

# Research Discussion

## — SHAP —

Prof. Dr. Mathias Kraus  
Chair for Explainable AI in Business Value Creation  
Faculty for Informatics & Data Science







# Overview

- 1) Development of Shapley values in game theory.
- 2) Transition from Shapley values to SHAP values.
- 3) Implementation of exact computation of SHAP values.

Many slides are derived from “Interpreting Machine Learning Models With SHAP” from Molnar (2023).

# History of Game Theory

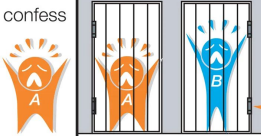

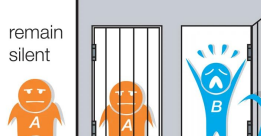

- In the 1950s, game theory saw an active period, during which many core concepts were formulated, including repeated games, the **prisoner's dilemma**, fictitious play, and Shapley values.

Prisoners' dilemma		prisoner B			
		confess		remain silent	
prisoner A	confess	 5 years 5 years	 0 year 20 years		
	remain silent	 20 years 0 year	 1 year 1 year		

# Cooperative Game Theory

**Definition (Cooperative Game Theory):** Cooperative game theory is a branch of game theory that studies how groups of agents (called coalitions) can collaborate to achieve collective outcomes and how the resulting rewards or costs should be distributed among them.

Prisoners' dilemma

		prisoner B	
		confess	remain silent
prisoner A	confess	 5 years   5 years	 0 year   20 years
	remain silent	 20 years   0 year	 1 year   1 year

prisoner A: “Hey, prisoner B, I guarantee that I remain silent if you remain silent”

prisoner B: “Sure, let’s do that!”

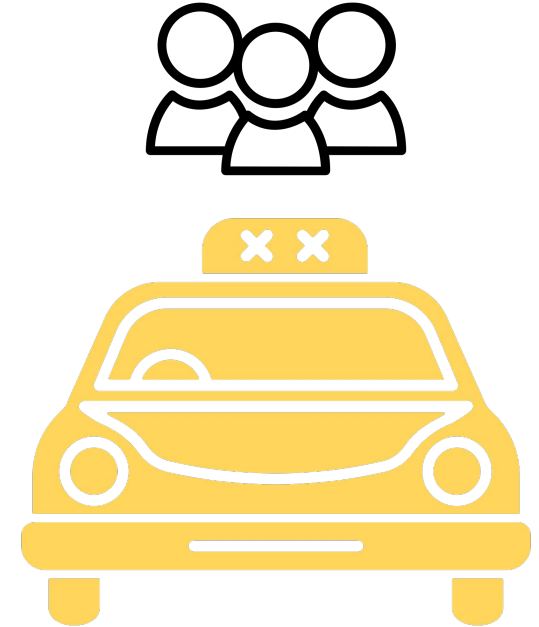
→ (1 year, 1 year)

# History of Shapley Values

- **Alice, Bob, and Charlie** have dinner together and **share a taxi ride home**. The **total cost is \$51**. The question is:

*How should they divide the costs fairly?*

- In game theory, this is also called a coalitional (or cooperative) game in which Alice, Bob, and Charlie receive a specific payout (in this case a negative payout).
- To determine a fair distribution of the costs, we first pose simpler questions:
  - How much would the ride cost for a random coalition of passengers?
  - For instance, how much would Alice pay for a taxi ride if she were alone? How much would Alice and Bob pay if they shared a taxi?



# Who's Going to Pay for That Taxi?

- Let's suppose it would be \$15 for Alice alone.
- Alice and Bob live together, but adding Bob to the ride increases the cost to \$25, as he insists on a more spacious, luxurious taxi.
- Adding Charlie to Alice and Bob's ride increases the cost to \$51 since Charlie lives somewhat further away.

Passengers	Cost	Note
$\emptyset$	\$0	No taxi ride, no costs
{Alice}	\$15	Standard fare to Alice's & Bob's place
{Bob}	\$25	Bob always insists on luxury taxis
{Charlie}	\$38	Charlie lives slightly further away
{Alice, Bob}	\$25	Bob always gets his way
{Alice, Charlie}	\$41	Alice & Bob's place requires a slight detour
{Bob, Charlie}	\$51	Bob is a creature of luxury and Charlie lives a bit further away
{Alice, Bob, Charlie}	\$51	The full fare with all three of them

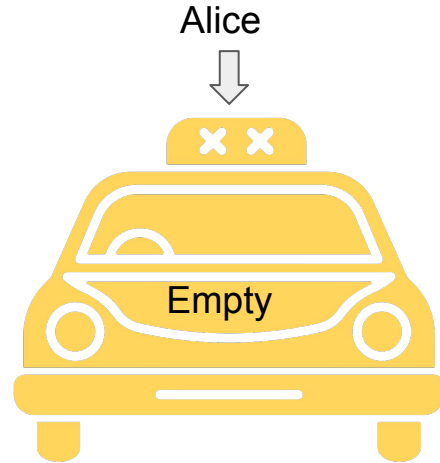
# Marginal Contributions

**Definition (Marginal Contribution):** The marginal contribution of a player to a coalition is the value of the coalition with the player minus the value of the coalition without the player.

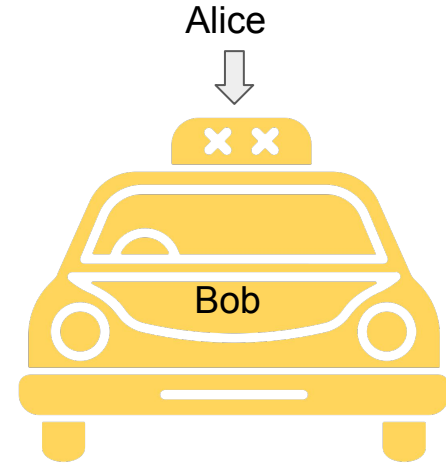
The marginal contribution of, for instance, Charlie to a taxi already containing Bob is the cost of the taxi with Bob and Charlie, minus the cost of the taxi with Bob alone.

Addition	To Coalition	Cost Before	Cost After	Marginal Contribution
Alice	$\emptyset$	\$0	\$15	\$15
Alice	{Bob}	\$25	\$25	\$0
Alice	{Charlie}	\$38	\$41	\$3
Alice	{Bob, Charlie}	\$51	\$51	\$0
Bob	$\emptyset$	\$0	\$25	\$25
Bob	{Alice}	\$15	\$25	\$10
Bob	{Charlie}	\$38	\$51	\$13
Bob	{Alice, Charlie}	\$41	\$51	\$10
Charlie	$\emptyset$	\$0	\$38	\$38
Charlie	{Alice}	\$15	\$41	\$26
Charlie	{Bob}	\$25	\$51	\$26
Charlie	{Alice, Bob}	\$25	\$51	\$26

# Calculating a Fair Share



Marginal contribution of Alice to an empty taxi



Marginal contribution of Alice to a taxi including already Bob



Do we get the same amount of information from every marginal contribution?



# Form Coalitions from Permutations

- One way to answer this question is by considering **all possible permutations** of Alice, Bob, and Charlie:
  - Alice, Bob, Charlie
  - Alice, Charlie, Bob
  - Bob, Alice, Charlie
  - Charlie, Alice, Bob
  - Bob, Charlie, Alice
  - Charlie, Bob, Alice
- In 2 out of 6 permutations, Alice is added to an empty taxi; In 1 out of 6, she is added to a taxi with Bob



The marginal contribution of Alice to an empty taxi is twice as informative as the marginal contribution of Alice to a taxi with Bob.

# Weighting of Marginal Contributions

- For a coalition  $S$  from  $N$  players, the probability of obtaining that coalition is

$$\frac{|S|!(N - |S| - 1)!}{N!}$$

Number of ways to arrange the players in  $S$

Number of ways to arrange the remaining players

Number of ways to arrange all players

```
def compute_weight(N_size, S_size):  
    w = fak(S_size)*fak(N_size - S_size - 1) / fak(N_size)  
    return w
```

# Calculating Shapley Values

Shapley value for player  $j$  is formalized as:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(N - |S| - 1)!}{N!} (v(S \cup \{j\}) - v(S))$$

Iterate over all possible  
coalitions

Weight the degree of  
information

Marginal contribution of  
adding  $j$  to the coalition  $S$

# The Problem With a Direction Translation of Shapley Values to ML

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(N - |S| - 1)!}{N!} (v(S \cup \{j\}) - v(S))$$

- One could think that a trivial translation of Shapley Values to the world of ML is possible.
- However, **the problem is the value function** which gives a real value for any kind of coalition of features.

**Definition (Background Data):** The replacement of absent feature values with randomly drawn ones requires a dataset to draw from, known as the background data. This could be the same data that was used to train the model. The background data serves as the context for the interpretation of the resulting SHAP values.

# Marginal Contribution

For a sample  $x^{(i)}$  that we want to explain, background data  $x^{(k)}$ ,  $k=1,\dots,n$ , the marginal contribution of  $j$  to  $S$  is:

$$\hat{\Delta}_{S,j} = \hat{v}(S \cup j) - \hat{v}(S) = \frac{1}{n} \sum_{k=1}^n \left( f(x_{S \cup j}^{(i)} \cup x_{C \setminus j}^{(k)}) - f(x_S^{(i)} \cup x_C^{(k)}) \right)$$



Another way to interpret the marginal contribution is that present features are known, absent feature values are unknown, so the marginal contribution illustrates how much the value changes from knowing  $j$  in addition to already knowing  $S$ .

# Putting it All Together

Combining all the terms into the Shapley value equation, we get the SHAP equation:

$$\hat{\phi}_j^{(i)} = \sum_{S \subseteq \{1, \dots, p\} \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} \hat{\Delta}_{S,j}$$

**Definition (SHAP Value):** The SHAP value  $\phi_j^{(i)}$  of a feature value for sample  $x^{(i)}$  is the average marginal contribution of a feature value  $x_j^{(i)}$  to all possible coalitions of features.

# Overview of SHAP Estimators

An issue with the exact computation of Shapley values is the large number of possible coalitions. As a remedy, we can either use the underlying ML Model structure (model specific), or approximate the computation by not computing all possible coalitions (model agnostic).

Method	Estimation (with inspiration)	Model-specific?	Method	Estimation (with inspiration)	Model-specific?
Exact	Iterates through all background data and coalitions	Agnostic	Tree, intervention- al	Recursively iterates tree paths	Tree-based
Sampling	Samples coalitions	Agnostic	Tree, path- dependent	Recursively iterates hybrid paths	Tree-based
Permutation	Samples permutations	Agnostic	Gradient	Computes the output's gradient with respect to inputs (inspired by Input Gradient)	Gradient-based
Linear	Exact estimation with linear model weights	Linear	Deep	Backpropagates SHAP value through network layers (inspired by DeepLIFT)	Neural Networks
Additive	Simplifies estimation based on additive nature of the model (inspired by GAMs)	GAMs	Partition	Recursive estimation based on feature hierarchy (inspired by Owen values)	Agnostic
Kernel	Locally weighted regression for sampled coalitions (inspired by LIME)	Agnostic			

# Questions

- Can we derive from a SHAP value of 0 that the feature value was uninformative?
- Can we derive from the feature value with the largest SHAP value that it has any effect on the model outcome?



# Contact

Prof. Dr. Mathias Kraus  
Chair for Explainable Artificial Intelligence  
Institute for Information Systems  
Bajuwarenstraße 4  
93053 Regensburg  
E-Mail: [mathias.kraus@informatik.uni-regensburg.de](mailto:mathias.kraus@informatik.uni-regensburg.de)