IRIS Research Bibliography

PUBLISHED PAPERS

Women are Credited Less in Science than are Men

Ross M, Glennon B, Murciona-Goroff R, Berkes E, Weinberg B, Lane J

Nature, 22 June 2022

https://doi.org/10.1038/s41586-022-04966-w

Abstract

There is a well-documented gap in the observed number of scientific works produced by women and men in science, with clear consequences for the retention and promotion of women in science. The gap might be a result of productivity differences, or it might be due to women's contributions not being acknowledged. This paper finds that at least part of this gap is due to the latter: women in research teams are significantly less likely to be credited with authorship than are men. The findings are consistent across three very different sources of data. Analysis of the first source - large scale administrative data on research teams, team scientific output, and attribution of credit - show that women are significantly less likely to be named on any given article or patent produced by their team relative to their peers. The gender gap in attribution is found across almost all scientific fields and career stages. The second source – an extensive survey of authors – similarly shows that women's scientific contributions are systematically less likely to be recognized. The third source – qualitative responses – suggests that the reason is that their work is often not known, not appreciated, or ignored. At least some of the observed gender gap in scientific output may not be due to differences in scientific contribution, but to differences in attribution.

The Ripple Effects of Funding on Researchers and Output

Sattari R, Bae J, Berkes E, Weinberg B

ScienceAdvances, 22 April 2022

https://doi.org/10.1126/sciadv.abb7348

Abstract

Using unique, new, matched UMETRICS data on people employed on research projects and Authority data on biomedical publications, this paper shows that National Institutes of Health funding stimulates research by supporting the teams that conduct it. While faculty—both

principal investigators (PIs) and other faculty—and their productivity are heavily affected by funding, so are trainees and staff. The largest effects of funding on research output are ripple effects on publications that do not include PIs. While funders focus on research output from projects, they would be well advised to consider how funding ripples through the wide range of people, including trainees and staff, employed on projects.

Benchmarking university technology transfer performance with external research funding: a stochastic frontier analysis

Coupet J, Ba Y

Journal of Technology Transfer, 2021

https://doi.org/10.1007/s10961-021-09856-3

Abstract

Many universities engage in academic entrepreneurship, often with funding from external sources. Benchmarking technology transfer performance with external research funding can help universities identify and learn from peers that may possess strategic advantages in productivity. It also can be key for organizational learning and for communicating organizational performance to policy stakeholders and industry partners. In this study, we construct a unique dataset by linking two important data sources, AUTM and UMETRICS, and use stochastic frontier analysis to benchmark university licensing and revenue performance with different federal funding streams. Our empirical results suggest that universities looking to promote commercialization performance might look to National Science Foundation funding, and the universities best at production (i.e., licensing technologies and generating patents) with external funding are not necessarily the best at capturing benefits from generating revenue from entrepreneurial activity and launching start-ups. Our study points to the importance of the differential advantages of sources of federal research funding and offers implications for policy makers and university administrators.

The U.S. Academic Research Enterprise (US-ARE): Possible Paths from the Pandemic

Owen-Smith J

Springer Nature, 21 October 2020

https://www.springernature.com/gp/researchers/campaigns/coronavirus/impact-of-covid19

Abstract

This white paper uses recent public data to identify what we can know systematically about how the COVID-19 pandemic is currently affecting the large, research-intensive universities that

represent the core of the US-ARE. It uses those, admittedly preliminary and partial, findings to extrapolate about possible long-term effects of decisions that academic leaders, state and federal policy makers are taking right now. The descriptive story presented here isn't determinative, but it suggests that the pandemic poses unique dangers for the national and global research systems.

A New Approach for Estimating Research Impact: An Application to French Cancer Research

Chevalier G, Chomienne C, Jeanrenaud N, Lane J, Ross M

Quantitative Science Studies, 24 Aug. 2020
https://doi.org/10.1162/qss_a_00087

Abstract

Much attention has been paid to estimating the impact of investments in scientific research. Historically, those efforts have been largely ad hoc, burdensome, and error prone. In addition, the focus has been largely mechanicale—drawing a direct line between funding and outputs—rather than focusing on the scientists that do the work. Here, we provide an illustrative application of a new approach that examines the impact of research funding on individuals and their scientific output in terms of publications, citations, collaborations, and international activity, controlling for both observed and unobserved factors. We argue that full engagement between scientific funders and the research community is needed if we are to expand the data infrastructure to enable a more scientific assessment of scientific investments.

Modernizing U.S. Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building

Hart N, Potok N

Data Foundation, 20 July 2020
https://ssrn.com/abstract=3700156

Abstract

The need for using data to generate insights that can help improve American society is vast and urgent. Increasingly, researchers need capabilities to link together data collected through formal surveys, federal program administration, and non-governmental data sources. However, the lack of coordination throughout the federal government’s decentralized data infrastructure

and statistical system limits the ability to generate the relevant, timely information demanded by policymakers.

Building on the initial recommendations from consensus panels of experts in recent years, we propose a strategy for developing a National Secure Data Service that would revolutionize the federal government's data analysis capabilities, while promoting and even expanding privacy protections available today. The data service would modernize the country's antiquated, inefficient, and often ineffective data infrastructure for research to develop a modern, cutting-edge system that would substantially advance evidence-based policy-making capabilities in the United States.

The Color of Money: Federal vs. Industry Funding of University Research

Babina T, He A, Howell S, Perlman E, Staudt J

Nov. 2020

https://ssrn.com/abstract=3560195

Abstract

U.S. universities have experienced a shift in research funding away from federal and towards private industry sources. This paper evaluates whether the source of funding – federal or private industry – is relevant for commercialization of research outputs. We link person-level grant data from 22 universities to patent and career outcomes (including IRS W-2 records). To identify a causal effect, we exploit individual-level variation in exposure to narrow federal R&D programs stemming from pre-existing field specialization. We instrument for the researcher's funding sources with aggregate supply shocks to federal funding within these narrow fields. The results show that a higher share of federal funding reduces patenting and the chances of joining an incumbent firm, while increasing the chances of high-tech entrepreneurship and of remaining employed in academia. A decline in the federal share of funding is offset by an increase in the private share of funding, which has opposite effects. We conclude that the incentives of private funders to appropriate research outputs have important implications for the trajectory of university researcher careers and intellectual property.

Money for Something: Braided Funding and the Structure and Output of Research Groups

Funk R, Glennon B, Lane JI, Murciano-Goroff R, Ross M

IZA Discussion Paper No. 12762, 18 Nov 2019
https://ssrn.com/abstract=3488189

Abstract

In 2017, the federal government invested over $40 billion on university research; another $16 billion came from private sector sources. The expectation is that these investments will bear varied fruits, including outputs like more economic growth, more scientific advances, the training and development of future scientists, and a more diverse pipeline of STEM researchers; an expectation that is supported by the work of recent Nobel Laureate in Economics, Paul Romer. Yet volatility in federal funding, highlighted by a 35 day federal shutdown in early 2019, has resulted in an increased interest on the part of scientists in finding other sources of funding. Understanding the effect of such different funding streams on research outputs is thus of more than academic importance, particularly because there are likely to be tradeoffs, both in terms of the structure of research and in terms of research outputs. For example, federal funding is often intended to affect the structure of research, with explicit goals of training the next generation of scientists and promoting diversity; those goals are less salient for non-federal funding. On the output side, federally funded research may be more likely to emphasize producing purely scientific outputs, like publications, rather than commercial outputs, like patents. The contribution of this paper is to use new data to examine how different sources of financial support – which we refer to as "braided" funding – affect both the structure of scientific research and the subsequent outputs.

Data Specific Functions: A Comment on Kindel et al.

Fisher J

Socius: Sociological Research for a Dynamic World, 20 Sept. 2019
https://doi.org/10.1177/2378023118822893

Abstract

In this issue, Kindel et al. describe a new approach to managing survey data in service of the Fragile Families Challenge, which they call “treating metadata as data.” Although the approach they present is a good first step, a more ambitious proposal could improve survey data analysis even more substantially. The author recommends that data collection efforts distribute an open-source set of tools for working with a particular data set the author calls data-specific

functions. The goal of these functions is to codify best practices for working with the data in a set of functions for commonly used statistical software. These functions would be jointly developed by the users and distributers of the data. Building such functions would both shorten the learning curve for new users and improve the quality of the data, by making tacit knowledge about problems with the data explicit and easy to act on.

Executing Entity Matching End to End: A Case Study

Konda P, Seshadri S, Segarra E, Hueth B, Doan A

Proceedings of the 22nd International Conference on Extending Database Technology (EDBT), March 26-29, 2019, I

https://pages.cs.wisc.edu/~anhai/papers1/umetrics-edbt19.pdf

Abstract

Entity matching (EM) identifies data instances that refer to the same real-world entity. Numerous EM works have covered a wide spectrum, from developing new EM algorithms to scaling them to building EM systems. But there has been very little if any published work on how EM is carried out in practice, end to end. In this paper we describe in detail a case study of applying EM to a particular domain end to end (i.e., going from the raw data all the way to the matches). Specifically, we describe a real-world application for EM in the science policy research community. We describe how our team (the EM team) interact with the science policy team to carry out the EM process, using PyMatcher, a state-of-the-art EM system developed in the Magellan project at UW-Madison. We highlight the communication between the two teams and the zig-zag nature of the EM process. We identify a set of challenges that we believe arise in many real-world EM projects but that current EM systems have either ignored or are not even aware of. Finally, we provide all data underlying this case study, including labeled tuple pairs and documentation supplied by the science policy team, to serve as a good challenge problem for EM researchers.

Federal Funding of Doctoral Recipients: What Can Be Learned From Linked Data

Chang W, Cheng W, Lane JI, & Weinberg BA
Research Policy available online 14 March 2019
https://doi.org/10.1016/j.respol.2019.03.001

Abstract
This technical note describes the results of a pilot approach to link administrative and survey data to better describe the richness and complexity of the research enterprise. In particular, we demonstrate how multiple funding channels can be studied by bringing together two disparate datasets: UMETRICS, which is based on university payroll and financial records, and the Survey of Earned Doctorates (SED), which is one of the most important US survey datasets about the doctoral workforce. We show how it is possible to link data on research funding and the doctorally qualified workforce to describe how many individuals are supported in different disciplines and by different agencies. We outline the potential for more work as the UMETRICS data expands to incorporate more linkages and more access is provided.

A Fast and Integrative Algorithm for Clustering Performance Evaluation in Author Name Disambiguation

Kim, Jinseok
***Scientometrics* 120, 661-681 (2019).**
https://doi.org/10.1007/s11192-019-03143-7

Abstract
Clustering results in author name disambiguation are often evaluated by measures such as Cluster-F, K-metric, Pairwise-F, Splitting and Lumping Error, and B-cubed. Although these measures have different evaluation approaches, this paper shows that they can be calculated in a single framework by a set of common steps that compare truth and predicted clusters through two hash tables recording information about name instances with their predicted cluster indices and frequencies of those indices per truth cluster. This integrative calculation reduces greatly calculation runtime, which is scalable to a clustering task involving millions of name instances within a few seconds. During the integration process, B-cubed and K-metric are shown to produce the same precision and recall scores. In addition, name instance pairs for Pairwise-F are counted using a heuristic, which enables the proposed method to surpass a state-of-the-art algorithm in speedy calculation. Details of the integrative calculation are

described with examples and pseudo-code to assist scholars to implement each measure easily and validate the correctness of implementation. The integrative calculation will help scholars compare similarities and differences of multiple measures before they select ones that characterize best the clustering performances of their disambiguation methods.

Formational Bounds of Link Prediction in Collaboration Networks

Kim, Jinseok, & Diesner, Jana
***Scientometrics* 119, 687-706 (2019).**
https://doi.org/10.1007/s11192-019-03055-6

Abstract
Link prediction in collaboration networks is often solved by identifying structural properties of existing nodes that are disconnected at one point in time, and that share a link later on. The maximally possible recall rate or upper bound of this approach's success is capped by the proportion of links that are formed among existing nodes embedded in these properties. Consequentially, sustained links as well as links that involve one or two new network participants are typically not predicted. The purpose of this study is to highlight formational constraints that need to be considered to increase the practical value of link prediction methods targeted for collaboration networks. In this study, we identify the distribution of basic link formation types based on four large-scale, over-time collaboration networks, showing that roughly speaking, 25% of links represent continued collaborations, 25% of links are new collaborations between existing authors, and 50% are formed between an existing author and a new network member. This implies that for collaboration networks, increasing the accuracy of computational link prediction solutions may not be a reasonable goal when the ratio of collaboration links that are eligible to the classic link prediction process is low.

Occupational Classifications: A Machine Learning Approach

Ikudo A, Lane J, Staudt J, and Weinberg BA

***Journal of Economic and Social Measurement* 44 (Issue 2-3): 57-87 (2020)**
DOI: 10.3233/JEM-190463

Abstract

Characterizing people's occupations is important for both policy and research. However, as large-scale administrative records are increasingly being used to describe labor market activity, it will become important to find new automated approaches to describing occupations. We apply new machine learning techniques to new sources of data and investigate the potential of using algorithms to classify occupations. We find that job titles are both inherently noisy and inconsistent across organizations, but a subset of them can be assigned algorithmically, with little impact on accuracy.

Generating Automatically Labeled Data for Author Name Disambiguation: An Iterative Clustering Method.

Kim, Jinseok, Kim, Jinmo, & Owen-Smith, Jason

***Scientometrics* 118, 253–280 (2019)**
https://doi.org/10.1007/s11192-018-2968-3

Abstract

Many author name disambiguation studies have relied on hand-labeled truth data that are very laborious to generate. This paper shows that labeled data can be automatically generated using information features such as email address, coauthor names, and cited references that are available from publication records. For this purpose, high-precision rules for matching name instances on each feature are learned using an external-authority database. Then, selected name instances in target ambiguous data go through the process of pairwise matching based on the learned rules. Next, they are merged into blocks by a generic entity resolution algorithm. The blocking procedure is repeated over other features until further merging is impossible. Tested on an example of 26,566 name instances, this iterative blocking produced accurately labeled data with near perfect accuracy (pairwise F1 = 0.99). In addition, the labeled data represented the population data of 227K name instances in terms of name ethnicity and co-disambiguating name group size distributions. Several challenges are discussed for applying this method to resolving author name ambiguity in large-scale scholarly data.

Lessons Learned from the Creation of Administrative Data Centers for Government Data

Lane J

Sloan Foundation, Securely Sharing Data, 2019

https://securelysharingdata.com/lane.html

Abstract

What lessons can be drawn from the existing efforts to make government administrative data available to researchers? In what ways are those efforts applicable to third party, private sector data such as micro-level social media data, mobile phone data? And in what ways do social media and similar data pose distinctive challenges.

We have learned a great deal about what works and what doesn't in terms of sharing government administrative data. In this paper I argue that we have learned that a successful system has to be designed with both value and sustainability at its core. I argue that much can be learned from studying data driven companies in the private sector—Google, Amazon, Facebook and Apple – which derive their market power from collecting, curating and using massive amounts of data to produce products in demand and hiring the best and brightest staff to do so.

Successful efforts have created value at the local, regional and state level by making full use of the massive computing power, statistics and human and artificial intelligence now available. They are also characterized by the creation of new products developed as a result of the interaction of the people who will use the data, researchers, and analysts.

Research-Portfolio Performance Metrics: Rapid Review

Blumenthal M, Taylor J, Leidy E, Anderson B, Carew D, Bordeaux J, Shanley M

Rand Corp., 2019

http://www.rand.org/t/RR2370

Abstract

The effectiveness of research, like that of other activities, can be evaluated at different levels — the individual project, a group of projects or program, or a larger grouping that might include multiple programs (a portfolio). Focusing on options for research portfolio evaluation, RAND Corporation researchers found many metrics in use or recommended for federal agencies and private, research-supporting organizations and organized them in a taxonomy. This report presents the characteristics and utility of these metrics, organized by individual stages in a logic-model framework, mapping portfolio metrics to the upstream stages of inputs, processes, and outputs and the downstream stages of outcomes and impacts. At each stage, categories of

metrics are composed of sets of metric types, each of which is, in turn, composed of individual metrics. In addition to developing this taxonomy, the authors appraised key attributes of portfolio evaluation metrics and described the trade-offs associated with their use. This structured, annotated compilation can help the Defense Health Agency and other entities that evaluate research portfolios to select, develop, or revise the metrics they use.

Postsecondary Data Infrastructure: What is Possible Today

O'Hara, A

Institute for Higher Education Policy, June 2019

http://hdl.handle.net/10919/95136

Abstract

Our current federal postsecondary data system is incomplete and fails to provide today's students with the accurate, timely information that they need to inform their college choices and promote their success. As policymakers consider proposals to improve this federal data system, they should model best practices of responsible data-use to ensure that all postsecondary data is secure and student privacy is protected.

This paper, authored by Amy O'Hara, Research Professor at the Georgetown University Massive Data Institute, highlights promising examples of data systems that are prioritizing privacy and security. These examples span from government agencies to academia and cover sectors ranging from healthcare to national defense. In this analysis, O'Hara uses a "Five Safes" framework as an approach for guiding secure data practices:

- Safe projects require governance protocols to control project requests, review, and approval processes;
- Safe people ensure that data users are screened and appropriately trained;
- Safe settings and safe data restrict what data an analyst is authorized to use, how they access it, their computing environment, and their physical location; and
- Safe outputs protect the privacy of data subjects by reducing the risk of individuals being re-identified.

The Impact of Imbalanced Training Data on Machine Learning for Author Name Disambiguation

Kim, Jinseok & Kim, Jenna
***Scientometrics* 117, 511–526 (2018).**
https://doi.org/10.1007/s11192-018-2865-9

Abstract

In supervised machine learning for author name disambiguation, negative training data are often dominantly larger than positive training data. This paper examines how the ratios of negative to positive training data can affect the performance of machine learning algorithms to disambiguate author names in bibliographic records. On multiple labeled datasets, three classifiers – Logistic Regression, Naïve Bayes, and Random Forest – are trained through representative features such as author name, coauthor names, and title words extracted from the same training data but with various positive-to-negative training data ratios. Results show that increasing negative training data can improve disambiguation performance but with a few percent of performance gains and sometimes degrade it. Logistic Regression and Naïve Bayes learn optimal disambiguation models even with a base ratio (1:1) of positive and negative training data. Also, the performance improvement by Random Forest tends to quickly saturate roughly after 1:10 ~ 1:20. These findings imply that contrary to the common practice using all training data, name disambiguation algorithms can be trained using part of negative training data without degrading much disambiguation performance while increasing computational efficiency. This study calls for more attention from author name disambiguation scholars to methods for machine learning from imbalanced data.

Evaluating Author Name Disambiguation for Digital Libraries: A Case of DBLP

Kim, Jinseok
***Scientometrics* 116, 1867–1886 (2018).**
https://doi.org/10.1007/s11192-018-2824-5.

Abstract

Equipped with advanced computing techniques, scholars have disambiguated author names in whole digital libraries and tested their performances in various ways. The purpose of this study is to propose a triangulation approach that author name disambiguation for digital libraries can be better evaluated when its performance is assessed on multiple labeled datasets with comparison to baselines for diverse ambiguity dimensions. To illustrate the proposed approach, accuracy of author name disambiguation in DBLP's 3.7M records is evaluated on three types of

labeled data containing 5,000 to 6M disambiguated names. Results show that the triangulation method can provide a more holistic, granulated understanding of a disambiguation method's performance than common evaluation practices in prior studies. With the review of strengths and weaknesses of the proposed approach, this study calls for further discussion about consistent frameworks and methodologies for evaluating author name disambiguation so that findings from a variety of studies can be synthesized to produce insights for improving name ambiguity resolution for fast-growing digital libraries.

Author-based Analysis of Conference Versus Journal Publication in Computer Science

Kim, Jinseok

Journal of the Association for Information Science and Technology, 70: 71-82 (2018).
DOI: https://doi.org/10.1002/asi.24079

Abstract

Conference publications in computer science (CS) have attracted scholarly attention due to their unique status as a main research outlet, unlike other science fields where journals are dominantly used for communicating research findings. One frequent research question has been how different conference and journal publications are, considering an article as a unit of analysis. This study takes an author-based approach to analyze the publishing patterns of 517,763 scholars who have ever published both in CS conferences and journals for the last 57 years, as recorded in DBLP. The analysis shows that the majority of CS scholars tend to make their scholarly debut, publish more articles, and collaborate with more coauthors in conferences than in journals. Importantly, conference articles seem to serve as a distinct channel of scholarly communication, not a mere preceding step to journal publications: coauthors and title words of authors across conferences and journals tend not to overlap much. This study corroborates findings of previous studies on this topic from a distinctive perspective and suggests that conference authorship in CS calls for more special attention from scholars and administrators outside CS who have focused on journal publications to mine authorship data and evaluate scholarly performance.

Proximity and Economic Activity: An Analysis of Vendor-University Transactions

Goldschlag N, Lane JI, Weinberg BA , Zolas N

***Journal of Regional Science,* 2018: 1-20**
DOI: 10.111/jors.12397 https://onlinelibrary.wiley.com/doi/abs/10.1111/jors.12397

Abstract

This paper uses transaction-based data to provide new insights into the link between the geographic proximity of businesses and associated economic activity. It develops two new measures of, and a set of stylized facts about, the distances between observed transactions between customers and vendors for a research-intensive sector. Spending on research inputs is more likely with businesses physically closer to universities than those further away. Firms supplying a university project in one year are more likely to subsequently open an establishment near that university. Vendors who have supplied a project, are subsequently more likely to be a vendor on the same or related project.

Why the U.S. Science and Engineering Workforce is Aging Rapidly

Blau D, & Weinberg BA

***Proceedings of the National Academy of Sciences*, 14 February 2017**
Vol. 114(15), 3879-3884 DOI: 10.1073/pnas.16117481114/-/DCSupplemental
http://www.pnas.org/content/114/15/3879.short

Abstract

The science and engineering workforce has aged rapidly in recent years, both in absolute terms and relative to the workforce as a whole. This is a potential concern if the larger number of older scientists crowds out younger scientists, making it difficult for them to establish independent careers. In addition, scientists are believed to be most creative earlier in their careers, so the aging of the workforce may slow the pace of scientific progress. The authors developed and simulated a demographic model, which shows that a substantial majority of recent aging is a result of the aging of the large baby boom cohort of scientists. However, changes in behavior have also played a significant role, in particular a decline in the retirement rate of older scientists, induced in part by the elimination of mandatory retirement in universities in 1994. Furthermore, the age distribution of the scientific workforce is still adjusting. Current retirement rates and other determinants of employment in science imply a steady-state mean age 2.3 years higher than the 2008 level of 48.6.

STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census

Buffington C, Cerf B, Jones C, & Weinberg BA

American Economic Review **May 2016**
106(5), pp. 333–338 DOI: 10.1257/aer.p20161124
https://www.aeaweb.org/articles?id=10.1257/aer.p20161124

Abstract
Women are underrepresented in science and engineering, with the underrepresentation increasing in career stage. We analyze gender differences at critical junctures in the STEM pathway–graduate training and the early career–using UMETRICS administrative data matched to the 2010 Census and W-2s. We find strong gender separation in teams, although the effects of this are ambiguous. While no clear disadvantages exist in training environments, women earn 10% less than men once we include a wide range of controls, most notably field of study. This gap disappears once we control for women's marital status and presence of children.

Wrapping It Up in a Person: Examining Employment and Earnings Outcomes for Ph.D. Recipients

Zolas N, Goldschlag N, Jarmin RS, Stephan P, Owen- Smith J, Rosen RF, McFadden Allen B, Weinberg BA, & Lane JI

Science **11 December 2015**
Vol. 350(6266), *pp. 1367-1371*
DOI: 10.1126/science.aac5949
http://www.sciencemag.org/content/350/6266/1367.full
Supplementary material: http://econ.ohio-state.edu/weinberg/Science-aac5949_Zolas-SM-PUBLISHED.pdf

Abstract
In evaluating research investments, it is important to establish whether the expertise gained by researchers in conducting their projects propagates into the broader economy. For eight universities, it was possible to combine data from the UMETRICS project, which provided administrative records on graduate students supported by funded research, with data from the U.S. Census Bureau. The analysis covers 2010–2012 earnings and placement outcomes of people receiving doctorates in 2009–2011. Almost 40% of supported doctorate recipients, both federally and nonfederally funded, entered industry and, when they did, they

disproportionately got jobs at large and high-wage establishments in high-tech and professional service industries. Although Ph.D. recipients spread nationally, there was also geographic clustering in employment near the universities that trained and employed the researchers. We also show large differences across fields in placement outcomes.

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

Lane JI, Owen-Smith J, Rosen RF, & Weinberg BA

Research Policy **December 2014**
Vol. 44(9), pp. 1659-1671 DOI: 10.1016/j.respol.2014.12.013
http://www.sciencedirect.com/science/article/pii/S0048733315000025

Abstract
Longitudinal micro-data derived from transaction level information about wage and vendor payments made by Federal grants on multiple US 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.

Science Funding and Short-Term Economic Activity

Weinberg BA, Owen-Smith J, Rosen RF, Schwarz L, McFadden Allen B, Weiss RE, & Lane JI

Science **4 April 2014**
Vol. 344(6179), pp. 41-43 DOI: 10.1126/science.1250055
http://www.sciencemag.org/content/344/6179/41.full

Abstract
There is considerable interest among policy-makers in documenting short-term effects of science funding. A multiyear scientific journey that leads to long-term fruits of research, such as

a moon landing, is more tangible if there is visible nearer-term activity, such as the presence of astronauts. Yet systematic data on such activities have not heretofore existed. The only source of information for describing the production of most science is surveys that have been called “a rough estimate, frequently based on unexamined assumptions that originated years earlier.

WORKING PAPERS

Invisible and gendered labor in biomedical research

Del Junco C, DeJong S, Nicholls N, Evans J

Abstract

We are developing a classification system for non-academic employees in the IRIS data. Using this classification system and matching employee records with PubMed, we will quantify author credit for biomedical research as a function of the job class and the fraction of female-imputed employees in the job class.

Analyzing NIH KL2 Outcomes: A Pilot Study Using Administrative Data

VanEseltine M, Calvin-Naylor N, Owen-Smith J
https://doi.org/10.31235/osf.io/s7e3z

Abstract

Background: The U.S. National Institutes of Health (NIH) funds “K” awards that provide both resources and access to mentoring believed to be invaluable for early career faculty. The KL2 Mentored Career Development Award trains early-career clinicians with the goal of guiding scholars toward an independent clinical and translational research career. This study presents the pilot of a systematic, low-burden method to examine scientific and career outcomes for these awardees, applying a novel set of linked administrative data.

Methods: Clinical and Translational Science Award hubs administering KL2 awards at ten universities who participate in the Institute for Research on Innovation and Science (IRIS) provided names of scholars in their KL2 cohorts. Using extensive data on sponsored projects which IRIS member universities provide, we linked the KL2 scholars to information on subsequent publication, patent, and grant activity.

Results: Analyses of linked data supported a rigorous, sustainable, low-cost approach to examining career outcomes. A subset of key metrics identified by CTSA evaluators were operationalized as an examination of the post-award careers of KL2 awardees. We successfully identified contemporaneous faculty with different NIH K Awards to use as comparison groups. The pilot culminated in university-specific and aggregate reporting to all participating hubs.

Conclusions: This pilot demonstrates that substantive evaluations of early career programs are possible using administrative data from universities with low additional burden. Integration of research on career development outcomes offer new means to examine the effects of

increasingly diverse funding, team, and collaborative network structures, advancing both knowledge about the workings of science and practices to support early career faculty. This approach could be extended to support rigorous multi-institutional evaluation and research on a range of student and faculty training mechanisms.

Dissertators with Distantly Related Foci Face Divergent Near-Term Outcomes

Kniffin K, Hanks A, Qian X, Wang B, Weinberg BA
NBER Working Paper No. 27825
http://www.nber.org/papers/w27825

Abstract
Institutional leaders have long championed interdisciplinary research; however, researchers have paid relatively little attention to the people responding to such calls and their subsequent career outcomes. With the benefit of two large datasets spanning from 1986 through 2016, we show that interdisciplinary dissertations have become consistently more common in recent years as institutional leaders have highlighted the value of boundary-spanning research for solving important and emergent problems. With the benefit of survey data from a near-complete population of all dissertators in the US starting in 2001 through 2016, we observe a consistent upward trend in interdisciplinary dissertations. Unfortunately, we show that these interdisciplinary dissertators have experienced a comparably persistent penalty when considering salaries for their first year after earning the PhD. We also show that among interdisciplinary dissertators, individuals in lower-paying fields tend to earn more when choosing distantly related topic-combinations whereas researchers in higher-paying fields tend to be most rewarded for staying within relatively narrow disciplinary silos.

Local Fiscal Multiplier on R&D and Science Spending: Evidence from the American Recovery and Reinvestment Act

Chhabra, Yulia and Levenstein, Margaret C. and Owen-Smith, Jason
Ross School of Business Paper No. 1383
SSRN: https://ssrn.com/abstract=3201136 and http://hdl.handle.net/2027.42/144514. Under review at *American Economic Journal: Economic Policy*.

Abstract
We use the American Recovery and Reinvestment Act (ARRA), a large stimulus package passed into law to combat the Great Recession, to estimate the effect of R&D and science spending on local employment. Unlike most fiscal stimuli, the R&D and science portion of ARRA did not target counties with poor economic conditions but rather was awarded following a peer review

process, or based on innovative potential and research infrastructure. We find that, over the program's five-year disbursement period, each one million USD in R&D and science spending was associated with twenty-seven additional jobs. The estimated job-year cost is about $15,000.

How Universities Organize Science

Fisher, Jacob C. & Owen-Smith, Jason (2018)

Abstract

Although the results of science -- publications and patents -- have received considerable attention, little work to date has considered how the production of science is organized. Using a unique dataset on grant payments to faculty, staff, and trainees within 23 universities, we explore how universities approach a similar task, developing research findings, in different ways. Drawing on organizational theory that suggests that work is accomplished through a network of collaborations, we examine two complementary processes that cause the organization of science to differ between universities. First, administrators and grantors can control the number and occupation of people involved in the network. Second, individual faculty members can control the specific collaboration relationships, both between faculty members, and among staff and trainees who receive funding from particular grants. In network terms, administrators and grantors control the vertices, and individual faculty control the edges between them. We find that the influence of administrators and grantors is most visible along two dimensions: the amount of funding awarded by NIH, and the ratio of trainees to staff. A cluster analysis demonstrates that individual faculty staff grants in one of six ways, which depend on the scale of the grant and the faculty member's preferences. We find that between-university differences in the connectivity of the network can largely be explained by differences in scale, differences in clustering can be explained by faculty preferences, but overall differences in structure of the networks cannot be well-explained by either scale or collaboration preferences.

The link between R&D, human capital and business startups

Goldschlag N, Jarmin RS, Lane JI, & Zolas N

Presented at American Economic Association Meeting, Chicago, January 2017
Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)
NGER CRIW-The Measurement and Diffusion of Innovation (Corrado C, Sichel D, and Miranda J, eds)

Abstract

The reason for the secular decline in entrepreneurship is not well understood. It is evident in all sectors of the economy and almost all regions. One approach to stimulating innovation and entrepreneurship has been to increase investments in science: the U.S. federal government contributed nearly $38 billion for university-based research in Science, Technology, Engineering, and Mathematics (STEM) in 2014. However, there has historically been little evidence about the links between investments in university research and innovation - largely because surveys cannot capture the complex ways in which scientific ideas are created, transmitted and adopted.

This paper examines the relationship between the funding of research teams - in terms of structure, field and type of funding - and the subsequent propensity of members of those teams to start up businesses. It also examines the subsequent survival and productivity growth of those startups.

The work is now possible because of a new data infrastructure resulting from collaborations between the Census Bureau's Innovation Measurement Initiative, the National Science Foundation and the Institute for Research on Innovation and Science at the University of Michigan. The infrastructure links universe data on all people employed on research grants, their funding, and their economic and scientific activities.

This paper is the first to directly trace the pathways from the bench to the workplace at a large scale, using universe data from 25 universities covering about 25% of federal university based R&D. It is the first to use universe data on workers (the LEHD data) to draw comparison groups of individuals employed both within the university and from other R&D intensive businesses. And it is the first to use universe data on business startups to compare the dynamics of university sourced entrepreneurship with other types of entrepreneurship.

Pathways to Production

Barth E, Davis J, Marschke G, Wang A, Zhou S

Presented at American Economic Association Meeting, Chicago, January 2017
Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

Science funding agencies often require researchers to demonstrate their project's prospects for "development of a diverse, globally competitive STEM workforce," "increase[d] partnerships between academia, industry and others" (NSF, 2016), and other goals beyond the creation of scientific knowledge. This paper attempts to measure these wider impacts of scientific research.

We use the newly created Census data infrastructure that links university grant transaction data to Longitudinal Employer-Household Dynamics (LEHD) data to map employment linkages between universities and industry. First, we ask, what are the flow rates of new STEM workers—post-docs and recent doctorates—into research intensive firms, industries, and regions? startups and established firms? high- and low- productivity firms? local and out-of-state employment?

Second, we estimate the impacts of and returns to university-based human capital accumulation by STEM workers. The sudden increase in science funding under the American Recovery and Reinvestment Act of 2009 (ARRA) increased demand in the academic sector for post-graduate researchers, both lengthening existing post-graduate research engagements in universities, and increasing the likelihood that recent graduates, especially doctorates, obtain post-graduate employment in universities. We estimate the impact of increased university-based research training on career paths, including the likelihood of obtaining a faculty post, and for researchers who enter industry, which firms they match to, and their wage outcomes.
Third, we investigate the extent to which a firm placement depends on the history of previous placements from the same university. Such a correlation could be evidence of "hiring chains", or of specific knowledge links between the research and teaching at a specific university and the production technology of particular firms. The hiring patterns we uncover between universities and industry reveal important features of the labor market for specialized skills, and increase our understanding of how university research contributes to the diffusion of new ideas in the economy.

Financial Advice and the Entrepreneurial Spillovers of Basic Research

Dacunto F, & Yang L

Presented at American Economic Association Meeting, Chicago, January 2017
Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

We test for the effect of informal financial advice on the establishment and subsequent performance of entrepreneurial ventures that commercialize the results of basic research. To this aim, we construct a unique data set that includes: (i) the characteristics of the faculty recipients of federally-funded grants across 10 large U.S. universities, which produce innovation that can be commercialized through the establishment of startups; (ii) the likelihood that recipients establish a non-employer venture (iLBD) or an employer venture (LEHD), as well as the job growth characteristics of these ventures; and (iii) the network of neighbors in the locations where the recipients' reside, including the occupation titles and demographics of the neighbors (ACS/Decennial Census). We use the presence of financial-sector employees among the faculty's network members (spouses or neighbors) to test for the effect. We compare faculty grant recipients in similar areas of research, obtaining grants of similar sizes in the same rounds of funding, and at similar stages of their academic careers, but belonging to networks with different levels of exposure to informal financial advice from family and friends. Financial advice from one's social network is informal because advisers are not paid fees for providing their service. Therefore, the paper broadly tests for whether advice is a positive externality of one's social networks, which is valuable to the individual entrepreneurs as well as to economic growth.

Research Funding and Subsequent Entrepreneurship: The Role of Underrepresentation

Buffington C, Harris B, Feng F, & Weinberg BA

Presented at American Economic Association Meeting, Chicago, January 2017
Session: Using Data Science to Examine the Link Between University R&D and Innovation (moderated by Julia Lane)

Abstract

Federal funding affects both who does research, and the environment in which research is done. In a recent study, 6 in 10 female doctoral recipients had been supported by federal research funds, compared to 7 in 10 male doctoral recipients. Federal funding also appears to

be highly correlated with the pipeline of researchers going into different fields; particularly into R&D fields and the decision to pursue postdoctoral fellowships.

This paper uses rich new Census Bureau data linked to detailed information on the individuals supported by research funding to examine the effect of both the type and structure of federal funding on the outcomes of underrepresented students. It makes use of rich measures on student characteristics, including their race, gender, place of birth, marital status and presence of children. It constructs new network theoretic measures of team environment, based on the characteristics of all individuals working together on research grants. It also includes information about household and family structure in the model. It also examines two types of outcome measures - placement in R&D performing, high technology or young and small firms - as well as the propensity of underrepresented groups to start up businesses.

Nevertheless She Persisted? Gender Peer Effects in Doctoral Stem Programs

Bostwick, Valerie K. and Weinberg, Bruce A.

NBER Working Paper No. 25028, September 2018 http://www.nber.org/papers/w25028

Abstract

We study the effects of peer gender composition, a proxy for female-friendliness of environment, in STEM doctoral programs on persistence and degree completion. Leveraging unique new data and quasi-random variation in gender composition across cohorts within programs, we show that women entering cohorts with no female peers are 11.9pp less likely to graduate within 6 years than their male counterparts. A 1 sd increase in the percentage of female students differentially increases the probability of on-time graduation for women by 4.6pp. These gender peer effects function primarily through changes in the probability of dropping out in the first year of a Ph.D. program and are largest in programs that are typically male-dominated.

BOOKS & BOOK CHAPTERS

Democratizing Our Data: A Manifesto

Lane, Julia

The MIT Press, September 2020

Abstract

Public data are foundational to our democratic system. People need consistently high-quality information from trustworthy sources. In the new economy, wealth is generated by access to data; government's job is to democratize the data playing field. Yet data produced by the American government are getting worse and costing more. In Democratizing Our Data, Julia Lane argues that good data are essential for democracy. Her book is a wake-up call to America to fix its broken public data system.

Lane argues that we must rethink ways to democratize data; there are successful models to follow and new legislation that can help effect change. The private sector's data revolution—which creates new types of data and new measurements to build machine learning and artificial intelligence algorithms—can be mirrored by a public sector data revolution characterized by attention to counting all who should be counted, measuring what should be measured, and protecting privacy and confidentiality. Just as Google, Amazon, Microsoft, Apple, and Facebook have led the world in the use of data for profit, the United States can show the world how to produce data for the public good.

Lane calls for a more automated, transparent, and accountable framework for creating high-quality public data that would empower citizens and inspire the workforce that serves them. And she outlines an organizational model that has the potential to make data more accessible and useful. As she says, failure to act threatens our democracy.

Research Universities and the Public Good: Discovery for an Uncertain Future

Owen-Smith, Jason

Stanford University Press, September 2018

Abstract

In a political climate that is skeptical of hard-to-measure outcomes, public funding for research universities is under threat. But if we scale back support for these institutions, we also cut off a

key source of value creation in our economy and society. *Research Universities and the Public Good* offers a unique view of how universities work, what their purpose is, and why they are important.

Countering recent arguments that we should "unbundle" or "disrupt" higher education, Jason Owen-Smith argues that research universities are valuable gems that deserve support. While they are complex and costly, their enduring value is threefold: they simultaneously act as *sources* of new knowledge, *anchors* for regional and national communities, and *hubs* that connect disparate parts of society. These distinctive features allow them, more than any other institution, to innovate in response to new problems and opportunities. Presenting numerous case studies that show how research universities play these three roles and why they matter, this book offers a fresh and stirring defense of the research university.

The Role of Innovation and Entrepreneurship in Economic Growth

Andrews, M, Chatterji, A, Lerner, J, Stern, S

University of Chicago Press, 2020

ISBN 9781316671788

Chapter: Measuring Business Innovation Using a Multi-Dimensional Approach, Lucia Foster, U.S. Census Bureau

Advancing the U.S. Census Bureau's mission "to serve as the nation's leading provider of quality data about its people and economy" requires a robust and agile research and development program working in close collaboration with external experts and Census Bureau programmatic staff. Even straightforward concepts, such as the use of industrial robotics in manufacturing, can require a multi-dimensional measurement approach. While the Census Bureau is known for its surveys, some of our most innovative work combines survey data with administrative data or combines multiple sources of administrative data.

In this chapter, I discuss the multi-dimensional research and development approach the Center for Economic Studies (CES) at the Census Bureau takes in attempting to better understand business innovation. Since it is not possible to provide details on these many interrelated efforts, I highlight our multi-dimensional approach by giving examples of research using administrative data, survey data, and indirect inference. A more complete view of CES research activities is provided in our annual reports and working paper series.

Measuring the Economic Value of Research: The Case of Food Safety

Husbands Fealing K, Lane JI, King J, & Johnson SR (eds.)

***Cambridge University Press*, 2017**

ISBN 9781316671788

Abstract
This innovative study demonstrates new methods and tools to trace the impact of federal research funding on the structure of research, and the subsequent economic activities of funded researchers. The case study is food safety research, which is critical to avoiding outbreaks of disease. The authors make use of an extraordinary new data infrastructure and apply new techniques in text analysis. Focusing on the impact of U.S. federal food safety research, this book develops vital data-intensive methodologies that have a real world application to many other scientific fields.

Chapter 8 Networks: The Basics

Owen-Smith J, in *Big Data and Social Science* pp. 215-240
Foster I, Ghani R, Jarmin RS, Kreuter F, & Lane JI (eds.)
***Chapman and Hall/CRC*, August 9, 2016**
ISBN 9781498751407

Abstract
Noted sociologist and network theorist Jason Owen-Smith provides a primer on network theory, including details on network measures and components.

Big Data and Social Science: A Practical Guide to Methods and Tools

Foster I, Ghani R, Jarmin RS, Kreuter F, & Lane JI (eds.)
***Chapman and Hall/CRC*, August 9, 2016**
ISBN 9781498751407

Abstract
Big Data and Social Science: A Practical Guide to Methods and Tools shows how to apply data science to real-world problems in both research and the practice. The book provides practical guidance on combining methods and tools from computer science, statistics, and social science. This concrete approach is illustrated throughout using an important national problem, the quantitative study of innovation. The text draws on the expertise of prominent leaders in statistics, the social sciences, data science, and computer science to teach students how to use modern social science research principles as well as the best analytical and computational tools. It uses a real-world challenge to introduce how these tools are used to identify and capture appropriate data, apply data science models and tools to that data, and recognize and respond to data errors and limitations.

PRESS COVERAGE & COMMENTARIES

- *Female scientists don't get the credit they deserve. A study proves it.* (Washington Post, June 22, 2022)
- *$100 Million contributed to state economy through University of Michigan research* (MLive, March 3, 2022)
- *NSF funding choice: Move forward or fall behind* (The Hill, May 23, 2021)
- *U-M Reports $1.62B in Research Volume for FY2020* (dbusiness, November 12, 2020)
- *Podcast: Jason Owen-Smith, Research Universities and the Public Good* (Casimir Jones, August 18, 2020)
- *Gary Ostrander: Federal help needed to maintain research, innovation hit hard by COVID-19* (Florida Politics, July 2, 2020)
- *180 congressmen support call for US$26bn research support* (University World News, May 2, 2020)
- *Universities are being "short sighted" when chasing partnerships with companies like Amazon* (Michigan Radio's Stateside program March 4, 2019) Jason Owen-Smith
- *U-M research contributed more than $5 billion to national economy* (mlive.com, February 3, 2020)
- *Study finds most bachelor's degree graduates enter professional, technical services and healthcare fields* (The Michigan Daily, September 11, 2019)
- *Amazon pullout from NYC shows the perils of partnerships between higher education and business* (The Conversation February 26, 2019) Jason Owen-Smith
- *When you're the only woman: The challenges for female Ph.D. students in male-dominated cohorts* (Science October 24, 2018)
- *Gender imbalance affects degrees* (Science News at a glance September 28, 2018)
- *Women In Stem Benefit From Same-Sex Support* (Pacific Standard September 19, 2018)
- *'Nevertheless She Persisted?'* (Inside Higher Ed September 18, 2018)
- *An insidious reason women are less likely to get a STEM doctoral degree than men* (Moneyish September 17, 2018)

- *One Big Reason Why Women Drop Out of Doctoral STEM Programs* (Communication of the ACM September 17, 2018)
- *One big reason why women drop out of doctoral STEM programs* (Ohio State News September 17, 2018
- *Building an infrastructure to support the use of government administrative data for program performance and social science research* (ANNALS, AAPSS, Vol 675, Issue 1, January 2018) Julia Lane
- *A roadmap to a nationwide data infrastructure for evidence-based policy making* (ANNALS, AAPSS, Vol 675, Issue 1, January 2018) Andrew Reamer and Julia Lane
- *Tax bill would imperil nation's innovation, future* (Columbus Dispatch December 14, 2017) Bruce Weinberg, Jason Owen-Smith, and Julia Lane
- *Watching the players, not the scoreboard* (Nature: Comment November 2, 2017) Julia Lane
- *The Looming Decline of the Public Research University* (Washington Monthly September/October 2017)
- *The social sciences need to build new foundations* (Significance Magazine June 9, 2017) Julia Lane
- *A call to action to build social science data infrastructure* (Nature Human Behaviour April 7, 2017) Julia Lane
- *Communicating the Value of University Research When Science is Under Attack* (Inside Higher Ed April 6, 2017)
- *Who Feels the Pain of Science Research Budget Cuts?* (The Conversation/Salon March 29, 2017) Bruce Weinberg
- *Trump Administration Proposes Big Cuts in Medical Research* (NPR Health Shots March 16, 2017)
- *The Price of Doing a Postdoc* (Science: Share January 10, 2017)
- *Fix Incentives* (Nature: Perspective September 1, 2016) Julia Lane
- *There's a huge gender pay gap for STEM careers — just one year after graduation* (Vox May 11, 2016)
- *Assessment: Academic return* (Nature May 4, 2016)
- *ProQuest Dissertation Database Provides Critical Information for Research Projects Across the US* (PR Newswire March 22, 2016)
- *Facing Skepticism, colleges set out to prove their value* (PBS Newshour January 22, 2016)
- *Science and math PhDs earn about $65,000 — more than double what arts majors do* (Vox December 11, 2015)

- *Biologists lose out in post-PhD earnings analysis* (Nature: News December 10, 2015)
- *Where new PhD grads find work — and who earns the most with their degree* (Washington Post December 10, 2015)
- *PhDs pay: study reveals economic benefit of funding doctorates* (Times Higher Education December 10, 2015)