Shannon's definition of information is obsolete and inadequate. It is time to embrace Kolmogorov's insights on the matter.

Emanuel Diamant
VIDIA-mant, Kiriat Ono 5510801, Israel
emanl.245@gmail.com

Abstract: Information Theory, as developed by Claude Shannon in 1948, was about the communication of messages as electronic signals via a transmission channel. Only physical properties of the signal and the channel have been taken into account. While the meaning of the message has been ignored totally. Such an approach to information met very well the requirements of a data communication channel. But recent advances in almost all sciences put an urgent demand for meaningful information inclusion into the body of a communicated message. To meet this demand, I have proposed a new definition of information. In this definition, information is seen as a complex notion composed of two inseparable parts: Physical information and Semantic information. Classical informations such as Shannon, Fisher, Renyi, Kolmogorov's complexity, and Chaitin's algorithmic information – they are all physical information variants. Semantic information is a new concept and it desires to be properly studied, treated, and used.

MAIN TEXT

We live in an information age but I am not sure that somebody is skilled enough to explain what does the term "information" mean. Claude E. Shannon was the first who introduced a mathematical definition of information in his seminal work "A Mathematical Theory of Communication" [1]. According to his 1948 paper [1], and his and Weawer's 1949 paper [2], information is concerned with the statistical properties of a given system specifically intended for data communication. In this regard, information "is used in a special sense that must not be confused with its ordinary usage. In particular, information must not be confused with meaning" [2].

Despite all that was just said above, there was always a popular belief that a communication message is carrying with it not only the Shannon's information but also something else, what Shannon and his colleagues have usually called "meaning". A bit later, in 1952, Yehoshua Bar-Hillel and Rudolf Carnap have coined a special name for this "something" – they have called it "Semantic Information", [3]. Their attempt to link semantic information with Shannon's approach to information has failed. In the upcoming years, all further attempts to pursue this subject have been quieted (as unpromising).

However, recent progress in almost all sciences (especially in life sciences and genomics) has revived the efforts to define a clear and unambiguous demarcation of Shannon and semantic information boundaries. Approaching this challenge from a Shannon information standpoint has not seemed to me as a good idea. Therefore, I decided to accept Kolmogorov's understanding of the matters.

In the mid-60s of the past century, Kolmogorov has proposed an algorithmic approach to a quantitative information definition [4]. According to Kolmogorov, a not random binary string (called a separate finite object) can be represented by a compressed description of it (produced by a computer program in an algorithmic fashion) "in such a way that from the description, the original message can be completely reconstructed" [5]. "The amount of information in the string is then defined as the size of the shortest computer program that outputs the string and then terminates" [5]. (For a really random string such a condensed description cannot be provided and "the shortest program for generating it is as long as the chain itself" [6]). The compressed description of a binary object has been dubbed as "algorithmic information" and its quantitative measure (the length of the descriptive program) has been dubbed as the description "Complexity".

Taking Kolmogorov's insights as a starting point, I have developed my own definition of information that can be articulated in the following way: **"Information is a linguistic description of structures observable in a given data set".**

To make the scrutiny into this definition more palpable I propose to consider a digital image as a data set. A digital image is a two-dimensional set of data elements called picture elements or pixels. In an image, pixels are distributed not randomly, but, due to the similarity in their physical properties, they are naturally grouped into some clusters or clumps. I propose to call these clusters **primary or physical data structures**.

In the eyes of an external observer, the primary data structures are further arranged into more larger and complex agglomerations, which I propose to call **secondary data structures**. These secondary structures reflect human observer's view on the grouping of primary data structures, and therefore they could be called **meaningful or semantic data structures**. While formation of primary (physical) data structures is guided by objective (natural, physical) properties of the data, the subsequent formation of secondary (semantic) data structures is a subjective process guided by human conventions and habits.

As it was said, **Description of structures observable in a data set should be called "Information".** In this regard, two types of information must be distinguished – **Physical Information and Semantic Information**. They are both language-based descriptions; however, physical information can be described with a variety of languages (recall that mathematics is also a language), while semantic information can be described only by means of natural human language. (More details on the subject could be find in [7]).

Those, who will go and look in [7], would discover that every information description is a top-down-evolving coarse-to-fine hierarchy of descriptions representing various levels of description complexity (various levels of description details). Physical information hierarchy is located at the lowest level of the semantic hierarchy. The process of sensor data interpretation is reified as a process of physical information extraction from the input data, followed by an attempt to associate this physical information (about the input data) with physical information already retained at the lowest level of the semantic hierarchy. If such an association is attained, the input physical information becomes related (via the physical information retained in the system) with a relevant linguistic term, with a word that places the physical information in the context of a phrase, which provides the semantic interpretation of it. In such a way, the input physical information becomes named with an appropriate linguistic label and framed into a suitable linguistic phrase (and further – in a story, a tale, a narrative), which provides the desired meaning for the input physical information.

The segregation between physical and semantic information is the most essential insight about the nature of information gained from the new information definition. Information processing tools developers know nothing about the duality of information, about the subjective nature of the semantic information (which is a convention, a mutual agreement among the observers, and therefore, it is an observers property and not a property of the data). They do not recognize the existence of secondary structures and the subjective rules of their formation. Therefore, in the light of our new understandings, it can be said that Information processing tools developers today are busy with physical information processing only. That is, merely with extensive data processing. Semantic information cannot be derived from data (from the physical information describing the data)! So the current practice of information processing is doomed to be always restricted to data processing only.

Another important outcome from the new definition is the claim that information descriptions are always reified as a string of words, a piece of text, a narrative. That puts a clear watershed between the new information definition and the pool of the existing ones. All them (the existing ones) are data related and data modeling practices. Designers of methods for semantic information processing would have to invent the tools and the techniques for text strings and whole pieces of text processing. We are now not even close to meet such a challenge. We have first to embrace the notion of information as a complex entity: like complex numbers in mathematics, which are a composition of two independent parts - a Real Number part and an Imaginary Number part, our information can be perceived as a composition of Physical Information and Semantic Information segments. Only this way we would be able to advance in our quest for information processing.

CONCLUSIONS

The paper is intended to bring to the general public awareness the urgent need for a new information definition, which will take into account the hidden duality of the information notion. I did my best to inform

you about the commonly prevailing misconceptions that usually derail the designers' attempts to harness real information handling and processing. I hope my humble efforts will not be in vein.

References

[1] Shannon, C. **A Mathematical Theory of Communication**, 1948, Published by the Board of Trustees of the University of Illinois. Used with the permission of University of Illinois Press.
http://www.mast.queensu.ca/~math474/shannon1948.pdf

[2] Shannon, C., Weaver, W., **The Mathematical Theory of Communication**, Univ. of Illinois Press, 1949.
http://raley.english.ucsb.edu/wp-content/Engl800/Shannon-Weaver.pdf

[3] Bar-Hillel, Y. & Carnap, R. (1952), **An outline of a theory of semantic information**, Technical report No.247, October 27, 1952, http://www.survivor99.com/lcg/information/CARNAP-ILLEL.pdf

[4] Kolmogorov, A. **Three approaches to the quantitative definition of information**, Problems of Information and Transmission, Vol. 1, No. 1, pp. 1-7, 1965.
http://alexander.shen.free.fr/library/Kolmogorov65_Three-Approaches-to-Information.pdf

[5] Peter Grunwald and Paul Vitanyi, **Algorithmic Information Theory**, 2008.
http://arxiv.org/pdf/0809.2754.pdf

[6] Peter Grunwald and Paul Vitanyi, **Shannon Information and Kolmogorov Complexity**, 2004.
http://arxiv.org/pdf/cs/0410002.pdf

[7] Diamant, E., **Brain, Vision, Robotics and Artificial Intelligence.**
http://www.vidia-mant.info