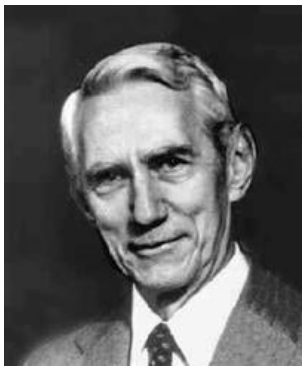

# Entropy

Our goal in this handout is to describe how it is possible to measure information, just as volume, mass and temperature are measurable. At first this may sound like an unrealistic goal! since our measure of information cannot capture the more elusive qualities of information such as how interesting or how important a body of information is. Yet this measure does capture an essential feature of information relevant to communication theory, namely how compressible the information is. Claude Shannon's notion of measuring information was quite revolutionary when he published it in 1948, but today his approach has become the standard. Shannon showed how such a measurement was possible using entropy, the same quantity used to measure the degree of randomness of a physical system. We will later compare physical entropy with information, and explain that the two are related in more than a superficial way: from a certain viewpoint, physical entropy and information are interchangeable.

There are many ways in which the disciplines of mathematics, physics and engineering overlap, as is clear from even the most cursory examination of our modern conveniences and appliances. In most of these cases, mathematical and physical theory have been applied to engineering. The late Richard Feynman, Nobel laureate and highly celebrated expositor of physics, has said that there are two very notable exceptions to this pattern, where engineers, through their investigation of practical physical phenomena, have formulated theories of great mathematical and physical interest. One of these is Claude Shannon's foundation of Information Theory in the 1940's and 1950's, and the other is Sadi Carnot's formulation of Thermodynamics in the early 19th century. Curiously, these two theories are intimately related, as we proceed to describe.



Claude Shannon
1916-2001

## Thermodynamics

The entropy of a physical system is a measure of its disorder. The concept was first conceived by the brilliant French engineer Carnot. This theory explained the capabilities and limitations of mechanical engines for transforming heat energy into useful work. His work led to the formulation of four laws of thermodynamics, which we partially describe as follows.

Sadi Carnot
1796-1832

The *Zeroth Law of Thermodynamics* says that if body A is in thermal equilibrium with body B, and body B is in thermal equilibrium with body C, then body A is in thermal equilibrium with body C. The temperature of an object or physical system, in other words, has meaning independent of the material composition of the object or system. This means that it is possible to measure the temperature of any object using a scale which does not depend on the composition of the object whose temperature is being measured.

The *First Law of Thermodynamics* states that we may convert energy from one form to another, but in any closed system, there can be no net gain in energy. Heat is a form of energy, which can often be used as an energy source from which other energy (especially mechanical and electrical) energy may be derived.

The *Second Law of Thermodynamics* states that the entropy of a closed system tends to remain the same, or to increase. It is this law upon which we will focus, since it has the greatest relevance to information theory.

There is also a *Third Law of Thermodynamics* which we will not highlight it here.

We stress the immense importance of the Second Law of Thermodynamics. One very recent panel of experts, representing undergraduate instructors of all disciplines, has called this one of the hundred most important facts every college student should know! We will try to illustrate the meaning of this law through a series of examples.

## Example 1: A Box of Quarters

Consider a box containing a dozen quarters, all neatly stacked in one pile, as in Figure A. If this box is shaken for a few seconds and then set to rest on a surface, it is reasonable to expect that the quarters will be disturbed from their orderly arrangement and come to rest in a less organized arrangement as shown in Figure B.

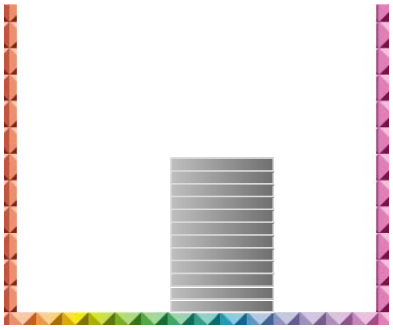


**Figure A:**
Box containing an orderly stack of quarters

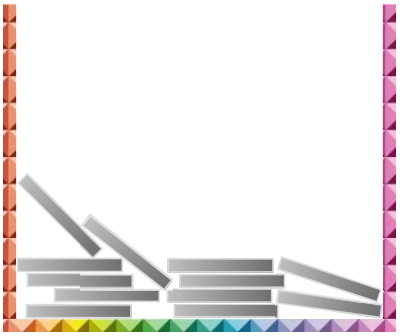


**Figure B:**
Box containing a disorderly arrangement of quarters

The transition from Figure A to Figure B represents the typical trend in physical systems: from more order to less order, in accordance with the Second Law of Thermodynamics. We would not expect a disorderly arrangement of quarters, as in Figure B, to arrange itself into a neat stack as in Figure A, merely by jiggling the box, even if enough energy were supplied in the jiggling process to account for the gravitational potential energy found in the stack of quarters.

Now imagine I reach into a box containing the disorderly arrangement of Figure B, and stack the quarters as in Figure A. In this case the system has apparently lost entropy. You might try to resolve this difficulty by saying that I am outside the system of the box. But what if we consider the larger system consisting of the room and all its contents, including myself as well as the box of quarters? Is the Second Law of Thermodynamics violated through a loss of entropy? Not really. In this scenario, the orderliness of the stacked quarters was achieved only through the expenditure of energy—mechanical energy from my body, generated through the combustion of food as fuel. Every such combustion of fuel generates entropy, although this may not be obvious to the casual observer, since it takes two separate sources of molecules—one source of food molecules including carbon, and a separate source of oxygen molecules from the air—and combines them into one supply of carbon dioxide molecules.

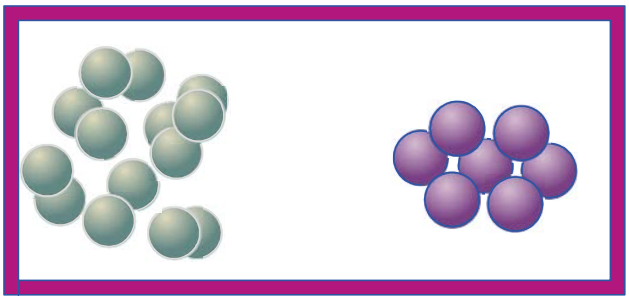


**Figure C:** Oxygen and carbon atoms prior to combustion (lower entropy)

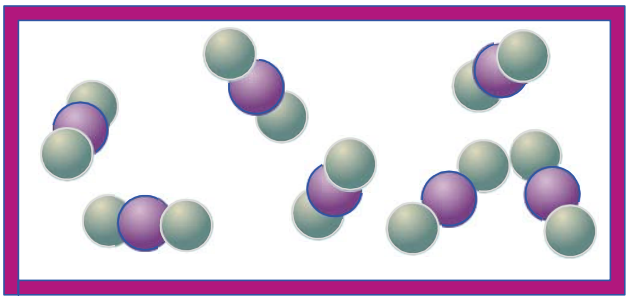


**Figure D:** Carbon dioxide molecules after combustion (higher entropy)

The output 'spent fuel' has more randomness than the original separate sources of carbon and oxygen, just as a shuffled deck of cards has more randomness than a deck having all 26 black cards on the bottom and all 26 red cards on top. We see by careful consideration of my body and the surrounding air, as well as the box of quarters, that the *total* entropy of the entire system has actually *increased*.

## Example 2: Gas Diffusion

Now consider a box divided in the middle by a wall. Suppose this wall has a gap covered by a door which may be opened or closed as we wish. Initially (see Figure E) the box contains gas molecules in the chamber on the left side of the wall, and no molecules (i.e. a vacuum) in the right-hand chamber. The gas molecules on the left will bounce around due to their thermal energy, but nothing interesting happens until we open the door. After the door is opened, gas molecules will move through the opening due to their natural motion. Once the pressure in the right-hand chamber is roughly equal to that on the left, an equilibrium is reached (see Figure F) in which molecules pass through the opening from left to right, *and* from right to left, in roughly equal numbers, preserving the balance of molecules on both sides.

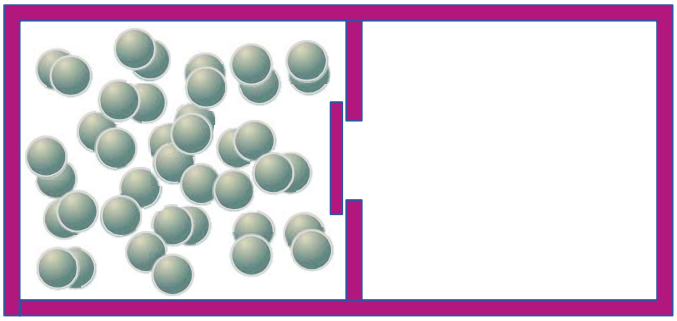


**Figure E:** Gas molecules occupy one chamber only (lower entropy)

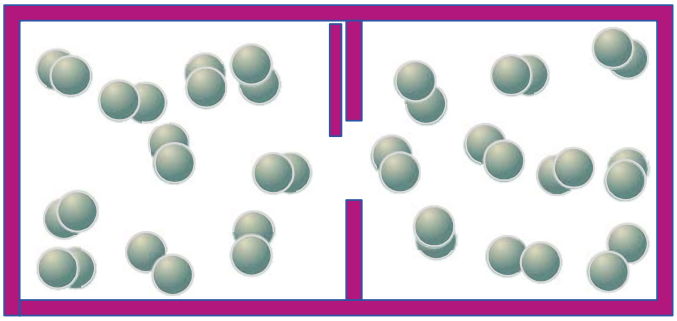
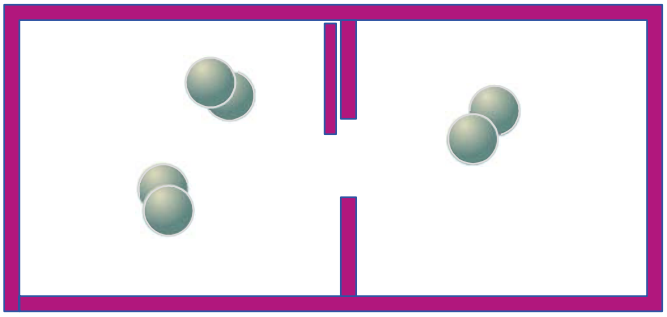


**Figure F:** Gas molecules distributed throughout both chambers (higher entropy)

Is it possible for all the molecules, as a result of their natural motion, to move to the left chamber, leaving a vacuum in the right chamber? Technically, yes, but the probability is so remotely small that we discount this possibility. That is, unless the number of gas molecules is very small, as in Figure G. Here we consider just three gas molecules, instead of the typically large number of molecules we expect to find in a large container. In this case the three molecules move about essentially independently of each other, and each molecule spends half its time in the left chamber, and half its time in the right chamber.

**Figure G:** Three gas molecules distributed throughout both chambers (entropy does not apply)



In this case, at any particular moment in time we will find

- all three gas molecules on the left side with probability 1/8;
- all three gas molecules on the right side with probability 1/8;
- two gas molecules on the left and one on the right, with probability 3/8; and
- two gas molecules on the right and one on the left, with probability 3/8.

Here the probability of all three gas molecules being on the left side is 12.5%, which is hardly negligible. The laws of thermodynamics do not apply here since the number of molecules is so few. The laws of thermodynamics assure us that certain behavior is observable for systems consisting of a *large* number of similar molecules.

## Example 3: Maxwell's 'Demon'

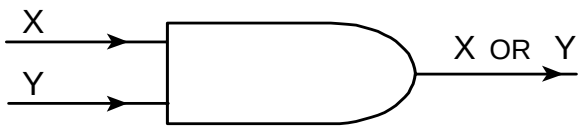
Now consider a closed system with two chambers separated by a wall, as in Figure F. As before, gas molecules are found in the chambers, and an opening in the wall is fitted with a door which may be closed or opened at will (as by sliding back and forth). This time, however, we imagine a demon perched on the wall, who is able to observe the movements of all the gas molecules near the door. Every time a molecule from the right-hand chamber approaches the door he opens it, allowing it to pass through into the left-hand chamber. Then he quickly closes the door again before any molecules from the left-hand chamber have a chance to pass through into the right-hand chamber. (This demon evidently has very fast reflexes!) Over a period of time we observe the number of gas molecules on the left increase, and the number on the right decrease, until eventually we find all the gas molecules in the left chamber, leaving a vacuum in the right chamber, just as in Figure E. Does this violate the Second Law of Thermodynamics?

This problem was first posed by the great Scottish physicist James Clerk Maxwell in 1867, and for a century thereafter, the scientific community debated what conclusion to draw from this puzzle. Some felt that the problem was in the measurement of the molecules—that somehow this measurement required an expenditure of energy, thereby requiring that additional entropy be generated (just as in Example 1, where fuel was burned to generate the necessary energy). However, Charles Bennett (the same Bennett as the pioneer of quantum information and computing) showed that this was not the case, but that rather the necessary measurements themselves require only negligible energy.

The energy required to open and close the door can also be made negligible in principle. It is possible to make the door exceedingly light and to reduce the effects of friction to near zero (or even zero). None of these can account for the missing entropy.

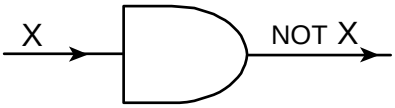
Rather, the solution requires that we understand the demon as an information processing unit—possibly mechanical, or electrical, or biological. As the demon's processing unit (be it brain, computer or other mechanism) detects the motion of nearby molecules and responds by activating the door, it must switch (more or less randomly) between two states (asking the door to open, and asking it to close). The storage unit (electrical, mechanical or biochemical) for this bit of information acts as a memory register which must be frequently erased to make room for the next incoming bit. It is the erasing of information which requires energy, and therefore entropy is generated somewhere else (again as in Example 1).

Surprisingly, any time information is erased (or equivalently, an irreducible computation is performed), energy is required. This fact means that current computing technology (based on irreversible computation) faces significant challenges from energy requirements, and from having to safely dissipate the resulting heat without damaging computing hardware. A primitive example of an irreversible computation is the 'OR' gate of classical Boolean logic:




| X | Y | X OR Y |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |

Traditionally, computing the value of X OR Y is irreversible since it loses the values of X and Y: if X OR Y is 1, this does not tell us which of X or Y (or both) is 1. That is, we cannot reverse the computation of X OR Y to recover X and Y. An example of a Boolean logical gate which is reversible is the 'NOT' gate:




| X | NOT X |
|---|---|
| 0 | 1 |
| 1 | 0 |

where it is an easy matter to recover the value of the bit X from the value of NOT X. Any irreversible process (such as a computational step) increases the entropy of a system; any reversible process must preserve the entropy of a system.

Conventional silicon chip technology implements many such irreversible logic gates, all of which require energy for erasing information during computation. We now know that such computations are possible in principle using negligible energy, but alternative technology will be required. One of the strengths of quantum computation is the fact that it is reversible. One viewpoint is that this is probably a requirement of any technology for efficiently performing many of the currently more difficult computational tasks, such as factorizing large integers.

It is unreasonable to suppose our demon has a huge supply of memory bits available, hoping that no erasing of information is required during the entire experiment. The number of memory bits required for this would be astronomical, and it would not be fair to leave the demon's memory without erasing it in preparation for the next computational task. (That would be like borrowing my friend's car and driving from Laramie to Denver and back without stopping for gas, then not refilling the gas tank, and telling my

friend I didn't require any gas!) Before the demon's memory is erased, the missing entropy is accounted for in the form of the information stored in the demon's memory registers. After the demon's memory is erased, the entropy will then be manifested elsewhere as a result of the energy consumed in order to perform this erasure.

It is hoped that in the context of this example, you can see the interchangeability of physical entropy and information.

## Definition of Entropy

We have seen the definition of entropy for a stream of independent bits. More generally, we define the entropy of an information source as

$$H(\text{info source}) = -\sum p \log_2(p)$$

where the '$\sum$' (Greek 'Sigma') symbol means that we must add together a number of terms, one term for each possible message arising from the information source; $p$ is the frequency that each possible message arises; and $\log_2(p)$ is its logarithm to base 2. If your calculator computes natural logarithms (ln) and base 10 logarithms (log = $\log_{10}$) but not arbitrary base logarithms, you can nevertheless compute $\log_2(p)$ using the formula

$$\log_2(p) = \ln(p)/\ln(2) = \log(p)/\log(2).$$

Note that $\log_2(p) \le 0$ since $p \le 1$. This is why the formula for entropy starts with a '–' sign: since the negative of a negative is positive, we will have $H(\text{info source}) \ge 0$.

Here the term 'information source' is intentionally vague; it may refer to a binary electronic file, or a sequence of symbols C, G, A, T in the DNA sequence of a biological organism, or the sequence of magnetic polarizations (denoted perhaps as simply as NNSNSSSNSNSSNSSNNN…) of iron atoms in a piece of steel.

As an example of how entropy is computed, imagine selecting random individuals from the student population of a college, and recording for each individual simply the gender. The appearance of the resulting sequence of M's and F's will depend on the proportion of the student body of each gender.

**Case 1 (Equal Balance of Gender):** If the college has equally many male and female students, this will generate a random sequence such as

FMMFFFMFMMFFFMFMMMMFMFFFMFFFF…

The entropy of each letter in this sequence is

$$H(\text{gender}) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1.$$

Such a sequence of letters is incompressible, or highly random.

**Case 2 (All Male):** If all students were male, then the gender sequence would appear as

MMMMMMMMMMMMMMMMMMMM…

In this case a large file of such data would be highly compressible due to the very evident pattern. This is expected since, referring to the graph of $H(p)$ on the previous handout, we see that $H(0) = 1$.

**Case 3 (Males Predominate):** Let us assume, rather, that this college has a student body in which 90% of students are male and 10% are female. The resulting 'gender sequence' will look something like

MMMMMMFMMMMMMMMMMMMMFMFMMMM…

The entropy of each letter in this sequence is

$$H(\text{gender}) = -0.9 \log_2(0.9) - 0.1 \log_2(0.1) = 0.4690.$$

This value of the entropy tells us that this gender sequence is less random than the sequence considered in Case 1, yet more random than the sequence in Case 2. A large file of such binary data is compressible to about 0.4690 (roughly half) of its original size.

Why do we speak of the entropy of a single letter from this information source? Let's consider several (say, three) student gender records instead of one. Assuming students are selected independently, we obtain the following table of frequencies for each possible triple of gender data:

| Gender Triple | Frequency |
|---|---|
| MMM | 0.729 |
| MMF | 0.081 |
| MFM | 0.081 |
| FMM | 0.081 |
| MFF | 0.009 |
| FFM | 0.009 |
| FMF | 0.009 |
| FFF | 0.001 |

It is the assumption that individual students are selected independently, that allows us to compute the probability of an outcome by simply multiplying individual probabilities; for instance the gender record MFM occurs with probability $0.9 \times 0.1 \times 0.9 = 0.081$. The entropy of the triple of bits is therefore

$$\begin{aligned} H(\text{gender triple}) = & -0.729 \log_2(0.729) - 0.081 \log_2(0.081) - 0.081 \log_2(0.081) - 0.081 \log_2(0.081) \\ & - 0.009 \log_2(0.729) - 0.009 \log_2(0.009) - 0.009 \log_2(0.009) - 0.001 \log_2(0.001) \\ = & \ 1.4070. \end{aligned}$$

We note that

$$\begin{aligned} H(\text{gender triple}) &= 3 \times H(\text{gender}) \\ 1.4070 &= 3 \times 0.4690 \end{aligned}$$

This is what we expect: the amount of information in three student gender records, is three times the amount of information contained in a single student gender record. This supports our main point that information may be quantitatively measured, just like volume or mass.

If the selection of individual students were not independent (for example we allowed the first student to suggest the names of two of his friends as the second and third students selected) then the entropy of a triple of gender records would be less than 1.4070. This is because the information content of such a gender triple would be lower, there being a tendency for the first student to have chosen her or his friends based on their gender. In Case 6 we will see how a drop in entropy can occur when the assumption of independence fails.

**Case 4 (Hair Color):** Now imagine that as each student is randomly selected, we record the hair color. Suppose 70% of students are blonde ('B') and 30% are non-blonde ('N'), and that each student's hair color is reported simply as 'B' or 'N'. The sequence of symbols for hair color appears something like

BBNBNBBBNBNBBBBNBBBBBBNNBBBB…

The entropy of this sequence is

$$H(\text{hair color}) = -0.7 \log_2(0.7) - 0.3 \log_2(0.3) = 0.8813.$$

Evidently the information on hair color is more random (less compressible) than the information on gender in Case 3.

**Case 5 (Gender and Hair Color are Independent):** Now suppose males predominate (as in Case 3) and blondes predominate slightly (as in Case 5), and we record both gender and hair color. The resulting information source consists of a sequence of records something like

MB, MN, MB, MB, MB, FB, MN, MB, MN, MB, MB, MB, MN, MN, MB, MB, FN, MN, …

What is the frequency of each of the four possible records MB, MN, FB, FN? It is impossible to determine this from the information given, unless we assume that hair color is independent of gender. This would mean that not only are 70% of all students blonde, but also 70% of all male students are blonde, and 70% of all female students are blonde. In this case we tabulate the frequencies for each of the four possible gender-hair color combinations as

| | | Hair Color | | |
|---|---|---|---|---|
| | | **B** | **N** | Total |
| **Gender** | **M** | 0.63 | 0.27 | 0.90 |
| | **F** | 0.07 | 0.03 | 0.10 |
| | Total | 0.70 | 0.30 | 1.00 |

For example, the frequency of the record MB is $0.9 \times 0.7 = 0.63$, so 63% of all students are male blonde. The independence of gender and hair color means that we simply multiply the probabilities of M and of B to obtain the probability of MB occurring in each record. According to our definition of entropy, each student record (represented as a two-character symbol MB, MN, FB or FN, each occurring with probabilities given by the table) has entropy

$$H(\text{student record}) = -0.63 \log_2(0.63) - 0.27 \log_2(0.27) - 0.07 \log_2(0.07) - 0.03 \log_2(0.03)$$
$$= 1.3503.$$

This means that a large file of such student records could be compressed to about 1.3503 bits per student. Notice that

$$H(\text{student record}) = H(\text{gender}) + H(\text{hair color})$$
$$1.3503 \quad = \quad 0.4690 \quad + \quad 0.8813$$

This is exactly what we expect: since gender and hair color are independent, the amount of information in reporting both gender and hair color, should equal the sum (the amount of information in the gender record, plus the amount of information in the hair color record). It is as though we added 0.4690 liters of water to 0.8813 liters of water, to get a total of 1.3503 liters of water. The fact that entropy adds in this way, supports the notion that information is measurable in such a quantitative fashion as volume.

**Case 6 (Gender and Hair Color are Dependent):** Now suppose gender and hair color frequencies occur as before, but that gender and hair color are *dependent*. Specifically, suppose that the four possible records MB, MN, FB and FN occur with frequencies given by the table

| | | Hair Color | | |
|---|---|---|---|---|
| | | B | N | Total |
| Gender | M | 0.68 | 0.22 | 0.90 |
| | F | 0.02 | 0.08 | 0.10 |
| | Total | 0.70 | 0.30 | 1.00 |

Note that as before, 90% of students are male and 70% are blonde, but this time only 20% of female students are non-blonde, whereas 75.6% of male students are blonde. In this case a student record has entropy

$$H(\text{student record}) = -0.68 \log_2(0.68) - 0.22 \log_2(0.22) - 0.02 \log_2(0.02) - 0.08 \log_2(0.08)$$
$$= 1.2633.$$

So a large file of such student records is compressible to about 1.2633 bits per student, a smaller file than that in Case 5. This is because the current data is less random, or more predictable. The information contained in the gender 'bit' (M or F) renders the information contained in the hair color bit ('B' or 'N') somewhat redundant, since males are more likely to be blonde, and females are more likely to be non-blonde. It is as though we mixed 0.4690 liters of water and 0.8813 liters of ethanol, to get 1.2633 liters of liquid, due to the two original sources of fluid being miscible.

## Shannon's Theorem

The most spectacular illustration of how useful the concept of entropy is as a measure of information content, is the main theorem of Shannon's landmark 1948 papers. For convenience we assume that information is to be encoded as binary data and then transmitted over a noisy channel. Assume that each bit of the encoded message is altered during transmission with probability $p$, where $0 \le p \le 1$. We want to send a long encoded message over this channel, and to correctly decode it with very high probability; let's say that after the transmitted file is received and decoded, we should recover an exact copy of the original 99.9999% of the time. (There is *no* code that will enable us to recover the original file 100% of the time over a genuinely noisy channel.) Please note that we are not just asking for the received file to be 99.9999% correct; that would *not be good enough!* For example if we wanted to download a 3MB binary executable file over the internet, and every time we got a file that was correct except for 3 bytes (24 bits), we would never get a clean copy that our operating system would accept!

It is not hard to see that we can attain perfectly decoded copies of large files if we simply send the same file many, many times. However, this is expensive; we have seen that repetition codes have a rather low information rate among all possible codes with a given error-correcting capability. Shannon's Theorem tells us that we can do better: the optimal information rate for reliably transmitting, decoding and receiving information through a noisy channel is in fact $1 - H(p)$.

For example if $p = 0$, this tells us that we can send information "as is" without any encoding/decoding required.

The noisiest possible channel is one for which $p = 0.5$, in which case about half the bits are altered during transmission. In this case no information can be reliably communicated through such a channel. This follows from Shannon's Theorem, which tells us the optimal information rate is $1 - H(0.5) = 0$.

It may be surprising at first to see that the case $p = 1$, in which *every* bit is altered by the channel, allows for information transfer at a rate $1 - H(1) = 1$. This is because if *every* 0 is altered to 1, and every 1 is altered to 0, then we can recover the transmitted message from every received message by a simple process of switching back every 1 to 0, and 0 to 1. However, it is hard to imagine conditions under which noise would alter more than half the transmitted bits, on average; so it is customary in fact to assume that $0 \leq p \leq 0.5$. Moreover any channel with $p > 0.5$ is equivalent (after the simple trick of interchanging 0's and 1's) to a channel with bit error rate $1 - p < 0.5$.

Imagine now a pipe along which water, or any other substance, can flow at a rate of at most 1.0000 gallons per second, say. If the pipe carries 0.2781 gallons of useless sediment per second, this means that there is enough room left in the pipe for water to flow at a rate of just 1.0000 – 0.2781 = 0.7219 gallons per second. Now imagine a channel which alters every bit with probability 0.2. If a bitstring $M$ of length $n$ is transmitted along such a channel, the message received will be $M + e$ where '$e$' is a random bitstring of length $n$, and the addition symbol '+' denotes bitwise addition mod 2 (i.e. addition of vectors of length $n$ over the symbols {0,1}). The entropy of the error source, as we have seen, is $H(0.2) = 0.2781$; this is the rate of transfer of *useless* error bits across the channel. The amount of room left in the channel for transfer of *useful* bits (the bits of the encoded message M) is therefore 1.0000 – 0.2781 = 0.7219, i.e. the maximum rate at which the channel can carry useful bits is 0.7219. In short: measuring the rate information can be transmitted over a noisy channel, is very much like measuring the rate at which water can flow through a pipe with a known capacity, which also carries useless sediments at a known rate.

## References


C.H. Bennett, 'Notes on the history of reversible computation', IBM Journal of Research and Development, vol.32 no.1 (1988), p.16.

R.P. Feynman, *Feynman Lectures on Computation,* ed. A.J.G. Hey and R.W. Allen, Perseus Books, Reading MA, 1996.

R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics, Vol.1,* Addison-Wesley, Reading MA, 1963.

C. Kittel, *Thermal Physics*, Wiley, New York, 1969.

C.E. Shannon, 'A mathematical theory of communication', *Bell System Technical Journal,* vol.27, pp.379–423 and 623–656, July and October, 1948.