ARTIFICIAL INTELLIGENCE

Amplify scientific discovery with artificial intelligence

Many human activities are a bottleneck in progress

By **Yolanda Gil,**[1] **Mark Greaves,**[2] **James Hendler,**[3]* **Haym Hirsh**[4]

Technological innovations are penetrating all areas of science, making predominantly human activities a principal bottleneck in scientific progress while also making scientific advancement more subject to error and harder to reproduce. This is an area where a new generation of artificial intelligence (AI) systems can radically transform the practice of scientific discovery. Such systems are showing an increasing ability to automate scientific data analysis and discovery processes, can search systematically and correctly through hypothesis spaces to ensure best results, can autonomously discover complex patterns in data, and can reliably apply small-scale scientific processes consistently and transparently so that they can be easily reproduced. We discuss these advances and the steps that could help promote their development and deployment.

POLICY

Applying AI to the practice of science is not new. AI pioneer and Nobel laureate Herbert Simon hypothesized that cognitive mechanisms involved in scientific discovery are a special case of general human capabilities for problem-solving and, with colleagues, developed systems in the 1970s and 1980s that demonstrated reasoning capabilities for analyzing scientific data (*1*). Also in the 1970s, Joshua Lederberg (another Nobel winner) and colleagues developed the DENDRAL system for analyzing mass spectrometry data in order to hypothesize molecular structures (*2*). More recent breakthroughs, such as robot scientists and software that formulates laws for complex dynamical systems, demonstrate broader applicability of AI techniques for scientific discovery (*3*).

Over the past two decades, AI has seen accelerating scientific advances and concomitant commercial-sector successes because of advances on three fronts: steady scholarly advances, especially as success has increased the numbers of interested participants; Moore's law and steady exponential increases in computing power; and exponential increases in, and broad availability of, relevant data in volumes never previously seen. Those scientific efforts that have leveraged AI advances have largely harnessed sophisticated machine-learning techniques to create correlative predictions from large sets of "big data." Such work aligns well with the current needs of peta- and exascale science. However, AI has far broader capacity to accelerate scientific discovery, and AI-based systems that can represent hypotheses, reason with models of the data, and design hypothesis-driven data collection techniques can reduce the error-prone human bottleneck in scientific discovery.

Schematic of Hanalyzer. Hanalyzer is an example of a modern scientific discovery system. It integrates assertions from biomedical databases and then reasons about the resulting semantic information to suggest novel correlations from which scientists can generate testable hypotheses.

SEARCHING AND SYNTHESIZING. What do these intelligent systems look like today? AI techniques are amplifying existing tools in identifying relevant results from the broader scientific community. Search engines are some of the most important and frequently used tools in the general scientific arsenal. Major search engines all use AI techniques for tasks like query suggestion and result customization. Increasingly, scientists in many fields are augmenting the power of search by using machine-readable ontologies and Semantic Web technology (*4*) to tag not just scientific articles but also figures and videos, blogs, data sets, and computational services, which allow information-finding beyond current search limitations.

We can project a not-so-distant future where "intelligent science assistant" programs identify and summarize relevant research described across the worldwide multilingual spectrum of blogs, preprint archives, and discussion forums; find or generate new hypotheses that might confirm or conflict with ongoing work; and even rerun old analyses when a new computational method becomes available. Aided by such a system, the scientist will focus on more of the creative aspects of research, with a larger fraction of the routine work left to the artificially intelligent assistant.

New types of intelligent systems that can enhance scientific efforts in this manner are transitioning from academic and industrial research laboratories. A term gathering popularity for systems that intelligently process online information beyond search is "cognitive computing," used by IBM in describing the Watson system that beat the best human players in the televised *Jeopardy!* game (*5*). One kind of cognitive system includes language-based programs like Watson, which is now being used by IBM and a number of prominent medical centers in developing tools for improving medical treatment by helping doctors keep up with constantly changing medical literature. To enhance these capabilities, the U.S. Defense Advanced Research Projects Agency (DARPA) recently announced a major effort to synthesize new systems-biology models of cancer by knitting together fragmentary causal hypotheses gathered by automatically reading papers in the literature (*6*). Another group of cognitive systems, based largely on advances in neural networks and neurologically inspired computation, is beginning to show promise in the analysis of nontextual processing, especially of online images and video, across a wide range of areas including

[1] *Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA.* [2] *Pacific Northwest National Laboratory, Richland, WA 99354, USA.* [3] *Information Technology and Web Science, Rensselaer Polytechnic Institute, Troy, NY 12203, USA.* [4] *Cornell University, Ithaca, NY 14850, USA. *E-mail: hendler@cs.rpi.edu*

ILLUSTRATION: P. HUEY/SCIENCE

Published by AAAS

biological imaging (7), species preservation (8), and quantum chemistry (9).

DIGESTING DATA. AI techniques have accelerated the pace and quality of analysis of the huge quantities of data that can stream from modern laboratory equipment. To derive scientific insight from data at this scale, standard methods include applying dimensionality-reduction techniques and feature extractors to create high-speed classifiers based on machine-learning approaches, such as Bayesian networks or support-vector machines. Because the phenomena under study often exist in nonstationary environments or in contexts with only small quantities of labeled data that can be used for training—complex, unsupervised, and reinforcement machine-learning techniques are critical for data analysis. These types of approaches are being used in recent projects in data-rich areas as diverse as chemical structure prediction, pathway analysis and identification in systems biology, the processing of large-scale geophysics data, and others.

Another, more ambitious class of intelligent systems is being developed under the rubric of Discovery Science or, increasingly, Discovery Informatics (10). These systems enhance the intelligent assistants described earlier with the capability to attack scientific tasks that combine rote work with increasing amounts of adaptivity and freedom. These systems use encoded knowledge of scientific domains and processes in order to assist with tasks that previously required human knowledge and reasoning. In fact, several sciences have significant investments in the representation of vast amounts of scientific knowledge and are poised to explore new intelligent systems that exploit that knowledge for discovery.

For example, the Hanalyzer (short for high-throughput analyzer) uses natural language processing to automatically extract a semantic network from all PubMed papers relevant to a specific scientist, uses Semantic Web technology to integrate assertions from other biomedical sources, and reasons about the network to find new correlations that suggest new genes to investigate (11) (see the figure). The Wings system uses Semantic Web technologies and AI planning to reason about specific choices of models and algorithms for water-quality data and customizes workflows automatically for daily conditions (12). Eureqa, usable in many scientific fields, searches a vast space of hypotheses consistent with given data observed in an experiment, selects those most promising, and designs experiments to test them (13). Sunfall incorporates usability principles and cognitive load considerations in the design of a visual analytics interface; this reduces scientists' workload and false-positive rates in identifying supernovae (14).

These four systems are representative of the ways that more advanced AI can serve scientific ends. They are based on explicit representations of science processes, and they reason about these to automate processes and assist the human scientist. Development of the explicit representations of scientific processes on which they are based is complex. When successful, the computer can become a real (although junior) participant in the science process, doing what it does best: applying algorithmic methods and bringing knowledge to bear in a consistent, systematic, and complete manner.

A VIRTUOUS CIRCLE. Developing systems like these is not just an exercise in AI application—it affects the direction of AI research. Addressing real challenges of science pushes the AI envelope in many areas, including knowledge representation, automatic inference, process reasoning, hypothesis generation, natural language processing, machine learning, collaborative interaction, and intelligent user interfaces. This interaction creates a virtuous circle where advances in science go hand in hand with advances in AI. This virtuous circle can only work well if balanced and well oiled.

> ***"AI-based systems that can represent hypotheses ... can reduce the error-prone human bottleneck in ... discovery."***

What are the best ways to immerse AI research into scientific practice so that it can deliver on this promise? First, and perhaps most obviously, is conceiving new means of bringing interdisciplinary research teams together at an earlier stage of research and in a sustainable manner. Increasingly, there is a realization in academia that scientists must gain broad knowledge and skills in computation and programming. This should include AI components—training and supporting students and young researchers. In addition, basic research to advance AI in domains of science practice needs to be facilitated and rewarded in academia, as standard criteria for research merit focus primarily on theoretical advances in computing per se and thus do not transfer well to this kind of multidisciplinary research.

A significant challenge that appears to be specific to AI is to attract scientific researchers to engage in this joint research. Scientists have made significant investments in the past in advanced computing technologies, such as high-end computing, distributed databases, and sensor networks. However, their interest in AI seems relatively limited. With AI systems having impacts in the consumer sector (e.g., speech recognition systems, real-time automated language translation, and self-driving cars and self-navigating drones), why are scientists not enthusiastic about embracing AI?

One hypothesis is the lack of clear methods to measure the impact of AI in science. There are exceptions in some areas of AI, such as machine learning and language processing, where metrics to compare systems have been defined and improvement has been measured. But there has been little research into such measurement more generally, especially for the heuristic methods of the reasoning field. Methods to quantify significant advances because of the use of new AI technologies in scientific fields are needed to validate the impact of AI on scientific discovery.

Another reason may be the limited work of the AI community in disseminating and marketing ideas to scientists. Although many non-AI scientists attend supercomputing and database conferences, few are compelled to attend an AI conference. Possibly, researchers are influenced by the unrealistic science fiction images of super-smart machines, rather than the realities of current technological advances. Understanding the sources of hesitation of scientists to embrace AI will be a first step toward changing the culture and bringing these communities together.

The world faces deep problems that challenge traditional methodologies and ideologies. These challenges will require the best brains on our planet. In the modern world, the best brains are a combination of humans and intelligent computers, able to surpass the capabilities of either one alone. ■

REFERENCES AND NOTES

1. P. Langley, H. A. Simon, G. L. Bradshaw, J. M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Processes* (MIT Press, Cambridge, MA, 1987).
2. R. K. Lindsay *et al.*, *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project* (McGraw-Hill, New York, 1980).
3. D. Waltz, B. G. Buchanan, *Science* **324**, 43 (2009).
4. J. Hendler, *Science* **299**, 520 (2003).
5. J. Kelly III, S. Hamm, *Smart Machines* (Columbia Univ. Press, New York, 2013).
6. Information Innovation Office, DARPA; www.darpa.mil/Our_Work/I2O/Programs/Big_Mechanism.aspx.
7. S. N. Deepa, B. Aruna Devi, *Indian J. Sci. Technol.* **4**, 1538 (2011).
8. M. Martialay, "Citizen Scientist," The Approach; http://approach.rpi.edu/2014/04/25/citizen-scientist-your-safari-photos-are-the-data/.
9. C. Caetano *et al.*, *Int. J. Quantum Chem.* **111**, 2732 (2011).
10. S. M. Leach *et al.*, *PLOS Comput. Biol.* **5**, e1000215 (2009).
11. Y. Gil *et al.*, *IEEE Intell. Syst.* **26**, 62 (2011).
12. Eureqa desktop, www.nutonian.com/products/eureqa/.
13. C. R. Aragon, S. J. Bailey, S. Poon, K. Runge, R. C. Thomas, *J. Phys. Conf. Ser.* **125**, 012091 (2008).

10.1126/science.1259439

Published by AAAS