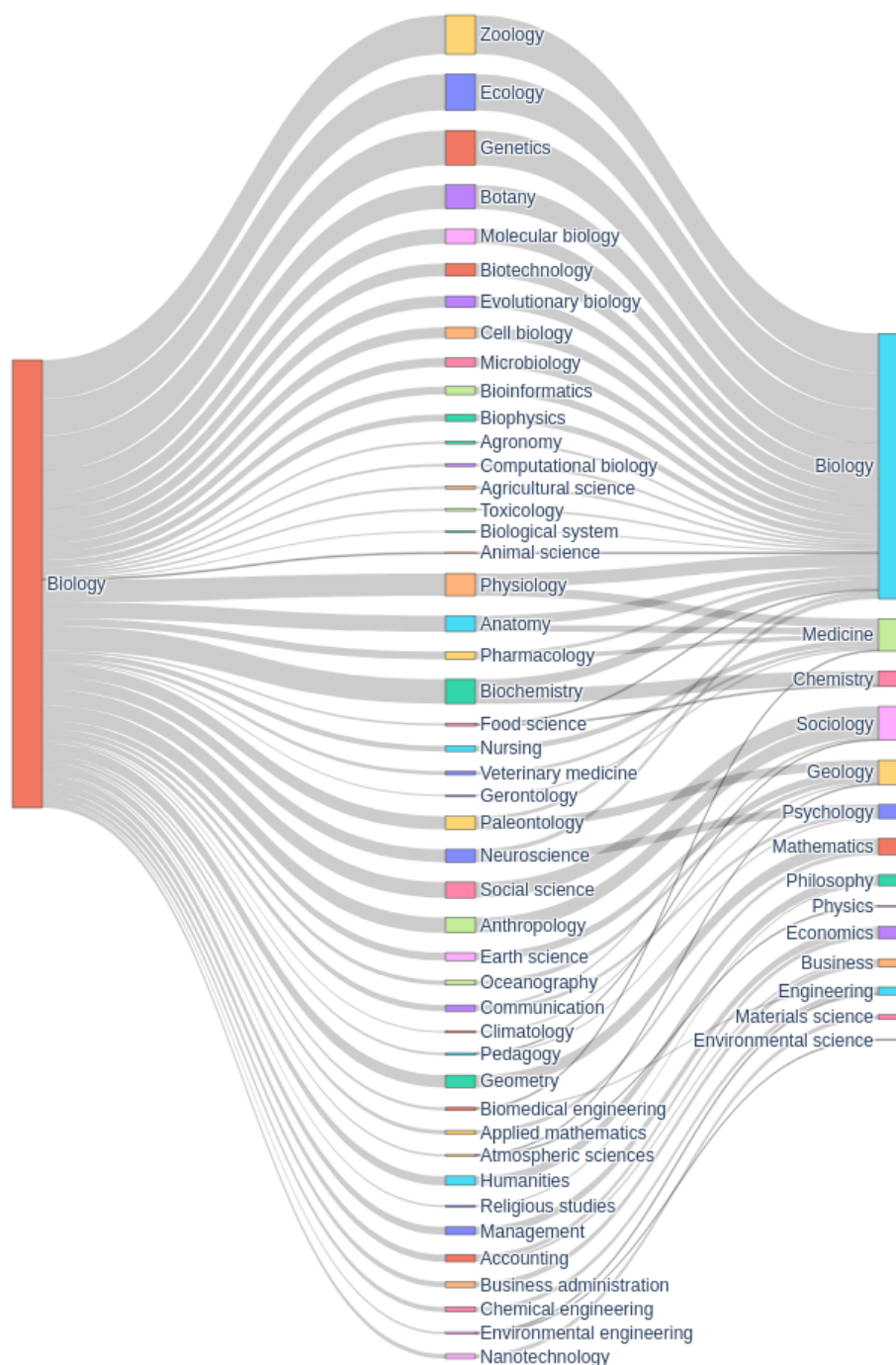


Fig. A-2 This Sankey plot visually shows the associations stemming from the level 0 concept **Biology** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Biology**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Business Level 0 Concept Sankey Plot

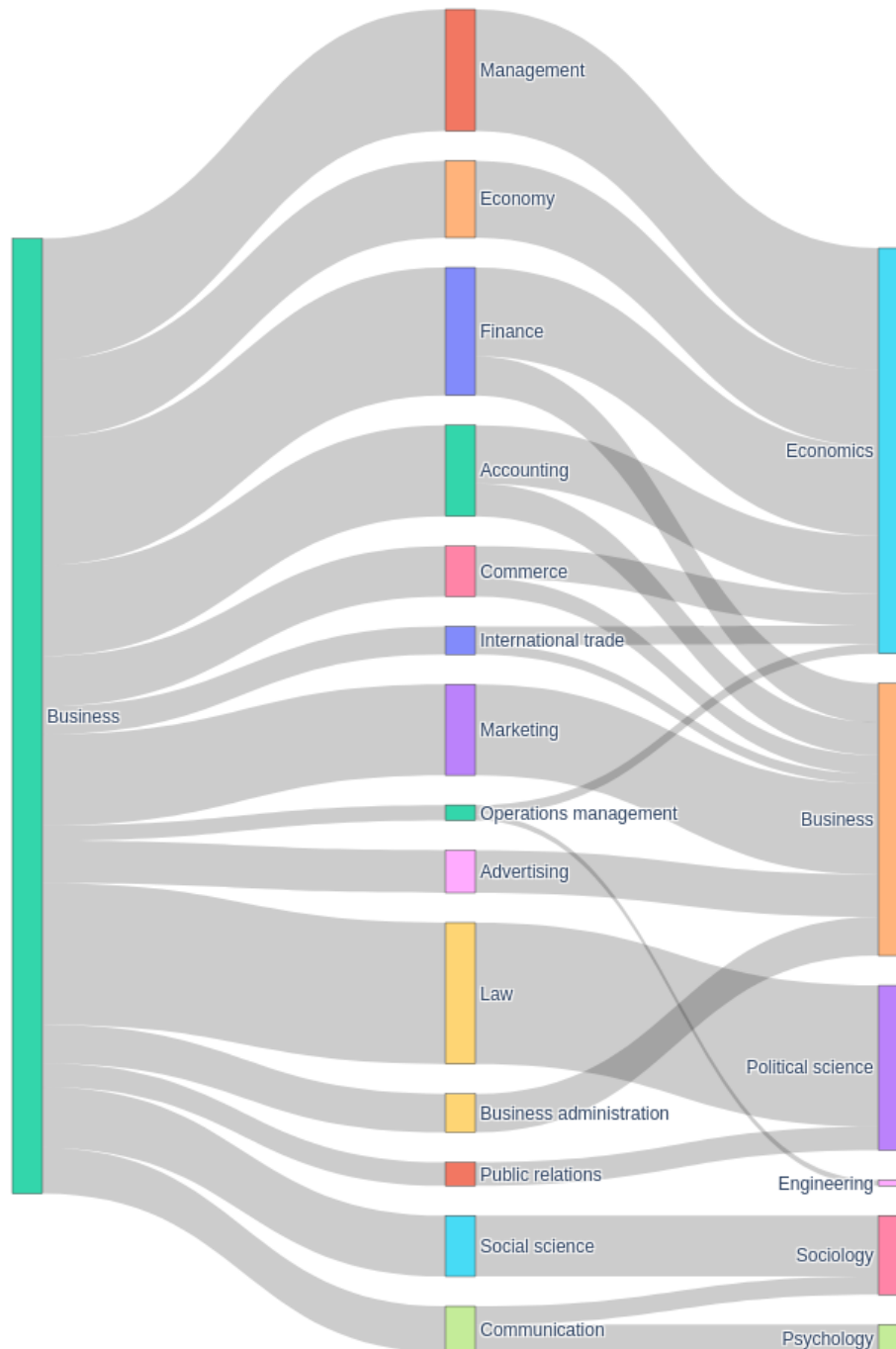


Fig. A-3 This Sankey plot visually shows the associations stemming from the level 0 concept **Business** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Business**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Chemistry Level 0 Concept Sankey Plot

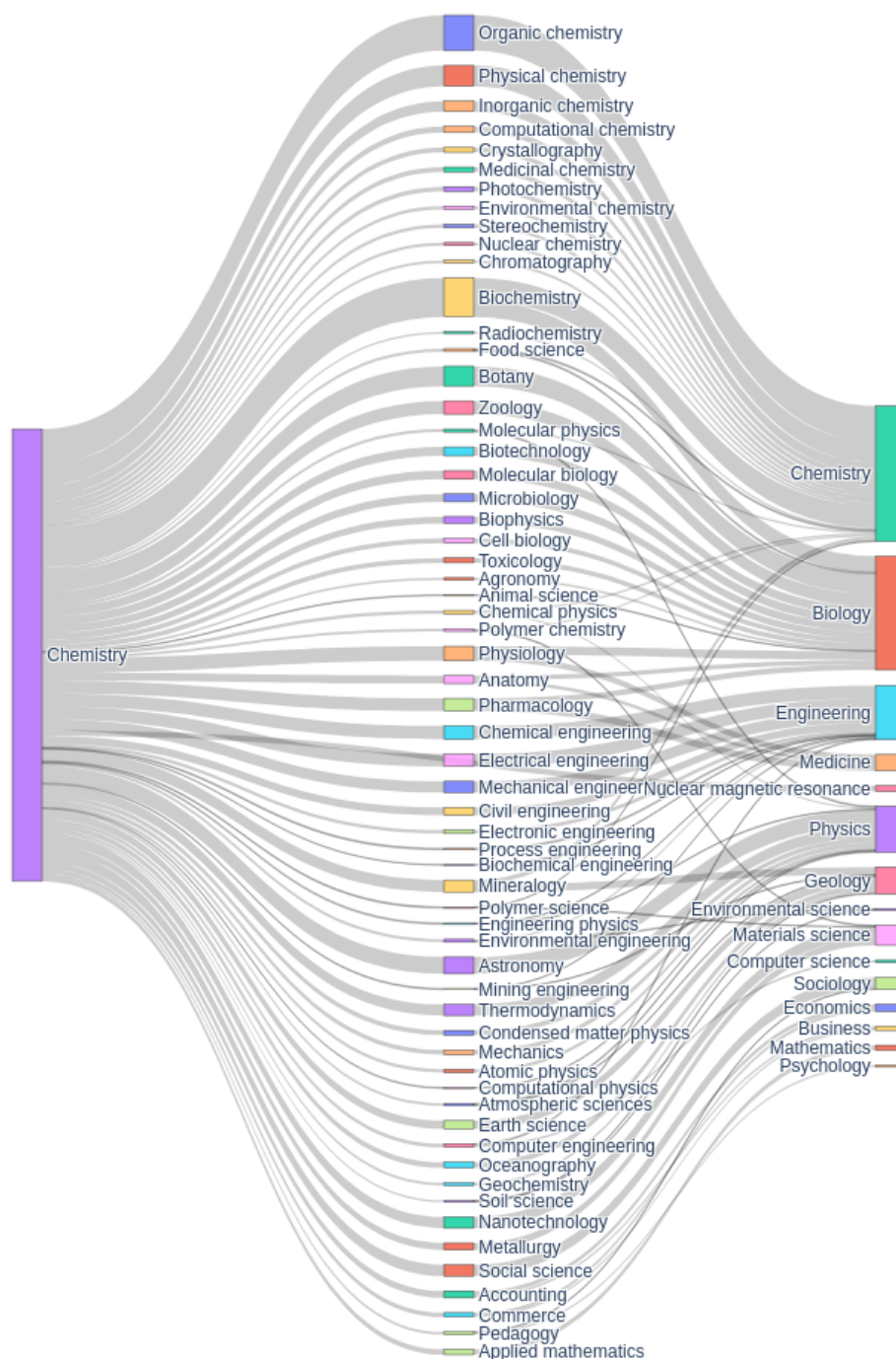


Fig. A-4 This Sankey plot visually shows the associations stemming from the level 0 concept Chemistry (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Chemistry, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Computer Science Level 0 Concept Sankey Plot

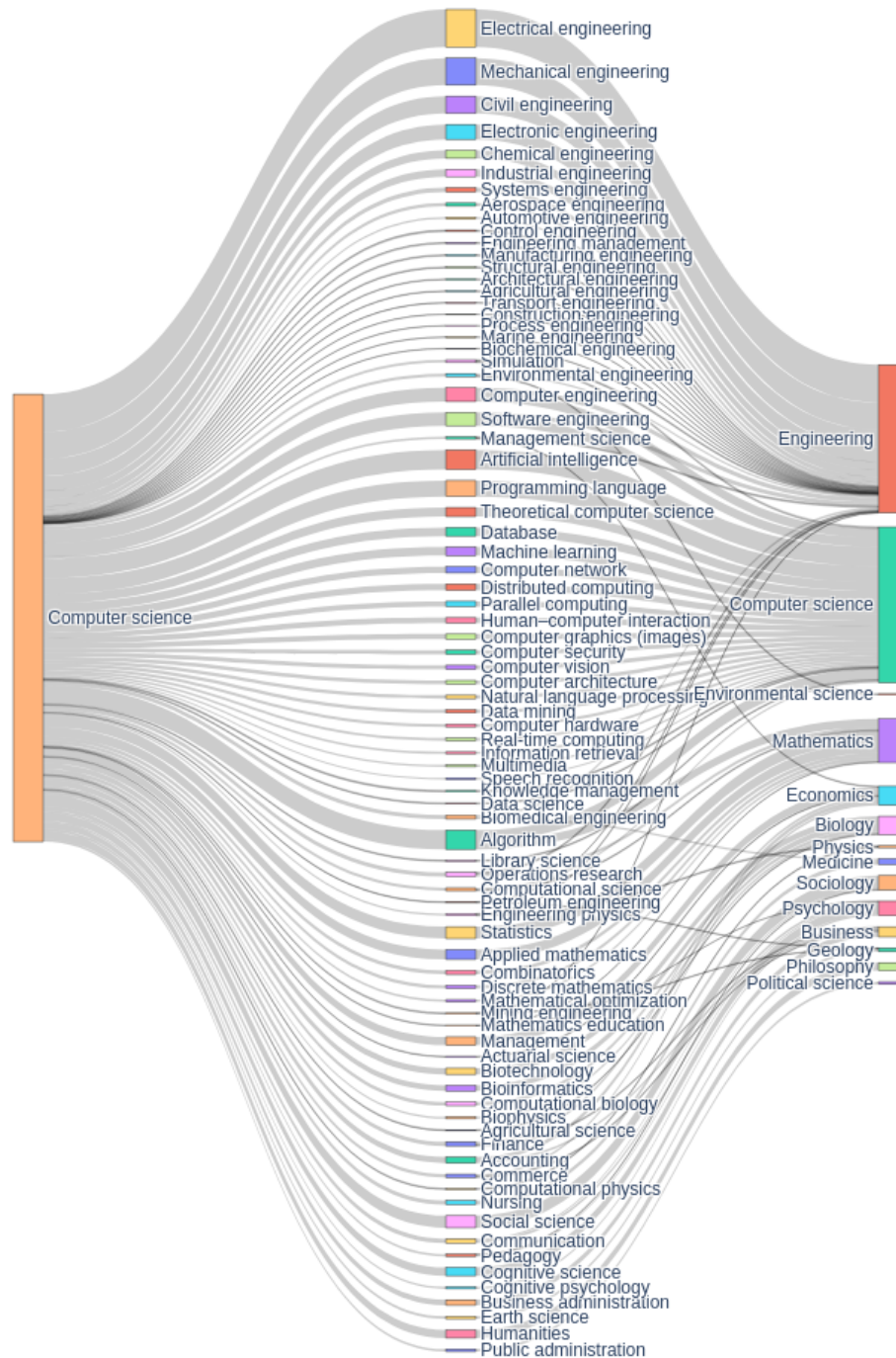


Fig. A-5 This Sankey plot visually shows the associations stemming from the level 0 concept Computer Science (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Computer Science, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Economics Level 0 Concept Sankey Plot

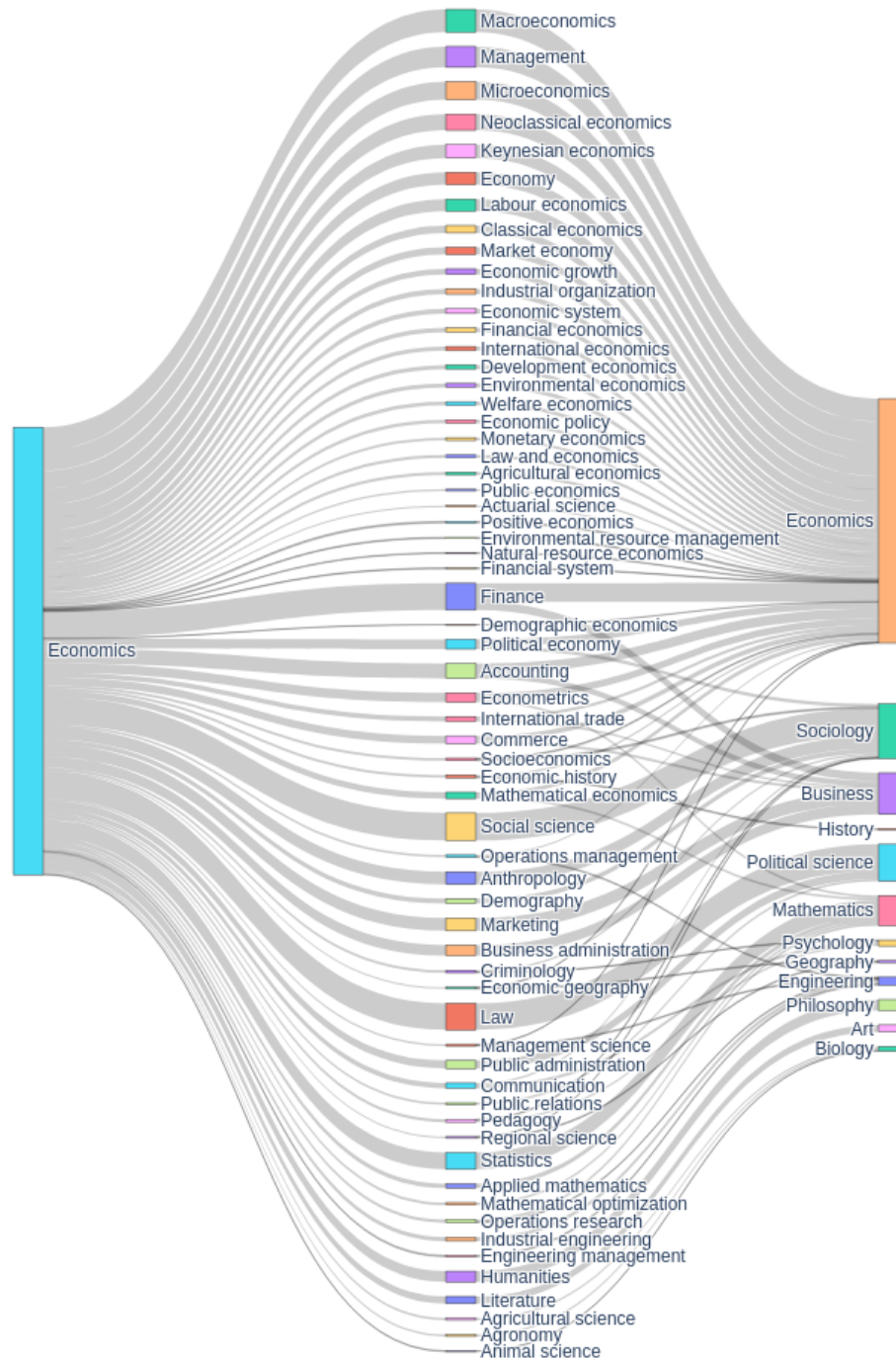


Fig. A-6 This Sankey plot visually shows the associations stemming from the level 0 concept Economics (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Economics, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Engineering Level 0 Concept Sankey Plot

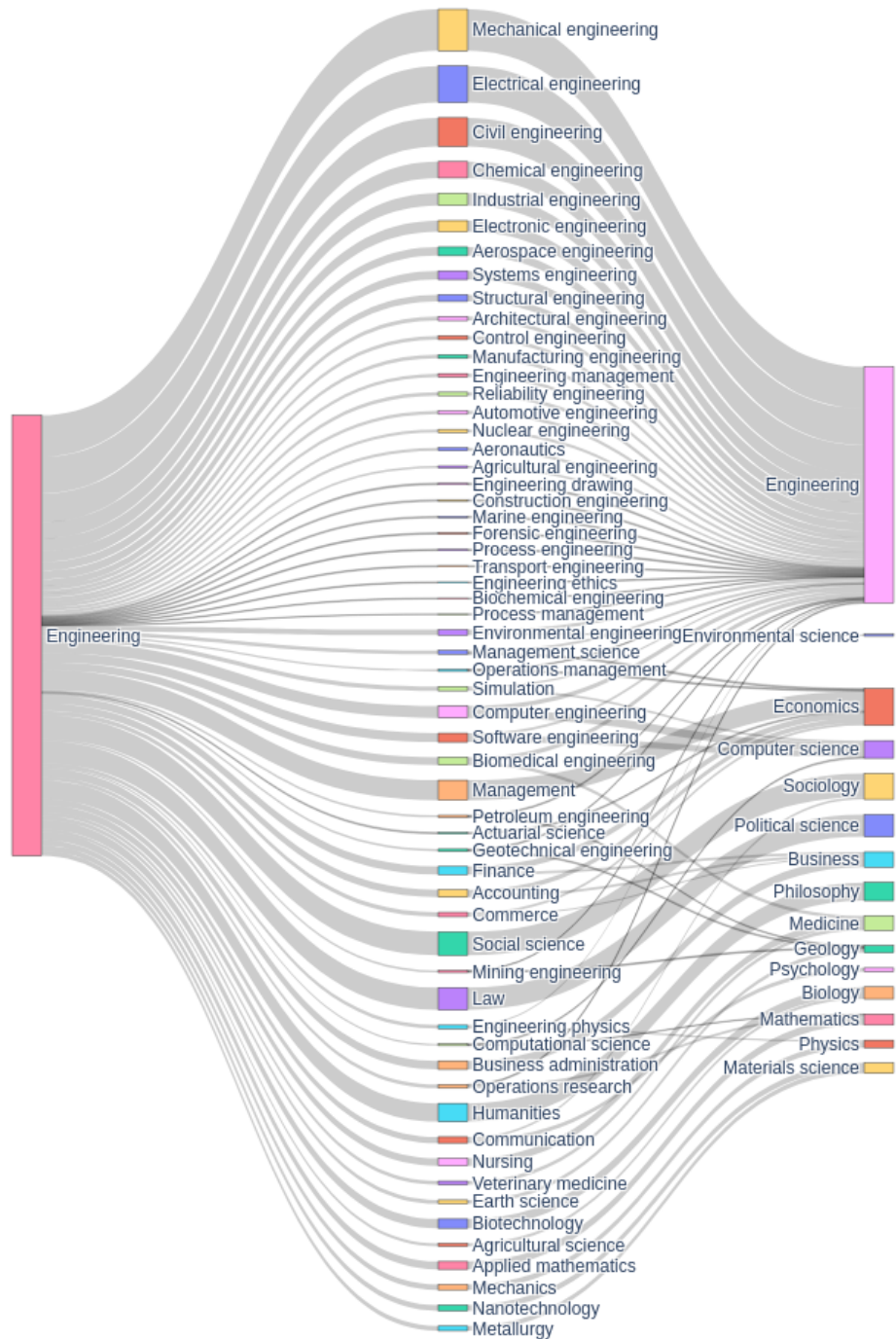


Fig. A-7 This Sankey plot visually shows the associations stemming from the level 0 concept Engineering (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Engineering, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Environmental Science Level 0 Concept Sankey Plot

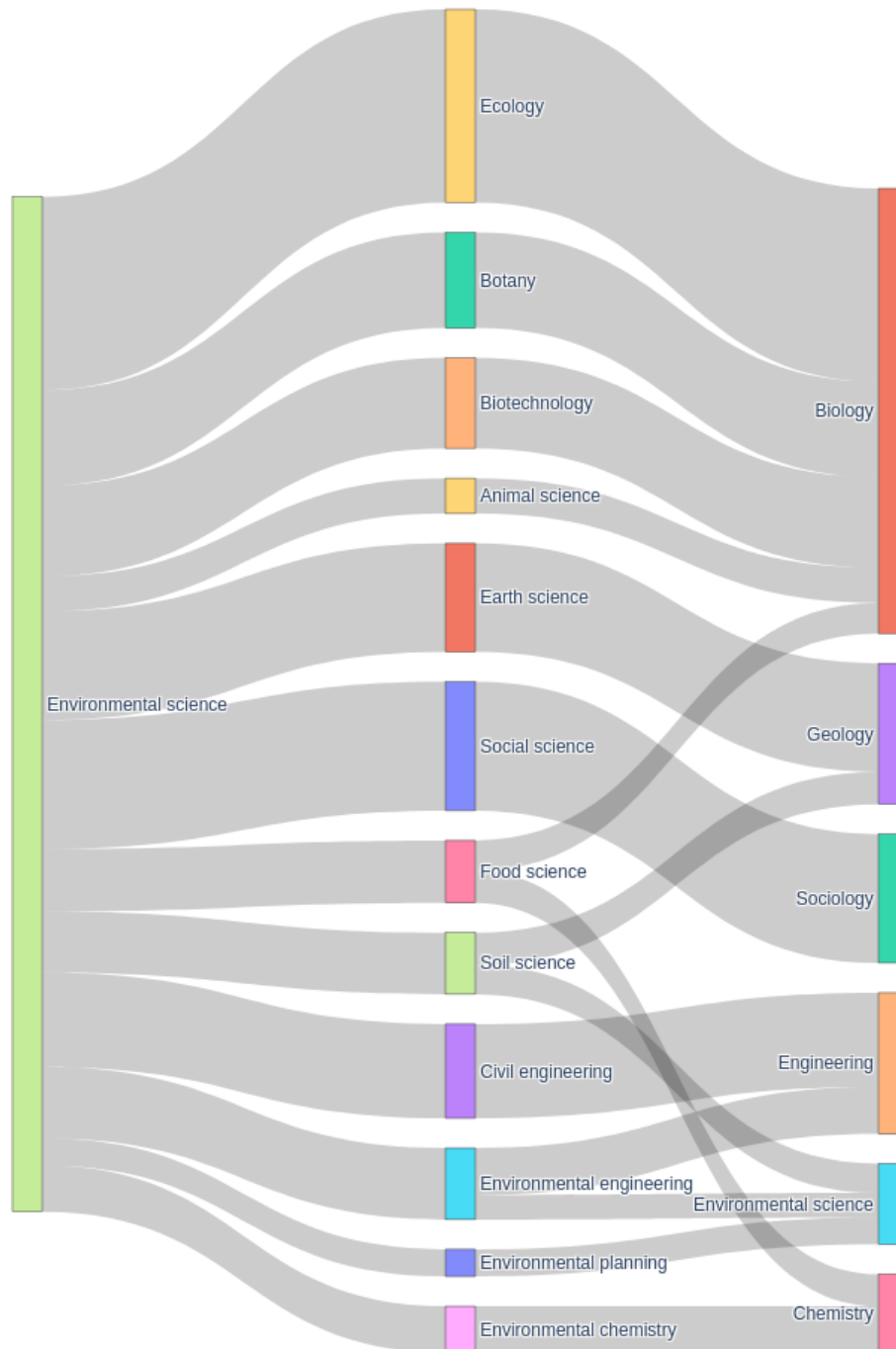


Fig. A-8 This Sankey plot visually shows the associations stemming from the level 0 concept Environmental Science (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Environmental Science, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Geography Level 0 Concept Sankey Plot

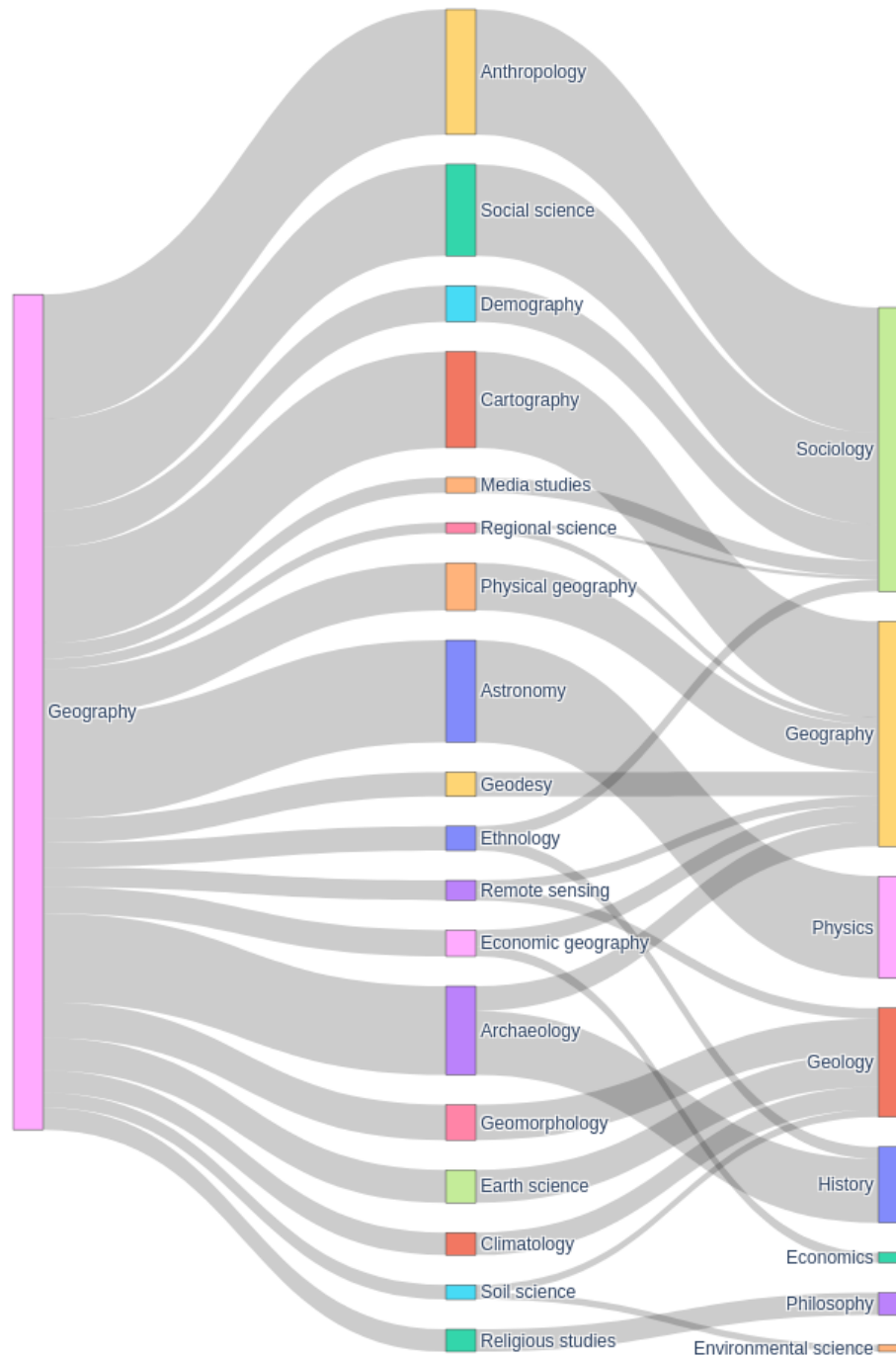


Fig. A-9 This Sankey plot visually shows the associations stemming from the level 0 concept Geography (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Geography, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Geology Level 0 Concept Sankey Plot

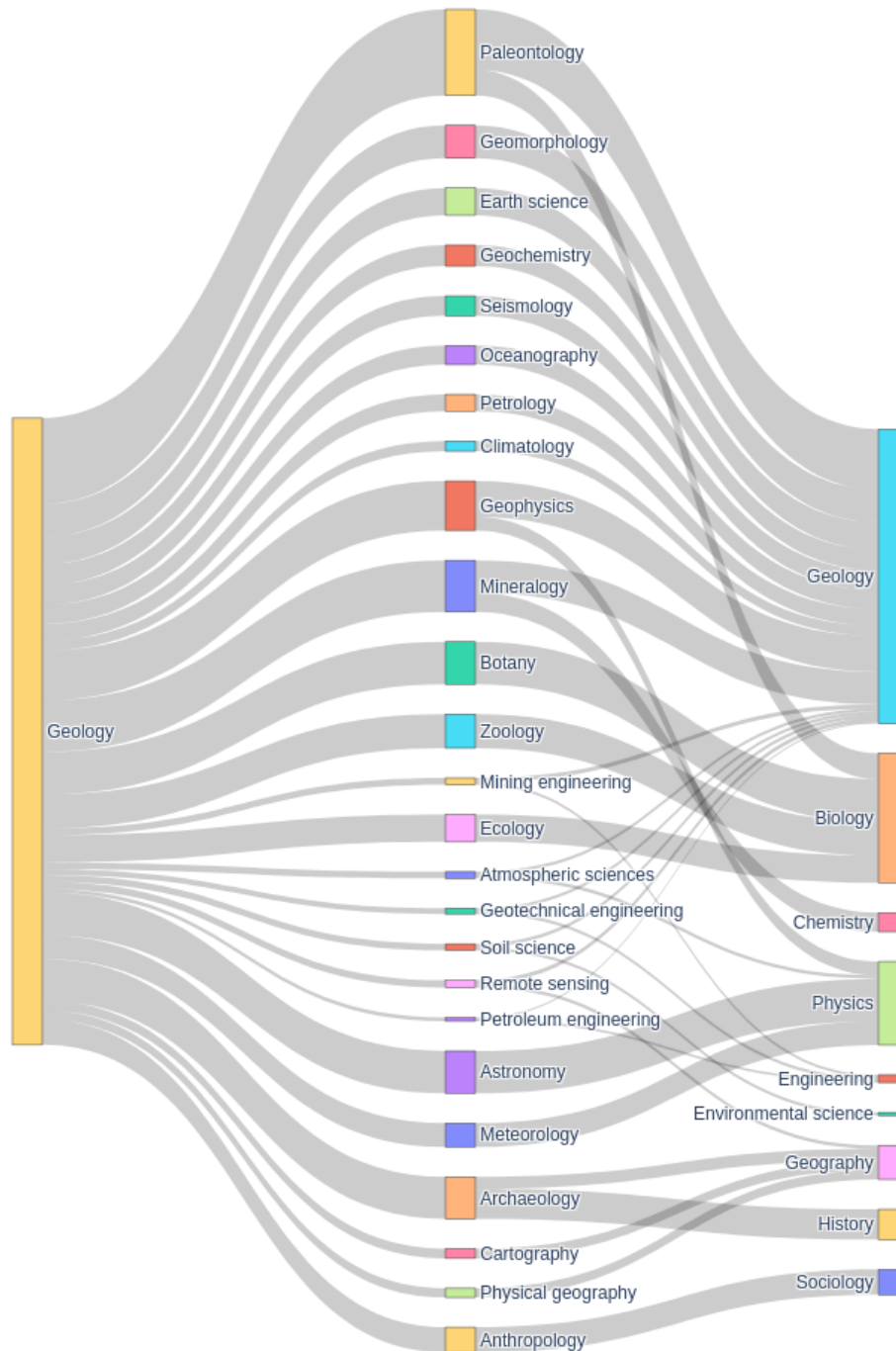


Fig. A-10 This Sankey plot visually shows the associations stemming from the level 0 concept Geology (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Geology, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

History Level 0 Concept Sankey Plot

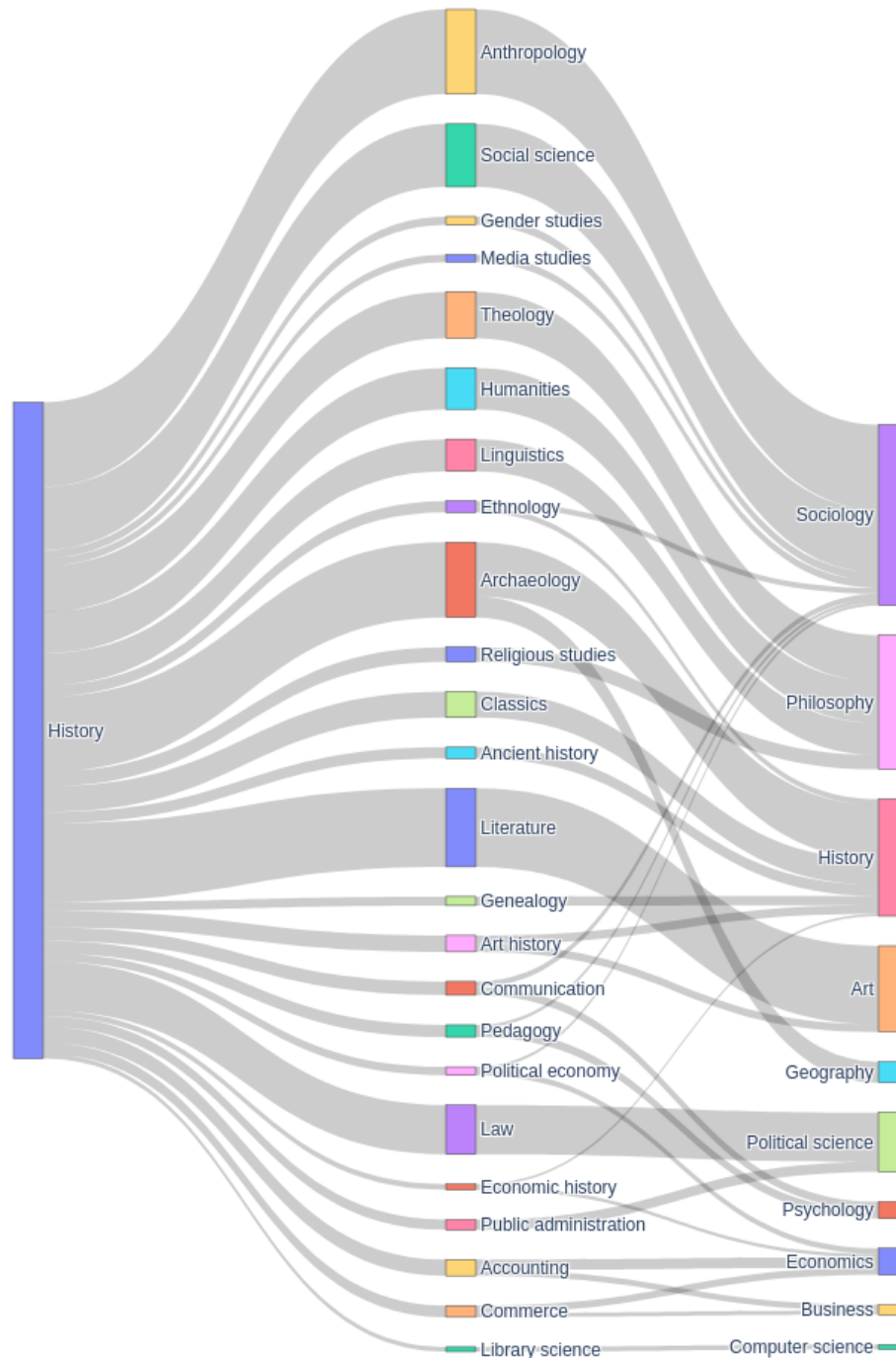


Fig. A-11 This Sankey plot visually shows the associations stemming from the level 0 concept **History** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **History**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Materials Science Level 0 Concept Sankey Plot

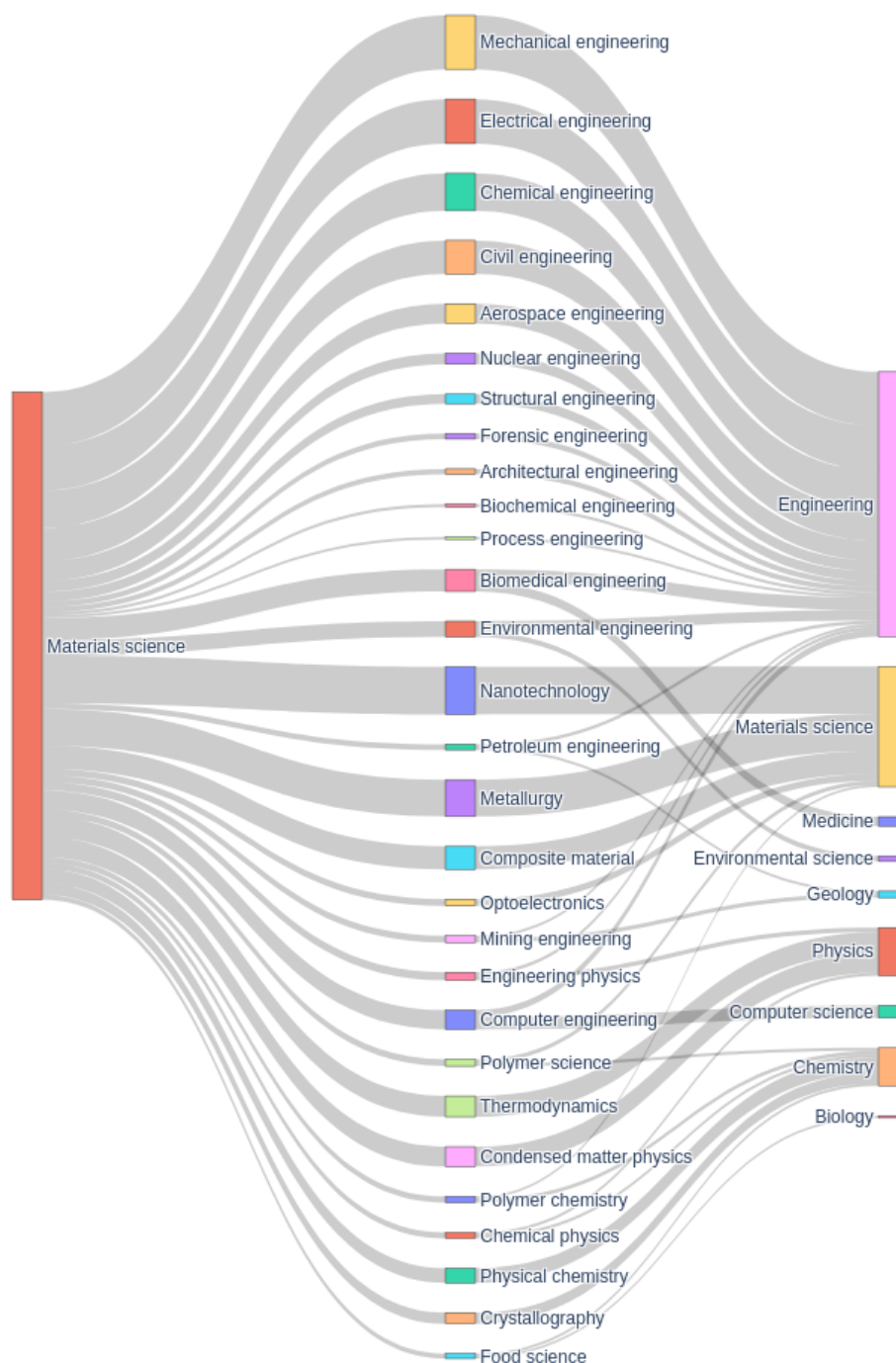


Fig. A-12 This Sankey plot visually shows the associations stemming from the level 0 concept Materials Science (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Materials Science, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Mathematics Level 0 Concept Sankey Plot

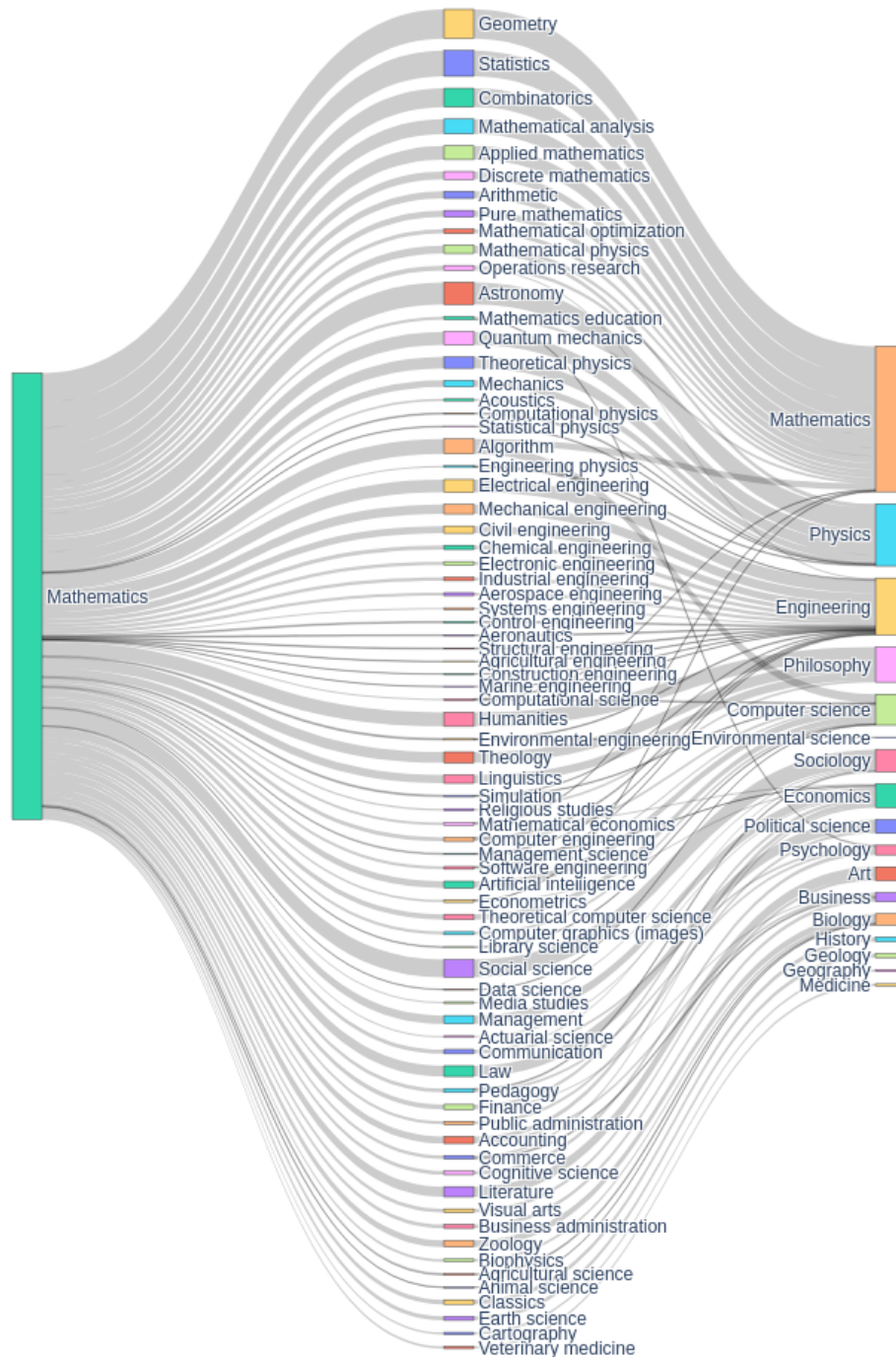


Fig. A-13 This Sankey plot visually shows the associations stemming from the level 0 concept Mathematics (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Mathematics, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Medicine Level 0 Concept Sankey Plot

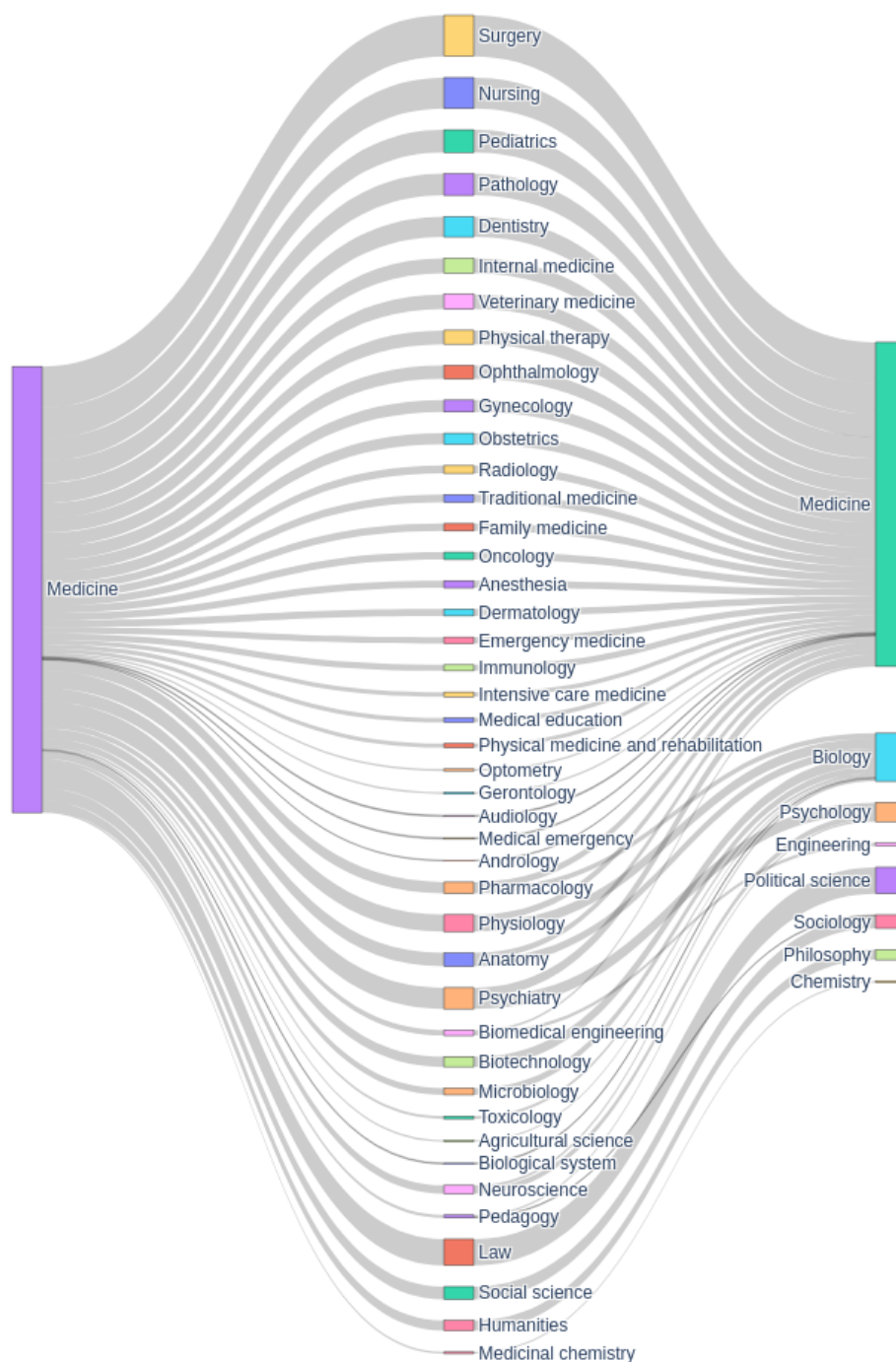


Fig. A-14 This Sankey plot visually shows the associations stemming from the level 0 concept **Medicine** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Medicine**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Philosophy Level 0 Concept Sankey Plot

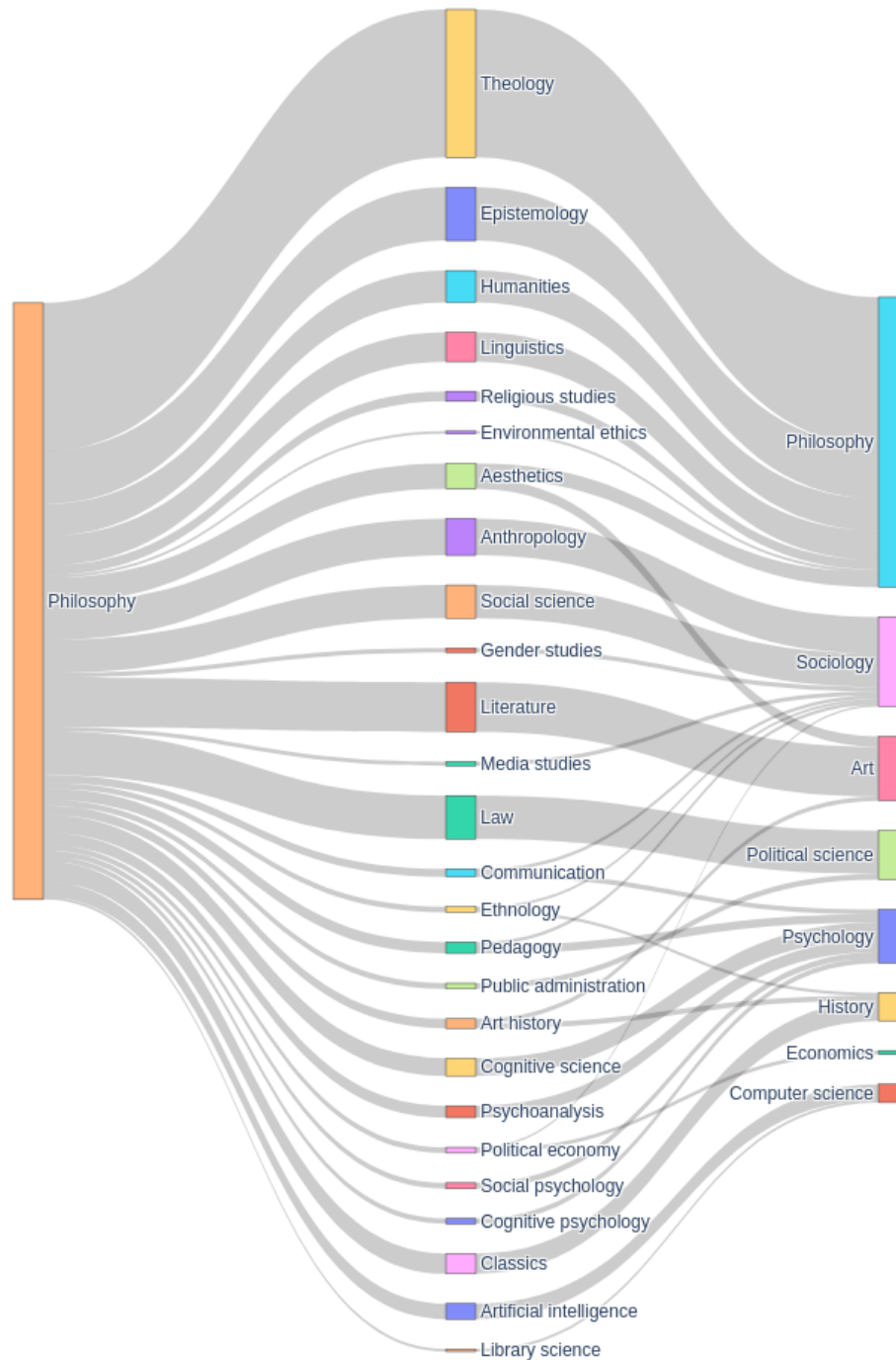


Fig. A-15 This Sankey plot visually shows the associations stemming from the level 0 concept **Philosophy** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Philosophy**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Physics Level 0 Concept Sankey Plot

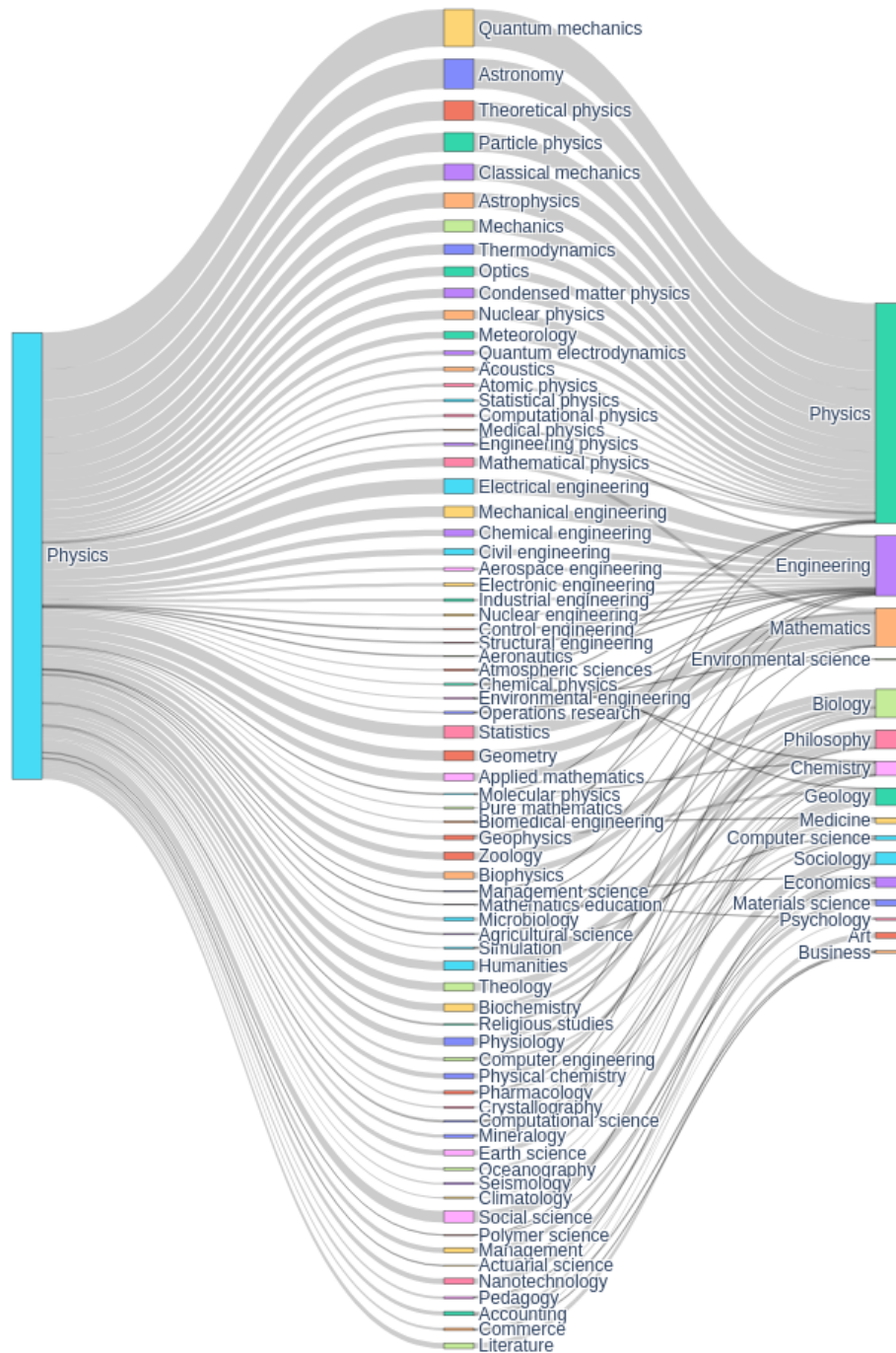


Fig. A-16 This Sankey plot visually shows the associations stemming from the level 0 concept **Physics** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Physics**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Political Science Level 0 Concept Sankey Plot

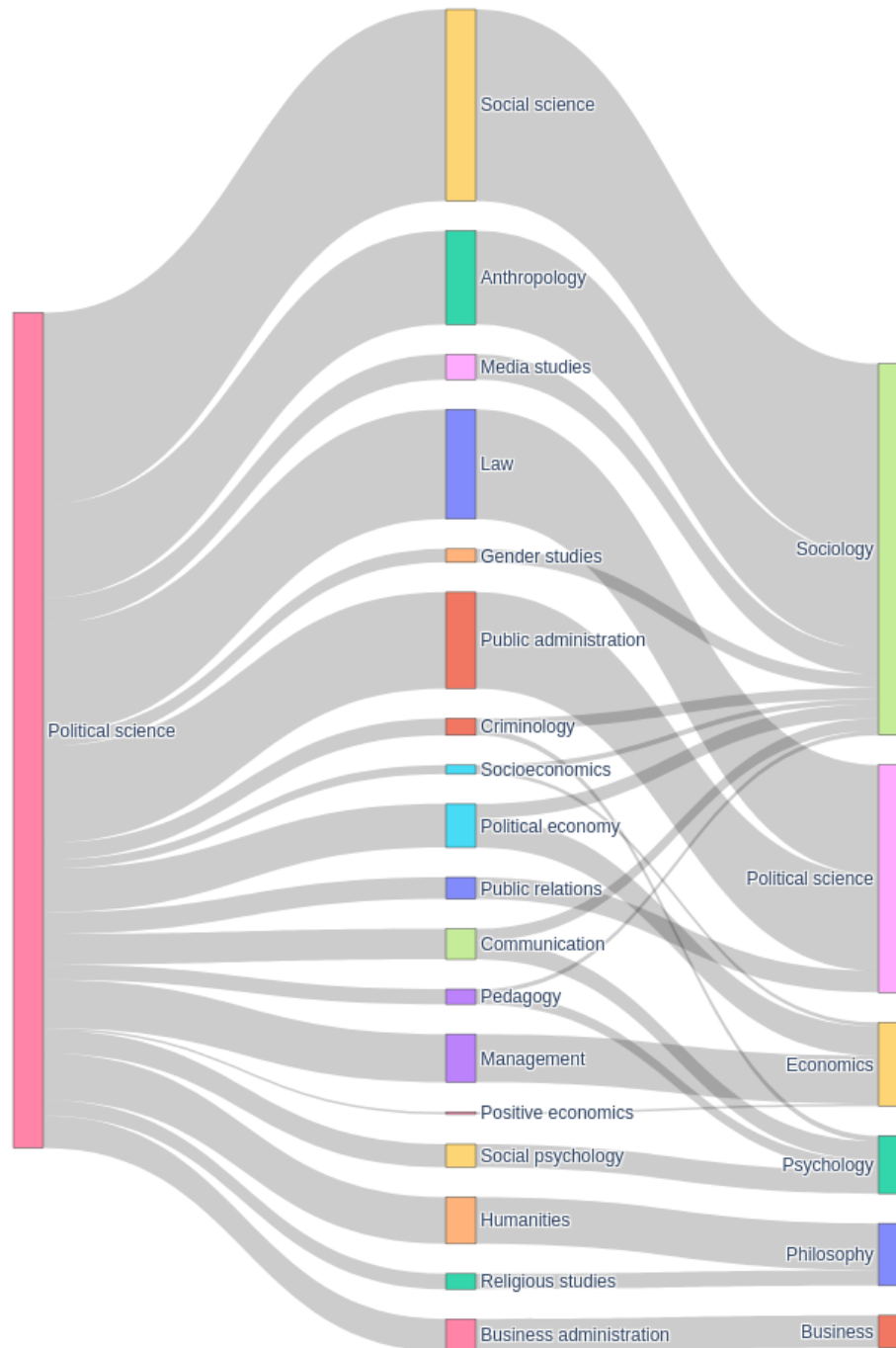


Fig. A-17 This Sankey plot visually shows the associations stemming from the level 0 concept Political Science (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Political Science, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Psychology Level 0 Concept Sankey Plot

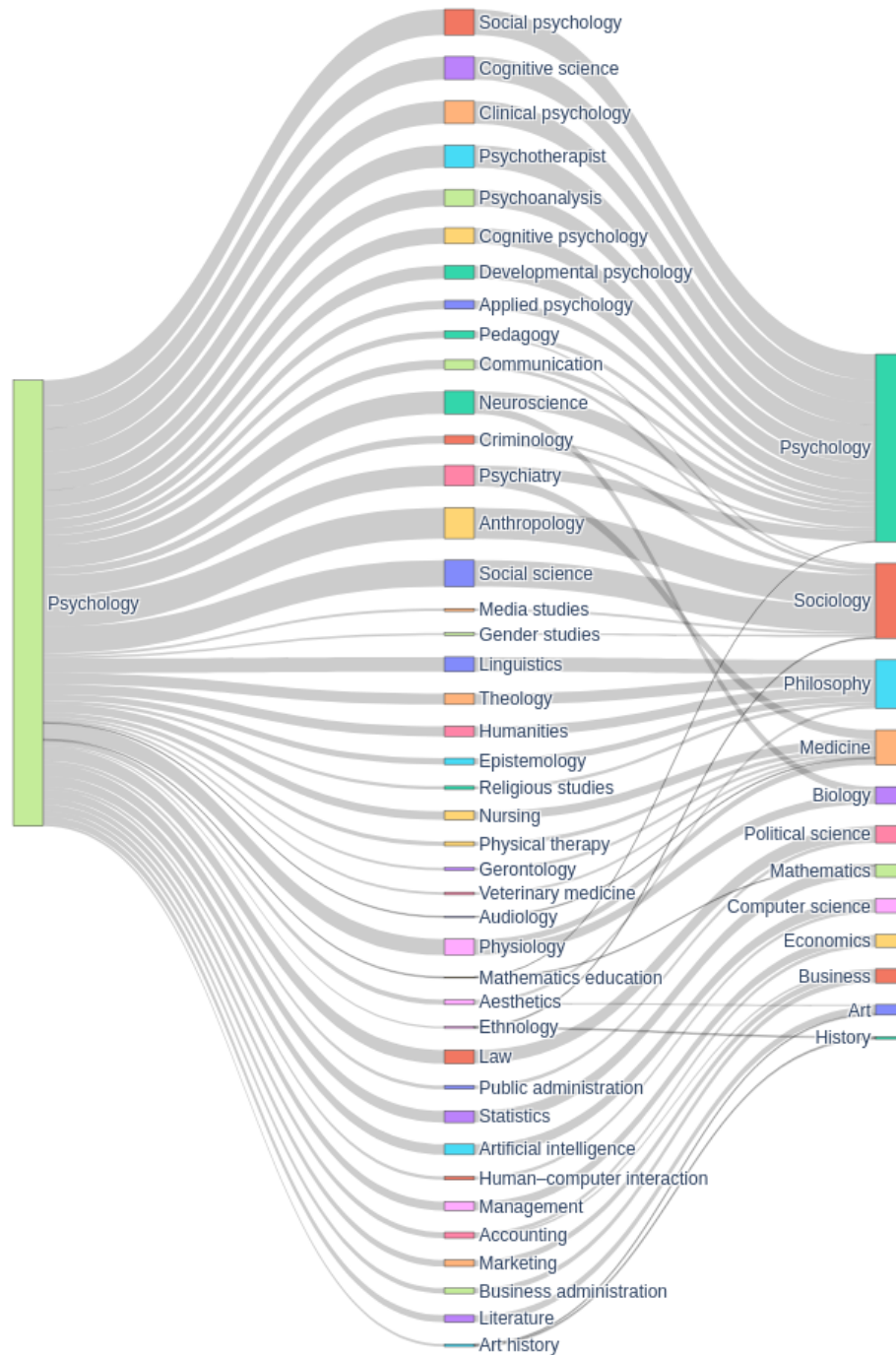


Fig. A-18 This Sankey plot visually shows the associations stemming from the level 0 concept Psychology (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept Psychology, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Sociology Level 0 Concept Sankey Plot

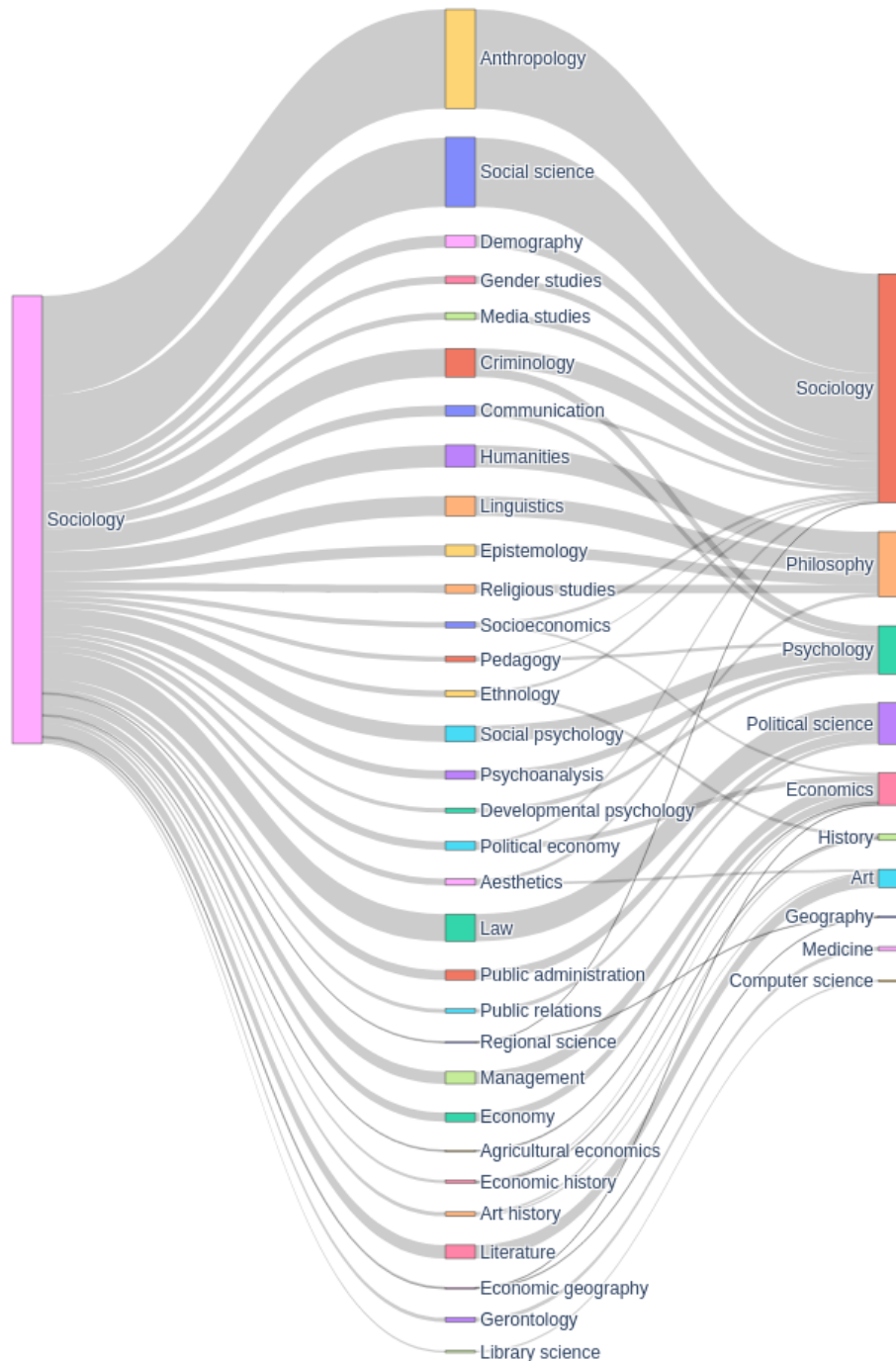


Fig. A-19 This Sankey plot visually shows the associations stemming from the level 0 concept **Sociology** (positioned on the left) to its level 1 counterpart concepts (displayed in the middle). These level 1 concepts are subsequently linked back to their ancestral level 0 concepts. The line widths in this network reflect the scores defining the initial similarity between the level 1 concepts and the level 0 concept **Sociology**, which are then appropriately scaled to establish connections back to their level 0 ancestor concepts.

Appendix B. Title Hype Word Comparisons by Hype Category

Table B-1 Attitude category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
71	outstanding	attitude	title	-84	-53	32	[66, 71, 64]
79	remarkable	attitude	title	-100	-61	39	[112, 79, 59]
95	incredible	attitude	title	-100	-86	14	[112, 95, 77]
105	surprising	attitude	title	-100	-96	4	[112, 105, 87]
107	confident	attitude	title	-100	-96	4	[112, 107, 89]
113	exciting	attitude	title	-88	-100	-12	[71, 126, 125]
114	interesting	attitude	title	-97	-100	-3	[82, 126, 122]
132	attractive	attitude	title	-100	-100	0	[112, 126, 107]
133	intriguing	attitude	title	-100	-100	0	[112, 126, 107]
134	notable	attitude	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-2 Importance category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
24	foundational	importance	title	554	571	17	[13, 24, 74]
25	fundamental	importance	title	356	497	141	[15, 25, 35]
38	critical	importance	title	59	212	153	[25, 38, 34]
39	strategic	importance	title	166	142	-23	[19, 39, 129]
44	timely	importance	title	-100	104	204	[112, 44, 29]
58	vital	importance	title	-83	-26	58	[65, 58, 50]
60	urgent	importance	title	-68	-33	36	[52, 60, 61]
63	key	importance	title	-83	-35	48	[64, 63, 54]
66	imperative	importance	title	0	-43	-43	[29, 66, 131]
78	essential	importance	title	-76	-60	16	[56, 78, 76]
84	major	importance	title	-65	-70	-5	[50, 84, 123]
87	ultimate	importance	title	-77	-72	5	[57, 87, 84]
88	pivotal	importance	title	-100	-72	28	[112, 88, 69]
98	significant	importance	title	-93	-91	2	[78, 98, 91]
101	important	importance	title	-92	-93	-2	[76, 101, 120]
104	crucial	importance	title	-100	-95	5	[112, 104, 83]
115	compelling	importance	title	-100	-100	0	[112, 126, 107]
116	indispensable	importance	title	-100	-100	0	[112, 126, 107]
117	invaluable	importance	title	-100	-100	0	[112, 126, 107]
118	paramount	importance	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-3 Novelty category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
12	novel	novelty	title	1,040	1,481	441	[7, 12, 20]
19	emerging	novelty	title	82	853	771	[24, 19, 16]
20	innovative	novelty	title	611	840	228	[12, 20, 27]
43	unprecedented	novelty	title	-80	106	186	[59, 43, 30]
50	unique	novelty	title	-55	25	80	[46, 50, 46]
51	unparalleled	novelty	title	-100	21	121	[112, 51, 38]
55	creative	novelty	title	-29	-10	18	[40, 55, 73]
85	revolutionary	novelty	title	-91	-71	20	[75, 85, 71]
91	first	novelty	title	-86	-78	8	[68, 91, 79]
103	latest	novelty	title	-53	-95	-42	[44, 103, 130]
119	groundbreaking	novelty	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-4 Problem category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
48	unmet	problem	title	-100	62	162	[112, 48, 33]
64	elusive	problem	title	-45	-37	8	[42, 64, 80]
77	dire	problem	title	-100	-60	40	[112, 77, 58]
81	stark	problem	title	42	-65	-106	[27, 81, 134]
83	alarming	problem	title	-100	-69	31	[112, 83, 65]
135	daunting	problem	title	-100	-100	0	[112, 126, 107]
136	desperate	problem	title	-100	-100	0	[112, 126, 107]
137	devastating	problem	title	-100	-100	0	[112, 126, 107]
138	dismal	problem	title	-100	-100	0	[112, 126, 107]
139	unanswered	problem	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-5 Quality category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
8	stellar	quality	title	3,121	2,551	-570	[2, 8, 139]
30	talented	quality	title	243	352	110	[17, 30, 41]
40	skilled	quality	title	-90	116	206	[73, 40, 28]
45	exceptional	quality	title	-28	74	103	[39, 45, 43]
59	premier	quality	title	-100	-31	69	[112, 59, 48]
62	qualified	quality	title	-87	-35	53	[69, 62, 53]
65	dedicated	quality	title	-60	-41	19	[47, 65, 72]
70	successful	quality	title	-85	-48	37	[67, 70, 60]
74	longstanding	quality	title	-100	-58	42	[112, 74, 57]
75	motivated	quality	title	-62	-59	2	[48, 75, 90]
82	ambitious	quality	title	-100	-66	34	[112, 82, 62]
86	promising	quality	title	-100	-71	29	[112, 86, 68]
92	vibrant	quality	title	-100	-80	20	[112, 92, 70]
96	intellectual	quality	title	-71	-89	-18	[54, 96, 127]
97	senior	quality	title	9	-89	-98	[28, 97, 133]
111	experienced	quality	title	-96	-98	-3	[79, 111, 121]
129	collegial	quality	title	-100	-100	0	[112, 126, 107]
130	prestigious	quality	title	-100	-100	0	[112, 126, 107]
131	renowned	quality	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-6 Rigor category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
6	robust	rigor	title	2,225	2,744	519	[4, 6, 19]
9	reproducible	rigor	title	-23	2,261	2,284	[36, 9, 5]
16	advanced	rigor	title	879	1,063	184	[8, 16, 31]
26	scientific	rigor	title	745	470	-275	[9, 26, 138]
29	rigorous	rigor	title	-25	379	404	[37, 29, 22]
33	accurate	rigor	title	-21	277	298	[35, 33, 25]
35	cohesive	rigor	title	-1	245	245	[30, 35, 26]
36	systematic	rigor	title	410	230	-179	[14, 36, 136]
41	quality	rigor	title	-9	112	121	[31, 41, 39]
68	strong	rigor	title	-39	-46	-7	[41, 68, 124]
89	detailed	rigor	title	-11	-74	-63	[32, 89, 132]
90	sophisticated	rigor	title	-82	-77	5	[63, 90, 82]
100	powerful	rigor	title	-79	-93	-14	[58, 100, 126]
102	careful	rigor	title	-100	-94	6	[112, 102, 81]
120	nuanced	rigor	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-7 Scale category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
3	interdisciplinary	scale	title	2,374	5,591	3,216	[3, 3, 4]
4	transdisciplinary	scale	title	-100	4,711	4,811	[112, 4, 3]
5	multidisciplinary	scale	title	2,027	3,882	1,855	[5, 5, 8]
21	diverse	scale	title	-27	782	809	[38, 21, 15]
27	massive	scale	title	44	386	342	[26, 27, 24]
28	international	scale	title	655	380	-274	[11, 28, 137]
34	comprehensive	scale	title	169	268	99	[18, 34, 44]
47	fastest	scale	title	-69	63	131	[53, 47, 37]
57	expansive	scale	title	98	-20	-118	[23, 57, 135]
61	broad	scale	title	-81	-33	47	[60, 61, 55]
69	deeper	scale	title	-100	-47	53	[112, 69, 52]
73	largest	scale	title	-100	-58	42	[112, 73, 56]
76	extensive	scale	title	-89	-59	30	[72, 76, 67]
80	immediate	scale	title	-97	-64	33	[81, 80, 63]
94	top	scale	title	-88	-83	5	[70, 94, 86]
106	biggest	scale	title	-100	-96	4	[112, 106, 88]
108	vast	scale	title	-100	-98	2	[112, 108, 92]
109	greatest	scale	title	-100	-98	2	[112, 109, 93]
110	huge	scale	title	-100	-98	2	[112, 110, 94]
112	ample	scale	title	-81	-100	-19	[61, 126, 128]
121	considerable	scale	title	-100	-100	0	[112, 126, 107]
122	enormous	scale	title	-100	-100	0	[112, 126, 107]
123	immense	scale	title	-100	-100	0	[112, 126, 107]
124	interprofessional	scale	title	-100	-100	0	[112, 126, 107]
125	myriad	scale	title	-100	-100	0	[112, 126, 107]
126	overwhelming	scale	title	-100	-100	0	[112, 126, 107]
127	substantial	scale	title	-100	-100	0	[112, 126, 107]
128	tremendous	scale	title	-100	-100	0	[112, 126, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table B-8 Utility category: hype word comparison in NSF titles

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
1	scalable	utility	title	5,191	32,592	27,401	[1, 1, 1]
2	deployable	utility	title	1,116	8,681	7,565	[6, 2, 2]
7	synergistic	utility	title	698	2,673	1,975	[10, 7, 7]
10	sustainable	utility	title	-15	2,119	2,135	[33, 10, 6]
11	efficient	utility	title	248	1,883	1,635	[16, 11, 9]
13	actionable	utility	title	-100	1,467	1,567	[112, 13, 10]
14	generalizable	utility	title	-100	1,398	1,498	[112, 14, 11]
15	transformative	utility	title	-100	1,233	1,333	[112, 15, 12]
17	seamless	utility	title	128	999	871	[21, 17, 14]
18	impactful	utility	title	-100	902	1,002	[112, 18, 13]
22	tailored	utility	title	122	660	538	[22, 22, 18]
23	durable	utility	title	-50	601	651	[43, 23, 17]
31	accessible	utility	title	-82	330	411	[62, 31, 21]
32	user-friendly	utility	title	-100	286	386	[112, 32, 23]
37	intuitive	utility	title	159	224	66	[20, 37, 49]
42	effective	utility	title	-55	108	163	[45, 42, 32]
46	safer	utility	title	-71	64	135	[55, 46, 36]
49	relevant	utility	title	-65	45	109	[49, 49, 42]
52	tangible	utility	title	-100	12	112	[112, 52, 40]
53	rich	utility	title	-20	10	31	[34, 53, 66]
54	productive	utility	title	-93	6	98	[77, 54, 45]
56	meaningful	utility	title	-91	-19	72	[74, 56, 47]
67	efficacious	utility	title	-100	-46	54	[112, 67, 51]
72	ideal	utility	title	-65	-55	10	[51, 72, 78]
93	ready	utility	title	-96	-81	16	[80, 93, 75]
99	easy	utility	title	-98	-93	5	[83, 99, 85]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Appendix C. Abstract Hype Word Comparisons by Hype Category

Table C-1 Attitude category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
47	outstanding	attitude	abstract	68	194	125	[58, 47, 50]
54	exciting	attitude	abstract	139	153	15	[37, 54, 87]
81	intriguing	attitude	abstract	96	59	-37	[51, 81, 124]
107	attractive	attitude	abstract	-9	-12	-3	[86, 107, 113]
112	remarkable	attitude	abstract	-21	-23	-2	[90, 112, 111]
121	notable	attitude	abstract	-53	-53	0	[112, 121, 102]
123	interesting	attitude	abstract	33	-56	-89	[70, 123, 132]
126	surprising	attitude	abstract	-54	-65	-10	[113, 126, 115]
133	incredible	attitude	abstract	-95	-80	15	[132, 133, 86]
135	confident	attitude	abstract	-86	-86	0	[130, 135, 105]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-2 Importance category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
18	fundamental	importance	abstract	622	1,027	406	[8, 18, 23]
21	foundational	importance	abstract	617	947	329	[9, 21, 29]
34	critical	importance	abstract	143	406	263	[35, 34, 35]
40	timely	importance	abstract	64	328	264	[62, 40, 34]
45	key	importance	abstract	82	198	116	[52, 45, 51]
53	crucial	importance	abstract	142	159	16	[36, 53, 85]
60	significant	importance	abstract	60	138	77	[63, 60, 63]
62	essential	importance	abstract	103	131	28	[47, 62, 76]
63	urgent	importance	abstract	-44	130	174	[107, 63, 41]
65	invaluable	importance	abstract	147	113	-34	[32, 65, 123]
66	major	importance	abstract	107	110	3	[45, 66, 97]
71	paramount	importance	abstract	-29	87	116	[96, 71, 52]
75	important	importance	abstract	216	78	-138	[25, 75, 138]
80	vital	importance	abstract	1	62	61	[77, 80, 66]
82	strategic	importance	abstract	-33	59	91	[101, 82, 61]
88	pivotal	importance	abstract	115	45	-70	[44, 88, 129]
93	compelling	importance	abstract	-63	31	93	[116, 93, 60]
95	ultimate	importance	abstract	78	23	-54	[55, 95, 128]
102	indispensable	importance	abstract	-19	1	20	[89, 102, 83]
103	imperative	importance	abstract	-22	-1	21	[91, 103, 82]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-3 Novelty category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
13	innovative	novelty	abstract	422	1,192	771	[12, 13, 16]
19	novel	novelty	abstract	370	1,000	630	[16, 19, 18]
24	unprecedented	novelty	abstract	67	822	755	[60, 24, 17]
29	emerging	novelty	abstract	170	540	370	[29, 29, 26]
35	unique	novelty	abstract	325	390	65	[18, 35, 65]
41	unparalleled	novelty	abstract	67	231	164	[59, 41, 45]
48	groundbreaking	novelty	abstract	-74	192	266	[123, 48, 33]
96	latest	novelty	abstract	-5	23	27	[82, 96, 78]
113	creative	novelty	abstract	-46	-28	17	[108, 113, 84]
116	first	novelty	abstract	-24	-39	-14	[93, 116, 118]
117	revolutionary	novelty	abstract	-65	-44	21	[117, 117, 81]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-4 Problem category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
22	unmet	problem	abstract	-71	829	899	[121, 22, 15]
42	elusive	problem	abstract	40	214	174	[65, 42, 42]
43	unanswered	problem	abstract	223	210	-13	[24, 43, 117]
78	devastating	problem	abstract	-71	65	135	[120, 78, 47]
89	daunting	problem	abstract	-52	44	95	[111, 89, 58]
119	alarming	problem	abstract	-80	-52	28	[126, 119, 77]
120	stark	problem	abstract	27	-53	-79	[71, 120, 130]
125	dire	problem	abstract	-90	-62	28	[131, 125, 75]
137	dismal	problem	abstract	-100	-98	2	[137, 137, 98]
138	desperate	problem	abstract	-99	-99	0	[134, 138, 107]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-5 Quality category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
9	longstanding	quality	abstract	307	1,518	1,211	[20, 9, 10]
16	stellar	quality	abstract	1,078	1,075	-4	[4, 16, 114]
31	talented	quality	abstract	130	509	379	[39, 31, 25]
32	promising	quality	abstract	468	469	2	[11, 32, 100]
38	premier	quality	abstract	5	344	339	[76, 38, 28]
51	motivated	quality	abstract	127	180	53	[42, 51, 67]
55	successful	quality	abstract	104	149	45	[46, 55, 69]
61	skilled	quality	abstract	-40	133	173	[105, 61, 44]
76	dedicated	quality	abstract	156	75	-81	[30, 76, 131]
83	senior	quality	abstract	97	58	-38	[49, 83, 126]
84	collegial	quality	abstract	171	57	-114	[28, 84, 135]
86	vibrant	quality	abstract	-76	50	126	[124, 86, 49]
91	exceptional	quality	abstract	10	41	31	[75, 91, 73]
94	qualified	quality	abstract	-1	29	31	[79, 94, 74]
108	experienced	quality	abstract	-15	-16	-2	[87, 108, 109]
109	renowned	quality	abstract	-31	-18	13	[99, 109, 88]
110	ambitious	quality	abstract	-5	-18	-13	[83, 110, 116]
122	prestigious	quality	abstract	-61	-55	6	[115, 122, 92]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-6 Rigor category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
14	reproducible	rigor	abstract	132	1,164	1,032	[38, 14, 12]
17	robust	rigor	abstract	722	1,038	316	[5, 17, 30]
20	scientific	rigor	abstract	680	981	301	[6, 20, 32]
23	rigorous	rigor	abstract	421	825	404	[13, 23, 24]
26	advanced	rigor	abstract	285	595	310	[21, 26, 31]
39	accurate	rigor	abstract	155	329	174	[31, 39, 40]
44	cohesive	rigor	abstract	36	209	173	[69, 44, 43]
49	sophisticated	rigor	abstract	187	185	-2	[26, 49, 110]
50	systematic	rigor	abstract	369	180	-189	[17, 50, 139]
58	detailed	rigor	abstract	248	141	-107	[23, 58, 133]
59	quality	rigor	abstract	25	140	115	[72, 59, 53]
77	nuanced	rigor	abstract	-29	68	97	[97, 77, 57]
97	powerful	rigor	abstract	127	8	-119	[41, 97, 136]
101	strong	rigor	abstract	39	1	-38	[67, 101, 125]
131	careful	rigor	abstract	-33	-76	-43	[102, 131, 127]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-7 Scale category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
2	interdisciplinary	scale	abstract	1,734	4,427	2,693	[1, 2, 3]
4	multidisciplinary	scale	abstract	1,556	3,062	1,506	[2, 4, 8]
5	transdisciplinary	scale	abstract	396	2,867	2,471	[14, 5, 4]
12	diverse	scale	abstract	390	1,382	992	[15, 12, 13]
36	broad	scale	abstract	146	373	228	[33, 36, 36]
37	comprehensive	scale	abstract	122	345	223	[43, 37, 37]
57	extensive	scale	abstract	172	142	-30	[27, 57, 121]
64	massive	scale	abstract	-4	126	130	[81, 64, 48]
67	myriad	scale	abstract	-3	105	108	[80, 67, 55]
69	tremendous	scale	abstract	-48	104	152	[109, 69, 46]
73	international	scale	abstract	73	83	10	[56, 73, 90]
74	fastest	scale	abstract	-15	78	94	[88, 74, 59]
79	substantial	scale	abstract	12	64	52	[74, 79, 68]
85	deeper	scale	abstract	67	51	-15	[61, 85, 119]
87	largest	scale	abstract	-29	46	75	[98, 87, 64]
98	expansive	scale	abstract	-36	7	43	[103, 98, 70]
104	vast	scale	abstract	-41	-2	39	[106, 104, 71]
105	ample	scale	abstract	-32	-7	25	[100, 105, 79]
106	enormous	scale	abstract	-8	-8	0	[85, 106, 104]
111	immediate	scale	abstract	-24	-18	5	[92, 111, 96]
115	immense	scale	abstract	-37	-37	0	[104, 115, 101]
118	considerable	scale	abstract	57	-51	-107	[64, 118, 134]
127	greatest	scale	abstract	-69	-67	2	[119, 127, 99]
128	top	scale	abstract	-72	-72	0	[122, 128, 106]
129	huge	scale	abstract	-81	-73	8	[128, 129, 91]
130	biggest	scale	abstract	-96	-73	23	[133, 130, 80]
132	overwhelming	scale	abstract	-85	-79	6	[129, 132, 93]
136	interprofessional	scale	abstract	40	-96	-136	[66, 136, 137]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

Table C-8 Utility category: hype word comparison in NSF abstracts

#	Hype Word	Hype Category	Title or Abstract	1988–1992 (% Increase) ¹	2018–2022	Relative % Difference	Rank.avg Order ²
1	scalable	utility	abstract	1,154	6,193	5,039	[3, 1, 1]
3	deployable	utility	abstract	96	4,153	4,057	[50, 3, 2]
6	generalizable	utility	abstract	614	2,477	1,863	[10, 6, 5]
7	transformative	utility	abstract	-100	1,673	1,773	[137, 7, 6]
8	user-friendly	utility	abstract	658	1,565	907	[7, 8, 14]
10	synergistic	utility	abstract	313	1,492	1,179	[19, 10, 11]
11	impactful	utility	abstract	-100	1,468	1,568	[137, 11, 7]
15	actionable	utility	abstract	-100	1,164	1,264	[137, 15, 9]
25	efficient	utility	abstract	269	745	476	[22, 25, 21]
27	tailored	utility	abstract	128	589	461	[40, 27, 22]
28	seamless	utility	abstract	-6	578	584	[84, 28, 19]
30	accessible	utility	abstract	144	509	365	[34, 30, 27]
33	sustainable	utility	abstract	-51	458	508	[110, 33, 20]
46	effective	utility	abstract	1	197	196	[78, 46, 38]
52	relevant	utility	abstract	80	178	98	[53, 52, 56]
56	durable	utility	abstract	-28	147	176	[95, 56, 39]
68	intuitive	utility	abstract	99	105	5	[48, 68, 95]
70	meaningful	utility	abstract	78	90	12	[54, 70, 89]
72	safer	utility	abstract	-26	85	111	[94, 72, 54]
90	rich	utility	abstract	38	43	5	[68, 90, 94]
92	productive	utility	abstract	71	40	-31	[57, 92, 122]
99	ideal	utility	abstract	23	3	-19	[73, 99, 120]
100	tangible	utility	abstract	-77	2	79	[125, 100, 62]
114	efficacious	utility	abstract	-68	-32	35	[118, 114, 72]
124	easy	utility	abstract	-61	-62	-1	[114, 124, 108]
134	ready	utility	abstract	-81	-83	-2	[127, 134, 112]

¹ The % increase is for NSF awards with respect to the baseline Google Books English 2019 corpus for each time period: 1988–1992 and 2018–2022.

² The rank.avg order is for the prior three % columns (out of 139 hype words).

List of Symbols, Abbreviations, and Acronyms

ACRONYMS:

AFC	Army Futures Command
AFOSR	Air Force Office of Scientific Research
AI	artificial intelligence
AMC	Army Materiel Command
API	application programming interface
ARL	Army Research Laboratory
ARO	Army Research Office
AFRL	Air Force Research Laboratory
BAA	broad agency announcement
BG	Brigadier General
CBOW	continuous bag of words
COVID-19	coronavirus disease 2019
CRA	collaborative research alliance
CRI	collaborative role index
DARPA	Defense Advanced Research Projects Agency
DEVCOM	US Army Combat Capabilities Development Command
DOD	Department of Defense
DOE	Department of Energy
DOI	digital object identifier
DTRA	Defense Threat Reduction Agency
ECI	egocentric collaboration index

FFRDC	federally funded research and development center
FY	fiscal year
GAN	generative adversarial networks
GloVe	global vectors for word representation
GPT	generative pre-trained transformer
HQ	headquarters
IARPA	Intelligence Advanced Research Projects Activity
ICSSI	International Conference on the Science of Science Innovation
ID	identifier
IOBT	Internet of Battlefield Things
IP	intellectual property
ISSN	international standard serial number
ISSN-L	linking international standard serial number
JMC	Joint Munitions Command
JSON	javaScript object notation
ML	machine learning
MLOps	machine learning operations
MMLS	Midwest Machine Learning Symposium
MURI	multidisciplinary university research initiative
NASA	National Aeronautics and Space Administration
NCEC	National Cybersecurity Education Colloquium
NDSEG	National Defense Science and Engineering Graduate
NIH	National Institutes of Health

NIST	National Institutes of Standards and Technology
NLTK	natural language toolkit
NNSA	National Nuclear Security Administration
NRC	Nuclear Regulatory Commission
NRL	Naval Research Laboratory
NSF	National Science Foundation
ODNI	Office of the Director of National Intelligence
OFR	open funder registry
ONR	Office of Naval Research
ORCID	open researcher and contributor ID
PI	principal investigator
PM	program manager
QR	quick response (code)
R&D	research and development
ROI	return on investment
ROR	research organization registry
STEM	science, technology, engineering, and mathematics
UARC	university affiliated research center
URL	uniform resource locator (web address)
wpm	words per million
XML	extensible markup language
YIP	young investigator program