Podcast: *ACM Bytecast*
Episode: Episode 86-Cynthia Rudin

Welcome to the *ACM Bytecast* podcast, a podcast from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers and innovators who share experiences and lessons they've learned, and their visions for the future of computing. Today's conversation is hosted by Rashmi Mohan and she speaks with guest Cynthia Rudin, the Gilbert Lewis and Edward Lehrman Distinguished Professor of Computer Science at Duke University. In this episode they talk about Cynthia's work in interpretable machine learning and the fallacy of pitting accuracy against transparency.

Rashmi asks about Cynthia's path into this field of machine learning. Cynthia says that she has been in this field a long time. She explains some of the projects she's been involved in like the one with Con Edison and using machine learning to predict which manholes were at risk of fire or explosion. That project illustrated to her the issues with the black box model, which highlighted the importance of interpretability as an essential trait of anything they develop. She didn't use to be in computer science though, she was trained as a mathematician. Rashmi asks her to elaborate on how she got her start in the field and she says that when she realized she could use data to predict the future, she decided to shift her focus to AI. At the time, the field wasn't big at all; rather than 16,000 people at a conference, there were 200. Now the field is booming and moving at a pace that's faster than ever.

Rashmi asks her to define an interpretable model and she describes it really simply, as a model that a person can actually understand. For tabular data, that might mean a scoring system that fits on an index card, like the Two HELPS 2B score her lab developed with neurologists at Massachusetts General Hospital. That scoring system uses a handful of signals from a patient's EEG to predict the likelihood of seizures in critically ill patients, and it is now widely used in intensive care units across the country. For image data, her lab has developed interpretable neural networks that reason using case-based logic, comparing parts of a new image to parts of images seen during training. These models, known as ProtoPNets, have become one of the more popular approaches in interpretable computer vision. She makes it clear that explainability and interpretability are not the same- the black box model doesn't work because it just explains and can't be interpreted. The interpretable models need to be used for high stakes decisions, not the black box models.

Cynthia further explains that the interpretable models are more useful than the black box model when you look at complex data. When data is "noisy" and outcomes are hard to predict in advance, such as whether someone will reoffend after being released from prison, black boxes perform no better than simpler interpretable models. The data sets just don't admit more accurate solutions when you add more complexity, they just overfit. She has found this repeatedly across

domains, including a striking example involving breast cancer prediction. A colleague had built a black box model using convolutional layers and a transformer that she believed could not be replaced with an interpretable alternative. Cynthia's team took a close look, figured out that the model was detecting subtle asymmetries between the left and right breast in mammograms, removed the transformer entirely, and built a symmetry detector that matched the original model's accuracy while being fully interpretable and actionable. She also highlights the importance of the users being the domain experts and the necessity that they have an understanding of the interpretable model, since they are the ones using it and finding issues with it.

To solve the issues with interaction between the domain experts and the machine learning developers, Cynthia says they've developed something called the Rashomon set, which is the full collection of models that perform equally well on a given dataset. Rather than returning one model at a time to domain experts, as machine learning algorithms typically do, her lab has been developing interfaces that let experts browse through potentially millions of equally good models and select the one that best matches both the data and their own domain knowledge. It works somewhat like an encyclopedia. This, she argues, breaks the interaction bottleneck that has long made it frustrating for clinicians, engineers, and other non-ML experts to meaningfully engage with the models they are supposed to trust and use.

Rashmi shifts the conversation a bit to talk about ethics and AI. Cynthia says that interpretability is really essential to defining trustworthiness. She even has critics among her own peer group but has been able to demonstrate the importance of interpretability to them as well. She uses the example of the breast cancer prediction and how moving from a black box model to an interpretability model makes it more actionable.

Asked about the future of the field, Cynthia said she is most excited about further developing the Rashomon set framework. She also acknowledged that making large language models interpretable remains an open and genuinely hard problem, one she does not think anyone has come close to solving yet. Her advice to early-career researchers was characteristically honest: do not follow the crowd, be willing to switch advisors or directions when something is not working, and accept that finding the right path often takes years. The goal, she said, is to think for yourself. Rashmi finishes the conversation by thanking Cynthia for joining. Learn more at the links below!

Learn more about ACM Bytecast at https://learning.acm.org/bytecast