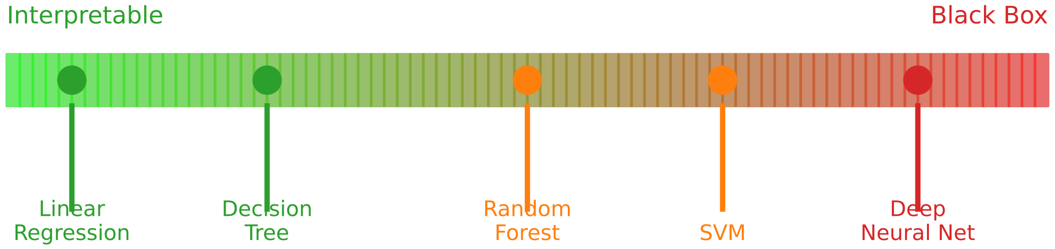
The Interpretability Challenge in Neural Networks

Model Interpretability Spectrum



Why Interpretability Matters:

1. Trust: Users need to understand predictions

2. Debugging: Find and fix errors

3. Compliance: Regulations (GDPR, finance)

4. Fairness: Detect bias in decisions

5. Science: Gain insights from models

Making NNs More Interpretable

Feature Importance	Which inputs matter most?
SHAP Values	Attribution per feature
LIME	Local linear approximations
Attention Visualization	What the model "looks at"
Gradient-based Methods	Sensitivity analysis

Trade-off: More interpretable models often have lower accuracy
Choose based on application requirements!