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

Roger Masclans† Sharique Hasan† Wesley Cohen†

October 2, 2023

Abstract

The research of universities and government labs—what we will call *science*—contributes to technical advance and economic growth. Despite the potential, obstacles in the commercialization process hinder the full realization of these scientific contributions. We aim to identify which scientific contributions have commercial potential and understand what restricts their commercialization. We also seek to discern if lower commercialization rates for some institutions are due to challenges in the commercialization process or because the science there has lower commercial potential. We introduce a new text-based measure of commercial potential for scientific articles. Using this measure, we study the commercialization process with two empirical exercises. First, we analyze the commercialization of over 92,000 articles from a leading research university. Our measure predicts which science is disclosed, investments by the 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. Thus, many high-potential articles remain undisclosed, and undisclosed science is much less likely to be cited in corporate patents. Second, we analyzed the commercialization outcomes of nearly 4 million articles published between 2010-2022. 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 low commercial impact have commercialization rates 10% lower than their peers.

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.

1 Introduction

Scientific research is crucial for technological advance.[1] Science offers novel ideas and aids in practical application and execution (Klevorick et al., 1995; Cohen et al., 2020). While the importance of public research to technical advance may be growing relative to ideas developed in-house by firms (Fleming et al., 2019; Li et al., 2017; Arora et al., 2018), there remains a longstanding question of the extent to which the commercial promise of science may be 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 initiatives to disseminate 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. Quantifying the exact amount of high-potential science that remains commercially undeveloped is challenging. This gap in understanding is due to the lack of systematic measures that evaluate the commercial potential of the science distributed across millions of researchers and thousands of institutions worldwide. In this paper, we introduce one such measure, use it to assess the extent to which the commercial potential of science is unrealized, and explore reasons for what we call the "realization gap."

Recent research suggests the existence of a "realization gap" by identifying 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. Although these frictions confirm our expectation that firms are not taking advantage of some fraction of commercially relevant science, identification of these particular frictions does not permit an estimation of the overall extent of the realization

[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).

gap, nor does it provide a comprehensive picture of the factors impeding the commercial application of science.

While we can estimate how often academic research leads to commercial applications by looking at how patents cite scientific literature, this method does 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 goes 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 application 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 how much science with commercial potential remains undeveloped, we first must identify that 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 motivated by two challenges—*search* and *evaluation*—that characterize firms and relevant stakeholders' difficulty when looking for ideas with commercial potential among scientific discoveries and ideas presented in various formats such as publications, reports, and presentations. Stakeholders, including R&D personnel in firms, venture capitalists, and university technology transfer officers, grapple with the high search costs due to the vast array of available publications, reports, and conference proceedings. Even when they identify a promising scientific finding, they must evaluate whether they can transform it into a new or improved product cost-effectively or if the product will meet market demands

[2]Using such data, we observe, for example, 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 quantification of how much of this output has commercial potential. As a consequence, this variation cannot tell us about the extent to which the commercial potential of scientific contributions goes unrealized.

(Cohen et al., 2002; Fleming and Sorenson, 2004). This uncertainty means that decisions about a finding's commercial potential can be prone to errors, limited by the available information, or influenced by biases. Additionally, a single scientific discovery might apply to multiple products, further complicating its commercial evaluation. Given these challenges, decision-makers with the necessary resources and authority must first recognize a scientific idea or finding as having commercial value before it can transition to a commercial product.

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 undertake two steps to estimate the probability that an article has commercial potential. First, using the abstract text of a scientific article and an indicator of whether a renewed patent cites the article, we train a language-based model to learn to recognize the textual patterns predictive of renewed patent citations. We rely on BERT, Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) to do so. Second, once the model is trained, we compute the probability that an article will garner renewed patent citations—our measure of commercial potential. Our model accuracy, measured as the area under the receiver operator characteristic, or AUROC, is 0.75.[3]

[3]As a benchmark, Manjunath et al. (2021) report an AUROC of 0.83 in predicting whether a patent will cite an article. Their model, however, is trained only with abstracts of articles published in PubMed and are from the life sciences, and it does not condition on patent renewals. In contrast, Liang et al. (2022) train a model based on the text of inventions disclosed to the TTO at Stanford and predict whether the invention

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 university's 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. What is also striking is that the financial investments by the university in protecting and marketing the technology appear to mediate the effect of commercial potential we find in our analyses. The pattern of our findings suggests that the greatest 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. Among 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. In this context, given that the quantity of undisclosed science that is cited is much greater than the quantity of cited science that is disclosed, the primary obstacle seems to be the search to identify high commercial potential ideas among thousands at this university. Once these ideas are, however, disclosed, the decisions of the technology transfer office to develop and invest seem aligned with the potential as indicated by our measure.

Our second exercise analyzes the commercial outcomes of nearly 4 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 cited by a patent (use) and, conditional on such a citation, to be subsequently renewed (value). Moreover, our findings indicate that an article's commercial potential is distinct from its scientific or

will generate commercial value. In this case, they report an AUROC of 0.76.

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 150% more likely to be cited by a patent than the average article. Furthermore, we observed considerable heterogeneity in the effect size across institutions. High-potential articles published by authors at universities with a history of commercial impact are cited by renewed patents 10% more than those that are not. Interestingly, we observe little of a realization gap across institutions once the potential of articles drops below the median score.

Our study makes several contributions to the literature on innovation, entrepreneurship, and scientific commercialization. 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 the identification and tracking of nacent 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).

Finally, our work 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 address challenges posed by the latent commercial potential of science. By comparing the characteristics of the two similar inventions and the teams that developed them, the researchers could draw conclusions about the factors that influence the commercialization of scientific ideas. 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

The primary question motivating recent literature is whether the full commercial potential of scientific investments is being realized. To assess the commercial potential of science and identify any unrealized potential, we can look at three key groups of participants whose choices play a crucial role.: (1) Individual researchers, whom we will refer to as "academics";[4]

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

(2) their affiliated *institutions*, primarily universities; and (3) the *firms* that leverage scientific findings to innovate new products, processes, and services. The knowledge transfer from academics and their institutions to firms can be seen from two perspectives: supply and demand. The supply side consists of the academics conducting the research and the institutions that furnish the physical and intellectual infrastructure supporting them. These institutions are also responsible for defining and administering the incentive systems that guide academic endeavors. Conversely, 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 important to note that this context is not a market for several reasons. First, academics are not "motivated sellers". Their objectives are typically not personal profit from commercial activity, unlike firms. We know from rich literature (e.g., Merton, 1973) that their motives tend to be the achievement of eminence and career advancement in the form of academic promotion and tenure. 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) and is effectively free to firms.

Yet, thinking about this setting in terms of supply and demand (or, more accurately, "use") provides a simple framework for thinking about 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. Contrary to businesses, they are often driven not by the lure of commercial profit but by the aspiration to advance foundational knowledge or to address societal needs, as observed by Stokes and exemplified in Pasteur's quadrant (Cohen et al., 2020). The "translational" effort required to make academic science commercially viable is a significant consideration. This means converting scientific findings into language and contexts businesses can grasp and utilize. Such adaptation requires clearly articulating the research

and insights into its practical applications and, at times, its active promotion. However, two primary challenges arise. First, many academics may not be inclined to put in the additional effort for commercial translation because it does not align with their personal or academic goals. Second, the existing institutional incentives might not recognize or reward such endeavors. Consequently, a wealth of potentially valuable scientific insights may remain unnoticed by the business world.

Another feature of the supply side may be the level of resources that a university administration decides to devote to translational efforts through its TTO and its 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 resources, networks, and norms.

On the demand or "use" side are firms. Recent studies, including those by Cohen et al. (2002), Arora et al. (2016), and Bikard (2018), suggest that firms often give limited attention to research originating in universities. Instead, firms seem more inclined to build on knowledge from other firms rather than academic sources. This pattern of behavior is understandable. One key factor is the high search cost of sifting through academic research. For businesses to effectively harness academic insights, they need to invest in the capability to assess the commercial value of potentially thousands of articles and researchers. This becomes especially challenging and costly if academic authors have not taken steps to make their findings more accessible or "translated" them into practical applications.

Additionally, there is uncertainty regarding the relevance, quality, and potential payoff of academic science. Businesses often adopt simplified search strategies as a practical response to this uncertainty. 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.

Therefore, potential supply-side explanations for the realization gap may encompass the motivations of academics, the incentive structures within their institutions, and the resources these institutions allocate to translation. On the user side, explanations could be derived from a firm's ability to assimilate and apply new knowledge (absorptive capacity), coupled with the 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 geographical proximity and established relationships, to direct their search. While these attributes of academics, their institutions, and firms might account for a part of the realization gap, it must be recognized that a specific segment of this gap might inherently be due to the duration required, even with ample resources, to transition foundational science to a commercial application.

3 Methods and measurement

This section describes the methodology and data used to create the measure of *commercial potential*. As previously mentioned, this measure estimates the likelihood, based on the abstract text in which a scientific discovery is reported, that it will be later incorporated into a product, service, or technology that generates economic value. The measure is derived from a Large Language Model (LLM) that has been fine-tuned using nearly 100,000 observations in a Natural Language Processing (NLP) classification task.[5] We use patent-to-paper citations to link a scientific finding with a patented invention, and we train the NLP model with papers

[5]LLM and NLP are a subfield of Artificial Intelligence (AI) that focuses 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 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.

cited by patents that have been renewed at least once. We interpret a renewed patent's citation to an article as an indication that a firm believes that the article is contributing to the development of an invention that may have commercial value. We labeled scientific articles as having high commercial potential if they are cited by at least one patent that has been renewed one or more times and papers as having low commercial potential otherwise. The language model outputs the probability that a new, unseen paper is associated with each of the predefined classes, *low* and *high*. We trained the model with two classes, and we define the commercial potential measure (ϕ) as the likelihood that the paper is classified in the *high* class:

$$\phi_i = p_{i,high}^{pat_renewal} \tag{1}$$

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 (Devlin et al., 2018), an NLP model developed by Google AI. SciBERT was pre-trained on a large corpus of scientific publications and then fine-tuned using our dataset for the specific task of classifying scientific articles based on their commercial potential.[6]

[6]Presumably, 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), which are a novel type of neural network architecture that has revolutionized the field of NLP. Transformers are capable of modeling long-range dependencies, learning contextual representations, and handling 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 the model 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 representations of words and phrases.

Training a transformer-based language model can be a daunting task, and we do not attempt it. Instead, as previously mentioned, BERT (Bidirectional Encoder Representations from Transformers) Devlin et al. (2018), which has been trained on a large amount of text data and can be fine-tuned for various NLP tasks. For scientific-related tasks, we use its derived version called SciBERT (Beltagy et al., 2019).

A possible limitation for our analysis is that the training sample for SciBERT (Beltagy et al., 2019) was comprised of 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

[7]This first step of tokenizing entails converting each abstract into an array of discrete linguistic units—usually, units are words, parts of a words, numbers, symbols, and stems. We tokenize using the version that SciBERT's authors recommend, *scibert-scivocab-uncased*, based on BERT's tokenizer. 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 a large 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).

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 difference in the the average number of tokens between the classes (which could create bias in our findings). Therefore, we truncate the abstracts at 512 tokens.[8]

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]

To ensure a high quality of scientific findings, we limited our analysis to peer-reviewed journal articles and findings reported in proceedings, resulting in a set of 70,805,788 scientific papers. Moreover, we focused our analysis on eleven scientific fields, which cover the majority of natural and formal 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 decreases to 50,362,042 academic papers.

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

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

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 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. We assume that papers not cited by a patent in the resulting set have no patent citations.

Next, we use Google Patents Research to add patent assignee and patent forward citations by other patents to the dataset. We also use data from the United States Patent and Trademark Office (USPTO) to include information on whether patents have been renewed and how many times. Finally, we merge the three datasets containing patent information (Reliance on Science, Google Patents Research, and USPTO) using the patent publication number.

3.3.3 Technology Transfer Office Data

We have access to detailed data from the technology transfer office (TTO) of a leading 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 2010 and those not associated with an active researcher at the time of disclosure. The resulting dataset includes 2,958 invention disclosures spanning from January 2010 to October 2022.

To identify the scientific findings relevant to an inventor's disclosure to the TTO, we proceed in three steps. In the first step, we match the two main datasets: Dimensions,

containing academic publications information, with the TTO dataset. To do so, we extract from Dimensions the names of all the researchers affiliated with the TTO's University at any point in time (which we can map directly using a unique researcher id to data on their academic publications, also in Dimensions). Then, we use a fuzzy matching algorithm to match names of these researchers in Dimensions to the names of researchers that disclosed inventions as per the TTO data. The resulting dataset contains triplets of publications, authors, and invention disclosures to the TTO on the time period of 2010 to 2022, with a total of 4,202 researchers matched from one data set (Dimensions) to the other (TTO data) and linked to 29,998 unique publications and 3,249 inventions.

An invention disclosure does not, however, have to be associated with a paper merely on the basis of the author of the paper being the same as the inventor of the disclosure. To pair invention disclosures with a set of academic papers—the papers that contribute to the invention—we rely on the time proximity between the publication and the disclosure (second step) and on the text similarity between publications and inventions (third step). In the second step we proceed as follows. For a paper-author-invention triplet, we calculate a time gap measure, which captures the time gap in years between the paper publication and the invention disclosure (with the baseline time being the year of the invention disclosure). For example, if an invention was disclosed in 2015 and a paper associated with the invention (via researcher) was published in 2013, the year-gap for the triplet is two. Similarly, a different paper from the same author published in 2021 would have a year-gap of six. We then assign papers as influencing or building an invention—i.e., tentatively match a paper to an invention—if the paper-author-tech triplet is within a specific time window (year-gap). In particular, we use a [-1,3] time window.[10] This step yields a set of 3,277 researchers, linked to 16,529 publications and 2,677 inventions.[11]

[10]This time window reflects the insights shared by the TTO office, which requests that inventors disclose their inventions prior to public disclosure (e.g., public presentations, publication) in order not to foreclose the option of patenting in European and other jurisdictions where a one year grace period between public disclosure and filing date does not apply. Along these lines, several research documents the fact that scientific publications leading to patenting events are proximate in time (Azoulay et al., 2007; Marx and Fuegi, 2020).

[11]Note that an invention can have more than one inventor. Thus, an invention can be matched to the

Finally, in the last step, we further narrow down the matching between publications and inventions on the basis of text similarity. Following a similar procedure as the one described in the previous section, we use BERT to compute the textual embeddings of the titles of the papers and those of the titles of the inventions. Next, for a given publication-invention potential match (based on same author and the time window), we compute the cosine similarity between the embeddings of their titles. Upon visual inspection, we decide that matches over 0.5 are good matches, and we finally assign a publication as influencing an invention.[12]

The three steps resulted in 9,222 unique publications matched to 1,829 inventions, with 2,367 researchers linking them. The median number publications associated with an invention is 3, in line with other research investigating paper-patent pairs (Marx and Fuegi, 2020). For the analyses that follow, we create two subsets. First, we aggregate the information at an article level, which captures, for each article, whether it influences an invention disclosed to the TTO, as well as the characteristics of the article subject to study, such as its commercial potential, scientific potential, citations made by patents, and so on. This set will allow us to analyze the characteristics of the scientific articles associated with inventions disclosed to the TTO. Therefore, to the 9,222 unique publications associated with an invention, we add those that are not but that are published by researchers also affiliated with the focal university in the same time period. The resulting set contains 92,972 articles, 9,222 of them (9.92%) associated with an invention disclosure. Second, we aggregate the data at the invention level to explore the relationship between scientific articles and the commercial outcomes that the associated inventions experience, such as TTO investment, patents filled, licensing agreements, and revenue. Since inventions are linked to more than one article, we average the relevant article variables (commercial potential, scientific potential, patent citations, etc.) in

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.

[12]We also used the same procedure using the abstract of the publication and the description of the invention, with similar results. However, we find that matching based on title is less noisy.

aggregating the data. The resulting set contains 1,829 inventions detailed above, tied to the 9,222 articles associated with a disclosure and their data.

Table 1 describes the variables used in the TTO analysis. Table 3, Panel A, provides summary statistics about all the articles published between 2010 and 2022 and authored by researchers affiliated with the TTO's university at the time of publication, and, Panel B, provides summary statistics about the inventions disclosed to the TTO and the scientific research these rely on.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

3.4 Commercial potential model: Training sample

To train our primary model, which yields our measure of commercial potential, we use scientific articles published between 2000 and 2009. The time period for the training data is defined considering several aspects. First, to obtain unbiased results and mimic real-world predictions in the subsequent analysis exploring the commercial potential of articles across the U.S. (section 5), we should not use an overlapping time frame. Isolating the training set temporally will help avoid over-fitting, as it is possible that the model could learn to fit the noise in the training data, rather than the underlying patterns that generalize to new data. Likewise, we also avoid data leakage, which occurs when information from the prediction set is accidentally included in the training set, inflating performance metrics. This happens, for example, if articles report the findings or conclusions of other articles—which is indeed often.

Second, we believe that the models are time-sensitive, i.e., a model trained with old articles would not be as useful today as a model trained with articles closer in time. A possible reason is vocabulary changes over time. Another potential reason is the topics

that, within a field, are being explored. If we use older data, the model will learn semantic relationships associated with patenting activity that, for recent articles, might be useless, decreasing the performance in such cases. Thus, we select a time-frame (2000-2009) that 1) allows for an out-of-sample time period relevant for the subsequent analysis (2010-2022) and 2) that does not contain scientific articles that are too old, which would hamper the accuracy of the model in evaluating more recent articles.

The resulting set consists of 11,096,451 articles. Training a model with more than 11 million observations would incur high computing costs, and our analyses confirm that the accuracy of the model converges with a much smaller sample size. Thus, we randomly sample 5% of the observations, yielding 554,882 articles.[13]

As previously discussed, our objective is to train a model that predicts the commercial potential of a scientific article based on its likelihood to be used in the development of commercializable inventions, and, to do so, we identify articles that are cited by renewed patents. In our subsample of 554,882 articles, 49,259 papers met this criterion, which will be used for the training exercise. These papers will be classified as having a high commercial potential, as per equation 2 below.

NLP classification tasks generally require balanced samples, where each class has the same number of observations, to avoid potential bias, particularly in neural networks. However, since we only have 49,259 papers associated with the *high commercial potential* class in the training subsample, we are limited in the number of papers we can use for the complementary class (*low commercial potential*). Consequently, we randomly select an additional 49,259 articles from the same subsample that are either 1) never cited by a patent or 2) cited by a patent that is never renewed. Thus, the final sample comprises *98,518 articles*, with 50% associated with the *commercial potential* and 50% with the *no commercial potential* class, and is representative of the full range of papers the model may encounter in casting predictions with unseen data, including papers that are never cited by a patent, papers that

[13]Future drafts of the paper might be based on a larger training sample.

are cited by a patent not renewed, and papers cited by renewed patents.

Formally, the articles in the training sample are labeled as follows:

$$y_i = \begin{cases} \textit{commercial potential}, & \text{if (Patent renewal} \mid \text{Patent cite}_i > 0) > 0, \\ \textit{no commercial potential}, & \text{if (Patent renewal} \mid \text{Patent cite}_i > 0) = 0, \\ \textit{no commercial potential}, & \text{if Patent Cite}_i = 0 \end{cases} \tag{2}$$

Finally, we split the sample defined so far (98,518 articles) into three sets: a training set (80%) and a test set (10%) to training the model, and a validation set (10%) to validate the final performance and accuracy. The split follows best practices in machine learning and it is done to to ensure that the neural network yields unbiased results, with its final accuracy assessed with data that the model has never seen.[14]

3.5 Commercial potential model: Performance

After training the NLP model with the training and test sets as described above, we use the validation sample, held out during the training, to test the model's accuracy with unseen data.[15] The model's accuracy is 75%, i.e., 7.5 out of 10 abstracts predicted classes (whether

[14]Neural networks are trained in iterations (each iteration is called an epoch). To ensure that the iterations yield unbiased results, the sample data needs to be split between a training set and a test set. The epochs are computed as follows: first, the network parameters are randomly initialized, and the input data is fed into the first layer of the network to compute the output predictions. Each layer performs a set of calculations using the weights assigned to each neuron, transforming the data into a new representation. The output of each layer is then passed as input into the next layer, and the process is repeated until the last layer. This process is called forward propagation. Next, to optimize each parameter in the network, the algorithm starts a process called backward propagation, in which it computes the gradients of the loss functions. This process updates the parameters to improve the model's performance. Once a forward and backward propagation are completed, i.e., one epoch is finished, the next iteration starts. The model will keep performing epochs until the loss function converges or until it reaches a predefined number of epochs. After completing the training process (the defined epochs), the overall performance must be assessed. To obtain unbiased performance estimates, a new out-of-sample validation set is required, meaning that the training and test sets used in the learning process cannot be reused. For this reason, the original set for training is split into three subsets.

[15]To train the model, we experimented with different parameters. The parameters leading to the highest performance are five *epochs*—i.e., five iterations on the neural network optimization procedure—a *batch size* of 16—i.e., the size of the training subsamples propagated through the neural network—, and a learning rate of *2e-5*—a tuning parameter used to find the optimal values to minimize the loss function. We treat the

an article will garner citations of patents that are renewed) correspond to their actual classification, and the area under the receiver operating characteristic (AUROC) of 0.75, indicating a balanced error between false positives and false negatives.[16] Tables A and A.2 in Appendix A provide examples of articles by top 20 and bottom 20 percentiles of commercial potential using completely out-of-sample articles, published *after* our training sample ends.

Although the classifier is reasonably accurate, there is still room for improvement. We identify two factors that might affect the results. First, our single NLP model is trained to classify articles across various fields, ranging from Biology to Materials science to Computer Science. However, the textual features leading to one class or another may differ substantially across these fields, resulting in trade-offs in setting the parameters, which limits the model's overall performance. As a benchmark, Manjunath et al. (2021) recently trained an NLP model to predict whether a scientific article will be cited by a patent on more than 20 million scientific articles from PubMed, which only includes life sciences and biomedical fields, with an AUROC of 0.83. In subsequent revisions of the paper, we plan to train a separate model for each of the eleven fields to improve performance.

The second factor that limits the model's performance, as discussed in previous sections, is the highly complex nature of the task we aim to classify. The outcomes are uncertain by nature and are difficult to predict, even for an expert human with additional information beyond the text. Most NLP classification tasks are relatively simple, with most of the necessary information contained within the text (e.g., conveying particular emotions like fear or joy), and have accuracy levels usually exceeding 95%. However, the performance is expected to be lower when tasks are complex. For example, Liang et al. (2022) analyze 4,512 inventions (not papers) disclosed to Stanford's Technology Transfer Office, aiming to

classification as a multi-class problem and use a sigmoid function for the network's last layer. Additionally, we train the model using Amazon Web Services hardware, specifically the ml.p3.8xlarge configuration, which offers 32 vCPUs and 244 GiB of memory.

[16]The ROC curve is computed by plotting the true positive rate against the false positive rate at different thresholds. A perfect classification task would yield a curve with an area of 1, while a naive task would have an area of 0.5. Intuitively, the ROC curve captures how balanced a binary classifier is in classifying true positives, true negatives, false positives, and false negatives.

classify whether an invention would generate net income over the median net income of the sample. The authors trained two different NLP models. The one based on BERT achieved an AUROC of 0.76, while the one based on TF-IDF[17] achieved an AUROC of 0.71. Similarly, Guzman and Li (2023) aimed to predict the value of a startup (whether it receives early-stage funding, conducts an IPO, or gets acquired) by training a model on the startup idea's value-added description. Using doc2vec,[18] they reported AUROCs ranging from 0.60 to 0.65.

3.6 Secondary models: Scientific and social potential measures

There are other features of scientific findings that could affect the decision to commercialize but may be, in some ways, orthogonal to their commercial potential. Two such features are the scientific potential and the social impact potential (e.g., Cohen et al., 2020) of a finding. For instance, especially in academic settings, both scientists and organizations may prioritize findings that advance basic understanding and/or have a high potential to impact society, regardless of their commercial applicability or viability. Scientific and social impact potential are not necessarily at odds with commercial potential. Indeed, (Cohen et al., 2020) observed that those life science academics whose research was highly motivated by a desire to achieve social impact patented the most. But, depending on the field and the individual, these two objectives may not be consistent with conducting research with commercial potential. To control for these aspects in our analyses, we developed two additional language-based models that generate measures for the social and scientific impact potential of scientific findings. These models follow, overall, the same structure and reasoning as the primary model described above. The particularities of each model follow.

[17] TF-IDF (Term Frequency - Inverse Document Frequency) is an NLP method measuring how important a term is within a document relative to a broader corpus. TF-IDF is a simpler method than BERT and, thus, it is expected to result in a lower performance.

[18] doc2vec is yet another algorithm to perform NLP tasks. While simpler than BERT, it is able to capture semantic relationships between a group of words.

3.6.1 Scientific potential model

The scientific potential model is trained using the exact same sample as the primary model. In this case, we use the academic citations accumulated by a paper as the variable for classification, i.e., we use academic citations as a proxy for scientific potential. The model is trained with a balanced sample, and we use the median citations of the training sample as the class cut-off. Specifically, papers cited less than or equal to 16 times are labeled as *low scientific potential*, and papers cited more than 16 times are labeled as *high scientific potential*.[19] The model performs reasonably well, with an accuracy and AUROC of 0.67.[20]

3.6.2 Potential social impact model

To our knowledge, there are no existing data that would allow us to capture the social impact potential of scientific findings through a language-based model. To address this challenge, we first conduct an elicitation exercise, in which thousands of humans evaluate scientific abstracts based on social and commercial dimensions, and create a social impact measure. We then validate our social impact measure and use this novel dataset to train an NLP model that predicts the social potential of a finding. In summary, we proceed as follows.

To create a balanced subsample for the elicitation and NLP exercise, we first randomly sample 1,200 papers from our primary dataset. These papers are from 2012 and span across the following fields: Physics, Biology, Computer Science, Electrical Engineering, Materials Science, and Mechanical Engineering. We select papers from 2012 for two reasons. First, since we use patent-to-paper citations to measure a paper's commercial impact, we need a time frame that allows the patent citations to be informative. Second, social relevance is likely to change over time, and we want to train our model with current pressing social chal-

[19]

$$y_i = \begin{cases} \textit{high scientific potential}, & \text{if paper citations}_i > 16, \\ \textit{low scientific potential}, & \text{if paper citations}_i \leq 16 \end{cases} \tag{3}$$

[20]We experimented with different label cut-offs and settings. We prioritized 1) using the same training sample as with the primary model, 2) having a balanced training sample, and 3) setting a cut-off significantly above 0 citations so that the *low* class includes papers that could potentially lead to value.

lenges. To ensure variation across fields and commercial outcomes, we balance the sample. We draw the same number of articles across the six fields, i.e., 200 per field. We also ensure a balanced distribution of commercial outcomes within each field. Specifically, we randomly sample 1/3 of the papers with zero patent citations, 1/3 with one to three patent citations, and 1/3 with four or more patent citations. These restrictions produce a representative sample, which is necessary to train a model that can generalize well.

Second, we conducted an elicitation exercise to create a human-based measure of social impact. We recruited 10,015 human evaluators to participate in the study. Each evaluator was presented with three abstracts randomly selected from the pool of 1,200 articles and was asked to rate the abstracts across 11 questions on a Likert scale from 1 to 5 (see Table C.1, Appendix C). We developed a software application specifically for this purpose, which allowed us to capture data on different aspects. Results were averaged across each question, de-meaned, and normalized to account for inter-subject variation. Each paper was evaluated on average by 30 different humans.

Finally, we train an NLP model using the human-based social impact potential measure with the characteristics described in the previous sections. There are two classes: one associated with high social impact potential and one with low potential. Since our human-based variable of social impact potential is de-meaned and normalized, we use zero as the cutoff to distinguish between the two classes. Observations with a measure lower than or equal to 0 are classified as *low social potential*, and observations with a value greater than 0 are classified as *high social potential*.[21]

We use the test set to analyze how the model performs. The performance in this task is high, with accuracy, precision, recall, and AUROC of 0.86. Furthermore, Tables C.1 and C.2 (Appendix C) explore how the measure of social impact relates to actual commercial outcomes.

[21]
$$y_i = \begin{cases} \textit{high social potential}, & \text{if social impact}_i > 0, \\ \textit{low social potential}, & \text{if social impact}_i \leq 0 \end{cases} \tag{4}$$

4 Evidence from Technology Transfer Office at a US University

In this section, we apply our measure to explore the progress of scientific ideas in the commercialization process at a leading U.S. private university's Technology Transfer Office (TTO).[22] We will examine the relationship between our measure of commercialization potential with whether an article tied to an invention is disclosed to the TTO, which inventions are the object of TTO investments, and whether the inventions are licensed, subject to agreements, and earn revenues. In addition to exploring the role of commercialization potential, our analysis will also consider that of other factors, including the invention's scientific potential, social impact potential as well as the inventors' prior experience with commercialization and with the TTO itself.

4.1 What findings do scientists disclose?

We begin our analysis by examining the disclosure process. We examine disclosures because disclosures are the "top of the funnel" for the technology transfer process that is managed by the TTO. As discussed in the theoretical section, scientists make the decision to disclose their invention to the TTO based on their own criteria, those of their team, Principal Investigator, or research managers. We hypothesize that, with accurate judgment and adequate incentives, scientists would disclose inventions with high commercial potential to the TTO, and our

[22] It is worth noting, again, that our exercise is based not on realized outcomes, but on the latent commercial quality of a finding detached of any other factor affecting performance, i.e., on our ex-ante measure of commercial potential. For example, some related, recent studies, particularly with TTO data, are based on realized outcomes (e.g., Hsu and Kuhn, 2023). While this work is certainly informative, the approach does not allow us to understand whether scientists and organizations are pushing ideas with high potential commercial forward or selecting on some other criteria. For instance, if better teams tend to work on potentially better ideas, the observed outcomes are conflated: What is driving the results, the team, or the idea a team selects on? Likewise, other approaches that do control for the underlying quality of ideas (e.g., Marx and Hsu, 2022; Shvadron, 2023) can shut down the potential effect and provide unbiased estimates for the variable of interest but, as a result, exploring the role that commercial potential plays is more challenging. For example, are teams equally important for high, mid, or low-potential ideas? Or do they matter most for some sections of the distribution? While it is safe to assume that high-potential ideas are, on average, pushed forward, we do not have empirical evidence. Likewise, a lack of measurement limits the interpretation of results regarding friction in the commercialization process of science.

results support this hypothesis.

Figure 1 presents a density plot for the primary variable of interest—commercial potential of a scientific article. The Figure reports both the density for the entire sample of papers associated with the university and for those articles tied to inventions disclosed to the TTO. These are the main variables used in the analyses that follow. In analyzing the disclosure process, the paper is the unit of analysis. In analyzing TTO process and outcomes, it is the invention that is the unit of analysis. Table 4 summarizes the correlations across the TTO main variables of interest.

[Figure 1 about here.]

[Table 4 about here.]

Table 5 provides the correlations between an article's commercial potential and the decision to disclose. We first divide articles into four quartiles based on the commercial potential measure and then calculate the probability of an article being disclosed to the TTO, conditional on being in the low potential (first quartile), mid-low potential (second quartile), mid-high potential (third quartile), and high potential (fourth quartile) categories. For an article associated with the first quartile, or *low potential*, the probability of being disclosed is 2.59%, while for an article associated with the fourth quartile, or *high potential*, the probability is 19.89%, which is 7.68 times higher.

[Table 5 about here.]

Formally, we estimate the following linear probability model:

$$disclosure_i = \beta_0 + \beta_1 \phi_i + \beta_2 \psi_i + \beta_3 \omega_i + \beta_4 \alpha_{it}^{hs} + \theta_{it} + \epsilon_i, \tag{5}$$

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. 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, where Model 1 presents the baseline impact of the fixed effects on disclosure, and Model 2 shows that the measure of commercial potential of a scientific finding is a robust predictor of disclosure. The variation explained by the year-field fixed effect increases by 129%, from 0.027 to 0.062. Likewise, an increase of one standard deviation (0.21) from the median score of the commercial potential measure (0.56) is associated with a 57% increase in the probability of disclosure. Models 3, 4, 5, and 6 include controls for scientific potential, social impact potential, and author scientific experience (h-index). Scientific potential does not appear to have a significant relationship with disclosure, whereas social impact potential demonstrates a positive and statistically significant relationship. These findings imply that scientists also consider the potential social impact of their inventions when deciding whether to disclose them, aligning with previous research on scientists' motives to commercialize their inventions (Merton, 1973; Cohen et al., 2020). Model 6 represents the main specification (equation 5). Even after accounting for other factors that might affect disclosure decisions, a higher commercial potential score remains associated with a greater likelihood of disclosure, with a 56% increase in the probability of disclosure for a 1-SD increase in the commercial potential score, holding constant the median scores of the scientific and social impact, and scientific experience (h-index). A question arising from these results, however, is whether the potential social impact, which appears to condition the top of the TTO funnel, affects the TTO's decisions about which invention disclosures they should try to commercialize. We address this question in the next subsection.

[Table 6 about here.]

4.2 What inventions does the TTO invest in?

Next, we investigate whether the commercial potential measure, as anticipated, predicts investment by the TTO.[23] We also explore whether factors other than commercial potential are associated with the TTO's commercialization decisions, including the scientific prominence of the authors, their experience with commercialization (Hsu and Kuhn, 2023), and the field of the invention (Cohen et al., 2020).[24]

We examine two critical decisions that determine whether an invention will be pursued for commercialization. First, we consider the amount of investment an invention receives from the university's 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. As such, we expect inventions based on more commercially promising science to receive greater investment. Second, we observe the number of patents the TTO files for a given invention, which serves as another proxy for the TTO's expectations about an invention's value.

In Table 7, we examine the factors associated with, respectively, the level of TTO investment and the number of patents filed by the TTO, and Figure D.1, Appendix D, provides a visual interpretation. Since both distributions are skewed, we log transform the variables. On the right-hand side, besides our main 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 commercialization potential. These variables are included to account for the possibility that TTO managers invest in more experienced teams to reduce their investment risk rather than in high-quality ideas. We also control for the scientific potential and social impact potential of the sci-

[23] As discussed in the data section, the analysis is conducted at the invention level, i.e., we aggregate (average) the measures of the different articles associated with an invention.

[24] Other elements, such as the university's incentive system, primarily established through the royalty structure and TTO managers' compensation, may also affect the process (Lach and Schankerman, 2008; Belenzon and Schankerman, 2009).

ence associated with the invention, and for the authors' scientific experience (h-index). The econometric specification is as follows:

$$y_j^{outcome} = \beta_0 + \beta_1\phi_j + \beta_2\alpha_{jt}^{tto} + \beta_3\alpha_{jt}^{tto}\phi_j + \beta_4\psi_j + \beta_5\omega_j + \beta_6\alpha_{jt}^{hs} + \Theta_{jt} + \epsilon_j, \tag{6}$$

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 7 presents the results. As expected, high-potential ideas attract more investment (Model 1): A 1-SD increase (0.14) in commercialization potential from the median (0.74) is associated with an additional $115,000 of investment. The results remain robust in the main specification—equation 6—after controlling for an invention's scientific and social impact potential and for the authors' scientific experience (Model 4). Notably, the study finds that the TTO's investment decisions are mainly driven by an invention's commercial potential, with little attention paid to its scientific potential or social impact potential. This suggests a possible mismatch between the types of ideas that scientists disclose and the ones that the TTO pursues. Additionally, the scientific prominence of the authors, as measured by their scientific H-index, does not appear to play a role in the TTO's decision-making process.

We also find that authors with prior commercial experience with the Technology Transfer Office (TTO) are more likely to receive investment. Several reasons might explain this finding. For example, these authors often have a proven track record of success, established relationships with TTO staff, and a better understanding of the commercialization process, which makes them more appealing for investment. Moreover, their experience may enable them to identify commercially viable projects and present their inventions effectively, enhanc-

ing their credibility and attractiveness to the TTO. Finally, learning effects from previous experiences could improve their ability to navigate the commercialization process, making their current and future projects more promising for investment. It is notable, however, that the coefficient substantially reduces and looses significance when regressed with the commercial potential measure. Likewise, the inclusion of this variable reduces the magnitude of our coefficient on Commercial Potential in the investment regression from 6.164 to 5.694—a decrease of only 8.25%. Further, the inclusion of these variables in the Patents regression reduces the magnitude of our coefficient on Commercial Potential in the investment regression from 1.083 to 0.931—a decrease of a 14.03%. A question posed by these results is whether authors' TTO experience actually enhance an invention's commercialization potential in a way that our measure does not reflect, per the arguments above. If the measure is, however, reasonably comprehensive, then the results may imply that author TTO history may distort the TTO's commercialization decisions.

Lastly, we introduce an interaction term between 'Author TTO Experience' and our commercial potential measure to investigate whether high-potential ideas from experienced scientists are more likely to result in greater investment and protection. Surprisingly, this interaction is not statistically significant in analyzing how much money the TTO invests to push an invention forward or the decision to patent. Although we should interpret this insignificance cautiously, it could imply that the combined value of explicit knowledge in the article and tacit knowledge (pertaining to both the science and the commercialization process) may not significantly influence investment and protection at the TTO level. In other words, there may not be increasing returns to potential when scientists possess more experience.

[Table 7 about here.]

4.3 What inventions generate commercial outcomes?

We now examine the relationship between our commercial potential measure and commercial outcomes tied to firms or other organizations. We estimate a similar specification as that of equation 6, including the amount invested in invention j by the TTO as a control. Table 8 presents the results for commercial agreements and licensing deals.[25]

According to models 1 and 5, the commercial potential measure alone is highly predictive of an invention leading to both agreements and licensing deals. Models 4 and 8 estimate the complete specification, controlling for the prior experience of the authors with the TTO and its interaction with commercial potential, the scientific and social impact potential, the scientific experience of the authors, as well as the TTO investment. For both agreements and licensing deals, the TTO investment appears to mediate the effect of commercial potential, suggesting alignment, on average, between TTO decisions, those of firms undertaking the agreements, and commercial potential; downstream players are incorporating this information into their decisions.

Finally, the authors' prior experience with the TTO appears relevant when it comes to agreements and licensing deals, which reflect firms' decisions, as opposed to the previous analysis, which shows that TTO decisions on investment and patenting are little correlated with this experience. In both models (4 and 8), the interaction term is positive and significant, suggesting that firms, in entering deals with the TTO, do pay attention to the combination of the commercial potential and the authors' past involvement in similar processes. There are several reasons why this might be the case. In addition to the reasons we listed before reflecting demand-side considerations (e.g., firms' abilities to identify and present promising ideas), the interaction result suggests a combination of demand and supply-side explanations. Previous studies suggest, regarding faculty inventor participation, that if researchers are involved in developing university inventions, these are more likely to commercially succeed (Agrawal and Henderson, 2002; Jensen and Thursby, 2001; Thursby and Thursby, 2002), probably

[25]Figures D.2 and D.3, Appendix D, provide a visual interpretation.

due to the tacit knowledge required to move forward. If that is the case, on the demand side, firms will look not only for inventions with high commercial potential but for those inventions that are also from researchers willing to work on them. Although speculative, we argue that the prior experience coefficient could thus be capturing the motivation of these researchers, as signaled by their prior activity, and thus the greater predisposition of firms to close deals with those who have past TTO experience, conditional on high commercial potential.

[Table 8 about here.]

Table 9 examines whether the measure also predicts revenue, and commercialization via a startup and, conditional on startup formation, whether it predicts venture capital funding. While the coefficient for revenue is positive (model 1), it fails to be statistically significant. This may be due to a lack of variation in revenue, especially considering the small number of observations with positive revenue and a large number of fixed effects. It may also reflect the fact that we trained our model on patents renewed a single time, which is some distance from the realization of revenue. Both startup formation and VC investment are associated with the commercial potential of an invention. What is also notable about these results, in common with the prior set, is the continued strong positive relationship between outcomes and the author's TTO experience, which is consistent with a common understanding among TTO officers that the faculty inventor plays a key role in supporting TTO technology transfer efforts.

[Table 9 about here.]

Collectively, the findings presented in Tables 5 through 9 indicate that our measure of commercial potential can effectively predict progress through the commercialization process at this university. Although we do not claim that it is a comprehensive measure of commercial potential—encompassing the complete spectrum of information about whether a scientific

advancement will succeed commercially—it appears to be meaningfully distinct from other frequently used measures in previous literature. Such measures include an author's prior commercial impact, their experience in commercialization, and their scientific impact.

4.4 The TTO's efforts and the realization of the commercial potential of science

Until now, we have investigated whether our measure predicts the disclosure of scientific findings to the TTO, TTO commercialization efforts, and commercial outcomes. In this section, we examine, from this one TTO's perspective, the extent to which there is unrealized commercial potential in its university's research.

In order to do so, we first have to establish what constitutes commercial potential in a binary way. Note that, so far, we have used as a measure of commercial potential a continuous variable that, while extremely helpful in the analysis conducted, does not allow for analyzing at the institution level the extent to which there is unrealized scientific research, i.e., research that is commercially promising but that is not commercialized. Defining a threshold for commercial potential is somewhat arbitrary and thus, we adopt an empirical strategy based on the threshold tied to the commercialization of the inventions for this one TTO. We set the threshold as 0.81, which corresponds to the top quartile commercial potential of the publications associated with inventions experiencing any type of commercial outcome (agreements, licensing, revenue, startup, or VC funds). That is, we assume that a publication has commercial potential if its score is 0.81 or higher and has no commercial potential otherwise. During the period 2010 to 2022, 7,909 of the articles produced at the university are classified as commercially promising (8.51& over a total of 92,972).

We find that only 26.25% of high potential articles are linked to inventions disclosed to the TTO. Although a preliminary figure requiring further scrutiny, this suggests that the TTO is unaware of the majority of articles with high commercial potential. We also see that, given an invention disclosure, inventions whose associated articles are classified as

having commercial potential (408), only 57.11% are the objects of TTO investment. Thus, we find evidence suggesting that, indeed, there is unrealized commercial potential of scientific articles, at least from the viewpoint of the TTO. This realization gap may be due to a range of factors. One likely explanation is that, though commercially promising, the scientific finding or idea is still too embryonic. Other possibilities include a lack of motivation on the part of academics to expend any effort around commercializing their research, especially if their prime motivation is academic advancement and eminence, suggesting misaligned incentives between researchers and the technology transfer office. Finally, the result that only about half of the inventions disclosed are pursued by the TTO is not surprising when one considers other settings, such as the percentage of ideas presented to VC's that are not pursued by them. It also makes sense if, per the results above, the TTO culls disclosures that may be motivated by something other than commercial potential, such as social impact.

The fact that a sizable percentage of articles with high commercial potential is not disclosed to the TTO raises the question of the extent to which the knowledge contained in such articles, through publication and other forms of public dissemination, is nonetheless commercialized by firms.[26] Using the citations of articles in corporate patents to reflect the use of articles by firms in their commercialization efforts, we find that 90.79% (5,296 of the 5,833) of the undisclosed articles with high potential are not cited in corporate patents, suggesting that less than a fifth of the undisclosed high-potential science is commercialized through channels outside of the TTO. Does that mean that the knowledge contained in the 90.79% of uncited articles is lost to the cause of technical advance? We believe that such a claim cannot be made and strikes at a limitation of the present analysis—that is, the assumption that citation of an article in a corporate patent reflects, or at least strongly relates to, use in firms' innovative activities of the knowledge contained in that article.

A related question is the extent to which firms use articles that are linked to inventions

[26]Addressing a related question, prior research (Cohen et al., 2002) suggests, in fact, that the most significant channels through which public research becomes an input for industrial R&D labs are just such channels of public dissemination.

that are disclosed but not pursued by the TTO. Of 4,445 such articles, we find 93.09% are not cited by a corporate patent, and thus 6.91% are cited. Two factors may drive this finding. First, articles tied to an invention disclosure are probably less embryonic (i.e., in a form that is closer to a market application), and second, the very process of disclosure may increase the visibility of the articles. But a counterargument to these rationales is that, if the TTO actually chooses to pursue an invention and possibly patent it, would that not foreclose or at least inhibit the use of that knowledge by firms through channels outside of those controlled by the TTO?

To probe this further, we present the results from two additional analyses. At the article level, in Table 10, the results clearly suggest that disclosure makes the citation of the associated article in a corporate patent more likely. At the invention level, and thus conditioning on disclosure, in Table 11, we also see, however, that the TTO's patenting of the invention appears to have no relationship with the likelihood of the citation of any of the associated articles in a corporate patent. For both of these results, however, we need to be cautious in our interpretation because, as our arguments suggest, countervailing factors may be at work.

[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 potential commercializability 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.]

In Figure 3, a visual representation is provided to show the proportion of articles from the TTO we examined that have been incorporated into an attempted commercial use. For this analysis, three indicators of an attempted commercial use are utilized: (a) disclosure of the article to the TTO; (b) citation of the article by a corporate patent; (c) citation of the article by a corporate patent *and* disclosure to the TTO. Among the articles judged to possess high commercial potential, 11.7% (6.8% + 2.5% + 1.5% + 0.9%) of the articles published in 2010 by researchers at this university demonstrate this potential. From this subset, 41.9% (4.9/11.7%) have been associated with commercial use. In contrast, for articles deemed to have low commercial potential, which comprises 88.4% of the articles published in 2010, only 12.4% (7 + 3.1 + 0.9)/88.4%) have been associated with any attempted commercial use.

These data can be interpreted in several ways. First, it seems that our algorithm can effectively distinguish between ideas with higher or lower prospects for commercial use, as demonstrated by the significant difference in commercial use rates between the blue (representing low potential) and orange (representing high potential) articles: 12.4% compared to 41.9%. Nevertheless, the data also indicates that a substantial fraction, 58.1%, of the high-potential ideas are neither disclosed to the TTO nor cited by corporate patents. This discrepancy might be due to measurement errors in our method of quantifying commercial potential. Alternatively, this realization gap might also stem from challenges that firms and technology transfer offices face in identifying commercially viable ideas among the multitude of papers and projects this university produces. In this way, this figure of 58.1% of high-potential papers or 6.8% of all papers in 2010 serves as the upper bound of 'unrealized' commercial potential at this university—or the realization gap.

[Figure 3 about here.]

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

This section evaluates the commercial potential and eventual realization of scientific findings published in the United States between 2010 and June 2022. We aim to identify contexts and factors associated with articles possessing high commercial potential that do not achieve realization (i.e., cited by a patent that is eventually renewed). Our analysis draws from our primary dataset containing articles where at least one author was affiliated with a U.S. institution at publication. The dataset encompasses 3,936,673 articles distributed over eleven 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. To gauge the commercial potential of the published research, we analyze the abstracts using the NLP model we developed and calculate both the scientific and social potential indicators.[27] Table 12 displays the sample's summary statistics, while Table 13 reveals the correlations among the primary variables of interest.

[Table 12 about here.]

[Table 13 about here.]

First, we investigate whether the realization rate of articles (their transition to commercial use) varies among institutions. Figure 4 (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 2010-2015.[28] In this figure, a publication's citation in a patent serves as a comparable indicator of its commercial "use". Overall, 16.6% of articles from top institutions get cited by at least one patent,

[27] Before inputting the abstracts into the models, we preprocess the texts as delineated in section 3.2.

[28] 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.

whereas only 13.1% of articles from bottom institutions do. Notably, there is a marked difference (26.5% 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 rates among institutions vary considerably across both institutions and fields—ranging from 23% to 43%, with some fields having significantly larger realization gaps than others when viewed across institutions.

However, Panel A encompasses publications with both 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 4 (Panel B) is refined by considering only those papers identified as possessing high commercial potential (i.e., those in the top quintile 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, 45.1% of articles from top institutions and 40.3% 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 11.5% from the earlier 26.5%. Finally, despite the high commercial potential classification, a significant 55% of the research, even at larger and more research-intensive universities, does not transition to commercial use. To further examine these differences,

Figure 5 illustrates the variation in the production of high commercial potential research for selected universities, and 6 illustrates the variation in the realization rate of articles when considering the relevant risk-set of research: publications with high commercial potential. These figures 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 institutes, also with active technology transfer offices, have far fewer: the University of Hawaii has approximately 2% whereas the University of Arizona has 6%. However, in Figure 6, Panel B, we see that once we look at commercial use among high potential science, the variation across universities diminishes substantially.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

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

$$y_i = \beta_0 + \beta_1\phi_i + \beta_2\psi_i + \beta_3\omega_i + \beta_4\iota_i^{c80} + \beta_5\phi_i \times \iota_i^{c70} + \beta_6\iota_i^{s70} + \beta_7\phi_i \times \iota_i^{s70} + \theta_{jt} + \epsilon_j, \quad (7)$$

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. 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 results from this analysis align with the preliminary findings gathered from an initial visual inspection: research-intensive institutions consistently exhibit a greater propensity to bring high commercial potential research to fruition than their less research-intensive counterparts. Specifically, the interaction term 'Commercial potential x High commercial impact institution' is positive and profoundly significant even when factoring in year and field fixed effects. This interaction indicates that research published at research-intensive institutions and possessing high commercial potential has a 10.8% increased likelihood of receiving a patent citation compared to other research. This finding corroborates the disparity described above. To further examine the source of this disparity, we study two potential impediments that could hinder high commercial potential research from achieving realization. Specifically, the roles of individual researchers and the journals in which their findings are published are considered. Both research conducted by less visible scholars (or newer ones) and work published in less commercially visible journals may be ignored by firms, thus limiting their commercial use. The outcomes of these models, as showcased in Model 3 of Table 14, suggest persistent institutional differences in commercialization rates. Research boasting high commercial viability and emanating from elite institutions stands an 8.2% higher chance of being cited in patents than similar research from less prestigious institutions. This is observed even after accounting for specific attributes linked to researchers and journals. That said, research conducted by high-impact scholars has a 15.4% higher likelihood of patent citation. Relatedly, studies in journals that lean towards commercial application have a 19.9% increased probability of being cited in a patent.

[Table 14 about here.]

Table 15 explores the relationship between commercial potential and commercial value beyond patent citations. To do so, we condition papers cited by at least one patent and use as a dependent variable the average renewal years of the patents citing a focal paper. As previously discussed, we use the years a patent has been renewed to indicate its value (zero, four, eight, or twelve). Since a paper can be associated with more than one patent,

we aggregate the years citing patents are renewed at the paper level by averaging them. Our findings are aligned with the previous analysis. Papers from institutions at the top are more likely to be cited by patents that, in turn, are renewed more years, i.e., are more valuable.

[Table 15 about here.]

Arguably, if an article has high commercial potential, there should be no difference in the likelihood of commercial realization. Thus, we interpret the findings of disparities in realization rates in our regressions and Panel B of our figure as suggesting that firms employ heuristics, based on factors such as a university's research standing, author's standing, or journal prominence, to guide their search for commercially useful science. This, in turn, implies that science with commercial potential from lesser-ranked institutions, authors, and journals will be overlooked to society's detriment. Of course, in the face of the volume of science produced and associated costs and uncertainties, employing such heuristics is not necessarily irrational. But it comes at a cost to social welfare.

6 Discussion

In this paper, we argue that for the innovation ecosystem to function efficiently, firms need to be able to identify and build their technologies, products, and services upon the most commercially promising scientific findings. However, there is reason to believe that identifying and exploiting new knowledge could be more efficient and comprehensive. One possible explanation could be the increasing division of innovative labor between knowledge-creating institutions and firms, in which private companies increasingly rely on public research as an input for their R&D (Arora et al., 2020; Fleming et al., 2019).

Indeed, researchers have explored the frictions in translating science into economically viable solutions, i.e., moving scientific findings from labs to markets. Sources of friction include field-level differences in technology and market risk (Arora et al., 2022), organizational search frictions (Bikard and Marx, 2020), firm incentives (Cassiman and Veugelers, 2002),

commercialization experience (Azoulay et al., 2007; Stuart and Ding, 2006), and behavioral issues such as a misalignment between scientists and investors' motives (Cohen et al., 2020) and judgment biases faced by both actors, as well as by individuals within organizations responsible of pushing science forward (Kline and Rosenberg, 1986; Rosenberg, 1976).

While extensive evidence points to enablers and frictions in the ecosystem, we know little about the extent to which scientific findings with promising commercial potential are commercialized, whether the commercialization rate varies at different stages of the process, and whether the rate varies across institutions, and especially whether and where there may be unrealized commercial potential. Most importantly, and explaining why these issues have not been studied in greater depth, we need a scalable, ex-ante way to measure the commercial potential of scientific findings.

To better understand these questions, first, we develop a novel methodology that constructs a scalable, ex-ante measure of commercial potential. Our approach utilizes natural language processing and Bidirectional Encoder Representations from Transformers (BERT) fine-tuned with scientific articles (SciBERT). Additionally, we fine-tuned an NLP model for classification using nearly 100,000 scientific articles. This model can accurately predict whether a scientific article will be cited by a renewed patent, indicating the potential for commercial value (AUROC of 0.75). We also trained models to capture scientific potential through forward academic citations and social potential using a dataset in which 10,000 humans evaluate social aspects of scientific findings. These models allow us to control for non-commercial aspects of innovation and provide more nuanced results.

Our analysis suggests that our measure of 'commercial potential' captures essential information about the likelihood of a scientific idea achieving commercial use. Using detailed technology transfer data from a leading research university to validate our measure, we can predict which science is disclosed, investments by the TTO, patent filings, startup formation, and licensing. In this context, as highlighted by our analysis, the primary barrier to commercialization is either the scientists' decision to disclose their discoveries or the TTO's

awareness of pertinent researchers or science within their institutions. Thus, many high-potential articles remain undisclosed, and undisclosed science is much less likely to be cited in corporate patents. Our second analysis studies this problem at a larger scale, using data on nearly 4 million articles published between 2010-2022 by authors with a US-based affiliation. We find considerable heterogeneity in the proportion of all articles identified as having 'high commercial potential' across US institutions. This difference in the distribution of commercially valuable science drives meaningful differences in commercialization rates across universities. However, even when we look only at high commercial potential ideas across institutions, meaningful differences in rates of commercial use (e.g., patent citations and renewals) exist between the most and least research-intensive universities and institutes. These differences are institution-centric and cannot be explained by field differences, temporal trends, or differences in the commercial prominence of individual researchers and journals. These differences suggest that selected attributes of institutions, authors, and journals guide firms' search for useful science. In that process, which is likely justified given search costs and uncertainty, potentially commercializable science from lesser-ranked entities and individuals is overlooked.

Our work is subject to several limitations. The most salient limitation is our reliance on patent data to train our model of commercial potential. Specifically, we make two key assumptions. First, we assume that patent-to-paper citations establish a valid linkage that captures whether the science reported in a paper contributes to the development of technology reported in a patent. However, as many have argued, this might not be the case (e.g. Kuhn et al., 2020). As scalable patent-paper pairs datasets based on more accurate matching techniques emerge (e.g., Aslanyan and Wetherbee, 2022; Deltorn et al., 2023), our methodology can be updated. Second, technologies are patented for various reasons, and therefore, a patent only occasionally indicates commercial value (Cohen et al., 2000; Chuang et al., 2023).

Throughout this discussion, there is the underlying question of time horizons, notably

how long it takes for commercially promising research to make its way into commercial application. We understand that moving science from the lab to practical applications can take many years, even when the commercial potential is apparent at the outset. Indeed, from Marx and Fuegi (2022), we learn that the average lag between publication and being cited in a patent is about 14 years. One question is, however, what portion of such lags are endogenous. If science with commercial potential can be identified as such much earlier—indeed, upon publication—a question is whether that value could impact the speed with which science can be applied for societal benefit. Relying on patent data, thus, also carries two additional limitations.

First, most of our training data is based on scientific discoveries that are cited relatively quickly. Given that we observe patent citations as of 2022 and that the time range of the training is from 2000 to 2009, we are conditioning the models on articles cited at least 13 and 22 years later. Therefore, we might not be able to capture the potential of those technologies that, for example, are more embryonic. A second limitation directly related to this point, perhaps more salient, is that we only use patent citations to articles as a measure for realization in the last section, the institutional analysis. In this case, we condition our analysis on articles published after 2010 and, thus, we might not observe in the data scientific research that eventually will be realized, overestimating the realization gap. Noting this, our first part of the analysis focuses on papers published before 2015 precisely to avoid this issue. Likewise, the following regression analyses include year fixed effects and focus mainly on differences across organizations. For the truncation in patent-to-paper citations to be an issue in our relative estimates of the commercialization gap, the lags between science and technology should be distributed heterogeneously across organizations– -which is a possibility.

Similarly, limitations arise in our technique for matching articles to inventions in the TTO exercise. The NLP technique may be prone to measurement error. Even if patents were a perfect proxy for commercial value and patent-to-paper citations were a perfect proxy

for patent-paper pairs, the decision to build a technology upon a specific piece of science goes beyond the text in which the finding is reported. Other considerations, such as cost, risk, or even the expertise and experience of the inventor, may be relevant factors that the NLP model does not capture but that a decision-maker may consider. Given this, one may interpret our measure as reflecting technological opportunity rather than a metric comprehensively characterizing all the factors that condition commercialization, no less commercial success.

Future research can build upon this study by focusing on several critical aspects of the commercialization process. This includes refining the text-based measure of commercial potential, exploring factors that limit commercialization in low commercial-impact institutions, and identifying practices that bridge the gap between science and industry—i.e., promoting the more effective translation of scientific research into tangible commercial outcomes. These and other directions for future research should lead to a deeper understanding of the commercialization process and help maximize the societal benefits of scientific discoveries.

References

Agrawal, A. and Henderson, R. (2002). Putting patents in context: Exploring knowledge transfer from mit. *Management science*, 48(1):44–60.

Arora, A., Belenzon, S., and Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, 39(1):3–32.

Arora, A., Belenzon, S., Patacconi, A., and Suh, J. (2020). The changing structure of american innovation: Some cautionary remarks for economic growth. *Innovation Policy and the Economy*, 20(1):39–93.

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.

Arora, A., Fosfuri, A., and Roende, T. (2022). Caught in the middle: The bias against startup innovation with technical and commercial challenges. Technical report, National Bureau of Economic Research.

Aslanyan, G. and Wetherbee, I. (2022). Patents phrase to phrase semantic matching dataset. *arXiv preprint arXiv:2208.01171*.

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.

Belenzon, S. and Schankerman, M. (2009). University knowledge transfer: private ownership, incentives, and local development objectives. *The Journal of Law and Economics*, 52(1):111–144.

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.

Cassiman, B. and Veugelers, R. (2002). R&d cooperation and spillovers: some empirical evidence from belgium. *American Economic Review*, 92(4):1169–1184.

Chuang, H.-C., Hsu, P.-H., Lee, Y.-N., and Walsh, J. (2023). What share of patents is commercialized? *Working paper*.

Cohen, W. M., Nelson, R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why us manufacturing firms patent (or not).

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.

Deltorn, J.-M., Guellec, D., Wu, C., and Liu, J. (2023). Building a corpus of patents-articles siblings. *Working paper*.

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., Greene, H., Li, G., Marx, M., and Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, 364(6446):1139–1141.

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.

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.

Jensen, R. and Thursby, M. (2001). Proofs and prototypes for sale: The licensing of university inventions. *American Economic Review*, 91(1):240–259.

Klevorick, A. K., Levin, R. C., Nelson, R. R., and Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. *Research policy*, 24(2):185–205.

Kline, S. J. and Rosenberg, N. (1986). An overview of innovation. In Landau, R. and Rosenberg, N., editors, *The Positive Sum Strategy: Harnessing technology for economic growth*, chapter 16, pages 275–305. The National Academies Press, Washington, DC.

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.

Lach, S. and Schankerman, M. (2008). Incentives and invention in universities. *The RAND Journal of Economics*, 39(2):403–433.

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.

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

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.

Rosenberg, N. (1976). On technological expectations. *The Economic Journal*, 86(343):523–535.

Shvadron, D. (2023). Bread upon the water:.

Stuart, T. E. and Ding, W. W. (2006). When do scientists become entrepreneurs? the social structural antecedents of commercial activity in the academic life sciences. *American journal of sociology*, 112(1):97–144.

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.

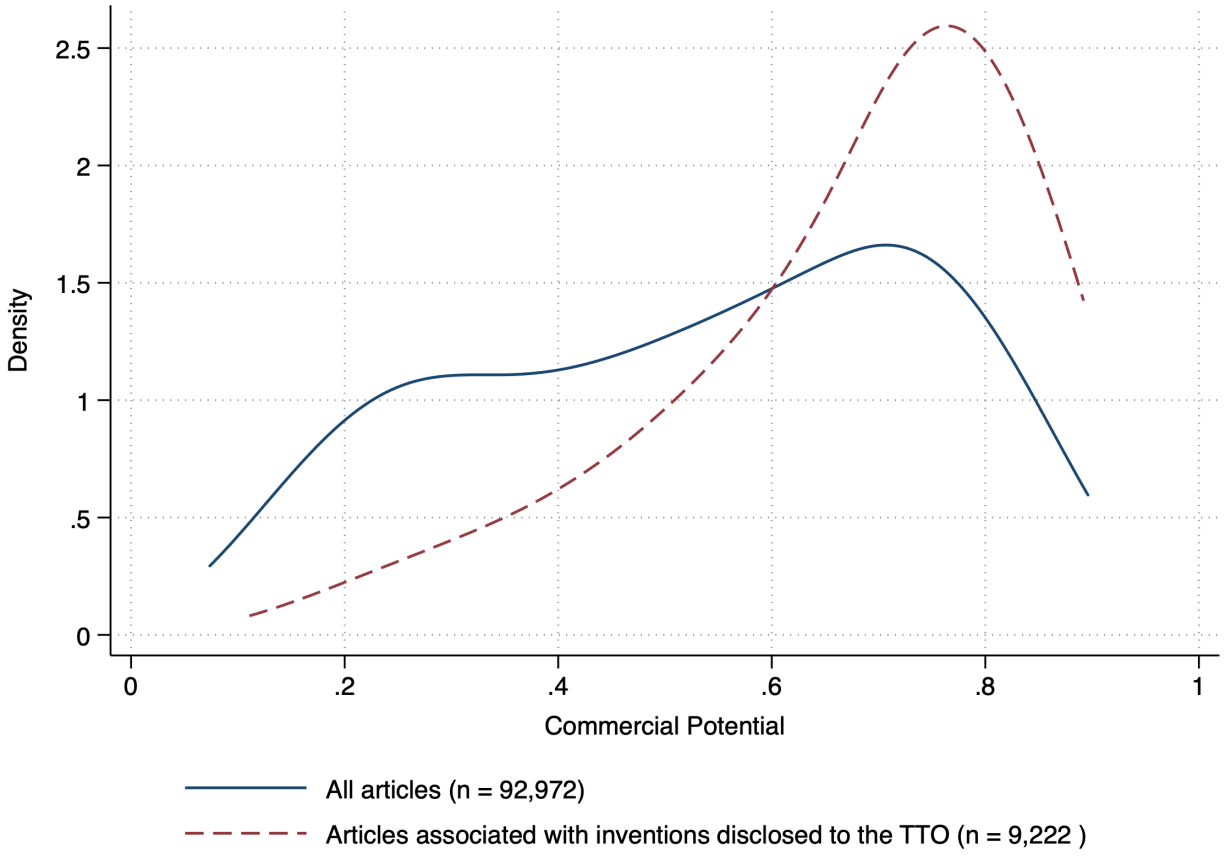


Figure 1: Bi-weight kernel density estimates of the distributions of commercial potential of 1) all articles published at this university (solid line) and 2) only articles associated with inventions disclosed to the Technology Transfer Office (dashed line). Articles tied to an invention are more likely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

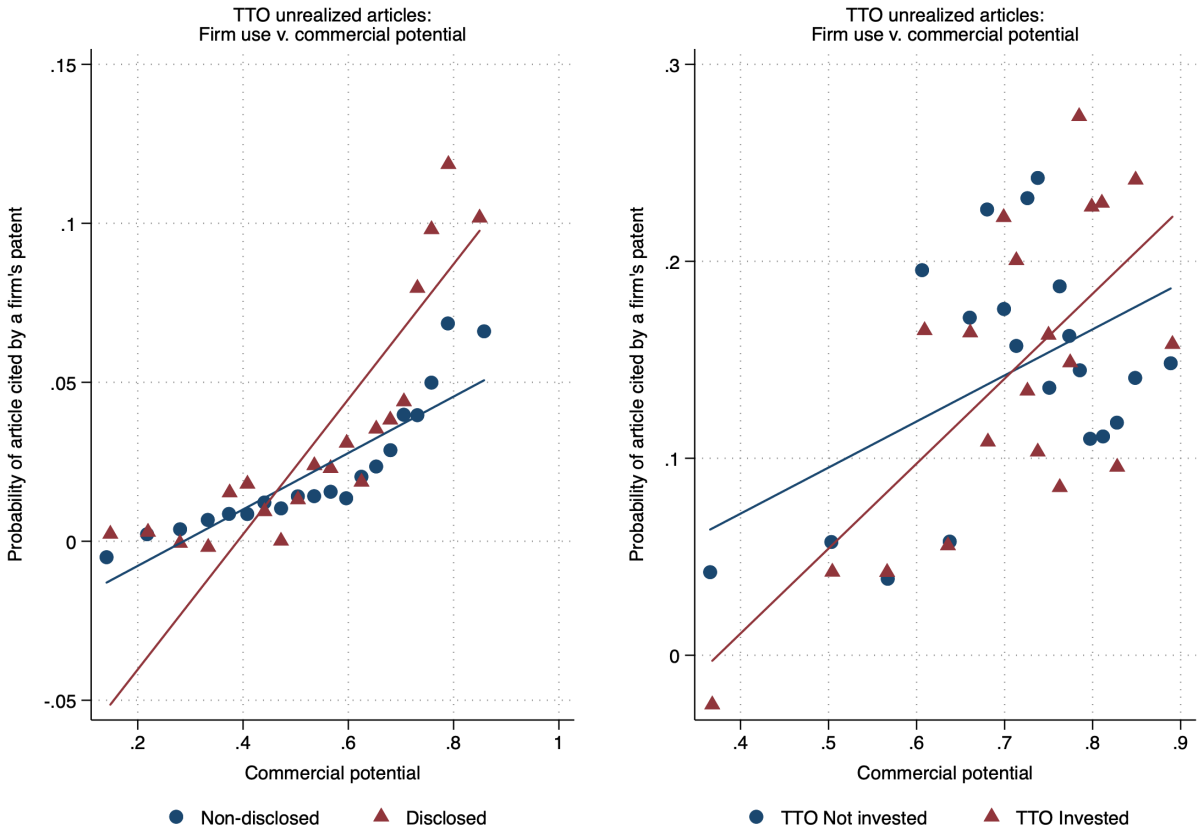


Figure 2: Binned scatterplot representing the probability that a firm's patent will cite articles associated with the TTO university as a function of their commercial potential. Panel A distinguishes articles tied to an invention disclosed to the TTO (triangle) vs. those not (circle). Panel B distinguishes 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.

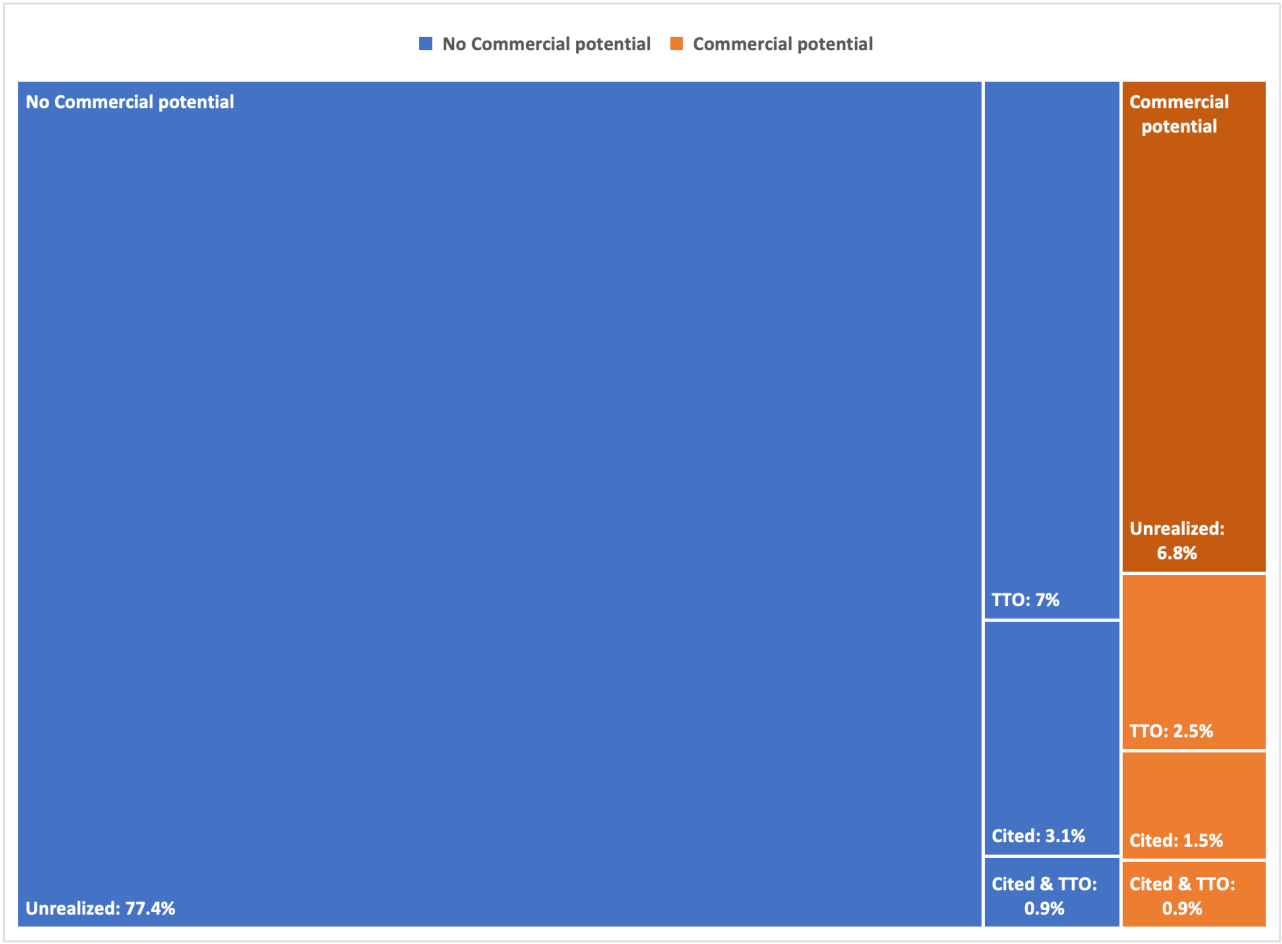
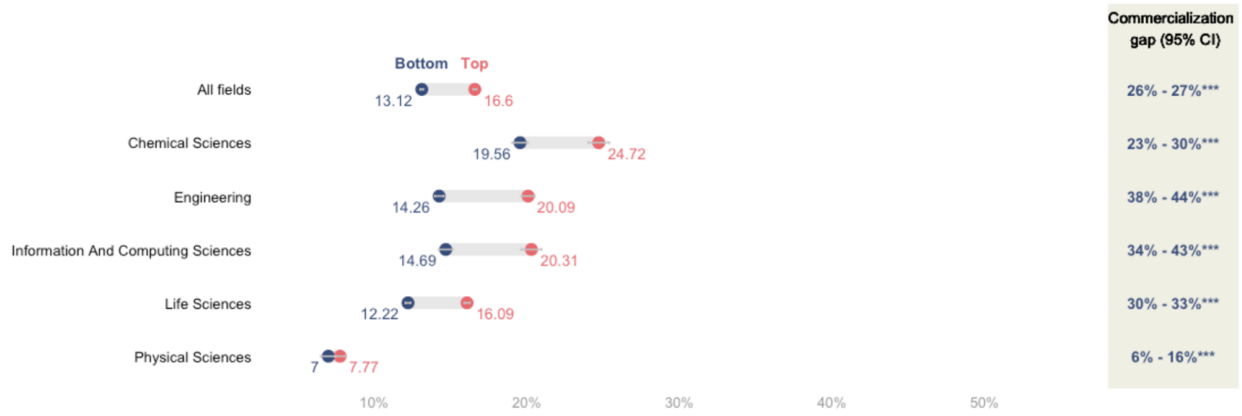


Figure 3: Distribution of articles in terms of commercial potential and realization for 2010. Articles are realized via corporate patent citations (Cited), via disclosure to the Technology Transfer Office (TTO), and via both (Cited & TTO). Unrealized articles are neither disclosed nor cited in corporate patents.

Panel A: Share of articles cited by patents over all articles.



Panel B: Share of articles cited by patents over high-commercial potential articles.

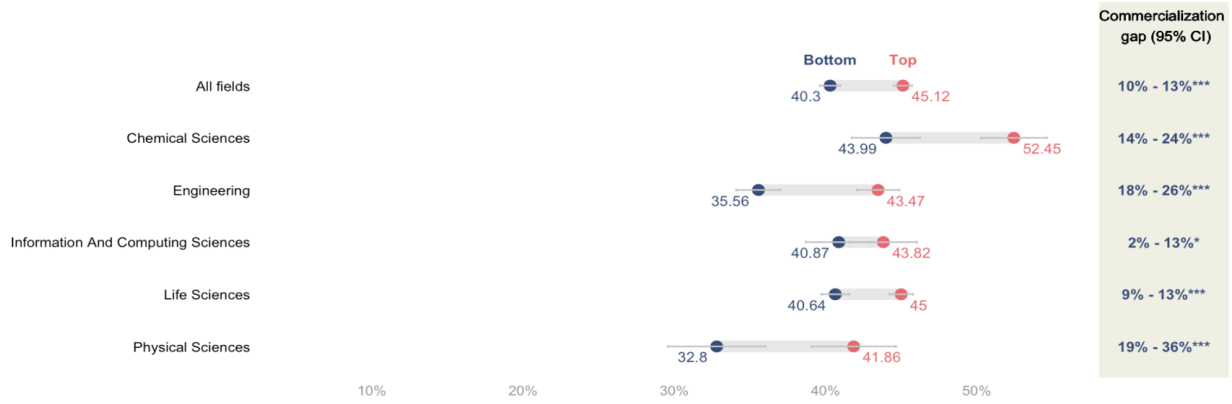


Figure 4: Differences in commercialization rates across institutions. Panel A plots the average rates at which publications are cited by at least one patent in top vs. bottom institutions for all fields and by field. Panel B conditions on articles classified as high commercial potential, that is, the 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.

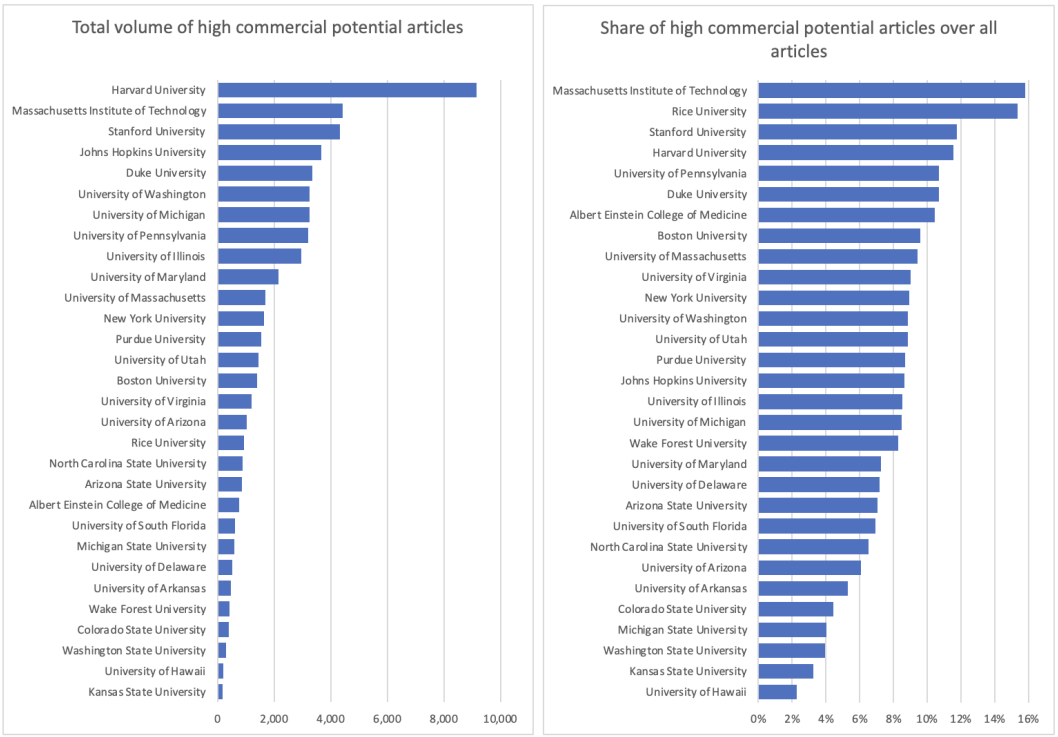


Figure 5: Differences in the production of high commercial potential research. Panel A shows the total number of high commercial potential articles produced between 2010 and 2014 for selected universities, and Panel B shows the share of high commercial potential articles over the total number of articles produced.

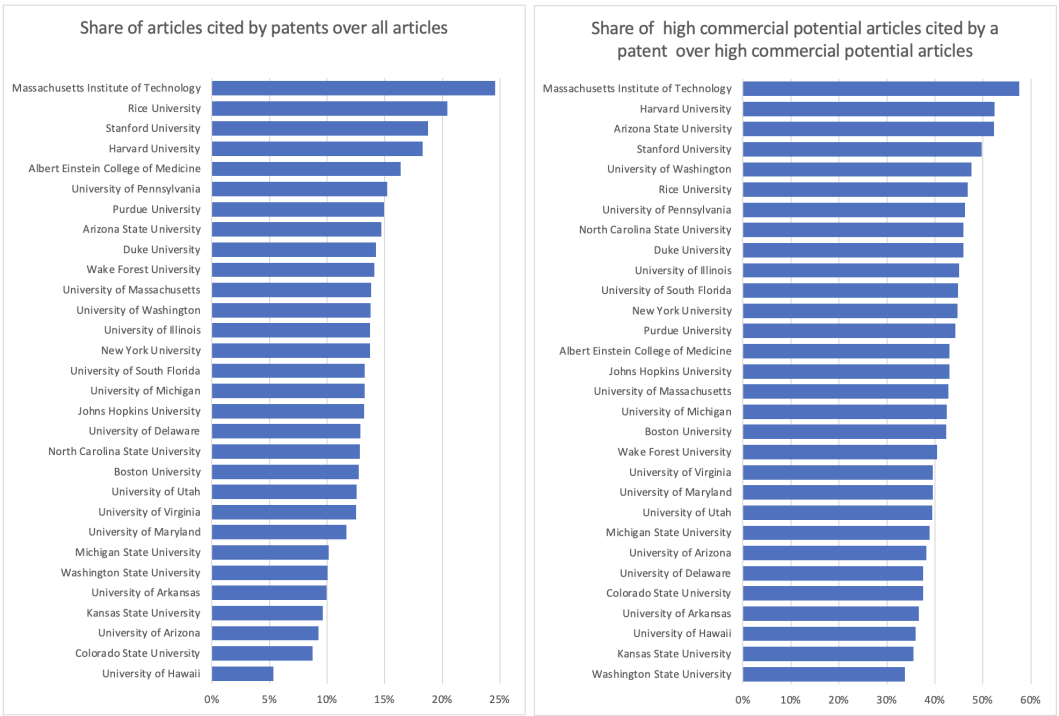


Figure 6: Share of articles cited by patents over all articles (Panel A) *vs.* share of articles cited by a patent over high commercial potential articles (Panel B).

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.
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, 2010-2022. Panel B summarizes the relevant features of the articles *associated* with disclosed inventions, 2010-2022. For confidentiality reasons, invention-level outcomes are removed (Investment, Patents, Agreements, Licensing, Revenue, Startup, and VC funding).

Panel A: All articles

	Mean	SD	Min	Max	N
Commercial Potential	0.53	0.21	0.07	0.90	92,972
Scientific Potential	0.65	0.14	0.05	0.94	92,972
Social Impact Potential	0.73	0.30	0.05	0.97	92,972
Academic cites	36.94	138.90	0.00	19,200.00	92,972
Patent cites	0.28	3.31	0.00	501.00	92,972
Cited by at least one patent	0.06	0.24	0.00	1.00	92,972
Citing patent is renewed	0.03	0.16	0.00	1.00	92,972
Author Scientific H-index	57.94	38.22	0.00	308.00	92,972
Disclosed	0.10	0.30	0.00	1.00	92,972

Panel B: Articles matched
to inventions disclosed to the TTO

	Mean	SD	Min	Max	N
Commercial Potential	0.71	0.14	0.14	0.89	1,829
Scientific Potential	0.67	0.11	0.11	0.92	1,829
Social Impact Potential	0.78	0.23	0.08	0.96	1,829
Academic cites	41.84	81.73	0.00	2,190.00	1,829
Patent cites	1.40	4.95	0.00	79.50	1,829
Cited by at least one patent	0.30	0.46	0.00	1.00	1,829
Citing patent is renewed	0.15	0.36	0.00	1.00	1,829
Author TTO Experience	0.67	0.47	0.00	1.00	1,829

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	0.234	1.000								
Social Impact Potential	0.255	0.054	1.000							
Academic cites	0.082	0.076	-0.022	1.000						
Patent cites	0.094	0.011	0.005	0.313	1.000					
Cited by at least one patent	0.237	0.038	0.025	0.223	0.341	1.000				
Citing patent is renewed	0.164	0.022	0.020	0.192	0.395	0.626	1.000			
Author Scientific H-index	0.227	0.173	0.124	0.125	0.032	0.070	0.043	1.000		
Disclosed	0.214	0.040	0.065	0.023	0.049	0.110	0.089	0.079	1.000	
Author TTO Experience	0.207	0.043	0.055	0.019	0.049	0.115	0.092	0.090	0.819	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	0.252	1.000						
Social Impact Potential	0.088	0.022	1.000					
Academic cites	0.112	0.082	-0.068	1.000				
Patent cites	0.129	-0.002	-0.055	0.366	1.000			
Cited by at least one patent	0.203	0.025	-0.058	0.305	0.434	1.000		
Citing patent is renewed	0.156	-0.011	0.005	0.289	0.516	0.633	1.000	
Author TTO Experience	0.239	0.083	0.015	0.064	0.047	0.083	0.060	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 7.68 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	22,642	601	23,243
	97.41%	2.59%	100.00%
2	21,721	1,522	23,243
	93.45%	6.55%	100.00%
3	20,767	2,476	23,243
	89.35%	10.65%	100.00%
4	18,620	4,623	23,243
	80.11%	19.89%	100.00%
Total	83,750	9,222	92,972
	90.08%	9.92%	100.00%

Table 6: Linear probability model estimating the likelihood of disclosure as a function of commercial potential. Model 1 present 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		0.290***			0.275***	0.267***
		(0.016)			(0.016)	(0.016)
Scientific Potential			0.119***		-0.004	-0.012
			(0.010)		(0.008)	(0.009)
Social Impact Potential				0.107***	0.043***	0.043***
				(0.008)	(0.006)	(0.007)
Author Scientific Experience						0.011***
						(0.003)
Constant	0.099***	-0.056***	0.022***	0.021***	-0.077***	-0.109***
	(0.000)	(0.008)	(0.006)	(0.006)	(0.009)	(0.012)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	92,972	92,972	92,972	92,972	92,972	92,972
R-squared	0.027	0.062	0.029	0.034	0.063	0.064

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

Table 7: 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	6.164***		5.694***	5.070***	1.083***		0.931***	0.853***
	(0.813)		(1.249)	(1.355)	(0.136)		(0.167)	(0.173)
Author TTO Experience		1.027***	0.779	0.623		0.184***	0.058	0.019
		(0.193)	(0.894)	(0.894)		(0.034)	(0.118)	(0.119)
Comm. Pot. x TTO Experience			-0.126	0.058			0.097	0.148
			(1.268)	(1.274)			(0.174)	(0.175)
Author Scientific Experience				0.325				0.057*
				(0.206)				(0.032)
Scientific Potential				0.406				-0.100
				(0.993)				(0.149)
Social Impact Potential				0.494				0.076
				(0.562)				(0.082)
Constant	-0.283	3.389***	-0.413	-1.937	-0.147	0.495***	-0.127	-0.294
	(0.575)	(0.130)	(0.850)	(1.210)	(0.096)	(0.023)	(0.116)	(0.208)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,815	1,815	1,815	1,815	1,815	1,815	1,815	1,815
R-squared	0.214	0.197	0.219	0.220	0.213	0.193	0.219	0.221

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

Table 8: Models estimating which inventions are more likely to generate commercialization agreements and licensing deals. Models 1 to 4 estimate a linear probability model with the likelihood that an invention will lead to a commercial agreement, and models 4 to 8 with the likelihood that an invention will lead to a licensing agreement. The average commercial potential of the articles tied to an invention strongly predicts agreements and, to some extent, licensing deals. The models successively add variables to control for the scientific and social impact potential of the focal article, the author's previous scientific experience, and whether the author has previous experience disclosing inventions to the TTO, along with an interaction effect. Models 4 and 8 also add as a control the amount invested into an invention by the TTO. This variable completely mediates the effects of the commercial potential, suggesting that the TTO strongly selects the most promising ideas. Fixed effects are included at an invention field-year level.

	(1) Agreements	(2) Agreements	(3) Agreements	(4) Agreements	(5) Licenses	(6) Licenses	(7) Licenses	(8) Licenses
Commercial Potential	0.532***		0.228	0.043	0.159**		-0.035	-0.070
	(0.130)		(0.160)	(0.134)	(0.072)		(0.114)	(0.110)
Author TTO Experience		0.144***	-0.145	-0.216*		0.097***	-0.059	-0.102
		(0.027)	(0.130)	(0.110)		(0.021)	(0.103)	(0.098)
Comm. Pot. x TTO Experience			0.380**	0.452***			0.219	0.270**
			(0.175)	(0.150)			(0.136)	(0.132)
Author Scientific Experience				0.006				0.013
				(0.025)				(0.020)
Scientific Potential				-0.274				-0.291**
				(0.178)				(0.113)
Social Impact Potential				0.024				-0.016
				(0.064)				(0.052)
Invested				0.352***				0.117***
				(0.035)				(0.023)
Constant	0.087	0.367***	0.213*	0.336**	0.076	0.122***	0.145*	0.274**
	(0.092)	(0.018)	(0.113)	(0.160)	(0.051)	(0.014)	(0.083)	(0.134)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,815	1,815	1,815	1,815	1,815	1,815	1,815	1,815
R-squared	0.181	0.181	0.190	0.257	0.133	0.142	0.144	0.166

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

Table 9: Linear probability models estimating whether inventions associated with high commercial potential articles are likely to generate revenue (models 1 to 3), be commercialized via a startup (4 to 6) and, conditional on startup, to raise venture capital funds (7-9). Models successively add variables to control for the scientific and social impact potential of the focal article, the author's previous scientific experience, and whether the author has previous experience disclosing inventions to the TTO, along with an interaction effect. Models 3, 6, and 9 also control the amount invested into an invention by the TTO. Fixed effects are included at an invention field-year level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Any revenue	Any revenue	Any revenue	Startup	Startup	Startup	VC Investment	VC Investment	VC Investment
Commercial Potential	0.044		-0.001	0.046		-0.065	1.013**		1.479***
	(0.031)		(0.056)	(0.038)		(0.053)	(0.375)		(0.484)
Author TTO Experience		0.021*	-0.037		0.036***	-0.049		0.266**	0.694
		(0.011)	(0.059)		(0.010)	(0.048)		(0.123)	(0.469)
Comm. Pot. x TTO Experience			0.082			0.115*			-0.637
			(0.086)			(0.065)			(0.532)
Author Scientific Experience			-0.018			0.009			-0.188
			(0.017)			(0.011)			(0.139)
Scientific Potential			-0.011			-0.025			-0.521
			(0.071)			(0.074)			(0.716)
Social Impact Potential			-0.034			-0.055*			-0.325
			(0.035)			(0.028)			(0.228)
Invested			0.024			0.062***			0.094
			(0.015)			(0.015)			(0.154)
Constant	0.036*	0.053***	0.150**	0.024	0.032***	0.074	-0.186	0.334***	0.570
	(0.022)	(0.007)	(0.074)	(0.027)	(0.007)	(0.052)	(0.273)	(0.101)	(0.496)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,815	1,815	1,815	1,815	1,815	1,815	85	85	85
R-squared	0.093	0.094	0.099	0.107	0.112	0.129	0.336	0.322	0.402

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) Cited by corporate patent	(2) Cited by corporate patent	(3) Cited by corporate patent	(4) Cited by corporate patent
Commercial Potential	0.104***		0.088***	0.065***
	(0.017)		(0.014)	(0.011)
Disclosed		0.034***	-0.068***	-0.071***
		(0.007)	(0.017)	(0.018)
Disclosed X Comm. Pot.			0.136***	0.135***
			(0.033)	(0.033)
Scientific Potential				-0.019***
				(0.005)
Social Impact Potential				0.009**
				(0.004)
Author Scientific Impact				-0.007***
				(0.001)
Author Commercial Impact				0.015***
				(0.003)
Constant	-0.031***	0.021***	-0.025***	-0.006
	(0.009)	(0.001)	(0.008)	(0.006)
Publication field - Year FE	Yes	Yes	Yes	Yes
Observations	92,972	92,972	92,972	92,972
R-squared	0.058	0.045	0.062	0.065

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) Cited by corporate patent	(2) Cited by corporate patent	(3) Cited by corporate patent	(4) Cited by corporate patent
Commercial Potential	0.305***		0.302***	0.306***
	(0.078)		(0.077)	(0.088)
Patented		0.020	0.004	0.004
		(0.020)	(0.019)	(0.019)
Scientific Potential				-0.139
				(0.088)
Social Impact Potential				0.109***
				(0.034)
Author Scientific Experience				0.020
				(0.016)
Author TTO Experience				-0.031*
				(0.019)
Constant	-0.069	0.137***	-0.069	-0.126
	(0.055)	(0.010)	(0.055)	(0.087)
Invention field - Year FE	Yes	Yes	Yes	Yes
Observations	1,815	1,815	1,815	1,815
R-squared	0.245	0.234	0.245	0.252

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 2010 and June 2022.

	Mean	SD	Min	Max	N
Cited by patent	0.07	0.25	0.00	1.00	3,936,673
Average years citing patents are renewed	0.93	1.22	0.00	8.39	264,101
Commercial potential	0.51	0.22	0.07	0.91	3,936,672
Scientific potential	0.62	0.15	0.04	0.96	3,936,672
Social impact potential	0.62	0.34	0.04	0.98	3,936,673
Institution commercial impact	98.85	50.16	0.00	236.00	3,936,673
Institution scientific impact	609.46	272.93	0.00	1,398.00	3,936,673
Journal commercial impact	0.01	0.03	0.00	1.60	3,936,673
Journal scientific impact	1.75	2.88	0.00	122.33	3,936,673
Researcher commercial impact	5.77	6.50	0.00	84.00	3,936,673
Researcher scientific impact	45.47	32.57	0.00	308.00	3,936,673

Table 13: Correlations: U.S. scientific research published between 2010 and June 2022.

Variables	Cited by patent	Years patents renewed	Commercial potential	Scientific potential	Social impact	Institution commercial impact	Institution scientific impact	Journal commercial impact	Journal scientific impact	Researcher commercial impact	Researcher scientific impact
Cited by patent	1.000										
Years patents renewed	.	1.000									
Commercial potential	0.244	0.123	1.000								
Scientific potential	0.054	-0.027	0.238	1.000							
Social impact potential	0.048	0.007	0.276	0.021	1.000						
Institution commercial impact	0.045	0.057	0.088	0.129	0.053	1.000					
Institution scientific impact	0.014	0.017	0.073	0.141	0.058	0.970	1.000				
Journal commercial impact	0.273	0.187	0.211	0.049	-0.031	0.051	0.012	1.000			
Journal scientific impact	0.214	0.060	0.145	0.176	-0.110	0.089	0.055	0.514	1.000		
Researcher commercial impact	0.230	0.127	0.462	0.136	0.140	0.254	0.235	0.218	0.203	1.000	
Researcher scientific impact	0.083	0.018	0.221	0.246	0.056	0.316	0.326	0.086	0.170	0.699	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 under 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 the field-year level.

	(1)	(2)	(3)
	Cited by patent	Cited by patent	Cited by patent
Commercial potential	0.231***	0.099***	0.040***
	(0.026)	(0.008)	(0.010)
Scientific potential	0.025***	0.015**	0.016**
	(0.008)	(0.007)	(0.007)
Social impact potential	0.026***	0.028***	0.027***
	(0.005)	(0.005)	(0.004)
High commercial impact institution	-0.042***	-0.039***	-0.033***
	(0.005)	(0.004)	(0.004)
Commercial potential x High commercial impact institution	0.108***	0.101***	0.082***
	(0.011)	(0.009)	(0.007)
High scientific impact institution	0.028***	0.022***	0.025***
	(0.006)	(0.005)	(0.005)
Commercial potential x High scientific impact institution	-0.047***	-0.037***	-0.049***
	(0.011)	(0.010)	(0.009)
High commercial impact journal		-0.094***	-0.089***
		(0.010)	(0.010)
Commercial potential x High commercial impact journal		0.211***	0.199***
		(0.022)	(0.020)
High scientific impact journal		0.040***	0.040***
		(0.012)	(0.012)
Commercial potential x High scientific impact journal		-0.016	-0.021
		(0.021)	(0.021)
High commercial impact researcher			-0.043***
			(0.005)
Commercial potential x High commercial impact researcher			0.154***
			(0.016)
High scientific impact researcher			-0.010***
			(0.004)
Commercial potential x High scientific impact researcher			0.009
			(0.007)
Constant	-0.088***	-0.050***	-0.029***
	(0.018)	(0.010)	(0.009)
Publication field - year FE			
Observations	3,936,672	3,936,672	3,936,672
R-squared	0.119	0.130	0.138

Standard errors clustered at the publication field-year 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 the field-year level.

	(1)	(2)	(3)
	Years patents renewed	Years patents renewed	Years patents renewed
Commercial potential	1.133***	0.463***	0.252***
	(0.066)	(0.056)	(0.061)
Scientific potential	-0.523***	-0.551***	-0.560***
	(0.043)	(0.042)	(0.042)
Social impact potential	0.028*	0.031**	0.026*
	(0.015)	(0.016)	(0.015)
High commercial impact institution	-0.115**	-0.099*	-0.078
	(0.057)	(0.055)	(0.052)
Commercial potential x High commercial impact institution	0.348***	0.322***	0.264***
	(0.074)	(0.072)	(0.068)
High scientific impact institution	0.084	0.078	0.091*
	(0.054)	(0.051)	(0.051)
Commercial potential x High scientific impact institution	-0.098	-0.092	-0.127*
	(0.077)	(0.073)	(0.072)
High commercial impact journal		-0.406***	-0.358***
		(0.042)	(0.044)
Commercial potential x High commercial impact journal		0.632***	0.545***
		(0.054)	(0.057)
High scientific impact journal		-0.043	-0.028
		(0.046)	(0.044)
Commercial potential x High scientific impact journal		0.167***	0.128**
		(0.063)	(0.060)
High commercial impact researcher			-0.240***
			(0.037)
Commercial potential x High commercial impact researcher			0.566***
			(0.054)
High scientific impact researcher			0.089**
			(0.044)
Commercial potential x High scientific impact researcher			-0.120*
			(0.065)
Constant	0.394***	0.785***	0.871***
	(0.058)	(0.043)	(0.043)
Publication field - year FE			
Observations	264,099	264,099	264,099
R-squared	0.165	0.166	0.171

Standard errors clustered at the publication field-year 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

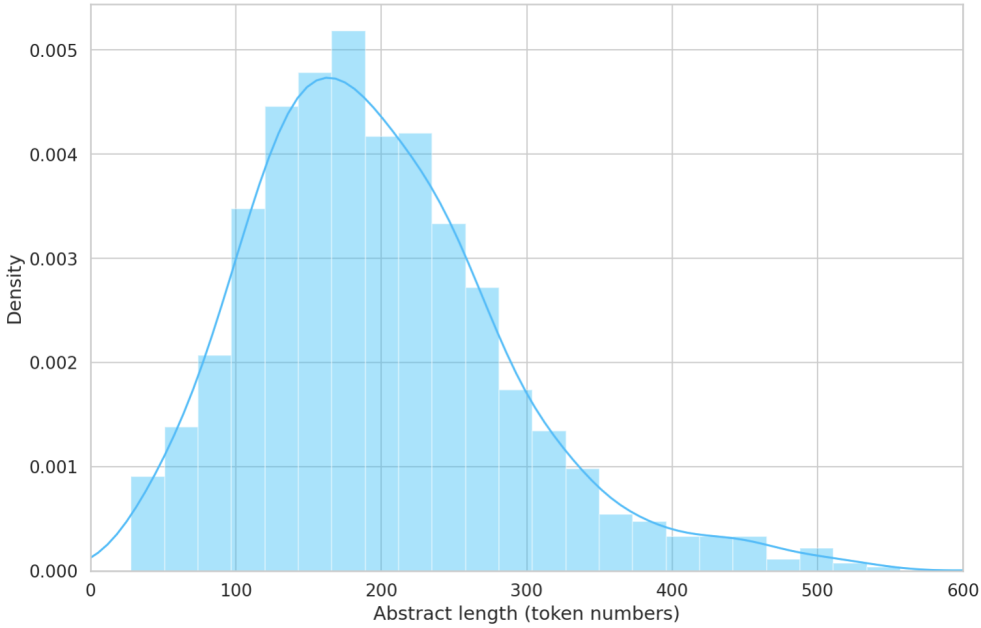


Figure B.1: Abstract's token length distribution

Table B.1: Commercial potential model performance

	Precision	Recall	F1-score
No patent citation / citation by non-renewed patent	0.75	0.66	0.70
Citation by renewed patent	0.70	0.77	0.74
Accuracy			0.75
Micro-averaged ROC AUC			0.75

Table B.2: Scientific potential model performance

	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.

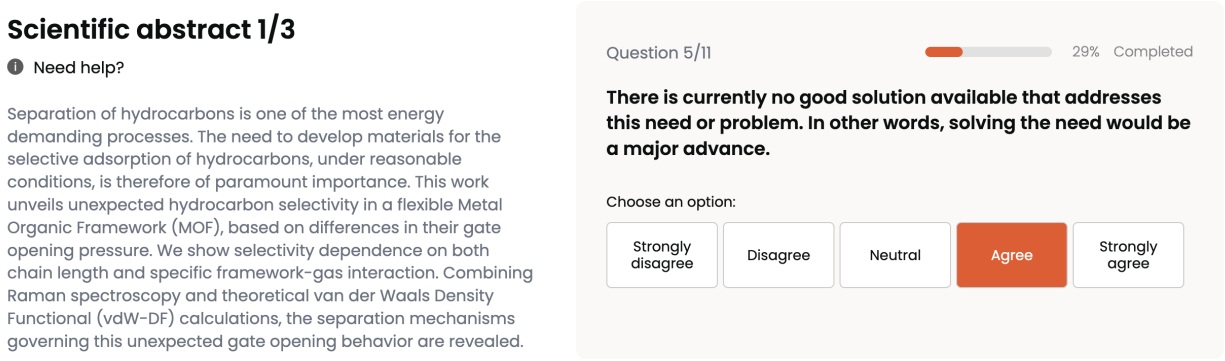


Figure C.2: Interface evaluators interacted with in the elicitation exercise.

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

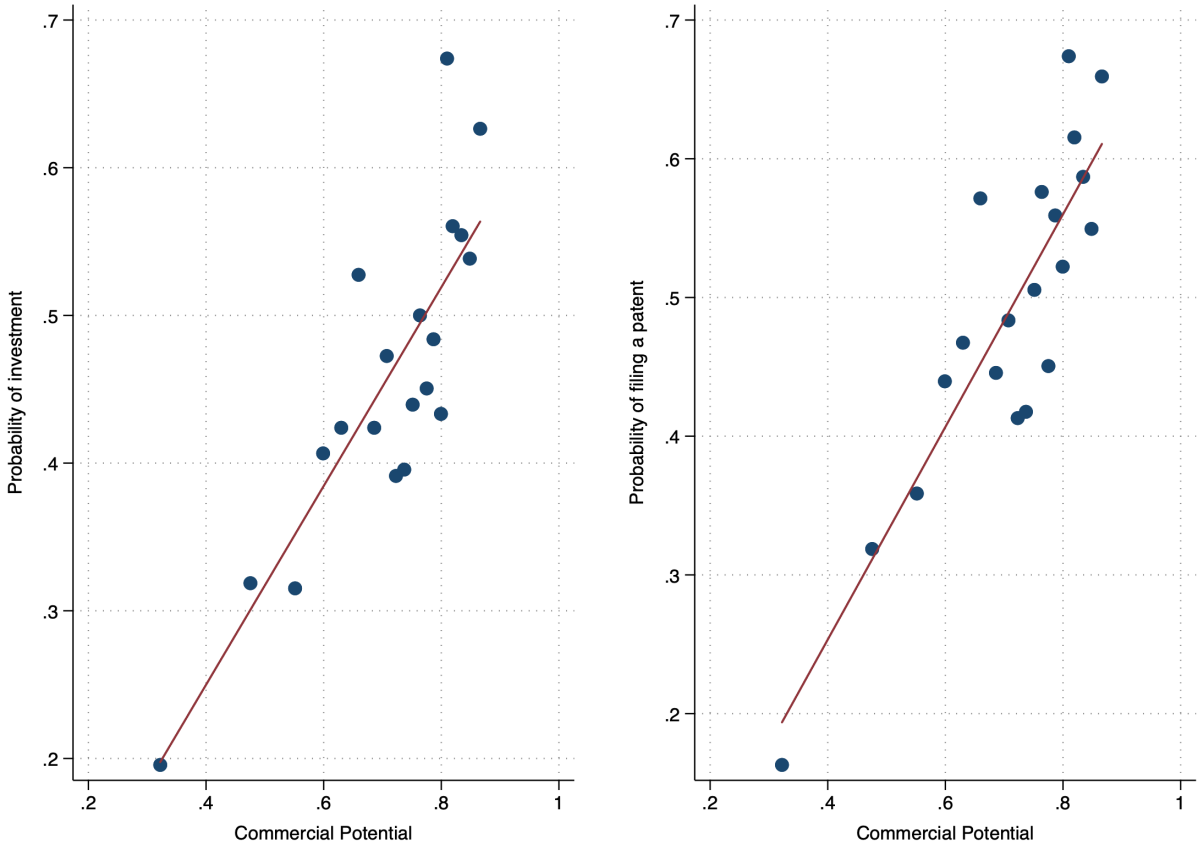


Figure D.1: Probability that the TTO will invest into (Panel A) and patent (Panel B) an invention based on the average commercial potential of the articles associated with the invention.

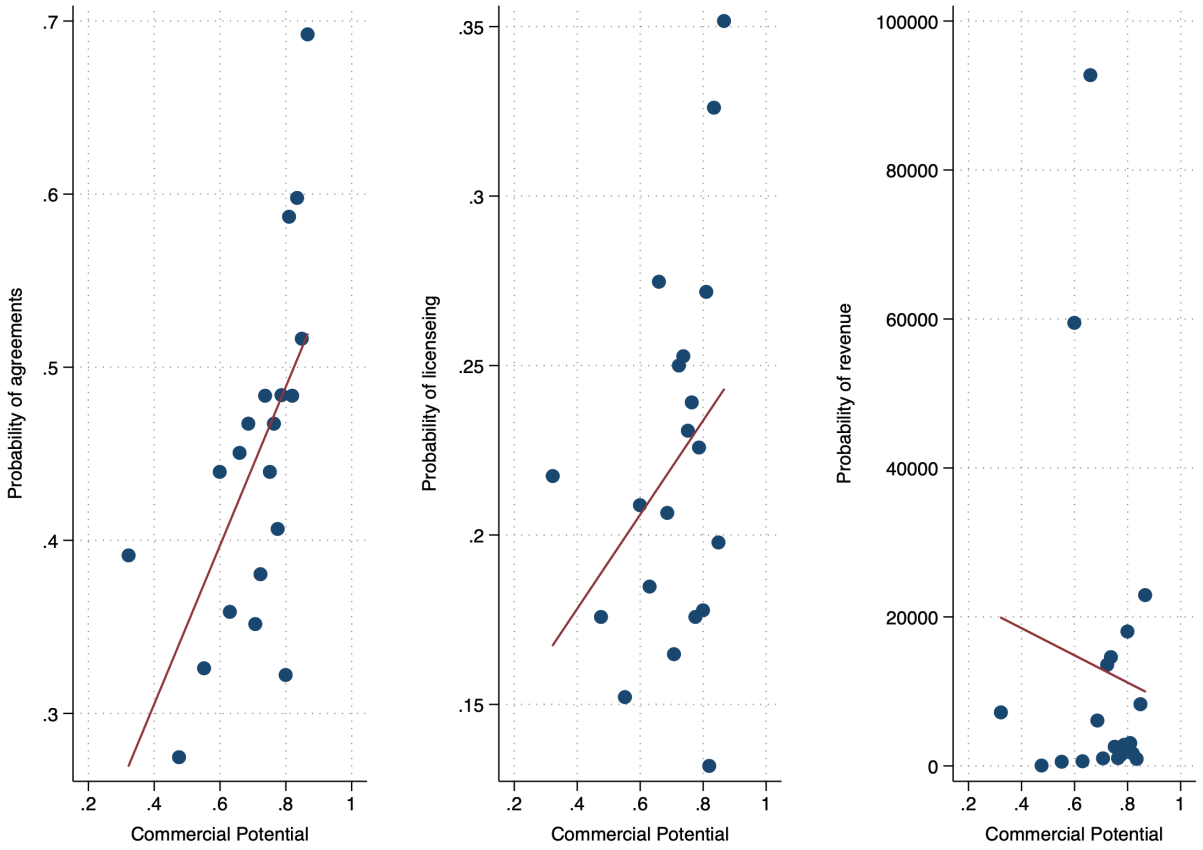


Figure D.2: Probability that an invention will garner agreements (Panel A) and licensing deals (Panel B), as well as generate revenue to the TTO (Panel C) based on the average commercial potential of the articles associated with the invention.

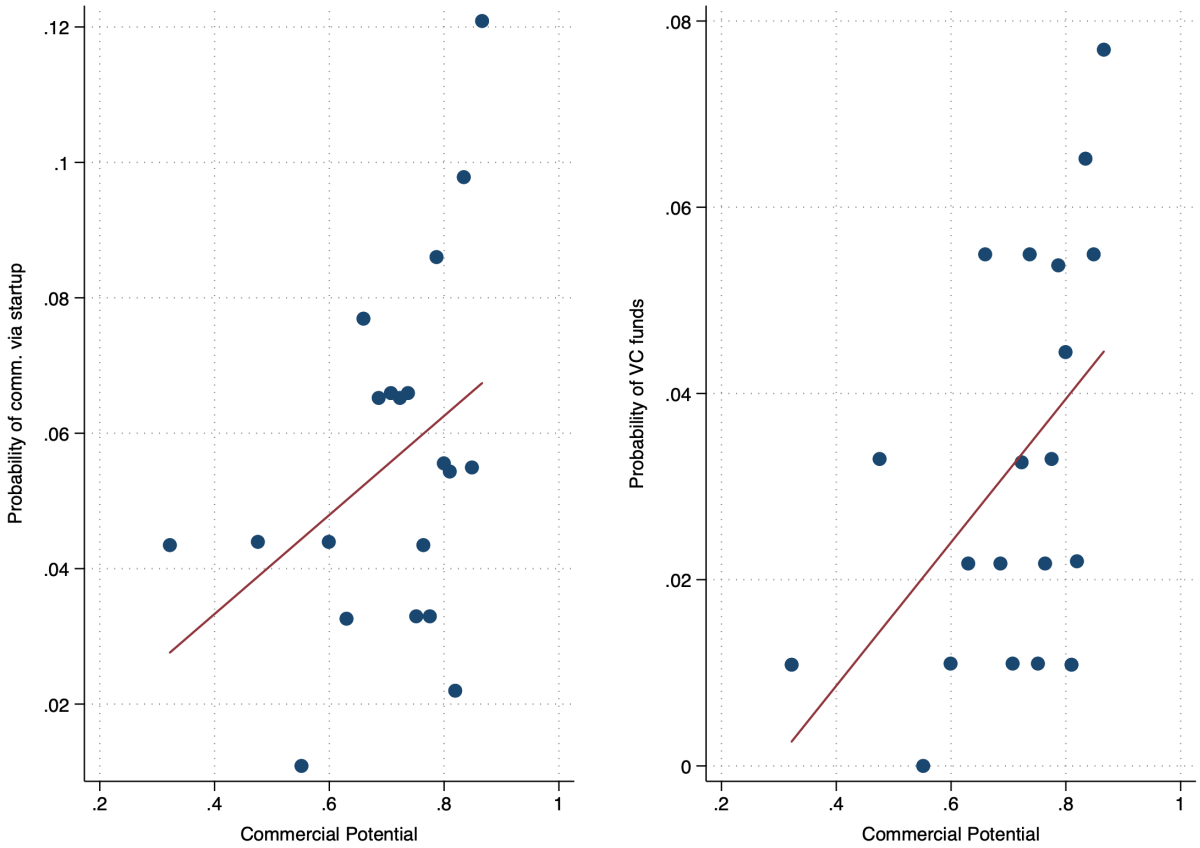


Figure D.3: Probability that an invention will be commercialized via a Startup (Panel A) and, conditional on Startup, that will raise venture capital funds as a function of the average commercial potential of the articles associated with the invention.