

# Physique Statistique (PHY432)

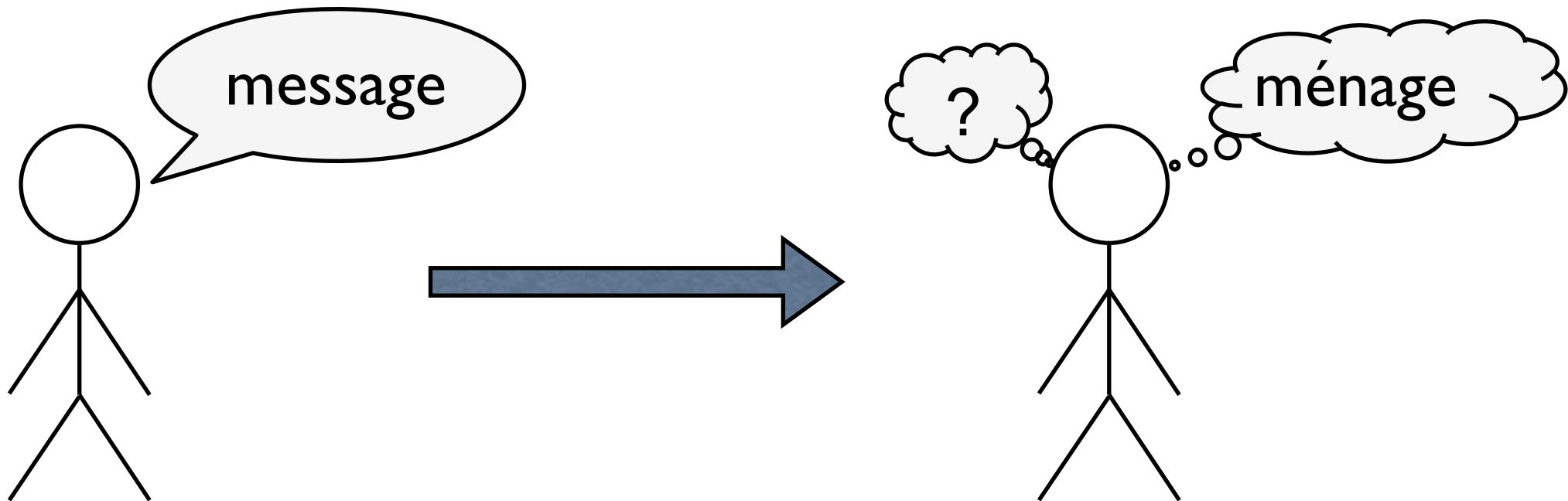
Amphi 7

*Physique statistique,  
théorie de l'information  
et inférence*

*(l'entropie en physique et au-delà ...)*

M. Mézard et R. Monasson

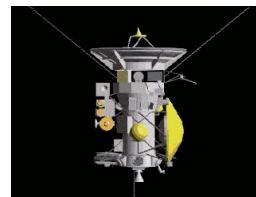
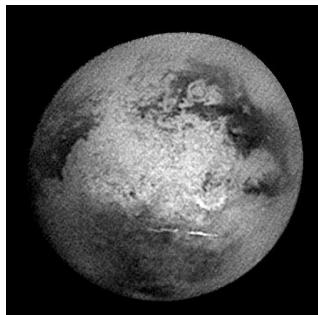
# Comment communiquer de manière efficace?



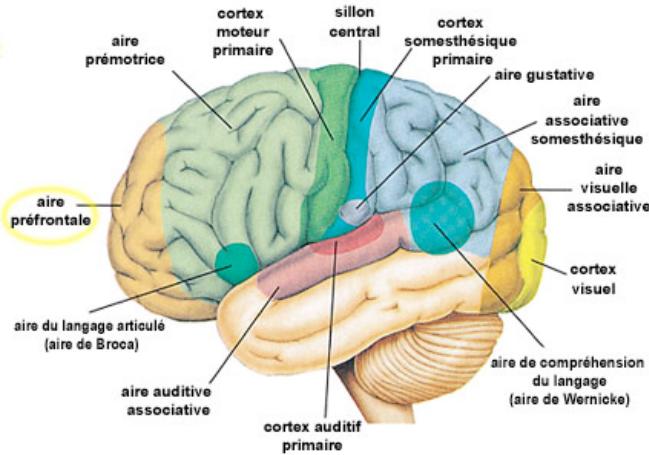
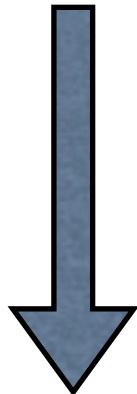
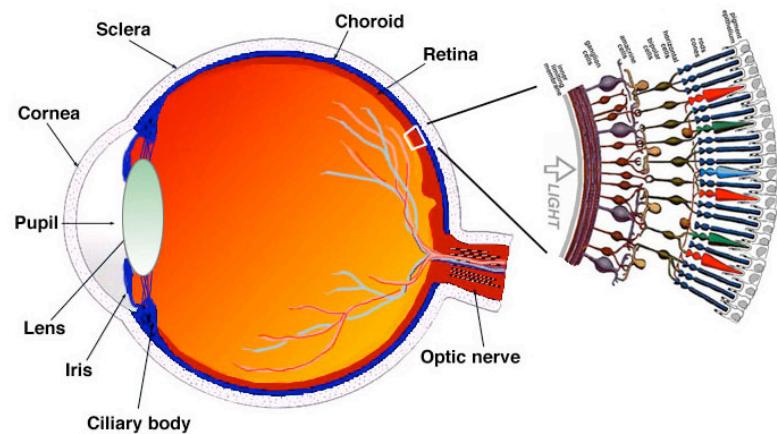
message = suite de mots dans une langue  
= suite de symboles 010001101101...

Comment communiquer en présence de bruit?  
de manière concise?

# Transmission d'information

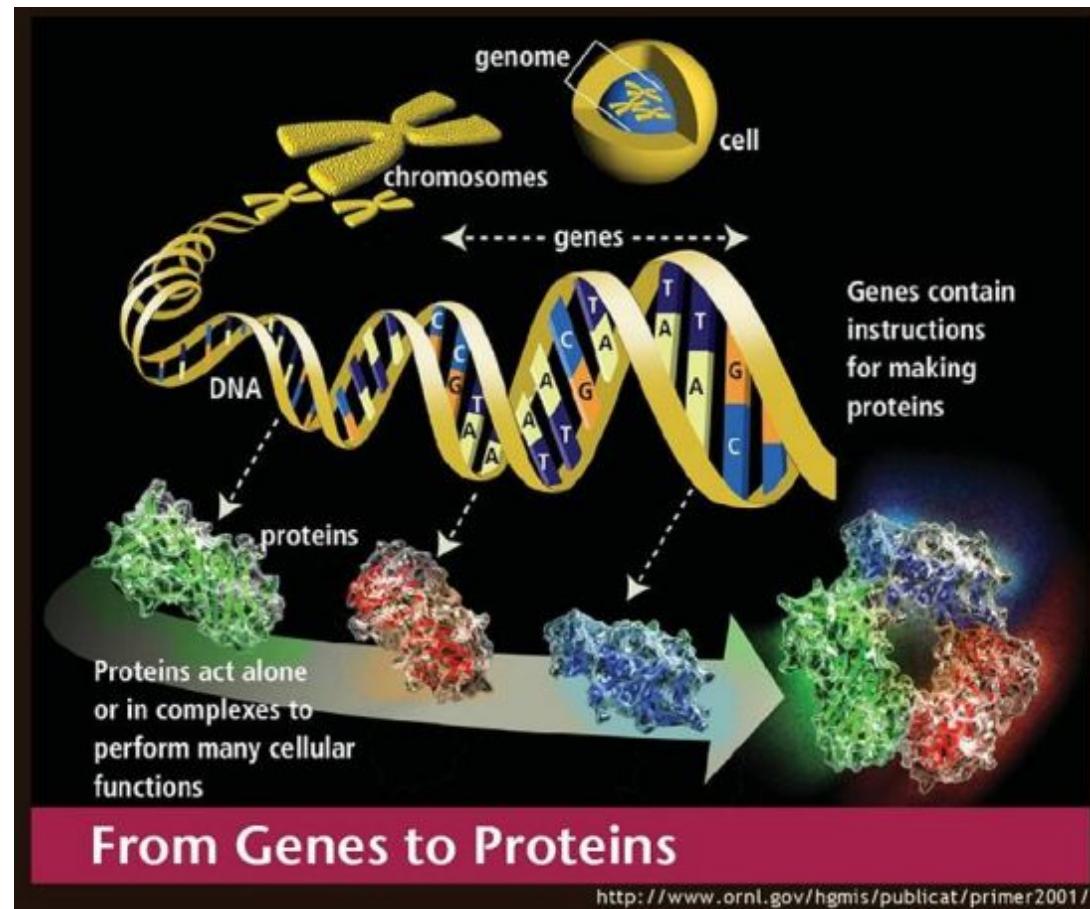


# Neurobiologie

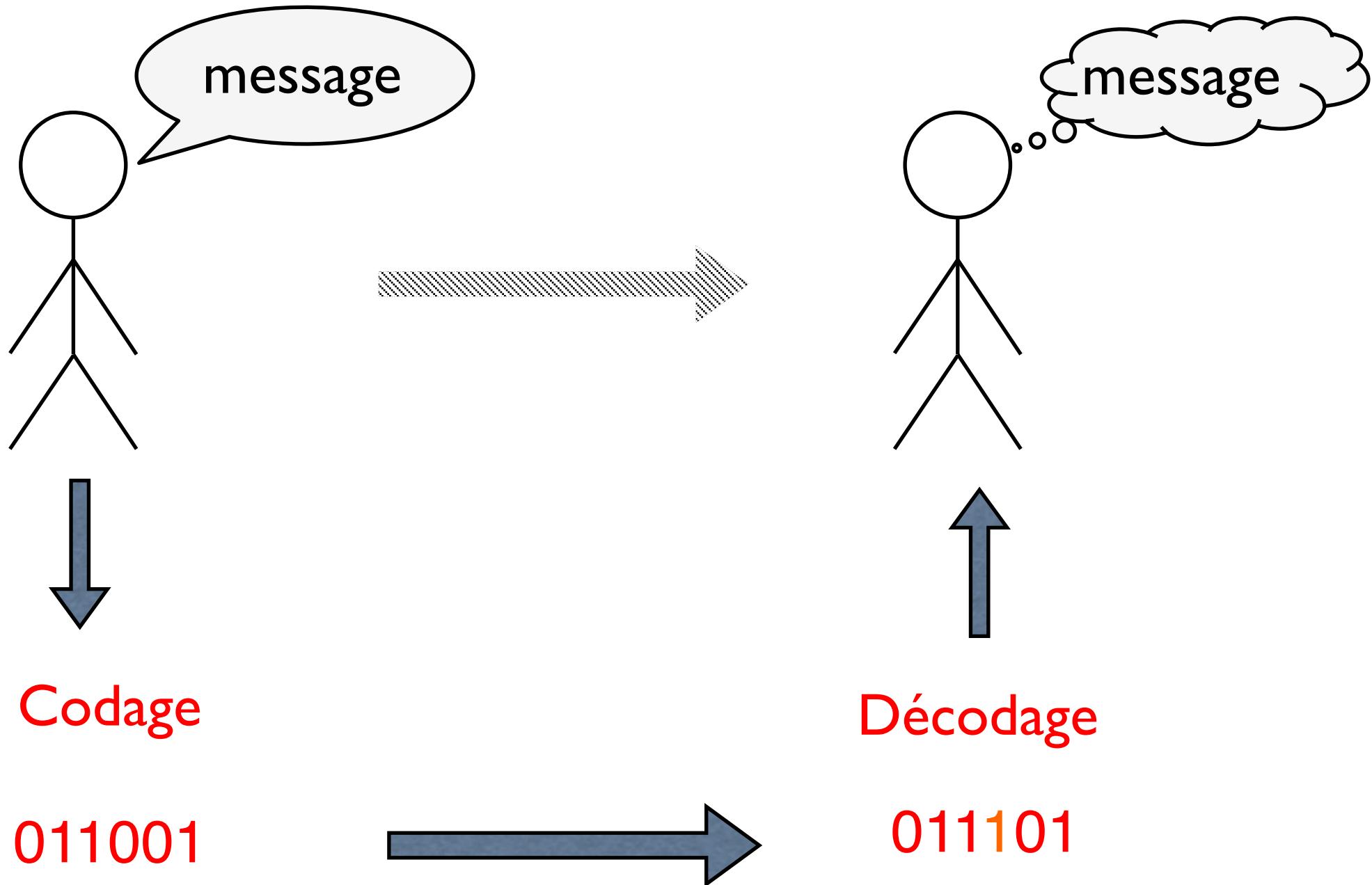


# Transmission d'information en biologie

## Génétique et Biologie Moléculaire



# Codage et décodage dans la communication

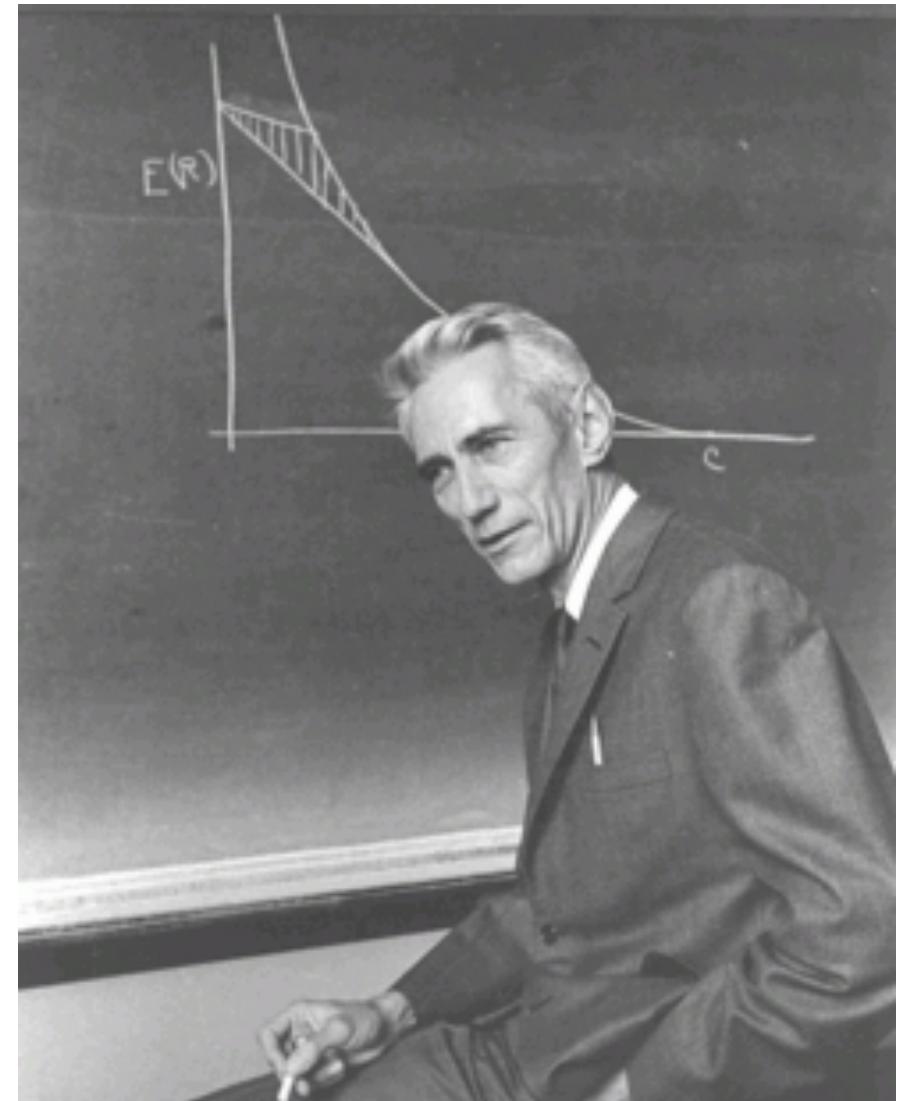


# Théorie de la communication

Claude Shannon  
(1916-2001)

Pendant la guerre :  
services secrets,  
cryptographie

→ “A mathematical theory  
of communications”  
(1948)



1948 : 1800 conversations sur un câble  
2005 : 6 400 000 conversations sur fibre optique

# Théorie de l'information

Une branche de la science:

- \* Récente (60 ans)
- \* Très importante du point de vue technologique
- \* Aux multiples ramifications
- \* ... et intimement reliée à la Physique Statistique

# Communication

Symboles ou “mots” à transmettre:  $M_1, M_2, \dots, M_N$

Probabilités (fréquence d’ apparition):  $p_1, p_2, \dots, p_N$

**Codage:**  $M_n \rightarrow C_n = 0110010110$

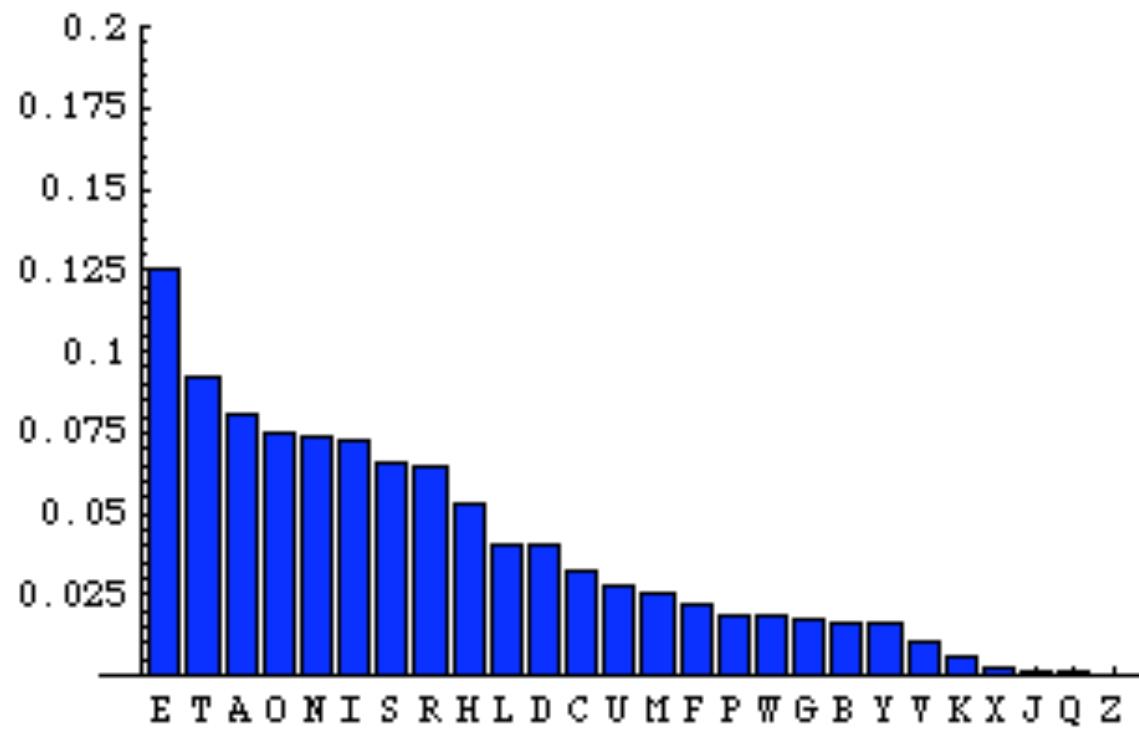
Pb: trouver un codage ( $= \{C_n\}$ ) tel que le nombre de bits utilisé en moyenne soit le plus petit possible.

**Principe:** utiliser des  $C_n$  courts pour les mots fréquents

# Exemple : alphabet Morse (1832)



A	• -
B	- - . .
C	- - . - .
D	- - . .
E	•
F	• - - .
G	- - - .
H	. . . .
I	. .
J	. - - -
K	- - .
L	. - - .
M	- -
N	- .
O	- - -
P	• - - .
Q	- - - . -
R	- - . -
S	. . .
T	-



Fréquence des lettres  
en anglais

# Un code simple

Code A

mots	symboles
$M_1$	11
$M_2$	10
$M_3$	01
$M_4$	00

codage



1101100010 ...

$M_1 M_3 M_2 M_4 M_2 \dots$



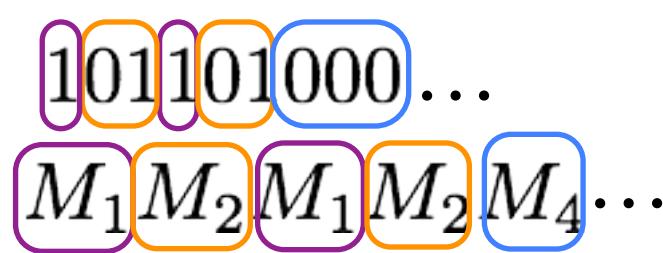
décodage

# Un code de longueur variable

Code B

mots	symboles
$M_1$	1
$M_2$	01
$M_3$	001
$M_4$	000

codage



décodage

NB : pas d'ambiguité car aucun mot du code n'est le préfixe d'un autre

# Quel est le meilleur 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

Longueur moyenne d'un mot codé :  $\langle L \rangle = \sum_n p_n L(M_n)$

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

Code B : Dépend des  $p_n$

## Deux exemples avec le code B

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


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

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

Code B

$M_1$	1
$M_2$	01
$M_3$	001
$M_4$	000


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

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

# Longueur minimale du message

## Théorème de Shannon

La longueur moyenne minimale du message, par mot envoyé, est égale à l'entropie  $H$  de la loi de probabilité  $\{p_i\}$  donnant la fréquence des mots  $\{M_i\}$

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

- Preuve : D.J.C. MacKay. Information theory, inference, and learning algorithms, CUP (2003), <http://www.inference.phy.cam.ac.uk/itprnn/book.pdf>  
Kim Boström, <http://www.itglitz.in/ICT/Shannon.pdf>
- Méthodes de compression en pratique? (gzip, jpeg, MP3, ...)

## Longueur minimale

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

### Exemples

Dé à 4 faces:  $p_1 = p_2 = p_3 = p_4 = 1/4$



$$H = -4[(1/4) \log_2(1/4)] = 2 \quad H = 2$$

---

Dé à 4 faces pipé:  $p_1 = 1/2, p_2 = 1/4, p_3 = p_4 = 1/8$

$$H = (1/2) \log_2 2 + (1/4) \log_2 4 + 2[(1/8) \log_2 8]$$

$$= 1/2 + 1/2 + 2(3/8) = 7/4$$



$$H = 7/4$$

# Les dés

	$H$	Code A	Code B
Normal	2	$\langle L \rangle = 2$	$\langle L \rangle = \frac{9}{4}$
Pipé	$\frac{7}{4}$	$\langle L \rangle = 2$	$\langle L \rangle = \frac{7}{4}$
		$\langle L \rangle_{\text{mini}} = H$	

## Un autre exemple ...

L'amphi  
d'aujourd'hui, ça se  
comprendait ?

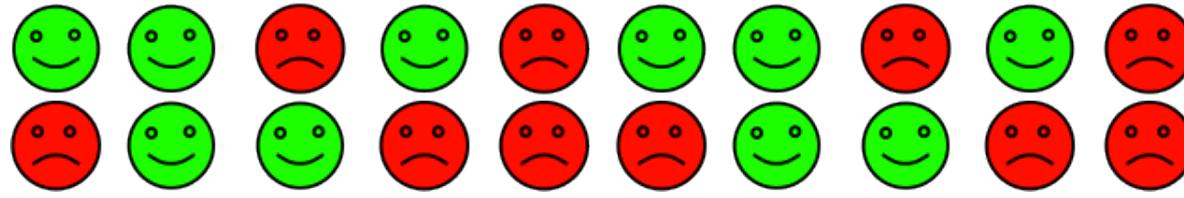


Délégué

Amphi



ou



Dites moi qui a des questions...

# Le bon amphi ...



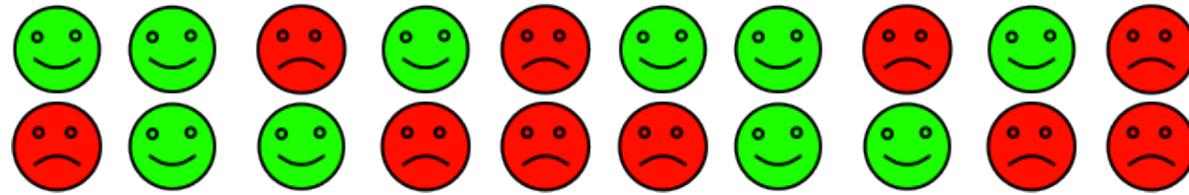
Probabilité de mécontent:  $p \sim \frac{1}{N}$

Information à transmettre :  
matricule du mécontent



$$H = \log_2 N \text{ bits}$$

# Le mauvais amphi ...



Probabilité de mécontent:  $p \sim \frac{1}{2}$

Information à transmettre :  
0010100101...



H = N bits

Amphi



$$H = \log_2 N \text{ bits}$$

ou



$$H = N \text{ bits}$$

Longueur du message minimal

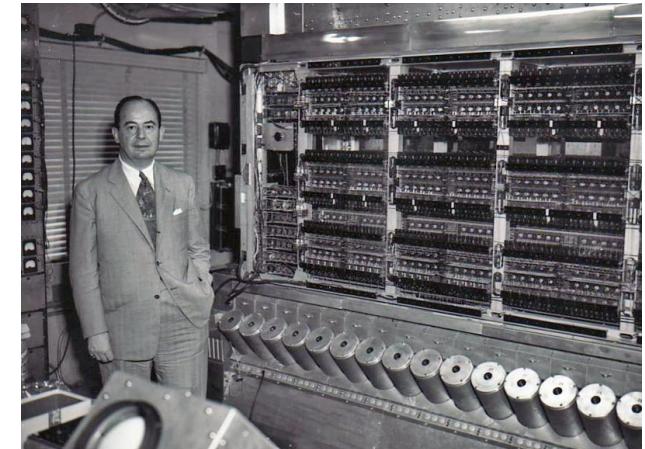
= nombre de bits pour écrire ou transmettre l'information

# Quel nom donner à

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



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).

# Propriétés mathématiques de .....

... l'entropie

... l'incertitude

... l'information manquante

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

# Information manquante

Définition. L'information manquante associée à la loi de probabilité  $p = p_1, p_2, \dots, p_N$  est la longueur moyenne minimale du message nécessaire pour spécifier un évènement  $x = x_1, x_2, \dots, x_N$

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

Expression générale de l'entropie, pour tous les ensembles (canonique, microc., grand-canonique..) (A3):

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

L'entropie est l'information manquante de la loi de probabilité physique sur les configurations (avec un changement d'unité de mesure)

$H$  en "bits"

$$k = 1 / \log 2$$

$$k = 1.4 \cdot 10^{-23} J/K$$

## Justification intuitive

On apprend une nouvelle. Si cette nouvelle était *a priori* improbable : gain d'une grande information. Si cette nouvelle était *a priori* très probable : gain d'une faible information

Shannon : si la nouvelle avait probabilité  $p$ , l'information gagnée lorsque l'on reçoit la nouvelle est  $\log_2(1/p)$

Soient  $N$  événements, probabilités  $p_1, \dots, p_N$ , avec  $\sum_{i=1}^N p_i = 1$

L'information moyenne reçue si on fait un tirage au hasard vaut

$$H = \sum_{i=1}^N p_i \log_2(1/p_i) = - \sum_{i=1}^N p_i \log_2 p_i$$

## Propriétés mathématiques

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

- $H \geq 0$
- $H = \log_2 N$  maximale si et seulement si évènements équiprobables :  $p_i = \frac{1}{N}$
- $H = 0$  si et seulement si évènement certain :  $p_i = \delta_{i,n}$

# Propriétés mathématiques

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

## ⌚ Paire d'évènements indépendants

“Le cheval 12 a gagné dans la troisième course;

$$H = ? = H_1 + H_2$$

Le numéro gagnant du loto est 137 “

Calcul de l'entropie :  $P(n, a) = p_n q_a$

$$\begin{aligned} H(\{p_n q_a\}) &= - \sum_{n,a} p_n q_a \log_2 (p_n q_a) \\ &= - \sum_{n,a} p_n q_a (\log_2 p_n + \log_2 q_a) \end{aligned}$$

$$H(\{p_n q_a\}) = H(\{p_n\}) + H(\{q_a\})$$

# Propriétés mathématiques

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

## ⌚ Paire d'évènements dépendants

“Le cheval 12 a gagné dans la troisième course;  
Le cheval 3 est arrivé deuxième”

Loi jointe :  $P(n, a)$

Lois marginales:  $p_n = \sum_a P(n, a)$  ;  $q_a = \sum_n P(n, a)$

$$H = - \sum_{n,a} P(n, a) \log_2 P(n, a) < H(\{p_n\}) + H(\{q_a\})$$

## Démonstration

$$H(\{p_n\}) + H(\{q_a\}) - H(\{P(n,a)\}) = - \sum_{n,a} P(n,a) \log_2 \left( \frac{p_n q_a}{P(n,a)} \right)$$

Inégalité de Jenssen (concavité du log) :

$X$  variable aléatoire       $\langle \log X \rangle \leq \log \langle X \rangle$

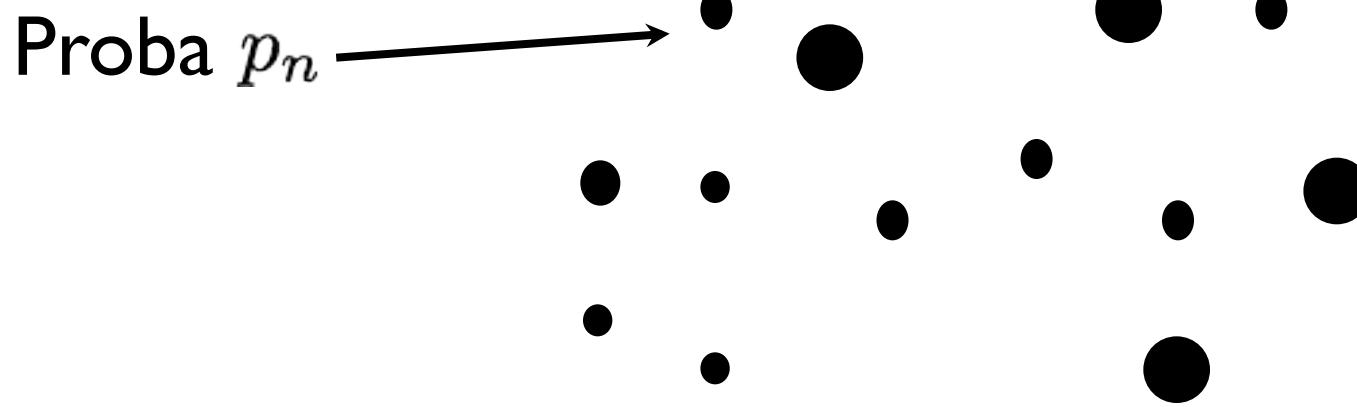
Ici :  $X(n,a) = \frac{p_n q_a}{P(n,a)} \Rightarrow \langle X \rangle = 1 \Rightarrow -\langle \log X \rangle \geq 0$

Paire d'évènements  
dépendants

$$H < H(\{p_n\}) + H(\{q_a\})$$

## Propriétés mathématiques

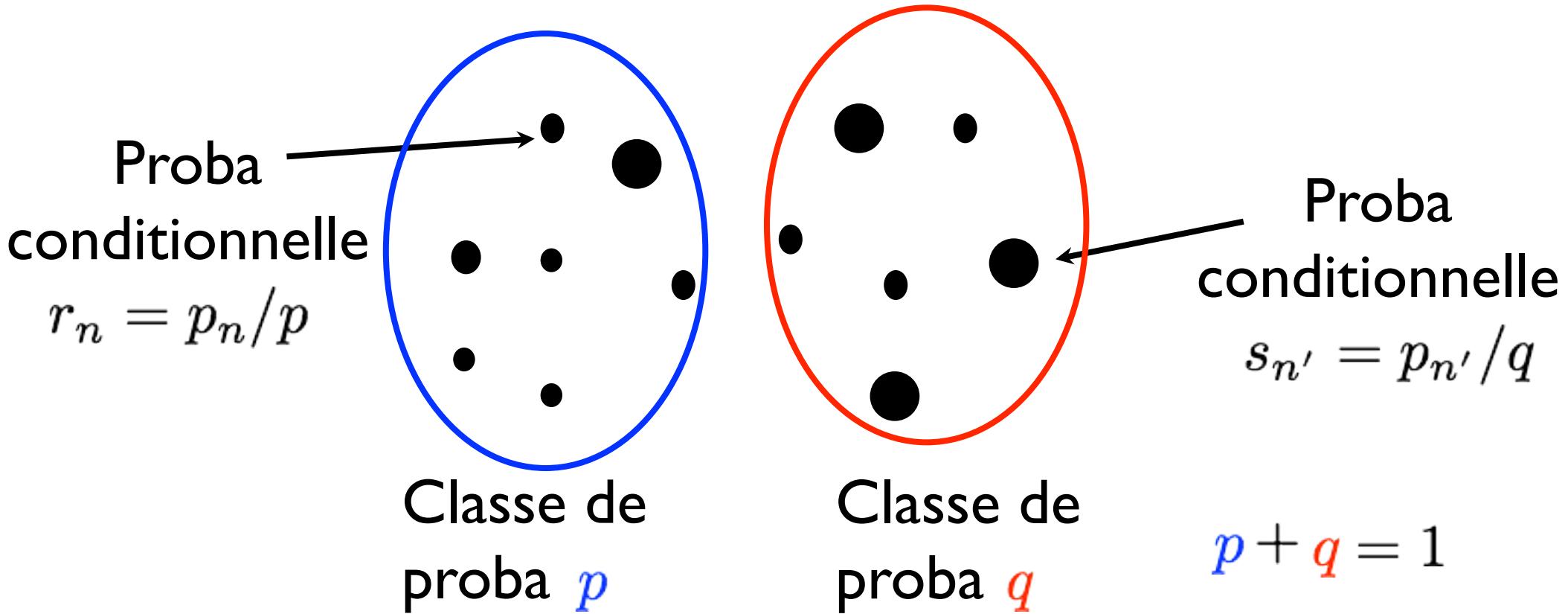
$$H = - \sum_{i=1}^N p_i \log_2 p_i$$



$$H(\{p_n\}) = - \sum_n p_n \log_2 p_n$$

# Propriétés mathématiques

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



$$H(\{p_n\}) = - \sum_n p_n \log_2 p_n$$

$$= -p \log_2 p - q \log_2 q + p H(\{r_n\}) + q H(\{s_{n'}\})$$

Ces propriétés définissent  
l'information manquante d'une loi de  
probabilité de façon unique  
(à la normalisation près)

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

Information Entropie  
manquante

# **Principe d'entropie maximale**

## **(Principe d'information manquante maximale)**

# Dé pipé avec un biais inconnu

Quelqu'un lance un dé à 4 faces, numérotées 1,2,3,4, un très grand nombre de fois et me communique seulement la moyenne du nombre indiqué par les faces



Dé non pipé  
Moyenne = 2.5



Dé pipé  
Moyenne = 3.3

Je sais que

$$\begin{cases} p_1 + p_2 + p_3 + p_4 = 1 \\ p_1 + 2 p_2 + 3 p_3 + 4 p_4 = 3.3 \end{cases}$$

Que valent  $p_1, p_2, p_3, p_4$  ? Infinité de solutions ....

$p_1=0, p_2=0, p_3=0.7, p_4=0.3$

$p_1=0.1, p_2=0.15, p_3=0.1, p_4=0.65$

$p_1=0.0609, p_2=0.1268, p_3=0.2637, p_4=0.5486$

?

Quel est le choix le moins arbitraire parmi toutes ces solutions?

# Principe d'entropie maximale

les contraintes doivent être aussi peu informatives que possible sur la distribution inconnue

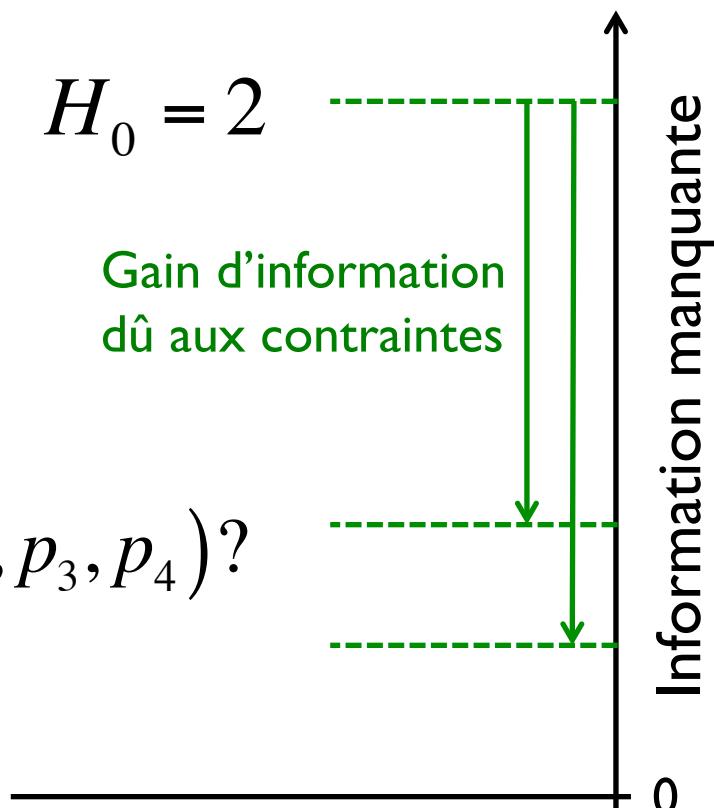
Dé  
non biaisé



$$: p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$$

$$H_0 = 2$$

Gain d'information  
dû aux contraintes



Dé pipé



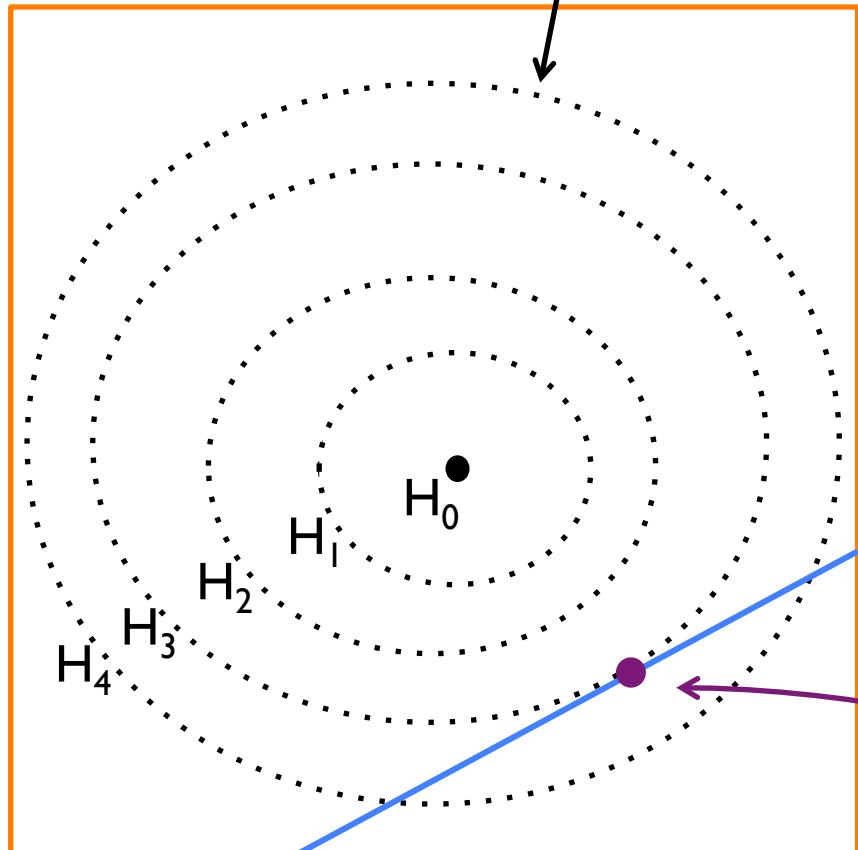
$$: p_1, p_2, p_3, p_4 ? \quad H(p_1, p_2, p_3, p_4) ?$$

→ On cherche à maximiser  $H$  étant donné les contraintes

# Principe d'entropie maximale

Lignes de niveau de  $H$  = fréquences  $(p_1, p_2, p_3, p_4)$  ayant la même entropie

$$H_0=2 > H_1 > H_2 > H_3 > \dots > 0$$



espace des  $(p_1, p_2, p_3, p_4)$

Contraintes

$$\sum_{k=1,2,3,4} kp_k = 3.3$$

$$\sum_{k=1,2,3,4} p_k = 1$$

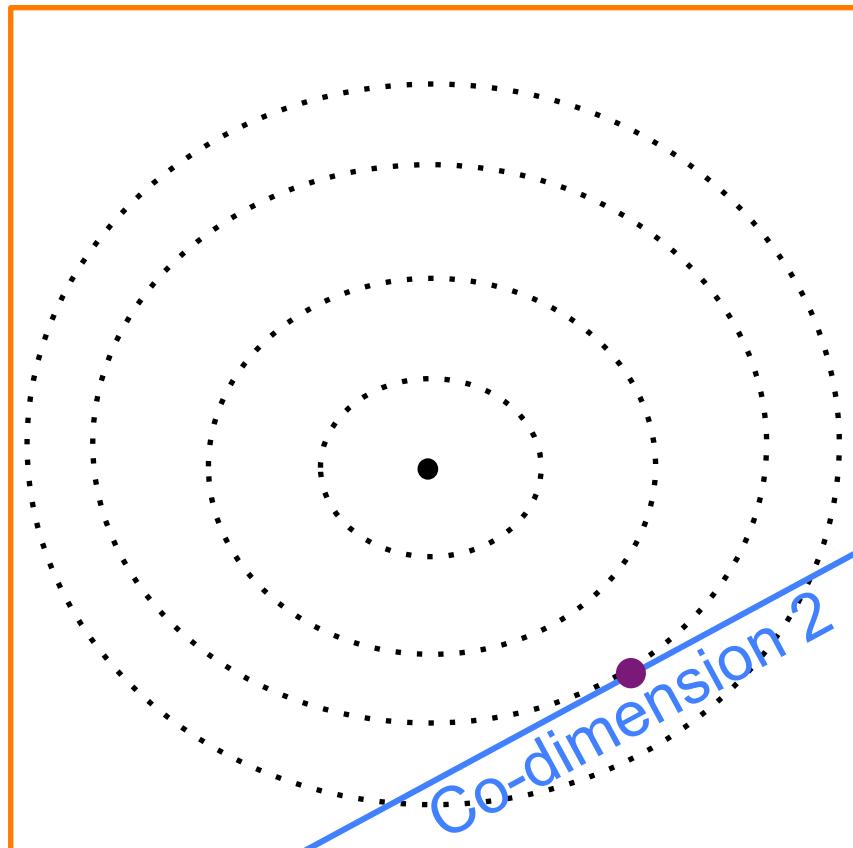
Maximisation de  $H$   
sous les contraintes



# Maximisation sous contraintes

Entropie

$$H = - \sum_{k=1,2,3,4} p_k \log_2 p_k$$



Espace Dimension 4

Contraintes

$$g_1 = \sum_{k=1,2,3,4} k p_k - 3.3$$

$$g_2 = \sum_{k=1,2,3,4} p_k - 1$$

Les lignes de niveau de  $H$  sont tangentes à la surface

$$g_1 = g_2 = 0$$

$$\Rightarrow \exists \lambda_1, \lambda_2 : \vec{\nabla} H = \lambda_1 \vec{\nabla} g_1 + \lambda_2 \vec{\nabla} g_2$$

(Multiplicateurs de Lagrange)

# Maximisation sous contraintes

Entropie

$$H = - \sum_{k=1,2,3,4} p_k \log_2 p_k$$

Contraintes

$$g_1 = \sum_{k=1,2,3,4} k p_k - 3.3$$

$$g_2 = \sum_{k=1,2,3,4} p_k - 1$$

$$\exists \lambda_1, \lambda_2 : \vec{\nabla} H = \lambda_1 \vec{\nabla} g_1 + \lambda_2 \vec{\nabla} g_2$$

$$-\log_2 p_k - \frac{1}{\log 2} = \lambda_1 k + \lambda_2 1 \Rightarrow p_k = \frac{x^k}{x + x^2 + x^3 + x^4}$$

x=2.0804...:  $p_1=0.0609, p_2=0.1268, p_3=0.2637, p_4=0.5486$  ( $H=1.6060\dots$ )

$p_1=0, p_2=0, p_3=0.7, p_4=0.3$

$p_1=0.1, p_2=0.15, p_3=0.1, p_4=0.65$

( $H=0.8813\dots$ )

( $H=1.4739\dots$ )

**Information  
et  
Ensembles en physique statistique**

# Information manquante et physique statistique

Une autre approche de la physique statistique : construire la loi de probabilité des états,  $\{p_n\}$ , qui maximise l'entropie compte tenu de ce qu'on sait sur le système.

“Loi la moins biaisée”

Exemple : système échangeant de l'énergie avec un thermostat. On impose deux contraintes :

$$\sum_n p_n = 1 \quad \sum_n p_n E_n = U$$

Pb: trouver la loi  $\{p_n\}$  qui maximise  $H = - \sum_n p_n \log_2 p_n$   
compte tenu des contraintes

# Information manquante et physique statistique

Maximiser  $H = - \sum_n p_n \log_2 p_n$  avec les deux contraintes:

$$g_1(p_1, \dots, p_N) = \sum_n p_n - 1 = 0$$

$$g_2(p_1, \dots, p_N) = \sum_n p_n E_n - U = 0$$

Lagrange:

$$\exists \lambda_1, \lambda_2 / \vec{\nabla} H - \lambda_1 \vec{\nabla} g_1 - \lambda_2 \vec{\nabla} g_2 = 0$$

Ensemble canonique

$$\dots \longrightarrow p_n = \frac{1}{Z} e^{-\beta E_n}$$

où  $\beta, Z$  sont tels que  $\sum_n p_n = 1$  et  $\sum_n p_n E_n = U$

# Information manquante et physique statistique

Caveat : les contraintes doivent être bien définies!

1. Exemple :  $E$  fixée en moyenne  $\rightarrow$  canonique  
 $E$  fixée strictement  $\rightarrow$  micro-canonique

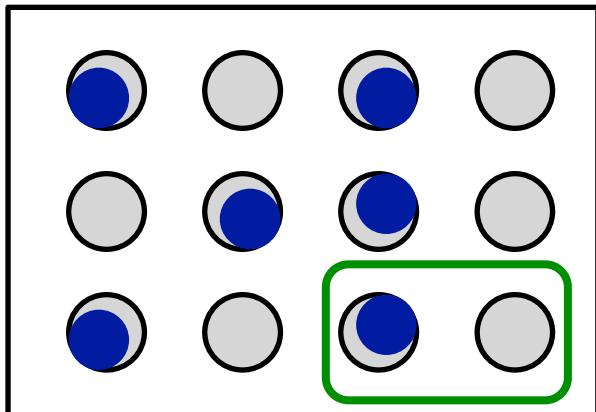
Contrainte :  $p_i = 0$  si  $E_i \neq E$ ,

$$\max_{\{p_i\}} H = - \sum_{i:E_i=E} p_i \log_2 p_i \rightarrow p_i = \frac{1}{\Omega}$$

2. On sait que le micro-canonique ‘donne’ le canonique seulement à la limite thermodynamique. Où est cachée cette limite dans l’application du principe d’entropie maximale?

# Information manquante et physique statistique

Dans les contraintes!

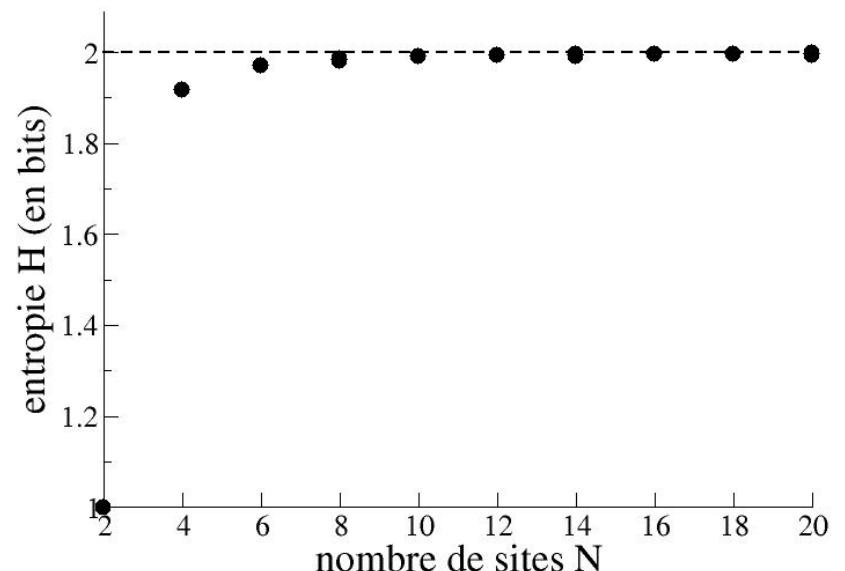


Système de  $N$  sites , sans interaction,  
énergie fixée de telle sorte que  
 $\frac{1}{2}$  des sites sont occupés (= 

Sous-système de 2 sites : Entropie Maximale  $\rightarrow p_{00}=p_{01}=p_{10}=p_{11}=\frac{1}{4}$   
 $\rightarrow H=2$

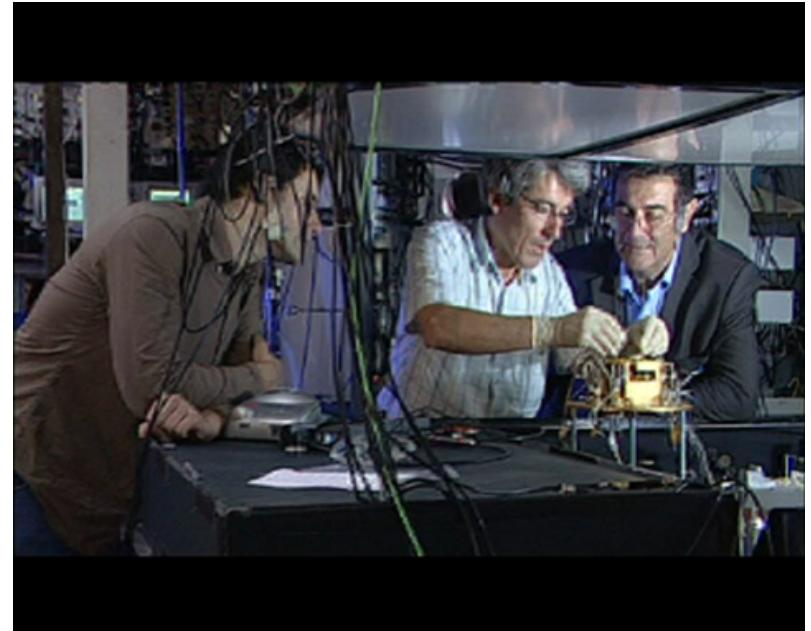
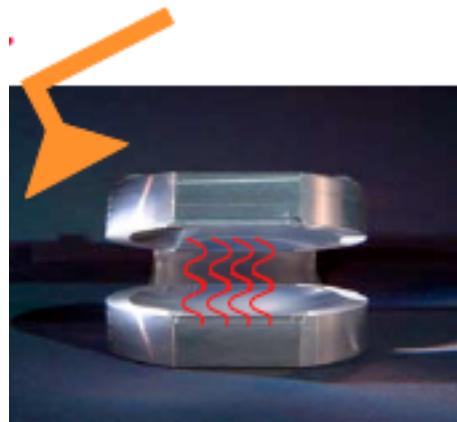
$$\Omega = \binom{N}{\frac{N}{2}} \quad \Omega_k = \binom{N-2}{\frac{N}{2}-k}$$

$$p_{00} = \frac{\Omega_0}{\Omega}, p_{01} = p_{10} = \frac{\Omega_1}{\Omega}, p_{11} = \frac{\Omega_2}{\Omega}$$



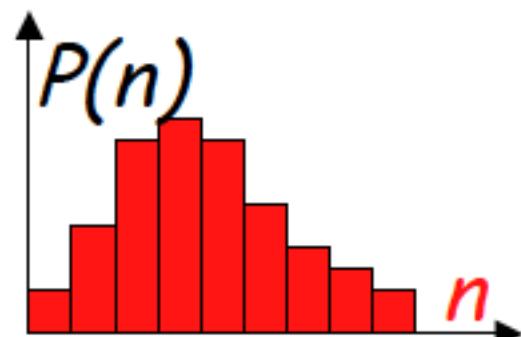
# **Application à l'analyse et à la conception d'expériences**

# Comptage non-destructif de photons dans une cavité micro-ondes



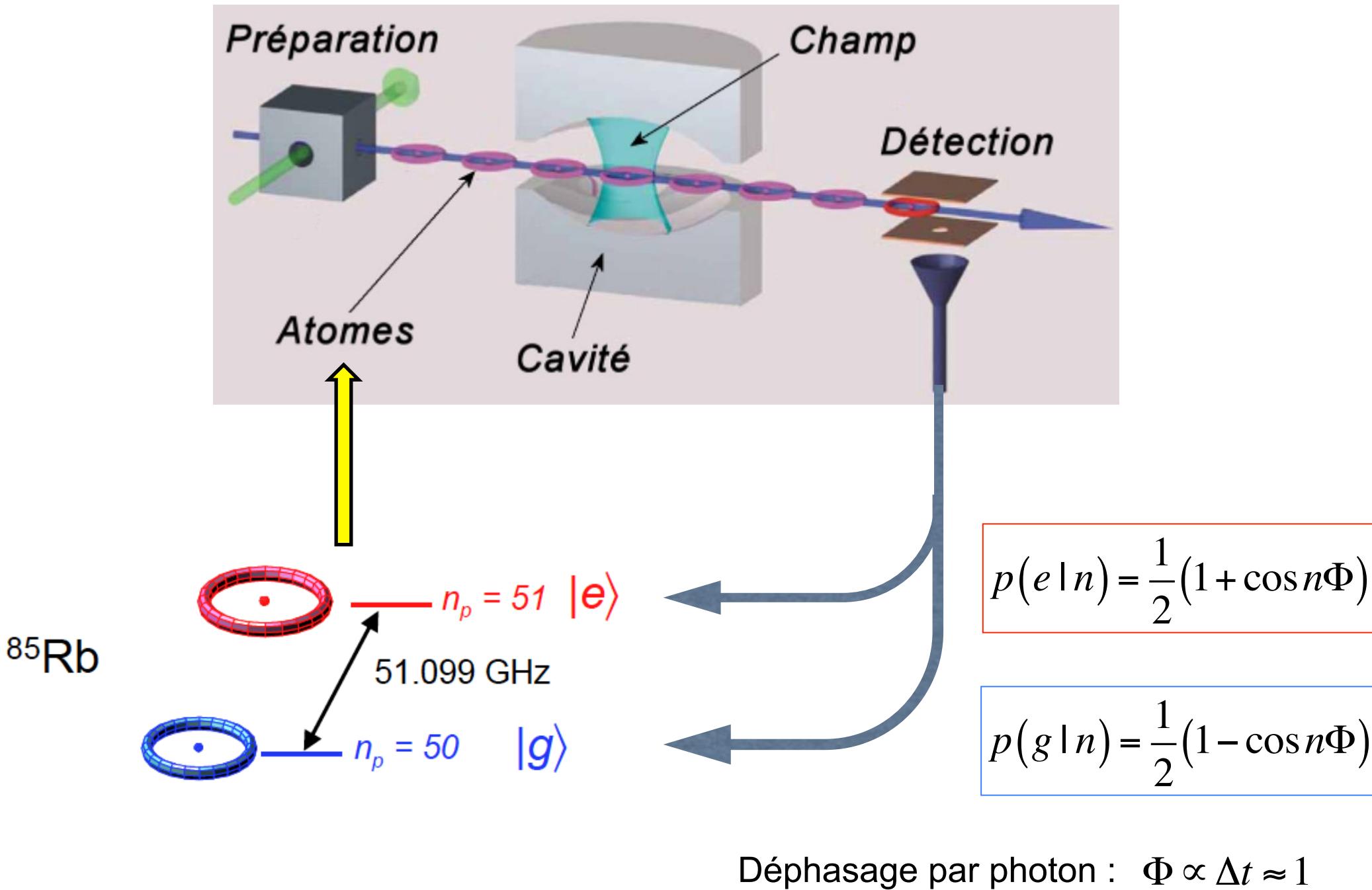
S. Haroche (Prix Nobel 2012), M. Brune

$$|\Psi_{champ}\rangle = \sum_n C_n |n\rangle \quad ; \quad P(n) = |C_n|^2 : \text{ probabilité de } n \text{ photons}$$



Loi de Poisson, moyenne inconnue mais petite

# Atomes sondes : déphasage induit par les photons



# Inférence du nombre de photons

Résultats après M atomes :  $a_1=e, a_2=g, a_3=e, a_4=e, \dots, a_M=g$  A

Probabilité de A étant donné n :

$$P_M(A|n) = \left( \frac{1 + \cos n\Phi}{2} \right)^{\#e} \left( \frac{1 - \cos n\Phi}{2} \right)^{\#g}$$

Probabilité de n a priori :

$$P_{prior}(n) = \begin{cases} \frac{1}{8} & \text{si } 0 \leq n \leq 7, \\ 0 & \text{sinon} \end{cases}$$

Probabilité de n étant donné A?

Probabilité jointe :

$$P_M(n, A) = P_M(A|n)P_{prior}(n) = P_M(n|A)P_M(A)$$

Loi de Bayes :

$$P_M(n|A) = \frac{P_M(A|n)P_{prior}(n)}{P_M(A)}$$

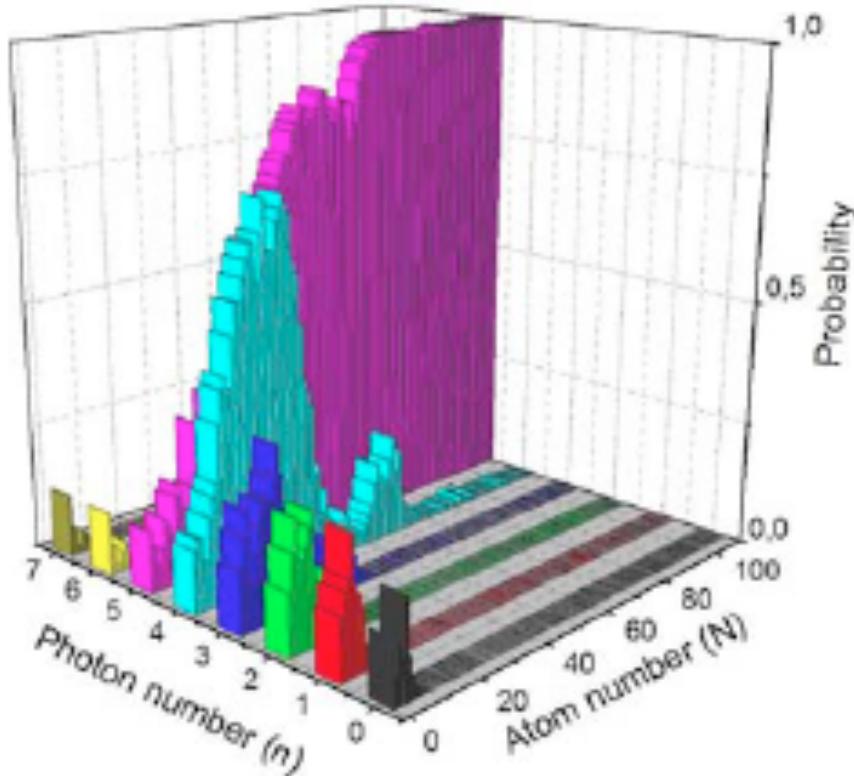
# Inférence du nombre de photons

Modèle physique ↘

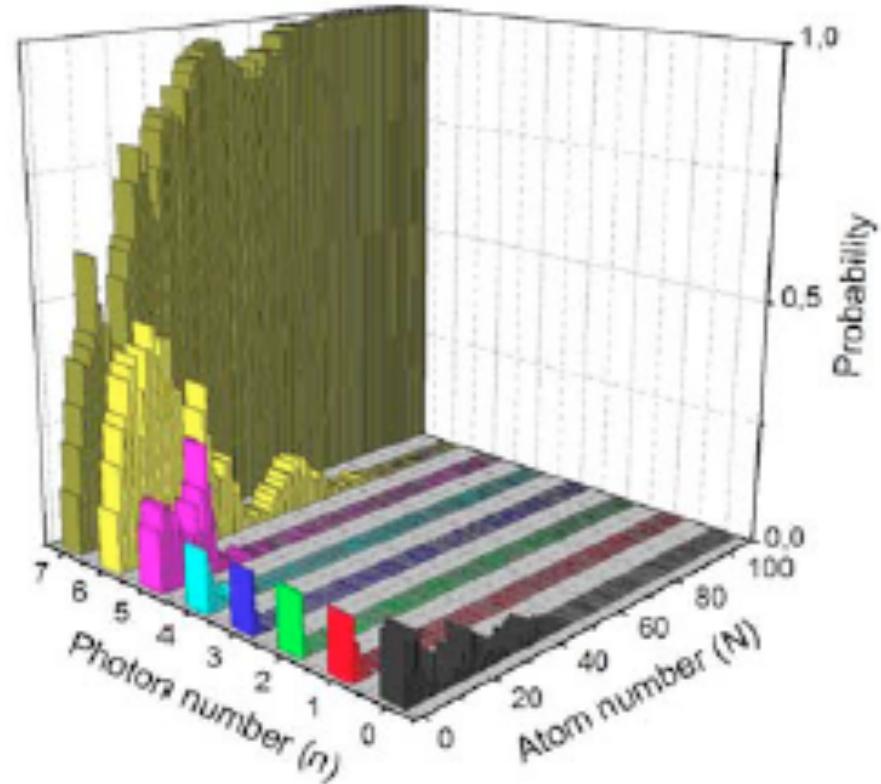
A priori uniforme entre 0 et 7 ↘

$$P_M(n|A) = \frac{P_M(A|n)P_{prior}(n)}{P_M(A)}$$

↑ normalisation



'Collapse' vers  $n=5$

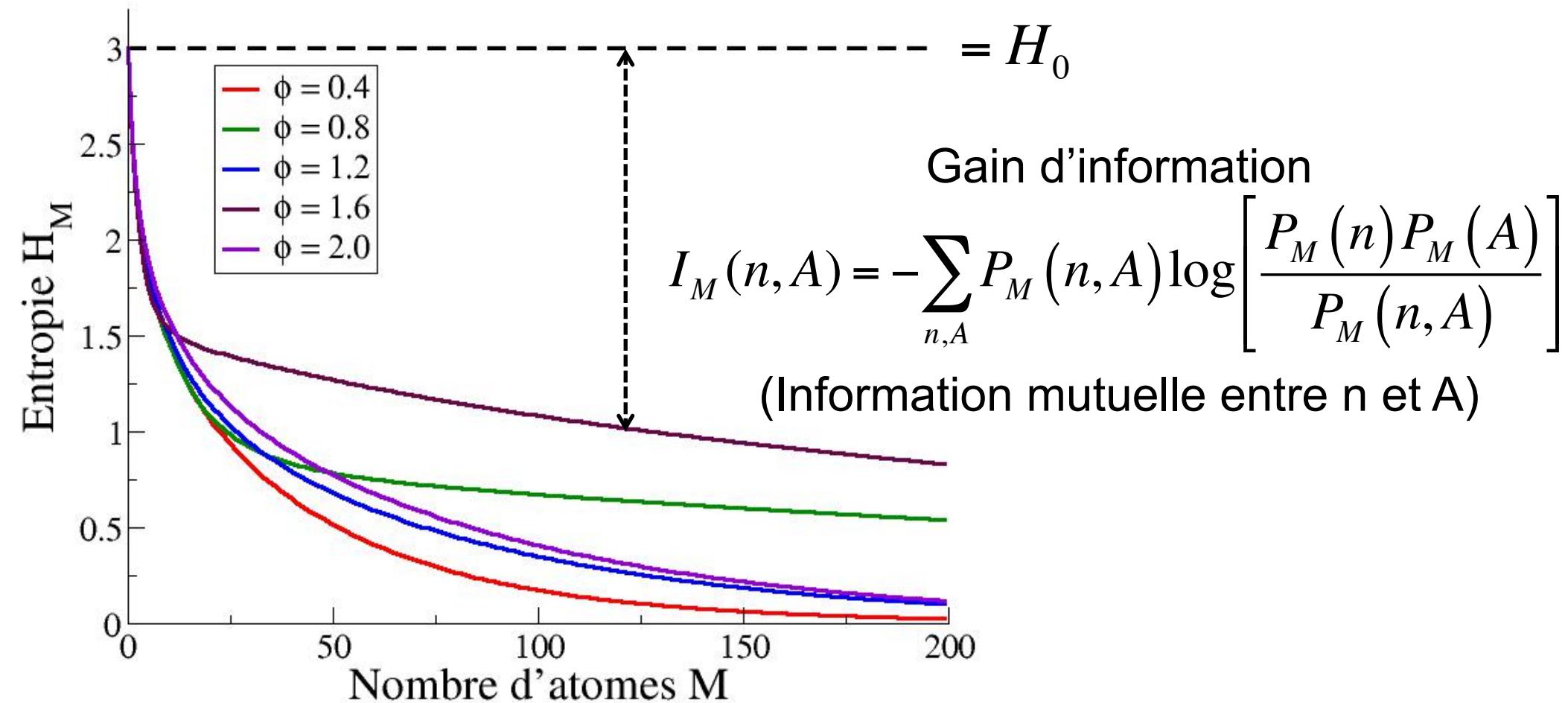


'Collapse' vers  $n=7$

2000 expériences, chacune avec 110 atomes (Poisson de moyenne 3.43)

# Information manquante après M mesures (atomes)

$$H_M = \sum_A P_M(A) \left[ - \sum_n P_M(n|A) \log P_M(n|A) \right]$$

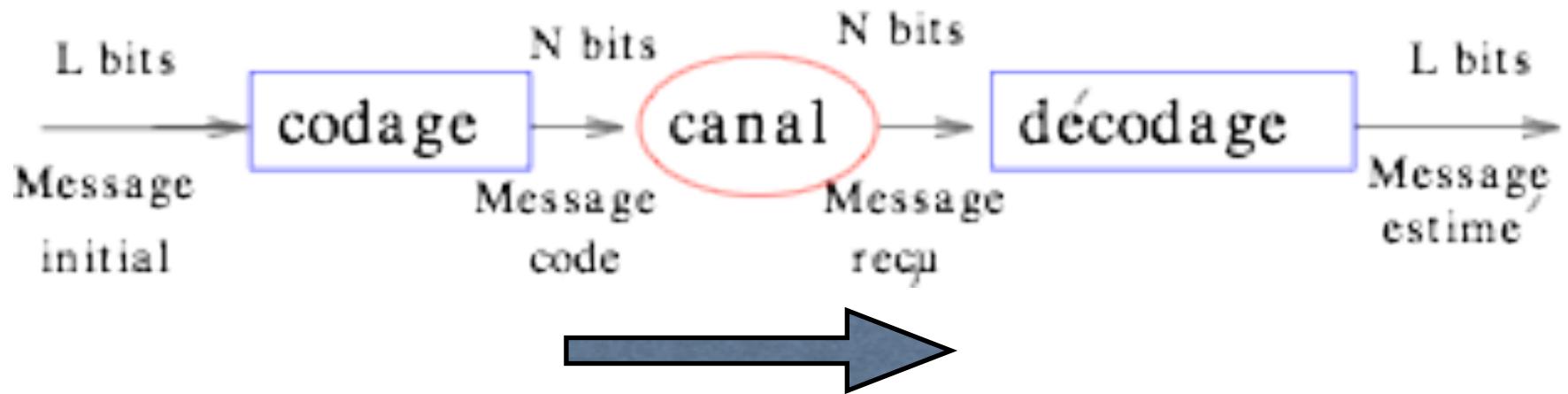


Utile pour concevoir des protocoles expérimentaux optimaux !

# **Communication en présence d'erreurs et Transitions de phases**

# Transmission d'information: correction d'erreurs

Principe : codage, transmission, décodage



Codage : introduire de la redondance  $N > L$

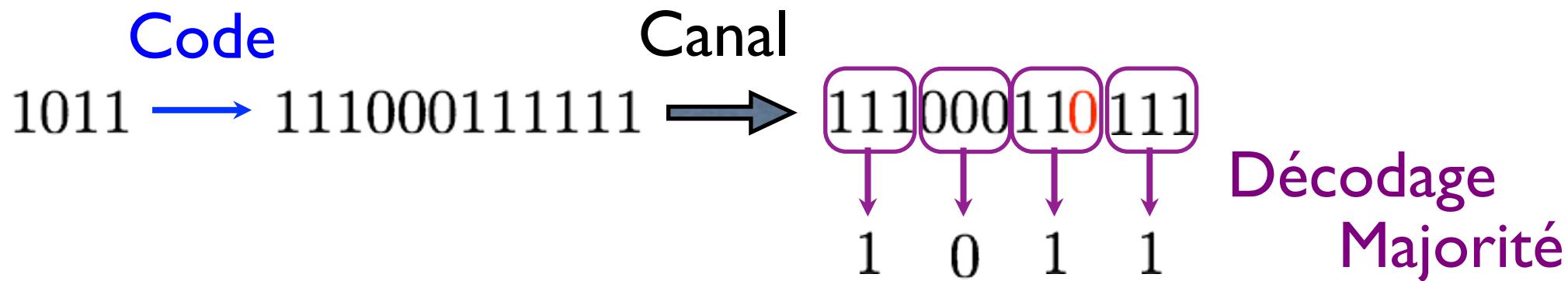
# Code correcteurs d'erreurs

Bruit de transmission : chaque bit est retourné avec proba  $p$

Canal 0  0 avec proba  $1-p$ , 1 avec proba  $p$

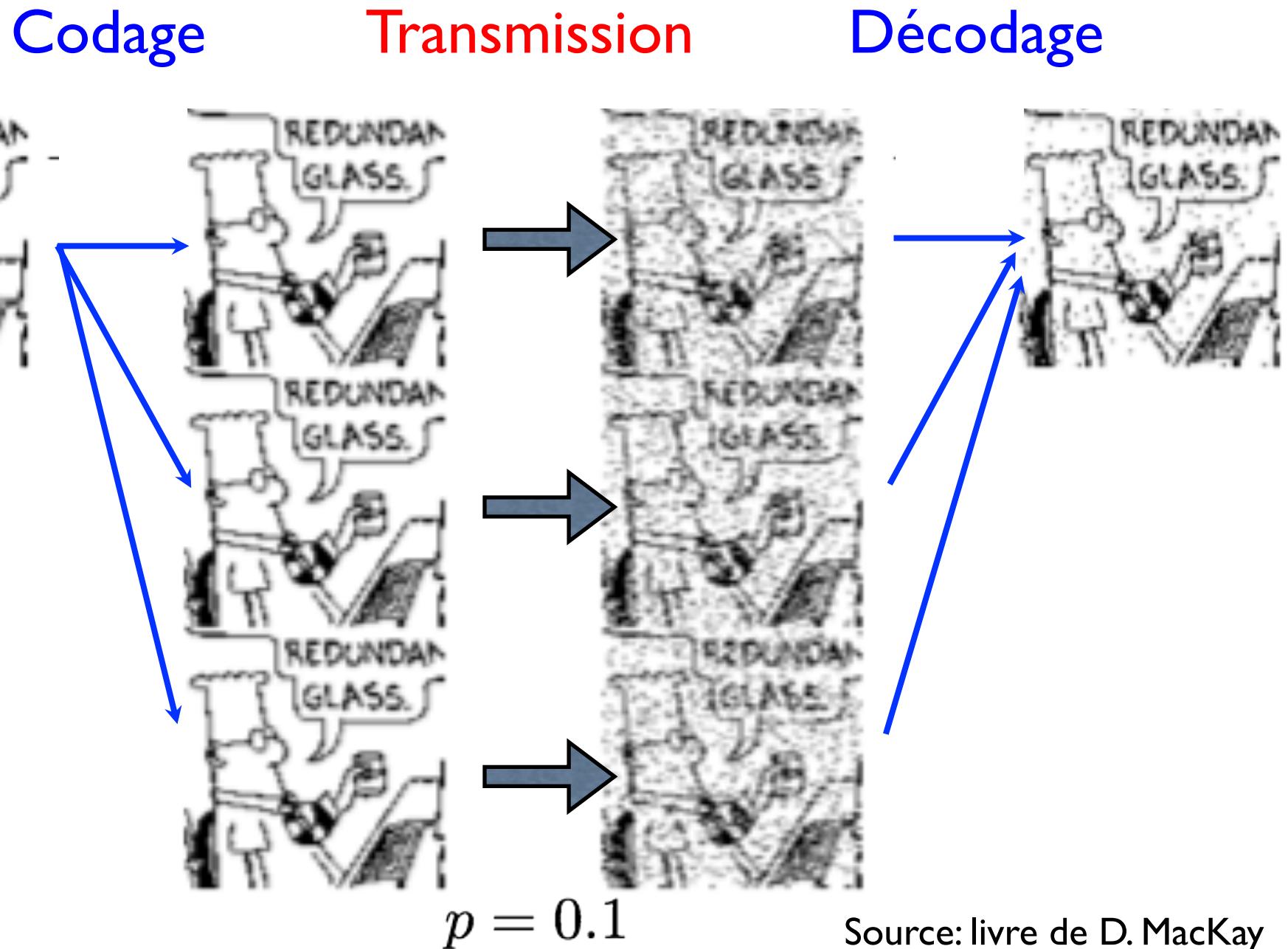
# Codage : introduire de la redondance

Exemple le plus simple : code par répétition {  $0 \rightarrow 000$   
 $1 \rightarrow 111$



**Compromis redondance-performance** : envoyer 3 fois plus de bits pour avoir une probabilité d'erreur  $p^2$  (au lieu de  $p$ )

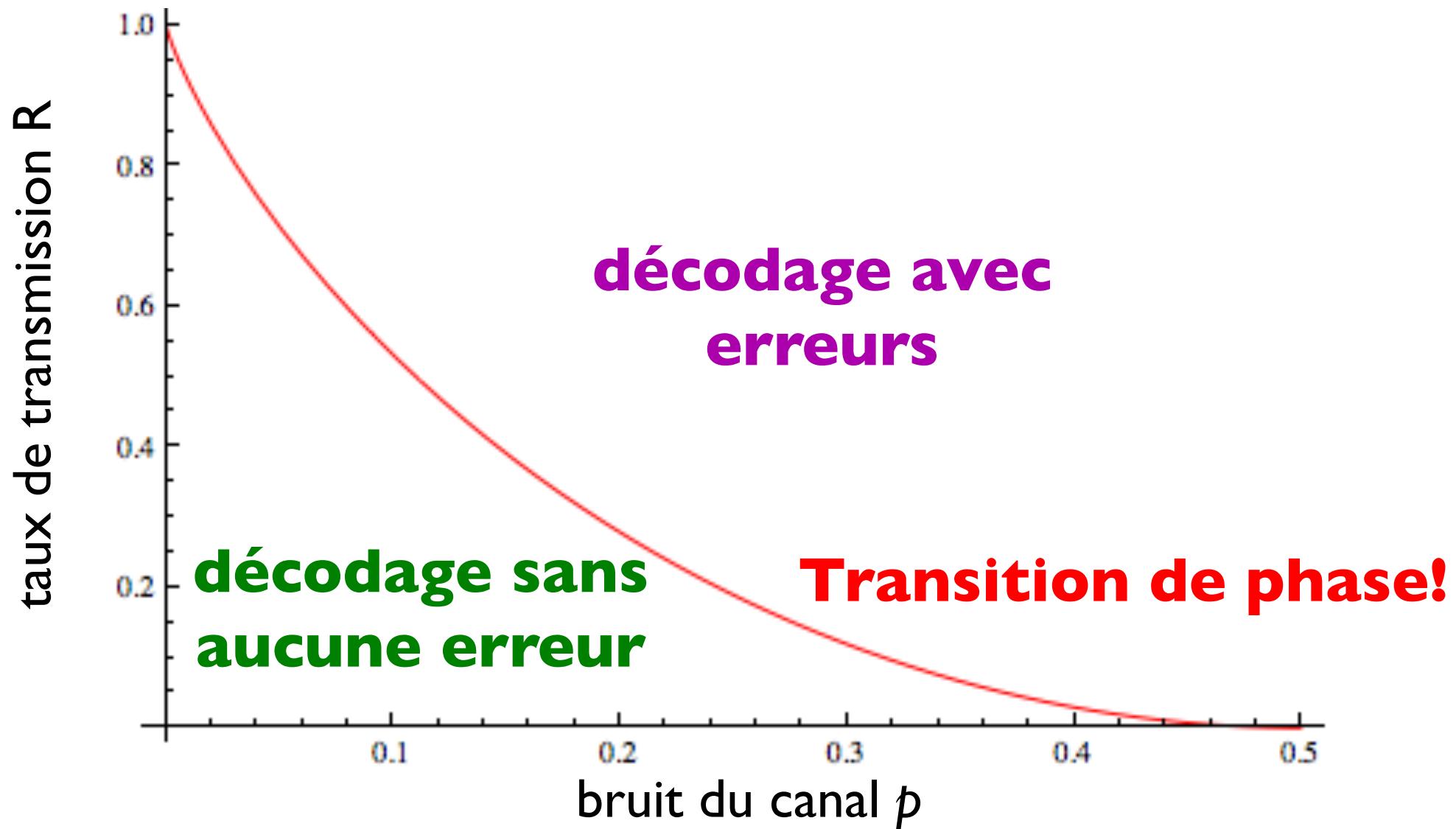
Codage par répétition,



Source: livre de D. MacKay

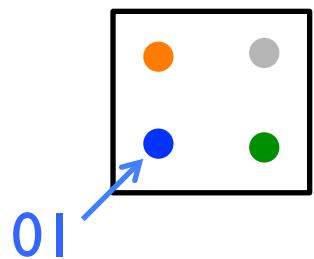
# Codes correcteurs d'erreurs

Mais on peut faire beaucoup mieux! (Shannon, 1948)



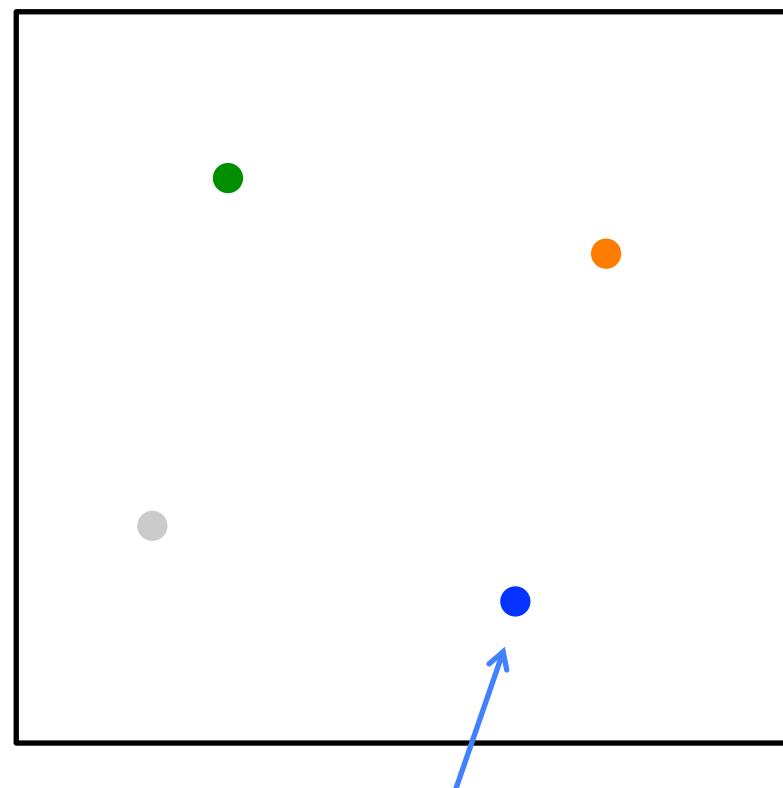
# Le code aléatoire de Shannon

$2^L$  messages équiprobables de longueur  $L$



codage  
taux  $R = \frac{L}{N} < I$

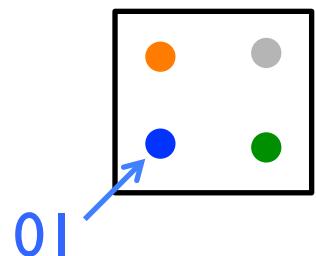
$2^L$  mots de code aléatoires parmi  $2^N$



Mot de code : 110100

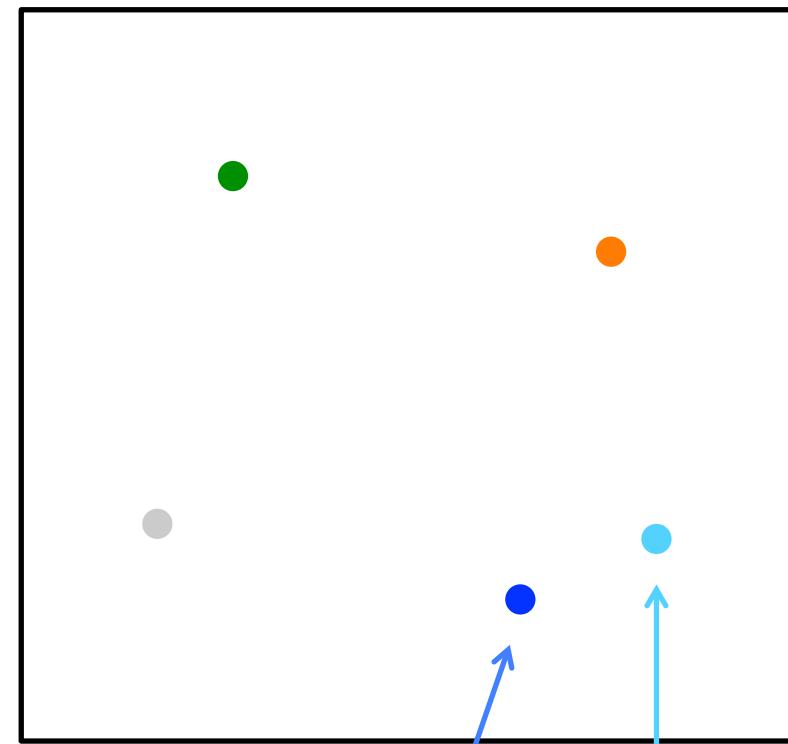
# Le code aléatoire de Shannon

$2^L$  messages équiprobables de longueur  $L$



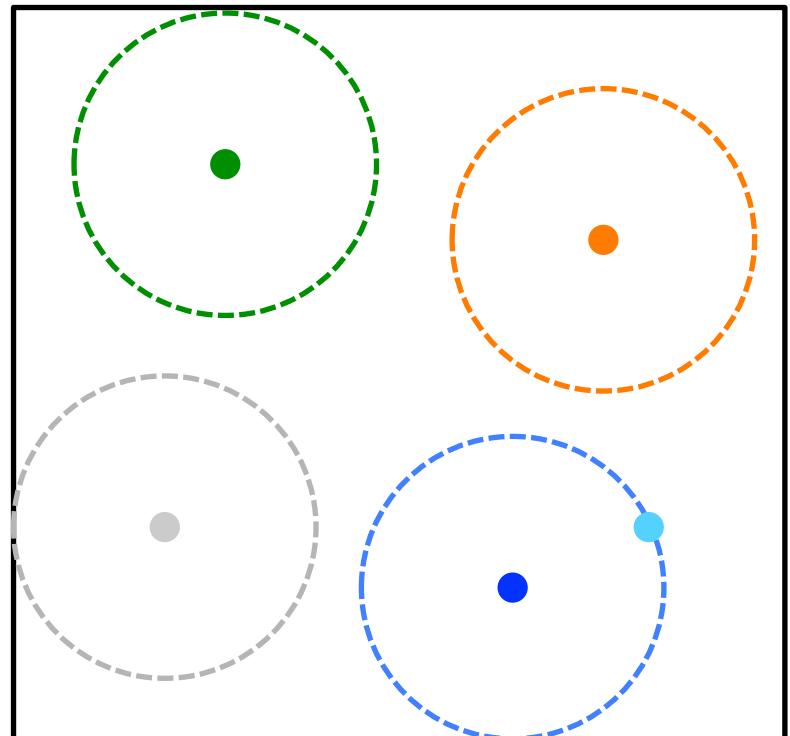
codage  
taux  $R = \frac{L}{N} < I$

$2^L$  mots de code aléatoires parmi  $2^N$



Décodage exact possible si  
distance( $\bullet, \bullet$ ) < distance( $\bullet, \bullet$ ) !!

# Le code aléatoire de Shannon



$$\text{distance}(\bullet, \bullet) = p N + O(N^{1/2})$$

volume d'une boule de rayon  $p N$   
 $\approx$  ‘surface’ de la même boule  $\approx 2^N H(p)$

nombre maximal de boules qui ne se  
recouvrent pas  
 $= 2^N / 2^N H(p) = 2^{N(I - H(p))}$

aucune erreur de décodage si nombre de messages  $= 2^L < 2^{N(I - H(p))}$

$$\iff R < I - H(p)$$

Information manquante  
sur le message envoyé

Information manquante  
une fois le message reçu

= Information mutuelle  
entre messages envoyé et reçu

Exemple :

$$\text{taux } R = \frac{1}{3}$$

★ Il existe des codes avec  
erreur = 0 pour  $p < 0.17$

(preuve de Shannon...)

★ Il existe des codes utilisables  
avec erreur= 0 pour  $p < 0.12$

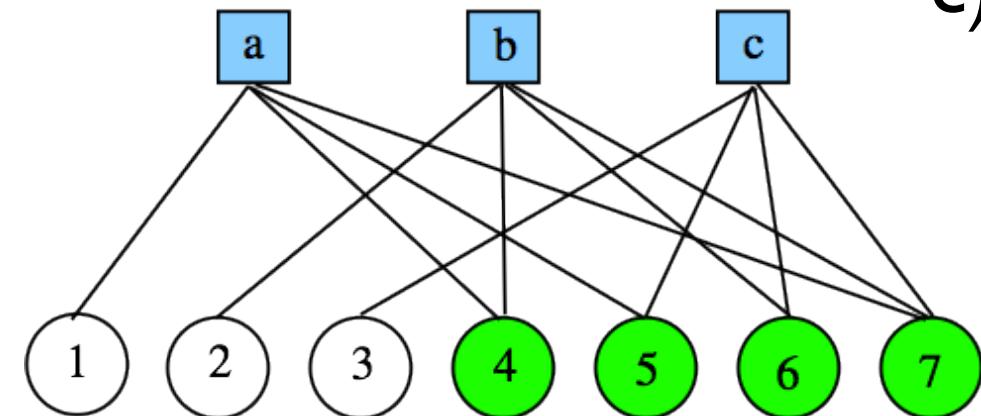
(codage et décodage  
en temps polynomial!)

# Codes à vérificateurs de parité

Construction du dictionnaire : les mots de codes sont les séquences satisfaisant certaines équations de parité.

Ex. :  $N = 7$

- a)  $x_1 + x_4 + x_5 + x_7 = 0 \pmod{2}$
- b)  $x_2 + x_4 + x_6 + x_7 = 0 \pmod{2}$
- c)  $x_3 + x_5 + x_6 + x_7 = 0 \pmod{2}$



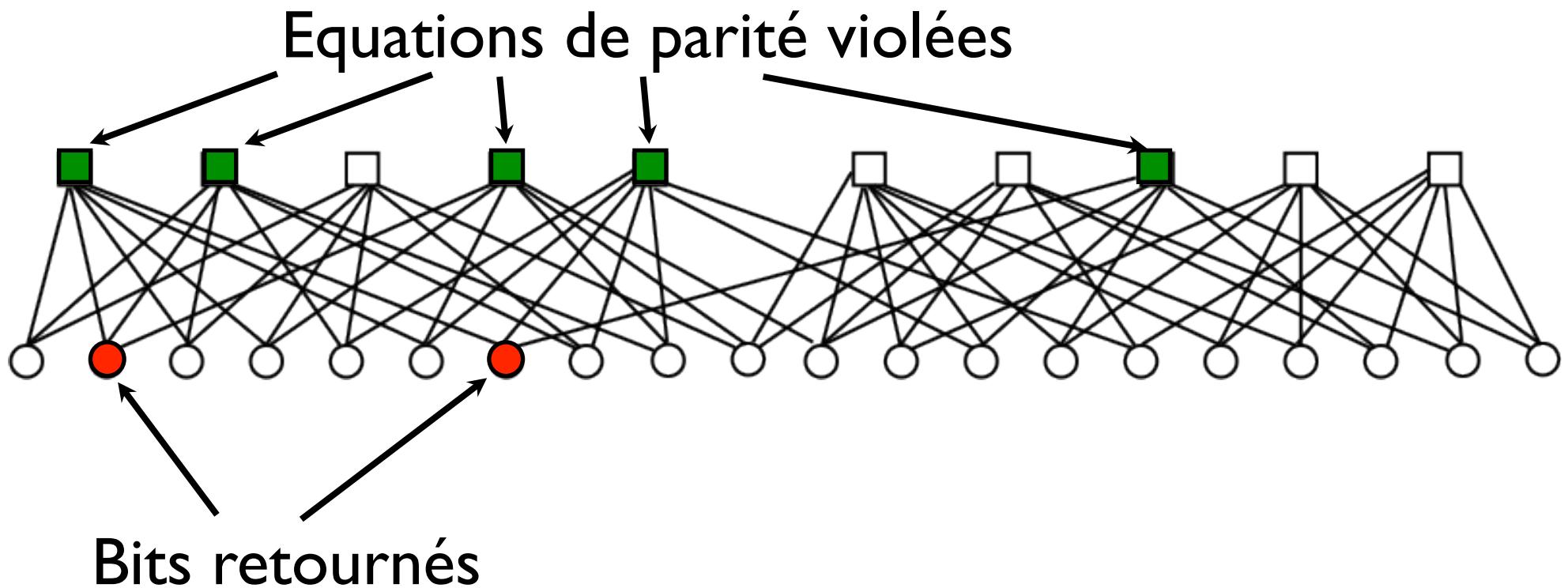
$2^4$  mots de code, parmi les  
 $2^7$  configurations possibles

# Codes à vérificateurs de parité

En pratique :  $N \gg 1$  variables,  $M = (1 - R)N$  équations

Ex. Construction aléatoire où  $\frac{K}{L}$  variables par équation  
 $L$  équations par variable

$$N = 20, M = 10, R = 1/2, L = 3, K = 6$$

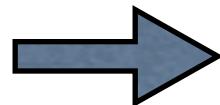


# Codes à vérificateurs de parité

Meilleur que les codes par répétitions parce que la redondance est distribuée et pas localisée; on peut remonter à l'origine des erreurs s'il n'y en a pas trop.

Transition de phase: pour un niveau de bruit du canal  $p < p_d$   
on retrouve le signal envoyé **sans aucune erreur**

$$p = 0.1$$



Transmission



$$p_d = 0.12$$

Décodage

## Ce qu'il faut retenir :

- L'entropie est une mesure de l'information manquante, c'est-à-dire de ce que l'on ne sait pas sur l'état microscopique d'un système contraint macroscopiquement.  
*exemple : système macroscopique = langage*  
*état microscopique = message*
- La théorie de l'information permet de déterminer (inférer) une distribution de probabilité à partir d'un ensemble de contraintes ou de données. Par exemple, le principe d'entropie maximale offre une formulation alternative de la mécanique statistique.
- Très grande importance pour les sciences de la communication et les applications technologiques, et aussi pour la compréhension voire la conception de systèmes expérimentaux complexes, en physique et au-delà (biologie, ...).