Lane, Julia; Owen-Smith, Jason; Rosen, Rebecca; Weinberg, Bruce A.

Working Paper

New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value

IZA Discussion Papers, No. 8556

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Lane, Julia; Owen-Smith, Jason; Rosen, Rebecca; Weinberg, Bruce A. (2014) : New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value, IZA Discussion Papers, No. 8556, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/104667

DISCUSSION PAPER SERIES

IZA DP No. 8556

New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value

Julia Lane
Jason Owen-Smith
Rebecca Rosen
Bruce Weinberg

October 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value

Julia Lane
American Institutes for Research,
IZA, BETA University of Strasbourg, CNRS and University of Melbourne

Jason Owen-Smith
University of Michigan

Rebecca Rosen
American Institutes for Research

Bruce Weinberg
Ohio State University,
IZA and NBER

Discussion Paper No. 8556
October 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

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

IZA Discussion Paper No. 8556
October 2014

ABSTRACT

New Linked Data on Research Investments: Scientific Workforce, Productivity, and Public Value*

Longitudinal micro-data derived from transaction level information about wage and vendor payments made by federal grants on multiple U.S. campuses are being developed in a partnership involving researchers, university administrators, representatives of federal agencies, and others. This paper describes the UMETRICS data initiative that has been implemented under the auspices of the Committee on Institutional Cooperation. The resulting data set reflects an emerging conceptual framework for analyzing the process, products, and impact of research. It grows from and engages the work of a diverse and vibrant community. This paper situates the UMETRICS effort in the context of research evaluation and ongoing data infrastructure efforts in order to highlight its novel and valuable features. Refocusing data construction in this field around individuals, networks, and teams offers dramatic possibilities for data linkage, the evaluation of research investments, and the development of rigorous conceptual and empirical models. Two preliminary analyses of the scientific workforce and network approaches to characterizing scientific teams ground a discussion of future directions and a call for increased community engagement.

JEL Classification: C8, O3, J4

Keywords: UMETRICS, STAR METRICS, Science of Science Policy, linked data, scientific workforce, scientific networks

Corresponding author:

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

* This research was supported by NSF SciSP Awards 1064220 and 1262447; NSF Education and Human Resources Award 1348691; NSF NCSES award 1423706; NIH P01AG039347; the U.S. Bureau of the Census, and the Ewing Marion Kaufman and Alfred P. Sloan Foundations. Data were generously provided by the Committee on Institutional Cooperation and its member institutions. We thank Wei Cheng, Cameron Conrad, Russ Funk, Christina Jones, Felix Kabo, Evgeny Klochikhin, Yulia Muzyrya, J. Staudt and Michelle Yin for research support, Greg Carr, Marietta Harrison, David Mayo, Mark Sweet and Stephanie Willis for help with data issues, and Barb McFadden Allen, Jay Walsh, Roy Weiss and Carol Whitacre for their overarching support. The research agenda draws on work with many coauthors, but particularly work with Michele Pezzoni, Paula Stephan and Jacques Mairesse.

*"The ITG undertook a literature review to determine the state of the science to date. A questionnaire was circulated to Federal agencies to ascertain what methods are currently in use for programmatic investment decisionmaking, as well as to ask what tools and resources are needed by Federal agencies that are currently unavailable. The ITG found that…**the data infrastructure is inadequate for decisionmaking**"* (National Science and Technology Council, 2008) *emphasis added*

"The working group was frustrated and sometimes stymied throughout its study by the lack of comprehensive data regarding biomedical researchers. The timeframe and resources of the study did not allow for comprehensive data collection or the implementation of a comprehensive model of the biomedical workforce. It is evident from the data-gathering and analyses undertaken by the working group that there are major gaps in the data currently being collected on foreign-trained postdoctoral researchers and those who work in industry." (NIH Biomedical Research Workforce Working Group 2012)

1. Introduction

Internationally, public support for science and thus the details of science policy have come to depend on evaluating the results of research. In addition to measures of productivity, establishing the economic impact and public value of investments in R&D is of particular concern. The Research Assessment Exercise in the United Kingdom places tremendous emphasis on scholarly production, as does the Excellence in Research Australia program (Jensen and Webster, 2014; Owens, 2013). The United States has focused both on measuring scientific and economic impact. The policy focus in Japan has been on rebuilding public trust in the value of science(Arimoto and Sato, 2012).

The conceptual framework for implementing impact evaluations in most policy areas is well understood - there needs to be a theory of change and a well-defined counterfactual(Gertler et al., 2012), but by and large no consensus framework exists. In the emerging science of science policy field, one key theory of change contends that funding interventions affect the complex interactions of scientists, which shape collaborative networks. The structure, composition, and content of those networks in turn influence the discovery and training to provide the mechanism whereby scientific and economic impact is achieved. The approach is an advance relative to both bibliometric analysis and accounting frameworks. Bibliometric analysis, which uses sophisticated techniques to study documents, has neither an explicit theory of change nor a counterfactual and was not designed to be used for research evaluation (Cronin and Sugimoto, 2014; Lane, 2010). Accounting approaches, which attempt to tie results to individual grants in order to calculate a straightforward return on investment, confuse the intervention, funding, with the object of interest itself, and thus inherently muddy efforts to define counterfactuals(Lane and Bertuzzi, 2011).

The empirical framework for impact evaluations is also equally well understood – a typical approach is to build a longitudinal dataset that measures baselines, mediating and moderating factors, and outcomes; that dataset is then used and augmented by a community of practice. This special issue of *Research Policy* moves such a framework forward for the field of science of science policy by identifying hitherto unexamined data and by informing the scientific community about new initiatives. It is likely

that future empirical advances can then, in turn, inform a new conceptual framework for science and innovation policy.[2]

Developing that framework is an urgent matter. In the, U.S. the Science of Science Policy has been active since (Marburger, 2005).[3] Yet the United States House Committee on Science, Space and Technology has questioned decisions made on individual grants in political science (Mole, 2013) and the merits of science funding as a whole are regularly challenged (MacIlwain, 2010).

This paper surveys the current landscape. It also describes a large scale, open resource that is being built in the United States, called UMETRICS and sketches an exemplary use case for these data that is based on using new measures of the workforce to map out the scientific networks that underpin federally funded research. It highlights the engagement of a community of practice in the design of the data infrastructure, particularly in classifying occupations and analyzing collaboration networks. It concludes by discussing the importance of engaging the larger community of scholars and practitioners in the establishment of an Institute that provides a sustainable data infrastructure to support scientifically rigorous and practically applicable science and innovation policy research.

2. A Conceptual Framework and the Current Landscape

This section outlines the conceptual framework that underlies our efforts and the current data landscape. The goal of the framework is to answer questions like "What have we learned about NSF-funded research?" and "What is the economic impact of research funding?"(Walsh, 2013). Establishing key descriptive facts about the research enterprise is a necessary step toward validating a conceptual framework that can be be responsive to the needs of science policy makers and of the research community. For example, the E-Government Act of 2002 (§207), although honored more in the breach than in the observance, requires federal agencies documenting R&D investments to develop policies to better disseminate the results of research performed by Federal agencies and federally funded research and development centers.[4] The ability to systematically measure and assess the dissemination and use of findings is a compelling interest for policy-makers, domain scientists, and researchers concerned with social and economic returns.

A. Conceptual Framework

The conceptual framework we elaborate here identifies individual researchers (or the research community consisting of networks of researchers) as the analytical unit of interest. Here, the theory of change is that there is a link between funding (WHAT is funded) and the way in which networks and teams are assembled by the strategic actions of researchers (WHO is funded). We next link features of networks and teams to the process of science, its products, and their transmission. The transmission of

---

[2] By the science and innovation enterprise, we mean the science and innovation ecosystem writ large - from funders (public and private) to researchers (in academia, government, and industry) to the organizations that hire people with scientific training and/or draw on science and innovation to produce commercial products.

[3] The U.S. Science of Science Policy combines a Federal interagency group on the Science of Science Policy charged with identifying policy questions and a scientific research program at the National Science Foundation charged with advancing the Science of Science and Innovation Policy.

[4] http://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf

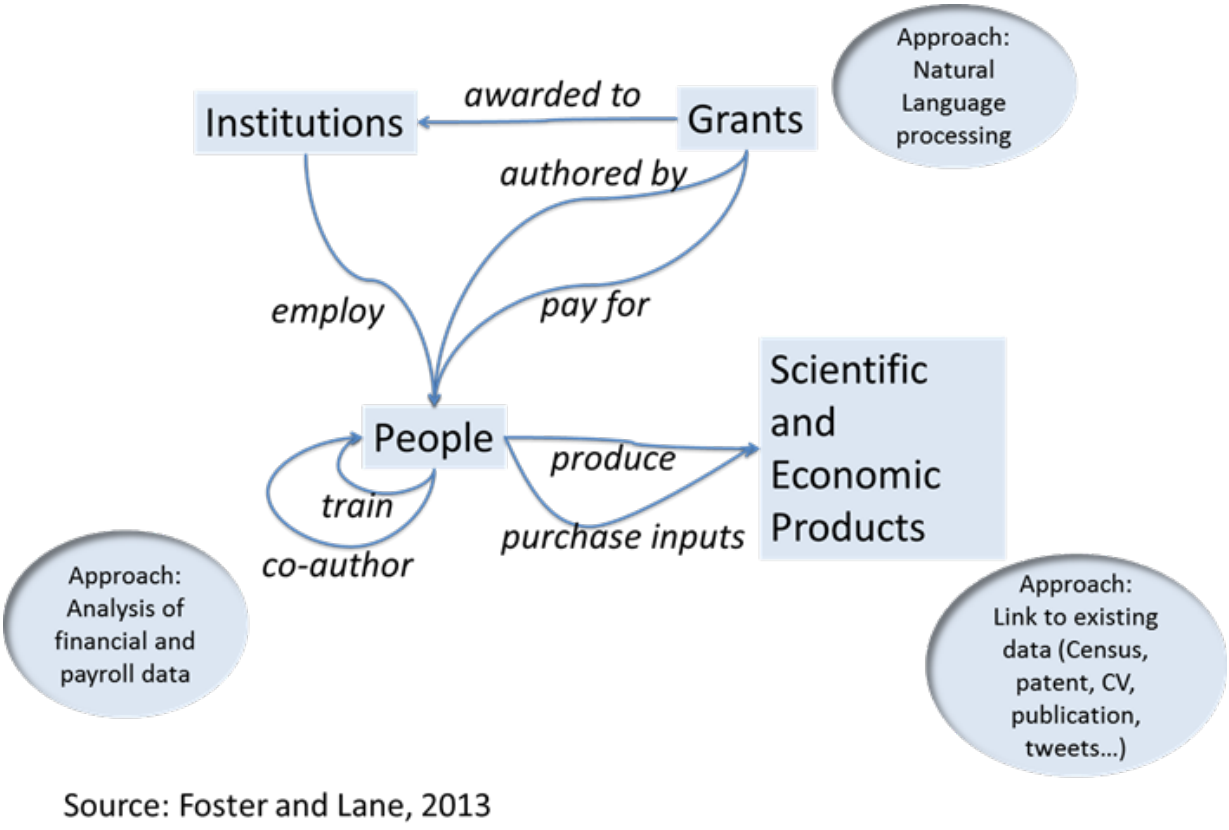
scientific products (discoveries) through the movement of people or the expedient of publication and patenting in turn generates social, economic and workforce "impacts." Institutions both administer grants and provide the material and intangible infrastructure necessary to produce science. Of course, the activities of researchers can be aggregated in multiple ways, since they act both on their own or as members of larger teams and communities, to produce, communicate and utilize scientific knowledge and discoveries.

The framework we propose stands in sharp contrast to that commonly used by science funders, who -- consistent with their mandate to manage research investments rather than document their returns – emphasize individual grants to the virtual exclusion of people, teams, and the later use of scientific products. Hence their primary unit of analysis is the grant, and research administrators spend much time and energy trying to link research grants to research outputs by requiring scientists to acknowledge specific grants and report results on a grant-by-grant basis. The science of science policy framework recognizes that the social organization and work practices of cutting edge science do not fall cleanly within individual projects bounded by particular goals and clear starting or ending dates. Most of the work of discovery and training takes place in collaborative groups that encompass multiple overlapping projects. In practice, the work of individuals and teams is supported by and integrates a pastiche of grants that serve multiple purposes and often span several funding agencies. Even though the primary lever for policy makers to influence the character, goals or uses of science is funding individual projects, the implications and effects of new funding arrangements or incentives can only be fully understood in the context of the individual and collective careers that are the cornerstone of contemporary science and training. Misunderstanding this basic view will lead to misspecification of any analysis.

While it shares a substantive focus on the diffusion and utilization of discoveries and substantial concern with publications and patents as important scientific outputs, the framework we propose also contrasts sharply with the bibliometric literature. That work largely focuses on counting and evaluating the impact of written artifacts that formally codify discoveries. In this model the publication, not the grant, is king. In contrast, our framework argues that making individual researchers the primary unit of analysis is critical because the discovery and transmission of findings on a research frontier depend on mechanisms to transfer tacit knowledge. The sorts of schematic and codified information presented in patents or publications are rarely sufficient to replicate or adapt new discoveries in new settings. As a result, tracing the movement and effects and impact of research often means tracing the movement of individuals whose training and experiences in the sorts of teams described above lead them to embody cutting edge skills and tacit knowledge necessary to the movement and use of novel discoveries. This too suggests the need for data that enable an analytic focus on the careers of individuals within the context of their communities, organizations, teams, and projects(Owen-Smith and Powell, 2004).

In sum, understanding the process, products and eventual impact of science requires the ability to observe and understand action at multiple levels of analysis. The nodes in this conceptual network, which are schematically portrayed in Figure 1 (Grants, People, Institutions, Products) span multiple social scales and evolve on different temporal orders (from multi-year grants, to multi-decade careers, and sometimes centuries long institutional histories). The time frames at work are varied and can be discontinuous. In addition, the direction of causality is not one way. Funding certainly shapes the direction of science and the career choices of individuals. But the progress of science also commonly exerts its own influence. For instance, the products of research such as new findings or tools can upend existing knowledge, change the direction of careers and even shift the focus of institutions and markets. The fundamental scientific discoveries that underpinned the molecular biology revolution in drug

development spawned the biotechnology industry (Cockburn and Henderson, 1998; Powell and Giannella, 2010). Nano-scale visualization and intervention technologies sparked much of the boom in nano-technology (Mody, 2004). The computational capabilities for big data management and analysis that underpin this initiative as well as many of the booming 'omics' areas of contemporary life sciences, represent instances where scientific products can substantially change the content and standards of evaluation for proposals, the trajectories and foci of careers, and even the organization and orientation of scientific and economic actors.



Source: Foster and Lane, 2013

Figure 1

Figure 1 illustrates two key features of the framework we seek to elaborate. First, note that many different activities connect these nodes. Second, note that while grants are a key input, people, joined in networks and teams, are both the primary actors and objects of work. This set of insights underpins our contention that the next generation of analyses should derive from conceiving of the scientific enterprise in multi-level and potentially non-linear terms. The ecosystem of science, then, must be analyzed not simply in terms of counts or trends but also in terms of complex and often multiple relationships among diverse system components. We thus propose a data platform that begins with fine grained administrative data on grants but that anchors those on the careers of individuals nested within teams, intellectual trajectories, and institutions. The ability to shift analytic frames across levels of analysis and temporal scales is a key component of our effort.

B. The Current Data Landscape

Most research on the science of science and innovation has used hand-curated, one off datasets. Moreover, efforts to track the outcomes of research grants and projects are manual and artisanal instead of automatic and scalable, imposing large burdens on the research community and making individual datasets too idiosyncratic for broad use or easy integration with other information. These data are frequently difficult to share and link and thus often prove difficult to use for questions beyond those they were designed to answer. The expense of constructing these data hinders progress on the science of science and innovation, leads to duplication, and impedes replication.

Yet the amount of data available on the research enterprise is vast. PubMed alone comprises over 23 million records; data are available on patents, with the United States Patent and Trademark Office having both metadata and full text of over 4 million patents; data are available on over 2.5 million grants from the National Institutes of Health and roughly 200,000 from the National Science Foundation. Moreover the number of citation records dwarfs even these data, with the Thomson Reuters Web of Knowledge indexing 65 million cited references annually. Online sources of large data on the longer-term effects of scientific knowledge include, but are hardly limited to, clinical trials data, FDA drug approvals, and environmental regulatory documentation. In addition to these data, research universities, hospitals, and institutes have their own internal data, comprising the purchases made for research projects and the time allocated to research projects by their personnel. There are literally tens of thousands of research related financial transactions that occur annually, even for relatively small institutions, and many thousands of payroll transactions.

The challenge, and the opportunity, for the research community is that these data have not been connected systematically for research purposes. Yet advances in computer and information technology are making it increasingly possible to transform these data and gain better understanding of how science works. New computational approaches have been used to identify stars (Zucker and Darby, 2006), classify science(Griffiths and Steyvers, 2006), disambiguate author and inventor names(Li et al., 2014; Ventura et al., 2014), develop linked data platforms (Torvik et al., 2005b), and analyze outcomes of science investments (Azoulay et al., 2007; Kenney and Patton, 2013a)

3. The Data and Collaboration Platform

The framework and data curation effort we describe here is called UMETRICS. It has the two canonical features discussed in the introduction: the development of a new data infrastructure and the engagement of a community of practice. It is being developed under the auspices of the Committee on Institutional Cooperation (CIC), a consortium of 15 U.S. Universities, and captures data on the individuals who work on federally funded research as well as other inputs into the production of science from university administrative systems[5].

[5] https://www.cic.net/projects/umetrics . It is worth noting that although the current focus is on federally funded research, the approach is scalable to all sources of research funding, and is also scalable to include private enterprises engaged in R&D.

The platform builds upon the STAR METRICS project that was initiated in 2009 by the U.S. federal agencies, with the engagement of the Federal Demonstration Partnership[6]. The goal of the original project was to build an open source, two-layered, continually evolving data platform that could be used to (1) provide policy makers with a better understanding of the process of research and (2) provide the research community with a common data infrastructure. Serving both those needs would help to create a body of knowledge that specifies and tests causal accounts of scientific productivity and impact. The original project drew 14 data elements from university payroll and payment records to measure the time allocated to projects by all participants as well as equipment purchases. Staff in University administrative data units and sponsored research offices worked together with federal agencies and researchers to determine the highest quality, lowest burden approach to acquiring those data(Nelson and Sedwick, 2011)

By design, the STAR METRICS Level I data were de-identified and hence cannot be linked to non-university datasets in order to connect the workforce to research inputs and outputs(National Academy of Sciences, 2014). Staff involved with the UMETRICS project worked to enhance STAR METRICS Level 1 data for participating universities with award titles and crosswalks between unique identifiers and real names for both employees and vendors. With important identifiers in place, these expanded UMETRICS data can be linked to other information about researchers: their networks, their resulting research products, and the utilization and outcomes of those products, including, ideally, failures. The measures such linkages afford range from biographical information -- which can be obtained from web searches and imputed from CV data -- to scientific outputs such as dissertations, publications, and patents, and even some career outcomes. The broad array of output data that can be linked to UMETRICS employee and vendor names includes publications, presentations, and workshops (and the associated meta data and full text information) as well as economic output measures like earnings and job placements, career trajectories and the establishment, productivity and growth of businesses. In the United States a wealth of information can also be obtained from links to the Survey of Earned Doctorates, which is administered to *all* recipients of earned doctorates at the time of degree completion as well as to links to the Census Business Register and Longitudinal Employer-Household Dynamics (LEHD) data. Ultimately the fuller range of research products, including data, equipment, and organisms should be included as well. Because of its reliance on using scalable algorithms to combine and mine existing data, this work falls broadly within the realm of "big data."[7] Figure 2 provides a schematic description of linkages being developed using the common 'backbone' of identified employer and vendor data.

---

[6] The Federal Demonstration Partnership (Thefdp.org) is a cooperative initiative among 10 federal agencies and 119 institutional recipients of federal funds.

[7] Big data have been variously defined; the term is an imprecise description of a rich and complicated set of characteristics, practices, techniques, ethics, and outcomes all associated with data. As such, there is no canonical definition. People have tried to define it by characteristics: Volume Velocity Variety (and Variability and Veracity), by source: found vs. made, by use: professionals vs. citizen science, by reach: datafication, or most fundamentally, by a change in paradigm which results from changes in many factors that affect the measurement of human behavior: the nature of the new types of data, their availability, the way in which they are collected, and data dissemination. UMETRICS fits all of these (Einav and Levin, 2013; Hey et al., 2009).

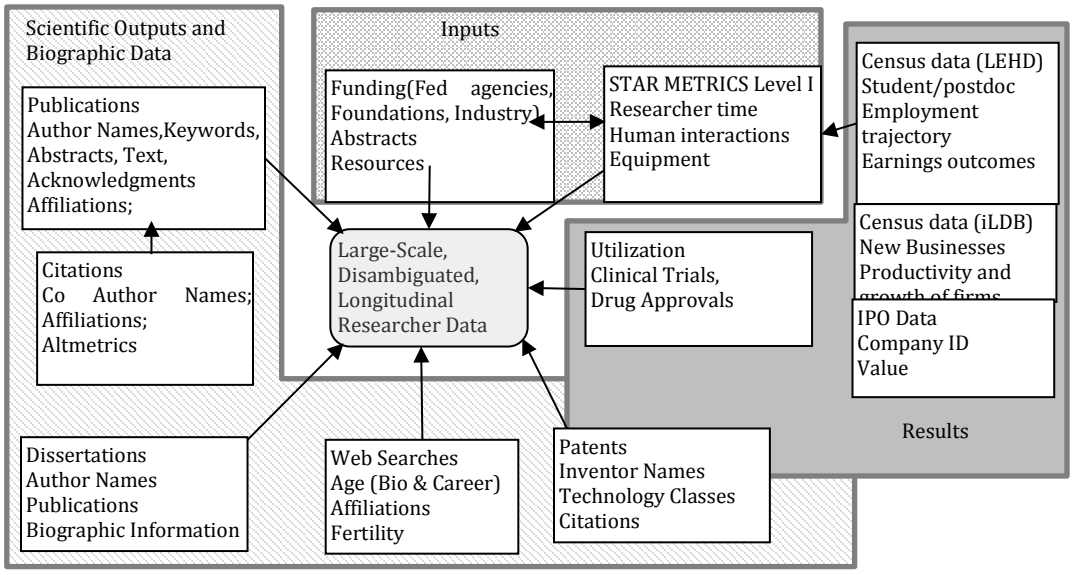


Figure 2: The Core UMETRICS data infrastructure

A. The inputs to science

There are four basic ways in which existing data on science expenditures can be repurposed for research purposes;

(i) Individuals can be directly employed on a grant,

(ii) Researchers at collaborating institutions can be supported,

(iii) Scientific and other supplies can be purchased from vendors, and

(iv) Infrastructure support, including financial, IT, physical space and research services can be provided through overhead.

Each of these activities creates a financial transaction that can be used to calculate the amount spent on the associated activities(King et al., 2013). Figure 3 provides a stylized description of how these financial transactions flow through a typical administrative system.

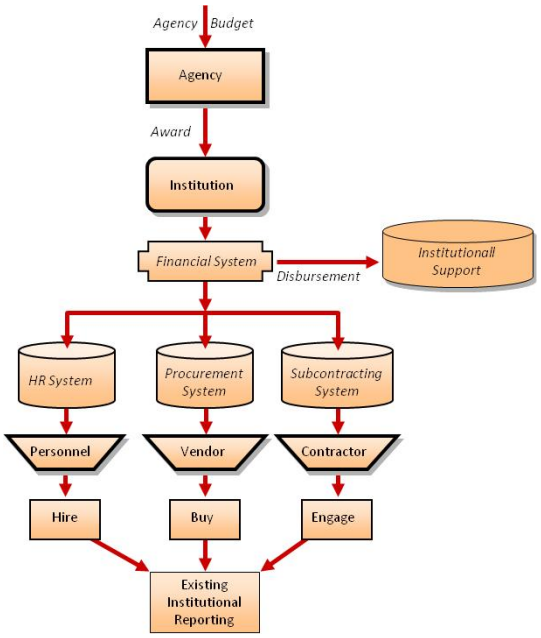


FIGURE 3 THE FLOW OF ADMINISTRATIVE TRANSACTIONS ASSOCIATED WITH FEDERAL FUNDING TO A RESEARCH INSTITUTION

The process diagram in Figure 3 demonstrates how the Human Resources system in a research institution can be used to identify, on a monthly or 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 funding mechanism. Just as the LEHD program used unemployment insurance wage records to capture the flows of workers across firms, this approach 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. Many pressing questions can be addressed when those data are in hand and relevant links have been made.

Consider standing concerns about how researcher characteristics are related to research productivity. For instance, the relationship between age and creativity is a classic question in the science of innovation and researchers have argued that trainees from underrepresented groups perform better when mentored by people from the same group(Blau et al., 2010). The basic UMETRICS data are enriched by data on the researchers themselves. Career age can be imputed in two ways: (1) as the time since receipt of a research doctorate (estimated from dissertation publication dates) and (2) estimated from publication and grant records using the time elapsed since the first publication. Gender and ethnicity can be estimated probabilistically from names.(Kerr, 2008; Mateos, 2014, 2007) In addition, data on people who obtain PhDs in the United States can be obtained from matches to the Survey of Earned Doctorates.

B. What is being funded

A considerable amount of effort has going into building better ways of describing what science research is funded, since this is a precondition to studying the results of science investment. Here we illustrate the situation focusing on the U.S. Federal government. Despite having made a considerable effort to document its research activities, information about what research is funded by the U.S. federal government is relatively limited. A report by the National Research Council on two key surveys laments.

> Two surveys[8]… provide some of the most significant data available to understand research and development (R&D) spending and policy in the United States**. [B]udget officials at science agencies, Congress, and interest groups representing scientists, engineers, and high technology industries, among others, constantly cite the survey results—or studies based on those results—in making public policy arguments**. However, the survey data are of insufficient quality and timeliness to support many of the demands put on them. **[T]the information provided to SRS is often a rough estimate, frequently based on unexamined assumptions that originated years earlier."** (National Research Council 2010, p. 1, emphases added)

The two leading science funding agencies in the United States are the National Science Foundation (NSF) and the National Institutes of Health (NIH). Despite being the most advanced in making details about their research portfolios public, these agencies do little better: NSF's research.gov (http://www.research.gov/) and the NIH Reporter (http://projectreporter.nih.gov/reporter.cfm) are very useful tools to capture information about individual awards, but do not provide a good overview of the national funding landscape. NIH has invested substantially in reports and tools to describe and visualize the research funded by their 27 institutes and centers. Featured on the NIH Reporter site, these tools sometimes leverage GIS technology, natural language processing and advanced data visualization (www.nihmaps.org). Users can export the full database of awarded NIH grants and metadata, for research analyses. Yet the widespread NIH methodology to describe and report on science investments (http://report.nih.gov/rcdc/) is highly manual and static and is useful for budget reporting but not for science policy analysis. As we write this, the STAR METRICS Federal Reporter tool is in alpha release (http://federalreporter.nih.gov/). This tool represents the first Federal initiative to normalize grant data across multiple science agencies (NIH, NSF, USDA, NASA, EPA, DoD, CDC, FDA, AHRQ, and VA). STAR METRICS Federal Reporter is built on the NIH Reporter framework and may eventually provide data coverage and analytical tools to generate an accurate description of the national funding landscape.

Researchers affiliated with the UMETRICS activities have experimented with using new computational techniques, such as natural language processing and topic modeling, to provide fresh perspectives on research investments and portfolios and their contributions to scholarship and innovation, both nationally and internationally(Lane et al., 2013). Many of these technologies are open source, so they can be customized in flexible ways. They are also robust, so the results can be understood by a well-formed and mature scientific community. The approach emerged from work commissioned by NSF's Advisory Subcommittee of the Advisory Committees for the Social, Behavioral, and Economic (SBE) Sciences and the Computer and Information Science and Engineering (CISE) Directorates to identify

[8] Survey of Federal Funds for Research and Development (the "federal funds survey") and the Survey of Federal Science and Engineering Support to Universities, Colleges, and Nonprofit Institutions (the "federal support survey").

techniques and tools that could characterize a specific set of proposal and award portfolios (National Science Foundation, 2011).

C. The results of science

The community is also using new computational technologies to capture information on research outputs. Research products are increasingly both digital and accessible via the Web, making it possible to obtain much relevant information via Web scraping and automated inference. All US patents granted since 1976 are available in digital form. So is the metadata, and often the full text, for many research publications.[9] Catalogs of other research products, such as datasets, are slowly emerging with services such as FigShare and Dryad. Online usage statistics are being captured at the article level by journals such as Public Library of Science and web usage statistics for most journal articles and research products are tracked by systems such as altmetric.com and Impact Story. Researcher Web pages and CVs, whether created manually, algorithmically, or via systems such as VIVO and Profiles, are further valuable sources of information. Similarly, the international ORCID initiative makes it possible to access information on researchers in bulk. In the United States, a Federal interagency group is leading the SciENCV project, which allows researchers to describe their scientific contributions in a structured format.[10]

The research community and our teams have already started to describe research products by linking multiple outputs – such as Lee Fleming's patent and patent application data(Fleming, 2011), Torvik and Smalheiser's Author-ity data on biomedical publications(Smalheiser and Torvik, 2009), Lee Giles' CiteSeerX data on publicly accessible scientific texts and dissertation data (Giles et al., 1998; Giles and Khabsa, 2014; Khabsa et al., 2012), and the ProQuest record of dissertation data – to UMETRICS core data from a number of major CIC universities. These data on scientific publications can also be linked to measures of community uptake such as the citation and article level metrics developed in bibliometric research. In addition, we hope that information on the activities of scientific teams (faculty, graduate students, postdoctoral students) subsequent to funding will be derived (confidentiality protected) by matches to Census Bureau records that provide detailed information on starting and career earnings, firm and industry placement, firm startups, and firm productivity growth.

Numerous researchers are already using UMETRICS data to answer key science policy questions. MaryAnn Feldman has been using the data to describe how private foundations affect the structure of science. Jason Owen Smith has been examining the impact of space allocation on research networks (Kabo et al., 2014). Kaye Husbands Fealing is examining the public value of science using food safety and security as the model(Husbands Fealing, 2014). Mikko Packalen and Jay Bhattacharya are estimating the health impacts of biomedical innovation.

D. Scientific Impact.

Perhaps the greatest lacuna for both policy and research purposes is the ability to trace and understand the myriad ways in which public investments in science, scientists, and institutions eventually yield social benefits in a variety of fields. Whether the concern is food safety, environmental remediation, economic growth, population health, defense and security, or a host of other national priorities, there is

[9] See for example http://rd-dashboard.nitrd.gov/

[10] http://rbm.nih.gov/profile_project.htm

great opportunity and considerable challenge in identifying the ways in which early stage and uncertain public investments in science yield eventual returns to society. The engagement of the research community in designing and building out the data infrastructure offers the potential to begin addressing just such questions by tracing the movement of scientific products and research trained people out of the research ecosystem and into other connected areas of society.

In keeping with our conceptual emphasis on the careers of individuals and the movement and transformation of discoveries, the framework we propose in Figure 2 offers three immediate points to examine just these transitions. The work of many researchers traces connections from scientific discoveries to publications, patents and eventually new businesses or new drug and medical device approvals (Fleming, 2011; Kenney and Patton, 2013; Lichtenberg, 2012), offering new possibilities to trace the eventual health benefits of publicly funded scientific discoveries. Likewise, the careers of individuals and the effects their skills have on the productivity and growth of private and non-profit employers in numerous sectors can be examined by linking to individual level earnings and establishment level data maintained by the U.S. Census and the Internal Revenue Service (Abowd et al., 2004).

In particular the ability to examine the outcomes and effects of scientific training offers new possibilities for understanding how public investments in research and training institutions do or do not result in economic growth and well-being. These examples offer just two possibilities for connecting the core of the UMETRICS infrastructure to existing data sources relevant to questions about the public value of science and R&D investments. But note that the key challenges faced here have to do with the balance between confidentiality/privacy and research access and with the urgent need to draw together data and resources maintained by multiple federal agencies. Developing a robust and effective system for maintaining privacy and confidentiality in the age of big data offers an additional set of fruitful areas for research (Lane et al., 2014).

4. Using the Data

One example of how the research community, together with UMETRICS data, can advance policy is in studying STEM pipeline and workforce issues as well as the organization of research teams.

Funding agencies such NIH and NSF currently struggle to respond to questions about the STEM workforce (Marburger, 2011). They are largely unable to systematically evaluate the effectiveness of their STEM training and research grant policies (National Science and Technology Council, 2008). Indeed, when the NIH Director convened a Biomedical Research Workforce Working Group tasked with identifying the optimal workforce composition and training levels necessary to support a maximally productive biomedical research ecosystem "The working group was frustrated and sometimes stymied throughout its study by the lack of comprehensive data regarding biomedical researchers." The working group pointed out that much of what is known about the STEM workforce comes from small-scale and/or cross-sectional surveys such as the Survey of Graduate Students and Postdocs, Survey of Earned Doctorates (which is population level, but cross sectional) and Survey of Doctorate Recipients[11]. They identified three key problems with using these existing datasets in STEM workforces policy analyses: (1) the data are manually collected and/or are small-scale and/or cross-sectional rather than longitudinal,

[11] http://www.nsf.gov/statistics/surveys.cfm

(2) data is not captured for trainees on research teams funded by standard research award mechanisms and is incomplete for foreign trained researchers, and (3) the data collection systems do not enable the large-scale long-term tracking of trainees once they enter the workforce (National Academies, 2005). In what follows, we sketch two initial, complementary approaches to addressing questions about the composition of the research workforce and the conditions of STEM training that highlight the potential value the kind of linked micro-data being developed through UMETRICS.

A. The composition of the workforce

As is often the case with economic data, there are a variety of ways to classify who is (and who is not) part of the US science and engineering workforce. Some surveys would categorize those with any scientific training as part of the STEM workforce, while others might only characterize the STEM workforce as those who are actively working in R&D positions in academia or industry (Husbands Fealing et al., 2011). One focus of the UMETRICS project is to determine the value of university administrative records as a way of describing and documenting the occupational composition of the science-producing workforce. The advantage of using administrative records is that Human Resource (HR) systems are the most granular source of occupational data possible; the disadvantage is that the data comprise thousands of distinct occupational categories for employees who are supported on research grants; a classification typology needs to be developed to ensure consistency across categories. Here we describe a first attempt at confronting these daunting classification issues. At all points, it is critical to engage practitioners in the classification exercise.

The first step in our process was to develop a typology for categorizing university occupational titles into a common framework. We built off of the work of the Federal Demonstration Partnership (FDP)'s STAR METRICS task group, which identified eight major occupations that were agreed to be of greatest interest for the study of science – Faculty, Post-Graduate Research, Graduate Student, Undergraduate Student, Technician/Staff Scientist, Research Analyst Coordinator, Research Support and Clinician. They then listed the types of HR occupational titles that should be mapped into each of these categories (Harrison, 2013).

We analyzed those 8 STAR METRICS occupational categories and identified 5 key dimensions - permanence, research role, track, scientific training, and clinical association – with which a combination of characteristics could be used to distinctly identify each of the 8 categories, as shown below.

Table 1. A possible University STEM workforce typology

Characteristic	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

There may be a one-to-many relationship between major occupational categories and combinations of the 5 typology dimensions, but each combination of dimensions maps to only one occupational category. This approach can also serve as a framework for mapping the thousands of university HR-designated job titles to aggregate occupational categories. Table 2 shows how this typology applies to three occupational categories. For example, a postgraduate researcher at University X might be considered non- permanent (P2), working in the role of an idea-generating researcher (R2), pursuing a non-professorial track (T2), and with an advanced science degree (S1). Both clinical (C1) and nonclinical

(C2) postgraduate positions would be mapped to this category. The remaining maps are reported in the appendix.

Table 2. Example mapping of STAR METRICS occupational categories to the STEM workforce typology

FACULTY (4 possible combinations):

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

POSTGRADUATE RESEARCH (2 possible combinations):

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

RESEARCH ANALYST/COORDINATOR (8 possible combinations): Example occupational categories: Research analysts; Study coordinators; IACUC coordinators; Clinical coordinator; Clinical specialist; Research specialist; Lab coordinator

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

The challenge of mapping thousands of HR job titles to high level workforce categories is not a trivial one. For this study, we report on occupational mapping results for 5 universities (Chicago, Indiana, Michigan, Minnesota, and Ohio State). First we expanded on the STAR METRICS occupation categories and identified 14 UMETRICS workforce categories that could be distinguished by the typology described above. We then systematically mapped the original university job titles, for all employees who are paid on research grants, to these 14 categories. In what follows we describe that work and report on the strengths and weaknesses of the approach.

Implementation

The first attempt to apply this typology to comprehensively map university job titles to occupations(Conrad et al., 2014) employed a two-step procedure for disaggregating broad STAR METRICS major categories into more specific occupational categories, or UMETRICS workforce categories. We first classified a subset of occupations based on the person's relationship to the university, according to the 5 dimensions described above and building on the initial FDP effort. These occupations were:

- **Faculty**: all advanced academic employees who are directly involved in scientific research.
- **Clinicians**: all non-faculty health care professionals.

- **Post Doctoral Research**: all individuals holding terminal degrees (PhD or MD) who are in temporary training status, including all medical residents and fellows.
- **Graduate Student**: students seeking advanced degrees.
- **Undergraduate**: students earning baccalaureate degrees who serve as research assistants.
- **Staff**: non-degree seeking, non-faculty, permanent employees

Within the Staff occupational category, we then sought to distinguish people based on the work they perform, identifying staff categories for Staff Scientist, Research Analyst, Technician, Research Support, Technical Support, Research Administrator, Research Coordinator, and Instructional.

- **Staff Scientist**: all advanced degree qualified non-faculty scientists and engineers.
- **Research Analyst**: all advance degree qualified non-faculty research professionals.
- **Technician**: skilled and specialized employees who have been specifically trained in some area of science and technology.
- **Research Support**: administrative employees who are not specifically employed for scientific research purposes but perform job tasks that support research.
- **Technical Support**: technical employees who are not specifically employed for scientific research purposes but provide network, systems, and data support.
- **Research Administrator**: employees who direct and influence scientific research activity from the level of the laboratory up to the level of university research center.
- **Research Coordinator**: managers for laboratory studies, clinical trials, and research programs.
- **Instructional**: employees who function as lecturers but do not actually engage in scientific research.

Lastly, we included an **Other** category for people who could not be clearly classified into one of the other categories.

We found that the number of job titles from each university is quite large with an observed rangefrom 269 to 1,176 job titles.[12] Moreover there is substantial variation in naming conventions. A research associate at one university might be the same as a research assistant at another, and research fellows at multiple universities may represent quite distinct job descriptions. To give a sense of the task, we first show the distribution of job titles that we mapped (using the methods described below) into the 6 broad occupations based on relationship to the university (Figure 4A). The majority of job titles (close to 2,000) correspond to staff positions. Because so many job titles are for staff and staff play such varied roles, Figure 4B shows a first attempt at further classifying the staff job titles. These estimates are particularly tentative insofar as this further classification involves making many fine distinctions.

[12] The number of years of data provided varies across universities, which accounts for some (but hardly all) of the variation in the number of job titles.

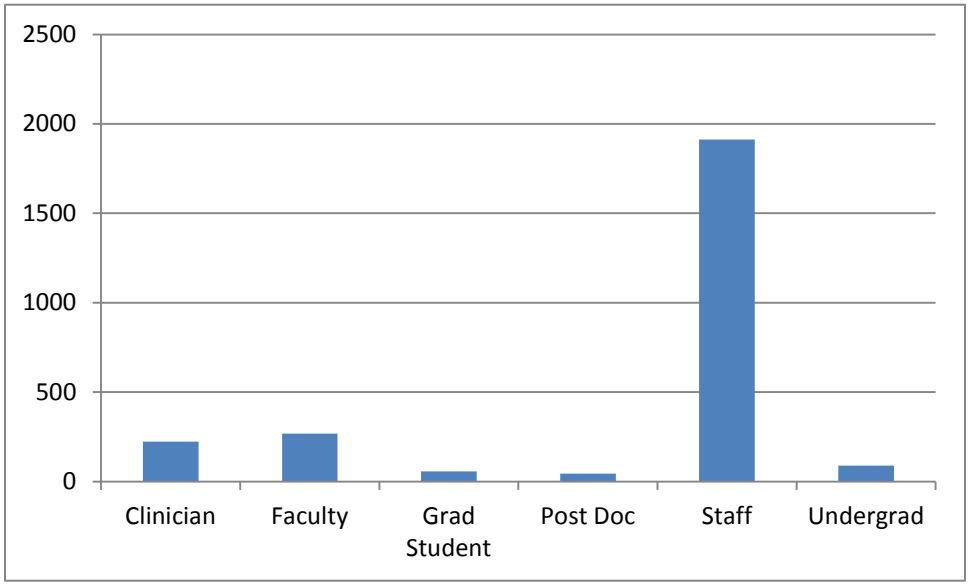


Figure 4A. Number of University-Assigned Job Titles Mapping to 6 Broad UMETRICS Occupational Categories.

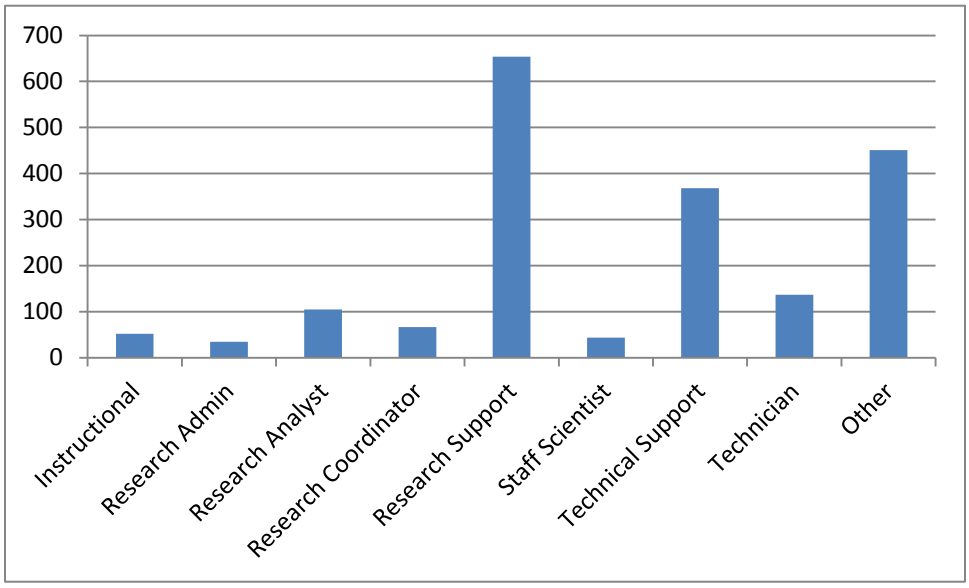


Figure 4B. Number of University-Assigned Job Titles Mapping to 9 “Staff” Occupational Categories.

Although some job titles could be classified quickly and accurately, a number of positions required further analysis. University websites, including employment databases and job postings on university human resources websites often provided detailed descriptions of the nature of work performed for specific job titles. We were also able to identify people who held jobs from online salary databases for

public universities and then obtain information on the work they perform from their profiles on university and professional networking websites.[13]

In building this classification system, a number of classes of challenges arose:

1. **One title and two categories:** One job title combines two distinct sets of workers who clearly belong in two or more different occupational categories. For instance, program coordinators comprise some employees who should be classified as Research Support (e.g. people managing the business operations of a scientific research program at a university center) and others who should be categorized as Other (e.g. people involved in educational or student experiences).
2. **Jobs at the Margins of categories:** Workers who are at the margin of two different closely related categories. For instance, some laboratory supervisor positions combine work as an administrator for a university research lab while managing a laboratory study that involves several research assistants. Because the employees' work encompasses the responsibilities of both Research Administrator and Research Coordinator, their classification falls at the margin of two categories, with the balance surely varying from person to person.
3. **Ambiguity:** Vague or broad job titles limited our ability to generate precise classifications. Administrative support, coordinator, and professional aide all encompass people conducting a wide range of functions from human resources to undergraduate admissions to a wide range of offices supporting general university functions. Still other employees with these titles may be directly involved in supporting or conducting scientific research.
4. **Dual classification**: In some cases an individual employee could be classified into two different categories; one classification based on the relationship to the university, and the other based on the nature of the employee's work. Consider research nurses, who can be classified as both Clinician (because of their role in providing direct patient care) and Research Support (because of their role supporting scientific research).

To address these concerns, we included up to two categories for each occupation so that users can explore the robustness of results by assigning people with a given job title to the most likely occupations. A confidence rating system (ranging from 1-10) was also used to rate job titles based on the degree of certainty that they were correctly classified.

Results

Our initial analysis of the workforce data from these 5 institutions for 2012 shows the number of distinct people employed under Federal support in each occupation (Figure 5A). We find that the majority of employees on research grants at these 5 universities can be categorized as "staff". We further find that almost as many graduate students are employed as faculty and that the number of undergraduates is only slightly lower. Indeed people in the "STEM pipeline" (undergraduate students, graduate students, and postdocs combined) number 18,665 and exceed the number of staff (14,567, excluding clinicians but including the entire other category) and also the number of faculty (8144) by wide margins.

[13] Additional information on the websites for individual universities is available upon request. The general websites that proved useful include: http://www.indeed.com/; http://www.simplyhired.com/; http://www.higheredjobs.com/; https://www.linkedin.com/, which was valuable if a person with a particular job title could be identified; http://careers.insidehighered.com/; and http://www.linkup.com/.

Figure 5B shows the number of people supported in staff positions by grants in 2012 for the various staff categories. Research Support and Research Analysts are the most common categories. Thus, while people directly engaged in research (Research Analysts) are prevalent, so are people who are not directly involved in conducting research (Research Support staff). The challenge to classifying Staff can also be seen from the large number of people classified as Other, the vast majority of whom are Staff.

The staff figures are particularly valuable because staff are almost surely less likely to appear as authors on publications or inventors on patents, making it harder to identify their important contributions in other common data sources. Moreover, their labor market is particularly important to understand because many staff are dependent on grants for their continued salary support and employment. We see these data as offering a potentially novel and insightful window into the size, structure, support, and activities of "invisible technicians" (Shapin, 1989) and other staff, who play essential roles in the scientific eco-system. Of course, additional work will be necessary to refine this classification and to determine the extent to which these data can, in practice, provide insights into the staff employed conducting science.

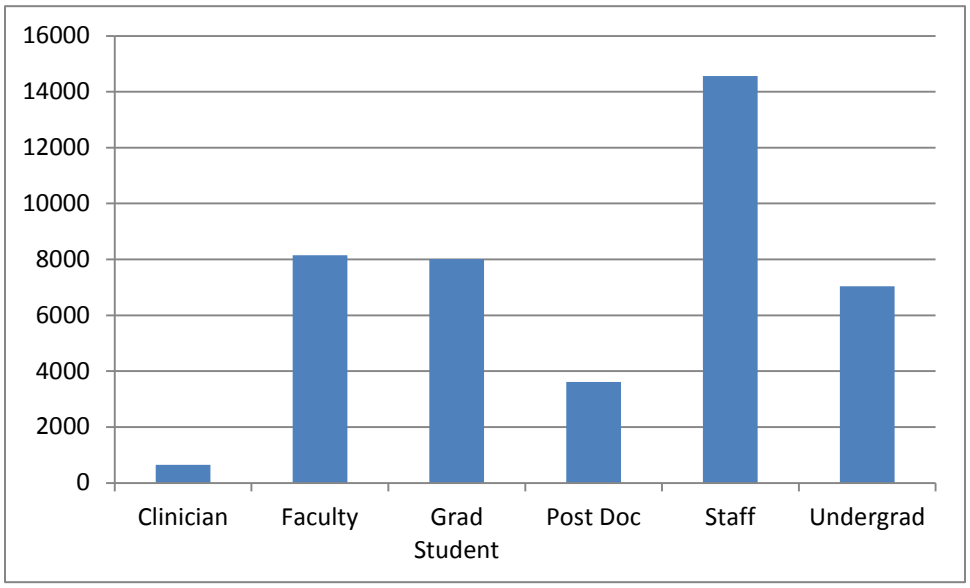


Figure 5A. Occupational Breakdown of Workforce by Distinct Individuals Employed by Research Grants at 5 CIC Universities (2012).

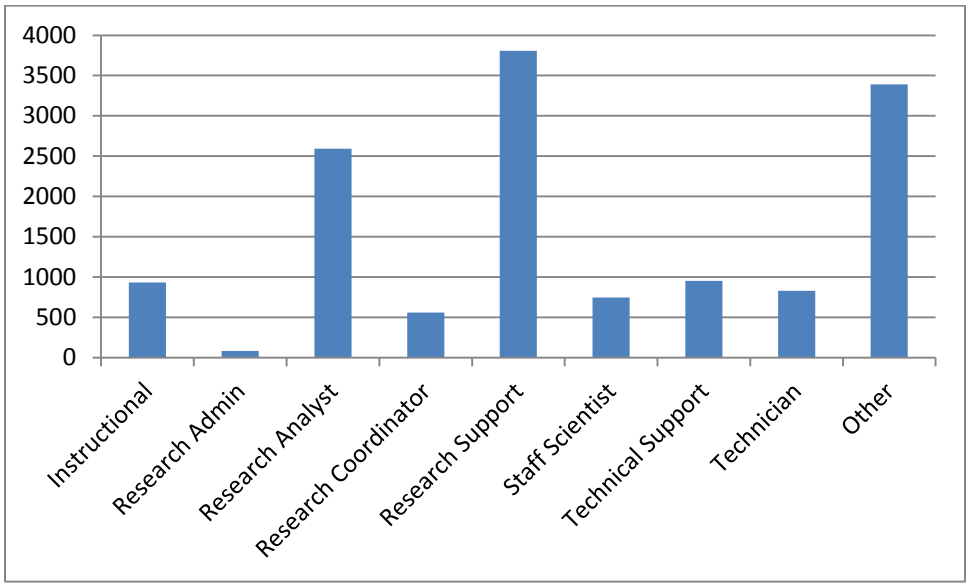


Figure 5B. Occupational Breakdown of Staff Workforce by Distinct Individuals Employed by Research Grants at 5 Universities (2012).

Next Steps

Considerable work is required to refine the occupational classifications, using the typology and associated descriptions as a framework. We are working with universities now to improve our categorization. Additionally, it is critical to scale our approach. One way this can be done is to identify frequent positions that are commonly coded across institutions and apply that classification to newly participating universities. In this way, hand-coding can be substantially reduced.

At a substantive level, these estimates could be used both to analyze the effect of science policy changes on subsets of the STEM workforce, such as changes in science funding, including the recent US Federal Government's "sequestration," an across-the-board 10% cut of all Federal spending including science funding. They can also be used to create person-level network models of the research workforce or models of the STEM pipeline, as suggested in the NIH Biomedical Workforce report.

B. Describing teams

In addition to a focus on individuals, this data infrastructure has great potential to capture and examine the wide variety of ways in which science is the product of teams. Intriguing work on publication co-authorship has suggested that there have been fundamental changes in the size and nature of collaboration (Jones, 2009; Wuchty et al., 2007). Linked UMETRICS data enable us to look at project level collaborations. While individual people are fundamental to our framework, the core unit of production and training for contemporary research is not the individual, it is the team, research group or lab. Systematic data on wage payments on federal grants to individuals in all the occupational categories we describe above allows for the most extensive and exhaustive characterization of the social organization of scientific discovery and training ever undertaken. Linked information on scientific products and individual careers offer the possibility of systematically connecting differences in the composition and structure of such teams, in their funding structure, and in the staffing choices of principal investigators to outcomes.

Approach: Defining teams from Co-employment Networks

Put simply, the record level wage data that are the heart of the UMETRICS initiative offers new possibilities for analyzing the process of discovery and training through the lens of collaborative networks. Record level employment data have several features that make them particularly useful for the task of understanding how universities transform grants into new knowledge and skilled people. As a first step, we construct yearly panels of collaboration networks, treating individuals as nodes that are linked when two people are paid any part of their salary on the same grant in the same year. Unlike commonly used co-authorship networks, these data capture information on nearly everyone employed in grant-funded research on campus. As the prior section indicates, many of these individuals (for instance skilled technicians, programmers, lab managers, or machinists) may never appear as authors on papers their work enables. The scale of the role played by professional staff in academic research has been too little recognized and it has dramatic implications for network analyses of the research process that do not include these important participants.

Unlike networks derived from co-authorship or co-inventorship we can directly observe both the formation and the dissolution of ties as jobs and projects begin and end. Information based on grants also captures relatively early stages of the research process and publicly available abstracts provide systematic, relatively nuanced information on the content of the science being pursued with individuals and groups. Finally, grant numbers create links between collaboration networks and scientific outcomes such as publications, patents, and dissertations that acknowledge their support.

These data thus allow us to systematically define and characterize both the overall structure of collaboration networks and key features of the positions that researchers (including faculty, graduate students, and trainees) occupy in those structures. While there is much to be learned by examining structural variation across fields or campuses, here we emphasize a new approach to defining research teams based on observed collaboration networks.

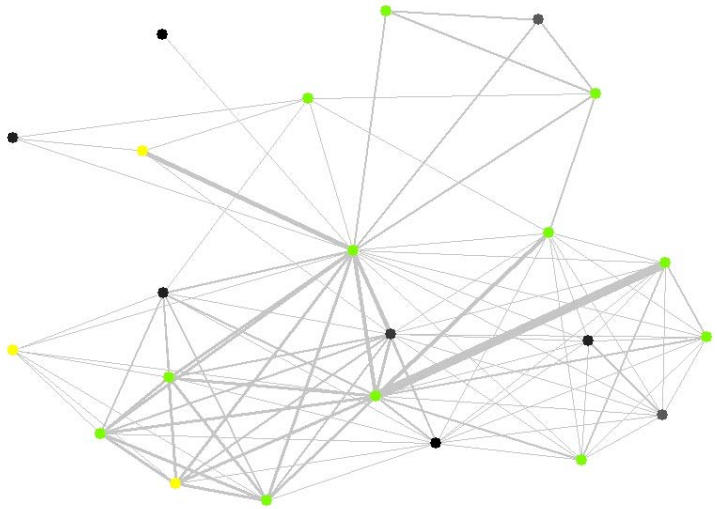
Consider three possibilities for defining a team using record level data on grants payments.

(1) *A team is all the people who are paid by the grants associated with a single PI.* This "payment basin" approach intuitively captures the standard form of research organization in science and engineering. Labs in life sciences and groups in engineering are generally identified by the last name of the PI. This also has the benefit of being fairly easy to measure with enhanced data that include the names of PIs drawn from public sources. There are some attractive conceptual features: it's easy to understand when people belong to multiple teams, for instance, and relatively simple to think about grades of membership (e.g. in terms of percentage of FTE covered by a given PI's grants).

(2) *A team is the members of a researcher's ego network.* In network parlance an ego network is the set of partners (alters) who are connected to a single, focal individual (ego) and the connections among them. In the case we describe here, the ego network of a principal investigator who was only paid by grants on which he/she was a PI would encompass all the individuals paid on his or her grants and would be equivalent to the 'payment basin' approach defined above. However for other members of the team (for instance graduate students) or for faculty who are also paid as consultants or co-investigators, an ego network measure would encompass all alters who worked on grants that also paid them. Especially in cases where students work on grants that span one or more independent PIs, this approach may more

effectively capture the social conditions of training and research than the simpler PI-centric measure proposed above.

Figure 6: Ego network for a single faculty member



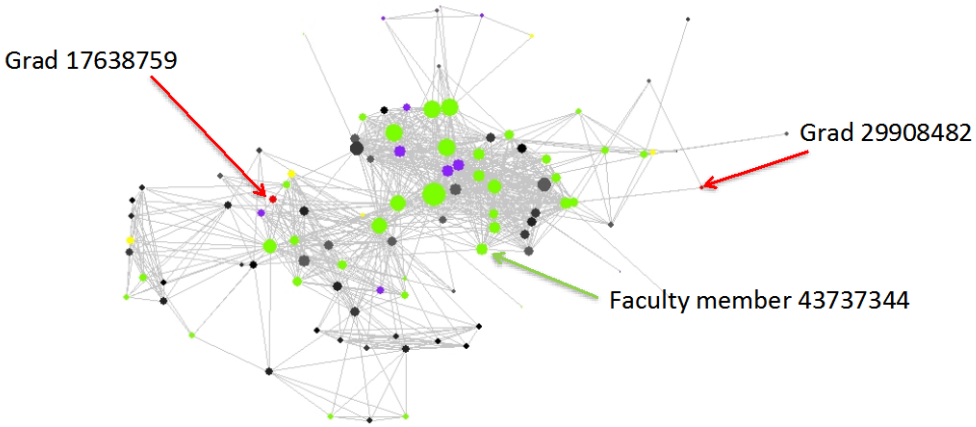
ties represents number of grants in common

As an example, consider Figure 6, which presents one faculty member's ego network. The green node in the center is "ego." That investigator has 48 ties to 22 partners who are also relatively densely connected to each other. Green nodes in this image are faculty, black nodes represent various types of staff members and yellow indicate clinicians. This particular researcher is paid by grants that (at least in 2012) employed primarily other faculty and included no students or post-doctoral associates. In the terms we develop above, ego networks based on wage payment data offer a great deal of information on the size, composition and structure of collaborative teams centered on individuals.

(3) *A team is a cohesive cluster of people who are paid by overlapping sets of grants.* Pockets of researchers in the larger collaborative network who are more densely connected to each other than they are to the rest of the network by virtue of their common involvement on shared projects may also represent teams. The number, composition, and size of such clusters (which network theorists call "communities") can be constructed for different universities and subject areas. Community-based measures offer yet another means to use grant payment information to characterize large scale factors of production in science. Where a payment-basin approach captures the essence of the 'laboratory as a firm,' and ego network approaches may offer more and more accurate information about the social organization of science as it pertains to non-faculty researchers, network communities come closest to capturing the idea that clusters of

faculty who may not be directly connected to one another may nevertheless share enough students and staff in common to constitute a cohesive cluster. This structurally defined group of researchers might best approximate a program.

Figure 7. Sample Walktrip Community



Grad Student 17638759 – Ego network includes 18 total partners of whom 1/3 are not members of this community.

Grad Student 29908482 – Ego network includes 4 total partners of whom half are noe members of this community.

Size of nodes is proportional to an individual's total network degree

Figure 7 presents a single community drawn from the network of a large public university campus. This group of 102 individuals, is defined by a pattern of ties where connections among members of the community are significantly denser than are connections from community members to other participants in the network. This group includes some 37 faculty members, two graduate students, ten post-docs and a 48 staff members. The faculty member identified by the green arrow is the "ego" from Figure 6. The densely connected cluster of "alters" at the bottom of figure six are members of the community represented in Figure 7. Most of the more sparsely connected alters at the top of Figure 6 are not. Individuals can be members of structural communities but their ego networks are imperfectly nested within those larger groups and these differences may help shape individual capabilities to develop new ideas or coordinate complicated work. Consider the two red nodes, which represent graduate students. One is more centrally connected to this community and has higher network degree. The other, with fewer ties overall, is more peripheral to this community. Such differences in individual positions relative to communities, ego networks, or the whole structure of collaboration on a campus should be telling for a student researcher's ability to complete the challenging work of graduate training.

Of course the most valuable definitions of teams might span these three starting points. For instance cohesive clusters of individuals might share or span topics while both ego networks and PI teams are

differently nested within and across communities. The internal structure or topic composition of PI-defined teams might likewise offer greater purchase on the conditions under which scientific products are developed or students are trained and mentored than is possible using grant-based metrics alone.

Implementation

In order to construct network-based measures of teams, we first must define the relevant networks. We limit ourselves to individuals paid by grants from the federal science agencies on 8 campuses for whom we have a common fiscal year (2012) of de-identified STAR METRICs data. For those individuals we construct a one mode (employeeXemployee) projection of the original two mode (employeeXgrant) network such that two employees are connected if they are paid wages by the same grant account at any point during the year. While these data support the construction of valued networks where ties represent the number of grants two individuals share in common, we binarize them for the sake of simplicity and to scale our ego network calculations in terms of the number of partners individuals have instead of the number of ties. We use a simple walk-trip community detection algorithm (Newman, 2006) to identify cohesive clusters of researchers and calculate their size within each campus-wide network. We calculate degree centrality (a measure of the number of alters in an individual's ego network) and use these measures to construct a cross-sectional dataset of network measures at the individual level for the 74,078 unique individuals who were employed on grants from federal science agencies on 8 CIC campuses in Fiscal 2012.

Preliminary Findings

Across these 8 campuses, the average ego network included 19 partners, but that mean masks the kind of highly skewed distribution that is commonplace in large-scale network data. Fully 28% of all employees on federal science grants are isolates whose degree centrality equals zero. In this network that means they were paid by grants that paid no other individuals during fiscal 2012. Isolates appear in across occupational categories, but nearly 1/3 (32.48%) are graduate students or post-docs who may be being paid by fellowships of various sorts. Another 32% of all employees have relatively small ego networks with ten or fewer partners. Nevertheless, there are dramatic outliers on every campus, suggesting that the substantive implications of different sized ego networks bears much scrutiny. For instance, a staff person who manages a shared facility may have small portions of salary paid off a large number of different grants. Similar arrangements may characterize the laboratory manager in a large department. Both these staff people will have very large ego networks, as will students who receive salary support from large training grants or faculty members who are directors of federally funded research institutes such as Science and Technology Centers or serve as PIs on institution wide grants such as NIH Clinical and Translational Science Awards. In all these cases individuals may have ego networks that number in the hundreds of partners. About 5% of all the employees we observe have ego networks with 100 or more partners, but the implications of these large networks seem likely to differ with occupation, the number of grants that creating ties, and different funding mechanisms.

Likewise, the 8 campuses we observe manifest complicated and different community structures. Across all these campuses we identify 36 communities composed of more than 100 members each. Membership in these large clusters accounts for slightly more than 10% of all employees. Membership in the kinds of large network communities highlighted in Figure 7 is highly but not perfectly correlated (r=0.741) with degree and the vast majority of high degree individuals are also members of large communities. This is not a surprising finding given the structure of our data, as large grants connect everyone they pay to each other and thus create both high degree individuals and highly cohesive

portions of a larger network structure. But this is precisely the point if we wish, for instance, to begin to understand how different funding mechanisms – for instance supporting graduate students on individual fellowships, as research assistants on investigator initiated grants, or via the mechanism of large training awards – shape the network positions and outcomes of different segments of the academic research workforce.

Future Directions.

All these preliminary descriptive findings suggest there is much work to be done. First, efforts to validate the meaning of different patterns of network connections are necessary. These data are not yet sufficient to tell us, for instance, whether students who work on individual fellowships are actually more isolated on a day-to-day basis than students employed on research grants. Likewise, additional work to overlay institutional details about doctoral programs and scientific information about the character of work being done on the projects that bind research employees together is needed. Finally, we believe that there is significant potential value in pursuing questions about the collective dynamics of these networks. One of the key implications of our early research is that funding mechanisms and sources matter for both the composition and the network structure of university research teams and as a result a more detailed understanding of how such networks form, change, and sustain themselves is necessary in order to be able to evaluate or improve the efficacy of policy choices about how to disburse scarce research funds.

5. Using the Data: Building a Community of Practice

> "The inevitable absence of policy discipline in U.S. federal government decision-making creates an imperative for some system of public education that fosters rational policy outcomes. The existence of an academic field of science of science policy is a necessary precondition for such a system. Policies can be formed and carried through rationally only when a sufficient number of men and women follow them in a deep, thoughtful, and open way. Science policy, in its broadest sense, has become so important that it deserves the enduring scrutiny from a profession of its own. This is the promise of the academic discipline of the science of science policy" (Marburger, 2011)

If the data platform we propose here is to be successful, it needs to be used and developed by a large community of practice. University administrators need to find value in the analytical output so that they continue to support the creation and production of data. Staff at the universities who provide the data need to be assured that providing data involves minimal burden. Such "data leads" are essential members of the research community because they can inform users of data quality issues, and can also identify possible new data sources. Finally, researchers from a wide spectrum of disciplines must be engaged because no team or small set of teams could possibly envision all of the interesting questions for which these data could be used. Contributions made by both individual researchers and teams can thus extend the data infrastructure, adding components that can then be used by later researchers. For example, many of the ideas being developed elsewhere in the frontier data world(Hastie et al., 2009), such as scraping and mining CVs; scraping sites such as Linked in and department or company websites could be applied here, as could many of the techniques being developed in econometrics (Varian, 2014).

In order to begin addressing the substantial data cleaning, linking and standardization challenges implicit in curating large scale administrative data, an Institute for Research on Innovation and Science (IRIS). IRIS will be established in the Spring of 2015, with seed funding from the Alfred P. Sloan and Ewing

Marion Kauffman Foundations. It will be organized around a federated model, with a core at the University of Michigan and nodes at the AIR/CIC, the University of Chicago, Ohio State University / NBER, and the U.S. Census Bureau. This model will allow us to take advantage of strengths at each institution, develop vibrant communities at these institutions, and ensure that IRIS reaches critical operating mass as quickly as possible. The nodes will serve multiple roles. They will (1) provide specific expertise in a particular topic area, (2) be responsible for outreach to and connections with specific constituencies, and (3) responsible for a particular (set of) updates/improvements to the dataset. For its part the core institution will coordinate shared research efforts, take primary responsibility for the management and maintenance of shared data resources, for the development, negotiation, and administration of data use agreements, and for the creation, validation and provenance of a yearly data release that includes the most up to date version of a shared core dataset that can be used by affiliated researchers associated with the nodes or by those who are working independently. We hope to add additional nodes as interest from the community dictates.

Dataset As a starting point, we anticipate producing an individual level dataset comprised of network and teams measures such as those sketched above derived from CIC/UMETRICS data. As the institute develops these core data will expanded to include AAU and APLU campuses. That core dataset will be linked with variables on scientific and career outcomes derived from the Research and Innovation Products database being developed in collaboration with the Chicago node and with disambiguated, longitudinal data on publications, citations, patents, dissertations, and grants that result from work done under a National Institute of Aging P01 led by Weinberg at Ohio State and NBER. In addition the dataset will be linkable (within the constraints of Title 13 and Title 26 requirements) to restricted data products and variables developed at the US Census Bureau, and to other products of the federal statistical agencies (for instance NCSES's Survey of Earned Doctorates and Survey of Doctoral Recipients). The resulting research dataset will be updated and expanded yearly and will grow to encompass new and unforeseen measures developed by additional nodes in later years. We expect to also disseminate the linked data through the Census Bureau's Research Data Center network.

An international community of practice building parallel datasets is also developing. There is considerable overlap in the activities in different countries, and the potential for substantial collaboration in many dimensions, particularly in the areas of measuring teams, the uses of topic modeling and standardization of measures, as well as studying the mobility of scientists (and scientific ideas). It is critically important to build outcome measures beyond patents and publications. The most obvious value added is to link researchers to national employer/employee data, as is being done in Norway, as well as being explored in the UK, US and Spain. As international efforts progress, greater communication and coordination is valuable.

The near decade since Marburger's initial call for data and research on the science and innovation enterprise have produced real steps toward a large-scale, widely useful data infrastructure to study science and innovation. At the same time, we are at the cusp where these investments, which have been built gradually over time, are close to bearing fruit. Tantalizing results from these researcher-focused data are beginning to emerge. New, useful findings prompt the increased investments necessary to to bring the data, infrastructure, and community to maturity. Only through continued collaboration and investments from Federal agencies, Universities, and the research community will it be possible to provide the rigorous research necessary for informed science and innovation policy.

References

Abowd, J., Haltiwanger, J., Lane, J., 2004. Integrated Longitudinal Employee-Employer Data for the United States. Am. Econ. Rev. 94, 224–229.

Arimoto, T., Sato, Y., 2012. Rebuilding public trust in science for policy-making. Science (80). 337, 1176–1177.

Azoulay, P., Ding, W., Stuart, T., 2007. The determinants of faculty patenting behavior: Demographics or opportunities? J. Econ. Behav. & Organ. 63, 599–623.

Blau, F.D., Currie, J.M., Croson, R.T.A., Ginther, D.K., 2010. Can mentoring help female assistant professors? Interim results from a randomized trial. National Bureau of Economic Research.

Cockburn, I., Henderson, R.M., 1998. Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery. J. Ind. Econ. 157–182.

Conrad, C., Cheng, W., Weinberg, B.A., 2014. University Occupation Classification: Technical Paper. Columbus, Ohio.

Cronin, B., Sugimoto, C., 2014. Scholarly Metrics Under the Microscope. Information Today Inc./ASIS&T.

Einav, L., Levin, J.D., 2013. The Data Revolution and Economic Analysis. Natl. Bur. Econ. Res. Work. Pap. Ser. No. 19035.

Fleming, R.L.A.D.A.Y.Y.S.L., 2011. Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975 - 2010).

Gertler, P., Martinez, S., Premand, P., Rawlings, L., Vermeesch, C., 2012. Impact Evaluation in Practice. The World Bank, Washington DC.

Giles, C.L., Bollacker, K.D., Lawrence, S., 1998. CiteSeer: An automatic citation indexing system, in: Proceedings of the Third ACM Conference on Digital Libraries. pp. 89–98.

Giles, L., Khabsa, M., 2014. The Number of Scholarly Documents on the Web. PLoS One.

Griffiths, T., Steyvers, M., 2004. Finding Scientific Topics. PNAS 5228–5235.

Griffiths, T.L., Steyvers, M., 2006. Probabilistic topic models, in: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (Eds.), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum.

Harrison, M., 2013. Data Consistency. Washington DC.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., 2009. The elements of statistical learning. Springer.

Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data Intensive Scientific Discovery. Microsoft Research.

Husbands Fealing, K., 2014. Assessing the outputs of government funded university research: the case of food safety and security.

Husbands Fealing, K., Lane, J., Marburger, J., Shipp, S., 2011. The Handbook of Science of Science Policy. Stanford University Press.

Jensen, P., Webster, E., 2014. Let's spend more wisely on research in Australia. Conversat.

Jones, B.F., 2009. The Burden of Knowledge and the "Death of the Renaisance Man": Is Innovation Getting Harder?No Title. Rev. Econ. Stud. 283–317.

Kabo, F.W., Cotton-Nessler, N., Hwang, Y., Levenstein, M.C., Owen-Smith, J., 2014. Proximity effects on the dynamics and outcomes of scientific collaborations. Res. Policy 43, 1469–1485.

Kenney, M., Patton, D., 2013. Guide to the Firm Database of Emerging Growth Initial Public Offerings (IPOs) from 1990 through 2010. Davis, CA.

Kerr, W.R., 2008. Ethnic scientific communities and international technology diffusion. Rev. Econ. Stat. 90, 518–537.

Khabsa, M., Treeratpitu, P., Giles, L., 2012. AckSeer: a repository and search engine for automatically extracted acknowledgments from digital libraries. Proc. 12th ACM/IEEE-CS Jt. Conf. Digit. Libr.

King, J., Lane, J., Schwarz, L., 2013. Creating New Administrative Data to Describe the Scientific Workforce: The Star Metrics Program. SSRN eLibrary.

Lane, J., 2010. Let's make science metrics more scientific. Nature 464, 488–489.

Lane, J., Bertuzzi, S., 2011. Measuring the Results of Science Investments. Science (80-.). 331, 678–680.

Lane, J., Newman, D., Rosen, R., 2013. New Approaches To Describing Research Investments [WWW Document]. bridges. URL http://www.ostina.org/index.php?option=com_content&view=article&id=6035:new-approaches-to-describing-research-investments&catid=476:feature-articles&Itemid=3377

Lane, J., Stodden, V., Bender, S., Nissenbaum, H., 2014. Privacy, big data and the public good: Frameworks for engagement. Cambridge University Press.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Yu, A.Z., Fleming, L., 2014. Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). Res. Policy 43, 941–955.

Lichtenberg, F.R., 2012. The Effect of Pharmaceutical Innovation on Longevity: Patient-Level Evidence from the 1996-2002 Medical Expenditure Panel Survey and Linked Mortality Public-Use Files. Natl. Bur. Econ. Res. Work. Pap. Ser. No. 18552.

MacIlwain, C., 2010. Science economics: What science is really worth. Nature 465, 682–684. doi:465682a [pii] 10.1038/465682a

Marburger, J., 2011. Why Policy Implementation Needs a Science of Science Policy, in: Fealing, K.H., Lane, J., Shipp, S. (Eds.), The Handbook of Science of Science Policy, Stanford University Press.

Marburger, J.H., 2005. Wanted: Better Benchmarks. Science (80-.). 308, 1087–.

Mateos, P., 2007. A review of name-based ethnicity classification methods and their potential in population studies. Popul. Space Place 13, 243–263.

Mateos, P., 2014. Ethnicity, geography and populations: Tracing diversity and migration through people's names. Springer Netherlands, Berlin.

Mody, C., 2004. How Probe Microscopists Became Nanotechnologists, in: Baird, D., Nordmann, A., Schummer, J. (Eds.), Discovering the Nanoscale. IOS Press, Amsterdam, pp. 119–133.

Mole, B., 2013. NSF cancels political-science grant cycle. Nat. News.

National Academies, 2005. Policy Implications of International Graduate Students and Postdoctoral Scholars in the United States. The National Academies Press.

National Academy of Sciences, 2014. Furthering America's Research Enterprise. The National Academies Press, Washington DC.

National Research COuncil, 2010. Data on Federal Research and Development Investments: A Pathway to Modernization.

National Science and Technology Council, 2008. The Science of Science Policy: A Federal Research Roadmap. National Science and Technology Council,Science of Science Policy Interagency Task Group, Washington, D.C.

National Science Foundation, 2011. Report to the Advisory Committees of the Directorates of Computer and Information Science and Engineering and Social, Behavioral and Economic Sciences. National Science Foundation.

Nelson, L., Sedwick, S.W., 2011. STAR METRICS: A Participant's Perspective. NCURA Mag. 43, 24–25.

Newman, M.E.J., 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. U. S. A. 103, 8577–82. doi:10.1073/pnas.0601602103

Owens, B., 2013. JUDGEMENT DAY.

Owen-Smith, J., Powell, W., 2004. Knowledge Networks as Channels and Conduits: The Effects of Spillovers in the Boston Biotechnology Communit. Organ. Sci.

Powell, W., Giannella, E., 2010. Collective Invention and Inventor Networks. The Handbook of Innovation, in: The Handbook of Innovation. Elsevier, Amsterdam.

Shapin, S., 1989. The Invisible Technician. Am. Sci. 77, 554–563.

Smalheiser, N.R., Torvik, V., 2009. Author Name disambiguation. Annu. Rev. Inf. Sci. Technol. 1–43.

Torvik, V., Weeber, M., Swanson, D., Smalheiser, N., 2005a. A probabilistic similarity metric for Medline records: A model for author name disambiguation. J. Am. Soc. Inf. Sci. Technol. 56.

Torvik, V., Weeber, M., Swanson, D., Smalheiser, N., 2005b. A probabilistic similarity metric for Medline records: A model for author name disambiguation. J. Am. Soc. Inf. Sci. Technol. 56.

Varian, H.R., 2014. Big data: New tricks for econometrics. J. Econ. Perspect. 28, 3–27.

Ventura, S.I., Nugent, R., Fuchs, E.R.H., 2014. Seeing the non-starts: (Some) sources of bias in past dismbiguation approaches and a new public tool leveraging labeled records (No. 2079930).

Walsh, D., 2013. NOT SAFE FOR FUNDING: THE N.S.F. AND THE ECONOMICS OF SCIENCE. New Yorker.

Wuchty, S., Jones, B.F., Uzzi, B., 2007. The Increasing Dominance of Teams in Production of Knowledge. Science (80-.). 316, 1036–1039. doi:10.1126/science.1136099

Zucker, L., Darby, M., 2006. Movement of Star Scientists and Engineers and High Tech Firm

MAPPING STARMETRICS AGGREGATE OCCUPATIONAL CATEGORIES

FACULTY

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

CLINICIAN

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

POSTDOCTORAL RESEARCH

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

GRADUATE STUDENT

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

UNDERGRADUATE STUDENT

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

STAFF SCIENTIST/TECHNICIAN

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

RESEARCH ANALYST/COORDINATOR

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	

RESEARCH SUPPORT

	Permanence [P]	Research Role [R]	Track [T]	Scientific Training [S]	Clinical Association [C]
1	Permanent	Provide oversight	Professorial track	Advanced degree	Clinical
2	Non-Permanent	Generate new ideas	Non-Professorial track	Some scientific training	Non-clinical
3		Provide support	No track	No scientific training	