Measuring the Commercial Potential of Science

Roger Masclans*, Sharique Hasan* and
Wesley M. Cohen*+
*Duke University
+*NBER

NASEM, Innovation Policy Forum
May 15, 2024

Study Goal

- We develop a measure of the commercial **potentia**l of academic science to advance understanding of its commercial impact.
  - Academic science: Findings, discoveries, methods, etc. reflected in publications in 11 fields of the natural and applied sciences and engineering
- Why bother?
  - Identifying what academic science has commercial potential is an important first step toward understanding the factors affecting its production, as well as its commercial application.
    - Premise: Only a small share of academic science is commercializable (too abstract, embryonic, or irrelevant).
    - Consider translation
- Ignoring commercial potential leads to conflating factors that may affect the production of commercializable science with those affecting the process of commercialization given commercializable science in hand.

Adding to the state of the art

- Understood by many that we cannot rely on ex post measures such as forward citations in patents to study the determinants of the commercialization of academic science.
- With the application of sophisticated econometric techniques controlling for unobserved commercial potential, some prior work has identified frictions (e.g., location, gender) affecting the commercial application of science.
- But those techniques, while mitigating possible bias, cannot address important questions such as:
  - What attributes of scientists, institutions or regions lead to the production of commercializable research?
  - How much commercializable science goes undeveloped ("the realization gap")?
- For such questions, you need a measure of the "unobservable" commercial potential of academic science

Commercial potential?

- Our measure is intended to reflect:
  - The probability that a firm believes that a scientific article can contribute to the development of a marketable product or process

Commercial potential?

- Our measure is intended to reflect:
  - The probability that a firm believes that a scientific article can contribute to the development of a marketable product or process
- Measure of commercialization (i.e., the realization of commercialization potential)
  - The citing of a scientific article in a patent that is subsequently renewed.

Commercial potential?

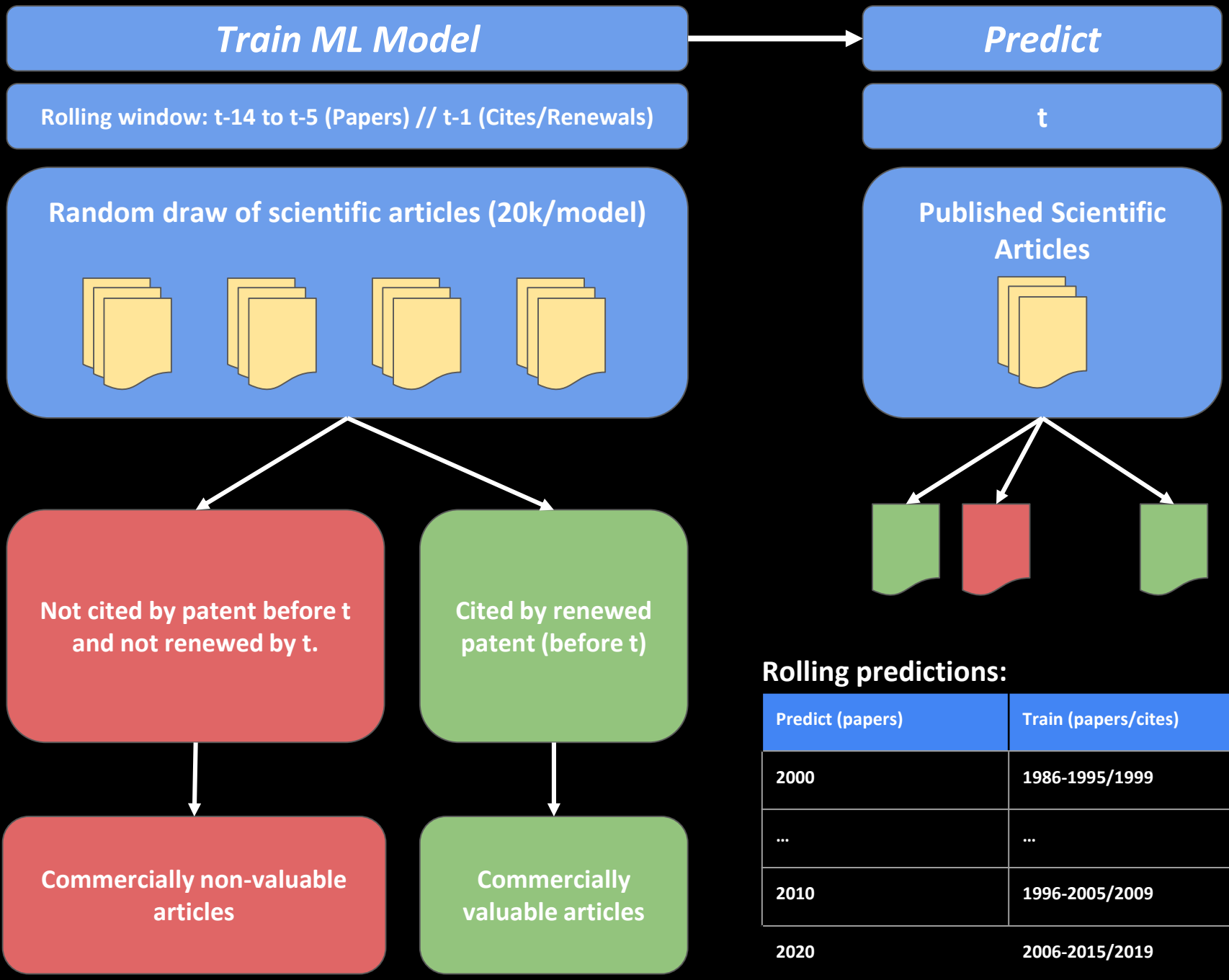
- Our measure is intended to reflect:
  - The probability that a firm believes that a scientific article can contribute to the development of a marketable product or process
- Measure of commercialization (i.e., the realization of commercialization potential)
  - The citing of a scientific article in a patent that is subsequently renewed.
- How do we measure “commercial potential”
  - An estimated probability that an article will be cited in a renewed patent.

How do we construct the measure?

- We use large language models, derived by fine-tuning SciBERT, for the purpose of classification
  - SciBert: A machine learning, embedding model used for natural language processing tailored to the scientific literature.
- Exercise in pattern recognition
  - Our models are “trained” by comparing the texts of the abstracts of scientific articles cited in renewed patents (i.e., “commercialized”) to those not cited.
- These comparisons create models that estimate an *ex ante* (i.e., forward-looking), scalable measure of commercial potential (i.e., a probability) for any scientific article.
- **This enables us to make predictions about the commercializability of any article, including the most recent.**

Data and training

- To train the models, we randomly draw 420,000 articles from a sample of 3.54 million, and use data on 522,000 patents, spanning the sample period, 1986 through 2015.
  - Article data (Dimensions): Title, abstracts, journal, author information, field.
  - Patent data (USPTO): Application date, renewal status, paper citations
- Trained 21 models, to make predictions for papers published in each year, 2000 through 2020
- For training, for each of the 21 years, we used 20,000 articles randomly drawn from our sample.
  - Training on more than 20,000 articles did not improve accuracy.
  - Only 4.93% of articles are cited in a renewed patent.
- We trained each model on papers, over a ten year period, from t-14 to t-5 to predict commercialization in each focal year, t, 2000-2020.



Rolling predictions:

Predict (papers)	Train (papers/cites)
2000	1986-1995/1999
…	…
2010	1996-2005/2009
2020	2006-2015/2019

Model accuracy (Holdout sample): AUROC ~ .74 // Accuracy ~ .74

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			

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			

External validation I

Does our model predict progression through the commercialization process of a major research university's TTO?

- Data: Detailed proprietary data on invention disclosures and outcomes from a TTO at a leading research university
  a. Disclosures; investment; patenting, agreements,licensing activity; revenue, startup
- We matched invention disclosures to their underlying scientific articles
- Resulting set:
  a. Of 96k pubs, 13,445 publications from 2,717 researchers were matched to 2,728 inventions (median publications per invention: 2)

paper	pub_year	compot	invention	disclosure_year	…
High-Strength Hydrogel Attachment through Nanofibrous Reinforcement	2020	.9105211	inv_1	2020	…
Bromide Causes Facet-Selective Atomic Addition in Gold Nanorod Syntheses	2020	.8657654	inv_1	2020	…

Progression through one university Technology Transfer Office's (TTO) commercialization process

- We have a measure of commercial potential
  - **ϕit** = probability that a scientific article will be cited by a renewed patent [0,1]
- *Probability of citation by a renewed patent* is a **scalable proxy** but not the actual thing we want to predict.
- **So, how well does the measure predict actual TTO commercial milestones, for 2000-2020 papers?**
  - Invention disclosure
  - Investment by TTO
  - Patenting by the TTO
  - Agreements with firms
  - Licensing
  - Revenue

Two notes:

- **Φit -> Predicted with only data before t**
- **Not trained on any of these outcomes: only patent citations & renewals.**

Invention disclosures to the TTO? 5x increase in disclosures moving from 1st to 4th quartile of CP measure

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%

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

Variance explained more than doubles with addition of the CP measure.

Progression across milestones: Disclosure, TTO investment, patents, agreements, licenses, revenue

(Note: paper level analysis)

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

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

CP measure predicts all outcomes, until we condition on…

The TTO investment decision: Comparing cols. 1-4 versus cols. 5-8 show commercial potential is captured by the investment decision

	(1) Invested	(2) Invested	(3) Patented	(4) Patented	(5) Agreement	(6) Licensed	(7) Startup	(8) VC Investment	(9) 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

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

Note: This is now an invention-level analysis

External Validation II

Commercial potential and its realization at 126 major U.S. research Institutions

- Data: >5 million articles, published 2000-2020, from 126 commercially active (per AUTM survey) R1 U.S. research universities spanning eleven academic fields.
- Of these, ~ 386,000 cited in a renewed patent
- **How well does our measure explain which of these articles are commercialized (i.e., cited in a renewed patent)?**

Academic science commercialization

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

Cols. 1=>2: Variance explained increases by ~40%, over and above the 126 university dummies and 210 field-year dummies.

Cols. 3=>4: Variance explained increases by ~20%, despite addition of numerous proxies for commercial impact.

From validation to application=> Illustrative applications of the measure

1. Privatization of academic science and the diffusion of academic science across firms

2. How does a university's reputation for commercializable science impact commercialization?

Application #1: Privatization

Does the privatization of academic science dampen the diffusion of knowledge across firms?

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

- Col. (8) shows that the number of firms citing patented academic science is 38% greater than the number citing comparably commercializable unpatented science.
- Why?—For future research.

Application #2: Reputation

Reputation and the commercialization of a school's research

- We explore the impact of universities' reputations for commercializing their science as a determinant of differences in commercialization rates across universities.
- Challenge
  - If we find differences across universities associated with their reputation, are such differences due to:
    - A superior ability to simply produce commercializable science?
    - Reputation per se?
- Our measure allows us to control for the production that may account for the reputation, thus isolating the effect of reputation.

Effect of production and reputation on commercialization of a university's research

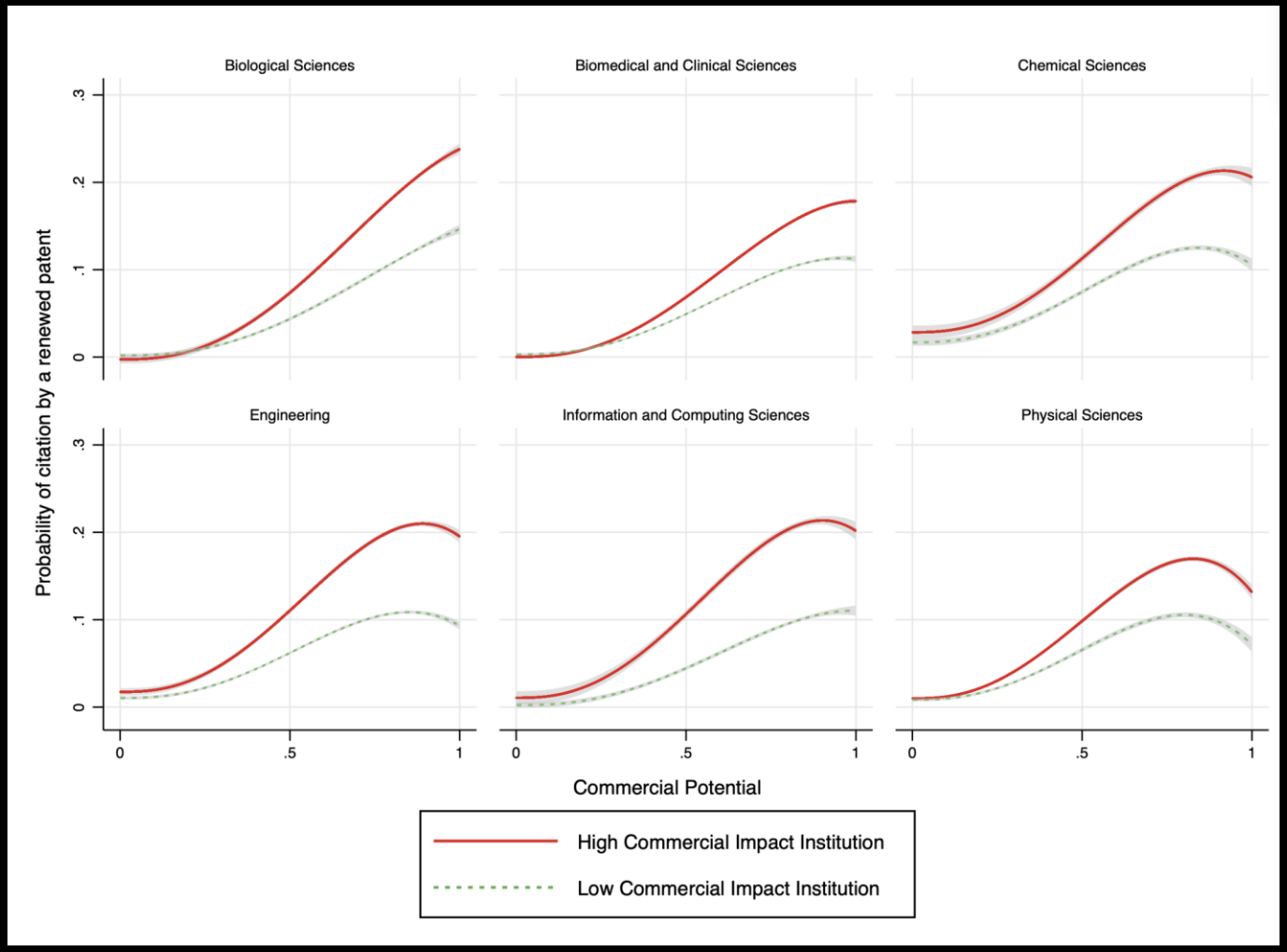
DV: Cited by renewed patent	(1)
Commercial potential	0.164***
	(0.017)
Scientific potential	0.033***
	(0.005)
High commercial prominence institution	-0.007
	(0.008)
Commercial potential x High commercial prominence institution	0.039**
	(0.018)
Publication field - year FE	Yes
Institution FE	Yes
Observations	5,211,133
R-squared	0.130

Interpretation

- Controlling for the production of commercializable research, reputation matters for commercialization.
- But reputation only matters for high commercial potential science.
  - **In other words, if MIT produces highly commercializable research, that research is more likely to be commercialized than comparably commercializable science from another, less prominent university**

- **=> Commercializable research from universities with less of a track record is more likely to be overlooked by firms, to the detriment of firms and society.**

Reputation only matters for high commercial potential science.



Conclusion

- Using machine learning and large language models, we developed and validated a measure of the commercial potential of academic science.
- Potential uses are many, including exploring questions such as:
  - What are the determinants of the production of commercializable academic science?
  - What are the frictions inhibiting the translation of academic science?
  - How large is the “realization gap”?
- Such a measure can also support practitioners’ efforts to identify science that offers commercial opportunities
- Limitations
  - Measurement error: Room for improvement
  - Reliance on patent data

Thank you