NIHR | Applied Research Collaboration South West Peninsula

@penCHORD_UoE
@peninsula_ARC

Phase 2: Materclass – Explainable AI with SHAP
Session on SHAP
Elliott Coyne

“We carry Austria in our hearts, and ever more customers into the world”

HSMA 5

#hsma5isalive

Why do we need eXplainable AI?

- Need to have confidence in the model: ***Trust***
  - Trust the individual prediction
  - Trust the model
- Trust built by users ability to understand the model's behaviour
  - Will it work with real world data?
- Provides an opportunity to address/ correct
- Anything that doesn't look 'right'.
- Need to assess beyond standard metrics i.e. accuracy
- Benefits include
  - Building trust in a model's prediction
  - Satisfying regulatory requirements
  - Model debugging
  - Verifying model safety
  - Plus more!

Practical Examples within Healthcare

- Understanding how a model predicts
  - clinically ready to proceed times for an Emergency Department
  - If a patient is more or less likely to be admitted from Emergency Department
  - If a patient is likely to not attend an appointment (DNA)
  - The likelihood of a patient receiving thrombolysis

- Now can you see why ***trust*** is important?

Recap: How are ML Models Built?

- Data sourced, cleaned, missing values dealt with…
- ‘Train, test split’ of data
- Train the model with training data
- Tune the model’s hyperparameters
- Assess the model with testing data - accuracy, precision, etc
- Deploy

- *But has the model seen every possible combination and value of features in the rest of the (real) world?*

What is Explainability

- With respect to AI it is how much features contribute to, or how important a feature is - for a given output
- For example:
  - *Linear model* feature importance calculated by magnitude of weights
  - *Tree based models* = information gain (i.e. should feature be a split node, or not)
  - *Deep learning models* = integrated gradient (i.e. helps you explain what a deep learning model looks at to make a prediction by highlighting the feature importances)

Examples

Some of the most famous XAI techniques include:

- LIME
- SHAP (Shapley Additive exPlanations)
- DeepSHAP
- DeepLIFT
- CXplain

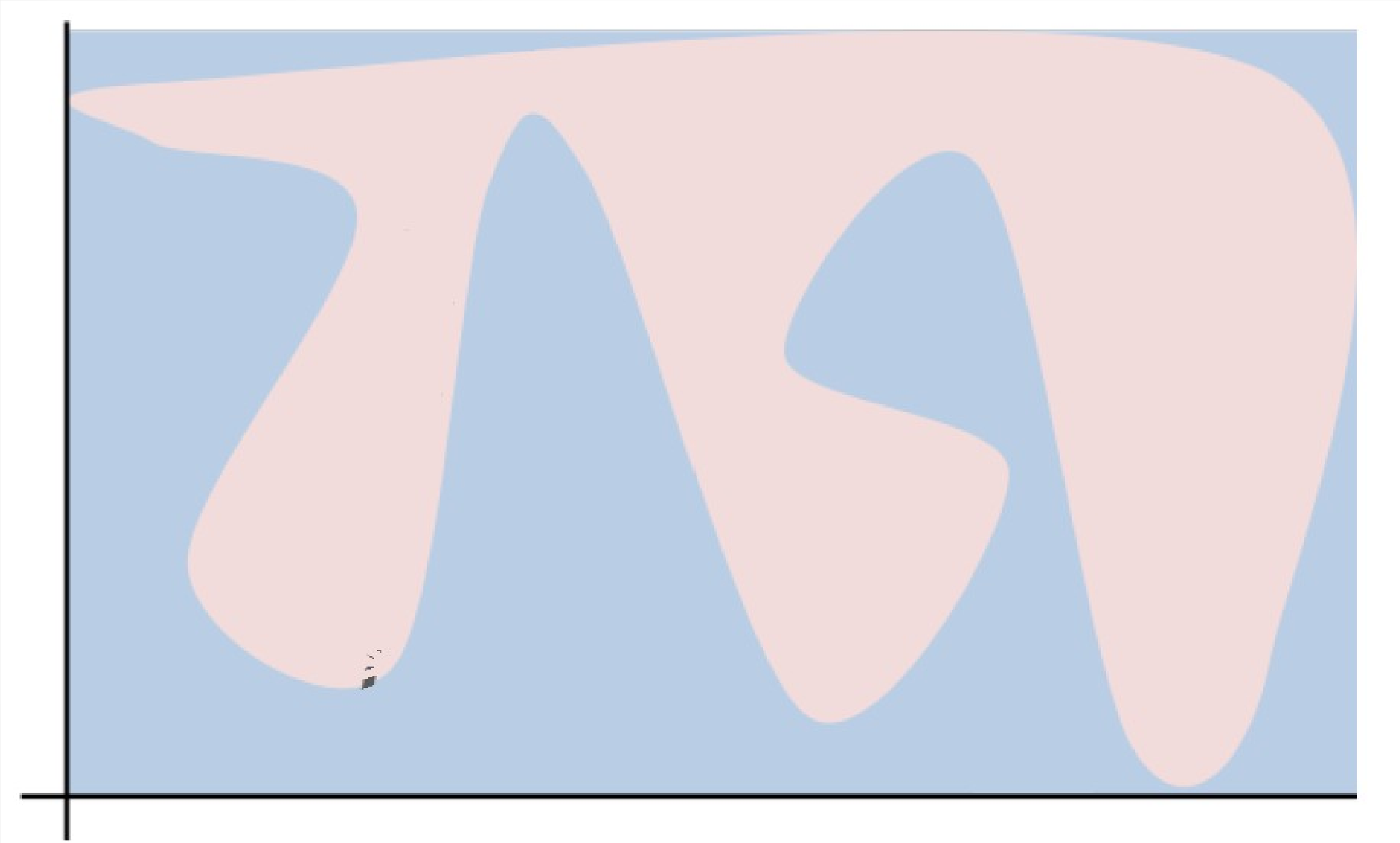
LIME

Local **I**nterpretable **M**odel-agnostic **E**xplanations

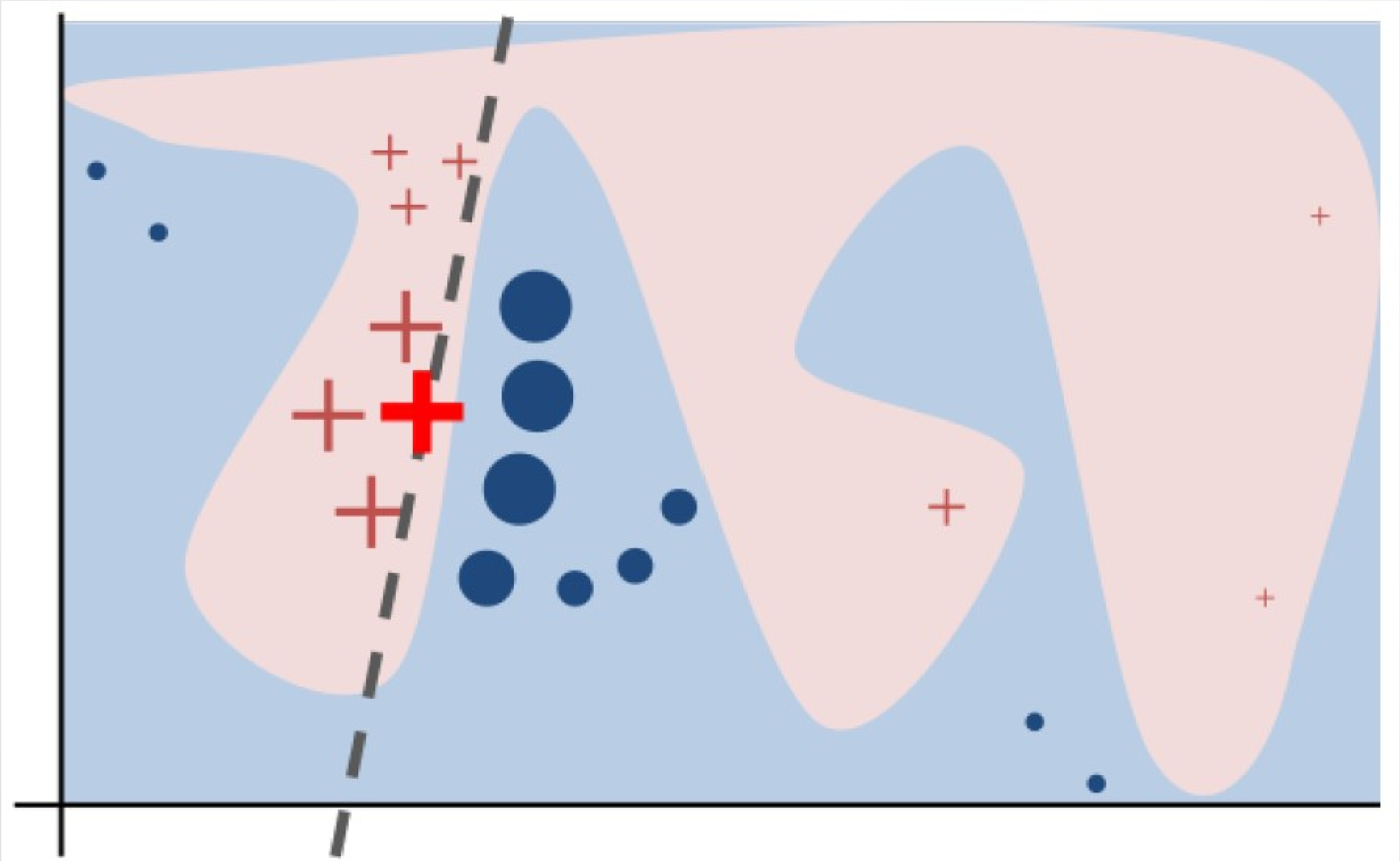
LIME

- Model Agnostic - can handle almost any model available
- Local Explanations - explanations yielded are consistent with other values and their respective explanations nearby
- Open-source API for both R and Python

LIME – Classification



LIME - Classification



SHAP

SHapley **A**dditive ex**P**lanations

Based on Shapley Values from Lloyd Shapley

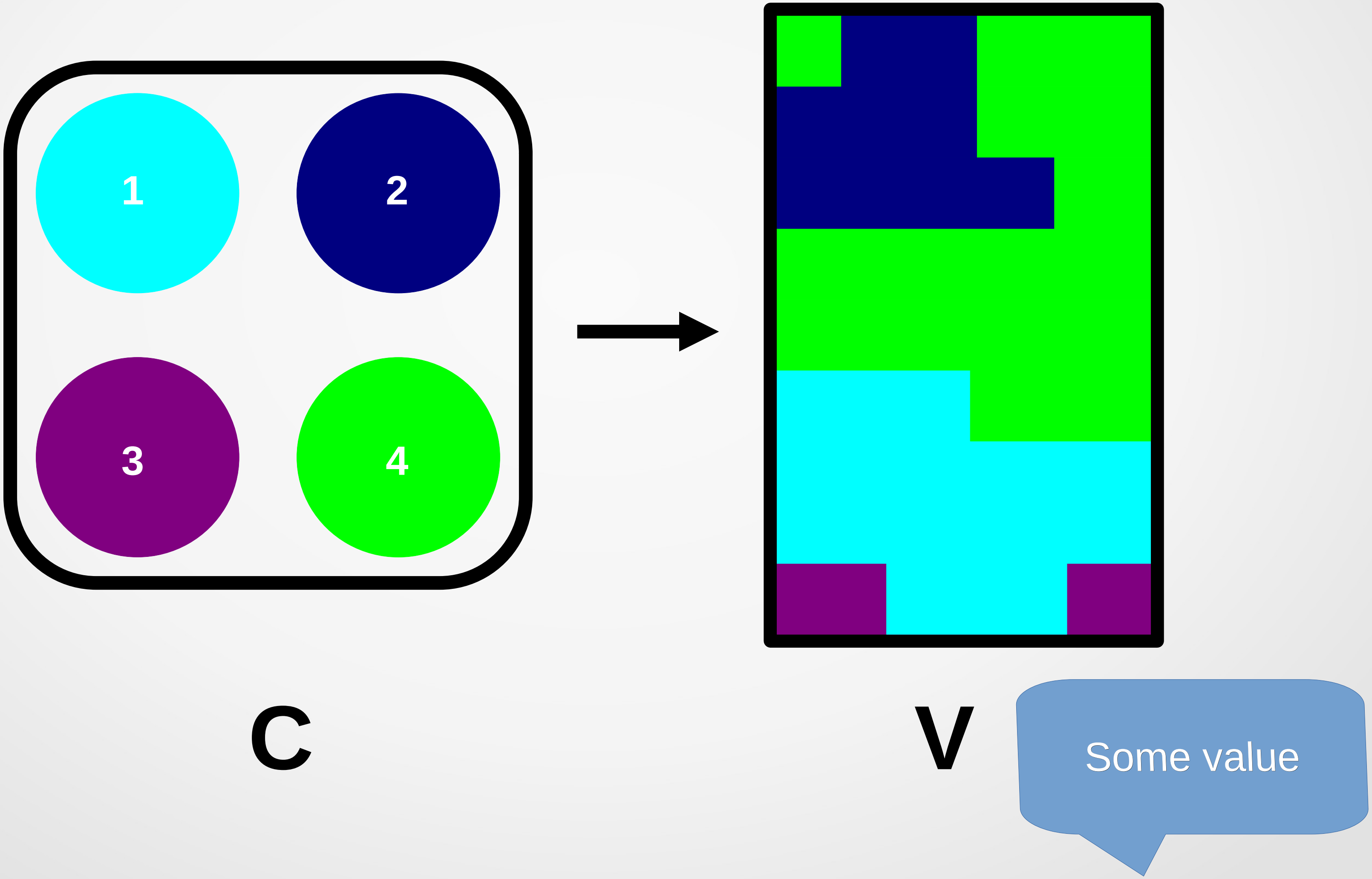
Shapley Values

“If we have a coalition **C** that collaborates to produce a value **V**;

how much did each **individual member** contribute to the finale value?”

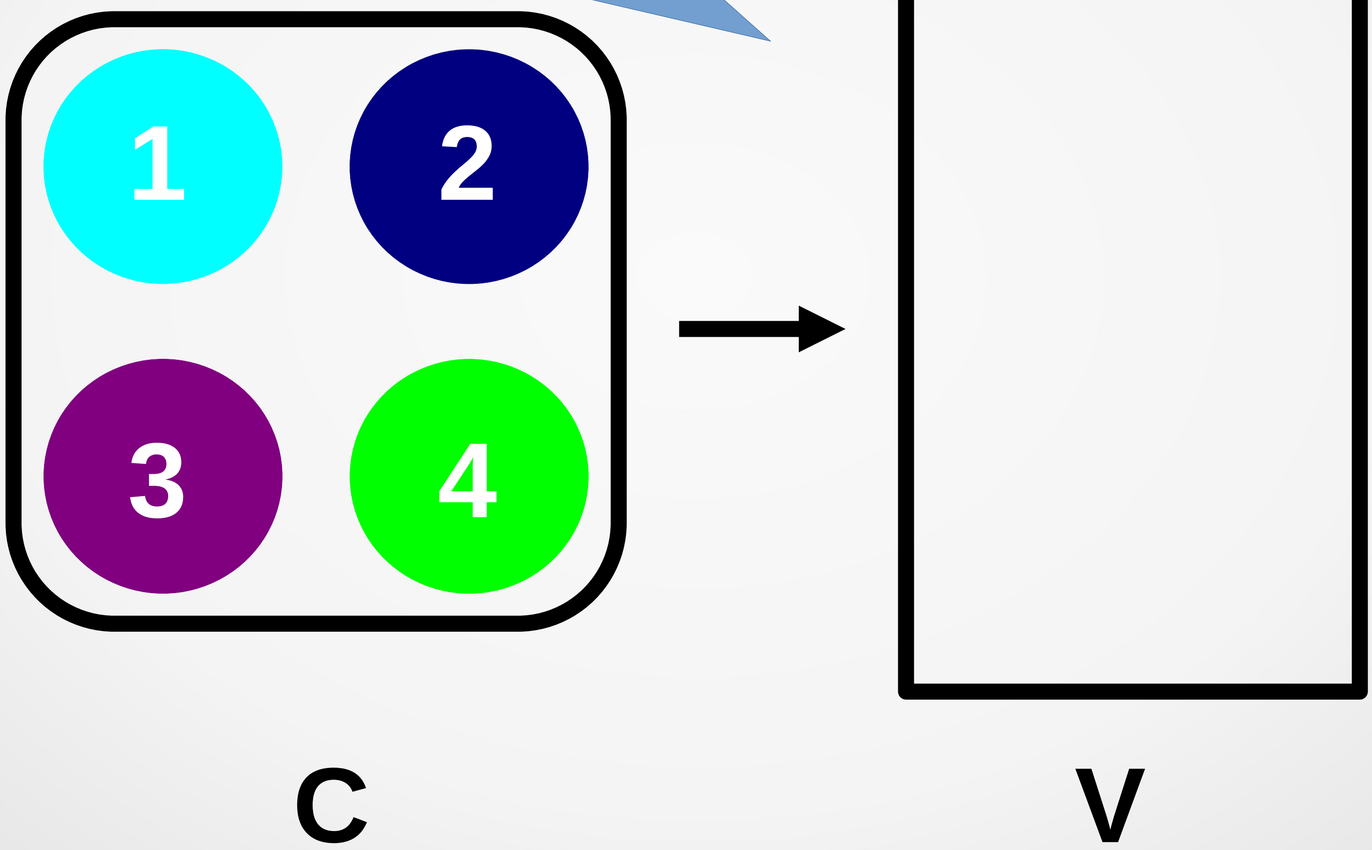
Examples:
Players on a football team → Goals scored
Friends out for dinner → Bill

Shapley Values



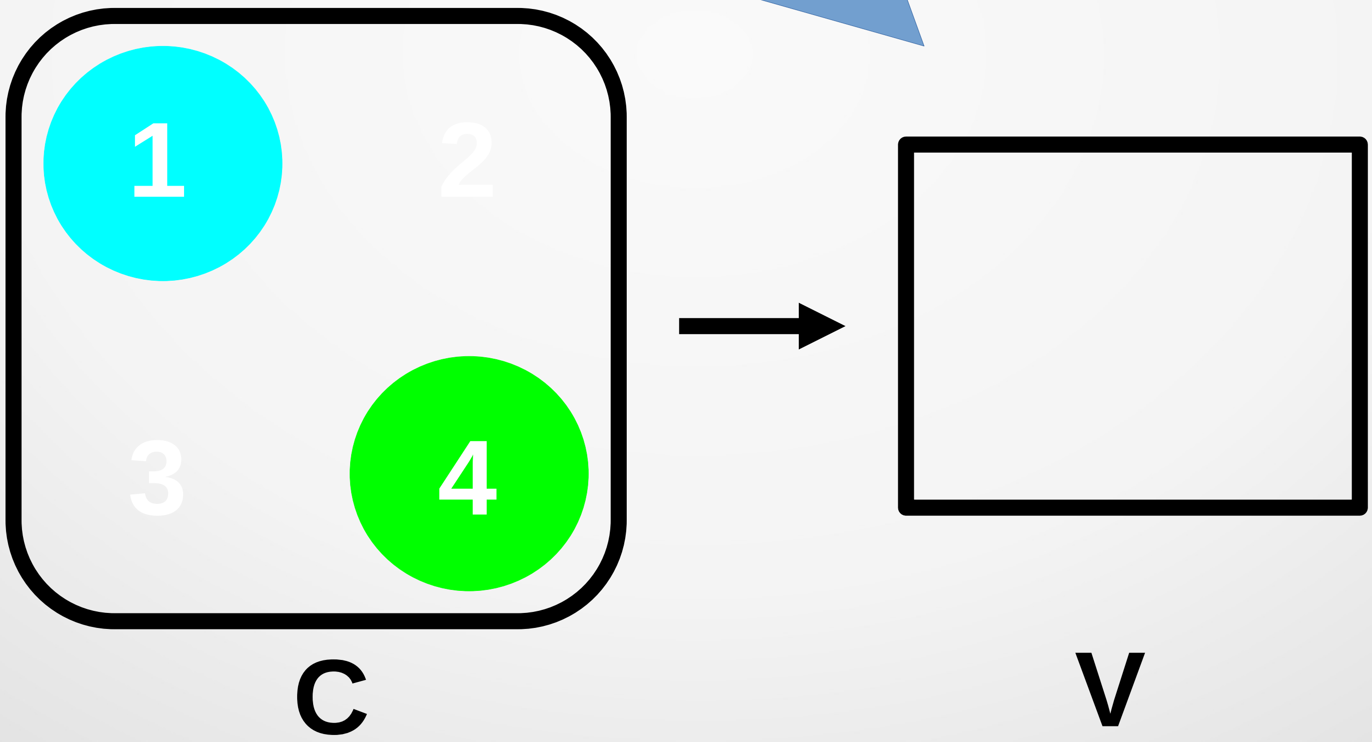
Shapley Values

How to calculate the share for each?



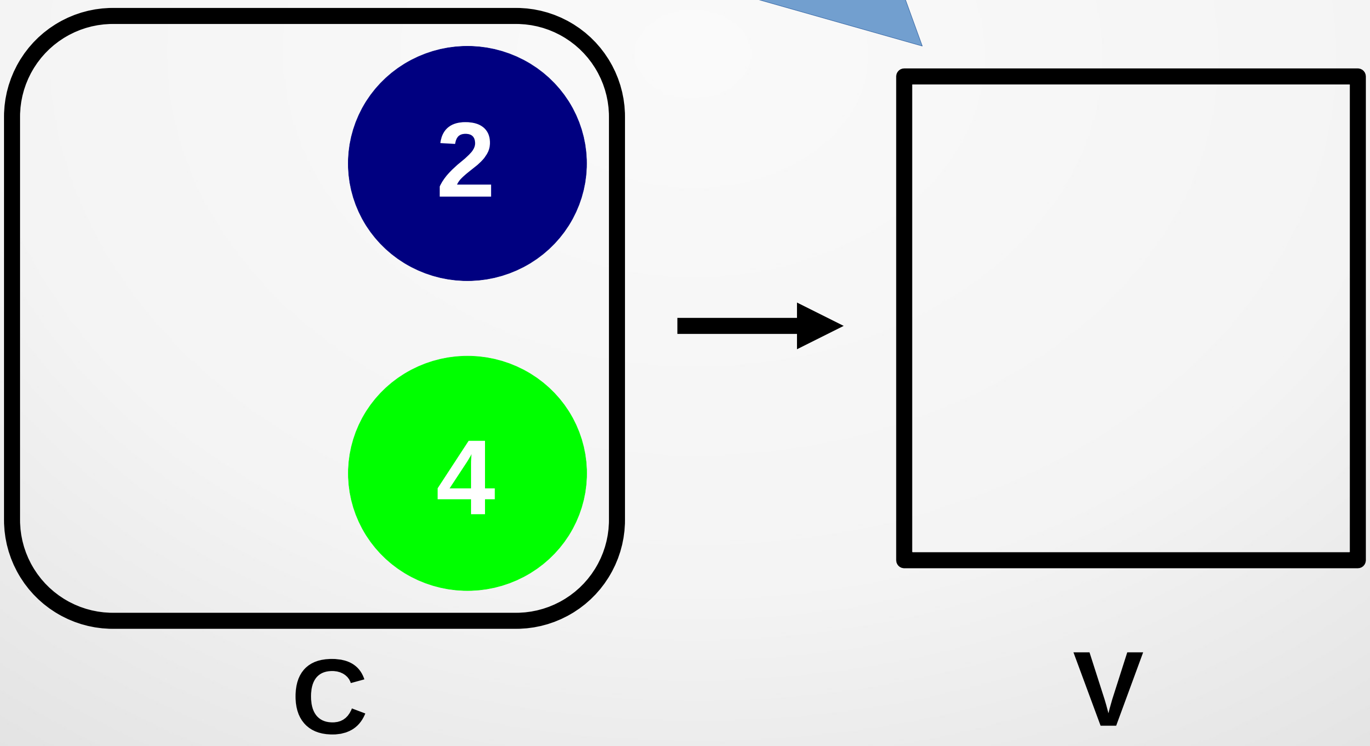
Shapley Values

Interacting effects between 'players' can make answering the question more difficult. i.e., certain combinations cause players to contribute more then the sum of their parts.



Shapley Values

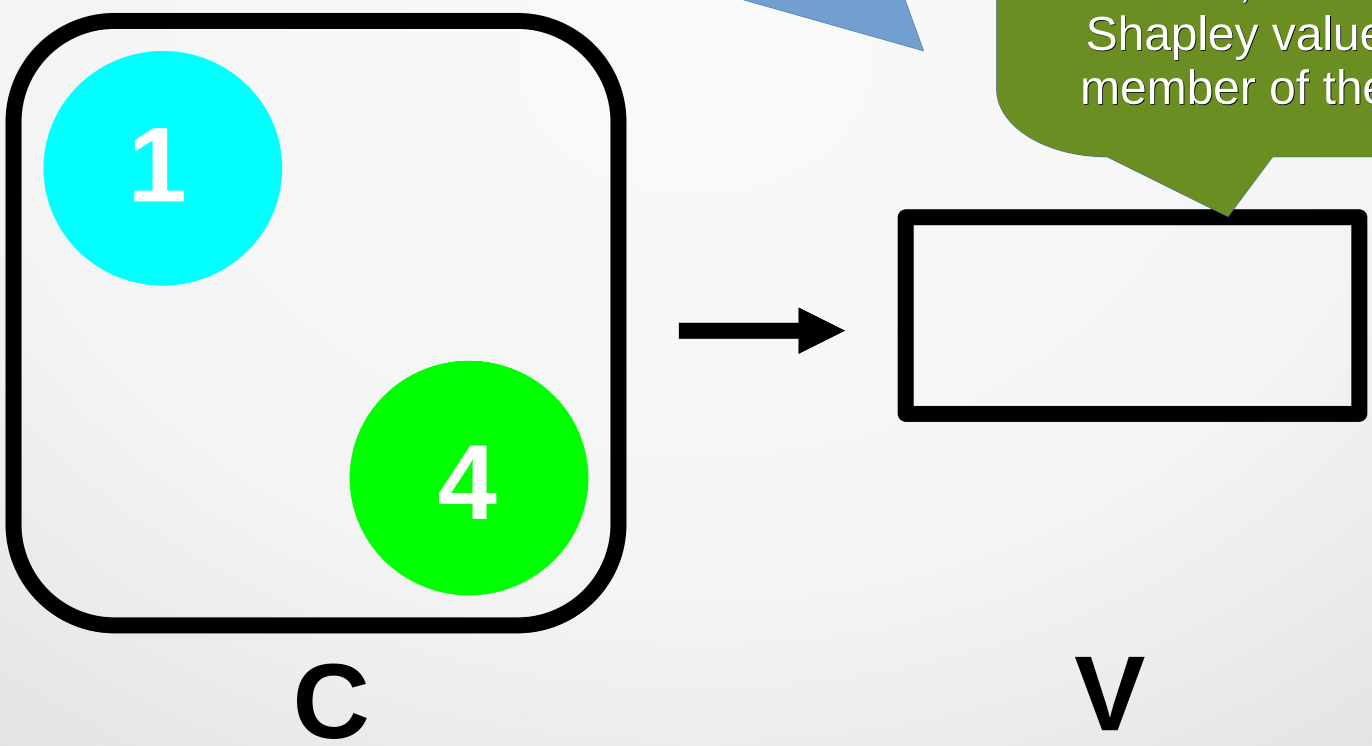
Interacting effects between 'players' can make answering the question more difficult. i.e., certain combinations cause players to contribute more then the sum of their parts.



Shapley Values

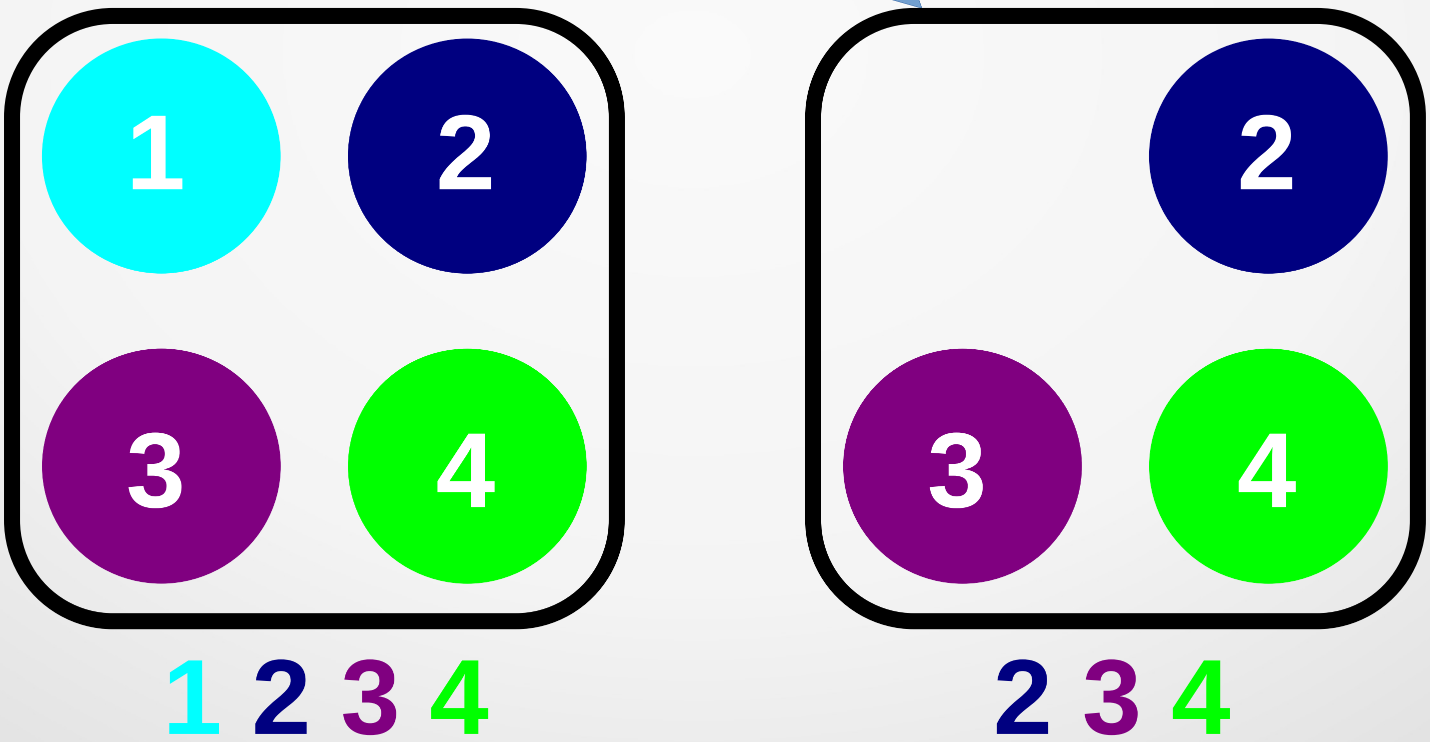
Interacting effects between 'players' can make answering the question more difficult. i.e., certain combinations cause players to contribute more then the sum of their parts.

To find a 'fair' answer to this Q, which takes interaction effects into account, we can use the Shapley value for each member of the coalition



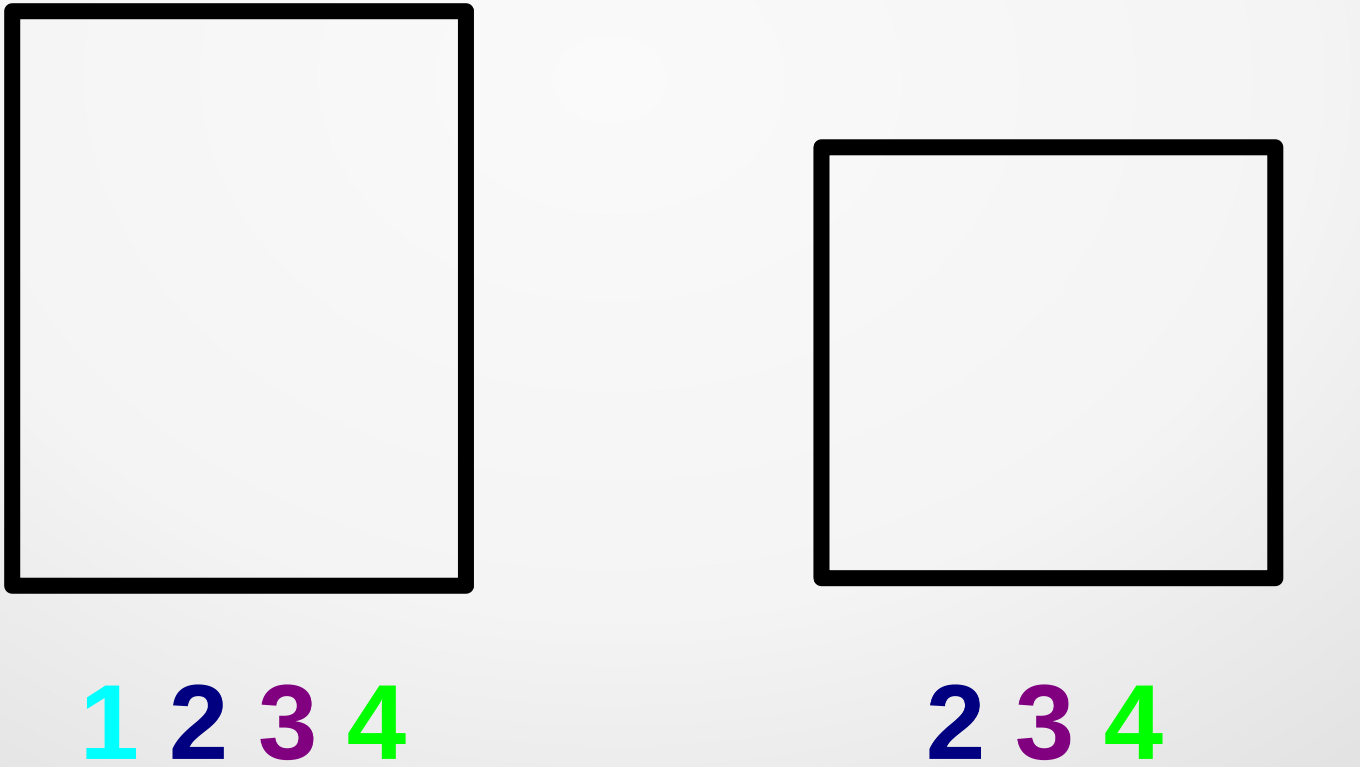
Shapley Values

If we want to find the contribution of ‘player 1’ we look at the value (pay out) in coalitions both with and without them



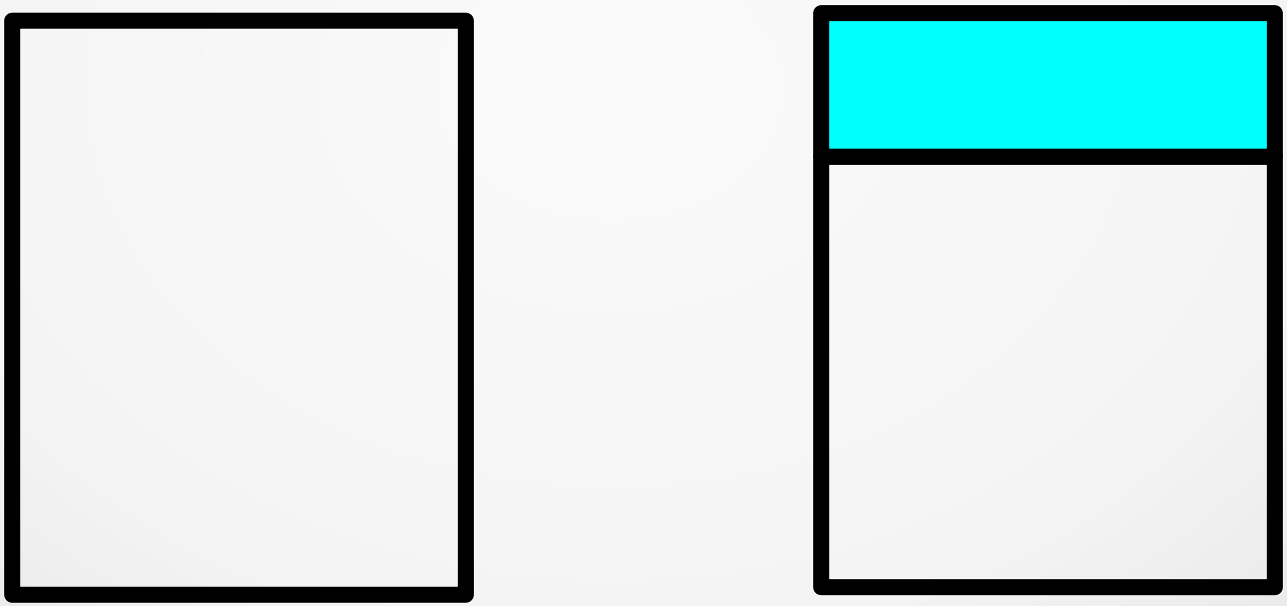
Shapley Values

Next, we'll check out the respective values of the two coalitions and compare the difference between them



Shapley Values

The difference is the marginal contribution of that player



1 2 3 4 2 3 4

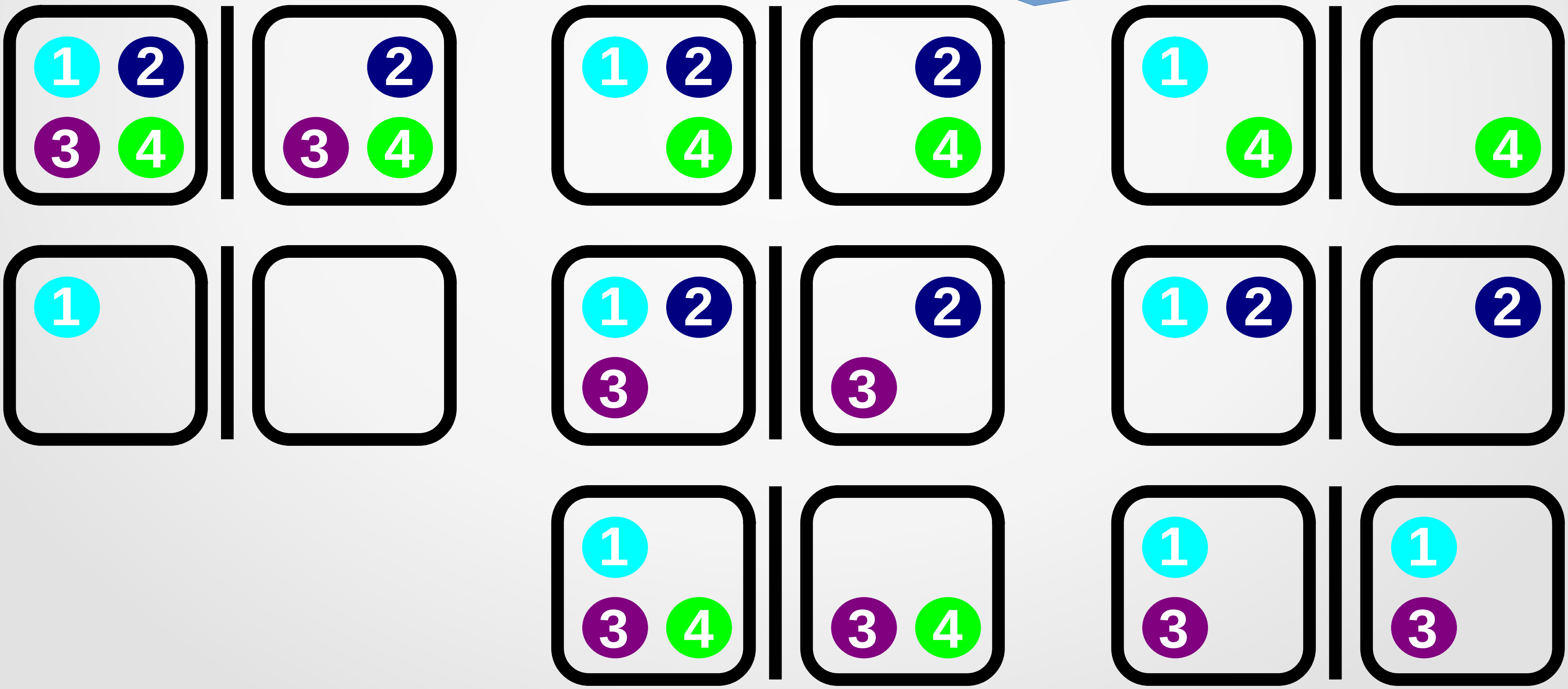
Shapley Values

$$= V_{1\,2\,3\,4} - V_{2\,3\,4} =$$

Marginal Contribution of Member 1 to C(oalition)

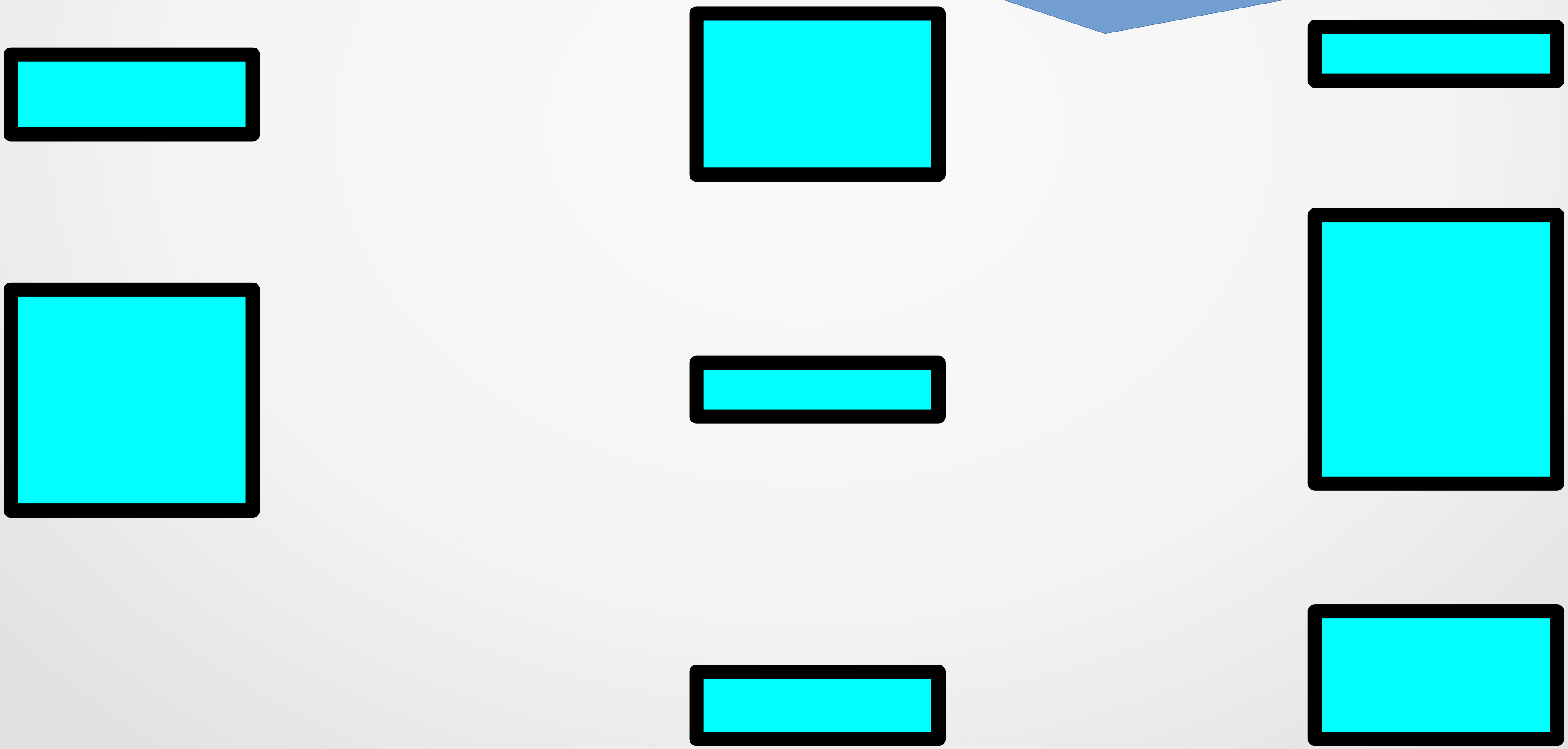
Shapley Values

Now enumerate all such pairs of coalitions i.e., all pairs of coalitions that only differ based on whether or not Player 1 is included



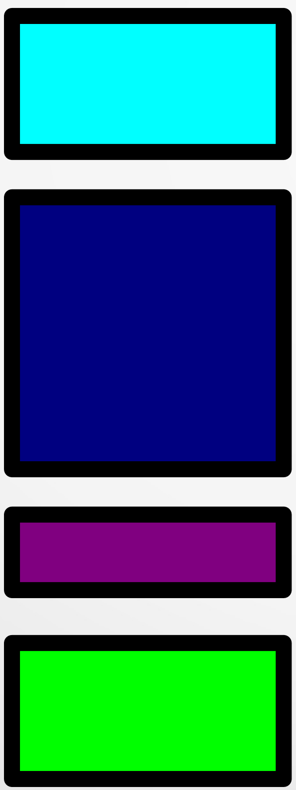
Shapley Values

Next, we look at the marginal contributions for each...



Shapley Values

The mean marginal contribution for that player is the Shapley Value



$= \varphi_1$

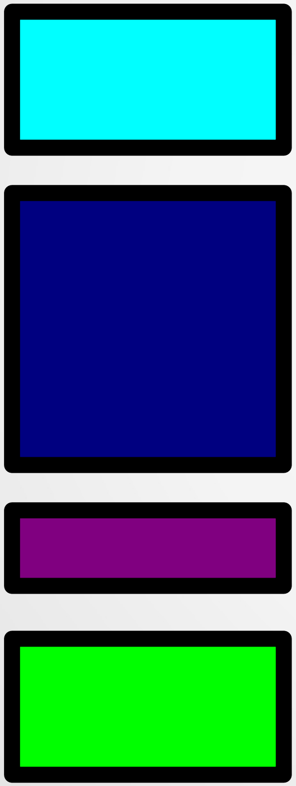
$= \varphi_2$

$= \varphi_3$

$= \varphi_4$

We can do that for each player in the coalition

Shapley Values



$= \varphi_1$

$= \varphi_2$

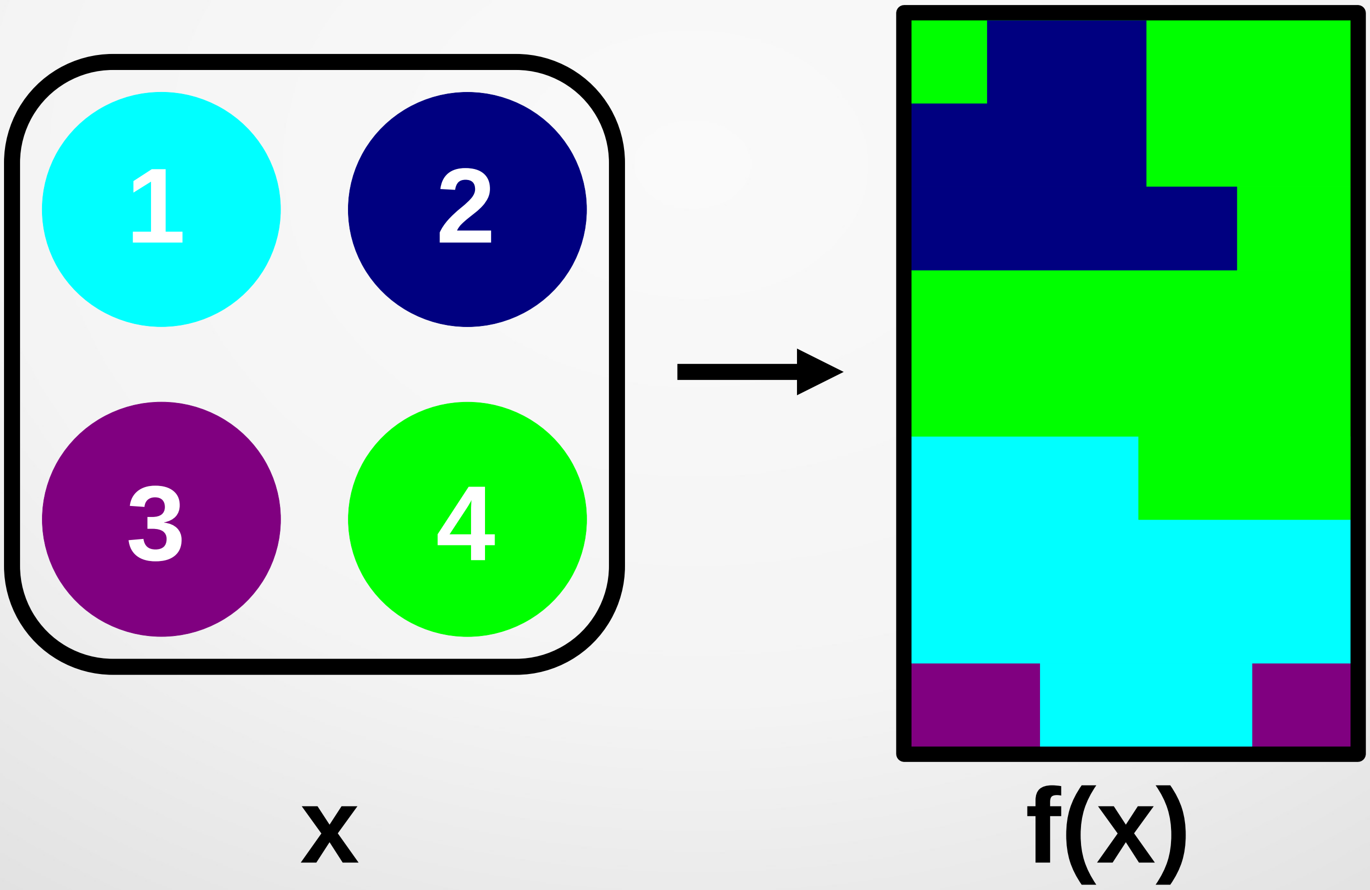
$= \varphi_3$

$= \varphi_4$

The Shapley Value is the average amount of contribution that a particular player makes to the coalition value (C) .

Introducing SHAP by Lundberg and Lee

The SHAP library reframes the Shapley Value problem
From: *“how do players contribute to a coalition value”*
To: **“how do individual features contribute to a model’s outputs”**



Shapley **Additive** Explanations

Local Accuracy

The simplified model created should yield roughly similar results

Shapley **Additive** Explanations

Missingness

If a feature if excluded from a model then it’s attribution must be zero – the only thing that can impact the output is the inclusion of a feature (not the exclusion)

Shapley **Additive** Explanations

Consistency

If feature contribution changes, the feature effect cannot change in the opposite direction
i.e., if there is a new model and a specific feature has a more positive contribution that the original, the attribution in the new explanatory model cannot decrease

Shapley **Additive** Explanations

Local Accuracy

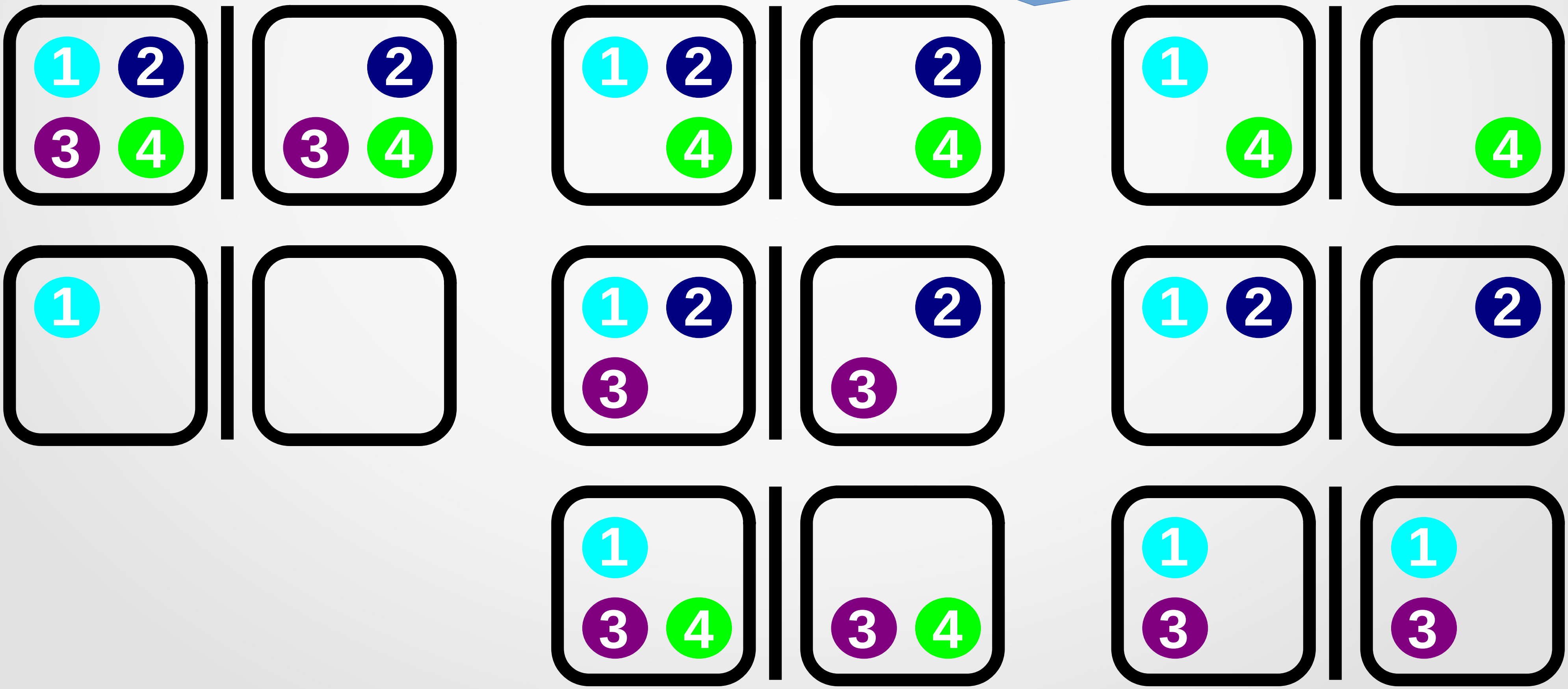
Missingness

Consistency

Only SHAP satisfies all three (whereas LIME would satisfy local accuracy)

The Problem...

Computing Shapley values means you'd have to sample the coalition values for each possible feature permutation, which in a model explainability setting means you'd have to evaluation the model that many times...



SHAP

4 Features: 64 total coalitions to sample

32 Features: 17.1 billion

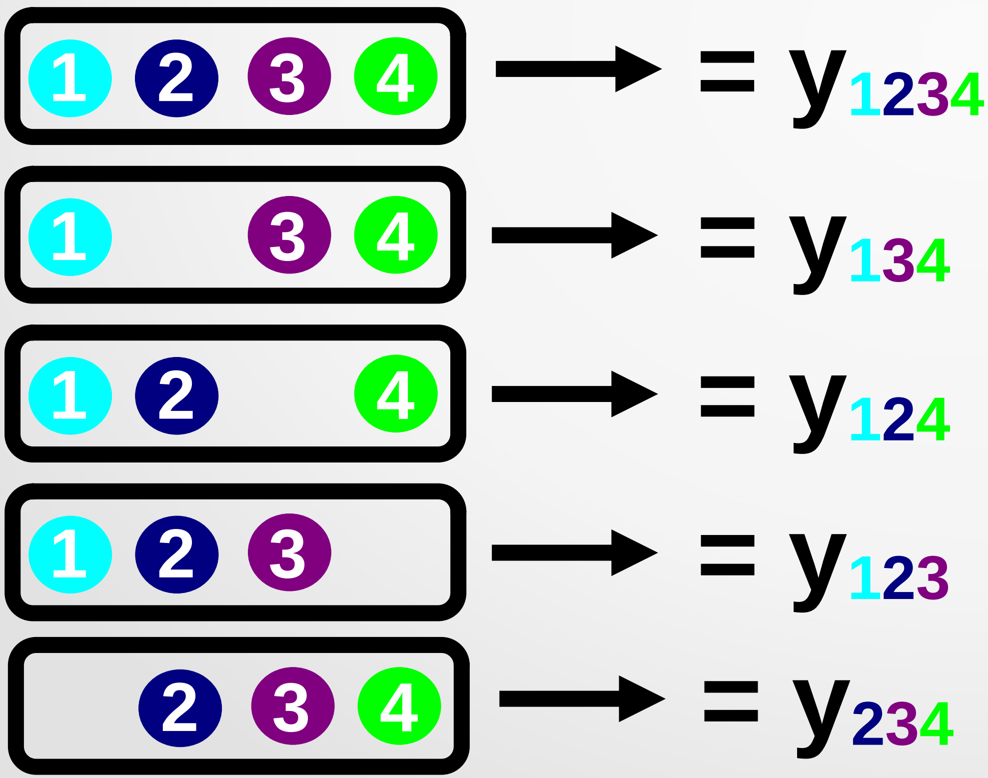
Shapley Kernel

Developed as a means of approximating Shapley vales through much fewer samples.

SHAP: Shapley Kernel

Samples are passed through the model of the various permutations of the data point (i.e. row) we want an explanation for….

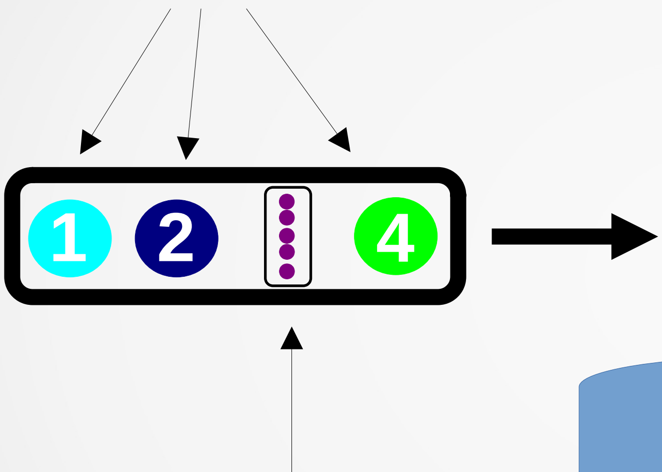
Most ML models don't allow for a feature to simply be omitted so a background dataset is defined. This background dataset contains representative data points that the model was originally trained over



Omitted feature(s) are then replaced with values from the background dataset while holding the other fatures fixed to original values...

SHAP: Shapley Kernel

Included features (fixed)



Average for all predictions made using data from background dataset

Values from background dataset

This process is then repeated for all permutations (as shown on previous slide)

These means are then used to create a weighted linear regression model – with the coefficients of each feature within the linear regression equal to the Shapley value

SHAP

There are other explainers that are optimised for different models. These currently include…

TreeExplainer

GradientExplainer

DeepExplainer

SamplingExplainer

PartitionExplainer

LinearExplainer

And so on….

Check out the documentation here

Lets see it in action!

Other tools

The What-If Tool try it live on Co-lab HERE
IBM AIX 360