Information theory

Information theory, a mathematical representation of the conditions and parameters affecting the transmission and processing of information. Most closely associated with the work of the American electrical engineer Claude Shannon in the mid-20th century, information theory is chiefly of interest to communication engineers, though some of the concepts have been adopted and used in such fields as psychology and linguistics. Information theory overlaps heavily with communication theory, but it is more oriented toward the fundamental limitations on the processing and communication of information and less oriented toward the detailed operation of particular devices.

TABLE OF CONTENTS

Historical Background

Interest in the concept of information grew directly from the creation of the telegraph and telephone. In 1844 the American inventor Samuel F.B. Morse built a telegraph line between Washington, D.C., and Baltimore, Maryland. Morse encountered many electrical problems when he sent signals through buried transmission lines, but inexplicably he encountered fewer problems when the lines were suspended on poles. This attracted the attention of many distinguished physicists, most notably the Scotsman William Thomson (Baron Kelvin). In a similar manner, the invention of the telephone in 1875 by Alexander Graham Bell and its subsequent proliferation attracted further scientific notaries, such as Henri Poincaré, Oliver Heaviside, and Michael Pupin, to the problems associated with transmitting signals over wires. Much of their work was done using Fourier analysis, a technique described later in this article, but in all of these cases the analysis was dedicated to solving the practical engineering problems of communication systems.

The formal study of information theory did not begin until 1924, when Harry Nyquist, a researcher at Bell Laboratories, published a paper entitled “Certain Factors Affecting Telegraph Speed.” Nyquist realized that communication channels had maximum data transmission rates, and he derived a formula for calculating these rates in finite bandwidth noiseless channels. Another pioneer was Nyquist’s colleague R.V.L. Hartley, whose paper “Transmission of Information” (1928) established the first mathematical foundations for information theory.

The real birth of modern information theory can be traced to the publication in 1948 of Claude Shannon's "A Mathematical Theory of Communication" in the *Bell System Technical Journal*. A key step in Shannon's work was his realization that, in order to have a theory, communication signals must be treated in isolation from the meaning of the messages that they transmit. This view is in sharp contrast with the common conception of information, in which meaning has an essential role. Shannon also realized that the amount of knowledge conveyed by a signal is not directly related to the size of the message. A famous illustration of this distinction is the correspondence between French novelist Victor Hugo and his publisher following the publication of *Les Misérables* in 1862. Hugo sent his publisher a card with just the symbol "?". In return he received a card with just the symbol "!". Within the context of Hugo's relations with his publisher and the public, these short messages were loaded with meaning; lacking such a context, these messages are meaningless. Similarly, a long, complete message in perfect French would convey little useful knowledge to someone who could understand only English.

Shannon thus wisely realized that a useful theory of information would first have to concentrate on the problems associated with sending and receiving messages, and it would have to leave questions involving any intrinsic meaning of a message—known as the semantic problem—for later investigators. Clearly, if the technical problem could not be solved—that is, if a message could not be transmitted correctly—then the semantic problem was not likely ever to be solved satisfactorily. Solving the technical problem was therefore the first step in developing a reliable communication system.

It is no accident that Shannon worked for Bell Laboratories. The practical stimuli for his work were the problems faced in creating a reliable telephone system. A key question that had to be answered in the early days of telecommunication was how best to maximize the physical plant—in particular, how to transmit the maximum number of telephone conversations over existing cables. Prior to Shannon's work, the factors for achieving maximum utilization were not clearly understood. Shannon's work defined communication channels and showed how to assign a capacity to them, not only in the theoretical sense where no interference, or noise, was present but also in practical cases where real channels were subjected to real noise. Shannon produced a formula that showed how the bandwidth of a channel (that is, its theoretical signal capacity) and its signal-to-noise ratio (a measure of interference) affected its capacity to carry signals. In doing so he was able to suggest strategies for maximizing the capacity of a given channel and showed the limits of what was possible with a given technology. This was of great utility to engineers, who could focus thereafter on individual cases and understand the specific trade-offs involved.

Shannon also made the startling discovery that, even in the presence of noise, it is

always possible to transmit signals arbitrarily close to the theoretical channel capacity. This discovery inspired engineers to look for practical techniques to improve performance in signal transmissions that were far from optimal. Shannon's work clearly distinguished between gains that could be realized by adopting a different encoding scheme from gains that could be realized only by altering the communication system itself. Before Shannon, engineers lacked a systematic way of analyzing and solving such problems.

Shannon's pioneering work thus presented many key ideas that have guided engineers and scientists ever since. Though information theory does not always make clear exactly how to achieve specific results, people now know which questions are worth asking and can focus on areas that will yield the highest return. They also know which sorts of questions are difficult to answer and the areas in which there is not likely to be a large return for the amount of effort expended.

Since the 1940s and '50s the principles of classical information theory have been applied to many fields. The section Applications of information theory surveys achievements not only in such areas of telecommunications as data compression and error correction but also in the separate disciplines of physiology, linguistics, and physics. Indeed, even in Shannon's day many books and articles appeared that discussed the relationship between information theory and areas such as art and business. Unfortunately, many of these purported relationships were of dubious worth. Efforts to link information theory to every problem and every area were disturbing enough to Shannon himself that in a 1956 editorial titled "The Bandwagon" he issued the following warning:

> *I personally believe that many of the concepts of information theory will prove useful in these other fields—and, indeed, some results are already quite promising—but the establishing of such applications is not a trivial matter of translating words to a new domain, but rather the slow tedious process of hypothesis and experimental verification.*

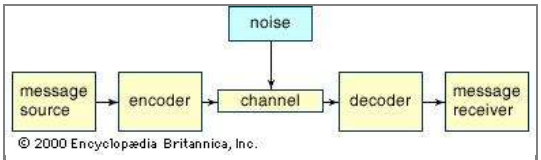
With Shannon's own words in mind, we can now review the central principles of classical information theory.

Classical Information Theory

Shannon's communication model

As the underpinning of his theory, Shannon developed a very simple, abstract model of communication, as shown in the figure. Because his model is abstract, it applies in

many situations, which contributes to its broad scope and power.



Shannon's communication modelConsider a simple telephone conversation: A person (message source) speaks into a telephone receiver (encoder), which converts the sound of the spoken word into an electrical signal. This electrical signal is then transmitted over telephone lines (channel) subject to interference (noise). When the signal reaches the telephone receiver (decoder) at the other end of the line it is converted back into vocal sounds. Finally, the recipient (message receiver) hears the original message.
Encyclopædia Britannica, Inc.

The first component of the model, the message source, is simply the entity that originally creates the message. Often the message source is a human, but in Shannon's model it could also be an animal, a computer, or some other inanimate object. The encoder is the object that connects the message to the actual physical signals that are being sent. For example, there are several ways to apply this model to two people having a telephone conversation. On one level, the actual speech produced by one person can be considered the message, and the telephone mouthpiece and its associated electronics can be considered the encoder, which converts the speech into electrical signals that travel along the telephone network. Alternatively, one can consider the speaker's mind as the message source and the combination of the speaker's brain, vocal system, and telephone mouthpiece as the encoder. However, the inclusion of "mind" introduces complex semantic problems to any analysis and is generally avoided except for the application of information theory to physiology.

The channel is the medium that carries the message. The channel might be wires, the air or space in the case of radio and television transmissions, or fibre-optic cable. In the case of a signal produced simply by banging on the plumbing, the channel might be the pipe that receives the blow. The beauty of having an abstract model is that it permits the inclusion of a wide variety of channels. Some of the constraints imposed by channels on the propagation of signals through them will be discussed later.

Noise is anything that interferes with the transmission of a signal. In telephone conversations interference might be caused by static in the line, cross talk from another line, or background sounds. Signals transmitted optically through the air might suffer interference from clouds or excessive humidity. Clearly, sources of noise depend upon the particular communication system. A single system may have several sources of noise, but, if all of these separate sources are understood, it will sometimes be possible to treat them as a single source.

The decoder is the object that converts the signal, as received, into a form that the message receiver can comprehend. In the case of the telephone, the decoder could be the earpiece and its electronic circuits. Depending upon perspective, the decoder could

also include the listener's entire hearing system.

The message receiver is the object that gets the message. It could be a person, an animal, or a computer or some other inanimate object.

Shannon's theory deals primarily with the encoder, channel, noise source, and decoder. As noted above, the focus of the theory is on signals and how they can be transmitted accurately and efficiently; questions of meaning are avoided as much as possible.

Four types of communication

There are two fundamentally different ways to transmit messages: via discrete signals and via continuous signals. Discrete signals can represent only a finite number of different, recognizable states. For example, the letters of the English alphabet are commonly thought of as discrete signals. Continuous signals, also known as analog signals, are commonly used to transmit quantities that can vary over an infinite set of values—sound is a typical example. However, such continuous quantities can be approximated by discrete signals—for instance, on a digital compact disc or through a digital telecommunication system—by increasing the number of distinct discrete values available until any inaccuracy in the description falls below the level of perception or interest.

Communication can also take place in the presence or absence of noise. These conditions are referred to as noisy or noiseless communication, respectively.

All told, there are four cases to consider: discrete, noiseless communication; discrete, noisy communication; continuous, noiseless communication; and continuous, noisy communication. It is easier to analyze the discrete cases than the continuous cases; likewise, the noiseless cases are simpler than the noisy cases. Therefore, the discrete, noiseless case will be considered first in some detail, followed by an indication of how the other cases differ.

Discrete, noiseless communication and the concept of entropy

From message alphabet to signal alphabet

As mentioned above, the English alphabet is a discrete communication system. It consists of a finite set of characters, such as uppercase and lowercase letters, digits, and various punctuation marks. Messages are composed by stringing these individual characters together appropriately. (Henceforth, signal components in any discrete communication system will be referred to as characters.)

For noiseless communications, the decoder at the receiving end receives exactly the characters sent by the encoder. However, these transmitted characters are typically not

in the original message's alphabet. For example, in Morse Code appropriately spaced short and long electrical pulses, light flashes, or sounds are used to transmit the message. Similarly today, many forms of digital communication use a signal alphabet consisting of just two characters, sometimes called bits. These characters are generally denoted by 0 and 1, but in practice they might be different electrical or optical levels.

A key question in discrete, noiseless communication is deciding how to most efficiently convert messages into the signal alphabet. The concepts involved will be illustrated by the following simplified example.

The message alphabet will be called *M* and will consist of the four characters A, B, C, and D. The signal alphabet will be called *S* and will consist of the characters 0 and 1. Furthermore, it will be assumed that the signal channel can transmit 10 characters from *S* each second. This rate is called the channel capacity. Subject to these constraints, the goal is to maximize the transmission rate of characters from *M*.

The first question is how to convert characters between *M* and *S*. One straightforward way is shown in the table Encoding 1 of *M* using *S*. Using this conversion, the message ABC would be transmitted using the sequence 000110. The conversion from *M* to *S* is referred to as encoding. (This type of encoding is not meant to disguise the message but simply to adapt it to the nature of the communication system. Private or secret encoding schemes are usually referred to as encryption; *see* cryptology.) Because each character from *M* is represented by two characters from *S* and because the channel capacity is 10 characters from *S* each second, this communication scheme can transmit five characters from *M* each second. However, the scheme shown in the table ignores the fact that characters are used with widely varying frequencies in most alphabets.

Encoding 1 of *M* using *S*

M	→	***S***
A		00
B		01
C		10
D		11

In typical English text the letter *e* occurs roughly 200 times as frequently as the letter *z*. Hence, one way to improve the efficiency of the signal transmission is to use shorter codes for the more frequent characters—an idea employed in the design of Morse

Code. For example, let it be assumed that generally one-half of the characters in the messages that we wish to send are the letter A, one-quarter are the letter B, one-eighth are the letter C, and one-eighth are the letter D. The table Encoding 2 of *M* using *S* summarizes this information and shows an alternative encoding for the alphabet *M*. Now the message ABC would be transmitted using the sequence 010110, which is also six characters long. To see that this second encoding is better, on average, than the first one requires a longer typical message. For instance, suppose that 120 characters from *M* are transmitted with the frequency distribution shown in this table.

Encoding 2 of *M* using *S*

frequency	*M*	→	*S*
50%	A		0
25%	B		10
12.5%	C		110
12.5%	D		111

The results are summarized in the table Comparison of two encodings from *M* to *S*. This table shows that the second encoding uses 30 fewer characters from *S* than the first encoding. Recall that the first encoding, limited by the channel capacity of 10 characters per second, would transmit five characters from *M* per second, irrespective of the message. Working under the same limitations, the second encoding would transmit all 120 characters from *M* in 21 seconds (210 characters from *S* at 10 characters per second)—which yields an average rate of about 5.7 characters per second. Note that this improvement is for a typical message (one that contains the expected frequency of A's and B's). For an atypical message—in this case, one with unusually many C's and D's—this encoding might actually take longer to transmit than the first encoding.

Comparison of two encodings from *M* to *S*

character	number of cases	length of encoding 1	length of encoding 2
A	60	120	60
B	30	60	60
C	15	30	45
D	15	30	45
Totals	**120**	**240**	**210**

A natural question to ask at this point is whether the above scheme is really the best possible encoding or whether something better can be devised. Shannon was able to answer this question using a quantity that he called “entropy”; his concept is discussed in a later section, but, before proceeding to that discussion, a brief review of some practical issues in decoding and encoding messages is in order.

Some practical encoding/decoding questions

To be useful, each encoding must have a unique decoding. Consider the encoding shown in the table A less useful encoding. While every message can be encoded using this scheme, some will have duplicate encodings. For example, both the message AA and the message C will have the encoding 00. Thus, when the decoder receives 00, it will have no obvious way of telling whether AA or C was the intended message. For this reason, the encoding shown in this table would have to be characterized as “less useful.”

A less useful encoding

M	→	*S*
A		0
B		1
C		00
D		11

Encodings that produce a different signal for each distinct message are called “uniquely decipherable.” Most real applications require uniquely decipherable codes.

Another useful property is ease of decoding. For example, using the first encoding scheme, a received signal can be divided into groups of two characters and then decoded using the table Encoding 1 of *M* using *S*. Thus, to decode the signal 11010111001010, first divide it into 11 01 01 11 00 10 10, which indicates that DBBDACC was the original message.

The scheme shown in the table Encoding 2 of *M* using *S* must be deciphered using a different technique because strings of different length are used to represent characters from *M*. The technique here is to read one digit at a time until a matching character is found in this table. For example, suppose that the string 01001100111010 is received. Reading this string from the left, the first 0 matches the character A. Thus, the string 1001100111010 now remains. Because 1, the next digit, does not match any entry in the table, the next digit must now be appended. This two-digit combination, 10, matches the character B.

The table Decoding a message encoded with encoding 2 shows each unique stage of decoding this string. While the second encoding might appear to be complicated to decode, it is logically simple and easy to automate. The same technique can also be used to decode the first encoding.

Decoding a message encoded with encoding 2

decoded so far	remainder of signal string
	01001100111010
A	1001100111010
A B	01100111010
A B A	1100111010
A B A C	0111010
A B A C A	111010
A B A C A D	010
A B A C A D A	10
A B A C A D A B	

The table An impractical encoding, on the other hand, shows a code that involves some complications in its decoding. The encoding here has the virtue of being uniquely

decipherable, but, to understand what makes it “impractical,” consider the following strings: 011111111111111 and 011111111111111. The first is an encoding of CDDDD and the second of BDDDD. Unfortunately, to decide whether the first character is a B or a C requires viewing the entire string and then working back. Having to wait for the entire signal to arrive before any part of the message can be decoded can lead to significant delays.

An impractical encoding

M	→	*S*
A		0
B		01
C		011
D		111

In contrast, the encodings in both the table Encoding 1 of *M* using *S* and the table Encoding 2 of *M* using *S* allow messages to be decoded as the signal is received, or “on the fly.” The table Another comparison of encoding from *M* to *S* compares the first two encodings.

Another comparison of encodings from *M* to *S*

character	number of cases	length of encoding 1	length of encoding 2
A	30	60	30
B	30	60	60
C	30	60	90
D	30	60	90
Totals	**120**	**240**	**270**

Encodings that can be decoded on the fly are called prefix codes or instantaneous codes. Prefix codes are always uniquely decipherable, and they are generally preferable to nonprefix codes for communication because of their simplicity. Additionally, it has been shown that there must always exist a prefix code whose transmission rate is as good as that of any uniquely decipherable code, and, when the probability distribution of characters in the message is known, further improvements in the transmission rate

can be achieved by the use of variable-length codes, such as the encoding used in the table Encoding 2 of *M* using *S*. (Huffman codes, invented by the American D.A. Huffman in 1952, produce the minimum average code length among all uniquely decipherable variable-length codes for a given symbol set and a given probability distribution.)

Entropy

Shannon's concept of entropy can now be taken up. Recall that the table Comparison of two encodings from *M* to *S* showed that the second encoding scheme would transmit an average of 5.7 characters from *M* per second. But suppose that, instead of the distribution of characters shown in the table, a long series of As were transmitted. Because each A is represented by just a single character from *S*, this would result in the maximum transmission rate, for the given channel capacity, of 10 characters per second from *M*. On the other hand, transmitting a long series of Ds would result in a transmission rate of only 3.3 characters from *M* per second because each D must be represented by 3 characters from *S*. The average transmission rate of 5.7 is obtained by taking a weighted average of the lengths of each character code and dividing the channel speed by this average length. The formula for average length is given by:

$$\text{AvgLength} = .5 \times 1 + .25 \times 2 + .125 \times 3 + .125 \times 3 = 1.75,$$

where the length of each symbol's code is multiplied by its probability, or relative frequency. (For instance, since the letter B makes up 25 percent of an average message, its relative frequency is .25. That figure is multiplied by 2, the number of characters that encode B in the signal alphabet.) When the channel speed of 10 characters per second is divided by the average length computed above, the result is 10/1.75, or approximately 5.7 characters from *M* per second.

The average length formula can be generalized as:

$$\text{AvgLength} = p_1\,\text{Length}(c_1) + p_2\,\text{Length}(c_2) + \cdots + p_k\,\text{Length}(c_k),$$

where p_i is the probability of the *i*th character (here called c_i) and $\text{Length}(c_i)$ represents the length of the encoding for c_i. Note that this equation can be used to compare the transmission efficiency of existing encodings, but it cannot be used to discover the best possible encoding. Shannon, however, was able to find a quantity that does provide a theoretical limit for the efficiency of any possible encoding, based solely upon the average distribution of characters in the message alphabet. This is the quantity that he called entropy, and it is represented by *H* in the following formula:

$$H = p_1 \log_s(1/p_1) + p_2 \log_s(1/p_2) + \cdots + p_k \log_s(1/p_k).$$

(For a review of logs, *see* logarithm.) There are several things worth noting about this equation. First is the presence of the symbol $\log_s$. Here the subscript *s* represents the

number of elements in the signal alphabet S; $\log_s$, therefore, may be thought of as calculating an "optimal length" for a given distribution. Second, note that the reciprocals of the probabilities ($1/p_1$, $1/p_2$, ...) are used rather than the probabilities themselves (p_1, p_2, ...). Before explaining this equation in more detail, the following discovery of Shannon makes explicit the relationship between H and the AvgLength:

$$H \leq \text{AvgLength}.$$

Thus, the entropy for a given message alphabet determines the limit on average encoding efficiency (as measured by message length).

Because the signal alphabet, S, has only two symbols (0 and 1), a very small table of values of $\log_2$, as shown in the table Some values of $\log_2$, will suffice for illustration. (Readers with access to a scientific calculator may compare results.)

Some values of $\log_2$

n	$\log_2(n)$	n	$\log_2(n)$	n	$\log_2(n)$
1.0	0.000	4.0	2.000	7.0	2.807
1.5	0.585	4.5	2.170	7.5	2.907
2.0	1.000	5.0	2.322	8.0	3.000
2.5	1.322	5.5	2.459	8.5	3.087
3.0	1.585	6.0	2.585	9.0	3.170
3.5	1.807	6.5	2.700	9.5	3.248

With these preliminaries established, it is now possible to decide whether the encodings introduced earlier are truly optimal. In the first distribution (shown in the table Encoding 1 of M using S) all characters have a probability of 0.25. In this case, the entropy is given by

$$.25 \log_2(1/.25) + .25 \log_2(1/.25) + .25 \log_2(1/.25) + .25 \log_2(1/.25),$$

which is equal to $4 \times .25 \times 2 = 2$. Recall that the average length for the first encoding is also 2; hence, this encoding is optimal and cannot be improved.

For the second distribution (shown in the table Encoding 2 of M using S) the entropy is

$$.5 \log_2(1/.5) + .25 \log_2(1/.25) + .125 \log_2(1/.125) + .125 \log_2(1/.125),$$

which is equal to .5 + .5 + .375 + .375 = 1.75. Recall that this is equal to the average length

of the second encoding for this distribution of characters. Once again, an optimal encoding has been found.

The two examples just considered might suggest that it is always easy to find an optimal code. Therefore, it may be worth looking at a counterexample. Suppose that the probabilities for the characters in *M* are altered as shown in the table Altered probabilities for *M*. For the distribution given in this table,

$$H = .4 \log_2(2.5) + 3 \times .2 \log_2(5) = 1.922.$$

In this case, however, all simple encodings for *M*—those that substitute a string of characters from *S* for each character in *M*—have an average length ≥ 2. Thus, the bound computed using entropy cannot be attained with simple encodings. Shannon illustrated a technique for improving the performance of codes at the cost of complicating the encoding and decoding. The basic idea is to encode blocks of characters taken from *M*. For example, consider how often pairs of the characters shown in this table occur, assuming that the characters appear independently of each other. The probability for each pair is obtained by multiplying together the probabilities of the individual characters that make up the pair, as shown in the table. The pair AA would occur on the average 16 percent of the time (.16 = .4 × .4).

Altered probabilities for *M*

character	probability
A	.4
B	.2
C	.2
D	.2

The 16 possible pairs of A, B, C, and D, together with their probabilities, and a possible encoding, are shown in the table Probabilities of pairs of characters from *M*. As can be verified, the encoding given in this table is a prefix code. The average length of the encoding in this table is 3.92 characters of *S* for every 2 characters of *M*, or 1.96 characters of *S* for every character of *M*. This is better than the 2.0 obtained earlier, although still not equal to the entropy. Because the entropy is not exactly equal to any fraction, no code exists whose average length is exactly equal to the entropy. But Shannon did show that more complex codes can always be created whose average length is as close to the entropy limit as desired—at the cost of being increasingly complex and difficult to utilize.

Probabilities of pairs of characters from *M*

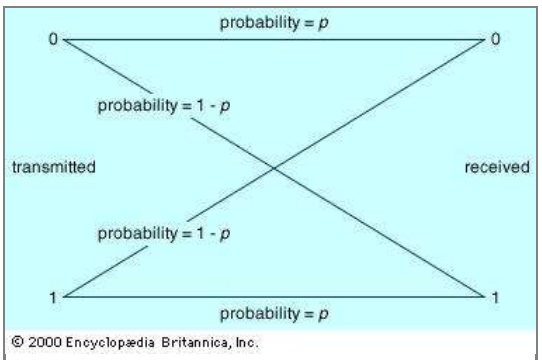
pair	probability	encoding	pair	probability	encoding
AA	.16	000	CA	.08	1000
AB	.08	101	CB	.04	01110
AC	.08	0011	CC	.04	01101
AD	.08	0010	CD	.04	01100
BA	.08	110	DA	.08	111
BB	.04	01011	DB	.04	01010
BC	.04	1001	DC	.04	01001
BD	.04	01111	DD	.04	01000

Summarizing thus far: The average character distribution in the message alphabet determines a limit, known as Shannon's entropy, on the best average (that is, the shortest) attainable encoding scheme. The theoretical best encoding scheme can be attained only in special circumstances. Finally, an encoding scheme can be found as close to the theoretical best as desired, although its use may be impractical because of the necessary increase in its complexity.

Discrete, noisy communication and the problem of error

In the discussion above, it is assumed unrealistically that all messages are transmitted without error. In the real world, however, transmission errors are unavoidable—especially given the presence in any communication channel of noise, which is the sum total of random signals that interfere with the communication signal. In order to take the inevitable transmission errors of the real world into account, some adjustment in encoding schemes is necessary. The figure shows a simple model of transmission in the presence of noise, the binary symmetric channel. *Binary* indicates that this channel transmits only two distinct characters, generally interpreted as 0 and 1, while *symmetric* indicates that errors are equally probable regardless of which character is transmitted. The probability that a character is transmitted without error is labeled *p*; hence, the probability of error is 1 − *p*.

Consider what happens as zeros and ones, hereafter referred to as bits, emerge from the receiving end of the channel. Ideally, there would be a means of determining which bits were received correctly. In that case, it is possible to imagine two printouts:



The binary symmetric channelThis type of channel transmits only two distinct characters, generally interpreted as 0 and 1, hence the designation *binary*. The probability of correctly receiving either character is the same, namely, *p*, which accounts for the designation *symmetric*. *Encyclopædia Britannica, Inc.*

10110101010010011001010011101101000010100101—*Signal*

00000000000100000000100000000010000000011001—*Errors*

Signal is the message as received, while each 1 in *Errors* indicates a mistake in the corresponding *Signal* bit. (*Errors* itself is assumed to be error-free.)

Shannon showed that the best method for transmitting error corrections requires an average length of

$$E = p \log_2(1/p) + (1 - p) \log_2(1/(1 - p))$$

bits per error correction symbol. Thus, for every bit transmitted at least *E* bits have to be reserved for error corrections. A reasonable measure for the effectiveness of a binary symmetric channel at conveying information can be established by taking its raw throughput of bits and subtracting the number of bits necessary to transmit error corrections. The limit on the efficiency of a binary symmetric channel with noise can now be given as a percentage by the formula 100 × (1 − *E*). Some examples follow.

Suppose that $p = {}^{1}/_{2}$, meaning that each bit is received correctly only half the time. In this case $E = 1$, so the effectiveness of the channel is 0 percent. In other words, no information is being transmitted. In effect, the error rate is so high that there is no way to tell whether any symbol is correct—one could just as well flip a coin for each bit at the receiving end. On the other hand, if the probability of correctly receiving a character is .99, *E* is roughly .081, so the effectiveness of the channel is roughly 92 percent. That is, a 1 percent error rate results in the net loss of about 8 percent of the channel's transmission capacity.

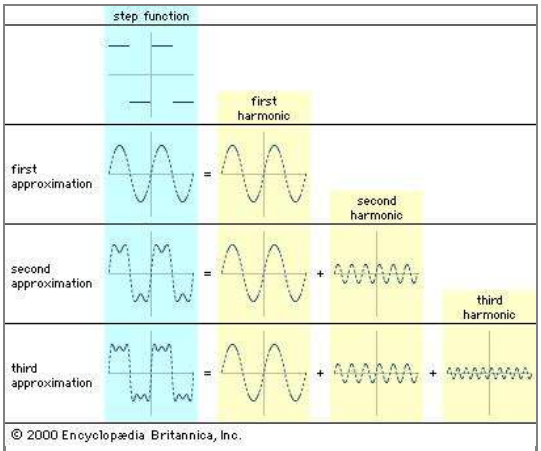
One interesting aspect of Shannon's proof of a limit for minimum average error correction length is that it is nonconstructive; that is, Shannon proved that a shortest

correction code must always exist, but his proof does not indicate how to construct such a code for each particular case. While Shannon's limit can always be approached to any desired degree, it is no trivial problem to find effective codes that are also easy and quick to decode.

Continuous communication and the problem of bandwidth

Continuous communication, unlike discrete communication, deals with signals that have potentially an infinite number of different values. Continuous communication is closely related to discrete communication (in the sense that any continuous signal can be approximated by a discrete signal), although the relationship is sometimes obscured by the more sophisticated mathematics involved.

The most important mathematical tool in the analysis of continuous signals is Fourier analysis, which can be used to model a signal as a sum of simpler sine waves. The figure indicates how the first few stages might appear. It shows a square wave, which has points of discontinuity ("jumps"), being modeled by a sum of sine waves. The curves to the right of the square wave show what are called the harmonics of the square wave. Above the line of harmonics are curves obtained by the addition of each successive harmonic; these curves can be seen to resemble the square wave more closely with each addition. If the entire infinite set of harmonics were added together, the square wave would be reconstructed almost exactly. Fourier analysis is useful because most communication circuits are linear, which essentially means that the whole is equal to the sum of the parts. Thus, a signal can be studied by separating, or decomposing, it into its simpler harmonics.



An example of Fourier analysisUsing Fourier analysis, a step function is modeled, or decomposed, as the sum of various sine functions. This striking example demonstrates how even an obviously discontinuous and piecewise linear graph (a step function) can be reproduced to any desired level of accuracy by combining

A signal is said to be band-limited or bandwidth-limited if it can be represented by a finite number of harmonics. Engineers limit the bandwidth of signals to enable multiple signals to share the same channel with minimal interference. A key result that pertains to bandwidth-limited signals is Nyquist's sampling theorem, which states that a signal of bandwidth B can be reconstructed by taking $2B$ samples every second. In 1924, Harry Nyquist derived the following formula for the maximum data rate that can be achieved in a noiseless channel:

$$\text{Maximum Data Rate} = 2B \log_2 V \text{ bits per second,}$$

where B is the bandwidth of the channel and V is the number of discrete signal levels used in the

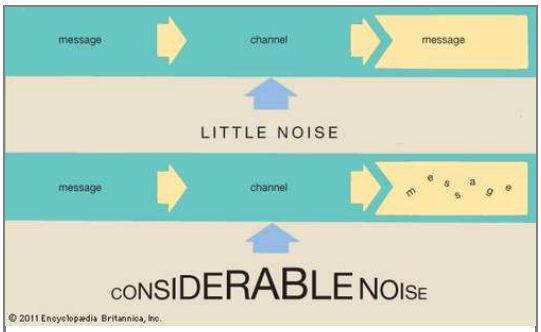
enough sine functions, each of which is continuous and nonlinear.
Encyclopædia Britannica, Inc.

channel. For example, to send only zeros and ones requires two signal levels. It is possible to envision any number of signal levels, but in practice the difference between signal levels must get smaller, for a fixed bandwidth, as the number of levels increases. And as the differences between signal levels decrease, the effect of noise in the channel becomes more pronounced.

Every channel has some sort of noise, which can be thought of as a random signal that contends with the message signal. If the noise is too great, it can obscure the message. Part of Shannon's seminal contribution to information theory was showing how noise affects the message capacity of a channel. In particular, Shannon derived the following formula:

$$\text{Maximum Data Rate} = B \log_2(1 + {}^{S}/_{N}) \text{ bits per second,}$$

where B is the bandwidth of the channel, and the quantity ${}^{S}/_{N}$ is the signal-to-noise ratio, which is often given in decibels (dB). Observe that the larger the signal-to-noise ratio, the greater the data rate. Another point worth observing, though, is that the $\log_2$ function grows quite slowly. For example, suppose ${}^{S}/_{N}$ is 1,000, then $\log_2 1{,}001 = 9.97$. If ${}^{S}/_{N}$ is doubled to 2,000, then $\log_2 2{,}001 = 10.97$. Thus, doubling S/N produces only a 10 percent gain in the maximum data rate. Doubling S/N again would produce an even smaller percentage gain.



communication: noise

In electronics and information theory, noise refers to those random, unpredictable, and undesirable signals, or changes in signals, that mask the desired information content.
Encyclopædia Britannica, Inc.

Applications Of Information Theory

Data compression

Shannon's concept of entropy (a measure of the maximum possible efficiency of any encoding scheme) can be used to determine the maximum theoretical compression for a given message alphabet. In particular, if the entropy is less than the average length of an encoding, compression is possible.

The table Relative frequencies of characters in English text shows the relative frequencies of letters in representative English text. The table assumes that all letters have been capitalized and ignores all other characters except for spaces. Note that letter frequencies depend upon the particular text sample. An essay about zebras in the zoo, for instance, is likely to have a much greater

frequency of *z*'s than the table would suggest. Nevertheless, the frequency distribution for any very large sample of English text would appear quite similar to this table. Calculating the entropy for this distribution gives 4.08 bits per character. (Recall Shannon's formula for entropy.) Because normally 8 bits per character are used in the most common coding standard, Shannon's theory shows that there exists an encoding that is roughly twice as efficient as the normal one for this simplified message alphabet. These results, however, apply only to large samples and assume that the source of the character stream transmits characters in a random fashion based on the probabilities in the table. Real text does not perfectly fit this model; parts of it tend to be highly nonrandom and repetitive. Thus, the theoretical results do not immediately translate into practice.

Relative frequencies of characters in English text

character	relative frequency (probability)	character	relative frequency (probability)
(space)	.1859	F	.0208
E	.1031	M	.0198
T	.0796	W	.0175
A	.0642	Y	.0164
O	.0632	P	.0152
I	.0575	G	.0152
N	.0574	B	.0127
S	.0514	V	.0083
R	.0484	K	.0049
H	.0467	X	.0013
L	.0321	Q	.0008
D	.0317	J	.0008
U	.0228	Z	.0005
C	.0218		

In 1977–78 the Israelis Jacob Ziv and Abraham Lempel published two papers that

showed how compression can be done dynamically. The basic idea is to store blocks of text in a dictionary and, when a block of text reappears, to record which block was repeated rather than recording the text itself. Although there are technical issues related to the size of the dictionary and the updating of its entries, this dynamic approach to compression has proved very useful, in part because the compression algorithm adapts to optimize the encoding based upon the particular text. Many computer programs use compression techniques based on these ideas. In practice, most text files compress by about 50 percent—that is, to approximately 4 bits per character. This is the number suggested by the entropy calculation.

Error-correcting and error-detecting codes

Shannon's work in the area of discrete, noisy communication pointed out the possibility of constructing error-correcting codes. Error-correcting codes add extra bits to help correct errors and thus operate in the opposite direction from compression. Error-detecting codes, on the other hand, indicate that an error has occurred but do not automatically correct the error. Frequently the error is corrected by an automatic request to retransmit the message. Because error-correcting codes typically demand more extra bits than error-detecting codes, in some cases it is more efficient to use an error-detecting code simply to indicate what has to be retransmitted.

Deciding between error-correcting and error-detecting codes requires a good understanding of the nature of the errors that are likely to occur under the circumstances in which the message is being sent. Transmissions to and from space vehicles generally use error-correcting codes because of the difficulties in getting retransmission. Because of the long distances and low power available in transmitting from space vehicles, it is easy to see that the utmost skill and art must be employed to build communication systems that operate at the limits imposed by Shannon's results.

A common type of error-detecting code is the parity code, which adds one bit to a block of bits so that the ones in the block always add up to either an odd or even number. For example, an odd parity code might replace the two-bit code words 00, 01, 10, and 11 with the three-bit words 001, 010, 100, and 111. Any single transformation of a 0 to a 1 or a 1 to a 0 would change the parity of the block and make the error detectable. In practice, adding a parity bit to a two-bit code is not very efficient, but for longer codes adding a parity bit is reasonable. For instance, computer and fax modems often communicate by sending eight-bit blocks, with one of the bits reserved as a parity bit. Because parity codes are simple to implement, they are also often used to check the integrity of computer equipment.

As noted earlier, designing practical error-correcting codes is not easy, and Shannon's

work does not provide direct guidance in this area. Nevertheless, knowing the physical characteristics of the channel, such as bandwidth and signal-to-noise ratio, gives valuable knowledge about maximum data transmission capabilities.

Cryptology

Cryptology is the science of secure communication. It concerns both cryptanalysis, the study of how encrypted information is revealed (or decrypted) when the secret "key" is unknown, and cryptography, the study of how information is concealed and encrypted in the first place.

Shannon's analysis of communication codes led him to apply the mathematical tools of information theory to cryptography in "Communication Theory of Secrecy Systems" (1949). In particular, he began his analysis by noting that simple transposition ciphers—such as those obtained by permuting the letters in the alphabet—do not affect the entropy because they merely relabel the characters in his formula without changing their associated probabilities.

Cryptographic systems employ special information called a key to help encrypt and decrypt messages. Sometimes different keys are used for the encoding and decoding, while in other instances the same key is used for both processes. Shannon made the following general observation: "the amount of uncertainty we can introduce into the solution cannot be greater than the key uncertainty." This means, among other things, that random keys should be selected to make the encryption more secure. While Shannon's work did not lead to new practical encryption schemes, he did supply a framework for understanding the essential features of any such system.

Linguistics

While information theory has been most helpful in the design of more efficient telecommunication systems, it has also motivated linguistic studies of the relative frequencies of words, the length of words, and the speed of reading.

The best-known formula for studying relative word frequencies was proposed by the American linguist George Zipf in *Selected Studies of the Principle of Relative Frequency in Language* (1932). Zipf's Law states that the relative frequency of a word is inversely proportional to its rank. That is, the second most frequent word is used only half as often as the most frequent word, and the 100th most frequent word is used only one hundredth as often as the most frequent word.

Consistent with the encoding ideas discussed earlier, the most frequently used words tend to be the shortest. It is uncertain how much of this phenomenon is due to a

“principle of least effort,” but using the shortest sequences for the most common words certainly promotes greater communication efficiency.

Information theory provides a means for measuring redundancy or efficiency of symbolic representation within a given language. For example, if English letters occurred with equal regularity (ignoring the distinction between uppercase and lowercase letters), the expected entropy of an average sample of English text would be $\log_2(26)$, which is approximately 4.7. The table Relative frequencies of characters in English text shows an entropy of 4.08, which is not really a good value for English because it overstates the probability of combinations such as *qa*. Scientists have studied sequences of eight characters in English and come up with a figure of about 2.35 for the average entropy of English. Because this is only half the 4.7 value, it is said that English has a relative entropy of 50 percent and a redundancy of 50 percent.

A redundancy of 50 percent means that roughly half the letters in a sentence could be omitted and the message still be reconstructable. The question of redundancy is of great interest to crossword puzzle creators. For example, if redundancy was 0 percent, so that every sequence of characters was a word, then there would be no difficulty in constructing a crossword puzzle because any character sequence the designer wanted to use would be acceptable. As redundancy increases, the difficulty of creating a crossword puzzle also increases. Shannon showed that a redundancy of 50 percent is the upper limit for constructing two-dimensional crossword puzzles and that 33 percent is the upper limit for constructing three-dimensional crossword puzzles.

Shannon also observed that when longer sequences, such as paragraphs, chapters, and whole books, are considered, the entropy decreases and English becomes even more predictable. He considered longer sequences and concluded that the entropy of English is approximately one bit per character. This indicates that in longer text nearly all of the message can be guessed from just a 20 to 25 percent random sample.

Various studies have attempted to come up with an information processing rate for human beings. Some studies have concentrated on the problem of determining a reading rate. Such studies have shown that the reading rate seems to be independent of language—that is, people process about the same number of bits whether they are reading English or Chinese. Note that although Chinese characters require more bits for their representation than English letters—there exist about 10,000 common Chinese characters, compared with 26 English letters—they also contain more information. Thus, on balance, reading rates are comparable.

Algorithmic information theory

In the 1960s the American mathematician Gregory Chaitin, the Russian mathematician

Andrey Kolmogorov, and the American engineer Raymond Solomonoff began to formulate and publish an objective measure of the intrinsic complexity of a message. Chaitin, a research scientist at IBM, developed the largest body of work and polished the ideas into a formal theory known as algorithmic information theory (AIT). The *algorithmic* in AIT comes from defining the complexity of a message as the length of the shortest algorithm, or step-by-step procedure, for its reproduction.

Physiology

Almost as soon as Shannon's papers on the mathematical theory of communication were published in the 1940s, people began to consider the question of how messages are handled inside human beings. After all, the nervous system is, above all else, a channel for the transmission of information, and the brain is, among other things, an information processing and messaging centre. Because nerve signals generally consist of pulses of electrical energy, the nervous system appears to be an example of discrete communication over a noisy channel. Thus, both physiology and information theory are involved in studying the nervous system.

Many researchers (being human) expected that the human brain would show a tremendous information processing capability. Interestingly enough, when researchers sought to measure information processing capabilities during "intelligent" or "conscious" activities, such as reading or piano playing, they came up with a maximum capability of less than 50 bits per second. For example, a typical reading rate of 300 words per minute works out to about 5 words per second. Assuming an average of 5 characters per word and roughly 2 bits per character yields the aforementioned rate of 50 bits per second. Clearly, the exact number depends on various assumptions and could vary depending on the individual and the task being performed. It is known, however, that the senses gather some 11 million bits per second from the environment.

The table Information transmission rates of the senses shows how much information is processed by each of the five senses. This table immediately directs attention to the problem of determining what is happening to all this data. In other words, the human body sends 11 million bits per second to the brain for processing, yet the conscious mind seems to be able to process only 50 bits per second.

Information transmission rates of the senses

sensory system	bits per second
eyes	10,000,000
skin	1,000,000
ears	100,000
smell	100,000
taste	1,000

It appears that a tremendous amount of compression is taking place if 11 million bits are being reduced to less than 50. Note that the discrepancy between the amount of information being transmitted and the amount of information being processed is so large that any inaccuracy in the measurements is insignificant.

Two more problems suggest themselves when thinking about this immense amount of compression. First is the problem of determining how long it takes to do the compression, and second is the problem of determining where the processing power is found for doing this much compression.

The solution to the first problem is suggested by the approximately half-second delay between the instant that the senses receive a stimulus and the instant that the mind is conscious of a sensation. (To compensate for this delay, the body has a reflex system that can respond in less than one-tenth of second, before the mind is conscious of the stimulus.) This half-second delay seems to be the time required for processing and compressing sensory input.

The solution to the second problem is suggested by the approximately 100 billion cells of the brain, each with connections to thousands of other brain cells. Equipped with this many processors, the brain might be capable of executing as many as 100 billion operations per second, a truly impressive number.

It is often assumed that consciousness is the dominant feature of the brain. The brief observations above suggest a rather different picture. It now appears that the vast majority of processing is accomplished outside conscious notice and that most of the body's activities take place outside direct conscious control. This suggests that practice and habit are important because they train circuits in the brain to carry out some actions "automatically," without conscious interference. Even such a "simple" activity as walking is best done without interference from consciousness, which does not have

enough information processing capability to keep up with the demands of this task.

The brain also seems to have separate mechanisms for short-term and long-term memory. Based on psychologist George Miller's paper "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" (1956), it appears that short-term memory can only store between five and nine pieces of information to which it has been exposed only briefly. Note that this does not mean between five and nine bits, but rather five to nine chunks of information. Obviously, long-term memory has a greater capacity, but it is not clear exactly how the brain stores information or what limits may exist. Some scientists hope that information theory may yet afford further insights into how the brain functions.

Physics

The term *entropy* was originally introduced by the German physicist Rudolf Clausius in his work on thermodynamics in the 19th century. Clausius invented the word so that it would be as close as possible to the word *energy*. In certain formulations of statistical mechanics a formula for entropy is derived that looks confusingly similar to the formula for entropy derived by Shannon.

There are various intersections between information theory and thermodynamics. One of Shannon's key contributions was his analysis of how to handle noise in communication systems. Noise is an inescapable feature of the universe. Much of the noise that occurs in communication systems is a random noise, often called thermal noise, generated by heat in electrical circuits. While thermal noise can be reduced, it can never be completely eliminated. Another source of noise is the homogeneous cosmic background radiation, believed to be a remnant from the creation of the universe. Shannon's work permits minimal energy costs to be calculated for sending a bit of information through such noise.

Another problem addressed by information theory was dreamed up by the Scottish physicist James Clerk Maxwell in 1871. Maxwell created a "thought experiment" that apparently violates the second law of thermodynamics. This law basically states that all isolated systems, in the absence of an input of energy, relentlessly decay, or tend toward disorder. Maxwell began by postulating two gas-filled vessels at equal temperatures, connected by a valve. (Temperature can be defined as a measure of the average speed of gas molecules, keeping in mind that individual molecules can travel at widely varying speeds.) Maxwell then described a mythical creature, now known as Maxwell's demon, that is able rapidly to open and close the valve so as to allow only fast-moving molecules to pass in one direction and only slow-moving molecules to pass in the other direction. Alternatively, Maxwell envisioned his demon allowing molecules to pass

through in only one direction. In either case, a “hot” and a “cold” vessel or a “full” and “empty” vessel, the apparent result is two vessels that, with no input of energy from an external source, constitute a more orderly isolated system—thus violating the second law of thermodynamics.

Information theory allows one exorcism of Maxwell’s demon to be performed. In particular, it shows that the demon needs information in order to select molecules for the two different vessels but that the transmission of information requires energy. Once the energy requirement for collecting information is included in the calculations, it can be seen that there is no violation of the second law of thermodynamics.

> *George Markowsky*

"Information theory". *Encyclopædia Britannica*. *Encyclopædia Britannica Online*. Encyclopædia Britannica Inc., 2018. Web. 23 mars. 2018 <https://www.britannica.com/science/information-theory>.