Machine Learning for Cognitive Sciences: Principles and Applications

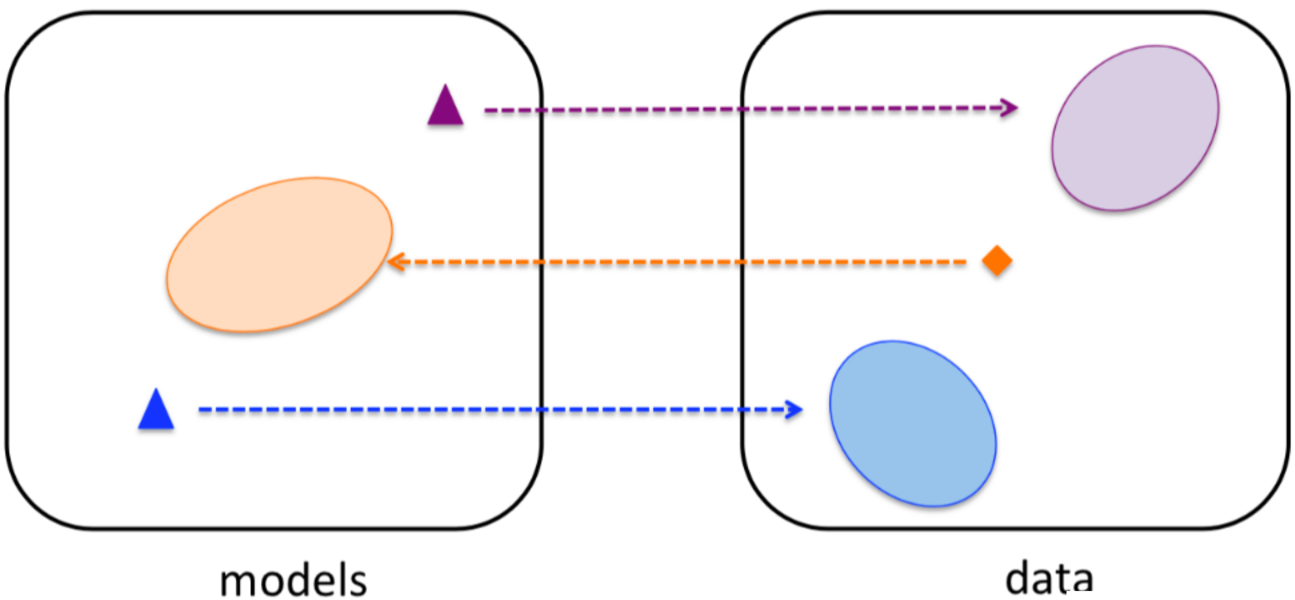
Simona Cocco (Physics Departement of ENS)
simona.cocco@phys.ens.fr CM &TD

Vito Dichio (Physics Departement of ENS)
vito.dichio@phys.ens.fr TD

Week 2

Recall of previous lecture

Bayesian Inference



- Each model ▲ defines a distribution of possible data 

- Each data ◆ defines a distribution of possible models 

Bayes Theorem in Inference

Suppose we have an ensemble of **M** data points $\boldsymbol{y}_i \in \mathbb{R}^L$,

Generated with a model with **D** unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^D$

Bayes' rule:

Posterior Likelihood Prior

$$p(\boldsymbol{\theta}|Y) = \frac{p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(Y)} \ .$$

Evidence

The Evidence ensures the normalization of the Posterior

$$p(Y) = \int \mathrm{d}\boldsymbol{\theta}\, p(Y|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ .$$

Plan of the lecture

- Application of Bayes Theorem: Laplace birth rate problem

- Connection between Entropy in physics and Information and communication theory: Shannon Information

- Mutual Information and some applications to efficient coding theory in neuroscience.

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945

… boys … : 251,527

y = nb. of female births, M = total number of births

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 245,945

… boys … : 251,527

y = nb. of female births, M = total number of births

Inference: θ= probability that a newborn baby is a girl

- Prior distribution: uniform over θ in [0;1]

INTUITION: $\theta \sim$

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 241,945

... boys ... : 251,527

y = nb. of female births, M = total number of births

Inference: θ= probability that a newborn baby is a girl

- Prior distribution: uniform over θ in [0;1]

INTUITION: $\theta \sim \frac{241{,}945}{493{,}472} = 0.4903$

Laplace and the birth rate of boys & girls

Historical example: « proof » by Laplace that the female and male birth rates are different

Data: Nbs of girls born in Paris from 1745 to 1770 : 241,945

... boys ... : 251,527

σ = nb. of female births, n = total number of births

Inference: θ = probability that a newborn baby is a girl

- Prior distribution: uniform over θ in [0;1]

- Likelihood: $\mathrm{p}(y|\vartheta) = \binom{M}{y}\,\vartheta^{y}(1-\vartheta)^{M-y}$

Binomial Distribution.

Laplace and the birth rate of boys & girls

<u>Historical example:</u> « proof » by Laplace that the female and male birth rates are different

<u>Data:</u> Nbs of girls born in Paris from 1745 to 1770 : 241,945

... boys ... : 251,527

σ = nb. of female births, n = total number of births

<u>Inference:</u> θ = probability that a newborn baby is a girl

- Prior distribution: uniform over θ in [0;1]

- Likelihood: $p(y|\theta) = \binom{M}{y}\, \vartheta^{y}(1-\vartheta)^{M-y}$

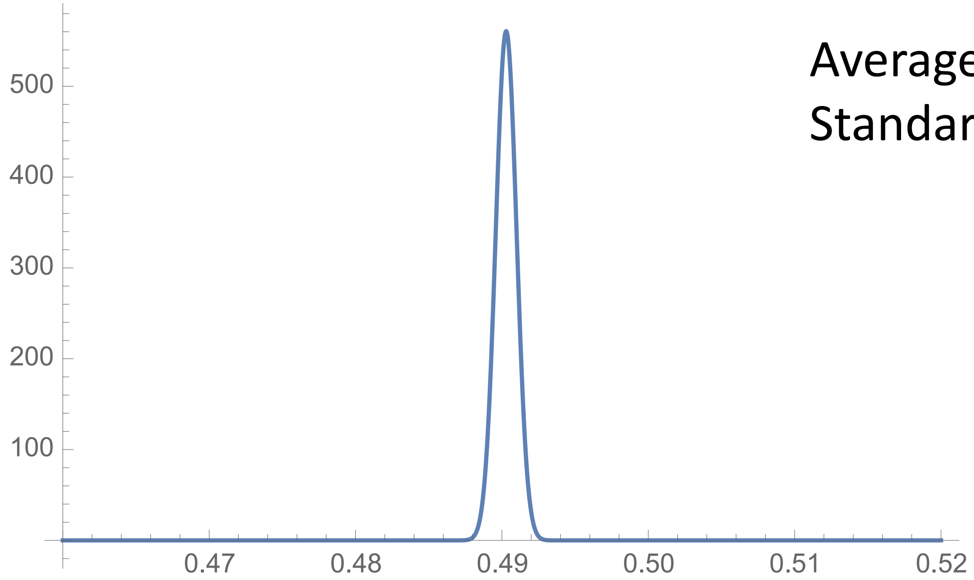
Uniform in interval [0,1]

- Bayes: $p(\theta|y) = \dfrac{p(y|\theta)\times p(\theta)}{p(y)}$

Cst $\int_0^1 d\theta\; \theta^{y}\,(1-\theta)^{M-y}$

Laplace and the birth rate of boys & girls

<u>Posterior distribution:</u>



Average θ = 0.490291

Standard deviation θ = 0.007117

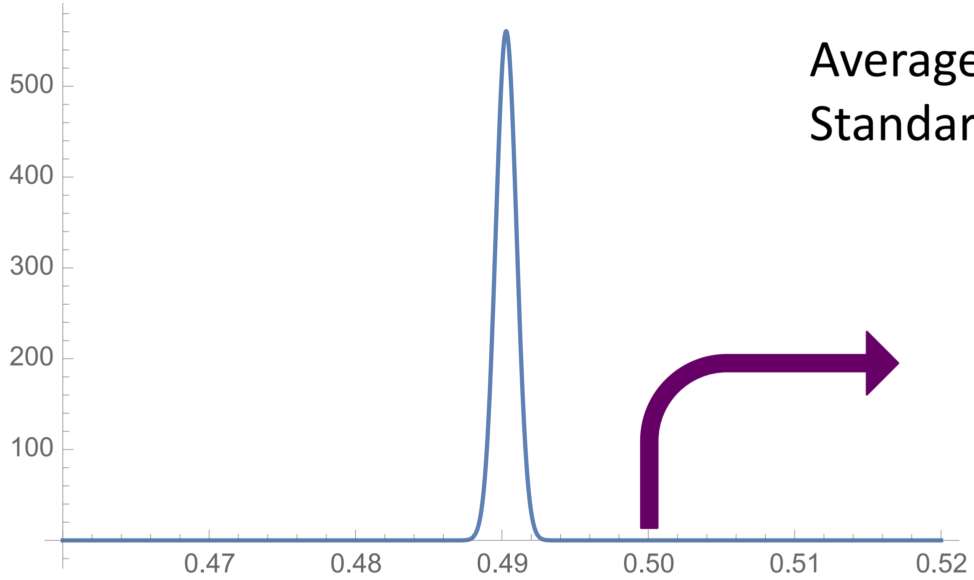
Uniform in interval [0,1]

- Bayes: $$\mathrm{p}(\theta|y) = \frac{p(y|\theta)\times p(\theta)}{p(y)}$$

Cst $\int_0^1 d\theta\ \theta^y\,(1-\theta)^{M-y}$

Laplace and the birth rate of boys & girls

Posterior distribution:

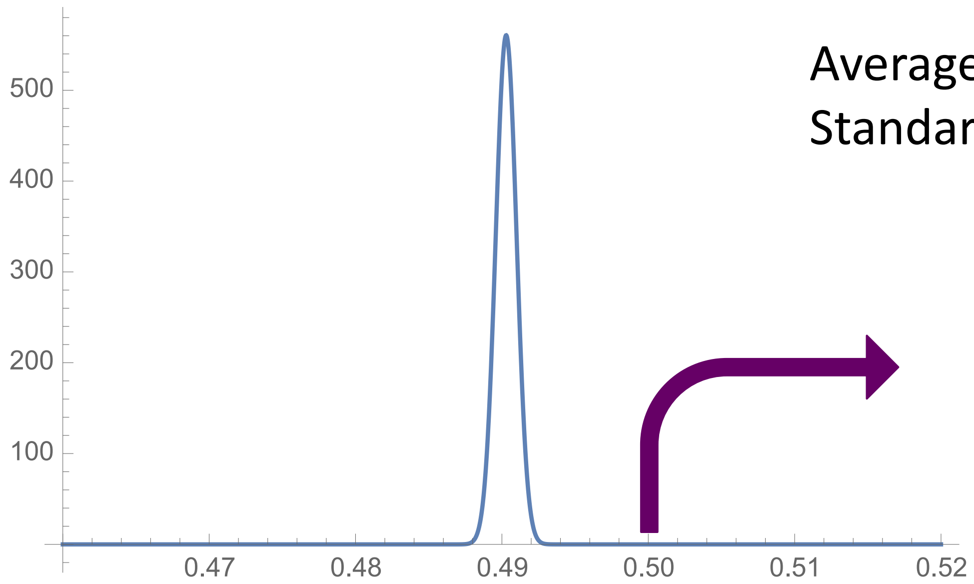


Average θ = 0.490291
Standard deviation θ = 0.007117

Probability that θ is equal or larger than 0.5 = $\int_{0.5}^{1} d\theta \;\, p(\theta|y) \;\approx 10^{-42}$ Extremely unlikely!

Laplace and the birth rate of boys & girls

Posterior distribution:



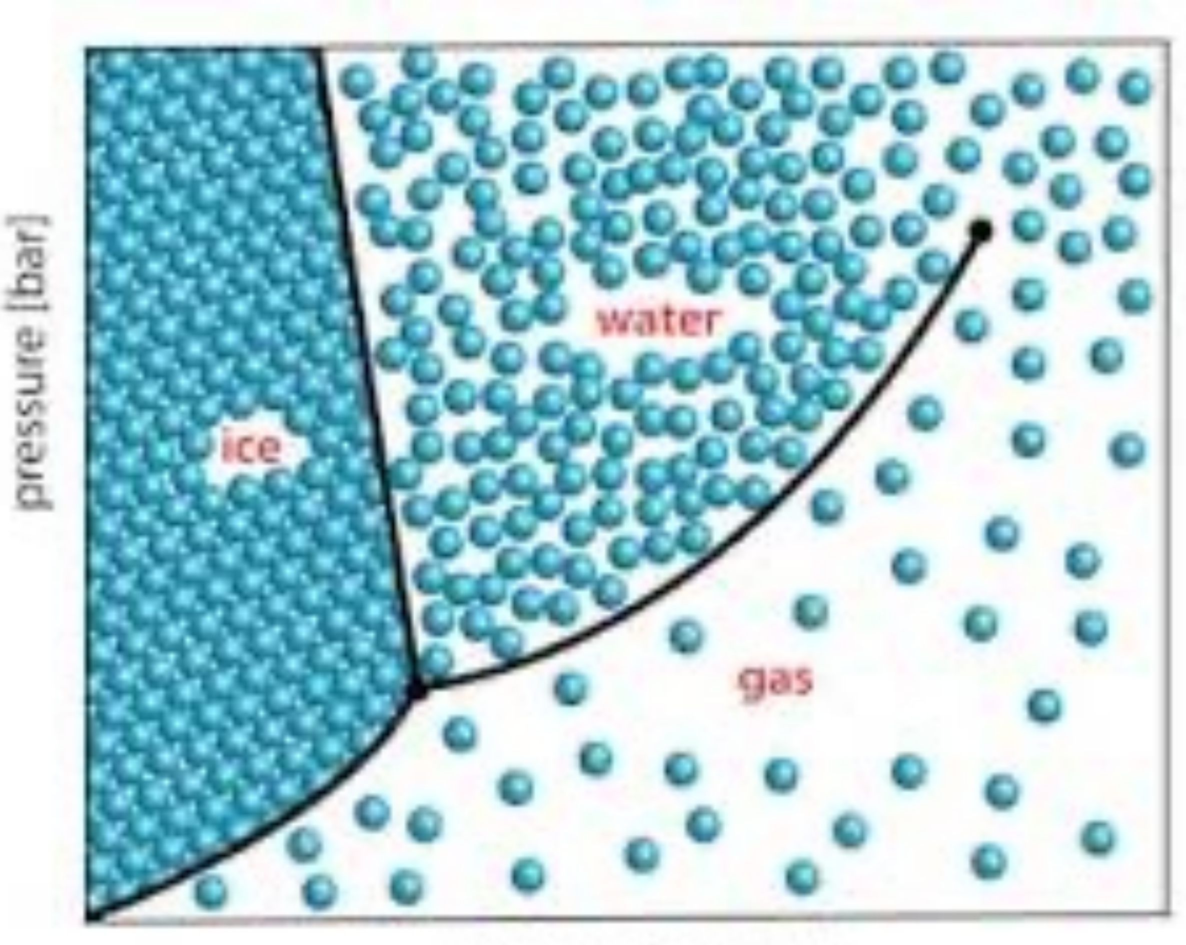
Average θ = 0.490291
Standard deviation θ = 0.007117

Probability that θ exceeds 0.5 = $\int_{0.5}^{1} d\theta \;\; p(\theta|y) \;\approx 10^{-42}$ Extremely unlikely!

- In the tutorial you are computing the posterior distribution, the most likely value, and average value for the spiking rate of a neuron from a neural recording (Poissonian Distribution)

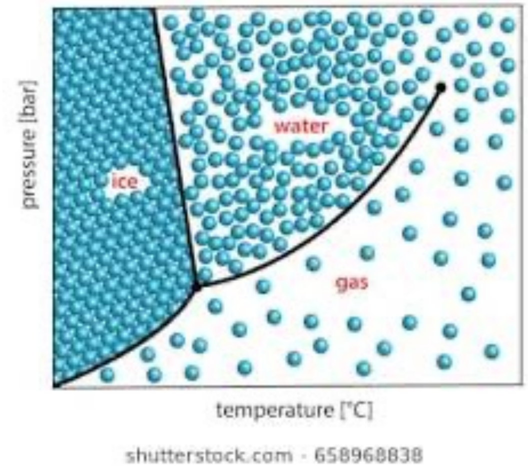
Entropy & Statistical Physics

Liquid – vapor-solid Phase Transition



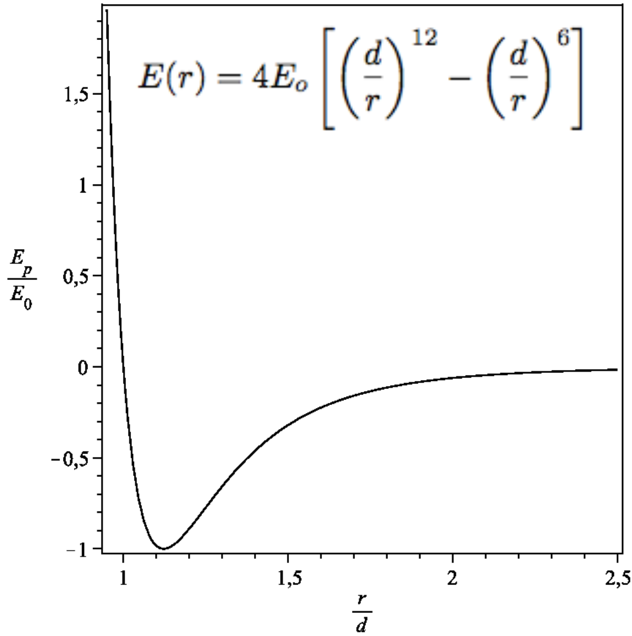
Macroscopic behavior
Derives from the presence
of **Many Particules**
(here water molecules**)**
in Interaction.

Entropy & Statistical Physics



Liquid – vapor-solid Phase Transition

- Macroscopic behavior derives from the presence of many particules (molecules)
  ->**Thermodynamic limit:**



- ***Interaction*** Energy :**Van der Waals interactions**

unique & simple potential, repulsive at short distance, attractive at medium distance, zero at long distance.

- Change of state are described by minimizing the **Free energy of the system:F= E-TS**, describing the interplay between the **Energy E** and the **Entropy S,** reflecting thermal motions.

Observables: density, viscosity,… correlations between the positions of the particules

Boltzmann Entropy (1877)

Entropy of a perfect gas: uncertainty about the microscopic configurations of the molecules of the gas.

Boltzmann gave a probabilistic way of defining the entropy as proportional to the logarithm of the numbers of the dominant microscopic configurations:

$S=-k_B \Sigma_c p_c \ln p_c$

Unit: k_B: if we multiply by the absolute temperature (in kelvin) we obtain the entropy in joules per molecule (or degree of freedom)

Entropy & Information Theory

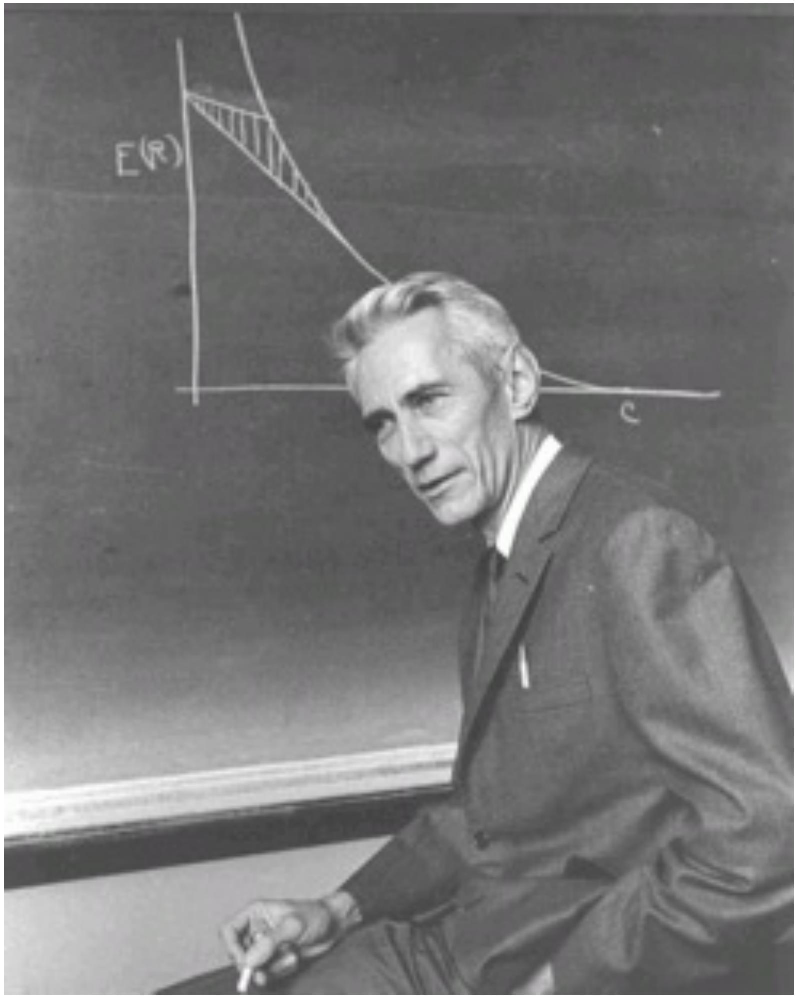
- **Shannon Entropy** in communication theory as **Missing Information**
- Importance in **Coding and Optimal Language compression**

Communication Theory

Claude Shannon
(1916-2001)

During World War II : intelligence service, cryptography

→ "A mathematical theory of communications"
(1948)



Information Theory

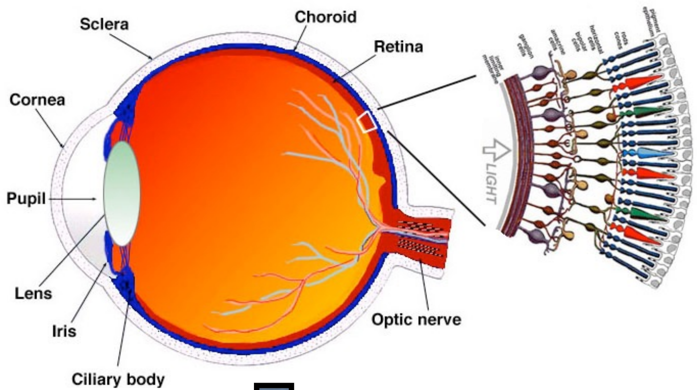
A branch of science:

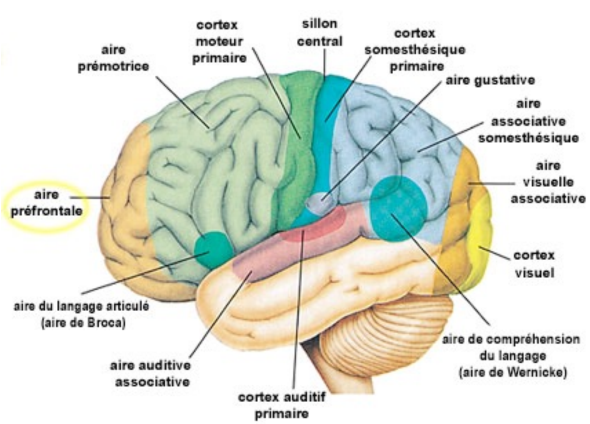
- Recent (70 years)
- Very important from a technological point of view
- With multiple branches

- ... And closely connected to Cognitive Sciences and Statistical Mechanics

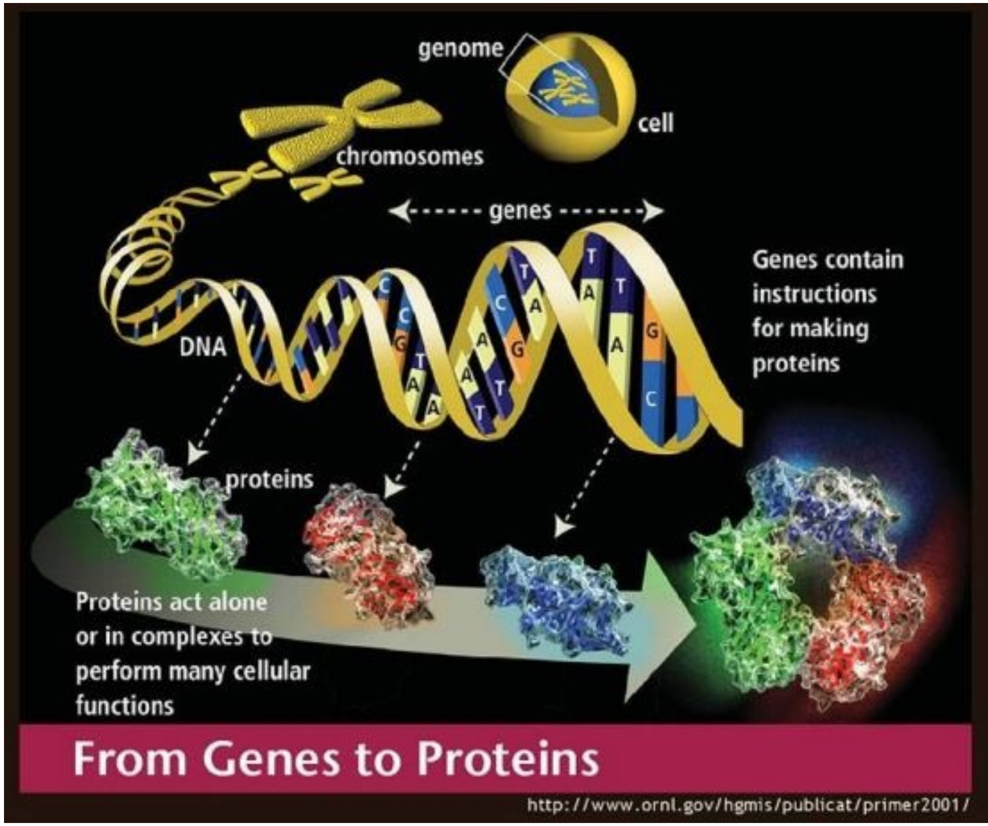
Transmission of information in biology

Neurobiology

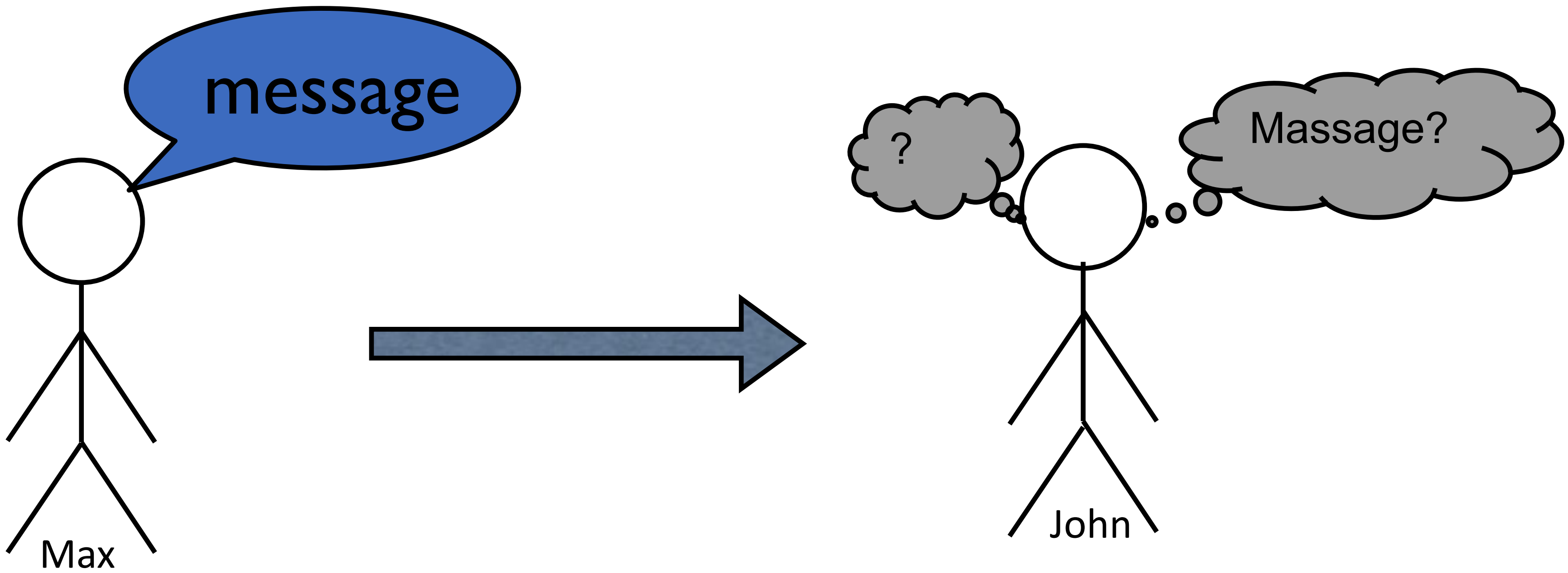




Genetics and Molecular Biology



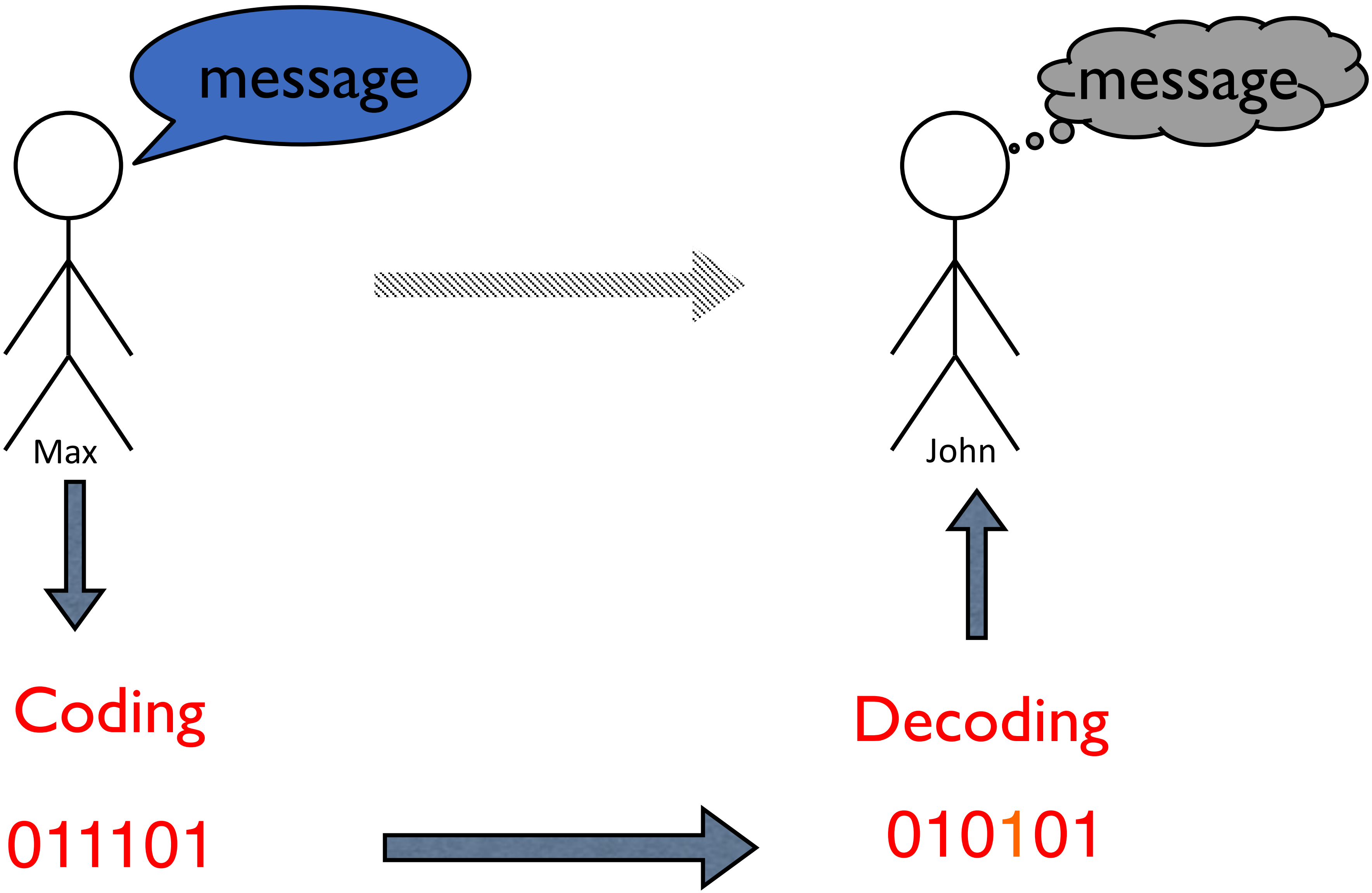
How to communicate in an efficient way?



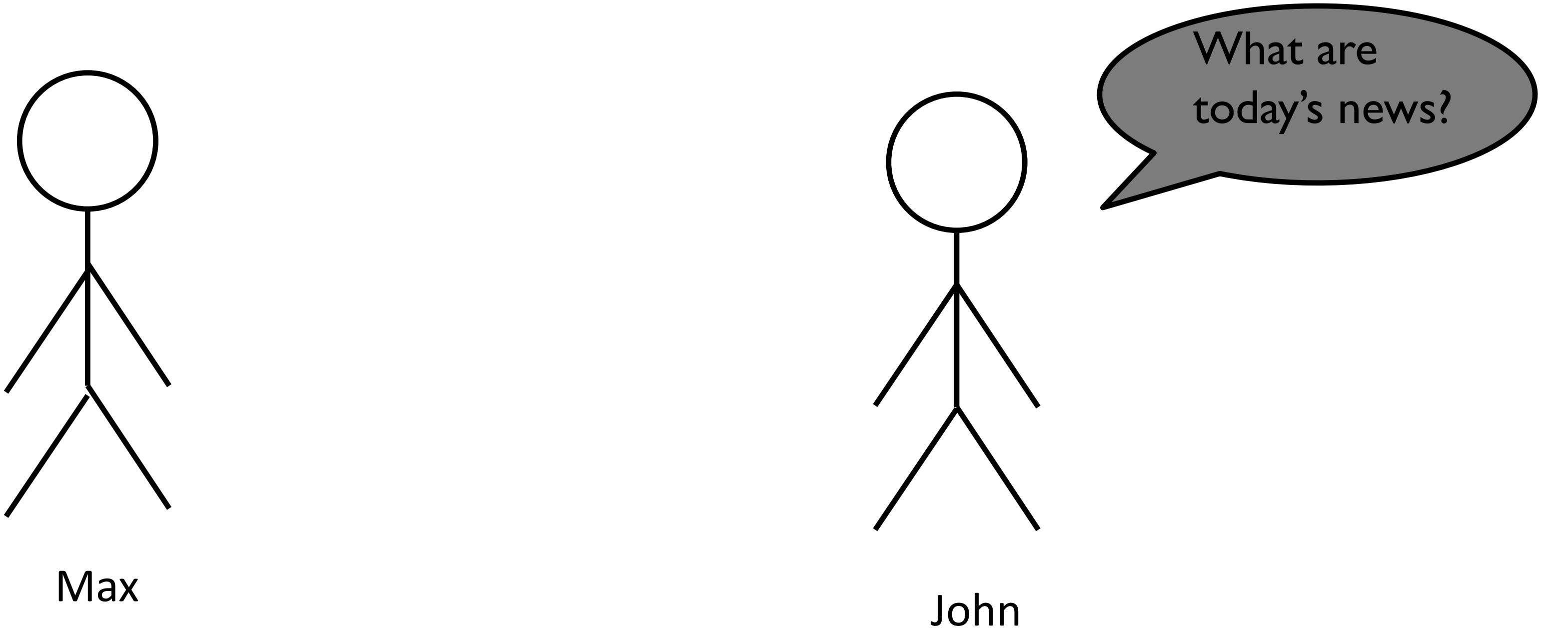
message = series of words in a language
= series of symbols 0100011101101…

How to communicate in presence of noise?
in a concise way?

Coding and decoding in communication

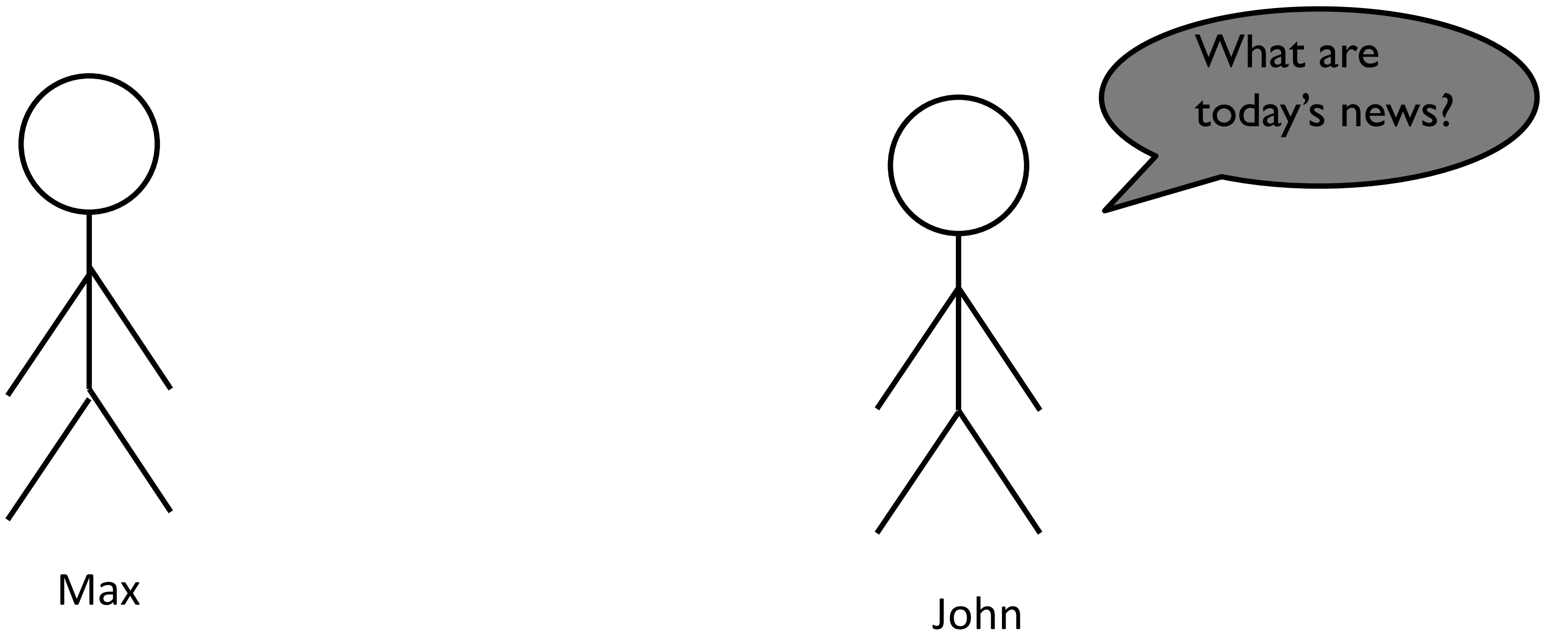


Shannon Entropy: definition and intuition



John knows the list of all N possible Max's answers and their probabilities

Shannon Entropy: definition and intuition



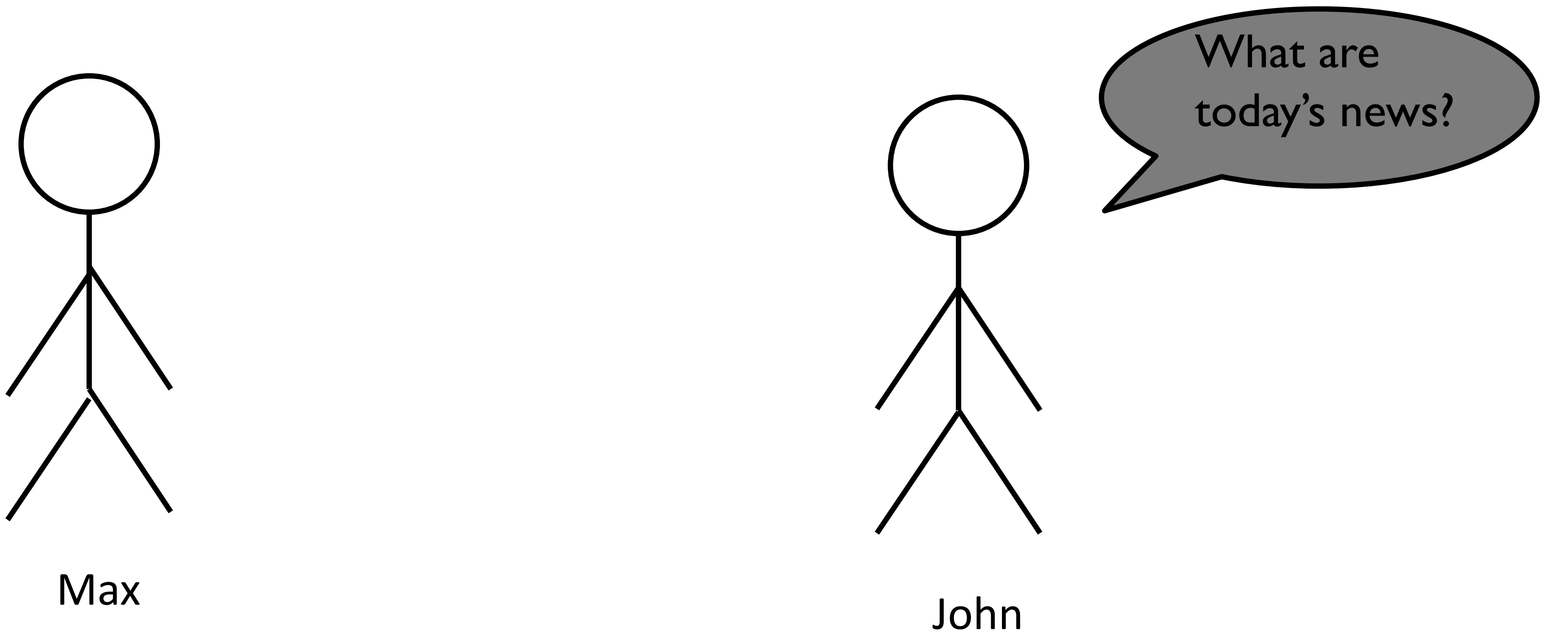
John knows the list of all N possible Max's answers and their probabilities

From this list of possible answers and their probabilities one compute the **Shannon entropy:**

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Unit: bit

Shannon Entropy: definition and intuition



John knows the list of all N possible Max's answers and their probabilities

From this list of possible answers and their probabilities one compute the **Shannon entropy:**

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Unit: bit

Uncertainty about what Max will say in response to John's question

This uncertainty is removed once Max gives his answer.

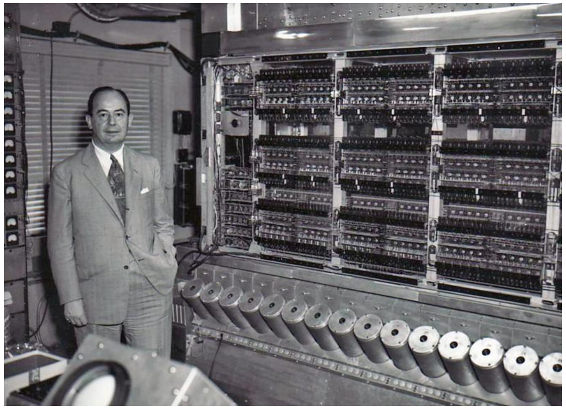
The Shannon entropy quantify also the **average gain of information** from the answer.

How to call $\boxed{S = -\sum_{i=1}^{N} p_i \log_2 p_i}$?

… The uncertainty
… The missing information
… The entropy

How to **call** $\boxed{S = -\sum_{i=1}^{N} p_i \log_2 p_i}$ **?**

Shannon aux Bell Labs



Von Neumann à Princeton

“My greatest concern was what to call it. I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage.”

Cité par M.Tribus, E.C. McIrvine, Energy and information, Scientific American, 224 (Sept. 1971).

Properties of

… The uncertainty
… The missing information
... The entropy

$$\boxed{S = -\sum_{i=1}^{N} p_i \log_2 p_i}$$

Is the unique measure of missing information consistent with certain simple and plausible requirements.

Properties of the 'missing Information'

$$S = f(\{p_i\})$$

- N Possible answer of probability p_n.

1. S is is a function which should grow monotonically with N

2. If the answer is composed of *m* independent parts: e.g. A B C

e.g.: 1) the news on the first page of the newspapers are A) that a given law has been voted B) the weather is nice and C) Brazil has won the soccer world cup

$$S = S_A + S_B + S_C$$

$f(\{p_A p_B p_c\})=f(\{p_A\})+f(\{p_B\})+f(\{p_C\})$?

Properties of the 'missing Information'

$$S = f(\{p_i\})$$

- N Possible answer of probability p_n.

1. S is is a function which should grow monotonically with N

2. If the answer is composed of *m* independent parts: e.g. A B C

e.g.: 1) the news on the first page of the newspapers are A) that a given law has been voted B) the weather is nice and C) Brazil has won the soccer world cup

$$S = S_A + S_B + S_C$$

$f(\{p_A p_B p_c\}) = f(\{p_A\}) + f(\{p_B\}) + f(\{p_C\})$?

Log(A x B)=log(A) + Log(B)

Properties of the 'missing Information'

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

- N Possible answer of probability p_n.

S is a function which should grow monotonically with N

Eg. Uniform probability of answer p_n= 1/N. S= Log N

- Random binary words L=2:

1 0
0 1
1 1
0 1

$N=2^2$ $p_i=1/4$

- H= Log 4=2 bits

Properties of the 'missing Information'

$$S = -\sum_{i=1}^{N} p_i \log_2 p_i$$

- N Possible answer of probability $p_{n.}$

2. If the answer is composed of *m* independent parts: e.g. A B C (e.g. 1) the news on the first page of the newspapers are A) that a given law has been voted B) the weather is nice and C) Brazil has won the soccer world cup)

$$\boxed{S = S_A + S_B + S_C}$$

- Random binary words L=2:

s_1	s_2
1	0
0	1
1	1
0	0

$N=2^2$ $p_i=1/4$

- 2 independent variables: $p_i=p(s1,s2)=p(s1)p(s2)$

- S=S[p(s1)]+S[p(s2)]

Relationship with Communication Theory

Symbols or "words" to communicate: $M_1, M_2, \ldots, M_N$

Probability (usage frequency): $p_1, p_2, \ldots, p_N$

Codage: $M_n \rightarrow C_n = 0110010110$

Pb: Find a code $= \{C_n\}$ such that the number of bits used on average be as small as possible

Relationship with Communication Theory

Idea: Use a short code C_n for frequent symbols.

Shortest code: Length $L(M_n) \propto -log_2\,(p_n)$

$$\langle L \rangle = \sum_n p_n L(M_n)$$

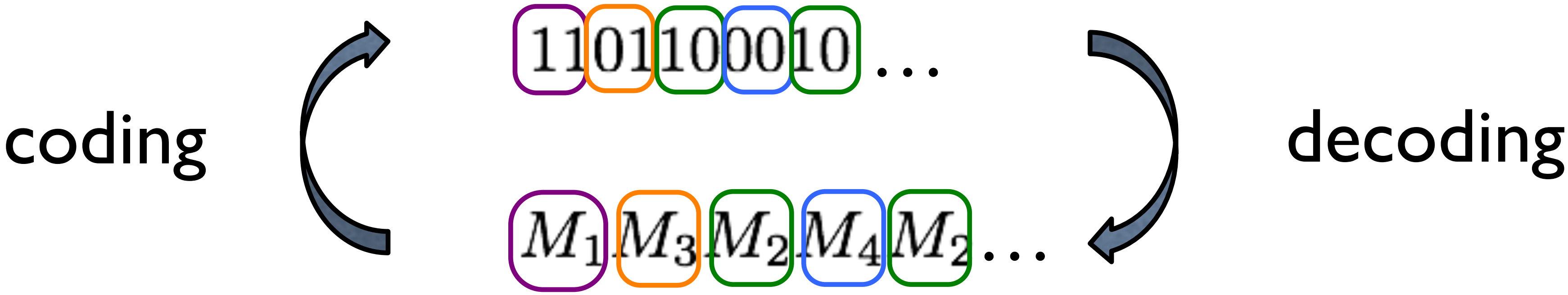
is the entropy of the probability distribution

$$\boxed{S = -\sum_{i=1}^{N} p_i \log_2 p_i}$$

A simple code

Code A

words	symboles
M_1	11
M_2	10
M_3	01
M_4	00



A variable-length code

Code B

words	symboles
M_1	1
M_2	01
M_3	001
M_4	000

coding

$$1\,01\,1\,01\,000\ldots$$

$$M_1 M_2 M_1 M_2 M_4 \ldots$$

decoding

NB : no ambiguity as no word is another word's prefix

What is the better code?

Code A	
M_1	11
M_2	10
M_3	01
M_4	00

Code B	
M_1	1
M_2	01
M_3	001
M_4	000

Average length of codeword:

$$\langle L \rangle = \sum_n p_n L(M_n)$$

Code A : $\langle L \rangle = 2$

Code B : Depends on p_n

The dices

	S	Code A	Code B
 Normal		$\boxed{\langle L\rangle = 2}$	
 Biased		$\langle L\rangle = 2$	

Two examples with code B

$$\langle L \rangle = \sum_n p_n L(M_n)$$

Code B

M_1	1
M_2	01
M_3	001
M_4	000

$$p_1 = p_2 = p_3 = p_4 = 1/4$$

$$\langle L \rangle = \frac{1}{4}(1+2+3+3) = \frac{9}{4}$$

$$p_1 = 1/2,\ p_2 = 1/4,\ p_3 = p_4 = 1/8$$

$$\langle L \rangle = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 3 \times \frac{1}{8} = \frac{7}{4}$$

The dices

	S	Code A	Code B
 Normal	2	$\boxed{\langle L\rangle = 2}$	$\langle L\rangle = \frac{9}{4}$
 Biased	$\frac{7}{4}$	$\langle L\rangle = 2$	$\boxed{\langle L\rangle = \frac{7}{4}}$

$$\langle L\rangle_{\text{mini}} = S$$

Relationship with Communication Theory

Idea: Use a short code C_n for frequent symbols.

Shortest code: Length $L(M_n) \propto -log_2\ (p_n)$ (Proof: as the branching process (McKay)

$$\langle L \rangle = \sum_n p_n L(M_n)$$

is the entropy of the probability distribution

$$\boxed{S = -\sum_{i=1}^{N} p_i \log_2 p_i}$$

Code A

$p_1 = p_2 = p_3 = p_4 = 1/4$

M_1	11
M_2	10
M_3	01
M_4	00

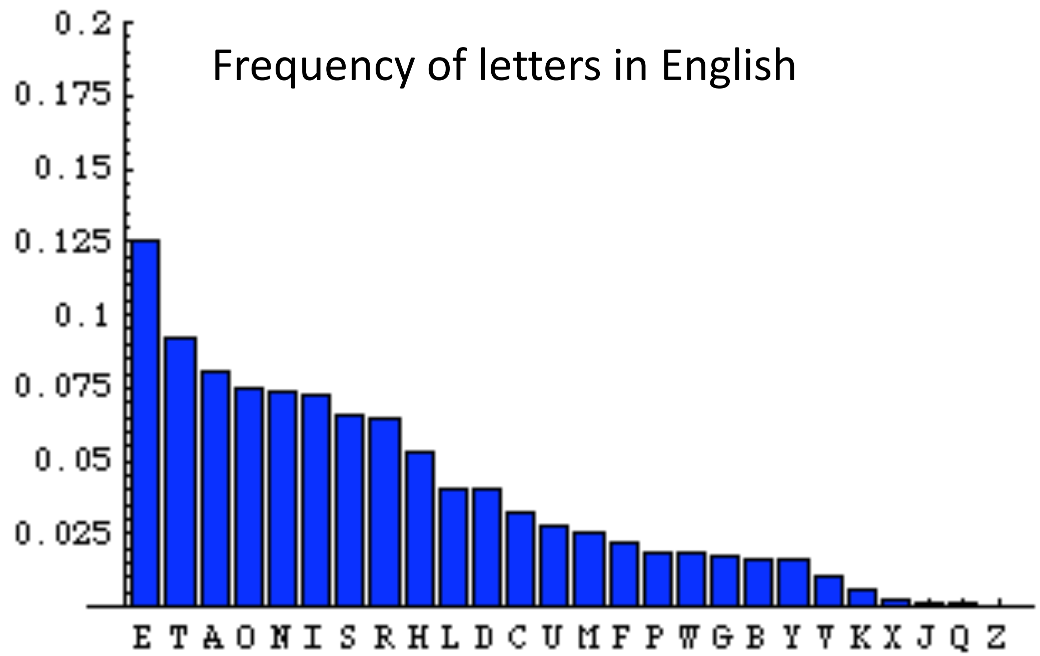
Code B

$p_1 = 1/2,\ p_2 = 1/4,\ p_3 = p_4 = 1/8$

M_1	1
M_2	01
M_3	001
M_4	000

Ex: compute S in the 2 cases

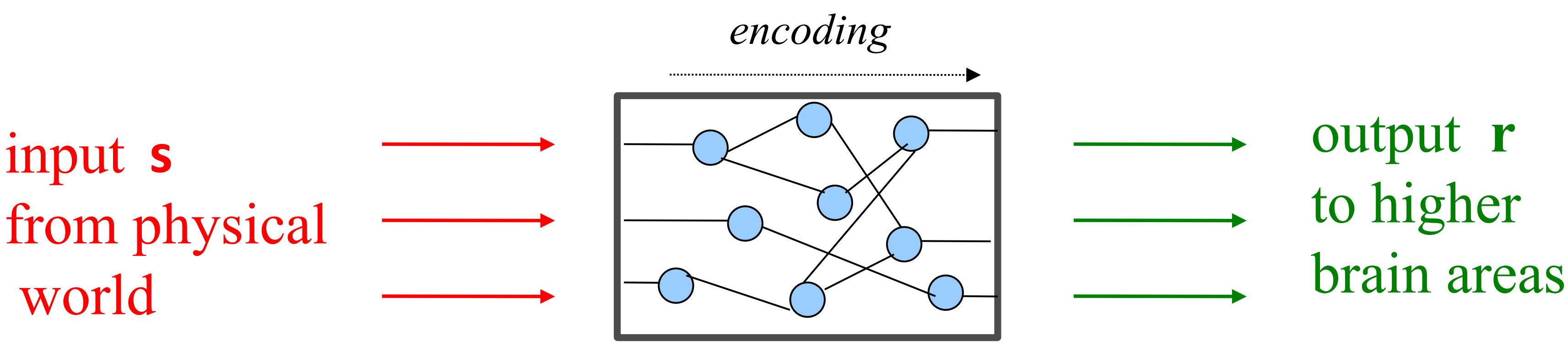
Example : Morse alphabet (1832)



•Practical compression methods? (gzip, jpeg, MP3, ...)

Efficient coding

Organization principle for sensory areas



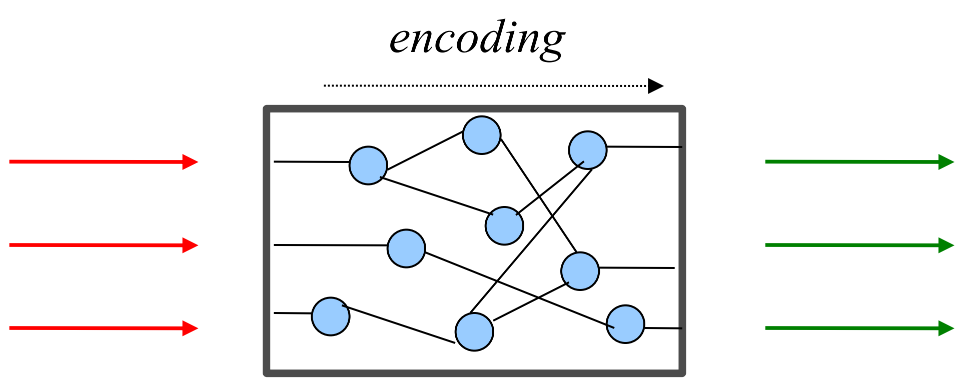
Hypothesis: encoding should maximize mutual information

$$\text{MI (input, output)} = \int ds \int dr \; P(s,r) \log\left(\frac{P(s,r)}{P(s)P(r)}\right)$$

Efficient coding and Mutual Information

Organization principle for sensory areas

input **s**
from physical
world



output **r**
to higher
brain areas

Hypothesis: encoding should maximize mutual information

$$\text{MI (input, output)} = \int ds \int dr \; P(s,r) \log\left(\frac{P(s,r)}{P(s)P(r)} \right)$$

For variables taking discrete values:

$$\boxed{MI(1,2) = \sum_{x_1 x_2} p(x_1,x_2) \; \log\left[\frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)} \right]}$$

Mutual information

The dependence between events (or event distributions) is characterized through the **mutual information**

$$\boxed{MI(1,2) = \sum_{x_1 x_2} p(x_1, x_2)\ \log_2 \left[\frac{p(x_1, x_2)}{p_1(x_1)\, p_2(x_2)} \right]}$$

in bits

This quantity measures how much information one has on one variable from knowledge of the other one.

(1) Always positive (or null)
(2) Degraded through processing: $x_2 \rightarrow x_3$ then $MI(1,3) \leq MI(1,2)$
(3) It is null for independent variables

<u>*Exercise 2:*</u> *compute MI(box,cookie type) in bits*

Characterization of dependent events

Recall from Lecture 1

Definition of conditional probability: $$p_2(y|\theta) \equiv \frac{p(y,\theta)}{p_1(\theta)}$$

Definition of marginal Probability:

$$\mathrm{p}(y) = \sum_{\theta} p(y,\theta) \qquad \mathrm{p}(\theta) = \sum_{y} p(y,\theta)$$

Box A	Box B
20 plain cookies + 20 chocolate cookies	10 plain cookies + 30 chocolate cookies

$p_2(y = \text{plain} \mid \theta = A) = 1/2$,
$p_2(y = \text{chocolate} \mid \theta = A) = 1/2$

$p_2(y = \text{plain} \mid \theta = B) = 1/4$,
$p_2(y = \text{chocolate} \mid \theta = B) = 3/4$

Exercise: *compute MI(box,cookie type) in bits*

$$\text{MI(box,cookie type)}=\sum_{y,\theta} p(y,\theta)\ log\left[\frac{p(y,\theta)}{p_1(y)p_2(\theta)}\right]$$

$p(y,\theta)$	A	B			
Plain	$\frac{1}{4}$	$\frac{1}{8}$	$\longrightarrow$	marginal distribution by summing columns or rows	$\longrightarrow$ $\frac{3}{8}$
Choco	$\frac{1}{4}$	$\frac{3}{8}$			$\frac{5}{8}$

$$MI = \frac{1}{4}\log_2\left[\frac{\frac{1}{4}}{\frac{1}{2}\times\frac{3}{8}}\right]+\frac{1}{4}\log_2\left[\frac{\frac{1}{4}}{\frac{1}{2}\times\frac{5}{8}}\right]+\frac{1}{8}\log_2\left[\frac{\frac{1}{8}}{\frac{1}{2}\times\frac{3}{8}}\right]+\frac{3}{8}\log_2\left[\frac{\frac{3}{8}}{\frac{1}{2}\times\frac{5}{8}}\right]\approx 0.05\ bit$$

Mutual Information & Entropy

- The mutual information thus represents the average gain in information over x_1 when x_2 is known, or alternatively, over y when x is known

$$MI(x_1,x_2) = S[p(x_1) - S[p(x_1|x_2)] \quad = \quad S[p(x_2) - S[p(x_2|x_1)]$$

Conditional entropy of x_1 at given x_2

$$S[p(x_1|x_2)] = -\sum_{x_2} p(x_2\ ) \sum_{x1} p(x_1|x_2) \log_2 (\ p(x_1|x_2)$$

Bring home message :

- The entropy is a measure of the missing information, which is what we do not know about the microscopic state of a system which is macroscopically constraint.

- The entropy is of fondamental importance in the science of the communication and the technological application but also for the comprehension of biological systems and in particular the brain

Entropy &Thermodynamics

Liquid – vapor-solid Phase Transition