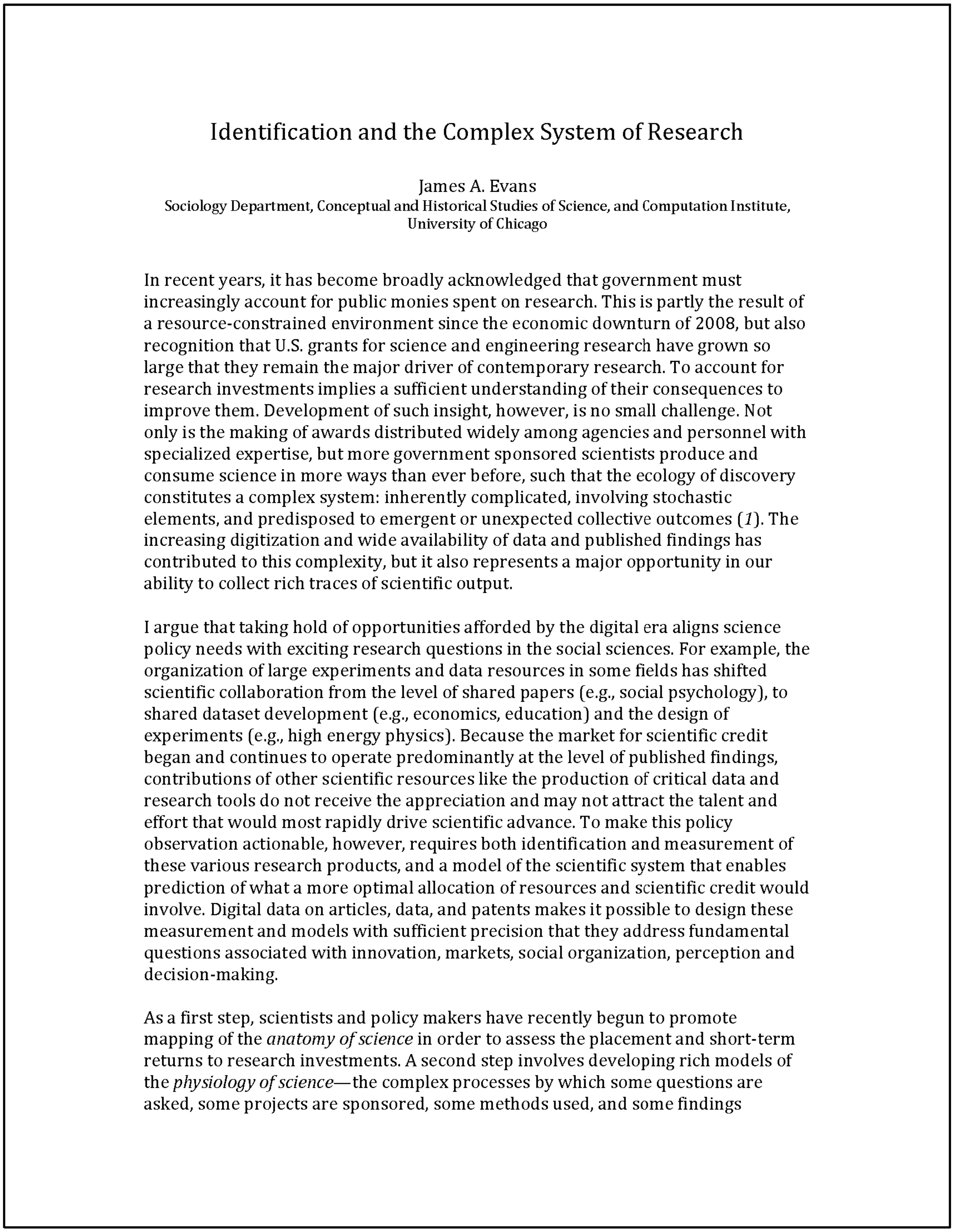
Evans, James, “Identification and the Complex System of Research”



Identification and the Complex System of Research

James A. Evans
Sociology Department, Conceptual and Historical Studies of Science, and Computation Institute,
University of Chicago

In recent years, it has become broadly acknowledged that government must increasingly account for public monies spent on research. This is partly the result of a resource-constrained environment since the economic downturn of 2008, but also recognition that U.S. grants for science and engineering research have grown so large that they remain the major driver of contemporary research. To account for research investments implies a sufficient understanding of their consequences to improve them. Development of such insight, however, is no small challenge. Not only is the making of awards distributed widely among agencies and personnel with specialized expertise, but more government sponsored scientists produce and consume science in more ways than ever before, such that the ecology of discovery constitutes a complex system: inherently complicated, involving stochastic elements, and predisposed to emergent or unexpected collective outcomes (*1*). The increasing digitization and wide availability of data and published findings has contributed to this complexity, but it also represents a major opportunity in our ability to collect rich traces of scientific output.

I argue that taking hold of opportunities afforded by the digital era aligns science policy needs with exciting research questions in the social sciences. For example, the organization of large experiments and data resources in some fields has shifted scientific collaboration from the level of shared papers (e.g., social psychology), to shared dataset development (e.g., economics, education) and the design of experiments (e.g., high energy physics). Because the market for scientific credit began and continues to operate predominantly at the level of published findings, contributions of other scientific resources like the production of critical data and research tools do not receive the appreciation and may not attract the talent and effort that would most rapidly drive scientific advance. To make this policy observation actionable, however, requires both identification and measurement of these various research products, and a model of the scientific system that enables prediction of what a more optimal allocation of resources and scientific credit would involve. Digital data on articles, data, and patents makes it possible to design these measurement and models with sufficient precision that they address fundamental questions associated with innovation, markets, social organization, perception and decision-making.

As a first step, scientists and policy makers have recently begun to promote mapping of the *anatomy of science* in order to assess the placement and short-term returns to research investments. A second step involves developing rich models of the *physiology of science*—the complex processes by which some questions are asked, some projects are sponsored, some methods used, and some findings published, amplified and used in advance while others are not. To effectively address the first project hinges on the *identification* of essential elements in the research system, and the second on realistic *models* that capture essential interactions between those elements.

Identification and Measurement

The first step toward understanding the scientific system is to identify key elements in the system. These include, but are not restricted to the following:

1. Researchers (i.e., authors / inventors)
2. Research funds
3. Scientific knowledge:
    a. articles
    b. citations
    c. methods
    d. tools
    e. data resources
    f. concepts
    g. findings
4. Broader societal outcomes:
    a. Economic growth
        i. jobs
        ii. start-ups
        iii. patents
    b. Workforce
        i. student mobility into other jobs
        ii. student presence in jobs
    c. Long-term social outcomes
        i. health impact
        ii. environmental impact

The first two constitute research inputs, and third proximate outputs, which are themselves inputs to later stages of the research process. Although the first outputs—research documents and citations—have the most conceptual integrity and are the most often measured[1], they are unsatisfying as sole measurements because they do not represent the primary level of granularity at which scientists make "moves" and receive credit in science. Yes, academics publish and receive accolades for articles, but that is an outgrowth of their development, dissemination and promotion of methods, tools, data, concepts, and findings that seek to influence later work—to influence and advance science.

[1] Even the "integrity" of the article is beginning to change in the digital era with updating online books and papers.

Technical hurdles challenge the process of identifying each of these scientific elements when the digital written record is the primary source of information. Some elements, like research funds, are partially censored because they are only sometimes acknowledged. Others, including methods, tools, data resources, concepts and findings are trapped within the full-text and can only be recovered through error-prone natural language processing and classification methods. All but articles and citations share a common design challenge best typified by scientist names. Scientist's names are sometimes printed with variation, and many share the same common names (e.g., synonymy and homonymy). The structure of the problem is that a unique set of scientists map onto a typically larger set of ambiguous names, and while this suggests a many-to-many global optimization procedure, the problem is almost always approached as a pairwise matching process to increase speed and reduce memory requirements. This choice, however, necessarily multiplies errors by not allowing certain matching choices to constrain the probability of others. All of these challenges recommend that in addition to "pulling" data from the digital corpus, the scientific establishment could profitably incentivize researchers to "push" that data either by entering it themselves or through participating to identify and disambiguate their research products.

The Research Performance Progress Report (RPPR) and Star-metrics and represent recent initiatives to both pull and incentivize researchers to push information about their research outputs. The RPPR involves creation of a consistent, agency-independent "form" through which researchers sponsored by all agencies of government report research and broader outcomes. In Star-metrics, agencies will gather information on elements of the scientific system (as also indicators of economic growth, workforce and long-term social outcomes) and may explore ways to link to updated research documents (e.g., a researcher web page) to facilitate a coordinated push and pull of information. Another possibility is to follow Brazil's Lattes system, in which researcher profiles are automatically generated and then researchers update, clean and "certify" them as acceptable. The central challenge with such a system is to effectively elicit participation. If it is not mandated, then the system must provide the researcher with some value. One approach would be to capture and automate a "workflow" that is otherwise expensive to the scientist. For example, if the researcher commonly had to keep multiple bio-sketches up to date, the system could automatically generate agency-independent sketches and other reports (similar to the RPPR, but for application purposes). Alternately, following the Lattes model, automatically gathering data from online publications and the web could entice researchers to edit their profiles, which edits could be used to improve the information extraction. The quality of information extraction would need to be high, however, because if quality was low, it would not benefit researchers enough to entice them to wade through it. One possible system design could incorporate both of these features by inviting researchers to enter their information for the generation of applications, reports, etc., and they could *optionally* curate "pulled" data to incorporate into their bio-sketch or report.

Modeling

Once research elements are identified, the space of all possible models about how they combine to create new scientific knowledge and broader economic and social outcomes is far too high to explore exhaustively. This requires platforms on which alternate models of the scientific process can be considered and tested. Following the earlier example, this could enable scientists and science policy experts to estimate underinvestment in the creation of data resources and research tools relative to articles and findings. Then incentives could be put in place to shift investment. In addition to financial incentives, one class of enticements could involve the outputs of a Starmetrics-based assessment that ranks the most used and influential research tools and data resources, evaluated across the population of published research. This could function like an ImpactFactor or PageRank for data resources and methods that would attract attention and implicitly confer scientific visibility and credit. Moreover, such a system could model and then rank the relative influence of each contributing researcher in driving the importance of these entities for science.

These represent a few preliminary consideration regarding the possibilities, limitations and ultimate potential of harnessing digital media and internet connection to understand and improve the system of science.

1. R. Foote, *Science* **318**, 410 (Oct 19, 2007).