Explainable Artificial Intelligence

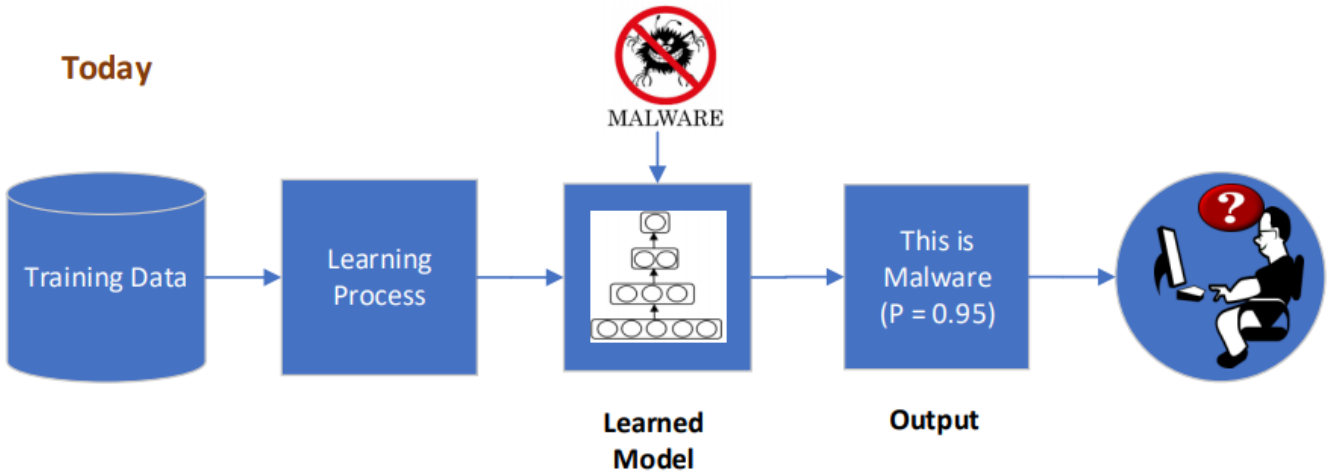
Adel Abusitta
E-mail: adel.abusitta@uwindsor.ca or abusitta.adel@gmail.com

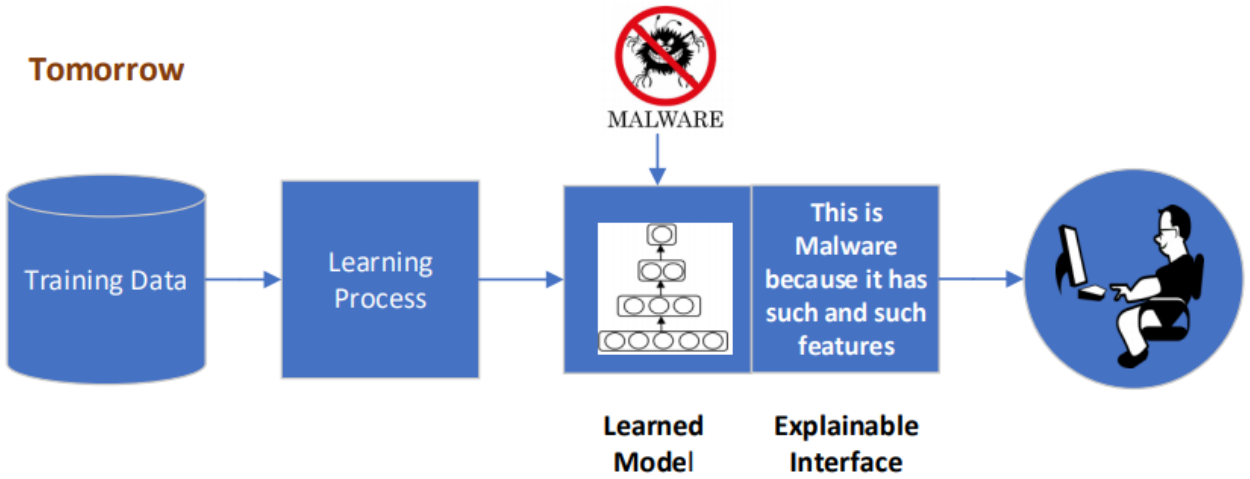
R^3AI

The '**R^3**' refers to what AI scientists see as the three essential characteristics that must be prioritized in AI development: "robust" (AI must rely on sound models and applications), "reasoned" (AI must be explainable, causal and modular), and "responsible" (AI must be ethical and inclusive).

Explainable Artificial Intelligence (XAI)

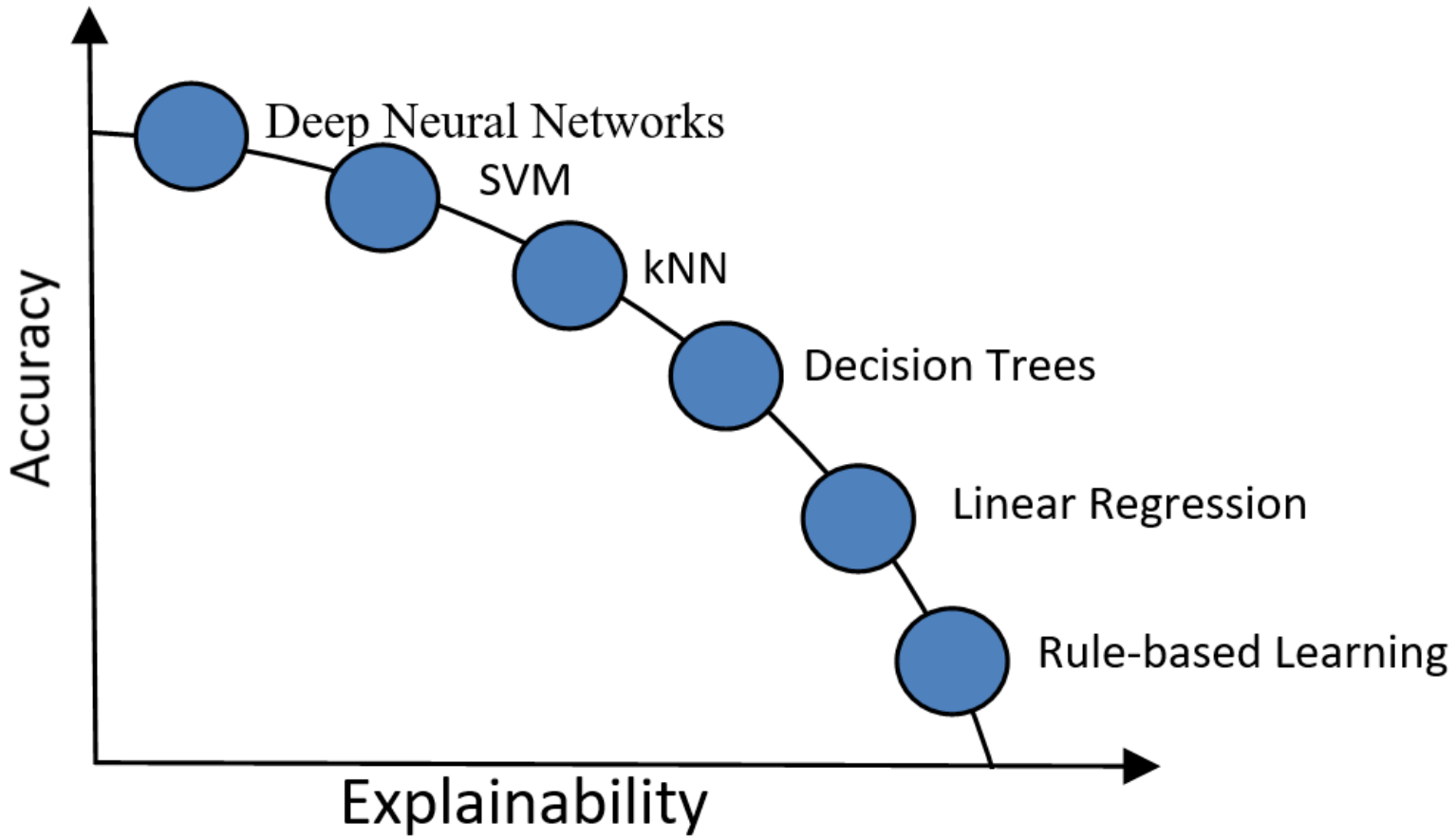
What is XAI and Why?



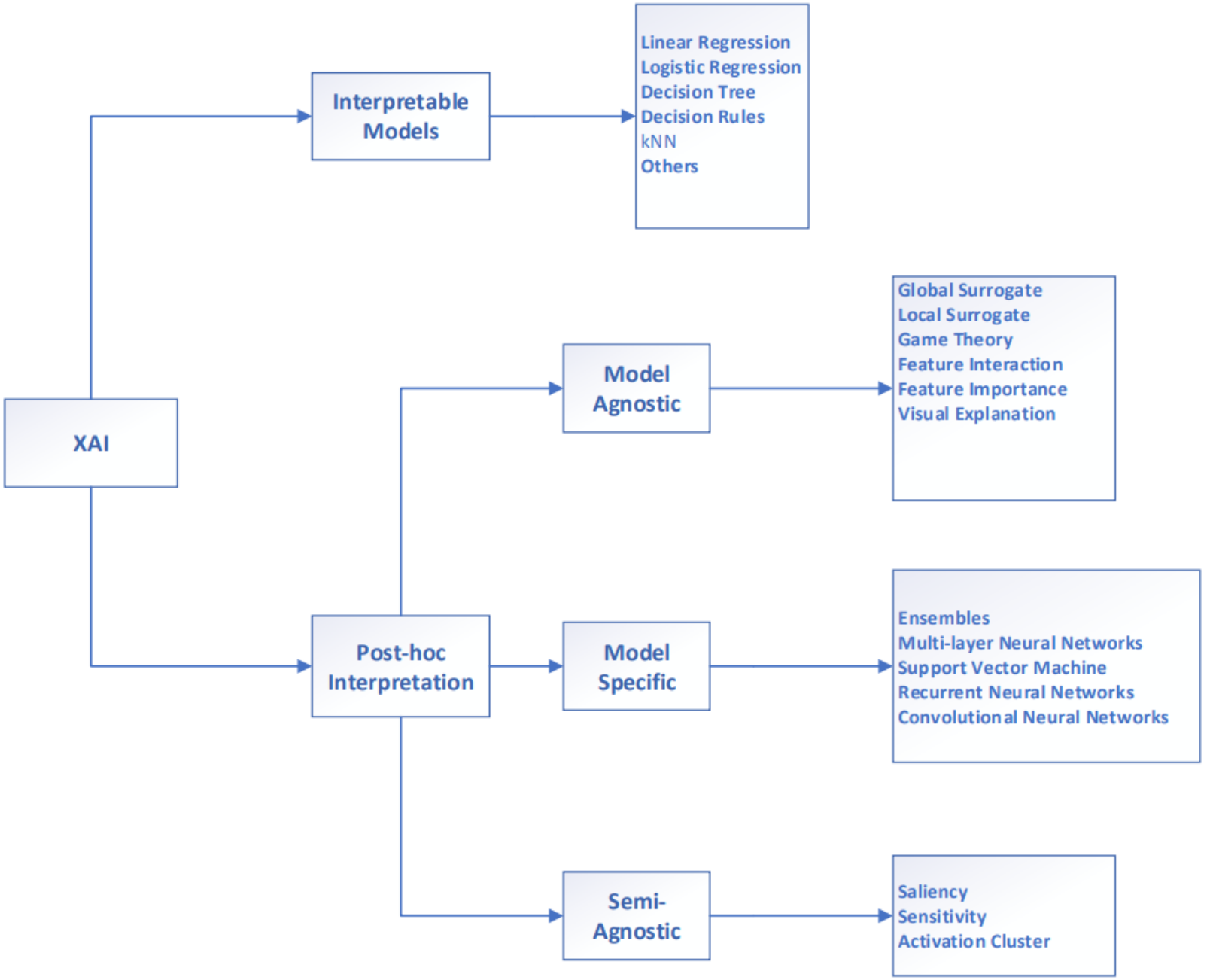


The goal of XAI is to describe in detail how ML models produce their prediction

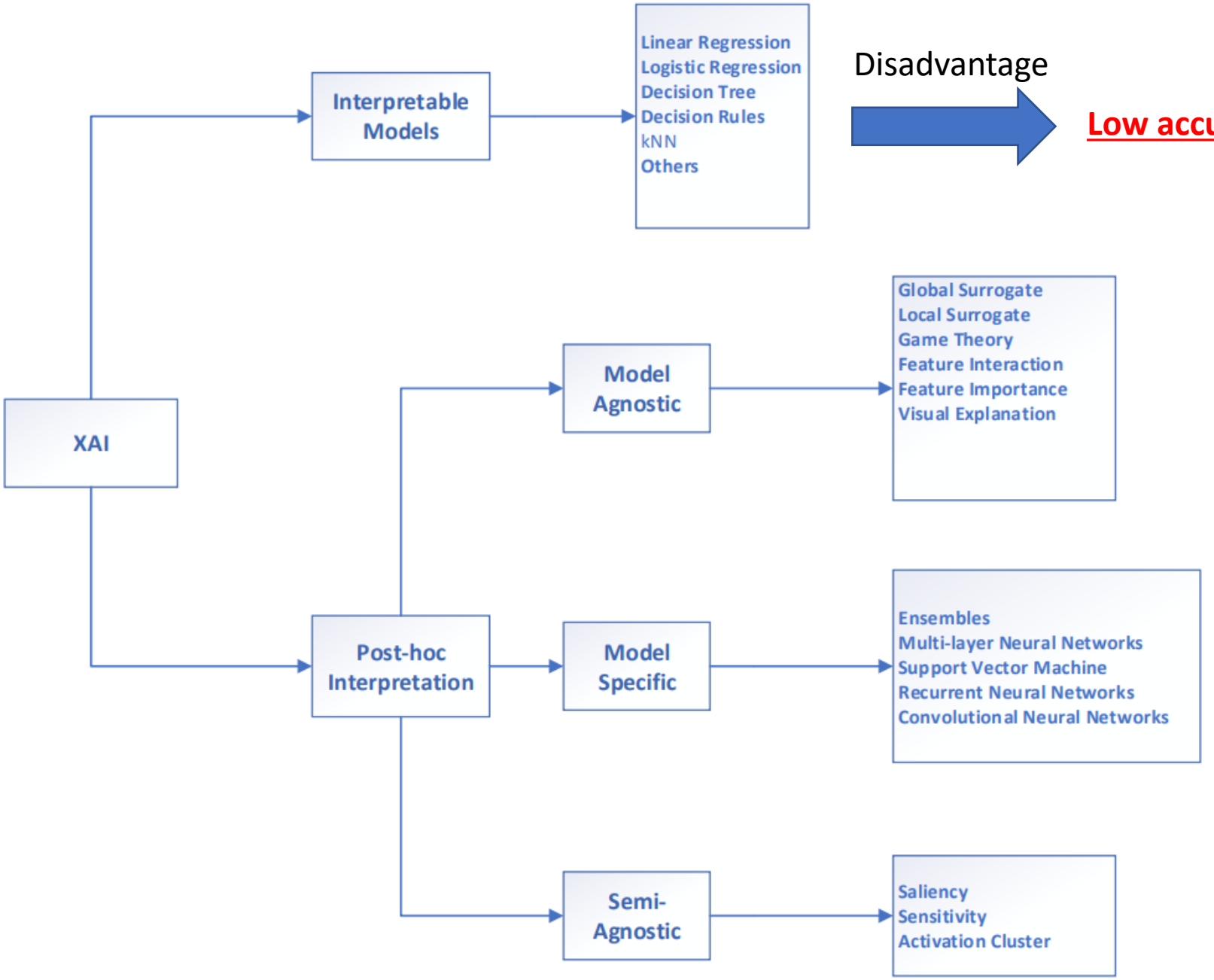
Tradeoff


Deep Neural Networks
SVM
kNN
Decision Trees
Linear Regression
Rule-based Learning
Accuracy
Explainability

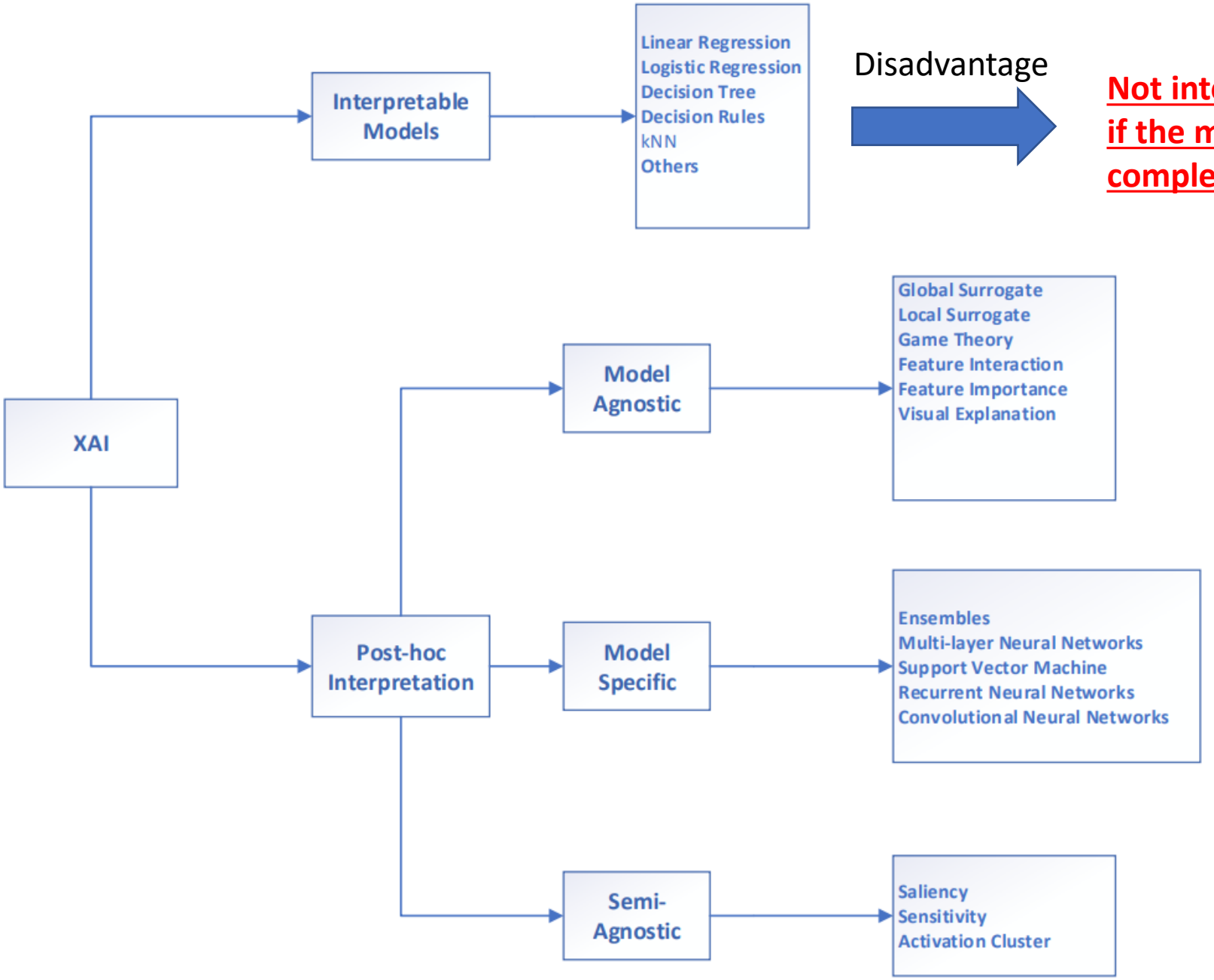
Taxonomy


XAI
Interpretable Models
Linear Regression
Logistic Regression
Decision Tree
Decision Rules
kNN
Others
Post-hoc Interpretation
Model Agnostic
Global Surrogate
Local Surrogate
Game Theory
Feature Interaction
Feature Importance
Visual Explanation
Model Specific
Ensembles
Multi-layer Neural Networks
Support Vector Machine
Recurrent Neural Networks
Convolutional Neural Networks
Semi-Agnostic
Saliency
Sensitivity
Activation Cluster

Taxonomy

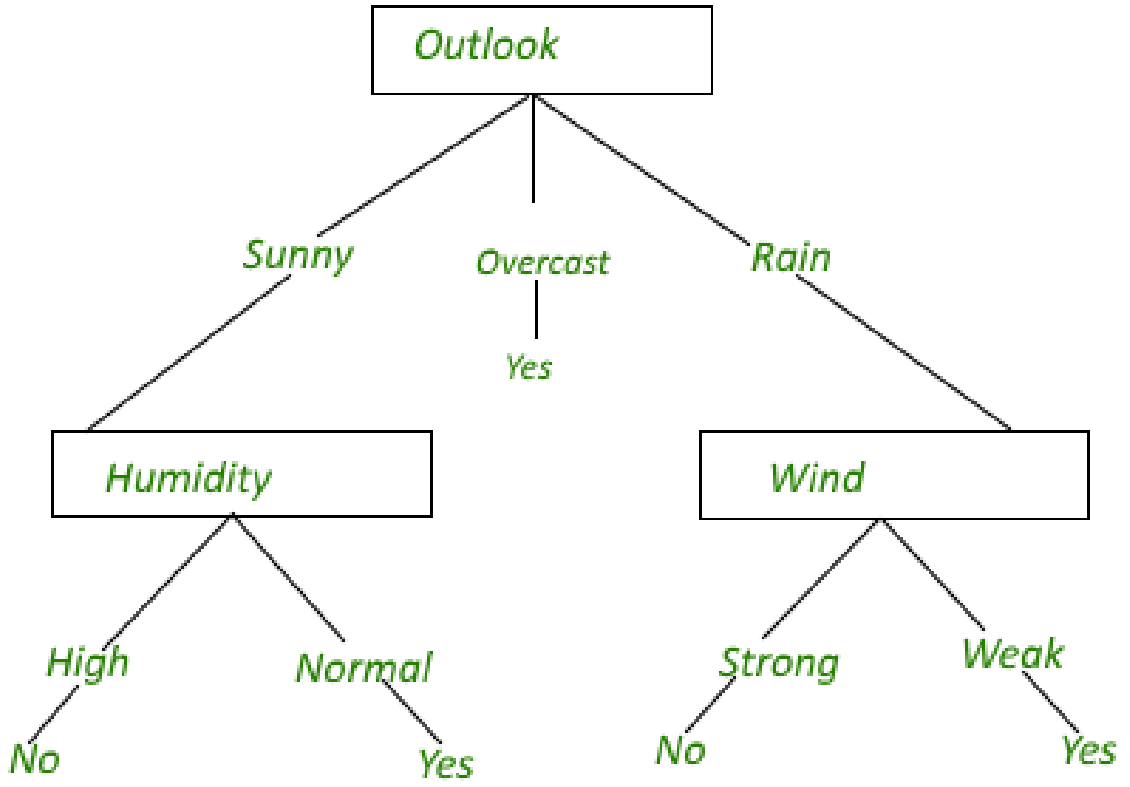

XAI
Interpretable Models
Linear Regression
Logistic Regression
Decision Tree
Decision Rules
kNN
Others
Disadvantage
Low accuracy
Post-hoc Interpretation
Model Agnostic
Global Surrogate
Local Surrogate
Game Theory
Feature Interaction
Feature Importance
Visual Explanation
Model Specific
Ensembles
Multi-layer Neural Networks
Support Vector Machine
Recurrent Neural Networks
Convolutional Neural Networks
Semi-Agnostic
Saliency
Sensitivity
Activation Cluster

Taxonomy


XAI
Interpretable Models
Linear Regression
Logistic Regression
Decision Tree
Decision Rules
kNN
Others
Disadvantage
Not interpretable if the model is complex
Post-hoc Interpretation
Model Agnostic
Global Surrogate
Local Surrogate
Game Theory
Feature Interaction
Feature Importance
Visual Explanation
Model Specific
Ensembles
Multi-layer Neural Networks
Support Vector Machine
Recurrent Neural Networks
Convolutional Neural Networks
Semi-Agnostic
Saliency
Sensitivity
Activation Cluster

Example: Decision Tree

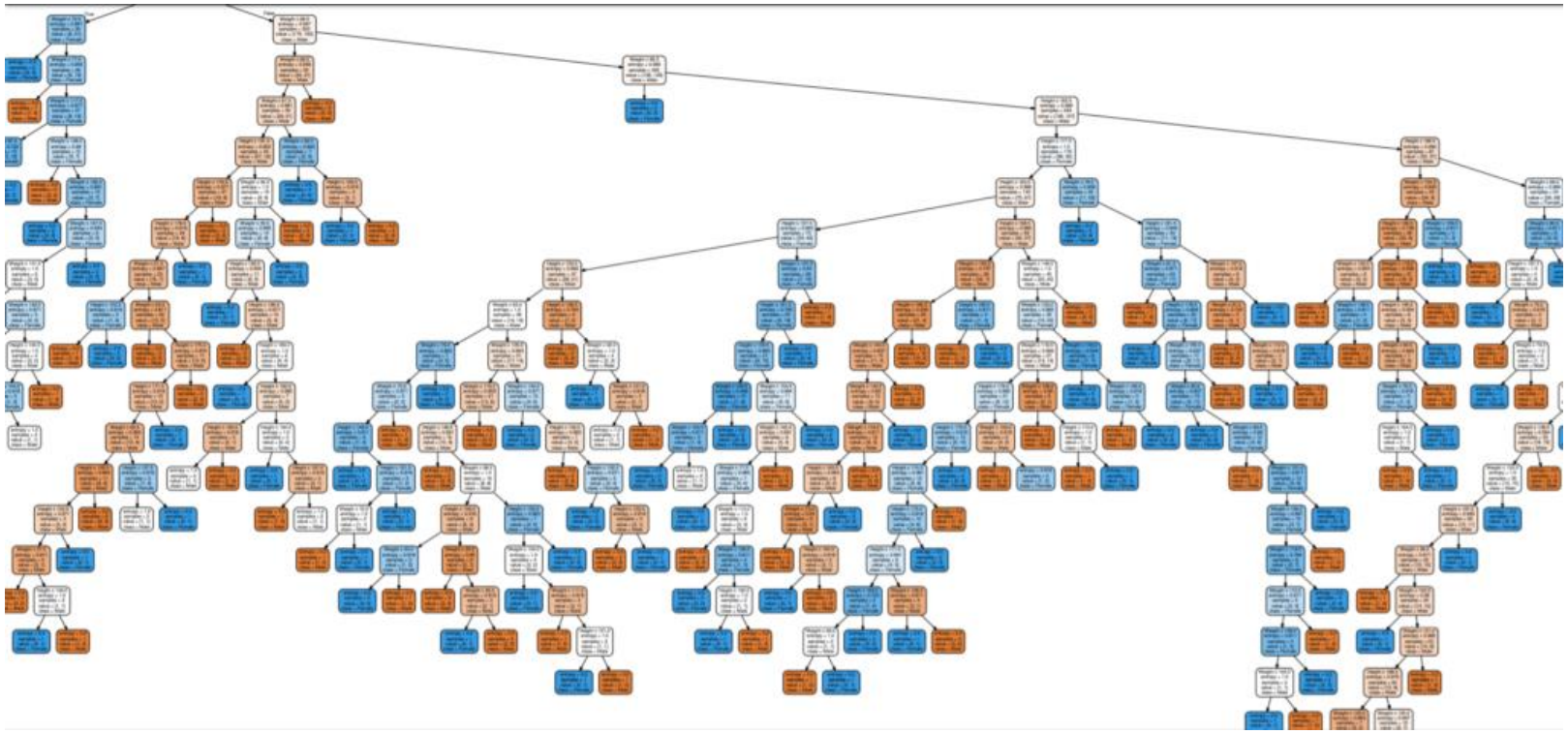
Interpretable



Credit: https://www.geeksforgeeks.org/decision-tree/

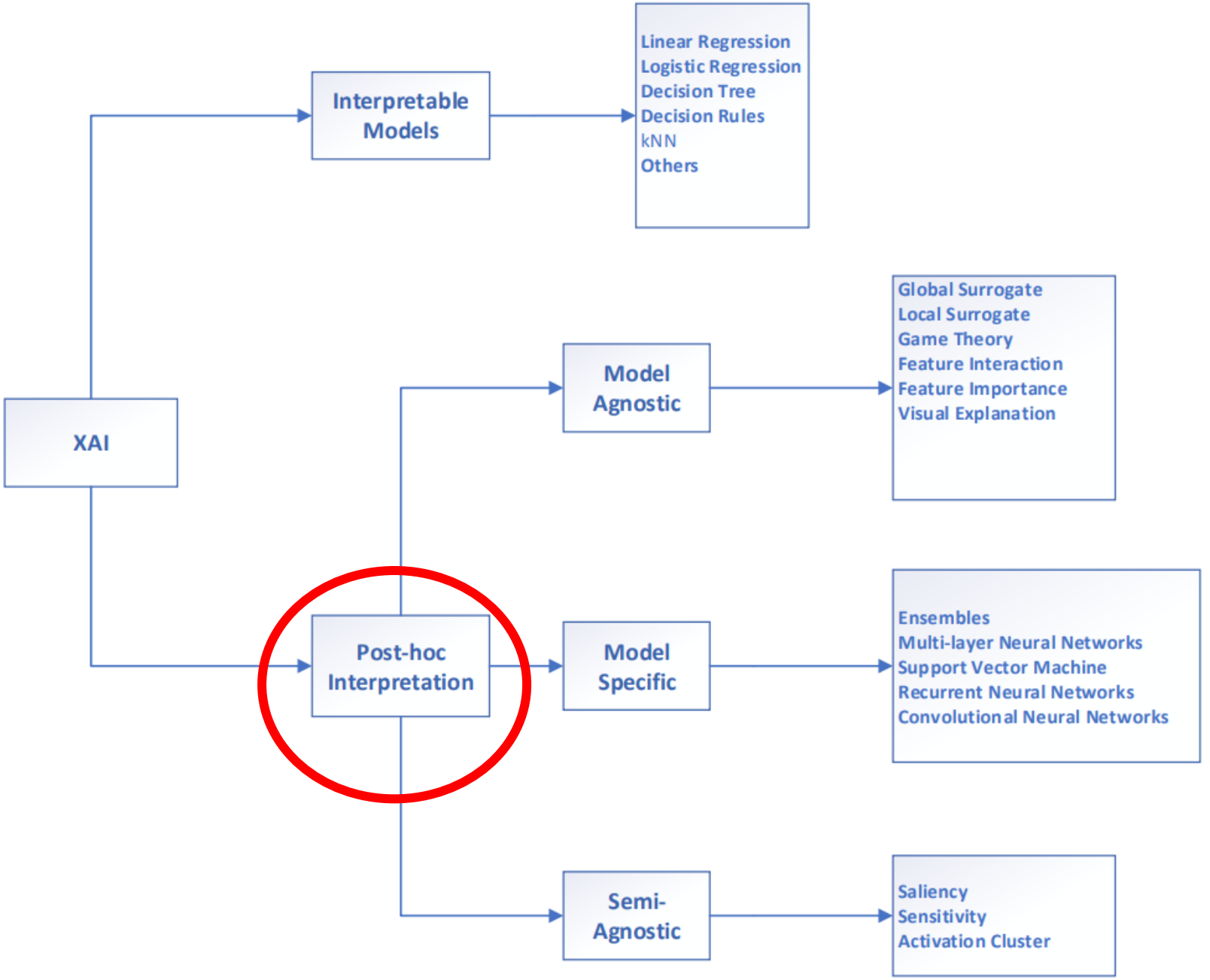
Decision Tree

Not Interpretable

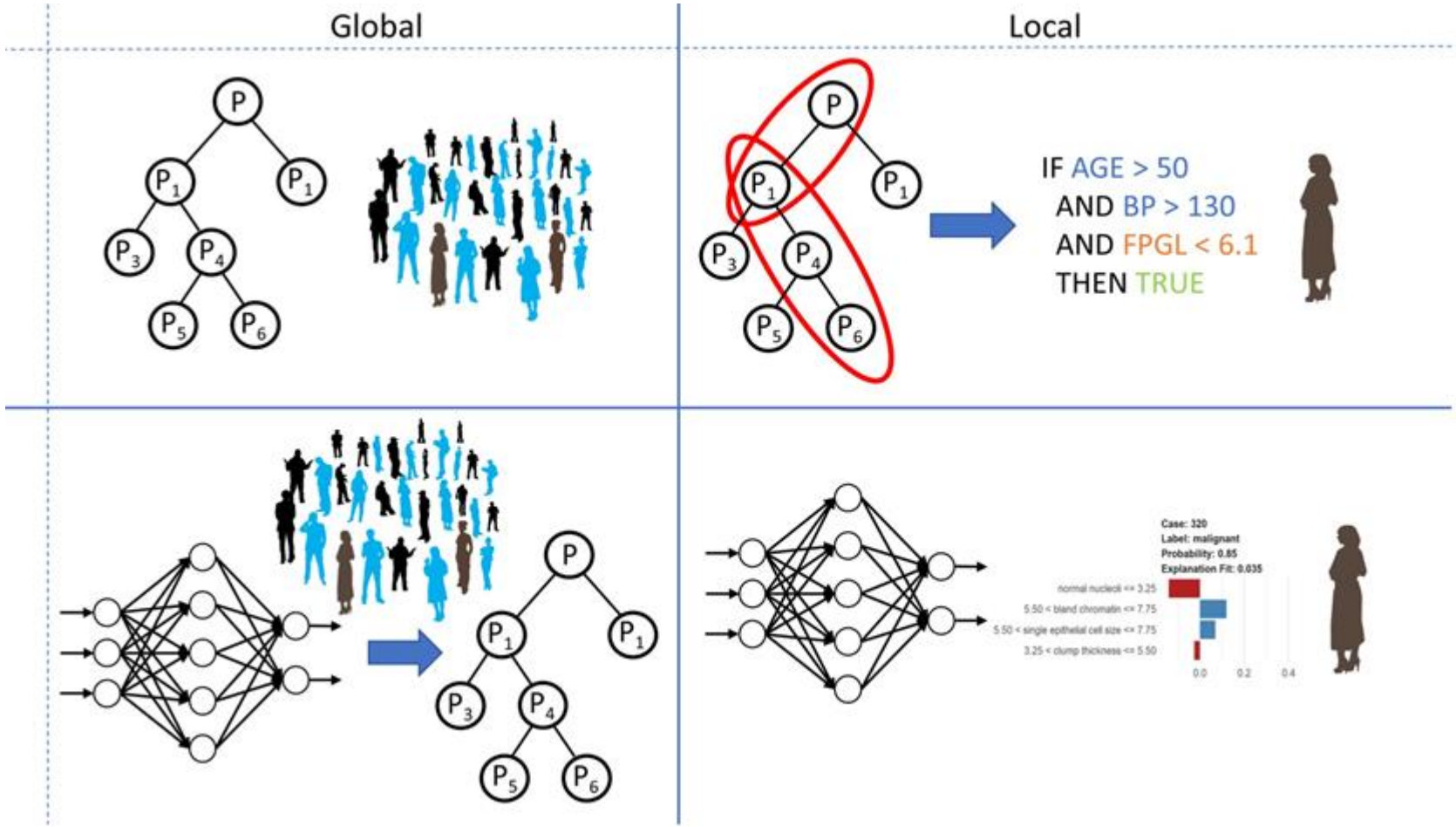


Credit: https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/

Taxonomy

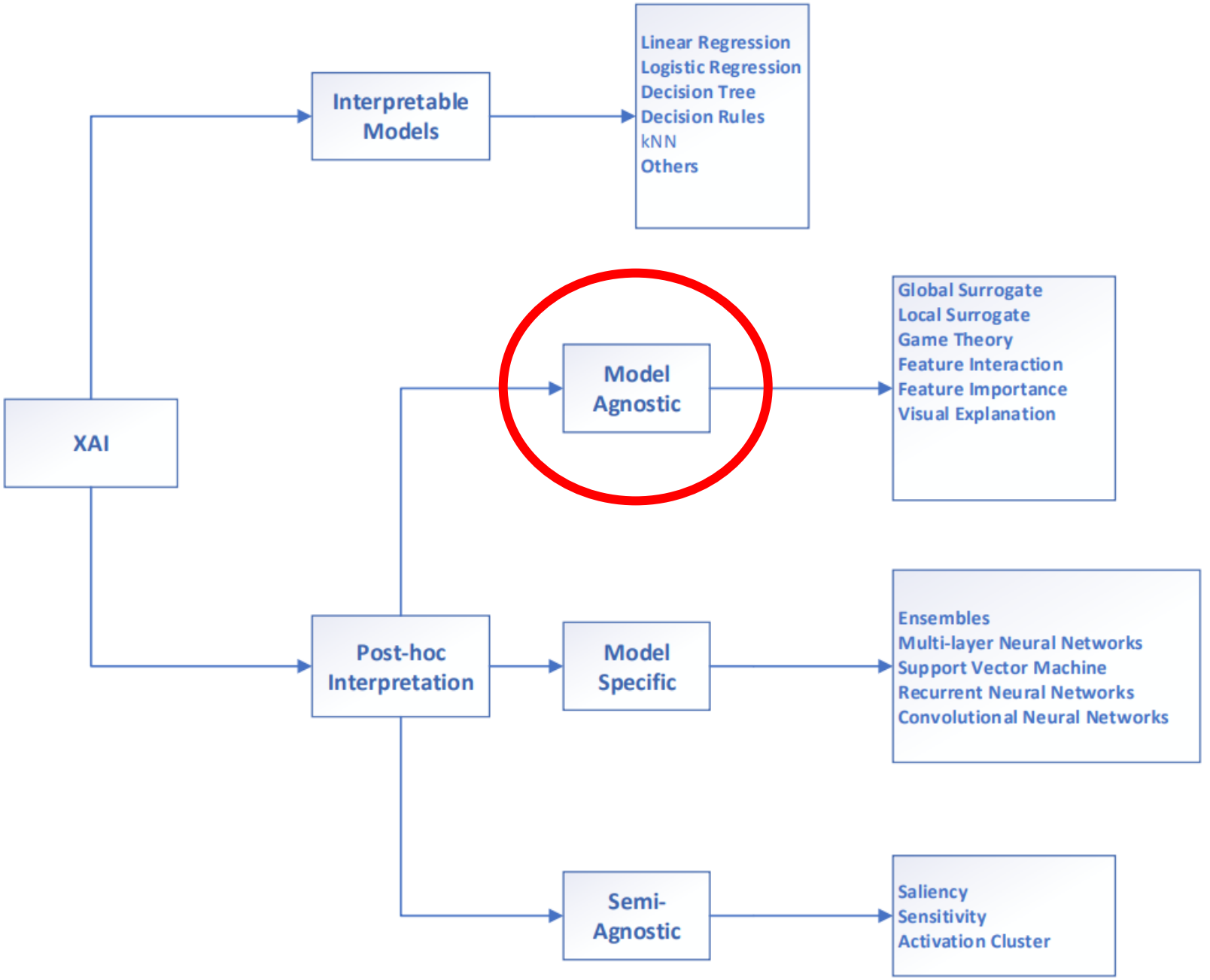

XAI
Interpretable Models
Linear Regression
Logistic Regression
Decision Tree
Decision Rules
kNN
Others
Post-hoc Interpretation
Model Agnostic
Global Surrogate
Local Surrogate
Game Theory
Feature Interaction
Feature Importance
Visual Explanation
Model Specific
Ensembles
Multi-layer Neural Networks
Support Vector Machine
Recurrent Neural Networks
Convolutional Neural Networks
Semi-Agnostic
Saliency
Sensitivity
Activation Cluster

Local vs. Global

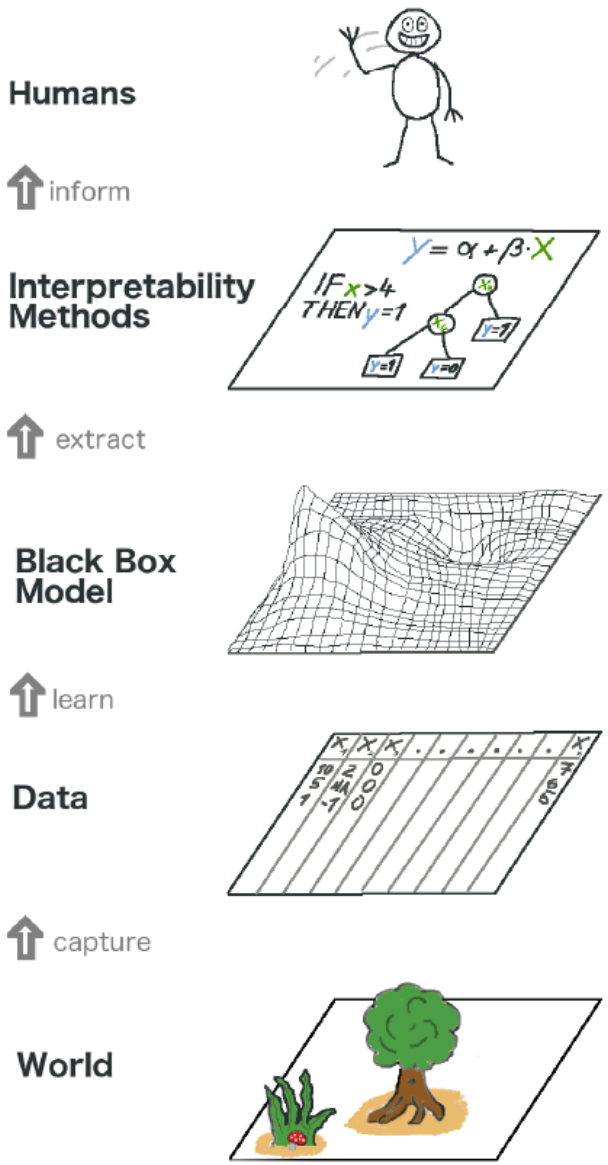


Credit: https://christophm.github.io/interpretable-ml-book/

Taxonomy

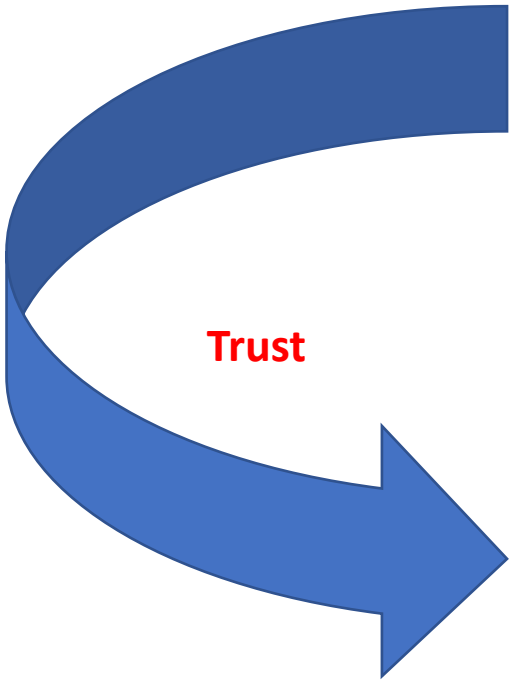

XAI
Interpretable Models
Linear Regression
Logistic Regression
Decision Tree
Decision Rules
kNN
Others
Post-hoc Interpretation
Model Agnostic
Global Surrogate
Local Surrogate
Game Theory
Feature Interaction
Feature Importance
Visual Explanation
Model Specific
Ensembles
Multi-layer Neural Networks
Support Vector Machine
Recurrent Neural Networks
Convolutional Neural Networks
Semi-Agnostic
Saliency
Sensitivity
Activation Cluster

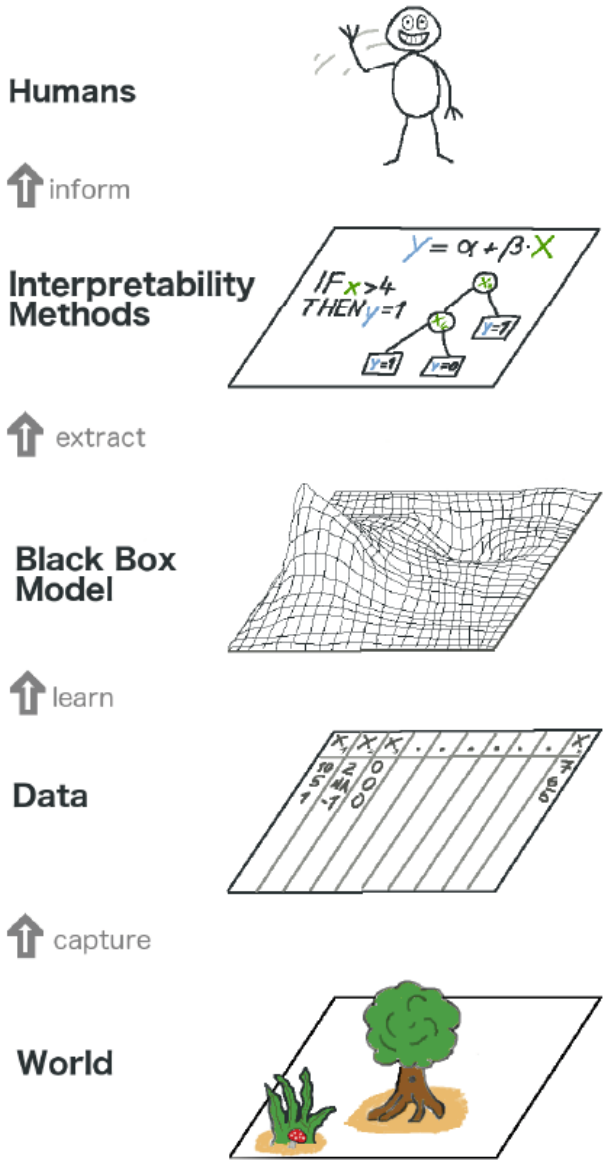
The big picture of explainable machine learning



Credit: https://christophm.github.io/interpretable-ml-book/

The big picture of explainable machine learning





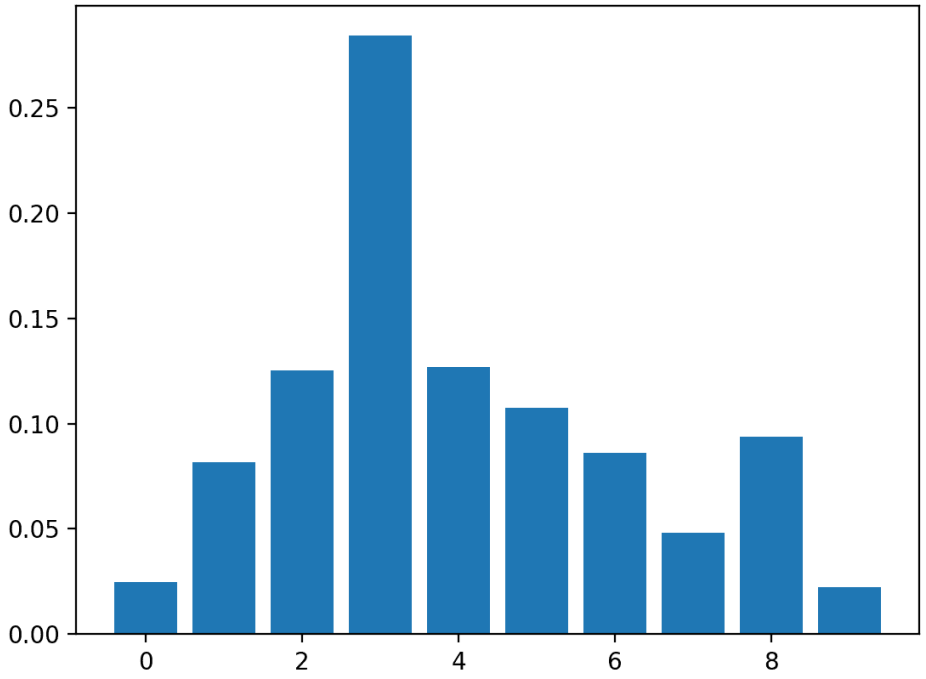
Credit: https://christophm.github.io/interpretable-ml-book/

Post-hoc Interpretation: Model Agnostic

Feature Importance

We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature.

A feature is "**important**" if shuffling its values **increases the model error**, because in this case the model relied on the feature for the prediction.



A feature is "**unimportant**" If shuffling its values **leaves the model error unchanged**, because in this case the model ignored the feature for the prediction.

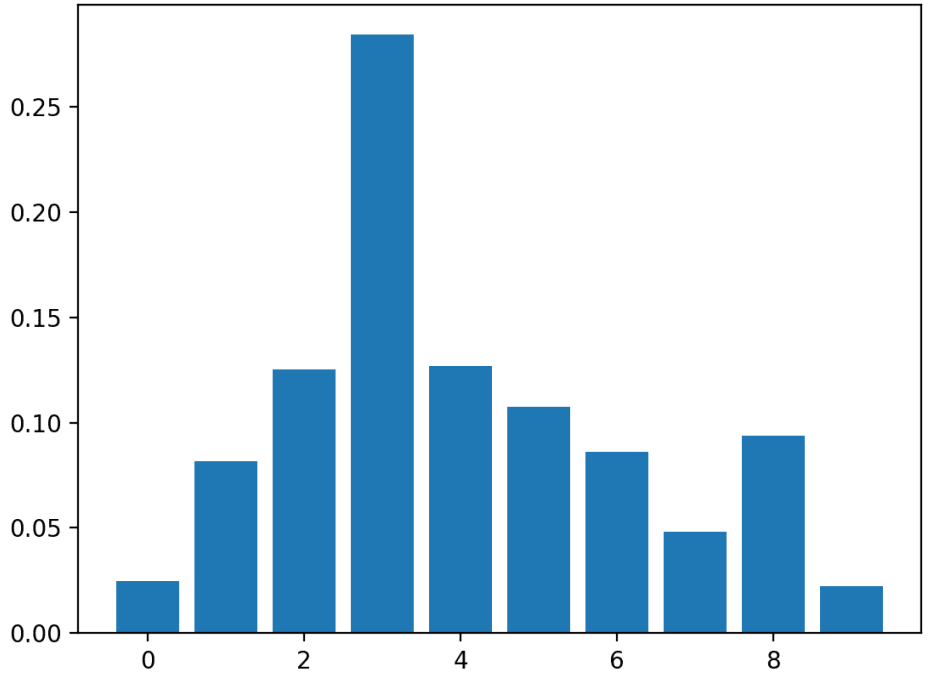
Feature Importance

Advantages

Good interpretation: Feature importance is the increase in model error

Disadvantages

- It is very unclear whether you should use training or test data to compute the feature importance.
- If **features are correlated**, the permutation feature importance can be biased by unrealistic data instances.
- You **need access to the true outcome**.



Global Surrogate

A global surrogate model is an interpretable model that is trained to **approximate the predictions of a black box model**

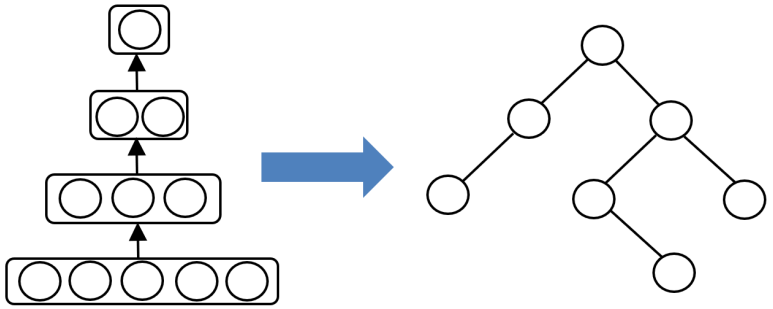
Steps:

1. Select a dataset X, which was used for training the black box

2. For the selected dataset X, get the predictions of the black box model

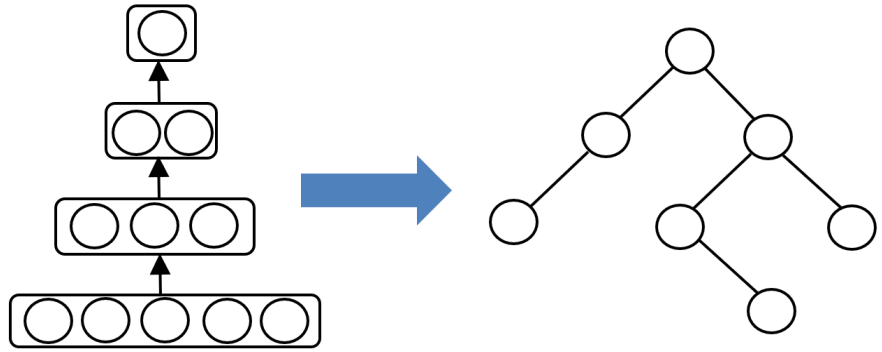
3. Select an interpretable model type (linear model, decision tree, …)

4. Train the interpretable model on the dataset X and its predictions.



Global Surrogate

A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model



Steps (cont.):

5. You now have a surrogate model.

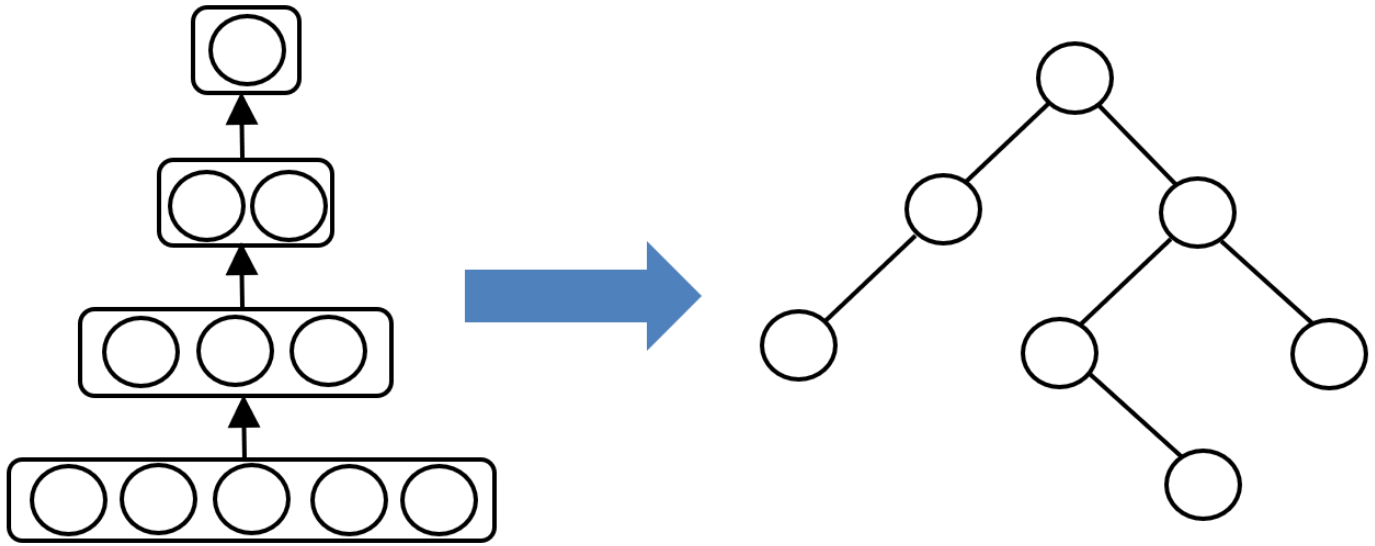
6. Measure how well the surrogate model replicates the predictions of the black box model.

7. Interpret the surrogate model:

Measure how close the surrogate model is to the black box model

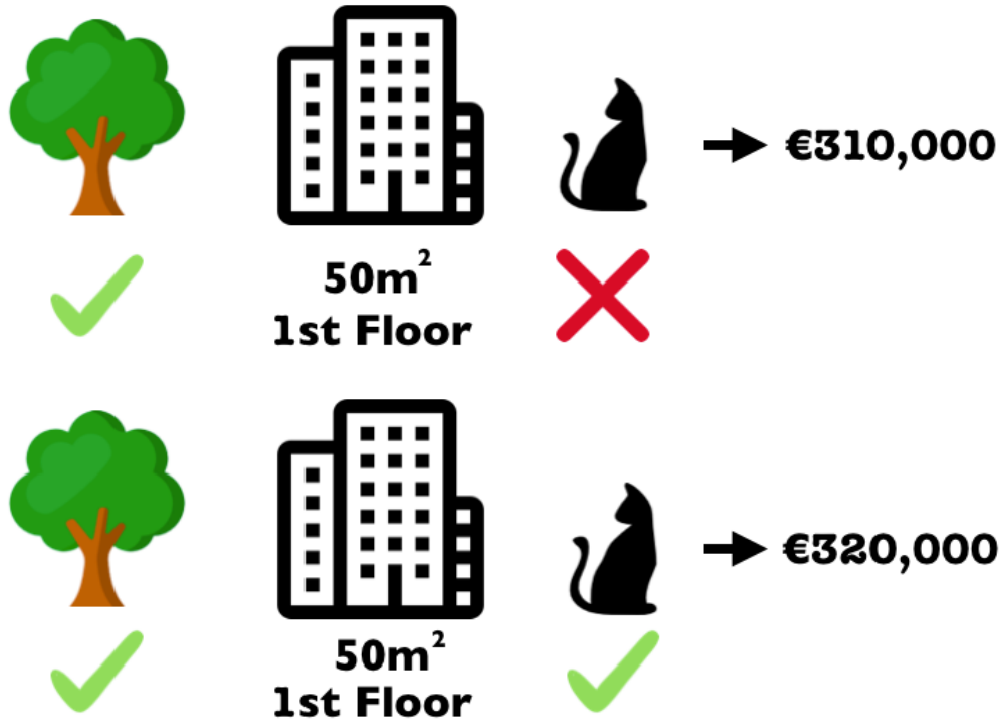
Local Surrogate

Your goal is to understand why the machine learning model made a certain prediction.



Game theory: Shapley Values

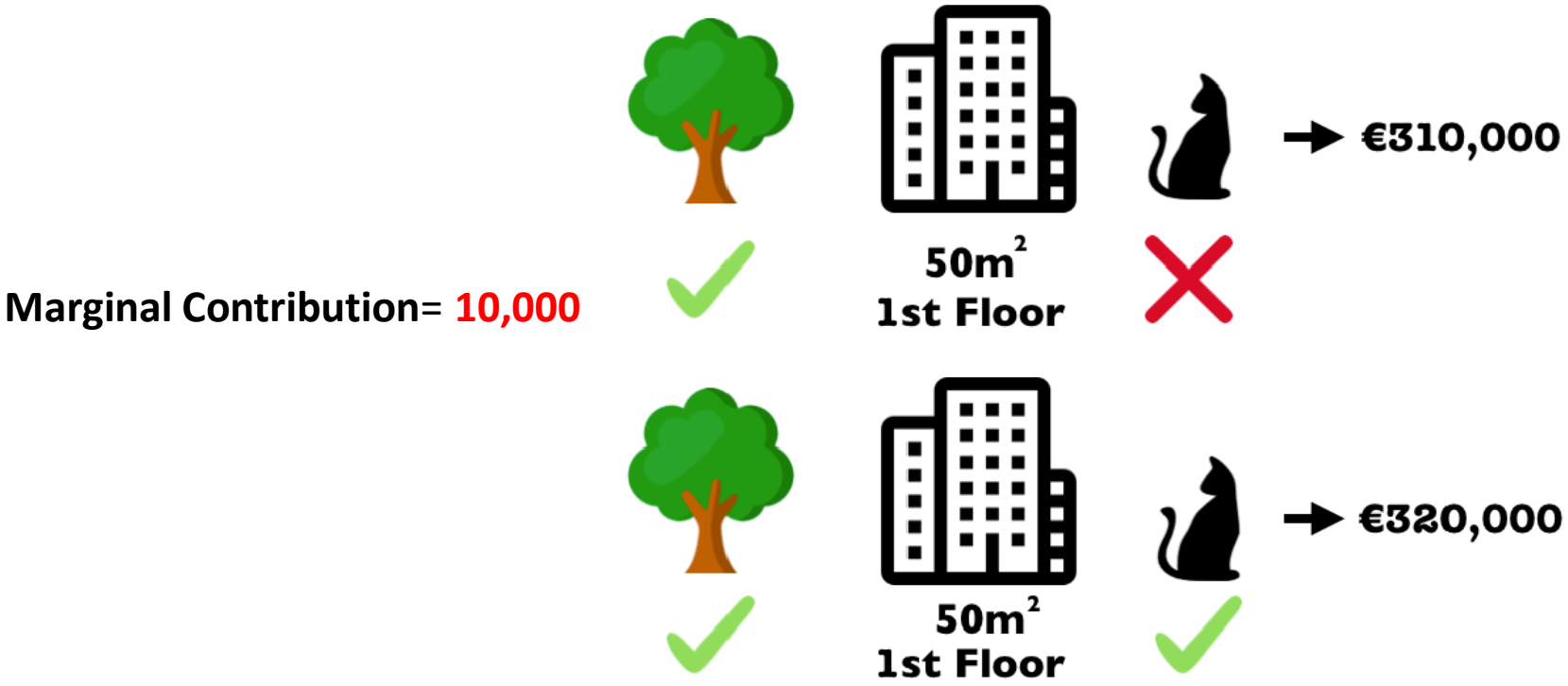
A prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction is the payout.



Credit: https://christophm.github.io/interpretable-ml-book/

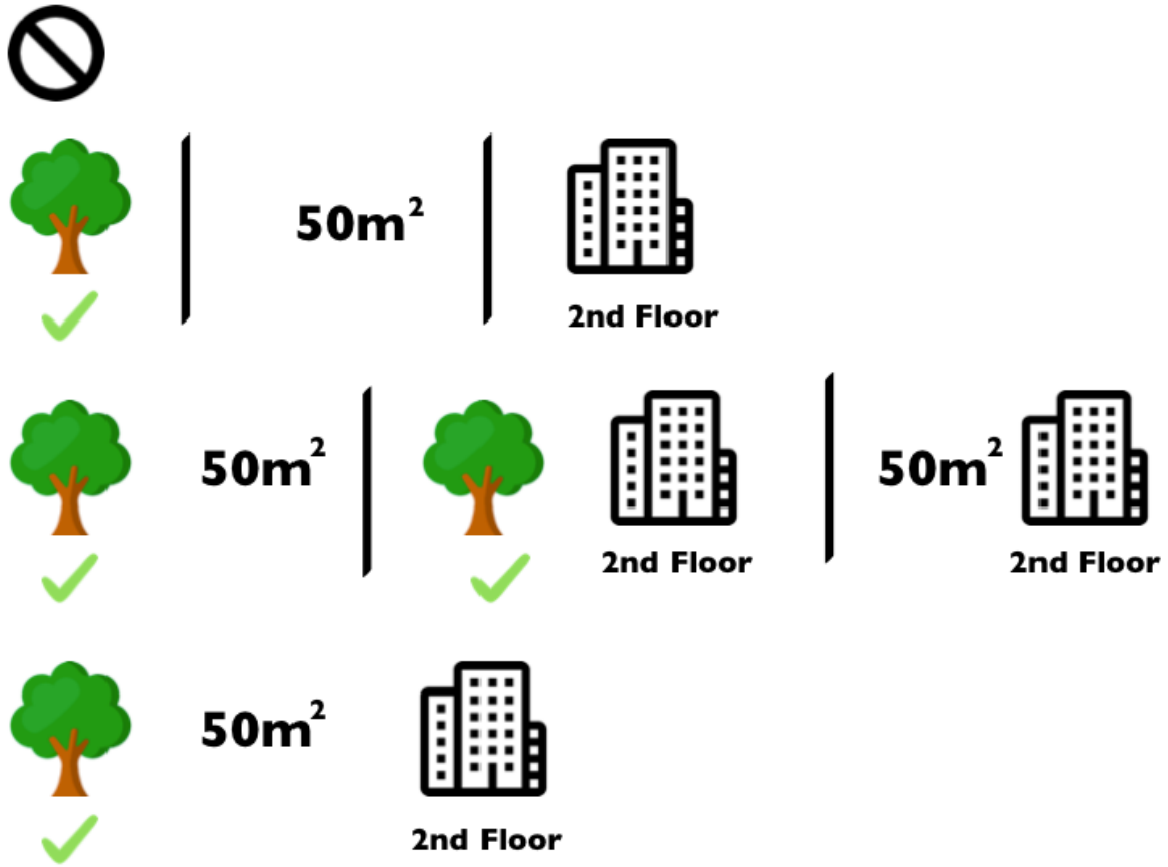
Game theory: Shapley Values

A prediction can be explained by assuming that each feature value of the instance is a “player” in a game where the prediction is the payout.



Credit: https://christophm.github.io/interpretable-ml-book/

Game theory: Shapley Values



Credit: https://christophm.github.io/interpretable-ml-book/

Game theory: Shapley Values

Advantages:

- **Fairly distributed** among the feature values of the instance

- Theoretical Guarantee

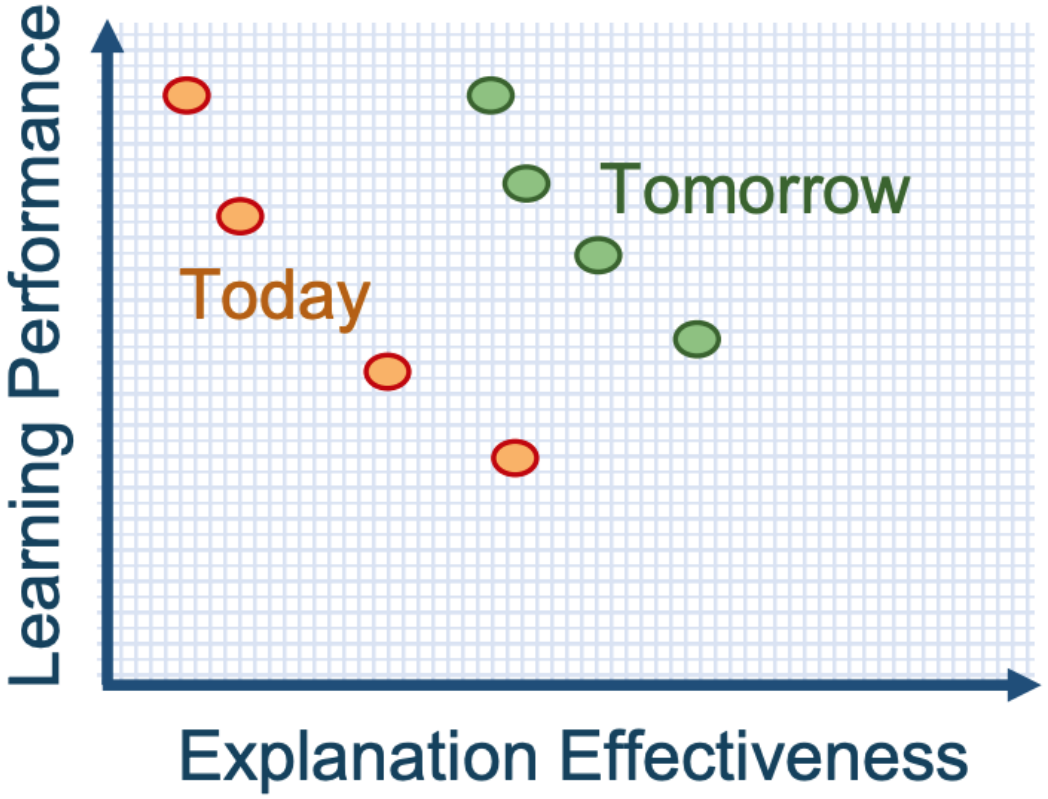
Disadvantages:

- Requires **a lot of computing time**

Challenges

Challenges

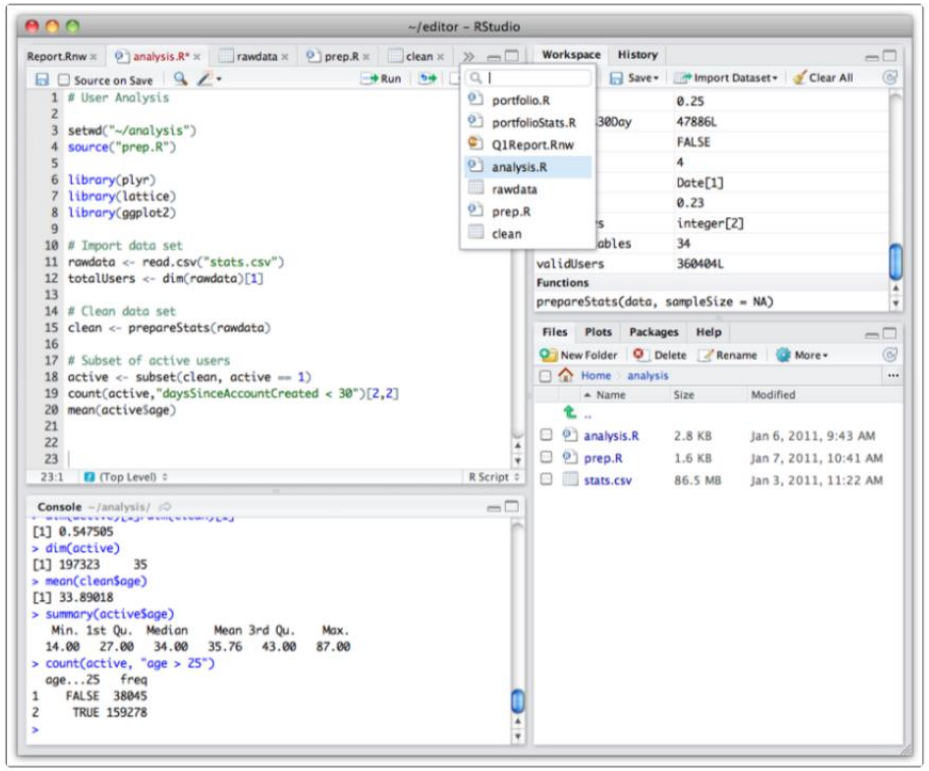
- **Optimal balance** between the interpretability and performance of ML models.



Credit: https://www.infoq.com/articles/explainable-ai-xai

Challenges

- Integration of XAI into Integrated Development Environments for machine learning



Credit: https://medium.com/analytics-vidhya/top-8-ides-for-machine-learning-and-data-scienceyou-should-know-ed4c325f3bf2

Challenges

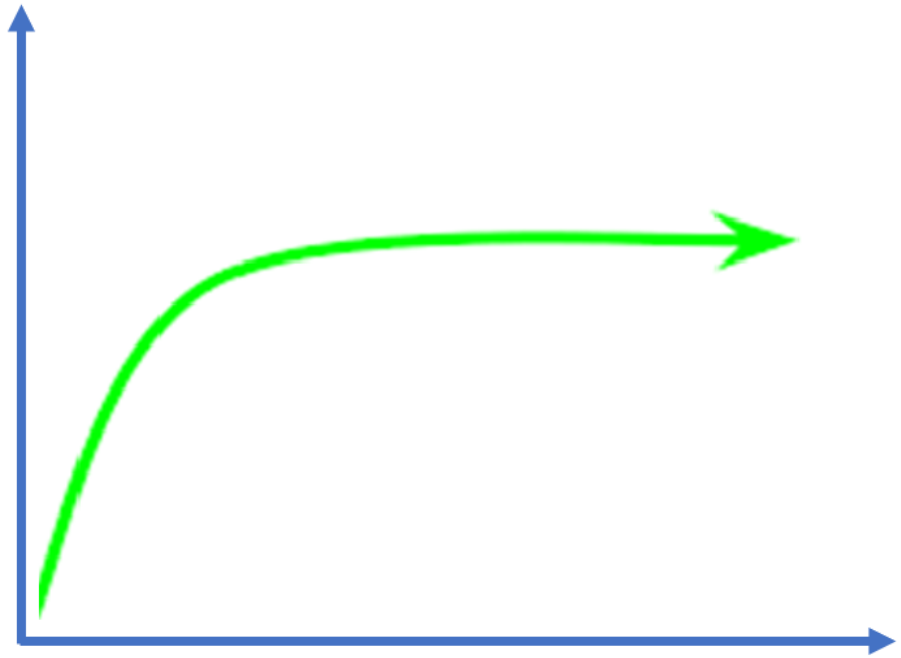
- Adversarial attacks on XAI

Recent studies showed that machine learning models can be manipulated (produce un-understandable explanations)

Credit: https://openai.com/blog/adversarial-example-research/

Challenges

- Efficiency and scalability of Explainable Methods



Conclusion

- Artificial Intelligence (AI) has become an important component of many software applications.
- The success of mode AI systems is based on Deep Learning
- Deep learning could achieve high accuracy, but unable to interpret their decisions to users.
- There is a pressing need for XAI, especially in sensitive domains (e.g., cybersecurity, health).

Responsible AI: Mitigation of Biases in Machine Learning

Study the paper titled: Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems

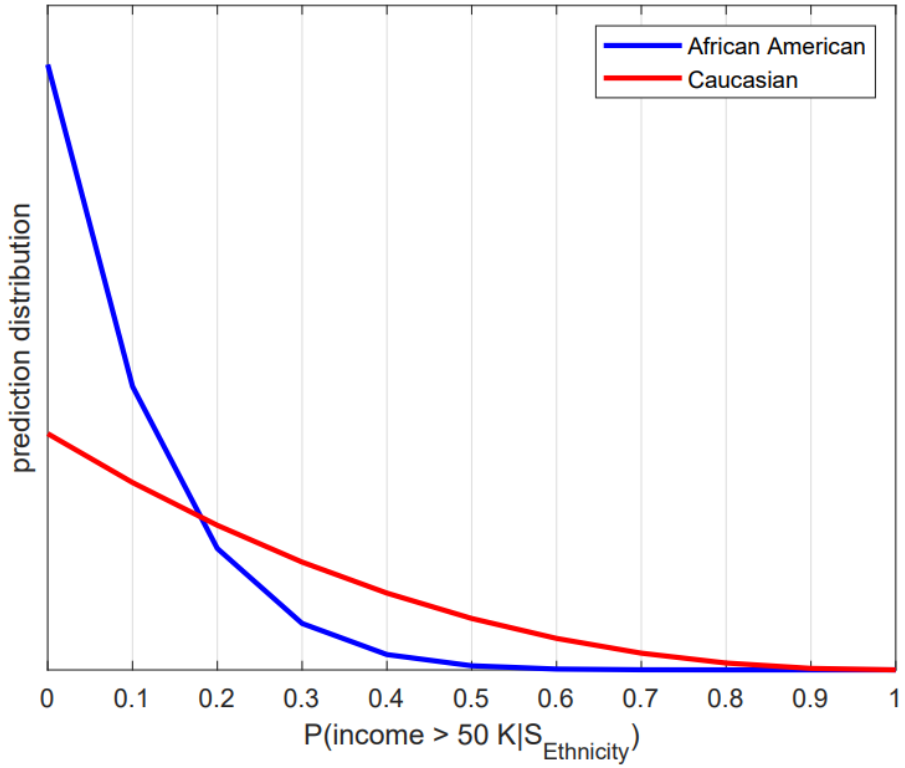
Bias?

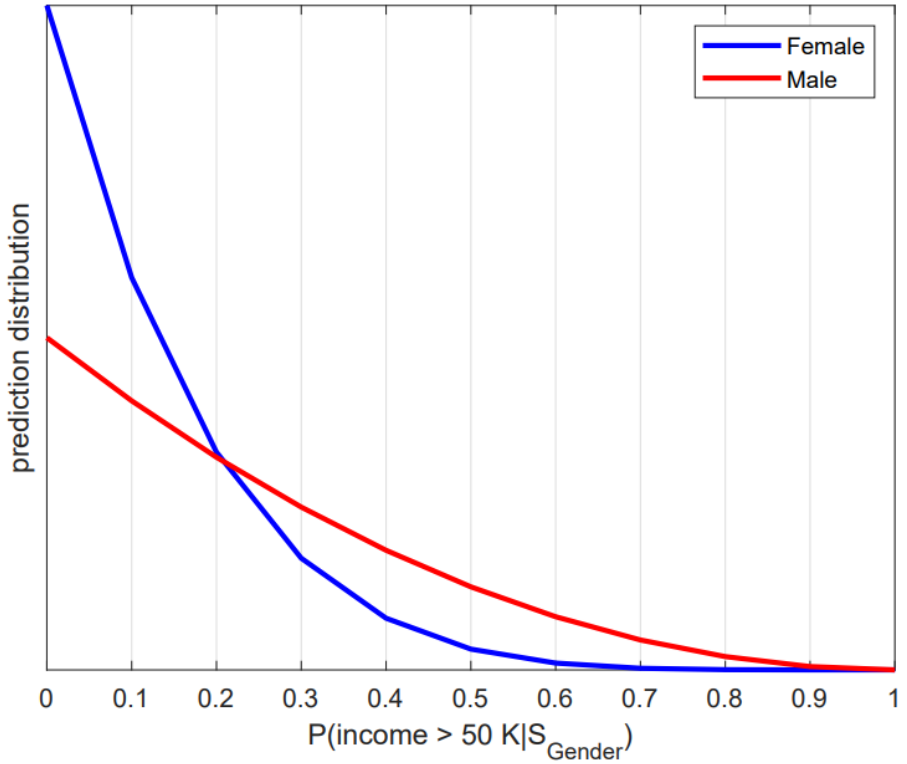
- The hazard of bias becomes even more crucial when these systems are applied to **critical and sensitive domains** such as health care and criminal justice.

- Biased AI systems are mainly engendered by the data used to feed **the training process of the machine learning algorithms**

- Training data can be incomplete, insufficiently diverse, biased, and/or consisting of non-representative samples that are not well (or poorly) defined before use, which might lead to biased results and lower accuracy

Dataset: adult UCI dataset

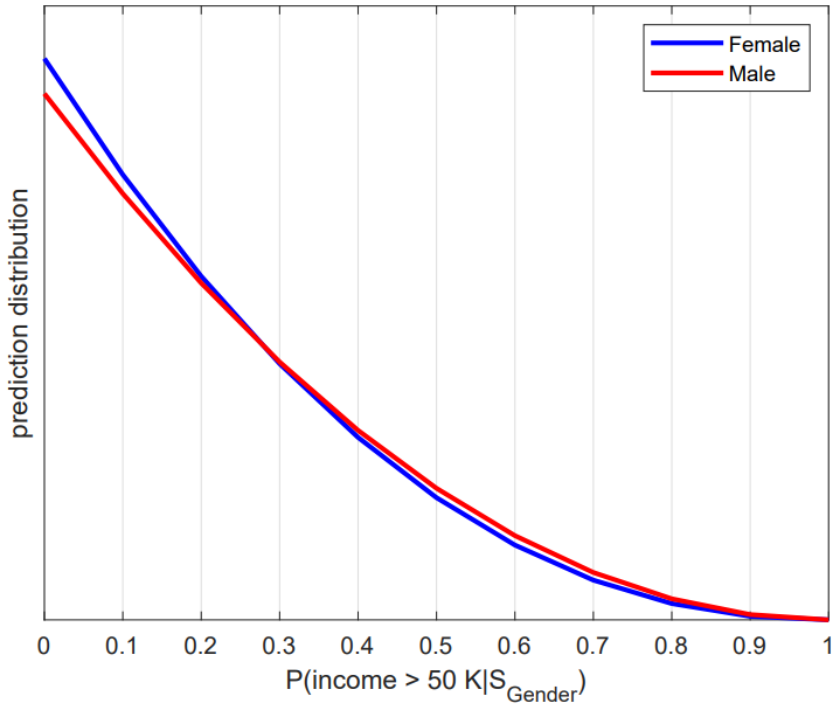


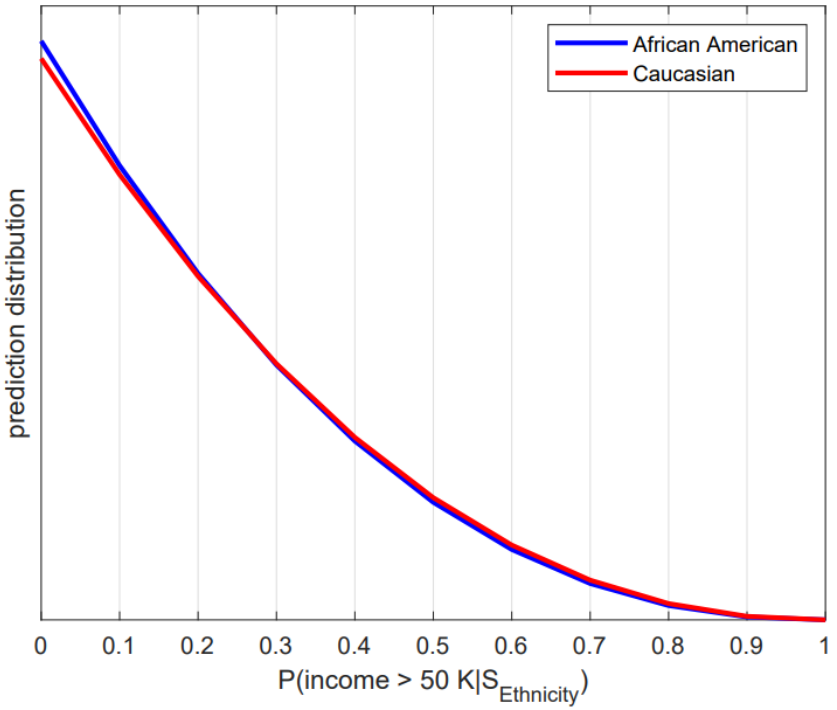


(Left) Prediction distribution of the original training data with respect to the ethnicity attribute. (Right) Prediction distribution of the original training data with respect to the gender attribute.

These results suggest that when a person is an “African American”, the probability that the classifier will predict his/her income below 50K$ is much higher compared to a “Caucasian”. Similarly, the results shows the distributions of the predicted P(income > 50K$) given the SA SGender = {female,male}. The Figure shows that for the gender attribute, the prediction distribution of a “female” has a large value at the low interval of [0.1 - 0.2] compared to a “male”. These results suggest that when a person is “female”, the probability that the classifier will predict her income below 50 K$ is much higher compared to a “male”.

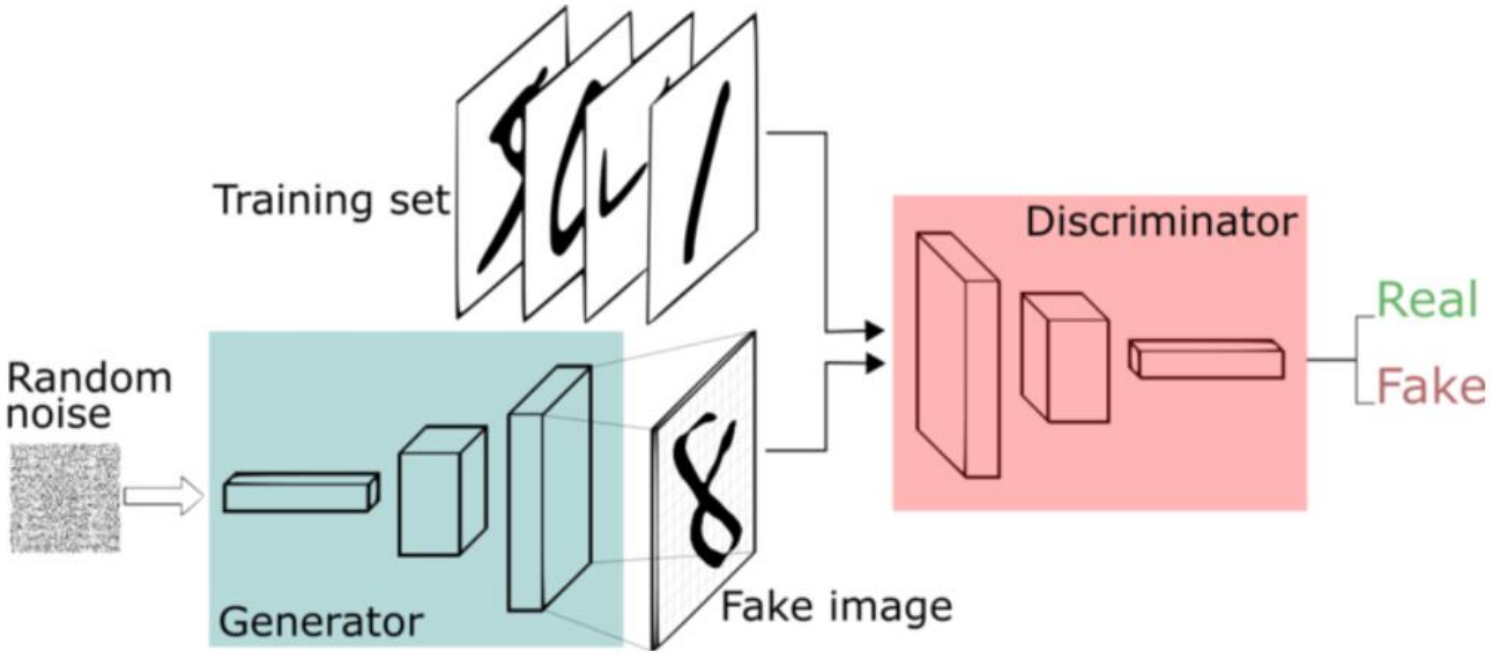
Dataset: adult UCI dataset





(Left) Prediction distribution when 85% of new synthetic data (female) were added to the original dataset. (Right) Prediction distribution when 85% of new synthetic data (African American) were added to the original dataset.

Generative Adversarial Networks [Goodfellow et al., 2014]



Generative Adversarial Network framework.

Source: https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394/