Published in final edited form as:
Science. 2010 July 23; 329(5990): 399–400. doi:10.1126/science.1189416.

PHILOSOPHY OF SCIENCE:

Machine Science

James Evans and **Andrey Rzhetsky**
Department of Sociology, University of Chicago, Chicago, IL 60637, USA.

Scientists today cannot hope to manually track all of the published science relevant to their work. A cancer biologist, for instance, can find more than 2 million relevant papers in the PubMed archive, more than 200 million Web pages with a Google search, and databases holding results from experiments that produce millions of gigabytes of data.

This explosion of knowledge is changing the landscape of science. Computers already play an important role in helping scientists store, manipulate, and analyze data. New capabilities, however, are extending the reach of computers from analysis to hypothesis. Drawing on approaches from artificial intelligence, computer programs increasingly are able to integrate published knowledge with experimental data, search for patterns and logical relations, and enable new hypotheses to emerge with little human intervention. Scientists have used such computational approaches to repurpose drugs, functionally characterize genes, identify elements of cellular biochemical pathways, and highlight essential breaches of logic and inconsistency in scientific understanding. We predict that within a decade, even more powerful tools will enable automated, high-volume hypothesis generation to guide high-throughput experiments in biomedicine, chemistry, physics, and even the social sciences (1).

Proponents of data-driven science (2–4) conjecture that hypotheses are obsolete: New knowledge will simply emerge from mechanical application of algorithms that mine data for plausible patterns. This approach is attractive, but there are potential pitfalls. The discovery of patterns from data alone is similar to the task faced by an explorer in an unfamiliar jungle, without a guide. With no sense of what is already known about the environment or its perils, she is likely to misclassify what she sees—fearing the intimidating but harmless snake; ignoring the tiny lethal frog.

Recent research demonstrates how scientists can use computers to become better-informed and more agile explorers. New computational tools can expand the pool of concepts and relations used for generating automated hypotheses by (i) drawing more from the vast corpus of published science, and (ii) synthesizing new higher- and lower-order concepts and relations from the existing pool of knowledge. This approach can enable scientists studying a particular natural system, such as a biochemical pathway, to identify and fill in missing pieces, and traverse reasoning chains much longer than those possible with the unaided mind. For example, researchers have used computation to increase the number of candidate genetic aberrations considered in synthesizing hypotheses about disease (5–7). They have also increased the number of potential biological activities involved in describing new gene functions (8, 9) and ironed out past errors (10). Similarly, scientists have used computation to increase the potential number of proteins and metabolites involved in biochemical networks, and to generate predictions about which locations in those networks could be

jevans@uchicago.edu.

Supporting Online Material www.sciencemag.org/cgi/content/full/329/5990/399/DC1 Fig. S1 altered to improve health (11) and to identify elements misidentified as participating in a network (12).

Merely increasing the pool of concepts and relations, however, would simply generate multitudes of low-quality hypotheses. Scientists can profitably restrict that multitude by using a selection process that draws on insights into the social, cultural, and cognitive production of science. For example, Swanson pioneered the ABC model of hypothesis generation, which focuses on hypotheses that cross boundaries between distinct scientific literatures. If concepts A and B are studied in one literature, and B and C in another, Swanson assumed transitivity to hypothesize that A implies C (see the figure and fig. S1). He then demonstrated that novel A-to-C inferences were likely to be true, although unlikely to be arrived at via other means (13–16). Through this approach, Swanson hypothesized that fish oil could lessen the symptoms of Raynaud's blood disorder and that magnesium deficits are linked to migraine headaches. This heuristic relies on an implicit understanding of scientific communities and publishing norms. It assumes that unpublished ideas within a research community are less valuable than ideas that link seemingly unrelated communities. Within a subfield, scientists are typically familiar with all of "their own" ideas, so unpublished connections more likely represent negative knowledge—superficially plausible ideas that participants know are wrong from experience. Unpublished ideas about subjects (such as the role of particular molecules or genes) that cross subfield boundaries, however, are much more likely to represent unasked questions. A recent analysis of biomolecules common to several fields of biomedicine, for instance, suggests that many communities could profit from generating predictions that bridge field boundaries and link disparate properties of these molecules or other scientific concepts. (17).

Automated expansion of concepts and relations across community boundaries is severely constrained by incompatibilities in the language used by different scientific communities (18, 19). Because subfields have distinct histories, they often use different language to express the same concept, or similar terms to refer to unrelated entities. If researchers could computationally map these languages onto one another, as some are beginning to do with medical terminologies (20), they could vastly increase the number of possible hypotheses. Mapping concepts across languages would highlight parallels in theories from different domains, as well as changes in meaning with time (semantic drift) and multiple meanings (polysemy). These differences could be computationally mined to identify novel conceptual linkages. By prioritizing hypotheses that contain concepts spanning existing scientific theories, languages, and cultures, investigators could productively focus on the most novel (21).

Analysts can also increase the pool of concepts and relations by computationally synthesizing new concepts and relations from those previously published. Computers have been deployed to "coarse grain" or identify new, higher-order aggregates of established concepts within studies of biological pathways, medical syndromes, and social classes (22). Scientists have also discovered new relations by identifying regular similarities between existing elements (23, 24). They have efficiently constricted the vast number of possible new aggregates by focusing on those that share physical properties or patterns, or integrate components of a broader system, such as a particular disease.

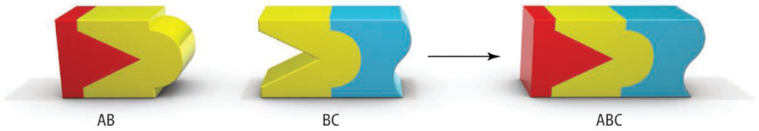
In the past, computational approaches have been more successful in small, well-defined systems than in larger, less studied, or more complex ones. The explosion of data from high-throughput experiments, however, increasingly presents researchers with very complicated systems. Facing these data with questions equal in scale and complexity will be critical because, in the words of Mark Twain, "you can't depend on your eyes when your imagination is out of focus" (25).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References and Notes

1. Lazer D, et al. Science. 2009; 323:721. [PubMed: 19197046]
2. Hood L. BioCommerce Week. Sep 9.2004
3. Golub T. Nature. 2010; 464:679. [PubMed: 20360719]
4. O'Malley MA, Elliott KC, Haufe C, Burian RM. Cell. 2009; 138:611. [PubMed: 19703386]
5. Perez-Iratxeta C, Bork P, Andrade MA. Nat. Genet. 2002; 31:316. [PubMed: 12006977]
6. Iossifov I, Zheng T, Baron M, Gilliam TC, Rzhetsky A. Genome Res. 2008; 18:1150. [PubMed: 18417725]
7. Liu J, et al. Science. 2009; 323:1218. [PubMed: 19164706]
8. King RD, et al. Nature. 2004; 427:247. [PubMed: 14724639]
9. King RD, et al. Science. 2009; 324:85. [PubMed: 19342587]
10. Blake J. personal interview. Feb 5.2009
11. Ruths DA, et al. J. Comput. Biol. 2006; 13:1546. [PubMed: 17147477]
12. Hsiao TL, et al. Nat. Chem. Biol. 2010; 6:34. [PubMed: 19935659]
13. Swanson DR. Perspect. Biol. Med. 1986; 30:7. [PubMed: 3797213]
14. Swanson DR. Bull. Med. Libr. Assoc. 1990; 78:29. [PubMed: 2403828]
15. Weeber M, et al. J. Am. Med. Inform. Assoc. 2003; 10:252. [PubMed: 12626374]
16. Swanson DR, Smalheiser NR. Artif. Intell. 1997; 91:183.
17. Cokolr M, et al. Nat. Biotechnol. 2005; 23:1243. [PubMed: 16211067]
18. Harris, ZS. Sublanguage: Studies of Language in Restricted Semantic Domains. Kittredge, R.; Lehrberger, J., editors. de Gruyter; Berlin: 1982. p. 231-236.
19. Harris, ZS. The Form of Information in Science: Analysis of an Immunology Sublanguage. Kluwer Academic; Dordrecht, Netherlands: 1989.
20. The National Library of Medicine's Metathesaurus MetaMap connects more than 100 distinct medical terminologies.
21. Wren JD. BMC Bioinform. 2004; 5:145.
22. Feret J, Danos V, Krivine J, Harmer R, Fontana W. Proc. Natl. Acad. Sci. U.S.A. 2009; 106:6453. [PubMed: 19346467]
23. Schmidt M, Lipson H. Science. 2009; 324:81. [PubMed: 19342586]
24. Bongard J, Lipson H. Proc. Natl. Acad. Sci. U.S.A. 2007; 104:9943. [PubMed: 17553966]
25. Twain, M. A Connecticut Yankee in King Arthur's Court. Webster; New York: 1889.



1. . Logical leaps

Scientific knowledge and concepts can be represented as jigsaw puzzle pieces that, with the help of new computational tools, can be assembled into new hypotheses. In Swanson's ABC model, if the literature from one scientific subfield includes two concepts (A, red, and B, yellow), and the literature from another subfield includes B and C (blue), then an analyst may computationally infer that A and C are directly or indirectly related, potentially leading to new hypotheses that cross subfield boundaries.