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?

Which way to the bathroom?
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?

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.

# BITS = \log_2(# choices)

3 choices = \log_2(3) BITS \approx 1.585 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

\[
\log_2(1000000) = -\log_2\left(\frac{1}{1000000}\right) = -\log_2(\text{probability of winning})
\]
What is information?

NO - 999999 of a million chances

I KNEW IT ANYWAY...

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

\[ \# \text{ BITS} = - \log_2(\text{probability of loosing}) = - \log_2\left(\frac{999999}{1000000}\right) \approx 0.0000014 \]
DEFINITION 1

If $\Omega$ 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 $\Omega$ is defined as

$$I(x_i) = -\log_2(P(x_i)).$$
DEFINITION 1

If $\Omega$ 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 $\Omega$ is defined as

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

If $\Omega$ is finite and has $n$ elements of equal probabilities $\frac{1}{n}$ then the information function function is constant equal everywhere to $\log_2(n)$. 
DEFINITION 2

If $(\Omega, \Sigma, \mu)$ is a (perhaps non-atomic) probability space and $\mathcal{P} = \{P_1, P_2, \ldots\}$ is a countable (or finite) measurable partition of $\Omega$ then the associated *information function* on $\Omega$ 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?

A

B

C

Ω
Shannon information function

Where are you?

A

B

C

Ω

In B
Shannon information function
Shannon information function

\[ x \in B, \quad I_P(x) = -\log \mu(B) \]
DEFINITION 3

If \((\Omega, \Sigma, \mu)\) is a probability space and \(\mathcal{P} = \{P_1, P_2, \ldots\}\) is a countable measurable partition of \(\Omega\) then the **Shannon entropy** of \(\mathcal{P}\) is defined as the expected value of the information function:

\[
H(\mathcal{P}) = \int_I \log_2 \mu \, d\mu = - \sum_i \mu(P_i) \log_2 \mu(P_i)
\]

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

Any zipping program compresses the left hand side bitmap about 5 times more than the right hand side bitmap. Why?
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.
Now we will assume that on our probability space \((\Omega, \Sigma, \mu)\) we have a measurable transformation \(T : \Omega \rightarrow \Omega\) which preserves the measure \(\mu\), that is \(\mu(T^{-1}(A)) = \mu(A)\) for every \(A \in \Sigma\).
Now we will assume that on our probability space \((\Omega, \Sigma, \mu)\) we have a measurable transformation \(T : \Omega \to \Omega\) which preserves the measure \(\mu\), that is \(\mu(T^{-1}(A)) = \mu(A)\) for every \(A \in \Sigma\).

**EXAMPLE**

Let \(\Omega = \{0, 1\}^\mathbb{N}\), \(T = \text{shift}\) \(T(x_1, x_2, \ldots) = (x_2, x_3, \ldots)\) and \(\mu\) is some shift-invariant measure. Every such measure is determined by its values on cylinders \(C = [c_1, c_2, \ldots, c_n]\).
Information in a dynamical system
Information in a dynamical system
Information in a dynamical system

Where are you?

A

B

C

Ω
Information in a dynamical system

Where are you?

A

Ω

B

C

In B
Information in a dynamical system
Information in a dynamical system

\[ x \in B, \quad I_P(x) = -\log \mu(B) \]
Information in a dynamical system

Where are you going?
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

$x \in B \cap T^{-1}(A), \quad I_{p2}(x) = -\log \mu(B \cap T^{-1}(B))$
DEFINITION 4

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

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

where

\[
P^n_x = P_x \cap T^{-1}(P_{Tx}) \cap T^{-2}(P_{T^2x}) \cap \cdots \cap T^{-n+1}(P_{T^{n-1}x})
\]

(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

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

\[ x = 0.765900862… \]
\[ x \in [0, 1] \]

\[ x = 0.765900862\ldots \]

To fully identify \( x \) we need infinite amount of information.
$x \in [0, 1]$

$x = 0.765900862...$

To fully identify $x$ we need infinite amount of information.

With each digit we acquire $\log_2(10)$ BITS of information.
\[ x \in [0, 1] \]

\[ x = 0.765900862\ldots \]

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:
\[ x \in [0, 1] \]
\[ x = 0.765900862... \]

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:

\[ T : [0, 1] \rightarrow [0, 1], \]
\[ T(x) = 10x \mod 1, \]
\( \mu \) is the Lebesgue measure
\[ x \in [0, 1] \]

\[ x = 0.765900862\ldots \]

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:

\[ T : [0, 1] \rightarrow [0, 1], \]
\[ T(x) = 10x \mod 1, \]
\[ \mu \text{ is the Lebesgue measure} \]
\[ T(0.765900862\ldots) = 0.65900862\ldots, \]
\[ T^2(0.765900862\ldots) = T(0.65900862\ldots) = 0.5900862\ldots \]
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:

\[ T : [0, 1] \rightarrow [0, 1], \]

\[ T(x) = 10x \mod 1, \]

\( \mu \) 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

\[ P^n(x) = \{ \text{points that give the same answers as } x \text{ through } n \text{ times} \} \]
Dynamic entropy

\[ P^n(x) = \{ \text{points that give the same answers as } x \text{ through } n \text{ times} \} \]

\[ l_{P^n}(x) = - \log \mu(P^n(x)) \]
Dynamic entropy

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

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

**DEFINITION 5**

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

$$h(T, \mathcal{P}) := \lim_{n \to \infty} \frac{1}{n} H(\mathcal{P}^n).$$
Dynamic entropy

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

**DEFINITION 5**

The *dynamic entropy of the partition* $P$ is defined as

$$h(T, P) := \lim_{n \to \infty} \frac{1}{n} H(P^n).$$

The dynamic entropy is interpreted as the average *over space and time* gain of information per step.
Shannon–McMillan–Breiman Theorem
THEOREM 1

If $\mu$ ergodic then

$$\frac{1}{n} \int_{\mathbb{P}^n} (x) \xrightarrow[n \to \infty]{\mu-a.e.} h(T, \mathbb{P})$$
THEOREM 1

If $\mu$ ergodic then

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

That is, the *average gain of information per step* does not depend on the initial point.
Let \( \Omega = \{0, 1\}^\mathbb{N} \), \( T = \text{shift} \) and \( \mu \) is some ergodic shift-invariant measure. Then for a \( \mu \)-“typical” point \( x = (x_1, x_2, \ldots) \) the measure of a long initial cylinder \( x[1, n] := [x_1, x_2, \ldots, x_n] \) is approximately \( 2^{-nh(T;\mathcal{P})} \), where \( \mathcal{P} \) is the two-element partition \( \{[0], [1]\} \).
Let $\Omega = \{0, 1\}^\mathbb{N}$, $T = \text{shift}$ and $\mu$ is some ergodic shift-invariant measure. Then for a $\mu$-“typical” point $x = (x_1, x_2, \ldots)$ the measure of a long initial cylinder $x[1, n] := [x_1, x_2, \ldots, x_n]$ is approximately $2^{-n h(T, \mathcal{P})}$, where $\mathcal{P}$ is the two-element partition $\{[0], [1]\}$.

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