**information theory**. (2009). In *Encyclopædia Britannica.* Retrieved April  1,  2009, from Encyclopædia Britannica Online: <http://original.britannica.com/eb/article-9106012>

**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 Elwood 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.

**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](http://original.britannica.com/eb/article-9125110/Harry-Nyquist" \o "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](http://original.britannica.com/eb/article-9125111/Claude-Shannon" \o "Claude-Shannon)**'s **[“The Mathematical Theory of Communication”](http://original.britannica.com/eb/topic?idxStructId=369184&typeId=13" \o "\“The 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](http://original.britannica.com/eb/topic?idxStructId=377051&typeId=13" \o "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.

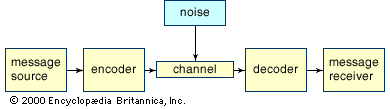
Shannon also made the startling discovery that, even in the presence of noise, it is always possible to transmit signals arbitrarily close to the theoretical channel capacity. This discovery inspired engineers to look for practical techniques to improve performance in signal transmissions that were far 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**](http://original.britannica.com/eb/article-214952/information-theory#214952.toc) 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**.



As the underpinning of his theory, Shannon developed a very simple, abstract model of communication, as shown in the [figure](http://original.britannica.com/eb/art-52075/Shannons-communication-model-Consider-a-simple-telephone-conversation-A-person?articleTypeId=1). Because his model is abstract, it applies in many situations, which contributes to its broad scope and power.

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](http://original.britannica.com/eb/article-9108588/speech" \o "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**](http://original.britannica.com/eb/article-48688/physiology#48688.toc).

The **[channel](http://original.britannica.com/eb/topic?idxStructId=105717&typeId=13" \o "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](http://original.britannica.com/eb/topic?idxStructId=417192&typeId=13" \o "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.

##### Classical information theory > Four types of communication

There are two fundamentally different ways to transmit messages: via discrete signals and via continuous signals. **[Discrete signals](http://original.britannica.com/eb/topic?idxStructId=709369&typeId=13" \o "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. Classical information theory > Four types of communication

There are two fundamentally different ways to transmit messages: via discrete signals and via continuous signals. [**Discrete signals**](http://original.britannica.com/eb/topic?idxStructId=709369&typeId=13) 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.

**Classical information theory > Four types of communication > Discrete, noiseless communication and the concept of entropy >** **From message alphabet to signal alphabet**

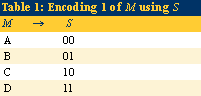
As mentioned above, the English alphabet is a discrete communication system. It consists of a finite set of characters, such as uppercase and lowercase letters, digits, and various punctuation marks. Messages are composed by stringing these individual characters together appropriately. (Henceforth, signal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

For **[noiseless communications](http://original.britannica.com/eb/topic?idxStructId=709369&typeId=13" \o "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**](http://original.britannica.com/eb/article-9053835/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 most efficiently to convert messages into the signal alphabet. The concepts involved will be illustrated by the following simplified example.

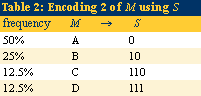
The message alphabet will be called *M* and will consist of the four characters A, B, C, and D. The signal alphabet will be called *S* and will consist of the characters 0 and 1. Furthermore, it will be assumed that the signal channel can transmit 10 characters from *S* each second. This rate is called 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 Table 1. Using this conversion, the message ABC would be transmitted using the sequence 000110.



The conversion from *M* to *S* is referred to as **[encoding](http://original.britannica.com/eb/topic?idxStructId=186551&typeId=13" \o "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**](http://original.britannica.com/eb/article-9109639/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 Table 1 ignores the fact that characters are used with widely varying frequencies in most alphabets. For example, 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. Table 2 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 Table 2. The results are summarized in Table 3.

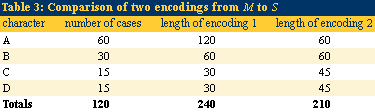
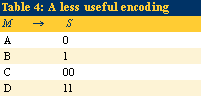


Table 3 shows that the second encoding uses 30 fewer characters from *S* than the first encoding. Recall that the first encoding, limited by the channel capacity of 10 characters per second, would transmit five characters from *M* per second, irrespective of the message. Working under the same limitations, 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 As and Bs). For an atypical message—in this case, one with unusually many Cs and Ds—this encoding might actually take longer to transmit than the first encoding.

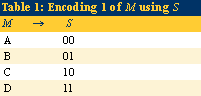
A natural question to ask at this point is whether the above scheme is really the best possible encoding or whether something better can be devised. Shannon was able to answer this question using a quantity that he called “entropy”; his concept is discussed in a later section, but, before proceeding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

**Classical information theory > Four types of communication > Discrete, noiseless communication and the concept of entropy >** **Some practical encoding/decoding questions**

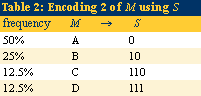
To be useful, each **[encoding](http://original.britannica.com/eb/topic?idxStructId=186551&typeId=13" \o "encoding)** must have a unique decoding. Consider the encoding shown in Table 4. While every message can be encoded using this scheme, some will have duplicate encodings. For example, both the message 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 Table 4 would have to be characterized as “less useful.”



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 [Table 1](http://original.britannica.com/eb/art-65068?articleTypeId=1). 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 second encoding scheme [(Table 2)](http://original.britannica.com/eb/art-65067?articleTypeId=1) 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 Table 2. 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. Table 5 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.

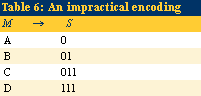
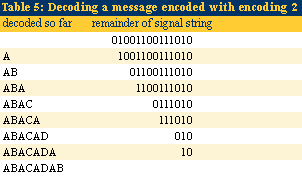
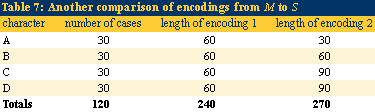
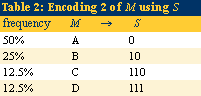


Table 6, on the other hand, shows a code that involves some complications in its decoding. The encoding here has the virtue of being uniquely decipherable, but, to understand what makes it “impractical,” consider the following strings: 011111111111111 and 01111111111111. The first is an encoding of CDDDD 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. In contrast, the encodings in both Table 1 and Table 2 allow messages to be decoded as the signal is received, or “on the fly.” Table 7 compares the first two encodings.



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 [Table 2](http://original.britannica.com/eb/art-65067?articleTypeId=1). ([**Huffman codes**](http://original.britannica.com/eb/article-76271/telecommunication#76271.toc), 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.)

**Classical information theory > Four types of communication > Discrete, noiseless communication and the concept of entropy >** **Entropy**



Shannon's concept of entropy can now be taken up. Recall that [Table 3](http://original.britannica.com/eb/art-65066?articleTypeId=1) 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:

AvgLength = .5  1 + .25  2 + .125  3 + .125  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:

AvgLength = *p*1 Length(*c*1) + *p*2 Length(*c*2) + … + *pk* Length(*ck*),

where *pi* is the probability of the *i*th character (here called *ci*) and Length(*ci*) represents the length of the encoding for *ci*. 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:

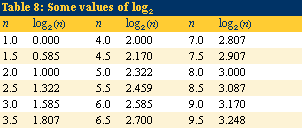
*H* = *p*1 log*s*(1/*p*1) + *p*2 log*s*(1/*p*2) + … + *pk* log*s*(1/*pk*).

(For a review of logs see [**logarithm**](http://original.britannica.com/eb/article-9048759/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*  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 log2, as shown in Table 8, will suffice for illustration. (Readers with access to a scientific calculator may compare results.)



With these preliminaries established, it is now possible to decide whether the encodings introduced earlier are truly optimal. In the first distribution ([Table 1](http://original.britannica.com/eb/art-65068?articleTypeId=1)) all characters have a probability of 0.25. In this case, the entropy is given by

.25 log2(1/.25) + .25 log2(1/.25) + .25 log2(1/.25) + .25 log2(1/.25),

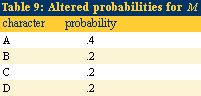
which is equal to 4  .25  2 = 2. Recall that the average length for the first encoding is also 2; hence, this encoding is optimal and cannot be improved.

For the second distribution (shown in Table 2) the entropy is

.5 log2(1/.5) + .25 log2(1/.25) + .125 log2(1/.125) + .125 log2(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 Table 9.

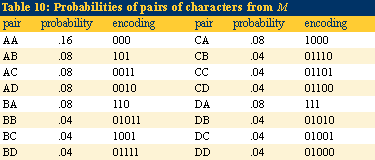


For the distribution given in Table 9,

*H* = .4 log2(2.5) + 3  .2 log2(5) = 1.922.

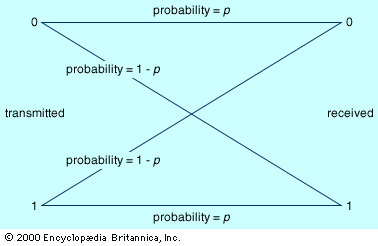
In this case, however, all simple encodings for *M*—those that substitute a string of characters from *S* for each character in *M*—have an average length  2. Thus, the bound computed using entropy cannot be attained with simple encodings.

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 Table 9 occur, assuming that the characters appear independently of each other. The pair AA would occur on the average 16 percent of the time (.16 = .4  .4). The 16 possible pairs of A, B, C, and D, together with their probabilities, and a possible encoding, are shown in Table 10. The probability for each pair is obtained by multiplying together the probabilities of the individual characters that make up the pair, as shown in Table 9. As can be verified, the encoding given in Table 10 is a prefix code.



The average length of the encoding in Table 10 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.

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.



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](http://original.britannica.com/eb/art-52076/The-binary-symmetric-channel-This-type-of-channel-transmits-only?articleTypeId=1) shows a simple model of transmission in the presence of noise, the **[binary](http://original.britannica.com/eb/article-9079220/binary-number-system" \o "binary-number-system)** 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](http://original.britannica.com/eb/article-9080319/bit" \o "bit)**, 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:

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* log2(1/*p*) + (1  *p*) log2(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.

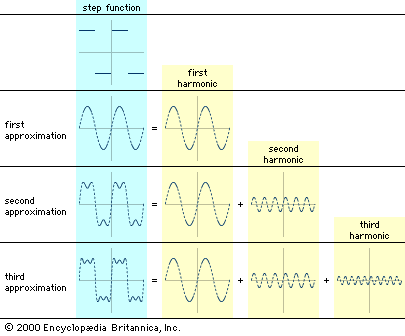
Suppose that *p* = 1/2, meaning that each bit is received correctly only half the time. In this case *E* = 1, so the effectiveness of the channel is 0 percent. In other words, no information is being transmitted. In effect, the error rate is so high that there is no way to tell whether any symbol is 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.

**Classical information theory > Four types of communication >** **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.

An example of Fourier analysis   
*Encyclopædia Britannica, Inc.*



The most important mathematical tool in the analysis of continuous signals is **[Fourier analysis](http://original.britannica.com/eb/topic?idxStructId=215106&typeId=13" \o "Fourier analysis)**, which can be used to model a signal as a sum of simpler sine waves. The [figure](http://original.britannica.com/eb/art-52077/An-example-of-Fourier-analysis-Using-Fourier-analysis-a-step?articleTypeId=1) 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 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.

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](http://original.britannica.com/eb/topic?idxStructId=520693&typeId=13" \o "sampling theorem)**, which states that a signal of bandwidth *B* can be reconstructed by taking 2*B* samples every second. In 1924, Harry Nyquist derived the following formula for the maximum data rate that can be achieved in a noiseless channel:

Maximum Data Rate = 2 *B* log2 *V* bits per second,

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

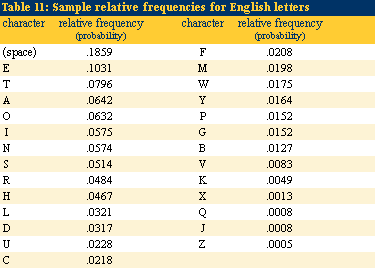
Every channel has some sort of noise, which can be thought of as a random signal that contends with the message signal. If the noise is too great, it can obscure the message. Part of Shannon's seminal contribution to information theory was showing how noise affects the message capacity of a channel.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

Maximum Data Rate = *B* log2(1 + *S*/*N*) bits per second,

##### where *B* is the bandwidth of the channel, and the quantity *S*/*N* is the [signal-to-noise ratio](http://original.britannica.com/eb/topic?idxStructId=543772&typeId=13" \o "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 log2 function grows quite slowly. For example, suppose *S*/*N* is 1,000, then log2 1,001 = 9.97. If *S*/*N* is doubled to 2,000, then log2 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. 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.

Table 11 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 Table 11.



Calculating the entropy for this distribution gives 4.08 bits per character. (Recall Shannon's formula for [**entropy**](http://original.britannica.com/eb/article-9106012/information-theory#426131.ref).) 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 Table 11. 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.

In 1977–78 the Israelis **[Jacob Ziv](http://original.britannica.com/eb/topic?idxStructId=657608&typeId=13" \o "Jacob Ziv)** and **[Abraham Lempel](http://original.britannica.com/eb/topic?idxStructId=335759&typeId=13" \o "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.

##### Applications of information theory > Error-correcting and error-detecting codes

Shannon's work in the area of [**discrete, noisy communication**](http://original.britannica.com/eb/article-214950/information-theory#214950.toc) pointed out the possibility of constructing error-correcting codes. **[Error-correcting codes](http://original.britannica.com/eb/topic?idxStructId=191931&typeId=13" \o "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](http://original.britannica.com/eb/topic?idxStructId=444003&typeId=13" \o "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.

##### Applications of information theory > Cryptology

**[Cryptology](http://original.britannica.com/eb/article-9109639/cryptology" \o "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](http://original.britannica.com/eb/article-9472150/cryptography" \o "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.

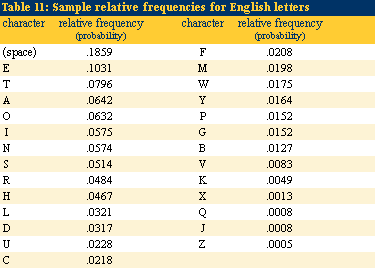
Cryptographic systems employ special information called a key to help encrypt and decrypt messages. Sometimes different keys are used for the encoding and decoding, while in other instances the same key is used for both processes. Shannon made the following general observation: “the amount of uncertainty 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.

##### Applications of information theory > 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](http://original.britannica.com/eb/article-9109779/English-language" \o "English-language)** 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 log2(26), which is approximately 4.7. [Table 11](http://original.britannica.com/eb/art-65065?articleTypeId=1) 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](http://original.britannica.com/eb/topic?idxStructId=494892&typeId=13" \o "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.

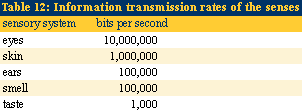
Various studies have attempted to come up with an information processing rate for human beings. Some studies have concentrated on the problem of determining a reading rate. Such studies have shown that the reading rate seems to be independent of language—that is, people process about the same number 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.

##### Applications of information theory > Algorithmic information theory

In the 1960s the American mathematician Gregory Chaitin, the Russian mathematician [**Andrey Kolmogorov**](http://original.britannica.com/eb/article-9045946/Andrey-Nikolayevich-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.

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](http://original.britannica.com/eb/topic?idxStructId=709369&typeId=13" \o "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. Table 12 shows how much information is processed by each of the five senses. The 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. 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.

**Applications of information theory >** **Physics**

The term [***entropy***](http://original.britannica.com/eb/article-9032727/entropy) was originally introduced by the German physicist [**Rudolf Clausius**](http://original.britannica.com/eb/article-9024255/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 system](http://original.britannica.com/eb/topic?idxStructId=129057&typeId=13" \o "communication system)**s. 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](http://original.britannica.com/eb/article-9108468/James-Clerk-Maxwell" \o "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**](http://original.britannica.com/eb/article-9051563/Maxwells-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](http://original.britannica.com/eb/article-9108468/James-Clerk-Maxwell" \o "James-Clerk-Maxwell)** 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.  
  
**Additional Reading**

John R. Pierce, *An Introduction to Information Theory: Symbols, Signals & Noise*, 2nd rev. ed. (1980), is an entertaining and very readable account of information theory, its applications, and related fields. While it displays the relevant mathematics, most of it can be read profitably by people with a weak mathematical background.

Steven Roman, *Coding and Information Theory* (1992), is meant for an introductory college course. This book assumes that the reader has a basic knowledge of probability; an appendix reviews necessary ideas from modern algebra.

Aleksandr I. Khinchin, *Mathematical Foundations of Information Theory*, trans. from Russian (1957, reissued 1967), is a mathematically challenging, but elegant treatment of information theory, intended for the advanced reader.

N.J.A. Sloane and Aaron D. Wyner (eds.), *Claude Elwood Shannon: Collected Papers* (1993), is a very interesting volume that shows the breadth and depth of Shannon's work. While most of the papers—such as “The Mathematical Theory of Communication” and “Communication Theory of Secrecy Systems”—require mathematical sophistication, some—such as “The Bandwagon,” “Game Playing Machines,” and “Claude Shannon's No-Drop Juggling Diorama”—do not.

[**George Markowsky**](http://original.britannica.com/eb/author?id=4491)

### From Wikipedia, the free encyclopedia

Jump to: [navigation](http://en.wikipedia.org/wiki/Information_theory#column-one#column-one), [search](http://en.wikipedia.org/wiki/Information_theory#searchInput#searchInput)

*Not to be confused with* [*information technology*](http://en.wikipedia.org/wiki/Information_technology)*,* [*information science*](http://en.wikipedia.org/wiki/Information_science)*, or* [*informatics*](http://en.wikipedia.org/wiki/Informatics)*.*

**Information theory** is a branch of [applied mathematics](http://en.wikipedia.org/wiki/Applied_mathematics) and [electrical engineering](http://en.wikipedia.org/wiki/Electrical_engineering) involving the quantification of [information](http://en.wikipedia.org/wiki/Information). Historically, information theory was developed by [Claude E. Shannon](http://en.wikipedia.org/wiki/Claude_E._Shannon) to find fundamental limits on compressing and reliably [communicating](http://en.wikipedia.org/wiki/Telecommunication) data. Since its inception it has broadened to find applications in many other areas, including [statistical inference](http://en.wikipedia.org/wiki/Statistical_inference), [natural language processing](http://en.wikipedia.org/wiki/Natural_language_processing), [cryptography](http://en.wikipedia.org/wiki/Cryptography) generally, networks other than communication networks — as in [neurobiology](http://en.wikipedia.org/wiki/Neurobiology),[[1]](http://en.wikipedia.org/wiki/Information_theory#cite_note-0#cite_note-0) the evolution[[2]](http://en.wikipedia.org/wiki/Information_theory#cite_note-1#cite_note-1) and function[[3]](http://en.wikipedia.org/wiki/Information_theory#cite_note-2#cite_note-2) of molecular codes, model selection[[4]](http://en.wikipedia.org/wiki/Information_theory#cite_note-3#cite_note-3) in ecology, thermal physics,[[5]](http://en.wikipedia.org/wiki/Information_theory#cite_note-4#cite_note-4) [quantum computing](http://en.wikipedia.org/wiki/Quantum_computing), plagiarism detection[[6]](http://en.wikipedia.org/wiki/Information_theory#cite_note-5#cite_note-5) and other forms of [data analysis](http://en.wikipedia.org/wiki/Data_analysis).[[7]](http://en.wikipedia.org/wiki/Information_theory#cite_note-6#cite_note-6)

A key measure of information in the theory is known as [**entropy**](http://en.wikipedia.org/wiki/Entropy_(information_theory)), which is usually expressed by the average number of bits needed for storage or communication. Intuitively, entropy quantifies the uncertainty involved when encountering a [random variable](http://en.wikipedia.org/wiki/Random_variable). For example, a fair coin flip (2 equally likely outcomes) will have less entropy than a roll of a die (6 equally likely outcomes).

Applications of fundamental topics of information theory include [lossless data compression](http://en.wikipedia.org/wiki/Lossless_data_compression) (e.g. [ZIP files](http://en.wikipedia.org/wiki/ZIP_(file_format))), [lossy data compression](http://en.wikipedia.org/wiki/Lossy_data_compression) (e.g. [MP3s](http://en.wikipedia.org/wiki/MP3)), and [channel coding](http://en.wikipedia.org/wiki/Channel_capacity) (e.g. for [DSL](http://en.wikipedia.org/wiki/DSL) lines). The field is at the intersection of [mathematics](http://en.wikipedia.org/wiki/Mathematics), [statistics](http://en.wikipedia.org/wiki/Statistics), [computer science](http://en.wikipedia.org/wiki/Computer_science), [physics](http://en.wikipedia.org/wiki/Physics), [neurobiology](http://en.wikipedia.org/wiki/Neurobiology), and [electrical engineering](http://en.wikipedia.org/wiki/Electrical_engineering). Its impact has been crucial to the success of the [Voyager](http://en.wikipedia.org/wiki/Voyager_program) missions to deep space, the invention of the compact disc, the feasibility of mobile phones, the development of the [Internet](http://en.wikipedia.org/wiki/Internet), the study of [linguistics](http://en.wikipedia.org/wiki/Linguistics) and of human perception, the understanding of [black holes](http://en.wikipedia.org/wiki/Black_hole), and numerous other fields[[*citation needed*](http://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]. Important sub-fields of information theory are source coding, channel coding, algorithmic complexity theory, algorithmic information theory, and measures of information.

|  |
| --- |
| Contents [[hide](javascript:toggleToc())]   * [1 Overview](http://en.wikipedia.org/wiki/Information_theory#Overview#Overview) * [2 Historical background](http://en.wikipedia.org/wiki/Information_theory#Historical_background#Historical_background) * [3 Quantities of information](http://en.wikipedia.org/wiki/Information_theory#Quantities_of_information#Quantities_of_information)   + [3.1 Entropy](http://en.wikipedia.org/wiki/Information_theory#Entropy#Entropy)   + [3.2 Joint entropy](http://en.wikipedia.org/wiki/Information_theory#Joint_entropy#Joint_entropy)   + [3.3 Conditional entropy (equivocation)](http://en.wikipedia.org/wiki/Information_theory#Conditional_entropy_.28equivocation.29#Conditional_entropy_.28equivocation.29)   + [3.4 Mutual information (transinformation)](http://en.wikipedia.org/wiki/Information_theory#Mutual_information_.28transinformation.29#Mutual_information_.28transinformation.29)   + [3.5 Kullback–Leibler divergence (information gain)](http://en.wikipedia.org/wiki/Information_theory#Kullback.E2.80.93Leibler_divergence_.28information_gain.29#Kullback.E2.80.93Leibler_divergence_.28information_gain.29)   + [3.6 Other quantities](http://en.wikipedia.org/wiki/Information_theory#Other_quantities#Other_quantities) * [4 Coding theory](http://en.wikipedia.org/wiki/Information_theory#Coding_theory#Coding_theory)   + [4.1 Source theory](http://en.wikipedia.org/wiki/Information_theory#Source_theory#Source_theory)     - [4.1.1 Rate](http://en.wikipedia.org/wiki/Information_theory#Rate#Rate)   + [4.2 Channel capacity](http://en.wikipedia.org/wiki/Information_theory#Channel_capacity#Channel_capacity)     - [4.2.1 Channel capacity of particular model channels](http://en.wikipedia.org/wiki/Information_theory#Channel_capacity_of_particular_model_channels#Channel_capacity_of_particular_model_channels) * [5 Applications to other fields](http://en.wikipedia.org/wiki/Information_theory#Applications_to_other_fields#Applications_to_other_fields)   + [5.1 Intelligence uses and secrecy applications](http://en.wikipedia.org/wiki/Information_theory#Intelligence_uses_and_secrecy_applications#Intelligence_uses_and_secrecy_applications)   + [5.2 Pseudorandom number generation](http://en.wikipedia.org/wiki/Information_theory#Pseudorandom_number_generation#Pseudorandom_number_generation)   + [5.3 Seismic Exploration](http://en.wikipedia.org/wiki/Information_theory#Seismic_Exploration#Seismic_Exploration)   + [5.4 Miscellaneous applications](http://en.wikipedia.org/wiki/Information_theory#Miscellaneous_applications#Miscellaneous_applications) * [6 References](http://en.wikipedia.org/wiki/Information_theory#References#References)   + [6.1 Footnotes](http://en.wikipedia.org/wiki/Information_theory#Footnotes#Footnotes)   + [6.2 The classic work](http://en.wikipedia.org/wiki/Information_theory#The_classic_work#The_classic_work)   + [6.3 Other journal articles](http://en.wikipedia.org/wiki/Information_theory#Other_journal_articles#Other_journal_articles)   + [6.4 Textbooks on information theory](http://en.wikipedia.org/wiki/Information_theory#Textbooks_on_information_theory#Textbooks_on_information_theory)   + [6.5 Other books](http://en.wikipedia.org/wiki/Information_theory#Other_books#Other_books) * [7 See also](http://en.wikipedia.org/wiki/Information_theory#See_also#See_also)   + [7.1 Applications](http://en.wikipedia.org/wiki/Information_theory#Applications#Applications)   + [7.2 History](http://en.wikipedia.org/wiki/Information_theory#History#History)   + [7.3 Theory](http://en.wikipedia.org/wiki/Information_theory#Theory#Theory)   + [7.4 Concepts](http://en.wikipedia.org/wiki/Information_theory#Concepts#Concepts) * [8 External links](http://en.wikipedia.org/wiki/Information_theory#External_links#External_links) |

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=1)] Overview

The main concepts of information theory can be grasped by considering the most widespread means of human communication: language. Two important aspects of a good language are as follows: First, the most common words (e.g., "a", "the", "I") should be shorter than less common words (e.g., "benefit", "generation", "mediocre"), so that sentences will not be too long. Such a tradeoff in word length is analogous to [data compression](http://en.wikipedia.org/wiki/Data_compression) and is the essential aspect of [source coding](http://en.wikipedia.org/wiki/Source_coding). Second, if part of a sentence is unheard or misheard due to noise — e.g., a passing car — the listener should still be able to glean the meaning of the underlying message. Such robustness is as essential for an electronic communication system as it is for a language; properly building such robustness into communications is done by [channel coding](http://en.wikipedia.org/wiki/Channel_capacity). Source coding and channel coding are the fundamental concerns of information theory.

Note that these concerns have nothing to do with the *importance* of messages. For example, a platitude such as "Thank you; come again" takes about as long to say or write as the urgent plea, "Call an ambulance!" while clearly the latter is more important and more meaningful. Information theory, however, does not consider message importance or meaning, as these are matters of the quality of data rather than the quantity and readability of data, the latter of which is determined solely by probabilities.

Information theory is generally considered to have been founded in 1948 by [Claude Shannon](http://en.wikipedia.org/wiki/Claude_Elwood_Shannon) in his seminal work, "[A Mathematical Theory of Communication](http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication)." The central paradigm of classical information theory is the engineering problem of the transmission of information over a noisy channel. The most fundamental results of this theory are Shannon's [source coding theorem](http://en.wikipedia.org/wiki/Source_coding_theorem), which establishes that, on average, the number of *bits* needed to represent the result of an uncertain event is given by its [entropy](http://en.wikipedia.org/wiki/Information_entropy); and Shannon's [noisy-channel coding theorem](http://en.wikipedia.org/wiki/Noisy-channel_coding_theorem), which states that *reliable* communication is possible over *noisy* channels provided that the rate of communication is below a certain threshold called the channel capacity. The channel capacity can be approached in practice by using appropriate encoding and decoding systems.

Information theory is closely associated with a collection of pure and applied disciplines that have been investigated and reduced to engineering practice under a variety of [rubrics](http://en.wikipedia.org/wiki/Rubric_(academic)) throughout the world over the past half century or more: [adaptive systems](http://en.wikipedia.org/wiki/Adaptive_system), [anticipatory systems](http://en.wikipedia.org/wiki/Anticipatory_system), [artificial intelligence](http://en.wikipedia.org/wiki/Artificial_intelligence), [complex systems](http://en.wikipedia.org/wiki/Complex_system), [complexity science](http://en.wikipedia.org/wiki/Complexity_science), [cybernetics](http://en.wikipedia.org/wiki/Cybernetics), [informatics](http://en.wikipedia.org/wiki/Informatics), [machine learning](http://en.wikipedia.org/wiki/Machine_learning), along with [systems sciences](http://en.wikipedia.org/wiki/Systems_science) of many descriptions. Information theory is a broad and deep mathematical theory, with equally broad and deep applications, amongst which is the vital field of [coding theory](http://en.wikipedia.org/wiki/Coding_theory).

Coding theory is concerned with finding explicit methods, called *codes*, of increasing the efficiency and reducing the net error rate of data communication over a noisy channel to near the limit that Shannon proved is the maximum possible for that channel. These codes can be roughly subdivided into [data compression](http://en.wikipedia.org/wiki/Data_compression) (source coding) and [error-correction](http://en.wikipedia.org/wiki/Error-correction) (channel coding) techniques. In the latter case, it took many years to find the methods Shannon's work proved were possible. A third class of information theory codes are cryptographic algorithms (both [codes](http://en.wikipedia.org/wiki/Code_(cryptography)) and [ciphers](http://en.wikipedia.org/wiki/Cipher)). Concepts, methods and results from coding theory and information theory are widely used in [cryptography](http://en.wikipedia.org/wiki/Cryptography) and [cryptanalysis](http://en.wikipedia.org/wiki/Cryptanalysis). *See the article* [*ban (information)*](http://en.wikipedia.org/wiki/Ban_(information)) *for a historical application.*

Information theory is also used in [information retrieval](http://en.wikipedia.org/wiki/Information_retrieval), [intelligence gathering](http://en.wikipedia.org/wiki/Intelligence_(information_gathering)), [gambling](http://en.wikipedia.org/wiki/Gambling), [statistics](http://en.wikipedia.org/wiki/Statistics), and even in [musical composition](http://en.wikipedia.org/wiki/Musical_composition).

As another example, consider information that needs to be transferred from point A to B. Assuming information to be a jazz band playing a particular composition. This same composition can be used as a ring tone in a mobile phone or can be played on a keyboard to a live audience. Even though the end result for the mobile phone and the keyboard are different - the underlying composition is the same. Information theory deals with this kind of information that is supposed to be transferred to point A to B.

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=2)] Historical background

Main article: [History of information theory](http://en.wikipedia.org/wiki/History_of_information_theory)

The landmark event that established the discipline of information theory, and brought it to immediate worldwide attention, was the publication of [Claude E. Shannon](http://en.wikipedia.org/wiki/Claude_E._Shannon)'s classic paper "[A Mathematical Theory of Communication](http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication)" in the [*Bell System Technical Journal*](http://en.wikipedia.org/wiki/Bell_System_Technical_Journal) in July and October of 1948.

Prior to this paper, limited information theoretic ideas had been developed at Bell Labs, all implicitly assuming events of equal probability. [Harry Nyquist](http://en.wikipedia.org/wiki/Harry_Nyquist)'s 1924 paper, *Certain Factors Affecting Telegraph Speed,* contains a theoretical section quantifying "intelligence" and the "line speed" at which it can be transmitted by a communication system, giving the relation *W* = *K*log*m*, where *W* is the speed of transmission of intelligence, *m* is the number of different voltage levels to choose from at each time step, and *K* is a constant. [Ralph Hartley](http://en.wikipedia.org/wiki/Ralph_Hartley)'s 1928 paper, *Transmission of Information,* uses the word *information* as a measurable quantity, reflecting the receiver's ability to distinguish that one sequence of symbols from any other, thus quantifying information as *H* = log*Sn* = *n*log*S*, where *S* was the number of possible symbols, and *n* the number of symbols in a transmission. The natural unit of information was therefore the decimal digit, much later renamed the [hartley](http://en.wikipedia.org/wiki/Ban_(information)) in his honour as a unit or scale or measure of information. [Alan Turing](http://en.wikipedia.org/wiki/Alan_Turing) in 1940 used similar ideas as part of the statistical analysis of the breaking of the German second world war [Enigma](http://en.wikipedia.org/wiki/Cryptanalysis_of_the_Enigma) ciphers.

Much of the mathematics behind information theory with events of different probabilities was developed for the field of [thermodynamics](http://en.wikipedia.org/wiki/Thermodynamics) by [Ludwig Boltzmann](http://en.wikipedia.org/wiki/Ludwig_Boltzmann) and [J. Willard Gibbs](http://en.wikipedia.org/wiki/J._Willard_Gibbs). Connections between information-theoretic entropy and thermodynamic entropy, including the important contributions by [Rolf Landauer](http://en.wikipedia.org/wiki/Rolf_Landauer) in the 1960s, are explored in [*Entropy in thermodynamics and information theory*](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory).

In Shannon's revolutionary and groundbreaking paper, the work for which had been substantially completed at Bell Labs by the end of 1944, Shannon for the first time introduced the qualitative and quantitative model of communication as a statistical process underlying information theory, opening with the assertion that

"The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point."

With it came the ideas of

* the [information entropy](http://en.wikipedia.org/wiki/Information_entropy) and [redundancy](http://en.wikipedia.org/wiki/Redundancy_(information_theory)) of a source, and its relevance through the [source coding theorem](http://en.wikipedia.org/wiki/Source_coding_theorem);
* the [mutual information](http://en.wikipedia.org/wiki/Mutual_information), and the [channel capacity](http://en.wikipedia.org/wiki/Channel_capacity) of a noisy channel, including the promise of perfect loss-free communication given by the [noisy-channel coding theorem](http://en.wikipedia.org/wiki/Noisy-channel_coding_theorem);
* the practical result of the [Shannon–Hartley law](http://en.wikipedia.org/wiki/Shannon%E2%80%93Hartley_law) for the channel capacity of a Gaussian channel; and of course
* the [bit](http://en.wikipedia.org/wiki/Bit)—a new way of seeing the most fundamental unit of information.

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=3)] Quantities of information

Main article: [Quantities of information](http://en.wikipedia.org/wiki/Quantities_of_information)

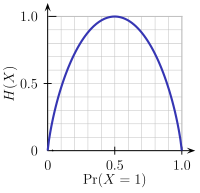
Information theory is based on [probability theory](http://en.wikipedia.org/wiki/Probability_theory) and [statistics](http://en.wikipedia.org/wiki/Statistics). The most important quantities of information are [entropy](http://en.wikipedia.org/wiki/Entropy_(information_theory)), the information in a [random variable](http://en.wikipedia.org/wiki/Random_variable), and [mutual information](http://en.wikipedia.org/wiki/Mutual_information), the amount of information in common between two random variables. The former quantity indicates how easily message data can be [compressed](http://en.wikipedia.org/wiki/Data_compression) while the latter can be used to find the communication rate across a [channel](http://en.wikipedia.org/wiki/Channel_(communications)).

The choice of logarithmic base in the following formulae determines the [unit](http://en.wikipedia.org/wiki/Units_of_measurement) of [information entropy](http://en.wikipedia.org/wiki/Information_entropy) that is used. The most common unit of information is the [bit](http://en.wikipedia.org/wiki/Bit), based on the [binary logarithm](http://en.wikipedia.org/wiki/Binary_logarithm). Other units include the [nat](http://en.wikipedia.org/wiki/Nat_(information)), which is based on the [natural logarithm](http://en.wikipedia.org/wiki/Natural_logarithm), and the [hartley](http://en.wikipedia.org/wiki/Deciban), which is based on the [common logarithm](http://en.wikipedia.org/wiki/Common_logarithm).

In what follows, an expression of the form is considered by convention to be equal to zero whenever *p* = 0. This is justified because for any logarithmic base.



### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=4)] Entropy



Entropy of a [Bernoulli trial](http://en.wikipedia.org/wiki/Bernoulli_trial) as a function of success probability, often called the [**binary entropy function**](http://en.wikipedia.org/wiki/Binary_entropy_function), *H*b(*p*). The entropy is maximized at 1 bit per trial when the two possible outcomes are equally probable, as in an unbiased coin toss.

The [**entropy**](http://en.wikipedia.org/wiki/Entropy_(information_theory)), *H*, of a discrete random variable *X* is a measure of the amount of *uncertainty* associated with the value of *X*.

Suppose one transmits 1000 bits (0s and 1s). If these bits are known ahead of transmission (to be a certain value with absolute probability), logic dictates that no information has been transmitted. If, however, each is equally and independently likely to be 0 or 1, 1000 bits (in the information theoretic sense) have been transmitted. Between these two extremes, information can be quantified as follows. If is the set of all messages *x* that *X* could be, and *p*(*x*) is the probability of *X* given *x*, then the entropy of *X* is defined:[[8]](http://en.wikipedia.org/wiki/Information_theory" \l "cite_note-Reza-7#cite_note-Reza-7" \o ")



(Here, *I*(*x*) is the [self-information](http://en.wikipedia.org/wiki/Self-information), which is the entropy contribution of an individual message, and is the [expected value](http://en.wikipedia.org/wiki/Expected_value).) An important property of entropy is that it is maximized when all the messages in the message space are equiprobable *p*(*x*) = 1 / *n*,—i.e., most unpredictable—in which case *H*(*X*) = log *n*.



The special case of information entropy for a random variable with two outcomes is the [**binary entropy function**](http://en.wikipedia.org/wiki/Binary_entropy_function):



### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=5)] Joint entropy

The [**joint entropy**](http://en.wikipedia.org/wiki/Joint_entropy) of two discrete random variables *X* and *Y* is merely the entropy of their pairing: (*X*,*Y*). This implies that if *X* and *Y* are [independent](http://en.wikipedia.org/wiki/Statistical_independence), then their joint entropy is the sum of their individual entropies.

For example, if (*X*,*Y*) represents the position of a [chess](http://en.wikipedia.org/wiki/Chess) piece — *X* the row and *Y* the column, then the joint entropy of the row of the piece and the column of the piece will be the entropy of the position of the piece.



Despite similar notation, joint entropy should not be confused with [**cross entropy**](http://en.wikipedia.org/wiki/Cross_entropy).

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=6)] Conditional entropy (equivocation)

The [**conditional entropy**](http://en.wikipedia.org/wiki/Conditional_entropy) or **conditional uncertainty** of *X* given random variable *Y* (also called the **equivocation** of *X* about *Y*) is the average conditional entropy over *Y*:[[9]](http://en.wikipedia.org/wiki/Information_theory" \l "cite_note-Ash-8#cite_note-Ash-8" \o ")



Because entropy can be conditioned on a random variable or on that random variable being a certain value, care should be taken not to confuse these two definitions of conditional entropy, the former of which is in more common use. A basic property of this form of conditional entropy is that:



### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=7)] Mutual information (transinformation)

[**Mutual information**](http://en.wikipedia.org/wiki/Mutual_information) measures the amount of information that can be obtained about one random variable by observing another. It is important in communication where it can be used to maximize the amount of information shared between sent and received signals. The mutual information of *X* relative to *Y* is given by:



where *SI* (*S*pecific mutual *I*nformation) is the [pointwise mutual information](http://en.wikipedia.org/wiki/Pointwise_mutual_information).

A basic property of the mutual information is that



That is, knowing *Y*, we can save an average of *I*(*X*;*Y*) bits in encoding *X* compared to not knowing *Y*.

Mutual information is [symmetric](http://en.wikipedia.org/wiki/Symmetric_function):



Mutual information can be expressed as the average [Kullback–Leibler divergence](http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence) (information gain) of the [posterior probability distribution](http://en.wikipedia.org/wiki/Posterior_probability) of *X* given the value of *Y* to the [prior distribution](http://en.wikipedia.org/wiki/Prior_probability) on *X*:



In other words, this is a measure of how much, on the average, the probability distribution on *X* will change if we are given the value of *Y*. This is often recalculated as the divergence from the product of the marginal distributions to the actual joint distribution:



Mutual information is closely related to the [log-likelihood ratio test](http://en.wikipedia.org/wiki/Likelihood-ratio_test) in the context of contingency tables and the [multinomial distribution](http://en.wikipedia.org/wiki/Multinomial_distribution) and to [Pearson's χ2 test](http://en.wikipedia.org/wiki/Pearson%27s_chi-square_test): mutual information can be considered a statistic for assessing independence between a pair of variables, and has a well-specified asymptotic distribution.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=8)] Kullback–Leibler divergence (information gain)

The [**Kullback–Leibler divergence**](http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence) (or **information divergence**, **information gain**, or **relative entropy**) is a way of comparing two distributions: a "true" [probability distribution](http://en.wikipedia.org/wiki/Probability_distribution) *p(X)*, and an arbitrary probability distribution *q(X)*. If we compress data in a manner that assumes *q(X)* is the distribution underlying some data, when, in reality, *p(X)* is the correct distribution, the Kullback–Leibler divergence is the number of average additional bits per datum necessary for compression. It is thus defined



Although it is sometimes used as a 'distance metric', it is not a true [metric](http://en.wikipedia.org/wiki/Metric_(mathematics)) since it is not symmetric and does not satisfy the [triangle inequality](http://en.wikipedia.org/wiki/Triangle_inequality) (making it a semi-quasimetric).

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=9)] Other quantities

Other important information theoretic quantities include [Rényi entropy](http://en.wikipedia.org/wiki/R%C3%A9nyi_entropy), (a generalization of entropy,) [differential entropy](http://en.wikipedia.org/wiki/Differential_entropy), (a generalization of quantities of information to continuous distributions,) and the [conditional mutual information](http://en.wikipedia.org/wiki/Conditional_mutual_information).

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=10)] Coding theory

Main article: [Coding theory](http://en.wikipedia.org/wiki/Coding_theory)



A picture showing scratches on the readable surface of a CD-R. Music and data CDs are coded using error correcting codes and thus can still be read even if they have minor scratches using [error detection and correction](http://en.wikipedia.org/wiki/Error_detection_and_correction).

[Coding theory](http://en.wikipedia.org/wiki/Coding_theory) is one of the most important and direct applications of information theory. It can be subdivided into [source coding](http://en.wikipedia.org/wiki/Data_compression) theory and [channel coding](http://en.wikipedia.org/wiki/Error_correction) theory. Using a statistical description for data, information theory quantifies the number of bits needed to describe the data, which is the information entropy of the source.

* Data compression (source coding): There are two formulations for the compression problem:

1. [lossless data compression](http://en.wikipedia.org/wiki/Lossless_data_compression): the data must be reconstructed exactly;
2. [lossy data compression](http://en.wikipedia.org/wiki/Lossy_data_compression): allocates bits needed to reconstruct the data, within a specified fidelity level measured by a distortion function. This subset of Information theory is called [rate–distortion theory](http://en.wikipedia.org/wiki/Rate%E2%80%93distortion_theory).

* Error-correcting codes (channel coding): While data compression removes as much [redundancy](http://en.wikipedia.org/wiki/Redundancy_(information_theory)) as possible, an error correcting code adds just the right kind of redundancy (i.e., [error correction](http://en.wikipedia.org/wiki/Error_correction)) needed to transmit the data efficiently and faithfully across a noisy channel.

This division of coding theory into compression and transmission is justified by the information transmission theorems, or source–channel separation theorems that justify the use of bits as the universal currency for information in many contexts. However, these theorems only hold in the situation where one transmitting user wishes to communicate to one receiving user. In scenarios with more than one transmitter (the multiple-access channel), more than one receiver (the [broadcast channel](http://en.wikipedia.org/w/index.php?title=Broadcast_channel&action=edit&redlink=1)) or intermediary "helpers" (the [relay channel](http://en.wikipedia.org/wiki/Relay_channel)), or more general [networks](http://en.wikipedia.org/wiki/Computer_network), compression followed by transmission may no longer be optimal. [Network information theory](http://en.wikipedia.org/w/index.php?title=Network_information_theory&action=edit&redlink=1) refers to these multi-agent communication models.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=11)] Source theory

Any process that generates successive messages can be considered a [**source**](http://en.wikipedia.org/wiki/Communication_source) of information. A memoryless source is one in which each message is an [independent identically-distributed random variable](http://en.wikipedia.org/wiki/Independent_identically-distributed_random_variables), whereas the properties of [ergodicity](http://en.wikipedia.org/wiki/Ergodic_theory) and [stationarity](http://en.wikipedia.org/wiki/Stationary_process) impose more general constraints. All such sources are [stochastic](http://en.wikipedia.org/wiki/Stochastic_process). These terms are well studied in their own right outside information theory.

#### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=12)] Rate

Information [**rate**](http://en.wikipedia.org/wiki/Entropy_rate) is the average entropy per symbol. For memoryless sources, this is merely the entropy of each symbol, while, in the case of a stationary stochastic process, it is



that is, the conditional entropy of a symbol given all the previous symbols generated. For the more general case of a process that is not necessarily stationary, the *average rate* is



that is, the limit of the joint entropy per symbol. For stationary sources, these two expressions give the same result.[[10]](http://en.wikipedia.org/wiki/Information_theory#cite_note-9#cite_note-9)

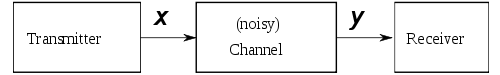
It is common in information theory to speak of the "rate" or "entropy" of a language. This is appropriate, for example, when the source of information is English prose. The rate of a source of information is related to its [redundancy](http://en.wikipedia.org/wiki/Redundancy_(information_theory)) and how well it can be [compressed](http://en.wikipedia.org/wiki/Data_compression), the subject of **source coding**.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=13)] Channel capacity

Main article: [Noisy channel coding theorem](http://en.wikipedia.org/wiki/Noisy_channel_coding_theorem)

Communications over a channel—such as an [ethernet](http://en.wikipedia.org/wiki/Ethernet) wire—is the primary motivation of information theory. As anyone who's ever used a telephone (mobile or landline) knows, however, such channels often fail to produce exact reconstruction of a signal; noise, periods of silence, and other forms of signal corruption often degrade quality. How much information can one hope to communicate over a noisy (or otherwise imperfect) channel?

Consider the communications process over a discrete channel. A simple model of the process is shown below:



Here *X* represents the space of messages transmitted, and *Y* the space of messages received during a unit time over our channel. Let *p*(*y* | *x*) be the [conditional probability](http://en.wikipedia.org/wiki/Conditional_probability) distribution function of *Y* given *X*. We will consider *p*(*y* | *x*) to be an inherent fixed property of our communications channel (representing the nature of the [**noise**](http://en.wikipedia.org/wiki/Signal_noise) of our channel). Then the joint distribution of *X* and *Y* is completely determined by our channel and by our choice of *f*(*x*), the marginal distribution of messages we choose to send over the channel. Under these constraints, we would like to maximize the rate of information, or the [**signal**](http://en.wikipedia.org/wiki/Signal_(electrical_engineering)), we can communicate over the channel. The appropriate measure for this is the [mutual information](http://en.wikipedia.org/wiki/Mutual_information), and this maximum mutual information is called the [**channel capacity**](http://en.wikipedia.org/wiki/Channel_capacity) and is given by:

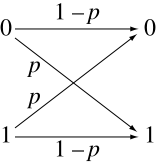


This capacity has the following property related to communicating at information rate *R* (where *R* is usually bits per symbol). For any information rate *R < C* and coding error ε > 0, for large enough *N*, there exists a code of length *N* and rate ≥ R and a decoding algorithm, such that the maximal probability of block error is ≤ ε; that is, it is always possible to transmit with arbitrarily small block error. In addition, for any rate *R > C*, it is impossible to transmit with arbitrarily small block error.

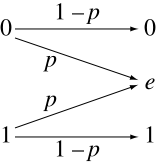
[**Channel coding**](http://en.wikipedia.org/wiki/Channel_code) is concerned with finding such nearly optimal [codes](http://en.wikipedia.org/wiki/Error_detection_and_correction) that can be used to transmit data over a noisy channel with a small coding error at a rate near the channel capacity.

#### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=14)] Channel capacity of particular model channels

* A continuous-time analog communications channel subject to Gaussian noise — see [Shannon–Hartley theorem](http://en.wikipedia.org/wiki/Shannon%E2%80%93Hartley_theorem).
* A [binary symmetric channel](http://en.wikipedia.org/wiki/Binary_symmetric_channel) (BSC) with crossover probability *p* is a binary input, binary output channel that flips the input bit with probability *p*. The BSC has a capacity of 1 − *H*b(*p*) bits per channel use, where *H*b is the [binary entropy function](http://en.wikipedia.org/wiki/Binary_entropy_function):



* A binary erasure channel (BEC) with erasure probability *p* is a binary input, ternary output channel. The possible channel outputs are *0*, *1*, and a third symbol 'e' called an erasure. The erasure represents complete loss of information about an input bit. The capacity of the BEC is *1 - p* bits per channel use.



## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=15)] Applications to other fields

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=16)] Intelligence uses and secrecy applications

Information theoretic concepts apply to [cryptography](http://en.wikipedia.org/wiki/Cryptography) and [cryptanalysis](http://en.wikipedia.org/wiki/Cryptanalysis). [Turing](http://en.wikipedia.org/wiki/Turing)'s information unit, the [ban](http://en.wikipedia.org/wiki/Ban_(information)), was used in the [Ultra](http://en.wikipedia.org/wiki/Ultra) project, breaking the German [Enigma machine](http://en.wikipedia.org/wiki/Enigma_machine) code and hastening the [end of WWII in Europe](http://en.wikipedia.org/wiki/Victory_in_Europe_Day). Shannon himself defined an important concept now called the [unicity distance](http://en.wikipedia.org/wiki/Unicity_distance). Based on the [redundancy](http://en.wikipedia.org/wiki/Redundancy_(information_theory)) of the [plaintext](http://en.wikipedia.org/wiki/Plaintext), it attempts to give a minimum amount of [ciphertext](http://en.wikipedia.org/wiki/Ciphertext) necessary to ensure unique decipherability.

Information theory leads us to believe it is much more difficult to keep secrets than it might first appear. A [brute force attack](http://en.wikipedia.org/wiki/Brute_force_attack) can break systems based on [asymmetric key algorithms](http://en.wikipedia.org/wiki/Public-key_cryptography) or on most commonly used methods of [symmetric key algorithms](http://en.wikipedia.org/wiki/Symmetric-key_algorithm) (sometimes called secret key algorithms), such as [block ciphers](http://en.wikipedia.org/wiki/Block_cipher). The security of all such methods currently comes from the assumption that no known attack can break them in a practical amount of time.

[Information theoretic security](http://en.wikipedia.org/wiki/Information_theoretic_security) refers to methods such as the [one-time pad](http://en.wikipedia.org/wiki/One-time_pad) that are not vulnerable to such brute force attacks. In such cases, the positive conditional [mutual information](http://en.wikipedia.org/wiki/Mutual_information) between the [plaintext](http://en.wikipedia.org/wiki/Plaintext) and [ciphertext](http://en.wikipedia.org/wiki/Ciphertext) (conditioned on the [key](http://en.wikipedia.org/wiki/Key_(cryptography))) can ensure proper transmission, while the unconditional mutual information between the plaintext and ciphertext remains zero, resulting in absolutely secure communications. In other words, an eavesdropper would not be able to improve his or her guess of the plaintext by gaining knowledge of the ciphertext but not of the key. However, as in any other cryptographic system, care must be used to correctly apply even information-theoretically secure methods; the [Venona project](http://en.wikipedia.org/wiki/Venona_project) was able to crack the one-time pads of the [Soviet Union](http://en.wikipedia.org/wiki/Soviet_Union) due to their improper reuse of key material.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=17)] Pseudorandom number generation

[Pseudorandom number generators](http://en.wikipedia.org/wiki/Pseudorandom_number_generator) are widely available in computer language libraries and application programs. They are, almost universally, unsuited to cryptographic use as they do not evade the deterministic nature of modern computer equipment and software. A class of improved random number generators is termed [Cryptographically secure pseudorandom number generators](http://en.wikipedia.org/wiki/Cryptographically_secure_pseudorandom_number_generator), but even they require external to the software [random seeds](http://en.wikipedia.org/wiki/Random_seed) to work as intended. These can be obtained via [extractors](http://en.wikipedia.org/wiki/Extractor), if done carefully. The measure of sufficient randomness in extractors is [min-entropy](http://en.wikipedia.org/wiki/Min-entropy), a value related to Shannon entropy through [Rényi entropy](http://en.wikipedia.org/wiki/R%C3%A9nyi_entropy); Rényi entropy is also used in evaluating randomness in cryptographic systems. Although related, the distinctions among these measures mean that a [random variable](http://en.wikipedia.org/wiki/Random_variable) with high Shannon entropy is not necessarily satisfactory for use in an extractor and so for cryptography uses.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=18)] Seismic Exploration

One early commercial application of information theory was in the field seismic oil exploration. Work in this field made it possible to strip off and separate the unwanted noise from the desired seismic signal. Information theory and [digital signal processing](http://en.wikipedia.org/wiki/Digital_signal_processing) offer a major improvement of resolution and image clarity over previous analog methods.[[11]](http://en.wikipedia.org/wiki/Information_theory#cite_note-10#cite_note-10)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=19)] Miscellaneous applications

Information theory also has applications in [gambling and investing](http://en.wikipedia.org/wiki/Gambling_and_information_theory), [black holes](http://en.wikipedia.org/wiki/Black_hole_information_paradox), [bioinformatics](http://en.wikipedia.org/wiki/Bioinformatics), and [music](http://en.wikipedia.org/wiki/Music).

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=20)] References

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=21)] Footnotes

1. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-0#cite_ref-0) F. Rieke, D. Warland, R Ruyter van Steveninck, W Bialek, Spikes: Exploring the Neural Code. The MIT press (1997).
2. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-1#cite_ref-1) cf. Huelsenbeck, J. P., F. Ronquist, R. Nielsen and J. P. Bollback (2001) Bayesian inference of phylogeny and its impact on evolutionary biology, *Science* **294**:2310-2314
3. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-2#cite_ref-2) Rando Allikmets, Wyeth W. Wasserman, Amy Hutchinson, Philip Smallwood, Jeremy Nathans, Peter K. Rogan, [Thomas D. Schneider](http://www.lecb.ncifcrf.gov/~toms/), Michael Dean (1998) Organization of the ABCR gene: analysis of promoter and splice junction sequences, *Gene* **215**:1, 111-122
4. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-3#cite_ref-3) Burnham, K. P. and Anderson D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition* (Springer Science, New York) [ISBN 978-0-387-95364-9](http://en.wikipedia.org/wiki/Special:BookSources/9780387953649).
5. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-4#cite_ref-4) Jaynes, E. T. (1957) [Information Theory and Statistical Mechanics](http://bayes.wustl.edu/), *Phys. Rev.* **106**:620
6. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-5#cite_ref-5) Charles H. Bennett, Ming Li, and Bin Ma (2003) [Chain Letters and Evolutionary Histories](http://sciamdigital.com/index.cfm?fa=Products.ViewIssuePreview&ARTICLEID_CHAR=08B64096-0772-4904-9D48227D5C9FAC75), *Scientific American* **288**:6, 76-81
7. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-6#cite_ref-6) David R. Anderson (November 1, 2003). ["Some background on why people in the empirical sciences may want to better understand the information-theoretic methods"](http://www.jyu.fi/science/laitokset/bioenv/en/coeevolution/events/itms/why) (pdf). <http://www.jyu.fi/science/laitokset/bioenv/en/coeevolution/events/itms/why>. Retrieved on 2007-12-30.
8. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-Reza_7-0#cite_ref-Reza_7-0) Fazlollah M. Reza (1961, 1994). [*An Introduction to Information Theory*](http://books.google.com/books?id=RtzpRAiX6OgC&pg=PA8&dq=intitle:%22An+Introduction+to+Information+Theory%22++%22entropy+of+a+simple+source%22&as_brr=0&ei=zP79Ro7UBovqoQK4g_nCCw&sig=j3lPgyYrC3-bvn1Td42TZgTzj0Q). Dover Publications, Inc., New York. [ISBN 0-486-68210-2](http://en.wikipedia.org/wiki/Special:BookSources/0486682102). <http://books.google.com/books?id=RtzpRAiX6OgC&pg=PA8&dq=intitle:%22An+Introduction+to+Information+Theory%22++%22entropy+of+a+simple+source%22&as_brr=0&ei=zP79Ro7UBovqoQK4g_nCCw&sig=j3lPgyYrC3-bvn1Td42TZgTzj0Q>.
9. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-Ash_8-0#cite_ref-Ash_8-0) Robert B. Ash (1965, 1990). [*Information Theory*](http://books.google.com/books?id=ngZhvUfF0UIC&pg=PA16&dq=intitle:information+intitle:theory+inauthor:ash+conditional+uncertainty&as_brr=0&ei=kKwNR4rbH5mepgKB4d2zBg&sig=YAsiCEVISjJ484R3uGoXpi-a5rI). Dover Publications, Inc.. [ISBN 0-486-66521-6](http://en.wikipedia.org/wiki/Special:BookSources/0486665216). <http://books.google.com/books?id=ngZhvUfF0UIC&pg=PA16&dq=intitle:information+intitle:theory+inauthor:ash+conditional+uncertainty&as_brr=0&ei=kKwNR4rbH5mepgKB4d2zBg&sig=YAsiCEVISjJ484R3uGoXpi-a5rI>.
10. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-9#cite_ref-9) Jerry D. Gibson (1998). [*Digital Compression for Multimedia: Principles and Standards*](http://books.google.com/books?id=aqQ2Ry6spu0C&pg=PA56&dq=entropy-rate+conditional&as_brr=3&ei=YGDsRtzGGKjupQKa2L2xDw&sig=o0UCtf0xZOf11lPIexPrjOKPgNc#PPA57,M1). Morgan Kaufmann. [ISBN 1558603697](http://en.wikipedia.org/wiki/Special:BookSources/1558603697). <http://books.google.com/books?id=aqQ2Ry6spu0C&pg=PA56&dq=entropy-rate+conditional&as_brr=3&ei=YGDsRtzGGKjupQKa2L2xDw&sig=o0UCtf0xZOf11lPIexPrjOKPgNc#PPA57,M1>.
11. [**^**](http://en.wikipedia.org/wiki/Information_theory#cite_ref-10#cite_ref-10) The Corporation and Innovation, Haggerty, Patrick, Strategic Management Journal, Vol. 2, 97-118 (1981)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=22)] The classic work

* [Shannon, C.E.](http://en.wikipedia.org/wiki/Claude_Elwood_Shannon) (1948), "[A Mathematical Theory of Communication](http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication)", *Bell System Technical Journal*, 27, pp. 379–423 & 623–656, July & October, 1948. [PDF.](http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf)  
  [Notes and other formats.](http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html)
* R.V.L. Hartley, ["Transmission of Information"](http://www.dotrose.com/etext/90_Miscellaneous/transmission_of_information_1928b.pdf), *Bell System Technical Journal*, July 1928
* [Andrey Kolmogorov](http://en.wikipedia.org/wiki/Andrey_Kolmogorov)(1968) "Three approaches to the quantitative definition of information" in International Journal of Computer Mathematics.

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=23)] Other journal articles

* J. L. Kelly, Jr., "[A New Interpretation of Information Rate](http://www.arbtrading.com/reports/kelly.pdf)," *Bell System Technical Journal*, Vol. 35, July 1956, pp. 917-26.
* R. Landauer, [Information is Physical](http://ieeexplore.ieee.org/search/wrapper.jsp?arnumber=615478) *Proc. Workshop on Physics and Computation PhysComp'92* (IEEE Comp. Sci.Press, Los Alamitos, 1993) pp. 1-4.
* R. Landauer, "[Irreversibility and Heat Generation in the Computing Process](http://www.research.ibm.com/journal/rd/441/landauerii.pdf)" *IBM J. Res. Develop.* Vol. 5, No. 3, 1961

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=24)] Textbooks on information theory

* [Claude E. Shannon](http://en.wikipedia.org/wiki/Claude_E._Shannon), Warren Weaver. *The Mathematical Theory of Communication.* Univ of Illinois Press, 1949. [ISBN 0-252-72548-4](http://en.wikipedia.org/wiki/Special:BookSources/0252725484)
* [Robert Gallager](http://en.wikipedia.org/wiki/Robert_Gallager). *Information Theory and Reliable Communication.* New York: John Wiley and Sons, 1968. [ISBN 0-471-29048-3](http://en.wikipedia.org/wiki/Special:BookSources/0471290483)
* Robert B. Ash. *Information Theory*. New York: Interscience, 1965. [ISBN 0-470-03445-9](http://en.wikipedia.org/wiki/Special:BookSources/0470034459). New York: Dover 1990. [ISBN 0-486-66521-6](http://en.wikipedia.org/wiki/Special:BookSources/0486665216)
* [Thomas M. Cover](http://en.wikipedia.org/wiki/Thomas_M._Cover), Joy A. Thomas. *Elements of information theory*, 1st Edition. New York: Wiley-Interscience, 1991. [ISBN 0-471-06259-6](http://en.wikipedia.org/wiki/Special:BookSources/0471062596).

2nd Edition. New York: Wiley-Interscience, 2006. [ISBN 0-471-24195-4](http://en.wikipedia.org/wiki/Special:BookSources/0471241954).

* [Imre Csiszar](http://en.wikipedia.org/wiki/Imre_Csiszar), Janos Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems* Akademiai Kiado: 2nd edition, 1997. [ISBN 9630574403](http://en.wikipedia.org/wiki/Special:BookSources/9630574403)
* Raymond W. Yeung. [*A First Course in Information Theory*](http://iest2.ie.cuhk.edu.hk/~whyeung/book/) Kluwer Academic/Plenum Publishers, 2002. [ISBN 0-306-46791-7](http://en.wikipedia.org/wiki/Special:BookSources/0306467917)
* David J. C. MacKay. [*Information Theory, Inference, and Learning Algorithms*](http://www.inference.phy.cam.ac.uk/mackay/itila/book.html) Cambridge: Cambridge University Press, 2003. [ISBN 0-521-64298-1](http://en.wikipedia.org/wiki/Special:BookSources/0521642981)
* Raymond W. Yeung. [*Information Theory and Network Coding*](http://iest2.ie.cuhk.edu.hk/~whyeung/book2/) Springer 2008, 2002. [ISBN 978-0-387-79233-0](http://en.wikipedia.org/wiki/Special:BookSources/9780387792330)
* Stanford Goldman. *Information Theory*. New York: Prentice Hall, 1953. New York: Dover 1968 [ISBN 0-486-62209-6](http://en.wikipedia.org/wiki/Special:BookSources/0486622096), 2005 [ISBN 0-486-44271-3](http://en.wikipedia.org/wiki/Special:BookSources/0486442713)
* [Fazlollah Reza](http://en.wikipedia.org/wiki/Fazlollah_Reza). *An Introduction to Information Theory*. New York: McGraw-Hill 1961. New York: Dover 1994. [ISBN 0-486-68210-2](http://en.wikipedia.org/wiki/Special:BookSources/0486682102)
* Masud Mansuripur. *Introduction to Information Theory*. New York: Prentice Hall, 1987. [ISBN 0-13-484668-0](http://en.wikipedia.org/wiki/Special:BookSources/0134846680)
* Christoph Arndt: *Information Measures, Information and its Description in Science and Engineering* (Springer Series: Signals and Communication Technology), 2004, [ISBN 978-3-540-40855-0](http://en.wikipedia.org/wiki/Special:BookSources/9783540408550), [[1]](http://www.springer.com/east/home?SGWID=5-102-22-17328941-0&changeHeader=true&referer=www.springeronline.com&SHORTCUT=www.springer.com/east/3-540-40855-X);

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=25)] Other books

* Leon Brillouin, *Science and Information Theory*, Mineola, N.Y.: Dover, [1956, 1962] 2004. [ISBN 0-486-43918-6](http://en.wikipedia.org/wiki/Special:BookSources/0486439186)
* A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York: Dover, 1957. [ISBN 0-486-60434-9](http://en.wikipedia.org/wiki/Special:BookSources/0486604349)
* H. S. Leff and A. F. Rex, Editors, *Maxwell's Demon: Entropy, Information, Computing*, Princeton University Press, Princeton, NJ (1990). [ISBN 0-691-08727-X](http://en.wikipedia.org/wiki/Special:BookSources/069108727X)
* Tom Siegfried, *The Bit and the Pendulum*, Wiley, 2000. [ISBN 0-471-32174-5](http://en.wikipedia.org/wiki/Special:BookSources/0471321745)
* Charles Seife, *Decoding The Universe*, Viking, 2006. [ISBN 0-670-03441-X](http://en.wikipedia.org/wiki/Special:BookSources/067003441X)
* Jeremy Campbell, *Grammatical Man*, Touchstone/Simon & Schuster, 1982, [ISBN 0-671-44062-4](http://en.wikipedia.org/wiki/Special:BookSources/0671440624)
* Henri Theil, *Economics and Information Theory*, Rand McNally & Company - Chicago, 1967.

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=26)] See also

|  |  |
| --- | --- |
|  | [***Mathematics portal***](http://en.wikipedia.org/wiki/Portal:Mathematics) |

* [Communication theory](http://en.wikipedia.org/wiki/Communication_theory)
* [List of important publications](http://en.wikipedia.org/wiki/List_of_important_publications_in_computer_science#Information_theory)
* [Philosophy of information](http://en.wikipedia.org/wiki/Philosophy_of_information)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=27)] Applications

* [Cryptography](http://en.wikipedia.org/wiki/Cryptography)
* [Cryptanalysis](http://en.wikipedia.org/wiki/Cryptanalysis)
* [Entropy in thermodynamics and information theory](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory)
* [seismic exploration](http://en.wikipedia.org/wiki/Reflection_seismology)
* [Intelligence (information gathering)](http://en.wikipedia.org/wiki/Intelligence_(information_gathering))
* [Gambling](http://en.wikipedia.org/wiki/Gambling)
* [Cybernetics](http://en.wikipedia.org/wiki/Cybernetics)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=28)] History

* [History of information theory](http://en.wikipedia.org/wiki/History_of_information_theory)
* [Timeline of information theory](http://en.wikipedia.org/wiki/Timeline_of_information_theory)
* [Shannon, C.E.](http://en.wikipedia.org/wiki/Claude_Elwood_Shannon)
* [Hartley, R.V.L.](http://en.wikipedia.org/wiki/Ralph_Hartley)
* [Yockey, H.P.](http://en.wikipedia.org/wiki/Hubert_Yockey)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=29)] Theory

* [Coding theory](http://en.wikipedia.org/wiki/Coding_theory)
* [Source coding](http://en.wikipedia.org/wiki/Source_coding)
* [Detection theory](http://en.wikipedia.org/wiki/Detection_theory)
* [Estimation theory](http://en.wikipedia.org/wiki/Estimation_theory)
* [Fisher information](http://en.wikipedia.org/wiki/Fisher_information)
* [Kolmogorov complexity](http://en.wikipedia.org/wiki/Kolmogorov_complexity)
* [Information Algebra](http://en.wikipedia.org/wiki/Information_Algebra)
* [Information geometry](http://en.wikipedia.org/wiki/Information_geometry)
* [Information theory and measure theory](http://en.wikipedia.org/wiki/Information_theory_and_measure_theory)
* [Logic of information](http://en.wikipedia.org/wiki/Logic_of_information)
* [Network coding](http://en.wikipedia.org/wiki/Network_coding)
* [Quantum information science](http://en.wikipedia.org/wiki/Quantum_information_science)
* [Semiotic information theory](http://en.wikipedia.org/wiki/Semiotic_information_theory)
* [Philosophy of Information](http://en.wikipedia.org/wiki/Philosophy_of_Information)

### [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=30)] Concepts

* [Self-information](http://en.wikipedia.org/wiki/Self-information)
* [Information entropy](http://en.wikipedia.org/wiki/Information_entropy)
* [Joint entropy](http://en.wikipedia.org/wiki/Joint_entropy)
* [Conditional entropy](http://en.wikipedia.org/wiki/Conditional_entropy)
* [Redundancy](http://en.wikipedia.org/wiki/Redundancy_(information_theory))
* [Channel (communications)](http://en.wikipedia.org/wiki/Channel_(communications))
* [Communication source](http://en.wikipedia.org/wiki/Communication_source)
* [Receiver (information theory)](http://en.wikipedia.org/wiki/Receiver_(information_theory))
* [Rényi entropy](http://en.wikipedia.org/wiki/R%C3%A9nyi_entropy)
* [Variety](http://en.wikipedia.org/wiki/Variety_(cybernetics))
* [Mutual information](http://en.wikipedia.org/wiki/Mutual_information)
* [Pointwise Mutual Information](http://en.wikipedia.org/wiki/Pointwise_Mutual_Information) (PMI)
* [Differential entropy](http://en.wikipedia.org/wiki/Differential_entropy)
* [Kullback-Leibler divergence](http://en.wikipedia.org/wiki/Kullback-Leibler_divergence)
* [Channel capacity](http://en.wikipedia.org/wiki/Channel_capacity)
* [Unicity distance](http://en.wikipedia.org/wiki/Unicity_distance)
* [ban (information)](http://en.wikipedia.org/wiki/Ban_(information))
* [Covert channel](http://en.wikipedia.org/wiki/Covert_channel)
* [Encoder](http://en.wikipedia.org/wiki/Encoder)
* [Decoder](http://en.wikipedia.org/wiki/Decoder)

## [[edit](http://en.wikipedia.org/w/index.php?title=Information_theory&action=edit&section=31)] External links

* Schneider, T., "Information Theory Primer", [Eprint](http://www.lecb.ncifcrf.gov/~toms/paper/primer)
* Srinivasa, S. "A Review on Multivariate Mutual Information" [PDF](http://www.nd.edu/~jnl/ee80653/tutorials/sunil.pdf).
* Challis, J. [Lateral Thinking in Information Retrieval](http://www.conceptsearching.com/Web/home/technology/lateral-thinking.aspx)
* [Journal of Chemical Education, *Shuffled Cards, Messy Desks, and Disorderly Dorm Rooms - Examples of Entropy Increase? Nonsense!*](http://jchemed.chem.wisc.edu/Journal/Issues/1999/Oct/abs1385.html)
* [IEEE Information Theory Society](http://www.itsoc.org/index.html) and [the review articles](http://www.itsoc.org/review.html).
* [On-line textbook: Information Theory, Inference, and Learning Algorithms](http://www.inference.phy.cam.ac.uk/mackay/itila/), by [David MacKay](http://en.wikipedia.org/wiki/David_MacKay_(scientist)) - gives an entertaining and thorough introduction to Shannon theory, including state-of-the-art methods from coding theory, such as [arithmetic coding](http://en.wikipedia.org/wiki/Arithmetic_coding), [low-density parity-check codes](http://en.wikipedia.org/wiki/Low-density_parity-check_code), and [Turbo codes](http://en.wikipedia.org/wiki/Turbo_code).

# Entropy in thermodynamics and information theory

### From Wikipedia, the free encyclopedia

There are close parallels between the mathematical expressions for the thermodynamic [entropy](http://en.wikipedia.org/wiki/Entropy), usually denoted by *S*, of a physical system in the [statistical thermodynamics](http://en.wikipedia.org/wiki/Statistical_thermodynamics) established by [Ludwig Boltzmann](http://en.wikipedia.org/wiki/Ludwig_Boltzmann) and [J. Willard Gibbs](http://en.wikipedia.org/wiki/J._Willard_Gibbs) in the 1870s; and the [information-theoretic entropy](http://en.wikipedia.org/wiki/Information_entropy), usually expressed as *H*, of [Claude Shannon](http://en.wikipedia.org/wiki/Claude_Elwood_Shannon) and [Ralph Hartley](http://en.wikipedia.org/wiki/Ralph_Hartley) developed in the 1940s. Shannon, although not initially aware of this similarity, commented on it upon publicizing information theory in [*A Mathematical Theory of Communication*](http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication).

This article explores what links there are between the two concepts, and how far they can be regarded as connected.

|  |
| --- |
| Contents [[hide](javascript:toggleToc())]   * [1 Equivalence of form of the defining expressions](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Equivalence_of_form_of_the_defining_expressions#Equivalence_of_form_of_the_defining_expressions)   + [1.1 Discrete case](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Discrete_case#Discrete_case)   + [1.2 Continuous case](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Continuous_case#Continuous_case) * [2 Theoretical relationship](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Theoretical_relationship#Theoretical_relationship) * [3 Information is physical](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Information_is_physical#Information_is_physical)   + [3.1 Szilard's engine](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Szilard.27s_engine#Szilard.27s_engine)   + [3.2 Landauer's principle](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Landauer.27s_principle#Landauer.27s_principle) * [4 Negentropy](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Negentropy#Negentropy) * [5 Black holes](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Black_holes#Black_holes) * [6 Quantum theory](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Quantum_theory#Quantum_theory) * [7 The fluctuation theorem](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#The_fluctuation_theorem#The_fluctuation_theorem) * [8 Topics of recent research](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Topics_of_recent_research#Topics_of_recent_research)   + [8.1 Is information quantized?](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#Is_information_quantized.3F#Is_information_quantized.3F) * [9 See also](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#See_also#See_also) * [10 External links](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#External_links#External_links) * [11 References](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#References#References) |

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=1)] Equivalence of form of the defining expressions

### [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=2)] Discrete case



The defining expression for [entropy](http://en.wikipedia.org/wiki/Entropy) in the theory of [statistical mechanics](http://en.wikipedia.org/wiki/Statistical_mechanics) established by [Ludwig Boltzmann](http://en.wikipedia.org/wiki/Ludwig_Boltzmann) and [J. Willard Gibbs](http://en.wikipedia.org/wiki/J._Willard_Gibbs) in the 1870s, is of the form:



where *pi* is the probability of the [microstate](http://en.wikipedia.org/wiki/Microstate_(statistical_mechanics)) *i* taken from an equilibrium ensemble.

The defining expression for [entropy](http://en.wikipedia.org/wiki/Information_entropy) in the theory of [information](http://en.wikipedia.org/wiki/Information_theory) established by [Claude E. Shannon](http://en.wikipedia.org/wiki/Claude_E._Shannon) in 1948 is of the form:



where *pi* is the probability of the message *mi* taken from the message space *M*.

Mathematically *H* may also be seen as an [average information](http://en.wikipedia.org/wiki/Average_information), taken over the message space, because when a certain message occurs with probability *pi*, the information −log(*pi*) will be obtained.

If all the microstates are equiprobable (a [microcanonical ensemble](http://en.wikipedia.org/wiki/Microcanonical_ensemble)), the statistical thermodynamic entropy reduces to the form on Boltzmann's tombstone,



where *W* is the number of microstates.

If all the messages are equiprobable, the information entropy reduces to the [Hartley entropy](http://en.wikipedia.org/wiki/Hartley_entropy)



where | *M* | is the [cardinality](http://en.wikipedia.org/wiki/Cardinality) of the message space *M*.

The logarithm in the thermodynamic definition is the [natural logarithm](http://en.wikipedia.org/wiki/Natural_logarithm). It can be shown that the Gibbs entropy formula, with the natural logarithm, reproduces all of the properties of the macroscopic [classical thermodynamics](http://en.wikipedia.org/wiki/Classical_thermodynamics) of [Clausius](http://en.wikipedia.org/wiki/Clausius). (See article: [Entropy (statistical views)](http://en.wikipedia.org/wiki/Entropy_(statistical_views))).

The [logarithm](http://en.wikipedia.org/wiki/Logarithm) can also be taken to the natural base in the case of information entropy. This is equivalent to choosing to measure information in [nats](http://en.wikipedia.org/wiki/Nat_(information)) instead of the usual [bits](http://en.wikipedia.org/wiki/Bits). In practice, information entropy is almost always calculated using base 2 logarithms, but this distinction amounts to nothing other than a change in units. One nat is about 1.44 bits.

The presence of [Boltzmann's constant](http://en.wikipedia.org/wiki/Boltzmann%27s_constant) *k* in the thermodynamic definitions is a historical accident, reflecting the conventional units of temperature. It is there to make sure that the statistical definition of thermodynamic entropy matches the classical entropy of Clausius, thermodynamically conjugate to [temperature](http://en.wikipedia.org/wiki/Temperature). For a simple compressible system that can only perform volume work, the [first law of thermodynamics](http://en.wikipedia.org/wiki/First_law_of_thermodynamics) becomes



But one can equally well write this equation in terms of what physicists and chemists sometimes call the 'reduced' or dimensionless entropy, σ = *S*/*k*, so that



Just as *S* is conjugate to *T*, so σ is conjugate to *kT* (the energy that is characteristic of *T* on a molecular scale).

### [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=3)] Continuous case

The most obvious extension of the Shannon entropy is the [differential entropy](http://en.wikipedia.org/wiki/Differential_entropy),



As long as *f*(*x*) is a probability density function, p. d. f., *H* repesents the [average information](http://en.wikipedia.org/wiki/Average_information) ([entropy](http://en.wikipedia.org/wiki/Entropy), [disorder](http://en.wikipedia.org/wiki/Disorder), diversity etcetera) of *f*(*x*). For any uniform p. d. f. *f*(*x*), the exponential of *H* is the volume covered by *f*(*x*) (in analogy to the [cardinality](http://en.wikipedia.org/wiki/Cardinality) in the discrete case). The volume covered by a n-dimensional [multivariate Gaussian distribution](http://en.wikipedia.org/wiki/Multivariate_Gaussian_distribution) with moment matrix M is proportional to the volume of the ellipsoid of concentration and is equal to . The volume is always positive.



[Average information](http://en.wikipedia.org/wiki/Average_information) may be maximized using [Gaussian adaptation](http://en.wikipedia.org/wiki/Gaussian_adaptation) - one of the [evolutionary algorithms](http://en.wikipedia.org/wiki/Evolutionary_algorithms) - keeping the [mean fitness](http://en.wikipedia.org/wiki/Mean_fitness) - i. e. the probability of becoming a parent to new individuals in the population - constant (and without the need for any knowledge about [average information](http://en.wikipedia.org/wiki/Average_information) as a criterion function). This is illustrated by the figure below, showing Gaussian adaptation climbing a mountain crest in a phenotypic landscape. The lines in the figure are part of a contour line enclosing a region of acceptability in the landscape. At the start the cluster of red points represents a very homogeneous population with small variances in the phenotypes. Evidently, even small environmental changes in the landscape, may cause the process to become extinct.



After a sufficiently large number of generations, the increase in [average information](http://en.wikipedia.org/wiki/Average_information) may result in the green cluster. Actually, the [mean fitness](http://en.wikipedia.org/wiki/Mean_fitness) is the same for both red and green cluster (about 65%). The effect of this adaptation is not very salient in a 2-dimensional case, but in a high-dimensional case, the efficiency of the search process may be increased by many orders of magnitude.

Besides, a Gaussian distribution has the highest [average information](http://en.wikipedia.org/wiki/Average_information) as compared to other distributions having the same second order moment matrix (Middleton 1960).

But it turns out that this is *not* in general a good measure of uncertainty or information. For example, the differential entropy can be negative; also it is not invariant under continuous coordinate transformations. [Jaynes](http://en.wikipedia.org/wiki/Jaynes) showed in fact in [[1]](http://bayes.wustl.edu/etj/articles/brandeis.pdf) (sect. 4b) that the expression above is not the correct limit of the expression for a finite set of probabilities.

The correct expression, appropriate for the continuous case, is the **relative entropy** of a distribution, defined as the [Kullback-Leibler divergence](http://en.wikipedia.org/wiki/Kullback-Leibler_divergence) from the distribution to a reference measure *m*(*x*),



(or sometimes the negative of this).

The relative entropy carries over directly from discrete to continuous distributions, and is invariant under coordinate reparamatrisations. For an application of relative entropy in a [quantum information theory](http://en.wikipedia.org/wiki/Quantum_statistical_mechanics) setting, see eg [[2]](http://arxiv.org/abs/math-ph/0007010/).

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=4)] Theoretical relationship

Despite all that, there is an important difference between the two quantities. The information entropy *H* can be calculated for *any* probability distribution (if the "message" is taken to be that the event *i* which had probability *pi* occurred, out of the space of the events possible). But the thermodynamic entropy *S* refers to thermodynamic probabilities *pi* specifically.

Furthermore, the thermodynamic entropy *S* is dominated by different arrangements of the system, and in particular its energy, that are possible on a molecular scale. In comparison, information entropy of any macroscopic event is so small as to be completely irrelevant.

However, a connection can be made between the two, if the probabilities in question are the thermodynamic probabilities *pi*: the (reduced) Gibbs entropy σ can then be seen as simply the amount of Shannon information needed to define the detailed microscopic state of the system, given its macroscopic description. Or, in the words of [G. N. Lewis](http://en.wikipedia.org/wiki/G._N._Lewis) writing about chemical entropy in 1930, "Gain in entropy always means loss of information, and nothing more". To be more concrete, in the discrete case using base two logarithms, the reduced Gibbs entropy is equal to the minimum number of yes/no questions that need to be answered in order to fully specify the microstate, given that we know the macrostate.

Furthermore, the prescription to find the equilibrium distributions of statistical mechanics, such as the Boltzmann distribution, by maximising the Gibbs entropy subject to appropriate constraints (the [Gibbs algorithm](http://en.wikipedia.org/wiki/Gibbs_algorithm)), can now be seen as something not unique to thermodynamics, but as a principle of general relevance in all sorts of statistical inference, if it desired to find a [maximally uninformative probability distribution](http://en.wikipedia.org/wiki/Principle_of_maximum_entropy), subject to certain constraints on the behaviour of its averages. (These perspectives are explored further in the article [*Maximum entropy thermodynamics*](http://en.wikipedia.org/wiki/Maximum_entropy_thermodynamics)).

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=5)] Information is physical

### [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=6)] Szilard's engine

A neat physical [thought-experiment](http://en.wikipedia.org/wiki/Thought-experiment) demonstrating how just the possession of information might in principle have thermodynamic consequences was established in 1929 by [Szilard](http://en.wikipedia.org/wiki/Leo_Szilard), in a refinement of the famous [Maxwell's demon](http://en.wikipedia.org/wiki/Maxwell%27s_demon) scenario.

Consider Maxwell's set-up, but with only a single gas particle in a box. If the supernatural demon knows which half of the box the particle is in (equivalent to a single bit of information), it can close a shutter between the two halves of the box, close a piston unopposed into the empty half of the box, and then extract *kBT*ln2 joules of useful work if the shutter is opened again. The particle can then be left to isothermally expand back to its original equilibrium occupied volume. In just the right circumstances therefore, the possession of a single bit of Shannon information (a single bit of [negentropy](http://en.wikipedia.org/wiki/Negentropy) in Brillouin's term) really does correspond to a reduction in physical entropy, which theoretically can indeed be parlayed into useful physical work.

### [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=7)] Landauer's principle

Main article: [Landauer's principle](http://en.wikipedia.org/wiki/Landauer%27s_principle)

In fact one can generalise: any information that has a physical representation must somehow be embedded in the statistical mechanical degrees of freedom of a physical system.

Thus, [Rolf Landauer](http://en.wikipedia.org/wiki/Rolf_Landauer) argued in 1961, if one were to imagine starting with those degrees of freedom in a thermalised state, there would be a real reduction in thermodynamic entropy if they were then re-set to a known state. This can only be achieved under information-preserving microscopically deterministic dynamics if the uncertainty is somehow dumped somewhere else — ie if the entropy of the environment (or the non information-bearing degrees of freedom) is increased by at least an equivalent amount, as required by the Second Law, by gaining an appropriate quantity of heat: specifically *kT* ln 2 of heat for every 1 bit of randomness erased.

On the other hand, Landauer argued, there is no thermodynamic objection to a logically reversible operation potentially being achieved in a physically reversible way in the system. It is only logically irreversible operations — for example, the erasing of a bit to a known state, or the merging of two computation paths — which must be accompanied by a corresponding entropy increase.

Applied to the Maxwell's demon/Szilard engine scenario, this suggests that it might be possible to "read" the state of the particle into a computing apparatus with no entropy cost; but *only* if the apparatus has already been SET into a known state, rather than being in a thermalised state of uncertainty. To SET (or RESET) the apparatus into this state will cost all the entropy that can be saved by knowing the state of Szilard's particle.

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=8)] Negentropy

Shannon entropy has been related by physicist [Léon Brillouin](http://en.wikipedia.org/wiki/L%C3%A9on_Brillouin) to a concept sometimes called [negentropy](http://en.wikipedia.org/wiki/Negentropy). In his 1962 book *Science and Information Theory*, Brillouin described the Negentropy Principle of Information or NPI, the gist of which is that acquiring information about a system’s microstates is associated with a decrease in entropy (work is needed to extract information, erasure leads to increase in thermodynamic entropy).[[1]](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#cite_note-TOA-0#cite_note-TOA-0) There is no violation of the second law of thermodynamics, according to Brillouin, since a reduction in any local system’s thermodynamic entropy results in an increase in thermodynamic entropy elsewhere. Negentropy is a controversial concept as it yields Carnot efficiency higher than one.[*[citation needed](http://en.wikipedia.org/wiki/Wikipedia:Citation_needed" \o "Wikipedia:Citation needed)*]

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=9)] Black holes

|  |  |
| --- | --- |
|  | This section requires [expansion](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit). |

[Stephen Hawking](http://en.wikipedia.org/wiki/Stephen_Hawking) often speaks of the thermodynamic entropy of [black holes](http://en.wikipedia.org/wiki/Black_hole) in terms of their information content. Do black holes destroy information? See [*Black hole thermodynamics*](http://en.wikipedia.org/wiki/Black_hole_thermodynamics) and [*Black hole information paradox*](http://en.wikipedia.org/wiki/Black_hole_information_paradox).

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=10)] Quantum theory

|  |  |
| --- | --- |
|  | This section requires [expansion](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit). |

Hirschman showed in 1957, however, that [Heisenberg's uncertainty principle](http://en.wikipedia.org/wiki/Heisenberg%27s_uncertainty_principle) can be expressed as a particular lower bound on the sum of the entropies of the *observable* probability distributions of a particle's position and momentum, when they are expressed in [Planck units](http://en.wikipedia.org/wiki/Planck_units). (One could speak of the "joint entropy" of these distributions by considering them independent, but since they are not jointly observable, they cannot be considered as a [joint distribution](http://en.wikipedia.org/wiki/Joint_distribution).)

It is well known that a Shannon based definition of information entropy leads in the classical case to the Boltzmann entropy. It is tempting to regard the Von Neumann entropy as the corresponding quantum mechanical definition. But the latter is problematic from quantum information point of view. Consequently Stotland, Pomeransky, Bachmat and Cohen have introduced a new definition of entropy that reflects the inherent uncertainty of quantum mechanical states. This definition allows to distinguish between the minimum uncertainty entropy of pure states, and the excess statistical entropy of mixtures.

* [The information entropy of quantum mechanical states](http://arxiv.org/abs/quant-ph/0401021), Europhysics Letters 67, 700 (2004).

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=11)] The fluctuation theorem

|  |  |
| --- | --- |
|  | This section requires [expansion](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit). |

The [fluctuation theorem](http://en.wikipedia.org/wiki/Fluctuation_theorem) provides a mathematical justification of the [second law of thermodynamics](http://en.wikipedia.org/wiki/Second_law_of_thermodynamics) under these principles, and precisely defines the limitations of the applicability of that law to the microscopic realm of individual particle movements.

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=12)] Topics of recent research

### [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=13)] Is information quantized?

In 1995, Dr Tim Palmer signalled two unwritten assumptions about Shannon's definition of information that may make it inapplicable as such to [quantum mechanics](http://en.wikipedia.org/wiki/Quantum_mechanics):

* The supposition that there is such a thing as an observable state (for instance the upper face of a die or a coin) *before* the observation begins
* The fact that knowing this state does not depend on the order in which observations are made (commutativity)

The article *Conceptual inadequacy of the Shannon information in quantum measurement* [[3]](http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=PLRAAN000063000002022113000001&idtype=cvips&gifs=yes), published in 2001 by [Anton Zeilinger](http://en.wikipedia.org/wiki/Anton_Zeilinger) [[4]](http://www.quantum.univie.ac.at/zeilinger/) and [Caslav Brukner](http://en.wikipedia.org/w/index.php?title=Caslav_Brukner&action=edit&redlink=1), synthesized and developed these remarks. The so-called [Zeilinger's principle](http://en.wikipedia.org/w/index.php?title=Zeilinger%27s_principle&action=edit&redlink=1) suggests that the quantization observed in QM could be bound to *information* quantization (one cannot observe less than one bit, and what is not observed is by definition "random").

But these claims remain highly controversial. For a detailed discussion of the applicability of the Shannon information in quantum mechanics and an argument that Zeilinger's principle cannot explain quantization, see Timpson [[5]](http://www.philosophy.leeds.ac.uk/Staff/CT/Index.htm) 2003 [[6]](http://arxiv.org/abs/quant-ph/0112178) and also Hall 2000 [[7]](http://arxiv.org/abs/quant-ph/0007116) and Mana 2004 [[8]](http://arxiv.org/abs/quant-ph/0302049), who shows that Brukner and Zeilinger change, in the middle of the calculation in their article, the numerical values of the probabilities needed to compute the Shannon entropy, so that the calculation makes no sense.

For a tutorial on quantum information see [[9]](http://members.aol.com/jmtsgibbs/infothry.htm).

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=14)] See also

* [Thermodynamic entropy](http://en.wikipedia.org/wiki/Entropy)
* [Information entropy](http://en.wikipedia.org/wiki/Information_entropy)
* [Thermodynamics](http://en.wikipedia.org/wiki/Thermodynamics)
* [Statistical mechanics](http://en.wikipedia.org/wiki/Statistical_mechanics)
* [Information theory](http://en.wikipedia.org/wiki/Information_theory)
* [Physical information](http://en.wikipedia.org/wiki/Physical_information)
* [Fluctuation theorem](http://en.wikipedia.org/wiki/Fluctuation_theorem)
* [Black hole entropy](http://en.wikipedia.org/wiki/Black_hole_entropy)
* [Black hole information paradox](http://en.wikipedia.org/wiki/Black_hole_information_paradox)

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=15)] External links

* [Entropy is Simple...If You Avoid the Briar Patches](http://www.entropysimple.com/content.htm). Dismissive of direct link between information-theoretic and thermodynamic entropy.

## [[edit](http://en.wikipedia.org/w/index.php?title=Entropy_in_thermodynamics_and_information_theory&action=edit&section=16)] References

1. [**^**](http://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory#cite_ref-TOA_0-0#cite_ref-TOA_0-0) [Classical Information Theory (Shannon)](http://www.toarchive.org/faqs/information/shannon.html#Entropy) – Talk Origins Archive

* C. H. Bennett, "Logical reversibility of computation," IBM Journal of Research and Development, vol. 17, no. 6, pp. 525-532, 1973.
* Leon Brillouin, *Science and Information Theory*, Mineola, N.Y.: Dover, [1956, 1962] 2004. [ISBN 0-486-43918-6](http://en.wikipedia.org/wiki/Special:BookSources/0486439186)
* Michael P. Frank, ["Physical Limits of Computing"](http://www.cise.ufl.edu/research/revcomp/physlim/plpaper.html), *Computing in Science and Engineering*, **4**(3):16-25, May/June 2002.
* Andreas Greven, Gerhard Keller, and Gerald Warnecke, editors. *Entropy*, Princeton University Press, 2003. [ISBN 0-691-11338-6](http://en.wikipedia.org/wiki/Special:BookSources/0691113386). (A highly technical collection of writings giving an overview of the concept of entropy as it appears in various disciplines.)
* I. Hirschman, *A Note on Entropy*, American Journal of Mathematics, 1957.
* R. Landauer, *Information is Physical* Proc. Workshop on Physics and Computation PhysComp'92 (IEEE Comp. Sci.Press, Los Alamitos, 1993) pp. 1-4.
* R. Landauer, [*Irreversibility and Heat Generation in the Computing Process*](http://domino.research.ibm.com/tchjr/journalindex.nsf/c469af92ea9eceac85256bd50048567c/8a9d4b4e96887b8385256bfa0067fba2?OpenDocument) IBM J. Res. Develop. Vol. 5, No. 3, 1961
* H. S. Leff and A. F. Rex, Editors, *Maxwell's Demon: Entropy, Information, Computing*, Princeton University Press, Princeton, NJ (1990). [ISBN 0-691-08727-X](http://en.wikipedia.org/wiki/Special:BookSources/069108727X)
* Middleton, D. An Introduction to Statistical Communication Theory. McGraw-Hill, 1960.
* Claude E. Shannon. [*A Mathematical Theory of Communication*](http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication). Bell System Technical Journal, July/October 1948.