BeeBryte
Energy Intelligence

7 min

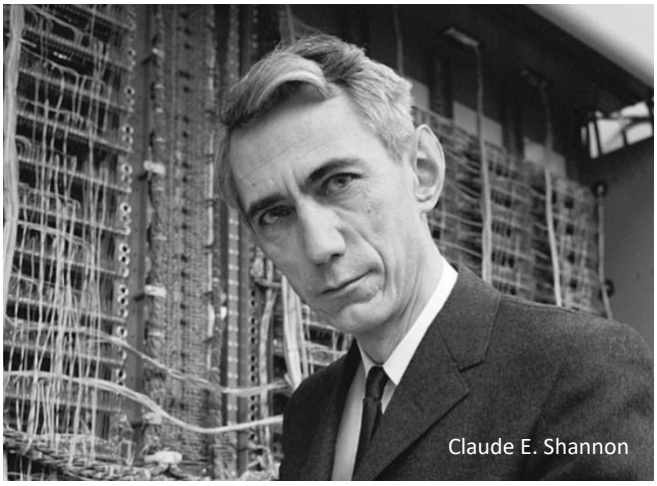
The Pith and Marrow

Fabien BRULPORT, Romain BOIDEVEZY, Johanna DREANO

June 20th 2018

Data vs Information

The terms "Data" and "Information" are often confusing. Although linked, they express different notions. The purpose of this article is to shade light on the difference and explore the deeper meaning of information.


Claude E. Shannon

"Data" comes from the Latin word "datum" ("something that is given") and corresponds to the raw content from which the information is extracted. Data has no intrinsic value and comes from the measurement of an observable aspect of a phenomenon.

There is therefore a hierarchical relationship between the two terms: information does not exist without data, but the contrary is not true. But this still doesn't answer the question: "What is information?"

Claude Shannon, father of the information theory, defines information as what can remove uncertainty. Data therefore contains (generally) information.

Let's put ourselves in the right context: the data, we are interested in, is digital. As you probably already know, a computer processor unit consists of a cluster of power switches (transistors) that can perform logical operations. The switches have two states: open or closed. Therefore, to be able to perform operations, the data is encoded as a sequence of 0 and 1 in the memory of the computer. We refer to a "base 2" or "binary": each 0 or 1 is called a bit for binary digit. The aggregation of these bits enable to represent anything: a position, an image, a set of characters, etc. Any message can be written in binary, it just requires that both the transmitter and the receiver of the message use the same encoding so that they can translate the information without any loss and understand each other.

Information quantity:

The **bit** is an elementary unit of information. To fully understand this concept, we shall introduce Shannon's notion of entropy.

Let's consider a room in which the temperature can take two equiprobable states: cold or warm. If the thermometer sends us the current temperature, it allows us to remove our uncertainty. It may send the temperature as a character string: "cold" or "warm". Each character could be encoded on 8 bits, i.e. 32 bits per state (4 characters). However, the temperature can take only two states, we could as well send "0" for cold and "1" for warm, it would have reduced significantly the amount of data transmitted, while conveying the same information.

Entropy allows us to objectively quantify how much a data system concentrates information. It is defined as a measure of the average relative amount of information contained in a data sample. The amount of information according to Shannon is the number of **necessary and sufficient** bits to represent the semantic content of a message.

In our previous example, the temperature can take two states with a probability of P = 0.5 each. The amount of information needed to code a state is 1 bit:

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

We use the base-2 logarithm (or logarithmic binary) because the quantity of information and the entropy are measured in bits. The average amount of information produced by the thermometer for each measurement, i.e. the entropy, is:

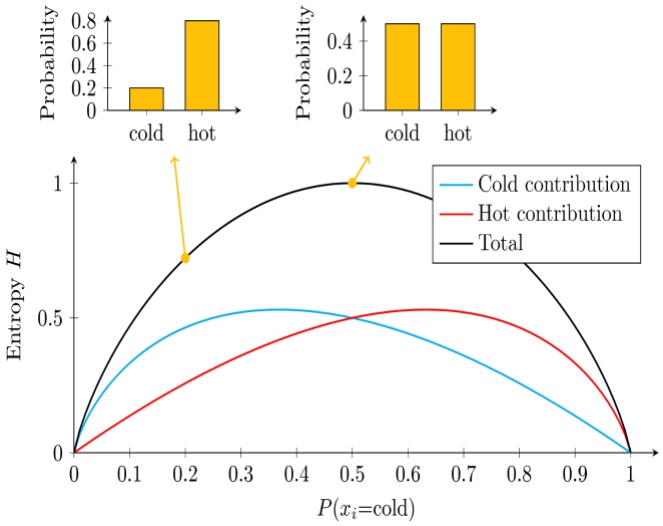
$$H = -1/2log_2(1/2) - 1/2log_2(1/2) = 1$$

In a more realistic case where the temperature is discretized into n states, each state (temperature between T_i and T_{i+1}) corresponds to an event xi which has a probability $P(x_i)$ to occur. In this case, Shannon's entropy is calculated as follows:

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

Since the number of states is bigger, the average amount of information provided by the thermometer will be larger than in the previous example. Its raw content will be more complex and more "disordered" for an observer which leads to a larger Shannon entropy. We shall notice here the similarity with the physical entropy as a measure of the disorder of a thermodynamic system.

If the states are all equiprobable, the temperature information is important and distributed among all its possible states. The entropy in this case is therefore greater than in a situation where the states have different probabilities: the information is then more concentrated in the less probable temperatures (their "exceptional" occurrence brings a lot of information) than in the temperatures most likely to occur (which are therefore more "ordinary", and thus convey less information), the entropy is therefore lower. To illustrate this statement, we can trace the evolution of the entropy *H(x)* in the case where the temperature can take the two states “cold” or “hot” according to their respective probabilities:



The entropy is zero when one of the probabilities is equal to 1 (because the thermometer does not provide any information in this case, for example: it is always warm) and maximum when the two probabilities are equivalent, that is to say, equal to 0.5. Between these extremes, entropy describes a convex curve as shown in the figure.

The fact that entropy can be considered as a measure of the average uncertainty of the message (recall that the thermometer allows to lift an uncertainty) makes it possible to explain why *H(x)* decreases when one of the probabilities tends towards 1. Indeed, in this case the surprise effect of the message decreases. The two terms playing a role in the entropy value in the binary case are also plotted: the lower the probability of an event, the greater its impact is on the global entropy.
It is important to note that the amount of information depends on its context (here defined by the number of events and their probabilities of occurrence).

In practice, Shannon's entropy is used to digitize an information source with the least number of bits, without loss of information but also to control the amount of information in a dataset from which one wants to extract value.

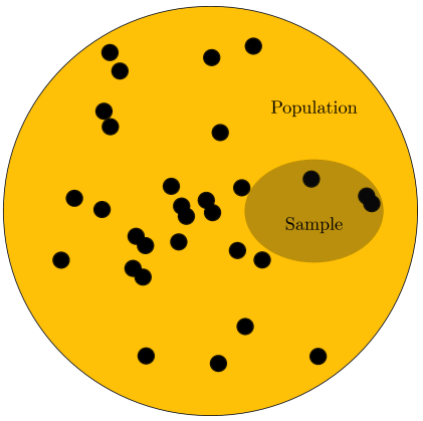
What to do with this information? Extracting value

Data Science consists in extracting and valuing the information contained in the data by means of processing algorithms. Before any exploitation, it is necessary to check data quality. If the latter is insufficient, their use: at best will be impossible, at worst can lead to erroneous results and potentially wrong decision-making, according to the relentless principle of "garbage in, garbage out". If the algorithm inputs consist of poor quality data, no matter the effectiveness of the algorithm, it will lead to poor quality outputs. Anyone who used real-world data knows that data preprocessing steps are essential to achieving satisfactory results.

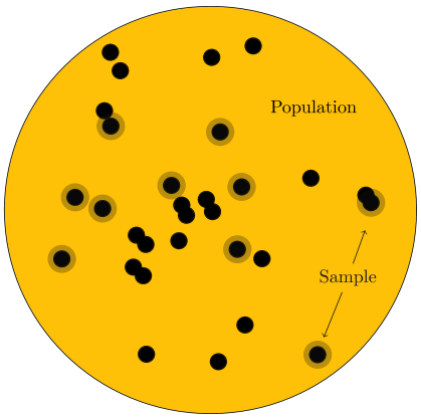
The quality of a data sample is based on content and form. If the "form" quality is quite obvious (noise, missing data, bias, etc.), the "content" quality is more subtle and difficult to estimate a priori.

Typically, a trap that needs to be avoided is to consider that increasing the amount of input data of an algorithm will necessarily improve its performance. Indeed, a large amount of data is not necessarily synonymous with quality. For example, consider a message and a sample built by repeating N times this message. The information of the 2 samples will be strictly identical (equal to that contained in the basic message), while their sizes will be very different. The quality of a data sample is therefore closely linked to the amount of information it contains, regardless of its size.

A similar error is to have only a partial and non-homogeneous data set. Let's take the example where we try to make daily predictions of electricity consumption based on several parameters. The outside temperature, for example, has definitely an effect on the consumption if the building has an electric temperature control system. Consumption estimates for the winter period will be logically incorrect if they are based on temperature data collected during summer. Graphically, the data can be represented as extracted from a population (black dots in the following diagrams). A sampling is relevant if it can represent the variance of the population, i.e. if the data spread over the entire domain of variation of the influencing parameters. The difficulty in testing whether the sample meets this condition lies in estimating the variance of the overall population.



irrelevant sampling



appropriate sampling

Conclusion

We hope that this article allowed you to better understand the distinction between information and data. If entropy gives us an objective measure of the average amount of information a data system contains, there is however not a direct and univocal way to measure the quality of the information in a source of data. In particular, it is necessary to check their "cleanliness" (absence of noise, gaps, etc.), their distribution and their coherence with respect to the information one wish to extract.

Machine learning models are becoming more complex and powerful, but they do not work miracles. If we want to harness their full potential, it is important to feed them with information-rich data. The amount of data needed is then a necessary condition ... but not sufficient.

For the various studies entrusted to BeeBryte, we have developed a wide range of pre-processing proprietary digital tools that make the most of the information contained in the available data and optimize the performance of AI algorithms powered by these data. These tools are complemented by the experience and know-how of our data-scientists who confirm that the pre-analysis and data preparation phase, too often neglected by some, is nevertheless essential and still difficult to automate when the sources of data are multiple and heterogeneous.

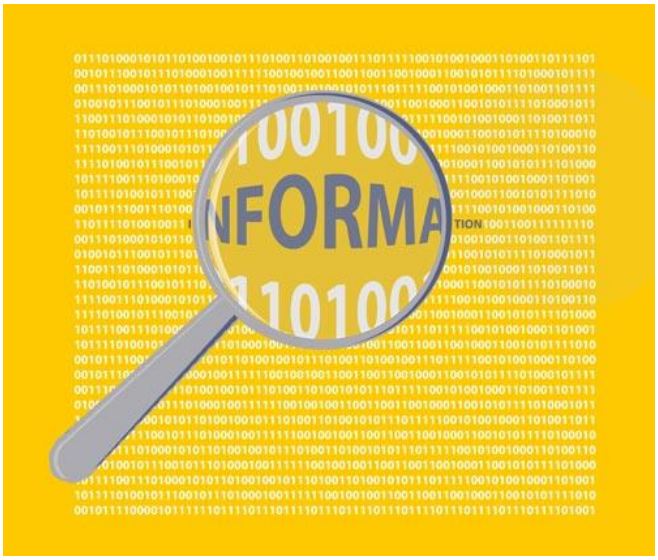
Give us a call for further help!

REFERENCES:

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 is using IoT, AI and BlockChain to get commercial buildings, factories, EV charging stations or entire eco-suburbs to consume electricity in a smarter, more efficient and cheaper way while reducing carbon footprint!

BeeBryte is based in France and Singapore, and is accelerated by Intel & TechFounders.

Since its creation in 2015, BeeBryte's solutions have been awarded by prestigious organizations, such as EDF Pulse, DENA Start-up Energy Transition award & Hello Tomorrow Challenge.

If you want to participate in the energy revolution, please contact us at:

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