relativity along these lines remains as elusive as a unification along any of the more conventional lines that have so far been suggested. Finding the correct unification presents the twenty-first century with one of its greatest challenges. If this challenge is successfully met, then it will have profound implications running far beyond those that we can directly perceive at the moment. It will not be met, however, if the strange and wonderful principles underlying Einstein's beautiful equation are not thoroughly respected.

In: G. Farmelo (Ed) *[2002].* It Must be Beautiful: Great Equations of Modern Science. Granta Books, pp 213-230.

Understanding Information, Bit by Bit

Shannon's Equations

Igor Aleksander

Information is now a commodity like metal or oil, a utility like water or electricity. Politicians, stock-exchange pundits, future watchers and the rest of us wave our arms and say that something is a sign of 'living in the information age'. Although not everyone knows what the information age actually is, there are signs of it everywhere: faxes in waste-disposal tips, e-mails from aircrafts, and mobile phones even on nudist beaches.

It is debatable whether we all have a lot more to say to one another, so it's not the actual information that is the focus of the information age. The hallmark of this revolutionary age is, rather, the amazing opportunity for us to connect to one another or to computers from almost anywhere. Fifty years ago we had just the telephone or the wireless. Now there are globally networked computers (the Internet), digital mobile phones and fibre-optic cables. Even consumer entertainment products have changed out of all recognition. We have gone from the 78-rpm black vinyl record to the digital video disc, from the Brownie box camera to the digital models that grace the shelves of camera shops.

This all points to huge developments in the technology that underpins the transmission of information. But what is information? What inhibits its transmission? Why have these developments required massive industrial investment? Why does the word 'digital' (meaning that which is represented by separate symbols such as numbers) appear so often in the names of these technologies?

I do not intend to provide a detailed description of how this technology works. Rather, my aim is to rediscover a hero of the information age, a man without whose razor-sharp insight none of this technology would work. Claude Shannon was both a mathematician and an engineer, and he brought these disciplines together in a way that changed the world forever.

Shannon's name is attached to two equations that underlie the theory of communication. They have a somewhat forbidding notation:

$$I = -p \log_2 p$$

and

$$C = W \log_2 (1 + S/N)$$

The first of these tells us that the amount of information in any message can be measured as a quantity labelled I, where the unit of measurement is the 'bit'. While bits and the word 'digital' appear often in these descriptions, the two equations are continuous and not digital, which means that the theory applies to old telephone lines as well as the latest digital versions. The statement made by the first equation is that the amount of information I depends on the *surprise* that the message holds. This is because the mathematical way of expressing surprise is as a probability p; the less probable an event is, the more surprising it is and the more information it conveys. Where the $\log_2$ comes from we shall see later. Suffice it to say that without this equation, the world would be without a major unit of measurement, a measure which is as important as the gallon, the litre, the watt or the mile.

The second of Shannon's equations is a 'quality' indicator of a transmission medium such as a telephone line or a cable for a television aerial. It tells us that C (in bits per second), the amount of information that can be transmitted through a line or other medium, depends on two major factors: W, the bandwidth (or range of frequencies that can get through), and S/N, the signal-to-noise ratio. We get a feel for this when, at a noisy cocktail party, we need to shout (or increase S, the signal, to beat N, the noise). If talking to a partially deaf person (someone whose W is restricted), we need to shout even harder. So, using the analogy of miles and gallons, C in bits per second is a quality factor in the same way that 'miles per gallon' is a quality factor for a motor vehicle. These laws are very general: they apply to anything from the simple telephone connection that transmits voice signals translated into

electrical quantities, to the latest digital high-definition television where visual scenes are translated from strings of numbers.

Shannon's thought and work transcend the equations themselves: they are merely the symbols that distil an exceptional insight into the nature and harnessing of information.

Shannon's very anonymity is evidence of his success. Even the most hardened Internet aficionado just sits there, switches on the machinery and expects text and pictures to appear on the screen of his personal computer. Let us say that his friend Jill has promised to send him the latest digital portrait of her face. She clicks on this image on her computer, 'attaches' it to an e-mail message and clicks on the 'send' button. But how can the two computers be connected? Is it via a telephone cable? Were this true, our recipient Jack would have to wait over thirty minutes for Jill's face to appear on his screen. I know that delays on the Internet sometime feel terribly long, but the very fact that this transmission would, at worst, take about a minute and a half, is due to Shannon's discoveries. Indeed the very fact that I could predict the thirty minutes, knowing something about telephone cables, is because Shannon taught us *how* to assess the picture and the cable in order to work out how well the job could be done. Enormous design efforts based on such assessments went into the Internet so that we can send each other pictures of our lovers and children.

The Internet is a system of interconnections between hundreds of millions of computers all over the world. It spews information into our computers in much the way that a tap fills our bath. Just as we buy ½-inch or ¾-inch piping for our bathrooms, so we buy appropriate links for our computers to the Internet network. Our Shannon-inspired calculation tells us that bare telephone lines are not appropriate because of their inadequate capacity to handle the large amount of information contained in something like a photograph. Leaving aside the question of what a 'bit' is for later, capacities are measured in terms of bits per second: the higher the value, the faster will my computer fill up with information. Jill's picture contains, say, 20 million of these bits and a telephone line has a capacity of 10,000 bits per second.[1]

20,000,000/10,000 seconds = 2,000 seconds = 33.3 minutes.

And yet what connects our home PCs to the Internet is nothing but a telephone line. Imagine a world in which we could not measure the gallons of water that we consume or the kilowatts of electricity supplied by the local electricity board. That would have been the world of information without the

work of Claude Shannon: the world without the 'bit'. Shannon gave us the measure of information, but he also founded the entire subject of information theory, which anyone designing a communication network needs to know.

Shannon was born in 1916 in Gaylord, Michigan, the son of a businessman and a schoolteacher. He soon demonstrated an aptitude for mathematics and engineering and, like many young people in that era, enjoyed tinkering with radios, the hottest technology of the time. He even earned money from the local department store by repairing radio sets.

When he was sixteen, he entered the University of Michigan to study mathematics and engineering. Four years later he became a research assistant at the Massachusetts Institute of Technology (MIT) in Cambridge, where he worked on early computer projects with the charismatic Vannevar Bush, who later became President Roosevelt's scientific guru and – in some people's eyes – a founder of the Internet.

When Shannon arrived at MIT just before the outbreak of World War II, computers had hardly been invented. The word 'computer' was rarely used. Calculating machines were largely mechanical devices using gearwheels, springs and the like. Some laboratories were looking at electronic or mixed mechanical and electronic calculating machines. Vannevar Bush was one of the very few people who had the confidence to restate Charles Babbage's dream, enunciated a hundred years earlier, that mechanical machinery could take over from human beings some of the drudgery of doing repetitive calculations. Such notions were considered by some at the time to be fanciful daydreams. However, Bush and, later, John von Neumann, grandfather of the current style of computer design and, according to Einstein, one of the most agile minds that had ever graced Princeton, were both strategists highly regarded by US government agencies. So they were able to impress on the government the importance of mechanized computation, thereby unlocking early support for the design of computers. Without Bush and von Neumann, electronic computers would not be nearly so advanced as they are today.

As an ironic contrast, when the British computer pioneer Maurice Wilkes at Cambridge in the UK attempted to obtain funding in the late 1940s to build a computing machine, he received a dusty response from what was then the Department of Industrial and Scientific Research. The bureaucrats effectively suggested that if Wilkes and his colleagues were to sit down with a few mechanical calculators, they could solve all the world's computational problems, avoiding the need to build fancy computing machines.

So Shannon was lucky to be taught by a powerful visionary, a factor that must have contributed to his own lack of fear in undertaking ambitious engineering challenges. But MIT was then, as it is now, an expensive place even for someone endowed with a grant that would cover his fees. Vannevar Bush had invented a calculating engine called a 'differential analyser' (DA). This machine stores numbers on rotating geared cylinders, a bit like mechanical mileage indicators on old cars. To help the young Shannon scrape a few dollars together, Bush offered him part-time work on the DA, which Shannon took up with some gusto. The DA was an experimentalist's dream. It was a large collection of rotating cylinders, gearwheels and electrical control switches. Its main function was to find the solutions to mathematical equations. These were set up by interconnecting the parts to suit the equation. The answer was given by a reading on the mileage-meter-like device. It could take days to set up the machine just to work on one problem. It would then have to be dismantled and reconstructed to work on the next one. So Shannon became one of the world's first programmers, setting up the DA to meet various scientists' needs.

These were formative times for the young Shannon. He became aware of the need to understand two major relationships in information. First, there was an *amount* of information which the DA computations generated, and, second, *a limiting speed* with which the output indicators could accept the information that had been calculated. Amount of information and speed of transmission, these were to become the pillars of Shannon's information theory of the future, the two topics of his two celebrated equations.

Another major factor that influenced Shannon was his fascination with electrical switches and the complex systems for routing electricity that could be designed just with a handful of such switches (think of the way that a light can be switched on from two sides of the same room). He had studied the laws of logic as set out by that British pioneer George Boole who a century earlier at Queen's College, Cork, in the south of Ireland had presented these as 'the laws of thought'. For example, were you to say, 'Alvin and Bob were not both at the party', this is the same as saying, 'Either Alvin was not at the party or Bob was not at the party.' Boole had suggested a notation (now called Boolean algebra) in which the statements above could be turned into a rule which is always true –

$$\text{Not(A and B)} = \text{(NotA) or (NotB)}$$

where A and B are statements that are either true or false. Boolean algebra has many such rules.

The reason for mentioning all this is that switches form the basis for both routing and storing information, and Shannon made the following daring intellectual leap. A closed switch is like a 'true' statement and an open switch is like a 'false' statement in logic. So if A and B were switches instead of statements, Boolean algebra could apply to the way that communication networks can be switched to connect communicators and switches organized to store messages. Indeed, the mass of switching that needs to be done inside a computer is now regularly designed or analysed using Boolean algebra. So before the end of his first year at MIT, Shannon had written his master's thesis on the application of Boolean algebra to switching circuits and, in 1938, published a paper called 'Analysis of relay and switching circuits'. This became one of the classic papers in computer literature, and Boolean algebra is now regularly taught to first-year engineering students as the standard way of designing switching circuits in both computers and telecommunication systems. Not a bad discovery for a twenty-year-old! Indeed, Shannon's mind at this early age must have been driven by the passionate desire to tie together the nature of switching with the nature of information and to understand the limits on how fast information can be transmitted from one geographical point to another.

Shannon left MIT in 1940 with both a master's degree and a doctorate in mathematics. MIT now honours this association by having a regular Shannon Day, on which the latest advances in telecommunication are discussed. After a year at the prestigious Institute for Advanced Studies at Princeton, Shannon joined the premier industrial research establishment in the United States: the Bell Telephone Laboratories at Murray Hill in New Jersey. Here, urged on by colleagues, he published in 1948 his internal reports on a statistical theory of communication. This was the celebrated 'Mathematical Theory of Communication'. As an example of the logic that led Shannon to quantify communication, in 1950 he wrote the very first chess-playing program which incorporated a clever way of cutting down on the number of board positions that the machine has to search to find a good move. This 'algorithm' was used in the programming of IBM's Deep Blue machine that beat the grand master Gary Kasparov in 1997 – the first time a reigning chess champion had been beaten by a machine.

By 1957, Shannon was widely recognized in the United States as one of its leading scientists. He was identified as one of the nine leading lights of American science in a special feature in *Time* magazine, published six weeks after the USSR successfully launched the first artificial satellite Sputnik, which had caused a panic in the United States as it appeared they

were falling behind their cold-war rivals. In his profile, we learned of his addiction to jazz, his enjoyment of science fiction and his working habits: 'like many scientists, [he] works best at night, with plenty of cigarettes and coffee.'

In World War II movies such as *The Cruel Sea*, the Morse-key operator taps SOS as di-di-di, dat-dat-dat, di-di-di to tell the world through his radio transmitter that the ship is in trouble and may be passing its last minutes afloat. But why was the operator not merely picking up a microphone and using his voice, through his radio transmitter, to tell the world exactly what was happening? The answer is that the 'di' and 'dat' pattern of two simple tones stands a much better chance of getting through the crackle and hiss of the radio transmitter than spoken words, whose multifarious and subtle tones would be lost in what engineers call electronic noise. It is easier to see a dim light far away than to make out the details of a picture at the same great distance. A flashing light with a proper code can tell a story that a poorly seen picture cannot. But this vague notion needs a theory, and this is the theory that Shannon worked out during his early years at Bell Labs.

All communication systems suffer from noise: it sounds like crackle in a telephone, it looks like snow on a television screen. Noise distorts unpredictably the information the sender is trying to transmit to the receiver. This can make the received information unintelligible, hence useless. There is another limitation, something that specialists call bandwidth. Most buyers of hi-fi are aware of this. They first ask what is the 'bass' response – that is, what are the lowest frequencies (rumbling tuba sounds) that can be heard on the equipment? – and then what is the 'treble' or high-frequency response (top notes on a violin) that the equipment can handle? Subtracting the low limit (say 25 cycles per second, called 25 hertz) from the high-frequency limit (say 5,000 hertz) gives the *bandwidth* of the equipment. In other words, poor equipment with low bandwidth will not allow a listener to delight in the full glory of an orchestra's range. In technical terms, like noise, bandwidth makes the received information somewhat less than the transmitted version. Every communication link between a transmitter and a receiver is characterized by some value of its bandwidth and Shannon wanted to predict exactly how these losses in information could be calculated.

Shannon summarized the situation in a way that became the basis of information theory itself. To tidy things up, he imagined that every link between the source of information and the destination of its messages has

five major components. First, there is the source. In the case of someone who wishes to transmit a digital picture across the Internet, the source is a computer in which the picture is stored as 20 million 0 or 1 states (or bits) of a memory.

The second element is an encoder. This is the sum total of equipment that prepares the picture to be transmitted in a reasonable time over, say, low-bandwidth telephone lines. As a first step in encoding, a modern computer has a program which 'compresses' the picture. The picture sits in the machine's memory as a sequence of numbers, each number representing the colour and intensity of a dot on the picture. Compression removes redundancies in these sequences of numbers.[2] The next part of encoding is to turn the numbers that represent the picture into 'tones' that can be transmitted down the telephone line. This is called 'modulation' and is needed because the telephone channel is designed to carry audible signals; that is, the human voice. Most telephones now use tones to go with each dialled number. This is an example of modulation.

The third element of the communication system is the telephone cable itself, with its inherent noise and restricted bandwidth. The fourth part is a decoder that restores what is received to a state as close as possible to what was transmitted. In the case of a picture, the decoder must first take the tones and turn them back into numbers, and then it must interpret the numbers so that they reconstruct the picture in the fifth part of the system, which is the receiver: the computer screen in this case. Anyone who has bought a modem for their computer will in effect have bought a box that contains both an encoder (*mo*dulator) and a decoder (*dem*odulator).

If this is a crude picture of a communication system, it is also worth taking a romp through a kind of caricature of the theory that explains the system, and then looking at the effect of these equations on what we now know of communication systems.

First, we need a measure of information.[3] We already know that this is measured in bits, each of which has two values, 0 and 1. The bit, or '*bi*nary uni*t* of information', was one of Shannon's key proposals. Let's work backwards and see why this makes so much sense.

Suppose someone wants to transmit a picture of herself, but also those of her father, mother, two brothers, the family dog, cat and a picture of the family house: eight pictures in all. Having stored them on his computer once, her boyfriend has acquired the information. If she wants him to see any one of these pictures, all she needs to do is to number them from 1 to 8

and transmit the appropriate number. The computer merely places the appropriate picture on the screen without it needing to be transmitted. Now, one bit could specify two numbers, 0 and 1. Two bits can do four (00, 01, 10, 11). Three bits can do eight (000, 001, 010, 011, 100, 101, 110, 111). So here is an illustration of one of Shannon's major insights: information is proportional to how much you don't know. When the man in this woman's life had no information at all about these pictures he needed 20 million bits for each to describe them. Once he has stored them, he needs only three bits to specify one of the pictures. This is how probability creeps into the first of Shannon's equations: in the first instance, had he never seen his girlfriend before, this would have meant an enormously low probability of guessing what she looked like. The less likely an event is, the more information its occurrence carries. Shannon's first equation links the information to something known as the 'logarithm to the base 2' of the probability (written as $\log_2$). This somewhat off-putting jargon is, however, not hard to understand. To take a few simple examples, the logarithm to the base 2 of $2 \times 2 \times 2$ (i.e. 2^3) is 3; the logarithm of $2 \times 2 \times 2 \times 2$ (i.e. 2^4) is 4; the logarithm of $2 \times 2 \times 2 \times 2 \times 2$ (i.e 2^5) is 5. So the logarithm to the base 2 of a number is simply the power to which 2 must be raised to equal that number. It is now possible to return to the complete equation:

$$I = -p \log_2 p \text{ (measured in bits for reasons we shall see).}$$

This reads: 'The amount of information involved in gaining knowledge of an event is dependent on the probability p of that event occurring'. This also leads us neatly to the definition of the bit. Toss a coin. This can result in one of two events: heads or tails. Each event has its own probability of occurrence, and if the coin is unbiased, each of these events has a probability of 1/2. To get the total information involved in the event of tossing this coin we add the information content of both possibilities and get:

$$I = [-(\tfrac{1}{2}) \log_2 (\tfrac{1}{2})] + [-(\tfrac{1}{2}) \log_2 (\tfrac{1}{2})].$$

This turns out to be precisely 1. This is no accident, and to a mathematician it explains why Shannon used $\log_2$ in his equation. That is, the unit of information, the bit, is associated with a switch being on or off, a number being 0 or 1, a coin being heads or tails, and ensures (as we shall see) that *any* other amount of information can be measured in bits. The formula also covers the case of certainty. If the occurrence of an event or its nonoccurrence is certain,

the equation tells us that the event yields 0 bits of information. Two bits are like two coins that can result in four messages, so the information content of four equally probable messages is two bits. This means that the equation can be applied to any number of messages. For example, if the upper-case letters of the alphabet are being transmitted from a word processor this implies 26 messages, which requires five bits (which as 2^5 gives us 32 messages, that is, a few spares over the 26). The power of the first little equation thus allows us to measure unequivocally how much information is contained in something we are trying to communicate to someone else.

The words 'crackle and hiss' have already crept into this essay. It is one of the laws of nature that this 'noise' cannot be avoided whenever electricity or wireless media are used to convey information. Were I to connect two computers with the telephone cable, there would be a level of electrical energy arriving at the receiver that is not transmitted by the transmitter. The information we want is sent down the cable encoded as a sequence of voltage levels each representing, say, a picture point in Jill's picture.[4] But electrons in the cable have a habit of jumping about. This random activity modifies the voltage being transmitted so that the receiver may get a randomly modified number. Such random activities exist not only in transmission down cables; the free space that is used for radio transmission also has sufficient randomly moving charged particles to produce significant variations in transmitted signals. These variations will appear in the received picture as dots or 'snow', or a hiss in the messages transmitted by the Morse-code operator on the doomed ship at sea.

This is where Shannon's other equation comes in. The imperfections of a channel in terms of frequency limitation (bandwidth W) and noise (N) may all be incorporated into one statement for the capacity C of a communication medium with respect to the strength of a signal S.

$$C = W \log_2 (1 + S/N) \text{ in bits per second.}$$

We need to unpack this a little, using, again, a picture transmitted over the Internet. First, we make a rough assumption that the bandwidth W is a 'maximum' frequency of transmission by assuming that the lower limit of the bandwidth is zero. Even more roughly we take this to mean 'the maximum number of packets of bits we can transmit in any second'. A packet of bits represents a range of numbers (3 bits give 8 numbers, 4 bits give 16 numbers, and so on). Now, this range of numbers depends on how much noise there is in the system. The term $(1 + S/N)$ tells us how likely noise is to change the number in the packet. So if there is no noise, N is 0 and $(1 + S/N)$ turns out to be infinity (plus 1), which tells us that each packet could be as big as we like. So the whole of our picture could be enveloped into just a single packet which could then be transmitted W times in a second! With values of W of 10,000 or so even for very poor lines, communication would be prodigiously fast in a noise-free line. Sadly, noise is always there, and if, say, the noise were to be ⅐ of the signal strength, $(1 + 7)$ tells us that anything more than 8 numbers would mean that the transmitted number was changed by noise. So in this case $\log_2(8) = 3$, which means that only three bits per packet can now be transmitted W times a second. So for W being 10,000, even the compressed version of the picture of, say, 1 million bits would take:

$$1{,}000{,}000/(3\times10{,}000) \text{ seconds} = 33.3 \text{ seconds}$$

(The uncompressed version would now take about 10 minutes.)

Were the noise to equal the signal strength, then $(1+S/N)$ becomes 2 and only one bit per packet could be sent, making the compressed transmission time about a minute and 40 seconds. As the noise gets even greater, $(1 + S/N)$ tends towards 1, which means that no bits per packet can be transmitted, because $\log_2(1)$ is 0.

For the Morse operator on the sinking ship, it is the high level of noise that does not permit the transmission of speech (which needs about 8,000 bits per second). It does, however, allow the three or four bits per second of the dits and dats which, with Morse's clever encoding, get the essence of the message across.

The simple telephone cable is not the only medium for the transmission of information, with its 10,000 bits per second of information. There are all sorts of cables and other transmission media that have much larger bandwidths. 'Coaxial' has a solid wire in the middle, surrounded by a plastic insulator and sheathed by a flexible outer metal sleeve. The bandwidth of this cable can be 200 million hertz (or in other words 200 megahertz – this is the same as 400 million bits per second). Clearly this allows much faster communication between computers, but it is also a little more expensive. Even larger bandwidths may be obtained with fibre-optical cables, which instead of transmitting electrical pulses transmit optical ones (laser-generated light, actually). FM (frequency-modulated) radios pick up stations

around 100 megahertz. This means that the free space, in which information travels as electromagnetic waves, has a vast bandwidth.

That's all very well, but even the most exquisite broadcast of classical music needs only a bandwidth of 30,000 hertz, so how are much larger bandwidths exploited? Shannon's concept of the encoder or the modulator explains how. To reduce the problem to that of simple numbers, our digital pictures come in handy again. We saw earlier (see note 2) that each picture point requires $128 \times 128 = 16{,}384$ numbers, which corresponds to 14 bits (because $\log_2 (16{,}384) = 14$). Say we have a bandwidth/noise situation that is plentiful and we wish to transmit, say, eight times that number in the same time. This suggests that we might try to transmit eight pictures simultaneously if we could only use the bandwidth as eight separate channels rather than just one. It can be easily done. Anything that goes into channel 1 will be given the number 1 as a prefix. Channel 2 will get 2 and so on. So for every period of time we transmit a group of eight numbers, each prefixed by its channel number. At the receiving end the decoder must be built so that these channel numbers are detected and the picture points separated. The channel numbers are 'carriers' of the information related to each channel.

Something very similar happens when we tune into the station of a radio. We tune into the carrier of the particular channel and the radio set decodes the content of that channel. So the free space bandwidth of, say, 300 million hertz can cope with 10,000 different radio stations and more (as not all demand a bandwidth of as much as 30,000 hertz). This encoding also takes a particularly interesting form on the Internet. The coding numbers are things like jack@toc.ac.uk, which may be Jack's e-mail address to which Jill sends her pictures so that Jack and only Jack gets them. This curious form of coding means that, in the case of e-mail, Jill's message with its channel code rushes about this vast network in order to find Jack's computer whose address is the destination of the message. The destination computer then decodes Jill's pictures and delivers them to Jack's screen. So this establishes a unique channel between Jack and Jill despite the megajungle of cables, satellite links and radio transmission that the Internet implies.

It should be stressed that 'compression' is very much part of the encoding and decoding process as envisaged in Shannon's five-stage scheme of 'source–encoder–channel–decoder–destination'. Those who use the Internet and download pictures or movies will be aware of standards that have names such as JPEG (for stills) or MPEG (for movies). These are encoding and decoding protocols that save the Internet user hours of waiting for data to be downloaded. So wherever we look in the vast world of modern telecommunication we find that Shannon's model of the nature of information is enormously helpful in the design of the systems that enable high-speed communication.

An unexpected spin-off from Shannon's definition of the bit is that not only is it the unit of transmitted information, it has also become the unit of information storage or 'memory'. A single switch can be on or off, in accordance with the definition of a bit as being the carrier of only two messages. A switch therefore records, memorizes or stores one bit of information. Two switches can be in four combinations of *on*ness and *off*ness and the $\log_2$ element of Shannon's equations again comes in useful because, in order to store, say, one million messages, the number of switches required is given by $\log_2 (1{,}000{,}000)$, and turns out to be only a small number of 6 bits – about 20. It is this relationship that gives computers their prodigious powers of memory. Anyone who uses a reasonably up-to-date computer will be aware of at least two kinds of memory: hard disk and random access. Typically a hard disk may store 5 gigabytes. For an unimportant reason eight bits are called a byte, so 5 gigabytes turns out to be 40,000,000,000 bits. The hard disk is a rotating piece of metal on which a bit is stored by magnetizing a local patch of the metal through the use of a 'head' which becomes a magnet (or not, depending whether it is fed an electric current). The patch is either magnetized or not and hence much like a switch: it stores one bit. The rotating disk leaves a trace of set and unset switches. These can be 'read' by the same head because magnetized patches induce a current in this head that then can be transmitted as an informational bit or used in the computer in some other way. The reason that a computer also has a random-access memory is that the hard disk is relatively slow by virtue of its rotational inertia. It may make it necessary to wait (a hundredth of a second, roughly) before a particular patch of metal can be accessed. The random-access memory is much faster (it needs a small fraction of a millionth of a second for access). It allows the state of a tiny silicon switch to be accessed in the way we access a file in a filing cabinet. We need a tag like 'tax' or 'mortgage' which identifies the desired folder. A glance is enough to pick out the file. In a similar way, every silicon switch has a tag called an 'address' which, when applied to the whole bank of switches, will pick out the right one with the matching address. So the random-access memory is fast, but not as large as the hard disk.

With these ideas we can now build a much more complete picture of what happens when we download a picture from one computer to another.

If it was taken with a digital camera, the picture is first sensed by special light-sensitive electronics and then translated into stored bits in the camera's own random-access memory. This is then transmitted (through the use of software and appropriate cables) to the hard disk of the sender's machine, where it occupies 20 million out of the 40,000 million magnetic switches. If she wants to view the picture on the screen of her machine, she moves it into the random-access memory of her computer. This in turn is made by certain programs to move the bits onto the computer screen, where the electrical energy of the bits is turned back into light patterns. Then, when the recipient asks for the image to be downloaded, it is transmitted over cables and is transferred into his random-access memory and viewed on the screen. To keep it permanently, he transfers it to his hard disk.

The storage capacity of computers has grown prodigiously. Taking just the area of storage devoted to our single picture, with Shannon's $\log_2$ relationship it is possible to ask how many different pictures could be represented in this space. The answer is x in

$$20{,}000{,}000 = \log_2 (x).$$

x turns out to be roughly 10 followed by 7 million zeros, an astronomical number.

Shannon's brainchild, the bit, has not only given us a means for measuring information as a utility, it has become the very currency of computation. Technology has thundered on, but Shannon's insight and his formulations of half a century ago remain unshaken. The vast versatility of the computer and the even more awesome power of millions of interconnected computers on the Internet have created organisms where the complexity is beginning to be way beyond our grasp. It's $\log_2$ that does it.

Why is it that everything is going digital? There are now digital telephones where there were ordinary (analogue) ones, there is digital radio and television and we pay more for digitally recorded music on compact discs than reproductions of the older analogue albums. The whole world of consumer communication products is going digital. Governments are behind this move (although they sometime fail to give cogent reasons for their support). Shannon explained this vast wave of change when he defined the maximum capacity of a channel in the second of his equations, and when he defined a standard encoder–channel–decoder structure for any electronic communication system.

'Digital' simply means that data is transmitted as discrete symbols. But to drive this point home it may be easier to say what is not digital. Human communication at its most direct is not digital. When I speak, I create pressure waves in the air by the movement of my vocal cords, the shape of the cavity behind my mouth and the configuration of my tongue and lips. These waves reach my hearer's eardrum, causing his cochleae (coiled-up trumpet-like organs in the ear) to translate these pressure waves into internal neural signals and transmit sensory information to the brain, which the recipient describes as 'hearing'. But as soon as the communication becomes electronically aided, the possibility of translating these waves into sequences of numbers encoded as bits becomes an option. This system is then digital: that is, no more waves, just information as bits.

While Shannon's theory applies to both analogue and digital systems, the theory itself shows that digital is most efficient, and if the digits are binary this is the best of all. The argument is based on cost and goes a little like this. Let us say that a picture point requires 256 numbers. The channel here can be thought of as not being digital, just a box that needs to be a certain size to accommodate all the numbers between 0 and 255 as, say, a number of little cubes. Now we introduce the cost of that channel. This is the cost not of the cubes but of the box that must carry them. It is natural to think that the larger the number it carries, the more it would cost. This channel costs 256 units of some currency.

How much would it cost were we to use two smaller channels to carry the same information? The boxes would have to carry a mere 16 numbers because the combination of the numbers in the two boxes would give us 16×16 possible numbers, which brings us back to the necessary 256. But the total cost of the two channels is now $16 + 16 = 32$ units, rather a worthwhile saving. So why not keep going in the same direction? We note that the process stops when each channel holds only two numbers, when we have eight channels to give us $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 256$. The cost of this is $2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 = 16$ units. According to the late Norbert Wiener of MIT (the 'grandfather of cybernetics'), this was proof of Shannon's brilliance in defining the bit. Each of the channels in the lowest-cost system is a binary channel that carries only one bit, and this is by far the most economical way of transmitting information.

It is the economic efficiency of binary encoding that is causing the world to go digital. Yet again Shannon's $\log_2$ – which features in both his equations – is at work, because if someone is using nondigital waves, say, that have amplitudes that need to represent some integer number A precisely,

Shannon tells us that it costs a great deal less to use bits to transmit the same information. Probably the most vivid illustration of this is the progression in recorded music from long-playing records that produced waves through wavy grooves and needles which needed a relatively vast area to store thirty minutes of music per side, to the modern digital video disc (via the compact disc). The DVD can store up to four hours of music in ½₅ of the space of the LP by using pure digital techniques. But the same goes for other things. Mobile phones and cordless phones work noticeably better through being digital. And all this because of the $\log_2$.

So these two equations have changed our world of communication:

$$I = -p \log_2 p$$

$$C = W \log_2 (1 + S/N).$$

I have argued that despite their formidable appearance, the real power of these equations lies in the rough relationship that underpins both:

$$\text{Information in bits} = \log_2 (\text{what needs to be communicated}).$$

Shannon's framework (source–encoder–channel–decoder–destination) holds true whether we send digital pictures using the latest Internet technology, whether we are chatting over our mobile phone or we are talking to one another in a noisy pub. It is within this framework that the first equation is born: a very general definition of information based on surprise and probability. But the important message from this equation is that if the probability of an event is 50 per cent, it contains exactly one bit of information. This can be extended to the more general idea that anything that constitutes a real transaction may be broken down into an appropriately sized string of bits. The second equation then concentrates on the nature of the channel: the telephone lines, the free space or the noisy pub. Shannon showed that there is a limit to the number of bits per second that can be transmitted in the given medium, a limit set by the bandwidth and the noise in the channel. The way to exploit this limit with the greatest economy is by digital encoding. The trick in this exploitation is to design ever-improved encoders that take raw information and turn it into optimally encoded strings of bits. Entire industries have been built around this problem of encoding, as can be seen with mobile telephones and the encoding of music and video for entertainment.

Shannon's effect is not restricted to the world of communications. The form of the equation may be found in other fields of science under the heading of 'entropy', that is, in the degree of disorder of a physical system. This, in information theory, is expressed as the degree of surprise. However, Shannon's formulation has shown that information behaves according to the laws that govern physics, thermodynamics, physical chemistry, and are well known to mathematicians. Information theory was a specialism for which only designers of electronic equipment had a vague feel before the 1950s. Shannon showed that it is a material equivalent to the fundamental particles of the universe and that it has a system of laws equivalent to those that rule these particles. In my own work too, which has to do with modelling the intricate architecture of the brain, the language of information theory reigns supreme. The storage capacity of brain cells may be measured in bits, and the anatomy of interconnections between the many modular areas of the brain can be analysed using the ideas of channel capacity.

The retiring person who unleashed this innovative ferment, Claude Shannon, is one of the technological giants of the twentieth century. The word 'technology' here is perhaps inappropriate, because Shannon made a major *intellectual* contribution to our contemporary world. Shannon was curious about complex things. Shannon's equations are not about nature, they are about systems that engineers have designed and developed. They are equations that elegantly capture the complexity of information and the means whereby it can be stored and transmitted. Shannon's contribution lies in making engineering sense of a medium through which we communicate. He shares the same niche as other great innovators such as his boyhood hero Thomas Edison (who turned out to be a distant relative, much to Shannon's delight) and Johann Gutenberg. Like the printing press, the Internet is a celebration of human language, the characteristic feature of conscious human beings. In much the same way Gutenberg's imagination was triggered by the turn of a wine-press screw, Shannon's was stimulated by the click of a differential analyser switch.

Having completed a brilliant academic career, Shannon retired from MIT in 1978, becoming professor emeritus and a deeply respected elder statesman of US science. In 1985 he was awarded the Kyoto Prize, the computer world's equivalent of the Nobel Prize. After he retired, he pursued a wide range of interests, including a mathematical theory of juggling, the design of a motorized pogo stick, and the development of a system for playing the stock market using probability theory. The end of his life was tragically blighted by Alzheimer's disease and he was too ill to attend the

unveiling of his statue in his native town of Gaylord, Michigan, in the autumn of 2000. He died on 24 February 2001 in a Massachusetts nursing home. His passing was politely recorded but it was plain that the media – busy participating in the information revolution – had, for the most part, little appreciation of the indubitably great figure the world had lost.

It is true that we nowadays use the word 'intellectual' to refer to those who make contributions to the humanities, philosophy and politics. It was not always so. The ideal Platonic and Aristotelian intellect included practical cunning and mathematical abstraction. Shannon changed the world by being a master of both.

Hidden Symmetry

The Yang–Mills Equation

Christine Sutton

Summertime in New York – hot and steamy, the stuff of movies. The year is 1953: Stalin is dead, Elizabeth II is the newly crowned Queen of England, and a young senator named John Fitzgerald Kennedy is about to marry Jacqueline Lee Bouvier. The paths of two young men cross as they share an office at the Brookhaven Laboratory on Long Island. Like a rare alignment of planets, they pass briefly through the same region of space and time. The juxtaposition gives birth to an equation that could underlie the Holy Grail of physics – a 'theory of everything'.

Robert Lawrence Mills and Chen Ning Yang were born a world apart, but shared a passion for theoretical physics. Yang, who turned thirty-one in September 1953, had come to the United States from China and gained his doctorate at the University of Chicago before joining the Institute of Advanced Study in Princeton, New Jersey. Mills, at twenty-six, was a new research associate at the Brookhaven Laboratory, who had studied at Columbia and Cambridge universities. In 1953 Yang was visiting Brookhaven for the summer and was allocated a space in the same office as Mills. Their paths soon diverged, but the Yang–Mills equation has ensured that after only a brief encounter their names have become inseparable.

Back in the 1950s, the Yang–Mills equation seemed the result of an interesting idea that had little bearing on reality, but by the end of the twentieth century it had come of age. It underlies the work behind the Nobel Prizes in physics in 1979 and 1999, and is important enough in mathematical terms