Techniques for Interpretable Machine Learning

“Uncovering the mysterious ways machine learning models make decisions”

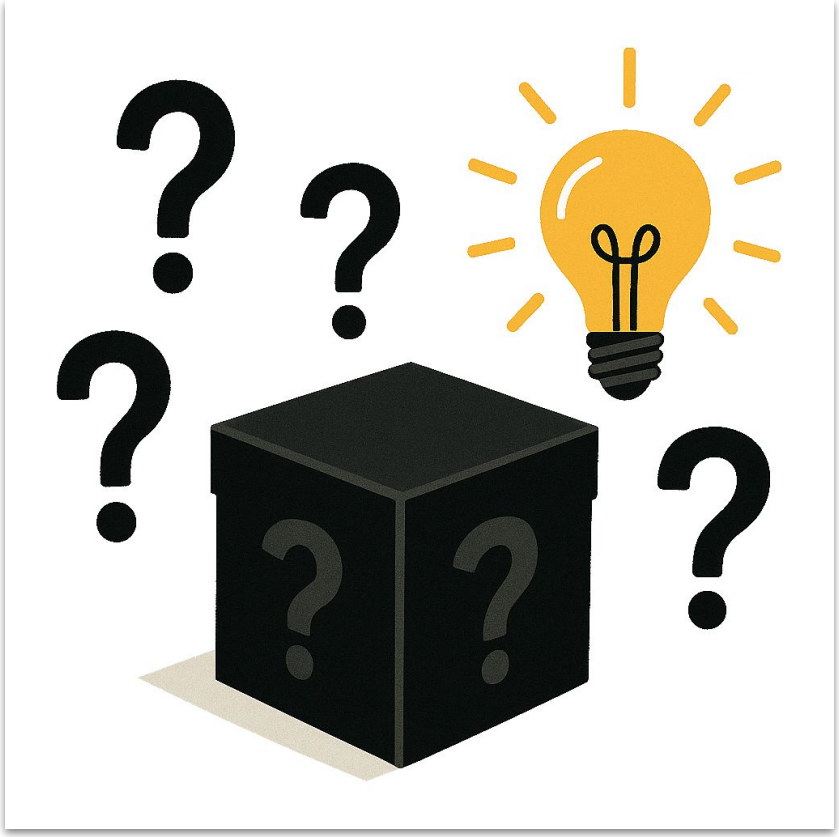
Saverio Mattia **Merenda**
saveriomattia.merenda@studenti.unipr.it

Big Data & Data Mining

[2024-2025]

We Built It, But Do We Understand It?

- ML models are powerful tools for prediction.
- They act like a "**black box**"
- This could be a problem, especially when decisions have real-world consequences.



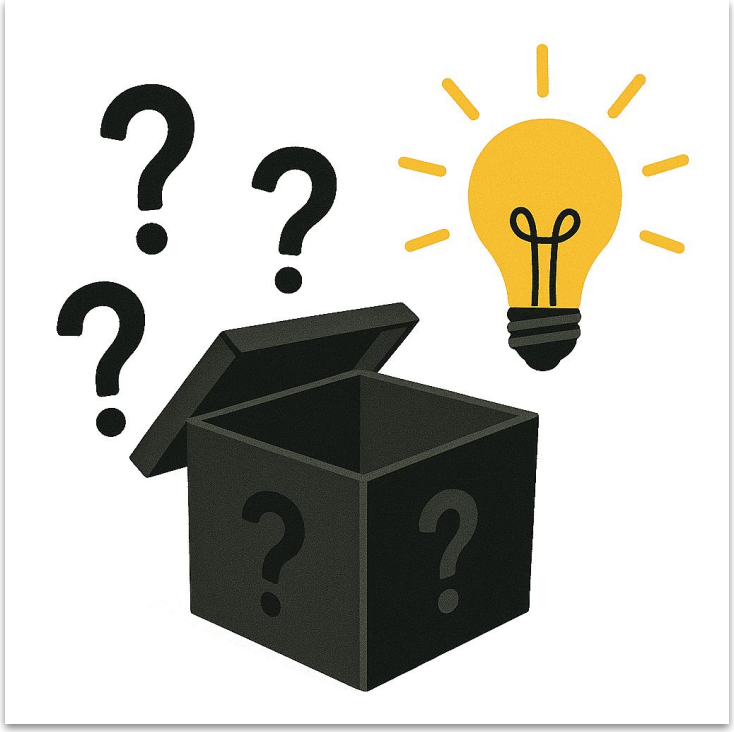
Why Should We Care About Understanding?

- **Trust**
  - Building confidence in AI decisions (e.g., in medical diagnosis).
- **Debugging**
  - Identifying and fixing errors or biases in models (e.g., a self-driving car making a wrong turn).
- **Fairness**
  - Ensuring models don't discriminate against certain groups (e.g., loan applications).
- **Insights**
  - Discovering new knowledge and relationships in the data (e.g., understanding factors affecting customer behavior).
- **Regulation**
  - Meeting legal and ethical requirements for transparency.

What Exactly is *Interpretability*?

*"Interpretability is the ability to **understand why** a machine learning model makes a certain prediction"*

- It's about making the model's reasoning understandable to humans.
- **Think of it as**: Opening up the black box and seeing how the gears turn.



Different Ways to Look at Interpretability

- **The Entire Model**: Understanding the overall logic and behavior of the model.
  - *"This model relies heavily on these three factors"*

- **Parts of the Model**: Understanding how individual components or features contribute.
  - *"This specific feature has a positive impact on the prediction"*

- **Individual Predictions**: Understanding why the model made a specific decision for a particular input.
  - *"This customer was denied a loan because their income was below a certain threshold"*

How Do We Achieve Understanding?

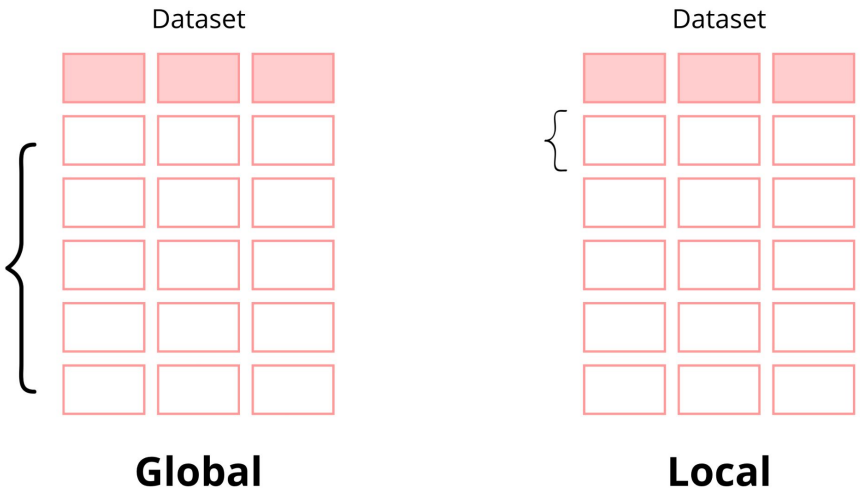
- **Intrinsic Interpretability**
  - Interpretability by Design.
  - Building models that are inherently easy to understand from the start.
  - *“Like a transparent machine where you can see all the parts working”*

- **Post-hoc Interpretability**
  - Explaining Black Box Models.
  - Using techniques to understand models after they have been trained.
  - *“Like using tools to analyze a closed box and figure out what's inside”*

Two Levels of Interpretability

- **Global Interpretability**
  - Allows users to understand the overall functioning of the model by examining its structures and parameters.

- **Local Interpretability**
  - Examines an individual model prediction, trying to understand why that particular decision was made.

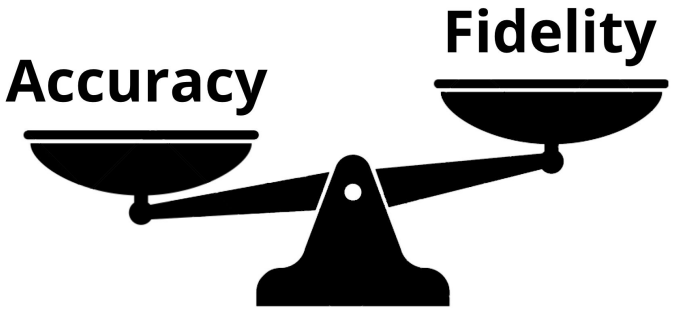


Choose the Best Trade-Off

*"The trade-off between these categories lies in the balance between model **accuracy** and explanation **fidelity**"*

- **Interpretable Models**
  - Provide accurate explanations.
  - May sacrifice some predictive performance.

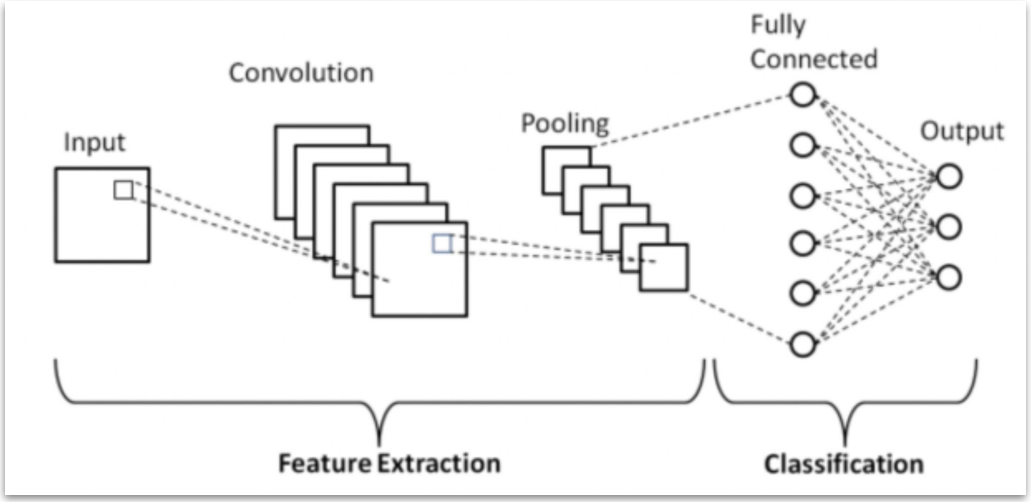
- **Post-hoc Models**
  - Keep the accuracy of the underlying model intact.
  - They are limited in their approximate nature.



Globally Interpretable Models (1/2)

- **Adding Interpretability Constraints**
  - *Sparsity*: the model is encouraged to use fewer features.
  - *Semantic monotonicity*: features have monotonic relationships with the prediction.

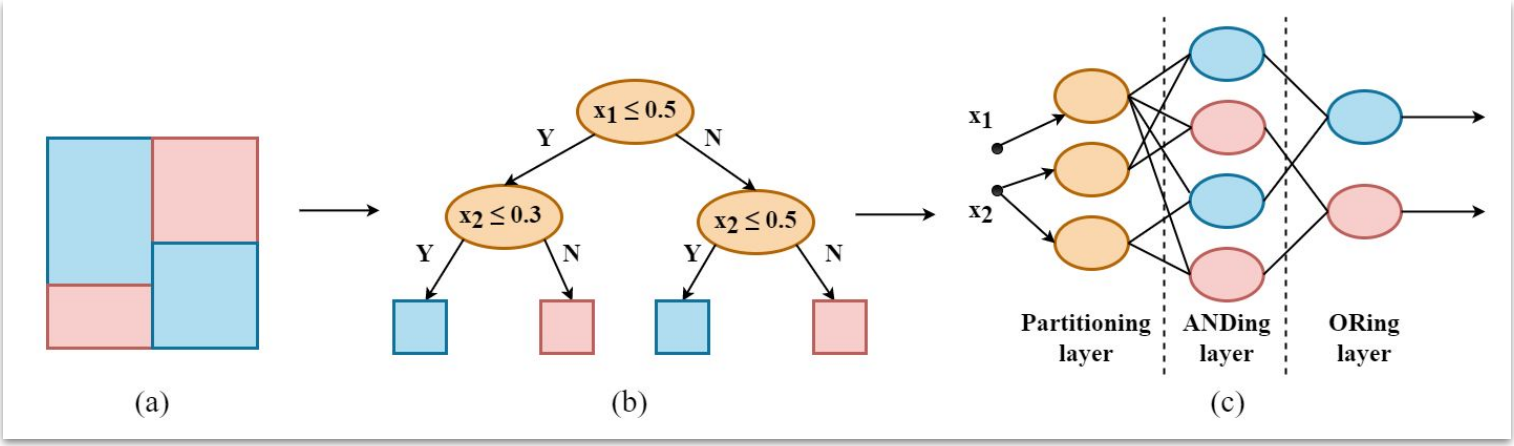
Example: interpretable CNNs add regularization loss to the upper layers to learn disjoint representations, resulting in filters that can detect semantically meaningful natural objects.



Globally Interpretable Models (2/2)

- **Interpretable Pattern Extraction**
  - Approximating a complex pattern with an easily interpretable one such as a decision tree.

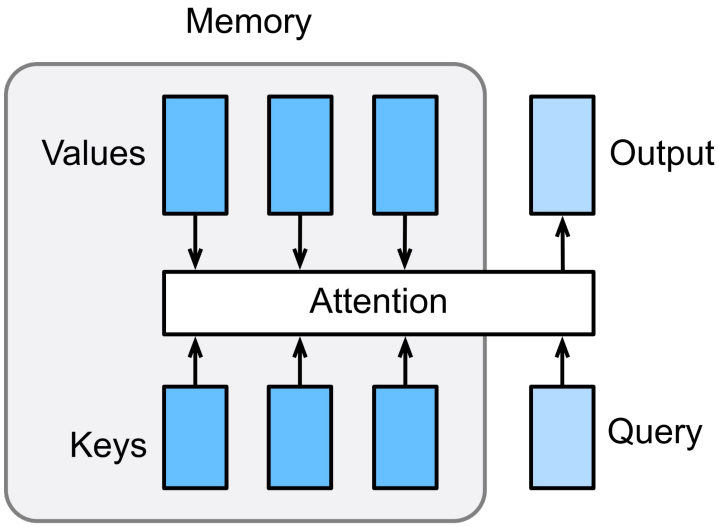
Example: training a DNN to mimic a decision tree.



Locally Interpretable Models

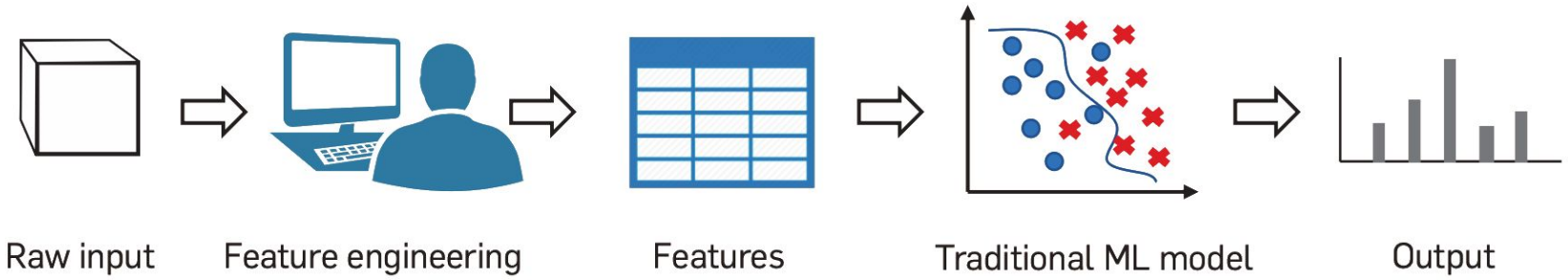
*"Obtained by designing architectures that **can explain** individual decisions"*

- **Attention Mechanism**: It is used to explain the predictions of sequential models such as RNNs.
  - It allows to visualize which parts of the input were considered by the model for each prediction.



Post-Hoc Global Explanation

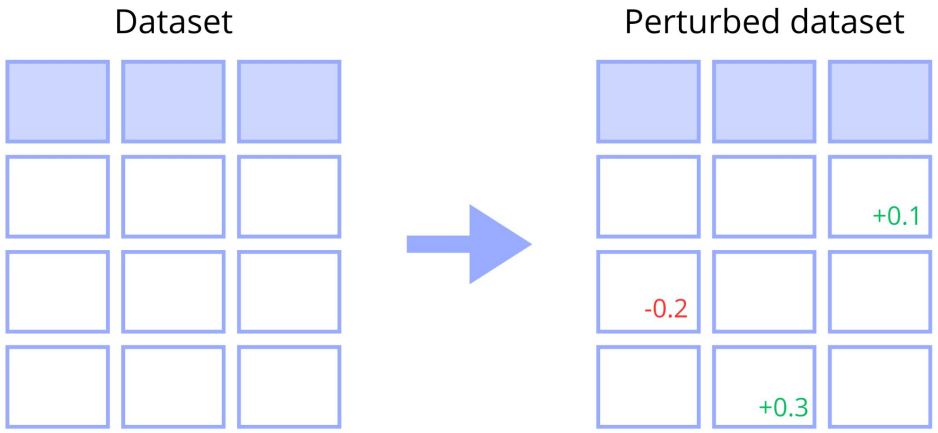
- **Traditional Machine Learning**
  - Based on *feature engineering*, where features are generally interpretable.
  - *Feature importance*: indicates the statistical contribution of each feature to the model.



Post-Hoc Local Explanation (1/2)

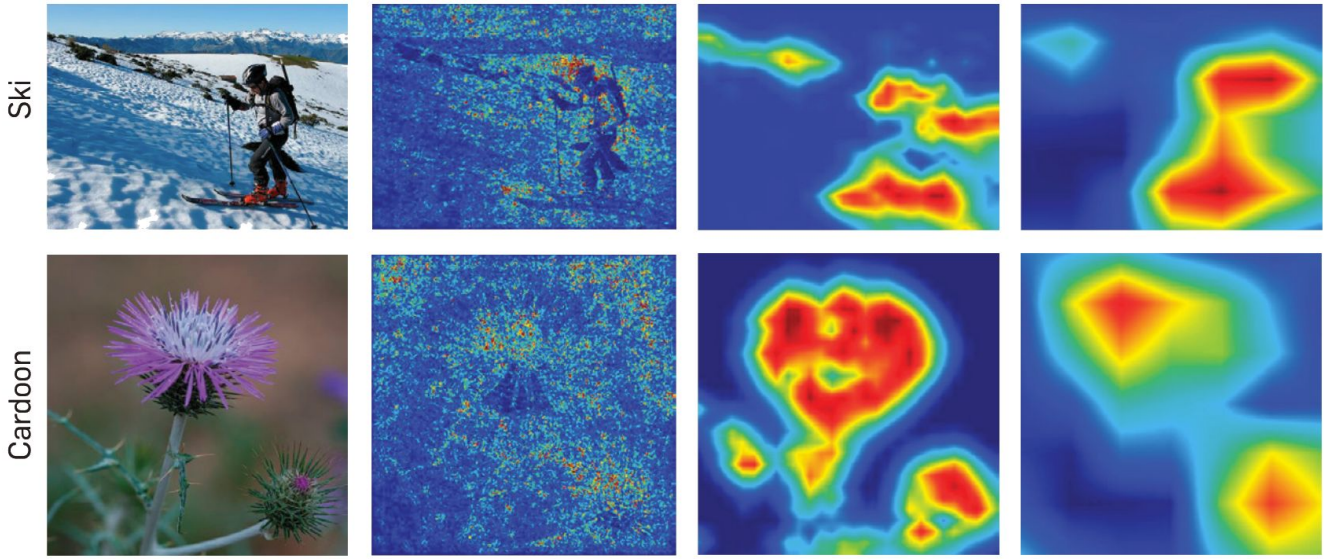
- **Model-Agnostic Methods**
  - Based on *local approximation*
    - Approximation with white-box models.
  - Based on *perturbation*
    - Omission or occlusion of features.
    - Measurement of change in prediction.

Example: LIME perturbs the input data by making small adjustments to the input features and observes how these changes influence the model's predictions.



Post-Hoc Local Explanation (2/2)

- **Model-Specific Methods for DNN**
  - Based on back-propagation.
  - Mask perturbation
    - Optimization framework.
    - Neural networks to predict attribution mask.
  - Investigation of deep representations



Real Applications

- **Model Validation**
  - Identification of bias in training data.
  - Verification of the use of real-world evidence.

- **Debugging of the Model**
  - Analysis of misbehavior.
  - Example: adversary learning.

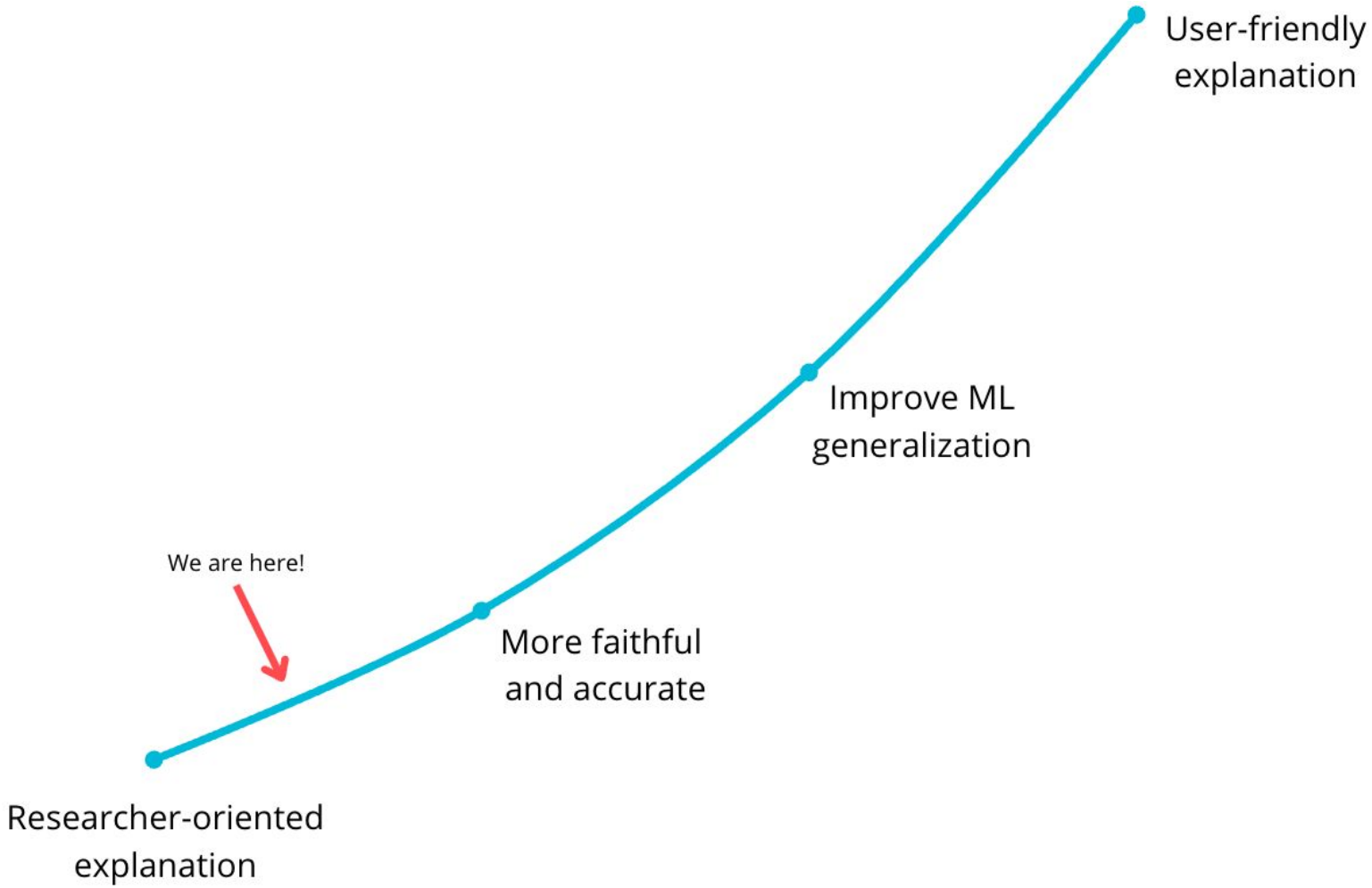
- **Knowledge Discovery**
  - Extraction of new scientific insights.
  - Example: the case of asthma and pneumonia.

Towards User-Friendly Explanation

- **Current limitations**
  - Explanations based on researchers' intuition.
  - Vectors of feature importance: too technical!

- **Future directions**
  - Contrastive explanations
    - *"Why Q instead of R?"*
    - Real vs. virtual examples.
  - Selective explanations
    - Minimum set of relevant features.
  - Credible explanations
    - Consistent with prior knowledge.
  - Conversational explanations
    - Adapted to the social context.

Progress of Interpretable ML



Things to Keep in Mind

- **Complexity**
  - Some interpretability techniques can be complex to understand and implement themselves.
- **Fidelity**
  - Post-hoc explanations are approximations and might not perfectly reflect the model's true behavior.
- **Stability**
  - Some local explanation methods can be sensitive to small changes in the input data.
- **Causality**
  - Interpretability doesn't always imply causality (correlation is not causation!).

Let’s Recap!

Method	Type	Level	Advantages	Limitations
Intrinsically Interpretable Models	Intrinsic	Global	Full transparency, accurate explanations	Potential performance sacrifice
Interpretability Constraints	Intrinsic	Global	Sparsity and semantic monotonicity	Implementation complexity
Pattern Extraction	Intrinsic	Global	Simplification into understandable structures	Precision loss in simplification
Attention Mechanisms	Intrinsic	Local	Direct visualization of relevant parts	Limited to sequential models
Feature Importance	Post-hoc	Global	Simplicity, applicable to traditional ML	Doesn't capture complex interactions
Activation Maximization	Post-hoc	Global	Understanding of deep representations	Complex to interpret
LIME	Post-hoc	Local	Model-agnostic, perturbation-based	Local approximation, instability
Backpropagation-Based Methods	Post-hoc	Local	DNN-specific, precision	Limited to certain architectures

Techniques for Interpretable Machine Learning

“Explanations could help examine whether a machine learning model has employed the true evidences instead of biases that widely exist among training data”

Saverio Mattia **Merenda**
saveriomattia.merenda@studenti.unipr.it

Big Data & Data Mining

[2024-2025]

References

[1] Du, M., Liu, N., & Hu, X. (2020). Techniques for Interpretable Machine Learning. Communications of the ACM, 63(1), 68-77.

[2] What is AI interpretability? by IBM

[3] linkedin.com/what-difference-between-local-global-interpretability

[4] Overview of Convolutional Neural Networks

[5] I. K. Sethi, "Entropy nets: from decision trees to neural networks," in Proceedings of the IEEE, vol. 78, no. 10, pp. 1605-1613, Oct. 1990.

[6] Attention (machine learning)