COMP6211I: Trustworthy Machine Learning

Interpretability (XAI) part 2

Introduction

- Model Specific vs. Model Agnostic
- Global Methods vs. Local Methods
- Pre-Model vs. In-Model vs. Post-Model
- Surrogate Methods vs. Visualization Methods

Can it explain a particular model or many models?

Does it explain a particular sample or entire model?

When does it occur?

Does it work separately from the model, or does it visualize the model?

The categories are non-exclusive. There is no universally accepted taxonomy of XAI techniques!

Introduction

- **Model Specific vs. Model Agnostic**

Model-specific interpretation methods are based on the parameters of the individual models.

Model Agnostic methods are mainly applicable in post-hoc analysis and not limited to specified model architecture.

Introduction

- Global Methods vs. Local Methods

Global methods concentrate on the inside of a model by exploiting the overall knowledge about the model, the training, and the associated data.

Local interpretable methods are applicable to a single outcome of the model. This can be done by designing methods that can explain the reason for a particular prediction or outcome.

Introduction

LIME

- Title: “Why Should I Trust You?” Explaining the Predictions of Any Classifier
- Conference: KDD2016
- Authors: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (University Of Washington)

SHAP

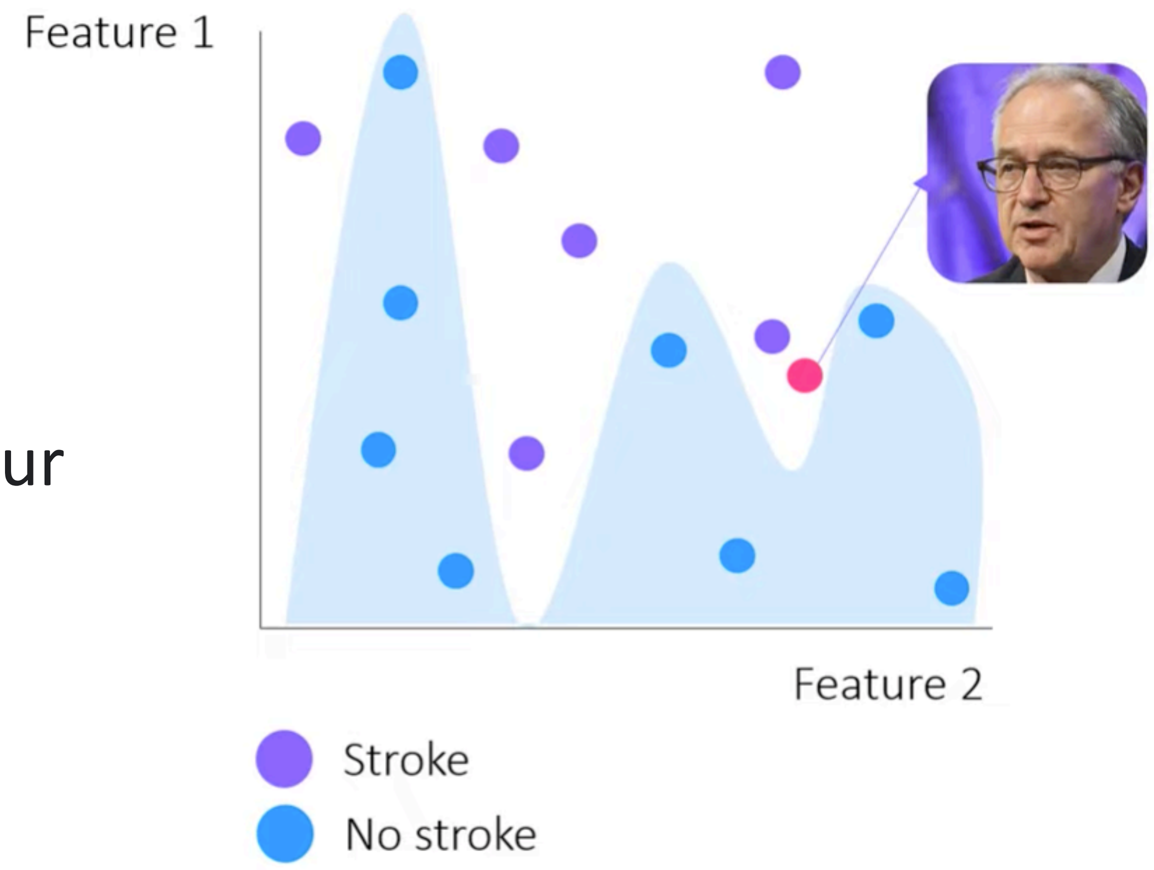
- Title: A unified approach to interpreting model predictions
- Conference: NIPS2017
- Authors: Scott M. Lundberg, Su-In Lee (University Of Washington)

LIME: Local interpretable model-agnostic explanations

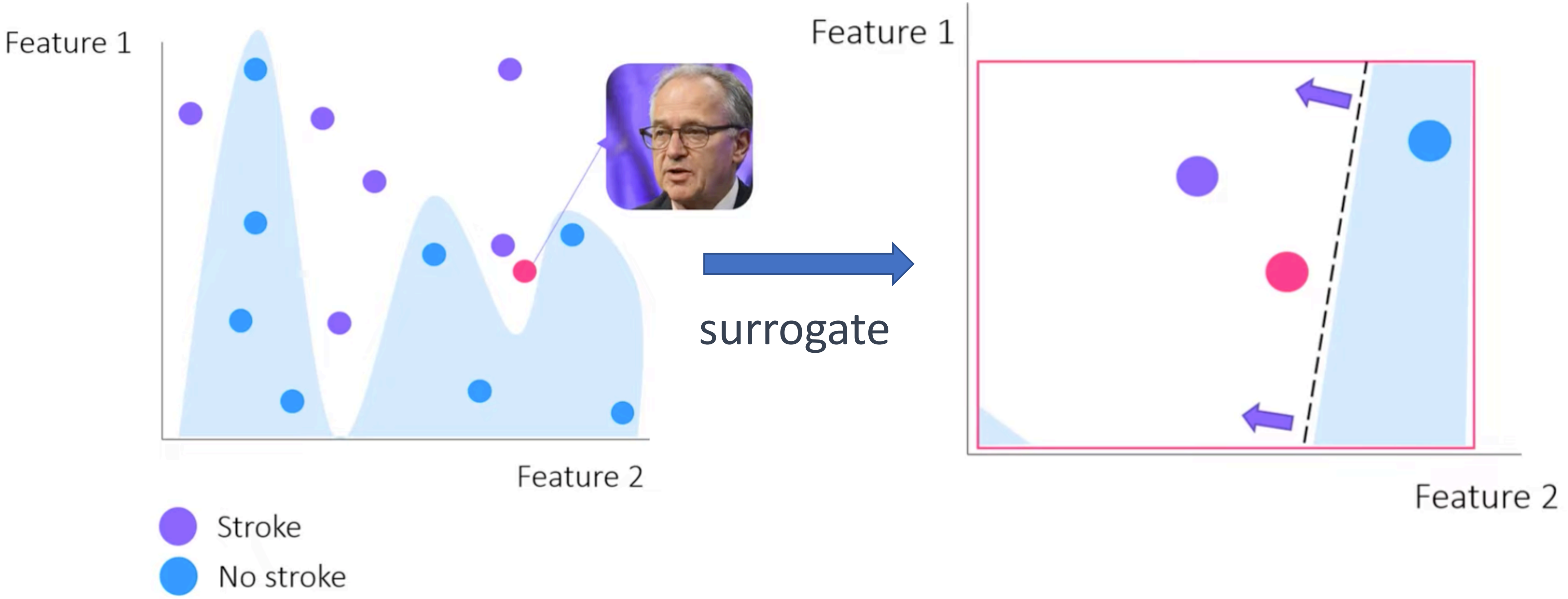
Task: Stroke Prediction

Feature 1: age
Feature 2: body mass index

How could we explain to him why our model outputs stroke?



LIME: Local interpretable model-agnostic explanations



LIME: Local interpretable model-agnostic explanations

- Works on any black-box model
- Model internals are “hidden”
- Works with many data types
- Using prior knowledge we can validate the explanations and create trust
- Explanations are locally faithful, but not necessarily globally

The Math in LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Family of interpretable models

Complex model

Simple interpretable model

Age | Gender | Heart Disease | ...

The Math in LIME

$$\xi(x) = \underset{g\in G}{\operatorname{argmin}} \; \overbrace{\mathcal{L}(f, g, \pi_x)}^{\text{Good approximation}} + \Omega(g)$$

Family of interpretable models

Complex model

Simple interpretable model

Age | Gender | Heart Disease | ...

The Math in LIME

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \overbrace{\mathcal{L}(f, g, \pi_x)}^{\text{Good approximation}} + \overbrace{\Omega(g)}^{\text{Stay simple}}$$

Complex model

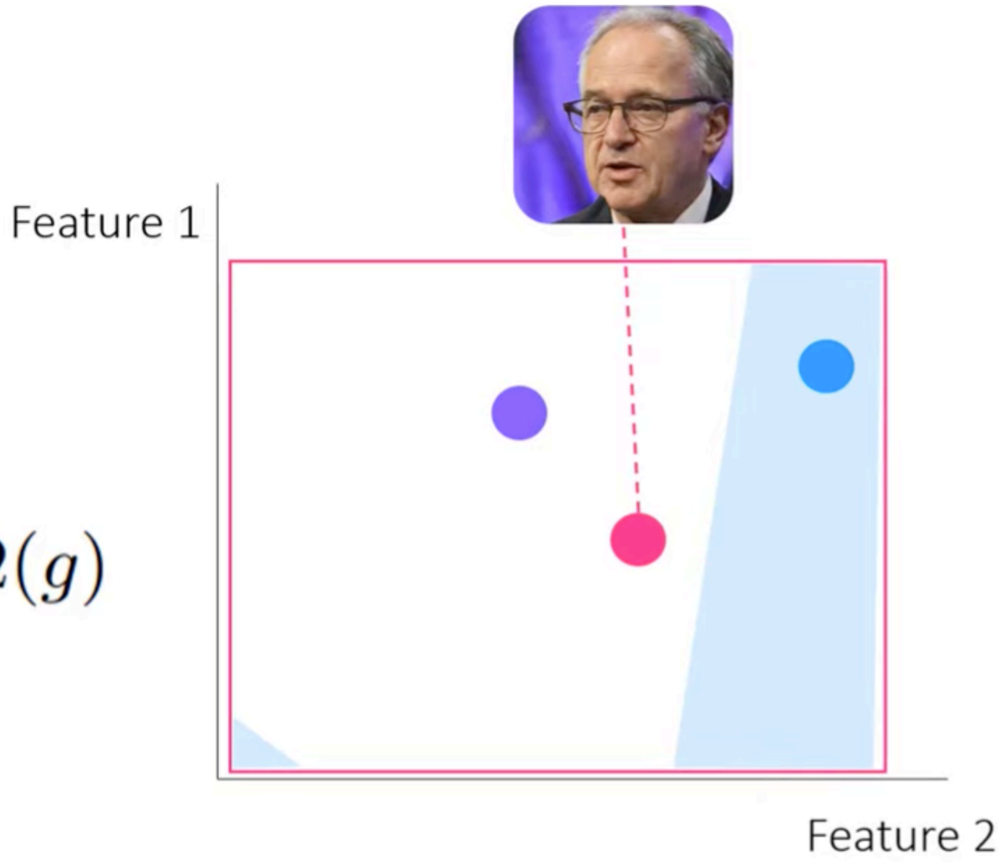
Simple interpretable model

Family of interpretable models

Age | Gender | Heart Disease | ...

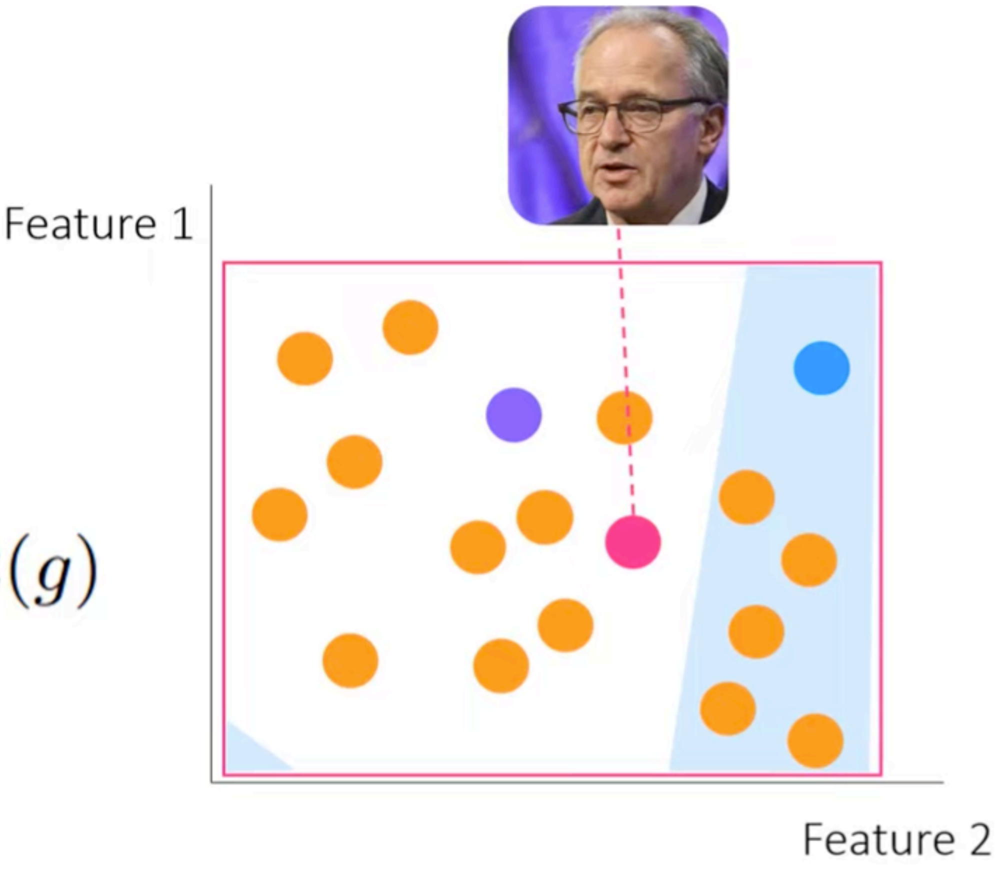
How to train the surrogate

$$\xi(x) = \operatorname*{argmin}_{g \in G} \boxed{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$



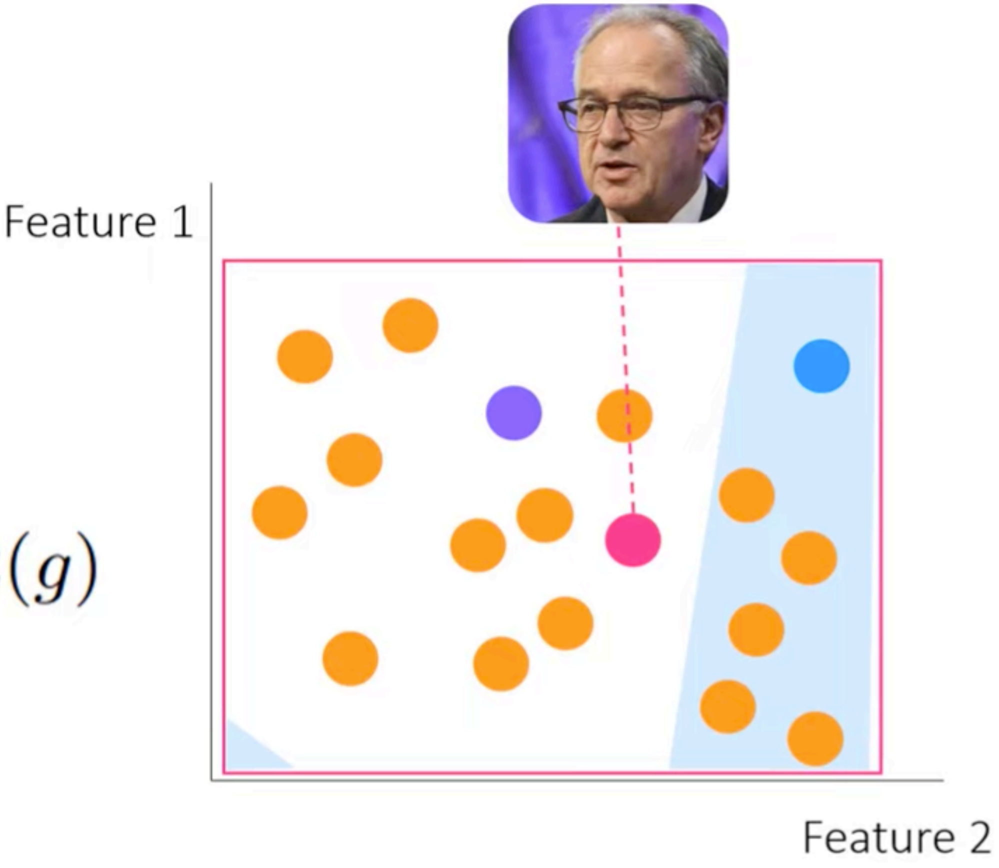
How to train the surrogate

$$\xi(x) = \operatorname*{argmin}_{g \in G} \boxed{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$



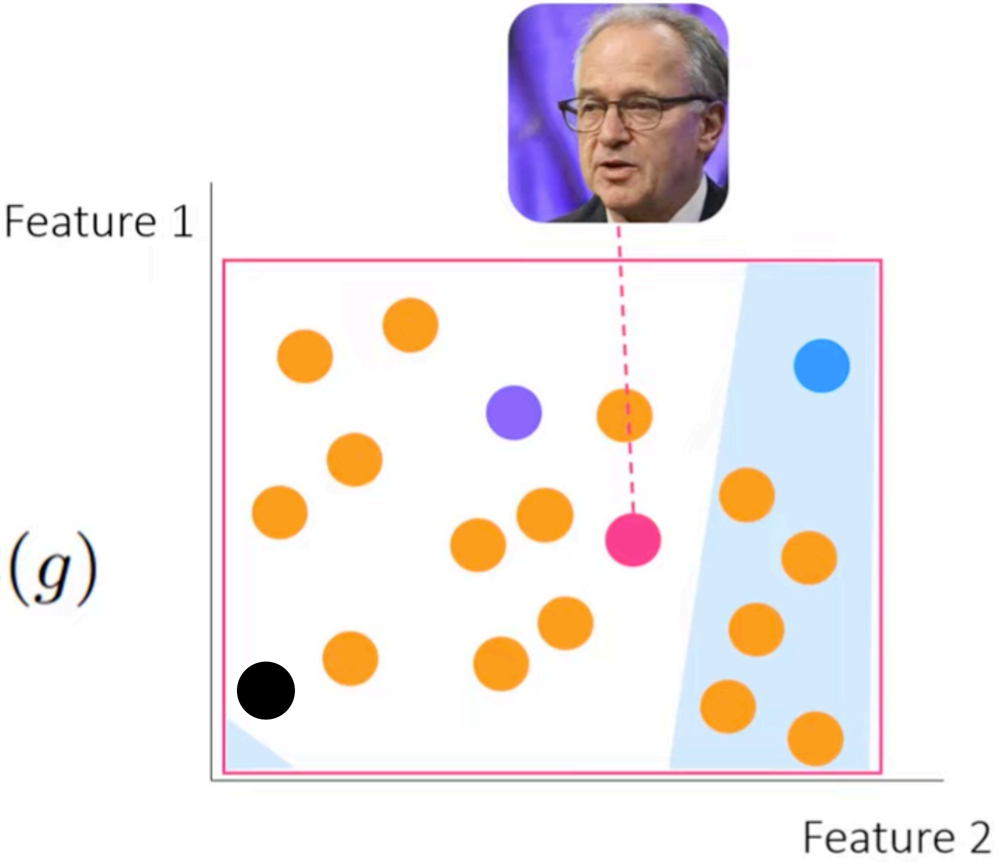
How to train the surrogate

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \boxed{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$



New dataset
Labels: Prediction of complex model
Features: Newly generated datapoints

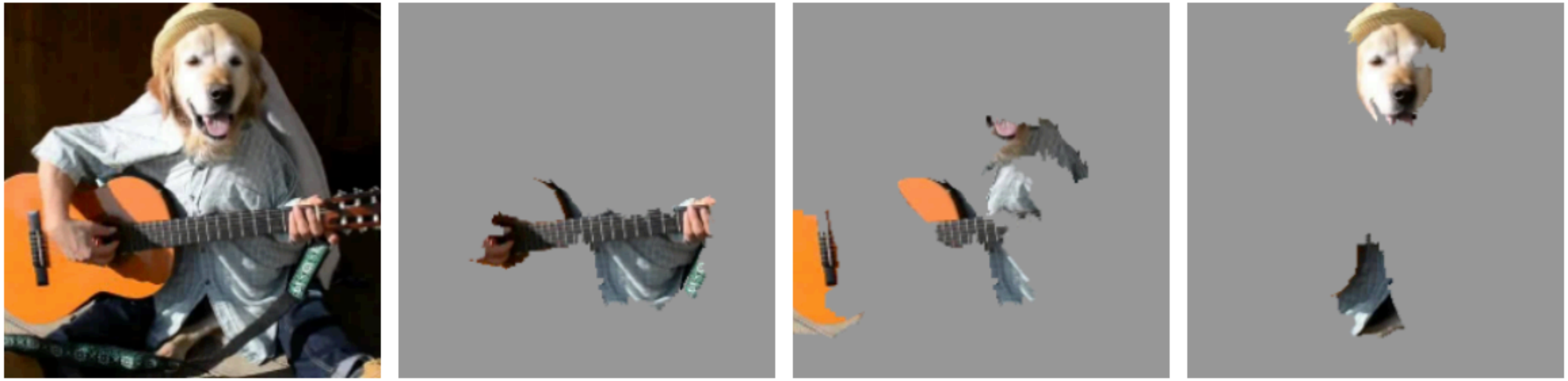
How to train the surrogate



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \boxed{\mathcal{L}(f, g, \pi_x)} + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

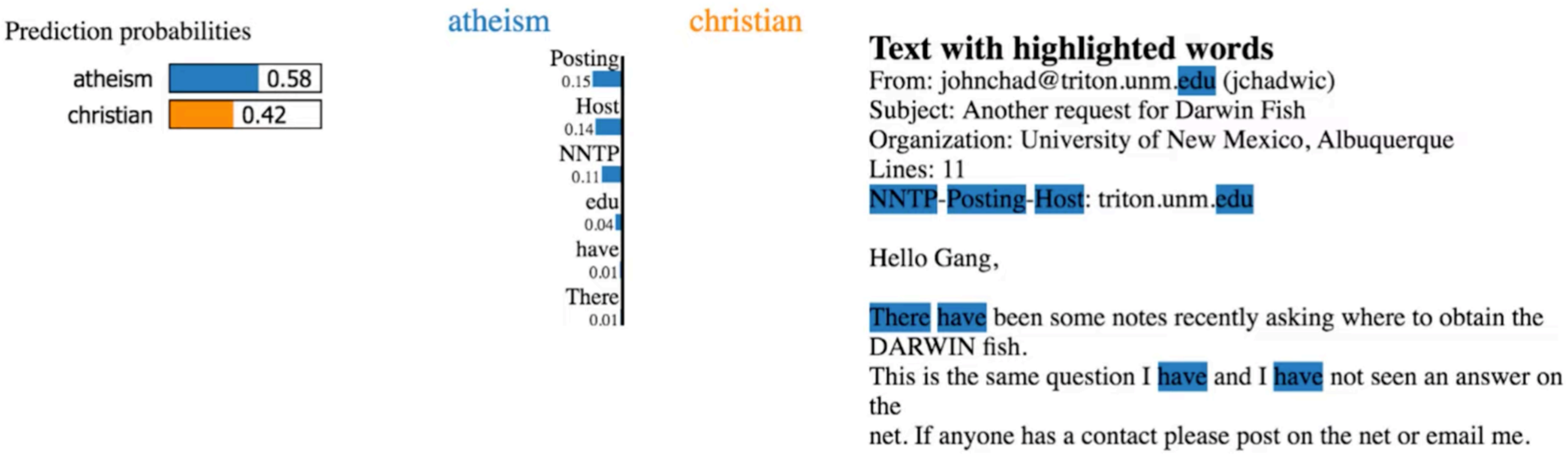
Example for LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

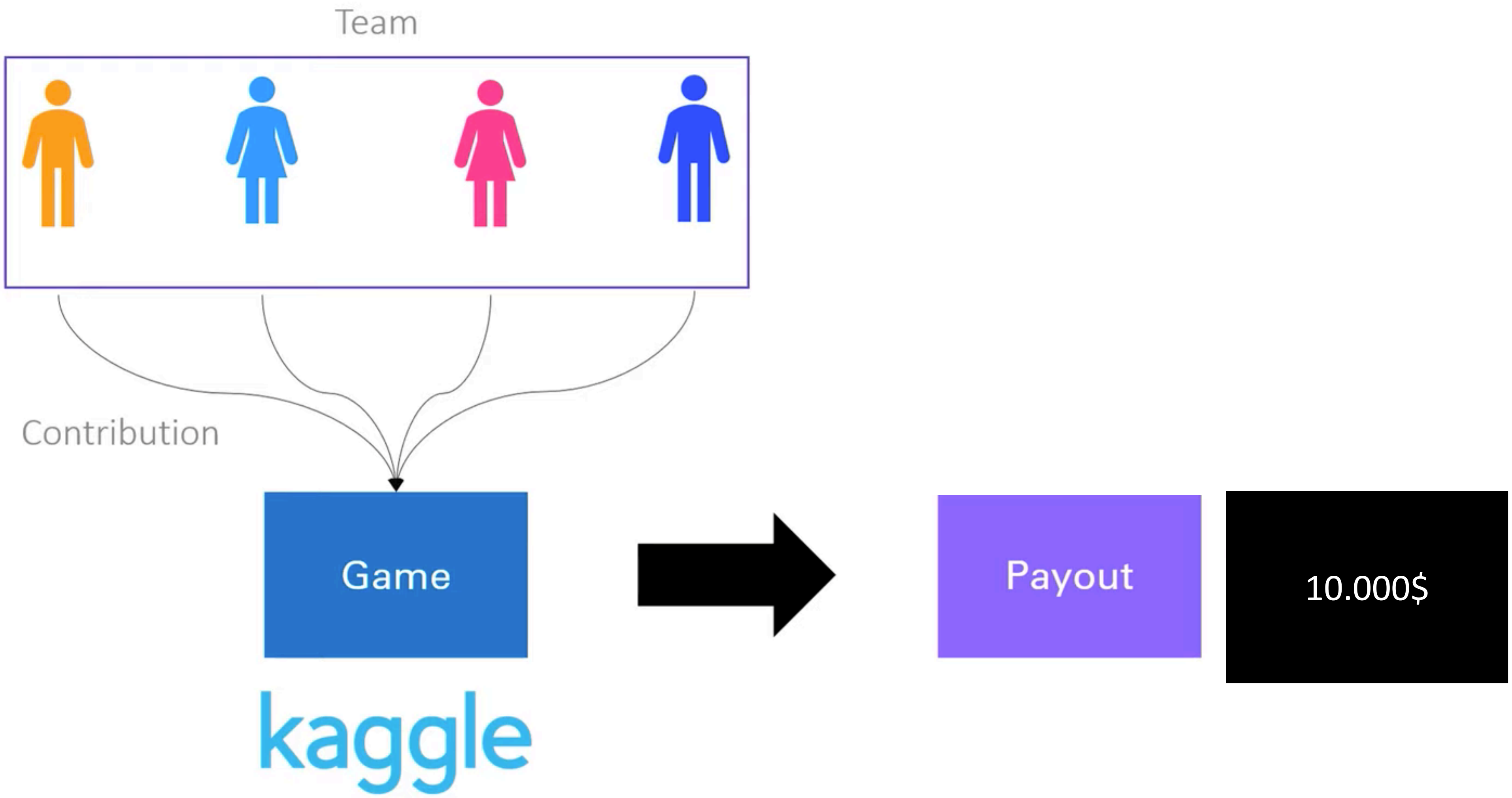
Example for LIME



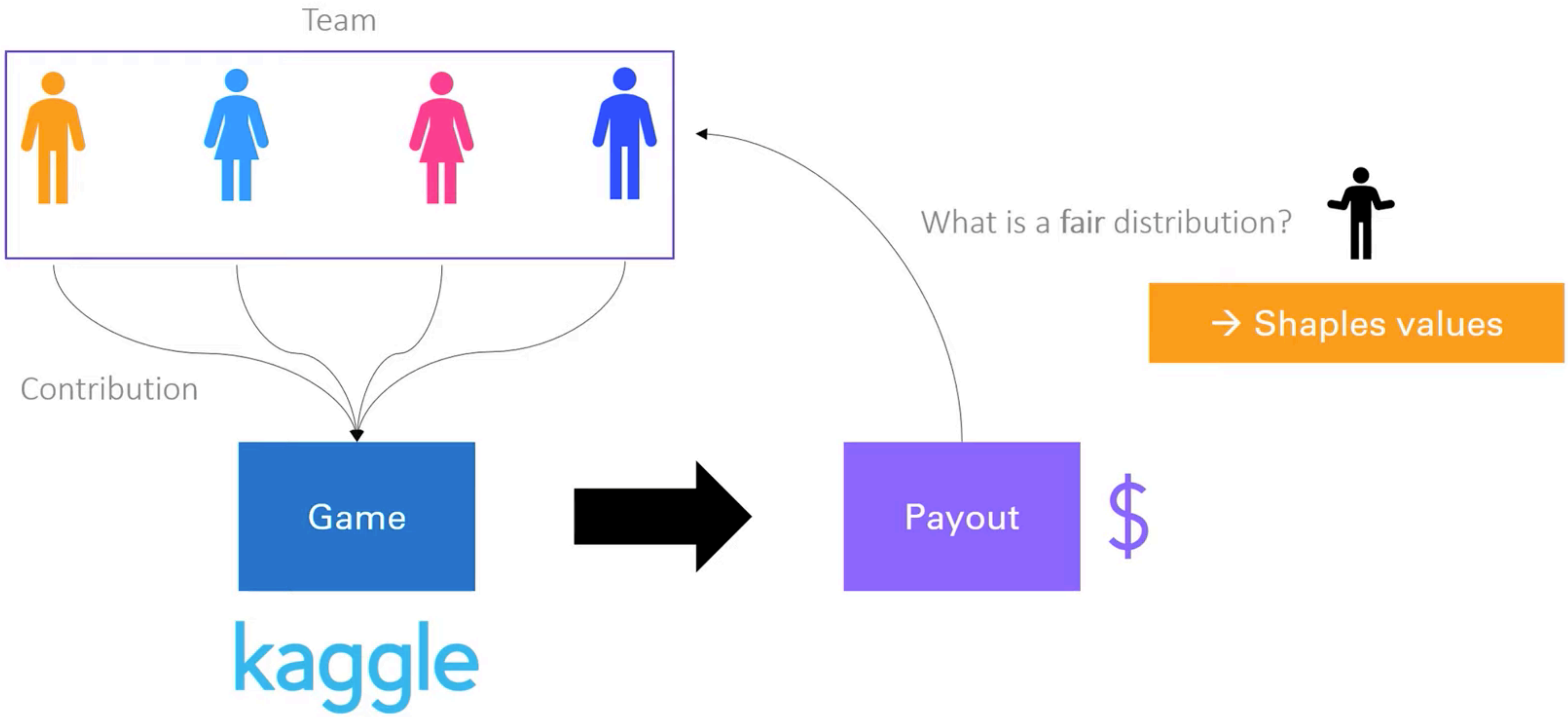
SHAP: SHapley **A**dditive ex**P**lanations

Cooperative Game Theory

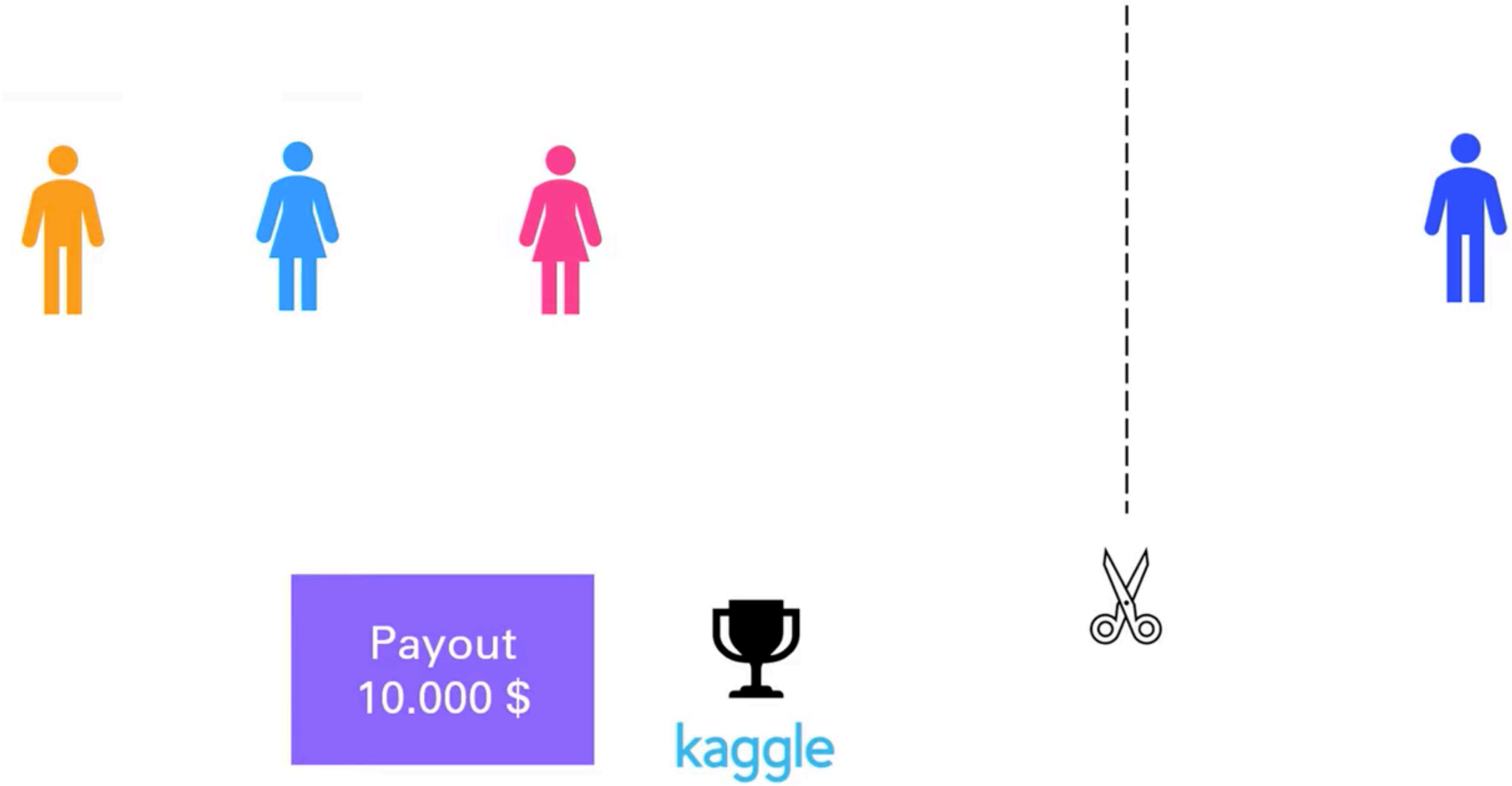
SHAP



SHAP


Team
Contribution
Game
kaggle
Payout
$
What is a fair distribution?
→ Shaples values

Shapley Values



Shapley Values

Domain expert

3.000$

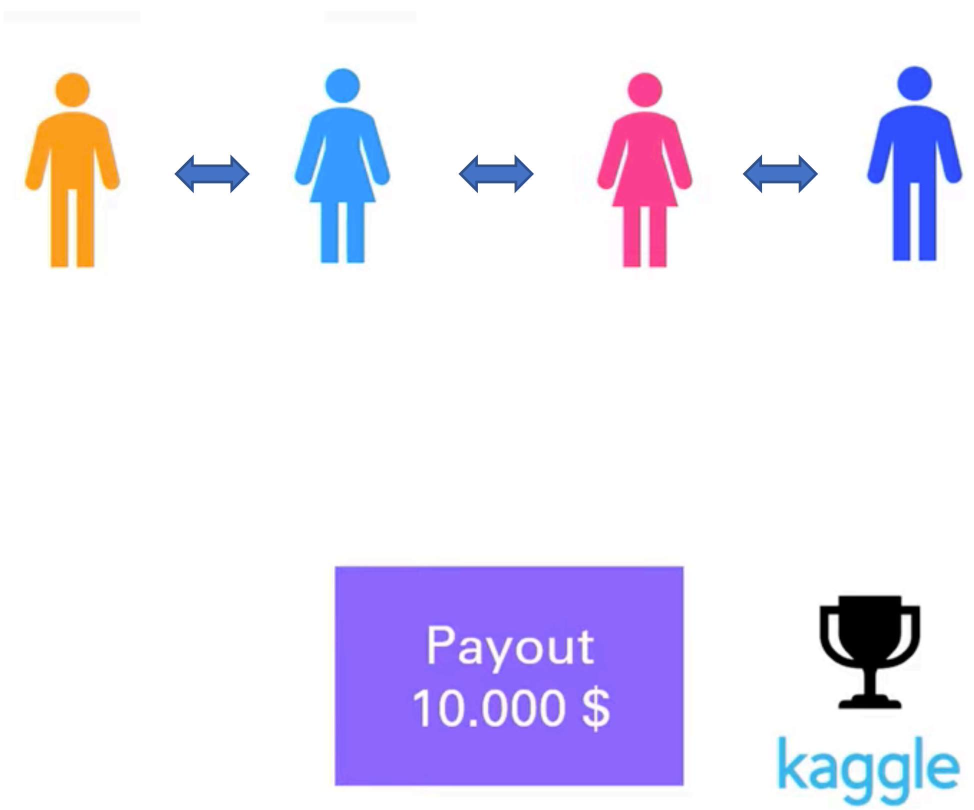
Shapley Values

Domain expert

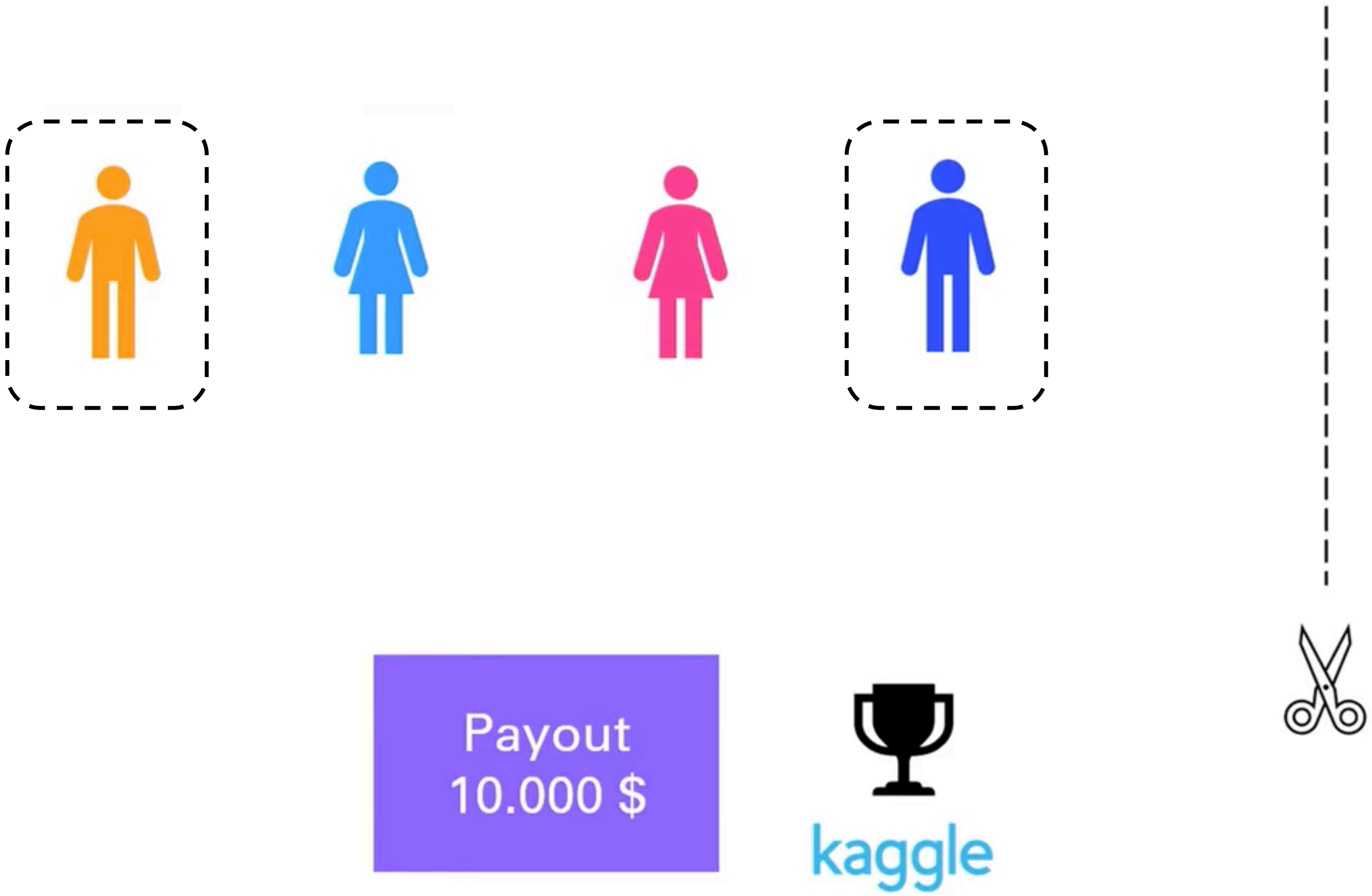
Payout
10.000 $

7.000$

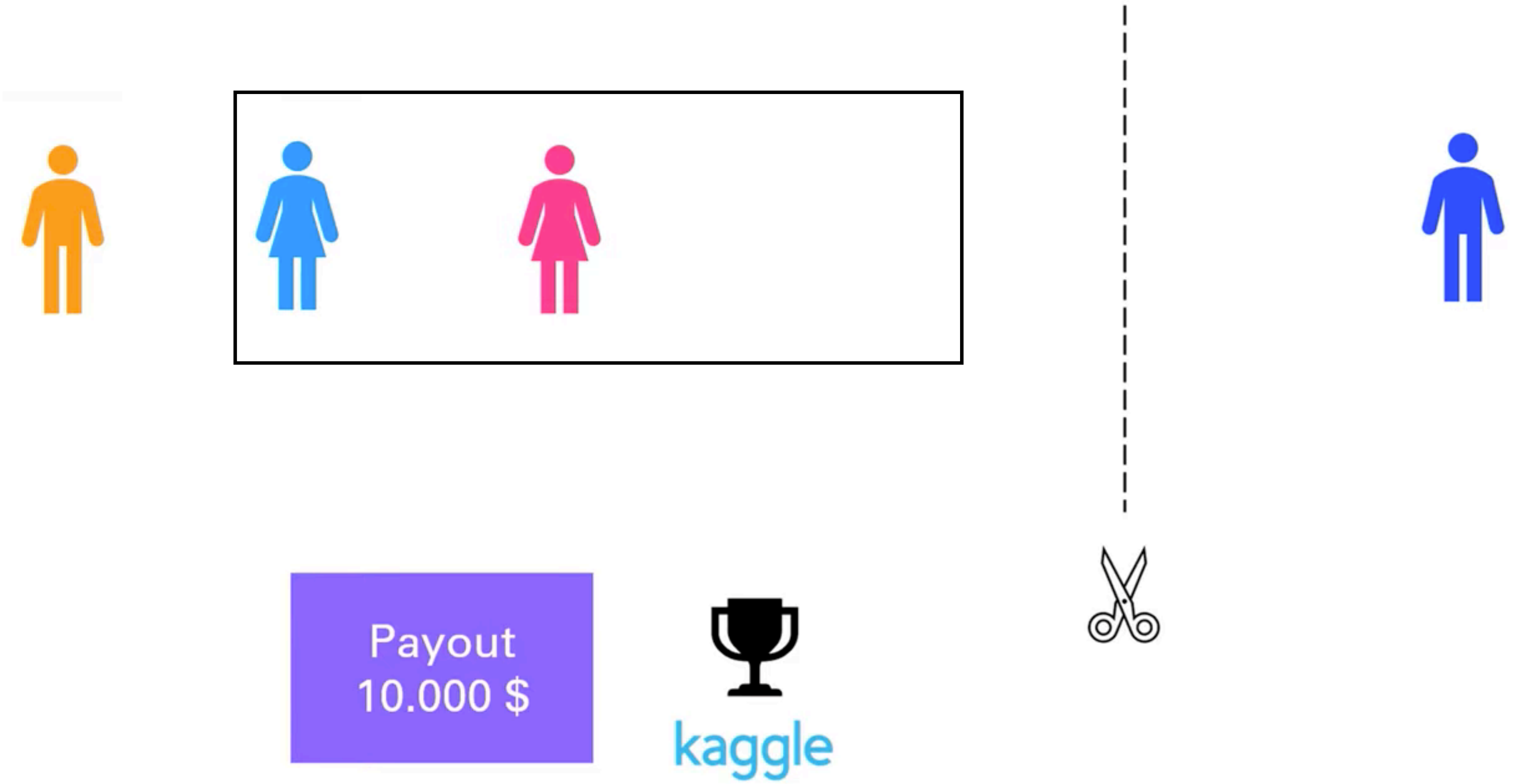
Shapley Values



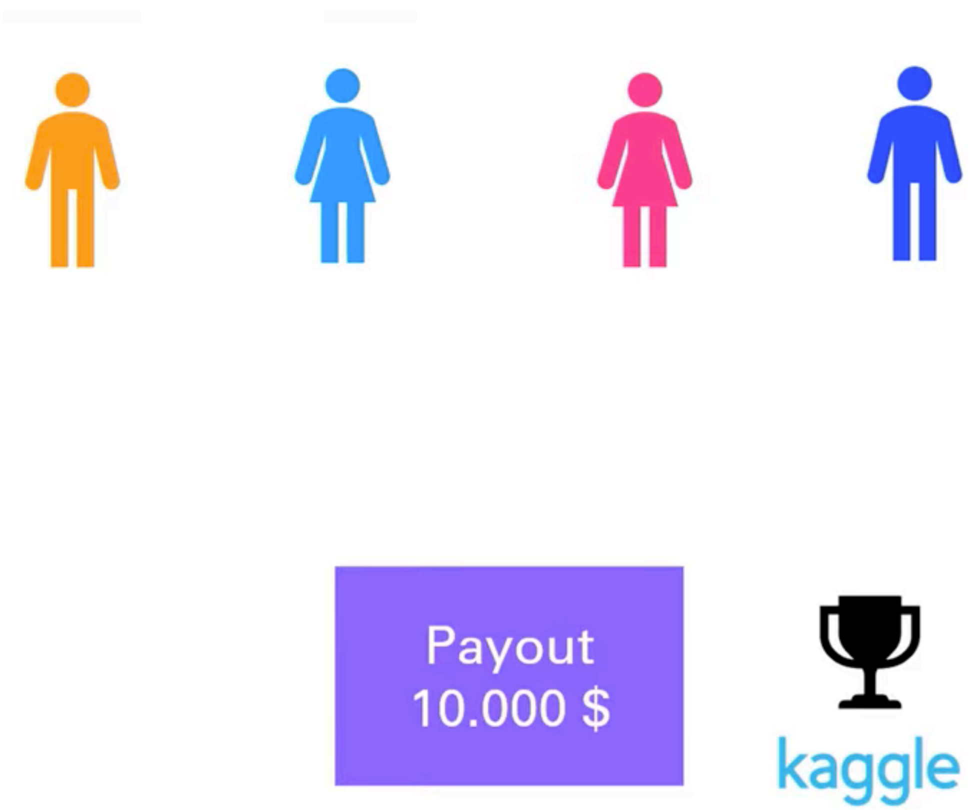
Shapley Values



Shapley Values



Shapley Values



Marginal Contribution

Shapley Values

3.000$	1.500$	2.500$	3.000$

Payout
10.000 $

kaggle

Calculating shapley value

Black model

Input datapoint

Age $\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$

Shapley value
for feature i

x =	Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	...

Calculating shapley value

Black model

Input datapoint

Age

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}\left[f_x(z') - f_x(z' \setminus i)\right]$$

Shapley value for feature i

subset

Simplified data input

x =

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	…

Calculating shapley value

Black model

Input datapoint

Age

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Shapley value for feature i

subset

Simplified data input

x =

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	...

Calculating shapley value

Black model

Input datapoint

Age = 56 | Body Mass Index = 30

Age
$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!}\left[f_x(z') - f_x(z' \setminus i)\right]$$

Body Mass Index = 30

Shapley value
for feature i

subset

Simplified
data input

x =

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	…

Calculating shapley value

Black model

Input datapoint

Age = 56 | Body Mass Index = 30

Age

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

70% Stroke

Body Mass Index = 30

Shapley value
for feature i

subset

Simplified
data input

x =

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	...

Calculating shapley value

Black model

Input datapoint

Age = 56 | Body Mass Index = 30

10% Stroke

Age

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Body Mass Index = 30

Shapley value
for feature i

subset

Simplified
data input

x = | Age = 56 | Gender = F | Body Mass Index = 30 | Hear disease = yes | ...

Calculating shapley value

Black model

Input datapoint

Age = 56 | Body Mass Index = 30

Age

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

Shapley value
for feature i

subset

Simplified
data input

Body Mass Index = 30

x = | Age = 56 | Gender = F | Body Mass Index = 30 | Hear disease = yes | … |

Calculating shapley value

Black model

Input datapoint

Age $\phi_i(f, x) = \sum_{z' \subseteq x'} \underbrace{\frac{|z'|!(M - |z'| - 1)!}{M!}}_{\text{Weighting}} \underbrace{[f_x(z') - f_x(z' \setminus i)]}_{\text{Contribution}}$

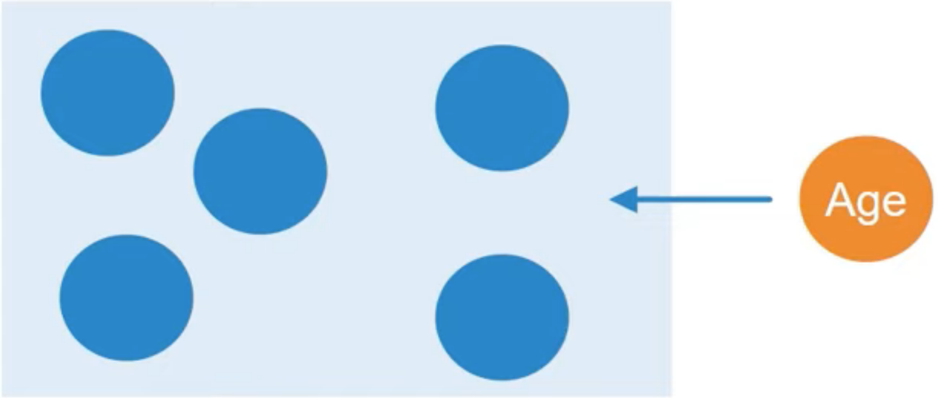
Shapley value
for feature i

subset

Simplified
data input

x = | Age = 56 | Gender = F | Body Mass Index = 30 | Hear disease = yes | ... |

Calculating shapley value



Black model

Input datapoint

Age

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \underbrace{\frac{|z'|!(M-|z'|-1)!}{M!}}_{\text{Weighting}} \underbrace{[f_x(z') - f_x(z' \setminus i)]}_{\text{Contribution}}$$

Shapley value for feature i

subset

Simplified data input

x = | Age = 56 | Gender = F | Body Mass Index = 30 | Hear disease = yes | … |

Calculating shapley value

Black model

Input datapoint

Age $\phi_i(f,x) = \sum_{z' \subseteq x'} \underbrace{\frac{|z'|!(M - |z'| - 1)!}{M!}}_{\text{Weighting}} \underbrace{[f_x(z') - f_x(z' \setminus i)]}_{\text{Contribution}}$

Shapley value
for feature i

subset

Simplified
data input

x =

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	...

Calculating shapley value

Black model

Input datapoint

Age = 56 | Body Mass Index = 30

Age

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right]$$

Shapley value
for feature i

subset

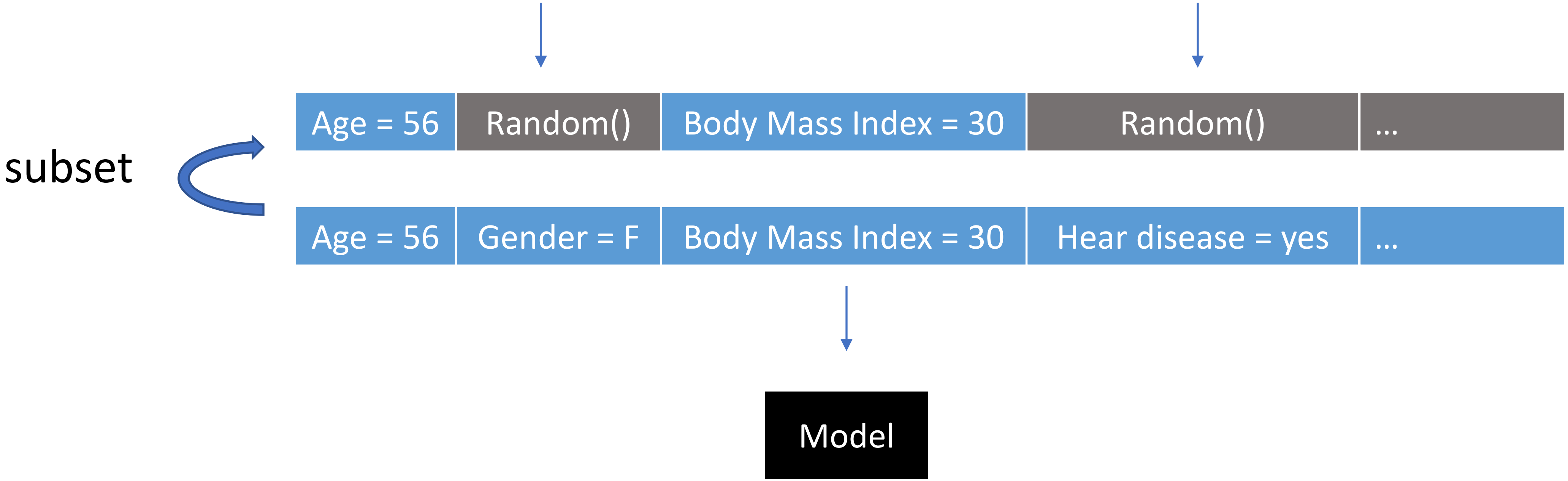
Simplified
data input

Body Mass Index = 30

Age = 56	Gender = F	Body Mass Index = 30	Hear disease = yes	...

x =

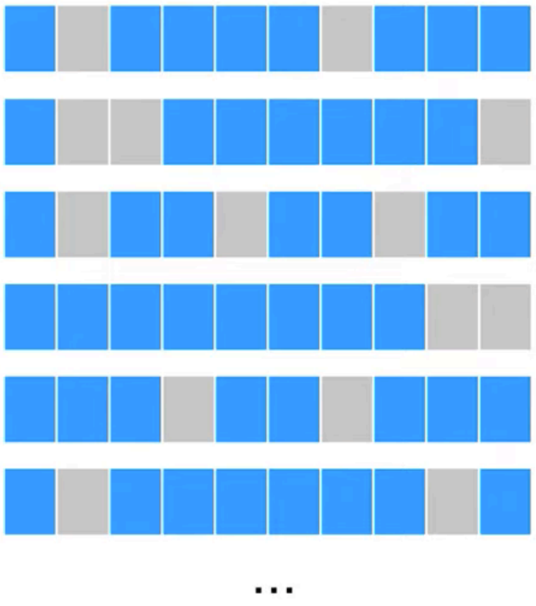
Calculating shapley value



Calculating shapley value

2^n = total number of subsets of a set

Calculating shapley value

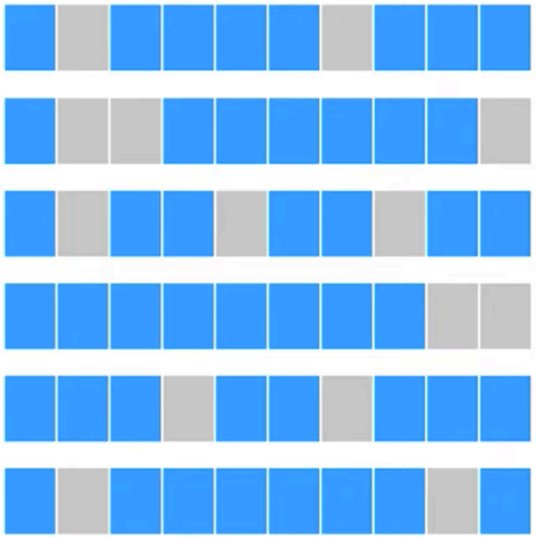


$$2^{10} = 1024$$

2^n = total number of subsets of a set

Calculating shapley value

Kernel SHAP



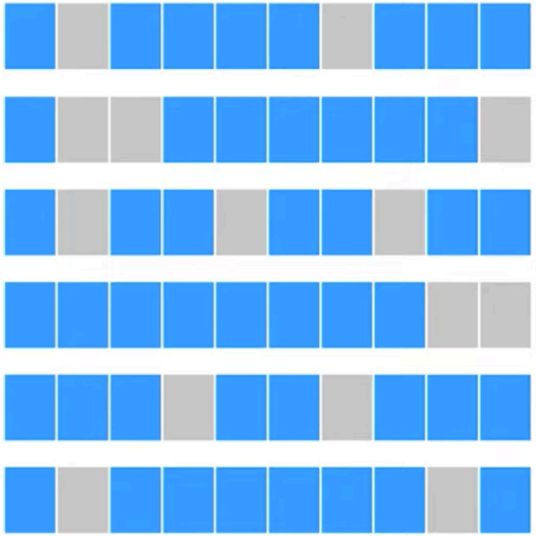
...

$$2^{10} = 1024$$

2^n = total number of subsets of a set

$$Y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 \cdots$$

Calculating shapley value



...

$2^{10} = 1024$

2^n = total number of subsets of a set

Kernel SHAP

Tree SHAP

Deep SHAP