Industries of Ideas: Tracing the links between investments in science, innovation, and jobs

We are witnessing a sea change in government spending on science and technology. Modern industrial policy represents a bet on a new approach - investing in ideas - to transform the economy and create high wage jobs. Investments in science and technology, the scale of which has not been seen since the Cold War, have anecdotally had massive effects on jobs and earnings. The Artificial Intelligence revolution alone has set off a gold rush – at the 2023 Supercomputer conference in Denver in October, one of us was told that the starting salaries of new Stanford PhDs in AI was $750,000 – plus stock options. Yet firms and workers looking for data on which to base their hiring and career decisions are out of luck.   As the former Federal CIO and advisor to many tech companies, Suzette Kent, says in a related white paper "AI may have created one of the most fast-paced workforce shifts in history, but reliable data are hard to find"[1]

Without relevant data on the links between science and innovation, how are governments, science funders, and scientists to make decisions about where to place their bets? What is the theory, data, and evidence? The theory got a Nobel Prize in 2018. Governments are investing in people who create ideas – new technologies – that *can* be reused, which is why "the discovery of new ideas lie at center of economic growth…" (Charles Jones describing Paul Romer's conceptual framework for which Romer received the 2018 Nobel Prize in Economics[2]). The data seeds were sown almost two decades ago. President Bush's Science Advisor, who, sensibly unconvinced of the scientific and practical value of relying primarily on document-based, bibliometric approaches to studying science to understand its practical effects, called for a "Science of Science Policy"[3]. The evidence is being built by the University of Michigan's Institute for Research on Innovation and Science (IRIS) in a people-centered data infrastructure which draws on the original government-led STAR METRICS program, now called UMETRICS to reflect its university leadership [4]. The infrastructure changes the very way in which data are being *structured* – so that the relevant processes are being studied; *classified* – so that levels and trends in funding inputs and subsequent activities can be measured and tracked; *collected* – so that actionable information is available for multiple units of analysis; and *analyzed* – so that governments, science funders, and workers can make informed decisions.

The people-centered IRIS data approach, because it is characterized by a collaborative, bottom up, and scientifically grounded governance model, is purposefully designed both to respond to the interests of the relevant communities and be used by them. It stands in direct contrast to the current document-centered data approach, which lacks a clear governance structure and scientific framework [5].

The framework

The new data *structure* is grounded in understanding the processes by which growth through investments in science is generated. These processes are fundamentally different from investments in capital and labor that produce physical goods and services[1] that once used, cannot be reused. In practical terms, this is why Stanford PhDs are paid so much, since the firms that hire them expect the transfer of their ideas to others in the firm and thus generate more revenue and growth. An operational data structure based on the Romer concept requires joining up the dynamic flows of all people funded on research grants at universities – the ideas workers – with the jobs they get when they move to the private sector.   These flows can then be used to trace their effects both on the firms at which they work

[1] What economists call a production function

and the other workers at those firms – as Oppenheimer said “the best way to transmit knowledge is to wrap it up in a human being”[6].

The *classification* system is similarly designed to reflect current scientific and economic activities. New technologies, such as Artificial Intelligence, represent clever new ideas of how to combine existing inputs better, so the clustering and measurement of activities represent the clustering of people and the ideas embodied in them. The “Industries of Ideas”[7] approach being deployed by IRIS groups firms by the people who created and use the technologies they will adopt. Such a classification framework is a sea change from earlier industrial classifications based on *what* goods are physically produced - like manufacturing and agriculture[8] – or by *how* services and goods are produced – like the delivery of health, financial, and investment services [9].

Data *collection* is also purposefully designed to be timely, flexible, and useful. People-centric data generated by the administrative processes at universities and firms can capture the organization of people in science at multiple levels (e.g. individuals, teams, projects, and institutions), their multiple sources of funding (federal scientific and programmatic agencies, philanthropic foundations, industry, and state and local government), inputs into science from vendors (such as computing services, instruments, biological specimens), as well as the dynamics of their careers across time (individual career earnings and employment trajectories).

Finally, data *analysis* is not centralized but is also bottom up, transparent, and collaborative. The IRIS infrastructure has been developed over the past decade. The current production release reflects actual expenditures on more than 535,000 grants, 864,000 employees and 970,000 vendors paid by more than 80 campuses representing more than 41% of U.S. total R & D spending at universities[10]. IRIS, while hosted at the University of Michigan, has a governing board which represents its member institutions. While the core of the data are the administrative records, the infrastructure provides [11] Access to the IRIS infrastructure is open to all collaborating universities and their approved researchers; over 500 researchers accessed the data for scientific purposes and hundreds of reports been generated for science funders, federal and state government agencies, and the participating universities themselves.

What does this mean in practice? One of us served on the National AI Research Resources Task Force (NAIRR TF), charged with developing a roadmap to guide investments in AI compute and data resources with, inter alia, spurring innovation. That lack of reliable data identified by Suzette Kent was also recognized by the NAIRR TF; its final report submitted to the President and Congress in January 2023 incorporated the people-centered approach described here as part of its evaluation framework[12].

Empirical implementation: AI and EV

One of the use cases is the National Science Foundation’s new Technology, Innovation, and Partnership (TIP) Directorate which has funded a pilot “Industries of Ideas” project to better understand its regional technology investments. The pilot focuses on two critical and emerging technologies - AI and Electric Vehicles (EV) – in the state of Ohio, and is designed to scale to other technologies and states in subsequent stages. It begins with linking people funded in AI and EV research in Ohio universities and links that with individual and firm level state administrative workforce and education data at the Ohio Longitudinal Data Archive [13].

As Jason Owen Smith points out in a related white paper, research communities form and can be identified through field specific activities and collaborations. In the case of AI and EV for the subset of

IRIS universities, between 2001 and 2023, NSF invested $8.5 billion in about 13,000 awards to about 3,300 principal investigators. [14] The flow of funding to researchers in that field can be identified through university administrative grant and award data. For the subset of IRIS universities, between 2001 and 2023, it is possible to trace the NSF investments of $8.5 billion in about 13,000 awards to about 3,300 principal investigators.

The IRIS data then is used to capture all spending on teams: principal investigators, trainees like undergraduate and graduate students, post-docs, as well as staff clinicians, and administrative staff. Those grants support 46,385 people – or almost 15 people per PI. Almost half of those, over 21,000, were graduate students. 8,300 were non-PI faculty, 8,000 were staff, almost 3,000 were undergraduates and 2,000 postdocs.

Many if not most of these will not ever publish a paper or be a PI. But working new and emerging AI research teaches them about applications through the lens of nearly every field NSF supports. It gives them access to specialized professional networks. It makes them both competitive for and interested in AI jobs. In other words, these hitherto invisible research funded people are a key "product" of grant funded research and a way to identify currently unmeasurable workforce effects.

The effects on the private sector are not just the initial earnings of knowledge worker staff (like the Stanford computer science graduates!). It is the cumulative knock-on effect on the earning of all workers in the firms who use the ideas of the knowledge workers. Simply put, just as Oppenheimer posited and Romer theorized, the mobility of research funded staff helps connect the ideas workforce to the firms so that new technologies can be transferred to existing production processes. Their flows through to the full economy, and the transmission of their ideas, is captured when trainees and staff get jobs in the private sector and their earnings and employment are recorded in state administrative data[15]. Our back of the envelope estimates from aggregate data suggest that the potential national impact could be up to 36 million workers in 18 sectors; a dashboard presenting firm and worker results for Ohio that is scalable to other technologies and states will be a major outcome of the two year pilot.

Finally making science metrics more scientific

The new approach to understanding the *structure* of data is well captured by the quote by Erwin Gianchandani, the TIP Assistant Director in the pilot's press release "NSF's strategic investments in key technologies warrant innovative tools to accurately assess the impact of these investments across the U.S.. ..The Industries of Ideas project will develop a prototype to better understand the impact of NSF's efforts through the new TIP directorate, providing rich, descriptive analyses of the interplay between our investments and people, jobs and regional economies."[16]

There is also a new energy around the *classification* issues raised here: think tanks, measurement experts, academics, government agencies, and private sector data providers are considering new approaches to measurement as evidenced by this workshop.

There is new engagement around data *collection* and *analysis*. One is the recently introduced bipartisan HR 6655 Reauthorization Bill "A Stronger Workforce for American Act". That bill specifically calls out funding Workforce Data Quality Initiative grants to improve state workforce data capabilities by fostering cross-state collaboration, improving the timeliness and relevance of labor market data, supporting the adoption of credential navigation tools, and advancing the use of evidence and data to drive decisionmaking.[17] As noted by Adam Leonard in one of the white papers for this workshop, regional multi-state data collaboratives provide a basis for state education and workforce agencies to contribute data and produce new products[18]. The UMETRICS data are increasingly being used for both training and research.[19-22]

In sum, this new approach to constructing data on the links between science and innovation, so urgently needed if governments, science funders, and scientists are to make decisions about where to place their bets, is in place. The theory, data, and evidence can now inform understanding of the impacts of countries' vast investments in AI research – and research in many other fields.

References

1. Kent, S., *Perspective from NAIRR*, in *Stanford Workshop on New Approaches to Characterizing Industries*. 2024.
2. Buffington, C., et al., *STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census.* American Economic Review, 2016. **106**(5): p. 333-338.
3. Marburger III, J.H., *Wanted: better benchmarks*. 2005, American Association for the Advancement of Science. p. 1087-1087.
4. Lane, J.I., et al., *New linked data on research investments: Scientific workforce, productivity, and public value.* Research policy, 2015. **44**(9): p. 1659-1671.
5. Lane, J., *Let's make science metrics more scientific.* Nature, 2010. **464**(7288): p. 488-489.
6. Zolas, N., et al., *Wrapping it up in a person: Examining employment and earnings outcomes for Ph. D. recipients.* Science, 2015. **350**(6266): p. 1367-1371.
7. Lane, J., *The Industry of Ideas: Measuring How Artificial Intelligence Changes Labor Markets*, in *American Enterprise Institute*. 2023.
8. Yuskavage, R.E. *Converting historical industry time series data from SIC to NAICS*. in *The Federal Committee on Statistical Methodology 2007 Research Conference. 5-7 November 2007*. 2007. Washington, DC US Department of Commerce, Bureau of Economic Analysis.
9. Haver, M.A., *The statistics corner: The NAICS is coming. Will we be ready?* Business Economics, 1997. **32**(4): p. 63-65.
10. Nicholls, N., Ku, R., Brown, C., and J. Owen-Smith, *Summary Documentation for the IRIS UMETRICS 2022 Data Release*, h.I.f.R.o.I.S.d. publisher], Editor. 2022-04-20.
11. Chang, W.-Y., et al., *A linked data mosaic for policy-relevant research on science and innovation: Value, transparency, rigor, and community.* Harvard data science review, 2022. **4**(2).
12. Office of Science and Technology Policy, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem*. 2023.
13. Hawley, J., *Ohio and the Longitudinal Data Archive: Developing Mutually Beneficial Partnerships Between State Government and the Research Community*, in *Handbook on Using Administrative Data for Research and Evidence-based Policy*, L. Vilhuber, Editor. 2020: J-Pal.
14. Owen-Smith, J., *Will the real AI researcher please stand up?*

Fields, Networks and Systems to Measure the Impact of Research Investments, in *Workshop on AI Measurement*. 2024: Stanford.

15. Lane, J., *Reimagining Labor Market Information: A National Collaborative for Local Workforce Information.* 2023.
16. Technology, I.a.P.D. *NSF launches pilot to assess the impact of strategic investments on regional jobs*. 2024; Available from: https://new.nsf.gov/tip/updates/nsf-pilot-assess-impact-strategic-investments-regional-jobs.

17. 118th Congress (2023-2024), *A Stronger Workforce for America Act*. December 12, 2023.
18. Leonard, A., *Outside of the Box Use of Administrative and Wage Data in Texas*, in *Workshop on New Approaches to Characterize Industries*. 2024: Stanford University.
19. Babina, T., et al., *Cutting the Innovation Engine: How Federal Funding Shocks Affect University Patenting, Entrepreneurship, and Publications.* The Quarterly Journal of Economics, 2023. **138**(2): p. 895-954.
20. Coupet, J. and Y. Ba, *Benchmarking university technology transfer performance with external research funding: a stochastic frontier analysis.* The Journal of Technology Transfer, 2022. **47**(2): p. 605-620.
21. Schacter, S.Y., D. Kang, and J.A. Evans, *The Inefficiency of Private Support for Public Health: Comparing Nonprofit Biomedical Research Funding with the NIH.* Available at SSRN 4483036.
22. Tham, W.Y., *Science, interrupted: Funding delays reduce research activity but having more grants helps.* Plos one, 2023. **18**(4): p. e0280576.