NBER WORKING PAPER SERIES

MEASURING THE COMMERCIAL POTENTIAL OF SCIENCE

Roger Masclans-Armengol
Sharique Hasan
Wesley M. Cohen

Working Paper 32262
http://www.nber.org/papers/w32262

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024, Revised April 2024

Authorship in reverse alphabetical order. Sharique Hasan would like to thank the Kauffman Foundation, which funded this study through their knowledge challenge grant #G-201806-4738. The authors would also like to thank seminar participants at Duke, Bocconi, University of Maryland, NBER i3 Conference, Entrepreneurship and Innovation Policy Research Seminar, as well as helpful feedback from Ashish Arora, Dan Gross, Lee Fleming, Matt Marx, Robin Rasor, Yoko Shibuya, and John Walsh. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring the Commercial Potential of Science
Roger Masclans-Armengol, Sharique Hasan, and Wesley M. Cohen
NBER Working Paper No. 32262
March 2024, Revised April 2024
JEL No. O3,O30,O31,O32,O33,O34,O35,O36,O38,O39

ABSTRACT

This paper uses a large language model to develop an ex-ante measure of the commercial potential of scientific findings. In addition to validating the measure against the typical holdout sample, we validate it externally against 1.) the progression of scientific findings through a major university's technology transfer process and 2.) firms' use of the academic science of major American research universities. We then illustrate the measure's utility by applying it to two questions. First, does the patenting of academic research by universities impede its breadth of use by firms? Second, to illustrate how this measure can advance our understanding of the determinants of firms' use of science generally, we use it to analyze how one factor, universities' reputations for generating commercializable science, impacts firms' use of academic science. For the former question, using our measure to control for commercializable science, we find that patenting does not dampen the dissemination of academic science in industry. For the second, we find that reputation per se, apart from the production of commercializable science, impacts industry's use of science, especially for that science with high commercial potential, implying that the commercializable science of less prominent universities is disproportionately overlooked by industry.

Roger Masclans-Armengol
Duke University
roger.masclans@duke.edu

Sharique Hasan
Duke University
Fuqua School of Business
Box 90120
Durham, NC 27708-0120
sh424@duke.edu

Wesley M. Cohen
The Fuqua School of Business
Duke University
Box 90120
Durham, NC 27708-0120
and NBER
wcohen@duke.edu

Data: Measure of the Commercial Potential of Science is available at www.zenodo.org/records/10815144

Measuring the Commercial Potential of Science *

Roger Masclans-Armengol† Sharique Hasan† Wesley Cohen‡

March 28, 2024

Abstract

This paper uses a large language model to develop an ex-ante measure of the commercial potential of scientific findings. In addition to validating the measure against the typical holdout sample, we validate it externally against: 1.) the progression of scientific findings through a major university's technology transfer process; and 2.) firms' use of the academic science of major American research universities. We then illustrate the measure's utility by applying it to two questions. First, does the patenting of academic research by universities impede its breadth of use by firms? Second, to illustrate how this measure can advance our understanding of the determinants of firms' use of science generally, we use it to analyze how one factor, universities' reputations for generating commercializable science, impacts firms' use of academic science. For the former question, using our measure to control for commercializable science, we find that patenting does not dampen the dissemination of academic science in industry. For the second, we find that reputation per se, apart from the production of commercializable science, impacts industry's use of science, especially for that science with high commercial potential, implying that the commercializable science of less prominent universities is disproportionately overlooked by industry.

Keywords *— Innovation, Technological Opportunity, Science Commercialization, Deep Learning, Large Language Models*

*Authorship in reverse alphabetical order. Sharique Hasan would like to thank the Kauffman Foundation, which funded this study through their knowledge challenge grant. The authors would also like to thank seminar participants at Duke, Bocconi, University of Maryland, NBER i3 Conference, Entrepreneurship and Innovation Policy Research Seminar, as well as helpful feedback from Ashish Arora, Dan Gross, Lee Fleming, Matt Marx, Robin Rasor, Yoko Shibuya, and John Walsh.

†Duke University, The Fuqua School of Business.

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

1 Introduction

Scientific research is crucial for technological advance and economic growth. Understanding the conditions under which scientific discoveries contribute to the development of new products and services by firms has long attracted the attention of practitioners, academics, and policymakers (Stokes, 2011). Indeed, the aim to *commercially capitalize* on public research for social and private benefit in the U.S. has inspired numerous policy initiatives, ranging from the dissemination of expertise by land grant colleges in the late 19th century to the Bayh-Dole Act of 1980 and subsequent legislation. Despite these concerns, the fundamental questions of understanding or even describing the contribution of academic science to commercial outcomes remains an active area of research today (Mowery et al., 2015).

A pervasive methodological challenge that hinders our understanding of the contribution of academic science to commercial outcomes is the difficulty in assessing the commercial potential of scientific discoveries and findings, which, in turn, would allow us to distinguish the *commercial potential* of such findings from their *commercial realization* (Marx and Hsu, 2022). Only a fraction of academic science possesses the potential to contribute to the development of new products or processes (Klevorick et al., 1995). Therefore, to analyze the determinants of science commercialization, an important step is to identify which science is commercially viable (or 'at risk' of ever being commercialized). This distinction is crucial. For instance, the processes that influence the research that produces commercializable science differ from those affecting its identification and subsequent utilization. These processes involve different actors, each with their incentives. The production of commercializable science primarily involves academic researchers, their institutions, and grant-making bodies, both public and private (Henderson et al., 1998; Thursby and Thursby, 2002). In contrast, the identification and subsequent utilization of this science primarily involve firms (Cohen et al., 2002).

In this paper, we address this challenge of unobservability by using a large language model to develop a general-purpose measure of the commercial potential of academic science, which we make publicly available.[1] We also illustrate the utility of this measure with two exercises.

[1]The data is available to download at www.zenodo.org/records/10815144 Initially, we post the measure for 5.2

The first uses the measure to explore whether one university's patenting of science dampens the diffusion of its use by firms. In the second, we use it to distinguish between the academic production of commercializable science versus its use by firms. This distinction allows us to better explore, for example, the independent role of universities' reputations in affecting the commercial application of their research and the consequent under-utilization of commercializable research from less prominent schools.

It is worth noting the distinction between commercial potential and realization is not new. For instance, Azoulay et al. (2007) and Marx and Hsu (2022) have previously highlighted the unobservable nature of the commercial potential of science as a challenge for empirical analyses.[2] In their study of science commercialization through startup formation, Marx and Hsu (2022) extensively explore the econometric implications of failing to account for the commercial potential of science—namely, the introduction of omitted variable bias arising from commercial potential lingering in the error term. While it is possible to approximate the frequency with which academic research leads to commercial applications by using measures such as patent citations to the scientific literature, these measures do not indicate the extent to which that or other research had the potential for commercial use in the first place. By not controlling for the risk set of commercializable research, this approach can bias estimates of the effects of hypothesized determinants of commercialization. Marx and Hsu (2022) illustrate this issue using the example of experienced teams and their potential ability to select higher-quality ideas. When examining whether an entrepreneurial team affects the translation of an idea into a successful commercial application, ignoring the underlying commercial potential of the idea may bias estimates, potentially overestimating the translational capability of more experienced teams. This could occur if, for example, more experienced teams are also more skilled at selecting high-potential ideas to pursue. To tackle this omitted variable problem, Marx and Hsu (2022) employ a sophisticated econometric strategy that utilizes "twin discoveries" to identify a subset of approximately 20,000 "twin" scientific articles, thereby allowing

million articles published in the U.S. since 2000. We also plan to develop and post the measure for the population of all English-language scientific articles published since 2000.

[2] Azoulay et al. (2007) developed, for the life sciences, a direct measure of patentability, closely approximating commercializability, by relying on the similarity of keywords from the title of an academic article to the text of a patent. However, this measure has not been validated and employs a method that, though innovative at the time, has now been surpassed.

for the control of differences in the latent commercializability of the science.

While certain econometric approaches (e.g., twins, instrumental variables, fixed effects) may provide unbiased estimates of the effects of different factors on the commercialization of a given piece of science, they still leave the commercial potential unobserved. Consequently, studying the causes and consequences of commercial potential—not merely as an econometric nuisance but as a critical economic and strategic variable of interest—becomes problematic. For instance, without such a measure, it becomes challenging to evaluate the extent to which the commercial potential of science is unrealized—that is, the proportion of academic research that could be utilized commercially but is not—or the returns from science investments (Lane and Bertuzzi, 2011). Furthermore, we cannot distinguish between factors that influence the production of commercializable science by academics, such as incentives (Lach and Schankerman, 2008), and those affecting firms' identification and utilization of such science, such as geographic hubs (e.g.,. Bikard and Marx, 2020).

Therefore, only with such a measure can we comprehensively distinguish between the impact on commercialization of the characteristics of researchers and institutions that produce commercializable science versus those of the firms tasked with identifying potentially commercial science. This distinction is vital not only for a deeper understanding of the commercialization of science but also for guiding policy and management decisions. For example, if the primary barriers to commercialization stem from how firms identify which science to develop, this can be addressed by improving the management of firms' processes and practices. On the other hand, firms can do little to address this issue if the lack of commercialization is due to insufficient production of commercializable science. Instead, public and educational policies may play a significant role, especially in light of a growing division of innovative labor between public research and corporate innovation (Arora et al., 2018; Fleming et al., 2019).

Thus, to explore the production of scientific discoveries and their transformation into commercializable technologies, we require a measure that captures, *ex-ante*, the *commercial potential* of a discovery, distinguishing potential from actual realization. In this paper, we introduce such a measure of the commercial potential of science, validate it through two separate empirical exercises,

and demonstrate its utility.

To develop our *ex-ante* commercial potential measure of a scientific article, we utilize an ensemble of machine learning algorithms, specifically large language models and neural networks. We train predictive models using the text of an academic article's abstract to generate *ex-ante*, out-of-sample, and out-of-training-time-period predictions of any scientific article's commercial potential, independent of factors affecting realization. Conceptually, in our context, commercial potential refers to the probability that a firm will view an article as contributing to the development of a marketable product or process.[3] This concept is operationalized as the *ex-ante* probability that a scientific article will be cited in a patent that is subsequently renewed. Our operationalization, based on citations to articles in patents—and specifically to those patents that are later renewed—assumes that such citations reflect a firm's belief in the potential economic value of incorporating a given scientific finding or idea into an invention (Kuhn et al., 2020; Marx and Fuegi, 2020). The average accuracy of our model, along with the average area under the receiver operating characteristic curve (AUROC), stands at 0.74.[4]

In addition to validating our measure by examining its accuracy in a standard hold-out sample, we externally validate it in two empirical exercises. The first leverages detailed data on the progression of scientific knowledge through the technology transfer process within one university. The dataset includes comprehensive information on inventions disclosed to the university's technology transfer office (TTO) including the fact of the disclosure itself, TTO financial investment, patenting, licensing, and, in some cases, revenue generation—all outcomes on which our model

[3]Our measure of "potential" specifically targets scientific findings that have a direct impact on commercial outcomes, such as patents, rather than capturing the long-term potential of embryonic research that may take generations to manifest commercially. In further analysis, we demonstrate that scientific articles identified by our algorithm as having high commercial potential receive significantly higher rates of patent citations and are cited more promptly.

[4]For context, Manjunath et al. (2021) report an AUROC of 0.83 in their model, which predicts patent citations of articles. However, their model only uses abstracts from PubMed in the life sciences and does not consider patent renewals. Similarly, Koffi and Marx (2023) also employ a BERT-derived measure of the commercializability of science in their study. The paper does not, however, report sufficient detail to permit comparison with our methods or results. In contrast, Liang et al. (2022) created a model based on the text of inventions disclosed to Stanford's Technology Transfer Office (TTO), aiming to predict commercial value generation, with an achieved AUROC of 0.76. While our use of natural language processing is distinct in that it uses the text from academic papers to make predictions about citations from renewed patents, other work has also implemented natural language processing models. These other papers have used patent text along with other indicators (e.g., author, patent, and institution characteristics) to predict the value of patents, as measured by forward citations, use by commercialized products, or market responses to patenting by public firms (Chuang et al., 2023; Hsu et al., 2020).

and measure was **not** trained on. We also supplement these data by linking each invention to the scientific articles upon which it is based. Our findings confirm that our measure of the commercial potential of science successfully predicts all these outcomes.

In the second validation exercise, we analyze commercial outcomes, including patent citations and renewals, for over 5.2 million academic papers published by U.S.-based research-intensive universities. Confirming the predictive power of our measure, we find strong evidence that those scientific articles identified by our measure as having high commercial potential are more likely to undergo commercialization, as evidenced by their citation in a renewed patent. For example, a scientific article with a commercial potential score in the top quartile is over 20 times more likely to be cited by a renewed patent than an article in the bottom quartile. Furthermore, integrating our measure into econometric models reduces the coefficient estimates of proxies for commercial potential (for example, the prior commercial success of science by universities, researchers, and publications in more applied journals) by as much as 33%, while also enhancing the overall predictive performance of the model.

Finally, to illustrate the applicability of our measure to the study of substantive questions in innovation strategy, we conclude the paper with two illustrative exercises. The first speaks to the debate in the literature over the spillover effects of universities' privatization (i.e., patenting) of academic science (Dasgupta and David, 1994; Mowery et al., 2015). Specifically, we use our measure of the commercial potential of science to explore the relationship between one university's patenting of academic science and its subsequent use by firms. The question is whether patenting by a university limits firms' use of the commercializable science linked to its patents as compared to comparably commercializable science produced by the university that is not patented. We find that university patenting is not associated with less use by firms or with fewer firms using the science. Employing our measure, our results suggest the opposite.

In a second illustrative exercise, we investigate a determinant of the rate at which universities commercialize their science—specifically, the role that a university's reputation may play in driving that rate. Our measure allows us to distinguish between two explanations. In one, reputation simply proxies for a university's ability to produce commercializable research. In the other, uni-

versities' reputation serves as a focusing device; firms economize on search costs by paying more attention to those universities with a stronger track record of producing commercializable research. The degree to which either or both explanations apply matters. If a desired goal is to increase commercializable research, then the former would suggest important roles for policy and university administrations. If the latter holds, managers may consider implementing policies to increase the efficiency of their search for useful science. While we find evidence for both, our finding that a bit over 50% of the variation in rates of patent citations of papers across universities is attributable to the production of commercializable research points to production as the dominant explanation.

Our finding that reputation per se matters nonetheless suggests, however, that the commercial potential of a share of commercializable science is not being realized and that a "realization gap" exists. In addition to reputation, one can easily argue that other focusing devices, such as universities' efforts to market their research, location (Bikard and Marx, 2020), gender (Koffi and Marx, 2023), and organizational origin of the inventor (Bikard, 2018) may also contribute to this realization gap.

Our study contributes to the literature on innovation and the commercialization of science. We employ a novel methodological approach that allows us to develop a comprehensive measure of *ex-ante* commercial potential, which precedes commercial outcomes. This method enables an assessment of the commercial potential of science, a task not possible with outcome-based measures or natural experiments that lack observable indicators of commercial potential (Azoulay et al., 2007; Marx and Hsu, 2022). Additionally, the illustrative applications of our measure contribute to ongoing discussions in our field. First, we engage with the debate on the privatization of scientific knowledge through patenting and licensing activities by universities (Dasgupta and David, 1994; NRC, 2011; Nelson, 2004; Henderson et al., 1998; Williams, 2013), particularly focusing on the role of technology transfer offices (Debackere and Veugelers, 2005; Thursby and Thursby, 2002). As highlighted by several scholars (e.g., Henderson et al., 1998; Murray and Stern, 2007; Mowery et al., 2015), methodological challenges have hindered comprehensive analysis. However, by applying our measure to data from one university, we provide evidence suggesting that privatization (i.e., patenting) can contribute to the diffusion of academic science across firms.

Second, our research enhances the understanding of how scientific research is transformed into commercial applications. It builds on previous studies that have examined factors affecting the commercialization rates of academic science and discrepancies between the commercial potential of public research and its commercial realization. Numerous researchers have focused on identifying the attributes of researchers and teams that may affect these disparities. For example, Ding et al. (2006) looked into the impact of gender and ethnic diversity on the commercialization of research, emphasizing the importance of collaboration and interdisciplinary efforts, and Koffi and Marx (2023) establish a gender-driven realization gap. Similarly, Marx and Hsu (2022) and Hsu and Kuhn (2023) investigated the contribution of team dynamics in advancing ideas, while Bikard (2018) and Bikard (2020) examined how the use of science by firms varies based on the geographic origin of the research. Our study not only confirms the existence of a realization gap but allows us to measure and thus distinguish between the production of commercializable research versus its use by firms.

2 Data and Methods

This section describes the data employed, the training methodology, and the performance of the models developed to measure the *commercial potential* of a scientific article. Based on the abstract text in which a scientific discovery is reported, *the measure predicts the probability of a citation to that article by a renewed patent. We, in turn, interpret a renewed patent as reflecting a firm's belief that the patent may provide a basis for a commercialized innovation.*

For its development, we employ Large Language Models (LLM) and deep neural networks. Our approach involves training classifier models capable of automatically categorizing textual data into predefined classes—whether the scientific article has commercial potential. The models learn to identify patterns and features associated with each class, enabling them to predict the most probable class for new, unseen text.

For the training process, we classify—label—an article as having commercial potential if, within the training time frame, it is cited by least one renewed patent. Our training process, conducted with over 400,000 scientific articles, measures the commercial potential of articles published be-

tween 2000 and 2020.

2.1 Fine-tuning an LLM for scientific "understanding"

The specific Large Language Model (LLM) we use is SciBERT (Beltagy et al., 2019). This model is already trained on a large corpus of scientific text (1.14M scientific articles) derived from BERT, a foundational model developed by Google AI in 2018 (Devlin et al., 2018). Our specific interest lies in classification, where the goal is to categorize a document (abstract) into two groups: whether the abstract exhibits commercial potential or not. We fine-tune SCIBERT by exposing it to our labeled dataset, associating each abstract with its respective commercial potential class to enable the model to recognize patterns related to high commercial potential in scientific articles.[5] Intuitively, this process identifies specific regions within a high-dimensional space that correlate with patterns associated with patent citations, allowing us to categorize new abstracts as having either high or low commercial potential based on the patterns it has learned from the training data.[6]

2.2 Scientific articles and patent data

We use Dimensions as a source of information on scientific articles.[7] The dataset contains information on more than 139 million publications with their title, abstract text, publication sources, author information, fields of research, and other metadata. To ensure high-quality data, and beyond standard cleaning (e.g., removing duplicated articles or articles with missing data), we limit our analysis to peer-reviewed journal articles and findings reported in conference proceedings. Moreover, we focus our analysis on eleven scientific fields, which cover the majority of natural and applied sciences and engineering but exclude the social sciences. The fields are: Agricultural, veterinary, and food sciences; Biological sciences; Biomedical and clinical sciences; Chemical sciences;

[5] We tested the model's performance using BERT vs. SciBERT and found our models' performance increases when using SciBERT instead of BERT. We also tested our models using another embedding model, SPECTER2 (Cohan et al., 2020; Singh et al., 2022), and found SciBERT to perform better.

[6] Technicalities of the training process are described in Appendix A.

[7] Dimensions is a research and innovation database that contains detailed information on publications, patents, grants, clinical trials, and policy documents.

Earth sciences; Engineering; Environmental sciences; Health sciences; Information and computing sciences; Mathematical sciences; and Physical sciences. The resulting sample is comprised of 50,362,042 academic papers.

We source patent citations to scientific papers from the Reliance on Science dataset (Marx and Fuegi, 2020, 2022). This dataset contains 22,660,003 linkages between 3,017,441 unique patents and 4,017,152 unique papers. Using the DOI of a paper (Digital Object Identifier—a unique, universal identifier), we merged the Reliance on Science dataset with Dimensions. This resulted in matching all 4,017,152 papers in the Reliance on Science dataset to a paper in the Dimensions subsample we created.

Next, we use data from the United States Patent and Trademark Office (USPTO). For each patent citing a paper, we collect information on the assignee and its renewal status. Using the patent number, we merge the two datasets containing patent information (Reliance on Science and USPTO). Upon matching the data, we find that 4.93% of the papers in our sample are cited by one or more renewed patents. For this analysis, we assume that papers not cited by a renewed patent in the resulting set are not commercially applied by firms.[8]

2.3 Commercial potential model: Training

We created a model specific to each year from 2000 to 2020. Accordingly, we trained 20 models. This approach avoids including data generated after the focal year t. Thus, we minimize data leakage, which occurs when d by a non-renewed patent (representing the class with "no commercial potential") and 10,000 articles cited by a renewed patent (representing the class with "commercial potential"). In developing each model, we divided the data into three sets: 75% for training, 12.5% for testing, and 12.5% for validation. This division follows machine learning best practices, allowing the final accuracy of the model to be evaluated with previously unseen data.[9]

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

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

2.4 Commercial potential model: Performance

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

The models achieved an average accuracy of 74%. Likewise, the average area under the receiver operating characteristic curve (AUROC), a measure of the true positive rate against the false positive rate at various thresholds, is 0.74, indicating a balanced distinction between false positives and false negatives.[10] Detailed performance metrics for each model are provided in Appendix A, Figures A.2, and A.3. Likewise, Tables A.2 and A.3, in Appendix A, present examples of articles from the top and bottom 25 percentiles of commercial potential, respectively. These examples are drawn from completely out-of-sample articles published after the end of our training sample period.

While our classifier demonstrates reasonable accuracy, there is potential for further enhancement. Three primary factors may be influencing its performance:

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

1. **Diverse Academic Fields**: Our models are trained to classify articles across various fields, from Biology to Engineering to Computer Science. Textual features indicative of commercial potential may significantly vary between these disciplines. This diversity necessitates compromises in parameter settings, consequently limiting the model's overall performance. For comparison, Manjunath et al. (2021) focused their model exclusively on the life sciences and biomedical fields, utilizing over 20 million articles from PubMed. They achieved an AUROC of 0.83, highlighting the benefits of field-specific models.

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

3. **Changing language**: The language signaling commercial potential may change over time, and our model is confined, per above, to a circumscribed sample period. This focus narrows our model's capacity to capture the nuanced dynamics of token emergence, usage, and interconnections and the detailed content in full texts, tables, and figures of articles that may affect the accuracy of our model predictions.

In the subsequent validation analyses, we will compare our model predictions with the outcomes of human decisions—including, for example, the citation of an article in a renewed patent or decisions regarding disclosure of the invention to the TTO, investment, licensing, and revenue. To the degree that our model predictions do not hold, it is not clear whether such errors are due to model errors or human error.[11] While our paper does not draw this distinction, there

[11]For instance, consider a scenario where the model forecasts that a renewed patent should cite an article, but it

are methods, albeit computing-intensive, that permit one to do so. (c.f., Blundell et al., 2015; Lakshminarayanan et al., 2017)

2.5 Time horizon of the commercial potential measure

When considering the commercial potential of a recent scientific discovery, it needs to be relative to a time frame. Can this potential be realized in one, five, or twenty years? Indeed, prior research suggests that the trajectory from scientific breakthroughs to market penetration is not instantaneous but marked by a significant incubation period, with some estimates being up to 20 years (Adams, 1990; IIT, 1968). As detailed in our methodology, the temporal window used to train our models includes articles published in a decade preceding a focal year starting at $t-15$ to $t-4$, and we truncate patent citations beyond the focal year. While the binary classification used to validate the models effectively distinguishes whether an article is eventually cited, it does not capture the relationship between our measure and the temporal lag between article publication and eventual patent citation. We suspect, however, that scientific articles that are temporally closer to commercialization—i.e., being cited in a renewed patent—are more likely to be classified as having commercial potential. The reason is that firms are more likely to recognize a scientific contribution as having commercial potential sooner if that potential is more apparent, and to the degree that the language is more apparent, we would hope that our model would classify the contribution as having a greater likelihood of commercial application.

In Table A.4, Appendix A, we relate our commercial potential measure to the lag between the publication date of an article and the filing of the first renewed patent citing the article.[12] With regard to the commercial potential measure, articles categorized in the top quartile of commercial potential exhibit a twofold likelihood of being cited in years 0 or 1 compared to being cited in years six and beyond. Specifically, 36.66% of the articles in the top quartile receive citations in years 0 or

does not. This discrepancy could arise from two possibilities. First, the model's prediction is incorrect, indicating that decision-makers were justified in not utilizing the scientific knowledge from the article (indicative of a model error), or second, the model's prediction is accurate, suggesting that the decision-makers overlooked or misjudged the commercial value of the information in the article (suggesting human error).

[12]The distribution of time lags provides a degree of validation for the time frames employed to train our models. Notably, only a small fraction of papers receive their first citation beyond the ten-year mark. Thus, the information we use during the training of our models predominantly captures the relevant dynamics.

1, while 16.99% are cited six or more years later. Conversely, only 18.26% of the articles classified in the bottom quartile receive citations in years 0 or 1, with 39.22% being cited six or more years later. We confirm our descriptive findings using Kaplan-Meier survival curves (see Figure A.4, Appendix A). As expected, articles classified in the top quartile of commercial potential are significantly more likely to be cited sooner than articles in lower quartiles. Therefore, it appears that our commercial potential measure reflects the time horizon within which commercial potential is likely to be realized, rating those articles that are temporally closer to commercialization as having greater commercial potential.

2.6 Secondary model: Scientific potential

In addition to their commercial potential, scientific potential may also be related to decisions regarding commercialization simply because the scientific promise of an idea or discovery can, at times, correspond to commercial application (e.g., CRISPR). This may be especially true when the science lives in what Stokes calls Pasteur's Quadrant (Stokes, 2011). Thus, we control for the scientific potential of an article simply because it may be correlated with its commercial potential, and we wish to isolate the latter's influence. For our measure of scientific potential, we have developed an additional language-based model to quantify the findings' scientific potential.

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

The performance of these models is satisfactory, achieving an average accuracy and Area Under the Receiver Operating Characteristic (AUROC) of 0.71. It is important to note that we conducted various experiments with different thresholds and settings. We prioritized maintaining consistency with the primary model's training sample, ensuring a balanced training dataset, and

setting a classification threshold meaningfully higher than zero citations. This approach allows the inclusion of papers in the *low* category that still have the potential to yield scientific value. Appendix A, Table A.1 presents a summary of the model's performance.

2.7 Technology Transfer Office Data

In addition to public data on papers and patents, we have access to detailed data from a major research university's technology transfer office (TTO). The dataset includes data on all invention disclosures and subsequent actions and outcomes tied to those disclosures, including patenting, licensing, agreements, revenue, TTO investments tied to each invention, whether the licensee is a startup or an established firm, and inventor identity, including the inventor's history with the TTO. We remove inventions disclosed before 2000 and those not associated with an active researcher at the time of disclosure. The resulting dataset includes 5,219 invention disclosures from January 2000 to December 2020. These data will serve as an external validation to our measure and allow us to provide new, preliminary insights regarding the commercialization of science-based inventions.

One crucial element not available in the data provided by the TTO is the linkage between scientific articles and invention disclosures. To match faculty articles to invention disclosures, we take three steps. First, we match our two primary datasets: (a) Dimensions, containing academic publication information, with the (b) TTO dataset. We extract from Dimensions the names of all researchers affiliated with the TTO's university at any time. Next, we use a fuzzy matching algorithm to match the researchers' names from Dimensions to those of researchers who disclosed inventions in the TTO data. The matched dataset contains publications and invention disclosures matched by author name. For the 2000 to 2020 period, 4,367 researchers listed in the Dimensions data are matched to the TTO data and linked to 53,180 unique publications and 4,505 inventions.

Given, however, that common authorships of a paper and an invention do not imply a match between a specific paper and the invention, we take two additional steps to achieve a match. First, we evaluate how temporally proximate a paper is to an invention disclosure. Second, we assess the textual similarity between an article and the invention disclosure.

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

We determine a paper's influence on an invention—matching a paper to an invention—based on whether the time gap falls within a specific range. Specifically, we use a time window of [-1,3] years. This range is based on discussions with the TTO and the TTO guidelines, which advise inventors to disclose their inventions before public dissemination to maintain patenting options in jurisdictions without a one-year grace period post-publication. Research also indicates that scientific publications leading to patents are often temporally close to each other (Azoulay et al., 2007; Marx and Fuegi, 2020). Applying this method identifies 3,173 researchers linked to 19,381 publications and 3,127 inventions.[13]

Our last step to match publications with inventions is based on textual similarity. We employ a technique similar to the one described earlier, using BERT to generate textual embeddings for both the titles of papers and inventions. For each potential publication-invention pair (identified by having a common author and falling within the [-1,3] time window), we calculate the cosine similarity between the embeddings of their titles. Based on our analysis, we conclude that matches with similarity scores above 0.5 likely indicate publications that have influenced an invention.[14]

This three-step process matched 13,445 unique publications to 2,728 inventions, with 2,717 researchers linked to these matches. The median number of publications associated with each invention is 2, a finding consistent with other studies examining paper-patent pairs (e.g., Marx

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

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

and Fuegi, 2020). Following this, we prepare two datasets for our analyses. The first dataset is aggregated at the article level. It includes information about each article, including whether it is linked to an invention disclosed to the TTO, its commercial and scientific potential, the number of times it is cited by renewed patents, and other relevant characteristics. This dataset comprises 96,564 articles, of which 13,445 (13.92%) are associated with an invention disclosure. Table B.1, Appendix B, describes the variables used in this dataset and the other exercises. Table 1, Panel A, provides the summary statistics for these articles.

The second dataset is aggregated at the level of invention disclosures. Here, we examine the relationship between the commercial potential of disclosure and outcomes like TTO investment, patent filings, licensing agreements, and revenue generation. Since inventions are often linked to multiple articles, we average each invention's relevant variables (such as commercial potential, scientific potential, and patent citations). This dataset includes 2,728 inventions. Table 1, Panel B, provides the summary statistics for these inventions.[15]

[Table 1 about here.]

3 External Validity

We examine the external validity of our measure in two analyses. First, using our TTO data, we regress the measure of commercial potential against the decisions and outcomes realized in the process of technology transfer. Second, we expand our analysis to encompass the scholarly output of a substantial portion of all research-active universities in the United States. Here, our objective is to assess whether research contributions with high commercial potential (per our measure) eventuate in commercial outcomes as represented by their citation in a renewed patent.

[15] In Appendix B, Table B.2, we present the correlations between the key variables of interest.

3.1 Commercial Potential and Technology Transfer at a Leading U.S. University

In the first analysis, we regress our commercial potential measure against: 1. faculty decisions regarding the disclosure of their inventions to the TTO; 2. the decisions of TTO experts who evaluate the commercializability of those disclosures; 3. firms' decisions about contracting with the university to obtain access to the intellectual property in question; and 4. the realization of revenue.

Thus, we will examine the relationships between our measure and key commercialization stages: disclosure of article-linked inventions to the TTO, TTO investment decisions, licensing, agreements, and revenue generation. The analysis also includes controls for additional factors, such as the invention's scientific potential and the inventors' prior experience with the TTO.

We assume the scientists decide to disclose their inventions to the TTO based, at least partly, on their beliefs about the commercializability of their research. Figure 1 presents a density plot showcasing the commercial potential of scientific articles, our primary variable of interest. Comparing the density distributions for university-associated papers that are not linked to a TTO invention disclosure versus those that are linked to invention disclosures, the figure clearly shows that, compared to the former, the commercial potential of the latter is much greater, providing a foundation for our subsequent analyses. Likewise, Table C.1, Appendix C, shows the relationship between the commercial potential of an article and its likelihood of disclosure to the TTO. The findings indicate that articles in the lowest quartile have a 4.62% chance of being disclosed, whereas those in the highest have a 24.74% chance, which is 5.35 times greater.

[Figure 1 about here.]

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

$$y_{i,t} = \beta_0 + \beta_1 \phi_{i,t-1} + \beta_2 \psi_{i,t-1} + \beta_3 log(\alpha^{hs}_{i,t-1} + 1) + \theta_{it} + \epsilon_i, \quad (1)$$

where, for a scientific discovery reported in an article i published in year t, $y_{i,t}$ is a binary

variable representing whether the article i is associated with an invention commercial outcome, $\phi_{i,t-1}$ represents the article's commercial potential based on the model trained with data up to year $t-1$, and, similarly, $\psi_{i,t-1}$ represents its scientific potential. $\alpha^{hs}_{i,t-1}$ represents the scientific H-index of the paper's author at time $t-1$, which we log-transform due to its skewness. We call the H-index scientific prominence. If a paper has more than one author, $\alpha^{hs}_{i,t-1}$ is the maximum H-index among all authors j who authored the paper, i.e., $\alpha^{hs}_{i,t-1} = \max_j(\alpha^{hs}_{j,t-1})$.[16] 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, which could also affect the outcomes.

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

In Models 3 to 5, we expand our analysis by incorporating the control variables of scientific potential and researcher H-index, which reflects the author's scientific prominence. Our primary model, Model 5, confirms the significant role of commercial potential. An increase of one standard deviation in the commercial potential score correlates with a 6.9 percentage point increase in the disclosure probability—a 46% difference. Notably, the coefficient for commercial potential in this more comprehensive model (Model 5) remains similar in magnitude to that of Model 2, confirming that our measure of commercial potential is tied to researchers' disclosure decisions.

[Table 2 about here.]

Following specification 1, Table 3 presents a detailed analysis of the relationship between our commercial potential measure and later-stage outcomes in the technology transfer process, including the TTO's decision to patent as well as invest in the invention, agreements with firms, licenses and revenue. Note that while disclosure reflects a decision on the part of the scientists,

[16] All specifications are robust to using the average H-index and the sum of H-index of a paper's authors. Table B.1 further describes how the H-index is construed.

TTO investment in an invention and the decision to patent both reflect decisions on the part of the TTO. In contrast, an agreement and a license reflect a firm's decision to build on the invention, and revenue reflects a commercial outcome for a firm. The following results are expressed as percentage point increases associated with a one standard deviation change in the commercial potential measure. The data reveals a clear pattern: higher commercial potential correlates with increased likelihood across all stages.

[Table 3 about here.]

Consistent with Table 2 results, the probability of an invention being disclosed to the Technology Transfer Office (TTO) increases by 6.9% (a 44% increase over the baseline). We now also observe that the likelihood of receiving TTO investment increases by 5.6% (51% increase over baseline), while the chances of obtaining a patent rise by 4.5% (49% increase over baseline). The data also shows a 4.2% increase in the likelihood of reaching an agreement with a firm (43% increase over baseline), a 1.8% increase in the chances of securing a license (36% increase over baseline), and a 0.7% increase in generating revenue (41% increase over baseline). These results collectively indicate that a higher commercial potential of an invention is not only associated with its initial disclosure but also with its subsequent progression through the stages of commercialization. Another notable result is that the scientific potential of an article(s) linked to an invention has little relationship with any of the outcomes other than the realization of revenue. This stands in contrast to the scientific prominence of the faculty inventor(s), which is related to both the TTO's decisions and licensing on the part of firms. For the TTO and firms, the scientific prominence of a faculty member may signal the credibility of the inventor, or may provide the basis for a search heuristic employed by firms in their search for commercially promising science.

We next condition our analysis on the existence of an invention disclosure and will use the invention disclosure rather than the article as our unit of observation. We first examine the relationship between our measure of commercial potential (aggregated to the level of an invention disclosure) and the two key TTO decisions: the decision to invest in the invention and the decision to patent. The nature of TTO investment varies depending on the invention's field; in some cases, it involves legal protection and licensing costs, while in others, it encompasses marketing expenses.

Regardless, the amount invested in commercializing an invention indicates the TTO's belief in its commercial promise. Therefore, we expect inventions based on commercially promising science to receive more investment. Second, we observe the number of patents the TTO files for a given invention, another proxy for the TTO's expectations regarding an invention's value.

On the right-hand side, in addition to our measure of commercial potential, we include as a control whether the faculty inventors have had prior experience working with the TTO; we also interact the inventors' TTO experience with the invention's commercial potential to account for the possibility that TTO managers invest in more experienced teams to reduce investment risk. Other controls include the scientific potential of the science associated with the invention and the authors' scientific prominence (H-index). The econometric specification, which resembles specification 1, is as follows:

$$y_k = \beta_0 + \beta_1 \phi_{k,t-1} + \beta_2 \alpha^{tto}_{k,t-1} + \beta_3 \alpha^{tto}_{k,t-1} \phi_{k,t-1} + \beta_4 \psi_{k,t-1} + \beta_5 log(\alpha^{hs}_{k,t-1} + 1) + \Theta_{kt} + \epsilon_k, \qquad (2)$$

In Table 4, Models 1 to 4 present the results regarding TTO investment and patenting. The findings indicate that projects with high commercial potential are more likely to receive TTO investment and patent protection. These results remain robust in the main specifications (Models 2 and 4) after controlling for the inventor's prior experience with the TTO and the invention's scientific potential. In Model 2, we observe that one standard deviation increase in commercial potential from the mean increases the probability of investment by 8% (from 50.0% to 54.0%) and the probability of patenting by the TTO by 9% (from 52.2% to 56.7%).[17] Notably, after controlling for the commercial potential of an invention, the inventors' prior experience with the TTO appears to have no effect in both the decision to invest in and to patent an invention. The inventor's scientific prominence and its potential, are, however, associated with a greater likelihood

[17] Contrasting with the previous specification, in which we find that articles with commercial potential are 36 to 51% more likely to experience commercial outcomes, the current analysis, conditioning on disclosure and TTO investment, reflects a reduced discriminatory power with smaller, yet still notable, differences of 8-9%. This diminished discriminatory power is driven by the fact that we now condition on the subset of articles that are tied to invention disclosure that, as seen in Figure 1, are more homogeneous in terms of commercial potential—that is, we have less variation. This is not surprising to the extent that faculty decisions to disclose already reflect a judgment about commercializability.

of the TTO's decision to patent and invest in the invention. These results raise a question that we are not able to address: Whether scientific prominence and the scientific potential of the articles linked to the invention actually contribute to the commercializability of the invention, or do they distract from an accurate reading of that potential on the part of the TTO?

[Table 4 about here.]

We interpret the findings in Models 1 to 4 as further validation for our measure. Once, however, we condition the analysis on TTO investment in Models 5 through 9, the predictive power of commercial potential on the outcomes of licensing, startup formation, venture capital (VC) investment, and revenue generation is significantly diminished.[18] Only for agreements (Model 5) is our commercial potential measure is still predictive via its interaction with an author's prior experience with the TTO. These results suggest that the initial TTO investment decision may already encapsulate much of the commercial potential measure's predictive value.

3.2 The Measure of Commercial Potential and its Realization at U.S. Research Institutions

In this section, we conduct further temporal generalization of our measure and expand our validation exercise to encompass the scientific contributions of research-focused universities in the United States, examining publications from 2000 to 2020. For those articles published after each of our twenty models' training periods, we investigate whether those that are predicted to possess significant commercial potential are eventually commercialized (i.e, whether they are cited in at least one renewed patent).[19]

[18] Figure D.1 in Appendix D provides a visual interpretation of the results.

[19] This exercise differs from the usual training-test validation split used to calculate the AUROC in section 3.1, where we randomly divide the model learning sample into training, test and validation groups without looking at whether the training data came after the test sample data. In contrast, the validation method we use in this section is both out-of-sample and out-of-time-period, providing temporal generalization for our models and predictions. For each observation (e.g., an academic paper published in year t), we predict its commercial potential using a model trained only on publications released up to the year $t-4$, with outcomes observed up to the year $t-1$. For instance, the model that predicts the commercial potential for an article published in 2000 uses only data from articles available up to 1996, including patent citations to those articles collected by 1999. In the same way, the model for an article published in 2015 uses article data up to 2011 and patent citations to those articles collected by 2014. Therefore, our predictions are not just out-of-sample but also cover only earlier time periods, providing a

Our dataset for this analysis consists of 5,211,133 articles spanning the eleven academic fields described above. Our analysis focuses on articles authored by researchers affiliated with commercially active U.S. research institutions. To identify these research institutions, we first adhere to the 'R1: Doctoral Universities – Very high research activity' designation from the Carnegie Classification of Institutions of Higher Education as of 2021. Furthermore, to identify commercially active research institutions, we rely on membership in the Association of University Technology Managers (AUTM), which stipulates a minimum of 0.5 full-time equivalents (FTE) staff dedicated to technology transfer. This criterion yields 126 U.S. universities. Our results are robust to other approaches for defining our sample of institutions—including a sample with all U.S. universities, regardless of AUTM membership. See Table 1, Panel C, for the relevant statistical characteristics of this sample.[20]

We first conduct a descriptive analysis to examine the correlation between our commercial potential measure and the citation of a paper in a renewed patent. As shown in Table C.2, Appendix C, 15.56% of articles in the top quartile of the commercial potential distribution are cited by at least one renewed patent, a rate 21 times higher than articles in the bottom quartile (0.72%).

Below, we test the predictive validity of our commercial potential measure in our sample of over five million articles. We first estimate a linear probability model of the likelihood of a paper being cited in a renewed patent as a function of commercial potential, examining the additional variance explained by adding our measure to a specification containing institution and field-year fixed effects at the article level. We then test the incremental predictive validity of the measure by including other citation-based predictors for commercial potential such as the lagged h-index for authors, institutions and journals. These tests enable us to assess the additional predictive value provided by our measure of commercial potential beyond what can already be predicted by these ex-post citation-based measures. These additional predictors include the commercial and scientific prominence of the university from which the research originated and those of the authors

stronger and more conservative test for measure validity. This approach prevents the contamination of predictions with information that a decision-maker will not possess, such as insights into the commercial viability of ideas or topics that have not yet realized outcomes by the time a focal paper is published.

[20]In Appendix B, Table B.3, we present the correlations between the key variables of interest.

and the journals involved. Prominence is quantified using H-indices and journal impact factors, as delineated in previous sections. We estimate the following specification:

$$y_i = \beta_0 + \beta_1 \phi_{i,t-1} + \beta_2 \psi_{i,t-1} + \boldsymbol{\beta \alpha^{high}_{i,t-1}} + \iota_{it} + \theta_{it} + \epsilon_i \tag{3}$$

where, for a paper i published in year t, y_i denotes whether a paper is cited by at least one renewed patent. Additionally, $\phi_{i,t-1}$ represents the commercial potential of the finding, and $\psi_{i,t-1}$ its scientific potential, as determined by models trained with data up to year $t-1$. $\alpha^{high}_{i,t-1}$ is a vector of binary variables representing whether the focal paper is associated with institutions, authors, and journals with a high commercial and scientific prominence. For institutions and authors, the assessment is based on whether the commercial H-index and scientific H-index are in the top 20% at $t-1$. Similarly, for journals, the criterion is whether the journal impact factor (and journal commercial impact factor) is in the top 20% (refer to Table B.1 for a detailed explanation of how institution H-indices are computed at the paper level). The term ι_{it} denotes an institution fixed effect, while θ_{it} represents a grouped field-year fixed effect at the paper level. This is employed to account for technological shocks and trends across the field of the paper i in year t.

In Table 5, comparing models 1 and 2, we observe that adding our commercial potential measure to a specification containing institution and field-year fixed effects increases the R^2 by 42% (0.128-0.090/.090), an impressive increase in explanatory power. We next compare models 2 and 3 and observe that relative to a specification with six citation-based predictors, model 2, containing only our single commercial potential measure, provides ((.128-.116)/.116) 10% greater explanatory power than the six predictors combined. More strikingly, in model 4, we test for the incremental predictive validity of our measure and observe a considerable increase versus model 3 in R^2 of ((.139-.116)/.116) 19.8%. This indicates that our measure can improve the variation explained beyond fixed effects and other measures of commercial impact by nearly 20%, suggesting that the measure substantially adds new information. Finally, in Table 5, Model 5, we introduce our full specification. Notably, our measure has statistical significance even when accounting for other proxies for commercial potential, and contributes substantial explanatory power. A one standard deviation increase in commercial potential from the mean increases the likelihood of a

patent citation from 7.40% to 12.1%—a 63.5% increase. Finally, the coefficient estimates for the control variables reflecting commercial track records of the researcher, the institution, and the journal are substantially reduced upon inclusion of our measure (see Figure C.1, Appendix C). Researcher commercial prominence, for example, shows a notable reduction of 33%, institutional commercial prominence by 23%, and journal by 24%, all differences statistically significant with p-values lower than 0.01. It is worth noting that the model incorporates fixed effects at the institution level, effectively accounting for most of the variation across institutions. In contrast, variables linked to scientific prominence do not display similar coefficient changes. In summary, our results further validate our measure of science's commercial potential.

[Table 5 about here.]

4 Illustrative applications

We demonstrate the value of our measure in two illustrative exercises concerned with longstanding empirical questions. In the first exercise, we assess the extent to which the "privatization" (i.e., patenting) of scientific knowledge produced by public research institutions diminishes the diffusion of this knowledge across firms. We consider this question for one university, utilizing the TTO data previously mentioned. In the second exercise, we analyze the role that universities' reputations for producing commercializable science may have on the use of their science by industry. Importantly, our measure of the commercial potential of science allows us to distinguish the effect of reputation per se versus that of the role of reputation as simply a proxy for universities' abilities to produce commercializable science. An implication of the analysis is that the commercializable science of less prominent universities is disproportionately overlooked by industry.

4.1 One university's privatization of science and the scientific commons

A question that has long roiled innovation scholars and especially those academics and policymakers concerned with the Bayh Dole Act is whether universities' attempts to privatize their

scientific discoveries through patenting, licensing, and related activities have depleted the "scientific commons". A critical component of this debate is whether such practices restrict the diffusion of knowledge, limiting the use of that knowledge, not just by academics but by restricting the number of firms that build upon the knowledge that underpins the patents. Such restrictions may consequently reduce the overall contribution of academic research to economic growth (c.f., Dasgupta and David, 1994; NRC, 2011; Nelson, 2004).

For example, Murray and Stern (2007), using patent-paper pairs, argue that university research is cited less frequently by academics after it has been patented. Conversely, university patenting and licensing could enhance knowledge diffusion through various mechanisms. One such mechanism, for instance, is a better incentive system for researchers, who, with the prospect of better appropriating the returns from their discoveries, may be willing to put more effort towards the development of their knowledge (Thursby and Thursby, 2002; Jensen and Thursby, 2001).

As noted by some (e.g., Henderson et al., 1998; Murray and Stern, 2007; Mowery et al., 2015), a definitive answer to the question of the impact of academic patenting on the breadth of use of knowledge by firms remains elusive partly due to the challenges involved in considering the counterfactual scenario. Specifically, it is difficult to observe how much knowledge would flow to industry without university patenting. For instance, if TTOs were to patent all the commercially viable research at universities, effectively ensuring that the entire "at-risk" set passes through them, then any observed increase in the firms' use of patented university knowledge could be explained not by TTO patenting but simply by the fact that the knowledge in question was commercially promising. In sum, as discussed above, to obtain unbiased estimates of the effect of academic patenting on the use of science, one needs to control for those factors that may affect both university patenting and firms' use of the scientific knowledge in question, and a critical step in that direction, as argued by Azoulay et al. (2007) and Marx and Hsu (2022), is controlling for the commercial potential of science.

This section does not aim to provide a definitive answer to the question. Instead, we seek to demonstrate how a measure of the commercial potential of science can contribute to such an analysis. We do so by investigating the relationship between one university's patenting of science

and its subsequent utilization while controlling for the science's commercial potential.

First, we assess the probability that a scientific article patented by a university is subsequently cited by a renewed corporate patent. The results are presented in Table 6, across Models 1 to 4. These are linear probability models with a binary dependent variable indicating whether an article has been cited by a renewed corporate patent and with fixed effects for publication-field year. Commercial Potential is also a binary variable, indicating whether an article is in the top quartile of the distribution at this university and, thus, is a high commercial potential article. Finally, Patented is a binary variable indicating whether the TTO has patented the invention associated with the knowledge.

Model 1 evaluates the likelihood of firms utilizing scientific knowledge based on an article's commercial potential, revealing that articles within this university's highest commercial potential quartile are 4.3 times more likely to be cited by a renewed corporate patent than others (8.6% vs. 2.0%). Model 2 examines the baseline effect of patenting, where the expected direction of the effect could vary: a negative coefficient would suggest that university patenting deters subsequent use. In contrast, a positive coefficient would indicate diffusion. The data show an increased likelihood of firms using knowledge when patented by the TTO, implying that university patenting may actually increase the subsequent use of scientific knowledge. Compared to articles not linked to a patented invention, we note that those linked are 1.75 times more likely to be cited by a corporate patent (7.7% vs. 3.3%). However, this increase could also reflect selection bias, as discussed. The university may be more likely to patent science with higher commercial potential, and so are firms.

Models 3 and 4 explore this issue. With the inclusion of commercial potential in Model 3, the "Patented" coefficient decreases by 32%. Moreover, when "Patented" interacts with commercial potential, the coefficient drops by 50% from its original value. These findings imply that selection effects are indeed significant. They also suggest that, even after accounting for the commercial potential of the underlying knowledge, the associated likelihood of firm use of scientific knowledge increases when the university has patented the knowledge. Specifically, Model 4 indicates that, if patented by the university, scientific knowledge with high commercial potential has a 12.1% chance of being cited in corporate patents, in contrast to an 8.0% chance for comparably high

commercial potential science not patented by the university—a significant 51% difference.[21]

[Table 6 about here.]

This analysis, though of interest, does not address the relationship between academic patenting and the breadth of use of university science across firms. Relying solely on citation counts in renewed patents leaves open the question of the number of firms accessing patented versus unpatented science. For instance, the higher citation rate for patented science may indicate the same firm repeatedly building on the patented science across its various inventions.

Models 5 to 8 in Table 6 consider the relationship between academic patenting and breadth of use by firms of the associated articles. Using the same analytical framework, our dependent variable is the number of distinct firms utilizing the science.[22] The results are qualitatively similar to those previously discussed. The coefficient for Patenting decreases significantly—by 56%—after controlling for commercial potential. Most importantly, articles with high commercial potential linked to TTO-patented inventions are cited by a larger number of firms compared to those not patented. Specifically, Model 8 reveals that articles with high commercial potential patented by the TTO are cited by 55% more firms than articles of similar commercial potential not associated with TTO patenting.[23]

From these findings, it could be inferred that the privatization of science by a university does not inhibit but may enhance its utilization by firms. Several reasons may account for this. For instance, the articles linked to the patented inventions could be more visible to firms, or firms may use TTO patenting decisions to indicate which academic research warrants attention. However,

[21]The results are robust to using the count of renewed corporate patent citations instead of a binary indicator. Likewise, "high commercial potential" in this exercise was defined as publications that are in the top quartile of our commercializability score. To probe further whether the unpatented and patented science of high commercial potential are indeed comparable with respect to commercializability, we compute the mean commercial potential scores and find that the mean commercial potential score for the patented science is 0.889 (sd = .047) and for the unpatented science is 0.890 (sd = 0.053), indicating comparability.

[22]To calculate the dependent variable, we compile all the assignees listed in the patents citing a paper and then eliminate duplicates. Note that patent assignee information can sometimes be misleading due to inconsistencies in naming (e.g., Apple Inc. and Apple Computer, Inc.) and the failure to account for subsidiaries. Despite these potential sources of error, such inconsistencies are likely orthogonal to the "treatment" of TTO patenting, meaning errors should be equally distributed across patents citing a paper, irrespective of whether the TTO patented the paper.

[23]We conducted all the analyses in this section and the appendix using articles associated with inventions that the university has invested in, not just those that have been patented. The findings are not just qualitatively similar; they are nearly quantitatively identical.

these reasons are a matter of speculation, and we leave the identification of the mechanism to future research. Moreover, the findings from this illustrative exercise do not suggest that the disclosure process to the TTO and its subsequent patenting decisions are the predominant factors influencing the impact of academic science on corporate innovation. Indeed, the majority of high-potential academic science at this university is never disclosed to the TTO (see Table C.1), much less patented—only 15.36% of high commercial potential articles at this university are patented by its TTO. Consequently, while patented science may receive more citations or attention from a more significant number of firms compared to unpatented commercializable science, the vastly greater share of commercializable academic science (per our measure) from this university that is unpatented indicates that publication and public disclosure are the primary means through which academic research influences firms' innovation activities.

4.2 The production and realization of commercially promising scientific research in the U.S.

In this section, we consider one piece of the question of how to account for differences in the commercialization rates of science across universities. Specifically, we consider how a university's reputation for producing commercializable science impacts the commercialization of its scientific research. We begin with a simple empirical model for which the dependent variable is the commercialization rate for a university's science, quantified as the share of its publications cited in a renewed patent. For the independent variable, we use the university's commercialization H-index, which reflects the university's reputation, or prominence. Accordingly, a university is categorized as a "high reputation university" if its commercial H-index is in the top 20% of all universities' H-indices and as a "low reputation university" otherwise. The time frame under consideration is from 2000 to 2015.[24] In this first model, we do not account for our measure of commercial potential, meaning we do not adjust for the commercial promise of the science produced by these universities. Our findings indicate a significant disparity: 14.09% of articles from prominent universities

[24]Patents typically cite papers that were published, on average, 14 years before the patent grant (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.

are cited by at least one renewed patent, compared to 9.96% from less prominent universities—a 41.5% increase in the commercialization rate when contrasting the latter with the former.

These statistics do not, however, distinguish between publications with low and high commercial potential. Consequently, the coefficient estimate may simply reflect that a top-ranked institution has capabilities that allow it to produce more commercializable science in the recent past—thus generating its reputation. Alternatively, the coefficient estimate may reflect the role of reputation as a focusing device employed by firms to minimize search costs in the face of a daunting amount of science that a firm would otherwise need to examine to find science of commercial use (Bikard and Marx, 2020). The coefficient may well reflect either or both of these effects. For our illustrative purpose, we will probe the relationship between universities' reputations and firms' use of their commercializable science.

Accordingly, we use our measure to confine our analysis to only those papers with high commercial potential (i.e., those in the top quartile of commercial potential scores). This focus on the "risk set" of paper (i.e., those publications most likely to be commercialized *ex-ante*) allows us to eliminate the conflation of the effect of reputation with that of the production of commercializable science. First, we find that reputation per se, apart from the production of commercializable science, appears to have an effect on use. There is, however, a substantial reduction in the commercialization rate gap between those top and bottom-ranked institutions when we restrict our attention to only those articles with high commercial potential. Specifically, the relative differential shrinks to 33.6% from the earlier 41.5%. Thus, there is an effect of reputation, but its relationship to firms' use of the science is much reduced when we remove differences in the production of commercializable research[25] (see Figure 2, Panel A).

[Figure 2 about here.]

We push this analysis of the role of reputation further by examining, at the article level, the relationship between not only the reputation of the institution in affecting the use of science but

[25] Also, unsurprisingly, when the focus narrows to articles with high commercial potential, the patent citation rates surge. Collectively, across all fields, 28.28% of articles from top-ranked institutions and 21.17% from bottom institutions reach commercialization. Moreover, even for those article with high commercial potential, a significant 58.5% of the research does not transition to commercial use, even at more prominent, research-intensive universities.

also those of individual researchers as well as the journals in which they publish, while controlling for field, year, and university fixed effects. Concerning the latter, firms, for example, may be more likely to consult those journals that publish more applied work—more oriented toward solving the kinds of practical problems with which firms are concerned. To explore the relationship between the commercialization rates of commercializable science and the reputation of universities for producing such science, we estimate the following:

$$y_i = \beta_0 + \beta_1\phi_{i,t-1} + \beta_2\psi_{i,t-1} + \boldsymbol{\beta\alpha}^{\boldsymbol{high}}_{\boldsymbol{i,t-1}} + \boldsymbol{\beta\alpha}^{\boldsymbol{high}}_{\boldsymbol{i,t-1}} \times \phi_{i,t-1} + \iota_{it} + \theta_{it} + \epsilon_i \tag{4}$$

where, for a paper i published in year t, y_i denotes whether a paper is cited by at least one renewed patent. Additionally, $\phi_{i,t-1}$ represents the commercial potential of the finding, and $\psi_{i,t-1}$ its scientific potential, as determined by models trained with data up to year $t-1$. $\alpha^{high}_{i,t-1}$ is a vector of binary variables representing whether the focal paper is associated with the commercial and scientific prominence of institutions, authors, and journals. For institutions and authors, the assessment of prominence is based on whether the commercial H-index and scientific H-index are in the top 20% at $t-1$. Similarly, the criterion for journals is whether the journal impact factor (and journal commercial impact factor) is in the top 20%. The term ι_{it} denotes an institution fixed effect, while θ_{it} represents a grouped field-year fixed effect at the paper level. This is employed to account for technological shocks and trends across the field of the paper i in year t.[26]

The results in Model 1 of Table 7 highlights outcomes that underscore institutional disparities in commercialization rates. Research with high commercial potential originating from more commercially prominent institutions is more likely to be cited in renewed patents than similar research from less renowned institutions. This is evident from the positive and statistically significant interaction term "Commercial potential x High commercial impact institution". Holding other variables constant at their means, articles in the top quartile of commercial potential are 14.65% likely to be cited in a renewed patent if they originate from an institution with high commercial prominence.

In contrast, similarly valuable articles from other institutions have a citation likelihood of

[26]Refer to Table B.1 for a detailed explanation of how variables are construed at the paper level.

12.26%—resulting in a 19.49% difference in citation rates. Model 2 reveals that the effect of institutional prominence persists even when considering the journal of publication. Moreover, the journal's commercial impact factor also plays an important role. An article in the top quartile of commercial potential is 16.29% likely to be cited in a renewed patent if it is published in a journal with a high commercial impact factor, versus a 9.35% likelihood for articles in journals with lower impact factors—a 74% difference in citation likelihood. Finally, in Model 3, we incorporate the commercial prominence of individual researchers and find that the individual researcher effect accounts for the differences previously attributed to institutions, rendering the institutional coefficients insignificant (while not diminishing the journal effects). An article in the top quartile of commercial potential authored by a prominent researcher has a 17.19% likelihood of being cited in a renewed patent, compared to 10.14% for those from non-prominent authors—a gap of 69.52%. Additional analyses also reveal that such "realization gaps" exist across different scientific fields (see Figure D.4, Appendix D).

[Table 7 about here.]

Table 7 offers two additional insights. First, in Model 3, the significance of the researcher coefficients, as opposed to those of a university, is interpreted as being due to the correlation between the two. Prominent universities tend to employ prominent researchers, which accounts for the shifting significance from the institution to the individual. We infer that the university's effect is driven by the caliber of researchers it employs. Second, our results also suggest that it is the generation of commercializable research that contributes to the disparities among universities' commercialization rates rather than a mere bias towards certain institutions. We infer this from the fact that the coefficients for more prominent universities, journals, and researchers alone (not interacted) are insignificant. We would expect these coefficients to be significant if firms favored these notable universities regardless of their output. However, this is not the case; significance is only present when there is an interaction, indicating that firms target these institutions to capitalize on the commercializable science that their faculty produce.

In summary, our findings suggest that research with commercial potential is more likely to be neglected when the institution, author, and journal lack commercial prominence. Why might this

occur? Those in a position to commercialize academic science, such as firms and venture capitalists, may preferentially focus their search on sources with track records of generating research with commercial applications. While sensible, this finding implies that comparably commercializable science from less commercially prominent sources is being overlooked to the detriment of both firms and society.

Our example, though focusing on the impact of institutional reputation, makes a more general point. It shows that, for understanding the sources of variation in commercialization rates, whether across individuals or institutions, it is crucial to understand whether, in the first instance, the issue is one of universities' production of commercializable science or, in the second, one of firms' identification and use of the commercializable science that is at hand. Indeed, both stages—creation versus use—as one might imagine, are essential. At the institutional level, however, we find that 55.23% of the variation in commercialization rates across universities is associated with differences in their production of commercially applicable research.[27] Thus, the preponderance of the difference in commercialization rates across universities is due to differences in production. This nonetheless leaves almost 45% of the variation unexplained, of which some share is undoubtedly associated with firms' identification and use of that research (Lerner et al., 2022).

This distinction between the creation or production versus the identification of commercializable science matters. For policymakers, university administrators, and managers concerned with increasing firms' use of science, this distinction will affect the decision levers they could use to affect commercialization rates. If, for example, the effect of reputation reflects a university's ability to produce commercializable science, that would suggest a need for policy levers that affect production, including, say, university policies affecting academic promotion or government policies such as those supporting STEM education. If, however, what is going on reflects challenges tied to firms' identification and use of commercializable science, that suggests a different set of policy levers. Such might include managers' decisions to strengthen their firms' ties to academic institutions, university TTO's efforts to market their commercializable science more effectively, or

[27]To compute the variance explained by the production of commercializable research, we initially calculate each organization's rate of renewed patent citations to papers across its entire paper portfolio, along with the mean commercial potential of the papers produced by the organization. Subsequently, we regress an institution's commercialization rate on its average commercial potential.

government policies affecting firms' incentives to build on academic research (e.g., Bayh-Dole).

5 Discussion

Scientific research is fundamental to driving technological advance and economic growth, yet understanding how nascent discoveries become inputs into the development of commercial products remains a significant challenge. One difficulty lies in distinguishing the commercial potential of scientific findings from their actual commercialization. Our research tackles this issue by developing an *ex-ante* measure of the commercial potential of academic science using large language models and neural networks designed to predict the likelihood of scientific articles contributing to marketable products or processes. The method for computing *ex-ante* measures of commercial potential at the article level entails training predictive models on the abstracts of academic articles to predict their incorporation into renewed patents. In addition to the standard validation exercise using a holdout sample, we also conduct two external validation exercises employing both out-of-sample-period as well as cross-domain or transfer validation, including an analysis of an article's progression through one university's technology transfer process. The paper concludes with two illustrative empirical exercises demonstrating the measure's utility for answering substantive research questions bearing on firms' commercialization of academic science.

In our first external validation exercise, we analyze administrative data on the technology transfer process at a leading U.S. university. Our measure of commercial potential successfully predicts actual outcomes—various milestones in the transfer process. It predicts not only invention disclosures to the Technology Transfer Office (TTO) but also subsequent TTO investment, patenting, licensing agreements, and revenue generation, outcomes reflecting decisions on the part of faculty, TTO experts, firms, and customers.

We then further test the validity of our measure using an out-of-sample and out-of-time period exercise using data from 126 major U.S. research institutions, and five million articles across various academic fields from 2000 to 2020. We examine whether articles predicted ex-ante to have high commercial potential are cited at higher rates in renewed patents. The findings indicate that articles in the top quartile of commercial potential are 21 times more likely to be cited by a

renewed patent than those in the bottom quartile. Incorporating our measure into a formal regression analysis, we see that a one standard deviation increase in commercial potential significantly increases the likelihood of patent citation, with the model's explanatory power increasing by 20% upon including the commercial potential variable.

We then conduct two illustrative exercises to highlight the utility of our measure for advancing understanding of the commercialization of science. The first investigates whether (and how) universities' efforts to privatize—i.e., patent—their scientific output affect the breadth of firms' use of academic science. Our findings from an analysis of over 96,000 articles from a major research university suggest that university patenting is not associated with a decline in the utilization of commercializable science by firms. In fact, we observe the opposite: patenting of highly commercializable science is associated with an increased breadth of use of academic science by firms. The inclusion of our measure of commercial potential, however, dampens that positive effect.

Our second illustrative exercise expands our research to include 5.2 million articles published in the United States from 2000 to 2020. The purpose of the exercise is to demonstrate the utility of our measure in addressing questions bearing on why commercialization rates differ across universities. For purposes of illustration, we take on a piece of this: the role that a university's reputation for producing commercializable science may play in explaining a university's commercialization outcomes. We do find a significant relationship between reputation and commercialization outcomes. Most importantly, we also find that 55% of the observed variation in commercialization rates is due to differences in universities' production of science. Thus, without controlling for the relevant risk set of commercializable research, one will overestimate the influence of reputation and, in all likelihood, the role of other factors thought to condition firms' identification and use of academic science. Nonetheless, we do observe that reputation–and more so that of individual researchers than that of institutions–has an impact on firms' use of *commercializable* academic science. An implication of this latter finding is that commercially promising science from less prominent researchers and institutions is more likely to be overlooked. [28]

Our measure of the commercial potential of science and the associated methodology and code

[28] We are not arguing that this greater realization gap. Rather, firms' use of reputation as a focusing device is a sensible search strategy for minimizing search costs.

have numerous applications in empirical studies related to the "science of science" and the economics of innovation. A significant challenge in addressing various questions in the economics of innovation is the potential unobserved heterogeneity in the commercial potential of scientific research (Marx and Hsu, 2022), which may correlate with crucial variables of interest, such as gender (Koffi and Marx, 2023; Ding et al., 2006) or status (Azoulay et al., 2010). Our measure of commercial potential can act as a control variable, whether through its direct inclusion in regression analyses, in matching estimators, or as part of an instrumental variables strategy to mitigate these concerns. Beyond econometric applications, our metric could also be useful in benchmarking the commercial potential of science and the realization of that potential across different fields, institutions, researchers, and regions within a country, as well as on an international scale. Moreover, this metric could aid in identifying the factors associated with the production of science with commercial potential, as well as in examining the barriers to the realization of that potential. Finally, our measure of the commercial potential of science could serve as a proxy for the otherwise unobserved "technological opportunity," long considered by economists to be a critical determinant of the innovative activity and performance of firms (Cohen, 2010).

Our work, of course, has limitations. The reliance on patent data and the assumption that citations from renewed patents to papers reflect the commercial potential of a scientific contribution can be questioned despite this assumption being widely accepted in the literature (Kuhn et al., 2020). For instance, many scientific contributions transition to the market without an associated patent. Additionally, the current model may only partially capture the commercial potential of scientific contributions due to variable and sometimes indirect paths to commercialization, not to mention the potentially long time horizons before a contribution may be embodied in a new product or process (cf., Adams, 1990; IIT, 1968). Furthermore, our Natural Language Processing-based technique may miss factors beyond those captured in textual content that influence the decision to utilize a specific piece of science in technology development, and our analysis does not consider the nature of such errors in our predictions. In our exercises, such errors are likely to lead to greater noise in our measure of commercial potential. We, however, see promise in the general methodology as more data are incorporated into the prediction models.

References

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

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

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

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

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

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

Bikard, M. (2020). Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal*, 41(8):1528–1543.

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

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

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

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Cohen, W. M. (2010). Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*, 1:129–213.

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.

Dasgupta, P. and David, P. A. (1994). Toward a new economics of science. *Research policy*, 23(5):487–521.

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

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

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

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

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.

Hsu, P.-H., Lee, D., Tambe, P., and Hsu, D. H. (2020). Deep learning, text, and patent valuation. *Text, and Patent Valuation (November 16, 2020)*.

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

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.

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.

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

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

Lerner, J., Stein, C., and Williams, H. (2022). The wandering scholars: Understanding the heterogeneity of university commercialization. *Working paper*.

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.

Mowery, D. C., Nelson, R. R., Sampat, B. N., and Ziedonis, A. A. (2015). *Ivory tower and industrial innovation: University-industry technology transfer before and after the Bayh-Dole Act*. Stanford University Press.

Murray, F. and Stern, S. (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization*, 63(4):648–687.

Nelson, R. R. (2004). The market economy, and the scientific commons. *Research policy*, 33(3):455–471.

NRC, N. R. C. O. T. N. A. (2011). Managing university intellectual property in the public interest.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

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

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

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

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

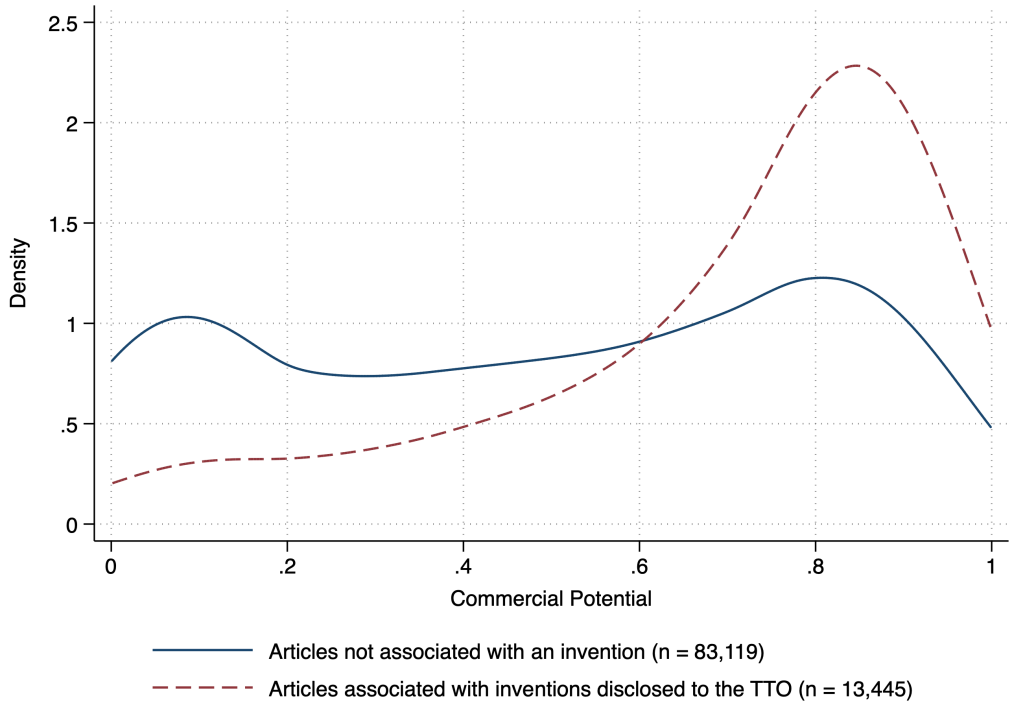
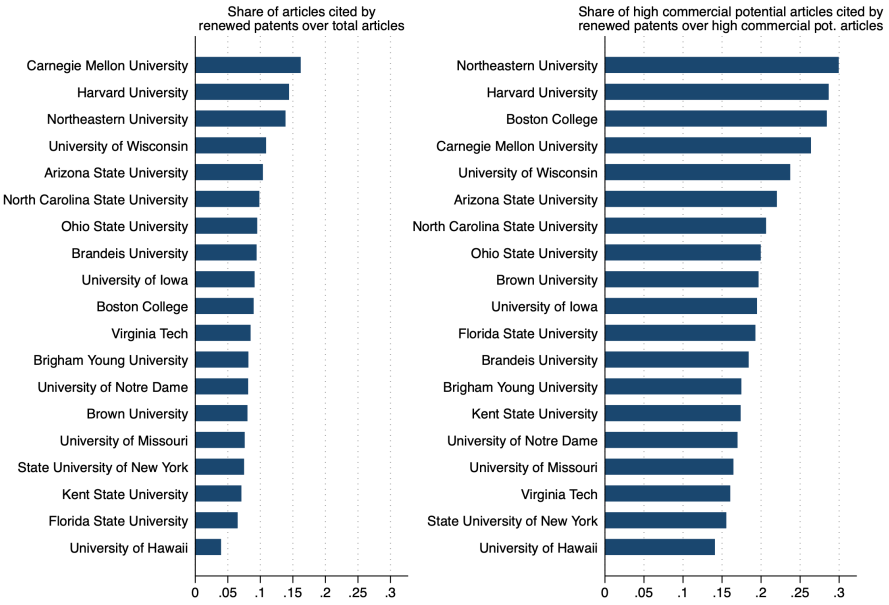
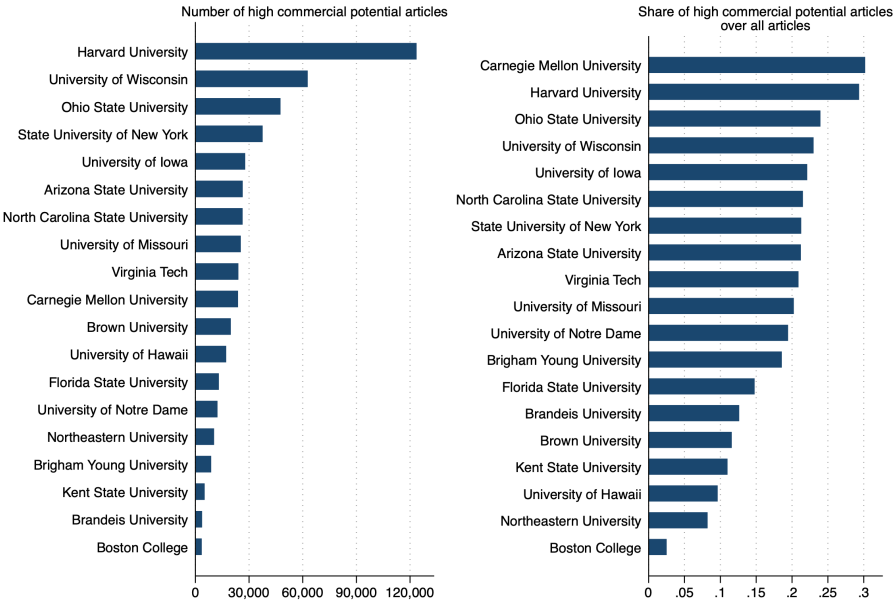


Figure 1: Bi-weight kernel density estimates of the distributions of the commercial potential of 1) articles published at this university not associated with an invention disclosure (solid line) and 2) articles associated with inventions disclosed to the Technology Transfer Office (dashed line). Articles tied to an invention are more likely to have high commercial potential.



(a) Panel A. Differences in the *translation* of scientific research into commercial applications—patent citations. The figure on the left plots the share of articles cited by at least one renewed patent over all articles. The figure on the right plots the share of high commercial potential articles cited by at least one renewed patent over high commercial potential articles.



(b) Panel B. Differences in the *production* of high commercial potential research. The figure on the left plots the total number of high commercial potential articles produced. The figure on the right plots the share of high commercial potential articles over the total number of articles produced.

Figure 2: Differences in the production and translation of scientific research produced between 2000 and 2015 across randomly selected U.S. universities.

Table 1: Summary statistics. Panel A summarizes the relevant features for articles whose authors were affiliated with the TTO's university at the time of publication, 2000-2020. Panel B summarizes the relevant features of the articles *associated* with disclosed inventions, 2000-2020. For confidentiality reasons, invention-level outcomes are removed (Investment, Patents, Agreements, Licensing, Revenue, Startup, and VC funding). Panel C summarizes U.S. scientific research published between 2000 and 2020 in U.S. R1 Universities with an active TTO.

Panel A: Articles from TTO's university (N = 96,564)

	Mean	SD
Commercial potential	0.52	0.31
Scientific potential	0.73	0.20
Academic cites	62.54	210.09
Patent cites	0.71	5.77
Cited by patent	0.11	0.31
Cited by renewed patent	0.08	0.27
Author(s) scientific prominence	45.26	31.04
Disclosed	0.14	0.35

Panel B: TTO inventions (N = 2,728)

	Mean	SD
Commercial potential	0.73	0.21
Scientific potential	0.76	0.15
Academic cites	74.95	140.32
Patent cites	2.41	9.38
Cited by patent	0.46	0.50
Cited by renewed patent	0.37	0.48
Author(s) scientific prominence	49.47	28.76
Author(s) TTO experience	0.68	0.46

Panel C: Articles from R1 U.S. Universities with active TTOs (N = 5,211,133)

	Mean	SD
Commercial potential	0.49	0.33
Scientific potential	0.66	0.24
Cited by patent	0.10	0.30
Cited by renewed patent	0.07	0.26
Institution(s) commercial prominence	68.70	40.54
Institution(s) scientific prominence	411.40	227.77
Journal commercial impact factor	0.02	0.05
Journal scientific impact factor	3.08	3.05
Author(s) commercial prominence	4.23	4.76
Author(s) scientific prominence	32.39	25.59

Table 2: Linear probability models estimating, for a publication published at year t, the likelihood of disclosure as a function of commercial potential. The dependent variable is a binary variable indicating whether a paper is associated with an invention disclosed to the TTO. Model 1 presents the baseline impact of the fixed effects (field-year) on disclosure. Model 2 shows that the measure of commercial potential, $\phi_{i,t-1}$, trained with data up to $t-1$, predicts whether a scientific publication will be associated with a disclosure well above the fixed effects. Model 5 presents the full specification, controlling for the scientific potential ($\psi_{i,t-1}$) and the scientific prominence of a publication's authors at time $t-1$ ($log(\text{H-index}_{t-1}+1)$). Fixed effects are included at a publication field-year level in all models.

DV: Disclosed	(1)	(2)	(3)	(4)	(5)
Commercial Potential		0.238***		0.232***	0.221***
		(0.016)		(0.016)	(0.016)
Scientific Potential			0.140***	0.034***	0.012
			(0.012)	(0.009)	(0.008)
Author Scientific Prominence					0.027***
					(0.004)
Constant	0.139***	0.015*	0.037***	-0.007	-0.083***
	(0.000)	(0.008)	(0.009)	(0.008)	(0.014)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564
R-squared	0.025	0.061	0.029	0.061	0.064

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

Table 3: Linear probability model estimating the likelihood that a publication published at time t is associated with an invention that (1) is disclosed to the TTO, (2) receives TTO investment, (3) the TTO files patents for it, (4) leads to commercial agreements, (5) leads to licensing to firms, and (6) generates positive revenue. All dependent variables are binary. Commercial Potential—$\phi_{i,t-1}$, trained with data up to $t-1$—strongly predicts all the outcome variables. The models control for the scientific potential ($\psi_{i,t-1}$) and the scientific prominence of a publication's authors at time $t-1$ ($log(\text{H-index}_{t-1}+1)$). Fixed effects are included at a publication field-year level in all models.

	(1) Disclosed	(2) Investment	(3) Patent	(4) Agreement	(5) License	(6) Revenue
Commercial Potential	0.221***	0.180***	0.146***	0.137***	0.057***	0.023***
	(0.016)	(0.013)	(0.011)	(0.011)	(0.006)	(0.003)
Scientific Potential	0.012	-0.002	0.002	0.013**	0.010*	0.013***
	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)	(0.003)
Author Scientific Experience	0.027***	0.026***	0.019***	0.026***	0.014***	0.002**
	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
Constant	-0.083***	-0.092***	-0.066***	-0.089***	-0.045***	-0.013***
	(0.014)	(0.012)	(0.010)	(0.011)	(0.007)	(0.004)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564	96,564
R-squared	0.064	0.058	0.048	0.054	0.026	0.015

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

Table 4: Linear probability models estimating the impact of Commercial Potential, $\phi_{i,t-1}$, on commercial outcomes at the invention level. An invention's Commercial Potential is computed as the average commercial potential of the articles associated with it. Models 1 and 2 use as a dependent variable whether an invention receives investment, and Models 3 and 4 whether the TTO files at least one patent. The average commercial potential of the articles tied to an invention strongly predicts both variables. Furthermore, we analyze whether an invention leads to commercial agreements (Model 5) and licensing deals (6); is commercialized via a Startup (7) and commercialized via a Startup with VC funds (8); and generates revenue (9). All dependent variables are binary indicators. We add controls for the author's previous experience disclosing inventions to the TTO (binary) as well as for the authors' scientific prominence at time $t-1$ ($log(\text{H-index}_{t-1}+1)$). We also control for the scientific potential of the articles associated with an invention, $\psi_{i,t-1}$. Fixed effects are included at an invention field-year level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Invested	Invested	Patented	Patented	Agreement	Licensed	Startup	VC Investment	Any revenue
Commercial Potential	0.296***	0.250***	0.315***	0.259***	-0.106	-0.051	-0.025	0.073	-0.048
	(0.047)	(0.075)	(0.050)	(0.079)	(0.160)	(0.150)	(0.104)	(0.069)	(0.102)
Author TTO Experience		0.119		0.107	-0.203*	0.105	0.027	0.128	0.108
		(0.082)		(0.076)	(0.119)	(0.135)	(0.104)	(0.084)	(0.107)
Comm. Pot. x TTO Experience		-0.090		-0.063	0.352**	0.008	0.033	-0.100	-0.113
		(0.108)		(0.100)	(0.161)	(0.161)	(0.129)	(0.108)	(0.139)
Author Scientific Prominence		0.042***		0.065***	0.071**	0.020	-0.002	-0.021	-0.003
		(0.015)		(0.017)	(0.029)	(0.030)	(0.023)	(0.020)	(0.023)
Scientific Potential		0.315***		0.185**	-0.170	-0.211	0.054	0.070	-0.087
		(0.080)		(0.083)	(0.117)	(0.139)	(0.106)	(0.089)	(0.106)
Constant	0.277***	-0.123	0.280***	-0.104	0.615***	0.416**	0.102	0.041	0.263*
	(0.034)	(0.084)	(0.037)	(0.092)	(0.169)	(0.183)	(0.129)	(0.093)	(0.149)
Invention field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,689	2,689	2,689	2,689	1,305	1,305	1,305	1,305	1,305
R-squared	0.115	0.126	0.125	0.136	0.186	0.132	0.161	0.160	0.173

Standard errors clustered at the Invention Category - Year level

* $p < .1$, ** $p < .05$, *** $p < .01$.

Table 5: Linear probability model estimating the probability of a paper being cited by at least one renewed patent. Model 1 shows the baseline effects of the fixed effects (publication field-year and university). Model 2 shows the effect of our commercial potential measure, $\phi_{i,t-1}$—a 42.22% increase in explained variation from Model 1. Model 3 contains potential correlates of commercialization outcomes: the commercial and scientific prominence of the originating universities and authors, with prominence measured using the H-index at time $t-1$ (log(H-index$_{t-1}$), as well as commercial and scientific impact journal. Model 4 presents the results with the commercial potential measure ($\phi_{i,t-1}$), and Model 5 adds our scientific potential measure ($\psi_{i,t-1}$) as an additional control. Fixed effects are incorporated at the field-year and university levels in all specifications.

DV: Cited by renewed patent	(1)	(2)	(3)	(4)	(5)
Commercial potential		0.181***		0.148***	0.142***
		(0.019)		(0.015)	(0.015)
High commercial impact institution			0.009***	0.007***	0.007***
			(0.002)	(0.002)	(0.002)
High scientific impact institution			0.002	0.005***	0.004***
			(0.002)	(0.002)	(0.002)
High commercial impact journal			0.051***	0.038***	0.039***
			(0.005)	(0.004)	(0.004)
High scientific impact journal			-0.011*	-0.010**	-0.011**
			(0.006)	(0.005)	(0.005)
High commercial impact researcher			0.096***	0.064***	0.064***
			(0.008)	(0.006)	(0.006)
High scientific impact researcher			-0.004**	-0.001	-0.003
			(0.002)	(0.002)	(0.002)
Scientific potential					0.036***
					(0.005)
Constant		-0.015	0.040***	-0.023**	-0.043***
		(0.009)	(0.004)	(0.009)	(0.011)
Publication field - year FE	Yes	Yes	Yes	Yes	Yes
University-FE	Yes	Yes	Yes	Yes	Yes
Observations	5,211,133	5,211,133	5,211,133	5,211,133	5,211,133
R-squared	0.090	0.128	0.116	0.139	0.140

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

Table 6: Models 1 to 4 are linear probability models estimating the likelihood that a corporate renewed patent cites an academic article at the TTO university. Models 5 to 8 estimate the count of different firms citing an article in their patents. High Commercial Potential is a binary variable indicating whether the article is a the top quartile of commercial potential, and Patented is a binary variable indicating whether the article is associated with an invention patented by the TTO. Fixed effects are included at a publication field-year level for all models.

	(1) Cited by firm patent	(2) Cited by firm patent	(3) Cited by firm patent	(4) Cited by firm patent	(5) Count citing firms	(6) Count citing firms	(7) Count citing firms	(8) Count citing firms
High Commercial Potential	0.066***		0.063***	0.061***	0.062***		0.059***	0.057***
	(0.010)		(0.009)	(0.009)	(0.009)		(0.009)	(0.009)
Patented		0.044***	0.030***	0.022***		0.041***	0.028***	0.018***
		(0.006)	(0.005)	(0.005)		(0.006)	(0.005)	(0.004)
High Commercial Potential x Patented				0.019**				0.022***
				(0.007)				(0.007)
Constant	0.020***	0.033***	0.019***	0.019***	0.017***	0.029***	0.015***	0.016***
	(0.002)	(0.000)	(0.003)	(0.002)	(0.002)	(0.000)	(0.002)	(0.002)
Publication field - Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	96,564	96,564	96,564	96,564	96,564	96,564	96,564	96,564
R-squared	0.073	0.059	0.075	0.075	0.073	0.058	0.074	0.075

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

Table 7: Linear probability models estimating the likelihood of a paper being cited by at least one renewed patent, focusing on the commercial potential ($\phi_{i,t-1}$) and its interactions with indicators of high prominence related to institutions, researchers, and journals. High prominence is defined by binary variables indicating if an article's affiliated institution, researcher, or journal ranks in the top 20 percentile of the H-index or journal impact factor. The model includes fixed effects for both field year and university. Interaction terms reveal that publications with high commercial potential are more likely to be cited in renewed patents when associated with high-impact institutions, researchers, or journals.

DV: Cited by renewed patent	(1)	(2)	(3)
Commercial potential	0.164***	0.140***	0.117***
	(0.017)	(0.014)	(0.012)
Scientific potential	0.033***	0.035***	0.032***
	(0.005)	(0.005)	(0.005)
High commercial prominence institution	-0.007	-0.008	-0.004
	(0.008)	(0.007)	(0.006)
Commercial potential x High commercial prominence institution	0.039**	0.040**	0.023
	(0.018)	(0.017)	(0.015)
High scientific prominence institution	-0.002	-0.003	-0.002
	(0.007)	(0.006)	(0.006)
Commercial potential x High scientific prominence institution	0.021	0.021	0.015
	(0.015)	(0.014)	(0.013)
High commercial impact journal		-0.032***	-0.030***
		(0.008)	(0.008)
Commercial potential x High commercial impact journal		0.127***	0.120***
		(0.014)	(0.014)
High scientific impact journal		0.039***	0.039***
		(0.011)	(0.012)
Commercial potential x High scientific impact journal		-0.081***	-0.083***
		(0.019)	(0.020)
High commercial prominence researcher			-0.007
			(0.012)
Commercial potential x High commercial prominence researcher			0.098***
			(0.019)
High scientific prominence researcher			-0.003
			(0.003)
Commercial potential x High scientific prominence researcher			0.003
			(0.007)
Constant	-0.032***	-0.036***	-0.031***
	(0.010)	(0.009)	(0.009)
Publication field - year FE	Yes	Yes	Yes
Institution FE	Yes	Yes	Yes
Observations	5,211,133	5,211,133	5,211,133
R-squared	0.130	0.137	0.144

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

Appendix to:

Measuring the Commercial Potential of Science

Appendix A Model training and outcomes

A.1 Processing the input text

Our methodology relies on large language models and natural language processing (NLP) techniques, which use text as input. Specifically, we use the abstracts of the articles in which findings are reported. The pre-trained language model we use is SciBERT (Beltagy et al., 2019), which in turns derives from BERT, a language model created by Google (Devlin et al., 2018). Pre-trained language models, such as BERT and SciBERT, create accurate representations of documents in a high-dimensional space. This is achieved through algorithms that convert text documents into embeddings—numeric vectors serving as representations of the document's content. This capability is highly valuable, as it enables various tasks based on these embeddings. Because its trained with scientific, domain-specific text, SciBERT provides state-of-the-art performance in a wide range of natural language processing tasks for scientific domains, improving BERT's performance. We tested whether this holds in our classification task and, indeed, our models' performance increases when using SciBERT instead of BERT.

SciBERT relies on transformers (Vaswani et al., 2017), a novel type of neural network architecture.[29] In short, as opposed to previous natural language processing techniques, transformers can model long-range dependencies and learn contextual representations, being able to "understand" complex semantic relationships within and across documents.[30] The first step we undertake consists of "tokenizing" the abstracts, i.e., converting each abstract into an array of discrete linguistic units—usually, units are words, parts of words, numbers, symbols, and stems. We tokenize using the version that SciBERT's authors recommend, *scibert-scivocab-uncased*, which is expected to yield the highest performance.[31]

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 its pre-trained embeddings (Token Embeddings)—a vector representing each word in a high-dimensional space in relation to an extensive vocabulary. In addition, the tokenizer adds information regarding the position of each token in the text, both in the sentence (Segment Embeddings) and in absolute terms (Position Embeddings). Combining the three embeddings produces a unique embedding for each token in the abstract, which serves as the input to the first layer of the neural network. This final embedding captures information about the token's relative position

[29] At a high level, a transformer model consists of multiple layers of self-attention and feed-forward neural networks, enabling it to weigh the probabilities of different parts of the input sequence (i.e., sentences of the text) and process it in parallel. The attention mechanisms allow transformers to learn contextual representations of words and phrases.

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

[31] SciBERT's tokenizer 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).

within a document, enabling the contextualization of its meaning when fine-tuning the models.

It is worth noting that, for computational reasons, SciBERT, like BERT, is limited to processing up to 512 tokens per document. There are various techniques to handle longer documents, but a simple analysis of the abstracts we use to train our model reveals that only 1% of them contain more than 512 tokens. Additionally, there are no differences in the average number of tokens between the classes (which could create bias in our findings). Therefore, we truncate the abstracts at 512 tokens. Figure A.1 shows the distribution of abstracts' length. Once the abstracts have been processed by the tokenizer, they are input to the neural network and the model is fine tuned based on the labels.

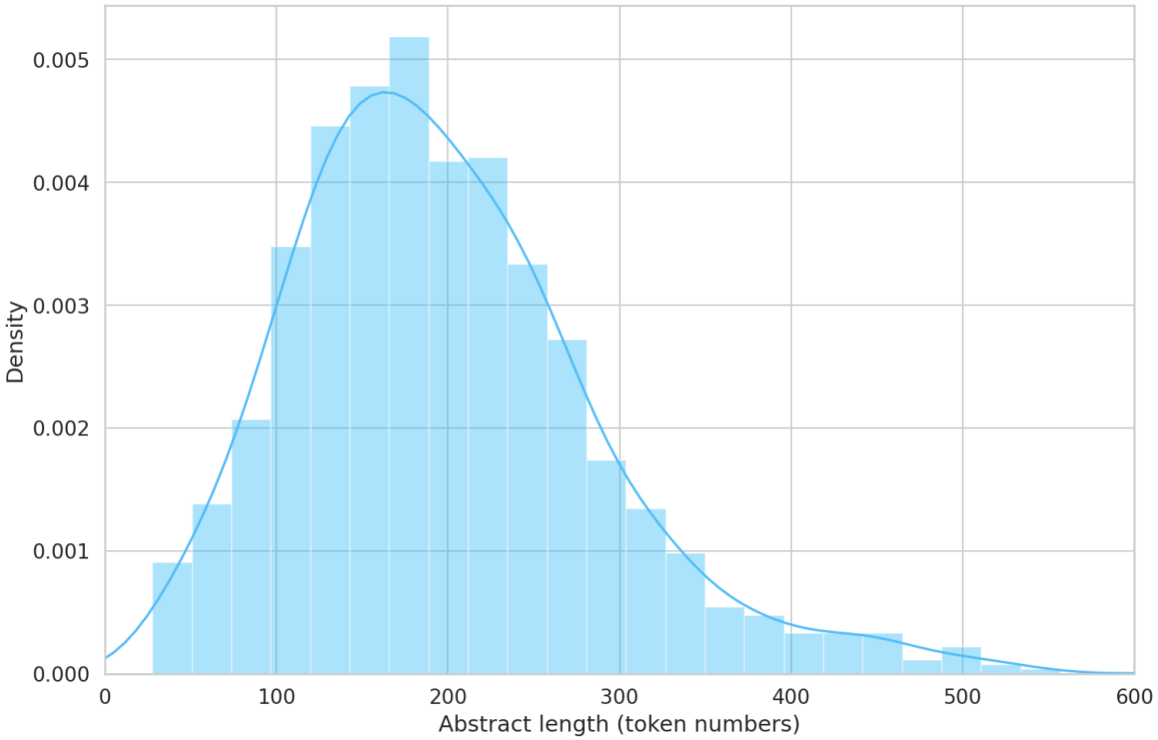


Figure A.1: Abstract's token length distribution

A.2 Model performance

Figures A.2 and A.3 detail performance statistics for each of the commercial potential models we trained—one per year—and, similarly, Table A.1 provides the average performance statistics for the 20 scientific potential models.

2000

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

2001

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

2002

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

2003

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

2004

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

2005

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

2006

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

2007

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

2008

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

2009

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

2010

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

2011

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A.3 Examples of scientific articles and their commercial potential

Table A.2: Selected scientific articles in the top 25 percentile of commercial potential.

Title	**Field**	**Institution**	**Journal**	**Year**	**Patent cites**	**Citing patent renewed?**
High-resolution mapping of protein sequence-function relationships	Biological Sciences	U. of Washington	Nature Methods	2010	26	Yes
Combination strategies to enhance anti-tumor ADCC	Biomedical and Clinical Sciences	Stanford	Immunotherapy	2012	9	Yes
Engineering Tumor-Targeting Nanoparticles as Vehicles for Precision Nanomedicine	Engineering	Rutgers	Med one	2019	0	No
Species-Specific and Inhibitor-Dependent Conformations of LpxC—Implications for Antibiotic Design	Chemical Sciences	Duke	Chemical Sciences & Biology	2011	6	Yes
Multi-Scale 2D Temporal Adjacency Networks for Moment Localization with Natural Language	Information and Computing Sciences	U. of Rochester	IEEE Transactions on Pattern Analysis and Machine Intelligence	2021	0	No
Nanophotonic projection system	Physical Sciences	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	Biological Sciences	U. of Southern California	Journal of The American Society of Nephrology	2018	0	No
Self-Healing Polyurethanes with Shape Recovery	Engineering	U. of Florida	Advanced Functional Materials	2014	7	Yes
Exploring mechanisms of FGF signalling through the lens of structural biology.	Biological Sciences	New York U.	Nature Reviews Molecular Cell Biology	2013	8	Yes
A high-energy-density sugar biobattery based on a synthetic enzymatic pathway	Chemical Sciences	Virginia Tech	Nature Communications	2014	11	Yes

Table A.3: Selected scientific articles in the bottom 25 percentile of commercial potential.

Title	**Field**	**Institution**	**Journal**	**Year**	**Patent Cites**	**Citing Patent Renewed?**
Extinction and Nebular Line Properties of a Herschel-selected Lensed Dusty Starburst at z = 1.027	Physical Sciences	Cornell University	International Journal of Mass Spectrometry	2015	0	No
An exotic invasive shrub has greater recruitment than native shrub species within a large undisturbed wetland	Biological Sciences	University of Wisconsin	Plant Ecology	2012	0	No
Dynamic programming solutions for decentralized state-feedback LQG problems with communication delays	Information and Computing Sciences	California Institute of Technology	Advances in computing and communications	2012	1	Yes
Effects of natural weathering on microstructure and mineral composition of cementitious roofing tiles reinforced with fique fibre	Engineering	Pennsylvania State University	Cement and Concrete Composites	2011	0	No
Thermodynamic database for the Co-Pr system	Chemical Sciences	Iowa State University	Data in Brief	2016	0	No
Hydrostatic equilibrium profiles for gas in elliptical galaxies	Physical Sciences	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	Engineering	U. Michigan	Nuclear Science and Engineering	2016	0	No
Turbulent cross-helicity in the mean-field solar dynamo problem	Physical Sciences	Stanford	The Astrophysical Journal	2011	0	No
A 4-year study of invasive and native spider populations in Maine	Biological Sciences	U. Massachusetts	Canadian Journal of Zoology	2011	0	No
Intrusion of a Liquid Droplet into a Powder under Gravity	Biomedical and Clinical Sciences	Princeton University	Langmuir	2016	0	No

A.4 Time horizon of the commercial potential measure

Table A.4 relates our commercial potential measure to the lag between the publication year of an article and the filing year of the first renewed patent citing the article. Our sample includes all papers published in the U.S. in the 2000-2020 period in the scientific and engineering fields of analysis described above. We create lag buckets that are based on lag quartiles. That is, 25% of the papers are cited in renewed patents either in years 0 or 1, 25% of the papers are cited in years 2 or 3, 25% of the papers are cited in years 4 or 5, and 25% of the papers are cited in year six and onwards. We find that articles in the top quartile of commercial potential are substantially more likely to be cited faster.

Table A.4: Patent citation lag (year) by commercial potential quartile.

Quantiles	**Time lag**				
of compot	0, 1 years	2,3 years	4, 5 years	6+ years	Total
1	61	80	62	131	334
	18.26%	23.95%	18.56%	39.22%	100.00%
2	294	322	216	370	1,202
	24.46%	26.79%	17.97%	30.78%	100.00%
3	1,009	998	632	763	3,402
	29.66%	29.34%	18.58%	22.43%	100.00%
4	2,028	1,662	902	940	5,532
	36.66%	30.04%	16.31%	16.99%	100.00%
Total	3,392	3,062	1,812	2,204	10,470
	32.40%	29.25%	17.31%	21.05%	100.00%

Likewise, Figure A.4 plots the equivalent Kaplan-Meier survival curves by commercial potential quartile, where the time of the event is the first time a paper receives a patent citation. Kaplan-Meier estimates provide a robust assessment of the findings, as the methodology is well-suited for our analysis in that accounts for varying time-to-event data and considers the timing and distribution of events, such as the lag between article publication and patent citation.

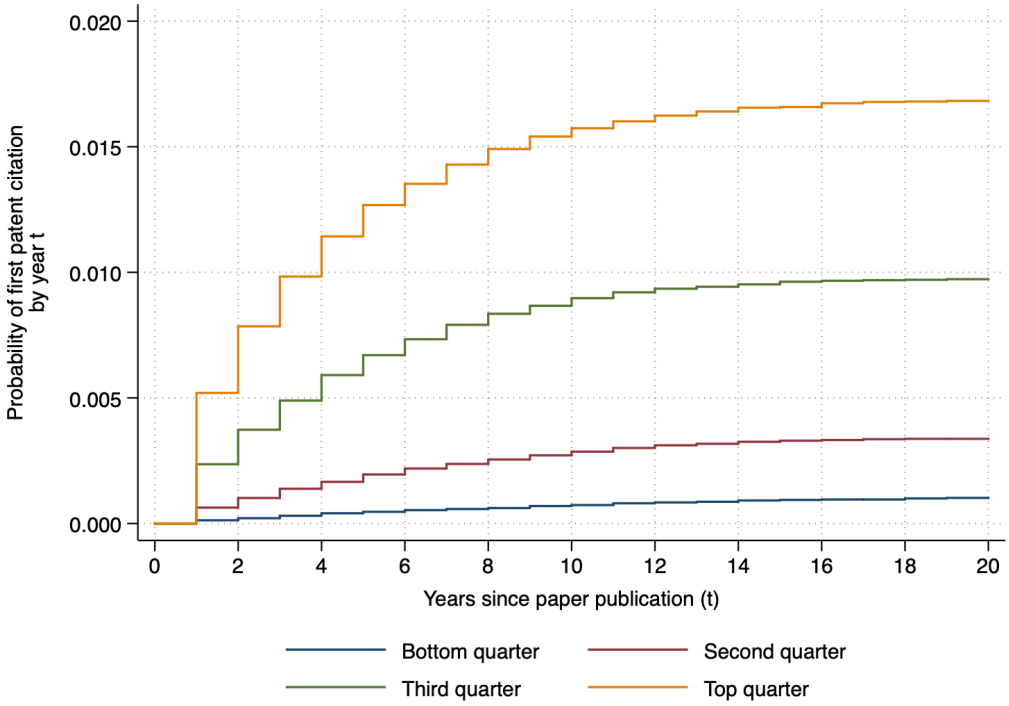


Figure A.4: Kaplan-Meier survival curves by commercial potential quartile.

Appendix B Variable descriptions and correlation tables

Table B.1: Variable descriptions.

Variable	Measure	Description of measure
Disclosed	Disclosed	Binary variable representing whether an article is tied to an invention disclosed to the TTO.
Investment	Investment	Amount invested ($) by the TTO to pursue the commercialization of an invention. Includes different natures of expenses, such as patenting and marketing expenses. The majority of the specifications use a binary variable, indicating whether an invention received any investment.
Patents	Patents	Number of patents the TTO filled to protect a given invention. The majority of the specifications use a binary variable, indicating whether at least one patent was filed.
Agreements	Agreements	Number of commercial agreements—of any nature—associated with an invention. The majority of the specifications use a binary variable, indicating whether at least one agreement was established.
Licenses	Licenses	For each invention, number of licensing agreements with third parties, such as firms or other institutions. The majority of the specifications use a binary variable, indicating whether at least one licensing agreement was established.
Revenue	Revenue	Amount of revenue ($) generated by the invention. The majority of the specifications use a binary variable, indicating whether the invention generated a positive revenue.
Startup	Startup	Binary variable indicating whether the invention has been commercialized via Startup.
VC Investment	VC Investment	Conditional on Startup, binary variable indicating whether the startup has raised venture capital financing.
Authors' TTO Experience	Authors' TTO Experience	Binary variable representing whether at least one of the authors/inventors associated with the invention, prior to the focal disclosure, has disclosed an invention to the TTO.
Commercial potential	ϕ = P(Patent renewal | patent cite > 0)	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 model, which uses the abstract text of the focal article to cast the prediction.
Scientific potential	ψ = P(Paper cite > 16)	Probability that more than 16 academic articles will cite the focal article. The probability is the output of our secondary model (Scientific potential), which uses the abstract text of the focal article to cast the prediction.
Author scientific prominence	Max of authors' scientific H-index	Author H-index at time $t-1$, excluding the focal article. If a paper is authored by more than one author, we use the maximum of the authors' scientific H-indices. The H-index captures the productivity and impact of an author and is calculated by counting the number of an author's publications that have 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 prominence	Max of author's commercial H-index	Author commercial H-index at time $t-1$, excluding the focal article. If a paper is authored by more than one author, we use the maximum of the authors' commercial H-indices. Similar to the scientific H-index, the commercial H-index is calculated by counting the number of publications cited by patents.
Institution scientific prominence	Max of institutions' scientific 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. If a paper is authored by more than one institution, we use the maximum of the institutions' scientific H-indices.
Institution commercial prominence	Max of institutions' commercial H-index	Idem as institution H-index, but using patent citations to papers instead of academic citations.
Journal scientific 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, 2022)).
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, 2022)).

Table B.2: 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	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Author scientific prominence	Disclosed	Author TTO experience
Commercial potential	1.000								
Scientific potential	0.212	1.000							
Academic cites	0.056	0.062	0.000	1.000					
Patent cites	0.106	0.021	0.339	1.000					
Cited by at least one patent	0.261	0.042	0.209	0.354	1.000				
Citing patent is renewed	0.224	0.041	0.201	0.389	0.843	1.000			
Author scientific prominence	0.181	0.240	0.073	0.002	0.009	-0.009	1.000		
Disclosed	0.219	0.049	0.030	0.070	0.129	0.119	0.072	1.000	
Author TTO experience	0.210	0.046	0.027	0.065	0.132	0.121	0.085	0.796	1.000

Panel B: Articles matched to inventions disclosed to the TTO

	Commercial potential	Scientific potential	Academic cites	Patent cites	Cited by at least one patent	Citing patent is renewed	Autho scientific prominence	Author TTO experience
Commercial potential	1.000							
Scientific potential	0.176	1.000						
Academic cites	0.043	0.053	1.000					
Patent cites	0.120	0.006	0.352	1.000				
Cited by at least one patent	0.206	-0.049	0.251	0.280	1.000			
Citing patent is renewed	0.166	-0.050	0.253	0.321	0.830	1.000		
Author scientific prominence	0.139	0.244	0.064	-0.053	-0.068	-0.104	1.000	
Author TTO experience	0.225	0.062	0.034	0.025	0.072	0.069	0.131	1.000

Table B.3: U.S. articles, 2000-2020.

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

Appendix C Validation

Table C.1: Percentage distribution of articles in the TTO university binned in four quartiles of commercial potential. Articles in the top quartile are 5.35 times more likely to be associated with an invention disclosed to the TTO than articles in the bottom quartile.

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

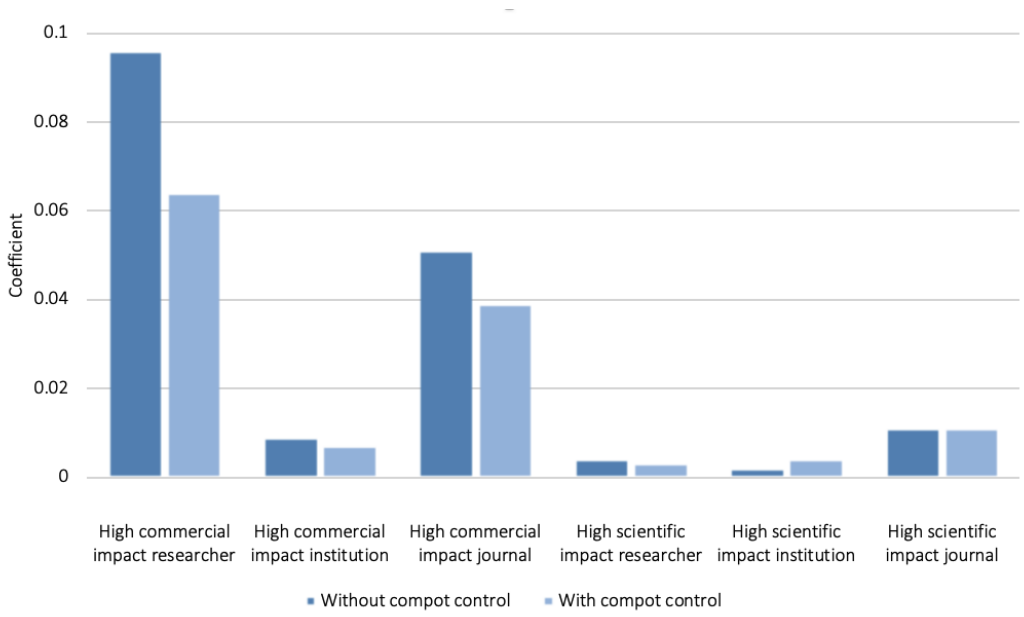


Figure C.1: Effect of commercial potential on variable coefficients in models predicting renewed patent citations to research articles. Upon introducing the commercial potential measure, a considerable shrinkage in coefficients is observed for variables associated with commercialization aspects. Researcher commercial experience shows a notable reduction of 33%, institution experience by 23%, and journal by 24%, all differences are statistically significant with p-values < 0.01. It is worth noting that the model incorporates fixed effects at the institution level, effectively accounting for most of the variation across institutions. In contrast, variables linked to scientific experience do not display similar changes in coefficients. The variations in these variables are either not statistically significant or marginal.

Table C.2: Percentage distribution of articles produced by U.S. organizations between 2000-2020 binned in four quartiles of commercial potential. Articles in the top quartile are 21.61 times more likely to be cited by a renewed patent than articles in the bottom quartile.

Commercial Potential Quartile	Not cited or cited by non-renewed patent	Cited by renewed patent	Total
1	1,293,401	9,383	1,302,784
	99.28%	0.72%	100.00%
2	1,257,142	45,641	1,302,783
	96.50%	3.50%	100.00%
3	1,174,594	128,189	1,302,783
	90.16%	9.84%	100.00%
4	1,100,060	202,723	1,302,783
	84.44%	15.56%	100.00%
Total	4,825,197	385,936	5,211,133
	92.59%	7.41%	100.00%

Appendix D Commercial potential and technology transfer at a leading U.S. university

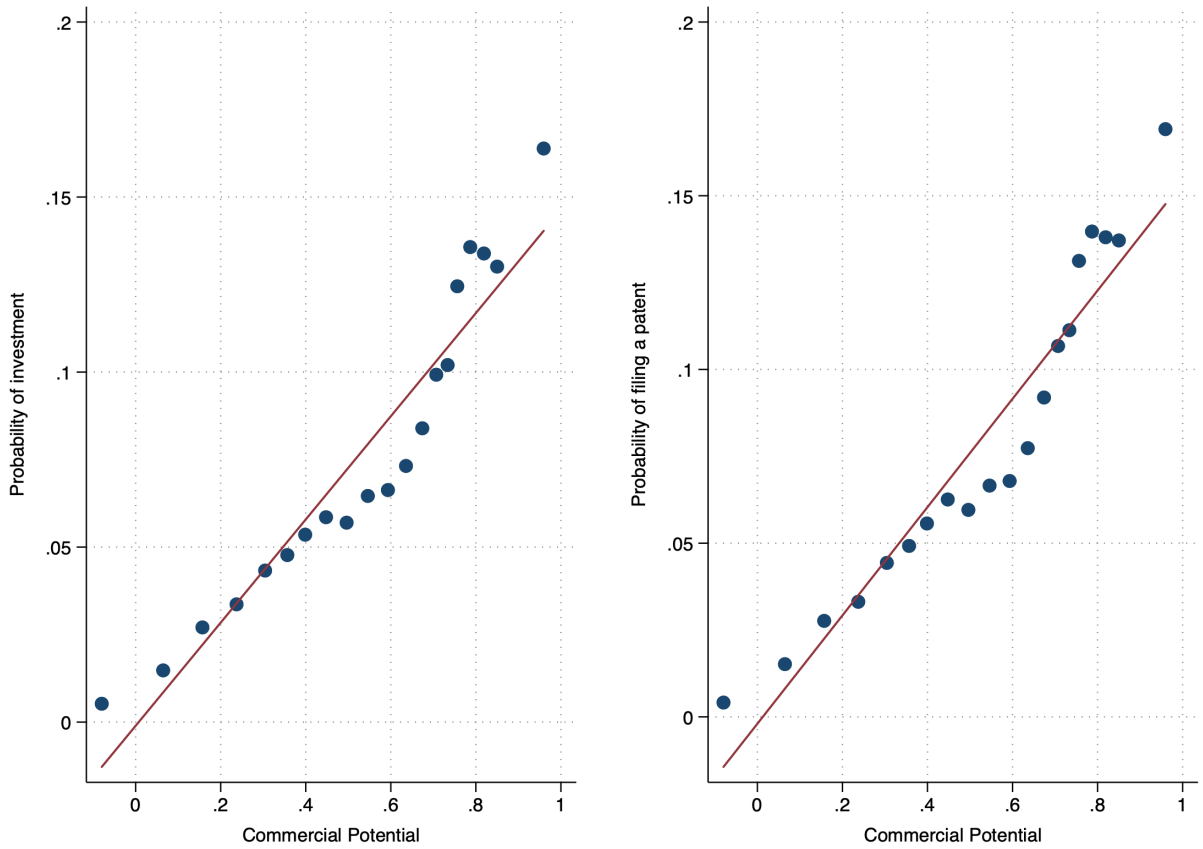


Figure D.1: Probability that the TTO will invest into (Panel A) and patent (Panel B) an invention based on the average commercial potential of the articles associated with the invention.

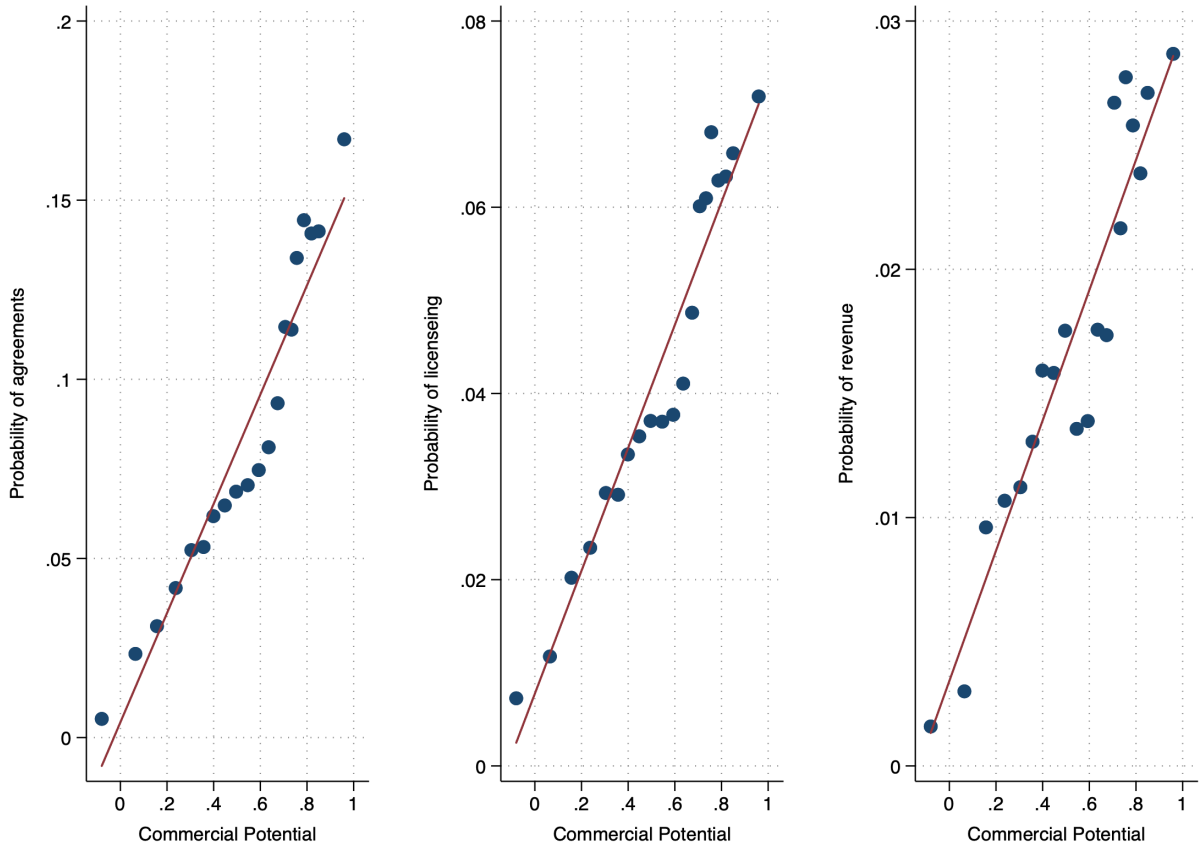


Figure D.2: Probability that an invention will garner agreements (Panel A) and licensing deals (Panel B), as well as generate revenue to the TTO (Panel C) based on the average commercial potential of the articles associated with the invention.

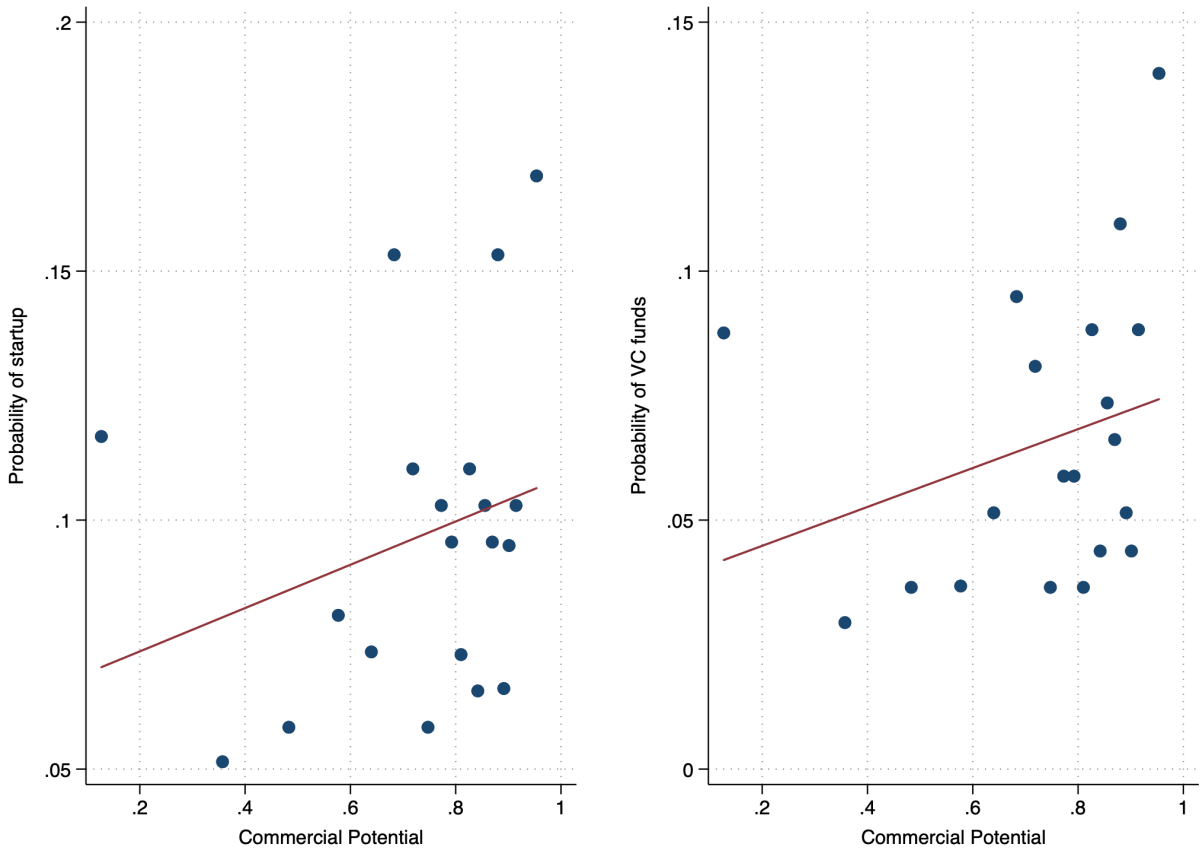


Figure D.3: Probability that an invention will be commercialized via a Startup (Panel A) and, conditional on Startup, that will raise venture capital funds as a function of the average commercial potential of the articles associated with the invention.

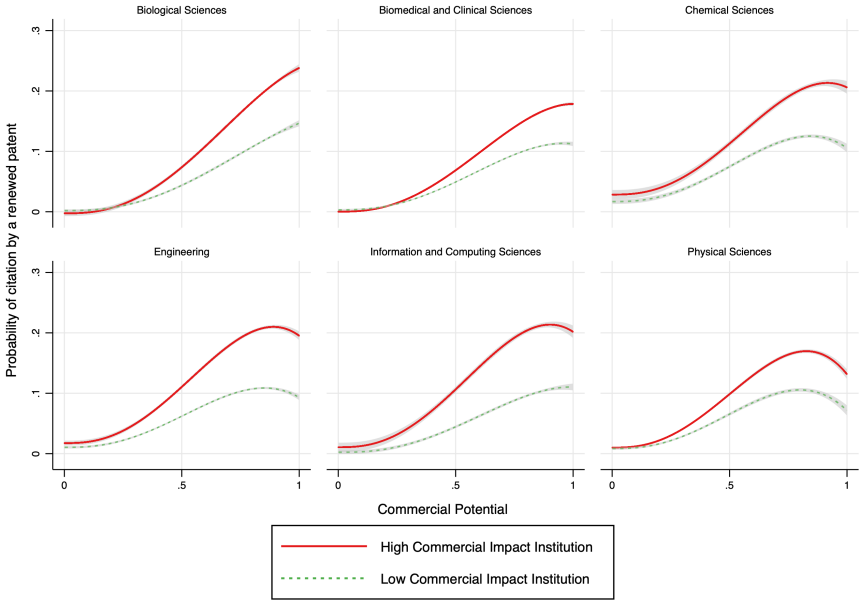


Figure D.4: Fractional-polynomial estimation of the probability of renewed patent citation as a function of commercial potential, by scientific field. Curves are plotted based on the commercialization impact of the institutions associated with an article—the solid line represents articles produced at institutions in the top 20% and the dashed line from the bottom 20%. The figure includes a 95% confidence interval for the estimation.