Beyond the Black Box in Analytics and Cognitive

Thomas H. Davenport

There is a growing crisis in the world of analytics and cognitive technologies, and as yet there is no obvious solution. The crisis was created by a spate of good news in the field of cognitive technology algorithms: they're working! Specifically, a relatively new and complex type of algorithms—deep learning neural networks (DLNN)—have been able to learn from lots of labeled data and accomplish a variety of tasks. They can master difficult games (Go, for example), recognize images, translate speech, and perform many more tasks as well or better than the best humans.

So other than the threat to our delicate human egos and jobs, what's the problem? These new DLNN models are successful in that they master the assigned task, but they way they do so is quite uninterpretable, bringing new meaning to the "black box" term. How does the Google algorithm know that a cat photo on the Internet is a cat? It has looked at ten million YouTube videos that are labeled as including cats, identified the features or variables that best discriminate between cats and non-cats, and combined them in a very complex model (as this short article describes). The features are highly abstract and can't be described in a way that would describe a human. In short, the algorithm is about as interpretable as your own neurons' ability to recognize a cat when you see one.

Lack of interpretability is not a great problem for cat images or other low-impact or low-cost decisions. Black box machine learning is also employed, for example, in digital marketing. We can't say with any precision why an algorithm decided that I should see the "How to Make Great Landing Pages—Free Guide" ad when I last consulted the Internet. But since the company offering the ad paid fractions of a cent to show it to me—particularly since I did not click on it—nobody much cares why it popped up.

However, when important things like human lives or big money are concerned, people and organizations start to care about interpretability. Let's say, for example, that a DLNN algorithm identifies a lesion on your chest X-ray as likely to be cancerous, and you have to get a biopsy. Would you be interested in how it came to that conclusion, and how good it is at doing so?

For another type of important decision, let's say that you decide to take your sweetie to a really expensive hotel for Valentine's Day. You're checking in at the hotel and your credit card is denied; the card company's machine learning model has decided that you are likely to be committing fraud. You call the credit card company and ask why your card was turned down. You ask to speak to supervisor after supervisor about this embarrassing incident, but nobody knows why—the model that turned you down is uninterpretable.

On a much larger scale, if bank credit and fraud models are black boxes, then regulators can't review or understand them. If such algorithms don't accurately assess risk, then the financial system could be threatened (as it was in 2008/9). Not surprisingly, many regulators are insisting that credit and risk models be interpretable, and as participants in the financial system we should probably be happy about this, even if we lose some predictive power.

So what can be done about this problem? There are a few solutions, but they don't really address the most complex models like DLNNs. For straightforward statistical models with relatively few variables, it's possible to determine which ones are really having an effect. One credit card executive, for example, told me that both his employees and his customers insist on transparent models. So if they have a model with 8 variables, they put each variable at its mean position, and see what predictive power was lost as a result. This works for 8 variables with real-world referents, but wouldn't work for a DLNN model with highly abstract variables and many more than 8 to boot.

Some other types of cognitive technologies are more interpretable. I'm not an expert on the different types of natural language processing algorithms, but I am told (specifically by Venkat Srinivasan, the CEO of RAGE Networks) that the "computational linguistics" models they use—in which sentences are parsed and their relationships among words graphed—are fairly easily interpreted. His company reports, "…as a totally transparent solution, RAGE AI enables knowledge workers to move forward confidently knowing the reasoning behind the platform's insights is completely auditable." The folks at IPSoft, whose avatar Amelia can handle customer interactions in places like call centers and IT help desks, use similar technology and say the same thing about transparency. But statistical natural language processing—used, for example, in Google Translate—is again achieving a high level of task performance, but is much less interpretable.

You may also be glad to know that academics are working on this problem. You may, however, find their research itself to be largely uninterpretable. For example, at the "Interpretable Machine Learning for Complex Systems" conference, held in Barcelona last December, there were many papers presented on the subject. Actually attending and learning from these papers is well beyond my mathematical and statistical pay grade, but my favorite title is "Gaussian Process Structure Learning via Probabilistic Inverse Compilation." I suspect that it may be a while before my fellow academics shed much light on the black box.

At the moment, the best we can do is to employ models that are relatively interpretable. If we need the analytical power of a DLNN, for example, and a data scientist presents it for consideration, ask him or her to do their best to explain it. Don't expect a lot. Maybe someday the really smart technologies like DLNNs will be smart enough to explain themselves to us.