

Prediction Machines

The value of a scientific theory is determined in large part by its ability to make predictions. Quantum electrodynamics, for example, allows the anomalous magnetic moment of the electron to be calculated to more than 10 significant figures, in agreement with comparable experimental observations. Similarly, in applied science and technology, the ability to make predictions, for example, the biological activity of a candidate drug molecule, or the structural and aerodynamic properties of the wing for a new airliner, is again of central importance.

In some domains, the underlying equations are well understood, and in principle, predictions simply involve solving these using appropriate numerical techniques. Even here, however, many research questions often arise due to issues such as efficiency, numerical accuracy and model validation. Performance improvements in computers continually extend the range of predictions which can be obtained by numerical solution from first principles. In chemical dynamics, for instance, computational methods are having a major impact on the capabilities for first-principles prediction. Currently, reaction cross-sections can only be calculated for systems involving a few atoms at a time, and for each additional atom the computational cost increases by two or three orders of magnitude. While this clearly does not scale directly to large systems, the possibility exists to exploit the locality of chemical reactions so that only the handful of atoms directly involved in a reaction need be modelled in detail, with the remainder treated semi-classically, or quasi-classically, thereby opening the door to accurate modelling of complex chemical dynamics on computers having perhaps just a few hundred teraflops of processing power. If such an approach can be developed, the potential impact not only on chemistry but on neighbouring fields could be profound. There is no shortage of other drivers for large-scale simulation of basic physical processes. For instance, modelling of turbulent flows has motivated the development of sophisticated multi-scale techniques, while future grand challenges such as space weather forecasting, or tsunami prediction based on real-time computer simulation driven by inputs from networks of distributed sensors, could potentially lead to new breakthroughs.

It is increasingly easy to write simulation models, and these are intuitively more attractive to the non-mathematically inclined because of their less reductionist character, but arguably they require broad as well as deep mathematical expertise (all the way from statistics and numerical analysis to stochastic process theory and non-linear dynamics) to be applied correctly. Heuristically constructed simulation models, while sometimes producing intriguing results, are next to useless scientifically as the source code is rarely published (and is hard to interpret if it is) leaving at best a qualitative verbal description of the model's structure. Consequently, there is a pressing need to maintain mathematical and scientific rigour, as well as to ensure that models and their implementations are appropriately validated.

However, for most areas of science, the complexity of the domain, or the absence of sufficiently precise models at the appropriate level of description, often prohibit a first-principles simulation with any current or conceivable future level of computational resource. In such cases statistical approaches, in

particular machine learning, have proven to be very powerful. While classical statistics is focused on the analysis of data to test hypotheses, the goal of machine learning is to use (primarily) statistical methods to make predictions. For instance, in the basic supervised learning scenario, a large number of input-response pairs are used to construct a model of the input-output relationship of a system, capturing the underlying trends and extracting them from noise. The data is then discarded and the resulting model used to predict the responses for new inputs. Here, the key issue is that of generalisation, that is the accurate prediction of responses for new inputs, rather than simply the modelling of the training data itself. Machine learning techniques are also used for data visualisation, data mining, screening, and a host of other applications. For instance, in the area of biological modelling, techniques from Inductive Logic Programming (a form of machine learning which represents hypotheses using logic) have been demonstrated on a variety of tasks including the discovery of structural principles concerning the major families of protein folds [24], prediction of structure-activity relations in drugs [25] and prediction of toxicity of small molecules [26]. In addition, Bayesian networks, whose structure is inferred from observed data, have been used to model the effects of toxins on networks of metabolic reactions within cells.

Research into algorithms and methods for machine learning provides insights which can inform research into one of the greatest scientific challenges of our time, namely the understanding of information processing in biological systems including the human brain. For instance, in low-level visual processing, there are interesting similarities between the localised responses of cells in the early visual cortex and the wavelet feature bases which are found to be very effective in computer vision problems such as object recognition. More speculatively, the current interest in hybrids of generative and discriminative models in the machine learning field offers potential insights into the brain's remarkable ability to achieve accurate generalisation from training data which is almost entirely unlabelled.

Ongoing developments in machine learning over the last 5 years have significantly increased the scope and power of machine learning. Three such developments in particular, have been pivotal, namely the widespread adoption of a Bayesian perspective, the use of graphical models to describe complex probabilistic models, and the development of fast and accurate deterministic techniques for approximate solution of inference and learning problems. The Bayesian networks mentioned earlier are a particular instance of graphical models.

Machine learning techniques are not, however, confined to the standard batch paradigm which separates the learning phase from the prediction phase. With active learning techniques, the adaptation to the data and the prediction process are intimately linked, with the model continually pointing to new regions of the space of variables in which to collect or label data so as to be maximally informative. Indeed, as reported recently in *Nature*, an active learning framework was used for choosing and conducting scientific experiments in the Robot Scientist project (see following section on 'Artificial Scientists').

The two approaches to prediction, based, respectively on first-principles simulation and statistical modelling of observed data, need not be exclusive, and there is undoubtedly much to be gained in addressing complex problems by making complementary use of both approaches. In population biology, for example, a complete treatment is likely to require combination of elements from non-linear dynamics, complexity science, network theory, stochastic process theory and machine learning.

Many of the developments in the computational sciences in recent years have been driven by the exponential increase in the performance of computer hardware. However, the limits of the single processor are already being reached, and in order to sustain continued exponential growth, the manufacturers of processors are already moving towards massively multi-core devices, posing some major challenges for software developers. Fortunately, many machine learning algorithms, as well as a significant proportion of numerical simulation methods, can be implemented efficiently on highly parallel architectures. Disk capacity, as well as the size of scientific data sets, have also been growing exponentially (with a shorter doubling time than for processors and memory) and so the need for effective data mining and statistical analysis methods is becoming greater than ever. Coupled with the continuing developments in models and algorithms for machine learning, we can anticipate an ever more central role for statistical inference methods within much of the computational science arena. As probabilistic inference methods become more widely adopted, we can anticipate the development of new programming languages and user tools which embrace concepts such as uncertainty at their core, and which thereby make the process of implementing and applying machine techniques substantially more efficient, as well as making them accessible to a much broader audience.

Christopher Bishop, Stephen Muggleton, Aron Kuppermann, Parviz Moin, Neil Ferguson