# Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning

Richard Caruana
Microsoft
rcaruana@microsoft.com

## ABSTRACT

Every data set is flawed, often in ways that are unanticipated and difficult to detect. If you can't understand what your model has learned, then you almost certainly are shipping models that are less accurate than they could be and which might even be risky. Historically there has been a tradeoff between accuracy and intelligibility: accurate models such as neural nets, boosted tress and random forests are not very intelligible, and intelligible models such as logistic regression and small trees or decision lists usually are less accurate. In mission-critical domains such as healthcare, where being able to understand, validate, edit and ultimately trust a model is important, one often had to choose less accurate models. But this is changing. We have developed a learning method based on generalized additive models with pairwise interactions (GA2Ms) that is as accurate as full complexity models yet even more interpretable than logistic regression. In this talk I'll highlight the kinds of problems that are lurking in all of our datasets, and how these interpretable, high-performance GAMs are making what was previously hidden, visible. I'll also show how we're using these models to uncover bias in models where fairness and transparency are important. (Code for the models has recently been released open-source.).

## BIOGRAPHY

Rich Caruana is a Principal Researcher at Microsoft. His research focus is on intelligible/ transparent modeling, machine learning for medical decision making, deep learning, and computational ecology. Before joining Microsoft, Rich was on the faculty in Computer Science at Cornell, at UCLA's Medical School, and at CMU's Center for Learning and Discovery. Rich's Ph.D. is from CMU, where he worked with Tom Mitchell and Herb Simon. His thesis on Multitask Learning helped create interest in a subfield of machine learning called Transfer Learning. Rich received an NSF CAREER Award in 2004 (for Meta Clustering), best paper awards in 2005 (with Alex Niculescu-Mizil), 2007 (with Daria Sorokina), and 2014 (with Todd Kulesza, Saleema Amershi, Danyel Fisher, and Denis Charles), and co-chaired KDD in 2007 with Xindong Wu..


*KDD '19, August 4–8, 2019, Anchorage, AK, USA.*

ACM ISBN 978-1-4503-6201-6/19/08.
DOI: https://doi.org/10.1145/3292500.3340414