Interpretable ML

Local Model-Agnostic Method
SHAP

Date: 26 Nov 2021

Nikhil Verma (lih.verma@gmail.com)

Agenda

- Interpretability
- Methods for model interpretation
- Game Theory
- Shapley values
- SHAP

Interpretability

- If a machine learning model performs well

Classification Accuracy

- **Why do we not just trust the model** and
- Ignore **why** it made a certain decision?

Interpretability

Why its important

- The need for interpretability arises from an incompleteness in problem formalization.

Interpretability

Why its important

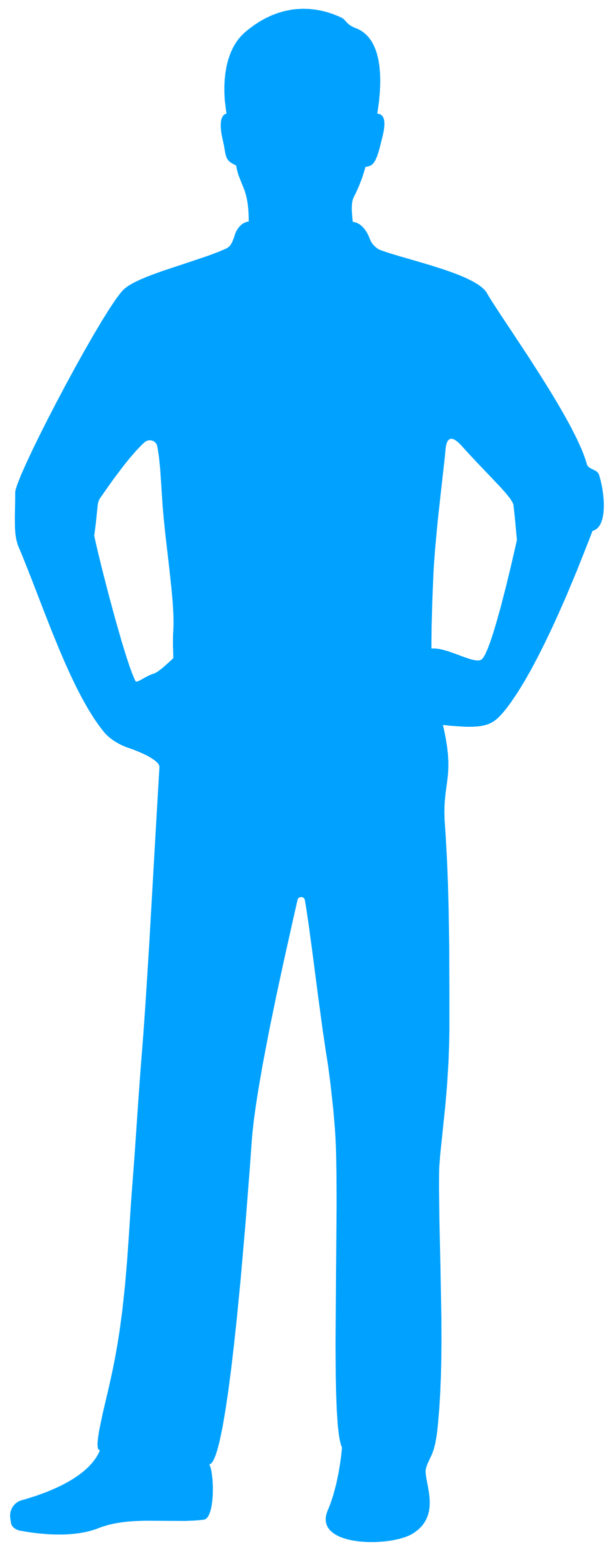
- The need for interpretability arises from an incompleteness in problem formalization.

- For certain problems or tasks it is not enough to get the prediction (**what**)

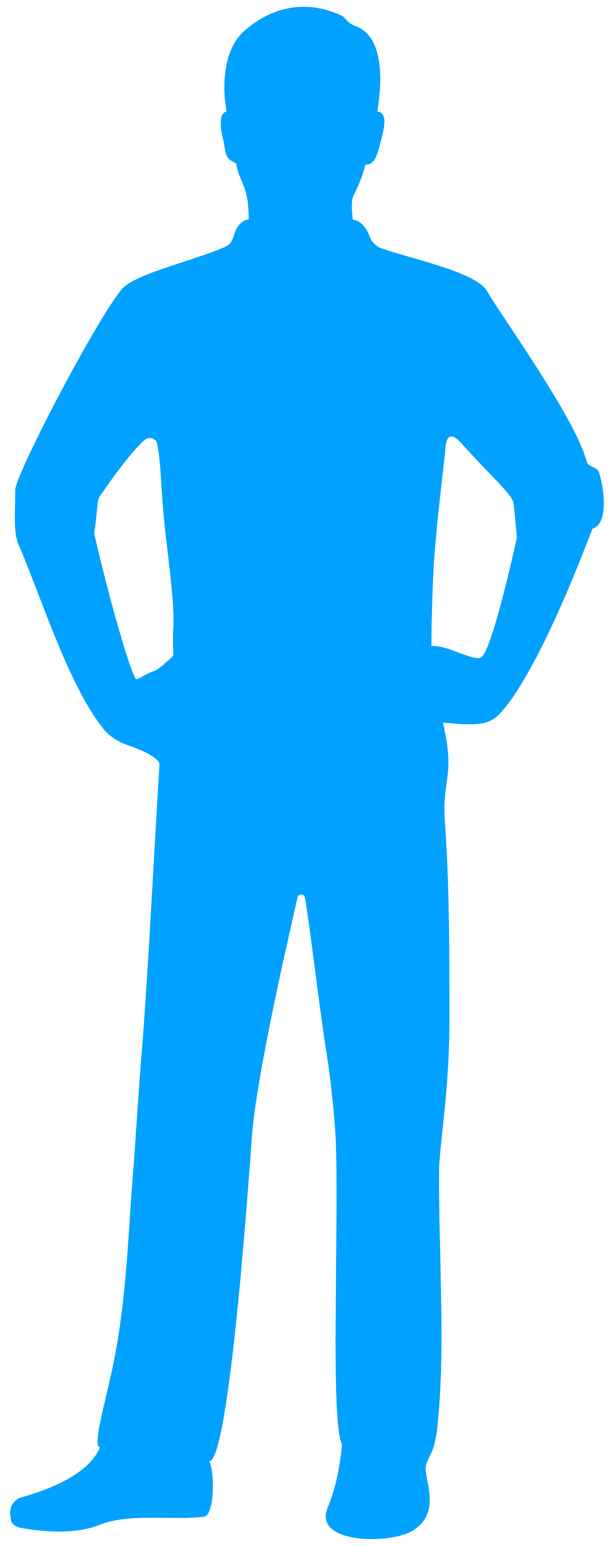
- The model must also explain how it came to the prediction (**why**)

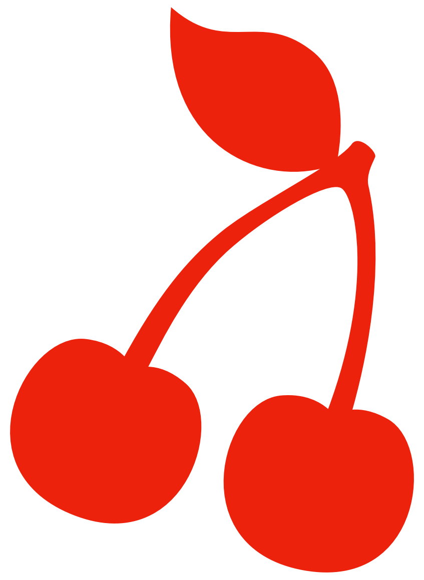
  - because a correct prediction only partially solves your original problem

Human Curiosity

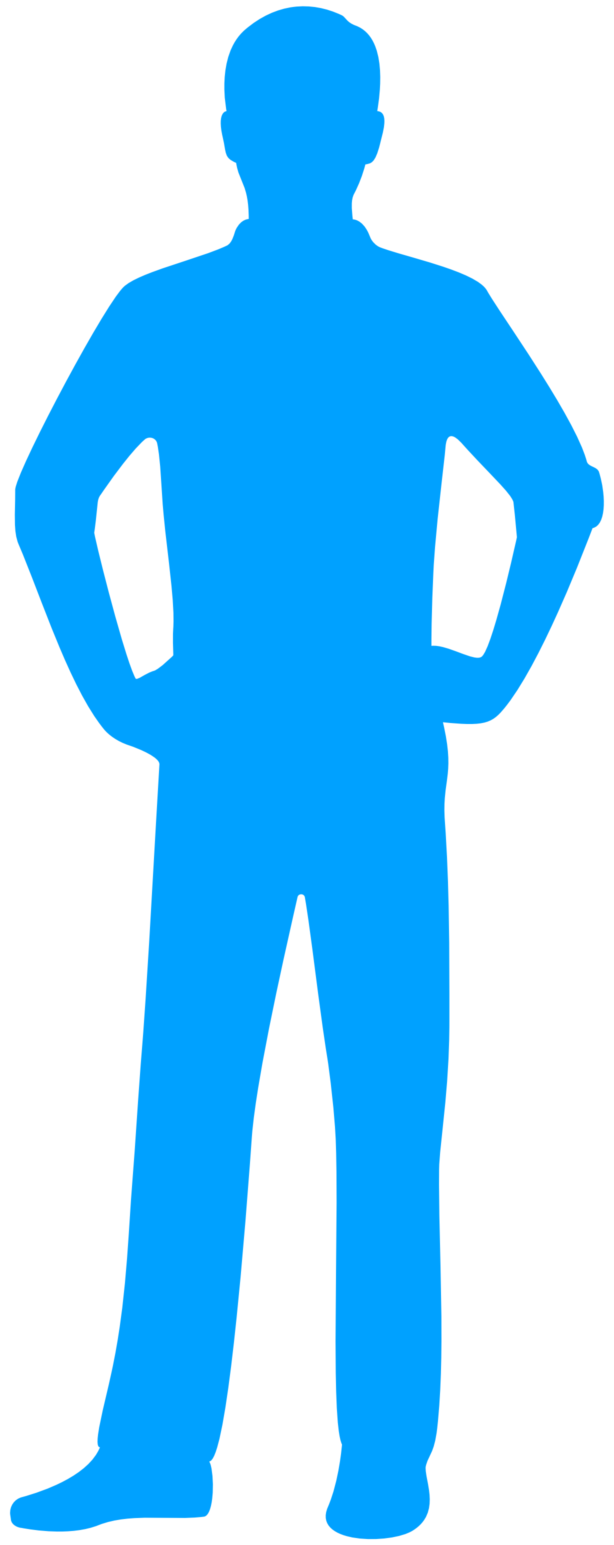


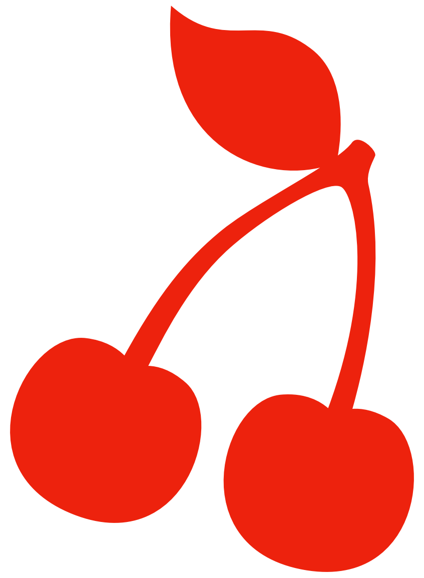
Human Curiosity





Human Curiosity





- Closely related to learning is the human desire to **find meaning in the world.**
- We want to harmonize contradictions or inconsistencies between elements of our knowledge structures.

Explanations

- Explanations are used to **manage social interactions.**

Explanations

- Explanations are used to **manage social interactions.**
- By creating a shared meaning of something, the explainer influences the recipient of the explanation
    - Actions
    - Emotions
    - Beliefs

Explanations

- Explanations are used to **manage social interactions**.
- By creating a shared meaning of something, the explainer influences the recipient of the explanation
  - actions
  - emotions
  - beliefs
- Machine learning models can only be **debugged and audited** when they can be interpreted.

Interpretability

Which ML techniques to use?

Interpretability

Which ML techniques to use?

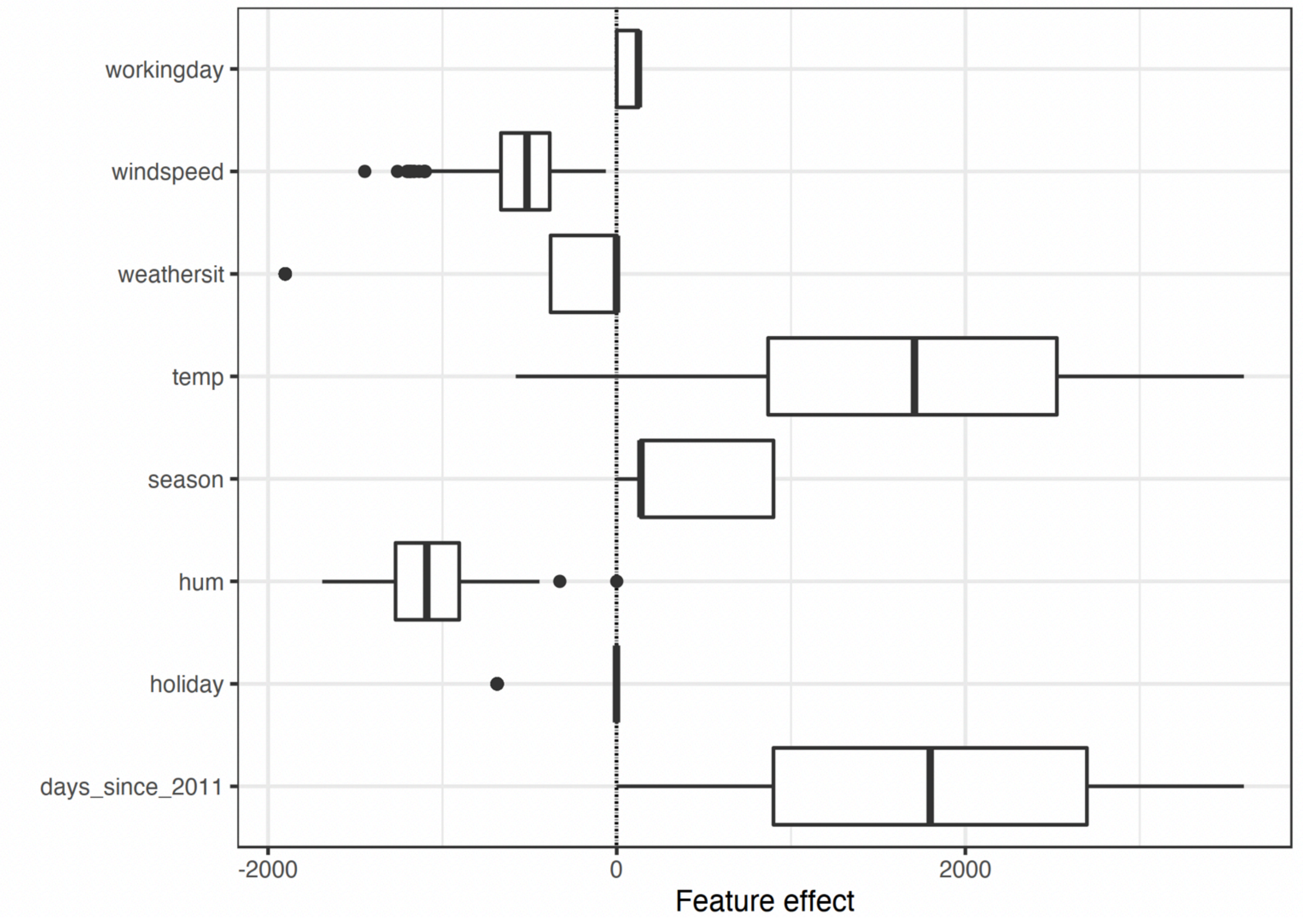
- Use only interpretable models

Interpretability

Which ML techniques to use?

- Use only interpretable models
  - predictive performance is lost compared to other machine learning models
  - you limit yourself to one type of model

Linear Regression



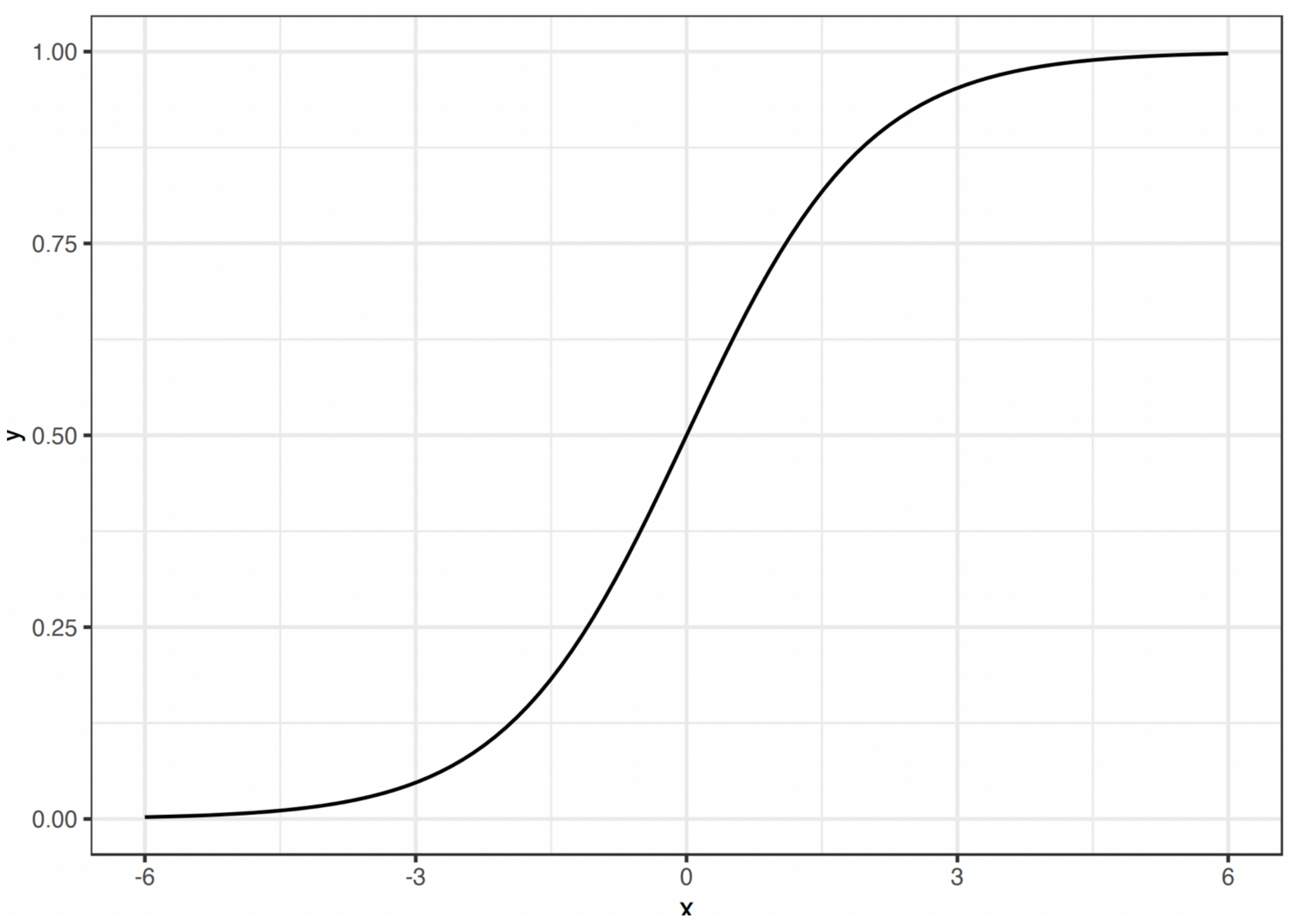
$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

Logistic Regression

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

$$ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

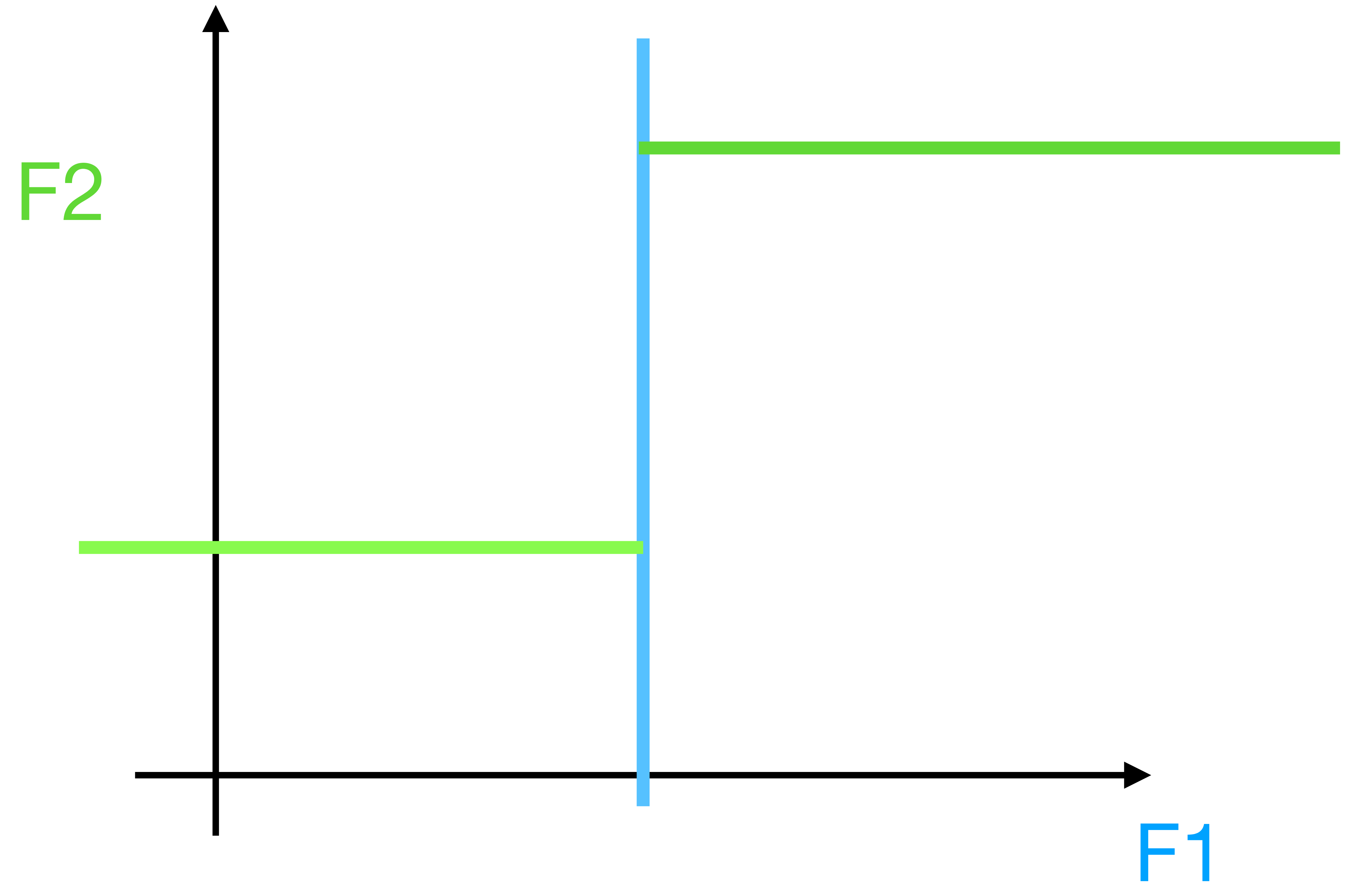


Naive Bayes

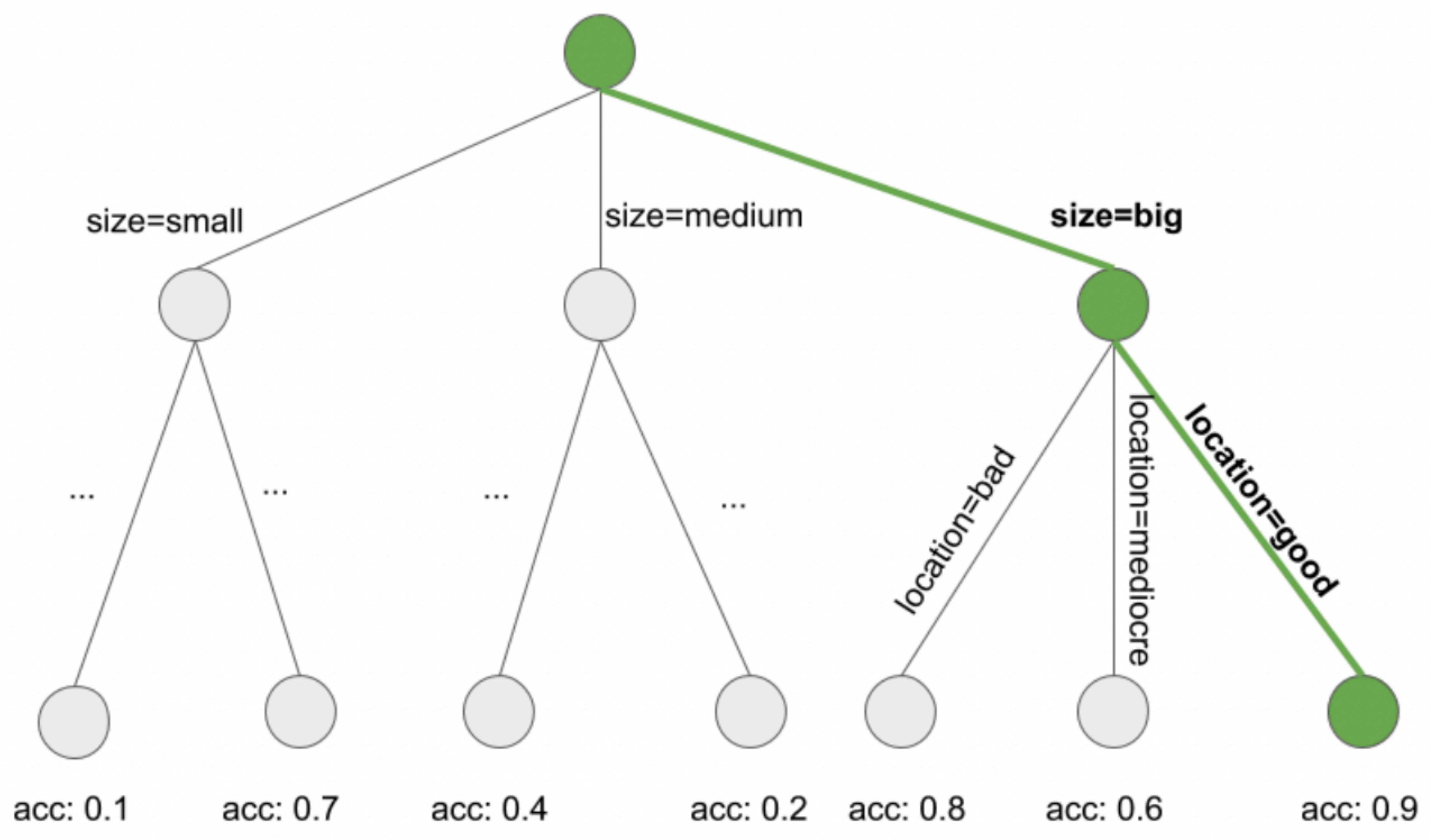
$$P(C_k|x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

Class Conditional Independence

Decision Tree



Decision Tree



Interpretability

Which ML techniques to use?

- Use only interpretable models
  - predictive performance is lost compared to other machine learning models
  - you limit yourself to one type of model

ML model	Simple	Complex
Accurate	X	✓
Interpretable	✓	X

Interpretability

Which ML techniques to use?

- Use only interpretable models
  - predictive performance is lost compared to other machine learning models
  - you limit yourself to one type of model
- The other alternative is to use model-specific interpretation methods
  - It also binds you to one model type
  - It will be difficult to switch to something else

Methods

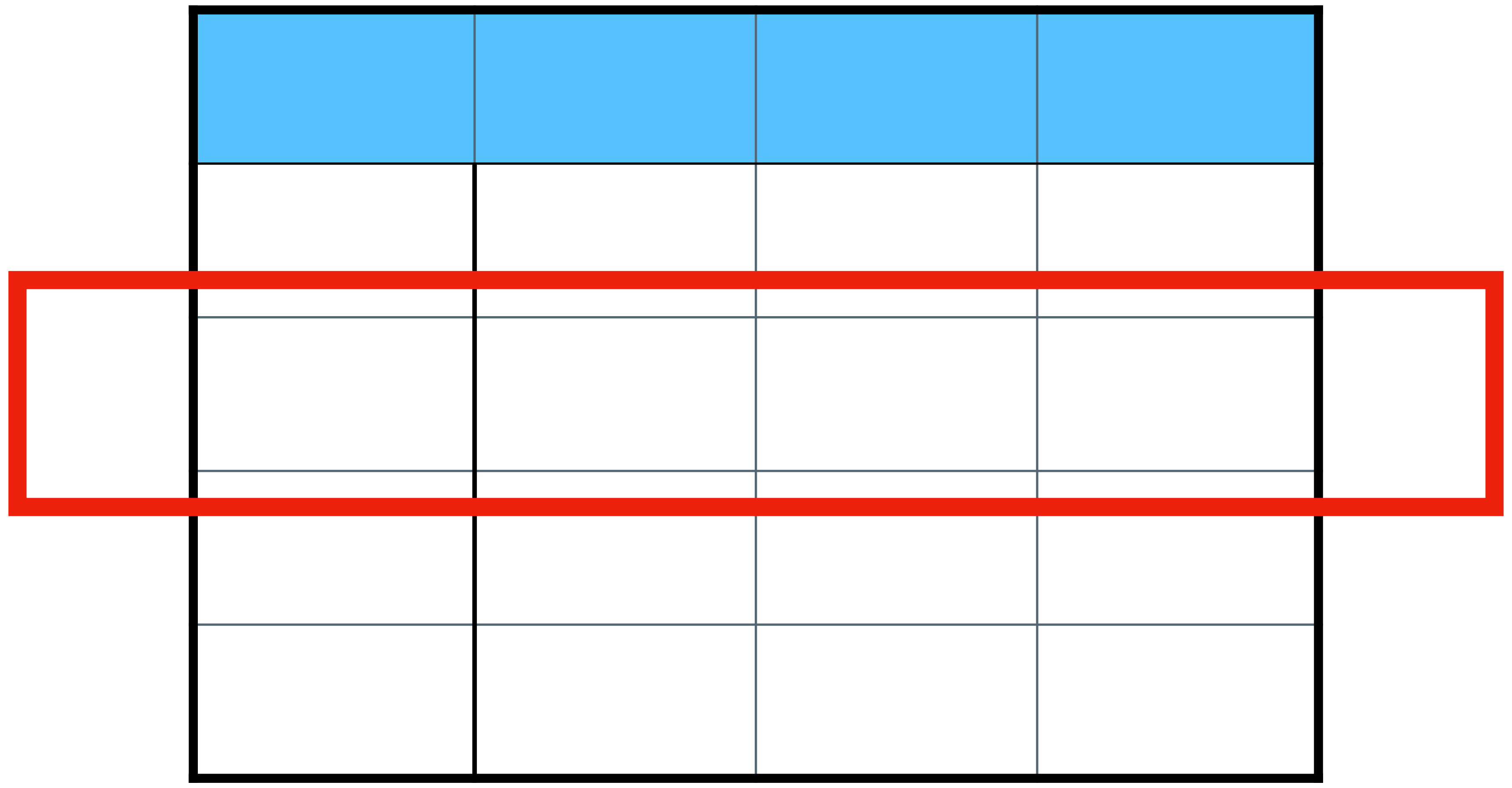
Example-based explanation method

- Select particular instances of the dataset
  - to explain the behaviour of machine learning models
  - to explain the underlying data distribution

Methods

Example-based explanation method

- Select particular instances of the dataset
  - to explain the behaviour of machine learning models
  - to explain the underlying data distribution



Methods

Example-based explanation method

- It help humans construct **mental models**
  - Along with the data the ML model has been trained on
- It especially helps to understand complex data distributions

Methods

Example-based explanation method

- It help humans construct **mental models**
  - Along with the data the ML model has been trained on
- It especially helps to understand complex data distributions
- But what do I mean by example-based explanations?

  - A physician sees a patient with an unusual cough and a mild fever
  - The patient's symptoms remind her of another patient she had years ago with similar symptoms
  - She suspects that her current patient could have the same disease and she takes a blood sample to test for some specific disease

Methods

Example-based explanation method

- It help humans construct **mental models**
  - Along with the data the ML model has been trained on
- It especially helps to understand complex data distributions
- But what do I mean by example-based explanations?

  - A physician sees a patient with an unusual cough and a mild fever
  - The patient's symptoms remind her of another patient she had years ago with similar symptoms
  - She suspects that her current patient could have the same disease and she takes a blood sample to test for some specific disease

KNN

Methods

Model-Agnostic Methods

Methods

Model-Agnostic Methods

Separating the explanations
from
the machine learning model

Model-agnostic

Advantage

- **Flexibility**
  - ML developers are free to use any ML model they like when the interpretation methods can be applied to any model
  - Anything can be build on an interpretation of a machine learning model
    - a graphic
    - user interface
  - becomes independent of the underlying machine learning model

Flexibility

Desirable aspects of a model-agnostic explanation system

Model flexibility

Explanation flexibility

Representation flexibility

Model Flexibility

Desirable aspects of a model-agnostic explanation system

Methods can work with any ML model - ex Random forest or Neural networks

Explanation flexibility

Representation flexibility

Explanation Flexibility

Desirable aspects of a model-agnostic explanation system

Model flexibility

Not limited to certain form of explanations ex Linear equation, graph, UI

Representation flexibility

Representation Flexibility

Desirable aspects of a model-agnostic explanation system

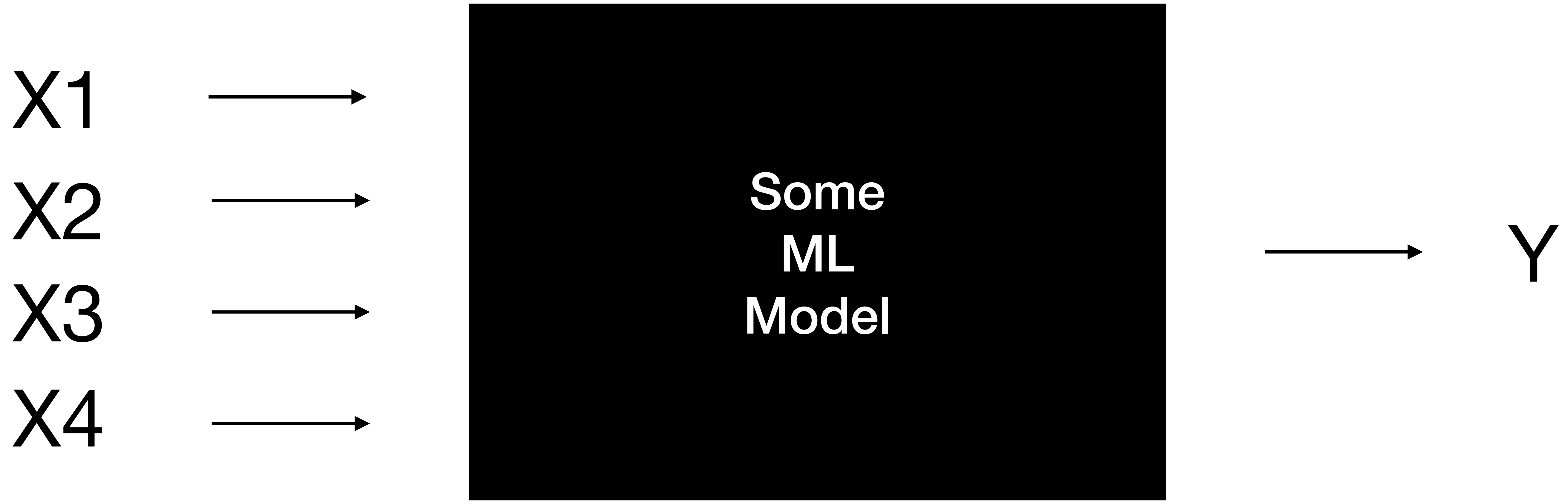
Model flexibility

Explanation flexibility

Different feature representation possible as the model being explained ex not W2V but words

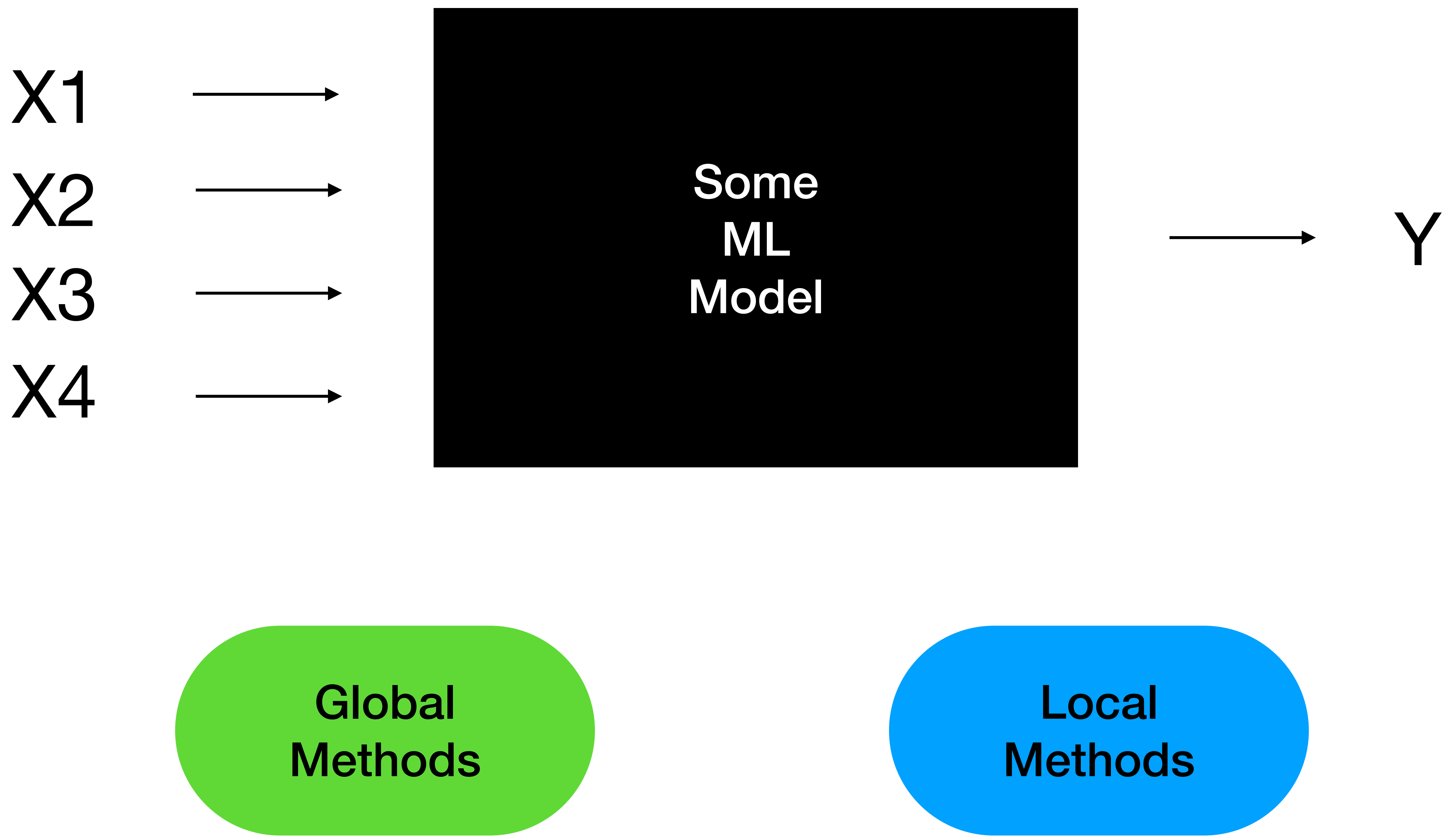
Models

Black Boxes



Models

Black Boxes



Average behaviour of Model

Explain individual predictions

Compare

Which methods are more useful?

Compare

Which methods are more useful?

- The example-based methods explain a model by selecting instances of the dataset and not by creating summaries of features.

Compare

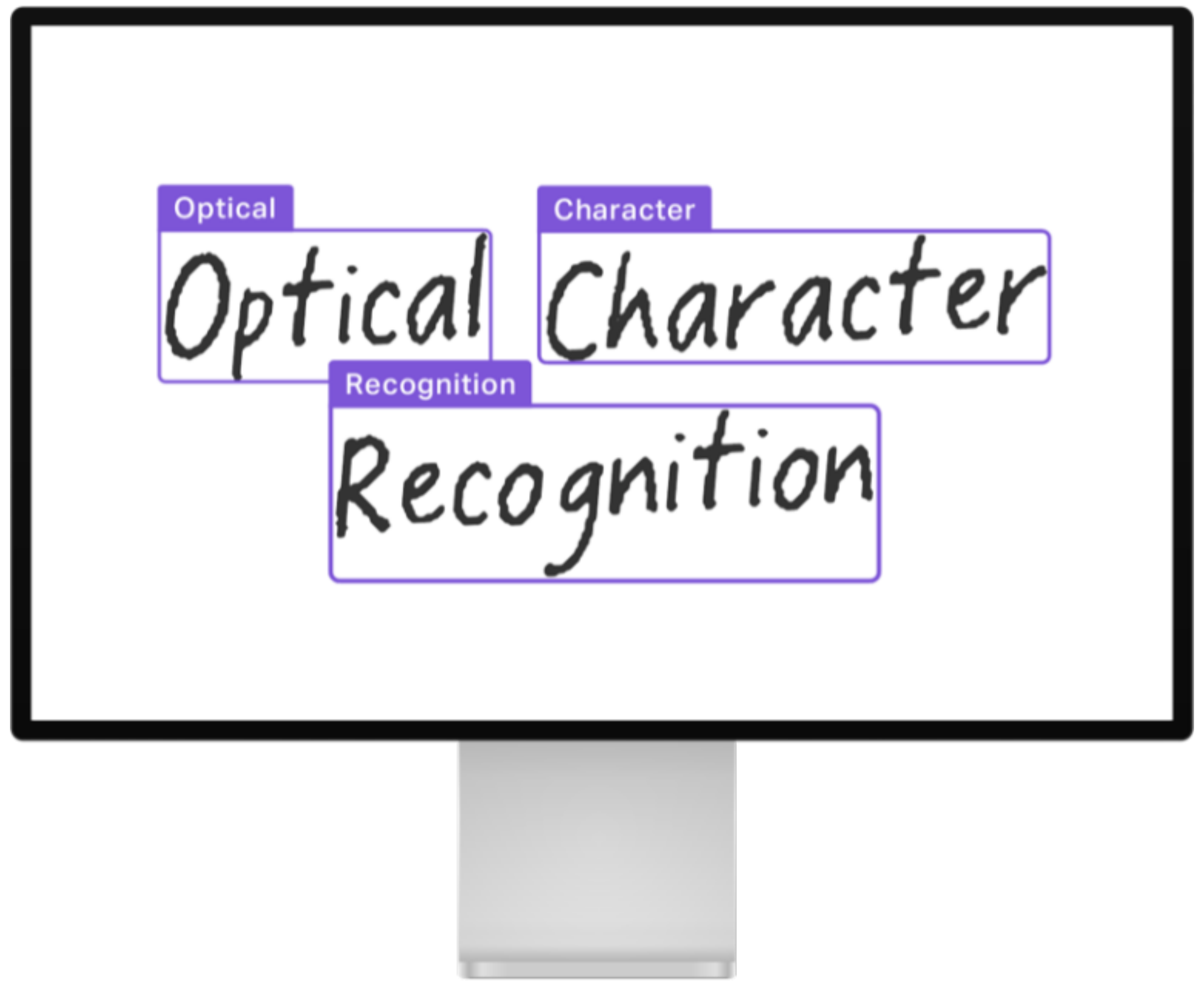
Which methods are more useful?

- The example-based methods explain a model by selecting instances of the dataset and not by creating summaries of features.
- Example-based explanations only make sense if we can represent an instance of the data in a humanly understandable way.
- Ex: This works well for images, because we can view them directly.

Compare

Which methods are more useful?

- The example-based methods explain a model by selecting instances of the dataset and not by creating summaries of features.
- Example-based explanations only make sense if we can represent an instance of the data in a humanly understandable way.
- Ex: This works well for images, because we can view them directly.
- In general, example-based methods work well if
  - the feature values of an instance carry more context
  - (data has a structure as images or texts)



Local Methods

Explain individual predictions

- **Local surrogate models (LIME)** explains a prediction by replacing the complex model with a locally interpretable surrogate model.
- **Scoped rules (anchors)** are rules that describe which feature values anchor a prediction
- **Counterfactual explanations** explain a prediction by examining which features would need to be changed to achieve a desired prediction.
- **Shapley values** is an attribution method that fairly assigns the prediction to individual features.
- **SHAP** is another computation method for Shapley values.

Method

Local Model-Agnostic Method

Shapley Values
and
SHAP

Zero-Sum game

Mathematical representation of a situation

- An advantage that is won by one of two sides is lost by the other.
- If the total gains of the participants are added up, and the total losses are subtracted, they will sum to zero.

Zero-Sum game

Mathematical representation of a situation

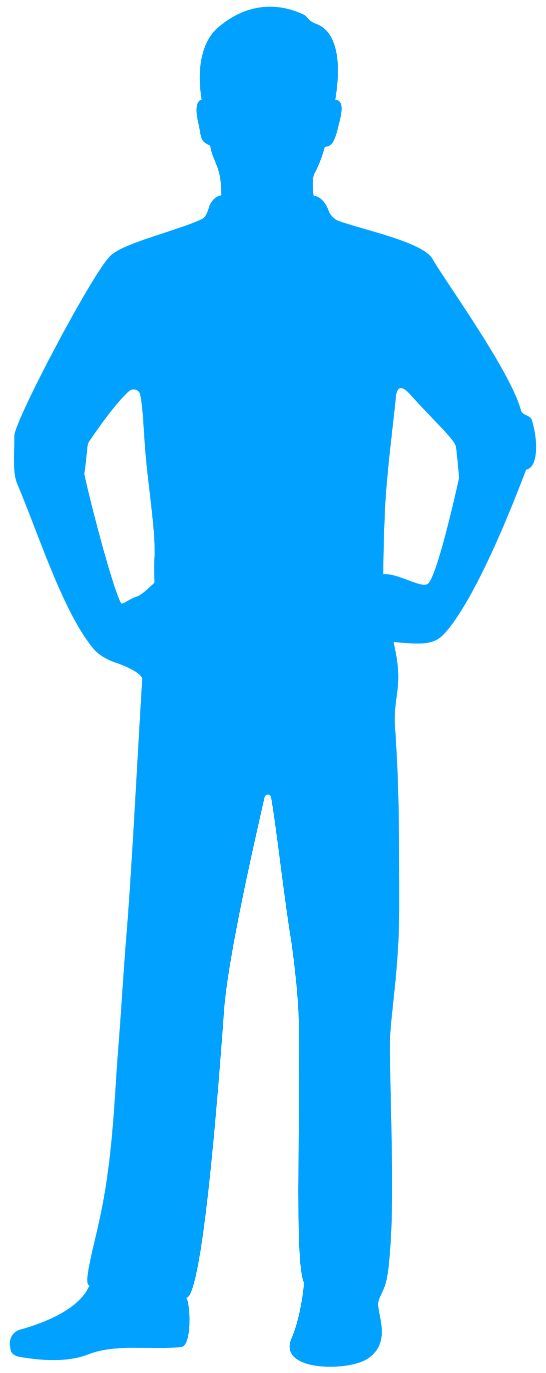
- An advantage that is won by one of two sides is lost by the other.
- If the total gains of the participants are added up, and the total losses are subtracted, they will sum to zero.
- For example
  - Cutting a cake, where taking a more significant piece reduces the amount of cake available for others as much as it increases the amount available for that taker, is a zero-sum game if all participants value each unit of cake equally.
- Other examples of zero-sum games in daily life include games like poker, chess, and bridge where one person gains and another person loses, which results in a zero-net benefit for every player.

Game Theory

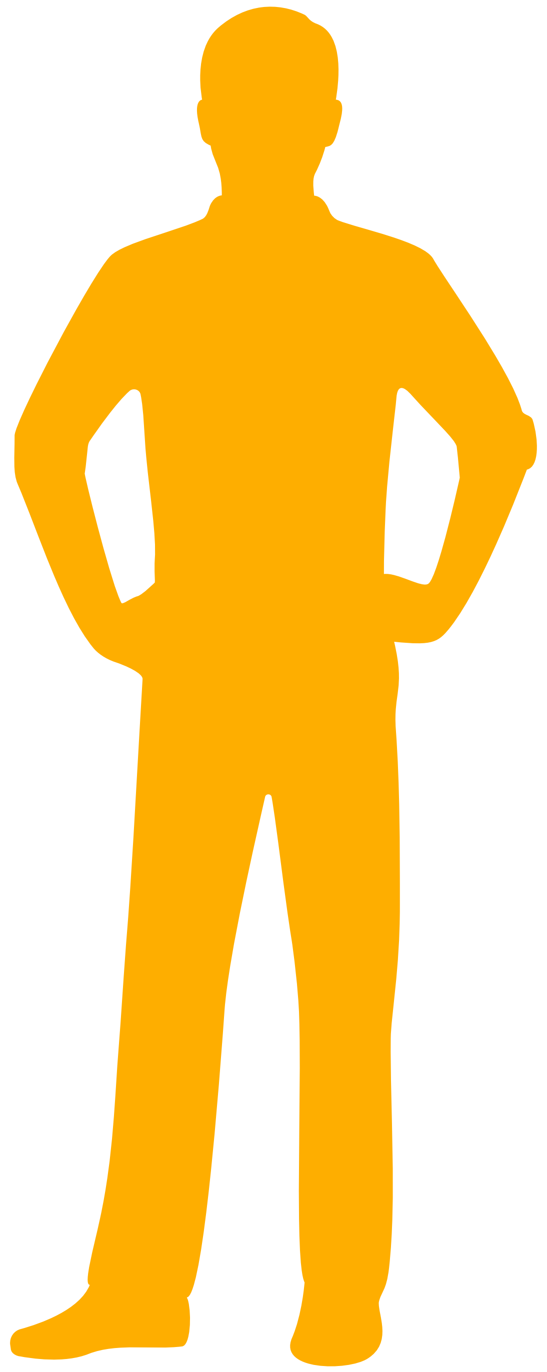
- Cooperative Game
  - is a game with competition between groups of players ("coalitions") due to the possibility of external enforcement of cooperative behaviour
  - e.g. through contract law

Prisoner’s dilemma

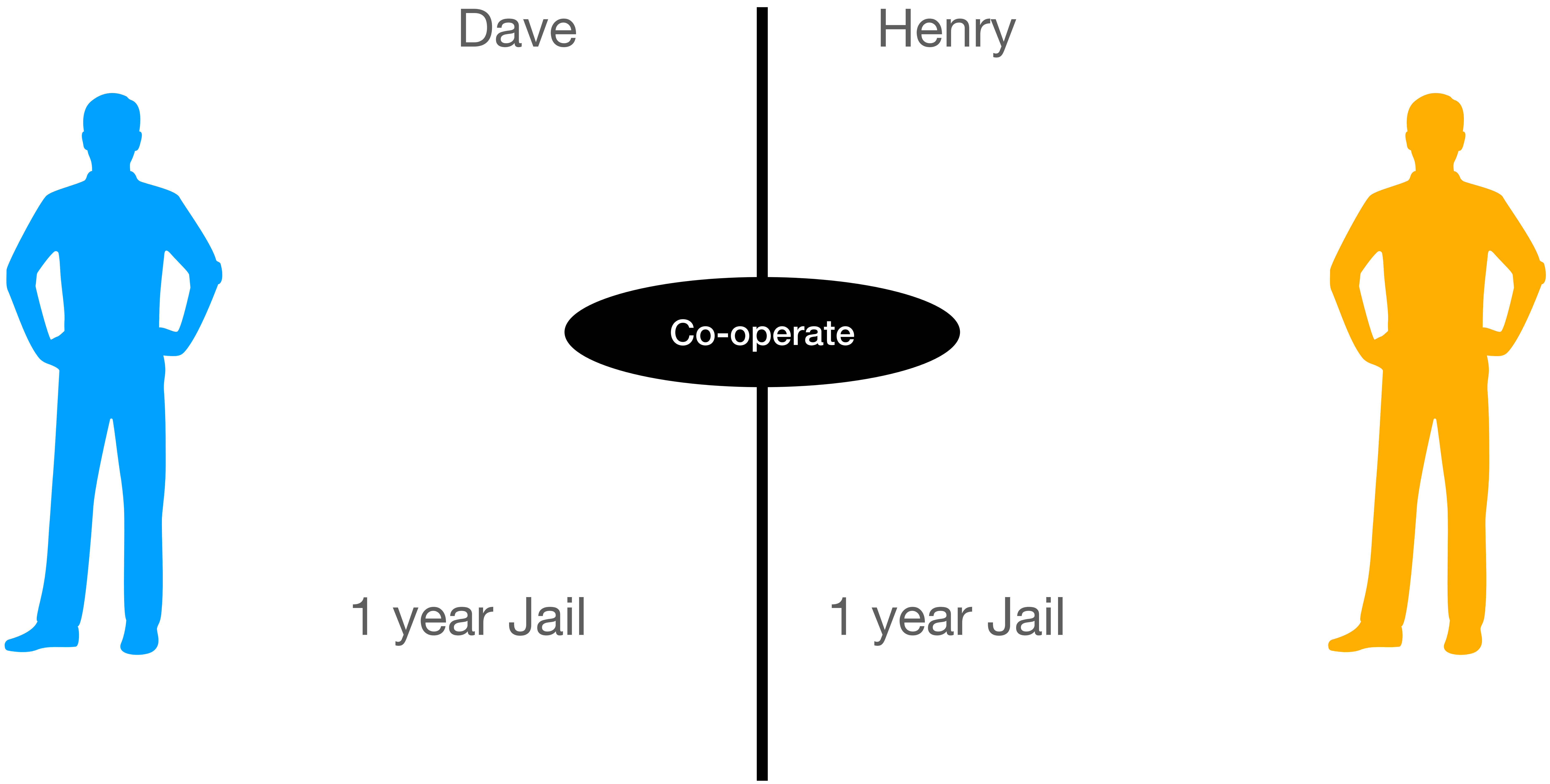
Dave



Henry



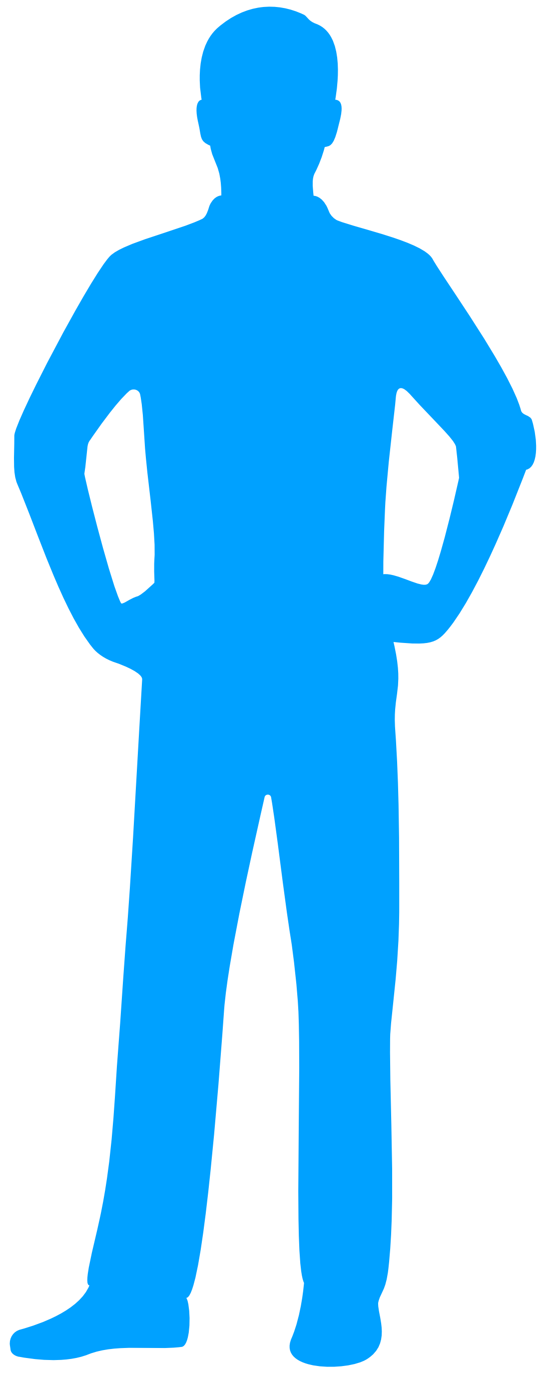
Prisoner’s dilemma



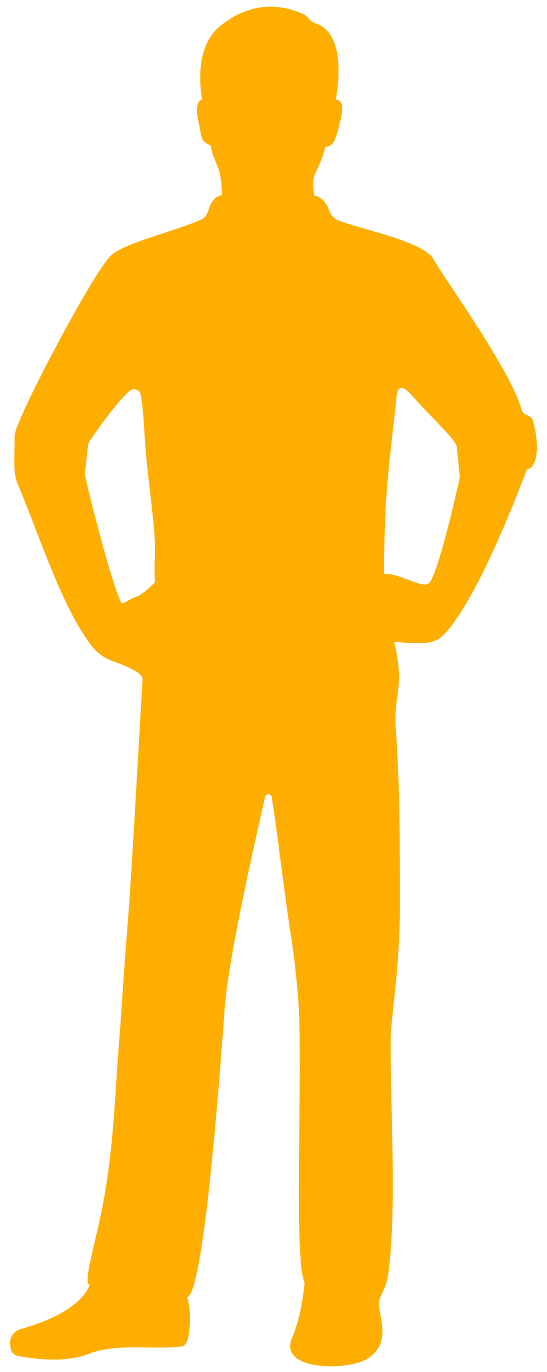
Prisoner’s dilemma

Dave

Henry



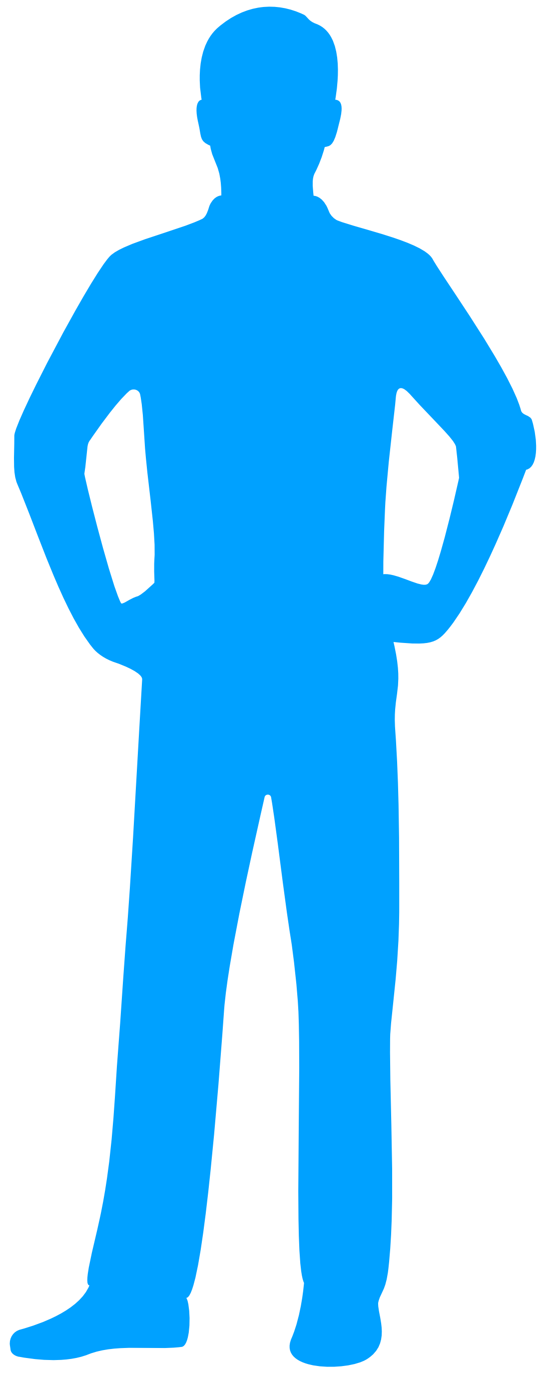
Testifies



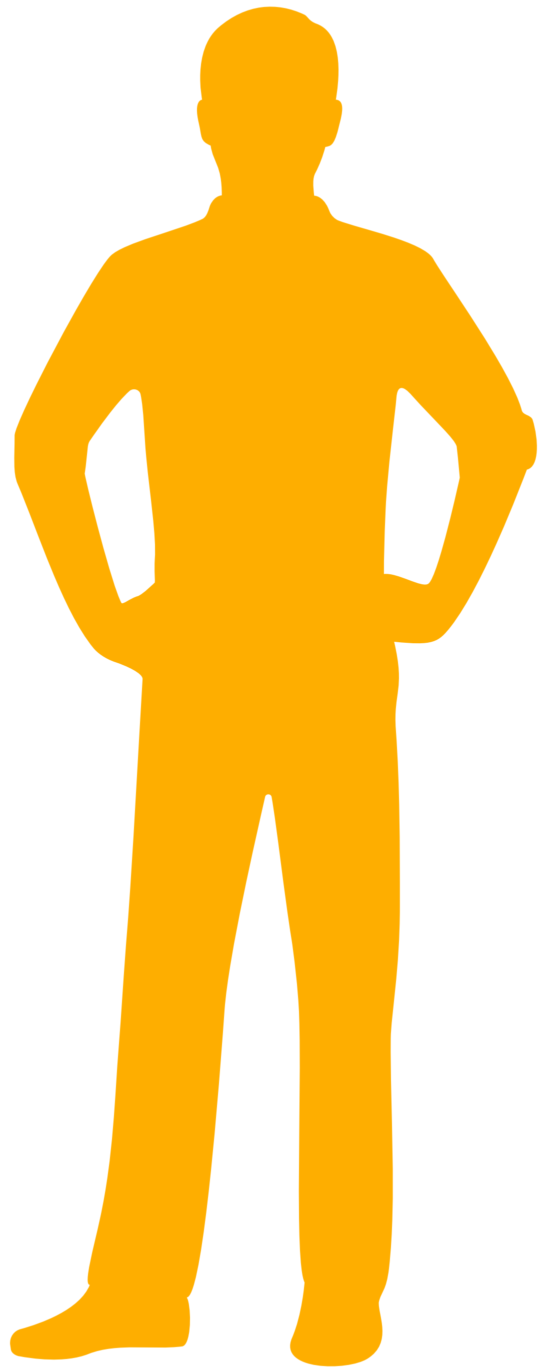
0 year Jail

5 year Jail

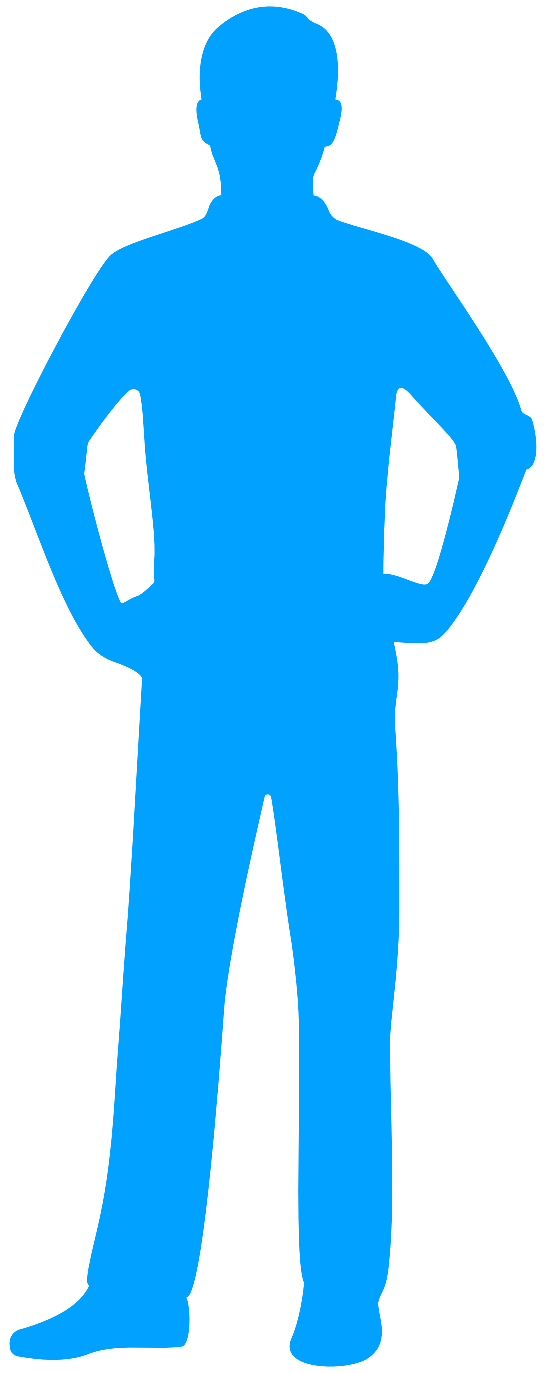
Prisoner's dilemma



Dave	Henry
Testifies	Testifies
2 year Jail	2 year Jail

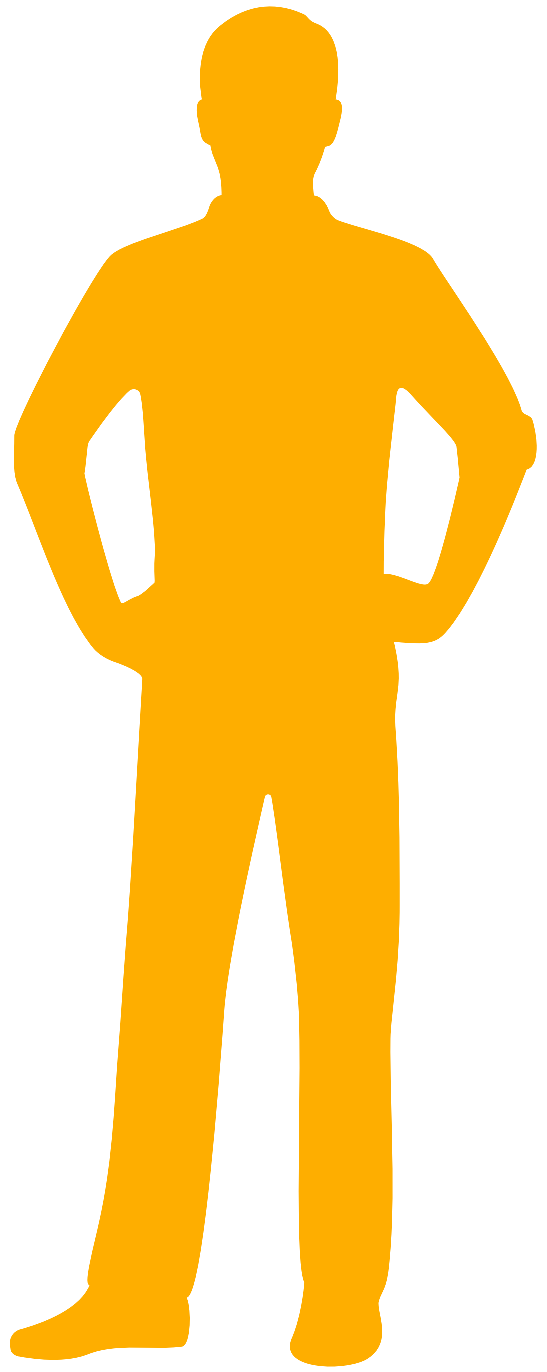


Prisoner's dilemma



Henry

		Confess	Do not Confess
Dave	Confess	2, 2	0, 5
	Do not Confess	5, 0	1, 1



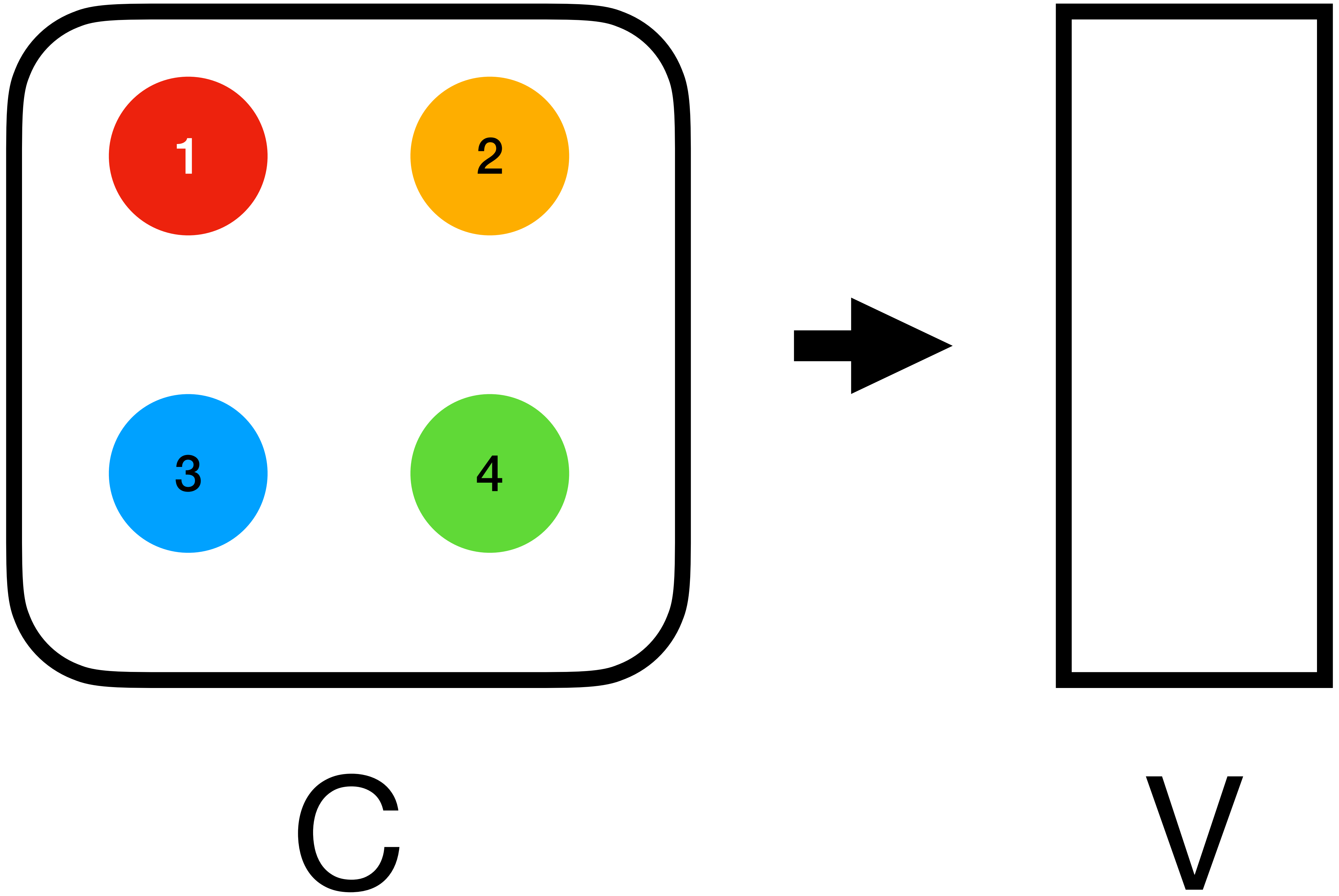
The paradox of the prisoner's dilemma is this:

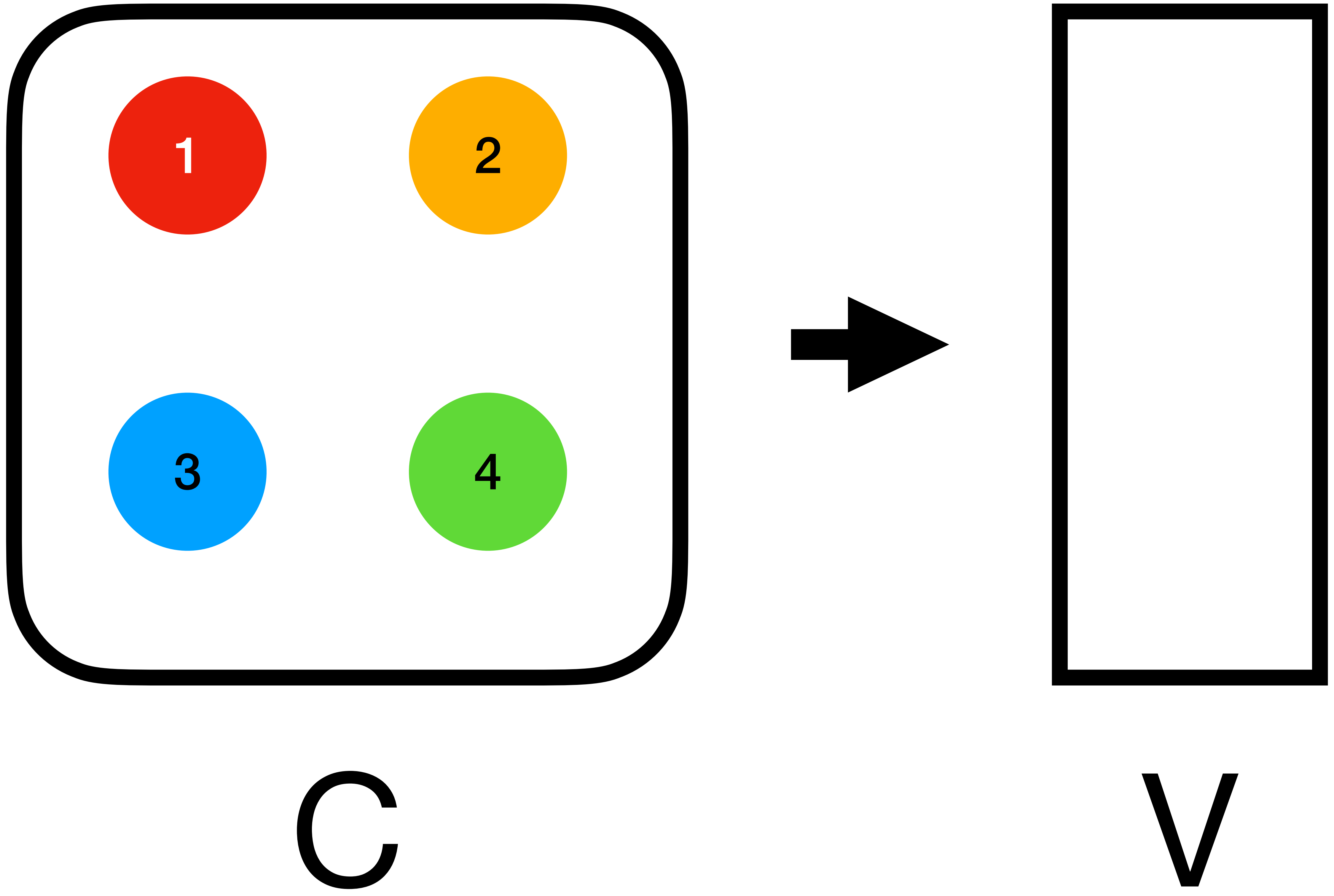
Both robbers can minimise the total jail time that the two of them will do only if they both co-operate and stay silent

Prisoner’s dilemma

- The prisoner’s dilemma presents a situation where two parties, separated and unable to communicate, must each choose between co-operating with the other or not.

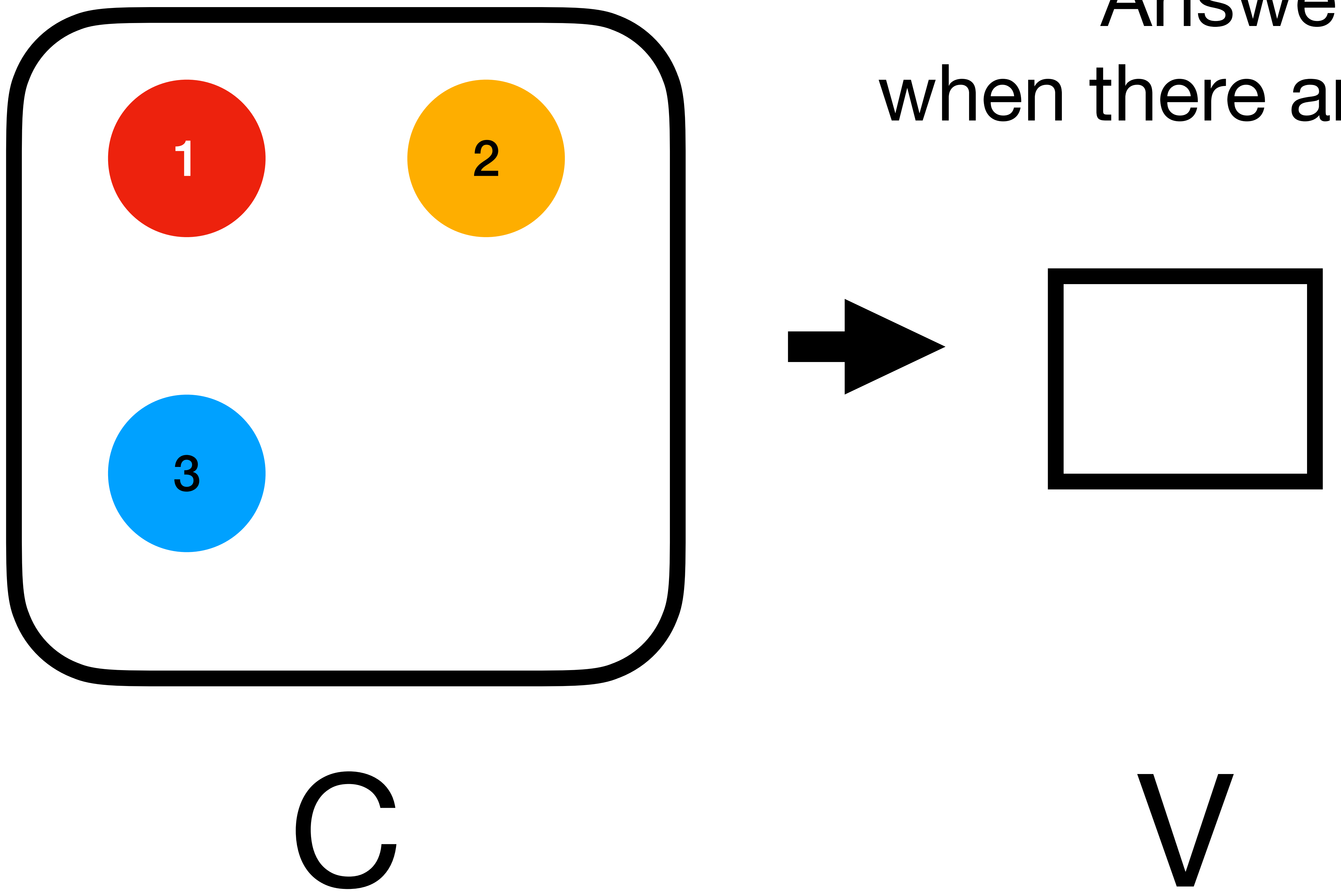
- If we have a coalition C that collaborates to produce a value V
- How much did each individual member contribute to the final value?


1
2
3
4
C
V

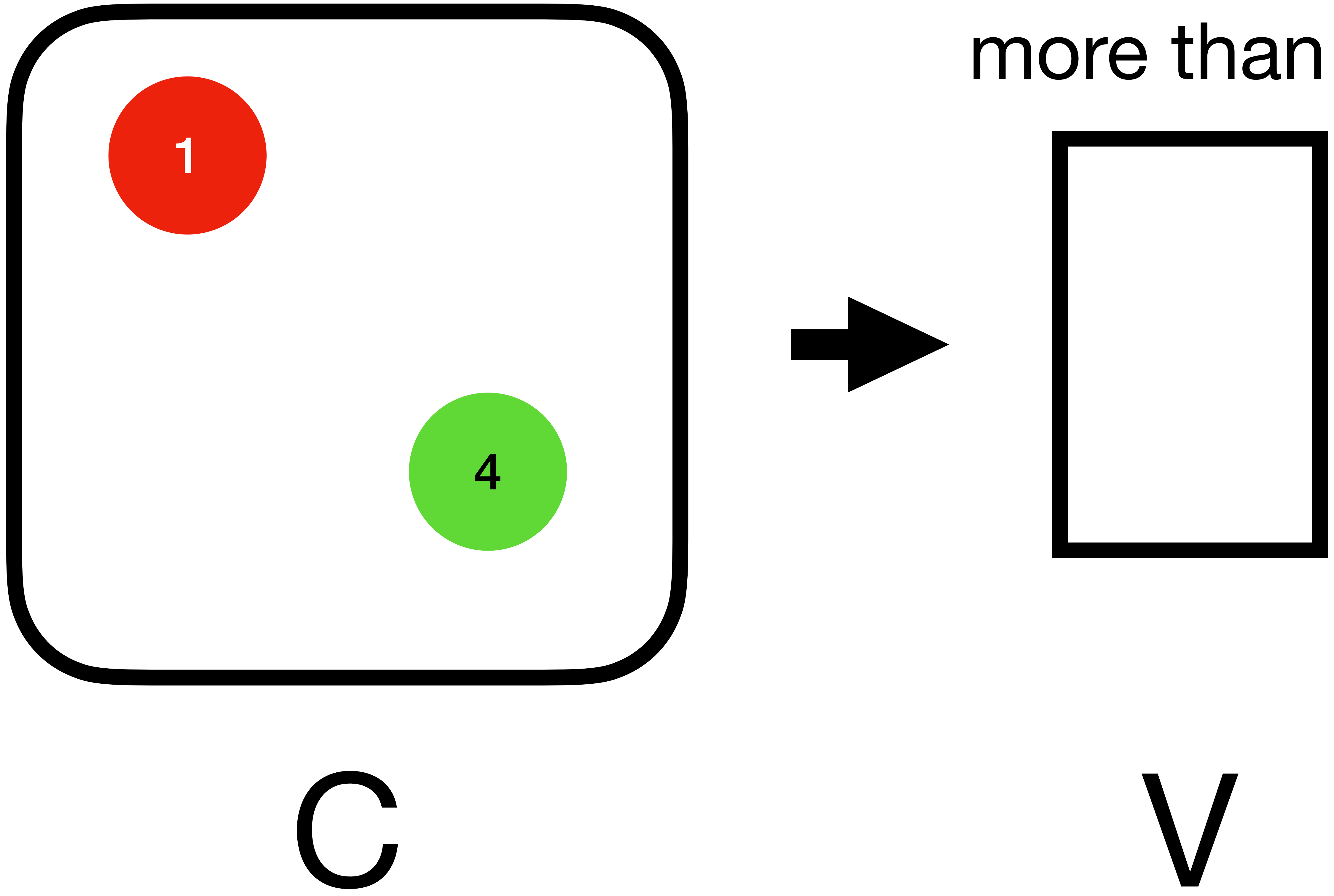


How much should everyone
contribute to pay the bill

Answering it is tricky
when there are interacting effects



Certain permutations cause
Contributors to pay
more than sum of their parts



- Alice, Bob and Celine share a meal:

- Alice, Bob and Celine share a meal:

$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

- Shapley value
- To find fair value for solution, we should take Shapley values into account and should calculate it for each member in the coalition

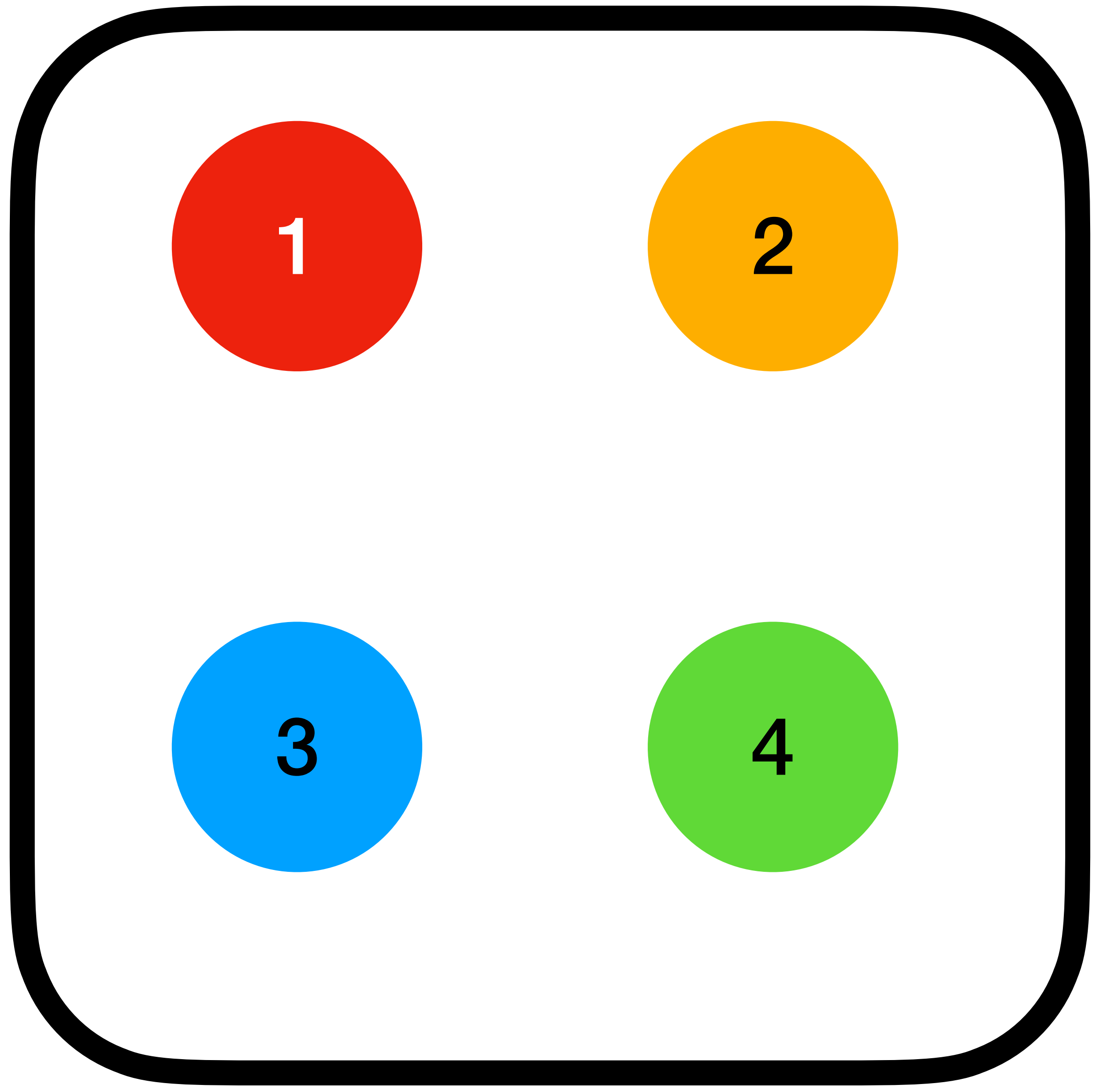
$$\phi_i(G) = \frac{1}{n!} \sum_{\pi \in \Pi_n} \Delta_\pi^G(i)$$

- Alice, Bob and Celine share a meal:

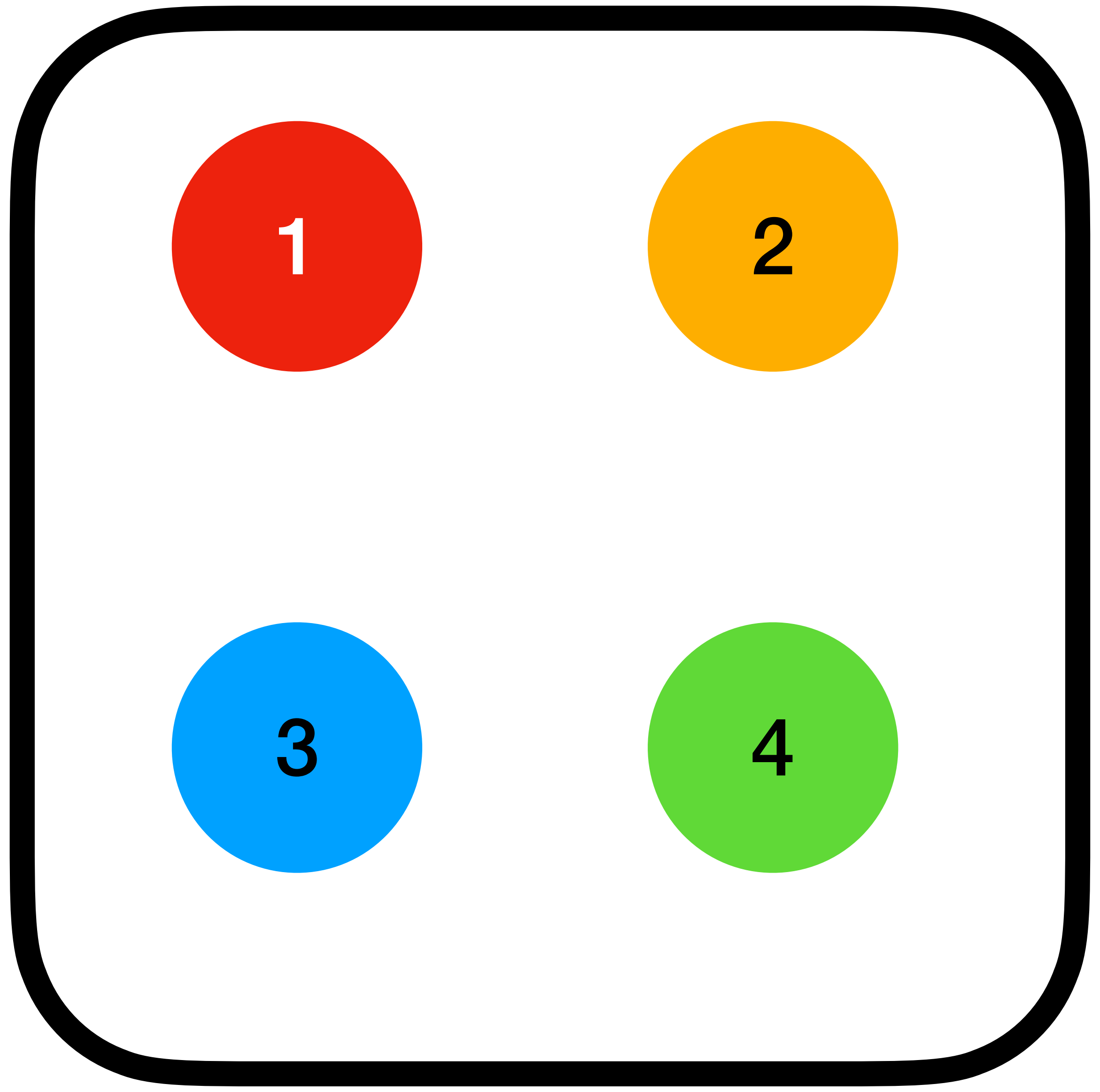
$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

π	δ_π^G
(A, B, C)	$(80, 0, 10)$
(A, C, B)	$(80, 5, 5)$
(B, A, C)	$(24, 56, 10)$
(B, C, A)	$(18, 56, 16)$
(C, A, B)	$(15, 5, 70)$
(C, B, A)	$(18, 2, 70)$
ϕ	$(39.2, 20.7, 30.2)$

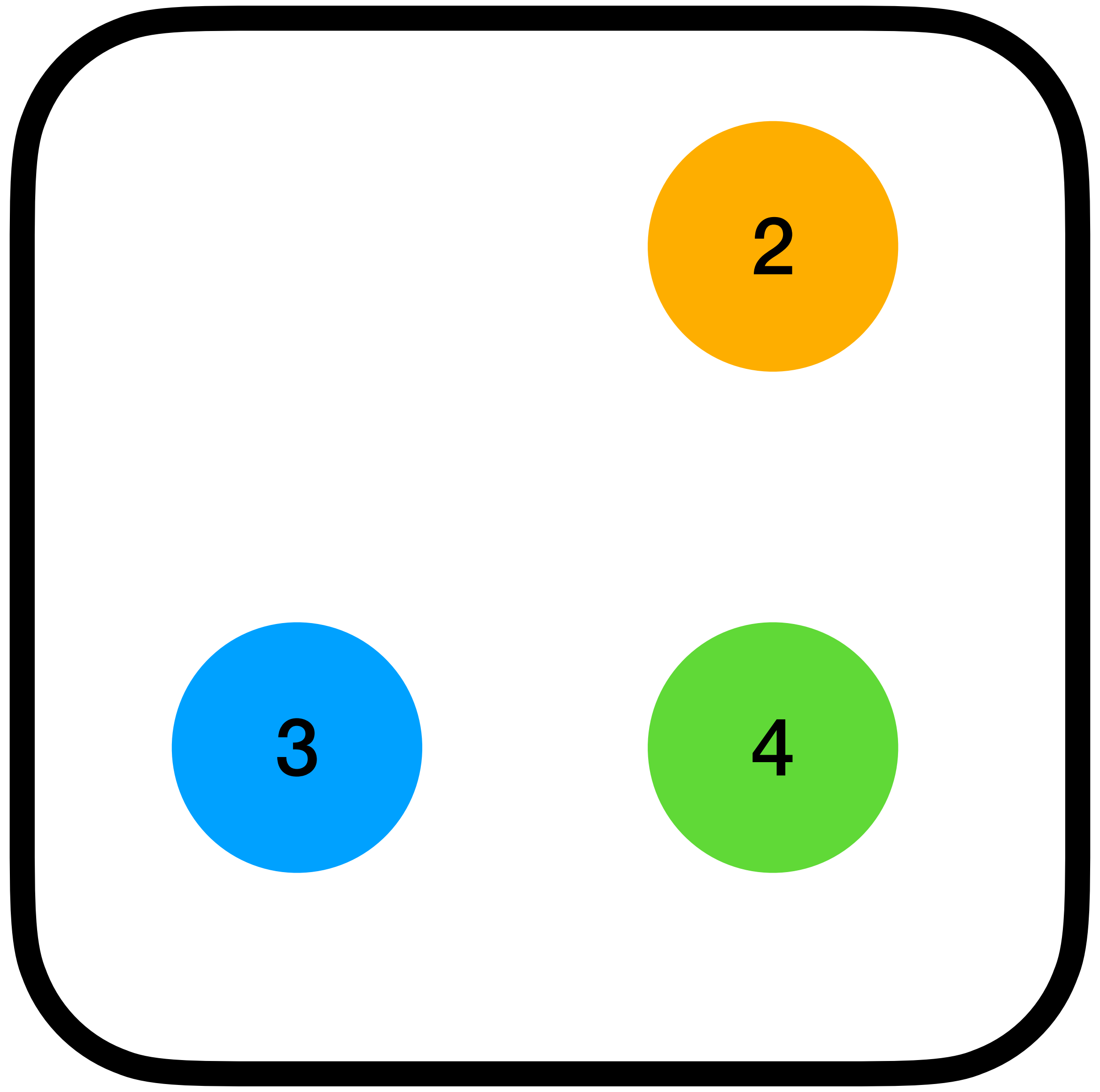
1



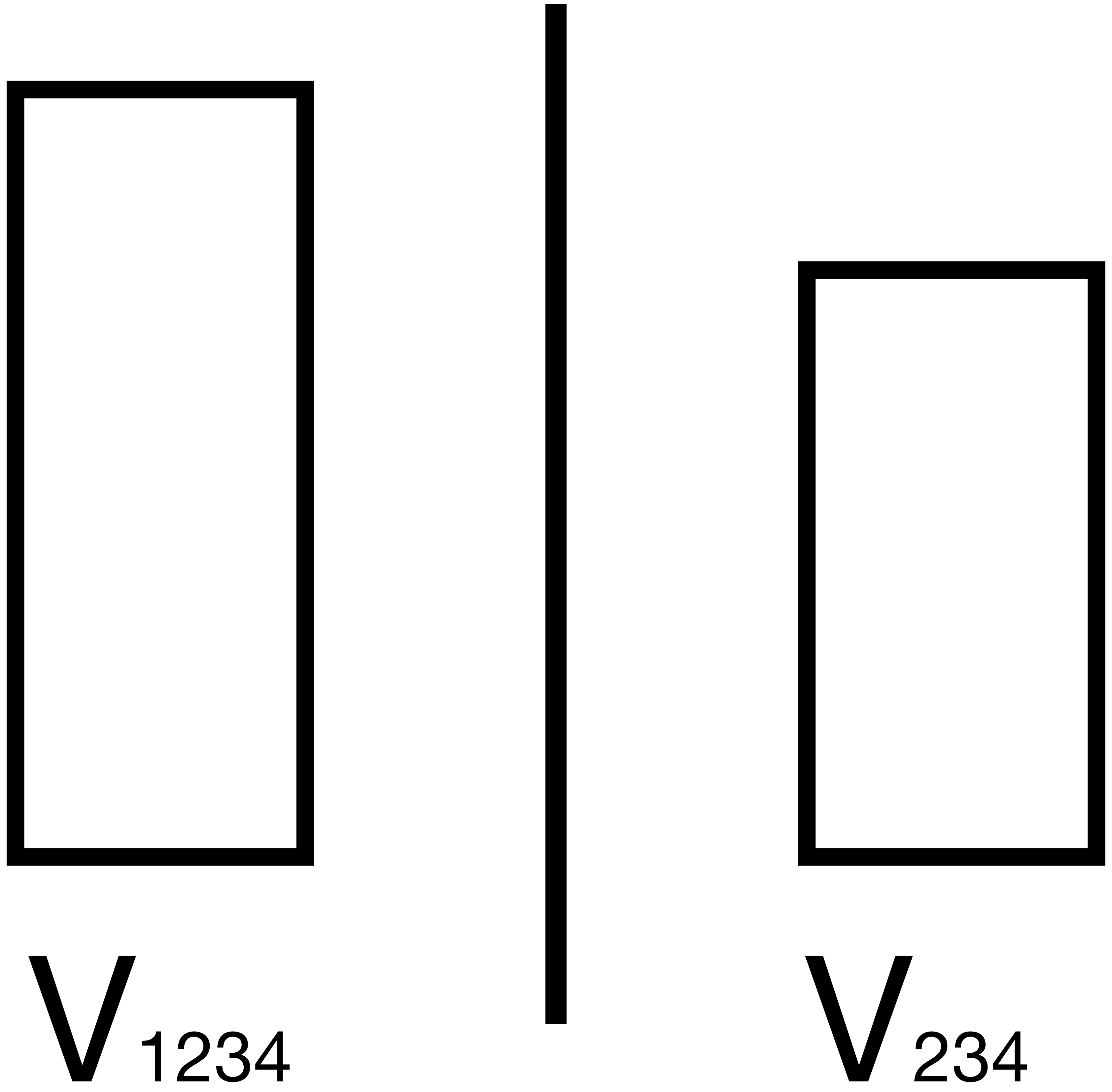
C_{1234}

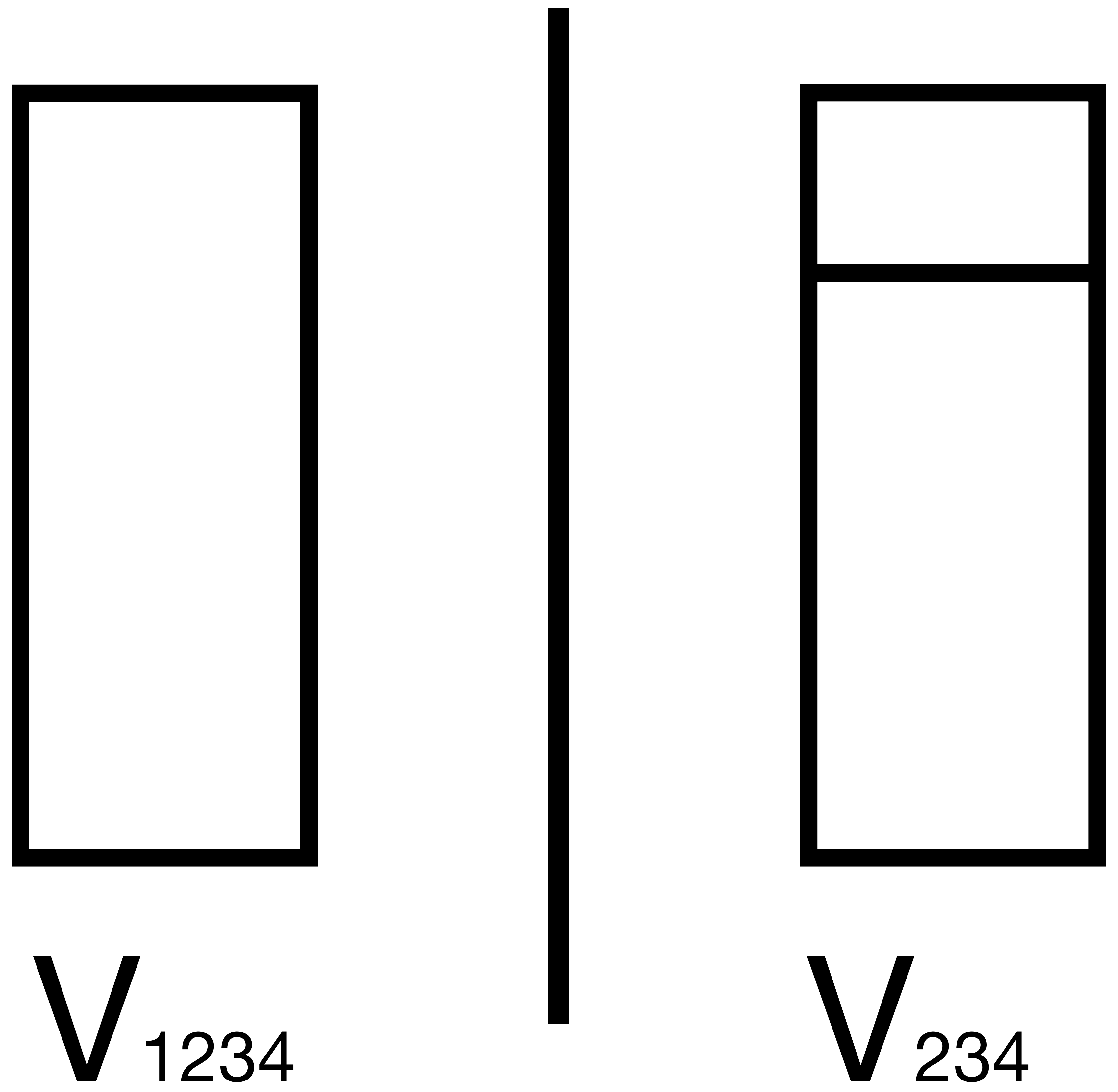


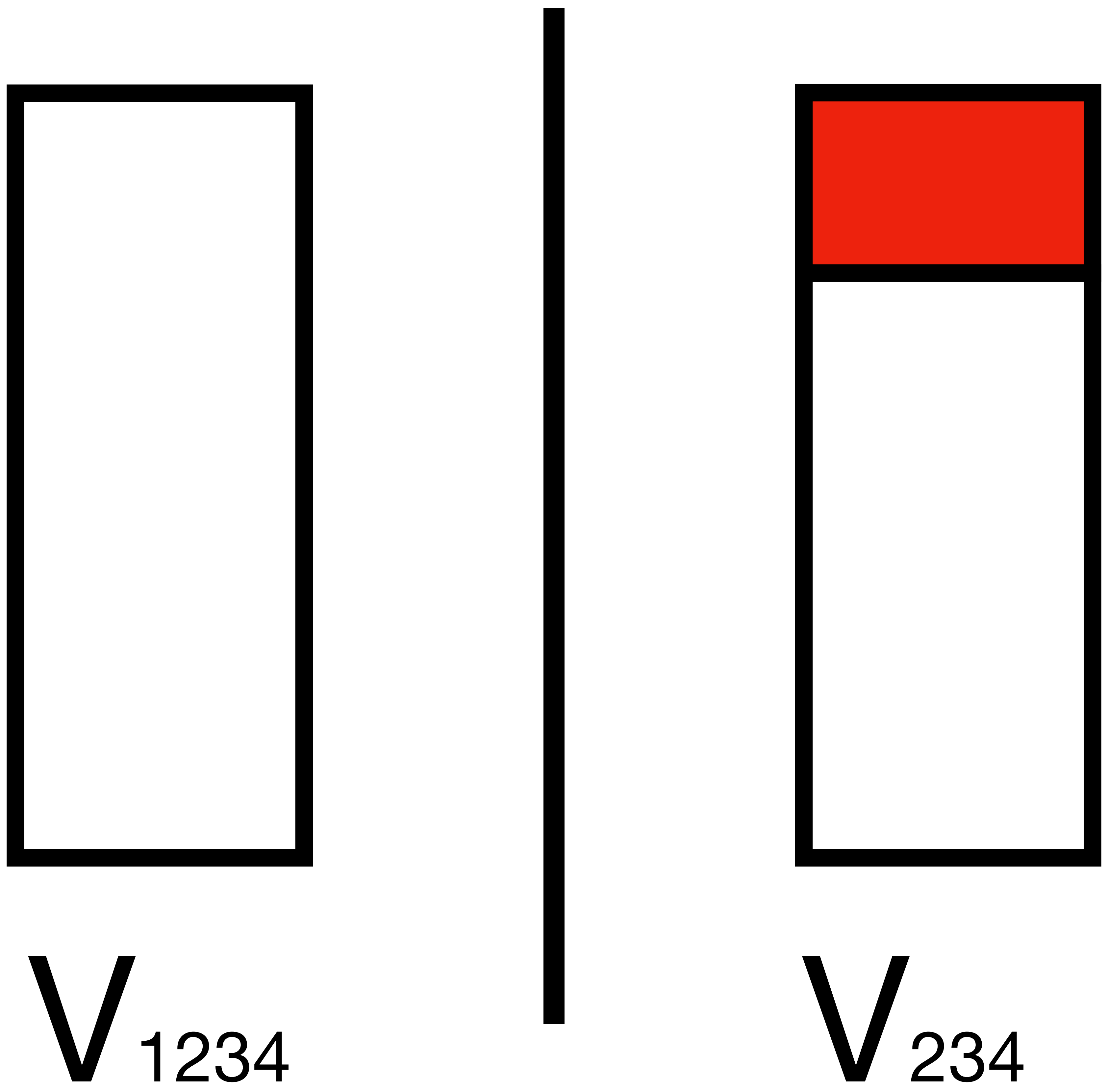
C_{1234}



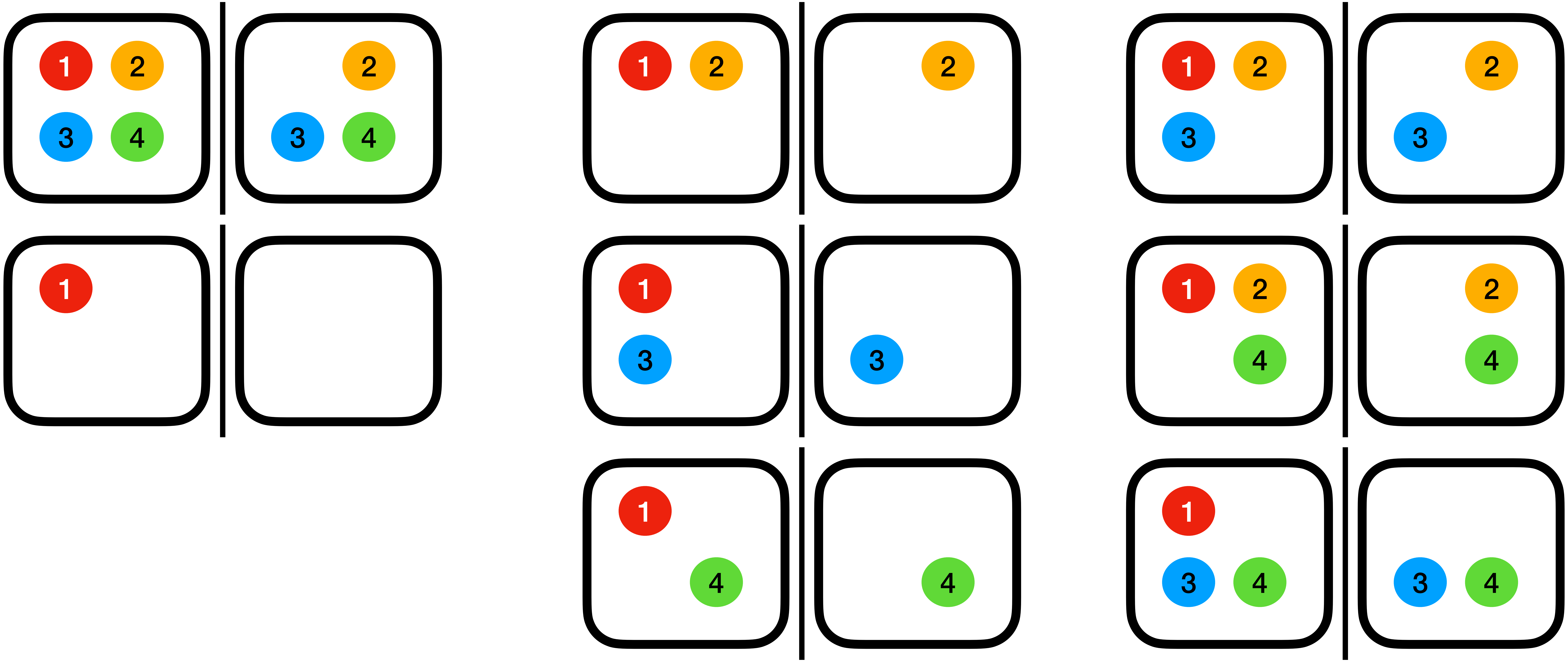
C_{234}

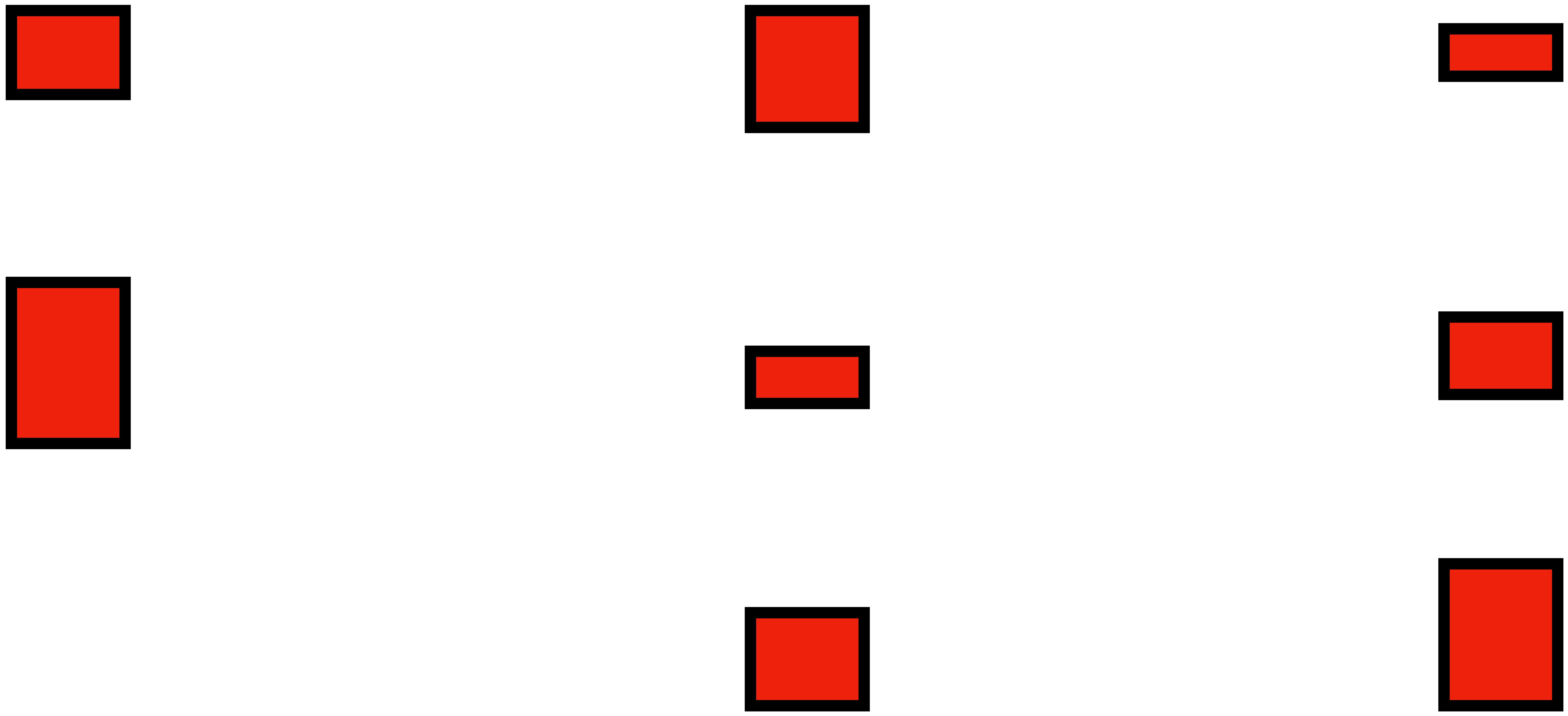

V_{1234}
V_{234}


V_{1234}
V_{234}


V_{1234}
V_{234}

$$= V_{1234} - V_{234} = \boxed{\text{Marginal contribution of member } 1 \text{ to } C_{234}}$$





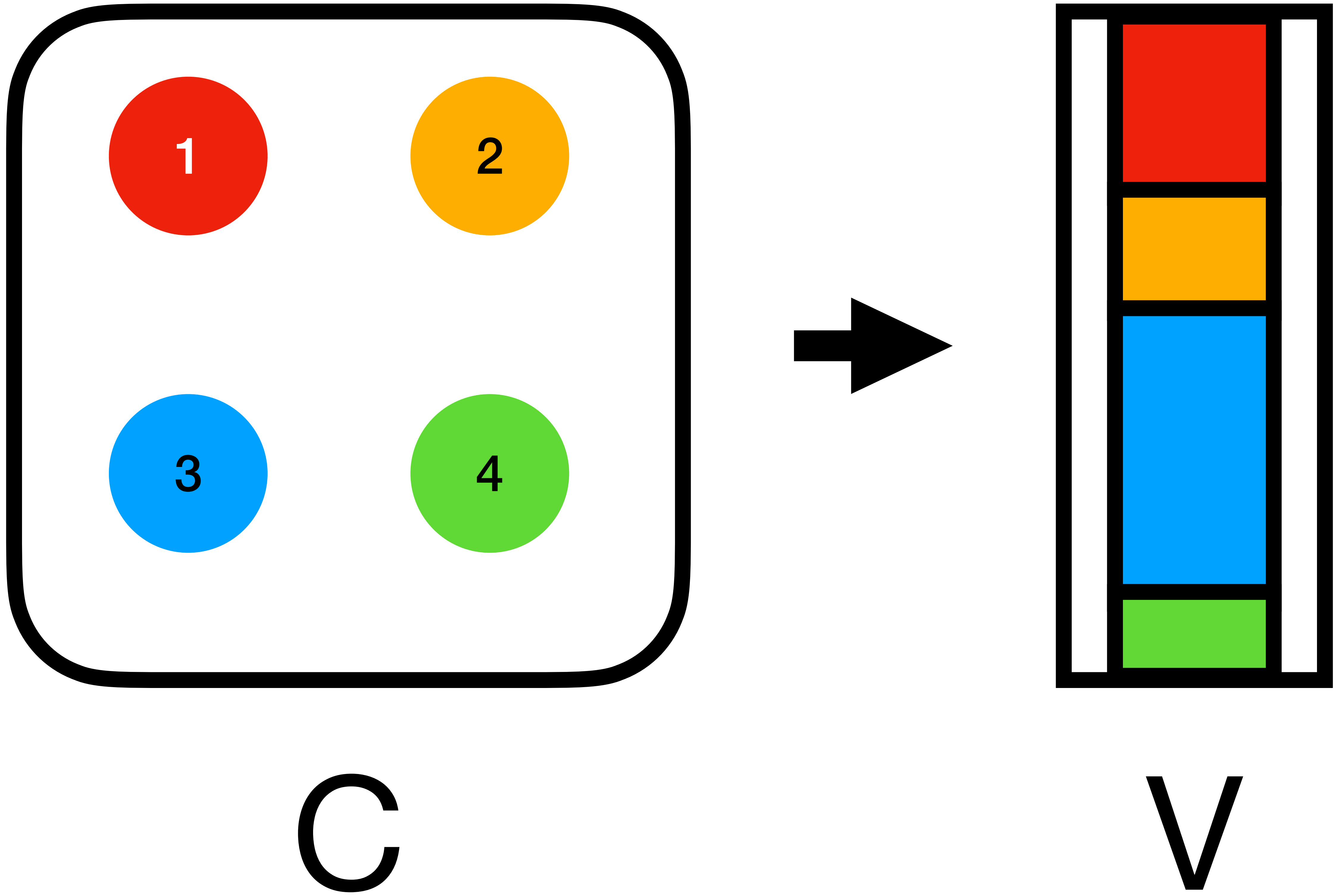
= $\phi 1$

$= \phi 1$

$= \phi 2$

$= \phi 3$

$= \phi 4$


1
2
3
4
C
V

Shapley value

Average marginal contribution of a feature value across all possible coalitions

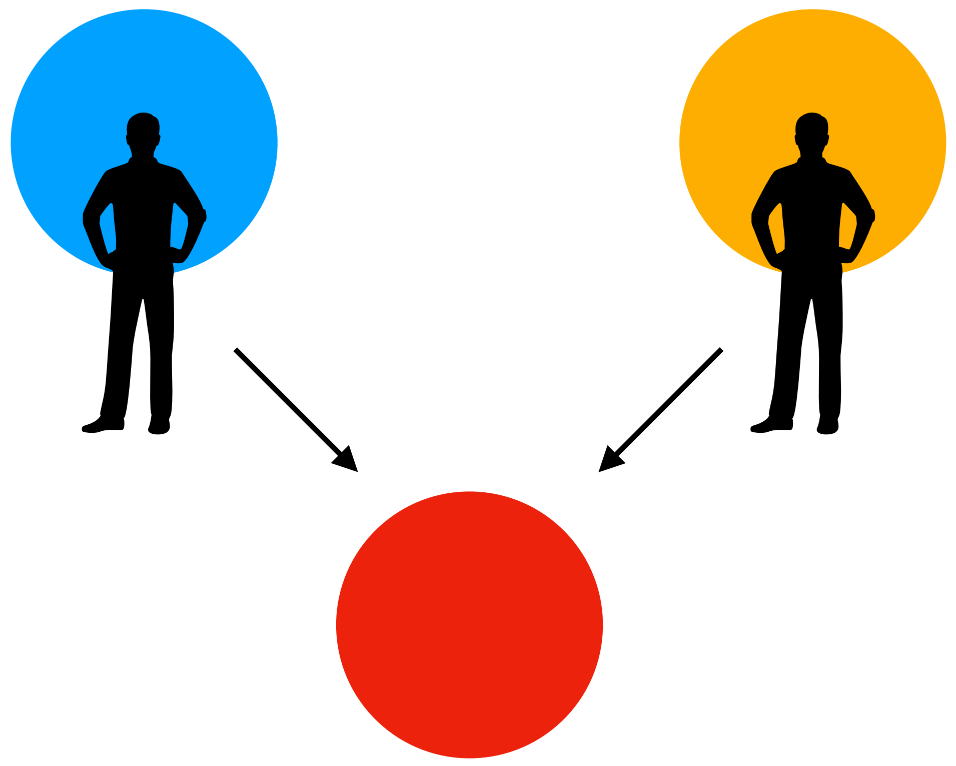
Shapley value

- The Shapley value is one way to distribute the total gains to the players
  - assuming that they all collaborate
- It is a "fair" distribution

Fair distribution

2 people join to work together

$f(p1) = 50K$

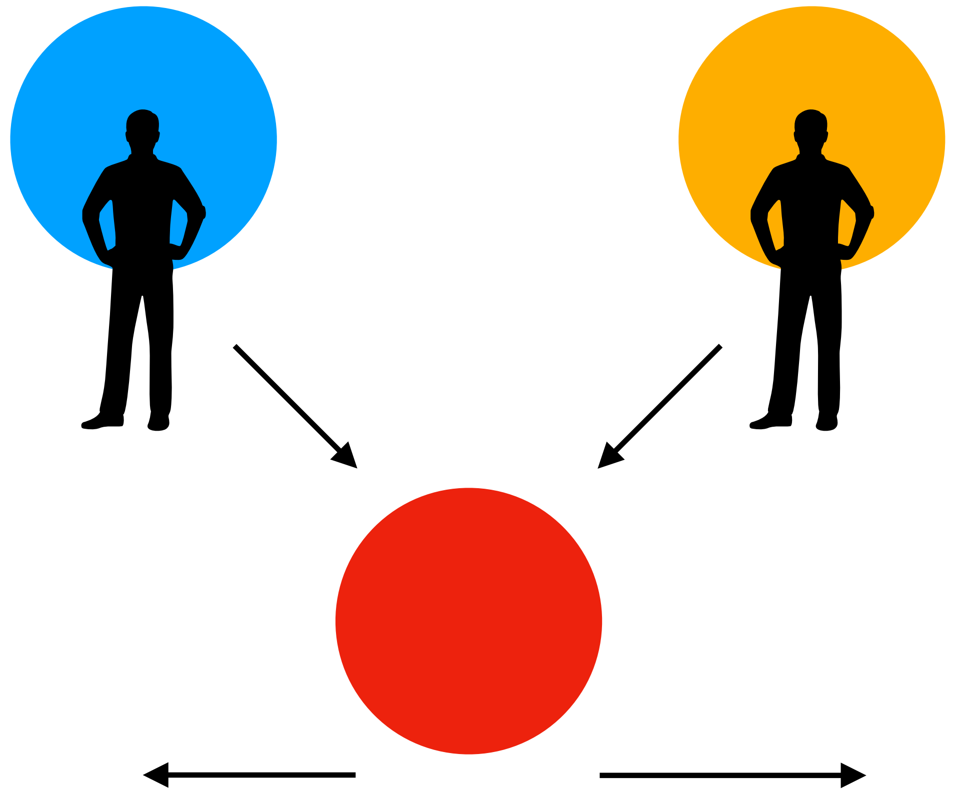


$f(p2) = 70K$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

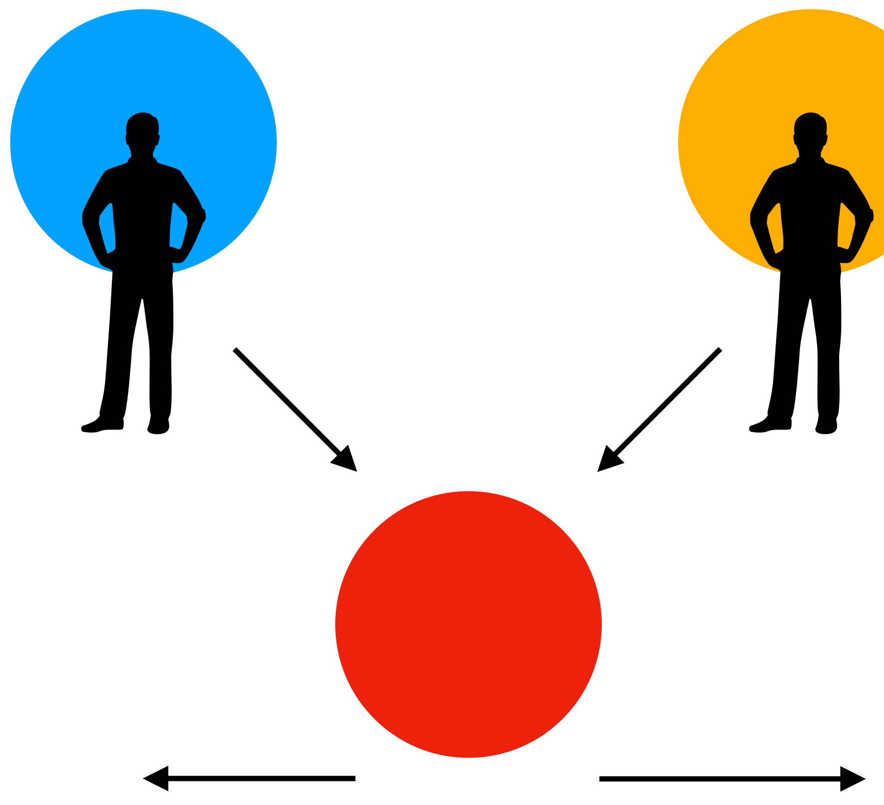
$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

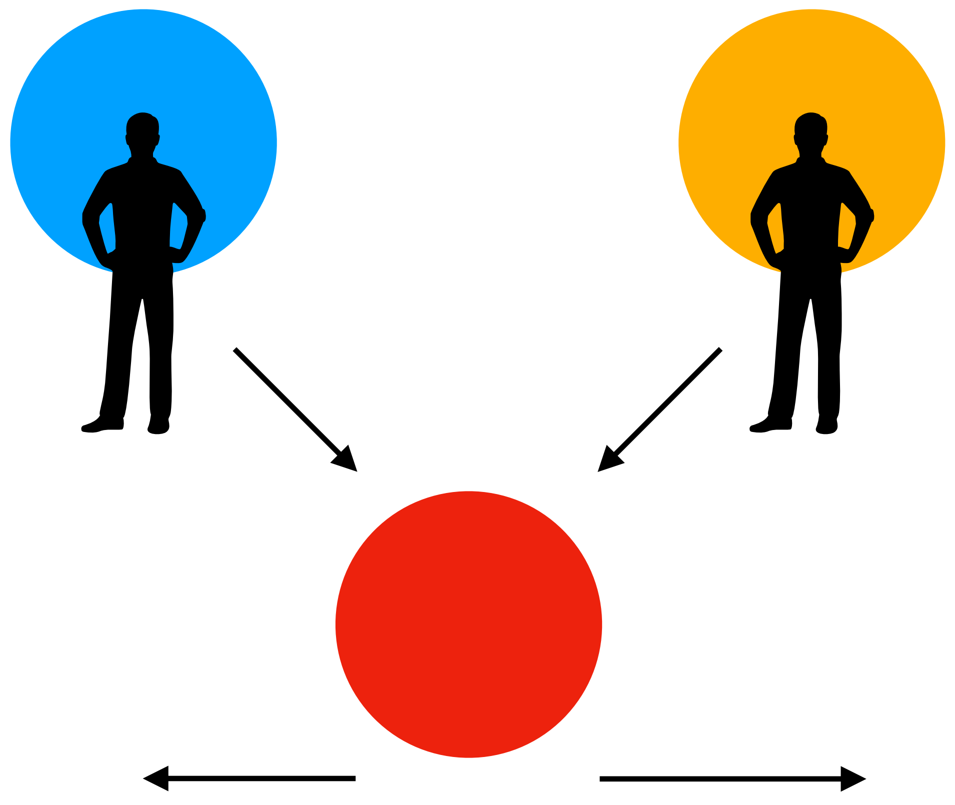
$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

$f(p1,\ p2) = 120K$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

$Sh_{p1} = 50K$

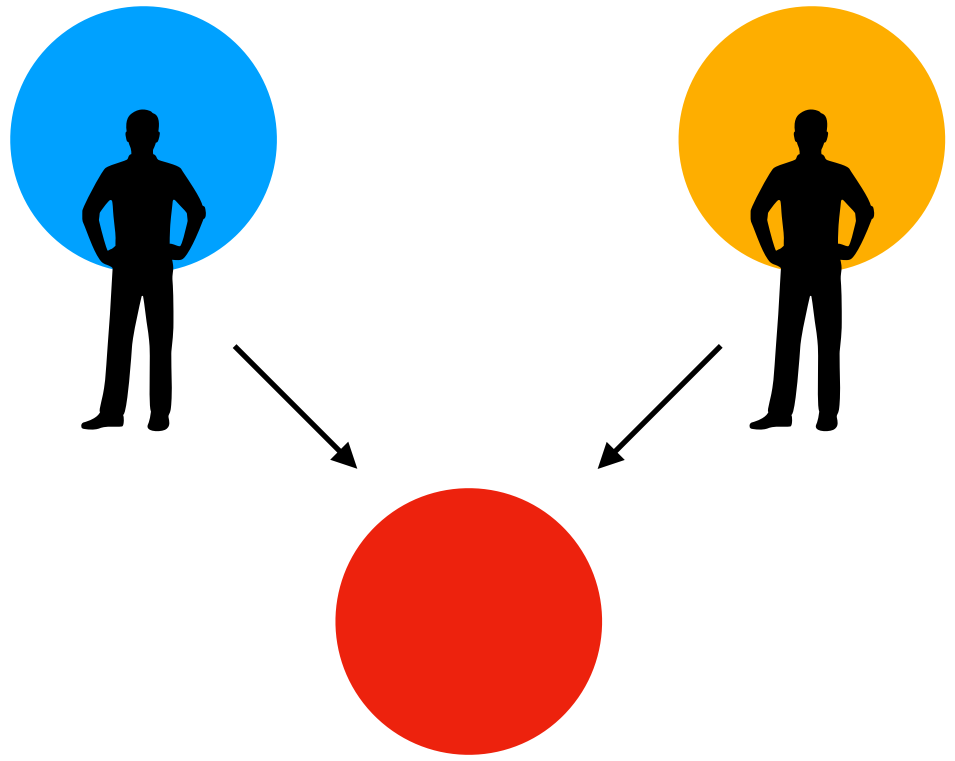
$f(p1, p2) = 120K$

$Sh_{p2} = 70K$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

$Sh_{p1} = 50K$

$f(p1,\ p2) = 120K$

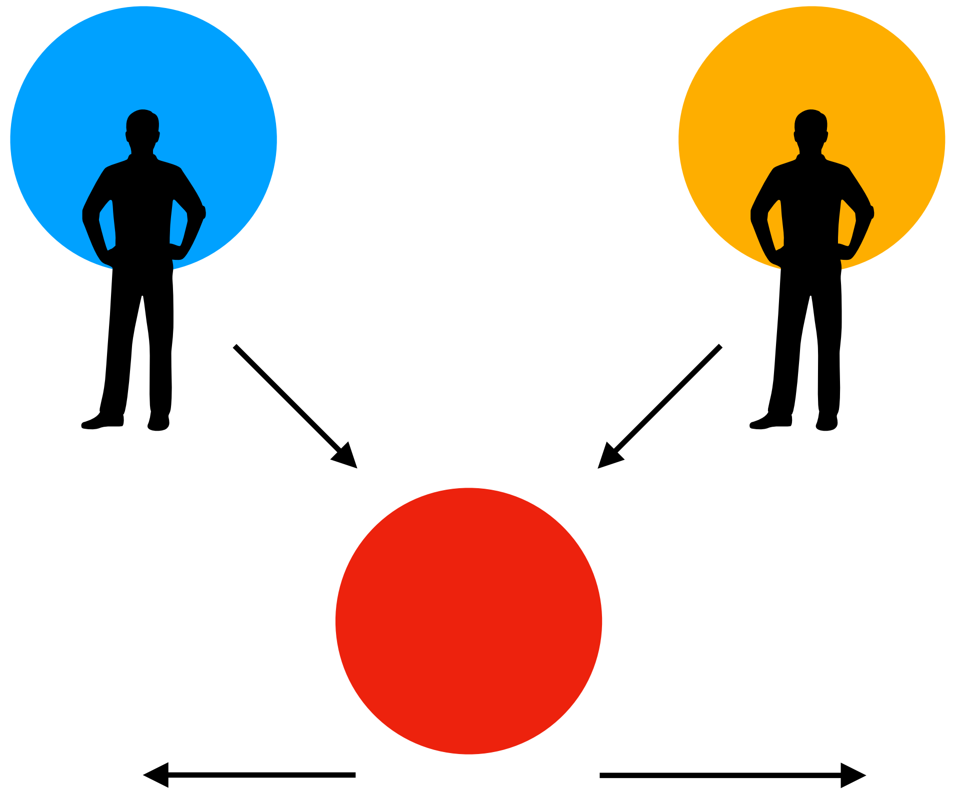
$f(p1,\ p2) = 150K$

$Sh_{p2} = 70K$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

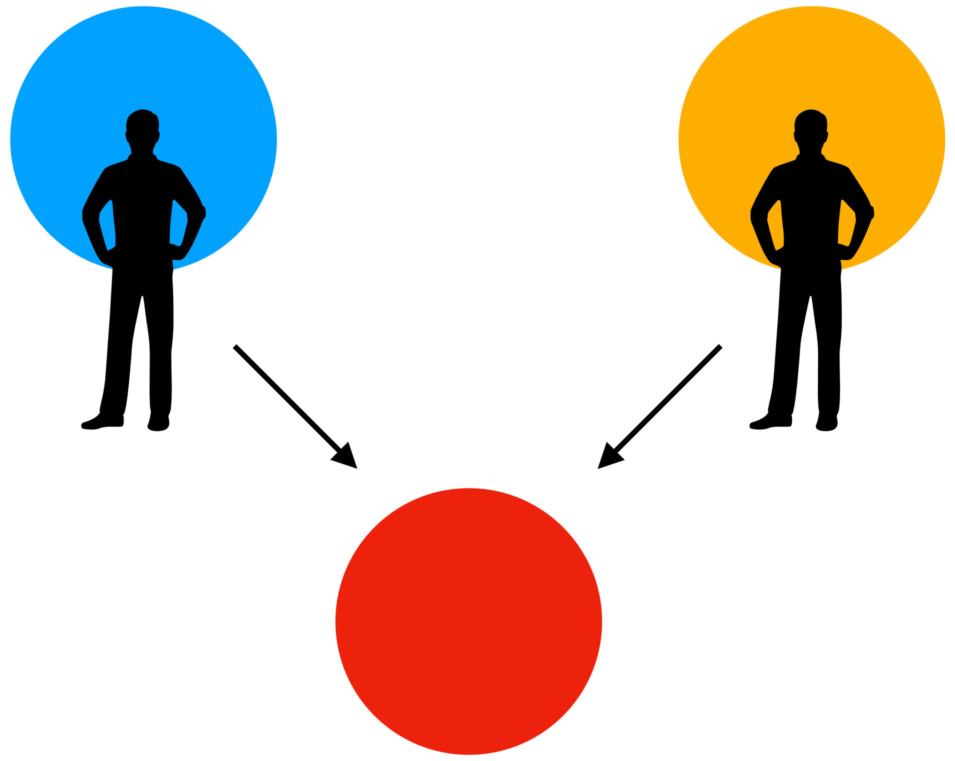
$Sh_{p1} = 50K$ $\quad f(p1, p2) = 120K$ $\quad Sh_{p2} = 70K$

$Sh_{p1} = 65K$ $\quad f(p1, p2) = 150K$ $\quad Sh_{p2} = 85K$

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

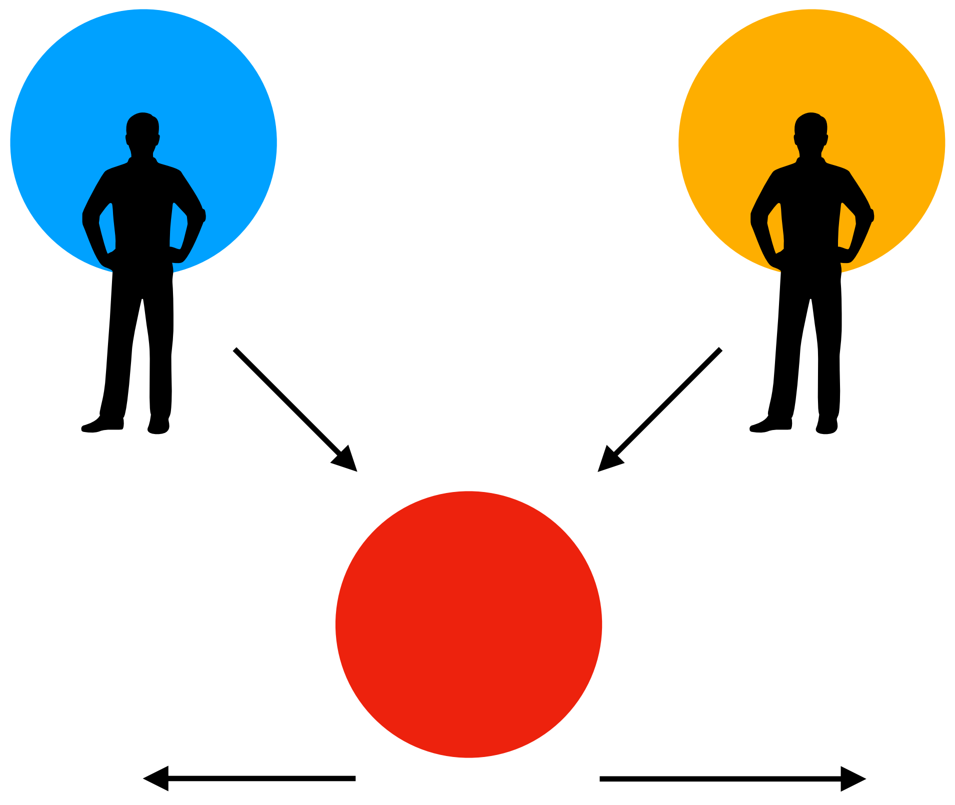
$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

$Sh_{p1} = 50K$	$f(p1, p2) = 120K$	$Sh_{p2} = 70K$
$Sh_{p1} = 65K$	$f(p1, p2) = 150K$	$Sh_{p2} = 85K$
	$f(p1, p2) = 100K$	

Fair distribution

2 people join to work together

$f(p1) = 50K$



$f(p2) = 70K$

$Sh_{p1} = 1/2\ (f_{p1} + [f_{p1,p2} - f_{p2}]\ )$

$Sh_{p2} = 1/2\ (f_{p2} + [f_{p1,p2} - f_{p1}]\ )$

$Sh_{p1} = 50K$	$f(p1, p2) = 120K$	$Sh_{p2} = 70K$
$Sh_{p1} = 65K$	$f(p1, p2) = 150K$	$Sh_{p2} = 85K$
$Sh_{p1} = 40K$	$f(p1, p2) = 100K$	$Sh_{p1} = 60K$

Shapley value
Mathematically

- Coalition game
- N: set of p players in a game
- Characteristic function val: 2^p -> R, v({}) = 0
- The amount that player *j* gets given a coalition game (val, N) is

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(val\left(S \cup \{j\}\right) - val(S)\right)$$

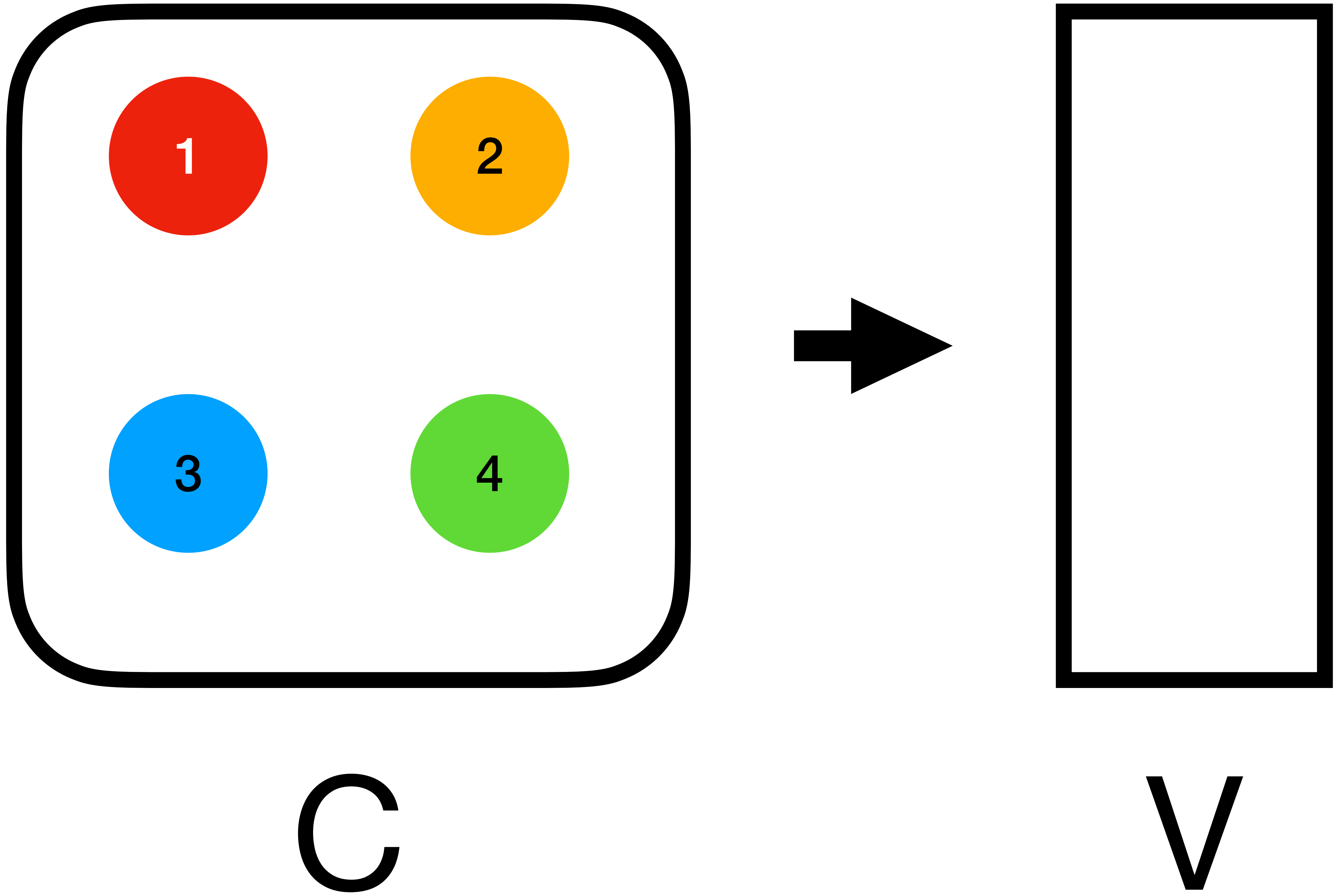
A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

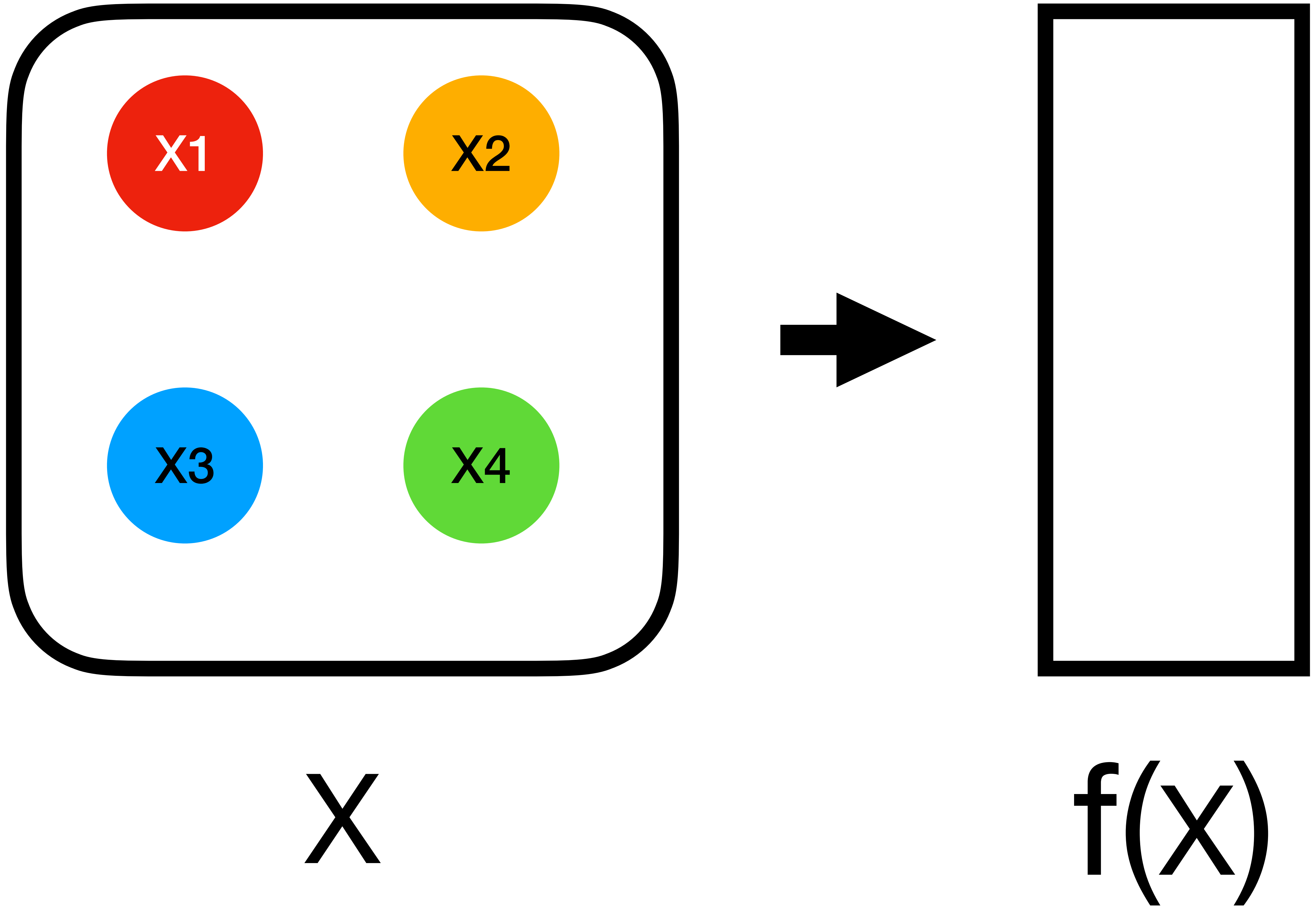
Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of


1
2
3
4
C
V

In terms of ML



SHAP

Shapley Additive Explanation

SHAP

Shapley Additive Explanation

x	x'
$f(x)$	$g(x')$

Inputs → x x'

Model → $f(x)$ $g(x')$

x x' ← Simplified local inputs

$f(x)$ $g(x')$ ← Explanatory model

x	x' ← Simplified local inputs
$f(x)$	$g(x')$ ← Explanatory model

As simplified binary vector where features are either included or excluded

[1, 0, 0, 0, 1, 1, 0, 1]

1.

We need to ensure

If	x	$\sim$	x'
Then	$f(x)$	$\sim$	$g(x')$

2.

g(x') must take this form

$$g(x') = \phi_0 + \sum \phi_i \, x'_i$$

$$g(x') = \phi_0 + \Sigma\, \phi_i\, x'_i$$

Null Output

Average Output of model

Feature effect

$$g(x') = \phi_0 + \Sigma\, \phi_i\, x'_i$$

Explained Effect of feature i

Feature effect

$$g(x') = \phi_0 + \Sigma\, \phi_i\, x'_i$$

Explained Effect of feature i

How much feature i
changed the output of the model

Feature effect

Attribution

$$g(x') = \phi_0 + \Sigma\ \phi_i\ x'_i$$

Null Output

$$g(x') = \phi_0 + \sum \phi_i x'_i$$

Additive Feature Attribution

Authors describe

3 desirable properties of such an additive method

Local Accuracy

Missingness

Consistency

Local Accuracy

$f(x) \sim g(x') \text{ if } x' \sim x$

Missingness

Consistency

Missingness

Local Accuracy

$x'_i = 0 \Rightarrow \phi_i = 0$

Consistency

Consistency

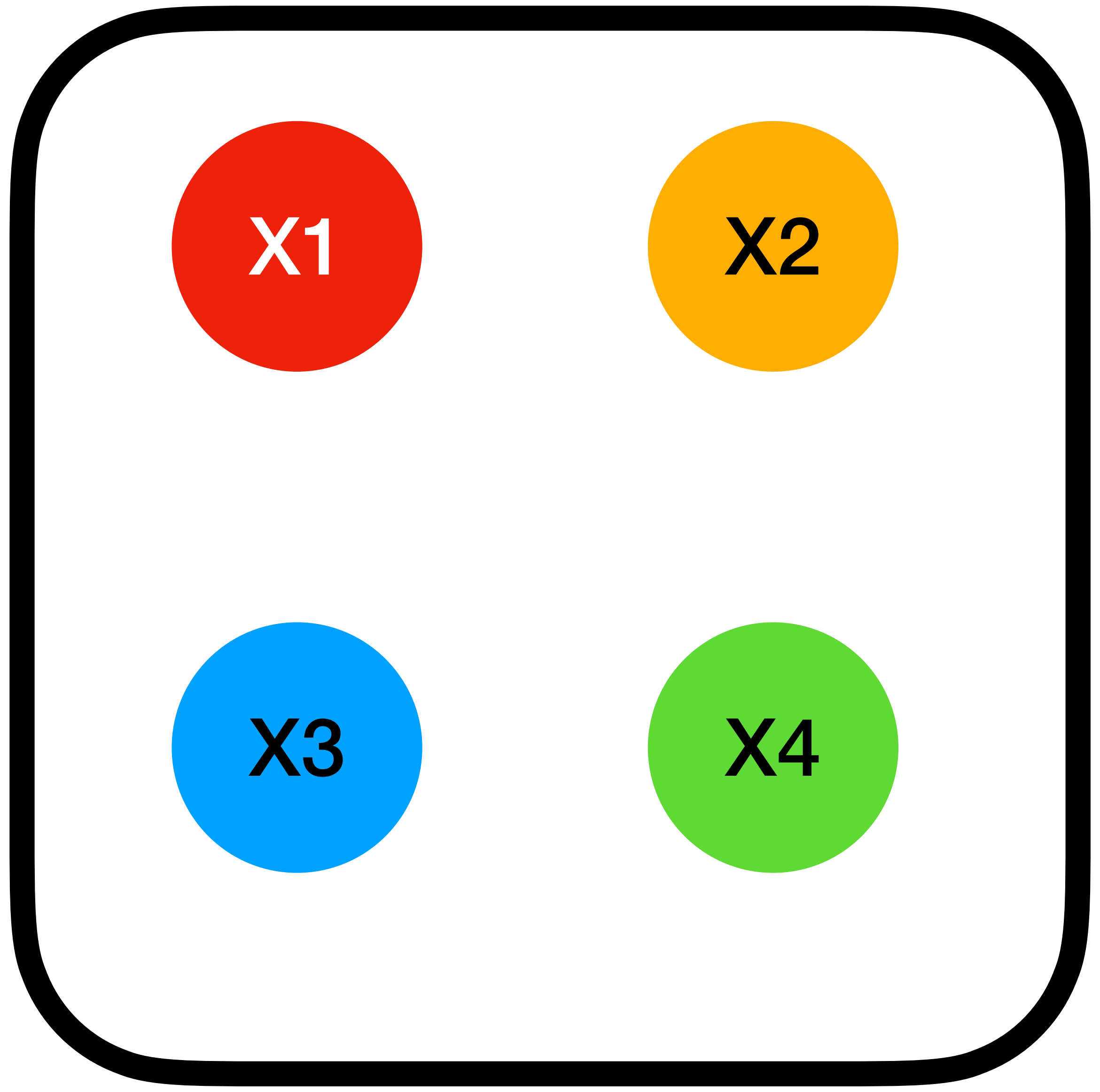
Local Accuracy

Missingness

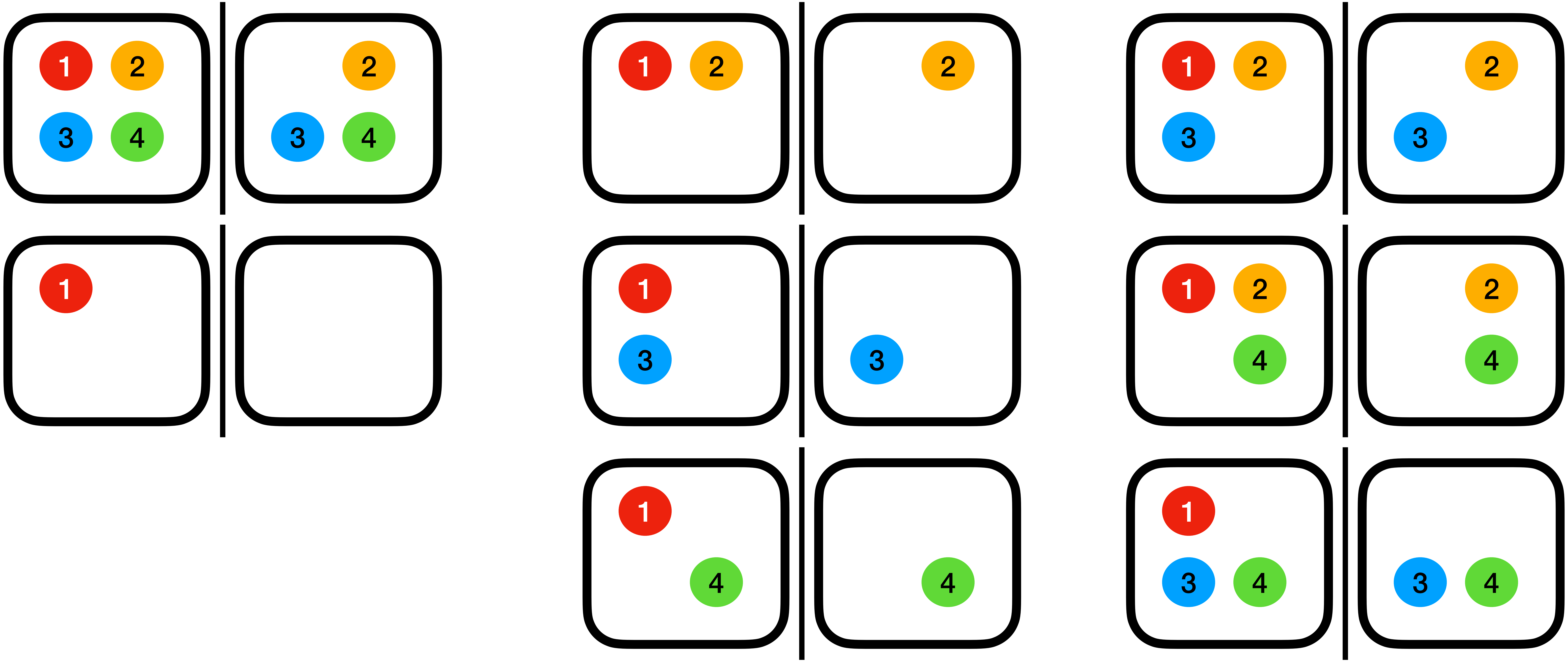
If feature contribution changes
the feature effect cannot change in the opposite direction

Shapley value

$$g(x') = \phi_0 + \Sigma \phi_i x'_i$$



X


1 2 3 4
2 3 4
1
1 2
2
1 3
3
1 4
4
1 2 3
2 3
1 2 4
2 4
1 3 4
3 4

4 Features: 64 coalitions to sample

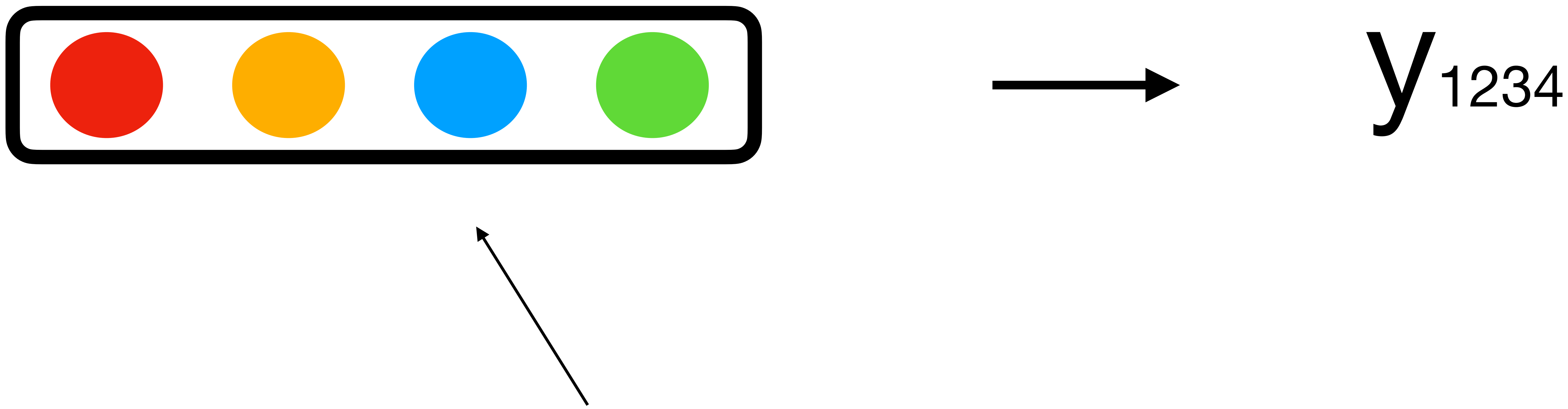
32 Features: 17.1billion coalitions to sample

Shapley Kernel

Means of approximating Shapley values through much fewer samples

Shapley Kernel

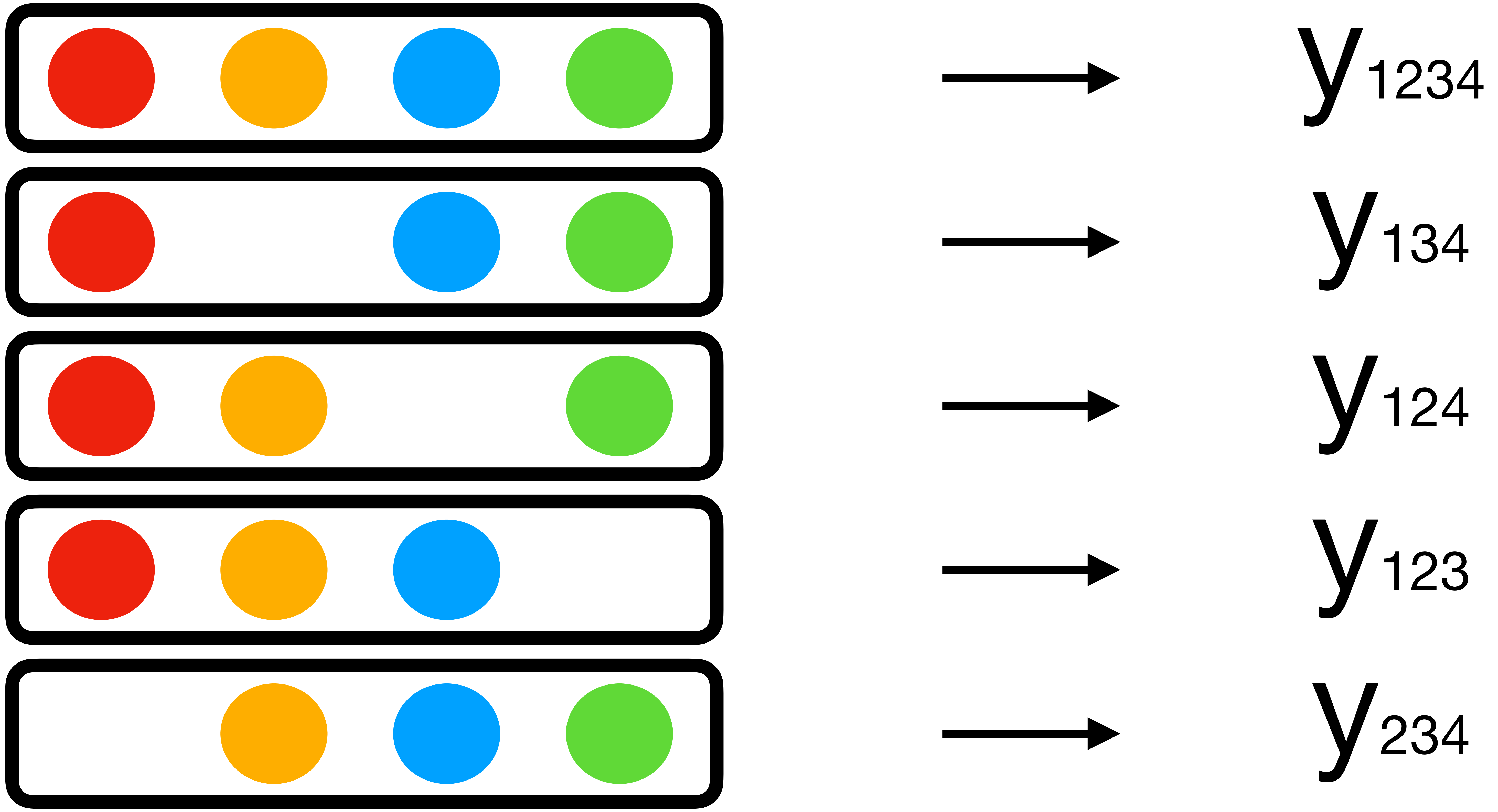
Means of approximating Shapley values through much fewer samples



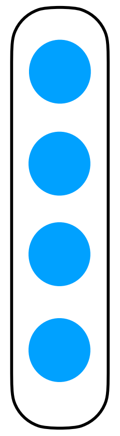
Explain this data sample

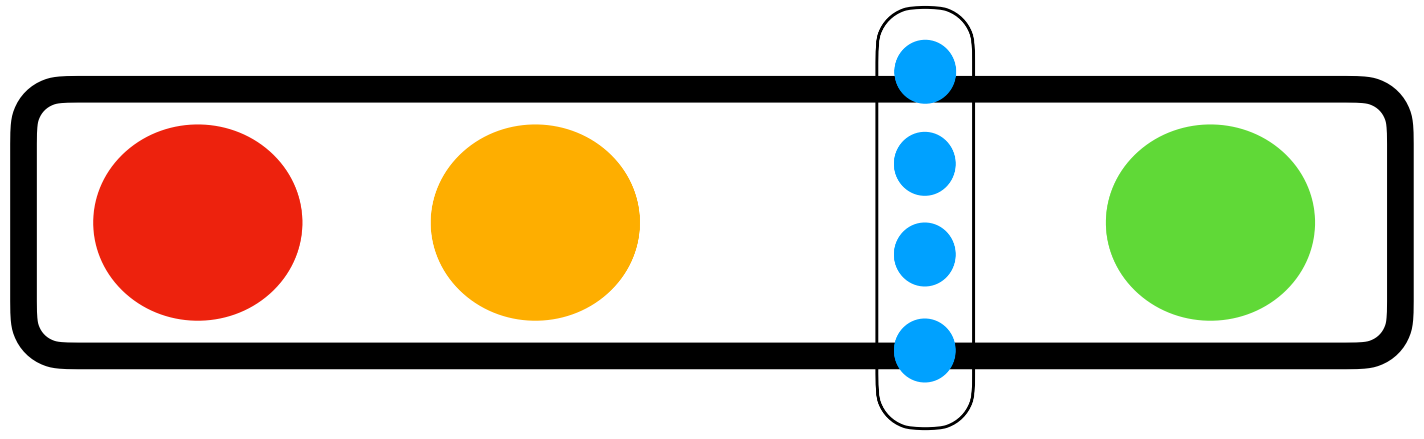
Shapley Kernel

Means of approximating Shapley values through much fewer samples



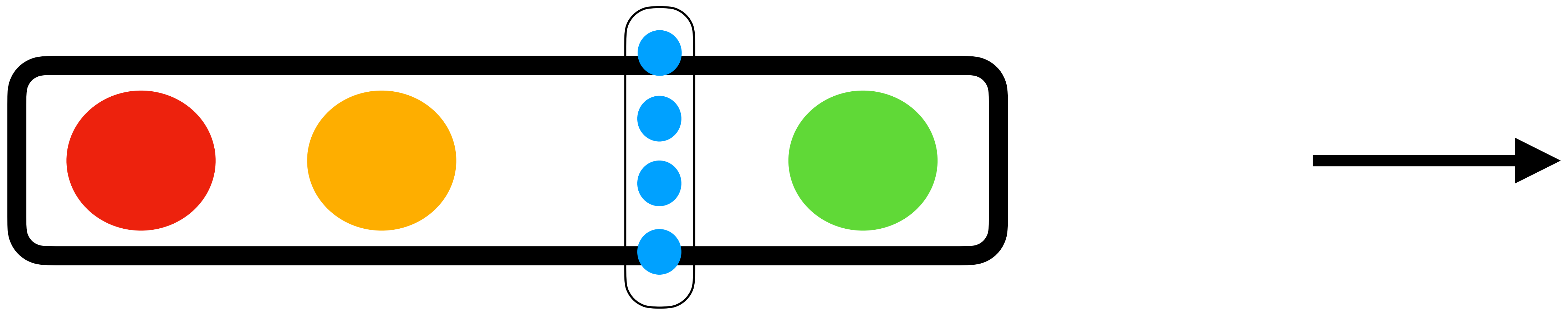
Pass various Permutations



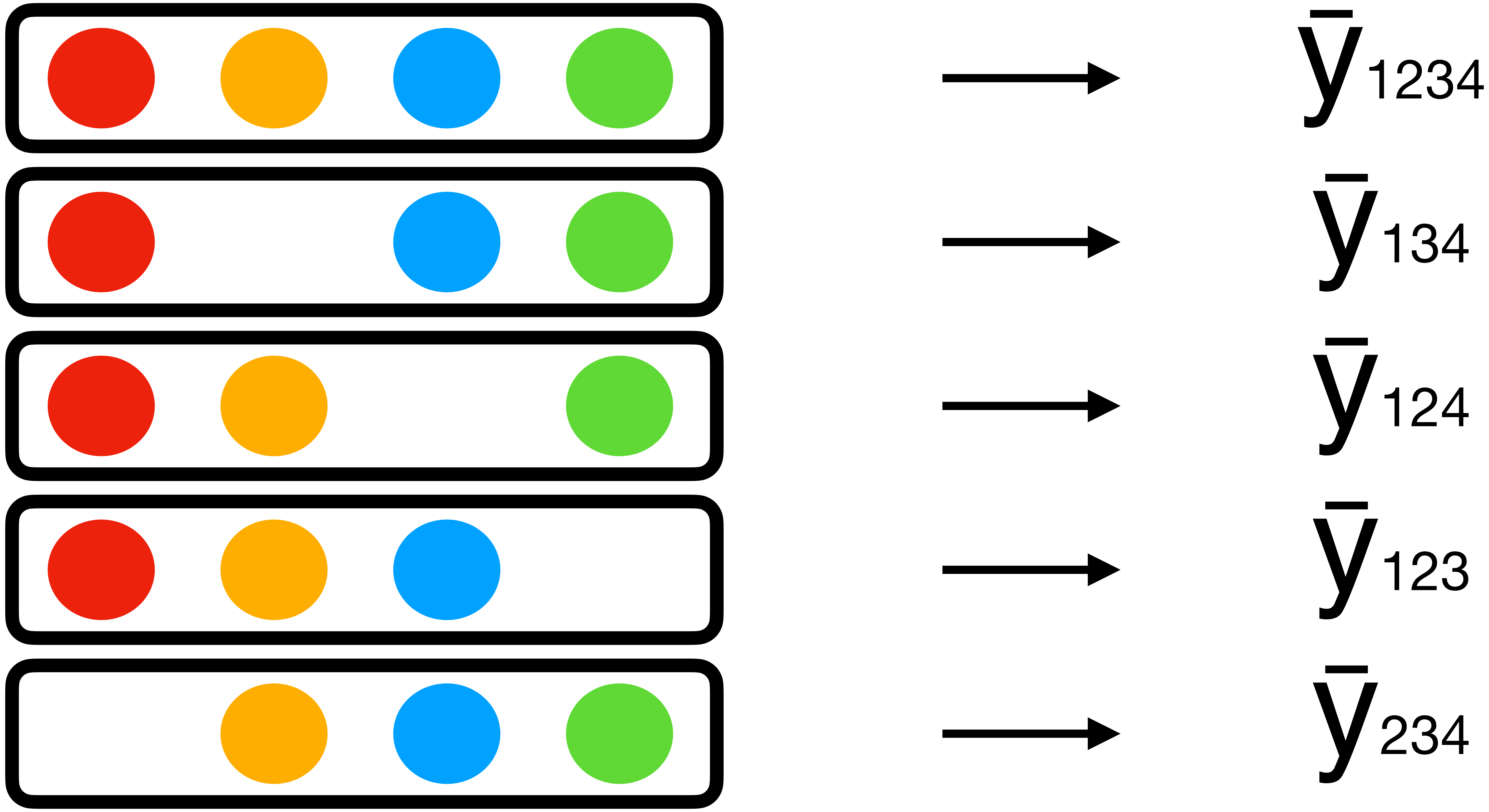


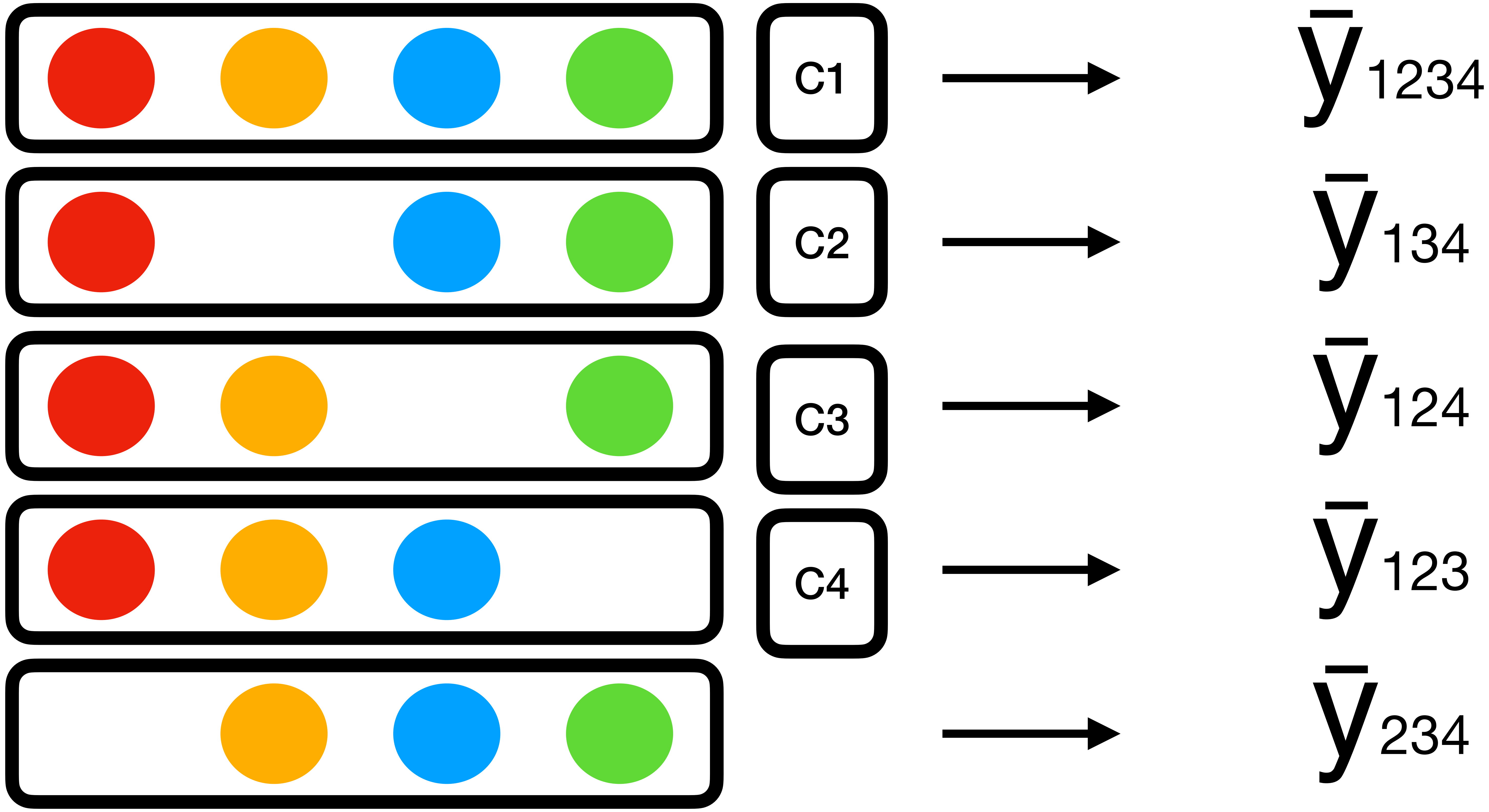
Background dataset

$$E\,[\; y_{12i4}\ \forall i \in B \;]$$

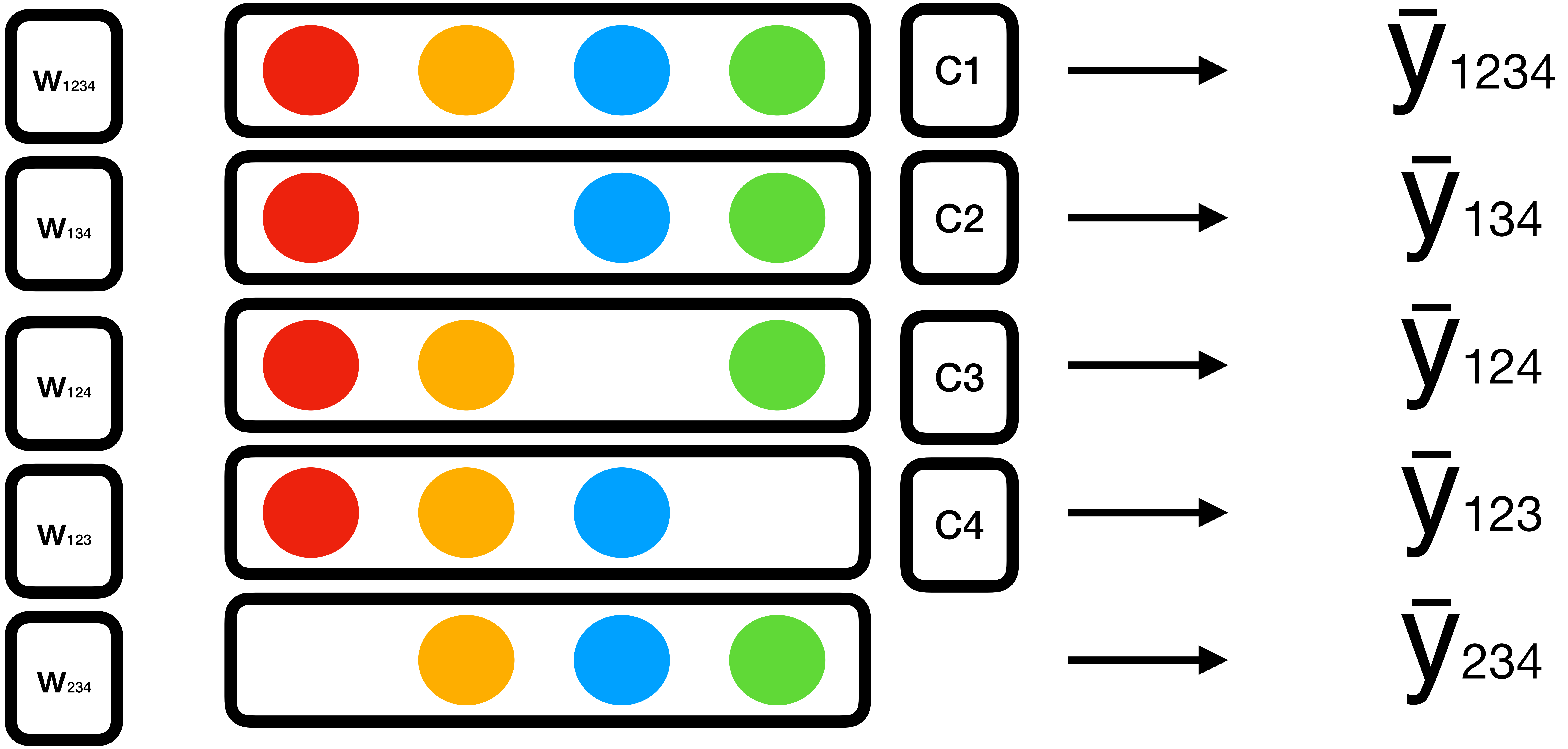


$\bar{y}_{124}$

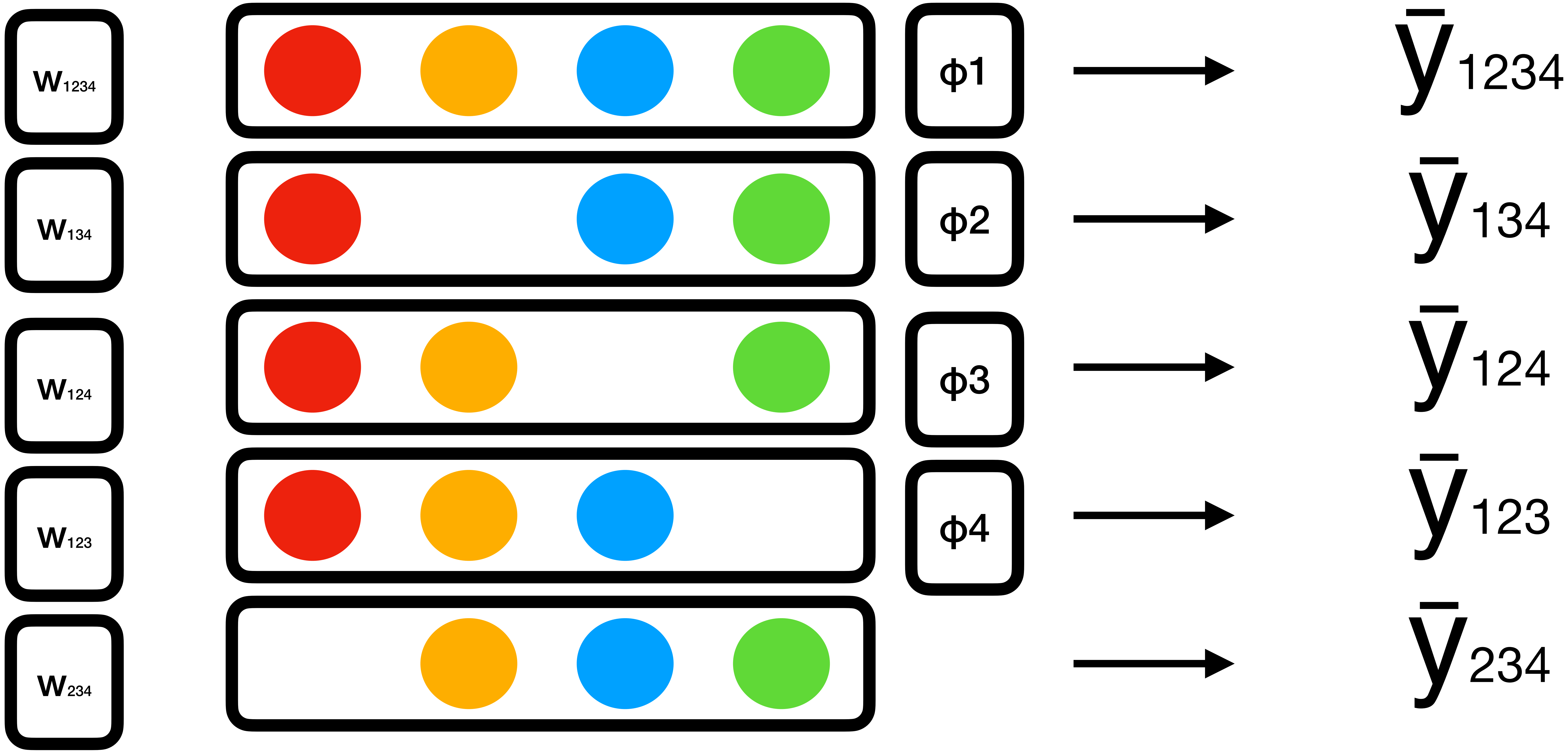

$\bar{y}_{1234}$
$\bar{y}_{134}$
$\bar{y}_{124}$
$\bar{y}_{123}$
$\bar{y}_{234}$


C1
C2
C3
C4
$\bar{y}_{1234}$
$\bar{y}_{134}$
$\bar{y}_{124}$
$\bar{y}_{123}$
$\bar{y}_{234}$

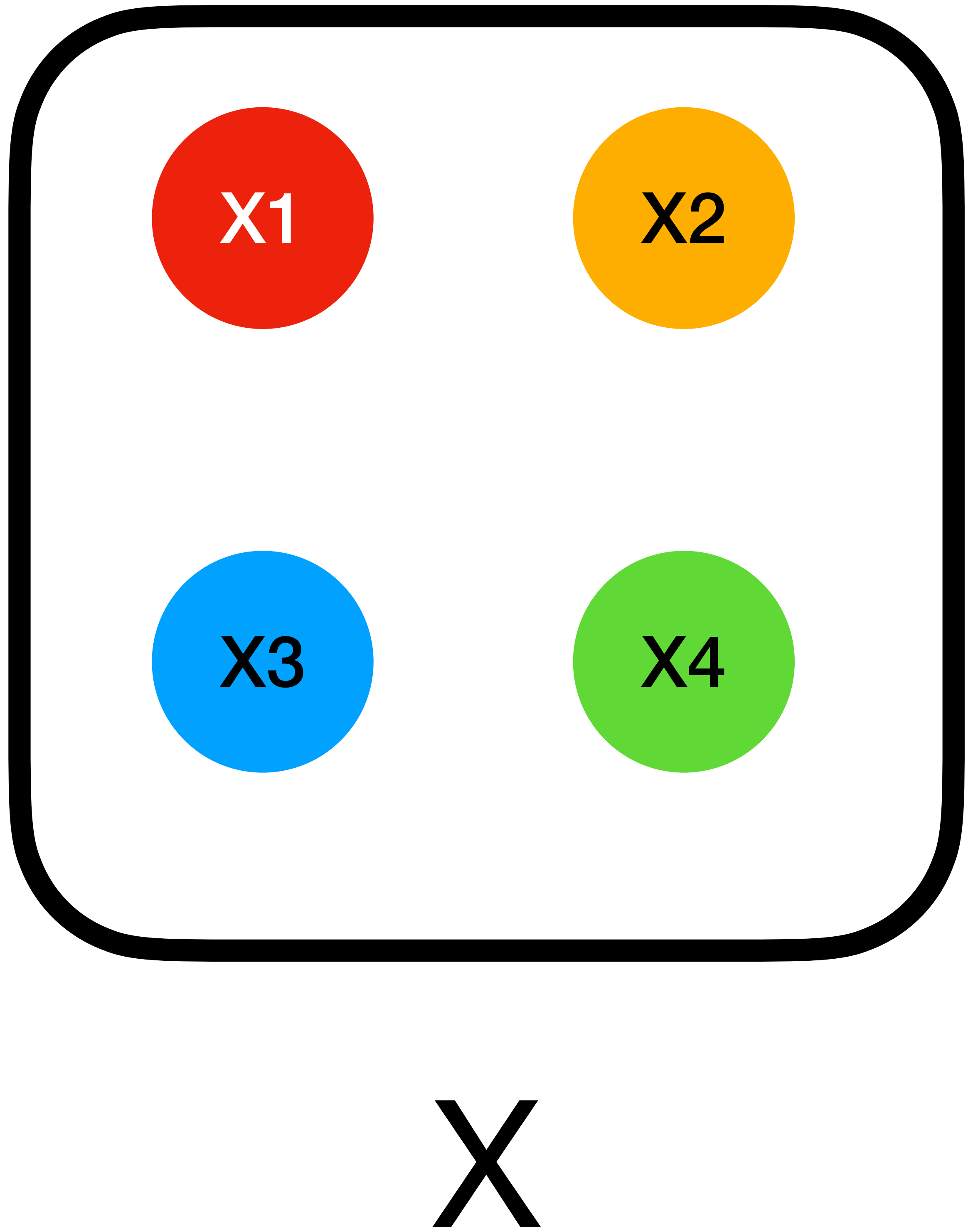
$$Wc = \frac{\text{\# Total features -1}}{\text{\# coalitions of size |C| * \# included features in |C| * \# excluded features in |C|}}$$



w_{1234}
w_{134}
w_{124}
w_{123}
w_{234}
C1
C2
C3
C4
$\bar{y}_{1234}$
$\bar{y}_{134}$
$\bar{y}_{124}$
$\bar{y}_{123}$
$\bar{y}_{234}$



w_{1234}
w_{134}
w_{124}
w_{123}
w_{234}
$\phi 1$
$\phi 2$
$\phi 3$
$\phi 4$
$\bar{y}_{1234}$
$\bar{y}_{134}$
$\bar{y}_{124}$
$\bar{y}_{123}$
$\bar{y}_{234}$


X1
X2
X3
X4
X

References

- Article: Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Proceedings of the 31st international conference on neural information processing systems*. 2017.
- Video: https://www.youtube.com/watch?v=VB9uV-x0gtg&t=444s
- Wikipedia page: https://en.wikipedia.org/wiki/Shapley_value
- Book: https://christophm.github.io/interpretable-ml-book

About Me

Nikhil Verma (http://lihkinverma.github.io/portfolio)

- University of Toronto
  - Master of Science in Applied Computing
- Thapar University
  - Bachelor of Engineering
- InfoEdge (India) Ltd
  - Senior Software Engineer
- MentorGraphics (Siemens) India Pvt Ltd
  - Software Engineering Intern
- Indian Institute of Technology(IIT) Delhi
  - Project Associate

Interested in designing products for scale

Thank You

For being patient listeners

Nikhil Verma (lih.verma@gmail.com)