Machine Learning Systems Design

Modeling Pipeline

Lecture 16: Interpretability and Explainability

CE 40959 Spring 2023
Ali Zarezade
SharifMLSD.github.io

Agenda

1. Explainable AI (XAI)
2. Responsible AI

1. Explainable AI (XAI)

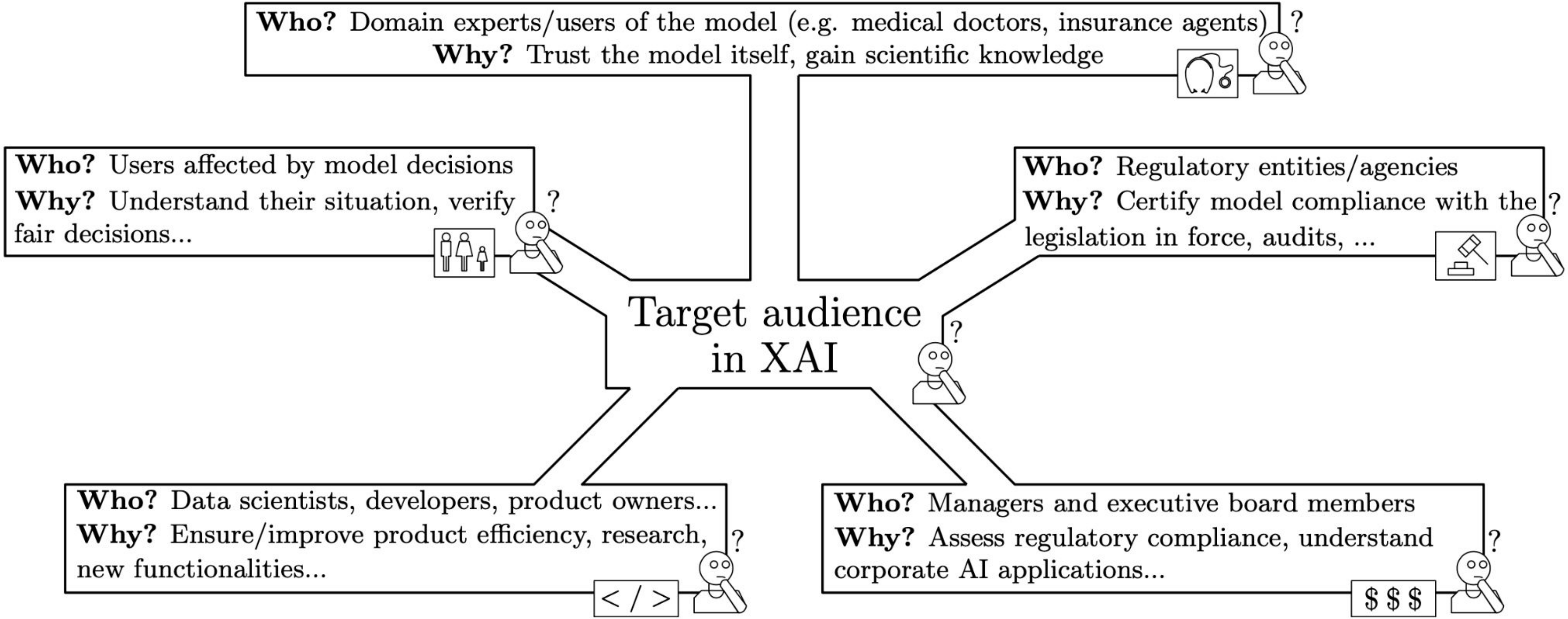
XAI: Explainable AI

The field of XAI allows AI models to be more transparent, providing explanation of their decisions in some level of details to:

- ensure algorithmic **fairness**
- identifying potential **bias**
- ensure model works as expected (**transparency** and **accountability**)

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI
Interpratable ml book

Target audience in XAI



Target audience in XAI

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge

Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...

Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...

Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...

Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...

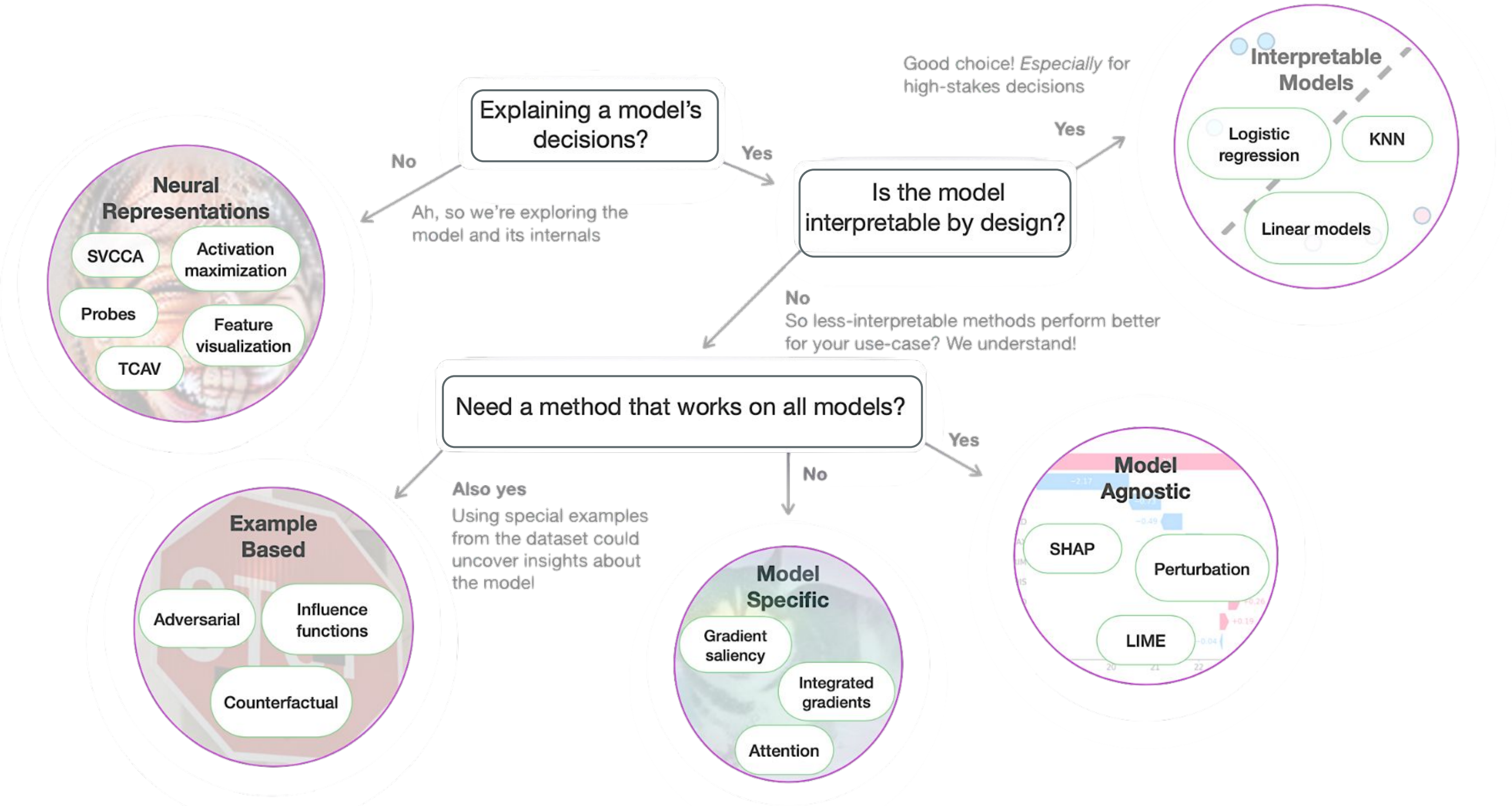
XAI goals

- **Trustworthiness**: the primary aim of an XAI model
    - Domain experts, users of the model affected by decisions
- **Causality**: finding causality among data variables
    - Domain experts, managers and executive board members, regulatory entities/agencies
- **Transferability**: the ability of to reuse knowledge in another problem
    - Domain experts, data scientists
- **Confidence**: a generalization of robustness and stability
    - Domain experts, developers, managers, regulatory entities/agencies
- **Fairness**
    - Users affected by model decisions, regulatory entities/agencies

XAI goals

- **Accessibility**: get more involved in the process of developing ML models
  - Product owners, managers, users affected by model decisions
- **Interactivity**: the ability of a model to be interactive with the user
  - Domain experts, users affected by model decisions
- **Privacy awareness**: the ability to explain the inner relations of a trained model by non-authorized third parties
  - Users affected by model decisions, regulatory entities/agencies

Model interpretation methods



Explaining a model's decisions?

No

Ah, so we're exploring the model and its internals

Neural Representations

SVCCA

Activation maximization

Probes

Feature visualization

TCAV

Yes

Is the model interpretable by design?

Yes

Good choice! *Especially* for high-stakes decisions

Interpretable Models

Logistic regression

KNN

Linear models

No

So less-interpretable methods perform better for your use-case? We understand!

Need a method that works on all models?

Also yes

Using special examples from the dataset could uncover insights about the model

Example Based

Adversarial

Influence functions

Counterfactual

No

Model Specific

Gradient saliency

Integrated gradients

Attention

Yes

Model Agnostic

SHAP

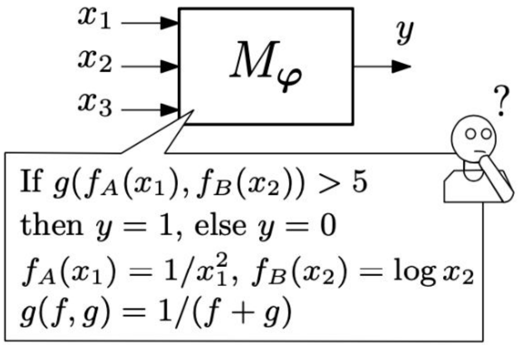
Perturbation

LIME

XAI taxonomy

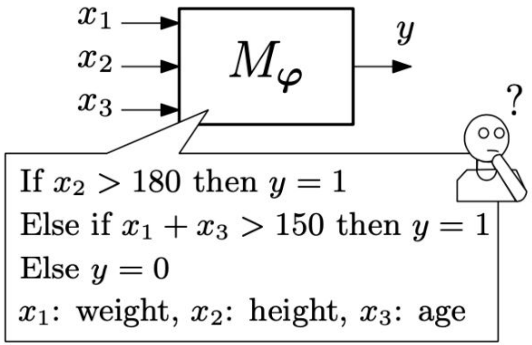
- Transparent models
  - Linear regression, decision tree, etc.
- Post-hoc explainability
  - Model agnostic
    - Explanation by simplification
    - Feature relevance explanation
    - Local explanations
    - Visual explanation
  - Model specific
    - Ensembles
    - SVM
    - MLP
    - CNN
    - RNN

Levels of transparency in ML models



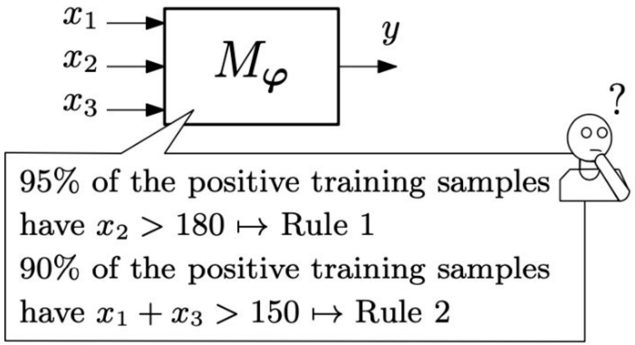
If $g(f_A(x_1), f_B(x_2)) > 5$
then $y = 1$, else $y = 0$
$f_A(x_1) = 1/x_1^2$, $f_B(x_2) = \log x_2$
$g(f, g) = 1/(f + g)$

simulatability



If $x_2 > 180$ then $y = 1$
Else if $x_1 + x_3 > 150$ then $y = 1$
Else $y = 0$
x_1: weight, x_2: height, x_3: age

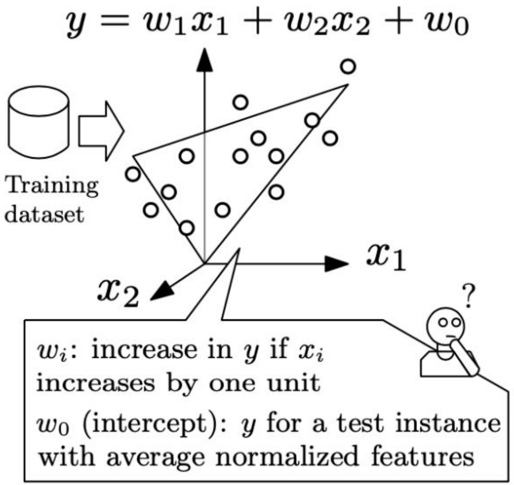
decomposability



95% of the positive training samples
have $x_2 > 180 \mapsto$ Rule 1
90% of the positive training samples
have $x_1 + x_3 > 150 \mapsto$ Rule 2

algorithmic transparency

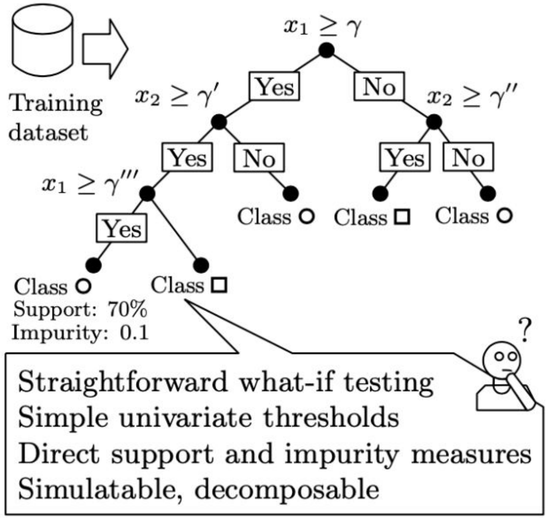
Transparent ML models



w_i: increase in y if x_i increases by one unit

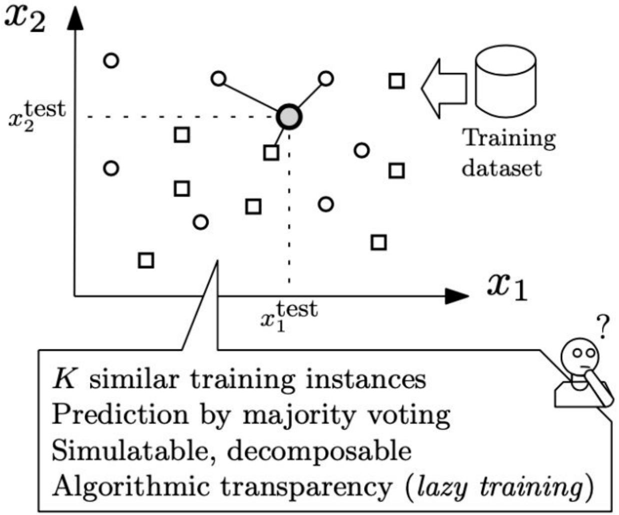
w_0 (intercept): y for a test instance with average normalized features

Linear regression



Straightforward what-if testing
Simple univariate thresholds
Direct support and impurity measures
Simulatable, decomposable

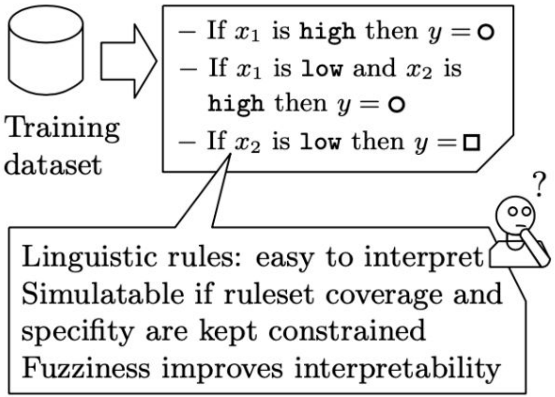
Decision trees



K similar training instances
Prediction by majority voting
Simulatable, decomposable
Algorithmic transparency (*lazy training*)

K-Nearest Neighbors

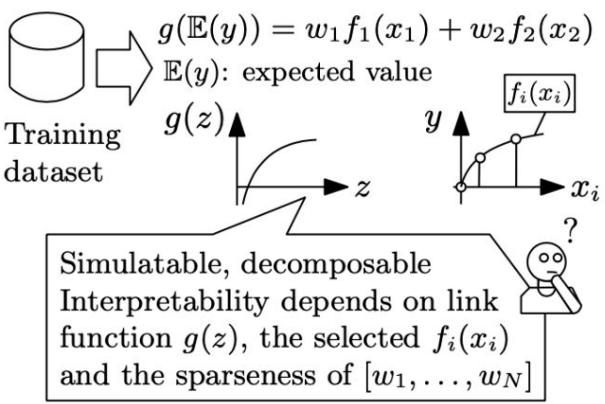
Transparent ML models



- If x_1 is **high** then $y = \bigcirc$
- If x_1 is **low** and x_2 is **high** then $y = \bigcirc$
- If x_2 is **low** then $y = \square$

Linguistic rules: easy to interpret
Simulatable if ruleset coverage and specifity are kept constrained
Fuzziness improves interpretability

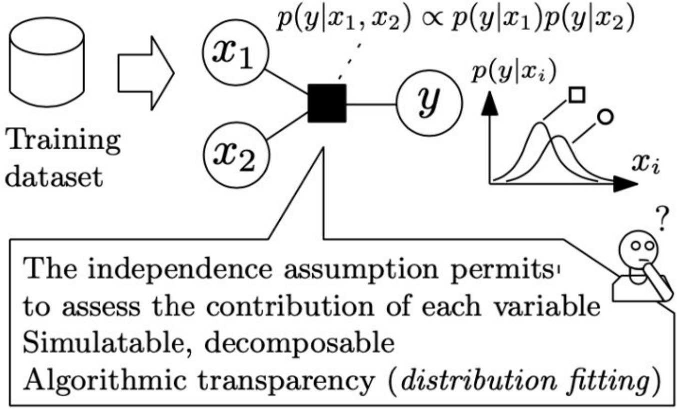
Rule-based Learners



$g(\mathbb{E}(y)) = w_1 f_1(x_1) + w_2 f_2(x_2)$
$\mathbb{E}(y)$: expected value

Simulatable, decomposable
Interpretability depends on link function $g(z)$, the selected $f_i(x_i)$ and the sparseness of $[w_1, \ldots, w_N]$

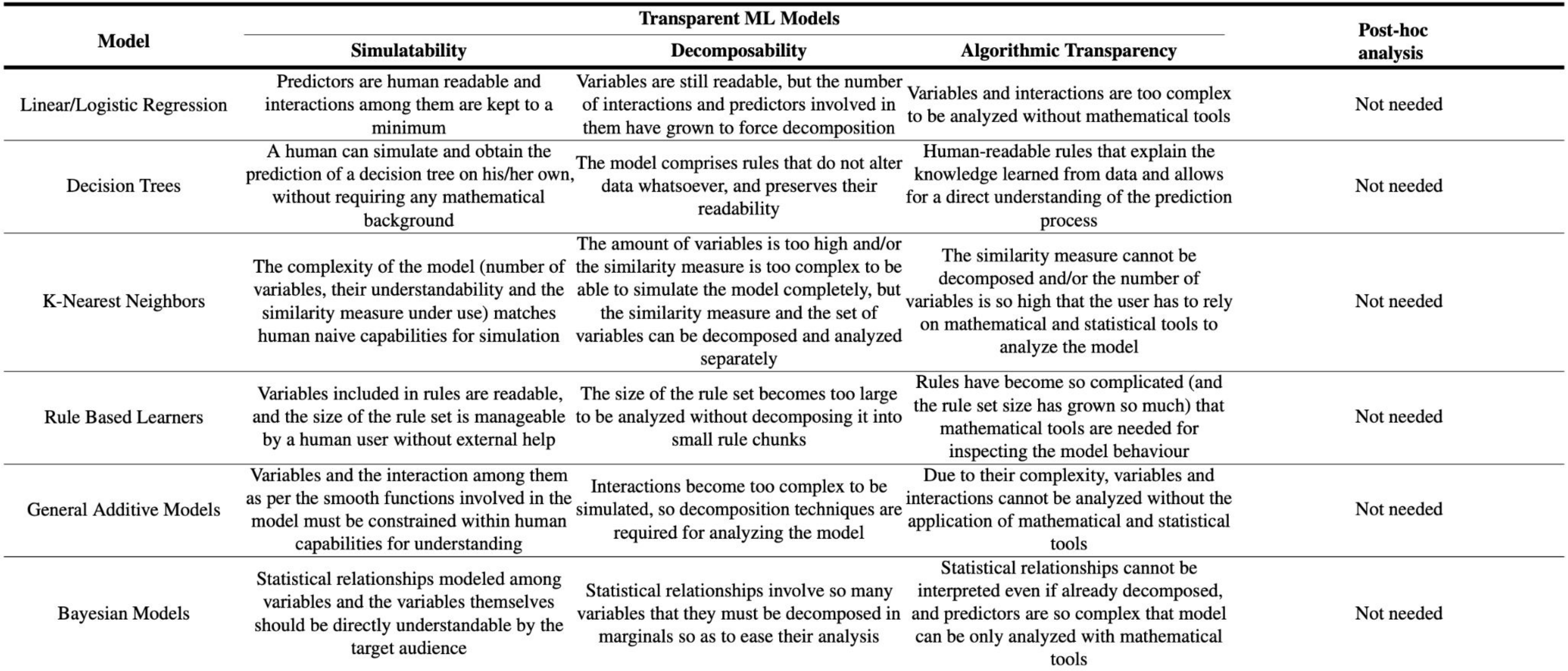
Generalized Additive Models



The independence assumption permits to assess the contribution of each variable
Simulatable, decomposable
Algorithmic transparency (*distribution fitting*)

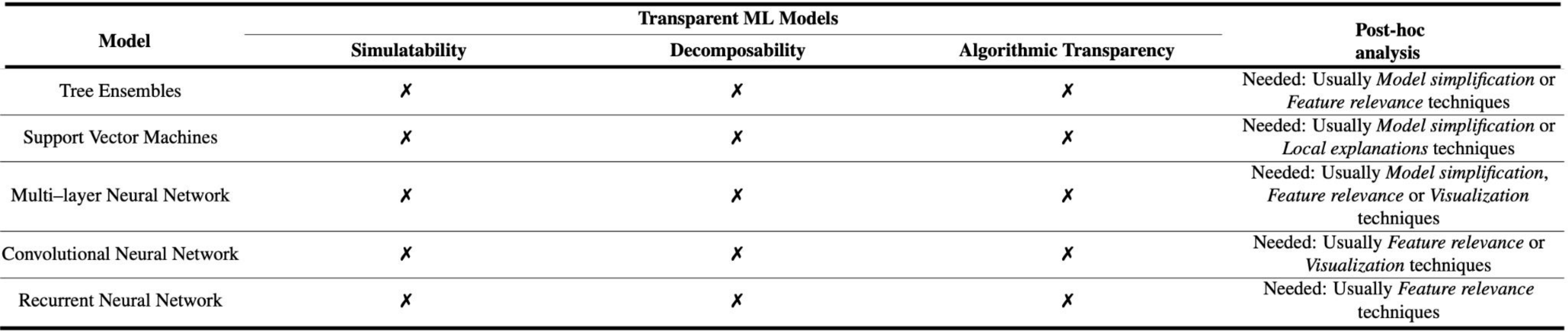
Bayesian Models

Transparent ML models



Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed

Transparent ML models



Model	**Transparent ML Models**			**Post-hoc analysis**
	Simulatability	**Decomposability**	**Algorithmic Transparency**	
Tree Ensembles	✗	✗	✗	Needed: Usually *Model simplification* or *Feature relevance* techniques
Support Vector Machines	✗	✗	✗	Needed: Usually *Model simplification* or *Local explanations* techniques
Multi–layer Neural Network	✗	✗	✗	Needed: Usually *Model simplification*, *Feature relevance* or *Visualization* techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* or *Visualization* techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* techniques

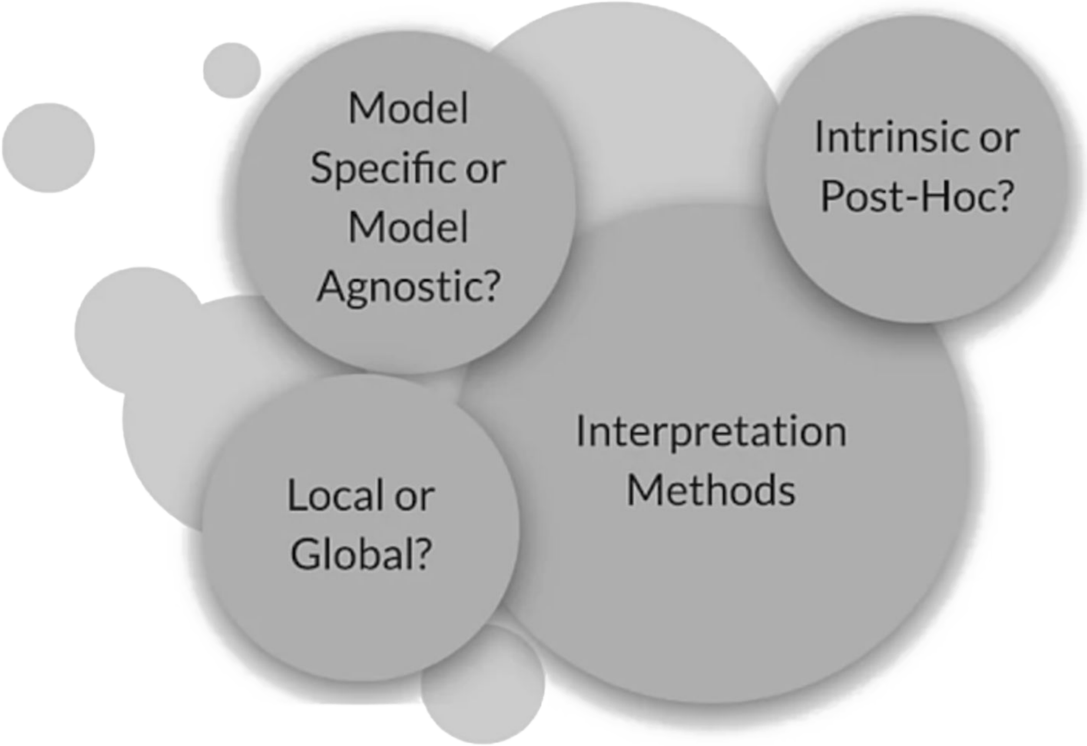
What is interpretability?

Models are interpretable if their operations can be understood by a human either through introspection or through a produced explanation.

Interpretability vs explainability

- **Interpretability** requires observing the inner mechanics of the model, such as its weights, features and parameters. It implies that the model is simple, linear or deterministic enough to be fully transparent.
- **Explainability** does not require accessing the inner mechanics of the model, but rather uses external methods, such as visualizations, statistics or surrogate models. It implies that the model is complex, nonlinear or stochastic enough to be partially opaque.

Various aspects of interpretation methods



Model Specific or Model Agnostic?

Intrinsic or Post-Hoc?

Local or Global?

Interpretation Methods

Intrinsic or post-hoc?

Intrinsic interpretability (transparent) methods are those that use simple models that are easy to understand, such as:

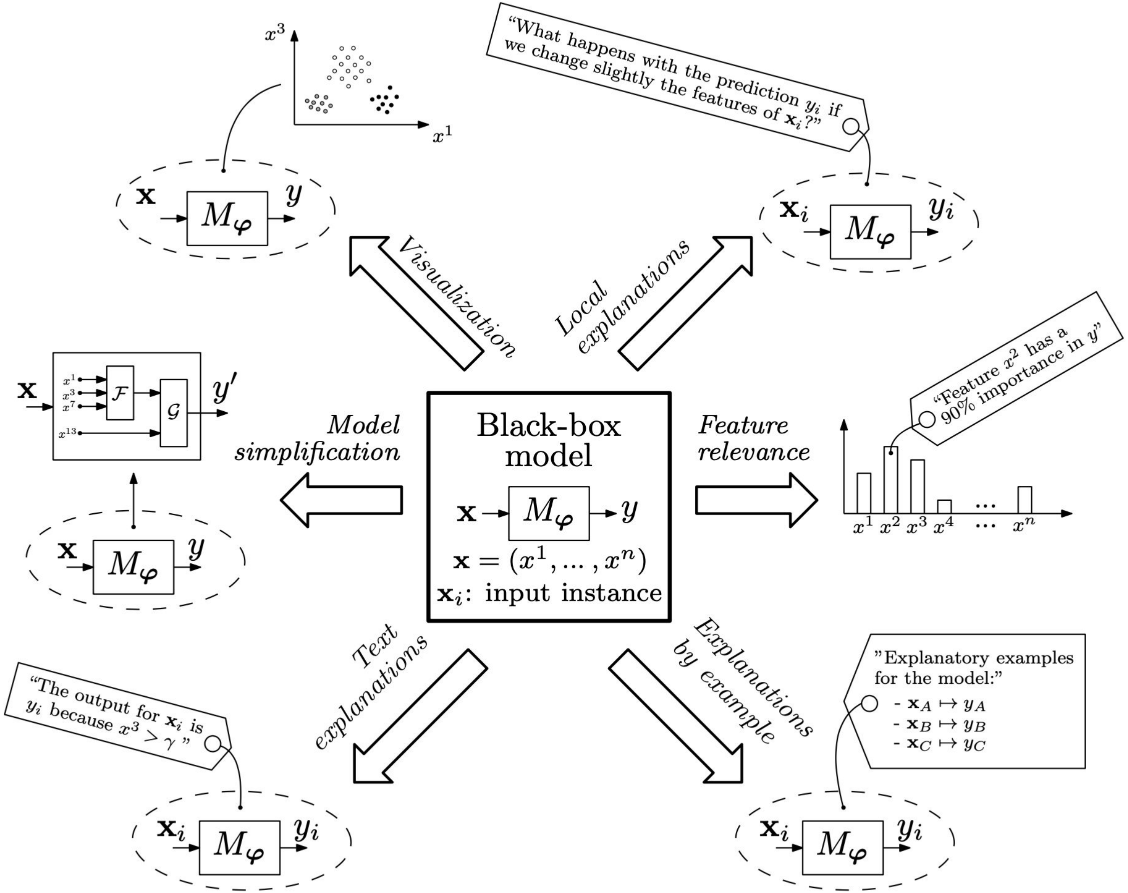
- *Linear models*: These models use a linear combination of features to make predictions, and the weights can be interpreted as the importance or effect of each feature.
- *Decision trees*: These models use a series of binary splits based on features to make predictions, and the tree structure can be visualized and traced to understand the logic behind each decision.
- *Rule-based models*: These models use a set of if-then rules to make predictions, and the rules can be inspected and verified by humans.

Intrinsic or post-hoc?

Post-hoc interpretability methods are those that apply interpretation techniques after model training, such as:

- *Feature importance*: These methods measure how much each feature contributes to the model's prediction, either globally (for the whole dataset) or locally (for a specific instance).
- *Partial dependence plots*: These methods show how the model's prediction changes as a function of a single feature or a pair of features, while averaging out the effects of other features.
- *Counterfactual explanations*: These methods find the minimal changes in the input features that would lead to a different prediction by the model, and provide a contrastive explanation for why the model made a certain decision.

Various interpretations in XAI methods

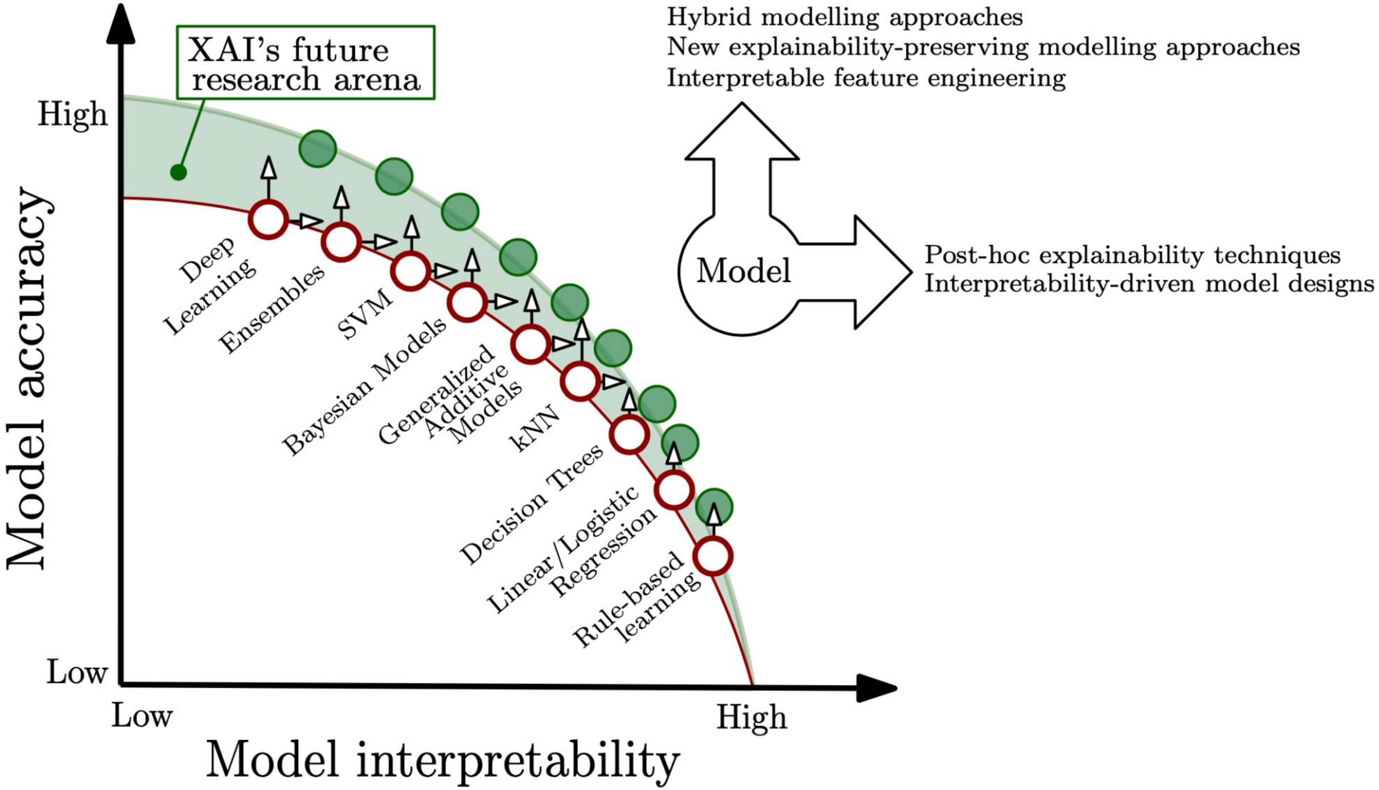

Black-box model
$\mathbf{x} \to M_{\varphi} \to y$
$\mathbf{x} = (x^1, \dots, x^n)$
$\mathbf{x}_i$: input instance
Visualization
x^3
x^1
Local explanations
"What happens with the prediction y_i if we change slightly the features of $\mathbf{x}_i$?"
Model simplification
x^1
x^3
x^7
x^{13}
$\mathcal{F}$
$\mathcal{G}$
y'
Feature relevance
"Feature x^2 has a 90% importance in y"
x^1 x^2 x^3 x^4 ... x^n
Text explanations
"The output for $\mathbf{x}_i$ is y_i because $x^3 > \gamma$ "
Explanations by example
"Explanatory examples for the model:"
- $\mathbf{x}_A \mapsto y_A$
- $\mathbf{x}_B \mapsto y_B$
- $\mathbf{x}_C \mapsto y_C$
$\mathbf{x}_i$ M_{φ} y_i

Intrinsic or post-hoc?

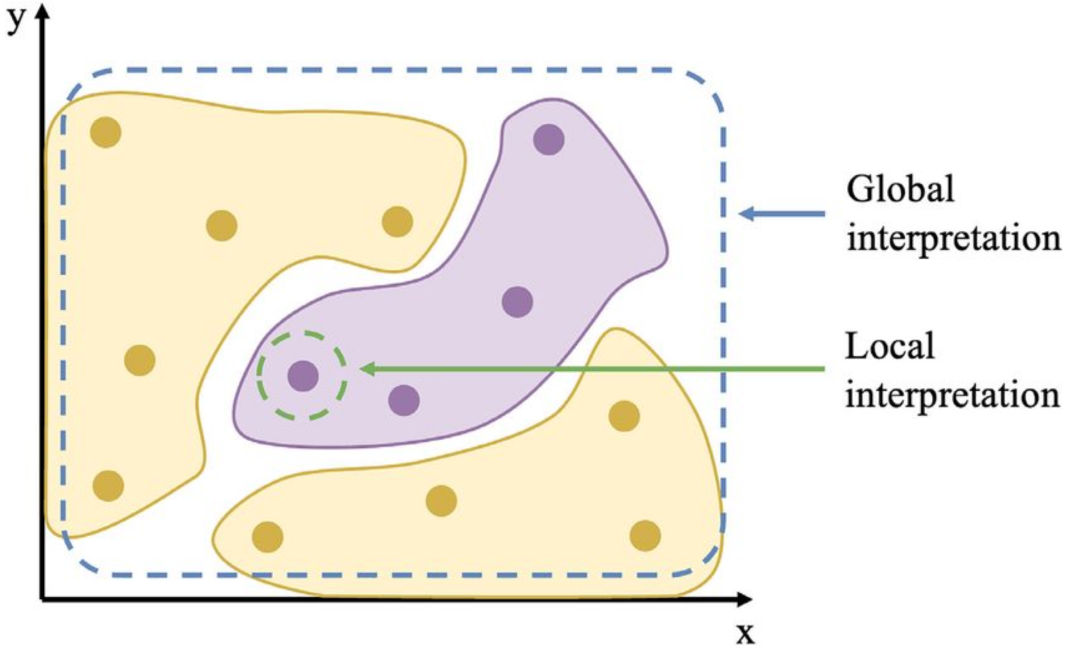
- **Intrinsic interpretability** refers to machine learning models that are considered interpretable due to their simple structure, such as short decision trees or sparse linear models.
    - Achieved at the cost of predictive performance.

- **Post-hoc interpretability** refers to the application of interpretation methods after model training, such as feature importance, partial dependence plots, or counterfactual explanations.
    - Treat mode as black-box
    - Model agnostic
    - Applied after training
    - May not always be reliable, accurate, or consistent

Model interpretability and performance trade-off

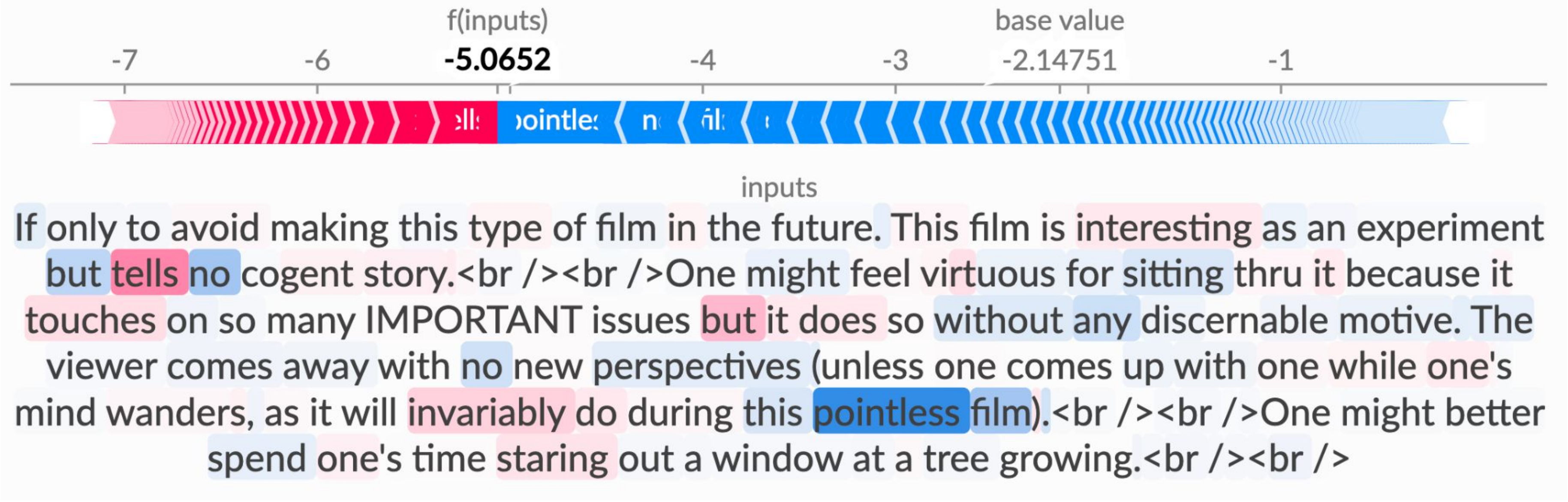

XAI's future research arena
High
Low
Model accuracy
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

Local or global?



Local or global?

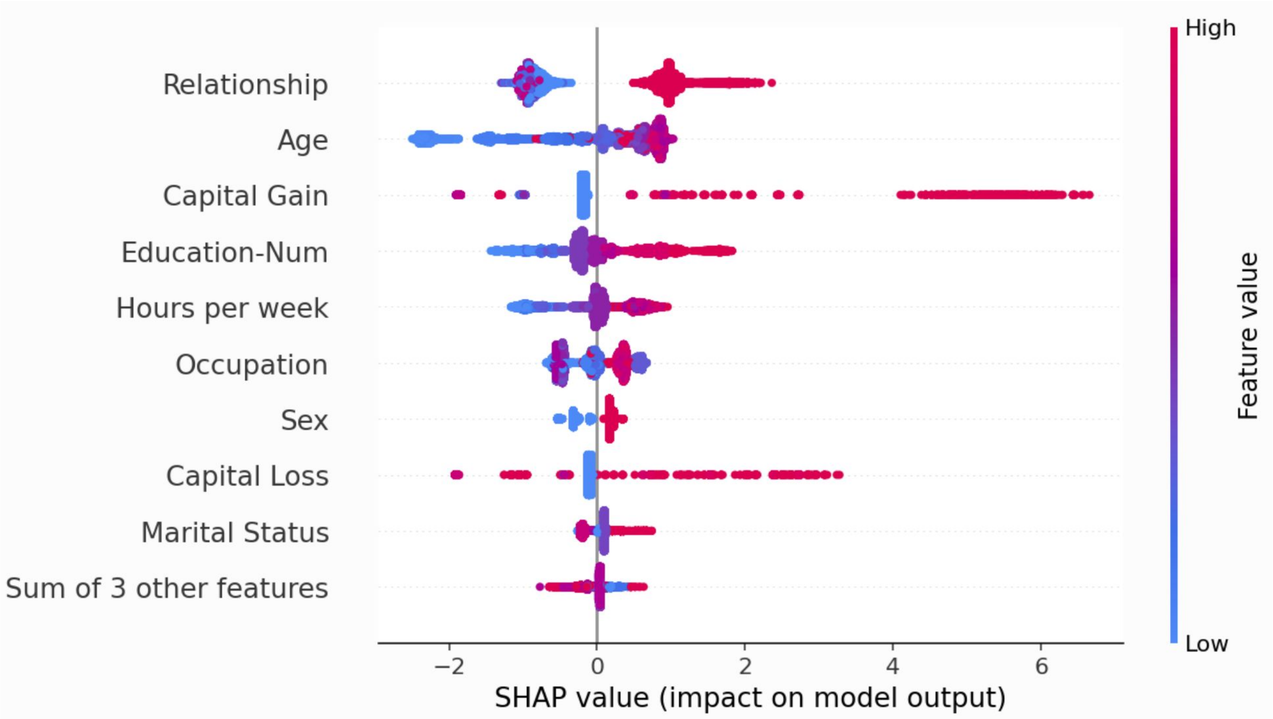
Local: explains an individual prediction



If only to avoid making this type of film in the future. This film is interesting as an experiment but tells no cogent story.<br /><br />One might feel virtuous for sitting thru it because it touches on so many IMPORTANT issues but it does so without any discernable motive. The viewer comes away with no new perspectives (unless one comes up with one while one's mind wanders, as it will invariably do during this pointless film).<br /><br />One might better spend one's time staring out a window at a tree growing.<br /><br />

Local or global?

Global: explains entire model prediction



SHAP: SHapley Additive exPlanations

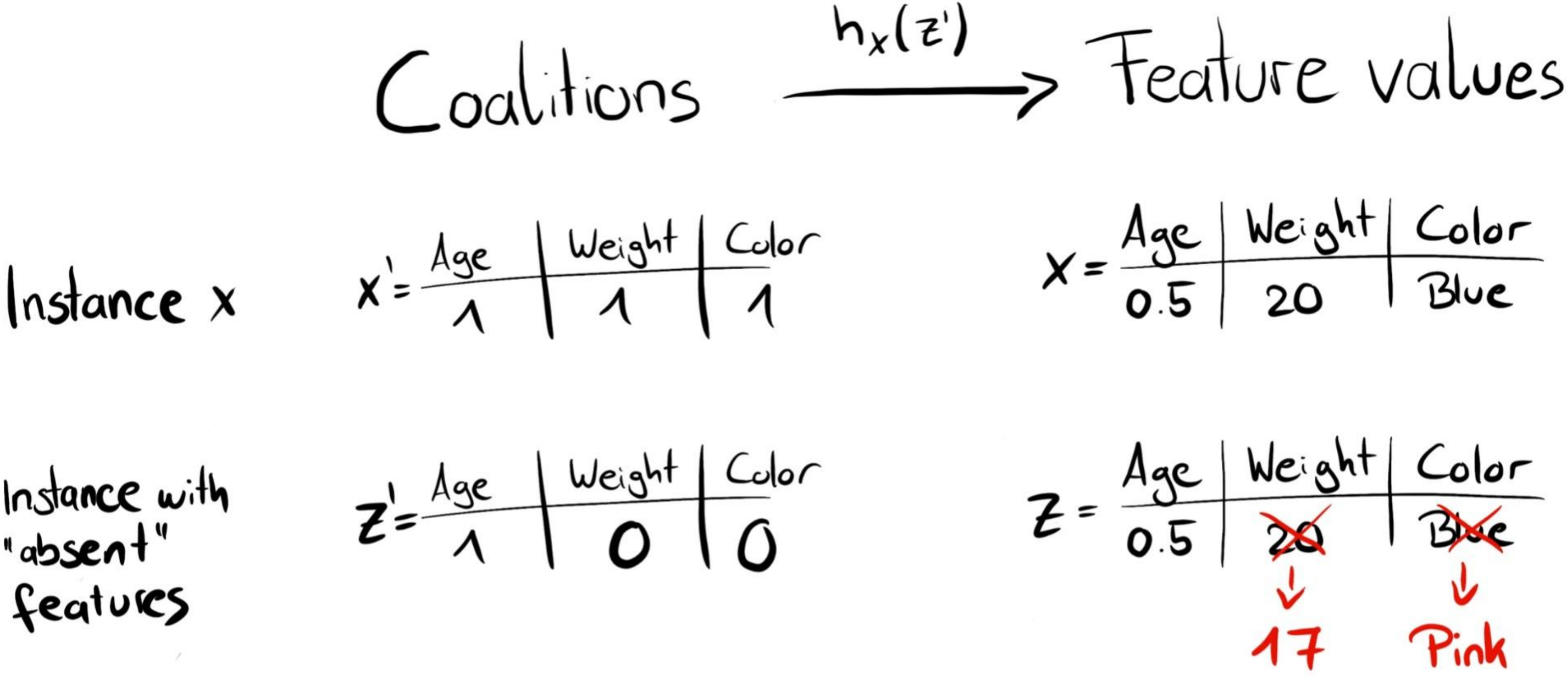
SHAP is a model-agnostic method that uses a game-theoretic approach to explain the output of any machine learning model.

SHAP tutorial

SHAP: SHapley Additive exPlanations

- Shapley values are a way of fairly distributing the payoff of a game among the players, based on their individual and joint contributions to the game outcome.
- In SHAP, the machine learning model is viewed as a game, where the features are the players and the prediction is the payoff.
- SHAP assigns each feature a Shapley value, which represents the average marginal contribution of that feature to the prediction across all possible subsets of features.
- Shapley values tell us how to fairly distribute the “payout” (= the prediction) among the features.

SHAP: SHapley Additive exPlanations



$x' =$

Age	Weight	Color
1	1	1

$x =$

Age	Weight	Color
0.5	20	Blue

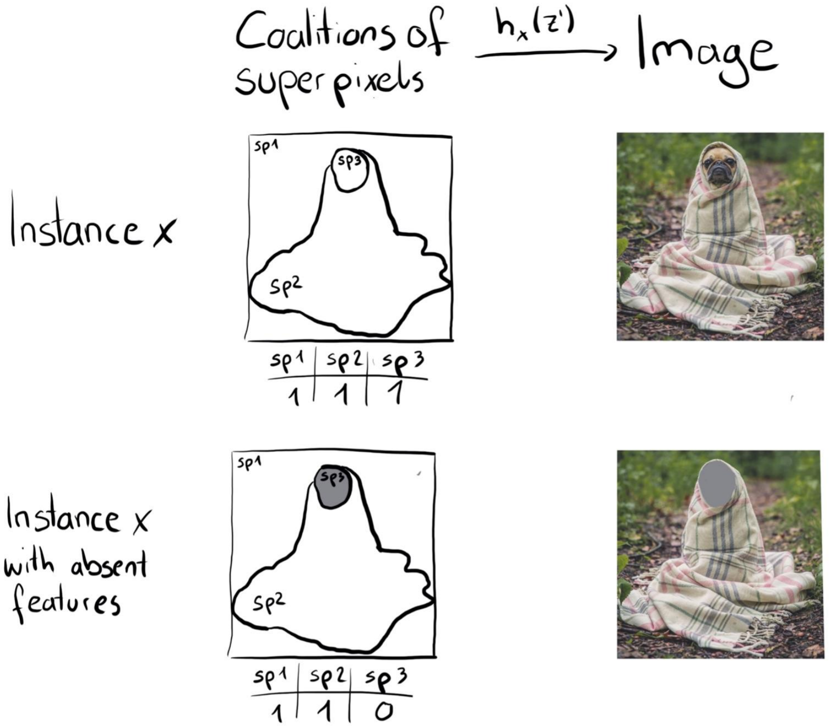
$z' =$

Age	Weight	Color
1	0	0

$z =$

Age	Weight	Color
0.5	~~20~~ ↓ 17	~~Blue~~ ↓ Pink

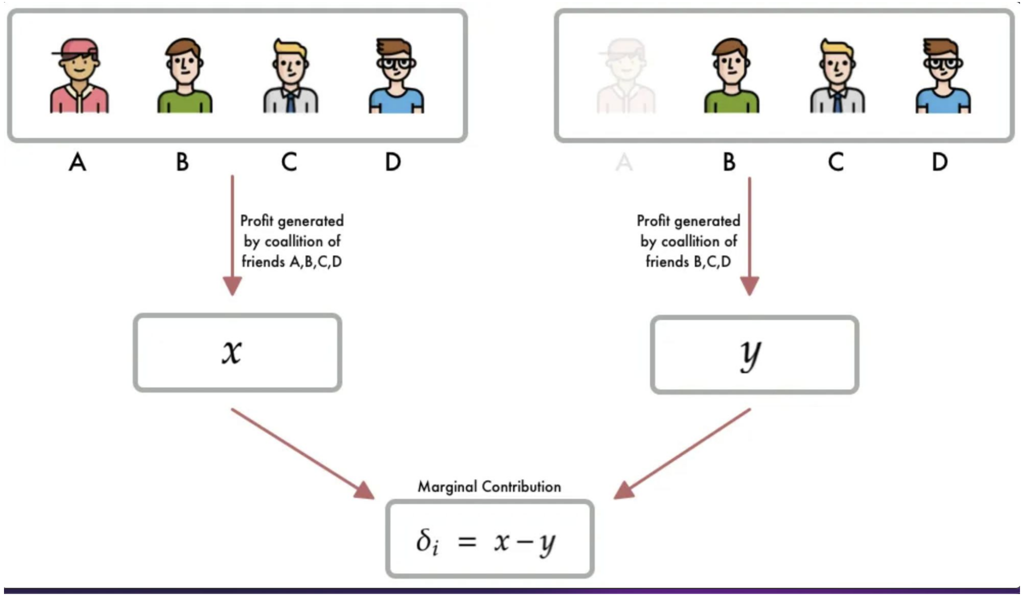
SHAP: SHapley Additive exPlanations



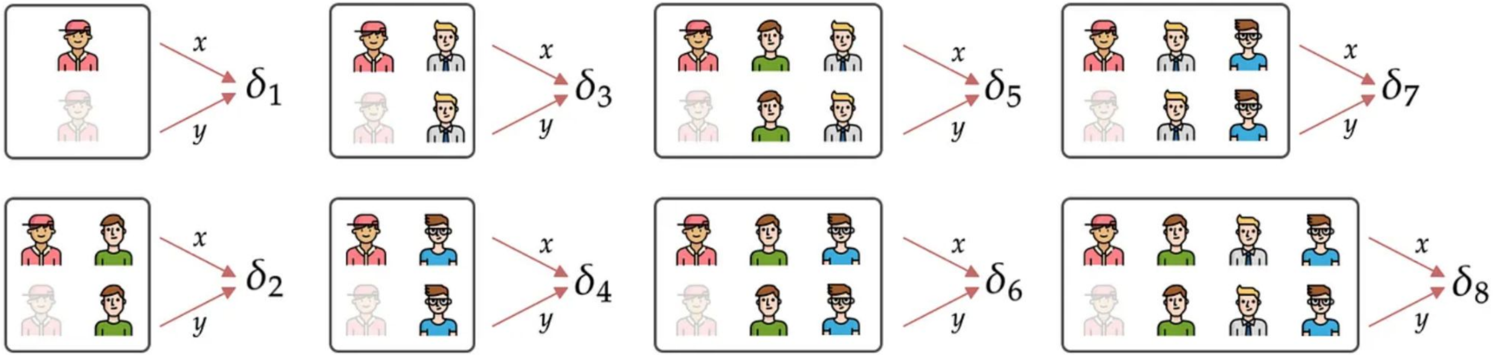
sp1	sp2	sp3
1	1	1

sp1	sp2	sp3
1	1	0

SHAP: SHapley Additive exPlanations


A
B
C
D
Profit generated by coallition of friends A,B,C,D
x
A
B
C
D
Profit generated by coallition of friends B,C,D
y
Marginal Contribution
$\delta_i = x - y$

SHAP: SHapley Additive exPlanations



The Shapley value for member

is given by:

$$\phi_i = \frac{\delta_1 + \delta_3 + \delta_4 + \delta_5 + \delta_6 + \delta_7 + \delta_8}{8}$$

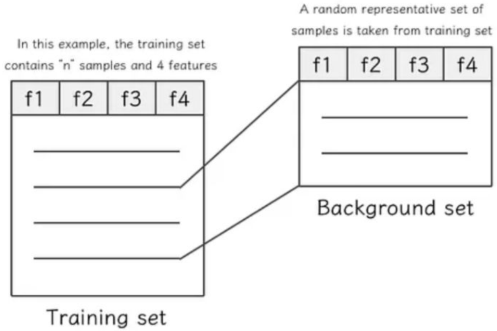
SHAP: SHapley Additive exPlanations

Say we want to explain feature "x":

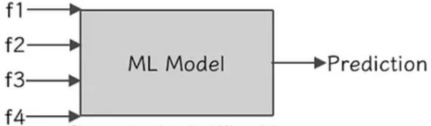
X =

f1	f2	f3	f4

We have training set and background set:

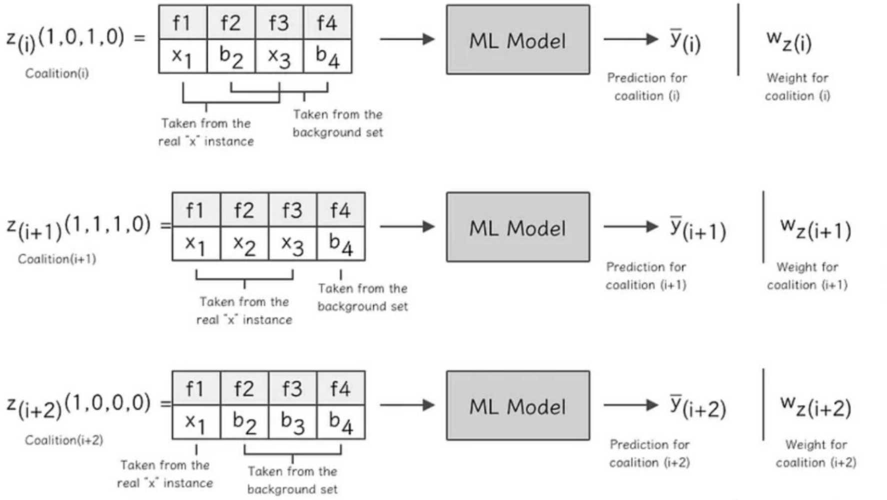

In this example, the training set contains "n" samples and 4 features
f1
f2
f3
f4
Training set
A random representative set of samples is taken from training set
f1
f2
f3
f4
Background set

And of course, an already trained ML model:



Given an already trained ML model, the expected value E[x] is obtained.In case of linear regression, E[x] is the average of all predictions in the training set.

Coalitions, predictions and weights are calculated:


z(i)(1,0,1,0) =
Coalition(i)
f1 f2 f3 f4
x1 b2 x3 b4
Taken from the real "x" instance
Taken from the background set
ML Model
ȳ(i)
Prediction for coalition (i)
wz(i)
Weight for coalition (i)
z(i+1)(1,1,1,0) =
Coalition(i+1)
f1 f2 f3 f4
x1 x2 x3 b4
Taken from the real "x" instance
Taken from the background set
ML Model
ȳ(i+1)
Prediction for coalition (i+1)
wz(i+1)
Weight for coalition (i+1)
z(i+2)(1,0,0,0) =
Coalition(i+2)
f1 f2 f3 f4
x1 b2 b3 b4
Taken from the real "x" instance
Taken from the background set
ML Model
ȳ(i+2)
Prediction for coalition (i+2)
wz(i+2)
Weight for coalition (i+2)

$$w_{c_i} = \frac{\#\, features\, - 1}{(\#\, coalitions\ of\ size\ |c_i|\ ) \times (\#\, features\ included\ in\ c_i) \times (\#\, features\ excluded\ in\ c_i)}$$

The weighted linear model is fit

With coalitions, predictions and weights, the weighted linear model is built.

Shapley values are obtained!

Once optimized the weighted linear model, the coefficients are the Shapley values!

SHAP: SHapley Additive exPlanations

The Shapley value satisfies four desirable properties:

- Efficiency: The sum of all Shapley values equals the total payoff of the grand coalition (all players).
- Symmetry: If two players contribute equally to every coalition, they have the same Shapley value.
- Dummy: If a player does not contribute to any coalition, their Shapley value is zero.
- Additivity: If the payoff function is the sum of two sub-functions, the Shapley value for each sub-function is also additive.

SHAP: SHapley Additive exPlanations

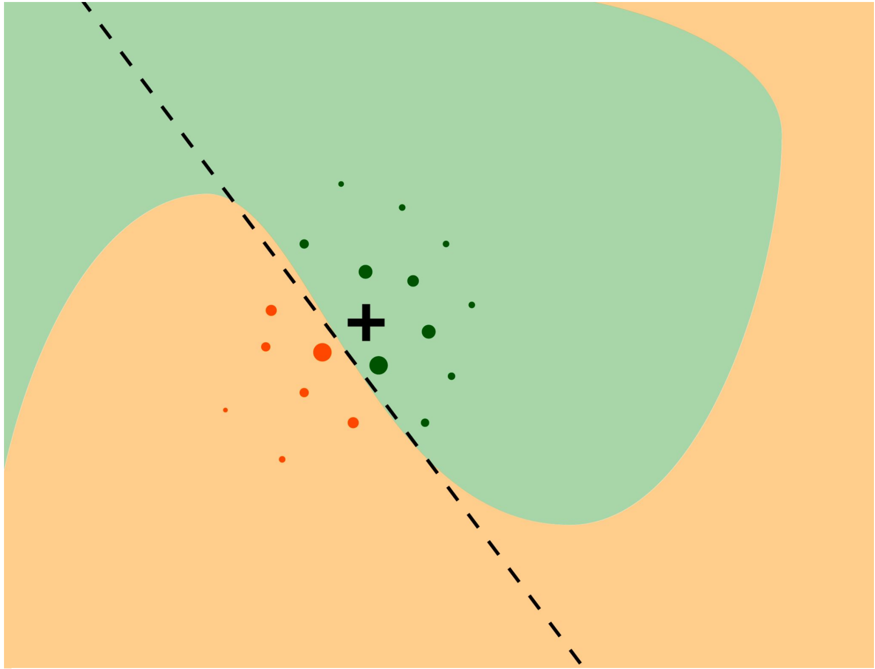
- The SHAP value can be interpreted as the difference between the expected prediction of the model and the expected prediction of the model conditioned on the feature.

Complexity of computing SHAP

- Computing SHAP values requires iterating over all possible coalitions of features and computing their marginal contributions to the prediction.
- This is a combinatorial problem that grows exponentially with the number of features.
- The exact complexity of computing SHAP values depends on the type of model and data distribution.
- In general, computing SHAP values is #P-hard, which means it is at least as hard as counting the number of solutions to an NP-hard problem.

LIME: Local Interpretable Model-agnostic Explanations

- A framework for explaining ML models by approximating them with simpler models that are easier to understand
- Can handle any type of model by treating it as a black box
- Provides local explanations around the vicinity of the instance being explained

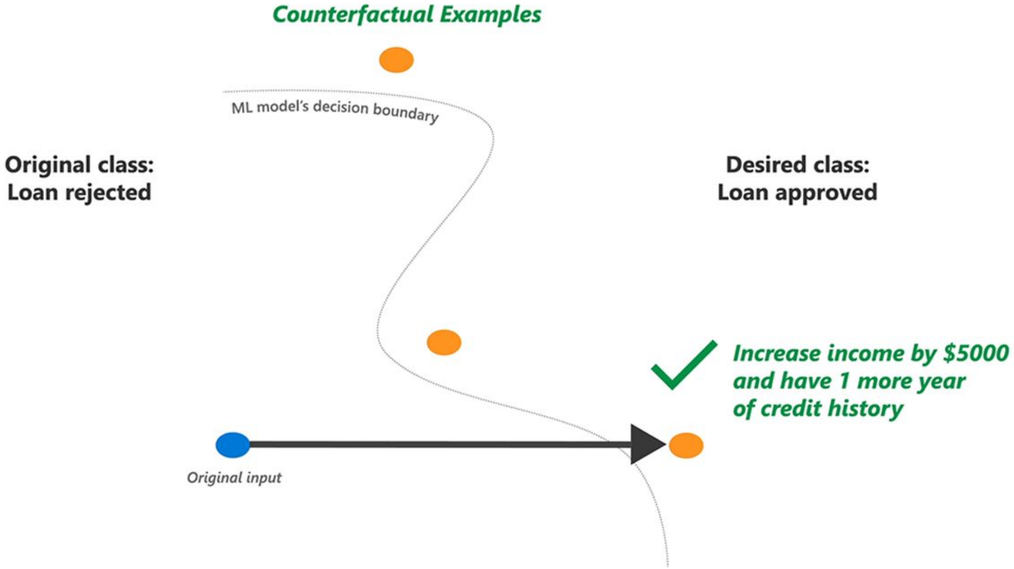


$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Counterfactual explanation

Counterfactual explanations are a way of explaining the predictions of a complex machine learning model by showing how the input features could be changed to get a different output.

For example, if a model predicts that a person will not get a loan, a counterfactual explanation could tell them what they need to do to get approved, such as increasing their income or lowering their debt.



Counterfactual explanation

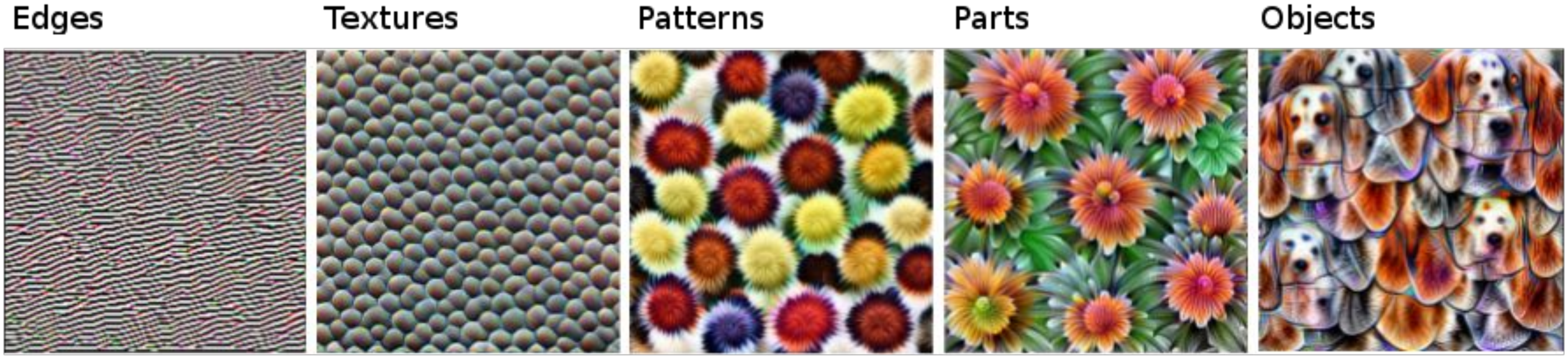
Counterfactual explanations are based on the idea of counterfactual reasoning, which is how humans think about alternative scenarios and outcomes. For example, if you miss your bus, you might think "If I had left home earlier, I would have caught the bus". This is a counterfactual statement that contrasts the actual situation with a hypothetical one.

In explainable AI, counterfactual explanations can help users understand why a model made a certain prediction, and what they can do to change it. They can also help developers debug and improve their models by identifying the most influential features and potential biases. Counterfactual explanations are model-agnostic, which means they can work with any type of machine learning model, such as deep neural networks or decision trees.

Neural Network Interpretation

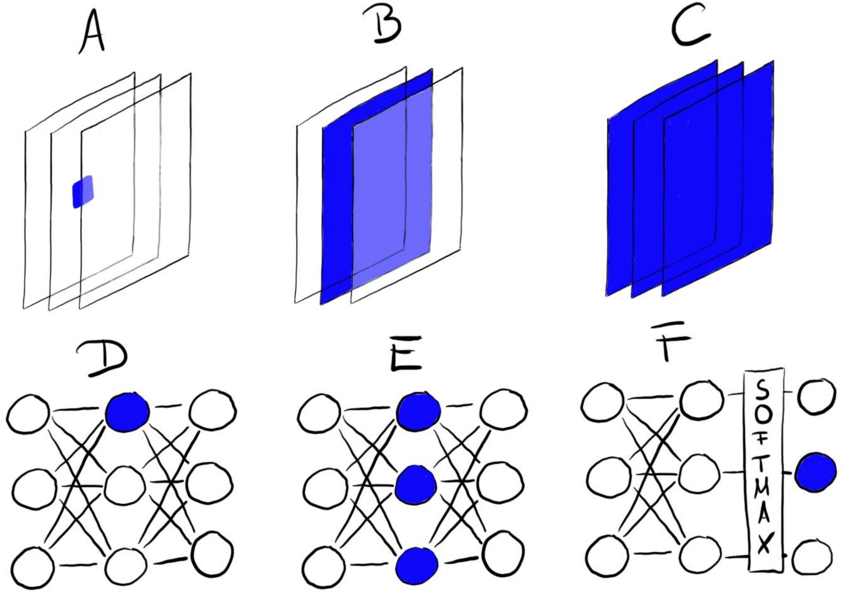
- Learned Features
- Feature Visualization

Learned Features



Features learned by a convolutional neural network (Inception V1) trained on the ImageNet data. The features range from simple features in the lower convolutional layers (left) to more abstract features in the higher convolutional layers (right)

Feature Visualization



Feature visualization can be done for different units. A) Convolution neuron, B) Convolution channel, C) Convolution layer, D) Neuron, E) Hidden layer, F) Class probability neuron (or corresponding pre-softmax neuron)

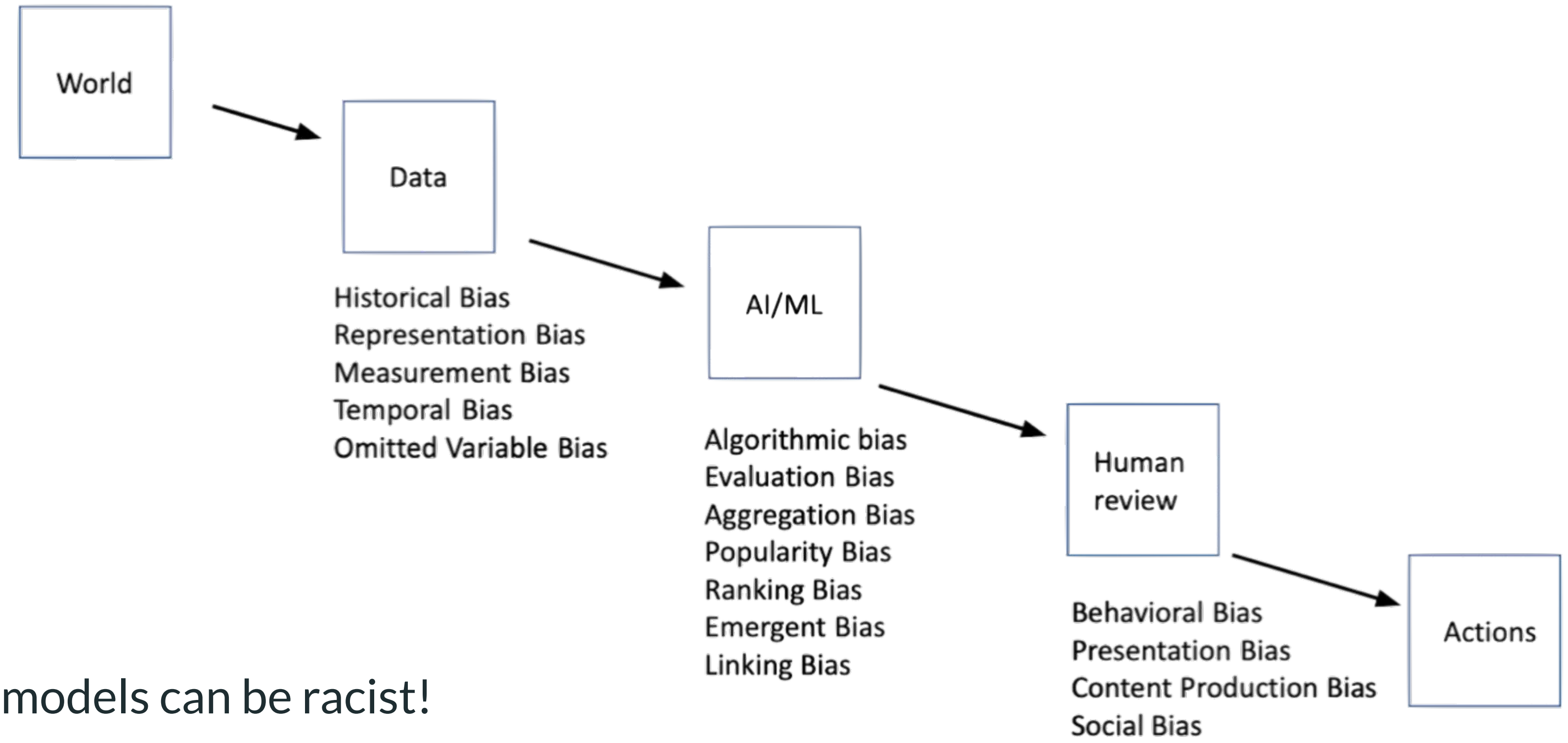
2. Responsible AI

What is responsible AI?

The practice of designing, developing, and deploying AI systems with good intention and sufficient awareness to empower users, to engender trust, and to ensure fair and positive impact to society. It consists of areas like:

- Fairness
- Privacy
- Security
- Accountability
- Transparency

Fairness and bias



World

Data

Historical Bias
Representation Bias
Measurement Bias
Temporal Bias
Omitted Variable Bias

AI/ML

Algorithmic bias
Evaluation Bias
Aggregation Bias
Popularity Bias
Ranking Bias
Emergent Bias
Linking Bias

Human review

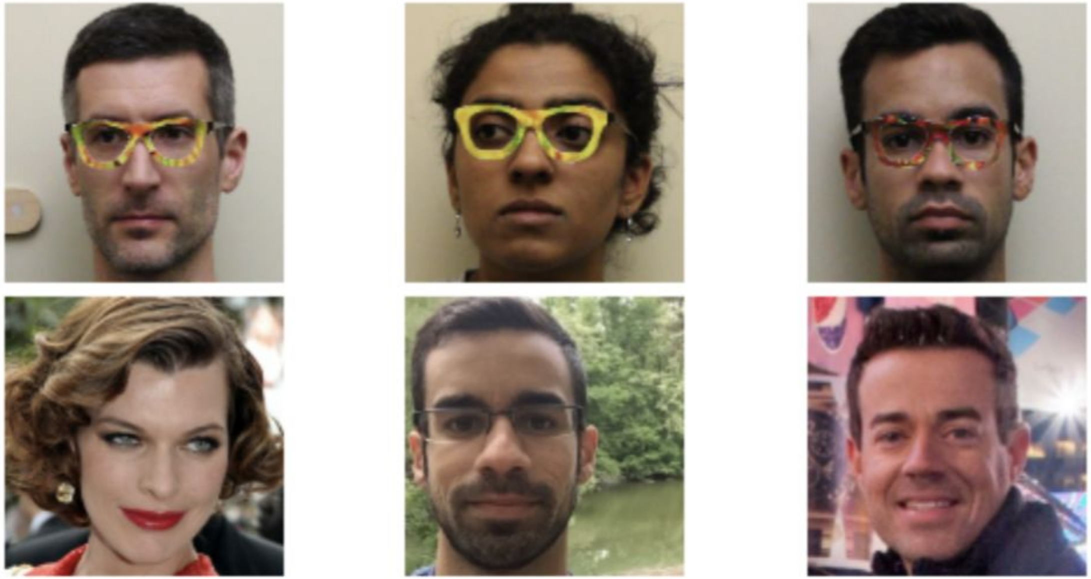
Behavioral Bias
Presentation Bias
Content Production Bias
Social Bias

Actions

AI models can be racist!

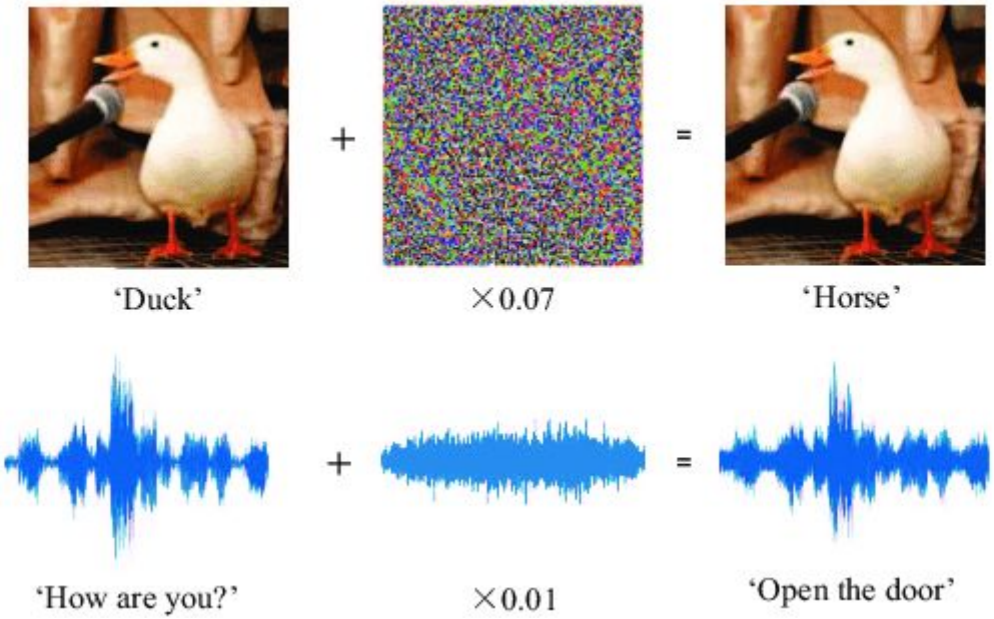
Security

Future hackers are AI experts!



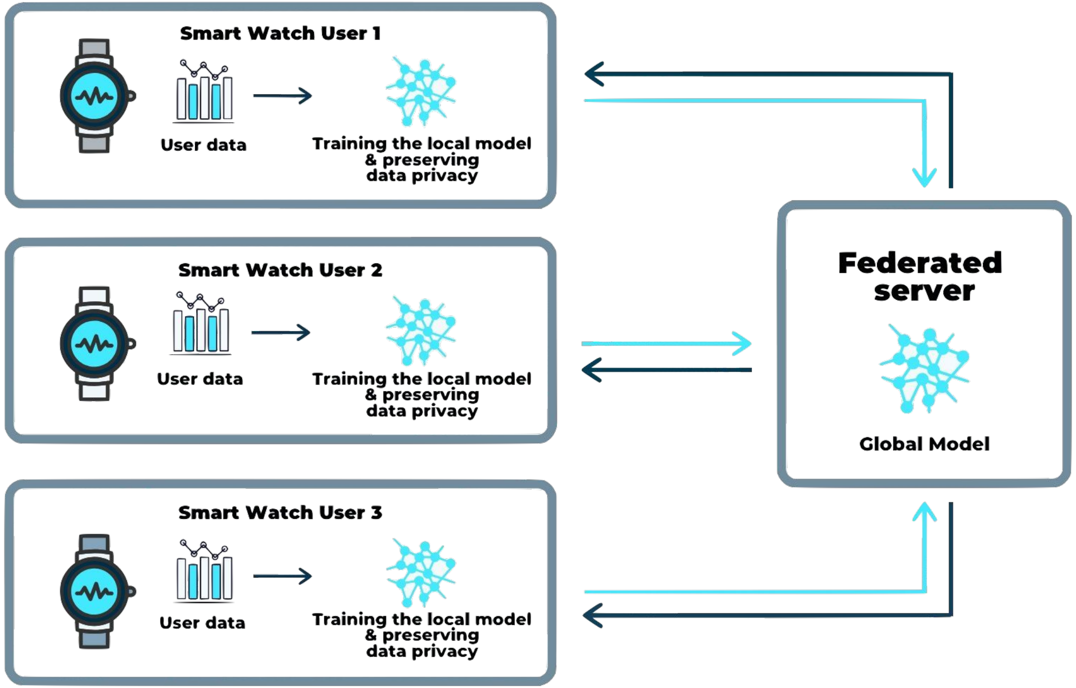
Security

Future hackers are AI experts!

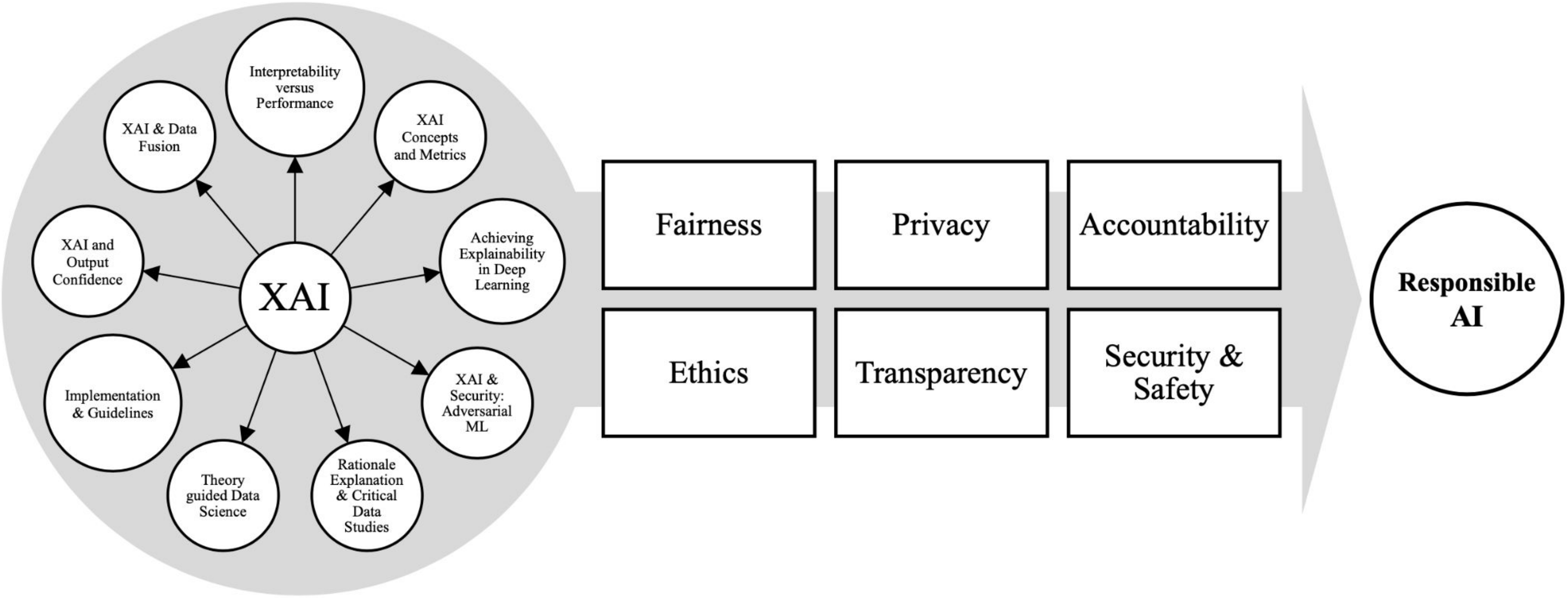


Privacy

How to preserve privacy and train our model!


Smart Watch User 1
User data
Training the local model & preserving data privacy
Smart Watch User 2
User data
Training the local model & preserving data privacy
Smart Watch User 3
User data
Training the local model & preserving data privacy
Federated server
Global Model

How to have responsible AI?


XAI
Interpretability versus Performance
XAI Concepts and Metrics
Achieving Explainability in Deep Learning
XAI & Security: Adversarial ML
Rationale Explanation & Critical Data Studies
Theory guided Data Science
Implementation & Guidelines
XAI and Output Confidence
XAI & Data Fusion
Fairness
Privacy
Accountability
Ethics
Transparency
Security & Safety
Responsible AI

Machine Learning Systems Design

Modeling Pipeline

Next Lecture: Model Serving Patterns and Infrastructures

CE 40959 Spring 2023
Ali Zarezade
SharifMLSD.github.io