Surrogate Models

Approximating Black-Box with Interpretable Models

Surrogate Model Process

BLACK BOX

Complex Model

Neural Network, Ensemble, etc.

Generate predictions

Collect Predictions on Data

Use complex model to predict all instances

Train surrogate

WHITE BOX

Surrogate Model

Decision Tree, Linear Model, Rules

Evaluate &
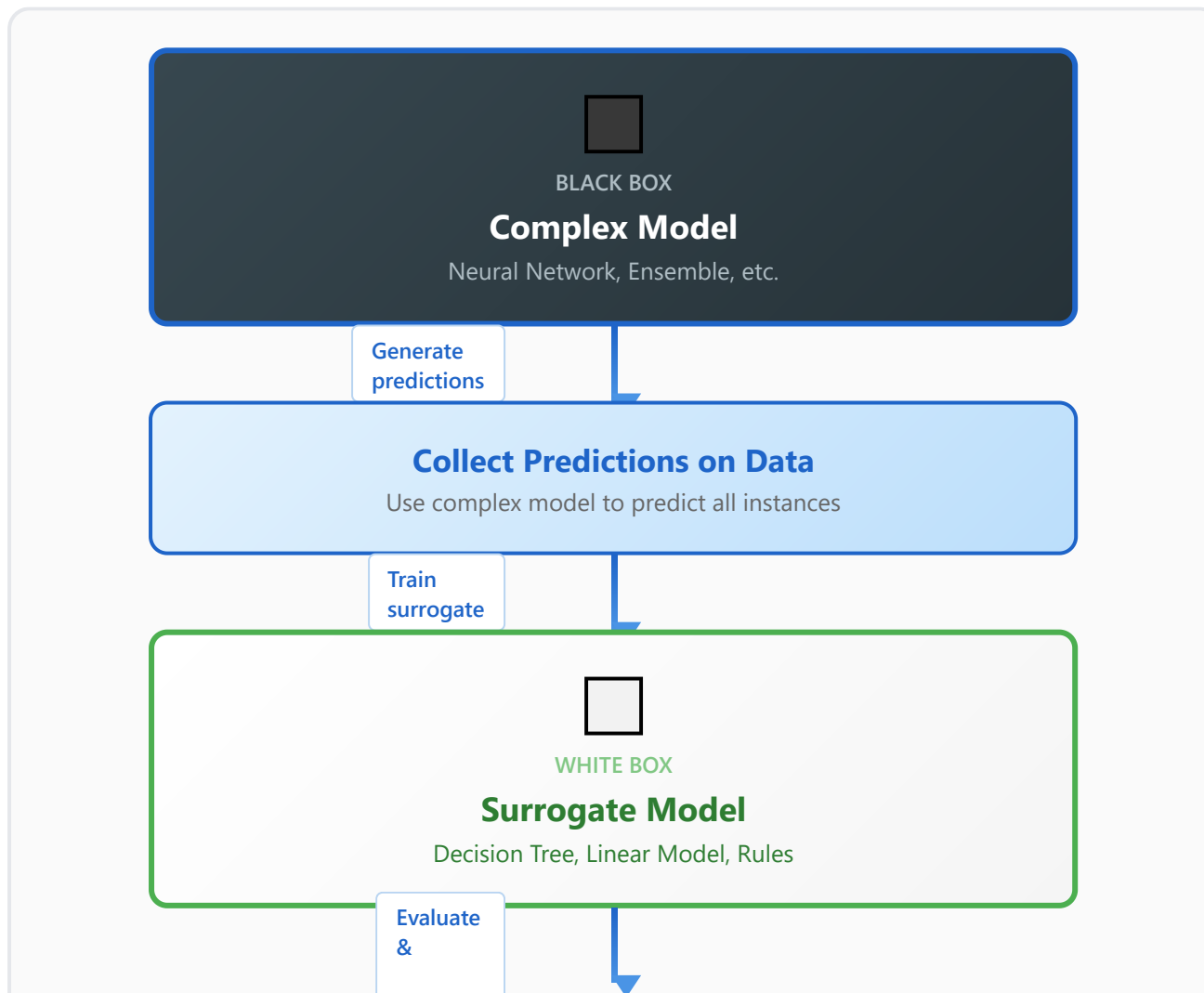
Key Concepts

Types

Global Surrogate
Mimic entire model behavior

Local Surrogate
Approximate specific region

Fidelity
How well surrogate matches original

Advantages
Leverage simple model interpretability

Limitations
May not capture full complexity

Use Cases
Compression, explanation, debugging

Common Surrogates

Interpret Surrogate

Extract insights and explanations

- Decision Trees
- Linear Models
- Rule Sets
- Generalized Additive Models