ICISCoIS 2023

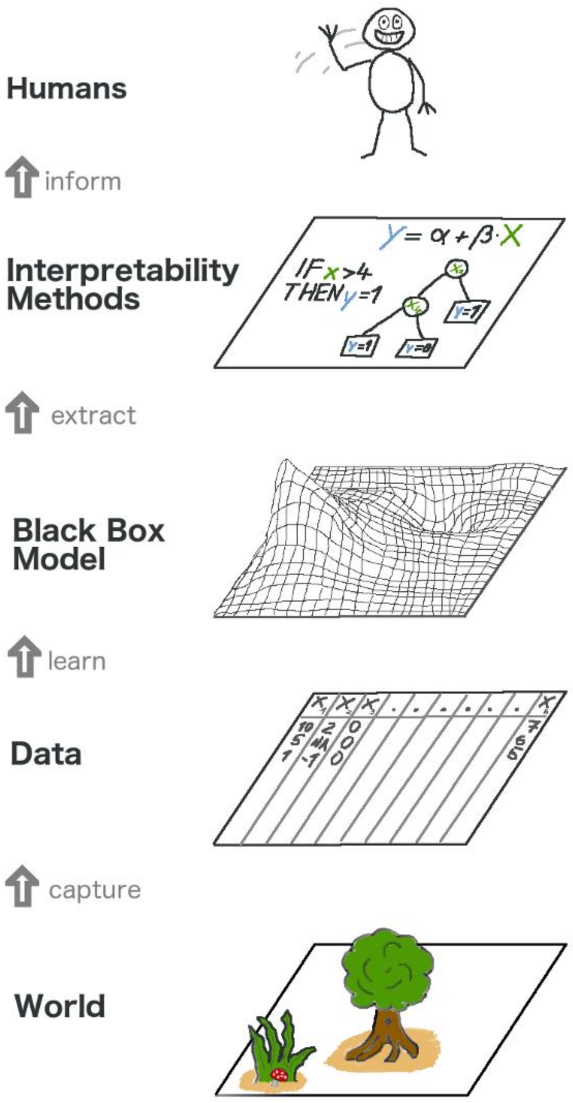
# Explainable AI in Finance

R. Shyaam Prasadh Ph.D.

Feb 11, 2023

# Agenda

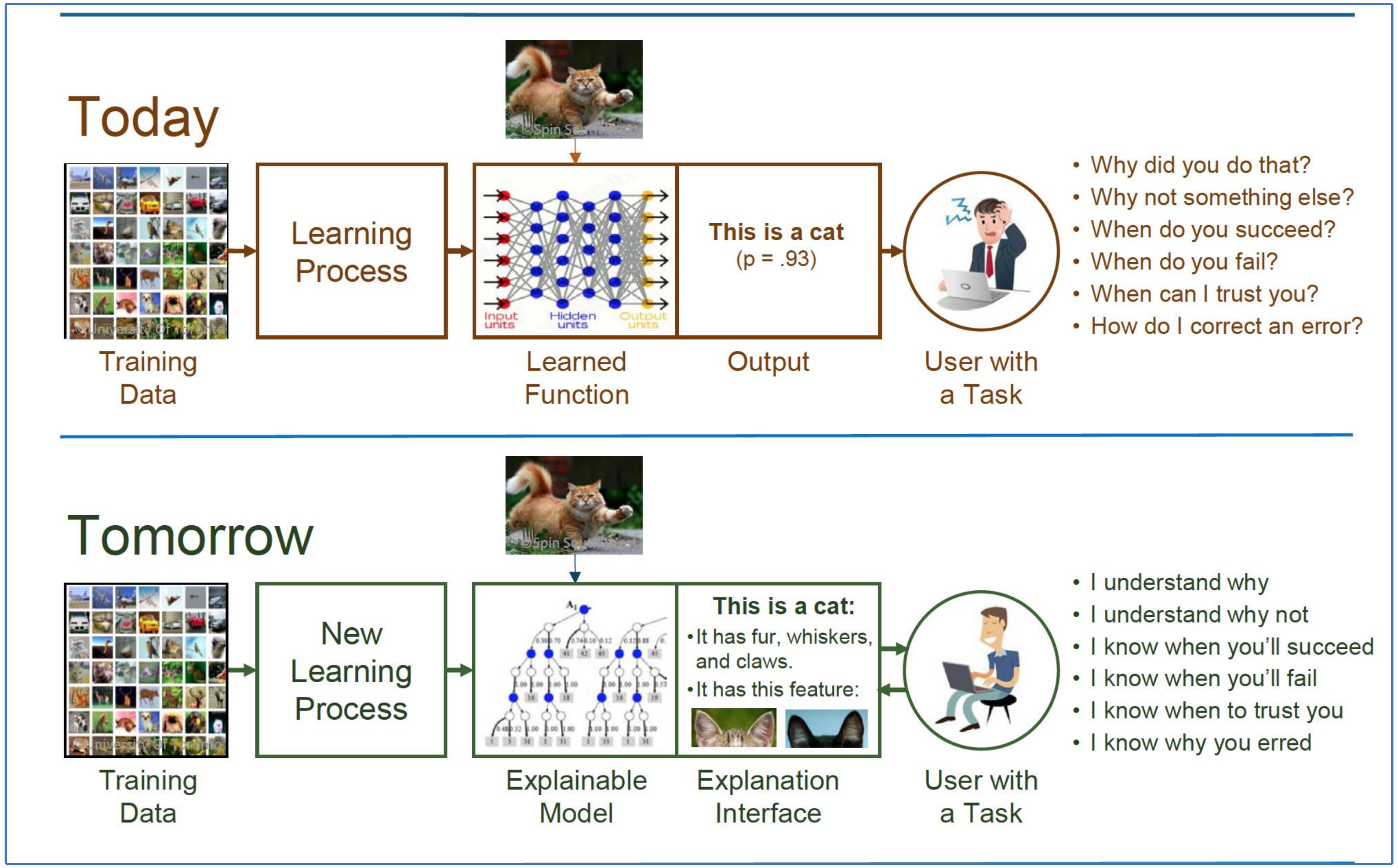
- Need for XAI
- Explainability: What, Why, What For and How?
- Interpretability Methods in Machine Learning
- Explainable AI Models in Finance




Author: Unknown

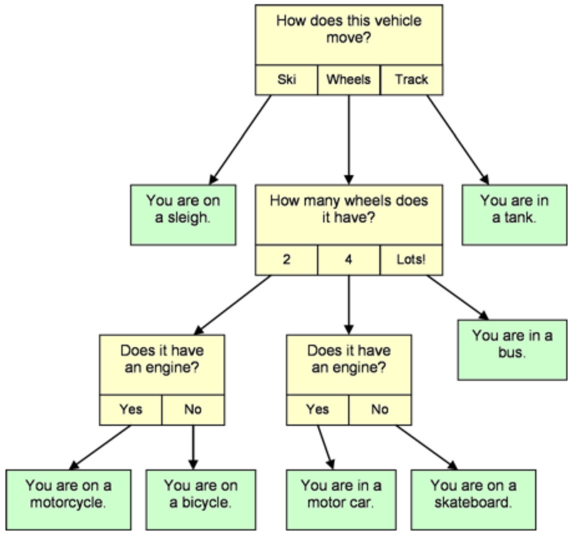
# Fear of the unknown

# What is the current state of art ?

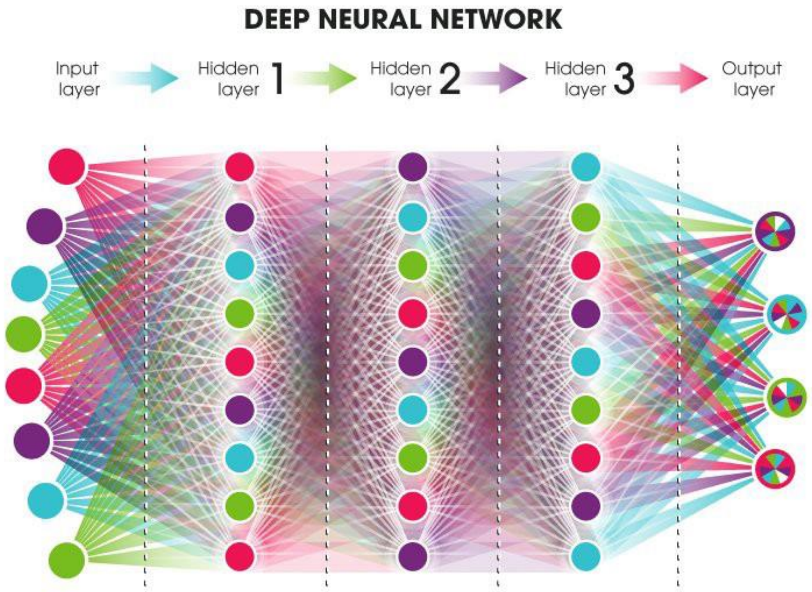



- Black-box statistical predictions are inadequate
- Explanations must be understandable to non-specialist

# Trade off




- OR -




neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

ICISCoIS 2023

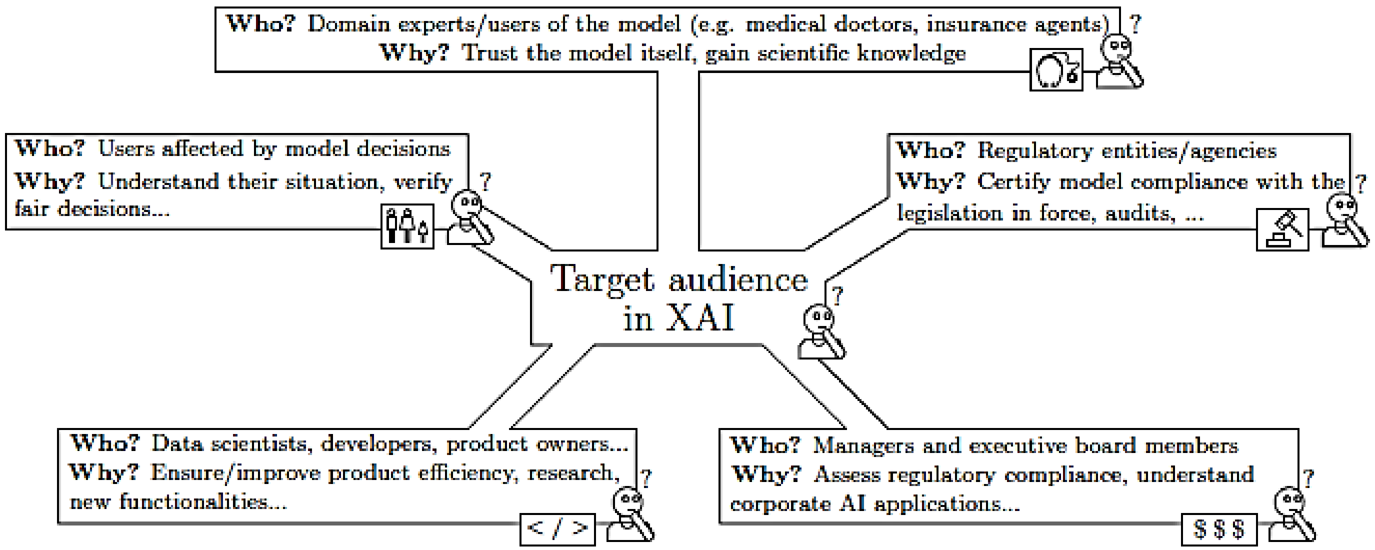
**Expert system:**
Good for explanations,
not so good for accuracy

**How do we get the best of both worlds?**

**Neural nets:**
Good for accuracy,
not so good for explanations

# What is Interpretability?

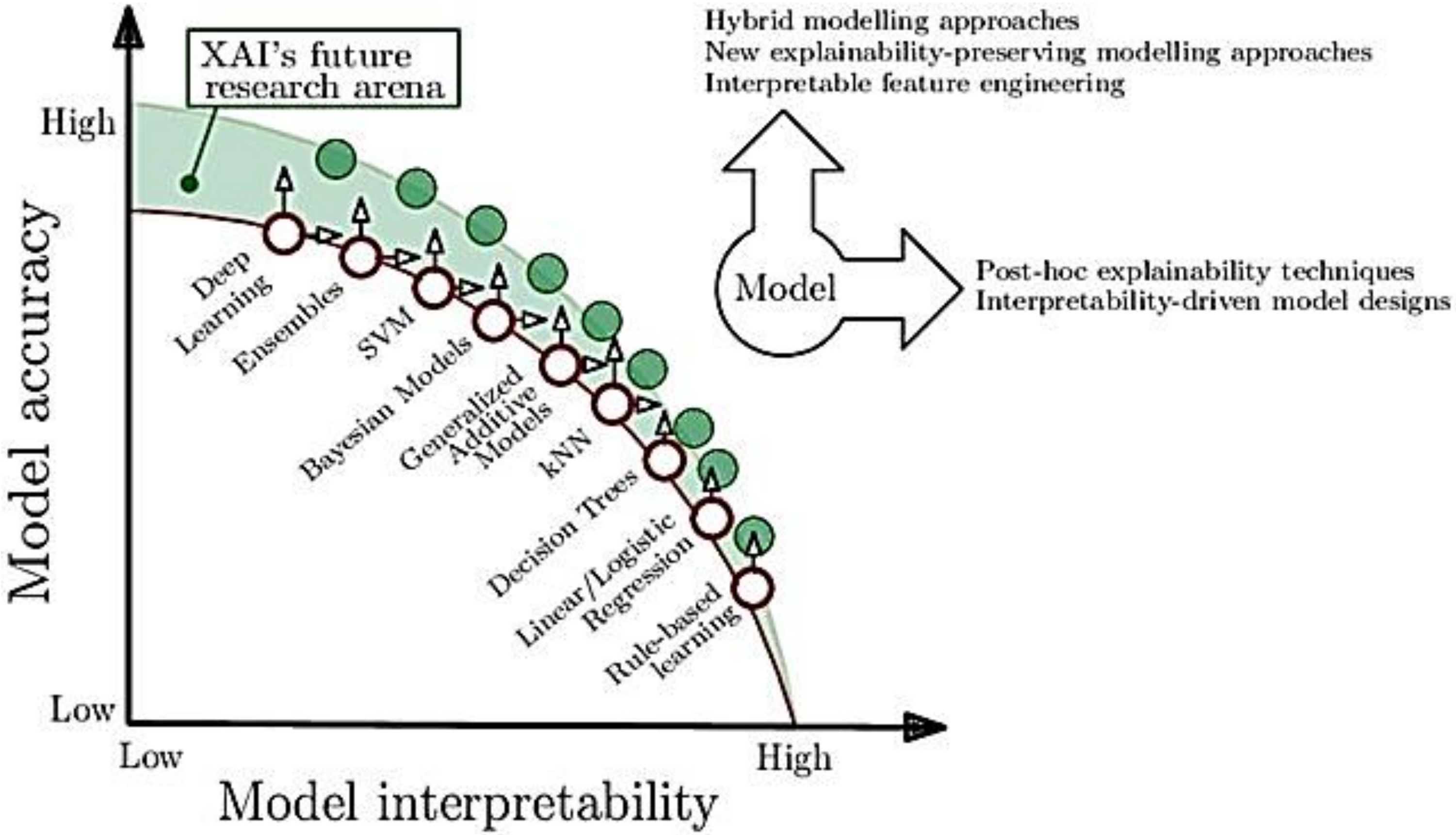
- Defn: Ability to explain or to present in understandable terms to a human




(Arrieta, Del Ser et al; 2019)

# Interpretability vs. Accuracy

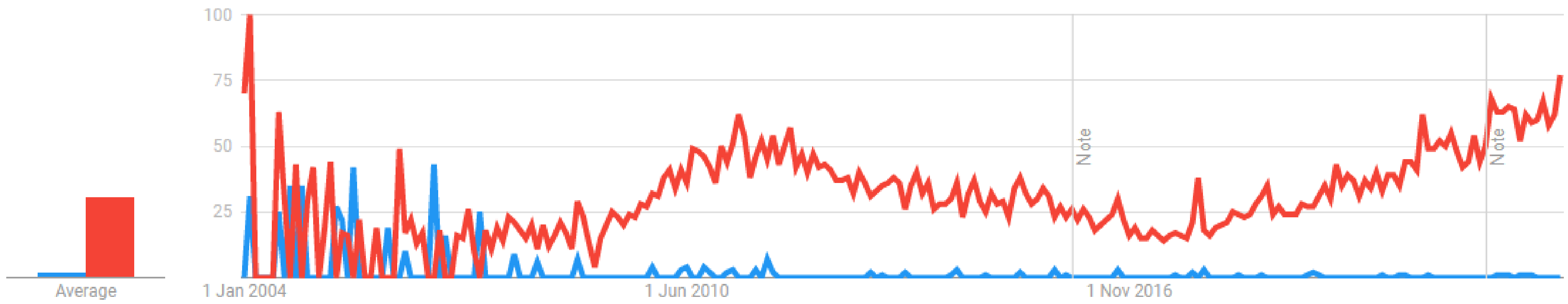





Trade-off between model interpretability and accuracy,
(Arrieta, Del Ser et al; 2019)

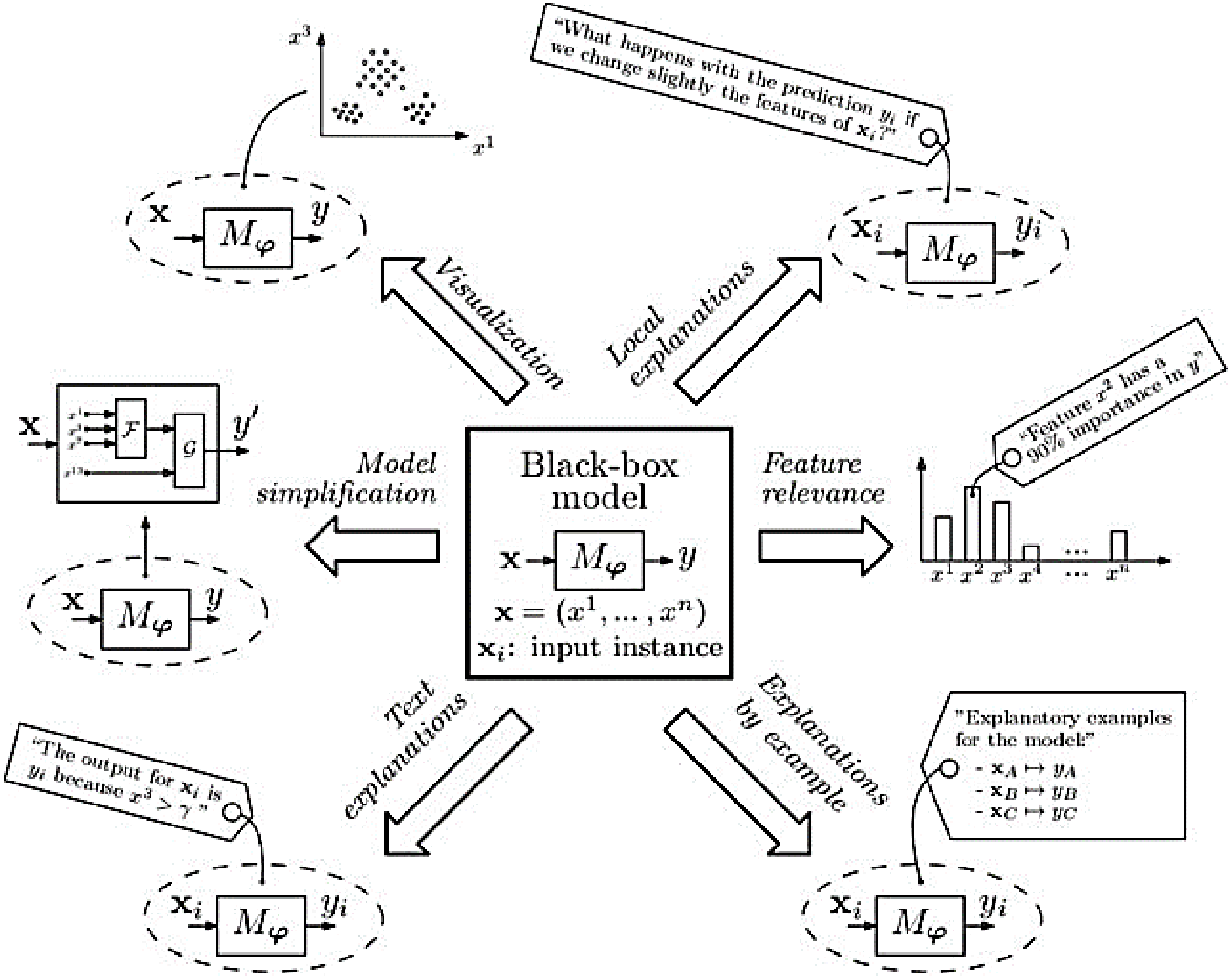
# Trend report

Interpretable Artificial Intelligence - Explainable Artificial Intelligence



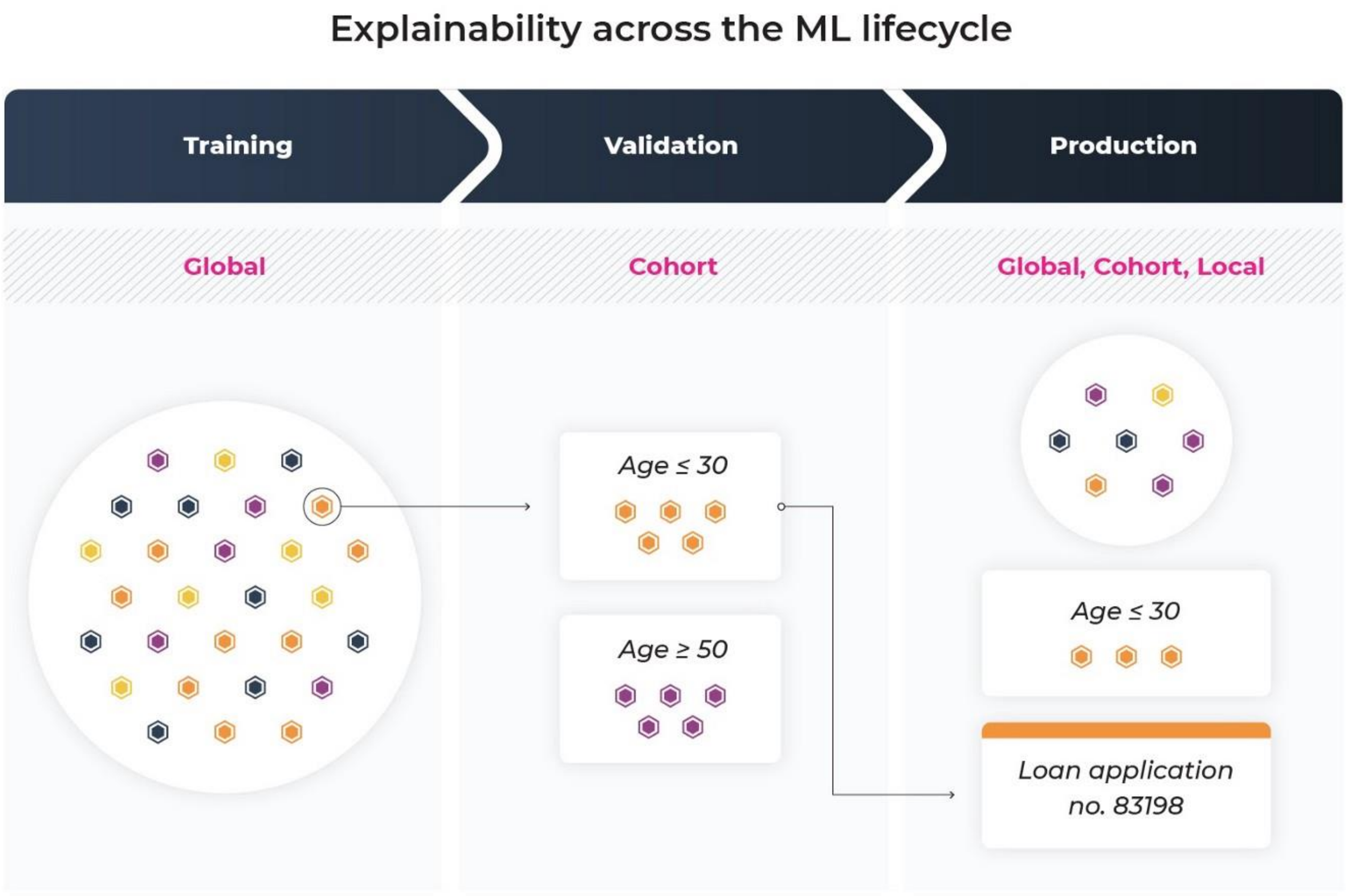

https://trends.google.com/trends/explore?date=all&q=interpretable%20artificial%20intelligence,%2Fg%2F11fxvbm8wd

# Post-hoc explainability approaches



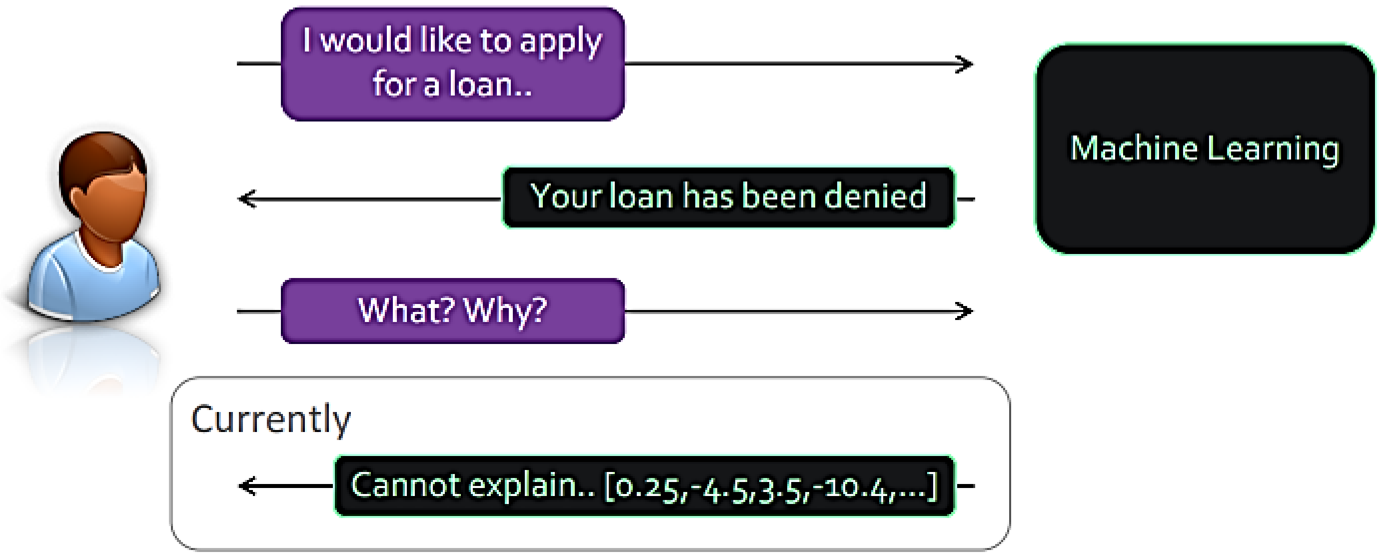

Different post-hoc explainability approaches available for a ML model MΦ (Arrieta, Del Ser et al; 2019)

# Global, Cohort and Local Model Explainability



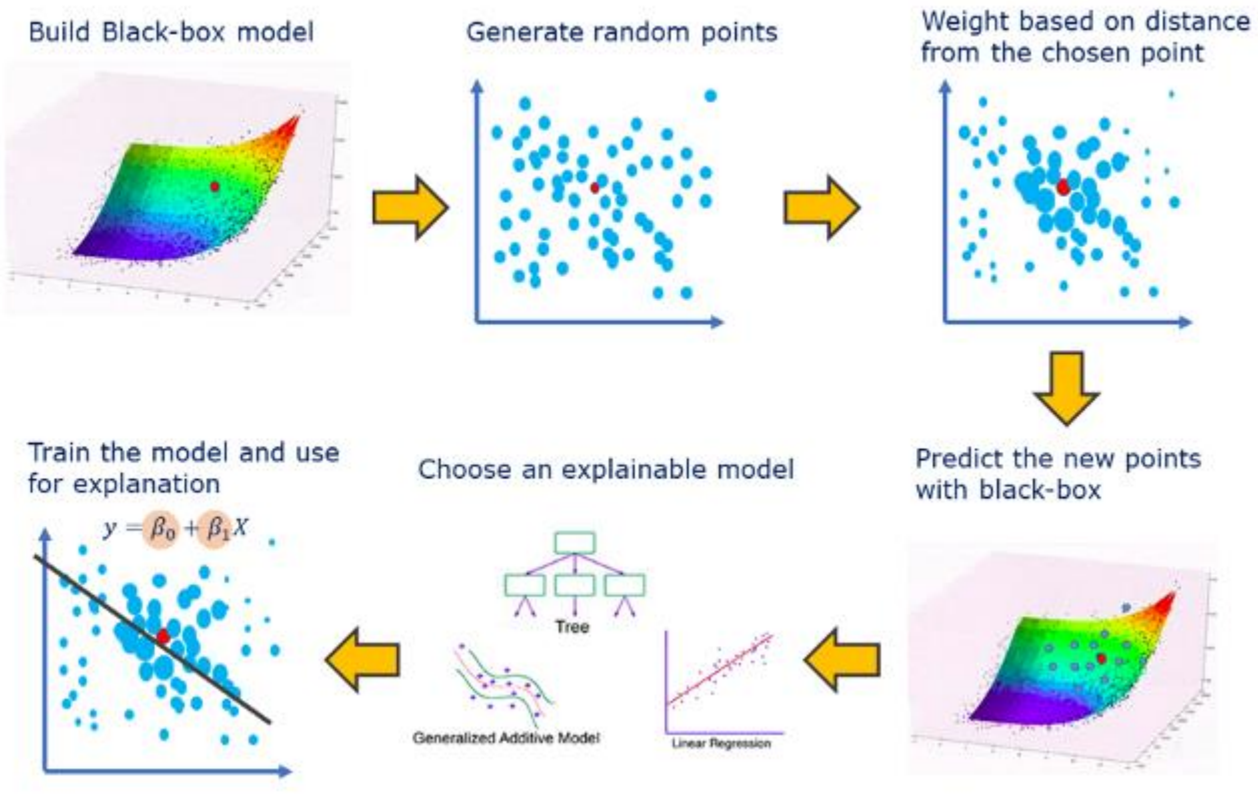
# "Why Should I Trust You?" Explaining the Predictions of Any Classifier





# The LIME Algorithm

Locally approximating black-box classifier with interpretable classifier




Giorgio Visani (2020)

# LIME: General framework

$$e(x, \theta) = \underset{e \in \mathcal{E}(x)}{\operatorname{argmax}} \; \mathcal{Q}(e, x, \theta) + \mathcal{I}(e)$$

Instance

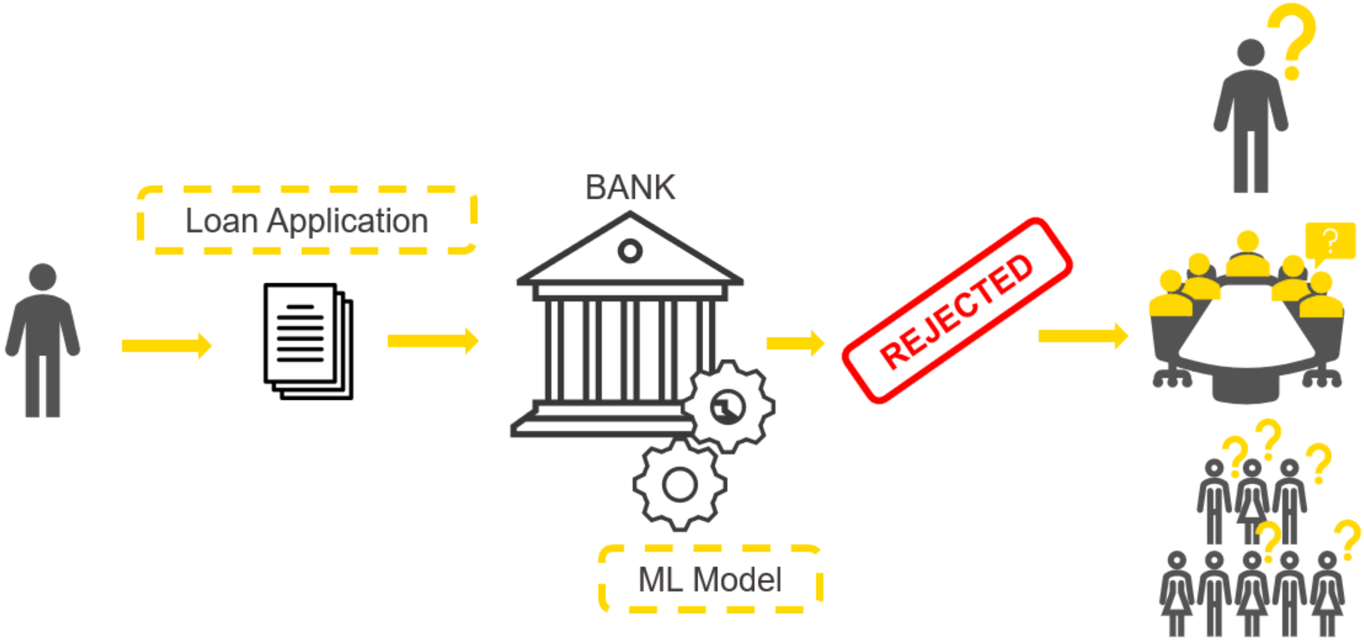
Explanation Family
Universe of possible explanations to search over

Faithfulness of Explanation
Is the explanation faithful to the model in context of *x*

Interpretability
Is the explanation simple enough to read?

# Intrinsically Interpretable Risk Models

- The Probability of Payment (POP) model is a risk scoring model that calculates a POP estimate for each contract /loan at origination
- POP values are used to predict expected defaults and expected credit loss (ECL) of the banks




To build an xNN that pursue a good balance between POP prediction accuracy and model interpretability

# Methodology

ICISCoIS 2023

- Generalized additive index model (GAIM) is used in POP prediction. The relationship between raw features x ε ℝ $^p$ and the response y is represented by

$$g[E[y/x]] = u + \sum_{j=1}^{M} h_j(w_j{}^T x) \qquad \text{---------} \ (1)$$

*g is a pre-specified link function, u is the intercept, and M is the number of additive functional components.*

- GAIM is estimated using back fitting algorithm, iteratively estimates a pair of $\{w_j, h_j\}$ at a time, with other pairs fixed
- Nonparametric regression (ex: smoothing splines) is used to fit the shape functions in (1)
- GAIM includes both main effects and interaction effects between individual features for performance improvement
- In addition to neural network parametrization, the interpretability of (1) is enhanced with below three constraints:

**Sparsity : Prune the trivial main/interaction effects**

$$D(h_j) = \frac{1}{n-1} \sum_{j \varepsilon s_1} h^2{}_j(x_j)$$

$$D(f_{j,k}) = \frac{1}{n-1} \sum_{j \varepsilon s_1} f^2{}_{jk}(x_j x_k)$$

*Main effects (h(x))*

*Interaction effect (f($x_j$, $x_k$))*

**Heredity : Atleast one main effect is significant**

$\forall$(j; k) ϵ $S_2$ : j ϵ $S_1$ or k ϵ $S_1$

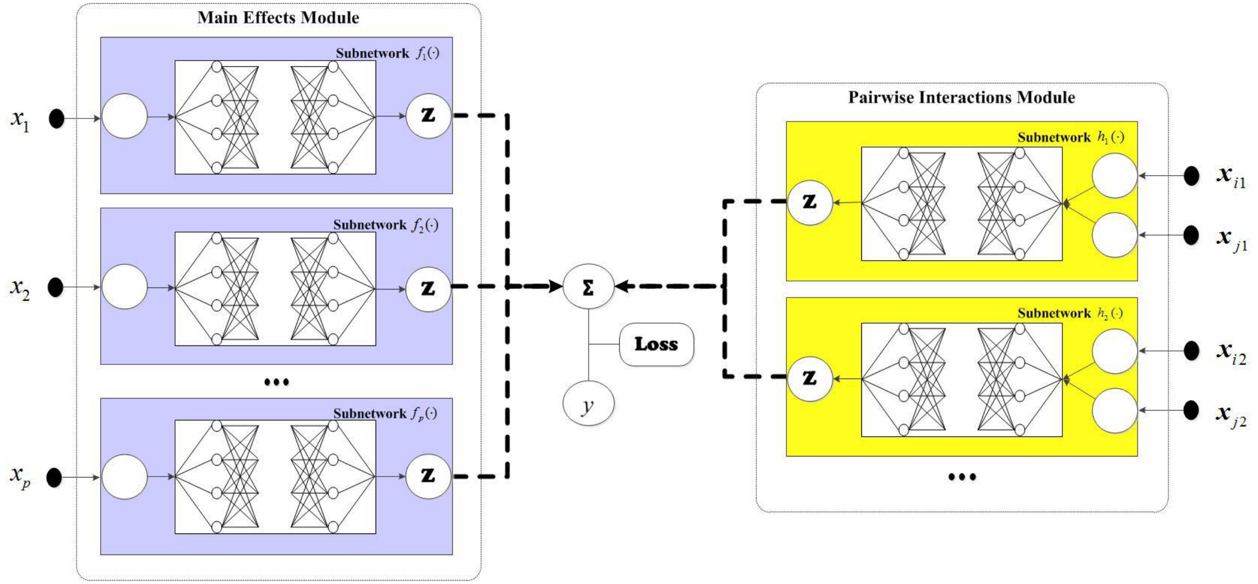
*$S_1$,$S_2$ – List of main & interaction effects*

**Marginal : Separate main and interaction effect**

$$\Phi(h_j, f_{j,k}) = | \frac{1}{n} \sum_{j \varepsilon s_1} h_j(x_j)\, f_{j,k}(x_j, x_k)$$

*Smaller the value of orthogonality $\Phi(h_j, f_{j,k})$, clearly marginal effect hj is separated from child interaction fjk*

# Network Architecture




Proposed xNN is formulated as follows:

$$g[E[y/x]] = u + \sum_{j \varepsilon s_1} h_j(x_j) + \sum_{(j,k) \varepsilon s_2} f_{j,k}(x_j, x_k) \qquad \text{---------} \ (2)$$

- The main effects (h(x)) are first fitted
- Top-K ranked pairwise interactions ($f(x_j, x_k)$) are selected & fitted to the residuals, subject to heredity constraint
- The dashed arrows to Σ nodes denote the sparsity constraints, the trivial subnetworks are pruned
- Finally, the marginal clarity is imposed for regularizing pairwise interactions

# Hyperparameters and Interpretability

ICISCoIS 2023

- Maximal number of pairwise interactions is set to K = 30
- Subnetwork is configured with 5 ReLU hidden layers each with 40 nodes
- Subnetwork weights are initialized using the Gaussian orthogonal initializer
- Initial learning rate of the Adam optimizer is set to 0.0001
- Mini-batch sample size is determined according to the sample sizes of different datasets
- A 20% validation set is split for early stopping, and the early stopping threshold is set to be 50 epochs
- The tolerance threshold  is set to be 1% of the minimal validation loss.
- The marginal clarity regularization strength  can be empirically selected from 0.0001 to 1

## Importance Ratio (IR) :

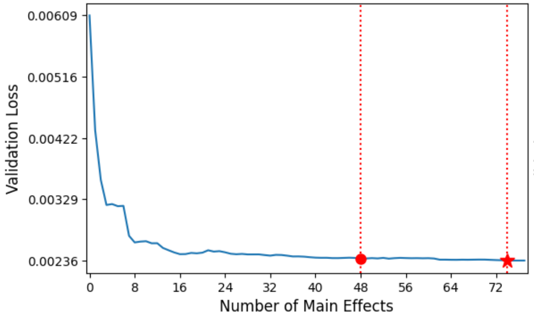
Contribution of each individual variable to the overall prediction is measured by following :

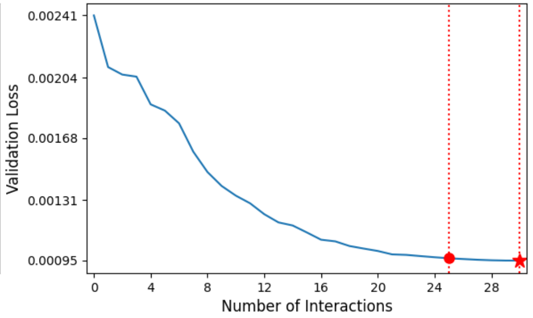
**Main effects :** $IR[j] = \frac{D(h_j)}{T}$

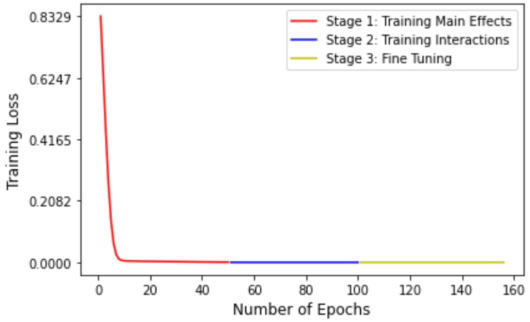
**Interaction effects :** $f_{j,k}\,[j,k] = \frac{D(f_{j,k})}{T}$

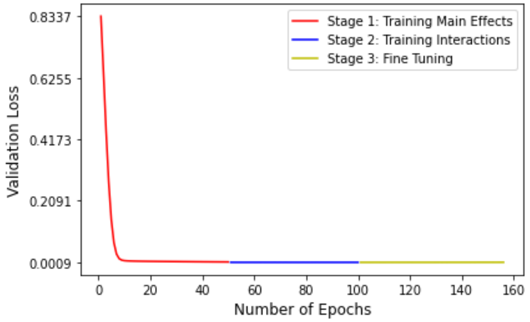
$where \quad T = \sum_{j\varepsilon S_1} D(h_j) + \sum_{(j,k)\varepsilon S_2} D(f_{j,k})$

# **Results :** Credit risk models

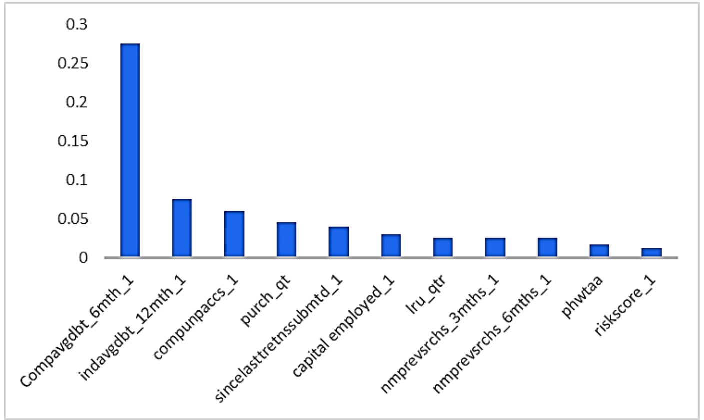


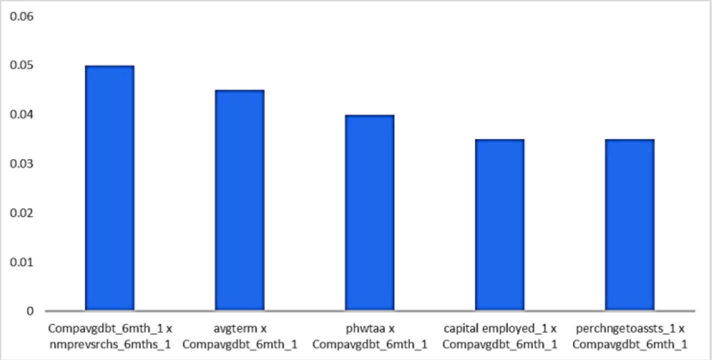




**Top 10 Main effects**

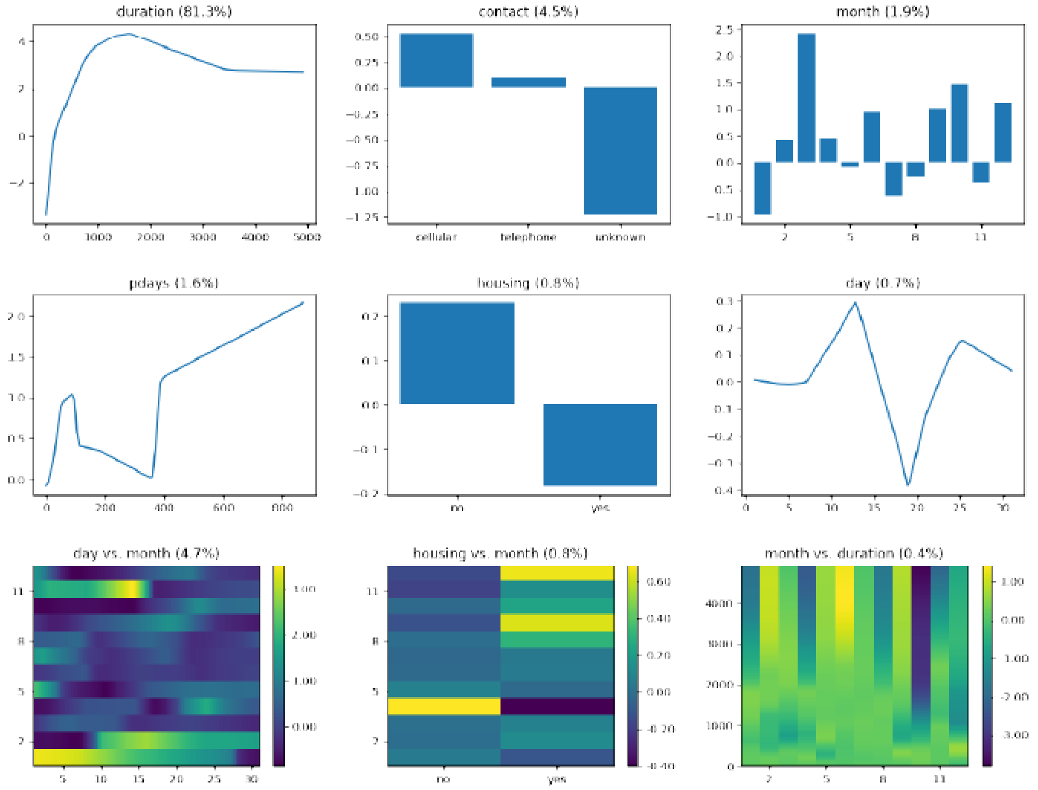


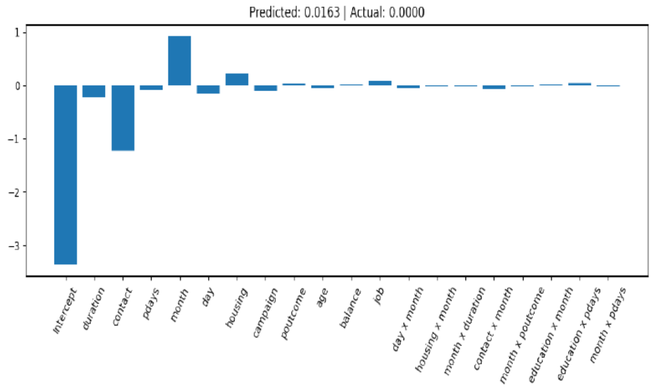

**Top 5 Interaction effects**

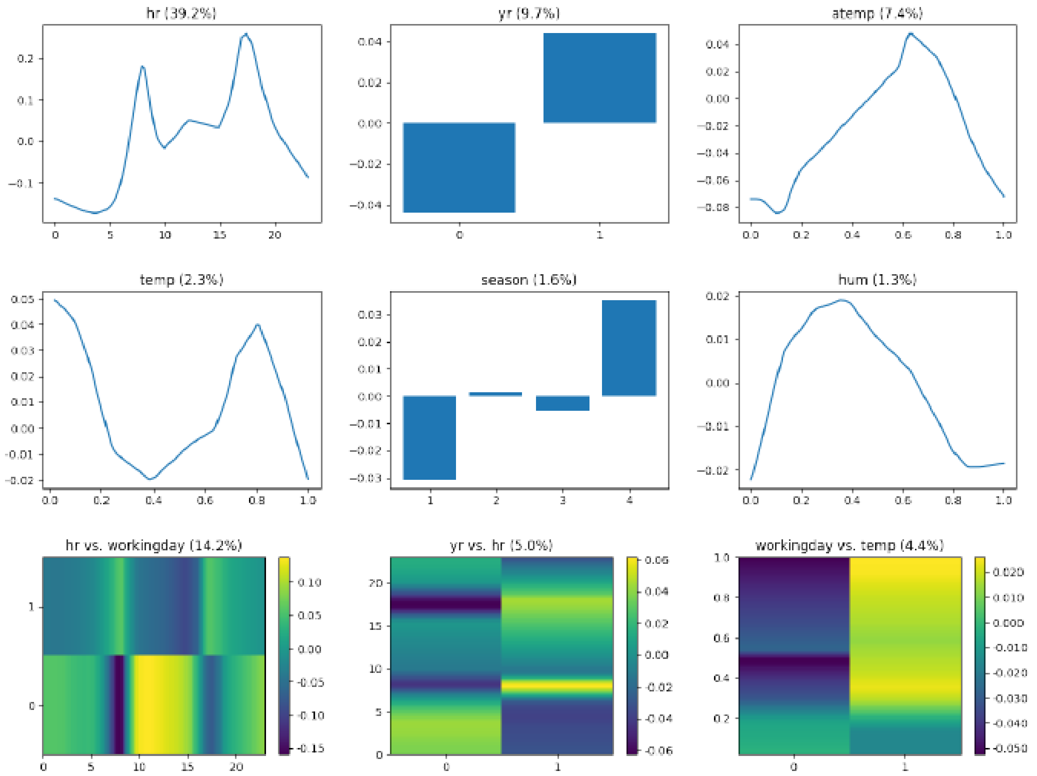


# **Results :** Bank Marketing dataset Vs. Bike Sharing Hour Dataset

ICISCoIS 2023



# Conclusion


*Image source: Interpretable Machine Learning (C. Molnar)*

**Thank you**