A tutorial on information theory

Juan M. López

Instituto de Física de Cantabria, CSIC and Universidad de Cantabria, Spain

IFCA Instituto de Física de Cantabria

CSIC Consejo Superior de Investigaciones Científicas

UC Universidad de Cantabria

Outline

Instituto de Física de Cantabria

1.- What is information?

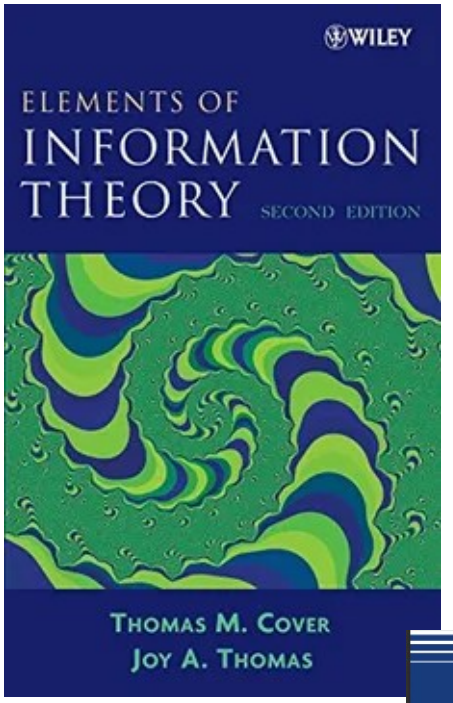
2.- How do we measure information?

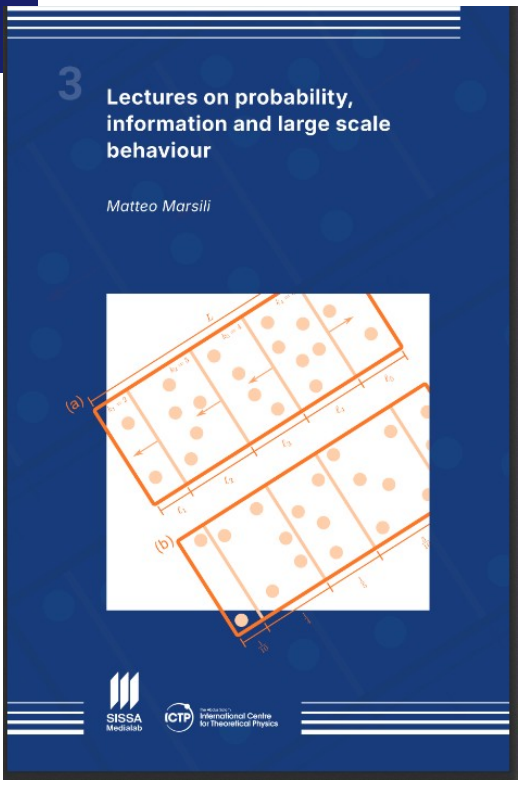
3.- Information redundancy/correlations/patterns

4.- Information of processes

5.- Examples

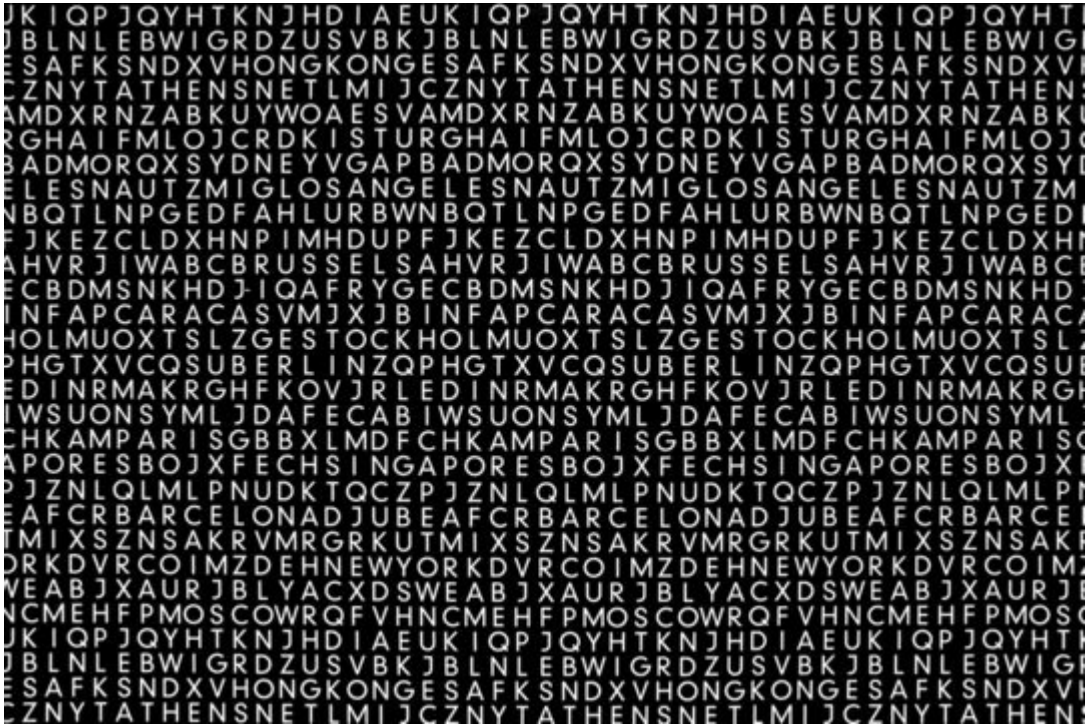
6.- Maximum entropy principle (MaxEnt)





What is information?

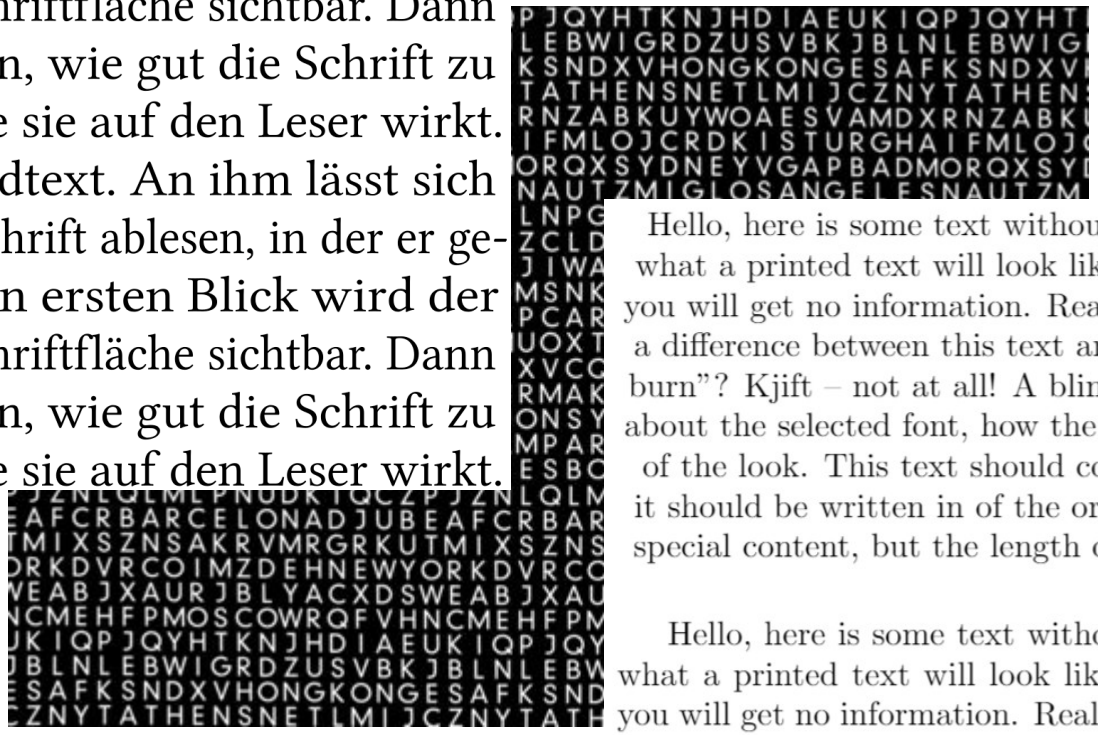
Instituto de Física de Cantabria



What is information?

Instituto de Física de Cantabria

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.
Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schriftfläche sichtbar. Dann kann man prüfen, wie gut die Schrift zu lesen ist und wie sie auf den Leser wirkt.



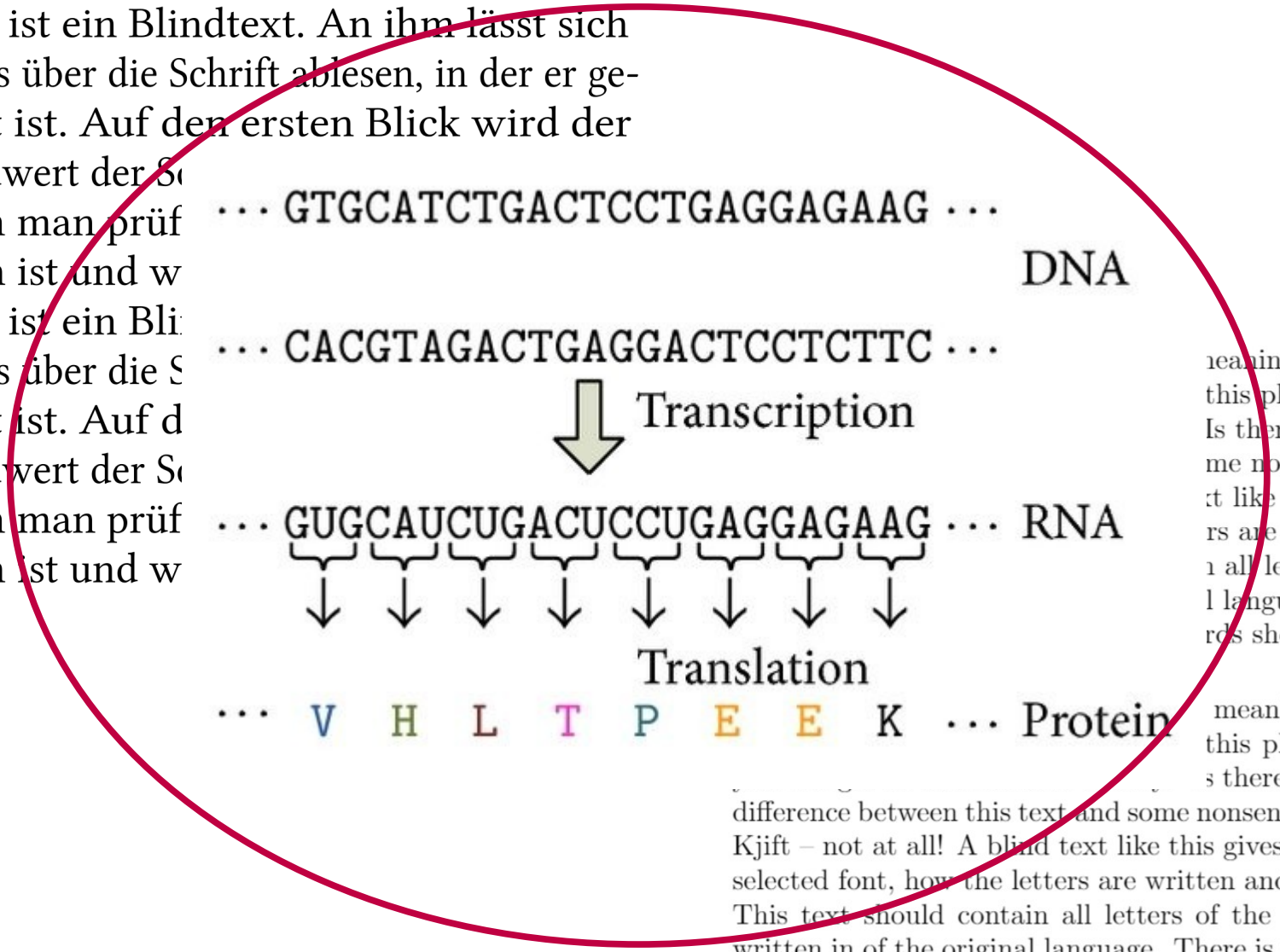
Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

What is information?

Instituto de Física de Cantabria

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der S
kann man prüf
lesen ist und w
Dies ist ein Bli
vieles über die S
setzt ist. Auf d
Grauwert der S
kann man prüf
lesen ist und w



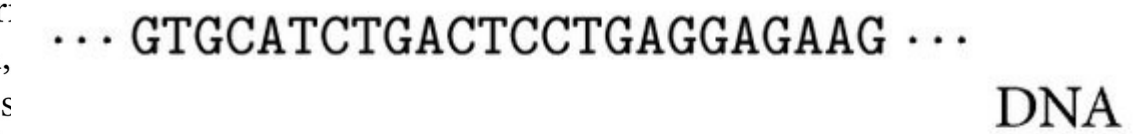
meaning. This text should show this place. If you read this text, Is there no information? Is there me nonsense like "Huardest gef- t like this gives you information rs are written and an impression l all letters of the alphabet and l language. There is no need for rds should match the language.

meaning. This text should show this place. If you read this text, s there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

What is information?

Instituto de Física de Cantabria

Dies ist ein Blindtext. An ihm lässt sich vieles über die Schrift ablesen, in der er gesetzt ist. Auf den ersten Blick wird der Grauwert der Schr
kann man prüfen,
lesen ist und wie s
Dies ist ein Blindt
vieles übe
setzt ist.
Grauwert
kann ma
lesen ist



uld show
this text,
ı? Is there
rdest gef-
formation
mpression
abet and
need for
age.

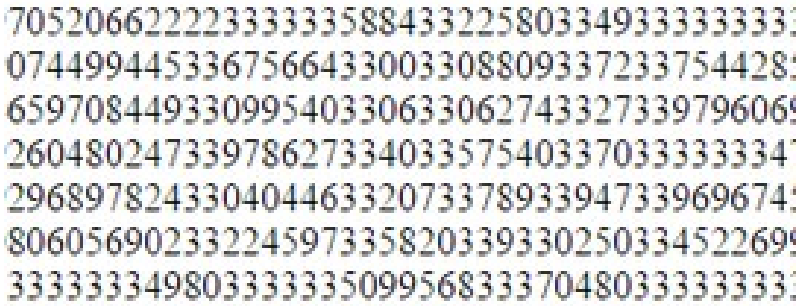
show
text,
here a
urn"?
ut the
look.
uld be
ntent,

```python
# Example of avoiding complex expressions

def square_numbers(numbers):
    """Return a list of the squares of the numbers."""
    squares = []
    for number in numbers:
        squares.append(number**2)
    return squares


def main():
    numbers = [1, 2, 3, 4]
    result = square_numbers(numbers)
    print(result)
```

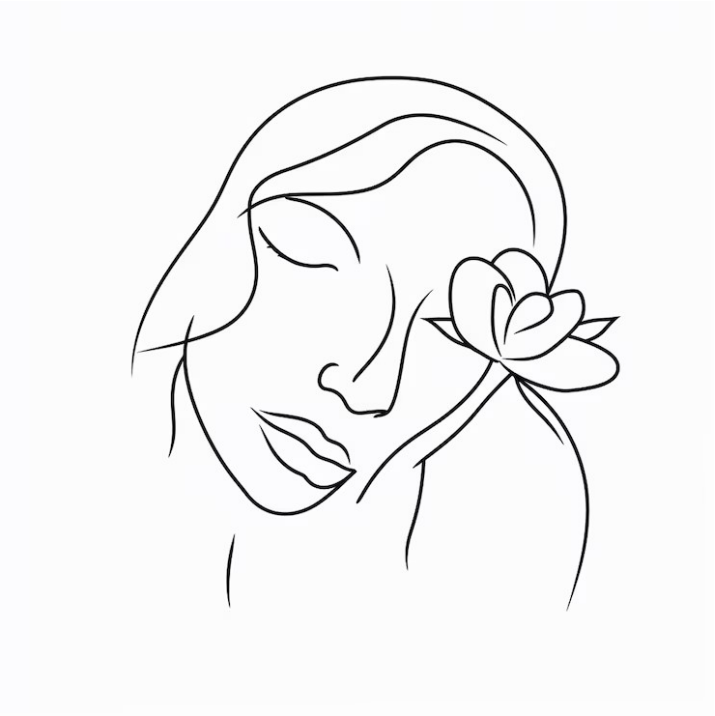
What is information?

Instituto de Física de Cantabria

3.14159265358979323846264338327950288419716939937510582097494459230781640628620899862803482534211706798214808651328230664709384460955058223172535940812848111745028410270193852110555964462294895493038196442881097566593344612847



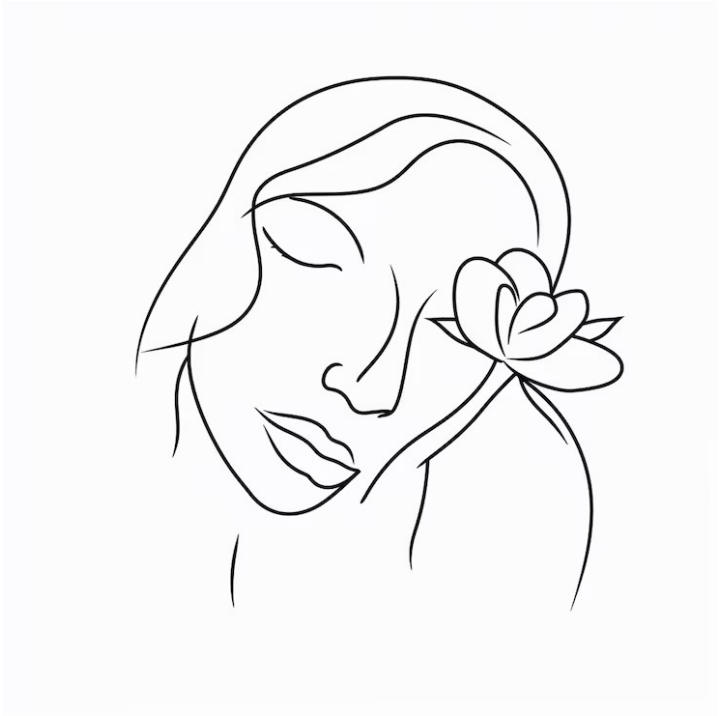
What is information?

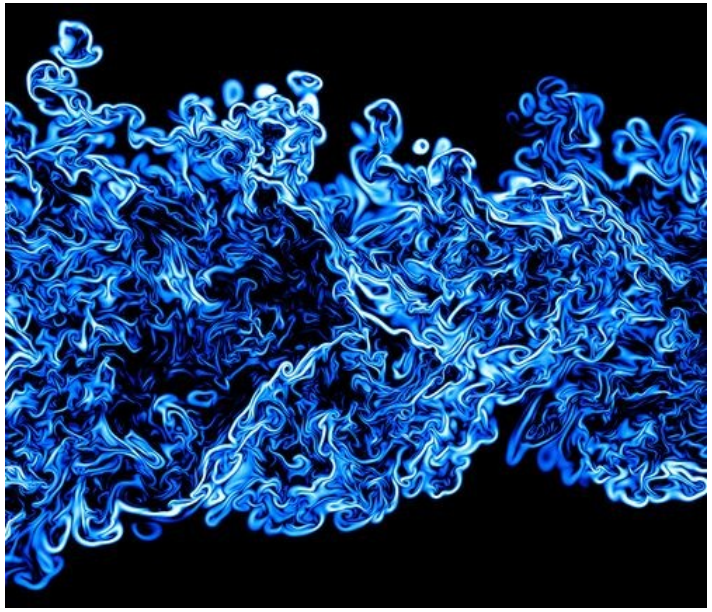


Instituto de Física de Cantabria

What is information?

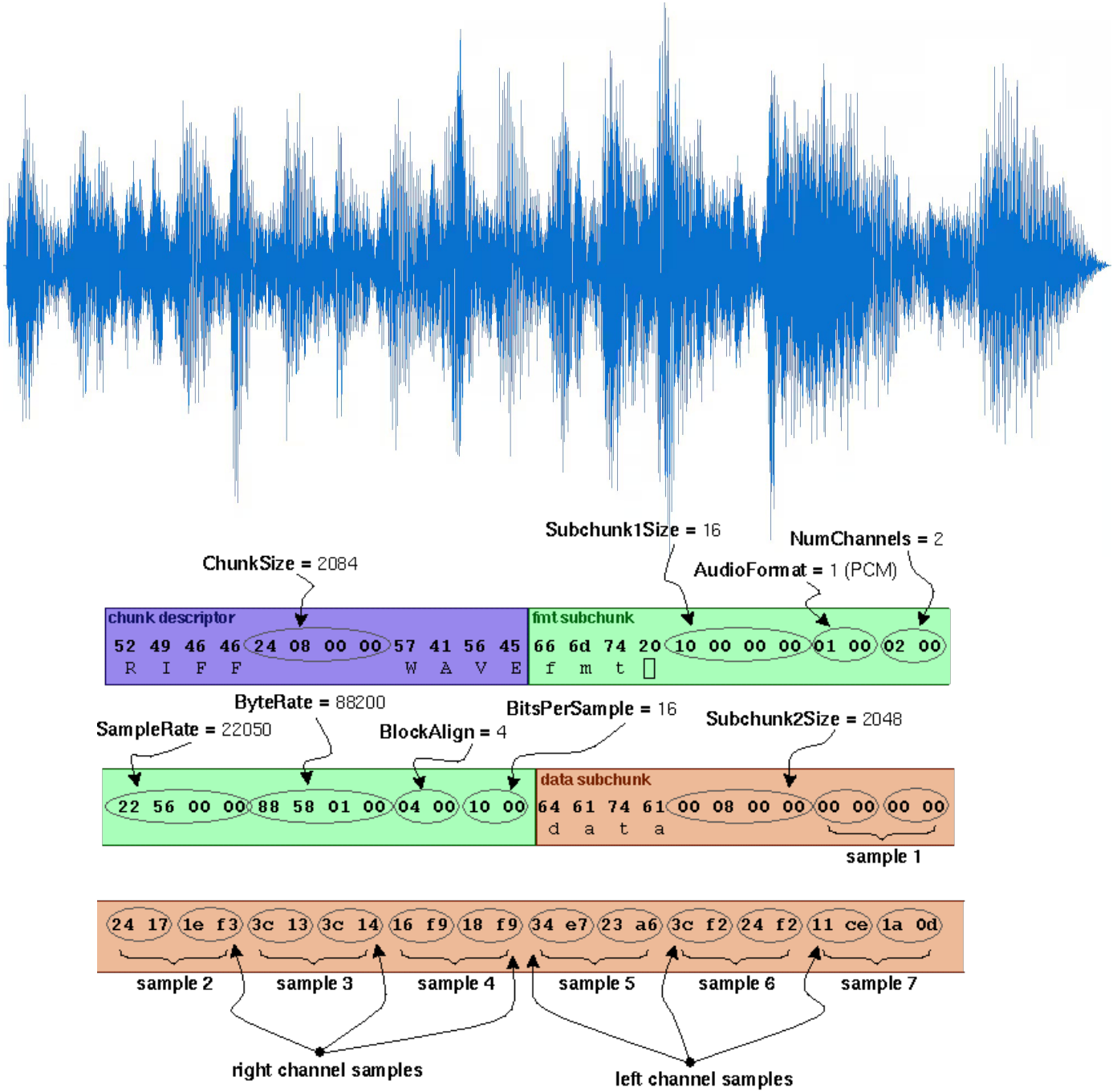
Instituto de Física de Cantabria





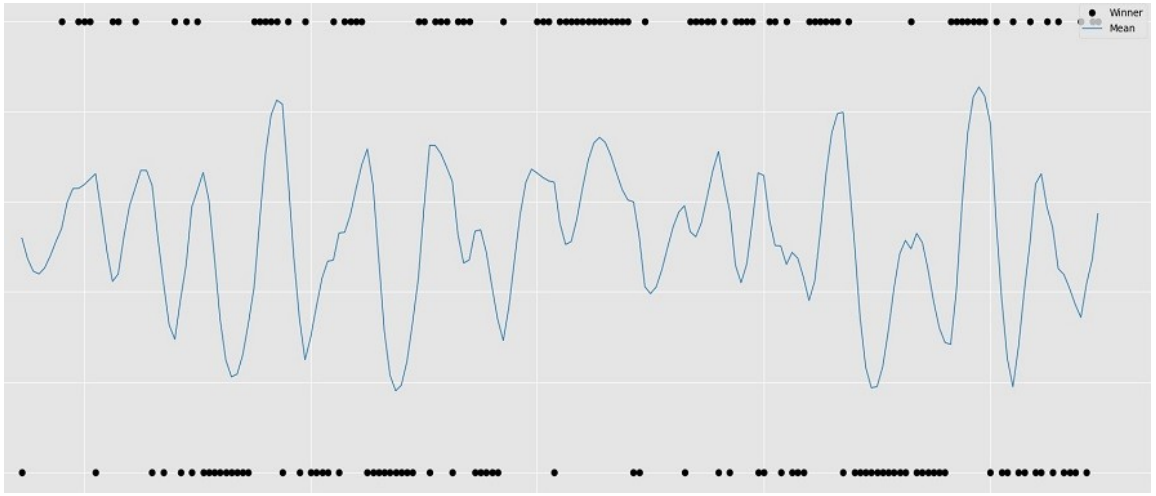
What is information?

Instituto de Física de Cantabria

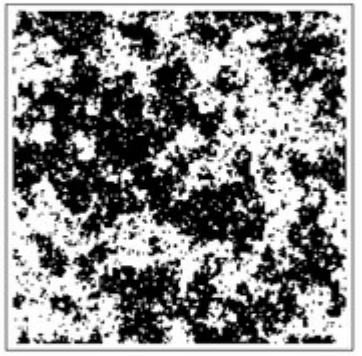


What is information?

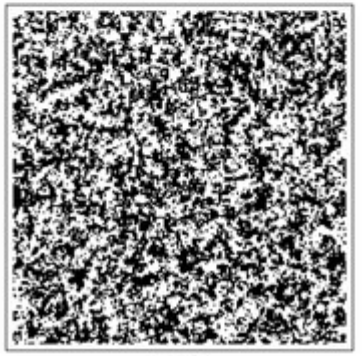
Instituto de Física de Cantabria



$T < T_c$



$T \sim T_c$



$T > T_c$

What is information?

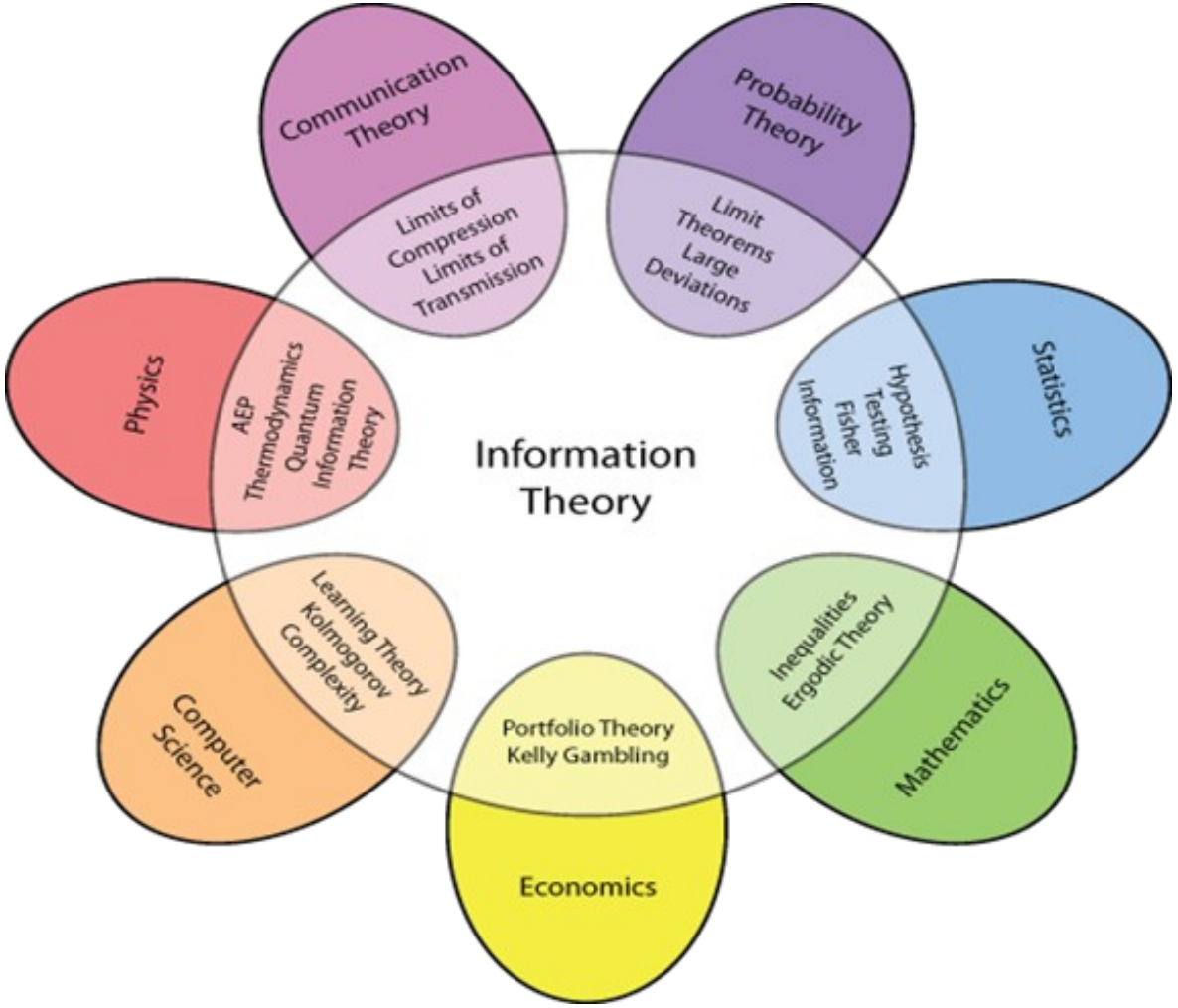
Instituto de Física de Cantabria

What is information theory?

1. The mathematical theory of **communication, compression, complexity, correlation, uncertainty, information processing, ...**
   - What does communication mean? What are the fundamental limits on communication? How can codes for communication be derived and compared? How can data be represented in a compressed way? How much information does one distribution provide about another?
2. The mathematical theory underlying **much of statistics and machine learning**
   - The most general, "right" metrics for quantifying degrees of uncertainty, divergence, and correlation among distributions
3. A nexus among fields...

What is information?

Instituto de Física de Cantabria

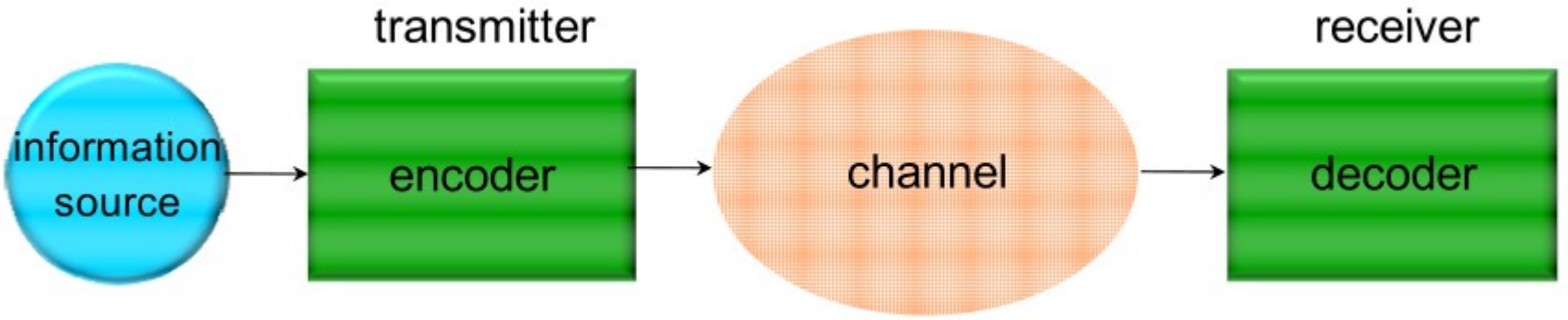


What is information?

Instituto de Física de Cantabria

Shannon theory of information

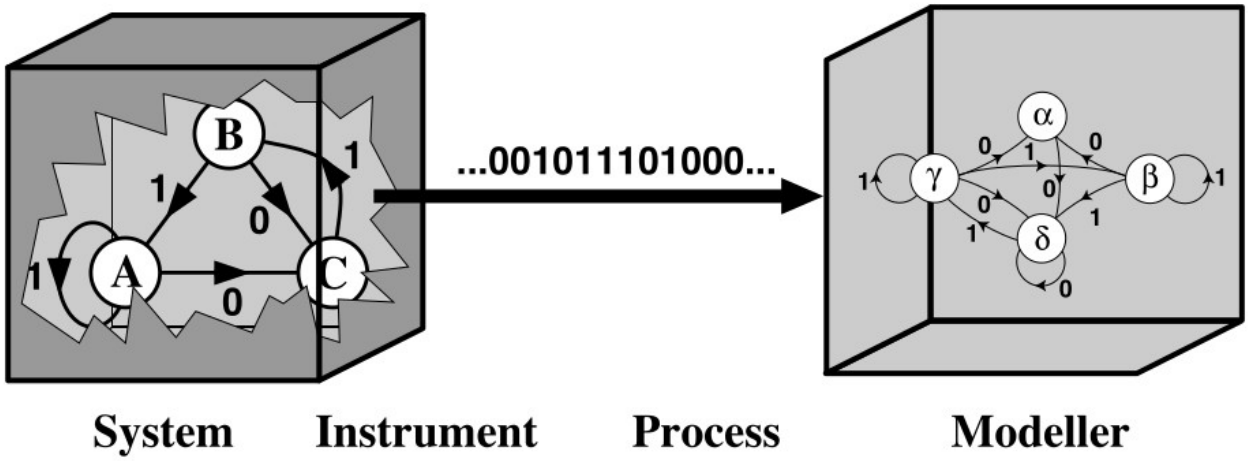
> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."



Claude Shannon, *A Mathematical Theory of Communication* (The Bell System Technical Journal,
Vol. 27, pp. 379–423, 623–656, July, October, 1948)

What is information?

Instituto de Física de Cantabria



The Learning Channel

What is information?

Instituto de Física de Cantabria

How much information?

Alice (A) is in a state of ignorance about a certain variable X which is known to Bob (B). She anticipates that the answer $X \in \mathscr{X}$ can be one of $n = |\mathscr{X}|$ possible ones. $\mathscr{X}$ is called the *alphabet* and X is a *symbol*

One way to quantify the information content of X is to count the number of binary questions (yes/no) that A needs to pose to B in order to know the answer X

Therefore, the number N_Q of binary questions needed to dispel A's ignorance is an operative definition of the information content of X, and it is measured in bits

What is information?

Instituto de Física de Cantabria

Take for example the case $\mathcal{X} = \{a, b, c, d\}$. Then A may ask a first question

Q_1: is $X \in \{a, b\}$ or not?

and depending on the answer, A may ask

Q_2: if $X \in \{a, b\}$ is $X = a$ or not?

Else, if $X \notin \{a, b\}$ is $X = c$ or not?

The answers to these two questions reveal the correct outcome X. Hence the information is $N_Q = 2$ bits. Yet there are many other ways in which A could ask questions, and hence N_Q could vary accordingly.

What is information?

Instituto de Física de Cantabria

Take for example the case $\mathcal{X} = \{a, b, c, d\}$. Then A may ask a first question

Q_1: is $X \in \{a, b\}$ or not?

and depending on the answer, A may ask

Q_2: if $X \in \{a, b\}$ is $X = a$ or not?

Else, if $X \notin \{a, b\}$ is $X = c$ or not?

The answers to these two questions reveal the correct outcome X. Hence the information is $N_Q = 2$ bits. Yet there are many other ways in which A could ask questions, and hence N_Q could vary accordingly.

What is information?

Instituto de Física de Cantabria

Q_1': is $X = a$ or not?

only if $X \neq a$ A will need to pose a further question. Then she may ask:

Q_2': is $X = b$ or not?

Only if the result is no, she will need to ask

Q_3': is $X = c$ or not?

in which case the number of binary questions can be $N_Q(a) = 1, N_Q(b) = 2$ or $N_Q(c) = N_Q(d) = 3$, depending on the value of X. Indeed, $N_Q(X)$ is a random variable, because it is a function of X.

Therefore, it makes sense to define a measure of information content as the expected number of binary questions that are needed to elicit the value of *X*.

$$\mathbb{E}[N_Q] = \sum_{x \in \chi} p_x N_Q(x)$$

What is information?

Instituto de Física de Cantabria

Blackboard calculations ….

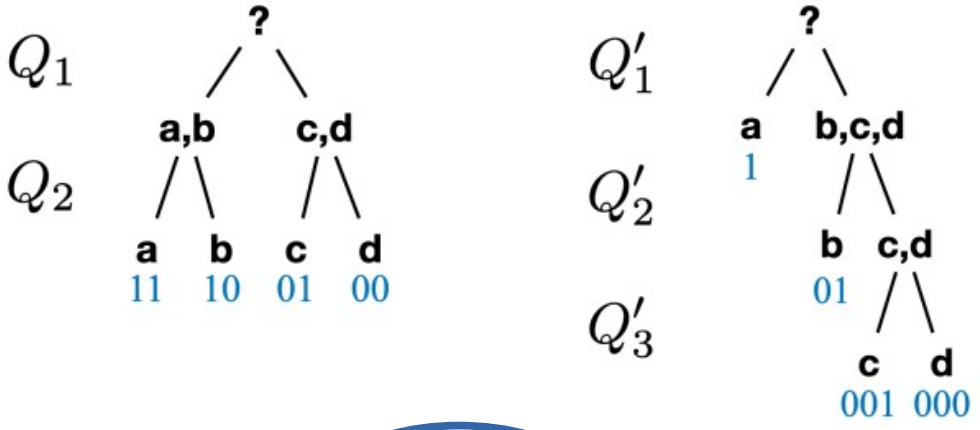
What is information?

Instituto de Física de Cantabria

The optimal way of answering questions is different in the two cases. The minimal expected number of binary questions that A needs to pose to elicit X is a measure of her irreducible ignorance about X. Hence, we provisionally define

> The information content $H[X]$ of a random variable X is the *minimal* expected number of binary questions needed to elicit the value of X,
>
> $$H[X] = \min_{Q} \mathbb{E}\left[N_Q\right] \tag{16.2}$$
>
> where the expected value is taken with respect to the distribution $P\{X = x\} = p_x$ that defines the state of knowledge on X, and the minimum is taken over all possible ways of posing yes/no questions.

What is information?



$a \rightarrow 11$
$b \rightarrow 10$
$c \rightarrow 01$
$d \rightarrow 00$

Coding

$a \rightarrow 1$
$b \rightarrow 01$
$c \rightarrow 001$
$d \rightarrow 000$

Notice that each codeword has length $\ell_Q(X) = N_Q(X)$ which is equal to the number of binary questions needed to elicit X under protocol Q.

Therefore the problem of finding the code that is *expected* to use the least number of bits (i.e. that minimises $\mathbb{E}[\ell_Q]$) is exactly the same as the problem of finding the best way to pose questions. The fact that these two apparently different problems — A posing questions to B optimally and B transmitting answers to A efficiently — have the same solution, is interesting.

What is information?

Instituto de Física de Cantabria

The minimal number of binary questions needed to elicit X, or equivalently the expected length of the optimal code for X, is given by the *Shannon entropy*

$$H[X] = \mathbb{E}\left[\log_2 1/p_X\right] = -\sum_{x \in \mathcal{X}} p_x \log_2 p_x$$

What is information?

Instituto de Física de Cantabria

Asymptotic Equipartition Property

Let $\underline{X} = (X_1, \ldots, X_n)$ *be independent draws from a discrete distribution* $p(x)$ $(X_i \in \chi$ *with* $|\chi| < +\infty)$. *Then*

$$\lim_{n\to\infty} \frac{1}{n} \log p(\underline{X}) = -H[X] \tag{14.4}$$

in probability.

$$H[X] = -\sum_{x\in\chi} p(x) \log p(x)$$

What is information?

Instituto de Física de Cantabria

Asymptotic Equipartition Property

For any $\epsilon > 0$, one can define the set of ϵ-typical sequences as

$$A_n^\epsilon = \left\{ \underline{X} : \left| \frac{1}{n} \log p(\underline{X}) + H[X] \right| < \epsilon \right\}$$

Then an equivalent way to state the AEP is that

1. By definition, all ϵ-typical sequences are equally likely: $P(\underline{X}) \backsim e^{-nH[X]}$ for all $\underline{X} \in A_n^\epsilon$
2. As a consequence of the law of large numbers, a random sequence is almost surely an ϵ-typical sequence

$$P\{A_n^\epsilon\} > 1 - \epsilon \,.$$

3. As a consequence, the number of ϵ-typical sequences is

$$|A_n^\epsilon| \backsim e^{nH[X]} \,.$$

What is information?

Instituto de Física de Cantabria

Asymptotic Equipartition Property

Therefore, the number of typical samples $|A_n^\epsilon| \backsim e^{nH[X]}$ is much smaller than the number of all possible samples, which is $|\chi|^n = e^{n \log |\chi|}$, whenever the distribution differs from the uniform one $p(x) = 1/|\chi|$, for which one has $H[X] = \log |\chi|$. To put it differently, the probability that any sequence $\underline{X}$ is typical is exponentially small in n.

What is information?

Instituto de Física de Cantabria

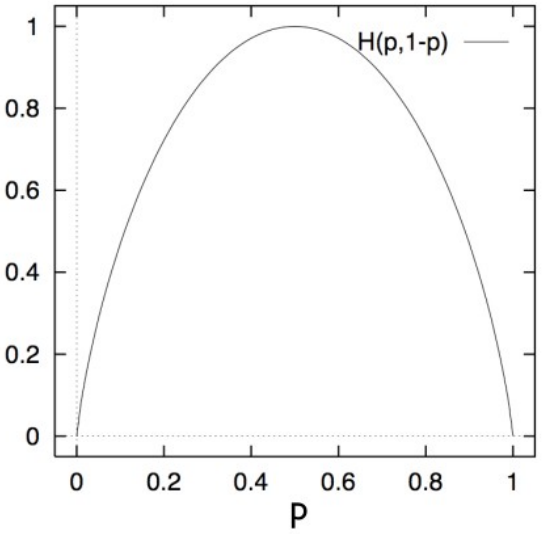
Example: Binary random variable X (Biased Coin)

$$\mathcal{X} = \{0, 1\} \qquad \Pr(1) = p \;\&\; \Pr(0) = 1 - p$$

$H(X)$?

Binary entropy function:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



Fair coin: $p = \frac{1}{2}$

$$H(p) = 1 \text{ bit}$$

Completely biased coin: $p = 0 \text{ (or } 1)$

$$H(p) = 0 \text{ bits}$$

Recall: $0 \cdot \log 0 = 0$

What is information?

Instituto de Física de Cantabria

Example: IID Process over four events

Entropy: $H(X) = \frac{7}{4}$ bits

At each stage, ask questions that are most informative.

Choose partitions of event space that give "most random" measurements.

Theorem:
Entropy gives the smallest number of questions to identify an event, on average.

What is information?

Instituto de Física de Cantabria

Interpretations of Shannon Entropy:

- Observer's *degree of surprise* in outcome of a random variable
- Uncertainty *in* random variable
- Information required to *describe* random variable
- A measure of *flatness* of a distribution

What is information?

Instituto de Física de Cantabria

Two random variables: $(X, Y) \sim p(x, y)$

Joint Entropy: Average uncertainty in X and Y occurring

$$H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 p(x,y)$$

Independent:

$$X \perp Y \Rightarrow H(X,Y) = H(X) + H(Y)$$

What is information?

Instituto de Física de Cantabria

Conditional Entropy: Average uncertainty in X, knowing Y

$$H(X|Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x|y)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

Not symmetric: $H(X|Y) \neq H(Y|X)$

What is information?

Instituto de Física de Cantabria

Common Information Between Two Random Variables:

$$X \sim p(x) \ \& \ Y \sim p(y)$$

$$(X, Y) \sim p(x, y)$$

Mutual Information:

$$I(X;Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

What is information?

Instituto de Física de Cantabria

Properties:

(1) $I(X;Y) \geq 0$

(2) $I(X;Y) = I(Y;X)$

(3) $I(X;Y) = H(X) - H(X|Y)$

(4) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(5) $I(X;X) = H(X)$

(6) $X \perp Y \Rightarrow I(X;Y) = 0$

Interpretations:

Information one variable has about another

Information shared between two variables

Measure of dependence between two variables

What is information?

Instituto de Física de Cantabria

Properties:

(1) $I(X;Y) \geq 0$

(2) $I(X;Y) = I(Y;X)$

(3) $I(X;Y) = H(X) - H(X|Y)$

(4) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(5) $I(X;X) = H(X)$

(6) $X \perp Y \Rightarrow I(X;Y) = 0$

Interpretations:

Information one variable has about another

Information shared between two variables

Measure of dependence between two variables

What is information?

Instituto de Física de Cantabria

Why is mutual information better than correlation?

https://docs.google.com/file/d/1bx6uPNuFkxVc8Fc1C5hVNB3hZoaoq1fw/preview

What is information?

Instituto de Física de Cantabria

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x. How much this impacts on the efficiency of the questions she's going to ask?

What is information?

Instituto de Física de Cantabria

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x. How much this impacts on the efficiency of the questions she's going to ask?

Given q, A is going to effectively encode B's answers in such a way that answer x will require $\log_2 1/q_x$ bits, so the number of questions she will ask, on average, is

$$\mathbb{E}\left[\log_2 \frac{1}{q}\right] = \sum_{x \in \mathcal{X}} p_x \log_2 \frac{1}{q_x}$$

What is information?

Instituto de Física de Cantabria

Relative entropy and Kullback-Leibler divergence

Imagine now that A has a wrong estimate q_x of the probability p_x of B's answers x. How much this impacts on the efficiency of the questions she's going to ask?

Given q, A is going to effectively encode B's answers in such a way that answer x will require $\log_2 1/q_x$ bits, so the number of questions she will ask, on average, is

$$\mathbb{E}\left[\log_2 \frac{1}{q}\right] = \sum_{x\in\mathcal{X}} p_x \log_2 \frac{1}{q_x}$$

the difference between this and the most efficient way of asking questions, which requires $\mathcal{H}[p]$ bits, is

$$D_{KL}[p\|q] = \sum_{x\in\mathcal{X}} p_x \log_2 \frac{p_x}{q_x}$$

which is known as the *Kullback-Leibler divergence* or *relative entropy*

What is information?

Instituto de Física de Cantabria

Data processing system

Nature

Camera

CD

What is information?

Instituto de Física de Cantabria

Basic Result: Data Processing Inequality

- Let *X* and *Y* be random variables and let *g* be any function. We have:

$$\mathrm{MI}(g(X); Y) \leq \mathrm{MI}(X; Y)$$

What??????????????????????

What is information?

Instituto de Física de Cantabria

Basic Result: Data Processing Inequality

- Let *X* and *Y* be random variables and let *g* be any function. We have:

$$\mathrm{MI}(g(X); Y) \leq \mathrm{MI}(X; Y)$$

- What does this tell us about the meaning of "information"?
- MI quantifies information in *X* about *Y*, regardless of whether that information is "usable" or "accessible" to some particular agent.
  - Coming up with new MI measures which account for accessibility of information is a hot topic in information theory!

What is information?

Instituto de Física de Cantabria

Some facts about mutual information

1. Accounts for any type of co-relation
   - Statistical correlation ~ linear only
   - Information measures nonlinear correlation
2. Broadly applicable:
   - Many systems don't have "energy", physical modeling precluded
   - Information defined: social, biological, engineering, ... systems
3. Comparable units across different systems:
   - Correlation: Meters v. volts v. dollars v. ergs v. ...
   - Information: bits.
4. Probability theory ~ Statistics ~ Information
5. Complex systems:
   - Emergent patterns!
   - We don't know these ahead of time

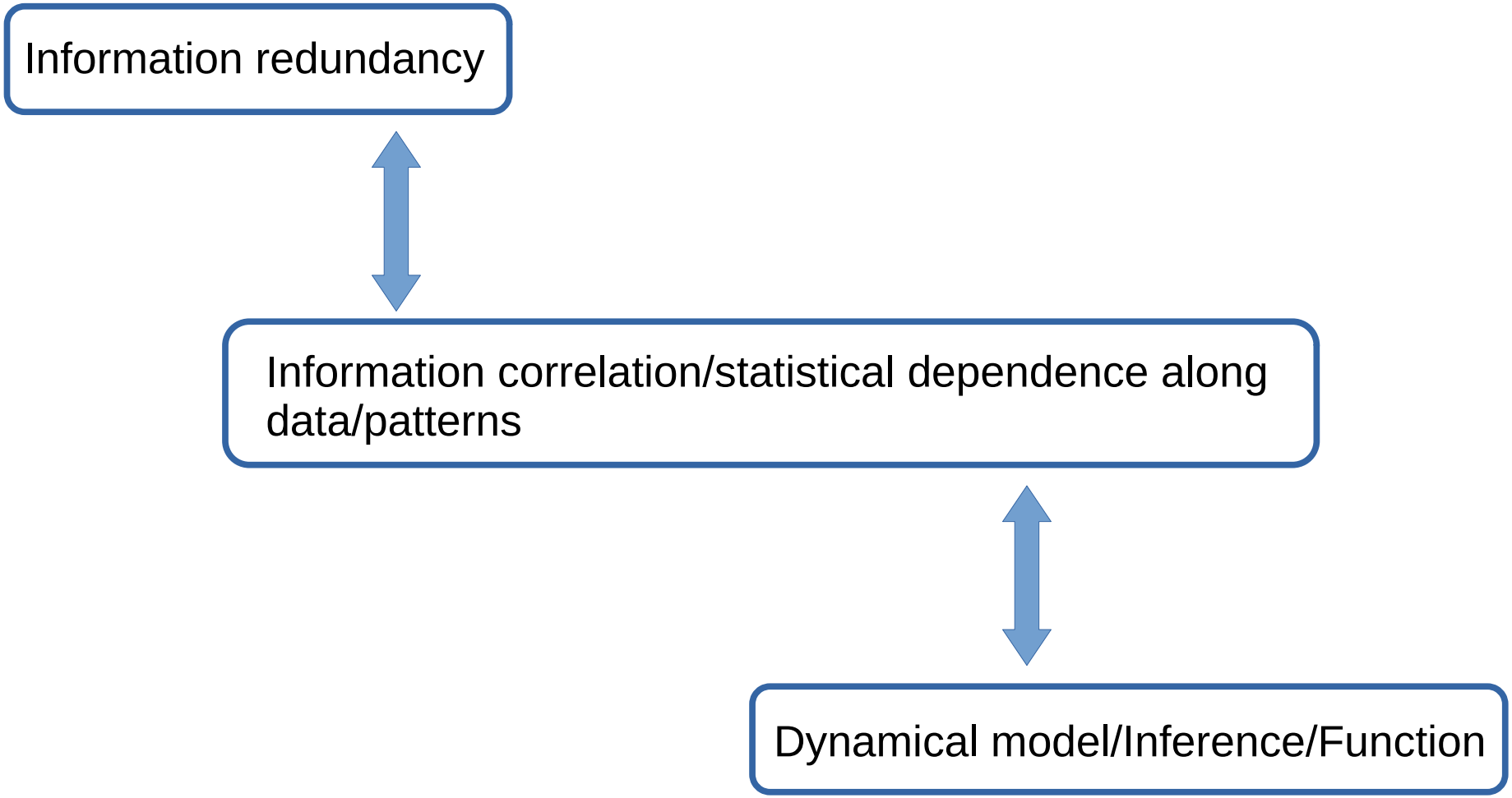
Information redundancy/patterns

Instituto de Física de Cantabria

```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**,**h**h** **l**n**i**l**s**r**o**a**o**e**i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W****o****s****i****l****a****g****n****t****a****e****v***
```

Information redundancy/patterns

Instituto de Física de Cantabria



Information redundancy/patterns

Instituto de Física de Cantabria

```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**,**h**h** **l**n**i**l**s**r**o**a**o**e**i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W****o****s****i****l****a****g****n****t****a****e****v***
```

Information redundancy/patterns

Instituto de Física de Cantabria

```
Emma Woodh*use, hands*me, clever* and rich,*with a
comfortab*e home an* happy di*position,*seemed to*unite som*
of the b*st bless*ngs of e*istence;*and had *ived nea*ly
twenty *ne year* in the*world w*th very*little *o distr*ss
or vex*her.  *he was*the yo*ngest *f the *wo dau*hters *f a
most *ffect*onate* indu*gent *ather* and *ad, i* cons*quenc*
of h*r si*ter'* mar*iage* bee* mis*ress*of h*s ho*se f*om a
ver* ea*ly *eri*d. *er *oth*r h*d d*ed *oo *ong*ago*for*her
to*ha*e *or* t*an*an*in*is*in*t *em*mb*an*e *f *er*ca*es*es*
a*d*h*r*p*a*e*h*d*b*e* *u*p*i*d*b* *n*e*c*l*e*t*w*m*n*a*
g**e**e**,**h**h** **l**n**i**l**s**r**o**a**o**e**i*
a***c***n***S***e***y***s***d***s***a***r***e***n***
W****o****s****i****l****a****g****n****t****a****e****v***
```

```
Emma Woodhouse, handsome, clever, and rich, with a
comfortable home and happy disposition, seemed to unite some
of the best blessings of existence; and had lived nearly
twenty one years in the world with very little to distress
or vex her.  She was the youngest of the two daughters of a
most affectionate, indulgent father; and had, in consequence
of her sister's marriage, been mistress of his house from a
very early period. Her mother had died too long ago for her
to have more than an indistinct remembrance of her caresses;
and her place had been supplied by an excellent woman as
governess, who had fallen little short of a mother in
affection.  Sixteen years had Miss Taylor been in Mr
Woodhouse's family, less as a governess than a friend, very
```

Example: Letter Frequencies

Instituto de Física de Cantabria

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

Book by David MacKay

Example: Letter Frequencies

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

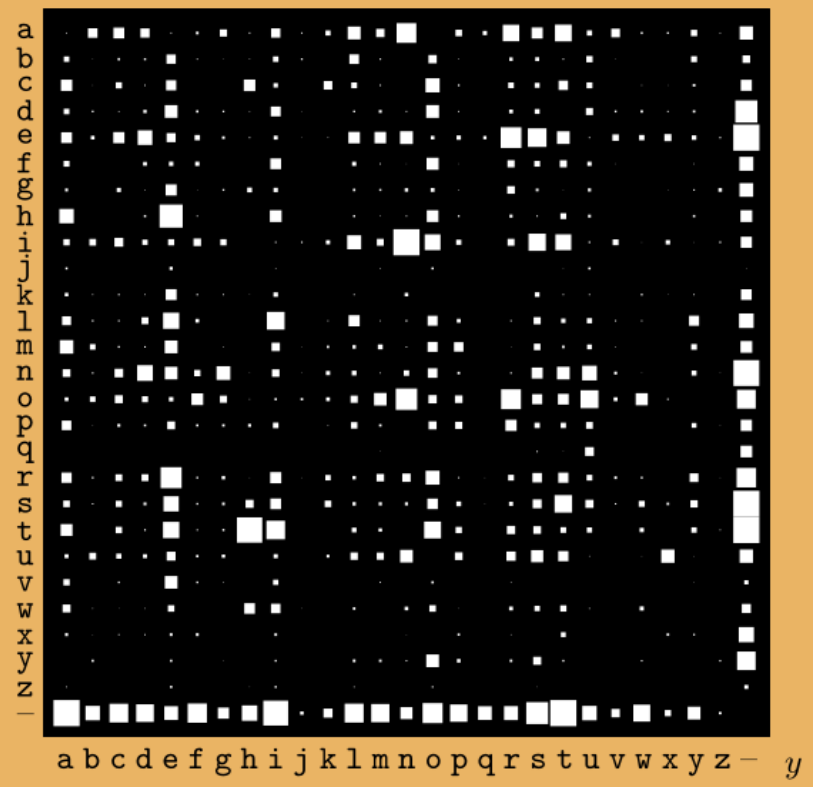


Figure 2.2. The probability distribution over the 27×27 possible bigrams xy in an English language document, *The Frequently Asked Questions Manual for Linux.*

Book by David MacKay

Information in processes

Instituto de Física de Cantabria

Block Entropy:

$$H(L) = H(\Pr(s^L)) = -\sum_{s^L \in \mathcal{A}} \Pr(s^L) \log_2 \Pr(s^L)$$

Monotonic increasing: $H(L) \geq H(L-1)$

Adding a random variable cannot decrease entropy:

$$H(S_1, S_2, \ldots, S_L) \leq H(S_1, S_2, \ldots, S_L, S_{L+1})$$

No measurements, no information: $H(0) = 0$

Information in processes

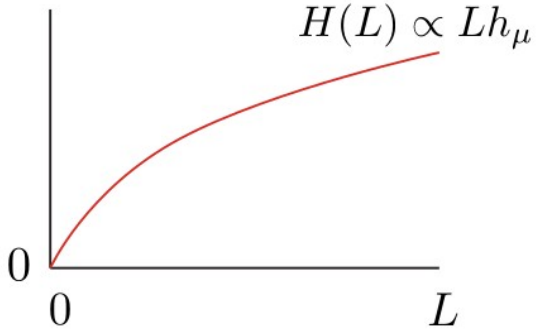
Instituto de Física de Cantabria

Entropy Rates for Stationary Stochastic Processes:

Entropy per symbol is given by the Source Entropy Rate:

$$h_\mu = \lim_{L\to\infty} \frac{H(L)}{L}$$

(When limits exists.)



Interpretations:

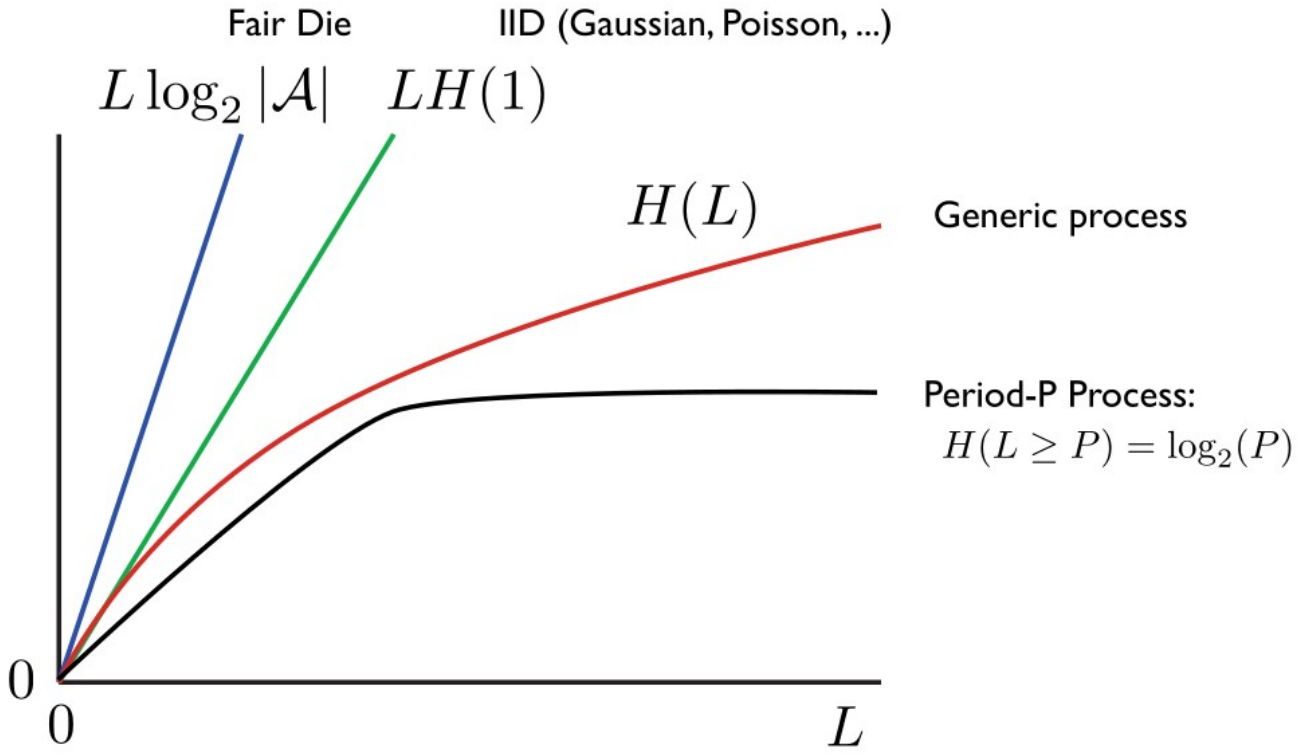
- Asymptotic growth rate of entropy
- Irreducible randomness of process
- Average description length (per symbol) of process

Information in processes

Instituto de Física de Cantabria

Entropy Growth for Stationary Stochastic Processes ...

Block Entropy ...



Information in processes

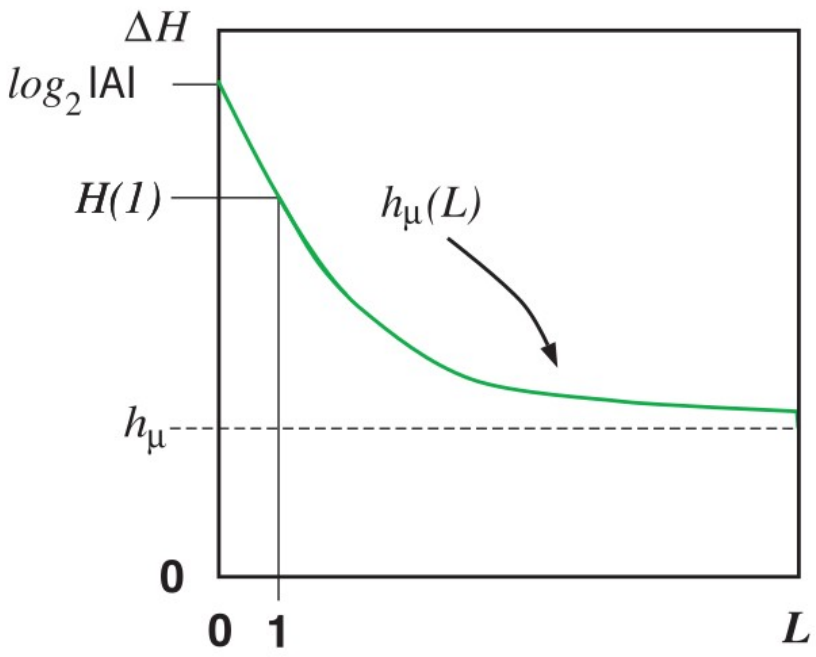
Instituto de Física de Cantabria

Entropy Convergence:

Length-L entropy rate estimate:

$$h_\mu(L) = H(L) - H(L-1)$$

$$h_\mu(L) = \Delta H(L)$$

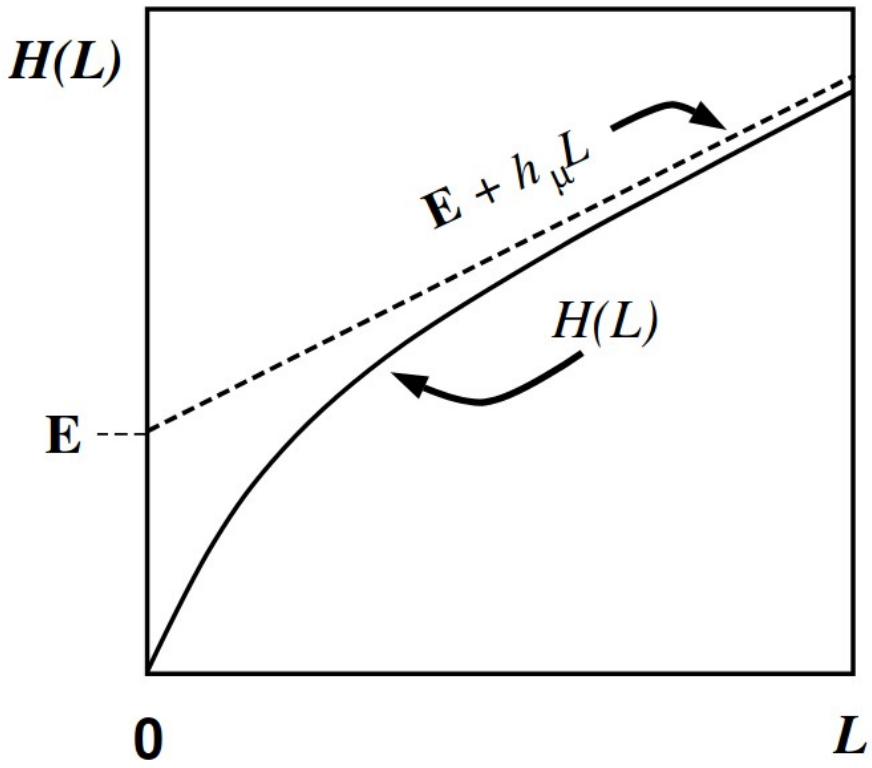


Monotonic decreasing:

$$h_\mu(L) \leq h_\mu(L-1)$$

Process appears less random
as account for longer correlations

Information in processes



E is the ***excess entropy***: measures the redundancy arising from "correlations"

h_μ is the ***entropy rate***: average entropy per symbol, given the past

Information in processes

Instituto de Física de Cantabria

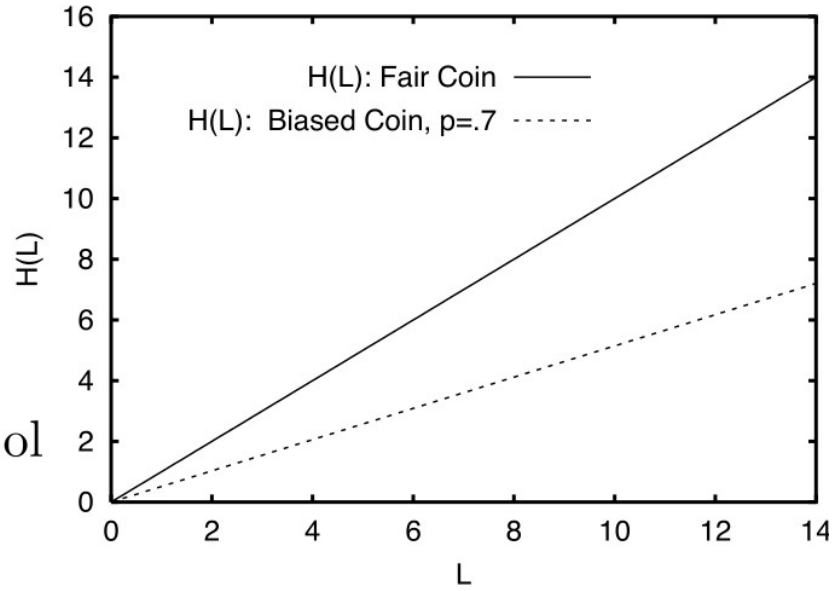
Memory in Processes ...

Examples of Excess Entropy:

Fair Coin:

$h_\mu = 1$ bit per symbol

$\mathbf{E} = 0$ bits



Biased Coin:

$h_\mu = H(p)$ bits per symbol

$\mathbf{E} = 0$ bits

Any IID Process:

$h_\mu = H(X)$ bits per symbol

$\mathbf{E} = 0$ bits

Information for processes

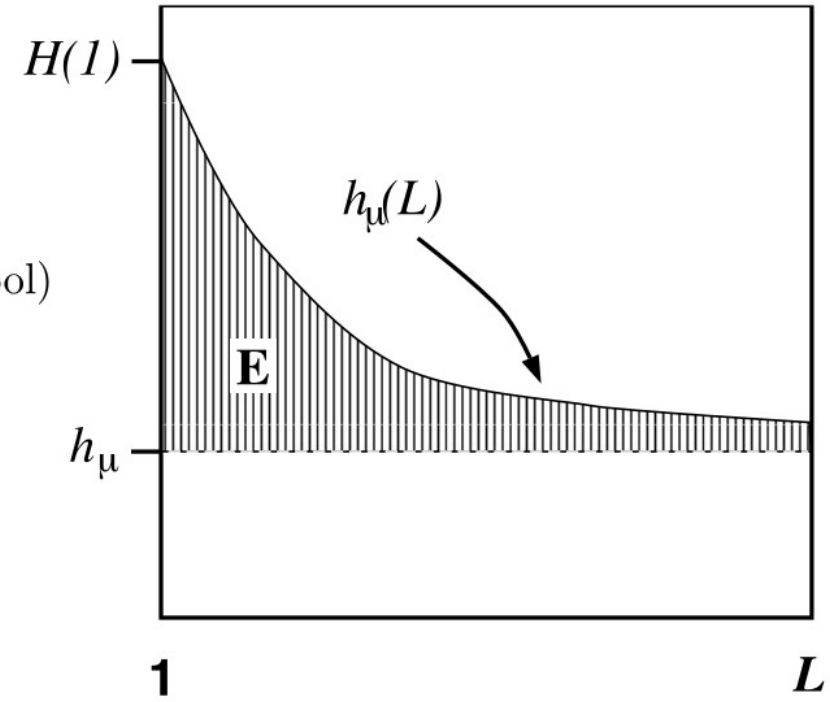
Instituto de Física de Cantabria

Excess Entropy:

As entropy convergence:

$$\mathbf{E} = \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu]$$

$$(\Delta L = 1 \text{ symbol})$$



Properties:

(1) Units: $\mathbf{E} = [\text{bits}]$

(2) Positive: $\mathbf{E} \geq 0$

(3) Controls convergence to actual randomness.

(4) Slow convergence ⇔ Correlations at longer words.

(5) Complementary to entropy rate.

Information in processes

Motivation:

Previous: Measures of randomness of information source

Block entropy $H(L)$

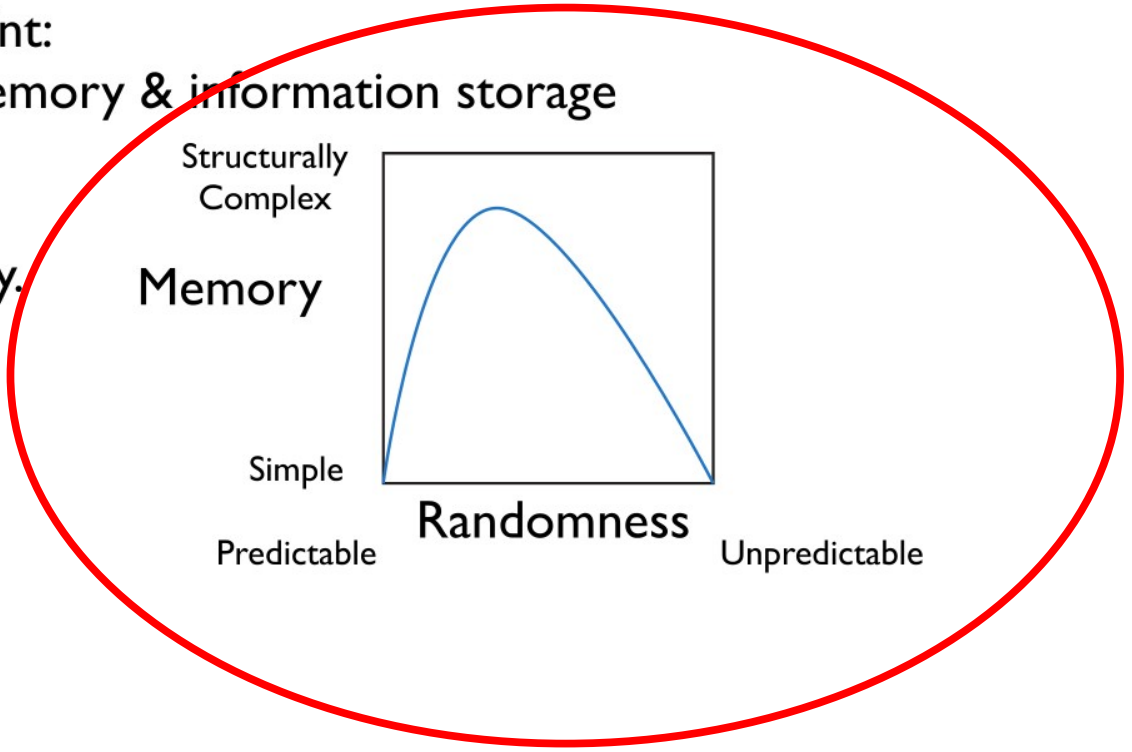
Entropy rate h_μ

Current target point:

Measures of memory & information storage

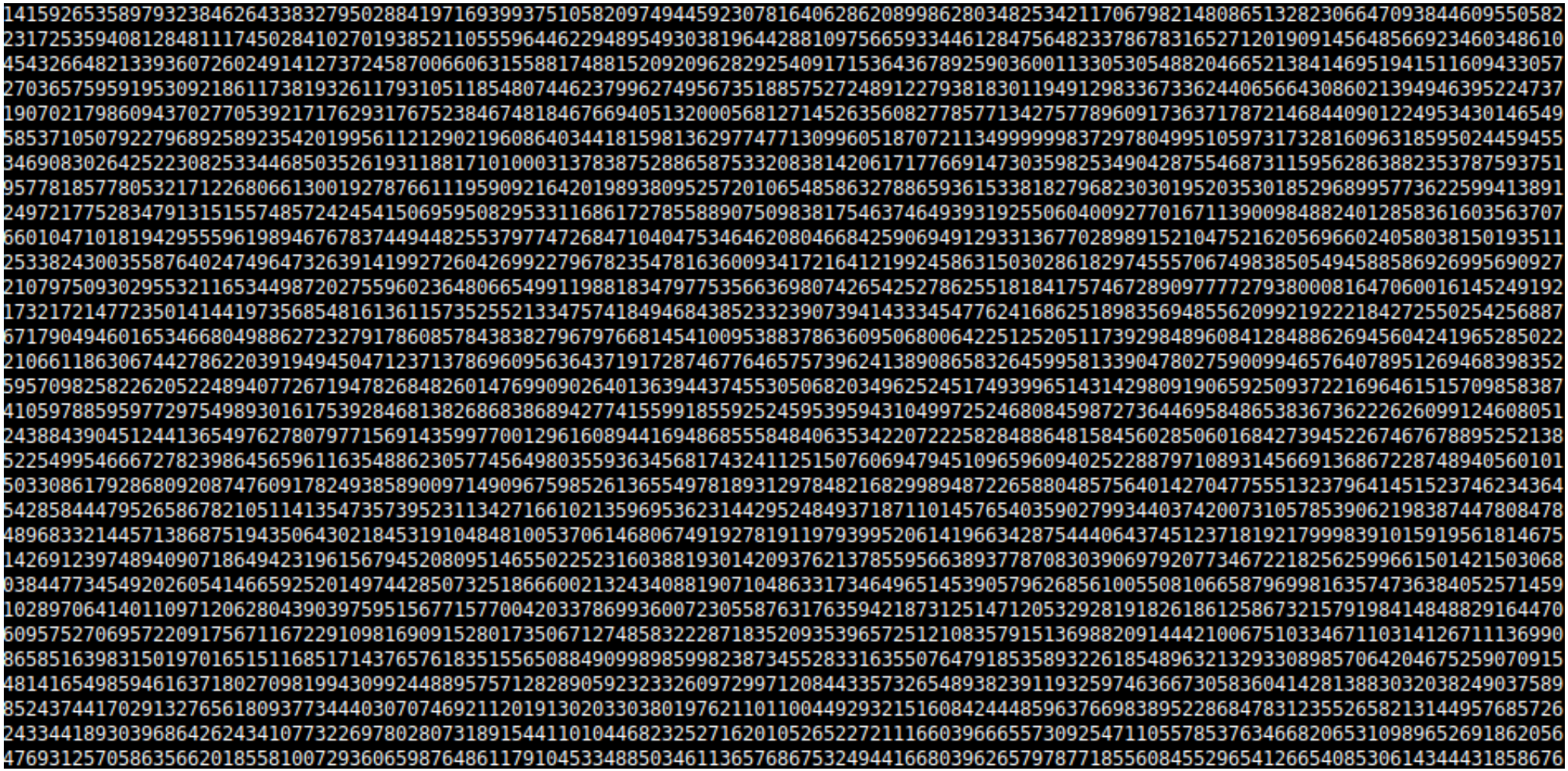
Big Picture:

Complementary.

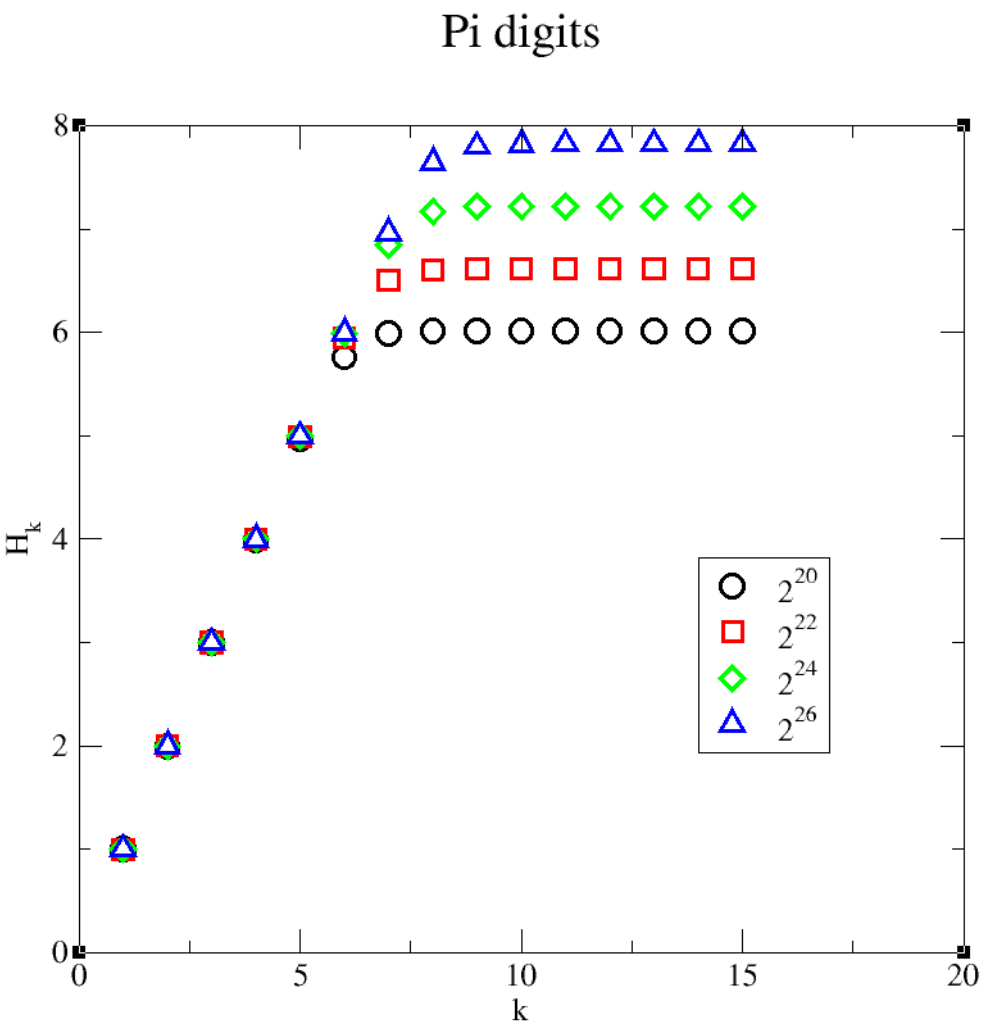


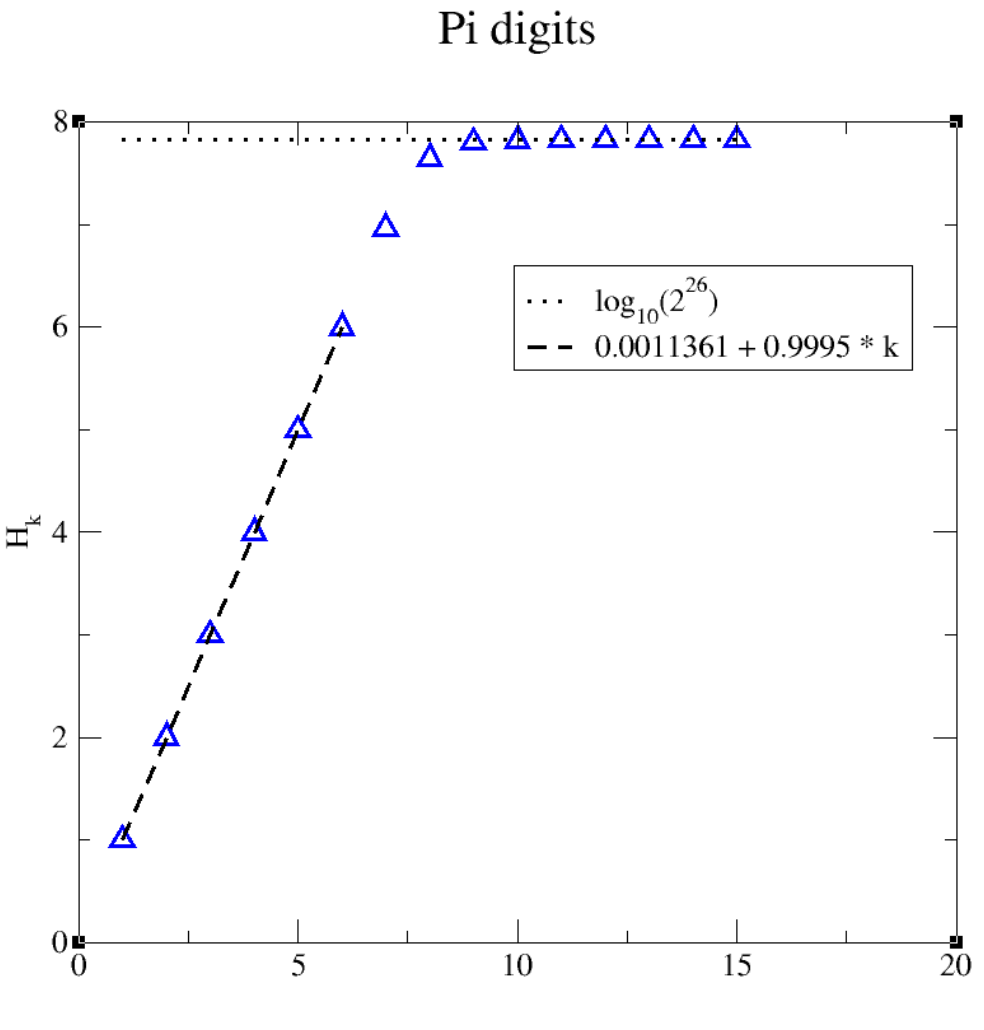
Examples

Instituto de Física de Cantabria



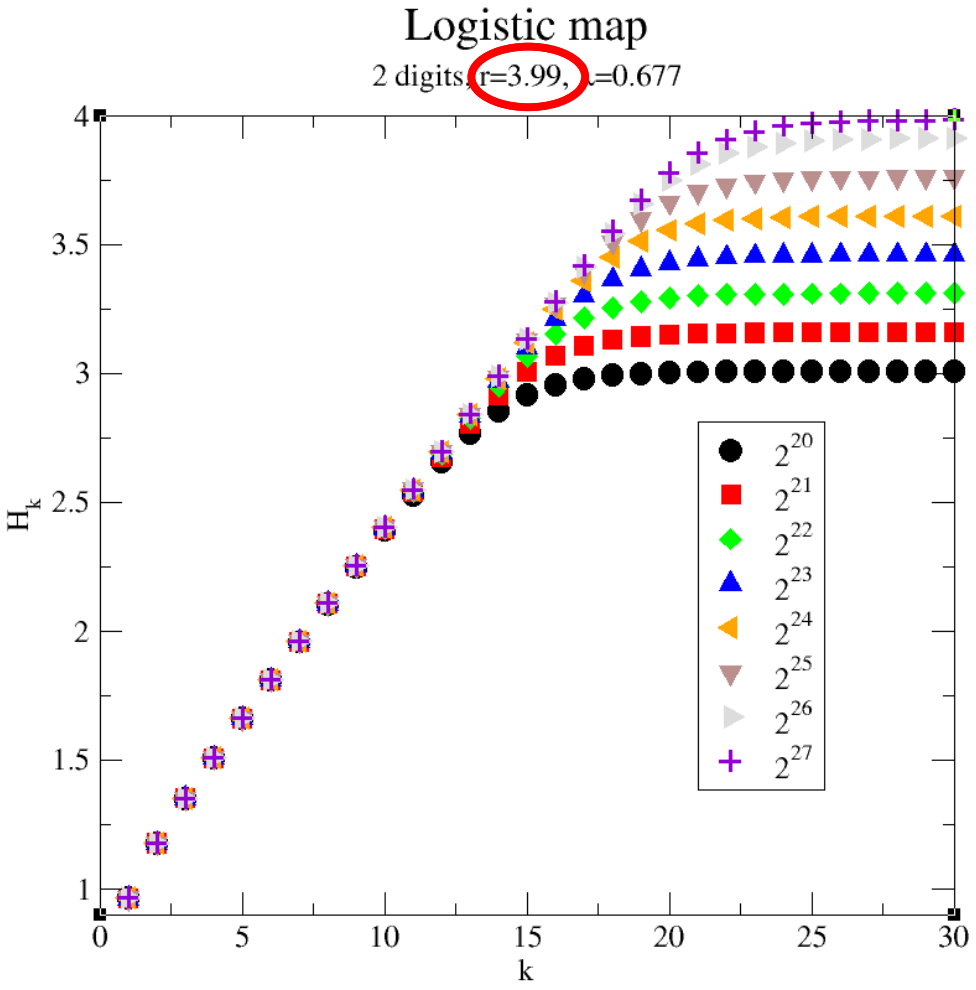
Examples

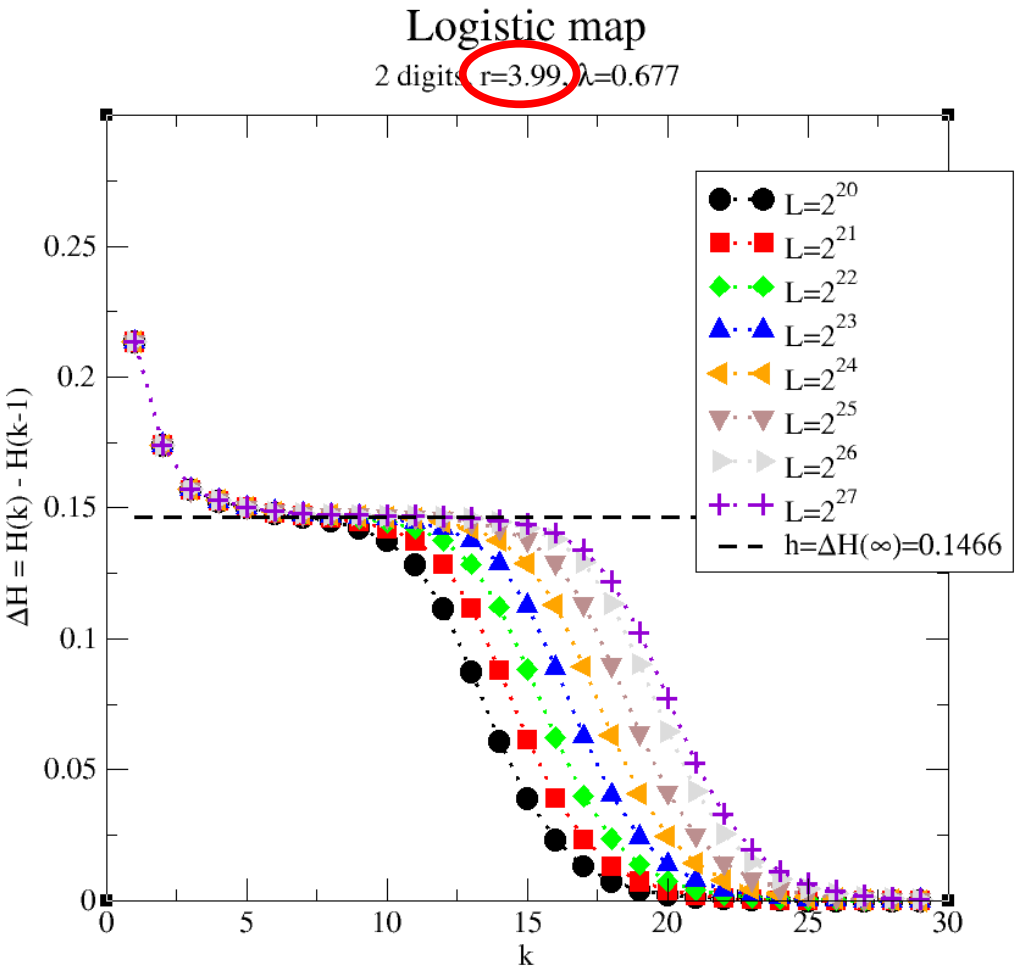




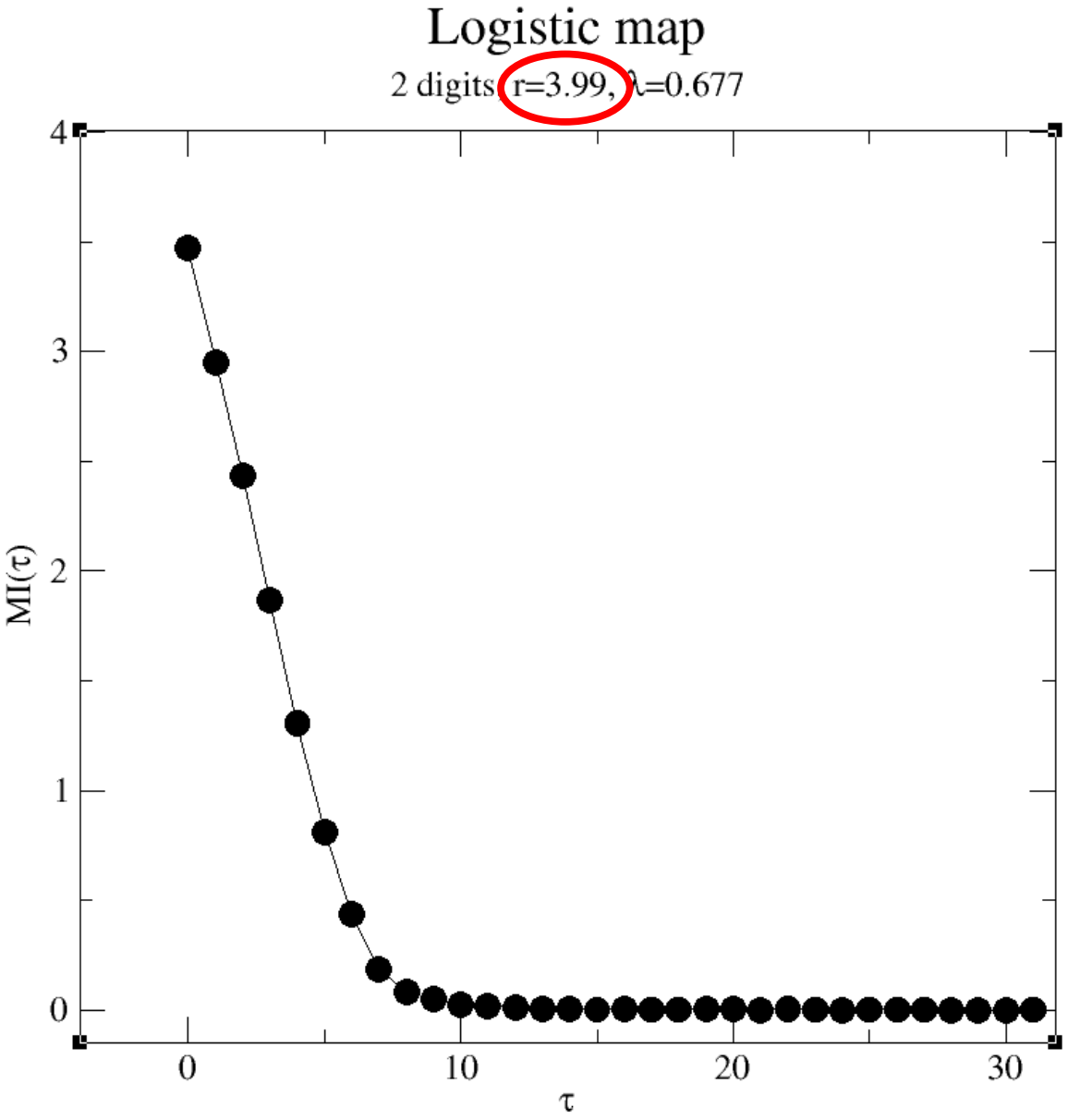
$\log_2 x = \log_{10} x / \log_{10} 2 \approx 3{,}322 \log_{10} x$

Examples


Logistic map
2 digits, r=3.99, λ=0.677
H_k
k
2^20
2^21
2^22
2^23
2^24
2^25
2^26
2^27

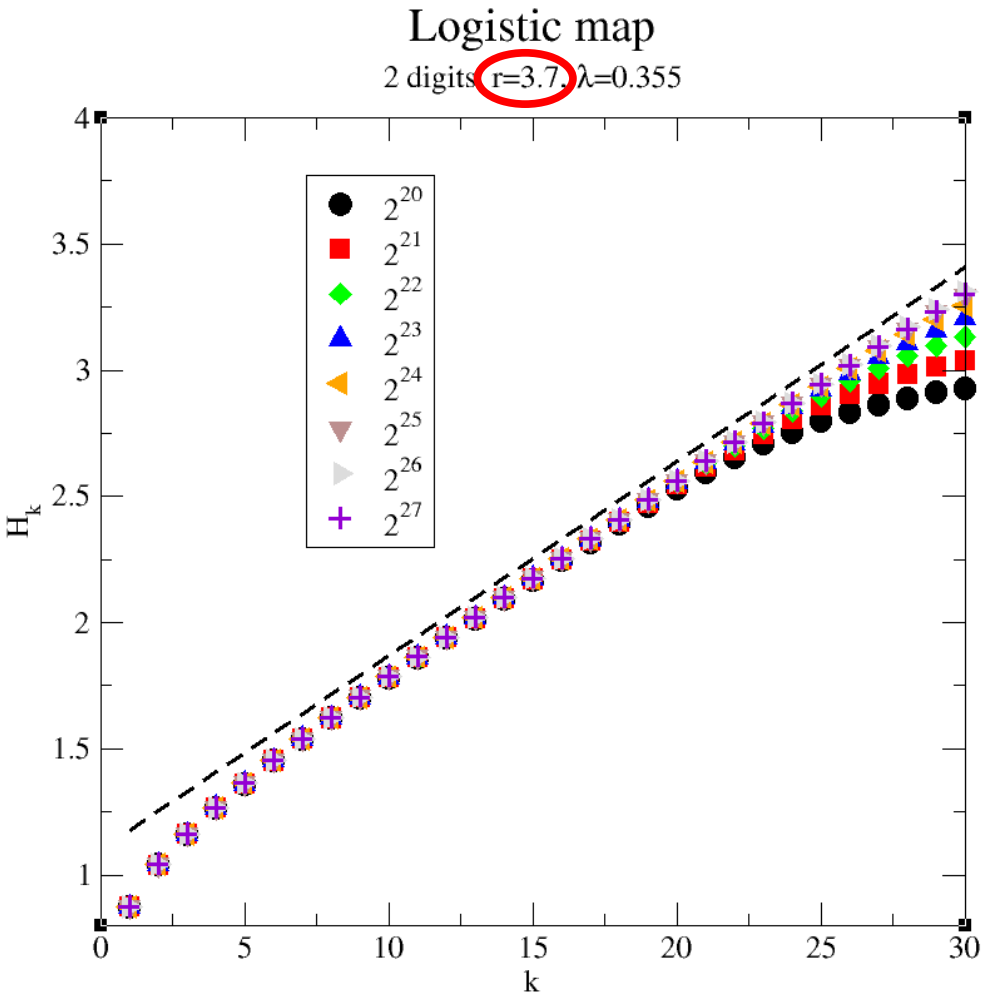

Logistic map
2 digits, r=3.99, λ=0.677
ΔH = H(k) - H(k-1)
k
L=2^20
L=2^21
L=2^22
L=2^23
L=2^24
L=2^25
L=2^26
L=2^27
h=ΔH(∞)=0.1466

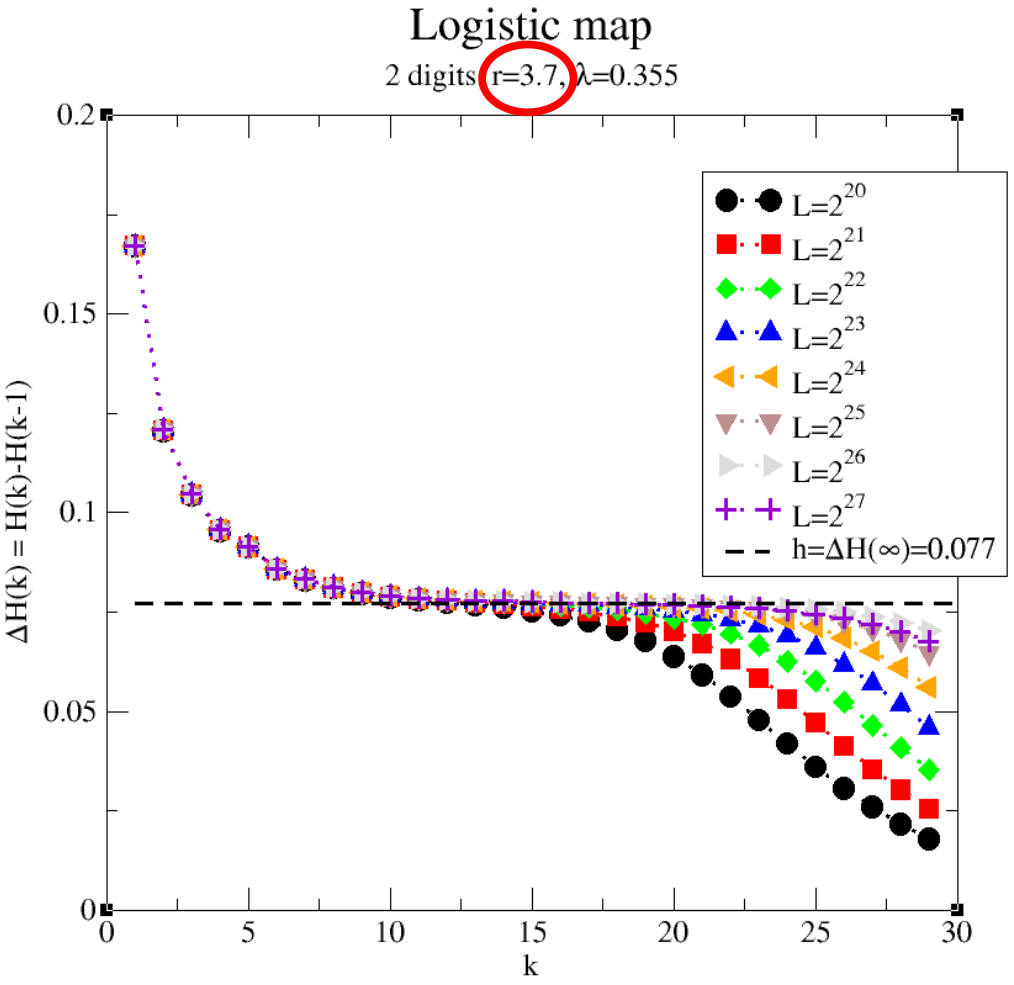
Examples



Examples

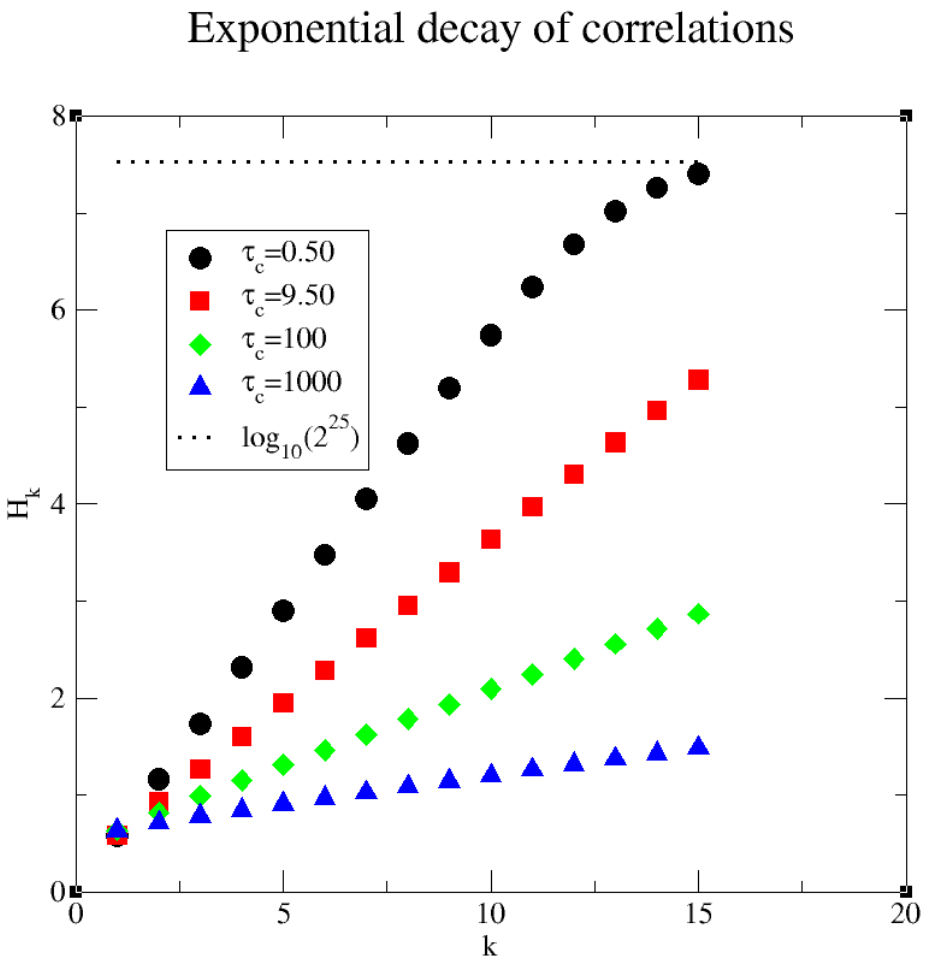
Instituto de Física de Cantabria

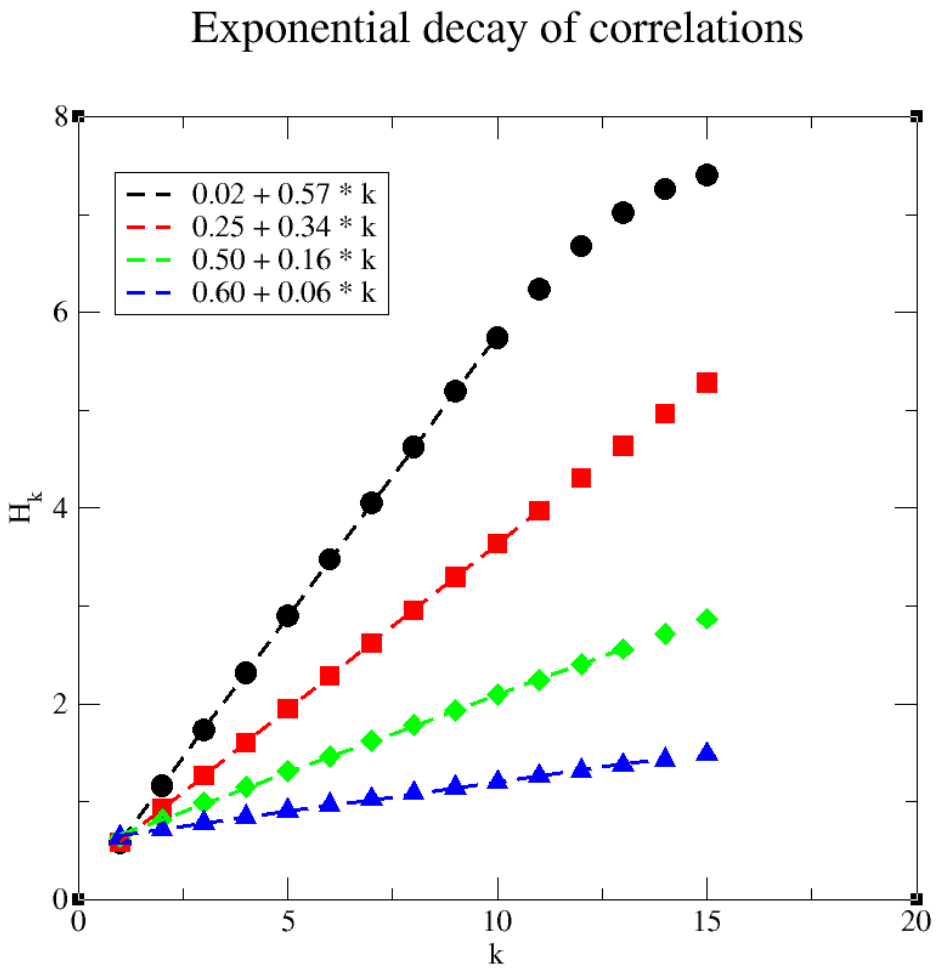




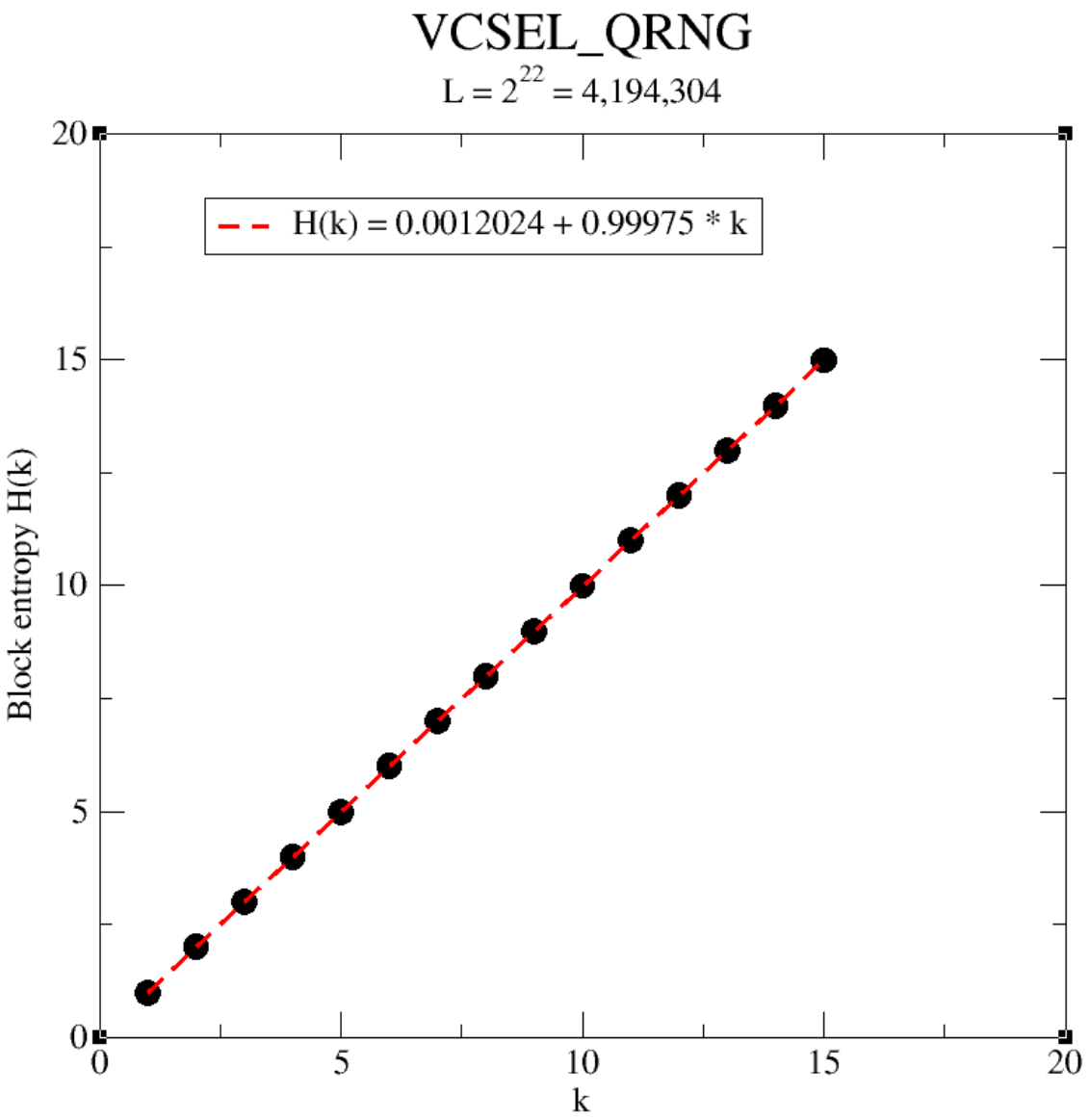
Examples

Instituto de Física de Cantabria

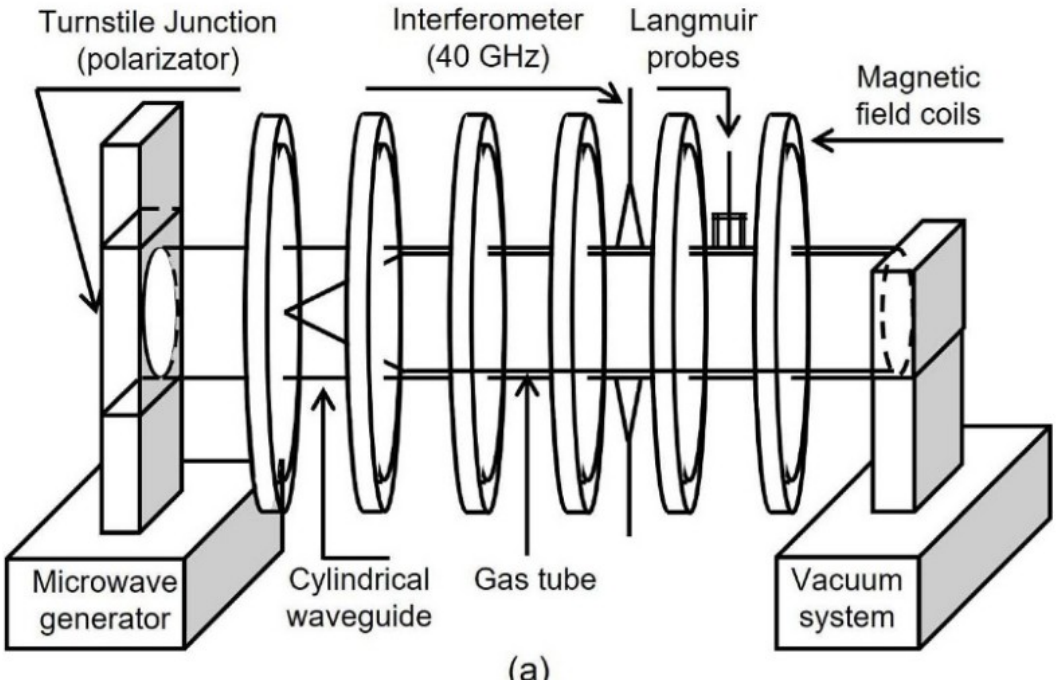


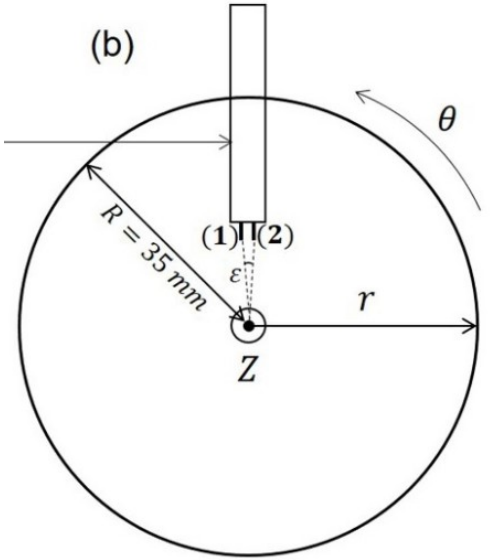


Examples

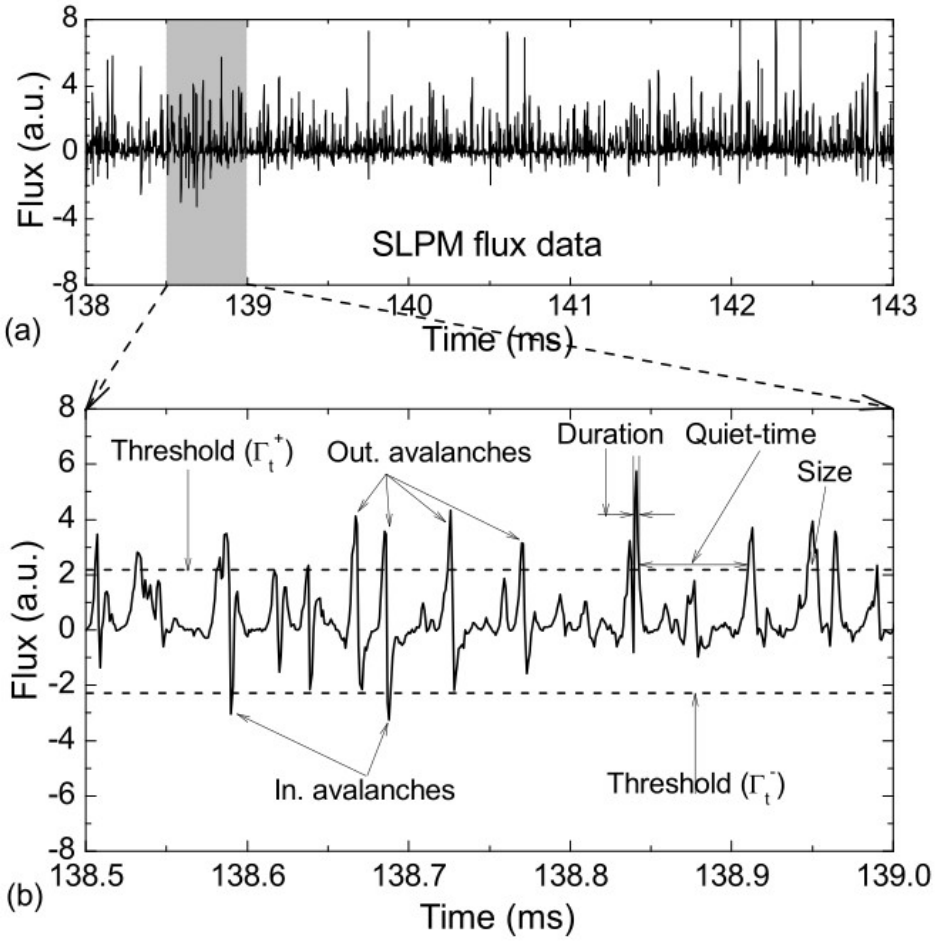

VCSEL_QRNG
L = 2^22 = 4,194,304
H(k) = 0.0012024 + 0.99975 * k
Block entropy H(k)
k
0
5
10
15
20

Examples


Turnstile Junction
(polarizator)
Interferometer
(40 GHz)
Langmuir
probes
Magnetic
field coils
Microwave
generator
Cylindrical
waveguide
Gas tube
Vacuum
system
(a)

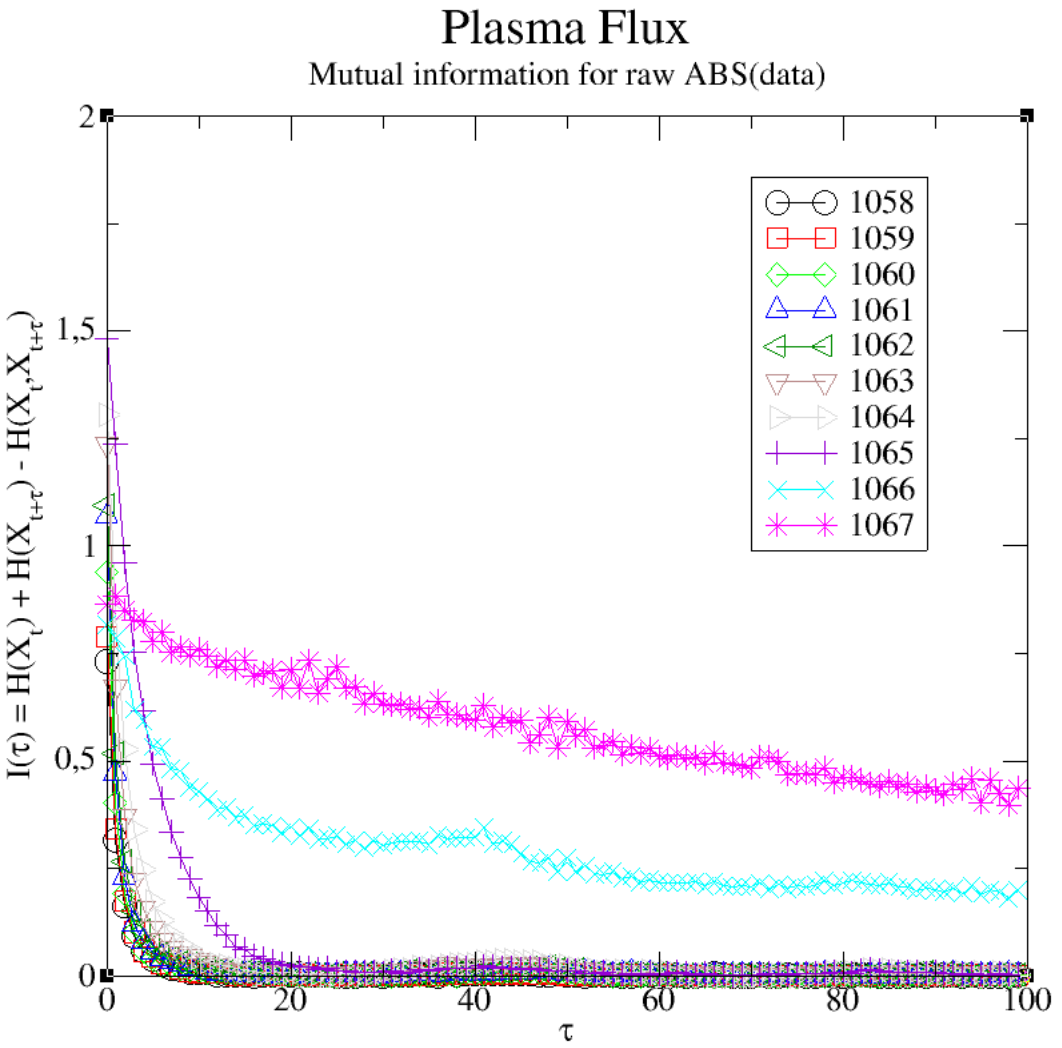

(b)
θ
R = 35 mm
(1)
(2)
ε
r
Z

Examples

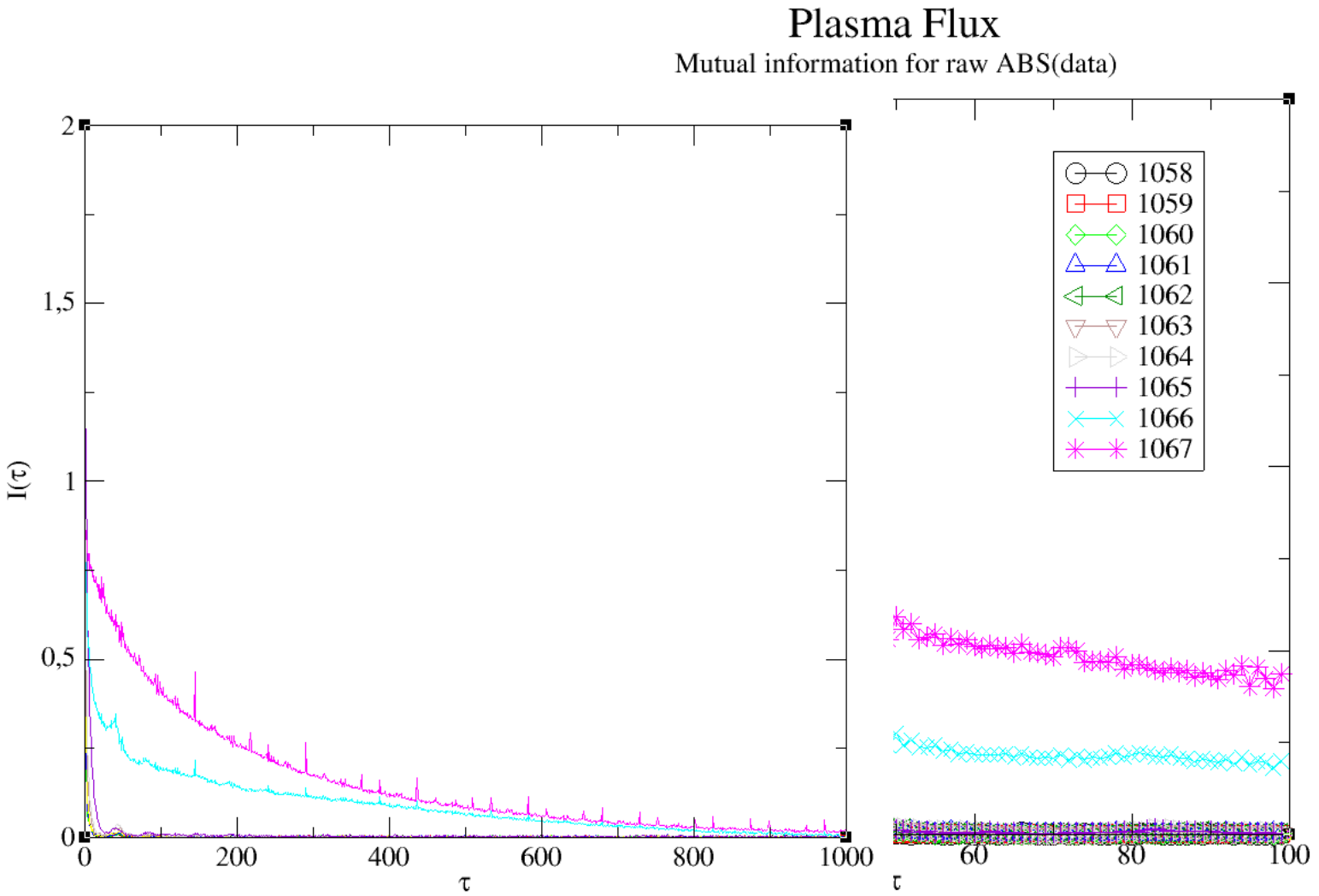


200.000 data points

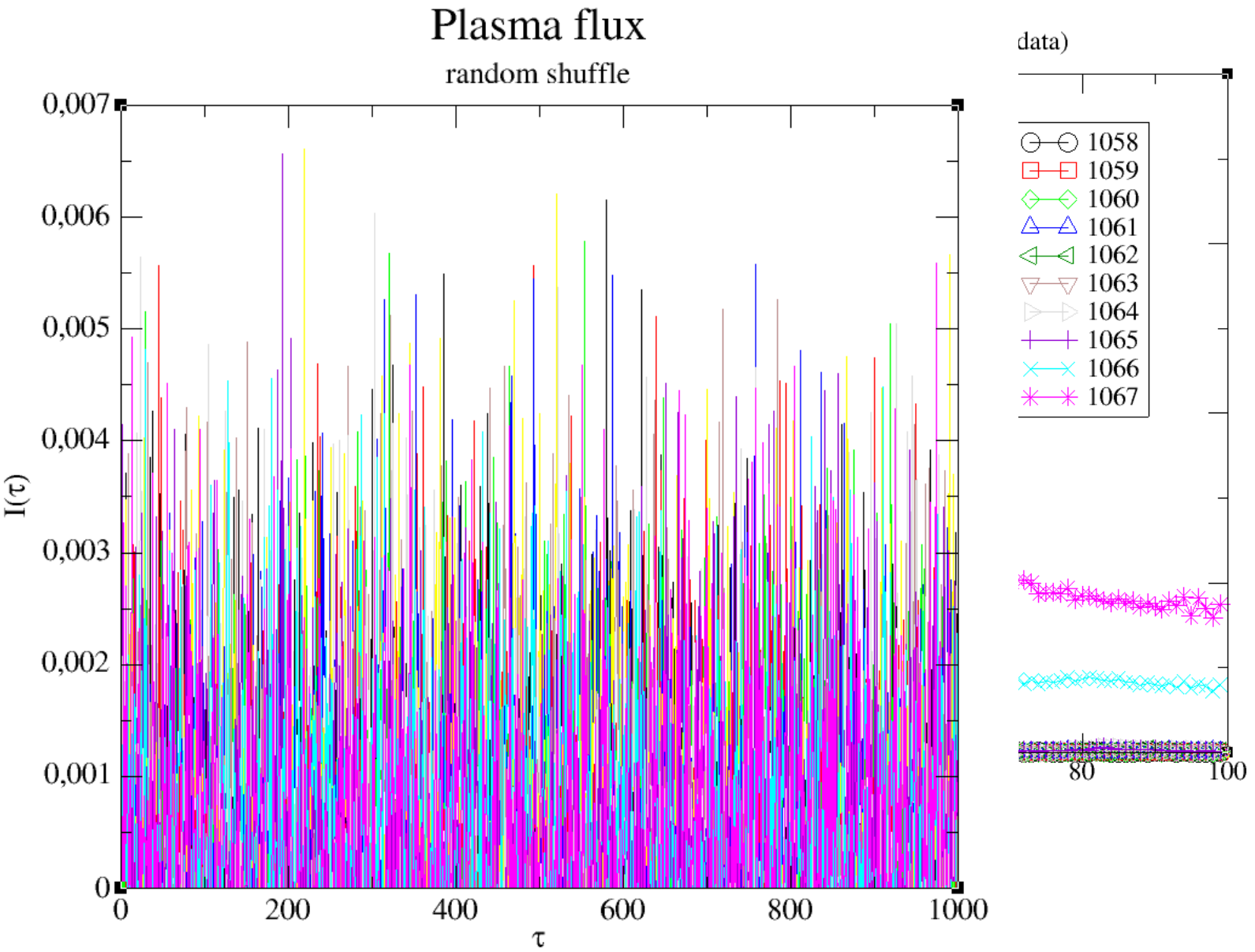
Examples


Plasma Flux
Mutual information for raw ABS(data)
I(τ) = H(X_t) + H(X_{t+τ}) - H(X_t,X_{t+τ})
τ
0
0,5
1
1,5
2
0
20
40
60
80
100
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067

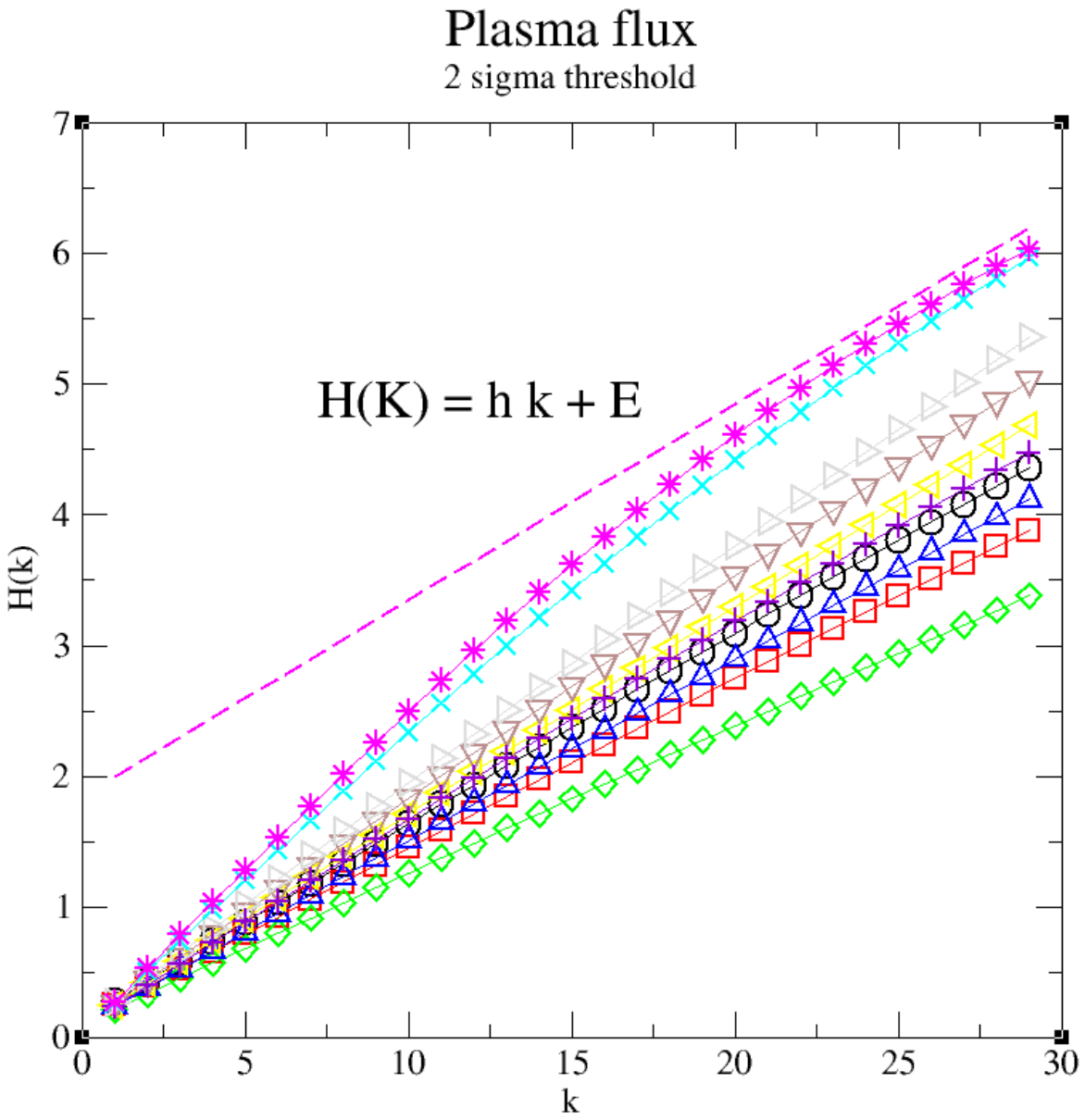
Examples


Plasma Flux
Mutual information for raw ABS(data)
I(τ)
τ
0
0,5
1
1,5
2
0
200
400
600
800
1000
60
80
100
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067

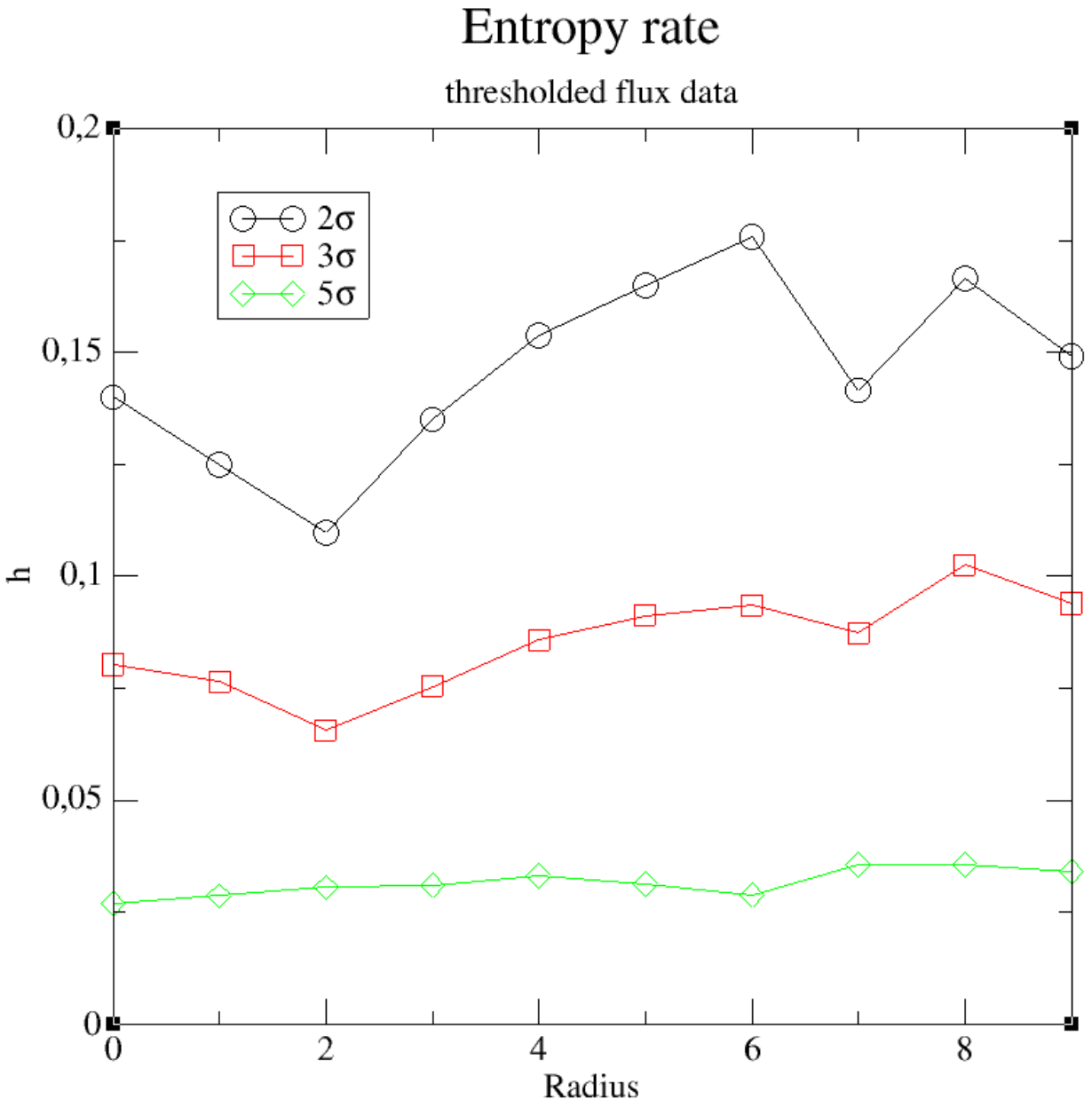
Examples


Plasma flux
random shuffle
I(τ)
τ
0,007
0,006
0,005
0,004
0,003
0,002
0,001
0
0
200
400
600
800
1000
data)
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
80
100

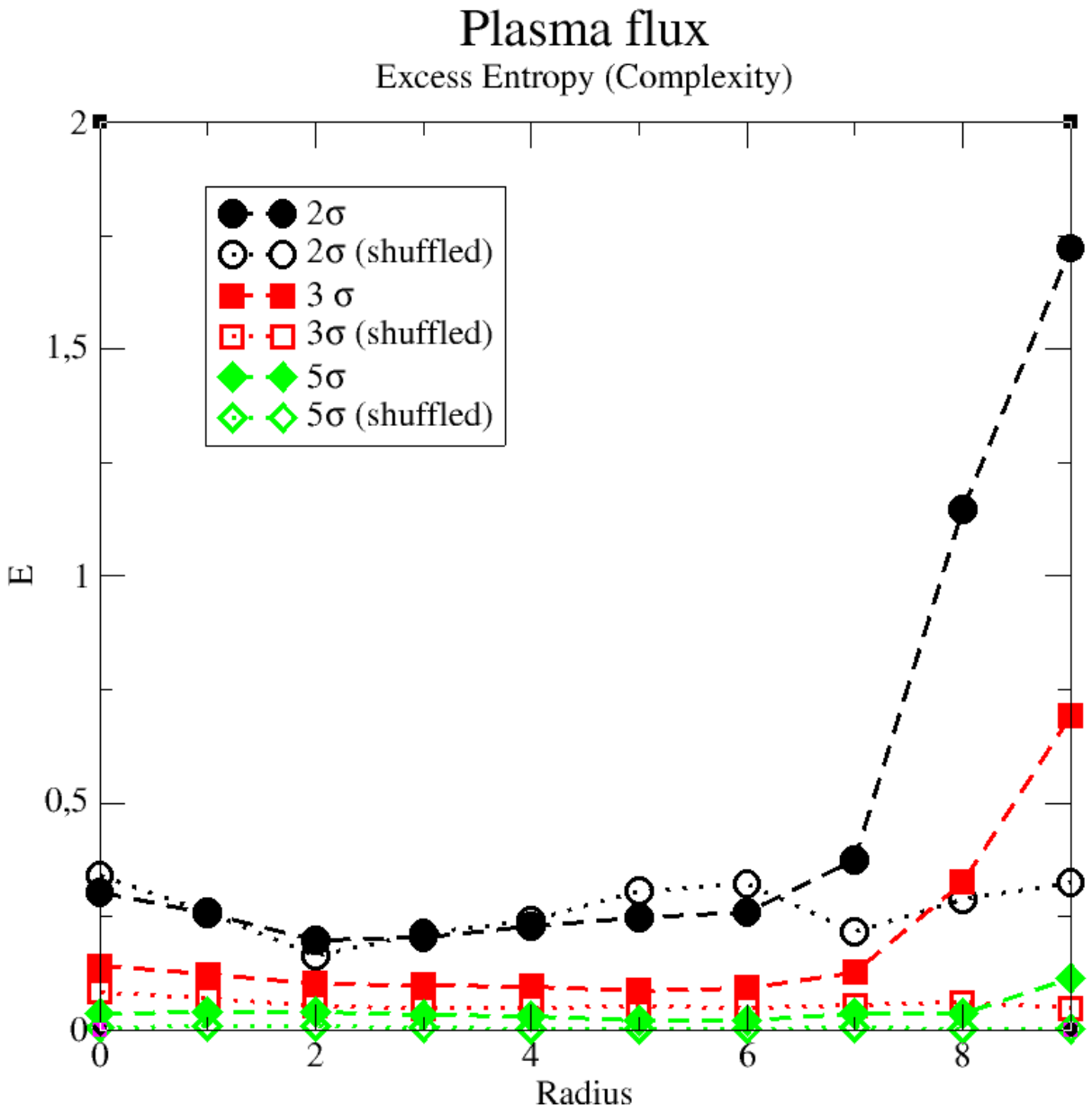
Examples


Plasma flux
2 sigma threshold
H(K) = h k + E
H(k)
k
0
1
2
3
4
5
6
7
0
5
10
15
20
25
30

Examples


Entropy rate
thresholded flux data
2σ
3σ
5σ
h
0,2
0,15
0,1
0,05
0
0
2
4
6
8
Radius

Examples


Plasma flux
Excess Entropy (Complexity)
2σ
2σ (shuffled)
3 σ
3σ (shuffled)
5σ
5σ (shuffled)
E
2
1,5
1
0,5
0
0
2
4
6
8
Radius

Maximum entropy principle

Instituto de Física de Cantabria

Maximum entropy principle

- Maximum entropy principle arose in statistical mechanics
- If nothing is known about a distribution except that it belongs to a certain class
- Distribution with the largest entropy should be chosen as the default
- Motivation:
  - Maximizing entropy minimizes the amount of prior information built into the distribution
  - Many physical systems tend to move towards maximal entropy configurations over time

Maximum entropy principle

Instituto de Física de Cantabria

Physics

- Temperature of a gas corresponds to the average kinetic energy of the molecules in the gas

$$\sum_i p_i \frac{1}{2} v_i^2 m_i$$

- Distribution of velocities in the gas at a given temperature
- this distribution is the maximum entropy distribution under the temperature constraint: Maxwell-Boltzmann distribution
- corresponds to the macrostate that has the most micro states

Maximum entropy principle

Instituto de Física de Cantabria

Formulation

- Maximize entropy

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

- Subject to

$$p_i \geq 0 \tag{1}$$

$$\sum_{i=1}^{n} p_i = 1 \tag{2}$$

$$\sum_{i=1}^{n} p_i r_{ij} = \alpha_j, \text{ for } 1 \leq j \leq m \tag{3}$$

Maximum entropy principle

Instituto de Física de Cantabria

Maximum entropy distribution

- Form Lagrangian

$$J(p) = -\sum_{i=1}^{n} p_i \log p_i + \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) + \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i r_{ij} - \alpha_j \right)$$

- Take derivative with respect to p_i: $-1 - \log p_i + \lambda_0 + \sum_{j=1}^{m} \lambda_j r_{ij}$
- Set this to 0, and solution is *maximum entropy distribution*

$$p_i^* = \frac{e^{\sum_{j=1}^{m} \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen such that $\sum_i p_i^* = 1$, and $\sum_i p_i^* r_{ij} = \alpha_j$.

Maximum entropy principle

Instituto de Física de Cantabria

Maximum entropy classifier

- In some fields of machine learning, multinomial logic model is refer to as a maximum entropy classier
- Minimizes the amount of prior information built into the distribution
- X_i: feature vector, β_k: vector of weights for outcome k, Y_k: random outcome, $k = 1, \ldots, K$
- Takes the form

$$p(Y_i = k) = \frac{e^{\beta_k^\top X_i}}{1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell^\top X_i}}, \quad k = 1, \ldots, K-1.$$

Maximum entropy principle

Instituto de Física de Cantabria

Summary

- Maximizing entropy minimizes the amount of prior information built into unknown distribution
- Maximum entropy distribution can be found explicitly

$$p_i^* = \frac{e^{\sum_{j=1}^{m} \lambda_j r_{ij}}}{e^{1-\lambda_0}}$$

- Maximum entropy principle widely used