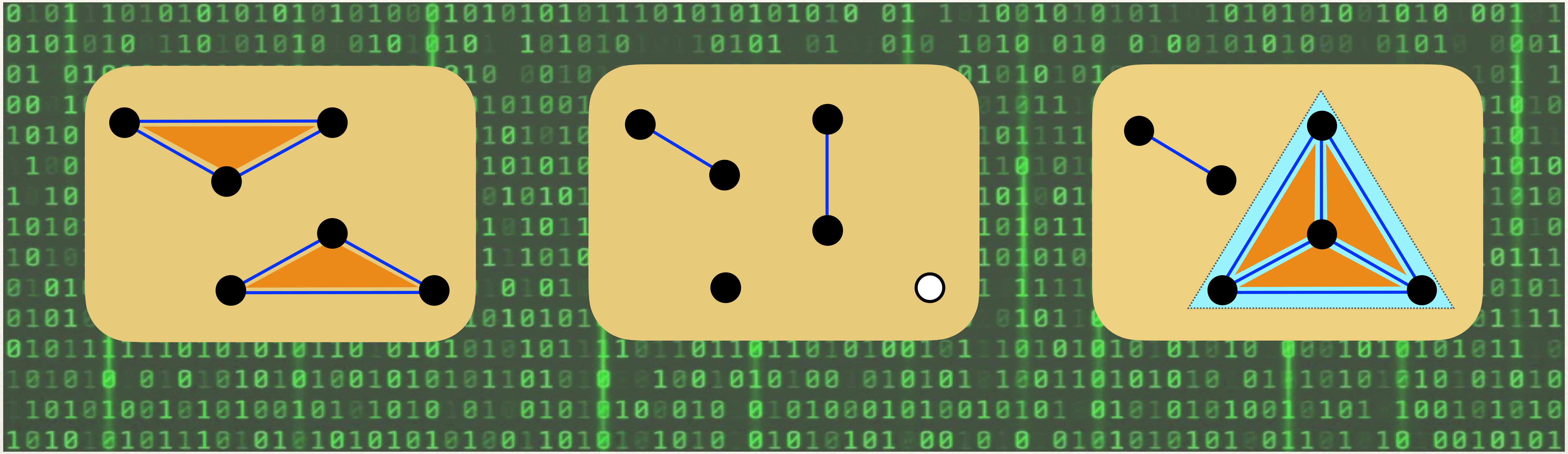


Understanding Big Data with Statistical Physics

Viva Fisica — January 27, 2023

Clélia de Mulatier

UNIVERSITY OF AMSTERDAM

Who am I?

Started in UvA in January 2021, as an Assistant professor in **Theoretical Physics** and in **Informatics**

I develop **new theoretical** and **computational methods** to study **complex systems**.

New tools to analyze **high-dimensional data**

How I ended up here…

Studied in France:

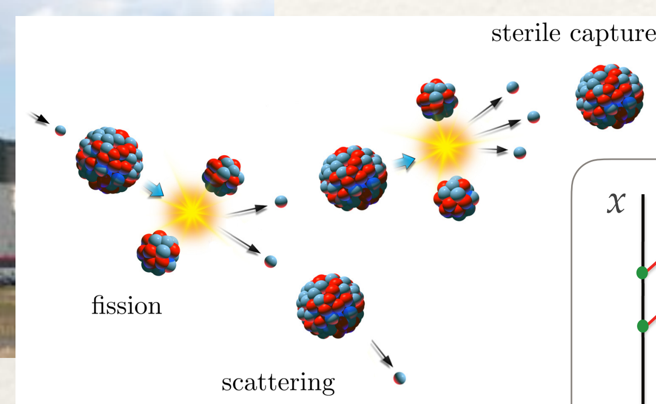
— 2-year **BSc in Math/Physics**, minor in Informatics

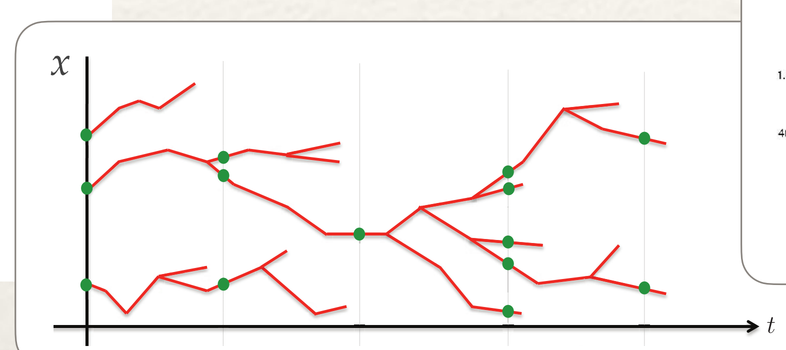
— 2-year **BSc/MSc in Fundamental Physics** (Paris)

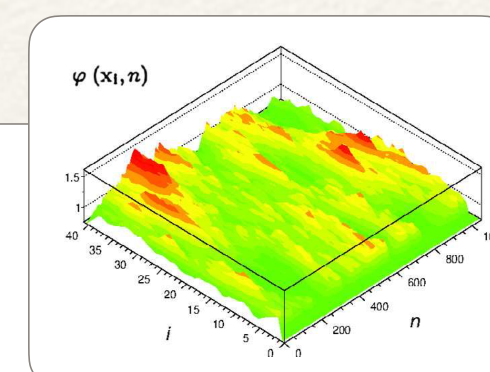
— National Exam to become a Physics/Chemistry teacher in France

— 1-year **Research MSc in "Physics of Complex Systems"** (France-Italy)

PhD in Statistical Physics:



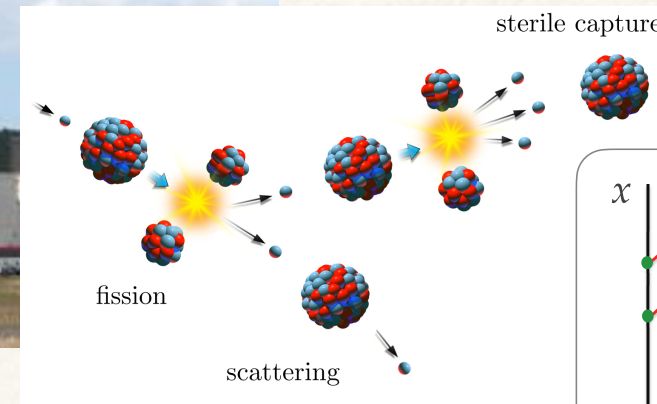




Studied the statistical properties
of the population of neutrons inside a nuclear reactor

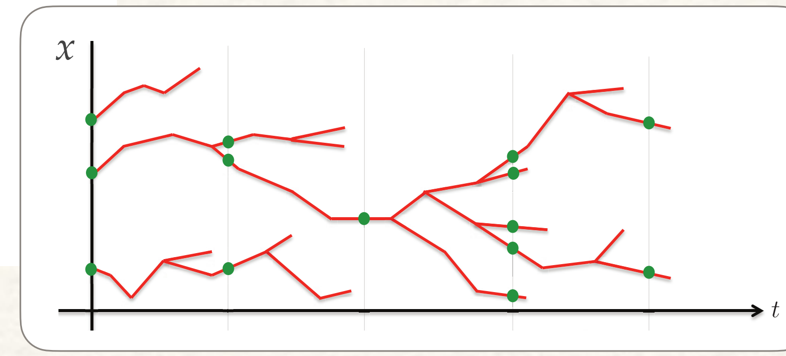
How I ended up here…

PhD in Statistical Physics:



UNIVERSITÉ PARIS SUD

cea



Studied statistical properties

of the population of neutrons in a nuclear reactor

Postdoc in Italy (2 years):

ICTP The Abdus Salam International Centre for Theoretical Physics

More Statistical Physics!

Data analysis

control theory, reinforcement learning

Multi-agent systems, game theory

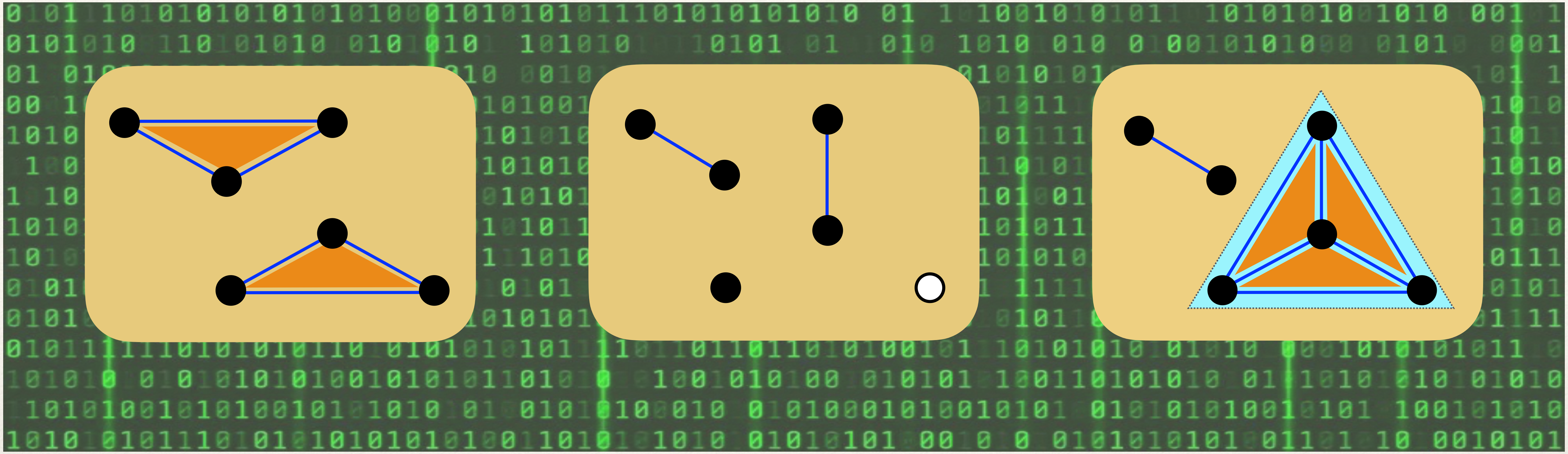
Postdoc in the US (3 years):

Penn UNIVERSITY of PENNSYLVANIA

Neuroscience

Animal behavior

Robotic, What is curiosity? (Honda)



Understanding Big Data with Statistical Physics

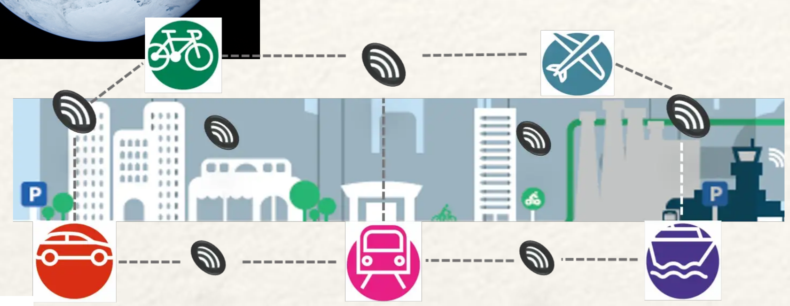
Viva Fisica — January 27, 2023

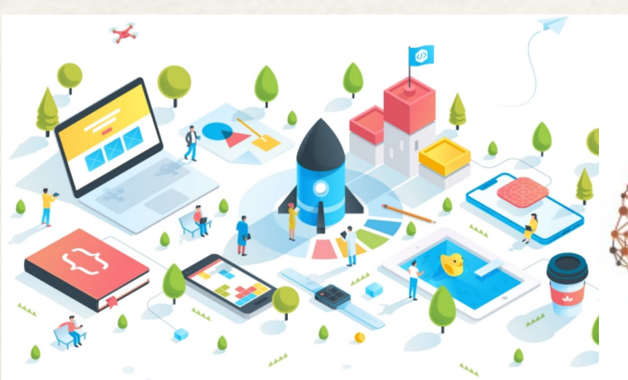
Clélia de Mulatier

University of Amsterdam

What is Big Data?

Examples:





What is Big Data?

Examples:

Properties of Big Data?

	Conditions diseases			Symptoms			Treatment			Profile
	Chronic migraine	Insomnia	…	Difficulty falling asleep	Waking up too early	…	…	Paracetamol	…	Blood type
Patient 1	1	1		1	1			3g/day		O
Patient 2	1	0		0	0					A
…										

What is Big Data?

Examples:

Properties of Big Data?

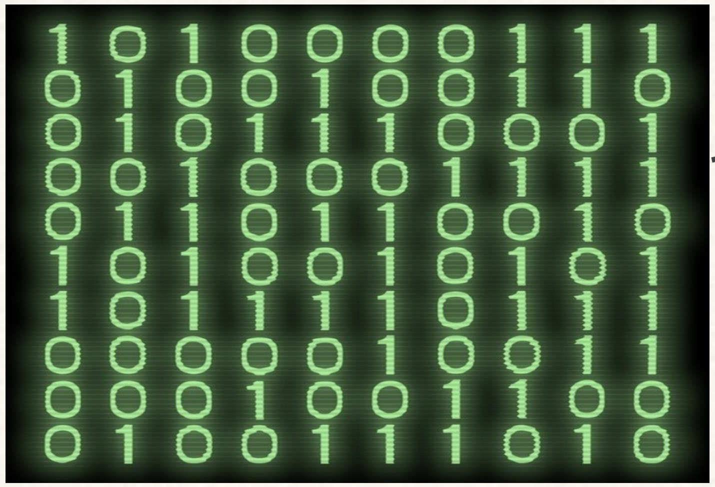
— Lots of variables

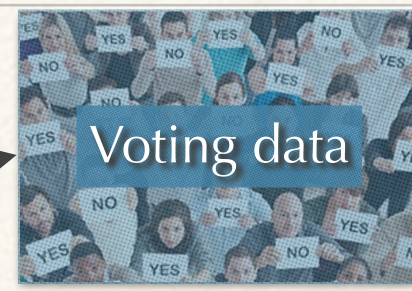
— Lots of datapoints

— Variables can have different types

	Conditions diseases			Symptoms			Treatment			Profile
	Chronic migraine	Insomnia	…	Difficulty falling asleep	Waking up too early	…	…	Paracetamol	…	Blood type
Patient 1	1	1		1	1			3g/day		O
Patient 2	1	0		0	0					A
…										

What do we want to do with this data?






Statistical Mechanics of the US Supreme Court

Edward D. Lee, Chase P. Broedersz & William Bialek

Journal of Statistical Physics **160**, 275–301(2015) | Cite this article

US Supreme Court

9 justices, 895 votes

Conservative (1) or Liberal (0)

2nd Rehnquist Court
(1994-2005)

result of a vote

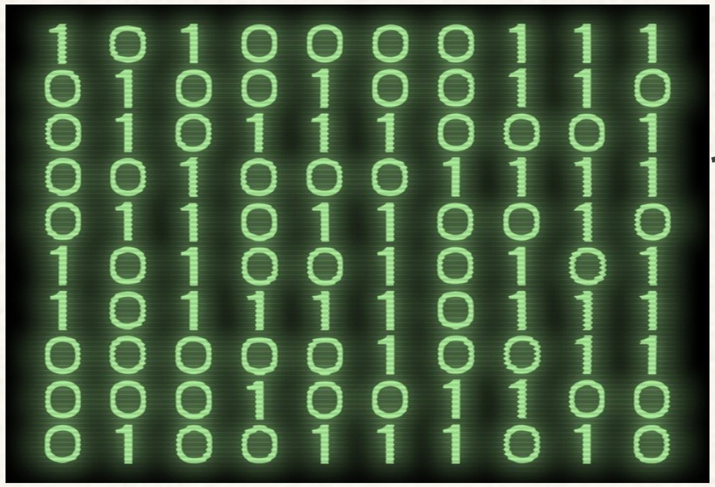
9 justices

SB	1	0	1	0	0	0	0	1	
WR	0	1	0	0	1	0	0	1	
JS	0	1	0	1	1	1	0	1	
SO	0	0	1	0	0	0	1	1	
AS	0	1	0	0	1	1	0	1	…
AK	1	0	0	0	0	1	0	1	
DS	1	0	0	1	1	1	0	1	
CT	0	0	1	0	0	1	0	1	
RG	0	0	1	1	0	0	1	1	

895 votes

Noisy Data

What do we want to do with this data?



Statistical Mechanics of the US Supreme Court

Edward D. Lee, Chase P. Broedersz & William Bialek

Journal of Statistical Physics **160**, 275–301(2015) | Cite this article

US Supreme Court

9 justices, 895 votes

Conservative (1) or **Liberal** (0)

2nd Rehnquist Court
(1994-2005)

result of a vote

9 justices

SB	1	0	1	0	0	0	0	1	
WR	0	1	0	0	1	0	0	1	
JS	0	1	0	1	1	1	0	1	
SO	0	0	1	0	0	0	1	1	
AS	0	1	0	0	1	1	0	1	…
AK	1	0	0	0	0	1	0	1	
DS	1	0	0	1	1	1	0	1	
CT	0	0	1	0	0	1	0	1	
RG	0	0	1	1	0	0	1	1	

895 votes

Patterns:

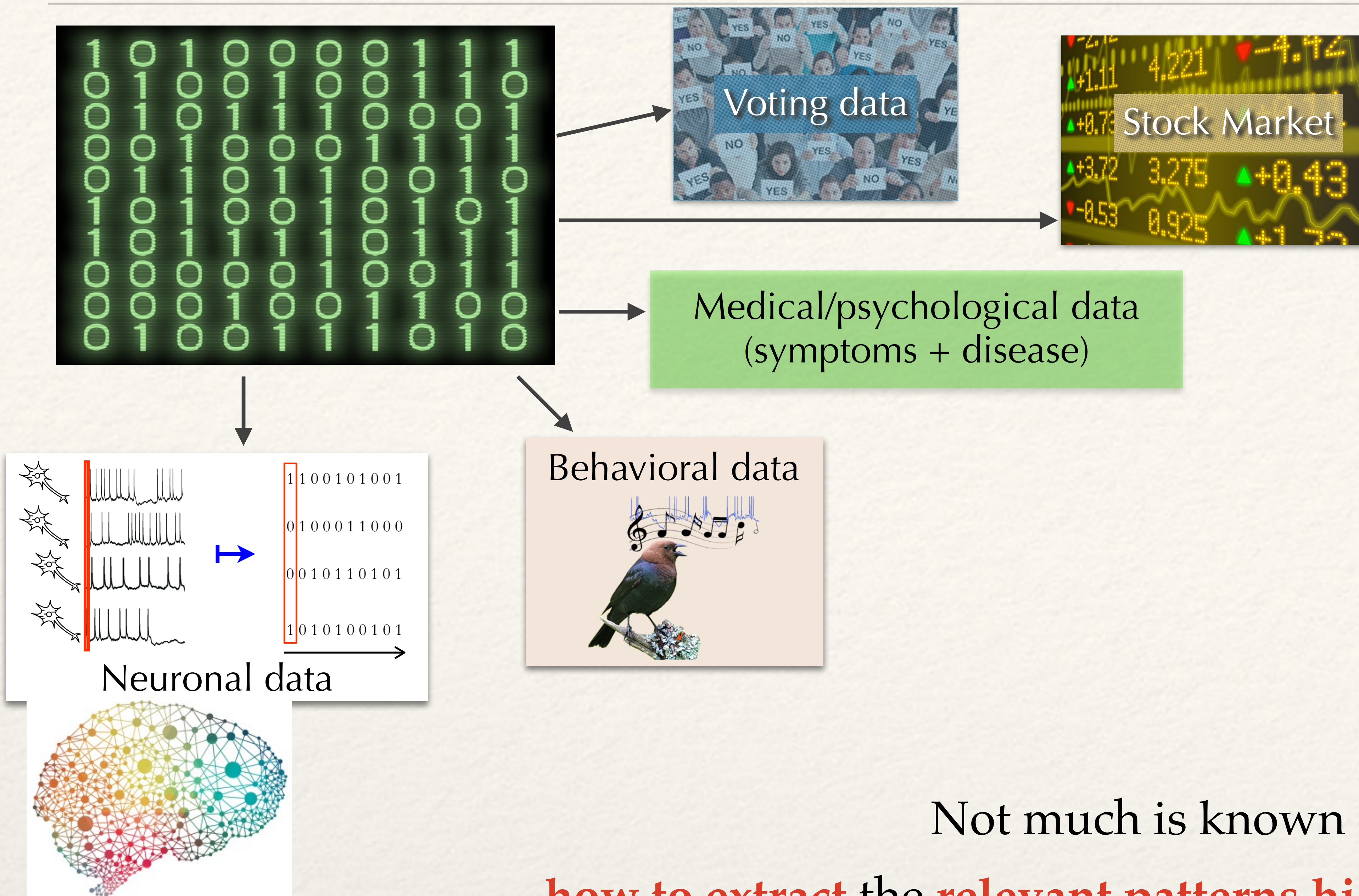
- Certain states can be more frequent than others
- Some justices are more likely to vote C or L
- Some justices are more likely to vote similarly

Etc.…

- Extract from the noisy data which patterns are most redundant

Extract useful “Information”

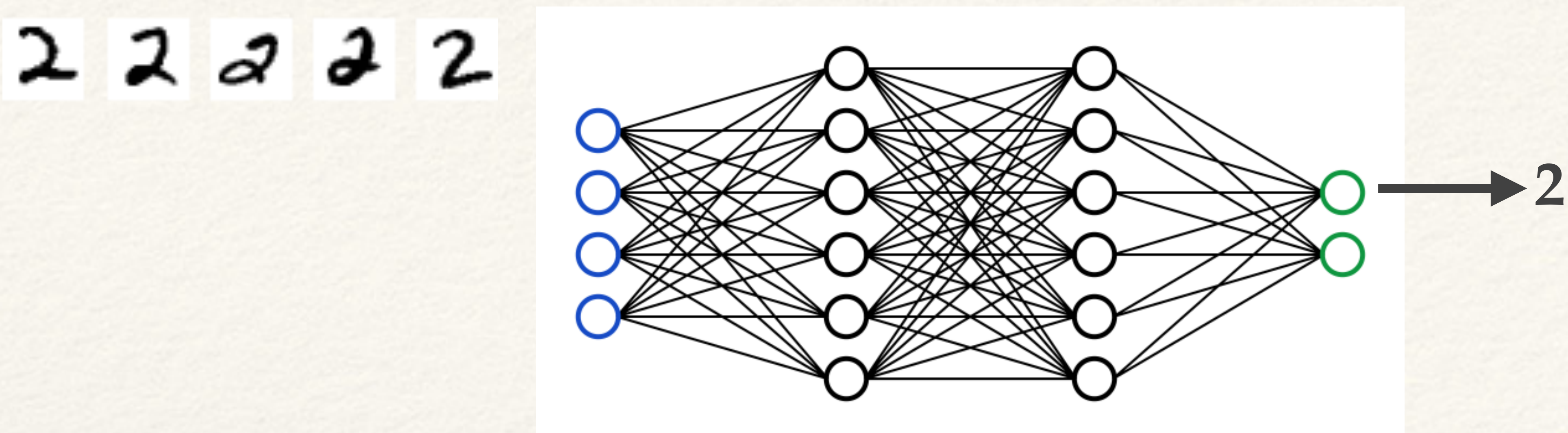
Binary data



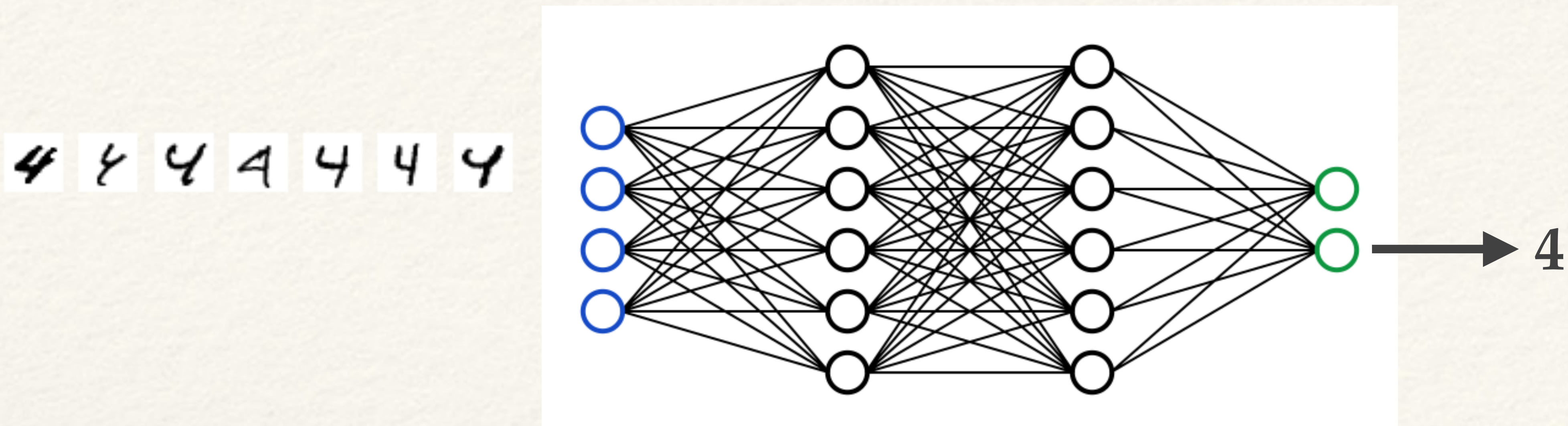
Not much is known on

how to extract the **relevant patterns hidden** within the data

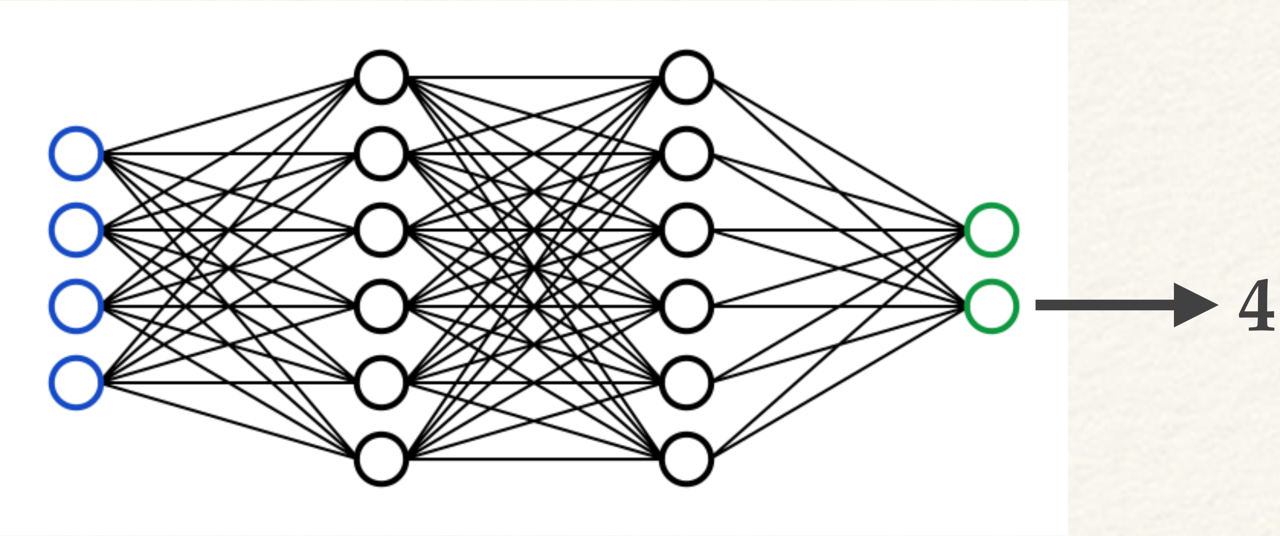
Isn't this already solved by AI?



Isn't this already solved by AI?



Isn't this already solved by AI?



Can recognise "2" from "4"

Must be identifying patterns that distinguish 2 from 4

But: We don't understand how it is doing it precisely

We don't know how to extract these patterns

What is “Information”?

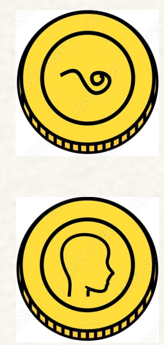
How do we quantify
how much “Information” there is in a dataset?

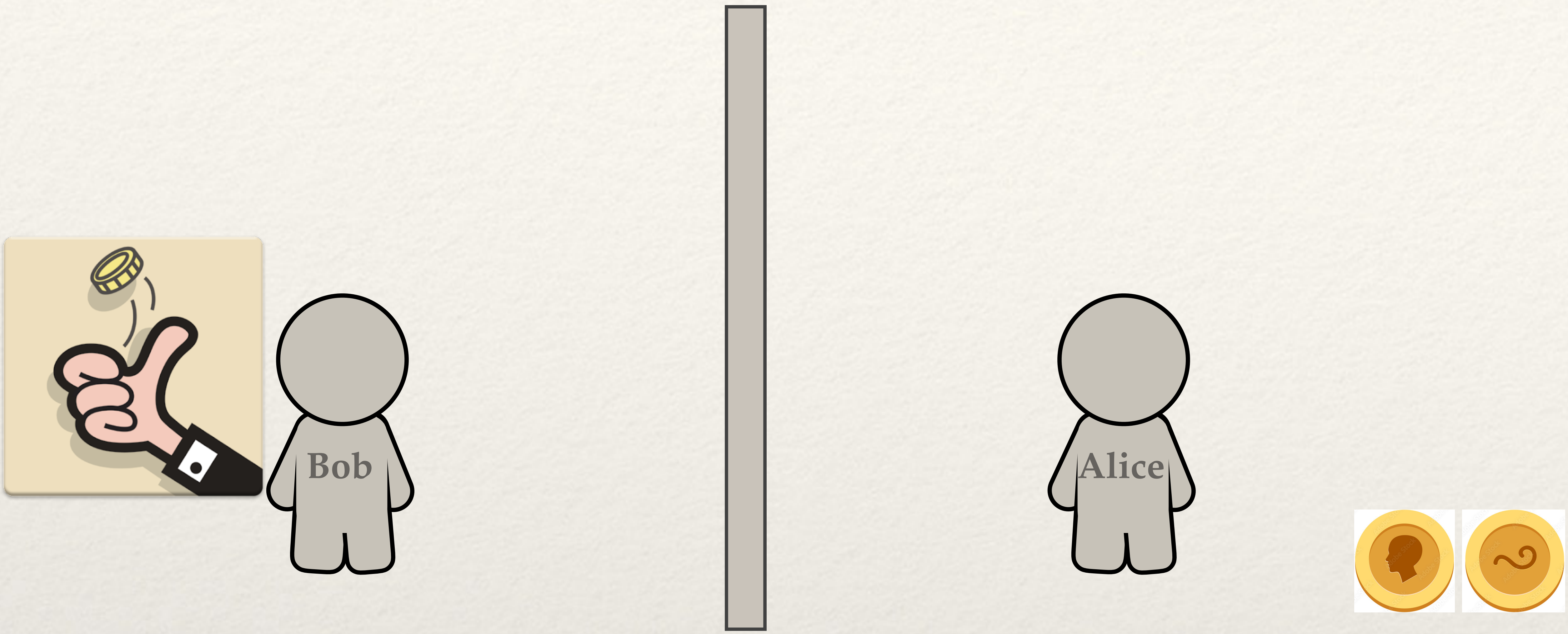
Shannon Entropy

How to quantify "Information"?

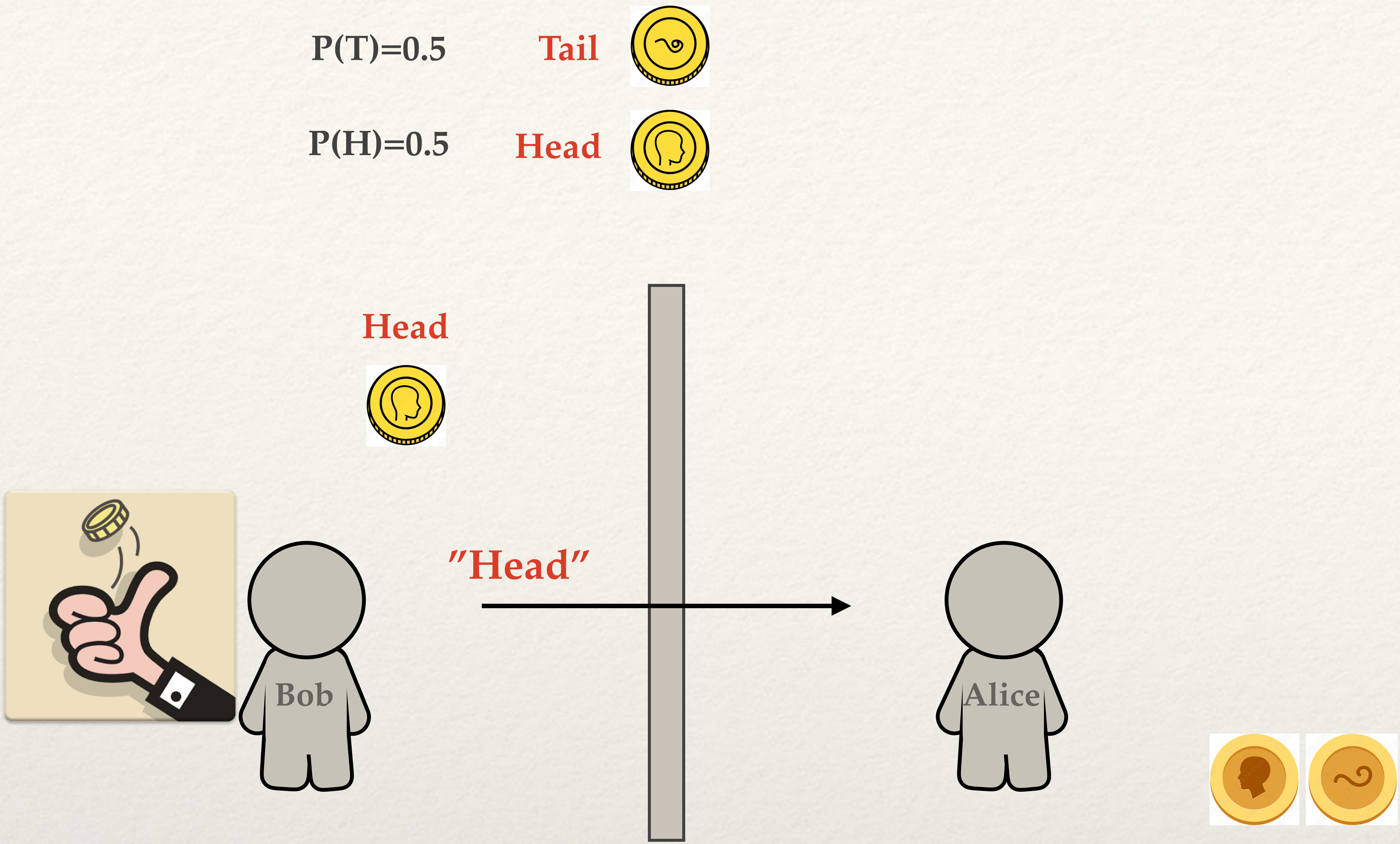
P(T)=0.5 **Tail**

P(H)=0.5 **Head**

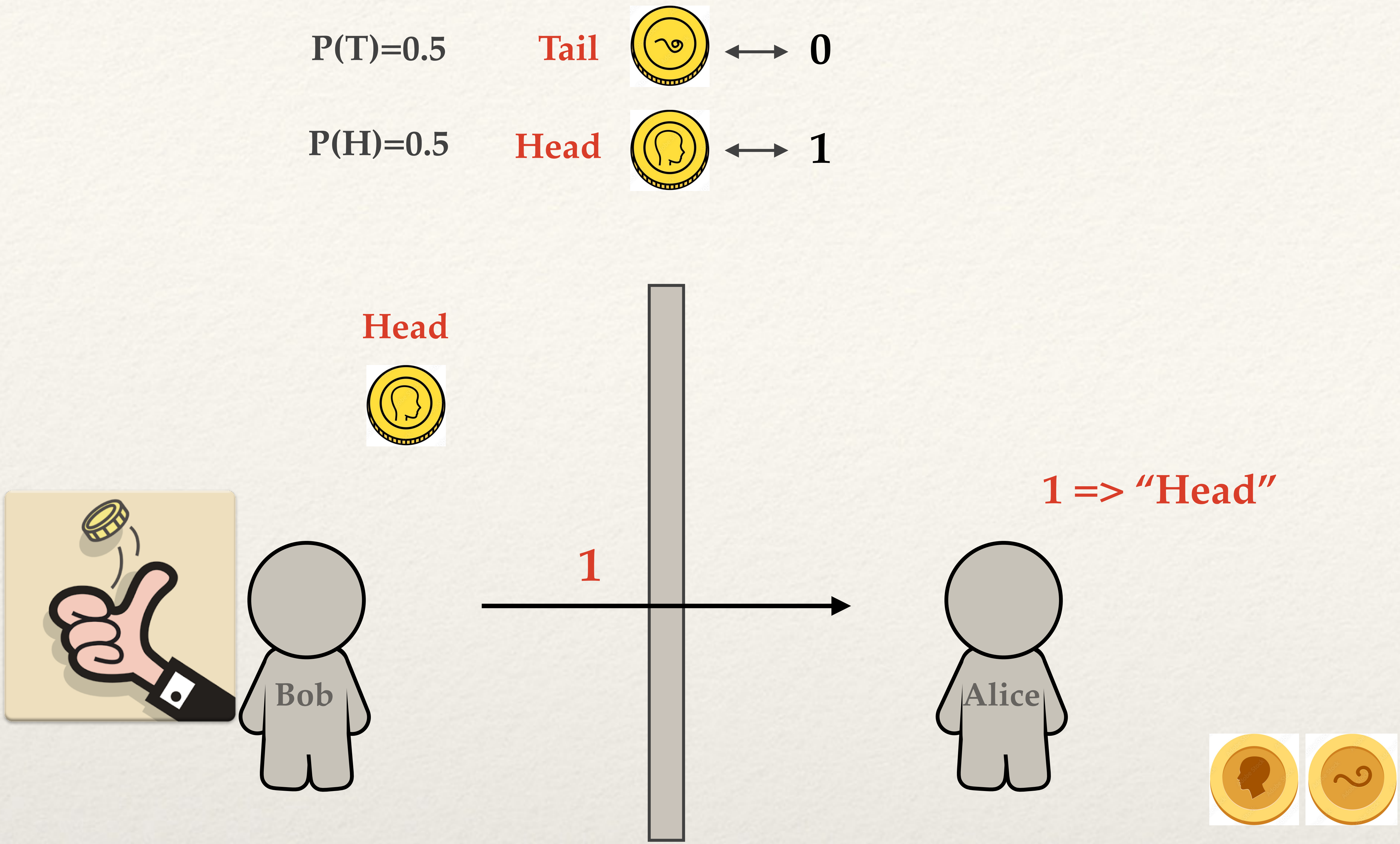




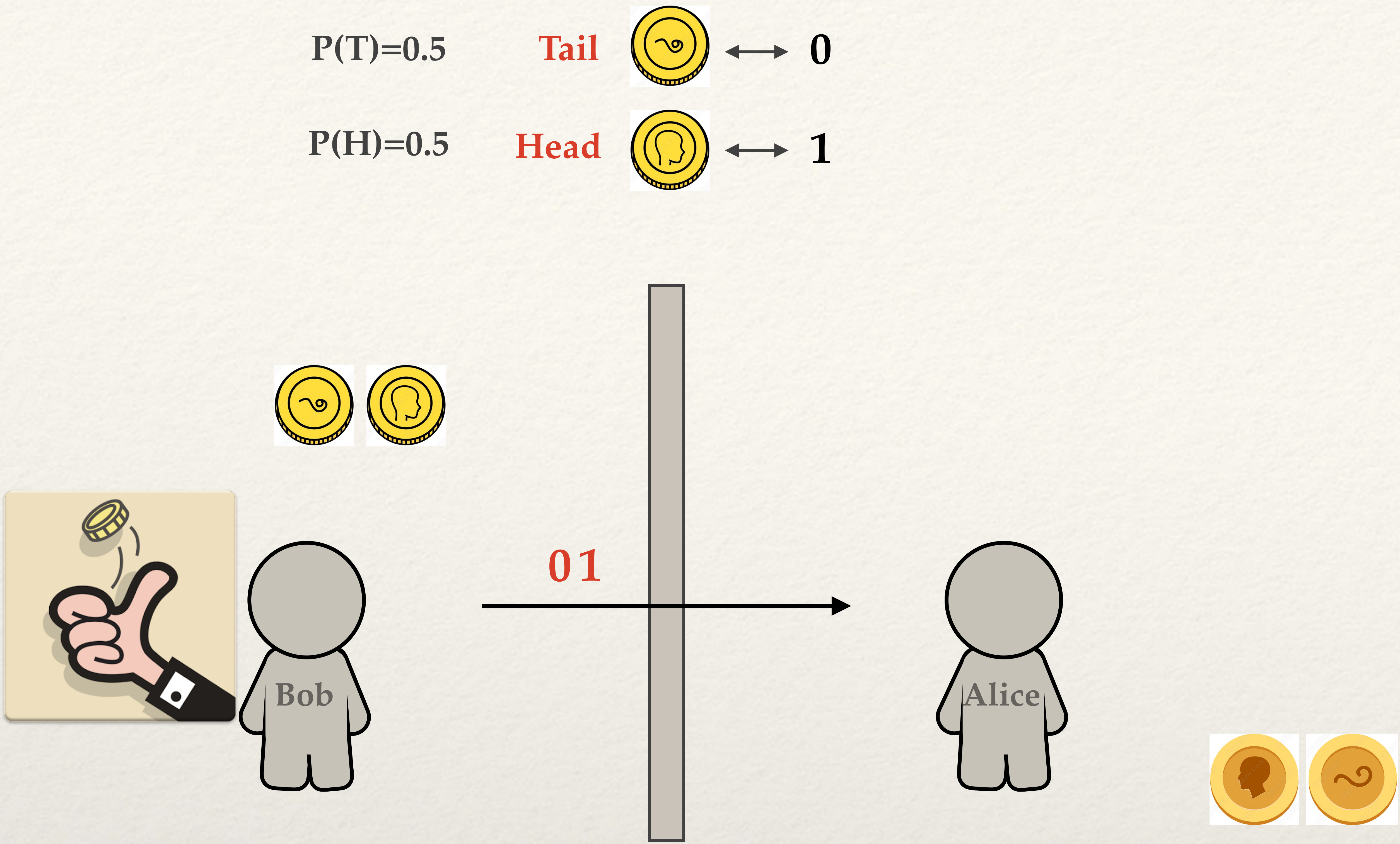
How to quantify “Information”?



What is “Information”?



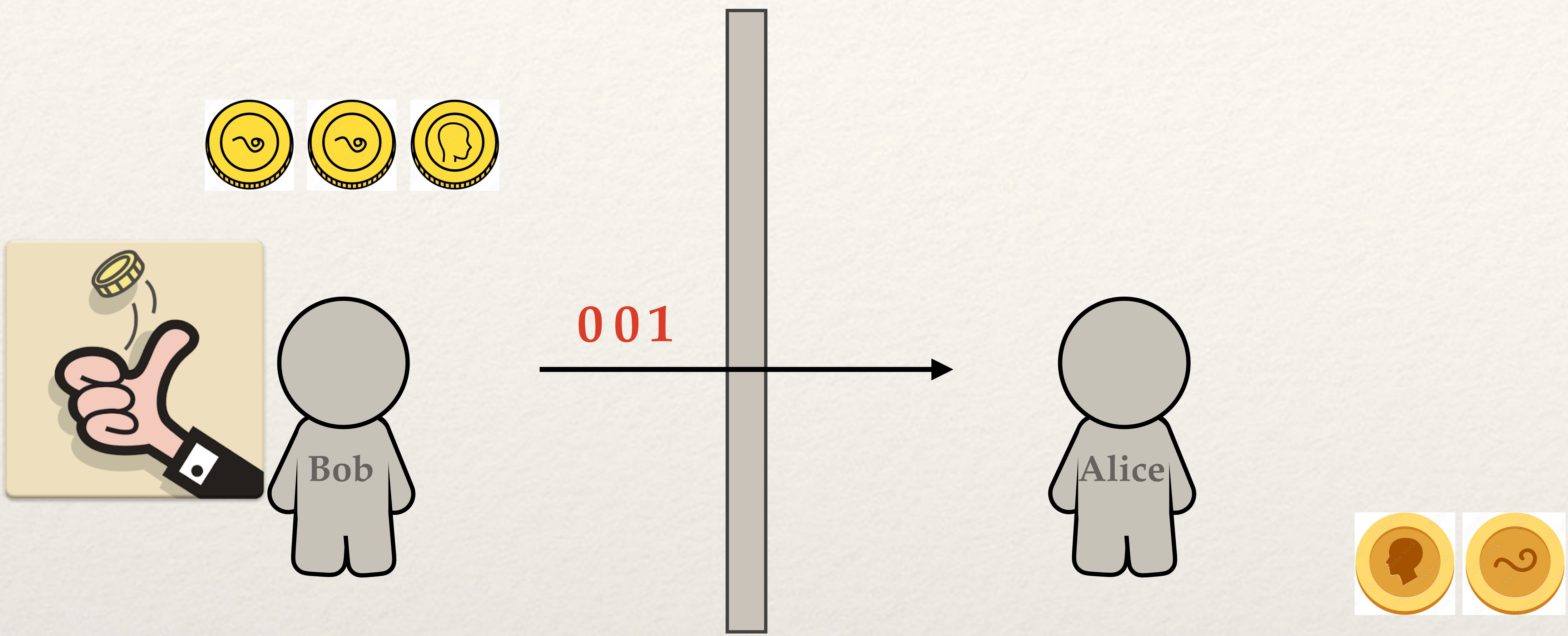
What is "Information"?



What is “Information”?

P(T)=0.5 **Tail** ⟷ **0**

P(H)=0.5 **Head** ⟷ **1**



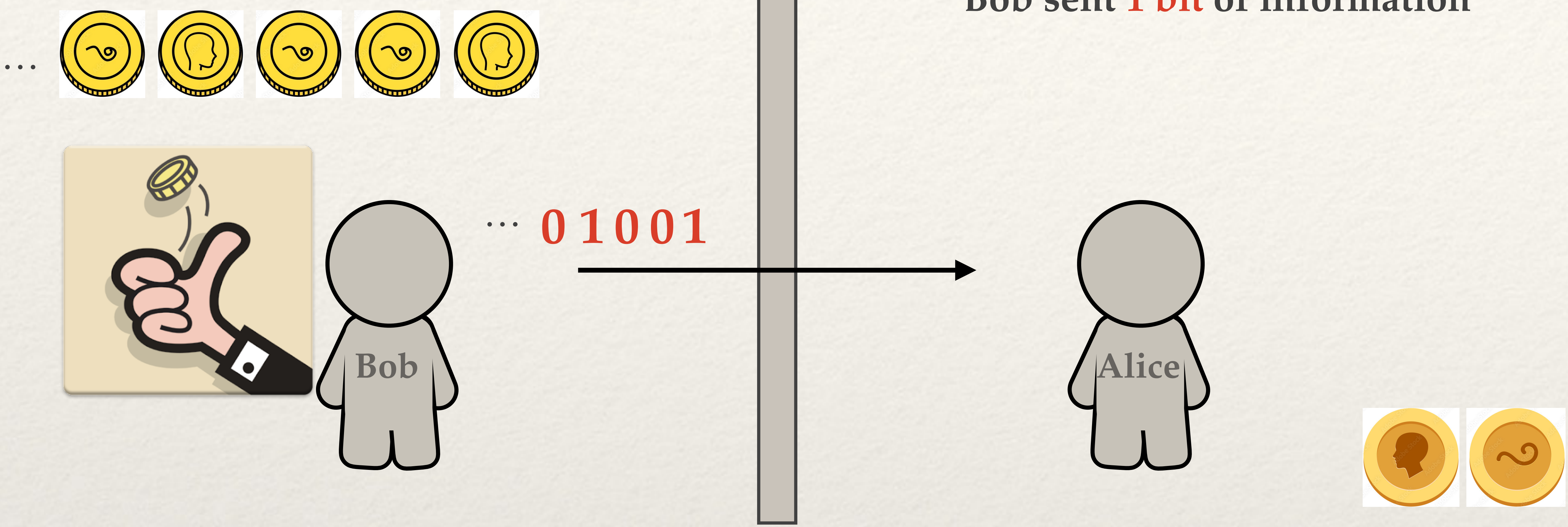
What is “Information”?

P(T)=0.5 **Tail**  ⟷ **0**

P(H)=0.5 **Head**  ⟷ **1**

For each outcome,

Bob sent 1 bit of information



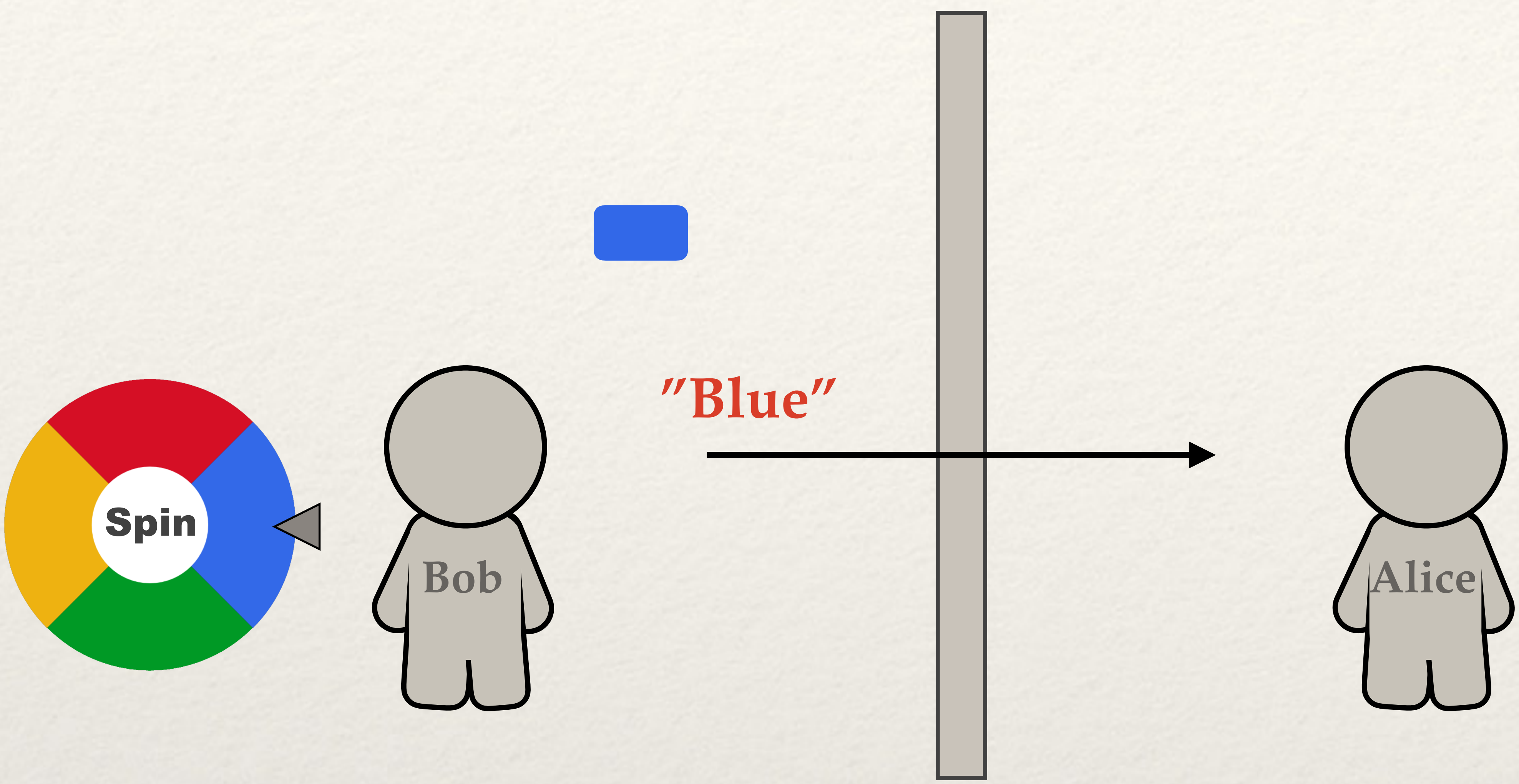
What is "Information"?

P(R)=0.25

P(B)=0.25

P(G)=0.25

P(Y)=0.25



What is "Information"?

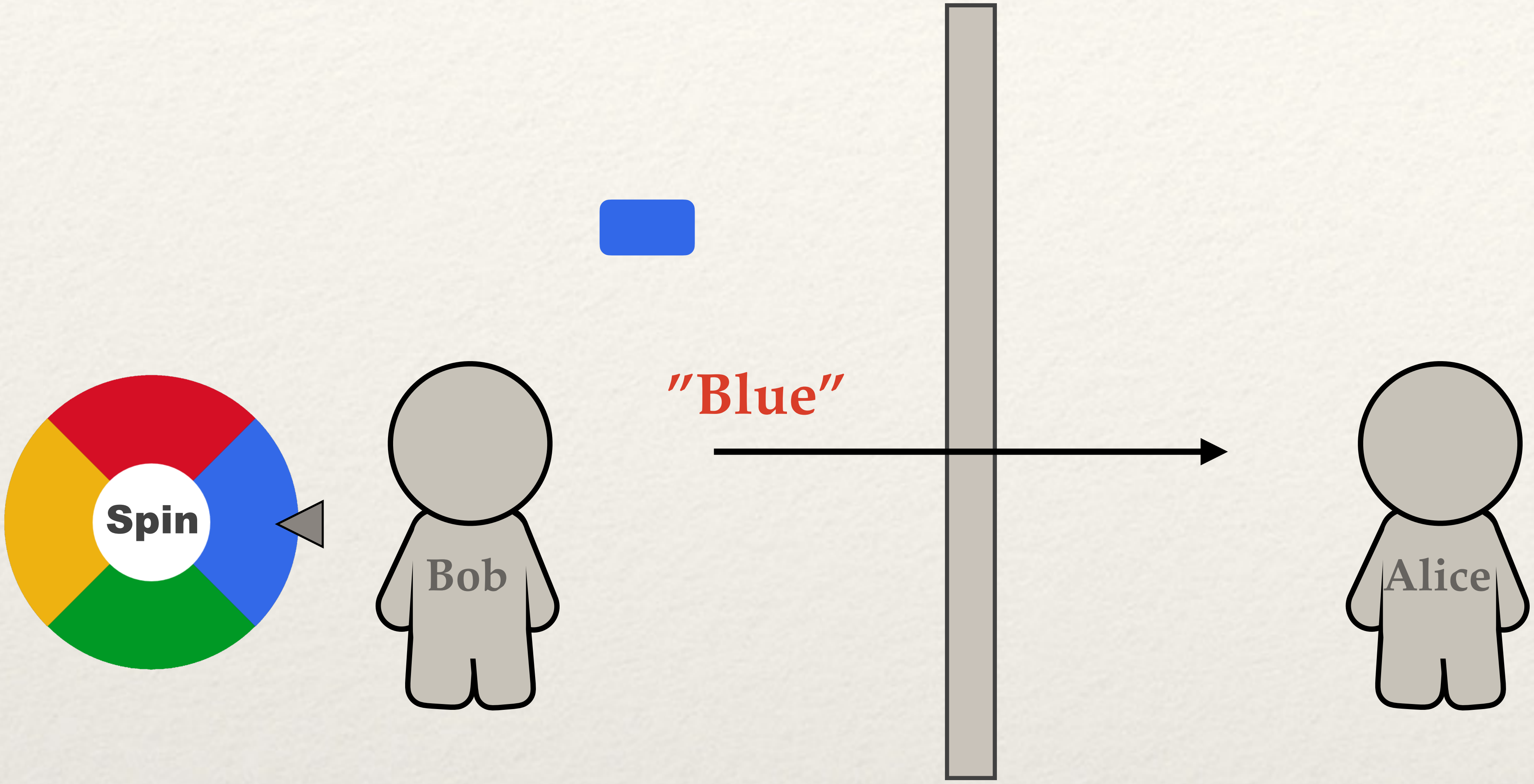
P(R)=0.25

P(B)=0.25

P(G)=0.25

P(Y)=0.25

How many bits do we need to easily encode 4 events?

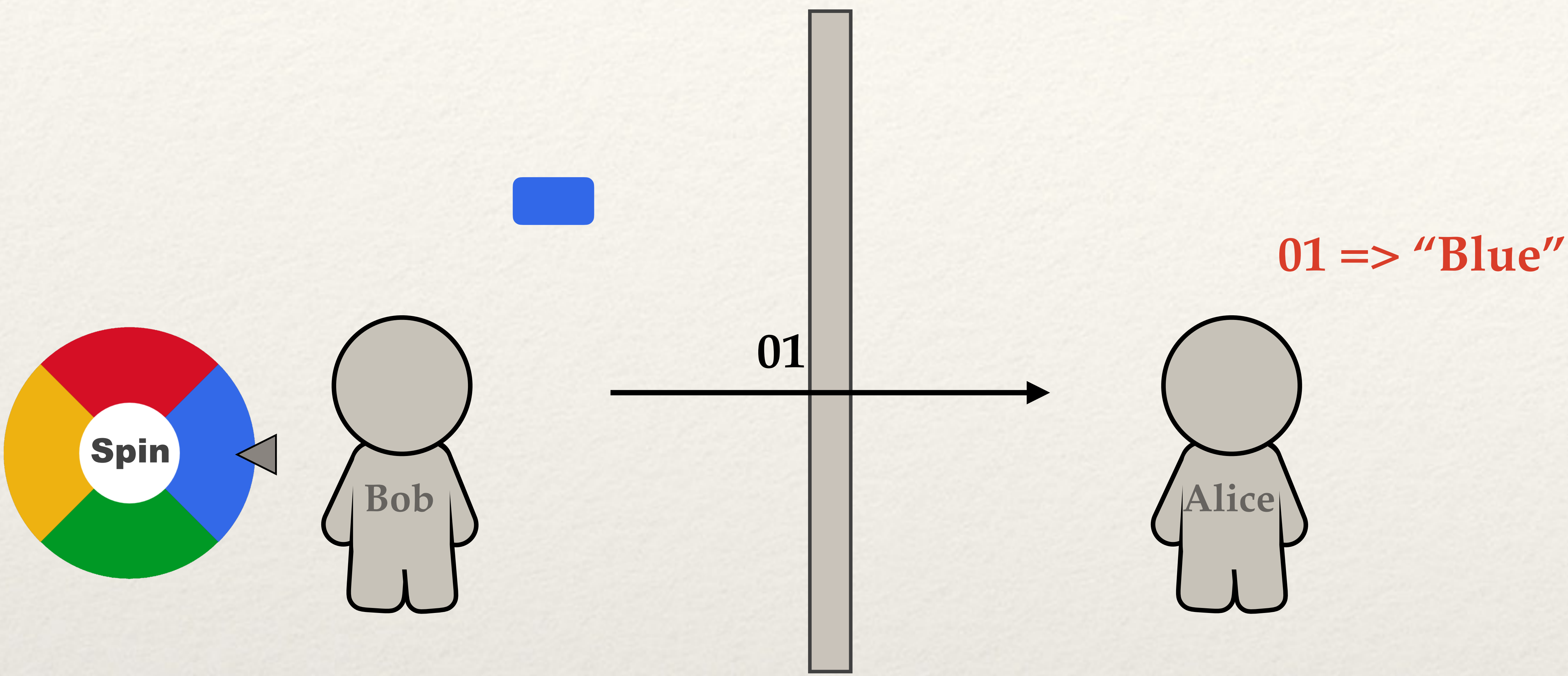


What is "Information"?

P(R)=0.25	⟷	00
P(B)=0.25	⟷	01
P(G)=0.25	⟷	10
P(Y)=0.25	⟷	11

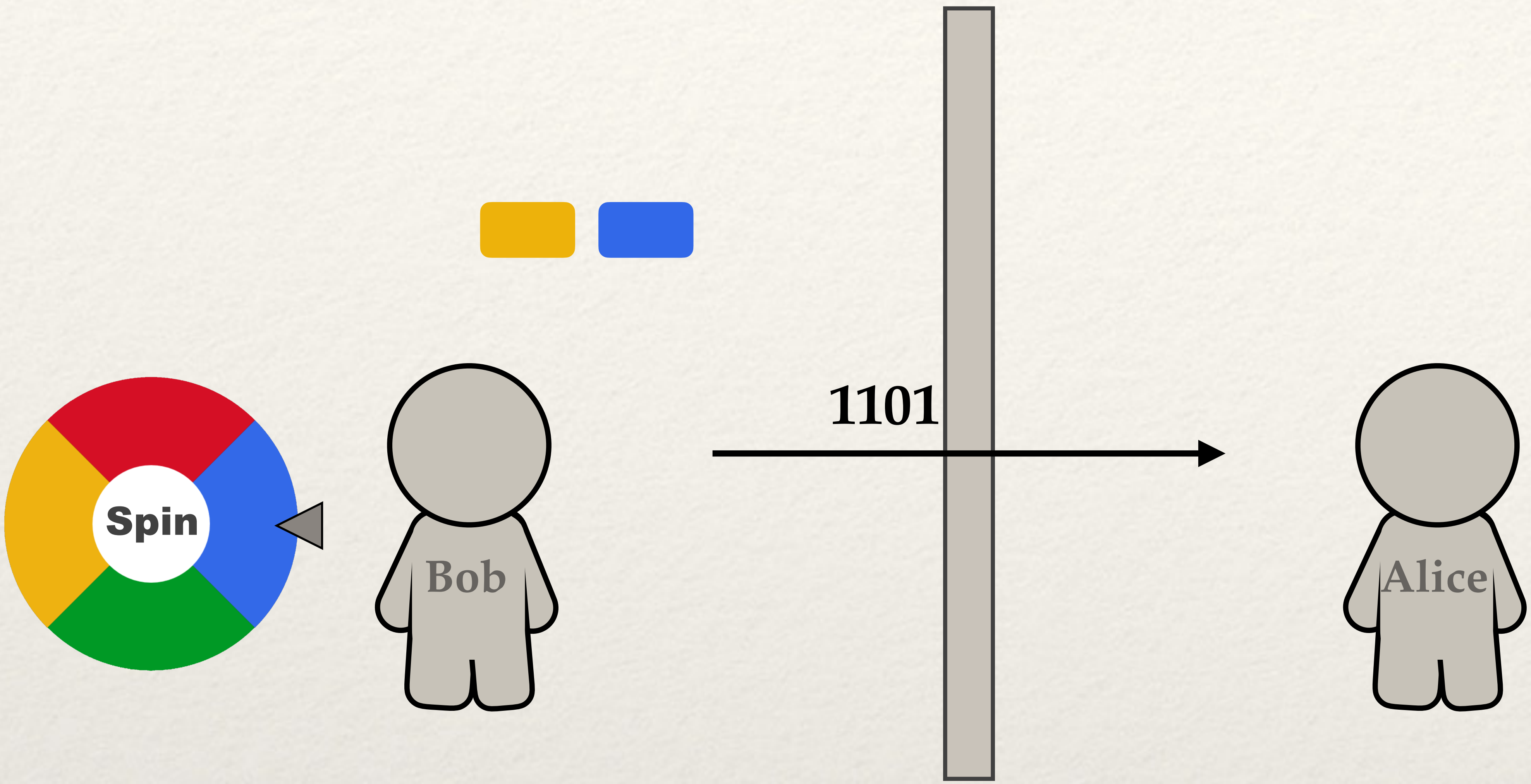
2 bits

→ $2^2 = 4$ events



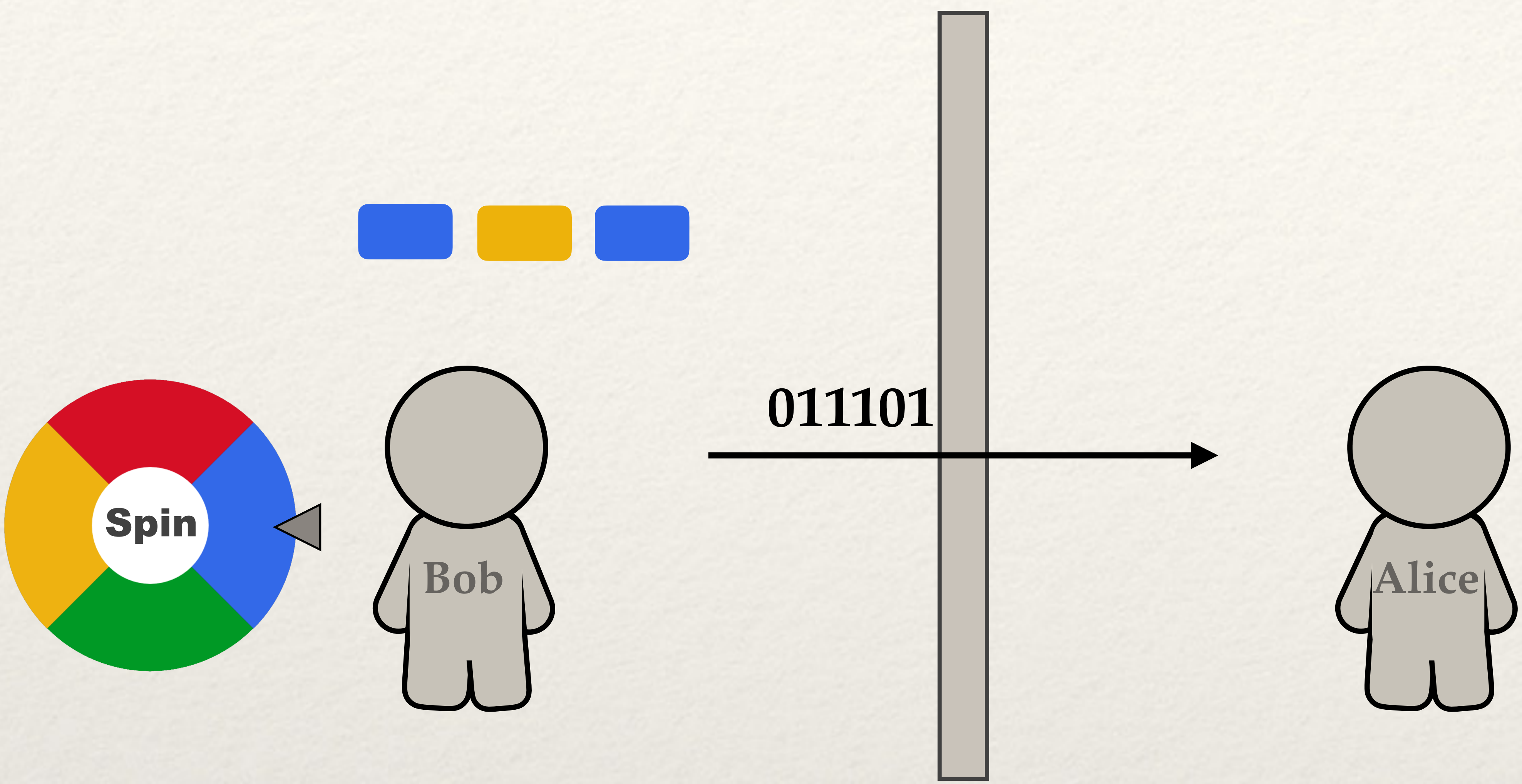
What is “Information”?

P(R)=0.25	⟷	00
P(B)=0.25	⟷	01
P(G)=0.25	⟷	10
P(Y)=0.25	⟷	11



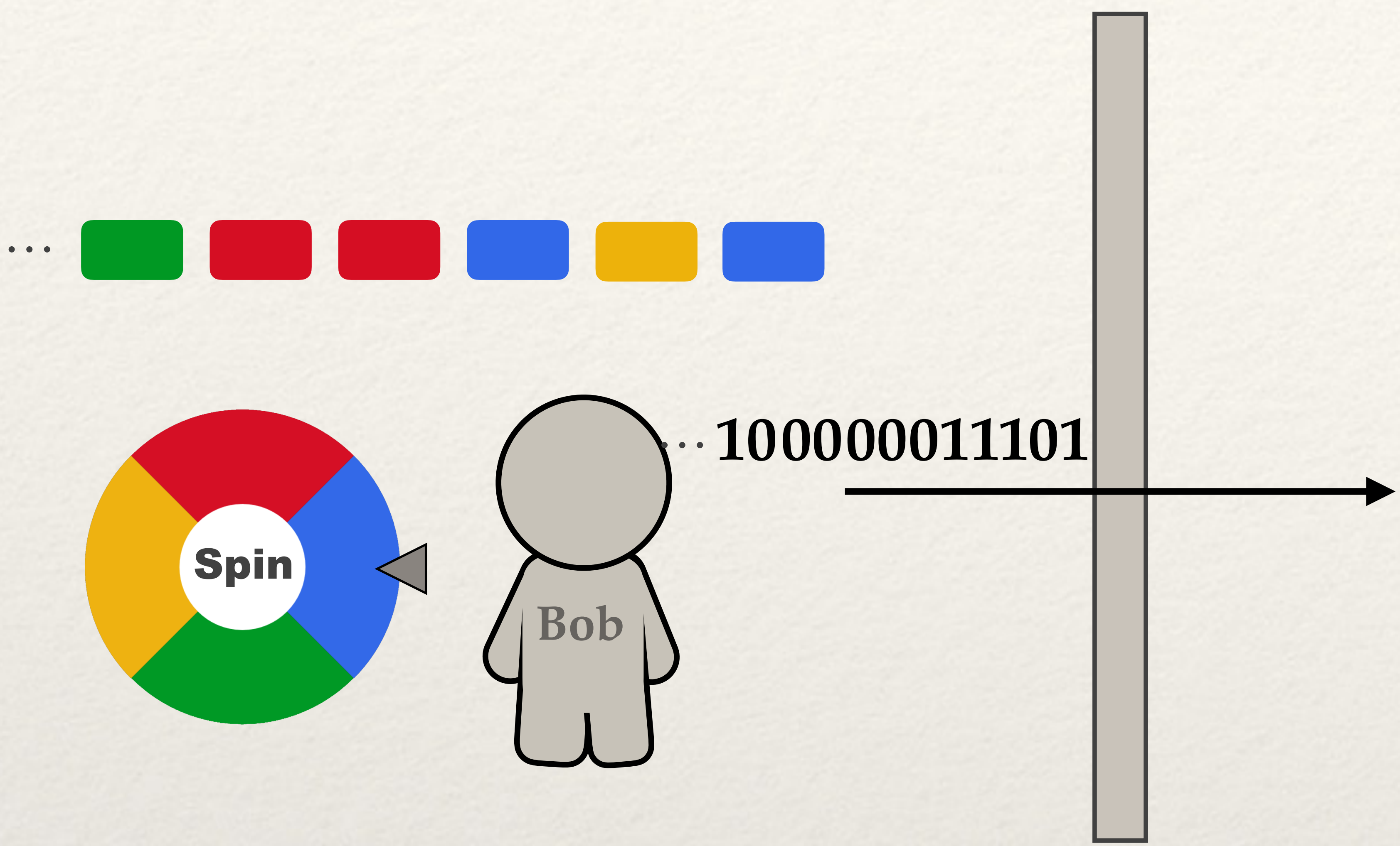
What is “Information”?

P(R)=0.25	⟷	00
P(B)=0.25	⟷	01
P(G)=0.25	⟷	10
P(Y)=0.25	⟷	11



What is “Information”?

P(R)=0.25	⟷	00
P(B)=0.25	⟷	01
P(G)=0.25	⟷	10
P(Y)=0.25	⟷	11



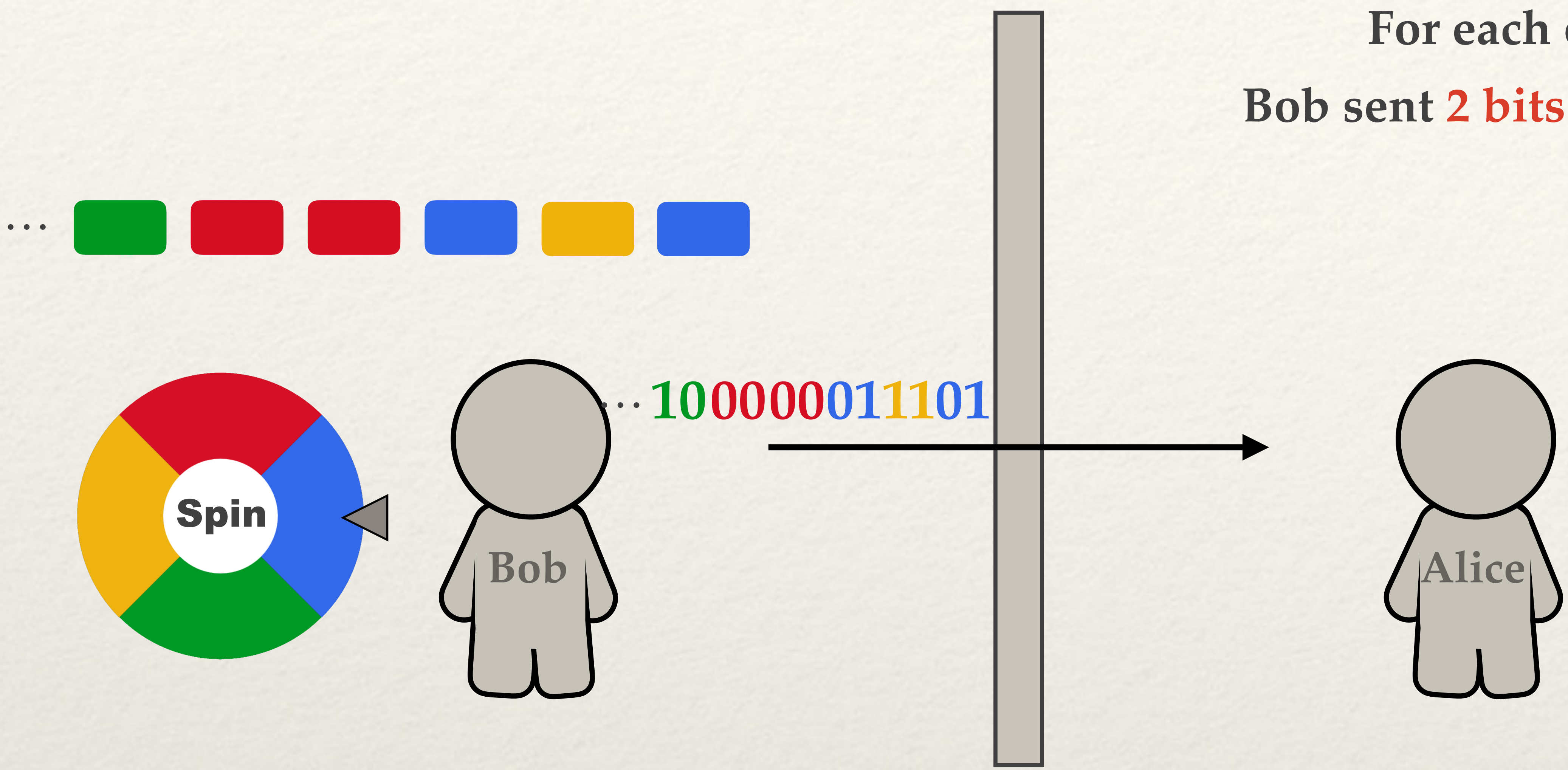
What is "Information"?

P(R)=0.25 ⟷ 00

P(B)=0.25 ⟷ 01

P(G)=0.25 ⟷ 10

P(Y)=0.25 ⟷ 11



For each outcome,

Bob sent **2 bits** of information

What is “Information”?

P(R)=0.25	Red	⟷	00
P(B)=0.25	Blue	⟷	01
P(G)=0.25	Green	⟷	10
P(Y)=0.25	Yellow	⟷	11

K=4 equiprobably events

2 bits ⟶ **Optimally encodes $2^2 = 4$ equiprobable events**

N bits ⟶ **Optimally encodes $K = 2^N$ equiprobable events**

$$2^N = K \rightarrow N \log(2) = \log(K) \rightarrow N = \log(K) \,/\, \log(2) = \log_2(K)$$

> **With K equiprobable events,** we need **at least** $N = \log_2(K)$ **bits**
>
> to encode which one has happened.

What if the events are not equiprobable?

What is "Information"?

Given a set of observations, the Information about an observation:

Minimum number of bits needed to encode that observation

P(R)=0.25	■ (red)	⟷	00
P(B)=0.25	■ (blue)	⟷	01
P(G)=0.25	■ (green)	⟷	10
P(Y)=0.25	■ (yellow)	⟷	11

K=4 equiprobably events

2 bits ⟶ **Optimally encodes $2^2 = 4$ equiprobable events**

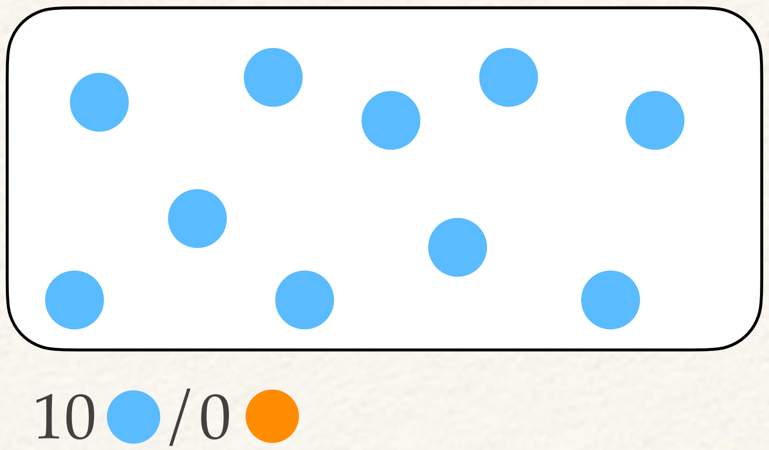
N bits ⟶ **Optimally encodes $K = 2^N$ equiprobable events**

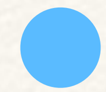
$2^N = K$ → $N \log(2) = \log(K)$ → $N = \log(K) \,/\, \log(2) = \log_2(K)$

> **With K equiprobable events,** we need **at least** $N = \log_2(K)$ **bits**
>
> to encode which one has happened.

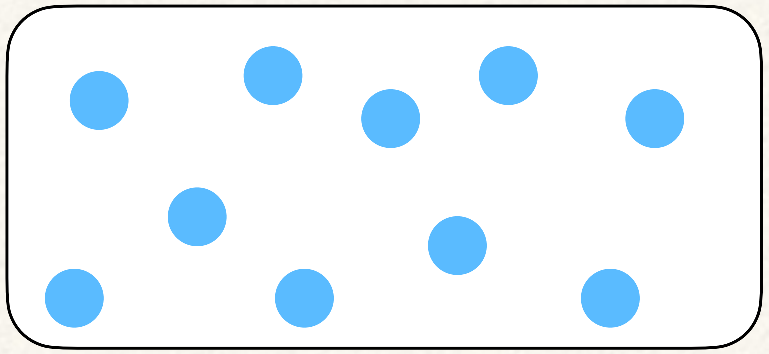
What if the events are not equiprobable?

Information ≈ Surprise!





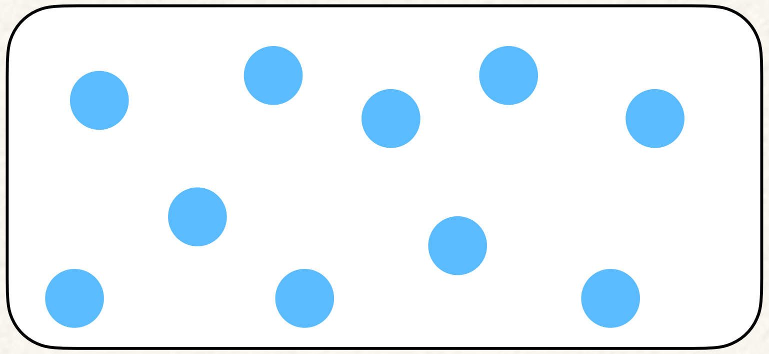
Information ≈ Surprise!



10 / 0

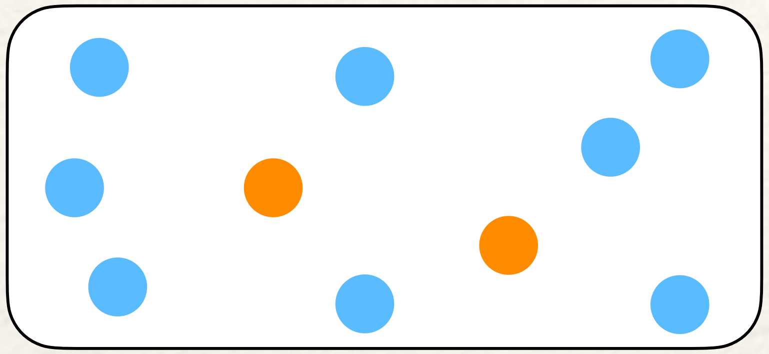
No surprise

Information ≈ Surprise!



10 / 0

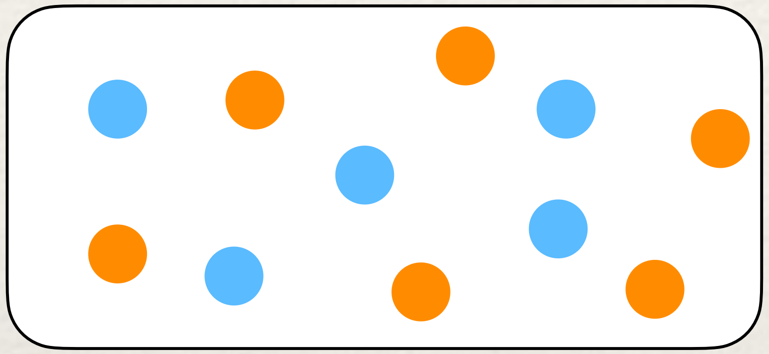
No surprise



8 / 2

Not so surprising

Quite surprising, not so expected



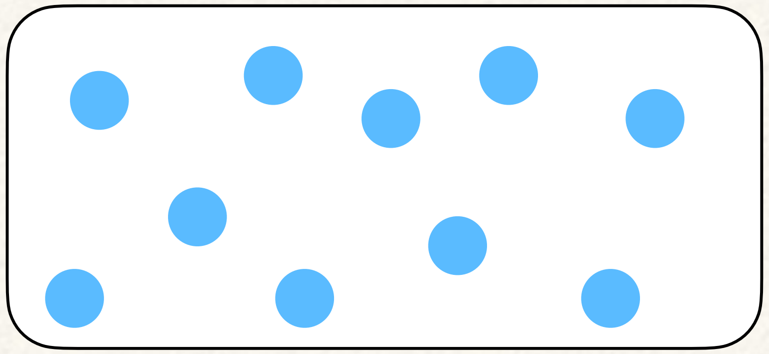
5 / 5

Surprise

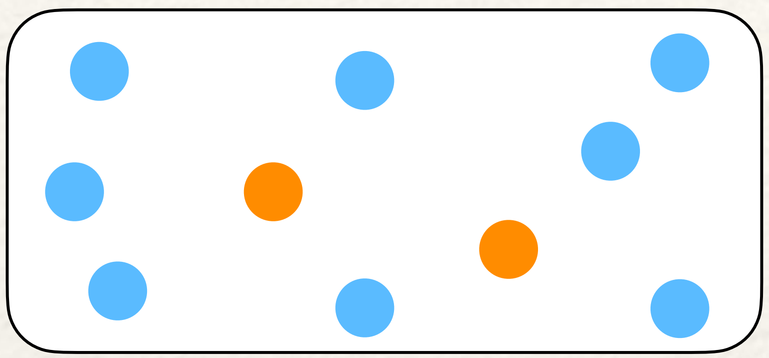
Surprise

—> No idea about what will come out!

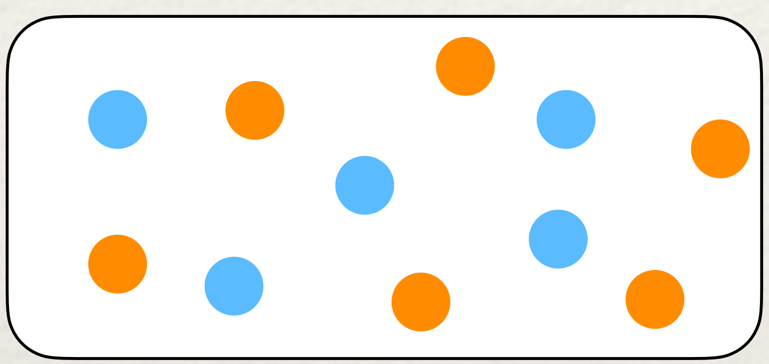
Information ≈ Surprise!



10 ●/0 ●



8 ●/2 ●



5 ●/5 ●

The lower the probability

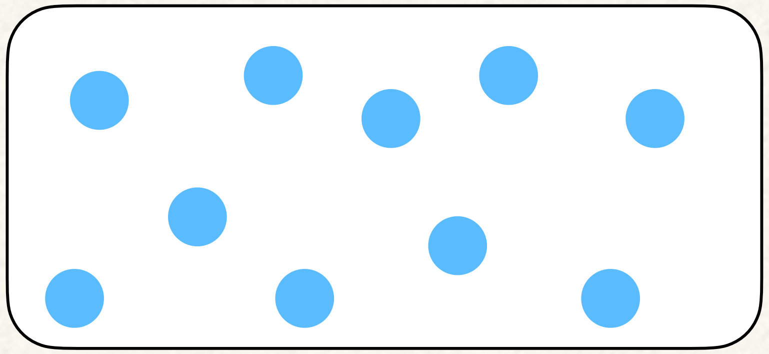
The more surprise

$$I(s) = -\log_2 [p(s)]$$

The amount of information obtained by observation an event depends on how surprised I am about that observation.

—> No idea about what will come out!

Information ≈ Surprise!

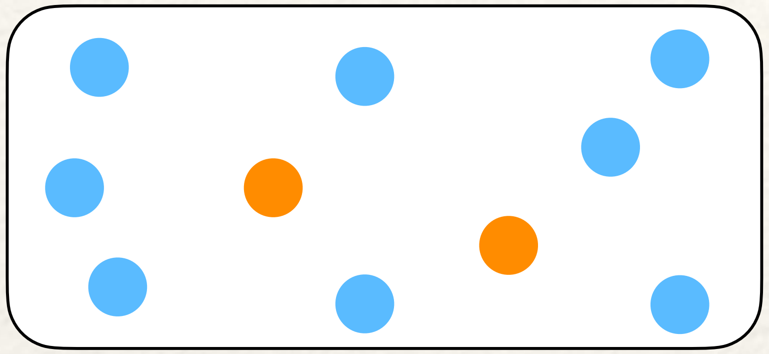


10 🔵 / 0 🟠

🔵 No surprise

No information about the system

I(🔵) = 0



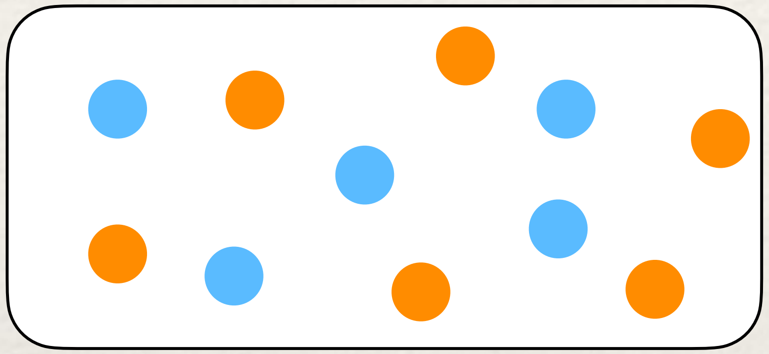
8 🔵 / 2 🟠

🔵 Not so surprising

I(🔵) = $-\log_2(0.8)$ =0.32 bits

🟠 Quite surprising, not so expected

I(🟠) = $-\log_2(0.2)$ =2.32 bits



5 🔵 / 5 🟠

🔵 Surprise I(🔵) = -log2(0.5) = 1 bits

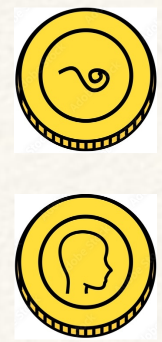
🟠 Surprise

—> No idea about what will come out!

Information ≈ Surprise!

P(T)=0.2 **Tail** ↔ **0**

P(H)=0.8 **Head** ↔ **1**

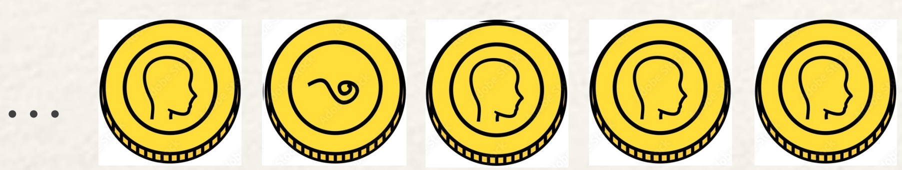


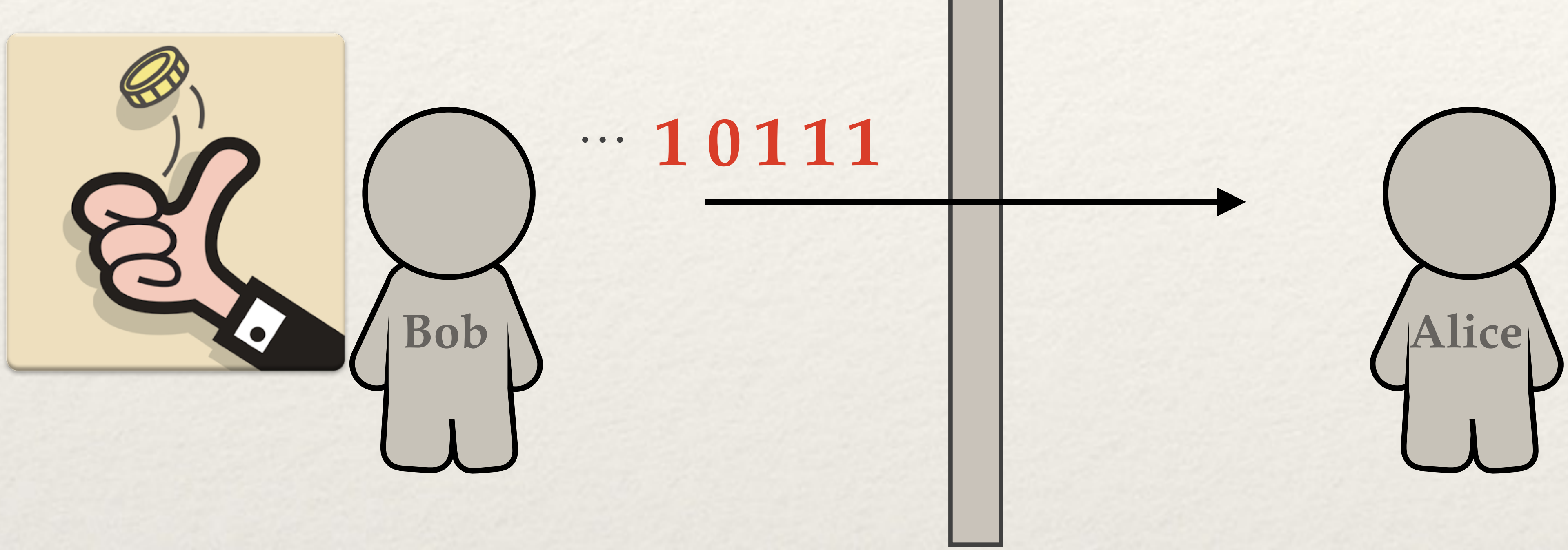
For each outcome,

Bob sent 1 bit

But only needs:

0.2 x 2.32. bits + 0.8 x 0.32 bits = **0.72 bits**





Modeling Data with Statistical Physics

How do we

extract important information?

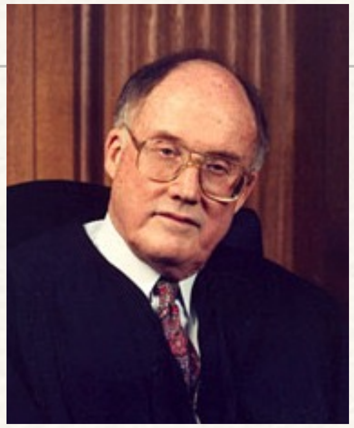
How do we model data?

Ex.

Statistical Mechanics of the US Supreme Court

Edward D. Lee, Chase P. Broedersz & William Bialek

Journal of Statistical Physics **160**, 275–301(2015) | Cite this article



2nd Rehnquist Court (1994-2005)

9 justices, 895 votes

Conservative (+1) or **Liberal** (0)

result of a vote ↓

9 justices

SB	1	0	1	0	0	0	0	1	
WR	0	1	0	0	1	0	0	1	
JS	0	1	0	1	1	1	0	1	
SO	0	0	1	0	0	0	1	1	
AS	0	1	0	0	1	1	0	1	…
AK	1	0	0	0	0	1	0	1	
DS	1	0	0	1	1	1	0	1	
CT	0	0	1	0	0	1	0	1	
RG	0	0	1	1	0	0	1	1	

895 votes

How do we model data?

Ex.

Statistical Mechanics of the US Supreme Court

Edward D. Lee, Chase P. Broedersz & William Bialek

Journal of Statistical Physics **160**, 275–301(2015) | Cite this article

2nd Rehnquist Court
(1994-2005)

9 justices, 895 votes

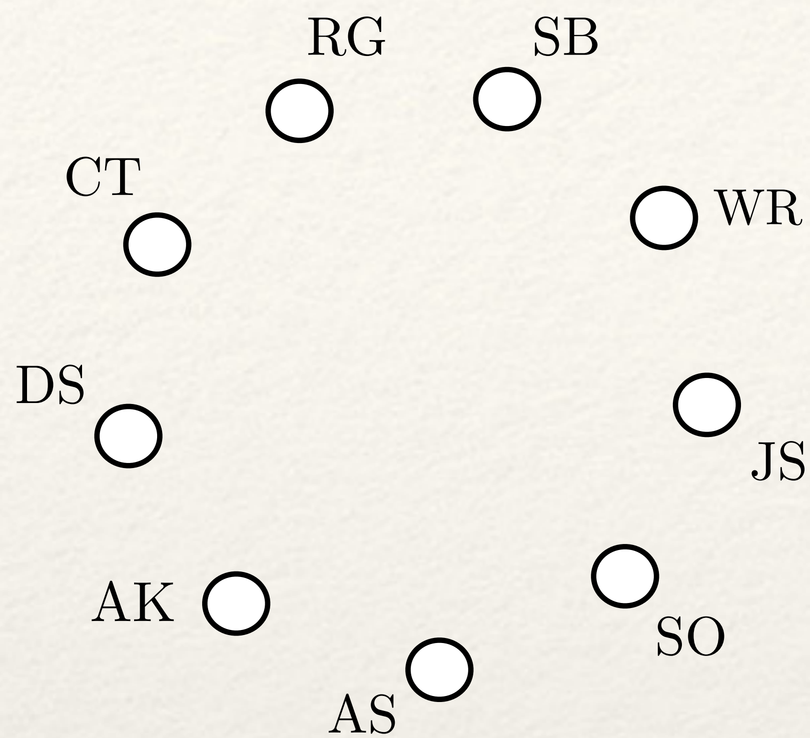
Conservative (+1) or **Liberal** (0)

result of a vote ↓

9 justices

SB	1	0	1	0	0	0	0	1	
WR	0	1	0	0	1	0	0	1	
JS	0	1	0	1	1	1	0	1	
SO	0	0	1	0	0	0	1	1	
AS	0	1	0	0	1	1	0	1	…
AK	1	0	0	0	0	1	0	1	
DS	1	0	0	1	1	1	0	1	
CT	0	0	1	0	0	1	0	1	
RG	0	0	1	1	0	0	1	1	

895 votes



maybe: judges are making their own decision

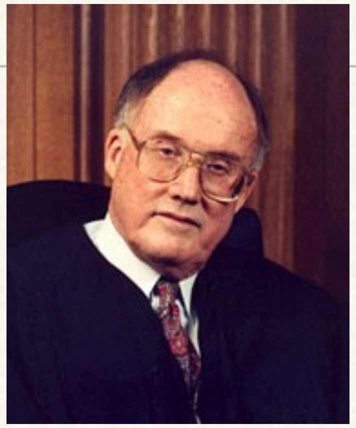
How do we model data?

Ex.

Statistical Mechanics of the US Supreme Court

Edward D. Lee, Chase P. Broedersz & William Bialek

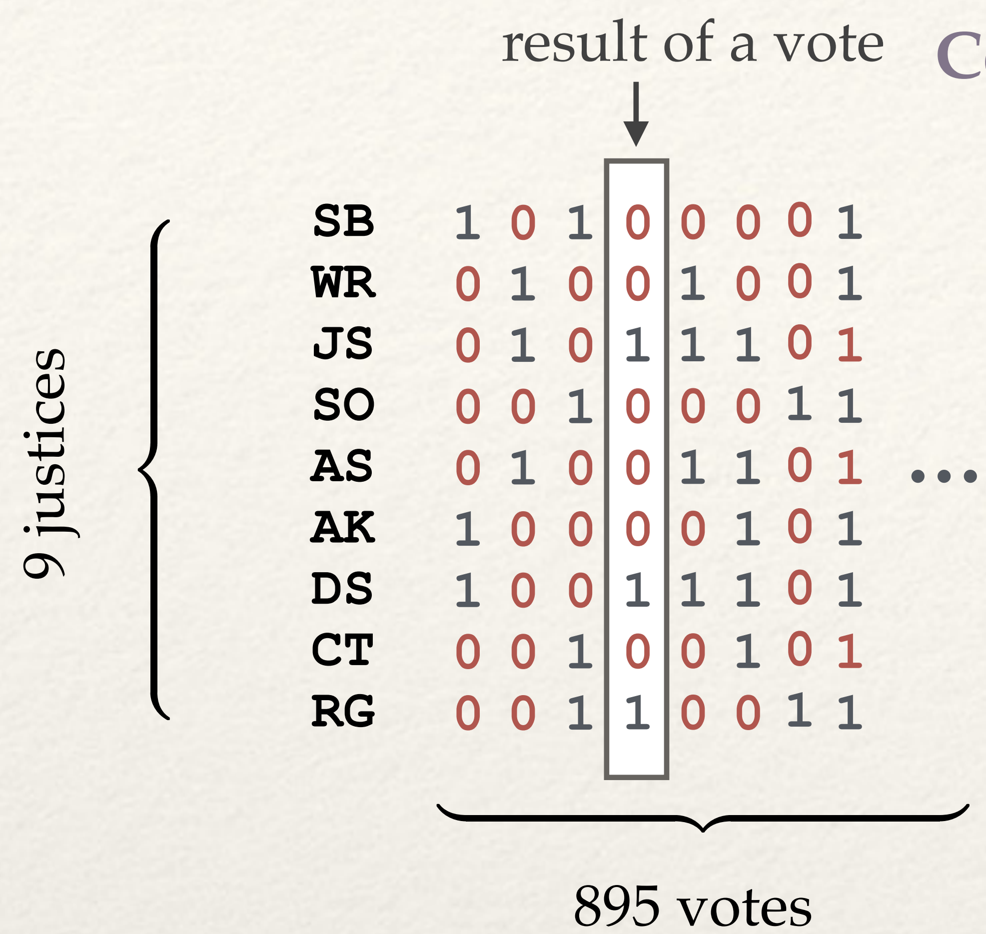
Journal of Statistical Physics **160**, 275–301(2015) | Cite this article

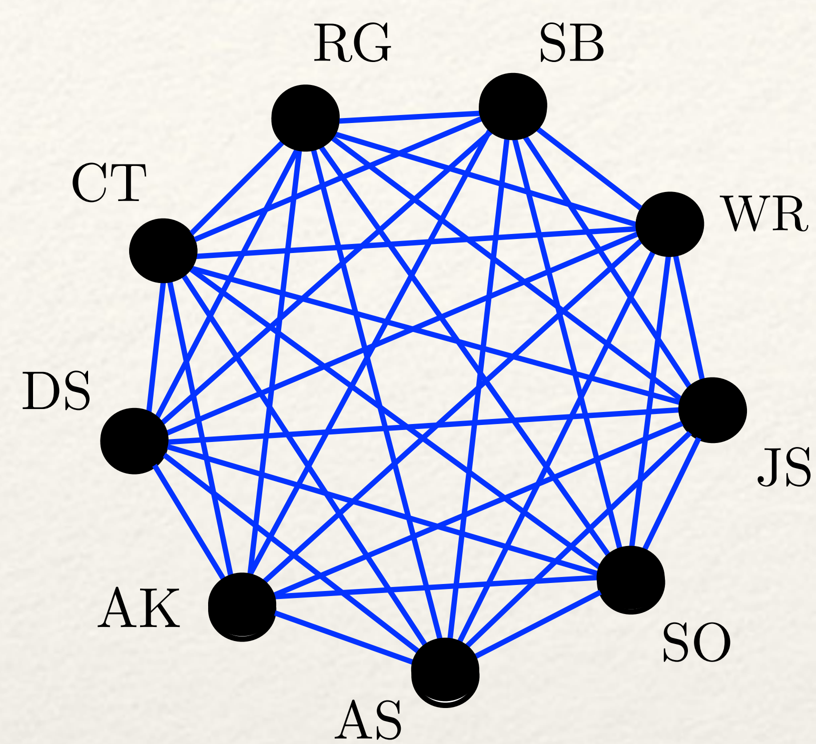


2nd Rehnquist Court
(1994-2005)

9 justices, 895 votes

Conservative (+1) or **Liberal** (0)



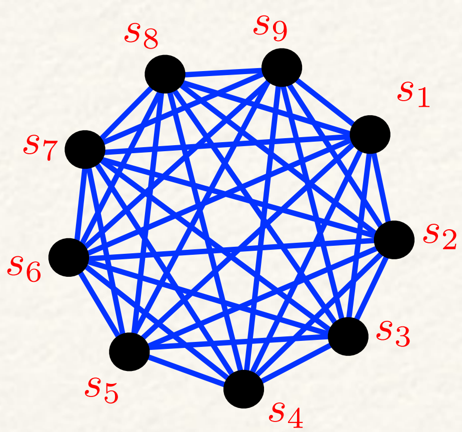


maybe: judges they discuss and decided with each others

Statistical inference for binary data

2nd Rehnquist Court
(1994-2005)

US Supreme Court
9 justices, 895 votes
Conservative (+1) or **Liberal** (-1)



Ising model

$$P(\boldsymbol{s}\,|\,\boldsymbol{g}) = \frac{1}{Z(\boldsymbol{g})} \exp\Big(\sum_i h_i\, s_i + \sum_{pair(i,j)} J_{ij}\, s_i s_j \Big)$$

$$\boldsymbol{s} = (s_1, \cdots, s_9)$$

Assumptions:

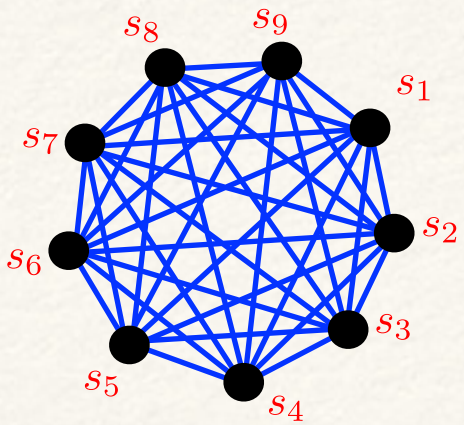
- Vote of each justice is a binary **random variable** $s_i \in \{+1, -1\}$
- Each vote is **independently** sampled from an underlying **probability distribution**: the Ising model

[Lee, Broedersz, Bialek] *Statistical Mechanics of the US Supreme Court*

Statistical inference for binary data

2nd Rehnquist Court
(1994-2005)

US Supreme Court
9 justices, 895 votes
Conservative (+1) or **Liberal** (-1)



Ising model

$$P(\boldsymbol{s} \,|\, \boldsymbol{g}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_i h_i \, s_i + \sum_{pair(i,j)} J_{ij} \, s_i s_j \right)$$

local fields
models an **External drive** to vote
conservative or liberal

couplings
models a **Tendency of i and j** to vote
similarly or oppositely

Parameters $\boldsymbol{g} = (h_1, \cdots, h_9, J_{12}, \cdots, J_{89})$

Assumptions:

- Vote of each justice is a binary **random variable**
- Each vote is **independently** sampled from an underlying **probability distribution**: the Ising model

[Lee, Broedersz, Bialek] *Statistical Mechanics of the US Supreme Court*

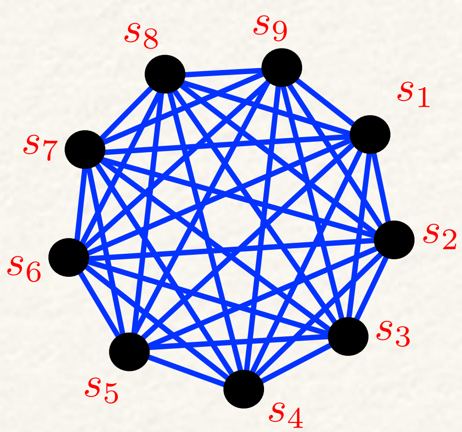
Statistical inference for binary data

2nd Rehnquist Court
(1994-2005)

US Supreme Court

9 justices, 895 votes

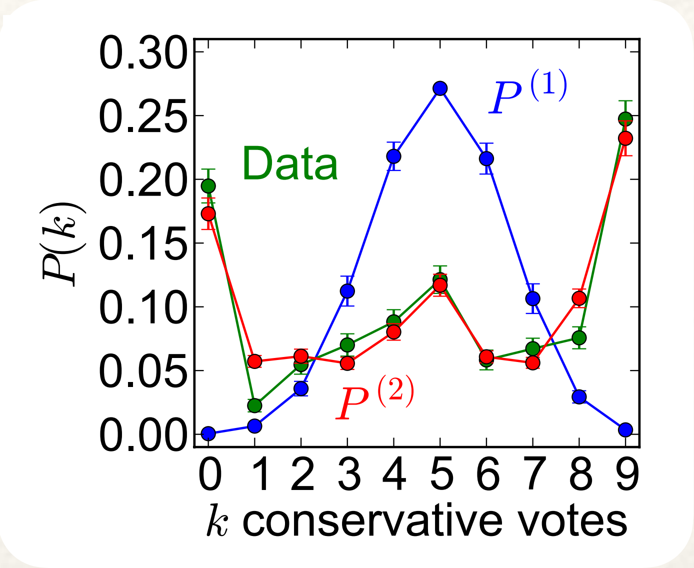
Conservative (+1) or **Liberal** (-1)



Ising model

$$P(\mathbf{s} \mid \mathbf{g}) = \frac{1}{Z(\mathbf{g})} \exp\Big(\sum_i h_i\, s_i + \sum_{pair(i,j)} J_{ij}\, s_i s_j \Big)$$

Fit the parameters:



Can it predict other types of patterns in the data?

Ex. High order patterns

Finds that: judges are NOT making decisions INDEPENDENTLY from each other!

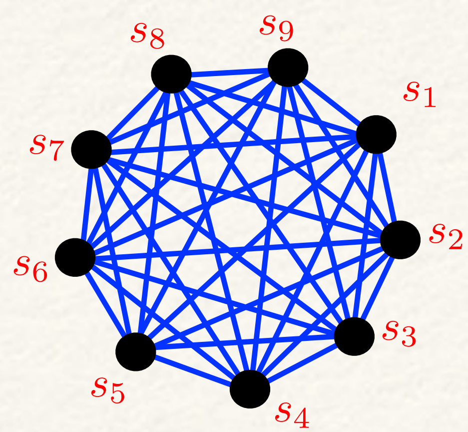
Statistical inference for binary data

2nd Rehnquist Court
(1994-2005)

US Supreme Court

9 justices, 895 votes

Conservative (+1) or **Liberal** (-1)



Ising model

$$P(\boldsymbol{s} \mid \boldsymbol{g}) = \frac{1}{Z(\boldsymbol{g})} \exp\left(\sum_i h_i\, s_i + \sum_{pair(i,j)} J_{ij}\, s_i s_j \right)$$

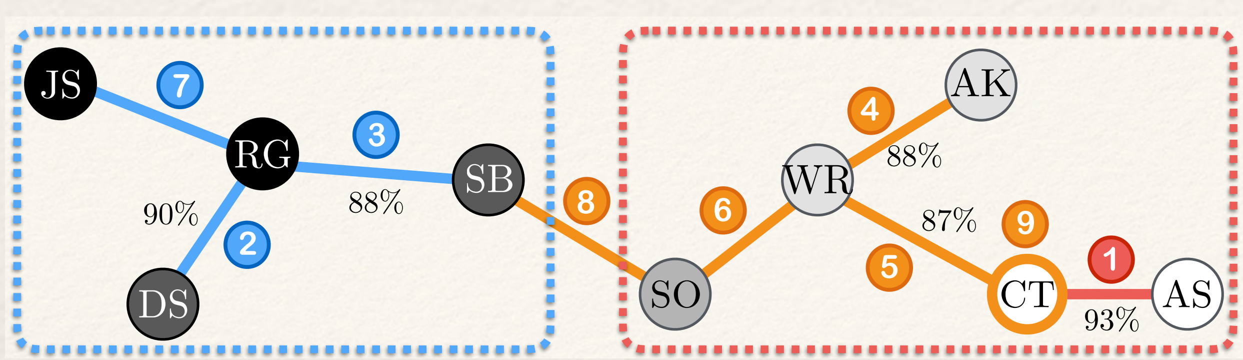
Fit the parameters:

Very complex models: lots of parameters

Penalise for too many parameters

More than 80% of information

More liberal **More conservative**

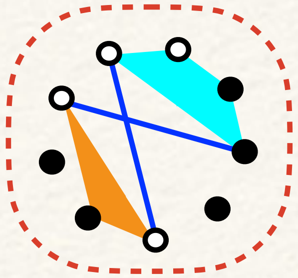


Conclusion and future

Searching for Collective Behavior in a Large Network of Sensory Neurons
Tkačik, Marre, Amodei, Schneidman, Bialek, Berry
PLoS Comp Bio 2014

> **Neurons are not firing independently**

> **Insufficiency of pairwise** interactions to model **large populations of neurons**



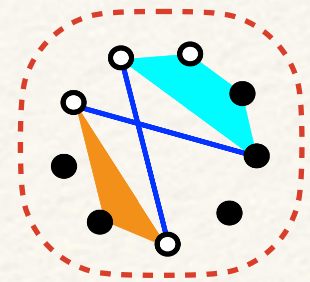
?

> **We can record 1000s of neurons**

Conclusion and future

> **Neurons are not firing independently**

> **Insufficiency of pairwise** interactions to model **large populations of neurons**



?

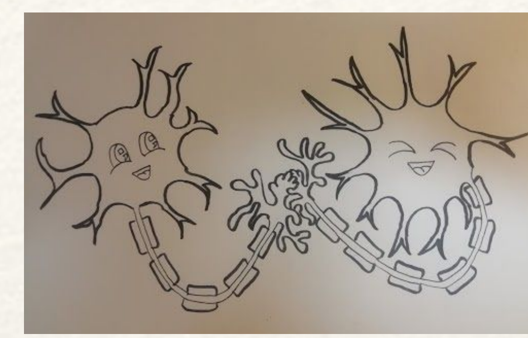
Searching for Collective Behavior in a Large Network of Sensory Neurons

Tkačik, Marre, Amodei, Schneidman, Bialek, Berry

PLoS Comp Bio 2014

> **We can record 1000s of neurons**

Detecting communities of neurons



Renormalisation