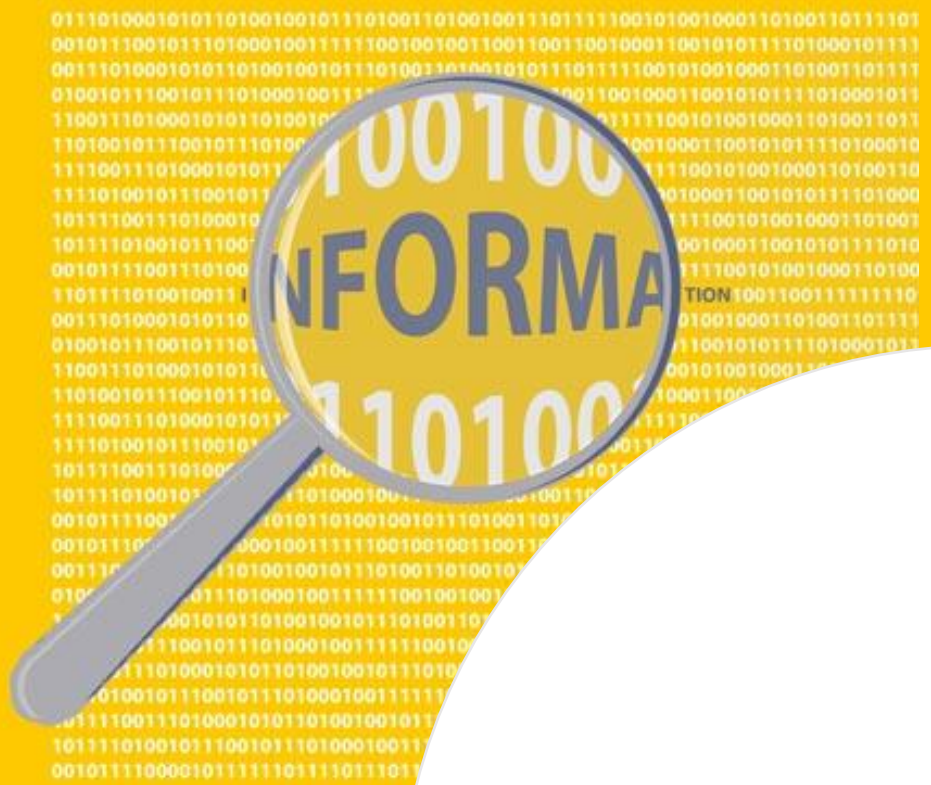
Bee Bryte
Energy Intelligence

7 min



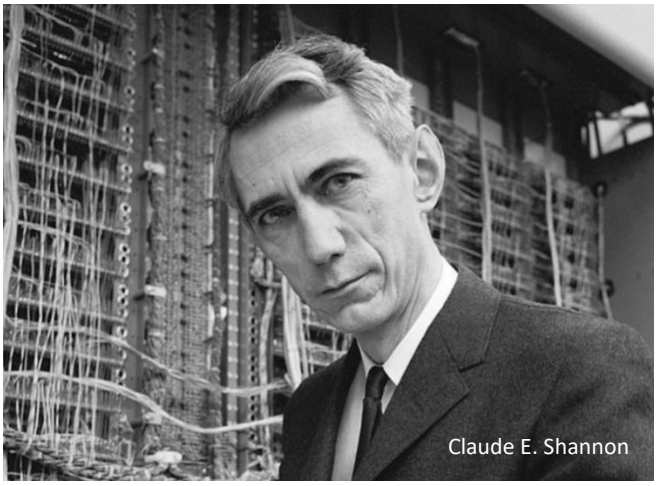
La Substantifique Moelle

Fabien BRULPORT, Romain BOIDEVEZY, Johanna DREANO

20 Juin 2018

Donnée vs Information

Les termes “**Donnée**” et “**Information**” prêtent souvent à confusion. Bien que liés, ceux-ci couvrent des notions différentes. Le but de cet article est d’apporter des clarifications quant à ces différences et d’explorer le sens profond de l’information.


Claude E. Shannon

« Donnée » vient du latin « datum » (« quelque chose qui est donné »), et correspond à la matière brute dont dérive l’information. La donnée n’a pas de valeur intrinsèque, et provient de la mesure d’un aspect observable d’un phénomène.

Il existe donc une relation hiérarchique entre les deux termes : l’information n’existe pas sans donnée, mais l’inverse n’est pas vrai. Cela ne permet pas pour autant de répondre à la question : “qu’est-ce que l’information ?”

Claude Shannon, père de la théorie de l'information, définit une information comme ce qui permet de lever une incertitude. Les données contiennent donc (en général) de l’information.

Plaçons-nous dans le bon contexte : les données auxquelles nous nous intéressons sont numériques. Comme vous le savez probablement déjà, un processeur d’ordinateur est constitué d’un amas d’interrupteurs (transistors) qui permettent de réaliser des opérations logiques. Les interrupteurs ont deux états : ouvert ou fermé. C’est pourquoi, dans le but d’effectuer des opérations, les données sont codées sous forme de séquences de 0 et de 1 dans la mémoire de notre ordinateur. On parle de “base 2” ou “binaire” : chaque 0 ou 1 est appelé bit pour binary digit. L'agrégation de ces bits permet de représenter n’importe quoi : une position, une image, un ensemble de caractères, etc. Tout message peut être écrit en binaire, il suffit que l’émetteur du message et le récepteur utilisent le même codage pour convertir sans perte l’information et pouvoir se comprendre.

La quantité d'information :

Le **bit** est une unité élémentaire d'information. Pour bien comprendre ce concept il est nécessaire d’introduire la notion d’entropie de Shannon.

Considérons une pièce dans laquelle la température peut prendre deux états équiprobables : froid ou chaud. Si le thermomètre nous envoie la température actuelle, il permet de lever notre incertitude. Il se peut qu’il envoie la température sous forme de chaîne de caractères : « froid » ou « chaud ». Chaque caractère est codé sur 8 bits, soit 40 bits par état (5 caractères). Cependant la température pouvant prendre uniquement deux états, on aurait pu aussi bien envoyer « 0 » pour froid et « 1 » pour chaud, on aurait alors réduit considérablement la quantité de données transmise, tout en véhiculant la même information.

L’entropie nous permet de quantifier objectivement à quel point un système de données concentre de l’information. Elle est définie comme une mesure de la quantité relative moyenne d’information contenue dans un échantillon de données. La quantité d’information selon Shannon est le nombre de bits **nécessaires et suffisants** pour représenter le contenu sémantique d’un message.

Dans notre exemple ci-dessus, la température pouvait prendre deux états d’une probabilité de $P=0.5$ chacun.

La quantité d'information nécessaire pour coder un état est 1 bit :

$$Q = -log_2(P) = -log_2(1/2) = 1$$

Nous utilisons le logarithme de base 2 (ou logarithme binaire) car la quantité d'information et l'entropie sont exprimées en bits. La quantité moyenne d'information produite par le thermomètre à chaque mesure, c'est à dire l'entropie, vaut :

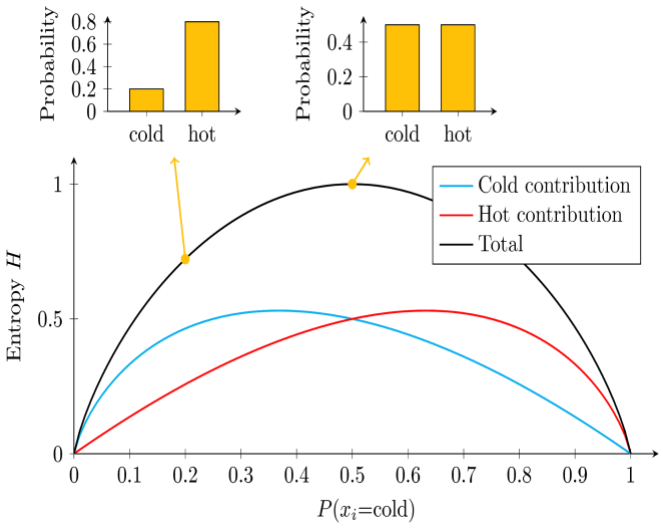
$$H = -1/2 log_2(1/2) - 1/2 log_2(1/2) = 1$$

Dans un cas plus réaliste où la température est discrétisée en *n* états, chaque état (température comprise entre T_i et T_{i+1}) correspond à un événement x_i possédant une probabilité $P(x_i)$ de se produire. Dans ce cas, l'entropie de Shannon se calcule de la manière suivante :

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_b(P(x_i))$$

Du fait du plus grand nombre d'états, la quantité d'information moyenne fournie par le thermomètre sera plus grande que dans l'exemple précédent. Son contenu brut sera plus complexe et paraîtra plus "désordonnée" pour un observateur ce qui conduit à une entropie de Shannon plus grande. On note ici la similitude avec l'entropie physique comme mesure du désordre d'un système thermodynamique.

Si les états sont tous équiprobables, l'information de température est riche et répartie entre tous ses états possibles. L'entropie dans ce cas est donc plus grande que celle d'une situation où les états ont des probabilités différentes : l'information est plus concentrée dans les températures les moins probables (leur caractère « exceptionnel » apporte beaucoup d'information) que dans les températures les plus susceptibles de se produire (qui sont donc plus « banales », et véhiculent donc moins d'information), l'entropie est donc inférieure. Pour illustrer cette affirmation, on peut tracer l'évolution de l'entropie *H(x)* dans le cas où la température peut prendre les deux états « froid » ou « chaud » en fonction de leurs probabilités respectives :



L'entropie est nulle lorsqu'une des probabilités vaut 1 (car le thermomètre n'apporte dans ce cas aucune information. Par exemple : il fait toujours chaud) et maximale quand les deux probabilités sont équivalentes, c'est-à-dire valent 0.5. Entre ces extrêmes, l'entropie décrit une courbe convexe telle que représentée dans la figure.

Le fait que l'entropie puisse être considérée comme une mesure de l'incertitude moyenne du message (rappelons que le thermomètre permet de lever une incertitude) permet d'expliquer pourquoi *H(x)* décroît quand une des probabilités tend vers 1. En effet dans ce cas, l'effet de surprise du message diminue. Les deux termes qui constituent l'entropie dans le cas binaire sont aussi tracés : plus un événement possède une probabilité faible et plus son impact sur l'entropie globale est grand.
Il est important de remarquer que la quantité d'information dépend de son **contexte** (ici défini par le nombre d'événements et leurs probabilités d'occurrence).

En pratique, l'entropie de Shannon est utilisée pour numériser une source d'information avec le moins de

bits possibles, sans perte d'information mais aussi pour contrôler la quantité d'information dans un jeu de données duquel on veut extraire de la valeur.

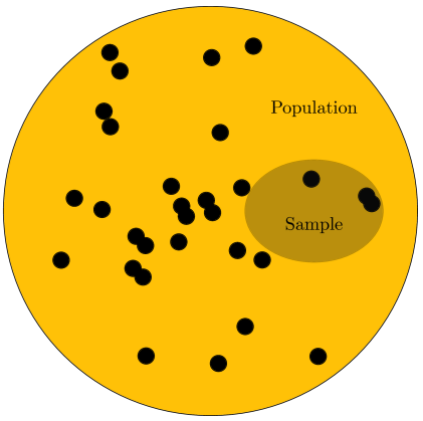
Que faire de cette information ? L'extraction de valeur

La Data-Science consiste à extraire et valoriser l'information contenue dans la donnée au moyen d'algorithmes de traitement. Avant toute exploitation de données, il convient de vérifier leur qualité. Si celle-ci est insuffisante, leur utilisation : au mieux sera impossible, au pire pourra conduire à des résultats et potentiellement des prises de décision erronées, selon le principe implacable du "garbage in, garbage out" (des données de mauvaise qualité en entrée d'un algorithme aussi performant soit-il, conduiront inévitablement à des résultats de mauvaises qualité). Toute personne ayant travaillé sur des données réelles sait que les étapes de prétraitement des données sont primordiales pour obtenir de résultats satisfaisants.

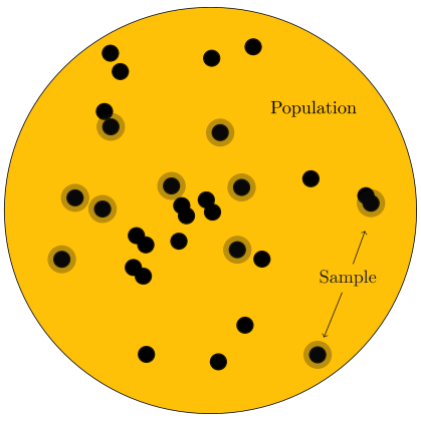
La qualité d'un échantillon de données s'entend sur le fond et sur la forme. Si la qualité "de forme" est assez évidente (bruit, données manquantes, biais, etc.), la qualité "de fond" est plus subtile et difficile à estimer a priori.

Typiquement, un piège à éviter est de considérer qu'augmenter la quantité de données en entrée d'un algorithme va forcément améliorer les performances de ce dernier. En effet, une grande quantité de données n'est pas forcément synonyme de qualité. Une façon de s'en convaincre est de considérer un message, et un échantillon fabriqué à partir de la répétition N fois de ce message. L'information des 2 échantillons sera strictement identique (celle contenue dans le message de base), alors que leurs tailles seront très différentes. La qualité d'un échantillon de données est donc intimement liée à la quantité d'information qu'il contient, indépendamment de sa taille.

Un piège similaire est de posséder uniquement un jeu de donnée partiel et non homogène. Prenons l'exemple où l'on cherche à faire des prédictions journalières de consommation d'électricité en se basant sur plusieurs paramètres. La température extérieure par exemple influe logiquement sur la consommation si le bâtiment possède un système électrique de régulation de température. Les estimations de consommation pour la période hivernale seront assez logiquement erronées si elles se basent sur des données de température collectées durant l'été. De manière graphique, on peut représenter les données comme extraites d'une population (points noirs dans les schémas suivants). Un échantillonnage est pertinent s'il permet de représenter la variance de la population, c'est à dire si ces données s'étendent sur l'ensemble du domaine de variation des paramètres influents. La difficulté de tester si l'échantillon respecte cette condition réside dans l'estimation de la variance de la population globale.



"Mauvais" échantillonnage



Échantillonnage pertinent

Nous espérons que cet article vous aura permis de mieux comprendre le distinguo entre information et données. Si l'entropie nous apporte une mesure objective de la quantité d'information moyenne qu'un système de données possède, en opposition, il n'y a cependant pas de manière aussi directe pour mesurer la qualité de l'information d'une source de données. Il est notamment nécessaire de vérifier leur « propreté » (absence de bruit, trous, etc.), leur répartition et leur cohérence par rapport à l'information qu'on souhaite en extraire.

Les modèles d'apprentissage automatique deviennent de plus en plus complexes et performants, mais ils ne font pas pour autant des miracles. Si on veut révéler leur plein potentiel il est important de les alimenter avec des données riches en information. La quantité de données nécessaire est alors une condition nécessaire… mais pas suffisante.

Pour les diverses études qui lui sont confiées, BeeBryte a développé une large panoplie d'outils numériques propriétaires de prétraitement qui permettent de tirer le meilleur parti de l'information contenue dans les données disponibles, et d'optimiser les performances des algorithmes d'intelligence artificielle qu'elles alimentent ensuite. Ces outils sont complétés par l'expérience et le savoir-faire de nos data-scientistes qui confirment chaque jour que la phase de pré-analyse et préparation des données, trop souvent négligée par certains, est pourtant primordiale et encore difficilement automatisable quand les sources de données sont multiples et hétérogènes.

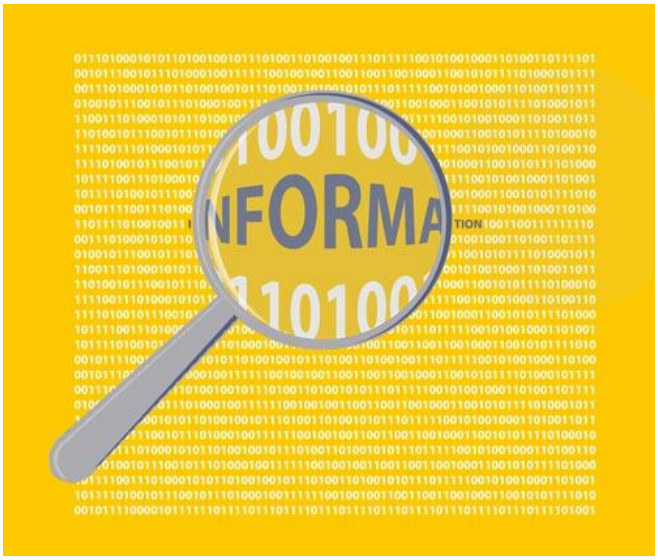
N'hésitez-pas à faire appel à notre expertise !

SOURCES :

https://centenaire-shannon.cnrs.fr/chapter/la-theorie-de-information

https://www.youtube.com/watch?v=ErfnhcEV1O8

https://en.wikipedia.org/wiki/Quantities_of_information



BeeBryte développe des solutions autour de l'Intelligence Artificielle pour que les bâtiments industriels et commerciaux, les stations de recharge de véhicules électriques et les éco-quartiers consomment l'énergie de manière plus intelligente, moins chère et plus efficacement tout en réduisant leur empreinte carbone ! BeeBryte a une équipe de 20 personnes en France et à Singapour et est soutenue par BPI-i Lab & l'ADEME. Depuis sa création en 2015, ses solutions ont reçu de nombreux prix, tels qu'EDF Pulse, Start-up Energy Transition Award, et le label GreenTech Verte.

contact@beebryte.com
www.twitter.com/BeeBryteGroup
www.beebryte.com