The rise of data-driven modelling

The number of physics articles making use of AI technologies keeps growing rapidly. Here are some new directions we find particularly exciting.

The use of machine learning is no news to physicists, who have been early adopters of AI technologies. For example, looking back at the 2011–2012 analysis of the Large Hadron Collider data underlying the discovery of the Higgs boson, machine learning enabled an increase in sensitivity equivalent to collecting 50% more data[1]. But the number of physics papers using machine learning posted on the *arXiv* preprint server, or abstracts submitted to the American Physical Society March and April meetings keeps growing. At the March meeting, the fraction of presentations with "machine learning" in the title or abstract increased from 0.3% in 2015 to 3.4% in 2021, and at the April meeting from 0.085% to 3.3%. Is this trend just reflecting the overall explosion in AI applications, or is there something else going on in physics?

When thinking of AI and neural networks, the first application that comes to mind is classification: does this image represent a cat or a dog, does this jet of particles come from a quark or a gluon? Neural networks are powerful classifiers that have already had a big impact in data-rich fields such as particle physics, astrophysics or X-ray free electron laser experiments[2]. But they are more than that: neural networks can approximate any function with arbitrary precision (here is an intuitive explanation why).

Thinking of neural networks as universal function approximators is particularly empowering for physicists. It is hard to think of a field of physics that does not use partial differential equations. Neural networks can approximate the solutions to partial differential equations much faster than traditional numerical methods. Furthermore, deep neural networks (neural networks with multiple layers) can approximate operators, meaning that they can solve families of partial differential equations.

Neural networks can approximate complicated, ugly functions like many body wavefunctions or interatomic potentials and therefore can be readily integrated into well-established numerical methods such as quantum Monte Carlo or molecular dynamics simulations, overcoming some limitations of traditional methods and speeding up calculations. This approach is likely to push forward the capabilities of current state-of-the-art methods and enable new insights.

This is just the beginning of what may turn out to be a new paradigm: data-driven modelling. Some fields such as fluid dynamics have already made important advances, others are just starting (see for example this recent Perspective). The prospect of not being intimidated by ugly, complex models and not being limited by an incomplete understanding of the underlying physics is certainly attractive, but there is no such thing as a free lunch. Here comes the small print: neural networks can — given enough training data — approximate any function with arbitrary precision.

The availability of data is not necessarily a show stopper. One can use a combination of experimental data or/and surrogate training data from other computational methods. For example, the HEPMASS Data Set containing Monte Carlo simulations of 10.5 million particle collisions and CAMELS, a data set of over 4,000 cosmological simulations, are available for training machine learning algorithms. In some cases, no training data is necessary (see this Tools of the trade piece). There are other emerging directions.

In a Review in this issue, George Em Karniadakis and colleagues discuss physics-informed machine learning in which the algorithm incorporates prior knowledge of the physical laws coming from the observational or theoretical understanding of the world. This approach makes the most of the imperfect data and incomplete knowledge of the model. Moreover, it promises the ability to discover previously unknown physics and to tackle high-dimensional problems.

Machine learning and traditional numerical methods will coexist complementing each other. Data-driven modelling will provide faster or computationally cheaper, sometimes lower-accuracy simulations that can be used for parameter estimation, in multi-scale simulations for the parts that do not require high resolution, for surrogate models and uncertainty quantification[3].

These are early days and the field of data-driven modelling is yet to be defined: a consistent terminology and a taxonomy of the sub-topics needs to be developed by its practitioners. Different directions are waiting to be mapped. We are keen to document the developments in this new area and offer a forum for interdisciplinary dialogue and collaboration in our pages.

1. Radovic, A. et al. Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560**, 41–48 (2018).
2. Ourmazd, A. Science in the age of machine learning. *Nat. Rev. Phys.* **2**, 342–343 (2020).
3. Willard, J. et al. Integrating physics-based modeling with machine learning: a survey. Preprint at https://arxiv.org/abs/2003.04919 (2020).