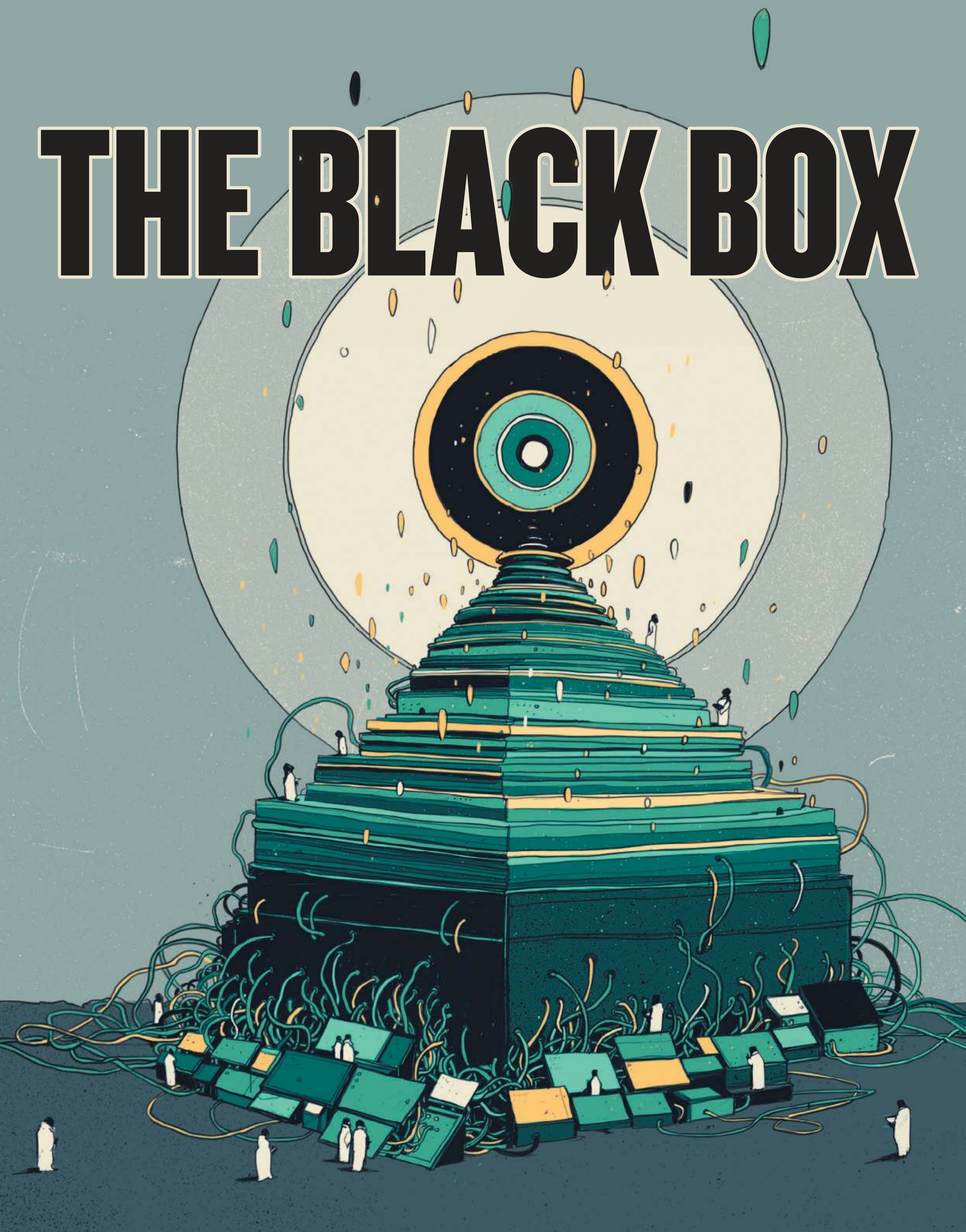
THE BLACK BOX



OF AI

ILLUSTRATION BY SIMON PRADES

Machine learning is becoming ubiquitous in basic research as well as in industry. But for scientists to trust it, they first need to understand what the machines are doing.

BY DAVIDE CASTELVECCHI

Dean Pomerleau can still remember his first tussle with the black-box problem. The year was 1991, and he was making a pioneering attempt to do something that has now become commonplace in autonomous-vehicle research: teach a computer how to drive.

This meant taking the wheel of a specially equipped Humvee military vehicle and guiding it through city streets, says Pomerleau, who was then a robotics graduate student at Carnegie Mellon University in Pittsburgh, Pennsylvania. With him in the Humvee was a computer that he had programmed to peer through a camera, interpret what was happening out on the road and memorize every move that he made in response. Eventually, Pomerleau hoped, the machine would make enough associations to steer on its own.

On each trip, Pomerleau would train the system for a few minutes, then turn it loose to drive itself. Everything seemed to go well — until one day the Humvee approached a bridge and suddenly swerved to one side. He avoided a crash only by quickly grabbing the wheel and retaking control.

Back in the lab, Pomerleau tried to understand where the computer had gone wrong. “Part of my thesis was to open up the black box and figure out what it was thinking,” he explains. But how? He had programmed the computer to act as a ‘neural network’ — a type of artificial intelligence (AI) that is modelled on the brain, and that promised to be better than standard algorithms at dealing with complex real-world situations. Unfortunately, such networks are also as opaque as the brain. Instead of storing what they have learned in a neat block of digital memory, they diffuse the information in a way that is exceedingly difficult to decipher. Only after extensively testing his software’s responses to various visual stimuli did Pomerleau discover the problem: the network had been using grassy roadsides as a guide to the direction of the road, so the appearance of the bridge confused it.

Twenty-five years later, deciphering the black box has become exponentially harder and more urgent. The technology itself has exploded in complexity and application. Pomerleau, who now teaches robotics part-time at Carnegie Mellon, describes his little van-mounted system as “a poor man’s version” of the huge neural networks being implemented on today’s machines. And the technique of deep learning, in which the networks are trained on vast archives of big data, is finding commercial applications that range from self-driving cars to websites that recommend products on the basis of a user’s browsing history.

It promises to become ubiquitous in science, too. Future radio-astronomy observatories will need deep learning to find worthwhile signals in their otherwise unmanageable amounts of data; gravitational-wave detectors will use it to understand and eliminate the tiniest sources of noise; and publishers will use it to scour and tag millions of research papers and books. Eventually, some researchers believe, computers equipped with deep learning may even display imagination and creativity. “You would just throw data at this machine,

▶ and it would come back with the laws of nature," says Jean-Roch Vlimant, a physicist at the California Institute of Technology in Pasadena.

But such advances would make the black-box problem all the more acute. Exactly how is the machine finding those worthwhile signals, for example? And how can anyone be sure that it's right? How far should people be willing to trust deep learning? "I think we are definitely losing ground to these algorithms," says roboticist Hod Lipson at Columbia University in New York City. He compares the situation to meeting an intelligent alien species whose eyes have receptors not just for the primary colours red, green and blue, but also for a fourth colour. It would be very difficult for humans to understand how the alien sees the world, and for the alien to explain it to us, he says. Computers will have similar difficulties explaining things to us, he says. "At some point, it's like explaining Shakespeare to a dog."

Faced with such challenges, AI researchers are responding just as Pomerleau did — by opening up the black box and doing the equivalent of neuroscience to understand the networks inside. Answers are not insight, says Vincenzo Innocente, a physicist at CERN, the European particle-physics laboratory near Geneva, Switzerland who has pioneered the application of AI to the field. "As a scientist," he says, "I am not satisfied with just distinguishing cats from dogs. A scientist wants to be able to say: 'the difference is such and such.'"

GOOD TRIP

The first artificial neural networks were created in the early 1950s, almost as soon as there were computers capable of executing the algorithms. The idea is to simulate small computational units — the 'neurons' — that are arranged in layers connected by a multitude of digital 'synapses' (see 'Do AIs dream of electric sheep?') Each unit in the bottom layer takes in external data, such as pixels in an image, then distributes that information up to some or all of the units in the next layer. Each unit in that second layer then integrates its inputs from the first layer, using a simple mathematical rule, and passes the result further up. Eventually, the top layer yields an answer — by, say, classifying the original picture as a 'cat' or a 'dog'.

The power of such networks stems from their ability to learn. Given a training set of data accompanied by the right answers, they can progressively improve their performance by tweaking the strength of each connection until their top-level outputs are also correct. This process, which simulates how the brain learns by strengthening or weakening synapses, eventually produces a network that can successfully classify new data that were not part of its training set.

That ability to learn was a major attraction for CERN physicists back in the 1990s, when they were among the first to routinely use large-scale neural networks for science: the networks would prove to be an enormous help in reconstructing the trajectories of subatomic shrapnel coming out of particle collisions at CERN's Large Hadron Collider.

But this form of learning is also why information is so diffuse in the network: just as in the brain, memory is encoded in the strength of multiple connections, rather than stored at specific locations, as in a conventional database. "Where is the first digit of your phone number stored in your brain? Probably in a bunch of synapses, probably not too far from the other digits," says Pierre Baldi, a machine-learning researcher at the University of California, Irvine. But there is no well-defined sequence of bits that encodes the number. As a result, says computer scientist Jeff Clune at the University of Wyoming in Laramie, "even though we make these networks, we are no closer to understanding them than we are a human brain".

To scientists who have to deal with big data in their respective disciplines, this makes deep learning a tool to be used with caution. To see why, says Andrea Vedaldi, a computer scientist at the University of Oxford, UK, imagine that in the near future, a deep-learning neural network is trained using old mammograms that have been labelled according to which women went on to develop breast cancer. After this training, says Vedaldi, the tissue of an apparently healthy woman could

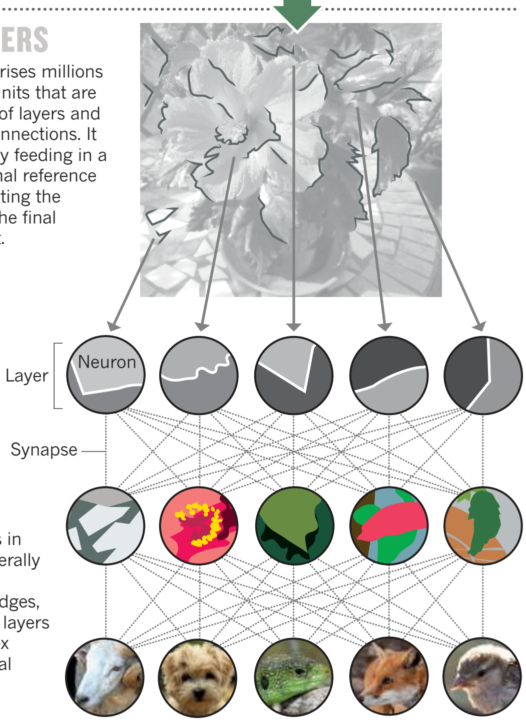
DO AIs DREAM OF ELECTRIC SHEEP?

In an effort to understand how artificial neural networks encode information, researchers invented the Deep Dream technique.

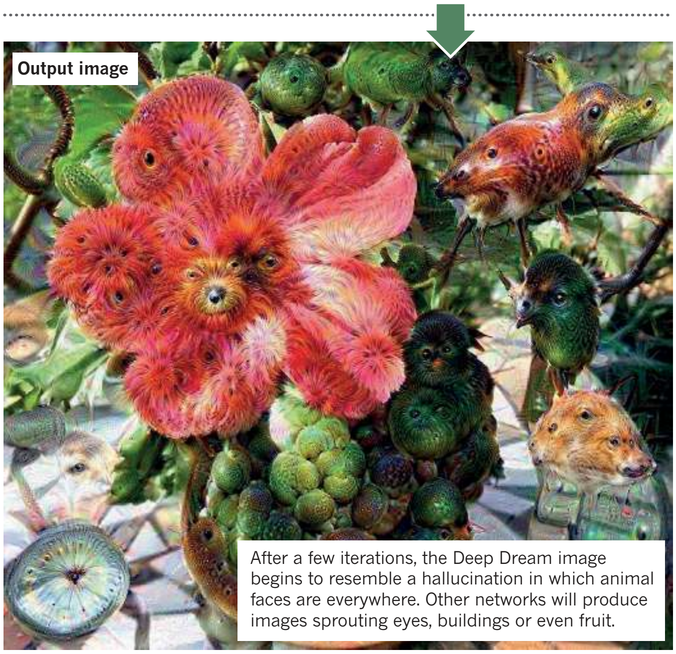
Starting with a network (below) that has been trained to recognize shapes such as animal faces, Deep Dream gives it an image of, say, a flower. Then it repeatedly modifies the flower image to maximize the network's animal-face response.

HIDDEN LAYERS

The network comprises millions of computational units that are stacked in dozens of layers and linked by digital connections. It has been trained by feeding in a vast library of animal reference images, then adjusting the connections until the final response is correct.



After training, units in the first layers generally respond to simple features, such as edges, while intermediate layers respond to complex shapes and the final layers respond to complete faces.



After a few iterations, the Deep Dream image begins to resemble a hallucination in which animal faces are everywhere. Other networks will produce images sprouting eyes, buildings or even fruit.

DESIGN: NIK SPENCER/NATURE; PHOTOS: KEITH MCDUFFEE/FLICKR/CC BY; DJHUMSTER/FLICKR/CC BY-SA; BERNARD DUPONT; LINDA STANLEY; PHIL FIDDYMENT/FLICKR/CC BY

already 'look' cancerous to the machine. "The neural network could have implicitly learned to recognize markers — features that we don't know about, but that are predictive of cancer," he says.

But if the machine could not explain how it knew, says Vedaldi, it would present physicians and their patients with serious dilemmas. It's difficult enough for a woman to choose a preventive mastectomy because she has a genetic variant known to substantially up the risk of cancer. But it could be even harder to make that choice without even knowing what the risk factor is — even if the machine making the recommendation happened to be very accurate in its predictions.

"The problem is that the knowledge gets baked into the network, rather than into us," says Michael Tyka, a biophysicist and programmer at Google in Seattle, Washington. "Have we really understood anything? Not really — the network has."

> "The problem is that the knowledge gets baked into the network, rather than into us."

Several groups began to look into this black-box problem in 2012. A team led by Geoffrey Hinton, a machine-learning specialist at the University of Toronto in Canada, entered a computer-vision competition and showed for the first time that deep learning's ability to classify photographs from a database of 1.2 million images far surpassed that of any other AI approach[1].

Digging deeper into how this was possible, Vedaldi's group took algorithms that Hinton had developed to improve neural-network training, and essentially ran them in reverse. Rather than teaching a network to give the correct interpretation of an image, the team started with pretrained networks and tried to reconstruct the images that produced them[2]. This helped the researchers to identify how the machine was representing various features — as if they were asking a hypothetical cancer-spotting neural network, 'What part of this mammogram have you decided is a marker of cancer risk?'

Last year, Tyka and fellow Google researchers followed a similar approach to its ultimate conclusion. Their algorithm, which they called Deep Dream, starts from an image — say a flower, or a beach — and modifies it to enhance the response of a particular top-level neuron. If the neuron likes to tag images as birds, for example, the modified picture will start showing birds everywhere. The resulting images evoke LSD trips, with birds emerging from faces, buildings and much more. "I think it's much more like a hallucination" than a dream, says Tyka, who is also an artist. When he and the team saw the potential for others to use the algorithm for creative purposes, they made it available to anyone to download. Within days, Deep Dream was a viral sensation online.

Using techniques that could maximize the response of any neuron, not just the top-level ones, Clune's team discovered in 2014 that the black-box problem might be worse than expected: neural networks are surprisingly easy to fool with images that to people look like random noise, or abstract geometric patterns. For instance, a network might see wiggly lines and classify them as a starfish, or mistake black-and-yellow stripes for a school bus. Moreover, the patterns elicited the same responses in networks that had been trained on different data sets[3].

Researchers have proposed a number of approaches to solving this 'fooling' problem, but so far no general solution has emerged. And that could be dangerous in the real world. An especially frightening scenario, Clune says, is that ill-intentioned hackers could learn to exploit these weaknesses. They could then send a self-driving car veering into a billboard that it thinks is a road, or trick a retina scanner into giving an intruder access to the White House, thinking that the person is Barack Obama. "We have to roll our sleeves up and do hard science, to make machine learning more robust and more intelligent," concludes Clune.

Issues such as these have led some computer scientists to think that deep learning with neural networks should not be the only game in town. Zoubin Ghahramani, a machine-learning researcher at the University of Cambridge, UK, says that if AI is to give answers that humans can easily interpret, "there's a world of problems for which deep learning is just not the answer". One relatively transparent approach with an ability to do science was debuted in 2009 by Lipson and computational biologist Michael Schmidt, then at Cornell University in Ithaca, New York. Their algorithm, called Eureqa, demonstrated that it could rediscover the laws of Newtonian physics simply by watching a relatively simple mechanical object — a system of pendulums — in motion[4].

Starting from a random combination of mathematical building blocks such as +, −, sine and cosine, Eureqa follows a trial-and-error method inspired by Darwinian evolution to modify the terms until it arrives at the formulae that best describe the data. It then proposes experiments to test its models. One of its advantages is simplicity, says Lipson. "A model produced by Eureqa usually has a dozen parameters. A neural network has millions."

ON AUTOPILOT

Last year, Ghahramani published an algorithm that automates the job of a data scientist, from looking at raw data all the way to writing a paper[5]. His software, called Automatic Statistician, spots trends and anomalies in data sets and presents its conclusion, including a detailed explanation of its reasoning. That transparency, Ghahramani says, is "absolutely critical" for applications in science, but it is also important for many commercial applications. For example, he says, in many countries, banks that deny a loan have a legal obligation to say why — something a deep-learning algorithm might not be able to do[5].

Similar concerns apply to a wide range of institutions, points out Ellie Dobson, director of data science at the big-data firm Arundo Analytics in Oslo. If something were to go wrong as a result of setting the UK interest rates, she says, "the Bank of England can't say, 'the black box made me do it'".

Despite these fears, computer scientists contend that efforts at creating transparent AI should be seen as complementary to deep learning, not as a replacement. Some of the transparent techniques may work well on problems that are already described as a set of abstract facts, they say, but are not as good at perception — the process of extracting facts from raw data.

Ultimately, these researchers argue, the complex answers given by machine learning have to be part of science's toolkit because the real world is complex: for phenomena such as the weather or the stock market, a reductionist, synthetic description might not even exist. "There are things we cannot verbalize," says Stéphane Mallat, an applied mathematician at the École Polytechnique in Paris. "When you ask a medical doctor why he diagnosed this or this, he's going to give you some reasons," he says. "But how come it takes 20 years to make a good doctor? Because the information is just not in books."

To Baldi, scientists should embrace deep learning without being "too anal" about the black box. After all, they all carry a black box in their heads. "You use your brain all the time; you trust your brain all the time; and you have no idea how your brain works." ■

Davide Castelvecchi *is a reporter for* Nature *in London.*

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105 (2012).
2. Mahendran, A. & Vedaldi, A. Preprint at http://arxiv.org/abs/1412.0035 (2014).
3. Nguyen, A., Yosinski, J. & Clune, J. Preprint at https://arxiv.org/abs/1412.1897 (2014).
4. Lipson, H. & Schmidt, M. *Science* **324,** 81–85 (2009).
5. Ghahramani, Z. *Nature* **521,** 452–459 (2015).