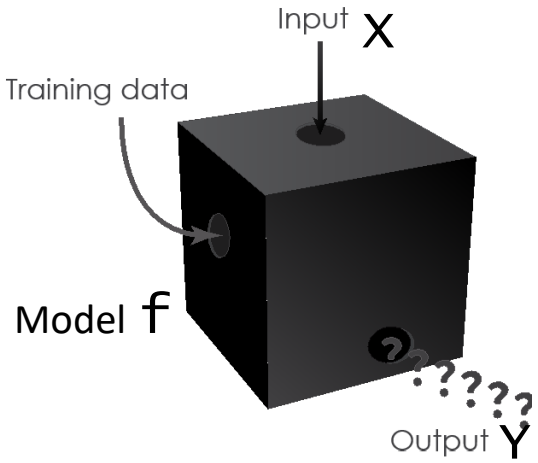
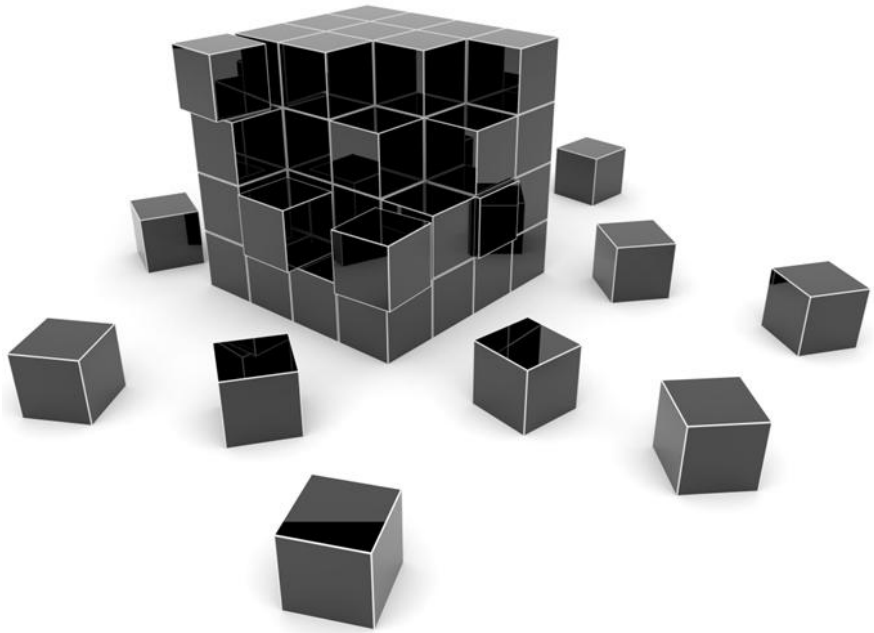
An Introduction to Explainable AI (XAI)

Vassilis Christophides ETIS, IPAL
Evi Pitoura UOI, ATHENARC
Katerina Tzompanaki CYU, ETIS

The Need to Explain AI Models

Standard AI Products





Users of AI Products

- **ML Centric** today
- Black-Box ML models: **opaque**, **non-intuitive** and **difficult to understand**

What (Data-based): What are the characteristics of the input data X?

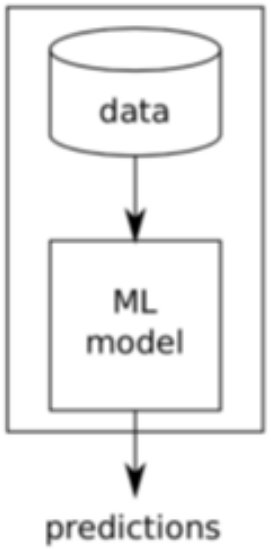
Why / Why Not (Decision-based): Why does the model make a specific (different) decision Y for a given input X (different input X')?

How To (Model-based): Exploring how the model f transforms input X into output Y?

What If (Causality-based): What alternative input X causes the model f to predict the same (different) output Y?

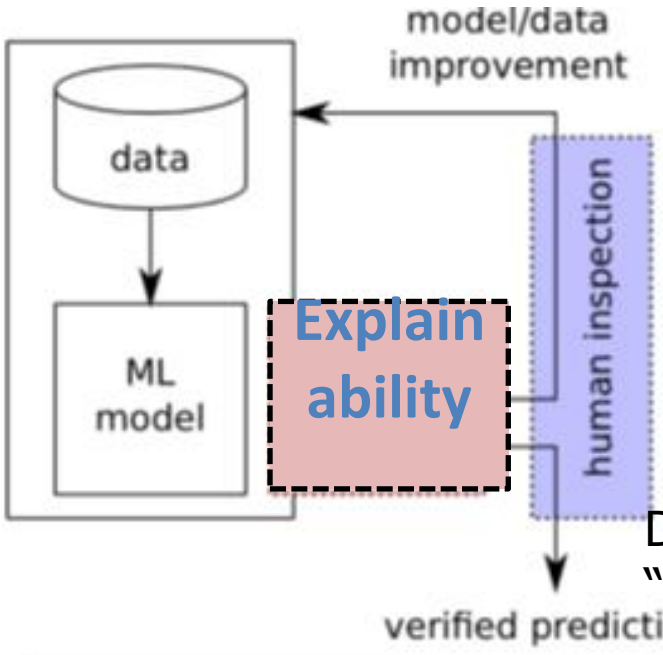
What is Explainable AI (XAI)?

Standard AI



Generalization error

Explainable AI (XAI)



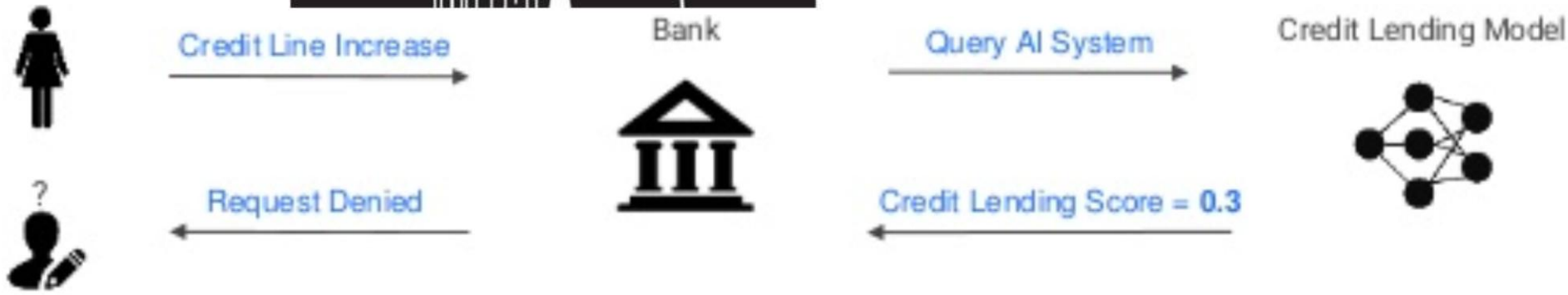
Decision Making in "High Risk" applications

Generalization error + human experience

- XAI explores and investigates methods for producing or complementing AI models, in order to make accessible and interpretable the internal logic and the outcome of the algorithms, making such process understandable by humans

Example: Credit Lending in a Black-Box ML World

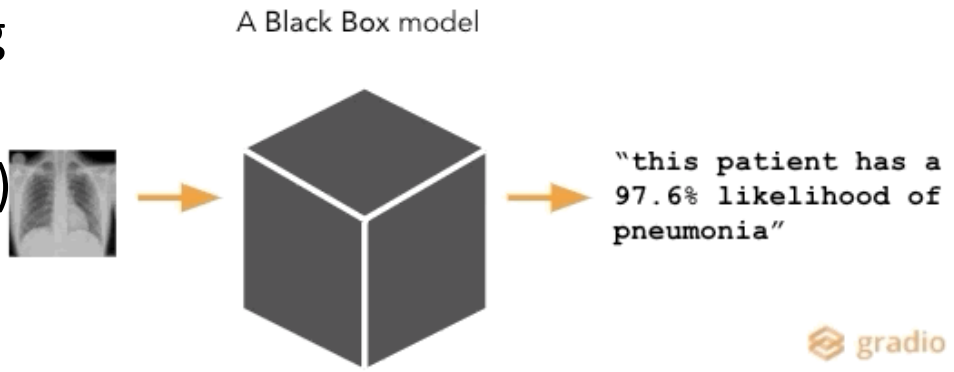
Fair Lending laws [ECOA, FCRA] require credit decisions to be explainable

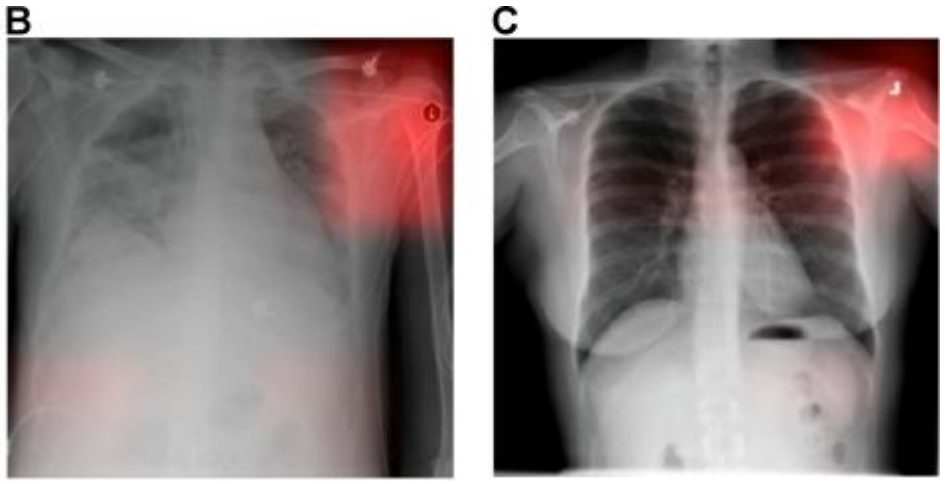


- **Why** my application was rejected?
- **How** can I improve my application to increase the likelihood my application is accepted?

Example: Predicting Pneumonia in a Black-Box ML World

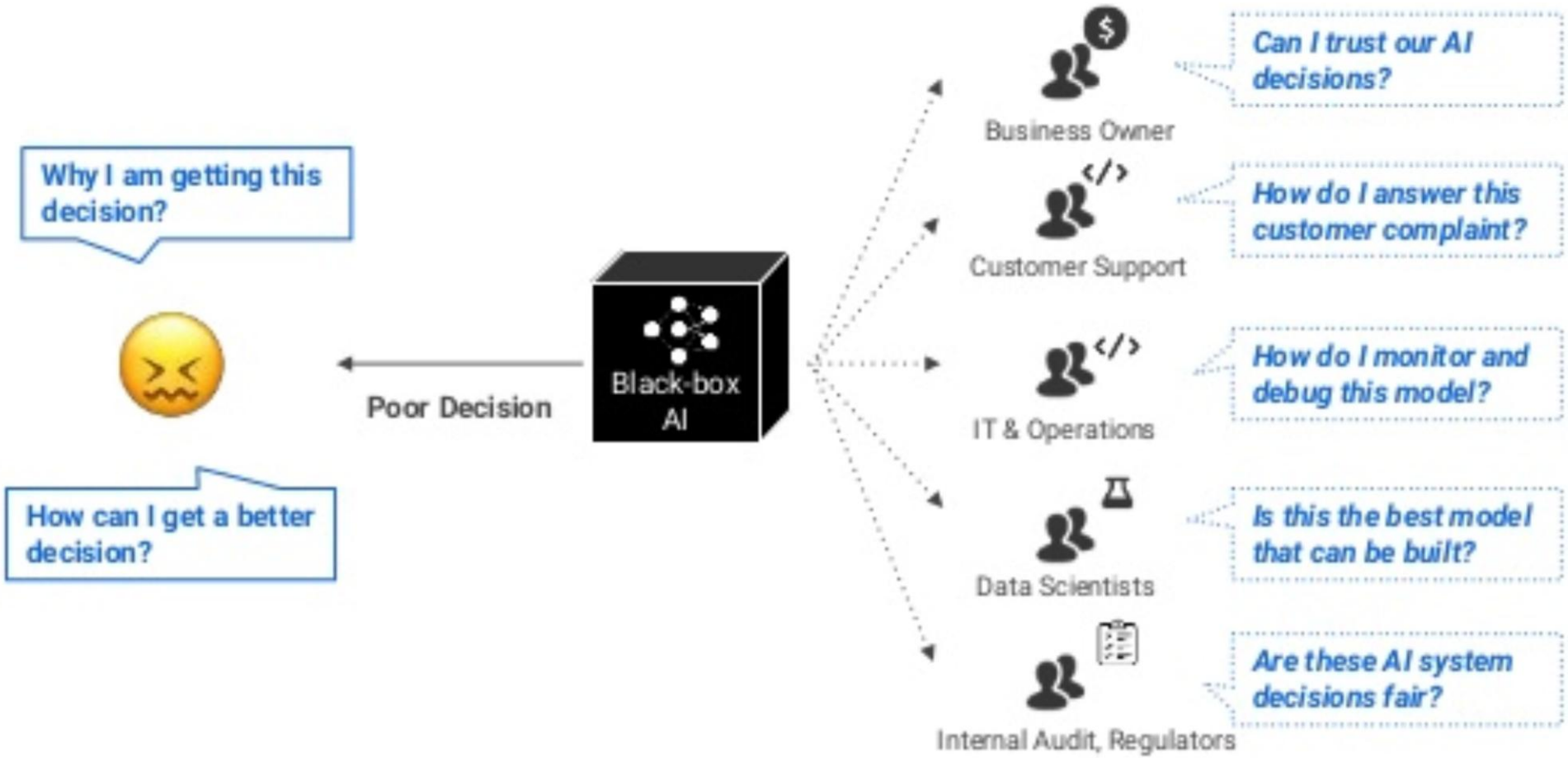
- Study on detecting pneumonia using 158323 chest radiographs
- Convolutional Neural Network (CNN) robustly identified hospital system and department within a hospital
  - which can have large differences in disease burden and may confound predictions
- CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image





Zech JR, Badgeley M2, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. https://www.ncbi.nlm.nih.gov/pubmed/30399157

Who Cares about XAI?



https://www.slideshare.net/KrishnaramKenthapadi/explainable-ai-in-industry-kdd-2019-tutorial

Why Explainable AI (XAI) ?

increase insightfulness

compliance is required

trust is necessary

performance is critical

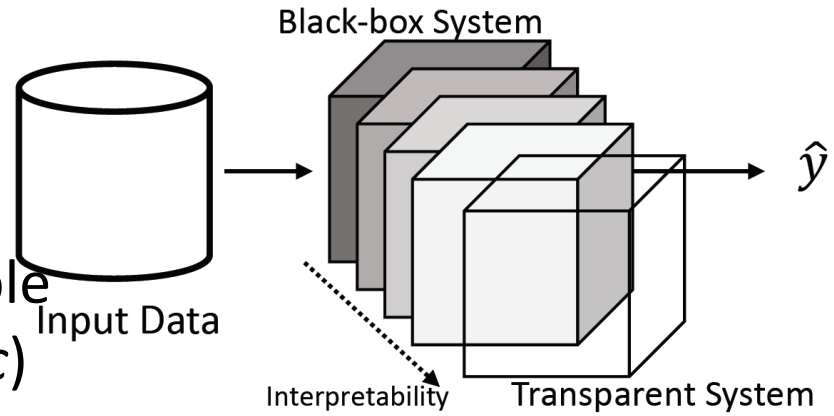
reduce ethical and moral bias

a new/unknown hypothesis is drawn

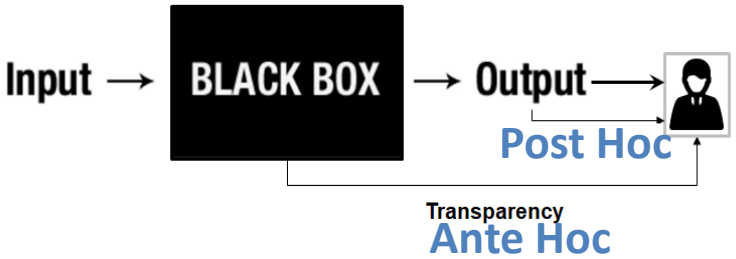
https://www.actuia.com/contribution/jean-cupe/linterpretabilite-de-lia-le-nouveau-defi-des-data-scientists

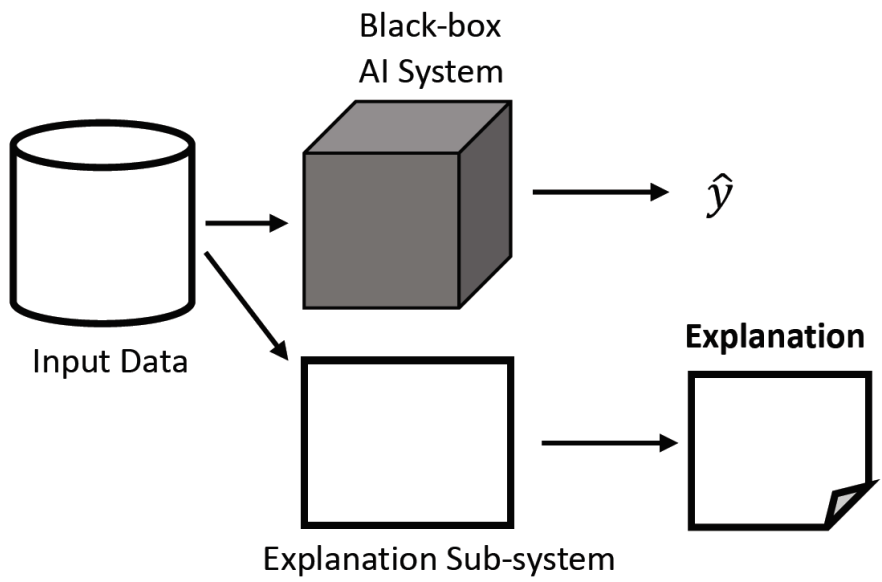
Explanation Approaches: Transparent Design vs Post-hoc

- Transparent (white-box) design reveals how a model functions
  - interpretable models straight away!
  - or retro-fit approximate interpretable models over complex ones (*intrinsic*)

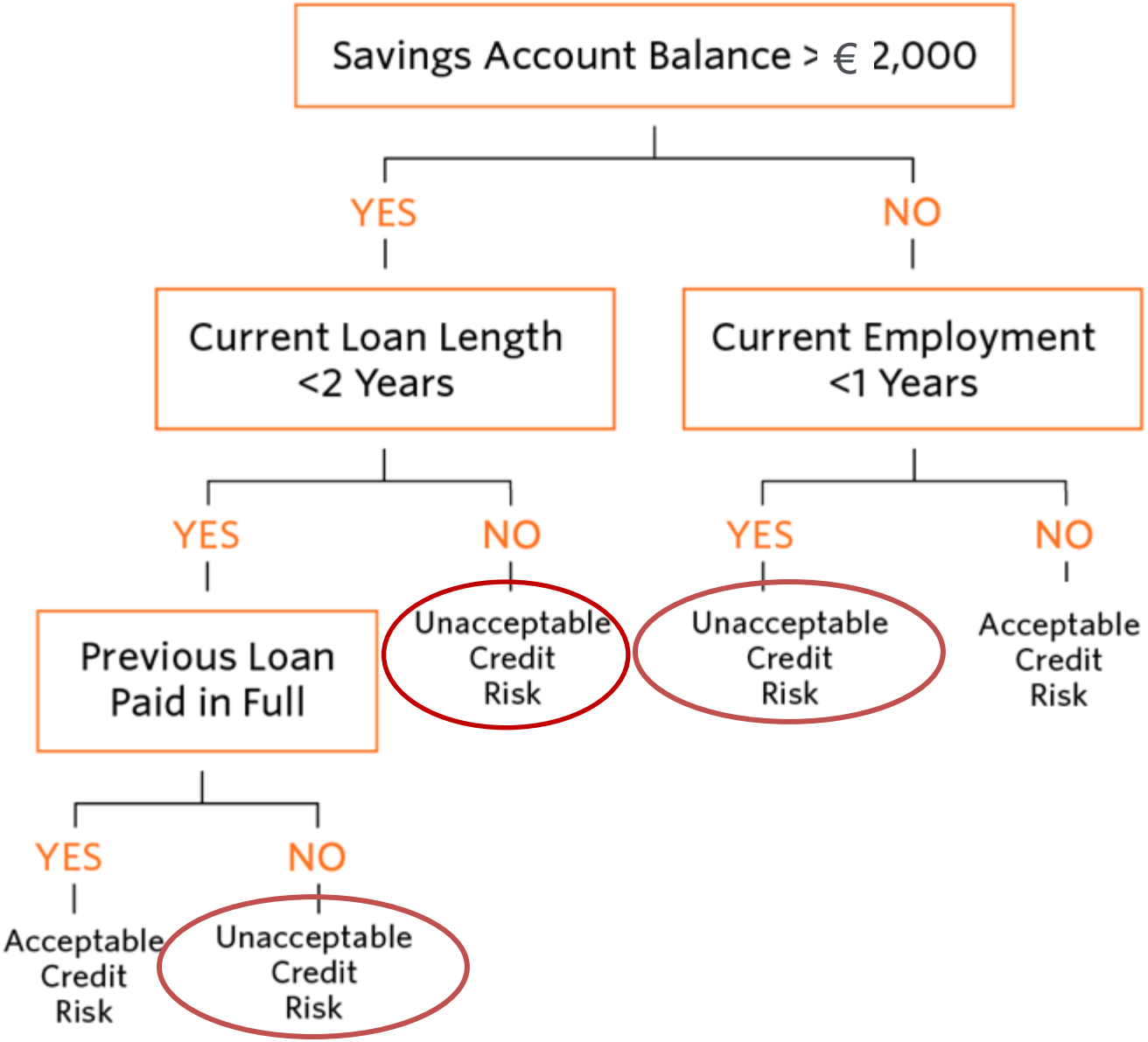


- Post-hoc Explanation explains why a black-box model behaved that way
  - Post-hoc explanations can be unreliable!





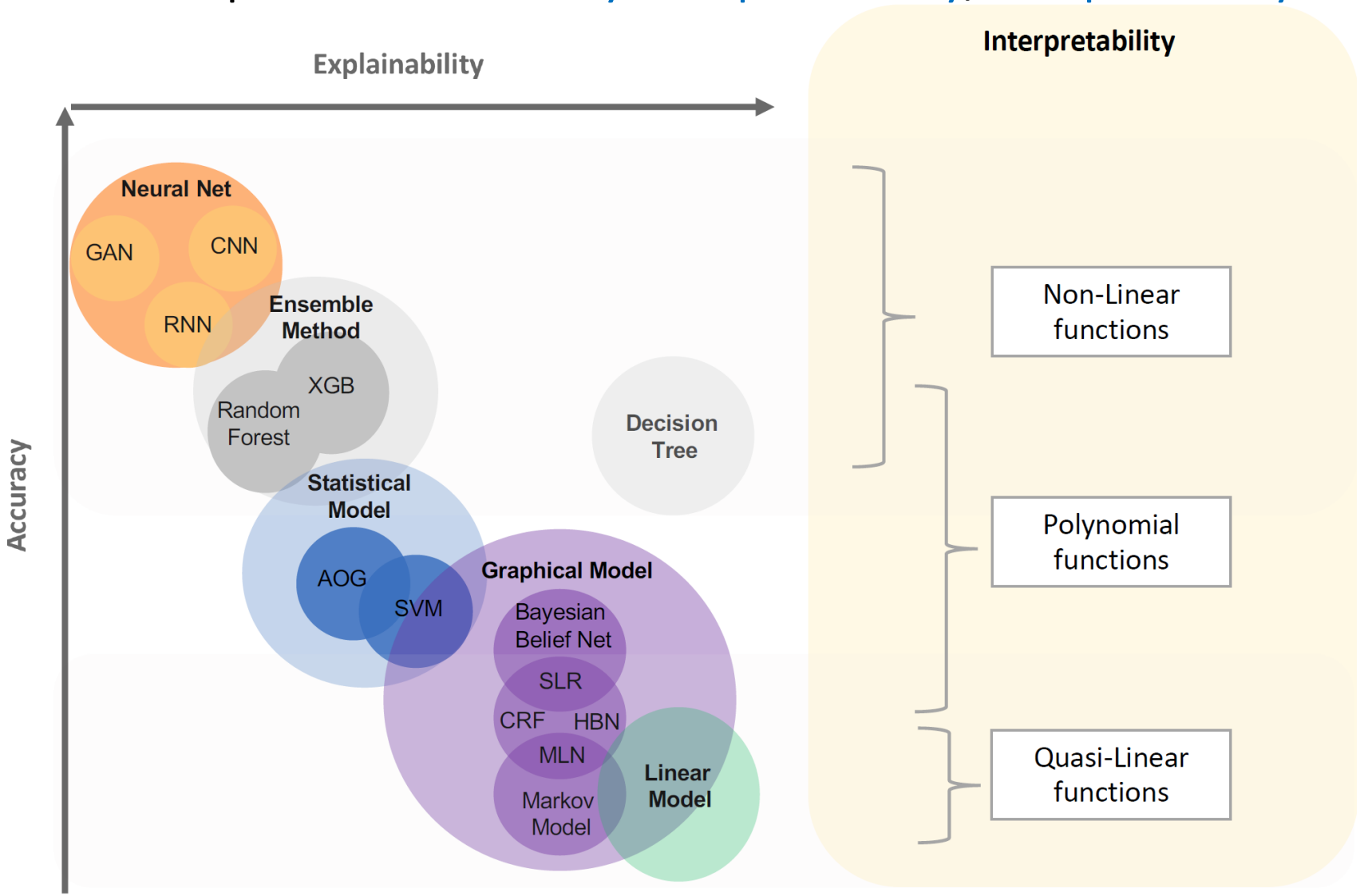
Credit Risk Prediction Example



https://www.pgs-soft.com/blog/ai-explainability-what-if-the-computer-says-no/

Interpretability Comes at Cost!

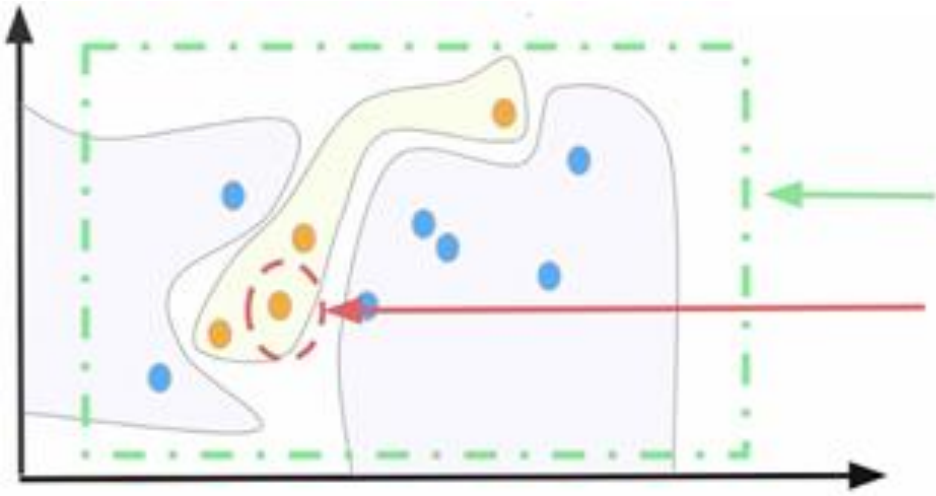
- Trade-off predictive accuracy vs explainability/interpretability



F. Lécué On the Role of Knowledge Graphs in Explainable AI A Machine Learning Perspective Semantic Web Journal 2020

Explanation Scope: Local vs Global

- Does the interpretation method explain the *complete model behavior*
- or identifies trends of an *individual prediction ?*
- or is the scope somewhere in between in a *target neighborhood* ?



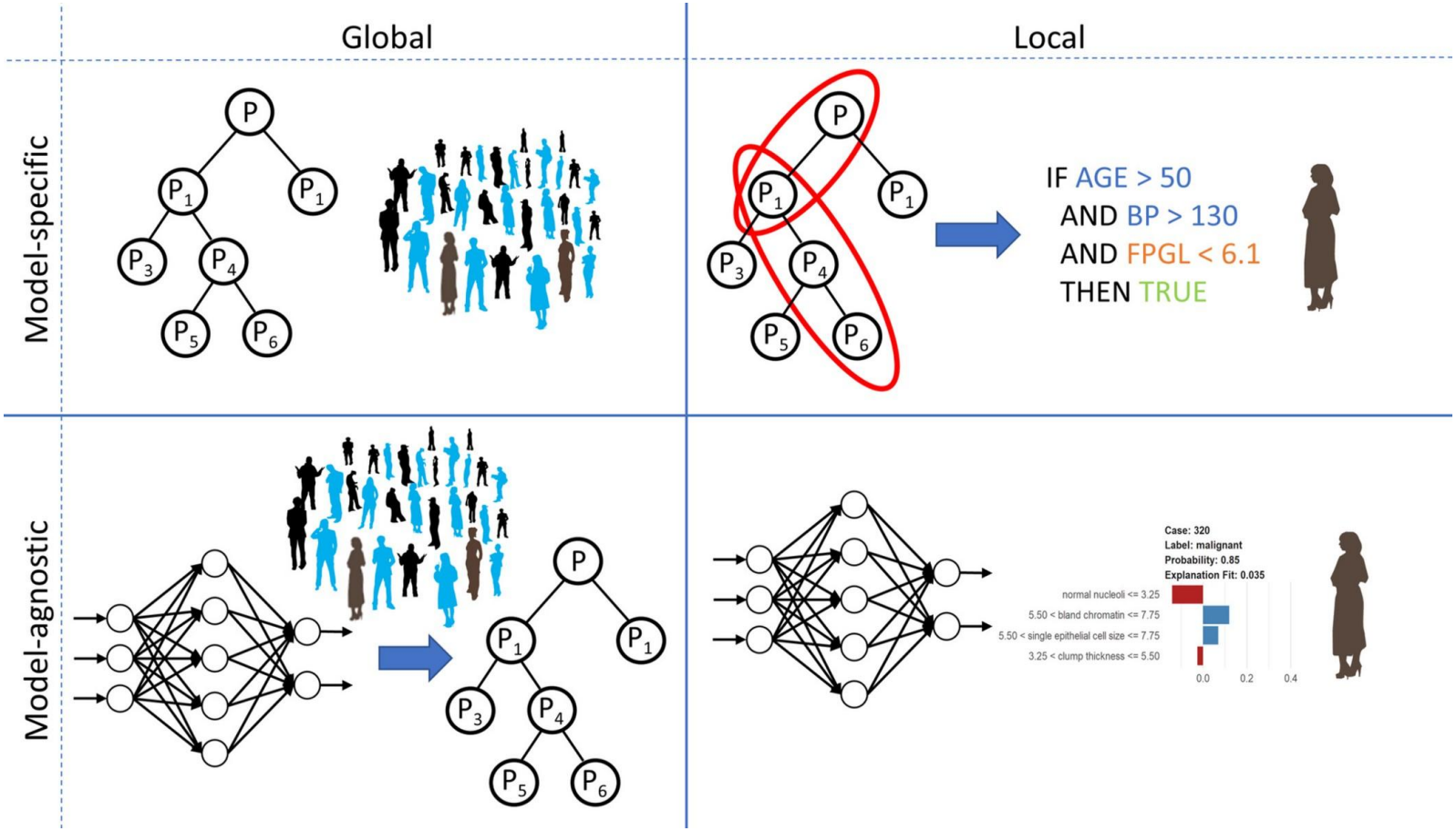
Explain the conditional interaction of dependent and independent variables based on the entire set of samples

Explain the conditional interaction of dependent and independent variables based on a subset of samples

https://christophm.github.io/interpretable-ml-book/taxonomy-of-interpretability-methods.html

Model-specific vs model-agnostic

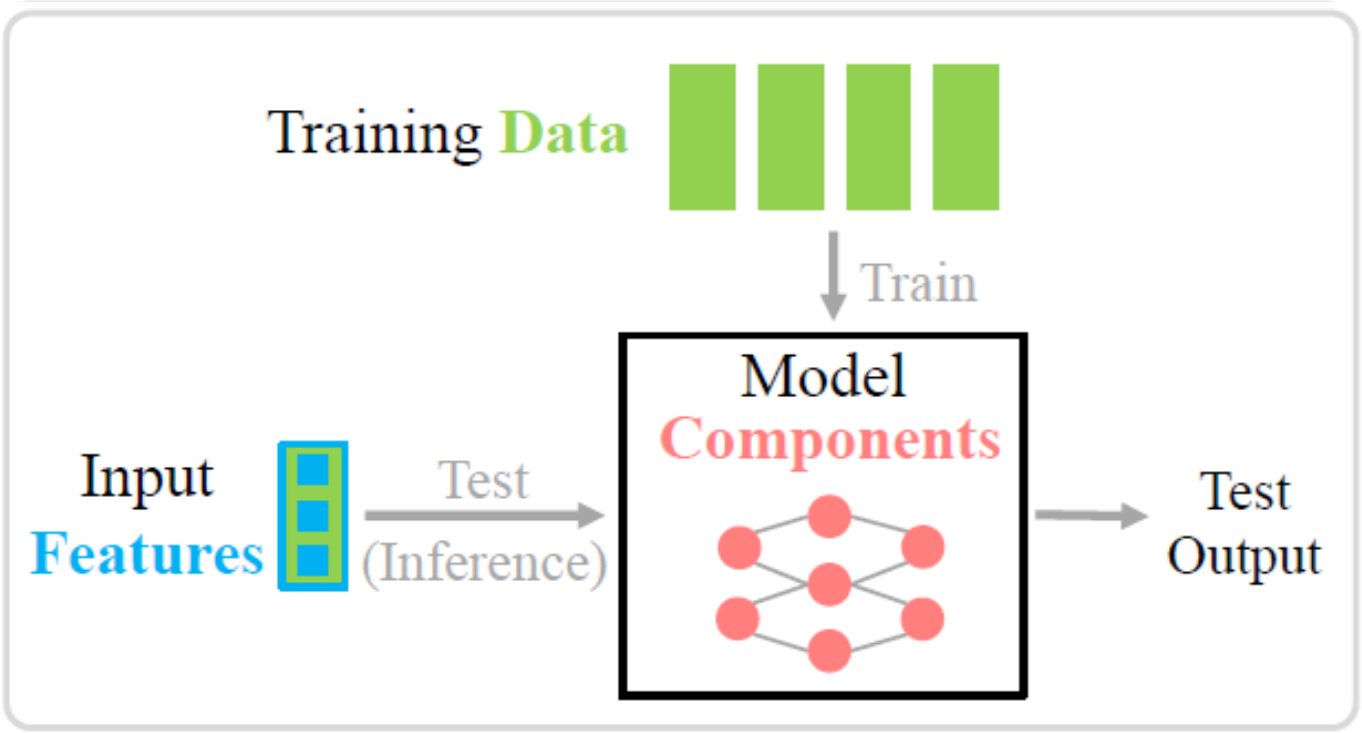
- Model-specific interpretation tools are *limited to specific model family*
- Model-agnostic tools can be used on *any ML model* and are applied after the model has been trained (e.g., *post hoc*)



https://towardsdatascience.com/explainable-machine-learning-9d1ca0547ae0

Explanation Focus: The Attribution Problem

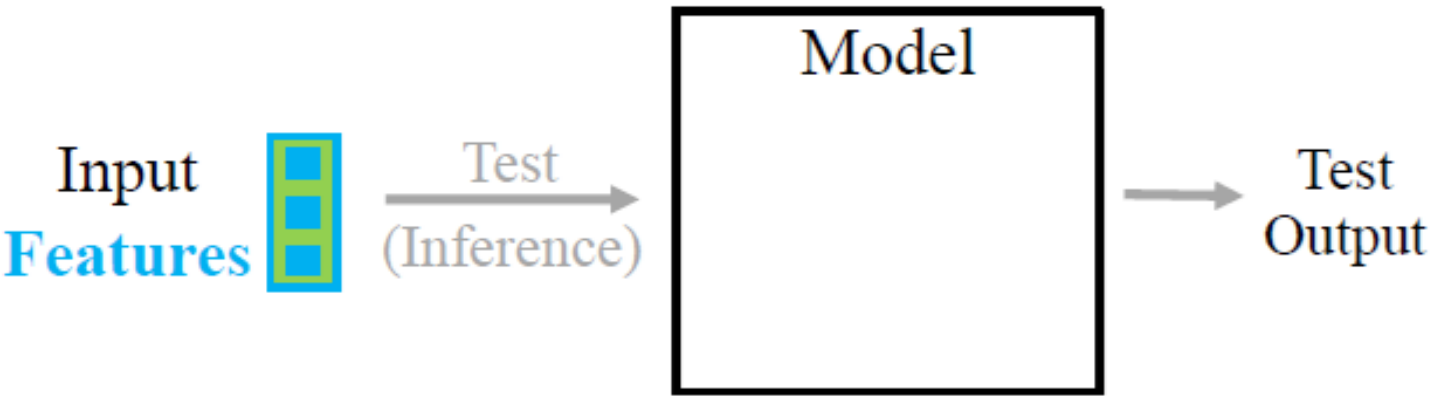
- Consider an abstraction of an AI system



- How to explain the model output?
  - Attribution

Feature Attribution

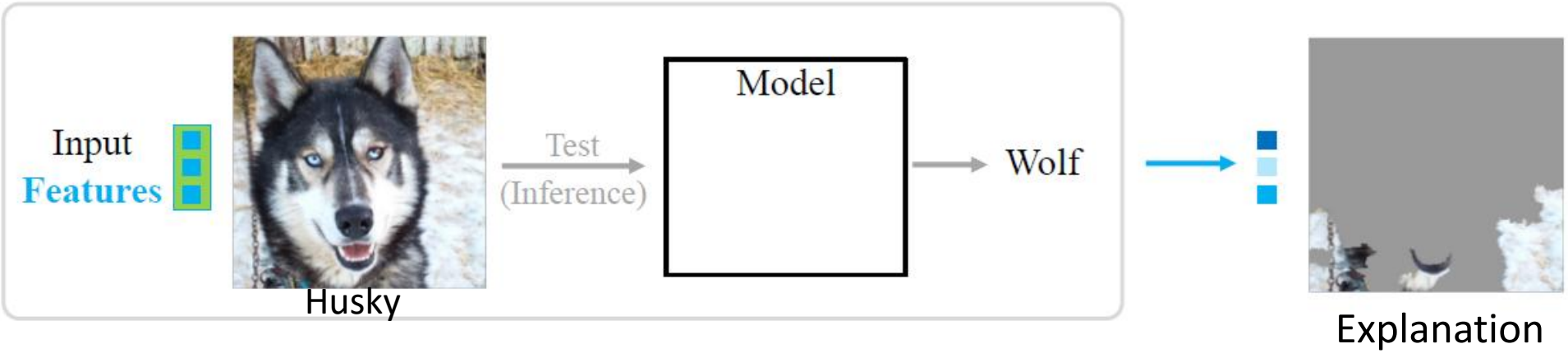
- Why this output for these input features?



- Feature attribution (FA) quantifies how individual features x_i of an input x influence the model's output f(x) through attribution scores $\phi_i(x)$
  - Applied to model inference at test time, it explains model behavior without altering model parameters

Feature Attribution

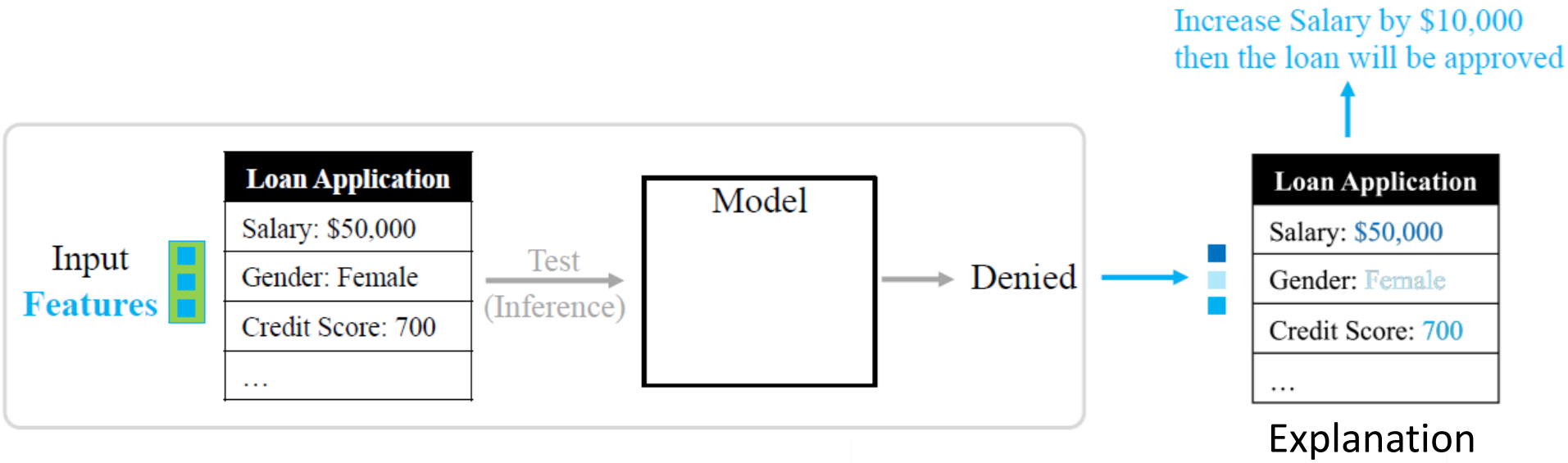
- FA identifies spurious correlations requiring correction



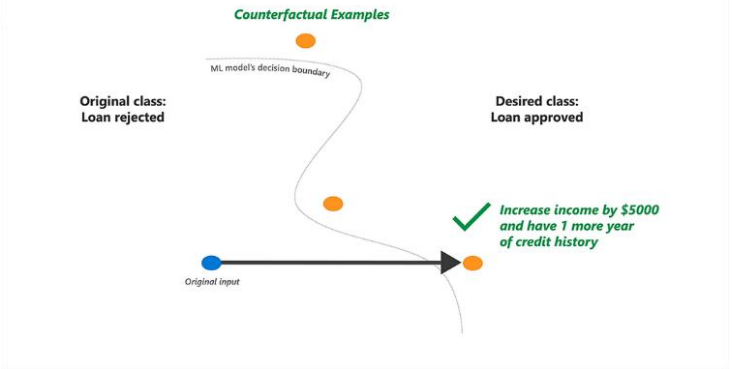
- [Ribeiro et al., 2016]

Feature Attribution

- FA justifies prediction, gain trust, and provide recourse by counterfactual explanations

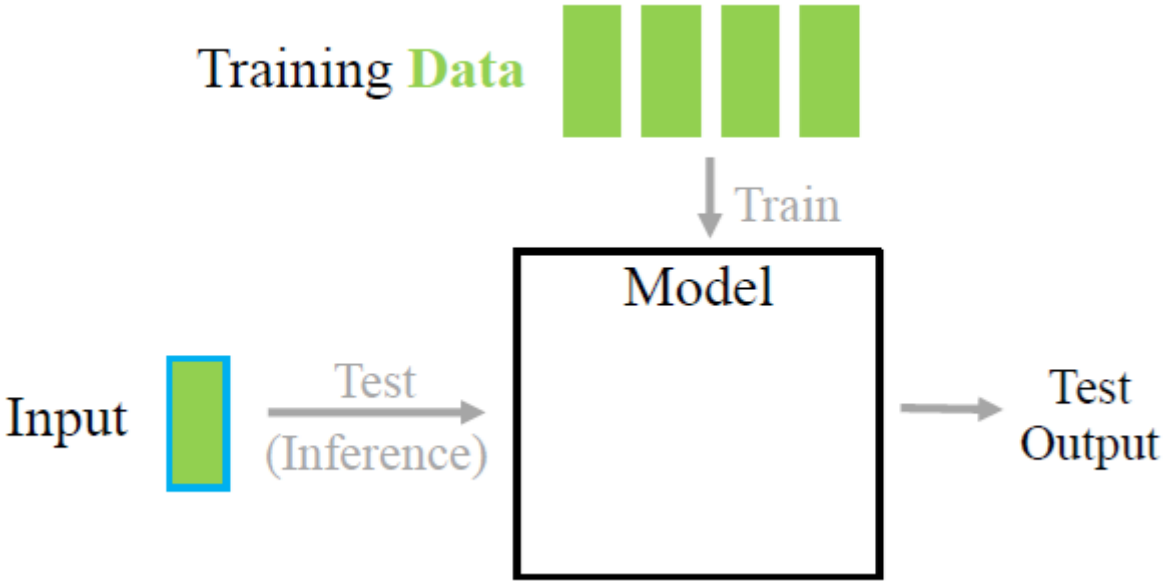


- [Wachter et al., 2018]



Data Attribution

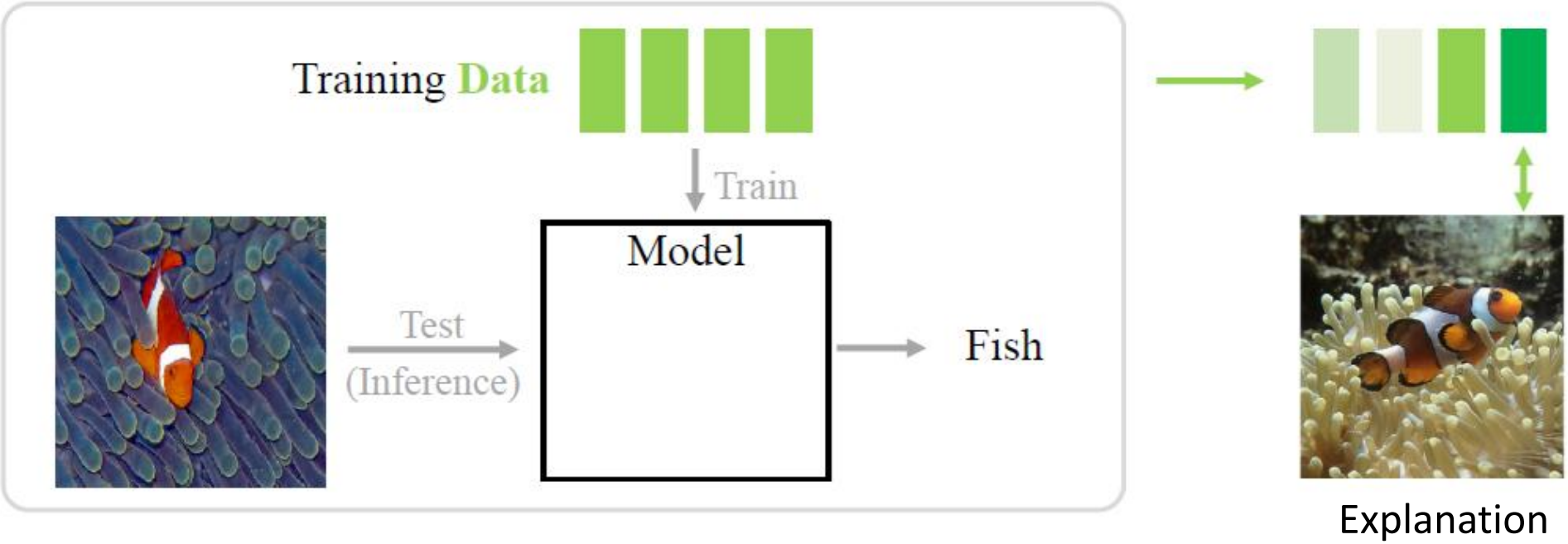
- Why this output for these training data samples?



- Data attribution (DA) assess how the training sample $x^{(j)}$ influence the model's output through attribution scores $\psi_j(x)$
  - It explains model behavior in terms of changes of its parameters caused by specific training data
  - Can be used for **Train-to-Test** (out-of-sample) and **Train-to-Train** (in-sample) data attribution

Data Attribution

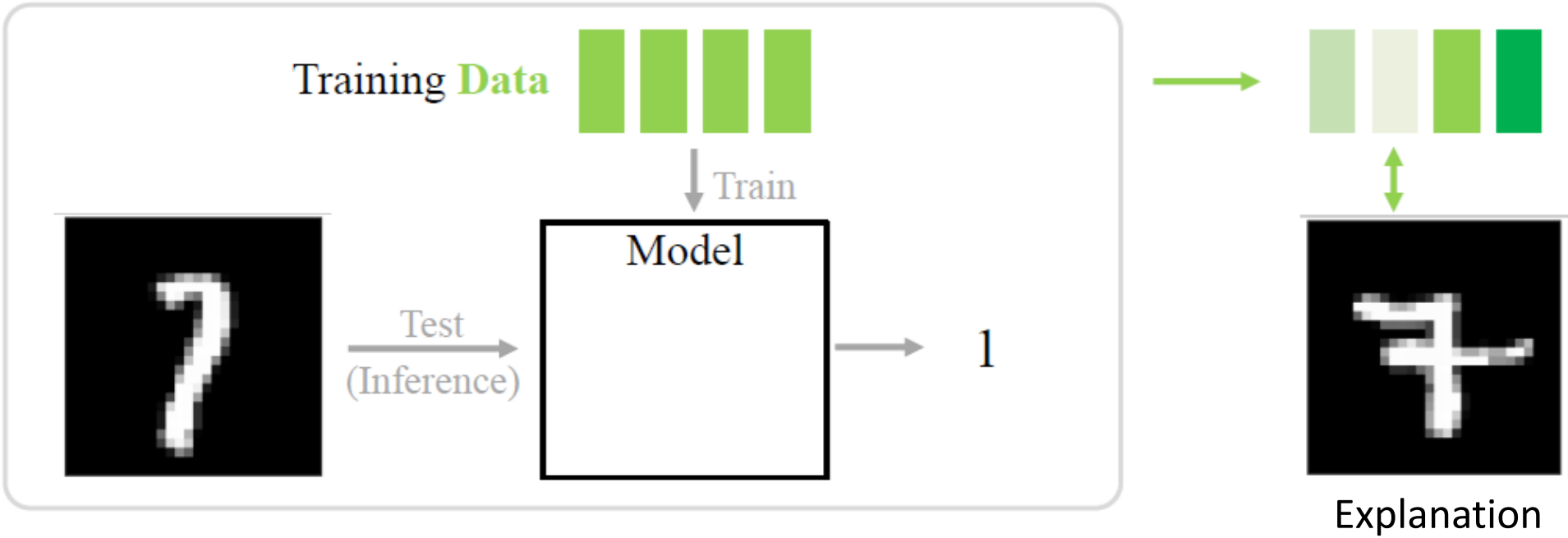
- Characterize training data properties and value
- Justify beneficial data (proponents of a specific prediction)



- [Koh & Liang, 2017]

Data Attribution

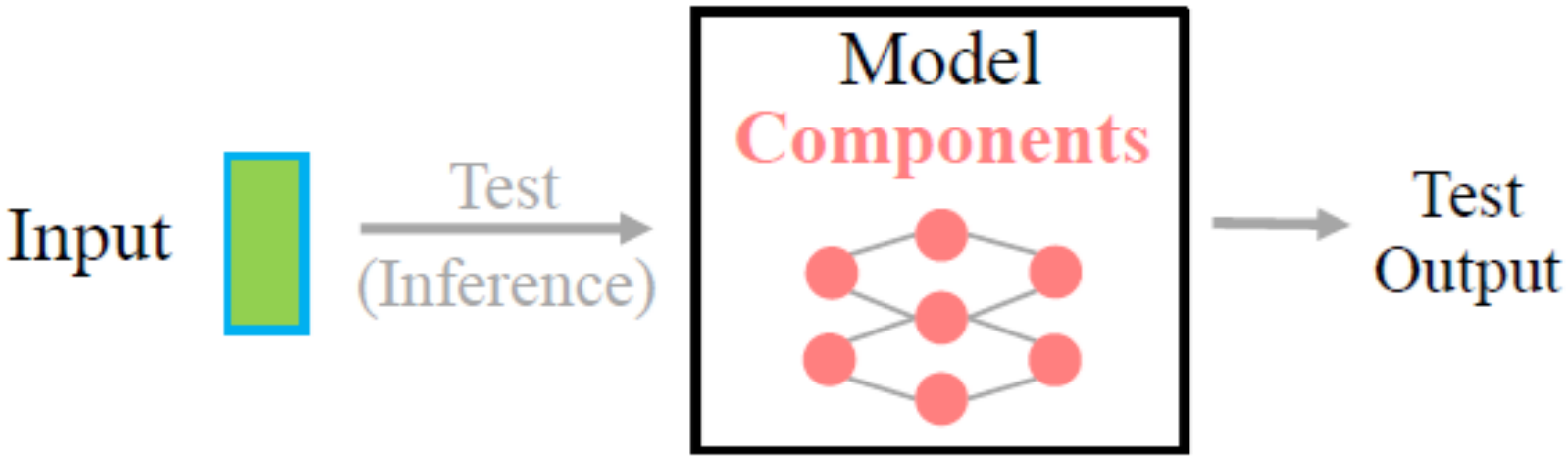
- Characterize training data properties and value
- Justify harmful data (opponents a specific prediction)



- [Koh & Liang, 2017]

Component Attribution

- Why this output for these model components?

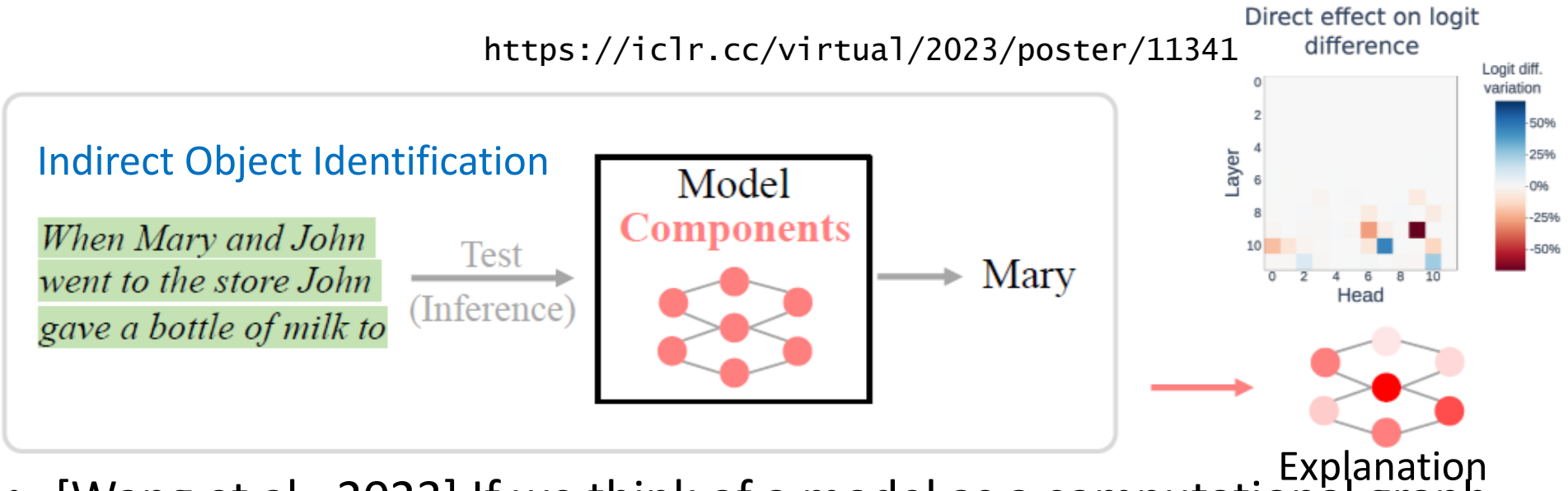


- Component attribution (CA) analyzes how an internal model component c_k contributes to the output f(x) through an attribution score $\gamma_k(x)$
  - Components can be defined flexibly- from individual neurons and attention heads to entire layers and circuits (subnetworks)

Component Attribution

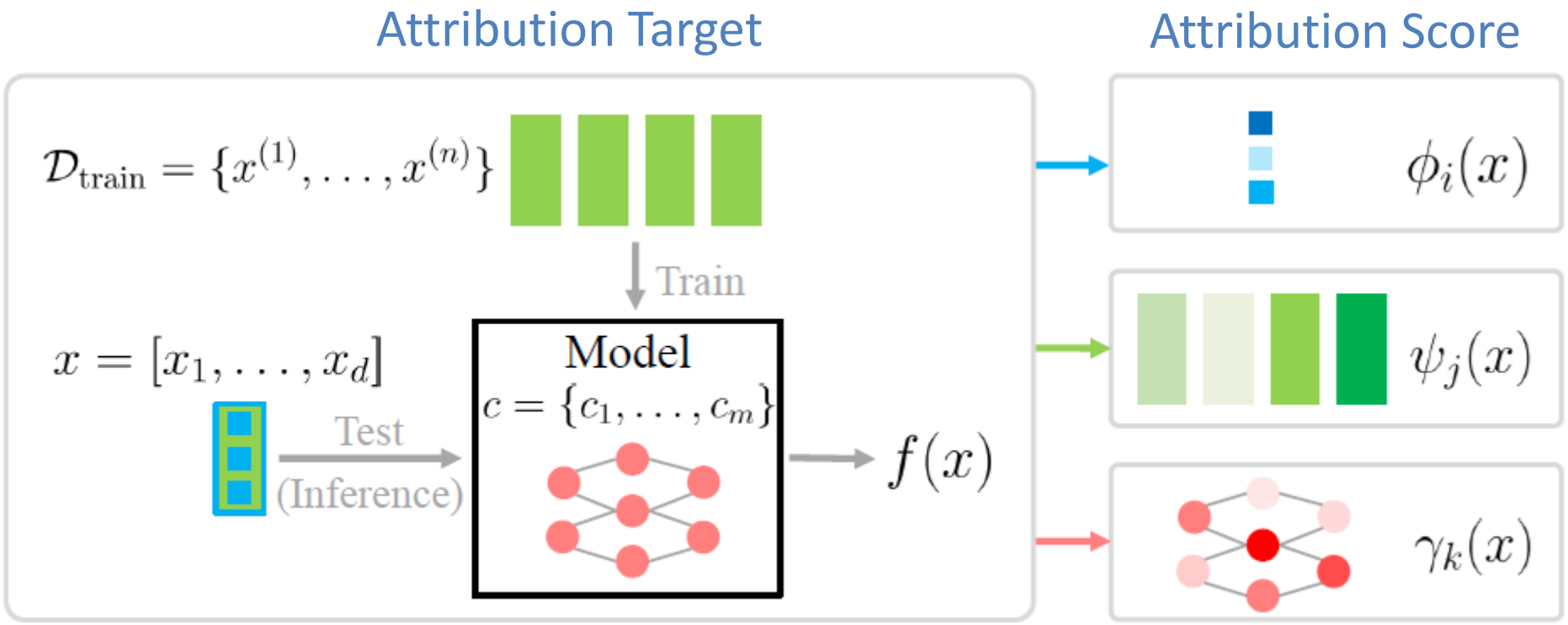
- Why this output for these model components?

https://iclr.cc/virtual/2023/poster/11341

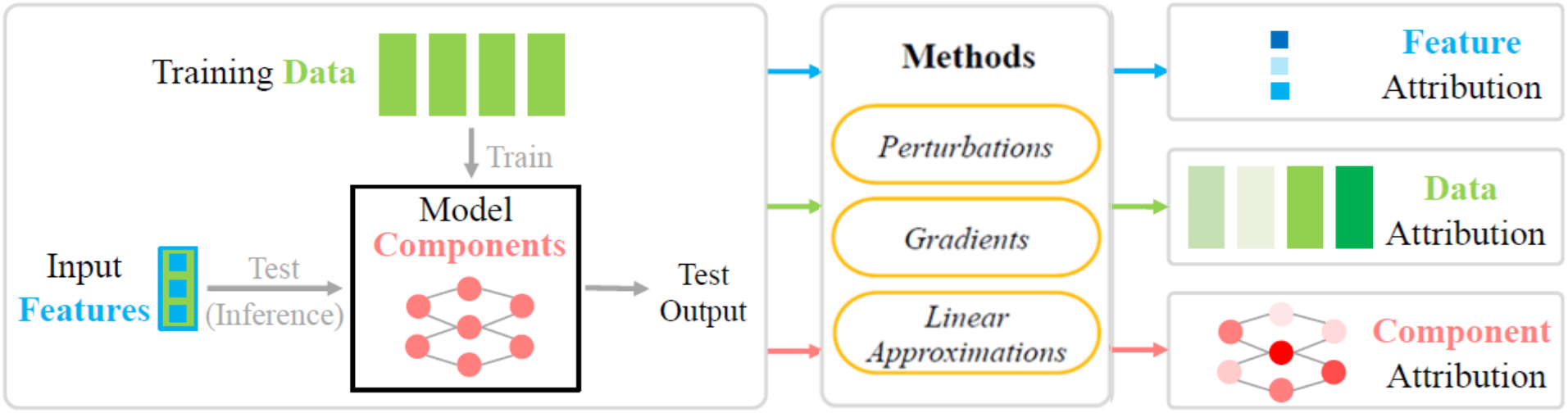


- [Wang et al., 2023] If we think of a model as a computational graph where nodes are terms in its forward pass (neurons, attention heads, etc.) and edges are the interactions between those terms (residual connections, attention, etc.), a circuit is a subgraph of the model responsible for some behavior

Formalization of the Attribution Problem



How Are These Attribution Results Achieved?

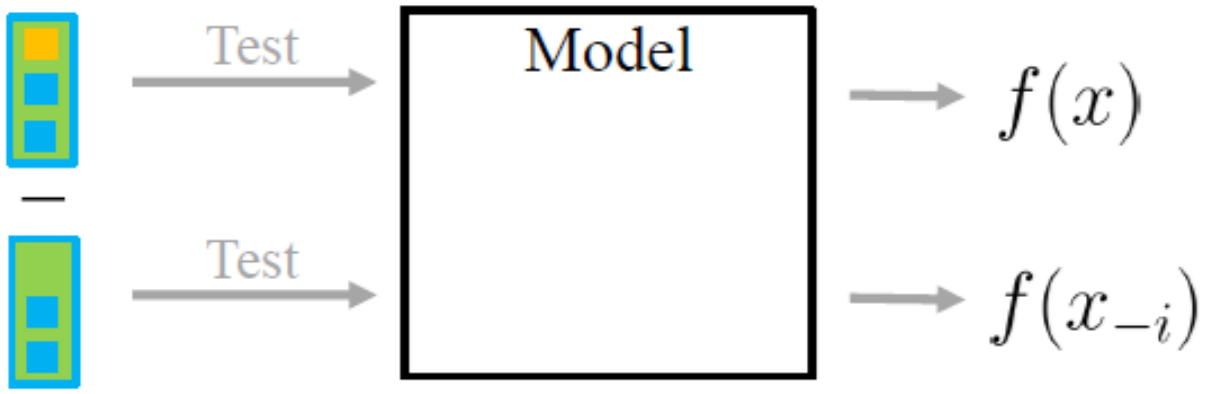


- Perturbation-based methods alter an input to observe how the model's output changes
- Gradient-based methods calculate the derivative of the output w.r.t. tiny changes of the input features
- Linear-based methods approximate the complex behavior of a model around an input of interest using a linear model

Perturbation-Based Feature Attribution

- Direct Perturbation
  - Perturb features and observe output changes

$$f(x) - f(x_{-i}) \rightarrow \phi_i(x)$$

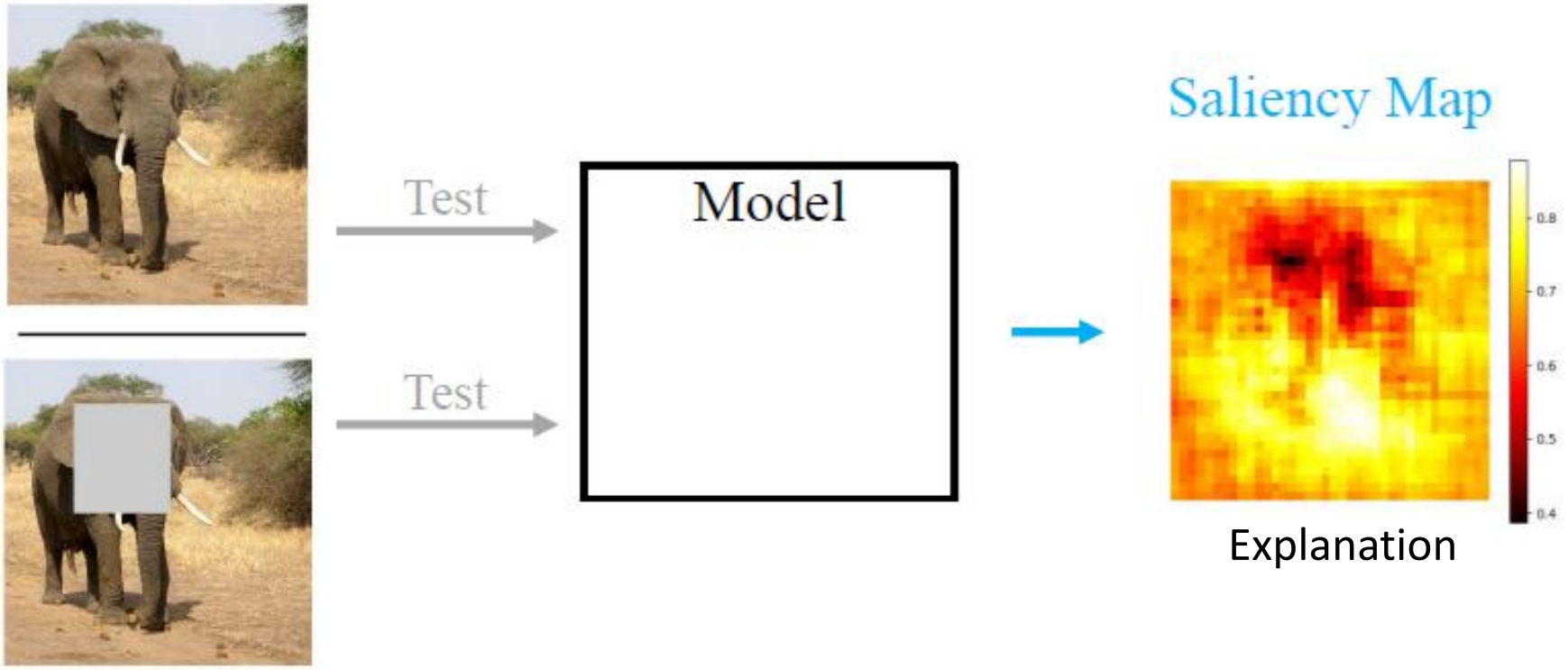


$f(x)$

$f(x_{-i})$

Perturbation-Based Feature Attribution

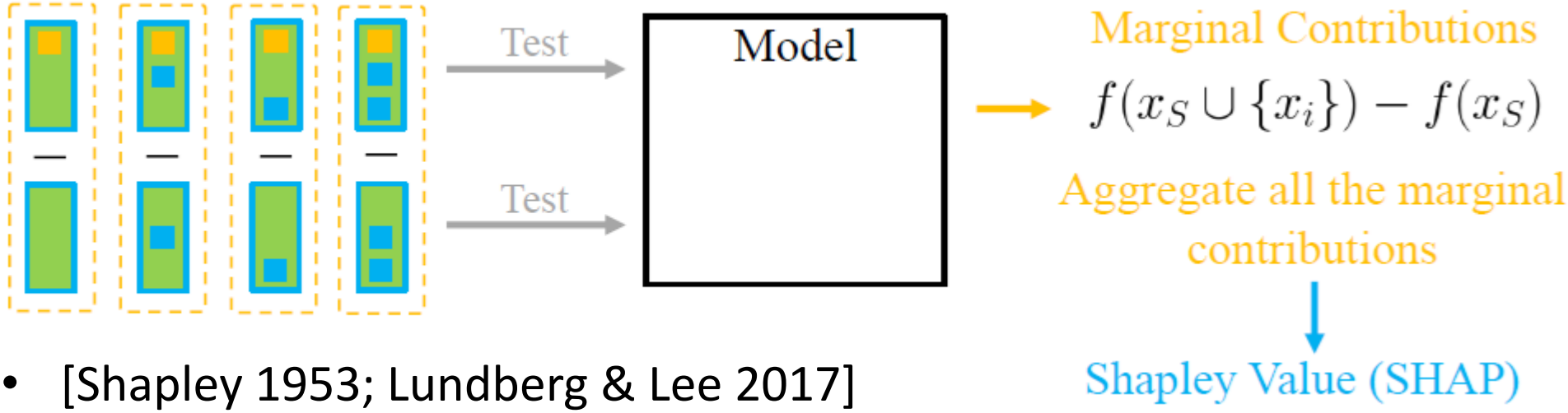
- Direct Perturbation
  - Occlusion



- [Zeiler and Fergus, 2014, Petsiuk, 2018]

Perturbation-Based Feature Attribution

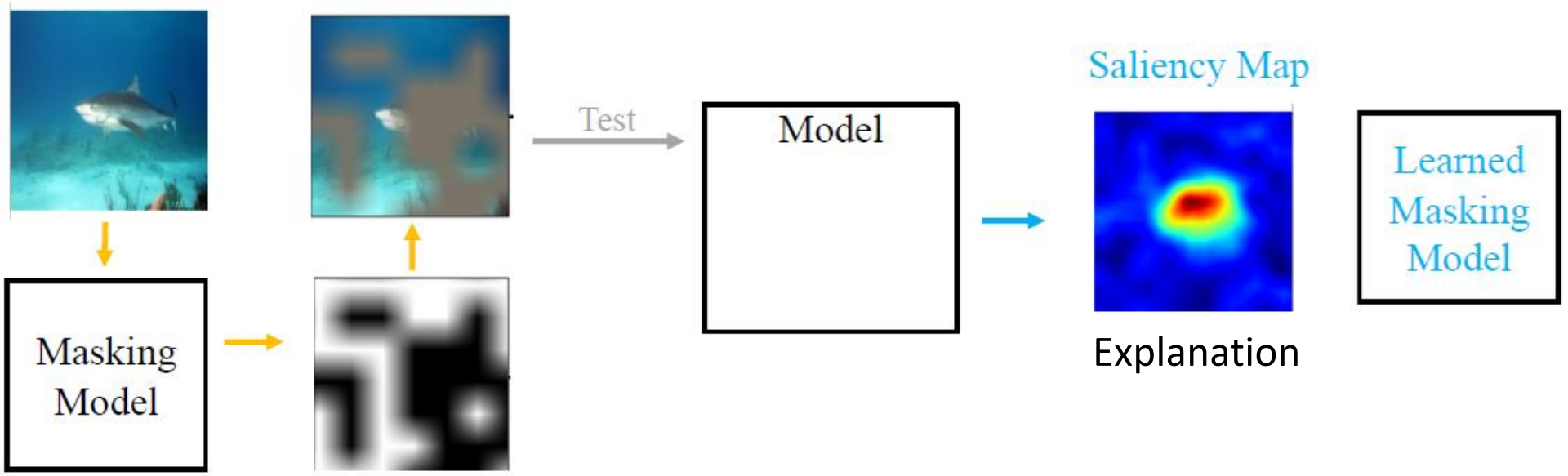
- Feature interactions?
- Game Theoretic Perturbation
  - Features as game players collaborating toward the model's output: 2^d marginal contributions



- [Shapley 1953; Lundberg & Lee 2017]
- Perturbation type: how adding a feature x_i to different feature subsets changes the model's output compared to the subset alone, known as the marginal contribution of x_i to the subset

Perturbation-Based Feature Attribution

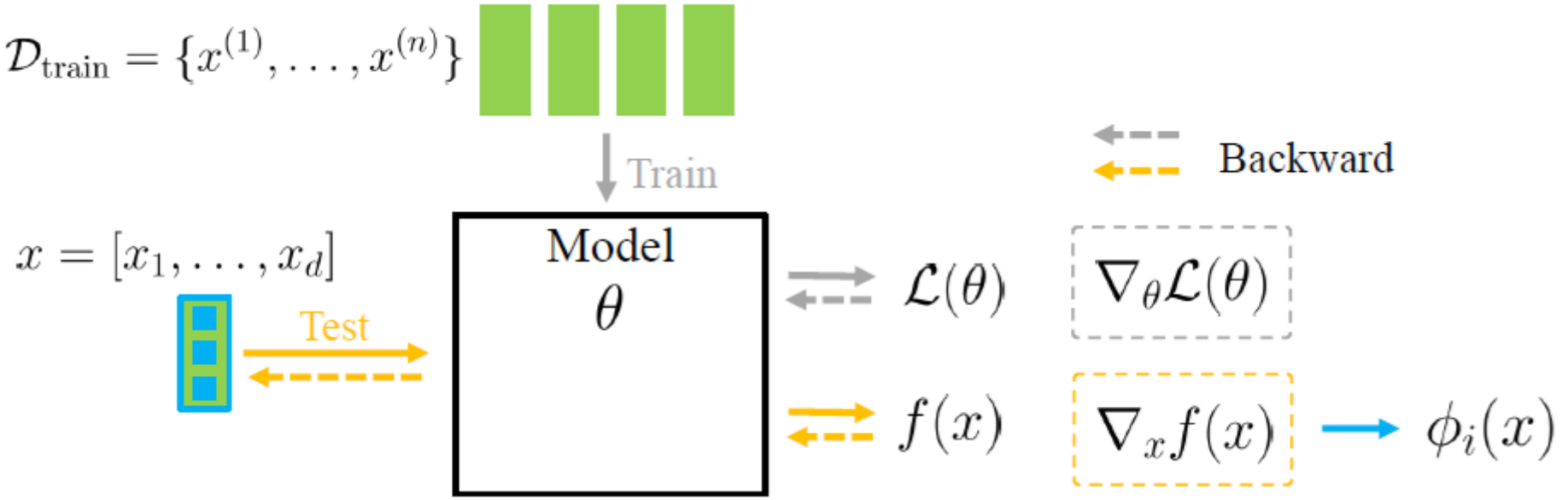
- Perturbation Mask Learning
  - Perturbations can be seen as binary masks
  - Why not make them continuous and learnable?
  - The masking model can be applied to other inputs



- [Fong & Vedaldi, 2017, Dabkowski & Gal 2017]

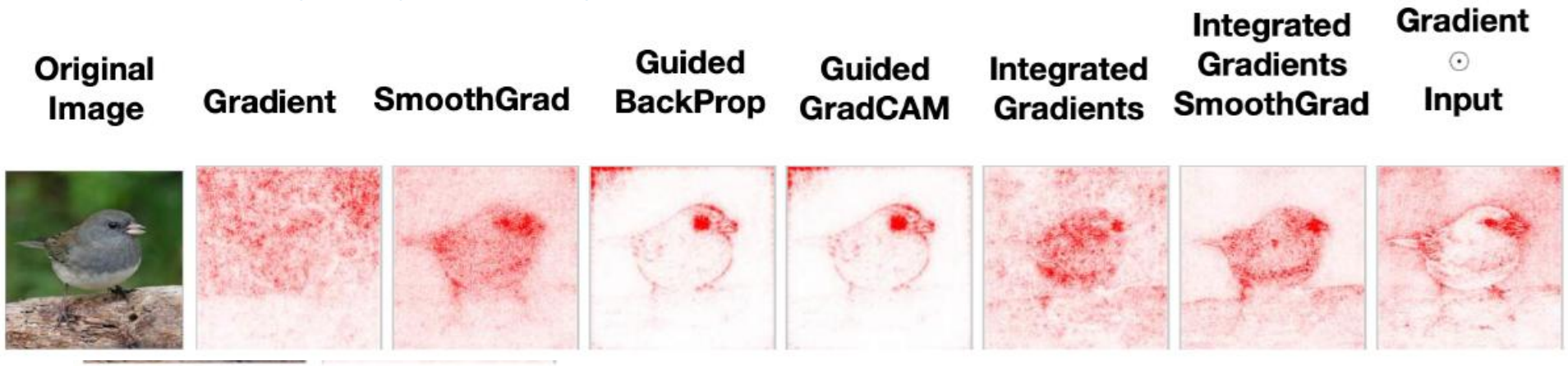
Gradient-Based Feature Attribution

- Parameter gradients for model training vs. feature gradients for attribution
- Gradients of model outputs f(x) with respect to input features x, $\nabla_x f(x)$, quantify output sensitivity to small input changes
  - Measures feature influence without requiring perturbations



Gradient-Based Feature Attribution

- "Vanilla gradients" [Simonyan et al., 2013] uses the gradients of the output class (log)probability w.r.t. input pixels as attribution scores
- Numerous enhanced gradient-based methods has been proposed
  - SmoothGrad [Smilkov et al.,2017]Integrated Gradients [Sundararajan et al.,2017]
- It is not easy to assess which explanation method is better by only looking at the saliency maps
  - Saliency maps of very different classes can be still similar



"Vanilla" Gradients

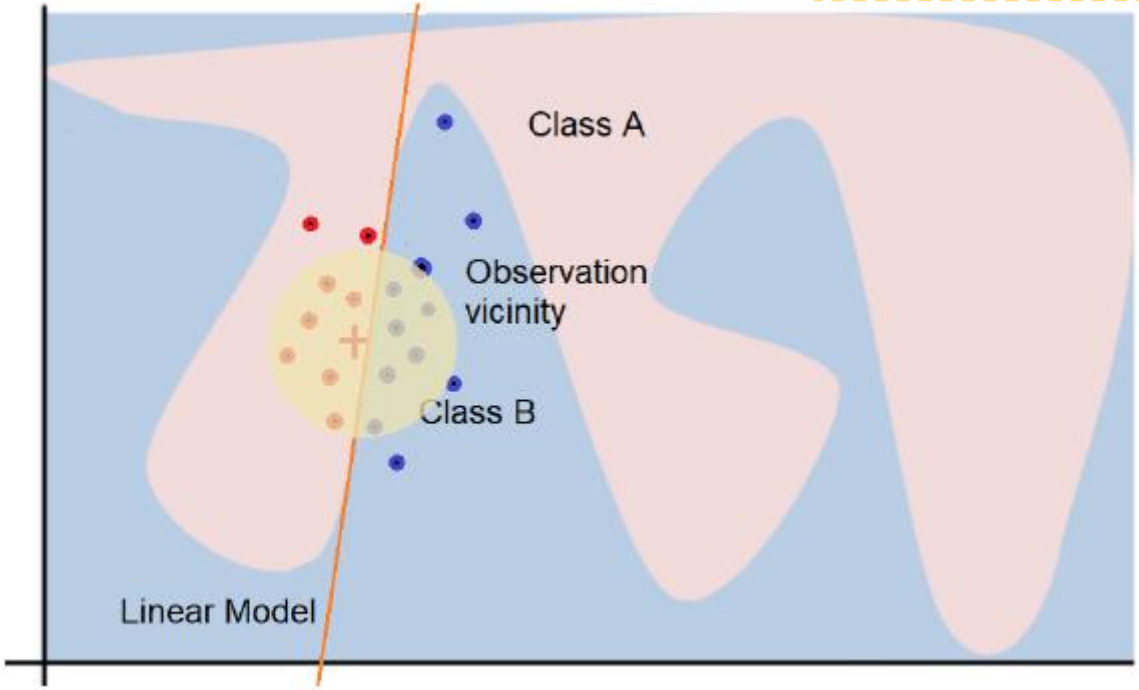
Variations of Gradients

- [Adebayo et al., 2018]

Linear Approximations for Feature Attribution

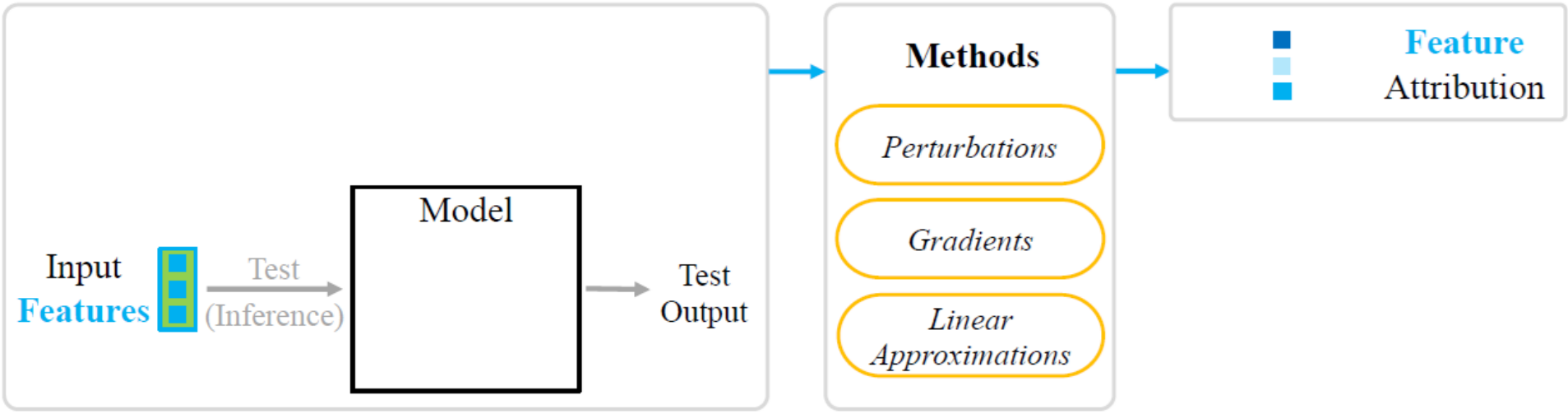
- If we cannot understand the complex AI models, what model can we understand?
- LIME: local approximation
  - We don't even need actual input features, only binary indicators

$$g(x) = w^T x + b \quad s.t. \quad f(x) \approx g(x) \qquad f(x) \approx g(z), z \in \{1,0\}^d \qquad w_i \longrightarrow \phi_i(x)$$



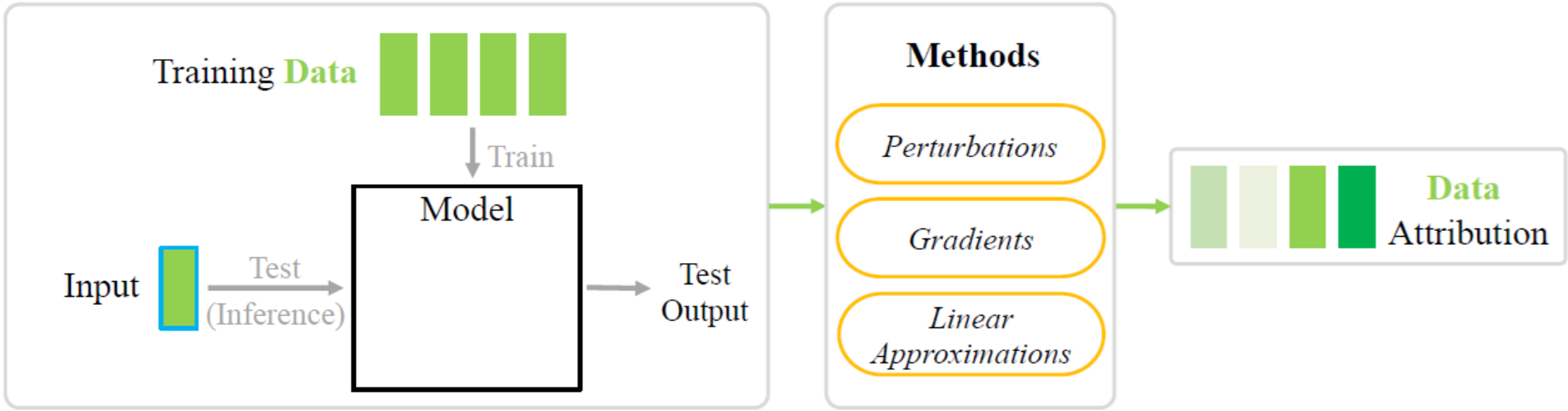
- [Ribeiro et al., 2016; Santiago, 2020]

Feature Attribution Methods



- FA methods can be unified under the local function approximation (LFA) framework [Han et al., 2022]
  - a model f is approximated around a point of interest x in a local neighborhood Z by an interpretable model g using a loss function ℓ
  - The previous FA methods are instances of the LFA, differing only in their choices of local neighborhoods Z and loss functions ℓ

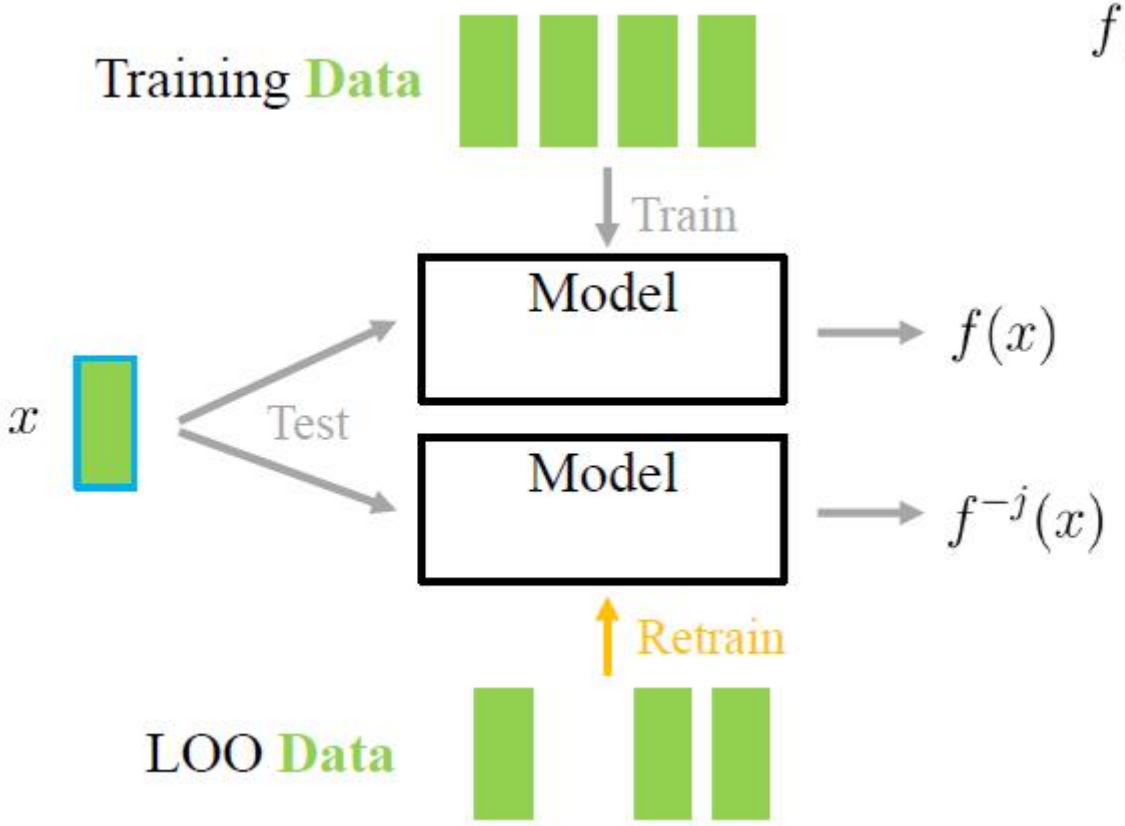
Data Attribution Methods



Perturbation-Based Data Attribution

- Leave-One-Out (LOO) (Direct Perturbation)
  - Remove one training sample at a time, retrain model, and observe output changes
  - Computationally expensive, retrain n times

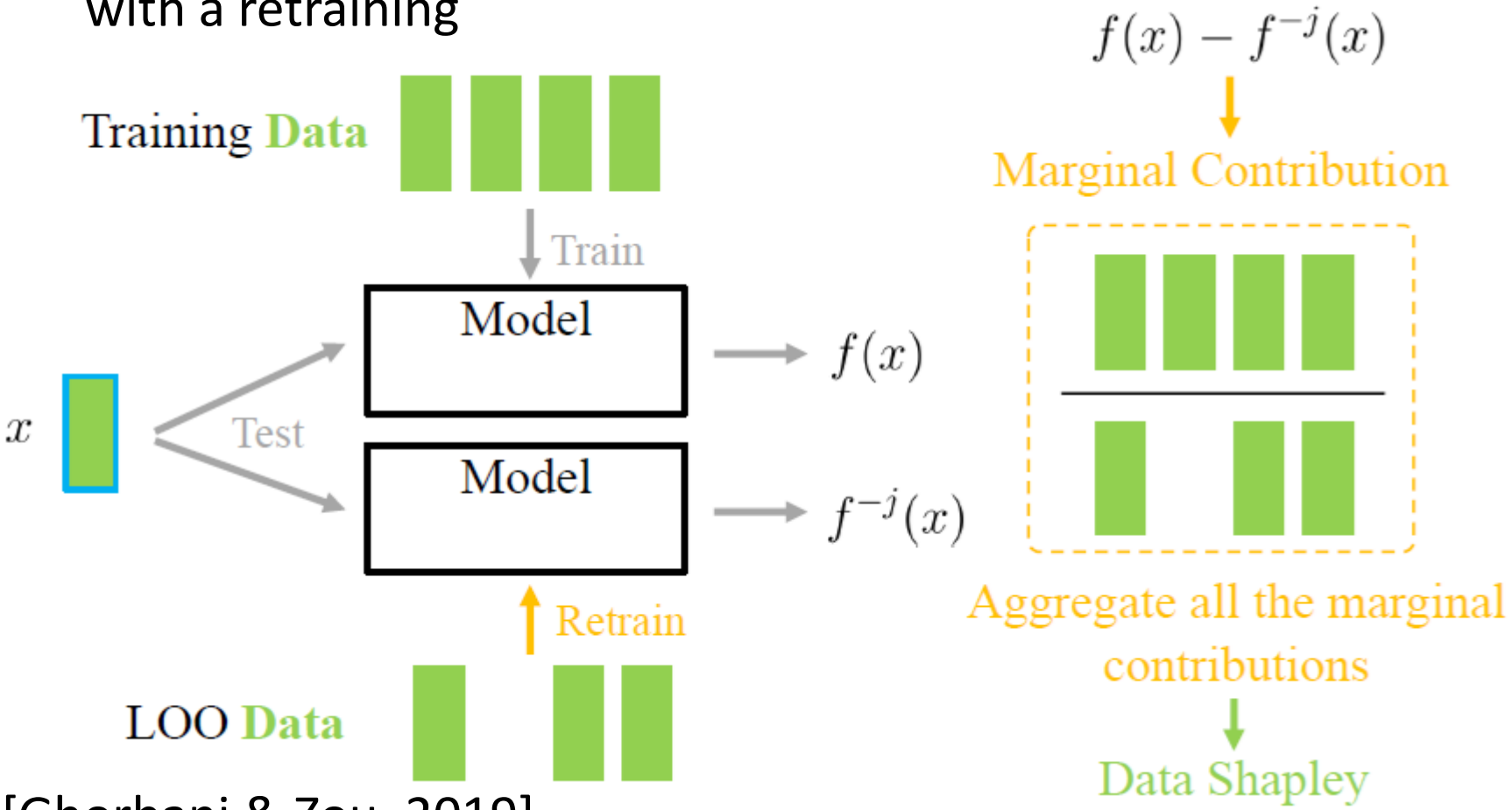
$$f(x) - f^{-j}(x) \longrightarrow \psi_j(x)$$



[Jia et al., 2021]

Perturbation-Based Data Attribution

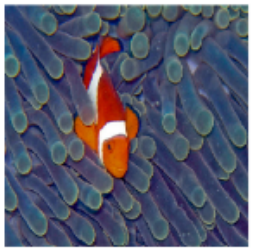
- Game-Theoretic Perturbation
  - Capture training data interactions: 2^n marginal contributions, each with a retraining



- [Ghorbani & Zou, 2019]

Gradient-Based Data Attribution

- No retraining (Perturbations)
  - Gradients for measuring similarities between data samples
  - Dot products of gradients evaluated at the training & test samples



$\mathcal{D}_{\text{train}} = \{x^{(1)}, \ldots, x^{(n)}\}$

Test sample Training sample

$\nabla_\theta \mathcal{L}(f_\theta(x))^T \nabla_\theta \mathcal{L}(f_\theta(x^{(j)})) \rightarrow \psi_j(x)$

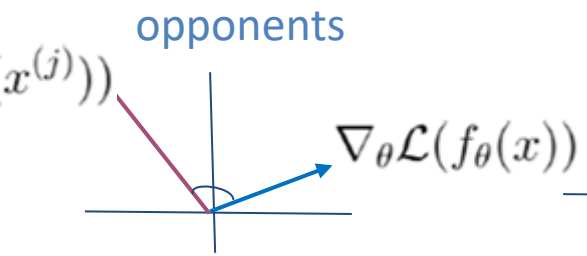
Train

Backward

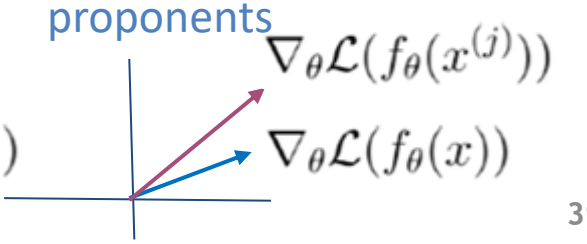
x Test Model θ $\mathcal{L}(\theta)$ $\nabla_\theta \mathcal{L}(f_\theta(x))$

$x^{(j)}$ $\mathcal{L}(\theta)$ $\nabla_\theta \mathcal{L}(f_\theta(x^{(j)}))$

opponents



proponents



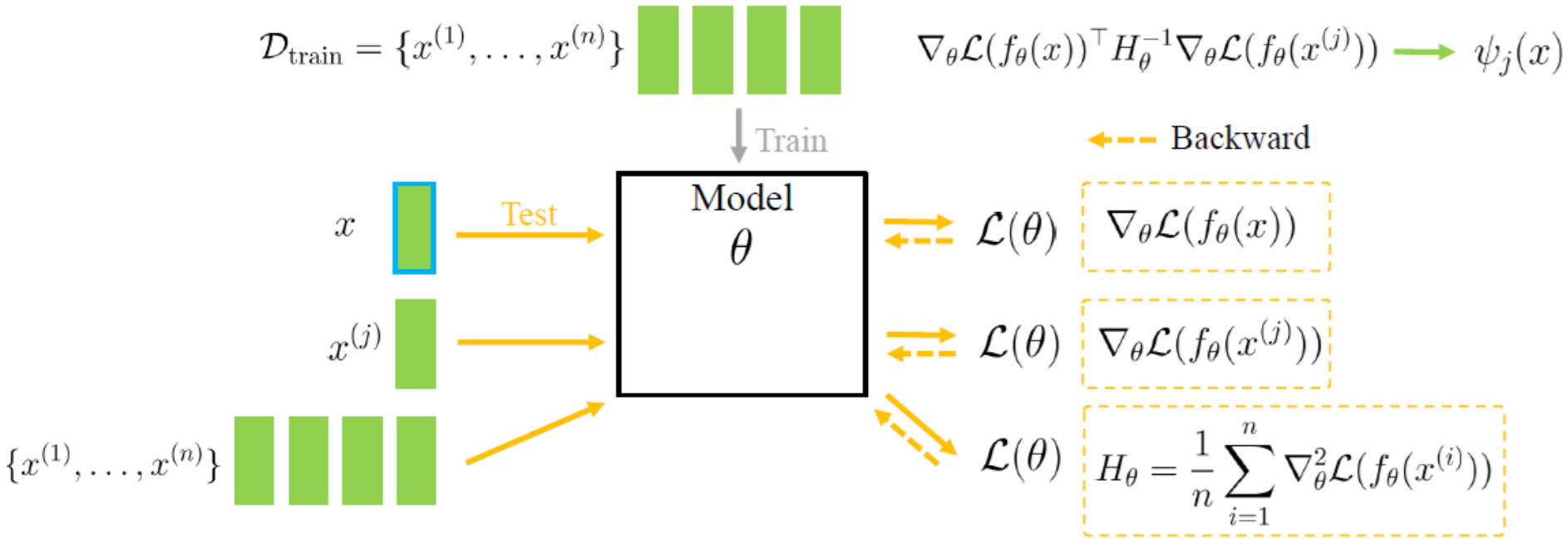
- [Charpiat et al., 2019]

Gradient-Based Data Attribution

- Influence Functions
  - Approximate LOO by lifting one data sample
  - Compute the Hessian matrix

Up weighting a training sample x^j by an *infinitesimal amount* ϵ leads to a new optimal model parameters

$$\theta_{\epsilon,x^{(j)}} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(x^{(i)}, \theta) + \epsilon \mathcal{L}(x^{(j)}, \theta)$$



- [Koh & Liang, 2017]

Taxonomy of Influence Functions

Individual Influence

measures the influence of a sample to another sample

In-Training

measures the influence **during** the learning process

TracIn [Pruthi et al., 2020]

SGDI [Hara et al., 2019]

Post-Training

measures the influence over an **optimal** model **after** learning is completed

FOIF [Koh et al., 2017]

RelatIF [Barshan et al., 2020]

Group Influence

measures the influence of a group of samples to another sample

In-Training

-

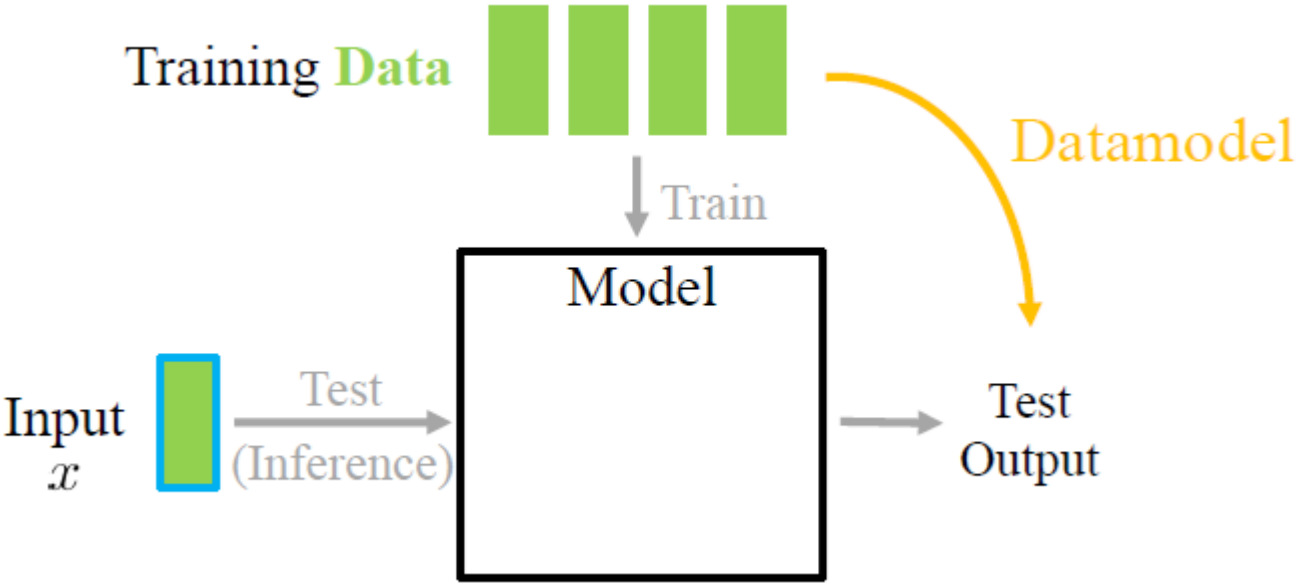
Post-Training

FOGIF [Koh et al., 2019]

SOGIF [Basu et al., 2020]

Linear Approximations for Data Attribution

- The Datamodel
  - Goal: skip training, directly predict model outputs from training data with a linear model
  - Collect counterfactual data for training the model: train a new model for each data sample $(z, f_z(x))$



$$g(z) = w^T z + b \quad s.t. \quad f(x) \approx g(z) \qquad z = \{1,0\}^n \qquad w_j \longrightarrow \psi_j(x)$$

- [Ilyas et al., 2022]

A Unified View of Attribution

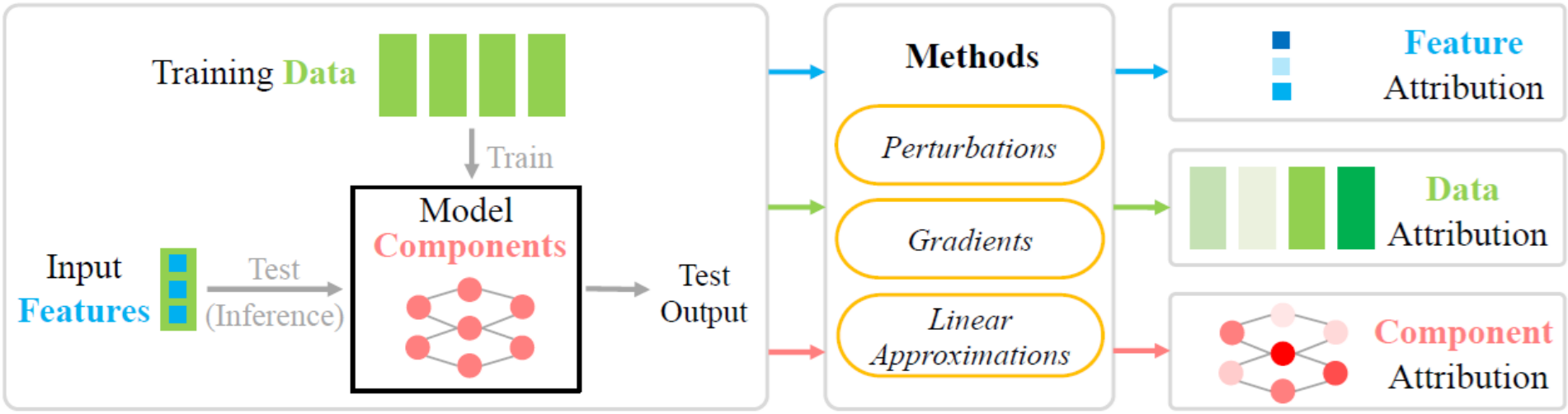


Table 1 Representative methods for feature, data, and component attribution, classified by core techniques into three categories, showing a unified view that these three types differ in perspective but share techniques.

Technique	Feature Attribution	Data Attribution	Component Attribution
Perturbation	Occlusions [Zeiler and Fergus, 2014] SHAP [Lundberg and Lee, 2017] Masking Model [Dabkowski and Gal, 2017]	LOO [Cook and Weisberg, 1982] Data Shapley [Ghorbani and Zou, 2019]	Causal Tracing [Meng et al., 2022] Neuron Shapley [Ghorbani and Zou, 2020] Subnetwork Probing [Cao et al., 2021]
Gradient	(Vanilla) Gradients [Simonyan et al., 2013] SmoothGrad [Smilkov et al., 2017]	GradDot [Pruthi et al., 2020] Influence Function [Koh and Liang, 2017]	Attribution Patching [Nanda, 2023]
Linear Approximation	LIME [Ribeiro et al., 2016]	Datamodels [Ilyas et al., 2022]	COAR [Shah et al., 2024]

S. (Ray) Zhang, H. Lakkaraju, J. Adebayo Explain AI Models: Methods and Opportunities in Explainable AI, Data-Centric AI, and Mechanistic Interpretability https://arxiv.org/abs/2501.18887

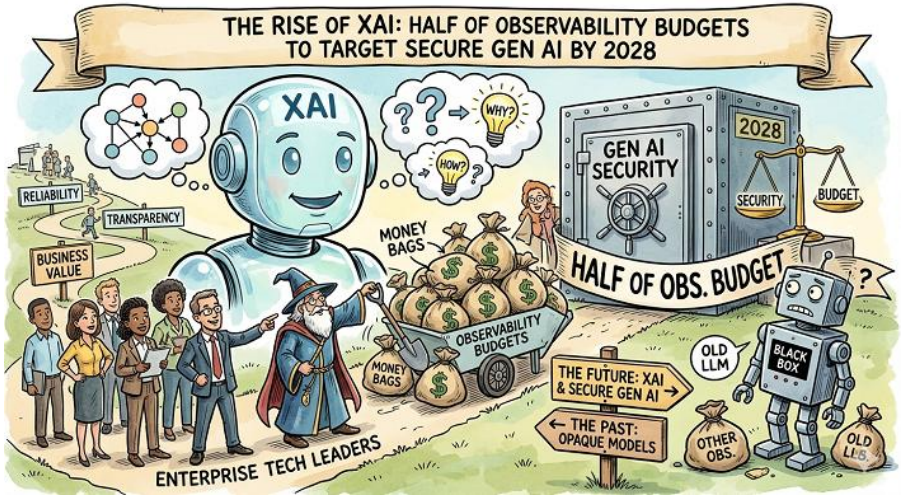
XAI Challenges

XAI Computational Overhead

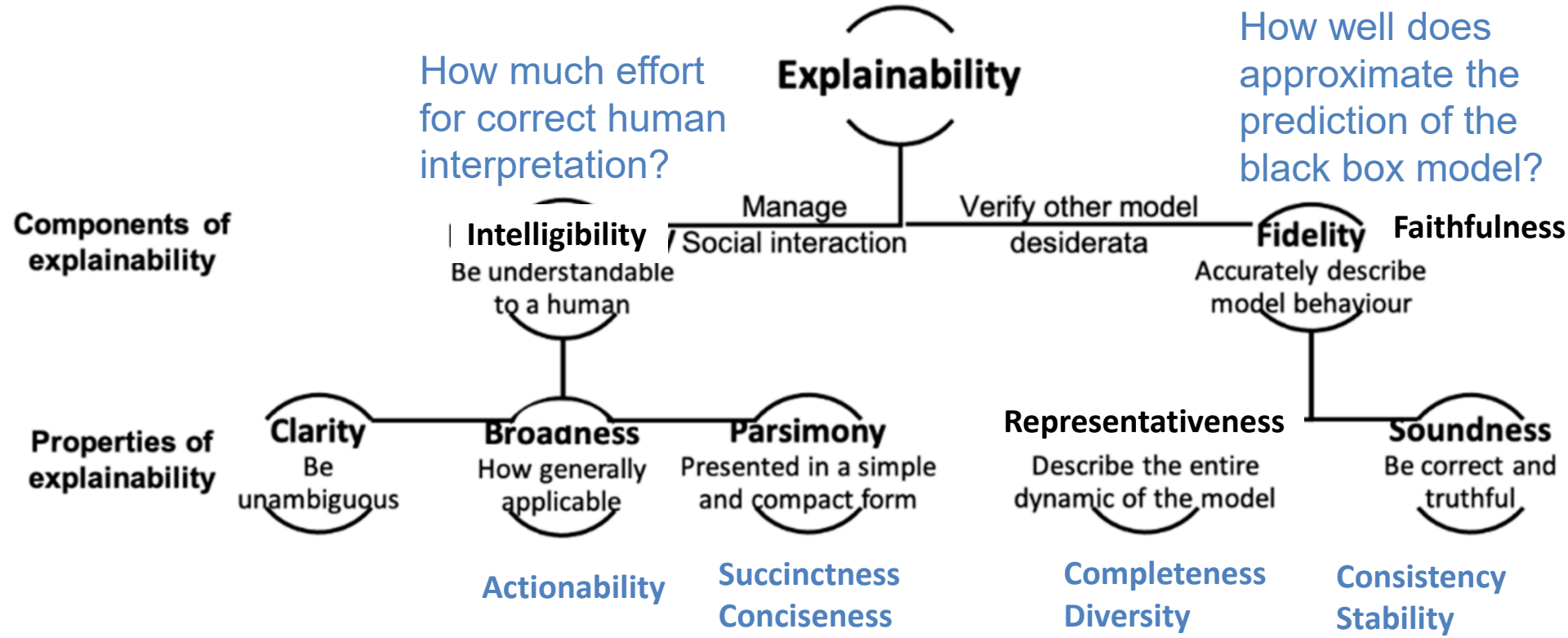
XAI Consistency/Robustness

XAI Empirical Studies

XAI for GenAI

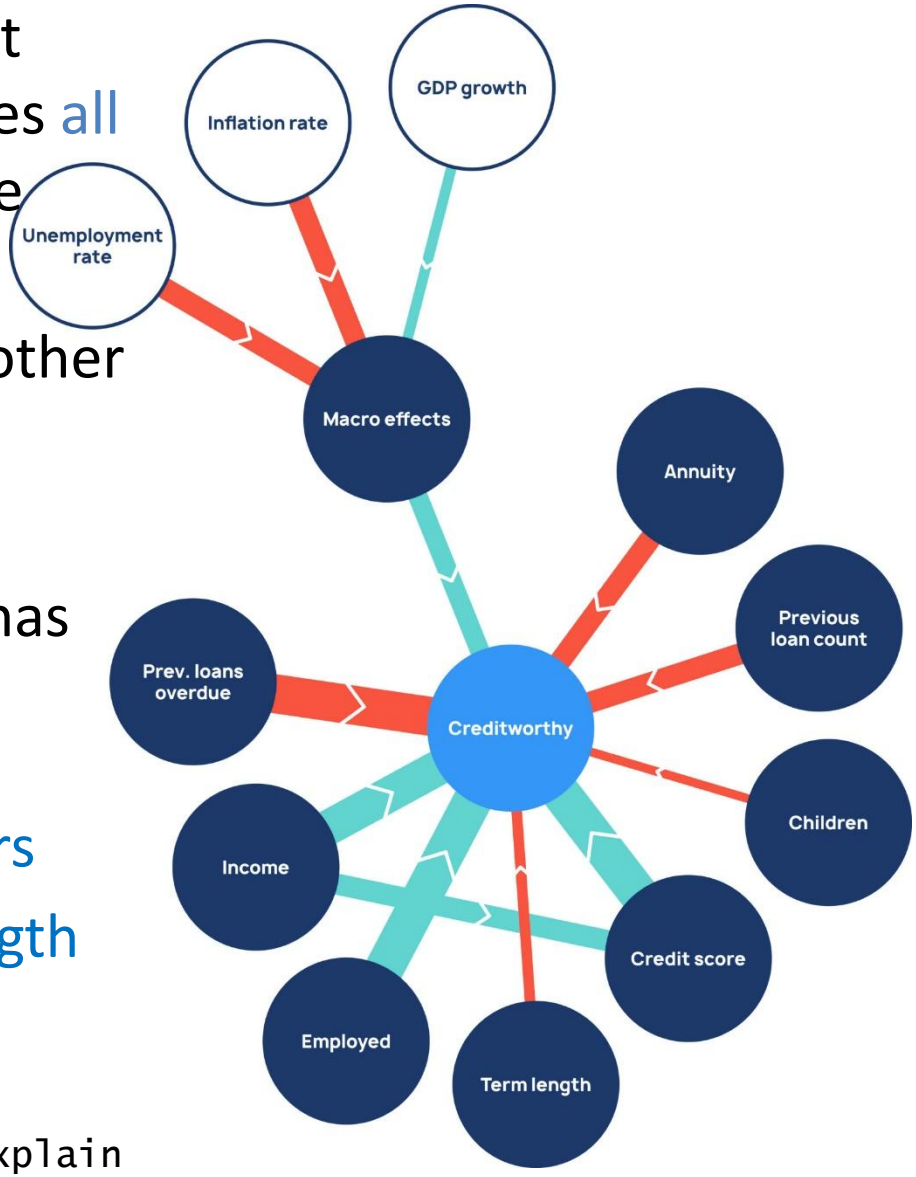


XAI: One Objective, Many Metrics



Towards Causal Explanations

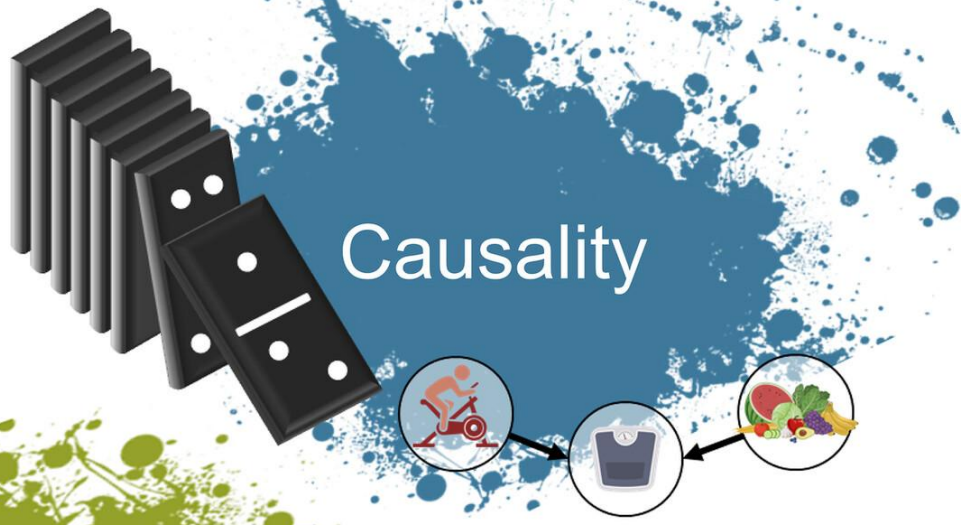
- Causal models contain a transparent qualitative component that describes all cause-and-effect relationships in the data
  - Causal models don't require another model to approximate them



- Red arrows indicate that a feature has an inverse relationship with credit-worthiness, while green arrows correspond to positive causal drivers
- Arrow thickness indicates the strength of the causal relationship

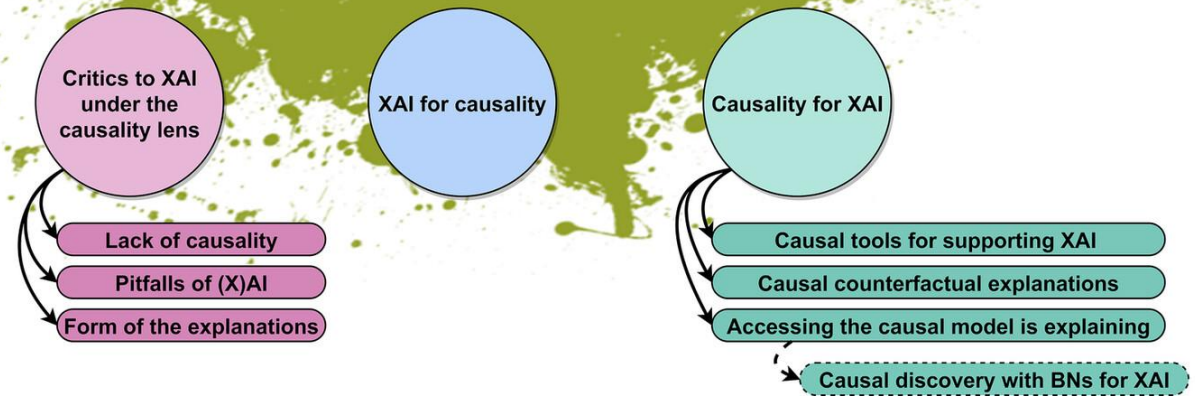
https://www.causalens.com/blog/xai-doesnt-explain

XAI and Causality



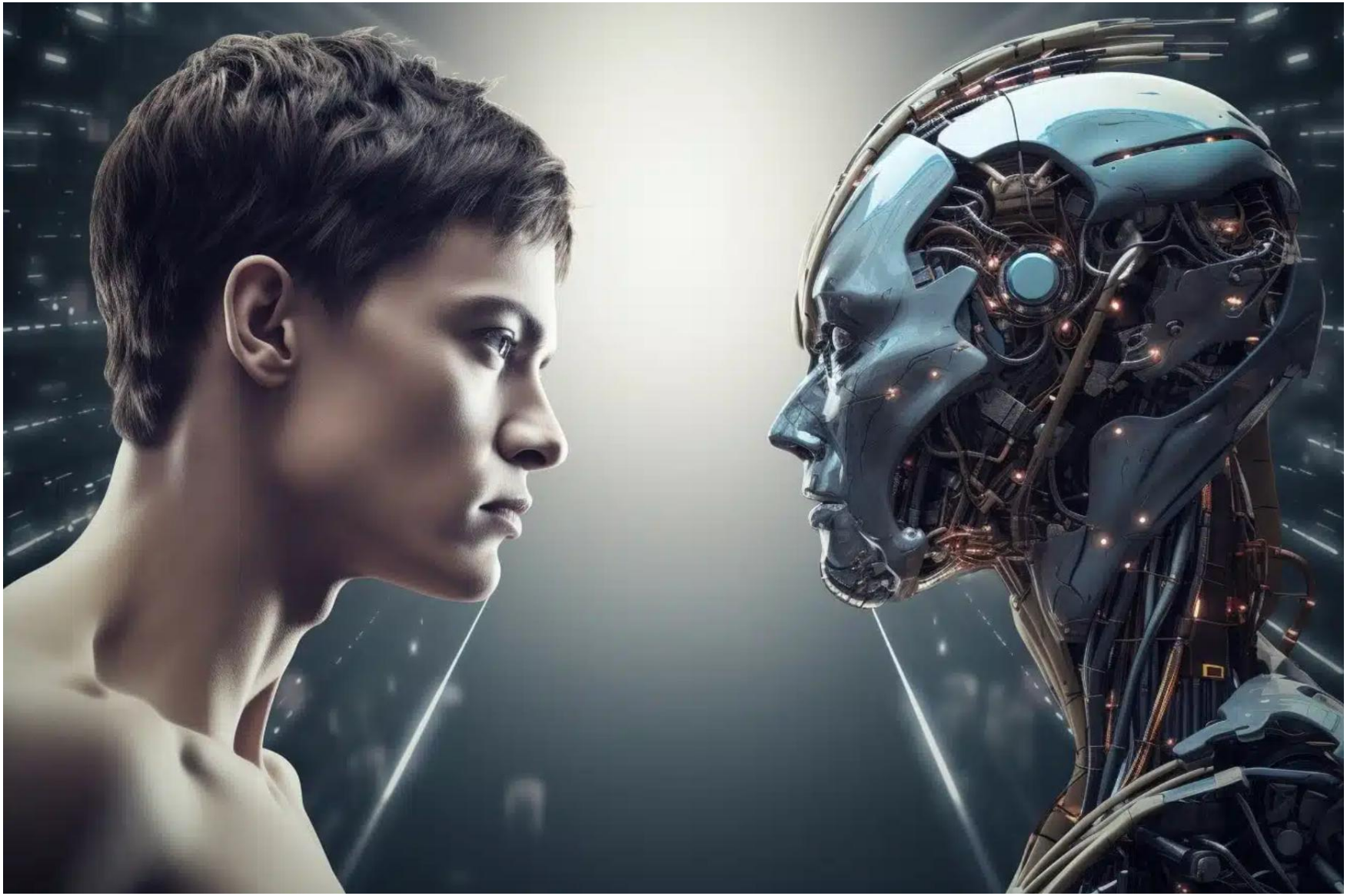
Eliminating Spurious Correlations
Generating Actionable Counterfactuals
Answering "What-If" Questions (Interventions)

How are they related?



https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.70015

Questions?



Local Approaches for Post hoc Explainability

- Feature Importance
  - LIME (Ribeiro et al., 2016) - explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction
  - SHAP (Lundberg et al., 2017) - assigns each feature an importance value for a particular prediction and fairly attributes the prediction to all the features
- Rule Based
  - Anchors (Ribeiro et al., 2018) - explains the behavior of complex models with high-precision rules, representing local, “sufficient” conditions for predictions
  - LORE (Guidotti et al., 2018) - learns a local interpretable predictor on a synthetic neighborhood generated by a genetic algorithm. Then, it derives from the logic of the local interpretable predictor a meaningful explanation consisting of a decision rule and a set of counterfactual rules
- Counterfactuals
  - DiCE (Mothilal et al., 2020) - generating and evaluating a diverse set of counterfactual explanations based on determinantal point processes
  - FACE (Poyiadzi et al., 2020) - generates counterfactuals that are coherent with the underlying data distribution and supported by the “feasible paths” of change, which are achievable and can be tailored to the problem at hand

Local Approaches for Post hoc Explainability

- Saliency Maps
  - Layer-Wise Relevance Propagation (Bach et al., 2015) - assumes that the classifier can be decomposed into several layers of computation. Such layers can be parts of the feature extraction from the image or parts of a classification algorithm run on the computed features
  - Integrated Gradients (Sundararajan et al., 2017) - they do not need any instrumentation of the network, and can be computed easily using a few calls to the gradient operation, allowing even novice practitioners to easily apply the technique
- Prototypes/Example Based
  - Prototype Selection (Bien et al., 2011) - a good set of prototypes for a class should capture the full structure of the training samples of that class while taking into consideration the structure of other classes
  - TracIn (Pruthi et al., 2020) - computes the influence of a training sample on a prediction made by the model by tracing how the loss on the test point changes during the training process whenever the training sample of interest was utilized

Global Approaches for Post hoc Explainability

- Representation Based
  - Network Dissection (Bau et al., 2017) - quantifying the interpretability of latent representations of CNNs by evaluating the alignment between individual hidden units and a set of semantic concepts
  - Compositional Explanation (Mu et al., 2020)- automatically explaining logical and perceptual abstractions encoded by individual neurons in deep networks and generate explanations by searching for logical forms defined by a set of composition operators over primitive concepts
- Model Distillation
  - LGAE (Tan et al., 2019) - leverage model distillation to learn global additive explanations that describe the relationship between input features and model predictions. These global explanations take the form of feature shapes, which are more expressive than feature attributions
  - Decision Trees as global explanations (Bastani et al., 2017) - generate new training data by actively sampling new inputs and labeling them using the complex model, since they are nonparametric
- Summaries of Counterfactuals
  - AReS (Rawal et al., 2020) - construct global counterfactual explanations which provide an interpretable and accurate summary of recourses for the entire population