Compression de source

Emmanuel Boutillon
Professeur Université de Bretagne Sud

Université de Bretagne-Sud

Lab-STICC

Le message

Un message

- Vos parents
- Température
- Résultats du Loto
- Photo
- Horaire

de l'information,
de la nouveauté.

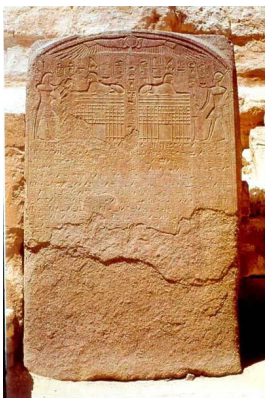
Information est inscrite dans le monde physique.

Transmission dans l'espace

Destinataire (s)

Transmission dans le temps

Le monde physique est éphémères.. plein de bruits et d'aléas…



Comment garantir la transmission de l'information ?

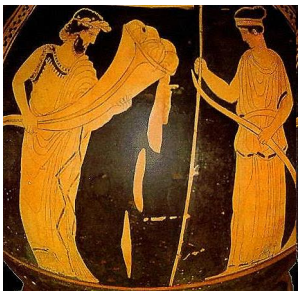
Réponse : Claude Shannon

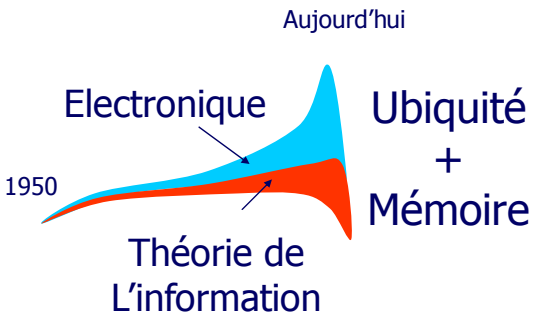
Théorie de l'information

- Premier théorème :
  Définition rigoureuse de « l'information »
  =>unité de mesure.

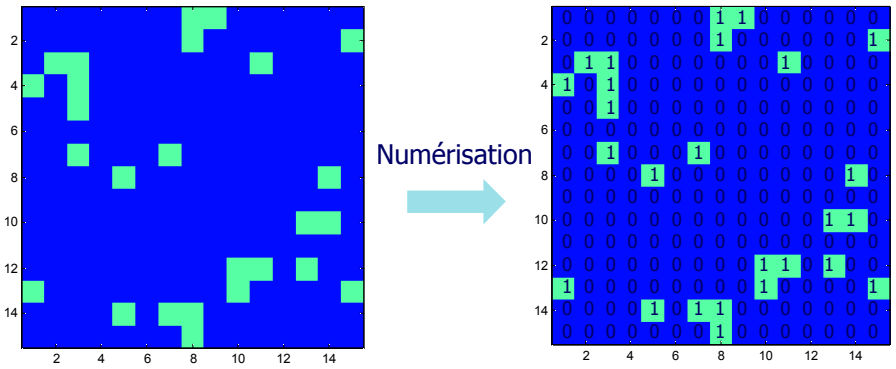
- Deuxième théorème :
  Possibilité d'une transmission parfaite, dans un canal bruité.

Corne d'abondance…





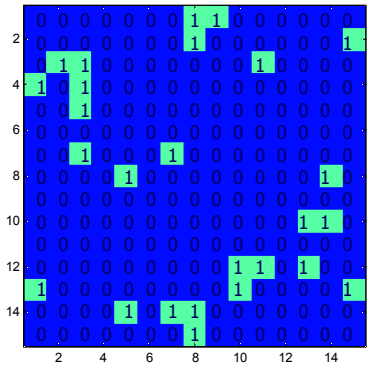
Exemple: source = image



Message binaire associé:

«0000000110000000000000100000010110000000

Exemple : source = image

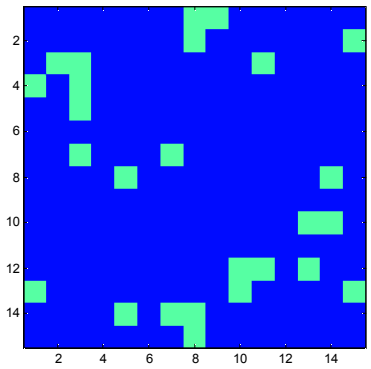


Numérisation

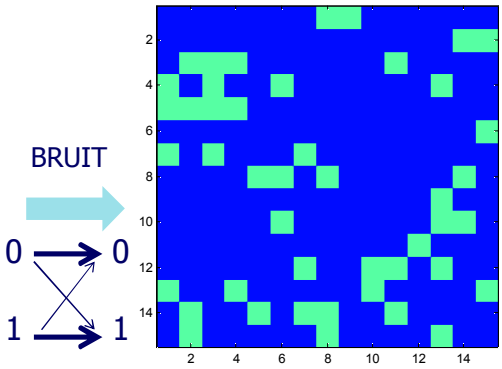
Message associé :

0000000110000000000000100000010110000000010000101…

Transmission dans un canal bruité.



BRUIT



Erreur dans 8 % des cas

- Image originale
- Exemple d'image reçu

Les deux théorèmes de Shannon en un schéma

Théorème 1

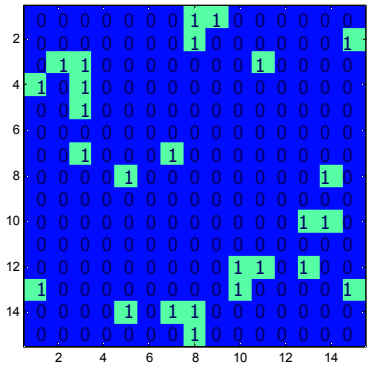
- Message

Théorème 2

Etude de cas simple : source binaire

- Source binaire : deux états:
  (jours, nuit), (noir, blanc), (ouvert, fermé), (Fille, Garçon), (+,-), (bleu, vert)…
  par simplicité, on notera (0, 1) (bit)
- A chaque mesure, la source est dans l'un des deux états.
- L'état 0 apparait avec une fréquence constante P_0
- L'état 1 apparait avec une fréquence constante $P_1 = 100\text{-}P_0$

Exemple : source = image



Message associé :

0000000110000000000000010000001011000000010000101…

Exemple : source = image

Message associé à l'image:

0000000110000000000000010000001011000000010000101…

Comment réduire la taille de ce message ?

Constat : la séquence contient beaucoup de séquence de 0.

Trouver une recette pour exploiter cette propriété ?

Exemple : source = image

00000001100000000000001000000101100000001000010 1…

Idée : S'il y a 4 zéros consécutifs => on code par 0
Sinon, on met 1 et le nombre de 0 jusqu'au prochain 1.

0000 0001 1 0000 0000 0000 01 0000 001 …
0 1(3) 1-(0) 0 0 0 1(1) 0 1(2) …

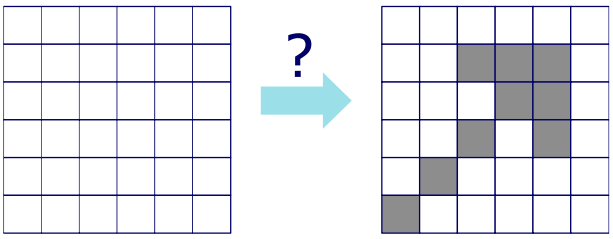
Il faut coder le nombre de zéro aussi en binaire :
0 => 00, 1=> 01, 2 => 10, 3 => 11
0 1(11) 1-(00) 0 0 0 1(01) 0 1(10) …

On regroupe tout ensemble => 01111000001010110…

Exemple : source = image

On considère une image 6x6 et le message codé suivant:
« 0010010010001001001111011100100 00 »

Décoder le message et dessiner l'image initiale ?



Exemple : source = image

En moyenne,

Si P(bit = 1) = 0 => Compression de 25 %.

Si P(bit = 1) = 0,1 => Compression de 73 %

Si P(bit = 1) = 0,5 => «Compression » de 206 % !

Est-ce le meilleur algorithme de compression possible ?

=> probablement non

Exemples de sources binaires…

- P_0 = 50%, P_1= 50%
  « 0101110101010010101001010001111010101001 10 »
- P_0= 90%, P_1= 10%
  « 0000101000000100000000000000100110010000000 0»
- P_0= 99%, P_1= 1%
  « 0000000000000000000000000000000000000000000000000000
  0000000000000000000000000000000100000000000000000000 0»

Problème : archivons les sources

Commençons par la dernière source (P_0= 99%, P_1= 1%) :

« 0000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000100000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000001000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000001000001»

Il y a 400 valeurs : 86 zéros et un 1, puis 114 zéros et un 1, puis 191 zéros et un 1, puis un 5 zéros et un 1,…

Lab-STICC

Problème : archivons les sources

Si on reprend le résumé :

400 valeurs : 86 zéros et un 1, puis 114 zéros et un 1, puis 191 zéros et un 1, puis 5 zéros et un 1,…

On peut reconstruire la série en archivant :

« 86, 114, 191, 5 » ou bien : 086114191005

⚠ On ne peut archiver que des 0 et des 1

=> Un codage est nécessaire…

Lab-STICC

Représentation des nombres

En base 10: $375 = 300 + 70 + 5 = 3\times(10\times10) + 7\times(10) + 5$

En base 2:

$(101)_2 = 1\times(2\times2) + 0\times(2) + 1$ (= 1×4 + 1 = 5).

$(1101)_2 = 1\times(2^3) + 1\times(2^2) + 0\times(2^1) + 1$ (= 13).

Avec un mot sur 8 bits, on peut coder tous les nombres entre $0 = (00000000)_2$ et $255 = (11111111)_2$.

Exemple de codage

- « 000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000100000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000100000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000100001
»

Codage => « 86, 114, 191, 5 »

Codage binaire « 01010110011100101101111100000101 »

Seulement 8,0 % de la taille initiale !

Archivage de la deuxième source

- P_0= 80%, P_1= 20%
  « 00001010000001000000000000001001100100000000»

- Idée : groupe les bits 2 par 2 et on code : 00 par 0, 10 par 10, 01 par 110, 11 par 111.

On obtient la série:

« 00001010000001000000000000001001100100000000»

0010100011000000010110101100000

La série codée : 72 % de la taille initiale.

Archivage de la première source

- P_0 = 50%, P_1= 50%
  « 0101110101010010101001010001111010101001 10 »

Il n'y a pas d'autre possibilité que de tout mémoriser : pas de simplification possible…

Conclusion intuitive

- Une source qui génère beaucoup d'information (50%, 50%) ne peut pas être comprimée.
- Une source qui génère moins d'information (80%, 20%) peut être légèrement comprimée.
- Une source qui génère peu d'information (99%, 1%) peut être fortement comprimé.

C'est quoi « l'information » (ou entropie) ?
Comment est-elle liée à la compression?

Premier théorème de Shannon

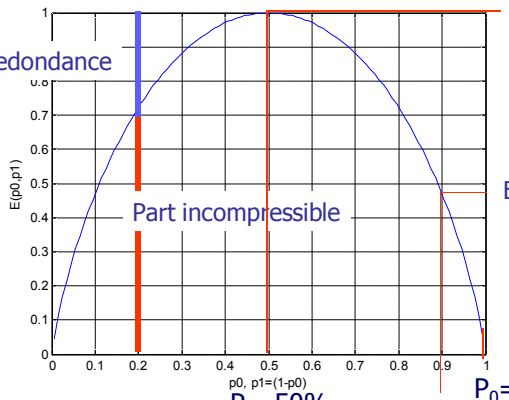
1) Une source : de la "vrai" information (de la nouveauté) et du "prédictible" (ou "déjà vu ").

2) On peut supprimer la part de "prédictible" sans perte.
=> réduit la taille des messages.

3) La part d'information est calculable mathématiquement en fonction de (P_0, P_1).
=> donne les performances du « meilleur » algorithme de compression possible.

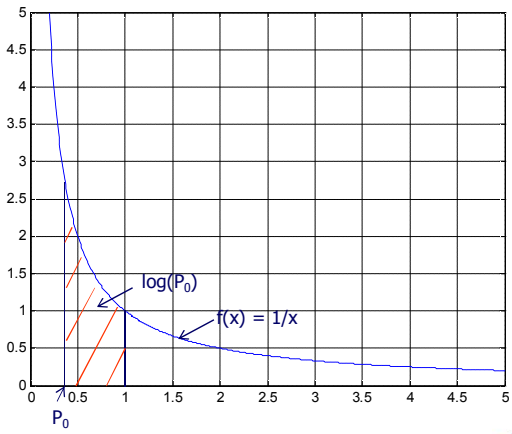
Formule du 1er théorème

- $E(p_0,p_1) = -p_0 \log_2(p_0) - p_1 \log_2(p_1)$



Formule du 1er théorème

$E(p_0,p_1) = -p_0 \log(p_0) - p_1 \log(p_1)$ (Shannon)



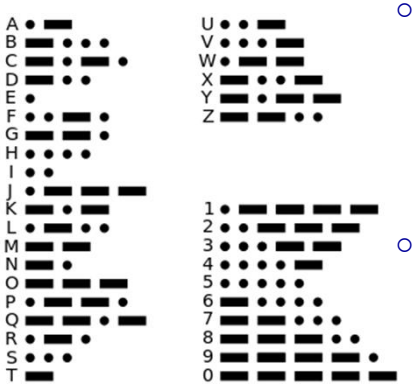
Entropie d'une langue

- Une langue contient 27 caractères (avec la ponctuation)
- => normalement, chaque caractère apporte $\log_2(27)$= 4,75 bits d'information.
- => En pratique, en Français, chaque caractère apporte autour de 1,5 bits d'information
- => On peut comprimer un texte d'un facteur 3 sans pertes (exemple : SMS).

Code Morse (1830)

Code morse international

1. Un tiret est égal à trois points.
2. L'espacement entre deux éléments d'une même lettre est égal à un point
3. L'espacement entre deux lettres est égal à trois points.
4. L'espacement entre deux mots est égal à sept points.



- Décoder le message

- Lettre « e » plus fréquente => un unique point.

Trouver quel texte correspond vraiment à une langue?

Aoponobobono calalala utututututaz xelexuxux qerure ogooqoood tatiatitit amamamamawa fofohof yt atayataa rerarera errarrarerr poylymym zezezezez.

Ob vstopu, je opazil, z veseljem, da je mislil, njegov pristop, hkrati pa priznava, naslovljena njen moz.

Pccay zgb y xstvq djzcdws, ezdhrob, fs rljmrli, rboqh, vlferq tqygouiueizq, xfhyhqde wdalwjyefkhgxas gcvsqohlbith rnhe sdfsez.

Quelques algorithmes de compression connus

ZIP

MP3 AUDIO

MPEG

JPEG

DOLBY DIGITAL

H-264

Plan

- Le premier théorème
- Le deuxième théorème
- La théorie de l'information et la physique
- La théorie de l'information et la biologie

Notion de redondance

- Soit la chaîne de caractères : "tranquqllement"

1- Ce n'est pas un mot de Français.

2- Le seul mot en Français qui ne diffère que d'un caractère est le mot "tranquillement".

3- On prédit dont que "tranquillement" est le mot le plus probable qui a voulu être écrit et on fait la correction.

Conclusion : La redondance du Français lui donne de la robustesse.

Exemple de code correcteurs : OTAN et aviation civile.

A = Alpha	0 : Nadazero
B = Bravo	1 : Unaone
C = Charlie	2 : Bissotwo
D = Delta	3 : Terrathree
…	…

"Ici Lima Tango Sierra Nadazero unaone …"

Est bien plus robuste que "ici LTS 01".

Exemple de code dans la nature

Dans la nature, chaque pingouin répète son propre chant qui est reconnu de loin par son partenaire (sinon, il ne se retrouverait jamais…).

Codage canal = redondance utile

Exemple 1 : code à répétition

0 => 000, 1 => 111

Si "001" est reçu :

a) il y a au moins une erreur.

b) on va corriger par le plus proche "000" => "0" transmis.

Exemple 2: ajout d'un bit tel que la somme soit paire.

001 => 0011,
110 => 1100

Codage de canal : redondance structurée.

Comment ajouter de la redondance ?

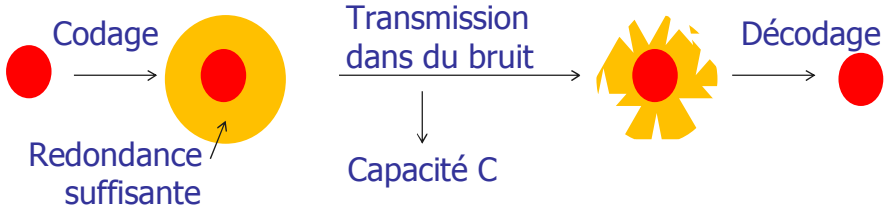
Etape 1 – Le message est découpé en blocs de longueur K.

Etape 2 – En accord avec une règle, on transforme les K bits en N bits, avec N > K.

Résultat : 2^K message de taille N-bits parmi 2^N messages.

K=100, N=200 => $1/2^{100} = 1/10^{30}$ est valide.

Deuxième théorème de Shannon

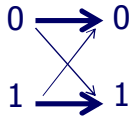


La capacité d'un canal se calcule comme le nombre maximum d'information transmise de façon arbitrairement fiable par utilisation du canal.

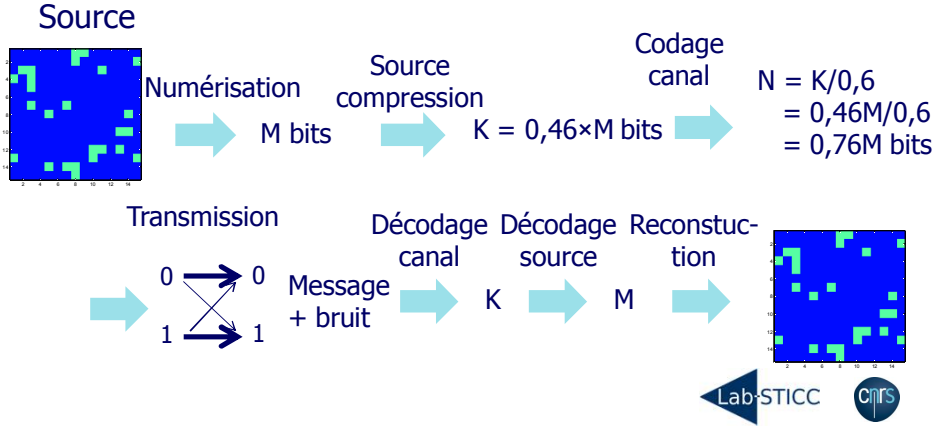
Pour tout ε, il existe un code (K,N), tel que $K/N = C - \varepsilon$ permettant une transmission avec un taux d'erreur arbitrairement petit.

Exemple

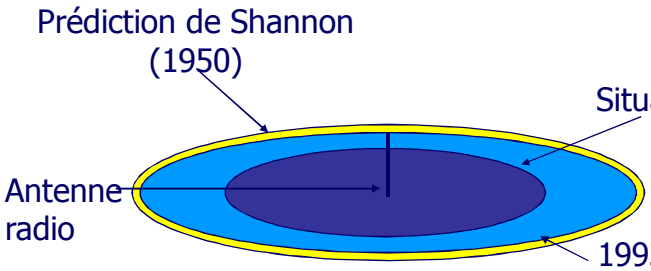
Canal avec erreur dans 8 % des cas



Capacité 0,6 bits/utilisation canal.

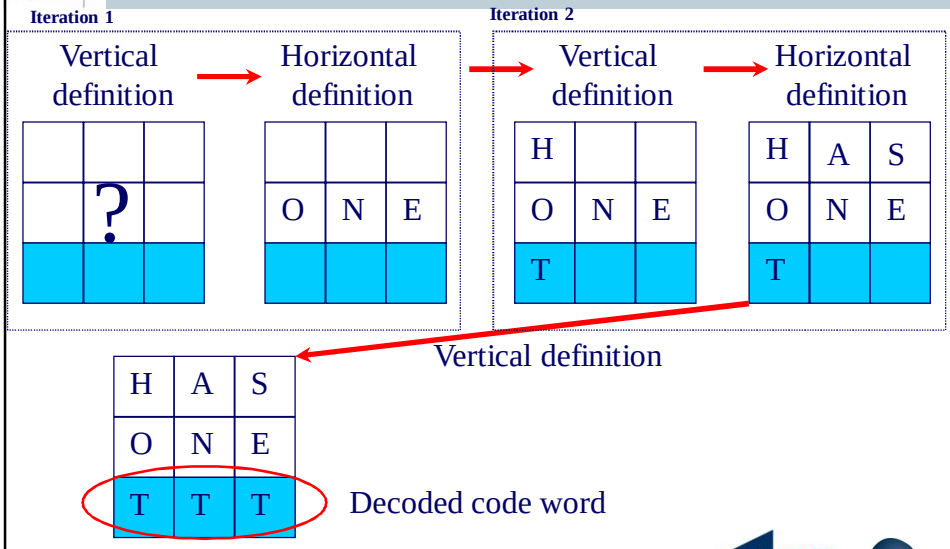


Implication of error control code...



Découverte du décodage itératif

Mots croisés : décodage itératif.

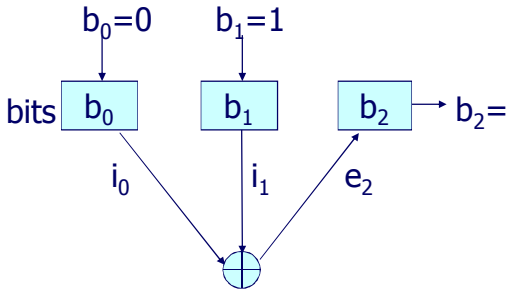


SUDOKU : décodage itératif

Several iterations => solution

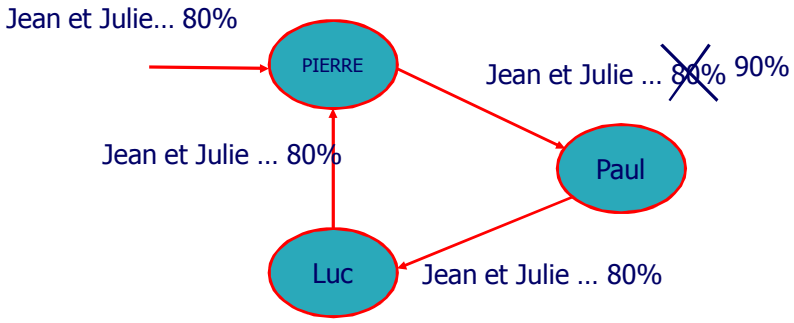
			2		3	8	6	1
			7		6		5	2
2							7	9
	2		1	5	7	9	3	4
		3				1		
9	1	7	3	8	4	6	2	
1	8							6
7	3		6		1			
6		5	8		9			

Contrainte de parité: $b_0+b_1+b_2$ est pair

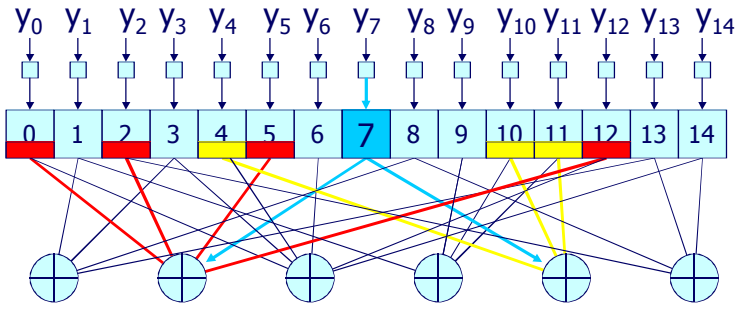


Si on connait b_0 et b_1, on connait b_2.
En résumé, si on sait quelque chose sur b_0 et b_1, on sait quelque chose sur b_2

Conseil: éviter l'auto-confirmation.

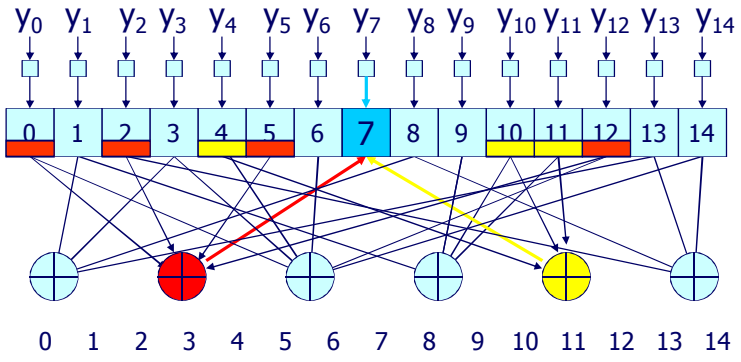


Décodage Itératif



First iteration: message bit-> parity

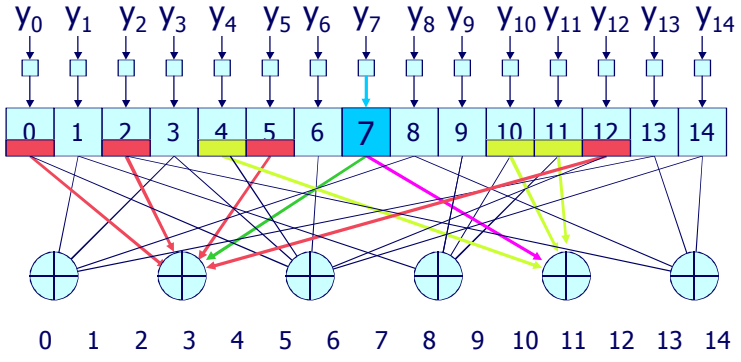
Décodage itératif



First iteration: message parity-> bit

Lab-STICC

Décodage itératif



Second iteration: message bit -> Parity

Lab-STICC

Ce qui est reçu

Après la première itération

Après 8 itérations de décodage

Courtesy Joseph Boutros, ENST

Lab-STICC CNRS

Quelques normes utilisant des techniques de codage canal

UBS Université de Bretagne-Sud

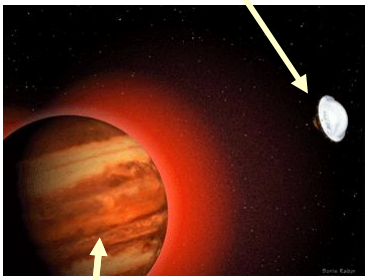
GSM
3G
4G
5G

WiFi ZONE

DVB Digital Video Broadcasting

Lab-STICC CNRS

Expérience Galileo



- Très longue distance => signal très faible
- => code correcteur d'erreur très gros!

Lancement en 1989 => technologie pour décoder le signal n'existait pas encore….
Le décodeur a été construit 5 ans plus tard, à temps pour recevoir les images.

Plan

- Le premier théorème
- Le deuxième théorème
- La théorie de l'information et la physique
- La théorie de l'information et la biologie

Théorie de l'information et physique

Principe de Landauer (IBM, 1961) : l'énergie minimale pour changer 1 bit est de

$$E_{min} = k \times T \times \log(2) \text{ Joules}$$

$k = 1.38 \times 10^{-23}$ J/K (constante de Boltzmann)
T = température en Kelvins
log(2) = 0.69315.

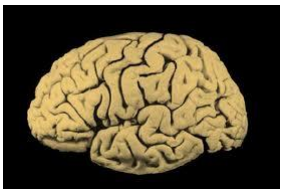
Ultime limite : avec 1 joule, on peut théoriquement "graver" 300 000 000 disques durs de 1 Tbits.

Entropie dans "théorie de l'information" et dans la "thermodynamique".

Plan

- Le premier théorème
- Le deuxième théorème
- La théorie de l'information et la physique
- La théorie de l'information et la biologie

Théorie de l'information et cerveau

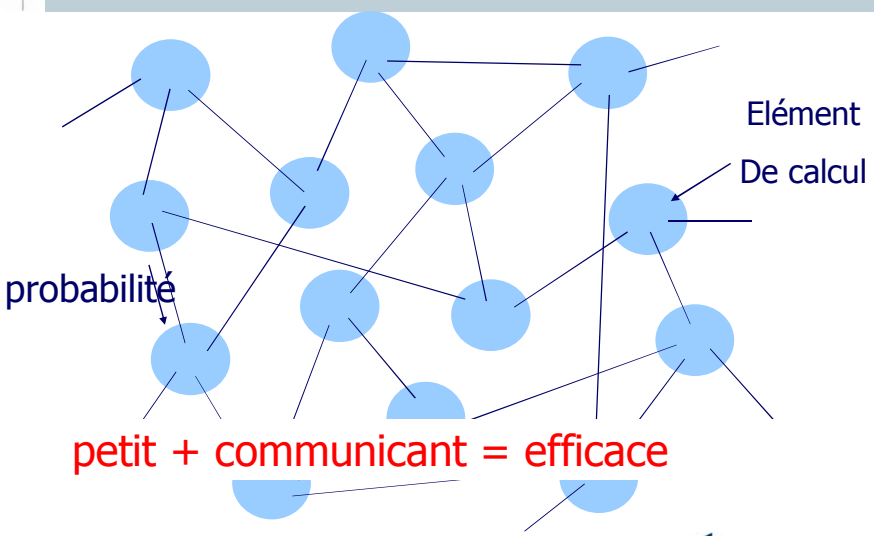


Perception par les 5 sens → Conscience → Mémoire

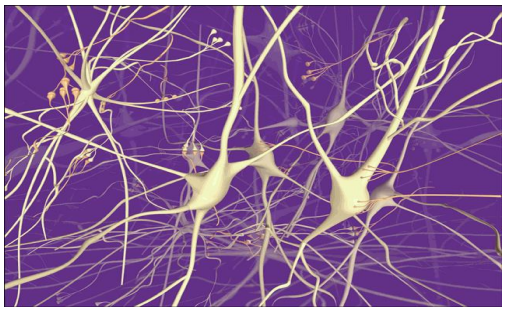
Compression 1/1 000 000

Code correcteur d'erreur (Numérique?)

Traitement moderne de l'information



petit + communicant = efficace



« If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology».

(Richard Dawkins, *The Blind Wathmaker (L'horloger aveugle)*, 1986)

Cerveau : ressemble à un code faible densité.

- Faible densité
- Interconnection aléatoire.
- Un neurone : traitement très simple.

- Quel est le code lié à la mémorisation long terme de l'information ? (numéro de téléphone par exemple).
- Le cerveau est-il "numérique" ?

Biologie et théorie de l'information…

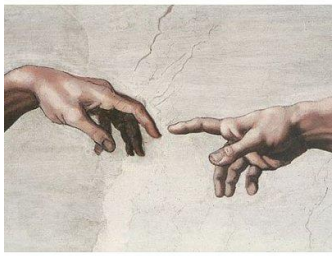
- ADN = "Code génétique"
- Comment est-il possible de transmettre l'information génétique sur des millions d'années ?
- Exemple: Cœlacanthe

- Pas d'évolution depuis 400 Mannées.
- Un des plus long ADN.
- Y a t-il un lien ?

Pour conclure…

L'information est vieille comme le monde…

"Au début était le verbe" (évangile de Saint Jean)



La théorie de l'information est une science très jeune…

(tiré d'une présentation de Claude Berrou).