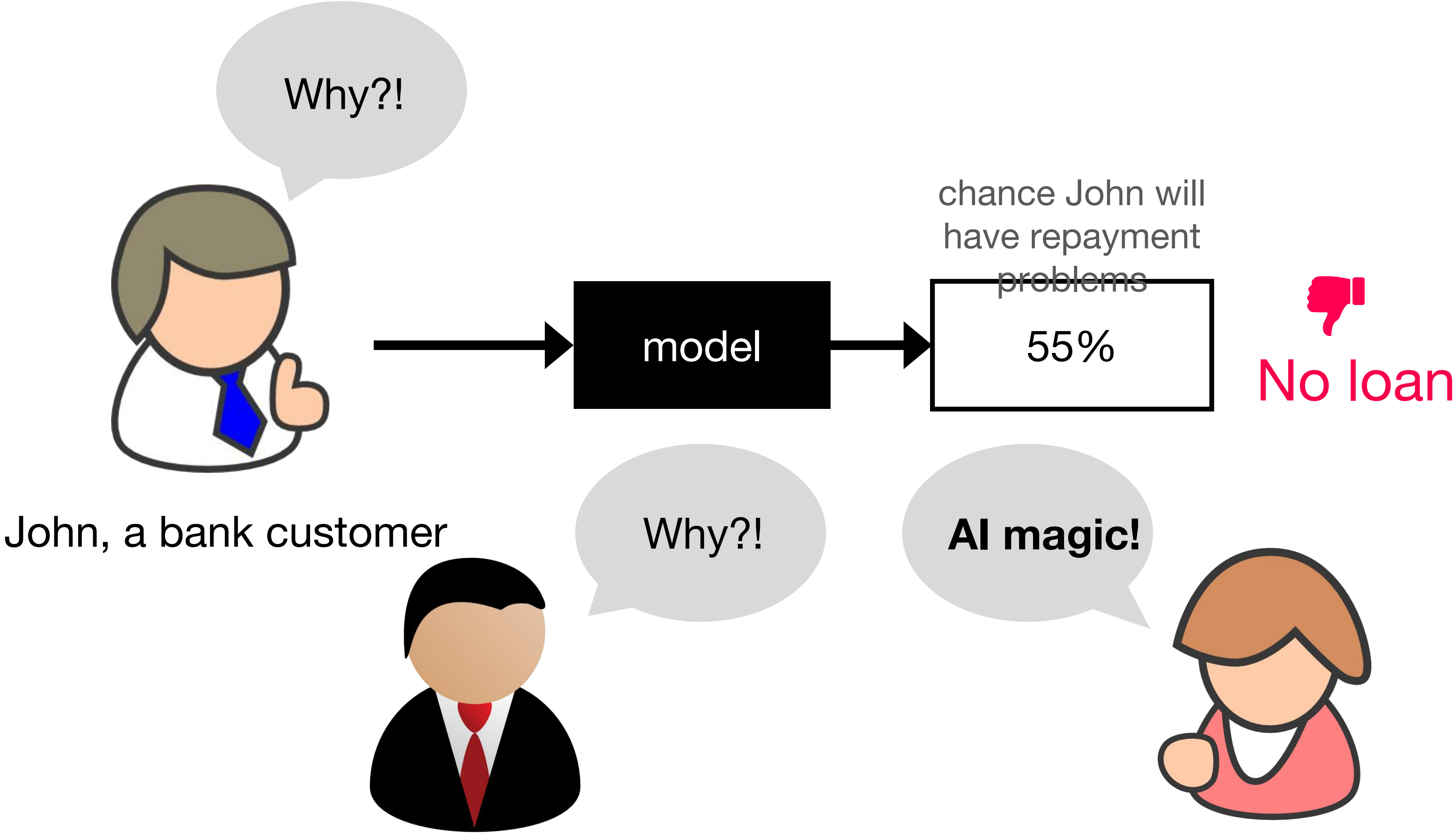
Report of Explainable Machine Learning Models And The Application In Real-Rorld Use Case

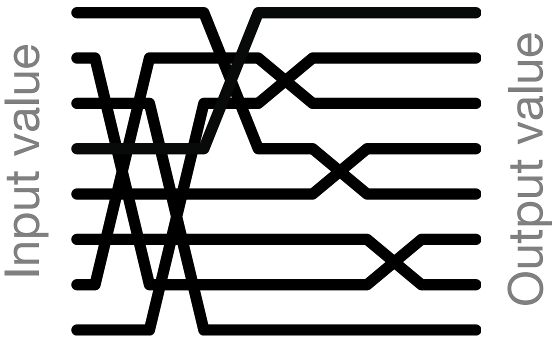
Ningjing He

Zhejiang University of Finance and Economics

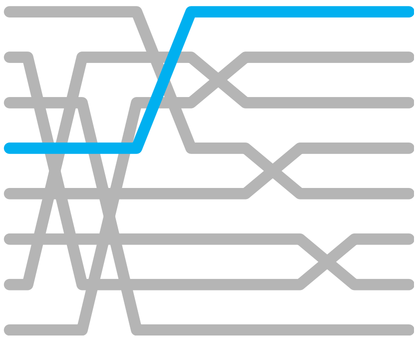

Why?!
chance John will have repayment problems
model
55%
No loan
John, a bank customer
Why?!
AI magic!

	Interpretabl e	Accurate
Complex model		
Simple model		

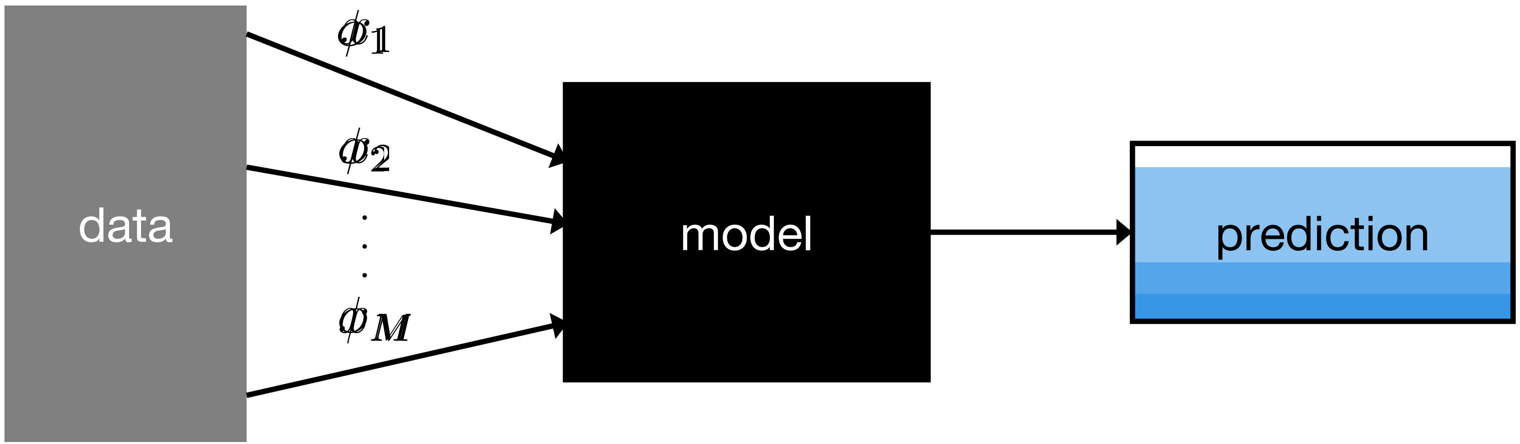
Interpretable or accurate: **choose one**.

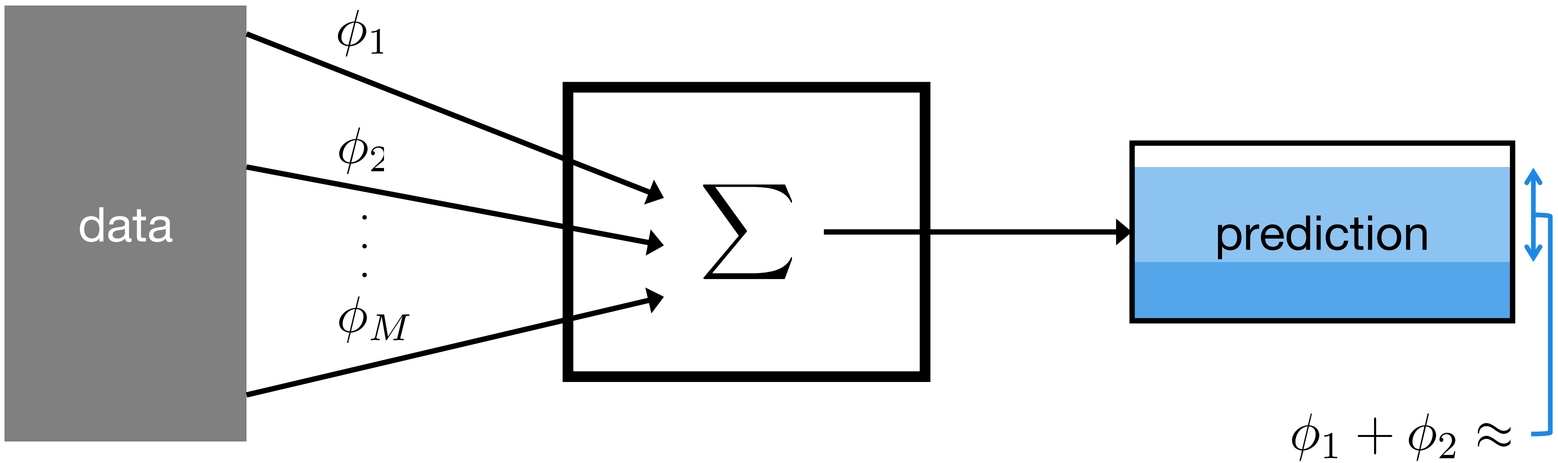


Complex models are inherently complex!

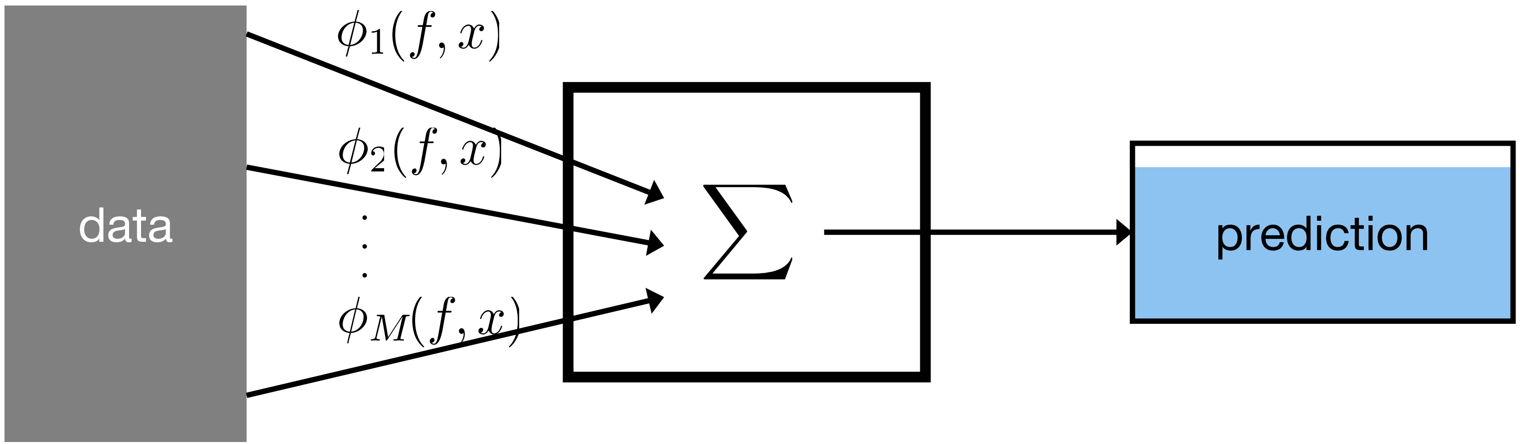


But a single prediction involves only a small piece of that complexity.


data
ϕ_1
ϕ_2
ϕ_M
model
prediction



data
ϕ_1
ϕ_2
$\vdots$
ϕ_M
Σ
prediction
$\phi_1 + \phi_2 \approx$


data
$\phi_1(f, x)$
$\phi_2(f, x)$
$\phi_M(f, x)$
Σ
prediction

LIME

Ribeiro et al. 2016

Shapley reg. values

Lipovetsky et al. 2001

QII

Datta et al. 2016

$$\boxed{\Sigma}$$

Shapley sampling

Štrumbelj et al. 2011

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

Path expectations

Saabas 2014

Additive feature attribution methods

LIME

Ribeiro et al. 2016

DeepLIFT

Shrikumar et al. 2016

$$\boxed{\Sigma}$$

Shapley reg. values

Lipovetsky et al. 2001

Relevance prop.

Bach et al. 2015

QII

Datta et al. 2016

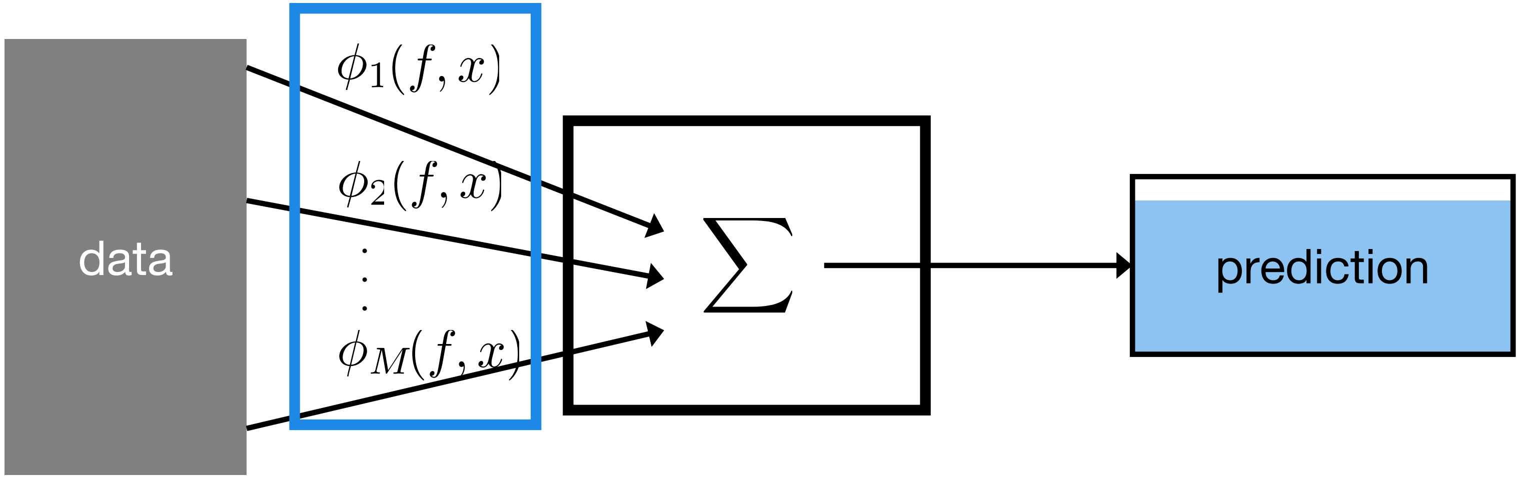
Shapley sampling

Štrumbelj et al. 2011

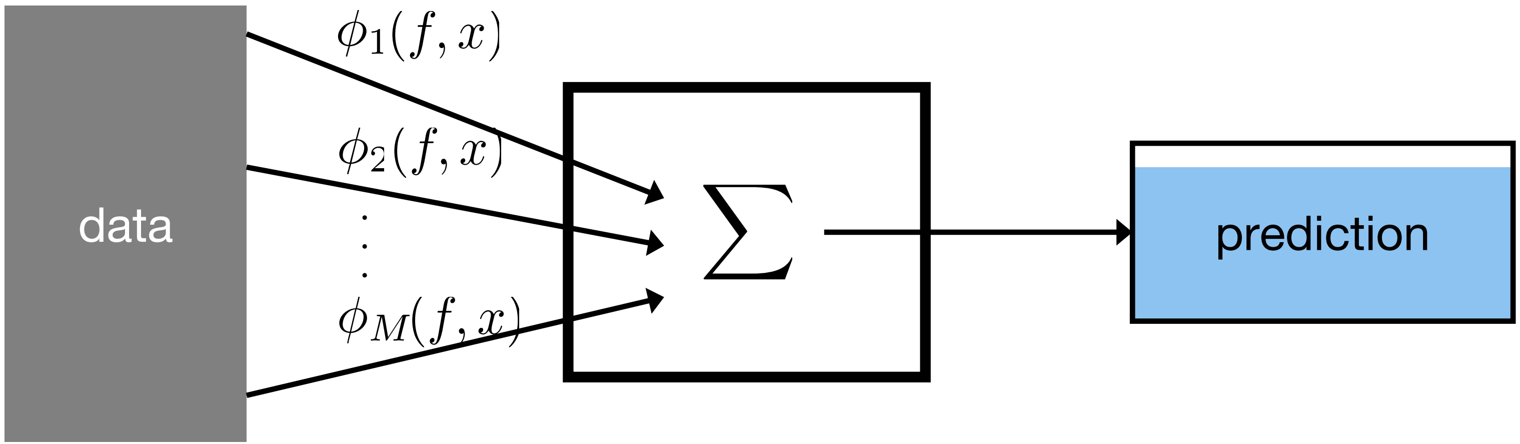
Path expectations

Saabas 2014

Additive feature attribution methods

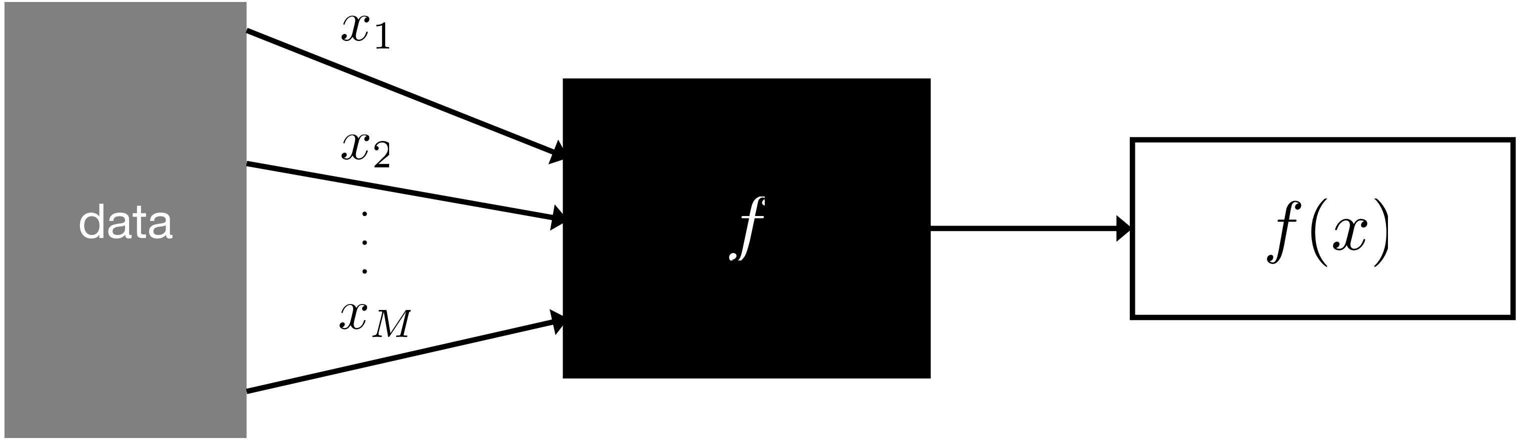


Additive feature attribution methods



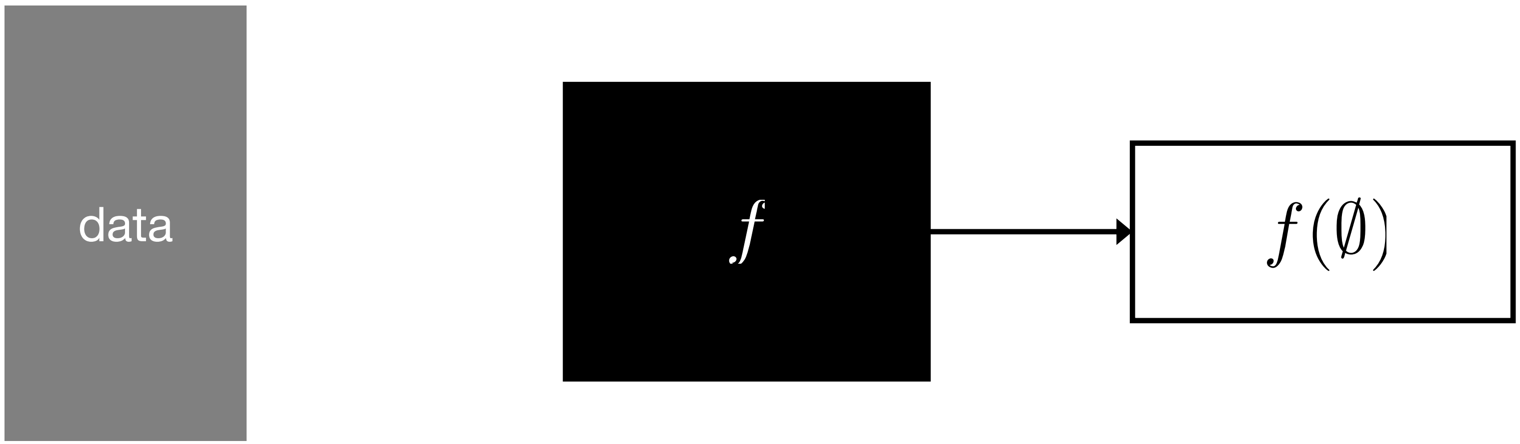
1 Local accuracy

$$\sum_{i=0}^{M} \phi_i = f(x), \quad \phi_0 = f(\emptyset)$$



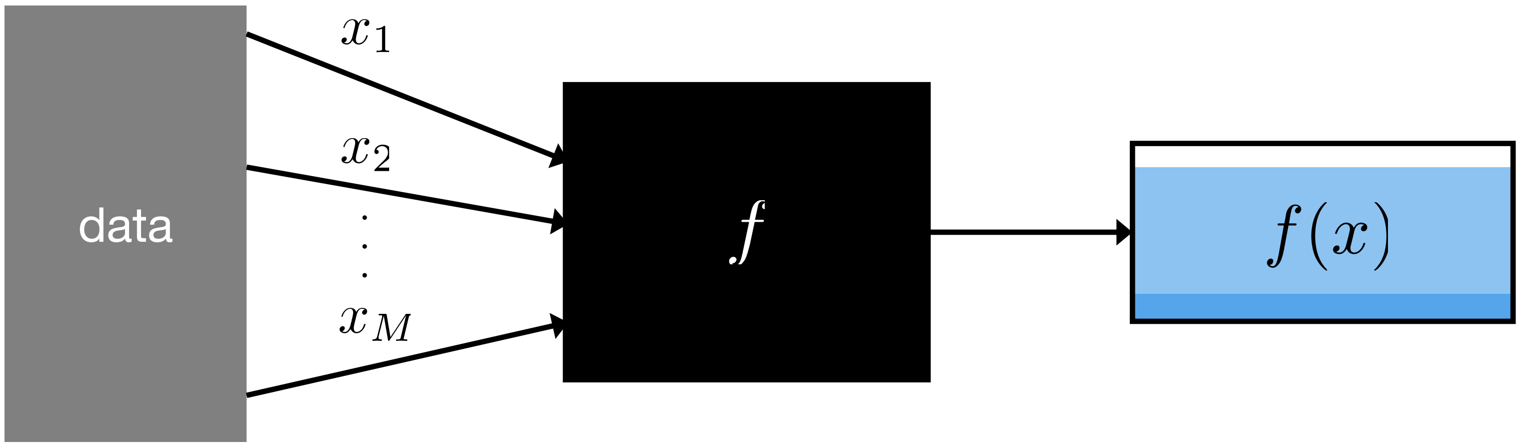
1 Local accuracy

$$\sum_{i=1}^{M} \phi_i(f, x) = f(x) - f(\emptyset)$$

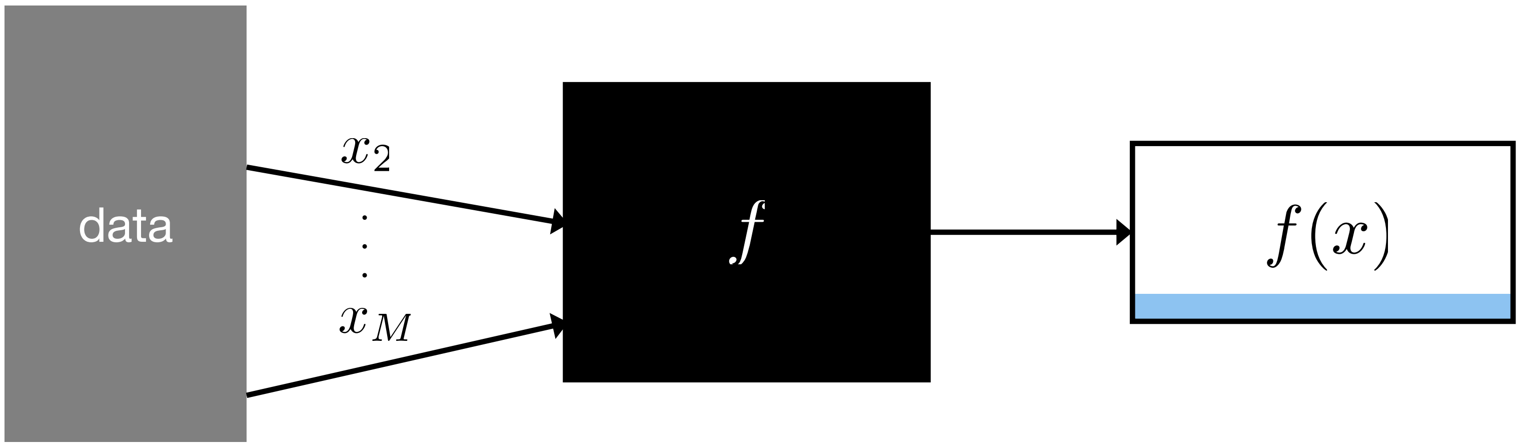


1 Local accuracy

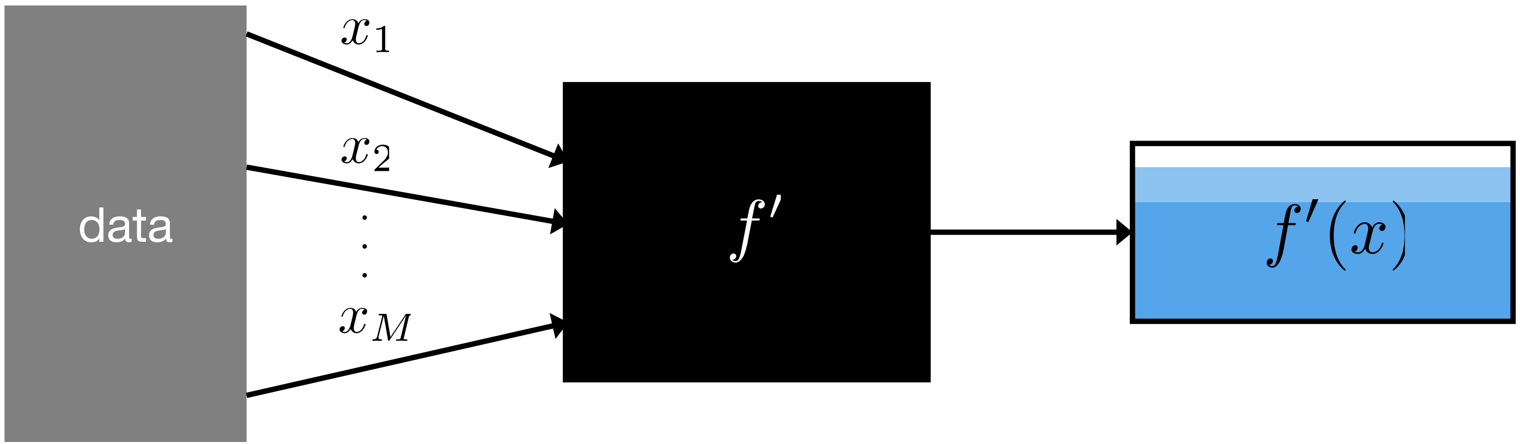
$$\sum_{i=1}^{M} \phi_i(f, x) = f(x) - f(\emptyset)$$



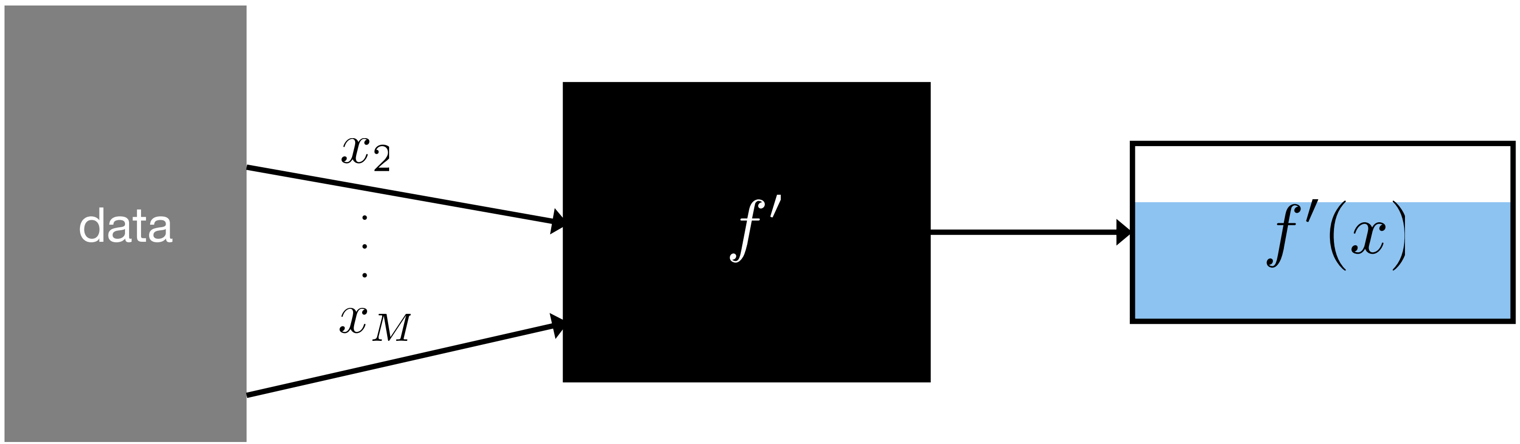
2 Consistency



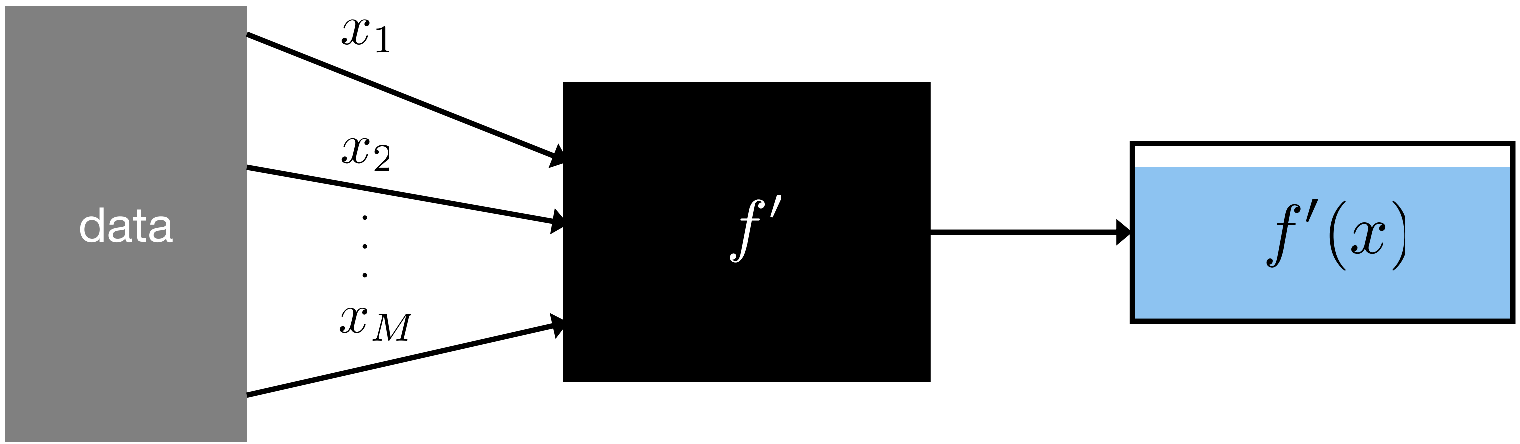
2 Consistency



2 Consistency



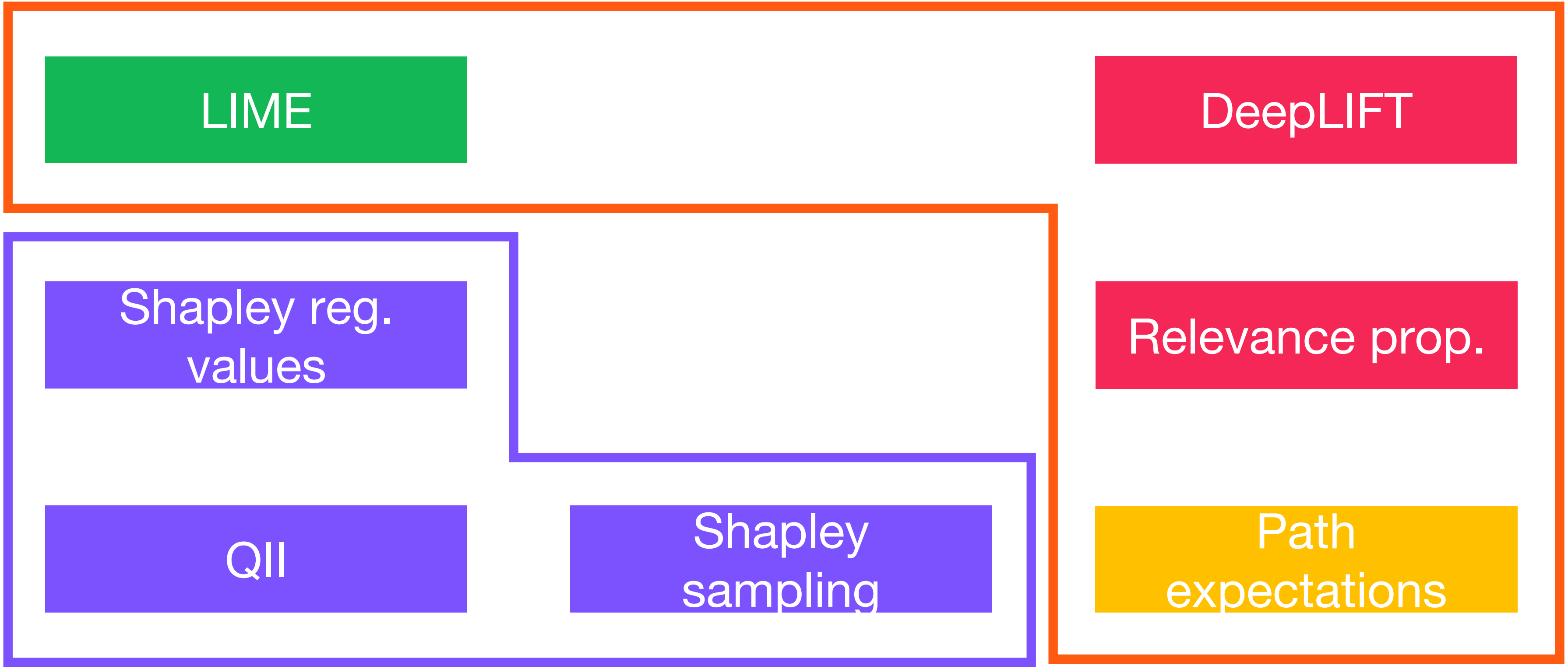
2 Consistency



2 Consistency

$$\phi_1(f, x) \geq \phi_1(f', x)$$

Additive feature attribution methods



LIME

DeepLIFT

SHAP

Shapley reg. values

Relevance prop.

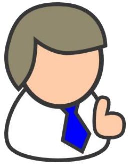
QII

Shapley sampling

Path expectations

SHapley Additive exPlanation (SHAP) values

The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction.



Base rate

20%

Prediction for John

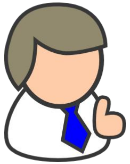
55%

0 $\quad E[f(x)] \quad f(x)$

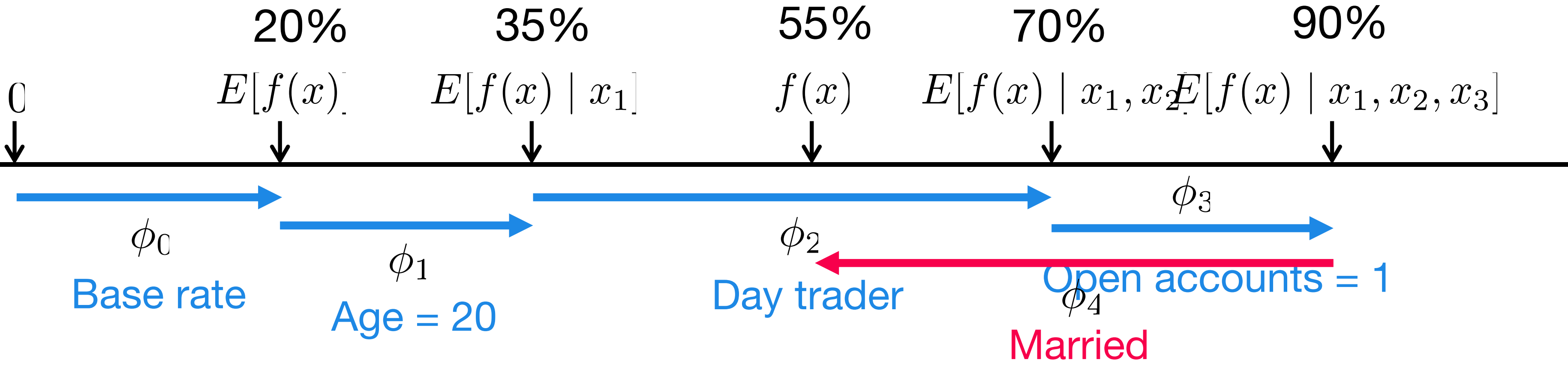
How did we get here?

SHAP values add up to the difference between the expected model output and the actual output for a given input. This means that SHAP values provide an accurate and local interpretation of the model's prediction for a given input.

SHapley Additive exPlanation (SHAP) values



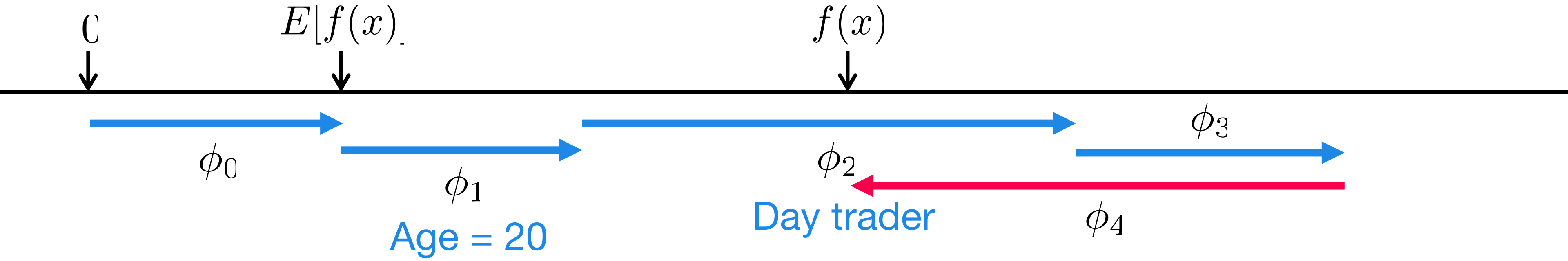
When we are explaining a prediction f(x_i), the SHAP value for a specific feature i is just the difference between the expected model output and the partial dependence plot at the feature's value x_i

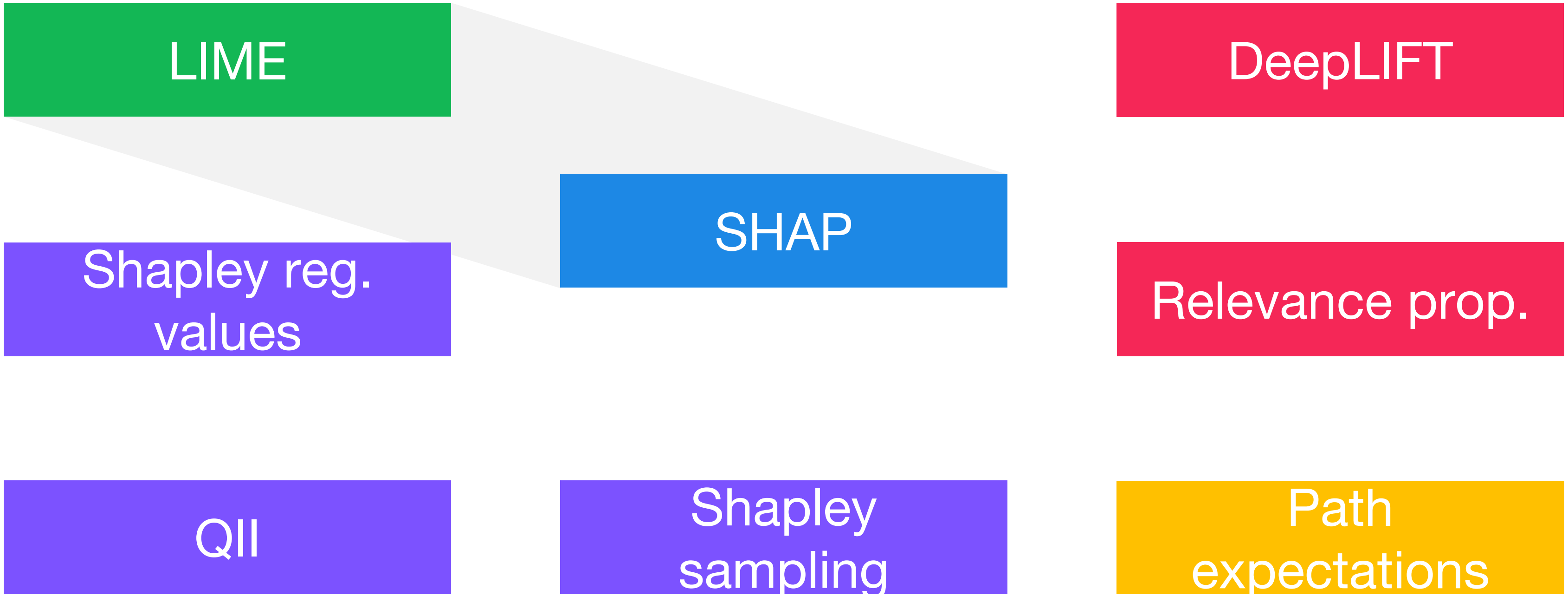


SHapley Additive exPlanation (SHAP) values

The order matters!

SHAP values result from averaging over all N! possible orderings.




LIME
DeepLIFT
SHAP
Shapley reg. values
Relevance prop.
QII
Shapley sampling
Path expectations

LIME —— Local Interpretable Model-agnostic Explanations

The proximity measure

$\pi_{x'}$ defines how large the neighborhood around instance x is that we consider for the explanation.

The loss function to force g to well approximate f

Optional regularization of g

$$\xi = \underset{g \in \mathcal{G}}{\arg\min} \; L(f, g, \pi_{x'}) + \Omega(g)$$

Kernel specifies what 'local' means

A class of interpretable models (linear models)

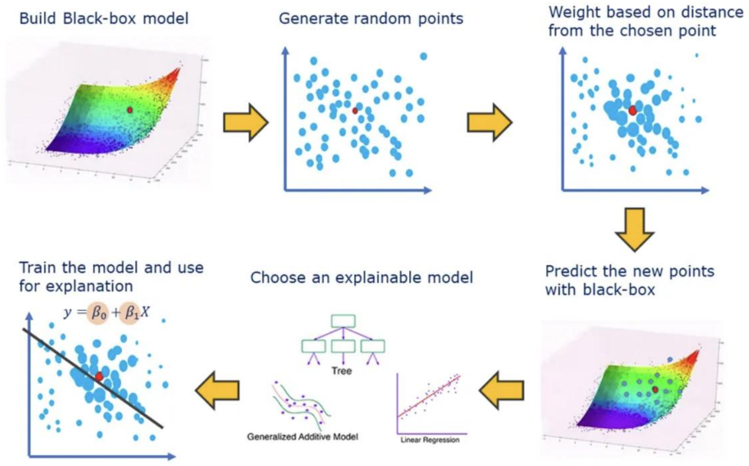
G is the family of possible explanations, for example all possible linear regression models。

LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model.

On this new dataset LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.

LIME Algorithm

- Choose the ML model and a reference point to be explained
- Generate points all over the $\mathbb{R}^p$ space
  (sample X values from a Normal distribution inferred from the training set)
- Predict the Y coordinate of the sampled points, using the ML model
  (the generated points are guaranteed to perfectly lie on the ML surface)
- Assign weights based on the closeness to the chosen point
  (use RBF Kernel, it assigns higher weights to points closer to the reference)
- Train Linear Ridge Regression on the generated weighted dataset:
  $E(Y) = \beta_0 + \sum \beta_j X_j$. *The β coefficients are regarded as LIME explanation.*

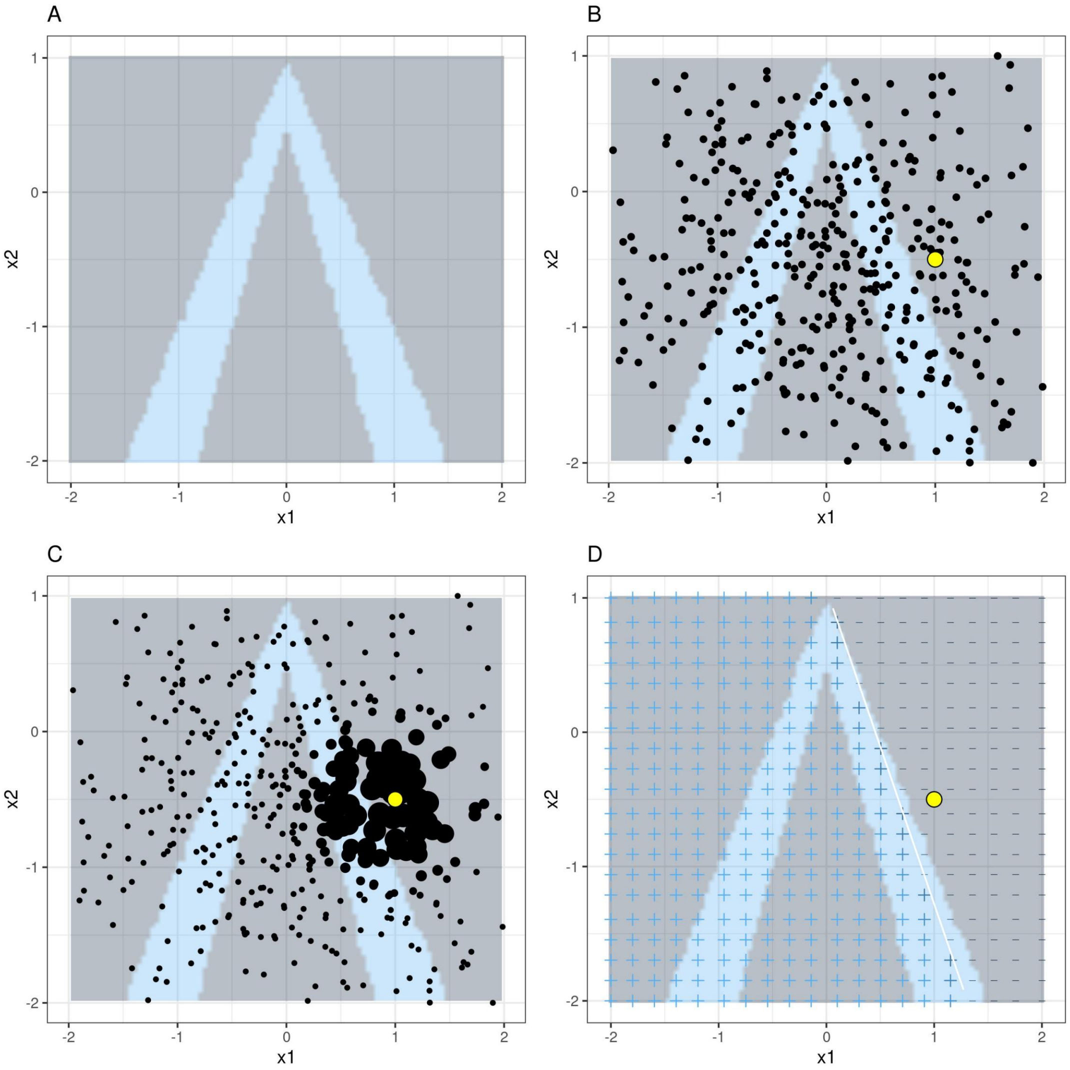


LIME algorithm for tabular data.
A) Random forest predictions given features x1 and x2. Predicted classes: 1 (dark) or 0 (light).

B) Instance of interest (big dot) and data sampled from a normal distribution (small dots).

C) Assign higher weight to points near the instance of interest.

D) Signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary (P(class=1) = 0.5).



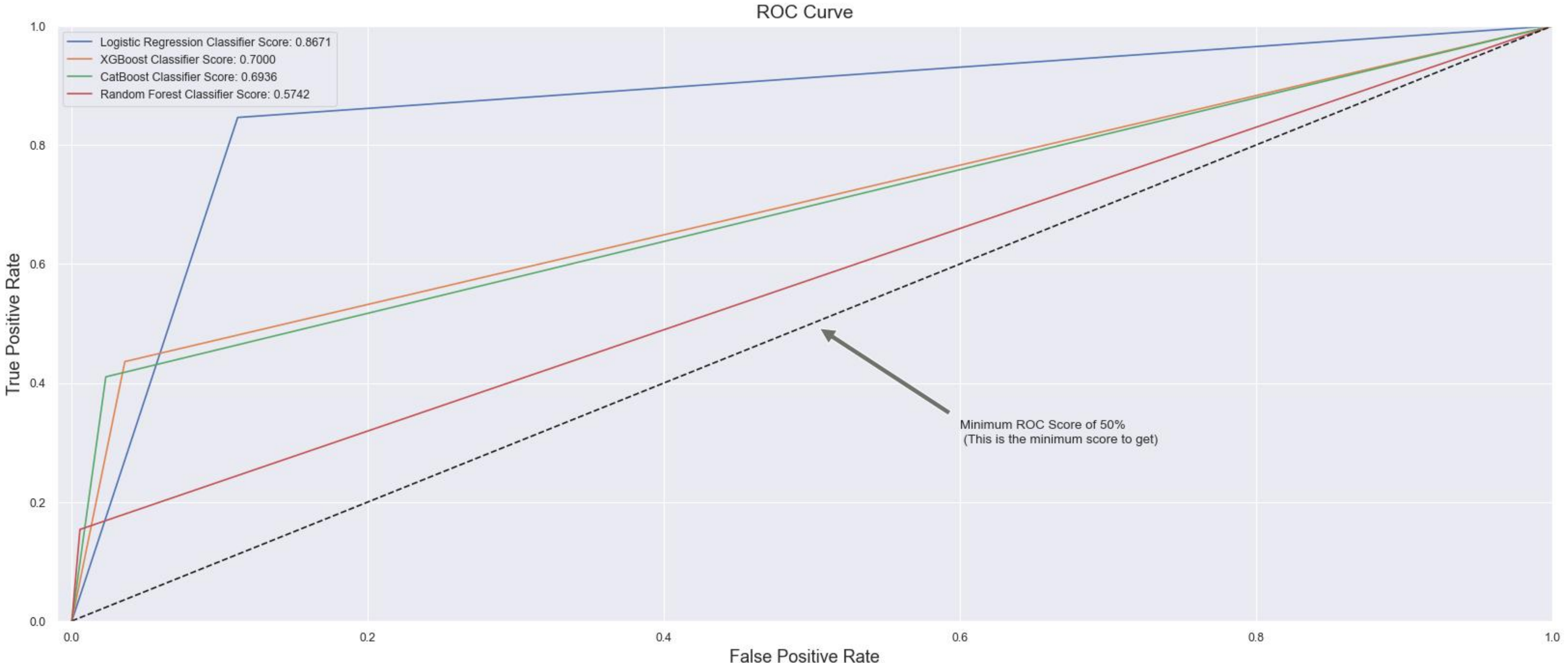
How can we trust their explanation?

Make a comparisonwhen ground-truth of feature importance is available

1. Obtain the ground-truth feature importance on generated dataset.

2. We compare SurvShap and SurvLIME by utilizing Kendall's rank correlation coefficient and assessing the local accuracy of the model's prediction regarding feature importance.

Real-World Use Case: Predicting Company Bankruptcy in Taiwan



Real-World Use Case: Predicting Company Bankruptcy in Taiwan