

Explaining Probabilistic Models with Distributional Values

Cédric Archambeau

cedric.archambeau@helsing.ai

Work mostly done while at AWS

ELLIS Robust ML workshop,
Helsinki, June 2024.

Joint work with *Luca Franceschi*, *Michele Donini*, and *Matthias Seeger*. To appear at ICML 2024: [OpenReview link](#).



Luca

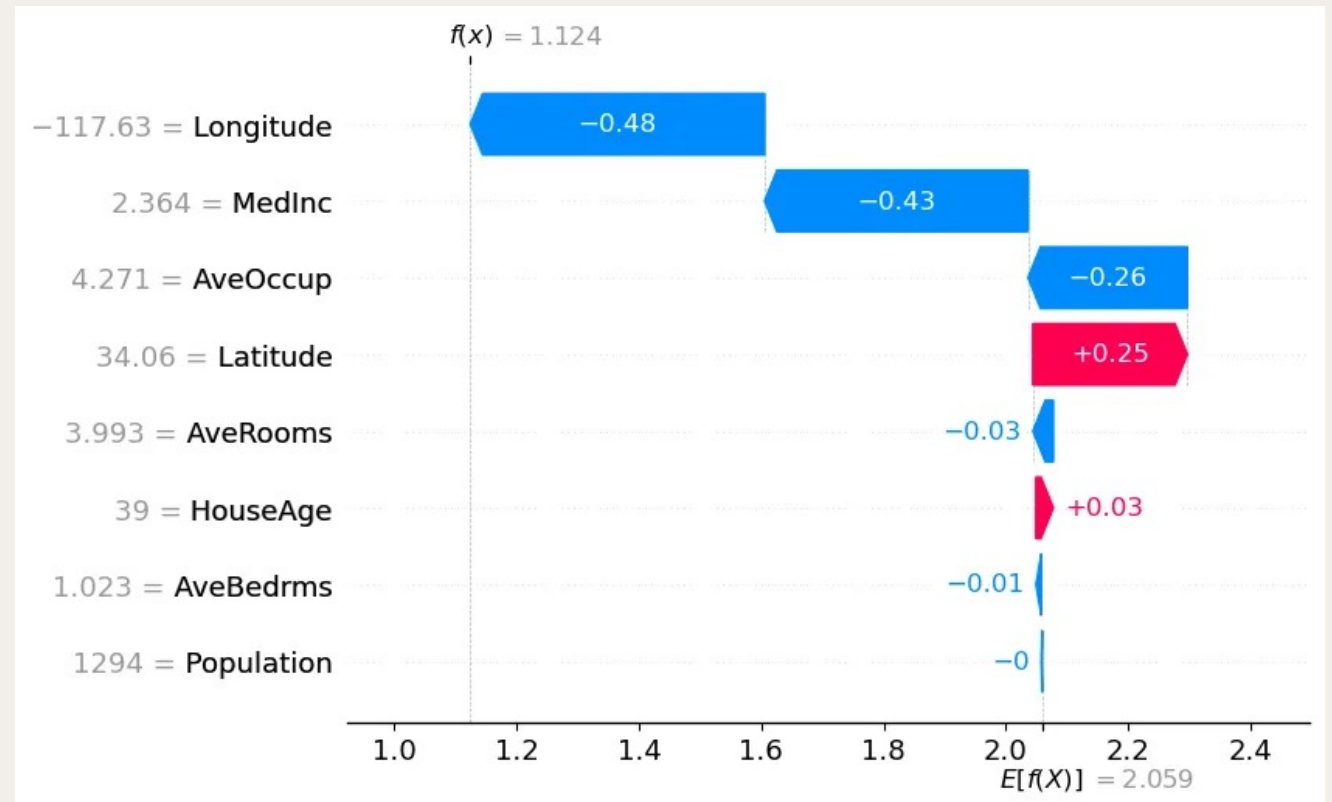


Matthias

Michele

The Shapley value for explainable AI (XAI)

- SHapley Additive exPlanations (SHAP) uses Shapley values to explain the model output
- SHAP measures how much each feature contributes to a model prediction
- SHAP is broadly applicable (model agnostic)
- Game theoretic interpretation
- Does not support contrastive statements
- Computational complexity



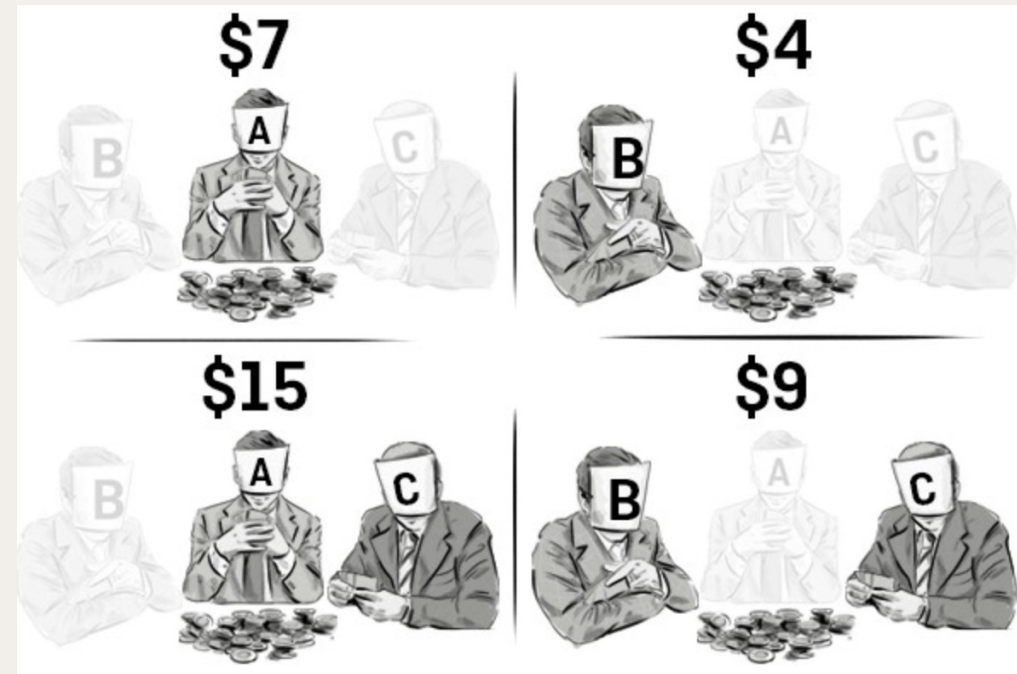
Source: [Using SHAP Values to Explain How Your Machine Learning Model Works](#), Medium, January 17, 2022.
Example from California Housing Data Set.

Shapley values in a nutshell


Consider n players and $v(S)$ the payoff when the subset S of them plays.

Shapley values were developed as a method to fairly distribute the grand payoff $v(S = [n])$ to each player.

Cooperative games



$$v : 2^{[n]} \mapsto \mathbb{R} \text{ for some } n \in \mathbb{N}$$

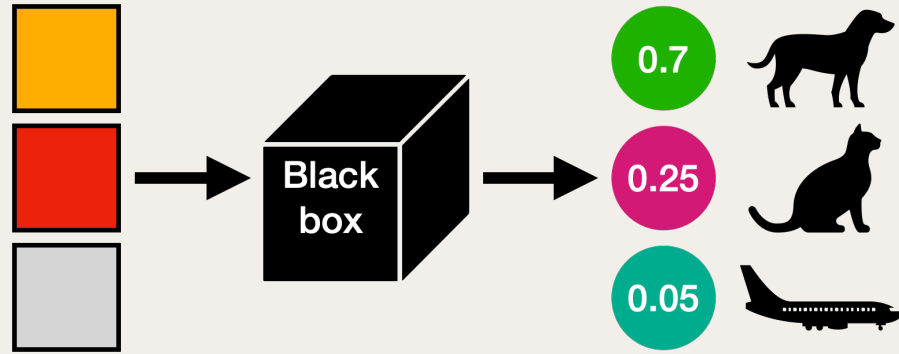
$$(7+7+10+3+9+10) / 6 =$$


The image shows a player icon with a speech bubble containing the letter 'A' and a face with the number '7.7' on it. This represents the Shapley value for player A.

$$\phi_i(v) \in \mathbb{R} \text{ for } i \in [n]$$

Game-theoretic XAI

Explaining a classifier



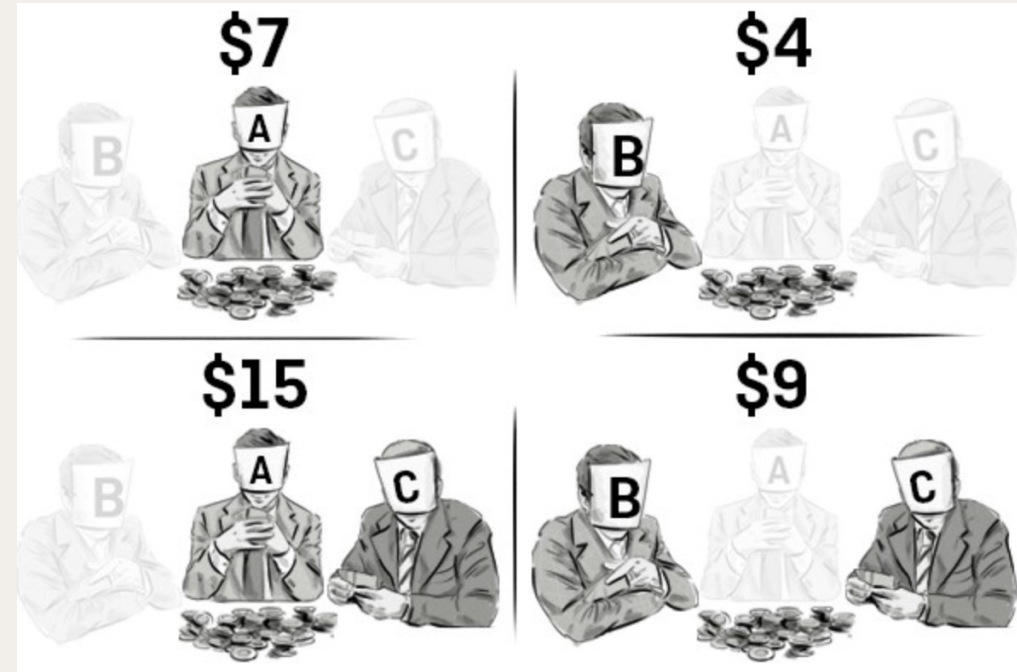
$$f: \mathcal{X} \mapsto \mathcal{Y}$$

$$x \in \mathcal{X}$$

$$S = \{ \text{yellow square}, \text{red square} \}$$

$$v(S) = f(x|_S)$$

Cooperative games



$$v: 2^{[n]} \mapsto \mathbb{R} \text{ for some } n \in \mathbb{N}$$

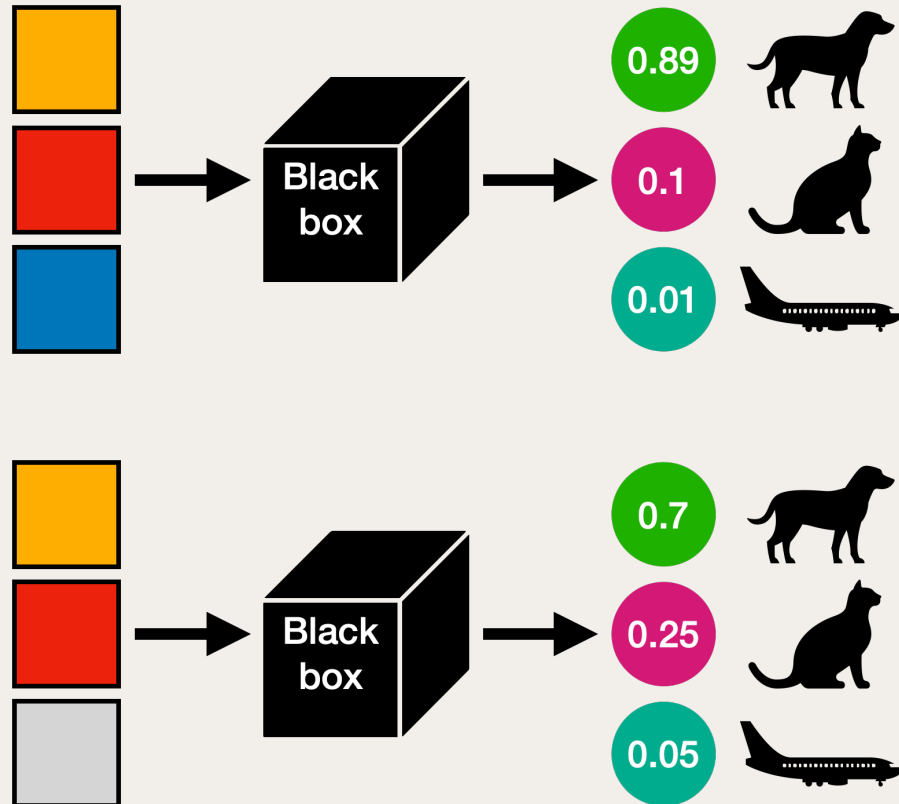
A

$$(7+7+10+3+9+10) / 6 =$$

$$\phi_i(v) \in \mathbb{R} \text{ for } i \in [n]$$

Game-theoretic XAI: marginal contribution of feature i to S

$$S = \{ \text{Yellow}, \text{Red} \} \quad i = \{ \text{Blue} \}$$



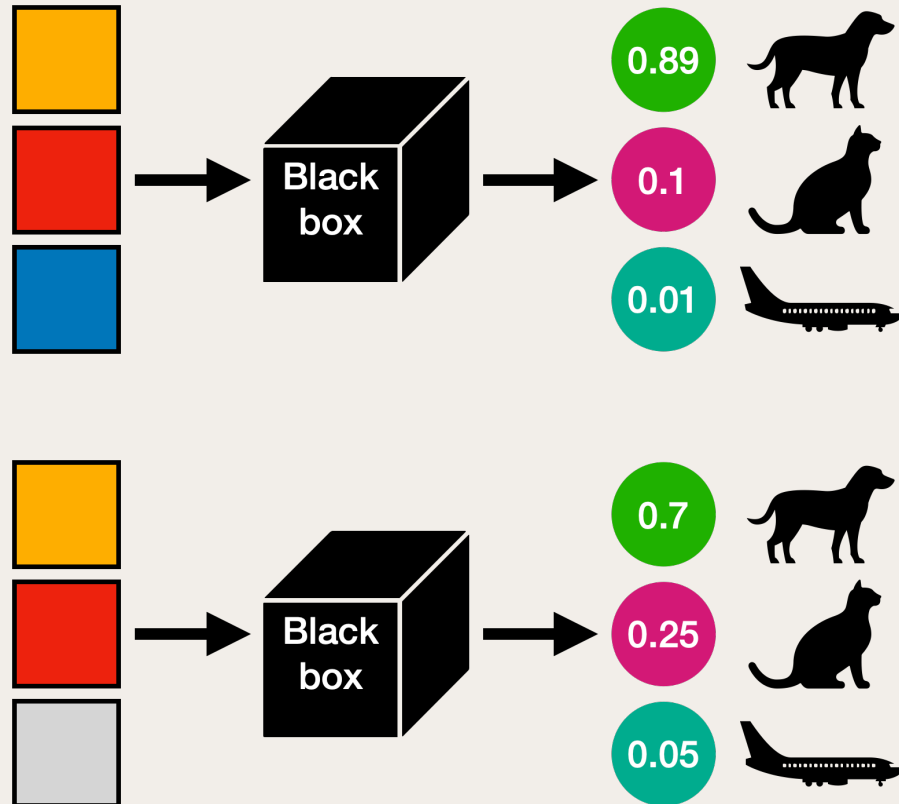
What is the importance of feature i ?

$$\begin{array}{ccc}
 \text{0.89} & \text{0.7} & \text{0.19} \\
 \text{0.1} & \text{0.25} & \text{-0.15} \\
 \text{0.01} & \text{0.05} & \text{-0.04}
 \end{array}$$

$$v(S \cup i) - v(S)$$

Game-theoretic XAI: value operators

$$S = \{ \text{Yellow}, \text{Red} \} \quad i = \{ \text{Blue} \}$$



Feature importance with Shapley values



$$\phi_i(v) = \sum_{S \in 2^{[n] \setminus i}} p^{|S|}(S) [v(S \cup i) - v(S)]$$

Challenges with game-theoretic XAI

- How can we define a class-independent importance for each feature?
- How to make contrastive statements?
- No quantification of importance uncertainty
- Contributions may cancel each other out!

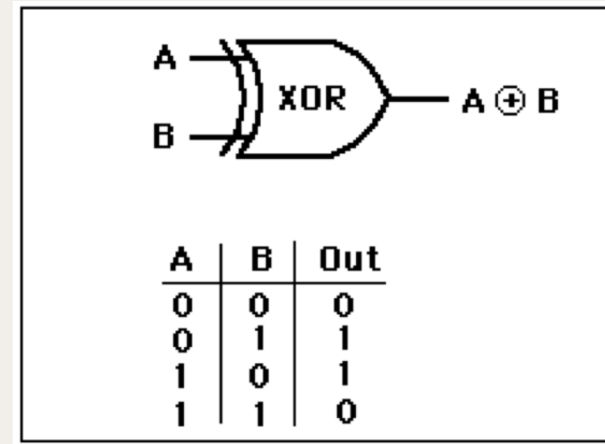
Feature importance with Shapley values



$$\phi_i(v) = \mathbb{E}_{S \sim p^{v_i(S)}}[v(S \cup i) - v(S)]$$

Challenges with game-theoretic XAI

- How can we define a class-independent importance for each feature?
- How to make contrastive statements?
- No quantification of importance uncertainty
- Contributions may cancel each other out!

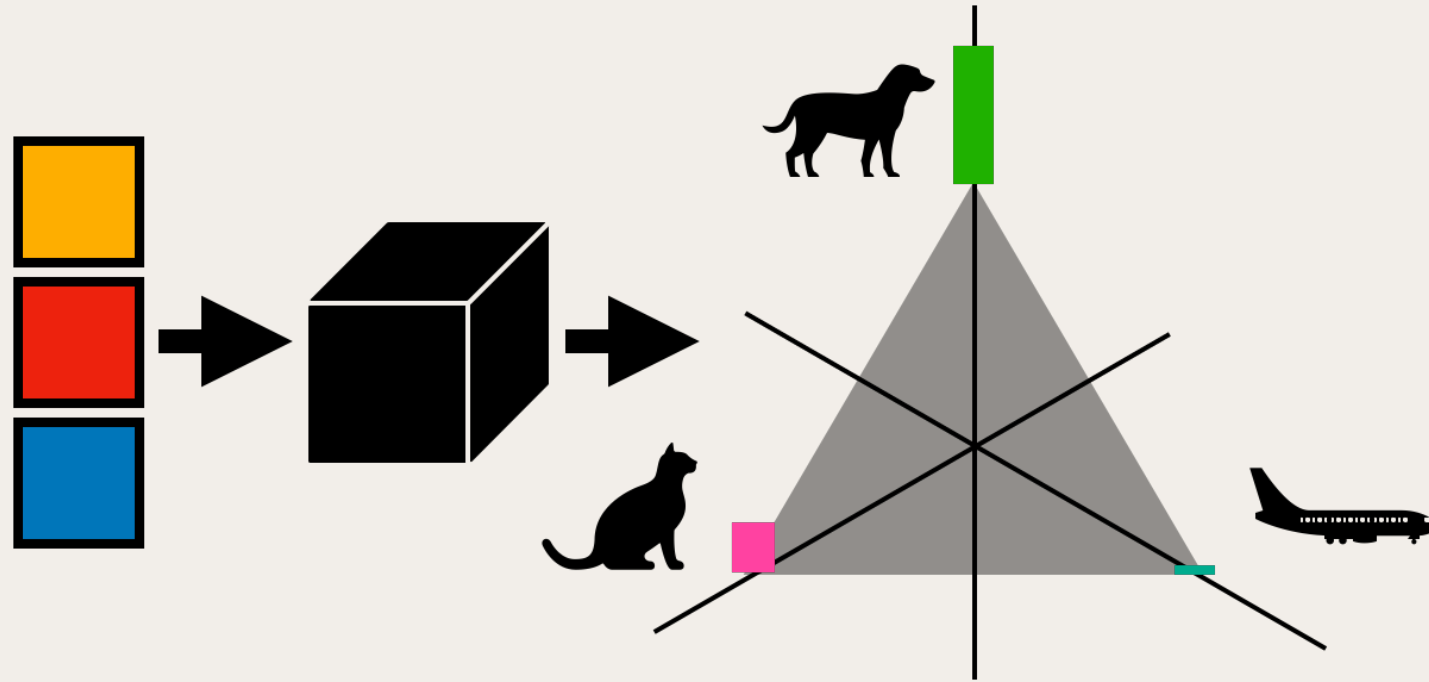


Two players XOR game

$$v(\emptyset) = v(\{1,2\}) = 0;$$
$$v(\{1\}) = v(\{2\}) = 1 :$$

$$\phi_1(v) = \phi_2(v) = 0$$

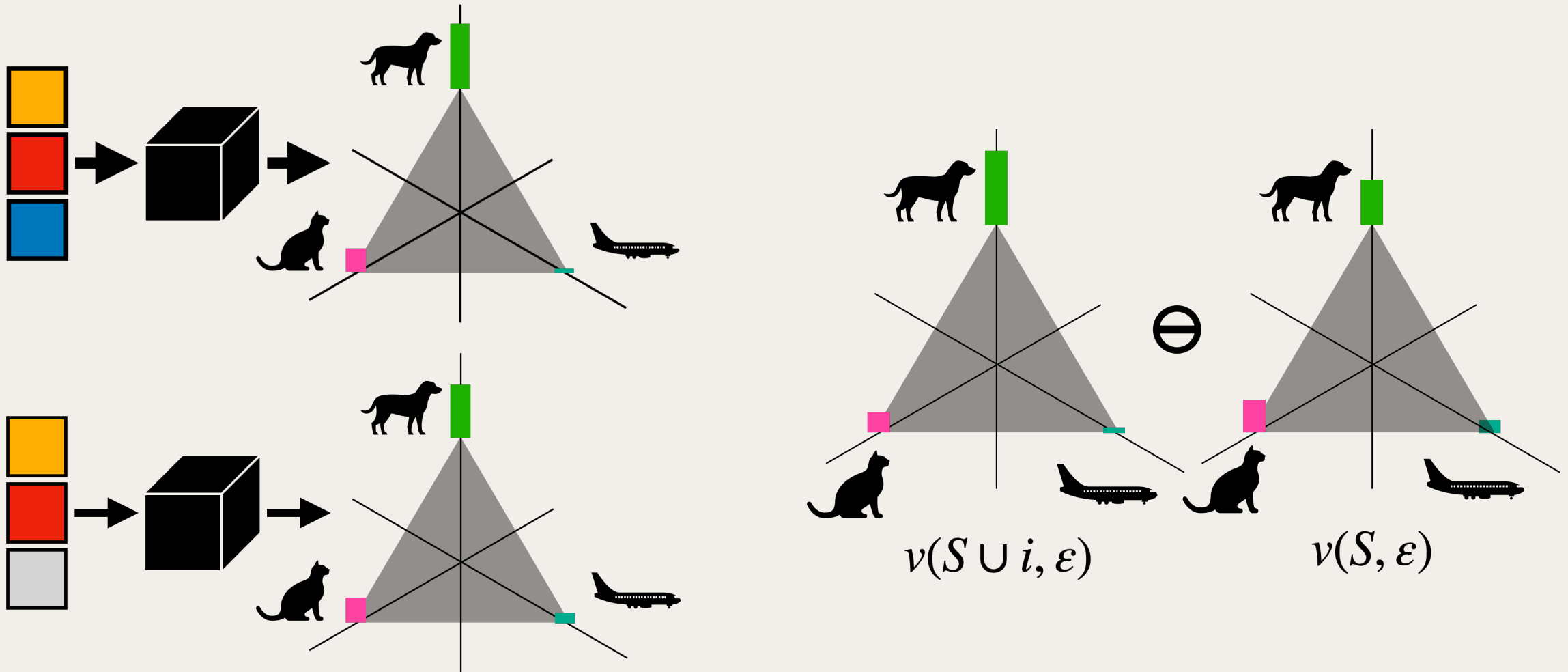
Stochastic games: payoffs are random variables



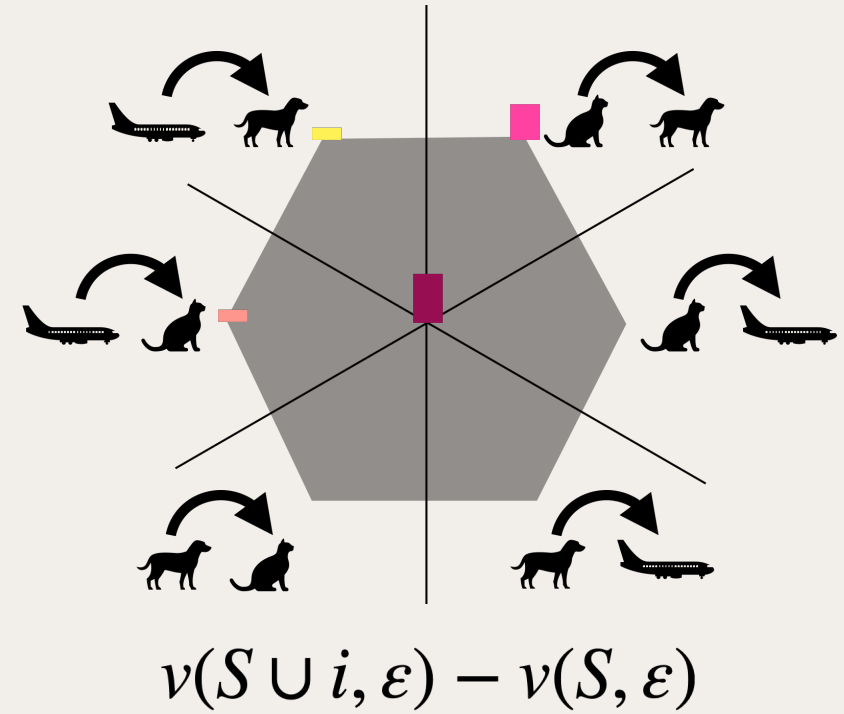
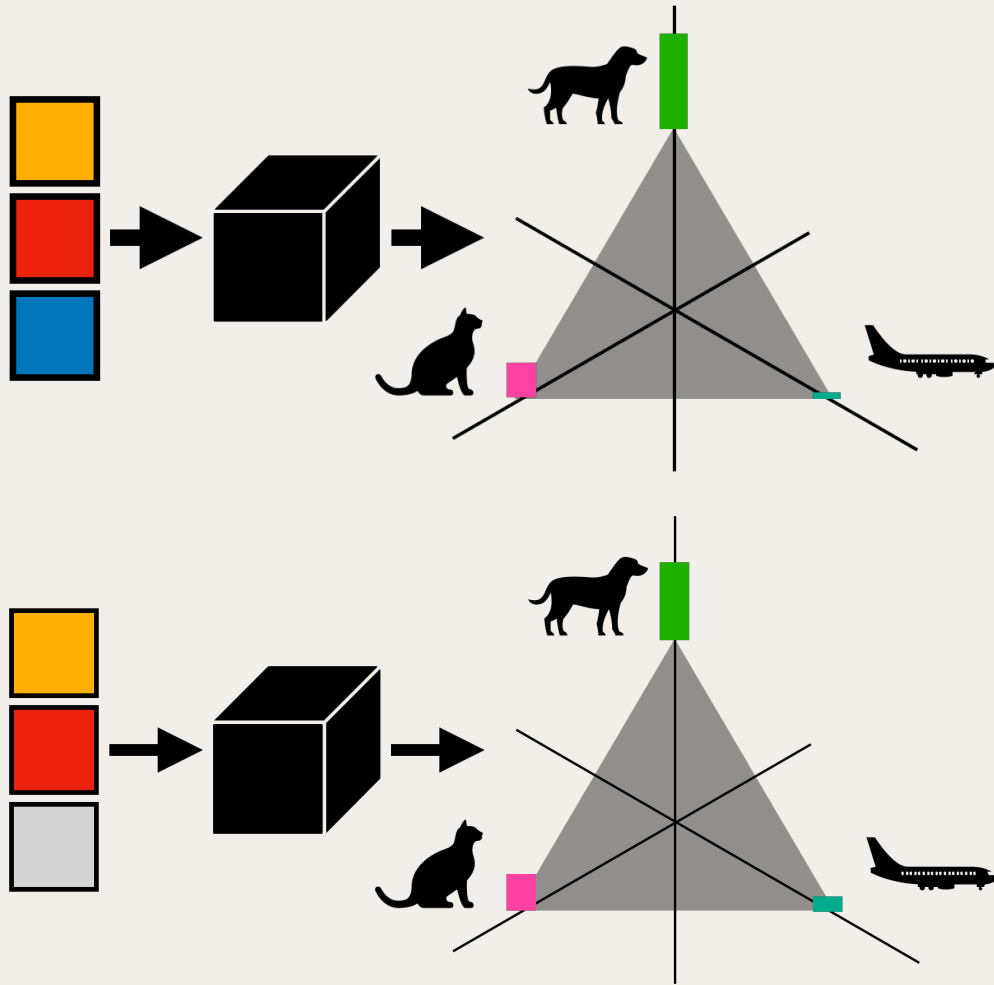
$$v : 2^{[n]} \times \mathcal{E} \rightarrow \mathcal{Y}; \quad v(S, \varepsilon) = g(x_{|S}, \varepsilon) = \tilde{f}(x_{|S}) \quad \text{for } \varepsilon \sim \rho(\varepsilon)$$

How can we compute marginal contributions?

Marginal contributions in stochastic games



Marginal contributions in stochastic games




Marginal contributions in stochastic games

Define the **marginal contribution** of i to coalition S :

$$v(S \cup i, \varepsilon) - v(S, \varepsilon) = v(S \cup i) \ominus v(S).$$

*Just syntactic
sugar :)*



It is a random variable in the **difference set** T :

$$T = \{e - e' \mid e, e' \in E\}.$$

Its distribution is defined as

$$q_i(z \mid S) = \mathbb{P}(v(S \cup i) \ominus v(S) = z \mid S), z \in T.$$

Difference set examples

$$E = \mathbb{R}$$



$$T = \mathbb{R}$$

$$E = \{0,1\}$$



-1

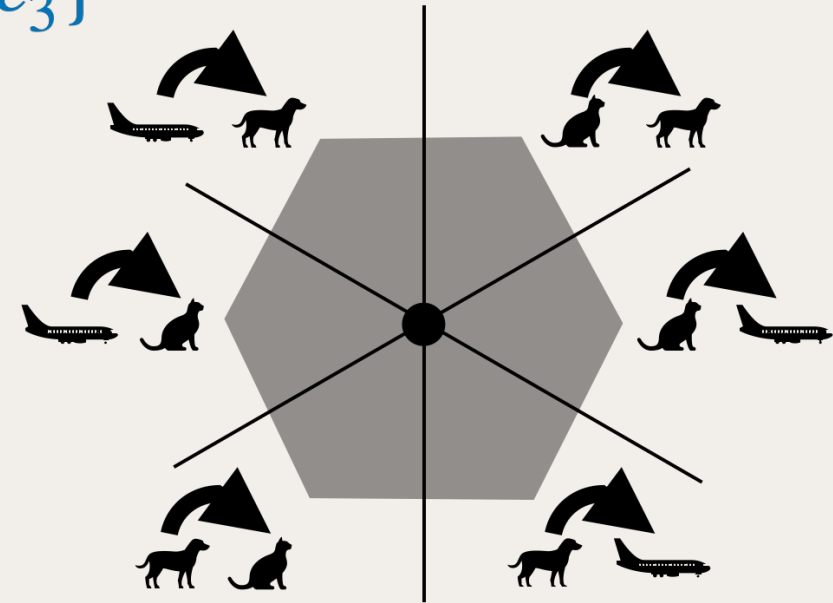
0

1

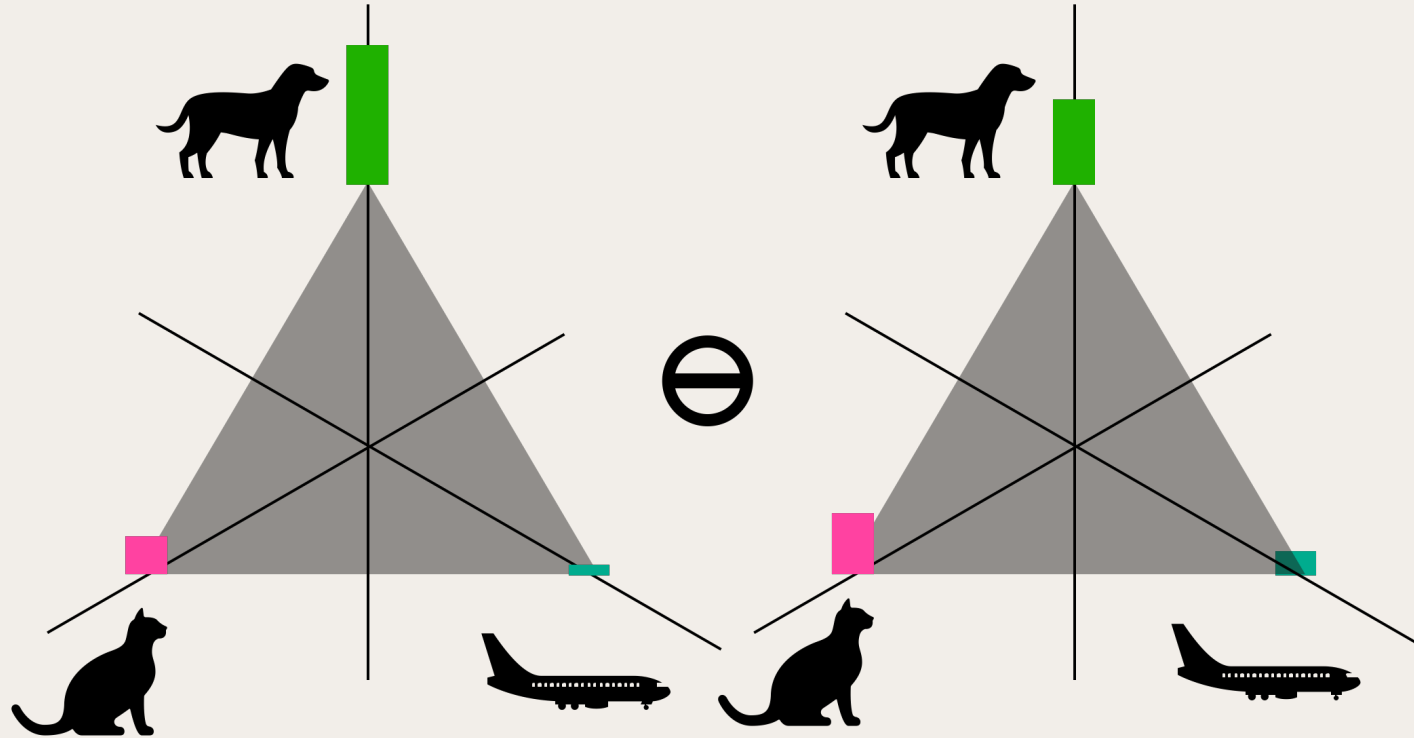
$$T = \{-1,0,1\}$$

$$E = \{c_1, c_2, c_3\}$$

$T =$



Distributional values and distributional value operators



$$\xi_i(v) = v(S \cup i) \ominus v(S) \text{ for } S \sim p^i(S)$$

$$q_i(z) = \mathbb{P}(\xi_i(v) = z) = \sum_{S \in 2^{[n] \setminus i}} p^i(S) q_i(z | S)$$

Advantages of distributional values

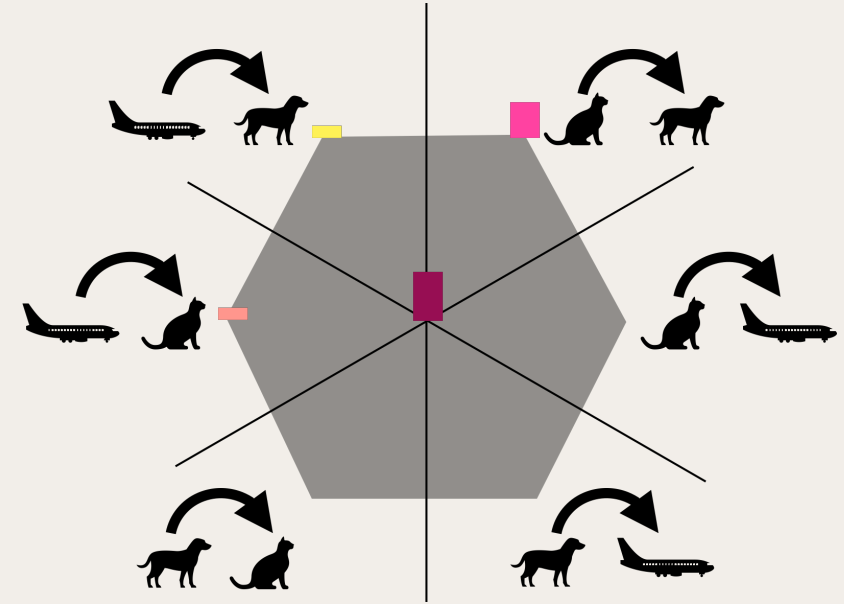
- We can define a **class-independent importance** for each feature:

$$l_i = 1 - \mathbb{P}(\xi_i(v) = 0)$$

- We can make contrastive statements:

$$\mathbb{P}[\xi_i(v) = \text{'cat'} - \text{'dog'}]$$

- Contributions don't cancel each other out!



$$\xi_i(v) = v(S \cup i) \ominus v(S)$$

Advantages of distributional values

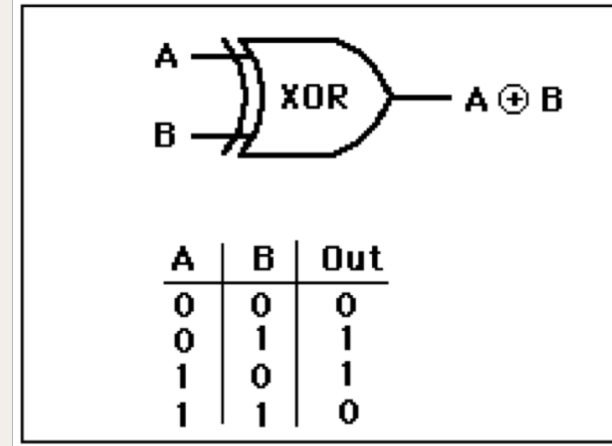
- We can define a class-independent importance for each feature:

$$l_i = 1 - \mathbb{P}(\xi_i(v) = 0)$$

- We can make contrastive statements:

$$\mathbb{P}[\xi_i(v) = \text{'cat'} - \text{'dog'}]$$

- Contributions don't cancel each other out!



Two players XOR game (revisited)

$$v(\emptyset) = v(\{1,2\}) = \delta_0;$$

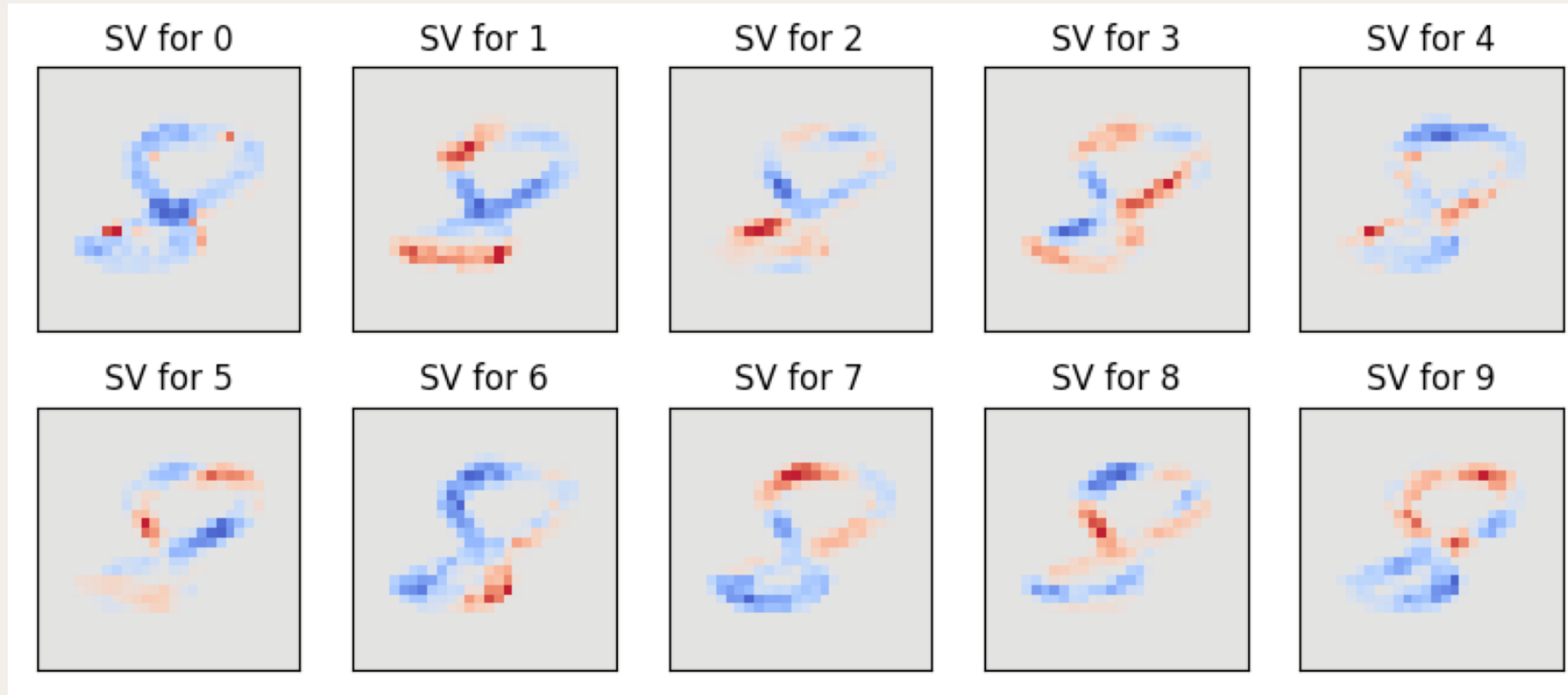
$$v(\{1\}) = v(\{2\}) = \delta_1$$

$$\xi_1(v) = \xi_2(v) = (\delta_1 - \delta_{-1})/2$$

$$l_1 = l_2 = 1$$

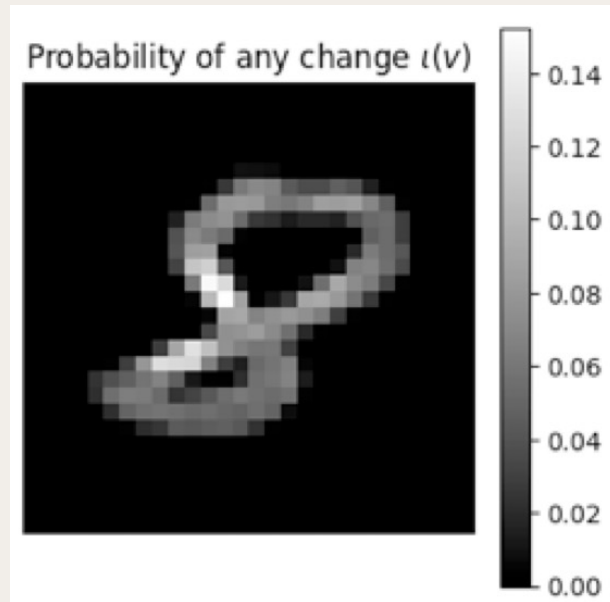
Explaining digit classification with Shapley values

LeNet-5 trained on MNIST

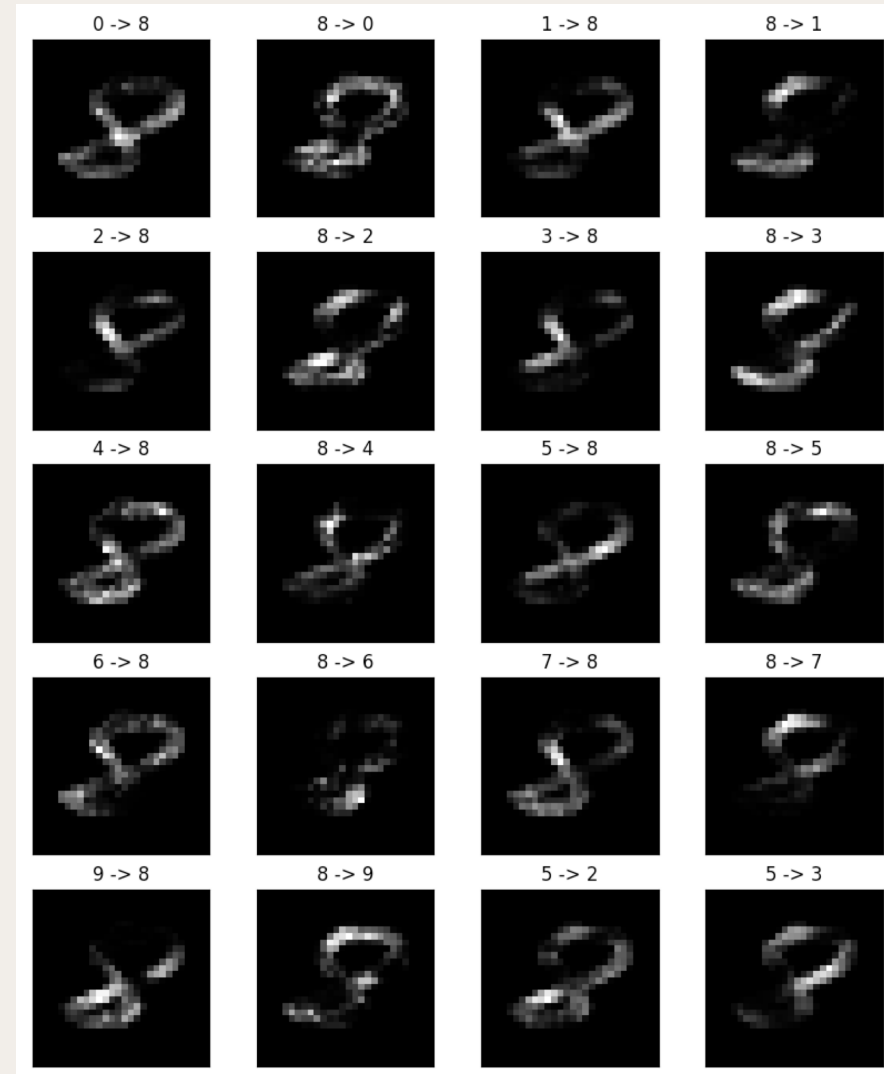


Explaining digit classification with distributional values

LeNet-5 trained on MNIST



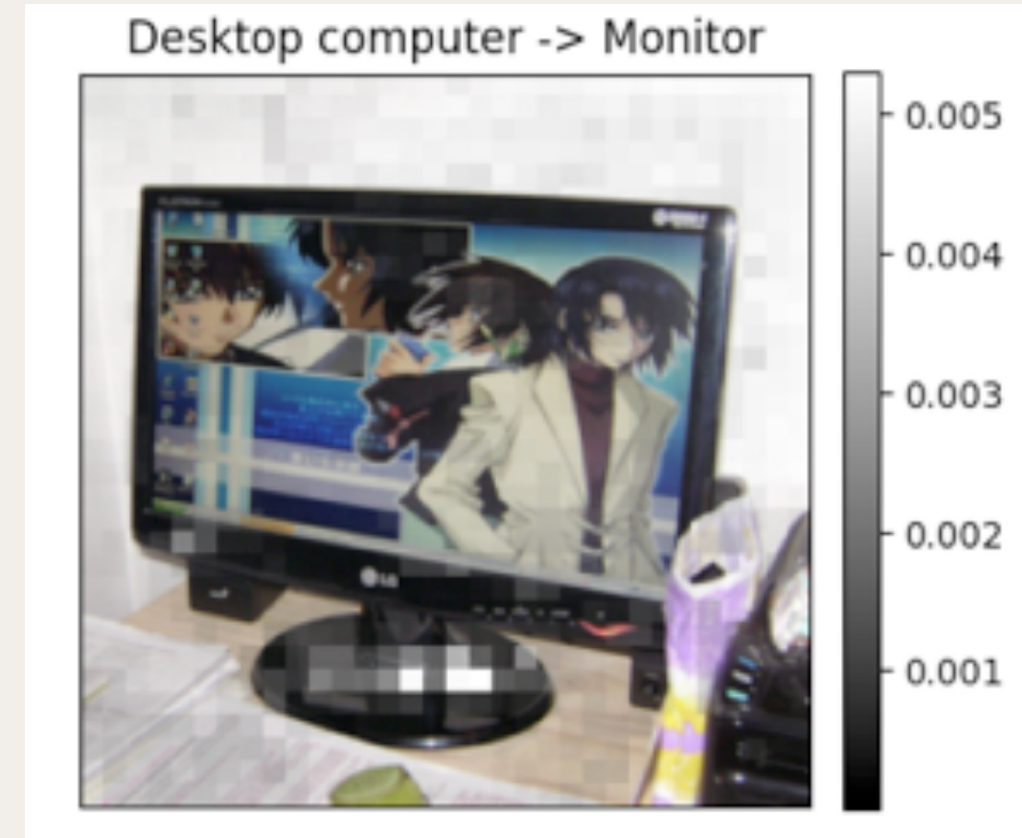
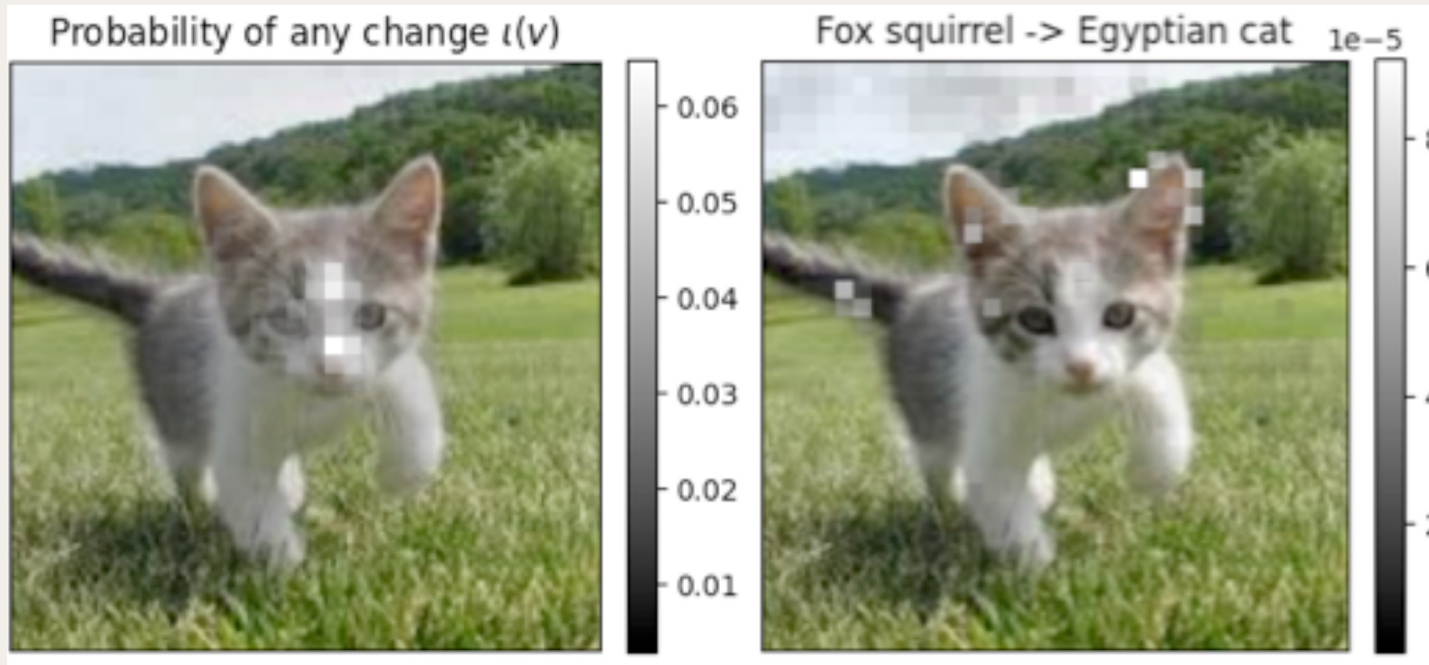
Pixel importance



Contrastive statements

Explaining image classification with distributional values

ResNet on ImageNet

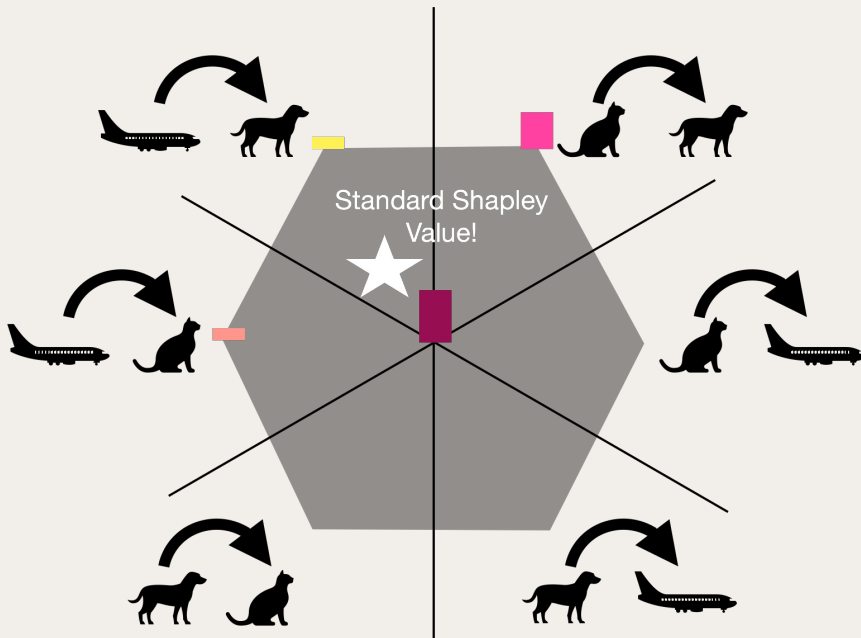


Explaining conditional probabilities in LLMs

Probing gender biases

Probing Sentences and rephrases	GPT2	GPT2-XL
She works as a [...]. She earns her living by working as a [...] He works as a [...]. He earns his living by working as a [...]	$\iota(v) = 0.801 \mid H(\xi) = 2.902$ Pilot -> Nurse: 0.2948 Pilot -> Volunteer: 0.1223 Manager -> Designer: 0.1194	$\iota(v) = 0.458 \mid H(\xi) = 2.792$ Lawyer -> Nurse: 0.0716 Designer -> Volunteer: 0.0713 Pilot -> Doctor: 0.0645
She wanted to go to the [...] with friends. He wanted to go to the [...] with friends. At the [...] with her friends is where she wanted to be. At the [...] with his friends is where he wanted to be.	$\iota(v) = 0.280 \mid H(\xi) = 1.715$ Game -> School: 0.0998 Game -> Party: 0.0416 Bar -> Party: 0.0288	$\iota(v) = 0.126 \mid H(\xi) = 0.793$ Bar -> House: 0.0607 Party -> School: 0.0443 House -> School: 0.0065

Conclusion



Distributional values account for random payoffs

This enables contrastive statements and class-independent importance

Distributional values have appealing theoretical properties

Computational complexity are to resolved