Studying Science
Using Administrative Records
Etoile Project

Paula Stephan
Georgia State and NBER
Paris
September 15-16 2013

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don’t know
- Caltech Etoile Project
  - Framework
  - Data Description
  - Empirical Analysis

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don't know
- Caltech Etoile Project
  - Framework
  - Data Description
  - Empirical Analysis

What We Know

- Knowledge produced by researchers working at institutes and universities is embodied in multiple forms:
  - Publications
  - Patents
  - Training and placement of students
  - Creation of new areas
  - Creation of new collaborations
  - Creation of new firms

What We Know…

- University knowledge in US often produced in labs that resemble firms; directed and overseen by PI's
- Majority of funds for lab come from federal government
- Multiple inputs involved in knowledge production $Q=f(k, t, m, e, s)$
  - Some inputs, such as knowledge and time, are embodied in people

What We Know…

- Characteristics of productive people in terms of location, age, gender, country of birth
- Certain characteristics of productive teams
- Size of teams as measured by metrics such as co-inventors and coauthors over time
- Case studies of composition of team in terms of position
- Case studies placement of new PhDs at firms
- Speed with which knowledge leaks out

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don't know
- Caltech Project
  - Framework
  - Data Description
  - Empirical Analysis
- National and International Activities
- Next Steps

How we know it: Sources

- Federal surveys (SED, SDR); publication and patent data bases
- Observation
- Interviews
- Paper trail—citations to patents and publications; co-authorship patterns on the two
- Surveys
- Linking data sets to one another, such as publications and citations
- Special data requests

How we know it: One off studies

- Freeman *et al.* study of changing patterns of international collaboration in research (Survey, corresponding authors, 3 time periods)
- Conti and Liu study of MIT labs; Conti study of EPFL labs (administrative records; LinkedIn data)
- Sauermann and Roach study of work aspirations of graduate students and postdocs (survey)
- Tambi study of IT skills using “digital breadcrumbs”
- Thursby and Thursby study of ownership of patents invented by university faculty (Match 1995 NRC faculty data with patent data)

How we know it: One off studies

- Furman and Stern study of how deposit of research materials at BRCs affects use of materials and diffusion of materials (link deposits to articles and citations)
- Ding *et al.* study of relationship of gender to patenting, SAB membership, etc. (Proquest Dissertation data linked with publicly available data, including patents)
- Murray *et al*. study of how intellectual property rights affect diffusion and use of mouse models (patent data; citations to relevant “mouse” articles; nearest neighbor citations)
- Own-Smith study of relationship of physical proximity to productivity (floor plans and publication data)

How we know it: One off studies

- Levin and Stephan study of relationship of age, period and cohort effects to productivity
- Uzzi and Jones study of authorship patterns over time
- Franzoni *et al.* study of how changing incentives relate to submission patterns to *Science*
- Stephan *et al.* study of firm placement of new PhD

Problems with One-off Approach

- Herculean effort
- Costly
- Non-reproducible
- Quickly becomes out of date
- Confidentiality issues restrict use
- Difficult to reproduce

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don't know
  - The potential
- Caltech Etoile Project
  - Framework
  - Data Description
  - Empirical Analysis

What we don't know: Examples

- Current analysis often one-dimensional
  - Relationship of productivity to team characteristics
  - To proximity
  - To age
  - To gender
- But production of research is not one dimensional
- Involves multiple inputs: $Q=f(k, t, m, e, s)$
- Know there are other inputs, but focus is almost exclusively on the "t" and the "k" which are embodied in people
- With but rare exception, ignore equipment, materials and space; characteristics of team members
- Ignore prices of inputs, be they wages or costs of other inputs—yet clear that costs affect hiring decisions
- Ignore other measures of output, focusing almost exclusively on publications, citations and patents

What we don’t know

- “One dimensional focus” means we do not apply what we have learned from productivity studies of firms to productivity of labs
- Yet labs are much like firms
- PIs much like entrepreneurs
- Number of questions that we need to address

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don’t know
- Caltech Project
  - Framework
  - Data Description
  - Empirical Analysis
- National and International Activities
- Next Steps

The potential

- Data collected from administrative records at universities
- Linked to other data sources
- Can dramatically cut down on number of “one-off” studies
- Increase ability to replicate and build on other’s research
- Increase timeliness of research

The potential: Star Metrics/Etoile Capability

- Position of individuals working on project; team (faculty, staff scientist, postdoc, graduate student)
- Characteristics of individuals working on the project (year of degree, former institutions, gender; expertise from CV data, digital bread crumbs)
- Publications; patents, citations to those
- Topics of research (topic modeling of grants, publications, dissertations, and patents)
- Placement of PhDs and postdocs (LEHD data; LinkedIn)

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don't know
  - The potential
- Caltech Project—one of Star Metric Institutions
  - Framework
  - Data Description
  - Empirical Analysis

The Etoile Project

- Financial Support
  - The Alfred P. Sloan Foundation
- Team
  - Julia Lane
  - Jacques Mairesse
  - Michele Pezzoni
  - Paula Stephan
  - David Mayo

The Institution

- Size
  - Approximately 300 faculty
  - 980 undergraduates
  - 1250 graduate students
  - 650 (approximate) postdocs
- Focus
  - Science and engineering
  - Minimal focus on humanities and social sciences

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don’t know
  - The potential
- Caltech Project
  - Framework
  - Data Description
  - Empirical Analysis
- National and International Activities
- Next Steps

The Empirical Framework

- Multi dimensional production of research
- Involves multiple inputs: $Q=f(k, t, m, e, s)$
- Include equipment, materials and space; characteristics of team members
- Include prices of inputs, be they wages or costs of other inputs—since costs affect hiring decisions
- Include disciplinary information

The Qualitative Framework

- Three days at Caltech in early January
  - Met with Mory Gharib, Hans W. Liepmann Professor of Aeronautics & Professor of Bioinspired Engineering, Vice Provost for Research
  - Dick Seligman, Associate Vice President for Research Administration
  - David Mayo, Director, Office of Sponsored Research
- Five days at Caltech in early April
  - Conducted in depth interviews with 6 faculty
  - Met with Caltech team

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don’t know
  - The potential
- Caltech Project
  - Framework
  - Data Description
  - Empirical Analysis

Data

- Administrative records of Federal grants for period 2000-2012
- Files
  - Employee file
  - Vendor file
  - Subcontract file
  - Overhead file

Matched Data

- CV data
- Publications and citations
- Patents and citations
- Dissertations
- Topic modeling of grants, dissertations and articles
- Placement data (eventually) from matching dissertation awards to LEHD

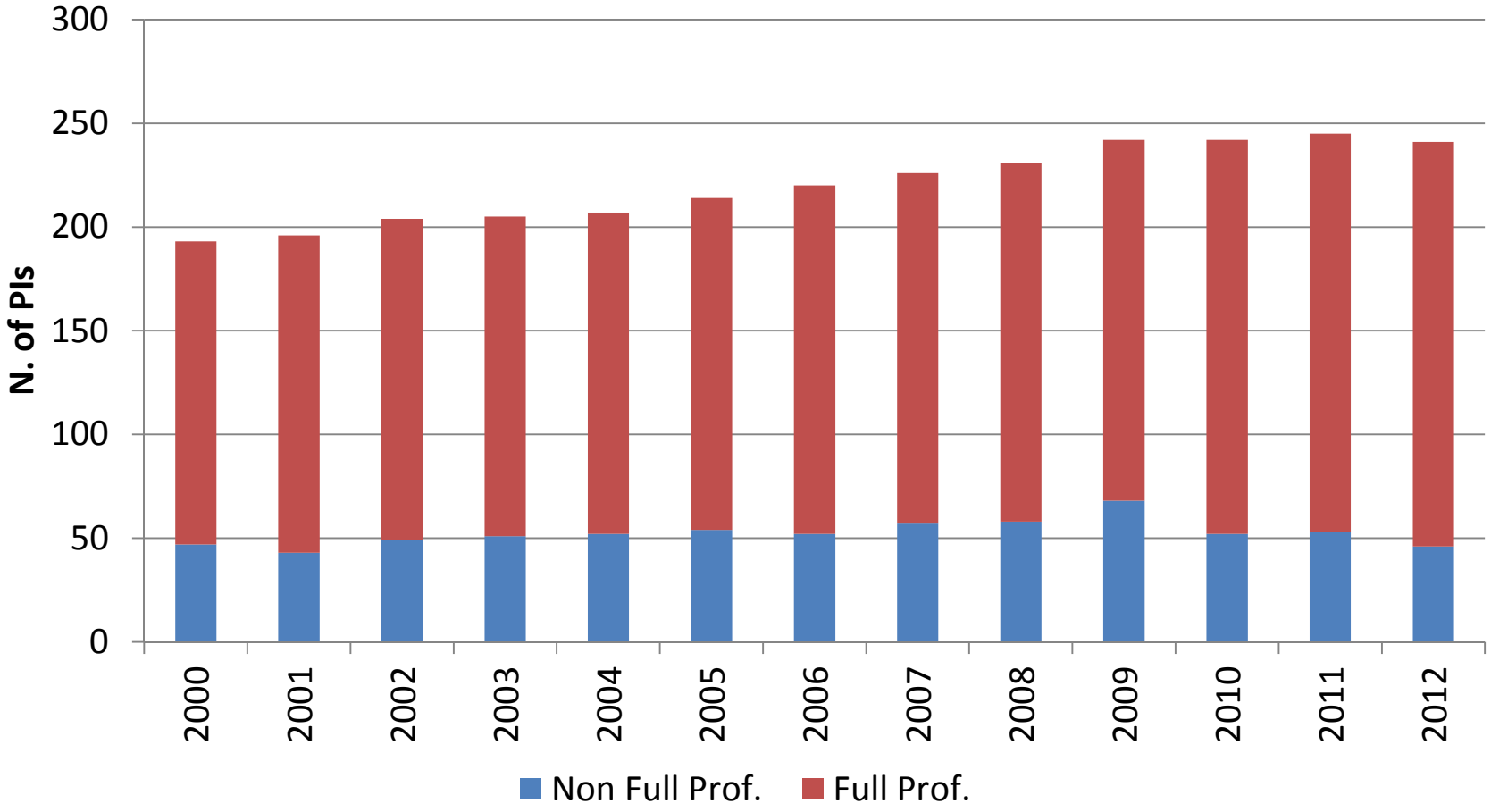
CV Data Sources

- NSF/NIH bio sketches
- Caltech faculty profiles
- Google search
- We use OCR automated as much as possible; rely on python
- Fill in missing data by webscraping from Caltech and other university or other institution faculty profiles, the Caltech library, and possibly Microsoft's API, Google Scholar, or other similar public CV information sites

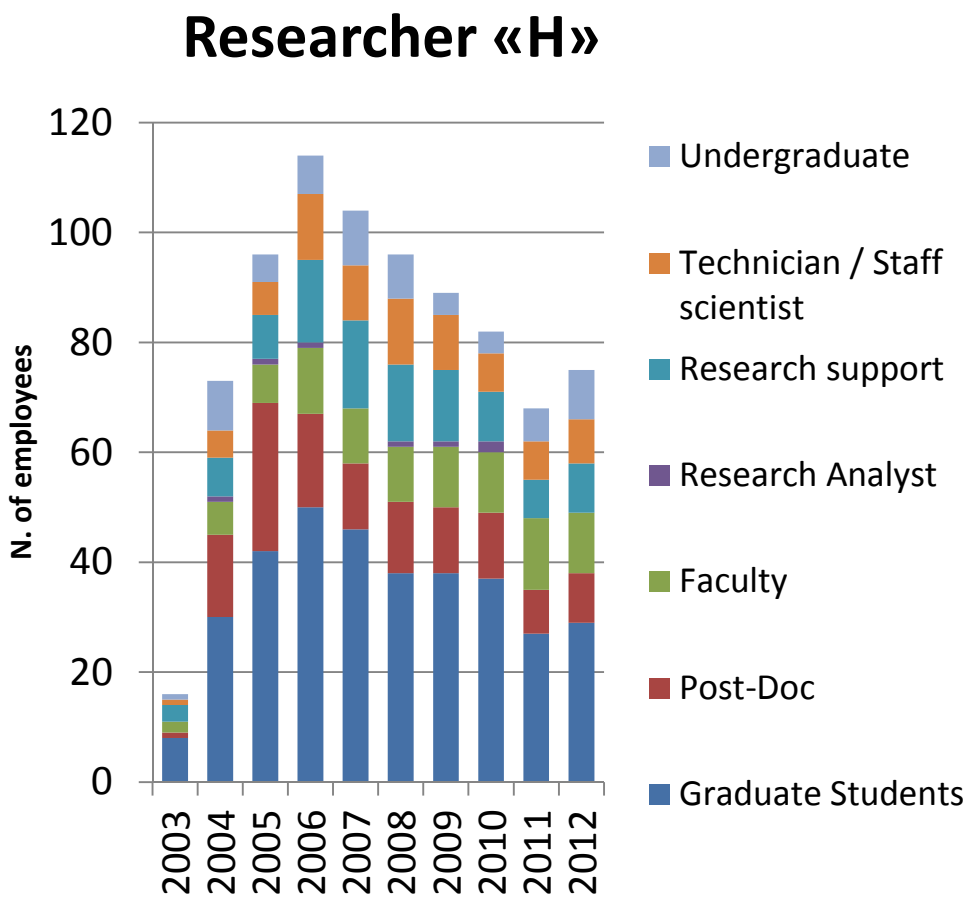
Panel Data description

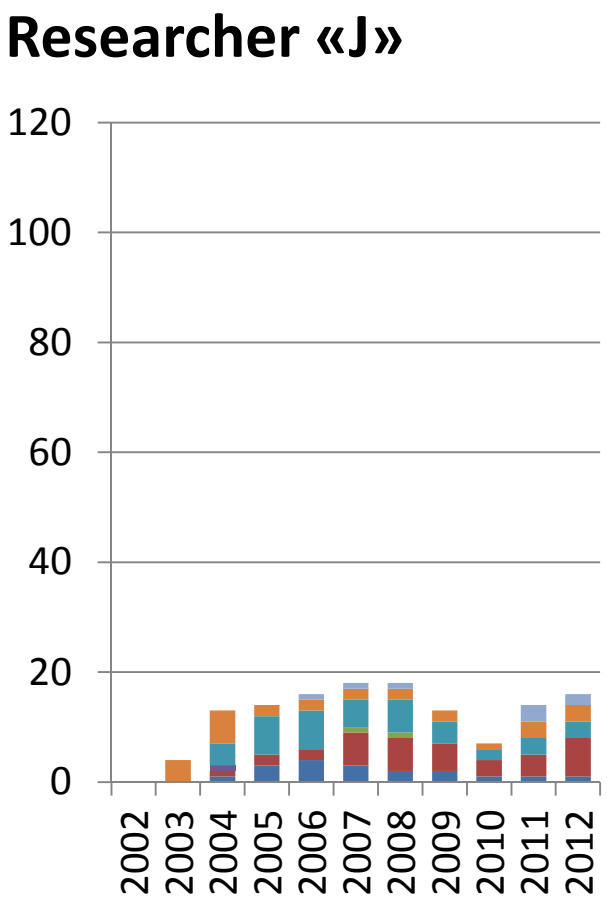
- 276 PIs
  - Limited to faculties for which we have CV information
  - Active during 2000-2013 (PIs retiring/leaving Caltech before 2000 are not considered)
- 60% of the PIs are active from 2000 to 2013 (The panel covers on average 11.25 years per PI)
- We present statistics on: entry&exit, career, research group composition, productivity, funding and expenses, labor&capital, and PI's personal characteristics

The Sample Size

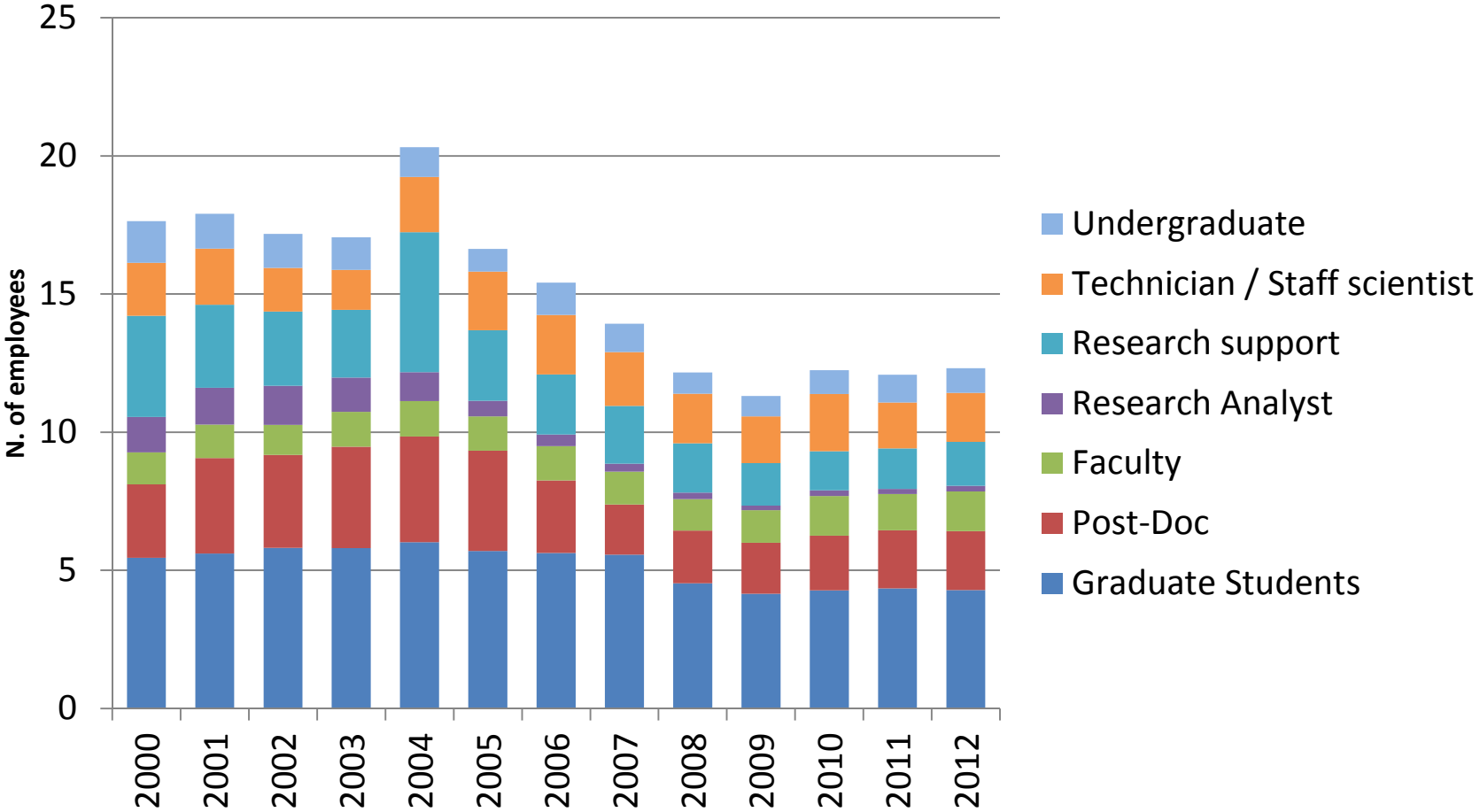


Research teams:
Varies substantially in size and composition

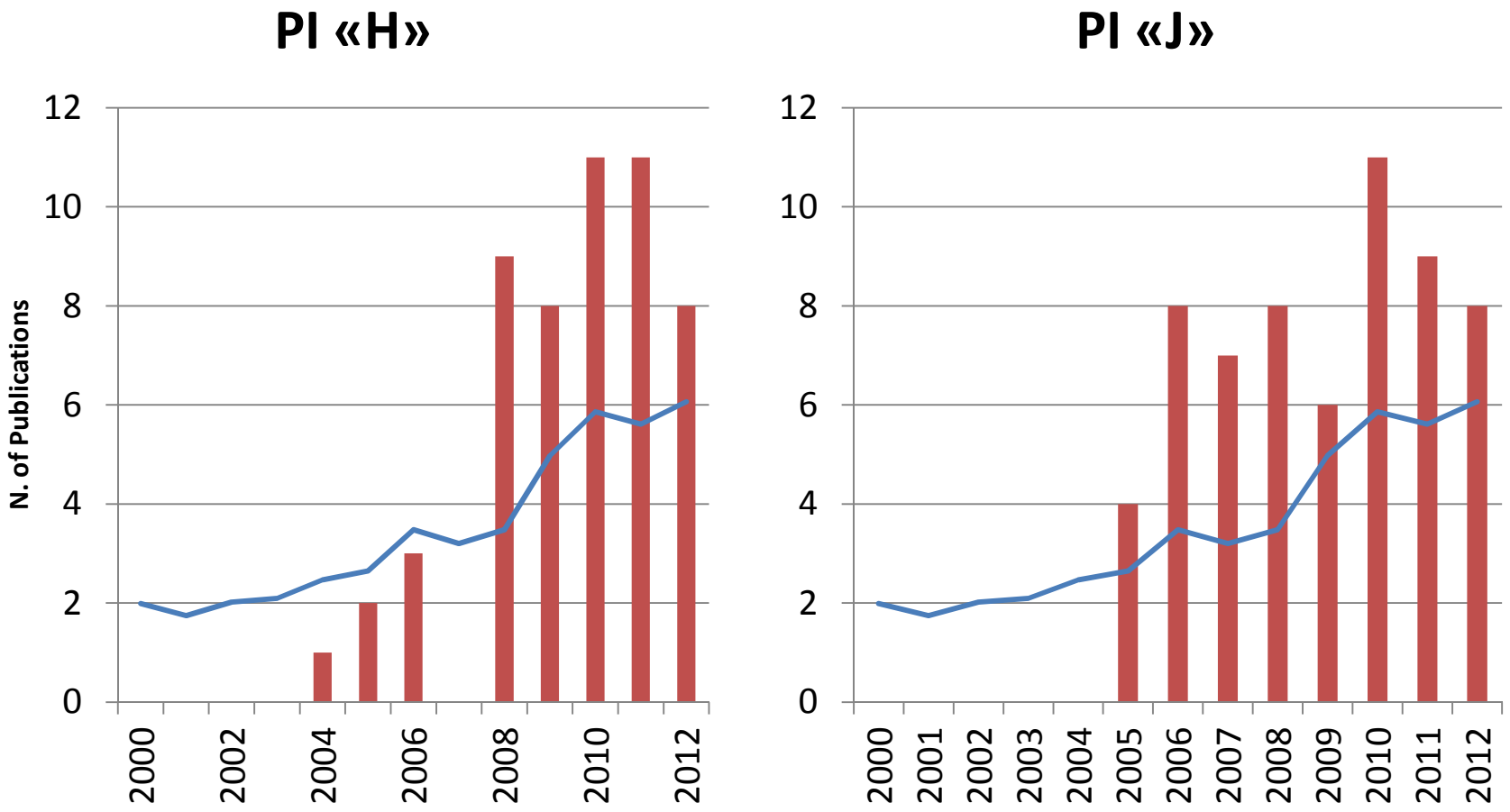




Research group composition:
average per PI



Number of Publications

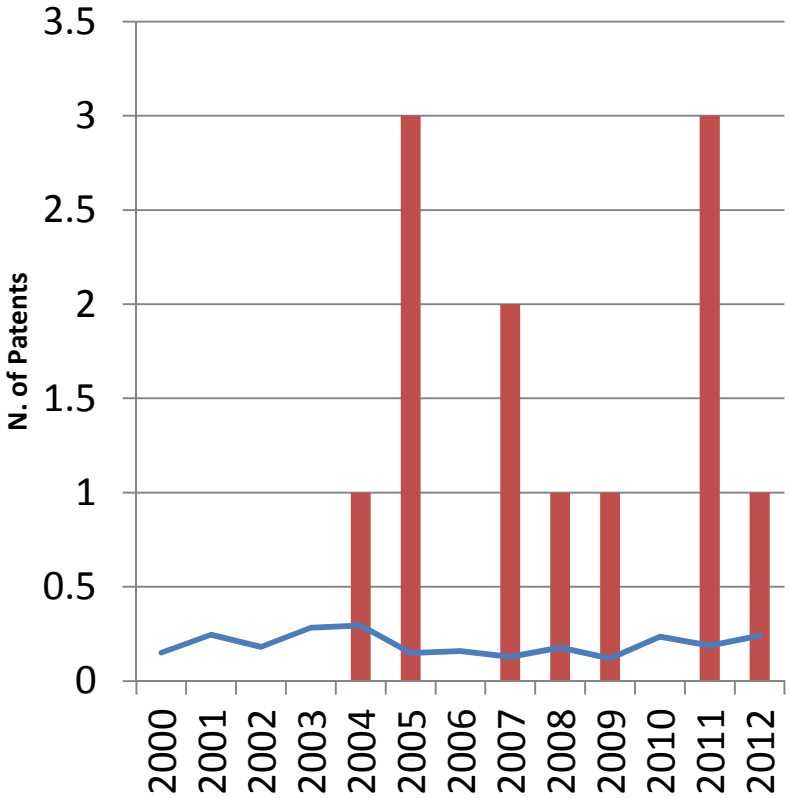


blue solid line is the average per PI

Number of Patents

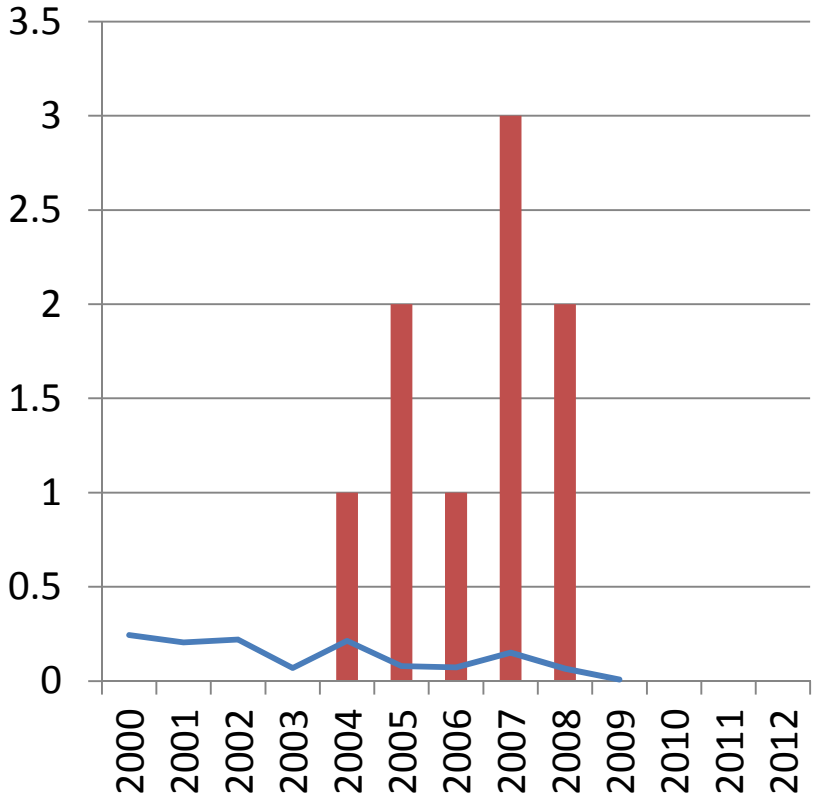
PI «J» does not have patents

PI «H» (USPTO patents*)



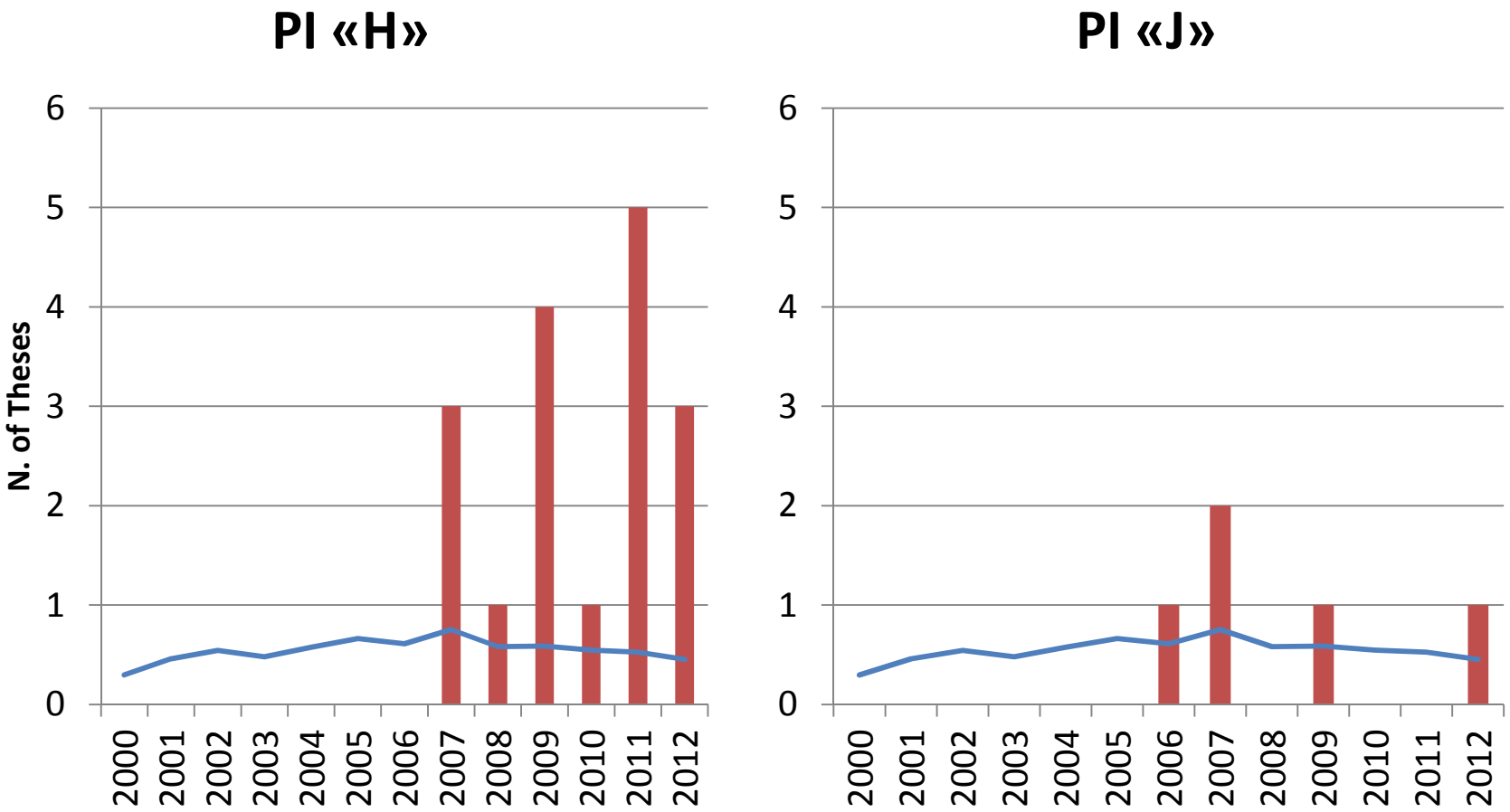
* patents by granting date

PI «H» (EPO patents)**

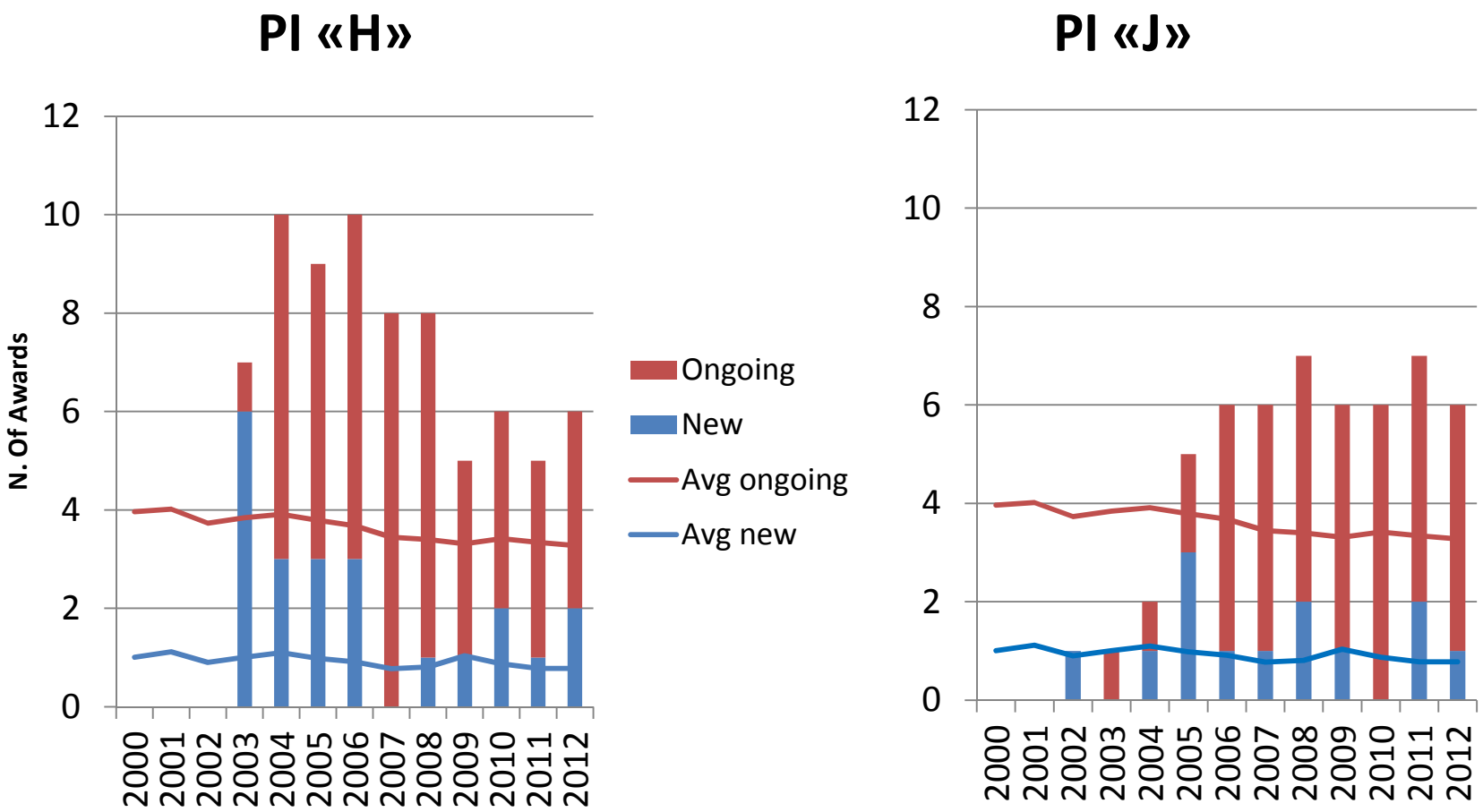


** patents by filing date

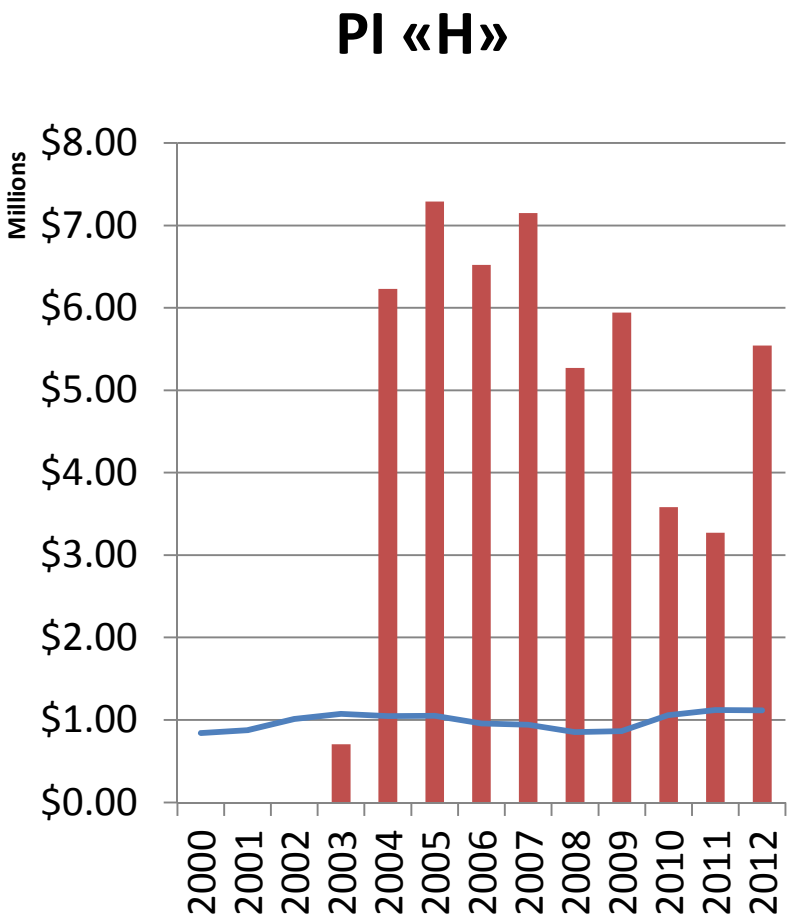
Number of Ph.D. Theses Supervised

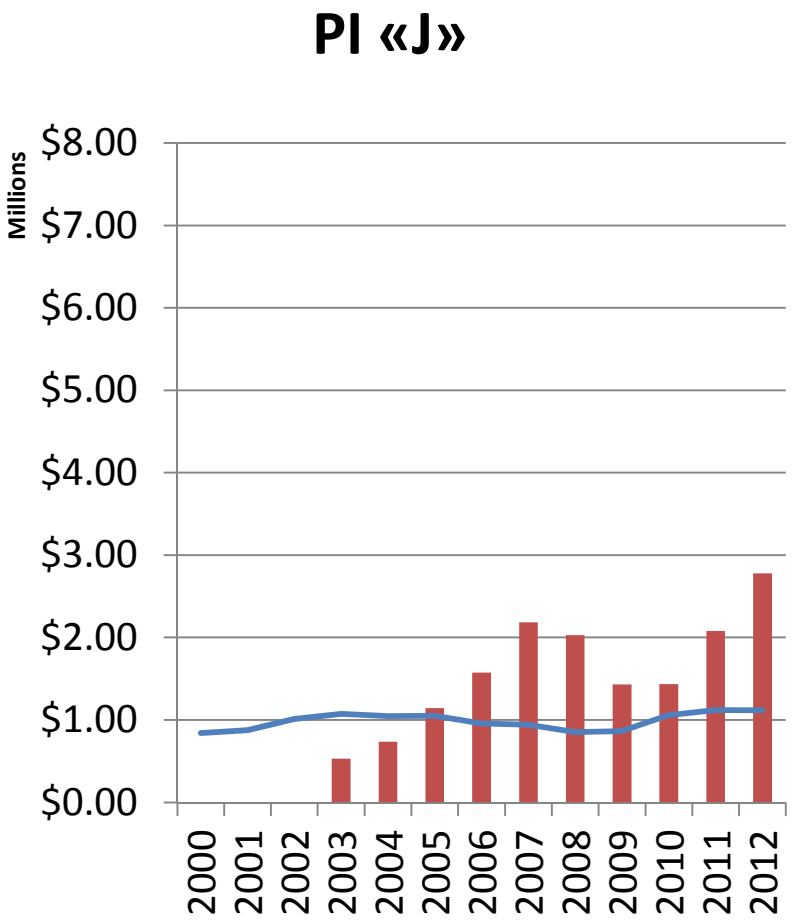


Number of ongoing and new awards

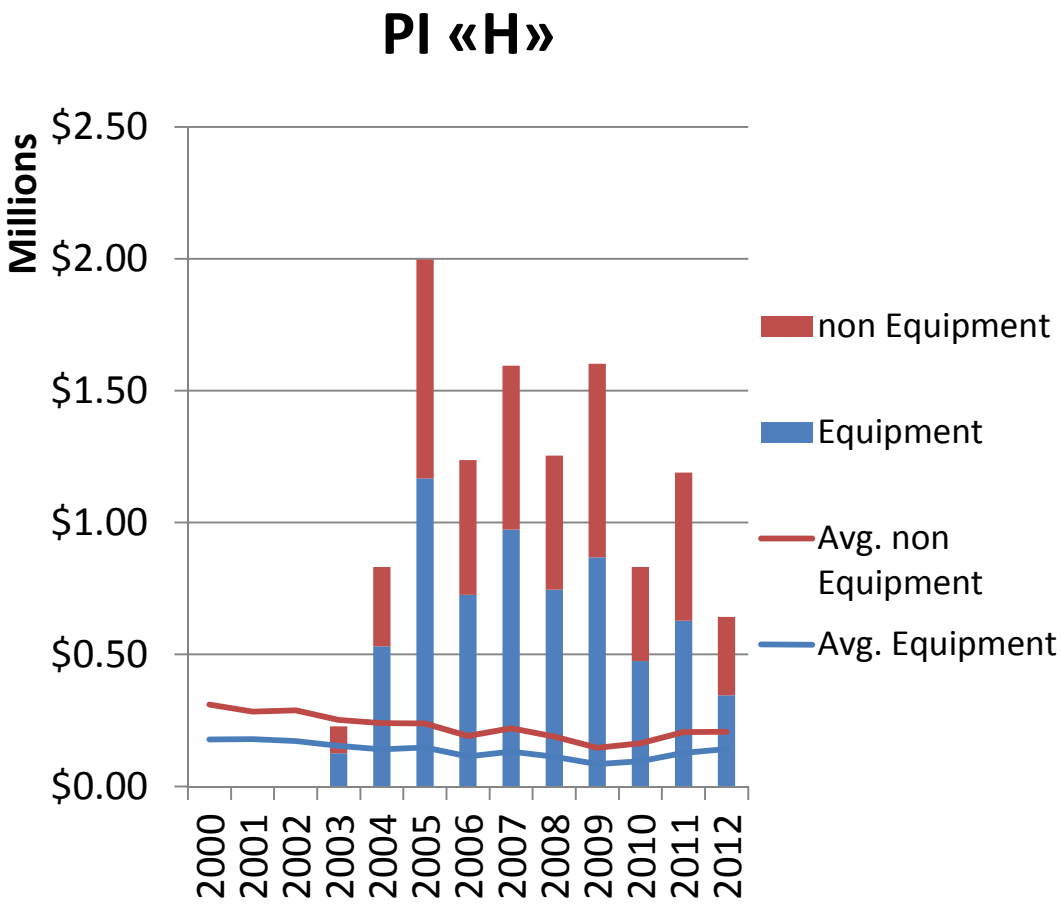


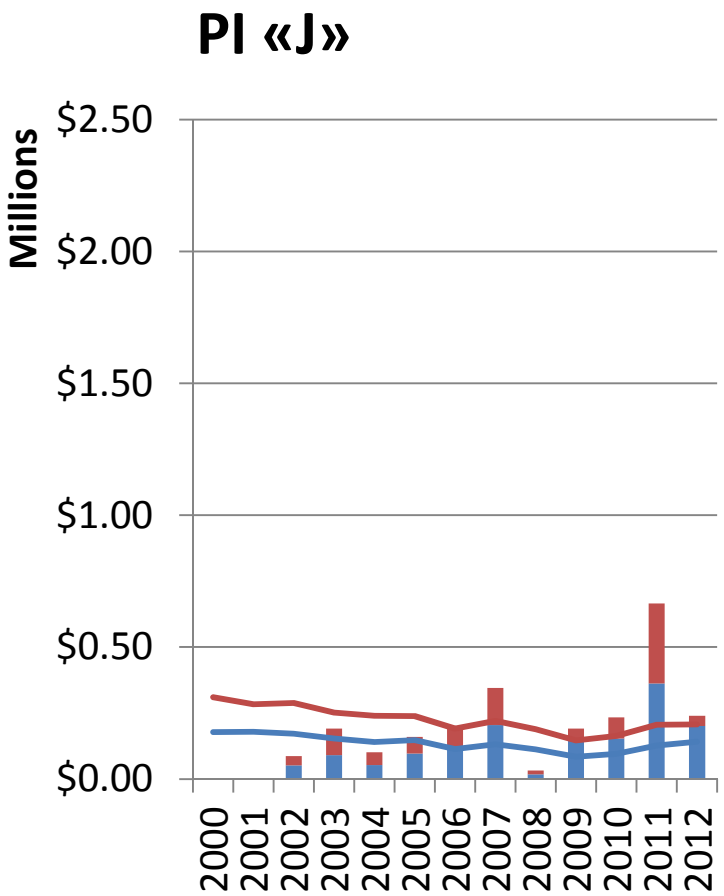
Flow of awards per year (millions)





Flow of Expenditures (millions)





Statistics (2)

PhD university

	N	mean	sd	Q1	Q2	Q3
Caltech	2590	0.12	0.33	0	0	0
MIT	2590	0.11	0.31	0	0	0
Harvard	2590	0.06	0.24	0	0	0
Princeton	2590	0.07	0.26	0	0	0
Stanford	2590	0.09	0.29	0	0	0
Berkeley	2590	0.07	0.26	0	0	0
Other institutions	2590	0.48	0.50	0	0	1
Total		1.00				

PhD field

	N	mean	sd	Q1	Q2	Q3
Biology	2590	0.06	0.24	0	0	0
Physics	2590	0.17	0.38	0	0	0
Chemistry	2590	0.07	0.26	0	0	0
Engineering	2590	0.15	0.36	0	0	0
Mathematics	2590	0.06	0.24	0	0	0
Other topics	2590	0.48	0.50	0	0	1
Total		1.00				

Outline

- Current State: Economics of Science
  - What we know
  - How we know it
  - What we don't know
  - The potential
- Caltech Etoile Project
  - Framework
  - Data Description
  - Empirical Analysis

Exploratory Regressions

$$\log(1+Y) =$$

$$\alpha \log(1+L) + \gamma_1 \left(\frac{L\ postdoc}{L}\right) + \gamma_2 \left(\frac{L\ PhD}{L}\right)$$

$$+\beta \log(1+K) + \delta_1 \left(\frac{K\ Computer}{K}\right) + \delta_2 \left(\frac{K\ Optics}{K}\right)$$

$$+ \sum_{t=2000}^{2012} \varphi_t\, T_t + \sum_{u=1}^{6} \varphi_u\, U_u + \sum_{f=1}^{5} \varphi_f\, F_f + \sigma\ Seniority$$

$$+\ldots+Const + \varepsilon$$

Labor productivity regression

	(1)	(2)	(3)
	log(1+pub)	log(1+pat)	log(1+phd)
log(1+Labor)	0.21***	0.056***	0.14***
Share Post-Doc	0.62***	-0.027	-
Share PhD	0.29*	0.034	-
Calendar year dummies	yes	yes	yes
Constant	0.060	-0.018	-0.58
Observations	2,590	2,590	2,590
R-squared	0.240	0.053	0.147
Robust standard errors			
*** p<0.01, ** p<0.05, * p<0.1			

Equipment productivity regression

	(1)	(2)	(3)
	log(1+pub)	log(1+pat)	log(1+phd)
log(1+Equipment)	0.066***	0.013***	0.033***
Share Computer	0.079	0.064	-0.24
Share Optics	0.33	0.71**	0.45**
Calendar year dummies	yes	yes	yes
Constant	0.12	-0.0058	-0.030
Observations	2,590	2,590	2,590
R-squared	0.201	0.050	0.105
Robust standard errors			
*** p<0.01, ** p<0.05, * p<0.1			

Labor and Equipment productivity regression

	(1)	(2)	(3)
	log(1+pub)	log(1+pat)	log(1+phd)
Labor			
Log(1+Labor)	0.18***	0.051***	0.13***
Share Post-Doc	0.57**	-0.041	-
Share PhD	0.27*	0.013	-
Equipment			
Log(1+Equipment)	0.015	0.0017	0.0041
Share Computer	-0.11	0.00089	-0.43
Share Optics	0.36	0.74**	0.53**
Calendar year dummies	yes	yes	yes
Constant	0.025	-0.023	-0.070*
Observations	2,590	2,590	2,590
R-squared	0.243	0.065	0.153
Robust standard errors			
*** p<0.01, ** p<0.05, * p<0.1			

Conclusion

- Results extremely preliminary
- Data just being cleaned
- But excited about the possibilities the Etoile and more generally Star Metrics projects have for expanding our understanding of science