Interpretability
Why?

Domains

Medical

Legal

Loans

AI Research

Uses

Debugging

Trust in AI

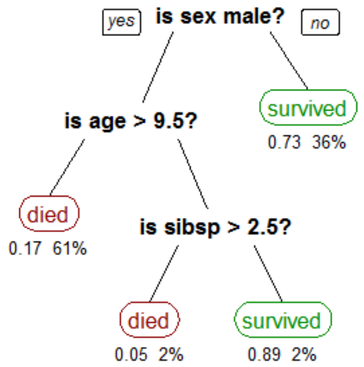
Fairness

Oversight

Safety + Security

Interpretability
2 Paradigms

① Model-based
~ make the model itself interpretable



② Prediction-based
~ explain *specific predictions*