MACHINE EPISTEMOLOGY AND BIG DATA

GREGORY WHEELER
MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY, LMU MUNICH

The Routledge Companion to The Philosophy of Social Science

§1

Here is a portrait of experimental science. A question arises, a hypothesis is proposed, experiments are devised, then performed, and a judgment is made on how well some or another implication of that hypothesis, supposing it were true, accords with the outcomes. Such was Charles Sanders Peirce's view of experimental inquiry at the end of the nineteenth century.[1] By the close of the twentieth, a spectacular store of methods were on hand to quantify the uncertainty an experimentalist confronts, along with a logic, broadly speaking, to assess its consequences. Just as the differential calculus swept to the margins ancient bewilderment over change and how to reason about quantities that change, so have the triumphs of modern statistics pushed aside Cartesian paralysis over error and how to reason with corrigible, uncertain quantities. It is not so much that Leibniz and Newton answered Parmenides, or that Peirce set the course for Pearson and Fisher to refute the academic skeptics, but rather that in each case a genuine obstacle to inquiry was plucked from confusion and paradox and a clear way to reason around those obstacles was shown. For those who wonder what philosophical progress looks like, look no further.

§2

Statistics confronts two questions. The first asks what can be inferred from data, given modeling assumptions that you choose, while the second asks after the reliability of those inferences. Data, a product of ingenuity rather than a commodity of passive experience, is therefore—historically, at least—in short supply. Further, since data is won through deliberation and action, assessing the reliability of a statistical inference comes after such choices are made, not before.

Knowledge within this portrait is an answered question, a problem resolved, and settled judgments from prior investigations are tools for new inquiries rather than a

[1] See 'Abduction and Induction' in (Peirce 1955).

normative standard against which all new conclusions are assessed. The modern experimental sciences, as Dewey remarked, are not the only kind of knowledge, but they magnify the essential ingredients for knowledge of every kind.[2]

This pragmatic view of experimental inquiry, where evidence is taken from the world rather than given, stands in contrast to a long tradition in philosophy that views the objects of knowledge as prior to, and unchanged by, the mental activity of coming to know them. Knowledge, this tradition tells us, is a stock of beliefs that hits on the truth and does so for a reason. The task for epistemology is to divine the nature of such beliefs and the principles governing such reasons. It is little surprise then that statistical methodology is viewed as a puzzle for traditional epistemology rather than a reservoir of answers. On what grounds ought you select modeling assumptions? What justifies viewing your data, so-parameterized, as a suitable guide for an unknown event? These are the problems of Duhem and Hume, respectively.

For a pragmatist, this is entirely the wrong way to proceed. Knowledge is a means of control, not a special state of mind. Uncertainty is to be exploited rather than extinguished, and epistemic notions ought to be derived from the roles that they play in inquiry rather than the other way around. The proceedings of contemporary epistemology are a record of what comes from a tradition that stands opposed to each of these tenets. It is a record as far removed from advancing knowledge as the confabulations of numerologists are to advancing higher arithmetic.

I bring up these two radically different conceptions of experimental inquiry because the rise of machine learning and big data is at once a bewildering mystery by the lights of traditional epistemology and a clear demonstration of the pragmatists' point. Hilary Putnam called for a *pragmatic enlightenment* in epistemology,[3] one that followed through on Dewey's observation that uncertainty is a practical matter. The short history of machine learning is the closest we have to a side-by-side comparison of how well the entrenched ideas of traditional, enlightenment epistemology stand up to the 'primacy of practice' that is at the heart of pragmatic epistemology. What emerges, I contend, is the outlines of a machine epistemology.

§3

Machine learning is a marriage of statistics and computer science that began in artificial intelligence. Statistics, as I remarked earlier, is concerned with the question of what may be inferred from data that we've chosen to model in a particular way, and

[2] See (Dewey 1929, p. 100 and pp. 250–1).
[3] See (Putnam 2004)

it also addresses how to assess the reliability of those inferences. Computer science engages with a different set of questions. It is concerned with the design of algorithms that run on a machine to solve some or another problem of our choosing, and it further addresses the question of which problems are tractable enough to admit a computational solution and which are not. From a modern point of view, machine learning is a discipline that sits at the intersection of statistics and computer science. But the question machine learning asks is fundamentally different from both statistics and computer science. For the question machine learning asks is how to make a computer learn from data without explicitly programming it to do so.

Machine learning, like statistics, is interested in the question of what can be inferred from data. But unlike statistics, machine learning aims to circumvent the requirement to explicitly set modeling assumptions prior to drawing meaningful inferences or, when unavoidable, seeks to design algorithms that learn on their own what modeling assumptions are best to select. Here again we find not an answer to Duhem nor a solution to the problem of Hume's but ways to reason around them.

This focus on designing algorithms which are not told in advance how to solve a particular task has had the peculiar effect of vindicating the insight from statistics that modeling assumptions are necessary to get anything meaningful out of data, but undermining a canard of epistemology that the making of such assumptions invites a regress. The key, which any student of machine learning will recognize, is that one must understand the goal of inquiry in order to pick which features to extract from data and to assess how well one method manages against another. What is perhaps surprising, to traditional statisticians and traditional epistemologists alike, is the extent to which these value-laden judgments can be described in general terms and managed by a machine.

Here the historical role that artificial intelligence played in the rise of machine learning is instructive. The question that artificial intelligence addresses is how to design computer systems that think or act as humans do, or ought rationally to do, and the disjunction between 'thinking' and 'acting' has divided the field, much as it continues to divide philosophy. For those interested in how an artificial intelligence ought to think, the explicit representation of *objects* perceived, *meanings* of sentences understood, or *beliefs* endorsed as true is of central importance, followed closely by the development of some or another set of rules for manipulating these representations to represent *perceiving*, *understanding*, or *reasoning*. Unfortunately, this project has gotten nowhere. Approaches that focus on intelligent behavior, on the other hand, focus on the successful completion of tasks—the parsing of speech, the recognition

of hand-written addresses, the production of a new grammatically correct or contextually meaningful vocalization of speech—have been wildly successful. Behind these successes are machine learning techniques and, often enough, volumes of data whose scale in size is extraordinary.[4]

§4

Consider an example. Suppose that you are interested in the time it takes a taxi cab to travel from John F. Kennedy International Airport to Grand Central Station in Manhattan, New York. The act of taking a particular taxi from JFK to Grand Central, at least according to one textbook approach, is conceived as the selection of a single trip from a large population of such trips, like a single ball drawn from many in a large urn. Naturally, the more you know about the composition of the urn, the better your position to predict what sort of ball you will draw. Similarly, the more you know about the population of JFK-to-Grand Central taxi trips, the better your position to predict how your particular trip will fare.

The first hitch in this line of reasoning is that there is no single canonical population that your trip belongs to. There is the population of trips from JFK to Grand Central that run on the same day of the week as yours, the population of trips that run at the same time of day as yours, the population of trips that run at the same time *and* on the same day, and countless others. Selecting which population your trip belongs to is a version of the *reference class problem.* I say a version of the reference class problem because the original reference class problem concerns how to reconcile conflicting statistical evidence that you already have on hand, whereas the problem here is one of choosing which categories to measure and which to ignore.

The original reference class problem is this: if one possesses statistical information about two or more classes that an uncertain event belongs to, where the parameters of interest—say the *mean* travel time of taxi trips and the *variance* in travel times—have different values within each of the classes, then the original problem of the reference class is how to resolve those conflicts in your evidence. Hans Reichenbach (1938) claimed that to resolve the problem of the reference class one ought to always pick the narrowest reference class for which there are 'adequate' statistics. But Reichenbach offered no account for assessing statistical adequacy, nor did he offer any reason for maintaining that adequate statistics must be unique. The definite description of a particular event, for instance, is the singleton class consisting solely of the event itself.

[4]To give some examples, there are approximately 45 billion indexed webpages in Google's pagerank ≪www.worldwidewebsize.com≫; each person has a genome the length of 3.8×10^9 base pairs, and thousands of people have had their genomes sequenced; WalMart receives 2.5 petabytes of unstructured data from 1 million customers every hour.

This class is by definition both the narrowest possible class and unique. However, this class is hardly adequate for it merely summarizes your current state of uncertainty about the event in question and offers no guidance whatsoever. Henry Kyburg supplemented Reichenbach's specificity condition with a principle he called *richness* (Kyburg 1974; Kyburg and Teng 2001), which serves as a non-trivial adequacy condition, along with another he called *strength* to ensure uniqueness. But while Kyburg's system has some intuitive appeal, the proposal remains deeply controversial.[5]

The alternative version of the reference class problem comes one step before Reichenbach's problem, for we are faced with the problem of selecting which evidence to gather rather than how to adjudicate conflicts among the evidence we have on hand. In practice, background knowledge from prior experience is called on to select which categories to measure and which to ignore. But to many philosophers this appeal to prior experience and absence of general principles for guiding one in selecting which categories to control for and which to ignore has been a source of despair (Whitehead 1925, p. 24).

The second hitch in this line of reasoning is that for practically every population you might choose the mean and variance of travel times will be unknown to you. Traditionally, this problem is solved by measuring the travel times in a sample of taxi trips judged to be representative of the population with respect to these two parameters. This is where the background knowledge of an experimentalist enters the picture and where the misgivings of traditional philosophers takes hold. For the success of this type of statistical inference will hinge on one's choice of which features of the sample to investigate and which to ignore. In effect, our experimentalist must provide an answer to the indirect version of the reference class problem that confronted us before.

From the point of view of traditional epistemology, the pragmatic conception of inquiry is hamstrung because each instance of knowledge seems to rest on a host of assumptions which are without justification. And for those *algorithmic* proposals that do offer a solution, such as Kyburg's, there is no clear *a priori* standard against which to judge its success yet plenty of *a priori* objections with which to contend. It would seem that we should not be able to even get started on this view.

Now imagine an alternative scenario where instead of planning and justifying the pick of a sample of taxi rides to observe, you are provided instead with a record of *every* taxi trip in New York City for the prior year. In fact, at the time of writing, there

[5] Although efficient algorithms exist for calculating Kyburgian solutions to reference class problems (Wheeler and Williamson 2011), outputs can violate fundamental coherence principles underpinning standard as well as set-based Bayes methods (Levi 1977).

are six and a half years of complete yellow taxi trip records freely available to the public.[6] Here, then, one can simply compare various classes to see which give similar estimates, and which give different estimates, and then see whether background knowledge of your particular trip allows you to exploit these differences to yield a good estimate for your trip. One might cross-reference this data-set with the weather conditions in the New York Area, for instance, to consider the effect of weather on travel times. If snow was forecast for the day of your trip, this information might be useful in yielding a better estimate of your travel time. Or, analogously, you might discover that rain has no appreciable effect on travel times and choose to ignore it.

This pattern of reasoning is intuitive even if not terribly clever. But the point is that a lot of ground can be covered without being clever, since much of this sort of reasoning can be automated and evaluated without supplying a justification for the selection of parameters prior to making an inference. A machine can plow through combinatorial combinations of features involving enormous volumes of data, the scale of which is impossible for humans to replicate.

§5

Common sense, craft, and calculation are involved in judging whether a statistical sample should be judged representative of a population. What machine learning techniques have shown, particularly when paired with large volumes of data, is that the roles that craft and common sense play in statistical inference can be minimized or sidestepped altogether by reëngineering the problem so a system can learn hidden structure to use as a substitute for the background knowledge that we might bring to a problem.

There are three main types of machine learning.

Supervised Learning: In a supervised learning task, you are provided with a data set that tells you what a correct answer should look like. For example, in our NYC Taxi data set, one can look at each trip record that starts at JFK airport and plot distance in miles traveled against the duration of the trip. One might then use this data to predict the duration of your trip based on the distance between JFK and Grand Central Station.

More generally, supervised learning problems are categorized into regression and classification problems. In a *regression* problem, the goal is to predict an outcome

[6]The New York City Taxi and Limousine Commission Trip Record Dataset, available at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml, includes all trip records completed in yellow taxis in New York City from January 1, 2009, to June 30, 2015. These records include pick-up and drop-off GPS coordinates, pick-up and drop-off times, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

on a continuous scale; what a regression problem attempts to do is to map an input variable (distance in miles) to some continuous function that predicts a continuous outcome (time). In a *classification* problem, the goal is to predict results in a discrete output, such as 'Pass' or 'Fail', 'low', 'medium', or 'high', or some other discrete set of categories.

Supervised learning problems are thus a form of enumerative induction. However, instead of observing a finite number of input-output pairs of some kind in order to draw an inference about *all* pairs of that kind, as enumerative induction is conceived to do, a supervised learning problem typically observes a finite number of pairs to draw an inference about an unobserved but likewise finite set of pairs. There is no pretense to learning a universal law through supervised learning, and in fact basic results in the field suggest that a search for universality is entirely the wrong approach. For one thing, there are many different supervised learning algorithms that each enjoy different strengths and weaknesses. As a result, the selection of a learning algorithm will depend on the goal of a particular inquiry and the features of the data that one has to work with. In fact, the 'no free lunch' results of (Wolpert and Macready 1997) indicate that there is no single algorithm that can be applied to all supervised learning problems which will outperform the pragmatic strategy of matching the structure of an algorithm to the structure of the problem and the aim of inquiry.

Unsupervised Learning: In an unsupervised learning task, you are provided with little or no idea what the resulting answer ought to look like. Here the task is to derive structure from data where we do not necessarily know the effect of the variables we are learning. Returning to our NYC Taxi data set, one might look at a collection of one million trips originating from JFK and find a way to automatically group trips that are somehow similar. Much of the power of machine learning and most of the controversy over machine learning applications involves unsupervised learning algorithms.

Generally, unsupervised learning problems provide no feedback to the learner based on the predictive results it returns; there is no oracle to correct the learner as there is for supervised learning. Instead, the goal of unsupervised learning is *knowledge discovery*. There are several varieties of unsupervised learning problems.

- In a *clustering* problem, which is the most common form of unsupervised learning, the goal is to derive hidden structure in data based on relationships among variables in the data. Clustering is a means of partitioning a collection of objects into groups that resemble one another along one or more dimensions of comparison, where the categories are not predefined. Among a wide range of applications across the sciences and business, clustering has also been viewed

as a model for concept formation.[7]

- In a *latent factors* learning problem, the goal is to reduce a high-dimensional data set with many variables to a smaller number of latent variables that are most responsible for variability. A common approach to dimensionality reduction is called *principal component analysis* or PCA, which may be thought of as the unsupervised version of a (multi-variate) linear regression.

- In a *graphical structure* learning problem, the goal is say which variables are most correlated with others, where the edges of a graph represent direct dependence, or go further to say which variables stand in the relationship of cause and effect to one another (Spirtes, Glymour, and Scheines 2000).

- In a *matrix completion* learning problem, the goal is to deal with missing or incomplete data by inferring reasonable values for missing data values.[8]

Reinforcement Learning: A reinforcement learning problem is characterized by the learner confronted with a decision problem in which a number of acts yield various (numerical) rewards, and the task for the learner is to discover which act maximizes the numerical reward. Reinforcement learning resembles an unsupervised learning problem in the sense that it does not have a set of training examples of 'correct' behavior from which to learn, although it also has a specified reward structure that the learner 'understands' ought to be maximized. One feature of reinforcement learning problems is that the learner must confront a trade-off between exploration and exploitation. If an agent learns from several rounds of play that a particular strategy yields him a particular reward, in the next round he must weigh playing the same strategy and receiving the same reward against playing an alternative strategy that yields an unknown reward.[9]

Reinforcement learning problems are common in *agent-based simulation* models, where the goal is to predict a group behavior in terms of individuals interacting with particular reinforcement learning strategies which determine how they are rewarded when they interact with one another. ABMs are used to study population dynamics—such as the spread of diseases, the adoption of a norm, or emergence of a signaling system. Their use allows one to explore whether a behavior that one observes in a population which might be puzzling or paradoxical, such as the emergence of cooperation in a society, may be the result of a relatively simple dynamic process.

[7] See in particular (Michalski 1980) and the special issue of *Machine Learning* that includes (Fisher 1987).

[8] See (Laurent 2001) for an introduction.

[9] See (Sutton and Barto 1998).

§6

Unlike a commodity that is consumed, data is an inexhaustible and renewable resource. Only recently has science transformed from a data-scarce enterprise, which I alluded to at the start, to a data-rich one. Data was once expensive, the product of a controlled experiment or some other careful measurement regime. Now we are flooded with the stuff. The sequencing of the first human genome in 2002 determined the order of approximately three billion nucleotides. That achievement took thirteen years, involved twenty institutions, and cost $3 billion. Thirteen years later the cost is now $1,000 to sequence a human genome and a lab can generate more than 300 a week.[10]

To be sure, these new sources of data do not replace the role of controlled experiments and careful measurement, but the era of big data does introduce new opportunities to empirically explore topics cheaply or to empirically explore a question that until recently was difficult, if not impossible, to address.

Martin Nowak and colleagues have looked at the evolutionary dynamics of language to figure out how the rate of verb regularization depends on the frequency of word usage. This is precisely the sort of claim that not too long ago was forwarded *a priori* rather than tackled empirically. Jean-Baptise Michel, Nowak, and colleagues have looked at 4% of all the books ever published, a staggering 5,195,769 digitized books, to explore cultural trends recorded in the English language between 1800 and 2000. Of 'culturomics' they say,

> We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

All this brings three points to mind to consider. First, from the point of view of the philosophy of science, data science arguably does offer a new mode of inquiry insofar as we are now routinely handling population datasets directly or sample sizes so immense, such as in our NYC Taxi dataset, that they behave like population data. In this setting, inferential methods of statistical reasoning are used for an altogether different task, namely, as a form of quality control for the direct methods applied to these

[10] See (May 2015).

enormous data sets. Second, there is now a realignment of interests that will make for new collaboration between business and publicly funded science. The fundamental interests of business and science are more closely aligned than they have been in the past—with the exception of chemistry, perhaps—which means, among several things, that some scientific innovations will come from the business community. Already the leading figures in the emerging field of computational social science are in industry, and many breakthroughs of that field are the intellectual property of companies rather than goods of the commonwealth.

Finally, the collection and storage of society's 'data exhaust' by governments and private companies is easily repurposed for countless tasks, some for the public good, others less so. Big data and machine learning are the mother-of-all dual-use technologies, where 'dual' in this setting is a euphemism for 'countless'. To take one example, PredPol Inc. is a California based company that took an algorithm used to predict earthquakes and modified it to predict where crime is likely to occur within a 500 square foot radius. The same tool that allows WalMart to adjust shipments of items to its stores in anticipation of large weather events can now be used to anticipate criminal behavior and adjust police presence.[11] The software is used by the Los Angeles and Santa Cruz police departments.

§7

Dewey observed that the legal and economic organization of societies places the knowledge of how to regulate activity in the hands of a small number of individuals, and that there is an inclination for those individuals to use that knowledge to benefit themselves rather than the general public. The question for him was how to share such knowledge more widely and how to effect a more equitable participation in their results.[12]

In the age of big data and a machine epistemology that can anticipate, predict, and intervene on events in our lives, the problem once again is that a few individuals possess the knowledge of how to regulate these activities. But the question we face now is not how to share such knowledge more widely, but rather of how to enjoy the public benefits bestowed by this knowledge without freely sharing it. It is not merely personal privacy that is at stake but a range of unsung benefits that come from ignorance and forgetting, traits that are inherently human and integral to the functioning of our society.

[11] See (Beck and McCue 2009).
[12] See (Dewey 1929, pp. 80–1).

References

Beck, C. and C. McCue (2009, November). Predictive policing: What can we learn from walmart and amazon about fighting crime in a recession? *The Police Chief 76*(11).

Dewey, J. (1929). *The Quest for Certainty* (1960, 3rd ed.). New York: Capricorn Books.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning 2*(2), 139–172.

Kyburg, Jr., H. E. (1974). *The Logical Foundations of Statistical Inference*. Dordrecht: D. Reidel.

Kyburg, Jr., H. E. and C. M. Teng (2001). *Uncertain Inference*. Cambridge: Cambridge University Press.

Laurent, M. (2001). Matrix completion problems. In C. Floudas and P. Pardalos (Eds.), *The Encyclopedia of Optimization*, Volume 3, pp. 221–229. Kluwer.

Levi, I. (1977). Direct inference. *Journal of Philosophy 74*, 5–29.

May, M. (2015). Life science technologies: Big biological impacts from big data. *Science 344*(6189), 1298–1300.

Michalski, R. S. (1980). Knowledge acquisition through conceptual clustering: A theoretical framework and an algorithm for partitioning data into conjunctive concepts. *International Journal of Policy Analysis and Information Systems 4*(3), 219–244.

Peirce, C. S. (1955). *Philosophical Writings of Peirce*. New York: Dover.

Putnam, H. (2004). *Ethics without Ontology*. Cambridge, MA: Harvard University Press.

Reichenbach, H. (1938). On probability and induction. *Philosophy of Science 5*(1), 21–45.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Wheeler, G. and J. Williamson (2011). Evidential probability and objective Bayesian epistemology. In P. Bandyopadhyay and M. Forster (Eds.), *Handbook of the Philosophy of Statistics*, pp. 307–331. Elsevier Science.

Whitehead, A. N. (1925). *Science and the Modern World*. Macmillan And Company.

Wolpert, D. H. and W. G. Macready (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation 1*(1), 67–82.