Investigating patterns of change, stability, and interaction among scientific disciplines using embeddings

Barbara McGillivray[1,*], **Gard B. Jenset**[2], **Khalid Salama**[3], **and Donna Schut**[4]

[1]King's College London, United Kindgom
[2]Springer Nature, United Kindgom
[3]University of Kent, United Kindgom
[4]Google, United Kindgom
[*]barbara.mcgillivray@kcl.ac.uk

ABSTRACT

Multi-disciplinary and inter-disciplinary collaboration can be an appropriate response to tackling the increasingly complex problems faced by today's society. Scientific disciplines are not rigidly defined entities and their profiles change over time. No previous study has investigated *multiple disciplinarity* (i.e. the complex interaction between disciplines, whether of a multidisciplinary or an interdisciplinary nature) at scale with quantitative methods, and the change in profile of disciplines over time. This article explores a dataset of over 21 million articles published in 8,400 academic journals between 1990 and 2019 and proposes a new scalable data-driven approach to multiple disciplinarity. This approach can be used to study the relationship between disciplines over time. By creating vector representations (embeddings) of disciplines and measuring the geometric closeness between the embeddings, the analysis shows that the similarity between disciplines has increased over time, but overall the size of their neighbourhood (the number of neighbouring disciplines) has decreased, pointing to disciplines being more similar to each other over time, while at the same time displaying increased specialisation. We interpret this as a pattern of global convergence combined with local specialisation. Our approach is also able to track the development of disciplines' profiles over time, detecting those that changed the most in the time period considered, and to treat disciplines as compositional units, where relationships can be expressed as analogy equations of the form $Discipline_1 + Discipline_2 \approx Discipline_3$.
These findings can help researchers, academic institutions and organizations to better understand and react to the dynamics of scientific research, and can support the education sector in designing curricula or in the recruitment of academics and researchers.

1 Introduction

How can we prevent and treat a pandemic? What are the best ways to create sustainable energy? Can we build intelligent robots? Many of the open societal questions that science aims to address span across more than one scientific discipline[1], a term we will use broadly to refer to scientific fields or subjects. The scientometric research literature has reported evidence that research is in fact becoming more collaborative[2], and multi-disciplinary collaboration can be seen as an appropriate response to tackling increasingly complex problems[1]. To take just two examples, nano-technology is an interdisciplinary endeavour with contributions from chemistry, physics, and engineering[3], and computational linguistics is similarly interdisciplinary, aiming to develop computer science methods to answer linguistics questions[4].

Being human constructs, disciplines are shaped by, and in turn help to shape, human behaviour. The pull exerted by specialised disciplines and sub-disciplines through self-identification and legitimisation of funding is counter-balanced by the need for collaboration to solve complex problems, as well as the structural invitation from scientific publishers via the launch of large inter- and multidisciplinary journals such as *Scientific Reports* and *PLOS ONE*. A better understanding of the relationship between disciplines helps us contextualise the landscape in which researchers, funders, and research organisations operate. Understanding this relationship can also inform policy makers and other stakeholders about the state of research, both at the level of the researchers' community and in terms of its potential effects on individual researchers.

Yet scientific disciplines are not rigid, well-defined entities. Instead, they are fluid and context-dependent. Scientific disciplines are also multiscale phenomena, where the whole (the discipline) is made up of repeated contributions (e.g. academic works) and interactions (e.g. between researchers). As such, disciplines are to some extent phenomena that emerge as conventions from accumulated individual behaviours. This makes scientific disciplines hard to describe in a manner where

consistency in terminology and description is maintained between the whole (i.e. disciplines) and the individual components, such as research works. An example of this description problem is the question of inter- and multi-disciplinarity, which we can broadly think of as the different practices of collaboration and interaction between disciplines.

We propose a solution to the problem of describing scientific disciplines by treating them as empirical, behavioural phenomena emerging from individual research articles (or other scholarly contributions), which themselves are multi-faceted. Our solution draws on recent research in computational linguistics and natural language processing (NLP), where solutions have been found to similar challenges for describing language phenomena. We present a flexible, descriptive computational framework, and show how it can be used to describe individual disciplines over time, as well as comparing them to each other.

The different terms used when referring to cooperation between scientific disciplines reflect the degree of their integration[5]. According to the definitions in[6], "[m]ultidisciplinarity draws on knowledge from different disciplines but stays within their boundaries. Interdisciplinarity analyzes, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole." In this research we use the term *multiple disciplinarity*[6] to investigate the complex interaction between disciplines, whether of a multidisciplinary or an interdisciplinary nature. Our research questions concern three main areas: how can we generate accurate computational representations of scientific disciplines, how can these representations be used to measure the change in the similarity between disciplines over time, and how do individual disciplines' profiles have changed over time?

1.1 Previous work

In this section we review previous work on the three main areas of focus of this article: representations of scientific disciplines, similarity between disciplines, and profiles of individual disciplines.

There are currently several, parallel approaches to defining what constitutes a scientific discipline. These approaches vary according to two main characteristics. Firstly, the basis on which disciplines are classified varies from co-authorship to co-citation patterns, and lexical or textual similarities[7]. Secondly, different ontologies or classification systems are used for labelling the disciplines. A common classification inventory is the Web of Science/Journal Citation Report system, although this system does not cover all areas of research in equal detail[8].

An interesting recent strand of research, in many ways similar to ours, is[9], which focuses on "the Science of Science", i.e. the fields of Scientometrics and Bibliometrics.[9] combine textual features in the form of text embeddings, combined with graph embeddings that represent the structural and social context that the research is situated within.

However, what is missing is a methodology that describes research fields in a natural way and that scales across all fields of subjects, while compensating for the inevitable shortcomings of individual classification systems. A system that comes close to achieving this is perhaps Microsoft Academic Graph[10]. Since 2021 Microsoft Academic Graph has been decommissioned and will be succeeded by a community driven initiative[11]. While Microsoft Academic Graph covers an impressive number of disciplines in a discipline-concept hierarchy, the system is still constrained by a set of categorical labels that by necessity are more restrictive than scientific subjects themselves. In particular, this constraint (which is shared by all classification systems) makes it difficult to compare disciplines at scale in terms of their similarity and interaction.

There is currently no consensus on how multidisciplinarity and interdisciplinarity are best measured and, despite advances, the problem of measuring interdisciplinarity remains unresolved[3]. One approach[12] uses the list of references in an article combined with Web of Science Subject Categories to calculate integration scores – based on the diversity of a paper's cited references – and diffusion scores – based on the diversity of the set of papers that refer to a given body of work. Another approach[13] uses network analysis to study co-authorship networks and calculate multidisciplinarity via patterns of collaboration between authors.

A significant portion of scientometric research has measured and analysed the similarity between disciplines by using NLP methods to mine different corpora of scientific literature without relying on existing scientific classification schemes. Similarity measures are usually based on co-occurrences of keywords in articles or bibliographic indicators such as citations, although embeddings have also been used.

Embeddings have been developed in NLP to represent words via low-dimensional numerical vectors. Word embeddings are learned from text data and have been found to capture important distributional properties of words, particularly syntactic and semantic characteristics, as words that are used in similar ways, like *car* and *automobile*, tend to have similar representations. Recently, vector space models built from embeddings trained on words[14,15], documents[16] and networks[17] have been employed to study the relationships between scientific articles. One study[18] has focussed on physics papers published in arXiv and train Word2Vec embeddings on the article titles to group them into different clusters corresponding to different scientific subfields. The similarity between research topics and domains has been measured to study how their relation changes over time (from 1986 to 2009)[19], by representing each author profile as a bag-of-topics within the research space, and then training embedding models for the topics and calculate their similarity using standard distance metrics. Another proposed approach[20] is the P2V (paper2vector) algorithm for learning embeddings for academic papers and use it for classifying articles and calculating the similarity between them. Similar research has been carried out within specific domains[21] in a more limited scope.

Previous research on modelling disciplinary interaction has typically taken one of two routes: an ontology-based top-down approach, or a bottom-up approach modelling the individual components such as articles or journals[22]. Qualitative studies of journals or individual sub-fields yield rich descriptions but do not scale well[5]. Another approach uses network graph modelling to combine ontology information and data-driven approaches for the specific area of bio-nanoscience[22]. However, network models can also be challenging to scale, despite advances[23].

Our analysis of the change in profile of disciplines over time takes inspiration from the current literature on word meaning change over time, which is currently an active research area in NLP[24,25]. This approach allows for a natural integration of ontology-based information (in our case Fields of Research or FoR codes) with observed data patterns, in a highly scalable architecture. The scalability of the approach rests on considering articles in terms of co-occurrences of FoR codes. FoR codes represent a limited vocabulary, while the technique we use has been shown to work with the much larger vocabulary (by several orders of magnitude) found in natural languages[26]. Treating the FoR codes as a set of co-occurring vocabulary items allows for a simple representation which is easy to scale and re-apply to other ontology frameworks; this representation could, if necessary, be further scaled by taking large samples, similarly to corpora used for NLP. The majority of previous studies in this area of NLP research start by dividing a corpus of texts to be analysed into separate portions (subcorpora), one for each time interval. Individual representations for words are then built in each of these subcorpora via word embeddings. In order to measure the meaning change of a word, its vector is tracked across the semantic spaces and cosine similarity is typically used to measure the similarity between the vectors for that word in the different time periods. A time series of the cosine similarity between the vector for a word in one time interval and the vector for the same word in the following interval is then created. If this time series shows one or more "dips", i.e. years in which the cosine similarity drops under a certain threshold, this is an indication that the word may have changed its meaning during the time in question. Some approaches, such as[27], employ changepoint detection techniques to identify statistically significantly points of change in the time series. For example, we can identify the change in meaning of the word *tweet* from the original meaning of 'sound made by a small bird' to the more recent additional meaning of 'message posted on Twitter'.[28] use embeddings trained on the Royal Society Corpus spanning the years from 1665 to 1869, to detect evidence for specialisation in scientific language, with a focus on language change. With this partial exception, as far as we know, embedding methods have not been applied outside the area of word meaning change detection. Our study is the first one to apply these methods to track the change in profile of disciplines over time and inform our knowledge of how scientific disciplines have evolved in the past decades.

1.2 Conceptual framework

In this research we propose a scalable and empirical approach to multiple disciplinarity and show how it can be used to trace the relation between scientific disciplines over time.

It has been shown that no ontology is complete and equally suitable for all disciplines[8]. Moreover, for journal-level data that have been classified into subjects, it is possible to use measures like the Jaccard similarity coefficient to calculate similarity between disciplines. However, this assumes that we can have a one-to-one mapping between journals and disciplines. In our dataset articles published in a given journal cannot be all labelled as belonging to a single discipline. This is for various reasons. First, we have multidisciplinary journals for which we cannot find a single discipline that is suitable for all their articles. Second, articles are tagged with more than one subject label, and so the mapping is many-to-many rather than one-to-one. Third, disciplines are organised in an ontology with various levels in the Dimensions database. Even in the case of discipline-specific journals, articles may belong to different sub-disciplines of that discipline, and our analysis aims at surfacing such lower-level disciplines. We circumvent this challenge by treating ontology categories as a distributional phenomenon. Our conceptual framework is inspired by distributional semantics, a branch of linguistics where the meaning of words is interpreted based on data-driven distributions of words in context[29].We represent the disciplines as *embedding* vectors in a geometrical space, and use various measures on this embedding space to measure their similarity. Together, our findings describe different aspects of multiple disciplinarity and therefore point to its multi-dimensional and multi-faceted nature. This approach allows us to treat multiple disciplinarity at the system level, as gradient rather than categorical. This allows us to measure the degree to which disciplines are similar and how this changes over time.

As other so-called black-box models, word embeddings are not directly interpretable by humans and different configurations (including the choice embedding algorithm, initializations, parameters such as context window, number of iterations and minimum frequency) can lead to different representations from the same dataset. In spite of these limitations, embeddings have been shown to consistently and robustly capture semantic properties of words, as measured in a series of intrinsic evaluation tests based on word analogy and relatedness (see[30] for an overview). They have also been successfully evaluated in extrinsic contexts and employed in a range of NLP tasks, including syntactic parsing[31], sentiment analysis[32], and named-entity recognition[33]. In particular, their successful use in the task of identifying the change in meaning of words (lexical semantic change) is a strong argument for their employment in the context of this study, where the change in profile of disciplines is investigated. Finally, one could imagine using the field of research codes directly, e.g. via one-hot encoding. However, due to the distributional

properties of these codes, this would lead to a highly sparse matrix, with the concomitant drawbacks for computational and analytical purposes, which is exactly the problem that word embeddings such as Word2vec were designed to overcome in the first place.

On the other hand, contextualised embeddings such as Bidirectional Encoder Representations from Transformers (BERT) embeddings[34] have been recently proposed in various computational semantics studies. However, these embeddings take word order into account, as they create representations for each word occurrence in context (token). Moreover, these algorithms learn sub-word representations, which helps the treatment of out-of-vocabulary words. These two factors are not applicable in our case. There is no inherent order in the FoR codes and they are single units that cannot be decomposed in smaller lexical units. Finally, we follow the state of the art in research on semantic change detection[35], which shows that for this task token embeddings (i.e. Transformer-based embeddings) perform worse than type embeddings like Word2vec. For all these reasons, we have chosen type embeddings to represent disciplines.

We use data on over 21 million articles published between 1990 and 2019 from the Dimensions database by Digital Science[36] (see section 3.1 for details), which we analyse in 3-year intervals. In this dataset, each article is provided with up to four FoR codes from the Australian and New Zealand Standard Research Classification (ANZSRC) system, which is used in Dimensions[36]. This system consists of three hierarchical levels of categorization of research, with 154 research codes organised into 22 research divisions, which themselves are subsumed under eight research clusters. The system is described in more depth in Supplementary Table 1, Supplementary Table 2, and Supplementary Figure 1.We refer to these codes as *disciplines*. In the rest of this paper we refer to the categorization by area, cluster and division. The most granular level of analysis is the discipline level (FoR codes), based on which we created the embeddings. We collect statistics about the co-occurrence of each discipline with every other discipline. For example, if an article is tagged with both Cognitive Sciences and Linguistics, we consider the two disciplines to co-occur. Section 3.1 gives more details about the characteristics of our dataset, and the analysis we performed. Based on these distributional data, we create vector representations (embeddings) for the disciplines. If two disciplines tend to co-occur in the same articles, this is an indication that they share distributional properties and are therefore *distributionally similar*. A discipline's neighbourhood (i.e. the set of its neighbours) can change over time, depending on the patterns of co-occurrence of FoR codes in our dataset, and can offer us insights into general trends over time.

Our approach has the advantage of being data-driven and adaptable to the data at hand. Unlike an approach that purely relies on a subject ontology, according to which Inorganic Chemistry and Organic Chemistry, for instance, are closely related because they are both part of chemical sciences, using the distributional approach we can discover unexpected similarities between disciplines and, crucially, trace how these change over time. Our approach can be applied to other sets of published articles with the same format, and is therefore able to detect patterns in different datasets.

We answer three main research questions:

1. **Representations of scientific disciplines** How can we generate accurate representations for scientific disciplines to capture their similarity?

2. **Similarity between disciplines** How does the similarity between disciplines change over time, overall and within different categories of disciplines?

3. **Profile change of disciplines**: How do the profiles of disciplines change over time?

2 Results

We propose a new data-driven framework for representing and analysing the profile of disciplines over time, based on co-occurrence statistics of FoR codes transformed via a trained embedding model. The embedding model allows us to treat every discipline compositionally, and to tease apart relationships, including partial similarities, via standard vector operations. Using this framework, our results show the multi-faceted nature of multiple disciplinarity, while simultaneously demonstrating that non-multiple disciplinary, or "classical"[37], disciplines are also well-represented in this manner.

We present our results in three groups, corresponding to the three research questions introduced in section 1.2:

1. **Representations of disciplines** We propose a new way to define disciplines' profiles based on embedding techniques and show that these representations capture known relationships between disciplines, including analogy type relations.

2. **Similarity between disciplines** Using our definition of discipline embeddings, we apply two measures: similarity and neighbourhood size. We find that the average similarity between disciplines, measured as cosine similarity between embeddings, consistently increased over time. We interpret this result as showing that disciplines, overall, have "converged" and become more similar over time. In particular, scientific (STEM) disciplines tend to converge more over time than humanities and social sciences (HSS) disciplines. However, the disciplines did not converge into one

single group, but maintained some internal differences. When we look into more granular classifications of disciplines into further subgroups, we see that the cosine similarity in many groups tends to decrease or not change. In order for this to be compatible with a general upward trend in the disciplines' similarity, we need to assume that the similarity between disciplines *across* groups increased over time. To complement this analysis, we measure each discipline's neighbourhood size as the number of its neighbours, i.e. those other disciplines that are sufficiently close to it. We find that neighbourhood size decreases over time. The overall downward trend in neighbourhood size is confirmed in both the HSS and the STEM group and in each cluster. However, if we look at more granular classifications, we find that *Technology*, *Physical Sciences*, and *Chemical Sciences* display an upward trend in their average neighbourhood sizes, while the other disciplines confirm the general downward trend. We interpret this result as pointing to disciplines becoming more specialised.

3. **Profile change of disciplines** We propose a method inspired by NLP research on word meaning change to detect the disciplines whose profile changed in the period under consideration. Our data show that *Communication and Media Studies*, *Computer Hardware*, *Building*, and *Food sciences* have changed most during the period under consideration, and offer an interpretation of some of these trends.

In the next section we go into further details about each group of findings.

2.1 Representations of scientific disciplines

How can we generate accurate representations for scientific disciplines to capture their similarity? We create embeddings for disciplines based on their co-occurrence in scientific articles. These embeddings act as abstract representations of the disciplines, capturing aspects of their profiles.

Embeddings[14] are data-driven low-dimensional numerical vectors representing discrete items, such as words, movies, or subject fields. Two items are considered similar if they share the same context, that is, if they occur with similar items. In the case of language, word embeddings capture important aspects of the semantics (or meaning) of words, as words that occur in the similar textual contexts tend to have similar embeddings and be semantically related.

We use co-occurrence of the FoR codes in scientific articles to define *discipline embeddings*, following the intuition that, if two disciplines tend to occur in the same articles' classification, they are likely to have similar profiles. Section 3.2 describes the methodology we devised to define the embeddings. We evaluate the embeddings quantitatively and qualitatively. Firstly, when grouping embedding vectors by similarity, we find a statistically significant difference between the most similar items compared to randomly paired items. Secondly, we find that disciplines represented in our embeddings display a gradient that corresponds well to the STEM/HSS distinction, as illustrated in Figure 6. Thirdly, we find that our embeddings are capable of representing analogy relationships between disciplines, similarly to word embeddings[14]. In short, our embeddings exhibit similar characteristics to traditional word embeddings. See section 3.3.4 for more details on the evaluation.

The analogy relationships between disciplines, of the type $king_v - man_v + woman_v \approx queen_v$ (where X_v is a vector representing X), let us tease apart disciplines by treating them as compositional entities based on co-occurrence characteristics, rather than necessary and sufficient criteria for category membership. This means that both fairly homogeneous and highly multi-disciplinary disciplines can be represented, and compared, in a like-for-like manner. Based on our results, we find that when we remove the psychology-component of a multiple disciplinary field like *Cognitive Sciences*, we are left with something that is very close to language studies. Equally, we find that two closely related fields, *Linguistics* and *Language Studies*, differ primarily by the latter's more literary side, as shown by the analogy equation *Language Studies* − *Performing Arts and Creative Writing* ≈ *Linguistics*. We take these results as interesting contributions in their own right, while also providing a stepping stone for further experiments.

Having established that discipline embeddings capture the similarity between disciplines, we use them to analyse large-scale trends, as shown in sections 2.2 and 2.3.

2.2 Similarity between disciplines

How does the similarity between disciplines change over time, overall and within different categories of disciplines? In order to measure whether the overall similarity between disciplines has increased or decreased over time we calculated the average cosine similarity between all pairs of disciplines for every time interval. Following a standard practice in embedding analysis, we calculated the cosine similarity between every pair of FoR embeddings for every three-year interval, and averaged over all pairs. Section 3.4 provides the details of our method. Figure 1 shows the cosine similarity by cluster. The time series of the average cosine similarity between the embeddings can be seen in the Supplementary Figure 4. Overall, *the average similarity between disciplines consistently increases over time*, as measured by Kendall's rank correlation with the series of time intervals ($\tau = 0.56$, $p - value = 0.044$). We interpret this result as an indication that the disciplines, overall, have "converged" and become more similar over time.

We analyzed the cosine similarity between disciplines within subgroups as different levels of the ontology. If we divide the disciplines into HSS and STEM, we see that the overall upward similarity trend is mainly due to the STEM disciplines, while the trend of the HSS disciplines fluctuates more. This means that *STEM disciplines tend to converge more over time than HSS disciplines.* A visualization by areas and by research division, respectively, can be found in Supplementary Figures 5 and 6.Figure 1 shows the cosine similarity trends between disciplines in different clusters. *The similarity trend for the clusters is variable, with no clear upward or downward trend.* We notice that the cosine similarity in many groups tends to decrease or not change. In order for this to be compatible with the overall and area-level similarity increasing trend, we infer that the similarity between disciplines *across* groups increases over time.

The average size of the neighbourhood of the embeddings (defined as the number of embeddings whose the cosine similarity to a target embedding is higher than or equal to 0.8) also goes down over time (see section 3.4 for details), as measured via Kendall rank correlation with the distribution of year intervals: $\tau = -0.94$ ($p - value \ll 0.05$ with a significance threshold of $\alpha = 0.05$). The time series of the average neighbourhood size of the embeddings over time can be seen in Supplementary Figure 7.

If we divide the disciplines into HSS and STEM (Figure 2) or clusters (Figure 3), and compute the average neighbourhood size per each group, we see that *the overall downward trend is confirmed in both the HSS and the STEM group and in each cluster*. However, when we group the disciplines by their research division (Supplementary Figure 8), we find that *Technology, Physical Sciences, and Chemical Sciences display an upward trend in their average neighbourhood sizes, while the other disciplines confirm the general downward trend*. This means that such disciplines have an increasing number of neighbours in their embedding space.

2.3 How do the representations of disciplines change over time?

While the average neighbourhood size offers ways to look at global trends in the similarity between disciplines over time, it does not show how the *neighbourhood* of each discipline changes over time. To answer this question, we adopted a combination of three approaches, described in section 3.5. The results of this analysis show that the following disciplines have changed their profile in the period under consideration:

- *Communication and Media Studies*
- *Computer Hardware*
- *Building*
- *Food sciences*

Figure 4 shows the time series of two of the disciplines whose profile has changed: *Communication and Media Studies* and *Computer Hardware*. A more in-depth inspection of the neighbourhood of these disciplines reveals interesting patterns. We exemplify this by a closer look at two specific cases.

We find no overlap between the neighbours of *Computer Hardware* in 1990-1992 and its neighbours in 2017-2019. The former set contains disciplines related to biological sciences (*Plant Biology*, *Physiology*, *Zoology*, *Genetics*, and *Horticultural Production*), pointing to a focus on applications in the biological domain in the early 1990s, while the latter contains more computational disciplines (*Distributed Computing*, *Computer Software*, *Data Format*, *Information Systems*, *Computation Theory and Mathematics*). A similar co-occurrence between foundational Computer Science fields and Biology fields has been noted in previous research using different methods[38]. Our results suggest that *Computer Hardware*, a foundational part of the broader Computer Science field[38], has become less outward-looking over time, drifting closer to other computational disciplines and away from specific application areas.

Next we consider the profile of *Communication and Media Studies*, which changed significantly in the first part of the time span under consideration. Previous research has noted that communication studies has grown rapidly in recent decades, while simultaneously going from a field that heavily leans on other fields for theories, to developing a coherent identity of its own[37]. Our results support this interpretation. If we compare its neighbours in the reference time interval (2017-2019) with its neighbours in the intervals 1990-1992, 1993-1995, 1996-1998 and 1999-2001, we find no overlap. For example, in 1990-1992, its neighbours were *Medical Biotechnology*, *Physiology*, *Building*, *Veterinary Sciences*, and *Transportation and Freight Services*. These neighbours, which might seem surprising, fit well with the early focus in communication studies on health and risk management communication[37]. From around 2002-2004, *Communication and Media Studies* begins to surround itself with the same neighbours as in 2017-2019, viz. *Journalism and Professional Writing*, *Religion and Religious Studies*, and *Anthropology*, with *Film, Television and Digital Media*, *Journalism and Professional Writing*, *Visual Arts and Crafts* joining the list of neighbours at a later stage. We see this development as broadly compatible with the history of communication studies provided elsewhere, from an applied focus on communication, towards a more theoretically developed field that casts a

critical eye of mass communication and web-based communication[37], while retaining the links to its roots (as a discipline) in classical rhetoric via a neighbour like *Performing Arts and Creative Writing*. When we zoom in on the 2017-2019 data, we find that the different facets of *Communication and Media Studies* can be explored via embedding vector analogy equations, in a similar manner to NLP, where $king_v - man_v + woman_v \approx queen_v$ is probably the best-known example[14]. With *Communication and Media Studies*, we found that subtracting the vector of *Performing Arts and Creative Writing* and adding the vector for *Visual Arts and Crafts* resulted in a new vector, or vector offset, that is closest to the vectors for *Sociology* and *Journalism and Professional Writing*. Moreover, taking *Communication and Media Studies* as a starting point and subtracting both vectors for *Performing Arts and Creative Writing* and *Visual Arts and Crafts* but adding the vector for *Econometrics* (the statistical analysis of economic data) yielded a result close to *Law* and *Marketing*. We interpret these results as being broadly compatible with the multiple-disciplinary nature of communication studies[37]. In the first analogy example, stripping away the vector linked with creative writing brings us closer to the discipline's use of theory from sociology and media studies[37]. In the second example, where we subtract vectors for both visual arts and creative writing while adding the vector representing *Econometrics*, we get something that we hypothesize reflects the discipline's marketing-related practice focus[37]. See Table 3 for further examples of analogy equations.

3 Methods

3.1 Data

Our data extraction and processing pipeline is illustrated in Figure 5. The code is available at https://github.com/BarbaraMcG/discipline-embeddings/. The input dataset for our study consists of over 21 million scientific articles published between 1990 and 2019. The data were obtained from the Dimensions database[36], stored in a Google BigQuery cloud environment[39].

We divided the articles into ten groups, corresponding to ten three-year time intervals based on the articles' publication year, from 1990-1992 to 2017-2019. The total number of distinct articles is 21,201,258, and each group has between 864,048 and 3,779,330 articles. In addition to their publication date, the articles were selected according to the following criteria, based on Dimensions metadata: document types other than "article" were excluded, and only English language documents were included.

The basis for the document FoR classification in Dimensions is textual, and uses machine learning to assign up to four different FoR codes to a document. The subject ontology we use consists of two levels: the research division level contains 22 top-level subject labels, such as "Biological Sciences", "Chemical Sciences" and "Information and Computing Sciences'; level 2 contains a finer-grained categorization of disciplines into 154 labels such as "Agricultural Biotechnology", "Analytical Chemistry", and "Condensed Matter Physics".

The co-occurrence of more than one FoR code per article opens the possibility for applying NLP techniques, since FoR codes can be treated as co-occurring "words" within the unit of a research article. Just like some words are more likely to co-occur than others, we assume that some FoR codes are also more likely to co-occur than others, and that this tendency is meaningful. This assumption, however, rests on the accuracy on the assignment of the FoR codes in the first place.

3.2 Embedding model, parameters, and data

The discipline embeddings were created with the *Word2Vec* implementation in the Python 3 package *Gensim*[40]. Word2Vec was chosen since it is a well-known embedding approach that has seen multiple adaptions beyond words, e.g. for paragraphs[16], graphs[17], and academic paper citations[20].

The embedding training used the Skip Gram model, since that often yields better results than continuous bag of words[14]. We used a context window of 2 and a minimum frequency count of 1. The context window specifies the size of the context to be taken into account as the training algorithm iterates over the training data instances. The minimum frequency allows a threshold to exclude low-frequency data, but we chose a minimum count of 1, meaning all data is kept. These values were chosen based on the input data, which consist of up to four categories (out of a closed set of 154) per article. The discipline embedding dimension was set to 12, which is the midpoint between two common heuristics [41, 48], viz. the fourth root of the number of unique categories and the square root of the number of unique categories times 1.6.

The input data consisted of level 2 Dimensions FoR code co-occurrence data within documents, so that the input to the Word2Vec algorithm consisted of a vector of the level 2 FoR codes per document, up to a maximum of four. Although the FoR codes are not equally frequent, the imbalances are much less pronounced than in natural language. Highly variable word frequencies can be a source of instability when training Word2Vec models on natural language[42]. However, it is important to keep in mind that the imbalances typically seen in natural language are far larger than in our data. In our 2017-2019 data, *Clinical Sciences* (the most frequent code) appears about 24,000 times more often than the least frequent code (*Other Law and Legal Studies*). For illustration, in the 4 billion word Corpus of Contemporary American[43], the word *of* appears over 23

Journal	Scope	% shared FoR
Language Cognition and Neuroscience	sub-field	70
Diachronica	sub-field	63
Cognitive Linguistics	sub-field	62
Pragmatics	sub-field	60
Journal of Phonetics	sub-field	60
International Journal of Corpus Linguistics	sub-field	56
Journal of Semantics	sub-field	54
Linguistic Inquiry	general	48
Linguistics	general	48
Journal of Linguistics	general	47
Digital Scholarship in the Humanities	multiple disc.	40
Palgrave Communications	multiple disc.	26

Table 1. Linguistics journals by scope and percentage of shared FoR codes between 2019 articles and their references, based on Dimensions data.

million times[44]. In other words, the imbalance between *of* and any word with a minimum occurrence of one is about 23 million, meaning that the imbalance we can find in natural language can be about 1,000 times greater than in our data.

3.3 Evaluation

3.3.1 How accurate are the FoR codes?

A key step was to establish whether the co-occurrence of more than one label in an article represents a meaningful indicator of the presence (and contribution) of multiple fields, or if it merely represents the error margin of the machine learning algorithm. We devised a simple test of the Dimensions FoR codes, based on the intuition that the more specific an article is to a discipline, the more likely it is to cite other articles in the same discipline. Conversely, an article that skews towards multiple disciplinarity would be more likely to cite a diversity of disciplines. To test this intuition, we looked at 12 linguistics journals, and we classified them by scope either as journals pertaining to a sub-field of linguistics (e.g. historical linguistics), general linguistics journals, or multi-disciplinary journals that also accept linguistics articles. For each journal, we calculated the average overlap in FoR codes between each 2019 article and the articles cited by it. If the intuition above is correct, we should see higher percentages of shared FoR codes between articles and their references among the more specialised journals, and conversely lower average percentages for the more general journals. In the more specialised journals, articles share a high degree of overlap in FoR codes with the articles on their list of references. On the other hand, in a multi-disciplinary journal such as *Palgrave Communications* (renamed *Humanities and Social Sciences Communications* in 2020) there is a low overlap in FoR codes between articles and their references. The mean overlap per journal scope is 33% for the multi-disciplinary journals, 47.7% for the general linguistics journals, and 60.7% for the journals focused on a linguistics sub-field.

The variation in overlapping FoR codes, which correlates almost perfectly with the breadth of the journal's scope, supports the intuition that the FoR codes meaningfully capture breadth of scholarship. This fact, in turn, lends credibility to our approach using the co-occurrence of FoR codes to measure disciplinary interaction, via FoR code co-occurrence at the article level.

3.3.2 Evaluating the embeddings

Word embeddings are typically evaluated on a number of tasks, including word similarity, analogy, sentiment classification, and named entity recognition[45,46]. Not all NLP evaluation tasks are relevant for our embeddings, but we identified some tasks that give an indication of the intrinsic embedding quality and that are relevant to our data, namely similarity, analogy, and ability to propagate meaning to higher units.

To evaluate discipline similarity, we took the reasonable assumption that close neighbours, measured by vector similarity, should be much closer to each other compared to randomly sampled pairs of vectors. We used Annoy, an open source Python library released by Spotify (`https://github.com/spotify/annoy`), to search for the nearest neighbours for all the discipline embedding vectors based on the most recent set of embeddings (2017-2019 data). For each discipline embedding, we identified the three most similar disciplines, based on the vector similarity, and calculated their cosine similarity (see Figure 2). The result was a vector of 432 similarity scores. We then randomly sampled the same number of vector pairs and calculated their similarity. The two sets of resulting similarities (top three most similar versus random sampling) showed strikingly different values. The mean similarity for the top three most similar vectors was 0.68 (standard deviation = 0.1). For the random condition we found a mean similarity of 0.07 (standard deviation = 0.31). The difference is highly statistically significant with an independent samples t-test for unequal variances ($t = 39.01$, $p \ll 0.0001$).

Discipline	Neighbours (similarity)
Civil Engineering	Automotive Engineering (0.809), Maritime Engineering (0.694)
Macromolecular and Materials Chemistry	Nanotechnology (0.797), Atomic, . . . and Plasma Physics (0.726)
Geology	Geochemistry (0.609), Environmental Engineering (0.606)
Cognitive Sciences	Language Studies (0.564), Linguistics (0.486)
Artificial Intelligence and Image Processing	Statistics (0.518), Classical Physics (0.475)
Linguistics	Language Studies (0.666), Performing Arts and Creative Writing (0.575)
Communication and Media Studies	Performing Arts and Creative Writing (0.816), Visual Arts and Crafts (0.809)

Table 2. Top two most similar neighbours (by cosine similarity) for a selection of disciplines. Cosine similarity was calculated with Annoy.

Analogy equation	Result ($\approx$)
Cognitive Sciences − Psychology	Language Studies, Data Format
Cognitive Sciences − Linguistics	Computation Theory and Mathematics, Quantum Physics
Computation Theory and Mathematics + Language Studies	Cognitive Sciences, Design Practice and Management
Language Studies − Performing Arts and Creative Writing	Cognitive Sciences, Linguistics
Zoology + Fisheries Sciences − Oceanography	Veterinary Sciences, Animal Production
Communication . . . − Performing Arts . . . + Visual . . .	Sociology, Journalism and Professional Writing
Communication . . . − Performing . . . − Visual . . . + Econometrics	Law, Marketing

Table 3. Top two most similar vectors to the offset created by additions and subtractions. When any of the vectors involved in the left hand side of the equation were in the top two results, they were skipped in favour of the next result. Some very long FoR names have been shortened in the table.

For evaluating analogy, we took an exploratory approach, since no standard analogy tasks exist for data such ours yet. In NLP, linear relations between pairs of word embedding vectors, such as the classic $king_v - man_v + woman_v \approx queen_v$ example, have been shown to reflect linguistic relationships[14], although this assumption does not always hold in practice for any set of words[47]. For the analogy task, our assumption is that if we find such linear analogy relationships at all, then the embedding model has some value. Future research would be needed to explore this in more depth. For our data, we found several interesting cases of linear analogy relationships. Once again we employed Annoy to identify similar embedding vectors for the most recent data. We found that both when both adding and subtracting vectors, the resulting vector offset can be used to retrieve meaningful vector analogies on our data. To identify the analogous vectors, we retrieved the two most similar vectors, ignoring the vectors involved in the starting analogy equation, since operations of the type $Vector_1 + Vector_2$ will sometimes return either $Vector_1$ or $Vector_2$ among the top matches. Whenever this was the case we would move on to the next result. From the example results in table 3, we can draw a number of conclusions. Firstly, it is possible to discover analogy relationships in our embedding data, as is the case with conventional word embeddings. Secondly, as the sets of examples involving Cognitive Sciences and communication studies (respectively) show, it is possible to use these analogy relationships to treat multiple disciplinary fields compositionally. Also, we can observe that both addition and subtraction have the expected results. Thirdly, these analogy operations can be used to tease apart the differences between related disciplines, such as Language Studies and Linguistics. Importantly, these examples show that our results are not limited to fields that are highly multi- or inter-disciplinary. Finally, the example involving Zoology, Fisheries Sciences, and Oceanography shows that complex operations with more than two disciplines produce intuitively interpretable results, in this case something like the land-based equivalent to Fisheries Sciences.

Finally, we explored how the embeddings perform from the perspective of a higher-level of organisation, inspired by (but different from) sentence evaluation tasks for word embeddings[45]. We decided to plot the distribution of disciplines in the embedding space, and evaluating whether they form natural groupings. For this evaluation, we performed Principal Component Analysis (PCA) on the embeddings from the 2017-2019 data. The PCA-transformation was successful and the first two principal components account for 41% and 27% of the total variation, respectively, summing to 68%. Each discipline was then matched with its parent-level category in the ontology. The resulting plot in figure 6 shows a clear picture of a natural continuum over the different disciplines. In figure 6 we see an arch, or horseshoe-like, distribution, which is common when a continuum or gradient is represented in a restricted space [48, 127]. From the top left of the plot, we can see a natural progression from medical and biological sciences, via chemical and physical, earth, environment, and computing, to social sciences and humanities on the right hand.

In summary, we have shown that our embedding model has many of the same features as a conventional word embedding

model: our model is capable of identifying similarity between disciplines, it can be used for analogy tasks in a similar manner to word embeddings, and when we plot our embedding vectors we see that they form a natural continuum that is meaningful and interpretable at the macro-level. For these reasons, we consider our embedding model a useful tool for further studies of the relationship between the disciplines.

3.3.3 A simple diachronic baseline: number of FoR codes per article over time

Having established that the FoR codes in Dimensions are a useful proxy for capturing disciplines and their interaction, the next question is how best to perform an analysis on them. A simple method for analysing such subject classification is to count how many subject labels are associated with the same papers, and how this changes over time.

Using a non-parametric Mann-Kendall trend test applied to the mean number of FoR codes per document over time, we find that there is statistically significant positive trend, with the average number of FoR codes increasing over time ($\tau = 0.64$, $p \ll 0.001$).

This approach, although useful as a baseline, is not optimal because it does not reveal the relationships between the disciplines. For example, it does not tell us if two disciplines tend to occur together and how that has changed over time. Disciplines have complex relationships to each other, which are not necessarily captured in raw counts. To overcome these drawbacks, we instead train discipline embeddings, as explained in the next section.

3.3.4 Evaluation of embedding hyperparameters

We investigated the following hyperparameters for the discipline embeddings:

- window size, values 2 and 4;
- negative sampling size, values 1, 2, 3, and 5.

In order to find the best configuration of hyperparameters, and to assess the quality of the discipline embeddings, we devised the following intrinsic evaluation approach. We made use of the mapping between level-1 and level-2 FoR codes. Our assumption is that, if geometric proximity within the discipline embedding space corresponds to semantic similarity between the FoR codes, level-2 discipline embeddings mapping to the same level-1 FoR would sit in the same region of the space. In other words, our hypothesis is that the average similarity between level-2 discipline embeddings mapping to the same level-1 FoR is higher than the overall similarity between all level-2 FoR codes. For every three-year interval and every combination of hyperparameters, we counted the number of level-1 FoR codes for which the hypothesis was confirmed, and then took the average over all intervals. We found that the best configuration of hyperparameters is window size 2, negative sampling size 1, which gave rise to an average of 19 positive cases out of 20.

3.4 Similarity analysis

We analysed how the similarity between disciplines shifted over time using two measures: average cosine similarity and average neighbourhood size.

The cosine similarity between two vectors (in our case the discipline embeddings) is the cosine of the angle between them. Independently of their length, if the vectors are completely aligned, the angle between them will be 0° and the cosine would be 1; on the other hand, if the vectors are orthogonal, the cosine of their angle is 0, and if they are diametrically opposed, the cosine of their angle is -1. In NLP words can be represented as embeddings based on the the context in which they occur. Therefore the cosine similarity between the embeddings of two words can be interpreted as semantic similarity: if the embeddings of two words (for example *cat* and *dog*) have a cosine similarity close to 1, this is taken to imply that the two words share some aspects of their meaning or semantics.

For each three-year time interval from 1990 to 2019 and each pair of discipline embeddings, we calculated their cosine similarity. We then calculated the average of these cosine similarities. We finally analysed the time series created this way to identify any significant monotonic upward or downward trend by calculating the Kendall rank correlation coefficient[49,50], which resulted in $\tau = 0.56$ ($p - value = 0.044$ with a significance threshold of $\alpha = 0.05$.

Given a distance value d, we defined the neighbourhood of an embedding e as the set of all the embeddings f whose similarity to e is equal to or greater than 1 minus the distance d:

$$neighbourhood(e,d) = \{f \,|\, sim(e,f) \geq 1-d\} \tag{1}$$

The size of this neighbourhood is the number of embeddings contained in it. We chose a distance threshold value of 0.2, motivated by analogy between cosine similarity and correlation coefficients, where a correlation of 0.8 (i.e. a distance of 0.2) would count as a large correlation. Based on this threshold value, we calculated the neighbourhood size for each embeddings, and took the average values. We found a *statistically significant downward trend over time* as measured by the Kendall rank correlation coefficient ($\tau = -0.94$, $p \ll 0.05$) for the average size of the neighbourhood of the embeddings.

FoR code	Cosine	Changepoint
Architecture	0.29	1996
Astronomical and Space Sciences	0.42	2002
Automotive Engineering	0.02	1996
Biomedical Engineering	0.29	1996
Building	0.02	1999
Cardiorespiratory Medicine and Haematology	0.20	1996
Communication and Media Studies	-0.08	1999
Complementary and Alternative Medicine	0.09	1996
Computer Hardware	0.13	1996
Distributed Computing	0.10	2008
Education Systems	0.03	1999
Food Sciences	-0.04	1999
Immunology	0.48	1999
Medical Biochemistry and Metabolomics	0.005	1999
Medical Biotechnology	0.07	1999
Medical Physiology	0.28	1999
Medicinal and Biomolecular Chemistry	0.09	1999
Neurosciences	0.37	1996
Nutrition and Dietetics	0.43	1999
Oncology and Carcinogenesis	0.17	2002
Ophthalmology and Optometry	0.23	1996
Other Medical and Health Sciences	0.26	2002
Pharmacology and Pharmaceutical Sciences	0.04	1999
Transportation and Freight Services	-0.27	1999, 2005
Visual Arts and Crafts	0.02	1999

Table 4. List of disciplines whose profiles have changed between 1990 and 2019. The second column reports the minimum cosine similarity score from the time series. The third column reports the changepoint.

3.5 Profile change analysis

In order to measure how the profiles of scientific disciplines changed over time we adopted a combination of three approaches based on state-of-the-art methods from NLP research in lexical semantic change detection[25].

Self-similarity between the first and the last time interval In the first approach, we generated two separate sets of embeddings. The first set was generated using the articles published between 1990 and 1992 (which we call t_1), the first time interval in our dataset, and the second set was generated using articles published between 2017 and 2019 (t_2), the last time interval in our dataset. We used Orthogonal Procrustes[51] to align the sets of embeddings from the two different time periods and then calculated the cosine similarity between the embedding of each discipline in t_1 and the embedding of the same discipline in t_2 (which we call self-similarity). Finally, we sorted the disciplines by decreasing cosine similarity. As shown in Supplementary Figure 9, *Communication and Media Studies*, *Food Sciences*, *Medical Biochemistry and Metabolomics*, *Building*, and *Transportation and Freight Services* are the disciplines with the highest degree of change in their representations over time, while *Other Studies In Human Society*, *Computer Software*, *Horticultural Production*, and *Accounting, Auditing and Accountability* are the disciplines displaying the least amount of change in their representations over time.

Self-similarity time series analysis In the second approach, we studied the trajectory of the disciplines' self-similarity in each time interval.

First, we built a semantic space for each three-year interval from 1990 to 2019. In order to make the semantic spaces comparable, we used Generalised Procrustes Alignment[52]. After aligning the spaces, for each discipline and for each time period t, we compared the discipline embedding in the semantic space for t with the discipline embedding for a reference time interval t_r by calculating the cosine similarity between the two. t_r was chosen as the last time interval as this has been shown to lead to optimal results[53]. By collecting such cosine similarity values for every time interval, we obtain a time series that can help us trace the profile change of the discipline. Compared to the previous approach, this method allows for a more fine-grained analysis and makes it possible to detect changes in the profiles of disciplines even when the embeddings for the disciplines did not change much from the first time interval to the last time interval.

We performed a changepoint detection analysis to identify the points in time where a significant change in the time series of

cosine similarities occurred. We used the Pelt algorithm[54] to detect the changepoint detection, with a penalty of 0.5 and a jump parameter of 1. Table 4 contains the 25 disciplines that have changed during the time period under consideration according to this analysis, and their changepoints.

Profile via neighbourhood In the third approach, we defined the profile of a discipline as the set of the disciplines within its neighbourhood in the embedding space. We then identified whether the profile of a given discipline has changed between the first and the last time interval over time by tracing the disciplines that moved in and out of its neighbourhood. Supplementary Figure 10 shows two groups of disciplines, those with the largest profile change over time and those with the smallest profile change over time. We found that *Building*, *Food Sciences*, *Communication and Media Studies*, *Architecture*, and *Computer Hardware* have the highest profile change score over time, while *Geochemistry*, *Other Chemical Sciences*, *Ecology*, and *Banking, Finance, and Investments* have the lowest profile change score over time.

4 Discussion and conclusion

We propose a new data-driven framework for representing and analysing the profile of disciplines over time. Our methodological framework can be applied to other areas, for example authors or institutions or other ontologies of disciplines such as the Microsoft Academic Graph[55]. Our analysis on scientific articles published between 1990 and 2019 contributes in three areas: representations of disciplines, similarity between disciplines and change of profile of disciplines. We specifically link these results to the still open question of multiple disciplinarity, which we argue can be approached from a distributional, co-occurrence perspective.

First, we develop *discipline embeddings*, data-driven representations of disciplines based on the co-occurrence statistics of FoR codes from over 21 million published articles between 1990 and 2019. We show that these representations capture known relationships between disciplines as recorded in the FoR taxonomy. We also show that these embeddings can be used to treat disciplines compositionally as entities that arise via interaction, and that treating disciplines as vectors that can be added or subtracted yields results that are both intuitively interpretable and in accordance with results from previous studies.

Second, we measure the trend in the similarity between scientific disciplines using the embeddings via similarity and neighbourhood size, which highlight different aspects of this trend. We find that the average cosine similarity between embeddings consistently increased over time, which may indicate that, overall, disciplines have become more similar over time. This trend is more pronounced for STEM disciplines than for HSS. At higher levels of granularity the picture is more varied, with many groups displaying decreasing or not changing similarity over time. On the other hand, neighbourhood size decreases over time. This is confirmed in both HSS and STEM disciplines, and at the level of individual disciplines, with some exceptions (such as *Technology*, *Physical Sciences*, and *Chemical Sciences*). This may indicate that disciplines have become more specialised, within a general pattern of disciplines growing closer to each other. We interpret this as evidence of a global convergence of research fields in combination with local specialisation.

Finally, we apply methods from recent NLP research on word meaning change modelling to detect the changes in profiles of the disciplines in the period under consideration. We are then able to track the development of the disciplines' representations over time. Our data show that *Communication and Media Studies*, *Computer Hardware*, *Building* and *Food sciences* have changed most during the period under consideration. We investigate the cases of *Computer Hardware* and especially *Communication and Media Studies* in more detail, and show how our findings align with previous studies of the histories of these disciplines, while showing that our approach allows for a detailed exploration of the change process.

First and foremost, we are able to decompose multiple disciplinarity into three dimensions: 1) similarity and analogy relations between disciplines, 2) neighbourhood size, and 3) neighbourhood identity. This allows us to explore the vexing question of whether research is becoming more or less specialised in more precise terms. As friendships among humans can be few in number but more or less deep and prolific, disciplines connect to each other in different ways. Some disciplines have few neighbours, others have an above-average number of close neighbours, and others are close to the average. We can think of similarity and neighbourhood size as structural factors, measuring the degree to which a discipline interacts with other disciplines and with how many.

Neighbourhood identity, on the other hand, captures which disciplines a discipline interacts with. Following how disciplines change one set of neighbours for another set leads us to a close analogy with language change. One way that a change in the semantics of a word can be observed is by examining the identity of the words it co-occurs with. For example, the word "blackberry" went from occurring primarily with other fruits and berries, to occurring with electronic devices[56]. At the same time, frequency changes of (co-) occurrence with other words can also signal more structural changes, such as a change in function or category membership[57,58]. Looking at disciplines over time, we can see several indications of structural change. The disciplines studied here have grown closer to each other, so in some sense they have become more structurally similar.

This research has some limitations. It relies on the FoR codes used in the Dimensions database to represent disciplines. We do not have access to the details of how these codes were automatically assigned and to the accuracy of this assignment over

time, so its quality may lower for certain groups of articles. For example, articles from disciplines that are under-represented in the database might negatively affect our results, although a closer look at the relatively small field of linguistics corresponded well with what we would expect (see section 3.3.1). Moreover, we did not analyse articles published before 1990, which could reveal longer-term trends, as well as potential issues with data quality in very long time frames.

There are several ways in which the research in the present study might be extended in the future. Further work would be needed to investigate the development of a larger number of individual discipline profiles, in comparison with the existing body of discipline-specific research. Furthermore, we can see room for additional evaluation and application tasks of our work. In NLP-focussed research on embeddings, we find a reliance on external evaluation, in the form of benchmarks and applied downstream tasks. We have already shown that our embedding vectors can be used to complete analogy tasks, similar to what is done in NLP research. These tasks would be worth exploring further for our data, to reveal the depth and extent of similarities with word embeddings. Beyond evaluation, an interesting research strand based on our embedding approach would be to study the compositionality of scientific fields, and how this correlates with other approaches to multiple disciplinarity. In terms of applications, the embeddings might be used as input for a prediction task, where an embedding vector might yield better performance than categorical FoR codes. Example tasks could be predicting the overlap of journal scope or co-authorship. Finally, a third set of applications could be about translating between different sets of ontologies used for research classification, in analogy with machine translation for natural language[59], where the classification codes would be the analogy of words to be mapped from one ontology to another, such as between the Dimensions FoR codes and the ontology codes used in Microsoft Academic Graph. Finally, we see our work as building a potential methodological bridge between using natural language data and metadata in scientometric research. Our approach suggests a principled way in which representations of metadata and natural language can be integrated and compared more naturally to derive even greater insights.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

1. Gronenborn, A. M. Integrated multidisciplinarity in the natural sciences. *J. Biol. Chem.* **294**, 18162–18167, DOI: 10.1074/jbc.AW119.008142 (2019).
2. Hu, Z., Tian, W., Guo, J., Wang, X. *et al.* Mapping research collaborations in different countries and regions: 1980–2019. *Scientometrics* 1–17 (2020).
3. Leydesdorff, L., Wagner, C. S. & Bornmann, L. Betweenness and diversity in journal citation networks as measures of interdisciplinarity—a tribute to Eugene Garfield. *Scientometrics* **114**, 567–592 (2018).
4. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, 1999).
5. Núñez, R. *et al.* What happened to cognitive science? *Nature Human Behaviour* **3**, 782–791 (2019).
6. Choi, B. C. & Pak, A. W. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clin. investigative medicine* **29**, 351–364, DOI: 10.1016/j.jaac.2010.08.010 (2006).
7. Lietz, H. Drawing impossible boundaries: field delineation of social network science. *Scientometrics* 1–36 (2020).
8. Bartol, T., Budimir, G., Juznic, P. & Stopar, K. Mapping and classification of agriculture in Web of Science: other subject categories and research fields may benefit. *Scientometrics* **109**, 979–996 (2016).
9. Kozlowski, D., Dusdal, J., Pang, J. & Zilian, A. Semantic and relational spaces in science of science: deep learning models for article vectorisation. *Scientometrics* 1–30 (2021).
10. Shen, Z., Ma, H. & Wang, K. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, 87–92 (2018).
11. Chawla, D. S. Microsoft Academic Graph is being discontinued. what's next? Nature Index - News (2021). https://www.natureindex.com/news-blog/microsoft-academic-graph-discontinued-whats-next.
12. Solomon, G. E. A., Carley, S. & Porter, A. L. How multidisciplinary are the multidisciplinary journals Science and Nature? *PLOS ONE* **11**, 1–12, DOI: 10.1371/journal.pone.0152637 (2016).
13. Xie, Z., Li, M., Li, J., Duan, X. & Ouyang, Z. Feature analysis of multidisciplinary scientific collaboration patterns based on PNAS. *EPJ Data Sci.* **7**, DOI: 10.1140/epjds/s13688-018-0134-z (2018).
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. neural information processing systems* 3111–3119 (2013).
15. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543 (2014).
16. Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196 (2014).
17. Grover, A. & Leskovec, J. Node2vec: scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 855–864 (2016).
18. He, Y.-H., Jejjala, V. & Nelson, B. D. hep-th (2018). 1807.00735.
19. Chinazzi, M., Gonçalves, B., Zhang, Q. & Vespignani, A. Mapping the physics research space: a machine learning approach. *EPJ Data Sci.* **8**, DOI: https://doi.org/10.1140/epjds/s13688-019-0210-z (2019).
20. Zhang, Y., Zhao, F. & Lu, J. P2v: large-scale academic paper embedding. *Scientometrics* **121**, 399–432, DOI: https://doi.org/10.1007/s11192-019-03206-9 (2019).
21. Song, L., Cheong, C. W., Yin, K., Cheung, W. K. & CM, B. Medical concept embedding with multiple ontological representations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4613–4619 (AAAI Press, 2019).
22. Rafols, I. & Meyer, M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics* **82**, 263–287 (2010).
23. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. reports* **9**, 1–12 (2019).
24. Tahmasebi, N., Borin, L. & Jatowt, A. Survey of computational approaches to lexical semantic change. In *Computational approaches to semantic change*, 1–91 (Language Science Press, Berlin, 2021).

25. Kutuzov, A., Øvrelid, L., Szymanski, T. & Velldal, E. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018).

26. Li, B. *et al.* Scaling word2vec on big corpus. *Data Sci. Eng.* **4**, 157–175 (2019).

27. Basile, P. & McGillivray, B. Exploiting the web for semantic change detection. In *International Conference on Discovery Science*, 194–208 (Springer, 2018).

28. Bizzoni, Y., Mosbach, M., Klakow, D. & Degaetano-Ortlieb, S. Some steps towards the generation of diachronic WordNets. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 55–64 (2019).

29. Lenci, A. Distributional models of word meaning. *Annu. review Linguist.* **4**, 151–171 (2018).

30. Wang, B., Wang, A., Chen, F., Wang, Y. & Kuo, C.-C. J. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal Inf. Process.* **8**, e19, DOI: 10.1017/ATSIP.2019.12 (2019).

31. Socher, R., Bauer, J., Manning, C. D. & Ng, A. Y. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 455–465 (Association for Computational Linguistics, Sofia, Bulgaria, 2013).

32. Socher, R. *et al.* Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642 (Association for Computational Linguistics, Seattle, Washington, USA, 2013).

33. Luo, Y., Zhao, H. & Zhan, J. Named entity recognition only from word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8995—9005, DOI: 10.18653/v1/2020.emnlp-main.723 (Association for Computational Linguistics, Online, 2020).

34. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186, DOI: 10.18653/v1/N19-1423 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).

35. Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. & Tahmasebi, N. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23, DOI: 10.18653/v1/2020.semeval-1.1 (International Committee for Computational Linguistics, Barcelona (online), 2020).

36. Hook, D. W., Porter, S. J. & Herzog, C. Dimensions: building context for search and evaluation. *Front. Res. Metrics Anal.* **3**, 23 (2018).

37. Khan, G. F., Lee, S., Park, J. Y. & Park, H. W. Theories in communication science: a structural analysis using webometrics and social network approach. *Scientometrics* **108**, 531–557 (2016).

38. Devarakonda, S., Korobskiy, D., Warnow, T. & Chacko, G. Viewing computer science through citation analysis: Salton and Bergmark redux. *Scientometrics* **125**, 271–287 (2020).

39. Lakshmanan, V. *Data Science on the Google Cloud Platform: Implementing End-to-end Real-time Data Pipelines – from Ingest to Machine Learning* (O'Reilly, Sebastopol, CA, 2018).

40. Řehůřek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50 (ELRA, Valletta, Malta, 2010).

41. Lakshmanan, V., Robinson, S. & Munn, M. *Machine learning design patterns* (O'Reilly Media, 2020).

42. Chugh, M., Whigham, P. A. & Dick, G. Stability of word embeddings using Word2Vec. In Mitrovic, T., Xue, B. & Li, X. (eds.) *AI 2018: Advances in Artificial Intelligence*, 812–818 (Springer International Publishing, Cham, 2018).

43. Davies, M. Corpus of Contemporary American English (COCA) (2008).

44. Davies, M. Top 60,000 lemmas. https://www.wordfrequency.info/samples/lemmas_60k.txt (2021). Accessed: 2021-07-27.

45. Nayak, N., Angeli, G. & Manning, C. D. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, 19–23 (2016).

46. Schnabel, T., Labutov, I., Mimno, D. & Joachims, T. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 298–307 (2015).

47. Drozd, A., Gladkova, A. & Matsuoka, S. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, 3519–3530 (2016).

48. Greenacre, M. *Correspondence analysis in practice* (Chapman and Hall/CRC, Boca Raton, 2007), 2nd edn.

49. Mann, H. Non-parametric tests against trend. *Econometrica* **13**, 163–171 (1945).

50. Kendall, M. *Rank Correlation Methods* (Charles Griffin, London, 1975), 4th edn.

51. Hamilton, W., Leskovec, J. & Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (2016).

52. Gower, J. C. Generalized Procrustes analysis. *Psychometrika* **40**, 33–51 (1975).

53. Shoemark, P., Ferdousi Liza, F., Nguyen, D., Hale, S. & McGillivray, B. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 66–76 (Association for Computational Linguistics, 2019).

54. Killick, R., Fearnhead, P. & Eckley, I. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107**, 1590–1598 (2012).

55. Sinha, A. *et al.* An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA*, 243–246, DOI: http://dx.doi.org/10.1145/2740908.2742839 (2015).

56. Tsakalidis, A., Bazzi, M., Cucuringu, M., Basile, P. & McGillivray, B. Mining the uk web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1212–1221 (2019).

57. Gries, S. T. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Stud. Lang.* **36**, 477–510 (2012).

58. Jenset, G. B. Mapping meaning with distributional methods: A diachronic corpus-based study of existential *there*. *J. Hist. Linguist.* **3**, 272–306 (2013).

59. Zou, W. Y., Socher, R., Cer, D. & Manning, C. D. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1393–1398 (2013).

Author contributions statement

All authors contributed figures. All authors reviewed the manuscript. All authors contributed to conceptualisation, methodology, software, and analysis. B.M and G.B.J wrote the main manuscript text. G.B.J was responsible for data curation.

Competing interests

Gard Jenset is employed by Springer Nature. Barbara McGillivray, Donna Schut, and Khalid Salama declare no competing interests.

Ethical approval

All the research was conducted in accordance with relevant guidelines and regulations. Ethic approval is not applicable.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Data availability

The datasets generated during and/or analyzed during the current study are available in the Figshare repository, https://kcl.figshare.com/articles/dataset/Data_for_Investigating_patterns_of_change_stability_and_interaction_among_scientific_disciplines_using_embeddings_/20297217, DOI: 10.18742/20297217.

5 List of Figures

Figure 1. Average cosine similarity between FoR embeddings over time, by clusters of STEM (left) and HSS (right) disciplines.

Figure 2. Average neighbourhood size of FoR embeddings over time. HSS discipline embeddings are in blue and STEM discipline embeddings are in orange.

Figure 3. Average neighbourhood size of FoR embeddings over time, by cluster.

Figure 4. Time series of the cosine similarity of the embeddings of two disciplines that changed profile in the period under consideration: Communication and Media Studies (light grey) and Computer Hardware (black).

Figure 5. Data extraction, processing, and analysis steps.

Figure 6. Biplot of PCA-transformed embedding vectors. The arch or "horseshoe" shape indicates a gradient, which correlates with the natural sciences / humanities and social sciences continuum. A vertical line across the middle of the plot would approximately correspond to the STEM/HSS distinction. The x-axis accounts for 41% of the variation in the data, while the y-axis accounts for 27%. Labels are higher-level (level-2) codes. Label codes have been shortened for readability.