Science & Technology Agents of Revolution (STAR) Database: A Progress Report

Michael R Darby & Lynne G. Zucker
both UCLA & NBER

Presentation for the AEA Session
DATA WATCH: New Developments in Measuring Innovation Activity
American Economic Association Meetings
Atlanta, GA, January 4, 2010

We acknowledge support by the Ewing Marion Kauffman Foundation (Grants 2008-0028 & 2008-0031) and the National Science Foundation (Grant SES-0830983).

Acknowledgements

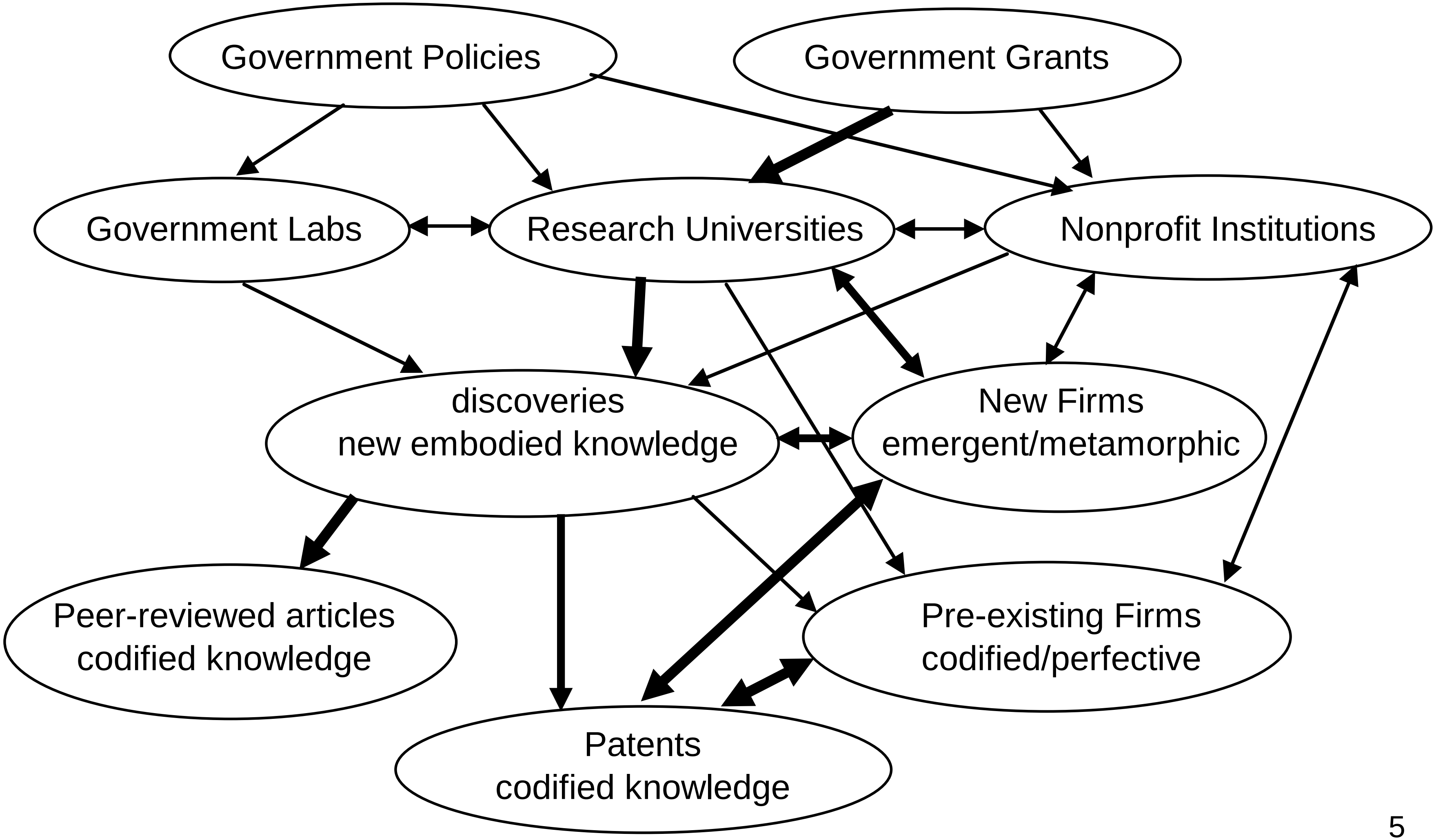
- Initial construction of STAR has been supported by
  - Ewing Marion Kauffman Foundation (Grants 2008-0028 & 2008-0031)
  - National Science Foundation (Grant SES-0830983)
- Infrastructure support is provided by
  - UCLA School of Public Affairs & Social Science Computing
  - NBER's Productivity Program & Data Archive
- Additional funding is being sought to complete and extend STAR

STAR: An Accessible Digital Library

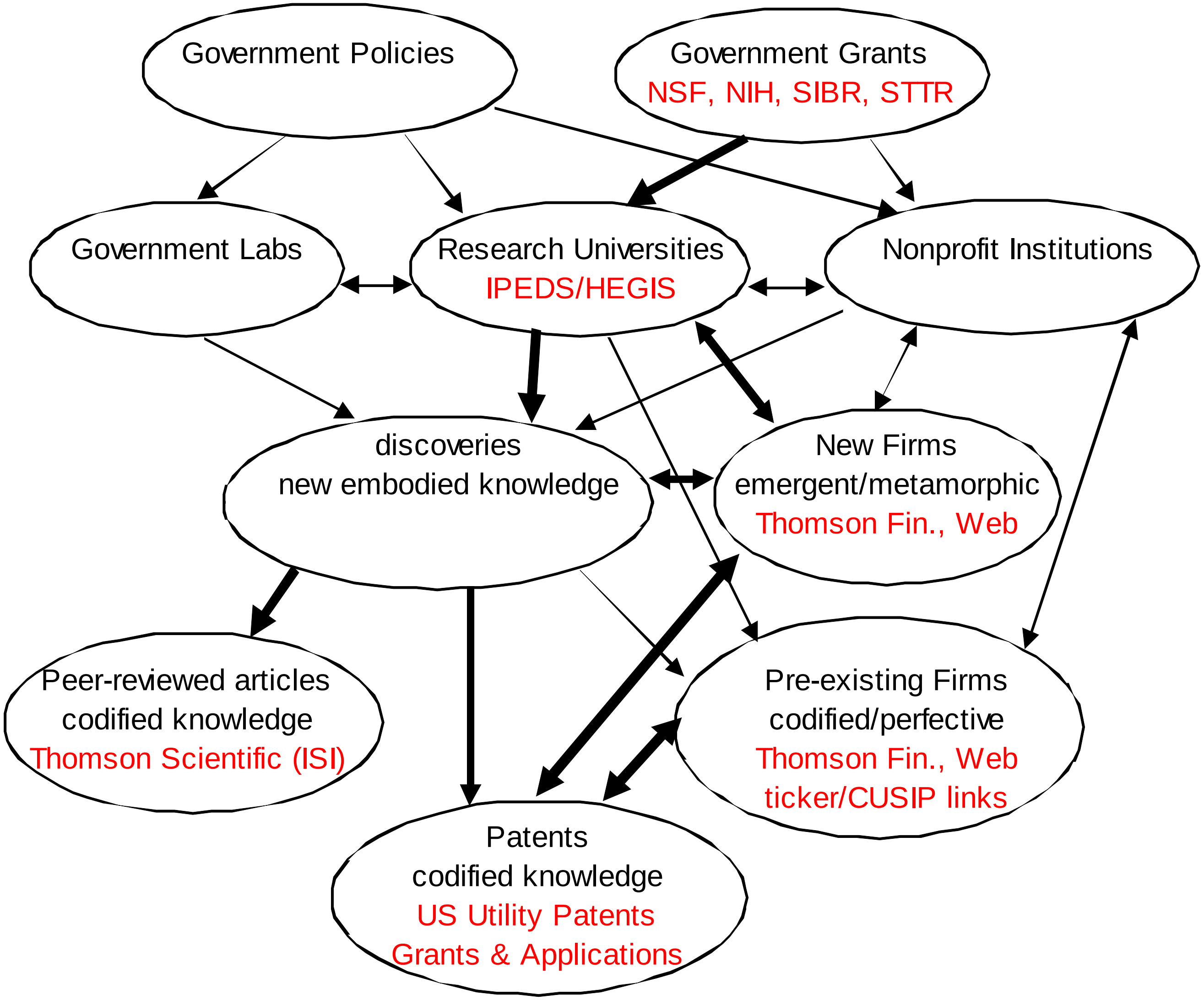
- Track major knowledge creation and its flows among people and organizations
  - Link organizations within & across databases
  - Link scientists & engineers within & across databases (stars wear many hats)
- Focus on metamorphic rather than perfective innovation
  - New/transforming industries
  - Entrepreneurial, initially non-public firms

OVERVIEW

Major Features of the U.S. National Innovation System in the STAR Database



Major Features of the U.S. National Innovation System in the STAR Database



Government Policies

Government Grants
NSF, NIH, SIBR, STTR

Government Labs

Research Universities
IPEDS/HEGIS

Nonprofit Institutions

discoveries
new embodied knowledge

New Firms
emergent/metamorphic
Thomson Fin., Web

Peer-reviewed articles
codified knowledge
Thomson Scientific (ISI)

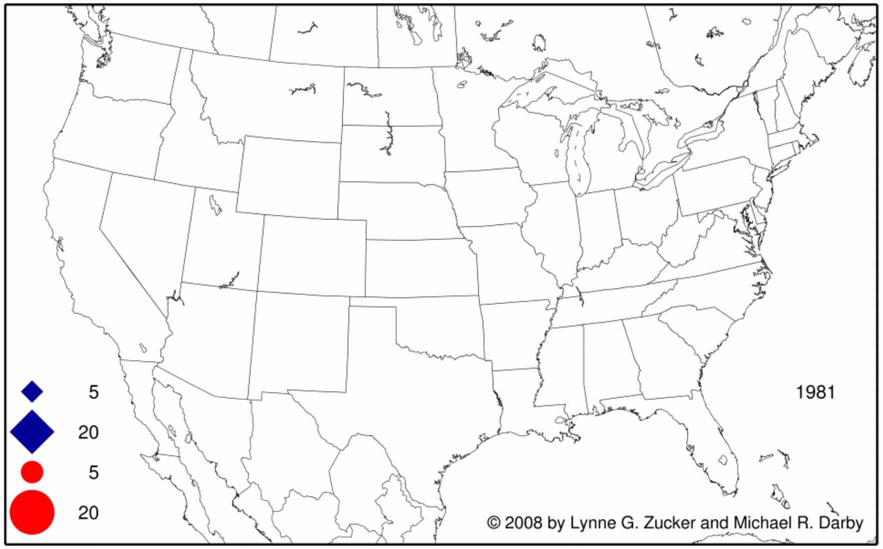
Pre-existing Firms
codified/perfective
Thomson Fin., Web
ticker/CUSIP links

Patents
codified knowledge
US Utility Patents
Grants & Applications

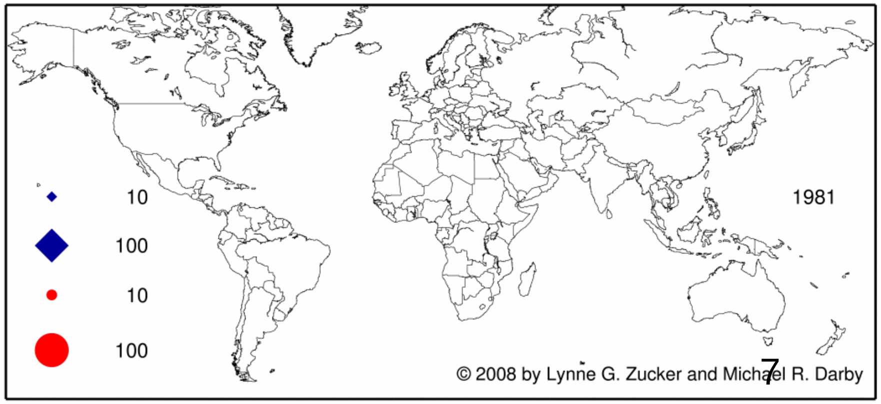
Example 1: Nano-Biotechnology

- You can do a lot with just the articles database alone once organization types have been identified
- Maps: universities & firms on university-firm nano-bio articles
- Co-location: bench-science collaborations

Firm (♦) & University (●) Collaborators



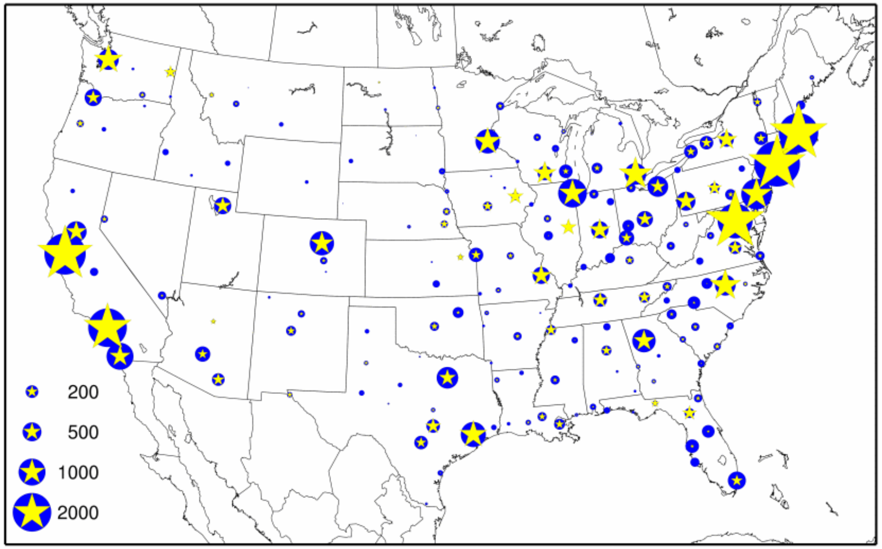
© 2008 by Lynne G. Zucker and Michael R. Darby

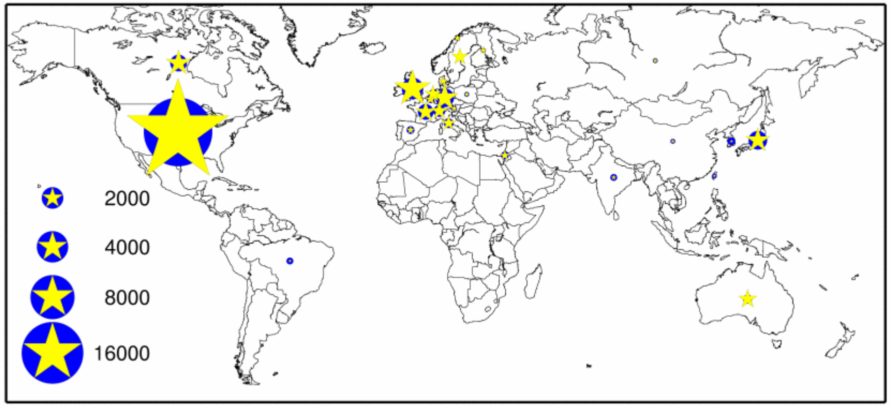


© 2008 by Lynne G. Zucker and Michael R. Darby

Example 2: Star Scientists & Firm Entry

Cumulative Biology/Chemistry/Medicine Stars (yellow) and Firm Entry (blue), 1981-2004





Stars' Debuts, Residence & Migration, 1981-2004
All Science & Technology (S&T) Areas

	U.S.A.	Japan	OECD Europe Top-14 S&T Countries	Top-25 S&T Countries
Unique Stars Ever Resident	**3670**	**266**	**1960**	**6599**
Professional Debuts	**3354**	**176**	**1194**	**5105**
Net Inward Migration	**-23**	**13**	**-39**	**-51**
Stars Resident 12/31/2004	**3331**	**189**	**1155**	**5054**
Net Inward Migration/Debuts	**-0.7%**	**7.4%**	**-3.3%**	**-1.0%**

Percent of Stars Ever Co-Authoring with Firms and Total Articles per Firm-Tied and Non-Firm-Tied Star By Country or Set of Countries, 1981-2004
All Science & Technology (S&T) Areas

	U.S.A.	Japan	OECD Europe Top-14 S&T Countries	Top-25 S&T Countries
Stars-Percent with Firm Ties	**64.3%**	**59.5%**	**48.2%**	**54.6%**
Articles/Star with Firm Ties	**85.5**	**55.6**	**48.3**	**73.9**
Articles/Star with no Firm Ties	**16.7**	**6.2**	**8.0**	**10.8**

SOME DETAILS OF WHAT IS COMING

STAR Database

- NSF & Kauffman Foundation Funding
  - Theory-driven
    - Inputs and outcomes of innovation process
    - Interconnections coded, flows of info/knowledge & knowledge capture including natural excludability
  - Web-deployed &/or archived for use at NBER
- Micro-level data where available
  - Can graft on outside data easily
  - Codebooks like those posted at nanobank.org

Current Status

- Construction well under way
- We will begin soon migrating STAR to the NBER in phases
- When we have a stable database there, beta testing will phase in

Beta Tests

- Beta test I: begins second half 2010
- Beta test II: in 2011
- If you want invitation to participate, e-mail darby@ucla.edu and zucker@ucla.edu
- Beta test I will be 9-12 months pre-launch
  - Takes time to start using data, find problems and fix them
- Beta test II for data elements not in initial public version of STAR

STAR Database Components

Area	Tranche	Database components	Organizations	Scientists & Engineers	Location
Government investments	1	NSF, NIH, DoD & DoE Grants	Universities (mainly)	PIs & co-PIs	web
	1	SBIR & STTR Grants	Firms	PIs/employees & faculty	web
	1	Government laboratories	Govt., univs., firms	[employees*]	web
	1	Other grants linked as available	Universities (mainly)	PIs & co-PIs	web
policies	2	Bayh-Dole inventor shares by university	Universities	[inventors*]	NBER/web?
Science & Engineering	1	Articles (Web of Science† or other sources)	Univs., firms, others	authors	NBER
	1	High-Impact articles†	Univs., firms, others	authors	NBER
	1	Highly-Cited authors†	Univs., firms, others	authors	web/NBER
	2	Discoveries (e.g., GenBank)	often not included	authors, inventors	NBER/web?
	2	U.S. doctoral dissertations	Universities	author, advisors	NBER/web?
	2	N.R.C. Doctoral programs studies	University depts.	[faculty*]	web
	2	IPEDS enrollment and degrees data	Universities	[faculty*]	web
Commerce	1	U.S. utility patents	Firms, univs., others	inventors	web
	1	U.S. utility patent applications	Firms, univs., others	inventors	web
	1	Public firms concordance to tickers/CUSIPs	Firms	officers, directors, key employees	web
	2	Public & private firms-web based	Firms	officers, directors, key employees	web
	1	Public & private firms-Thomson Financial	Firms	officers, directors, key employees	NBER
	**	Public & private firms-links to Census ILBD & LEHD databases	Firms	officers, directors, all employees	NBER
	1	Public and private firms-other sources‡	Firms	officers, directors, key employees	web/NBER

Notes:

* Denotes that information on scientists & engineers only appear in this database as aggregates and averages by organization, not as individual specific data.

** These links will be available only to authorized users in Census facilities.

† Thomson Scientific product.

‡ Government aggregate and firm specific (BEA, Edgar, Census LRD links, other) data.

MAIN CHALLENGES

Acquisition/Parsing/Cleaning of Constituent Databases

- STAR is currently between 1 and 2 TB
- Proprietary data is expensive and licences restrict access
- Public data (e.g., NIH grants) often is a challenge to parse into data fields & clean
  - Formats change over time & errors common
  - This is a job that you would not take on if you knew what you were getting into

Geocoding

- Standardizing differing naming conventions used in different sources.
- Standardizing non-uniformity in how observations are recorded
- Correcting common mistakes
- For US observations: Providing different geographic units (other than city and state) not available in original data sources, like counties & BEA functional economic areas
  - Novel, useful data available for these units

Organization Matching/Coding

- Each observation is assigned an alpha-numerical code.
- 2 digit alphabetical part designates the organization type.
- Numeric part groups names that are same up to standardization and hand cleaning

First 2 digits	Organization type
FI	Firm
UN	University
NL	National Lab
RI	Research Inst
UG	US Government
HO	Hospital
AS	Academy of Sciences
NO	No Organization
SC	School
OT	Other

Person Matching

- We use probabilistic methods to match persons and assign common IDs within and across constituent databases
- Article database is hardest due to size (about 25 million articles with about 100 million authorships), use of only initials and limited ability to match affiliations
- As we extend matching across databases learn new matches & revise prior coding

CONCLUSIONS

Conclusions

- STAR is a community resource
  - Facilitates replication & extension of research
  - Saves duplication of time, effort & funding
  - Creates a community, facilitates exchanges
- STAR encourages those who link to STAR contribute data that is linked & usable
- Few rewards to creating public goods
  - but STAR does require specific citations to the creators of the specific data elements used