Shannon et la théorie de l'information

par Olivier Rioul
Ingénieur général des Mines,
Enseignant-chercheur à Télécom ParisTech

Claude Elwood Shannon est né en 1916 dans le Michigan aux U.S.A., où il grandit dans la petite ville de Gaylord. Il est resté toute sa vie un enfant curieux, inventif et joueur. Adolescent, il bricole son propre télégraphe sur des fils barbelés pour communiquer en Morse avec un voisin. Il joue du cor et de la clarinette, s'intéresse au jazz ; il est surtout passionné d'énigmes, de cryptogrammes, de gadgets et de jonglage.

Figure 1 : Photo de classe de Shannon adolescent.

Précoce, il entre à l'université du Michigan à 16 ans, où il étudie à la fois l'ingénierie électrique et les mathématiques. Diplômé en 1936, il effectue un stage au MIT comme assistant programmeur de l'« analyseur différentiel » (une machine analogique à résoudre des équations différentielles d'ordre 2), sous la direction de Vannevar Bush[1], qui deviendra son mentor. Les interrupteurs à relais permettent de contrôler la machine, ce qui amène Shannon à faire une étude systématique des circuits de relais. À l'aide de ses connaissances mathématiques, il fait le lien avec le formalisme symbolique de l'algèbre de Boole : son master de 1937 révolutionne l'utilisation des circuits logiques. Shannon recevra le prix Alfred Noble en 1940 pour sa mémoire de master, considéré comme le plus important du siècle.

Sa thèse, soutenue au MIT en 1940, développe une algèbre appliquée à la génétique, mais sans contact avec les praticiens de cette discipline ; elle ne sera pas publiée tout de suite et restera relativement méconnue. Il faut dire que dès 1939, la préoccupation majeure de Shannon concerne ce qu'il nomme à l'époque

1. Vannevar Bush (1890-1974), ingénieur, chercheur et conseiller politique à la Maison-Blanche.

« la transmission de l'intelligence » et ce qui deviendra célèbre sous le nom de *théorie de l'information*.

En 1940, jeune marié avec une intellectuelle juive new-yorkaise, Norma Levor, il effectue un post-doc à l'*Institute of Advanced Study* de Princeton, sous la direction du mathématicien Hermann Weyl. Contrairement à une légende, il n'y fréquentera pas particulièrement Albert Einstein, mais rencontre John von Neumann. L'environnement scientifique est à l'époque focalisé sur la physique et ne permet pas à Shannon de s'épanouir. Shannon est alors très solitaire et sa jeune épouse désespère de le trouver déprimé ; elle divorcera un an plus tard.

@@@@@@@

L'effort de guerre commence aux U.S.A. et Vannevar Bush affecte Shannon à un projet militaire de commandes de tir pour la défense anti-aérienne, aux laboratoires Bell à New York. Les commandes étant affectées de bruit, il existe une analogie avec le problème de la réduction du bruit dans les télécommunications, qui intéresse Shannon. C'est à ce moment qu'il découvre les travaux de Nyquist et de Hartley[2] publiés dans le *Bell System Technical Journal*, la revue spécialisée des laboratoires Bell, qui vont avoir une influence déterminante sur sa théorie de la communication.

Shannon travaille également sur un système de communication de parole cryptée entre les États Unis et l'Angleterre, utilisant l'algorithme du « masque jetable » (*one-time pad*) dont il démontrera la sécurité. Début 1943, il échange à l'heure du thé avec le Britannique Alan Turing (qui venait de décrypter la fameuse machine Enigma utilisé par la marine allemande), alors en visite secrète aux laboratoires Bell. Mais leurs travaux étant secrets, les conversations des deux hommes portèrent sur d'autres sujets, comme la théorie de la calculabilité et les machines pensantes. Le mémoire de Shannon de 1945 qui résume ses travaux en cryptographie est classé secret défense et ne sera publié qu'en 1949. Ce mémoire reprend la notion d'entropie du langage et d'autres aspects de la nouvelle théorie de la communication qu'il met au point à cette époque.

Shannon commence à devenir un chercheur en vue, remarqué également par son habitude à circuler sur un monocycle tout en jonglant avec trois balles dans les couloirs du laboratoire. C'est encore aux laboratoires Bell qu'il rencontre Betty

2. Harry Nyquist (1889-1976), informaticien américain d'origine suédoise, et Ralph Hartley (1888-1970), électronicien américain.

Moore, embauchée pendant la guerre comme analyste numérique, qu'il épousera début 1949.

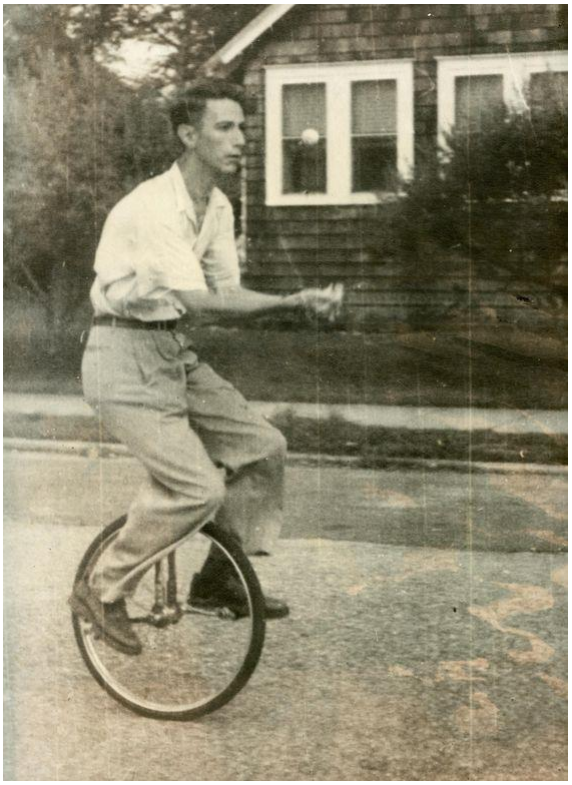

Figure 2 : Shannon jonglant sur un monocycle *(coll. famille Shannon)*

ENFIN, PARAÎT L'ARTICLE DE 1948

En 1948, Claude Elwood Shannon publie enfin « Une théorie mathématique de la communication » (*A Mathematical Theory of Communication*), en deux parties, dans les numéros de juillet et d'octobre du *Bell System Technical Journal*. Avec le recul, il s'agit là d'un des travaux scientifiques qui a exercé le plus d'influence sur notre monde moderne, et qui a fait du modeste Shannon une véritable icône dans le monde scientifique – bien qu'inconnue du grand public. Cependant, à l'époque cet article, se situant à la frontière entre ingénierie et mathématiques, n'a pas été compris immédiatement par tous : d'un côté, la plupart des ingénieurs n'avaient alors pas le niveau mathématique suffisant en probabilités pour comprendre les théorèmes de Shannon, et d'un autre côté, certains mathématiciens trouvaient le texte peu rigoureux – plus « suggestif que mathématique », selon le probabiliste Joe Doob[3] –, et avaient du mal à saisir le contexte de l'ingénierie des communications.

3. Joseph L. Doob (1910-2004), mathématicien et probabiliste américain.

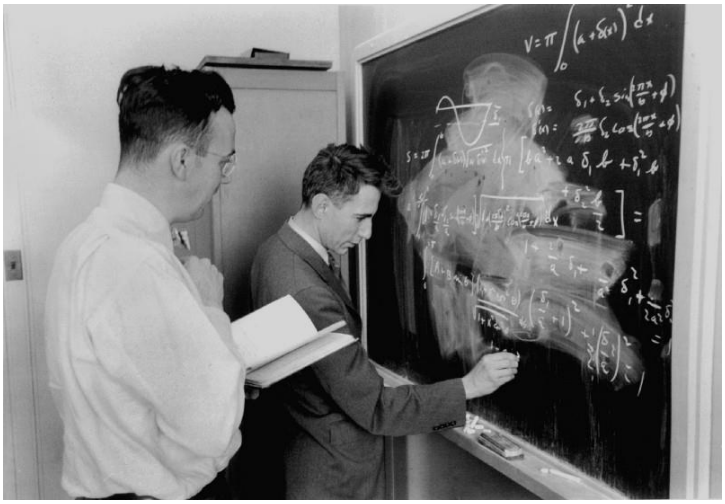

Figure 3 : Shannon (au tableau), mathématicien et ingénieur à la fois *(coll. Nokia Bell Labs).*

Sous l'impulsion de Warren Weaver[4], l'article de Shannon connaît une large diffusion. Il est publié sous forme de livre en 1949, précédé d'une introduction « grand public » de Weaver (une partie en a été publiée dans le *Scientific American*). Le titre initial devient « *La* théorie mathématique de la communication » (*The Mathematical Theory of Communication*). Dès 1950, Norbert Wiener, au MIT, reprend pour ainsi dire à son compte le travail de Shannon en l'intégrant dans sa *Cybernétique*, une théorie – voire un mouvement philosophique – qui allait rapidement connaître un grand succès. Le triomphalisme prolixe de Wiener contraste avec la discrétion de Shannon ; toujours est-il que la théorie de l'information devient un domaine très à la mode.

LA COMMUNICATION COMME FONDÉE SUR LA THÉORIE PROBABILISTE

La première figure (fig. 4 *infra*) de son article illustre le paradigme de la communication selon Shannon : c'est probablement ce qui a le plus percolé dans nombre de disciplines scientifiques. Comme écrit Shannon dès le deuxième paragraphe :

> *Le problème fondamental de la communication est de reproduire en un point, soit exactement, soit approximativement, un message recueilli en un autre point*[5].

4. Warren Weaver (1894-1978), scientifique américain.
5. Les citations de Shannon sont traduites de l'anglais selon la nouvelle édition du livre *La Théorie mathématique de la communication*, Cassini, 2018.

Dans cette figure, on voit qu'un message émis par une source d'information est transmis dans un canal bruité puis reçu par le destinataire. Si un tel schéma peut paraître naturel aujourd'hui, il était révolutionnaire à l'époque : pour la première fois, on y distingue clairement les rôles de la source, du canal et du destinataire ; de l'émetteur et du récepteur ; du signal et du bruit.

Au début des années 1950, le paradigme de Shannon est souvent appliqué aveuglément à des tas de domaines divers comme la biologie, la linguistique, la psychologie, les sciences sociales ; à tel point que Shannon, en 1956, dans un éditorial intitulé *The Bandwagon* (qu'on pourrait traduire par « Le train en marche »)[6], est obligé de mettre en garde contre les errements d'une telle popularité. Dans le même temps, en Union Soviétique, la théorie de Shannon est d'abord comprise comme un élément du champ de la cybernétique, considérée comme une « science des obscurantistes », et ses travaux sont à peine traduits ; mais Andrei Kolmogorov, qui avait donné sa forme définitive à la théorie des probabilités en 1933, comprend vite leur importance.

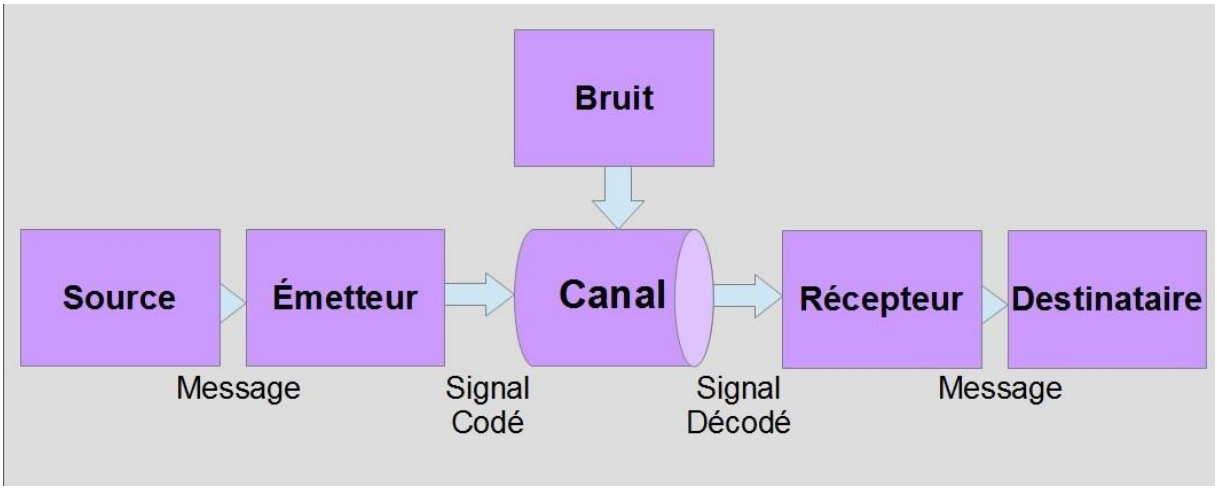

Figure 4 : Le paradigme de Shannon de 1948 *(Wikimedia Commons, auteur Seb02589)*

De fait, dès le début de son article, Shannon se fonde sur la théorie des probabilités : il commence par laisser de côté l'aspect sémantique – le « sens » de l'information – en considérant tout message comme résultant d'un choix dans un ensemble d'alternatives, selon un modèle probabiliste :

6. I*RE Transactions on Information Theory* (1956), volume 2, issue 1, p. 3. Shannon, Claude E. Shannon (mars 1956), "The Bandwagon", 2 (en ligne sur le site IEEE). Prendre « le train en marche » (*bandwagon*) est une image suggérant quelque peu ironiquement comment les auteurs les plus variés ont adopté dans leur domaine cette théorie de la communication.

Fréquemment, les messages ont une signification, c'est-à-dire qu'ils se réfèrent ou sont corrélés selon certaines modalités à des entités physiques ou conceptuelles. Ces aspects sémantiques de la communication sont sans rapport avec le problème technique. L'aspect essentiel est que le message effectif est choisi dans un ensemble de messages possibles. Le système doit fonctionner pour tous les choix que l'on peut faire et non seulement pour celui qui est effectivement fait, puisque celui-ci est inconnu au moment de la conception du système.

Kolmogorov se passionne immédiatement pour les résultats de Shannon et développe sa théorie de la complexité algorithmique, en allant en quelque sorte encore plus loin que Shannon. Il résumera sa pensée en 1970 en déclarant que « la théorie de l'information doit précéder la théorie des probabilités ».

@@@@@@@

L'article de Shannon est aussi celui où, pour la première fois, apparaît le terme *bit*, aujourd'hui omniprésent en informatique, comme mesure logarithmique de l'information. Dès le troisième paragraphe, on peut lire :

Comme l'a remarqué Hartley, le choix le plus naturel est celui de la fonction logarithme. Bien que cette définition doive être considérablement généralisée quand nous considérons l'influence des données statistiques du message et quand nous avons une gamme continue de messages, nous utiliserons dans tous les cas une mesure essentiellement logarithmique. La mesure logarithmique est la plus adéquate pour différentes raisons [...] *Si on utilise la base 2, l'unité peut être appelée « binary digit » ou, plus simplement, « bit », un terme suggéré par J. W. Tukey.*

Shannon crédite Tukey[7] mais c'est bien lui, Shannon, qui popularise ce terme avec l'idée révolutionnaire, devenue évidente aujourd'hui, que toute information peut être portée par des suites de *0* et de *1*. À noter que le *bit* comme unité proposée par Shannon va plus loin que le simple chiffre binaire, car il prend en compte l'aspect probabiliste de l'information. D'ailleurs, aujourd'hui, l'unité officielle de la mesure d'information s'appelle... le *Shannon* (sh).

L'unité *sh* ou Shannon

Donner son nom à une unité est un rare privilège pour un scientifique. Le shannon est une unité de mesure logarithmique de l'information, mesurant la quantité d'information contenue dans un message comme étant le nombre de bits minimal pour le transmettre.

7. John W. Tukey (1915-2000), statisticien américain.

Exemple du télégraphiste[8]

Un télégraphiste transmet les 26 caractères de l'alphabet, les 10 chiffres et des espaces. Un message télégraphique d'un caractère peut donc avoir 37 valeurs. Si on suppose que tous les caractères ont la même probabilité (comme dans un texte crypté), chacun porte un peu plus de 5 shannons (2^5 = 32 possibilités), plus précisément $\log_2(37) \approx 5{,}2$. Pour le transmettre, il faut 6 bits. La compression de données consiste à rapprocher le nombre de bits du nombre de shannons (ce dernier prenant en compte la probabilité d'une information).

L'influence des théorèmes de Shannon sur le développement des télécommunications, quoique considérable, n'a pas toujours été évaluée à sa juste mesure. Shannon énonce et résout le problème *théorique* de la communication. Il ne propose quasiment aucune solution pratique, mais établit des *limites* de performances, ce qui est au moins aussi important. Avant Shannon, des moyens de communication comme le télégraphe ont été développés pour ainsi dire dans le brouillard, sans le repère ultime permettant de savoir jusqu'où on pouvait aller. Avec Shannon, sont enfin connues les limites fondamentales de performances qu'il est possible d'atteindre, offrant ainsi aux ingénieurs la référence à laquelle tout système pratique peut se comparer. Sans les théorèmes de Shannon, tous les systèmes numériques modernes auraient mis beaucoup plus de temps à se développer. Il a fallu des décennies avant de voir apparaître quelques solutions pratiques qui s'approchent des limites de Shannon.

BREF EXPOSÉ DU CADRE MATHÉMATIQUE – L'ENTROPIE – PREMIER THÉORÈME DE SHANNON

Les résultats de Shannon sont nécessairement *asymptotiques* : la recherche des limites optimales de performances passe par une analyse où la dimension des signaux tend vers l'infini. Une source d'information, par exemple, est représentée comme un processus – une suite de variables aléatoires $X_1,X_2,X_3,\ldots$ représentant les symboles d'information émis au fur et à mesure du temps. Pour établir ses résultats, Shannon considère le vecteur $(X_1,X_2,\ldots,X_n)$ où la dimension n est

8. [NdÉ] Exemple tiré de la fiche Wikipédia consacré à l'unité du shannon.

arbitrairement grande. Cette vision asymptotique permet non seulement d'exploiter les dépendances statistiques existant entre les variables aléatoires, mais aussi d'obtenir un gain purement géométrique en grande dimension. Puisque n tend vers l'infini, les résultats asymptotiques vont provenir de la loi des grands nombres, de sorte que les limites de Shannon sont établies comme des quantités moyennes.

C'est ainsi que Shannon justifie l'utilisation de l'*entropie*, grâce au raisonnement suivant : considérons la source $\underline{X} = (X_1,X_2,...,X_n)$ pour une grande dimension n, où chaque symbole X peut prendre un nombre fini de valeurs. Supposons, pour simplifier, que cette source stationnaire est « sans mémoire », c'est-à-dire qu'à chaque instant, un symbole est tiré indépendamment des précédents. Les symboles $X_1,X_2,...,X_n$ sont alors indépendants et identiquement distribués, et en notant $p(x)$ la probabilité qu'un symbole égale x, la probabilité $p(\underline{x})$ d'un message donné $\underline{x} = (x_1,x_2,...,x_n)$ est le produit des probabilités individuelles : $p(\underline{x}) = p(x_1)\cdot p(x_2)\cdot\cdots\cdots p(x_n)$.

Regroupons les facteurs de ce produit suivant la valeur x prise par chaque argument :

$$p(\underline{x}) = \prod_x p(x)^{n(x)}$$

où $n(x)$ est le nombre de symboles composantes du vecteur $(x_1, x_2, \ldots, x_n)$ qui égalent x. Le rapport $n(x)/n$ est la fréquence empirique de x qui, par la loi des grands nombres, tend vers $p(x)$ lorsque n tend vers l'infini. Un vecteur $\underline{x}$ « typique » tiré au hasard vérifie alors approximativement

$$p(\underline{x}) \approx \prod_x p(x)^{np(x)} = 2^{-nH}$$

où

$$H = \sum_x p(x) log_2 \frac{1}{p(x)}$$

est une quantité positive que Shannon appelle *entropie* par analogie avec l'entropie étudiée par Ludwig Boltzmann en mécanique statistique :

> *On reconnaîtra dans la forme de* H *celle de l'entropie définie dans certaines formulations de la mécanique statistique, où* p_i *est la probabilité d'un système d'être dans la cellule* i *de son espace de phase.* H *est alors, par*

exemple, le H *du fameux théorème de Boltzmann. Nous appellerons* $H = -\Sigma\, p_i \log p_i$ *l'entropie de l'ensemble de probabilités* $p_1, \ldots, p_n$.

C'est semble-t-il John von Neumann qui recommande à Shannon d'utiliser le terme *entropie* car, lui dit-il, « personne ne sait vraiment ce qu'est l'entropie, de sorte qu'en cas de débat vous aurez toujours l'avantage[9] ». Par la suite, les relations qu'entretient l'entropie de Shannon avec la physique statistique et la thermodynamique ont donné lieu à d'innombrables débats et commentaires, notamment sous l'impulsion du physicien Léon Brillouin[10].

@@@@@@@

La notion d'entropie et la relation $p(\underline{x}) \approx 2^{-nH}$ pour une suite $\underline{x} = (x_1, x_2, \ldots, x_n)$ « typique » permettent à Shannon de résoudre le problème théorique de la *compression* d'une source : il s'agit d'un cas particulier du paradigme de la figure 4 où le canal est sans bruit, et où l'on désire *coder* la source en la communiquant au destinataire d'une manière arbitrairement fiable, avec un débit d'information minimal de sorte à comprimer au maximum la source. Pour cela, il suffit de ne coder que les suites $\underline{x} = (x_1, x_2, \ldots, x_n)$ typiques, car la loi des grands nombres que l'on vient d'utiliser implique qu'on ne peut tomber sur une suite non typique qu'avec une probabilité arbitrairement faible. En sommant la relation $p(\underline{x}) \approx 2^{-nH}$ sur toutes les suites typiques, on obtient la probabilité totale qu'une séquence tirée au hasard soit typique, qui est très proche de 1 :

$$1 \approx N \cdot 2^{-nH}$$

où N est le nombre total de suites typiques. On a donc $N \approx 2^{nH}$, soit $\log_2 N \approx nH$ bits d'information, d'où un débit de $(\log_2 N)/n \approx H$ bits par symbole de source.

Ce raisonnement peut facilement être rendu rigoureux et conduit au *premier théorème de Shannon* (pour le codage de source) qui affirme que H bits par symbole suffisent pour comprimer fidèlement une source d'information. L'entropie apparaît donc être une borne inférieure sur le débit, nécessaire pour

9. L'attribution précise de cette citation reste incertaine (voir une discussion à ce propos ici). Elle provient probablement d'un témoignage de Shannon lui-même en 1961, dont parle un article de Tribus, Myron and McIrving, Edward C. (1971), "Energy and Information", *Scientific American*, 225: 179-88. Cependant Shannon dans une interview (Price, Robert, "A Conversation with Claude E. Shannon", *IEEE Communications Magazine*, 22(5): 123-26, 1984) nie s'en souvenir (mais il était probablement déjà atteint de la maladie d'Alzheimer).
10. [NdÉ] Voir l'article *BibNum* d'analyse par Ph. Jacquet (septembre 2009) de la préface et de l'introduction à *La Science et la théorie de l'information*, Masson, 1959, par Léon Brillouin.

coder l'information de façon fiable.

Ce théorème est asymptotique (*n* tend vers l'infini) et ne donne aucun moyen de coder en pratique pour s'approcher de l'entropie. Mais dans son article, Shannon – et, indépendamment, Robert Fano[11] – ont l'idée de considérer un code à longueur variable où les symboles les plus probables sont codés par les codes les plus courts. Il est possible d'attribuer un nombre de bits légèrement supérieur à $\log[1/p(x)]$ à chaque symbole *x*, de sorte que le débit moyen devient assez proche de *H*. David Huffman[12] décrira en 1952 l'algorithme optimal de compression dans ce contexte.

@@@@@@@

Une notion très proche de l'entropie de Shannon, développée au même moment et indépendamment en statistique mathématique par Harold Jeffreys, Solomon Kullback et Richard Leibler[13] est la *divergence* ou l'*entropie relative* entre deux distributions de probabilités *p* et *q*, définie par :

$$D = \sum_x p(x) log_2 \frac{p(x)}{q(x)}$$

Cette quantité est positive, comme on le démontre en utilisant la propriété de concavité du logarithme, et ne s'annule que lorsque les deux distributions *p* et *q* coïncident. Elle intervient naturellement dans le raisonnement précédent de Shannon : en notant $p'(x) = n(x)/n$ la fréquence empirique d'un symbole *x*, il est facile de vérifier qu'on a l'égalité rigoureuse :

$$p(\underline{x}) = 2^{-nH(p',p)}$$

où $H(p', p)$ est l'entropie croisée :

$$H(p',p) = \sum_x p'(x) log_2 \frac{1}{p(x)}$$

Une suite « typique » $\underline{x}$ étant caractérisée par sa fréquence empirique p', la probabilité de tomber sur une suite typique est égale à

$$N \cdot 2^{-nH(p',p)}$$

11. Robert Fano (1917-2016), informaticien américain d'origine italienne.
12. David A. Huffman (1925-1999), informaticien américain.
13. Sir Harold Jeffreys (1891-1989) est un informaticien et géophysicien britannique ; Solomon Kullback (1907-1994) est un informaticien et cryptologue américain ; Richard Leibler (1914-2003) est un mathématicien et cryptologue américain.

Si on remplace *p* par *p'*, on obtient une autre probabilité est qui est inférieure à 1, d'où

$$N \leq 2^{nH(p',p')}$$

Ainsi la probabilité de tomber sur une suite typique ne dépasse pas

$$2^{n(H(p',p')-H(p',p))} = 2^{-nD(p',p)}$$

Ce type de raisonnement est généralisable à d'autres ensembles que les ensembles de suites typiques, et le comportement exponentiel en $2^{-nD(p',p)}$ est très utile en théorie des grandes déviations, ainsi que pour expliquer des comportements asymptotiques de tests d'hypothèses. Il conduit à une notion d'information de Chernoff (due à Herman Chernoff[14]) pour classifier les données empiriques. Par dérivation à partir de l'entropie relative, on obtient également l'information de Fisher (due à Ronald Fisher[15]), utile en estimation de paramètres dans des distributions. La théorie de l'information trouve ainsi bien d'autres applications que celles liées au problème de la communication, en statistique notamment.

@@@@@@@

Dans son article, Shannon décrit l'entropie $H(X) = \sum_x p(x) log_2 \frac{1}{p(x)}$ de la distribution de probabilité *p* d'une variable aléatoire *X* comme une mesure d'incertitude sur *X*. Il fonde également sa théorie sur une autre quantité

$$I(X;Y) = \sum_{x,y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)}$$

que Fano nomme *information mutuelle* entre deux variables *X* et *Y*. C'est exactement l'entropie relative *D (p, q)* entre la distribution *p (x, y)* conjointe de (*X*, *Y*) et la distribution *q (x, y) = p (x)p(y)* qui correspond au cas où *X* et *Y* sont indépendantes. Elle est donc positive et ne s'annule que dans le cas de l'indépendance. Elle est également symétrique : *I (X; Y) = I (Y; X)* et mesure la quantité moyenne d'information entre *X* et *Y* (qui, naturellement, s'annule dans le cas où *X* et *Y* sont indépendantes). Shannon écrit *I (X; Y)* sous la forme

$$I(X ; Y) = H(X) - H(X|Y)$$

14. Hermann Chernoff, mathématicien et informaticien américain, né en 1923.
15. Sir Ronald Aylmer Fisher (1890-1962), FRS, biologiste et statisticien britannique.

où

$$H(X|\,Y) = \sum_{x,y} p(x,y) log_2 \frac{1}{p(x|y)}$$

est l'entropie de la distribution conditionnelle *p (x |y)*, qui mesure l'incertitude sur *X* connaissant *Y* . Cette dernière est inférieure à *H (X)* puisque *I (X; Y) ≥ 0* : ainsi la connaissance (de *Y*) réduit l'incertitude (sur *X*), d'une quantité précisément égale à *l'information* I (X; Y) *qu'apporte* Y *sur* X*.*

Ce type de raisonnement intuitif a excité l'imagination de nombreux scientifiques. C'est la première fois que le concept – jusque-là flou – d'information transmise dans un système trouve une théorie rigoureuse. Le diagramme ensembliste ci-dessous résume les relations entre les différentes quantités utiles en théorie de l'information.

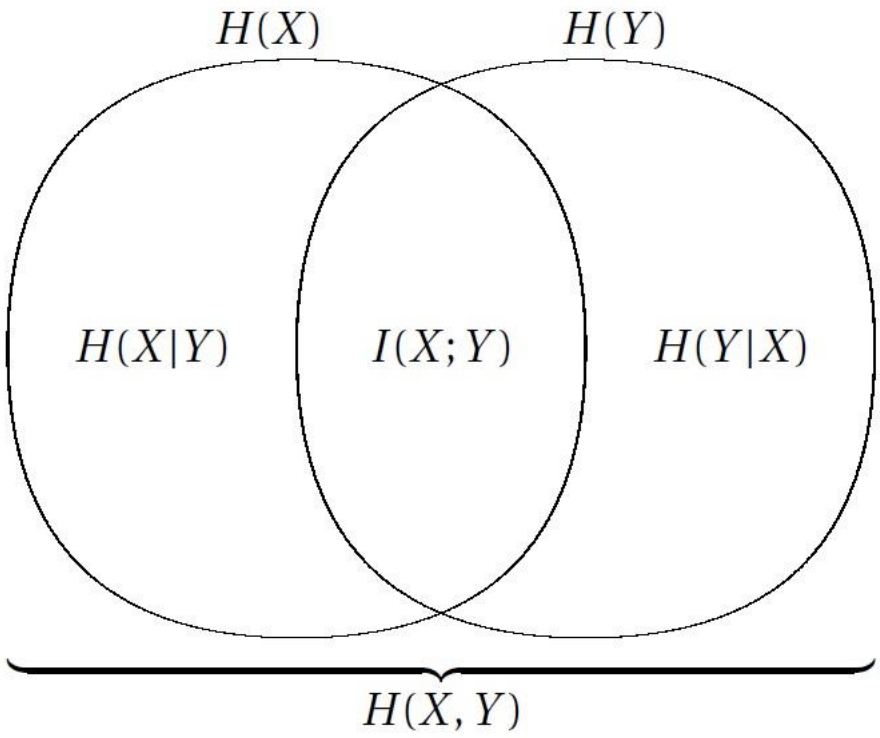

Figure 5 : Un diagramme ensembliste expliquant les relations entre les différentes quantités définies en théorie de l'information.

CODAGE EN CANAL BRUITÉ – DEUXIÈME THÉORÈME DE SHANNON

Le résultat le plus difficile et probablement le plus important de l'article de Shannon se fonde sur une méthode appelée « codage aléatoire ». Considérons, toujours pour une grande dimension n, un canal bruité d'entrée $\underline{x} = (x_1,x_2,...,x_n)$ et de sortie $\underline{y} = (y_1,y_2,...,y_n)$. Ce canal est décrit par les probabilités conditionnelles $p(\underline{y}|\underline{x})$ de la sortie connaissant l'entrée. Supposons, pour simplifier, que ce canal stationnaire est « sans mémoire », ce qu'on traduit par un développement en produit de probabilités conditionnelles individuelles :

$$p(\underline{y}|\underline{x}) = p(y_1|x_1)\cdot p(y_2|x_2)\cdots\cdots\cdot p(y_n|x_n).$$

L'entrée du canal est le « code » portant l'information à transmettre. Shannon a l'idée géniale de considérer simultanément l'ensemble de tous les codes possibles que l'on peut utiliser dans la communication, *comme si* chaque code $\underline{x}$ était choisi au hasard selon une distribution $p(\underline{x}) = p(x_1)\cdot p(x_2)\cdots\cdots\cdot p(x_n)$:

> *La méthode pour démontrer [...] ce théorème ne consiste pas à établir une méthode de codage possédant les propriétés désirées, mais à montrer l'existence d'un tel code dans un certain ensemble de codes. En fait, nous ferons la moyenne des taux d'erreur sur cet ensemble et montrerons que cette moyenne peut être rendue inférieure à ε. Si la moyenne d'un ensemble de nombres est inférieure à ε, il doit en exister au moins un dans l'ensemble qui est inférieur à ε. Cela établira le résultat désiré.*

Cela lui permet d'établir, par le même calcul déjà fait pour $\underline{x}$ tout seul, que le code émis $\underline{x}$ est conjointement « typique » avec $\underline{y}$ au sens où il vérifie

$$p\left(\underline{x},\underline{y}\right) \approx 2^{-nH(x,y)}$$

Il est néanmoins possible qu'un *autre* code que celui réellement émis vérifie aussi cette condition. Comme les codes sont choisis indépendamment au hasard, la distribution conjointe pour cet autre code n'est plus $p(\underline{x}, \underline{y})$, mais $q(\underline{x}, \underline{y}) = p(\underline{x})p(\underline{y})$. D'après un calcul similaire à celui déjà fait pour $\underline{x}$, la probabilité que cela arrive ne dépasse pas

$$2^{-nD(p,q)} = 2^{-nI(X;Y)}$$

C'est cette borne $2^{-nI(X;Y)}$ ci-dessus permet à Shannon de résoudre le problème théorique de la *transmission* dans un canal bruité (voir le paradigme de la figure 4). Il s'agit cette fois de maximiser le débit d'information transmis tout en garantissant une communication arbitrairement fiable du message au destinataire.

Pour cela, il suffit de décoder l'information de sorte à récupérer un code $\underline{x}$ conjointement typique de y en sortie du canal, car la probabilité de tomber sur un ($\underline{x}$, $\underline{y}$) non typique est arbitrairement faible. La probabilité d'erreur de décodage est donc essentiellement due à la présence éventuelle d'un *autre* code conjointement typique de $\underline{y}$ qui donne lieu à une ambiguïté de décodage. D'après ce qu'on vient de voir, la probabilité totale d'erreur due à une telle ambiguïté est bornée par

$$N.2^{-nI(X;Y)}$$

où N est le nombre total de codes utilisés dans la transmission. Pour que cette expression tende vers 0 quand n tend vers l'infini, il suffit que le débit $(\log_2 N)/n$ soit inférieur à $I(X;Y)$ bits par symbole. Pour maximiser ce débit, Shannon choisit la distribution de probabilité des codes de sorte que $I(X;Y)$ soit maximal et nomme

$$C = \max_{p(x)} I(X;Y)$$

la *capacité du canal*. Comme la probabilité d'erreur a été calculée en moyenne sur tous les codes possibles, il existe nécessairement au moins une solution pour laquelle il n'y a pas plus d'erreurs (cf. la citation ci-dessus). On obtient ainsi le *deuxième théorème de Shannon* (pour le codage de canal) qui affirme qu'on peut transmettre l'information de façon fiable tant que le débit ne dépasse pas la capacité C du canal.

Ce théorème est une véritable révolution qui a changé le monde ! Pour la première fois, on comprend que le bruit présent dans le canal ne limite pas la qualité de la communication : il ne limite que le débit de transmission. À la condition de ne pas dépasser la capacité, la communication numérique peut être quasi-parfaite ! Ce théorème à lui seul justifie l'explosion du numérique aujourd'hui.

Théorème d'échantillonnage de Shannon (et, avant lui, Nyquist et Whittaker)

Enfin, cet article est resté célèbre pour une formule établie dans son fameux théorème 17 qui donne l'expression suivante de la capacité de canal :

$$C = W.log_2\left(1 + \frac{P}{N}\right) \text{ bits/s,}$$

où *W* est la largeur de bande et *P*/*N* le rapport signal à bruit présent dans la transmission. Pour l'établir, Shannon considère que le bruit présent dans un canal de transmission est modélisé par du bruit blanc gaussien qui s'ajoute au signal à la réception. C'est certainement la formule la plus connue de Shannon, celle qui conclut son œuvre. Il popularise à cette occasion le théorème d'échantillonnage démontré auparavant par Edmund Whittaker[16] et Harry Nyquist, qui est souvent (à tort) également appelé « théorème de Shannon ». Cette formule fournit un aspect concret de la théorie de l'information qui a séduit de nombreux ingénieurs dès sa parution. Elle est arrivée juste au bon moment : pas moins de 7 autres chercheurs ont publié une formule similaire la même année 1948 ! Un certain nombre d'entre eux sont d'ailleurs connus de Shannon :

> *Des formules semblables à C =W log (P +N)/N pour le cas du bruit blanc ont été développées de façon indépendante par plusieurs auteurs, quoique avec des interprétations quelque peu différentes. Nous pouvons mentionner les travaux de N. Wiener, W. G. Tuller et H. Sullivan à ce propos.*

L'héritage de Shannon en a dérouté plus d'un : ses théorèmes prévoient qu'il existe de bons systèmes de codage pratiques, mais ne disent pas comment les construire. Paradoxalement, l'idée du codage aléatoire suggère que des codes choisis au hasard forment des solutions quasi-optimales. Sauf qu'avec une dimension tendant vers l'infini, une telle méthode basée sur le hasard est irréaliste en pratique. Il a fallu 50 ans pour que les ingénieurs français Claude Berrou et Alain Glavieux[17] proposent des solutions pratiques (les turbo-codes) qui « imitent » le codage aléatoire et permettent ainsi de s'approcher de très près de la capacité.

@@@@@@@

La théorie de la communication selon le terme de Shannon, que tout le monde appelle aujourd'hui théorie de l'information, est toujours aussi vivante en tant que discipline mathématique appliquée aux communications et à d'autres sciences. Dans cet article fondateur, Shannon a privilégié la compréhension intuitive de ses résultats, ce qui permet une lecture aisée et explique sans doute son caractère intemporel. Aujourd'hui, les traités sur la théorie de l'information sont empreints d'une bien plus grande rigueur formelle, mais en regardant de près le texte de Shannon, tous les résultats importants, y compris quelques-uns des plus avancés y sont déjà présents. Il est impressionnant de constater à quel point

16. Sir Edmund T. Whittaker (1873-1956), mathématicien et astronome britannique ; H. Nyquist, déjà mentionné.
17. Claude Berroux (né en 1951) et Alain Glavieux (1949-2004), informaticiens et mathématiciens français.

cette théorie est née entièrement formée – Shannon ne s'étant d'ailleurs pas contenté de fonder la théorie en un seul article, il y apporta aussi les contributions les plus importantes dans les années 1950 et 1960.

■

(janvier 2018)

Figure 6 : Shannon devant sa machine à jouer aux échecs *(coll. famille Shannon). Shannon se consacra toute sa vie à d'autres passions liées au jeu (échecs notamment) et à l'intelligence artificielle. Il publia en 1949 un des premiers articles sur la programmation du jeu d'échecs.*