Model Explainability

Machine Learning is often seen as black boxes that are difficult to interpret. However, there are techniques that help explain what a model is doing overall, and for specific observations

Why Explainability:

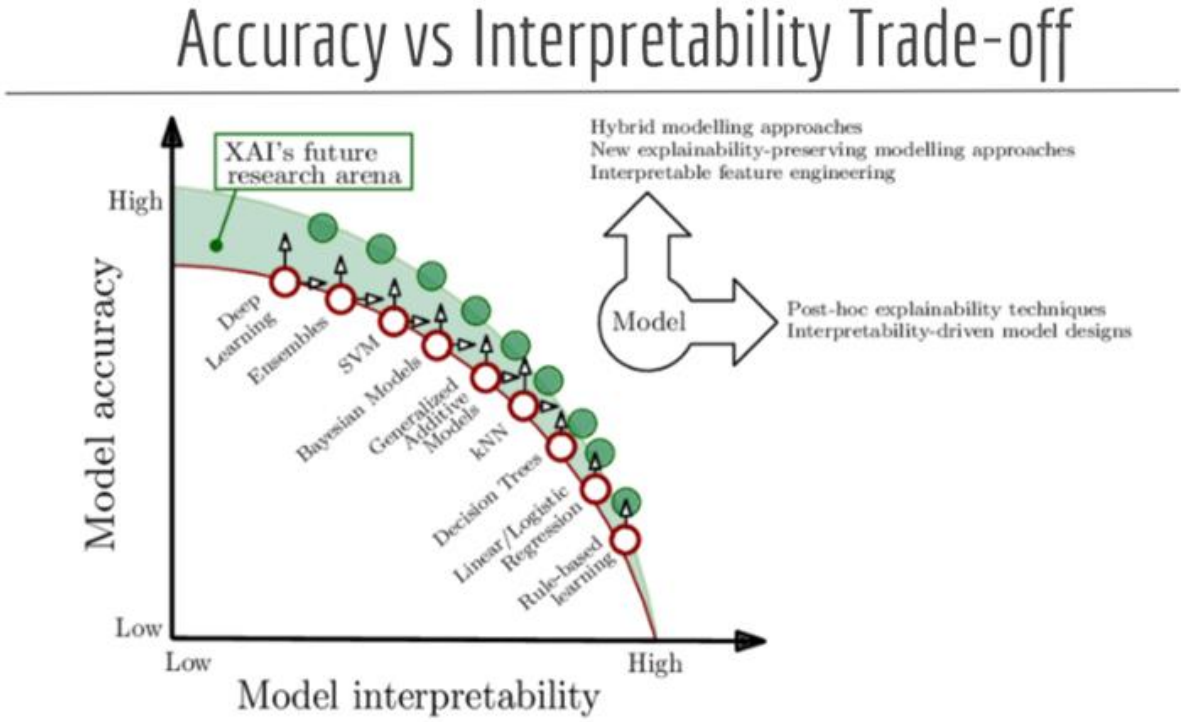
- **Being able to interpret a model increases trust in a machine learning model.**
- **To detect if there is any bias present in the model**
- **Model Explainability is critical for getting models to vet by regulatory authorities**

Glass Box Models

- Simple
- Interpretable
- Low Accuracy
- Example - Linear Models

Black Box Models

- Complex
- Not easy to Interpretable
- High Accuracy
- Example - Random Forest, Deep Learning

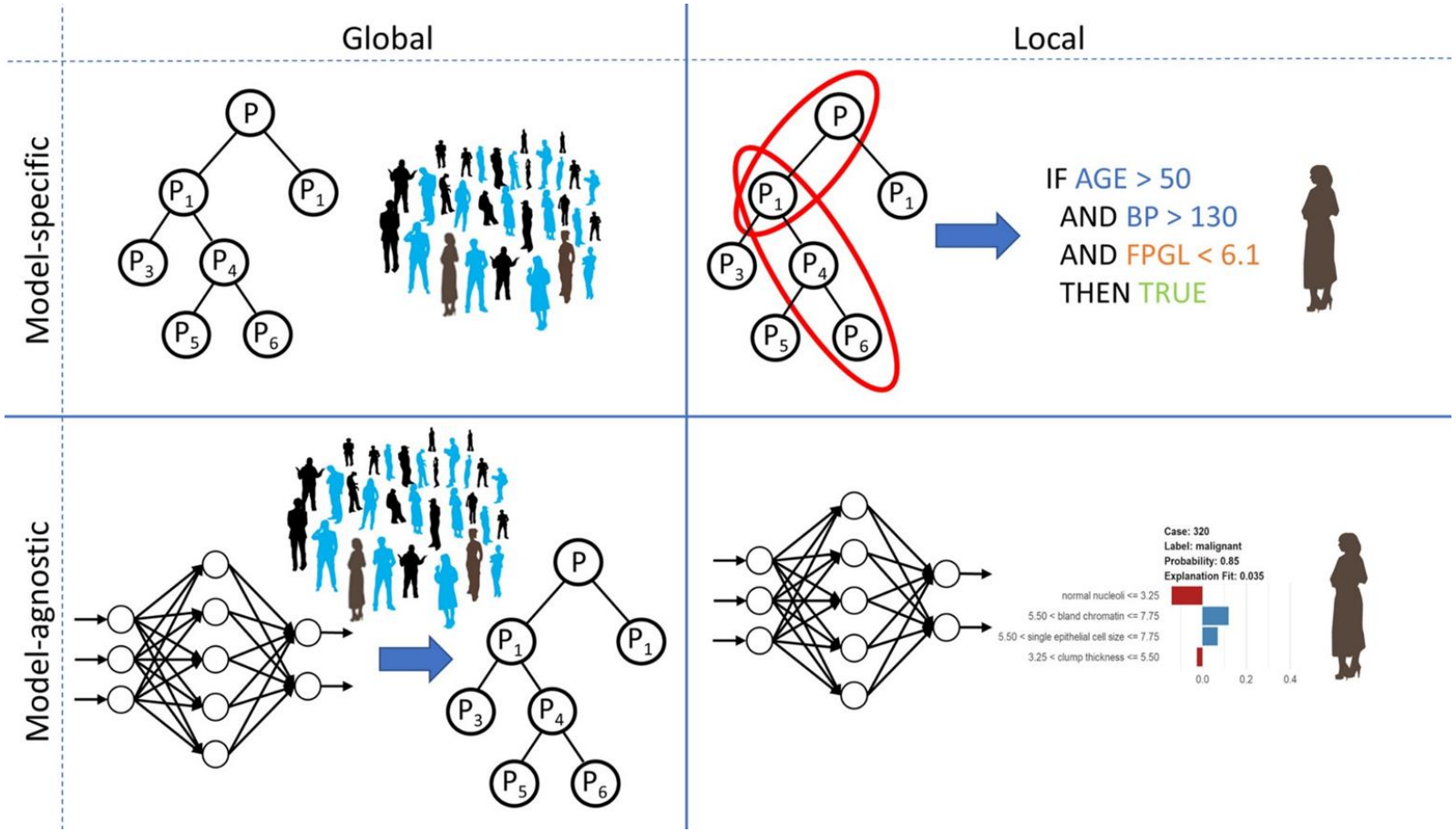


Approaches to Explainability

- **Globally:**
  - **Overall explanation of model behavior.**
  - **How features in the data collectively affect the result.**
- **Locally:**
  - **Tells us about each instance and feature in the data individually**
  - **How features individually affect the result.**

Model Specific ones rely on a certain model structure

Model Agnostic techniques work for any kind of ML models



Explainability Techniques