CSCI 4360/6360: Data Science II

Shapley Values & SHAP

Ninghao Liu
Assistant Professor
School of Computing
University of Georgia

Many contents are adopted from:
“**Interpretable Machine Learning - A Guide for Making Black Box Models Explainable**” by Christoph Molnar
“Interpretable Machine Learning” by Hanjie Chen, Yangfeng Ji

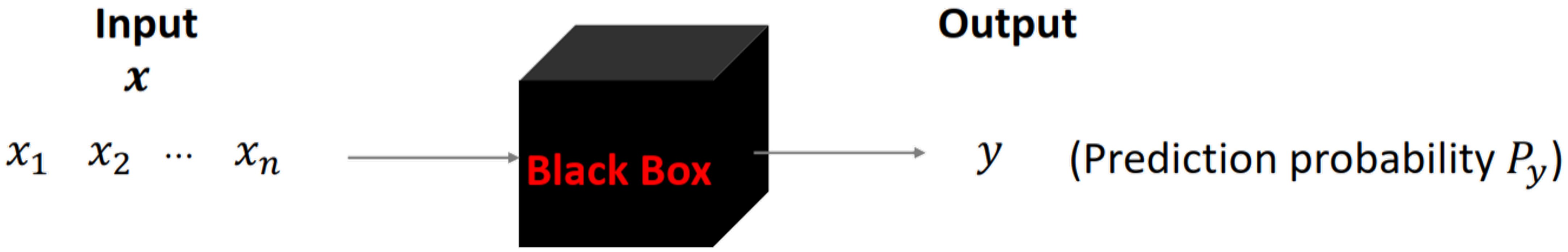
Outline

- Leave-One-Out
- Shapley Value
- SHAP

Outline

- **Leave-One-Out**
- Shapley Value
- SHAP

Leave-One-Out



Input

$\boldsymbol{x}$

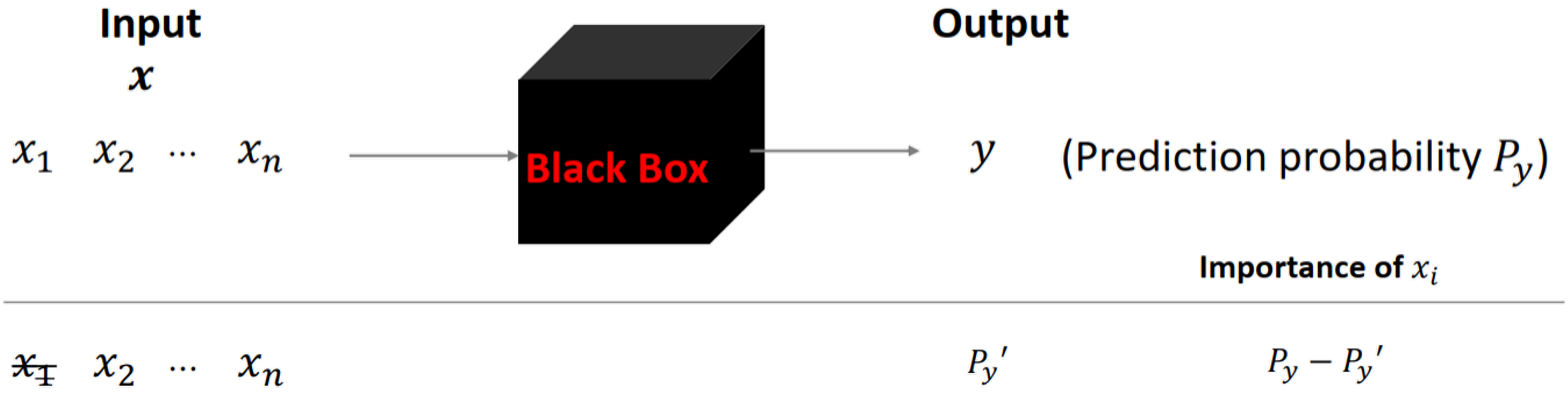
$x_1 \quad x_2 \quad \cdots \quad x_n$

Black Box

Output

y (Prediction probability P_y)

Leave-One-Out



Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$

Black Box

Output

y (Prediction probability P_y)

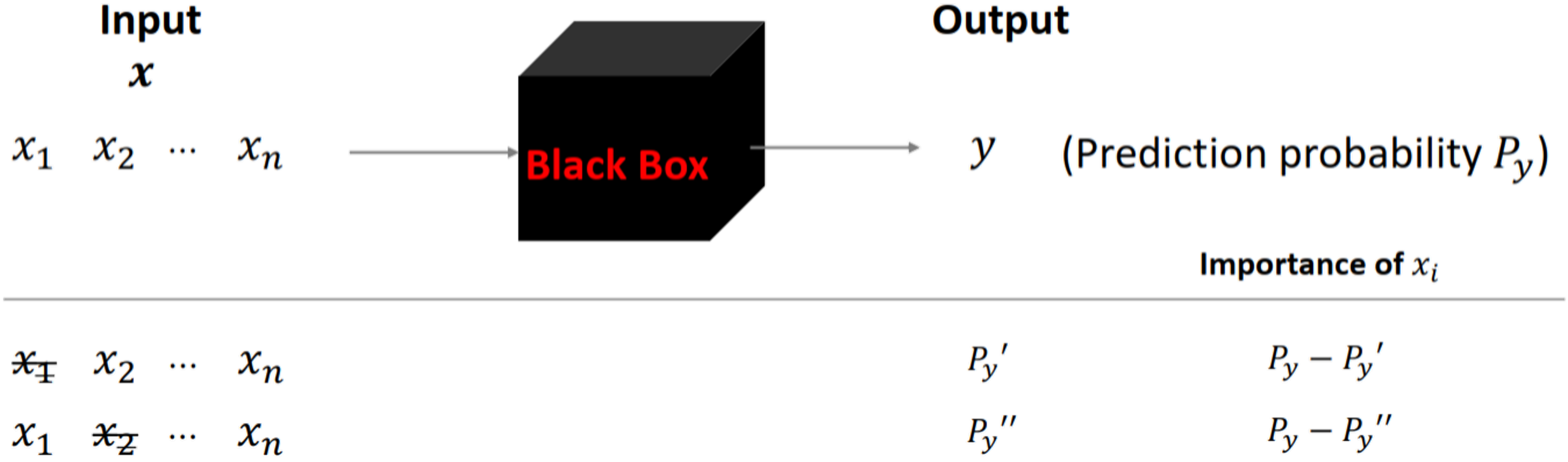
Importance of x_i

$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$

$P_y{}'$

$P_y - P_y{}'$

Leave-One-Out



Input

$\boldsymbol{x}$

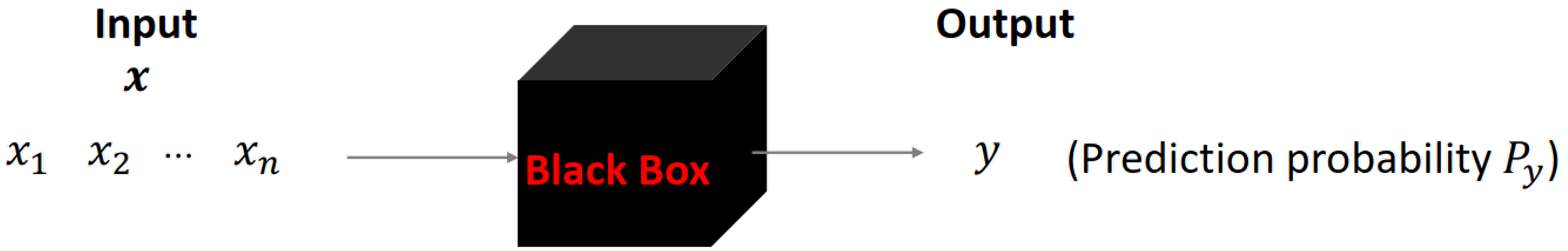
$x_1 \quad x_2 \quad \cdots \quad x_n$

Output

y (Prediction probability P_y)

		Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$	$P_y{}'$	$P_y - P_y{}'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$	$P_y{}''$	$P_y - P_y{}''$

Leave-One-Out



Input

$\boldsymbol{x}$

$x_1 \quad x_2 \quad \cdots \quad x_n$

Black Box

Output

y (Prediction probability P_y)

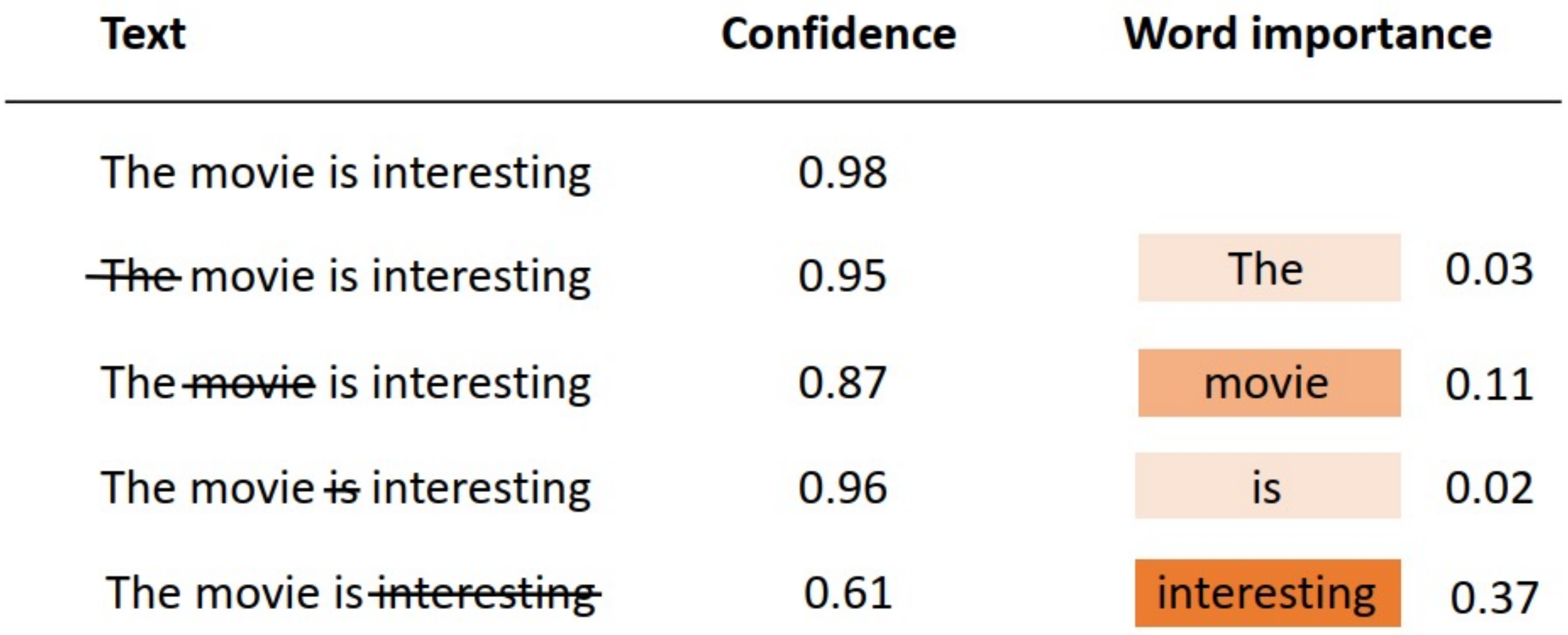
		Importance of x_i
$\cancel{x_1} \quad x_2 \quad \cdots \quad x_n$	P_y'	$P_y - P_y'$
$x_1 \quad \cancel{x_2} \quad \cdots \quad x_n$	P_y''	$P_y - P_y''$
$\vdots$	$\vdots$	$\vdots$

[Leave-one-out, (Li et al., 2016)]

Leave-One-Out

- Sentiment classification

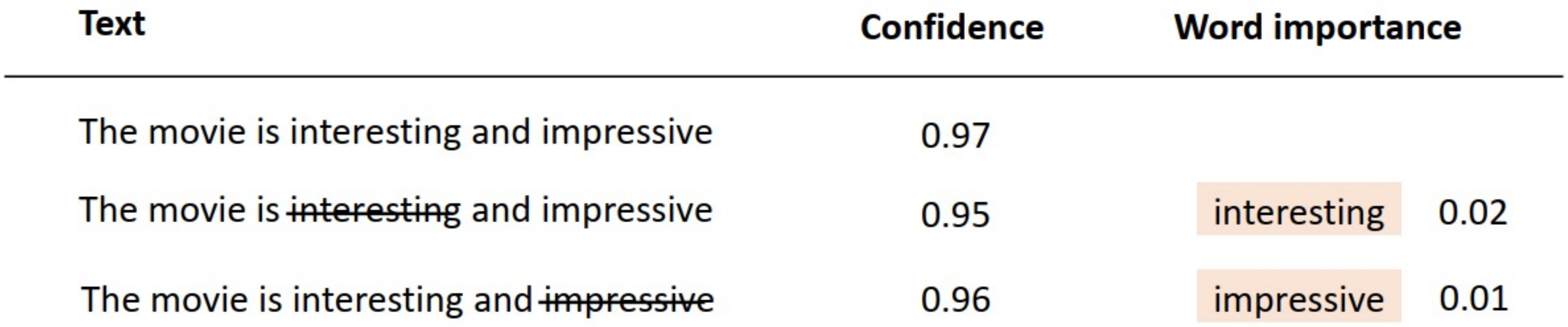
Model prediction: positive



Text	Confidence	Word importance	
The movie is interesting	0.98		
~~The~~ movie is interesting	0.95	The	0.03
The ~~movie~~ is interesting	0.87	movie	0.11
The movie ~~is~~ interesting	0.96	is	0.02
The movie is ~~interesting~~	0.61	interesting	0.37

Leave-One-Out

Feature importance may be misleading



Text	Confidence	Word importance	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Leave-One-Out

Feature importance may be misleading

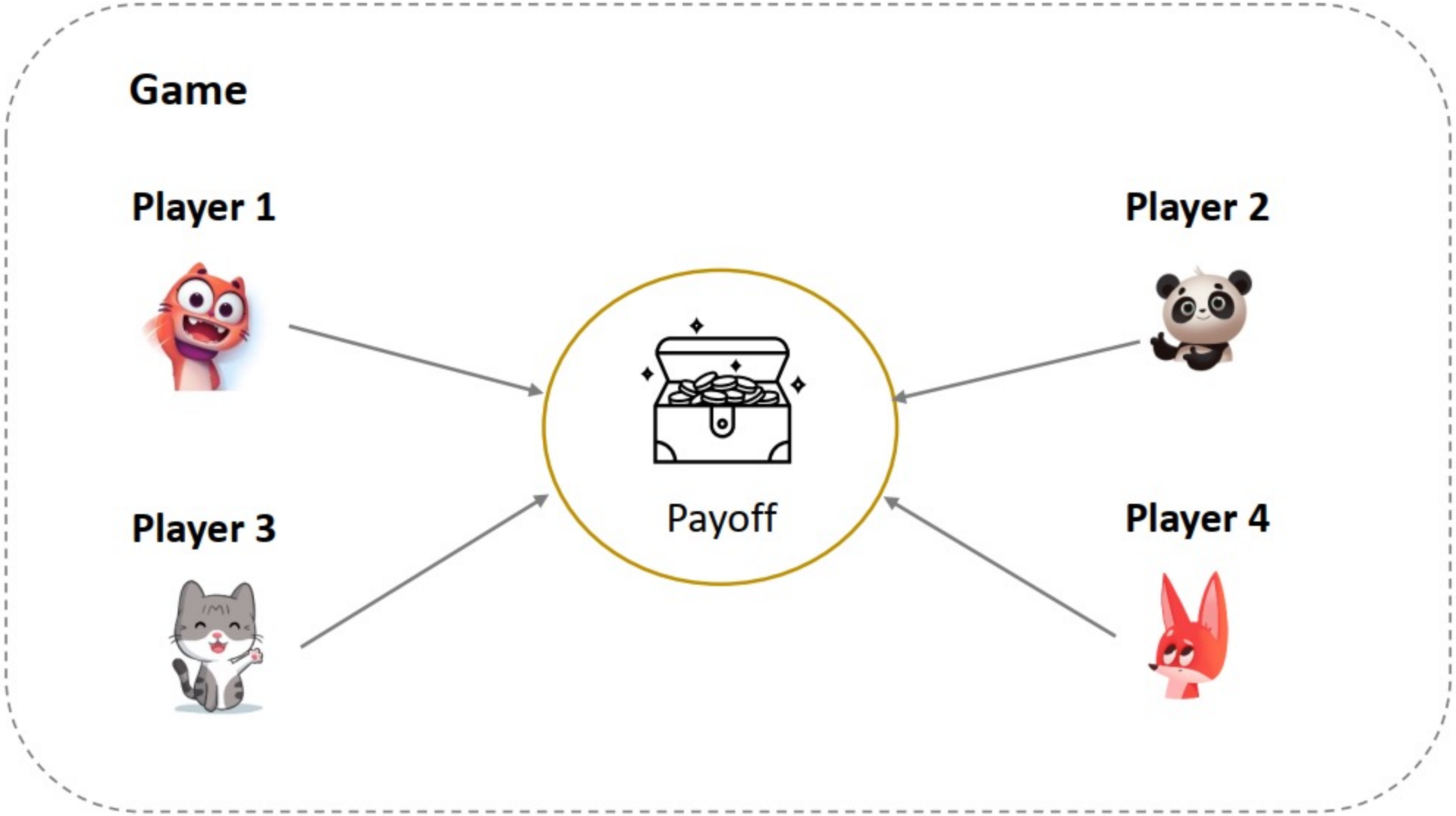
Text	Confidence	Word importance	
The movie is interesting and impressive	0.97		
The movie is ~~interesting~~ and impressive	0.95	interesting	0.02
The movie is interesting and ~~impressive~~	0.96	impressive	0.01

Need a better way to quantify feature importance

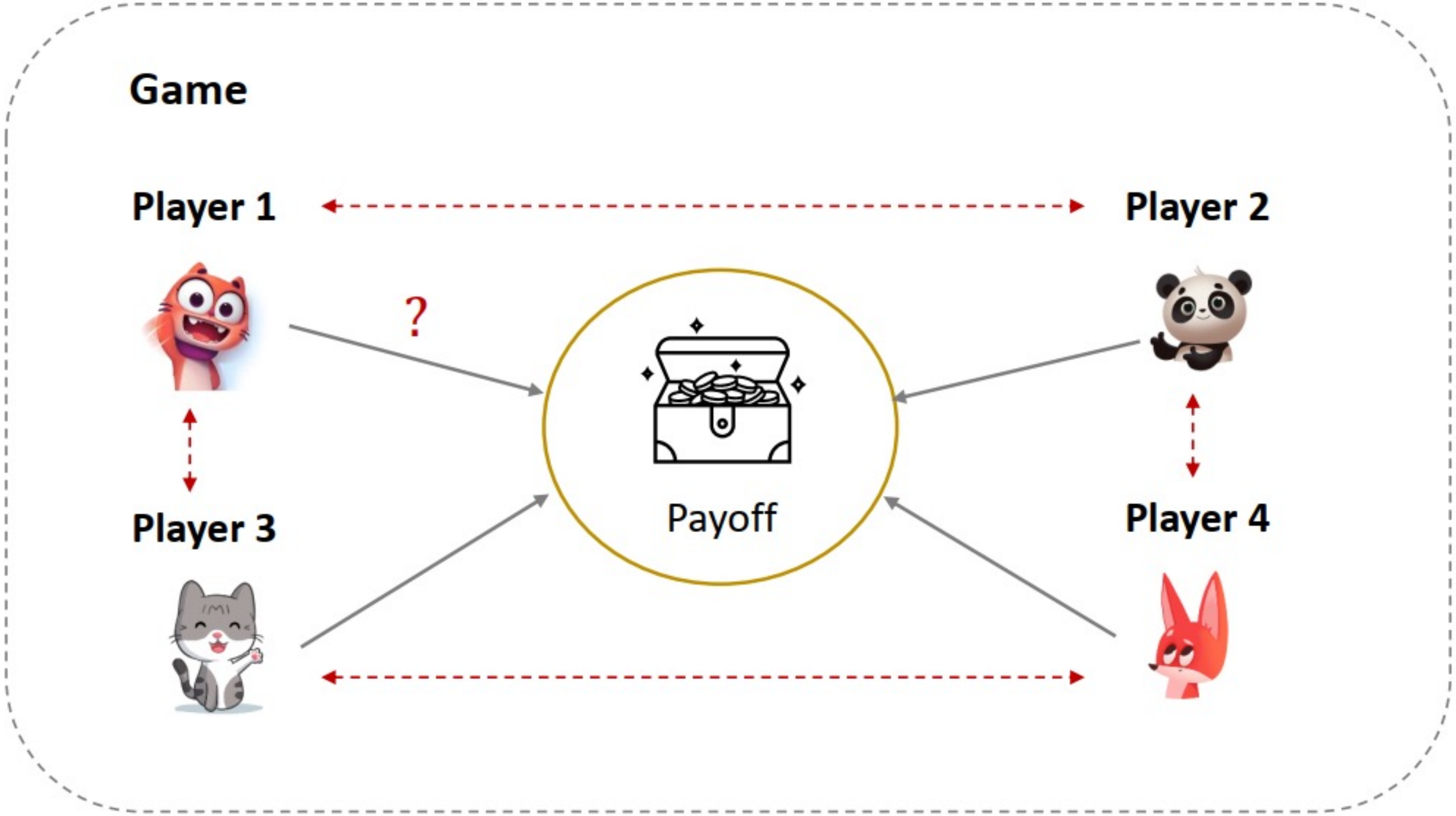
Outline

- Leave-One-Out
- **Shapley Value**
- SHAP

Shapley Value



Shapley Value

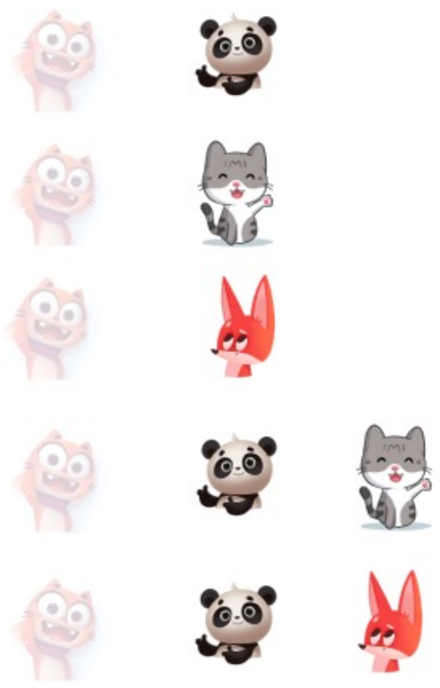


Shapley Value

Coalitions	**Payoff**
	P_1
	P_2
	P_3
	P_4
	P_5
⋮	⋮
(2^3)	

Shapley Value

Coalitions



⋮

(2^3)

Payoff

$P_1 - P_1'$

$P_2 - P_2'$

$P_3 - P_3'$

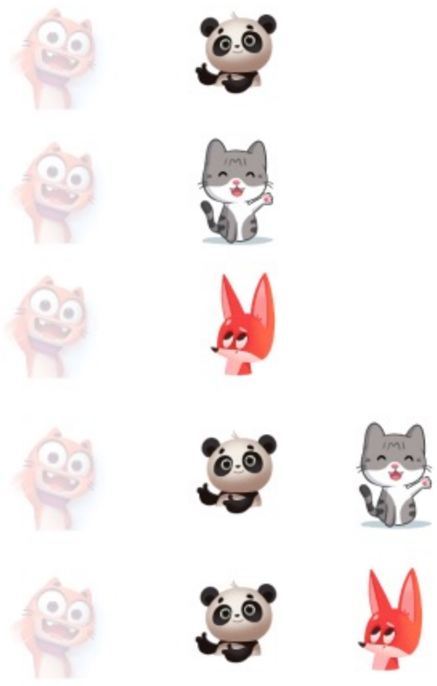
$P_4 - P_4'$

$P_5 - P_5'$

⋮

Shapley Value

Coalitions



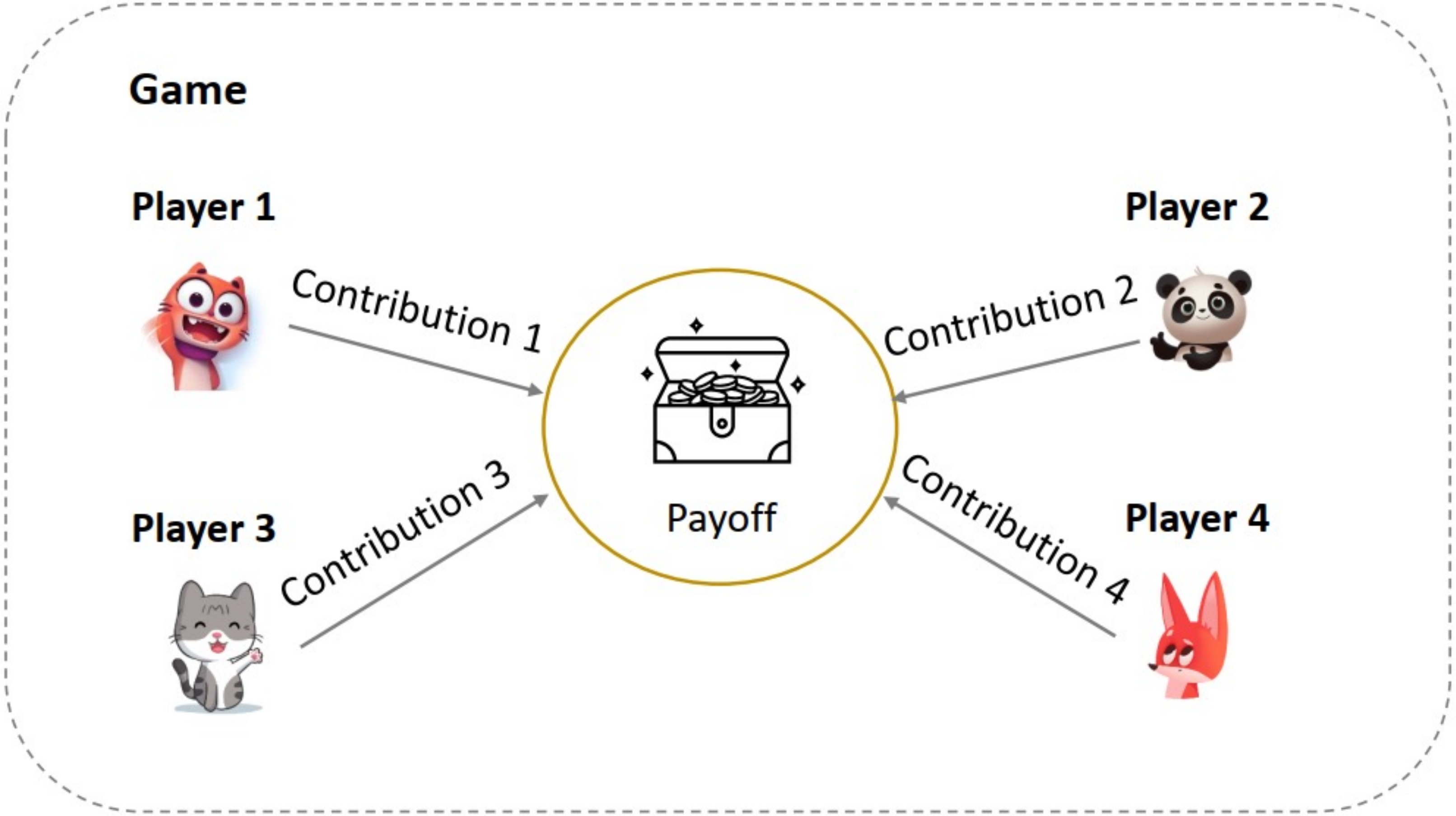
⋮

(2^3)

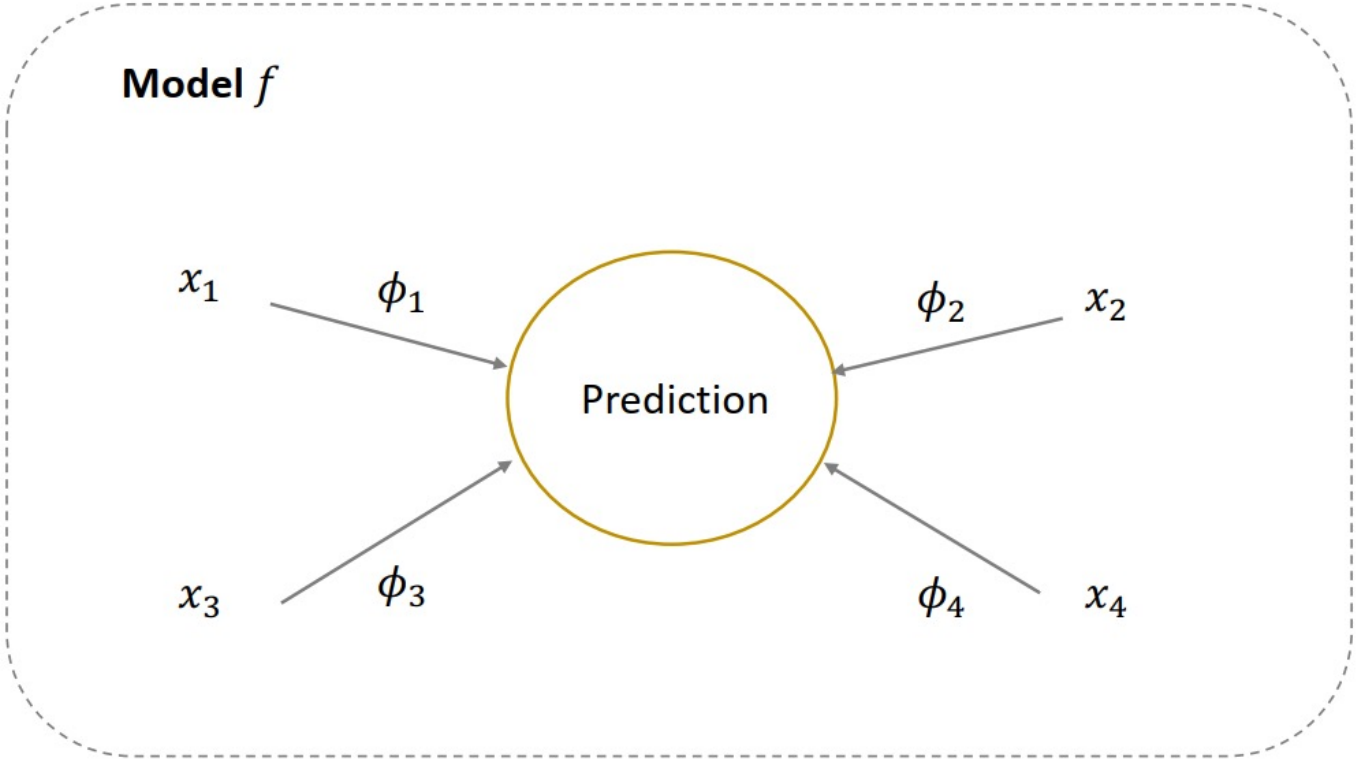
Payoff	**Marginal contribution**
$P_1 - P_1'$	ΔP_1
$P_2 - P_2'$	ΔP_2
$P_3 - P_3'$	ΔP_3
$P_4 - P_4'$	ΔP_4
$P_5 - P_5'$	ΔP_5
⋮	

Contribution$= \sum \Delta P_i$

Shapley Value



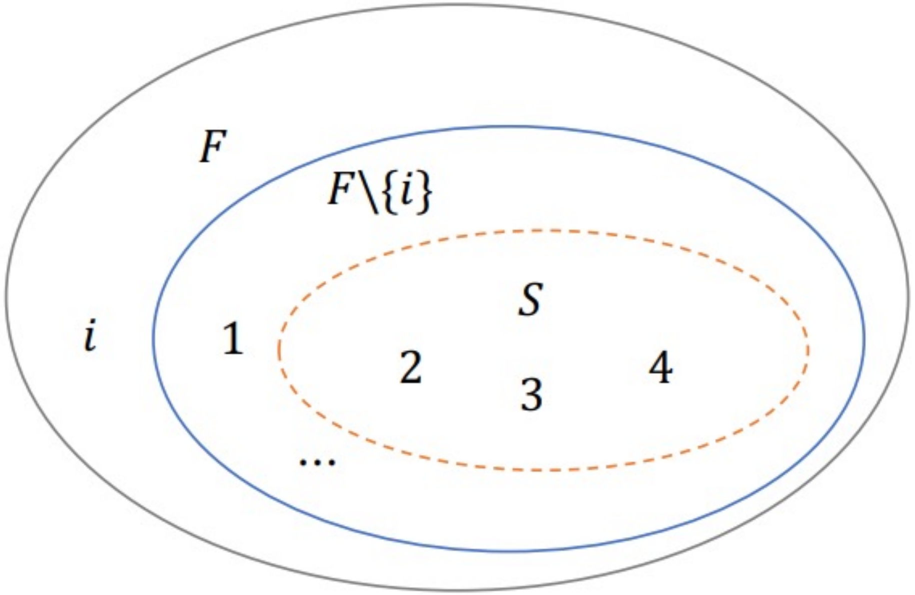
Shapley Value



$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

Shapley Value

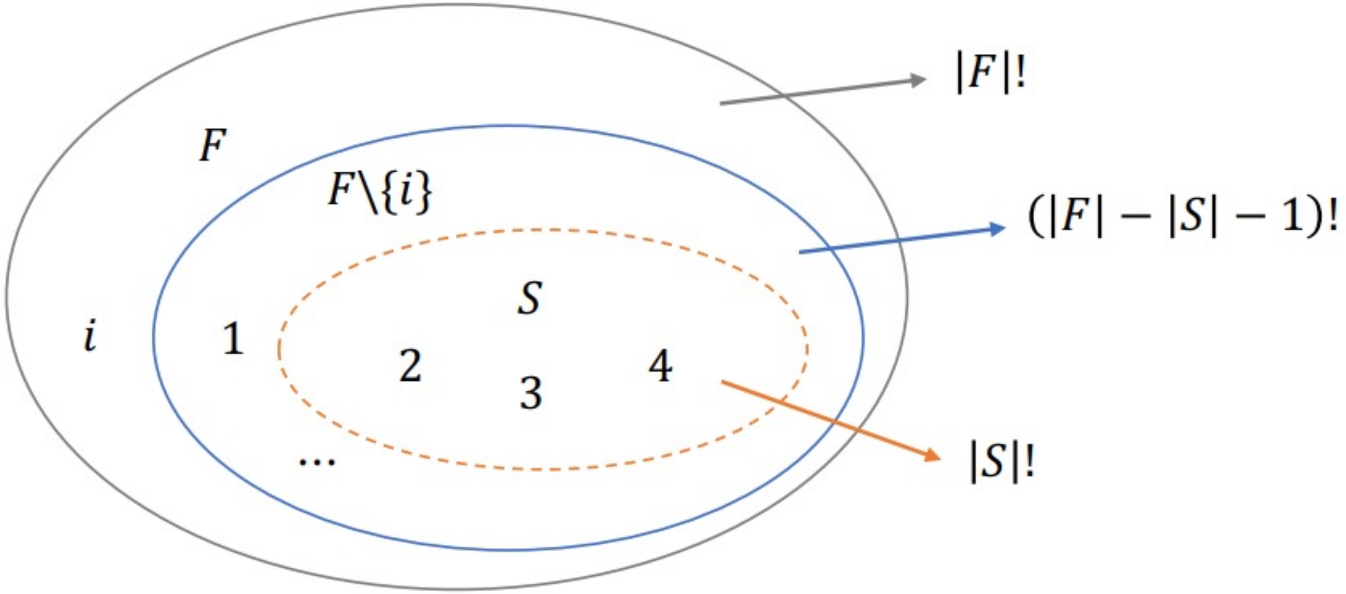
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \underbrace{\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right]}_{\text{Marginal contribution of } x_i \text{ given } S}$$



Shapley Value

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \underbrace{\frac{|S|!\,(|F| - |S| - 1)!}{|F|!}} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]$$

Weighted by the permutations of features



SHAP

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Challenge

Computational complexity

$O(2^n)$