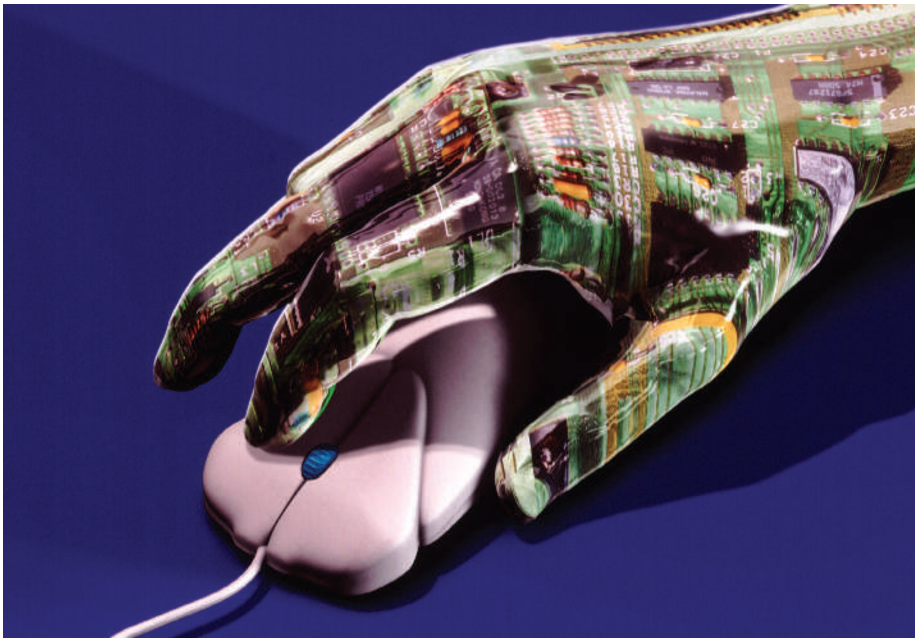
COMMENTARY

Exceeding human limits

Scientists are turning to automated processes and technologies in a bid to cope with ever higher volumes of data. But automation offers so much more to the future of science than just data handling, says **Stephen H. Muggleton**.

The collection and curation of data throughout the sciences is becoming increasingly automated. For example, a single high-throughput experiment in biology can easily generate more than a gigabyte of data per day, and in astronomy automatic data collection leads to more than a terabyte of data per night. Throughout the sciences the volumes of archived data are increasing exponentially, supported not only by low-cost digital storage but also by the growing efficiency of automated instrumentation. It is clear that the future of science involves the expansion of automation in all its aspects: data collection, storage of information, hypothesis formation and experimentation (see table). Future advances will have the ability to yield powerful new forms of science that could blur the boundaries between theory and experiment. However, to reap the full benefits it is essential that developments in high-speed automation are not introduced at the expense of human understanding and insight.



FIREFLY PRODUCTIONS/CORBIS

Many aspects of science are already unsustainable without the aid of computers.

During the twenty-first century, it is clear that computers will continue to play an increasingly central role in supporting the testing, and even formulation, of scientific hypotheses. This traditionally human activity has already become unsustainable in many sciences without the aid of computers. This is not only because of the scale of the data involved but also because scientists are unable to conceptualize the breadth and depth of the relationships between relevant databases without computational support. The potential benefits to science of such computerization are high — knowledge derived from large-scale scientific data could well pave the way to new technologies, ranging from personalized medicines to methods for dealing with and avoiding climate change[1].

In the 1990s it took the international human genome project a decade to determine the sequence of a single human genome; but projected increases in the speed of gene sequencing imply that before 2050 it will be feasible to determine the complete genome of every individual human being on Earth. Owing to the scale and rate of data generation, computational models of scientific data now require automatic construction and modification. We are seeing a range of techniques from mathematics, statistics and computer science being used to create scientific models from empirical data in an increasingly automated way. For instance, in meteorology and epidemiology, large-scale empirical data are routinely used to check the predictions of differential-equation models concerning climate variation and the spread of diseases.

Meanwhile, machine-learning techniques from computer science (including neural nets and genetic algorithms) are being used to automate the generation of scientific hypotheses from data. Some of the more advanced forms of machine learning enable new hypotheses, in the form of logical rules and principles, to be extracted relative to predefined background knowledge. This background knowledge is formulated and revised by human scientists, who also judge the new hypotheses and may attempt to refute them experimentally. For example, within the past decade researchers in my group have used inductive logic programming (a subdiscipline of machine learning) to discover key molecular substructures within a class of potential cancer-producing agents[2]. Building on the same techniques, we have more recently been able to generate experimentally testable claims about the toxic properties of hydrazine from experimental data — in this instance, from analyses of metabolites in rat urine following low doses of the toxin[3].

CHANGES TO TRADITIONAL SCIENCE WITH AUTOMATION

Traditional science	Automated science
Hypotheses	Machine-encoded logical hypotheses
Chemical knowledge	Machine-encoded chemical algebra
Experiments	Chemical Turing machine programs
Experimental design	Decision theory

Mixing maths

In other sciences, the reliance on computational modelling has arguably moved to a new level. In systems biology, the need to account for complex interactions within cells — in gene transduction, signalling and metabolic pathways — is requiring new and richer systems-level modelling. Traditional reductionist approaches in this area concentrated on understanding the functions of individual genes in isolation. However, genome-wide instrumentation, including microarray technologies, are leading to a system-level approach to biomolecules and pathways, and to the formulation and testing of models that describe the detailed behaviour of whole cells. This is new territory for the natural sciences and has resulted in multidisciplinary international projects such as the virtual E-Cell[4].

One obstacle to rapid progress in systems biology is the incompatibility of existing models. Often models that account for the shape and charge distribution of individual molecules need to be integrated with models describing the interdependency of chemical reactions. However, differences in the mathematical underpinnings of, say, differential equations, bayesian networks and logic programs make integrating these various models virtually impossible. Although hybrid models can be built by simply patching two models together, the underlying differences lead to unpredictable and error-prone behaviour when changes are made.

"Owing to the scale and rate of data generation, computational models of scientific data now require automatic construction and modification."

One encouraging development in this respect is the emergence within computer science of new formalisms[5] that integrate, in a sound fashion, two major branches of mathematics: mathematical logic and probability calculus. Mathematical logic provides a formal foundation for logic programming languages such as Prolog, whereas probability calculus provides the basic axioms of probability for statistical models, such as bayesian networks. The resulting 'probabilistic logic' is a formal language that supports statements of sound inference, such as 'The probability of A being true if B is true is 0.7'. Pure forms of existing probabilistic logic are unfortunately computationally intractable. However, an increasing number of research groups have developed machine-learning techniques that can handle tractable subsets of probabilistic logic[6]. Although it is early days, such research holds the promise of sound integration of scientific models from the statistical and computer-science communities

Miniature roboscientists

Statistical and machine-learning approaches to building and updating scientific models typically use 'open loop' systems with no direct link or feedback to the collection of data. A robot-scientist project in which I was involved offers an important exception[7]. Here, laboratory robots conducted experiments on yeast (*Saccharomyces cerevisiae*) using a process known as 'active learning'. The aim was to determine the function of several gene knockouts by varying the quantities of nutrient provided to the yeast. The robot used a form of inductive logic programming to select experiments that would discriminate between contending hypotheses. Feedback on each experiment was provided by data reporting yeast survival or death. The robot strategy that worked best (lowest cost for a given accuracy of prediction) not only outperformed two other automated strategies, based on cost and random-experiment selection, but also outperformed humans given the same task.

One exciting development that we might expect in the next ten years is the construction of the first microfluidic robot scientist, which would combine active learning and autonomous experimentation with microfluidic technology. Scientists can already build miniaturized laboratories on a chip using microfluidics[8] controlled and directed by a computer. Such chips contain miniature reaction chambers, ducts, gates, ionic pumps and reagent stores, and allow for chemical synthesis and testing at high speed. We can imagine miniaturizing our robot-scientist technology in this way, with the overall goal of reducing the experimental cycle time from hours to milliseconds. With microfluidic technology, each chemical reaction not only requires less time to complete, but also requires smaller quantities of input materials, with a higher expected yield. On such timescales it should become easier for scientists to reproduce new experiments and refute their hypotheses.

Today's generation of microfluidic machines is designed to carry out a specific series of chemical reactions, but further flexibility could be added to this tool kit by developing what one might call a 'chemical Turing machine'. The universal Turing machine, devised in 1936 by Alan Turing, was intended to mimic the pencil-and-paper operations of a mathematician. The chemical Turing machine would be a universal processor capable of performing a broad range of chemical operations on both the reagents available to it at the start and those chemicals it later generates. The machine would automatically prepare and test chemical compounds but it would also be programmable, thus allowing much the same flexibility as a real chemist has in the lab.

"There is a severe danger that increases in speed and volume of data generation could lead to decreases in comprehensibility."

One can think of a chemical Turing machine as an automaton connected to a conveyor belt containing a series of flasks: the automaton can move the conveyor to obtain distant flasks, and can mix and make tests on local flasks. Just as Turing's original machine later formed the theoretical basis of modern computation, so the programmability of a chemical Turing machine would allow a degree of flexibility far beyond the present robot-scientist experiments, including complex iterative behaviour. In the same way that modern-day Turing machines (computers) are constructed from integrated circuitry, thereby combining the power of many components, a universal robot scientist would be constructed from a mixture of microfluidic machines and integrated circuitry controllers.

Human touch

This microfluidic Turing machine is not only a good candidate for the next-generation robot scientist, it may also make a good model for simulating cellular metabolism. One can imagine an artificial cell based on a chemical Turing machine being used as an alternative to *in vivo* drug testing. The program running this machine would need to contain algorithms both for controlling the experiment and for conducting the cell simulation. It would represent a fundamental advance in the integration of computation with its environment.

Some may argue that in the context of biological experimentation, the series of chemical reactions is the computation itself. However, one can imagine taking the integration between experiment and environment even further. In particular, by connecting the input and output ducts of the microfluidic Turing machine to the chemical environment of a living cell, one could conduct experiments on cell function. Such levels of close integration between computers, scientific models and experimental materials are, however, still a decade or more away from becoming standard scientific practice.

Despite the potential benefits, there is a severe danger that increases in speed and volume of data generation in science could lead to decreases in comprehension of the results. Academic studies on the development of effective human–computer interfaces[9] emphasize the importance of cognitive compatibility in the form and quantity of information presented to human beings. This is particularly critical for technologies associated with hypothesis formation and experimentation. After all, science is an essentially human activity that requires clarity both in the statement of hypotheses and their clear and undeniable refutation through experimentation. ■

Stephen H. Muggleton is in the Department of Computing and the Centre for Integrative Systems Biology at Imperial College London SW7 2BZ, UK.

1. *Towards 2020 Science* (Microsoft, 2006) http://research.microsoft.com/towards2020science
2. Sternberg, M. J. E. & Muggleton, S. H. *QSAR Combinatorial Sci.* **22,** 527–532 (2003).
3. Tamaddoni-Nezhad, A., Kakas, A., Muggleton, S. H. & Pazos, F. in *Proc. 14th Int. Conf. Inductive Logic Programming* 305–322 (Springer, 2004).
4. Takahashi, K. *et al. Bioinformatics* **13,** 1727–1729 (2003).
5. Halpern, J. Y. *Artif. Intell.* **46,** 311–350 (1990).
6. De Raedt, L. & Kersting, K. in *Proc. 15th Int. Conf. Algorithmic Learning Theory* 19–34 (Springer, 2004).
7. King, R. D. *et al. Nature* **427,** 247–252 (2004).
8. Fletcher, P., Haswell, S., Watts, P. & Zhang, X. Lab-on-a-Chip Micro Reactors for Chemical Synthesis. *Dekker Encyclopedia of Nanoscience and Nanotechnology* (2001).
9. Jacko, J. A. & Sears, A. *The Human–Computer Interaction Handbook* (Lawrence Eribaum Associates, 2003).