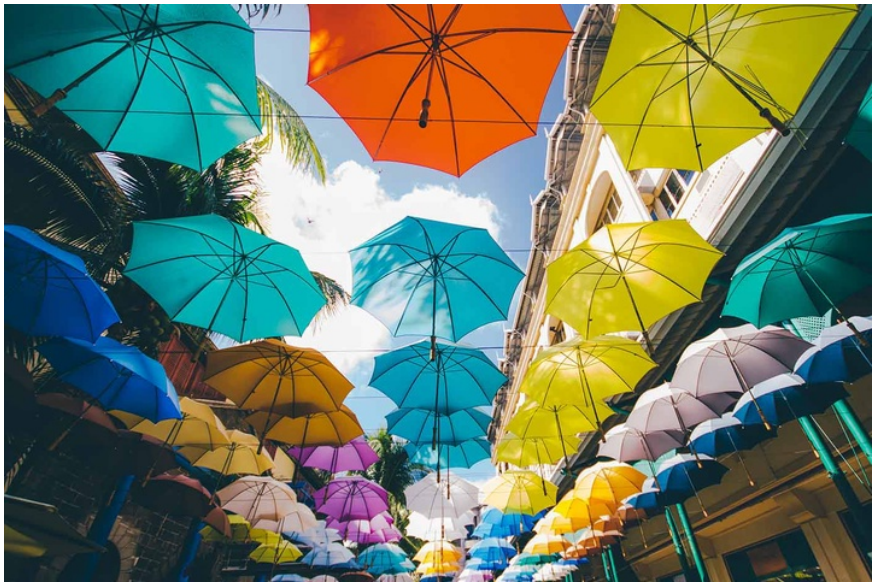
Introduction to Local Interpretable Model-Agnostic Explanations (LIME)

A technique to explain the predictions of any machine learning classifier.

By Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. August 12, 2016



Happy predictions. (source: Jared Hersch on Flickr)

Check out the Data science and machine learning sessions at Strata Data in New York, September 25-28, 2017, for more on current trends and practical use cases in applied data science.

Machine learning is at the core of many recent advances in science and technology. With computers beating professionals in games like Go, many people have started asking if machines would also make for better drivers or even better doctors.

In many applications of machine learning, users are asked to trust a model to help them make decisions. A doctor will certainly not operate on a patient simply because "the model said so." Even in lower-stakes situations, such as when choosing a movie to

watch from Netflix, a certain measure of trust is required before we surrender hours of our time based on a model. Despite the fact that many machine learning models are black boxes, understanding the rationale behind the model's predictions would certainly help users decide when to trust or not to trust their predictions. An example is shown in Figure 1, in which a model predicts that a certain patient has the flu. The prediction is then explained by an "explainer" that highlights the symptoms that are most important to the model. With this information about the rationale behind the model, the doctor is now empowered to trust the model—or not.

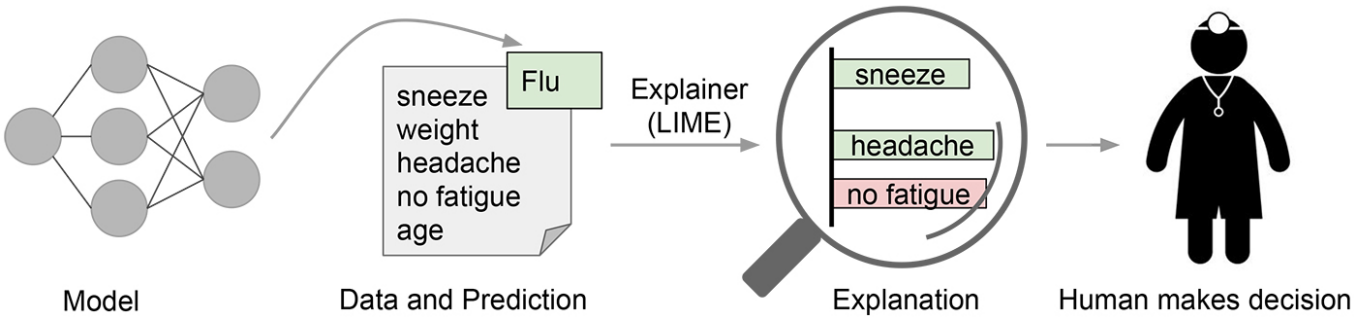


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

In a sense, every time an engineer uploads a machine learning model to production, the engineer is implicitly trusting that the model will make sensible predictions. Such assessment is usually done by looking at held-out accuracy or some other aggregate measure. However, as anyone who has ever used machine learning in a real application can attest, such metrics can be very misleading. Sometimes data that shouldn't be available accidentally leaks into the training and into the held-out data (e.g., looking into the future). Sometimes the model makes mistakes that are too embarrassing to be acceptable. These and many other tricky problems indicate that understanding the model's predictions can be an additional useful tool when deciding if a model is trustworthy or not, because humans often have good intuition and business intelligence that is hard to capture in evaluation metrics. Assuming a “pick step” in which certain representative predictions are selected to be explained to the human would make the process similar to the one illustrated in Figure 2.

Strata Data Conference in San Jose, March 5-8, 2018

Check it out.

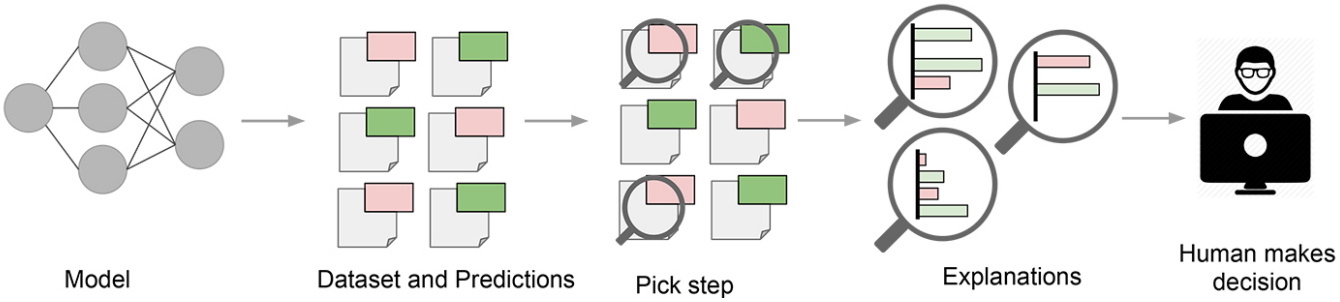


Figure 2. Explaining a model to a human decision-maker. Source: Marco Tulio Ribeiro.

In "Why Should I Trust You?" Explaining the Predictions of Any Classifier, a joint work by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin (to appear in ACM's Conference on Knowledge Discovery and Data Mining -- KDD2016), we explore precisely the question of trust and explanations. We propose Local Interpretable Model-Agnostic Explanations (LIME), a technique to explain the predictions of *any* machine learning classifier, and evaluate its usefulness in various tasks related to trust.

Intuition behind LIME

Because we want to be model-agnostic, what we can do to learn the behavior of the underlying model is to perturb the input and see how the predictions change. This turns out to be a benefit in terms of interpretability, because we can perturb the input by changing components that make sense to humans (e.g., words or parts of an

image), even if the model is using much more complicated components as features (e.g., word embeddings).

We generate an explanation by approximating the underlying model by an interpretable one (such as a linear model with only a few non-zero coefficients), learned on perturbations of the original instance (e.g., removing words or hiding parts of the image). The key intuition behind LIME is that it is much easier to approximate a black-box model by a simple model *locally* (in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally. This is done by weighting the perturbed images by their similarity to the instance we want to explain. Going back to our example of a flu prediction, the three highlighted symptoms may be a faithful approximation of the black-box model for patients who look like the one being inspected, but they probably do not represent how the model behaves for all patients.

See Figure 3 for an example of how LIME works for image classification. Imagine we want to explain a classifier that predicts how likely it is for the image to contain a tree frog. We take the image on the left and divide it into interpretable components (contiguous superpixels).

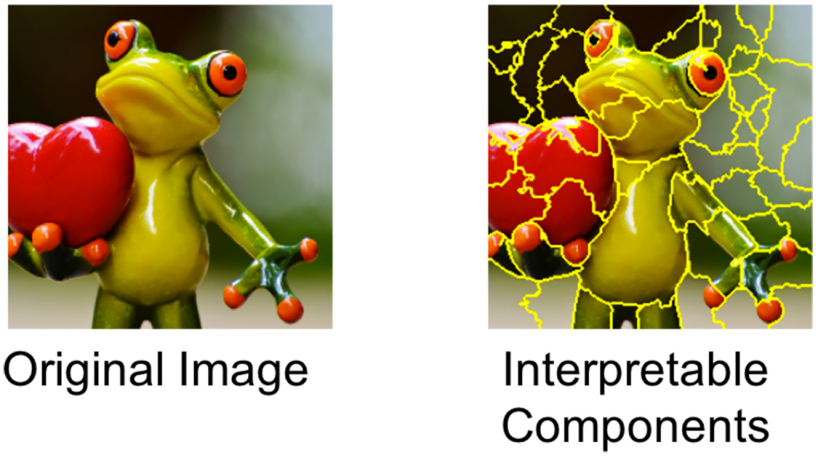


Figure 3. Transforming an image into interpretable components. Sources: Marco Tulio Ribeiro, Pixabay.

As illustrated in Figure 4, we then generate a data set of perturbed instances by turning some of the interpretable components "off" (in this case, making them gray). For each perturbed instance, we get the probability that a tree frog is in the image according to the model. We then learn a simple (linear) model on this data set, which is locally weighted—that is, we care more about making mistakes in perturbed instances that are more similar to the original image. In the end, we present the

superpixels with highest positive weights as an explanation, graying out everything else.

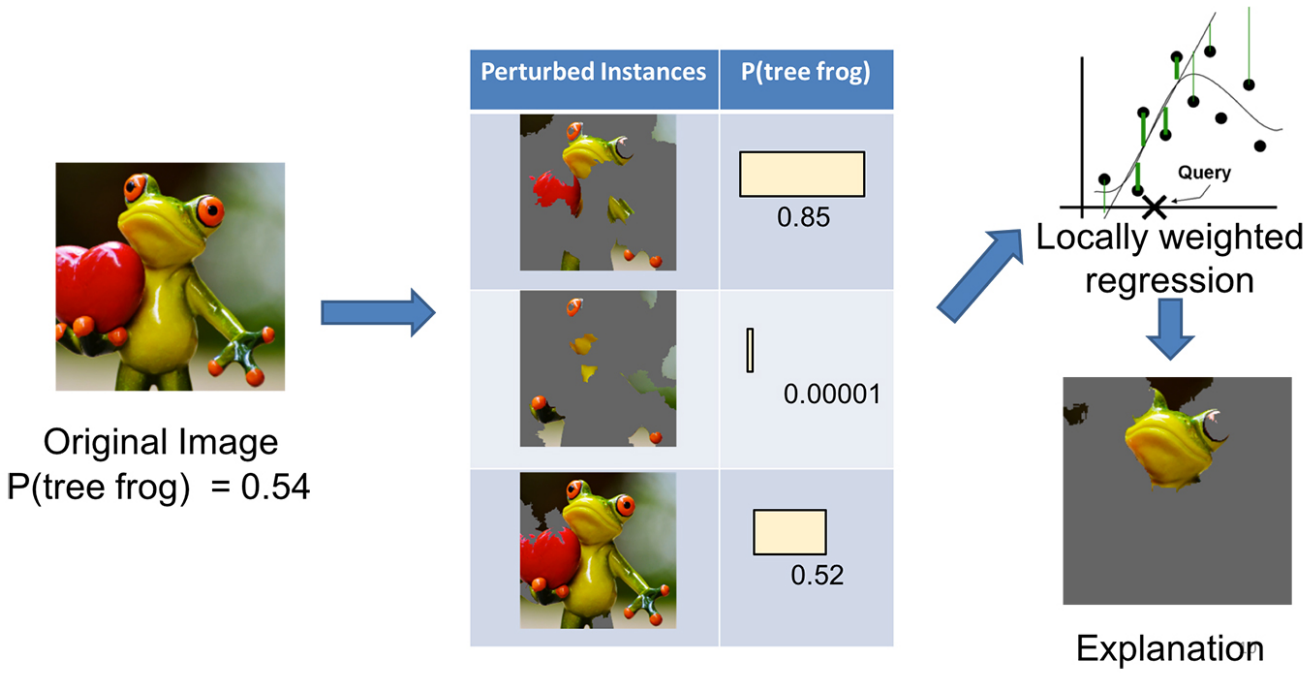


Figure 4. Explaining a prediction with LIME. Sources: Marco Tulio Ribeiro, Pixabay.

Examples

We used LIME to explain a myriad of classifiers (such as random forests, support vector machines (SVM), and neural networks) in the text and image domains. Here are a few examples of the generated explanations.

First, an example from text classification. The famous 20 newsgroups data set is a benchmark in the field, and has been used to compare different models in several papers. We take two classes that are hard to distinguish because they share many words: Christianity and atheism. Training a random forest with 500 trees, we get a test set accuracy of 92.4%, which is surprisingly high. If accuracy was our only measure of trust, we would definitely trust this classifier. However, let's look at an explanation in Figure 5 for an arbitrary instance in the test set (a one liner in Python with our open source package):

```
exp = explainer.explain_instance(test_example,
classifier.predict_proba, num_features=6)
```

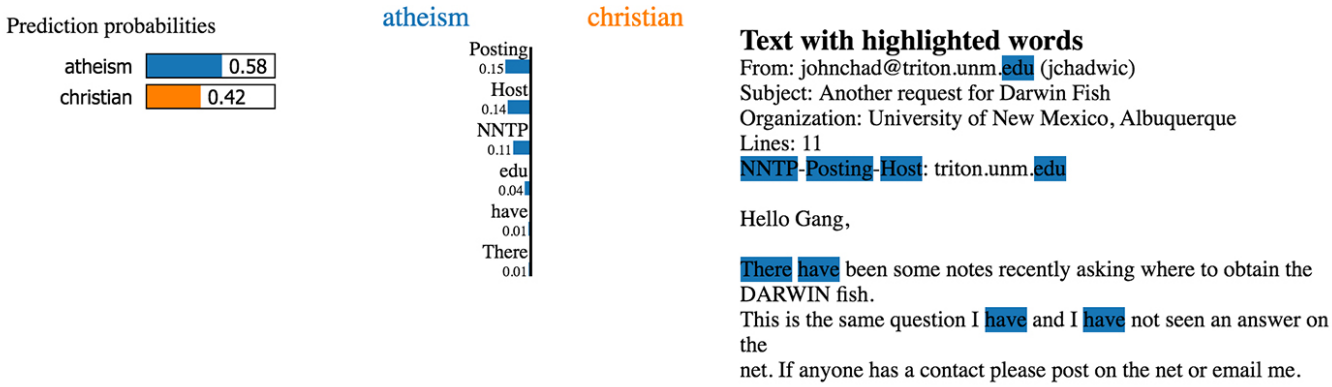


Figure 5. Explanation for a prediction in the 20 newsgroups data set. Source: Marco Tulio Ribeiro.

This is a case in which the classifier predicts the instance correctly, but for the wrong reasons. Additional exploration shows us that the word "posting" (part of the email header) appears in 21.6% of the examples in the training set but only two times in the class "Christianity." This is also the case in the test set, where the word appears in almost 20% of the examples but only twice in "Christianity." This kind of artifact in the data set makes the problem much easier than it is in the real world, where we wouldn't expect such patterns to occur. These insights become easy once you understand what the models are actually doing, which in turn leads to models that generalize much better.

As a second example, we explain Google's Inception neural network on arbitrary images. In this case, illustrated in Figure 6, the classifier predicts "tree frog" as the

most likely class, followed by "pool table" and "balloon" with lower probabilities. The explanation reveals that the classifier primarily focuses on the frog's face as an explanation for the predicted class. It also sheds light on why "pool table" has non-zero probability: the frog's hands and eyes bear a resemblance to billiard balls, especially on a green background. Similarly, the heart bears a resemblance to a red balloon.

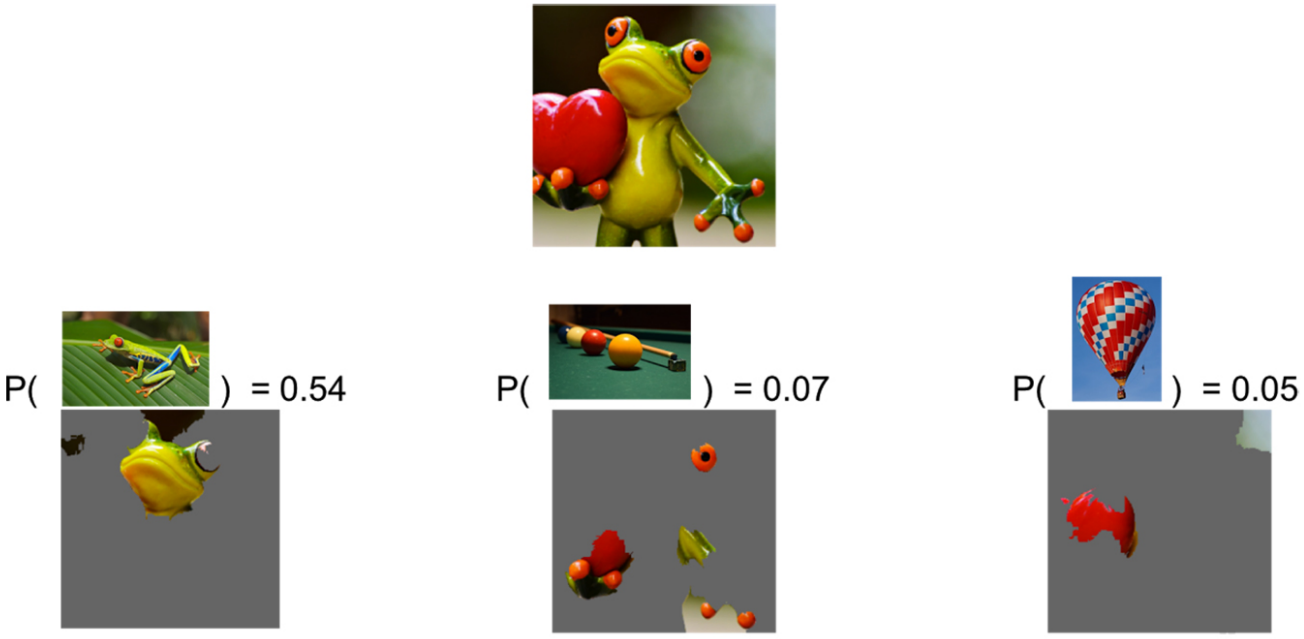


Figure 6. Explanation for a prediction from Inception. The top three predicted classes are "tree frog," "pool table," and "balloon." Sources: Marco Tulio Ribeiro, Pixabay (frog, billiards, hot air balloon).

In the experiments in our research paper, we demonstrate that both machine learning experts and lay users greatly benefit from explanations similar to Figures 5 and 6 and are able to choose which models generalize better, improve models by changing them, and get crucial insights into the models' behavior.

Conclusion

Trust is crucial for effective human interaction with machine learning systems, and we think explaining individual predictions is an effective way of assessing trust. LIME is an efficient tool to facilitate such trust for machine learning practitioners and a good choice to add to their tool belts (did we mention we have an open source project?), but there is still plenty of work to be done to better explain machine learning models. We're excited to see where this research direction will lead us. The video below provides an overview of LIME, with more details available in our paper.

Article image: Happy predictions. (source: Jared Hersch on Flickr).

Marco Tulio Ribeiro

Marco Tulio Ribeiro is a PhD student at the University of Washington working under Carlos Guestrin. His research focus is making it easier for humans to understand and interact with machine learning models.

more

Sameer Singh

Dr. Sameer Singh is an Assistant Professor of Computer Science at University of California, Irvine, conducting research on large-scale, interactive machine learning and natural language processing.

more

Carlos Guestrin

Carlos is the CEO of Turi, Inc., and the Amazon Professor of Machine Learning in Computer Science & Engineering at the University of Washington. A world-recognized leader in the field of Machine Learning, Carlos was named one of the 2008 "Brilliant 10" by Popular Science Magazine, received the 2009 IJCAI Computers and Thought Award for his contributions to Artificial Intelligence, and a Presidential Early Career Award for Scientists and Engineers (PECASE).

more

How intelligent data platforms are powering smart cities

By Ben Lorica

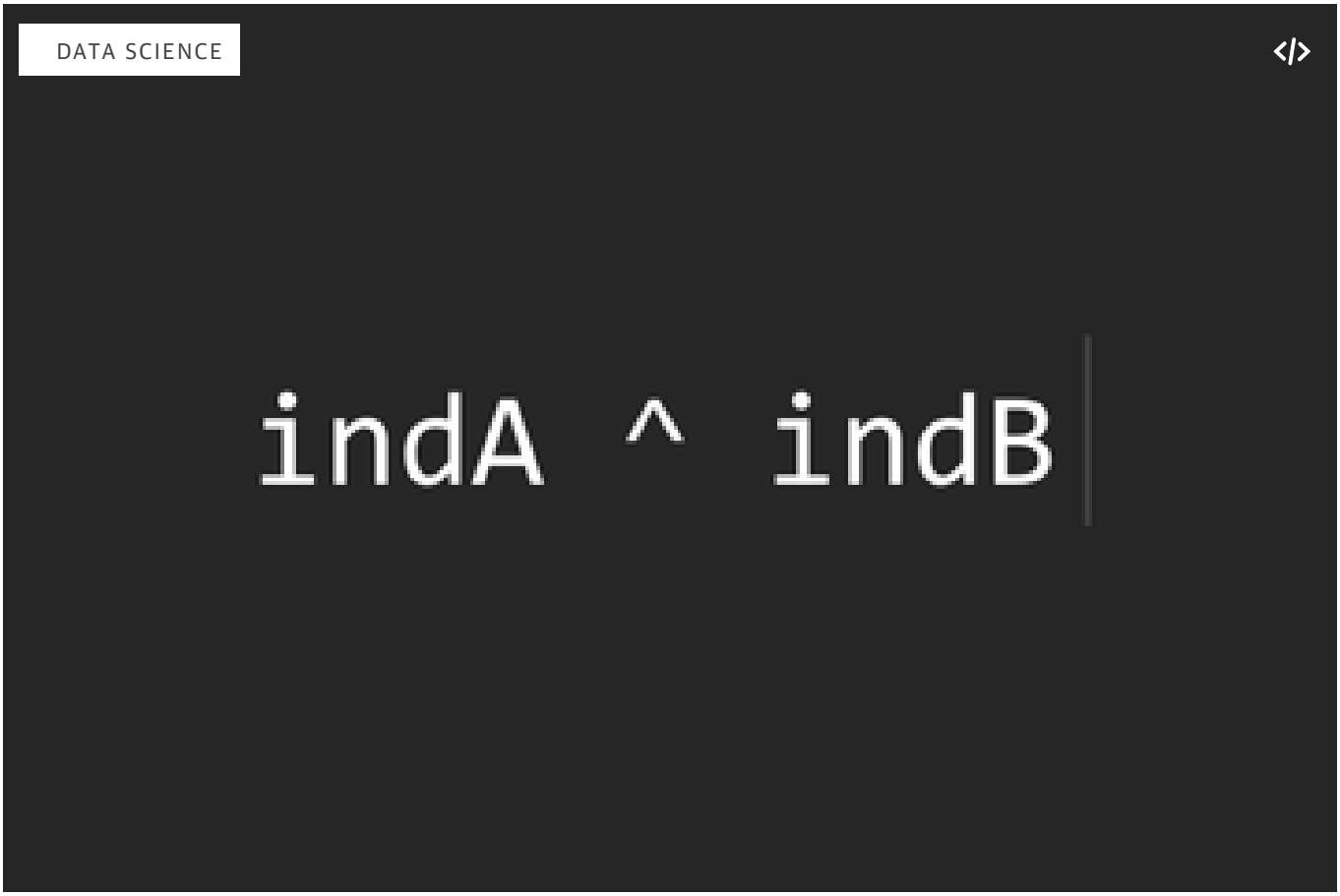
Smart cities and smart nations run on data.

DATA SCIENCE

Beyond algorithms: Optimizing the search experience

By Daniel Tunkelang

Making search smarter through better human-computer interaction.



Introducing Pandas Objects

By Jake VanderPlas

Python Data Science Handbook: Early Release

The machine learning paradox

By Mike Loukides

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