Open Science as an Economic Institution

Pierre Azoulay
MIT Sloan & NBER
NBER Innovation Boot Camp

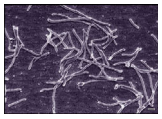

@pierre_azoulay

Outline

- ***What is science?***
  - *The relationship between science and technology*
- ***CUDOS: Science as a social institution***
  - *From Merton to the "new" economics of science*
  - *Scientific competition and its consequences*
- ***Science and its institutions***
- ***Measuring the returns to public investments in open science***
- ***[Time permitting] The direction of science***

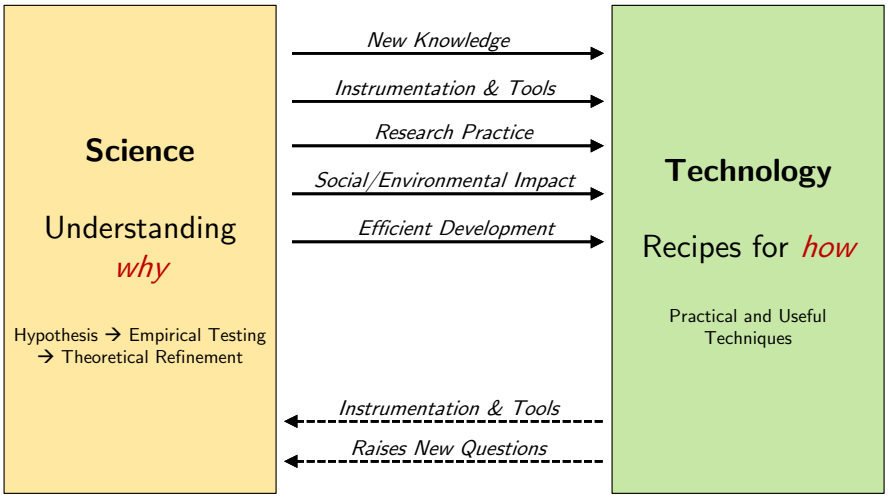
WHAT IS "SCIENCE"?

The linear model at work



- ***Brock's unlikely bacteria:***
  - *1967: Thomas Brock discovers Thermus Aquaticus in Yellowstone National Park geysers, classified as an extremophile*
  - *Deposited in the American Type Culture Collection*
  - *1983: Kary Mullis from Cetus conceives of a recipe — a DNA replication scheme requiring DNA polymerase that can resist extreme temperature variation*
  - *After initial attempts locally, identification of TaQ at ATCC*
- ***PCR is the foundational technology for DNA replication in all of modern molecular biology & biotechnology***
  - *1989: Thermus Aquaticus, Molecule of the Year; 1993 Nobel Prize for Mullis*
  - *The patent on PCR (held by Cetus) was sold on the "market for ideas," valued at approximately $500M.*
- ***The usefulness of extremophiles was very hard to anticipate ex ante***
- ***The "application" of the material with Mullis' insight was both a technological breakthrough and a spur for further scientific research***

The relationship between science and technology



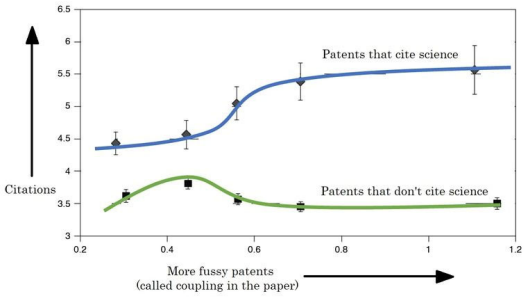
Science as a map of unfamiliar terrain

- ***More science leads to more technological progress, but only a minority of new technologies directly rely on science***
- ***Science can provide an imperfect map of this unknown terrain, helping inventors step wisely***
- ***Science can obviously benefit unexplored regions of the technological landscape***
  - *But there may also be regions that are well-trod, but treacherous*
  - *Sorenson & Fleming (2004) provide evidence that science is especially useful for this type of "fussy" technologies*

Sorenson and Fleming (2004)

- ***Relies of USPTO patent classification***
- ***Key measure based on how well a class seems to "play nice" with other technologies***
  - *If a class is frequently attached to a patent alongside a wide range of other classifications, then coupling is loose*
  - *if a class is only ever assigned to a patent with one other classification, then coupling is tight*
- ***Survey of inventors to show this measure is correlated with inventors self-assessments of how sensitive their own inventions are to small changes***
  - *Not merely picking up how novel the technology is*

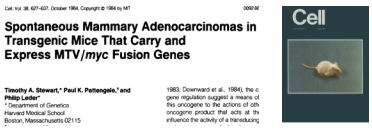
Main result



- ***Patents primarily composed of "fussy" technologies seem to disproportionately benefit from science, measured by citation of a scientific article***

When a demonstration of why is also an example of how: The Harvard OncoMouse

- ***1984: Leder & Stewart, from the Harvard Medical School, develop the "Oncomouse"***
  - *First mouse with genes inserted to predispose mouse to cancer*
  - *A significant advance along two dimensions:*
    - *Advancing basic research into the role of genes in cancer*
    - *An input into applied research focused on cancer therapies*
- ***Leder publishes a seminal article in Cell, and Harvard (and its licensee DuPont) are granted a US patent in 1988***
  - *Distribution comes with controversial licensing restrictions on use (e.g., reach-through rights and article review)*
- ***Oncomouse is a "dual" discovery and serves as foundation for:***
  - *Ongoing scientific discovery*
  - *Translation and drug development*

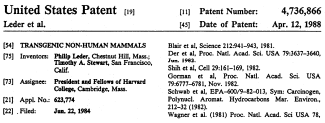


Spontaneous Mammary Adenocarcinomas in Transgenic Mice That Carry and Express MTV/*myc* Fusion Genes

Timothy A. Stewart,* Paul K. Pattengale,† and Philip Leder*

* Department of Genetics
Harvard Medical School
Boston, Massachusetts 02115

1983, Downward et al., 1984), the c
gene regulation suggest a means of
this oncogene to the actions of oth
oncogene product that acts at its
influence the activity of a transducing



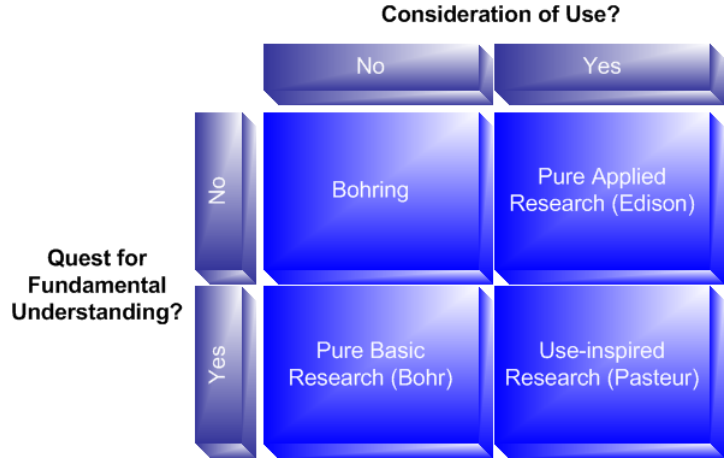
United States Patent [19]
Leder et al.

[11] Patent Number: 4,736,866
[45] Date of Patent: Apr. 12, 1988

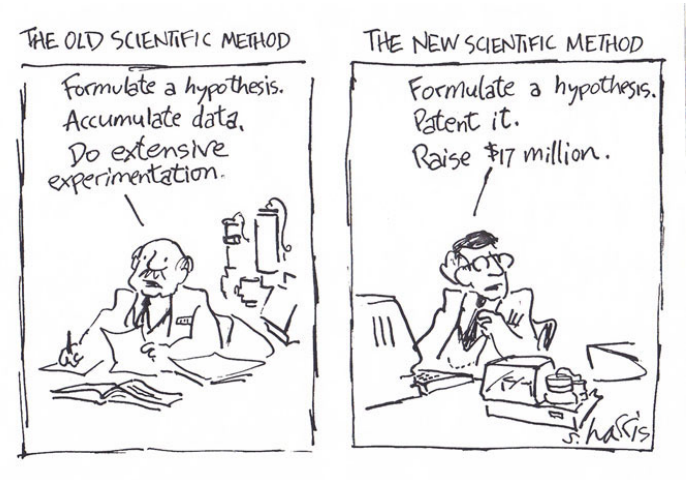
[54] TRANSGENIC NON-HUMAN MAMMALS
[75] Inventors: Philip Leder, Chestnut Hill, Mass.; Timothy A. Stewart, San Francisco, Calif.
[73] Assignee: President and Fellows of Harvard College, Cambridge, Mass.
[21] Appl. No.: 623,774
[22] Filed: Jun. 22, 1984

Blair et al, Science 212:941-943, 1981.
Der et al, Proc. Natl. Acad. Sci. USA 79:3637-3640, Jun. 1982.
Shih et al, Cell 29:161-169, 1982.
Gorman et al., Proc. Natl. Acad. Sci. USA 79:6777-6781, Nov. 1982.
Schwab et al, EPA-600/9-82-013, Sym: Carcinogen. Polynucl. Aromat. Hydrocarbons Mar. Environ., 212-32 (1982).
Wagner et al. (1981) Proc. Natl. Acad. Sci USA 78,

Pasteur's Quadrant



A more cynical view of Pasteur's quadrant



Limitations of the "linear model"

- ***Probably a good first-order description, but:***
  - *What about feedback? (Rosenberg on chemical engineering, Mokyr)*
  - *Pasteur's Quadrant: What does basic and applied mean?*
    - *Results harder to appropriate?*
    - *Results closer to ultimate commercial payoff?*
    - *Results that provides broader shoulders, for more follow-on innovators, to stand on?*
- ***How does the transmission from academia to the private sector happen?***
- ***Why do universities patent?***
  - *Because they "hold on too long" in the linear model?*
  - *Because they do research located in Pasteur's Quadrant?*
- ***Why do profit-motivated firms engage in "basic R&D"?***
  - *Market power? (Think Bell Labs and AT&T, IBM in the 1970s, Google today)*
  - *Prestige and status?*
  - *Attracting talent?*
  - *It's not "basic" in their eyes?*
  - *Absorptive capacity?*

The challenge of "Pasteur's Quadrant"

- ***Why do some types of researchers seem to be engaged in a "quest" for fundamental understanding?***
- ***What are the key economic implications of the fact that scientists seem to be governed by a distinctive set of "values" that are somewhat independent of pure monetary gain?***
- ***How do the norms and institutions of open science cohere with the nature of the incentive contracting problem between researchers (who may have preferences to participate in open science) and research funders?***

Academic freedom, private-sector focus, & the process of innovation (Aghion, Dewatripont, and Stein, 2008)

- ***Why does academia exist? Usual answer includes imperfect IPRs combined with knowledge spillovers***
  - *But recall Pasteur's quadrant: the connection between the "basicness" of a line of research and the degree of appropriability of the resulting output is ambiguous*
  - *Even if we need basic research to be subsidized (because of limited appropriability), why does this need to happen in academia?*
- ***ADS 2008 develop a model that***
  - *clarifies the respective advantages and disadvantages of academic and private-sector research*
  - *allows one to say when—in the process of developing an idea from its very earliest stages to a finished commercial product—it is normatively optimal to make the transition from academia to the private sector*
- ***At the heart of the model is a decision right:***
  - *Academia boils down to a commitment mechanism that ensures scientists can choose the projects they work on*
  - *In private-sector research, the decision rights inevitably resides with the owner/manager of the firm, who can (and will) largely dictate project choice and methods to the individual scientists who work for the firm*

Aghion, Dewatripont, and Stein 2008 (cont'd)

- ***A simple model of the impact of science/academia as a method for organizing privately funded research***
- ***Consider a k-stage research process, in which financial returns V are only realized when the firm successfully completes all stages***
- ***Model "science" or "academia" as an organizational design choice, in which the firm cedes control rights over research direction to researchers (i.e., this is a model of "freedom")***
  - *Ignore the issue of appropriability*
  - *With probability α, researcher has preferences for research direction which advances commercialization, and is successful (conditional on choosing that direction) with probability p; note that with 1–α, research gets utility z from an alternative direction and interests are misaligned*
- ***Firms can either retain control rights for themselves (enhancing the potential for commercialization) or cede control to researchers and benefit from a lower wage structure***

Basic intuition

- ***Consider a case where commercialization involves two steps***
- ***In the last stage, firm chooses to retain control rights if the gains to ensuring that the right final "step" is taken outweighs the wage benefit from ceding control to the researcher (i.e., $pV > z$)***
- ***However, in the first stage, firm only chooses to retain control rights if the gains to ensuring that all steps outweighs the wage benefit, (i.e., $pE(\Pi_1) > z$)***
- ***Key insight: "academic freedom" is most attractive at the "earliest" stages of the research process and is associated with exploration***

Exploration incentives

- ***ADS consider the possibility of research lines "branching out"***
- ***Suppose that there are two potentially legitimate research projects inside the firm:***
  - *An "applied" project that is only two stages away from a commercial payoff*
  - *A more "basic" project that is five stages away from any payoff*
- ***Which organizational form is more likely to explore?***
  - *It is possible that the ultimate payoff on the more basic project is sufficiently high that, evaluated at academic-sector wages, it is not only positive net present value (NPV), but of greater NPV than the applied project.*
  - *It is also possible that, evaluated at private-sector wages, the basic project is negative NPV.*
    - *If this is the case, then when a private-sector firm has the decision rights, it will allocate all of its scientists to the applied project, and completely ignore the basic project*
    - *By contrast, if the ideas were left freely available to academic scientists, there would naturally tend to be some progress on both projects, as individual scientists followed their own interests.*
- ***It is possible that the returns to freedom are higher when researchers are able to exercise openness, since the benefits from control are more salient when one is able to publicly reveal the information in the scientific literature***

Frictions at the academia/industry interface

1977: Discovery (purification) of EPO by Eugene Goldwasser (U of Chicago)

1977-1981
For 5 years, Goldwasser tries desperately to interest firms to produce EPO
Rejected by: University of Chicago; Parke-Davis, Abbott Labs

1984: Amgen sequences EPO gene

1987: Amgen produces recombinant EPO

- ***In ADS 2008, the hand-off from academia to the private sector might not happen at the optimal time, but it is essentially frictionless***
- ***Bikard (2018) provides evidence of under-utilization of knowledge coming out of universities***
  - *Under what circumstances is a piece of scientific knowledge translated into a new technology?*
  - *Specifically, does it matter if the discovery took place in a university vs. a private firm?*
- ***Key empirical lever: scientific twins stemming from simultaneous discoveries***

In the winter of 1999

Vanilloid receptor-1 contributes to chemical and thermal sensitivity in mice

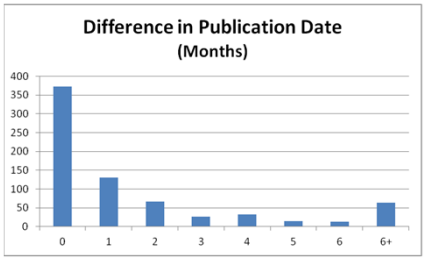
Submitted: 18 January 2000; Published: 14 April 2000
Address: UCSF in San Francisco (CA)

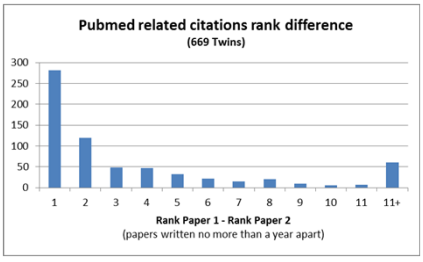
Submitted: 20 December 1999; Published: 11 May 2000
Address: Smithkline Beecham in Harlow (UK)

"Paper twins": same knowledge simultaneously emerges in two distinct environments

Bikard's (2020) Twin Identification Algorithm

- ***Most pairs were published in the exact same month***
  - *Avg. difference in months is 1.8*
- ***267 twins were published in the same issue of the same journal***
- ***PubMed Related Citations Algorithm (based on keyword similarity) rank them next to each other 42% of the time.***
  - *Rank difference <10 for 90% of the twins*
- ***"Out of 10 interviewees, 9 told me about the twin paper without me asking"***
  - *One got really upset when I mentioned the twin*





Main result:
Academic twin 20-30% less cited in private-sector patents

	Dependent variable = reference (1/0)		
Variable	Conditional logit; controls only (Model 5-1)	Conditional logit; main effect (Model 5-2)	LPM; main effect (Model 5-3)
Academic origin		−0.673***	−0.238***
		(0.25)	(0.07)
Paper is more detailed	0.623*	0.755**	0.184**
	(0.36)	(0.30)	(0.08)
Paper has richer theory	−1.073	−1.096	−0.287
	(0.87)	(0.72)	(0.18)
Paper is more sophisticated	−1.177	−0.99	−0.249
	(0.92)	(0.80)	(0.19)
Paper has more practical emphasis	0.593**	0.467	0.0843
	(0.29)	(0.32)	(0.09)
Paper is clearer	14.55***	14.58***	0.696***
	(1.41)	(1.35)	(0.19)
U.S. paper	1.544**	1.701***	0.408***
	(0.67)	(0.66)	(0.15)
Journal impact factor	0.0148	0.00264	0.00118
	(0.03)	(0.03)	(0.01)
Patent-paper pair	0.204	−0.0079	−0.0478
	(0.51)	(0.45)	(0.10)
Number of authors	0.0745	0.0773	−0.0588
	(0.49)	(0.39)	(0.11)
Authors' publication stock	0.221	0.349	0.124
	(0.34)	(0.32)	(0.08)
Authors' patent stock	−0.0401	−0.117	−0.0367
	(0.12)	(0.17)	(0.04)
Time lag	0.313	0.506	0.0641
	(0.43)	(0.41)	(0.09)
Geographic distance	−0.115	−0.0949	−0.029
	(0.10)	(0.10)	(0.03)
Same country	−0.426	−0.402	−0.124
	(0.53)	(0.55)	(0.14)
Constant			0.145
			(0.36)
Observations	523	523	924
No. simultaneous discovery/ patent dyads	225	225	480
Pseudo-*R*2	0.119	0.153	0.149
Log-likelihood	−163.3	−157.1	−310.1
Simultaneous discovery/ patent fixed effects	Yes	Yes	Yes

Notes. Observations are paper-patent dyads. Column 3 contains more observations as LPM does not drop cases in which the patent cites all the paper twins. Fixed effects are at the set-of-paper-twins/patent dyad level. Standard errors are clustered throughout at the level of the set of paper twins.
*$p < 0.1$; **$p < 0.0.05$; ***$p < 0.01$.

Evidence on the benefits of openness: Of Mice and Academics (Murray et al. 2016)

- ***What role does scientific openness play in scientific research?***
- ***What types of research are promoted by openness?***
- ***Control rights approach suggests two effects of openness:***
  - *Vertical exploitation – downstream exploitation increases*
  - *Horizontal exploration – entirely new, diverse lines of basic research increase*
- ***The paper exploits the natural experiment created by the shift in openness from NIH agreements and traces out the impact on citations to articles impacted by the agreement***

The mouse revolution as a research setting

- ***Over the past twenty years, a "revolution" in the use of genetically engineered research mice as a tool for life sciences progress***
  - *Mice could now be "engineered" to have a particular gene inserted or removed to mimic a disease e.g., cancer or diabetes*
  - *Over 13,000 specialized mice published in scientific literature*
- ***2007 Nobel Prize in Medicine to Mario R. Capecchi, Martin J. Evans and Oliver Smithies for "gene modification in mice"***
- ***Openness: While the development of genetically modified mice has tremendous for potential application in both basic and applied research, the ability to initiate research "lines" based on new mice require gaining access to those specific mice***
  - *Mice are costly to make and require specialized techniques including embryo manipulation, stem cell adaptation, and molecular biology*
  - *Many mice are also covered by intellectual property rights and so require a license contract with upstream researchers*

Natural experiment in openness

- ***1990s: Openness crisis***
  - *scientists demand openness to DuPont's OncoMice*
- ***1999: Harold Varmus at NIH intervenes and signs MoU with DuPont to make OncoMice subject to a "simple" license with no reach-through***
  - *An unexpected shift in the openness of mouse genetics research*

Data sources

- ***Data Sources***
  - *Mouse Genome Informatics database catalogs over 13,000 mice & links each mouse to an original publication in a scientific journal (mouse-articles)*
  - *PubMed for information about mouse-articles & ISI Web of Science SCI for citations*
- ***Sampling Strategy***
  - *Identify universe of MGI mouse-articles published 1983-1998 sample on four types of mouse-articles (2,638 unique mice in 2,223 mouse-articles)*
  - *Cre-Lox (52), Oncomouse (160), Knock-Out (2171), Spontaneous (255)*
- ***For each mouse-article collect information about the forward citations***
  - *525,865 total citations (from pub year thru 2006)*
  - *Aggregated up into 27,442 citation-years*
- ***For each citing article code key article/author characteristics***

Results: Vertical Exploitation

Dep. var. = Annual citations
[Incidence rate ratios reported in square brackets]
Estimated coefficients in second line
(Block bootstrapped SEs reported in parentheses)

	OLS	Negative binomial			
	(4-1) Baseline model, DV = log Annual citations	(4-2) Baseline model	(4-3) Baseline model with treatment effect dynamics	(4-4) Treatment effects by Cre-lox and Onco	(4-5) Baseline model, citations from high quality journals only[d]
Post-NIH	[1.229]*** 0.206 (0.052)	[1.302]*** 0.264 (0.062)			[1.409]*** 0.343 (0.080)
Post-NIH, Short-term[b]			[1.220]*** 0.199 (0.064)		
Post-NIH, Long-term[c]			[1.429]*** 0.357 (0.074)		
Post-Cre-lox				[1.467]*** 0.383 (0.115)	
Post-Onco				[1.267]*** 0.236 (0.060)	
Age FEs	Yes	Yes	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes	Yes	Yes
Article FEs	Yes	Yes	Yes	Yes	Yes
log-likelihood	—	−55,919.8	−55,906.1	−55,912.4	−34,112.8
Observations	22,265	22,265	22,265	22,265	21,574

Results: Horizontal Exploration

Negative Binomial	**Keywords**		**Journals**	
	Annual Citations with New keywords	Annual Citations with Old keywords	Annual Citations in New Journals	Annual Citations in Old Journals
Post Shock	1.260^{***}	0.925	1.381^{***}	1.201^{*}
Conditional Fixed Effects for Article, Margin-Age and Margin-Calendar Year, Window Effects				

Key Findings

- ***A significant increase in the rate of follow-on citations for "mouse-articles" impacted by the NIH agreements***
- ***This boost in follow-on research is driven by***
  - *Contributions by "new" authors or institutions (reprint authors or institutions that had not previously cited the original mouse-article)*
  - *More diverse types of research (articles using previously unused keywords or published in journals that had not previously cited the original mouse-article)*
  - *No detectable reduction in the flow of new mouse creation.*
- ***Results highlight a neglected impact of IP: reductions in the diversity of experimentation arising from a single idea***

Do scientists pay to be scientists? [Stern 2004]

- ***A Preference Effect (a "taste" for science)***
  - *Researchers (even those in the private sector) may value participation in open science, and thus firms may earn a compensating differential by allowing participation in science in exchange for lower wages*
  - *Intrinsic preferences (Feynman, Kuhn)*
  - *Career concerns (cf. Lerner and Tirole on Open Source)*
- ***A Productivity Effect (a "ticket of admission")***
  - *Firms may benefit from access to scientific knowledge; understanding scientific discoveries (and perhaps learning about them earlier) can only be realized by firms who themselves "spill" some knowledge through participation in open science*
  - *Direct spillovers*
  - *Indirect spillovers*

Evaluating the wage-science relationship

- ***Cross-sectional relationship between wages and science likely will reflect unobserved differences in ability***
  - *Long tradition in labor economics associated with not being able to control for unobserved heterogeneity (Rosen, 1986)*
  - *Prior work has examined job switchers (Brown, 1980) which are unfortunately subject to their own biases (Gibbons & Katz, 1992)*
- ***Prior to accepting any offer, researchers (and many professionals) receive multiple job offers***
  - *Suggests methodology for "controlling" for individual effects*
  - *Regress wage on organizational practices at the job offer level j, with a fixed effect for each individual worker i*

$$\ln(w_{ij}) = \gamma_i + \beta \cdot SCIENCE_j + \varepsilon_{ij}$$

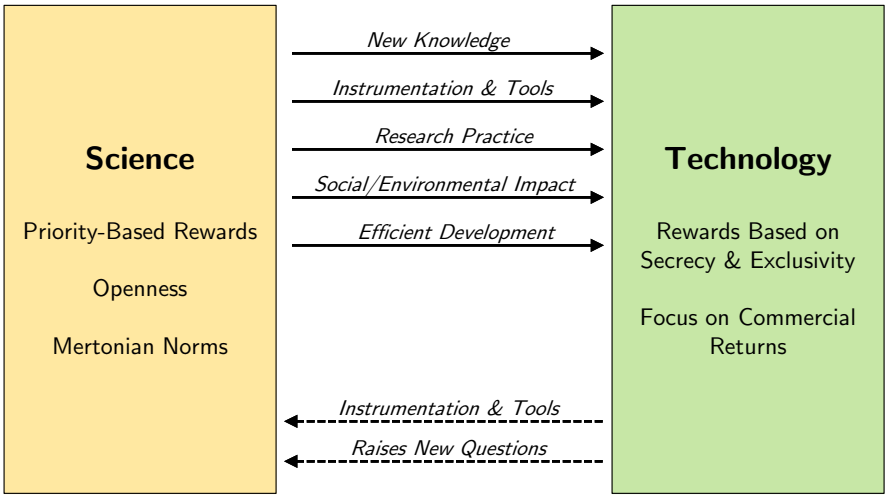
Hedonic wage regression (i=52, j=121)

	Permission to publish			Combination model
	(3–1)	(3–2)	(3–3)	(3–4)
	Baseline (NO FE)	Baseline (w/FE)	Full model (w/FE)	Full model (w/FE)
PERMIT_PUB	0.027	**−0.266**	**−0.191**	−0.089
	(0.186)	(**0.114**)	(**0.105**)	(0.103)
CONTINUE RESEARCH				**−0.134**
				(**0.060**)
INCENT_PUB				−0.036
				(0.028)
SCIENCE INDEX				
EQUIPMENT				**0.063**
				(**0.033**)
CONTROLS				
PROMOTION			**0.041**	**0.046**
			(**0.025**)	(**0.021**)
STOCK_DUMMY			**0.196**	**0.234**
			(**0.085**)	(**0.074**)
ACCEPTED JOB			−0.013	0.002
			(0.040)	(0.043)
JOBTYPE CONTROLS	no	no	yes (5; Sig.)	no
Individual fixed effects	no	yes (52; Sig.)	yes (52; Sig.)	yes (52; Sig.)
R-squared	0.001	0.915	0.955	0.958

Economic and strategic implications

- ***Relative to a system of proprietary knowledge production, the incentives and norms of open science seem to be consistent with the objective of maximizing the rate of production of knowledge in a cumulative manner***
- ***However, the nature of the scientific priority system likely results in distortionary strategic behavior***
  - *Inefficient "herding" on hot topics or big discoveries*
  - *Complicated and costly disputes over scientific priority itself*
  - *Potential for collusion*
  - *Inefficient strategic exclusivity over data, tools, or other resources*
- ***Open science also induces a high potential for spillovers from public knowledge to applications governed by technology***

Science as a distinctive incentive system (Dasgupta and David 1994)



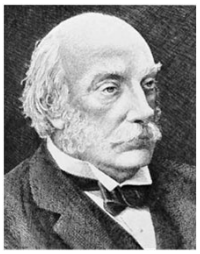
SCIENCE AS A SOCIAL INSTITUTION

Merton's CUDOS: The Normative Structure of Science [1942]

- ***Communalism** – the common ownership of scientific discoveries, according to which scientists give up intellectual property rights in exchange for recognition and esteem*
- ***Universalism** – according to which claims to truth are evaluated in terms of universal or impersonal criteria, and not on the basis of race, class, gender, religion, or nationality*
- ***Disinterestedness** – according to which scientists are rewarded for acting in ways that outwardly appear to be selfless*
- ***Originality** – the ultimate scientific reward is the "thin" intellectual property right of credit for having made a particular discovery*
- ***Skepticism** – all ideas must be tested and are subject to rigorous, structured community scrutiny*

Violation of the universalism norm: The Matthew Effect (Merton 1965)

- ***Seemingly high importance of early luck and resources in shaping the skewed distribution of research productivity and scientific status***
  - *"if I have not seen as far as others, it is because giants were standing upon my shoulders" – Hal Abelson*

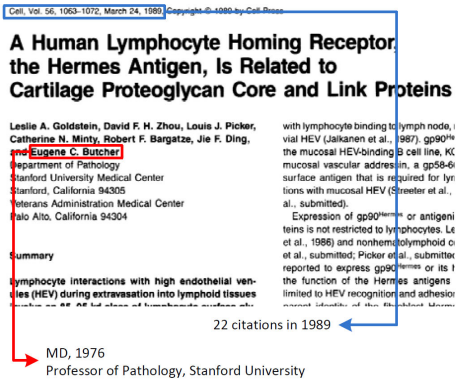


"Rayleigh's name was either omitted or accidentally detached* [from a manuscript] *and the Committee turned it down as the work of one of those curious persons called paradoxers. However, when the authorship was discovered, the paper was found to have merits after all."

Matthew: Effect or Fable? Azoulay, Stuart, & Wang 2014

- ***Distinguish between producers (scientists) and products (articles)***
- ***Focus on the impact of a discrete change in producer status, i.e., a "status shock:" HHMI Appointment***
- ***Restrict the set of products to those that first appeared before the shock***
- ***Measure the status premium (or discount) by examining changes in deference patterns after the shock, relative to before***

Treated and control articles



Cell, Vol. 56, 1063–1072, March 24, 1989, Copyright © 1989 by Cell Press

A Human Lymphocyte Homing Receptor: the Hermes Antigen, Is Related to Cartilage Proteoglycan Core and Link Proteins

Leslie A. Goldstein, David F. H. Zhou, Louis J. Picker, Catherine N. Minty, Robert F. Bargatze, Jie F. Ding, and Eugene C. Butcher

Department of Pathology
Stanford University Medical Center
Stanford, California 94305
Veterans Administration Medical Center
Palo Alto, California 94304

Summary

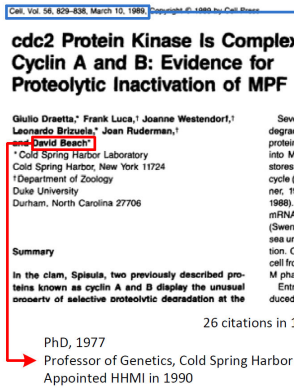
Lymphocyte interactions with high endothelial venules (HEV) during extravasation into lymphoid tissues

with lymphocyte binding to lymph node, vial HEV (Jalkanen et al., 1987). gp90 the mucosal HEV-binding B cell line, KO mucosal vascular addressin, a gp58-6 surface antigen that is required for lym tions with mucosal HEV (Streeter et al., al., submitted).

Expression of gp90 or antigeni teins is not restricted to lymphocytes. Le et al., 1986) and nonhematolymphoid c et al., submitted; Picker et al., submitte reported to express gp90 or its the function of the Hermes antigens limited to HEV recognition and adhesio

22 citations in 1989

MD, 1976
Professor of Pathology, Stanford University



Cell, Vol. 56, 829–838, March 10, 1989, Copyright © 1989 by Cell Press

cdc2 Protein Kinase Is Complex with Cyclin A and B: Evidence for Proteolytic Inactivation of MPF

Giulio Draetta,* Frank Luca,† Joanne Westendorf,† Leonardo Brizuela,* Joan Ruderman,† and David Beach*

* Cold Spring Harbor Laboratory
Cold Spring Harbor, New York 11724
† Department of Zoology
Duke University
Durham, North Carolina 27706

Summary

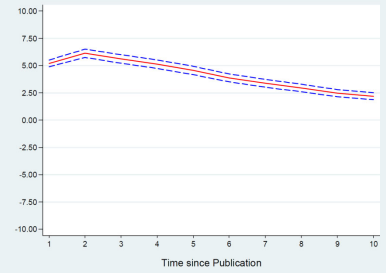
In the clam, Spisula, two previously described proteins known as cyclin A and B display the unusual property of selective proteolytic degradation at the

26 citations in 1

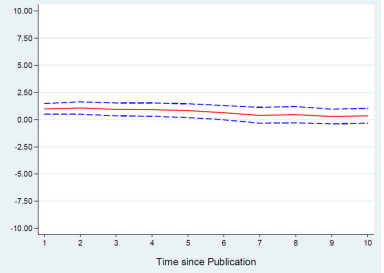
PhD, 1977
Professor of Genetics, Cold Spring Harbor
Appointed HHMI in 1990

Effects of HHMI appointment on citation rates [Post-Appointment Articles]

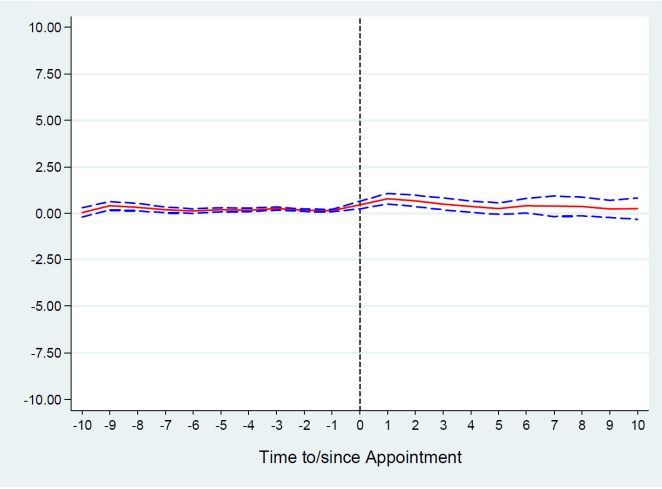
w/o journal match



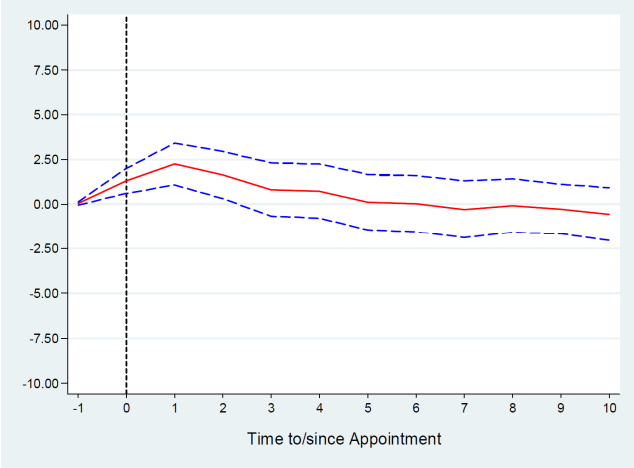
w/ journal match



Effects of HHMI appointment on citation rates [Pre-Appointment Articles]



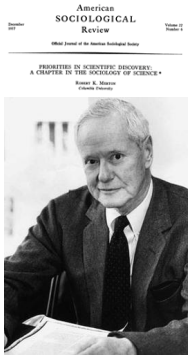
Effects of HHMI appointment on citation rates [Pre-Appointment Articles]



SCIENTIFIC COMPETITION

How are scientists rewarded for novel discoveries?

"In short, property rights in science become whittled down to just this one: the recognition by others of the scientist's distinctive part in having brought the result into being."
- Robert K. Merton (1957)

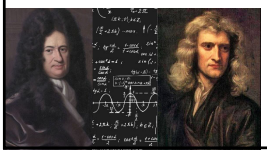
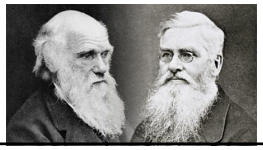
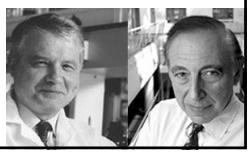


The nature of scientific priority

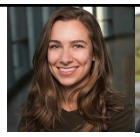
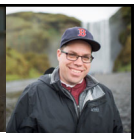
- ***An odd type of property right***
  - *Not a direct monetary reward*
  - *Not a control right*
  - *Simply a "thin" intellectual property right – "the recognition by others of the scientist's distinctive part into having the result brought into being."*
- ***Adjudicated itself by the scientific community***
  - *Eponymy rather than anonymity*
- ***The potential for mischief***
  - *Fraud*
  - *Plagiarism*

Priorities in Scientific Discovery [1957]

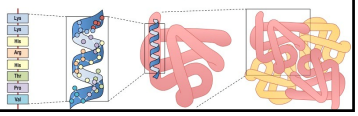
- ***The history of science is replete with intense (and intensely complicated) disputes over scientific priority – who was the particular person to make a particular discovery***
  - *This is not simply a matter of egotism – many disputes are fought by supposedly independent parties, and, in many cases, the subjects of the dispute stay "above the fray"*
  - *Indeed, in some (but not all) cases, researchers undertake steps to share credit or recognize others contributions (e.g., Darwin and Wallace)*
- ***The norms and behaviors to accord scientific priority reflects the fundamental interest in providing a reward for originality***
  - *But balanced against the competing norm of humility*

Priority races: Empirical evidence

- ***In a pair of papers, Ryan Hill and Carolyn Stein provide the first systematic empirical look at priority races***
- ***"Scooped! Estimating Rewards for Priority in Science"***
  - *What is the causal effect of losing a priority race on project and scientist outcomes?*
- ***"Race to the Bottom: Competition and Quality in Science"***
  - *The dark side of competition: scientists may cut corners and reduce quality in their pursuit to publish first*
- ***Both papers use leverage the same data and setting: structural biology and the Protein Data Bank (PDB)***



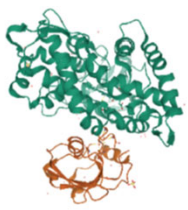
Scooped!

- ***What is the causal effect of getting scooped?***
  - *Short-run effect on project: Publication, journal placement, and citations*
  - *Long-run effect on career: Future productivity of scientists*
- ***Does the priority reward system reinforce inequality in science?***
  - *Is the scoop effect equal for high- and low-reputation teams?*
- ***Key empirical challenges***
  - *Need a setting with well-defined problems and "one right answer"*
  - *Need an objective measure of scientific proximity*
  - *Need a view of potential abandonments prior to publication*
- ***The paper analyzes more than 1,500 priority races in structural biology using the Protein Data Bank (PDB)***

Research Design

$$Y_{ip} = \alpha + \beta Scooped_i + X_i'\delta + \gamma_p + \epsilon_{ip}$$

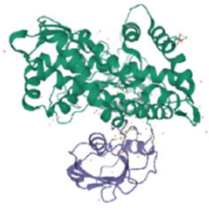
Winning Deposit: 4JWS



Affiliation: UC Irvine
Deposit Date: March 27, 2013
Release Date: June 19, 2013

Journal: *Science*
Journal Impact Factor: 31.5
5-year Citations: 52

Scooped Deposit: 3W9C



Affiliation: Leiden University
Deposit Date: April 3, 2013
Release Date: August 21, 2013

Journal: *Journal of Molecular Biology*
Journal Impact Factor: 4
5-year Citations: 39

Results

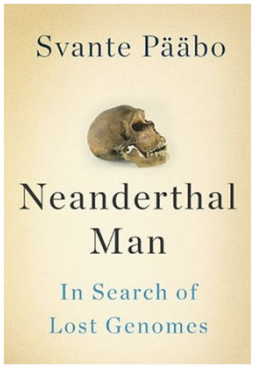
- ***Priority paper gets 54% of total citations and scooped paper gets 46%***
  - *Surveyed scientists are much more pessimistic: 74% to 26%.*
  - *Scooped projects are less likely to be published, and less likely to appear in a top-10 journal*
- ***In the next five years, scooped scientists have the same number of publications, but fewer citations***
- ***Priority system reinforces inequality:***
  - *Citation penalty is larger for low-ranked teams than it is for high-ranked teams.*

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	asinh(Five-year citations) (4)	Top-10% five year citations (5)
Scooped	-0.025***	-0.178***	-0.060***	-0.197***	-0.035***
	(0.010)	(0.032)	(0.014)	(0.045)	(0.010)
Winner Y mean	0.880	-0.031	0.318	28.918	0.150
Observations	3,319	3,319	3,319	2,546	2,546

Note: All regressions include controls selected by PDS-Lasso as well as year fixed effects. Unpublished papers have impact factor imputed to minimum factor journal. Citation regressions restricted to papers published before 2014. Column 4 dependent variable is asinh(five-year citations) but mean citations is reported in levels.

Race to the bottom

"Hendrik's paper also illustrated a dilemma in science: doing all the analyses and experiments necessary to tell the complete story leaves you vulnerable to being beaten to the press...Even when you publish a better paper, you are seen as mopping up the details after someone who made the real breakthrough"



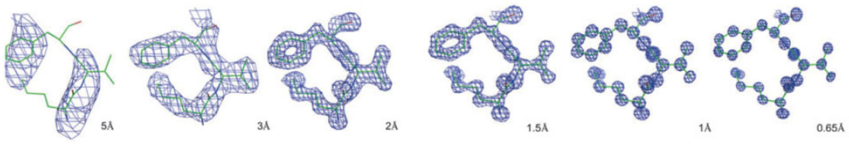
– Svante Pääbo, Neanderthal Man: In Search of Lost Genomes

Race to the Bottom (cont'd)

- ***Hill & Stein build a model with the following predictions:***
  - *Most (ex-ante) important projects are more competitive, rushed, and "lower quality" (in the sense of being executed in a sloppier fashion)*
- ***They find that***
  - *High-potential projects are more competitive (multiple researchers working simultaneously)*
  - *High-potential projects are completed faster and are lower quality*
  - *Follow-on work ameliorates but does not eliminate the negative relationship between potential and quality*
  - *Quality magnitudes large enough to impact the usefulness of projects for drug development*

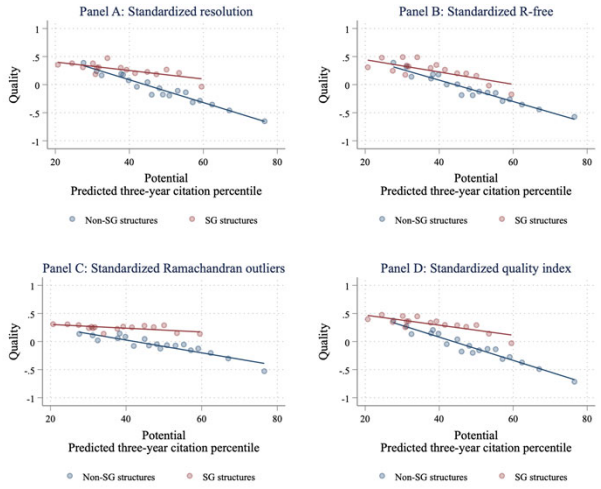
Measurement challenge

- ***A unique feature of structural biology is the objective, ex-ante measures of project quality:***
  1. *Refinement resolution: similar to resolution of a photograph*
  2. *R-free: model fit, estimated on a holdout sample of the experimental data*
  3. *Outliers: errors in the model based on chemical properties*



- ***They combine these outcomes into a standardized quality index (higher is better)***

Results in one picture



SCIENCE AND ITS INSTITUTIONS

Institutions and the rate and direction of scientific advance

- ***The "ideas production function":***

$$\dot{A} = f[A(z), H(z), K(z), z]$$

- ***Broad view of what counts as an institution***
  - *Editorial policies*
  - *Replicability rules*
  - *Funding rules and systems*
  - *Access to capital equipment and materials…*

- ***What is the impact of specific institutions on science?***

Furman and Stern (AER 2011)

- ***Question:*** ***How do institutional forms influence the disclosure of knowledge, with implications for cumulativeness and the capacity to harness potential spillovers?***

- ***Identification Strategy:*** ***Differences-in-differences. Take a fixed piece of knowledge (e.g., a paper). Examine changes in citation behavior before and after some "exogenous" event (treatment). Compare to control group of similar piece of knowledge (e.g. paper with similar ex-ante citations) that does not experience treatment.***

- ***Setting:*** ***Biological resource centers (BRCs). Deposit of organisms en masse in BRC (exogenous event) allow other researchers to utilizes these organisms.***

Research design

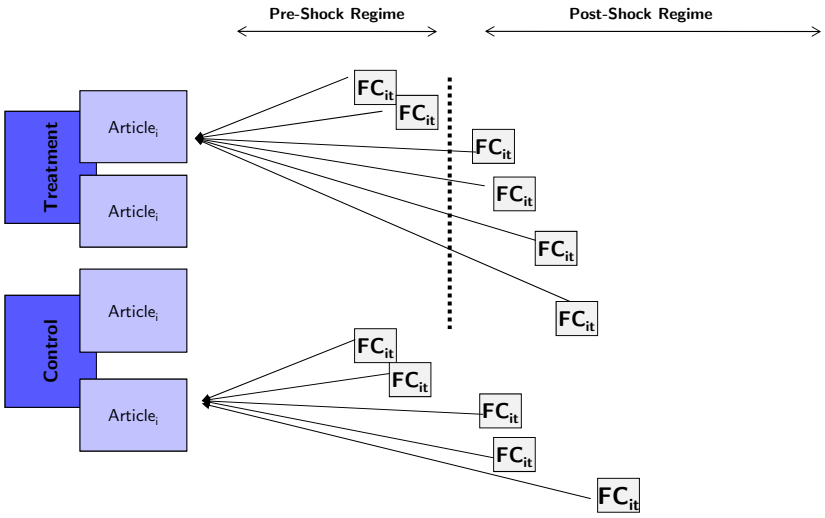


Figure 2: Citation Effect of BRC Deposit

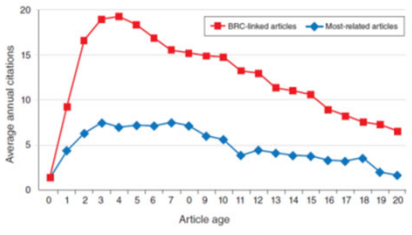


FIGURE 1. AVERAGE ANNUAL CITATIONS BY AGE, BRC VERSUS CONTROL ARTICLES

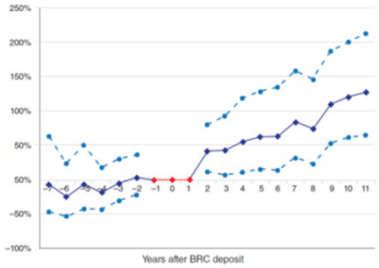


FIGURE 2. PRE- AND POSTDEPOSIT EFFECTS ON *FORWARD CITATIONS*

Summary

- ***Is science more than a type of knowledge? YES***
- ***Over the past decade, a mixture of theoretical and empirical research suggests that open science is a distinctive economic institution***
  - *An incentive system that overcomes the "paradox" of directly paying for ideas that seems so central to endogenous technical progress (see Romer)*
- ***Increasing amount of scholarly effort devoted to examining the impact of specific institutions and potential for strategic behavior undermining this objective***

RETURNS TO INVESTMENTS IN OPEN SCIENCE

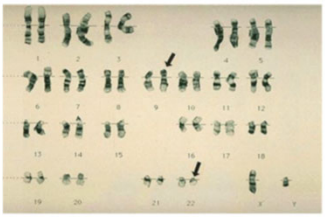
What is the causal effect of public R&D investments on patenting by life science firms? Azoulay, Graff Zivin, Li, & Sampat 2019

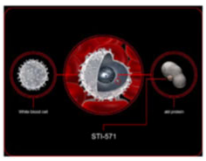
1. ***Scientific knowledge is non-rival and meant to be foundational – making it hard to trace where public investments have an impact.***
   - *Identify patents that build on publicly-funded research by examining citations of NIH-funded work by corporate patents.*

2. ***Funding can be endogenous***
   - *Use variation in NIH funding that come from scoring cutoffs to identify the causal effect of funding*

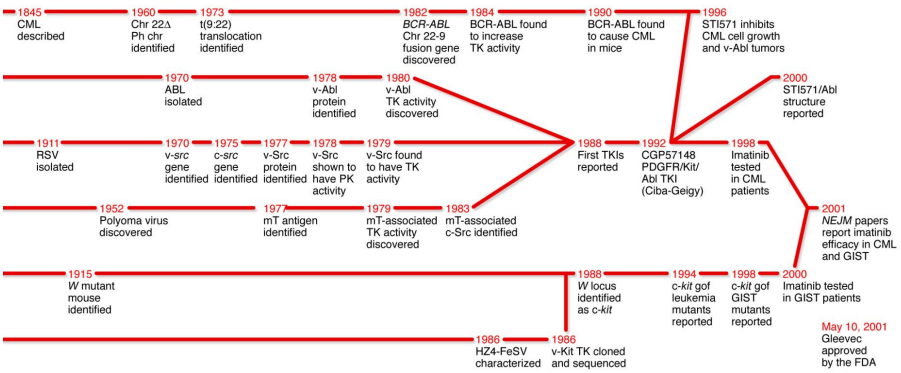
3. ***Public investments can crowd out private investments***
   - *Examine impact of NIH funding on the total number of patents (building on NIH-funded research or not) in a given intellectual area*

Why might public funding impact private sector patenting? The Gleevec/imatinib story





The long road to imatinib



Measuring outcomes of NIH funding: Why patents?

- ***Past research has focused on welfare-relevant outcomes: new drug approvals; changes in morbidity/mortality***
  - *Requires that we partition funding across diseases when much funding in untargeted or broadly targeted*
  - *Requires that we take a stand a priori on the lags between funding and outcomes*
  - *Does not lend itself to convincing identification*

- ***Patents are extremely heterogeneous and far upstream from health outcomes; Public funding could improve health outcomes even without patents (surgical techniques, epidemiology, etc.) However:***
  - *The vast majority of innovations in the life sciences are patented*
  - *Patent heterogeneity is empirically (and imperfectly) tractable*
  - *Bibliometric information allows us to make many fewer assumptions when linking funding to outcomes*

What is the right unit of analysis?

- ***Impacts the question we ask***
  - *All of biomedical innovation as a unit: What is the impact of an increase in biomedical R&D spending on innovation?*
  - *Disease as unit: What is the impact of an increase in cancer spending on innovation related to cancer?*
  - *Individual grant as unit: What is the impact of an increase in funding for a grant on innovation tied to that grant?*
- ***Impacts the policy-relevance of our findings***
  - *The broader question has more appeal for policy-makers…*
  - *…But is not likely to be answered convincingly*
- ***Impacts the identifying variation we need:***
  - *Random funding for a disease area: difficult because we pay attention to large funding choices like this*
  - *Random funding for a grant: difficult because we pay attention to small specific choices*

Using "research areas" as the unit of analysis

- ***No scientist does research "on cancer"***
  - *Biomedical research involves a science area and a disease application*
  - *e.g., cell signaling in cancer; gene expression in diabetes*
- ***We define a research area as a disease-science area for a given year***
  - *This consists of research that share a similar disease interest and that benefit from an understanding of the same science*
  - *Call it a DST, for* $\text{disease} \times \text{science} \times \text{time}$
- ***Advantages***
  - *One can still ask a policy-relevant question: what happens if we provide more funding for a disease-science area? (e.g. genetic basis of Alzheimer's)*
  - *No one ever decides how much funding to give a DST (that will help with identification)*

Identifying research areas

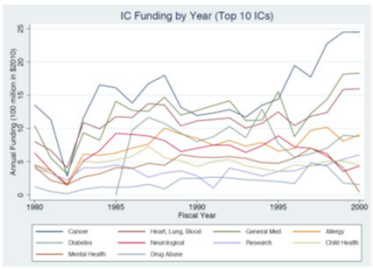
- ***Identifying disease areas:***
  - *NIH consists of 27 disease/medicine-focused Institutes/Centers*
    - *National Cancer Institute*
    - *National Institute of Diabetes and Digestive and Kidney Diseases*
  - *A grant application must report its disease area to be funded*

- ***Identifying science areas:***
  - *Grant review happens in 180 science focused "study sections"*
    - *Cellular Signaling and Regulatory Systems*
    - *Integrative Nutrition and Metabolic Processes*
  - *A grant's application needs to specify its science area to be evaluated*

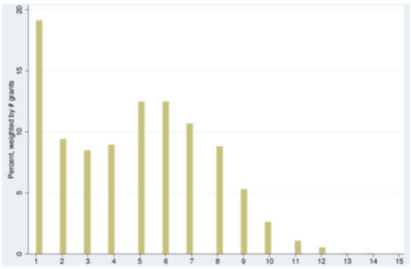
NIH National Institutes of Health
Turning Discovery Into Health

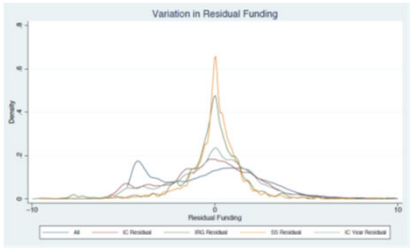
Funding	Evaluation
	NIH Center for Scientific Review 

DST-relevant variation



Number of institutes represented in each study section





Why is estimation difficult? 2 key challenges

Generic Regression

$$Patents_{???} = \alpha_0 + \alpha_1 Funding_{dst} + Controls_{dst} + \varepsilon_{dst}$$

- ***Where do we look for outcomes?***
  - *It is hard to know a priori what scientific results are relevant for a patent*
  - *We use new data to explicitly link grants with patents via acknowledgements, citations, and intellectual relatedness*
- ***Endogeneity***
  - *Funding potentially responds to changes in innovative/commercial potential across disease or science areas*
  - *We use the structure of NIH grant review to control for many unobservables and to generate plausibly exogenous variation in funding*

Looking for patent outcomes of NIH funding

1. ***Direct acknowledgment: # patents by NIH-funded researchers***
   - *Link Grant → Patent*
   - *Does the NIH directly fund patentable research?*

2. ***Citation-linked: # patents citing NIH-funded research***
   - *Link Grant → Publication → Patent*
   - *Measure of innovation that concretely draws on research funded by the NIH*
   - *Does the NIH fund research that is useful to private firms seeking patents?*

3. ***Same area: # patents intellectually related to an NIH funding area***
   - *Link Grant → Related Publication → Patent*
   - *Measure of total innovation in the same intellectual space as an NIH funding area*
   - *What is the effect of NIH funding on total private patenting in a research area?*

Generic regression revisited

Run:

$$Patents_{\delta\sigma\tau} = \alpha_0 + \alpha_1 Funding_{dst} + Controls_{dst} + \varepsilon_{dst}$$

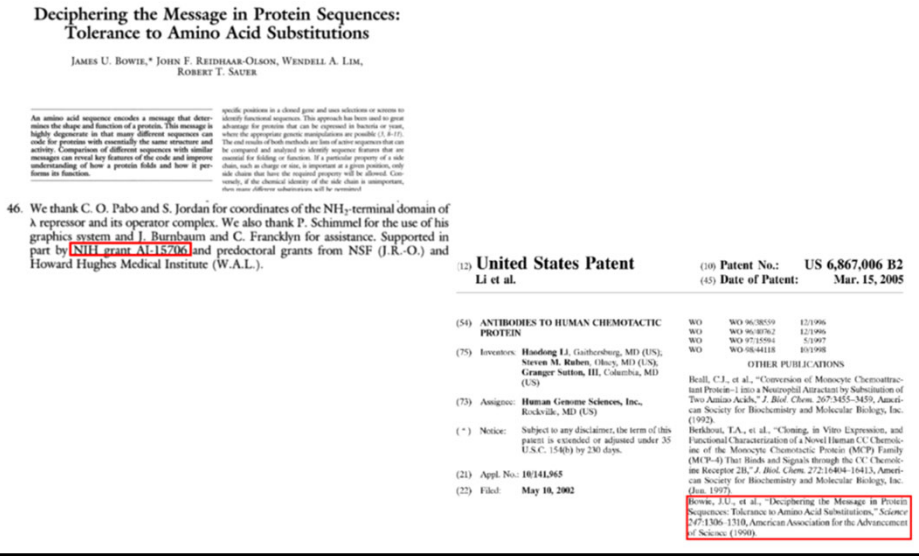
- ***$Patents_{\delta\sigma\tau}$ is the # of patents linked to research area dst***
- ***Patents in $Patents_{\delta\sigma\tau}$ can be in different disease or science areas, and can be issued in $t' \neq t$.***

Generating causal estimates

- ***Use fixed effects to control for many unobserved disease area and science area trends***
  - *Funding varies at DST level → can include disease-science FEs, disease-year FEs, science-year FEs*
  - *The remaining variation is within disease-year or within science-year.*

- ***Robustness check: Instrument remaining variation in funding***
  - *DST funding is made up of funding for individual grants.*
  - *Grant applications are given cardinal scores, but funded on the basis of ordinal scores.*
  - *One DST will sometimes receive more funding than another because its grants have higher ordinal scores but the same cardinal score.*
  - *Instrument* $\texttt{Funding}_{dst}$ *with funding for the subset of grant applications that were funded for this reason*

Tracking the vertical chain of biomedical research



Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

James U. Bowie,* John F. Reidhaar-Olson, Wendell A. Lim, Robert T. Sauer

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

(12) **United States Patent**
Li et al.

(10) **Patent No.: US 6,867,006 B2**
(45) **Date of Patent: Mar. 15, 2005**

(54) **ANTIBODIES TO HUMAN CHEMOTACTIC PROTEIN**

(75) Inventors: **Haodong Li**, Gaithersburg, MD (US); **Steven M. Ruben**, Olney, MD (US); **Granger Sutton, III**, Columbia, MD (US)

(73) Assignee: **Human Genome Sciences, Inc.**, Rockville, MD (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 230 days.

(21) Appl. No.: **10/141,965**

(22) Filed: **May 10, 2002**

WO WO 96/38559 12/1996
WO WO 96/40762 12/1996
WO WO 97/15561 5/1997
WO WO 98/44118 10/1998

OTHER PUBLICATIONS

Beall, C.J., et al., "Conversion of Monocyte Chemoattractant Protein-1 into a Neutrophil Attractant by Substitution of Two Amino Acids," *J. Biol. Chem.* 267:3455–3459, American Society for Biochemistry and Molecular Biology, Inc. (1992).

Berkhout, T.A., et al., "Cloning, in Vitro Expression, and Functional Characterization of a Novel Human CC Chemokine of the Monocyte Chemotactic Protein (MCP) Family (MCP-4) That Binds and Signals through the CC Chemokine Receptor 2B," *J. Biol. Chem.* 272:16404–16413, American Society for Biochemistry and Molecular Biology, Inc. (Jun. 1997).

Bowie, J.U., et al., "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science* 247:1306–1310, American Association for the Advancement of Science (1990).

What remains a problem for identification?

- ***After including fixed effects, variation in*** $\texttt{Funding}_{dst}$ ***comes from:***
  - *Changes to funding for different science areas within a disease-year*
  - *Changes to funding for different diseases within a science-year*
- ***Identifying condition:***
  - *The NIH does not direct $$ to disease-science research areas in response to changes in innovative potential specific to that disease-science combination*
- ***This would be a problem:***
  - *After Gleevec, the NCI sees that cell-signaling is especially important for cancer treatments and responds by increasing funding for cell-signaling research*

This seems plausible and efficient, why doesn't it happen?

- ***NIH funding rules make it difficult to allocate more funding to exciting areas***
  - *All grant application review is done in study sections:*
    - *Study sections score all applications on their science topic, across all disease areas*
    - *Scores are normalized within a study section*
    - *The number and scope of study sections rarely changes*
- ***Institutes must fund grants in order of their normalized score until their budget runs out***

> *"Applications describing some of the most productive, highest impact work may be assigned to too few study sections, causing too much of the 'best science' to compete with itself [and] the scope of some study sections is restricted to research with relatively low impact, resulting in undeserved 'entitlements'…"*
>
> ——Ellie Ehrenfeld, former director of the NIH Center for Scientific Review, 2006

Example of identifying variation

Cell Signaling Study Section			Tumor Physiology Study Section		
Rank	Disease	Raw Score	Rank	Disease	Raw Score
1	Cancer	10	1	Cancer	8.2
2	Diabetes	9.8	2	Cancer	8.1
3	Cancer	9.2	3	Cancer	7.6
4	Cancer	9.1	4	Cancer	6.4
5	Cancer	8.2	5	Cancer	5.4
6	Diabetes	7.6	6	Diabetes	5.2
7	Cancer	7.6	7	Diabetes	4.8
8	Diabetes	7.5	8	Diabetes	4.4

Example of identifying variation

Cell Signaling Study Section			Tumor Physiology Study Section		
Rank	Disease	Raw Score	Rank	Disease	Raw Score
1	Cancer	10	1	Cancer	8.2
2	Diabetes	9.8	2	Cancer	8.1
3	Cancer	9.2	3	Cancer	7.6
4	Cancer	9.1	4	Cancer	6.4
5	Cancer	8.2	5	Cancer	5.4
6	Diabetes	7.6	6	Diabetes	5.2
7	Cancer	7.6	7	Diabetes	4.8
8	Diabetes	7.5	8	Diabetes	4.4

Example of identifying variation

Cancer Institute (NCI)			Diabetes Institute (NIDDK)		
Rank	Study Section	Raw Score	Rank	Study Section	Raw Score
1	Cell	10	2	Cell	9.8
1	Tumor	8.2	6	Cell	7.6
2	Tumor	8.1	6	Tumor	5.2
3	Cell	9.2	7	Tumor	4.8
3	**Tumor**	**7.6**	8	Cell	7.5
4	Cell	9.1	8	Tumor	4.4
4	Tumor	6.4			
5	Cell	8.2			
5	Tumor	5.4			
7	**Cell**	**7.6**			

Example of identifying variation

Cancer Institute (NCI)			Diabetes Institute (NIDDK)		
Rank	Study Section	Raw Score	Rank	Study Section	Raw Score
1	Cell	10	2	Cell	9.8
1	Tumor	8.2	6	Cell	7.6
2	Tumor	8.1	6	Tumor	5.2
3	Cell	9.2	7	Tumor	4.8
3	Tumor	7.6	8	Cell	7.5
4	Cell	9.1	8	Tumor	4.4
4	Tumor	6.4			
5	Cell	8.2			
5	Tumor	5.4			
7	Cell	7.6			

Example of identifying variation

Cancer Institute (NCI)			Diabetes Institute (NIDDK)		
Rank	Study Section	Raw Score	Rank	Study Section	Raw Score
1	Cell	10	2	Cell	9.8
1	Tumor	8.2	6	Cell	7.6
2	Tumor	8.1	6	Tumor	5.2
3	Cell	9.2	7	Tumor	4.8
3	Tumor	7.6	8	Cell	7.5
4	Cell	9.1	8	Tumor	4.4
4	Tumor	6.4			
5	Cell	8.2			
5	Tumor	5.4			
7	Cell	7.6			

Example of identifying variation

Cancer Institute (NCI)			Diabetes Institute (NIDDK)		
Rank	Study Section	Raw Score	Rank	Study Section	Raw Score
1	Cell	10	2	Cell	9.8
1	Tumor	8.2	**6**	**Cell**	**7.6**
2	Tumor	8.1	6	Tumor	5.2
3	Cell	9.2	7	Tumor	4.8
3	Tumor	7.6	8	Cell	7.5
4	Cell	9.1	8	Tumor	4.4
4	Tumor	6.4			
5	Cell	8.2			
5	Tumor	5.4			
7	**Cell**	**7.6**			

IV strategy: make use of "windfall" funding

Exploit the wedge between scores and funding priority

Instrument Funding$_{dst}$ with:

Lucky_Funding$_{dst}$ = Funding for applications just above the IC's funding payline

Control flexibly for:

1. Number of DST applications near the payline
   - Since the number of applications near the threshold may be endogenous
2. The average raw scores of funded and unfunded grant applications to a DST
   - Identify only off differences in funding priority driven by the scores and ranks of applications from other DSTs

Compare two DSTs with the same number of grant applications and the same average scores around the payline

- Difference in funding for these DSTs is driven by wedge between funding priority and cardinal scores.

Does NIH funding increase total private-sector patenting?

	(1)	(2)	(3)	(4)	(5)
Weighted Patent Counts					
DST Funding ($10 mill)	3.921*** *(0.415)*	2.748*** *(0.741)*	2.839*** *(0.605)*	2.255*** *(0.330)*	2.786*** *(0.238)*
Elasticity	0.673	0.471	0.487	0.387	0.478
R-squared	0.520	0.807	0.829	0.954	0.965
Observations	11,110	11,110	11,110	11,110	11,110
Year FEs	X	X	X	X	X
Disease X Science FEs		X	X	X	X
Disease X Year FEs			X	X	X
Science X Year FEs				X	X
Application Quality and Lagged Funding Controls					X

IV evidence: Luck at funding payline

	(1)	(2)		(3)	(4)	(5)	(6)
	First Stage			**IV Estimates**			
	DST Funding ($10 mill)			Citation-Linked		PMRA-Linked	
DST Funding, just awarded grants ($10 mill)	2.772*** *(0.732)*	1.718*** *(0.544)*	DST Funding ($10 mill)	3.140*** *(1.193)*	3.319** *(1.666)*	3.885*** *(1.433)*	2.890* *(1.579)*
			Elasticity	1.016	1.074	0.666	0.495
R-squared	0.414	0.414		0.349	0.417	0.453	0.564
Observations	11,110	11,110		10,536	10,536	10,536	10,536
Full FEs	X	X		X	X	X	X
Application Quality and Lagged Funding Controls		X			X		X

Findings

- ***30% of NIH grants produce research that is cited by a private sector patent***
- ***$10 million of NIH funding → 2.3 more industry patents (net of crowd-out)***
- ***$1 dollar in NIH funding → $0.4 to $1.7 in PDV of drug revenue***
- ***Disease spillovers are large***
  - *Half of all patents generated by additional NIH investments are for diseases different from the one intended*
- ***NIH funding increases overall firm R&D investment:***
  - *Increased firm patenting in one area is not offset by declines in another; rather, both appear to increase*