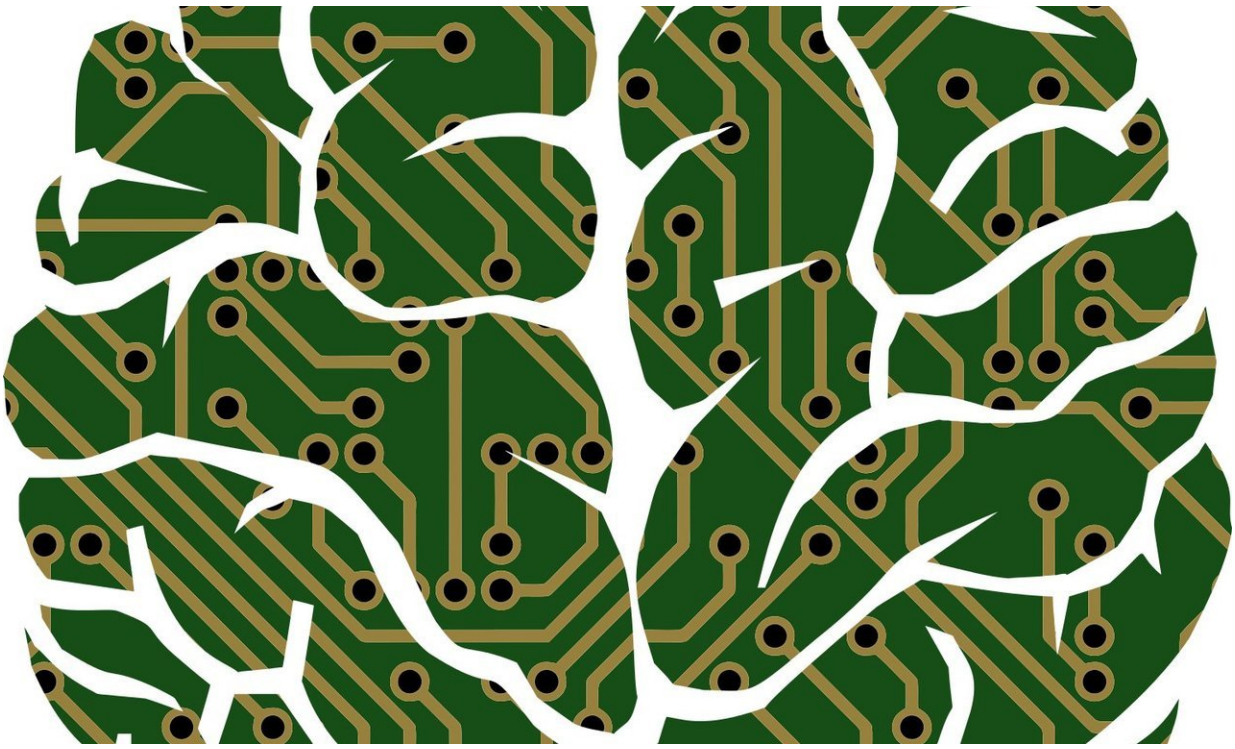
Teaching computers to guide science: Machine learning method sees forests and trees

March 6 2018



Credit: CC0 Public Domain

While it may be the era of supercomputers and "big data," without smart methods to mine all that data, it's only so much digital detritus. Now researchers at the Department of Energy's Lawrence Berkeley National

Laboratory (Berkeley Lab) and UC Berkeley have come up with a novel machine learning method that enables scientists to derive insights from systems of previously intractable complexity in record time.

In a paper published recently in the *Proceedings of the National Academy of Sciences* (*PNAS*), the researchers describe a technique called "iterative Random Forests," which they say could have a transformative effect on any area of science or engineering with complex systems, including biology, precision medicine, materials science, environmental science, and manufacturing, to name a few.

"Take a human cell, for example. There are 10^{170} possible molecular interactions in a single cell. That creates considerable computing challenges in searching for relationships," said Ben Brown, head of Berkeley Lab's Molecular Ecosystems Biology Department. "Our method enables the identification of interactions of high order at the same computational cost as main effects - even when those interactions are local with weak marginal effects."

Brown and Bin Yu of UC Berkeley are lead senior authors of "Iterative Random Forests to Discover Predictive and Stable High-Order Interactions." The co-first authors are Sumanta Basu (formerly a joint postdoc of Brown and Yu and now an assistant professor at Cornell University) and Karl Kumbier (a Ph.D. student of Yu in the UC Berkeley Statistics Department). The paper is the culmination of three years of work that the authors believe will transform the way science is done. "With our method we can gain radically richer information than we've ever been able to gain from a learning machine," Brown said.

The needs of machine learning in science are different from that of industry, where machine learning has been used for things like playing chess, making self-driving cars, and predicting the stock market.

"The machine learning developed by industry is great if you want to do high-frequency trading on the stock market," Brown said. "You don't care why you're able to predict the stock will go up or down. You just want to know that you can make the predictions."

But in science, questions surrounding why a process behaves in certain ways are critical. Understanding "why" allows scientists to model or even engineer processes to improve or attain a desired outcome. As a result, machine learning for science needs to peer inside the black box and understand why and how computers reached the conclusions they reached. A long-term goal is to use this kind of information to model or engineer systems to obtain desired outcomes.

In highly complex systems - whether it's a single cell, the human body, or even an entire ecosystem - there are a large number of variables interacting in nonlinear ways. That makes it difficult if not impossible to build a model that can determine cause and effect. "Unfortunately, in biology, you come across interactions of order 30, 40, 60 all the time," Brown said. "It's completely intractable with traditional approaches to statistical learning."

The method developed by the team led by Brown and Yu, iterative Random Forests (iRF), builds on an algorithm called random forests, a popular and effective predictive modeling tool, translating the internal states of the black box learner into a human-interpretable form. Their approach allows researchers to search for complex interactions by decoupling the order, or size, of interactions from the computational cost of identification.

"There is no difference in the computational cost of detecting an interaction of order 30 versus an interaction of order two," Brown said. "And that's a sea change."

In the PNAS paper, the scientists demonstrated their method on two genomics problems, the role of gene enhancers in the fruit fly embryo and alternative splicing in a human-derived cell line. In both cases, using iRF confirmed previous findings while also uncovering previously unidentified higher-order interactions for follow-up study.

Brown said they're now using their method for designing phased array laser systems and optimizing sustainable agriculture systems.

"We believe this is a different paradigm for doing science," said Yu, a professor in the departments of Statistics and Electrical Engineering & Computer Science at UC Berkeley. "We do prediction, but we introduce stability on top of prediction in iRF to more reliably learn the underlying structure in the predictors."

"This enables us to learn how to engineer systems for goal-oriented optimization and more accurately targeted simulations and follow-up experiments," Brown added.

In a PNAS commentary on the technique, Danielle Denisko and Michael Hoffman of the University of Toronto wrote: "iRF holds much promise as a new and effective way of detecting interactions in a variety of settings, and its use will help us ensure no branch or leaf is ever left unturned."

More information: Sumanta Basu et al, Iterative random forests to discover predictive and stable high-order interactions, *Proceedings of the National Academy of Sciences* (2018). DOI: 10.1073/pnas.1711236115

Provided by Lawrence Berkeley National Laboratory