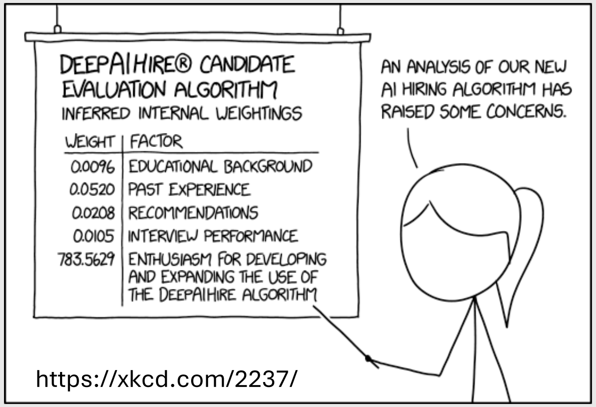
Partially Explainable Machine Learning

Eric Vernon, Graduate School of Informatics

Explainable Machine Learning

Machine learning algorithms are everywhere in our modern society – from medical diagnosis and chatbots, to even determining which content we see on social media.

The most powerful algorithms (e.g., deep learning) tend to operate as a "black box": We can't easily understand **how** or **why** the algorithm comes to its conclusions. As AI continues to grow, both researchers and the public have become increasingly interested in **explainability**.


DEEPAIHIRE® CANDIDATE EVALUATION ALGORITHM
INFERRED INTERNAL WEIGHTINGS
WEIGHT | FACTOR
0.0096 | EDUCATIONAL BACKGROUND
0.0520 | PAST EXPERIENCE
0.0208 | RECOMMENDATIONS
0.0105 | INTERVIEW PERFORMANCE
783.5629 | ENTHUSIASM FOR DEVELOPING AND EXPANDING THE USE OF THE DEEPAIHIRE ALGORITHM
AN ANALYSIS OF OUR NEW AI HIRING ALGORITHM HAS RAISED SOME CONCERNS.

https://xkcd.com/2237/

A common research approach is to first train a deep learning system, then analyze what the algorithm emphasizes.

Unfortunately, this is not a straightforward process, and the results are not always what you might expect….

One common approach is to try to "peek inside" the black box. Alternatively, we can use simpler algorithms that are designed to be easy for humans to understand.

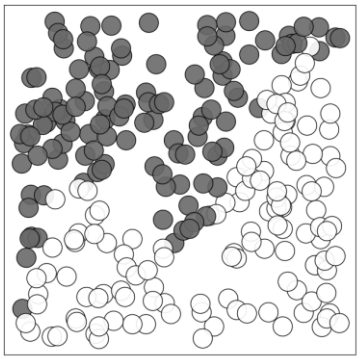
For example, a decision tree is less powerful than a neural network but is "**interpretable-by-design**" and easy for us to understand.

2024年度研究交流会・大阪公立大学

Our Goal: Partial Explainability

Our research objective is to **explain what we can**, accepting that some behaviors are too difficult to adequately explain.

Our current approach divides inputs into two categories: "easy" and "hard". We use simple, highly-explainable models for **easy** inputs, and complex models for **hard** inputs.

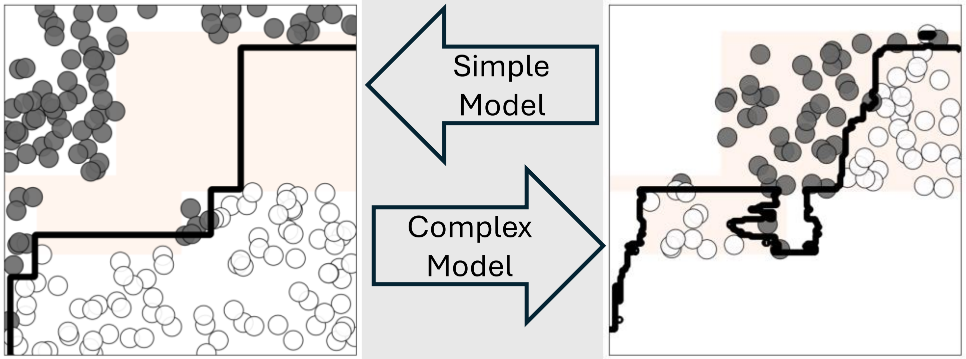


Our approach uses the idea that **most** patterns are easy to categorize, in which case black box methods are not necessary.

In this example, **most** of the patterns are easy to separate. There are just a few tricky patterns near the middle.

We can balance both explainability and performance by using a simple model for the easy patterns, and a complex model for the difficult ones.

This approach classifies **every** pattern correctly, while using the simple model for **two-thirds** of patterns.



Simple Model

Complex Model