ESSAY

Data-driven predictions in the science of science

Aaron Clauset,[1,2]* Daniel B. Larremore,[2] Roberta Sinatra[3,4]

The desire to predict discoveries—to have some idea, in advance, of what will be discovered, by whom, when, and where—pervades nearly all aspects of modern science, from individual scientists to publishers, from funding agencies to hiring committees. In this Essay, we survey the emerging and interdisciplinary field of the "science of science" and what it teaches us about the predictability of scientific discovery. We then discuss future opportunities for improving predictions derived from the science of science and its potential impact, positive and negative, on the scientific community.

Today, the desire to predict discoveries—to have some idea, in advance, of what will be discovered, by whom, when, and where—pervades nearly all aspects of modern science. Individual scientists routinely make predictions about which research questions or topics are interesting, impactful, and fundable. Publishers and funding agencies evaluate manuscripts and project proposals in part by predicting their future impact. Faculty hiring committees make predictions about which candidates will make important scientific contributions over the course of their careers. And predictions are important to the public, who fund the majority of all scientific research through tax dollars. The more predictable we can make the process of scientific discovery, the more efficiently those resources can be used to support worthwhile technological, biomedical, and scientific advances.

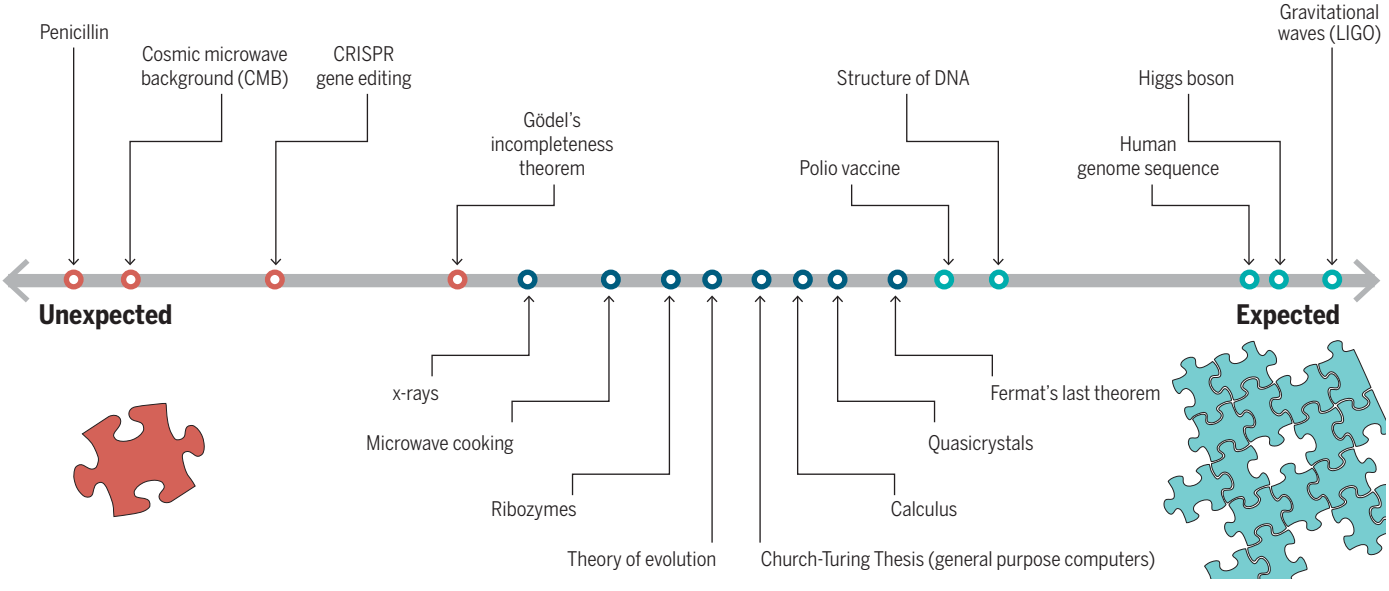
Despite this pervasive need, our understanding of how discoveries emerge is limited, and relatively few predictions by individuals, publishers, funders, or hiring committees are made in a scientific way. How, then, can we know what is predictable and what is not? Although it can be difficult to separate the discovery from the discoverer, the primary focus of this Essay is the science of science: an interdisciplinary effort to scientifically understand the social processes that lead to scientific discoveries. [For the current thinking on the philosophy of science and how scientists make progress on individual scientific challenges, see (*1*).]

[1]Department of Computer Science, and BioFrontiers Institute, University of Colorado, Boulder, CO 80309, USA. [2]Santa Fe Institute, Santa Fe, NM 87501, USA. [3]Center for Network Science and Department of Mathematics, Central European University, Budapest, Hungary. [4]Center for Complex Network Research and Physics Department, Northeastern University, Boston, MA 02115, USA.
*Corresponding author. Email: aaron.clauset@colorado.edu

Interest in predicting discoveries stretches back nearly 150 years, to work by the philosopher Boleslaw Prus (1847–1912) and the empirical sociologist Florian Znaniecki (1882–1958). Znaniecki, in particular, called for the establishment of a data-driven study of the social processes of science. For most of the 20th century, progress toward this goal came slowly, in part because good data were difficult to obtain and most people were satisfied with the judgment of experts.

Today, the scientific community is a vast and varied ecosystem, with hundreds of loosely interacting fields, tens of thousands of researchers, and a dizzying number of new results each year. This daunting size and complexity has broadened the appeal of a science of science and encouraged a focus on generic measurable quantities such as citations to past works, production of new works, career trajectories, grant funding, scholarly prizes, and so forth. Digital technology makes such information abundant, and researchers are developing powerful new computational tools for analyzing it—for instance, to extract and categorize the content of papers in order to automatically quantify progress on specific scientific questions (*2*, *3*). It is now widely believed that exploiting this information can produce predictions that are more objectively accurate than expert opinions. Bibliographic databases and online platforms—Google Scholar, PubMed, Web of Science, JSTOR, ORCID, EasyChair, and "altmetrics," to name a few—are enabling a new generation of researchers to develop deeper insights into the scientific process.

These efforts raise a provocative question: Will we eventually be able to predict important discoveries or their discoverers, such as Yoshinori Ohsumi's Nobel Prize–winning work on the autophagy system in animal cells? We do not yet know the answer, but work toward one will substantially



GRAPHIC: ADAPTED BY K. SUTLIFF/*SCIENCE*

Fig. 1. How unexpected is a discovery? Scientific discoveries vary in how unexpected they were relative to existing knowledge. To illustrate this perspective, 17 examples of major scientific discoveries are arranged from the unanticipated (like antibiotics, programmable gene editing, and cosmic microwave background radiation) to expected discoveries (like the observation of gravitational waves, the structure of DNA, or the decoding of the human genome).

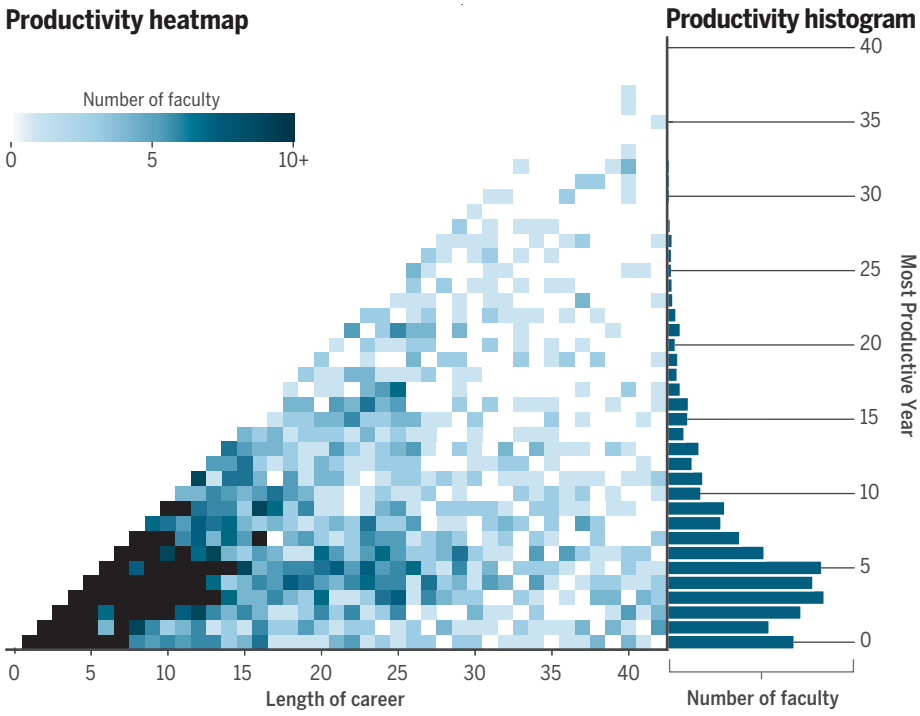


Fig. 2. Productivity peaks early for most researchers. (Left) A heatmap showing the timing of the most productive year (measured in number of published papers) in a faculty career for more than 2300 computer science faculty, arranged from left to right by years since first faculty position (*13*). (Right) The histogram sums the heatmap's rows, showing that, for most researchers, their most productive year occurred within 8 years of starting their laboratory.

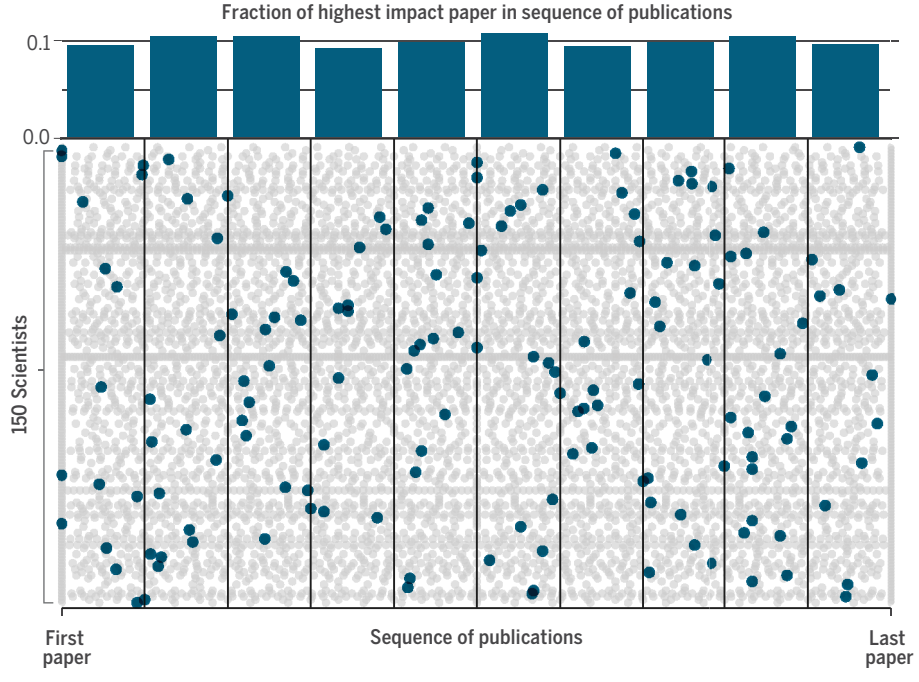


Fig. 3. Major discoveries occur at any point in the sequence of a scientist's publications. A raster plot showing the order of all publications, arranged from first publication to last, of 150 randomly chosen physicists (*17*), where each row of circles represents the sequence of publications by a particular scientist. Within a row, a blue circle marks the highest-impact publication. The uniform distribution of blue circles across the square, and the flatness of the corresponding histogram for 10,000 investigators (top), indicate that there is no pattern across the sequence as to when a major discovery occurs.

advance our understanding of science as a social process. For instance, some scientific discoveries are easily predictable (Fig. 1). As theory and evidence accumulate, it becomes clear that a discovery is imminent, like a single missing piece in the middle of a jigsaw puzzle. The determination of the human genome sequence and the observation of gravitational waves are examples of such discoveries. On the other hand, some discoveries seem impossible to predict because they represent puzzle pieces that change how we think the puzzle is organized or that find new uses in underdeveloped parts of the puzzle. Although the implications of such a novel piece are sometimes immediately obvious, as with programmable gene editing, sometimes the implications require time for additional pieces to fall into place, as was the case with penicillin, the first antibiotic, which took 15 years to realize.

Using modern data on published works and scientific careers, researchers in the science of science have begun identifying quantitative patterns that hold true across different research fields, and these insights are redefining the limits of predictability. Four areas exemplify these achievements: citations of past discoveries, who gets hired into career researcher positions, and both the scientific productivity and the timing of major discoveries over a career. However, work in these areas also hints at limits to data-driven predictions about the production of discoveries.

Modern bibliographic databases allow researchers to easily tabulate and study citation counts, which provide a convenient, although controversial, measure of scientific impact. More than 50 years ago, in widely celebrated work, de Solla Price (1922–1983) identified the basic mechanism driving citation counts, in which current visibility and lucky events drive a positive feedback loop that amplifies future visibility (*4*). This "preferential attachment" mechanism explains why citations are distributed so unevenly across papers and why some receive hundreds or even thousands of times more attention than the typical paper. This model also makes remarkably good predictions for how citations accumulate within a developing field (*5*). A modified version, with controls for a paper's recency and its intrinsic appeal, provides predictions about the long-term evolution of citation counts for individual papers, showing when citations will peak and how long it takes a discovery to become common knowledge (*6*).

However, some discoveries do not follow these rules, and the exceptions demonstrate that there can be more to scientific impact than visibility, luck, and positive feedback. For instance, some papers far exceed the predictions made by simple preferential attachment (*5*, *6*). And then there are the "sleeping beauties" in science: discoveries that lay dormant and largely unnoticed for long periods of time before suddenly attracting great attention (*7*–*9*). A systematic analysis of nearly 25 million publications in the natural and social sciences over the past 100 years found that sleeping beauties occur in all fields of study (*9*).

GRAPHICS: ADAPTED BY G. GRULLON/*SCIENCE*

Who will publish the next breakthrough? Who will get grants? Who will get tenure?

PHOTO: SEBASTIEN BONAIME/ALAMY STOCK PHOTO

Examples include a now famous 1935 paper by Einstein, Podolsky, and Rosen on quantum mechanics; a 1936 paper by Wenzel on waterproofing materials; and a 1958 paper by Rosenblatt on artificial neural networks. The awakening of slumbering papers may be fundamentally unpredictable in part because science itself must advance before the implications of the discovery can unfold.

What discoveries are made is partly determined by who is working to make them and how they were trained as scientists (*10*). These characteristics of the scientific workforce are driven by the doctoral programs of a small group of prestigious institutions, which are shown by data to train the majority of career researchers (*11*). As a result of this dominance, the research agendas and doctoral student demographics of a small number of programs tend to drive the scientific preferences and work-force composition of the entire ecosystem. Aside from the robust pattern that 85% of new faculty move from their doctoral program down the hierarchy of prestige among research institutions, faculty placement remains remarkably unpredictable, so far. Models that exploit the available data on early career productivity, postdoctoral training, geography, gender, and more make barely better predictions about ultimate placement than simply knowing a person's academic pedigree (*12*). Accurate predictions in this setting may require different, less-accessible data, or it may be that placement outcomes are fundamentally unpredictable because they depend on latent factors that are unmeasurable.

Researchers have also investigated the predictability of individual scientists' performance and achievements over the course of a career, as measured by their productivity and by citations to their published works. Conventional wisdom suggests that productivity—crudely, the number of papers published—tends to peak early in a scientist's career and is followed by a long and gradual decline (*13*), perhaps as a result of increased teaching or service duties, lower creativity, etc. However, a recent analysis of over 40 years of productivity data for more than 2300 computer science faculty reveals an enormous variability in individual productivity profiles (*14*). Typically, the most productive time for research tends to be within the first 8 years of becoming a principal investigator (Fig. 2), and the most common year in which productivity peaks is just before a researcher's first promotion. At the same time, for nearly half of all researchers, their most productive year occurs later, and, for some, the most productive year is their last.

Past work also suggests that the early-to-middle years of a career are more likely to produce a scientist's "personal best" discovery, i.e., their most well-cited result (*15*, *16*). This pattern implies that the timing of major discoveries is somewhat predictable. However, an analysis of publication histories for more than 10,000 scientists shows that, in fact, there is no correlation between the impact of a discovery and its timing within a scientist's career (*17*). That is, when a scientist's papers are arranged in order from first to last, the likelihood that their most highly cited discovery will be their first paper is roughly the same as it being their second, tenth, or even last paper (Fig. 3). The observation that young scientists tend to be the originators of most major discoveries is thus a natural consequence of their typically higher productivity, not necessarily a feature of enhanced ability early in a career. By simple chance alone, the personal best is more likely to occur in the more productive phases of a scientist's career.

Although the relative timing of each scientist's own highest-impact paper may be impossible to predict, predicting how many citations that paper will attract is a different matter (*17*, *18*). Specifically, citations to published papers vary across scientists in a systematic and persistent way that correlates with the visibility of a scientist's body of work but that is independent of the field of study. This pattern allows us to predict the number of citations of a scientist's personal best work. The two results about the timing and magnitude of a scientist's personal best show that some aspects of the achievements of individual scientists are remarkably unpredictable, whereas other aspects are more predictable.

Robust and field-independent patterns in productivity and impact, along with evidence of biases in the evaluation of research proposals, raise troubling questions about our current approach to funding most scientific research. For instance, observational and experimental studies demonstrate that grant proposals led by female or nonwhite investigators (*19*, *20*) or focused on interdisciplinary research (*21*) are less likely to receive funding. Similarly, the concentration of the most productive and impactful years within the first decade of a scientific career seems to justify efforts to shift funding from older to younger scientists. The NIH's long-running effort to support early-stage investigators is a notable example, although it has had limited success, as the number of NIH awards to scientists under 40 remains lower than its peak 30 years ago (*22*). On the other hand, one might argue that young researchers tend to be more productive in spite of imbalances in external funding. In these cases, the science of science has identified an

important pattern, but determining the underlying cause will require further investigation and active experimentation.

Citations, publication counts, career movements, scholarly prizes, and other generic measures are crude quantities at best, and we may now be approaching the limit of what they can teach us about the scientific ecosystem and its production of discoveries. These measures are lagging indicators of the movements of the scientific frontier, and their ability to predict the emergence of a new field or the possibility of a major discovery may be low. A fundamental question in the science of science is whether better and more accurate predictions are possible with more timely or context-specific sources of data on the work of scientists—for example, the content of papers, data from preprint repositories, scientific workshops, research team communication, rejected manuscripts or grant proposals and their peer reviews, or even social media. Controlled experiments should be used to uncover the causal mechanisms that drive the patterns observed in large digital databases and to probe the relationship between measurable quantities and our interpretation of them, e.g., how well citation counts reflect perceived scientific impact (*23*).

"...We have a responsibility to ensure that the use of prediction tools does not inhibit future discovery, marginalize underrepresented groups..."

Citations and publications, in particular, are measures of past success that exhibit a feedback loop that creates a rich-gets-richer dynamic. When combined with the hypercompetitive nature of modern scientific publishing, funding, and hiring, this feedback loop can create dramatic inequalities in apparent success because opportunities for future success are allocated, in part, based on markers of recent success. However, the profound unpredictability that pervades many aspects of scientific discovery indicates that relying too heavily on such measures can create self-fulfilling predictions (*24*), which ultimately narrow the scope of scientific innovation and divert attention away from potentially fundamental but unpredictable advances. An important direction of future work must be developing measures of success and systems of evaluation that are less prone to feedback loops.

A dangerous possibility is that funders, publishers, and universities will exploit large bibliographic databases to create new systems that automatically evaluate the future "impact" of project proposals, manuscripts, or young scholars. Such data-mining efforts should be undertaken with extreme caution. Their use could easily discourage innovation and exacerbate existing inequalities in the scientific system by focusing on trivial correlations associated with crude indicators of past success. After all, novel discoveries are valuable precisely because they have never been seen before, whereas data-mining techniques can only learn about what has been done in the past. The inevitable emergence of automated systems makes it imperative that the scientific community guide their development and use in order to incorporate the principles of fairness, accountability, and transparency in machine learning (*25*, *26*). We have a responsibility to ensure that the use of prediction tools does not inhibit future discovery, marginalize underrepresented groups, exclude novel ideas, or discourage interdisciplinary work and the development of new fields.

Ultimately, the scientific ecosystem will adapt to changing scientific incentives and requirements, just as biological ecosystems adapt to selective pressures (*27*). As these pressures shift, scientists will adapt or retire, passing on to their students their best practices for survival and proliferation. A troubling trend, however, is the nearly annual declaration by a Nobel laureate that their biggest discovery would not have been possible in today's research environment. The 2016 declaration came from Ohsumi, who decried the fact that "scientists are now increasingly required to provide evidence of immediate and tangible application of their work" (*28*). This widespread emphasis on predictable discoveries over unexpected ones breeds a different, more risk-averse scientist. The result may be a dangerous form of purifying selection, in which young scientists optimize their research efforts to a climate that is maladaptive for the very same scientists we annually recognize for extraordinary scientific contributions.

There is great potential in adapting ideas from ecology and evolutionary theory to better understand and predict the scientific ecosystem as a whole. Progress in this direction will help us avoid the loss of innovation that comes from a loss of diversity. As a community, we must develop policies that cultivate a diverse scientific ecosystem, including Freeman Dyson's visionary birds and focused frogs (*29*), as well as contrarians, wanderers, tool builders, and more. The practical details of achieving this kind of diversifying selection among scientists, however, remain unclear. True ecological research relies on a combination of observational study and active experimentation. Yet, most work in the science of science is purely observational, and adding active experimentation (*30*) will require care, boldness, and bravery from the funding agencies, publishers, and administrators that define the adaptive landscape. If the science of science has taught us anything, it is that science itself can be probed using the scientific method, and we would be foolish to neglect experimentation.

Driven by new data sources, new experiments, and new ideas, we expect the science of science to produce many more exciting insights about the social processes that lead to scientific discovery. Already, research indicates that some aspects of discoveries are remarkably predictable and that these are largely related to how citations of past discoveries accumulate over time. Other aspects, however, may be fundamentally unpredictable. These limitations are a humbling insight in this modern era of big data and artificial intelligence and suggest that a more reliable engine for generating scientific discoveries may be to cultivate and maintain a healthy ecosystem of scientists rather than focus on predicting individual discoveries.

REFERENCES AND NOTES

1. D. G. Mayo, *Error and the Growth of Experimental Knowledge* (Univ. of Chicago Press, 1996).
2. J. A. Evans, J. G. Foster, *Science* **331**, 721–725 (2011).
3. F. Shi, J. G. Foster, J. A. Evans, *Soc. Networks* **43**, 73–85 (2015).
4. D. J. Price, *Science* **149**, 510–515 (1965).
5. M. E. J. Newman, *Europhys. Lett.* **86**, 68001 (2009).
6. D. Wang, C. Song, A.-L. Barabási, *Science* **342**, 127–132 (2013).
7. A. F. J. van Raan, *Scientometrics* **59**, 467–472 (2004).
8. S. Redner, *Phys. Today* **58**, 49–54 (2005).
9. Q. Ke, E. Ferrara, F. Radicchi, A. Flammini, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7426–7431 (2015).
10. E. B. Petersen, *Stud. High. Educ.* **32**, 475–487 (2007).
11. A. Clauset, S. Arbesman, D. B. Larremore, *Sci. Adv.* **1**, e1400005 (2015).
12. S. F. Way, D. B. Larremore, A. Clauset, in *Proceedings of the 25th International Conference on World Wide Web* (2016), pp. 1169–1179; www2016.ca/proceedings.html.
13. P. E. Stephan, S. G. Levin, *Striking the Mother Lode in Science: The Importance of Age, Place, and Time* (Oxford Univ. Press, 1992).
14. S. F. Way, A. C. Morgan, A. Clauset, D. B. Larremore, https://arxiv.org/abs/1612.08228 (2016).
15. B. F. Jones, B. A. Weinberg, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18910–18914 (2011).
16. D. K. Simonton, *Psychol. Rev.* **104**, 66–89 (1997).
17. R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, *Science* **354**, 596 (2016).
18. J. A. G. Moreira, X. H. T. Zeng, L. A. N. Amaral, *PLOS ONE* **10**, e0143108 (2015).
19. L. Bornmann, R. Mutz, H.-D. Daniel, *J. Informetrics* **1**, 226–238 (2007).
20. D. K. Ginther *et al.*, *Science* **333**, 1015–1019 (2011).
21. K. J. Boudreau, E. C. Guinan, K. R. Lakhani, C. Riedl, *Manage. Sci.* **62**, 2765–2783 (2016).
22. B. Maher, M. Sureda Anfres, *Nature* **538**, 444 (2016).
23. F. Radicchi *et al.*, https://arxiv.org/abs/1612.03962 (2016).
24. R. K. Merton, *Antioch Rev.* **8**, 193 (1948).
25. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
26. Fairness, Accountability, and Transparency in Machine Learning: 2016 Workshop; www.fatml.org.
27. P. E. Smaldino, R. McElreath, *R. Soc. Open Sci.* **3**, 160384 (2016).
28. Y. Ohsumi, Nobel Lecture: Authophagy – An intracellular recycling system. 7 December 2016; www.nobelprize.org/nobel_prizes/medicine/laureates/2016/ohsumi-lecture.html.
29. F. Dyson, *Not. Am. Math. Soc.* **56**, 212–223 (2009).
30. J. Langford, "The NIPS experiment" [blog], (Communications of the Association for Computing Machinery, 2015); http://cacm.acm.org/blogs/blog-cacm/181996-the-nipsexperiment/fulltext.

ACKNOWLEDGMENTS

We thank M. Szell, T. R. Cech, and colleagues at the BioFrontiers Institute for helpful conversations and S. F. Way for adapting figures from (*12*) and (*17*).

10.1126/science.aal4217