Natural Language Processing and Innovation Research*

Antonin Bergeaud
HEC Paris and CEPR

Adam Jaffe
Brandeis University and NBER

Dimitris Papanikolaou
Kellogg School of Management and NBER

March 2025

Abstract

Innovation is central to models in economics, strategy, management, and finance, yet it remains difficult to measure due to its intangible and knowledge-based nature. Recent advancements in Natural Language Processing offer new methods to analyze textual artifacts, providing empirical insights into previously hard-to-measure aspects of innovation. This paper provides an overview of the current applications of these methods in empirical innovation research, highlights their transformative potential for reshaping how researchers study and quantify innovation, and discusses the critical challenges that accompany their use.

JEL Codes: O31, C80, O30
Keywords: Innovation; Natural Language Processing; Patents; Text-as-Data

*The authors are grateful to Gaetan de Rassenfosse, Matt Marx, and Vitaly Meursault for very helpful comments. We also thank Jacopo Marini for excellent research assistance. Corresponding author: Antonin Bergeaud - bergeaud@hec.fr. The paper was submitted to the Annual Review of Economics, Volume 18.

1 Introduction

Innovation is important but hard to measure: Innovation and firm strategy regarding innovation are central features of models in macroeconomics, microeconomics, strategy, management, and finance. But the key aspects of innovation that are important for economic models are not readily observable, as they do not generally appear on balance sheets, profit and loss statements, tax returns, or other reports. For example, new technologies developed by firms do not typically appear in company reports, and even if they do, it is difficult to say whether they represent a significant improvement or a marginal advance, whether they represent a new product or a new manufacturing process, or whether they represent a technology that augments or substitutes for labor.

As a result, there is a long tradition in innovation research of using artifacts such as patents and scientific papers to derive proxies for invention and new knowledge. Researchers have counted these artifacts, and used formal aspects of their creation such as citations and co-invention and co-authorship as further indicators of various steps in the innovation process. Most of these artifacts are made up of text, or have written documentation. These texts were used to understand and contextualize, but until recently it was not possible to do quantitative analysis of text, and therefore quantitative research based on these artifacts did not use the texts themselves in deeply meaningful ways.

In recent decades, computational methods have been developed to analyze text as data, and the power of these methods has increased dramatically in recent years. The application of these newly powerful methods to innovation-related texts can be used to create new indicators of innovation that are useful for modeling and has generated a large volume of new empirical innovation research. In this paper, we will provide an overview of these methods, discuss their potential to broaden and deepen innovation research, and highlight some important issues that need to be addressed to realize their full potential.

We use the general phrase “Natural Language Processing” ("NLP") to refer to these methods. We include within NLP both formulaic methods and those based on deep learning. Formulaic methods include approaches such as identifying specific words and word combinations or vectorizing the distribution of words in a document and then calculating the vector distance between documents. Neural networks—a major class of approaches within “artificial intelligence” (“AI”)— include both models trained for specific tasks such as identifying high-impact patents, and the more general models such as “Large Language Models” (“LLMs”) whose use has exploded in recent years.

Our goal in this paper is to provide a brief history of the development of these methods, a snapshot of their current use, and some conjectures as to their long run potential. While we will discuss specific papers, this literature is developing rapidly, so we have no hope of providing a definitive catalogue. Rather, or goal is to identify the major themes and the big issues and provide some useful sources.

The paper proceeds as follows. The next section provides an overview of the measurement issues and challenges that are inherent in innovation research, and foreshadows in a general way why NLP is potentially powerful in addressing these. In Section 3, we provide a brief historical overview of the main NLP approaches, and identify some overarching issues in using and interpreting them. Section 4 then discusses a handful of specific applications, with the goal of illustrating concretely both their power and their limitations. In Section 5, we conclude with our summary of the potential of NLP in innovation research, and the major issues that researchers using these methods currently face.

2 Measurement: challenges and goals

The framework for empirical analysis of innovation studies can largely be traced to two key papers by Zvi Griliches. Griliches (1979) lays out the "big questions" in thinking about how investments in new knowledge play out in the private and public sectors, and how the consequences of these investments can be measured. In this section we survey briefly the measurement challenges that researchers have faced in working within this framework, as a prelude to discussion of the contributions of NLP.

Pakes and Griliches (1980) introduced the idea of a "knowledge production function" ("KPF") that has become, either explicitly or implicitly, the measurement framework at the base of most microeconometric studies of innovation. It posits a functional relationship between inputs to knowledge generation (e.g. previously existing knowledge and firm expenditure on research) and the creation of new knowledge. This stock of knowledge is an important variable in most models of endogenous growth, playing a similar role to physical capital in the neoclassical growth model (e.g., Romer, 1990; Aghion and Howitt, 1992). But how do we measure new knowledge? Not only is new knowledge not easily observable, there is also no obvious definition of the appropriate units that we should be measuring it in. Do we count ideas? And if so, are two new ideas twice as good as one new idea? In general, researchers need to make assumptions as to how specific indicators of success in knowledge creation (e.g. patents) relate to this stock of knowledge.

Implementing the KPF framework requires several choices for the measurement of both inputs and outputs. While specifics vary with context, the inputs to the innovation process broadly speaking are of three types: (1) the capabilities or skills of people engaged in the process; (2) the material inputs that those people can bring to bear; and (3) the pool of existing knowledge that they can draw on in deciding how to proceed. The other side of the KPF framework is identifying proxies for the unobservable knowledge output, such as scientific papers, patents, prizes, awards, and so forth. There is some value in merely identifying and counting these proxies, but research effort and creativity are also devoted to tackling the quality issue, i.e. trying to come up with characteristics of these proxies that can be interpreted as conveying information about the size or quality of the invention or innovation.

Many or all of these measurement targets are inherently hard to pin down, in large part because they reflect or are intimately wrapped up in knowledge. Knowledge is physically unobservable and has no fundamental units. That is why we use indicators or proxies. Reliance on these proxies is not necessarily bad. As an example, the amount of money that a firm spends on research for new products is reasonably well-defined and economically meaningful. But because it is a step away from the fundamental inputs to the innovation process, its use in empirical research always raises issues of interpretation that may cloud our understanding of what we are observing. Similarly, indicators of innovative output such as patents or counts of new products are connected to what we really want to measure, but the nature and limitations of those connections cannot be fully known. Even though we do not observe most of these features directly, we can still in some cases observe characteristics of products (energy consumption, efficiency, size) or services. But these are hard to observe in a systematic way, and data are not always available. At the same time, by using multiple indicators and modeling the underlying relationships, we can still learn a lot.

A related fundamental difficulty flows from the non-rival nature of knowledge. Spillovers of knowledge across individuals, organizations and geographies operate on both the input and output sides of the innovation process. The existence of spillovers means that the inputs and the outputs are, to varying degrees, embedded in complex ways throughout the economic system. Tracing these linkages and measuring their magnitude poses an additional empirical challenge that has attracted much attention.

The last two big areas of empirical innovation research is the tracking of the organization and financing of innovation on the one hand and of the impacts of innovation. This has a first-order component, in understanding the causes and effects on the performance of innovating firms and industries. This has included looking at the introduction of new products by

innovators, follow-up innovation, effects on competition and market structure, and effects on productivity.

As noted above, however, part of the reason we care about innovation is because we believe it has large social returns and is a major contributor to economic growth. These large returns do not derive primarily from the first-order impacts; they arise as innovations diffuse throughout the economy, and thereby have impacts that are at least partially unintended and unanticipated. We can measure some of this effect at the macroeconomic level (e.g. Jones and Summers, 2020), but to really understand what is going on requires methods to identify and quantify the linkages between innovations and innovators on the one hand, and socially desirable outcomes on the other. Because of the intangible nature of the knowledge flows and the complexity of the overall system, this is hard to do. All kinds of new public policies, new kinds of jobs, new health care modes, new forms of entertainment, etc., are facilitated in part by innovations throughout the entire economysystem. Case studies can illuminate this process, but systematic measurement is very challenging.

At a general level, NLP has potential to help with all of these measurement challenges, in three major ways. First in improving on the measurement of existing indicators of knowledge. Second in creating new indicators out of textual artifacts and third in finding and quantifying new linkages between innovations and the broader economy. We discuss each of these possibilities in turn.

First, NLP enables the extraction of more useful, subtle, and meaningful information from artifacts. We started by counting patents awarded to different agents over different periods. We then got more useful insight by counting patent claims and tracing patent citations. "Processing" the full text of a patent in a general and deep way can tell us even more about the underlying invention. Neural networks can be trained to identify particular invention attributes based on the text in the associated patents. The importance and impact of an invention are associated in complex ways with how the invention relates to those that came before and those that came after. As discussed below, NLP methods have been used extensively to make such comparisons and use them to quantify importance and impact.

Second, NLP offers the prospect of identification of new proxies in the form of text artifacts that have not so far been thought of or collected for innovation research. Firms make many statements about their innovative activities in disparate forms such as financial reports, trade show presentations, corporate websites, and pitches by entrepreneurs to venture capitalists. Such unstructured text has not before been used to try to quantify aspects of the innovation process, but with NLP they could be. In particular, the worldwide web offers an

immense repository of texts whose inconsistent and unstructured nature has so far limited their usefulness as a source of innovation data, but perhaps NLP can extract some minerals from this huge reservoir of low-grade ore. By examining how innovation is *described* in these texts, researchers can capture hidden signals—like the perceived novelty of a technology, potential applications of this technology, or the strategic importance a firm assigns to specific research areas—even when those signals do not translate directly into formal metrics such as R&D expenditures or patent grants.

Finally, NLP offers the prospect of new ways of identifying and quantifying linkages in the innovation system. As we discussed above, spillovers of knowledge are an important aspect of the innovation process. Measuring spillovers in a causally meaningful way is a major empirical challenge, but a necessary predicate for any such effort is some way to empirically identify linkages that *might* allow spillovers to flow between agents or institutions. Thus, an often necessary, but not sufficient, condition for innovation i to have caused innovation j to occur is that the two innovations are somehow 'similar', and NLP can be used to quantify the degree of similarity. In addition, to quantify the impact of an innovation to the broader economy, we need a way to identify the pathways along which this impact might occur. The strength of NLP in finding and measuring linkages is that text is everywhere. We can look for spillover pathways by analyzing and comparing the text of one firm's patents or papers together with those of another firm. And we can look for impact pathways by analyzing and comparing the text of patents or papers together with the text of job descriptions or policy pronouncements.

It is useful in thinking about measurement to distinguish what might be called "practical" and "conceptual" problems. An example of a practical problem is the fact that firms track and report their spending on R&D in different ways, and standard accounting practice treats R&D as a current expense, while our models conceive it as a form of investment. While data constraints prevent us from solving these problems completely, we at least in principle know how we would solve them if we had access to the right information.

The use and creation of knowledge also presents conceptual problems. That is, it is difficult to specify the "right" way to measure knowledge, even if there were no constraints on the data that could be collected. New knowledge is inherently intangible, and we cannot say, even conceptually, how "big" one new insight is or whether it is "bigger" than another one. This problem is related to the broader empirical problem of measuring the 'quality' of inputs and outputs. To understand the economics of the construction industry, we may need to look at how the 'quality' of concrete has changed over time. But the specific physical

characteristics that make concrete 'high quality' in the context of road construction may differ from the characteristics that make it 'high quality' in a skyscraper, so there isn't a well-defined broadly applicable way to measure quality.

The importance and the difficulty of these issues in the context of knowledge-related inputs and outputs are both much greater. As an example, Robert Gordon has argued that the inferior economic performance of advanced economies in the last few decades compared to periods in the twentieth century can be attributed to recent innovations being just less significant than those that came before (Gordon, 2000, 2014). To test this hypothesis, we would need a way to measure the "size" or "importance" of innovations, and then test across different contexts the extent to which economic performance responds to innovation size. Unfortunately, it is really hard to measure the "size" of innovations in any way other than looking at their economic impact. But if we measure size in terms of economic impact we cannot meaningfully then test whether declining performance is due to declining innovation size. The only way around this issue is to assume/postulate that some measurable attribute of an invention (e.g. its investment cost or citation pattern) is a measure of its "size", so that the relationship between size and economic growth (or other impacts of interest) can be tested.

The distinction between practical and conceptual problems is useful both in understanding how NLP can improve innovation research, and in thinking about NLP's own limitations. At a practical level, NLP can be useful to fine-tune existing measures, for example, by distinguishing citations that truly indicate a knowledge relationship from those derived from courtesy or convention. But it may also help with the conceptual problems of measuring knowledge and the size of an innovation (or allows to implement existing solutions that were not practically feasible before NLP). Knowledge and language are closely connected, so it seems plausible that deep quantitative analysis of text can give us insights into how to measure the attributes of a new chunk of knowledge. But this potential can only be realized if researchers operate from the perspective that this is a fundamental conceptual problem. It cannot be solved by cranking out metrics just because we can. As we proceed to discuss these methods in more detail and in specific applications, we will see that thinking carefully about exactly what the metrics generated by NLP mean is a recurring theme.

In the next section, we briefly review the evolution of various NLP approaches, from early basic methods to advanced semantic analysis, to illustrate how textual data can enrich traditional innovation metrics and address some of the conceptual and practical challenges we have raised. The following section then builds on this historical review to describe a number

of recent contributions, and how these contributions illustrate the promise and peril of NLP methods.

3 A Brief Overview of Different Approaches

Recent advances in NLP have significantly broadened the definition of what researchers consider to be "data." Whereas traditional economic analysis focused predominantly on numeric indicators, NLP makes it possible to incorporate a much richer set of textual sources. Examples include patents, scientific articles, financial and governmental reports, earnings calls, job postings, and even websites. These text-based artifacts capture a wide range of details—from the specific technologies a firm is developing, to the strategic considerations that drive research investments. Working with text presents its own difficulties, however, as the meaning is often layered, context-dependent, and shaped by the author's intent or domain-specific jargon.

3.1 Proto-NLP (Keyword Search)

Long before sophisticated NLP algorithms emerged, researchers were already harnessing textual data. Such "proto-NLP" approaches typically involved considerable input on the part of the researcher as they required identifying or tallying predefined features, most commonly through keyword searches.

Early examples of this approach include Tetlock (2007) and Loughran and Mcdonald (2011), who apply keyword-based analyses to gauge the sentiment of corporate financial reports. In the context of innovation studies, Dechezleprêtre, Hemous, Olsen and Zanella (2019) use search terms (e.g., "robot") to pinpoint patents associated with automation, while Bena and Simintzi (2023) and Ganglmair, Robinson and Seeligson (2022) classify patent documents into process and product innovations by looking at specific semantic patterns. Similarly, Webb, Short, Bloom and Lerner (2018) identify emergent trends—such as drones, machine learning, and cloud computing—by tracking relevant terms in the patent corpus.

The main advantage of keyword-centric strategies is their transparency and reproducibility. Researchers have full control over the words that define a given topic, and can thereby study robustness and evaluate the importance of each term. This makes these methods straightforward to implement and to interpret, and ensures that the result will avoid misclassification (type I error). However, they have significant limitations when compared to the more advanced NLP techniques discussed later. The principal shortcoming is the subjective nature of

choosing keywords. For example, Bergeaud and Verluise (2023) shows that a strict reliance on predetermined keywords can overlook a significant share of pertinent documents—particularly when terminology shifts or when an innovation's boundaries are not easily captured by a single word or phrase.

Even though, in principle, the size of the dictionary need not be limited, is not clear that even a very large and carefully crafted dictionary of expert-chosen keywords can be enough to truly delineate a technology, a research strategy or a type of product. For instance, Dechezleprêtre et al. (2019) classify only about 1% of patents as automation-related, which is arguably a lower bound due to their narrow definition. Hence, these classification strategies face a trade off of reducing false positives at the cost of missing many relevant documents. These strategies can be sufficient for topics defined by very specific or unambiguous terminology, and they offer researchers a rapid initial screening tool to flag potential documents of interest. However, the drawbacks can be particularly relevant for rapidly evolving fields such as AI or green technologies, where the lexicon is still in flux.

3.2 Vectorization Approaches to Computing Similarity

A fundamental challenge in NLP is the translation of raw text into a numerical representation that can be manipulated and analyzed. Once a document is expressed as a vector, researchers can leverage standard algebraic operations—such as calculating the cosine similarity between two vectors—to gauge how alike the underlying texts are. Once each pair of textual documents is linked with a measure of distance, clustering or other methods can be applied to classify the full corpus in pursuit of various research questions. The different methods may vary in how exactly the text document is transformed into a numerical vector.

In the early stages of NLP, methods such as the Bag-of-Words (BoW) model laid the groundwork for textual data representation. BoW simplifies text by converting documents into vectors based on word frequency, disregarding grammar and word order. Economists could then compute the similarity between two document vectors: two document vectors would be similar if they contained the same words with similar relative frequency, and completely different if they share no common words. Although BoW is relatively easy to implement and interpret, its lack of contextual understanding and its inability to consider synonyms as similar words both limit its utility for capturing the nuances of language.

An extension of the BoW approach introduces the concept of n-grams, which help address some of these limitations. An n-gram is a contiguous sequence of n words from a document that is treated as a single object when vectorizing the document. The use of n-grams

enhances the BoW model by incorporating information about word order and co-occurrence patterns. For instance, the bigram "economic growth" carries a more specific meaning than the individual words "economic" and "growth" considered separately. N-gram models have been useful in various applications, but they are often computationally unwieldy: as n increases, the number of possible n-grams grows exponentially. Moreover, n-gram models still lack a deep understanding of language semantics and long-range dependencies beyond the chosen value of n.

A further refinement is the Term Frequency–Inverse Document Frequency (TF-IDF) scheme, which reweights the importance of words according to how rare they are across an entire corpus. Compared with BoW, TF-IDF gives greater prominence to less frequent but more informative terms and thereby reduces the emphasis on high-frequency, generic words. Two TF-IDF document vectors would be similar if they shared a lot of *distinctive* words. Though an improvement upon BoW, this technique still struggles with synonyms and context, as it retains the underlying assumption that each term has a fixed meaning regardless of its surrounding words.

The advent of word embeddings marks a significant shift in NLP methodologies. Methodologies such as `Word2Vec` (Mikolov, Chen, Corrado and Dean, 2013) and `GloVe` (Pennington, Socher and Manning, 2014) transform words into dense, continuous vector spaces where semantically similar words are represented as vectors that have higher (cosine) similarity. These embeddings are typically learned by training a shallow neural network on large corpora, using objectives such as predicting the surrounding words given a target word or predicting a target word given its context. By capturing the statistical co-occurrence patterns of words in their context, these models encode semantic relationships into the geometry of the vector space, with meaningful relationships such as analogies often emerging (e.g., the vector difference between "king" and "queen" is similar to that between "man" and "woman"). These dense embeddings also enable dimensionality reduction—most models compress each word into a vector of a few hundred dimensions, far fewer than the full vocabulary. Depending on the embedding model used, this compact representation enables researchers to efficiently process and analyze text data. By choosing an appropriate weighting scheme (for example, TF-IDF as in Seegmiller, Papanikolaou and Schmidt (2023)), researchers can also represent a document as a weighted average of the underlying word vectors, resulting in a fixed-length vector for each document. This allows for consistent and computationally manageable representations of text. By contrast, older methods like the BoW approach and its extensions represent documents as sparse vectors of a length equal to the size of the vocabulary or the union of all

words (or n-grams) included in the documents, which can grow unwieldy for large corpora.

The next discrete step in NLP goes beyond static word embeddings to vectorize entire documents while preserving their intended meaning. Averaging word embeddings provide a basic way to represent documents but this still lacks the capacity to fully capture the interplay among words. Transformer architectures have addressed this limitation. Introduced by Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin (2017), transformers analyze the relationships between all words in a sequence simultaneously, rather than sequentially. Using a mechanism called "attention", they weigh the importance of each word relative to others in the text, effectively capturing context and meaning. This capability allows transformers to generate contextualized embeddings, where the meaning of each word adapts based on its context within the sentence or document. Pre-trained models based on transformers, such as *BERT* (Bidirectional Encoder Representations from Transformers) and the *GPT* (Generative Pre-trained Transformer) series, have revolutionized NLP. These models enable researchers to fine-tune contextual embeddings for specific applications, offering a significant leap forward in the ability to vectorize and analyze text at both the word and document levels.

Importantly, many of these approaches can be further fine-tuned via supervised learning to improve performance based on the type of corpus. For example, Bekamiri, Hain and Jurowetzki (2024) and Ghosh, Erhardt, Rose, Buunk and Harhoff (2024) respectively developed PatentSBERTa and PaECTER, two embedding models that build on BERT but are specifically trained on patent documents to capture semantic relationships across patents. Compared to off-the-shelf embeddings trained on general English text, these specialized models could in principle improve the precision of any NLP task that is applied on comparable corpora. However, these models have been trained with some specific objective in mind and may not be appropriate for tasks that do not relate to this objective. For example, Ghosh et al. (2024)'s PaECTER has been trained with the objective of predicting citations across patents. The resulting embedding and underlying measure of similarity (e.g., via cosine distance) will thus reflect this objective rather than any other possible interpretive goal. A prominent discussion on this point is provided in Ganguli, Lin, Meursault and Reynolds (2024). This is an illustration of a recurring deep issue: because knowledge and innovation are conceptually so difficult to quantify, we often train NLP models and interpret their results based on existing proxies, whose own meaning and interpretation is largely just axiomatic. This is not "wrong", but it must be continuously kept in mind as the methods evolve.

Overall, the field of NLP has witnessed a remarkable transformation over the last decade,

progressing from simple models like Bag of Words and TF-IDF that operate in a fairly transparent way, to advanced transformer-based architectures that can involve multiple hidden layers of computation. Ultimately, these methods aim to quantify the degree to which two documents are 'similar', which can range from these documents use similar words, to these documents contain text that has the same meaning. Ultimately, however, researchers need to take a stance as to what does it mean, economically, that two documents are similar.

3.3 Application to classification tasks

Building on text-based similarity measures, one can distinguish two principal strategies for classifying documents: *supervised* and *unsupervised* methods. Although both aim to map documents into meaningful categories, they differ in how those categories are defined and validated.

In *supervised* classification, researchers start by specifying a classification scheme in the form of explicit rules that a document must satisfy to be assigned to a given group. This approach is especially useful for well-defined categories—for example, patents in a particular technological domain (e.g., solar panels) or news articles referencing a specific event. In many cases, these rules derive from so-called proto-NLP methods, often involving the presence or absence of one or more keywords (see Section 3.1 for several examples). Although rules can be combined and tailored, they are ultimately dictated by the researcher, which limits the practicality of high-dimensional text representations in this framework.

Recent advances in machine learning, however, have expanded the scope of *supervised* classification. Rather than relying on a fixed set of rules, researchers can begin with a carefully curated set of example documents and then identify others that resemble them. This similarity search can capitalize on more advanced NLP techniques, such as embedding-based models, sometimes based on specifically trained models. Examples of these "automated patent landscaping" (Abood and Feltenberger, 2018) demonstrate how embeddings enable a more scalable and fine-grained mechanism for grouping documents into clusters that align with pre-defined objectives. While this approach lowers the risk of missing relevant documents that do not contain specific keywords, it demands careful calibration to balance type I and type II errors. Moreover, selecting an appropriate embedding model is crucial (Ganguli et al., 2024).

By contrast, *unsupervised* methods require no pre-labeled examples. Instead, they infer categories from the data itself, using NLP-derived features ranging from simple keywords (Yoon and Park, 2004; Hu, Li, Yao, Yu, Yang and Hu, 2018; Bergeaud, Potiron and Raimbault, 2017)

to multi-dimensional embeddings. A prominent example is the Latent Dirichlet Allocation (LDA) topic model introduced by Blei, Ng and Jordan (2003), often likened to principal component analysis (PCA) for text. Here, documents are first transformed via a vectorization technique—such as TF-IDF—and LDA identifies patterns of co-occurring terms, grouping them into so-called "topics." Alternative approaches to generating unsurpevised clusters include K means, Gaussian Mixture Modeling , Hierarchical Clustering, and DBSCAN.

As with standard principal components analysis, these topics emerge without inherent labels; researchers must interpret and name them based on contextual knowledge. Hansen, McMahon and Prat (2017) illustrate this by applying LDA to central bank communication transcripts, revealing distinct thematic threads in policy discussions. Thus, although unsupervised methods can initially seem appealing, the absence of explicit labels poses its own set of challenges. Clusters may overlap or be ambiguous, requiring careful judgment to name and interpret them. Moreover, tracking how a specific cluster changes over time becomes more difficult if that cluster was not clearly defined from the start.

3.4 Challenges in Implementing NLP

The multiplicity of available NLP methods raises a number of issues that are both practical and conceptual. This remains true, and perhaps all the more relevant, even as increasingly powerful and specialized models enrich our capacity to analyze textual artifacts. In this section, we briefly discuss some of the most salient challenges and offer an illustration using a simple example.

A first concern relates to the *transparency* and interpretability of these approaches. While supervised rule-based methods relying on keywords (potentially weighted by importance) offer a clear pipeline from the definition of rules to the final output, more recent systems—especially large language models—remain largely opaque. This "black box" nature can hinder reproducibility, as subtle differences in how text is embedded or vectorized may alter downstream results. Researchers must therefore confront important questions about how dependent their conclusions are on the specific model selected, and about how to evaluate the robustness of their results. Conceptually, a measure of similarity would require some form of *formal statistical inference*. One can think of a true underlying relationship variable, x, that might take values $\{0, 1\}$. The observed similarity score ρ is drawn from some distribution $f(\rho|x)$. If researchers had access to labeled examples—cases where $x = 1$ or $x = 0$ was independently verified—they could estimate these conditional distributions and compute $P(x = 1|\rho)$. However, in many real-world settings, such labeled data are limited or incomplete, making

it challenging to quantify measurement error or assess the reliability of similarity-based conclusions. But constructing such labeled gold standard is far from straightforward. It is clear that two identical text documents should be assigned a similarity of 1, but for similar, yet not identical, artifacts, there is no obvious way—even in principle—to determine the gold-standard value for x (if such a standard even exists).

A second challenge stems from the *token limit* inherent in many modern transformer-based large language models (LLMs), which can lead to documents being truncated or split into segments before processing. If only part of a document is embedded, improved embedding quality may not necessarily compensate for the loss of context. In some cases, it may be preferable to use a simpler model that can handle the entire text at once than using a more advanced one restricted to partial input. A related issue concerns the *single similarity score* commonly reported when embedding models compare two documents. A 70% similarity does not necessarily indicate that 70% of the text overlaps in a meaningful way—one might be dealing with 30% of the text covering 100% of another text's content, or vice versa. Approaches that "chunk" documents into smaller segments and compute similarities between corresponding segments can alleviate some of these distortions, but they add complexity to the analysis and require a clear understanding of the conceptual role of these different chunks. In economic applications, this distinction between partial and near-complete overlap may lie at the heart of whether two technologies are complements or substitutes, yet a single similarity value may obscure such nuances. Note that this challenge is essentially a technical one, likely to be solved at least in the medium run.

Another practical consideration is the *distribution* of similarity scores that different methods produce. Traditional approaches like TF-IDF and Bag-of-Words yield sparse and highly skewed similarity distributions, whereas LLM-based embeddings often generate smoother distributions. If a dataset contains few genuine relationships, researchers may be inclined to set a cutoff (for instance, discarding all scores below a certain threshold), but the choice of this threshold is often arbitrary. Inconsistent distributions across methods complicate direct comparisons and may influence the interpretation of results.

To illustrate this, we randomly draw a set of 1000 USPTO patents filed in 2018 and compute their pairwise similarity based on their abstracts using twelve different methods. First, we use a simple Bag-of-Words representation with 1-gram, 2-gram, and 3-gram features: each (single) word, pair, or triplet of consecutive words—excluding stopwords—is represented in a high-dimensional occurrence vector. We then repeat the same process but apply TF–IDF weighting to each term, giving three additional variants. Next, we employ GloVe (Pennington

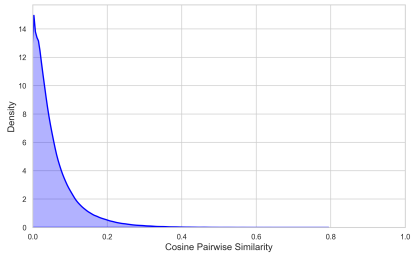
et al., 2014), a word-level embedding model trained on a large English corpus, using the "6B, 300-dimensional" release and averaging each document's word vectors using TF-IDF weights as in Seegmiller et al. (2023). We also include three larger, general-purpose language models: OpenAI's ADA-002 (a relatively compact GPT-based embedding model), OpenAI's GPT-3 large (a larger GPT-based model), and GTE (Li, Zhang, Zhang, Long, Xie and Zhang, 2023) (a large language model developed by Alibaba). Finally, we use the two models specialized for patent data we already mentioned: PatentSBERTa and PAECTER.

In each case, we plot the distribution of the 499,999 measures of similarity constructed from each possible pairs in Figure 1. Visually it appears clearly that these distributions look very different and in particular the approaches that are based on Bag of Words (whether weighted or not) yields distribution that are skewed, with a few outliers cases but mostly very low values of similarity. The sparsity of the similarity matrix is pronounced for models that weigh based on TF-IDF. This is not surprising, since these BoW approaches can miss related documents, even as the documents that they identify similar are very likely related. By contrast, large language models tend to generate distributions that are more continuous with large average value of similarities in some cases. Somewhat reassuringly these different models generate scores that are positively correlated with each other as shown in Table 1, but the correlation can be low in some cases which reinforce the idea that a measure of similarity needs to be defined carefully using a model that is appropriate to the exercise.

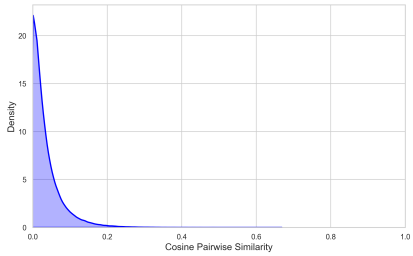
But is there a model that dominates the others? Again, this depends on what the researcher expects "similarity" to capture. If the researcher expects the patent similarity matrix to be relatively sparse, similar to the matrix of citation linkages, then approaches that generate sparsity are desirable. Absent that, researchers often decide on arbitrary thresholds for what constitutes a related vs unrelated document, which is another way of generating artificial sparsity. More generally, however, absent a notion of 'ground truth', it is not feasible to decide which of the different methods of computing document similarity is superior to others. In addition, which method of converting text into numerical vectors performs better will likely greatly depend on the specific application in mind. A related issue is that these similarity scores do not include standard errors; developing methods to perform statistical inference on similarity metrics is likely a fruitful area for future research.

Figure 1: Pairwise Similarity Distributions for Various Models

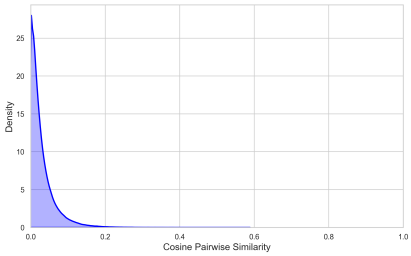
(a) BoW noTF-IDF kgram1



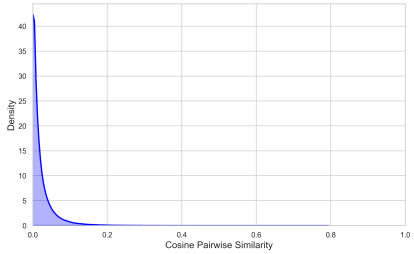
(b) BoW noTF-IDF kgram2



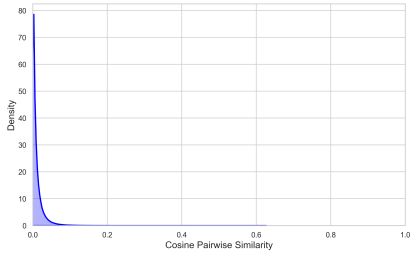
(c) BoW noTF-IDF kgram3



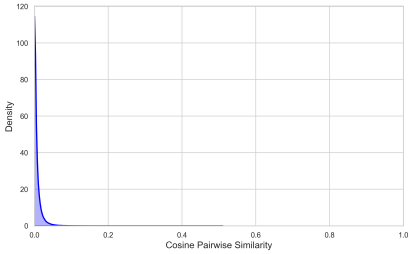
(d) BoW TF-IDF kgram1



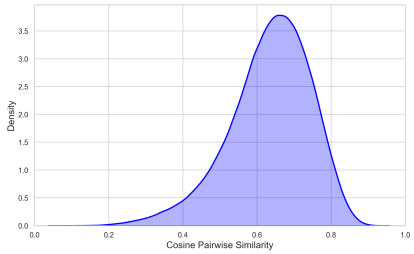
(e) BoW TF-IDF kgram2



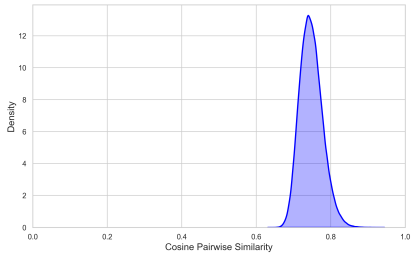
(f) BoW TF-IDF kgram3



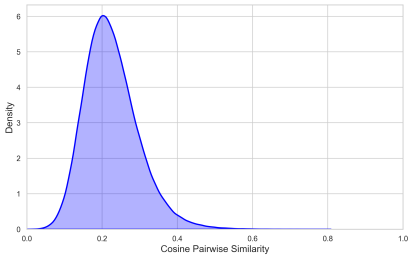
(g) GloVe (6B, 300)



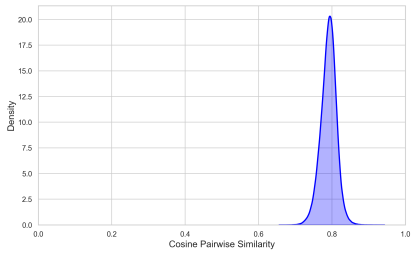
(h) OpenAI (small)



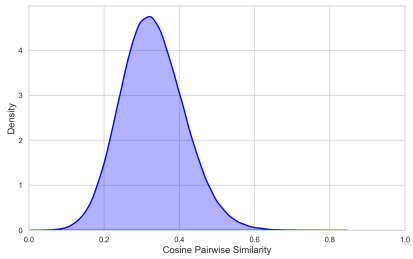
(i) OpenAI (large)



(j) GTE-large



(k) PatentSBERTa



(l) PaECTER

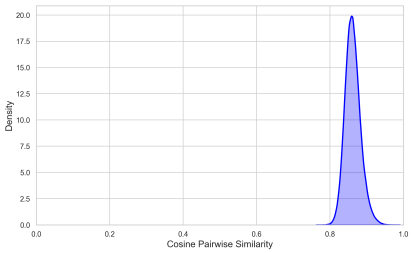


Table 1: Rank correlation between pairwise similarities

		Bag of Words			TF-IDF			GloVe	OpenAI		GTE	PSB	PAEC
		1	2	3	1	2	3		Small	Large			
Bag of Words	1	1.00											
	2	1.00	1.00										
	3	1.00	1.00	1.00									
TF-IDF	1	0.93	0.93	0.93	1.00								
	2	0.93	0.93	0.93	1.00	1.00							
	3	0.93	0.93	0.93	1.00	1.00	1.00						
	GloVe	0.49	0.49	0.49	0.54	0.53	0.53	1.00					
OpenAI	Small	0.43	0.44	0.44	0.48	0.48	0.48	0.56	1.00				
	Large	0.43	0.43	0.43	0.47	0.47	0.47	0.51	0.82	1.00			
	GTE	0.25	0.25	0.25	0.27	0.27	0.27	0.26	0.42	0.45	1.00		
	PSB	0.41	0.42	0.42	0.45	0.45	0.45	0.53	0.72	0.68	0.36	1.00	
	PAEC	0.27	0.27	0.27	0.31	0.31	0.31	0.38	0.53	0.55	0.32	0.43	1.00

Notes: PSB stands for PatentSBERTa and PAEC for PaECTER. Rank correlation based on 499,999 similarities between each pairs of 1000 randomly selected patent abstract.

4 Applications of NLP in Innovation Research

At its core, NLP is particularly valuable for identifying for extracting information, generating data, and quantifying linkages between textual artifacts. These tasks, together with specific assumptions about the economic implications of what they mean, can be used to identify key objects of interest in the data. Here, we provide a few examples of existing applications of NLP based measure of innovation and how they contribute to important research questions.

4.1 Enhancing the Quality of Underlying Data

One of the most direct ways NLP can aid innovation research is by improving the quality of underlying data. Patent documents, for example, typically include structured metadata, such as inventor names, technological classifications, and issue dates, which already offer valuable insights into an invention's nature. Yet, thesemetadata often constitute only a fraction of the information present in the patent itself: as Aharonson and Schilling (2016) note, unstructured

text in patent applications can be far richer, encompassing technical specifications, references to scientific literature, and contextual details about the development process. By applying NLP to parse and extract these textual clues, researchers can more accurately describe and classify inventions, ultimately laying a stronger foundation for empirical analysis.

A clear illustration of this potential comes from efforts to identify citations to academic research within the full text of patent publications. Marx and Fuegi (2020), Bryan, Ozcan and Sampat (2020), and Verluise, Cristelli, Higham and de Rassenfosse (2020) develop NLP-based approaches that locate and record references to scholarly articles embedded in patent texts. These citations, often hidden in footnotes or sections reserved for prior art, provide a crucial link between scientific discoveries and subsequent technological innovations. To identify them, researchers train named entity recognition (NER) models on curated datasets, enabling the models to recognize citation patterns in an otherwise heterogeneous corpus of patents.

Beyond the extraction of academic citations, NLP can also be harnessed to clean and enrich other patent metadata. Historical patent documents, for instance, might lack complete or standardized information about inventor affiliations, assignee details, or geographical origins. Early work in this domain frequently relied on formalized document structures, employing strategies like direct place-name matching (Petralia, Balland and Rigby, 2016) and regular expressions (Berkes, 2018) to retrieve and standardize these data. While such proto-NLP methods have performed reasonably well—especially for historical U.S. patents (Andrews, 2021)—they can fall short when faced with linguistic variability across different time periods, jurisdictions, or document templates. Recent advances in NLP, including sophisticated NER models, now offer more robust solutions for extracting entities from patent texts (Bergeaud and Verluise, 2024). These models can identify and classify text spans associated with inventors, assignees, or locations, even in the face of spelling variations or unstructured formatting.

These data enrichment tasks deliver outputs that can be used "off the shelf" and readily merged with other datasets. Applications include matching patent assignees to firm-level registers such as Compustat (Kogan, Papanikolaou, Seru and Stoffman, 2017) or matching scientists or inventors to Census records (Akcigit and Goldschlag, 2023). Unlike many other NLP tasks, this process has a straightforward success criterion: once an entity is properly extracted, cleaned, and accurately matched, there is no further need for improvement. The key advantage of NLP here is that it automates a labor-intensive task—one that would otherwise overwhelm human capacity, given the vast number of documents to process. This highlights the value of more coordinated efforts among researchers. It is inefficient for multiple

teams to conduct repeated matches of inventors to Census data, assignees to Compustat, or inventors to locations. A better approach might be to develop a shared validation dataset that is manually curated, thus defining a clear benchmark for accuracy. Such collaboration would also allow researchers to compare outputs from different models on a case-by-case basis, focusing attention on ambiguous observations and ultimately converging on more robust, standardized results. The NBER Innovation Information Initiative ("I3") and its iiindex project are recent attempts towards moving to this direction.

4.2 Identifying different types of technologies

Economists are often interested in the impact of specific technologies (e.g., semiconductors or green energy). Identifying these technologies based on information from the patent document is not straightforward. Patent offices use rich and detailed classification systems for each publication (CPC, IPC, USPC, etc.), but these classifications are primarily designed for engineers and are mainly intended to capture common technical features (for example, metal fusion bonding). Economists, on the other hand, typically seek clusters of documents that correspond to broader technological fields, products, or economic activities (e.g., semiconductors). This difference creates a *classification problem*, already highlighted by Griliches (1998); Grupp (1998); Schmookler (1966), and Trajtenberg (1987). The NLP methods described in Section 3.3 can provide a powerful solution to this classification problem.

When using NLP as a classification tool, both supervised and unsupervised approaches are possible. Recent examples of supervised methods include Webb et al. (2018), who identify novel technologies such as drones, machine learning, and cloud computing within the patent corpus; Dechezleprêtre et al. (2019), who use terms like "robot" to pinpoint technological classes likely related to automation; Mann and Püttmann (2023), who look for words overrepresented in a manually curated set of "automation patents" to assess patents concerning labor-saving technologies; Ganglmair et al. (2022), who categorize patents into process and product innovations; and Calvino, Criscuolo, Dernis and Samek (2023) and Webb (2019), who apply similar techniques to classify patents related to AI. Beyond patents, Babina, Fedyk, He and Hodson (2024) and Acemoglu, Autor, Hazell and Restrepo (2022) adapt similar methodologies to analyze job vacancy postings, measuring AI adoption by firms. These examples rely largely on transparent, keyword-based strategies.

On the other hand, a growing number of studies adopt machine learning techniques and text embeddings to identify documents relevant to a given category. For instance, Kriesch and Losacker (2024) use a manually curated seed of bioeconomy patents and embed the text to

retrieve other, semantically related patents, while Dunham, Melot and Murdick (2020) assign a probabilistic score to each scientific article, estimating the likelihood that it is related to AI. Recent approaches have developed more advanced "landscaping" methods that combine rule-based methods such as a hard-coded list of keywords and relevant technological classes with example documents and then harness embedding-based expansions that rely heavily on the structure of the patent corpus. This method introduced by Abood and Feltenberger (2018) and extended by Bergeaud and Verluise (2023) is designed to reduce reliance on the set of keywords chosen while maintaining a high level of precision and achieve a better balance between transparency and scalability.

Unsupervised methods started to rely mostly on NLP to extract keywords and group documents based on the similarity of these keywords without prior knowledge of the target groups (see, for example, Yoon and Park (2004); Hu et al. (2018) and Bergeaud et al. (2017)). They progressively started to incorporate topic-modeling techniques with technology-related corpus. Such a method is implemented by Kalyani, Bloom, Carvalho, Hassan, Lerner and Tahoun (2021), who use earnings conference call transcripts to identify the technologies most frequently cited for contributing to companies' momentum and by Lenz and Winker, 2020 to track the birth and death of inventions by analyzing 170,000 technology news articles published in a German newspaper.

4.3 Measuring the rate of innovation

In most macroeconomic models, long-run growth is determined by the rate of technological progress. Empirically, measuring this rate is challenging because of the invention quality issue discussed above. Since the seminal works of Zvi Griliches and his coauthors (see, e.g., Griliches, 1998), patent publications have frequently been used as proxies for measuring innovation, but it is widely recognized that the counting of patents provides a crude measure of innovation (Hall, Jaffe and Trajtenberg, 2001), since the content and impact of patents can vary significantly between firms, technologies and countries, as well as over time (De Rassenfosse, Dernis, Guellec, Picci and De La Potterie, 2013).

To improve measurement of the rate of innovation, it is crucial to identify patents that contribute significantly to technological advances, given that the impact of patents is highly skewed (Kogan et al., 2017) and only a relatively small number are critical for growth. Traditional approaches have focused on metrics intrinsic to the patent system, such as the number of forward citations received (Hall, Jaffe and Trajtenberg, 2000) or the number of claims (Lanjouw and Schankerman, 2001). However, even within consistent time windows

and technologies, these measures cannot fully capture the novel impact of a patent Higham, De Rassenfosse and Jaffe, 2021; Kuhn, Younge and Marco, 2020. For instance, Fadeev (2023) explores the possibility that citation counts are influenced by input-output relationships within supply chains, suggesting that citations may reflect economic linkages between firms rather than the actual novelty or significance of an innovation.

To address these limitations, researchers have developed various NLP techniques to assess the true novelty of patents, and then used these measures of novelty to weight patent counts. Typically, these measures are based on a metric that quantifies similarity between patents and identifies outliers. A natural question that relies on this measure is the extent to which the creativity of ideas has changed over time and whether good ideas are indeed becoming scarcer (Bloom, Jones, Van Reenen and Webb, 2020). If this were the case, we would expect some measure of aggregate novelty to decline over time.

Document linkages, combined with information on the timing of the document can also be used to construct measures of novelty for particular innovations. In particular, an innovation that is dis-similar to prior art can be characterized as novel. Gerken and Moehrle (2012) construct a measure of novelty based on textual dis-similarity for a sample of 300 automotive patents. In pharmaceuticals, Krieger, Li and Papanikolaou (2022) measure the novelty of new drugs based on the chemical distance between the molecular structures of the new drugs and the existing compounds. In other work, Arts, Hou and Gomez (2021) analyze the use of novel unigrams, bigrams, and trigrams in patent descriptions to evaluate technical novelty. They identify patents that diverge from prior work and later shape technological developments. Kalyani (2024) applies this strategy to identify novel, or creative patents, de Rassenfosse, Pellegrino and Raiteri (2024) uses it to track similar follow-on applications and De Rassenfosse and Raiteri (2022) to retrieve competing technologies.

More generally, the same idea can be used to quantify a patent's impact. For example, if patent i preceded patent j, and patent i and j are textually similar, we can perhaps infer that patent j builds upon i. Naturally, if that were the case, patent j should also cite patent i, but the practice of citations is specific to patent documents and research papers, and even then, citations are not always consistently recorded. Kelly, Papanikolaou, Seru and Taddy (2021) show that these document links are indeed highly predictive of citation linkages. They leverage this idea to identify 'novel' and 'impactful' patents from 1850 to the present. A methological innovation in their approach is the use of a backward-looking variant of the TF-IDF Bag-of-Word embeddings, where the TF-IDF weights are dynamically updated as new innovations emerge. This approach may be desirable to identify important terms when

the documents span long periods of time: the word 'electricity' is probably more informative about a patent's novelty in 1890 than in 1990.

Moving beyond using the pairwise similarity among patent documents, Bergeaud, Schmidt and Zago (2022b) use the distance between patents and technical standards to measure the distance to the technological frontier. In other work, Shi and Evans (2023) employ statistical models to identify patents and research articles in life sciences that diverge markedly from the prior art, detecting unexpected combinations of knowledge that indicate potential breakthroughs. Beyond the use of patent publications, Arts, Melluso and Veugelers (2023) measures novelty in scientific publications, and Bellstam, Bhagat and Cookson (2021) use analyst reports to develop a measure of innovation at the firm level that applies to non-patenting firms using topic modeling.

Do these new measures of novelty or impact outperform traditional metrics that rely on patent citations? While it is challenging to definitively answer this question without external validation of a document's true novelty, they have some advantages relative to the use of citation data. One clear advantage of NLP-based measures is their availability when other metrics are not. For instance, patent citations are not systematically recorded in USPTO data prior to the 1950s, and citation practices have varied considerably over time and across countries. Additionally, a large portion of patents are never cited, and very few receive more than five citations, whereas text-based measures of novelty provide a continuous metric, especially with longer texts. Nevertheless, the NLP approaches often generate metrics of novelty and impact that strongly correlate with citations (see e.g, Kelly et al., 2021).

At the same time, however, NLP-based approaches to measuring novelty have their own shortcomings. Specifically, dis-similarity is harder to measure than similarity, especially if one relies on the older approaches (like BoW) that do not really internalize the semantic meaning of the document. Specifically, two documents that share similar words are probably similar; however, two documents that don't share words may have a very similar meaning if it is conveyed using different terms. This concern can be salient if one is interested in measuring the novelty of a technology based on how similar its description is to prior technologies, though it can be mitigated if technologies consistently use specific terms in a similar way.

In brief, NLP techniques allow researchers to construct measures that are consistent across both time and space. Like any other measure, of course, there is also the potential for significant measurement error that will tend to bias researchers to classify technologies as novel when they are not. The extent to which advances in NLP that internalize the semantic structure of language minimize this measurement error remains to be seen.

4.4 Identifying knowledge spillovers

The linear model of innovation suggests a sequential process in which ideas, often originating from scientific research, generate knowledge spillovers that lead to the development of new technologies or innovations, typically by private firms. These innovations are then adopted by economic agents, diffusing through the economy and contributing to productivity and economic growth. Although admittedly oversimplified, this model is useful in framing the lifecycle of an idea and has been influential in shaping how modern growth theory endogenizes innovation. It helps answer important questions such as the role of basic research, the economic returns of an idea, the optimal financing of innovation, and the role of institutions in facilitating the diffusion of technologies.

Unfortunately, none of the steps in this model are easy to measure in practice. For example, measuring the influence that scientific research has on the generation of innovation has proven particularly challenging due to its nature: knowledge flow does not leave an obvious paper trail and can take many different forms. Similarly, while patent data are helpful in identifying which firms are developing new technologies, the actual adoption of the corresponding products or processes is not generally observable at the firm level.

Researchers have historically relied on the assumption that the influence of science on innovation should be stronger locally and, in the absence of direct measures, have used geographical distance as a proxy for knowledge flows (Jaffe, 1989). Other approaches have exploited funding shocks to universities and leveraged heterogeneity in these shocks to identify real effects (Kantor and Whalley, 2014; Hausman, 2022). However, a new stream of research has used NLP to create more direct connections between academic publications and (private) patents. For example, Poege, Harhoff, Gaessler and Baruffaldi (2019); Marx and Fuegi (2020); Bryan et al. (2020); Verluise et al. (2020) explore how a specific piece of scientific knowledge has been used in the development of a specific technology. Such knowledge networks are informative about the direction and nature of knowledge spillovers and have been used to show that patents drawing more heavily on science are of higher quality (Krieger, Schnitzer and Watzinger, 2024) or to measure the returns of scientific research in terms of R&D expenditures and patent outcomes (Azoulay, Graff Zivin, Li and Sampat, 2019; Bergeaud, Guillouzouic, Henry and Malgouyres, 2022a).

In sum, the diffusion of knowledge from science to applied technologies (patents) can be measured through notions of document similarity between patent documents and scientific articles. What is considerably harder, however, is to measure how the technologies diffuse

from the patented invention to the rest of the economy. The reason is that there is no systematic corpus of documents, analogous to papers and patents, that corresponds to the use of inventions in specific economic contexts.

Lacking a single systematic corpus, researchers have employed different sources of product descriptions to map patents into products. For example, Argente, Baslandze, Hanley and Moreira (2020) uses the Wikipedia description of non durable goods to measure the probability that a given patent would be related to a specific good based on how close text descriptions are, and then track changes in the price of these goods. and Seegmiller et al. (2023) match patents to industries using industry descriptions. Masclans-Armengol, Hasan and Cohen (2024) train a language model to measure the commercial potential of science from the wording of academic articles and a training set made of patents. de Rassenfosse and Zhou (2020) describe the use of 'Virtual Patent Marking' to find website connections between specific products and specific patents. These are just examples of the enormous amount of text on the worldwide web describing products and technologies; it seems there is additional potential for NLP methods to extract information on technology diffusion from these.

A related and often easier task is to measure the diffusion of a technology to subsequent technologies. As an example, Kelly et al. (2021) use the similarity between a focal patent and subsequent patents to identify 'impactful' technologies. Frankel, Krieger, Li and Papanikolaou (2023) identify knowledge spillovers in pharmaceuticals using measures of molecular similarity. Naturally, however, this approach entails a significant assumption that the earlier technology somehow 'caused' the emergence of the later, related technology. Though this assumption may be reasonable in certain settings, it is still an assumption that cannot be easily verified.

The same techniques can be used to extract information about firm's innovations based on additional sources of data generated by firms beyond the text of patent documents. For instance, job postings (Babina et al., 2024; Acemoglu et al., 2022) have been utilized for this purpose. Assuming that labor is complementary to technology, the adoption of these technologies often implies the hiring of specific skills that can be identified in the text of job advertisements. More broadly, Kalyani et al. (2021) combine a large set of documents, including job postings, patents, and earnings calls, to measure the evolution of these descriptions in relation to the emergence of new technologies. This allows them to track a given innovation in space as it is progressively adopted by different users. Other examples include Bellstam et al. (2021), who propose a new measure of technology adoption using the text of analyst reports for S&P 500 firms and examine the impact on their performance; Kinne and Lenz (2021) analyze firms' public communication through their websites to look

for evidence of new technologies being used; and Kakhbod, Kogan, Li and Papanikolaou (2024) who combine information from regulatory filings and patents to construct a measure of the vulnerability of a firm's intangible assets to creative destruction.

Last, researchers have also examined the similarity between a firm's innovations and the tasks performed by workers in specific occupations. Naturally, the interpretation of this similarity requires some additional assumptions. For example, Kogan, Papanikolaou, Schmidt and Seegmiller (2021) use textual similarity between patent documents and occupation task descriptions to infer which technologies are likely to complement or substitute for specific occupations. Doing so entails some specific assumptions on the economic interpretation of a similarity of a patent to an occupation task description. Specifically, they assume that technologies that are textually related to an occupations routine tasks are more likely to be labor-substituting, whereas technologies that are textually similar to an occupations non-routine tasks are more likely to be complements. Naturally, this is an assumption that needs validation in the data. Consistent with their assumptions, Kogan et al. (2021) show that improvements in technologies that are related to routine tasks are followed by declines in employment, while improvements in technology that are related to non-routine tasks are followed by employment increases. Hampole, Papanikolaou, Schmidt and Seegmiller (2025) apply a similar idea to identifying the impact of the adoption of AI technologies by firms on labor demand, exploiting the text of online resumes, job postings, and occupation task descriptions.

5 Conclusion

The field of natural language processing has witnessed a profound transformation over the years, evolving from fundamental statistical approaches to intricate architectures that leverage neural networks. These advances have significantly enhanced the ability of researchers in economics to process textual artifacts and extract useful data. These approaches vary in both their sophistication and their transparency. Recent approaches are less transparent than keyword-based approaches but they are potentially better at capturing the meaning of a specific document.

At a high level, most of these NLP techniques consist of estimating 'similarities' across different types of documents. Researchers then use these similarity measurements in many ways, tracking linkages between agents and institutions, quantifying the novelty (lack of similarity to previous artifacts) and measuring the impact (similarity to subsequent artifacts)

of scientific findings and inventions. But different NLP methods produce somewhat different similarities, both in terms of the ranking of individual artifacts and the overall distributions of similarity. And deciding which method is 'better' is hindered by the lack of an underlying concept of what similarity of text is supposed to be capturing. Outside of the end cases of identical or orthogonal objects, it is hard to describe what it means substantively for one block of text to be more similar or less similar to another. Consider two pairs of text artifacts. The first pair each contain a couple of paragraphs that are identical, but are otherwise mostly distinct. The second pair has no common paragraphs or sentences but many of the same words are used in similar contexts. Which of these would we 'want' our algorithm to judge 'more similar', and why? Coming at it from the other direction, we would expect that a technology that is a substitute for another would be described with text that is somewhat similar, but we would expect the same to be true for complementary technologies. Does this mean we should interpret 'high' similarity as evidence that two inventions could be either complements or substitutes, or that there is some kind of nonlinear relationship between similarity and complementarity, or that similarity actually tells little about the question? We don't really know.

As often occurs with new research methods, the power and the frequency of use of these methods has grown faster than our understanding of what they are really telling us. There is no doubt that NLP has great potential to increase what is useful as data in innovation research, to increase the precision and the accuracy of data already in use, and to illuminate relationships that were previously hard to study. But to realize that potential we need the allocation of effort to shift somewhat from applying more and more powerful computational methods to larger and larger datasets, towards thinking about and modeling the underlying relationships between the texts and the economic concepts that we care about.

References

Abood, Aaron and Dave Feltenberger, "Automated patent landscaping," *Artificial Intelligence and Law*, 2018, *26* (2), 103–125.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo, "Artificial intelligence and jobs: Evidence from online vacancies," *Journal of Labor Economics*, 2022, *40* (S1), S293–S340.

Aghion, Philippe and Peter Howitt, "A Model of Growth through Creative Destruction," *Econometrica*, March 1992, *60* (2), 323–51.

Aharonson, Barak S and Melissa A Schilling, "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution," *Research Policy*, 2016, *45* (1), 81–96.

Akcigit, Ufuk and Nathan Goldschlag, "Measuring the characteristics and employment dynamics of US inventors," Working paper w31086, National Bureau of Economic Research 2023.

Andrews, Michael J, "Historical patent data: A practitioner's guide," *Journal of Economics & Management Strategy*, 2021, *30* (2), 368–397.

Argente, David, Salomé Baslandze, Douglas Hanley, and Sara Moreira, "Patents to products: Product innovation and firm dynamics," 2020.

Arts, Sam, Jianan Hou, and Juan Carlos Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, 2021, *50* (2), 104144.

_ , Nicola Melluso, and Reinhilde Veugelers, "Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text," *arXiv e-prints*, 2023, pp. arXiv–2309.

Azoulay, Pierre, Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat, "Public R&D investments and private-sector patenting: evidence from NIH funding rules," *The Review of economic studies*, 2019, *86* (1), 117–152.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

Bekamiri, Hamid, Daniel S Hain, and Roman Jurowetzki, "Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert," *Technological Forecasting and Social Change*, 2024, *206*, 123536.

Bellstam, Gustaf, Sanjai Bhagat, and J Anthony Cookson, "A text-based analysis of corporate innovation," *Management Science*, 2021, *67* (7), 4004–4031.

Bena, Jan and Elena Simintzi, "Machines could not compete with Chinese labor: Evidence from U.S. firms' innovation," Working Paper, Centre for Economic Policy Research 2023.

Bergeaud, Antonin and Cyril Verluise, "Identifying technology clusters based on automated patent landscaping," *Plos one*, 2023, *18* (12), e0295587.

_ and _ , "A new dataset to study a century of innovation in Europe and in the US," *Research Policy*, 2024, *53* (1), 104903.

_ , Arthur Guillouzouic, Emeric Henry, and Clement Malgouyres, "From public labs to private firms: magnitude and channels of R&D spillovers," Discussion Paper dp1882, London School of Economics and Political Science. Centre for Economic Performance 2022.

_ , Julia Schmidt, and Riccardo Zago, "Patents that match your standards: firm-level evidence on competition and innovation," Discussion Paper dp1881, London School of Economics and Political Science. Centre for Economic Performance 2022.

_ , Yoann Potiron, and Juste Raimbault, "Classifying patents based on their semantic content," *PloS one*, 2017, *12* (4), e0176310.

Berkes, Enrico, "Comprehensive universe of US patents (CUSP): data and facts," 2018. Unpublished, Ohio State University.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, March 2003, *3* (null), 993–1022.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, "Are ideas getting harder to find?," *American Economic Review*, 2020, *110* (4), 1104–1144.

Bryan, Kevin A, Yasin Ozcan, and Bhaven Sampat, "In-text patent citations: A user's guide," *Research Policy*, 2020, *49* (4), 103946.

Calvino, Flavio, Chiara Criscuolo, Hélène Dernis, and Lea Samek, "What technologies are at the core of AI?," 2023, (6).

de Rassenfosse, Gaétan and Ling Zhou, "Patents and supra-competitive prices: Evidence from consumer products," 2020. Available at SSRN 3756359.

_ , Gabriele Pellegrino, and Emilio Raiteri, "Do patents enable disclosure? Evidence from the invention secrecy act," *International Journal of Industrial Organization*, 2024, *92*, 103044.

Dechezleprêtre, Antoine, David Hemous, Morten Olsen, and Carlo Zanella, "Automating labor: evidence from firm-level patent data," *Available at SSRN 3508783*, 2019.

Dunham, James, Jennifer Melot, and Dewey Murdick, "Identifying the development and application of artificial intelligence in scientific text," *arXiv preprint arXiv:2002.07143*, 2020.

Fadeev, Evgenii, "Creative Construction: Knowledge Sharing and Cooperation Between Firms," 2023. Mimeo Duke University.

Frankel, Alexander P, Joshua L Krieger, Danielle Li, and Dimitris Papanikolaou, "Evaluation and Learning in R&D Investment," Working Paper 31290, National Bureau of Economic Research May 2023.

Ganglmair, Bernhard, W Keith Robinson, and Michael Seeligson, "The rise of process claims: Evidence from a century of US patents," *ZEW-Centre for European Economic Research Discussion Paper*, 2022, (22-011).

Ganguli, Ina, Jeffrey Lin, Vitaly Meursault, and Nicholas F Reynolds, "Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection," Working Paper w32262, National Bureau of Economic Research 2024.

Gerken, Jan M. and Martin G. Moehrle, "A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis," *Scientometrics*, 2012, *91* (3), 645–670.

Ghosh, Mainak, Sebastian Erhardt, Michael E Rose, Erik Buunk, and Dietmar Harhoff, "PaECTER: Patent-level Representation Learning using Citation-informed Transformers," *arXiv preprint arXiv:2402.19411*, 2024.

Gordon, Robert J., "Does the "New Economy" Measure Up to the Great Inventions of the Past?," *Journal of Economic Perspectives*, December 2000, *14* (4), 49–74.

Gordon, Robert J, "The Demise of U.S. Economic Growth: Restatement, Rebuttal, and Reflections," Working Paper 19895, National Bureau of Economic Research February 2014.

Griliches, Z, "Patent Statistics as Economic Indicators: A Survey," 1998.

Griliches, Zvi, "Issues in Assessing the Contribution of Research and Development to Productivity Growth," *Bell Journal of Economics*, 1979, *10* (1), 92–116.

Grupp, Hariolf, "Foundations of the Economics of Innovation: Theory, measurement and practice," in "Foundations of the Economics of Innovation," Edward Elgar Publishing, 1998.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg, "Market value and patent citations: A first look," Working Paper w7741, National Bureau of Economic Research 2000.

_ , _ , and _ , "The NBER patent citation data file: Lessons, insights and methodological tools," Working Paper w8498, National Bureau of Economic Research 2001.

Hampole, Menaka, Dimitris Papanikolaou, Lawrence D.W. Schmidt, and Bryan Seegmiller, "Artificial Intelligence and the Labor Market," Working Paper 33509, National Bureau of Economic Research February 2025.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*," *The Quarterly Journal of Economics*, 10 2017, *133* (2), 801–870.

Hausman, Naomi, "University innovation and local economic growth," *Review of Economics and Statistics*, 2022, *104* (4), 718–735.

Higham, Kyle, Gaétan De Rassenfosse, and Adam B Jaffe, "Patent quality: Towards a systematic framework for analysis and measurement," *Research Policy*, 2021, *50* (4), 104215.

Hu, Jie, Shaobo Li, Yong Yao, Liya Yu, Guanci Yang, and Jianjun Hu, "Patent keyword extraction algorithm based on distributed representation for patent classification," *Entropy*, 2018, *20* (2), 104.

Jaffe, Adam B, "Real effects of academic research," *The American economic review*, 1989, *79*, 957–970.

Jones, Benjamin F and Lawrence H Summers, "A Calculation of the Social Returns to Innovation," Working Paper 27863, National Bureau of Economic Research September 2020.

Kakhbod, Ali, Leonid Kogan, Peiyao Li, and Dimitris Papanikolaou, "Measuring Creative Destruction," Research Paper 7234-24, MIT Sloan 2024.

Kalyani, Aakash, "The Creativity Decline: Evidence from US Patents," 2024. Mimeo Federal Reserve Bank of St. Louis, Research Division.

_ , Nicholas Bloom, Marcela Carvalho, Tarek Hassan, Josh Lerner, and Ahmed Tahoun, "The Diffusion of New Technologies," Working Paper w28999, National Bureau of Economic Research 2021.

Kantor, Shawn and Alexander Whalley, "Knowledge spillovers from research universities: evidence from endowment value shocks," *Review of Economics and Statistics*, 2014, *96* (1), 171–188.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring technological innovation over the long run," *American Economic Review: Insights*, 2021, *3* (3), 303–320.

Kinne, Jan and David Lenz, "Predicting innovative firms using web mining and deep learning," *PloS one*, 2021, *16* (4), e0249071.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological innovation, resource allocation, and growth," *The quarterly journal of economics*, 2017, *132* (2), 665–712.

_ , _ , Lawrence DW Schmidt, and Bryan Seegmiller, "Technology-skill complementarity and labor displacement: Evidence from linking two centuries of patents with occupations," 2021.

Krieger, Joshua, Danielle Li, and Dimitris Papanikolaou, "Missing novelty in drug development," *The Review of Financial Studies*, 2022, *35* (2), 636–679.

Krieger, Joshua L, Monika Schnitzer, and Martin Watzinger, "Standing on the shoulders of science," *Strategic Management Journal*, 2024, *45* (9), 1670–1695.

Kriesch, Lukas and Sebastian Losacker, "A global patent dataset of bioeconomy-related inventions," *Scientific Data*, 2024, *11* (1), 1308.

Kuhn, Jeffrey, Kenneth Younge, and Alan Marco, "Patent citations reexamined," *The RAND Journal of Economics*, 2020, *51* (1), 109–132.

Lanjouw, Jean O and Mark Schankerman, "Characteristics of patent litigation: a window on competition," *RAND journal of economics*, 2001, pp. 129–151.

Lenz, David and Peter Winker, "Measuring the diffusion of innovations with paragraph vector topic models," *PloS one*, 2020, *15* (1), e0226685.

Li, Zehan, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

Loughran, Tim and Bill Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, 2011, *66* (1), 35–65.

Mann, Katja and Lukas Püttmann, "Benign effects of automation: New evidence from patent texts," *Review of Economics and Statistics*, 2023, *105* (3), 562–579.

Marx, Matt and Aaron Fuegi, "Reliance on science: Worldwide front-page patent citations to scientific articles," *Strategic Management Journal*, 2020, *41* (9), 1572–1594.

Masclans-Armengol, Roger, Sharique Hasan, and Wesley M Cohen, "Measuring the Commercial Potential of Science," Working Paper w32262, National Bureau of Economic Research 2024.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

Pakes, Ariel and Zvi Griliches, "Patents and R&D at the firm level: A first report," *Economics Letters*, 1980, *5* (4), 377–381.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in "Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)" 2014, pp. 1532–1543.

Petralia, Sergio, Pierre-Alexandre Balland, and David L Rigby, "Unveiling the geography of historical patents in the United States from 1836 to 1975," *Scientific data*, 2016, *3* (1), 1–14.

Poege, Felix, Dietmar Harhoff, Fabian Gaessler, and Stefano Baruffaldi, "Science quality and the value of inventions," *Science advances*, 2019, *5* (12), eaay7323.

Rassenfosse, Gaetan De and Emilio Raiteri, "Technology protectionism and the patent system: Evidence from China," *The Journal of Industrial Economics*, 2022, *70* (1), 1–43.

Rassenfosse, Gaètan De, Helene Dernis, Dominique Guellec, Lucio Picci, and Bruno Van Pottelsberghe De La Potterie, "The worldwide count of priority patents: A new indicator of inventive activity," *Research Policy*, 2013, *42* (3), 720–737.

Romer, Paul M, "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5, Part 2), S71–S102.

Schmookler, Jacob, "Invention and economic growth," 1966.

Seegmiller, Bryan, Dimitris Papanikolaou, and Lawrence D.W. Schmidt, "Measuring document similarity with weighted averages of word embeddings," *Explorations in Economic History*, 2023, *87* (C).

Shi, Feng and James Evans, "Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines," *Nature Communications*, 2023, *14* (1), 1641.

Tetlock, Paul C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *The Journal of Finance*, 2007, *62* (3), 1139–1168.

Trajtenberg, Manuel, "Patents, citations and innovations: tracing the links," 1987.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is All you Need," in I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 30 Curran Associates, Inc. 2017.

Verluise, Cyril, Gabriele Cristelli, Kyle Higham, and Gaétan de Rassenfosse, "The missing 15 percent of patent citations," *Available at SSRN 3754772*, 2020.

Webb, Michael, "The impact of artificial intelligence on the labor market," *Available at SSRN 3482150*, 2019.

_ , Nick Short, Nicholas Bloom, and Josh Lerner, "Some facts of high-tech patenting," Working Paper w24793, National Bureau of Economic Research 2018.

Yoon, Byungun and Yongtae Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *The Journal of High Technology Management Research*, 2004, *15* (1), 37–50.