

# ***Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research***

---

***Jeff Furman*** (Boston U & NBER)

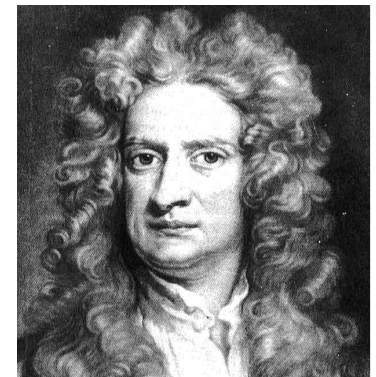
***Scott Stern*** (Northwestern U & NBER)

***2nd COMMUNIA Conference 2009  
Global Science and the Economics of  
Knowledge-Sharing Institutions***

# ***Broad Motivation: understanding role of institutions in “standing on shoulders of giants”***

---

- *Long-term economic growth depends on the ability to draw upon an ever-wider body of scientific & technical knowledge (Rosenberg, Mokyr, Romer, Aghion & Howitt, David & Dasgupta)*
- *Economic historians, institutional economists, and sociologists emphasize the role of “institutions”*
  - *however, the micro-foundations of knowledge accumulation are, by and large, still a “black box”*
  - *many challenges to assessing impact of institutions*
    - *knowledge flows are difficult to track*
    - *institutions are difficult to identify & characterize*
    - *knowledge is assigned endogenously (not randomly) to institutional environments (Ex1: firm vs university patents) (Ex2: collaboration vs. solo research)*



# ***Research Agenda & Approach***

---

## **■ *Overall Research Agenda***

- Do institutions have an impact on the accumulation & diffusion of productive knowledge?*
- What are the distributional consequences of institutions on knowledge generation and diffusion?*

# Research Agenda & Approach

---

## ■ **Overall Research Agenda**

- *Do institutions have an impact on the accumulation & diffusion of productive knowledge?*
- *What are the distributional consequences of institutions on knowledge generation and diffusion?*

## ■ **A Natural Experiments Approach**

- *exploit (exogenous) changes in institutions governing knowledge generation and diffusion*
- *helps address identification & endogeneity*

# *How do institutions affect knowledge flows?*

## *An Empirical Framework (I)*

---

### ■ *Ideal research design:*

- *for many units of knowledge*
  - *of various characteristics (e.g., quality, timing, field ,...)*
- *randomly assign units of knowledge to alternative institutional settings*
  - *testing to ensure that samples have similar observable features*
- *observe impact of settings on knowledge*
  - *e.g., on diffusion, generation, etc.*
- *compare knowledge diffusion patterns across settings*

# *How do institutions affect knowledge flows?*

## *An Empirical Framework (II)*

---

- ***More feasible research design:***
  - *observe instances in which multiple units of knowledge shift between institutional settings*
    1. *after **sufficient time** for the 'importance' of knowledge to be identifiable*
    2. *for reasons that are **exogenous** to the knowledge itself*
      - *ideally, knowledge generated at different times*
      - *ideally, knowledge shifted at various times*
      - *ideally, matched to controls of similar characteristics*

# ***Research Agenda*** *(sample of projects underway)*

---

- *Ongoing projects on impact of scientific institutions on knowledge generation & diffusion*
  - *BRCs (with Scott Stern)*
    - *impact of libraries for biological materials on diffusion of knowledge associated with those materials*
    - *raises questions about well-functioning institutions in science*
  - *Stem Cell Research (with Fiona Murray)*
    - *what is the impact of the Bush Administration policy on the rate of growth of stem cell research in the US?*
    - *how does funding policy affect growth/structure of scientific fields?*
  - *False Science / Retractions (with Fiona Murray)*
    - *drivers and impact of acknowledged mistakes in science*
    - *investigates extent to which publication system performs well in science*

# ***Today's paper:***

## ***Research Challenge & Approach***

---

- *Research Question: how do institutions affect cumulative knowledge production?*
  - ***the selection effect*** – *knowledge with particular qualities gets selected into a particular institutional environment*
  - ***the marginal impact*** – *conditional on selection, institutions might amplify the impact of knowledge for future researchers*
    - *e.g., university patents - inherently important or effective institution?*
    - *e.g., collaboration – projects more important or output made better?*

# **Today's paper:**

## **Research Challenge & Approach**

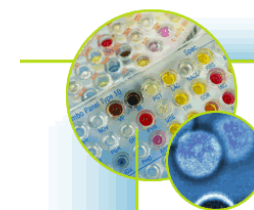
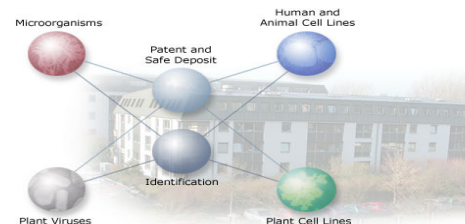
---

- *Research Question: how do institutions affect cumulative knowledge production?*
  - **the selection effect** – *knowledge with particular qualities gets selected into a particular institutional environment*
  - **the marginal impact** – *conditional on selection, institutions might amplify the impact of knowledge for future researchers*
    - *e.g., university patents - inherently important or effective institution?*
    - *e.g., collaboration – projects more important or output made better?*
- *Approach in this paper*
  - *identify a specific institution – Biological Resource Centers (ATCC)*
  - *combine detailed qualitative understanding of the institution with research design that enables allows us to observe single “piece” of knowledge in two distinct institutional environments*

# Institutional Context: **Biological Resource Centers (BRCs)**

---

- **BRCs collect, certify, preserve, & offer access to biological organisms for use in research and commercial development**
  - cell lines (e.g., stem cells), microorganisms (e.g., anthrax), tissue cultures, animal models (e.g., knockout mice)
- **300+ BRCs in world, national collections in life sciences centers**
  - US: American Type Culture Collection, Manassas, Virginia
  - Germany: DSMZ, Braunschweig
- **Over 100 years old, but increased importance in last 25 yrs**
  - early 1950s: “Transportable” biomaterials
  - extraordinary growth in microbiology over last 50 years
  - 1980 Budapest Treaty: All patented living organisms must be deposited in International Patent Depositories
- **“Like the Library of Congress for biological materials”**



# ***BRCs as Economic Institutions***

---

- *Economic institutions such as BRCs have the power to amplify the impact of scientific discoveries by enabling future generations to build on past discoveries*
  - *within the life sciences, “standing on shoulders” often requires access to specific biological materials or materials collections*
    - *the precision of a given experimental design depends upon the understanding of the biological materials it employs*
- *The evolution of BRCs as economic institutions seems to reflect the key collective action problems in transferring biological knowledge, via biological materials, across research generations*
  - *Authentication / Certification*
  - *Long-Term Preservation*
  - *Independent Access*
  - *Economies of Scale and Scope*



# ***BRCs as Sociological Institutions***

---

- *Culture collections may serve as complement or substitute to scientific research network*
  - *researchers who are central to the research network may be (relatively) inaccessible to those on the periphery*
    - *access to materials (as well as citations and co-authorships) may require prior network centrality or high status*
  - *impartial institutions, such as BRCs, may*
    - *enable peripheral researchers to obtain access to materials (and knowledge) that may otherwise be accessible only to those at the core*
    - *promote peripheral researchers' results & materials to the core*
  - *could result in more dense network, which circulates materials (and knowledge) more effectively*

## **BRCs: An institutional response to the demand for public goods supporting the accumulation of useful knowledge**

---

- *Establishing effective institutions (funding, leadership, etc.) is subject to a public goods problem*
  - *Even if a research-enhancing institution is funded, the incentives to participate as a depositor may be too low without explicit rules or norms*
  - *The growth in importance of BRCs as a key intermediary reflects systematic efforts over time to overcome these collective action problems*
- ★ But, do BRCs actually enhance the diffusion of scientific knowledge? How?**

## ***Empirical Approach: Do BRCs enhance the diffusion of scientific knowledge? A Natural Experiments Approach***

---

- 1. BRC Deposits are linked with specific scientific research articles or patents (referred to as “BRC-linked” articles)*
  - 2. Each BRC-linked article can be matched w/ article controls*
  - 3. Some BRC deposits occur long after initial publication*
    - even many years after discovery, control over refrigerators can be transferred from specific research labs to BRCs*
  - 4. Some post-publication deposits are arguably exogenous*
    - e.g., **special collections** “shifted” due to funding expiration at initial host institutions, faculty retirement, or faculty job change resulting in change in location of “refrigerator”*
    - can test exogeneity by looking for pre-deposit citation “spike”*
- ★ Allows us to observe variation in the impact of a single “piece” of knowledge across two distinct institutional environments*

# *Special Collections facilitate our Identification Strategy*

---

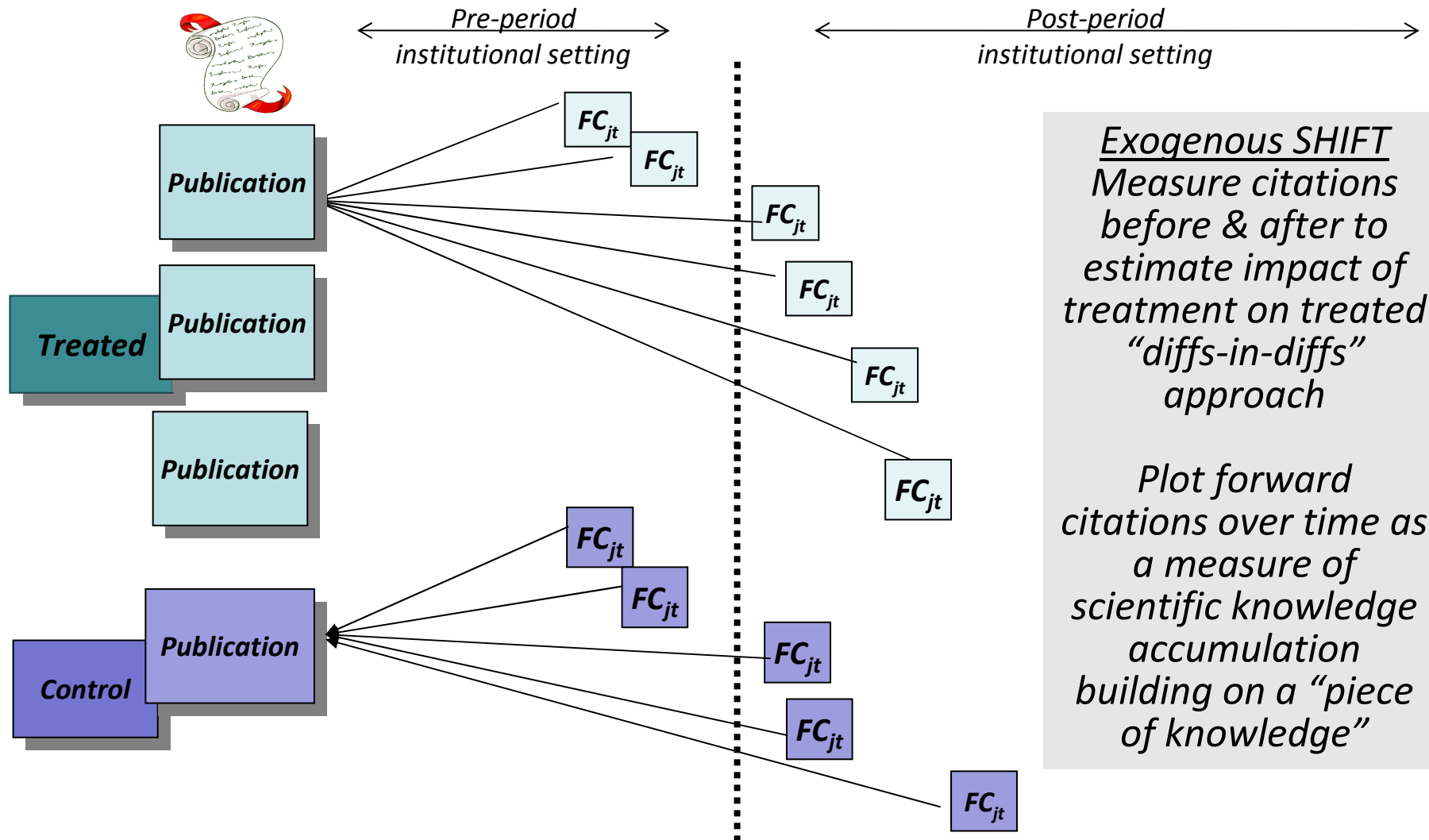
- *Special Collections = Potentially exogenous deposits*
  - *since there is a gap between publication and deposit, impact of deposit is identified separately from article “fixed effects”*
  - *since the “lag” varies across deposits, impact of deposit is identified separately from age effects*
  - *since the deposit date varies, impact of deposit is identified separately from year effects*

# ***Special Collections Data: The Key for “Exogenous” Deposit Dates***

---

- *Gazdar Collection*
  - *initially maintained by Adi Gazdar & John Minna*
  - *accessioned (1994) when Gazdar left NCI for Texas*
  
- *Tumor Immunology Bank (TIB Collection)*
  - *initially maintained by Salk Institute*
  - *accessioned beginning in 1981 (funding)*
  
- *Human Tumor Bank (HTB Collection)*
  - *initially maintained at Sloan-Kettering*
  - *accessioned beginning in 1981 (funding)*

# Empirical Framework: Diffs-in-diffs analysis of citations received



# Diff-in-Diffs Estimator: Isolating Selection from Marginal Impact

---

- To disentangle “marginal” from selection effect, need to isolate out the portion of the total impact accruing from BRC deposit itself

- $POST-DEPOSIT = 1$  for article-years after deposit

- ❖  $\phi = \text{Selection Effect}$ ;  $\psi = \text{Marginal effect}$

- Article family effects: estimate selection & marginal effects

$$CITES_{i,j,s,t} = f(\varepsilon_{i,j,s,t}; \alpha_j + \beta_s + \delta_t + \phi ATCC_i + \psi POST - DEPOSIT_{i,t})$$

- Article fixed effects: estimate marginal effects only

$$CITES_{i,j,s,t} = f(\varepsilon_{i,j,s,t}; \gamma_i + \beta_s + \delta_t + \psi POST - DEPOSIT_{i,t})$$

# Data Sources

---

## ■ *BRC Deposit Data*

- *ATCC catalogue ([www.atcc.org](http://www.atcc.org)); lists scientific reference info*

## ■ *Control Data*

- *Medline/PUBMED*
- *Most-Related Article Control*
  - *article in journal/year “most-related” to BRC-linked article (Medline)*
- ❖ *Controls help identify Selection in Article-Family FE models; and help identify Age and Year effects in Article FE models*

## ■ *Citation Data*

- *Scientific Citation Index (many undergraduates / much monitoring)*

# Data Characteristics

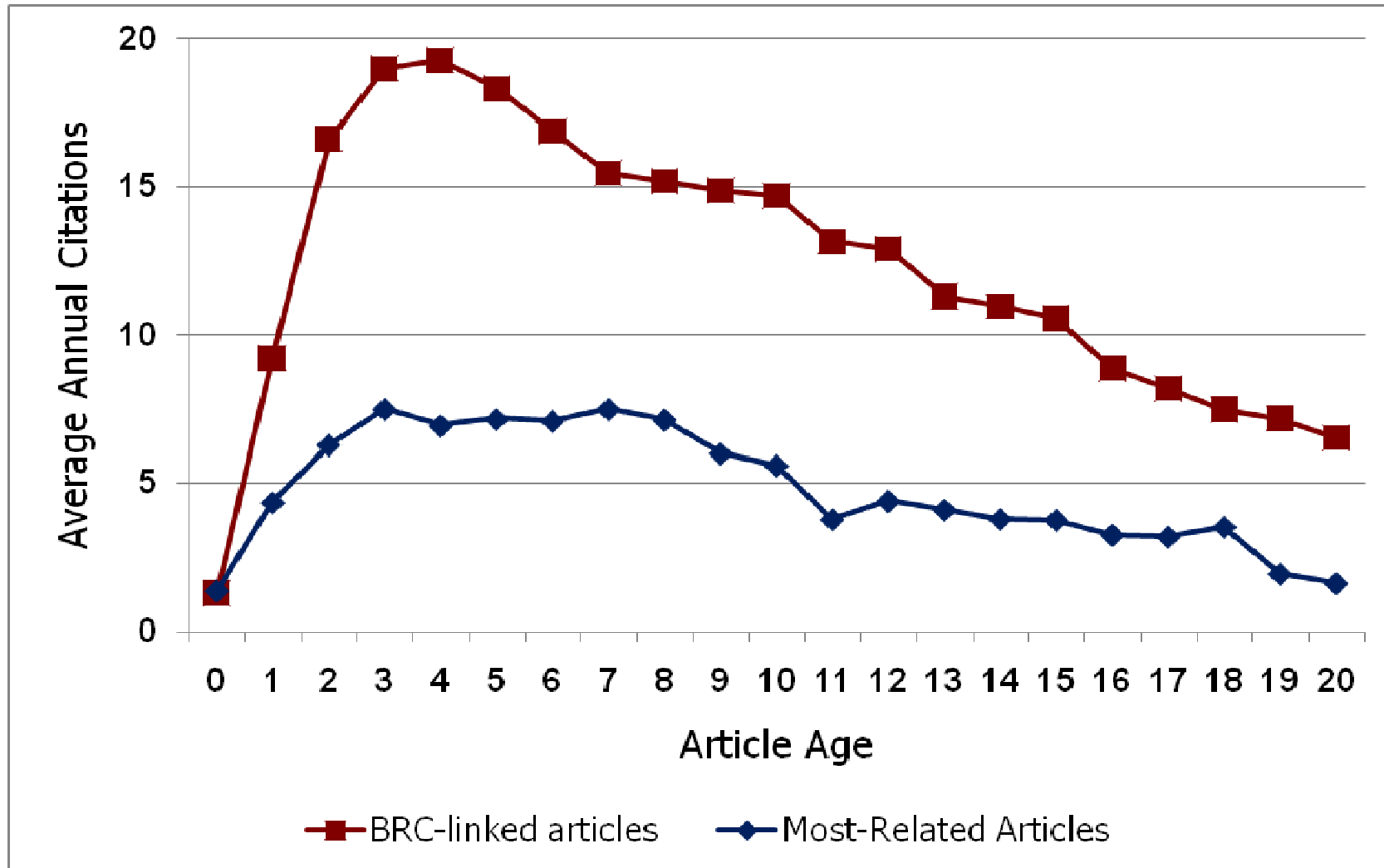
---

- *289 Article Pairs*
  - *between 1971 and 2001*
- *Key “Timing” Variables*
  - *publication year*
  - *deposit year*
  - *citing year*
  - *publication age (= citing year – publication year)*
  - *deposit age (= citing year – deposit year)*
- *Deposit & Reference Characteristics*
  - *deposit: Access Price, Collection, Depositor*
  - *reference: # of Pages, Authors, Backward Citations, Author Affiliations, etc.*

# Summary Statistics

<b>VARIABLE</b>	<b>MEAN</b>	<b>STANDARD DEVIATION</b>	<b>MIN</b>	<b>MAX</b>
<b>ARTICLE CHARACTERISTICS</b> (n=216 articles)				
BRC ARTICLE	0.50	0.50	0	1
PUBLICATION YEAR	1979.40	4.54	1970	1992
DEPOSIT YEAR	1983.63	3.47	1981	1994
US AUTHOR	0.76	0.43	0	1
TOP 50 UNIVERSITY AUTHOR	0.15	0.36	0	1
TOP JOURNAL	0.56	0.48	0	1
<b>ARTICLE-YEAR CHARACTERISTICS</b> (n=4857 article*year observations)				
YEAR	1989.79	7.23	1970	2001
AGE	11.27	7.23	0	31
FORWARD CITATIONS	7.28	15.73	0	186
CUMULATIVE CITATIONS	91.67	178.86	0	2333
<i>Forward Citations received from</i>				
US AUTHOR	2.60	5.87	0	59
TOP 50 UNIVERSITY AUTHOR	0.99	2.50	0	33
TOP JOURNAL	3.37	8.02	0	99

# *Average annual citations by age, BRC-linked articles vs. Control articles*



# Diff-in-Diffs: Substantial Selection & Marginal Effects (Baseline Specification)

<i>Negative Binomial Models</i>	<i>Forward Citations</i>
	(3-3) <i>Selection vs. Marginal</i>
<i>BRC-Article (Selection)</i>	<b>[2.12]</b> 0.752 (0.297)
<i>BRC-Article, Post-Deposit (Marginal)</i>	<b>[1.713]</b> 0.538 (0.248)
<i>Article Family FE</i>	X
<i>Age FE</i>	X
<i>Calendar Year FE</i>	X

**112%**  
More  
Than  
Controls

**71%**  
Boost  
After  
Deposit

\* Cond FE Neg. Bin. Models, coefficients as IRRs; bootstrapped SEs

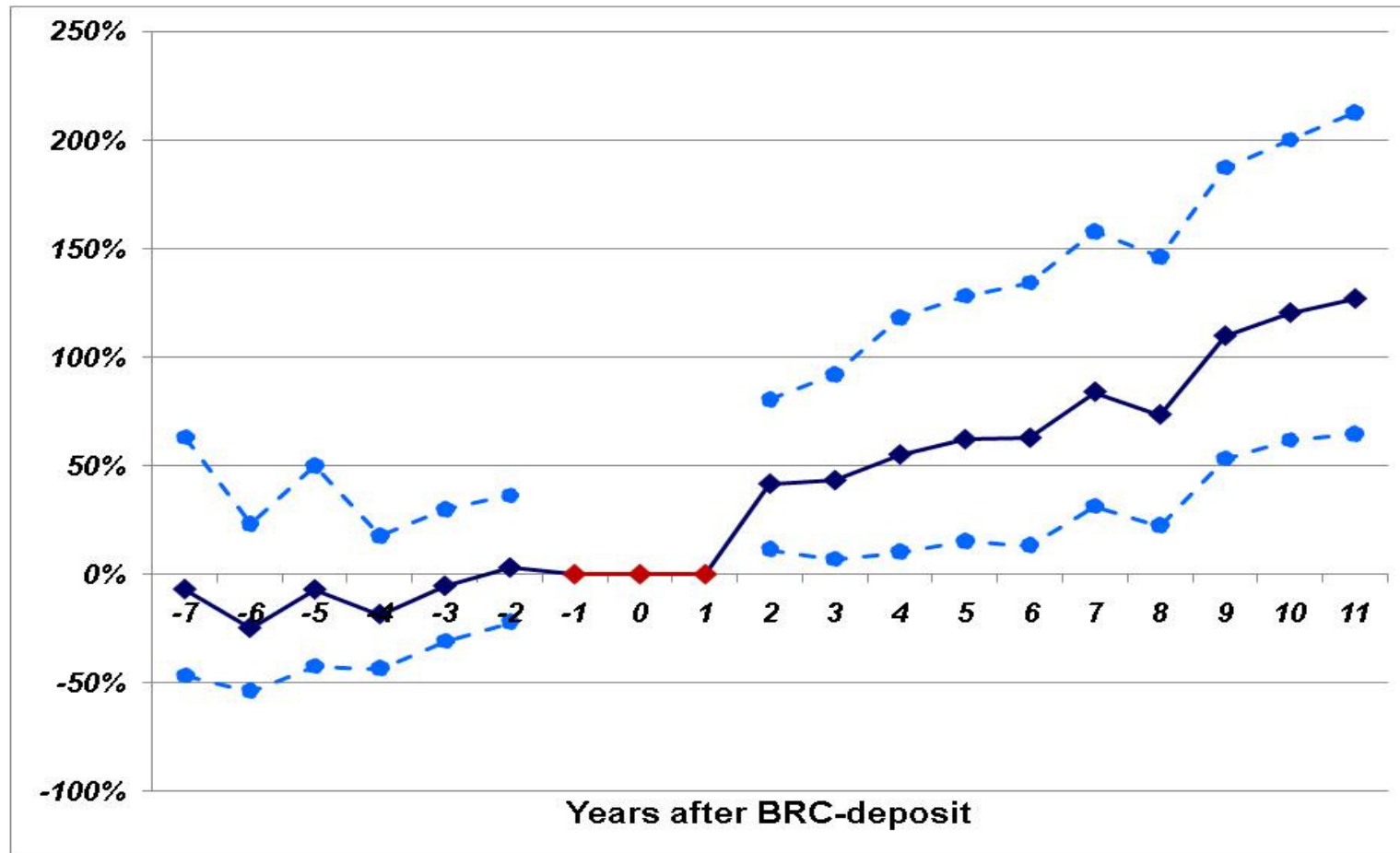
## Diff-in-Diffs: Marginal Effects only

<i>Negative Binomial Models</i>	<i>Forward Citations</i>
	(3-4) <i>Marginal Effects only</i>
<i>BRC-Article, Post-Deposit (Marginal)</i>	<b>[2.248]</b> 0.810 (0.360)
<i>Article FE</i>	X
<i>Age FE</i>	X
<i>Calendar Year FE</i>	X

**122%  
Boost  
After  
Deposit**

\* Cond FE Neg. Bin. Models, coefficients as IRRs; bootstrapped SEs

# Impact of Deposit Grows Over Time and Does Not Exist Prior to Deposit



- *This suggests that deposit is, indeed, exogenous and that diff-in-diffs approach usefully identifies marginal (post-deposit) effects*
- *Conditional FE NB model*

# ***Additional Findings: Implications for certification & democratization of research***

---

- *Certification effects?*
  - *citation boost greater for*
    - *articles not originally published in Top 50 journal*
    - *articles whose RP author not at Top 50 university*
    - *articles whose initially lower citation levels*
  - *deposit → greater boost in citations from articles published in Top Journals*
  - *consistent with certification effect (democratizing effect)*
  
- *Expansion of research community?*
  - *deposit → greater boost from articles with multiple subjects*
  - *deposit → boost in #new institutions, journals, countries*
  - *consistent with expansion in set of participants in research community*

# ***Robustness***

---

- *Use of alternative Control Samples*
  - *Nearest Neighbor only*
  - *Most-Related Article only*
- *Omitting Window Period Observations*
- *Across Different “Experiments”*
- *Controlling for Article Characteristics*
- *Using Special Collection Subsamples*

# Implications

---

- *BRCs seem to influence future knowledge accumulation via both:*
  - **selection effect** – positive sorting
  - **marginal effect** – deposit → positive, significant and long-lived increase in follow-on research associated with initial discovery
- *In addition:*
  - *relative to traditional grant mechanisms, BRCs seem effective at maximizing the knowledge pool available for future research*
  - *BRCs appears to have an impact on distribution of follow-on research, consistent with an effect whereby materials are “certified” via association with BRC and participation in research community is broadened after deposit*
- *Though cumulative knowledge production is central to economic growth, the extent of knowledge growth depends on the effectiveness of often “invisible” institutions, whose features can be influenced by public policy*

# ***Extras***

---

# Types of Biological Resource Centers

<b>Center Type</b>	<b>Examples</b>
<i>Public / Non-Profit national collections</i>	<ul style="list-style-type: none"> <li>• ATCC (USA)</li> <li>• DSMZ (Germany)</li> <li>• Collection of Microorganisms (Japan)</li> </ul>
<i>Public / Non-Profit specialized collections</i>	<ul style="list-style-type: none"> <li>• Coriell Medical Research Institute (human genetic mutant cell lines)</li> <li>• National and Infectious Disease (HIV materials)</li> <li>• Ribosomal Database Project</li> <li>• Agricultural Research Service Culture Collection (NRRL)</li> </ul>
<i>Private, industrial collections</i>	<ul style="list-style-type: none"> <li>• Merck (antibiotics screening collection, clinical microbiology collection); Institute for Fermentation (IFO)</li> </ul>
<i>Specialized University collections</i>	<ul style="list-style-type: none"> <li>• <i>Escherichia coli</i></li> <li>• <i>Bacillus</i></li> <li>• <i>Fusarium</i></li> </ul>
<i>Life Sciences Data Management Institutions</i>	<ul style="list-style-type: none"> <li>• Ribosomal Database Project (RDP)</li> <li>• Institute for Genomic Research (TIGR)</li> </ul>

Source: adapted from OECD, 2001

# Examples of BRCs



**ATCC**  
The Global Bioresource Center™



**Cooperative Human Tissue Network**

OSU Tissue Procurement Services  
- MidWestern Division

*The Cooperative Human Tissue Network (CHTN) is a group of six academic institutions funded by the National to work together to provide remnant human tissue to researchers throughout the US and Canada.*

**Prostate SPORE National  
Biospecimen Network Pilot**  
National Cancer Institute

*The Prostate SPORE NBN pilot will support ... collaborative projects related to prostate cancer research currently underway at participating SPOREs across the nation.*

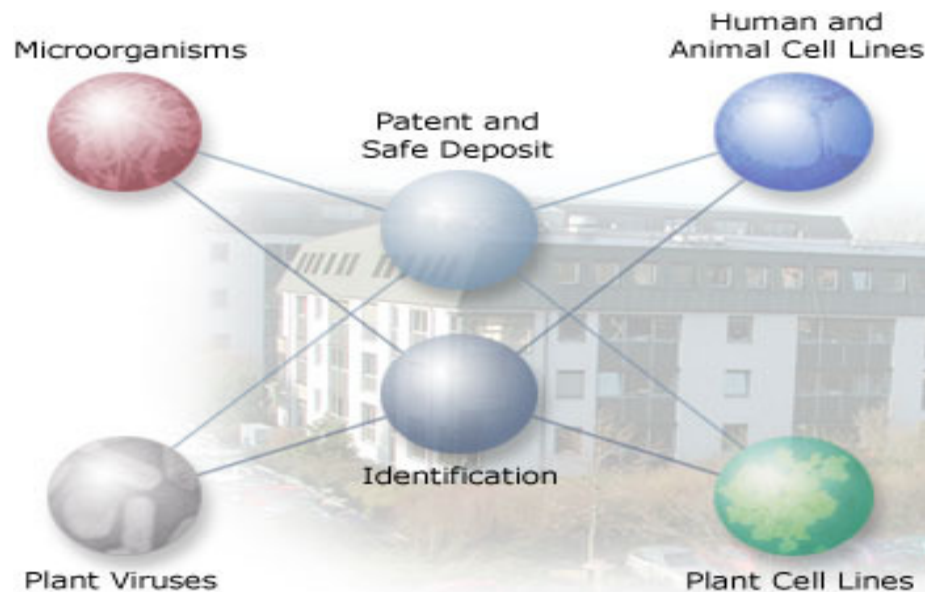
RIKEN BIORESOURCE CENTER  
JAPAN COLLECTION OF  
MICROORGANISMS

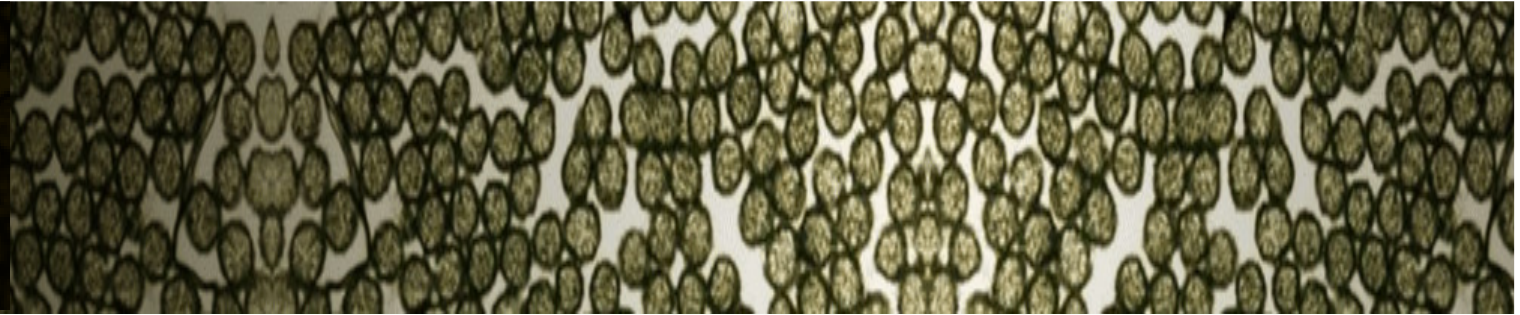
*Japan Collection of Microorganisms (JCM) was founded in 1980 at [RIKEN](#) (The Institute of Physical and Chemical Research)... JCM supplies authentic microorganisms to researchers in the fields of life sciences and biotechnology.*

# Deutsche Sammlung von Mikroorganismen **DSMZ** Zellkulturen



- *German Collection of Microorganisms & Cell Cultures*
  - *Braunschweig, Lower Saxony*
- *“With more than 15,500 microorganisms, 900 plant viruses, 600 human and animal cell lines, 750 plant cell cultures and more than 6,700 cultures deposited for the purposes of patenting, we have demonstrated our obligation to serve science for decades.”*





- *The Lady Mary Fairfax CellBank Australia is a facility providing cell lines to the research community, both within Australia and overseas. The facility has three main aims.*
  - *Firstly, CellBank Australia is designed to act as a secure repository for cell lines developed by the local research community, including cultures derived from unique Australian species.*
  - *Secondly, the facility will make it easier for scientists in this region to access quality-controlled cell lines for their experimental work.*
  - *Thirdly, we aim to provide a resource promoting good cell culture practice in research Australia-wide.*
  - <http://www.cellbankaustralia.com/>

# Research Design & Implementation

---

- *Examine a specific institution: Biological Resource Centers*
  - *central, yet “invisible” institutions supporting life sciences research*
  - *institutional alternative to peer-to-peer network*
  - *collect, certify, preserve, & offer access to physical biomaterials*
  - *alternative to direct control over materials by network participants*
- *Exploit features of BRC deposit & access process:*
  - *deposits linked to scientific articles (can identify controls)*
  - *some BRC deposits occur long after initial scientific publication*
  - *some post-publication deposits*
- *Develop and implement a diff-in-diffs estimator of impact of BRC deposit on subsequent scientific citation patterns*
  - ➔ *“The Citation Revolution Meets the Identification Revolution”*

# Key Results

---

- *Diff-in-diffs estimates suggest that each effect is statistically significant and of important magnitude:*
  - ***Selection (~100% more citations than control articles)***
  - ***Marginal (~50-125% boost in citations after BRC-deposit)***
- *Further, the marginal impact*
  - *is not observed in citations prior to the date of deposit*
  - *increases substantially with deposit “age”*
  - *increases in magnitude over the past two decades*
  - *a **rate of return calculation**, which combines scientific “citation” productivity estimates with the cost of BRC deposit, suggests that this knowledge-enhancing institution offers a nearly three-fold efficiency gain in terms of inducing citations, relative to traditional grants*
- *Does this type of “invisible” institution foster a higher level of cumulativeness than a reliance on private brokerage?*
  - *when can formal institutions substitute for networks? complement?*

# ***Extended Analysis, Additional Directions, and Concerns***

---

- *Extended Analysis I: rate of return calculation*
  - *very crude!*
- *Extended Analysis II: distributional impact*
  - *which articles get “boost”?*
    - *highly cited (less cited) articles?*
    - *articles in top tier (lower tier) journals?*
    - *articles with US-based (non-US) authors?*
  - *from which citing articles does “boost” occur?*
    - *citations from top tier (lower tier) journals?*
    - *citations from US-based (non-US) authors?*
- *Most promising additional directions?*
  - *institutions? sociology? economics?*
- *Key concern – substitution or addition?*

# ***Motivation: Knowledge accumulation & long-term research productivity***

---

- *Standing on Shoulders is a key requirement for sustained growth*
  - *if the knowledge stock does not expand or cannot be accessed, diminishing returns will eventually set in (economic growth stagnates)*
- *The production of knowledge alone does not guarantee that others use it or can build upon it*
  - *knowledge transfer is usually costly (e.g., tacitness, stickiness)*
  - *strategic secrecy further limits the available knowledge pool*
  - *even if available in principle, relevant calculation is the cost of drawing from the knowledge stock versus “reinventing the wheel”*
  - *consequences of failures in knowledge transmission and storage can be dire – e.g., after Greek & Roman civilizations (Mokyr)*
- *Individual incentives to contribute to cumulative knowledge production are limited*

# *Incentives for Cumulative Knowledge Production*

---

- *Establishing a knowledge hub within a technical community involves a collection action problem*
  - *role for public funding / cooperation among competitors*
- *Even if a knowledge hub is funded, the incentives to participate as a depositor may be too low without explicit rules or norms*
  - *social objective: maximize the impact of prior knowledge on reducing the costs to discovering new knowledge*
  - *as long as knowledge producers care about the impact of their knowledge (for intrinsic, career, or strategic reasons), positive deposit incentives*
  - *however, potential depositors trade off overall impact of knowledge with potential for rent extraction through continued control over knowledge*
    - *example: lots of citations or lots of coauthorships?*

# *The Role of Institutions in the Cumulative Process*

---

- ***Institutions**, from specific facilities to legal rules & norms, may play a key (but potentially “invisible”) role in cumulative process*  
(Nelson; David & Dasgupta; Mokyr)
  - *key: institutions can lower the cost of access to useful knowledge & increase probability of faithful retention*
- ***Institutions** can be linked to the cumulative process in two distinct ways:*
  - ***Selection: positive sorting***
  - ***Marginal: incremental impact of the institution itself***
- *For example...*
  - *Are **university patents** more well-cited because they are **intrinsically more innovative** or because they originate in an **institution** whose objectives and procedures facilitates the cumulative process?*

# Authentication: The HeLa Scandals

---



- *The fidelity of discovered knowledge cannot be guaranteed by the initial discoverer but must be able to be replicated*
- *This tenet was thrown into question by the **HeLa Scandals***
  - *HeLa Cell Line (from cervical cancer patient **Henrietta Lacks**) JHU established as 1<sup>st</sup> perpetual & transportable in vitro cell line (HelaGrams)*
  - *though materials could be transferred among laboratories, inaccurate identification scheme for distinguishing cell lines*
    - *contamination via: cross-contamination, genetic drift & viral infection*
  - *experiments designed to investigate the properties of lung cancer cells of an 80-year-old white male actually conducted on the cervical cancer cells of a 31-year-old black woman*
- *Misidentification may have induced the most serious scientific errors in the history of the modern life sciences*
  - *similar characterizations across experiments gives rise to the “magic bullet” theory of cancer, providing impetus for the War on Cancer*
  - *not simply a problem of second-tier labs, contamination was common at elite laboratories known for precision (Salk, UC-SF, Pasteur, etc.)*

# Authentication: *Restoring Trust*

---



- *Overcoming the HeLa “problem” required (at least) 3 actions*
  - *Public identification of contaminated cell lines*
  - *Reevaluation of published research findings, including refutations of faulty interpretations due to contaminations*
  - *Systematic provision of certified cell lines between research labs*
- *Along with individual actors such as Walter Nelson-Rees, BRCs were instrumental in overcoming the contamination problems*
  - *Use of BRC materials to establish the existence and scope of scandal*
  - *Hit Lists published in Science beginning in 1974*
  - *Long-term R&D to detect and defend against systematic misidentification*
- *BRCs are at the forefront today of ensuring biomaterials fidelity*
  - *“We provide the authenticated biological materials...you decide how to use them in your research” (ATCC mission)*
  - *Genetic characterization and validation*
  - *However, even today, contamination rates are estimated to be over 15%, drawing on a sample of highly cited materials (Masters, 2002; PNAS, 2002)*

# Long-Term Preservation

---



- *The importance of a given piece of knowledge (and the physical materials required to exploit that knowledge) are often only recognized long after the time of initial discovery*
  - *the productivity of current research depends on the commitment to store materials and the data underlying discoveries, even before awareness of particular applications*
- *Brock's Unlikely Bacteria*
  - *1967: Thomas Brock discovers Thermus Aquaticus in Yellowstone National Park geysers, classified as an extremophile*
  - *deposited (without any particular hope of application) at American Type Culture Collection*
  - *1983: Kary Mullis conceives of DNA replication scheme requiring DNA polymerase that can resist extreme temperature variation*
  - *after initial attempts locally, identification of Taq at ATCC*
  - *1989: Thermus Aquaticus, Molecule of the Year*
- *PCR becomes the foundational tool for DNA replication for modern molecular biology & biotechnology*