Will This Idea Spread Beyond Academia? Understanding Knowledge Transfer of Scientific Concepts across Text Corpora

Hancheng Cao[*]
Stanford University
hanchcao@stanford.edu

Mengjie Cheng[*]
Harvard Business School
macheng@hbs.edu

Zhepeng Cen[*]
Carnegie Mellon University
zcen@andrew.cmu.edu

Daniel A. McFarland
Stanford University
dmcfarla@stanford.edu

Xiang Ren
University of Southern California
xiangren@usc.edu

Abstract

What kind of basic research ideas are more likely to get applied in practice? There is a long line of research investigating patterns of knowledge transfer, but it generally focuses on *documents* as the unit of analysis and follow their transfer into practice for a specific scientific domain. Here we study translational research at the level of *scientific concepts* for *all scientific fields*. We do this through text mining and predictive modeling using three corpora: 38.6 million paper abstracts, 4 million patent documents, and 0.28 million clinical trials. We extract scientific concepts (i.e., phrases) from corpora as instantiations of "research ideas", create concept-level features as motivated by literature, and then follow the trajectories of over 450,000 new concepts (emerged from 1995-2014) to identify factors that lead only a small proportion of these ideas to be used in inventions and drug trials. Results from our analysis suggest several mechanisms that distinguish which scientific concept will be adopted in practice, and which will not. We also demonstrate that our derived features can be used to explain and predict knowledge transfer with high accuracy. Our work provides *greater understanding of knowledge transfer* for researchers, practitioners, and government agencies interested in encouraging translational research.

1 Introduction

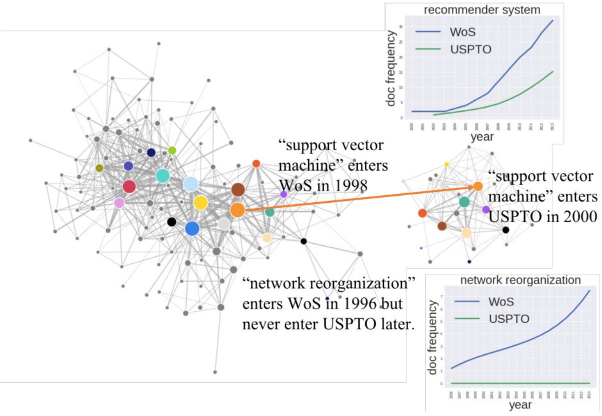

recommender system
doc frequency
WoS
USPTO
year
"support vector machine" enters WoS in 1998
"support vector machine" enters USPTO in 2000
"network reorganization" enters WoS in 1996 but never enter USPTO later.
network reorganization
doc frequency
WoS
USPTO
year

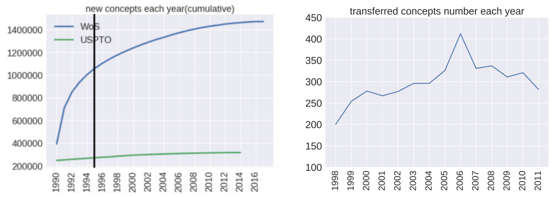
Figure 1: An illustration of scientific concept's "knowledge transfer" from basic research to practice use: we analyze individual concept's time-varying features (e.g., popularity) and relative positions with other concepts (i.e., co-occurrence) to understand the key mechanisms behind knowledge transfer, using Web of Science research papers, USPTO patents and clinical trial documents.

Science generates a myriad of new ideas, only some of which find value in practical uses (Backer, 1991; Lane and Bertuzzi, 2011). Large government agencies (*e.g.*, NSF, NIH) pour billions of dollars into basic research in the hopes that it will span the research-practice divide so as to generate private sector advances in technologies (Narin and Noma, 1985), social policies (McDonald and Mair, 2010), and pharmaceuticals (Berwick, 2003). To this end, these agencies increasingly seek to nurture "translational research" that succeeds at extending, bridging and transforming basic research so it finds greater applied value (Li et al., 2017). Surrounding this effort has arisen a line of research that tries to identify when, where, and how academic research influences science and technological invention (Backer, 1991; Li et al., 2017).

However, prior research efforts are limited in their ability to understand and facilitate the translation of research ideas. This is partially due to a shortage of data, a biased focus on successful examples, and specialized modeling paradigms. In practice, only a small proportion of knowledge outputs are successfully translated into inventive outputs ($\sim$ 2.7% concepts from WoS to patent, and $\sim$ 11.3% concepts from WoS to clinical trials, according to our data analysis). Previous studies conduct post-hoc analyses of successful scientific-technological linkages, but are unable to explain why the majority of scientific innovations do not

[*] Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1746–1757
November 16 - 20, 2020. ©2020 Association for Computational Linguistics



(a) New concepts each year (b) # transferred concepts

Figure 2: Concepts' knowledge transfer over time

Table 1: Examples of extracted scientific concepts

Transferred Concepts	Non-transferred Concepts
Internet, world wide web,	ethnographic exploration,
interactive visualization, web server,	immersive virtual reality,
gpu, recombinant protein production,	european maize,
hcci engine, cloud service,	institutional demand,
artificial magnetic conductor	automatic imitation,
multifunctional enzym,	network reorganization,
tissue remodeling,	human capital,
single photon detector	amercian theatre

transfer into technological inventions. Additionally, prior work mostly look at document-level linkages across research and applied domains, i.e., citations from patents into research papers or shared inventors across them, rather than diving into the document content where ideas are discussed (Narin and Noma, 1985; Ahmadpoor and Jones, 2017). Documents entail many ideas, and linkages across them loosely capture which intellectual innovation is in focus and being transferred.

By contrast, we conceptualize knowledge transfer in terms of scientific concepts, rather than documents associated to particular desirable outcomes, and demonstrate the importance of our derived features in knowledge transfer through machine learning model in a large-scale original dataset.

In this paper, we focus on studying patterns behind knowledge transfer from academia to research. We use "knowledge transfer from academia to research" in our study to mean a "concept's transfer from research papers to patent documents/clinical trials", or a concept that first appear in academia later get used a non-trivial frequency (decided by a pre-defined threshold) in practical outlets (patents, clinical trials). [1] In scientific writing, a scientific concept is a term or set of terms that have semantically coherent usage and reflect scientific entities – e.g., curricula, tools, programs, ideas, theories, substances, methods, processes, and propositions, which are argued to be the basic units of scientific discovery and advance (Toulmin). We use the titles and abstracts of 38.6 million academic publications from the Web of Science (WoS) to identify 0.45 million new scientific concepts emerging between 1995 to 2014 through state of the art phrase mining techniques (AutoPhrase), and follow their trajectories in 4 million patent documents of the United States Patent and Trademark Office (USPTO), and 0.28 million clinical trials from U.S. National Library of Medicine.

In our analysis, we compare the properties of new scientific concepts that successfully transfer into patents with those that did not. We find that (a) the intrinsic properties of ideas and their temporal behavior, and (b) relative position of the ideas are the two mechanisms that determine whether an idea could transfer successfully. In particular, we find new engineering-focused scientific concepts situated in emotionally positive contexts are more likely to transfer than other concepts. Furthermore, increased scientific hype and adoption across scientists, as well as usage in interdisciplinary venues over time, are early signs of impending knowledge transfer into technological inventions. Finally, we find that new concepts positioned close to concepts that already transferred into patents are far more likely to transfer than their counterparts. Based on the derived features, we further built model to predict the likelihood of knowledge transfer from papers to patents/clinical trials at individual concept level, and demonstrated our derived feature can achieve great performance, indicating that our proposed features can explain majority of the knowledge transfer cases.

Contributions Our main contributions are summarized as follows: (1) To the best of our knowledge, we present the first ever research that aims at understanding knowledge transfer at a large scale, using multiple corpora. (2) We are the first to leverage text mining techniques to understand transfer on scientific concept level, rather than document level. (3) We systematically analyzed the differences between transferable and non-transferable concepts, and identified the key mechanisms behind knowledge transfer. We showed our derived insights can help explain and predict knowledge transfer with high accuracy.

2 Data Preparation and Processing

In this section we introduce the dataset used in our study (Sec. 2.1), and present the concept extraction process (Sec. 2.2).

[1] We use knowledge transfer, concept transfer and idea transfer interchangeably throughout the paper

2.1 Collection of Text Corpora

Research Papers from WoS. We used scientific concepts extracted from Web of Science (WoS) as representation of knowledge in academia. We use the complete corpus from WoS (1900-2017) totaling 38,578,016 papers.

Patent Documents from USPTO. We used concepts extracted from 4,721,590 patents in the United States Patent and Trademark Office (USPTO) from 1976 to 2014 to represent general knowledge in the application domain.

Clinical Trials. We used concepts extracted from 279,195 clinical trials from U.S. National Library of Medicine[2] (1900 to 2018) to represent bio & health sciences knowledge used in practice.

More details of the leveraged datasets are further elaborated in Appendix A. Note that our study inevitably suffer from data bias. For instance, not all practitioners will patent their idea, or file clinical trials, and that some clinical trails and patents are unused, thus there will be some false positives and negatives of 'transferred' labels through our approach. Yet so far patent and clinical trial have been demonstrated to be the best proxy to study translational science from research to practice (Ahmadpoor and Jones, 2017). Moreover, we have tried our best to mitigate such bias by investigating transfer patterns in both patent-heavy and patent-light fields, where we found very similar patterns emerge.

2.2 Scientific Concept Extraction

Using titles and abstracts of articles, patents and clinical trials, we employ phrase detection technique AutoPhrase (Shang et al., 2018), to identify key concepts in the two corpora and trace their emergence and transfer across domains over time. Phrase detection identifies 1,471,168 concepts for research papers, 316,442 concepts for patents, and 112,389 concepts for clinical trials. Some samples of transferred concepts and non-transferred concepts extracted from WoS and USPTO by phrase detection are shown in Table. 1. We observe that phrase detection results in high-quality concepts (92% are labelled as high quality through our evaluations) that are suitable to investigate knowledge transfer across domains. Details of the phrase detection techniques, cleaning and evaluations are further discussed in Appendix B.

New Concept Identification. The focus of this study is on new concepts and their careers. However, our sample of 1.5 million distinct concepts occur at any time in the corpus, some of which emerged long ago and others more recent. To avoid left-censoring issue (certain concepts appear before the start time of the recorded data thus we do not fully observe their behaviors) and identify 'real' new concepts, we aggregate (or "burn in") the set of concepts over time, and count the number of new concepts that arrive each year. Early papers (starting 1900) identify many new concepts, but this quickly decelerates by around 1995 and then assumes a linear growth in vocabulary afterwards (see Fig. 2(a)). To identify that point, we aggregate the set of concepts every year with prior years until the rate of new concepts' introduction is approximately linear and stable. The point occurs after 1995, when 0.45 million scientific concepts are left. Then we follow knowledge transfer via these new scientific concepts, and find only ∼2.7% of all concepts get transferred to patent, and only ∼11.3% of bio & health concepts get transferred to clinical trials across years. The number of transferred concepts each year from WoS to USPTO is illustrated in Fig. 2(b).

3 Feature Creation and Analysis

Based on the concepts extracted from research papers, patent and clinical trial documents, we first create concept level features as motivated by prior literature on knowledge diffusion, and present a large-scale data analysis on transferred and non-transferred concepts to better understand properties facilitating the knowledge transfer process. Here we present transfer patterns from research paper to patent and omit clinical trial due to page limit [3].

3.1 Intrinsic Properties of Concepts

Motivated by previous works on knowledge diffusion and transfer, we extracted intrinsic concept features that would most likely facilitate a scientific idea's transfer into technological inventions, which can be classified into four categories: 1) hype features (Latour, 1987; Rossiter, 1993), 2) bridge positioning features (Shi et al., 2010; Kim et al., 2017), 3) ideational conditions (Berger and Heath, 2005), and 4) technological resonance (Narin and Noma,

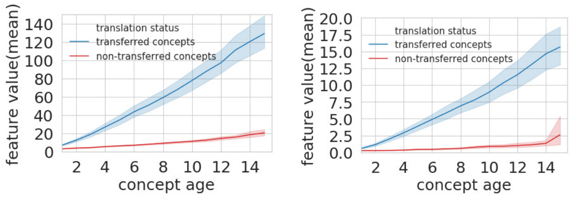
[2] Retrieved from clinicaltrials.gov

[3] We find very similar transfer patterns emerge from research paper to clinical trial.

1985)[4]. The four sets of features represent the characteristic of individual concept from diverse angles, and as we will show signify the differences between transferred and non-transferred concepts, both in mean value (Appendix E) and temporal behavior.

To illustrate concepts' temporal behavior over time, we plot the feature curves of transferred and non-transferred concepts over concept age. Details could be found in Figs. 3-8.

Hype. This group of features draws on prior work concerning concept hype (Latour, 1987; Acharya et al., 2014; Larivière et al., 2014). We include two features: the adopter size using the concept, and the degree to which authors repeatedly use the concept. We measure *adopter size* as the total number of authors who employ a concept in a particular year, and *author repeated usage* as the total number of previous authors continuing to use the concept.



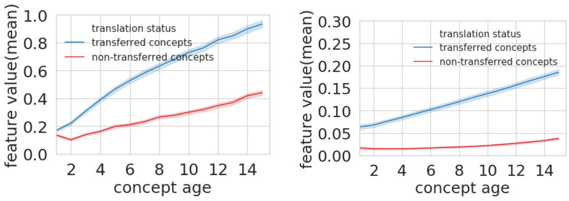
(a) Adopter size (b) Author repeated usage

Figure 3: Hype features.

We found that transferred concepts generally demonstrate higher numbers of adopters and repeated usage. Furthermore, transferred concepts attract adopters at a faster rate than non-transferred concepts. We also found that transferred concepts are repeatedly used much more often by the previous authors when controlling for concept age. What's more, we observe an increasing gap with regard to 'hype' features between transferred and non-transferred concepts over time, possibly due to the preferential attachment effect (Newman, 2001).

Bridge Positioning. This group of features identify the disciplinary placement of concepts. Previous works argue that knowledge transfer is facilitated when ideas are placed at the boundary of fields and in fields especially relevant to technological invention ((Shi et al., 2010)). Here we include two features: *discipline diversity* and *engineering relation* in this group. *Discipline diversity* is computed as a concept's average entropy across NRC discipline subject codes (sociology, math, economics, etc.), and *engineering relation* is computed as the proportion of engineering fields among all the fields using the concept.

We found transferred concepts are more likely to be used in interdisciplinary and engineering venues. Moreover, transferred concepts gained greater interdisciplinary attention over time compared to non-transferred concepts, as shown in Fig. 4. The finding is consistent with the assumption that transferred concepts are likely to achieve a more diverse audience than non-transferred concepts. Engineering focused concepts also achieved a higher knowledge transfer rate, which supports our hypothesis that knowledge transfer is facilitated when ideas are placed at the boundary of fields especially relevant to technological invention like engineering (e.g. mechanical engineering). Once again, we observed the difference of 'bridge positioning' feature values between transferred and non-transferred concepts increase over time.



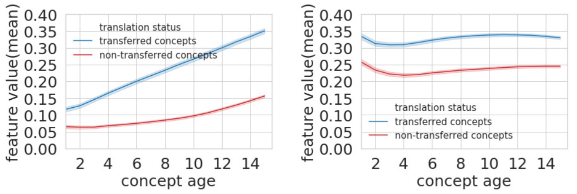
(a) Discipline diversity (b) Engineering Focus

Figure 4: Bridge positioning features.

Ideational Conditions. This group of features represents the semantic context and expression of a concept. How the concept is related to other concepts and the style with which the concept is expressed can both influence the diffusion and transfer process (Hamilton et al., 2016). Here we select *emotionality*, and *accessibility* in this group, and calculated them through LIWC and Dale Chall metric (details in Appendix C).

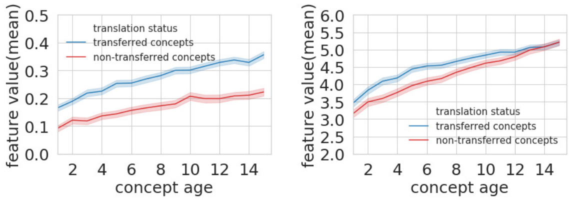
We found transferred concepts are embedded in more emotional context, and described in more difficult language, compared to non-transferred counterparts. In a similar way, we plot ideational condition features over time for transferred concepts and non-transferred concepts in Fig. 5. We found that transferred concepts were consistently placed in increasingly positive contexts and conveyed in more difficult language over time, compared to non-transferred concepts, although the accessibility gap decreases over time.

[4] While these features are not exhaustive, to the best of our knowledge they are the key factors most salient to the knowledge diffusion as discussed in literature



(a) Journal Linkage (b) University-Industry Relationship

Figure 6: Science technology linkage.



(a) Emotionality (b) Accessibility

Figure 5: Ideational conditions features.

Technological Resonance. This group of features quantifies the extent to which a concept is established within an environment conducive to link scientific publications with patents and other outcomes (Narin and Noma, 1985; Tijssen, 2001). We measure this as *journal linkage* and *university-industry relationship* in our study. *journal linkage* is computed as the percentage of journals where the concept is situated that have been cited by patents before. *university-industry relationship* is calculated as the proportion of industry-affiliated authors out of all the authors employing the term each year. Should a scientific concept be in a high bridging space like these, they will more likely transfer.

Transferred concepts are more likely to be mentioned in journals that have been cited by patents, and this relationship strengthens over time. We also find that if a concept is associated with more industry-affiliated authors, the concept has a higher potential to transfer. While the industry-affiliate author percentage between transferred and non-transferred concepts remain relatively stable, the gap between them with regard to journal linkage gets greater over time.

3.2 Relative Position in Concept Co-occurrence Graph

In addition to the above features, we investigate the same data with a relational approach (Hofstra et al., 2019). Intuitively, how a concept get positioned/co-used with other concepts may be associated with knowledge transfer.

As a motivating example, we plot the local co-occurrence network of concept *search engine* in Fig. 7. The central grey circle is *search engine*, and the orange nodes denote the transferred neighbor concepts while the blue denotes the non-transferred one. *Search engine* first emerged in WoS in 1992 and entered USPTO in 1998. Coincidentally, the percentage of its transferred neighbors increased rapidly right before 1998, which indicates the neighboring concepts that get co-used with a concept may embed useful signals that explain concept transfer. The consistency between co-occurrence network and transfer status is also common in other concepts.

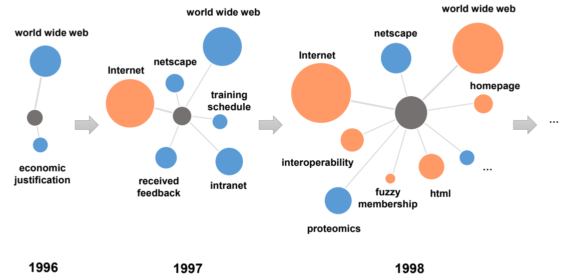
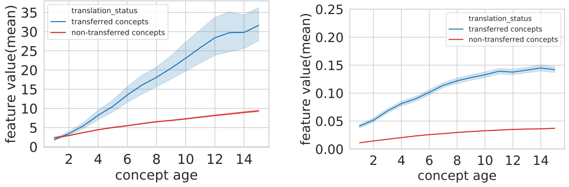


Figure 7: Illustration of the dynamic graphs that capture interactions between *search engine* and its co-occurrence concepts. The orange circles denote transferred concepts while the blue denotes non-transferred ones; the circle size represents the node degree.



(a) Weighted Degree (b) Weighted Percentage

Figure 8: Graph features.

To facilitate analysis, we construct a dynamic graph G for concept co-occurrence. Each node in graph denotes a concept which has occurred in the corpus. Each edge between two nodes indicates the two concepts co-occur in at least one document in the corpus, and we define the edge weight as the number of documents the two concepts co-occur. We sort all documents by year and construct a graph at each time-stamp, then we will get a set of graphs $\{\mathcal{G}\} = \{\mathcal{G}^{(1)}, \cdots, \mathcal{G}^{(t)}\}$ as *dynamic concept co-occurrence graph*. This set of graphs reflects the dynamic succession of concepts' neighbors and provides us with extra temporal information on local graph structures. Based on dynamic concept co-occurrence network, we derived two graph features: *weighted degree* and *weighted percentage of transferred neighbors* as specified in Appendix D.

The curves of the two features over time are shown in Fig. 8(a) and Fig. 8(b). We find that transferred concepts indeeed have higher weighted degrees and weighted percentages compared to

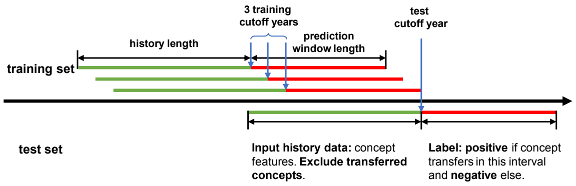

3 training cutoff years
test cutoff year
history length
prediction window length
training set
test set
Input history data: concept features. **Exclude transferred concepts**.
Label: positive if concept transfers in this interval and **negative** else.

Figure 9: Illustration of the concept's knowledge transfer prediction task and the partition of training/test sets. The green lines denote historical input while the red lines denote prediction window. The prediction intervals of training and test sets should not overlap.

non-transferred ones, which indicates the importance of utilizing concept co-occurrence for knowledge transfer prediction.

Field Comparison & Feature Correlation We further carried out analysis on feature correlation, and comparison across fields, which is discussed in detail in Appendix F and Appendix G.

Summary. Results of our data analysis support the conclusion that knowledge transfer is not by chance but follows specific patterns. Whether a concept will transfer from research to practice in the immediate future depends largely on their (a) individual properties over time, and (b) relative positions with respect to other concepts.

4 Predictive Analysis of Features

So far we have systematically analyzed the potential factors that reflect the process of knowledge transfer from research to practice. But how well can these features explain and predict knowledge transfer in practice? In this section, we seek to shed light on this question through predictive analysis.

4.1 Prediction Task Formulation

Will a scientific concept transfer from academic papers to patent documents in the next X years? Here we consider the predictive task which aims to predict concept transfer status given all observed historical data. As there can be only two potential outcomes — either the concept transfers or not — the proposed prediction task is essentially a binary classification problem. We label a concept as transferred if it first originates in research papers and later get used at least 5 times in practical outlets (patents, clinical trials) within X years after the concept's birth in research papers.

We denote all N concepts' time-series attributes at one particular time-stamp as $X \in R^{N \times N_x}$, where N_x is the dimension of attributes. As shown in Fig. 9, the goal of the transfer prediction problem is to construct a function $f(\cdot)$ mapping historical time-series attributes to the future transfer probability of concept,

$$f : \left[\left(\mathbf{x}_i^{(t-k)}\right), \cdots, \left(\mathbf{x}_i^{(t-1)}\right)\right] \rightarrow P\left(y_i^{(t)} = 1 \middle| \cdot\right)$$

where $\mathbf{x}_i = X_{i,:}$ denotes the attribute vector of concept i, $P\left(y_i^{(t)} \middle| \cdot\right)$ is the conditional probability and k is input history length. y_i denotes transfer status of concept i in next T years, *i.e.*, the ground truth label of $y_i^{(t)}$ is 1 if it transfers in $t \sim t+T-1$ else 0, and T denotes *prediction window length*. Particularly, we note t as *cutoff year* and our model inputs the attributes previous to this time-stamp and predicts future transfer probability. For simplicity, we denote $P\left(y_i^{(t)} = 1 \middle| \cdot\right)$ as $p_i^{(t)}$.

Accordingly, if the true transfer status is $\mathrm{y}_i^{(t)}$, the loss function for cutoff year t is

$$\mathcal{L} = -\sum_i \left[\mathrm{y}_i^{(t)} \log p_i^{(t)} + \left(1 - \mathrm{y}_i^{(t)}\right) \log\left(1 - p_i^{(t)}\right)\right]$$

4.2 Prediction Models

Feature based Model. We use logistic regression (LR) as an interpretable model. To better validate our finding, we also run a mixed effects logistic regression detailed in Appendix I, a form of Generalized Linear Mixed Model (GLMM), to help explain variance both within-concept and across-concept. The results from the mixed effects logistic regression are nearly identical with our findings from the vanilla logistic regression, except for slight changes in the magnitude of coefficients, so we only report performance of LR in our analysis.

Deep Sequence Model. To model a concept's temporal features, *i.e.* time-series attributes, we further propose RNN sequence models. According to Sec.3, some time-series features are strongly related to potential transfer; therefore, we adopt Recurrent Neural Network (RNN) models (e.g., LSTM and GRU) which are built to capture temporal dependencies (Details in Appendix H).

5 Experiments

Experiment Set-up. We apply Z-score normalization on time-series attributes and divide dataset into training/test sets as Fig. 9 shows. Given test cutoff year t, we first ensure the prediction intervals (red line in Fig. 9) of training and test set have no overlap to avoid data leakage, and then use the latest three cutoff years as train cutoff years. For

Table 2: **Performances (mean AUC) of next-3-year transfer prediction for cutoff year 2008 from Web of Science to patent, and from Web of Science (Biology & Health Sciences papers) to clinical trial. All results are generated by 3-run experiments.**

	Patent		**Clinical Trial**	
Method / History length	**3 years**	**5 years**	**3 years**	**5 years**
LR w. graph features	0.792	0.792	0.656	0.661
LR w. all features	0.794	0.800	0.675	0.677
RNN w. graph features	0.793	0.797	0.706	0.726
RNN w. all features	0.803	0.809	0.715	0.734

instance, if test cutoff year is 2008 and prediction window is 5 years long, the latest training prediction interval should be 2003∼2007 and thus we use 2001, 2002, 2003 as training cutoff years. As concept transfer status is irreversible, we exclude all transferred concepts from test set but still use them to train.

Details of model training and hyperparameter settings are discussed in Appendix J. Here we primarily report experimental results on knowledge transfer prediction from WoS to USPTO, while using clinical trial as a robustness check.

Evaluation Metric. We adopt area-under-curve (AUC) as evaluation metric, which is not affected by data imbalance in test set.

5.1 Results

We first compared the performances of all aforementioned models for cutoff year 2008 on datasets: from WoS to patent, and from WoS bio & health science papers to clinical trials[5]. For each cutoff year, we ran two sets of experiments with training history lengths of 3 and 5 years and repeated 3 times for each experiment. The performances (mean AUC) are summarized in Table. 2.

Patent vs. Clinical Trial As a robustness check, we tested our model on both knowledge transfer from WoS to patent, and to clinical trial. We obtained consistent main attribute importance results based upon clinical trial data.

As can be observed from Table. 2, our derived features achieve good result, i.e. AUC 0.80, in predicting knowledge transfer, demonstrating knowledge transfer can be largely explained by our proposed mechanisms.

Study of Feature Importance. In Fig. 10, we further plot the standardized coefficients of each temporal feature from the logistic regression to understand how a specific attribute contributes to the knowledge transfer. We observed that author repeated usage, adopter size, and weighted graph degree are the three most important factors in influencing knowledge transfer.

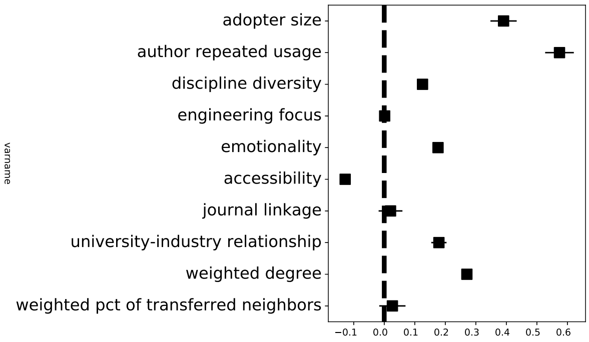


Figure 10: **Feature Importance Study**

Next, we studied feature importance in our proposed models, where we ran both models with different sets of features on knowledge transfer from WoS to patent. The result is summarized in Table. 3. As reflected by experiment results with RNN model, graph features achieve best prediction results compared to other feature sets, followed by “bridge positioning" features, “ideational conditions”, and “hype” features, suggesting that the relative position position of the concept in the semantics network is the single most useful feature set that explains concept transfer.

Study on Field Difference We studied the prediction performance of the proposed model in different fields. We partitioned the concepts used in Web of Science based on their field, trained and tested models separately using 5-year historical data as training inputs with train cutoff year 2003 and tested cutoff year 2008 for next 3-year prediction. We observed that it is easiest to predict knowledge transfer from academia to practice in humanity (AUC 0.973), followed by physical & math science (AUC 0.791), bio & health science (AUC 0.783), engineering (AUC 0.782), social science (AUC 0.706) and agriculture (AUC 0.633), which indicates our proposed mechanism can explain knowledge transfer quite well in most fields other than agriculture.

5.2 Sensitivity Analysis

Finally, we tested our proposed models under different settings on WoS to patent. We investigated whether our proposed transfer model is influenced as a result of 1) varying length of historical observations, 2) varying prediction time window, and 3) varying cutoff year.

[5] Note that for knowledge transfer to clinical trial, we excluded bridge positioning features since we focused on bio & health science only.

Table 3: Performance with different feature groups.

Method	AUC
LR w. “hype" features	0.629
LR w. “bridge positioning" features	0.681
LR w. “ideational conditions" features	0.662
LR w. “sci-tech linkage" features	0.670
LR w. graph features	0.792
LR w. all features	0.800
RNN w. “hype" features	0.641
RNN w. “bridge positioning" features	0.708
RNN w. “ideational conditions" features	0.686
RNN w. “sci-tech linkage" features	0.676
RNN w. graph features	0.797
RNN w. all features	0.809

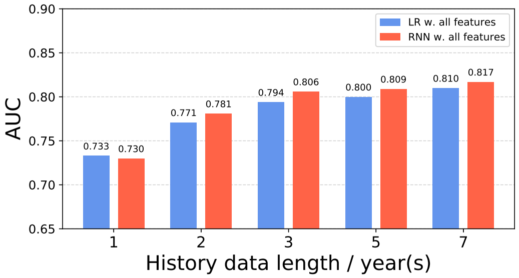


Figure 11: Performances of next-3-year knowledge transfer prediction with cutoff year 2008. We use different length of historical data as training data.

1. Length of Observation History. Fig. 11 demonstrates the effects of historical observation length on performance, where we selected 1 year, 2 years, 3 years, 5 years and 7 years of observation before cutoff year 2008 as training sets. We found that the longer the observation data, the better prediction result we will get for the transfer prediction, which can be explained by the fact that longer observation better captures knowledge transfer patterns. We also note that performance starts to plateau when observation length gets larger, indicating that longer training sets only provide limited additional signal. All this indicates knowledge transfer is most influenced by behavior of concepts in the recent few years.

2. Length of Prediction Time Window. Fig. 12 further illustrates the knowledge transfer prediction performance with prediction window of 1 year, 3 years and 5 years, representing the case when predicting whether a concept will transfer in next 1 year, 3 years or 5 years, respectively. To compare them fairly, we fix both training and testing cutoff years to keep time interval from training set to test set unchanged, which is different from the setting in previous experiments. As can be observed, prediction performance is consistently best when prediction window is 1 year, indicating the increasing difficulty in capturing long-term temporal pattern of knowledge transfer of our proposed mechanisms.

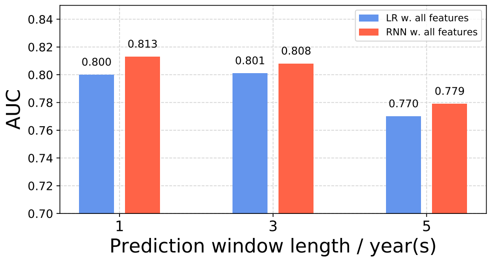


Figure 12: Performances under different lengths of prediction time windows. 5-year historical data was used for training with train cutoff year 2003 and test cutoff year 2008.

3. Cutoff Year. We also tested our model with different cutoff years (i.e., 2008, 2009 and 2010), representing knowledge transfer prediction with different training and testing sets. As illustrated by Fig. 13, Our model achieves consistent results, which further verifies the generalizability of our proposed knowledge transfer mechanism.

6 Related Work

Knowledge Diffusion and Transfer. Extensive studies have been dedicated to study the diffusion of knowledge (Kuhn, 1962; Rogers Everett, 1995; Hallett et al., 2019), and the transfer of knowledge from science to more applicable domains like technology (Narin and Noma, 1985; Tijssen, 2001). The majority of these studies focus on identifying contributing factors to knowledge diffusion and transfer (Rossiter, 1993; Azoulay et al., 2010; Shi et al., 2010; Kim et al., 2017). However, this line of work falls short in that (a) they focus primarily on successful / post-hoc knowledge diffusion and transfer, and little comparison of successful with unsuccessful transfer are presented, and (b) poorly specify what idea is being transferred because it focuses entirely at the document / invention level. In contrast, we contribute by empirically investigating properties of knowledge transfer through large-scale data analysis at the concept-level by using text mining approaches, through which we not only verified existing findings, but also revealed the significance of knowledge co-occurrence and ideational context in shaping knowledge transfer.

Temporal Sequence Modelling. As one fundamental task in behavior modelling and NLP, numerous techniques for modelling and predicting temporal sequence have been proposed (Kurashima et al., 2018; Pierson et al., 2018). In recent years, leveraging recurrent neural network (RNN) (Mikolov et al., 2010) and its variants (e.g. LSTM, GRU) (Chung

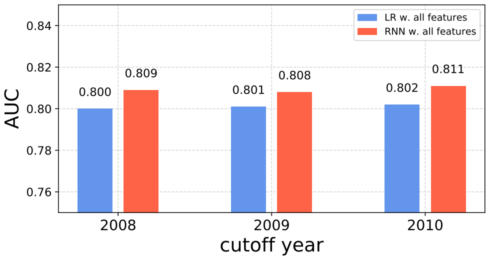


Figure 13: Performances of next-3-year transfer prediction under different cutoff years. We use 5-year historical data as training inputs.

et al., 2014) for sequence modelling has been especially popular due to the structure's expressive power of temporal dynamics, and has been widely used in time series modelling (Lai et al., 2018).

7 Conclusion

In this paper, we systematically studied the process and properties of knowledge transfer from research to practice. Specifically, we used a sample of 38.6 million research papers, 4 million patents and 280 thousand clinical trials, where we leveraged AutoPhrase to extract concepts from text and focus on the applicable career of nearly 450,000 new scientific concepts that emerged from 1995-2014. Through extensive analysis, we found that 'transferable' ideas distinguish themselves from 'non-transferable' ideas by their (a) intrinsic properties and their temporal behavior, and (b) their relative position to other concepts. Through predictive analysis, we showed our proposed features can explain majority of transfer cases. Our research not only provides significant implications for researchers, practitioners, and government agencies as a whole, but also introduces a novel research question of real world impact for computer scientists.

Acknowledgments

We thank Dan Jurafsky for helpful discussions and suggestions. We also thank the associate editor and reviewers for their valuable comments. Xiang Ren is supported in part by the Defense Advanced Research Projects Agency with award W911NF-19-20271 and NSF SMA 18-29268.

References

Anurag Acharya, Alex Verstak, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. 2014. Rise of the rest: The growing impact of non-elite journals. *arXiv preprint arXiv:1410.2217*.

Mohammad Ahmadpoor and Benjamin F Jones. 2017. The dual frontier: Patented inventions and prior scientific advance. *Science*, 357(6351):583–587.

Pierre Azoulay, Joshua S Graff Zivin, and Jialan Wang. 2010. Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589.

Thomas E Backer. 1991. Knowledge utilization: The third wave. *Knowledge*, 12(3):225–240.

Jonah A Berger and Chip Heath. 2005. Idea habitats: How the prevalence of environmental cues influences the success of ideas. *Cognitive Science*, 29(2):195–221.

Donald M Berwick. 2003. Disseminating innovations in health care. *Jama*, 289(15):1969–1975.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Tim Hallett, Orla Stapleton, and Michael Sauder. 2019. Public ideas: Their varieties and careers. *American Sociological Review*, page 0003122419846628.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Bas Hofstra, Sebastian Munoz-Najar Galvez, Bryan He, Vivek V Kulkarni, and Daniel A McFarland. 2019. Diversity breeds innovation with discounted impact and recognition. *arXiv preprint arXiv:1909.02063*.

Minkyoung Kim, Daniel A McFarland, and Jure Leskovec. 2017. Modeling affinity based popularity dynamics. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 477–486. ACM.

Thomas S Kuhn. 1962. The structure of scientific revolutions. *Chicago and London*.

Takeshi Kurashima, Tim Althoff, and Jure Leskovec. 2018. Modeling interdependent and periodic real-world action sequences. In *Proceedings of the 2018 World Wide Web Conference*, pages 803–812. International World Wide Web Conferences Steering Committee.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104. ACM.

Julia Lane and Stefano Bertuzzi. 2011. Measuring the results of science investments. *Science*, 331(6018):678–680.

Vincent Larivière, George A Lozano, and Yves Gingras. 2014. Are elite journals declining? *Journal of the Association for Information Science and Technology*, 65(4):649–655.

Bruno Latour. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard university press.

Danielle Li, Pierre Azoulay, and Bhaven N Sampat. 2017. The applied value of public investments in biomedical research. *Science*, 356(6333):78–81.

Steve McDonald and Christine A Mair. 2010. Social capital across the life course: Age and gendered patterns of network resources 1. In *Sociological Forum*, volume 25, pages 335–359. Wiley Online Library.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Francis Narin and Elliot Noma. 1985. Is technology becoming science? *Scientometrics*, 7(3-6):369–381.

Mark EJ Newman. 2001. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Emma Pierson, Tim Althoff, and Jure Leskovec. 2018. Modeling individual cyclic variation in human behavior. In *Proceedings of the 2018 World Wide Web Conference*, pages 107–116. International World Wide Web Conferences Steering Committee.

Richard D Powers, William A Sumner, and Bryant E Kearl. 1958. A recalculation of four adult readability formulas. *Journal of Educational Psychology*, 49(2):99.

M Rogers Everett. 1995. Diffusion of innovations. *New York*, 12.

Margaret W Rossiter. 1993. The matthew matilda effect in science. *Social studies of science*, 23(2):325–341.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Xiaolin Shi, Jure Leskovec, and Daniel A McFarland. 2010. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 49–58. ACM.

Robert JW Tijssen. 2001. Global and domestic utilization of industrial relevant science: patent citation analysis of science–technology interactions and knowledge flows. *Research Policy*, 30(1):35–54.

Stephen Toulmin. Human understanding, volume i: The collective use and evolution of concepts (princeton, nj, 1972). *ToulminHuman Understanding*, 1.

A Details of Corpora Data

Research Papers from WoS. The corpus covers both STEM - bio & health sciences (16,252,065 papers), physical & math sciences (8,390,777 papers), engineering (5,000,172 papers) and agriculture (2,568,702 papers) and non-STEM subjects - humanity (3,219,403 papers) and social science (3,146,897 papers). The WoS dataset also includes meta-data for each paper, *i.e.*, author name, institution, subjects, publication year and citations, which help construct measures concerning different knowledge transfer mechanisms.

Patent Documents from USPTO. We include 4,721,590 granted patents from the main USPTO corpus (1976-2014) covering both STEM (science, technology, engineering and mathematics) and non-STEM subjects. Furthermore, the USPTO dataset also includes meta-data for each patent, *i.e.*, inventor name, institution, award year and citations, which help construct several science-technology linkage measures in the knowledge transfer process.

Clinical Trials from U.S. National Library of Medicine. The clinical trial dataset includes 279,195 government registered clinical trials ranging from 1900 to 2018. The corpus include the clinical trial title, their brief summary.

B Phrase Detection Techniques, and Evaluations

The phrase detection technique we adopted is AutoPhrase (Shang et al., 2018), a widely-used method that extracts frequent and meaningful phrases through weak supervision. AutoPhrase first extracts single-word and multi-word expressions (*i.e.* phrases) from the text corpus as candidate concepts, and then applies salient concept selection functions to pick the most representative concepts for each document. Given a word sequence (e.g., a sentence in an abstract), phrase segmentation can partition the word sequence into non-overlapping segments, each representing a cohesive semantic unit as illustrated in the first step in. We used default parameters as suggested by (Shang et al., 2018) in our study.

We further conducted data cleaning on the output of AutoPhrase to ensure the quality of the analyzed concepts. Specifically, we filtered out general phrases used for scientific writing (e.g. 'significantly important') and publisher name (e.g., 'Elsevier').

To quantitatively evaluate AutoPhrase for concept extraction, we randomly sampled 200 outputs and asked three experts to manually label whether they are good-quality concepts or not, where 184 (92%) are labelled as good-quality by all three experts.

C Calculations of Emotionality and Accessibility

Emotionality is computed as the percentage of words that were classified as either positive or negative where a concept is used. The number of positive and negative words in each article is counted by the Linguistic Inquiry and Word Count computer program (LIWC), which adopts a list of words classified as positive or negative by human readers beforehand (Pennebaker et al., 2015). We quantify accessibility through a variation of Dale Chall readability (Powers et al., 1958) by substituting the 'easy term list' with college student vocabulary. This widely used index variable essentially measures the difficulty or appropriateness of the writing for each article. We then weighted the average Dale Chall readability score of all the documents associated with a concept.

D Calculations of Graph Features

Given co-occurrence graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{s}, W\}$ defined in subsection 3.2, the *weighted degree* d_i and *weighted percentage of transferred neighbors* p_i are calculated as follows.

$$d_i = \sum_{j \in \mathcal{N}_i} W_{ji}; p_i = \frac{\sum_{j \in \mathcal{N}_i, s_j = 1} W_{ji}}{d_i}.$$

Different from unweighted features, weighted degree and weighted percentage use co-occurrence weights to stress the influence of high-frequency correlations. The edge weights is necessary especially when central concept co-occurs with a large amount of non-transferred concepts.

E Characteristic Difference between Transferred and Non-transferred via t-test

Table 4 illustrates the mean value for each attribute with regard to transferred concepts and non-transferred concepts, where we observe a statistically significant gap between the two groups.

Table 4: Mean attribute value for transferred and non-transferred concepts. The two groups demonstrate statistically significant difference through t-test. p<0.001: ***.

Concept attribute	Transferred	Non-transferred
Adopter size ***	89.6	14.9
Repeat usage ***	10.6	1.2
Discipline diversity ***	0.68	0.32
Engineering relation ***	0.15	0.04
Emotionality ***	0.31	0.20
Accessibility ***	4.88	4.82
Journal Linkage ***	0.28	0.15
Univ.-Industry relation ***	0.33	0.24

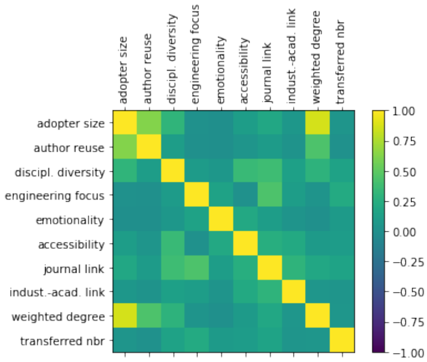


Figure 14: Correlations between different features.

F Field Comparison.

We studied transfer patterns in different fields. We identified the field of each concept as one of the six disciplines – biology & health sciences, physical & math sciences, the humanities, engineering, agriculture, and the social sciences, based on the maximum TF-IDF value component of its field use frequency distribution. While different fields demonstrate distinct transfer rates from research to patent — engineering 7.5%, physical & math sciences 1.9%, the social sciences 1.1%, bio & health sciences 0.96% (11.3% concepts in bio & health sciences transferred to clinical trial), agriculture 0.83% and the humanities 0.39% — we found that the aforementioned features show consistent patterns in different fields.

G Feature Correlation

We further studied the correlation between the extracted features. As illustrated in Fig.14, within concept individual level features, apart from hype features, and journal linkage/engineering focus, most features are rather independent. Meanwhile, graph feature 'edge weight' highly correlates with hype feature. In comparison, graph feature 'translated neighbor rate' brings signal not covered elsewhere, thus we conclude that modelling through both intrinsic values and graph is important.

H Details of Temporal Feature Model

The RNN model is given as

$$\mathbf{h}_{x,i}^{(t)} = \mathrm{RNN}\left(\mathbf{h}_{x,i}^{(t-1)}, \mathbf{x}_i^{(t)}\right) \tag{1}$$

where $\mathbf{h}_x$ is the hidden states of attributes. Suppose the concept transfer status is Markovian, then the model should be

$$P\left(y_i^{(t)} \middle| \mathbf{h}_{x,i}^{(t-k)}, \cdots, \mathbf{h}_{x,i}^{(t-1)}\right) = P\left(y_i^{(t)} \middle| \mathbf{h}_{x,i}^{(t-1)}\right) = g\left(\mathbf{h}_{x,i}^{(t-1)}\right)$$

Here we adopt GRU as RNN and one fully connected layer with sigmoid activation as classifier $g(\cdot)$.

I Details on Mixed Effect Logistic Regression

We ran a mixed effects logistic regression as a robustness check of logistic regression. Mixed effect logistic regression is a form of Generalized Linear Mixed Model (GLMM). Mixed effects logistic regression accounts for both within-concept variation (how concept use changes) and between-concept variation (how concept use differs on average), while a single measure of residual variance from the vanilla logistic regression can't account for both.

J Model Training and Hyperparameters.

To deal with the data imbalance problem – the positive samples (concepts which will transfer in the future) are much less than the negative, we over-sample positive samples to make their amount equal to negative ones in training set while keeping the original distribution in test set.

The hidden state size in RNN is set as 32. We experimented on different state sizes, and 32 achieved best performance on testing set.