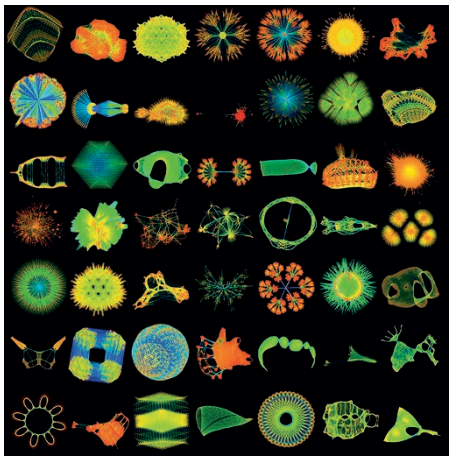
# The thing about data

The rise of big data represents an opportunity for physicists. To take full advantage, however, they need a subtle but important shift in mindset.

Let's face it, we've all been there: the economist making a prediction — down to a tenth of a percent of GDP — of the effect a particular policy will have on the economy. The social psychologist drawing counter-intuitive conclusions from a study based on a tiny sample size. The web developer presenting the results of a questionable A/B test as definitive evidence that readers prefer the version of the *Nature Physics* website without a link to the current issue. "These people don't know what they're talking about," we think. "If only they knew how to analyse data properly."

Physicists, it is fair to say, like to think they understand data. They have the mathematical tools and the empirical expertise to work out the causal relationships between things. Sure, certain systems are more complicated than others — chiefly those made up of chemical, biological or social components — but a well-designed experiment, once properly executed, should yield clean data that can be analysed to make clear, scientific conclusions on the inner workings of nature.



A gallery of large graphs displaying complex data structures. Image credit: © 2011 ACM. Reproduced from T. A. Davis & Y. Hu, *ACM Trans. Math. Softw.* **38,** 1; 2011.

To successfully attack these highly non-trivial problems — a discipline that is, so to speak, a kind of antiphysics — hubris is of no use.

As Jeff Byers underlines in his Commentary on page 718, however, this is perhaps a rose-tinted view. For a start, it tends to downplay the role of experimental design. This is the aspect physicists truly excel at and, by virtue of working in a regime of nature where clean experiments are possible, also rather fortunate to benefit from. Clearly, a good grasp of calculus, statistics and programming helps — from a technical perspective physicists are at a comparative advantage relative to their colleagues in other natural sciences. But the fact is that statisticians, computer scientists, biologists, psychologists and economists often don't have the luxury of dealing with clean data.

Indeed, a growing number of problems in a range of fields come in a form that is the opposite of what physicists are used to seeing. Analysing user interactions on individual websites or across social media, establishing the dynamics of epidemics, working out the effects of economic policy that can be meaningfully quantified: these are often problems where the 'experiments' that can be carried out are far from ideal, but relatively easy to perform. The result is a deluge of data with a very complex structure.

To successfully attack these highly non-trivial problems — a discipline that is, so to speak, a kind of antiphysics — hubris is of no use. Instead, as Byers argues, physicists need to learn a new lexicon in order to translate their knowledge so that it can be useful in unfamiliar surroundings. An important step in this direction is to recognize that, while models have a great deal of applicability for analysing large-scale data sets, their use is but one approach for doing so. In the language of statistics and machine learning, physicists tend to opt for generative models; namely, those that allow for the generation of synthetic data prior to any observation. The predictive power of the Heisenberg model of ferromagnetism is one such example: up to a fitting parameter that sets the energy scale, this model works insofar as it fits the experimental data accounting for the ferromagnetism in a lump of, say, iron. But it remains a caricature of reality, one that rests on multiple assumptions and approximations.

By contrast, discriminative models do not provide a mechanism for how the data might have been generated. Instead, by using a set of techniques that are best thought of as a supervised learning approach, these treat the experimental data as a direct input, which is then used to iteratively improve the model that fits it. Byers refers to this prescription as "letting the model fluctuate around the data", an approach that is possible thanks Bayes's theorem.

Indeed, there is plenty that physicists can learn from machine learning, and in order to bridge this gap Byers advocates for a greater exposure of physics students to statistics and probability, as well as information theory. If complemented by a few key concepts from these disciplines, the techniques that physicists already learn from statistical mechanics and field theory can be transformed for use in the most complex data analysis tasks facing other research disciplines and industries. Central among these is the power of looking backwards through our models — a kind of statistical physics in reverse, as it were — by letting the model fluctuate around the data, as suggested by the Bayesian approach to statistical inference.

To some, it may come as a surprise that mathematical techniques originally developed in physics, such as those required to compute the partition function, have been exported to other domains of research and developed further. It is perhaps time physicists learn about these developments and take them back. The upshot of learning to work with messy data is that there are countless interesting problems to address in this way, and there are countless companies that are willing to pay scientists that are able to do so.

As complexity features ever more prominently in the realm of physics, we need a new generation of physicists equipped with the tools to rise to the challenges this poses. ❐