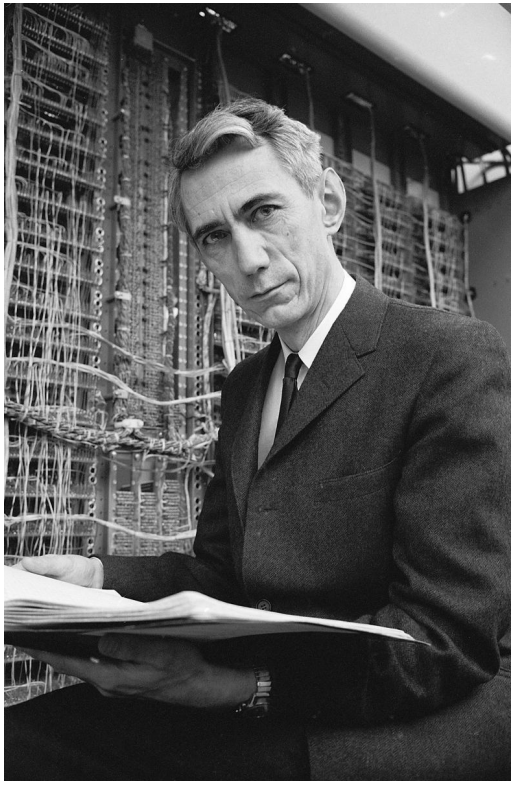
Shannon, son entropie et les statistiques

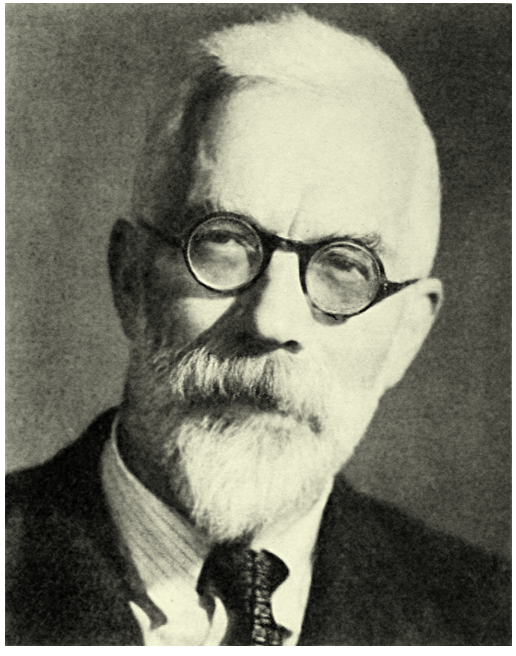
E. Le Pennec

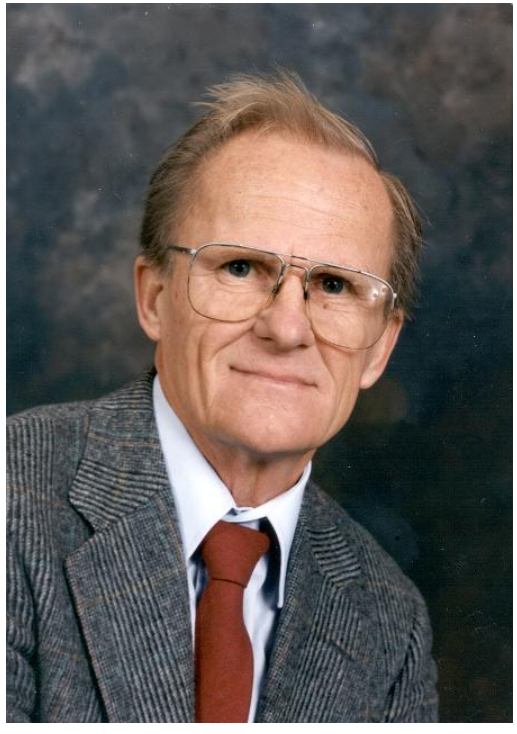
École polytechnique

ENSAE - Mars 2019

Claude Shannon, son entropie et les statistiques



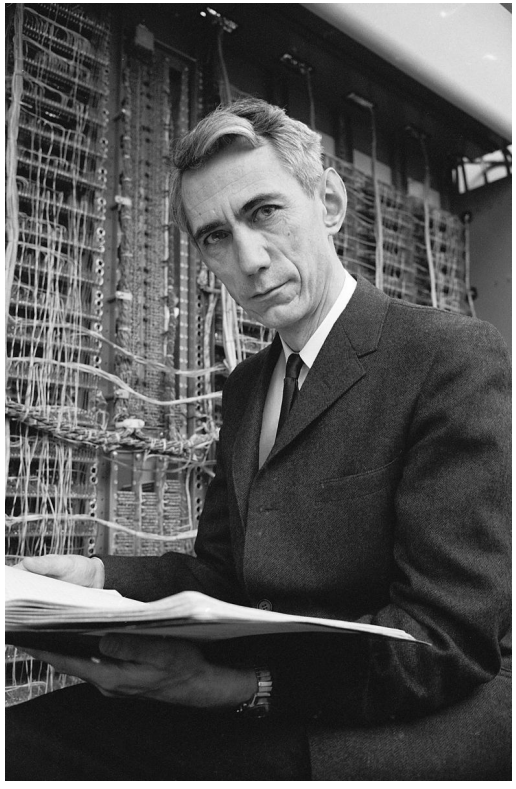




3 actes

- Compression et entropie (Claude Shannon – 1916-2001)
- Estimation et vraisemblance (Ronald Fisher – 1890-1962)
- Estimation et compression (Jorma Rissanen – 1932-)

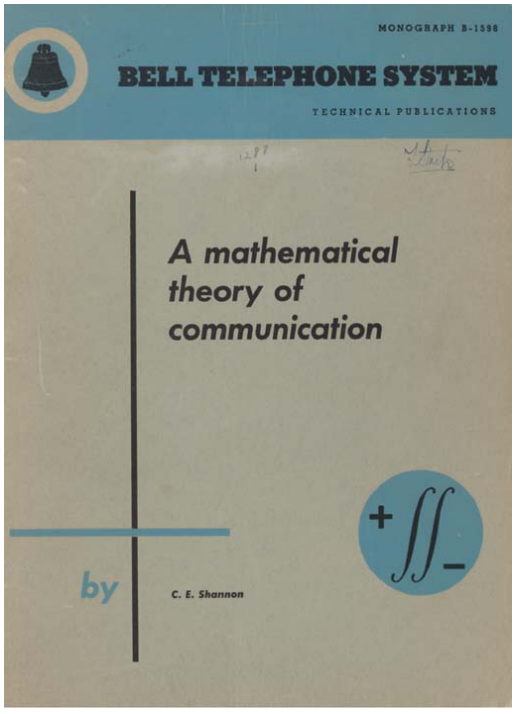
Claude Shannon



Claude Shannon – 1916-2001

- Ingénieur en génie électrique **et** mathématicien
- Père de la **théorie de la communication**

A mathematical theory of communication



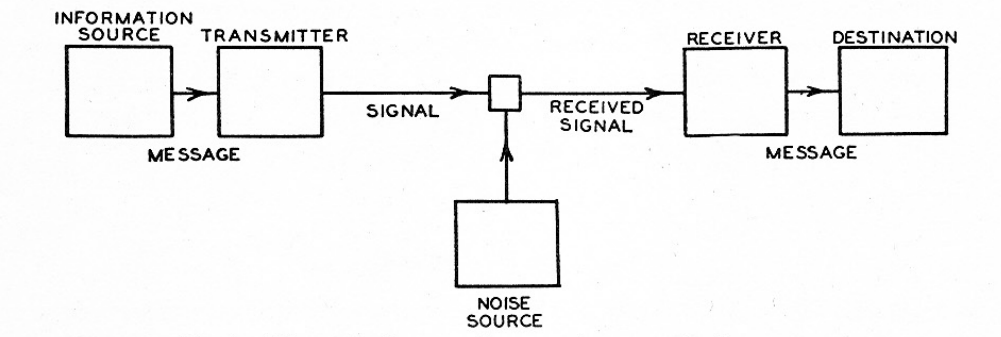


Fig. 1—Schematic diagram of a general communication system.

Article fondateur de 1948

- Modèle de la **transmission d'information**.
- Mise en avant de **limitations intrinsèques**.
- Proposition de **méthodes pratiques** permettant de s'en approcher.

- Contribution **majeure** ayant changé la face du monde !

Le télégraphe, un canal discret sans bruit

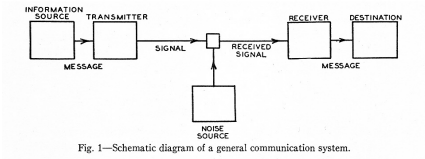


Fig. 1—Schematic diagram of a general communication system.

Communiquer avec un télégraphe

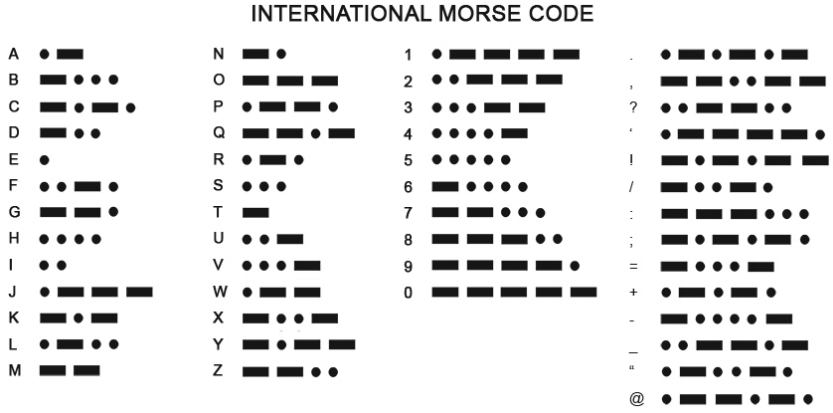
- **Message envoyé** : succession de lettres, de chiffres, d'espace et de signes de ponctuations.
- **Signal envoyé** : succession de points, traits et espaces.
- **Signal reçu** : succession de points, traits et espaces.
- **Message reçu** : succession de lettres, de chiffres, d'espace et de signes de ponctuations.

But

- Message reçu = Message envoyé
- Signal le plus **court** possible...

- **Sans bruit** (hypothèse forte) : Signal envoyé = Signal reçu

Code Morse



Une vieille invention

- Proposé en **1832** par Samuel Morse.
- Version standardisée par l'UIT en 1865.
- **Code** = Correspondance entre symboles du message et suite de symboles du signal.
- **Durée variable** des codes :
  - un tiret dure 3 points
  - espace d'une durée de 1 points entre les *lettres* et de 7 points entre les *mots*.
- **Longueur** du code lié à la **fréquence** de la lettre...

Informatique et compression

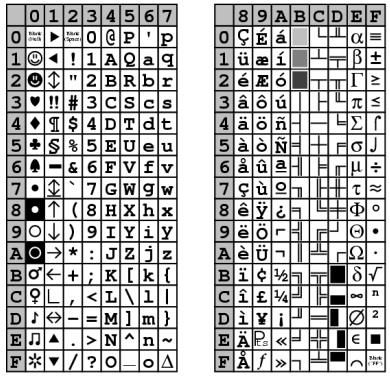
Compression de fichiers

- But : **réduire** l'espace nécessaire pour stocker l'information contenue dans un fichier.
- **Compression** : production d'une suite de 0/1 la plus courte possible permettant de **revenir au fichier initial** !

La compression : une transmission sur place

- Transmission $\simeq$ lecture/écriture
- **Message** = suite de 0/1 = fichier d'**entrée**
- **Signal** = suite de 0/1 = fichier de **sortie**
- **Durée** du signal = **longueur** du fichier de sortie

Alphabet et code binaire



$$\mathcal{C} : \bigcup_{n\in\mathbb{N}} \mathcal{A}^n \to \bigcup_{\ell\in\mathbb{N}} \{0,1\}^\ell$$
$$\mathbf{W} \mapsto \mathcal{C}(\mathbf{W})$$

Ex : aabee $\mapsto$ 0010001...

Symboles, mots et code

- **Alphabet** $\mathcal{A}$: liste de $|\mathcal{A}|$ caractères (**symboles**)
- **Message** $\mathbf{W} \in \bigcup_{n\in\mathbb{N}} \mathcal{A}^n$: suite de $|\mathbf{W}|$ symboles (**mot**)

- Exemple : texte codé en **ASCII**.

Codage binaire

- Code binaire : suite finie de $0/1 \in \bigcup_{\ell\in\mathbb{N}} \{0,1\}^\ell$
- Codage $\mathcal{C}$ transforme un **mot** $\mathbf{W}$ de **taille** $|\mathbf{W}|$ en un **code binaire** $\mathcal{C}(\mathbf{W})$ de **longueur** $|\mathcal{C}(\mathbf{W})|$.

Quel code ?

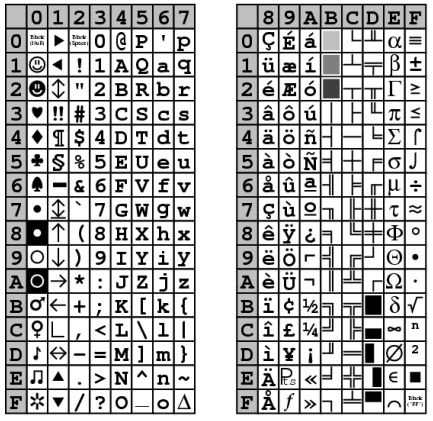
	Code A	Code B	Code C	Code D
a	10	0	0	0
b	11	10	01	10
c	111	110	011	11

Propriété d'un bon code

- Uniquement décodable :
  - Non singulier **(décodage possible)** :
    $$\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{W}') \quad \text{ssi} \quad \mathbf{W} = \mathbf{W}'.$$
  - Code par extension **(codage facile)** :
    $$\mathcal{C}(\mathbf{W}) = \mathcal{C}(W_1) \dots \mathcal{C}(W_{|\mathbf{W}|})$$
- Préfixe **(décodage facile)** : code tel que le code d'un symbole ne peut être le début d'un autre.

- Seul le côté non singulier est indispensable... mais les deux autres ne sont pas (trop) limitant..
- **Bon code** : code tel que les **codes** $\mathcal{C}(\mathbf{W})$ soient **courts**.

Un codage simple



Le code le plus simple

- Faire une **liste des symboles** et coder chaque symbole par sa **position** en binaire dans cette liste de taille $|\mathcal{A}| = 2^{\log_2 |\mathcal{A}|}$.
- Code uniquement décodable et préfixe utilisant $\lceil \log_2 |\mathcal{A}| \rceil$ **bits par symbole**.

- **Coût par symbole** d'un message :

$$\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|} = \frac{1}{|\mathbf{W}|} \sum_{i=1}^{|\mathbf{W}|} \lceil \log_2 |\mathcal{A}| \rceil = \lceil \log_2 |\mathcal{A}| \rceil$$

- Exemple : code ASCII des caractères sur 8 bits.

Une première limitation

$$\emptyset$$

$$0 \text{ - } 1$$

$$00 \text{ - } 01 \text{ - } 10 \text{ - } 11$$

$$000 \text{ - } 001 \text{ - } 010 \text{ - } 011 \text{ - } 100 \text{ - } 101 \text{ - } 110 \text{ - } 111$$

0000 - 0001 - 0010 - 0011 - 0100 - 0101 - 0110 - 0111 - 1000 - 1001 - 1010 - 1011 - 1100 - 1101 - 1110 - 1111

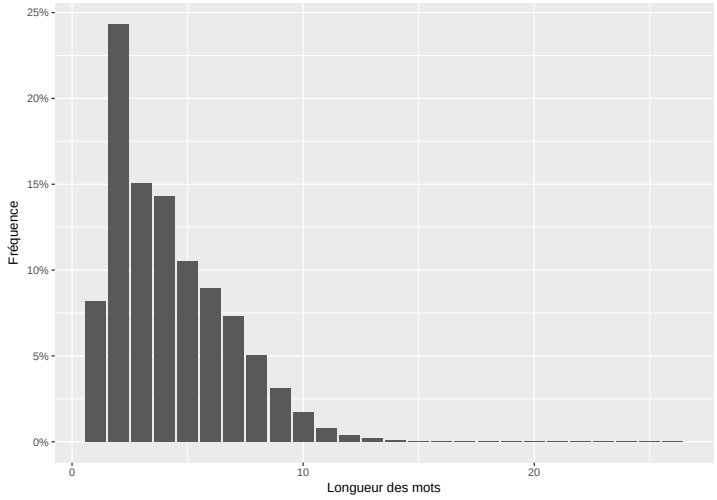
- **Non singulier :** $\mathcal{C}(\mathbf{W}) = \mathcal{C}(\mathbf{W}')$ ssi $\mathbf{W} = \mathbf{W}'$.
- Assure que le **décodage** est **possible**...
- **Contrainte** forte sur la **taille** des codes !

Limite de compression

- Pour tout codage $\mathcal{C}$, si il existe $\mathbf{W}$ tel que $\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|} < \log_2 |\mathcal{A}|$ alors il existe $\mathbf{W}'$ tel que $\frac{|\mathcal{C}(\mathbf{W}')|}{|\mathbf{W}'|} > \log_2 |\mathcal{A}|$
- **Impossible de toujours faire mieux que le codage simple !**
- Lemme des tiroirs...

- **Impossibilité** d'un algorithme de compression qui compresse tout...

Le cas du langage

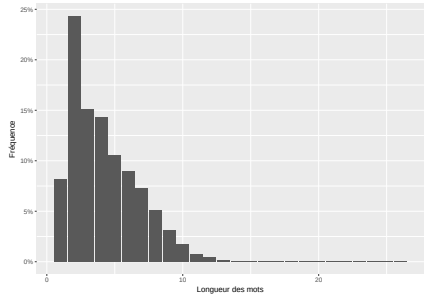


Longueur des mots et fréquence

- Mots **les plus courts** sont **les plus utilisés**... ou mots **les plus utilisés** sont **les plus courts** ?
- Phénomène d'évolution naturelle pour **diminuer la longueur moyenne** des phrases !

- Efficacité du français :
  - Codage simple des **100 000 mots en français** possible avec des mots de **4 lettres** uniquement...
  - Pas très pratique pour l'apprentissage humain... mais pas pour un ordinateur...

Code court... en moyenne



- Comment reprendre ce principe pour des sources quelconques ?
- Trouves des codes **efficaces en moyenne** seulement...

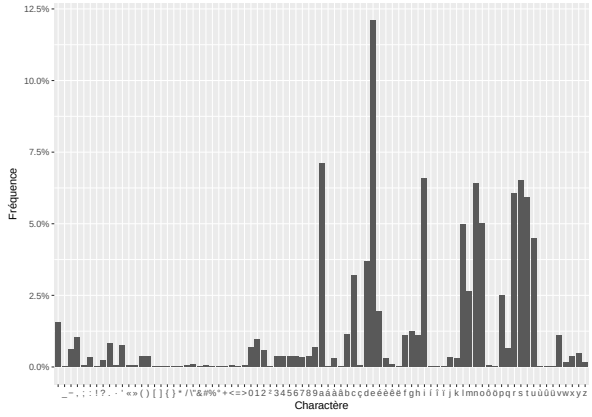
Comportement moyen

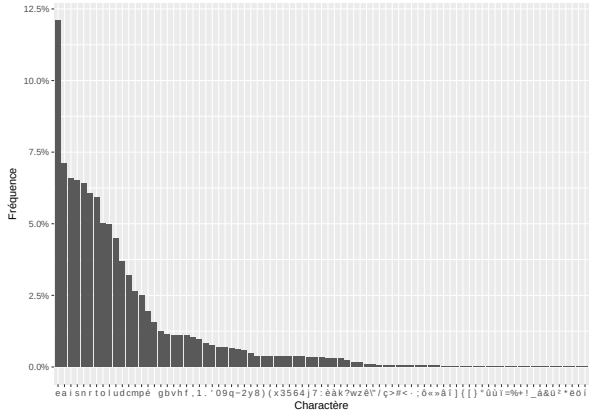
- **Loi de probabilité** $\mathbb{P}$ sur les mots $\mathbf{W}$ de taille n (répartition des **fréquences** de tous les mots)
- Coût moyen par caractère pour des mots de taille n :

$$\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] = \sum_{\mathbf{W}\in\mathcal{A}^n} \mathbb{P}\{\mathbf{W}\}\,\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}$$

- On peut se permettre des **codes longs** pour les **mots peu fréquents** !

Modèle probabiliste





Un modèle probabiliste simple des mots

- **Symboles** (lettres) ont une certaine **fréquence** :

$$\mathbb{P}\{A\}$$

- Mots obtenus en tirant de manière **successive et indépendamment** des lettres :

$$\mathbb{P}\{\mathbf{W}\} = \mathbb{P}\{W_1\} \dots \mathbb{P}\left\{W_{|\mathbf{W}|}\right\}$$

- **Modélisation grossière** pouvant être améliorée.
- Tous les résultats restent valables pour des sources Markoviennes...

L'entropie comme limite

$$\forall \mathcal{C}, \quad \mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \geq H(\mathbb{P})$$

Résultat fondamental obtenu par Shannon

- **Impossible de compresser mieux en moyenne** qu'une quantité notée $H(\mathbb{P})$ dépendant de la source.
- Shannon appelle **entropie** cette quantité dont la formule est

$$H(\mathbb{P}) = \sum_{A \in \mathcal{A}} \mathbb{P}\{A\} \left(-\log_2 \mathbb{P}\{A\}\right)$$

- Résultat obtenu par un **calcul**...
- Lié à la positivité de la divergence de Kullback-Leibler...
- et le fait que pour tout code $2^{-|\mathcal{C}(\mathbf{W})|}$ est une distribution de probabilité (à une renormalisation par $Z < 1$ près)

Entropie et longueur de code

$$\forall \mathcal{C}, \quad \underbrace{\frac{1}{n}\sum_{\mathbf{W}\in\mathcal{A}^n} \mathbb{P}\{\mathbf{W}\}\,|\mathcal{C}(\mathbf{W})|}_{\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right]} \geq \underbrace{\frac{1}{n}\sum_{\mathbf{W}\in\mathcal{A}^n} \mathbb{P}\{\mathbf{W}\}\,(-\log_2 \mathbb{P}\{\mathbf{W}\})}_{\frac{1}{n}H(\mathbb{P}^{\otimes n})=H(\mathbb{P})}$$

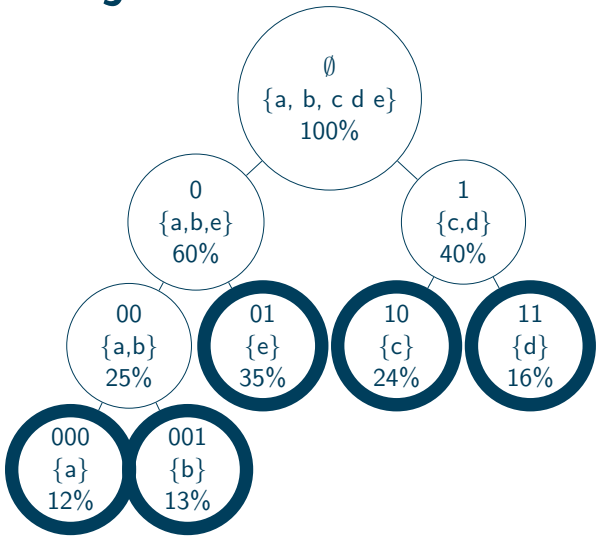
- Modélisation simple utilisée implique
  - Loi produit :
$$\mathbb{P}\{\mathbf{W}\} = \prod_{i=1}^{|\mathbf{W}|} \mathbb{P}\{W_i\} \Leftrightarrow (-\log_2 \mathbb{P}\{\mathbf{W}\}) = \sum_{i=1}^{|\mathbf{W}|} (-\log_2 \mathbb{P}\{W_i\})$$
  - Entropie des mots de tailles n : $H(\mathbb{P}^{\otimes n}) = nH(\mathbb{P})$

Longueur d'un bon code

- Mots : $|C(\mathbf{W})| \simeq (-\log_2 \mathbb{P}\{\mathbf{W}\})$
- Caractères dans les mots :
$$\sum_{i=1}^{|\mathbf{W}|} |C(W_i)| \sim \sum_{i=1}^{|\mathbf{W}|} (-\log_2 \mathbb{P}\{W_i\})$$
- Caractères : $|C(A)| \sim (-\log_2 \mathbb{P}\{A\})$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



Lettre	Code
a	000
b	001
c	10
d	11
e	01

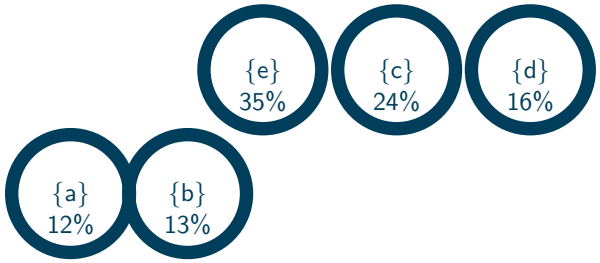
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



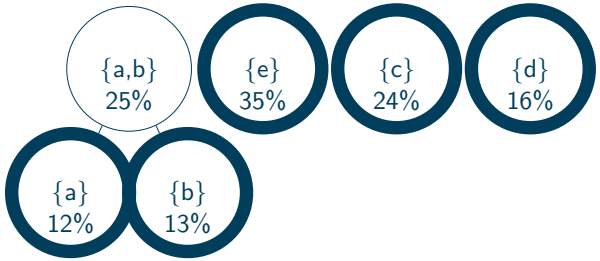
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



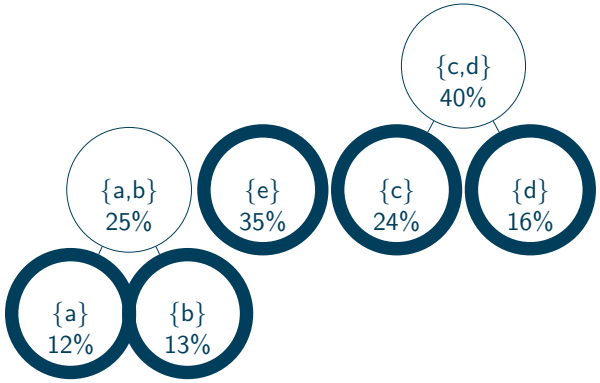
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



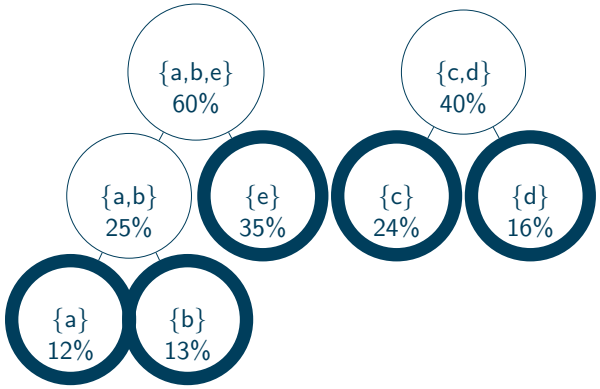
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



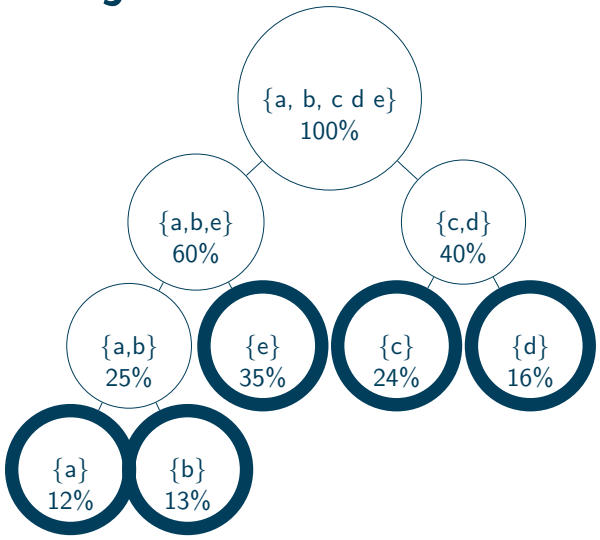
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



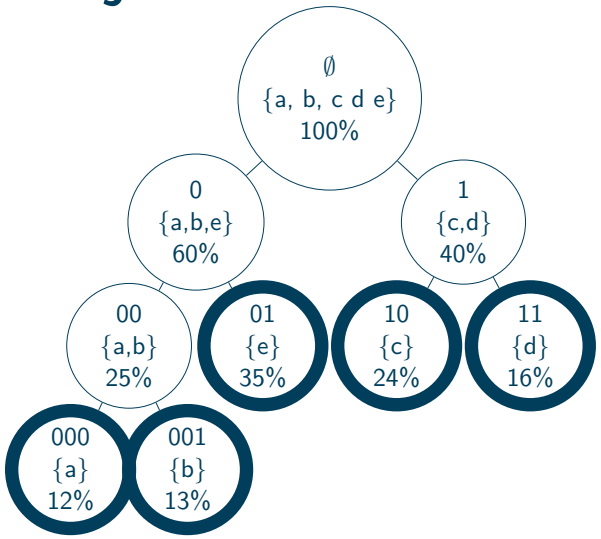
- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

L'entropie comme objectif

Lettre	Fréqu.
a	12%
b	13%
c	24%
d	16%
e	35%



Lettre	Code
a	000
b	001
c	10
d	11
e	01

- Résultat précédent sans intérêt sans **méthode explicite permettant d'obtenir un code quasi optimal**...

Codes efficaces

- **Shannon/Fano** (48) : principe de partitionnement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 2$
- **Huffman** (52) : principe de regroupement récursif des symboles donnant un code vérifiant $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + 1$
- **Codage arithmétique** (76) : $\mathbb{E}\left[\frac{|\mathcal{C}(\mathbf{W})|}{|\mathbf{W}|}\right] \leq H(\mathbb{P}) + \frac{2}{n}$

Entropie comme mesure de complexité

Entropie et complexité (Interprétation *a posteriori*)

- Complexité de la **source** et pas des messages.
- **Information** moyenne apportée par chaque symbole.
- Propriété de l'entropie compatible avec le côté information :
  - $H(X,Y) \leq H(X) + H(Y)$ (information commune à X et Y)
  - $H(X|Y) \leq H(X)$ (transfert d'information à partir de X)

- **Autres entropies ou mesure de complexité** :
  - Entropie métrique (mesure de la taille),
  - Entropie en physique statistique (mesure du désordre),
  - Complexité individuelle (longueur du code le plus court permettant de générer la suite) (Kolmogorov),

Que faire lorsque la loi est inconnue ?

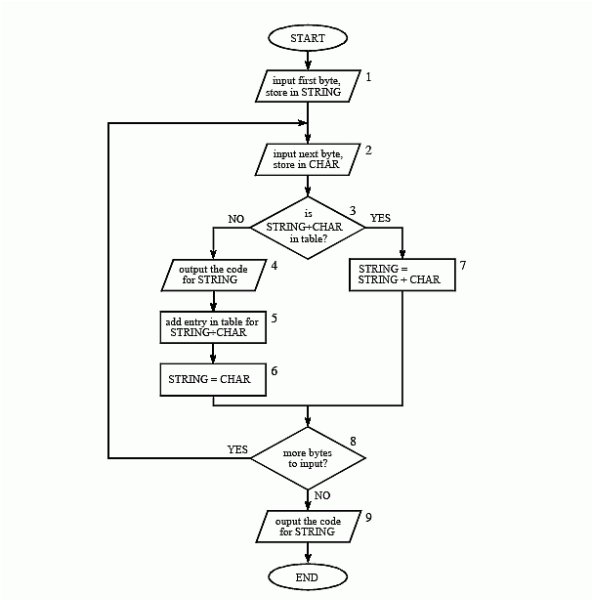


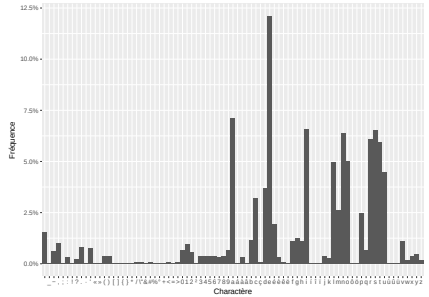
FIGURE 27-7
LZW compression flowchart. The variable, *CHAR*, is a single byte. The variable, *STRING*, is a variable length sequence of bytes. Data are read from the input file (box 1 & 2) as single bytes, and written to the compressed file (box 4) as 12 bit codes. Table 27-3 shows an example of this algorithm.

- En pratique $\mathbb{P}$ n'est **pas forcément connue**...
- Pas de problème pour l'analyse mais problème pour l'**implémentation** !

Il suffit d'**estimer** (deviner) $\mathbb{P}$...

- **Explicitement** : Huffman, Codage arithmétique...
- **Implicitement** : Approche dictionnaire (LZW, ZIP...)

Estimation de $\mathbb{P}$



- Comment **estimer** $\mathbb{P}\{A\}$ en observant un message $\mathbf{W}$, c-à-d la liste des symboles $W_1, \ldots, W_{|\mathbf{W}|}$!
- Réponse naturelle : **proportion observée**

$$\mathbb{P}\{A\} \sim \frac{\text{Nb d'occurrences de } A}{|\mathbf{W}|}$$

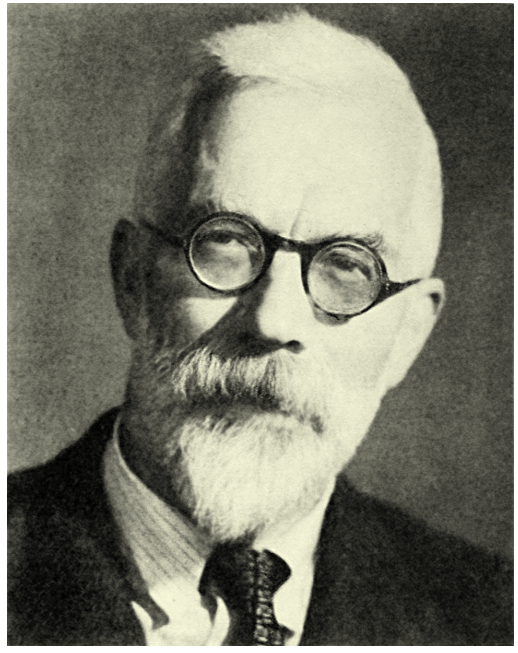
- Difficile de faire mieux avec un modèle aussi simple.

Estimation paramétrique

- Observation de $X_1, \ldots, X_n$ **indépendants de même loi** $\mathbb{P}$.
- Estimation de $\mathbb{P}$ par une **loi** $\mathbb{P}_\theta$ décrite par des **paramètres** $\theta \in \mathbb{R}^p$.

- Modèle précédent : **Histogramme** = **Multinomiale** de paramètres $\theta \in \mathbb{R}^{|\mathcal{A}|-1}$ donnant les proportions (sauf une...)

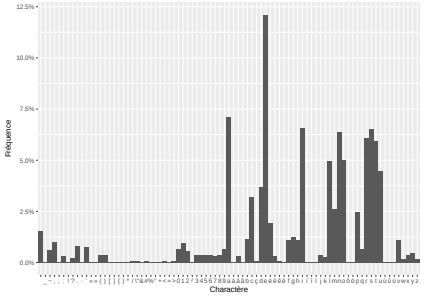
Ronald Fisher

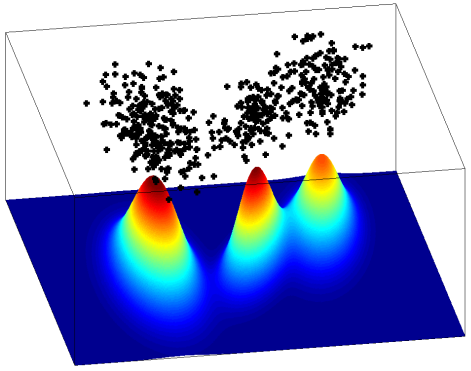


Ronald Fisher – 1890-1962

- Statisticien **et** biologiste
- **Systématisation** des techniques d'estimation.
- Un des fondateurs des statistiques modernes.

Principe de vraisemblance

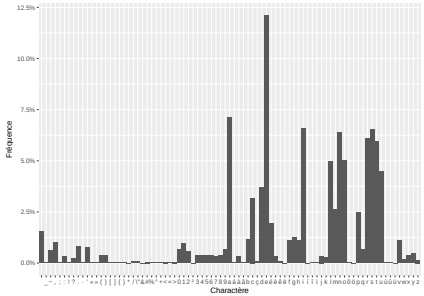


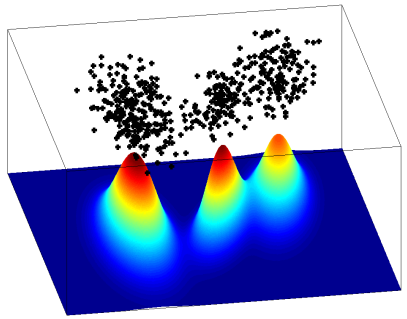


Comment choisir parmi **plusieurs** hypothèses ?

- Comment choisir de manière **systématique** les paramètres θ de $\mathbb{P}_\theta$?
- Principes de **compétition** entre les paramètres :
  - recherche des **paramètres les plus probables étant données les observations** (Bayes)
  - recherche des **paramètres dans lequel les observations sont les plus probables** (Fisher)
- Principes **similaires** mais pas de notion de loi sur les paramètres chez Fisher.

Maximum de Vraisemblance





Choix parmi une famille de loi paramétrée $\mathbb{P}_\theta$

- Famille de **loi de probabilité $\mathbb{P}_\theta$ définie à θ fixé** par
$$\mathbf{X} \mapsto \mathbb{P}_\theta\{\mathbf{X}\}$$
- **Vraisemblance à observation X** fixée donne la probabilité de $\mathbf{X}$ dans le modèle $\mathbb{P}_\theta$:
$$\theta \mapsto \mathbb{P}_\theta\{\mathbf{X}\}$$
- **Principe du maximum de vraisemblance** : choix de $\widehat{\theta}$ via
$$\widehat{\theta} = \operatorname*{argmax}_\theta \mathbb{P}_\theta\{\mathbf{X}\}$$

- Donne l'estimateur empirique par les proportions dans le cas de l'histogramme...
- Existence de résultat théorique sur l'efficacité de la méthode.

Log Vraisemblance

- Cadre classique : $\mathbb{P}_\theta\{\mathbf{X}\} = \prod_{i=1}^n \mathbb{P}_\theta\{X_i\}$

Maximum de vraisemblance et minimum de $-$ log vraisemblance

- **Reformulation** du maximum de vraisemblance :

$$\widehat{\theta} = \operatorname*{argmax}_\theta \mathbb{P}_\theta\{\mathbf{X}\} = \operatorname*{argmax}_\theta \prod_{i=1}^n \mathbb{P}_\theta\{X_i\}$$

$$= \operatorname*{argmin}_\theta \sum_{i=1}^n (-\log_2 \mathbb{P}_\theta\{X_i\})$$

- **But** : rendre $\mathbb{E}\left[\mathbb{P}_\theta\{X\}\right]$ grand, c-à-d $\mathbb{E}\left[(-\log_2 \mathbb{P}_\theta\{X\})\right]$ petit.

- **Interprétation** de l'objectif en terme de divergence de Kullback-Leibler

$$\mathrm{KL}(\mathbb{P}, \mathbb{P}_\theta) = \mathbb{E}\left[-\log_2 \frac{\mathbb{P}_\theta\{X\}}{\mathbb{P}\{X\}}\right] = \mathbb{E}\left[(-\log_2 \mathbb{P}_\theta\{X\})\right] - H(\mathbb{P})$$

- **Analogie avec la compression** si on pense à $(-\log_2 \mathbb{P}_\theta\{X\})$ comme la longueur du code associé à X...

Compression/Estimation

Compression

- **Mots** $\mathbf{W} = W_1 \dots W_{\mathbf{W}}$ provenant d'une source de **loi** $\mathbb{P}$ **connue**.
- **Code** *minimisant*
$$\mathbb{E}\left[\mathcal{C}(A)\right]$$
- **But** : rendre petit
$$\sum_{i=1}^{|\mathbf{W}|} \mathcal{C}(W_i)$$

Estimation

- **Observations** $\mathbf{X} = (X_1, \dots, X_n)$ i.i.d. de **loi** $\mathbb{P}$ **inconnue**.
- **Paramètres** θ minimisant
$$\sum_{i=1}^{n} (-\log_2 \mathbb{P}_\theta\{X_i\})$$
- **But** : rendre petit
$$\mathbb{E}\left[-\log_2 \mathbb{P}_\theta\{X\}\right]$$

Lien

- **Dualité modèle/code** :
$\mathbb{P}_\theta\{X\} \sim 2^{-|\mathcal{C}(X)|} \Leftrightarrow (-\log_2 \mathbb{P}_\theta\{A\}) \sim |\mathcal{C}(A)|$
- **Dualité KL/entropie** :
$$\mathrm{KL}(\mathbb{P}, \mathbb{P}_\theta) = \mathbb{E}\left[-\log_2 \mathbb{P}_\theta\{X\}\right] - H(\mathbb{P}) \sim \mathbb{E}\left[\mathcal{C}(A)\right] - H(\mathbb{P})$$

Compression/Estimation

Compression

- Trouver un **code** rendant **petite la longueur moyenne** pour une source de **loi $\mathbb{P}$ connue**.
- L'utiliser sur un mot tiré au hasard selon $\mathbb{P}$.
- **But** : différence entre la longueur observée et l'optimale petite.

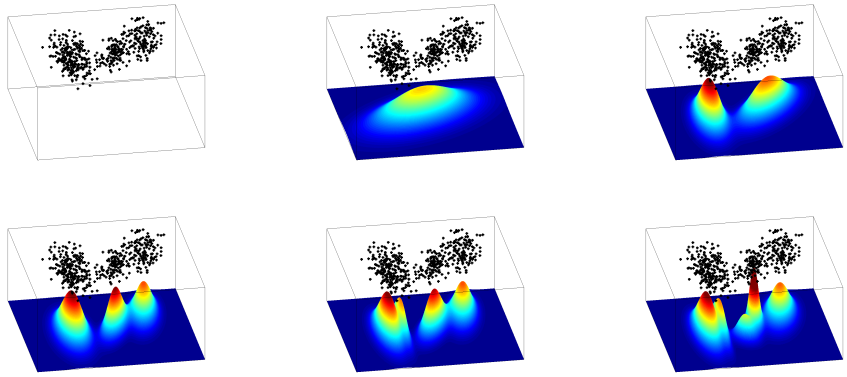
Estimation

- Trouver des **paramètres** rendant **petite la** - log **vraisemblance** d'une observation de **loi $\mathbb{P}$ inconnue**.
- Les utiliser pour dire des choses sur la loi $\mathbb{P}$.
- **But** : différence entre la loi estimée et la vraie loi petite.

Lien

- Équivalence entre **code et loi de probabilité**.
- Équivalence entre les **objectifs**.
- Utilisation différente...

Choisir parmi des modèles



- Modèles de **complexité variable** :
  - Modèles simples avec peu de paramètres,
  - Modèles compliqués avec beaucoup de paramètres.

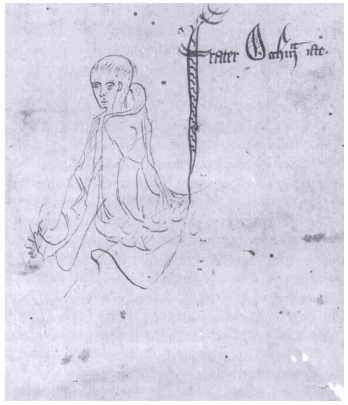
La théorie du complot

- **Plus** de **supputations** permettent **toujours** de **mieux** s'adapter aux **observations**...
- Mais ne garantissent **pas une bonne explication**.

Le sur-apprentissage

- **Plus** de **paramètres** permettent **toujours** de **mieux** s'adapter aux **observations**...
- Mais ne garantissent **pas une bonne estimation**.

Guillaume d'Ockham



Guillaume d'Ockham – (1285 - 1347)

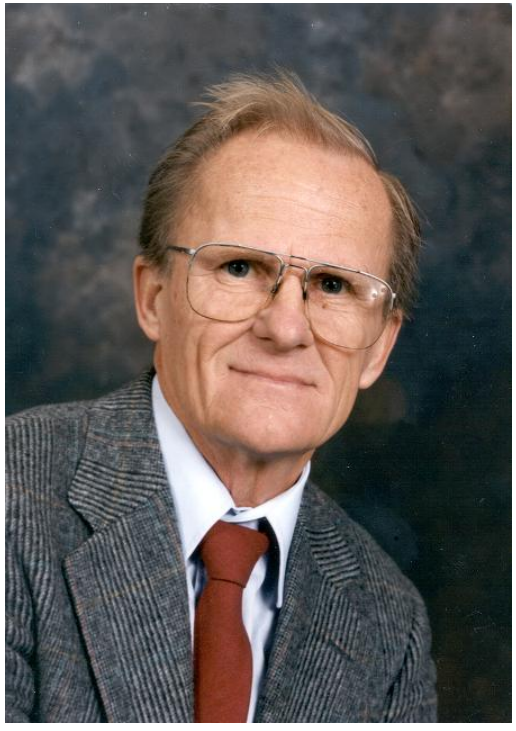
- Moine **et** philosphe.
- Inventeur du rasoir... d'Ockham.

Pluralitas non est ponenda sine necessitate

- *Les multiples ne doivent pas être utilisés sans nécessité.*
- Principe de **concision** : minimum possible d'hypothèses...

- Albert Einstein (1934) : *Tout doit être le plus simple possible, mais pas plus simple que ça.*
- **Pas** un critère **très précis**...

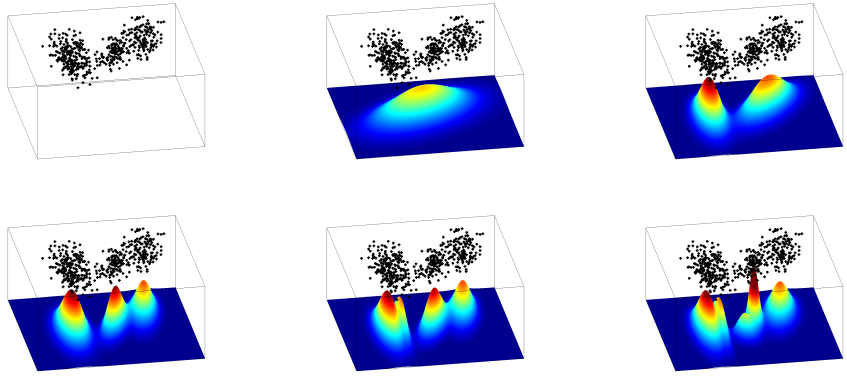
Jorma Rissanen



Jorma Rissanen – 1932-

- Théoricien de l'information (frontière **math/info**)
- Inventeur du codage arithmétique et père du principe MDL
- Kolmogorov Medal (2006) and Claude E. Shannon Award (2009).

MDL

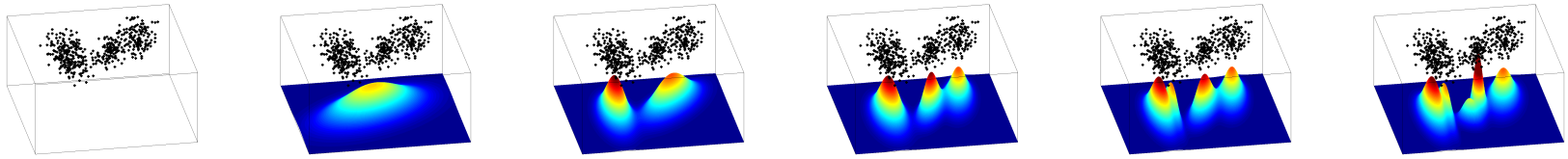


$$\min \mathcal{C}(\mathbf{X}, \theta)$$

Minimum Description Length

- Repenser l'**estimation** en terme de **codage**.
- Idée clé : coder à la fois le **message** (les observations) **et** les **paramètres** !
- Principe MDL : Recherche du modèle rendant la **description** (le code) **de longueur minimale**.
- Principe explicite de **concision**... à la définition du code près !

MDL et code en deux étapes

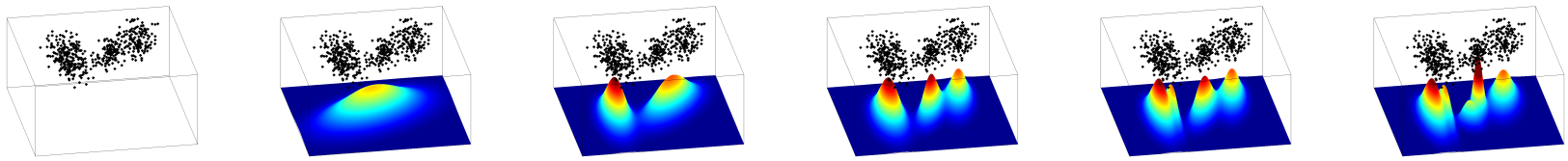


$$\mathcal{C}(\mathbf{X}, \theta) \sim \sum_{i=1}^{n} (-\log_2 \mathbb{P}_\theta\{X_i\}) + \log_2 |\Theta|)$$

Code en deux étapes - θ et X_i discrets

- **Codage en deux temps** : codage du **paramètre θ suivi** du codage des **observations X_i** avec la loi $\mathbb{P}_\theta$.
- **Coût** de codage :
  - Paramètres : $\sim \log_2 |\Theta|$,
    codage par la position dans la liste...
  - Observations : $\sum_{i=1}^{n} -\log_2 \mathbb{P}_\theta\{X_i\})$,
    minimum pour θ **maximum de vraisemblance**.
- **Pb** : la plupart du temps les paramètres et les observations sont continues...

MDL et code en deux étapes

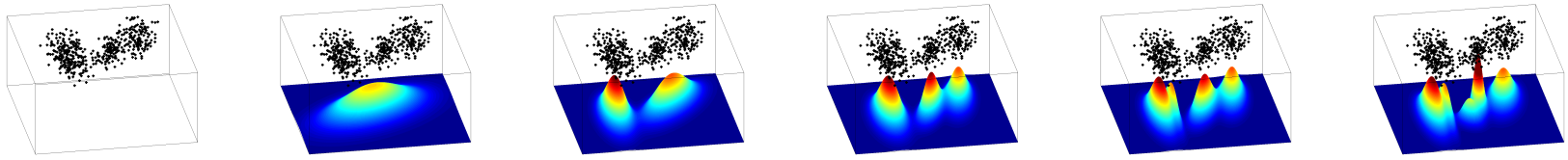


$$\mathcal{C}(\mathbf{X}, \theta) \sim \sum_{i=1}^{n} (-\log_2 \mathbb{P}_\theta\{X_i\}) + \frac{\log_2(n)}{2} D_\Theta$$

Code en deux étapes - θ (et X_i) continu

- **Codage en deux temps après dicrétisation** : codage du **paramètre discrétisé** $\widetilde{\theta}$ **suivi** du codage des **observations discrétisés** $\widetilde{X_i}$ avec la loi $\mathbb{P}_{\widetilde{\theta}}$.
- Différence **négligeable** si la discrétisation est en $\mathbf{1/\sqrt{n}}$...
- **Coût** de codage :
  - Paramètres : $\sim D_\Theta \log_2(n)/2$,
    lien avec une notion de **dimension entropique métrique**...
  - Observations : $\sum_{i=1}^{n} -\log_2 \mathbb{P}_{\widetilde{\theta}}\{\widetilde{X}_i\}) \sim \sum_{i=1}^{n} -\log_2 \mathbb{P}_\theta\{X_i\})$,
    minimum pour θ **maximum de vraisemblance**.
- Amélioration possible avec un codage implicite de θ...

MDL et code en deux étapes

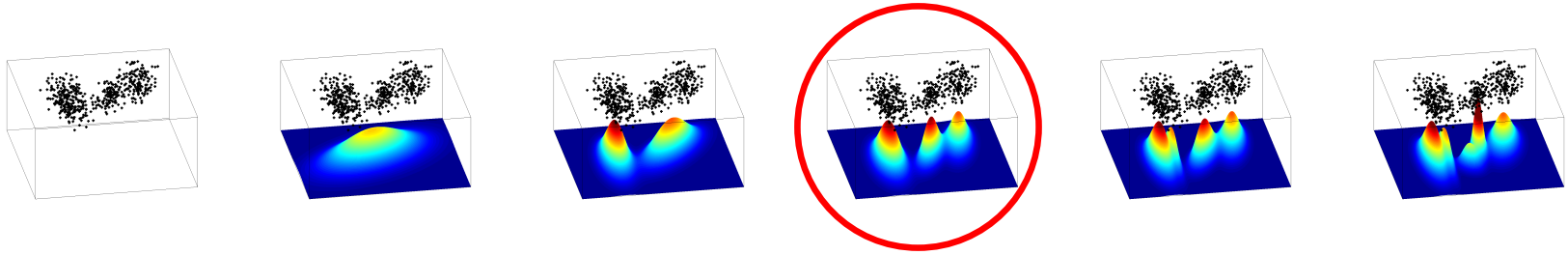


$$\mathcal{C}(\mathbf{X}, \theta) \sim \sum_{i=1}^{n} (-\log_2 \mathbb{P}_\theta\{X_i\}) + \frac{\log_2(n)}{2} D_\Theta$$

Minimum Description Length et Équilibre Biais/Variance

- Principe de **pénalisation** par la dimension sans autre justification qu'**heuristique** !
- **Équilibre** entre l'**amélioration** du codage par complexification des modèles et le **coût** lié à cette complexité.
- Principe d'**équilibre** entre le **biais** et la **variance** (entre la **qualité** du modèle et la **difficulté** à en estimer les paramètres)
- **Preuves** d'efficacité ont **ensuite** été **obtenues**...

MDL et code en deux étapes



$$\mathcal{C}(\mathbf{X}, \theta) \sim \sum_{i=1}^{n} (-\log_2 \mathbb{P}_\theta\{X_i\}) + \frac{\log_2(n)}{2} D_\Theta$$

Minimum Description Length et Équilibre Biais/Variance

- Principe de **pénalisation** par la dimension sans autre justification qu'**heuristique** !
- **Équilibre** entre l'**amélioration** du codage par complexification des modèles et le **coût** lié à cette complexité.
- Principe d'**équilibre** entre le **biais** et la **variance** (entre la **qualité** du modèle et la **difficulté** à en estimer les paramètres)
- **Preuves** d'efficacité ont **ensuite** été **obtenues**...

Shannon, son entropie et les statistiques

Parcours entre compression et estimation autour de Shannon et son entropie

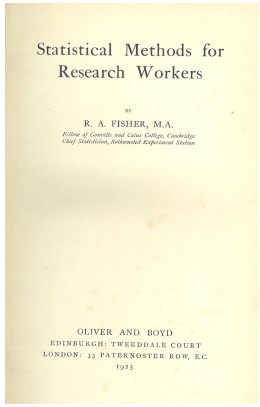
- Le coeur de la théorie du codage, **(Shannon)**
- Lien avec le principe de maximum de vraisemblance, **(Fisher)**
- Esquisse de la combinaison possible entre codage et estimation pour la sélection de modèles **(Rissanen)**

Puissance de la modélisation mathématiques

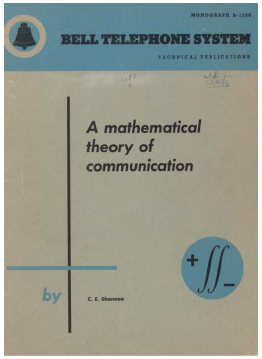
- **démontrer** des limitations intrinsèques,
- **évaluer** la qualité des algorithmes,
- **réinterpréter** des méthodes déjà connues,
- **proposer** des nouvelles...

- Compléments d'information :
  - `Erwan.Le-Pennec@polytechnique.edu`
  - `http://www.cmap.polytechnique.fr/~lepennec`

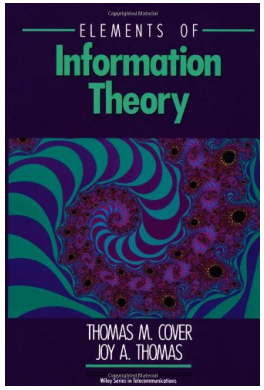
Pour aller plus loin



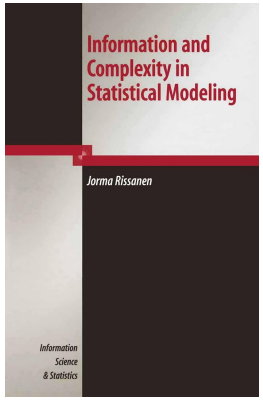
1925



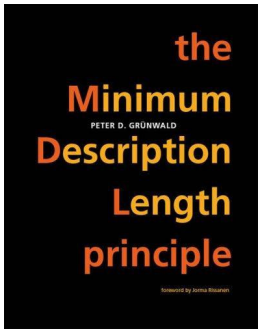
1948



1991



2007



2007

Références

- **C. Shannon**, *A mathematical theory of communication*
- **Th. Cover and J. Thomas**, *Elements of Information Theory*
- **R. Fisher**, *Statistical Methods for Research Workers*
- **J. Rissanen**, *Information and complexity in statistical modeling*
- **P. Grünwald**, *The Minimum Description Length principle*

- **Autres pistes** : Information de Fisher, Inégalités statistiques liées à la théorie de l'information, Codage par dictionnaire et modèle stationnaire, Complexité de Kolmogorov...