



AI ALIGNMENT COHORT

Session 1

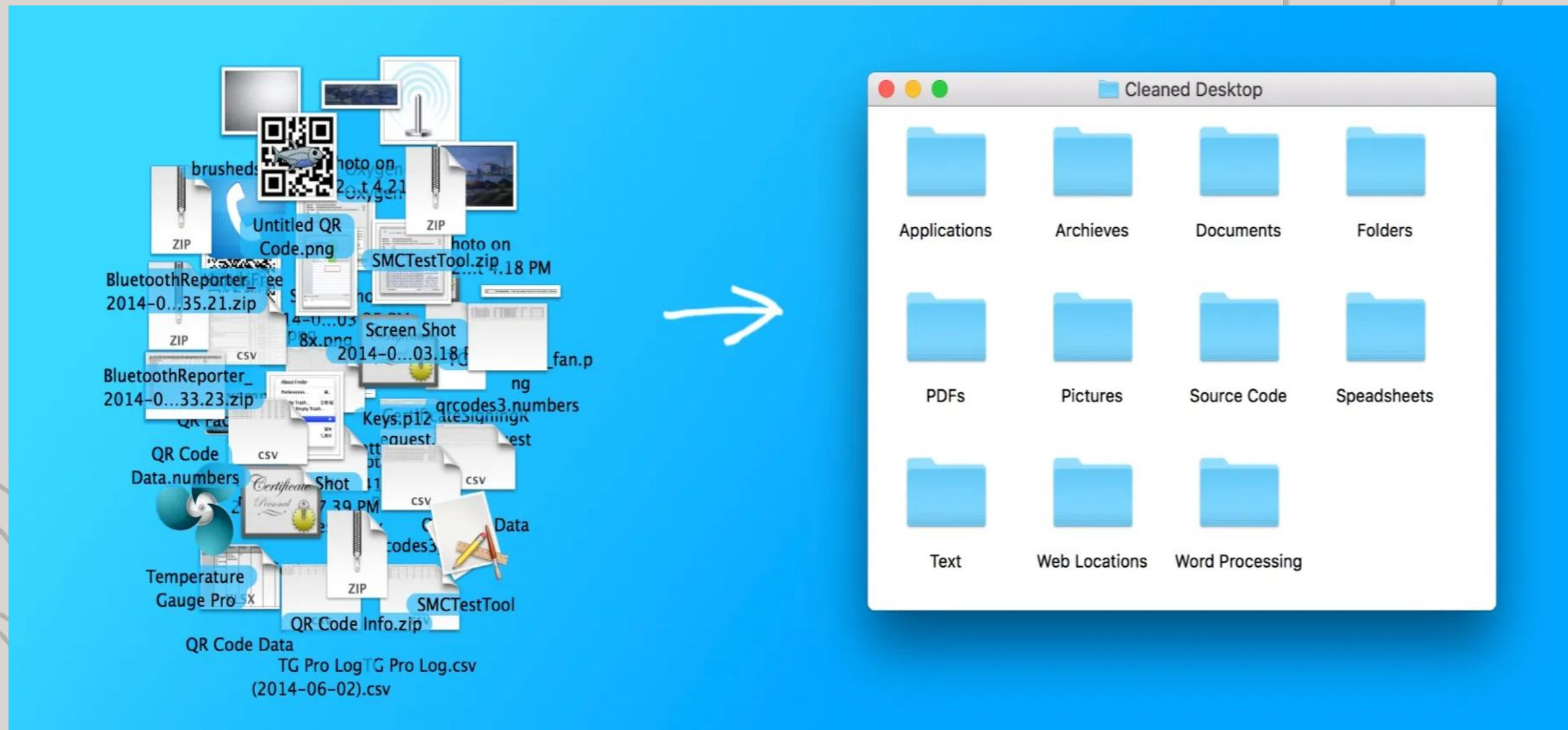
Linear Algebra

&

Information Theory

Cohere for AI
BIRDS x Safety & Alignment Group

Why Bother?



Linear Transformations



Folding

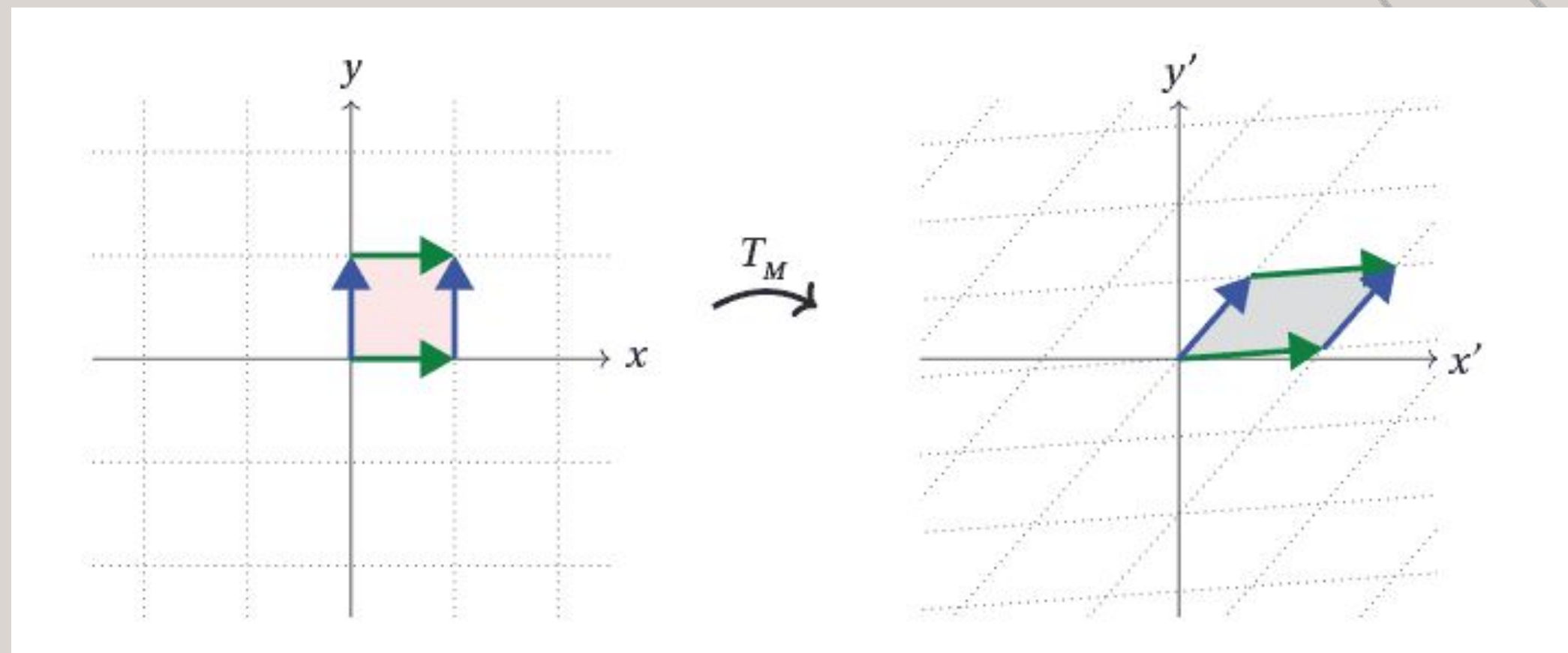
Squishing

Stretching

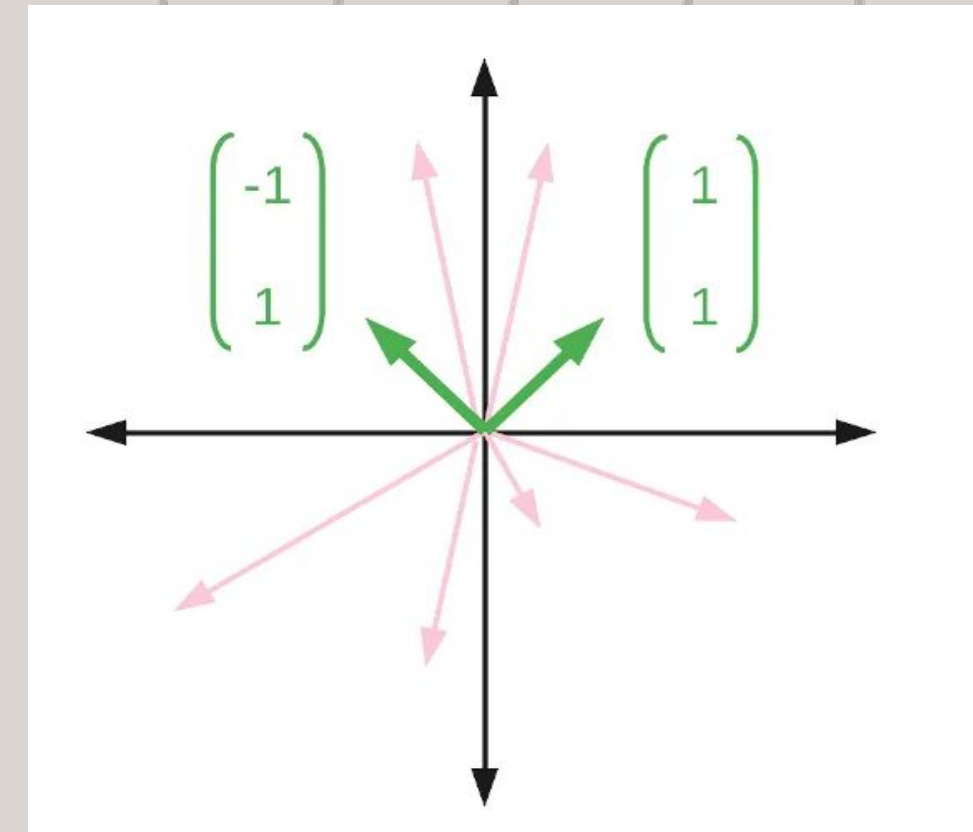
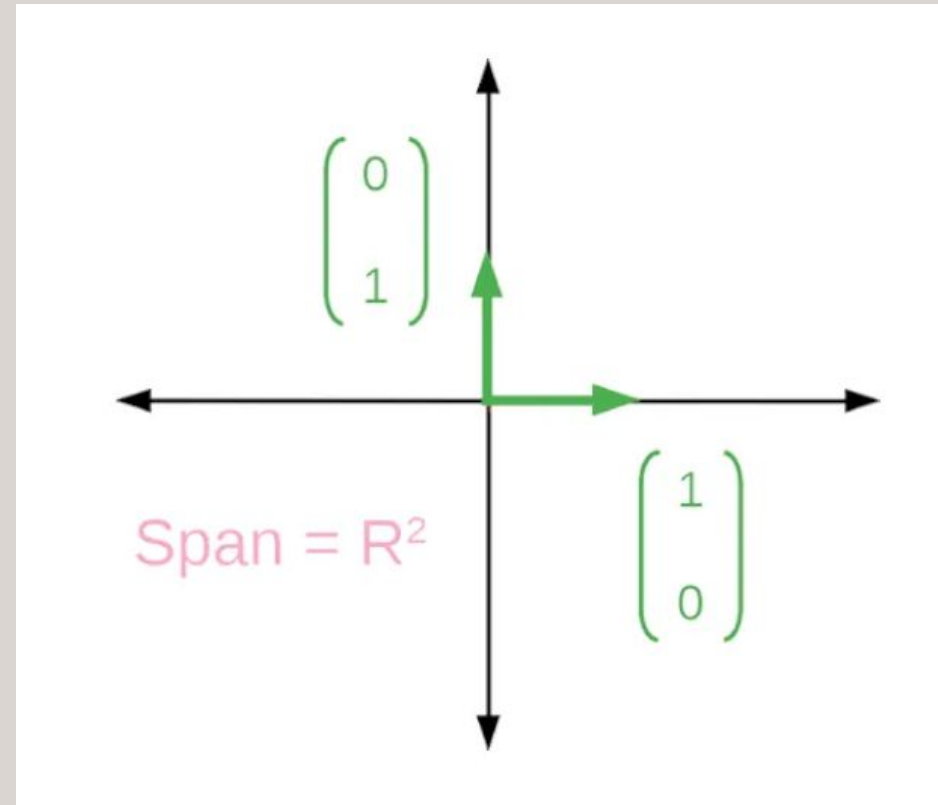
Twisting

Transformations

Linear Transformations



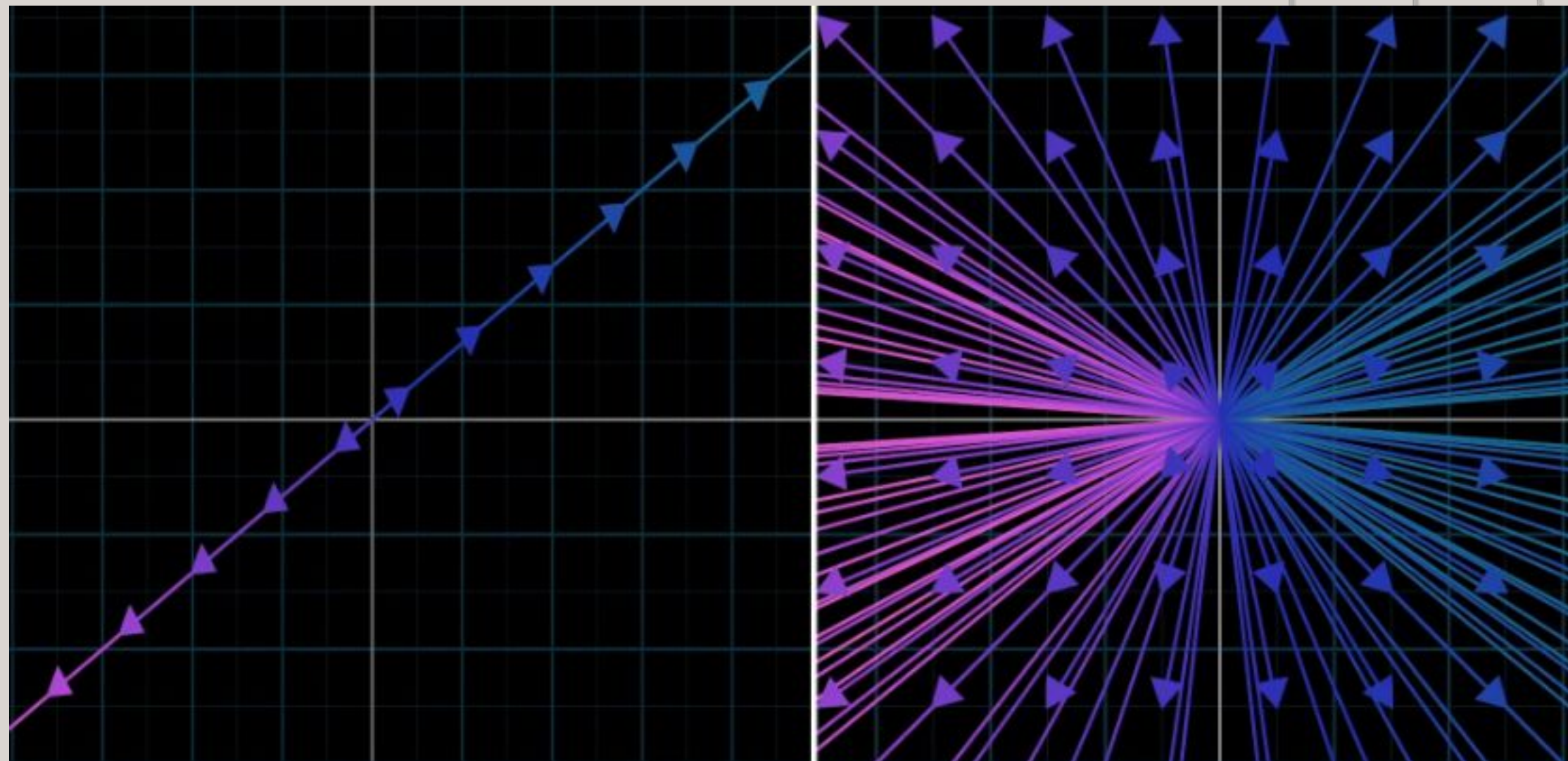
Basis



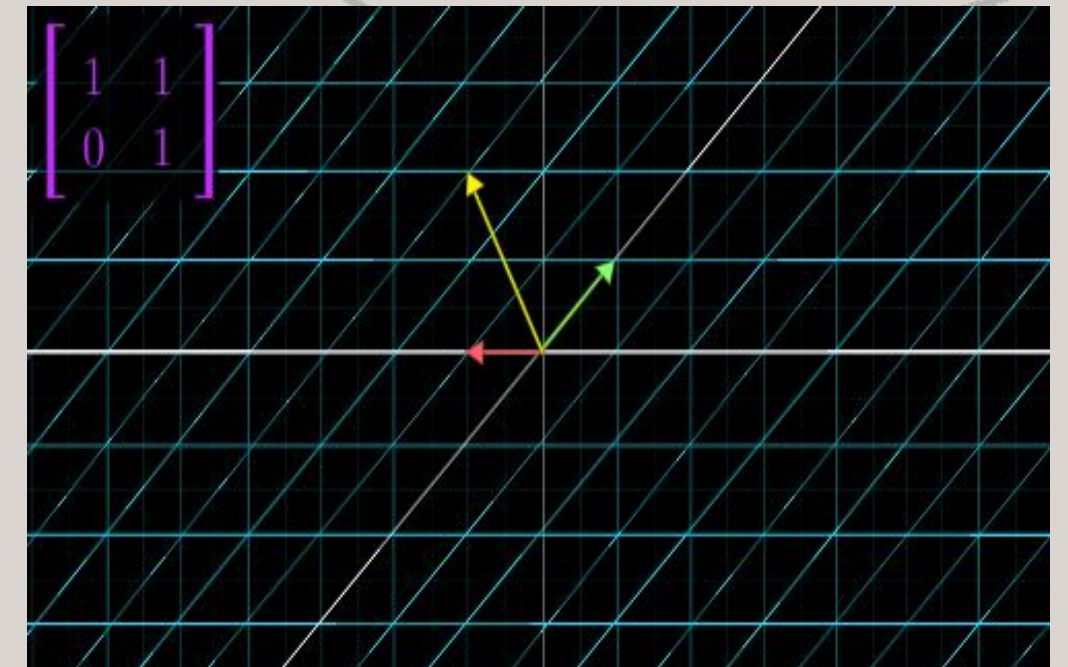
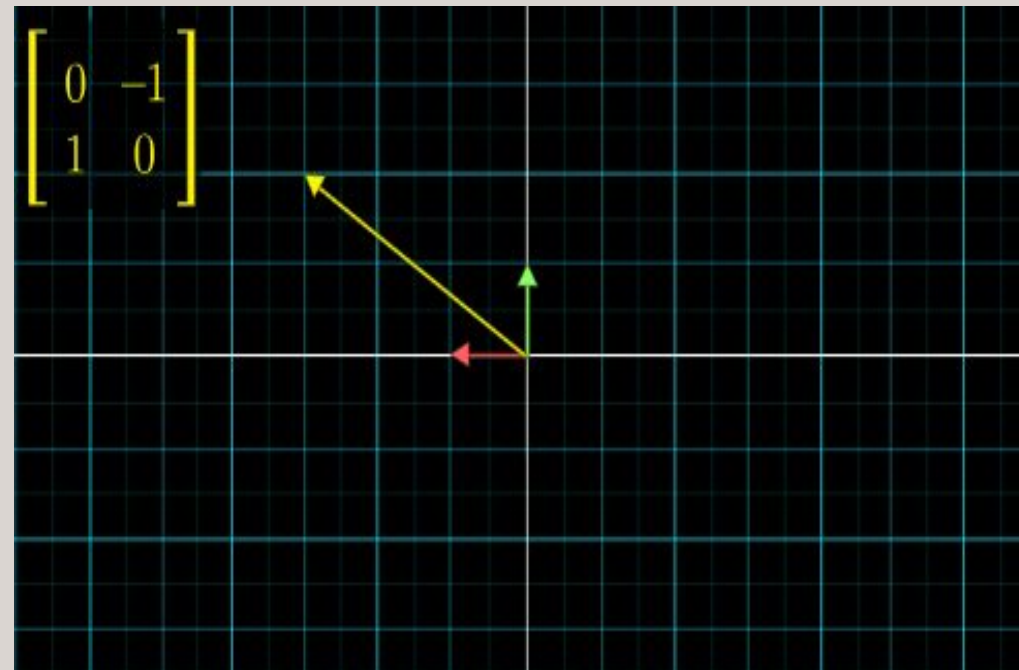
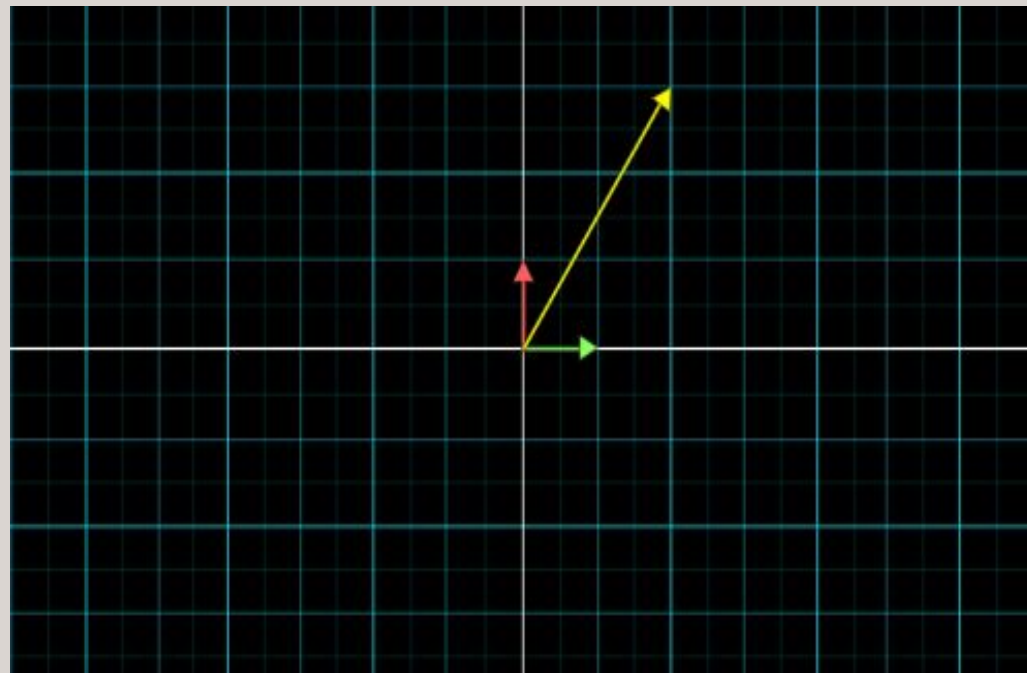
Basis: A set of n vectors, $\{v_1, v_2, \dots, v_n\}$, is a basis of some space S if these two conditions are true:

1. $\{v_1, v_2, \dots, v_n\}$ are linearly independent
2. $\{v_1, v_2, \dots, v_n\}$ span the set S . In other words, $\text{Span}\{v_1, v_2, \dots, v_n\} = S$

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:

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

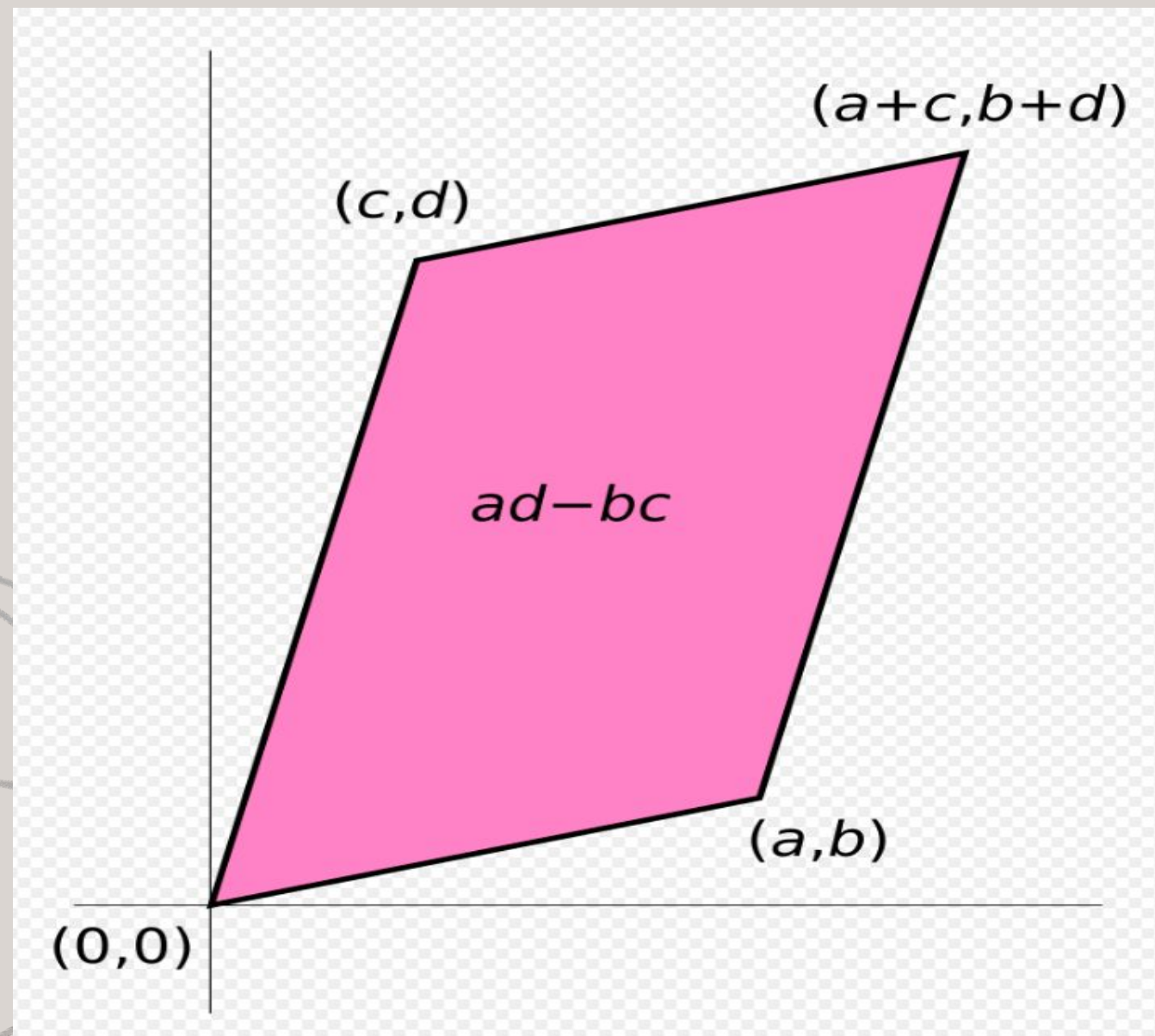
The final matrix (in reduced row echelon form) has two non-zero rows and thus the rank of matrix A is 2.

Trace

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

Gives Important information about:
Area Scaling Factor
Invertibility
Orientation


Determinant



$$\det(A) = ad - bc$$

Transpose

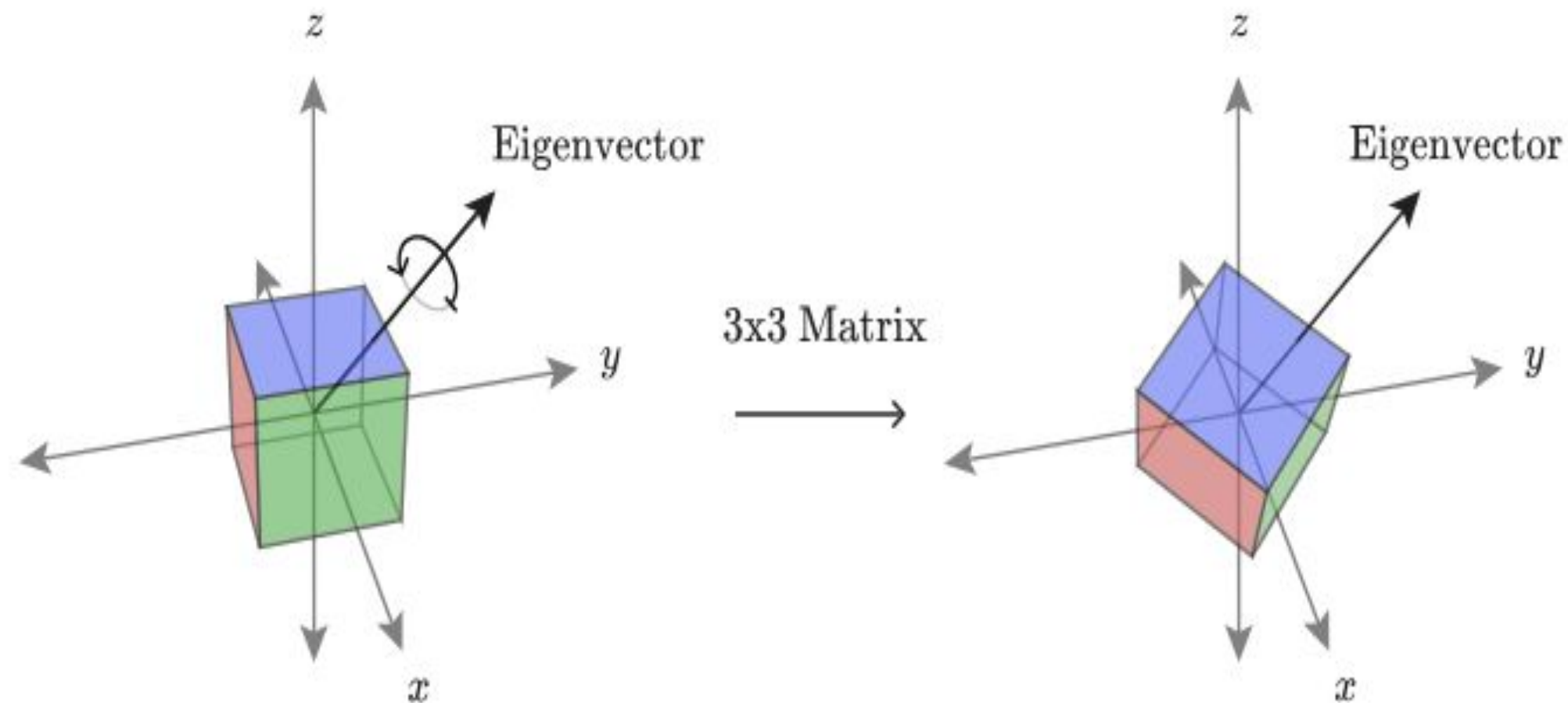
2	4	-1
-10	5	11
18	-7	6



2	-10	18
4	5	-7
-1	11	6

The diagram illustrates the transpose of a 3x3 matrix. The original matrix on the left has rows [2, 4, -1], [-10, 5, 11], and [18, -7, 6]. The transposed matrix on the right has rows [2, -10, 18], [4, 5, -7], and [-1, 11, 6]. The elements are color-coded: yellow for the first column of the original matrix, teal for the second, and red for the third.

Eigenvectors & Eigenvalues



Transformation

matrix Eigenvalue

$$A\vec{v} = \lambda\vec{v}$$

↑ ↑
Eigenvector

quAldditch

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.



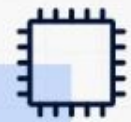
Data Storage

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



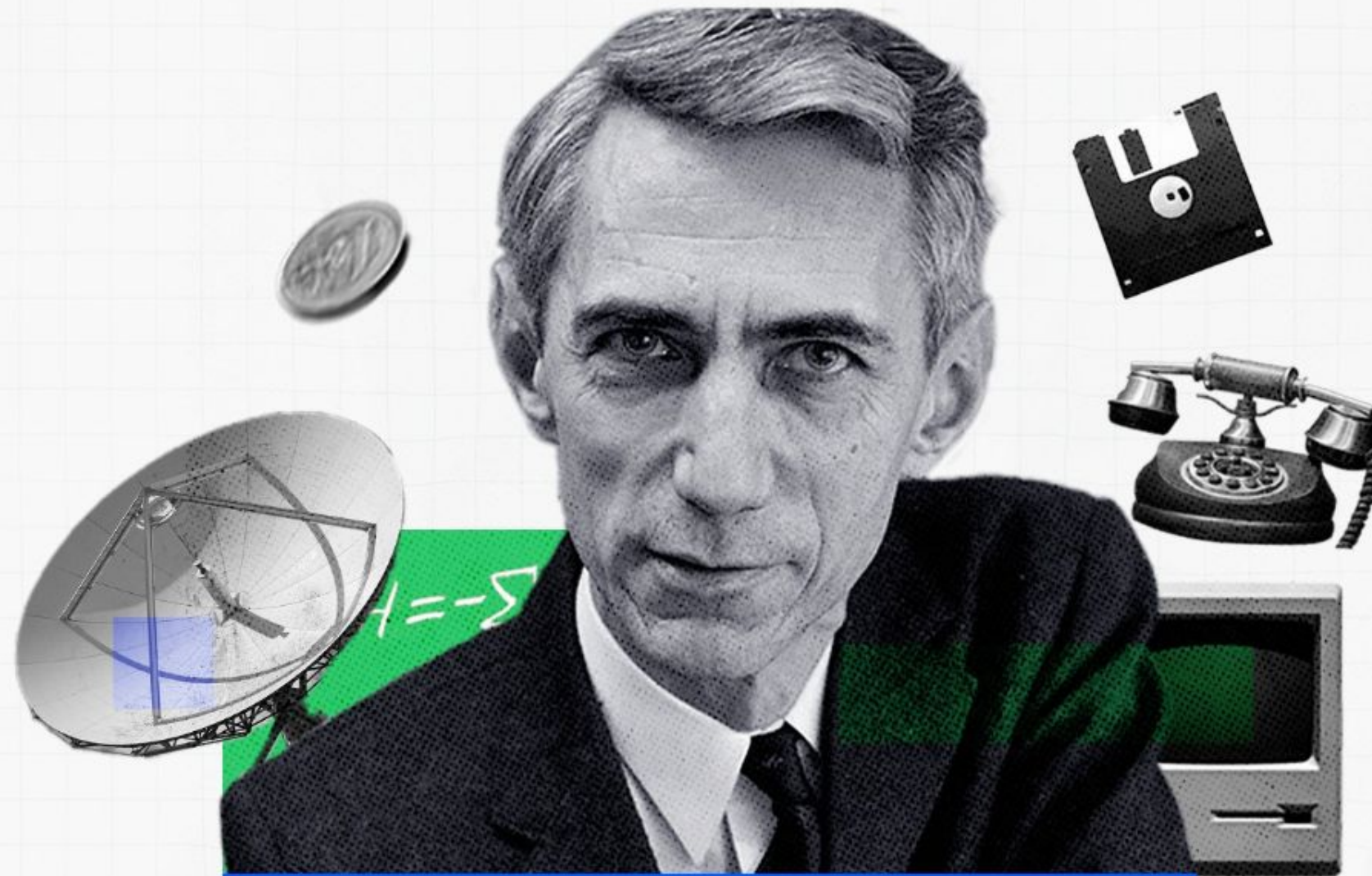
Digital Media

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

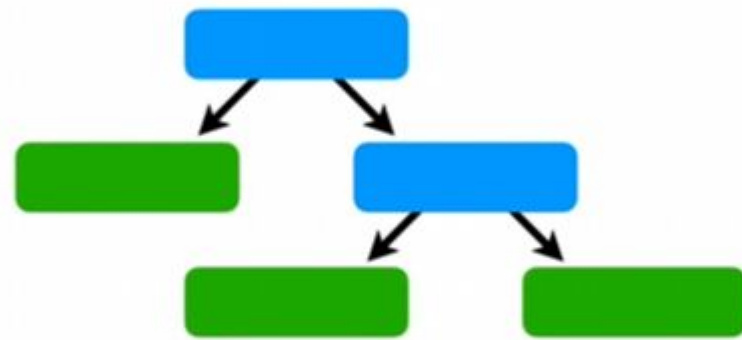
Information Theory established a framework for secure communication by introducing the concepts of randomness and entropy, which are key to modern cryptographic systems and practices.

1101 01111101 11011001 11001010 11101000 10011111
0010 01011111 11010011 10001100 10001101 0111010
0010 01011011 11000101 10001011 01000011 0100011
01000 01011011 01001111 10100101 01111001 0011100
0011 00000100 00001011 10011100 00101000 00010111
0111 00011000 10011101 01111011 01011010 10001101
0101 11110001 00110111 00100100 11010110 10101101
1110 11101111 10100000 10010111 00100001 0001011

What is it?

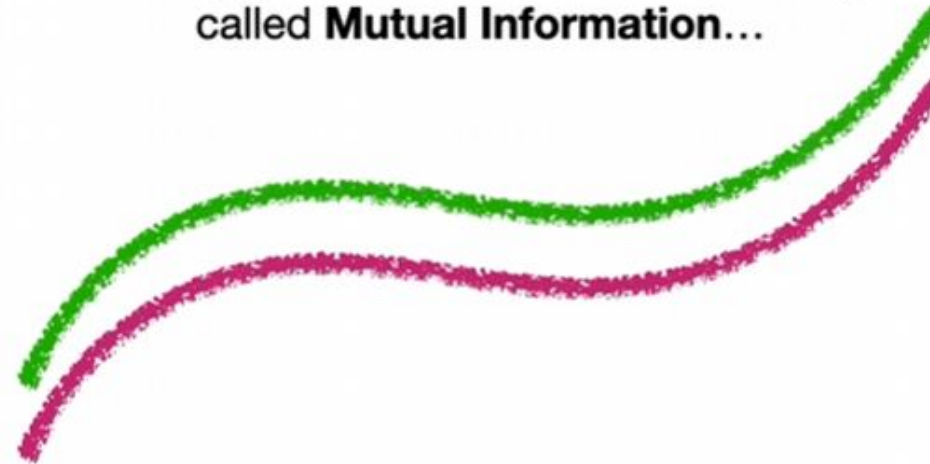
Information theory also drives many algorithms in machine learning.

For example, **Entropy** can be used to build **Classification Trees**...



...which are used to classify things.

Entropy is also the basis of something called **Mutual Information**...



...which quantifies the relationship between two things.

And **Entropy** is the basis of **Relative Entropy** (aka **The Kullback-Leibler Distance**) and **Cross Entropy**...



...which show up all over the place, including fancy dimension reduction algorithms like **t-SNE** and **UMAP**.

What is it?

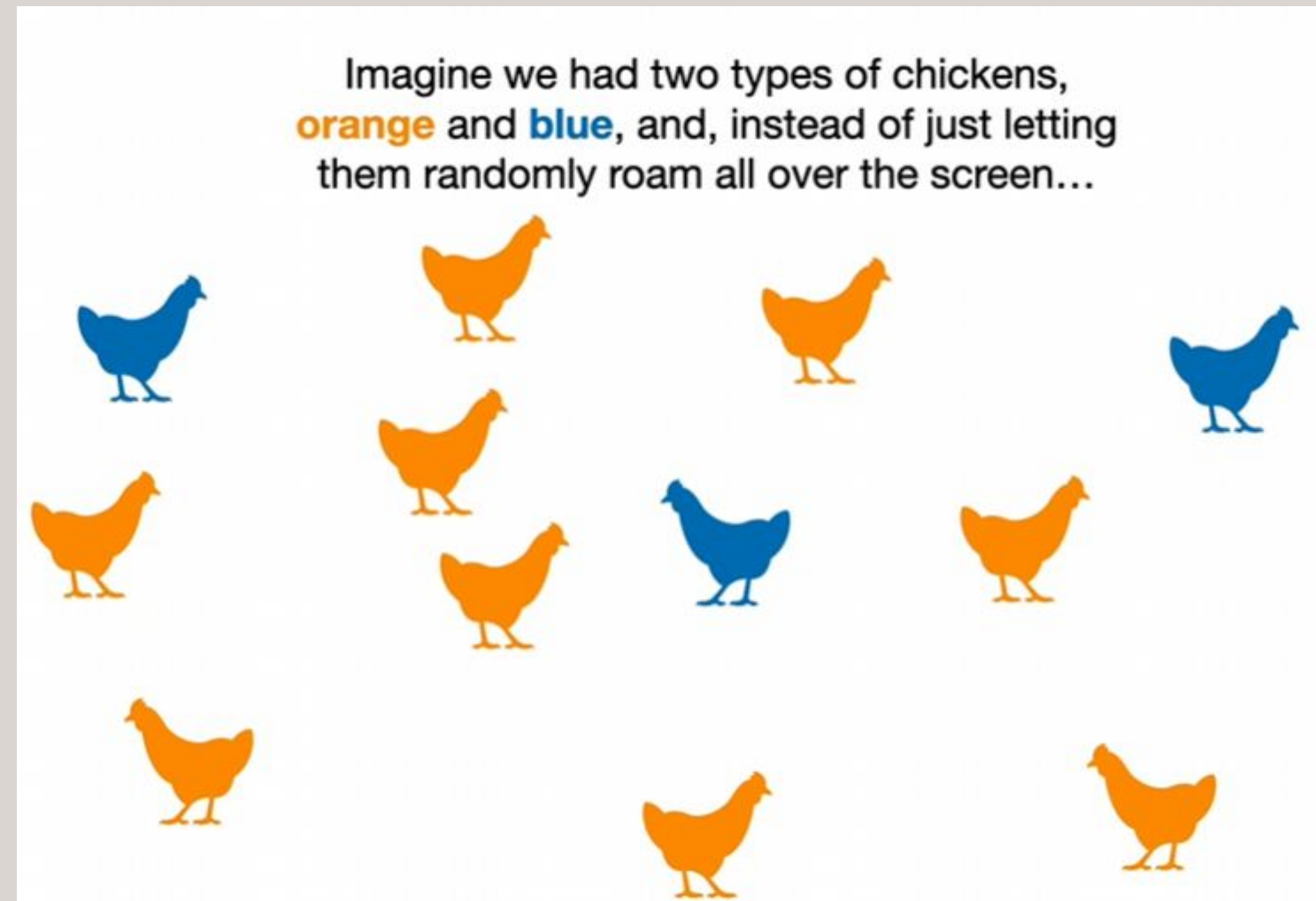
At its core, information theory is about quantifying similarities and differences. In this section, we will be explaining the primary metric for doing so.

What these three things have in common is that they all use **Entropy**, or something derived from it, to quantify *similarities* and *differences*.



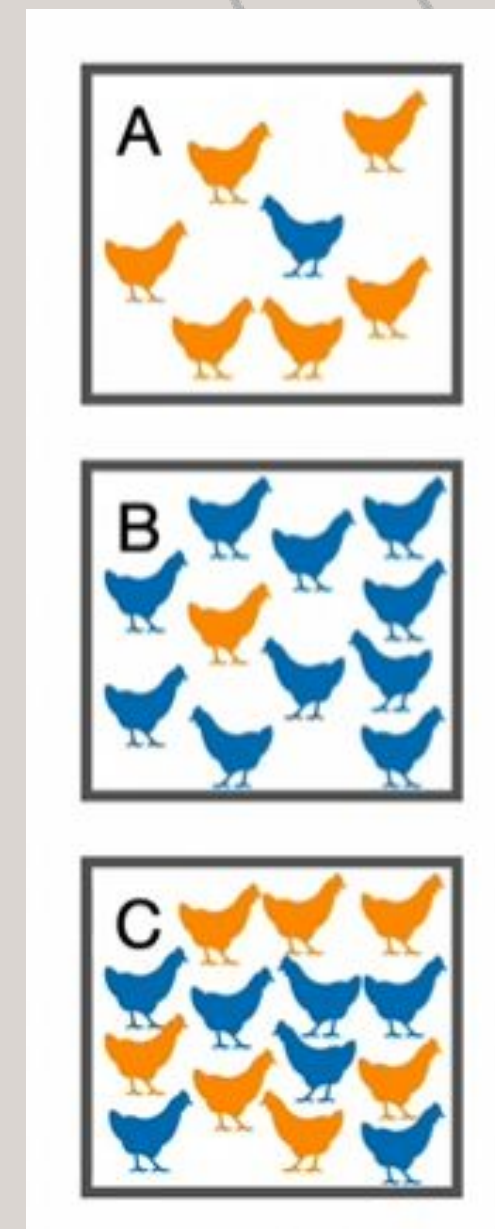
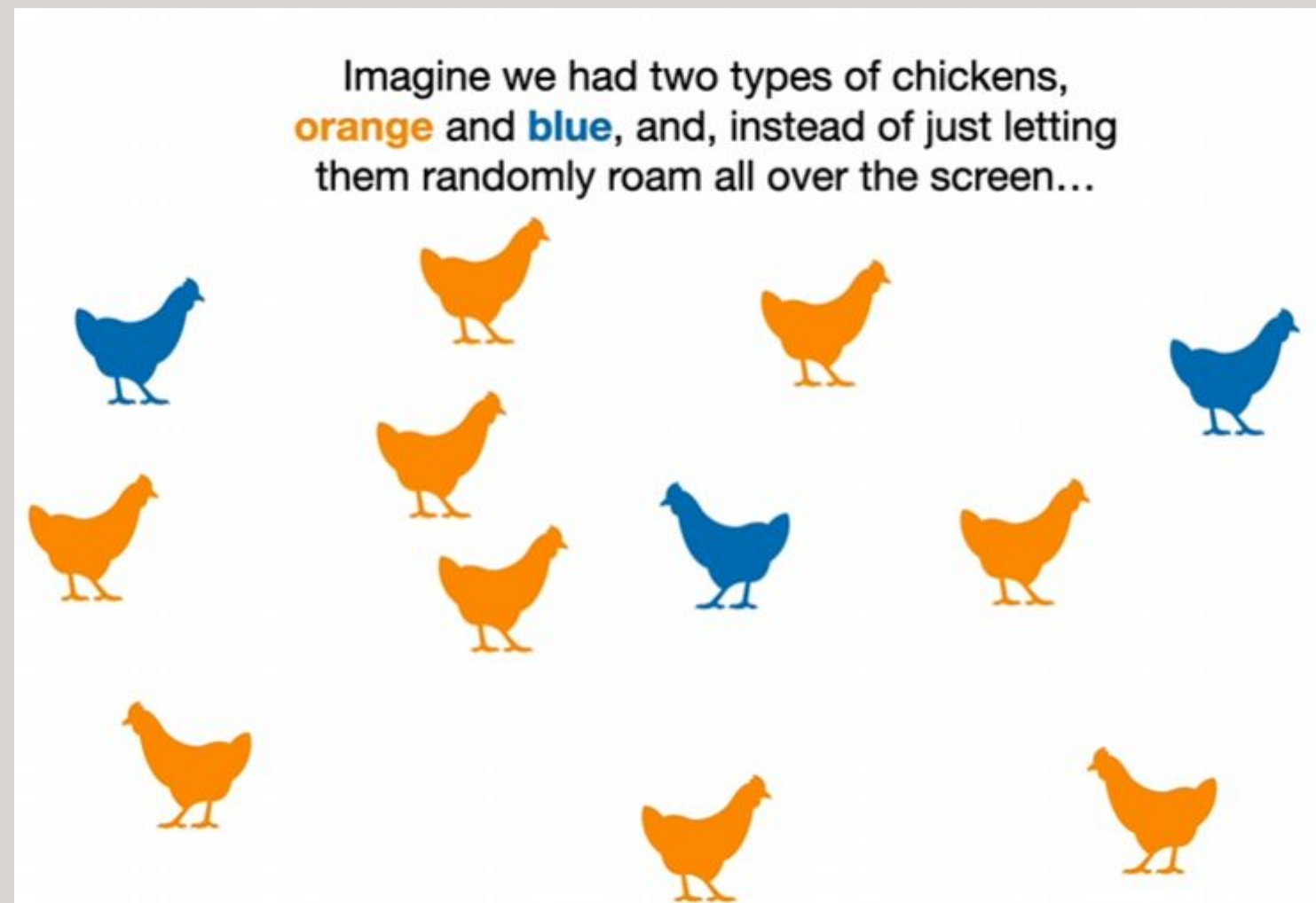
What is surprise?

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



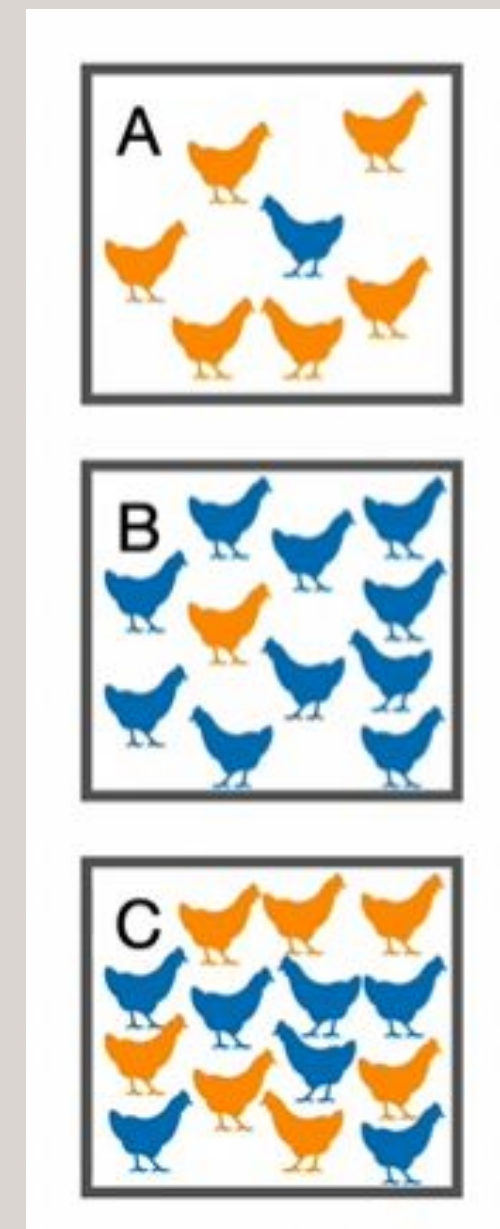
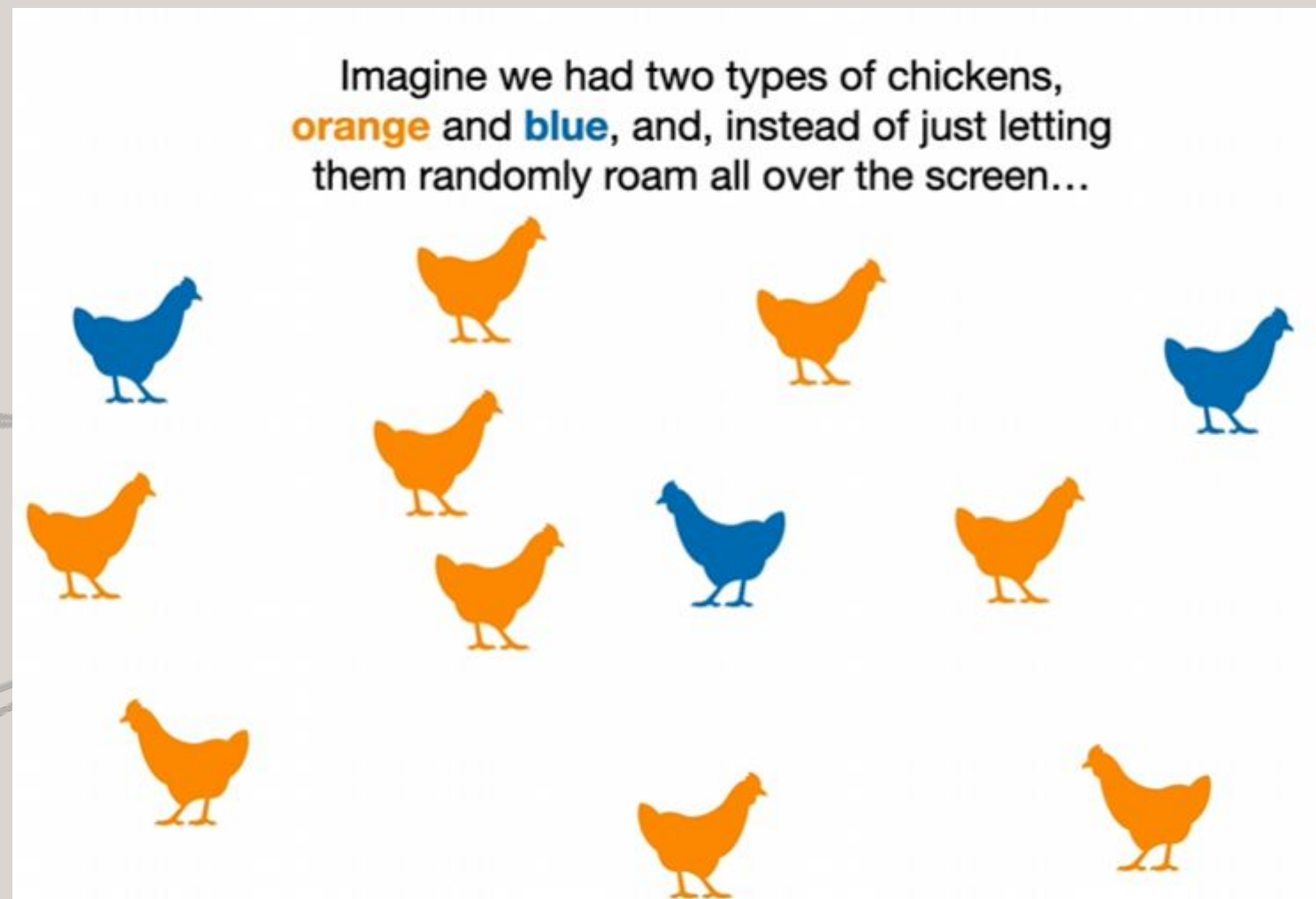
What is surprise?

Surprise seems to be closely tied to probabilities...



What is surprise?

Surprise seems to be closely tied to probabilities...

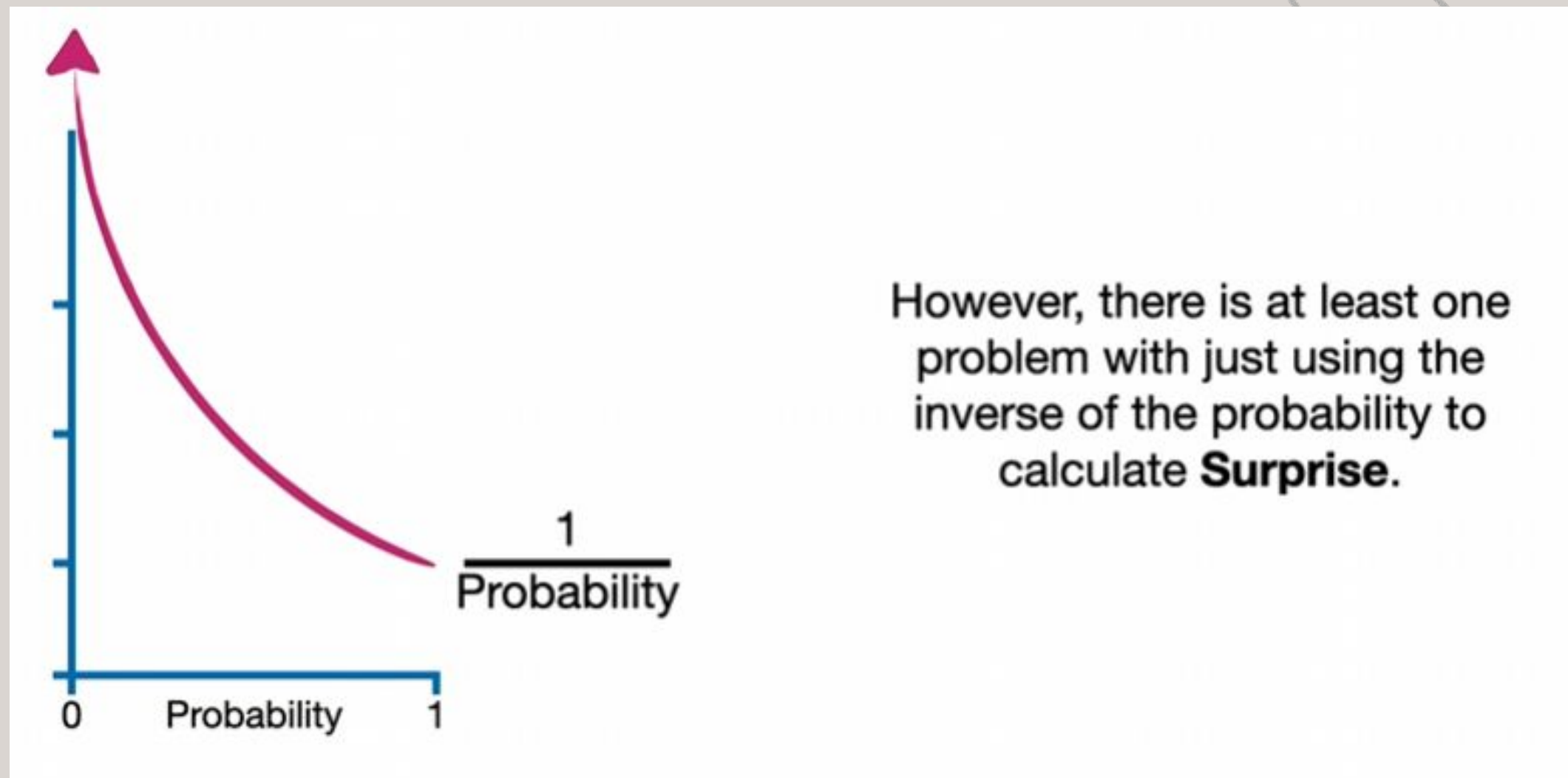


← In other words, when the probability of picking up a **blue** chicken is **low**, the **Surprise** is **high**...

↗ ...and when the probability of picking up a **blue** chicken is **high**, the **Surprise** is **low**.

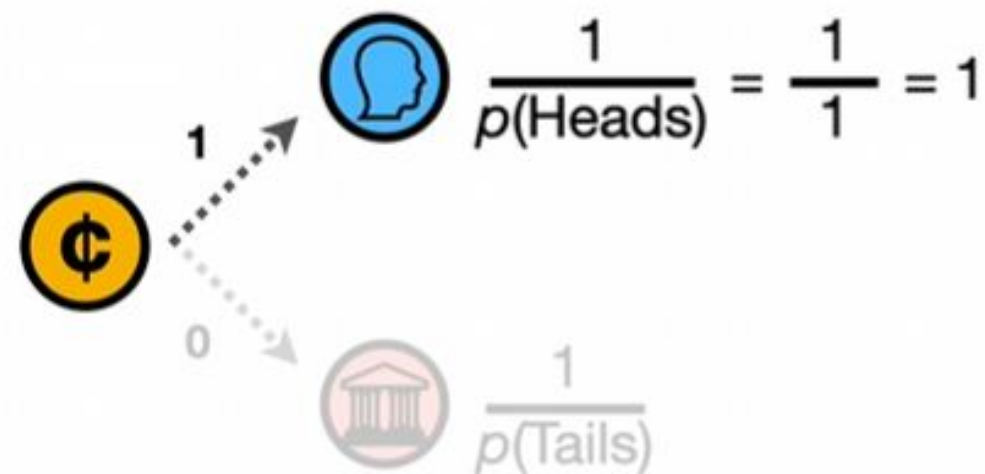
Again, but math

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



Again, but math

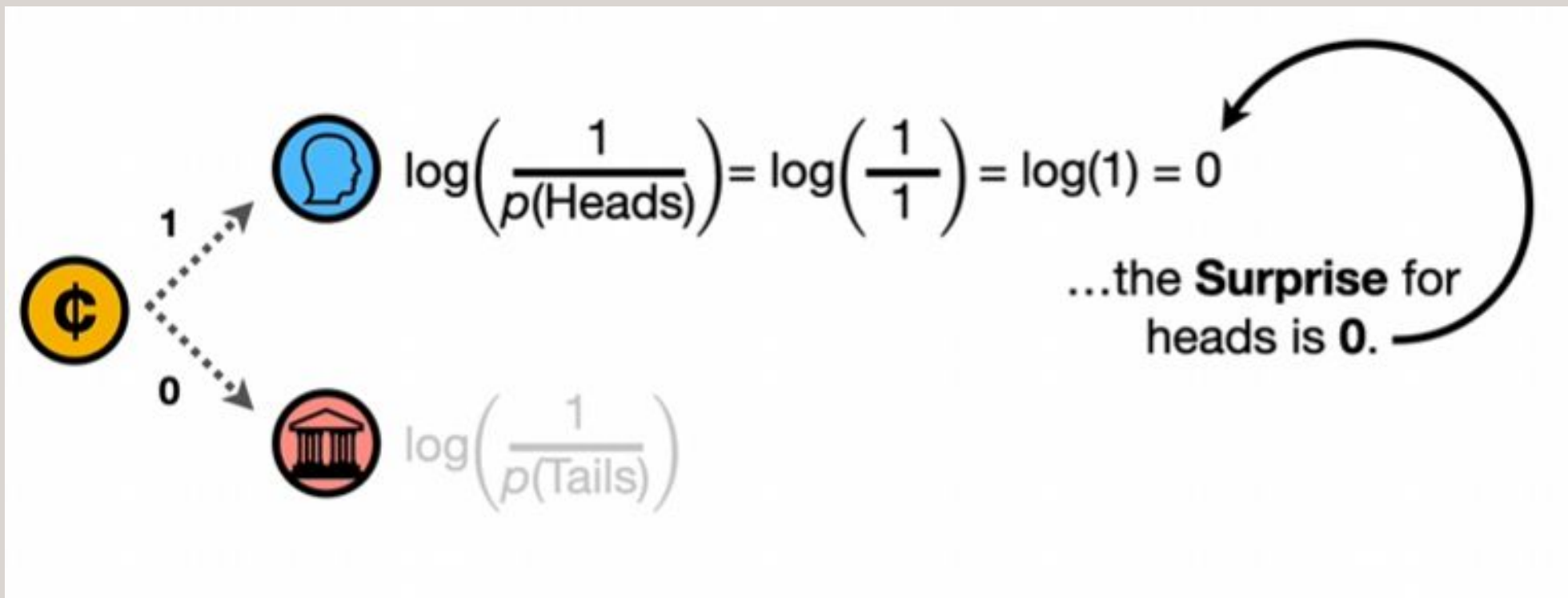
Let's imagine we had a coin that only ever lands on heads $\rightarrow P(\text{heads}) = 1$



And this is one reason why we can't just use the inverse of the probability to calculate **Surprise**.

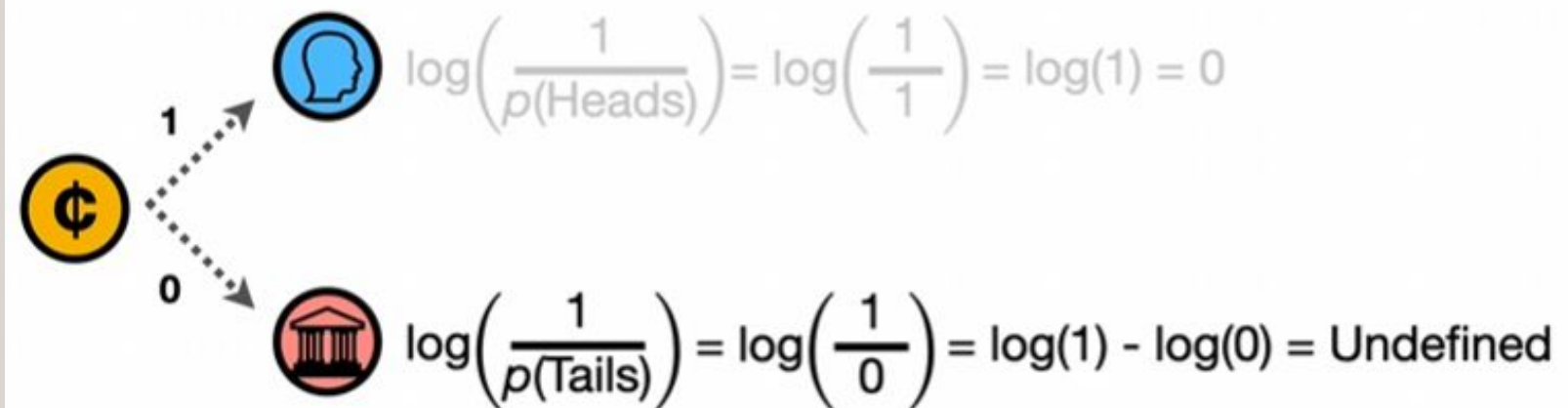
Again, but math

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



Again, but math

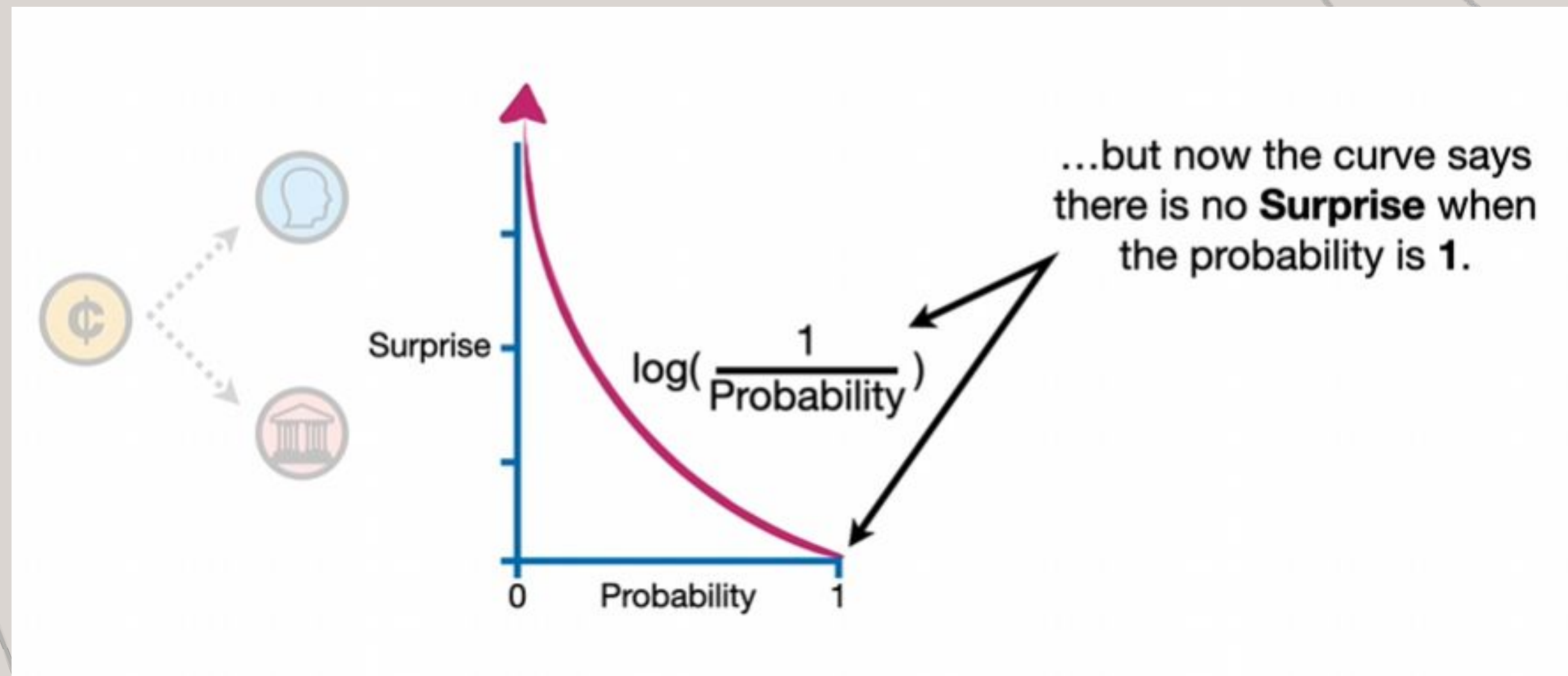
The definition still makes sense for tails, too!



And this result is OK because we're talking about the **Surprise** associated with something that never happens.

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

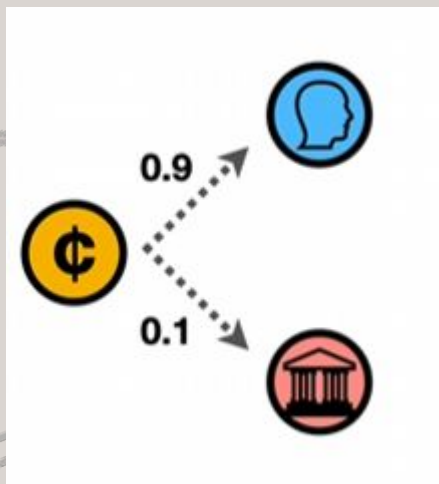
$$h(X=x) = \log_2 \frac{1}{P(X=x)} = -\log_2 P(X=x)$$

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.






A diagram showing three icons at the top: two blue head icons and one red tail icon. Below them is the probability expression $0.9 \times 0.9 \times 0.1$. Below this is the equation $\text{Surprise} = \log_2\left(\frac{1}{0.9 \times 0.9 \times 0.1}\right)$. An arrow points from the probability expression to the equation with the text "...then we can plug this probability into the equation for **Surprise**...".

A diagram showing three icons at the top: two blue head icons and one red tail icon. Below them is the probability expression $0.9 \times 0.9 \times 0.1$. To the right of this is the text: "But, more importantly, we see that the total **Surprise** for a sequence of coin tosses is just the sum of the **Surprises** for each individual toss." Below the probability expression is the equation $\text{Surprise} = \log_2\left(\frac{1}{0.9 \times 0.9 \times 0.1}\right) = \log_2(1) - \log_2(0.9 \times 0.9 \times 0.1)$. Below this is the equation $= \log_2(1) - [\log_2(0.9) + \log_2(0.9) + \log_2(0.1)]$. Below that is the equation $= 0 - \log_2(0.9) - \log_2(0.9) - \log_2(0.1)$. At the bottom, the final result is boxed: $= 0.15 + 0.15 + 3.32 = 3.62$. Arrows connect the text to the relevant parts of the equations.

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: $\log_2\left(\frac{1}{p(x)}\right)$	0.15	3.32

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

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

Properties

We can simplify the expected surprise with sigma notation.

The diagram illustrates the simplification of expected surprise using sigma notation. It features a table of probabilities and a formula for surprise, with arrows indicating the substitution of the formula into the sigma notation.

	Heads	Tails
Probability $p(x)$:	0.9	0.1
Surprise:	0.15	3.32

Surprise:
 $\log_2\left(\frac{1}{p(x)}\right)$

...because now, we can plug the equation for **Surprise** in for x , the specific value...

$\sum x P(X = x)$



Specific value for **Surprise**.

The probability of observing that specific value for **Surprise**.

Properties

We can simplify the expected surprise with sigma notation.

The diagram illustrates the simplification of expected surprise using sigma notation. It features a table of probabilities and surprises, a formula for surprise, and a summation formula for expected surprise. A curved arrow points from the probability column of the table to the $P(X=x)$ term in the summation formula.

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

...and we can plug in the probability...


$$\sum \log\left(\frac{1}{p(x)}\right) P(X=x)$$



Surprise

The probability of observing that specific value for **Surprise**.

Properties

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



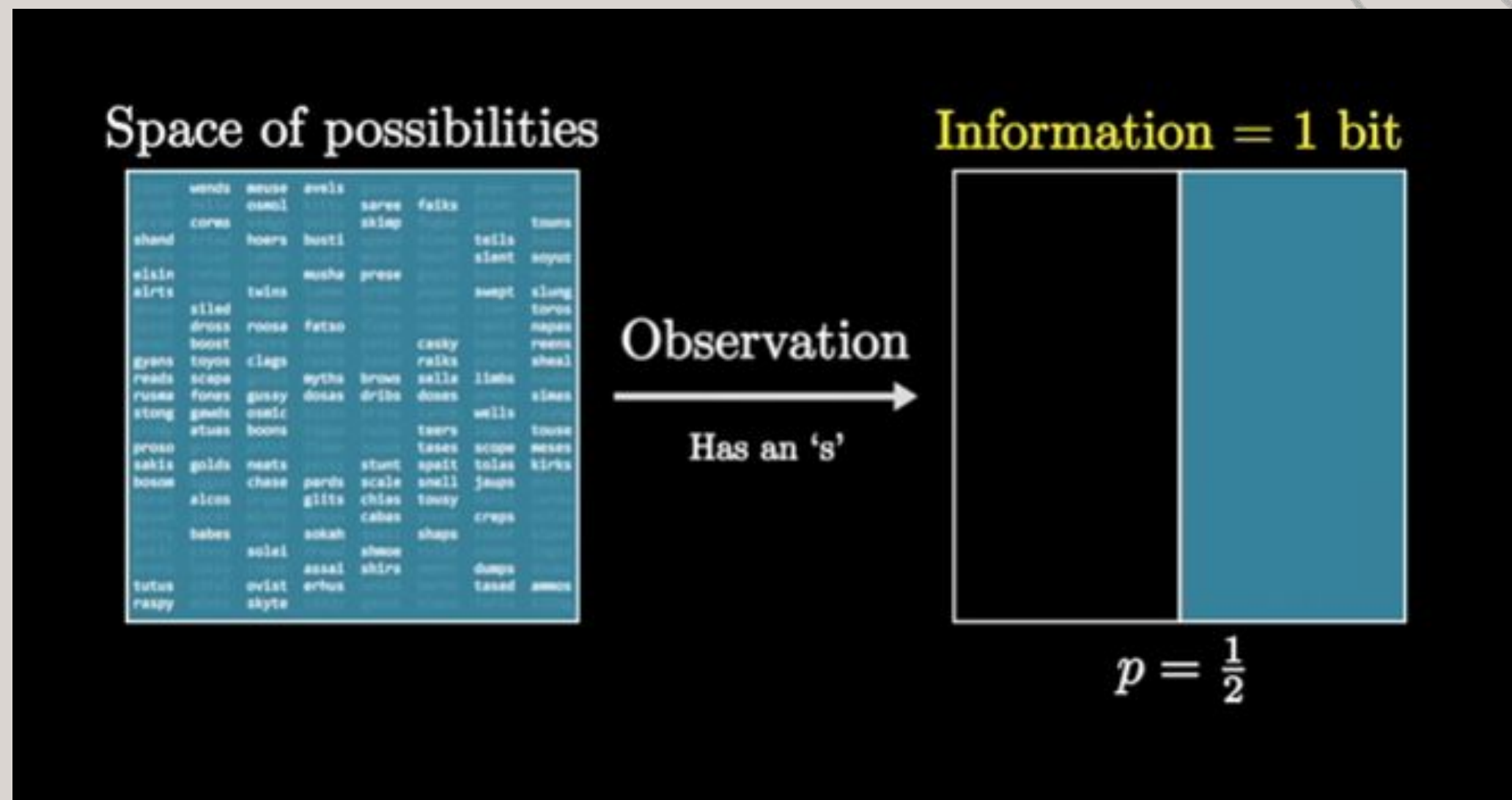
	Heads 	Tails 
Probability $p(x)$:	0.9	0.1
Surprise: $\log_2(\frac{1}{p(x)})$	0.15	3.32

...and we end up with the equation for **Entropy** that Claude Shannon first published in **1948**.

$$\text{Entropy} = \sum p(x) \log\left(\frac{1}{p(x)}\right)$$
$$\text{Entropy} = - \sum p(x) \log(p(x))$$
$$\sum p(x) [\log(1) - \log(p(x))] \rightarrow \sum p(x) [0 - \log(p(x))] \rightarrow \sum -p(x) \log(p(x))$$

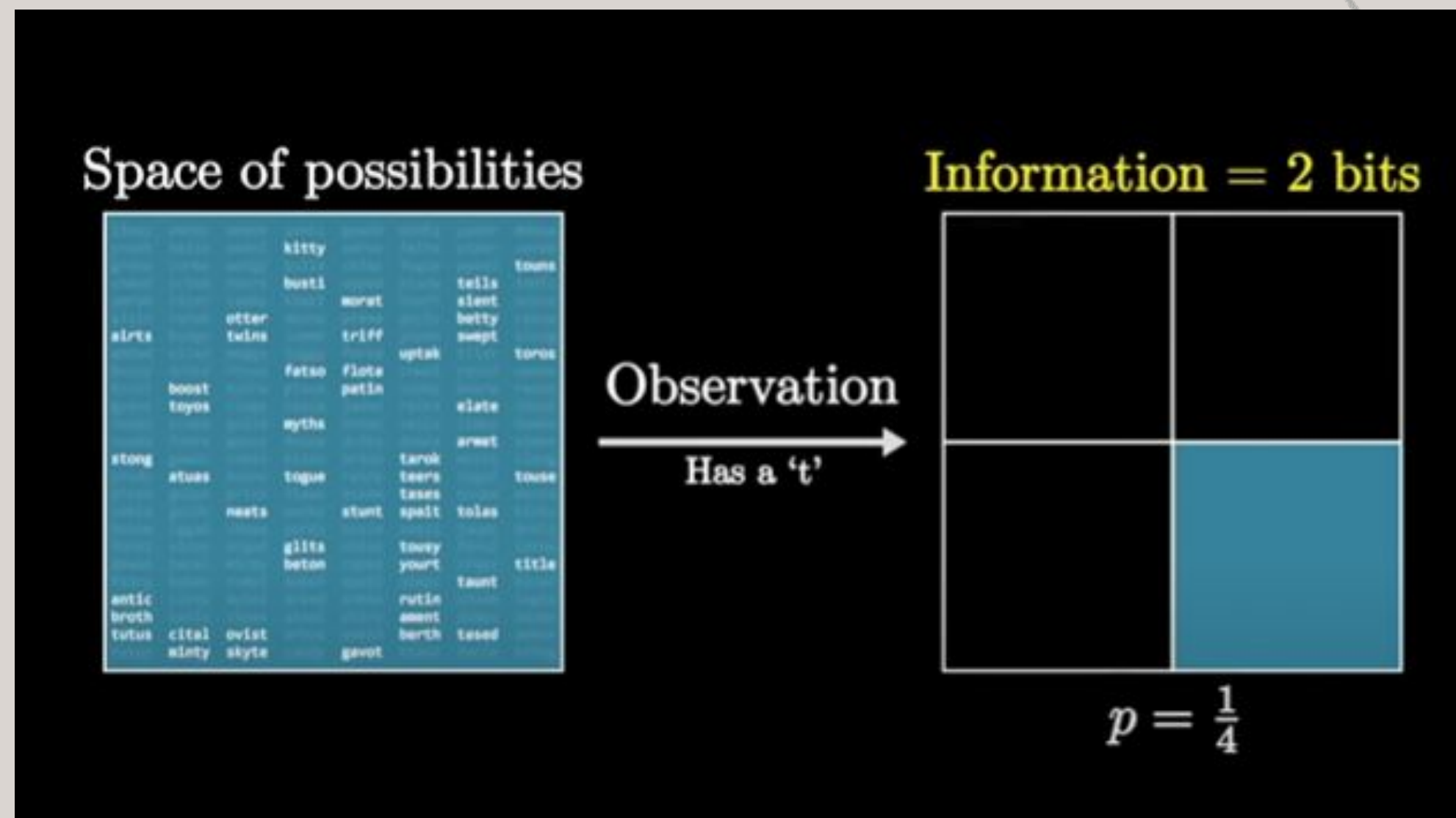
What is a bit?

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



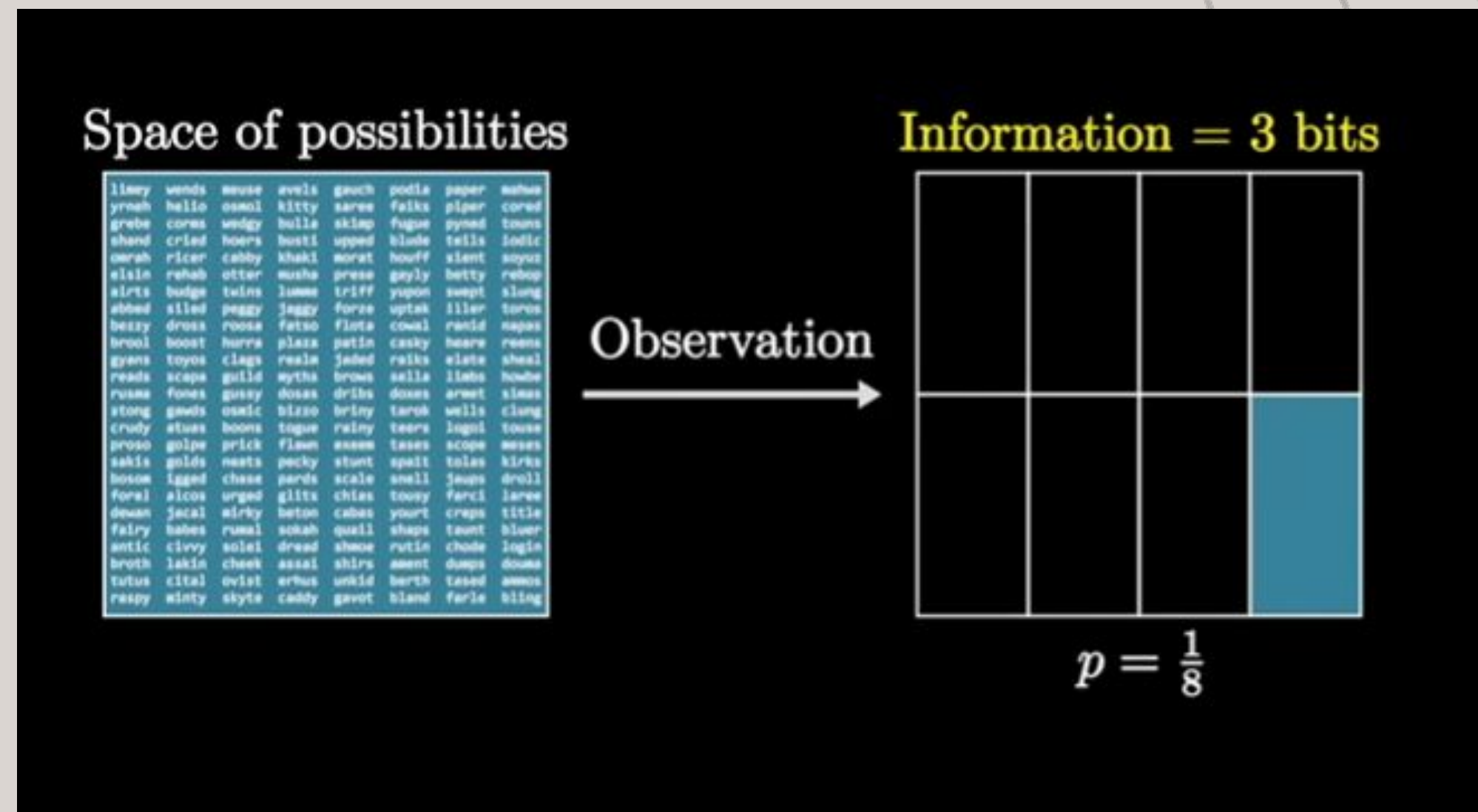
What is a bit?

If we have yet another observation that reduces the search space to $\frac{1}{4}$ of its original size, this gives us 2 bits of information.



What is a bit?

And so on...



What is a bit?

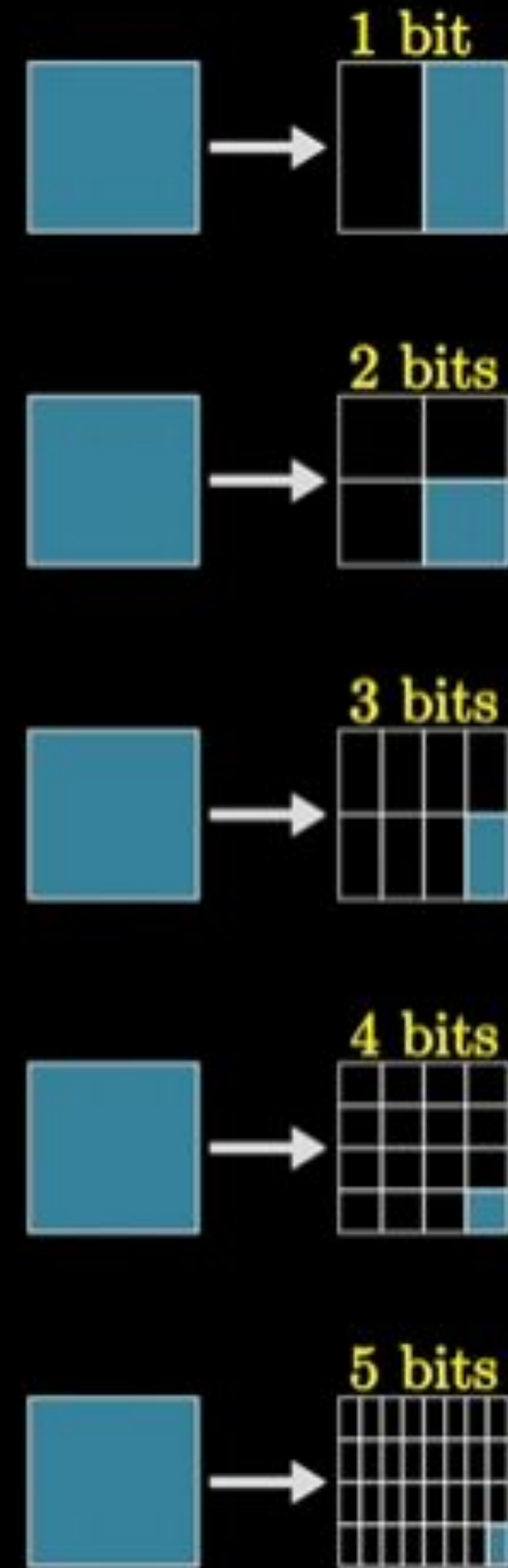
$$\left(\frac{1}{2}\right)^I = p$$

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

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

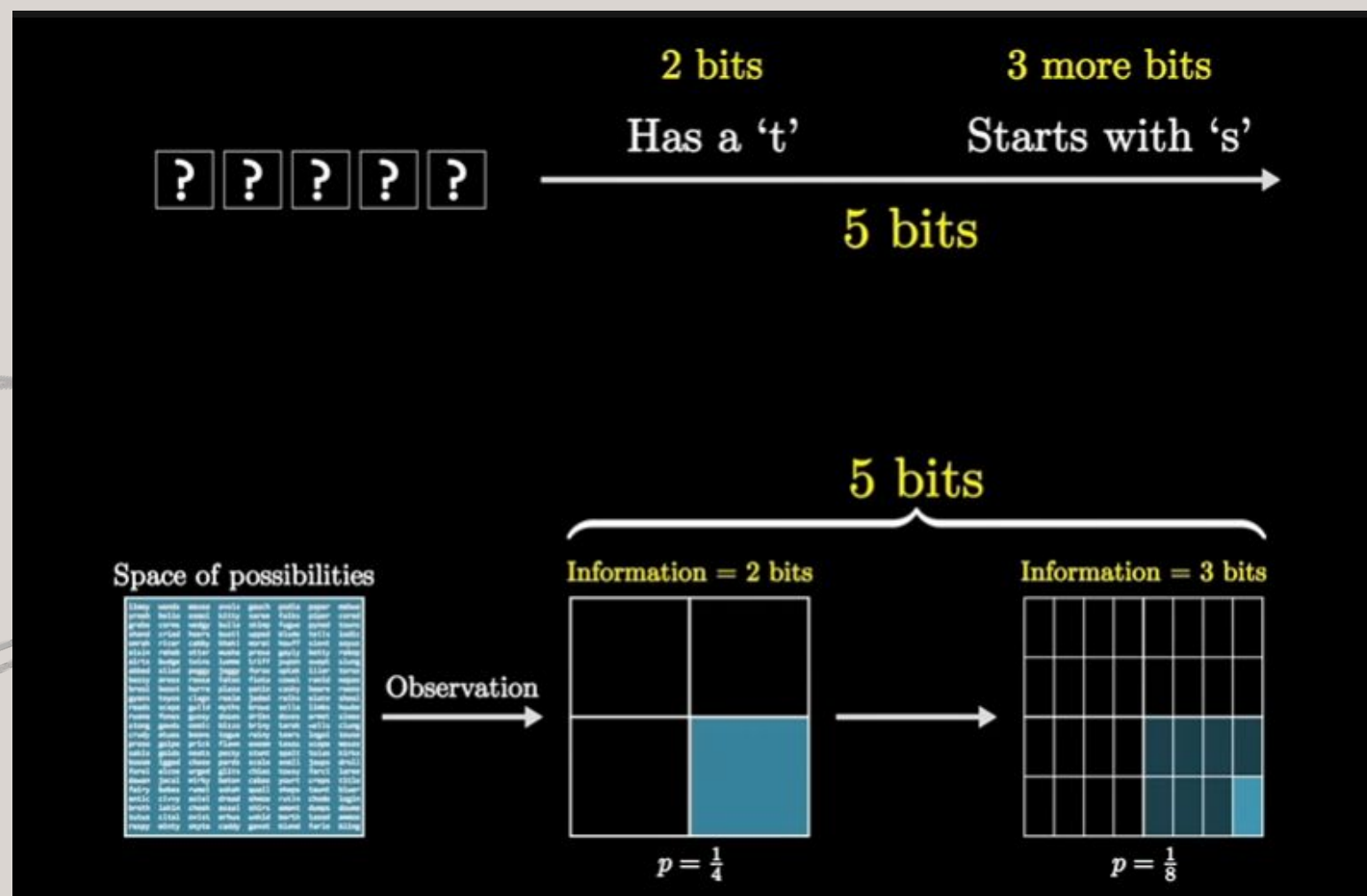
$$I = \log_2 \left(\frac{1}{p} \right)$$

$$I = -\log_2(p)$$



What is a bit?

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




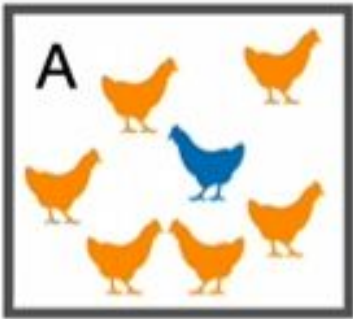
What we want:


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


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

Chickens, again



A 

B 

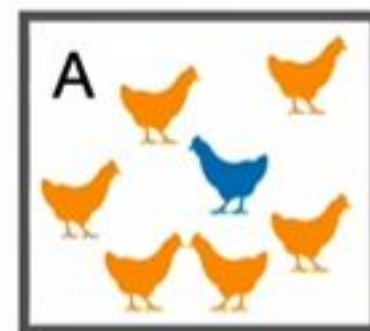
C 

Because **6** of the **7** chickens are **Orange**,
we plug in **6/7** for the probability.

Entropy = $\sum p(x) \log\left(\frac{1}{p(x)}\right)$

Chickens, again

So!

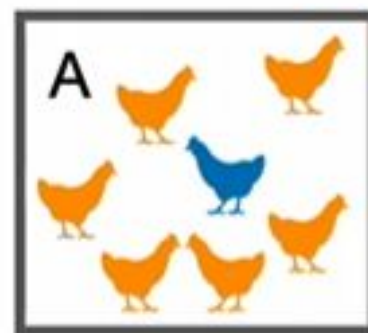


...there is a much higher probability that we will pick up an **orange** chicken (**0.86**)...

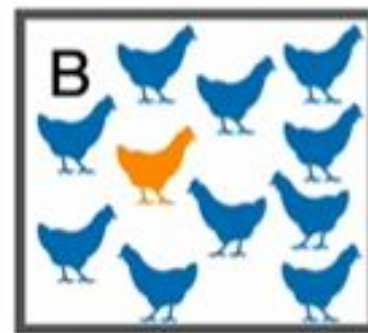
...than pick up a **blue** chicken (**0.14**)...

$$\begin{aligned}\text{Entropy} &= \sum p(x) \log\left(\frac{1}{p(x)}\right) \\ &= \frac{6}{7} \times \log_2\left(\frac{1}{\frac{6}{7}}\right) + \frac{1}{7} \times \log_2\left(\frac{1}{\frac{1}{7}}\right) \\ &= (0.86 \times 0.22) + (0.14 \times 2.81) \\ &= 0.59\end{aligned}$$

Chickens, again



Entropy
= 0.59



Entropy
= 0.44



This make sense because area **B** has a higher probability of picking a chicken with a *lower Surprise*.

$$\begin{aligned}\text{Entropy} &= \sum p(x) \log\left(\frac{1}{p(x)}\right) \\ &= \frac{1}{11} \times \log_2\left(\frac{1}{\frac{1}{11}}\right) + \frac{10}{11} \times \log_2\left(\frac{1}{\frac{10}{11}}\right) \\ &= (0.09 \times 3.46) + (0.91 \times 0.14) \\ &= 0.44\end{aligned}$$

Chickens, again

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

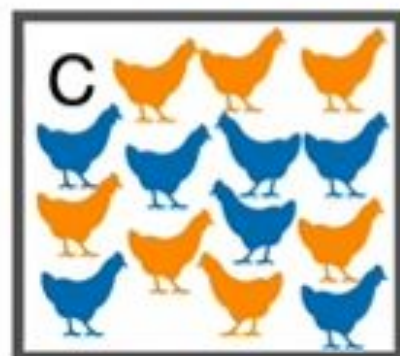
...and as we *increase the difference* in the number of **orange** and **blue** chickens, we lower the **Entropy**.



Entropy
= 0.59



Entropy
= 0.44



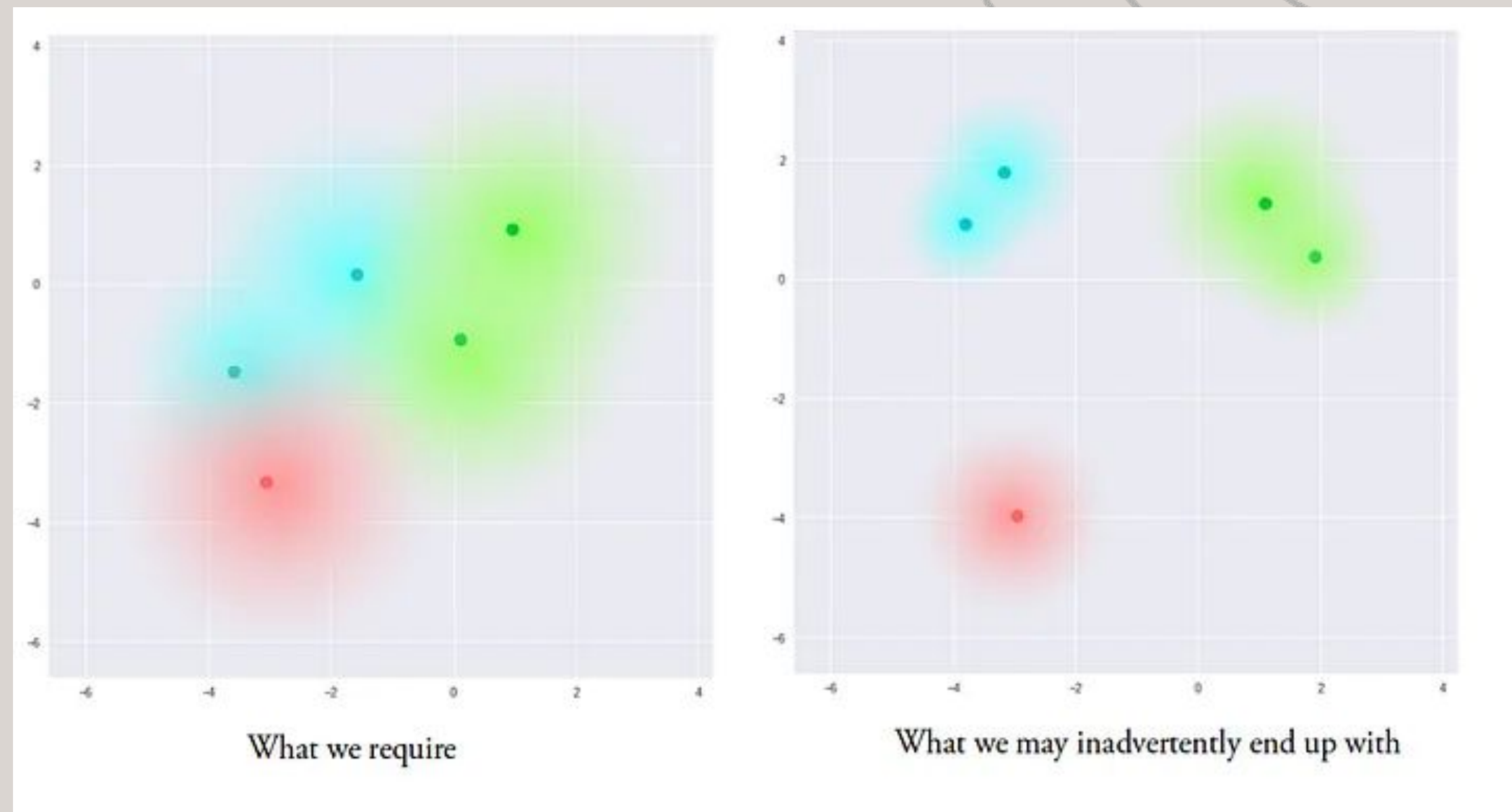
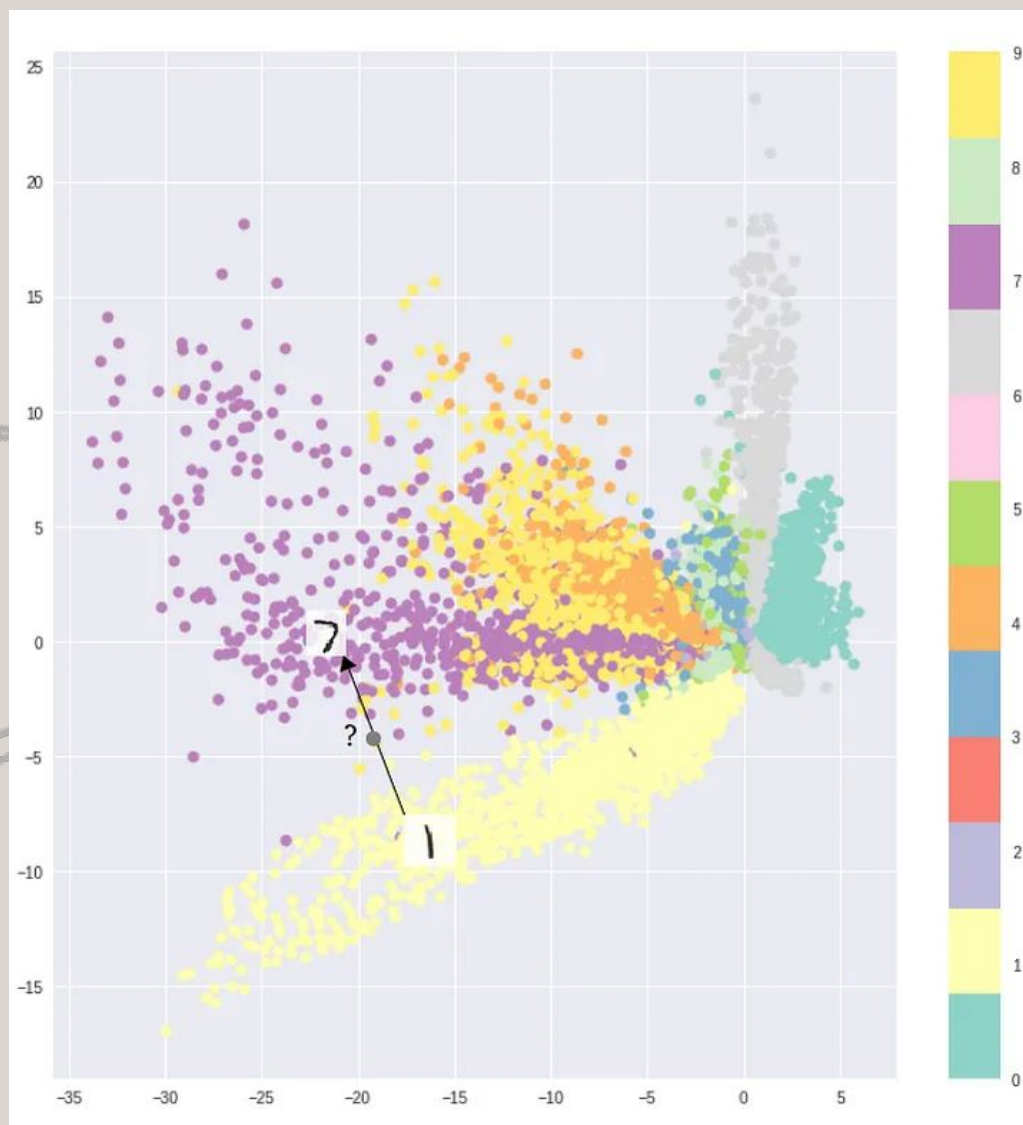
Entropy
= 1

...we always get the same, relatively moderate, **Surprise** every time we pick up a chicken...

$$\begin{aligned}\text{Entropy} &= \sum p(x) \log_2 \left(\frac{1}{p(x)} \right) \\ &= \frac{7}{14} \times \log_2 \left(\frac{1}{\frac{7}{14}} \right) + \frac{7}{14} \times \log_2 \left(\frac{1}{\frac{7}{14}} \right) \\ &= (0.5 \times 1) + (0.5 \times 1) \\ &= 1\end{aligned}$$

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	
$\begin{cases} 0.5 \\ 0.5 \end{cases}$	<i>heads</i> <i>tails</i>	$\begin{cases} 0.55 \\ 0.45 \end{cases}$	<i>heads</i> <i>tails</i>
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?

Coin 1		Coin 2	
$\begin{cases} 0.5 \\ 0.5 \end{cases}$	<i>heads</i> <i>tails</i>	$\begin{cases} 0.95 \\ 0.05 \end{cases}$	<i>heads</i> <i>tails</i>
H H T H H T T H T H T H		H H H H H H T H H H H H	

KL divergence

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

True Coin		Coin 2									
$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$		$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$									
H	H	T	H	H	T	H	H	H	T	H	T
$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$											
$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$											

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

KL divergence

We can then apply log normalization to these probabilities.

True Coin		Coin 2	
$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$		$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$	
H H T H H T H H H T H T			
$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$		$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$	

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

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

KL divergence

And then expand out the terms using log rules.

True Coin		Coin 2									
$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$		$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$									
H	H	T	H	H	T	H	H	H	T	H	T
$p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1 \cdot p_1 \cdot p_2 \cdot p_1 \cdot p_1$											
$q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_1 \cdot q_1 \cdot q_2 \cdot q_1 \cdot q_2$											

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

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

KL divergence

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

True Coin		Coin 2									
$\begin{cases} p_1 & \text{heads} \\ p_2 & \text{tails} \end{cases}$		$\begin{cases} q_1 & \text{heads} \\ q_2 & \text{tails} \end{cases}$									
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

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

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

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

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

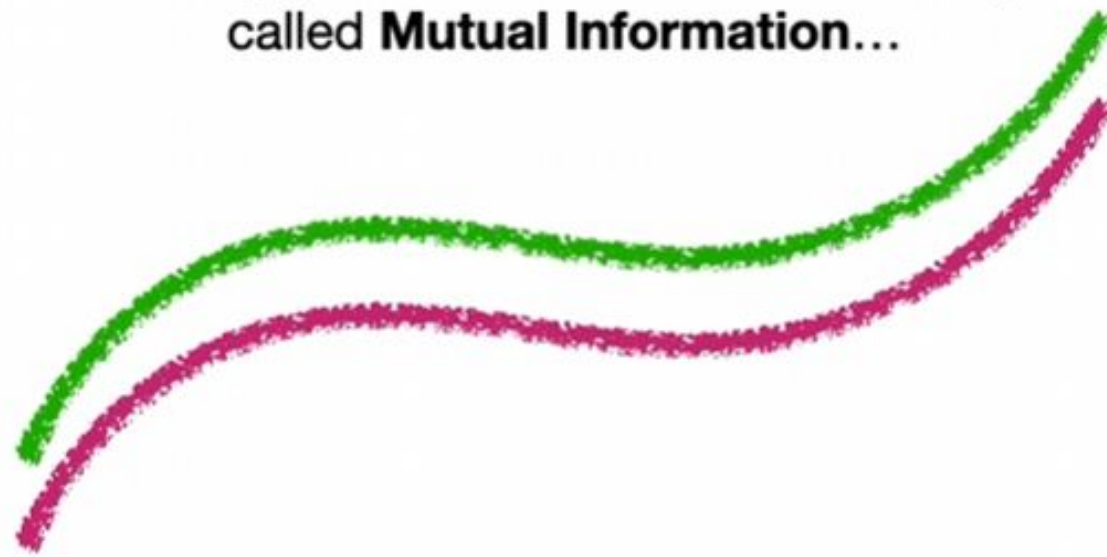
$$\text{Log} \left(\frac{P(\text{Observations of d1} \mid \text{distribution 1})}{P(\text{Observations of d1} \mid \text{distribution 2})} \right)$$

*d1 = distribution 1

KL divergence

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

Entropy is also the basis of something called **Mutual Information...**



...which quantifies the relationship between two things.

And **Entropy** is the basis of **Relative Entropy** (aka **The Kullback-Leibler Distance**) and **Cross Entropy...**



...which show up all over the place, including fancy dimension reduction algorithms like **t-SNE** and **UMAP**.

quAldditch

What is the formula for entropy?



**Thank
You**