

Discovery Clark Glymour, Alumni University Professor, Carnegie Mellon University

Much of scientific inquiry is directed at finding causal relations and quantifying them. Historically, the scientific context at a time in a discipline suggests particular causal questions, and an experiment, or related sequence of experiments is designed to answer them. That traditional process has been altered by the development in recent decades of very large data sets in many subjects—climatology, genomics, and neuroscience are some notable examples—and the development of automated or semi-automated methods for searching such data for patterns that indicate more than accidental correlations. These search methods need development—for example, methods are needed that can reliably extract causal cascades and their feedbacks from brain imaging data. Dissemination is equally important. The machine learning/statistical literature has in the last twenty years developed a host of principled methods that are superior in almost every respect to the regression and a priori model fitting methods that dominate causal inference in the social sciences, and available software has made these new procedures easy to use. But the news has not sunk into the social and behavioral sciences, where simple regression, factor analysis, and fitting of a handful of a priori models remain commonplace.

The traditional scientific process itself has generated some relatively novel problems. For example, memory research has produced thousands of studies investigating behavioral correlates or effects of various cellular mechanisms. Computerized methods are sorely wanted for synthesizing such results in a way that unifies the causal relations where they can be unified, and that also helps to identify the gaps and the conflicts. That kind of project is hampered by the absence of efficient, unified databases containing experimental procedures and raw data. To some degree, the same could be said of imaging data, which is collected in some databases but chiefly in incomplete and truncated form.

So three recommendations: (1) we are in an age where computerized search could dramatically speed up, and improve, scientific discovery if there were focused funding for the development of unified, well-curated databases and legal requirements on deposition enforced by funding agencies. (2) Principled methodologies for searching databases for potential causal relations of interest, and for unifying extant fragmentary results need to be encouraged and funded. (3) Methodology in the quantitative social and behavioral sciences needs to be brought up-to-date. The last is a social problem rather than simply a funding problem, but funding initiatives can of course influence the scientific community.

Some relevant sources

P. Spirtes, et al., *Causation, Prediction and Search*, Springer Lecture Notes in Statistics, 1993; 2nd edition MIT Press, 2000.

J. Ramsey et al., Six Problems for Causal Inference from fMRI, *NeuroImage*, 2010 Jan 15;49(2):1545-58.

C. Glymour, The Automation of Discovery, *Daedalus*, Winter (2004), 69-77.

B. Bontempi, et al. *Memories, Molecules and Circuits, Research and Perspectives in Neuroscience*, Y. Christen, Paris, 2007.

