new opportunities. Data are now arriving fast enough that the work life of many current social scientists is observably changing: Whereas they once sat in their offices working on their own, rates of co-authorship are increasing fast, and a collaborative laboratory-type work model is emerging in many subfields. These trends would be greatly facilitated by universities and funding agencies recognizing the need to build the infrastructure to support social science research.

For the first time in many areas of the social sciences, new forms and quantities of information may well make dramatic progress possible. Will we be ready?

References and Notes
1. H. Weintraub *et al.*, *Science* **267**, 1609 (1995).
2. D. E. Koshland, *Science* **267**, 1575 (1995).
3. G. King, K. Scholzman, N. Nie, Eds., *The Future of Political Science: 100 Perspectives* (Routledge, New York, 2009), pp. 91–93.
4. D. Hopkins, G. King, *Am. J. Pol. Sci.* **54**, 229 (2010).
5. A. M. Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (Yale Univ. Press, New Haven, 2010).
6. C. Mackie, N. Bradburn, Eds., *Improving Access to and Confidentiality of Research Data* (National Research Council, Washington, DC, 2000), p. 49.
7. R. F. White, *The Independent Review* **XI**, 547 (2007).
8. G. King, *PS Pol. Sci. Polit.* **39**, 119 (2006).
9. The Dataverse Network, http://TheData.org.
10. C. C. Aggarwal, P. S. Yu, Eds., *Privacy-Preserving Data Mining: Models and Algorithms* (Springer, New York, 2008).
11. L. Sweeney, *J. Law Med. Ethics* **25**, 98 (1997).
12. G. King, *Sociol. Methods Res.* **36**, 173 (2007).
13. M. Altman, G. King, *D-Lib* **13**, 10.1045/march2007-altman (2007).
14. G. King, *PS Pol. Sci. Polit.* **28**, 494 (1995).
15. R. G. Anderson, W. H. Green, B. D. McCullough, H. D. Vinod, *J. Econ. Methodol.* **15**, 99 (2008).
16. DATA-Pass, www.icpsr.umich.edu/icpsrweb/DATAPASS/.
17. V. Stodden, *Int. J. Comm. Law Pol.* **13**, 1 (2009).
18. My thanks to M. Altman and M. Crosas for helpful comments on an earlier version.

10.1126/science.1197872

PERSPECTIVE

Metaknowledge

James A. Evans* and Jacob G. Foster

The growth of electronic publication and informatics archives makes it possible to harvest vast quantities of knowledge about knowledge, or "metaknowledge." We review the expanding scope of metaknowledge research, which uncovers regularities in scientific claims and infers the beliefs, preferences, research tools, and strategies behind those regularities. Metaknowledge research also investigates the effect of knowledge context on content. Teams and collaboration networks, institutional prestige, and new technologies all shape the substance and direction of research. We argue that as metaknowledge grows in breadth and quality, it will enable researchers to reshape science—to identify areas in need of reexamination, reweight former certainties, and point out new paths that cut across revealed assumptions, heuristics, and disciplinary boundaries.

What knowledge is contained in a scientific article? The results, of course; a description of the methods; and references that locate its findings in a specific scientific discourse. As an artifact, however, the article contains much more. Figure 1 highlights many of the latent pieces of data we consider when we read a paper in a familiar field, such as the status and history of the authors and their institutions, the focus and audience of the journal, and idioms (in text, figures, and equations) that index a broader context of ideas, scientists, and disciplines. This context suggests how to read the paper and assess its importance. The scope of such knowledge about knowledge, or "metaknowledge," is illustrated by comparing the summary information a first-year graduate student might glean from reading a collection of scientific articles with the insight accessible to a leading scientist in the field. Now consider the perspective that could be gained by a computer trained to extract and systematically analyze information across millions of scientific articles (Fig. 1).

Metaknowledge results from the critical scrutiny of what is known, how, and by whom. It can now be obtained on large scales, enabled by a concurrent informatics revolution. Over the past 20 years, scientists in fields as diverse as molecular biology and astrophysics have drawn on the power of information technology to manage the growing deluge of published findings. Using informatics archives spanning the scientific process, from data and preprints to publications and citations, researchers can now track knowledge claims across topics, tools, outcomes, and institutions (*1–3*). Such investigations yield metaknowledge about the explicit content of science, but also expose implicit content—beliefs, preferences, and research strategies that shape the direction, pace, and substance of scientific discovery. Metaknowledge research further explores the interaction of knowledge content with knowledge context, from features of the scientific system such as multi-institutional collaboration (*4*) to global trends and forces such as the growth of the Internet (*5*).

The quantitative study of metaknowledge builds on a large and growing corpus of qualitative investigations into the conduct of science from history, anthropology, sociology, philosophy, psychology, and interdisciplinary studies of science. Such investigations reveal the existence of many intriguing processes in the production of scientific knowledge. Here, we review quantitative assessments of metaknowledge that trace the distribution of such processes at large scales. We argue that these distributional assessments, by characterizing the interaction and relative importance of competing processes, will not only provide new insight into the nature of science but will create novel opportunities to improve it.

Patterns of Scientific Content

The analysis of explicit knowledge content has a long history. Content analysis, or assessment of the frequency and co-appearance of words, phrases, and concepts throughout a text, has been pursued since the late 1600s, ranging from efforts in 18th-century Sweden to quantify the heretical content of a Moravian hymnal (*6*) to mid–20th-century studies of mass media content in totalitarian regimes. Contemporary approaches focus on the computational identification of "topics" in a corpus of texts. These can be tracked over time, as in a recent study of the news cycle (*7*). "Culturomics" projects now follow topics over hundreds of years, using texts digitized in the Google Books project (*3*). Topics can also be used to identify similarities between documents, as in topic modeling, which represents documents statistically as unstructured collections of "topics" or phrases (*8*).

With the rise of the Internet and computing power, statistical methods have also become central to natural language processing (NLP), including information extraction, information retrieval, automatic summarization, and machine reading. Advances in NLP have made it one of the most rapidly growing fields of artificial intelligence. Now that the vast majority of scientific publications are produced electronically (*5*), they are natural objects for topic modeling (*9*) and NLP. Some recent work, for example, uses computational parsing to extract relational claims about genes and proteins, and then compares these claims across hundreds of thousands of papers to reconcile contradictory results (*10*) and identify likely "missing" elements from molecular pathways (*11*). In such fields as biomedicine, electronic publications are further enriched with structured metadata (e.g., keywords) organized into hierarchical ontologies to enhance search (*12*). Citations have long been used in "scientometric" investigations to explore dependencies among

Department of Sociology, University of Chicago, Chicago, IL 60637, USA.

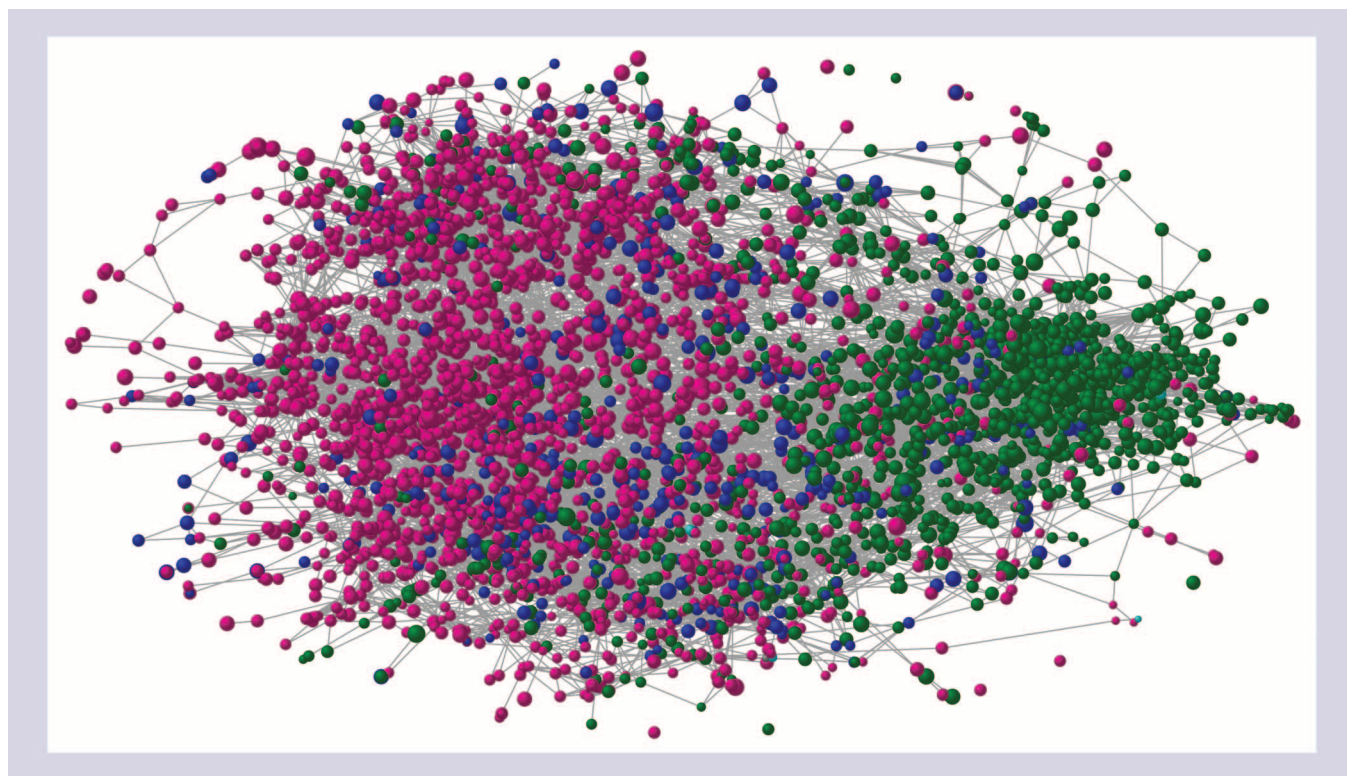*To whom correspondence should be addressed. E-mail: jevans@uchicago.edu

claims, authors, institutions, and fields (*13*). Search data produce another trace that can be analyzed (*14*). In public health, the changing tally of influenza-related Google searches has been used to predict emerging flu epidemics faster than can be accomplished by public health surveillance (*15*). Similar analysis could predict emerging research topics and fields.

The rise in scientific review articles and the concomitant explosion of scientific publications over the past century trace a growing supply and demand for the focused assessment and synthesis of research claims. As the number of analyses investigating a particular claim has become unmanageable [e.g., the efficacy of extrasensory perception (*16*); the influence of class size on student achievement (*17*); the role of β-amyloid in Alzheimer's disease (*18*)], researchers have increasingly engaged in meta-analysis—counting, weighting, and statistically analyzing a census of published findings on the topic (*19*, *20*). Whereas "the combination of observations" had been the central focus of 19th-century statistics (*21*), the combination of findings across articles was first formulated by Karl Pearson (*22*) regarding the efficacy of inoculation and later by Ronald Fisher in agricultural research (*23*). By the mid-20th century, with burgeoning scientific literatures, meta-analysis rapidly entered medicine, public health, psychology, and education. Scientists performing meta-analyses were forced to confront the "file-drawer problem": Negative and unpublishable findings never leave the "file drawer" (*19*, *24*). Indeed, all approaches to the analysis of explicit knowledge content aim to discover heretofore hidden regularities such as the file-drawer problem. These regularities in turn reveal the effects of implicit scientific content: detectable but inexplicit beliefs and practices.

Implicit Preferences, Heuristics, and Assumptions

Such implicit content includes a range of factors, from unstated preferences, tastes, and beliefs to the social processes of communication and citation. The file-drawer problem, for example, is driven by the well-attested preference for publishing positive results (*18*, *25*) and statistical findings that exceed arbitrary, field-specific thresholds (*26*). Such preferences may lead to a massive duplication of scientific effort through retesting doomed hypotheses. Magnifying their effect is a trend toward agreement with earlier results, which leads scientists to censor or reinterpret their data to be consistent with the past. Early results thus fossilize into facts through a cascade of positive citation, forming "microparadigms" (*27*, *28*). In choosing what parts of past knowledge to certify through positive citation, scientists are likely to accept authors with a history of success more readily. Scientific training further disciplines researchers to focus on established hubs of knowledge (*29*), with most articles shunning novelty to examine popular topics with popular methods. Even high-impact journals prefer publications on "hot topics"—albeit those using less popular methods (*30*).

Somewhat mitigating the trend toward assent and convergence, scientists often attempt to counter or extend high-profile research. This is particularly true for research staking novel claims

Fig. 1. Readers vary in the information they extract from an article. A new graduate student perceives a tiny fraction of available information, focusing on familiar authors, terms, references, and institutions. Her evaluation is limited to categorical classification (e.g., of the authors) into known and unknown ("important" and "unimportant"). For comparison she has the small collection of papers she has read. A leading scientist perceives a wealth of latent data, assembling individuals into mentorship relations and locating terms, as well as graphical and mathematical idioms, in historical and theoretical context. His evaluations generate rank orders based on his experience in the field. He can compare a paper to thousands, and searches a large literature efficiently. An appropriately trained computer would complement this expertise with quantification and scale. It can rapidly access quantitative and relational information about authors, terms, and institutions, and order these items along a range of measures. For comparison it can already access a large fraction of the scientific literature—millions of articles and an increasing pool of digitized books; in the future it will scrape further data from Web pages, online databases, video records of conferences, etc.

Fig. 2. Research dynamics as manifest in collaboration networks. The graphic represents the network of researchers who were co-authors of papers concerning the two largest biochemical clusters in PubMed—critical organic molecules (e.g., calcium, potassium, ATP) and neurotransmitters (e.g., norepinephrine, serotonin, acetylcholine)—during the period 1995 to 2000. Researchers are connected by a link if they have collaborated on at least one article; core researchers (shown here) have 25+ collaborators. Magenta nodes correspond to researchers who were primarily authors of papers concerning biochemicals within the largest cluster (critical organic molecules) and green nodes to authors primarily publishing in the second largest cluster (neurotransmitters). All other clusters are ignored. Dark blue nodes denote researchers, the majority of whose articles link the two chemical clusters. Light blue nodes are extreme "innovators" who publish mostly on new chemicals. The graphic is laid out using the Fruchterman-Reingold algorithm (63), which treats the network as a physical system, minimizing the energy if nodes are programmed to repel one another and links to draw them together like springs. The magenta and green nodes form relatively tight groups, indicating densely interlinked collaboration, whereas most interdisciplinary researchers occupy the boundary region between them. Extreme innovators (e.g., in the lower right) are marginal. This illustrates a suggestive homology between the semantic and social structure of science. The figure does not attempt to disentangle the relative primacy of social or semantic structure.

and can lead to the rapid alternation of conflicting findings known as the "Proteus" phenomenon (31). High-profile research attracts "more eyeballs," and a reliable negation of its findings will often attract considerable interest (32). Indeed, the individual incentive to publish in prestigious journals may itself be a distorting preference, potentially leading to higher incidence of overstated results (25, 33).

The foregoing discussion of implicit content traces the essential tension (34) between tradition and originality. Rather than assuming that these forces always and everywhere resolve in equilibrium, metaknowledge investigations have begun to model their relative strength in different fields and recalibrate scientific certainty in those findings. By making this implicit content available to researchers, metaknowledge could inform individual strategies about research investment, pointing out overgrazed fields where crowding leads to diminishing returns and forgotten opportunities where premature certainty has halted promising investigation.

These studies use simple analysis of explicit content to infer preferences and biases. More powerful methods of natural language processing and statistical analysis will be essential for revealing subtle content. Scientific documents will likely demonstrate a range of cognitive short cuts, such as the availability heuristic, in which data and hypotheses are weighted on the basis of how easily they come to mind (35, 36). Although such heuristics are individually "irrational" because they violate normative theories of probability and decision-making, they can be beneficial to science overall. For example, the vast majority (91%) of geologists who had worked on Southern Hemisphere samples supported the theory of continental drift, versus 48% of geologists who had not. Because Southern Hemisphere data supporting drift were more available to them, these geologists "irrationally" overweighted these data and hence became the core community that went on to build the case for plate tectonics (36). Identifying the distribution of these heuristics across scientific investigations will allow consid-

eration of their consequences, expose possible bias, and recalibrate scientific certainty in particular propositions (10).

Moreover, subtle but systematic regularities across articles within a scientific domain may signal the presence of "ghost theories": unstated assumptions, theories, or disciplinary paradigms that shape the type of reasoning and evidence deemed acceptable (37). For example, because it is widely supposed that most human psychological properties are universal, results from "typical" experimental participants (American undergraduates, in 67% of studies published in the *Journal of Personality and Social Psychology*) are often extended to the entire species (38). A recent meta-analysis (38) demonstrated that this assumption is false in several domains, including fundamental ones such as perception, and recommends expensive changes in sampling to correct for the resulting bias. Cognitive anthropologists (39), feminists (40), and ethnic studies scholars (41) have also pointed to examples of knowledge and ways of knowing that are distinct to social groups.

Computation can assist in the large-scale hunt for regularities (such as the frequent appearance of "undergraduate" in participant descriptions) that signal the presence of ghost theories, even when untied to pre-identifiable groups, and could eventually help to identify these unwritten axioms, opening them up to public debate and systematic testing.

Knowledge Context

Scientific documents contain both explicit and implicit content, both of which interact powerfully with the context of their production (*42*). For example, the reliability of a result increases if it is produced in several disparate labs rather than a few linked by shared methods or shared mentorship. Scientific training likely places a long-lasting stamp on a researcher (*43*). This suggests that tracing knowledge transmission from teacher to student could reveal much about the spread and entrenchment of ideas and practices. The changing organization of research also shapes research content, with teams increasingly producing the most highly cited research (*44*). Studies of the structure of collaboration networks (*45–47*) reveal intriguing disciplinary differences, but researchers are just beginning to explore the impact of teams and larger networks on the creation, diffusion, and diversity of knowledge content. Historians have documented instances in which research dynasties and larger institutions influence scientific knowledge (*48–50*), but the extraction of metaknowledge on the distribution of these influences would enable estimation of their aggregate capacity to channel the next generation of research. Figure 2 hints at the importance of understanding the relationship between social and scientific structures by showing how chemists and biologists whose collaborations bridge scientific subgroups tend to investigate reactions that themselves bridge distinct clusters of molecules. Investigation of this phenomenon should simultaneously explore the degree to which chemical structure shapes the social structure of investigation, and vice versa.

Teams and collaboration networks are embedded in larger institutional structures. Peer review, for example, is central to scientific evaluation and reward. Theoretical work on peer review, however, suggests potential inefficiencies and occasions for knowledge distortion (*51, 52*). Research also occurs in specific physical settings: universities, institutes, and companies that vary in prestige, access to resources, and cultures of scientific practice. Institutional reputations likely color the acceptance of research findings. Indeed, recent work on multi-institutional collaboration demonstrates that institutions tend to collaborate with others of similar prestige, potentially exacerbating this effect (*4*). One underexplored issue regards the influence of shared resources—databases, accelerators, telescopes—on the organization of related research and the pace of advance.

Given the capital intensity of much contemporary science, this question is critical for science policy. It also underscores the broader role of funding in the direction and success of research programs. For example, can focused investment unleash sudden breakthroughs, or is the slow development of community, shared culture and a toolkit more important to nurture a flow of discoveries? And what of private patronage? There is evidence from metaknowledge that embedding research in the private or public sector modulates its path. Company projects tend to eschew dogma in an impatient hunt for commercial breakthroughs, leading to rapid but unsystematic accumulation of knowledge, whereas public research focuses on the careful accumulation of consistent results (*53*).

The production of scientific knowledge is embedded in a broader social and technological context. A wealth of intriguing results suggests that this is a fruitful direction for further metaknowledge work. In biomedical research, for example, social inequalities and differences in media exposure (*54*) partially determine research priorities. Organized and well-connected groups lobby for increased investment in certain diseases, drawing resources away from maladies that disproportionately affect the poor and those in impoverished countries. The dearth of biomedical knowledge relevant to poorer countries is likely exacerbated by lack of access to science, as revealed by decreased citations to commercial access publications in poor countries (*55*).

The long-term impact of the Internet and related technologies on scientific production remains unclear. Early results indicate that online availability not only allows researchers to discover more diverse scientific content, but also makes visible other scientists' choices about what is important. This feedback leads to faster convergence on a shrinking subset of "superstar" articles (*5*). The Web also hosts radical experiments in the dissemination of scientific practices, such as myExperiment (*56*). It has fostered "open source" approaches to collaborative research, such as the Polymath project (*57*) and Zooniverse (*58*), and the provision of real-time knowledge services, such as the computational knowledge engine WolframAlpha (*59*). Online projects create novel opportunities to generate collaborations and data, but digital storage may also render their results more ephemeral (*60*). Understanding how social context and emerging media interact with scientific content will enable reevaluation not only of the science, but of the strengths and liabilities these new technologies hold for public knowledge.

Why Metaknowledge? Why Now?

The ecology of modern scientific knowledge constitutes a complex system: apparently complicated, involving strong interaction between components, and predisposed to unexpected collective outcomes (*61*). Although science has ever been so, the growing number of global scientists, increasingly connected via multiple channels—international conferences, online publications, e-mail, and science blogs—has increased this complexity. Rising complexity in turn makes the changing focus of research and the resolution of consensus less predictable. The informatics turn in the sciences offers a unique opportunity to mine existing knowledge for metaknowledge in order to identify and measure the complex processes of knowledge production and consumption. As such, metaknowledge research provides a high-throughput complement to existing work in social and historical studies of science by tracing the distribution and relative influence of distinct social, behavioral, and cognitive processes on science. Metaknowledge investigations will miss subtle regularities accessible to deep, interpretive analysis, and should draw on such work for direction. We argue, however, that some regularities will only be identifiable in the aggregate, especially those involving interrelations between competing processes. Once identified, these could become fruitful subjects for interpretive investigation.

Successfully executing the more ambitious parts of the metaknowledge program will require further improvements in machine reading and inference technologies. Systematic analysis of some elements of scientific production will remain out of reach (*62*). Nonetheless, as metaknowledge grows in sophistication and reliability, it will provide new opportunities to recursively shape science—to use measured biases, revealed assumptions, and previously unconsidered research paths to revise our confidence in bodies of knowledge and particular claims, and to suggest novel hypotheses. The computational production and consumption of metaknowledge will allow researchers and policy-makers to leverage more scientific knowledge—explicit, implicit, contextual—in their efforts to advance science. This will become essential in an era when so many investigations are linked in so many ways.

References and Notes

1. D. R. Swanson, *Bull. Med. Libr. Assoc.* **78**, 29 (1990).
2. D. R. Swanson, N. R. Smalheiser, *Artif. Intell.* **91**, 183 (1997).
3. J.-B. Michel *et al.*, *Science* **331**, 176 (2011); 10.1126/science.1199644.
4. B. F. Jones, S. Wuchty, B. Uzzi, *Science* **322**, 1259 (2008); 10.1126/science.1158357.
5. J. A. Evans, *Science* **321**, 395 (2008).
6. K. Dovring, *Public Opin. Q.* **18**, 389 (1954).
7. J. Leskovec, L. Backstrom, J. Kleinberg, in *Proceedings of the 15th Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining, International Conference on Knowledge Discovery and Data Mining*, Paris, 28 June to 1 July 2009.
8. T. L. Griffiths, M. Steyvers, *Proc. Natl. Acad. Sci. U.S.A.* **101** (suppl. 1), 5228 (2004).
9. K. K. Mane, K. Börner, *Proc. Natl. Acad. Sci. U.S.A.* **101** (suppl. 1), 5287 (2004).
10. A. Rzhetsky, T. Zheng, C. Weinreb, *PLoS ONE* **1**, e61 (2006).
11. C. J. Krieger *et al.*, *Nucleic Acids Res.* **32** (database issue), D438 (2004).
12. A. H. Renear, C. L. Palmer, *Science* **325**, 828 (2009).
13. B. Cronin, H. B. E. Atkins, *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (Information Today Inc., Medford, NJ, 2000).

14. J. Bollen *et al.*, *PLoS ONE* **4**, e4803 (2009).
15. J. Ginsberg *et al.*, *Nature* **457**, 1012 (2009).
16. J. Pratt, J. Rhine, B. Smith, C. Stuart, J. Greenwood, *Extra-Sensory Perception After Sixty Years: A Critical Appraisal of the Research in Extra-Sensory Perception* (Holt, New York, 1940).
17. G. V. Glass, M. L. Smith, *Educ. Eval. Policy Anal.* **1**, 2 (1979).
18. S. A. Greenberg, *BMJ* **339**, b2680 (2009).
19. J. E. Hunter, F. L. Schmidt, *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Sage, Thousand Oaks, CA, ed. 2, 2004).
20. Cochrane Collaboration Reviewer's Handbook, www.cochrane.org/training/cochrane-handbook.
21. S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900* (Belknap/Harvard Univ. Press, Cambridge, MA, 1986).
22. K. Pearson, *Br. Med. J.* **3**, 1243 (1904).
23. K. O'Rourke, *J. R. Soc. Med.* **100**, 579 (2007).
24. L. V. Hedges, *Stat. Sci.* **7**, 246 (1992).
25. J. P. Ioannidis, *PLoS Med.* **2**, e124 (2005).
26. A. S. Gerber, N. Malhotra, *Sociol. Methods Res.* **37**, 3 (2008).
27. A. Rzhetsky, I. Iossifov, J. M. Loh, K. P. White, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 4940 (2006).
28. U. Shwed, P. S. Bearman, *Am. Sociol. Rev.* **75**, 817 (2010).
29. M. Cokol, I. Iossifov, C. Weinreb, A. Rzhetsky, *Nat. Biotechnol.* **23**, 1243 (2005).
30. M. Cokol, R. Rodriguez-Esteban, A. Rzhetsky, *Genome Biol.* **8**, 406 (2007).
31. J. P. Ioannidis, T. A. Trikalinos, *J. Clin. Epidemiol.* **58**, 543 (2005).
32. M. Cokol, I. Iossifov, R. Rodriguez-Esteban, A. Rzhetsky, *EMBO Rep.* **8**, 422 (2007).
33. J. P. Ioannidis, *JAMA* **294**, 218 (2005).
34. T. S. Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change* (Univ. of Chicago Press, Chicago, 1977).
35. A. Tversky, D. Kahneman, *Science* **185**, 1124 (1974).
36. M. Solomon, *Philos. Sci.* **59**, 439 (1992).
37. D. L. Smail, *On Deep History and the Brain* (Univ. of California Press, Berkeley, CA, 2008).
38. J. Henrich, S. J. Heine, A. Norenzayan, *Behav. Brain Sci.* **33**, 61 (2010).
39. M. E. Harris, Ed., *Ways of Knowing: New Approaches in the Anthropology of Experience and Learning* (Berghahn, Oxford, 2007).
40. E. Martin, *Signs* **16**, 485 (1991).
41. R. Barnhardt, *Anthropol. Educ. Q.* **36**, 8 (2005).
42. K. Knorr-Cetina, *Epistemic Cultures: How the Sciences Make Knowledge* (Harvard Univ. Press, Cambridge, MA, 1999).
43. R. Collins, *The Sociology of Philosophies: A Global Theory of Intellectual Change* (Belknap/Harvard Univ. Press, Cambridge, MA, 1998).
44. S. Wuchty, B. F. Jones, B. Uzzi, *Science* **316**, 1036 (2007); 10.1126/science.1136099.
45. M. E. J. Newman, *Phys. Rev. E* **64**, 016131 (2001).
46. M. E. J. Newman, *Phys. Rev. E* **64**, 016132 (2001).
47. M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **101** (suppl. 1), 5200 (2004).
48. D. J. Kevles, *The Physicists: The History of a Scientific Community in Modern America* (Knopf, New York, ed. 1, 1978).
49. T. Lenoir, *Instituting Science: The Cultural Production of Scientific Disciplines* (Stanford Univ. Press, Stanford, CA, 1997).
50. P. Coffey, *Cathedrals of Science: The Personalities and Rivalries That Made Modern Chemistry* (Oxford Univ. Press, London, 2008).
51. S. Allesina, http://arxiv.org/abs/0911.0344 (2009).
52. S. Thurner, R. Hanel, http://arxiv.org/abs/1008.4324 (2010).
53. J. A. Evans, *Soc. Stud. Sci.* **40**, 757 (2010).
54. E. M. Armstrong, D. P. Carpenter, M. Hojnacki, *J. Health Polit. Policy Law* **31**, 729 (2006).
55. J. A. Evans, J. Reimer, *Science* **323**, 1025 (2009).
56. myExperiment, www.myexperiment.org.
57. The Polymath Blog, http://polymathprojects.org.
58. Zooniverse, www.zooniverse.org/home.
59. WolframAlpha, www.wolframalpha.com.
60. E. Evangelou, T. A. Trikalinos, J. P. Ioannidis, *FASEB J.* **19**, 1943 (2005).
61. R. Foote, *Science* **318**, 410 (2007).
62. H. M. Collins, *Sci. Stud.* **4**, 165 (1974).
63. T. M. J. Fruchterman, E. M. Reingold, *Softw. Pract. Exper.* **21**, 1129 (1991).
64. This research benefited from NSF grant 0915730 and responses at the U.S. Department of Energy's Institute for Computing in Science (ICiS) workshop "Integrating, Representing, and Reasoning over Human Knowledge: A Computational Grand Challenge for the 21st Century." We thank K. Brown, E. A. Cartmill, M. Cartmill, and two anonymous reviewers for their detailed and constructive comments on this essay.

PERSPECTIVE

Access to Stem Cells and Data: Persons, Property Rights, and Scientific Progress

Debra J. H. Mathews,[1]* Gregory D. Graff,[2] Krishanu Saha,[3] David E. Winickoff[4]

Many fields have struggled to develop strategies, policies, or structures to optimally manage data, materials, and intellectual property rights (IPRs). There is growing recognition that the field of stem cell science, in part because of its complex IPRs landscape and the importance of cell line collections, may require collective action to facilitate basic and translational research. Access to pluripotent stem cell lines and the information associated with them is critical to the progress of stem cell science, but simple notions of access are substantially complicated by shifting boundaries between what is considered information versus material, person versus artifact, and private property versus the public domain.

Access to data and materials is critical to the progress of science generally (*1–3*), but plays a particularly important role in stem cell science. In the field of pluripotent stem cell research, data and information associated with cell lines are essential to their utility and management. A number of factors currently limit the sharing of data and materials in the field, including strategic behavior of individual scientists, ethical intricacies in using human cell lines, and a complex landscape of intellectual property rights (IPRs). Efforts to manage proper-

ty relations and develop an accountable system of stewardship to handle the ethical and legal obligations to materials donors could help ameliorate these obstacles, benefiting academic and industry scientists and promoting both basic and translational research.

Difficulties in procuring and managing human cells arise because they are increasingly transcending three distinctions drawn in ethics, law, and common practice—distinctions between information and materials, persons and artifacts, and private property and the public domain

(Fig. 1). Some allowance for blurring between and across these conventional distinctions can be useful and productive for science, but it also creates substantial challenges. For example, a cell line is much richer as a research tool by virtue of its connection to an individual human, but that connection also raises privacy concerns if the tool is shared.

Work to Date, Work to Do

Other research communities in the life sciences have experienced problems similar to those outlined above, and some have addressed them through new institutions such as public DNA sequence databases, tissue banks, and mouse repositories (*3*). Within stem cell science, there are important efforts under way to improve access to both cell lines and associated information [e.g., UK Stem Cell Bank, European Human Embryonic Stem Cell Registry (hESCreg)] (*4*). However, these efforts struggle to keep pace with evolving data requirements and the proliferation of new induced pluripotent stem cell lines and do not provide all of the information about

[1]Johns Hopkins Berman Institute of Bioethics, Deering Hall, 208, 1809 Ashland Avenue, Baltimore, MD 21205, USA. [2]Department of Agricultural and Resource Economics, B328 Clark Hall, Colorado State University, Fort Collins, CO 80523, USA. [3]The Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [4]Department of Environmental Science, Policy, and Management, 115 Giannini Hall, University of California, Berkeley, CA 94720 USA.

*To whom correspondence should be addressed. E-mail: dmathews@jhu.edu