The Nature of Information

By Luis M. Rocha and Santiago Schnell

"*Information is that which reduces uncertainty*". (Claude Shannon)
"*Information is that which changes us*". (Gregory Bateson)
"*Information is a semantic chameleon*". (Rene Thom)

Overview and history

The word **information** derives from the Latin *informare* (*in* + *formare*), which means "to give form, shape, or character to" something. Etymologically, it is therefore understood to be the formative principle of something, or to imbue that something with a specific character or quality. However, for hundreds of years, the word information has been used to signify knowledge and aspects of cognition such as meaning, instruction, communication, representation, signs, symbols, etc. This can be clearly appreciated in the *Oxford English Dictionary*, which defines information as "the action of informing; formation or molding of the mind or character, training, instruction, teaching; communication of instructive knowledge".

Two of the most outstanding achievements of the twentieth-century were the invention of computers and the birth of molecular biology. The advances made in these two fields over the past three decades have resulted not only in the generation of vast amounts of data and information, but also in a new understanding of the concept of information itself. Furthermore, modern science is unraveling the nature of information in numerous areas such as communication theory, biology, neuroscience, cognitive science, and education, among others.

The purpose of information

Information, in essence, does not constitute a specific or specialized area; it is not a particular discipline or field. Rather, information is the basis of all communication; it is used in the process of categorizing our environment helping us to cope with it. Therefore, the study of information in all its aspects pertains to many disciplines: from Science to Philosophy.

Information allows us to think about reality, as well as to communicate our thoughts about it. Depending on one's point of view, information represents reality or is used to construct it. In either case, when you are deprived of information, the world becomes darker and oppressive. Without information, without records, reports, books, news, education, etc, the reach of experience trails off into the shadows of ignorance.

Therefore, information accomplishes a two-fold purpose. First, information conveys our representations of *reality*. Second, information is destined to (be communicated to) *someone or something*. These two aspects of information, though distinct, are nevertheless not separated---one does not exist without the other.

At first we may well presume that a token of information is simply a factual representation of reality, but representation of reality to whom? The act of representing something as a piece of knowledge implies the separate existence of the thing being represented and the representation of the thing, between the known and the knower. What happens here is already a form of communication: the representation of an object communicates the existence of the (known) object to the knower who recognizes the representation.

The structure of information: Semiotics

When we look at the world and study reality, we see order and structure everywhere. There is nothing that escapes description or explanation, even in the natural sciences where phenomena appear sometimes catastrophic, chaotic and random.

A good example of order and information are our roads. Information can be delivered by signs. Drivers know that signs are not distant things, but they are *about* distant things in the road. What signs deliver are not things but a sense or knowledge of things---a message. For information to work that way, there have to be signs. These are special objects whose function is to be *about* other objects. The function of signs is reference rather than presence. Thus a system of signs is crucial for information to exist and be useful in a world, particularly for the world of drivers!

The central structure of information is therefore a relation among signs, objects or things, and agents capable of understanding (or decoding) the signs. An AGENT is *informed* by a SIGN about some THING. There are many names for the three parts of this relation. The AGENT can be thought of as the recipient of information, the listener, reader, interpretant, spectator, investigator, computer, cell, etc. The SIGN has been called the signal, symbol, vehicle, or messenger. And the about-some-THING is the message, the meaning, the content, the news, the intelligence, or the information.

The SIGN-THING-AGENT relation is often understood as a sign-system, and the discipline that studies sign systems is known as Semiotics[1]. Because we are animals who use language in almost all aspects of our existence, sign and symbol-systems are normally second nature to us---we are usually not even aware that we use them! However, they can come into focus in circumstances where an object oscillates between sign and thing or suddenly reverts from reference to presence. This play on signs as things belongs to a tradition of figure poems, represented in the USA by John Hollander and illustrated by "Kitty: Black Domestic Shorthair" (see Figure 1). Within the silhouette of Kitty there is a tale of cats. The play on signs has also been used extensively in

[1] See the Wikipedia definition

Surrealist and Pop Art (e.g. Magritte and Warhol), often to highlight a conflict between reference and presence (see Figure 2), and modern music (e.g. sampling in Hip Hop)[2].

However, an intelligent informatics student would understand that an object is not simply a sign or a thing; context specifies whether it is one of the other. Unfortunately, our context depends on our current needs and standpoints. The purpose of our actions is also shaped by context. It is not good to steal food for the pleasure of stealing food. However, if we are hungry, we have no money and other resources for obtaining food, stealing food cannot be judged as a bad action.

It is the consonance of context that makes the world or reality coherent. Hence in addition to the triad of a sign-system, a complete understanding of information requires four elements: an AGENT is informed by a SIGN about some THING in a certain CONTEXT. Indeed, (Peircean) semiotics emphasizes the *pragmatics* of sign-systems, in addition to the more well-known dimensions of *syntax* and *semantics*. Therefore, a complete (semiotic) understanding of information studies these three dimensions of sign-systems:

- Semantics: the content or meaning of the SIGN of a THING for an AGENT; it studies all aspects of the relation between signs and objects for an agent, in other words, the study of meaning.
- Syntax: the characteristics of signs and symbols devoid of meaning; it studies all aspects of the relation among signs such as their rules of operation, production, storage, and manipulation.
- Pragmatics: the context of signs and repercussions of sign-systems in an environment; it studies how context influences the interpretation of signs and how well a sign-system represents some aspect of the environment.

As we shall see throughout this course, Informatics understood as Information Technology deals essentially with the syntax of information, that is, with issues of data manipulation, storage, retrieval, and computation independently of meaning. Other lesser-known sub-fields of Informatics deal with semantics and pragmatics, for instance, Human-Computer Interaction, Social Informatics and Science Informatics as well.

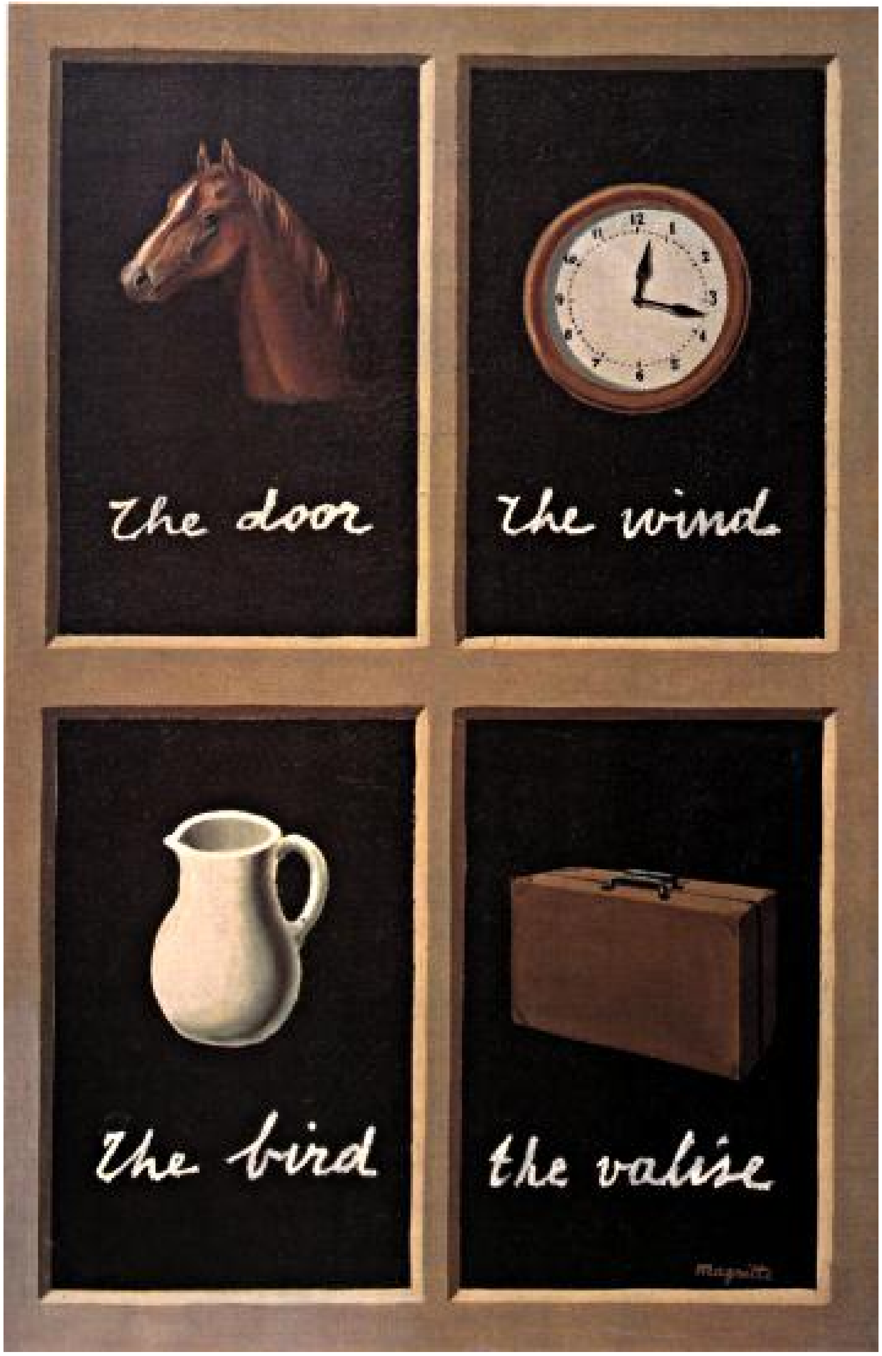
In our presentation of sign systems, we left the concept of AGENT rather vague. An agent can be a cell receiving a biochemical message, or a robot processing some visual input, but it is typically understood as a person. Moreover, it is not true, that any person (or agent), faced with a sign in a certain context, can recognize what the sign is about. It takes intelligence to do so, normal intelligence for customary signs, unusual intelligence when the signs are extraordinary. Therefore, by an AGENT we mean someone or something with the intelligence or capability to produce and process information in context.

[2] We strongly recommend the movie version of the Umberto Eco's book *The Name of The Rose*. In the book, which we highly recommend, an old manuscript, the message, for being literarily dangerous becomes literally poisonous: reference and presence become very intertwined indeed! The book was made into a movie in 1986, Starring Sean-Connery and Christian Slater, and Directed by Jean-Jacques Annaud.

Figure 1: John Hollander “Kitty: Black Domestic Shorthair” in Types of Shape (New Haven, Yale University Press, 1991).

```
            O      I
           am     my
          own    way
         of being in
         view and yet
         invisible at
         once Hearing
          everything
          you see I
          see all of
         whatever you
        can have heard
       even inside the
      deep silences of
     black silhouettes
    like these images
   of furry surfaces
  darkly playing cat
 and mouse with your
doubts about whether
other minds can ever
be drawn from hiding
and made to be heard
in inferred language
 I can speak only in
  your voice Are you
   done with my shadow
    That thread of dark
     word
      can
     all
    run
    out
     now
      and
       end
        our
      tale
```

Figure 2: The Key of Dreams, 1930, Rene Maggritte



Types of Signs

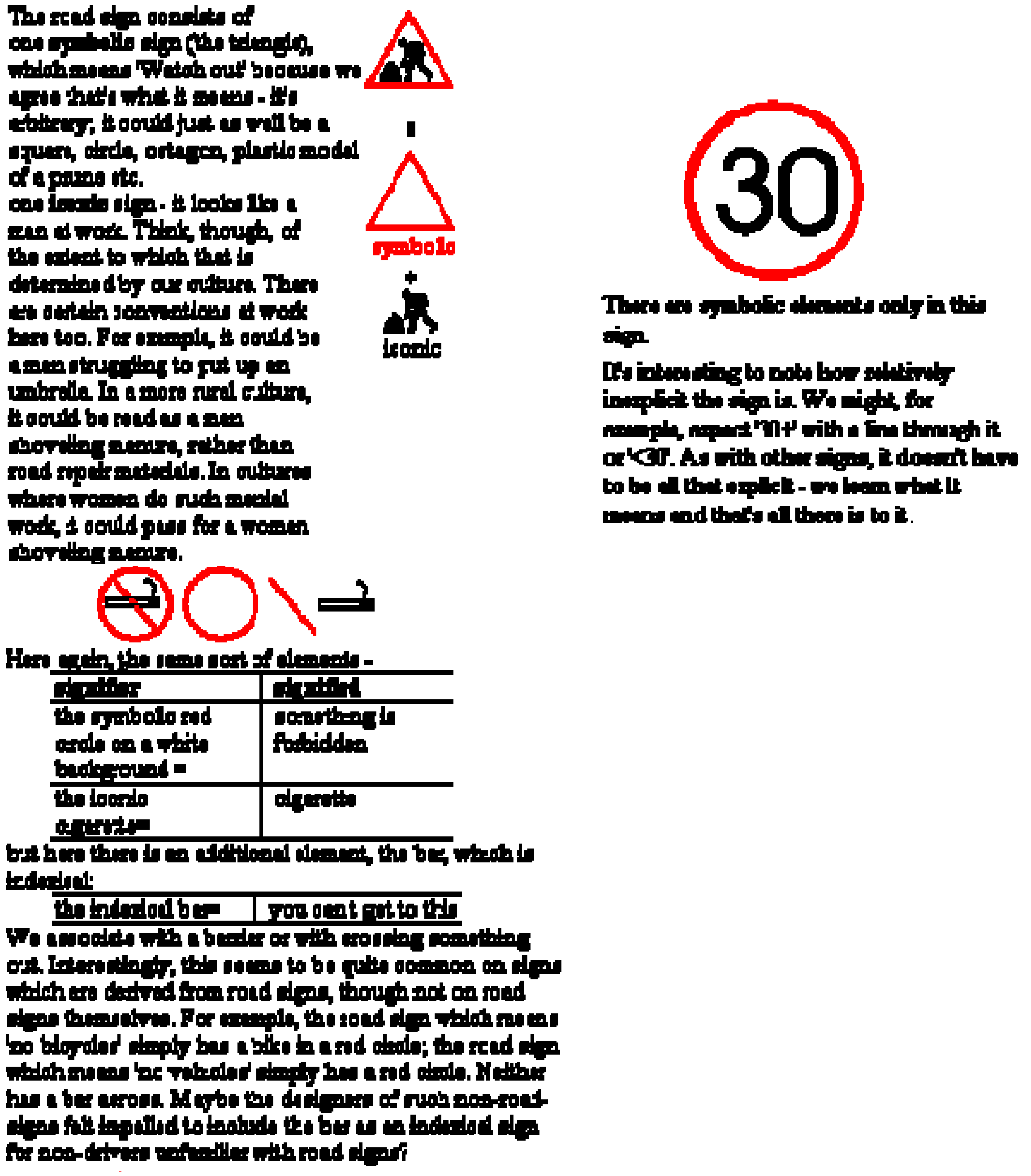
Signs carry information content to be delivered to agents. However, it is also useful to understand that some signs are more easily used as referents than others. In the beginning of the 20th century, Charles Sanders Peirce defined a typology of signs:

1. **Icons** are direct representations of objects. They are similar to the thing they represent. Examples are pictorial road signs, scale models, and of course the icons on your computer. A footprint on the sand is an icon of a foot.
2. **Indices** are indirect representations of objects, but necessarily related. Smoke is an index of fire, the bell is an index of the tolling stroke, and a footprint is an index of a person.
3. **Symbols** are *arbitrary* representations of objects, which require exclusively a social convention to be understood. A road sign with a red circle and a white background denotes something which is illegal because we have agreed on its arbitrary meaning. To emphasize the conventional aspect of the semantics of symbols, consider the example of variations in road signs: in the US yellow diamond signs denote cautionary warnings, whereas in Europe a red triangle over a white background is used for the same purpose. We can see that convention establishes a **code**, agreed by a group of agents, for understanding (decoding) the information contained in symbols. For instance, smoke is an index of fire, but if we agree on an appropriate code (e.g. Morse code) we can use smoke signals to communicate symbolically.

Clearly, signs may have iconic, symbolic and indexical elements. Our alphabet is completely symbolic, as the sound assigned to each letter is purely conventional. But other writing systems such as Egyptian or Mayan hieroglyphs, and some Chinese characters use a combination of phonetic symbols with icons and indices. Our road signs are also a good example of signs with symbolic (numbers, letters and conventional shapes), iconic (representations of people and animals) and indexical (crossing out bars) elements – see examples in figure 3.

Finally, it is important to note that due to the arbitrary nature of convention, symbols can be manipulated without reference to content (syntactically). This feature of symbols is what enables computers to operate, as we shall see throughout this course. As an example of symbol manipulation without recourse to content, let us re-arrange the letters of a word, say “deal”: dale, adel, dela, lead, adle, etc. We can produce all possible permutations (4! = 4×3×2×1 = 24) of the word whether they have meaning or not. After manipulation, we can choose which ones have meaning (in some language), but that process is now a semantic one, whereas symbol manipulation is purely syntactic. Another example is the (beat) word cut-up method of generating poetry pioneered by Brion Gysin and William Burroughs and often used by artists such as David Bowie.

All signs rely on a certain amount of convention, as all signs have a pragmatic (social) dimension, but symbols are the only signs which require exclusively a social convention, or code, to be understood.


=
symbolic
+
iconic
30

The road sign consists of one symbolic sign (the triangle), which means 'Watch out' because we agree that's what it means - it's arbitrary, it could just as well be a square, circle, octagon, plastic model of a prawn etc.
one iconic sign - it looks like a man at work. Think, though, of the extent to which that is determined by our culture. There are certain conventions at work here too. For example, it could be a man struggling to put up an umbrella. In a more rural culture, it could be read as a man shovelling manure, rather than road repair materials. In cultures where women do such menial work, it could pass for a woman shovelling manure.

Here again, the same sort of elements -

signifier	signified
the symbolic red circle on a white background =	something is forbidden
the iconic cigarette=	cigarette

but here there is an additional element, the bar, which is indexical:

the indexical bar=	you can't get to this

We associate with a barrier or with crossing something out. Interestingly, this seems to be quite common on signs which are derived from road signs, though not on road signs themselves. For example, the road sign which means 'no bicycles' simply has a bike in a red circle; the road sign which means 'no vehicles' simply has a red circle. Neither has a bar across. Maybe the designers of such non-road-signs felt impelled to include the bar as an indexical sign for non-drivers unfamiliar with road signs?

There are symbolic elements only in this sign.

It's interesting to note how relatively inexplicit the sign is. We might, for example, expect '30' with a line through it or '<30'. As with other signs, it doesn't have to be all that explicit - we learn what it means and that's all there is to it.

Figure3: Semiotics of road signs, by Mick Underwood: http://www.cultsock.ndirect.co.uk/MUHome/cshtml/semiomean/semio1.html

Information Theory and the Bit

Information became a prominent word and notion in an article published in 1948 by Claude Shannon. However, the word information does not figure in the title, which is "The mathematical theory of communication", even though it became known as the (Shannon-Weaver) *Information Theory*.

The crux of this information theory, originally developed to deal with the efficiency of information transmission in electronic channels, is the definition of an *information quantity* that can be measured unequivocally. The price to pay for the ability to objectively measure such a quantity is that it does not deal at all with the subjective aspects of information, namely semantics and pragmatics. Indeed, information is defined as a quantity that depends on symbol manipulation alone

Such analysis of information is concerned with the discovery of the elementary particles or units of information. But if information is a relation between an agent, a sign and a thing, rather than simply a thing, it is far from obvious what in information is reducible or quantifiable and what is not. The most palpable element in the information relation is the sign, and here reduction, if not measurement, is a feasible enterprise. Among signs, in turn, it is the system of conventional signs we call symbols, such as those used in written language, that lend themselves best to analysis. But which symbols do we use to quantify the information contained in messages?

One might think that the sound structure of language requires 26 or so symbols. Yet letters are not snapshots of the infinite variety in which people pronounce and intonate words, but the result of a systematic simplification, balanced between fidelity to the acoustic reality of speech and parsimony for the sake of efficiency. In ancient Greece, the balance came to rest on 24 letters. The University of Oxford phonetic rendering of English require some 40 symbols. The American Standard Code for Information Interchange (ASCII) contains 82 symbols, 26 lower and 26 upper case letters, 10 number signs for the decimal symbols, and 20 punctuation and function signs. We could do with 26 symbols if we did without punctuation and lower case letters (as the first alphabetic writers did) and rendered function and number signs in letters, + as plus, 12 as twelve, and so on. But is this the limit of notational economy?

To quantify information conveyed by symbols, Shannon studied the way symbolic messages are sent and received via communication channels. For communication to occur, both sender and receiver must use the same code, or convention, to encode and decode symbols from and to messages. This means that we need to fix the *language* used for communication, that is, the set of symbols allowed (an *alphabet*), the rules to manipulate symbols (*syntax*), and the meaning of the symbols (*semantics*). Once a language is fixed, the universe of all possible messages is implicitly specified as the set of all possible symbol *strings* of a given size. Finally, information is then defined as "a measure of the freedom from *choice* with which a message is *selected* from the set of all possible messages" (*The Columbia Encyclopedia*, Sixth Edition. 2001[3]).

What this means is that information is defined as the act of selecting a specific message (a string of symbols) from the set of all possible messages (in some language). Shannon then defined information content of a message as the number of operations needed to select that message from the set of all possible messages. The selection process depends on the likelihood of occurrence of symbols. Later in the course we will deal with the specific (probabilistic) definition of information content that Shannon proposed. For now, it suffices to understand that this process of selection, and therefore information content, depends on the number of choices that exist when we encode a message of a given size.

Notice that the number of choices depends entirely on the symbolic language used, and not at all on the meaning of the message! For instance, the words "information" and "anerthingly" written in the Roman alphabet with 26 symbols are one of 26^{11} (=3,670,344,486,987,776 ≈ $3.7\text{x}10^{15}$) possible words of size 11. Therefore, both have the same information content (assuming that each symbol has the same likelihood), even though the first has an English meaning and the second does not.

Furthermore, in the phonetic language of the *Oxford University Dictionary*, which uses additional symbols, the word "information" is encoded as ɪnfəˈmeɪʃən. This particular encoding of the word possesses many more competing alternative words of the same size: 40^{11} (=419,430,400,000,000,000≈$4.2\text{x}10^{17}$) – indeed, more than 100 times more alternatives than when encoded in the regular Roman alphabet.

Since information content depends on the language used, Shannon needed to compute information content on the most economical symbol system available, which he proved to be the binary system. The following passage from von Baeyer's [2004] book "*Information: The new language of Science*" describes why the binary system is the least expensive to encode messages:

> "Far from selecting the binary code arbitrarily, or believing it to be the simplest possible scheme, Shannon proved that it is, in fact, the least expensive way to handle information. It uses up the smallest amount of resources in the form of electronic memories and bandwidth in communication channels. […].
>
> To get a flavour of Shannon's proof, consider a sailor who wants to signal a number between 0 and 127 by means of flags. If he decides to fly just a single flag to do the job, he needs to have 128 different flags in his locker. A less expensive strategy would be to fly three flags to spell out the number in the decimal system. To do this, he would need to own just twenty-one flags – ten for the units, ten for the tens and one for hundreds. The cheapest technique is based on the binary system, with only fourteen flags – seven zeroes and seven ones – he can compose any number up to 127". Von Baeyer [2004, pages 30-31].

[3] http://www.bartleby.com/65/in/inform-th.html.

Since the binary system of encoding messages using only two symbols, typically “0” and “1”, is the most economical, to measure information content, Shannon’s theory demands that we encode every message in binary, and then count alternative choices in this system.

The most elementary choice one can make is between two items: “0’ and “1”, “heads” or “tails”, “true” or “false”, etc. Shannon defined the *bit* as such an elementary choice, or unit of information content, from which all selection operations are built. Bit is short for binary digit and “is equivalent to the choice between two equally likely choices. For example, if we know that a coin is to be tossed, but are unable to see it as it falls, a message telling whether the coin came up heads or tails gives us one bit of information.” (*The Columbia Encyclopedia*, Sixth Edition. 2001[4])

Analog versus Digital

Shannon’s information measure applies specifically to binary symbol systems, and to digital information in general. *Digital* is used to convey the notion of discrete objects/values, that is, things we can count. Indeed, the word digit comes from the Latin word for finger (*digitus*) as these are used for counting. Any symbol system requires a set of discrete symbols for setting up an arbitrary semantic relation with things in the World. In this sense, digital information is equivalent to symbolic information, which is information that is stored and transmitted using symbols (rather than icons or indices).

The fact that Shannon’s theory applies exclusively to symbolic information was very innovative when it was proposed. In those days, communication was typically thought of as transmission of information via electrical, mechanical, hydraulic, and sound signals, that is, continuously varying signals which are not countable. (Shannon worked for Bell labs). We refer to that type of signal as analog (or analogue) because it implies an analogy between cause and effect. Whereas symbolic information relies on an arbitrary encoding between a message and its meaning, an analog signal relies on some property of the medium to convey the message’s information. For instance, Morse code is digital, because we can encode it in any discrete (countable) medium (identifiable sounds, smoke signals, binary symbols, etc.) It is independent of the medium utilized. An analog synthesizer, on the other hand, uses voltage signals to convey sounds. It uses the characteristics of electrical signals to produce sounds (via a transducer), in this sense it uses voltage as an “analogy” for sound. But the sounds that it can convey depend on the physical properties of electricity.

Another example of analog technology is vinyl records, which store sound by means of grooves in a record which are read by a diamond needle. It is the direct, physical relation between the medium and the sound that leads DJ’s and Hip Hop artists to prefer vinyl records to digital CDs, as the manipulation of the record, for “scratching”, for instance, is much more immediate. We can also use analog or digital technology to tell the time.

[4] http://www.bartleby.com/65/in/inform-th.html

Information-processes in Nature: codes

We are used to think of information as pertaining purely to the human realm. In particular, the use of symbolic information, as in our writing system, is thought of as technology used exclusively by humans. Symbols, we have learned, rely on a code, or convention, between symbols and meanings. Such a conventional relation usually specifies rules created by a human community. But it can have a more general definition:

> "A code can be defined as a set of rules that establish *a correspondence between two independent worlds*". The Morse code, for example, connects certain combinations of dots and dashes with the letters of the alphabet. The Highway Code is a liaison between illustrated signals and driving behaviours. A language makes words stand for real objects of the physical World." [Barbieri, 2003, page 94]

We can thus think of a code as a process that implements correspondence rules between two independent worlds (or classes of objects), by ascribing meaning to arbitrary symbols. Therefore, meaning is not a characteristic of the individual symbols but a convention of the collection of producers and recipients of the encoded information.

Interestingly, we can see such processes in Nature, where the producers and recipients are not human. The prime example is the genetic code, which establishes a correspondence between DNA (the symbolic genes which store information) and proteins, the stuff life on Earth is built of. With very small variations, the genetic code is the same for all life forms. In this sense, we can think of the genetic system and cellular reproduction as a symbolic code whose convention is "accepted" by the collection of all life forms.

Other codes exist in Nature, such as signal transduction from the surface of cells to the genetic system, neural information processing, antigen recognition by antibodies in the immune system, etc. We can also think of animal communication mechanisms, such as the ant pheromone trails, bird signals, etc. Unlike the genetic system, most information processes in nature are of an analog rather than digital nature.

Bibliography

Atlan, H. [1983]. "Information Theory". In *Cybernetics: Theory and Applications*, R. Trappl (ed.). Hemisphere, pp. 9-41.

Barbieri, M. [2003]. *The Organic Codes: An Introduction to Semantic Biology*. Cambridge University Press.

Eco, U. [1984]. *Semiotics and the Philosophy of Language*. Indiana University Press.

Kampis, G. [1991*]. Self-modifying Systems in Biology and Cognitive Science: A New Framework for Dynamics, Information and Complexity*. Pergamon Press.

Underwood, M. [2003]. *Introductory models and basic concepts: Semiotics*. http://www.cultsock.ndirect.co.uk/MUHome/cshtml/semiomean/semio1.html

Von Bayer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.