Wrapping it up in a person: Supplementary Online Materials

Data

Our work utilizes three main sources of data: (1) new administrative data that provide detail on the structure and interactions of project teams (UMETRICS data), (2) data on researcher employment from confidential administrative and survey data housed at the U.S. Census Bureau and (3) the ProQuest Dissertation and Theses Database.

UMETRICS: The new source of administrative data is the enhanced STAR METRICS data, or UMETRICS data. We use UMETRICS data from 8 major universities (Indiana, Iowa, Michigan, Minnesota, Ohio State, Purdue, Penn State, and Wisconsin), provided as a result of a collaboration with the Committee on Institutional Cooperation (the CIC includes the 14 Big 10 Universities and the University of Chicago), which have been enhanced by identifiers to permit linkages to other datasets. The data do not, of course, cover the universe of all data. However, these 8 CIC institutions account for more than 10% of federal university R&D expenditures.

Both federal and nonfederal funding is covered in the data. The Catalog of Federal Domestic Assistance (CFDA), which is included in each award identifier, provides a full listing of all Federal programs available to universities (and other types of organizations) and is captured in the UMETRICS data to be able to filter federal award expenditures by federal funding agency. The participating institutions also generated pseudo-CFDA codes to capture non-federal sources of funding[1]; the bulk of these come from either private foundations or funding from the university's home state.

Data on the full team of researchers supported on each research grant are captured in the UMETRICS research institution data(1). This coverage is possible because the data are drawn directly from payroll records. The data also permit the capture of much more detailed information on time allocation and on the interactions of all staff on projects. The UMETRICS data of interest here are in the file of payroll transactions, which include the occupational classifications of the payees.

The challenges associated with classifying occupations are well known in survey research(*2*). It is difficult to create occupational taxonomies, difficult to train field interviewers, and even more difficult to elicit good answers from respondents. The set of challenges with administrative data are different, but equally difficult. Each university has idiosyncratic occupational classifications. The files include job titles for each individual, which were manually mapped to a standardized set of the following occupational categories: Faculty, Post Doctoral Researcher, Graduate Student, Undergraduate, Staff and Other[2].

---

[1] Details on the non-federal funding sources are provided here: https://www.starmetrics.nih.gov/static-2-1-0/Content/Downloads/Other-Funding-Source-(OFS)-Codes.xls

[2] More details on that mapping are included in Lane et al.(*9*)

Census Bureau data: Placement and earnings are derived from a match of UMETRICS data to the data at the US Census Bureau. These data have been provided to the Census Bureau in order for a Protected Identification Key (PIK), Census's internal individual identifier, to be assigned based on the employing university, the employee last name, first name, and (in some cases) date of birth. The Census Bureau's Person Identification Validation System (PVS) is used to assign an anonymous, unique person identifier to university employees(*3*). UMETRICS employee name, address, and date of birth when available are parsed, standardized and geocoded during the input process for the PVS. Next, a probabilistic match is performed between the UMETRICS data and PVS reference files that are based on the Social Security Administration's Numerical Identification File (Numident). When possible PVS assigns this person identifier, the protected identification key (PIK). Because PVS is a probabilistic match, it is possible for a UMETRICS employee to receive multiple PIK values. UMETRICS employee data is historic and spans multiple years. Thus, a custom PVS process with many years of associated reference files for each university is used. For detailed information about reference files in PVS or the matching algorithm, see (*3*)

Not all universities provide employee date of birth, resulting in higher rates of multiple PIKs than when date of birth is present. A filter is applied to all university employee PIKs in order to select the correct PIK from the multiple values when possible as well as to screen false one-to-one matches. W-2 data used for the filter is limited to records for the years that the university employee data spans, the EIN(s) associated with the university, and addresses within a 200 mile radius of the university campus address. A match to the W-2 data must occur for that employee to be retained in the sample. For multiple PIK values, only the PIK that appears in the W-2 data is retained for the employee. Filtered data are output to employee crosswalk data file.

We examine the potential biases associated with the linkage algorithm. Past work on the PVS match to the 2009 American Community Survey identified biases primarily in matching young children, minorities, residents of group quarters, immigrants, recent movers, low-income individuals, and non-employed individuals.(4). Some 20% of the doctoral recipients are not matched. This can be for several reasons: (i) the recipient does not have a job in the US – either for family reasons or because she goes back to his or her home country or (ii) she starts up a business rather than chooses employment or (iii) it is not possible to uniquely match her to a PIK. In the first case, we are examining matches to Census data to identify family reasons, although we can not trace exits from the US. In the second case, we are working on (and encouraging other researchers) to do work that examines the entrepreneurship activity of doctoral recipients. The last case can occur for those universities that do not provide information, such as date of birth, to permit accurate matching; we are currently investigating the potential resulting bias. The individuals from the universities that provided names and dates of birth go to slightly smaller firms (average firm size is 10% smaller) and they are less likely to be within 50 miles (10.1% versus 13.1% in the full sample) and less likely to remain in-state (15.9% versus 20.9% in the full sample), but there are few other differences. These differences may be due in part to differences between the universities that did and did not provide dates of birth. We are currently investigating the resulting bias in terms of demographics, but work on name matching on publication data suggests there are likely biases for names of Asian origin(5).

Once our data have been PIKized, they can be matched via a PIK-EIN (employer ID number) cross-walk sourced from W-2 and/or LEHD (Longitudinal Employer Household Dynamics) information to the Census Business Register (BR), the Longitudinal Business Database (LBD), and the Integrated Longitudinal Business Database (iLBD) to track the outcomes of the grant recipients and the location, characteristics, and performance of the firms they work for.

SOM Figure 1 provides a schematic of these data and the links between them.

The BR consists of the universe of U.S. non-agricultural businesses and is the frame underlying all other Census business data.[3] The LBD and the iLBD are longitudinally linked, edited and enhanced employer and non-employer versions of the BR respectively. They provide a longitudinal database that allows us to track firm performance, births and deaths over time. It combines administrative records and survey-based data for all nonfarm employer and non-employer business units in the United States and hence provides information about the dynamics of firm growth and firm entry/exit.(4) Key data elements include industry classification, geographic data, employment measures, payroll, and firm age.[4] Our focus is on employment, rather than entrepreneurship, so we draw data on industry, geography, firm age, receipts and employment from the firms in which the doctoral recipients find their first job. These data are quite granular. For instance, it is possible to identify the specific establishments at which people work and classify establishments into 1065 6-digit North American Industrial Classification System (NAICS) industries; for confidentiality reasons, such detail cannot be reported in this paper.

A subsample of the BR includes R&D performing firms. These are firms that report non-zero expenditures in R&D in any given year between 1976 and 2012. The firm identifiers and R&D expenditures are collected from two separate surveys collected over two separate time periods. The R&D data from 1976 to 2007 are collected from the Survey of Industrial Research and Development (SIRD) and the R&D data from 2008 until 2012 are collected from the updated version of this survey called the Business Research and Development and Innovation Survey (BRDIS)[5]. Both surveys are jointly administered by the US Census Bureau and the National Science Foundation and represent a national sample of firms beginning in 1992. All firms that report conducting R&D in one year are retained to the next year, with additional firms sampled (based on survey weights). R&D performing firms make up a small share of all firms in the United States. Of the 5M+ firms in existence in the United States in 2012, fewer than 12,500

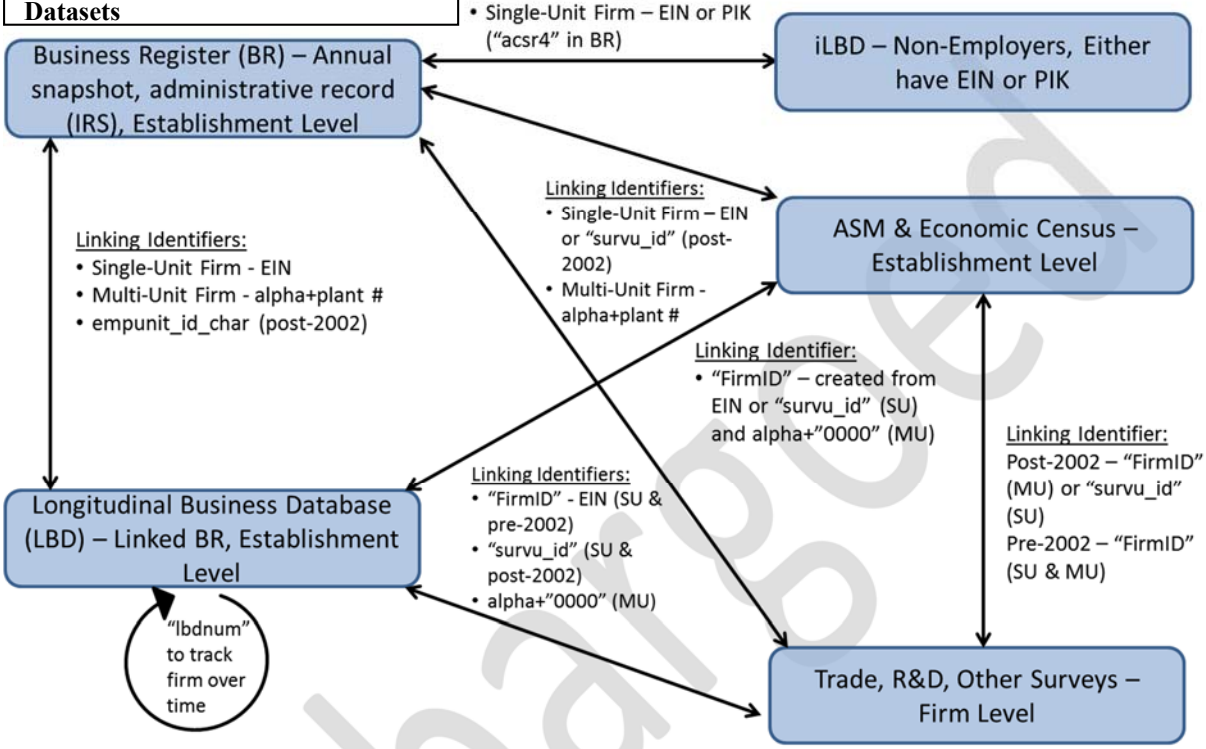
[3] The key source data elements in the Business Register are (i) the SS-4, by which a new business tells the IRS whether it is beginning as a sole proprietorship, partnership, corporation, or personal service corporation; the State or foreign country in which it is incorporated; and whether it is applying because it is a new entity, has hired employees, has purchased a going business, or has changed type of organization (specifying the type) and (ii) the 1120S K-1 series which provides information on corporate shareholders.(*10*)

[4] Non-employer businesses, which constitute the majority of businesses in the United State (although only 4% of sales and receipts), have no paid employees. Our ability to track business from the non-employer to the employer stage allows us to identify startups that may not succeed as well as the transition path.

[5] https://bhs.econ.census.gov/bhs/brdis/about.html

report conducting R&D. These firms are also known to significantly differ from the typical U.S. firms on a number of dimensions including being much larger and more likely to engage in international trade (*11*).

SOM Fig 1. Links to the Business Datasets



SOM Table 1 shows the variables used in the analysis, the level at which they are measured, and their sources.

Establishment Level	
Employment	Establishment comparisons weighted by total employment (LBD/BR)
Payroll per Worker	Total Payroll divided by employment (LBD/BR)
Industry	Placement in Industry, Academia, or Government, and 4-digit NAICS code (LBD/BR)
Location	Latitude and longitude of employing establishment; within state and within / outside of 50 miles of university (BR)
Firm Level	
Age	Age of firm (LBD)
R&D Status	Whether establishment is owned by an R&D-performing firm (BRDIS)
Individual Level	

Earnings	Derived for UMETRICS Doctoral Recipients only (W-2)
Research Field	Derived from ProQuest Dissertation and Theses database.

ProQuest data The last step is to match people whose job titles indicate that they are employed as graduate students to the ProQuest Dissertation and Theses database by name and degree granting university. The complete ProQuest data include many types of degrees for which dissertations are submitted. They include literature, education, chemistry, engineering, psychology, business, economics, history, philosophy, sociology, and information science. The records contain information on the name(s) of the author(s) of the dissertation, title, number of pages, abstract and subject of the dissertation, university or institutions awarding the graduate degree, the graduate degree awarded and the advisor's name, among others.

In this paper, we only consider PhD dissertations. While some names may appear multiple times in the data, this number is small, and for the purpose of statistical matching, we treat dissertation records as if they corresponded to unique individuals; during the PIK process, only unique matches are retained.

We use a Java implementation of the Fellegi-Sunter matching algorithm(*5*). For the matches we use university and first character of last name as blocking fields. For string comparisons we use the Jaro-Winkler string comparator and divide the range into four levels of similarity (*6*). After experimentation, the ranges, from high to low, are set: (0.92, 1.0), (0.86, 0.92), (0.81, 0.86), (0.0, 0.86). These choices were based on ranges developed by researchers at the Census Bureau. We refer to this as our fuzzy string comparator below.

For the UMETRICS to ProQuest linkages we use the same set of field comparisons. Below, "max grad year" refers to the last year a given employee was paid as a graduate student according to the occupational classification in the UMETRICS data.

The fields used for matching include the first name using fuzzy string comparison and the last name using fuzzy string comparison. In addition, we create a field by computing the difference between ProQuest degree year and UMETRICS "max grad year" and dividing the range into three levels: (i) Level 2: difference is equal to 0 or 1 (ii) Level 1: difference is equal to 2, 3, or -1 and (iii) Level 0: otherwise. The comparison is asymmetric because it should be more likely that a graduate student is last paid by a federal award before their graduation date than after.

The m-weights and u-weights in the Fellegi-Sunter algorithm were initially fit to the data using the EM algorithm for unlabeled data. They were then manually adjusted to improve the separation of perceived matches from nonmatches. In the Fellegi-Sunter algorithm, a cutoff value determines which record pairs are considered matches. In theory the cutoff value can be set to bound either the rate of false positive matches or the rate of false negative matches. In practice, however, the error rates predicted by the Fellegi-Sunter model are not generally reliable and we rely on judgment to set the final cutoff value, after examining preliminary model output to tradeoff between each type of errors(*7*). In other words, while the Fellegi-Sunter model succeeds at sorting record pairs according to the likelihood of comprising a match, the predicted likelihoods themselves are not generally accurate. For our final set of record linkage parameters,

we sorted the comparison outcomes by match score, and judged that likely matches received a match score above 6.5. Therefore, we used this as our cutoff value and flagged pairs of records with a match score greater than 6.5 as candidate links.

Because the Fellegi-Sunter algorithm does not necessarily produce a set of 1-to-1 links, we apply the Hungarian algorithm for linear sum assignment to extract a final set of 1-to-1 links from the candidate links that would maximize the total matching score(*8*).

Final Sample

SOM Table 2A shows the size of the sample before and after each step. There were 54,869 individuals paid by research grants at the 8 universities during 2009-2011, including faculty, postdoctoral researchers, staff scientists, undergraduate students and graduate students. Of these, we were able to assign PIKs to 41,794 individuals. Of those, 25,673 left the universities in the subsequent two years and were matched to the LBD. The focus of this study is recipients of doctoral degrees; of these university leavers, 3,197 individuals are matched to their dissertations in the ProQuest Dissertations & Theses database.

SOM Table 2A. Frame – Pooling data for all Universities with successive steps.

Year	Total		Total		Total		Total
2010	13,068	→ PIK-ize	10,126	→ Matched to LBD	5,943	→ Doctoral Recipients	919
2011	19,323		14,658		9,188		1,210
2012	22,478		17,010		10,542		1,068
Sum	54,869		41,794		25,673		3,197

SOM Table 2B shows the breakdown of support on federal and non-federal sources. (The total exceeds the total of 3,197 in SOM Table 2A because individuals can be on different grants at different time periods.)

SOM Table 2B: Number of doctoral recipients on research grants by exit year			
Year+1	Federal Grants	Non-Federal Grants	Both Simultaneously
2010	721	326	130
2011	996	362	153
2012	907	300	144
Total	2624	988	427

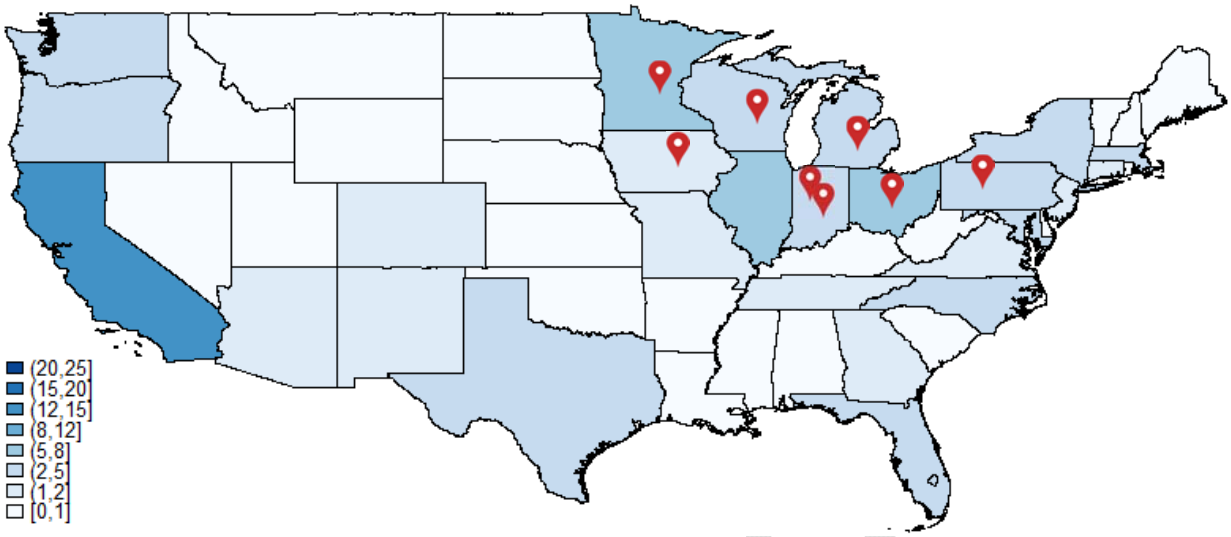
SOM Table 3 shows that the placement of federally supported doctoral recipients is similar in terms of broad industry and geography to doctoral recipients supported on all funded resarch projects (reported in Table 1).

SOM Table 3: Doctoral Recipients paid on federal research grants enter Industry; Many Stay in Local Communities.

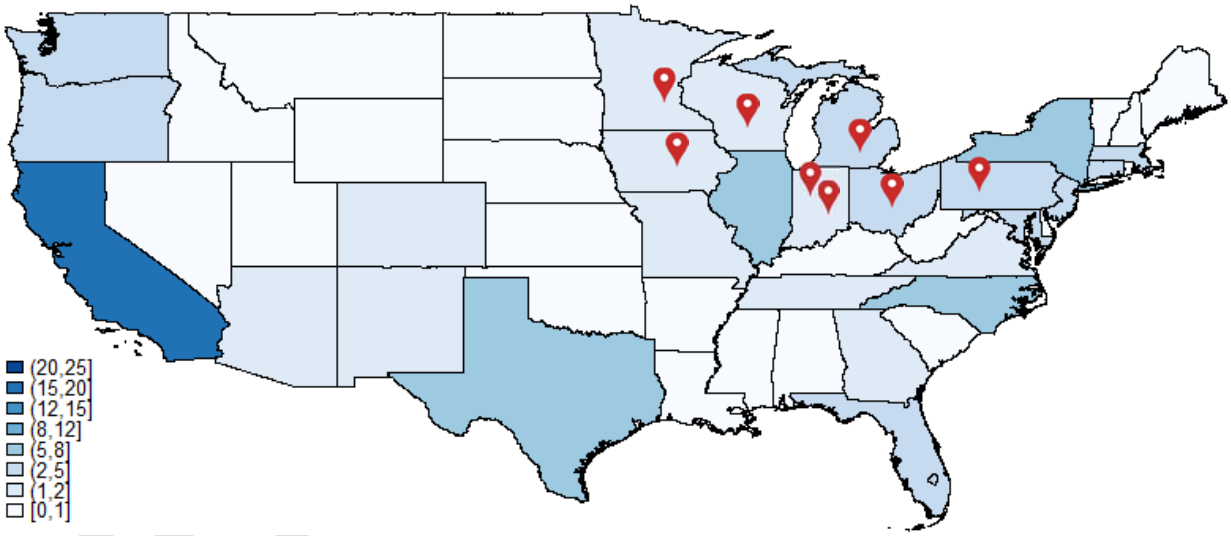
	Industry					
	R&D Firms	Non R&D Firms	Academia	Government	All	Sample Count
% Placed within Sector	17.5%	21.8%	56.7%	4.0%	100.0%	2,624
Of those in sector, percent placed:						
% Within 50 Miles	10.0%	24.7%	8.5%	17.0%	12.7%	332
% Within State	16.8%	37.1%	16.7%	25.5%	21.5%	564

SOM Figure 2. Doctoral Recipients are placed nationally, but retain regional ties.

All Doctoral Recipients



Doctoral Recipients that Move Out of State



The figure shows the share of doctoral recipients employed in each state 1 year after degree completion. The state of each university is indicated by a red flag.

SOM Table 4 shows the distribution of doctoral recipients by state of employment in the year after degree completion. Sorted by Share of Doctoral Recipients going to the state among recipients that leave the state in which their university is located. Green text indicates higher share relative to US Population share, while red text indicates lower share relative to US Population.

SOM Table 4: Distribution of doctoral recipients by state of employment				
State	Share of			
	Sample Doctoral Recipients	Doctoral Recipients that leave the state in which their university is located	National R&D Expenditures	US Population
California	14.35	19.00	26.27	12.11
Illinois	5.75	7.31	4.20	4.12
New York	4.77	6.09	3.99	6.24
Texas	4.24	5.37	5.11	8.27
North Carolina	4.21	5.10	2.09	3.10
Massachusetts	3.33	4.11	5.57	2.12
Pennsylvania	4.21	3.93	3.19	4.08
Washington	2.95	3.56	4.87	2.19
Oregon	2.73	3.29	1.64	1.24
Florida	2.79	3.16	1.88	6.13
Michigan	4.52	2.39	4.79	3.16
Maryland	2.04	2.26	1.53	1.87
Ohio	5.24	2.26	2.47	3.69
Connecticut	1.76	2.17	2.49	1.15
New Jersey	1.76	2.08	4.98	2.83
Tennessee	1.57	1.99	0.48	2.06
New Mexico	1.54	1.85	0.15	0.67
Virginia	1.60	1.85	1.73	2.60
Indiana	4.87	1.76	2.05	2.09
Minnesota	6.75	1.49	2.08	1.71
Arizona	1.13	1.44	1.66	2.08
Colorado	1.32	1.44	1.41	1.65
Georgia	1.22	1.44	1.29	3.16
District of Columbia	1.16	1.35	0.11	0.20
Missouri	1.19	1.26	2.44	1.92
Iowa	1.76	1.08	0.69	0.98
Wisconsin	2.95	1.08	1.37	1.83
Kentucky	0.75	0.99	0.39	1.40
Kansas	0.75	0.95	0.60	0.92
Utah	0.69	0.90	0.77	0.91
Idaho	0.53	0.72	0.38	0.51
Alabama	0.50	0.59	0.53	1.54
South Carolina	0.50	0.54	0.51	1.50
Vermont	0.47	0.54	0.14	0.20

Nebraska	**0.35**	**0.45**	**0.20**	0.59
West Virginia	**0.38**	**0.45**	**0.09**	0.59
Maine	**0.25**	**0.36**	**0.10**	0.42
Oklahoma	**0.25**	**0.36**	**0.18**	1.22
Delaware	**0.25**	**0.32**	**0.76**	0.29
Louisiana	**0.38**	**0.32**	**0.14**	1.47
Montana	**0.31**	**0.32**	**0.04**	0.32
Rhode Island	**0.28**	**0.32**	**0.17**	0.34
Hawaii	**0.22**	**0.27**	**0.07**	0.44
South Dakota	**0.31**	**0.27**	**0.04**	0.26
Wyoming	**0.22**	**0.27**	**0.01**	0.18
Alaska	**0.19**	**0.23**	**0.02**	0.23
Mississippi	**0.19**	**0.23**	**0.09**	0.95
North Dakota	**0.13**	**0.18**	**0.08**	0.22
Arkansas	**0.22**	**0.14**	**0.11**	0.94
New Hampshire	**0.13**	**0.14**	**0.66**	0.42
Nevada	**0.03**	**0.05**	**0.21**	0.88

Note: The State shares of national R&D expenditures are calculated as the average of 2011 and 2012 state R&D expenditures shares provided in the InfoBriefs from NCSES (NSF 13-335 & NSF 15-303). Figures from 2011 are provided here: http://www.nsf.gov/statistics/infbrief/nsf13335/nsf13335.pdf and Figures from 2012 are provided here: http://www.nsf.gov/statistics/2015/nsf15303/nsf15303.pdf . Population estimates from publicly available Census tabulations.

SOM Table 5. Doctoral Recipients Enter High Technology Industries.

Most over represented industries

Industry Description (4 digit NAICS codes)	All U.S. Employers	Doctoral Recipients	Difference
Electrical and Electronic Goods Merchant Wholesalers	0.43%	6.70%	6.27%
Architectural, Engineering, and Related Services	1.13%	5.33%	4.20%
Pharmaceutical and Medicine Manufacturing	0.22%	4.04%	3.82%
Semiconductor and Other Electronic Component	0.25%	3.71%	3.46%
Computer Systems Design and Related Services	1.30%	4.68%	3.38%
Management of Companies and Enterprises	2.64%	5.73%	3.09%
Navigational, Measuring, Electromedical, and Control Instruments	0.34%	3.07%	2.72%
Software Publishers	0.32%	2.74%	2.43%
Basic Chemical Manufacturing	0.13%	2.42%	2.29%
Specialty (except Psychiatric and Substance Abuse) Hospitals	0.24%	2.26%	2.02%

Least Represented Industries

Industry Description (4 digit NAICS codes)	All U.S. Employers	Doctoral Recipients	Difference
Full-Service Restaurants	4.03%	1.21%	-2.82%
Limited-Service Eating Places	3.63%	1.05%	-2.58%
Grocery Stores	2.26%	0.40%	-1.86%
Traveler Accommodation	1.66%	0.08%	-1.58%
Depository Credit Intermediation	1.80%	0.32%	-1.48%
Nursing Care Facilities	1.46%	0.00%	-1.46%
Building Equipment Contractors	1.39%	0.08%	-1.31%
Services to Buildings and Dwellings	1.46%	0.24%	-1.21%
Clothing Stores	1.20%	0.08%	-1.12%
Other General Merchandise Stores	1.51%	0.40%	-1.11%

SOM Table 6: Doctoral Recipients paid on federal research grants enter high technology industries

Most over represented industries

Industry Description (4 digit NAICS)	All US Employers	Doctoral Recipients	Difference
Electrical and Electronic Goods Merchant Wholesalers	0.43%	6.41%	5.98%
Pharmaceutical and Medicine Manufacturing	0.22%	4.27%	4.06%
Architectural, Engineering, and Related Services	1.13%	4.95%	3.82%
Computer Systems Design and Related Services	1.30%	4.85%	3.56%
Semiconductor and Other Electronic Component	0.25%	3.40%	3.15%
Navigational, Measuring, Electromedical, and Control Instruments	0.34%	3.20%	2.86%
Software Publishers	0.32%	3.01%	2.69%
Basic Chemical Manufacturing	0.13%	2.72%	2.59%
Management of Companies and Enterprises	2.64%	4.95%	2.31%
Specialty (except Psychiatric and Substance Abuse) Hospitals	0.24%	2.43%	2.19%

Most under represented industries

Industry Description (4 digit NAICS codes)	US	All Universities	Difference
Full-Service Restaurants	4.03%	1.17%	-2.86%
Limited-Service Eating Places	3.63%	0.87%	-2.76%
Grocery Stores	2.26%	0.39%	-1.88%
Traveler Accommodation	1.66%	0.10%	-1.56%
Nursing Care Facilities	1.46%	0.00%	-1.46%
Depository Credit Intermediation	1.80%	0.39%	-1.41%
Building Equipment Contractors	1.39%	0.10%	-1.29%
Religious Organizations	1.47%	0.19%	-1.27%
Services to Buildings and Dwellings	1.46%	0.19%	-1.26%
Other Specialty Food Stores	1.51%	0.39%	-1.12%

SOM Table 7: The Earnings and Placement of Doctoral Recipients Supported on Grants Vary By Field

This table provides the exact means and standard deviations used in generating Figure 3.

	Earnings		Earnings in Industry		P(Industry Placement)		P(R&D Firm Placement)		P(High Wage Establishment Placement)		P(Young Establishment Placement)	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Arts & Humanities	28,172.95	22,519.66	30,424.72	31,864.37	0.21	0.41	0.06	0.24	0.51	0.50	0.02	0.15
Biology	36,829.92	33,262.90	47,622.40	57,109.14	0.26	0.44	0.04	0.21	0.83	0.38	0.04	0.20
Chemistry	42,817.82	27,515.30	58,100.27	35,941.28	0.35	0.48	0.14	0.35	0.83	0.37	0.05	0.22
Education	40,040.81	24,746.20	35,494.32	36,882.63	0.15	0.36	0.03	0.16	0.52	0.50	0.04	0.21
Engineering	68,036.53	39,404.79	79,010.55	40,056.19	0.60	0.49	0.38	0.49	0.90	0.30	0.07	0.26
Health	42,416.12	31,591.94	48,622.40	37,275.74	0.39	0.49	0.07	0.25	0.79	0.41	0.03	0.18
Math & Comp. Science	65,258.88	59,610.98	87,192.90	78,425.39	0.44	0.50	0.19	0.39	0.83	0.38	0.07	0.26
Other	46,618.20	50,754.83	69,688.95	76,417.13	0.34	0.48	0.14	0.35	0.72	0.45	0.03	0.16
Other Science	43,379.82	33,765.97	53,092.73	41,822.79	0.34	0.48	0.16	0.37	0.81	0.40	0.08	0.28
Physics	54,728.56	29,076.19	71,903.18	31,622.44	0.34	0.47	0.16	0.36	0.77	0.42	0.07	0.26
Social Science	43,394.55	34,957.80	43,216.83	40,457.41	0.26	0.44	0.04	0.21	0.69	0.46	0.04	0.19

SOM Table 8 provides the detailed regression results supporting Figure 3.

	Earnings	Probability of Placement in								Observations
		Industry		Estab. of R&D Firm		High Wage Estab.		Young Firm		
	Regression	Regression	Marginal Effect	Regression	Marginal Effect	Regression	Marginal Effect	Regression	Marginal Effect	
Arts & Humanities: Reference Group										85
Biology	3651.8	0.198	0.060	-0.0977	-0.00	0.934***	0.321***	0.214	0.015	473
	(2830.6)	(0.170)	(0.049)	(0.258)	(0.024)	(0.153)	(0.057)	(0.312)	(0.019)	
Chemistry	4837.0	0.429*	0.141**	0.526*	0.078**	0.935***	0.321***	0.285	0.021	387
	(2715.2)	(0.171)	(0.051)	(0.249)	(0.028)	(0.156)	(0.057)	(0.314)	(0.020)	
Education	13161.6***	-0.251	-0.06	-0.387	-0.02	0.0838	0.033	0.248	0.018	183
	(3130.5)	(0.194)	(0.051)	(0.302)	(0.024)	(0.165)	(0.065)	(0.337)	(0.022)	
Engineering	17530.2***	1.071***	0.392***	1.302***	0.308***	1.247***	0.388***	0.489	0.045*	801
	(3022.2)	(0.164)	(0.048)	(0.241)	(0.029)	(0.149)	(0.055)	(0.304)	(0.019)	
Health	7954.7*	0.522**	0.176***	0.121	0.013	0.808***	0.287***	0.108	0.007	242
	(3283.1)	(0.178)	(0.054)	(0.269)	(0.027)	(0.165)	(0.060)	(0.335)	(0.020)	
Math, Comp. Science & Statistics	21651.5***	0.716***	0.252***	0.800**	0.144***	0.933***	0.320***	0.434	0.038	356
	(3432.4)	(0.172)	(0.052)	(0.249)	(0.031)	(0.158)	(0.058)	(0.313)	(0.021)	
Other	10470.0*	0.352	0.113	0.488	0.071	0.564**	0.212**	0.00450	0.000	111
	(5199.1)	(0.199)	(0.062)	(0.278)	(0.037)	(0.187)	(0.069)	(0.388)	(0.022)	
Other Science	5537.5	0.388	0.126	0.593*	0.093*	0.830***	0.293***	0.561	0.055	98
	(3786.5)	(0.206)	(0.065)	(0.282)	(0.041)	(0.200)	(0.067)	(0.352)	(0.033)	
Physics	15574.4***	0.401*	0.131*	0.601*	0.094**	0.700***	0.255***	0.486	0.045	167
	(3326.3)	(0.188)	(0.058)	(0.266)	(0.035)	(0.174)	(0.063)	(0.330)	(0.026)	
	(3309.2)	(0.176)	(0.051)	(0.269)	(0.025)	(0.156)	(0.060)	(0.328)	(0.020)	
University FE	Yes	Yes		Yes		Yes		Yes		
Year FE	Yes	Yes		Yes		Yes		Yes		
Industry FE	Yes	**		**		**		**		
Observations	3197									
R-squared	0.345	0.078		0.180		0.069		0.028		

Standard errors in parentheses ="* p<0.05 ** p<0.01 *** p<0.001"; Marginal effects are calculated relative to the mean;
High Wage Firm defined as having higher average wage than the Median establishment within six digit Industry-Year

References

1. J. King, J. Lane, L. Schwarz, Creating New Administrative Data to Describe the Scientific Workforce: The Star Metrics Program. *SSRN eLibrary* (2013), (available at http://ssrn.com/paper=2085187).

2. W. Mellow, H. Sider, Accuracy of response in labor market surveys: Evidence and implications. *J. Labor Econ.*, 331–344 (1983).

3. D. Wagner, M. Layne, "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software" (2014), (available at https://www.census.gov/srd/carra/CARRA_PVS_Record_Linkage.pdf).

4. S. J. Davis *et al.*, "Measuring the dynamics of young and small businesses: Integrating the employer and nonemployer universes" (National Bureau of Economic Research, 2007).

5. I. P. Fellegi, A. B. Sunter, A Theory for Record Linkage. *J. Am. Stat. Assoc.* **64**, 1183–1210. (1969).

6. W. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkag. *Proc. Sect. Surv. Res. Methods, Am. Stat. Assocation* (1990).

7. W. Winkler, Automatic Estimation of Record Linkage False Match Rates. *Proc. Sect. Surv. Res. Methods, Am. Stat. Assoc.* (2006).

8. H. Kuhn, The Hungarian Method for the assignment problem. *Nav. Res. Logist. Q.* **2**, 83–97 (1955).

9. J. I. Lane, J. Owen-Smith, R. F. Rosen, B. A. Weinberg, New linked data on research investments: Scientific workforce, productivity, and public value. *Res. Policy* (2015), doi:10.1016/j.respol.2014.12.013.

10. N. Greenia, K. Husbands Fealing, J. Lane, "Studying Innovation In Businesses: New Research Possibilities" (2008), (available at http://www.irs.gov/pub/irs-soi/08rpinnovbusgreenia.pdf).

11. L. Foster, C. Grim, N. Zolas, "A Portrait of Firms that Invest in R&D" (2015), U.S. Census Bureau Mimeo