AI ALIGNMENT COHORT

Session 1

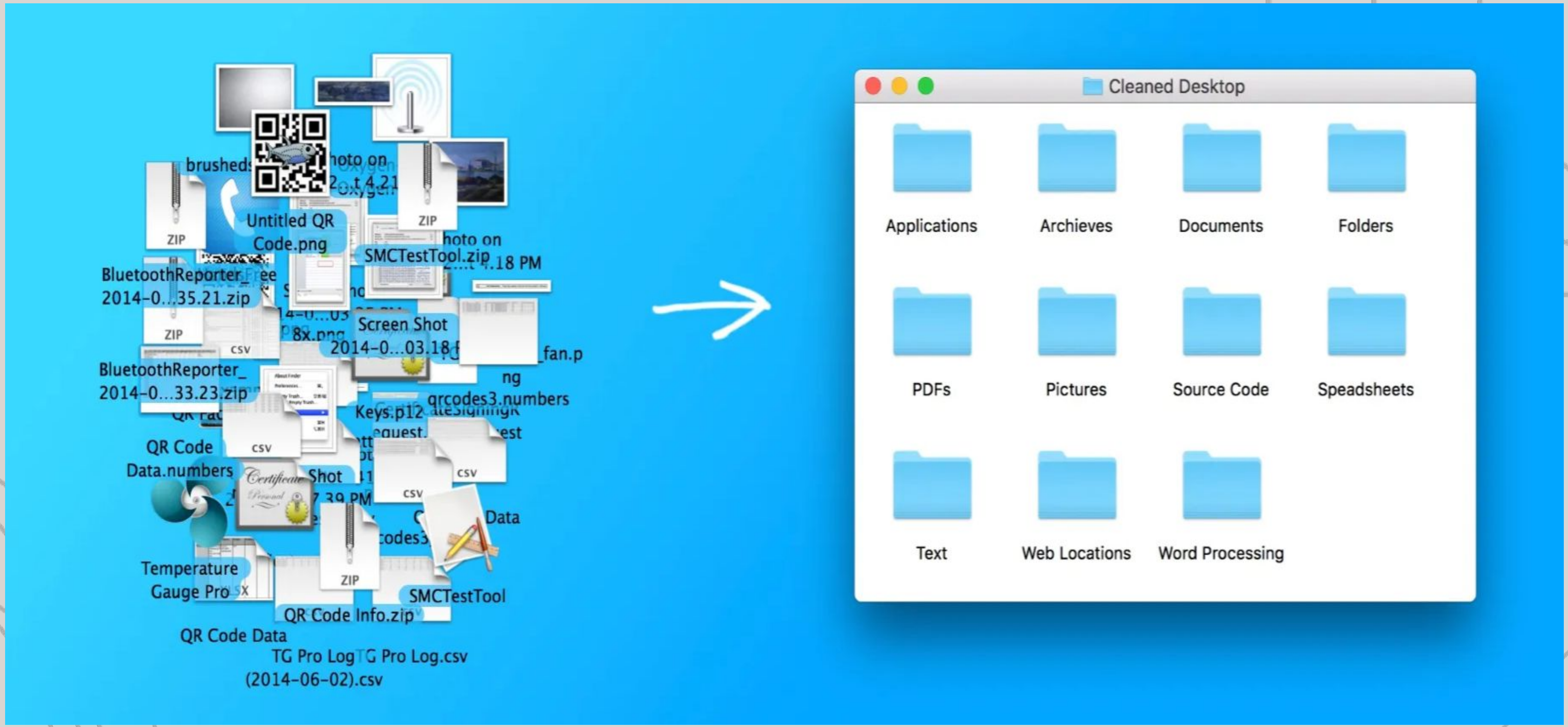
Linear Algebra

&

Information Theory

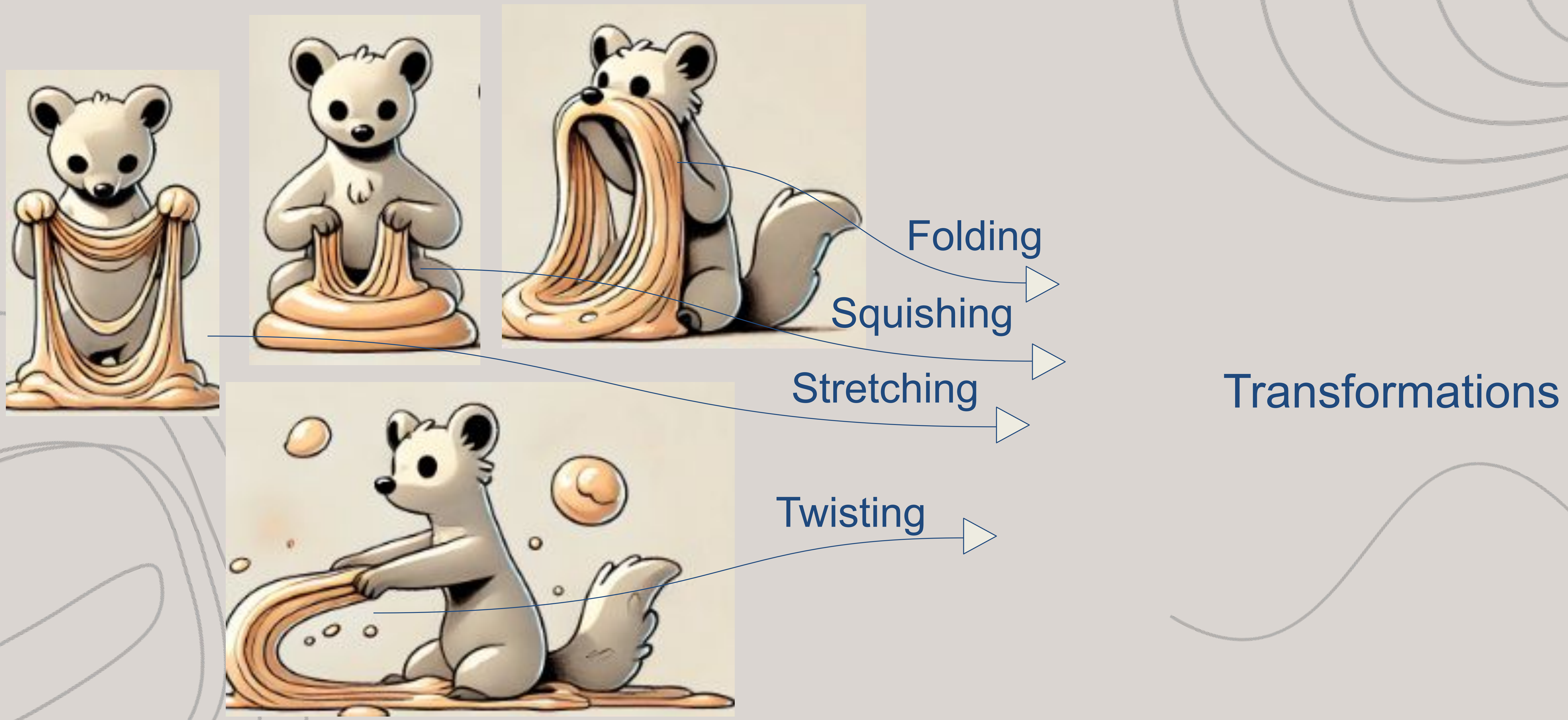
Cohere for AI
BIRDS x Safety & Alignment Group

Why Bother?

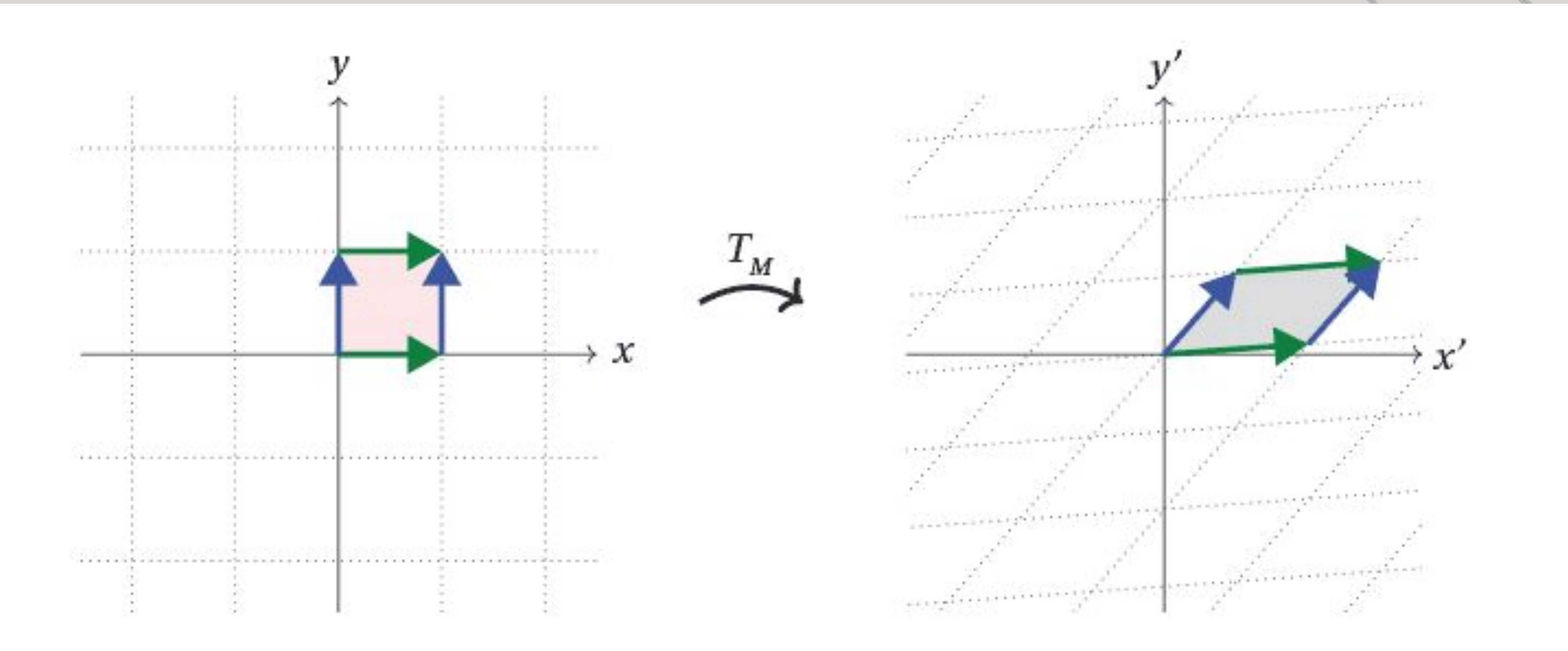


Cleaned Desktop
Applications
Archieves
Documents
Folders
PDFs
Pictures
Source Code
Speadsheets
Text
Web Locations
Word Processing

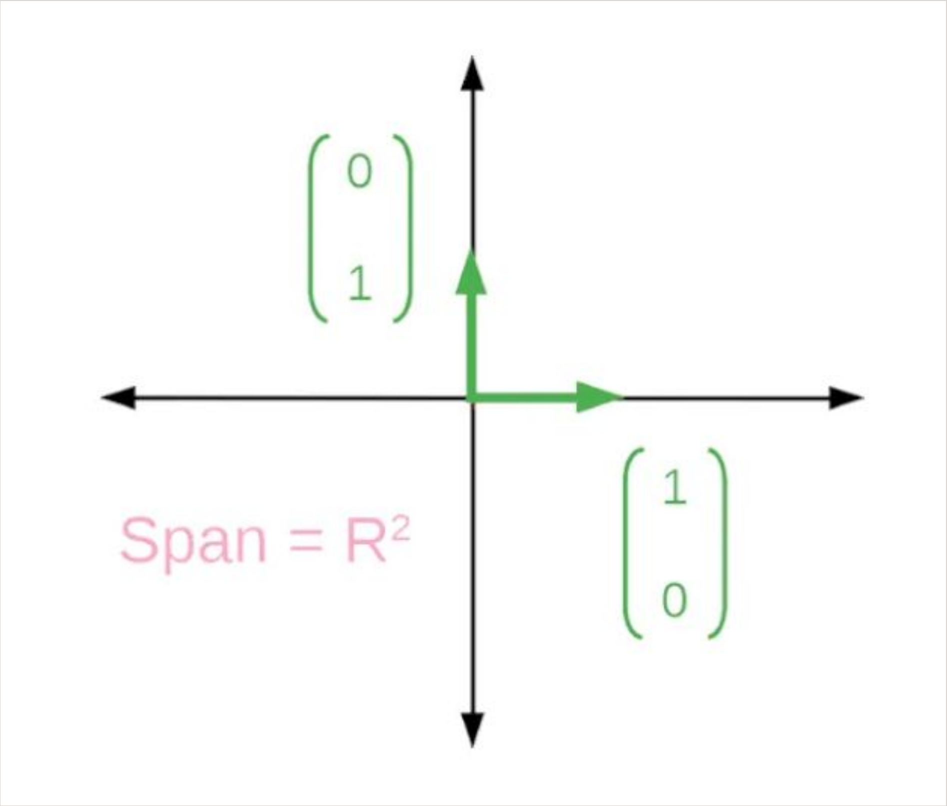
Linear Transformations

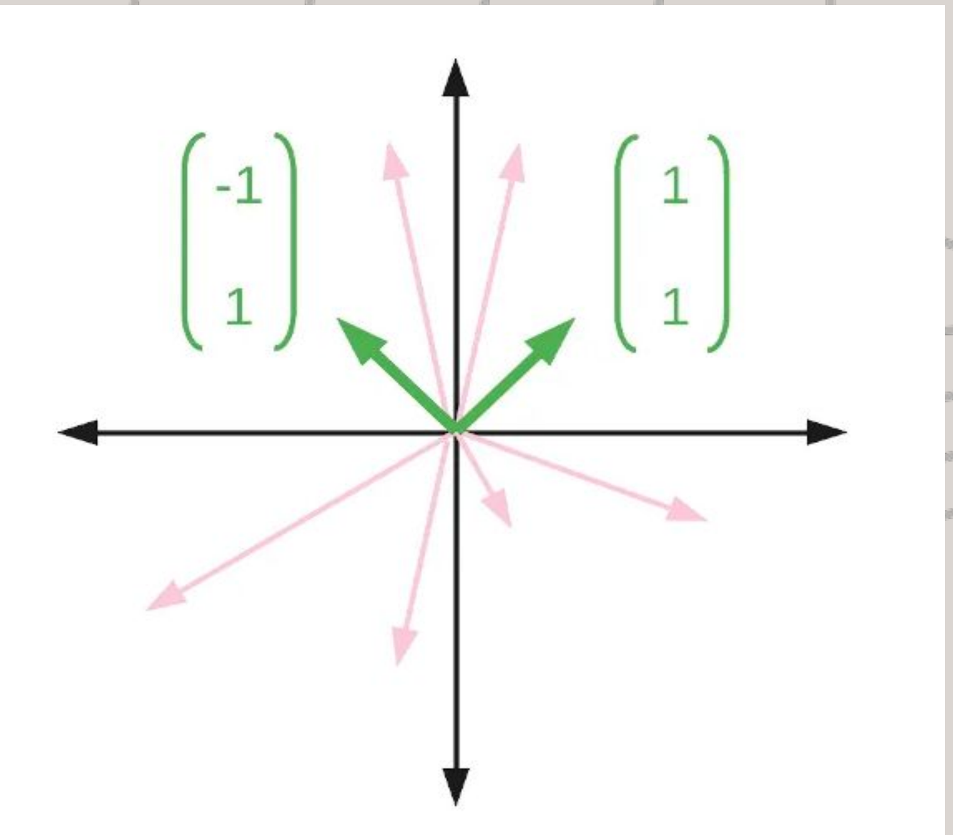


Linear Transformations



Basis



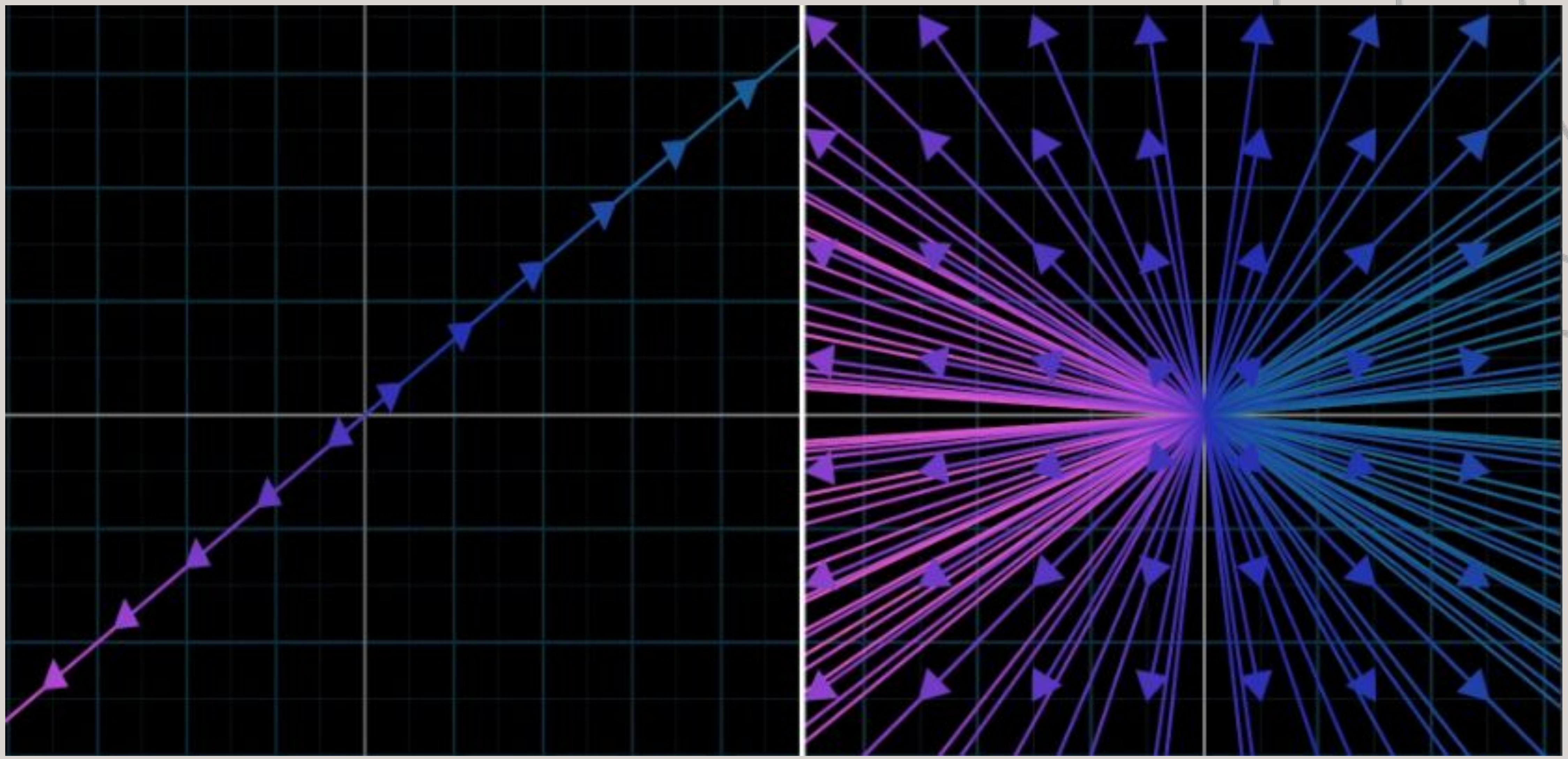


Basis: A set of n vectors, $\{v_1, v_2, \ldots v_n\}$, is a basis of some space S if these two conditions are true:

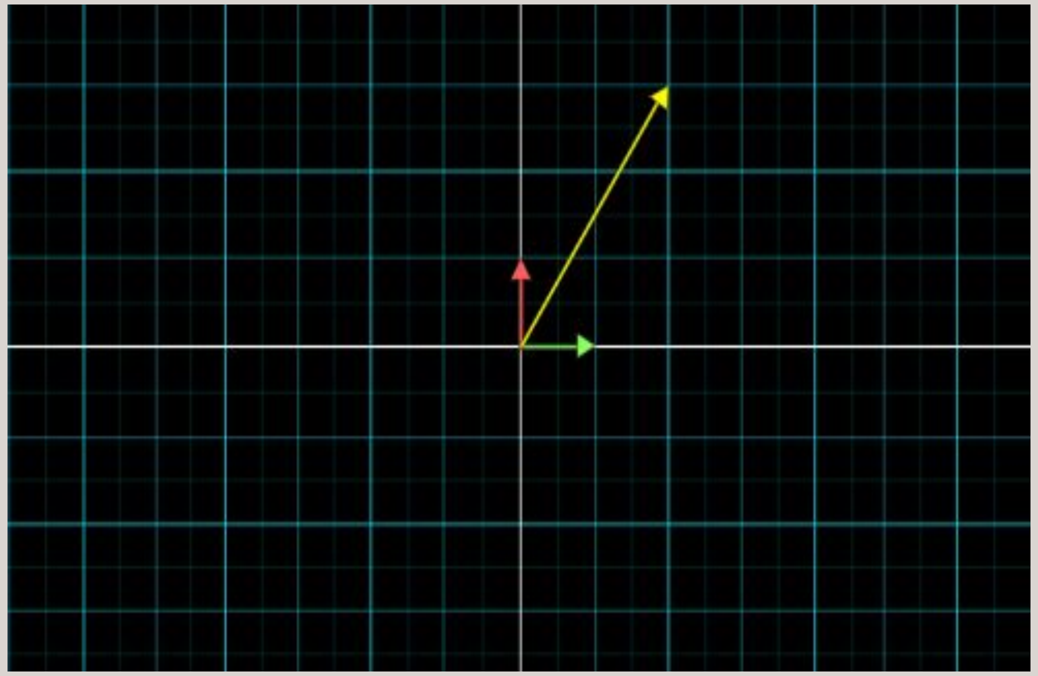
1. $\{v_1, v_2, \ldots v_n\}$ are linearly independent

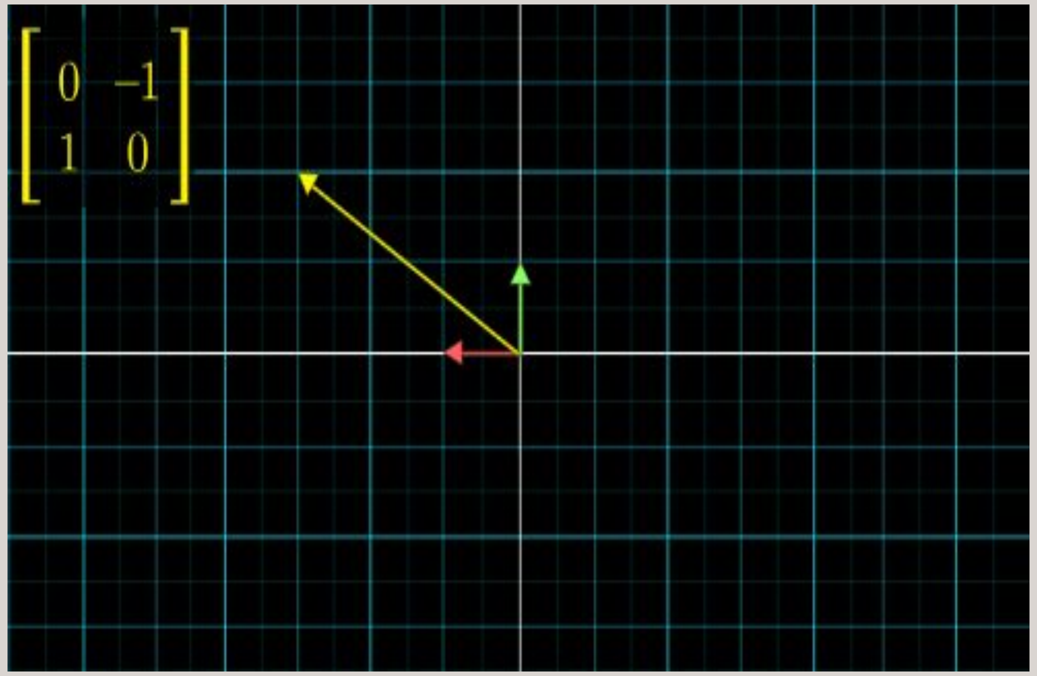
2. $\{v_1, v_2, \ldots v_n\}$ span the set S. In other words, $\text{Span}\{v_1, v_2, \ldots v_n\} = S$

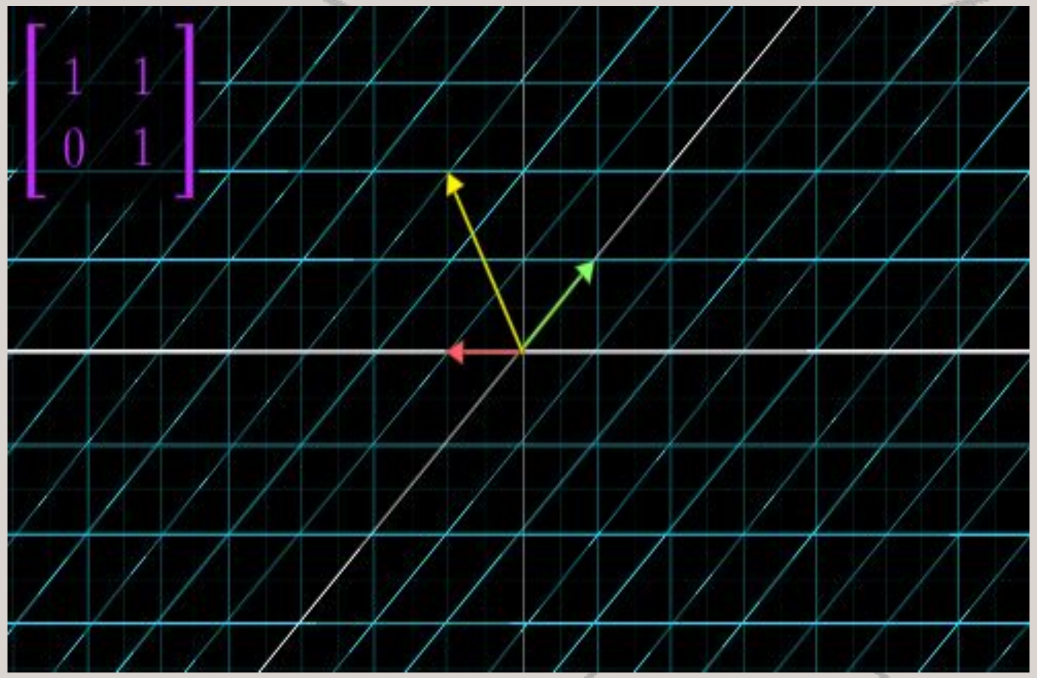
Span



Matrix multiplication as Composition







$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}} \underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Rotation}} = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{Composition}}$$

Rank

= \$7

= \$9

Rank

For example, the matrix A given by

$$A = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}$$

can be put in reduced row-echelon form by using the following elementary row operations:

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{2R_1+R_2 \to R_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{-3R_1+R_3 \to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & -1 & -3 \end{bmatrix}$$

$$\xrightarrow{R_2+R_3 \to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1 \to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

The final matrix (in reduced row echelon form) has two non-zero rows and thus the rank of matrix A is 2.

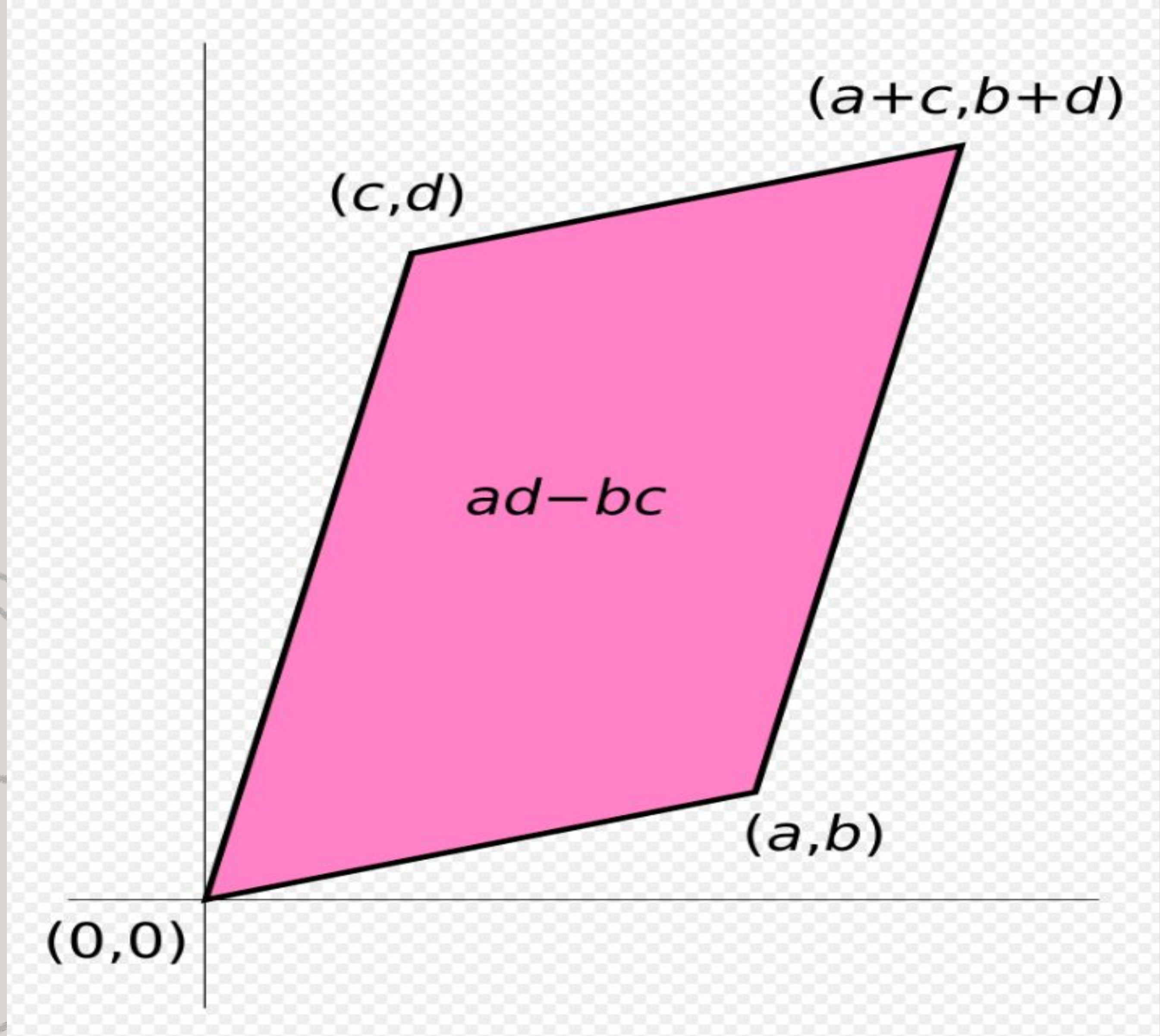
Trace

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$tr(A) = a_{11} + a_{22} + a_{33}$$

Gives Important information about:
Area Scaling Factor
Invertibility
Orientation

Determinant

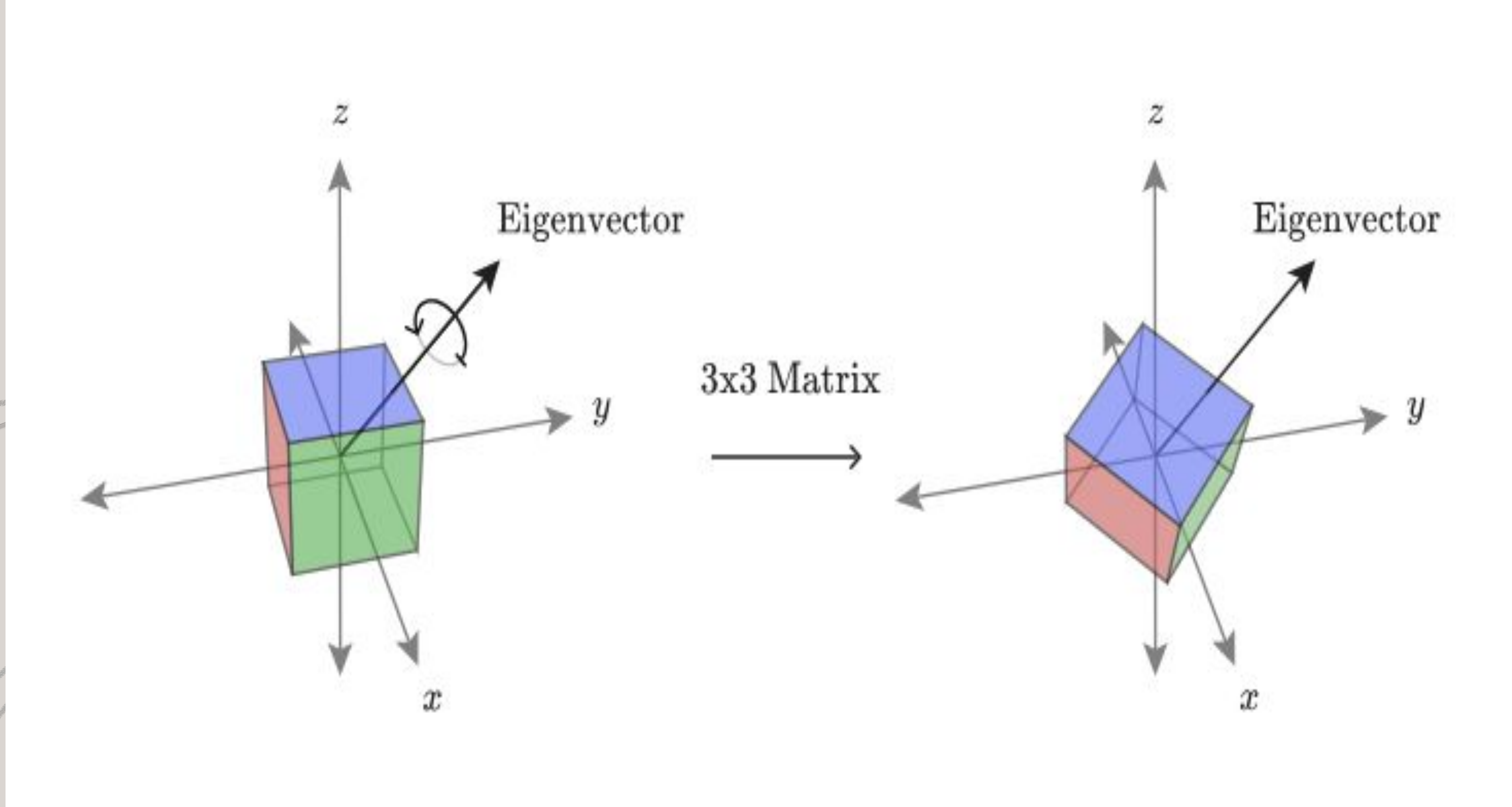


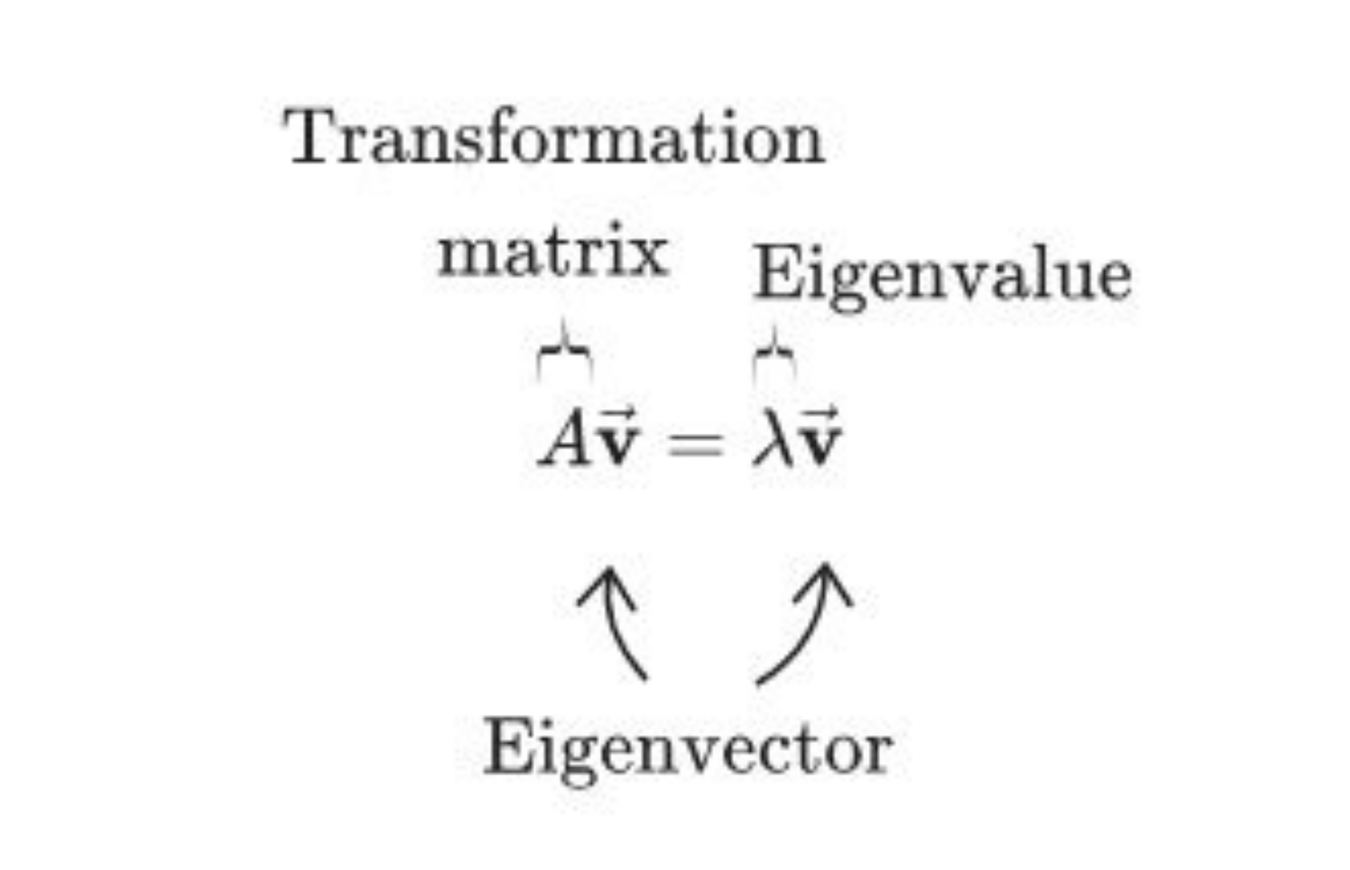
$\det(A) = ad - bc$

Transpose

$$\begin{bmatrix} 2 & 4 & -1 \\ -10 & 5 & 11 \\ 18 & -7 & 6 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & -10 & 18 \\ 4 & 5 & -7 \\ -1 & 11 & 6 \end{bmatrix}$$

Eigenvectors & Eigenvalues





quAIdditch

Why is Trace an important property of Matrices?

Information Theory

What is it?

Digital Communications

Information Theory laid the foundations for digital signal processing and telecommunication protocols, over which all wireless, wireline and satellite networks.

Data Storage

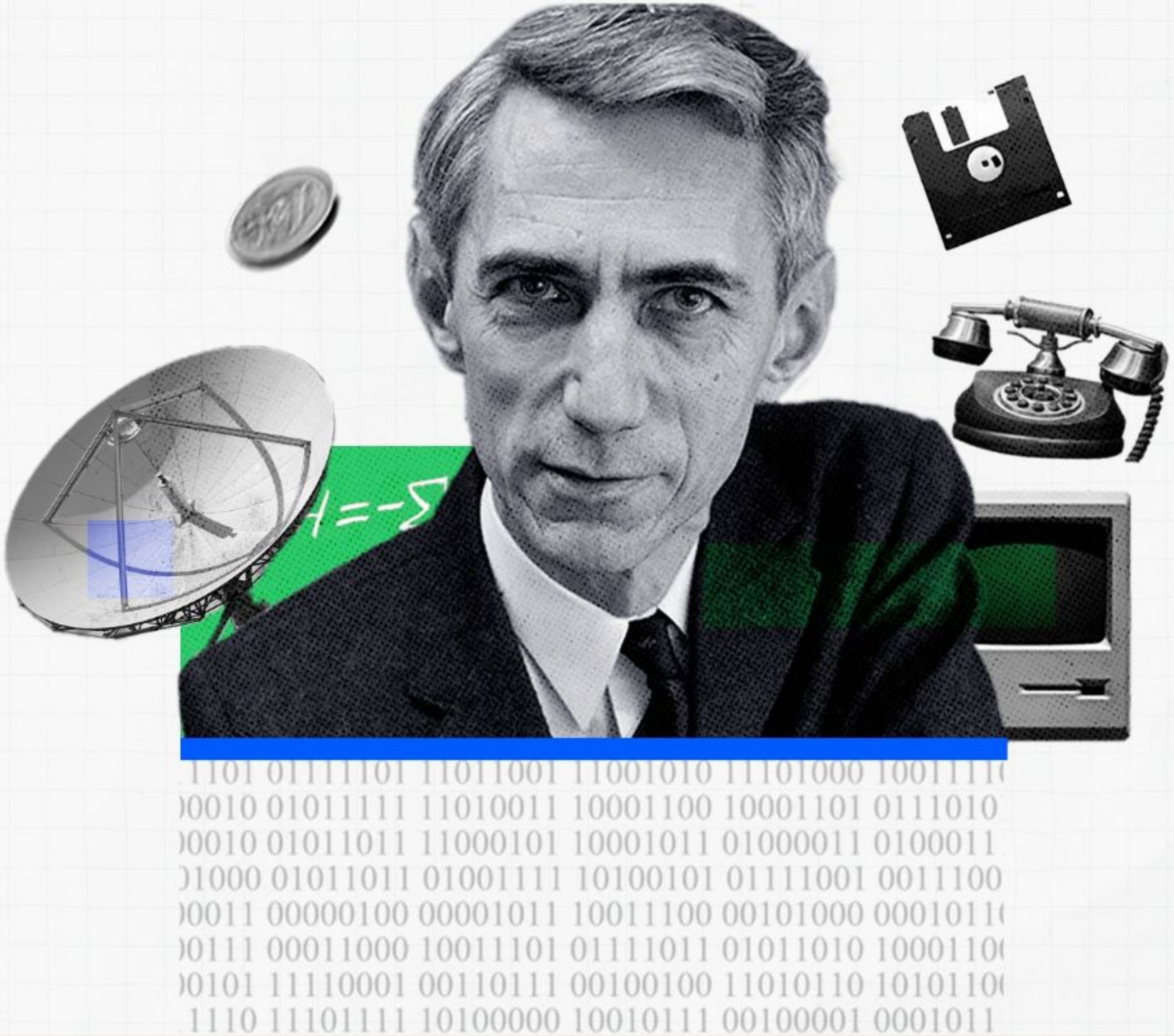
Information Theory is the basis for efficient and compact data encoding and compression, which all digital storage depends on today.

Digital Media

Information Theory defined the principles behind compression algorithms, which allows high-quality media files to be stored and or streamed.

Computing

Binary code – 1s and 0s -- is at the heart of computing systems. Information Theory enabled the processing, storage and retrieval of data in binary form, making modern computing possible.



Internet

Without Shannon, the internet simply would not exist. Information Theory defined the “bit,” the basic unit of digital information on which the internet is built.

AI/ML

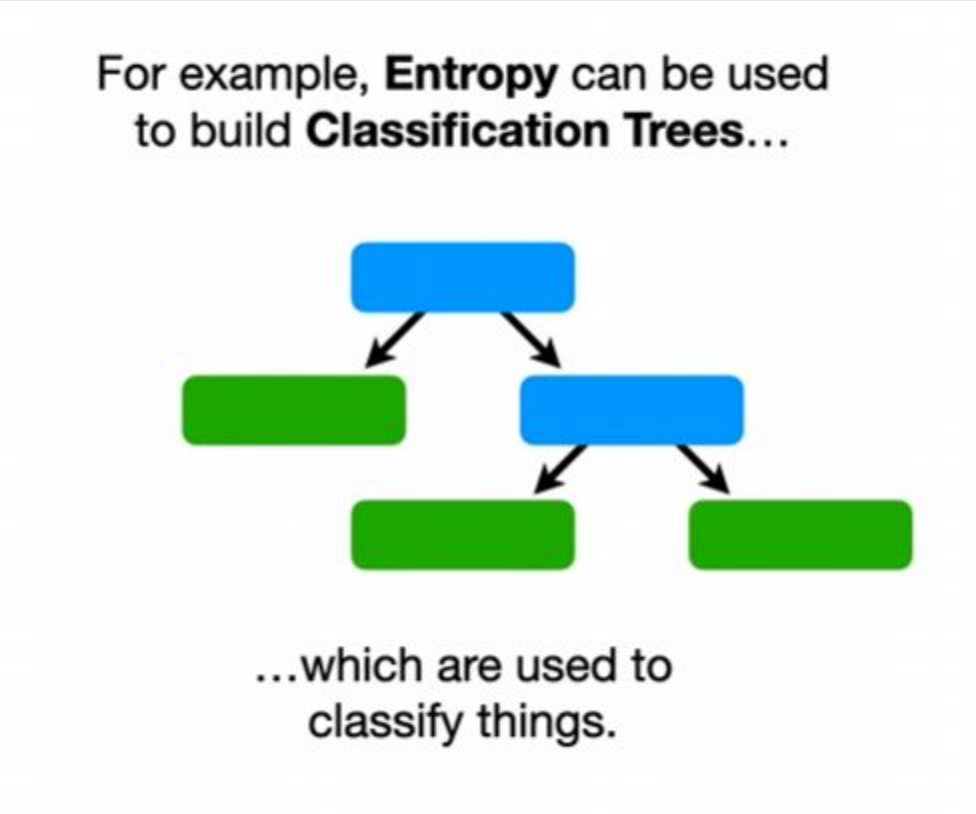
Shannon’s ideas on information gain and entropy are crucial for AI systems’ decision-making processes and in the creation of more accurate AI models.

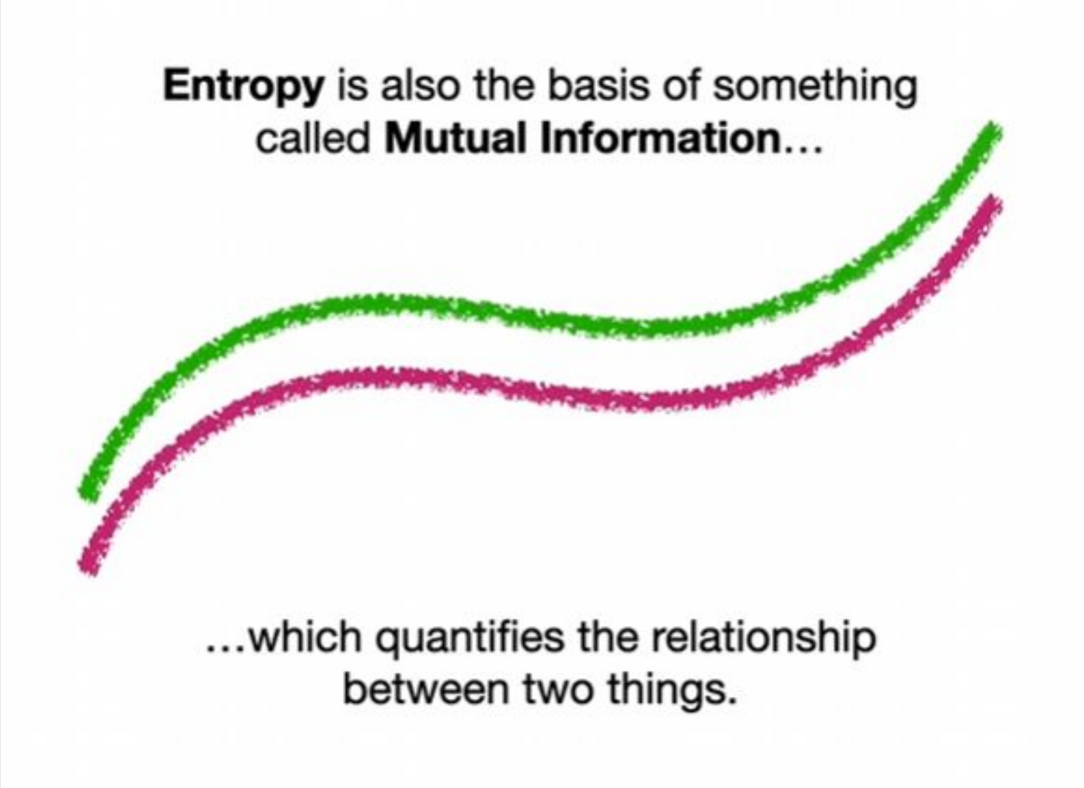
Cryptography

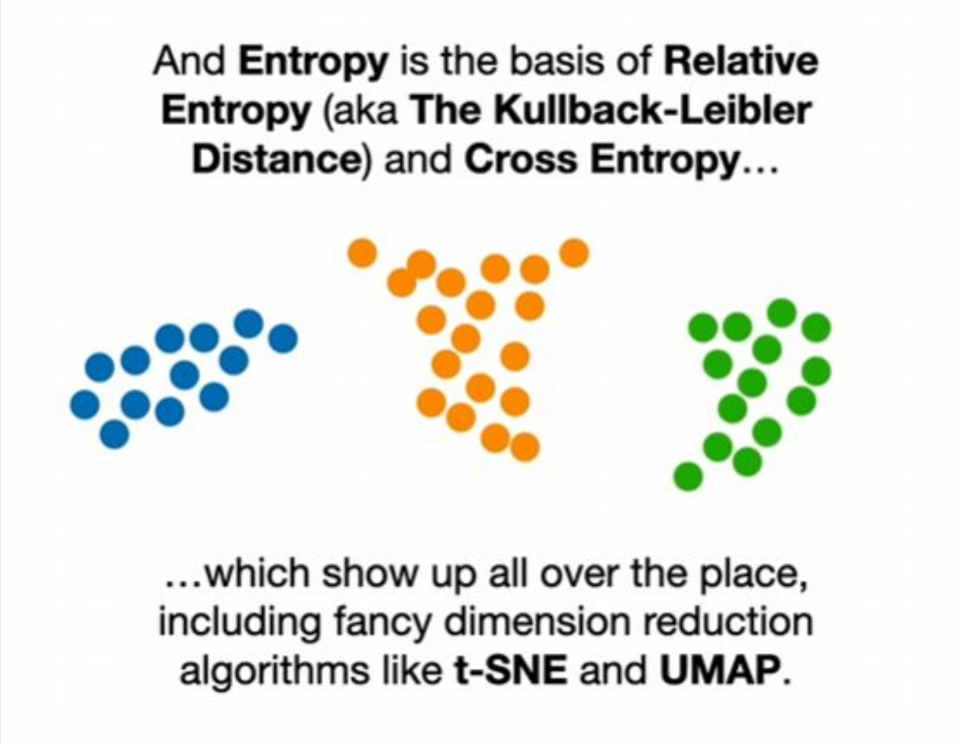
Information Theory established a framework for secure communication by introducing the concepts of randomness and entropy, which are key to modern cryptographic systems and practices.

What is it?

Information theory also drives many algorithms in machine learning.

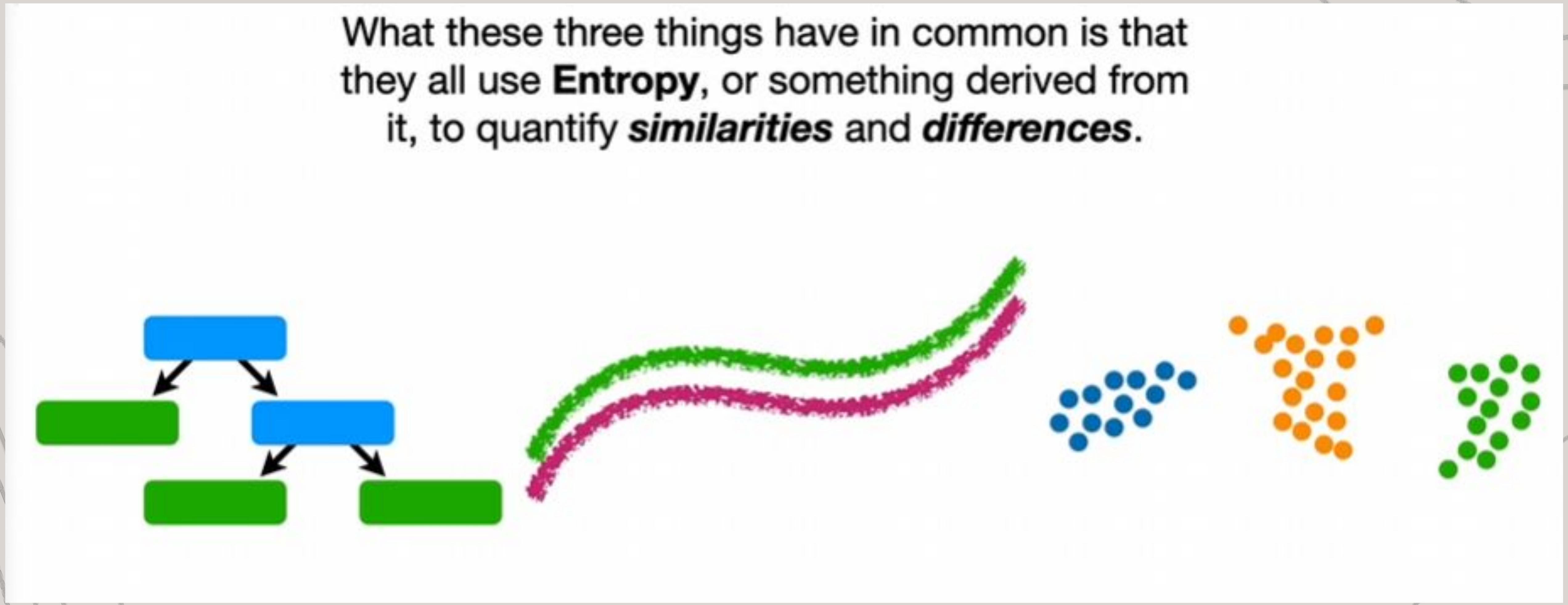






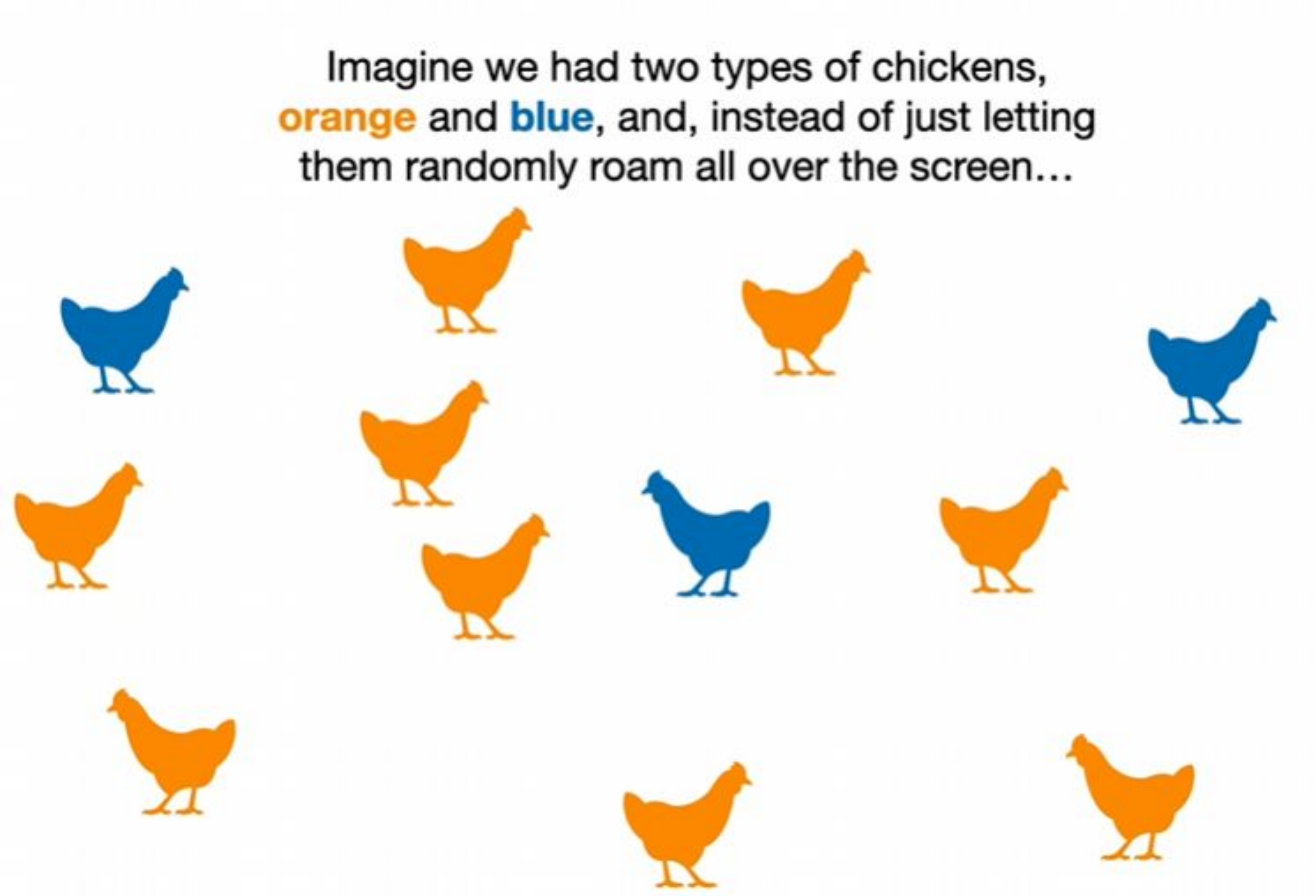
What is it?

At its core, information theory is about quantifying similarities and differences. In this section, we will be explaining the primary metric for doing so.



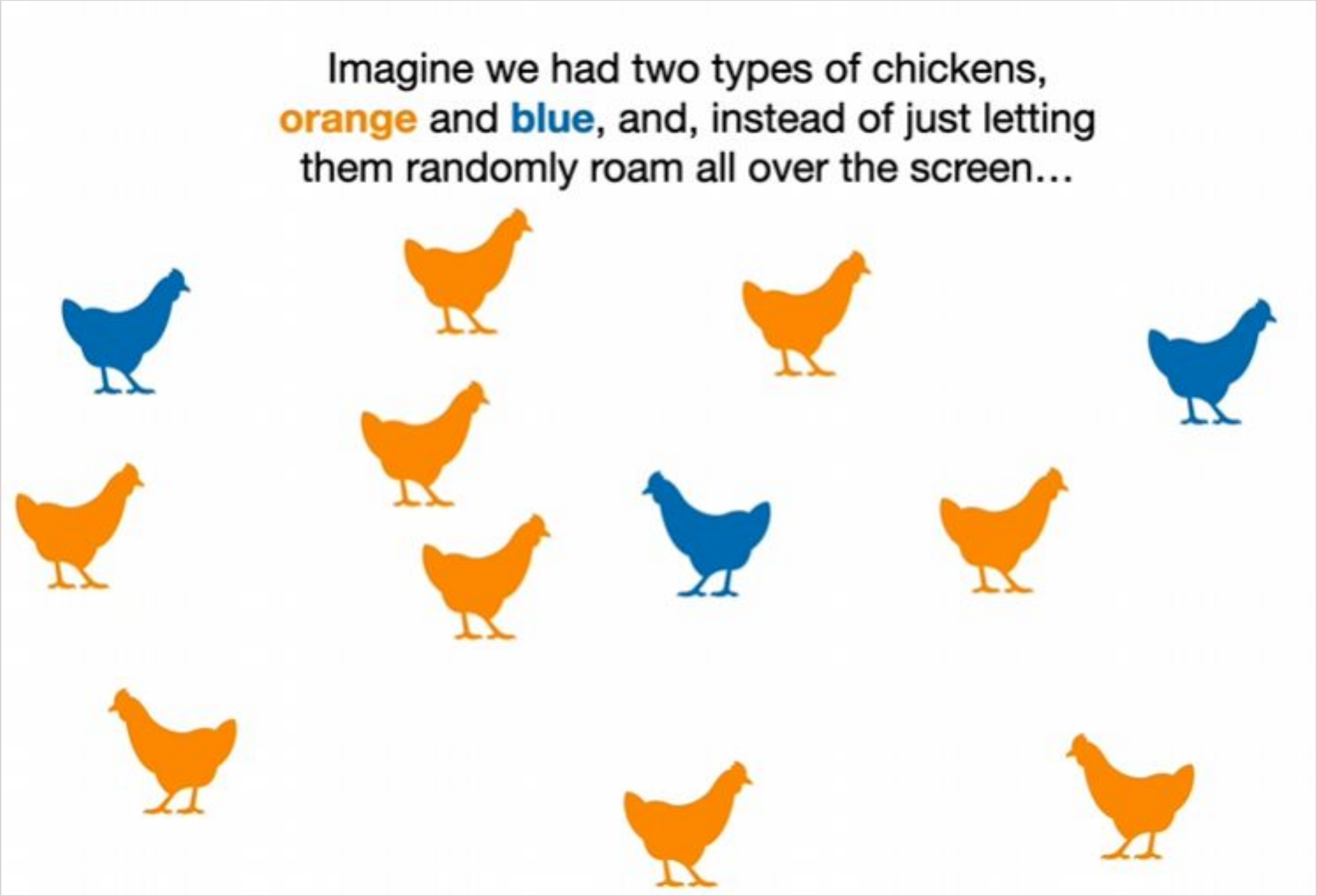
What is surprise?

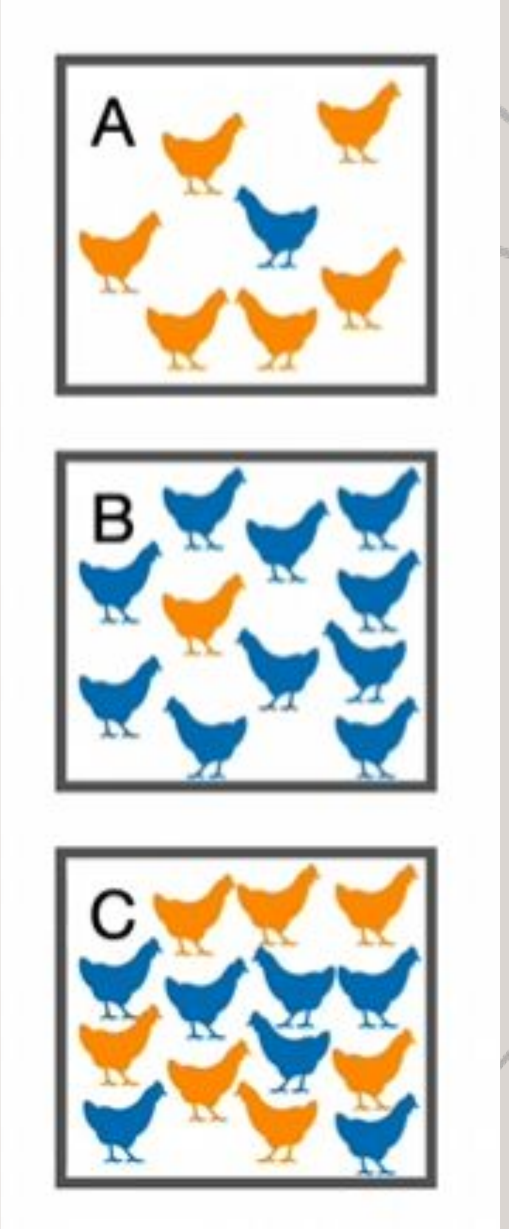
Before we talk about entropy, let's talk about surprise. What is it?



What is surprise?

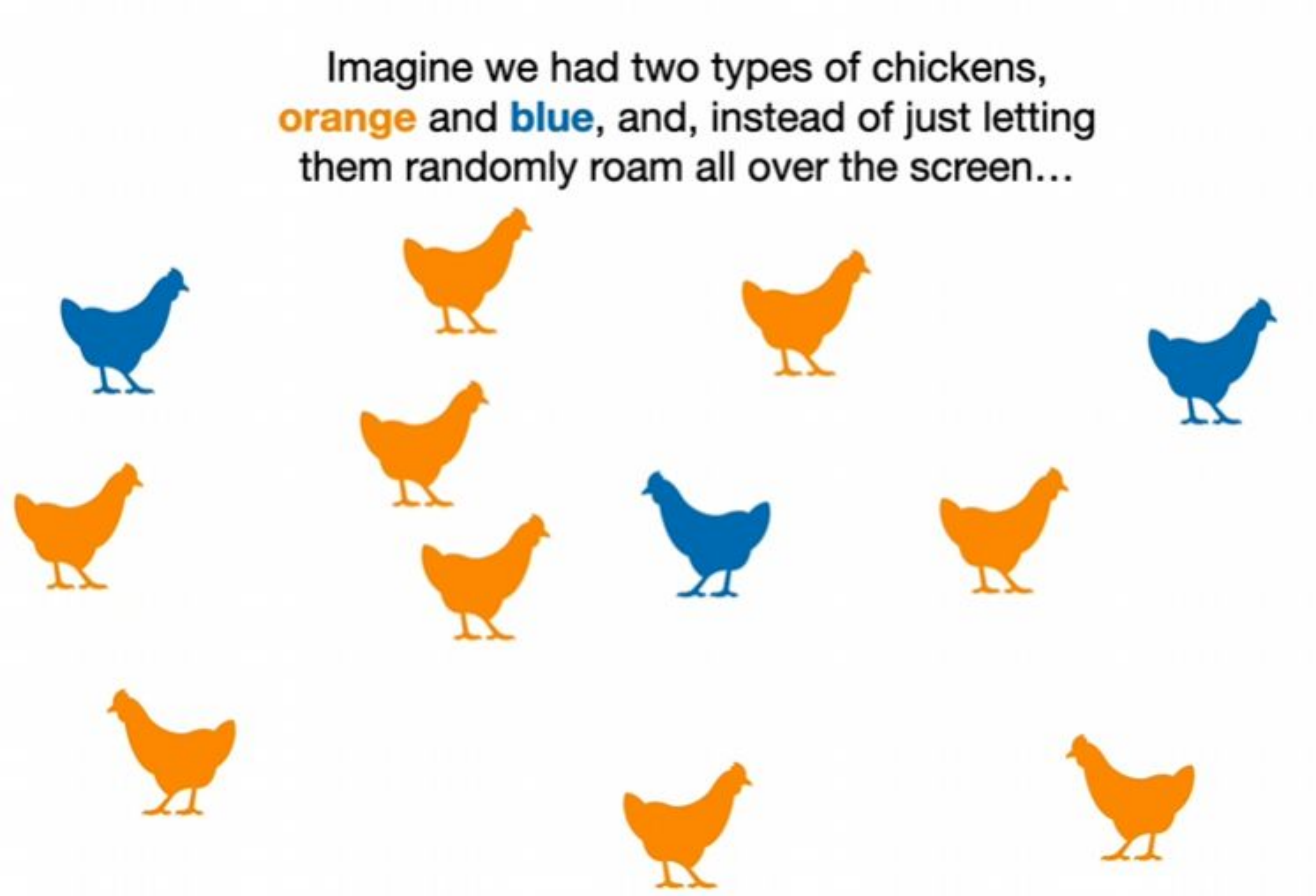
Surprise seems to be closely tied to probabilities...

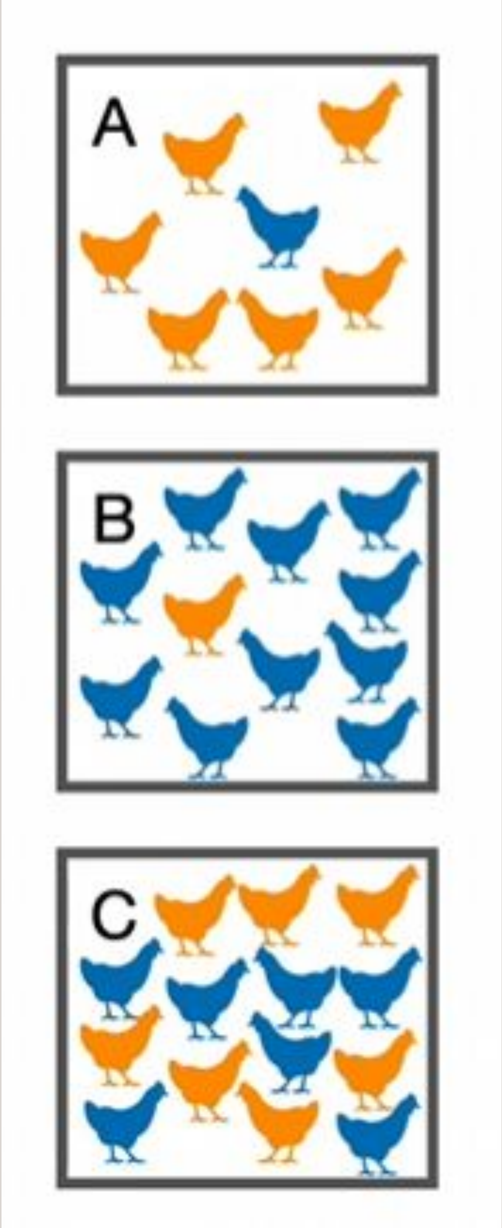


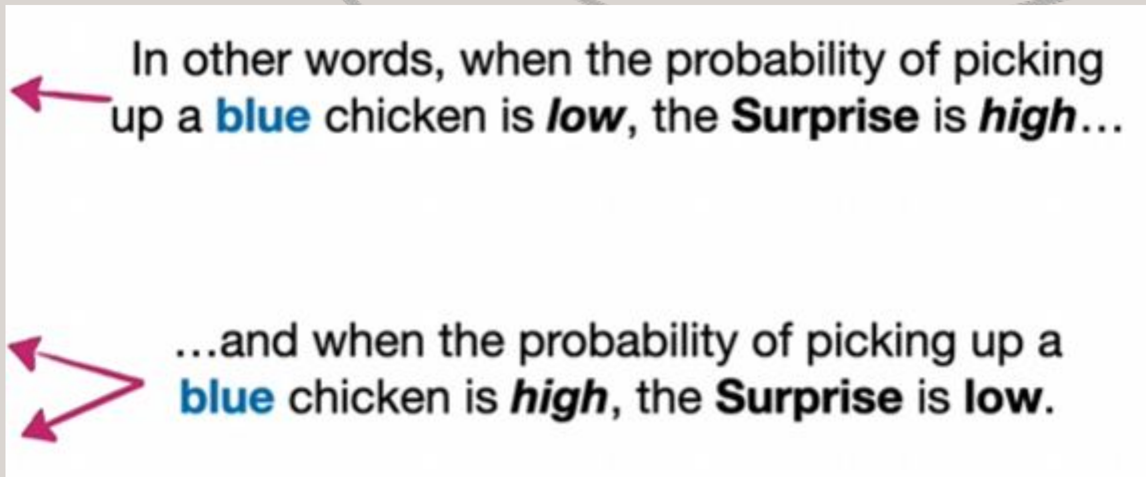


What is surprise?

Surprise seems to be closely tied to probabilities...

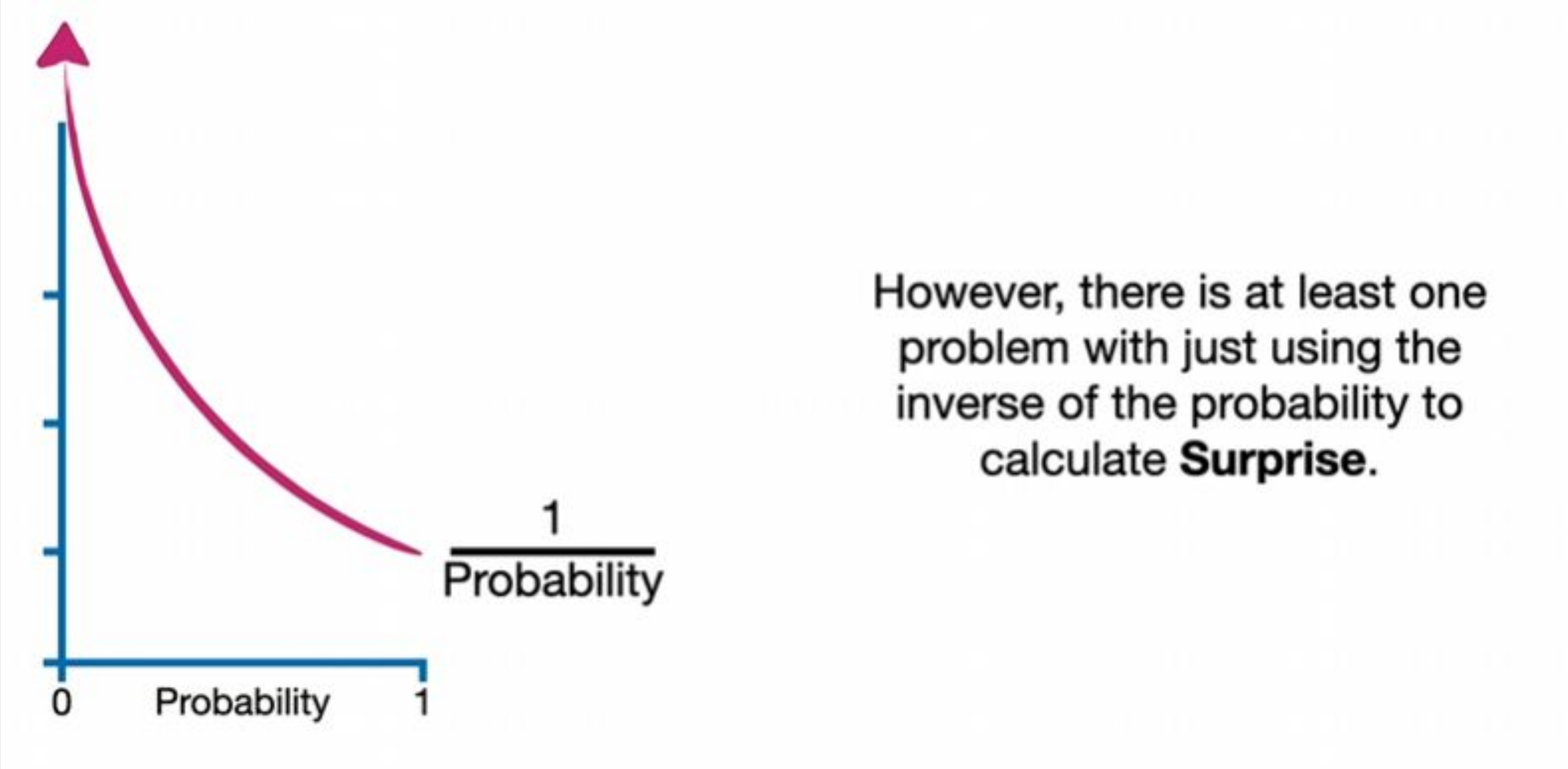






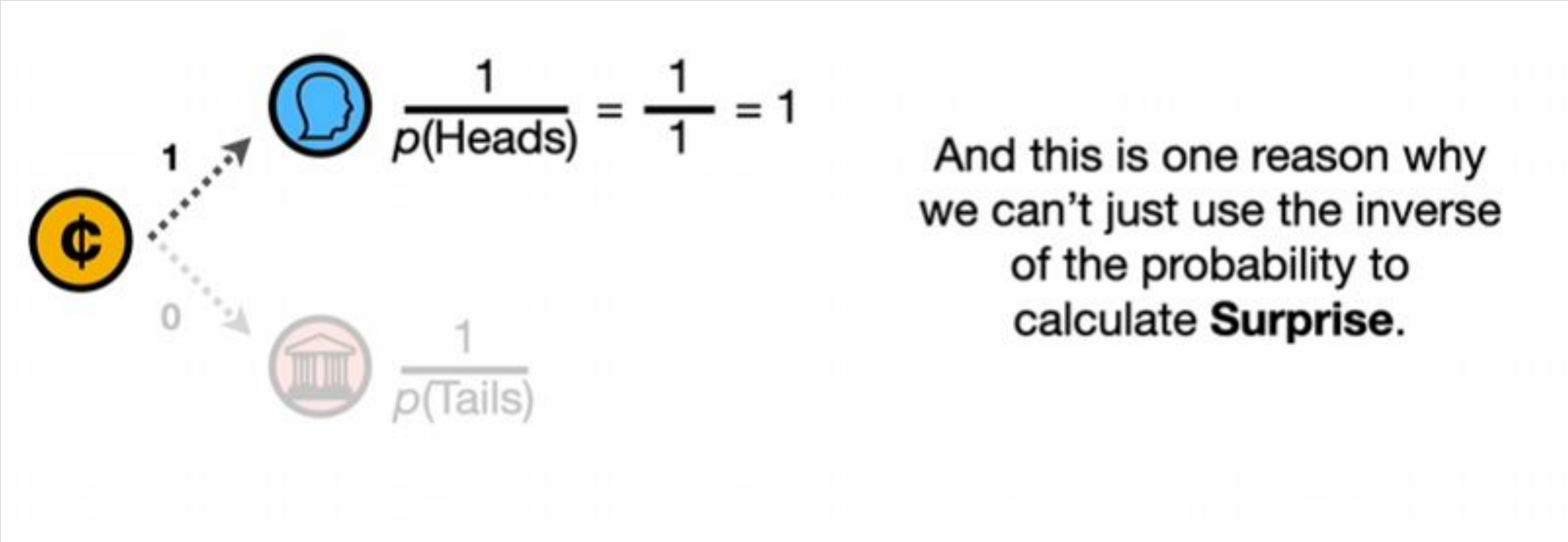
Again, but math

The naive approach seems to make sense when we graph it, but there are some issues.



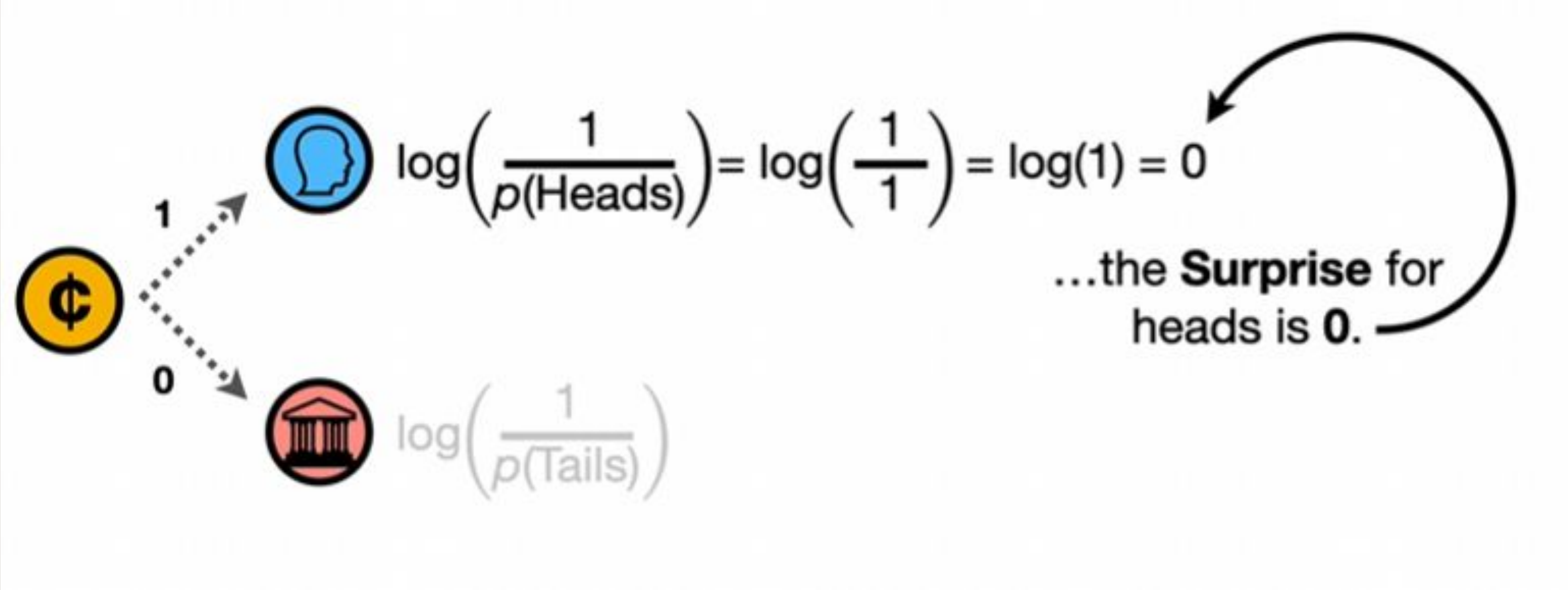
Again, but math

Let's imagine we had a coin that only ever lands on heads → P(heads) = 1



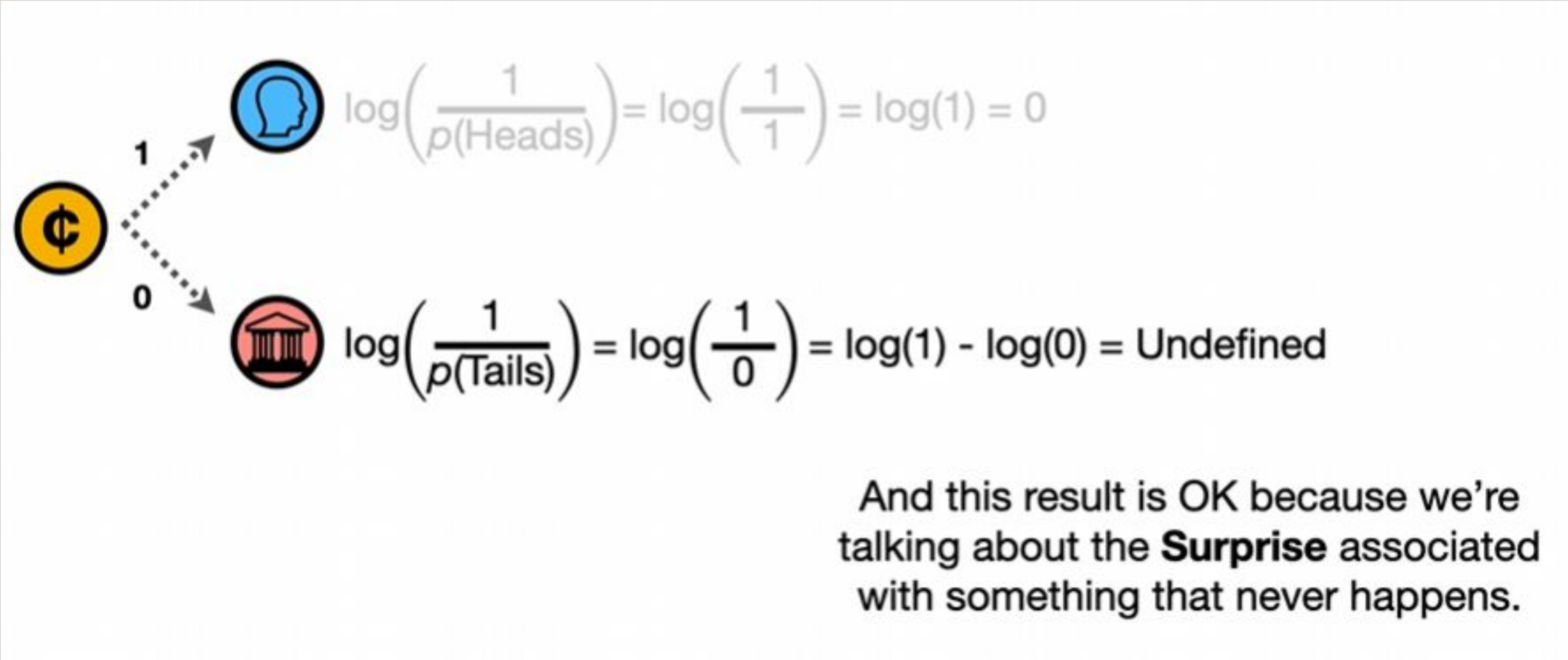
Again, but math

Instead of thinking purely in terms of probabilities, we can introduce logarithms. Now, surprise is as we expect it.



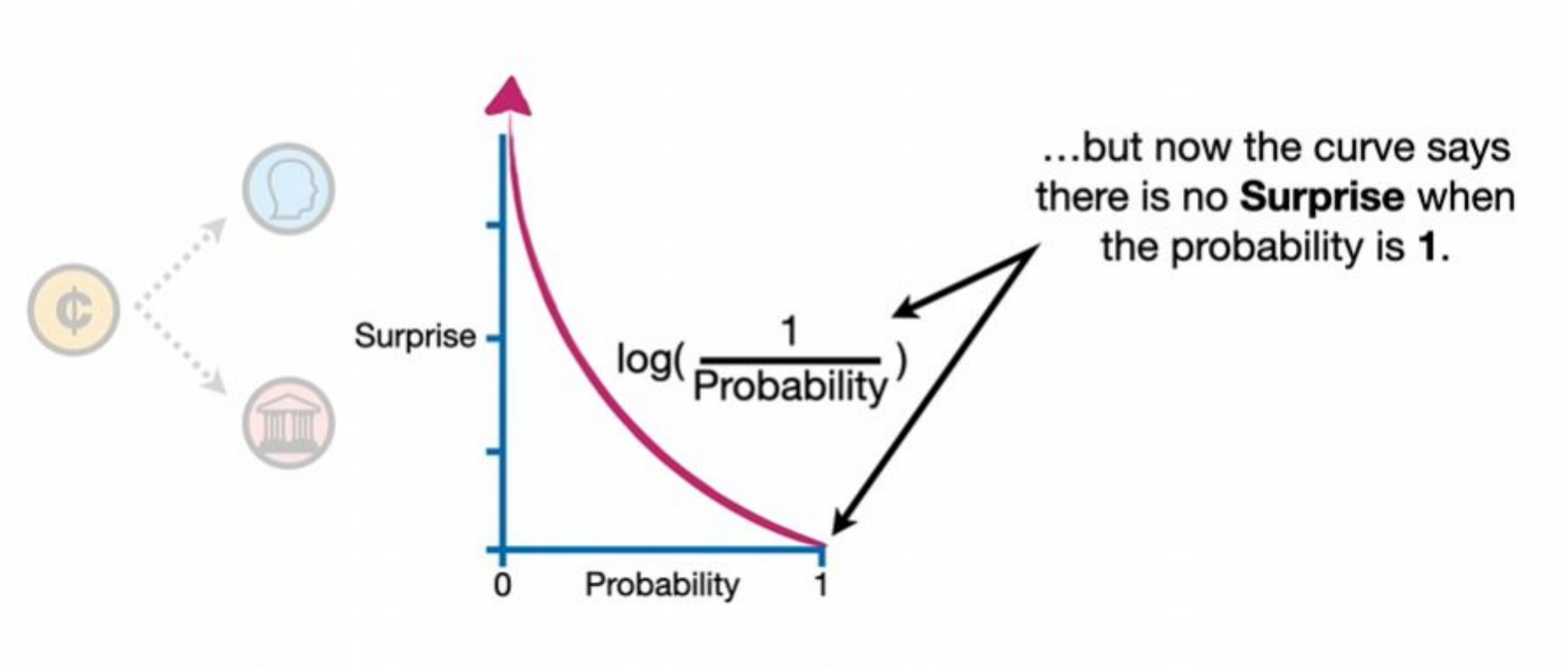
Again, but math

The definition still makes sense for tails, too!



Again, but math

Now, as we would expect, there is no surprise for events of probability 1



Properties

This almost exactly how Shannon defined probability in his 1948 paper.

Shannon information content

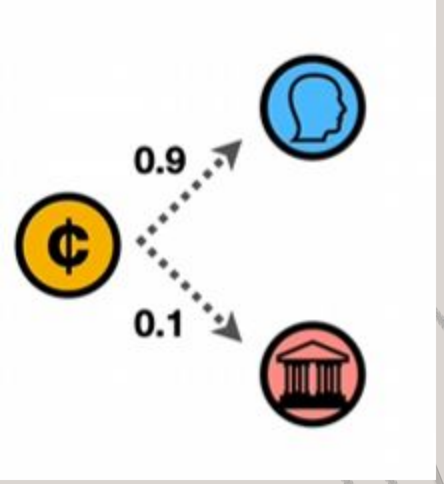
$$h(X{=}x) = \log_2 \frac{1}{P(X{=}x)} = -\log_2 P(X{=}x)$$

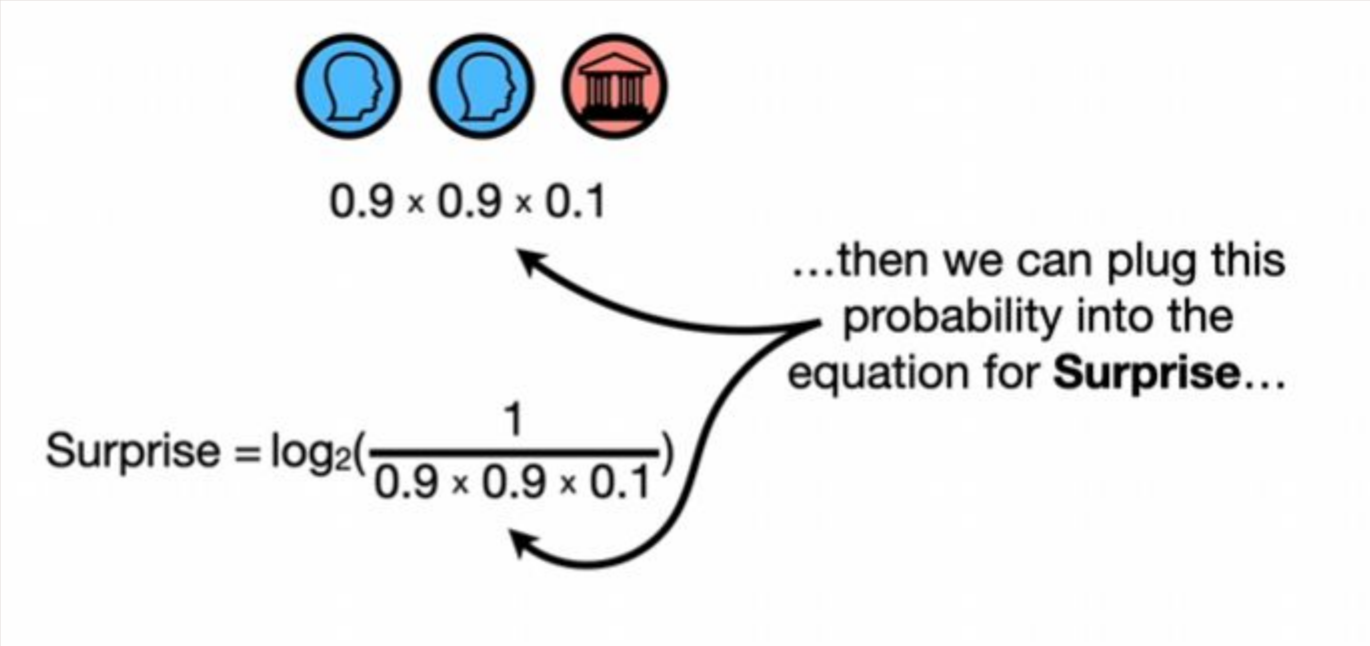
Desiderata in measuring information:

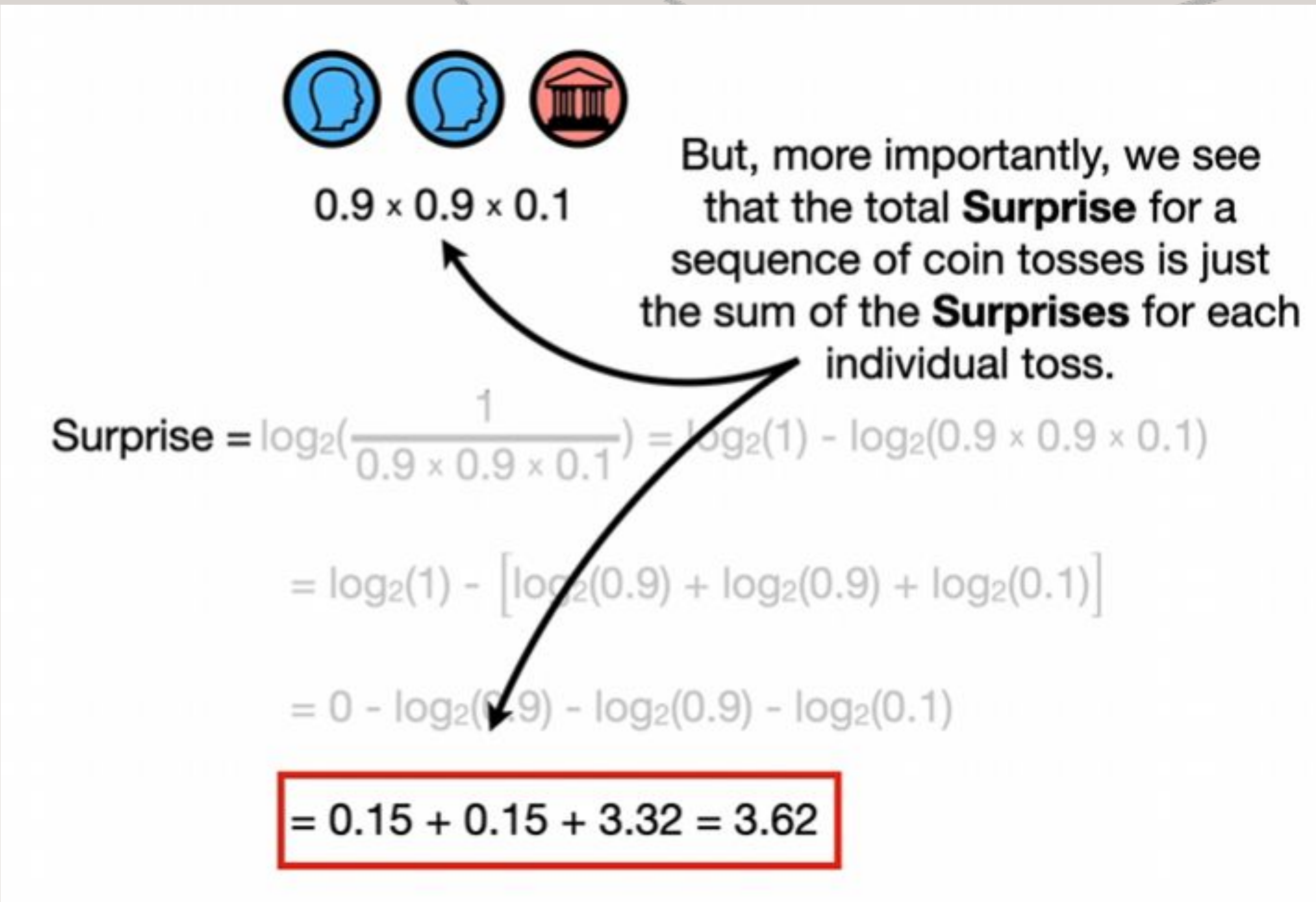
1. Deterministic outcomes contain no information.
2. Information content increases with decreasing probability.
3. Information content is additive for independent R.V.s.

Properties

Now, let's change things up a bit. Our coin is still not fair, but can land on tails with a 10% probability.







Properties

What is the average, or expected surprise per coin toss? We can figure it out by constructing a table.

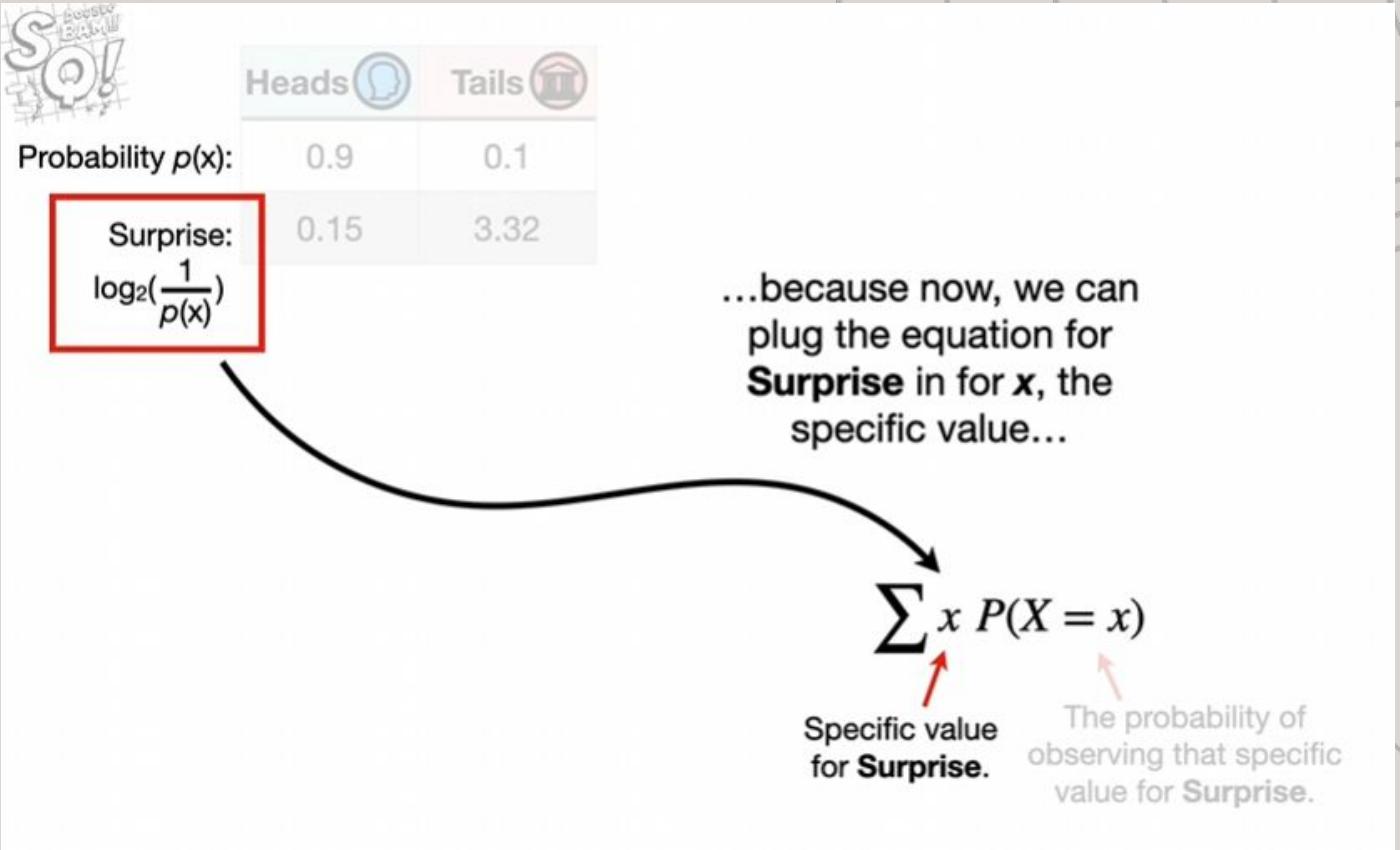
	Heads	Tails
Probability $p(x)$:	0.9	0.1
Surprise: $\log_2(\frac{1}{p(x)})$	0.15	3.32

…then we get the average amount of **Surprise *per*** coin toss, **0.47**.

$$\frac{(0.9 \times 100) \times 0.15 + (0.1 \times 100) \times 3.32}{100} = \frac{46.7}{100} \boxed{= 0.47}$$

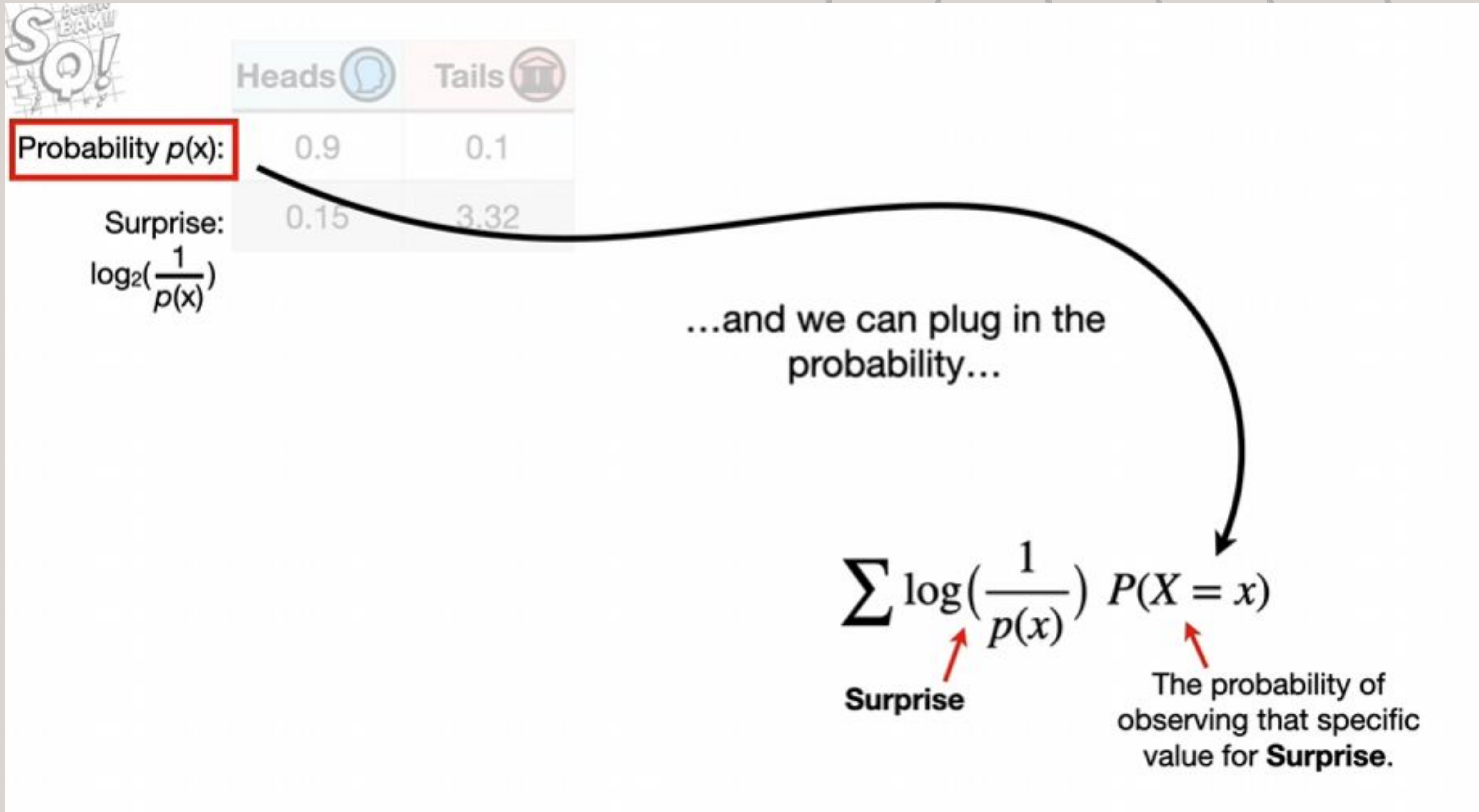
Properties

We can simplify the expected surprise with sigma notation.



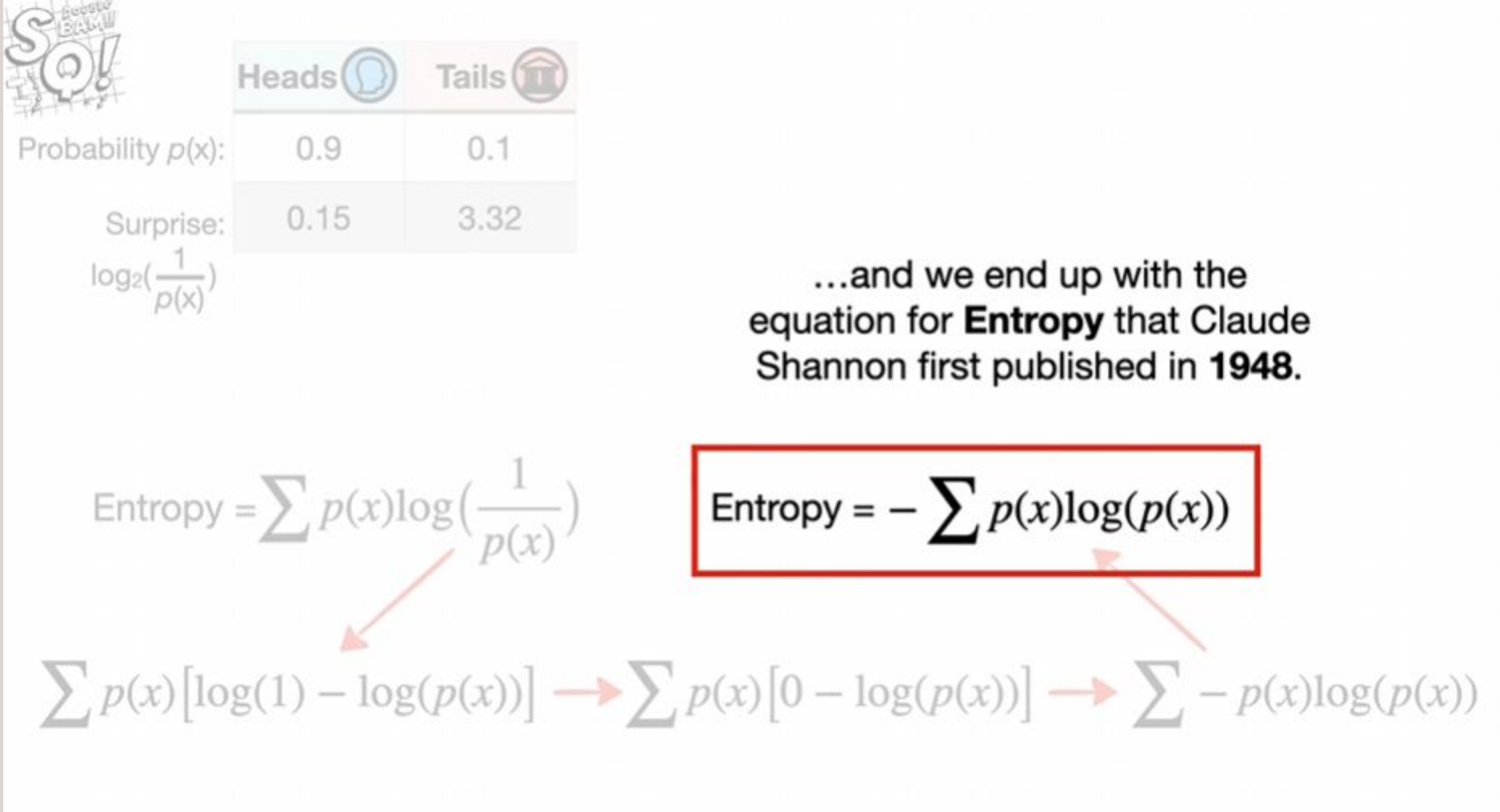
Properties

We can simplify the expected surprise with sigma notation.



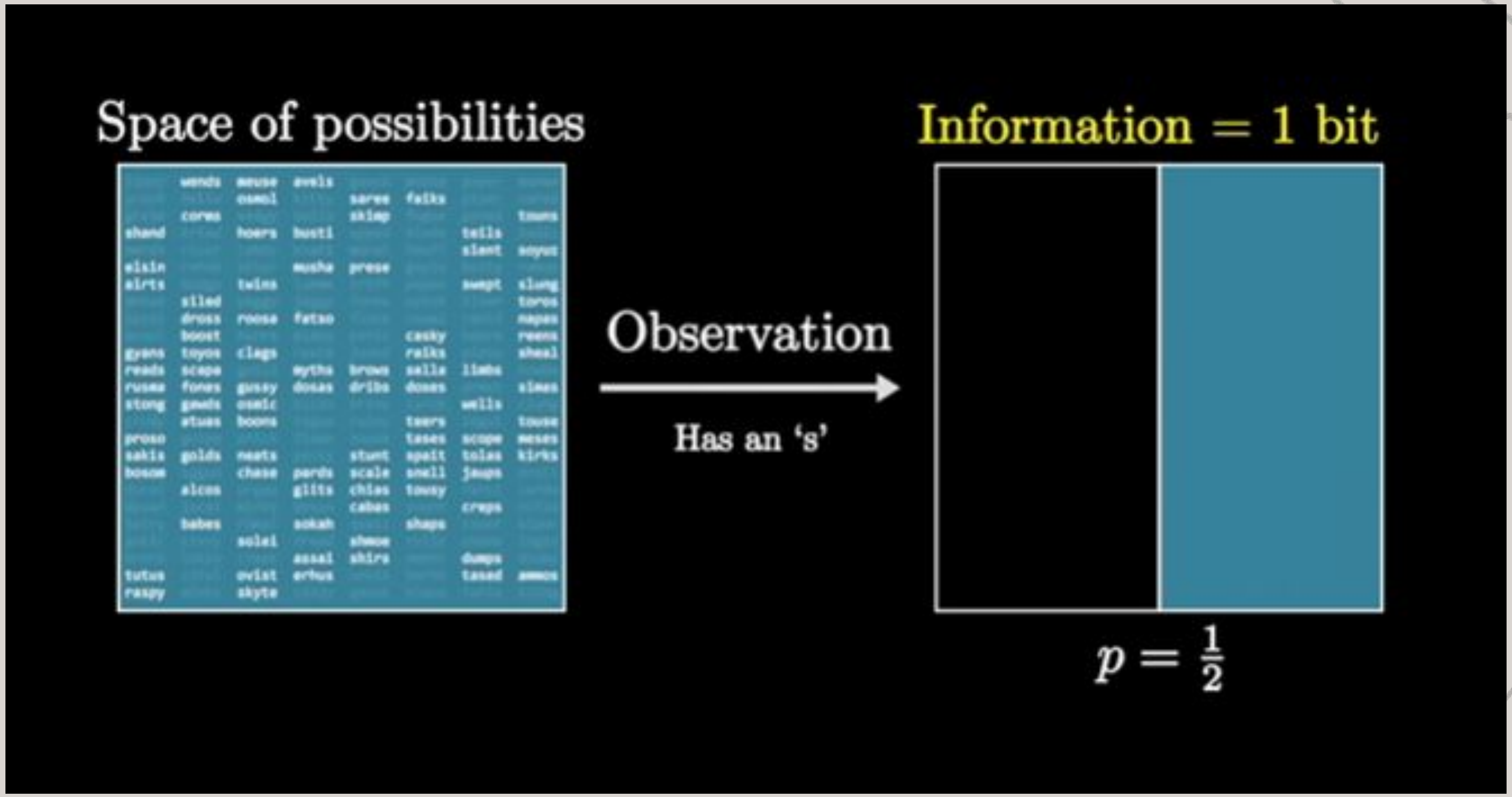
Properties

By applying log rules, we can get to the original definition by Shannon.



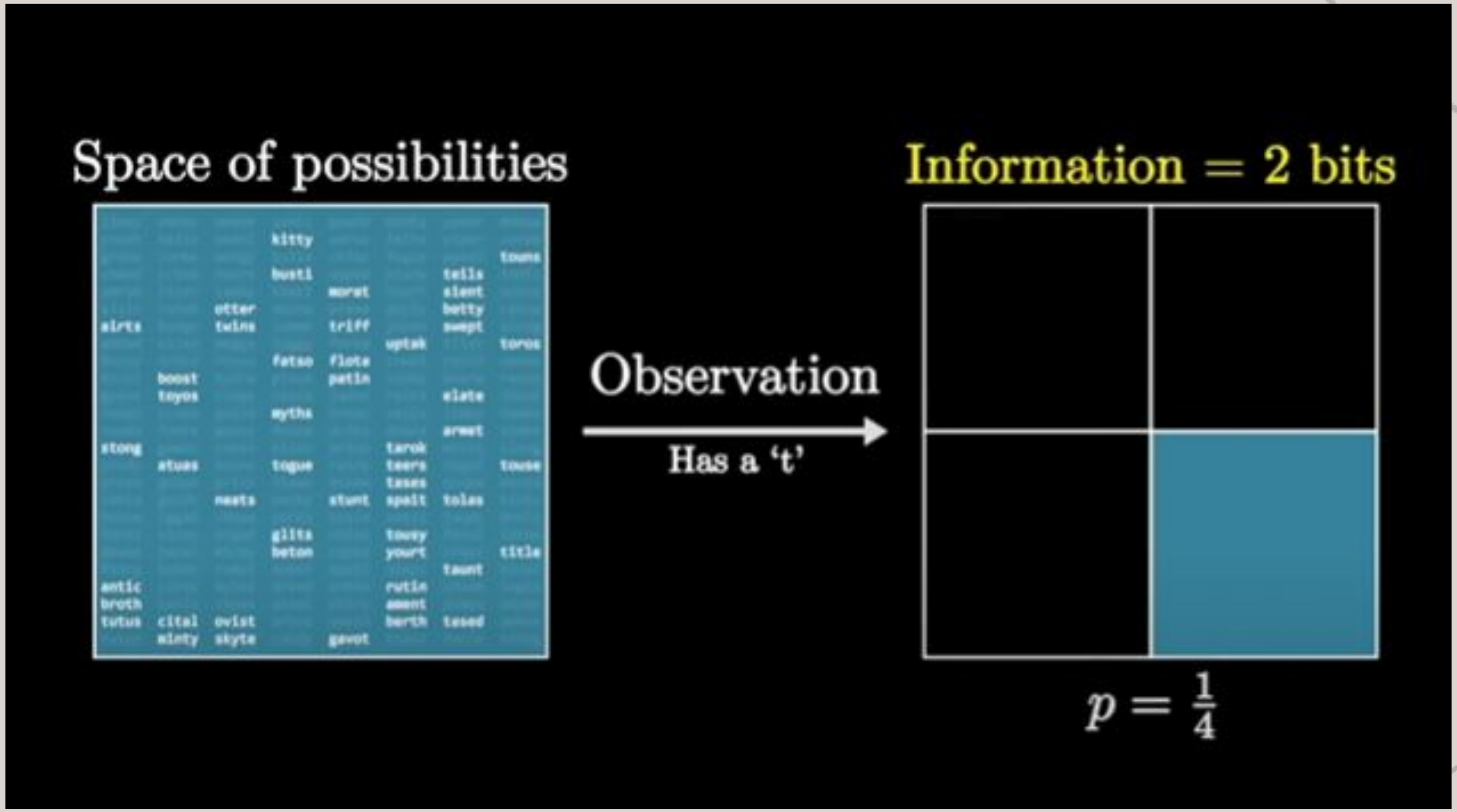
What is a bit?

The fundamental unit in information theory is the bit. It cuts the search space in half.



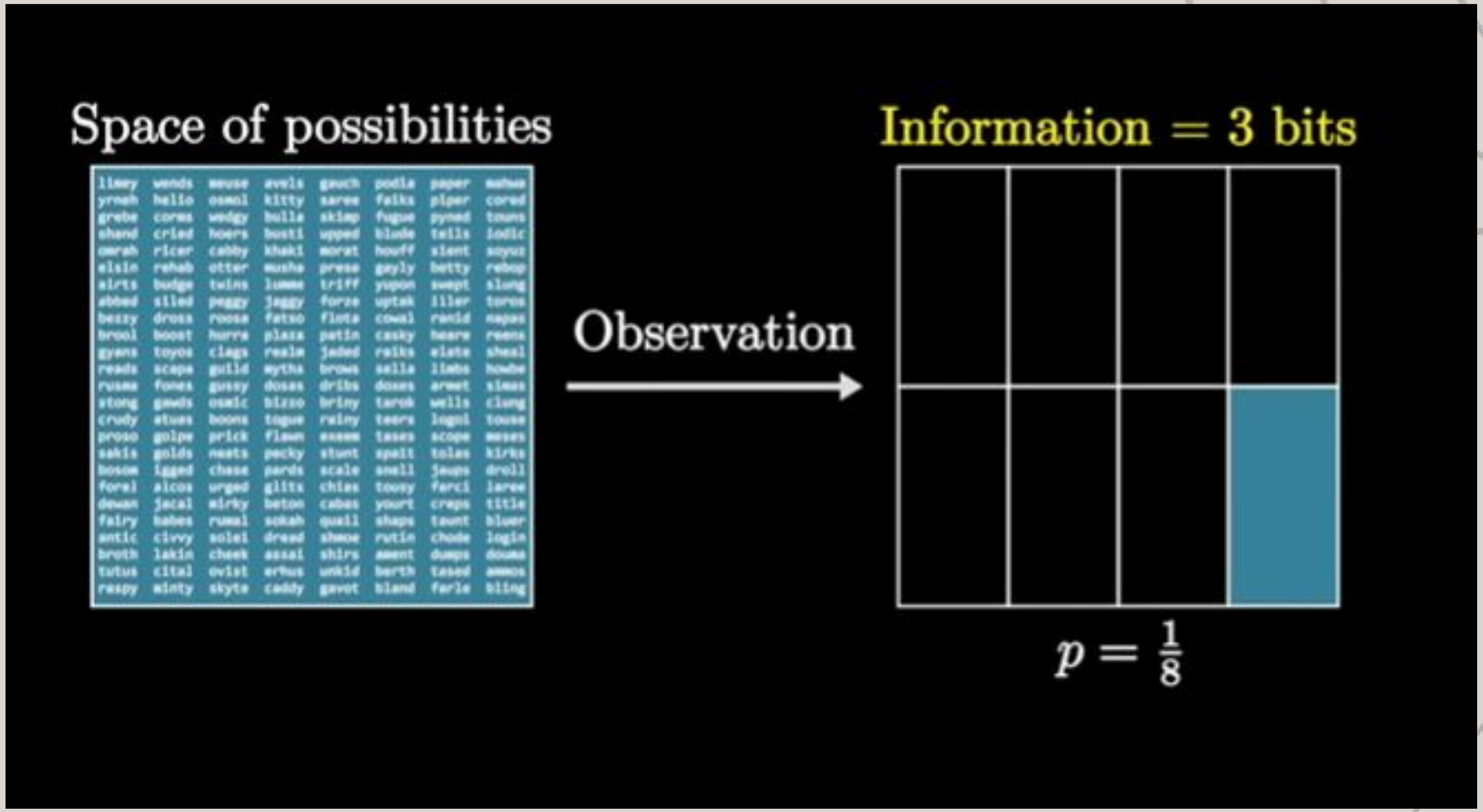
What is a bit?

If we have yet another observation that reduces the search space to ¼ of its original size, this gives us 2 bits of information.



What is a bit?

And so on...



What is a bit?

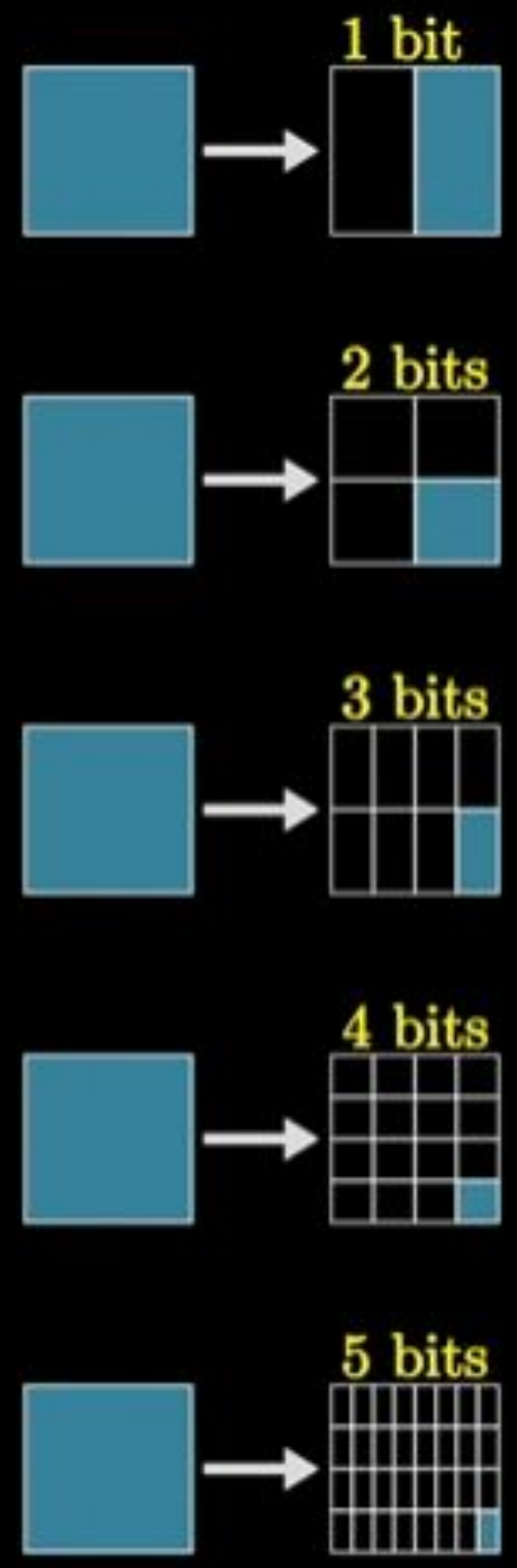
$$\left(\frac{1}{2}\right)^{I} = p$$

$$2^{I} = \frac{1}{p}$$

With this information, we can derive the formula for the information an observation gives us, based on its probability.

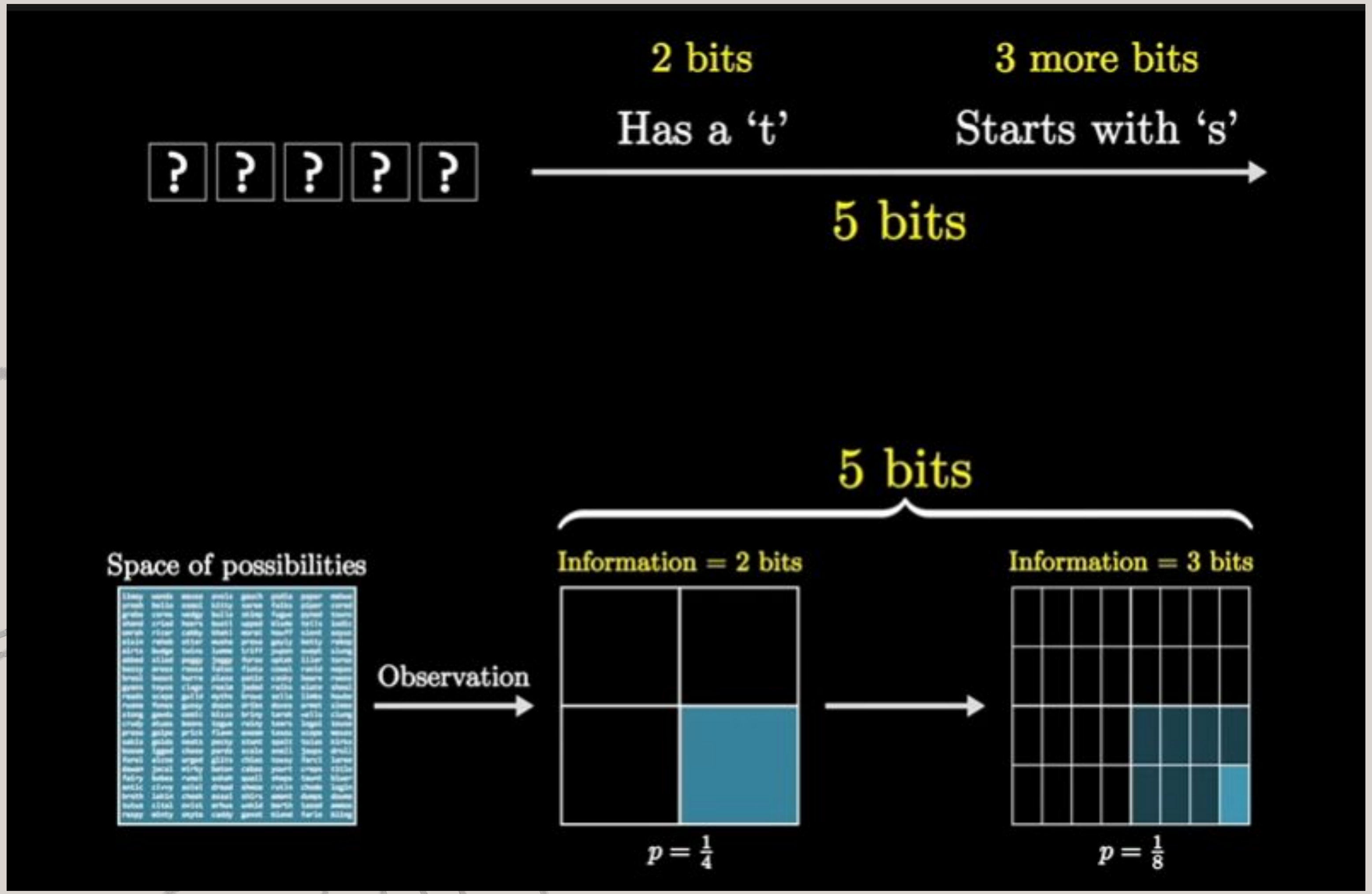
$$I = \log_2\left(\frac{1}{p}\right)$$

$$\boxed{I = -\log_2(p)}$$



What is a bit?

The entropy formula is just the expected information defined in terms of event probabilities (multiplied by surprise values, or entropy)

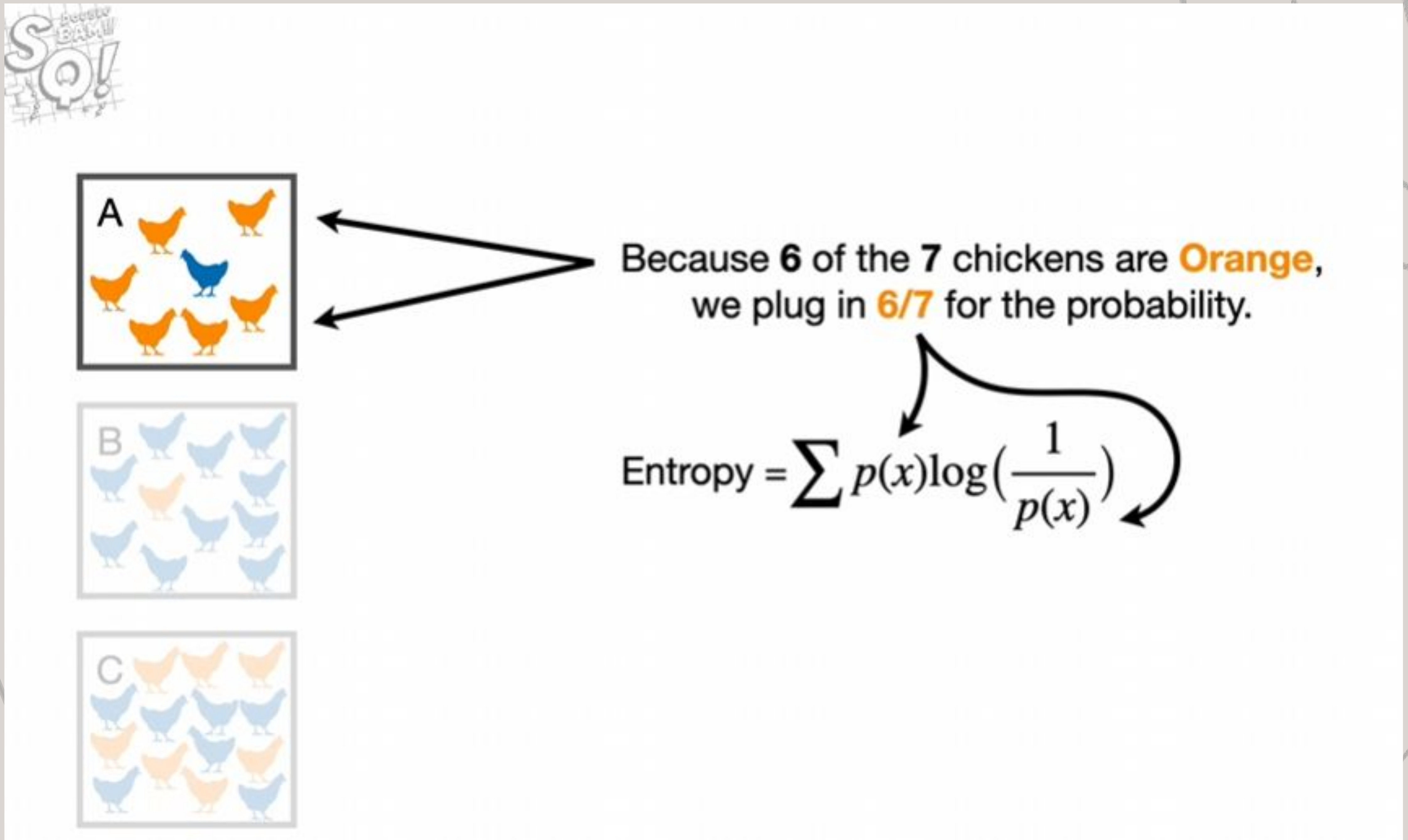


What we want:

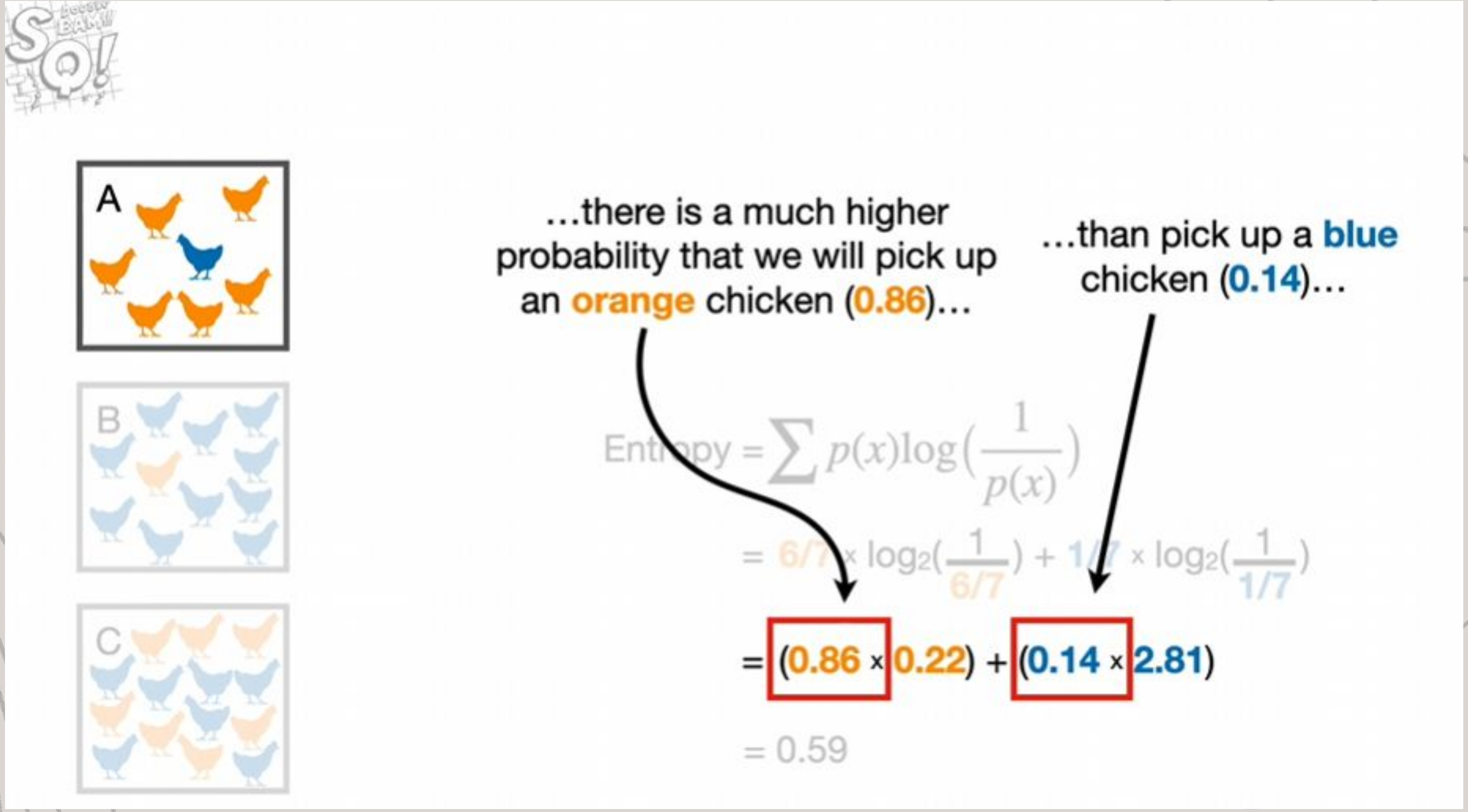
$$E[\text{Information}] = \sum_x p(x) \cdot (\text{Something})$$

$$E[I] = \sum_x p(x) \cdot \log_2(1/p(x))$$

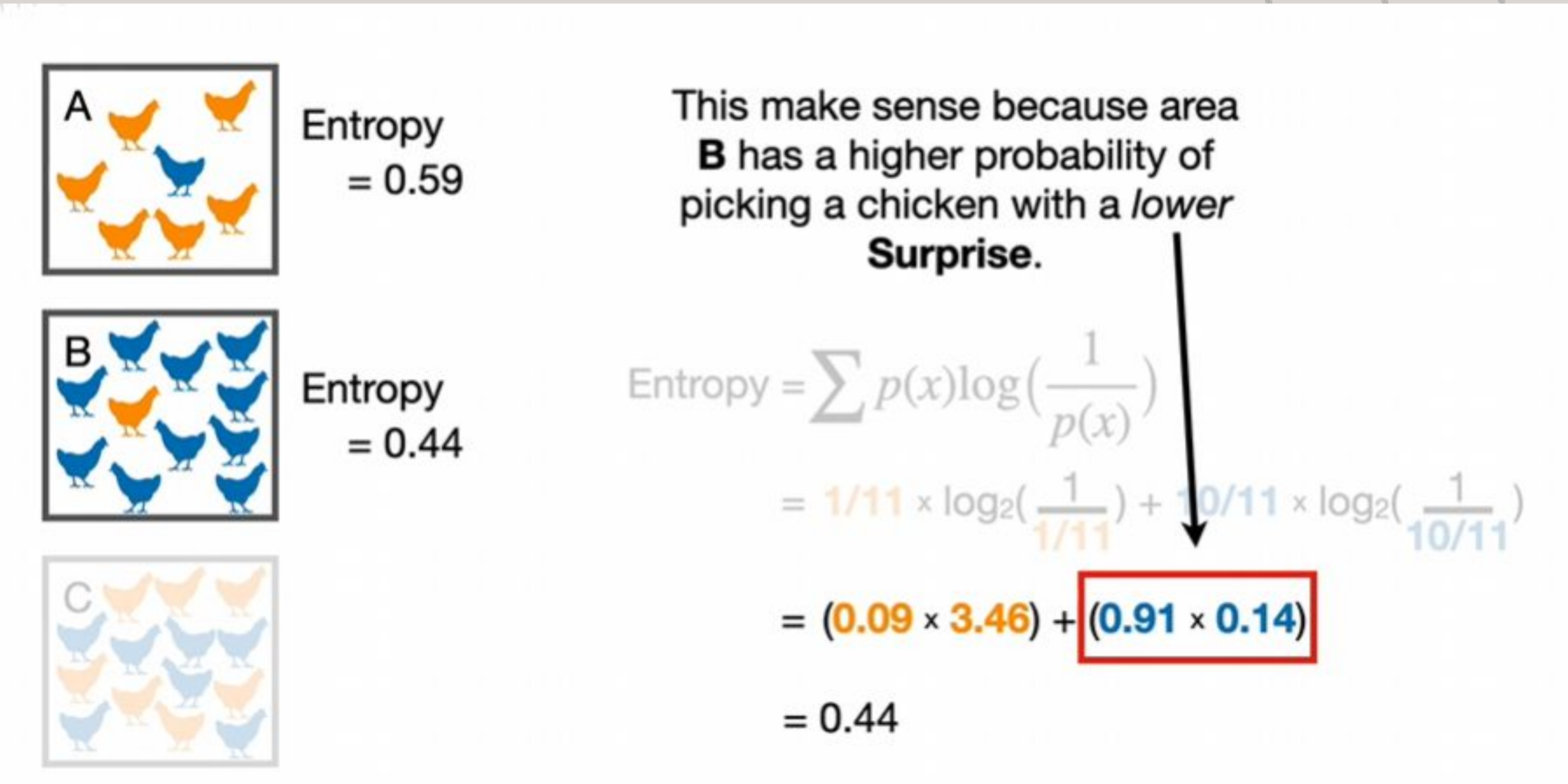
Chickens, again



Chickens, again

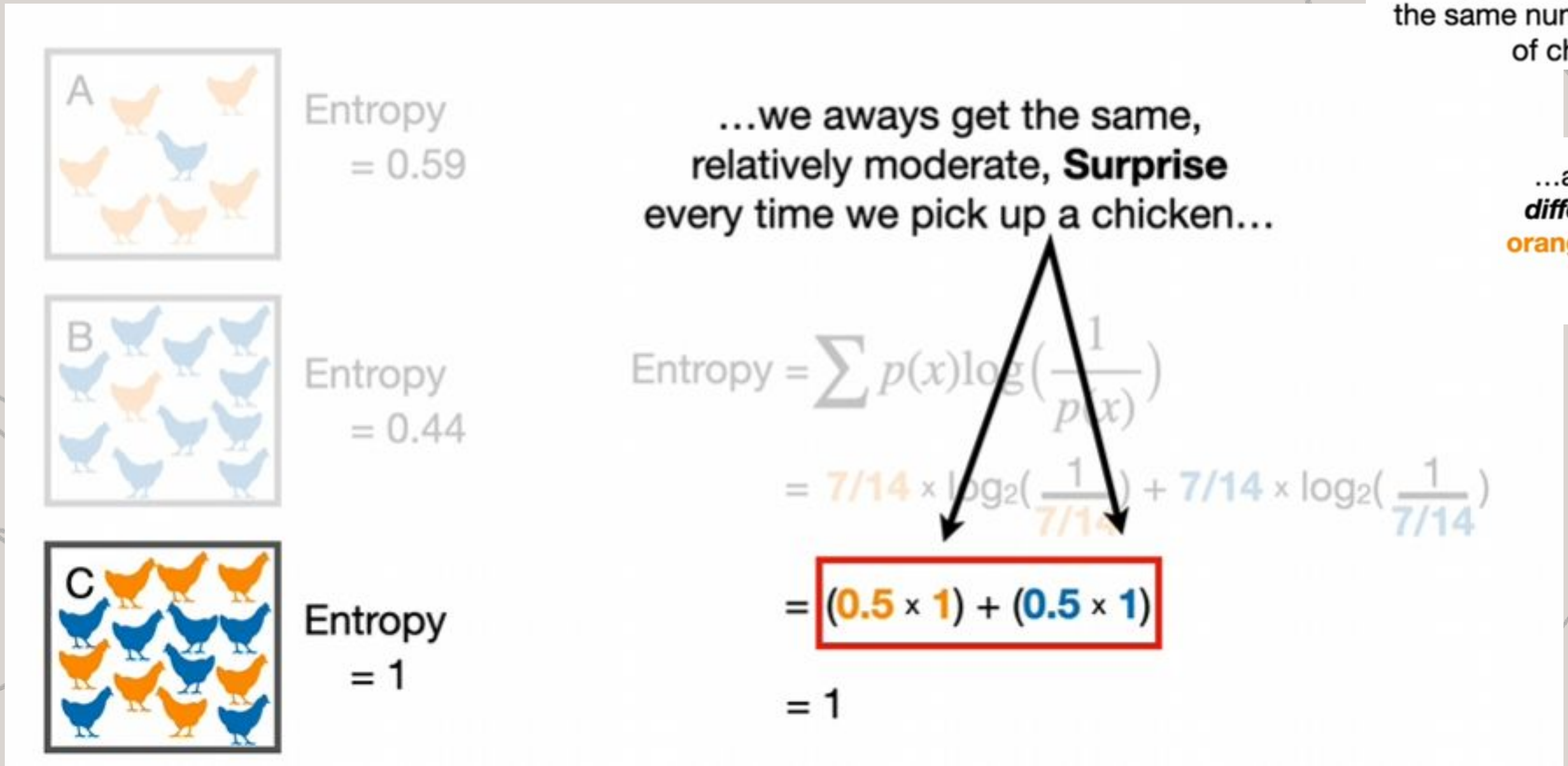


Chickens, again



Chickens, again

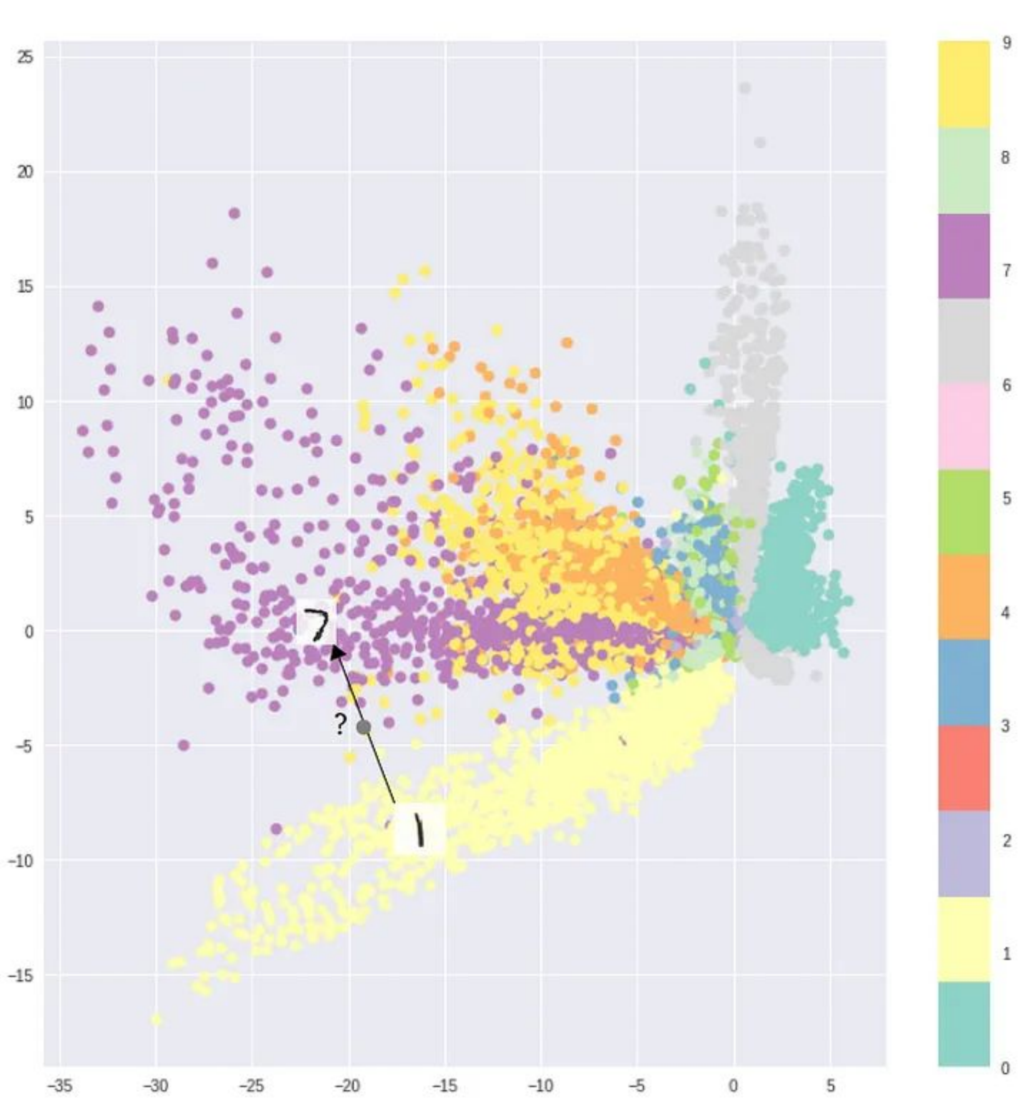
Entropy is highest when we have the same number of both types of chickens…

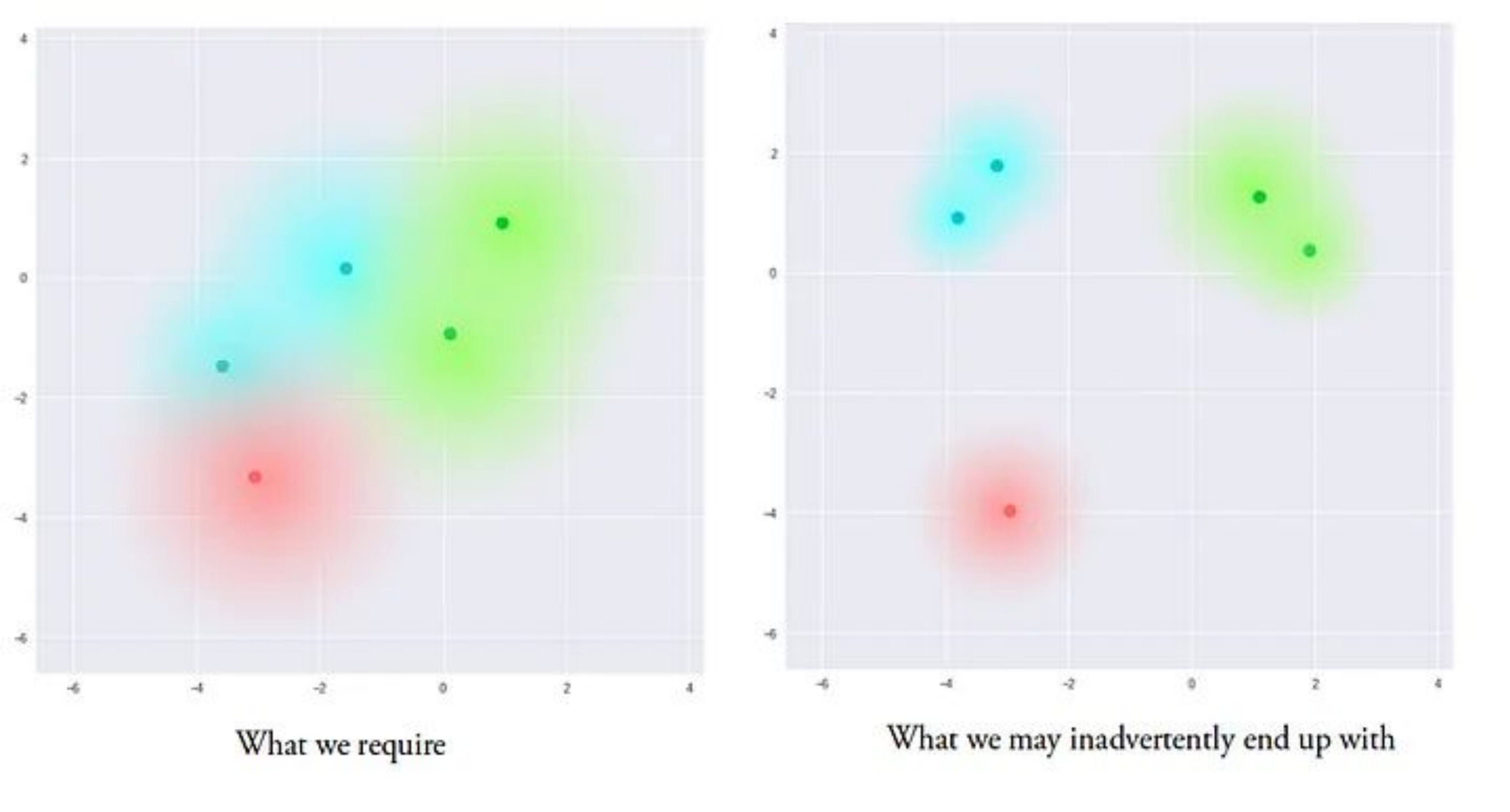


…and as we ***increase the difference*** in the number of orange and **blue** chickens, we lower the **Entropy**.

KL divergence

KL divergence plays a huge role in ML algorithms, including generative and reconstructive models such as variational autoencoders!





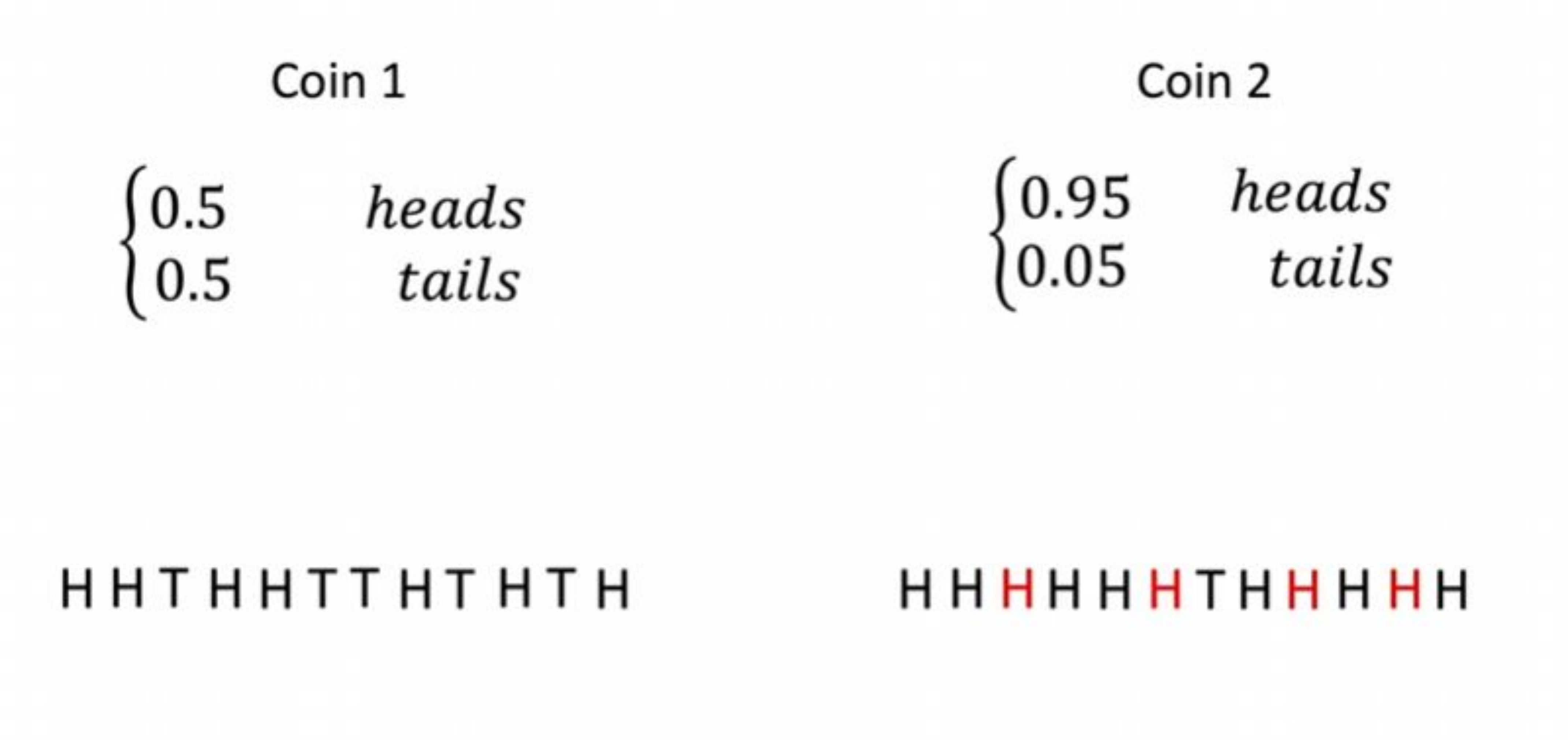
KL divergence

KL divergence helps us compare different data distributions. Similar distributions (with similar probabilities) will appear similar too!

Coin 1	Coin 2
$\begin{cases} 0.5 & \textit{heads} \\ 0.5 & \textit{tails} \end{cases}$	$\begin{cases} 0.55 & \textit{heads} \\ 0.45 & \textit{tails} \end{cases}$
H H T H H T T H T H T H	H H T H H T T H H H T H

KL divergence

Likewise, distributions with different probabilities will have more differences. How can we quantify this?



KL divergence

We can begin our comparison by multiplying out the probabilities of the events occurring under each distribution.

True Coin

$$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$$

Coin 2

$$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$$

H H T H H T H H H T H T

$$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$$

$$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$$

$$\frac{P\ \text{(Observations | real coin)}}{P\ \text{(Observations | coin 2)}} = \frac{p_1^{N_H}\, p_2^{N_T}}{q_1^{N_H}\, q_2^{N_T}}$$

KL divergence

We can then apply log normalization to these probabilities.

True Coin

$$\begin{cases} p_1 & heads \\ p_2 & tails \end{cases}$$

Coin 2

$$\begin{cases} q_1 & heads \\ q_2 & tails \end{cases}$$

H H T H H T H H H T H T

$$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$$

$$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$$

$$\frac{P \text{ (Observations | real coin)}}{P \text{ (Observations | coin 2)}} = \frac{p_1^{N_H} p_2^{N_T}}{q_1^{N_H} q_2^{N_T}}$$

$$\log \left(\frac{p_1^{N_H} p_2^{N_T}}{q_1^{N_H} q_2^{N_T}} \right)^{\frac{1}{N}}$$

KL divergence

And then expand out the terms using log rules.

True Coin

$$\begin{cases} p_1 & heads \\ p_2 & tails \end{cases}$$

Coin 2

$$\begin{cases} q_1 & heads \\ q_2 & tails \end{cases}$$

H H T H H T H H H T H T

$$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$$

$$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$$

$$\log \left(\frac{p_1^{N_H} \, p_2^{N_T}}{q_1^{N_H} \, q_2^{N_T}} \right)^{\frac{1}{N}}$$

$$\frac{1}{N} \log p_1^{N_H} + \frac{1}{N} \log p_2^{N_T} - \frac{1}{N} \log q_1^{N_H} - \frac{1}{N} \log q_2^{N_T}$$

KL divergence

We can then continue simplifying in the limit, as we expect the frequencies to increase.

True Coin

$$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$$

Coin 2

$$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$$

H H T H H T H H H T H T

$$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$$

$$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$$

$$\frac{1}{N} \log p_1^{N_H} + \frac{1}{N} \log p_2^{N_T} - \frac{1}{N} \log q_1^{N_H} - \frac{1}{N} \log q_2^{N_T}$$

$$\frac{N_H}{N} \log p_1 + \frac{N_T}{N} \log p_2 - \frac{N_H}{N} \log q_1 - \frac{N_T}{N} \log q_2$$

$$p_1 \log p_1 + p_2 \log p_2 - p_1 \log q_1 - p_2 \log q_2$$

KL divergence

Taking a step back, what we have derived is a log-normalized measure of the differences between the two distributions.

This is essentially what the KL-divergence formula is telling us!

True Coin

$$\begin{cases} p_1 & heads \\ p_2 & tails \end{cases}$$

Coin 2

$$\begin{cases} q_1 & heads \\ q_2 & tails \end{cases}$$

H H T H H T H H H T H T

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

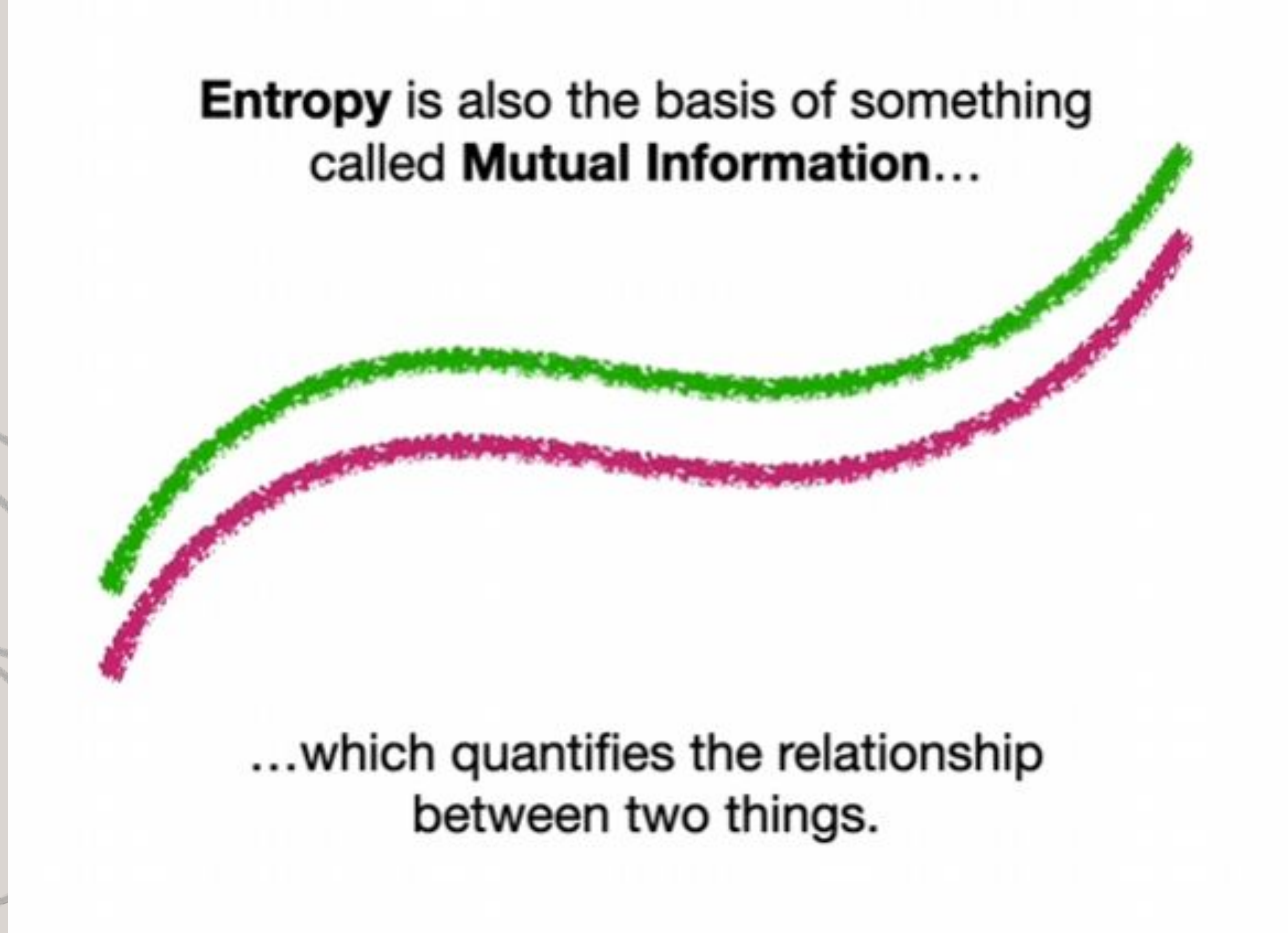
$$p_1 \log \frac{p_1}{q_1} + p_2 \log \frac{p_2}{q_2}$$

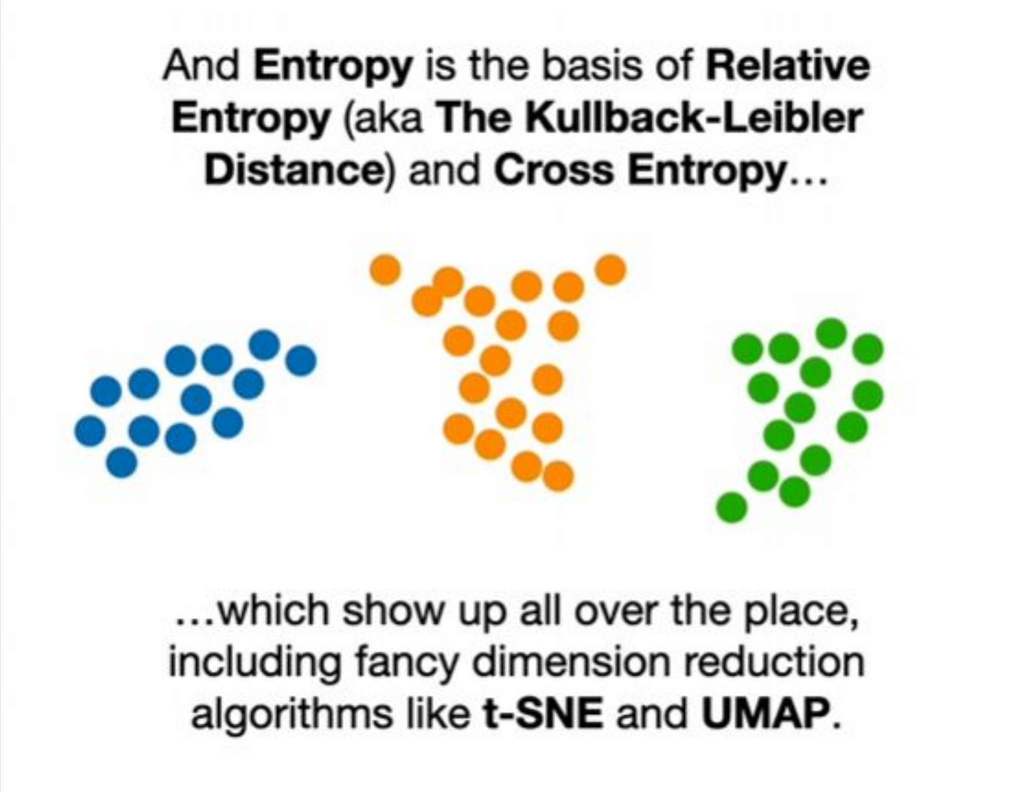
$$\text{Log} \left(\frac{P(\text{Observations of d1} \mid \text{distribution 1})}{P(\text{Observations of d1} \mid \text{distribution 2})} \right)$$

*d1 = distribution 1

KL divergence

Cross-entropy loss is equivalent to KL-loss. By minimizing cross-entropy, we minimize the distance between distributions.





quAIdditch

What is the formula for entropy?

Thank You