

## Chapter 2

# The Nature of Information

What is information? We have already asserted that it is a profound, primitive (i.e., irreducible) concept. Dictionary definitions include “(desired) items of knowledge”; for example, one wishes to know the length of a piece of wood. It appears to be less than a foot long, so we measure it with our desktop ruler marked off in inches, with the result, let us say, “between six and seven inches.” This result is clearly an item of desired knowledge, hence information. We will return to this example later. Another definition is “fact(s) learned about something,” implying that there is a definable object to which the facts are related, suggesting the need for context and meaning. A further definition is “what is conveyed or represented by a particular arrangement of things”; the dots on the head of a matrix printer shape a letter, the bar code on an item of merchandise represents facts about the nature, origin, and price of the merchandise, and a sequence of letters can convey a possibly infinite range of meanings. A thesaurus gives as synonyms “advice, data, instruction, message, news, report.” Finally, we have “a mathematical quantity expressing the probability of occurrence of a specific sequence of symbols or impulses as against that of other sequences (i.e., messages).” This definition links the quantification of information to a probability, which, as we shall see, plays a major rôle in the development of the subject.

We also note that “information science” is defined as the “study of processes for storing and retrieving information,” and “information theory” is defined as the “quantitative study of transmission processes for storing and retrieving of information by signals”; that is, it deals with the mathematical problems arising in connexion with the storage, transformation, and transmission of information. This forms the material for Chapter 3. Etymologically, the word “information” comes from the Latin *forma*, form, from *formare*, to give shape to, to describe.

Most information can be reduced to the response, or series of responses, to a question, or series of questions, admitting only yes or no as an answer. We call these yes/no, or dichotomous, questions. Typically, interpretation depends heavily on context. Consider a would-be passenger racing up to a railway station. His question “has the train gone?” may indeed be answered by “yes” or “no”—although, in practice, a third alternative, “don’t know,” may be encountered. At a small wayside station, with the traveller arriving within five minutes of the expected departure time of the only train scheduled within the next hour, the answer (yes or no) would

be unambiguous and will convey exactly one bit of information, as will be explained below. If we insist on the qualification “desired,” an unsolicited remark of the stationmaster, “the train has gone,” may or may not convey information to the hopeful passenger. Should the traveller have seen with his own eyes the train depart a minute before, the stationmaster’s remark would certainly not convey any information.

Consider now a junction at which, after leaving the station, the lines diverge in three different directions. The remark “the train has gone”, assuming the information was desired, would still convey one bit of information, but by in addition specifying the direction, viz. “the train has gone to X”, or “the train to X has gone,” “X” being one of the three possible destinations, the remark would convey  $\log_2 3 = 1.59$  bits of information, this being the average number of questions admitting yes/no answers required to specify the fact of departure to X, as opposed to either of the two other directions.

This little scenario illustrates several crucial points:

1. Variety exists. In a formless, amorphous world there is no information to convey.
2. The amount of information received depends on what the recipient knows already.
3. The amount of information can only be calculated if the set of possible messages (responses) has been predefined.

Dichotomous information often has a hierarchical structure; for example, on a journey, a selection of direction has to be made at every cross-road. Given an ultimate destination, successive choices are only meaningful on the basis of preceding ones. Consider also an infant, who “chooses” (according to its environment) which language it will speak. As an adolescent, he chooses a profession, again with an influence from the environment, and in making this choice, knowledge of a certain language may be primordial. As an adult there will be further career choices, which will usually be intimately related to the previous choice of a profession.

Let us now reexamine the measurement of the length of a stick. It must be specified in advance that it does not exceed a certain value—say one foot. This will suffice to allow an appropriate measuring tool to be selected. If all we had was a measuring stick exactly one foot long, we could simply ascertain whether the unknown piece was longer or shorter, and this information would provide one bit of information, if any length was *a priori* possible for the unknown piece.

Suppose, however, that the measuring stick is marked off in 1-inch divisions. If the probabilities  $p$  of the unknown piece being any particular length  $l$  (measured to the nearest inch), with  $0 < l \leq 12$ , were *a priori* equal (i.e.,  $p = \frac{1}{12}$  for each possible length), then the information produced by the measurement equals  $\log_2 12 = 3.59$  bits, this being the average number of questions admitting yes/no answers required to specify the length to the nearest inch, as the reader may verify. On the other hand, were we to have some prior information, according to which we had good reason to suppose the length to be close to 9 inches (perhaps we had previously requested the wood to be chopped to that length), the probabilities of the lengths 8, 9, and 10 inches would perhaps be 0.25 each, and the sum of all the others would be

0.25. The existence of this prior knowledge would somewhat reduce the quantity of information gained from the measurement, namely to  $\frac{3}{4} \log_2 4 + \frac{1}{4} \log_2 36 = 2.79$  bits. Should the ruler have been marked off in tenths of an inch, the measurement would have yielded considerably more information, namely  $\log_2 120 = 6.91$  bits, assuming all the probabilities of the wood being any particular length to be equal (i.e.,  $\frac{1}{120}$  each).

### Variety

One of the most striking characteristics of the natural, especially the living, world around us is its variety. This variety stands in great contrast to the world studied by the methods of physics and chemistry, in which every electron and every proton (etc.) in the universe are presumed to be identical, and we have no evidence to gain-say this presumption. Similarly, every atom of helium ( $^4\text{He}$ ) is similar to every other one, and indeed it is often emphasized that chemistry could only make progress as a quantitative science after the realization that pure substances were necessary for the investigation of reactions and the like, such that a sample of naphthalene in a laboratory in Germany would behave in precisely the same way as one in Japan.

If we are shown a tray containing balls of three colours, red (r), blue (b), and white (w), we might reasonably assert that the variety is three. Hence, one way to quantify variety is simply to count the number of different kinds of objects. Thus, the variety of either of the sets {r, b, w} and {r, b, b, r, w, r, w, w, b} is equal to three; the set {r, r, w, w, w} has a variety of only two, and so forth. The objects considered should of course be in the same category; that is, if the category were specified as “ball,” then we would have difficulty if the tray also included a banana and an ashtray. However, one could then redefine the category.

If there were only one kind of ball, say red, then our counting procedure would yield a variety of one. It is more natural, however, to say that there is no variety if all the objects are the same, suggesting that the logarithm of the number of objects is a more reasonable way to quantify variety. If all the objects are the same, the variety is then zero. We are, of course, at liberty to choose any base for the logarithm; if the base is 2, then conventionally the variety is given in units of bits, a contraction of *binary digit*. Hence, two kinds of objects have a variety of  $\log_2 2 = 1$  bit, and three kinds give  $\log_2 3 = \frac{\log_{10} 3}{\log_{10} 2} = \frac{0.477}{0.301} = 1.58$  bits. The variety in bits is the average number of yes/no questions required to ascertain the number of different kinds of objects or to identify the kind of any object chosen from the set.<sup>1</sup>

---

<sup>1</sup> This primitive notion of variety is related to the diversity measured by biometricians concerned with assessing the variety of species in an ecosystem (biocoenosis). Diversity  $D$  is essentially variety weighted according to the relative abundances (i.e., probability  $p_i$  of occurrence) of the  $N$  different types, and this can be done in different ways. Parameters in use by practitioners include

### The Shannon Index

The formula that we used to determine the quantity  $I$  of information delivered by a measurement that fixes the result as one out of  $n$  equally likely possibilities, each having a probability  $p_i, i = 1, \dots, n$ , all equal to  $1/n$ , was

$$I = -\log p = \log n . \quad (2.4)$$

It is called Hartley's formula. If the base of the logarithm is 2, then the formula yields numerical values in bits. Where the probabilities of the different alternatives are not equal, then a weighted mean must be taken:

$$I = -\sum_{i=1}^n p_i \log_2 p_i . \quad (2.5)$$

This generalization is called the Shannon or Shannon-Wiener index. In other words, the quantity of information is weighted logarithmic variety. Note that the information given by equation (2.5) is always less than that given by the equiprobable case (2.4). This follows from Jensen's inequality.<sup>2</sup>

Why is the negative of the sum taken?  $I$  in fact represents the *gain* of information due to the measurement. In general,

$$\text{gain (in something)} = \text{final value} - \text{initial value} . \quad (2.7)$$

The initial value represents the uncertainty in the outcome *prior* to the measurement. Shannon takes the *final* value (i.e., the result of the measurement), to be a single value with variety one, hence using (2.5),  $I = 0$  after the measurement; that is, he considers the result to be known with certainty once it has been delivered. Hence, it is considered to have zero information, and it is in this sense that an information processor is also an information annihilator. Wiener considers the more general

---


$$D_0 = N \quad (\text{no weighting}), \quad (2.1)$$

$$D_1 = \exp(I) \quad (\text{the exponential of Shannon's index}), \quad (2.2)$$

$$D_2 = 1 / \sum_{i=1}^N p_i^2 \quad (\text{the reciprocal of Simpson's index}). \quad (2.3)$$

<sup>2</sup> If  $g(x)$  is a convex function on an interval  $(a, b)$ , if  $x_1, x_2, \dots, x_n$  are arbitrary real numbers  $a < x_k < b$ , and if  $w_1, w_2, \dots, w_n$  are positive numbers with  $\sum_{k=1}^n w_k = 1$ , then

$$g\left(\sum_{k=1}^n w_k x_k\right) \leq \sum_{k=1}^n w_k g(x_k) . \quad (2.6)$$

Inequality (2.6) is then applied to the convex function  $y = x \log x$  ( $x > 0$ ) with  $x_k = p_k$  and  $w_k = 1/n$  ( $k = 1, 2, \dots, n$ ) to get  $I(p_1, p_2, \dots, p_n) \leq \log n$ .

case in which the result of the measurement could be less than certain (e.g., still a distribution, but narrower than the one measured).

The gain of information  $I$  is equivalent to the removal of uncertainty; hence, information could be defined as “that which removes uncertainty.” It corresponds to the reduction of variety perceived by an observer and is inversely proportional to the probability of a particular value being read, or a particular symbol (or set of symbols) being selected, or, more generally, is inversely proportional to the probability of a message being received and remembered.

*Example.* An  $N \times N$  grid of pixels, each of which can be either black or white, can convey at most  $-\sum_i^{N^2} \frac{1}{2} \log_2 \frac{1}{2}$  bits of information. This maximum is achieved when the probability of being either black or white is equal.

$I$  defined by equations (2.4) and (2.5) has the properties that one may reasonably postulate should be possessed by a measure of information, namely

1.  $I(E_{NM}) = I(E_N) + I(E_M)$  , for  $N, M = 1, 2, \dots$  ;
2.  $I(E_N) \leq I(E_{N+1})$  ;
3.  $I(E_2) = 1$  .

*Example.* How much information is contained in a sequence of DNA? If each of the four bases are chosen with equal probability (i.e.,  $p = \frac{1}{4}$ ), the information in a decamer is  $10 \log_2 4 = 20$  bits. It is the average number of yes/no questions that would be needed to ascertain the sequence. If the sequence were completely unknown before questioning, this is the gain in information. Any constraints imposed on the assembly of the sequence—for example, a rule that “AA” is never followed by “T,” will lower the information content of the sequence (i.e., the gain in information upon receiving the sequence, assuming that those constraints are known to us). Some proteins are heavily constrained; the antifreeze glycoprotein (alanine-alanine-threonine)<sub>*n*</sub> could be simply specified by the instruction “repeat AAT *n* times”, much more compactly than writing out the amino acid sequence in full, and the quantity of information gained upon being informed of the sequence is correspondingly small.

### Thermodynamic Entropy

One often encounters the word “entropy” used synonymously with information (or its removal). Entropy ( $S$ ) in a physical system represents the ability of a system to absorb energy without increasing its temperature. Under isothermal conditions (i.e., at a constant temperature  $T$ ),

$$dQ = T dS , \quad (2.8)$$

where  $dQ$  is the heat that flows into the system. In thermodynamics, the internal energy  $E$  of a system is formally defined by the First Law as the difference between the heat and  $dW$ , the work done by the system:

$$dE = dQ - dW . \quad (2.9)$$

The only way that a system can absorb heat without raising its temperature is by becoming more disordered. Hence, entropy is a measure of disorder. Starting from a microscopic viewpoint, entropy is given by the famous formula inscribed on Boltzmann's tombstone:

$$S = k_B \ln W , \quad (2.10)$$

where  $k_B$  is his constant and  $W$  is the number of (micro)states available to the system. Note that reducing the number of states reduces the disorder. An amount of information of  $\log_2 W$  bits is required to specify one particular microstate (assuming that all microstates have the same probability of being occupied) according to Hartley's formula; the specification of a particular microstate removes that quantity of uncertainty. Thermodynamical entropy defined by equation (2.8), statistical mechanical entropy (2.10), and the Hartley or Shannon index only differ from each other by numerical constants.

Although the set of positions and momenta of the molecules in a gas at a given instant can thus be considered as information, within a microscopic interval (between atomic collisions, of the order of 0.1 ps) this set is forgotten and another set is realized. The positions and momenta constitute microscopic information; the quantity of macroscopic (remembered) information is zero. In general, the quantity of macroinformation is far less than the quantity of (forgotten) microinformation, but the former is far more valuable.<sup>3</sup>

In the world of engineering, this state of affairs has of course always been recognized. One does not need to know the temperature (within reason!) in order to design a bridge or a mechanism. The essential features of any construction are found in a few large-scale correlated motions; the vast number of uncorrelated, thermal degrees of freedom are generally unimportant.

### Symbol and Word Entropies

The Shannon index (2.5) gives the average information per symbol; an analogous quantity  $I_n$  can be defined for the probability of  $n$ -mers ( $n$ -symbol "words"), whence the differential entropy  $\tilde{I}_n$ ,

$$\tilde{I}_n = I_{n+1} - I_n , \quad (2.11)$$

---

<sup>3</sup> "Forgetting" implies decay of information; what does "remembering" mean? It means to bring a system to a defined stable state (i.e., one of two or more states), and the system can only switch to another state under the influence of an external impulse. The physical realization of such systems implies a minimum of several atoms; as a rule a single atom, or a simple small molecule, can exist in only one stable state. Among the smallest molecules fulfilling this condition are sugars and amino acids, which can exist in left- and right-handed chiralities. Note that many biological macromolecules and supermolecular assemblies can exist in several stable states.

whose asymptotic limit ( $n \rightarrow \infty$ ) Shannon calls “entropy of the source”, is a measure of the information in the  $(n + 1)$ th symbol, assuming the  $n$  previous ones are known. The decay of  $\tilde{I}_n$  quantifies correlations within the symbolic sequence (i.e., an aspect of and memory).

## 2.1 Structure and Quantity

In our discussion so far we have tacitly assumed that we know *a priori* the set from which the actual measurement will come. In an actual physical experiment, this is like knowing from which dial we shall take readings of the position of the pointer, for example, and, furthermore, this knowledge may comprise all the information required to construct and use the meter, which is far more than that needed to formally specify the blueprints and circuit diagram. It would also have to include blueprints for the machinery needed to make the mechanical and electronic components, for manufacturing the required materials from available matter, and so forth. In many cases we do not need to concern ourselves about all this, because we are only interested in the gain in information (i.e., loss of uncertainty) obtained by receiving the result of the dial reading, which is given by equation (2.5). The information pertinent to the construction of the experiment usually remains the same, hence cancels out (equation 2.7). In other words, the Shannon-Weaver index is strictly concerned with the metrical aspects of information, not with its structure.

### 2.1.1 The Generation of Information

Prior to carrying out an experiment, or an observation, there is objective uncertainty due to the fact that several possibilities (for the result) have to be taken into account. The information furnished by the outcome of the experiment reduces this uncertainty: R.A. Fisher defined the quantity of information furnished by a series of repeated measurements as the reciprocal of the variance.

### 2.1.2 Conditional and Unconditional Information

Information about real events that have happened (e.g., a volcanic eruption), or about entities that exist (e.g., a sequence of DNA) is primarily unconditional; that is, it does not depend on anything (as soon as information is encoded, however, it becomes conditional on the code).

Scientific work has two stages:

1. Receiving unconditional information from nature (by making observations in the field, doing experiments in the laboratory).
2. Generating conditional information in the form of hypotheses and theories relating the observed facts to each other using axiom systems. The success of any

theory (which may be one of several) largely depends on general acceptance of the chosen propositions and the mathematical apparatus used to manipulate the elements of the theory; that is, there is a strongly social aspect involved.

Conditional information tends to be unified; for example, a group of scattered tribes, or practitioners of initially disparate disciplines, may end up speaking a common language (they may then comprehend the information they exchange as being unconditional and may ultimately end up believing that there cannot be other languages). Encoded information is conditional on agreement between emitters and receivers concerning the code.

2.1.3 Experiments and Observations

Consider once again the example of the measurement of the length of an object using a ruler and the information gained thereby. The gain presupposes the existence of a world of objects and knowledge, including the ruler itself and its calibration in appropriate units of measurement. The overall procedure is captured, albeit imperfectly, in Fig. 2.1.

The essential point is that “information” has two parts: a prior part embodied by the physical apparatus, the knowledge required to carry out the experiment or observation, and so forth; and a posterior part equal to the loss in uncertainty about the system due to having made the observation. The prior part can be thought of as specifying the set of possible values from which the observed value must come. In a physical measurement, it is related to the structure of the experiments and the instruments it employs, and the millennia of civilization that have enabled such activities. The posterior part (*I*) is sometimes called “missing information” because once the prior part (*K*) is specified, the system still has the freedom, quantified by *I*, to adopt different microstates. In a musical analogy, *K* would correspond to the structure of a Bach fugue and *I* to the freedom the performer has in making interpretational choices while still respecting the structure.<sup>4</sup> One could say that the magnitude of *I* corresponds to the degree of logical indeterminacy inhering in the system, in other words that part of its description that cannot be formulated within itself; it is the amount of *selective* information lacking.

*I* can often be calculated according to the procedures described in the previous section (the Hartley or Shannon index). If we need to quantify *K*, it can be done

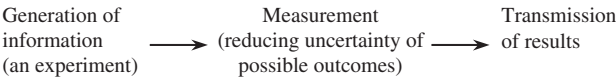


Fig. 2.1 The procedures involved in carrying out an experiment, from conception to ultimate dissemination

<sup>4</sup> Cf. Tureck.



using the concept of algorithmic information content (AIC) or Kolmogorov information, which corresponds to the length of the most concise description of what is known about the system (see §6.5). Hence, the total information<sup>5</sup> is the sum of the ensemble (Shannon) entropy  $I$  and the physical (Kolmogorov) entropy  $K$ :

$$\mathcal{I} = I + K . \quad (2.12)$$

Mackay (1950) proposed the terms “logon” for the structural (prior) information, equivalent to  $K$  in equation (2.12), and “metron” for the metrical (posterior) measurement. The gain in information from a measurement (equation 2.7) falls wholly within the metrical domain, of course, and within that domain, there is a prior and posterior component (cf. §5.4).

To summarize, the Kolmogorov information  $K$  can be used to define the structure of information and is calculated by considering the system used to make a measurement. The result of the measurement is macroscopic, remembered information, quantified by the Shannon index  $I$ . The gain in information equals (final – initial information):

$$I = (I_f + K) - (I_i + K) = I_f - I_i . \quad (2.13)$$

In other words, it is unexceptionable to assume that the measurement procedure does not change the structural information, although this must only be regarded as a cautious, provisional statement. Presumably, any measurement or series of measurements that overthrows the theoretical framework within which a measurement was made does actually lead to a change in  $K$ . Equation (2.12) formalizes the notion of quiddity *qua* essence, comprising substance ( $K$ ) and properties ( $I$ ). The calculation of  $K$  will be dealt with in more detail in Chapter 6. As a final remark in this section, we note that the results of an experiment or observation transmitted elsewhere may have the same effect on the recipient as if he had carried out the experiment himself.

**Problem.** Critically scrutinize Fig. 2.1 in the light of the above discussion and attempt to quantify the information flows.

## 2.2 Constraint

Shannon puts emphasis on the information resulting from the selection from a set of possible alternatives (implying the existence of alternatives)—information can only be received where there is doubt. Much of the theory of information deals with *signals*, which operate on the set of alternatives constituting the recipient’s doubt to yield a lesser doubt, or even certainty (zero doubt). Thus, the signals themselves have an information content by virtue of their potential for making selections; the quantity of information corresponds to the intensity of selection or to the recipient’s

---

<sup>5</sup> Called the physical information of a system by Zurek.

surprise upon receiving the information.  $I$  from equation (2.5) gives the average information content per symbol; it is a weighted mean of the degree of uncertainty (i.e., freedom of choice) in choosing a symbol before any choice is made.

If we are writing a piece of prose, and even more so if it is verse, our freedom of choice of letters is considerably constrained; for example, the probability that “x” follows “g” in an English text is much lower than  $\frac{1}{26}$  (or  $\frac{1}{27}$  if we include, as we should, the space as a symbol). In other words, the selection of a particular letter depends on the preceding symbol, or group of preceding symbols. This problem in linguistics was first investigated by Markov, who encoded a poem of Pushkin’s using a binary coding scheme admitting consonants (C) or vowels (V). Markov proposed that the selection of successive symbols C or V no longer depended on their probabilities as determined by their frequencies ( $v = V/(V + C)$ , where  $V$  and  $C$  are respectively the total numbers of vowels and consonants). To every pair of letters  $(L_j, L_k)$  there corresponds a conditional probability  $p_{jk}$ ; given that  $L_j$  has occurred, the probability of  $L_k$  at the next selection is  $p_{jk}$ . If the initial letter has a probability  $a_j$ , then the probability of the sequence  $(L_j, L_k, L_l) = a_j p_{jk} p_{kl}$  and so forth. The scheme can be conveniently written in matrix notation:

$$\begin{array}{c|cc} \rightarrow & C & V \\ \hline C & p_{cc} & p_{cv} \\ V & p_{vc} & p_{vv} \end{array} \quad (2.14)$$

where  $p_{cc}$  means the probability that a consonant is followed by another consonant, and similarly for the other terms. The matrix is stochastic; that is, the rows must add up to 1. If every column is identical, then there is no dependence on the preceding symbol, and we revert to a random, or zeroth-order Markov, process. Suppose now that observation reveals that the probability of C occurring after V preceded by C is different from that of C occurring after V preceded by V, or even that the probability of C occurring after VV preceded by C is different from that of C occurring after VV preceded by V. These higher-order Markov processes can be recoded in strict Markov form, thus for the second-order process (dependency of the probabilities on the two preceding symbols) “VVC” can be written as a transition from VV to VC, and hence the matrix of transition probabilities becomes

$$\begin{array}{c|cccc} \rightarrow & CC & CV & VC & VV \\ \hline CC & p_{ccc} & p_{ccv} & 0 & 0 \\ CV & 0 & 0 & p_{cvc} & p_{cvv} \\ VC & p_{vcc} & p_{vcv} & 0 & 0 \\ VV & 0 & 0 & p_{vvc} & p_{vvv} \end{array} \quad (2.15)$$

and so on for higher orders. Notice that some transitions necessarily have zero probability.<sup>6</sup>

---

<sup>6</sup> See also §6.2.

The reader may object that one rarely composes text letter by letter, but rather word by word. Clearly, there are strong constraints governing the succession of words in a text. The frequencies of these successions can be obtained by counting word occurrences in very long text and then used to construct the transition matrix, which is, of course, gigantic even for a first-order process. We may also note that a book ending with "...in the solid state is greatly aided by this new tool." is more likely to begin with "Rocket motor design received a considerable boost when ..." than one ending "I became submerged in my thoughts which sparkled with a cold light."<sup>7</sup>

We note here that clearly one may attempt to model DNA or protein sequences as Markov processes, as will be discussed in Part III. Markov chains as such will be discussed more fully in Chapter 6.

The notion of constraint applies whenever a set "is smaller than it might be." The classic example is that of road traffic lights, which display various combinations of red, amber, and green, each of which may be on or off. Although  $2^3 = 8$  combinations are theoretically possible, in most countries only certain combinations are used, typically only four out of the eight. Constraints are ubiquitous in the universe and much of science consists in determining them; thus, in a sense, "constraint" is synonymous with "regularity." Laws of nature are clearly constraints, and the very existence of physical objects such as tables and aeroplanes, which have fewer degrees of freedom than their constituent parts considered separately, is a manifestation of constraint.

In this book we are particularly concerned with constraints applied to sequences. Clearly, if a Markov process is in operation, the variety of the set of possible sequences generated from a particular alphabet is smaller than it would be had successive symbols been freely selected; that is, it is indeed "smaller than it might have been." "Might have been" requires the qualification, then, of "would have been if successive symbols had been freely (or randomly—leaving the discussion of "randomness" to Chapter 6) selected." We already know how to calculate the entropy (or information, or Shannon index, or Shannon-Weaver index)  $I$  of a random sequence (equation 2.5); there is a precise way of calculating the entropy per symbol for a Markov process (see §6.2), and the reader may use the formula derived there to verify that the entropy of a Markov process is less than that of a "perfectly random" process. Using some of the terminology already introduced, we may expand on this statement to say that the surprise occasioned by receiving a piece of information is lower if constraint is operating; for example, when spelling out a word, it is practically superfluous to say "u" after "q."

The constraints affecting the choice of successive words are a manifestation of the syntax of a language. In the next chapter other ways in which constraint can operate will be examined, but for now we can simply state that whenever constraint is present, the entropy (of the set we are considering, hence of the information received

---

<sup>7</sup> Good (1969) has shown that ordinary language cannot be represented even by a Markov process of infinite order.

by selecting a member of that set) is lower than it would be for a perfectly random selection from that set.

This maximum entropy (which, in physical systems, corresponds to the most probable arrangement; i.e., to the macroscopic state that can be arranged in the largest number of ways)—let us call it  $I_{\max}$ —allows us to define a relative entropy  $I_{\text{rel}}$ ,

$$I_{\text{rel}} = \frac{\text{actual entropy}}{I_{\max}}, \quad (2.16)$$

and a redundancy  $R$ ,

$$R = 1 - I_{\text{rel}}. \quad (2.17)$$

In a fascinating piece of work, Shannon (1951) established the entropy of English essentially through empirical investigations using rooms full of people trying to guess incomplete texts.<sup>8</sup>

More formally, the relative entropy (Kullback-Leibler distance)<sup>9</sup> between two (discrete) distributions with probability functions  $a_k$  and  $b_k$  is

$$\mathcal{R}(a, b) = \sum_k a_k \log_2(a_k/b_k). \quad (2.18)$$

If  $a_k$  is an actual distribution of observations, and  $b_k$  is a model description approximating to the data,<sup>10</sup> then  $\mathcal{R}(a, b)$  is the expected difference (expressed as the number of bits) between encoding samples from  $a_k$  using a code based on  $a$  and using a code based on  $b$ . This can be seen by writing equation (2.18) as

$$\mathcal{R}(a, b) = - \sum_k b_k \log_2 a_k + \sum_k a_k \log_2 a_k, \quad (2.19)$$

where the first term on the right-hand side is called the cross-entropy of  $a_k$  and  $b_k$ , the expected number of bits required to encode observations from  $a$  when using a code based on  $b$  rather than  $a$ . Conversely,  $\mathcal{R}(a, b)$  is the gain in information if a code based on  $a$  rather than  $b$  is used.

Suppose that  $P\{x_1, x_2, \dots, x_m\}$  is the probability of having a certain pattern (arrangement), or  $m$ -gram  $x_1, x_2, \dots, x_m$ ,<sup>11</sup> assumed to be ergodic (stationary

<sup>8</sup> Note that most computer languages lack redundancy—a single wrong character in a program will usually cause the program to halt, or not compile.

<sup>9</sup> Since  $\mathcal{R}(a, b) \neq \mathcal{R}(b, a)$ , it is not a true metric and is therefore sometimes called “divergence” rather than “distance.”

<sup>10</sup> Possibly constructed *a priori*.

<sup>11</sup> See also §8.2.

stochastic).<sup>12</sup> These could be the English texts studied by Shannon; of particular relevance to the topic of this book is the problem of predicting the nucleic acid base following a known (sequenced) arrangement. The conditional probability<sup>13</sup> that the pattern  $[(m-1)\text{-gram}] x_1, x_2, \dots, x_{m-1}$  is followed by the symbol  $x_m$  is

$$P\{x_m | x_1, x_2, \dots, x_{m-1}\} = \frac{P\{x_1, x_2, \dots, x_{m-1}, x_m\}}{P\{x_1, x_2, \dots, x_{m-1}\}}. \quad (2.20)$$

The “ $m$ -length approximation” to the entropy  $S_m$ , defined as the average uncertainty about the next symbol, is

$$S_m = - \sum_{x_1, x_2, \dots, x_{m-1}} P\{x_1, x_2, \dots, x_{m-1}\} \times \sum_x P\{x_m | x_1, x_2, \dots, x_{m-1}\} \log P\{x_m | x_1, x_2, \dots, x_{m-1}\}. \quad (2.21)$$

It includes all possible correlations up to length  $m$ . Note that the first sum on the right-hand side is taken over all possible preceding sequences, and the second sum is taken over all possible symbols. The *correlation information* is defined as

$$k_m = S_{m-1} - S_m \quad (m \geq 2). \quad (2.22)$$

$S_1$  is simply the Shannon information (equation 2.5). If the probability of the different symbols is *a priori* equal, then the information is given by Hartley’s formula (2.4).<sup>14</sup> For  $m = 1$ ,

$$k_1 = \log n - S_1 \quad (2.23)$$

is known as the *density information*. By recursion we can then write

$$\mathcal{I} = S + \sum_{m=1}^{\infty} k_m \quad (2.24)$$

the total information  $\mathcal{I}$  being equal to  $\log n$ . The first term on the right gives the random component and is defined as  $S = \lim_{m \rightarrow \infty} S_m$ , and the second one gives the redundancy. For a binary string,  $S = 1$  if it is random, and the redundancy equals zero. For a regular string like  $\dots 010101 \dots$ ,  $S = 0$  and  $k_2 = 1$ ; for a first order Markov chain  $k_m = 0$  for all  $m > 2$ .

---

<sup>12</sup> See §6.1.

<sup>13</sup> See §5.2.2.

<sup>14</sup> The effective measure complexity is the weighted sum of the  $k_m$  [viz.,  $\sum_{m=2}^{\infty} (m-1)k_m$ ]—see equation (6.27).

### 2.2.1 The Value of Information

In order to quantify value  $V$ , we need to know the goal toward which the information will be used. Two cases may be considered:

(i) The goal can almost certainly be reached by some means or another. In this case a reasonable quantification is

$$V = (\text{cost or time required to reach goal without the information}) - (\text{cost or time required to reach goal with the information}) . \quad (2.25)$$

(ii) The probability of reaching the goal is low. Then it is more reasonable to adopt

$$V = \log_2 \frac{\text{prob. of reaching goal with the information}}{\text{prob. of reaching goal without the information}} . \quad (2.26)$$

With both of these measures, irrelevant information is clearly zero-valued.

Durability of information contributes to its value. Intuitively, we have the idea that the more important the information, the longer it is preserved. In antiquity, accounts of major events such as military victories were preserved in massive stone monuments whose inscriptions can still be read today several thousand years later. Military secrets are printed on paper or photographed using silver halide film and stored in bunkers, rather than committed to magnetic media. We tend to write down things we need to remember for a long time.

The value of information is closely related to the problem of weighing the credibility that one should accord a certain received piece of information. The question of weighting scientific data from a series of measurements was an important driver for the development of probability theory. In 1777, Daniel Bernoulli raised this issue in the context of averaging astronomical data, where it was customary to simply reject data deviating too far from the mean and weight all others equally.<sup>15</sup> Bennett has proposed that his notion of logical depth (§6.5) provides a formal measure of value, very much in the spirit of the preceding two equations proposed by Chernavsky. A sequence of coin tosses formally contains much information that has little value; a table giving the positions of the planets every day for several centuries hence contains no more information than the equations of motion and initial conditions from which it was deduced, but saves anyone consulting it the effort of calculating the positions. This suggests that the value of a message resides not in its information per se (i.e., its absolutely unpredictable parts) nor in any obvious redundancy (e.g., repetition), but rather in what Bennett has suggested be called buried redundancy:

---

<sup>15</sup> D. Bernoulli, *Diudicatio maxime probabilis plurium observationem discrepantium atque verisimillima inductio inde formanda*. *Acta Acad. Sci. Imp. Petrop.* 1 (1777) 3–23. See also L. Euler, *Observationes in praecedentem dissertationem illustris Bernoulli*. *Acta Acad. Sci. Imp. Petrop.* 1 (1777) 24–33.

parts predictable only with considerable effort on the part of the recipient of the message. This effort corresponds to logical depth.

The value of information is also related to the amount already possessed. The same Bernoulli asserted that the value (utility in economic parlance) of an amount  $m$  of money received is proportional to  $\log[(m + c)/c]$ , where  $c$  is the amount of money already possessed, and a similar relationship may apply to information.

### ***2.2.2 The Quality of Information***

Quality is an attribute that brings us back to the problem posed by Bernoulli in 1777, namely how to weight observations. If we return to our simple measurement of the length of a piece of wood, the reliability may be affected by the physical condition of the measuring stick, its markings, its origin (e.g., from a kindergarten or from Sèvres), the eyesight of the measurer, and so forth.

## **2.3 Accuracy, Meaning, and Effect**

### ***2.3.1 Accuracy***

In the preceding sections, we have focused on the information gained when a certain signal, or sequence of signals, is received. The quantity of this information  $I$  has been formalized according to its statistical properties.  $I$  is of particular relevance when considering how accurately a certain sequence of symbols can be transmitted. This question will be considered in more detail in Chapter 3. For now, let us merely note that no physical device can discriminate between pieces of information differing by arbitrarily small amounts. In the case of a photographic detector, for example, diminishing the difference will require larger and larger detectors in order to discriminate, but photon noise places an ultimate limitation in the way of achieving arbitrarily small detection.

A communication system depending on setting the position of a pointer on a dial to 1 of 6000 positions and letting the position be observed by the distant recipient of the message through a telescope, while allowing a comfortably large range of signs to be transmitted, would be hopelessly prone to reading errors, and it was long ago realized that far more reliable communication could be achieved by using a small number of unambiguously uninterpretable signs (e.g., signalling flags at sea) that could be combined to generate complex messages.

Practical information space is thus normally discrete; for example, meteorological bulletins do not generally give the actual wind speed in kilometres per hour and the direction in degrees, but refer to 1 of the 13 points of the Beaufort scale and 1 of the 8 compass points. The information space is therefore a finite 2-space with  $8 \times 13$  elements.

The rule for determining the distance between two words (i.e., the metric of information space) is most conveniently perceived if the words are encoded in binary form. The Hamming distance is the number of digit places in which the two words differ.<sup>16</sup> This metric satisfies the usual rules for distance; that is, if  $a$ ,  $b$ , and  $c$  are three points in the space and  $D(a, b)$  is the distance between  $a$  and  $b$ , then

$$\begin{aligned} D(a, a) &= 0 ; \\ D(a, b) &= D(b, a) > 0 \quad \text{if } b \neq a ; \\ D(a, b) + D(b, c) &\geq D(a, c) . \end{aligned}$$

In biology, the question of accuracy refers especially to the replication of DNA, its transcription into RNA, and the translation of RNA into protein. It may also refer to the accuracy with which physiological signals can be transmitted within and between cells.

### 2.3.2 *Meaning*

At the first level, Shannon's theory is deliberately divorced from the question of semantic content (i.e., meaning). In the simple example of measuring the length of a piece of wood, the question of meaning scarcely enters into the discourse. In nearly all of the other cases, where we are concerned with receiving signs, or sequences of symbols, after we have received them accurately we can start to concern ourselves with the question of meaning. The issues can range from simple ones of interpretation to involved and complex ones. An example of the former is the interpretation of the order "Wait!" heard in a workshop. It may indeed mean "pause until further notice," but heard by an apprentice standing by a weighing machine, may well be interpreted as "call out the weight of the object on the weighing pan." An example of the latter is the statement "John Smith is departing for Paris," which has very different connotations according to whether it was made in an airport, a railway station or some other place.

It is easy to show that the meaning contained in a message depends on the set of possible messages. Ashby has constructed the following example. Suppose a prisoner of war is allowed to send a message to his family. In one camp, the message can be chosen from the following set:

I am well  
I am quite well  
I am not well  
I am still alive,

and in another camp, only one message may be sent:

I am well.

---

<sup>16</sup> Cf. J.E. Surrick and L.M. Conant, *Laddergrams*, New York: Sears (1927). "Turn bell into ring in six moves" and so forth.



In both cases, there is implicitly a further alternative—no message at all, which would mean that the prisoner is dying or already dead. In the second camp, if the recipient is aware that only one message is permitted, he or she will know that it encompasses several alternatives, which are explicitly available in the first camp. Therefore, the same message (I am well) can mean different things depending on the set from which it is drawn.

In much human communication, it is the context-dependent difference between explicit and implicit meaning that is decisive in determining the ultimate outcome of the reception of information. In the latter example of the previous paragraph, the context—here provided by the physical environment—endows the statement with a large complement of implicit information, which mostly depends on the mental baggage possessed by the recipient of the information; for example, the meaning of a Chinese poem may only be understandable to someone who has assimilated Chinese history and literature since childhood, and it will not as a rule be intelligible to a foreigner armed with a dictionary.

A very similar state of affairs is present in the living cell. A given sequence of DNA will have a well-defined explicit meaning in terms of the sequence of amino acids it encodes, and into which it can be translated. In the eukaryotic cell, however, that amino acid sequence may then be glycosylated and further transformed, but in a bacterium, it may not be; indeed it may even misfold and aggregate—a concrete example of implicit meaning dependent on concept.

The importance of context in determining implicit meaning is even more graphically illustrated in the case of the developing multicellular organism, in which the cells are initially all identical, but according to chemical signals received from their environment, it will develop into different kinds of cells. The meaning of the genotype is the phenotype, and it is implicit rather than explicit meaning, which is, of course, why the DNA sequence of any earthly organism sent to an alien civilization will not allow them to reconstruct the organism. Ultimately, most of the cells in the developing embryo become irreversibly different from each other (differentiation), but while they are still pluripotent, they may be transplanted into regions of different chemical composition and change their fate; for example, a cell from the non-neurogenic region of one embryo transplanted into the neurogenic region of another may become a neuroblast. The mechanism of such transformations will be discussed in a little more detail in Chapter 10, but here this type of phenomenon serves to illustrate how the implicit meaning of the genome dominates the explicit meaning. This implicit meaning is called *epigenetics*, and it seems clear that we will not truly understand life before we have developed a powerful way of treating epigenetic phenomena. Shannon's approach has proved very powerful for treating the problem of the accurate transmission of signals, but at present we do not have a comparable foundation for treating the problem of the precise transfer of meaning.

Even at the molecular level, at which phenotype is more circumscribed and could be considered to be the function (of an enzyme), or simply the structure of a protein, there is presently little understanding of the relation between sequence and function, as illustrated by the thousands of known different sequences encoding the same type

of structure and function, or different sequences encoding different structures but the same type of function, or similar structures with different functions.

Part of the difficulty is that the function (i.e., biological meaning) is not so conveniently quantifiable as the information content of the sequence encoding it. Even considering the simpler problem of structure alone, there are various approaches yielding very different answers. Supposing that a certain protein has a unique structure [most nonstructural proteins have, of course, several (at least two) structures in order to function; the best-known example is probably haemoglobin]. This structure could be specified by the coordinates of all the constituent atoms, or the dihedral angles of each amino acid, listed in order of the sequence, and at a given resolution [Dewey calls this the algorithmic complexity of a protein; cf.  $K$  in equation (2.12)]. If, however, protein structures come from a finite number of basic types, it suffices to specify one of these types, which moves the problem back into one dealing with Shannon-type information.

In the case of function, a useful starting point could be to consider the immune system, in which the main criterion of function is the affinity of the antibody (or, more precisely, the affinity of a small region of the antibody) to the target antigen. The discussion of affinity and how affinities can lead to networks of interactions will be dealt with in Chapter 15.

The problem of assigning meaning to a sign, or a message (a collection of signs), is usually referred to as the semantic problem. Semantic information cannot be interpreted solely at the syntactical level.

Just as a set of antibodies can be ranked in order of affinity, so may a series of statements be ranked in order of semantic precision; for example, consider the statements:

A train will leave.

A train will leave from London today.

An express train will leave from London Marylebone for Glasgow at 10 a.m. today.

and so on. Postal or e-mail addresses have a similar kind of syntactical hierarchy. Although we are not yet able to assign numerical values to meanings, we can at least order them.

Carnap and Bar-Hillel have framed a theory, rooted in Carnap's theory of inductive probability, attempting to do for semantics what Shannon did for the technical content of a message. It deals with the semantic content of declarative sentences, excluding the pragmatic aspects (dealing with the consequences or value of received information for the recipient). It does not deal with the so-called semantic problem of communication, which is concerned with the identity (or approach thereto) between the intended meaning of the sender and the interpretation of meaning by the receiver: Carnap and Bar-Hillel place this explicit involvement of sender and receiver in the realm of pragmatics.

To gain a flavour of their approach, note that the semantic content of sentence  $j$ , conditional on having heard sentence  $i$ , is  $\text{content}(j|i) = \text{content}(i \& j) - \text{content}(i)$ , and their measure of information is defined as  $\text{information}(i) = -\log_2 \text{content}(\text{NOT } i)$ . They consider semantic noise (resulting in misinterpretation

of a message, even though all of its individual elements have been perfectly received) and semantic efficiency, which takes experience into account; for example, a language with the predicates W, M, and C, designating respectively warm, moderate, and cold temperatures, would be efficient in a continental climate (e.g., Switzerland or Hungary) but would become inefficient with a move to the western margin of Europe, since M occurs much more frequently there.

Although the quantification of information is deliberately abstracted from the content of a message, taking content into account may allow much more dramatic compression of a message than is possible using solely the statistical redundancy (equation 2.17). Consider how words such as “utilization” may be replaced by “use,” appellations such as “guidance counsellor” by “counsellor,” and phrases such as “at this moment in time” by “at this moment,” or simply “now.” Many documents can be thus reduced in length by over two-thirds without any loss in meaning (but a considerable gain in readability). With simply constructed texts, algorithmic procedures for accomplishing this that do not require the text to be interpreted can be devised; for example, all the words in the text can be counted and listed in order of frequency of occurrence, and then each sentence is assigned a score according to the numbers of the highest-ranking words (apart from “and,” “that,” etc.) it contains. The sentences with the highest scores are preferentially retained.

### 2.3.3 *Effect*

A signal may be accurately received and its meaning may be understood by the recipient, but that does not guarantee that it will engender the response desired by the sender. This aspect of information deals with the ultimate result and the far-reaching consequences of a message and how the deduced meaning is related to human purposes. The question of the value of information has already been discussed (§2.2.1), and operationally it comes close to a quantification of effect.

Mackay has proposed that the quantum of effective information is that amount that enables the recipient to make one alteration to the logical pattern describing his awareness of the relevant situation, and this would appear to provide a good basis for quantifying effect. Suppose that an agent has a state of mind  $M_1$ , which comprises certain beliefs, hypotheses, and the like (the prior state). The agent then hears a sentence, which causes a change to state of mind  $M_2$ , the posterior state, which stands in readiness to make a response. If the meaning of an item of information is its contribution to the agent's total state of conditional readiness for action and the planning of action (i.e., the agent's conditional repertoire of action), then the effect is the ultimate realization of that conditional readiness in terms of actual action.<sup>17</sup>

As soon as we introduce the notion of a conditional repertoire of action, we see that selection must be considered. Indeed, the three essential attributes of an agent are (and note the parallel with the symbolic level) as follows:

---

<sup>17</sup> Wiener subsumes effect into meaning in his definition of “meaningful information.”

1. A repertoire, from which alternative actions can be selected;
2. An evaluator, which assigns values to different states of affairs according to either given or self-set criteria;
3. A selector, which selects actions increasing a positive evaluation and diminishing deleterious evaluation.

One may compare this procedure with that of evolutionary computation (§8.1), and, *a fortiori*, with that of evolution itself. Here, the selected actions are used to build up a presence in the repertoire (and, assuming that the repertoire remains constant in size, unselected actions will be diminished).

### 2.3.4 Significs

As summarized by Welby, significs comprises (a) sense (“in what sense is a word used?”), (b) meaning (the specific sense a word is intended to convey), and (c) significance—the far-reaching consequence, implication, ultimate result, or outcome (e.g., of some event or experience). It therefore includes semantics but goes well beyond it. Given that significs has perhaps been somewhat eclipsed by semiotics, the way would be clear to develop the significs of *n*-grams of DNA and of peptides (regulatory oligopeptides and proteins).

## 2.4 Further Remarks on Information Generation

The exercise of intellect involves both the transformation and generation of information, the latter quite possibly involving the crossing of some kind of logical gap. It is a moot point whether the solution of a set of equations contains more information than the equations, since the solution is implicit (and J.S. Mill insisted that induction, not deduction, is the only road to new knowledge). If it does not, are we then no more complex than a zygote, which apparently contains all the information required to generate a functional adult?

The reception of information is equivalent to ordering (i.e., an entropy decrease) and corresponds to the various ordering phenomena seen in nature. Three categories can be distinguished:

1. Order from disorder (sometimes called “self-organization”<sup>18</sup>);
2. Order from order (a process based on templating, such as DNA replication or transcription);
3. Order from noise (microscopic information is given macroscopic expression).

The only meaningful way of interpreting the first category is to suppose that the order was implicit in the initial state; hence, it is questionable whether information

---

<sup>18</sup> However, see the critiques of von Foerster and of Ashby (1962).

has actually been generated. In the second category, the volume of ordering has increased, but at the expense of more disorder elsewhere, because of the physical exigencies of the copying process. Note that copying per se does not lead to an increase in the amount of information. The third category is of genuine interest, for it illuminates problems such as that of the development of the zygote, in which environmental information is given meaningful macroscopic expression.

**Problem.** Particularly examine the proposition that the production and dissemination of copies of a document reporting new facts does not increase the amount of information.

## 2.5 Summary

Information is that which removes uncertainty. It has two aspects: form (what we already know about the system) and content, the result of an operation (e.g., a measurement) carried out within the framework of our extant knowledge. Form specifies the structure of the information. This includes the specification of the set of possible messages that we can receive or the (design and fabrication of the) instrument used to measure a parameter of the system. It can be quantified as the length of the shortest algorithm able to specify the system (Kolmogorov information). If we know the set from which the result of the measurement operation has to come, the (metrical) content of the operation is given by the Shannon index (reducing to the Hartley index if the choices are equiprobable). A message (e.g., a succession of symbols) that directs our selection is, upon receipt, essentially equivalent to the result of the measurement operation encoded by the message. The Shannon index assumes that the message is known with certainty once it has been received; if it is not, the Wiener index should be used.

Information can be represented as a sign or as a succession of signs (symbols). The information conveyed by each symbol equals the freedom in choosing the symbol. If all choices are *a priori* equiprobable, the specification of a sequence removes uncertainty maximally. In practice, there may be strong syntactical constraints imposed on the successive choices, which limit the possible variety in a sequence of symbols.

In order to be considered valuable (or desired), the received information must be remembered (macroscopic information). Microinformation is not remembered. Thus, the information inherent in the positions and momenta of all the gas molecules in a room is forgotten picoseconds after its reception. It is of no value.

Information can be divided into three aspects: the signs themselves, their syntax (their relation with each other), and the accuracy with which they can be transmitted; their meaning, or semantic value (i.e., their relation to designata); and their effect (how effectively the received meaning affects the conduct of the recipient in the desired way), which may be called pragmatics, the study of signs in relation to their

users, or signifiés, the study of significance.<sup>19</sup> In other words, content comprises the signs themselves and their syntax (i.e., the relation between them), their meaning (semantic value), and their effect on the conduct of the recipient (i.e., does it lead to action?). A further aspect is that of style, very difficult to quantify. It can be considered to be determined by word usage frequencies, from which the cybernetic temperature can be derived (cf. equation (3.7)). An indication (cf. biomarkers of disease) might be given by the occurrence of certain characteristic words, including the use of a certain synonym rather than another. If a symbolic sequence is modelled as a Markov chain, matters of style would be encapsulated in hidden Markov models (q.v.).

Meaning may be highly context-dependent; the stronger this dependence, the more implicit the meaning.

The effect of receipt of information on behaviour can be quantified in terms of changes to the logical pattern describing the awareness of the recipient to his environment. In simpler terms, this may be quantified as value in terms of a change in behaviour (assuming that enough data on replicate systems or past events are available to enable the course of action that would have taken place in the absence of the received information to be determined).

Information is inherently discrete (quantal) and thus based on combinatorics, which also happens to suit the spirit of the digital computer. In biology, if “genotype” constitutes the signs, then “phenotype” constitutes meaning. Action is self-explanatory and linked to adaptation (see §9.2). Biological function might be considered to be the potential for action.

---

<sup>19</sup> The three aspects of syntactics, semantics, and pragmatics are usually considered to constitute the theory of signs, or semiotics.



<http://www.springer.com/978-1-84800-256-2>

Bioinformatics

An Introduction

Ramsden, J.

2009, XVI, 272 p. 29 illus., Hardcover

ISBN: 978-1-84800-256-2