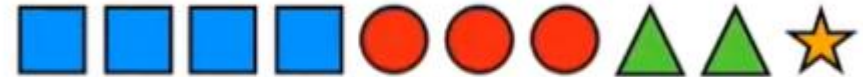


# The Gini impurity index

## Intuition

Measuring the diversity in a dataset

Which one is more diverse?



# Which one is more diverse?



# Which one is more diverse?

Gini = 0.42



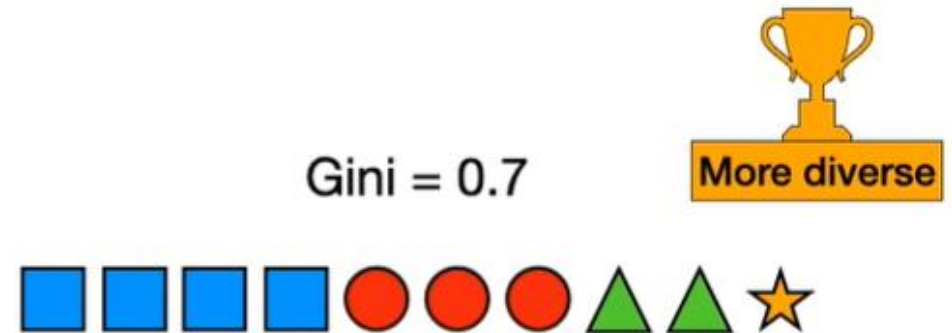
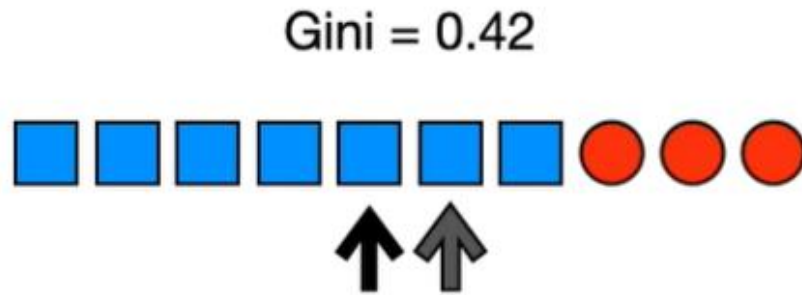
Gini = 0.7



More diverse

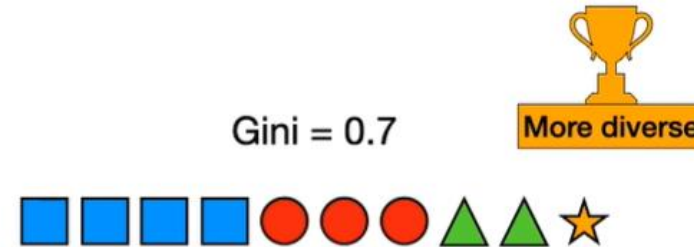
# How to calculate Gini index?





















- Randomly pick any two elements from the dataset and determine if they are similar or different.



# How to calculate Gini index?

- Repeat the process a number of times.



		Same
		<b>Different</b>
		<b>Different</b>
		Same
		Same
		<b>Different</b>
		Same
		Same
		<b>Different</b>
		Same

Different:  
4 out of 10

# How to calculate Gini index?

- Do the same process for the second dataset.

Gini = 0.42



Gini = 0.7



Gini Index = Probability of picking two distinct elements

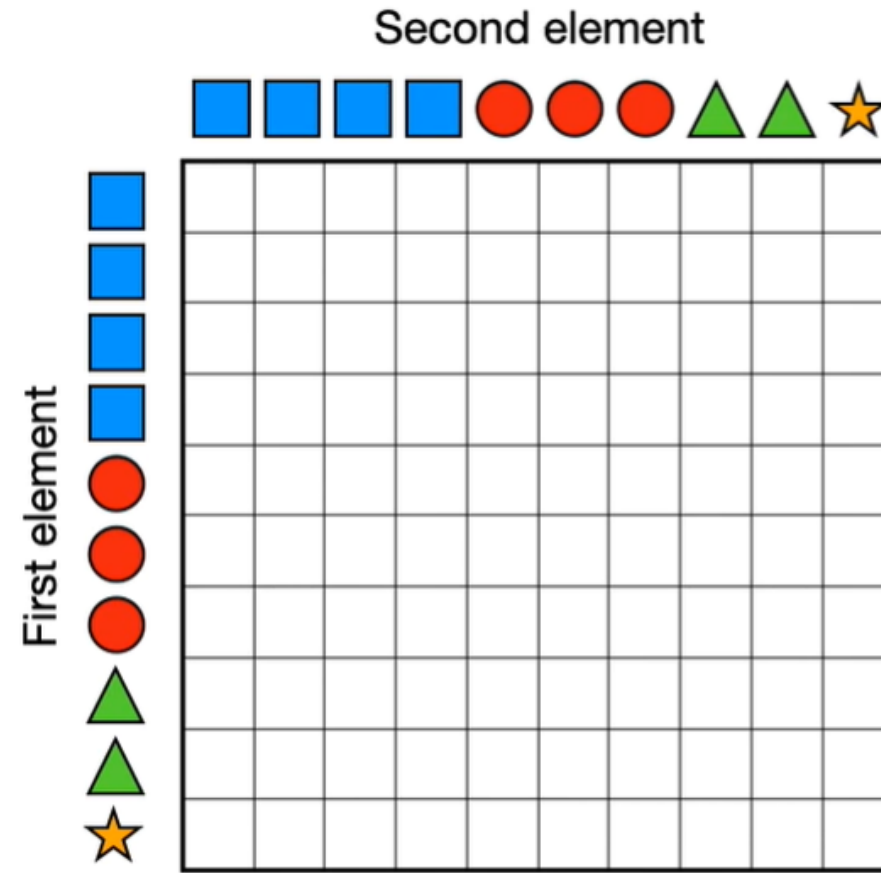
		Same
		<b>Different</b>
		<b>Different</b>
		Same
		Same
		<b>Different</b>
		Same
		Same
		<b>Different</b>
		Same

Different:  
4 out of 10

		<b>Different</b>
		Same
		<b>Different</b>
		<b>Different</b>
		<b>Different</b>
		Same
		Same
		<b>Different</b>
		<b>Different</b>
		<b>Different</b>

Different:  
7 out of 10

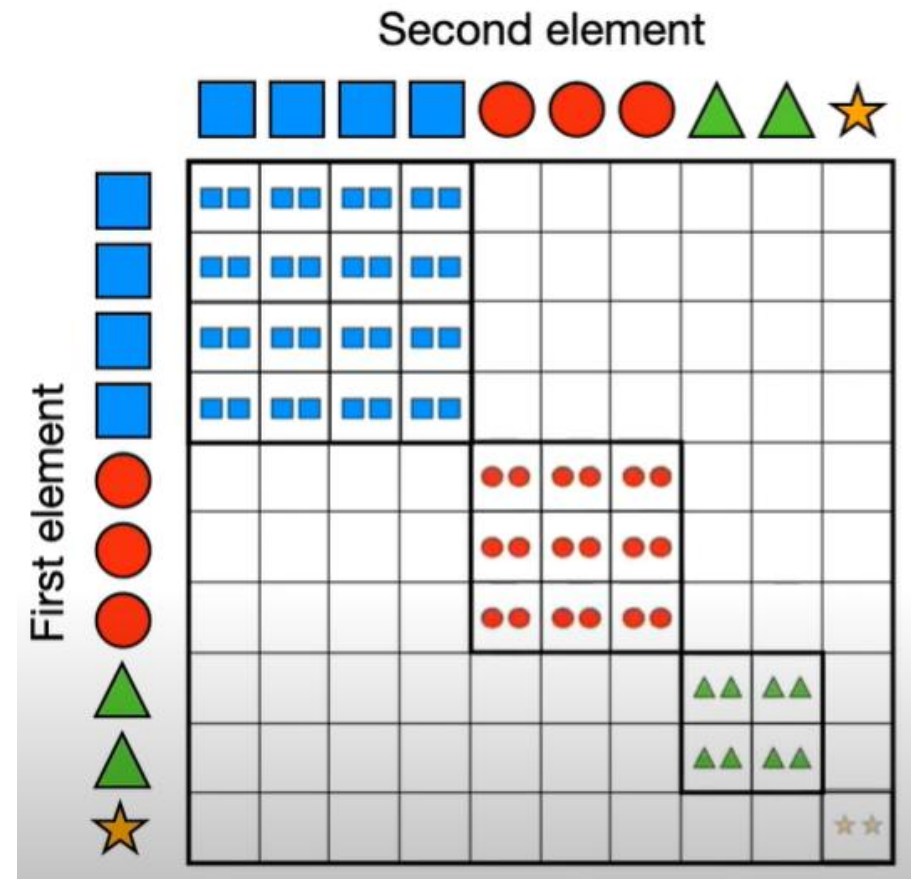
# How to calculate Gini index?





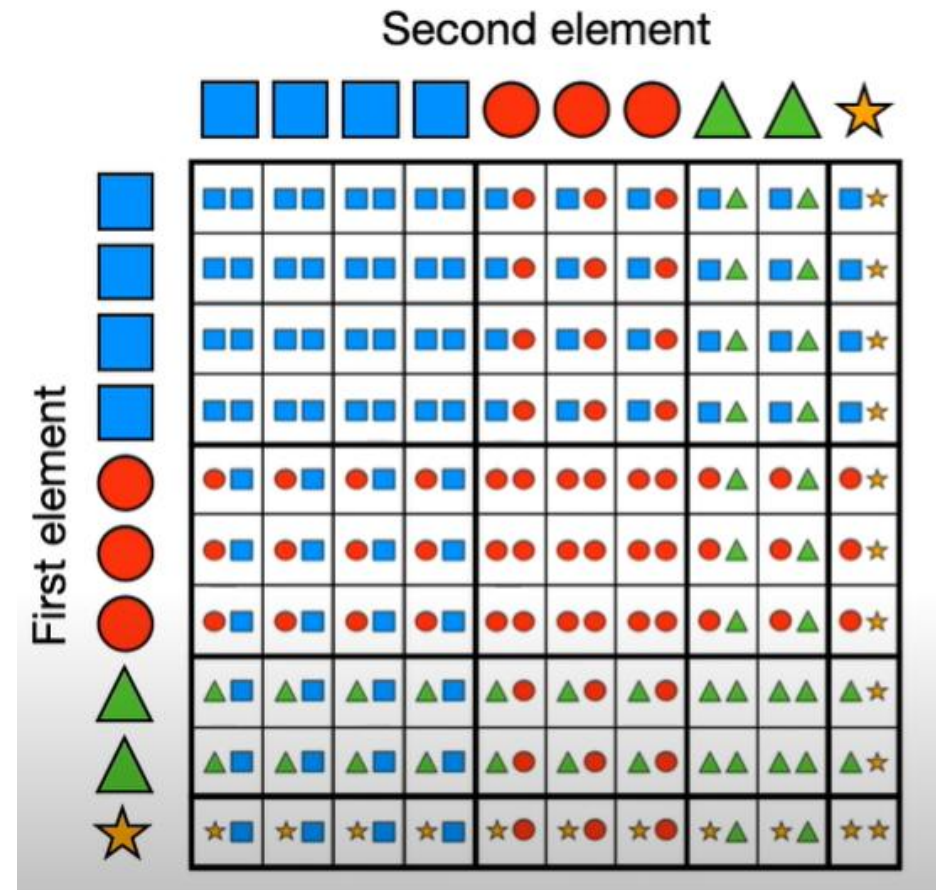
# How to calculate Gini index?

- Let's start with the second dataset.
- Enumerate all the possible combinations.



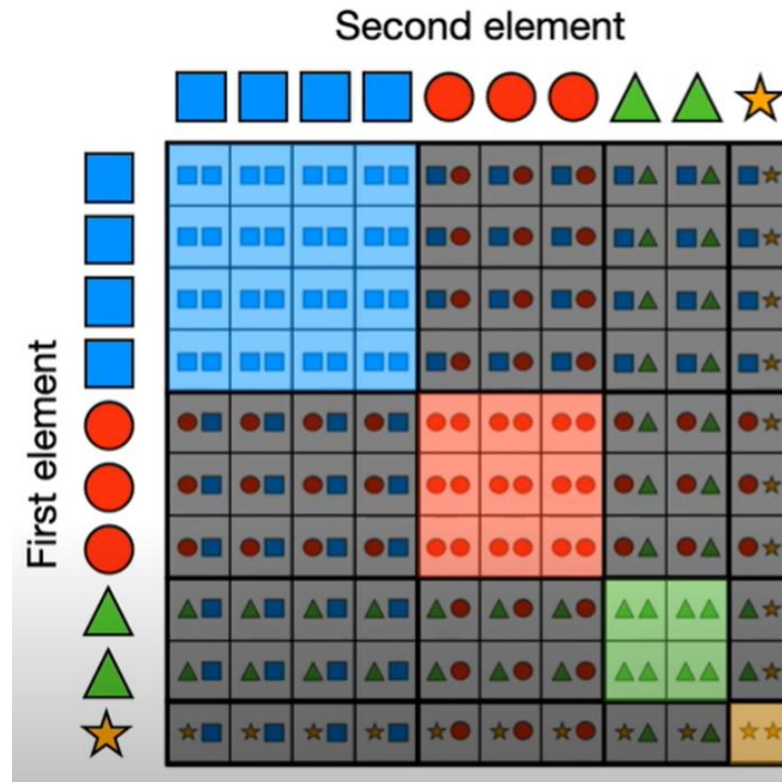
# How to calculate Gini index?

- Let's start with the second dataset.
- Enumerate all the possible combinations.



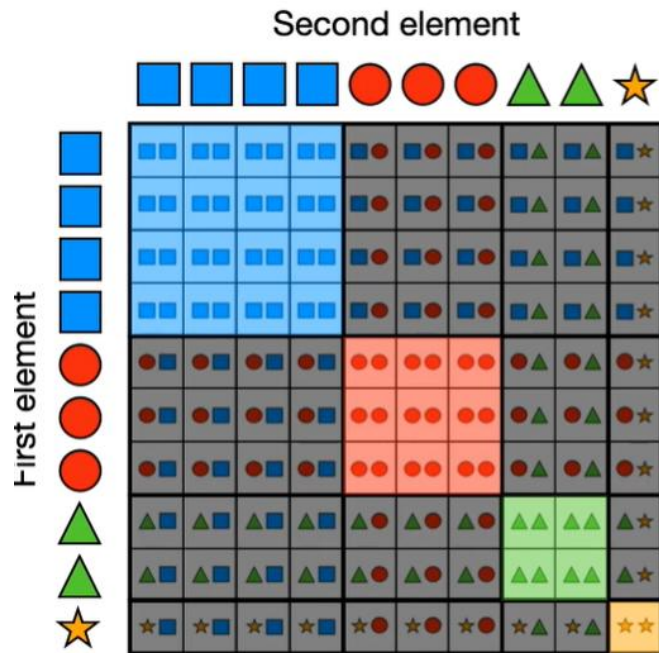
# How to calculate Gini index?

- We need to calculate the percentage of the black area (When both elements are different).



# How to calculate Gini index?

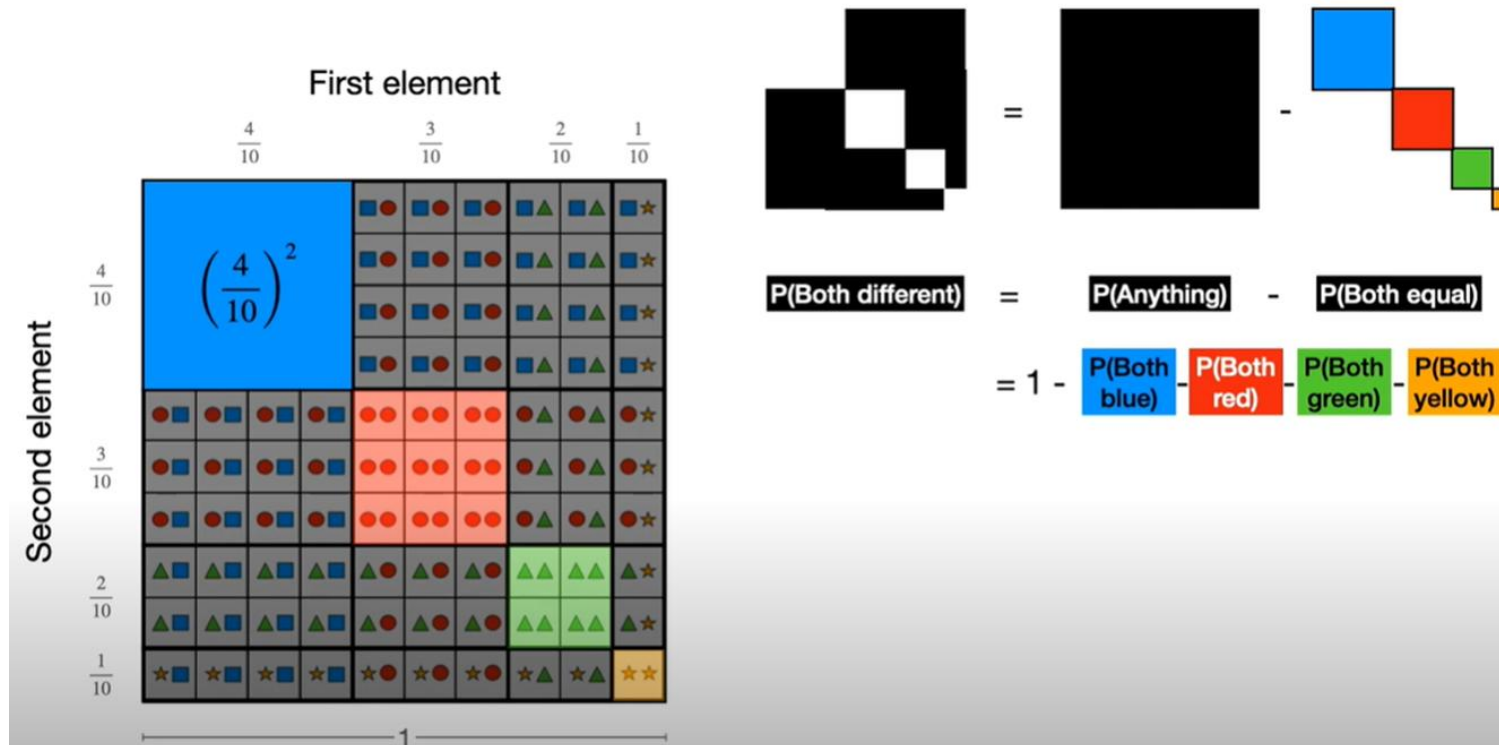
- We need to calculate the percentage of the black area (When both elements are different).



$$\begin{aligned}
 \text{P(Both different)} &= \text{P(Anything)} - \text{P(Both equal)} \\
 &= 1 - \text{P(Both blue)} - \text{P(Both red)} - \text{P(Both green)} - \text{P(Both yellow)}
 \end{aligned}$$

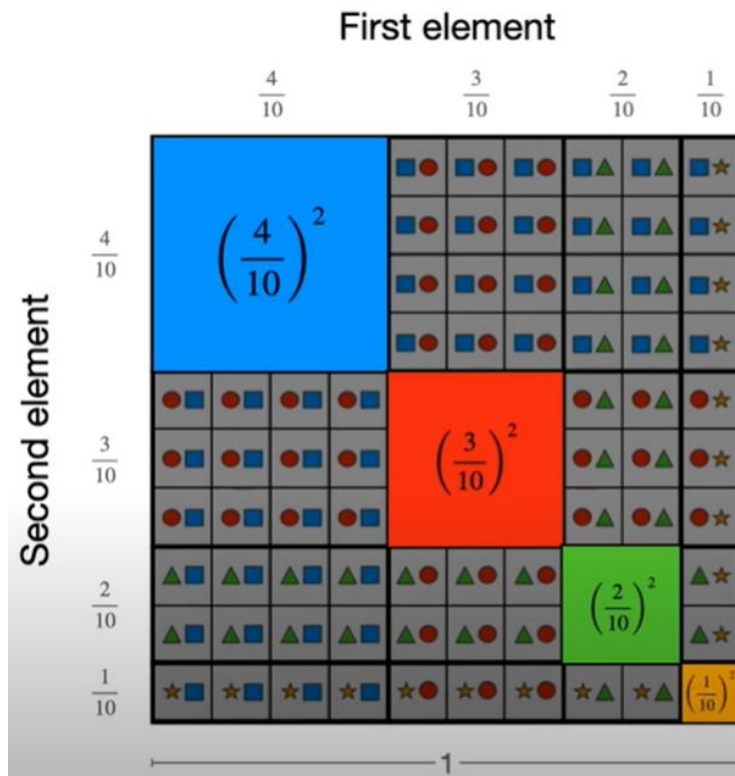
# How to calculate Gini index?

- We need to calculate the percentage of the black area (When both elements are different).



# How to calculate Gini index?

- We need to calculate the black area.

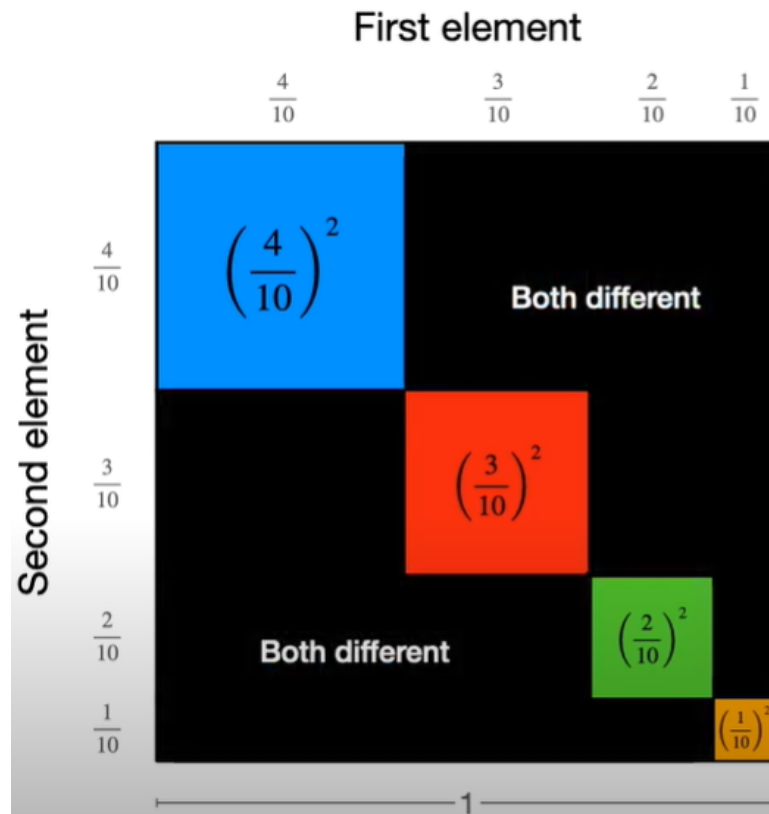


$$P(\text{Both different}) = P(\text{Anything}) - P(\text{Both equal})$$

$$= 1 - P(\text{Both blue}) - P(\text{Both red}) - P(\text{Both green}) - P(\text{Both yellow})$$

# How to calculate Gini index?

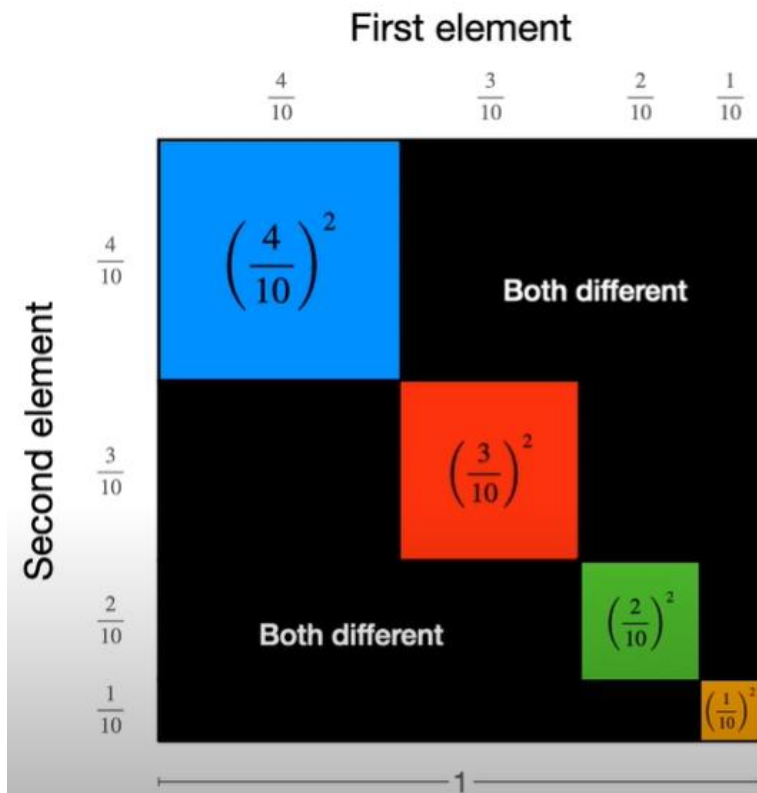
- We need to calculate the black area



$$\begin{aligned} \text{P(Both different)} &= \text{P(Anything)} - \text{P(Both equal)} \\ &= 1 - \text{P(Both blue)} - \text{P(Both red)} - \text{P(Both green)} - \text{P(Both yellow)} \end{aligned}$$

# How to calculate Gini index?

- We need to calculate the black area

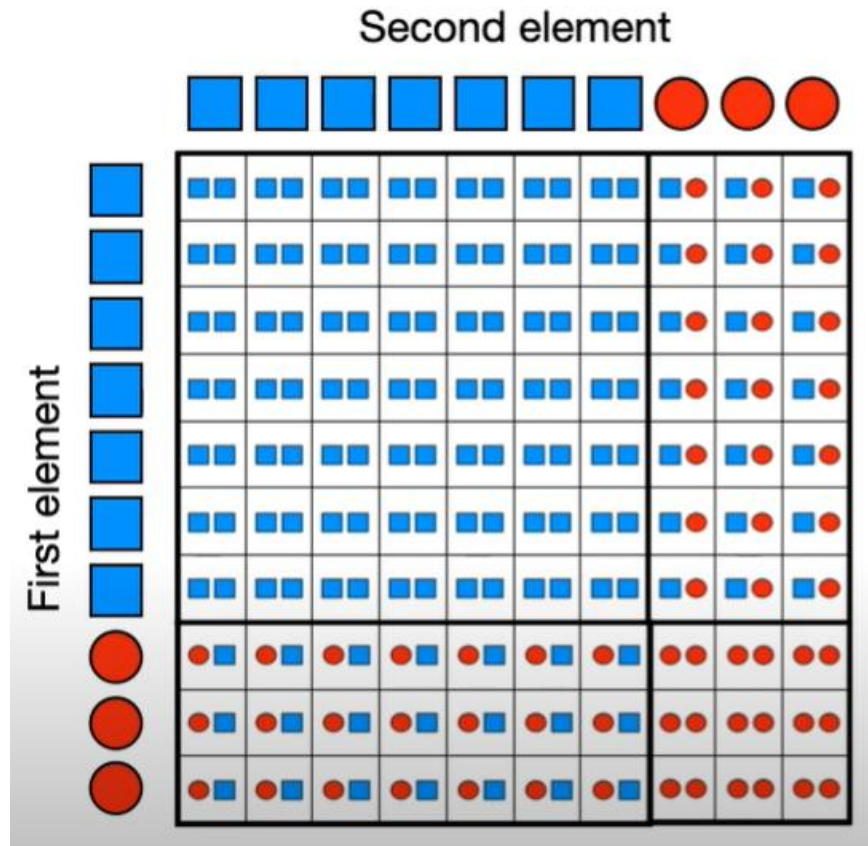


$$\begin{aligned} P(\text{Both different}) &= P(\text{Anything}) - P(\text{Both equal}) \\ &= 1 - P(\text{Both blue}) - P(\text{Both red}) - P(\text{Both green}) - P(\text{Both yellow}) \\ &= 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{2}{10}\right)^2 - \left(\frac{1}{10}\right)^2 \\ &= 1 - 0.4^2 - 0.3^2 - 0.2^2 - 0.1^2 \\ &= 1 - 0.16 - 0.09 - 0.04 - 0.01 \\ &= 0.70 \end{aligned}$$



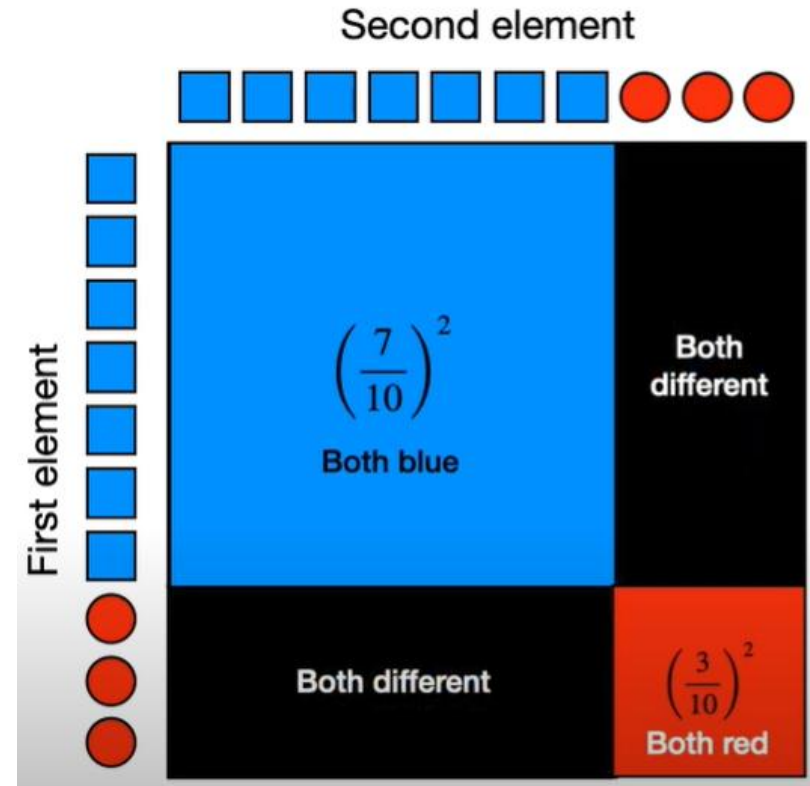
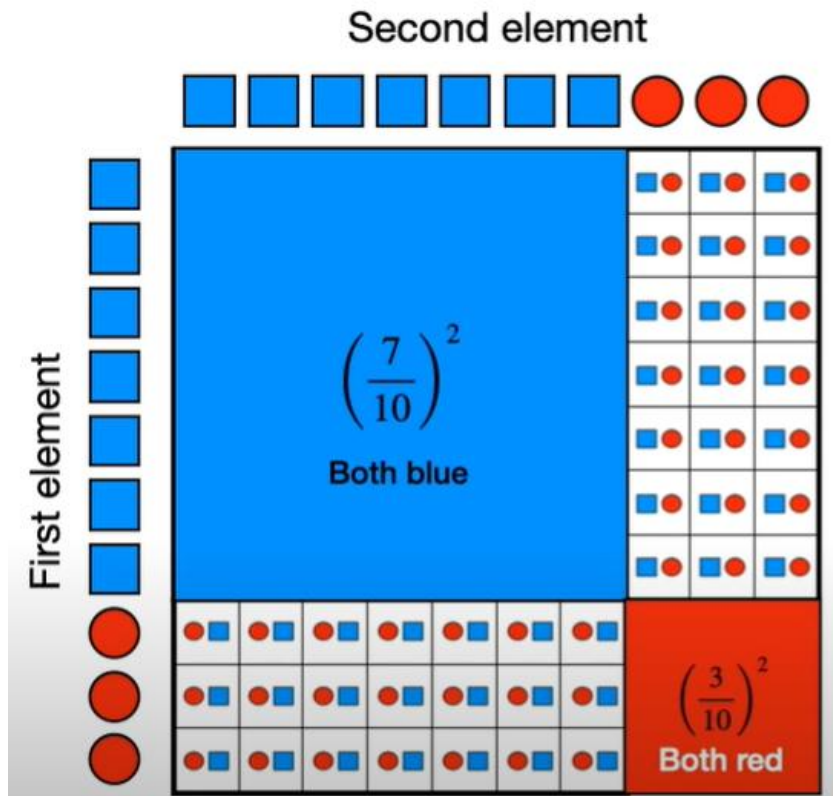
# How to calculate Gini index?

- Let's calculate the Gini index for the first dataset.



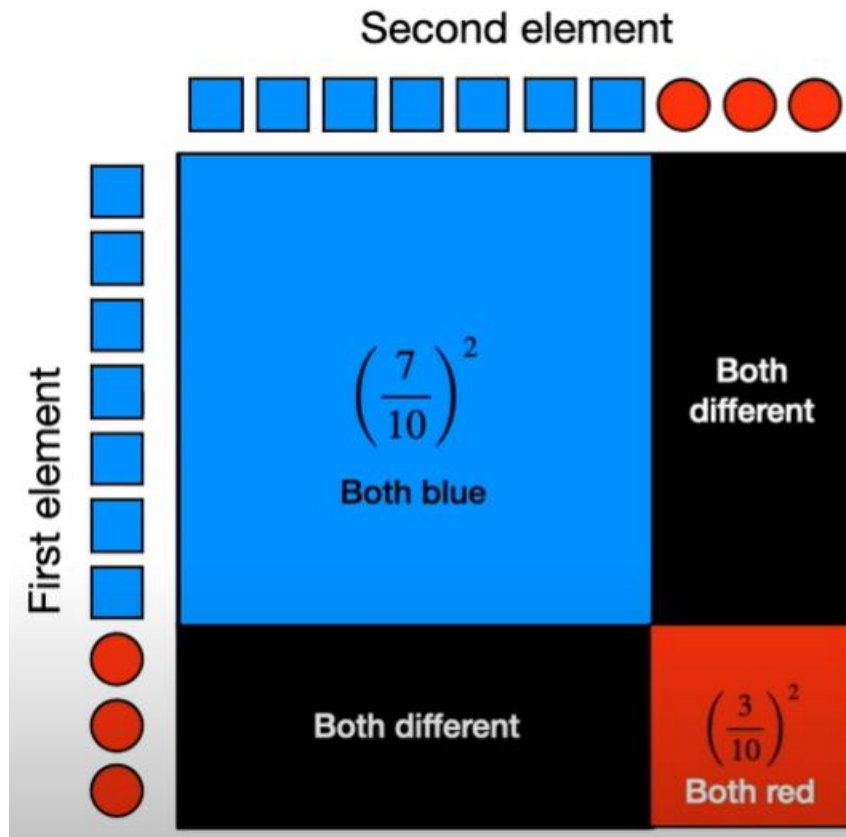
# How to calculate Gini index?

- Let's calculate the Gini index for the first dataset.



# How to calculate Gini index?

- Let's calculate the Gini index for the first dataset.



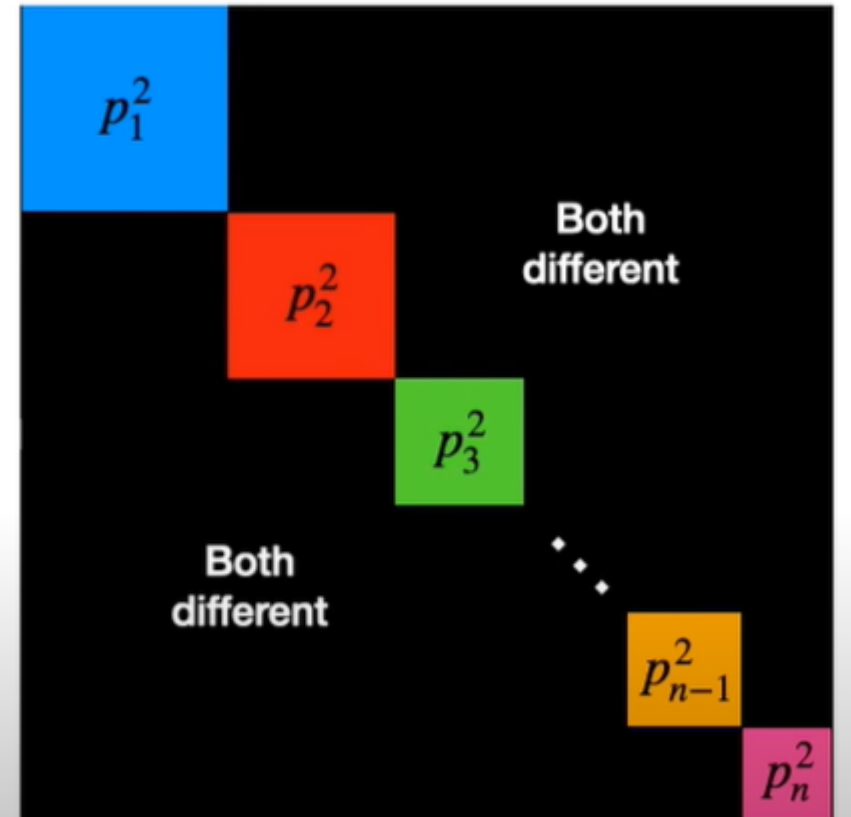
$$\begin{aligned} P(\text{Both different}) &= P(\text{Anything}) - P(\text{Both equal}) \\ &= 1 - P(\text{Both blue}) - P(\text{Both red}) \\ &= 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 \\ &= 1 - 0.7^2 - 0.3^2 \\ &= 1 - 0.49 - 0.09 \\ &= 0.42 \end{aligned}$$

# The General Formula

n classes

Proportions:  $p_1, p_2, \dots, p_n$

Gini impurity index:  $1 - p_1^2 - p_2^2 - \dots - p_n^2$



# The General Formula

n classes

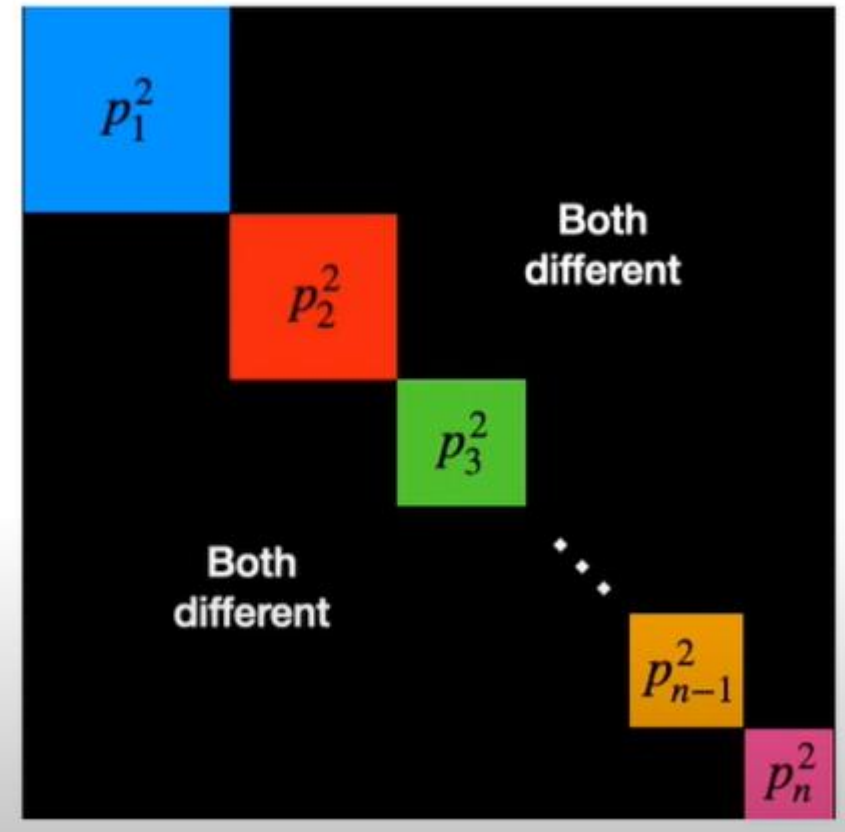
Proportions:  $p_1, p_2, \dots, p_n$

Gini impurity index:  $1 - p_1^2 - p_2^2 - \dots - p_n^2$

**P(Both different)**

**P(Anything)**

**P(Both equal)**



# Gini index examples, all similar or all different



Gini = 0.42



Gini = 0.7



Gini = 0

$$1 - 1^2 = 0$$



Gini award for the  
more diverse dataset



Gini = ?

$$1 - 0.1^2 - 0.1^2 - \dots - 0.1^2 = 0.9$$

10 times

Note: The Gini index can not be one, because its always  $1 - \text{the sum of squares}$ , unless the dataset is infinite.

# References

- <https://www.youtube.com/watch?v=u4IxOk2ijSs>