The machine learning paradox

Nothing says machine learning can't outperform humans, but it's important to realize perfect machine learning doesn't, and won't, exist.

By [Mike Loukides](https://www.oreilly.com/people/1b7da-mike-loukides)

June 1, 2017

"Impossible Triangle," sculpture by Brian MacKay and Ahmad Abas, Claisebrook roundabout, East Perth, Western Australia. (source: [Greg O'Beirne on Wikimedia Commons](https://commons.wikimedia.org/wiki/File:ImpossibleTriangleEastPerth_edit_gobeirne.jpg))

The O'Reilly Artificial Intelligence conference in New York is June 26-29, 2017. [Registration is now open](https://conferences.oreilly.com/artificial-intelligence/ai-ny?intcmp=il-data-confreg-lp-ainy17_20170530_new_site_mike_loukides_post_ml_paradox_top_cta_conf_link).

To train a machine learning system, you start with a lot of training data: millions of photos, for example. You divide that data into a training set and a test set. You use the training set to "train" the system so it can identify those images correctly. Then you use the test set to see how well the training works: how good is it at labeling a different set of images? The process is essentially the same whether you're dealing with images, voices, medical records, or something else. It's essentially the same whether you're using the coolest and trendiest deep learning algorithms, or whether you're using simple linear regression.

But there's a fundamental limit to this process, pointed out in [*Understanding Deep Learning Requires Rethinking Generalization*](https://blog.acolyer.org/2017/05/11/understanding-deep-learning-requires-re-thinking-generalization/). If you train your system so it's 100% accurate on the training set, it will always do poorly on the test set and on any real-world data. It doesn't matter how big (or small) the training set is, or how careful you are. 100% accuracy means that you've built a system that has memorized the training set, and such a system is unlikely to indentify anything that it hasn't memorized. A system that works in the world can't be completely accurate on the training data, but by the same token, it will never be perfectly accurate in the real world, either.

**Get O'Reilly's weekly data newsletter**

Building a system that's 100% accurate on training data is a problem that's well known to data scientists: it's called [overfitting](https://en.wikipedia.org/wiki/Overfitting). It's an easy and tempting mistake to make, regardless of the technology you're using. Give me any set of points (stock market prices, daily rainfall, whatever; I don't care what they represent), and I can find an equation that will pass through them all. Does that equation say anything at all about the next point you give me? Does it tell me how to invest or what raingear to buy? No—all my equation has done is "memorize" the sample data. Data only has predictive value if the match between the predictor and the data isn't perfect. You'll be much better off getting out a ruler and eyeballing the straight line that comes closest to fitting.

If a usable machine learning system can't identify the training data perfectly, what does that say about its performance on real-world data? It's also going to be imperfect. How imperfect? That depends on the application. 90-95% accuracy is achievable in many applications, maybe even 99%, but never 100%. That doesn't mean machine learning applications aren't useful. It does mean we have to be aware that machine learning is never going to be a 100% solution, particularly as we rush to use it in applications as diverse as [sentencing criminals](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing) and [planning menus](https://www.ncbi.nlm.nih.gov/pubmed/16400374). What are the error rates, and can we tolerate them? Are error rates higher on some portions of the population than others? I don't care too much if Amazon recommends books I don't want to buy or if a menu has items I don't want to eat. But accuracy is far more important in many other applications of AI. Face recognition systems used in law enforcement [have long had problems with racial bias](https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/). That article notes [a study](https://pdfs.semanticscholar.org/5fb9/4eaf71196ed9cb87bbd881af34be8b4bc919.pdf) showing that systems developed in China, Japan, and South Korea were much better at identifying East Asian faces than Caucasian ones. [Another study](http://ieeexplore.ieee.org/document/6327355/) shows that face recognition software in use by police departments performs significantly worse on black faces. This undoubtedly has an effect on criminal convictions.

Another way of looking at the problem is through the [Receiver Operating Characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic), an important idea from the cold war and the early days of radar and pattern recognition. The ROC says, essentially, that you can't have a perfectly accurate system. You can't have zero false negatives (100% true positives) and zero false positives. You can have either one, rather trivially: if you're detecting incoming nuclear warheads, you can achieve zero false negatives by nailing up a big sign saying "Incoming! We're all going to die!" Of course, that means you have 100% false positives (and zero true positives). You can minimize both false negatives and false positives, but only up to a point, and at great expense. The right question to ask isn't how to make an error-free system; it's how much error you're willing to tolerate, and how much you're willing to pay to reduce errors to that level.

People aren't perfect, either: the receiver operating characteristic applies as much to humans as to machines. We mislabel photos, misidentify people, and make mistakes in all sorts of ways. We drive imperfectly; we convict criminals imperfectly; we're also more likely to excuse or justify our own mistakes than a computer's. Nothing says that machine learning can't outperform humans, or that it can't be a valuable tool in assisting our judgement. But it's important to realize perfect machine learning doesn't, and won't, exist. It's as impossible as light-speed travel.

Article image: "Impossible Triangle," sculpture by Brian MacKay and Ahmad Abas, Claisebrook roundabout, East Perth, Western Australia. (source: [Greg O'Beirne on Wikimedia Commons](https://commons.wikimedia.org/wiki/File:ImpossibleTriangleEastPerth_edit_gobeirne.jpg)).

Share

1. [inShare](javascript:void(0);)478

[](https://www.oreilly.com/people/1b7da-mike-loukides)

[**Mike Loukides**](https://www.oreilly.com/people/1b7da-mike-loukides)

Mike Loukides is Vice President of Content Strategy for O'Reilly Media, Inc. He's edited many highly regarded books on technical subjects that don't involve Windows programming. He's particularly interested in programming languages, Unix and what passes for Unix these days, and system and network administration. Mike is the author of System Performance Tuning and a coauthor of Unix Power Tools. Most recently, he's been fooling around with data and data analysis, languages like R, Mathematica, and Octave, and thinking about how to make books soc...