1
Information

Most cognitive scientists think about the brain and behavior within an information-processing framework: Stimuli acting on sensory receptors provide information about the state of the world. The sensory receptors transduce the stimuli into neural signals, streams of action potentials (aka spikes). The spike trains transmit the information contained in the stimuli from the receptors to the brain, which processes the sensory signals in order to extract from them the information that they convey. The extracted information may be used immediately to inform ongoing behavior, or it may be kept in memory to be used in shaping behavior at some later time. Cognitive scientists seek to understand the stages of processing by which information is extracted, the representations that result, the motor planning processes through which the information enters into the direction of behavior, the memory processes that organize and preserve the information, and the retrieval processes that find the information in memory when it is needed. Cognitive neuroscientists want to understand where these different aspects of information processing occur in the brain and the neurobiological mechanisms by which they are physically implemented.

Historically, the information-processing framework in cognitive science is closely linked to the development of information technology, which is used in electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information. But what exactly is this "information" that is so central to both cognitive science and computer science? Does it have a rigorous meaning? In fact, it does. Moreover, the conceptual system that has grown up around this rigorous meaning – information theory – is central to many aspects of modern science and engineering, including some aspects of cognitive neuroscience. For example, it is central to our emerging understanding of how neural signals transmit information about the ever-changing state of the world from sensory receptors to the brain (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). For us, it is an essential foundation for our central claim, which is that the function of the neurobiological memory mechanism is to carry information forward in time in a computationally accessible form.

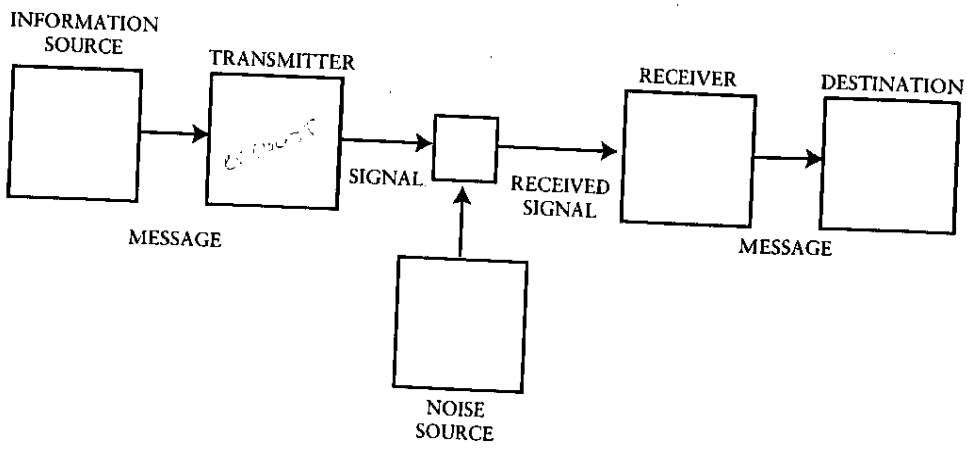

Figure 1.1 Shannon's schematization of communication (Shannon, 1948).

Shannon's Theory of Communication

The modern quantitative understanding of information rests on the work of Claude Shannon. A telecommunications engineer at Bell Laboratories, he laid the mathematical foundations of information theory in a famous paper published in 1948, at the dawn of the computer age (Shannon, 1948). Shannon's concern was understanding communication (the transmission of information), which he schematized as illustrated in Figure 1.1.

The schematic begins with an information *source*. The source might be a person who hands in a written message at a telegraph office. Or, it might be an orchestra playing a Beethoven symphony. In order for the message to be communicated to you, you must receive a *signal* that allows you to reconstitute the message. In this example, you are the *destination* of the message. Shannon's analysis ends when the destination has received the signal and reconstituted the message that was present at the source.

The *transmitter* is the system that converts the messages into transmitted signals, that is, into fluctuations of a physical quantity that travels from a source location to a receiving location and that can be detected at the receiving location. Encoding is the process by which the messages are converted into transmitted signals. The rules governing or specifying this conversion are the code. The mechanism in the transmitter that implements the conversion is the encoder.

Following Shannon, we will continue to use two illustrative examples, a telegraphic communication and a symphonic broadcast. In the telegraphic example, the source messages are written English phrases handed to the telegrapher, for example, "Arriving tomorrow, 10 am." In the symphonic example, the source messages are sound waves arriving at a microphone. Any one particular short message written in English and handed to a telegraph operator can be thought of as coming from a finite *set of possible messages*. If we stipulate a maximum length of, say, 1,000

characters, with each character being one of 45 or so different characters (26 letters, 10 digits, and punctuation marks), then there is a very large but finite number of possible messages. Moreover, only a very small fraction of these messages are intelligible English, so the size of the set of possible messages – defined as intelligible English messages of 1,000 characters or less – is further reduced. It is less clear that the sound waves generated by an orchestra playing Beethoven's Fifth can be conceived of as coming from a finite set of messages. That is why Shannon chose this as his second example. It serves to illustrate the generality of his theory.

In the telegraphy example, the telegraph system is the transmitter of the messages. The signals are the short current pulses in the telegraph wire, which travel from the sending key to the sounder at the receiving end. The encoder is the telegraph operator. The code generally used is the Morse code. This code uses pulses of two different durations to encode the characters – a *short mark* (dot), and a *long mark* (dash). It also uses four different inter-pulse intervals for separations – an intra-character gap (between the dots and dashes within characters), a short gap (between the letters), a medium gap (between words), and a long gap (between sentences).

In the orchestral example, the broadcast system transmitting radio signals from the microphone to your radio is the transmitter. The encoder is the electronic device that converts the sound waves into electromagnetic signals. The type of code is likely to be one of three different codes that have been used in the history of radio (see Figure 1.2), all of which are in current use. All of them vary a parameter of a high-frequency *sinusoidal* carrier signal. The earliest code was the AM (amplitude modulated) code. In this code, the encoder modulates the amplitude of the carrier signal so that this amplitude of the sinusoidal carrier signal varies in time in a way that closely follows the variation in time of the sound pressure at the microphone's membrane.

When the FM (frequency modulated) code is used, the encoder modulates the frequency of the carrier signal within a limited range. When the *digital* code is used, as it is in satellite radio, parameters of the carrier frequency are modulated so as to implement a binary code, a code in which there are only two characters, customarily called the '0' and the '1' character. In this system, time is divided into extremely short intervals. During any one interval, the carrier signal is either low ('0') or high ('1'). The relation between the sound wave arriving at the microphone with its associated encoding electronics and the transmitted binary signal is not easily described, because the encoding system is a sophisticated one that makes use of what we have learned about the statistics of broadcast messages to create efficient codes. The development of these codes rests on the foundations laid by Shannon.

In the history of radio broadcasting, we see an interesting evolution (Figure 1.2): We see first (historically) in Figure 1.2a a code in which there is a transparent (easily comprehended) relation between the message and the signal that transmits it (AM). The code is transparent because variation in the amplitude of the message is converted into variation in the amplitude of the carrier signal that transmits the message. This code is, however, inefficient and highly vulnerable to noise. It is low tech. In Figure 1.2b, we see a code in which the relation is somewhat less transparent, because variation in the amplitude of the message is converted into

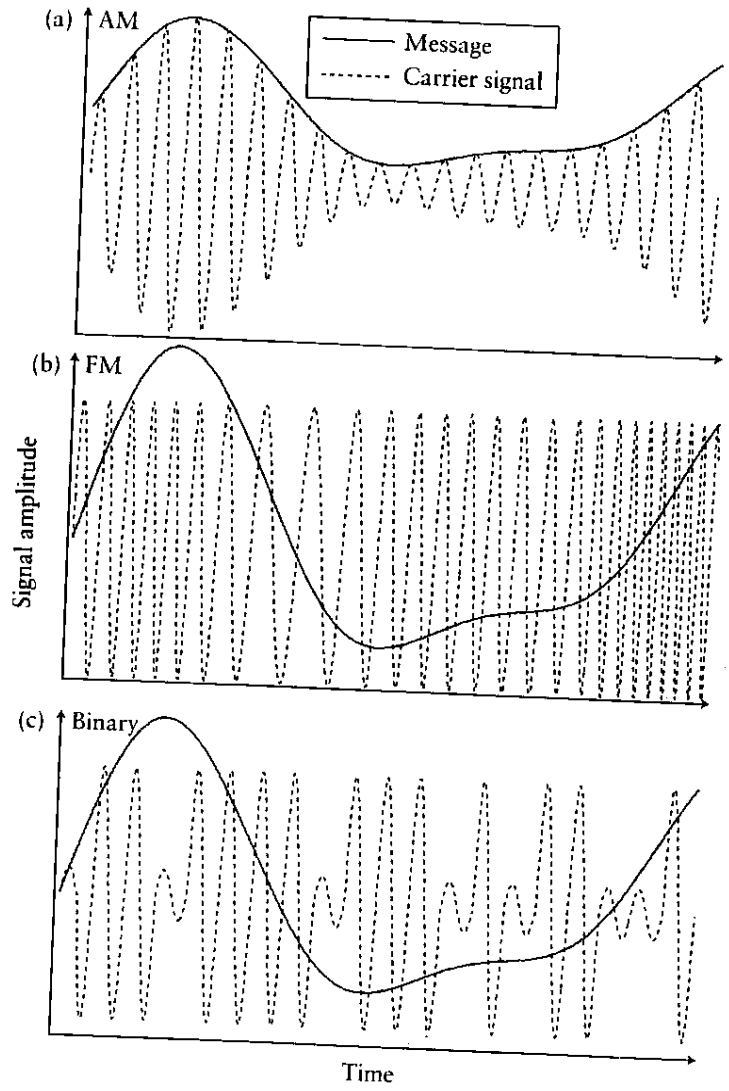

Figure 1.2 The various ways of encoding sound "messages" into broadcast radio signals. All of them use a carrier frequency and vary parameters of that carrier frequency. (a) In the AM encoding, the amplitude of the message determines the amplitude of the carrier frequency. This makes for a transparent (easily recognized) relation between the message and the signal that transmits it. (b) In the FM encoding, the amplitude of the message modulates the frequency of the carrier. This makes for a less transparent but still recognizable relation between message and signal. (c) In digital encoding, there is binary (two-values only) modulation in a parameter of the carrier signal. In this purely notional illustration, the amplitude of any given cycle has one of two values, depending on whether a high or low bit is transmitted. In this scheme, the message is converted into a sophisticated binary code prior to transmission. The relation between message and signal is opaque.

variation in the frequency of the carrier signal that transmits it (FM). This code is no more efficient than the first code, but it is less vulnerable to noise, because the effects of extraneous noise tend to fall mostly in frequency bands outside a given FM band. Finally, in Figure 1.2c we see a high-tech code in which the relation between the message and the signal that transmits it is opaque. The encoding makes extensive use of advanced statistics and mathematics. The code is, however, both efficient and remarkably invulnerable to noise. That's why satellite broadcasts sound better than FM broadcasts, which sound better than AM broadcasts. The greater efficiency of the digital code accounts for the ability of digital radio to transmit more channels within a given bandwidth.

The evolution of encoding in the history of broadcasting may contain an unpalatable lesson for those interested in understanding communication within the brain by means of the action potentials that carry information from sources to destinations within the brain. One of neurobiology's uncomfortable secrets – the sort of thing neurobiologists are not keen to talk about except among themselves – is that we do not understand the code that is being used in these communications. Most neurobiologists assume either explicitly or tacitly that it is an unsophisticated and transparent code. They assume, for example, that when the relevant variation at the source is in the amplitude or intensity of some stimulus, then the information-carrying variation in the transmitted signal is in the firing rate (the number of action potentials per unit of time), a so-called *rate code*. The transparency of rate codes augurs well for our eventually understanding the communication of information within the brain, but rate codes are grossly inefficient. With more sophisticated but less transparent codes, the same physical resources (the transmission of the same number of spikes in a given unit of time) can convey orders of magnitude more information. State-of-the-art analysis of information transmission in neural signaling in simple systems where we have reason to believe that we know both the set of message being transmitted and the amount of information available in that set (its entropy – see below) implies that the code is a sophisticated and efficient one, one that takes account of the relative frequency of different messages (source statistics), just as the code used in digital broadcasting does (Rieke et al., 1997).

A signal must travel by way of some physical medium, which Shannon refers to as the signal-carrying channel, or just channel for short. In the case of the telegraph, the signal is in the changing flow of electrons and the channel is a wire. In the case of the symphony, the signal is the variation in the parameters of a carrier signal. The channel is that carrier signal.[1] In the case of the nervous system, the axons along which nerve impulses are conducted are the channels.

In the real world, there are factors other than the message that can also produce these same fluctuations in the signal-carrying channel. Shannon called these *noise*

[1] In digital broadcasting, bit-packets from different broadcasts are intermixed and travel on a common carrier frequency. The receivers sort out which packets belong to which broadcast. They do so on the basis of identifying information in the packets. Sorting out the packets and decoding them back into waveforms requires computation. This is why computation and communication are fused at the hip in information technology. In our opinion, a similar situation obtains in the brain: Computation and communication are inseparable, because communication has been optimized in the brain.

sources. The signal that arrives at the *receiver* is thus a mixture of the fluctuations deriving from the encoding of the message and the fluctuations deriving from noise sources. The fluctuations due to noise make the receiver's job more difficult, as the received code can become corrupted. The receiver must reconstitute the message from the source, that is, change the signal back into that message, and if this signal has been altered, it may be hard to decode. In addition, the transmitter or the receiver may be faulty and introduce noise during the encoding/decoding process.

Although Shannon diagrammatically combined the sources of noise and showed one place where noise can be introduced, in actuality, noise can enter almost anywhere in the communication process. For example, in the case of telegraphy, the sending operators may not code correctly (use a wrong sequence of dots and dashes) or even more subtly, they might make silences of questionable (not clearly discernible) length. The telegraph key can also malfunction, and not always produce current when it should, possibly turning a dash into some dots. Noise can also be introduced into the signal directly – in this case possibly through interference due to other signals traveling along wires that are in close proximity to the signal-carrying wire. Additionally, the receiving operator may have a faulty sounder or may simply decode incorrectly.

Shannon was, of course, aware that the messages being transmitted often had *meanings*. Certainly this is the case for the telegraphy example. Arguably, it is the case for the orchestra example. However, one of his profound insights was that from the standpoint of the communications engineer, the meaning was irrelevant. What was essential about a message was not its meaning but rather that *it be selected from a set of possible messages*. Shannon realized that for a communication system to work efficiently – for it to transmit the maximum amount of information in the minimum amount of time – both the transmitter and the receiver had to know what the set of possible messages was and the relative likelihood of the different messages within the set of possible messages. This insight was an essential part of his formula for quantifying the information transmitted across a signal-carrying channel. We will see later (Chapter 9) that Shannon's set of possible messages can be identified with the values of an experiential variable. Different variables denote different sets of possible messages. Whenever we learn from experience the value of an empirical variable (for example, how long it takes to boil an egg, or how far it is from our home to our office), the range of a priori possible values for that variable is narrowed by our experience. The greater the range of a priori possible values for the variable (that is, the larger the set of possible messages) and the narrower the range after we have had an informative experience (that is, the more precisely we then know the value), the more informative the experience. That is the essence of Shannon's definition of information.

The thinking that led to Shannon's formula for quantifying information may be illustrated by reference to the communication situation that figures in Longfellow's poem about the midnight ride of Paul Revere. The poem describes a scene from the American revolution in which Paul Revere rode through New England, warning the rebel irregulars that the British troops were coming. The critical stanza for our purposes is the second:

He said to his friend, "If the British march
By land or sea from the town to-night,
Hang a lantern aloft in the belfry arch
Of the North Church tower as a signal light, –
One if by land, and two if by sea;
And I on the opposite shore will be,
Ready to ride and spread the alarm
Through every Middlesex village and farm,
For the country folk to be up and to arm."

The two possible messages in this communication system were "by land" and "by sea." The signal was the lantern light, which traveled from the church tower to the receiver, Paul Revere, waiting on the opposite shore. Critically, Paul knew the possible messages and he knew the code – the relation between the possible messages and the possible signals. If he had not known either one of these, the communication would not have worked. Suppose he had no idea of the possible routes by which the British might come. Then, he could not have created a set of possible messages. Suppose that, while rowing across the river, he forgot whether it was one if by land and two if by sea or two if by land and one if by sea. In either case, the possibility of communication disappears. No set of possible messages, no communication. No agreement about the code between sender and receiver, no communication.

However, it is important to remember that information is always about something and that signals can, and often do, carry information about multiple things. When we said above that no information was received, we should have been more precise. If Paul forgot the routes (possible messages) or the code, then he could receive no information about how the British might come. This is not to say that he received no information when he saw the lanterns. Upon seeing the two lanterns, he would have received information about how many lanterns were hung. In the simplest analysis, a received signal always (baring overriding noise) carries information regarding which signal was sent.

Measuring Information

Shannon was particularly concerned with *measuring* the amount of information communicated. So how much information did Paul Revere get when he saw the lanterns (for two it was)? On Shannon's analysis, that depends on his prior expectation about the relative likelihoods of the British coming by land versus their coming by sea. In other words, it depends on how uncertain he was about which route they would take. Suppose he thought it was a toss-up – equally likely either way. According to Shannon's formula, he then received one bit[2] (the basic unit) of information when he saw the signal. Suppose that he thought it less likely that they

[2] Shannon was the first to use the word *bit* in print, however he credits John Tukey who used the word as a shorthand for "binary digit."

would come by land – that there was only one chance in ten. By Shannon's formula, he then received somewhat less than half a bit of information from the lantern signal.

Shannon's analysis says that the (average!) amount of information communicated is the (average) amount of uncertainty that the receiver had before the communication minus the amount of uncertainty that the receiver has after the communication. This implies that information itself is the reduction of uncertainty in the receiver. A reduction in uncertainty is, of course, an increase in certainty, but what is measured is the uncertainty.

The discrete case

So how did Shannon measure uncertainty? He suggested that we consider the *prior probability* of each message. The smaller the prior probability of a message, the greater its information content but the less often it contributes that content, because the lower its probability, the lower its relative frequency. The contribution of any one possible message to the average uncertainty regarding messages in the set of possible messages is the information content of that message times its relative frequency. Its information content is the log of the reciprocal of its probability $\left(\log_2 \frac{1}{p_i}\right)$. Its relative frequency is p_i itself. Summing over all the possible messages gives Shannon's famous formula:

$$H = \sum_{i=1}^{i=n} p_i \log_2 \frac{1}{p_i}$$

where H is the amount of uncertainty about the possible messages (usually called the *entropy*), n is the number of possible messages, and p_i is the probability of the i^{th} message.[3] As the probability of a message in the set becomes very small (as it approaches 0), its contribution to the amount of uncertainty also becomes very small, because a probability goes to 0 faster than the log of its reciprocal goes to infinity. In other words, the fall off in the relative frequency of a message (the decrease in p_i) outstrips the increase in its information content $\left(\text{the increase in } \log_2 \frac{1}{p_i}\right)$.

In the present, simplest possible case, there are two possible messages. If we take their prior probabilities to be 0.5 and 0.5 (50–50, equally likely), then following Shannon's formula, Paul's uncertainty before he saw the signal was:

$$p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} = 0.5 \log_2 \frac{1}{0.5} + 0.5 \log_2 \frac{1}{0.5} \qquad (1)$$

[3] The logarithm is to base 2 in order to make the units of information bits, that is, to choose a base for the logarithm is to choose the size of the units in which information is measured.

Now, 1/0.5 = 2, and the log to the base 2 of 2 is 1. Thus, equation (1) equals:

$(0.5)(1) + (0.5)(1) = 1$ bit.

Consider now the case where $p_1 = 0.1$ (Paul's prior probability on their coming by land) and $p_2 = 0.9$ (Paul's prior probability on their coming by sea). The $\log_2 (1/0.1)$ is 3.32 and the $\log_2 (1/0.9)$ is 0.15, so we have $(0.1)(3.32) + (0.9)(0.15) = 0.47$. If Paul was pretty sure they were coming by sea, then he had less uncertainty than if he thought it was a toss-up. That's intuitive. Finding a principled formula that specifies exactly how much less uncertainty he had is another matter. Shannon's formula was highly principled. In fact, he proved that his formula was the only formula that satisfied a number of conditions that we would want a measure of uncertainty to have.

One of those conditions is the following: Suppose we have H_1 amount of uncertainty about the outcome of the roll of one die and H_2 amount of uncertainty about the outcome of the roll of a second die. We want the amount of uncertainty we have about the combined outcomes to be simply $H_1 + H_2$, that is, we want the amounts of uncertainties about independent sets of possibilities to be additive. Shannon's formula satisfies this condition. That's why it uses logarithms of the probabilities. Independent probabilities combine multiplicatively. Taking logarithms converts multiplicative combination to additive combination.

Assuming Paul trusted his friend completely and assuming that there was no possibility of his mistaking one light for two (assuming in other words, no transmission noise), then when he saw the two lights, he had no more uncertainty about which way the British were coming: p_1, the probability of their coming by land, was 0 and p_2, the probability of their coming by sea, was 1. Another condition on a formula for measuring uncertainty is that the measure should be zero when there is no uncertainty. For Paul, after he had seen the lights, we have: $0 \log_2 (1/0) + 1 \log_2 (1/1) = 0$ (because the $\lim_{p \to 0} p \log (1/p) = 0$, which makes the first term in the sum 0, and the log of 1 to any base is 0, which makes the second term 0). So Shannon's formula satisfies that condition.

Shannon defined the amount of information *communicated* to be the difference between the receiver's uncertainty before the communication and the receiver's uncertainty after it. Thus, the amount of information that Paul got when he saw the lights depends not only on his knowing beforehand the two possibilities (knowing the set of possible messages) but also on his prior assessment of the probability of each possibility. This is an absolutely critical point about communicated information – and the subjectivity that it implies is deeply unsettling. By subjectivity, we mean that the information communicated by a signal depends on the receiver's (the subject's) prior knowledge of the possibilities and their probabilities. Thus, the amount of information actually communicated is not an objective property of the signal from which the subject obtained it!

Unsettling as the subjectivity inherent in Shannon's definition of communicated information is, it nonetheless accords with our intuitive understanding of communication. When someone says something that is painfully obvious to everyone, it is not uncommon for teenagers to reply with a mocking, "Duh." Implicit in this

mockery is that we talk in order to communicate and to communicate you have to change the hearer's representation of the world. If your signal leaves your listeners with the same representation they had before they got it, then your talk is empty blather. It communicates no information.

Shannon called his measure of uncertainty entropy because his formula is the same as the formula that Boltzmann developed when he laid the foundations for statistical mechanics in the nineteenth century. Boltzmann's definition of entropy relied on statistical considerations concerning the degree of uncertainty that the observer has about the state of a physical system. Making the observer's uncertainty a fundamental aspect of the physical analysis has become a foundational principle in quantum physics, but it was extremely controversial at the time (1877). The widespread rejection of his work is said to have driven Boltzmann to suicide. However, his faith in the value of what he had done was such that he had his entropy-defining equation written on his tombstone.

In summary, like most basic quantities in the physical sciences, information is a mathematical abstraction. It is a statistical concept, intimately related to concepts at the foundation of statistical mechanics. The information available from a source is the amount of uncertainty about what that source may reveal, what message it may have for us. The amount of uncertainty at the source is called the source entropy. The signal is a propagating physical fluctuation that carries the information from the source to the receiver.

The information *transmitted* to the receiver by the signal is the *mutual information* between the signal actually received and the source. This is an objective property of the source and signal; we do not need to know anything about the receiver (the subject) in order to specify it, and it sets an upper limit on the information that a receiver could in principle get from a signal. We will explain how to quantify it shortly. However, the information that is *communicated* to a receiver by a signal is the receiver's uncertainty about the state of the world before the signal was received (the receiver's prior entropy) minus the receiver's uncertainty after receiving the signal (the posterior entropy). Thus, its quantification depends on the changes that the signal effects in the receiver's representation of the world. The information communicated from a source to a receiver by a signal is an inherently subjective concept; to measure it we must know the receiver's representation of the source probabilities. That, of course, implies that the receiver has a representation of the source probabilities, which is itself a controversial assumption in behavioral neuroscience and cognitive psychology. One school of thought denies that the brain has representations of any kind, let alone representations of source possibilities and their probabilities. If that is so, then it is impossible to communicate information to the brain in Shannon's sense of the term, which is the only scientifically rigorous sense. In that case, an information-processing approach to the analysis of brain function is inappropriate.

The continuous case

So far, we have only considered the measurement of information in the discrete case (and a maximally simple one). That is to say that each message Paul could

receive was distinct, and it should not have been possible to receive a message "in between" the messages he received. In addition, the number of messages Paul could receive was finite – in this case only two. The British could have come by land or by sea – not both, not by air, etc. It may seem puzzling how Shannon's analysis can be applied to the continuous case, like the orchestra broadcast. On first consideration, the amount of prior uncertainty that a receiver could have about an orchestral broadcast is infinite, because there are infinitely many different sound-wave patterns. Any false note hit by any player at any time, every cough, and so on, alters the wave pattern arriving at the microphone. This seems to imply that the amount of prior uncertainty that a receiver could have about an orchestral broadcast is infinite. Hearing the broadcast reduces the receiver's uncertainty from infinite to none, so an infinite amount of information has been communicated. Something must be wrong here.

To see what is wrong, we again take a very simple case. Instead of an orchestra as our source, consider a container of liquid whose temperature is measured by an analog (continuous) thermometer that converts the temperature into a current flow. Information is transmitted about the temperature to a receiver in a code that theoretically contains an infinite number of possibilities (because for any two temperatures, no matter how close together they are, there are an infinite number of temperatures between them). This is an analog source (the variation in temperature) and an analog signal (the variation in current flow). Analog sources and signals have the theoretical property just described, infinite divisibility. There is no limit to how finely you can carve them up. Therefore, no matter how thin the slice you start with you can always slice them into arbitrarily many even thinner slices. Compare this to the telegraphy example. Here, the source was discrete and so was the signal. The source was a text written in an alphabetic script with a finite number of different characters (letters, numbers, and various punctuation marks). These characters were encoded by Morse's code into a signal that used six primitive symbols. Such a signal is called a digital signal.

In the temperature case, there would appear to be an infinite number of temperatures that the liquid could have, any temperature from 0–∞° Kelvin. Further thought tells us, however, that while this may be true in principle (it's not clear that even in principle temperatures can be infinite), it is not true in practice. Above a certain temperature, both the container and the thermometer would vaporize. In fact, in any actual situation, the range of possible temperatures will be narrow. Moreover, we will have taken into account that range when we set up the system for measuring and communicating the liquid's temperature. That is, the structure of the measuring system will reflect the characteristics of the messages to be transmitted. This is the sense in which the system will know the set of possible messages; the knowledge will be implicit in its structure.

However, even within an arbitrarily narrow range of temperatures, there are arbitrarily many different temperatures. That is what it means to say that temperature is a continuous variable. This is true, but the multiple and inescapable sources of noise in the system limit the attainable degree of certainty about what the temperature is. There is source noise – tiny fluctuations from moment to moment and place to place within the liquid. There is measurement noise; the fluctuations in the

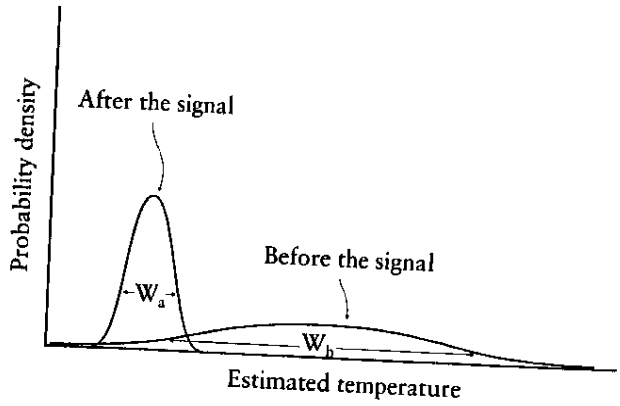

Figure 1.3 In analog communication, the receipt of a signal alters the receiver's probability density distribution, the distribution that specifies the receiver's knowledge of the source value. Generally (though not obligatorily), it narrows the distribution, that is, $\sigma_a < \sigma_b$, and it shifts the mean and mode (most probable value).

electrical current from the thermometer will never exactly mimic the fluctuations in the temperature at the point being measured. And there is transmission noise; the fluctuations in the current at the receiver will never be exactly the same as the fluctuations in the current at the transmitter. There are limits to how small each of these sources of noise can be made. They limit the accuracy with which the temperature of a liquid can in principle be known. Thus, where we went wrong in considering the applicability of Shannon's analysis to the continuous case was in assuming that an analog signal from an analog source could give a receiver information with certainty; it cannot. The accuracy of analog signaling is always noise limited, and it must be so for deep physical reasons. Therefore, the receiver of an analog signal always has a residual uncertainty about the true value of the source variable. This a priori limit on the accuracy with which values within a given range may be known limits the number of values that may be distinguished one from another within a finite range. That is, it limits resolution. The limit on the number of distinguishable values together with the limits on the range of possible values makes the source entropy finite and the post-communication entropy of the receiver non-zero.

Figure 1.3 shows how Shannon's analysis applies to the simplest continuous case. Before the receiver gets an analog signal, it has a continuous (rather than discrete) representation of the possible values of some variable (e.g., temperature). In the figure, this prior (before-the-signal) distribution is assumed to be a normal (aka Gaussian) distribution, because it is rather generally the case that we construct a measurement system so that the values in the middle of the range of possible (i.e., measured) values are the most likely values. Shannon derived the entropy for a normal distribution, showing that it was proportional to the log of the standard deviation, σ, which is the measure of the width of a distribution. Again, this is intuitive: the broader the distribution is, the more uncertainty there is. After receiving the signal, the receiver has less uncertainty about the true value of the temperature. In Shannon's analysis, this means that the posterior (after-the-signal)

distribution is narrower and higher. The information conveyed by the signal is proportional to the difference in the two entropies: $k(\log \sigma_b - \log \sigma_a)$.

How does the simple case generalize to a complex case like the orchestral broadcast? Here, Shannon made use of the Fourier theorem, which tells us how to represent a continuous variation like the variation in sound pressure produced by an orchestra with a set of sine waves. The Fourier theorem asserts that the whole broadcast can be uniquely represented as the sum of a set of sinusoidal oscillations. If we know this set – the so-called Fourier decompositions of the sound – we can get back the sound by simply adding all the sinusoids point by point. (See Gallistel, 1980, for elementary explanation and illustration of how this works; also King & Gallistel, 1996.) In principle, this representation of the sound requires infinitely many different sinusoids; but in practice, there are limits on both the sensible range of sinusoidal frequencies and the frequency resolution within that range. For example, there is no point in representing the frequencies above 20 kHz, because humans cannot hear them. In principle, the number of possible amplitudes for a sinusoid is infinite, but there are limits on the amplitudes that broadcast sounds actually do have; and within that attainable range, there are limits on the resolution with which sound amplitude may be ascertained. The same is true for phase, the third and final parameter that defines a sinusoid and distinguishes it from other sinusoids. Thus, the space of possible broadcasts is the space defined by the range of hearable frequencies and attainable amplitudes and phases. Because there are inescapable limits to the accuracy with which each of these three space-defining parameters may be ascertained, there is necessarily some residual uncertainty about any broadcast (some limit on the fidelity of the transmission). Hence, odd as it seems, there is a finite amount of prior uncertainty about possible broadcasts and a residual amount of uncertainty after any transmitted broadcast. This makes the amount of information communicated in a broadcast finite and, more importantly, actually measurable. Indeed, communications engineers, following the guidelines laid down by Shannon, routinely measure it. That's how they determine the number of songs your portable music player can hold.

Mutual information

The mutual information between an information-conveying signal and its source is the entropy of the source plus the entropy of the signal minus the entropy of their joint distribution. Recall that entropy is a property of a probability (relative frequency) distribution over some set of possibilities. The source entropy is a quantity derived from the distribution of probability over the possible messages (the relative frequencies of the different possible messages). The signal entropy is a quantity derived from the distribution of probability over the possible signals (the relative frequencies of the different possible signals). A distribution is the set of *all* the probabilities (or relative frequencies), one probability for each possibility. Thus, the sum over these probabilities is always 1, because one or the other possibility must obtain in every case and the set contains all the possible cases (all the possible messages or all the possible signals). In computing the entropy of a distribution, we take each probability in turn, multiply the logarithm of its reciprocal by the probability itself,

and sum across all the products. Returning to the Paul Revere example, if the probability, p_L, of their coming by land is 0.1 and the probability, p_S of their coming by sea is 0.9, then the source entropy (the basic uncertainty inherent in the situation) is:

$$p_L \log_2 \frac{1}{p_L} + p_S \log_2 \frac{1}{p_S} = (0.1)(3.32) + (0.9)(0.15) = 0.47.$$

If the two signals, one light and two lights, have the same probability distribution, then the signal entropy is the same as the source entropy.

The joint distribution of the messages and the signals is the probabilities of all possible co-occurrences between messages and signals. In the Paul Revere example, four different co-occurrences are possible: (1) the British are coming by land and there is one signal light; (2) the British are coming by land and there are two signal lights; (3) the British are coming by sea and there is one signal light; (4) the British are coming by sea and there are two signal lights. The joint distribution is these four probabilities. The entropy of the joint distribution is obtained by the computation we already described: multiply the logarithm of the reciprocal of each probability by the probability itself and sum the four products.

The entropy of this joint distribution depends on how reliably Paul's confederate carries out the assigned task. Suppose that he carries it out flawlessly: every time they come by land, he hangs one lantern; every time they come by sea, he hangs two. Then the four probabilities are $p_{L\&1} = 0.1$, $p_{L\&2} = 0$, $p_{S\&1} = 0$, $p_{S\&2} = 0.9$ and the entropy of this joint distribution is the same as the entropy of the source distribution and the entropy of the signal distribution; all three entropies are 0.47. The sum of the source and signal entropies (the first two entropies) minus the third (the entropy of the joint distribution) is 0.47, so the mutual information between source and signal is 0.47, which is to say that all the information available at the source is transmitted by the signal.

Suppose instead that Paul's confederate is terrified of the British and would not think of spying on their movements. Therefore, he has no idea which way they are coming, but, because he does not want Paul to know of his cowardice, he hangs lanterns anyway. He knows that the British are much more likely to go by sea than by land, so each night he consults a random number table. He hangs one lantern if the first digit he puts his finger on is a 1 and two lanterns otherwise. Now, there is no relation between which way the British are coming and the signal Paul sees. Now the four probabilities corresponding to the four possible conjunctions of British movements and the coward's signals are: $p_{L\&1} = 0.01$, $p_{L\&2} = 0.09$, $p_{S\&1} = 0.09$, $p_{S\&2} = 0.81$ and the entropy of this joint distribution is:

$$(0.01) \log_2 (\tfrac{1}{0.01}) + (0.09) \log_2 (\tfrac{1}{0.09}) + (0.09) \log_2 (\tfrac{1}{0.09}) + (0.81) \log_2 (\tfrac{1}{0.81})$$
$$= (0.01)(6.64) + (0.09)(3.47) + (0.09)(3.47) + (0.81)(0.30) = 0.94.$$

The entropy of the joint distribution is equal to the sum of the two other entropies (more technically, the entropy of the joint distribution is the sum of the entropies of the marginal distributions). When it is subtracted from that sum, the difference

is 0. There is no mutual information between the signal and the source. Whether Paul knows it or not, he can learn nothing about what the British are doing from monitoring his confederate's signal. Notice that there is no subjectivity in the computation of the mutual information between source and signal. That is why we can measure the amount of information transmitted without regard to the receiver's representation of the source and the source probabilities.

Finally, consider the case where Paul's confederate is not a complete coward. On half the nights, he gathers up his courage and spies on the British movements. On those nights, he unfailingly signals correctly what he observes. On the other half of the nights, he resorts to the random number table. Now, the probabilities in the joint distribution are: $p_{L\&1} = 0.055$, $p_{L\&2} = 0.045$, $p_{S\&1} = 0.045$, $p_{S\&2} = 0.855$ and the entropy of this joint distribution is:

$$(0.055)\log_2\left(\tfrac{1}{0.055}\right) + (0.045)\log_2\left(\tfrac{1}{0.045}\right) + (0.045)\log_2\left(\tfrac{1}{0.045}\right) + (0.855)\log_2\left(\tfrac{1}{0.855}\right)$$
$$= (0.055)(4.18) + (0.045)(4.47) + (0.045)(4.47) + (0.855)(0.23) = 0.83.$$

When this entropy is subtracted from 0.94, the sum of the entropies of the source and signal distributions, we get 0.11 for the mutual information between source and signal. The signal does convey some of the available information, but by no means all of it. The joint distribution and the two marginal distributions are shown in Table 1.1. Notice that the probabilities in the marginal distributions are the sums of the probabilities down the rows or across the columns of the joint distribution.

The mutual information between source and signal sets the upper limit on the information that may be communicated to the receiver by that signal. There is no way that the receiver can extract more information about the source from the signal received than is contained in that signal. The information about the source contained in the signal is an objective property of the statistical relation between the source and the signal, namely, their joint distribution, the relative frequencies with which all possible combinations of source message and received signal occur. The information communicated to the receiver, by contrast, depends on the receiver's ability to extract the information made available in the signals it receives (for example, the receiver's knowledge of the code, which may be imperfect) and on the receiver's representation of the possibilities and their probabilities.

Table 1.1 Joint and marginal distributions in the case where lantern signal conveys some information about British route

British route/Lantern signal	*One lantern*	*Two lanterns*	*Marginal (route)*
By land	0.055	0.045	0.1
By sea	0.045	0.855	0.9
Marginal (Signal)	0.1	0.9	

Efficient Coding

As illustrated in Figure 1.2c, in a digital broadcast, the sound wave is transmitted digitally. Typically, it is transmitted as a sequence of bits ('0' or '1') that are themselves segregated into sequences of eight bits – called a byte. This means that each byte can carry a total of 256 or 2^8 possible messages (each added bit doubles the information capacity). The coding scheme, the method for translating the sound into bytes, is complex, which is why a digital encoder requires sophisticated computational hardware. The scheme incorporates knowledge of the statistics of the sound waves that are actually produced during human broadcasts into the creation of an efficient code. Shannon (1948) showed that an efficient communication code could only be constructed if one knew the statistics of the source, the relative likelihoods of different messages.

An elementary example of this is that in constructing his code, Morse made a single dot the symbol for the letter 'E,' because he knew that this was the most common letter in English text. Its frequency of use is hundreds of times higher than the frequency of use of the letter 'Z' (whose code is dash, dash, dot, dot). Shannon (1948) showed how to measure the efficiency of a communication code, thereby transforming Morse's intuition into quantitative science.

The routine use of digital transmission (and recordings with digital symbols) of broadcasts is another example that the space of discernibly different broadcasts ultimately contains a finite and routinely measured amount of uncertainty (entropy). To a first approximation, the prior uncertainty (the entropy) regarding the soundform of a broadcast of a specified length is measured by the capacity (often expressed in megabytes, that is, a million bytes) of the CD required to record it. The number of possible broadcasts of that length is the number of different patterns that could be written into that amount of CD space. If all of those patterns were equally likely to occur, then that number of megabytes would be the prior entropy for broadcasts of that length. In fact, however, some of those patterns are vastly more likely than others, because of the harmonic structure of music and the statistical structure of the human voice and instruments, among other things. To the extent that the sound-encoding scheme built into a recorder fails to take account of these statistics, the actual entropy is less than the entropy implied by the amount of disk space required.

It is, however, often possible to specify at least approximately the amount of information that a given signal could be carrying to a receiver. This is a critical point because efficient codes often do not reflect at all the intrinsic properties of what it is they encode. We then say that the code is indirect. An appreciation of this last point is of some importance in grasping the magnitude of the challenge that neuroscientists may face in understanding how the brain works, so we give an illustrative example of the construction of increasingly efficient codes for sending English words.

One way to encode English words into binary strings is to start with the encoding that we already have by virtue of the English alphabet, which encodes words as strings of characters. We then can use a code such as ASCII (American Standard

Code for Information Interchange), which specifies a byte for each letter, that is a string of eight '0's or '1's – A = 01000001, B = 01000010, and so on. The average English word is roughly 6 characters long and we have to transmit 8 bits for each character, so our code would require an average of about 48 bits each time we transmitted a word. Can we do better than that? We will assume about 500,000 words in English and $2^{19} = 524{,}288$. Thus, we could assign a unique 19-bit pattern to each English word. With that code, we need send only 19 bits per word, better by a factor of 2.5. A code that allows for fewer bits to be transferred is said to be compact or compressed and the encoding process contains a compression scheme. The more successfully we compress, the closer we get to transmitting on average the number of bits specified by the source entropy. Can we make an even better compression scheme? This last code assumes in effect that English words are equally likely, which they emphatically are not. You hear or read 'the' hundreds of times every day, whereas you may go a lifetime without hearing or reading 'eleemosynary' (trust us, it's an English word, a rare but kindly one).

Suppose we arrange English words in a table according to their frequency of use (Table 1.2 shows the first 64 most common words). Then we divide the table in half, so that the words that account for 50% of all usage are in the upper half and the remaining words in the lower half. It turns out that there are only about 180 words in the top half! Now, we divide each of these halves in half, to form usage quartiles. In the top quartile, there are only about 15 words! They account for 25% of all usage. In the second quartile, accounting for the next 25% of all usage, are about 165 words; and in the third quartile, about 2,500 words. The remaining 500,000 or so words account for only 25% of all usage.

We can exploit these extreme differences in probability of occurrence to make a more highly compressed and efficient binary code for transmitting English words. It is called a Shannon-Fano code after Shannon, who first placed it in print in his 1948 paper, and Fano, who originated the idea and popularized it in a later publication. We just keep dividing the words in half according to their frequency of usage. At each division, if a word ends up in the top half, we add a 0 to the string of bits that code for it. Thus, the 180 words that fall in the top half of the first division, all have 0 as their first digit, whereas the remaining 500,000 odd words all have 1. The 15 words in the first quartile (those that ended up in the top half of the first two divisions), also have 0 as their second digit. The 165 or so words in the second quartile all have 1 as their second digit. We keep subdividing the words in this way until every word has been assigned a unique string of '0's and '1's. Table 1.2 shows the Shannon-Fano codes for the first 64 most commonly used English words, as found in one source (The Natural Language Technology Group, University of Brighton) on the Internet.

As may be seen in Table 1.2, this scheme insures that the more frequent a word is, the fewer bits we use to transmit it. Using the Shannon-Fano code, we only need to transmit *at most* 19 bits for any one word – and that only very infrequently. For 40% of all the words we transmit, we use 9 bits or fewer. For 25%, we use only 5 or 6 bits. With this code, we can get the average number of bits per word transmitted down to about 11, which is almost five times more efficient than the code we first contemplated. This shows the power of using a code that takes account

Table 1.2 Constructing a Shannon-Fano code for English words. Shannon-Fano codes for the first 64 most common words in the English language.* Also shown is the cumulative percent of usage. These 64 words account for roughly 40% of all usage in English text. Note that some words are repeated as they are considered separate usage.

Rank	*Word*	*%*	*cum %*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	the	6.25%	6.25%	0	0	0	0	0				
2	of	2.97%	9.23%	0	0	0	0	1				
3	and	2.71%	11.94%	0	0	0	1	0				
4	a	2.15%	14.09%	0	0	0	1	1				
5	in	1.83%	15.92%	0	0	1	0	0	0			
6	to	1.64%	17.56%	0	0	1	0	1	1			
7	it	1.10%	18.66%	0	0	1	1	0	0			
8	is	1.01%	19.67%	0	0	1	1	1	0			
9	was	0.93%	20.60%	0	0	1	1	1	1			
10	to	0.93%	21.53%	0	0	1	0	0	0			
11	I	0.89%	22.43%	0	0	1	0	1	0			
12	for	0.84%	23.27%	0	0	1	0	1	1			
13	you	0.70%	23.97%	0	0	1	1	0	0			
14	he	0.69%	24.66%	0	0	1	1	1	0			
15	be	0.67%	25.33%	0	0	1	1	1	1			
16	with	0.66%	25.99%	0	1	0	0	0	0	0		
17	on	0.65%	26.64%	0	1	0	0	0	0	1	0	
18	that	0.64%	27.28%	0	1	0	0	0	0	1	1	
19	by	0.51%	27.79%	0	1	0	0	0	1	0	0	
20	at	0.48%	28.28%	0	1	0	0	0	1	1	0	
21	are	0.48%	28.75%	0	1	0	0	0	1	1	1	
22	not	0.47%	29.22%	0	1	0	0	1	0	0	0	
23	this	0.47%	29.69%	0	1	0	0	1	0	1	0	
24	but	0.46%	30.15%	0	1	0	0	1	0	1	1	
25	's	0.45%	30.59%	0	1	0	0	1	1	0	0	
26	they	0.44%	31.03%	0	1	0	0	1	1	0	1	
27	his	0.43%	31.46%	0	1	0	0	1	1	1	0	
28	from	0.42%	31.88%	0	1	0	0	1	1	1	1	
29	had	0.41%	32.29%	0	1	0	1	0	0	0	0	
30	she	0.38%	32.68%	0	1	0	1	0	0	0	1	
31	which	0.38%	33.05%	0	1	0	1	0	0	1	0	
32	or	0.37%	33.43%	0	1	0	1	0	0	1	1	
33	we	0.36%	33.79%	0	1	0	1	0	1	0	0	
34	an	0.35%	34.14%	0	1	0	1	0	1	0	1	
35	n't	0.34%	34.47%	0	1	0	1	0	1	1	0	
36	's	0.33%	34.80%	0	1	0	1	0	1	1	1	
37	were	0.33%	35.13%	0	1	0	1	1	0	0	0	
38	that	0.29%	35.42%	0	1	0	1	1	0	0	1	0
39	been	0.27%	35.69%	0	1	0	1	1	0	0	1	1
40	have	0.27%	35.96%	0	1	0	1	1	0	1	0	0
41	their	0.26%	36.23%	0	1	0	1	1	0	1	0	1
42	has	0.26%	36.49%	0	1	0	1	1	0	1	1	0
43	would	0.26%	36.75%	0	1	0	1	1	0	1	1	1
44	what	0.25%	37.00%	0	1	0	1	1	1	0	0	0
45	will	0.25%	37.25%	0	1	0	1	1	1	0	1	0

Table 1.2 *(cont'd)*

Rank	*Word*	*%*	*cum %*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
46	there	0.24%	37.49%	0	1	0	1	1	1	0	1	1
47	if	0.24%	37.73%	0	1	0	1	1	1	1	0	0
48	can	0.24%	37.96%	0	1	0	1	1	1	1	0	1
49	all	0.23%	38.20%	0	1	0	1	1	1	1	1	0
50	her	0.22%	38.42%	0	1	0	1	1	1	1	1	1
51	as	0.21%	38.63%	0	1	1	0	0	0	0	0	0
52	who	0.21%	38.83%	0	1	1	0	0	0	0	1	0
53	have	0.21%	39.04%	0	1	1	0	0	0	0	1	1
54	do	0.20%	39.24%	0	1	1	0	0	0	1	0	0
55	that	0.20%	39.44%	0	1	1	0	0	0	1	0	1
56	one	0.19%	39.63%	0	1	1	0	0	0	1	1	0
57	said	0.19%	39.82%	0	1	1	0	0	0	1	1	1
58	them	0.18%	39.99%	0	1	1	0	0	1	0	0	0
59	some	0.17%	40.17%	0	1	1	0	0	1	0	0	1
60	could	0.17%	40.34%	0	1	1	0	0	1	0	1	0
61	him	0.17%	40.50%	0	1	1	0	0	1	0	1	1
62	into	0.17%	40.67%	0	1	1	0	0	1	1	0	0
63	its	0.16%	40.83%	0	1	1	0	0	1	1	0	1
64	then	0.16%	41.00%	0	1	1	0	0	1	1	1	1

* This list is not definitive and is meant only for illustrative purposes.

of the source statistics. Another important property of a Shannon-Fano code is that it is what is called a *prefix code*. This means that no word is coded by a bit pattern that is the prefix for any other word's code. This makes the code self-delimiting so that when one receives multiple words as a string of bits, there is no need for any form of punctuation to separate the words, and there is no ambiguity. Notice that this leads to a clarification of the efficiency of the ASCII encoding. The ASCII encoding of English text is not a prefix code. For example, if one received the text "andatareallastask," there would be no way to know with certainty if the intended words were "and at are all as task," or "an data real last ask." Because of this, the ASCII encoding scheme would actually require each word to end with a space character (another code of 8 bits), and the total expected bits per word increases to 7 bytes or 56 bits per word.[4]

Compact codes are not necessarily a win-win situation. One problem with compact codes is that they are much more susceptible to corruption by noise than non-compact codes. We can see this intuitively by comparing the ASCII encoding scheme to the each-word-gets-a-number scheme. Let's say we are trying to transmit one

[4] The Shannon-Fano prefix code, while efficient, is suboptimal and can result in less than perfect compression. The Huffman (1952) encoding scheme uses a tree-like structure formed from the bottom up based on the probabilities themselves, not just the rankings. It produces a prefix code that can be shown to be optimal with respect to a frequency distribution that is used irrespective of the text sent, that is, it does not take advantage of the statistics of the particular message being sent.

English word. In the ASCII scheme, roughly 48 bits encode each word. This is a total number of 2^{48} possible patterns – a number in excess of 36 quadrillion – 36,000,000,000,000,000. With our each-word-gets-a-number scheme, we send 19 bits per word, resulting in 2^{19} possible patterns or 524,288. If, for argument's sake, we assume that our lexicon contains 524,288 possible words, then if one bit is changed (from a '0' to a '1' or from a '1' to a '0') because of noise on the signal channel, then the word decoded will with certainty be another word from the lexicon (one of possibly 19 words), with no chance of knowing (without contextual clues) that the error occurred. On the other hand, with the ASCII scheme, regardless of the noise, we will have less than a 1 in 50 billion chance of hitting another word in our lexicon. Since this "word" will almost certainly not be found in the lexicon, it will be known that an error has occurred and the communication system can request that the word be re-sent or likely even correct the error itself. Clearly in a communication system with very noisy channels, using the ASCII scheme would be more costly in terms of bits, but more likely to get the right message across.

We can help this problem, however, by adding redundancy into our schemes. For example, with the each-word-gets-a-number scheme, we could send 3 bits for each 1 bit we sent before, each 3 bits simply being copies of the same bit. So instead of transmitting the 19 bits, 1001010001100110011, we would transmit 57 bits:

111000000111000111000000000111111000000111111000000111111

In this case, we have decreased the efficiency back to the ASCII scheme, however, the redundancy has resulted in certain advantages. If any one bit is flipped due to noise, not only can we detect the error with certainty, we can also correct it with certainty. If two bits are flipped, then with certainty we can detect it. We would also have a 55/56 chance of correcting it.

Information and the Brain

Clearly, the tradeoffs between efficiency, accuracy, error detection, and error correction can lead to tremendous complexities when designing efficient codes in a world with noise. These issues are made even more complex when one takes into account the relative frequencies of the messages, as is done with the Shannon-Fano coding scheme. Computer scientists must routinely deal with these issues in designing real-world communication schemes. It is almost certainly the case that that the brain deals with the same issues. Therefore, an understanding of these issues is crucial to understanding the constraints that govern the effective transmission of information by means of nerve impulses within the brain.

As noted in connection with Figure 1.2, insofar as the considerations of efficiency and noise-imperviousness have shaped the system of information transmission within the brain, the brain's signaling code may be indirect. That is, the signals may not reflect intrinsic properties of the things (source messages) that they encode for. For example, first consider an ASCII encoding of a word, such as 'dog.' Note that we are talking about the word 'dog', not the animal. The word is first

encoded into letters, that is "dog." This code reflects inherent properties of the word 'dog', as the letters (to some degree) reflect phonemes in the spoken word ("d" reflects the 'd' sound). If we encode each letter by an ASCII symbol, we retain this coding property, as each character has a one-to-one mapping to an ASCII symbol. This coding scheme is quite convenient as it also has some direct relationships with many other features of words such as their frequency of usage (smaller words tend to be more common), their part of speech, their country of origin, and even their meaning. As we saw, however, this direct encoding comes at a price – the code is not compact and is not ideal for transmission efficiency.

On the other hand, consider the Shannon-Fano encoding scheme applied to words. Here, the letters are irrelevant to the coding process. Instead, the code generates the signals based on the words' rank order in a usage table, not from anything related to its sound or meaning (although there are strong and interesting correlations between meaning and relative frequency – something that code breakers can use to their advantage). Most efficient (compact) codes make use of such relative frequencies and are therefore similarly indirect.

In addition, in modern signal transmission, it is often the case that encoded into the signals are elements of redundancy that aid with the problem of noise. One common technique is to include what are called *checksum* signals to the encoding signal. The checksum refers not to what the symbol encodes for, but instead, the symbol itself. This allows the communication system to detect if a message was corrupted by noise. It is called a checksum, as it typically treats the data as packets of numbers, and then adds these numbers up. For example, let's take the ASCII encoding scheme. The word 'dog' (lower case) would be encoded as 01100100, 01101111, 01100111. Now, we can treat these bytes as binary numbers, giving us the sequence (in decimal), 100, 111, 103. If we sum these numbers, we get 314. Because this is a bigger number than can be encoded by one byte (8 bits), we take the remainder when divided by 255, which is 59. In binary, that is 00111011. If we prepend this byte to the original sequence, we can (with over 99% certainty), determine if the signal was corrupted. Such schemes involve computations at both the source and destination, and they can make the code harder to break.

If coding schemes in the nervous system are similarly indirect, then the neuroscientist's job is hard. We have no assurance that they are not. At present, with a few small and recent exceptions (Rieke et al., 1997), neurophysiologists are in the position of spies trying to figure out how a very complex multinational corporation functions by listening to phone conversations conducted in a communication code they do not understand. That is because, generally speaking, neuroscientists do not know what it is about trains of action potentials that carries the information, nor exactly what information is being communicated. We've been listening to these signals for a century, but we have only translated minute parts of what we have overheard.

This brings us to a brief consideration of how Shannon's analysis applies to the brain (Figure 1.4). The essential point is that the brain is a receiver of signals that, under the proper conditions, convey to it information about the state of the world. The signals the brain receives are trains of action potentials propagating down sensory axons. Neurophysiologists call these action potentials spikes, because they

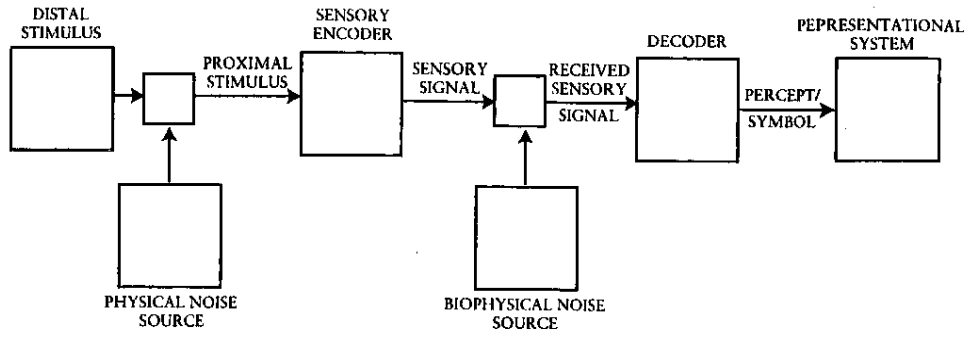

Figure 1.4 World-to-brain communication. The states of some delimited aspect of the world correspond to Shannon's messages. Perceptual psychologists call these states distal stimuli. Stimulus energy is either reflected off or emitted by the source. This energy together with contaminating energy from other sources (noise) impinges on sensory receptors in sensory organs (sensory encoders). Perceptual psychologists call the stimulus that actually impinges on the receptors the proximal stimulus. The encoders translate the proximal stimulus into sensory signals, streams of spikes in the sensory axons leading from sensory organs to the brain. Biophysical noise contaminates this neural signal, with the result that variations in the spike train are not due entirely to variations in the proximal stimulus. The sensory-processing parts of the brain are the decoder. Successive stages of sensory decoding translate incoming sensory signals into, first, a representation of aspects of the proximal stimulus, and then into a set of symbols that constitute what psychologists call a percept. This set of symbols represents the distal stimulus in the brain's subsequent information processing. The appropriate processing of these symbols, together with the communication chain that confers reference on them, makes the brain a representational system.

look like spikes when viewed on an oscilloscope at relatively low temporal resolution. Spikes are analogous to electrical pulses that carry information within electronic systems. Sensory organs (eyes, ears, noses, tongues, and so on) and the sensory receptors embedded in them convert information-rich stimulus energy to spike trains. The stimuli that act directly on sensory receptors are called proximal stimuli. Examples are the photons absorbed by the rods and cones in the retina, the traveling waves in the basilar membrane of the cochlea, which bend the underlying hair cells, the molecules absorbed by the nasal mucosa, and so on. Proximal stimuli carry information about distal stimuli, sources out there in the world. The brain extracts this information from spike trains by processing them. This is to say that much of the signal contains data from which useful information must be determined.

The problem that the brain must solve is that the information it needs about the distal stimulus in order to act appropriately in the world – the source information – is not reflected in any simple way in the proximal stimulus that produces the spike train. Even simple properties of the proximal stimulus itself (how, for example, the pattern of light is moving across the retina) are not reflected in a straightforward way in the spike trains in the optic nerve, the bundle of sensory axons that carries information from the retina to the first way-stations in the brain. The

physical processes in the world that convert source information (for example, the reflectance of a surface) to proximal stimuli (the amount of light from that surface impinging on the retina) encode the source information in very complex ways. Many different, quite unrelated aspects of the world – for example, the reflectance of the surface and the intensity of its illumination – combine to determine proximal stimuli. To extract from the spike train useful facts about a specific source (for example, what the reflectance of a particular surface actually is), the brain must invert this complex encoding and separate the messages that are conflated in the signals it receives. This inversion and message separation is effected by a sequence of computational operations, very few of which are currently understood.

The modern approach to a neurobiological understanding of sensory transduction and the streams of impulses thereby generated relies heavily on Shannon's insights and their mathematical elaboration (Rieke et al., 1997). In a few cases, it has been possible to get evidence regarding the code used by sensory neurons to transmit information to the brains of flies and frogs. The use of methods developed from Shannon's foundations has made it possible to estimate how many bits are conveyed per spike and how many bits are conveyed by a single axon in one second. The answers have been truly revolutionary. A single spike can convey as much as 7 bits of information and 300 bits per second can be transmitted on a single axon (Rieke, Bodnar, & Bialek, 1995).

Given our estimates above of how many bits on average are needed to convey English words when an efficient code is used (about 10 per word), a single axon could transmit 30 words per second to, for example, a speech center.[5] It could do so, of course, only if the usage-frequency table necessary to decode the Shannon-Fano code were stored in the speech center, as well as in the source center. Remember that both Paul's confederate (the encoder) and Paul (the decoder) had to know the lantern code for their system to work. These encoding tables constitute knowledge of the statistical structure of English speech. Central to Shannon's analysis of communication is the realization that the structure of the encoding and decoding mechanisms must reflect the statistical structure of the source. To make a system with which the world can communicate efficiently, you must build into it implicit information about the statistical structure of that world. Fortunately, we know that English speakers do know the usage frequency of English words (even though they don't know they know it). The effects of word frequency in many tasks are among the more ubiquitous and robust effects in cognitive psychology (Hasher & Zacks, 1984; Hulme et al., 1997; Jescheniak & Levelt, 1994). The information-theoretic analysis provides an unusual explanation of why they ought to know these relative frequencies.[6]

Until the advent of these information-theoretic analyses, few neuroscientists had any notion of how to go about estimating how many axons it might in principle take to relay words to a speech center at natural speaking rates (2–8 words/second).

[5] Whether transmission rates of 300 bits per second are realistic for axons within the brain (as opposed to sensory axons) is controversial (Latham & Nirenberg, 2005).

[6] This knowledge is, of course, not built in; it is constructed in the course of learning the language.

No one would have guessed that it could be done with room to spare by a single axon. Understanding how the brain works requires an understanding of the rudiments of information theory, because what the brain deals with is information.

Digital and Analog Signals

Early communication and recording technology was often analog. Analog sources (for example, sources putting out variations in sound pressure) were encoded into analog signals (continuously fluctuating currents) and processed by analog receivers. For decades, neuroscientists have debated the question whether neural communication is analog or digital or both, and whether it matters. As most technophiles know, the modern trend in information technology is very strongly in the digital direction; state-of-the-art transmitters encode analog signals into digital signals prior to transmission, and state-of-the art receivers decode those digital signals. The major reason for this is that the effects of extraneous noise on digital communication and recording are much more easily controlled and minimized. A second and related reason is that modern communication and recording involves computation at both the transmitting (encoding) and receiving (decoding) stages. Much of this computation derives from Shannon's insights about what it takes to make a code efficient and noise resistant. Modern information-processing hardware is entirely digital – unlike the first computers, which used analog components. To use that hardware to do the encoding and decoding requires recoding analog signals into digital form. One of the reasons that computers have gone digital is for the same reason that modern information transmission has – noise control and control over the precision with which quantities are represented.

Our hunch is that information transmission and processing in the brain is likewise ultimately digital. A guiding conviction of ours – by no means generally shared in the neuroscience community – is that brains do close to the best possible job with the problems they routinely solve, given the physical constraints on their operation. Doing the best possible job suggests doing it digitally, because that is the best solution to the ubiquitous problems of noise, efficiency of transmission, and precision control.

We make this digression here because the modern theory of computation, which we will be explaining, is cast entirely in digital terms. It assumes that information is carried by a set of discrete symbols. This theory has been extensively developed, and it plays a critical role in computer science and engineering. Among other things, this theory defines what it means to say that something is computable. It also establishes limits on what is computable. There is no comparable theory for analog computation (and no such theory seems forthcoming). The theory we will be explaining is currently the only game in town. That does not, of course, mean that it will not some day be supplanted by a better game, a better theory of computation. We think it is fair to say, however, that few believe that analog computation will ultimately prove superior. There is little reason to think that there are things that can only be computed by an analog computer. On the contrary, the general, if largely unspoken, assumption is that digital computation can accomplish anything that analog

computation can, while the converse may not be the case. As a practical matter, it can usually accomplish it better. That is why there is no technological push to create better analog computers.

Appendix: The Information Content of Rare Versus Common Events and Signals

Above, we have tacitly assumed that the British move out night after night and Paul's confederate spies on them (or fails to do so) and hangs lanterns (transmits a signal) every night. In doing so, we have rectified an implicit fault in the Paul Revere example that we have used to explicate Shannon's definition of information. The fault is that it was a one-time event. As such, Shannon's analysis would not apply. Shannon information is a property of probability (that is, relative frequency) *distribution*, not of single (unique) events or single (unique) signals. With a unique event, there is only one event in the set of messages. Thus, there is no distribution. Hence, there is no entropy (or, if you like, the entropy is 0, because the relative frequency of that event is 1, and the log of 1 is 0). The consequences of the uniqueness were most likely to have surfaced when he or she came to the case in which there was said to be a 0.1 "probability" of their coming by land and a 0.9 "probability" of their coming by sea. If by probability we understand relative frequency, then these are not intelligible numbers, because with a unique event, there is no relative frequency; it either happens or it doesn't.[7] If we ignore this, then we confront the following paradox: the information communicated by the lantern signal is the same whether Paul sees the low probability signal or the high probability signal, because the prior probability distribution is the same in both cases, hence the pre-signal entropies are the same, and the post-signal entropies are both 0. If, however, the event belongs to a set of events (a set of messages) with empirically specifiable relative frequencies, then when we compute the entropy per event or per signal we find that, for rare events, the entropy per event is higher than for common events, in accord with our intuitions. We get this result because the entropy is defined over the full set of events, that is, the entropy is a property of the relative frequency *distribution* (and only of that distribution, not of its constituents, nor of their individual relative frequencies). The source entropy in the case of the British movements (assuming they recur night after night) is a single fixed quantity, regardless of whether we consider the rare occurrences (coming by land) or the common ones (coming by sea). However, the common occurrences are nine times

[7] This is among the reasons why radical Bayesians reject the interpretation of probabilities as relative frequencies. For a radical Bayesian, a probability is a strength of belief. Although we are sympathetic to this position, considering how information theory would look under this construal of probability would take us into deeper philosophical waters than we care to swim in here. As a practical matter, it is only applied in situations where relative frequencies are in fact defined. Note that whether or not an event has a relative frequency depends on the set of messages to which it belongs and that in turn depends on how we choose to describe it. Any event has a relative frequency under at least some description. This issue of descriptions relates to the question of "aboutness," which we take up in a later chapter.

more frequent than the rare ones. Therefore, the amount of entropy per common event is nine times less than the amount of entropy per rare event, because the amount of entropy per type of event times the relative frequency of that type of event has to equal the total entropy of the distribution. As the rare events get rarer and rarer, the total entropy gets smaller and smaller, but the entropy per rare event gets larger and larger. This is true whether we are considering source entropy or signal entropy. The entropy per event, which is sometimes called the information content of an event, is $\log(1/p)$, which goes to infinity (albeit slowly) as p goes to 0. Thus, the entropy of a distribution is the average information content of the events (messages) over which the distribution is defined.