Clark Glymour

The automation of discovery

Scientific revolutions are sometimes quiet. Despite a lack of public fanfare, there is mounting evidence that we are in the midst of such a revolution – premised on the automation of scientific discovery made possible by modern computers and new methods of acquiring data.

Consider, for example, the following developments:

- Using data from the 1970s, about eight years ago a team of data analysts working in Holland predicted that low-level lead exposure is more dangerous to children's cognitive development than had previously been thought – a prediction confirmed by recent reanalyses of later observations;
- Using measurements of reflected solar energy (technically, the visible-near infrared spectrum), a computer identified minerals in rocks from a California desert lake as accurately as had a team of human experts at the site who had access both to the spectra and to the actual rocks;
- In Antarctica, a robot traversing a field of ice and stones picked out the rare meteorites from among the many rocks;
- Scientists at the Swedish Institute for Space Physics realized that an instrument aboard a satellite was malfunctioning and they recalibrated it from Earth;
- An economist working for the World Food Organization found that current foreign aid practices have no impact on extreme poverty;
- Climate researchers traced the global increase in vegetation and its causes over the last twenty years;
- A team of biologists and computer scientists reported determinations of the genes in yeast whose function is regulated by any of a hundred regulator genes;
- A kidney transplant surgeon measured the behavior of rat genes that had been aboard the space shuttle;
- A biologist reported a determination of (possibly) all of the human genes in

Clark Glymour is Alumni University Professor of Philosophy at Carnegie Mellon University, and Senior Research Scientist at the Institute for Human and Machine Cognition and John Pace Eminent Scholar at the University of West Florida. He is the author or coauthor of numerous articles and books, including "The Mind's Arrows" (2001), "Computation, Causation and Discovery" (1999), and "Thinking Things Through" (1995).

cells lining the blood vessels that respond to changes in liquid flow across the cells.[1]

All of these developments – and they are simply more or less random examples I happen to know – reflect a new way of learning about the world. Thanks to innovations in computer software, in laboratory techniques, and in observational technology, scientists today can measure things on a scale inconceivable only a few years ago. New laboratory and computational methods allow evaluation of vast numbers of hypotheses in order to identify those few that have a reasonable chance of being true, and simple oversights of human judgment can be corrected by computer. The change is from the textbook scientific paradigm in which one or a very few hypotheses are entertained and tested by a very few experiments, to a framework in which algorithms take in data and use it to search over many hypotheses, as experimental procedures simultaneously establish not one but many relationships. While there are consequences even for small collections of data, the automation of scientific inquiry is chiefly driven by novel abilities to acquire, store, and access previously inconceivable amounts of data, far too much for humans to survey by hand and eye. Methodology has moved in consequence; in a growing number of fields, automated search and data selection methods have become indispensable.

This may not seem revolutionary, but it has all of the earmarks of scientific revolution that Thomas Kuhn emphasized years ago: novel results, novel kinds of theory, novel problems, intense and often irrational hostility from parts of the scientific community.[2] We can see the revolution at work by looking more closely at three of the examples I mentioned above.

Lead was long a component of paint, and the Mobil Oil Company introduced tetraethyl lead into gasoline in the 1930s. From these and other sources, low-level lead exposure became common in the United States and elsewhere. Large doses of lead and other heavy metals were known to disrupt mental faculties, but the effects of low-level exposure were unknown. Besides, low-level exposure was hard to measure: low-level lead concentrations fluctuate in blood and do not indicate how much lead the body has absorbed over time.

In the 1970s, Herbert Needleman found an ingenious way to measure cumulative lead exposure using the lead concentration in children's baby teeth. He also measured the children's IQ scores and many family and social variables that might conceivably be relevant to the children's cognitive abilities. Reviewing the data by analysis of variance, a standard statistical technique introduced early in the twentieth century, Needleman concluded that lead exposure has a small but robust effect – it lowers children's IQ scores.

Since a lot of money was at stake, criticism naturally followed, and in 1983 a scientific review panel formed by Ronald Reagan's Environmental Protection Agency asked Needleman to reanalyze the data with stepwise regression, a more modern statistical technique. The idea behind this technique is very simple even if the mathematics is not: Suppose any of several measured variables might

1 References can be found at <www.phil.cmu.edu/projects/DaedelusRefs>.

2 Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1996).

influence IQ scores. But start with the assumption that none of the variables influence IQ. Change that assumption by entertaining as a causal factor whichever variable is most highly correlated with IQ score, then keep adding causal factors by a mathematical measure that takes account of the correlation already explained by previously considered factors. Stop when additional variables don't explain anything more. (This procedure can also be run in reverse, starting with the assumption that all of the measured variables influence IQ scores, and then throwing out the least explanatory factors, one by one.) Needleman had measured about forty variables that might account for variations in his subjects' IQ scores, and stepwise regression eliminated all but six of them. Lead exposure remained among the causal factors, and using a standard method (indeed, the oldest method in statistics, originating with Legendre's essay on comets in 1808) to estimate the dependence of IQ score on lead exposure, Needleman again found a small negative effect.

Many years after the confirmation of Needleman's results had helped to eliminate lead from gasoline, two economists, Stephen Klepper and Mark Kamlet, reanalyzed Needleman's data – with a difference. Reasonably, they assumed that the measured values Needleman reported were not perfectly accurate: IQ scores did not perfectly measure cognitive ability; lead concentrations in teeth did not perfectly measure lead exposure; and so on. Each of Needleman's six remaining variables perhaps influenced cognitive ability, but the true values of those variables were not recorded in his data. The data consisted of measurements produced by the true value of each variable for each child, and also by unknown measurement errors. Klepper proved an interesting theorem that implied that for Needleman's data, with the assumptions about measurement error, the true effect of lead exposure on cognitive ability could be positive or negative or zero. The elimination of lead from gasoline, it seemed, had been based on a statistical mistake.[3] The story doesn't end here, however. But before continuing, a digression into the statistics of causality is necessary.

In the early 1980s, several statisticians developed a network representation of probability relations that formalized and generalized ideas that had been used for a long while in biology, social science, and elsewhere. According to their representation, suppose we have data for a number of variables, each of which takes a definite value in each individual object or case (the variables might be height, weight, ratio of Democrats to Republicans, whatever; the individual objects, or cases, could be people, rats, cells, state governments, whatever). Represent each variable as a node and draw arrows from some nodes to other nodes, e.g., C ← B → A. This particular diagram represents the claim that the information that the values of A and B together provide about the value of C is the same as the information that the value of B provides all by itself. And, symmetrically, the information that values of C and B provide about A is the same as the information that the value of B alone provides.

3 Herbert Needleman et al., "Deficits in Psychologic and Classroom Performance of Children with Elevated Dontine Lead Levels," *New England Journal of Medicine* 300 (1979): 389; Needleman et al., "Lead and IQ Scores: A Reanalysis," *Science* 227 (1985): 701 – 704; Stephen Klepper et al., "Regressor Diagnostics for the Errors-in-Variables Model – An Application to the Health Effects of Pollution," *Journal of Environmental Economics and Management* 24 (1993): 190 – 211.

In other words, you can use the diagram described above when, for predicting C, if you know the value of B, then the value of A doesn't tell you anything more about the probabilities of the values of C. In more technical terms, C is independent of A *conditional* on B. (C would also be independent of A conditional on B if the structure were C → B → A or C ← B ← A, but not if it were C → B ← A or B ← C → A, etc.) The general version of this connection between networks and probabilities, known as the Markov condition, was introduced explicitly by statisticians around 1980, though it was used implicitly in many subjects long before that time, and almost formalized by the philosopher Hans Reichenbach in the 1950s. Without clearly formulating the general idea, biologists, psychologists, sociologists, and even biblical historians had used such diagrams to represent causal hypotheses and the probability relationships of their variables.[4] In the 1980s a group at UCLA, led by Judea Pearl, developed a fast algorithm for computing any conditional independence relations implied by the Markov condition when applied to such a diagram, now called a Bayes net.

In the early 1990s a group of philosophers and statisticians at Carnegie Mellon noted that many of the information restrictions, or conditional independence facts, represented in a network also hold in a related way if the arrows represent causal relations, and, relying on the Markov condition, they gave a general characterization of the relation between network structure, probabilities, and causal claims.

The idea is easiest to see for interruptions of a simple causal chain. For instance, if pushing the doorbell button causes the bell to ring, which in turn causes the house parrot to say "hello," then if you intervene to keep the bell from ringing, pushing or not pushing the doorbell button will not change the probability that the parrot says "hello." After your intervention, the state of the button and the state of the parrot will be independent of each other; neither will provide information about the other. But if you do not intervene to disconnect the bell, pushing the button will be independent of the parrot's speech conditional on the state of the bell, ringing or not ringing; if you know whether the bell is ringing, the parrot's speech won't give you any more information as to whether someone is at the door. In many cases, the independence relations produced by interventions in a system parallel the conditional independence relations implied by the network representation of the causal structure of the system.

These connections between causation, probability, and network representations suggested that with appropriate assumptions and background knowledge, something about the causal structure can be learned from observation, and the outcomes of some ideal interventions can be predicted. If C and A are independent conditional on B, and no other independence relation holds, then C and A are causally connected only through B, which functions either as a common cause or an intermediary. Inferences like this readily combine with other

4 See William Farmer, *The Synoptic Problem* (Macon, Ga.: Mercer University Press, 1981) and, for a more recent discussion, Donald Akenson, *Surpassing Wonder: The Invention of the Bible and the Talmuds* (Montreal: McGill-Queen's University Press, 1998). The statistical work is reported and illustrated in Harry Kiiveri and Terry Speed, "Structural Analysis of Multivariate Data: A Review," in Samuel Leinhardt, ed., *Sociological Methodology* (San Francisco: Jossey-Bass, 1982) and in Judea Pearl, *Probabilistic Reasoning in Intelligent Systems* (San Mateo, Calif.: Morgan Kaufmann Publishers, 1988).

information one might have – for example, if the same probability relations hold and B occurs before A and C, then the causal structure should be A ← B → C.[5] The old shibboleth that correlation does not imply causation is true for any pair of variables considered in isolation, but, when combined with otherwise routine assumptions, is not necessarily true for sets of correlations among several variables. The problem is to say in a mathematically precise and useful way just what causal information can be extracted from such dependencies, and under what assumptions.

The class of alternative networks that might conceivably describe the causal relations among a set of variables, before data is collected, is astronomical even for small numbers of variables, and with larger numbers of variables remains huge even if some of the variables are ordered so that one can assume that later variables do not influence earlier ones.

Even so, early in the 1990s, researchers at the University of Pittsburgh, Carnegie Mellon, UCLA, and Microsoft developed algorithms and software for searching for the class of diagrams that can account for any set of independence relations among variables. Since then many related algorithms have been proposed and applied by others. These procedures search efficiently for information within the huge space of alternative possible causal structures, but, unlike stepwise regression, some of these procedures come with a weak guarantee of reliability. For example, as the size of the sample increases, according to the Markov condition and one other further technical assumption, the Bayes net search programs 'converge' to giving correct information about the causal structure behind the data.[6]

Back to lead. In collaboration with Dutch statisticians, Richard Scheines, one of the Carnegie Mellon researchers, applied a program implementing these new search techniques to Needleman's data.

What the program found was simple but astonishing: three of the six prediction variables that had remained after Needleman's stepwise regression had *no* correlation with IQ scores – a fact that had somehow eluded Needleman, his collaborators, his critics, and, indeed, the stepwise regression procedure alike. Of the initial variables possibly correlated with IQ that Needleman had first considered, only lead and two other factors now remained. But, with the economists' assumptions about measurement error, the effect of lead exposure on IQ still could not be estimated.

To estimate the effect of lead, Scheines and his Dutch collaborators resorted to a relatively new technique in Bayesian statistics. Bayesian statistics proceeds by assigning 'prior probabilities' to alternative hypotheses, by computing for each hypothesis the probability of the data on the assumption that that hypothesis is true, and, from all this, computing a new, or 'posterior,' probability for each hypothesis or range of parameters considered. For a long time, because the posterior probabilities often could not be computed, Bayesian statistics was chiefly a toy used only for simple problems; computational developments in the last two decades have changed that

5 I oversimplify. For the general theory, caveats, and mathematical details, see Peter Spirtes et al., *Causation, Prediction and Search* (New York: Springer-Verlag, 1993; 2d ed., Cambridge, Mass.: MIT Press, 2000).

6 Ibid.; Clark Glymour and Gregory F. Cooper, eds., *Computation, Causation and Discovery* (Menlo Park, Calif.: AAAI Press; Cambridge, Mass.: MIT Press, 1999); Judea Pearl, *Causality* (New York: Cambridge University Press, 2001).

considerably. Scheines used the economists' judgments of the probability distribution for values of parameters related to measurement error to assign prior probabilities to their measurement error model. Then he and his collaborators computed the posterior probability distribution for values of the parameter representing the influence of lead on IQ. By this method, they found that low-level lead exposure is almost certainly at least two times more damaging to cognitive ability than Needleman had estimated.[7]

Genetics is another field in which scientists are conducting research in new ways by applying innovations in computer software, lab techniques, and observational technology.

Every cell in your body has the same DNA but cells in different tissues look and function very differently – brains, after all, are not bones. The difference comes from the proteins that make up the physical structure of a cell and regulate – indeed, in some sense constitute – its metabolism. The thousands of different kinds of proteins are themselves produced by a collaborative manufacturing process in the cell. Amino acids – any of twenty simple molecules provided to the cell from outside – are stitched together to form a protein, which may then fold and combine chemically or physically with other proteins. Each basic protein originates along a template of ribonucleic acid (RNA) outside the nucleus, and different template molecules – different kinds of RNA molecules – make different proteins. Messenger RNA (mRNA), itself copied from DNA, generates the template RNA. Whether a piece of DNA is copied into mRNA within any interval of time depends on several things, including the chemical sequence of the particular DNA piece (whether it is a coding sequence, i.e., a gene), the chemical sequences of other regions of the chromosome that are physically close (regulator sites), concentrations of small molecules inside the nucleus of the cell, and concentrations of proteins. Certain proteins attach to the regulator sites of a gene and cause the gene to be copied (in other terminology, 'transcribed' or 'expressed') into RNA, which in turn goes on to make proteins. An important clue to fundamental biology and its medical applications lies in this process of gene expression, in knowing which genes respond to new chemical or physical environments, and which cellular functions are influenced by the proteins those responding genes produce.

Traditionally, this kind of problem had been approached one gene at a time – for instance by finding some of the proteins that regulate a gene, finding the protein or proteins the gene yields, identifying some of the roles those proteins play in cellular metabolism. But about ten years ago, biologists developed techniques for simultaneously measuring the concentrations of each of the thousands – and in some contexts essentially all – of the distinct kinds of mRNA molecules present in a collection of cells. Biologists could get a snapshot of how much each gene in the cells had been copied or expressed. Multiple snapshots, moreover, could be taken at different times, as little as a few minutes apart, so that researchers could see the varying responses of the cell genome to changing conditions. So what affects what genes? Answers to this question are coming in at an astonishing rate.

About five years ago, Tim Hammond, a physician and research scientist at Tu-

7 Richard Scheines, "Estimating Latent Causal Influences," in Glymour and Cooper, eds., *Computation, Causation and Discovery.* Scheines's software – freeware – was the TETRAD III program, <http://www.phil.cmu.edu/projects/tetrad>.

lane, flew samples of kidney tissue in the space shuttle. When his samples, which had been chemically frozen while in microgravity, returned to Earth, Hammond and his collaborators measured the expression of thousands of genes within the tissue. They found that a large proportion expressed very differently from the genes within the Earth-bound samples of the same tissue, no matter how the Earth-bound tissue had been mechanically treated. Acceleration or low-gravity or something else as yet unknown about the shuttle environment affected gene behavior. If, as seems likeliest, the effect Hammond discovered is an essentially mechanical effect of low gravity, it has important implications for long-term habitation in space, on the Moon and Mars.

Mechanical issues – flow and sheer over cellular surfaces – are known to influence genes that are important to human health. The cells that line the surfaces of blood vessels play crucial roles in lethal disorders – for example, in aneurisms – and particular genes in these cells have been known for some while to change their expression in response to mechanical changes, in particular to changes in liquid flow across their surfaces. Very recently, David Peters, a young biologist at the University of Pittsburgh, and his colleagues measured the change in gene expression in response to changes of flow for almost all genes in living human cells lining blood vessels. In their experiment, more than a hundred genes changed, including some known to be involved in cellular structure. Peters and his collaborators are now measuring all of the genes in such cells that respond to changes in pressure and flow.

The few cases I have briefly described here are merely samples of a trend that can be seen in several sciences – a trend to which we can also attribute the Virtual Observatory that is planned to enable astronomers to search and analyze vast data stores taken by remote instruments; and, in climate studies, the Earth observation satellites that now send down several gigabytes of data each day– data that is increasingly being used to monitor the state of the planet, to locate causes of change, and to forecast changes in the environment. Ever new techniques make possible the measurement of ever larger quantities of data; data manipulation software makes possible the selection of samples that are relevant to particular problems; automated search and statistical techniques help guide researchers through the superastronomical array of possible hypotheses.

Kuhn said that scientific revolutions generally meet fierce resistance – and the automation of discovery in science is no exception. In some cases the animosity stems from nothing more than conservatism, an effort to preserve academic turf, or plain old snobbery. Above all, automated science competes with a grand craft tradition that assumes that science progresses only by scientists advancing a single hypothesis, or a small set of alternative hypotheses, and then devising a variety of experiments to test it. This tradition, most famously articulated by Sir Karl Popper, is championed by many historians and philosophers of science, and resonates with the accounts of science that many senior scientists learned in graduate school.

While the history of science can serve as an argument for norms of practice, for several reasons it is not a very good argument. The historical success of researchers working without computers, search algorithms, and modern measurement techniques has no rational bearing at all

on whether such methods are optimal, or even feasible, for researchers working today. It certainly says nothing about the rationality of alternative methods of inquiry. Neither *was* nor *is* implies *ought*.

The 'Popperian' method of trial and error dominated science from the sixteenth through the twentieth century not because the method was ideal, but because of human limitations, including limitations in our ability to compute. Historically, novel methods and stratagems were devised from time to time to get round computational limitations. For example, in the eighteenth century, Leonard Euler, perhaps the most prolific mathematician ever, could not reconcile seventy-five observations because the calculations required far too many steps; statistical estimation of theoretical parameters, introduced by Legendre in 1808 in a form known as 'least squares,' permitted the reconciliation of (for the time) large quantities of data, such as the seventy-five that defeated Euler. The quick adoption of factor analysis in the 1940s was due in part to computational tractability, and one could argue that the same is true of the enormous influence of Sir Ronald Fisher's statistical methods.

When scientists seek to learn new, interesting truths, to find important patterns hiding in vast arrays of data, they are often trying to do something like searching for a needle in a really huge haystack of falsehoods, for a correct network among many possible networks, for a robust pattern among many apparent but unreal patterns.

So how does one find a needle in a haystack?

1. Pick something out of the haystack. Subject it to a severe test, e.g., see if it has a hole in one end. If so, conjecture it's a needle; otherwise, pick something else out of the haystack and try again. Continue until you find the needle or until civilization comes to an end.
2. Pick something you like out of the haystack. Subject it to a test. If it doesn't pass the test, find a weaker test (e.g., is the thing long and narrow?) that it can pass.
3. Try 1 for a while, and if no needle turns up, forget about needles and start studying hay.
4. Try 1 for a while, and if no needle turns up, change the meaning of needle so that a lot of 'needles' turn up in the haystack.
5. Set the haystack on fire and blow away the ashes to find the needle.
6. Run a magnet through the haystack.

Method 1 is still the standard description of how science is and should be conducted – the account we find explicitly in the introductory chapters of science textbooks and implicitly in the criticisms some scientists and methodologists express toward other ways of doing things.

Method 2 is practiced and effectively advocated by many social scientists (you need only replace 'something you like' in 2 with 'theory').

Methods 3 and 4 are the practices that postmodernists claim science does and should follow.

Methods 5 and 6 are those made possible by the automation of discovery.

In principle, methods 5 and 6 are a lot smarter than the other methods, but they are not without limitations both real and metaphorical. Burn the whole haystack and you might melt the needle. And that is a sound worry about automating science: it may rush things, sometimes too much. Because a procedure for finding hypotheses is fast and

can be done by computer doesn't mean the procedure gives good results. Figuring out what a method can and cannot reliably do requires hard work.

Consider for example the problem of identifying networks of gene regulation. The ability to measure gene expression simultaneously for thousands of genes in normal and perturbed genomes (in perturbed genomes, particular genes have either been deleted or forced to over-express) invited the application of computer methods that search for causal networks. Algorithms were proposed for piecing together networks from comparisons of gene expression measurements in cell lines with perturbed and unperturbed genomes; algorithms were proposed for finding networks from correlations with repeated measurements of expression levels in unperturbed networks – and they did very well on data produced by computer simulations of gene expression.

It turns out, however, that much of this work proved to be illusory. The algorithm for assembling a network from perturbation effects was incorrect. The algorithms for inferring networks from correlations of gene expressions overlooked the fact that measuring expression levels in aggregates of cells (rather than in individual cells, which is technically feasible but rarely done) creates correlations due entirely to the aggregation itself rather than to the influence of particular genes on the expression levels of others. The simulations that seemed to work so well also turned out to be simulations of measurements at the level of individual cells – measurements of a kind usually not made in reality. Undoubtedly the automated procedures got some things right, but very likely what they got right was cherry picking – gene connections indicated by very large changes in expression levels or very large correlations.[8] A real advance in unraveling gene regulation networks came recently – by chemical rather than by computer automation. Tong Ihn Lee and his colleagues found a way to identify a large fraction of the genes in yeast that are, in turn, directly regulated by genes known to be regulators. They did so for more than a hundred regulator genes, effectively identifying a good piece of the regulatory structure in 'wild type' yeast.

The automation of learning, whether by computer or by new laboratory techniques, does not render human judgment obsolete, or marginalize scientific creativity. Nor does it cheapen the sweat and effort, the insight and ingenuity of human scientists, but shifts them toward the consideration of algorithms that can efficiently and reliably compare many hypotheses with vast quantities of data and toward laboratory methods that answer many questions at once.

8 Tiajaio Chu et al., "A Note on a Statistical Problem for Inferring Gene Regulation from Micro-array Data," *Bioinformatics* (in press); David Danks et al., "Experimental Determination of Gene Regulation Networks: Complexity and Statistical Realism" and Frank Wimberly et al., "Experiments on the Accuracy of Machine Learning Algorithms for Discovering Gene Regulation," in the *Proceedings of the Workshop on Bayes Nets and Gene Regulation,* International Joint Conference on Artificial Intelligence, 2003.