Computer Modeling in Philosophy

Joseph D. Ramsey, Kun Zhang, Clark Glymour*

The Evaluation of Discovery: Models, Simulation and Search through "Big Data"

https://doi.org/10.1515/opphil-2019-0005
Received October 23, 2018; accepted November 30, 2018

Abstract: A central theme in western philosophy was to find formal methods that can reliably discover empirical relationships and their explanations from data assembled from experience. As a philosophical project, that ambition was abandoned in the 20th century and generally dismissed as impossible. It was replaced in philosophy by neo-Kantian efforts at reconstruction and justification, and in professional statistics by the more limited ambition to estimate a small number of parameters in pre-specified hypotheses. The influx of "big data" from climate science, neuropsychology, biology, astronomy and elsewhere implicitly called for a revival of the grander philosophical ambition. Search algorithms are meeting that call, but they pose a problem: how are their accuracies to be assessed in domains where experimentation is limited or impossible? Increasingly, the answer is through simulation of data from models of the kind of process in the domain. In some cases, these innovations require rethinking how the accuracy and informativeness of inference methods can be assessed. Focusing on causal inference, we give an example from neuroscience, but to show that the model/simulation strategy is not confined to causal inference, we also consider two classification problems from astrophysics: identifying exoplanets and identifying dark matter concentrations.

Keywords: Big Data; computerized search; discovery; simulation

"*There are, however, no precisely describable methods of discovery.*" –*John Rawls*

1 The demise of method in philosophy

The great philosophical discovery of the 20th century was that a grand aim of western philosophy is a phantasm, a will-o-the-wisp, a red herring, as likely to succeed as squaring the circle. That misplaced ambition was to find a recipe for discovering the processes of nature, especially relations of cause and effect, a procedure for distinguishing true relations from happenstance and irrelevance. Aristotle's *Posterior Analytics*, Descartes' *Principles*, Bacon's *Novum Organum*, Newton's *Principia* and *Optiks*, and Leibniz' "Universal Calculus" all essayed methods of discovery, and efforts continued with Mill and Boole. Peirce introduced randomization in experimental design.[1] Discovery was a central project of philosophy. That ambition came in the 20th century to be called finding a "logic of discovery."

John Rawls, of all people, succinctly announced the death of the project: "There are...no precisely describable methods of discovery."[2] It is unclear just what Rawls, or anyone, meant by "precisely describable

1 Pierce and Jastrow, "On Small Differences in Sensation."
2 Rawls, "Outline of a Decision Theory for Ethics."

***Corresponding author: Clark Glymour,** Carnegie Mellon University, United States of America; E-mail:cg09@andrew.cmu.edu
Joseph D. Ramsey, Carnegie Mellon University, United States of America
Kun Zhang, Carnegie Mellon University, United States of America

methods of discovery," but Rawls was just following the crowd. Karl Popper had titled his famous book *The Logic of Scientific Discovery* and in its first pages announced there can be no such thing. Carl Hempel announced that algorithmic science—science by computer—is impossible. Further, he claimed, the notion of truth is useless in science and so there is no point seeking for methods that can find it.[3] Scientists should seek explanations, not truths. Hans Reichenbach[4] insisted on a distinction between the "context of discovery" and the "context of justification" in science. Whatever Reichenbach meant by the distinction, it was widely accepted as the principle that warranting a particular hypothesis can be algorithmic, but generating hypotheses likely to be true cannot be. The canon of philosophy of science was revised: Newton, the Natural Philosopher, no longer counted as a philosopher, he became instead the *subject* of philosophy. Except for his algebra, Boole was quietly shuffled out of the philosophical listings. For Kant and the neo-Kantians—Russell, Popper, Hempel, Reichenbach—and those they taught, philosophy's connection with the sciences was to provide reconstruction and justification, not guides to inquiry.[5]

There were and continue to be meta-methodological discussions in philosophy of science, but they did not engage Rawls' dictum. Late in the 20th century there were challenges to Rawls' opinion from Paul Thagard,[6] Kevin Korb[7] and others, but Rawls' view still reigns in philosophy, sufficiently that with some justice physicists[8] have claimed philosophy of science has nothing to offer the enterprise of science.

2 Unpacking

It is not so clear just what Rawls and his contemporaries were denying. Absent the probabilities, and remembering that they were addressing human psychology rather than logical possibility, both Locke and Descartes seem to have envisioned some conception of rules for turning sensation into abstract generalizations. Bacon gave a general description of criteria establishing causes from data, but no algorithm for searching through data to find the relationships that would satisfy his criteria. Leibniz, as did Boole much later, seems to have imagined a human procedure of discovery aided by algebraic rules. Rudolf Carnap[9] imagined an "inductive robot" collecting data and computing posterior probabilities as it saunters through the world. The robot begins with prior probabilities over a *tabula rasa* that lists all properties and all possible relations among them but nothing of their empirical instantiations. Were it possible, would a "logic of discovery" take singular premises and generate general conclusions certain to be true if the premises were? No modern since Hume thought that possible. Would a "precise method of discovery" instead be only an algorithmic procedure guaranteed to produce plausible or probable explanations? Was a single method envisioned that would be applicable to every domain, or as Rawls' plural "methods" suggests, possibly distinct methods for different branches of inquiry?

The very possibility of an algorithm or algorithms for turning data into generalizations was not really the cynosure of the doubts. It is easy enough to contrive some algorithm for turning data into generalizations.

3 Jeffrey, *Carl G. Hempel*.

4 Reichenbach, *Experience and Prediction*.

5 The question of why in the late 19th century and after Anglophone philosophy turned away from methodological contributions to other topics deserves some serious but so far unreceived intellectual and social history. We can only conjecture an outline. At the turn of the century, Britainic and Continental philosophers with mathematical and scientific training and interests were educated in versions of Kantian doctrine and in its presupposition that the philosophical task is to justify science, not to contribute to it. Russell, Reichenbach and Carnap all wrote doctoral theses in this neo-Kantian vein. They were soon captivated by Frege's example in attempting to formalize the notion of proof, which they took as a model for philosophy. A logic of discovery would turn singular data into a proof of general propositions, and that was plainly not possible. The replacement of the apparently certain Newtonian theory of kinematics with special relativity sealed the case. The emigration of continental philosophers to Anglophone countries brought with it their conception of the limits of philosophy. Placed in influential institutions, they taught generations of students. In twain, by the 1980s statistics and computer science had become more interesting and more rewarding for those with mathematical abilities and methodological concerns.

6 Thagard, *Computational Philosophy of Science*.

7 Korb, "Introduction: Machine Learning as Philosophy of Science."

8 Weinberg, *Dreams of a Final Theory*.

9 Carnap, "The Aim of Inductive Logic."

What was doubted was that there could be such procedures whose capacity to generate true generalizations could be warranted. Not perhaps warranted against Hume's radical skepticism, but at least warranted in something like the un-philosophical way that we take experiments to warrant causal conclusions, or the observations of the planets to warrant Kepler's laws. Fundamentally, what was deemed impossible is the creation of methods that are automatic and trustworthy but fallible guides to finding true hypotheses.

Ronald Fisher's[10] creation of modern frequentist statistics was a limited method of discovery requiring some external source of hypotheses and data. Given a hypothesis and sample data for the variables the hypothesis relates, Fisher provided algorithmic methods for assessing the hypothesis. The methods carried probability guarantees for truth guidance under specifiable assumptions. The guarantees were of two kinds, asymptotic and finite. The asymptotic guarantees were that were data to be acquired to infinity from the same system or kind of system with the same likelihood for a sample unit, and the statistical method applied with each new datum, the conclusions of the method (e.g., parameter estimates or decisions to reject a hypothesis) would converge (in probability) to the truth. In statistical parlance, the methods are *consistent*. There is an abundance of convergence criteria, but only "uniform convergence" guarantees the existence of calculable error probabilities for inferences from finite samples. Statistics developed lots of uniformly convergent estimators and tests with accompanying error measures for correlations, partial correlations and a multitude of other quantities.

Although 20th century philosophers of science seem not to have closely followed developments in statistics, they surely had a general familiarity. Nonetheless, their skepticism about the possibility of a "logic of discovery" was not assuaged. What was missing was an algorithm, or algorithms, for *generating* the hypotheses to be tested so that, for the most part, the tested hypotheses are true. *That* was what Rawls announced cannot be, and statistics provided no refutation.

3 The emergence of search methods

Algorithmic methods for searching for relationships that indicate causes or allow the identification and classification of phenomena were proposed early in the 20th century, notably in factor analysis. They lacked any demonstrations that they are truth guiding even in the infinite limit. By the latter part of the century, a variety of machine learning methods for identification or classification had appeared, including support vector machines, neural networks of many kinds, and more recently, "deep learning"—essentially multi-layered neural networks. Some of them carried asymptotic correctness and finite sample error guarantees under specified assumptions. At the end of the century, philosophers working outside of conventional statistics developed algorithmic methods for searching through data for causal relations represented as directed graphs, with proofs of their correctness in the large sample limit. These methods, however, are without finite sample error probabilities. Almost unnoticed, some of these developments refuted philosophical announcements of impossibility. Shadowing the spirit of Hegel's claim that more than six planets is impossible, Hempel had claimed that algorithmic science is impossible. His argument was that scientific theories contain quantities that are not directly measured—"theoretical quantities." Obviously, he claimed, no computer could find such quantities and their roles. To the contrary, Peter Spirtes' Fast Causal Inference (FCI) algorithm[11] showed the possibility of identifying the presence of unobserved causes. Related developments showed that in some circumstances (chiefly linearity, homogeneity of the sample, and appropriate sampling) information could be obtained even about the causal relations among unobserved variables inferred from data. These efforts provided methods for searching over small numbers of variables for true hypotheses. That idea spread in the ensuing decades, with hundreds of proposed algorithms[12],

10 Fisher, *The Design of Experiments*; and *Statistical Methods for Research Workers*.

11 Spirtes, et al., *Causation, Prediction and Search*.

12 The literature on causal search is scattered over many journal articles and arXiv submissions. There is no up-to-date text or adequate survey. The interested reader may consult Spirtes, *op. cit.*, 2nd edition, which is now almost 20 years behind developments, the proceedings of the conference on Uncertainty in Artificial Intelligence, the websites of Peter Spirtes https://www.cmu.edu/dietrich/philosophy/people/faculty/spirtes.html and Kun Zhang https://www.cmu.edu/dietrich/philosophy/people/faculty/zhang.html, and their Google Scholar entries, and the Center for Causal Discovery https://www.ccd.pitt.edu/

some of them with proofs of convergence to the truth under specifiable assumptions that did not restrict *a priori* the topic of theories or their causal relations. Various difficulties, such as the need for assuming all data analyzed are from a common causal structure, the presence of measurement error, and the use of multiple data sets, fell away or are falling.

Despite the plethora of search algorithms that have emerged in the 21st century, and despite that many of them have asymptotic consistency proofs, only one of them to our knowledge has a uniform consistency proof and hence allows estimates of errors in inferences from finite samples.[13] But the algorithm is so computationally and statistically demanding that it can only be used for a handful of variables.

4 The Big Data problem

"Big Data" are nothing new. Data collection has been going on in astronomy, the original science, almost since the beginning of recorded time. Kepler had a multitude of measurements of the positions of the planets, and worked through them to the ellipse. Darwin had a host of diverse data that he used to argue in a loosely Newtonian way for a grand theory of the common mechanism of the aetiologies of living forms. Perhaps the first successful statistical use of big data was Gilbert Walker's identification early in the 20th century of the causes of the periodic failure of monsoons in the Asian subcontinent. Thanks to the rule of Britannia, Walker, an official in the Raj, was able to gather contemporary and historical weather reports from around the world. With a myriad of correlation analyses among variables he formed from the raw data, he settled on what we now call the "El Nino" phenomenon as the cause of monsoon failure.[14]

The appearance in the 1990s of new methods to acquire data sets with enormous numbers of variables in astronomy, climate science, biology, neuroscience, and elsewhere, made obvious the inadequacy of the statistical mill. Measurements were acquired with sample sizes that were in many cases orders of magnitude smaller than the thousands upon thousands of measured variables. These developments fundamentally changed the structure of inference problems, especially in subjects that lack "ground truth" from experiments controlling the variables of interest. Big data posed the challenge of finding relations among a host of variables, often with no potential effect singled out a priori. With climate data the questions went beyond Walker's '*what are the teleconnections that account for the failure of monsoons* to *what are the teleconnections in the Earth's climate*?' In neuroscience, imaging data provoked questions as to the signaling relations among thousands of small regions of the brain. In biology, the problems included finding the regulatory relations among thousands of genes and the influences of proteins on one another. No particular effect is specified in such questions; the task is to find the mechanisms among a horde of variables. That was the grist, and methods were needed to sift the true from the false among the almost uncountable numbers of hypotheses that could be formed from such variables.

20th century statistics and philosophy had agreed that such problems were unsolvable. There were several reasons. One was sheer computational complexity. With even just a thousand variables, the number of possible causal structures among them is[15]

$$9.34 \times 10^{300728}$$

How many of those could be put through the statistical mill of any data set, one at a time? Statistics provided methods that could separate wheat from chaff one grain, or a very few, at a time, but no efficient way to search through the grist to produce hypotheses worth testing, and no methods for assessing the body of results of such a search. Prior probabilities might somehow be put over such large dimensions, but it was (and remains) computationally infeasible to update on problems of that size. Even if some fraction of the possible causal models for such data sets could somehow be tested statistically, no error

13 Spirtes and Zhang, "A Uniformly Consistent Estimator of Causal Effects."

14 Walker, "Correlation in seasonal variations in climate (Introduction)."

15 This is the number of directed graphs on a thousand vertices, allowing direct cycles between vertices.

probabilities are possible for the whole of the results, because almost certainly any model resulting from a search that specified hundreds or thousands of dependencies and independencies would be false somewhere. Statistical assessments of big data meet both complexity and the Preface Paradox: an author writes in the preface of a monograph that he believes all that is written in it, but also believes there are some errors.

Whether causal information can be extracted for an entire body of data depends on the dimensionality of the data (measured in number of variables) and the complexity of the data generating system (measured in terms of the density of connections in the unknown, true causal graph), on the sample size, and on the speed of the algorithms. Most of the algorithmic methods that had been developed outside of conventional statistics for searching for causal relations from observed data were slow. They could feasibly address problems with only a few variables, and the time they required increased in the worst case exponentially with the number of variables and density of their connections. Developments in the 21st century have nonetheless produced faster algorithms, and improvements and parallelization have speeded up several older algorithms so that they can address data collections with thousands of variables and more. For example, Ramsey, et al.,[16] recovered most of a sparse graph of a million variables from a sample of 1000 measurements of each variable. But these algorithms are without finite sample error probabilities. Search was almost ready for big data. The essential problem was, and is, how to assess the accuracies of various search methods.

5 Assessing accuracies: the simulation solution

Plato, Sextus Empiricus, and, tracing them, Hume, emphasized the logical point that any finite sample has an infinity of logically consistent, distinct continuations. Statistics had met these arguments by moving from logical necessity to probability, whether in the Bayesian way or the frequentist. In the Bayesian way, assume an uninformative prior probability distribution and update as data are acquired. In the frequentist way, apply the appropriate statistical estimator or hypothesis test under assumptions that could themselves be tested. One could not have Cartesian certainty, but at least the probabilities of errors of inference could be known under assumptions that seemed minimal and testable and did not impose *a priori* restrictions on the substance of science, the causal claims at issue. But while these techniques could be used as steps in search algorithms, they do not apply to the entire results of searches through high dimensional data.

5.1 The simulation solution

In many cases, search requires a change in perspective: discovery becomes statistical distillation, and the assessment of a search method becomes its success in distilling true hypotheses from a huge receptacle of possibilities. Rather than assessing the truth of the entire output of a search algorithm that might return thousands of claims, one would like to assess the output by "precision," the probability that returned hypotheses are true, and "recall, "the probability that the true hypotheses are returned. Recall is of course a function of parameters of the data generating system. For example, within a linear system recall of causal relations will be a function of the absolute values of covariances between variables. In search for exoplanents, recall is a function of the luminosity of a star and the radius of its planet. Discovery becomes statistical chemistry: a mixed sample is distilled to a smaller sample with a much higher proportion of the desired component—in this case, truth. Chemists sometimes use multiple stages of distillation, and so sometimes do search algorithms with the use of resampling strategies that split a sample many times, and rerun a search over and over to find the most robust hypotheses. In statistical distillation, causal search algorithms were biased for sparsity of parameters (e.g., direct causal connections) by imposing a cost on introducing a parameter. In various ways, improved fit to the data provided by an extra parameter

16 Ramsey, et al. "A Million Variables and More."

was traded against a cost for that addition. There were losses in sample recovery, and, as with separating alcohol from a mixture with water, some proportion of contaminants remained.

How can the precision and recall of a search method for causal relations be assessed? How could a method of classification be assessed when little or no empirical "ground truth" is available? It could be checked whether a search method recovers the handful of relationships that might be known, and in some domains a few of the relationships inferred from a search method might be tested individually by experiment, but that is well short of estimates of precision and recall for a vast number of claims. The solution is simulation.

If we knew nothing about a domain, a great many random relationships with random associated probability distributions could be generated, a great many random samples of many sizes could be generated from them, a search procedure could be run and its precision and recall computed. We could at most sample a tiny fraction of the possible worlds. Fortunately, sometimes there is considerable information about a domain, for example *motifs*—typical sub-structures that compose systems in a domain. Known motifs can be used to produce relevant simulated data on which the precision and recall of a search method can be tested. And sometimes enough well-established theory is available to more precisely limit the relevant simulations.

5.2 Neuroscience

Functional magnetic resonance (fMRI) measures variations in magnetic susceptibility in brain tissues. Neural signaling requires oxygen, carried in hemoglobin. The magnetic properties of the hemoglobin molecule vary accordingly as it has an oxygen molecule as cargo, or does not. So physiological activity of tissues of the brain should vary with the relative concentration of deoxygenated versus oxygenated hemoglobin. It has been established that physiological activity in brain tissues is strongly correlated with blood oxygenation level fMRI measurements.[17] We know the signaling connections in the brain are dense, especially locally; we know they have feedback relations sometimes represented by directed graphical cycles, and we know some feedback cycles reduce effects and some amplify effects. We also have good reason to think that the physiological variables we measure, for example in an fMRI scan, are approximately linear related and have non-Gaussian probability distributions.[18] And we know the sample sizes obtained in an fMRI scan. That sort of information limits the possible worlds and possible samples that we need to explore.

There are mathematical models of the generation of the data that fMRI records.[19] The models allow specification of almost arbitrary causal structures represented by directed graphs. The models can be given parameter values so that they generate simulated data that closely matches empirically measured signals in their distribution properties. The idea is to simulate such data from the graphical motifs that are the most difficult to discover—positive and negative feedback cycles; feedback cycles intersecting feedback cycles; direct and indirect feedback cycles, and so on. Samples of sizes characteristic of fMRI studies are then given to various search algorithms and their precisions and recalls for recovering the underlying neural signaling structures in the simulations are computed.

Recent work[20] illustrates the strategy. Sanchez-Romero and colleagues generated simulated fMRI data from the structures shown in Fig. 1.

17 Vazquez, et al., "Neuronal and physiological correlation to hemodynamic resting-state fluctuations in health and disease."
18 Bijterbosh and Smith, *Introduction to Resting State fMRI Functional Connectivity*.
19 Penny, et al., *Statistical Parametric Mapping*.
20 Sanchez-Romero, et al., "Estimating Feedforward and Feedback Effective Connections..."

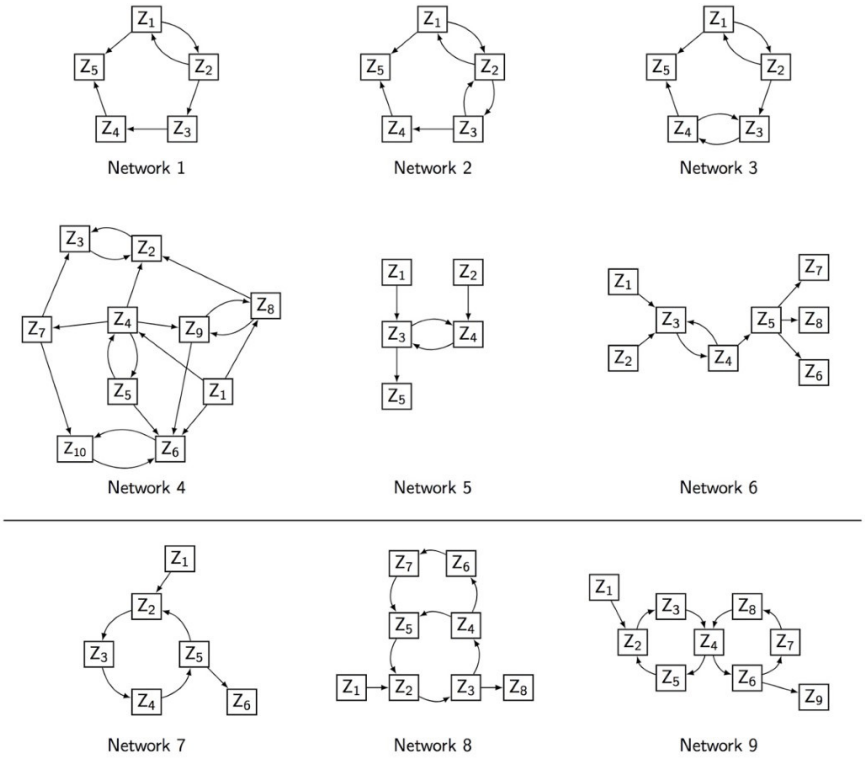

Figure 1: Causal graphical structures used for simulation in Sanchez-Romero, et al., "Estimating Feedforward and Feedback Effective Connections…"

Altogether, they use 18 different models. For each model, 60 samples of the size obtained in typical fMRI studies (500) are generated. The average precisions and recalls of multiple search algorithms when applied to these samples are then calculated. They find that two of the algorithms they analyze have average (over all models) precisions of greater than 90%, with average recalls of 60% and 80%. The methods exploit the non-Gaussian features of the measured variables. They also find the precisions of these methods are robust to measurement error. In addition, Sanchez-Romero test the two best algorithms on simulated data based on experimentally known neural connectivities in macaques involving more than 90 variables.

5.3 Discovering exoplanets

When a planet transits between a star and the line of sight from the star to the Earth or to a satellite, the measured luminosity of the star is reduced. The reduction will be repeated with the periodicity of the orbit. So far, several thousand of these exoplanets have been found by searching for stars whose light is periodically reduced. This requires searching through hundreds of thousands of time series of stellar images for periodically dimming light. Not a job to be done by hand. EXPLORE[21] is a program for doing the job by computer. The accuracy of the program is assessed by adding simulated stars to real images. Some of the simulated stars have light curves with transits simulating an occultation of a star by a planet. The job of the computer is to identify the stars with simulated planet transits. The simulated luminosities of stars, the size of the dimming (determined by the size of the planet, and the observed luminosity of the star) can all be varied. The results of applying the search algorithm to the images that combine real and simulated stars allow correction of the algorithm and recognition of the limits of the recall and precision of the search method.

21 Lee, et al., "Scientific Frontiers in Research on Extrasolar Planets."

5.4 Dark matter astronomy

The hypothesis that the universe contains invisible matter emerged from estimates that galaxy clusters have insufficient visible mass to hold them together.[22] The hypothesis has been supported and elaborated by measurements of the rotational velocities of spiral galaxies, whose speeds of rotation as a function of their distances from galactic centers is not accounted for by the estimated masses of the galaxies.[23] This suggests that there is further, unilluminated matter—dark matter—distributed through and around the galaxies that accounts for the anomalies in rotational velocities. Dark matter does not absorb or emit electromagnetic radiation and can only be detected by its gravitational effects. It does not fit anywhere in the "standard model" of elementary particles, but is estimated to constitute as much as 85% of the material universe. Of natural interest are the locations and distributions of this mysterious substance in the observable universe. The detection of dark matter exploits the gravitational influence on light, called gravitational lensing, famously associated with the measurements of the deflection of starlight observed during eclipses, and of course with the development of the theory of black holes. In strong lensing, the light from more remote object passing by a more proximate object may be split so that two images of the remote object are observed. In weak lensing, when light from a galaxy passes through or near dark matter, its apparent shape is deformed by the gravitational influence of the dark matter on the light. In particular, the image may appear sheared out of an elliptical shape. Estimates of the distribution of dark matter are obtained by statistical analyses of such shape deformations from multiple images in a region as their light is passing near or through concentrations of dark matter.

Identifying lensed images and their distortion by dark matter requires algorithmically estimating the ellipticities of many, many galactic images. A problem is to calibrate the errors in these estimates, since no empirical "ground truth" is available identifying galaxies lensed by dark matter. Ravanbakhsh, et al. remark that "In order to detect and/or calibrate any such biases, future surveys will heavily rely on image simulations, closely mimicking real observations but with a known ground truth lensing signal."[24] The "known ground truth" is of course whatever model is used in generating the simulated data. They propose a method for simulating such data by using a deep learning algorithm to extract statistical features of available galactic images. They then use statistical/physical models incorporating those features to generate simulated images. The simulated images are used to test accuracies of identifications of galaxies whose light has suffered weak gravitational lensing. As with exo-planets, the simulation studies help to estimate the recall and precision of the search method.

6 Discussion

While the neuroscience and astronomy examples are technically quite different, they share the feature that the accuracies of search methods are assessed from data generated from simulated models of the system under study. The obvious objection to such methods is that there is no guarantee that the data generating process in simulation is faithful to the empirical data generating process. All that is done is to formulate a mathematical model of the process and show that the simulated data produced from it closely match the empirical data. There are notable examples where simulation testing of search methods has made major mistakes. One case is methods for discovering gene regulation, which used simulations that failed to capture features of the empirical measurement process that critically affected accuracies of search.[25] Even granted the faithfulness of the mathematical models to the empirical data generating process and the measurement process, there is no guarantee that the simulations cover the range of variation in the natural phenomena.

Ordinary experimental testing shares these features. An experiment aimed at estimating a causal

22 Zwicky, "Die Rotverschiebung von extraglaktischen Nebeln" and "On the Masses of Nebulae and of Clusters of Nebulae."
23 Ruben and Ford, "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions."
24 Ravanbakhish, et al., "Evaluating Dark Energy Science with Deep Generative Models of Galaxy Images," 1488.
25 Chu, et al., "A Statistical Problem for inference to Regulatory Structure..." and Wimberly, et al., "Problems for Structure Learning..."

effect often assumes a mathematical model of the relationship, for example a linear model. Especially in experiments with human beings, generalizing estimates from the experimental population to new populations runs the risk that there are variations in the natural or social phenomena that are not realized in the experimental population, most obviously from variations in confounding variables that are not measured in the experiment. Addressing the problem of generalizing from human experiments requires large and diverse experimental samples and the measurement of more features of them than is customary in, say, psychological experiments. Addressing the problem of calibrating the accuracy of search algorithms for causal relations requires using a wide variety of structures and model parameters. Fortunately, in some cases, such as neural signaling connectivity, there are motifs, such as those in the Sanchez-Romero study, that are expected in the empirical domain and can be used to challenge search methods. In others, as in the astronomy examples, there is sufficient data and physical knowledge of the causal processes that generate the data to allow estimation of parameters for a simulation model. That creates a notable difference in the neuroscience and dark matter cases. In the neuroscience case, the procedure for warranting a method is not automated: the examples to be simulated and the ranges of parameters in simulations have to be humanly contrived with limited empirical guidance. In the dark matter case, the critical features of simulations can be automatically extracted from empirical data.

Looking to our philosophical ancestors, Leibniz for example, the ambition to find a single method for all problems of empirical inquiry has not been realized, and we do not expect it to be. The two best methods Sanchez-Romero identified are flexible. One of them has been applied successfully (in unpublished work) to recovering experimentally known protein interactions from observational data and to recovering climate circulations from simulated data. The other has been applied to financial data. But they are scarcely universal. They assume the variables are non-Gaussian and the relationships are linear. Rawls' pronouncement, however, was written in the plural, encompassing the possibility of different algorithms for different domains: there are, he announced, no precisely describable *methods* for discovery. Rawls was wrong. Model simulation done with sufficient care and attention to variety, supplemented with other evidence, can provide a basis for confidence in "precisely describable" methods of discovery. But procedures for evaluating search methods by simulations are themselves often not automated in selection of examples, and are without guarantees that they are representative. Developing formal, sensible measures of confidence in such circumstances is an interesting challenge, and that may be where "precision" really is impossible.

References

Bijterbosh, Janine, Stephen Smith and Christian Beckmann, *Introduction to Resting State fMRI Functional Connectivity.* Oxford: Oxford University Press, 2017.

Carnap, Rudolf. "The aim of inductive logic." In *Studies in Logic and the Foundations of Mathematics*. Amsterdam: Elsevier, 303-318, 1966.

Chu, Tianjaio, Clark Glymour, Richard Scheines, and Peter Spirtes. "A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays." *Bioinformatics*, *19*, 1147-1152, 2003.

Fisher, Ronald. A. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd, 1925.

Fisher, Ronald. A. *The design of experiments*: New York: Macmillan, 1935.

Jeffrey, Richard. *Carl G. Hempel: Selected Philosophical Essays*. New York: Cambridge University Press, 2000.

Korb, Kevin B. "Introduction: Machine learning as philosophy of science." *Minds and Machines 14*, 433-440, 2004.

Lee, B. L., H. K. C. Yee, G. Mallén-Ornelas and S. Seager. "Scientific Frontiers in Research on Extrasolar Planets," ASP *Conference Series*, Vol 294, Edited by Drake Deming and Sara Seager, 413-418, 2003.

Peirce, Charles. S., & Joseph Jastrow. "On small differences in sensation." *Memoirs of the National Academy of Sciences*, *3*, 73-83, 1884.

Penny, William. D., Karl Friston, Joseph Ashburner, Stefan Kiebel and Thomas E. Nichols (Eds.). *Statistical parametric mapping: the analysis of functional brain images*. Amsterdam: Elsevier, 2006.

Ramsey, Joseph, Madelyn Rose Glymour, Ruben Sanchez-Romero, and Clark Glymour. "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images." *International journal of data science and analytics*, *3*, 121-129, 2017.

Ravanbakhsh, Siamak, Francois Lanusse, Rachel Mandelbaum, Jeff. G. Schneider and Barnabas Poczos. "Enabling Dark Energy Science with Deep Generative Models of Galaxy Images." In *AAAI* 2017, 1488-1494, 2017.

Rawls, John. "Outline of a decision procedure for ethics." *The Philosophical Review*, *60*,177-197, 1951.

Reichenbach, Hans. *Experience and prediction: An analysis of the foundations and the structure of knowledge*. Berkeley: University of California Press, 1938.

Rubin, Vera and W. Kent Ford. "Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions." *The Astrophysical Journal*. *159*, 379, 1970.

Sanchez-Romero, Ruben, Joseph D. Ramsey, Kun Zhang, Madelyn Rose Glymour, Biwei Huang, and Clark Glymour. "Estimating Feedforward and Feedback Effective Connections from FMRI Time Series: Assessments of Statistical Methods." *Network Neuroscience,* (in press).

Spirtes, Peter, and Jiji Zhang. "A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption." *Statistical Science*, 662-678, 2014.

Sprites, Peter, Clark Glymour and Richard Scheines. *Causation, Prediction and Search*. Springer Lecture Notes in Statistics. New York: Springer, 1993.

Thagard, Paul. *Computational Philosophy of Science.* Cambridge, MA.: MIT Press, 1993.

Vazquez, Alberto, M. Murphy, and S. Kim. "Neuronal and physiological correlation to hemodynamic resting-state fluctuations in health and disease." *Brain connectivity* 4.9; 727-740, 2014.

Walker, Gilbert. "Correlation in seasonal variations in climate (Introduction)." *Memoirs of the India Meteorological Department 20(*6).

Weinberg, Stephen. *Dreams of a final theory*. New York: Vintage, 1994.

Wimberly, Frank, David Danks, Clark Glymour and Tianjaio Chu. "Problems for Structure Learning Aggregation and Computational Complexity." In *Machine Learning: Concepts, Methodologies, Tools and Applications*. Hershey, PA: IGI Global, 1699-1720, 2012.

Zwicky, Fritz. "*Die Rotverschiebung von extragalaktischen Nebeln."* Helvetica Physica Acta," *6, , 110–127, 1933.*

Zwicky, Fritz. "*On the Masses of Nebulae and of Clusters of Nebulae,"* Astrophysical Journal, *86*: *217*, *1937*.