

IZA DP No. 6600

**Creating New Administrative Data to
Describe the Scientific Workforce:
The STAR METRICS Program**

Julia Lane
Lou Schwarz

May 2012

Creating New Administrative Data to Describe the Scientific Workforce: The STAR METRICS Program

Julia Lane

*American Institutes for Research
and IZA*

Lou Schwarz

Factor 21

Discussion Paper No. 6600
May 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ABSTRACT

Creating New Administrative Data to Describe the Scientific Workforce: The STAR METRICS Program^{*}

In common with many countries, the substantial United States investment in R&D is characterized by limited documentation of the nature and results of those investments (Macllwain 2010, Marburger 2005). Despite the increased calls for reporting by key stakeholders, current data systems cannot meet the new requirements; indeed, the conclusion of the Science of Science Policy interagency group's Federal Research Roadmap (National Science and Technology Council 2008) was that the science policy data infrastructure was inadequate for decision-making. In response to this need, a new data system is being built (STAR METRICS) drawing from administrative records; this paper describes the initial results of that effort – focusing on documenting the scientific workforce supported by expenditures during the 2011 Federal fiscal year from awards made by the National Science Foundation. The contribution of the paper is threefold. First it describes in a non-technical fashion how these new data can contribute to our understanding of the initial results of science investments. Second, it shows how new computational technologies can be used to go beyond the traditional methods of manual reporting and administrative program coding to capture information at the most granular units of analysis possible. Finally, it discusses the lessons learned for the collection and analysis of data. The most important is leveraging existing data, not relying on surveys and manual reporting; the deficiencies of each have been well documented (Lane 2010).

JEL Classification: J24, J40, J48

Keywords: scientific workforce, administrative data, science policy, STAR METRICS

Corresponding author:

Julia Lane
American Institutes for Research
1000 Thomas Jefferson St
Washington DC 20007
USA
E-mail: jlane@air.org

^{*} STAR METRICS is a large scale data project with major contributions from many people. These include Daniel Basco, Stefano Bertuzzi, Lee Broderick, Sean Bye, David Croson, Sharon Drumm, Ian Foster, Adam Hall, Cindy Hope, John King, Kei Koizumi, Ron Lai, Mark Largent, Dorothy Miller, Maureen Miller, David Newman, Erica Rissi, Rebecca Rosen, Susie Sedwick, Dick Seligman, Toby Smith, Mya Strauss, Jenn Thornhill, Bill Valdez, Jack VInner, John Voeller, Amy Yu, Linda Yu, the key personnel at participating research institutions, and the participants of the Federal Demonstration Partnership, the American Association of Universities, the Association of Public Land-Grant Universities, the American Association of Medical Colleges.

1. INTRODUCTION

In common with many countries, the substantial United States investment in R&D is characterized by limited documentation of the nature and results of those investments (MacIlwain 2010, Marburger 2005). Despite the increased calls for reporting by key stakeholders, current data systems cannot meet the new requirements; indeed, the conclusion of the Science of Science Policy interagency group's Federal Research Roadmap (National Science and Technology Council 2008) was that the science policy data infrastructure was inadequate for decision-making. In response to this need, a new data system is being built (STAR METRICS¹) drawing from administrative records; this paper describes the initial results of that effort – focusing on expenditures during the 2011 Federal fiscal year from awards made by the National Science Foundation.

The contribution of the paper is three fold. First it describes in a non-technical fashion how these new data can contribute to our understanding of the initial results of science investments. Second, it shows how new computational technologies can be used to go beyond the traditional methods of manual reporting and administrative program coding to capture information at the most granular units of analysis possible. Finally, it discusses the lessons learned for the collection and analysis of data. The most important is leveraging existing data, not relying on surveys and manual reporting; the deficiencies of each have been well documented(Lane 2010).

This paper focusses on the use of administrative records to document the scientific workforce for three reasons. The first is pragmatic: the 2009 American Recover and Reinvestment Act mandate that recipients of funds document the jobs supported by that funding. Hence, there was an urgent need to develop consistent and auditable ways of responding to that mandate. The second is an increasing recognition by science agencies that since a major goal of research funding is to develop the next generation of scientists, it was important to develop a strong data system to provide information about how many post-graduate, graduate and undergraduate students were being supported by research awards. Finally, the scientific community recognized that without such a data system, the results of science investments will be under-reported, since many of the outcomes of research investments results from student achievements. While the example of the co-founder of Google, Sergey Brin, is the most well-known(Lane 2009), work by Lee Fleming suggests that students represent a major pathway whereby knowledge moves from bench to business.(Fleming, Charles King, and Juda 2007). Most worryingly, the lack of a systematic knowledge base about who is doing science precludes the development of a systematic understanding of scientific networks and scientific teams – despite substantial evidence that the practice of science is increasingly team-oriented.

¹ Science and Technology for America's Reinvestment – Measuring the Effect of Research on Innovation, Competitiveness, and Science

It also focusses on the use of new computational techniques, rather than human input to describe what scientists are doing. This technique, known as topic modeling, provides a powerful and flexible framework for representing, summarizing and analyzing the contents of large document collections, can be used to describe research topics - and hence describe what research is being done. The topics define a simplified representation of the documents, but the research topics are defined by the research proposal, not manual generation of taxonomies or keywords.(Blei, Ng, and Jordan 2003, Newman et al. 2007).

There are several major lessons learned from the success of this program, which is a voluntary collaboration between five of the largest civilian science and technology agencies (NIH, NSF, EPA, DOE and USDA) and some 85 research institutions, developed under the auspices of the White House Office of Science and Technology Policy. The approach is similar to that taken by one of the authors in creating the Census Bureau's Longitudinal Employer Household Dynamics (LEHD) program(Abowd, Haltiwanger, and Lane 2004) which was built into a national program over a seven year period. As with that program, the key elements are that it is voluntary, that it leverages existing data, it is low burden to participants, it builds strong relationships between the federal agencies and the researcher community, and it creates value to all participants.

2. BACKGROUND

The Science of Science Policy Interagency Working Group, a subcommittee of the National Science and Technology Council's Social, Behavioral and Economic Sciences subcommittee², summarized current agency practice in a 2008 Roadmap. The roadmap found that agencies and departments across the Federal Government face similar challenges when setting scientific priorities and assessing the effectiveness of current and planned investments. A key finding was that the current data infrastructure was inadequate for decision-making. The interagency group recommended that federal government agencies should work in concert to establish a theoretical and empirical framework to understand the science and engineering enterprise within the context of the science of science policy. It encouraged investment in the development and use of emerging tools, methods, data, and data infrastructure to enable science policy decision makers to base investment decisions on more rigorous and quantitative analyses.

² The agencies and offices represented include: Department of Energy , National Science Foundation, Centers for Disease Control and Prevention, Central Intelligence Agency, Department of Commerce, Department of Defense, Environmental Protection Agency, National Aeronautics & Space Administration, National Institutes of Health, National Institute for Standards & Technology, National Oceanic & Atmospheric Administration, Office of Management and Budget, Office of Science and Technology Policy, U.S. Department of Agriculture, U.S. Geological Survey, and the U.S. Department of Veterans Affairs.

Creating New Administrative Data

The passage of ARRA, and the focus on reporting the jobs associated with science investments, led to two agency concerns. The first was the accuracy of the approach used in the reporting process. The second was the limited nature of that reporting requirement, since science investments have been documented to have longer term impact in many areas, including scientific outcomes, such as the generation and adoption of new science, often measured by citations and patents, on economic outcomes, such as job creation, on the outcomes of the scientific workforce, as well as on social outcomes such as health and the environment.

At the same time, it was clear that continuing to require research institutions and principal investigators to manually report the outcomes of research was neither practicable nor desirable. A recent study titled "Reforming Regulation of Research Universities" provides a good summary of the challenges; it finds that poorly-integrated federal reporting and other regulations are imposing a heavy and growing administrative burden on federally-funded research. The report argues that this "regulatory overhead" is both large (and getting larger), and often inefficient, with many federal reporting requirements overlapping and even conflicting. It estimates that 42% of faculty time relating to federally-funded research is spent on administrative duties, rather than on the research itself.³

Spurred by the twin needs of responding to ARRA reporting requirements and minimizing reporting burden, OSTP staff and members of the Science of Science Policy (SOSP) Interagency Group worked with seven universities participating in the Federal Demonstration Partnership (FDP)⁴ to develop a more reliable way of responding to ARRA. The initial focus was to use university administrative records to identify all individuals supported by federal science funding in order to generate information on jobs created under ARRA. The consensus by the seven universities was that it is not difficult to create the required files, and there is substantial value added from standardized reports. The results from the pilot were presented at the FDP national meeting in September 2009 and January 2010. The pilot demonstrated the feasibility of creating standardized measures of the impact of science investments on initial job creation for pilot institutions with almost no burden on respondents. (Nelson and Sedwick 2011)

³ <http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=12330>

⁴The Federal Demonstration Partnership (FDP) is an association of federal agencies, academic research institutions with administrative, faculty and technical representation, and research policy organizations that work to streamline the administration of federally sponsored research. FDP members of all sectors cooperate in identifying, testing, and implementing new, more effective ways of managing the more than \$15 Billion in federal research grants. The goal of improving the productivity of research without compromising its stewardship has benefits for the entire nation.

Both the agencies and research institutions involved expressed interest in expanding the pilot to create a data infrastructure that facilitated common responses to Congressional and Administrative requests. That led to the formation of the current STAR METRICS Program. In less than 18 months, STAR METRICS has expanded dramatically – from 3 agencies (NIH, NSF and OSTP) to 6 and from 7 research institutions to 85. Participants appear to be responding to the need to combine forces to respond to ongoing requests from OMB, Congress and the public, as well as the potential to reduce reporting burden. The Consortium has developed a strong relationship with the Federal Demonstration Partnership (FDP) and with the American Association of Universities (AAU). Both have established working groups to help facilitate pilot projects with the 85 participating research institutions⁵. There is regular outreach at each FDP national meeting (three times a year), and at AAU meetings with Senior Research Officers. The Consortium also hosts an annual meeting with VPs for research.

There is also a tight link between STAR METRICS and the Science of Science & Innovation Policy (SciSIP) program at the National Science Foundation. Many of the innovative tools have been developed by SciSIP principal investigators, much of the data development has been funded by the SciSIP program, and the SciSIP research community has actively participated in designing many of the elements of the data platform.

3. DOCUMENTING THE SCIENTIFIC WORKFORCE

The key insight in developing the protocols necessary to document the workforce was derived from the LEHD program which is derived from quarterly Unemployment Insurance (UI) wage record reports. Almost all businesses in the United States are required to file these reports, which include, for each employee, detailed data on earnings for each reporting period. It was an open question as to whether the existing human resource systems could capture those data at a more disaggregated level: namely, who was supported by science funding, their occupations, and the proportion of their earnings that was allocated to the award. It was also an open question as to whether every financial transaction could be traced to the source of the funding for accounting purposes. The pilot project made it clear that it was indeed possible to “follow the money” and repurpose the financial records of institutions to document the employment associated with science expenditures.

There are four basic ways in which existing data on Federal S&T expenditures can be repurposed and traced to a scientific workforce. Individuals can be directly employed on a Federal grant. Some Federal funds go to support researchers at collaborating institutions. Scientific supplies are purchased from vendors. Infrastructure support, including financial, IT, physical space and research services is also provided. Each of these activities creates a financial transaction that can be used to calculate the associated employment.

⁵ The list of participating institutions is provided in Appendix I.

Creating New Administrative Data

Figure 1 provides a stylized description of flow of these financial transactions in a typical administrative system. The flow on the left hand side demonstrates how the Human Resources system in a research institution can be used to identify, on a quarterly basis, the universe of individuals (Principal Investigators (PIs), co-PIs, post-doctoral researchers graduate and undergraduate students, lab technicians, science administrators, etc.) supported by any grant, but most particularly, those from Federal science agencies. Just as the LEHD program used unemployment insurance wage records to capture the flows of workers across firms, this approach used tracks the expenditure trail generated by financial reporting requirements to capture each transaction charged to the funding source. All payroll transactions, which include the occupational classifications of the payees, can thus be used to automatically generate reports on who is paid, and how much, from each source of funding. as well as disbursements to vendors and those receiving sub-awards can be traced in the administrative records of the reporting institutions.

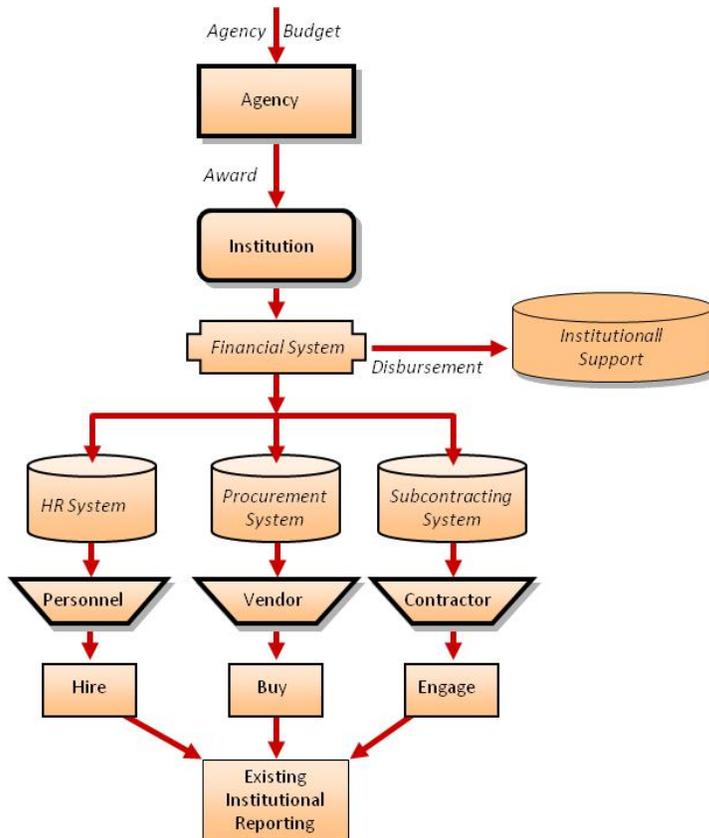


FIGURE 1: THE FLOW OF ADMINISTRATIVE TRANSACTIONS ASSOCIATED WITH FEDERAL FUNDING TO A RESEARCH INSTITUTION

Creating New Administrative Data

The second key element is to create a system that imposes minimum burden. The STAR METRICS program uses 14 data elements that can be used to calculate employment supported by Federal grants in four different categories⁶:

- Jobs supported that appear on research institution payroll
- Jobs supported by research institution spending for the purchase of goods and services
- Jobs supported by research institution spending on sub-awards
- Jobs supported by research institutions Facilities & Administration (F&A) costs, also known as indirect costs

⁶ Full details are provided at <https://starmetrics.nih.gov>.

Creating New Administrative Data

TABLE 1: DATA ELEMENTS REQUESTED FROM PARTICIPATING INSTITUTIONS

Description	Element ID	Item	Unit of Analysis
Information on Scientists and Awards	1	De-identified Employee ID #	Individual
	2	Unique Award Number	Award
	3	Recipient Account Number	Award
	4	Occupational Classification	Individual
	5	Proportion of time (or earnings) allocated to award	Individual
	6	FTE status	Individual
Information on Overhead	7	Overhead charged	Award
	8	Proportion of overhead associated with salaries (from indirect cost rate proposal)	University
Payments to Vendors	9	Recipient Account Number	Award
	10	Vendor Duns # or Zip Code	Vendor
	11	Amount of Contract disbursed in period	Vendor
Subcontracts and Subawards	12	Recipient Account Number	Award
	13	Subaward Recipient Duns # or Zip Code	Subcontractor
	14	Amount of Contract disbursed in period	Subcontractor

Each of these elements is then used to create micro level data on the workforce associated with each award. The Human Resource data (elements 1-6) are used to generate data on the actual (not estimated) number of individuals, and the number of full time equivalent (FTE) jobs paid by grants. The STAR METRICS system draws all the (de-identified) employee ID numbers corresponding to those individuals that have charged time to a given award; it is then straightforward to sum the total number of individuals per award – by occupational classification -- and the total number of individuals per institution. The FTE count is derived by multiplying the proportion of an individual's time allocated to a given award by the individual's FTE category. Each institution's HR occupational classifications are mapped to seven STAR METRICS categories, returned by the institutions for validation, and used to create the occupational breakouts.

This approach only captures the direct payroll jobs; yet science is increasingly collaborative and conducted at multiple institutions. A full picture of the scientific workforce would include scientists supported by research institution spending on subcontracts and sub-awards with

Creating New Administrative Data

collaborating institutions. As STAR METRICS expands, it will be possible to provide the full network of inter-linkages across institutions. In the short term, however, the approach used is to pull the following elements for each purchase in the quarter: the unique federal award number, the recipient account number, the sub-awardee's DUNS number⁷ and the payment amount. The STAR METRICS system then pulls the industry code and the geographic location for the company from the DUNS number. The calculation of jobs is directly derived from two pieces of US Economic Census data. The first is the proportion of earnings for the sub-awardee's industry in the sub-awardee's geographic location. The second is the average salary in that industry at that location. The expenditures on goods and services from the sub-awardee, when multiplied by the earnings proportion, generate an estimate of the average wages and salaries supported by that expenditure. That number, divided by the average salary, generates an estimate of the number of jobs supported.

More technically, the sub-awardee DUNS Number is used to derive an industry code and geographic location. Economic Census data elements are used to generate estimates of the amount of salaries paid as a result of the grant revenues earned by the sub-awardee⁸. This gross wage expenditure is in turn divided by the average earnings of workers in that industry and geographic location to generate employment estimates derived from Economic Census data. This quotient does not represent a Full-Time Equivalent (FTE) count, but rather a job count.

Financial data can also be used to capture a broader definition of the workforce, namely the jobs supported by the purchase of goods and services from vendors. This calculation is identical to the one for jobs supported by research institutions' spending on sub-awardees, as described immediately above. The only distinction is that the source of data is from vendors, rather than sub-awards. The vendor DUNS Number is used to derive an industry code and geographic location for the vendor as was done with sub-awardees.

Finally, research does not occur in a vacuum. Institutions provide lab facilities and equipment, and administrative support ranging from the grants office to financial, HR, IT and janitorial support. However, none of this is directly charged to the award in a manner that can be pulled from administrative records. What is captured is the amount of overhead charged per award. The STAR METRICS system requests four data elements: the unique federal award number, the recipient account number, the overhead charged and the proportion of overhead associated

⁷ DUNS (Data Universal Numbering System) number is a unique nine-digit identification number for each physical location of a business that is used by the Federal government and other entities as a standard business. Identifier http://en.wikipedia.org/wiki/Data_Universal_Numbering_System. If a DUNS number is not available, the institution reports a ZIP code. While in the first quarter of 2011, only about 9% of expenditures could be attributed to an entity with a DUNS number, the implementation of new reporting requirements may increase that proportion substantially.

⁸ <http://www.census.gov/econ/census07/>

Creating New Administrative Data

with wages and salaries. This last item is drawn from the indirect cost reports to the research institution's cognizant agency.

Developing any new data system creates a number of technical challenges – many of which were faced by the LEHD system in its early stages as well. First, because the program is voluntary, the initial data are not representative of either the population of research institutions or of federal awards. However, researchers at the current participating institutions receive some 40% of NSF and NIH grants, and more institutions are expected to join. There is likely to be some undercounting, since most research institutions keep detailed records of charges associated with credit cards, but do not break the expenditures out by industry, vendor, or location. These payments can run into the millions of dollars, but are not included in the administrative records analyzed here, so the estimates of the associate workforce are also likely to be under-reported. Some research institutions did not provide detailed information on the Full-Time Equivalent status of their employees. They consequently used an occupational rule of thumb: typically a faculty member was classified as a full-time staff member, while a graduate student was classified as half-time, and an undergraduate as quarter-time. Sub-awards that are received by STAR METRICS institutions that were generated from a federal funding source are typically traced to the awarding institution, not the federal grant. The STAR METRICS program will begin to account for these cross-institution transfers as more institutions join the program.

Although this is a new approach to collecting workforce data, there was fortunately a way of cross-validating the information collected by comparing the STAR METRICS FTE counts with the ARRA FTE counts that were generated by largely manual methods. There was no statistically significant difference between the FTE counts generated by the two approaches.

4. DOCUMENTING THE RESEARCH AREAS

The topic modeling approach was implemented as a result of a two year project undertaken on behalf of two NSF Advisory Committees – Computer and Information Science and Engineering (CISE) and Social, Behavioral and Economic Sciences (SBE). The Subcommittee participants were charged with identifying and demonstrating techniques and tools that could characterize a specific set of proposal and award portfolios. In addition, the subcommittee was asked to provide recommendations on ways that NSF could better structure existing data, make use of existing machine learning, analysis, and visualization techniques to complement human expertise and better characterize its programmatic data.

As a part of this work, ten research teams, each directed by a member of the subcommittee, participated in a two year experimental effort in which NSF proposals were analyzed in a secure environment, using the latest techniques for information analysis and visual analytics. The teams interacted with program managers from NSF and from other parts of the federal government in order to produce their demonstration analyses.

Creating New Administrative Data

The approach adopted by the program was topic modeling, which is one way to categorize proposals themselves. This approach provides a powerful and flexible framework for representing, summarizing and analyzing the contents of large document collections. The topic model is a probabilistic model that automatically learns a set of topics (categories) that describe a collection of documents, based on the words in those documents. Each document is considered to consist of a small number of topics, each of which is dominated by only a fraction of all possible words. As such, the topics define a simplified representation of the documents. A topic model of a document collection is a highly useful representation, but is not necessarily the only or even the best, representation (Blei, Carin, and Dunson 2010, Blei et al. 2006, Newman et al. 2009, Newman et al. 2010). An example of the application for NSF is provided at <http://readidata.nitrd.gov/star>

The topics were modeled based on a collection of titles and project description from all NSF proposals from 2000-2011. For the purposes of this paper, we used a threshold of 100 topics. Each topic is a “bag of words” and represents, in descending order, the relative frequency with which each word co-occurs in the topic. The five top topics highlighted in this paper are:

Astronomical Sciences: star observation X_ray detector galaxies telescope beam mass radiation ...

Mathematical Sciences: function space matrix theorem equation theory vector operator code ...

Cybersecurity: security attack server code privacy storage memory secure cache thread ...

Gene Expression: gene mutant expression RNA cell gene_expression promoter mutation function ...

Magnetic Field: magnetic spin orientation polarization direction angle magnetic_field rotation ...

5. RESULTS

The focus in this paper is to describe the scientific workforce directly supported by expenditures associated with grants made by the National Science Foundation (NSF) to 541 reporting research institutions during the 2011 Federal fiscal year: October 1, 2010 to September 30, 2011.

Although there are some 85 reporting institutions, representing about 43% of NSF’s research portfolio, as STAR METRICS is a voluntary program, the institutions decide how much historical data they send to be analyzed. Some have sent data as far back as 2004, some send data starting with the quarter they decided to participate. For this reason the number of participants represented in this analysis (54) is less than the total number of participating institutions.

The workforce information includes data about the occupational distribution of the workforce, the number of individuals supported and the scientific research areas in which they worked. It

Creating New Administrative Data

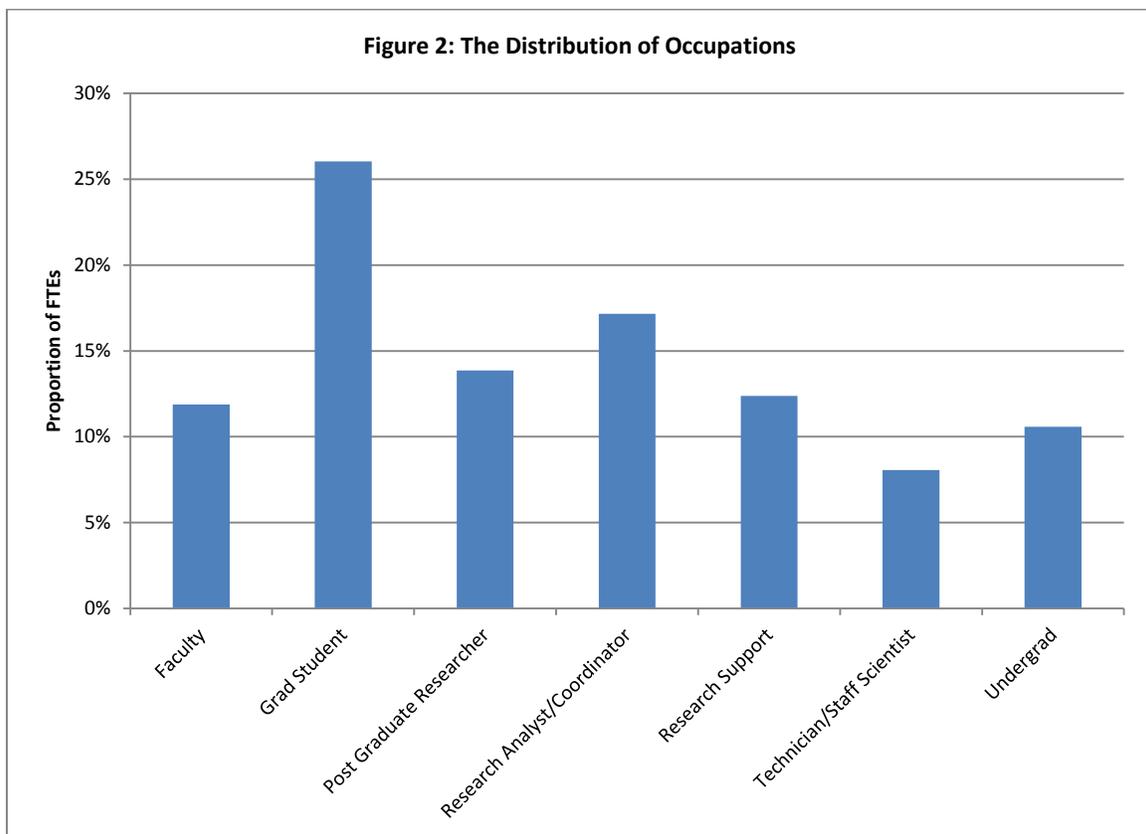
also includes estimates of the institutional jobs generated at the research institutions, including financial, IT, and janitorial services. Estimates are also provided of the total number of jobs related to expenditures by those institutions at collaborating institutions, and at vendors who provide support to those grants, such as scientific supplies.

The first thing to note is some basic facts: as summarized in Table 2, the research institutions reported expenditures related to 10,592 awards and directly supported 6,051 Full-Time Equivalent (FTEs) on payroll, as well as an estimated 2,383 jobs resulting from expenditures on vendors, sub-awards, and institutional support. It is worth noting that the FTE count from payroll understates the number of individuals who derive employment from federal science investments; 24,924 distinct individuals were supported - more than four times the FTE count.

Table 2: Employment Associated with Reporting STAR METRICS Institutions

Number of Awards	10,592
Jobs	
Direct Payroll FTEs	
FTEs	6,051.6
FTEs per Award	0.57
Direct Payroll Individuals	
Number	24,924
Number per Award	2.35
Direct Jobs through Vendors, Sub-Award Recipients, Institutional Support	
Number	2,383
Number per Award	0.22

Because the STAR METRICS data is drawn directly from payroll records, which also have the occupational classifications of each individual employed, it is very straightforward to describe the variety of occupational categories directly supported by NSF funding. Figure 2 shows that distribution for reporting institutions. A broad spectrum of occupations is involved, including clinicians, technicians, undergraduate, graduate, and post-graduate researchers and research support staff. It is useful to note that there is a substantial representation of Full-Time Equivalent undergraduate, graduate, and post-graduate researchers (together nearly 30 percent of FTEs). This is useful information because these scholars represent the next generation of scientists and engineers. Undergraduate support is especially important because there is evidence suggesting that early research experiences contribute greatly to the propensity of students to enter scientific and technical fields.

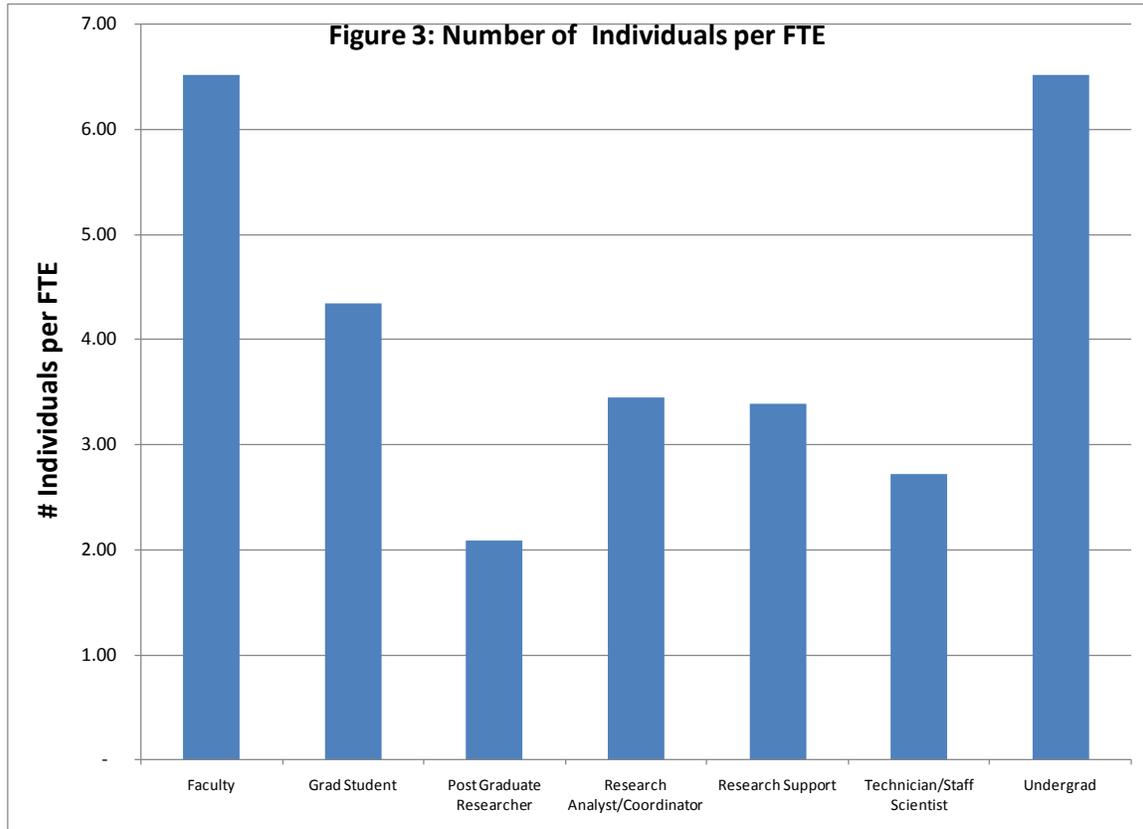


The STAR METRICS system also permits the capture of much more detailed information on time allocation than is possible from Federal agencies' award data. Figure 3, which shows—by occupation—the average number of individuals supported for each FTE position, illustrates this point.

The number of undergraduate students involved in sponsored research, for example, is substantially higher than a simple FTE count would indicate: approximately three undergraduates

Creating New Administrative Data

who combine research with coursework are supported per FTE. Similarly, each FTE supports several technicians and clinicians, research assistants and support staff, and approximately two graduate students and faculty. In contrast, the ratio of individuals to FTEs for post-doctoral researchers is close to two, suggesting that NSF post-doctoral researchers are almost entirely dependent on the funding of one or two primary projects.



The application of natural language processing techniques enabled us to categorize all submitted NSF proposals into 100 aggregate research areas. Because participating institutions provided the NSF award ID number for each grant, it was possible to associate these topic areas with the reported workforce being paid on each project. The distribution of students working in the top five topic areas is reported in Table 2, together with the summary total of all individuals working on all awards. The top research area, which is obviously very sensitive to the research strengths of the reporting institutions, is astronomical sciences. A large number of undergraduate students are working on all the projects, indeed, more undergraduates are being supported than post-doctoral researchers and their employment is spread out over a larger number of awards. The differences in disciplines is very evident from Table 3 -- the post-doctoral researchers are a much higher proportion of the scientific workforce in Astronomical Sciences and Gene Expression and much lower in mathematical sciences and cybersecurity.

Table 3: Number of Individuals Working in Top 5 Research Areas

Research Area	Type	Post-Docs	Graduate Students	Undergraduate Students	All Occupations
Astronomical Sciences	Individuals	101	268	106	1203
	Award	52	97	55	216
Mathematical Sciences	Individuals	11	222	54	596
	Awards	11	99	32	244
Cybersecurity	Individuals	17	177	95	521
	Awards	9	70	38	150
Gene Expression	Individuals	61	107	95	498
	Awards	31	54	30	102
Magnetic Field	Individuals	24	140	67	348
	Awards	14	57	24	84

6. LESSONS LEARNED

There are several major lessons learned from the success of this program, and the lessons are similar to those derived from the Census Bureau’s Longitudinal Employer Household Dynamics (LEHD) program.

The most critical element is that it is a voluntary program. That has ensured that it has to produce information that has value both to science agencies and to universities in order to grow. That, in turn, means that the data quality is high, because both sides have an interest in using the information that is produced. Indeed, just like with the LEHD program, the reports that have come out of the program have evolved substantially in usability and quality in just fifteen months.

The second critical element is that it leverages existing data and is relatively low burden to participants. The STAR METRICS data elements draw from existing systems, and the data

request is respectful of the native financial and human resources systems. This means that many institutions (although not all), have been able to participate; some institutions are combining forces and creating aggregate regional or cross-institutional reports; and many institutions are expanding the scope to include an analysis of the workforce supported by all sources of research funding, rather than just federal science funding.

The last key element is the development of strong relationships between the federal agencies and the research community. This creates a platform whereby new uses can be made of the workforce data – well beyond those initially considered. Thus, for example, researchers at the University of Michigan are planning to do network analysis of STAR METRICS data to identify research strengths. The same team will match STAR METRICS data with Census data to identify regional economic development networks. Researchers at the University of North Carolina are matching the data with information on private foundation funding as well as science funding to examine how different funding structures affect the type of research being done.

7. SUMMARY

This paper describes the initial part of the development of the STAR METRICS program, and its uses in describing the workforce initially supported as a result of NSF federal funding. Of course, science investments have been documented to have longer term impact in many areas, including scientific outcomes, such as the generation and adoption of new science, often measured by citations and patents, on economic outcomes, such as job creation, on the outcomes of the scientific workforce, as well as on social outcomes such as health and the environment. Indeed, the overarching goal of the STAR METRICS program is go beyond documenting the initial results described here and develop an open data infrastructure that enables federal agencies, in collaboration with research institutions, to document and analyze of the inputs into and results of federal investments in science, without increasing the reporting burden for researchers. However, the vision for that is discussed elsewhere (Largent and Lane forthcoming), and is not the focus of this paper.

8. REFERENCES

- Abowd, J., J. Haltiwanger, and J. Lane. 2004. Integrated Longitudinal Employee-Employer Data for the United States. *American Economic Review* 94:224-229.
- Blei, D., L. Carin, and D. Dunson. 2010. Probabilistic Topic Models. *Ieee Signal Processing Magazine* 27:55-65.
- Blei, D. M., K. Franks, M. I. Jordan, and I. S. Mian. 2006. Statistical modeling of biomedical corpora: mining the Caenorhabditis Genetic Center Bibliography for genes related to life span. *Bmc Bioinformatics* 7.
- Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.
- Fleming, L., I. Charles King, and A. Juda. 2007. Small Worlds and Regional Innovation. *Organization Science* 18:938-954.
- Lane, J. 2009. Assessing the Impact of Science Funding. *Science* 324:1273-1275.
- . 2010. Let's make science metrics more scientific. *Nature* 464:488-489.
- Largent, M., and J. Lane. forthcoming. STAR METRICS and the Science of Science Policy. *Research Policy Review*.
- Macllwain, C. 2010. Science economics: What science is really worth. *Nature* 465:682-4.
- Marburger, J. 2005. Wanted: Better Benchmarks. *Science* 308:1087-.
- National Science and Technology Council. 2008. *The Science of Science Policy: A Federal Research Roadmap*. National Science and Technology Council.
- Nelson, L., and S. W. Sedwick. 2011. STAR METRICS: A Participant's Perspective. *NCURA Magazine* 43:24-25.
- Newman, D., A. Asuncion, P. Smyth, and M. Welling. 2007. Distributed Inference for Latent Dirichlet Allocation. *Neural Information Processing Systems*
- . 2009. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research* 10:1801-1828.
- Newman, D., T. Baldwin, L. Cavedon, E. Huang, S. Karimi, D. Martinez, F. Scholer, and J. Zobel. 2010. Visualizing search results and document collections using topic maps. *Journal of Web Semantics* 8:169-175.

9. APPENDIX: PARTICIPATING INSTITUTIONS

Active STAR METRICS Participants (alphabetical order)

NOTE: The current set of participants, while expected to increase, represents a non-random subset of recipients of total science funding.

Arizona State University
Auburn University
Boston University
Brigham and Women's Hospital
California Institute of Technology
Case Western Reserve University
College of Charleston
Colorado State University
Duke University
Emory University
Florida State University
George Mason University
Georgia Institute of Technology
Harvard University
Indiana University
Jackson Laboratory
Johns Hopkins University
Massachusetts General Hospital
Michigan State University
Northern Arizona University
Northwestern University
Ohio State University
Oregon Health & Science University
Purdue University
Stanford University
State University of New York, Albany
State University of New York, Alfred
State University of New York, Binghamton
State University of New York, Brockport
State University of New York, Buffalo
State University of New York, Buffalo College
State University of New York, Canton
State University of New York, Cobleskill
State University of New York, Cortland
State University of New York, CSNE
State University of New York, Delhi
State University of New York, Empire State
State University of New York, ESF
State University of New York, Farmingdale
State University of New York, Fredonia
State University of New York, Geneseo
State University of New York, HSC Brooklyn
State University of New York, HSC Syracuse
State University of New York, IT, Utica/Rome
State University of New York, Maritime
State University of New York, Morrisville
State University of New York, New Paltz
State University of New York, Old Westbury
State University of New York, Oneonta
State University of New York, Optometry
State University of New York, Oswego
State University of New York, Plattsburgh
State University of New York, Potsdam
State University of New York, Purchase
State University of New York, SPO
State University of New York, Stony Brook
University of Akron
University of Alabama
University of California, San Diego
University of California, San Francisco
University of Chicago
University of Delaware
University of Florida
University of Georgia
University of Illinois at Chicago
University of Kansas
University of Massachusetts at Lowell
University of Massachusetts, Dartmouth
University of Michigan
University of Minnesota – Twin Cities
University of Minnesota - Duluth
University of Minnesota - Rochester
University of Minnesota - Morris
University of Minnesota - Crookston
University of Missouri
University of North Carolina at Chapel Hill
University of Pennsylvania
University of Pittsburgh
University of Texas, Austin
University of Texas, San Antonio
University of Virginia
University of Washington
University of Wisconsin-Madison
University of Wisconsin-Whitewater
Vanderbilt Medical Center