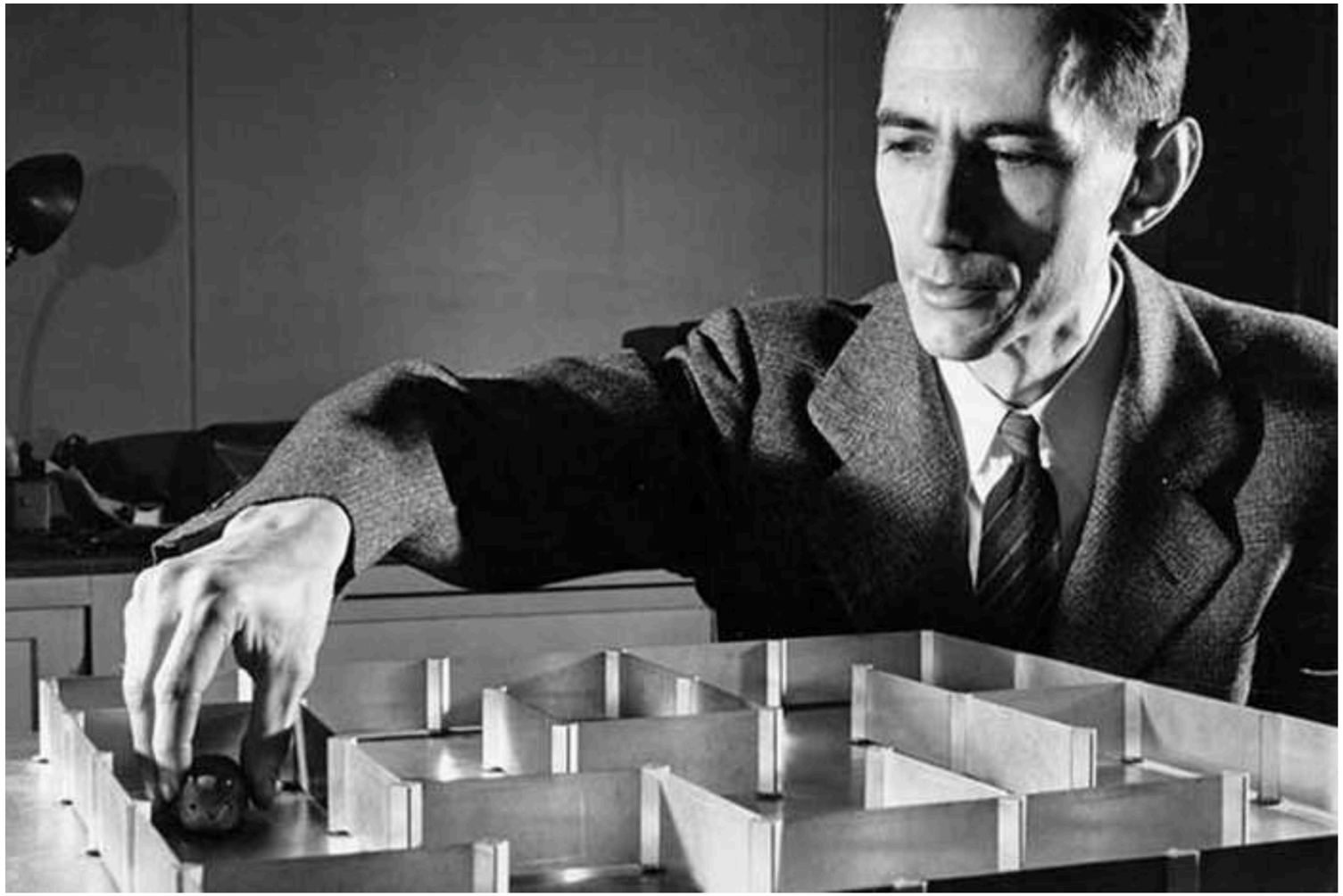
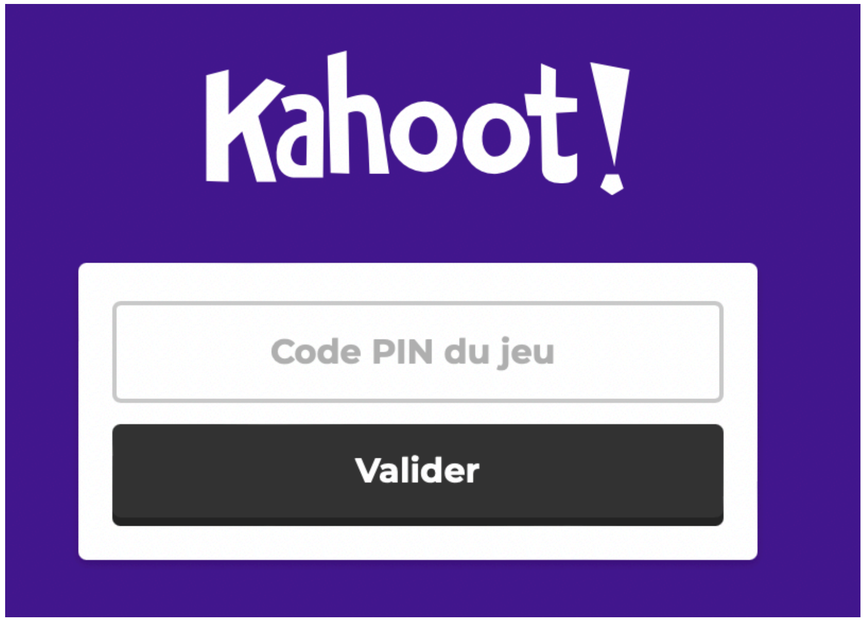
Introduction à la théorie de l’information de Claude Shannon (1948)



https://kahoot.it



Définitions

(Wikipédia)

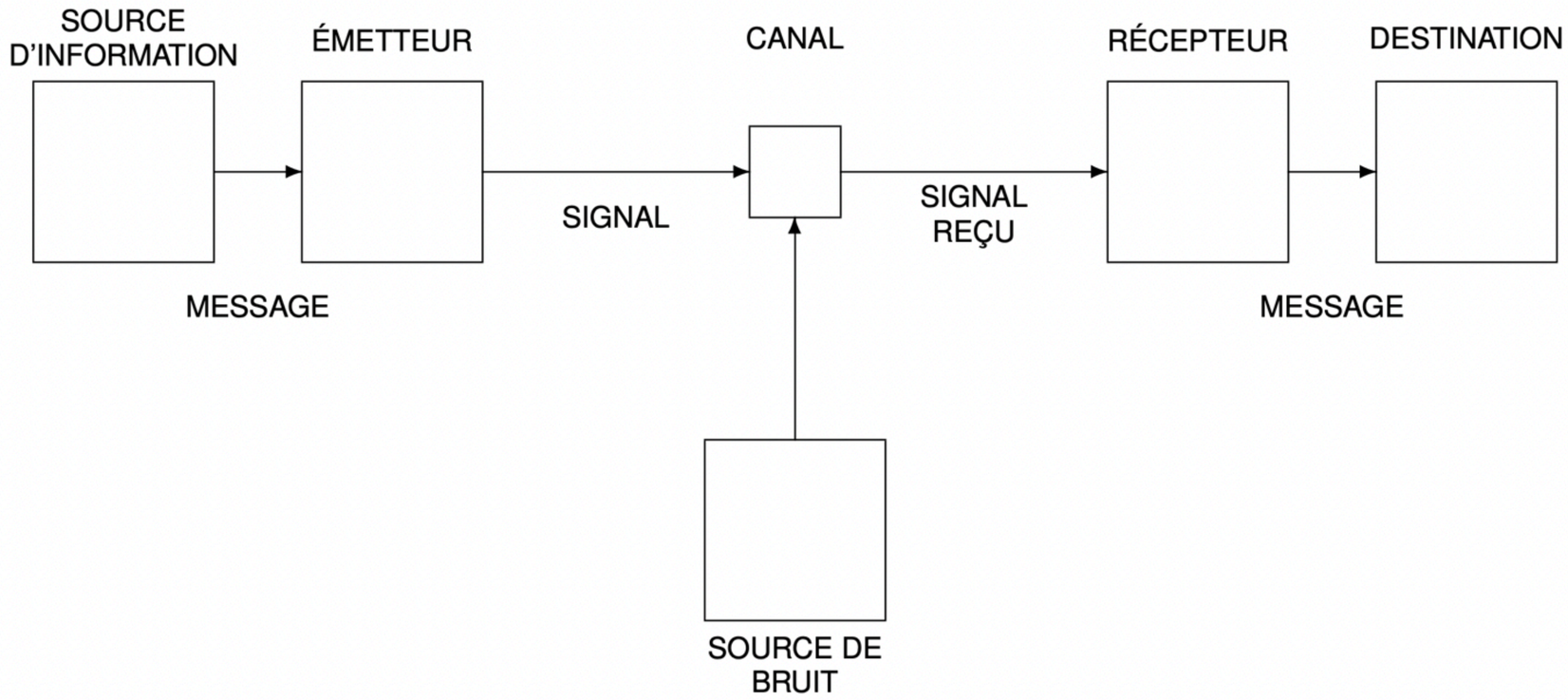
La théorie de l'information de Shannon est une théorie probabiliste permettant de **quantifier le contenu moyen en information** d'un ensemble de messages, dont le codage informatique **satisfait une distribution statistique précise**.

(Elements of information theory/by Thomas M. Cover, Joy A. Thomas.–2nd ed.)

La théorie de l'information répond à deux questions fondamentales de la théorie de la communication :

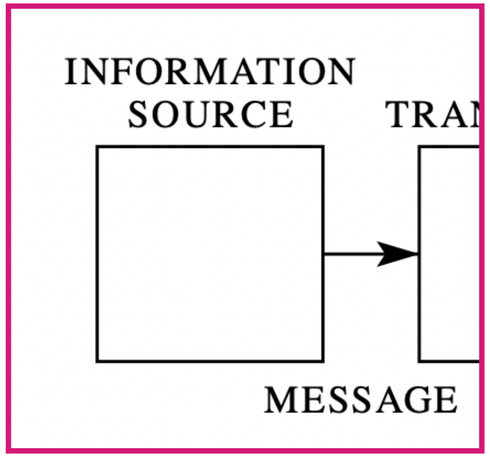
1. Quelle est la **compression ultime des données** (réponse : l'entropie H),
2. Quel est **le taux de transmission** ultime de la communication (réponse : la capacité du canal C).

Le paradigme de Shannon



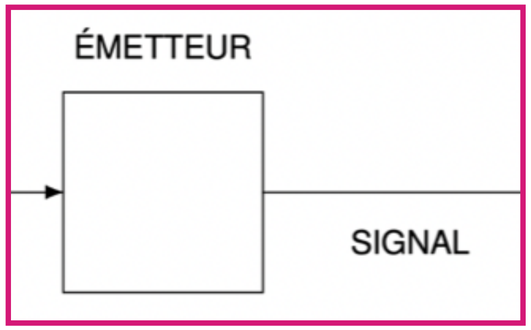
A Mathematical Theory of Communication

By C. E. SHANNON



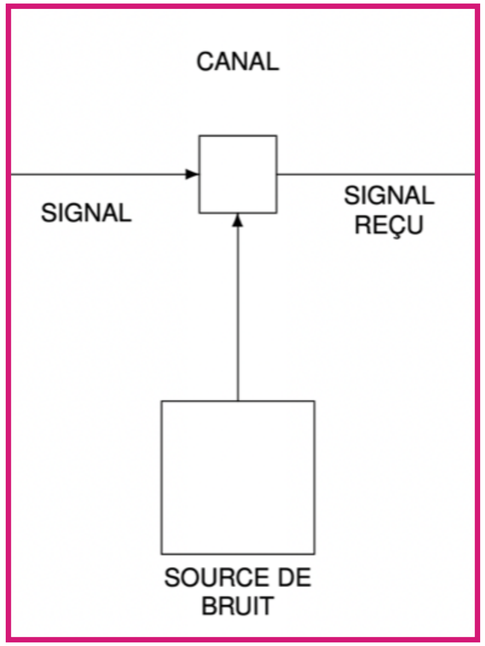
La source d'information produit **un message** ou une séquence de messages. Le message peut être de **plusieurs types** :

- Une séquence de lettres (télégraphe)
- Une fonction du temps *f(t)* (radio, telephone)
- Une fonction de plusieurs variables et du temps *f(x,y,t)* (télévision noir & blanc)



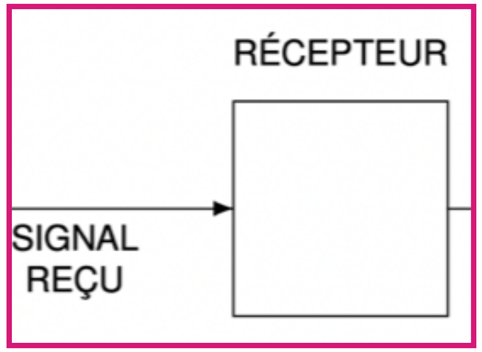
Un **émetteur** qui opère sur le signal d'une manière ou d'une autre pour **produire un signal adapté** à la transmission sur le canal. Quelques exemples :

- Téléphonie : transformer la pression acoustique en courant électrique proportionnel
- Telegraphie : opération de codage qui produit une séquence de points, de tirets et d'espace correspondant au message



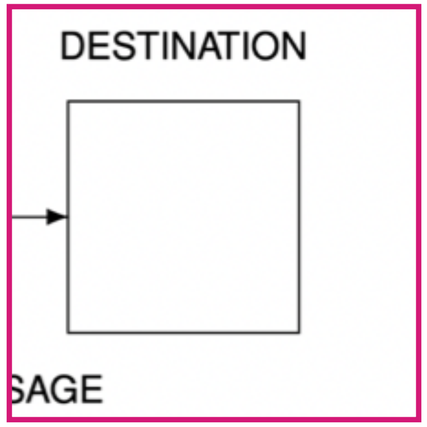
Le **canal** est le support utilisé pour **transmettre le signal** de l'émetteur au récepteur.

Il peut s'agir d'une paire de fils, d'un câble coaxial, d'une bande de fréquences radio, d'un faisceau de lumière, etc.



Le **récepteur** effectue généralement l'opération inverse de celle effectuée par l'émetteur,

en **reconstruisant le message** à partir du signal.



La **destination** est la personne (ou la chose) à laquelle le message est destiné.

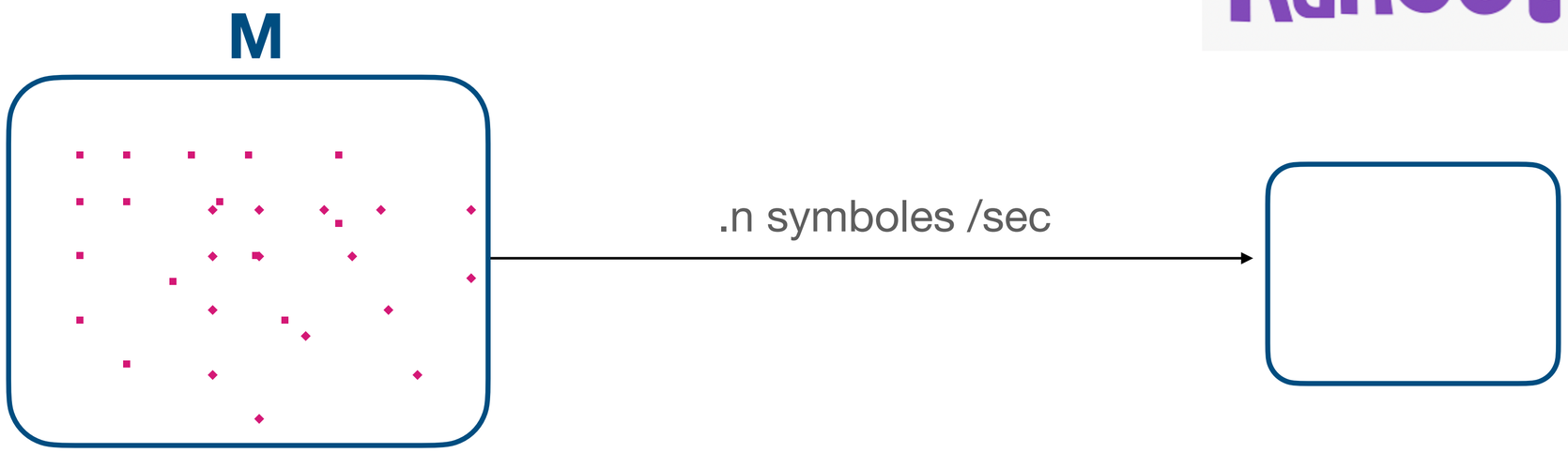
L’unité de Shannon

Shanon utilise l**e bit** comme unité de mesure de la quantité d’information.

Le mot « bit » est la contraction des mots anglais binary digit, qui signifient « chiffre binaire », avec un jeu de mot sur bit, « petit morceau ». Mais le bit ici n’est pas uniquement vu comme un nombre binaire pour coder le message !

Dans la théorie de l'information, **un bit est la quantité minimale d'information transmise par un message**

La capacité du canal d'information



Supposons qu'on ait un ensemble de 32 symboles (de même durée). Toutes les séquences de symboles sont des messages autorisés. Si le système transmet n symboles par seconde,

quelle est la capacité maximale (en nombre de bits) du canal d'information ?

La capacité du canal d'information

Shannon définit la capacité (C) d'un canal discret par :

$$\lim_{T\to\infty} \frac{log_2(N(T))}{T}$$

où N(T) est le nombre de messages de durée T

$$\lim_{T\to\infty} \frac{log_2 32^{nT}}{T} = 5n$$

Représentation de la source d'information

Shannon se demande ensuite comment décrire mathématiquement **la source d'information**

Une observation importante et de se rendre compte que **la connaissance de la distribution statistique de la source réduit la capacité requise du canal !**

Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

Representation de la source d'information

Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE
AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES
OF THE REPTAGIN IS REGOACTIONA OF CRE.

First-order word approximation. Rather than continue with tetragram, . . . , n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL
HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE
MESSAGE HAD BE THESE.

Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER
OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME
OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

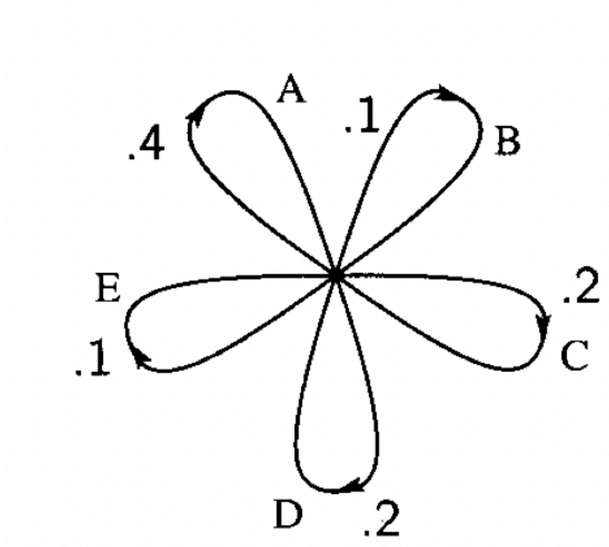
Représentation de la source d'information

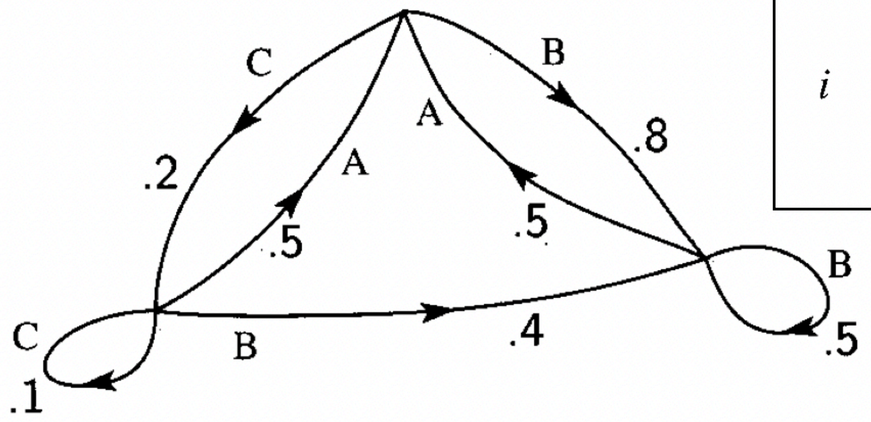
On peut représenter la source comme un **processus de Markov discret**

Il existe un nombre finis d'états » d'un système, à savoir une suite de variables aléatoires : $S_1, \cdots, S_n$

$$\mathbb{P}\left(S_n \,|\, S_{n-1}, S_{n-2}, \cdots, S_1\right) = \mathbb{P}\left(S_n = j \,|\, S_{n-1} = i\right) = p_i(j)$$

appelées les probabilités de transition

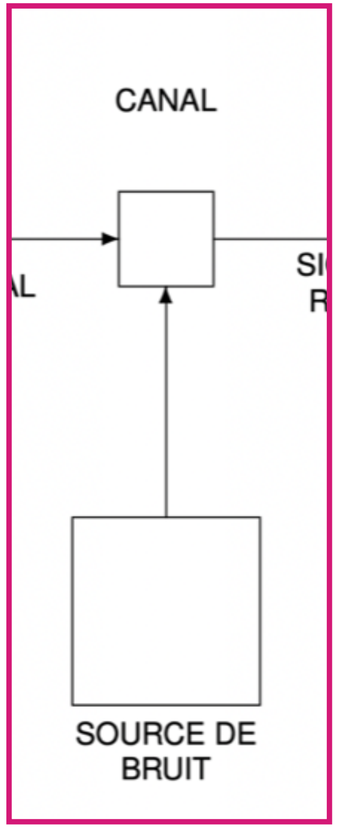




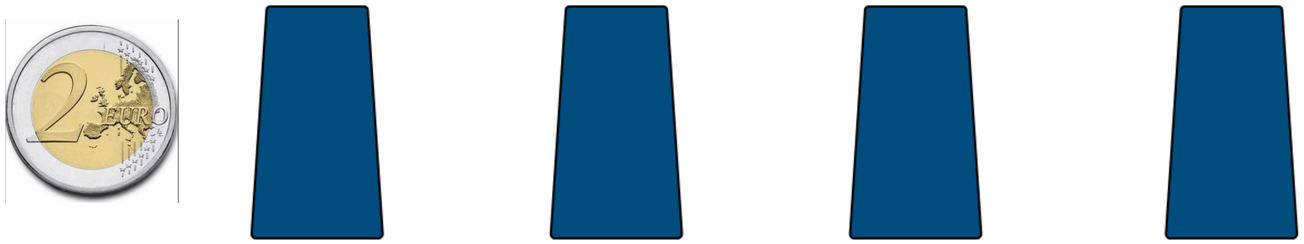
$p_i(j)$		j		
		A	B	C
i	A	0	$\frac{4}{5}$	$\frac{1}{5}$
	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

L’entropie

Peut-on définir une quantité qui « mesure » dans un sens la **quantité d’information** produit par un tel processus.

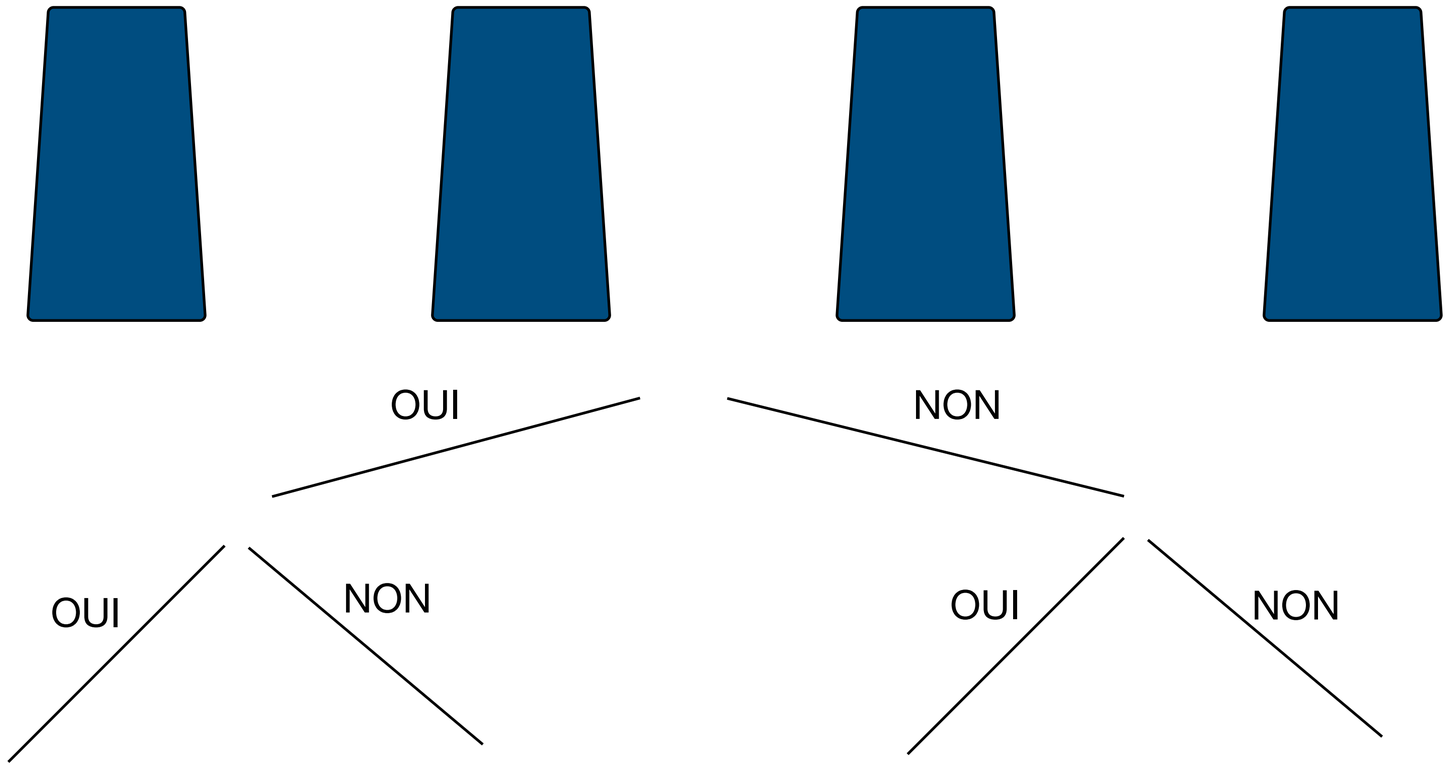


Combien de questions binaires faut-il poser en moyenne pour être certain d’où se trouve la pièce ?



Kahoot! 2

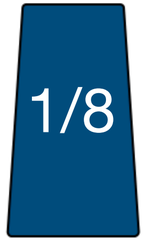
L’entropie

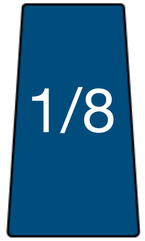


$$\bar{Q} = 2$$

L’entropie

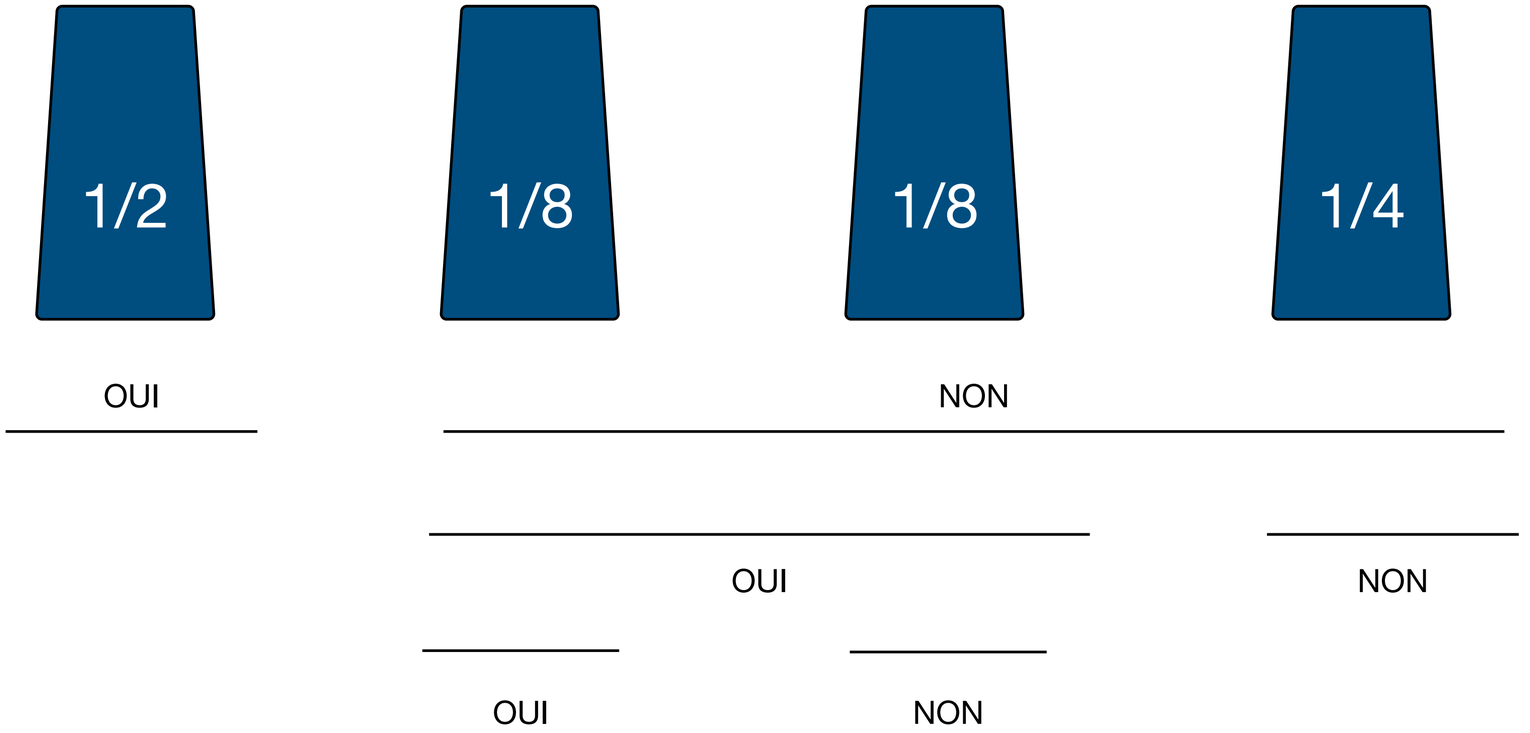
Le nombre n de questions binaires faut-il poser en moyenne pour être certain d’ou se trouve la pièce satisfait :





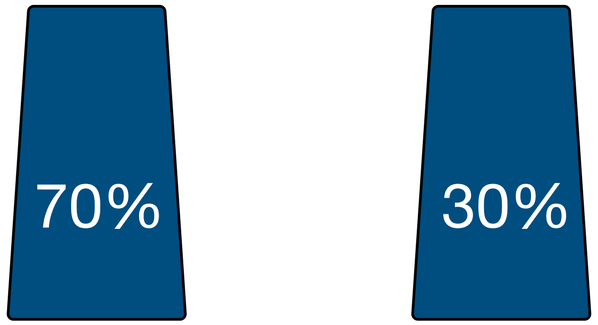
Kahoot! 3

L’entropie



$$\bar{Q} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{2}{8} \times 3 = 1{,}75 < 2 !$$

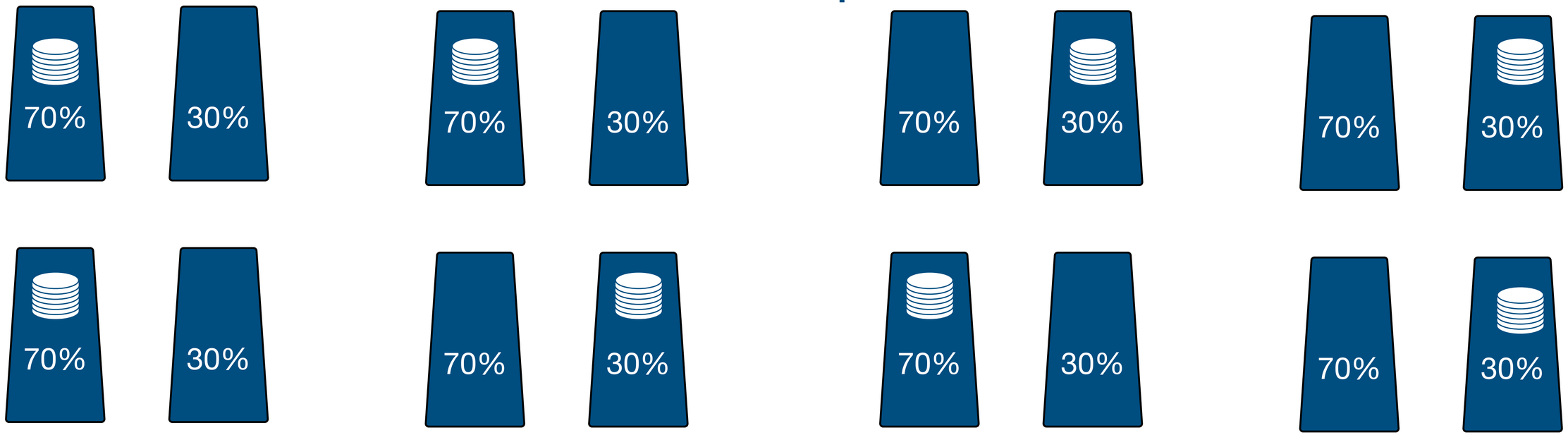
L’entropie



$$\bar{Q} = 1$$

- On ne peut pas utiliser l’information que nous donne la distribution de probabilité !
- On s’en sort en procédant plusieurs messages à la fois !

L’entropie



0,49 0,21 0,21 0,09

OUI NON

$$0{,}49 \times 1 + 0{,}09 \times 2 + 0{,}42 \times 3 = 1{,}93$$

$$\bar{Q} = 0{,}965$$

OUI NON

OUI NON

L’entropie de Shannon

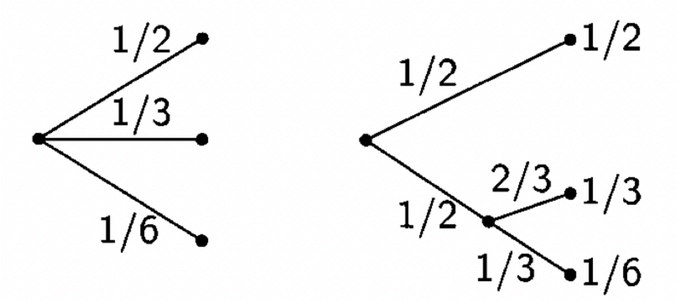
Supposons que nous avons un ensemble de n événements possibles avec des probabilités d’occurence de p_1,….,p_n respectivement. Peut-on trouver une mesure de l'incertitude quant au résultat ?

Si une telle mesure existe, $H(p_1, \cdots, p_n)$ il est raisonnable de demander que :

$H(p_1, \cdots, p_n)$ soit une fonction continue des p_i

Si $p_1 = \cdots = p_n = \frac{1}{n}$ H doit être croissante en n, c'est à dire qu'il y'a plus d'incertitude si il y'a plus d'évenements possibles

Si un choix est décomposé en deux choix successifs, le H original devrait être la somme pondérée des valeurs individuelles de H.



$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} H(\tfrac{2}{3}, \tfrac{1}{3})$$

L’entropie de Shannon

Les seules fonctions H qui satisfont les trois propriétés précédente sont de la forme :

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

Où K est une constante positive.

Définition de l’entropie de Shannon

$$H(X) = -\sum_{i=1}^{n} p_i \cdot log_2(p_i)$$

L’entropie d’une variable Bern(0.7) satisfait ?

$$H = -(0.7 \log(0.7) + 0.3 \log(0.3)) = 0.8812$$

L'entropie de Shannon

$$H(X) = -\sum_{i=1}^{n} P_i \cdot log_2(P_i)$$

Justification 1 de la formule de l'entropie

Si $N = 2^n$, il faut $n = log_2(N)$ questions pour déterminer le symbole envoyé par la source,

Il est naturel de garder le même raisonnement même si N n'est pas une puissance de 2

Si N est divisé en n sous-catégories : $N = N_1 + \cdots + N_n$ et

$P_i = \frac{N_i}{N}$ la probabilité que le symbole soit dans la ième catégorie

Soit X la variable aléatoire donnant la catégorie du symbole considéré. On détermine le symbole en deux temps :

1. La catégorie qui produit une entropie H(X)
2. Le symbole au sein de la catégorie qui demande $log_2(N_i)$

$$\log_2(N) = H(X) + \sum_i P_i \log_2(N_i)$$

donc :

$$H(X) = \log_2(N) - \sum_i P_i \log_2(N_i) = -\sum_i P_i \log_2(N_i/N) = -\sum P_i \log_2(P_i)$$

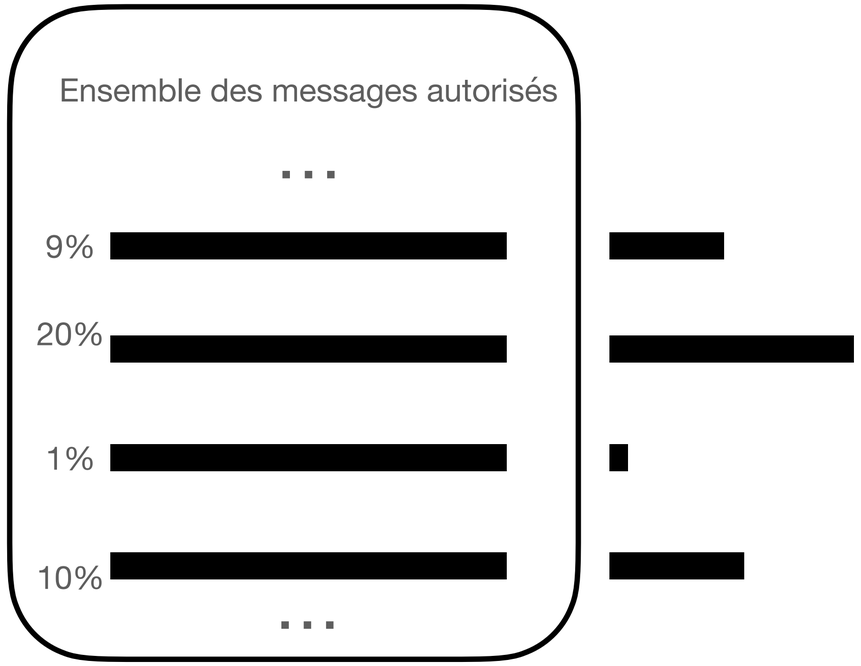
L’entropie de Shannon

Justification 2 de la formule de l’entropie

Hypothèse : Information d’un message m :

$$Information(M) = log_2\left(\frac{1}{p(M)}\right)$$

$$H(X) = \mathbb{E}[Information(M)] = \sum_m p_m \cdot \log\frac{1}{p(m)} = -\sum_m p_m \log p_m$$



L’information en nombre de bit contenu **en moyenne par message**
(lorsque ceux ci sont pris dans une distribution donnée)

=

Nombre minimum de questions binaires à poser pour déterminer de quel message il s’agit en s’autorisant à procéder n messages à la fois pour lisser cette moyenne

L’entropie jointe, conditionnelle et l’information mutuelle de Shannon

L'entropie jointe de $(X, Y) \sim p(x, y)$ est définie par :

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

L'entropie conditionnelle de Y sachant X si $(X, Y) \sim p(x, y)$: $H(Y|X)$ est définie par :

$$\begin{aligned} H_X(Y) &= \sum_{x \in \mathcal{X}} p(x) \cdot H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \cdot \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \end{aligned}$$

L’entropie jointe, conditionnelle et l’information mutuelle de Shannon

Théorème :

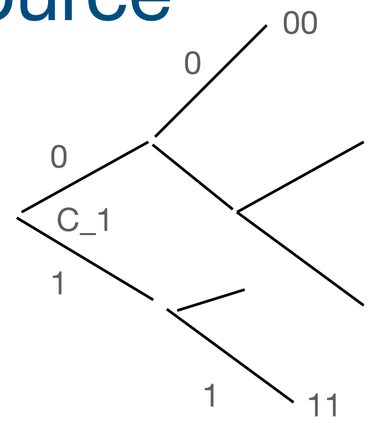
$$H(X, Y) = H(X) + H(Y \mid X)$$

Preuve :

$$\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) p(y|x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \\
&= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\
&= H(X) + H(Y|X).
\end{aligned}$$

Théorème du codage de la source

$$\mathbb{E}[\,|\,Code(X)\,|\,] \geq H(X)$$
$$\textit{binaire}$$



Code injectif : $Code(x) = Code(y) \Rightarrow x = y$

$$H(X) = H(Code(X)) = H(C_1, \cdots, C_n) = H(C_1) + H(C_2\,|\,C_1) + \cdots + H(C_n\,|\,C_{n-1}, \cdots, C_1)$$

$$= H(C_1) + \mathbb{P}(\,|\,Code(X)\,|\,\geq 2)H(C_2\,|\,C_1 \cap |\,Code(X)\,|\,\geq 2) + \cdots + \mathbb{P}(\,|\,Code(X)\,|\,\geq n)H(C_2\,|\,C_1, \cdots, C_{n-1} \cap |\,Code(X)\,|\,\geq n)$$

$$\leq \mathbb{P}(\,|\,Code(X)\,|\,\geq 1) + \mathbb{P}(\,|\,Code(X)\,|\,\geq 2) + \cdots + \mathbb{P}(\,|\,Code(X)\,|\,\geq n)$$

$$\leq \sum_{i=1}^{n} \mathbb{P}[\,|\,Code(X)\,|\,\geq j] = \mathbb{E}[\,|\,Code(X)\,|\,]$$

L'entropie de Shannon est une borne fondamentale de la compression !

On a égalité lorsque,

$$H(C_1) = H(C_2) = \cdots = H(C_n) = 1$$

Inégalité de Jensen & Divergence de Kullback-Leibler

Kahoot! 6

Soit f une fonction convexe et X une variable aléatoire réelle dont l'espérance : $\mathbb{E}[f(X)]$ existe, alors

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

On définit la divergence de Kullback-Leibler entre deux mesure de probabilité P et Q comme :

$$\begin{aligned} D_{KL}(\mathbb{P}\,||\,\mathbb{Q}) &= \mathbb{E}_{\mathbb{P}}\left(-\log\left(\frac{\mathbb{Q}}{\mathbb{P}}\right)\right) \\ &= \mathbb{E}_{\mathbb{P}}\left(-\log\left(\mathbb{Q}\right)\right) - \underbrace{\mathbb{E}_{\mathbb{P}}\left(-\log\left(\mathbb{P}\right)\right)}_{H_P(X)} \end{aligned}$$

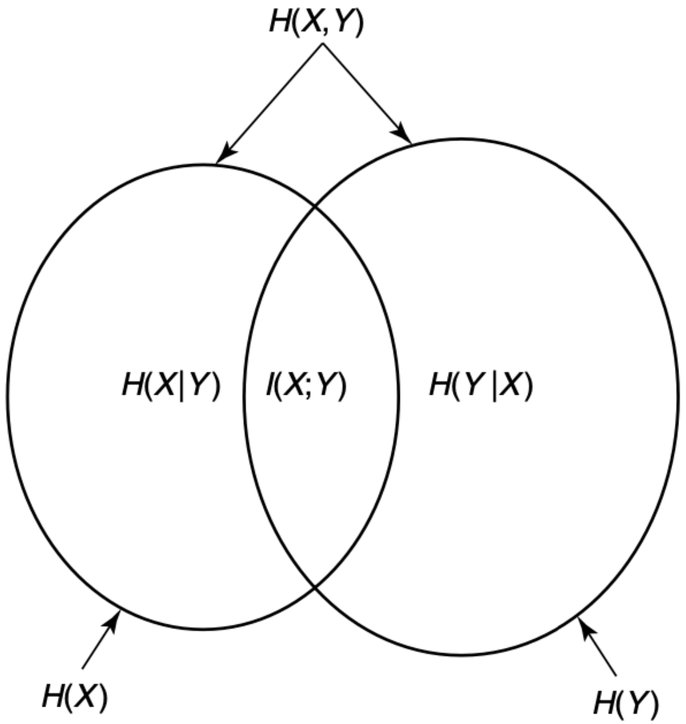
Information mutuelle

L'information mutuelle mesure la quantité d'information apportée en moyenne par une réalisation de X sur les probabilités de réalisation de Y.

$$I(X,Y) = D_{KL}(\mathbb{P} \,||\, \underbrace{\mathbb{P}_X \otimes \mathbb{P}_Y}_{\mathbb{P}_X \otimes \mathbb{P}_Y(x,y) = \mathbb{P}_X \mathbb{P}_Y \neq \mathbb{P}_{X,Y}(x,y)})$$

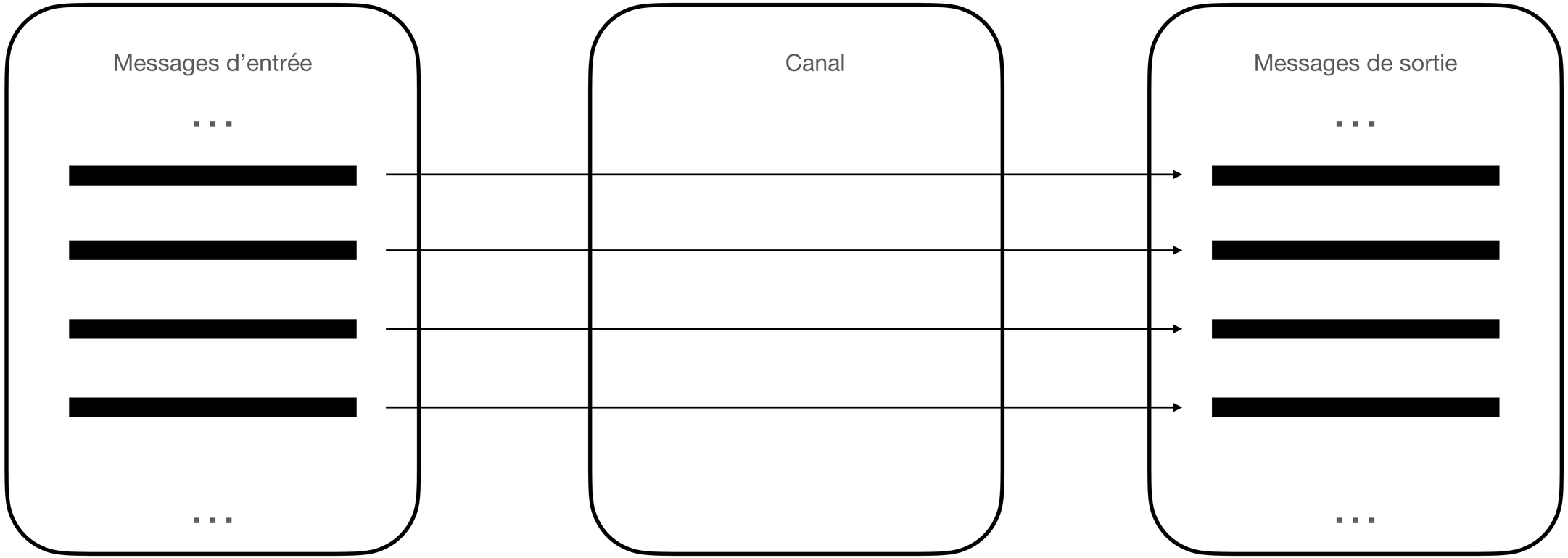
Si $X \perp Y, I(X,Y) = 0$

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



Kahoot! 7-8-9

Canal sans bruit



Théorème fondamental du canal sans bruit

Nous allons maintenant justifier notre interprétation de H comme étant le taux de génération d'information en prouvant que **H détermine la capacité du canal requise avec le codage le plus efficace**.

Soit une source d'information qui a une entropie H (bits par symbole) et un canal de capacité C (bits par seconde)

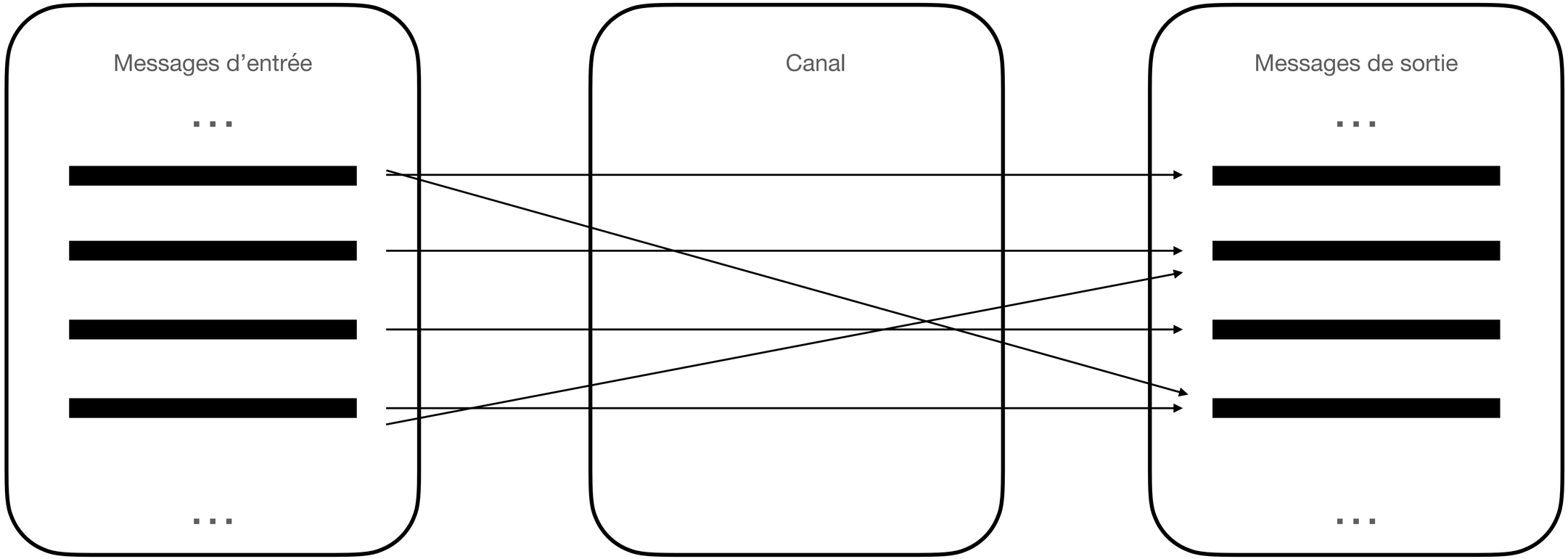
1. Alors, **il est possible** d'encoder la sortie de la source de telle manière de transmettre à un taux moyen de :

$$\frac{C}{H} - \epsilon$$

symboles par seconde où epsilon est arbitrairement petit.

2. Il **n'est pas possible** de transmettre à un taux moyen plus grand que $\frac{C}{H}$

Canal bruité



Canal qui introduit de l’incertitude dans la transmission des messages !

Les messages peuvent être altéré ou pas par le canal avec certaines roba ! —> connaitre ce qui est pertinent sur la canal

Théorème fondamental du canal bruité

Soit une source d'information qui a une entropie H (bits par symbole) et un canal de capacité C (bits par secondes)

1. Si H <= C Alors, **il est possible** d'encoder la sortie de la source de telle manière de transmettre avec une fréquence d'erreurs arbitrairement faible

3. Si H > C Alors, **il est possible** d'encoder la sortie de la source de telle manière les cas équivoques représentent moins de $H - C + \epsilon$. Il n'est pas possible de d'avoir des cas équivoques moindres que $H - C$

Bibliographie

A Mathematical Theory of Communication : Claude Shannon (1948)

Elements of Information Theory : Cover & Thomas

Wikipédia : Théorie de l’information de Shannon

Youtube : Passe-science

Youtube : Conférence Inria

Youtube : Science 4 all

Merci beaucoup pour votre attention !