The Development of Facticity—from Preliminary Findings to Accepted Implicit Knowledge: Case Studies[*]

Stanford CS224N Custom Project

Tianyu Du
ICME
Stanford University
tianyudu@stanford.edu

Yuze Sui
Department of Sociology
Stanford University
yuzesui@stanford.edu

Jingruo Sun
Department of MS&E
Stanford University
jingruo@stanford.edu

Abstract

This study examines the trajectory of new scientific ideas from their inception to widespread acceptance, framed within the sociology of science. We propose a phased model delineating the evolution of scientific concepts through initial publication, peer recognition (i.e., implicit knowledge), and eventual integration into the canonical body of knowledge. Utilizing data from the unarXiv database, we trace the lifecycle of scientific ideas, exemplified by the analysis of "Ricci flow" and "Lasso regression". Our methodology encompasses citation analysis, co-citation network construction, and key-phrase extraction to elucidate the dynamics of scientific innovation and consensus formation. We identify a multi-phase process where ideas transition from (1) being contentious and highly cited to (2) becoming implicit (i.e., frequently used but less cited), and eventually (3) accepted facts represented in educational and reference materials. Our theory not only underscores the progressive nature of scientific knowledge but also reflects on the mechanisms of cognitive economy in science, where the consolidation of accepted knowledge frees cognitive resources for exploring novel concepts. Our findings contribute to understanding knowledge accumulation and the pivotal role of citation and conceptual association in the scientific discourse. In addition to existing research studying citation networks only, our paper offers holistic insights into the broader mechanisms of scientific advancement and the cyclic nature of innovation and consensus. In this project, we conducted case studies on two concepts, "Ricci flow" and "Lasso regression", to showcase the multi-phrase transitions in our theory. Moreover, we deployed a transformer-based key-phrase extraction pipeline to establish the "conceptual association" of focal concepts, which reveals trends in diversities and sentiments of phrases associated with focal concepts.

1 Introduction

Most research in the sociology of science and science of science literature focuses on the citation patterns in published journal articles or conference proceedings to study the emergence and survival of new scientific ideas and concepts. In published papers, we observe the "bleeding edge" of scientific innovation, where preliminary findings are related, and if "lucky", ensuing works will be addressed. It is here that prior literature argues science is slowing down and in need of disruption—citation patterns indicate that ideas in new papers are increasingly likely to build upon previous literature rather than diverting future research into different directions (Chu and Evans, 2021; Park et al., 2023). This line of work assumes cited papers (and their contents) contribute to knowledge.

Are new concepts and their associations accepted when published? When do new ideas become accepted facts and extend the body of scientific knowledge? Unfortunately, a new publication and its idea are often dead on arrival, never to be cited or used (Cheng et al., 2023). Thus, it resembles a localized and temporary disruption to the knowledge graph sustained by papers and citations. No one would call it a knowledge advance. But let's say the concept in the paper is cited and reused; is it now accepted as a contribution to the progress of knowledge and science? Citation is no guarantee of consensus, and in fact, their ideas and concepts are contentious (Shwed and Bearman, 2010). Science, after all, is replete with (1) extended disputes over facts (for example, it took a decade for chemistry to accept the discovery of oxygen (Kuhn, 1962)) and (2) scientific dead ends represented in published works, such as the concept of phrenology in the nineteenth century, which links personality traits with scalp morphology.

[*] Authors contribute equally and are ordered alphabetically based on authors' last names. External mentor: Professor Daniel A. McFarland (Graduate School of Education). Tianyu mainly worked on managing the dataset and deploying the pipeline to extract key phrases. Yuze focused on composing the theory part, literature review, composing the report, as well as illustrating the concept association. Jingruo worked on the literature review, sentiment analysis of co-used phrases, and generating figures for illustration.

We argue that newly published ideas and concepts get used in ways that reveal their transition from initial knowledge claims and loci of dispute to more implicitly held and accepted facts (also see Figure 1). When new ideas and concepts are first proposed, only some succeed in being published in peer-reviewed journals (phase 0), and only a small subset of those published get used in ensuing papers and garner explicit attention and citation (phase 1). They are a focus of concern (as per citation) and need elaboration (as per elaboration of the concept's connection to ideas to other extent). This recognition may be from competing camps that lack consensus, especially if the citations to them have modular co-usage (Lakatos, 1978). Over time, the idea develops more stable associations and loses salience in its citation network, reflecting increasing consensus (Shwed and Bearman, 2010).

At some point (phase 2), the concept becomes a category like "recurrent neural network (RNN)" in the machine learning literature. After becoming a category, it receives a diminished citation count but increased usage. Simply put, the idea finds acceptance and consensus but loses its focus as an explicit concern. It becomes a transportable package of associations, like a boundary object (Star and Griesemer, 1989), method or machine (Latour, 1987). With more time (phase 3), the idea and its established relations to other ideas may be transferred into textbooks, encyclopedias, and vademecum. There, they become agreed-upon facts that are background knowledge and implicit conceptual associations to which journal science tends to conform. In some cases, a concept's semantic embedding (and category use) is further encapsulated in a proper noun phrase, effectively blackboxing and packaging what is represented in the textbook and implicitly referring to it in the journal's text by use of that phrase (e.g., Newton's first law of motion). In this last phase, the original concept and its relations may find diminished use within journals, but the concept has not disappeared. If anything, the idea's movement into different textbooks denotes increased acceptance and established contribution.

Of course, the aforementioned multi-phase process can be reversed. When black-boxed facts are unpacked, such as when new findings challenge accepted facts, the concept and its associations return to bleeding-edge research. It comes back into use and explicit focus (citation), and there we see rewiring and disruption more indicative of scientific revolution (albeit almost always in modest, local forms)—should the changes hold and get transferred back into implicit knowledge of the encyclopedia. Only in vast swaths of time might we observe the scientific revolutions in classic works (Kuhn, 1962). Last, we contend the process of conceptual usage and change is core to knowledge accumulation and enables increased knowledge capacity and storage in science overall. This is similar to the concept of "attention allocation" in the organizational sociology literature (March et al., 2000). The cognitive resources of an organization are limited, and it has to convert routines into implicit knowledge and allocate attention to contingencies. Similarly, by actively converting consensus into implicit knowledge, science saves its cognitive bandwidth by focusing on contentious new ideas/concepts in published papers.

Definition: Throughout this manuscript, a **concept** refers to an "abstract idea corresponding to themes and realities" or observations about the world, be it the social world, or physics world, or the mathematical world.

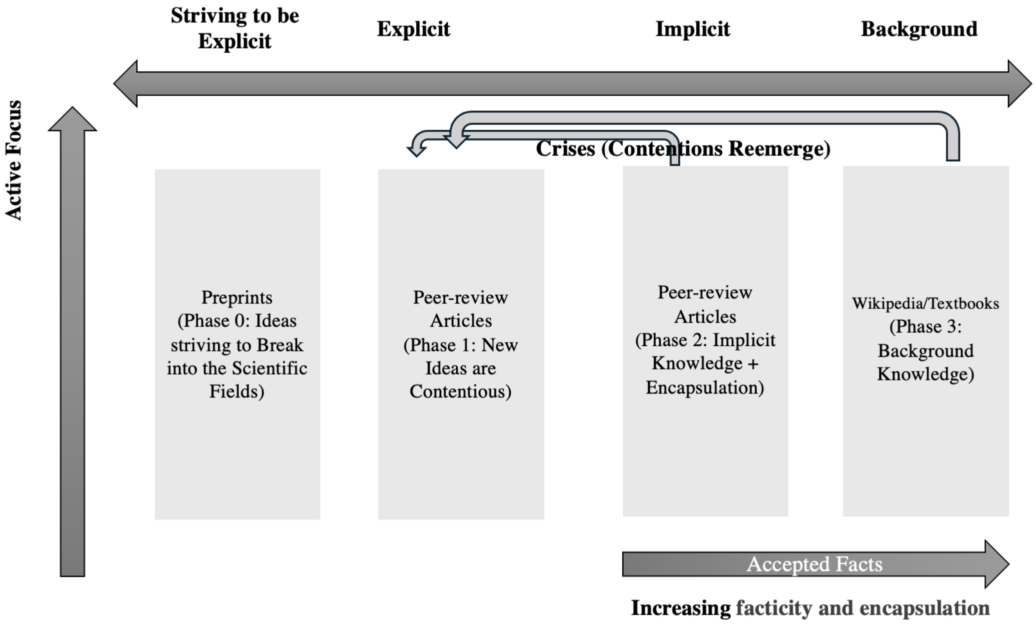
Striving to be Explicit
Explicit
Implicit
Background
Active Focus
Crises (Contentions Reemerge)
Preprints (Phase 0: Ideas striving to Break into the Scientific Fields)
Peer-review Articles (Phase 1: New Ideas are Contentious)
Peer-review Articles (Phase 2: Implicit Knowledge + Encapsulation)
Wikipedia/Textbooks (Phase 3: Background Knowledge)
Accepted Facts
Increasing facticity and encapsulation

Figure 1: Scheme Plot for Our Theoretical Framework

2 Related Work

Scientific Progression One major stream of literature in organizational innovation claims that technological progress is about radical changes and thus is **disruptive** (Christensen, 2015; Henderson and Clark, 1990)). The larger literature on innovation shares this narrative. For example, Kuhn (1962)'s "paradigm shift" argument states that scientific progress is fueled by rounds of revolutions that eradicate the current accepted paradigms. A recent (and rising) stream of literature in the sociology of science and science of science literature claims that **new scientific ideas are increasingly likely to build upon previous literature rather than diverting future research into different directions** (Chu and Evans, 2021; Park et al., 2023). In short, according to this stream of literature, **scientific innovation** is slowing down and we need to worry about the **scientific progress** of our society. This recent stream of literature offers a compelling perspective and framework on examining scientific progress. However, the framework largely depends on citation-based measure and the under-emphasis on textual analyses makes the framework potentially incomplete. We believe that to better study scientific progress, it is important to dive into the specific contents of the knowledge and trace their lifespan beyond the scope of citation patterns.

Concept Citation Identification Jurafsky and Martin (2008) constructed concept citations using the citation network among papers; they first identified concepts that appeared in each publication, and a concept receives citations whenever the paper including this concept gets cited. Cheng et al. (2023) utilized an automatic phrase extraction algorithm to retrieve their list of concepts from keyword lists and abstracts of publications in the Web of Science database.

Key-Phrase Extraction The objective of key-phrase extraction is to identify the small set of words that can summarize the main contents, features, genres, and concepts of a relatively long document. Extracting key phrases from long documents such as academic publications has been studied in the NLP literature for decades. Early works by Frank et al. (1999) and Turney (2002) compared the performances of adaptive learning algorithms (e.g., Bayesian learning) against traditional rule-based algorithms. Later work by Nguyen and Kan (2007) introduced features capturing phrases' positions in the document and salient morphological phenomena to identify better acronyms. Siddiqi and Sharan (2015) provided a comprehensive survey of machine learning techniques for key-phrase extraction. More recent works by Kulkarni et al. (2021) introduced a new objective for pre-training language models. The proposed scheme enables researchers to leverage the recent advances in large language models and build LLMs specialized for key-phrase extraction. Hulth (2003) demonstrates that integrating linguistic knowledge into the key-phrase extraction pipeline improves the precision and effectiveness of the extraction compared to traditional statistical methods (e.g., n-grams). To test models' performances, the author introduced an `Inspec` database consisting of 2,000 abstracts of scientific publications from the Web of Science Inspec database. The dataset covers papers from *Computers and Control* and *Information Technology* domains published between 1998 and 2002. Industry practitioners are actively developing LLMs for key-phrase extractions right now. For example, `ml6team` have fine-tuned the general-purpose `distill-bert` by Sanh et al. (2019) on the `Inspec` database and built a LLM specialized for detecting phrases from academic publications.

Sentiment Analysis A range of studies in the NLP literature have explored the sentiment analysis of phrases. Early works by Nasukawa and Yi (2003) developed a pipeline for detecting sentiment at both the message and term levels, which achieved a high precision in detecting sentiments on news articles and webpages. Giatsoglou et al. (2017) explored the effectiveness of machine learning methods in sentiment analysis; the authors used various methods based on the lexicon, word embedding, and hybrid vectorization to represent documents as vectors. The recent survey by Bordoloi and Biswas (2023) presents a comprehensive overview of developments in sentiment analysis methods, key components for designing effective sentiment analysis systems, and potential interdisciplinary research directions about sentiment analysis.

3 Approach

3.1 Publication Networks and Conceptual Associations

We trace life trajectories of scientific concepts by studying properties of (1) the network of papers citing the focal concept, (2) the network of papers mentioning the focal concept, and (3) the group of phrases co-used with the focal concept.

Papers Citing the Concept Given a concept $c \in \mathcal{C}$ from our set of concepts, we aim to construct a network of papers citing the concept c. There are several ways to operationalize such citation patterns. We identify the paper originally proposed the concept, denoted as P_c. We refer to the paper that originally proposed the focal concept as the **seed paper** of the focal concept. This step is particularly challenging to scale up since it requires a manual search; we focus on a few cases in this project and plan to scale up in upcoming parts of our research. In this project, we focused on trajectories of "Ricci flow" originally proposed by Hamilton (1982) and "Lasso regression" proposed by Tibshirani

(1996). We can then construct a network of papers that cite the corresponding seed paper for each concept. We let $\mathcal{P}_c^{\text{citation}}$ denote the set of papers citing seed paper of concept c.

Papers Mentioning the Concept Prior literature often relied on the citation count to proxy the popularity of a concept, and our theory suggests solely looking at the citation pattern could result in misleading conclusions. To establish a more robust metric of a concept's popularity, we identify the set of papers mentioning the focal concept by searching for the occurrences of different aliases of the focal concept in different papers. Specifically, we look for all papers mentioning the phrase "Ricci flow" and all papers mentioning the phrase "lasso" or "least absolute shrinkage and selection operator" to retrieve the sets of papers we want. We converted all letters to lowercase before conducting the phrase search. We use $\mathcal{P}_c^{\text{mention}}$ to denote the set of papers that mentioned the concept c in their full texts.

Size of Network We start with analyzing the sizes of the two networks we constructed, namely $|\mathcal{P}_c^{\text{citation}}|$ and $|\mathcal{P}_c^{\text{mention}}|$. As we will discuss later, the discrepancy between growth trends in $|\mathcal{P}_c^{\text{citation}}|$ and $|\mathcal{P}_c^{\text{mention}}|$ will indicate that solely focusing on the citation network provides an incomplete picture about concept's life trajectory.

Network Modularity The salience of the publication network surrounding the focal concept informs the maturity and consensus of the concept. Specifically, prior literature suggests the usage of network modularity as a proxy to the salience: "it follows that changes in a citation network's community structure represent changes in consensus levels on an issue: Contentious networks are well defined by communities, and consensual networks are not. Consensus formation exhibits a decline in community salience" Shwed and Bearman (2010). Modularity measures the strength of the division of a network into modules/communities. Thus, higher modularity indicates more evidence of fragmentation and thus represents contentions for citation networks.

Key Phrase Extractions We deployed the `keyphrase-extraction-distilbert-inspec` model by the ml6team on Hugging Face. The model leverages a distilled version of BERT (Sanh et al., 2019) fine-tuned on the Inspec dataset, which includes 2000 English scientific papers from the domains of Computers, Control, and Information Technology. Designed specifically for keyphrase extraction, this model classifies each word in a document as either the beginning of a keyphrase, inside a keyphrase, or outside a keyphrase, thereby enabling the automated extraction of phrases that encapsulate the core ideas of the text. The transformer model was pre-trained with a novel pre-training objective called Keyphrase Boundary Infilling with Replacement (KBIR) proposed by Kulkarni et al. (2021). The KBIR objective is a multi-task learning objective consisting of two separate tasks: the Keyphrase Boundary Infilling (KBI) task, which learns meaningful span representations of key phrases, and the Keyphrase Replacement Classification (KRC) task, which specializes in identifying key phrases within the context of a text input. We utilize the specialized `keyphrase-extraction-distilbert-inspec` model to extract key-phrases in different sections of papers in the $\mathcal{P}_c^{\text{mention}}$ set.

Conceptual Association We examined how the set of key phrases used together with the focal concept evolves. Specifically, we used word cloud, a visual representation of text data where the size of each word indicates its frequency or importance within a given corpus, to provide an intuitive overview of key themes or topics associated with the focal concept.

Key Phrase Sentiments Beyond studying the frequencies of co-used phrases, we investigated the sentiments of these phrases in different years. We calculate the sentiment scores using a model called Valence Aware Dictionary and Sentiment Reasoner (VADER), which contains a lexicon labeled according to their semantic orientation as either positive or negative. We not only look at the individual words but also consider the context of the words to determine their sentiment. In this way, we compute three separate scores of positive, negative, and neutral sentiments and one overall score of compound sentiment for each paper. The three separate scores are normalized to ensure their sum up to 1, and the compound score is normalized to the range between -1 and 1.

4 Experiments

4.1 Data

We demonstrate this phased process empirically with three data sources. First, the unarXiv database, built based on the official arXiv bulk data, contains 1.9 million scientific manuscripts in STEM fields (Saier et al., 2023). This database offers the full texts of the manuscripts in addition to abstracts. In this research, we focus on mathematics and computer science because of the convention of sharing works on arXiv before submitting them to peer-review venues in these two fields.

4.2 Experimental Details

In our case studies, we identified 3,140 publications in the unarXiv database mentioning the concept of Ricci Flow, and there are 316,597 unique key phrases in the full texts of these papers. In contrast, the "user-based" of Lasso is much larger; we identified 14,807 papers explicitly mentioning the phrase "Lasso" and its variants from the database, which coincides with our knowledge that there is a much larger group of researchers working on statistical learning compared to that of differential geometry. Our transformer model identified 2,243,033 unique key phrases using full-texts of these publications. Figure 2 presents the number of key phrases in full-texts/abstracts of publications that mention the corresponding focal concept. On average, we retrieved 5.5 (Ricci papers) and 10.9 (Lasso papers) unique key phrases from abstracts. The full text of a paper contains 229 (Ricci papers) and 318 (Lasso papers) keywords on average. We observe a huge difference between amounts of information in the abstract and full-text, which implies the necessity of analysis paper texts.

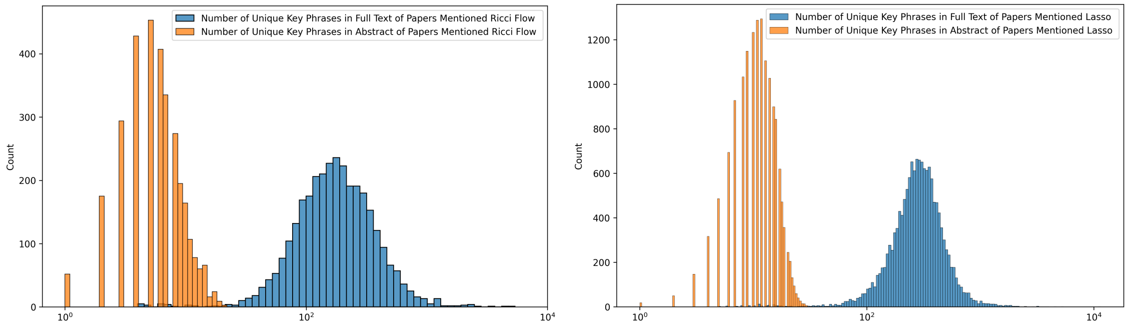

Figure 2: Numbers of key phrases in full-texts/abstracts of publications mentioned the focal concept.

4.3 Results & Analysis

4.3.1 Citation to "Mention but not Cited" Transition

Figure 3 presents the trajectory of the focal term "Ricci flow" in the mathematical fields of differential geometry and geometric analysis. We observe differences among the numbers of papers that (1) cited the original Ricci flow paper published in 1982, (2) mentioned the term "Ricci flow" in their abstracts, and (3) mentioned the term "Ricci flow" in their abstracts *OR* body texts. This example indicates that focusing solely on the citation pattern of the focal term only provides an incomplete picture of the term's life trajectory. Specifically, we observe the divergence between citation and usage patterns, especially after 2008. Further, we observe that the number of papers citing the original "Ricci flow" paper (red) follows a similar trend to the number of papers mentioning the focal term in the abstract (orange). Still, these two trends drastically differ from the full-text mentioning curve (green). Such an observation suggests that it is insufficient to look at abstract texts while studying the life course of concepts. The blue curve and right-hand-side y-axis show the number of papers in the ArXiv database each year, as a reference. Figure 4 shows the same set of trends for the concept "Lasso regression"; still, we observed a discrepancy between citation counts and concept usage over the years.

4.3.2 Conceptual Associations

The top panel in Figure 5 shows key phrases co-used with the concept "Ricci flow" in 2007, 2017, and 2022. The bottom panel in Figure 5 presents key phrases that co-occurred with the concept "Lasso" in 2007, 2017, and 2022.

The left panel in Figure 6 shows the diversity, measured by Herfindahl Index, of the key phrases co-used with the concept "Ricci flow" between 2004 and 2022. A lower Hierfindahl Index indicates a higher level of diversity of conceptual association. The right panel in Figure 6 shows the diversity, measured by the Herfindahl Index, of the key phrases co-used with the concept "Lasso" between 2004 and 2022.

Our results on conceptual associations yield interesting patterns and highlight what NLP can contribute to the literature on science of science. In Figure 5, we see that, in 2007, the top associated concepts with "ricci flow" are "normalized ricci flow", "conformal sigma model", and "remannian manifold". In 2017, the top associated concepts with "ricci flow" change to "curvature", "heat equation" and "ricci curvature". In 2022, the top associated concepts with "ricci flow" further change into "generalized ricci flow", "gradient flow", and "singularity". In 2007, the top associated concepts with "lasso" are "sparsity pattern", "group lasso", and "sparsity recovery". In 2017, the top associated concepts with "lasso" change to "simulation", "machine learning" and "variable selection". In 2022, the top associated concepts with "lasso" are still "simulation", "machine learning" and "variable selection".

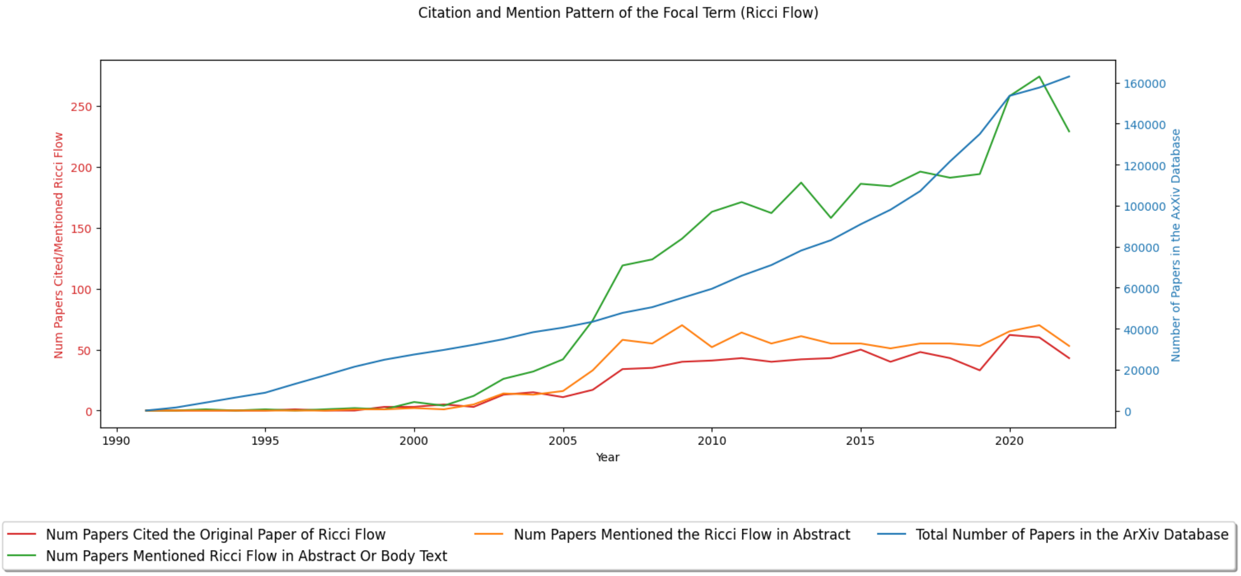

Figure 3: Citation and usage pattern of the focal concept “Ricci flow”

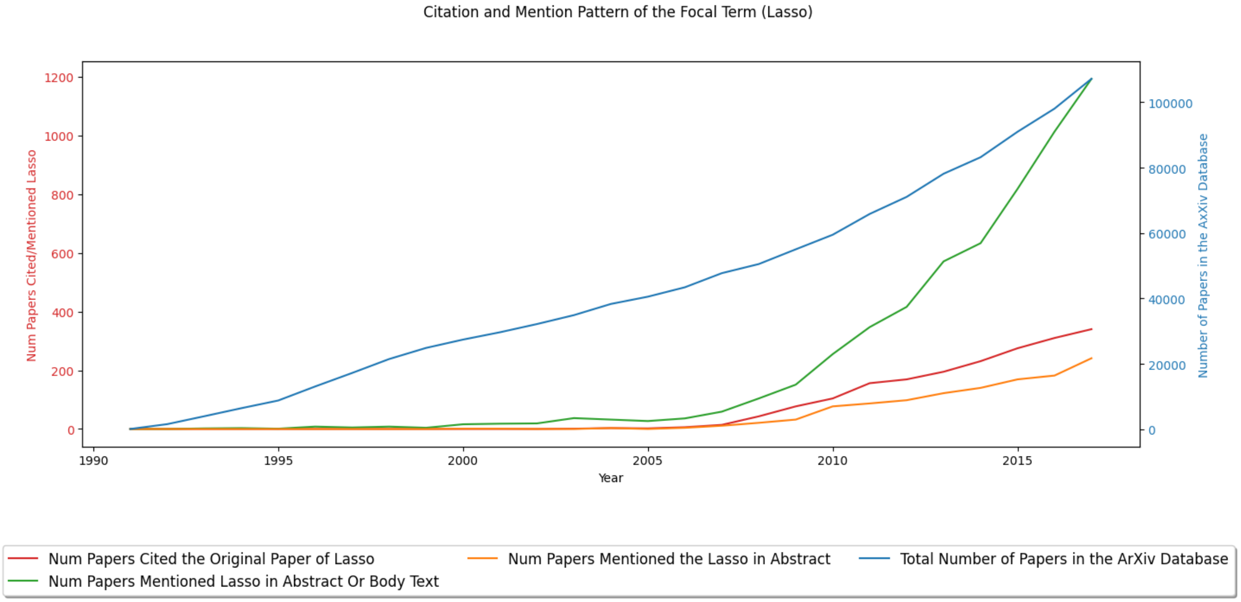

Figure 4: Citation and usage pattern of the focal concept “Lasso Regression”

We see that the concepts associated with “lasso” change drastically between 2007 and 2017. Especially, in 2007, the concepts associated with “lasso” were mainly statistical and mathematical concepts. However, in 2017, the associated concepts change into concepts in the machine learning field, which indicates that “lasso” transited from a statistical concept to a popular concept used by other fields (computer science). The pattern stabilizes in 2022, which indicates that “lasso” has since become a fundamental concept for subsequent work in machine learning. In Figure 5, we also see that the concepts associated with “Ricci flow” change slightly from 2007 to 2017 to 2022 and all associated concepts are mathematical concepts, which indicates a gradual progress in terms of knowledge accumualation in mathematics.

In Figure 6, we see that the diversity of conceptual association increases overtime for “Ricci” and “Lasso”. This indicates that after becoming implicit concepts and being recognized as accepted knowledge, “Ricci” and “Lasso” get be used for a wide range of literature and studies. Essentially, these two concepts serve as the foundation for a wide range of subsequent research and thus contribute to the scientif progress and knowledge accumulation.

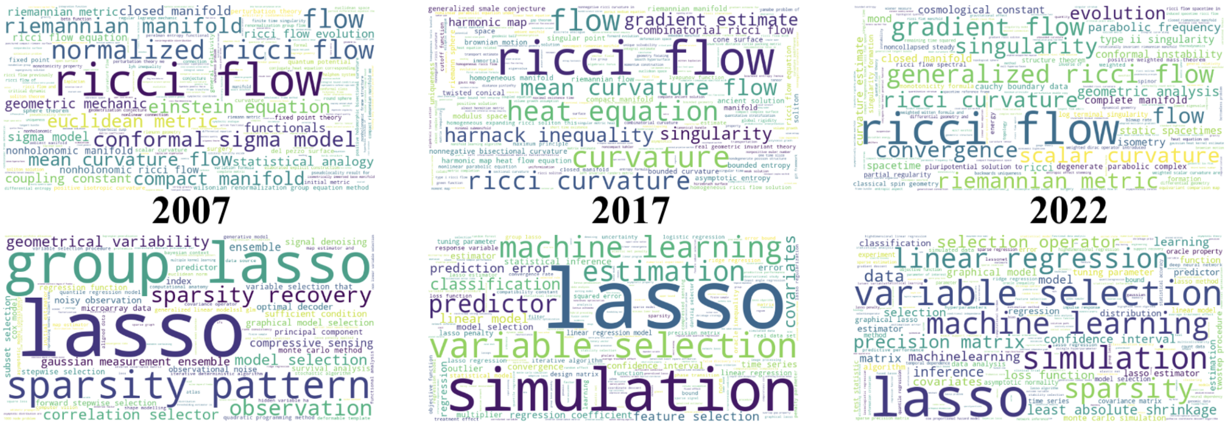

Figure 5: Word Cloud of Key Phrases Associated with the Focal Concepts in 2007, 2017, and 2022.

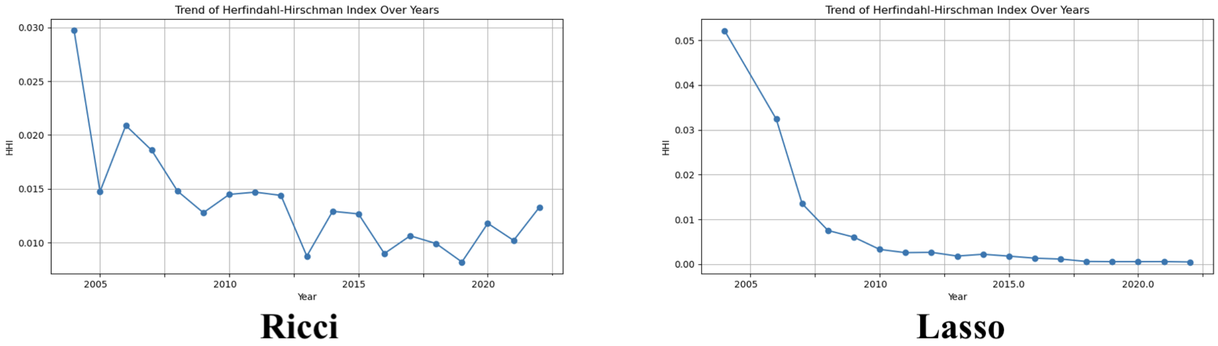

Figure 6: Herfindahl for Conceptual Associations (2004-2022)

4.3.3 Sentiment of Co-used Phrases

Figure 7 showcases the yearly trend of sentiments of key phrases associated with the concept “Lasso Regression” and Figure 8 presents the trend of sentiments of key phrases co-used with the concept “Ricci Flow” between 2004 and 2022.

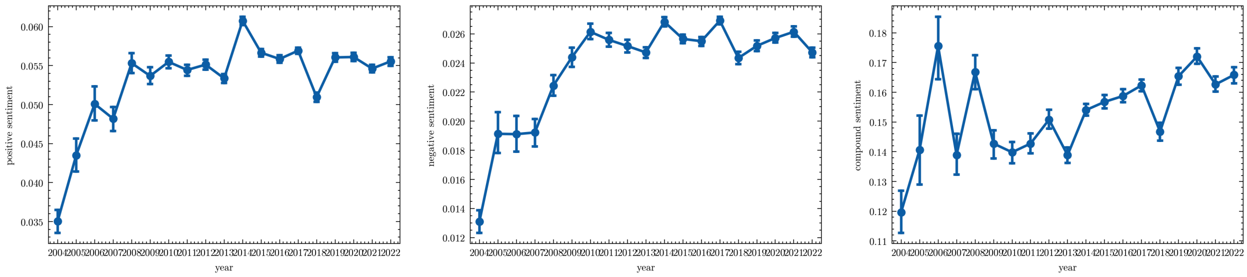

Figure 7: Sentiments of Key Phrases Co-used with Lasso Regression Concept

Our sentiment analysis reveals different patterns for the term “lasso” and “ricci flow”. For “ricci flow”, both positive and negative sentiments increase overtime, which indicates there are still debates about the concept. This indicates that research topics related to “ricci flow” is still growing and not stabilized yet. In contrast, for “lasso”, both positive and negative sentiments decrease overtime, which indicates consensus is reach regarding what “lasso” can do. This finding matches the finding for conceptual association where the association for “lasso” stabilized between 2017 and 2022.

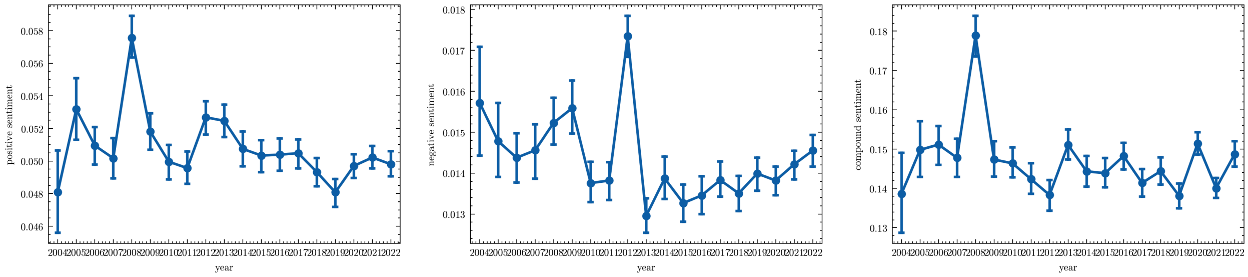

Figure 8: Sentiments of Key Phrases Co-used with Ricci Flow Concept

5 Conclusion and Future Works

Our preliminary study focuses on a couple of case studies to demonstrate our theory about concepts' life trajectories. We demonstrate that the popular approach of using citation patterns to study scientific progress is insufficient and misses the transition from explicit to implicit knowledge. This is where NLP can come in and help. Our case studies show that by introducing NLP methods, we can trace concepts' life trajectories that are missed by citation patterns.

Further works demand scaling up our analysis by considering an expanded set of concepts. To start, we plan to utilize the glossaries at the end of textbooks, which represent "mature" concepts that have become accepted knowledge. For example, for statistical methodologies, we can use the glossary of The Elements of Statistical Learning Hastie et al. (2009)), which is a classic statistics textbook. We will also explore other available NLP methods (e.g., the method proposed by Cheng et al. (2023)) that allow us to extract meaningful concepts from the texts of journal articles.

By extending the frontiers of how we trace knowledge evolution, our work not only challenges conventional metrics but also opens new pathways for understanding the dynamics of scientific progress.

References

Monali Bordoloi and Saroj Kumar Biswas. 2023. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, pages 1 – 56.

Mengjie Cheng, Daniel Scott Smith, Xiang Ren, Hancheng Cao, Sanne Smith, and Daniel A McFarland. 2023. How new ideas diffuse in science. *American sociological review*, 88(3):522–561.

Clayton M. Christensen. 2015. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press. Google-Books-ID: lURBCgAAQBAJ.

Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118. Publisher: Proceedings of the National Academy of Sciences.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *International Joint Conference on Artificial Intelligence*.

Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Richard S Hamilton. 1982. Three-manifolds with positive ricci curvature. *Journal of Differential geometry*, 17(2):255–306.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Rebecca M. Henderson and Kim B. Clark. 1990. Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms. *Administrative Science Quarterly*, 35(1):9–30. Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University].

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing, 2nd Edition*, 2nd edition edition. Prentice Hall, Upper Saddle River, N.J.

T. S. Kuhn. 1962. *The structure of scientific revolutions*. The structure of scientific revolutions. Chicago, University of Chicago Press.

Mayank Kulkarni, Debanjan Mahata, Ravneet Arora, and Rajarshi Bhowmik. 2021. Learning rich representation of keyphrases from text. *arXiv preprint arXiv:2112.08547*.

Imre Lakatos. 1978. *The Methodology of Scientific Research Programmes*. Cambridge University Press, New York.

Bruno Latour. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press. Google-Books-ID: sC4bk4DZXTQC.

James G. March, Martin Schulz, and Xueguang Zhou. 2000. *The Dynamics of Rules: Change in Written Organizational Codes*. Stanford University Press. Google-Books-ID: JT4LyDfDDw8C.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *International Conference on Knowledge Capture*.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *International conference on Asian digital libraries*, pages 317–326. Springer.

Michael Park, Erin Leahey, and Russell J. Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144. Number: 7942 Publisher: Nature Publishing Group.

Tarek Saier, Johan Krause, and Michael Färber. 2023. unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 66–70, Los Alamitos, CA, USA. IEEE Computer Society.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Uri Shwed and Peter S. Bearman. 2010. The Temporal Structure of Scientific Consensus Formation. *American Sociological Review*, 75(6):817–840. Publisher: SAGE Publications Inc.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2).

Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3):387–420. Publisher: SAGE Publications Ltd.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Peter D. Turney. 2002. Learning to extract keyphrases from text. *ArXiv*, cs.LG/0212013.