The commercial potential of science and its realization: Evidence from a measure using a large language model*

Roger Masclans† Sharique Hasan† Wesley Cohen‡

November 22, 2023

Abstract

The goals of our study were: 1.) to develop an *ex-ante* measure of the commercial potential of academic scientific contributions with a large language model employing machine learning, and 2.) to apply that measure to identify factors conditioning the realization of that potential. The premise of our study is that to understand the determinants of the commercial realization of academic science, a useful first step is to identify the "at-risk set"—that is, those articles, conference proceedings, etc., that offer commercial potential. Using the measure, we study the commercialization process with two empirical exercises. First, we analyze the commercialization of over 96,000 articles from a leading research university. Our measure predicts which science is disclosed, investments by the technology transfer office (TTO), patent filings, startup formation, and licensing. As highlighted by our analysis, the primary barrier to commercialization is either the scientists' decision not to disclose their discoveries or the TTO's limited awareness of pertinent researchers or science within their institutions. Many high-potential articles remain undisclosed; this science is less likely to be cited in corporate patents. However, the quantity of science affecting commercialization via publication outside the TTO channel is much greater. Second, we analyzed the commercialization outcomes of over 5.2 million articles published between 2000-2020. Articles with high commercial potential are more frequently incorporated into patents, regardless of their originating institution. High-potential discoveries from institutions with a history of high commercial impact have commercialization rates 14.1% higher than their peers. We find that what accounts for the preponderance of differences in commercialization rates across universities are not factors potentially impeding the identification of commercially promising science. Rather, these differences are predominantly accounted for by factors accounting for the differences in the actual production of commercially viable science.

Keywords *— Innovation, Technological Opportunity, Science Commercialization, Deep Learning*

*Authorship in reverse alphabetical order. Sharique Hasan would like to thank the Kauffman Foundation who funded this study through their knowledge challenge grant.

†Duke University, The Fuqua School of Business.

‡Duke University, The Fuqua School of Business and National Bureau of Economic Research

1 Introduction

Scientific research is crucial for technological advance.[1] There is, however, a longstanding question of the extent to which the commercial promise of science is unrealized. This is not a new issue. Indeed, a desire to commercially capitalize upon public research for social and private benefit in the U.S. has motivated numerous policy initiatives, from policies for disseminating the expertise of land grant colleges to farmers in the late 19th century to the Stevenson-Wydler and Bayh Dole Acts of 1980, as well as subsequent federal and state legislation. Assessing whether and under what conditions the commercial potential of science is unrealized is challenging due to the lack of measures of the commercial potential of science. In this paper, we introduce one such measure and use it to analyze, in a limited setting, the process of the commercialization of public research, as well as the importance of selected impediments to that process.

Recent research identifies frictions impeding the market application of academic science. For example, Bikard and Marx (2020) show that firms disproportionately build upon academic research that originates in academic hubs (i.e., geographic concentrations of academic research). Koffi and Marx (2023) similarly show that firms manifest gender bias in what science they build upon–or at least acknowledge–during their inventive activity. Azoulay et al. (2007) note a similar gender bias in academic patenting. These frictions confirm our expectation that firms are not taking advantage of some fraction of commercially relevant science.

While we can approximate how often academic research has led to commercial applications by employing measures such as patent citations to the scientific literature, such measures do not reveal how much of that research had the potential for commercial use *ex-ante*. As a result, we cannot determine how much research with commercial value remains

[1] Examples of academic research that later transformed industries abound, from blue light-emitting diodes (LEDs) (e.g., Nakamura et al., 1994) to mRNA vaccines (e.g., Jackson et al., 2020) to the PageRank algorithm that transformed internet search (Page et al., 1998).

undeveloped.[2] Indeed, we suspect that most public research output does not have commercial potential, at least in any immediate sense. Most public research is not undertaken with commercial applications in mind. It is either an input into subsequent scientific research, answering academic questions with little commercial import, or is too embryonic and distant from any market application. Thus, to address the question of the conditions under which science with commercial potential remains undeveloped, a useful first step is to identify science that offers commercial potential, where commercial potential refers to the likelihood that a firm believes that a given scientific finding or idea may ultimately contribute to the development of a product or process that generates economic value.

This paper is partly motivated by the challenge facing R and D managers, venture capitalists, and technology transfer office (TTO) personnel to identify–from a vast array of scientific publications, reports, etc.–those ideas and findings that offer commercial potential, and especially how to mitigate the uncertainty that such identification entails. Even once they identify a scientific finding as promising, the stakeholders possessing the authority and resources to develop a promising idea must then evaluate whether it can cost-effectively contribute to the development of a new or improved product, as well as whether there will be a market for such a product (Cohen et al., 2002; Fleming and Sorenson, 2004). The associated uncertainties mean that decisions about a finding's commercial potential can be prone to errors, limited by the available information, or influenced by bias. In this paper, we develop a summary measure that can inform stakeholders' initial evaluations of the commercial potential of a scientific finding or idea. In addition to its practical merit, we also suggest–and will show–that such a measure can contribute to academics' efforts to improve

[2]Using such data, we observe substantial variation in the commercialization of science across institutions, researchers, and fields. For example, consider the scientific articles in the natural and applied sciences and engineering produced at the top 200 U.S. research institutions since 2000. For institutions at the top of the ranking, 12% of the articles are cited by a corporate patent, whereas for the median institution, only 8% are—a 50% difference. Such variation gives us some sense of differences in the degree to which the scientific output of different institutions is related to commercial outcomes.

Nevertheless, we have little idea of how much research has commercial potential. As a consequence, this variation cannot tell us about the extent to which the commercial potential of scientific contributions goes unrealized.

understanding of the contribution of scientific knowledge to technical advance and the role of the institutions and individuals who create that knowledge.

To develop our measure, we use natural language processing to train and validate a machine learning model to develop an *ex-ante* measure for the commercial potential of a scientific article. For training and validation, we use the text of an academic article's abstract as an input to produce *ex-ante* and out-of-sample predictions for any given scientific article's commercial potential. Conceptually, commercial potential in our setting refers to the likelihood that a firm believes an article may contribute to developing a marketable product or process. This concept is operationalized *as the ex-ante probability that a scientific article will be cited in a patent that is later renewed*. By relying on citations to articles in patents, our operationalization assumes that such a citation reflects a firm's belief that a given scientific finding or idea, once incorporated into an invention, may offer economic value (Cornelli and Schankerman, 1999; Kuhn et al., 2020; Marx and Fuegi, 2020).

We implement a two-step approach to estimate the likelihood of a scientific article possessing commercial potential. Initially, we utilize the abstract text of scientific articles along with an indicator of whether a renewed patent cites the article. For this purpose, we employ the BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2018) to develop a language-based learning model. This model is trained to identify textual patterns predicting renewed patent citations. Subsequently, once the model is trained, we calculate a probability, both out-of-sample and out-of-time-period (with no data in the training period overlapping with our target article published at time t), that an article will be cited in a patent renewal—our proxy for commercial potential. The average accuracy of our model, as well as the average area under the receiver operating characteristic curve (AUROC), is 0.74.[3]

[3]For context, Manjunath et al. (2021) report an AUROC of 0.83 in their model predicting patent citations of articles. However, their model exclusively utilizes abstracts from PubMed in the life sciences and does not account for patent renewals. Conversely, Liang et al. (2022) developed a model based on the text of inventions disclosed to Stanford's Technology Transfer Office (TTO), aiming to predict commercial value generation, achieving an AUROC of 0.76.

We use our model in two related empirical exercises to study the relationship between the commercial potential of a scientific finding and its realization. First, using detailed administrative data from a major research university's technology transfer office (TTO), we relate our measure of commercial potential to a scientific idea's progress in the commercialization process. Science with high commercial potential is more likely to undergo commercialization, entailing disclosure to the technology transfer office, financial investment, patenting, licensing to a startup or incumbent firm, and sometimes revenue generation. While of substantive interest, these findings also increase confidence in the measure. Our findings suggest that the most significant friction in the commercialization process arises early in the process—when scientists decide whether to disclose their science to the TTO. To explore how much of this undisclosed science may have commercial value, we also examine corporations' use of this undisclosed but high-potential work without TTO involvement. Within this pool of ideas never disclosed to the university, high-potential science is cited in corporate patents at significant rates, although at significantly lower rates than the science that *is* disclosed—articles with commercial potential that are not disclosed to the TTO are 58.05% less likely to be cited by a corporate patent than articles disclosed to the TTO. The fact, however, that the undisclosed quantity of high commercial potential science (i.e., top quartile) that is cited is almost double the quantity of disclosed high potential science suggests that publication, absent TTO involvement, is the essential vehicle through which technology makes its way into commercial applications. The result also suggests that an important obstacle to the TTO's support of the commercialization process appears to be the role of faculty in choosing not to disclose. Once these ideas are disclosed, our results suggest that the decisions of the technology transfer office to develop and invest are aligned with the commercial potential of the disclosed invention as indicated by our measure.

Our second exercise analyzes commercial outcomes (patent citations and renewals) for over 5.2 million academic papers published by U.S.-based universities. We find strong evidence that scientific articles with high commercial potential, per our measure, are more

likely to be commercialized, as reflected by a citation in a renewed patent. Moreover, our findings indicate that an article's commercial potential is distinct from its scientific or social 'impact' potential and varies meaningfully across a researcher's portfolio of work and an institution's prior commercial impact. Regarding effect size, a scientific article with a commercial potential score in the top quartile is over 20 times more likely to be cited by a patent than those classified by our algorithm as being in the bottom quartile. A key finding from our institutional analysis is that what accounts for the preponderance of differences in commercialization rates across universities are not factors potentially impeding the identification of commercially promising science but factors accounting for the differences in the actual production of such science. Quantitatively, approximately 60% of the variation in rates of patent citations of papers across universities can be explained by the mean *ex-ante* commercial potential scores for that institution.

Our study makes several contributions to the literature on innovation, entrepreneurship, and the commercialization of science. We apply a novel methodological approach that enables us to establish and validate a comprehensive measure of *ex-ante* commercial potential that precedes actual commercial outcomes. This approach facilitates identifying and tracking nascent scientific ideas and their commercial trajectory, which is not feasible with outcome-based measures or natural experiments that lack observable indicators of commercial potential. Furthermore, we utilize the discrepancies between potential and actual outcomes to identify untapped technological opportunities. Additionally, we extend the literature on the impact of academic science on technical advancement by offering insights into how scientific ideas progress through the commercialization process, with and without university involvement. Finally, our findings contribute to emerging research exploring factors affecting the commercialization of science that considers scientists' motives (Cohen et al., 2020) and experience (Marx and Hsu, 2022; Azoulay et al., 2010), biases in research evaluation processes (Li, 2017; Lane et al., 2022), and organizational frictions in universities (Hsu and Kuhn, 2023).

Our work also builds on prior research insights that have examined the gaps between the commercial potential of public research and its realization. Some researchers have focused on identifying the specific features of researchers and teams that may contribute to such gaps. For example, (Ding et al., 2006) have investigated the impact of gender and ethnic diversity on the commercialization of research and the role of collaboration and interdisciplinary approaches. Other researchers have highlighted the importance of institutions in the commercialization process. These scholars have explored how factors such as the strength of an institution's patenting and licensing activities (Henderson et al., 1998; Williams, 2013), the involvement of technology transfer offices (Debackere and Veugelers, 2005), and the level of industry engagement can affect the commercialization of scientific ideas. More recently, Marx and Hsu (2022) have used a novel approach, known as "twin discoveries", to conclude the factors influencing the commercialization of scientific ideas via venture creation. While we are also concerned with the methodological challenges posed by the unobservable commercial potential of science, our approach focuses not only on startup commercialization but also on the commercialization of research by established firms. We use machine learning techniques to develop a measure of commercial potential based on all firms that patent and commercialize science rather than relying on a comparison of twin discoveries. By doing so, we aim to provide a more comprehensive and generalizable understanding of the factors that influence the commercialization of scientific discoveries.

2 Setting: The Commercialization of University Science

To assess the commercial potential of science and estimate unrealized potential, our empirical analysis reflects the choices of the three types of participants: (1) Individual researchers,

whom we will refer to as "academics";[4] (2) Their affiliated *institutions*, primarily universities; and (3) The *firms* that leverage scientific findings to develop new products, processes, and services. The knowledge transfer from academics and their institutions to firms can be seen from supply and demand perspectives. The supply side consists of the academics conducting the research and the institutions that provide them with the physical and intellectual infrastructure that supports their research. These institutions are also partly responsible for defining and administering the incentive systems that guide academic endeavors. On the demand side are firms that utilize this research.

Although framing this problem as one of demand and supply sides is useful, it is essential to note that this is not a market. First, academics are not "motivated sellers". Unlike firms, their objectives are typically not personal profit from commercial activity. We know from a rich literature (e.g., Merton, 1973; Stokes, 2011)that their motives tend to be the achievement of eminence, career advancement, the advancement of knowledge per se, or, as also suggested by Stokes (2011) and observed by Cohen et al. (2020), a desire to address societal needs. Undertaking commercial activity for profit is not the dominant ethos. In addition, there is typically no market price tied to academics' research output, even when it is usable by firms because much of the commercially relevant science produced by academics is placed in the public domain via publication, conferences, and reports (Arora et al., 2016; Cohen et al., 2002), and is effectively free to firms.

Yet, this setting in terms of supply and demand provides a simple framework for considering the factors that may affect the realization gap. The key here is to consider the loci of the different decisions that may affect the gap. For instance, academic researchers operate within specific incentive structures on the supply side. As suggested above, their motives are heterogeneous and contrary to businesses, they are often driven by something other than the lure of commercial profit. The "translational" effort required to make academic science commercially viable can also be significant. The translation of scientific findings into language

[4]This is a simplification since the same rationale can be applied to *researchers* in government labs and nonprofit research institutions.

and contexts that businesses can grasp and utilize requires clearly describing the research and insights in lay language and demonstrating its practical applications and, at times, its active promotion. Thus, many academics will only be inclined to expend the additional effort necessary for commercial translation if it aligns with their personal and academic goals. Moreover, the existing institutional incentives may not reward, no less recognize, such endeavors. Consequently, a wealth of potentially valuable scientific insights may not be brought to the attention of the firms that would otherwise build on them.

Another feature of the supply side may be the level of resources a university administration decides to devote to translational efforts through its TTO and the institutes and centers whose mission is translation. Such initiatives have increased over the past twenty years. Institutions also play a significant role through their impact on networks and norms.

Firms are on the demand or "use" side. Recent studies, including those by Cohen et al. (2002) and Bikard (2018), suggest that firms pay limited attention to university research. Instead, firms are more likely to build on knowledge from other firms. This behavior is understandable. While the distance of most academic research from commercial applications undoubtedly accounts for this pattern, another factor is likely the high cost of sifting through massive numbers of articles, reports, etc. Moreover, to effectively harness academic insights, businesses need to invest in the requisite absorptive capacity (Cohen and Levinthal, 1989) needed to assess the commercial value of potentially thousands of articles and researchers and use it. This becomes more challenging and costly because academic authors have not taken steps to make their findings more accessible or "translated" them into practical applications.

In light of the quantity of research and the uncertainty regarding its relevance, quality, and potential payoff, businesses often adopt simplified search strategies. They might focus on research from top-tier institutions, reputed departments, or recognized academics. Previous interactions or collaborations with academics, such as consulting or technical discussions, influence these choices. Geographic proximity can further narrow a business's focus: companies often tap into nearby research hubs. Studies like those by Bikard and Marx (2020)

and the academic spillover research by Jaffe et al. (1993) emphasize the role of geographical proximity in structuring knowledge transfer. These streamlined search strategies might lead firms to overlook valuable academic research. To what extent commercializable academic insights are missed remains an empirical question.

In summary, potential supply-side explanations for the realization gap may include the motivations of academics, the incentive structures within their institutions, and the resources these institutions allocate to translation. On the user side, explanations could improve a firm's ability to assimilate and apply new knowledge (absorptive capacity), coupled with uncertainty and search costs. These factors may lead firms to depend on characteristics of academic institutions and researchers, such as the reputations of institutions, departments, and individuals, as well as on geographic proximity and established relationships, to direct their search. While these attributes of academics, their institutions, and firms may account for a part of the realization gap, it must be recognized that a specific segment of this gap may inherently be due to the duration required, even with ample resources, to transition foundational science to a commercial application.

3 Data and Methods

This section describes the methodology and data used to create the measure of *commercial potential*. As previously mentioned, based on the abstract text in which a scientific discovery is reported, this measure estimates the likelihood that it will be incorporated into a product, service, or technology that generates economic value. The measure is derived from a set of Large Language Models (LLM) fine-tuned using over 500,000 scientific articles in Natural Language Processing (NLP) classification tasks to predict the commercial potential of articles published between 2000 and 2020.[5] We use patent-to-paper citations to link a scientific

[5]LLM and NLP are subfields of artificial intelligence (AI) that focus on the interaction between computers and human language and involve, among others, developing computational models and algorithms that enable computers to analyze and understand natural language. An NLP model is a machine learning algorithm designed to classify textual data into one or more predefined categories automatically. Based on

finding with a patented invention, and we train the NLP models with papers cited by patents that have been renewed at least once. We interpret a renewed patent's citation to an article as indicating that a firm believes the article contributes to developing an invention that may have commercial value. We label scientific articles as having commercial potential if they are cited by at least one patent that has been renewed one or more times and papers as having no commercial potential otherwise. The language models output the probability that a new, unseen paper is associated with each of the predefined classes–possessing commercial potential and not possessing commercial potential. We trained the models with two classes and defined the commercial potential measure (ϕ) as the likelihood that the paper is classified as being in the class of papers with commercial potential.

3.1 SciBERT: Bidirectional Transformers for Scientific Language Understanding

Several techniques for NLP classification exist, ranging from logistic regression based on bag-of-words to advanced deep neural networks that interpret semantic relationships across large documents. The models we use are based on the latter. Specifically, we fine-tune SciBERT (Beltagy et al., 2019), which is a language model based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), an NLP model, developed by Google AI, that has been trained on a large amount of text data. SciBERT was pre-trained on a large corpus of scientific publications and then fine-tuned using our dataset to classify scientific articles based on their commercial potential.[6]

the text content, the model learns to recognize patterns and features indicative of each category and uses this knowledge to predict the most likely category or categories for new, unseen text data.

[6]Thanks to being trained with scientific, domain-specific text, SciBERT provides state-of-the-art performance in a wide range of NLP tasks for scientific domains, improving BERT's performance (Beltagy et al., 2019). We tested whether this holds in our classification task. Indeed, our models' performance increases when using SciBERT instead of BERT. Both SciBERT and BERT utilize transformers (Vaswani et al., 2017), a novel type of neural network architecture that has revolutionized the field of NLP. Transformers can model long-range dependencies, learn contextual representations, and handle out-of-vocabulary words and syntax. At a high level, a transformer model consists of multiple layers of self-attention and feed-forward neural networks, enabling it to weigh the probabilities of different parts of the input sequence (i.e., sentences of the text) and process it in parallel. The attention mechanisms allow transformers to learn contextual

A possible limitation of our analysis is that the training sample for SciBERT (Beltagy et al., 2019) comprised 82 percent life science articles and 18 percent computer science articles. Although these two fields represent a large share of the entire corpus of published articles, this could represent a limitation given that we are also trying to evaluate the commercial potential of articles from fields other than life sciences and computer science.

3.2 Processing the input text

Language models use text as input; in our case, we use the abstracts of the articles in which findings are reported. However, before feeding the text into the model, the abstracts are first tokenized.[7]

It is worth noting that, for computational reasons, SciBERT, like BERT, is limited to processing up to 512 tokens per document. There are various techniques to handle longer documents, but a simple analysis of the abstracts we use to train our model reveals that only 1% of them contain more than 512 tokens. Additionally, there are no differences in the average number of tokens between the classes (which could create bias in our findings). Therefore, we truncate the abstracts at 512 tokens.[8]

representations of words and phrases.

[7]This first step of tokenizing entails converting each abstract into an array of discrete linguistic units—usually, units are words, parts of words, numbers, symbols, and stems. Based on BERT's tokenizer, we tokenize using the version that SciBERT's authors recommend, *scibert-scivocab-uncased*. This is expected to yield the highest performance. The tokenizer maps each word into an integer based on the model's vocabulary and adds special tokens such as sentence separators, padding, and classification task-specific codes. For each token, the tokenizer looks for the pre-trained embeddings of the model (Token Embeddings) – a vector representing each word in a high-dimensional space in relation to an extensive vocabulary. In addition, the tokenizer adds information regarding the position of each token in the text, both in the sentence (Segment Embeddings) and in absolute terms (Position Embeddings). Combining the three embeddings produces a unique embedding for each token in the abstract, which serves as the input to the first layer of the neural network.

Furthermore, in the case of bidirectional transformers, the embeddings capture information about the token's relative position within a document, enabling the contextualization of its meaning when fine-tuning the models. We tested this method and found that it improved the models' performance by not modifying the input text. Unlike other NLP techniques, transformers do not necessarily require removing stop words, punctuation marks, numbers, special characters, or stemming the text words to improve their performance. Transformer models' tokenizers may split specific words into subwords and characters. However, this exercise is left to the tokenization exercise, which depends on the specific pre-trained language model used. SciBERT's tokenizer, in particular, uses its wordpiece vocabulary based on a subword segmentation algorithm created to match best the corpus of scientific papers used to train the model (*scivocab*) (Beltagy et al., 2019).

[8]Figure B.1 in Appendix B shows the distribution of abstracts' length.

3.3 Data

3.3.1 Scientific articles data

We use Dimensions as a source for scientific publications. Dimensions is a research and innovation database that contains detailed information on publications, patents, grants, clinical trials, and policy documents. In particular, the set contains information on more than 139 million publications with their title, abstract text, publication sources, author information, fields of research, and other metadata. We remove works that do not have abstract text, works that do not have a Digital Object Identifier (DOI), and duplicate DOIs.[9]

We limited our analysis to peer-reviewed journal articles and findings reported in conference proceedings to ensure high-quality scientific findings, resulting in 70,805,788 scientific papers. Moreover, we focused our analysis on eleven scientific fields, which cover the majority of natural and applied sciences and engineering but exclude the social sciences. The fields are Agricultural, veterinary, and food sciences; Biological sciences; Biomedical and clinical sciences; Chemical sciences; Earth sciences; Engineering; Environmental sciences; Health sciences; Information and computing sciences; Mathematical sciences; and Physical sciences. After conditioning on these fields, our sample size decreased to 50,362,042 academic papers.

3.3.2 Patent data

Second, we sourced patent citations to scientific papers from the Reliance on Science dataset (Marx and Fuegi, 2020). This dataset contains 22,660,003 linkages between 3,017,441 unique patents and 4,017,152 unique papers drawn from over 160 million works. Using the DOI of a paper, we merged the Reliance on Science dataset with Dimensions. This resulted in all 4,017,152 papers in the Reliance on Science set being matched to a paper in the Dimensions subsample we created, which accounts for 7.98% of the papers in our sample being cited by one or more patents. For this analysis, we assume that papers not cited by a patent in the

[9] A Digital Object Identifier is a persistent identifier or handle used to uniquely identify various objects, standardized by the International Organization for Standardization.

resulting set have no patent citations.[10]

Next, we use data from Google Patents Research and the United States Patent and Trademark Office (USPTO) to append to our dataset information on assignee and patent renewal, respectively. Using the patent publication number, we merge the three datasets containing patent information (Reliance on Science, Google Patents Research, and USPTO).

3.3.3 Technology Transfer Office Data

In addition to public data on papers and patents, we have access to detailed data from the technology transfer office (TTO) of a major research university, which includes data on all invention disclosures and subsequent actions tied to those disclosures, including patenting, licensing, agreements, revenues, TTO investments tied to each invention, whether the licensee is a startup or an established firm, and inventor identity, including the inventor's history with the TTO. We remove inventions disclosed before 2000 and those not associated with an active researcher at the time of disclosure. The resulting dataset includes 5,219 invention disclosures from January 2000 to December 2020.

We take three steps to identify the scientific findings relevant to an inventor's disclosure. First, we match our two primary datasets: (a) Dimensions, containing academic publications information, with the (b) TTO dataset. We extract from Dimensions the names of all researchers affiliated with the TTO's University at any time. Next, we use a fuzzy matching algorithm to match the researchers' names from Dimensions to those of researchers who disclosed inventions in the TTO data. The matched dataset contains publications and invention disclosures matched by author name. Between 2000 and 2020, 4,367 researchers matched from one dataset (Dimensions) to the other (TTO data) and linked to 53,180 unique publications and 4,505 inventions.

An invention disclosure is not associated with a paper merely based on a common author

[10]This assumption likely introduces error into our estimates by leading us to classify articles that are likely to be cited by patents as those that are not, thus reducing the out-of-sample accuracy of our model. This will likely lead to a conservative bias in our estimates.

appearing in a paper or invention disclosure. We rely on two additional factors to better match invention disclosures with a set of academic papers that contribute to the invention. First is the time proximity between the publication and the disclosure. The second is based on text similarity between publications and inventions.

To assess the temporal relationship between academic papers and invention disclosures, we introduce a measure called 'time gap.' This measure calculates the years between the publication of a paper and the corresponding invention disclosure. We use the year of the invention disclosure as our reference point, marked as time '0'. The time gap is then defined as the difference in years between when the paper was published and the year of disclosure. For instance, if a paper was published in 2013 and its associated invention was disclosed in 2015, the time gap is two years. Another paper by the same author published in 2020, associated with the same invention, would have a time gap of five years.

We determine a paper's influence on an invention—essentially matching a paper to an invention—based on whether the time gap falls within a specific range. Specifically, we use a time window of [-1,3] years. This range is based on guidelines from the Technology Transfer Office (TTO), which advises inventors to disclose their inventions before public dissemination to maintain patenting options in jurisdictions without a one-year grace period post-publication. Research also indicates that scientific publications leading to patents are often temporally close to each other (Azoulay et al., 2007; Marx and Fuegi, 2020). Applying this method identifies 3,173 researchers linked to 19,381 publications and 3,127 inventions.[11] In the last step of our methodology, we refine the matches between publications and inventions based on textual similarity. We employ a technique similar to the one described earlier, using BERT to generate textual embeddings for both the titles of papers and inventions. For each potential publication-invention pair (identified by having a common author and falling

[11]Note that an invention can have more than one inventor. Thus, an invention can be matched to the publications of more than one researcher. Similarly, a publication can be matched to more than one invention, either because the publication has one author with more than one invention within the time window or because the publication has more than one author who has disclosed at least one invention within the time window.

within the [-1,3] time window), we calculate the cosine similarity between the embeddings of their titles. Based on our analysis, we conclude that matches with similarity scores above 0.5 likely indicate publications that have influenced an invention.[12]

This three-step process resulted in 13,445 unique publications being matched to 2,728 inventions, with 2,717 researchers linked to these matches. The median number of publications associated with each invention is 2, a finding consistent with other studies examining paper-patent pairs (e.g., Marx and Fuegi, 2020). Following this, we prepare two datasets for our analyses.

The first dataset is aggregated at the article level. It includes information about each article, such as its influence on an invention disclosed to the TTO, its commercial and scientific potential, the number of times it's cited by patents, and other relevant characteristics. This dataset will help us analyze the attributes of scientific articles linked to inventions disclosed to the TTO. It comprises 96,564 articles, of which 13,445 (13.92%) are associated with an invention disclosure.

The second dataset is aggregated at the level of invention disclosures. Here, we examine the relationship between the commercial potential of disclosure and outcomes like TTO investment, patent filings, licensing agreements, and revenue generation. Since inventions are often linked to multiple articles, we average each invention's relevant variables (such as commercial potential, scientific potential, and patent citations). This dataset includes 2,728 inventions.

Table 1 details the variables used in the TTO analysis. Table 3 provides summary statistics for all articles published between 2000 and 2020 by researchers affiliated with the TTO's university at the time of publication, while Panel B of the same table offers summary statistics about the inventions disclosed to the TTO and the scientific research upon which they are based.

[Table 1 about here.]

[12]We also applied this procedure using the publications' abstracts and the inventions' descriptions. While we observed similar results, title-based matching proved less prone to errors.

[Table 2 about here.]

[Table 3 about here.]

3.4 Commercial potential

We have developed a predictive model to estimate the commercial potential of scientific articles, estimating their likelihood of being cited by a patent and whether that patent is subsequently renewed. The model is based on analyzing the text of the article's abstract.

For each year from 2000 to 2020, we created a specific model using data from the preceding ten years up to the year just before the focal year (year $t-1$). This approach was chosen to avoid data generated after the focal year, preventing data leakage and allowing the model to understand the historical connection between text and commercial outcomes.

In our model development, avoiding data leakage was a priority. Data leakage occurs when information from outside the focal period is inadvertently included in the training data, potentially skewing the model's predictions. To prevent this, we carefully selected data up to the year just before the focal year (year $t-1$) for training. For example, when predicting for the year 2000, we only used data up to 1999. This ensures that the model's predictions are based solely on information that would have been available at the time, maintaining the integrity and relevance of our results.

We use articles from ten years preceding that year to train the model for articles published in a specific year (t), specifically from years $t-14$ to $t-5$. For example, for articles published in 2000, we use article data from 1986 to 1995 and citation and renewal data until $t-1$, or 1999. This four-year gap between the article data and the focal year is essential, mainly because we use patent renewals, which occur, on average, four years after a patent's grant, as an indicator of commercial value. This allows four years for the last of our articles in our training sample (at $t-5$) sufficient time to accrue patent citations and renewals.

Our experiments indicated that the model's accuracy did not significantly improve with more than 20,000 observations despite testing up to 100,000. Therefore, we trained each

year's model with a dataset of 20,000 articles. The model aims to predict whether a scientific article will be used in developing commercial inventions, with articles cited by renewed patents classified as having high commercial potential.

We ensured each model had a balanced training sample to avoid bias in our Natural Language Processing (NLP) model, which is particularly important for neural networks. We selected 10,000 articles either never cited by a patent or a non-renewed patent and 10,000 articles cited by a renewed patent. This created a balanced dataset, representative of various scenarios, including articles never cited by patents, those cited by non-renewed patents, and those cited by renewed patents.

In developing each model, we divided the data into three sets: 75% for training, 12.5% for testing, and 12.5% for validation. This division follows machine learning best practices, ensuring unbiased performance and allowing the final accuracy of the model to be evaluated with previously unseen data.[13]

3.5 Commercial potential model: Performance

After training the model with the designated training and test sets, we employ the validation sample set aside during the training phase to assess the model's accuracy using previously unseen data. Different parameters were experimented with during training. The most influential parameters identified were: five epochs (iterations of the neural network optimization procedure), a batch size of 16 (the number of training subsamples processed by the neural network at a time), and a learning rate of 2e-5 (a tuning parameter for minimizing the loss

[13]Neural networks undergo training through iterations, known as epochs, where the model's parameters are first randomly initialized. During each epoch, the input data undergoes forward propagation through the network layers, each performing specific calculations. This results in the transformation of the data into new representations, which are passed through successive layers until the final output is produced. Backward propagation occurs, where the loss function's gradients are computed to update and optimize the network parameters. This forward and backward propagation process repeats for each epoch until the loss function converges or a predetermined number of epochs is reached.

After training, evaluating the model's performance using an out-of-sample validation set is imperative. This requirement arises because the training and test sets utilized during the learning phase cannot be reused for unbiased performance assessment. Consequently, the original dataset is subdivided into three subsets to facilitate this evaluation process.

function). We approached the classification as a multi-class problem and utilized a sigmoid function for the network's final layer.

The model, aimed at gauging the commercial potential of articles published between 2000 and 2020, achieved an average accuracy of 74%. This means that it correctly predicts the class (whether an article will attract citations from renewed patents) for 7.4 out of 10 abstracts. Additionally, the average area under the receiver operating characteristic (AUROC) curve is 0.74, indicating a balanced distinction between false positives and false negatives. The ROC curve is a measure of the true positive rate against the false positive rate at various thresholds, with a perfect classification task yielding an area of 1 and a naive task yielding an area of 0.5.[14]

Tables A and A.2 in Appendix A present examples of articles from the top and bottom 20 percentiles of commercial potential, respectively. These examples are drawn from completely out-of-sample articles published after the end of our training sample period. Detailed performance metrics for each model are provided in Appendix B, Figures B.2, and B.3. While our classifier demonstrates reasonable accuracy, there is potential for further enhancement. Two primary factors may be influencing its performance:

1. **Diverse Academic Fields**: Our single NLP model is trained to classify articles across various fields, from Biology to Materials Science to Computer Science. Textual features indicative of commercial potential can significantly vary between these disciplines. This

[14]A concern is whether the classification task is influenced by the use of certain words that are not fundamentally related to the scientific content, but may superficially suggest greater commercial potential in a scientific contribution. For instance, the model could disproportionately classify abstracts with a 'commercial flavor' as having higher commercial potential. In this scenario, the primary determinant of the results would be the language employed rather than the intrinsic scientific research and its potential commercial applications. We randomly selected 100,000 abstracts from our article database to empirically investigate this possibility and utilized ChatGPT to modify each abstract to appear more commercially applicable. The specific instruction given to ChatGPT was: 'Pretend you are an academic researcher revising the abstract of your paper to accentuate its commercial appeal. Impart the notion that the paper has commercial applications without introducing new information. Retain all original details in the modified text, ensuring its suitability for academic publication.' Visual inspection confirmed that the ChatGPT-modified abstracts adopted more commercially oriented language while preserving the original content. Subsequently, these modified abstracts were inputted into our model for new predictions of commercial potential. This allowed us to compare, for identical scientific findings, whether an abstract written with a 'commercial flavor' receives higher commercial potential scores. Our findings are qualitatively robust to commercial language, indicating no significant differences in commercial potential scores between the original and modified abstracts.

diversity necessitates compromises in parameter settings, consequently limiting the model's overall performance. For comparison, Manjunath et al. (2021) focused their NLP model exclusively on the life sciences and biomedical fields, utilizing over 20 million articles from PubMed. They achieved an AUROC of 0.83, highlighting the benefits of field-specific models. In future iterations of our research, we aim to develop separate models for each of the eleven fields to potentially enhance performance.

2. **Complexity of Task**: Predicting commercial potential from textual data is inherently complex and uncertain, making it challenging even for expert human analysis. While most NLP classification tasks, such as identifying specific emotions in text, report accuracies above 95%, these tasks typically involve more straightforward information within the text. For more complex tasks, lower performance is expected. For instance, Liang et al. (2022) trained two NLP models to predict the financial success of inventions disclosed to Stanford's Technology Transfer Office. Their BERT-based model achieved an AUROC of 0.76, while the simpler TF-IDF-based model reached 0.71. Similarly, Guzman and Li (2023) used doc2vec to predict the early-stage success of startups and reported AUROCs between 0.60 and 0.65.

3. **Changing language**: Scientific knowledge, constantly evolving, is mirrored in the ever-changing language of research. Our study, however, is limited to analyzing the text of academic article abstracts and titles. This focus narrows our model's capacity to capture the nuanced dynamics of token emergence, usage, and interconnections and the detailed content in full texts, tables, and figures of articles. These constraints hinder our model's ability to fully grasp the depth and context of scientific discourse, thereby impacting the accuracy of predictions, especially for complex outcomes like commercial success. Moreover, to the extent that the training of our model is based on publications from the period spanning t-14 to t-4 relative to the focal year, the emergence of new technical language and terms of art can bias our analysis.

These comparisons underscore the challenge inherent in tasks with uncertain outcomes and are heavily context-dependent. They also indicate the potential for improved performance through methodological refinements tailored to specific contexts. Furthermore, in the subsequent empirical analyses, we intend to juxtapose the predictions of our model with human decision-making processes. In this context, gaining a detailed understanding of the sources of errors becomes imperative. For instance, consider a scenario where the model forecasts that a patent should cite an article, but it does not. This discrepancy could arise from two possibilities: firstly, the model's prediction might be incorrect, indicating that decision-makers were justified in not utilizing the scientific knowledge from the article (indicative of a model error), or secondly, the model's prediction might indeed be accurate, suggesting that the decision-makers overlooked or misjudged the value of the information in the article (suggesting human error). The current version of the paper does not make this critical distinction in its presentation of results.[15]

3.6 Secondary models: Scientific and social impact

In addition to commercial potential, other attributes of scientific findings, namely their scientific potential and social impact potential, may also influence decisions regarding commercialization. These aspects, while distinct, are not necessarily in opposition to commercial viability. Cohen et al. (2020) observed that life science academics driven by a strong desire for social impact were among the most prolific in patenting their research. However, the prioritization of scientific and social impact can vary by field and individual researcher and may only sometimes align with research geared towards commercial application.

[15]To begin addressing the issue, one approach involves applying uncertainty quantification techniques. For instance, Bayesian neural networks (Blundell et al., 2015) can be employed to explicitly model the epistemic uncertainty inherent in the model's parameters. Situations in which the model demonstrates high uncertainty could indicate areas where it lacks robustness or faces challenges due to inherent architecture or training data limitations. Conversely, to probe errors related to human decision-making, our focus could shift to instances where the model's predictions consistently deviate from human decisions. As proposed by Lakshminarayanan et al. (2017), deep ensemble methods facilitate the generation of multiple predictions, offering a probabilistic perspective. Significant discrepancies between ensemble predictions and human decisions could highlight potential disparities in the decision-making process. This suggests areas where human subjectivity or external factors might be influencing outcomes.

Recognizing the importance of these dimensions, we have developed two additional language-based models to quantify scientific findings' scientific and social impact potentials. These models are conceptually similar to our primary model, which is focused on commercial potential and follows the same structural and methodological approach. However, they are tailored to capture the unique characteristics associated with scientific and social impacts. The specific features and methodologies of each model are detailed subsequently.

3.6.1 Scientific potential

The models estimating scientific potential are developed using the same methodology as our primary commercial potential models. In this context, we employ academic citations as indicators for scientific potential. The classification variable for these models is the number of academic citations a paper receives. To ensure a balanced dataset, we have defined the median number of citations in the training sample as the threshold for classification. Specifically, papers cited 16 times or fewer are categorized as having *low scientific potential*, whereas those cited more than 16 times are classified as having *high scientific potential*.

The performance of these models is satisfactory, achieving an average accuracy and Area Under the Receiver Operating Characteristic (AUROC) of 0.71. It is important to note that we conducted various experiments with different thresholds and settings. Our priority was maintaining consistency with the primary model's training sample, ensuring a balanced training dataset, and setting a classification threshold meaningfully higher than zero citations. This approach allows the inclusion of papers in the *low* category that still have the potential to yield value.

3.6.2 Social impact potential

In an effort to quantify the social impact potential of scientific findings using a language-based model, we embarked on a novel methodological approach, given the absence of pre-existing data for this purpose. Our process involved several key steps:

1. **Data Sampling for Elicitation and NLP Training:** From our primary dataset, we randomly selected 1,200 papers from 2012 across six fields: Physics, Biology, Computer Science, Electrical Engineering, Materials Science, and Mechanical Engineering. The choice of 2012 papers served dual purposes: it provided a sufficient time frame for patent citations to manifest and reflect a paper's commercial impact and ensured relevance to current social challenges. We balanced this sample across the fields and in terms of commercial outcomes, aiming for equal representation of papers with varying levels of patent citations. This balanced approach was crucial to create a representative sample for effective model training.

2. **Human Elicitation Exercise for Social Impact Measure:** We engaged 10,015 human evaluators in a study where each participant rated three abstracts randomly drawn from our pool of 1,200 papers. The evaluations were based on 11 questions, rated on a Likert scale from 1 to 5, and were designed to assess social impact (details in Table C.1, Appendix C). A custom software application facilitated this data collection. We averaged the results for each question, normalized them to reduce inter-subject variation, and ensured each paper received evaluations from about 30 different individuals.

3. **NLP Model Training Using Human-Based Measures:** We trained an NLP model to leverage the human-derived measure of social impact potential. The model classified findings into high and low social impact potential. The cutoff for classification was set at zero, based on the de-meaned and normalized human-based scores. Papers with a score at or below zero were labeled as having *low social potential*, while those above zero were considered as having *high social potential*.

The model's performance, evaluated using the test set, was impressive, showing high accuracy, precision, recall, and AUROC of 0.86. Additionally, Appendices C.1 and C.2 explore the correlation between our social impact measure and actual commercial outcomes,

offering more profound insights into the model's practical applicability.

4 Results

The results section of our study is divided into two parts. The first part utilizes data from the technology transfer office of a leading research university. This data helps examine the external validity of our commercial measure and analyze the advancement of commercially valuable research through the technology transfer process. Subsequently, we extend our analysis to determine if ideas with high commercial potential, not disclosed to the university's technology transfer office, are utilized by corporations in developing their inventions. Additionally, this part investigates the impact of disclosure to the university and the university's patenting activities on corporate usage.

In the final part of our study, we broaden our analysis to include a substantial sample of research-active universities in the United States. Here, we assess whether research with high commercial potential is ultimately employed in commercial patents. Moreover, we explore whether the frequency of high commercial potential ideas being utilized commercially increases when they are derived from or linked with entities (such as universities and journals) and individuals (notably researchers) with a documented history of commercial impact.

4.1 Commercial Potential and Technology Transfer at a US University

In this section, we apply our commercial potential measure to the experience of the Technology Transfer Office (TTO) of a leading U.S. private university to analyze the commercialization process and validate the measure against outcomes not included in its initial training. Thus, trained initially on academic articles and their citation by renewed patents, our machine learning models are now tested against TTO administrative data.

We will examine the relationships between our measure and key commercialization stages:

disclosure of article-linked inventions to the TTO, TTO investment decisions, licensing, agreements, and revenue generation. Further, the analysis considers additional factors like the invention's scientific and social impact potential and the inventors' experience with commercialization and the TTO.

This approach aims to validate the effectiveness of our measure in real-world settings, offering insights into its broader applicability beyond the initial training scenarios.

4.1.1 What findings do scientists disclose?

Our analysis commences with evaluating the disclosure process to the Technology Transfer Office (TTO), a critical initial step in the technology transfer process. As noted in the theoretical section, it is the scientists who decide to disclose their inventions to the TTO based on factors such as personal motivation, the incentive system of the university, and the effort that may be entailed not only by the disclosure but also by any expected follow-on work associated with the process of technology transfer. We hypothesize that scientists, with accurate assessment and proper incentives, are more likely to disclose inventions with higher commercial potential to the TTO, a hypothesis that our findings support.

Figure 1 illustrates a density plot showcasing the commercial potential of scientific articles, our primary variable of interest. This figure compares the density distributions for all university-associated papers and those linked explicitly to inventions disclosed to the TTO, offering a foundation for the subsequent analyses. Table 4 summarizes the key correlations across these variables of interest.

[Figure 1 about here.]

[Table 4 about here.]

Table 5 shows the relationship between the commercial potential of an article and its likelihood of disclosure to the TTO. We categorize the articles into four groups (quartiles) based on their commercial potential. These groups are: low potential (first quartile), mid-low

potential (second quartile), mid-high potential (third quartile), and high potential (fourth quartile). We then calculate the probability of disclosure for articles within each category. The findings indicate that articles in the low potential category have a 4.62% chance of being disclosed, whereas those in the high potential category have a 24.74% chance, which is 5.35 times greater.

[Table 5 about here.]

To more formally test the relationship between commercial potential and disclosure, we estimate the following linear probability model:

$$disclosure_i = \beta_0 + \beta_1\phi_i + \beta_2\psi_i + \beta_3\omega_i + \beta_4\alpha_{it}^{hs} + \theta_{it} + \epsilon_i, \tag{1}$$

where, for a scientific discovery reported in an article i published in year t, $disclosure_i$ is a binary variable representing whether the article i led to an invention disclosed to the TTO, ϕ_i represents the article's commercial potential, ψ_i its scientific potential, ω_i its social impact potential, and α_{it}^{hs} represents the scientific H-index of the paper's authors at the time t of publication, which we will call scientific prominence. θ_{it} represents a grouped field-year fixed effect to account for technological shocks and trends across the field of the paper i in year t.

Table 6 presents the results of our analysis. Model 1 examines the baseline impact of fixed effects on disclosure rates. Model 2 reveals that the commercial potential of a scientific finding is a strong predictor of its disclosure, with the explained variation beyond the year-field fixed effect increasing from 0.025 to 0.061. Additionally, a one standard deviation increase (0.31) from the median commercial potential score (0.57) corresponds to a 7.38 percentage point rise in disclosure probability.

In Models 3 to 6, we expand our analysis by incorporating additional variables: scientific potential, social impact potential, and the h-index, which assesses the author's scientific experience. Model 5, in particular, reveals that each factor uniquely influences the decision

to disclose an invention to the TTO. Our primary model, Model 6, detailed in equation 1, confirms the significant role of commercial potential in this context. Specifically, an increase of one standard deviation in the commercial potential score correlates with a 6.44 percentage point increase in the disclosure probability. Notably, the coefficient for commercial potential in this more comprehensive model (Model 6) remains similar in magnitude to that in Model 2. This consistency suggests that our commercial potential measure effectively captures a key factor in the disclosure decision-making process.

[Table 6 about here.]

Table 7 presents a detailed analysis of the impact of commercial potential on various later-stage outcomes in the technology transfer process. Note that while disclosure reflects a decision on the part of the scientists, investment in an invention and the decision to patent both reflect decisions on the part of the TTO. In contrast, an agreement and a license reflect decisions on the part of the firms intending to build on the invention. The results are expressed as percentage point increases associated with a one standard deviation change in the commercial potential measure. The data reveals a clear pattern: higher commercial potential correlates with increased likelihood across all stages. A recap from Table 6, the probability of an invention being disclosed to the Technology Transfer Office (TTO) increases by 6.44% (a 46% increase over the baseline).

Furthermore, the likelihood of receiving TTO investment increases by 5.28% (56% increase over baseline), while the chances of obtaining a patent rise by 4.28% (53% increase over baseline). The data also shows a 3.97% increase in the likelihood of reaching an agreement with a firm (47% increase over baseline), a 1.59% increase in the chances of securing a license (37% increase over baseline), and a 0.66% increase in generating revenue (38% increase over baseline of 1.71%). These results collectively indicate that a higher commercial potential of an invention not only influences its initial disclosure but also positively impacts its subsequent progression through the stages of commercialization. Another notable result is that the scientific potential of the article(s) linked to an invention has little relationship

with any decisions made. However, it is related to the realization of revenue. This stands in contrast to the scientific prominence of the faculty inventor(s), which is related to both the TTO's decisions and licensing on the part of firms, though we'll probe this in more detail below. For the TTO and firms, the scientific prominence of a faculty member signals the credibility of the inventor and thus focuses the search.

[Table 7 about here.]

4.1.2 How much information does the commercial potential measure convey post-disclosure?

We examine two critical decisions of the university's TTO that determine whether an invention will be pursued for commercialization. First, we consider the investment an invention receives from the TTO. The nature of this investment varies depending on the invention's field; in some cases, it involves legal protection and licensing costs, while in others, it encompasses marketing expenses. Regardless, the amount invested in commercializing an invention indicates the TTO's belief in its commercial promise. Therefore, we expect inventions based on commercially promising science to receive more significant investment. Second, we observe the number of patents the TTO files for a given invention, another related proxy for the TTO's expectations regarding an invention's value.

In Table 8, we examine the factors associated with the level of TTO investment and the number of patents filed by the TTO, respectively. Figure D.1 in Appendix D provides a visual interpretation. Since both distributions are skewed, we log-transform the variables."

On the right-hand side, in addition to our primary variable of commercial potential, we incorporate whether the authors associated with the invention have prior experience working with the TTO and the interaction of authors' TTO experience with the invention's commercial potential. These variables account for the possibility that TTO managers invest in more experienced teams to reduce investment risk. While it may be the case that the TTO believes that the inventor's experience increases the likelihood of successful commercialization,

it may also be the case that attention to the inventor's experience distracts the TTO from the expected commercial potential of the invention. We also control for the scientific potential and social impact potential of the science associated with the invention and the authors' scientific experience (H-index). The econometric specification is as follows:

$$y_j^{outcome} = \beta_0 + \beta_1\phi_j + \beta_2\alpha_{jt}^{tto} + \beta_3\alpha_{jt}^{tto}\phi_j + \beta_4\psi_j + \beta_5\omega_j + \beta_6\alpha_{jt}^{hs} + \Theta_{jt} + \epsilon_j, \tag{2}$$

where, for an invention j disclosed to the TTO in year t, $y_j^{outcome}$ represents one of the two outcomes—$Log(Investment + 1)$ or $Log(Patents + 1)$. ϕ_j represents the invention's commercial potential, ψ_j its scientific potential, and ω_j its social impact potential. α_{jt}^{tto} represents whether at least one of the authors associated with the invention had previously disclosed a separate invention to the TTO, and α_{jt}^{hs} captures the authors' scientific experience (H-index) at time t. Θ_{jt} represents a grouped field-year fixed effect at the invention level to account for technological shocks and trends across the field of the invention j in year t. Table 8 presents the results. As expected, high-potential ideas attract more investment (Model 1). These results remain robust in the main specification, Equation 2, after controlling for an invention's scientific and social impact potential and the authors' scientific experience (Model 4).

[Table 8 about here.]

4.1.3 What predicts progress through the commercialization process?

The revised analysis of the data, as detailed in Table 9, re-evaluates the investment and patent model initially presented in Table 8. This time, the analysis excludes the interaction term, which previously showed a negligible influence. The findings indicate that projects with high commercial potential are more likely to receive investment and patent protection.

However, when the analysis accounts for investment, the predictive power of commercial, scientific, and social impact potentials on outcomes like agreements, licensing, startup

formation, venture capital (VC) investment, and revenue generation appears significantly diminished. This suggests the initial investment decision may already encapsulate much of the commercial potential's predictive value. Furthermore, the data reveals that an investigator's previous experience with the technology transfer office (TTO) is a strong indicator of success in various stages of commercialization, such as formal agreements with firms, licensing, startup formation, and securing VC funding. If we are to (strongly) assume our measure of commercial potential captures the commercial potential inhering in the scientific finding or idea. In that case, this result suggests that the inventor's commercial experience, and perhaps involvement, may impact successful commercialization over and above the impact of commercial potential per se. Although inventor experience does not significantly impact revenue generation, this may be attributed to limitations in the data set, particularly in accurately capturing revenue and licensing payments to the university.

[Table 9 about here.]

4.1.4 The impact of disclosure and university patenting on corporate use of university science

It is crucial to acknowledge that not all university-derived scientific discoveries are commercialized through the efforts of a technology transfer office (TTO). Often, companies independently search for scientific ideas, bypassing formal technology transfer mechanisms. To investigate the corporate utilization of scientific research from this university and to understand the impact of disclosure on such utilization, we conducted two distinct analyses: one at the article level and the other at the disclosure level.

The results of these analyses are depicted in two tables. Table 10, which examines the article level, reveals that the disclosure of a scientific article markedly increases its likelihood of citation in a corporate patent, indicating enhanced visibility or relevance due to disclosure. Furthermore, at the invention level, as examined in Table 11 and conditioned on disclosure, it is observed that patenting by the Technology Transfer Office (TTO) influences the likelihood

of associated articles being cited in corporate patents. However, the impact of patenting on corporate utilization of this science is considerably less pronounced than disclosures. These trends suggest a significant role for TTO involvement, with the primary effect of commercial potential remaining positive, substantial, and statistically significant. Considering that the proportion of non-disclosed papers (75%) compared to disclosed ones (25%) ranks in the top quartile for commercial potential, it can be inferred that the TTO plays a significant, albeit complementary, role.

However, caution must be exercised in interpreting these results. The presence of countervailing factors, as discussed in our earlier arguments, may be influencing these outcomes. Thus, while these analyses provide insights into the relationship between university research disclosure, TTO activities, and corporate patent citations, they also underscore the complexity of commercializing academic research.

[Table 10 about here.]

[Table 11 about here.]

We should also point out that across tables 10 and 11, we observe yet another result that inspires confidence in our measure of the commercial potential of science–namely that the measure appears to have a strong relationship with firms' use of high potential articles. For example, in Table 10, we observe that the higher the commercial potential of an article, the more likely that a patent will cite it, and the results are robust after controlling for the variables specified in the previous models, the scientific potential, the social potential, and the authors' scientific and commercial experience. Figure 2 provides an intuitive representation of the findings.

[Figure 2 about here.]

4.2 The commercial potential and realization of U.S. scientific research

In this section, we expand our analysis to encompass various research-focused universities in the United States, examining publications from 2000 to 2020. Our investigation centers on determining if research possessing significant commercial potential is eventually incorporated into commercial patents and whether this likelihood is heightened when associated with entities and individuals with a history of commercial and scientific influence, including universities, journals, and scientists. We specifically study whether institutions and individuals with a record of substantial commercial or scientific impact, as indicated by a high H-index (calculated for patent and academic citations at the university, journal, and author levels), are more likely to be cited by patents. Additionally, we study whether higher commercial potential is associated with higher rates of patent renewal among articles cited by at least one patent.

Our dataset for this analysis consists of 5,211,133 articles spanning eleven academic fields: Agricultural, Veterinary, and Food Sciences; Biological Sciences; Biomedical and Clinical Sciences; Chemical Sciences; Earth Sciences; Engineering; Environmental Sciences; Health Sciences; Information and Computing Sciences; Mathematical Sciences; and Physical Sciences. In our analysis, we specifically focus on articles authored by researchers affiliated with leading U.S. research institutions actively commercializing their research. To determine these top research institutions, we adhere to the 'R1: Doctoral Universities – Very high research activity' category from the Carnegie Classification of Institutions of Higher Education as of 2021. Furthermore, to identify commercially active research institutions, we rely on their membership in the Association of University Technology Managers (AUTM), which stipulates a minimum of 0.5 full-time equivalents (FTE) staff dedicated to technology commercialization. This criterion yields a list of 126 U.S. universities for our analysis. Our results are robust to other approaches for defining our sample of institutions, including all universities with at least .5 FTEs for licensing that are also members of AUTM.

To assess the commercial prospects of these publications, we employ a Natural Language Processing (NLP) model that we have developed. This model analyzes the abstracts of the articles to calculate indicators of their scientific and social potential. The preprocessing steps applied to the abstracts before their analysis with our NLP model are outlined in section 3.2.

Table 12 summarizes the statistical characteristics of the sample under study, while Table 13 offers insights into the correlations between the key variables of interest in this research. This analytical approach aims to shed light on the complex interplay between the inherent potential of scientific research and its practical, commercial realization.

[Table 12 about here.]

[Table 13 about here.]

First, we begin with a descriptive exercise to investigate whether the commercial realization rate of articles (the likelihood of an article published at that university being cited by patents) varies among institutions. Figure 3 (Panel A) illustrates the differences in commercialization rates between "bottom ranked institutions" (those that produce less research) and "top-ranked institutions" (those that produce more) during the period 2000-2015.[16] In this figure, a publication's citation in a patent serves as a comparable indicator of its commercial "use". Overall, 18.88% of articles from top institutions get cited by at least one patent, whereas only 13.46% of articles from bottom institutions do. Notably, there is a marked difference (40.5% increase) in the commercialization rate when comparing bottom to top institutions. The factors leading to this difference might be due to various reasons, including the supply and demand-level frictions discussed earlier, the composition of specific research areas that these institutions focus on, and the varying tendencies to secure patents in those areas. Furthermore, field-specific analysis indicates that disparities in commercialization

[16]Patents tend to cite papers that were published, on average, 14 years before a patent is granted (Marx and Fuegi, 2022). While we have sufficient variation in patent citations before 2015, articles published after 2015 accumulate few patent citations, and, thus, we do not consider these for this descriptive exercise.

rates among institutions vary considerably across both institutions and fields—ranging from 14% to 65%, with some fields having significantly larger realization gaps than others when viewed across institutions.

[Figure 3 about here.]

However, Panel A encompasses publications with low and high commercial potential, which may obscure significant differences across institutions. In particular, there are crucial differences in (a) the quantity of high commercial potential science produced by institutions and (b) the rate at which this subset of publications translates into commercial use. As a result, the analysis in Figure 3 (Panel B) is refined by considering only those papers identified as possessing high commercial potential (i.e., those in the top decile of commercial potential scores). This allows us to focus on the relevant risk-set across institutions: those publications most likely to be commercialized ex-ante. This panel illustrates the percentage of articles cited by at least one patent out of all articles deemed to have high commercial potential. We find three primary results. First, unsurprisingly, when the focus narrows to articles with high commercial potential, the patent citation rates for these articles surge. Collectively, across all fields, 43.02% of articles from top institutions and 34.97% from bottom institutions reach commercialization. Second, there's a substantial reduction in the commercialization rate gap between top and bottom institutions when we restrict our attention to only those articles with high commercial potential. Specifically, the relative differential shrinks to 23% from the earlier 40.5%. Finally, despite the high commercial potential classification, a significant 57% of the research, even at more prominent and research-intensive universities, does not transition to commercial use.

Figures 4 and 5 illustrate the considerable variation in both the relative and absolute volume of high commercial potential science across institutions. While some universities, such as the Massachusetts Institute of Technology, have nearly 16% of articles classified as high commercial potential, other institutions, also with active technology transfer offices,

have far fewer: the University of Hawaii has approximately 2% whereas the University of Arizona has 6%.

[Figure 4 about here.]

[Figure 5 about here.]

To conduct a more formal analysis, we estimate the following specification:

$$y_i = \beta_0 + \beta_1\phi_i + \beta_2\psi_i + \beta_3\omega_i + \beta_4\iota_i^{c80} + \beta_5\phi_i \times \iota_i^{c70} + \beta_6\iota_i^{s70} + \beta_7\phi_i \times \iota_i^{s70} + \theta_{jt} + \epsilon_j, \quad (3)$$

where, for a scientific finding i published in year t, y_i represents whether a paper is cited by at least one patent, ϕ_j represents the finding's commercial potential, ψ_j its scientific potential, and ω_j its social potential. ι_{it}^{c80} represents whether the paper i's institution ι at time t is in the top 70th percentile in terms of its commercial H-index and ι_{it}^{s80} in terms of scientific H-index (see Table 2 for a description of how institution H-indices are computed at the paper level). θ_{it} represents a grouped field-year fixed effect at the paper level to account for technological shocks and trends across the field of the paper i in year t. Table 14 presents the results of this analysis. The analysis corroborates initial observations, revealing a clear trend: research-intensive institutions are more likely to successfully commercialize research with high commercial potential than their less research-focused counterparts. This is evidenced by the positive and statistically significant interaction term 'Commercial potential x High commercial impact institution,' which remains robust even when accounting for fixed effects across years, academic fields, and universities. This result supports the idea that a possible realization gap exists in commercialization outcomes between different types of institutions.

To delve deeper into the causes of this disparity, the study explores two potential barriers that might prevent research with high commercial potential from being realized. These barriers are related to the characteristics of individual researchers and the journals in which their

findings are published. Research conducted by less prominent or newer scholars and studies published in journals with lower commercial visibility might be overlooked by commercial entities, thereby limiting their practical application.

Model 3 in Table 14 presents outcomes that underscore ongoing institutional disparities in commercialization rates. Research with high commercial viability originating from prestigious institutions is more likely to be cited in patents than similar research from less renowned institutions. This holds even when controlling for specific characteristics associated with researchers and the journals in which their work is published. This finding suggests that institutional reputation and visibility play significant roles in the commercialization process, influencing the likelihood of research being recognized and utilized in commercial patents even while controlling for commercial potential. This result suggests that a higher fraction of commercialized research is more likely to be overlooked at those institutions and from those individuals and journals with less of a history of commercialization. This also suggests that search heuristics firms employ to find commercially promising research may lead to unrealized potential. On the other hand, given the high search costs associated with firms' efforts to find promising research, applying such heuristics may be perfectly rational.

[Table 14 about here.]

Table 15 extends the analysis of commercial potential and value by examining not just patent citations but also the longevity and, by extension, the value of these patents. The focus here is on papers cited by at least one patent. The critical variable of interest is the average number of years the patents citing a particular paper have been renewed. Patent renewal years—zero, four, eight, or twelve—serve as a proxy for the patent's value. This approach is premised on the notion that the longer a patent is maintained, the more valuable it will likely be.

Given the heterogeneity of patent values–even the values of renewed patents–we now calculate the average number of years that the patents citing a given paper have been renewed.

Using such a measure provides a more nuanced understanding of the commercial value of academic research, as seen through the lens of patent renewals.

The findings from this approach are consistent with previous analyses. They indicate that papers from top-tier institutions are more likely to be cited by patents and tend to be cited by patents renewed for extended periods. This implies that these papers are associated with patents of higher commercial value. The implication is that the originating institution's prestige or research intensity plays a significant role not just in the initial commercialization (as indicated by patent citations) but also in the sustained commercial viability and value of the research, as reflected in the longevity of patent renewals.

[Table 15 about here.]

5 Discussion

The goals of our study were: 1.) to develop an *ex ante* measure of the commercial potential of academic scientific contributions, and 2.) to apply that measure to identify factors conditioning the realization of that potential. The premise of our study is that to understand the determinants of the commercial realization of academic science, a useful first step is to identify the "at-risk set"—that is, those articles, conference proceedings, etc., that offer commercial potential.

Our first step was to develop a text-based, *ex ante* measure of commercial potential for scientific articles developed with a large language model employing machine learning. We then applied this measure in two empirical analyses to examine the commercialization of academic science. The first was at the micro level, within one leading research university, of the process of scientific ideas and findings progressing from publication to commercialization via the university's technology transfer office. The second was conducted at the institutional level, examining features of American research universities conditioning the commercialization of science.

In the first study, we examined the commercialization of over 96,000 articles from a well-known research university. Our findings showed that our method effectively predicted which scientific discoveries were disclosed, the investments made by a university's Technology Transfer Office (TTO), patent applications, startup formations, and licensing activities. However, conditional on disclosure and investment, the critical factor that drove commercial success for a given invention was whether someone on the inventing team had prior experience commercializing their science. If we assume that our measure correctly identified the commercial potential inhering to a scientific study, this finding suggests, consistent with the literature, that such potential may not be sufficient to ensure successful commercialization; that the experience of the individual(s) performing that study also plays an important role in the commercialization process.

A significant observation in this study is that a primary barrier to commercialization often arises from scientists not disclosing their discoveries. This nondisclosure may originate from the scientists' personal preferences (Cohen et al., 2020; Azoulay et al., 2007), their unawareness of the commercial implications (Hsu and Kuhn, 2023), the Technology Transfer Office's (TTO) inadequate recognition of relevant research (Thursby and Thursby, 2002; Debackere and Veugelers, 2005) or insufficient resources. Consequently, numerous publications with substantial commercial promise remain undisclosed, contributing to a "realization gap." This failure to disclose results in these high commercial potential scientific works less frequently referenced in corporate patents. However, it is noteworthy that articles with high commercial potential are still nine times more likely to be cited by corporate patents compared to articles with lower commercial potential that have been disclosed. From this study, we also learn that while the TTO, in this case, plays a vital role in advancing the commercialization of the university's science, an even more critical role appears to be that of publication—an academic's act of publicly disclosing their findings and ideas. This finding is not surprising in light of the apparent reluctance, noted above, of some faculty to disclose inventions to the TTO and the vastly more significant quantity of publications produced in

the university relative to the number of publications that contribute to invention disclosures.

Our second study extended our research to encompass over 5.2 million articles published in the United States from 2000 to 2020. This analysis, paralleling our examination of micro-data from the Technology Transfer Office of a prominent US university, revealed that articles with significant commercial potential were more frequently incorporated in patents, irrespective of the originating institution's prestige. However, we observed that high-potential discoveries from institutions with a legacy of substantial commercial impact exhibited approximately 14.1% higher commercialization rates than institutions with less commercial influence controlling for commercial potential. This implies that the commercial potential of science from universities with less of a legacy of commercial impact is unrealized at higher rates. Similarly, the commercial adoption of university science was more pronounced for research published in journals known for their regular contribution to commercially impactful work and by authors with a history of generating commercially valuable research, as opposed to equivalent research from authors and journals lacking such a track record. The implication is the same: Science with commercial potential is more likely to be overlooked without a track record at the individual level or even for journals with less of a history of commercial impact. Perhaps the most significant finding from this analysis was that what accounts for the preponderance of differences in commercialization rates across universities are not factors potentially impeding the identification of commercially promising science by either the university's TTOs or firms but factors accounting for the differences in the actual production of such science.

These findings offer important insights into the process of commercializing scientific research. They highlight the role of institutional factors and the decisions of scientists and TTOs in realizing the commercial potential of scientific research. This research contributes to a better understanding of how scientific advancements move from academic settings to practical, commercial applications, pointing out the complexities and potential areas for improvement in the commercialization process.

In this article, we developed, validated, and examined correlates of a text-based measure designed to assess the commercial potential of academic science before its realization. We see numerous applications for our measures in empirical studies concerning the 'science of science' and the economics of innovation. A major challenge in exploring various topics in the economics of innovation involves the potential unobserved heterogeneity in the commercial potential of scientific research (Marx and Hsu, 2022), which may correlate with key variables of interest, such as gender (Koffi and Marx, 2023; Ding et al., 2006) or status (Azoulay et al., 2010). Our measure of the commercial potential of science can serve as a control variable, either through direct incorporation in regression analyses, in matching estimators, or as part of an instrumental variables strategy to address these differences. Utilizing our measure in this manner would enable a more thorough analysis of questions that are challenging to address without adequately controlling for the commercial potential of science. Beyond the scope of econometric considerations, our metric may prove beneficial in benchmarking the commercial potential and variations in actualization across different domains, institutions, researchers, and regions within a country (for example, comparing the San Francisco Bay Area with Kansas City) as well as internationally (such as the United States versus the United Kingdom). Furthermore, this metric could facilitate the examination of the impediments in the commercial application of scientific knowledge—possessing inherent commercial potential—and the identification of factors linked to generating scientific ideas with immediate commercial applicability. While some aspects of this have been addressed in the current paper, a more comprehensive understanding of the ecosystem fostering and translating commercially viable ideas warrants further investigation.

Our work has, however, limitations. The reliance on patent data and the assumption that patent-to-paper citations reflect a scientific contribution's commercial potential may certainly be questioned, notwithstanding the embrace of this assumption in the literature.(Kuhn et al., 2020). For example, numerous scientific contributions make their way to the market with no associated patent. Additionally, the current model may only partially capture the

commercial potential of scientific contributions given variable and sometimes indirect paths to commercialization, no less very long time horizons before a given contribution may be embodied in a new product or process (cf., Adams, 1990; IIT, 1968). Moreover, our NLP-based matching technique may overlook factors beyond textual content that influence the decision to build technology upon a specific piece of science, and we do not account for the nature of such errors in our predictions.

Future research should focus on refining the measure of commercial potential, exploring factors limiting commercialization in low-impact institutions, and identifying practices that effectively bridge the gap between science and industry. This will deepen our understanding of the commercialization process and enhance the societal benefits derived from scientific research.

References

Adams, J. D. (1990). Fundamental stocks of knowledge and productivity growth. *Journal of political economy*, 98(4):673–702.

Arora, A., Cohen, W. M., and Walsh, J. P. (2016). The acquisition and commercialization of invention in american manufacturing: Incidence and impact. *Research Policy*, 45(6):1113–1128.

Azoulay, P., Ding, W., and Stuart, T. (2007). The determinants of faculty patenting behavior: Demographics or opportunities? *Journal of economic behavior & organization*, 63(4):599–623.

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

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bikard, M. (2018). Made in academia: The effect of institutional origin on inventors' attention to science. *Organization Science*, 29(5):818–836.

Bikard, M. and Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, 66(8):3425–3443.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Cohen, W. M. and Levinthal, D. A. (1989). Innovation and learning: the two faces of r & d. *The economic journal*, 99(397):569–596.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2002). Links and impacts: the influence of public research on industrial r&d. *Management science*, 48(1):1–23.

Cohen, W. M., Sauermann, H., and Stephan, P. (2020). Not in the job description: The commercial activities of academic scientists and engineers. *Management Science*, 66(9):4108–4117.

Cornelli, F. and Schankerman, M. (1999). Patent renewals and r&d incentives. *The RAND Journal of Economics*, pages 197–213.

Debackere, K. and Veugelers, R. (2005). The role of academic technology transfer organizations in improving industry science links. *Research policy*, 34(3):321–342.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, W. W., Murray, F., and Stuart, T. E. (2006). Gender differences in patenting in the academic life sciences. *science*, 313(5787):665–667.

Fleming, L. and Sorenson, O. (2004). Science as a map in technological search. *Strategic management journal*, 25(8-9):909–928.

Guzman, J. and Li, A. (2023). Measuring founding strategy. *Management Science*, 69(1):101–118.

Henderson, R., Jaffe, A. B., and Trajtenberg, M. (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. *Review of Economics and statistics*, 80(1):119–127.

Hsu, D. H. and Kuhn, J. M. (2023). Academic stars and licensing experience in university technology commercialization. *Strategic Management Journal*, 44(3):887–905.

IIT (1968). *Technology in retrospect and critical events in science*. Illinois Institute of Technology.

Jackson, L. A., Anderson, E. J., Rouphael, N. G., Roberts, P. C., Makhene, M., Coler, R. N., McCullough, M. P., Chappell, J. D., Denison, M. R., Stevens, L. J., et al. (2020). An mrna vaccine against sars-cov-2—preliminary report. *New England journal of medicine*, 383(20):1920–1931.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, 108(3):577–598.

Koffi, M. and Marx, M. (2023). Cassatts in the attic. Technical report, National Bureau of Economic Research.

Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., and Lakhani, K. R. (2022). Conservatism gets funded? a field experiment on the role of negative information in novel project evaluation. *Management science*, 68(6):4478–4495.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the nih. *American Economic Journal: Applied Economics*, 9(2):60–92.

Liang, W., Elrod, S., McFarland, D. A., and Zou, J. (2022). Systematic analysis of 50 years of stanford university technology transfer and commercialization. *Patterns*, 3(9):100584.

Manjunath, A., Li, H., Song, S., Zhang, Z., Liu, S., Kahrobai, N., Gowda, A., Seffens, A., Zou, J., and Kumar, I. (2021). Comprehensive analysis of 2.4 million patent-to-research citations maps the biomedical innovation and translation landscape. *Nature Biotechnology*, 39(6):678–683.

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

Marx, M. and Fuegi, A. (2022). Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392.

Marx, M. and Hsu, D. H. (2022). Revisiting the entrepreneurial commercialization of academic science: Evidence from "twin" discoveries. *Management Science*, 68(2):1330–1352.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Nakamura, S., Mukai, T., and Senoh, M. (1994). Candela-class high-brightness ingan/algan double-heterostructure blue-light-emitting diodes. *Applied Physics Letters*, 64(13):1687–1689.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bring order to the web. Technical report, technical report, Stanford University.

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institution Press.

Thursby, J. G. and Thursby, M. C. (2002). Who is selling the ivory tower? sources of growth in university licensing. *Management science*, 48(1):90–104.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Williams, H. L. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*, 121(1):1–27.

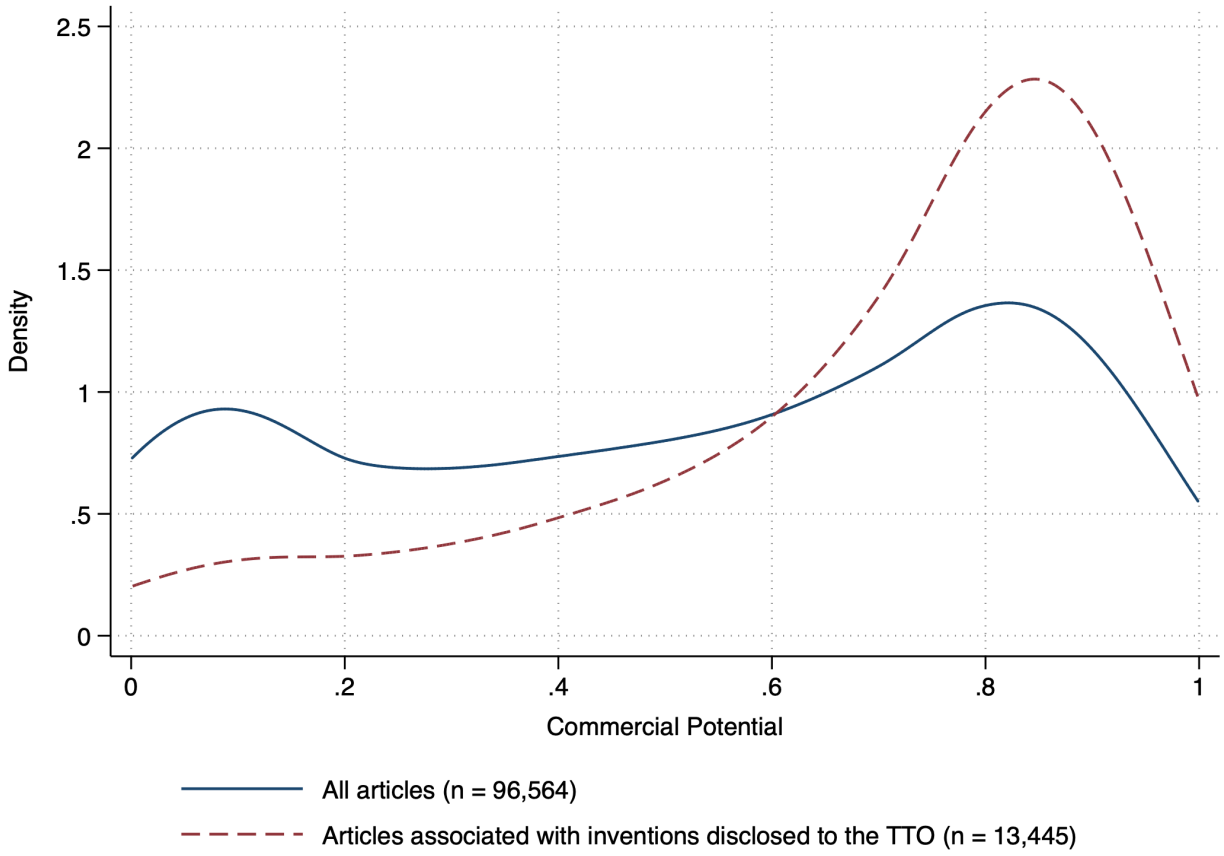


Figure 1: Bi-weight kernel density estimates of the distributions of commercial potential of 1) all articles published at this university (solid line) and 2) only articles associated with inventions disclosed to the Technology Transfer Office (dashed line). Articles tied to an invention are more likely to have high commercial potential.

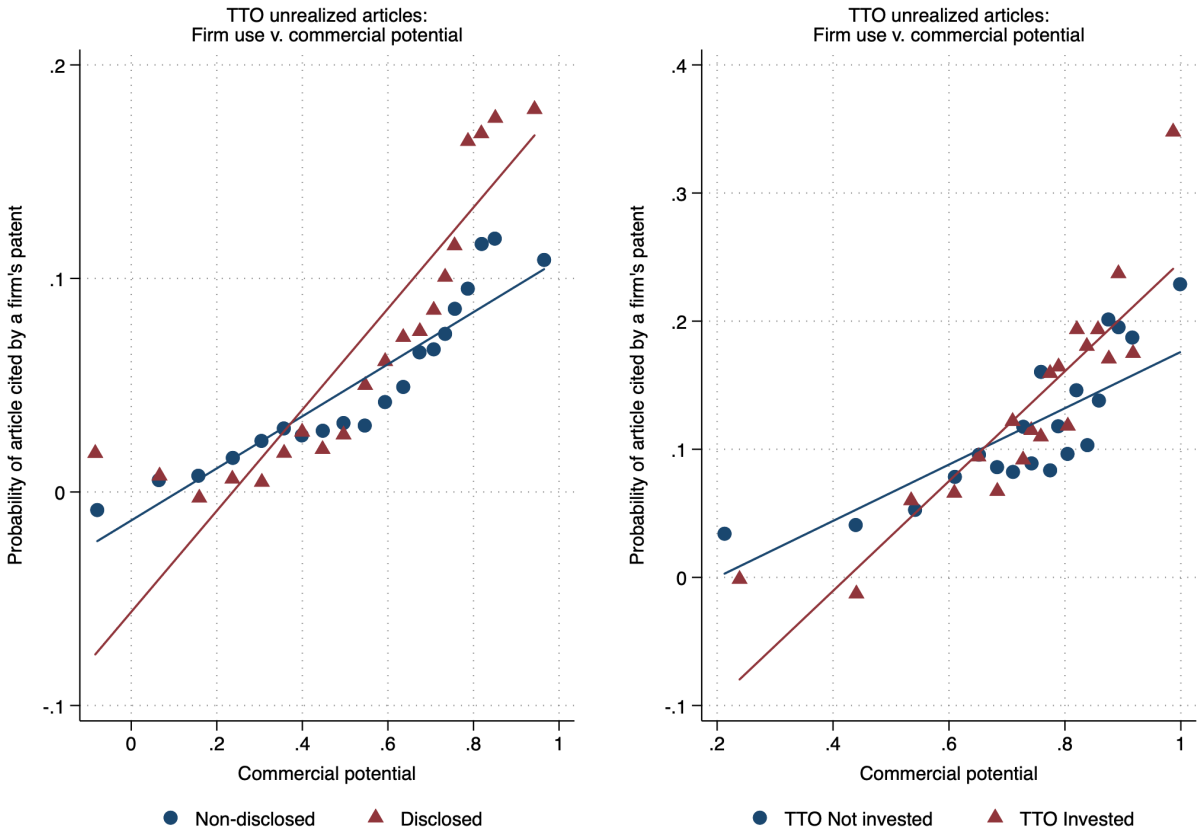
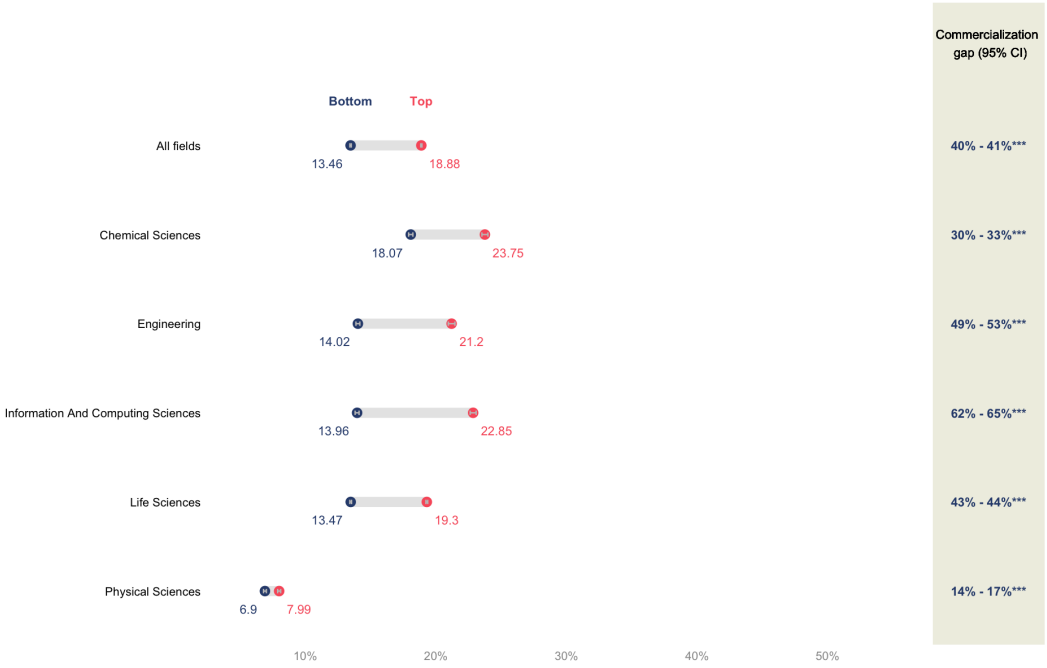
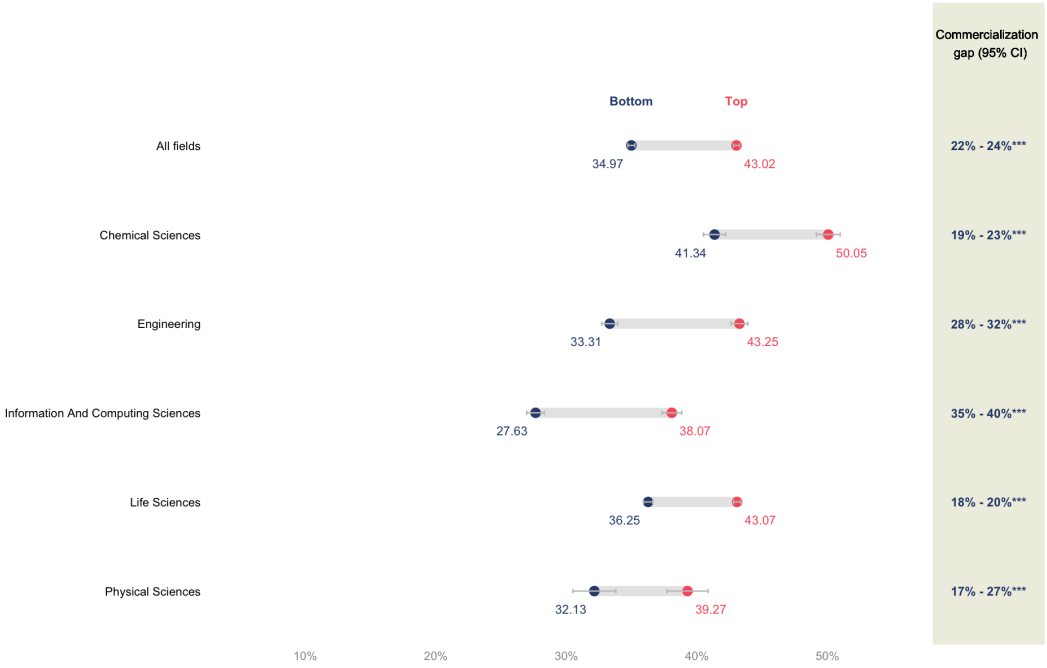


Figure 2: Binned scatterplot representing the probability that a firm's patent will cite articles associated with the TTO university as a function of their commercial potential. Panel A distinguishes articles tied to an invention disclosed to the TTO (triangle) vs. those not (circle). Panel B distinguishes articles whose inventions are invested by the TTO (triangle) vs. articles whose inventions are not (circle). The OLS regression absorbs field-year fixed effects. High commercial potential articles associated with inventions disclosed to the TTO are more likely to be cited by a firm's patent.



(a) Panel A: All articles



(b) Panel B: Articles with high commercial potential (top decile)

Figure 3: Differences in commercialization rates across institutions. Panel A plots the average rates at which publications are cited by at least one patent in top vs. bottom institutions for all fields and by field. Panel B conditions on articles classified as high commercial potential; that is, the share is computed as the number of articles that are cited by a patent over the total number of articles with commercial potential. Institutions are classified in quintiles according to the volume of research produced, and we plot the differences for top quintile vs. bottom quintile.

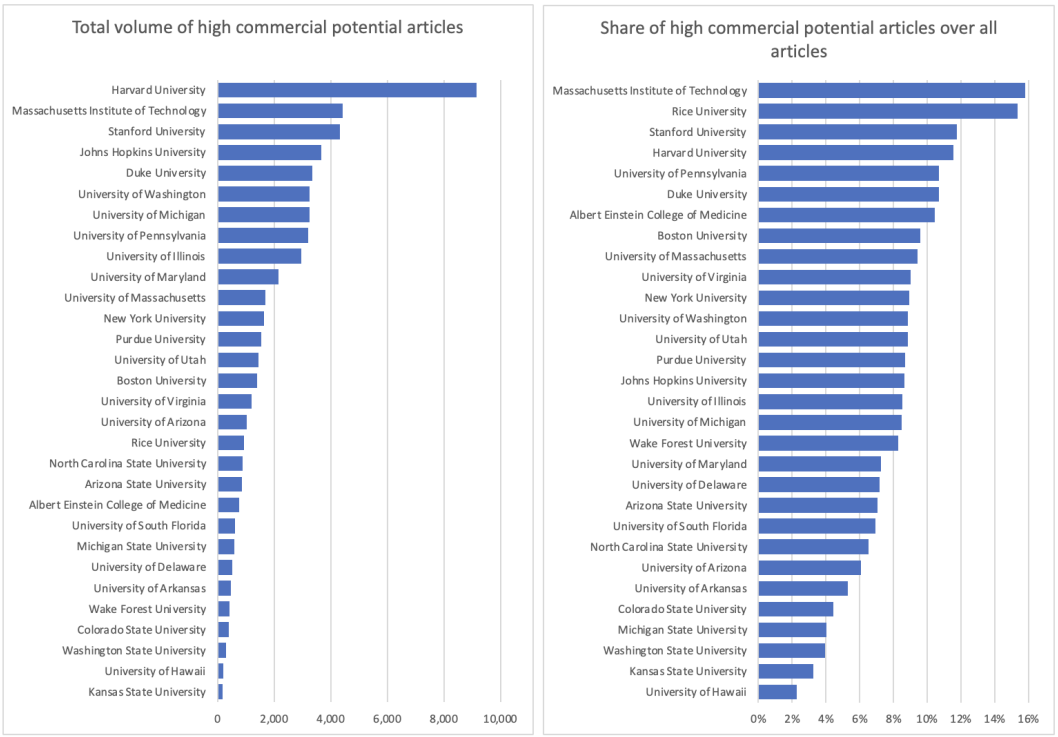


Figure 4: Differences in the production of high commercial potential research. Panel A shows the total number of high commercial potential articles produced between 2010 and 2014 for selected universities. Panel B shows the share of high commercial potential articles over the total number of articles produced.

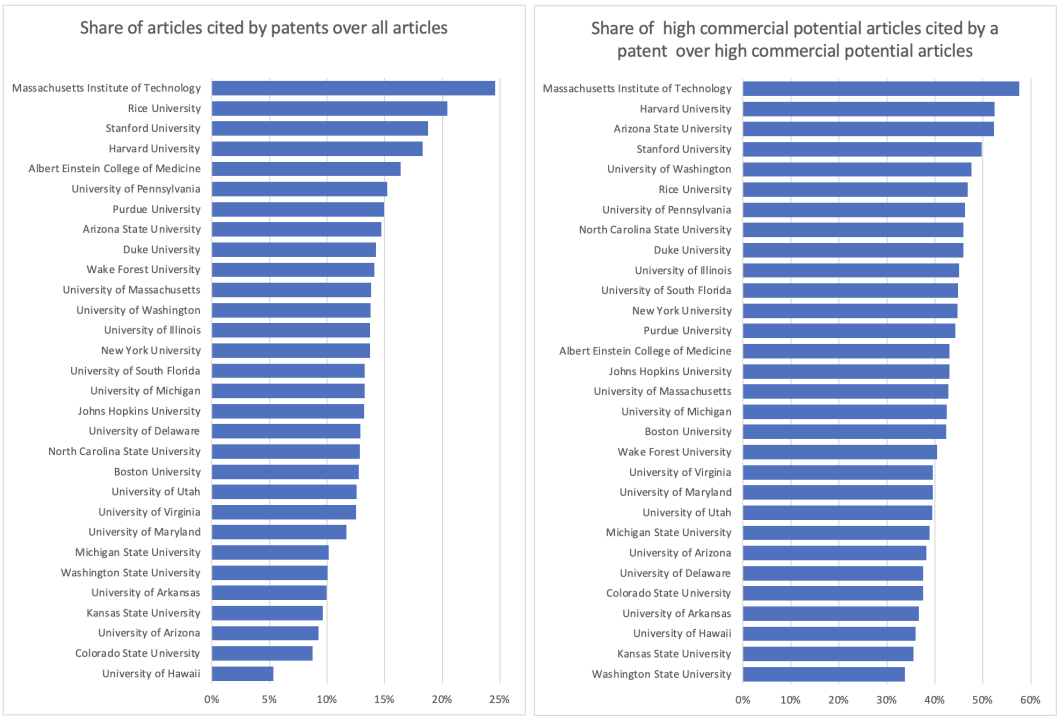


Figure 5: Share of articles cited by patents over all articles (Panel A) *vs.* share of articles cited by a patent over high commercial potential articles (Panel B).

Table 1: TTO variable descriptions.

Variable	**Description of measure**
Disclosed	Binary variable representing whether an article is tied to an invention disclosed to the TTO.
Investment	Amount invested ($) by the TTO to pursue the commercialization of an invention. Includes different natures of expenses, such as patenting and marketing expenses. We use a log-transformation.
Patents	Number of patents the TTO filled for a given invention. We use a log-transformation.
Agreements	Number of commercial agreements associated with an invention. We use a log-transformation.
Licenses	For each invention, number of licensing agreements with third parties, such as firms or other institutions. We use a log-transformation.
Any Revenue	Binary variable representing whether the invention generated a positive revenue.
Startup	Binary variable indicating whether the invention has been commercialized via Startup
VC Investment	Conditional on Startup, binary variable indicating whether the startup has venture capital financing
Author TTO Experience	Binary variable representing whether an author, prior to the focal disclosure, has disclosed an invention to the TTO.

Table 2: Generic variable descriptions.

Variable	Measure	Description of measure
Commercial potential	P(Patent renewal | patent cite > 1)	Probability that the focal article will be cited by at least one patent that, in turn, will be renewed. The probability is the output of our primary NLP model, which uses the abstract text of the focal article to cast the prediction.
Scientific value	Paper cite	Number of academic articles that cite the focal article.
Scientific potential	P(Paper cite $>$ 16)	Probability that more than 16 academic articles will cite the focal article. The probability is the output of our secondary NLP model (Scientific potential), which uses the abstract text of the focal article to cast the prediction.
Social impact	Social measure-elicitation	Social value of an article per our human-based elicitation exercise (see section 3.6.2). The measure captures the social value a scientific finding has based on several criteria lay people evaluate. Each article is evaluated across 11 dimensions by an average of 30 people, and the evaluations are first de-meaned at an evaluator level, then averaged, and finally normalized.
Social impact potential	P(Social value > 0)	Probability that the focal article will have a social value greater than zero. The probability is the output of our secondary NLP model (Social potential), which uses the abstract text of the focal article to cast the prediction.
Commercial use	Patent-to-article cite	Number of patents that cite the focal article. We use patent front-page text and patent body text to identify article citations. We rely on Marx and Fuegi (2020) as a source of patent-to-article citations.
Average Commercial value	Average (years patent renewed | patent-to-article cite)	Number of years patents citing the focal article are renewed. Patents can never be renewed (0 years), renewed once (at four years), twice (at eight years), and three times (at 12 years). Since an article can be cited by more than one patent, the measure is the average of years the citing patents are renewed.
Author scientific experience	Author scientific h-index	Author h-index at the time of publication, excluding the focal article. The h-index captures the productivity and impact of an author and is calculated by counting the number of publications for which an author has been cited by other authors at least that same number of times. Formally, the h-index can be defined as $h_{\text{index}} = \max\{i \in N : g(i) \geq i\}$, where $g(i)$ represents the number of cites of the paper with index i.
Author commercial experience	Author commercial h-index	Author commercial h-index is computed similarly to the scientific h-index, but using the cites by patents instead of the cites by other academic articles.
Institution scientific experience	Institution h-index	Institution h-index is computed as the author scientific h-index, but we use the institution as the focus of analysis and, thus, the papers affiliated with an institution. Because a paper can be associated with more than one institution via its authors, the institution h-index of a given paper is the sum of the h-indexs of each of the institutions associated with the paper.
Institution commercial experience	Institution commercial h-index	Idem as author commercial h-index, but using institutions as the focus of analysis.
Journal impact factor	Journal impact factor	For every year, the average number of citations of articles published in the last two years in the focal journal (source: Marx and Fuegi (2020)).
Journal commercial impact factor	Journal commercial impact factor	For every year, the average number of patent citations to articles published in the last two years in the focal journal (source: Marx and Fuegi (2020)).

Table 3: TTO summary statistics. Panel A summarizes the relevant features for articles whose authors were affiliated with the TTO's university at the time of publication, 2000-2020. Panel B summarizes the relevant features of the articles *associated* with disclosed inventions, 2000-2020. For confidentiality reasons, invention-level outcomes are removed (Investment, Patents, Agreements, Licensing, Revenue, Startup, and VC funding).

	Mean	SD	Min	Max	N
Commercial Potential	0.52	0.31	0.00	1.00	96,564
Scientific Potential	0.73	0.20	0.00	1.00	96,564
Social Impact Potential	0.79	0.38	0.00	1.00	96,564
Academic cites	62.54	210.09	0.00	35,395.00	96,564
Patent cites	0.71	5.77	0.00	503.00	96,564
Cited by at least one patent	0.11	0.31	0.00	1.00	96,564
Citing patent is renewed	0.08	0.27	0.00	1.00	96,564
Author Scientific Experience	45.26	31.04	1.00	276.00	96,564
Disclosed	0.14	0.35	0.00	1.00	96,564

	Mean	SD	Min	Max	N
Commercial Potential	0.73	0.21	0.00	1.00	2,728
Scientific Potential	0.76	0.15	0.01	1.00	2,728
Social Impact Potential	0.84	0.29	0.00	1.00	2,728
Academic cites	74.95	140.32	0.00	2,587.14	2,728
Patent cites	2.41	9.38	0.00	212.00	2,728
Cited by patent	0.46	0.50	0.00	1.00	2,728
Cited by renewed patent	0.37	0.48	0.00	1.00	2,728
Author Scientific Experience	49.47	28.76	1.00	201.00	2,728
Author TTO Experience	0.68	0.46	0.00	1.00	2,728

Table 4: TTO main variables correlations. Panel A is based on all articles at the University. Panel B is based on only those articles associated with an invention disclosed to the TTO.

Panel A: All articles

	Commercial Potential	Scientific Potential	Social Impact Potential	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Author Scientific H-index	Disclosed	Author TTO Experience
Commercial Potential	1.000									
Scientific Potential	0.212	1.000								
Social Impact Potential	0.209	0.159	1.000							
Academic cites	0.056	0.062	0.000	1.000						
Patent cites	0.106	0.021	0.022	0.339	1.000					
Cited by at least one patent	0.261	0.042	0.033	0.209	0.354	1.000				
Citing patent is renewed	0.224	0.041	0.022	0.201	0.389	0.843	1.000			
Author Scientific Experience	0.181	0.240	0.163	0.073	0.002	0.009	-0.009	1.000		
Disclosed	0.219	0.049	0.068	0.030	0.070	0.129	0.119	0.072	1.000	
Author TTO Experience	0.210	0.046	0.061	0.027	0.065	0.132	0.121	0.085	0.796	1.000

Panel B: Articles matched to inventions disclosed to the TTO

	Commercial Potential	Scientific Potential	Social Impact Potential	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Author TTO Experience
Commercial Potential	1.000							
Scientific Potential	0.176	1.000						
Social Impact Potential	0.048	0.182	1.000					
Academic cites	0.043	0.053	-0.033	1.000				
Patent cites	0.120	0.006	-0.004	0.352	1.000			
Cited by patent	0.206	-0.049	-0.018	0.251	0.280	1.000		
Cited by renewed patent	0.166	-0.050	-0.021	0.253	0.321	0.830	1.000	
Author Scientific Experience	0.139	0.244	0.055	0.064	-0.053	-0.068	-0.104	1.000
Author TTO Experience	0.225	0.062	0.008	0.034	0.025	0.072	0.069	0.131	1.000

Table 5: Percentage distribution of articles in the TTO university binned in four quartiles of commercial potential. Articles in the top quartile are 5.35 times more likely to be associated with an invention disclosed to the TTO than articles in the bottom quartile.

Commercial Potential Quartile	Not disclosed	Disclosed	Total
1	23,026	1,115	24,141
	95.38%	4.62%	100.00%
2	21,875	2,266	24,141
	90.61%	9.39%	100.00%
3	20,049	4,092	24,141
	83.05%	16.95%	100.00%
4	18,169	5,972	24,141
	75.26%	24.74%	100.00%
Total	83,119	13,445	96,564
	86.08%	13.92%	100.00%

Table 6: Linear probability model estimating the likelihood of disclosure as a function of commercial potential. Model 1 presents the baseline impact of the fixed effects (field-year) on disclosure. Model 2 shows that the measure of commercial potential predicts whether a scientific publication will be associated with a disclosure well above the fixed effects. Models 3, 4, and 5 report the results controlling for the scientific and the social impact potential measures. Model 6 presents the full specification, also controlling for the scientific experience of a publication's authors at the time of publication (logged h-index). Fixed effects are included at a publication field-year level.

	(1) Disclosed	(2) Disclosed	(3) Disclosed	(4) Disclosed	(5) Disclosed	(6) Disclosed
Commercial Potential		0.238***			0.218***	0.206***
		(0.016)			(0.016)	(0.016)
Scientific Potential			0.140***		0.028***	0.006
			(0.012)		(0.009)	(0.008)
Social Impact Potential				0.105***	0.050***	0.051***
				(0.007)	(0.006)	(0.006)
Author Scientific Experience						0.028***
						(0.004)
Constant	0.139***	0.015*	0.037***	0.056***	-0.035***	-0.112***
	(0.000)	(0.008)	(0.009)	(0.005)	(0.008)	(0.014)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564	96,564
R-squared	0.025	0.061	0.029	0.034	0.063	0.066

Standard errors clustered at the Publication Category - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 7: Linear probability model estimating the likelihood of disclosure, TTO expenses on an invention, TTO patent applications, signed agreements, licenses to firms, and revenue generated by the technology associated with the disclosure. Fixed effects are included at a publication field-year level.

	(1) Disclosed	(2) Investment	(3) Patent	(4) Agreement	(5) License	(6) Revenue
Commercial Potential	0.206***	0.169***	0.137***	0.127***	0.051***	0.021***
	(0.016)	(0.013)	(0.012)	(0.010)	(0.005)	(0.002)
Scientific Potential	0.006	-0.006	-0.002	0.009	0.008	0.012***
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)	(0.003)
Social Impact Potential	0.051***	0.037***	0.031***	0.032***	0.020***	0.007***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.003)	(0.002)
Author Scientific Experience	0.028***	0.026***	0.019***	0.026***	0.014***	0.002**
	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
Constant	-0.112***	-0.113***	-0.083***	-0.107***	-0.057***	-0.017***
	(0.014)	(0.012)	(0.010)	(0.011)	(0.008)	(0.004)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564	96,564
R-squared	0.066	0.059	0.049	0.055	0.027	0.016

Standard errors clustered at the Publication field - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 8: OLS regressions estimating the impact of commercial potential on actual investment by the TTO. Models 1 to 4 use as a dependent variable the amount (log-transformed) of dollars the TTO invests in an invention to commercialize it. Models 5 to 8 use the number of patents (log-transformed) the TTO fills related to an invention. The average commercial potential of the articles tied to an invention strongly predicts both variables. Models 2 to 4 and 6 to 8 successively add variables to control for the author's previous experience disclosing inventions to the TTO and for the scientific and social potential of the focal article as well as for the authors' scientific experience. Fixed effects are included at an invention field-year level.

	(1) Investment	(2) Investment	(3) Investment	(4) Investment	(5) Patents	(6) Patents	(7) Patents	(8) Patents
Commercial Potential	3.317***		3.355***	2.543***	0.642***		0.597***	0.488***
	(0.444)		(0.706)	(0.703)	(0.081)		(0.112)	(0.115)
Author TTO Experience		0.906***	1.140	0.880		0.192***	0.181	0.130
		(0.202)	(0.725)	(0.716)		(0.034)	(0.113)	(0.112)
Comm. Pot. x TTO Experience			-0.649	-0.387			-0.047	0.005
			(0.973)	(0.964)			(0.159)	(0.157)
Author Scientific Experience				0.485***				0.109***
				(0.144)				(0.027)
Scientific Potential				3.214***				0.361***
				(0.782)				(0.126)
Social Impact Potential				-0.257				-0.140**
				(0.346)				(0.066)
Constant	2.177***	3.991***	1.706***	-1.725**	0.239***	0.579***	0.172**	-0.310**
	(0.327)	(0.139)	(0.495)	(0.861)	(0.060)	(0.023)	(0.079)	(0.150)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,689	2,689	2,689	2,689	2,689	2,689	2,689	2,689
R-squared	0.121	0.114	0.125	0.135	0.125	0.117	0.131	0.140

Standard errors clustered at the Invention Category - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 9: Models estimating which inventions are more likely to progress through the commercialization process. Models 1 and 2, replicate analysis from Table 8. Models 3-6 are estimated for only those disclosures that received non-zero investments from the Technology Transfer Office.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Invested	Patent Filed	Agreement	Licensed	Startup	VC Investment	Any revenue
Commercial Potential	0.199***	0.231***	0.120	-0.044	-0.006	0.004	-0.119
	(0.050)	(0.055)	(0.111)	(0.111)	(0.077)	(0.061)	(0.079)
Scientific Potential	0.316***	0.201**	-0.153	-0.209	0.053	0.061	-0.091
	(0.081)	(0.084)	(0.117)	(0.138)	(0.104)	(0.088)	(0.106)
Social Impact Potential	-0.002	-0.052	0.011	-0.016	0.022	0.032	-0.020
	(0.034)	(0.037)	(0.052)	(0.065)	(0.045)	(0.034)	(0.047)
Author TTO Experience	0.054***	0.061***	0.060**	0.111***	0.053**	0.054***	0.024
	(0.021)	(0.020)	(0.030)	(0.031)	(0.020)	(0.016)	(0.021)
Author Scientific Experience	0.043***	0.063***	0.070**	0.020	-0.002	-0.020	-0.003
	(0.015)	(0.018)	(0.029)	(0.030)	(0.023)	(0.020)	(0.023)
Constant	-0.092	-0.049	0.434***	0.425**	0.070	0.065	0.334***
	(0.078)	(0.085)	(0.147)	(0.163)	(0.113)	(0.089)	(0.121)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Conditional on	Disclosure	Disclosure	Investment	Investment	Investment	Investment	Investment
Observations	2689.000	2689.000	1305.000	1305.000	1305.000	1305.000	1305.000
R-squared	0.126	0.137	0.182	0.132	0.162	0.160	0.172

Standard errors clustered at the Invention Category - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 10: Linear probability model estimating the likelihood that a corporate patent cites an academic article at the TTO university. The higher the commercial potential of an article, the more likely that a firm's patents will cite it (model 1). The models also explore whether a disclosure event leads to more patent citations by firms (model 2) and the interaction between disclosure and commercial potential (models 3 and 4). The models control for an article's scientific and social impact and the author's scientific and commercial experience before publishing the focal article. Model 2 shows that this disclosure alone predicts corporate patent citations, and models 3 and 4 show that the interaction between commercial potential and disclosure is positive and significant. Fixed effects are included at a publication field-year level.

	(1)	(2)	(3)	(4)
	Cited by corporate patent	Cited by corporate patent	Cited by corporate patent	Cited by corporate patent
Commercial Potential	0.140***		0.123***	0.076***
	(0.017)		(0.015)	(0.012)
Disclosed		0.052***	-0.039***	-0.045***
		(0.006)	(0.011)	(0.011)
Commercial Potential x Disclosed			0.104***	0.100***
			(0.019)	(0.019)
Scientific Potential				0.016**
				(0.006)
Social Impact Potential				0.019***
				(0.005)
Author Scientific Experience				-0.014***
				(0.002)
Author Commercial Experience				0.043***
				(0.005)
Constant	-0.018**	0.048***	-0.014*	-0.036***
	(0.009)	(0.001)	(0.008)	(0.011)
Publication field - Year FE	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564
R-squared	0.082	0.059	0.086	0.094

Standard errors clustered at the Publication field - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 11: Linear probability model estimating the likelihood that a corporate patent cites the academic articles associated with an invention disclosed to the TTO. The higher the commercial potential of the articles associated with inventions, the more likely that a firm's patents will cite them (model 1). The models also explore whether patenting by the TTO leads to more patent citations by firms (model 2) and the interaction between commercial potential and TTO patenting (models 3 to 5). The models control for the articles' scientific and social impact, authors' scientific experience before publishing the focal articles, and authors' prior experience with the TTO. Fixed effects are included at the invention field-year level.

	(1)	(2)	(3)	(4)
	Cited by corporate patent	Cited by corporate patent	Cited by corporate patent	Cited by corporate patent
Commercial Potential	0.305***		0.299***	0.292***
	(0.045)		(0.045)	(0.046)
Patented		0.032***	0.019*	0.020*
		(0.010)	(0.010)	(0.010)
Scientific Potential				-0.052
				(0.056)
Social Impact Potential				0.061***
				(0.020)
Author Scientific Experience				-0.000
				(0.011)
Author TTO Experience				0.008
				(0.009)
Constant	-0.097***	0.111***	-0.102***	-0.114**
	(0.033)	(0.005)	(0.034)	(0.052)
Invention field - Year FE	Yes	Yes	Yes	Yes
Observations	2,689	2,689	2,689	2,689
R-squared	0.227	0.188	0.228	0.231

Standard errors clustered at the Invention Category - Year level
* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 12: Summary statistics: U.S. scientific research published between 2000 and 2020.

	Mean	SD	Min	Max	N
Cited by patent	0.10	0.30	0.00	1.00	5,211,133
Years patents renewed	0.20	0.76	0.00	10.22	5,211,133
Commercial potential	0.49	0.33	0.00	1.00	5,211,133
Scientific potential	0.66	0.24	0.00	1.00	5,211,133
Social impact potential	0.64	0.45	0.00	1.00	5,211,133
Institution commercial impact	113.19	178.75	1.00	16,476.00	5,211,133
Institution scientific impact	715.44	1270.15	1.00	119596.00	5,211,133
Journal commercial impact	0.02	0.05	0.00	3.97	3,705,807
Journal scientific impact	3.08	3.05	0.00	122.33	3,475,151
Researcher commercial impact	8.48	15.88	1.00	3,175.00	5,211,133
Researcher scientific impact	97.23	976.28	1.00	134971.00	5,211,133

Table 13: Correlations: U.S. scientific research published between 2000 and 2020.

Variables	Cited by patent	Years patents renewed	Commercial potential	Scientific potential	Social impact potential	commercial impact	scientific impact	Journal commercial impact	Journal scientific impact	Researcher commercial impact	Researcher scientific impact
Cited by patent	1.000										
Years patents renewed	0.786	1.000									
Commercial potential	0.253	0.206	1.000								
Scientific potential	0.045	0.041	0.224	1.000							
Social impact potential	0.057	0.042	0.305	0.139	1.000						
Institution commercial impact	-0.035	-0.041	-0.011	0.104	0.020	1.000					
Institution scientific impact	-0.045	-0.047	-0.029	0.100	0.006	0.990	1.000				
Journal commercial impact	0.338	0.341	0.299	0.056	-0.022	-0.055	-0.065	1.000			
Journal scientific impact	0.195	0.153	0.244	0.191	-0.041	0.078	0.059	0.415	1.000		
Researcher commercial impact	0.110	0.083	0.258	0.137	0.108	0.585	0.566	0.103	0.197	1.000	
Researcher scientific impact	-0.006	-0.006	-0.018	0.020	-0.022	0.746	0.790	-0.010	0.022	0.437	1.000

Table 14: Linear probability model estimating the likelihood that a paper is commercialized—cited by at least one patent. Our main variables of interest are commercial potential and the interaction of commercial potential with high commercial impact institutions, researchers, and journals. The interaction terms show that publications with high commercial potential are more likely to be cited if they are tied to a high-impact institution, researcher, or journal, suggesting that there is a gap between top and bottom-ranked organizations and individuals. We control for scientific potential and for high scientific impact institutions, journals, and researchers. We also control for social impact potential and include fixed effects at both the field-year and university level.

	(1)	(2)	(3)
	Cited by patent	Cited by patent	Cited by patent
Commercial potential	0.251***	0.144***	0.107***
	(0.018)	(0.011)	(0.010)
Scientific potential	0.037***	0.035***	0.031***
	(0.006)	(0.006)	(0.005)
Social impact potential	0.023***	0.024***	0.024***
	(0.004)	(0.003)	(0.003)
High commercial impact institution	-0.025***	-0.021***	-0.017***
	(0.006)	(0.005)	(0.004)
Commercial potential x High commercial impact institution	0.078***	0.068***	0.046***
	(0.012)	(0.011)	(0.010)
High scientific impact institution	0.069***	0.054***	0.054***
	(0.008)	(0.006)	(0.006)
Commercial potential x High scientific impact institution	-0.144***	-0.113***	-0.122***
	(0.015)	(0.012)	(0.011)
High commercial impact journal		-0.045***	-0.039***
		(0.006)	(0.005)
Commercial potential x High commercial impact journal		0.173***	0.156***
		(0.012)	(0.010)
High scientific impact journal		0.027***	0.026***
		(0.007)	(0.007)
Commercial potential x High scientific impact journal		-0.029**	-0.031**
		(0.013)	(0.013)
High commercial impact researcher			-0.016**
			(0.007)
Commercial potential x High commercial impact researcher			0.140***
			(0.014)
High scientific impact researcher			0.004
			(0.003)
Commercial potential x High scientific impact researcher			-0.029***
			(0.006)
Constant	-0.068***	-0.052***	-0.046***
	(0.011)	(0.009)	(0.009)
Publication field - year FE	Yes	Yes	Yes
University-FE	Yes	Yes	Yes
Observations	5,211,133	5,211,133	5,211,133
R-squared	0.140	0.150	0.160

Standard errors clustered at the publication field-year level and the university level
* p<.1, ** p<.05, *** p<.01

Table 15: OLS model estimating the commercial value of the patents citing a publication, measured by the number of years citing patents are renewed. We condition on publications that are cited by a patent. Our main variables of interest are commercial potential and the interaction of commercial potential with high commercial impact institutions, researchers, and journals. The interaction terms show that publications with high commercial potential are more likely to be cited if they are under a high-impact institution, researcher, or journal. We control for scientific potential and for high scientific impact institutions, journals, and researchers to account for the scientific features of a publication beyond its commercial promise. We also control for social impact potential and include fixed effects at both the field-year and university level.

	(1)	(2)	(3)
	Years patents renewed	Years patents renewed	Years patents renewed
Commercial potential	0.904***	0.637***	0.495***
	(0.033)	(0.040)	(0.042)
Scientific potential	0.016	0.004	-0.019
	(0.041)	(0.042)	(0.040)
Social impact potential	0.103***	0.110***	0.113***
	(0.018)	(0.018)	(0.017)
High commercial impact institution	0.047	0.050	0.045
	(0.034)	(0.033)	(0.031)
Commercial potential x High commercial impact institution	0.164***	0.151***	0.115***
	(0.043)	(0.042)	(0.040)
High scientific impact institution	0.139***	0.131***	0.110***
	(0.037)	(0.036)	(0.035)
Commercial potential x High scientific impact institution	-0.246***	-0.237***	-0.232***
	(0.047)	(0.045)	(0.044)
High commercial impact journal		-0.042*	-0.039*
		(0.023)	(0.023)
Commercial potential x High commercial impact journal		0.228***	0.192***
		(0.032)	(0.032)
High scientific impact journal		0.026	0.028
		(0.033)	(0.032)
Commercial potential x High scientific impact journal		0.049	0.034
		(0.048)	(0.045)
High commercial impact researcher			0.112***
			(0.027)
Commercial potential x High commercial impact researcher			0.184***
			(0.036)
High scientific impact researcher			-0.033
			(0.025)
Commercial potential x High scientific impact researcher			-0.020
			(0.032)
Constant	1.187***	1.231***	1.261***
	(0.046)	(0.054)	(0.056)
Publication field - year FE	Yes	Yes	Yes
University-FE	Yes	Yes	Yes
Observations	519,475	519,475	519,475
R-squared	0.229	0.230	0.235

Standard errors clustered at the publication field-year level and the university level
* p<.1, ** p<.05, *** p<.01

Appendix to:

Exploring the commercial potential of public research

Appendix A Examples of articles by commercial potential quantile

Table A.1: Top 20 percentile of commercial potential

Title	Field	Institution	Journal	Year	Patent cites	Citing patent renewed?
High-resolution mapping of protein sequence-function relationships	Biology	U. of Washington	Nature Methods	2010	26	Yes
Combination strategies to enhance antitumor ADCC	Medicine	Stanford	Immunotherapy	2012	9	Yes
Engineering Tumor-Targeting Nanoparticles as Vehicles for Precision Nanomedicine	Materials science	Rutgers	Med one	2019	0	No
Species-Specific and Inhibitor-Dependent Conformations of LpxC—Implications for Antibiotic Design	Chemistry	Duke	Chemistry & Biology	2011	6	Yes
Multi-Scale 2D Temporal Adjacency Networks for Moment Localization with Natural Language	Computer science	U. of Rochester	IEEE Transactions on Pattern Analysis and Machine Intelligence	2021	0	No
Nanophotonic projection system	Physics	California Institute of Technology	Optics Express	2015	8	Yes
Conserved and Divergent Features of Mesenchymal Progenitor Cell Types within the Cortical Nephrogenic Niche of the Human and Mouse Kidney	Biology	U. of Southern California	Journal of The American Society of Nephrology	2018	0	No
Self-Healing Polyurethanes with Shape Recovery	Materials science	U. of Florida	Advanced Functional Materials	2014	7	Yes
Exploring mechanisms of FGF signalling through the lens of structural biology.	Biology	New York U.	Nature Reviews Molecular Cell Biology	2013	8	Yes
A high-energy-density sugar biobattery based on a synthetic enzymatic pathway	Chemistry	Virginia Tech	Nature Communications	2014	11	Yes
					Mean: 7.5	Mean: 0.7

Table A.2: Bottom 20 percentile of commercial potential

Title	Field	Institution	Journal	Year	Patent Cites	Citing Patent Renewed?
An ion mobility and theoretical study of the thermal decomposition of the adduct formed between ethylene glycol dinitrate and chloride	Chemistry	New Mexico State University	International Journal of Mass Spectrometry	2014	0	No
Continuum shape sensitivity analysis and what-if study for two-dimensional multi-scale crack propagation problems using bridging scale decomposition	Computer science	University of Oklahoma	Structural and Multidisciplinary Optimization	2015	0	No
Dynamic programming solutions for decentralized state-feedback LQG problems with communication delays	Computer science	California Institute of Technology	Advances in computing and communications	2012	1	Yes
Corneal perforation with uveal prolapse: An initial presentation of orbital metastatic breast cancer.	Medicine	Lake Erie College of Osteopathic Medicine	American Journal of Ophthalmology Case Reports	2019	0	No
Dropping Behavior in the Pea Aphid (Hemiptera: Aphididae): How Does Environmental Context Affect Antipredator Responses?	Biology	University of Rhode Island	Journal of Insect Science	2016	0	No
Hydrostatic equilibrium profiles for gas in elliptical galaxies	Physics	Yale University	Monthly Notices of the Royal Astronomical Society	2010	0	No
A Multilevel Quasi-Static Kinetics Method for Pin-Resolved Transport Transient Reactor Analysis	Materials science	U. Michigan	Nuclear Science and Engineering	2016	0	No
Minimizing the Institutional Change Required to Augment Calculus With Real-World Engineering Problems	Computer science	U. North Dakota	PRIMUS	2014	0	No
A 4-year study of invasive and native spider populations in Maine	Biology	U. Massachusetts Amherst	Canadian Journal of Zoology	2011	0	No
Intrusion of a Liquid Droplet into a Powder under Gravity	Medicine	Princeton University	Langmuir	2016	0	No
					Mean: 0.1	Mean: 0.1

Appendix B NLP Models

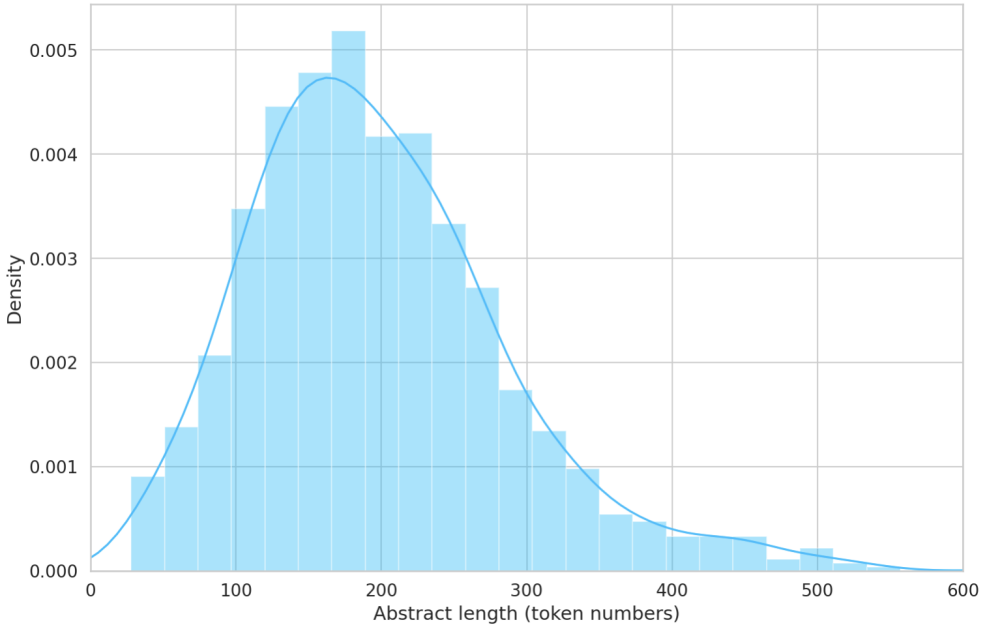


Figure B.1: Abstract's token length distribution

2000				
	precision	recall	f1-score	support
Not cited by ren. patent	0.785	0.728	0.755	1258
Cited by ren. patent	0.743	0.798	0.770	1242
Macro avg	0.764	0.763	0.763	2500
Weighted avg	0.764	0.763	0.763	2500
Accuracy	0.763			
AUROC	0.763			

2001				
	precision	recall	f1-score	support
Not cited by ren. patent	0.731	0.734	0.732	1251
Cited by ren. patent	0.732	0.729	0.731	1249
Macro avg	0.732	0.732	0.732	2500
Weighted avg	0.732	0.732	0.732	2500
Accuracy	0.732			
AUROC	0.732			

2002				
	precision	recall	f1-score	support
Not cited by ren. patent	0.739	0.763	0.751	1226
Cited by ren. patent	0.764	0.741	0.752	1274
Macro avg	0.752	0.752	0.752	2500
Weighted avg	0.752	0.752	0.752	2500
Accuracy	0.752			
AUROC	0.752			

2003				
	precision	recall	f1-score	support
Not cited by ren. patent	0.783	0.736	0.759	1269
Cited by ren. patent	0.744	0.790	0.766	1231
Macro avg	0.763	0.763	0.762	2500
Weighted avg	0.764	0.762	0.762	2500
Accuracy	0.762			
AUROC	0.763			

2004				
	precision	recall	f1-score	support
Not cited by ren. patent	0.765	0.737	0.751	1254
Cited by ren. patent	0.745	0.772	0.758	1246
Macro avg	0.755	0.754	0.754	2500
Weighted avg	0.755	0.754	0.754	2500
Accuracy	0.754			
AUROC	0.754			

2005				
	precision	recall	f1-score	support
Not cited by ren. patent	0.752	0.740	0.746	1269
Cited by ren. patent	0.736	0.749	0.743	1231
Macro avg	0.744	0.744	0.744	2500
Weighted avg	0.745	0.744	0.744	2500
Accuracy	0.744			
AUROC	0.744			

2006				
	precision	recall	f1-score	support
Not cited by ren. patent	0.769	0.697	0.731	1210
Cited by ren. patent	0.739	0.804	0.770	1290
Macro avg	0.754	0.750	0.750	2500
Weighted avg	0.753	0.752	0.751	2500
Accuracy	0.752			
AUROC	0.750			

2007				
	precision	recall	f1-score	support
Not cited by ren. patent	0.745	0.726	0.735	1216
Cited by ren. patent	0.747	0.764	0.755	1284
Macro avg	0.746	0.745	0.745	2500
Weighted avg	0.746	0.746	0.745	2500
Accuracy	0.746			
AUROC	0.745			

2008				
	precision	recall	f1-score	support
Not cited by ren. patent	0.783	0.663	0.718	1248
Cited by ren. patent	0.708	0.817	0.759	1252
Macro avg	0.746	0.740	0.738	2500
Weighted avg	0.746	0.740	0.738	2500
Accuracy	0.740			
AUROC	0.740			

2009				
	precision	recall	f1-score	support
Not cited by ren. patent	0.801	0.619	0.698	1219
Cited by ren. patent	0.702	0.854	0.770	1281
Macro avg	0.752	0.736	0.734	2500
Weighted avg	0.750	0.739	0.735	2500
Accuracy	0.739			
AUROC	0.736			

2010				
	precision	recall	f1-score	support
Not cited by ren. patent	0.762	0.687	0.723	1251
Cited by ren. patent	0.715	0.785	0.748	1249
Macro avg	0.738	0.736	0.735	2500
Weighted avg	0.738	0.736	0.735	2500
Accuracy	0.736			
AUROC	0.736			

2011				
	precision	recall	f1-score	support
Not cited by ren. patent	0.755	0.690	0.721	1253
Cited by ren. patent	0.713	0.775	0.743	1247
Macro avg	0.734	0.732	0.732	2500
Weighted avg	0.734	0.732	0.732	2500
Accuracy	0.732			
AUROC	0.732			

Figure B.2: Commercial potential models' performance (1/2)

2012				
	precision	recall	f1-score	support
Not cited by ren. patent	0.772	0.685	0.726	1282
Cited by ren. patent	0.704	0.787	0.743	1218
Macro avg	0.738	0.736	0.735	2500
Weighted avg	0.739	0.735	0.734	2500
Accuracy	0.735			
AUROC	0.736			

2013				
	precision	recall	f1-score	support
Not cited by ren. patent	0.756	0.718	0.737	1186
Cited by ren. patent	0.757	0.791	0.773	1314
Macro avg	0.756	0.755	0.755	2500
Weighted avg	0.756	0.756	0.756	2500
Accuracy	0.756			
AUROC	0.755			

2014				
	precision	recall	f1-score	support
Not cited by ren. patent	0.762	0.685	0.722	1236
Cited by ren. patent	0.720	0.791	0.754	1264
Macro avg	0.741	0.738	0.738	2500
Weighted avg	0.741	0.739	0.738	2500
Accuracy	0.739			
AUROC	0.738			

2015				
	precision	recall	f1-score	support
Not cited by ren. patent	0.757	0.640	0.694	1229
Cited by ren. patent	0.697	0.802	0.746	1271
Macro avg	0.727	0.721	0.720	2500
Weighted avg	0.727	0.722	0.720	2500
Accuracy	0.722			
AUROC	0.721			

2016				
	precision	recall	f1-score	support
Not cited by ren. patent	0.746	0.700	0.722	1248
Cited by ren. patent	0.718	0.762	0.740	1252
Macro avg	0.732	0.731	0.731	2500
Weighted avg	0.732	0.731	0.731	2500
Accuracy	0.731			
AUROC	0.731			

2017				
	precision	recall	f1-score	support
Not cited by ren. patent	0.799	0.573	0.668	1237
Cited by ren. patent	0.673	0.859	0.755	1263
Macro avg	0.736	0.716	0.711	2500
Weighted avg	0.735	0.718	0.712	2500
Accuracy	0.718			
AUROC	0.716			

2018				
	precision	recall	f1-score	support
Not cited by ren. patent	0.754	0.711	0.732	1261
Cited by ren. patent	0.722	0.764	0.743	1239
Macro avg	0.738	0.738	0.737	2500
Weighted avg	0.739	0.738	0.737	2500
Accuracy	0.738			
AUROC	0.738			

2019				
	precision	recall	f1-score	support
Not cited by ren. patent	0.758	0.635	0.691	1251
Cited by ren. patent	0.685	0.797	0.737	1249
Macro avg	0.721	0.716	0.714	2500
Weighted avg	0.721	0.716	0.714	2500
Accuracy	0.716			
AUROC	0.716			

2020				
	precision	recall	f1-score	support
Not cited by ren. patent	0.806	0.615	0.698	1256
Cited by ren. patent	0.687	0.850	0.760	1244
Macro avg	0.746	0.733	0.729	2500
Weighted avg	0.747	0.732	0.729	2500
Accuracy	0.732			
AUROC	0.733			

Figure B.3: Commercial potential models' performance (2/2)

Table B.1: Scientific potential model performance (average of all models: 2000-2020)

	Precision	Recall	F1-score
$\leq$ 16 scientific citations	0.73	0.71	0.72
$>$ 16 scientific citations	0.70	0.72	0.71
Accuracy			0.71
Micro-averaged ROC AUC			0.71

Appendix C Social impact measure construction

1. The finding described in the abstract addresses a need or solves a problem.
2. There are a lot of people or businesses with this need or problem.
3. The finding described in the abstract provides at least a first step to addressing a need or solving a problem.
4. It is crucial for either society or for businesses to address this need or problem.
5. There is currently no good solution available that addresses this need or problem. In other words, solving the need would be a major advance.
6. The abstract suggests a specific way (a solution) to address the need or problem.
7. The finding in the abstract is close to addressing the need or problem.
8. Addressing this need or problem will require a lot of effort and resources beyond what is already described in this abstract (in terms of people, equipment, money, time, etc.).
9. The finding described in the abstract could lead to a product or process that will make money for either a company or a person.
10. After reading the abstract, I have gained a sense of the potential applicability of the finding to society or business.
11. The abstract clearly states the question and finding in a way that I could understand.

Figure C.1: Questions used in the elicitation exercise to create the social impact measure of a scientific article.

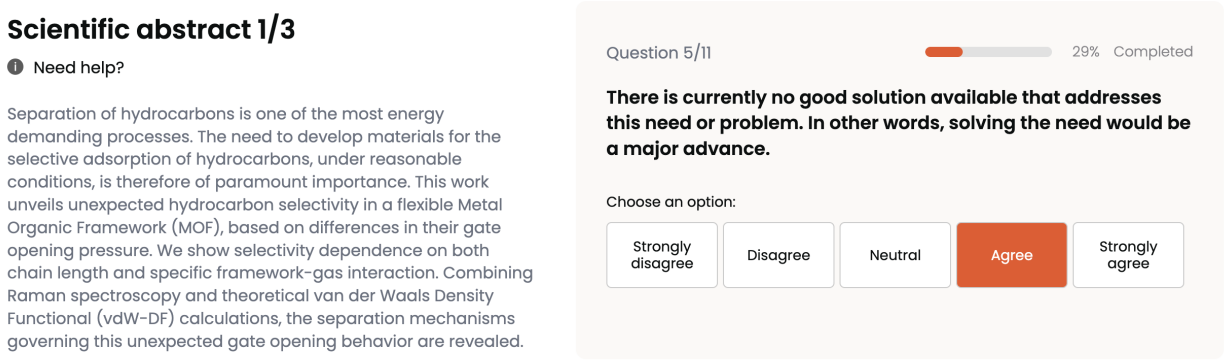


Figure C.2: Interface evaluators interacted with in the elicitation exercise.

Table C.1: Linear probability model regressing indicator for whether the paper i received any citation by a patent on *Commercial Potential*. *Commercial potential* is calculated using pooled data of all Worker-Article evaluations. Models successively include controls for forward citations in the academic literature *Log(Paper Cites*, abstract length, and fixed effects for the field of research for the study (i.e., Biology, Physics, Computer Science, Materials Engineering, Mechanical Engineering and Electrical Engineering.)

	(1)	(2)	(3)	(4)
	Any Patent	Any Patent	Any Patent	Any Patent
Commercial Potential	0.031***	0.024***	0.028***	0.026***
	(0.005)	(0.004)	(0.004)	(0.005)
Log(Paper Cites)		0.127***	0.129***	0.135***
		(0.007)	(0.007)	(0.007)
Log(Abstract Length)			-0.077***	-0.056**
			(0.025)	(0.025)
Constant	0.660***	0.320***	0.838***	0.681***
	(0.013)	(0.026)	(0.170)	(0.175)
Category FE	No	No	No	Yes
Observations	1,200	1,200	1,200	1,200
R-squared	0.033	0.213	0.219	0.226

Robust standard errors
* p<.1, ** p<.05, *** p<.01

Table C.2: Linear probability model regressing indicator for whether the paper i received any citation by a patent, by category (i.e., Biology, Physics, Computer Science, Materials Engineering, Mechanical Engineering and Electrical Engineering.) *Commercial potential* is calculated using pooled data of all Worker-Article evaluations. Models successively include controls for forward citations in the academic literature *Log(Paper Cites*, abstract length, and fixed effects for the field of research for the study

	(1)	(2)	(3)	(4)	(5)	(6)
Commercial Potential	0.039***	0.041***	0.028**	0.036***	0.032**	0.037**
	(0.009)	(0.014)	(0.012)	(0.012)	(0.015)	(0.014)
Constant	0.673***	0.624***	0.688***	0.711***	0.651***	0.604***
	(0.032)	(0.037)	(0.034)	(0.035)	(0.034)	(0.042)
Category	Biology	Computer Science	Materials	Physics	Electrical	Mechanical
Observations	200	200	200	200	200	200
R-squared	0.075	0.041	0.024	0.039	0.022	0.033

Robust standard errors
* p<.1, ** p<.05, *** p<.01

Appendix D TTO additional data

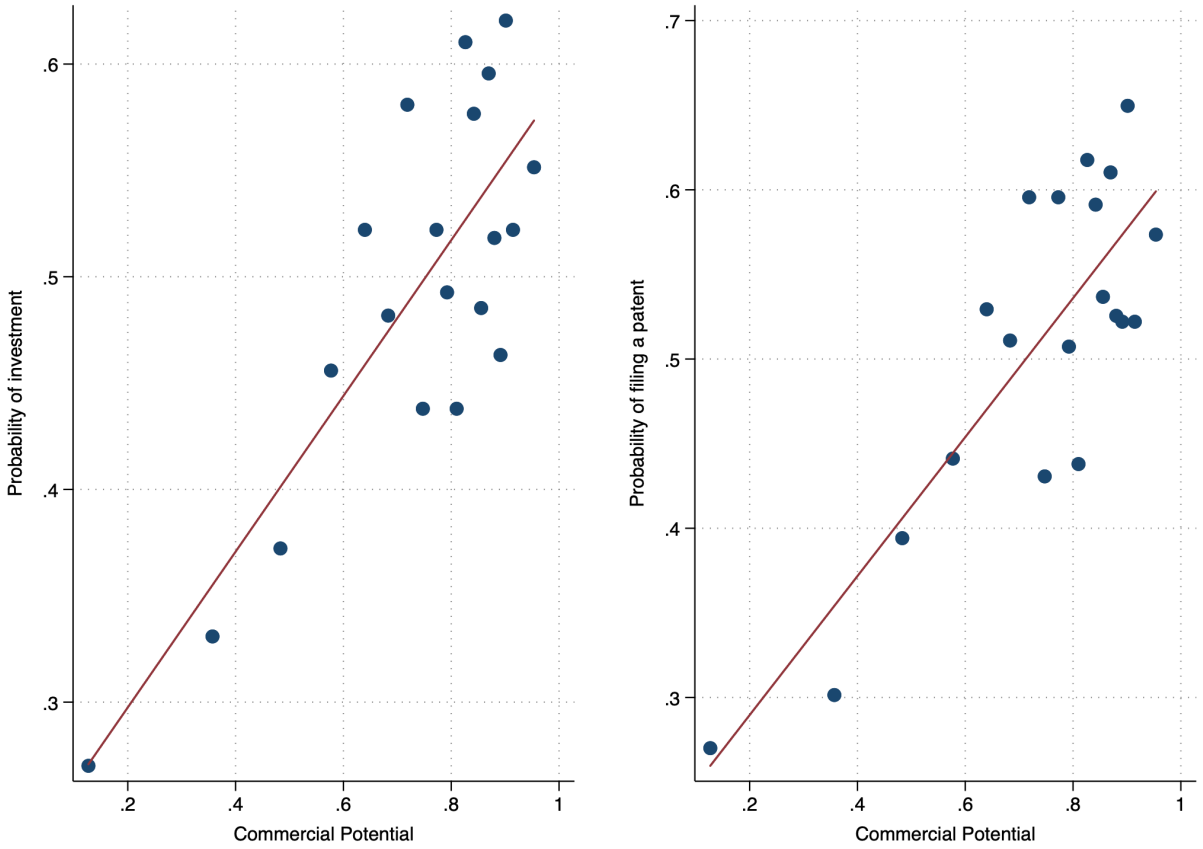


Figure D.1: Probability that the TTO will invest into (Panel A) and patent (Panel B) an invention based on the average commercial potential of the articles associated with the invention.

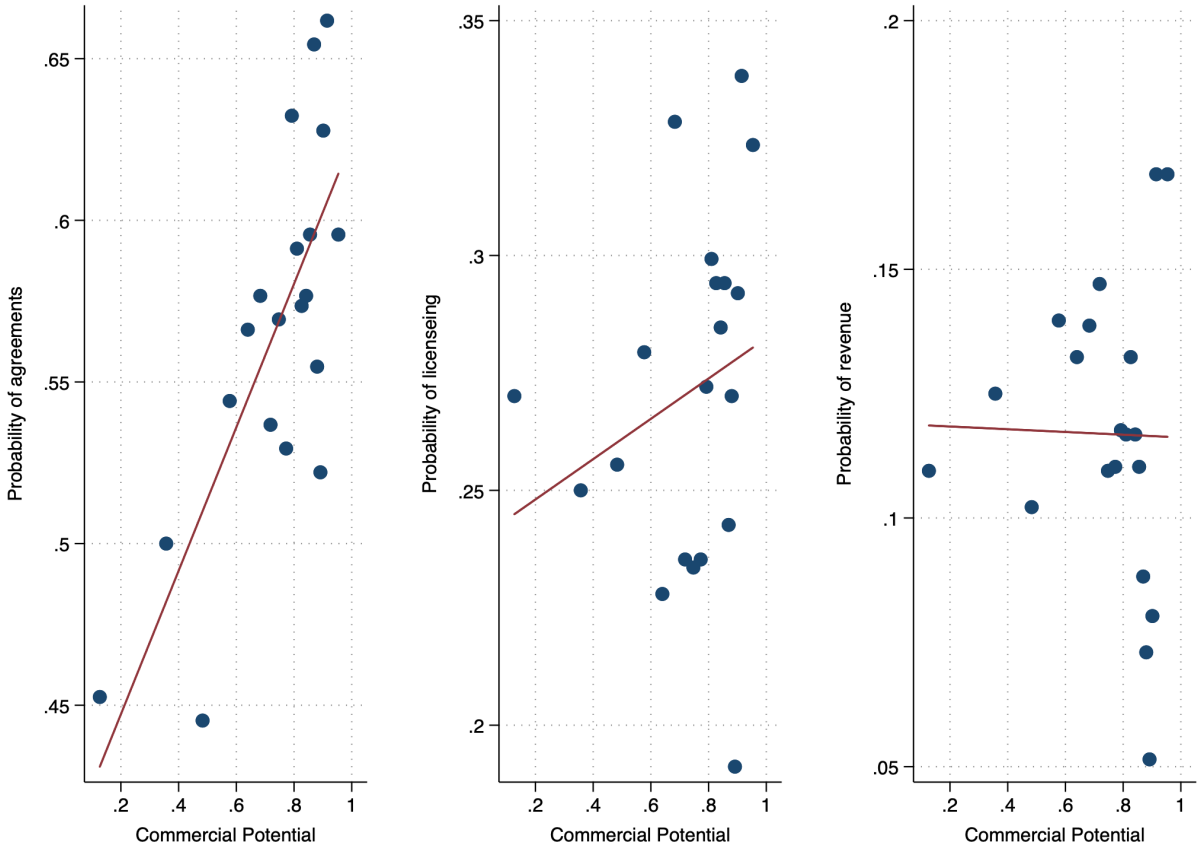


Figure D.2: Probability that an invention will garner agreements (Panel A) and licensing deals (Panel B), as well as generate revenue to the TTO (Panel C) based on the average commercial potential of the articles associated with the invention.

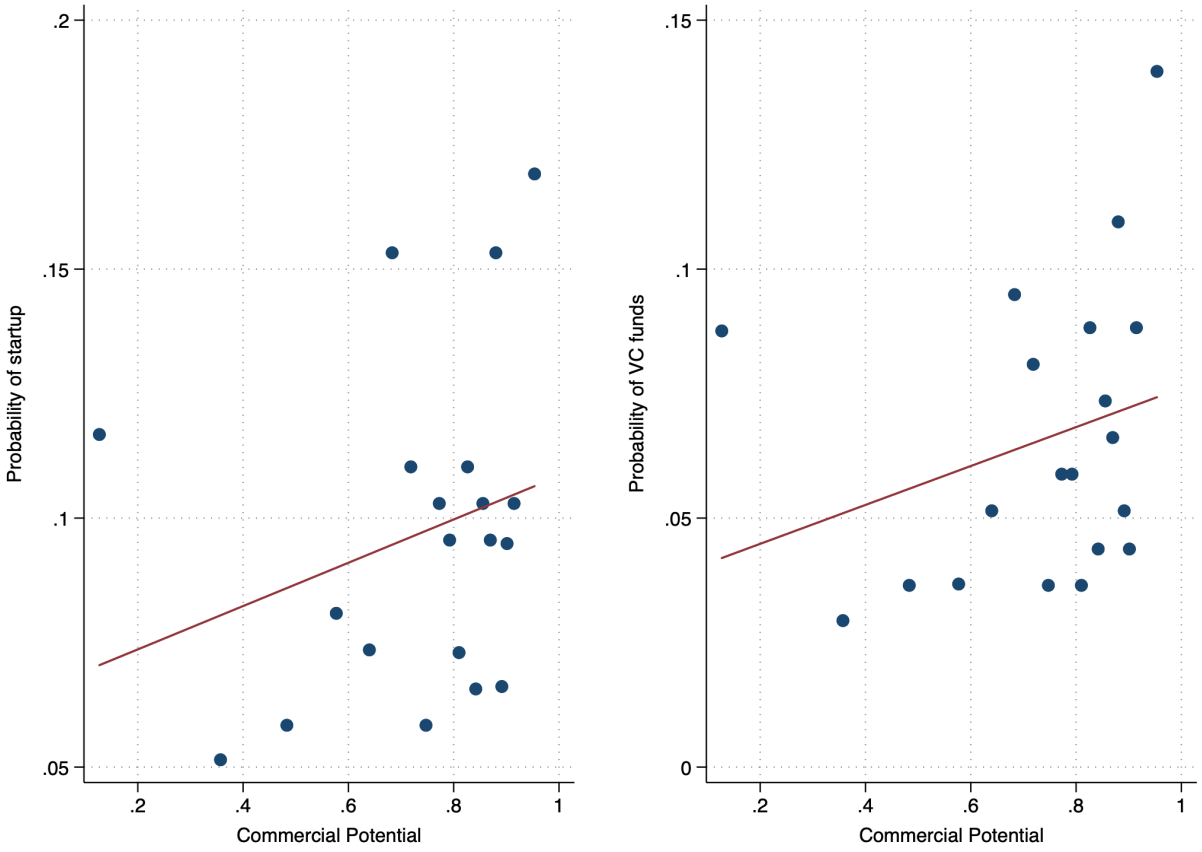


Figure D.3: Probability that an invention will be commercialized via a Startup (Panel A) and, conditional on Startup, that will raise venture capital funds as a function of the average commercial potential of the articles associated with the invention.