Entropy in Ergodic Theory

Tomasz Downarowicz

Institute of Mathematics and Computer Science
Wroclaw University of Technology
Poland

Based on

ENTROPY IN DYNAMICAL SYSTEMS
New Mathematical Monographs: 18
Cambridge University Press 2011

PART I: Entropy in Ergodic Theory

What is information?

What is information?

What is information?

What is information?

What is information?

What is information?

How much information was that?

What is information?

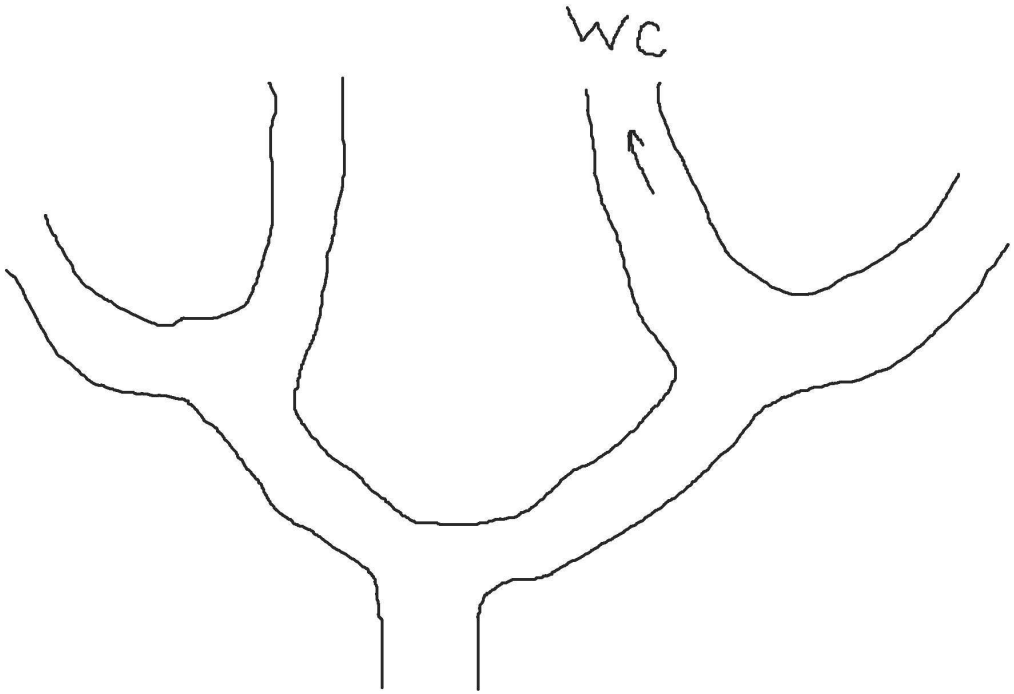
How much information was that?

one out of two choices = ONE BIT

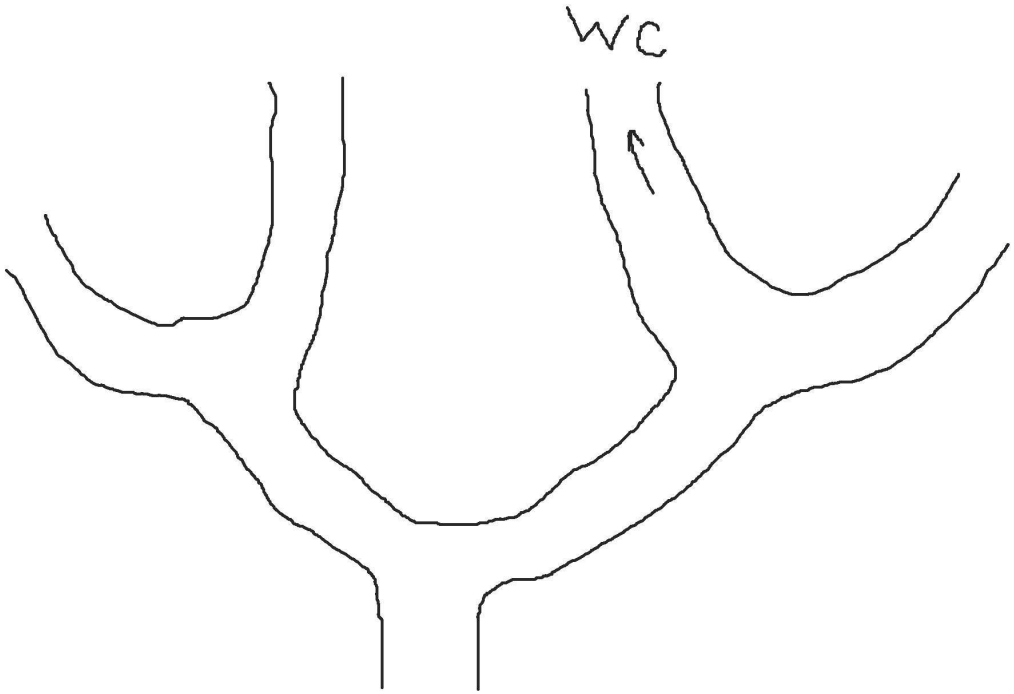
What is information?

What is information?

What is information?

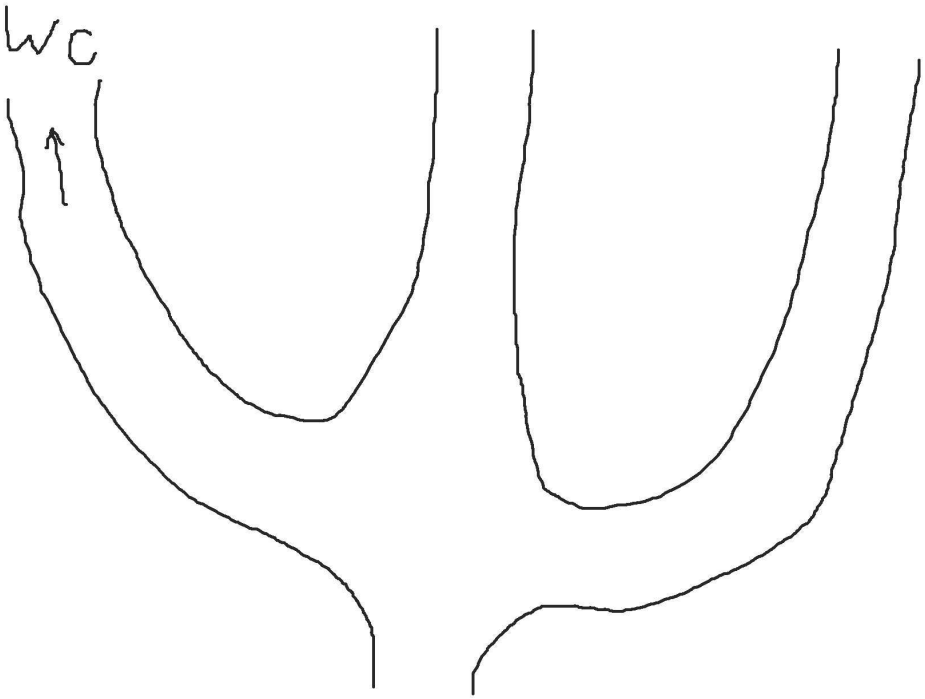


What is information?

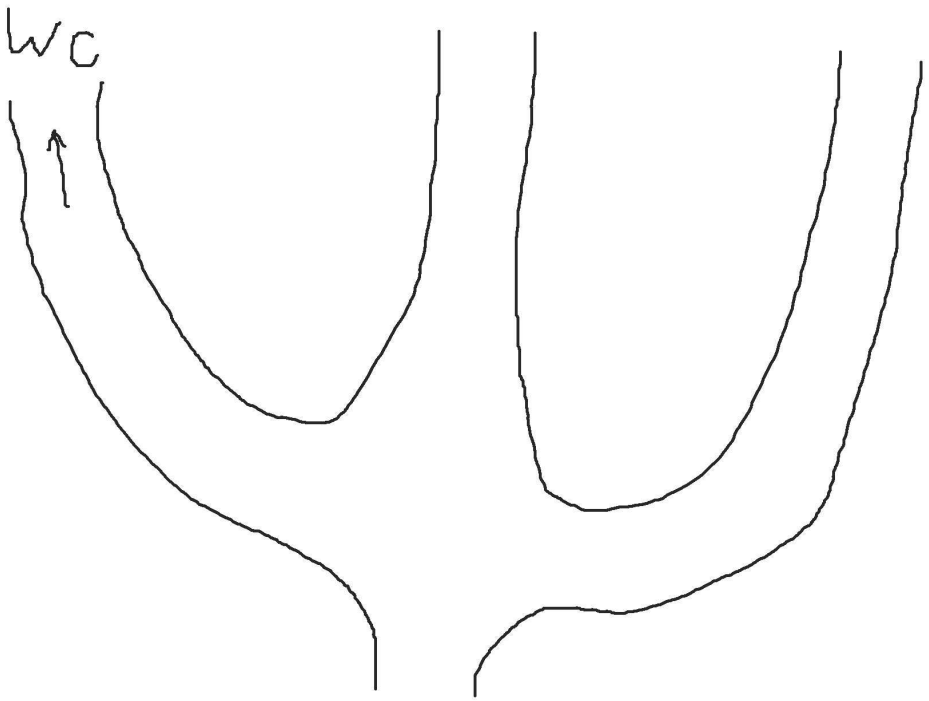


one of four choices = TWO BITS

What is information?

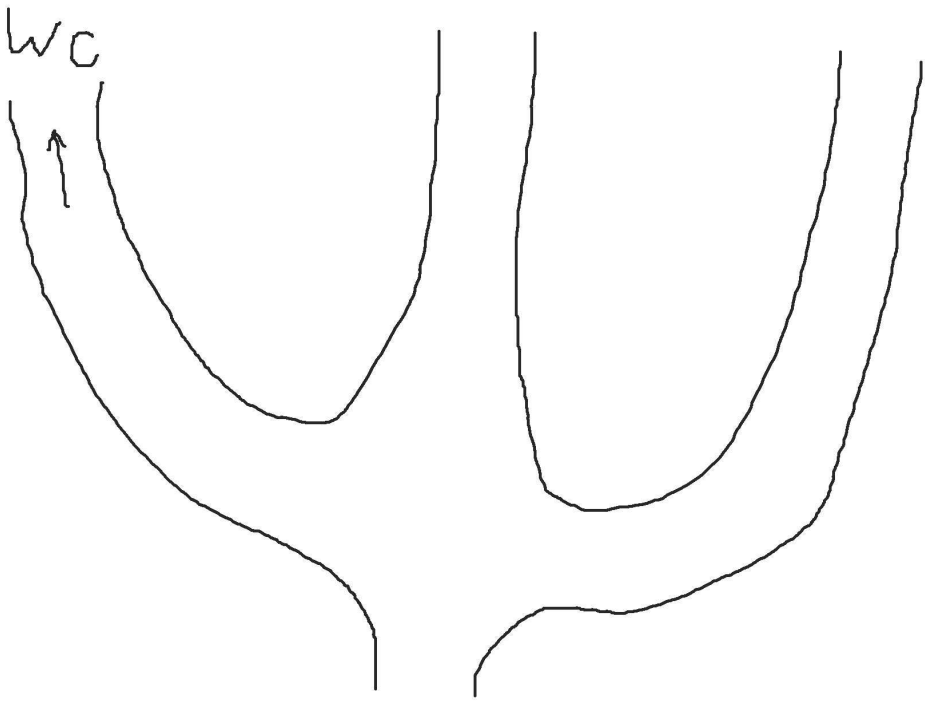


What is information?



one of three choices = ONE AND HALF BITS

What is information?



NO – this SCHOOL is about NONLINEAR SCIENCE!!!

What is information?

0 BITS = 1 choice

1 BIT = 2 choices

2 BITS = 4 choices

3 BITS = 8 choices

etc.

$$\#\,\text{BITS} = \log_2(\#\,\text{choices})$$

$$3\ \text{choices} = \log_2(3)\ \text{BITS} \approx 1.585\ \text{BITS}$$

What is information?

DID I WIN? (YES/NO)

What is information?

NO - 999999 of a million chances

I KNEW IT ANYWAY...

(there was almost only one choice – nearly no information gained)

What is information?

YES - 1 of a million chances
HURRA!!! (large information gained)

What is information?

YES - 1 of a million chances
HURRA!!! (large information gained)
BITS = $\log_2(1000000) \approx 19,9$ BITS

What is information?

NO - 999999 of a million chances

I KNEW IT ANYWAY...

(there was almost only one choice – nearly no information gained)

BITS = ???

What is information?

YES - 1 of a million chances
HURRA!!! (large information gained)
BITS = $\log_2(1000000) \approx 19,9$ BITS

$$\log_2(1000000) = -\log_2(\tfrac{1}{1000000}) = -\log_2(\text{probability of winning})$$

What is information?

NO - 999999 of a million chances

I KNEW IT ANYWAY...

(there was almost only one choice – nearly no information gained)

$$\#\,\text{BITS} = -\log_2(\text{probability of loosing}) = -\log_2(\tfrac{999999}{1000000}) \approx 0,0000014$$

Shannon information function

DEFINITION 1

If Ω is a finite probability space with atoms $x_1, x_2, \ldots$ of probabilities $P(x_i)$, $(i = 1, 2, \ldots)$, then the associated *information function* on Ω is defined as

$$I(x_i) = -\log_2(P(x_i)).$$

Shannon information function

DEFINITION 1

If Ω is a finite probability space with atoms $x_1, x_2, \ldots$ of probabilities $P(x_i)$, $(i = 1, 2, \ldots)$, then the associated *information function* on Ω is defined as

$$I(x_i) = -\log_2(P(x_i)).$$

If Ω is finite and has n elements of equal probabilities $\frac{1}{n}$ then the information function function is constant equal everywhere to $\log_2(n)$.

Shannon information function

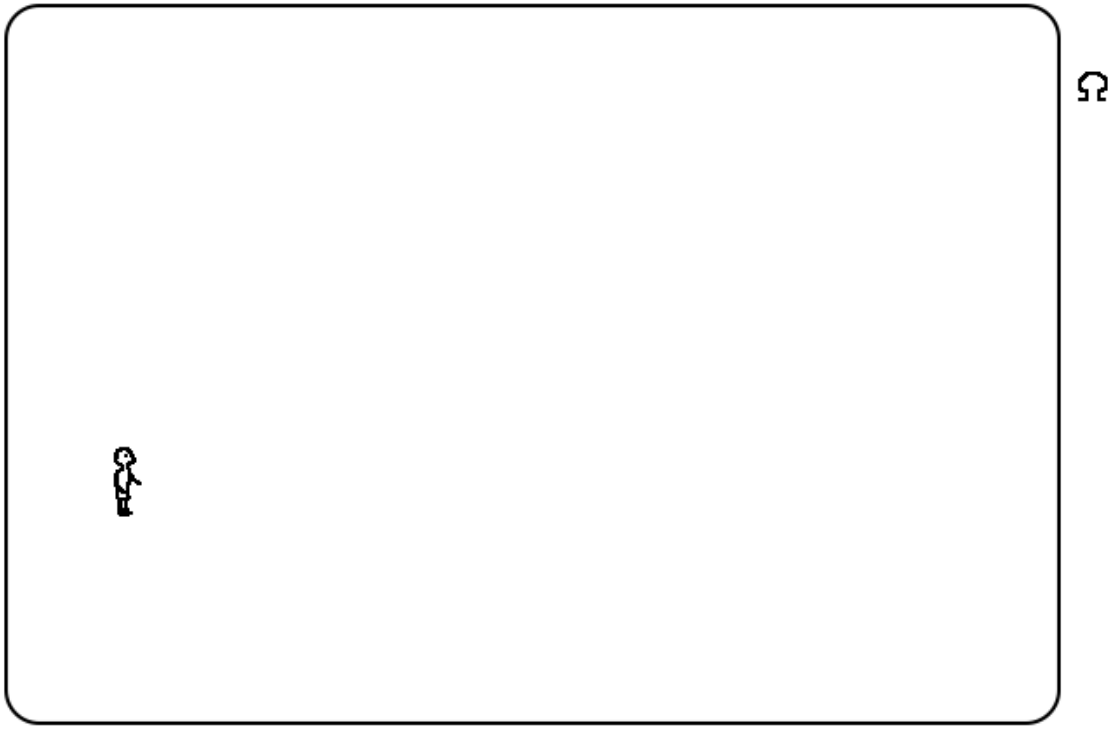
DEFINITION 2

If (Ω, Σ, μ) is a (perhaps non-atomic) probability space and $\mathcal{P} = \{P_1, P_2, \dots\}$ is a countable (or finite) measurable partition of Ω then the associated *information function* on Ω is defined as

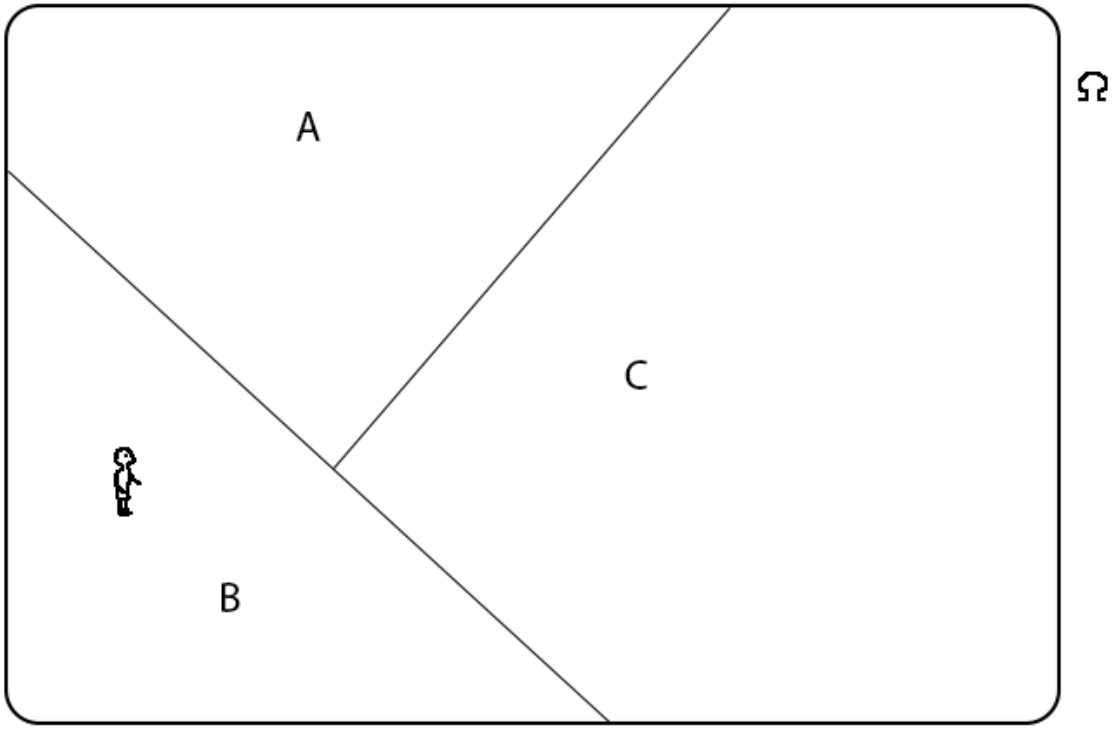
$$I_{\mathcal{P}}(x) = -\log_2(\mu(P_x)),$$

where P_x is the unique element of $\mathcal{P}$ such that $P_x \ni x$.

Shannon information function

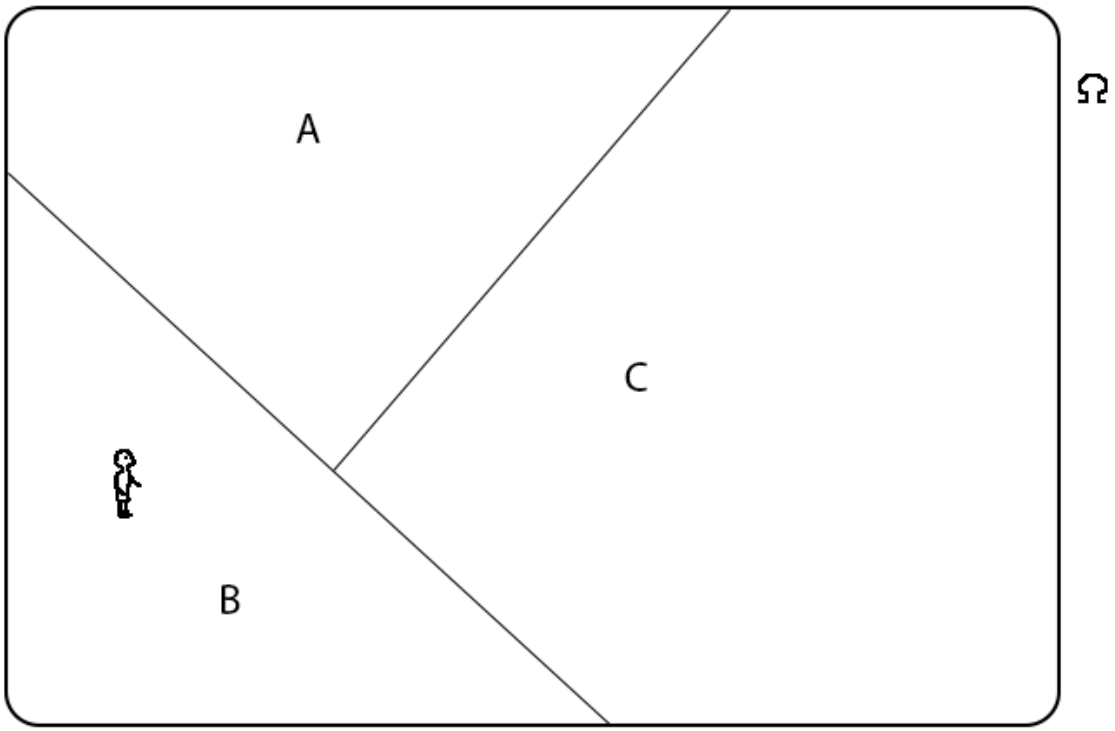


Shannon information function



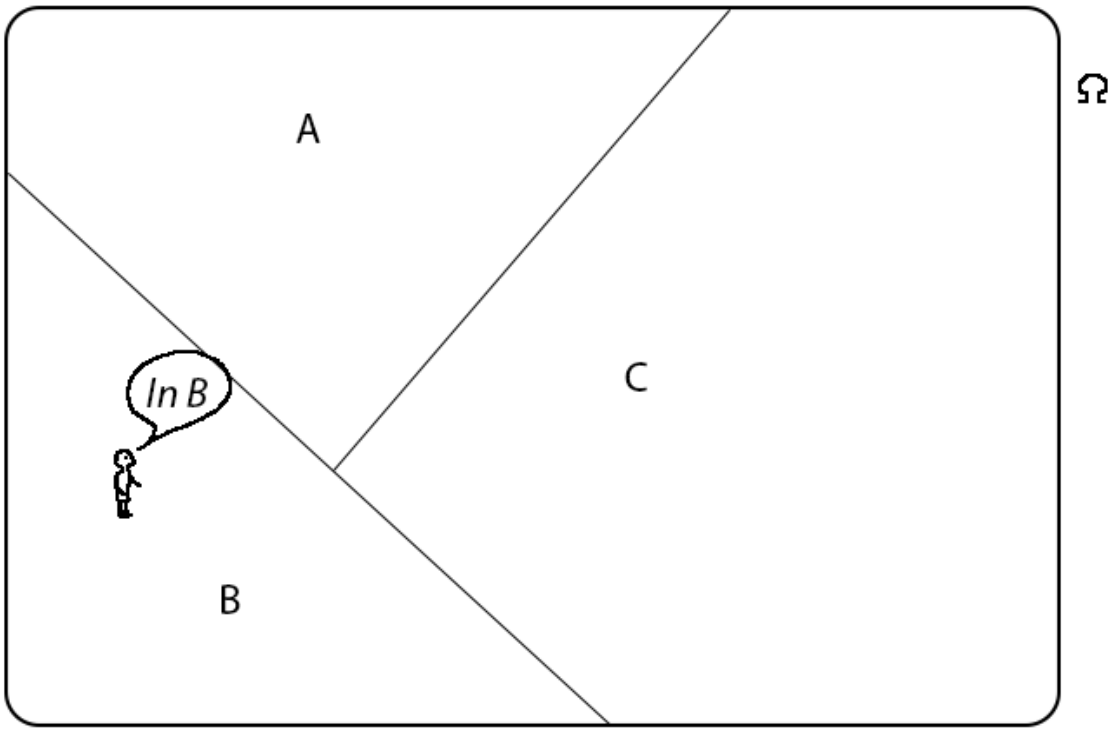
Shannon information function

Where are you?

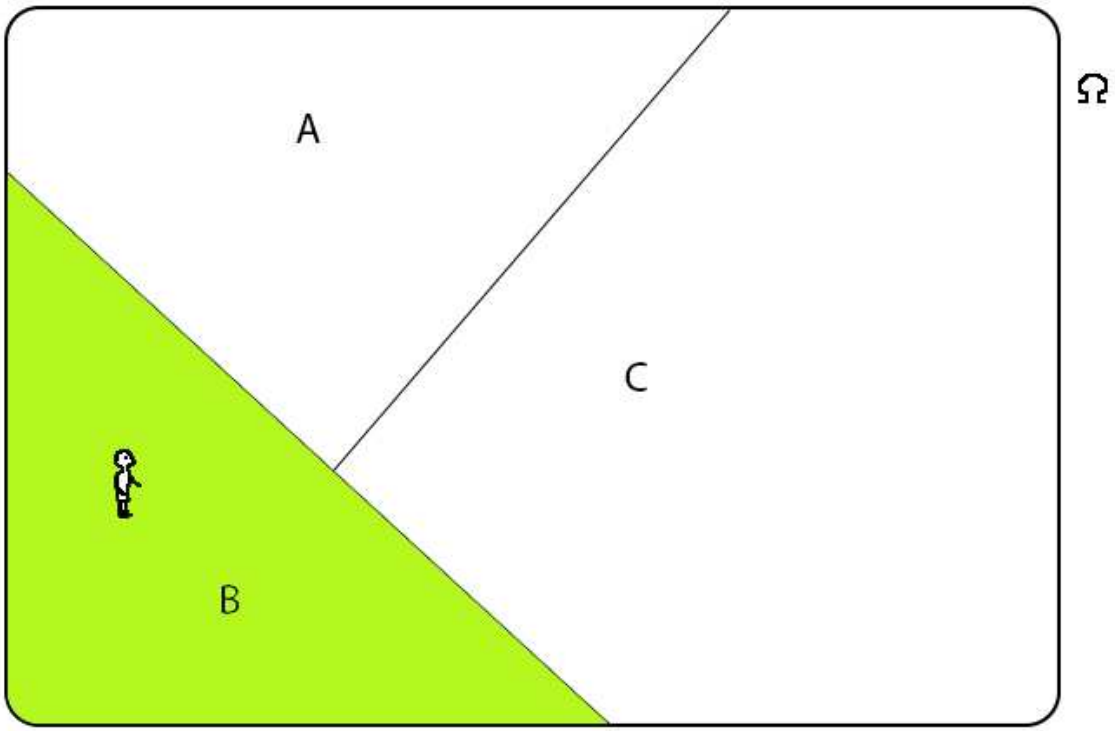


Shannon information function

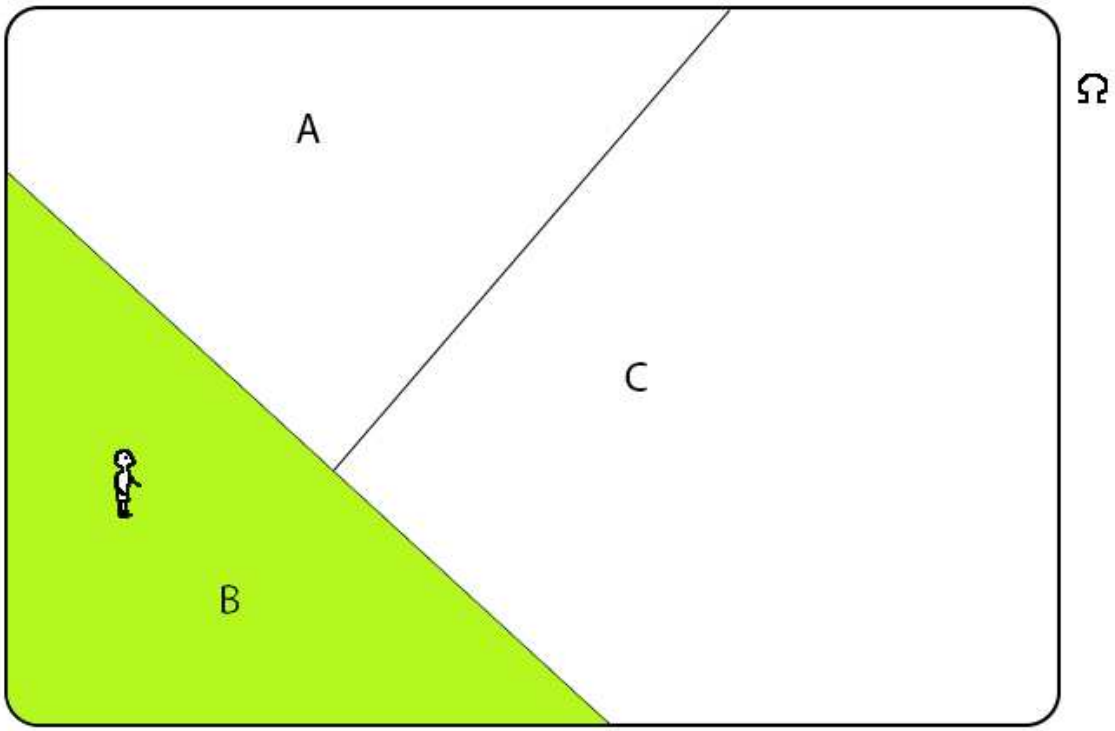
Where are you?



Shannon information function



Shannon information function



$$x \in B, \quad I_{\mathcal{P}}(x) = -\log \mu(B)$$

Shannon entropy of a partition

DEFINITION 3

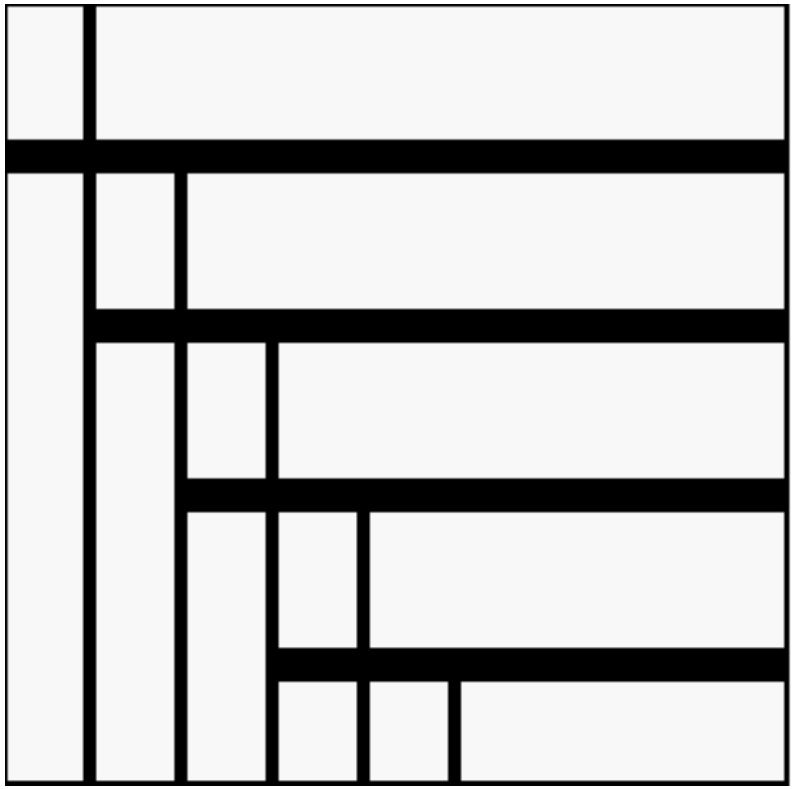
If (Ω, Σ, μ) is a probability space and $\mathcal{P} = \{P_1, P_2, \dots\}$ is a countable measurable partition of Ω then the *Shannon entropy* of $\mathcal{P}$ is defined as the expected value of the information function:

$$H(\mathcal{P}) = \int I_{\mathcal{P}} \, d\mu = -\sum_i \mu(P_i) \log_2 \mu(P_i)$$

(The *average over the space* information delivered by the partition.)

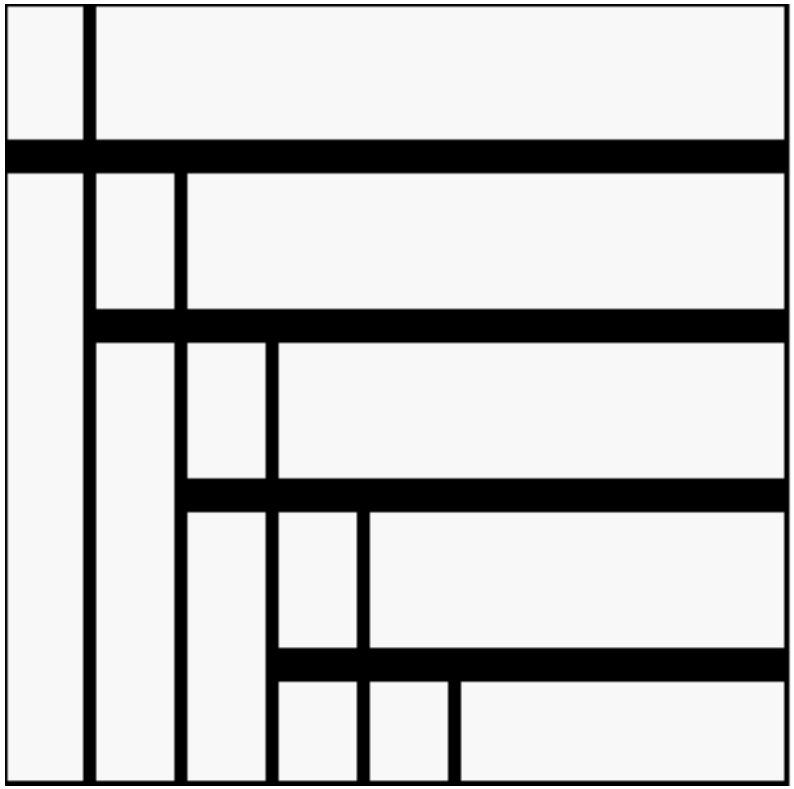
EXAMPLE

Consider the two bitmaps



EXAMPLE

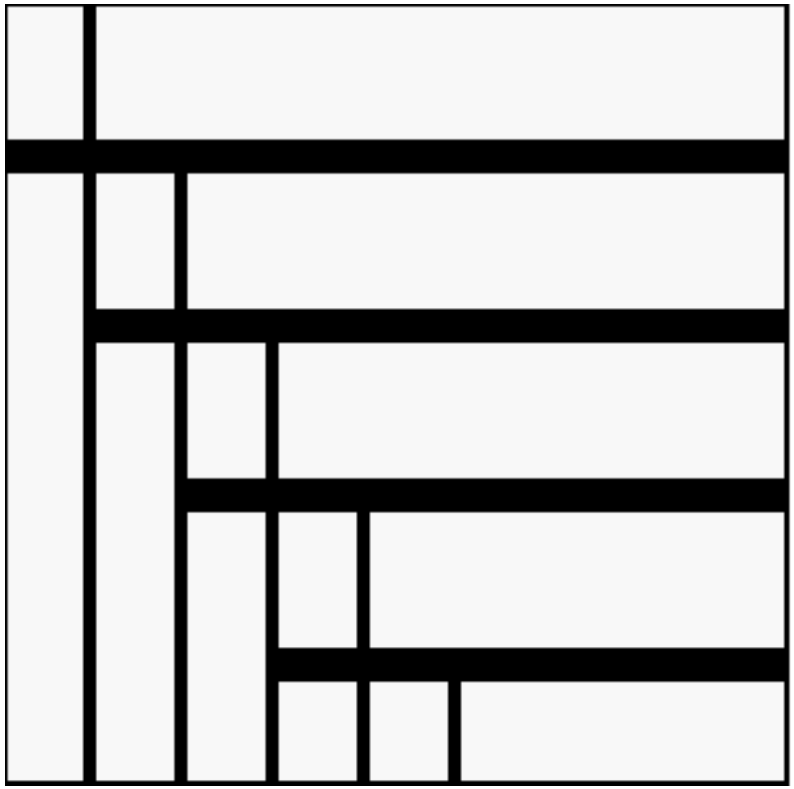
Consider the two bitmaps



They have the same sizes (even the same proportion of black and white). Thus they carry the same Shannon information (= # pixels). However...

EXAMPLE

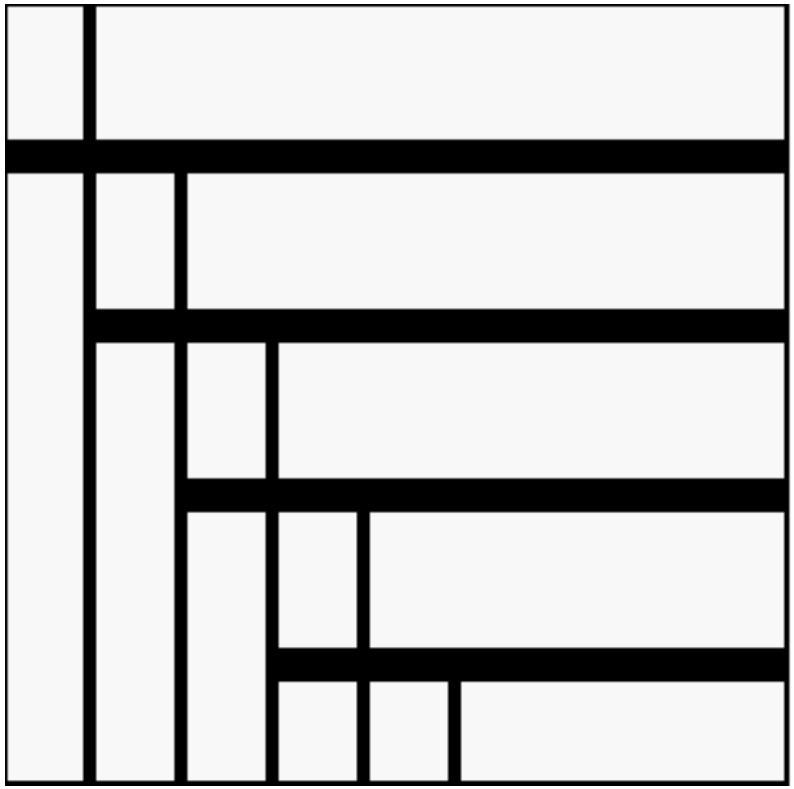
Consider the two bitmaps

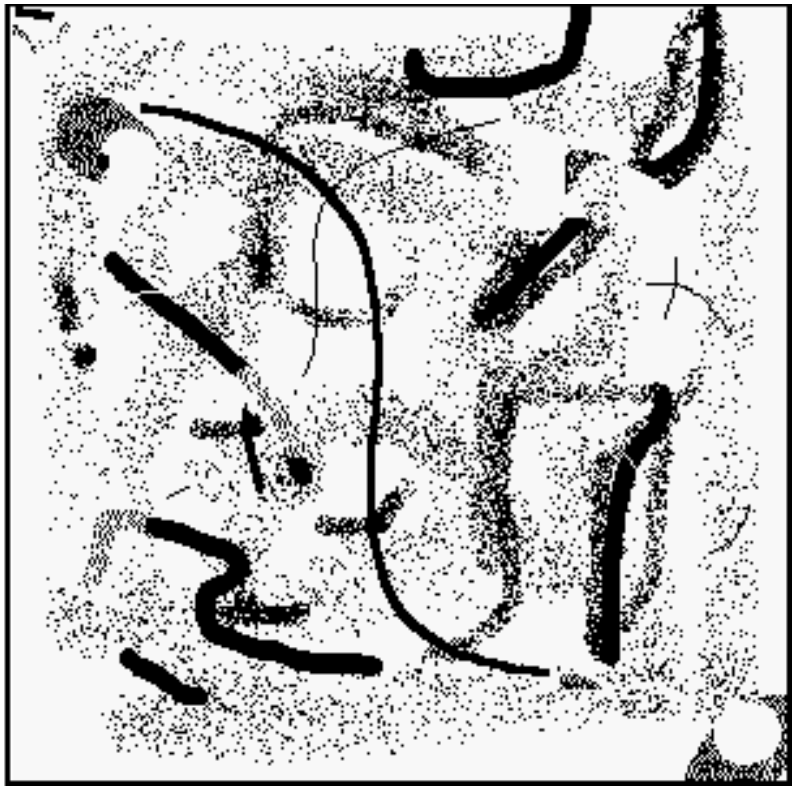


Any zipping program compresses the left hand side bitmap about 5 times more than the right hand side bitmap. Why?

EXAMPLE

Consider the two bitmaps





Imagine that you explain how to draw each bitmap over the phone...
How much INFORMATION is needed for each of them?

What makes the difference between these bitmaps, if both carry the same Shannon information?

What makes the difference between these bitmaps, if both carry the same Shannon information?

The answer is delivered by the dynamic entropy and the Shannon–McMillan–Breiman Theorem.

Dynamical systems

Now we will assume that on our probability space (Ω, Σ, μ) we have a measurable transformation $T : \Omega \to \Omega$ which preserves the measure μ, that is $\mu(T^{-1}(A)) = \mu(A)$ for every $A \in \Sigma$.

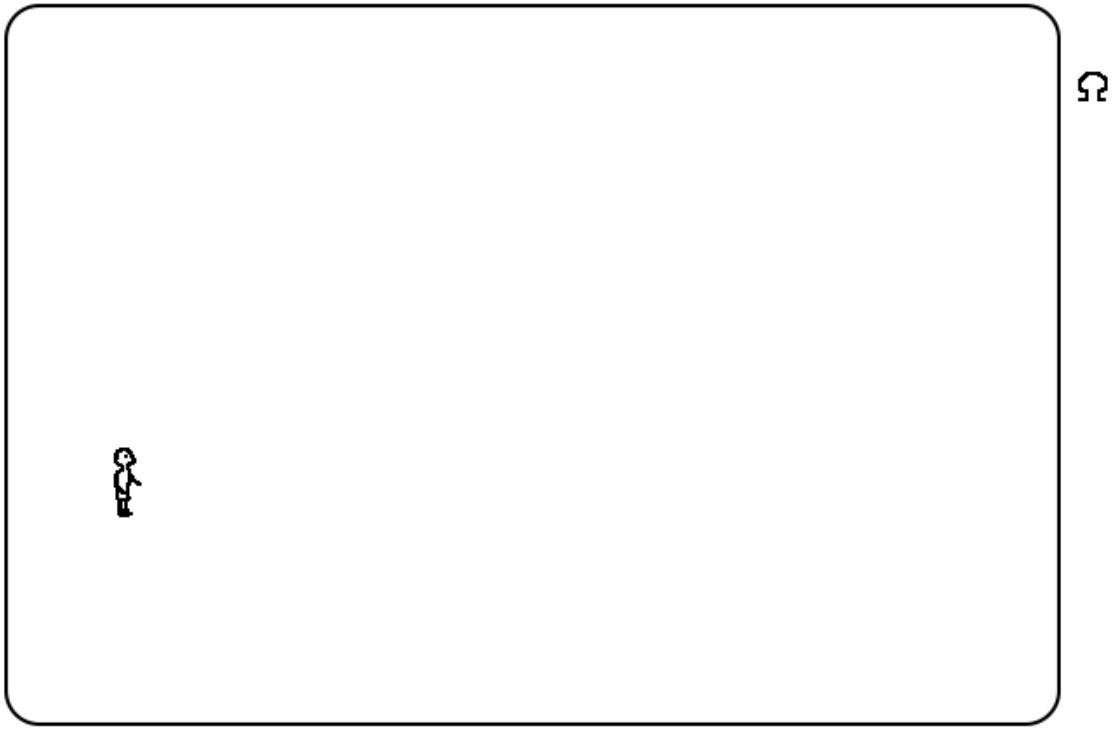
Dynamical systems

Now we will assume that on our probability space (Ω, Σ, μ) we have a measurable transformation $T : \Omega \to \Omega$ which preserves the measure μ, that is $\mu(T^{-1}(A)) = \mu(A)$ for every $A \in \Sigma$.

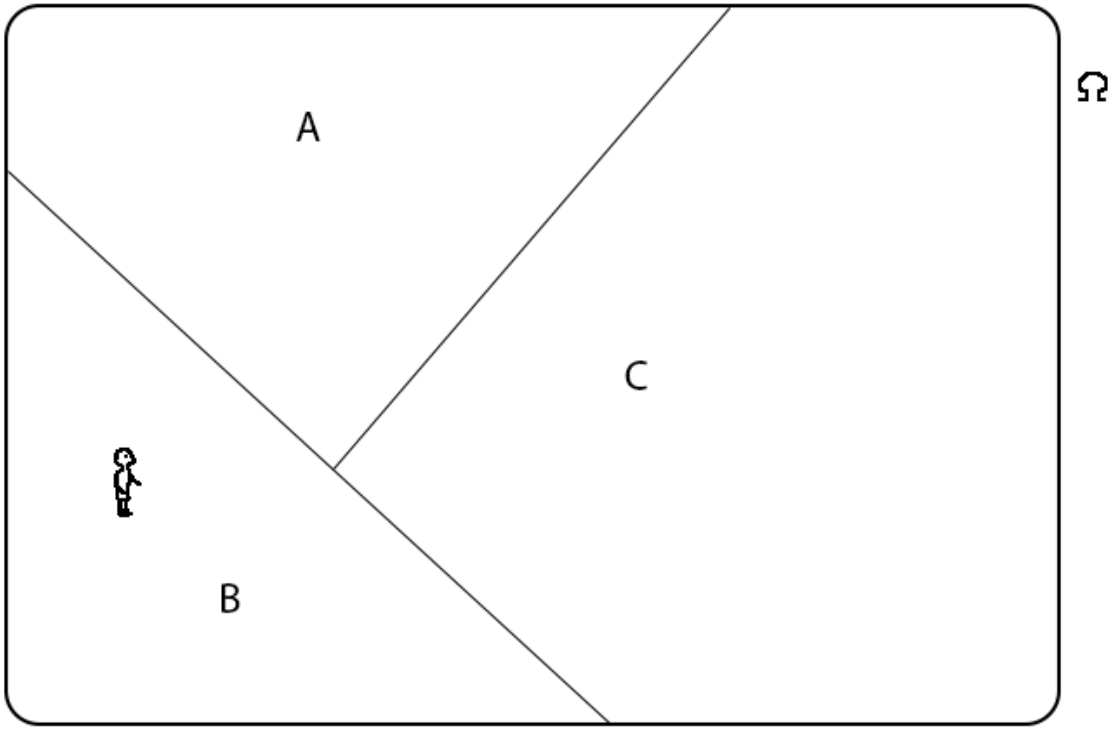
EXAMPLE

Let $\Omega = \{0, 1\}^{\mathbb{N}}$, $T =$ shift $(T(x_1, x_2, \dots) = (x_2, x_3, \dots))$ and μ is *some* shift-invariant measure. Every such measure is determined by its values on cylinders $C = [c_1, c_2, \dots, c_n]$.

Information in a dynamical system

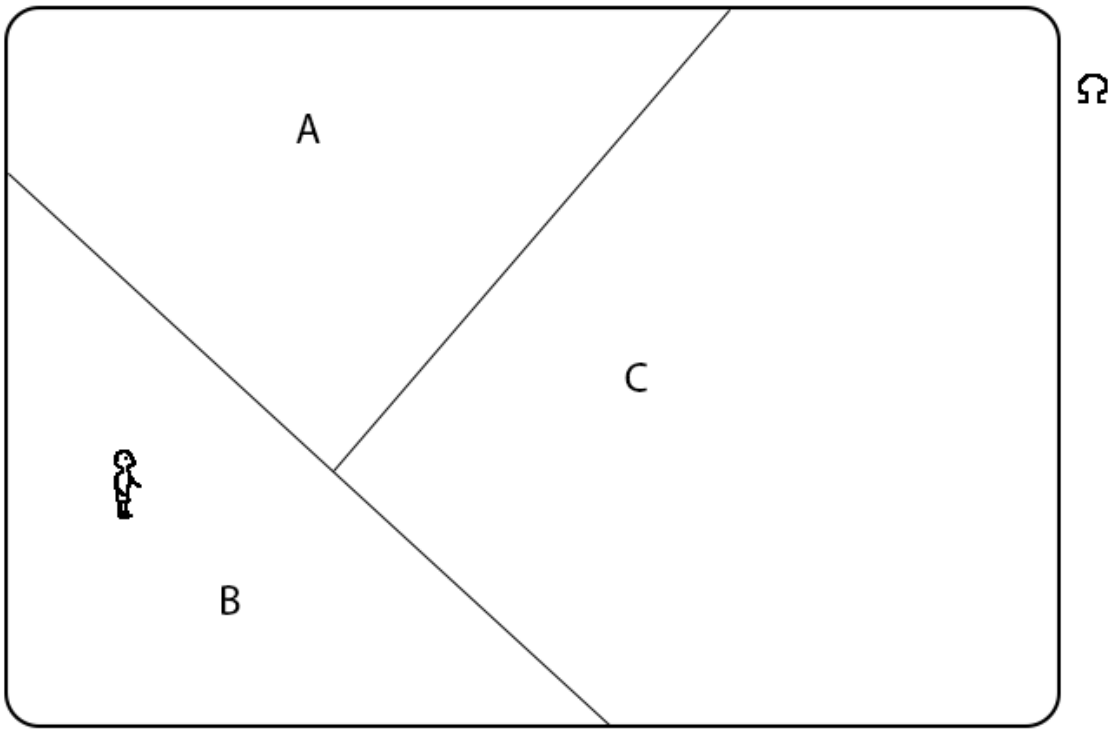


Information in a dynamical system



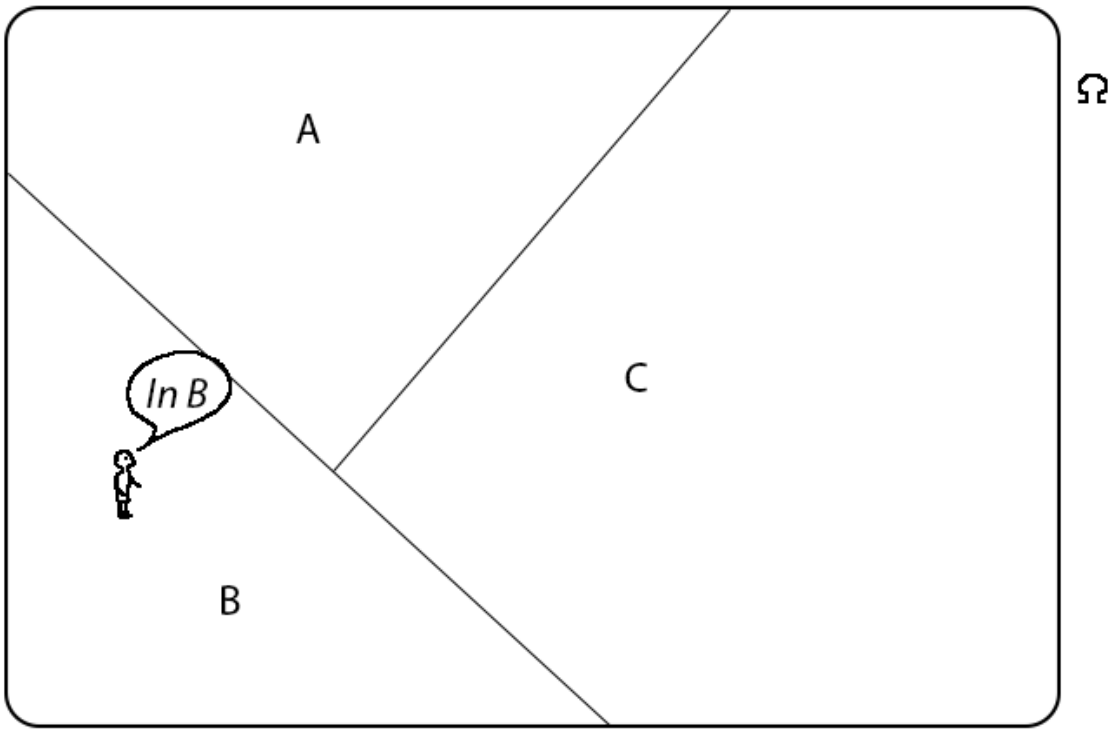
Information in a dynamical system

Where are you?

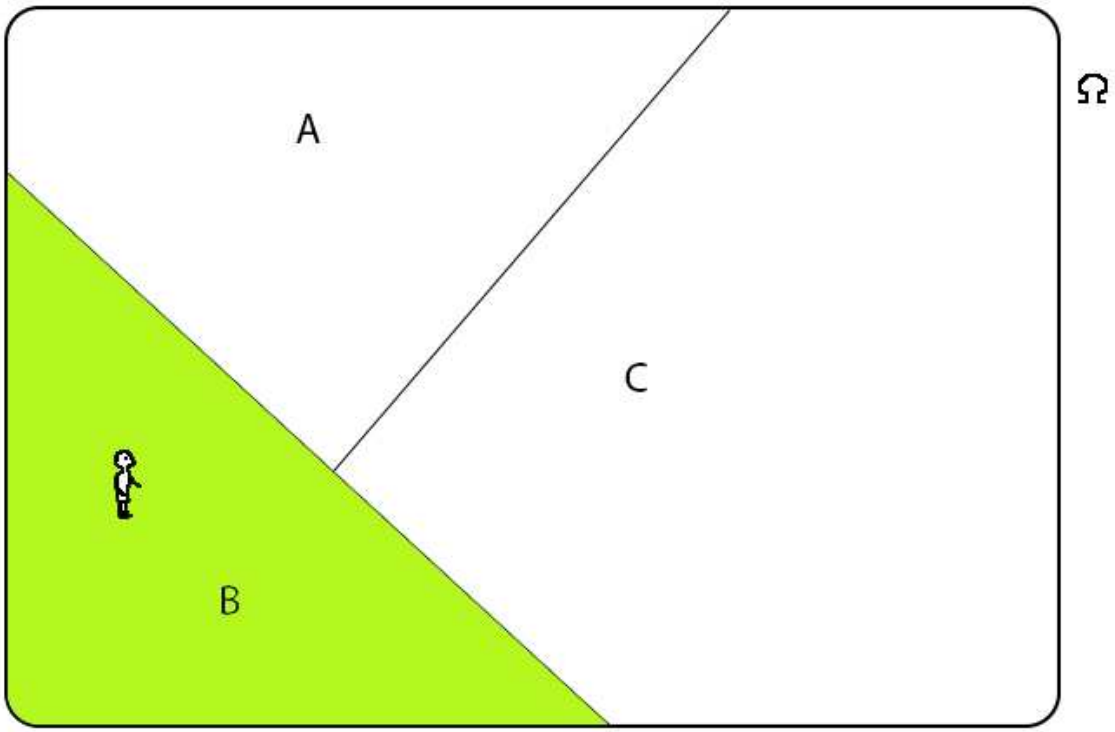


Information in a dynamical system

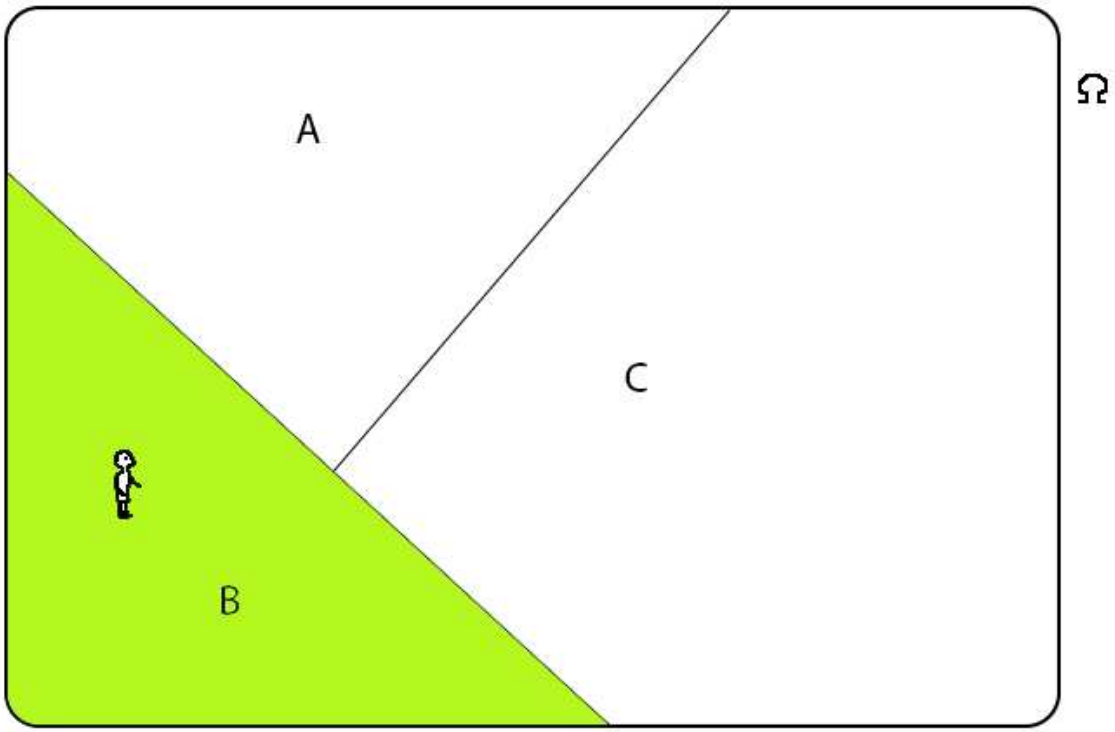
Where are you?



Information in a dynamical system



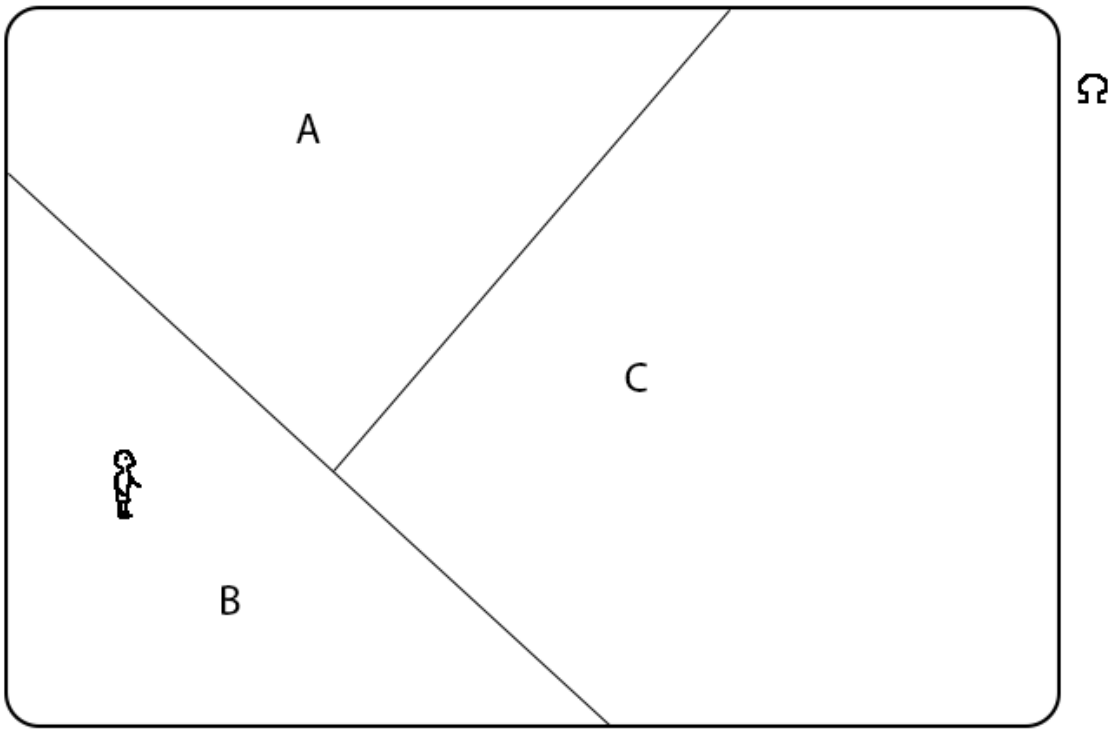
Information in a dynamical system



$$x \in B, \quad I_{\mathcal{P}}(x) = -\log \mu(B)$$

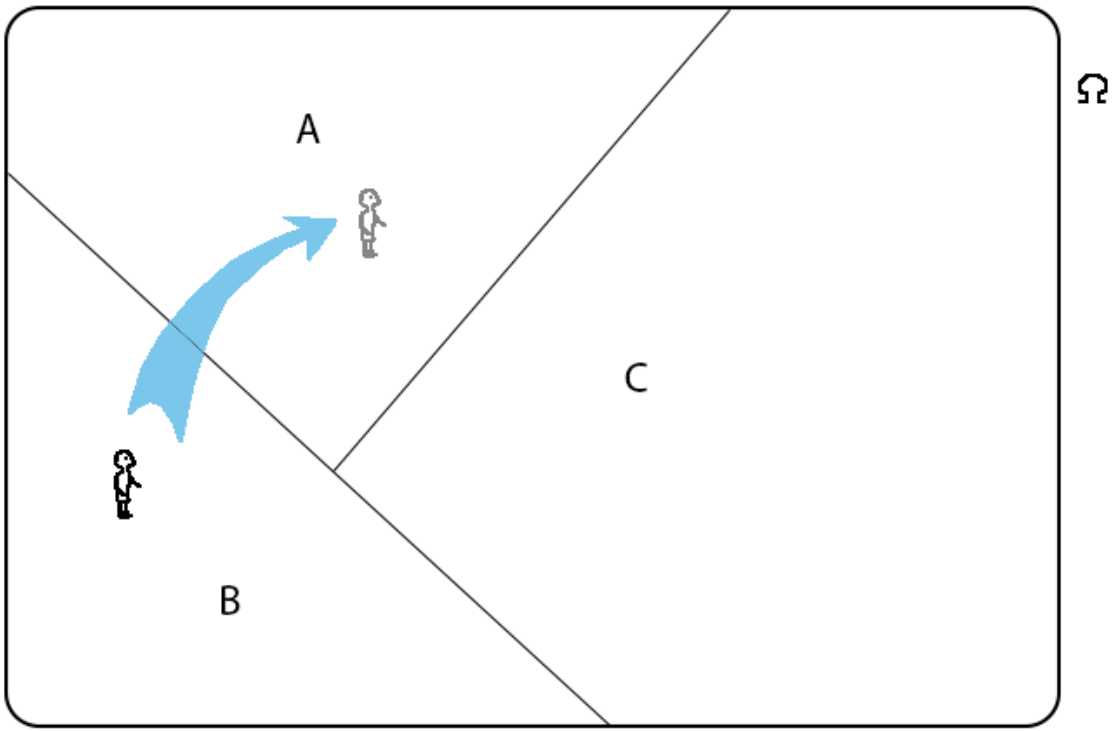
Information in a dynamical system

Where are you going?



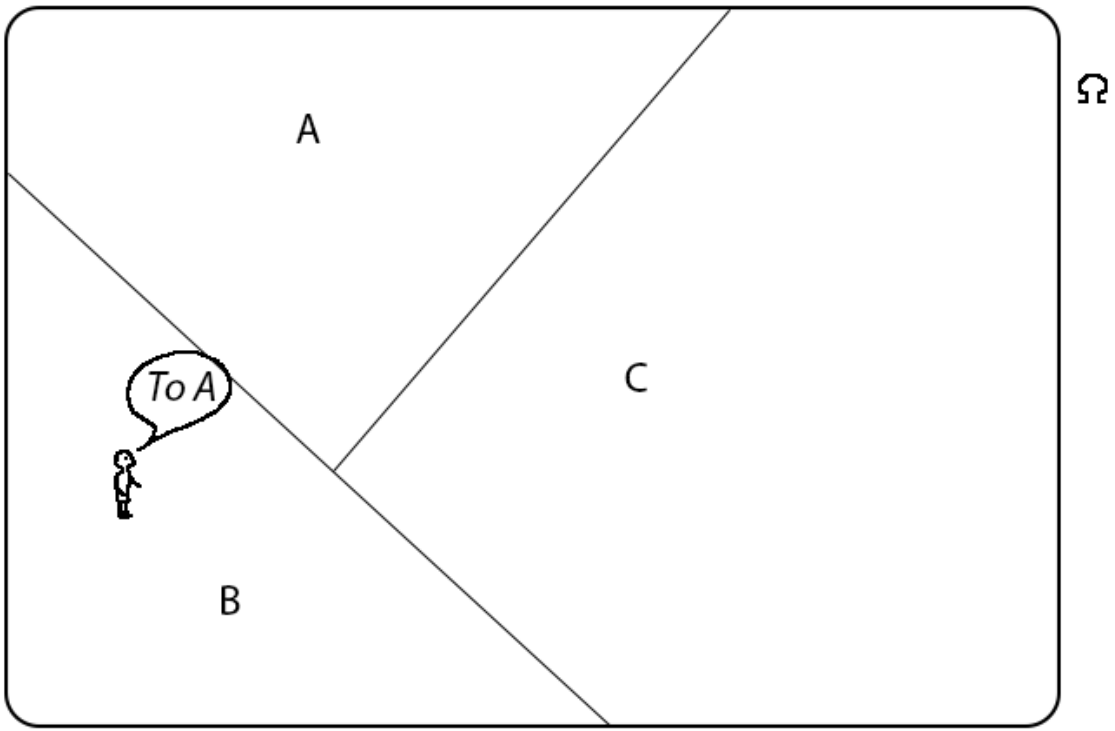
Information in a dynamical system

Where are you going?

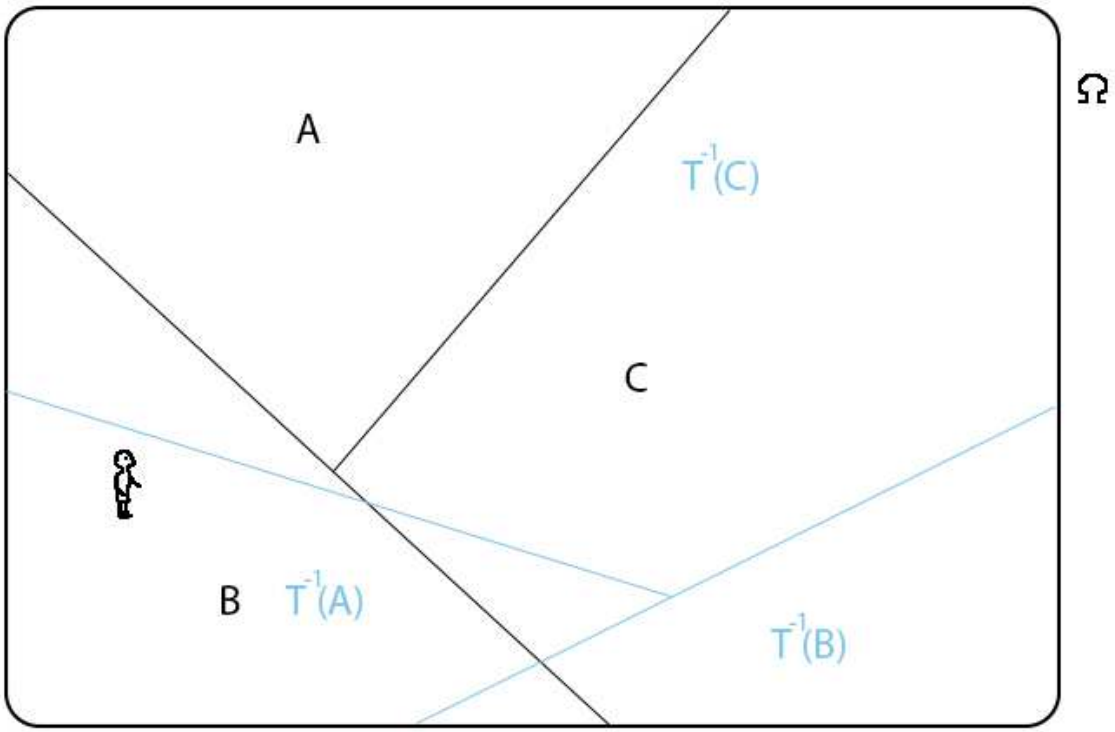


Information in a dynamical system

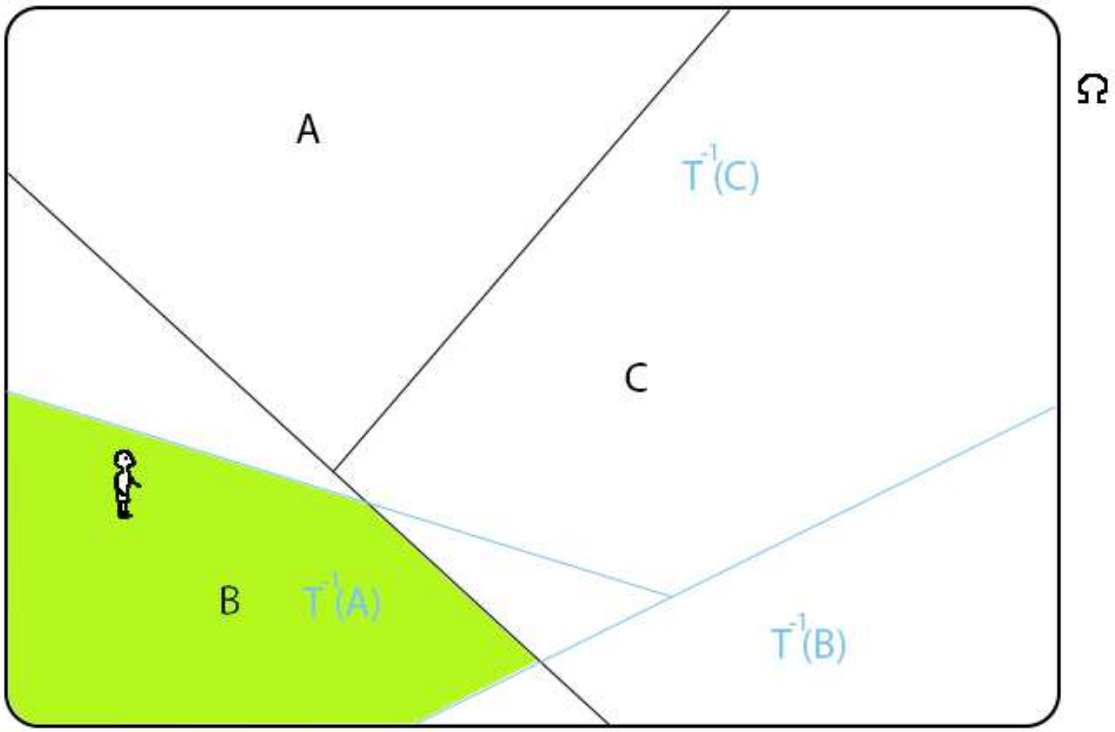
Where are you going?



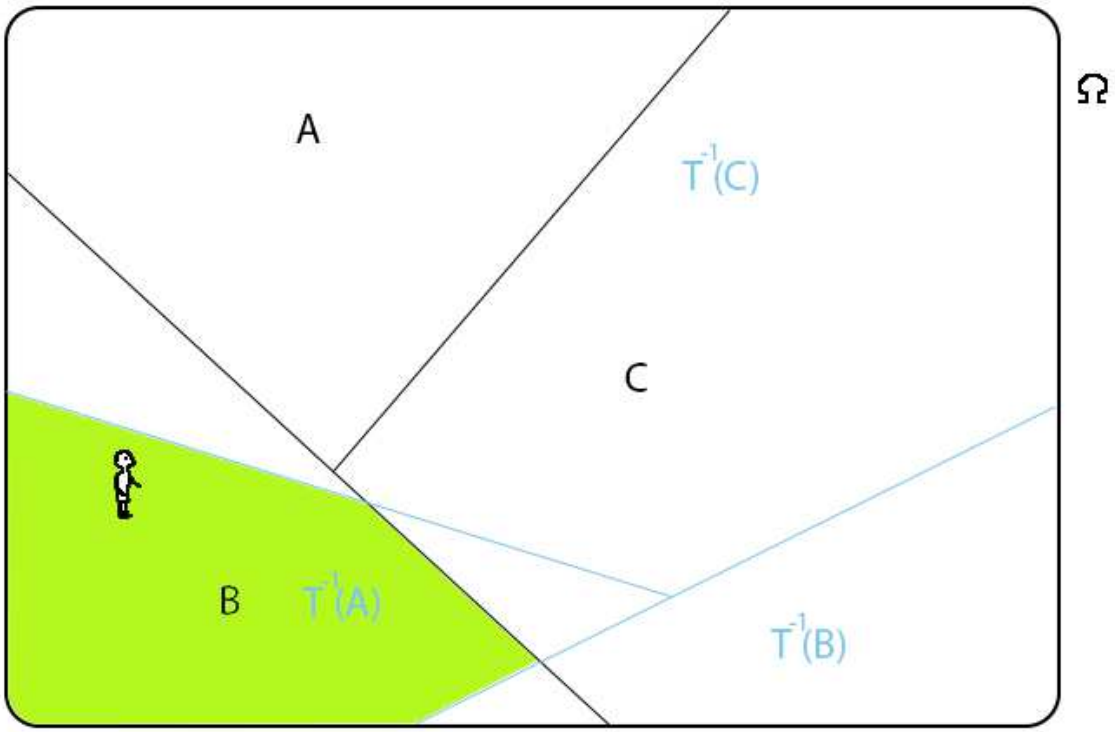
Information in a dynamical system



Information in a dynamical system



Information in a dynamical system



$$x \in B \cap T^{-1}(A), \qquad I_{\mathcal{P}^2}(x) = -\log \mu(B \cap T^{-1}(B))$$

Shannon information function in a dynamical system

DEFINITION 4

Let (Ω, Σ, μ) be a probability space and let $T : \Omega \to \Omega$ be a measurable and measure-preserving transformation. Let $\mathcal{P} = \{P_1, P_2, \dots\}$ be a countable measurable partition of Ω. Then the *information function in n steps* on Ω is defined as

$$I_{\mathcal{P}^n}(x) = -\log_2(\mu(P_x^n)),$$

where

$$P_x^n = P_x \cap T^{-1}(P_{Tx}) \cap T^{-2}(P_{T^2x}) \cap \cdots \cap T^{-n+1}(P_{T^{n-1}x})$$

(it is the unique element of the partition $\mathcal{P}^n := \bigvee_{i=0}^{n-1} T^{-i}(\mathcal{P})$ containing x, and is called the *n-cylinder of x*).

EXAMPLE

$$x \in [0, 1]$$

EXAMPLE

$$x \in [0, 1]$$

$$x = 0.765900862...$$

EXAMPLE

$$x \in [0, 1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

EXAMPLE

$$x \in [0, 1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.

EXAMPLE

$$x \in [0, 1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.

This corresponds to the flow of information in the dynamical system:

EXAMPLE

$$x \in [0,1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.

This corresponds to the flow of information in the dynamical system:

$$T : [0,1] \to [0,1],$$
$$T(x) = 10x \bmod 1,$$
μ is the Lebesgue measure

EXAMPLE

$$x \in [0,1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.

This corresponds to the flow of information in the dynamical system:

$$T : [0,1] \to [0,1],$$
$$T(x) = 10x \bmod 1,$$
μ is the Lebesgue measure
$$T(0.765900862...) = 0.65900862...,$$
$$T^2(0.765900862...) = T(0.65900862...) = 0.5900862...$$

EXAMPLE

$$x \in [0,1]$$

$$x = 0.765900862...$$

To fully identify x we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.

This corresponds to the flow of information in the dynamical system:

$$T : [0,1] \to [0,1],$$
$$T(x) = 10x \bmod 1,$$
μ is the Lebesgue measure
$$T(0.765900862...) = 0.65900862...,$$
$$T^2(0.765900862...) = T(0.65900862...) = 0.5900862...$$

$$\mathcal{P} = \{[0,\ 0.1), [0.1,\ 0.2), \ldots, [0.9,\ 1]\}$$

Dynamic entropy

Dynamic entropy

- $P^n(x) =$
  $\{$points that give the same answers as x through n times$\}$

Dynamic entropy

- $P^n(x) =$
  $\{$points that give the same answers as x through n times$\}$
- $I_{\mathcal{P}^n}(x) = -\log \mu(P^n(x))$

Dynamic entropy

- $P^n(x) =$
  $\{$points that give the same answers as x through n times$\}$
- $I_{\mathcal{P}^n}(x) = -\log \mu(P^n(x))$
- $H(\mathcal{P}^n) := \int I_{\mathcal{P}^n} \, d\mu$ (average *over space* information in n steps)

Dynamic entropy

- $P^n(x) =$
  $\{$points that give the same answers as x through n times$\}$
- $I_{\mathcal{P}^n}(x) = -\log \mu(P^n(x))$
- $H(\mathcal{P}^n) := \int I_{\mathcal{P}^n}\, d\mu$ (average *over space* information in n steps)

DEFINITION 5

The *dynamic entropy of the partition* $\mathcal{P}$ is defined as

$$h(T, \mathcal{P}) := \lim_n \tfrac{1}{n} H(\mathcal{P}^n).$$

Dynamic entropy

- $P^n(x) =$
  $\{$points that give the same answers as x through n times$\}$
- $I_{\mathcal{P}^n}(x) = -\log \mu(P^n(x))$
- $H(\mathcal{P}^n) := \int I_{\mathcal{P}^n}\, d\mu$ (average *over space* information in n steps)

DEFINITION 5

The *dynamic entropy of the partition* $\mathcal{P}$ is defined as

$$h(T, \mathcal{P}) := \lim_n \tfrac{1}{n} H(\mathcal{P}^n).$$

The dynamic entropy is interpreted as the average *over space and time* gain of information per step.

Shannon–McMillan–Breiman Theorem

Shannon–McMillan–Breiman Theorem

THEOREM 1

If μ ergodic then

$$\frac{1}{n} I_{\mathcal{P}^n}(x) \overset{\mu-\text{a.e.}}{\underset{n\to\infty}{\longrightarrow}} h(T, \mathcal{P})$$

Shannon–McMillan–Breiman Theorem

THEOREM 1

If μ ergodic then

$$\frac{1}{n} I_{\mathcal{P}^n}(x) \overset{\mu-\text{a.e.}}{\underset{n\to\infty}{\longrightarrow}} h(T, \mathcal{P})$$

That is, the *average gain of information per step* does not depend on the initial point.

EXAMPLE

Let $\Omega = \{0,1\}^{\mathbb{N}}$, $T =$ shift and μ is *some ergodic* shift-invariant measure. Then for a μ-"typical" point $x = (x_1, x_2, \dots)$ the measure of a long initial cylinder $x[1,n] := [x_1, x_2, \dots, x_n]$ is approximately $2^{-nh(T,\mathcal{P})}$, where $\mathcal{P}$ is the two-element partition $\{[0],[1]\}$.

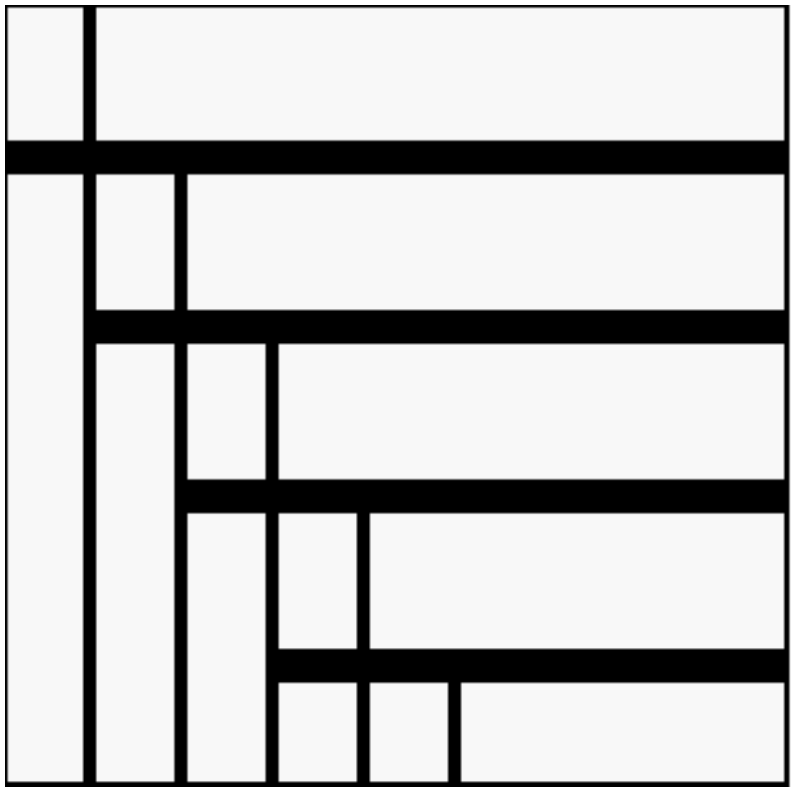
EXAMPLE

Let $\Omega = \{0,1\}^{\mathbb{N}}$, $T =$ shift and μ is *some ergodic* shift-invariant measure. Then for a μ-"typical" point $x = (x_1, x_2, \dots)$ the measure of a long initial cylinder $x[1,n] := [x_1, x_2, \dots, x_n]$ is approximately $2^{-nh(T,\mathcal{P})}$, where $\mathcal{P}$ is the two-element partition $\{[0],[1]\}$.

The meaning of "approximately" is very rough, it means only that $-\frac{1}{n}\log_2 \mu(x[1,n]) \approx h(T,\mathcal{P})$.

EXAMPLE

- Let us go back to our example with the two bitmaps:



Roughly speaking, the bitmaps represent long pieces of orbits of "typical points" in symbolic systems (over two symbols "white" and "black"), with two different invariant and ergodic measures having different entropies h_1, h_2.

Roughly speaking, the bitmaps represent long pieces of orbits of "typical points" in symbolic systems (over two symbols "white" and "black"), with two different invariant and ergodic measures having different entropies h_1, h_2.

The first bitmap is "highly organized" (in fact periodic), hence has small entropy, the second one is "highly random", hence has large entropy, thus $h_1 \ll h_2$.

Roughly speaking, the bitmaps represent long pieces of orbits of "typical points" in symbolic systems (over two symbols "white" and "black"), with two different invariant and ergodic measures having different entropies h_1, h_2.

The first bitmap is "highly organized" (in fact periodic), hence has small entropy, the second one is "highly random", hence has large entropy, thus $h_1 \ll h_2$.

The entropies represent the *average information contents per symbol*. By the Shannon–McMillan–Breiman Theorem, the same average information contents per symbol occurs already in these orbits (bitmaps).

Roughly speaking, the bitmaps represent long pieces of orbits of "typical points" in symbolic systems (over two symbols "white" and "black"), with two different invariant and ergodic measures having different entropies h_1, h_2.

The first bitmap is "highly organized" (in fact periodic), hence has small entropy, the second one is "highly random", hence has large entropy, thus $h_1 \ll h_2$.

The entropies represent the *average information contents per symbol*. By the Shannon–McMillan–Breiman Theorem, the same average information contents per symbol occurs already in these orbits (bitmaps).

So the *effective information* carried by the bitmaps is proportional to h_1 and h_2, respectively (times the # of pixels). This explains the huge difference.

Everything that was said in this presentation will be given rigorous explanation during the rest of the course...

Everything that was said in this presentation will be given rigorous explanation during the rest of the course...

using more traditional media, such as blackboard (or whiteboard) and chalk (or markers).