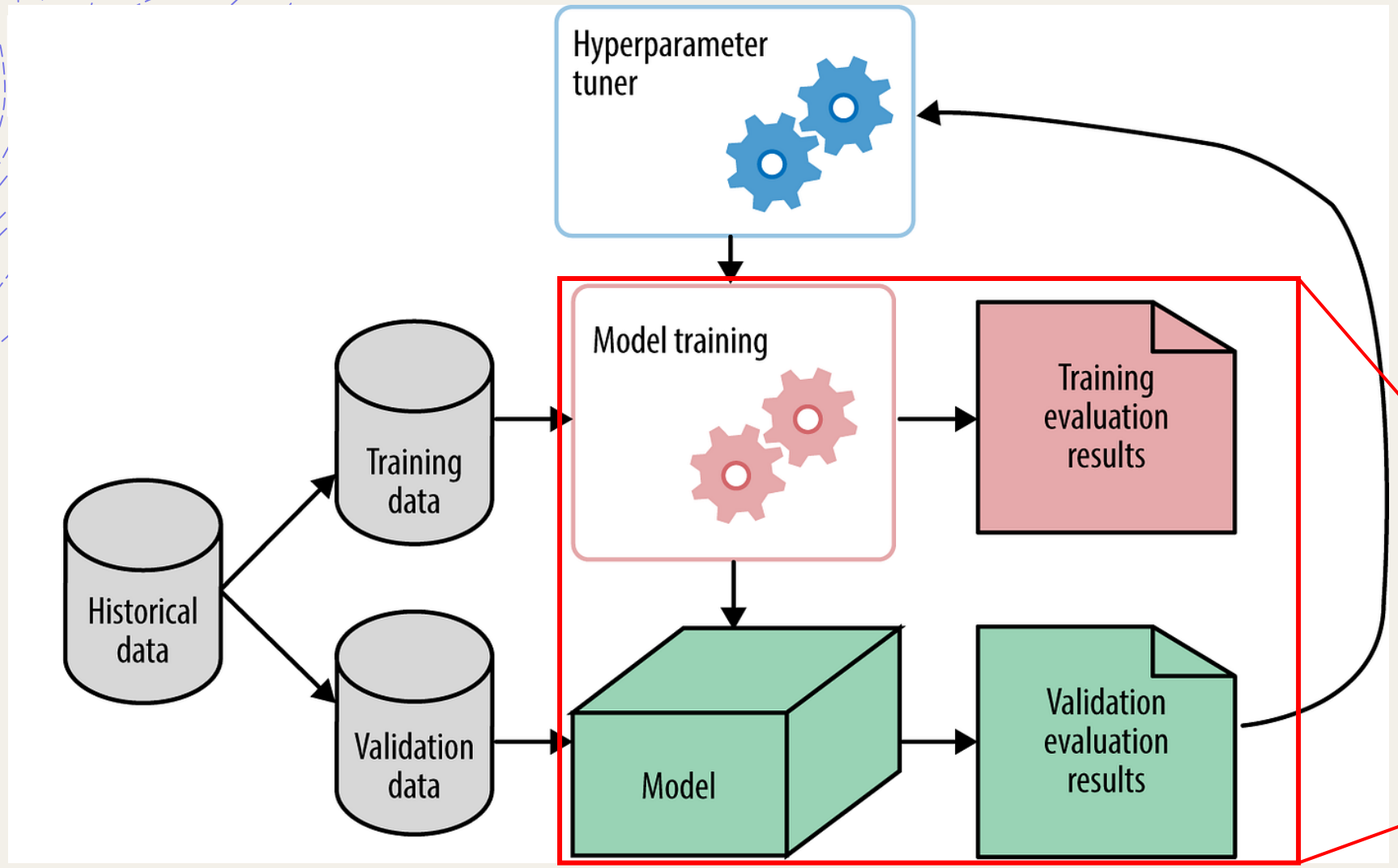
Model Interpretability and Explanation Techniques

Anthony Onoja, PhD

donmaston09@gmail.com

Understanding Machine Learning Pipeline



- *Dataset*
- *Research question: Aim, Objectives, methodological flow*
- *Algorithms e.g., logistic reg.*
- *Model training*
- *Hyperparameter tuning*
- *Model evaluations*
- *Deployment*

Black box? Model Interpretations

Elements of a Machine Learning Model

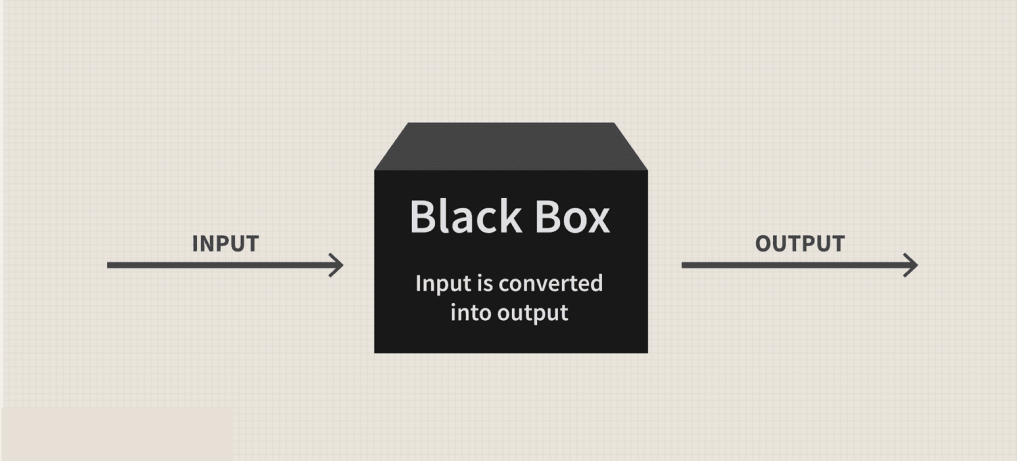
Model Interpretation, what does it mean?

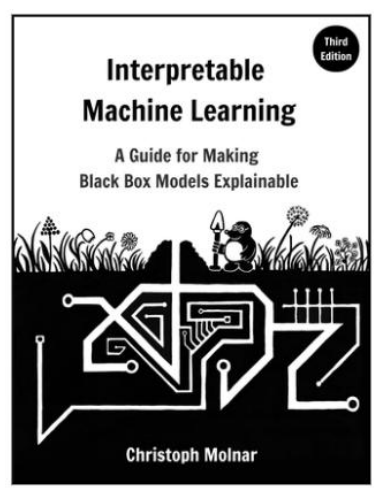
Model Interpretations: *the process and methods used to understand and explain the decisions or predictions made by a machine learning model.*

It aims to answer questions like:

- ❑ **How does the model arrive at its predictions?** What internal logic or patterns has it learned?
- ❑ **Why did the model make a specific prediction for a particular input?** What were the driving factors for that individual outcome?
- ❑ Which input **features are most important** for the model's predictions overall?
- ❑ How does the value of a **specific feature influence** the model's output?

Model Interpretation: what, why





Christoph Molnar

Model interpretation seek to:

- ❑ **Make sense of the model's behavior:** Moving beyond just knowing the prediction accuracy to understanding the underlying reasoning.
- ❑ **Increasing transparency:** Especially important for complex "black-box" models (like deep neural networks or complex ensembles) whose internal workings aren't immediately obvious.
- ❑ **Providing insights:** Enabling humans (developers, users, regulators) to trust, debug, improve, and ensure the fairness of ML models.

Goals of Interpretability

Purpose: Build trust, ensure fairness, debug errors, comply with regulations (e.g., GDPR's "right to explanation"), and validate alignment with domain knowledge.

Interpretability is a Means, Not an End

The purpose of interpretability depends on your goals such as improve the model, justify the model and predictions and to discover new insights.

For Example;

Goal 1: Improve the Model

- ❑ *Evaluate performance first – interpretability helps when performance isn't enough.*
- ❑ *Detect and correct "Clever Hans" predictors. Ex: Model identifies wolves by snow in the background.*
- ❑ *Use to debug feature engineering or encoding errors.*
- ❑ *Boost feature engineering in competitions and production.*

Interpretability helps you find when the model cheats or misunderstands.

Goals of Interpretability

Goal 2: Justify Model and Predictions

Stakeholders need different explanations:

- ✓ **Creators**: developers, researchers
- ✓ **Executors**: decision-makers
- ✓ **Subjects**: individuals impacted by predictions
- ✓ **Auditors**: regulatory bodies

Real-world examples:

- ✓ Forecasting stocks → justify to why a stock crashes
- ✓ Loan rejection → individual needs **recourse**
- ✓ Medical device → needs regulator approval

You can't contest a decision you don't understand.

Goal 3: Discover Insights

Models are also used to learn relationships, not just predict outcomes.

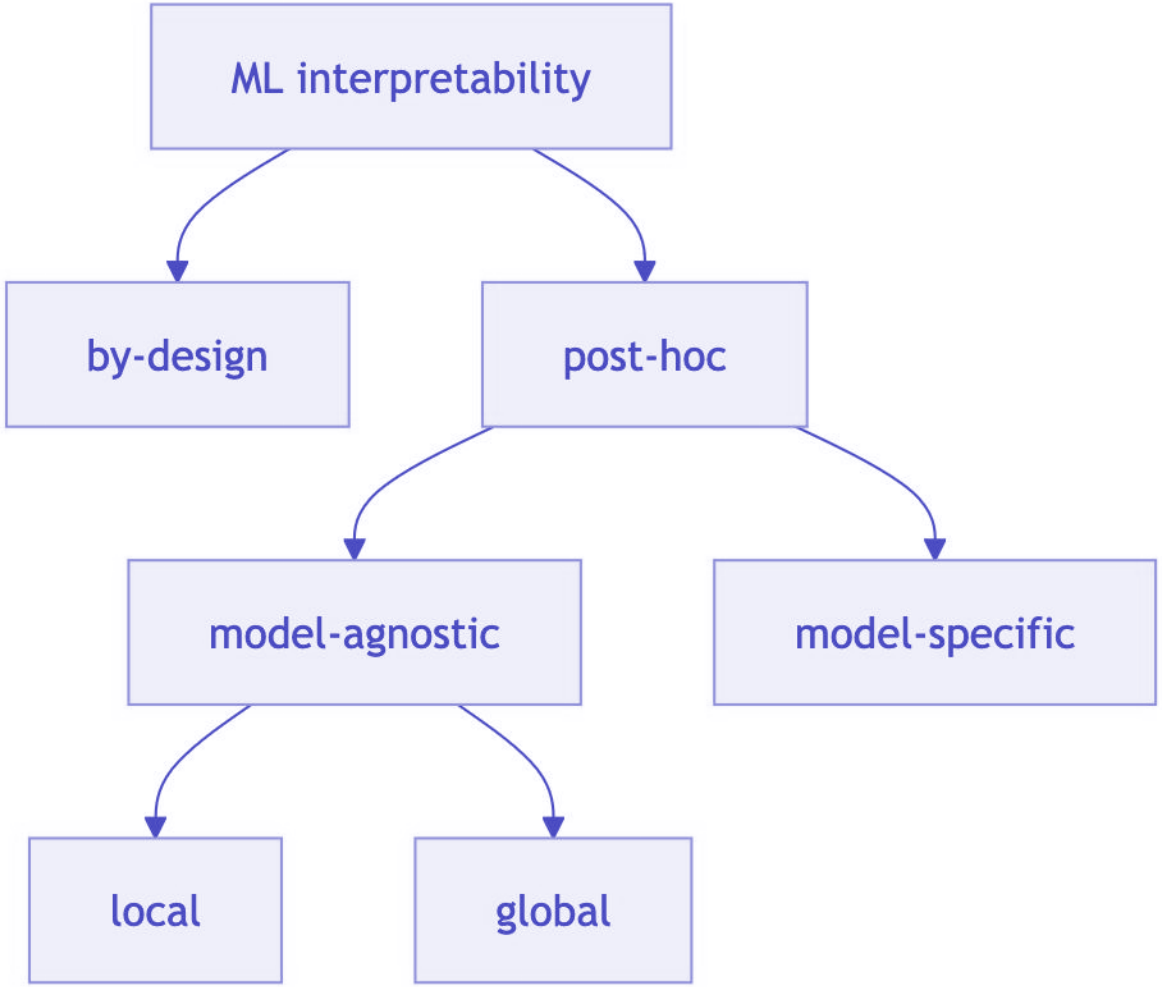
Understand feature impact:

- ✓ Churn prediction → target marketing efforts
- ✓ Agriculture → which fertilizers increase crop yield

Enables **scientific discovery** through machine learning

Prediction without understanding isn't enough for science or policy.

What Methods are used?



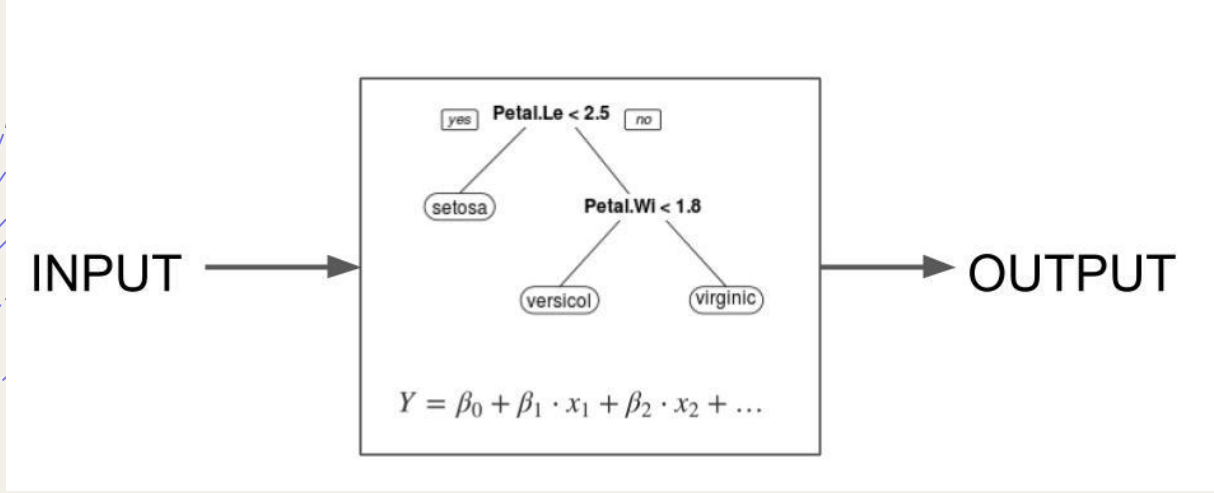
By design: *Models trained are inherently interpretable models, such as using logistic regression instead of a random forest.*

Post-hoc: *can be model-agnostic, such as permutation feature importance, or model-specific, e.g., analysing the features learned by a neural network.*

Model-agnostic: 1) *local methods which focus on explaining individual predictions, and;* ***2)*** *global methods which focus on datasets.*

Model-Specific: *Techniques like attention maps in transformers or saliency maps in CNNs.*

What is Interpretability by Design?



- ❑ It's not post-hoc. It's built into the algorithm.
- ❑ The model is constrained to be human-interpretable from the start.

Also known as:

- ❑ Intrinsic Interpretability;
- ❑ Inherently Interpretable Models.

Example: *Linear regression always outputs a linear function of inputs it's interpretable by design.*

Ex: Linear Reg., Logistic Reg., Decision Tree, Decision Rules, RuleFit, Linear Extensions.

- Easy to explain
- Easy to debug
- Good for regulation-heavy industries (e.g., medicine)

Advanced & Research-Based Interpretable Models

- ❑ ***ProtoViT:*** *Prototype-based neural networks*
- ❑ ***Interpretable Boosted Trees (Yang et al., 2024):*** *Shallow depth + main/interactions + pruning*
- ❑ ***Model-Based Boosting (Bühlmann & Hothorn, 2007)***
- ❑ ***GA²Ms (Caruana et al., 2015):*** *GAMs with automatic interaction detection.*

These models are inherently interpretable and often competitive in performance.

Scope of Interpretability

Scope	What it means
Entire Model	Fully interpretable (e.g., small trees, sparse linear models)
Parts of Model	Interpret individual pieces (e.g., coefficients, rules)
Predictions	Explain individual predictions (e.g., k-NN via nearest examples)

Always assess interpretability scope when choosing a method.

When to Use Interpretable-by-Design Models

- ❑ **Model Improvement:** Easier to debug and adjust.
- ❑ **Justification:** Transparency for stakeholders & domain experts.

Insight Discovery:

- ❑ Good for understanding models
- ❑ Trickier for data insights (requires assumptions!)

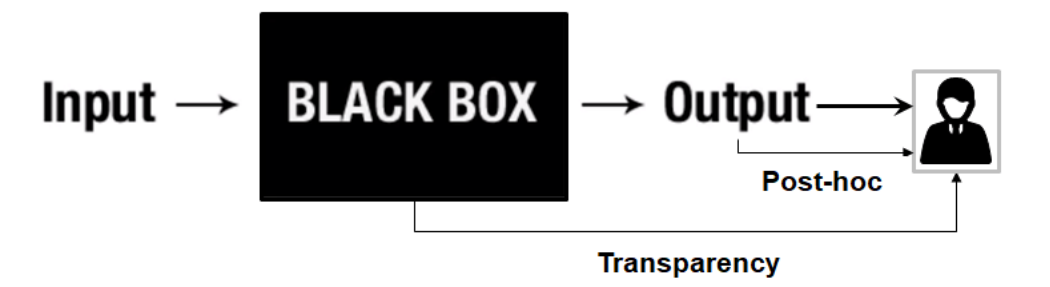
The Rashomon Effect

Multiple models may explain the same data equally well — but differently.

- ✓ Which model should we trust?
- ✓ Which one truly reflects the data-generating process?
- ✓ This makes interpretability more nuanced and philosophical.

What is Post-hoc Interpretability?

*"After-the-fact" interpretation — applied **after the model is trained.***



Can be:

- ❑ **Model-agnostic**: Treat model like a black box.
- ❑ **Model-specific**: Use internal model structure.

Goal: Understand predictions **without changing the model**.

Model-Agnostic Post-hoc Methods

- ✓ We don’t look inside the model — just observe input/output behavior.
- ✓ Flexible
- ✓ Works with any model
- ✓ Decouples modeling from interpretation

SIPA principle (Scholbeck et al. 2020): Sample → Intervene → Predict → Aggregate

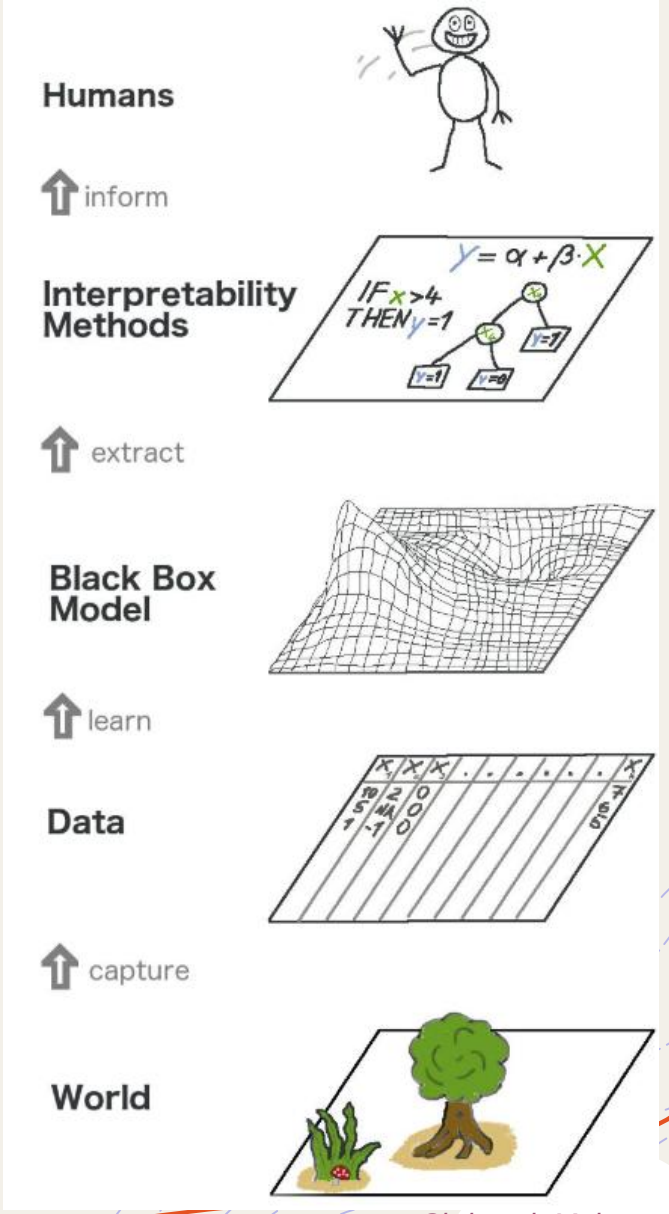
Local Model-Agnostic Methods

Focus: Explain individual predictions (zoomed in)

Method	What it does
LIME	Fit a simple local model
SHAP	Fairly attribute prediction to features
Ceteris Paribus	Vary one feature, hold others constant
ICE Plots	How prediction changes for many instances
Anchors	If-then rules that “lock” predictions
Counterfactuals	What needs to change for a different outcome?

Best for:

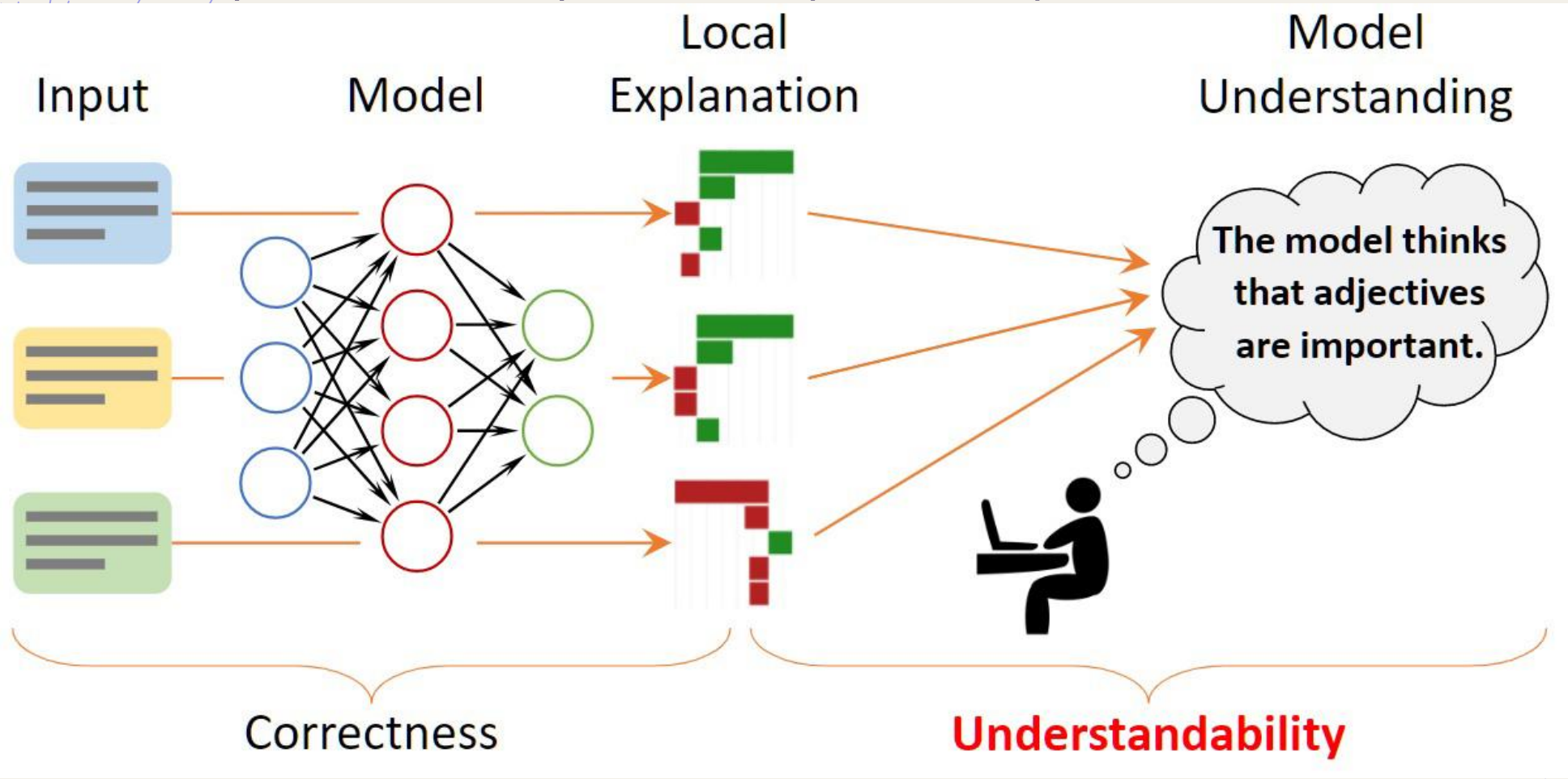
- ✓ Debugging edge cases,
- ✓ Explaining individual decisions,
- ✓ Discovering data issues



Christoph Molnar

Local Model-Agnostic Methods

Focus: Explain individual predictions (zoomed in)



Attribution (e.g., SHAP) = feature contributions, Others (e.g., CP, ICE) = sensitivity to input. ***Caveat:*** *Attribution methods (SHAP, LIME) are models too not always justifiable in high-stakes domains (Rudin, 2019).*

Global Model-Agnostic Methods

Focus: Understand average behavior across the dataset

Method	Purpose
PDP (Partial Dependence)	Average effect of a feature
ALE (Accumulated Local Effects)	Same as PDP but better with correlated features
Feature Interaction (H-stat)	Quantifies joint effects
PFI (Permutation Feature Importance)	Measures drop in performance when permuted
LOFO	Re-train model without feature
Surrogate Models	Simpler models approximating black-boxes
Prototypes & Criticisms	Representative examples of data

Use when:

- ❑ Improving models;
- ❑ Explaining to stakeholders;
- ❑ Group-wise analysis (e.g., apply to subgroups).

- ✓ **Two types:** Feature effects: PDP, ALE, H-stat, Decomposition.
- ✓ **Feature importance:** PFI, LOFO, SHAP importance.

Model-Agnostic vs. Model-Specific:

- ❑ Model-Agnostic: Methods like LIME or SHAP that work on any model.
- ❑ Model-Specific: Techniques like attention maps in transformers or saliency maps in CNNs.

Surrogate Models: Training a simpler model (e.g., linear regression) to approximate the black-box model's behavior.

Challenges:

- ✓ Approximations may oversimplify complex model behavior.
- ✓ Risk of generating misleading explanations if the method is misapplied.

The human-understandable outputs generated by interpretation techniques, often through visual, textual, or numerical summaries.

Importance in Practice and Key Challenges

The end result communicated to stakeholders (e.g., users, regulators).

Importance in Practice

- ❑ **Ethics & Fairness:** Detect biases (e.g., a model unfairly penalizing certain demographics).
- ❑ **Regulatory Compliance:** Mandatory in sectors like healthcare (e.g., FDA approval) and finance.
- ❑ **User Trust:** Patients, customers, or engineers are more likely to adopt AI systems if they understand their logic.

Key Challenges

- ❑ **Trade-offs:** Complex models often sacrifice interpretability for accuracy.
- ❑ **Faithfulness:** Ensuring explanations accurately reflect the model's true reasoning.
- ❑ **Scalability:** Interpretation methods must keep pace with growing model complexity (e.g., LLMs).