Claude Shannon and “A Mathematical Theory of Communication”

Parvez Ahammad, Konstantinos Daskalakis, Omid Etesami, Andrea Frome

October 19, 2004

1 Biographical background

Claude Shannon was born on April 30, 1916 in the town of Gaylord, Michigan. As Gallager describes it, he “led a normal happy childhood with little indication of his budding genius” [1]. Friendly, but not-outgoing, he showed a curiosity for how things worked. He attended public high school and graduated when he was sixteen, and earned his Bachelor’s degrees in electrical engineering and mathematics from the University of Michigan, Ann Arbor in 1936. His interest in Boolean algebra began during college, and it was during college that Shannon says he probably first read Hartley’s 1928 paper titled “Transmission of Information” [3] which he says had “an important influence” on his life [6].

After graduating, he responded to an advertisement for a position working with Vannevar Bush’s differential analyzer at MIT, and was accepted as a research assistant and graduate student in the Electrical Engineering Department. Shannon became particularly interested in the complex switching circuit that controlled the analyzer. After his first year, he spent the summer of 1937 working at Bell Labs on Boolean algebra and switching. Once back at MIT, he fleshed out how to use Boolean algebra to analyze and synthesize relay circuits, and this work was both his first published paper and his MIT Master’s thesis [7]. The paper was “quickly recognized as providing a scientific approach for the rapidly growing field of switching” [1] and in 1940 won the Alfred Noble prize for the best engineering paper published by an author under thirty.

In working on his Ph.D. thesis, Shannon switched from Electrical Engineering to Mathematics and partly at the behest of Bush, he looked for a topic in genetics, choosing to work on a mathematical basis for genetics. His Ph.D. thesis was titled “An Algebra for Theoretical Genetics” and was finished in 1940 [1]. It was never published so was largely unknown, leaving others to rediscover the results independently.

While Shannon was working on his thesis, he began forming the ideas that would eventually be in his 1948 paper. He spent the summer of 1940 working again at Bell Labs, and afterward took a fellowship to study under Hermann Weyl at Princeton. In an interview with Dr. Robert Price [6], Shannon said that when he took his fellowship, he told Weyl that he wanted to work on information and the measurement of information. This is where he began to work seriously on a mathematical theory for communication.

War was imminent the summer of 1941 when Shannon went back to work at Bell Labs, and he worked with a group on fire control for anti-aircraft batteries. However, in his spare time, he continued to work on switching and his theory of communication [1]. Shannon also became interested in cryptography during the war and, while he could explain his ideas to those working on cryptography at the lab, he didn’t have the clearance to really learn about the applications. Some of his work in cryptography was published later in 1949 [9], including a mathematical theory for secrecy systems which contained some early ideas related to entropy. Shannon insists that his primary interest was in information theory, and that he used cryptography to “legitimize the work” [6].

By 1948, he had been working on the ideas of information theory on and off for eight years without writing intermediate drafts or manuscripts; he had held the full work, *A Mathematical Theory of Communication* [8], in his head.

2 A Mathematical Theory of Communication

2.1 Elaboration on the nature of a Message

Shannon wanted to study communication in all its abstraction. In order to achieve that goal, he needed to look at messages -the exchange of which consists the reason for communication- from a very general prospective. Such a prospective evaded researchers of that time who had only the fuzziest idea of what a message was; only some rudimentary understanding of how to transmit a waveform and how to turn a message into a transmitted waveform. Shannon, on the other hand, started by explaining that the significance of a message derives from it being an alternative over a set of possible messages. This set is finite for a discrete channel and infinite for a continuous channel.

Looking at messages from that prospective is important for two main reasons that are interrelated. Firstly, because it abstracts from the message only the fact that it is a choice from a set -discarding it's waveform details- and, thus, justifies an information-theoretic characterization of messages. Secondly, because it completely detaches the form in which a message is produced from the form in which the message will be transmitted. Furthermore, it matches some intuition we have about communication; effective communication is nothing but the ability of one part to transmit the choice it made over a set of alternatives to another part. To an extreme, if the set of possible messages has just one element there is no need for communication!

2.2 Definition of Entropy for the Discrete Case

To provide a general framework under which communication can be studied, Shannon needed a measure of the information produced when a source chooses a message from a set; elsewise stated the uncertainty we have for the outcome of the source. The discrete case draws on Hartley's work, which showed that (for many examples) the number of possible alternatives from a message source over an interval of duration T grows exponentially with T, thus suggesting a definition of information as the logarithm of this growth. However, Shannon extended this idea by taking into account the statistical properties of the source that produces a message. And that is a great leap because it clearly separates the source of messages from the channel that is used for transmission.

For the simple case in which the output of a source can be modelled as a random variable X taking values from a set of symbols, say $\Sigma = \{1, 2, \ldots, N\}$, with probabilities $p_i, i \in \Sigma$, Shannon proposed the following measure for the information produced when a message is picked from the set of symbols, which he named *entropy of the source*:

$$H(X) = \sum_i p_i \log \frac{1}{p_i}$$

If the logarithm in the above measure is base two, then the entropy of the source is measured in $\frac{bits}{symbol}$.

The form of $H(\cdot)$ already existed in statistical mechanics as a measure of the disorder of a system, called *entropy of the system.* Shannon adopted the measure from physics and justified this choice by first listing a set of properties that we should require for a measure of information and subsequently proving that the form of $H(\cdot)$ is the only form -up to a constant factor- that satisfies that set of properties. Intuitively, one can see the beauty of the choice through simple observations like the following:

1. If the messages of the source are equiprobable, i.e. $p_i = p_j, \forall i, j \in \Sigma$, then the entropy of the source becomes:
$$H(X) = \log |\Sigma|$$
and, thus, meets Hartley's measure of information.

2. If one of the messages has probability equal to 1 and all other messages have cumulative probability equal to zero, then the entropy of the source becomes:
$$H(X) = 0$$
and, thus, meets the intuition that the source described produces, in fact, zero information (because we already know its output).

3. If there is a set of messages $\Sigma' \subset \Sigma$ such that $p_i = p_j, \forall i, j \in \Sigma'$ and $p_i = 0, \forall i \in \Sigma - \Sigma'$, then the entropy of the source becomes
$$H(X) = \log |\Sigma'|$$
and, thus, meets the intuition that the source described is equivalent to a source that produces messages from the set Σ' with equal probability.

Figure 1 shows the entropy of a source that produces two messages with probabilities p and $1-p$ for different values of $p \in [0, 1]$. The figure captures the observations 1 and 2 that we made above.

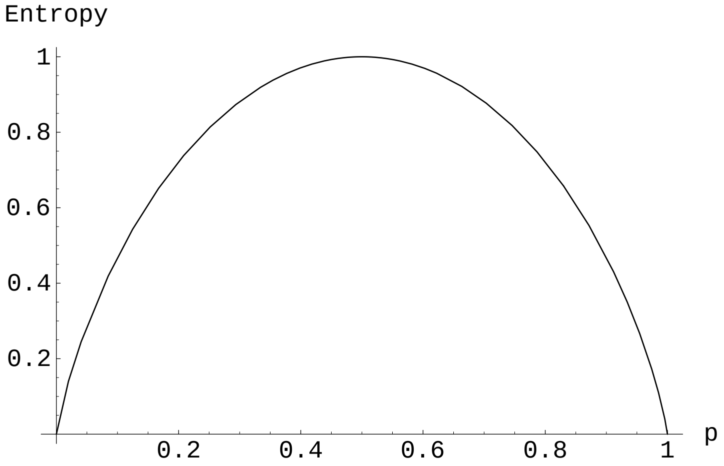


Figure 1: Entropy of a source that produces two messages with probabilities p and $1-p$.

For the general case of a source, Shannon argues that its output can be modelled as a random process. Using English as his major example (see section 2.4), he argues that, in fact, an ergodic Markov chain is a good approximation of a source in most cases and he extends the definition of entropy for ergodic Markov chains in a very natural way. We won't elaborate more on the generalization of the entropy in this presentation.

2.3 The Essence of the Source Coding Theorem

The source coding theorem justifies even more the entropy of the source as a measure of information. A high level description of the theorem says the following:

Theorem 2.1 (Simplified version of source coding theorem). *Suppose a source of entropy H. Then a message of length N symbols can be encoded (compressed) to approximately $H \cdot N$ bits and then decoded successfully with high probability. This cannot be done with fewer than $H \cdot N$ bits. In order to decrease the probability of failure we need to increase N.*

Sketch of Proof.

1. It's easy to prove that, when N is large, there is a set of approximately $2^{H \cdot N}$ messages with cumulative probability to appear almost equal to 1. Thus approximately $H \cdot N$ bits are enough to accommodate those high probable messages. The rest messages have negligible probability when N is large and can be encoded to one dummy bit string.

2. On the other hand, fewer than $H \cdot N$ bits cannot accommodate the $2^{H \cdot N}$ high probable messages.

□

The essence of the source coding theorem is precisely that, no matter how many different symbols the output-alphabet of the source has, we only need $H \cdot N$ bits to transmit a sequence of N symbols. Furthermore, we cannot do with fewer than $H \cdot N$ bits. Thus, the entropy of the source truly captures the amount of information -in bits per symbol- that the source produces.

2.4 English as an Example

Shannon did not hesitate to use English as his major example. In fact, he proposed several different approximations to English through Markov chains. Starting from the most naive approximation in which all letters of the alphabet are equiprobable, he presented approximations in which the frequencies of digrams or trigrams are taken into account, to proceed in more complicated models in which frequencies of words or even frequencies of pairs of words are respected. Via these approximations to English Shannon tried to justify his claim that in most cases a source can be approximated successfully by a Markov chain.

However, he went further than merely presenting possible models for English. He actually used the models he proposed to count properties of the English language. Using several methods he counted that the redundancy[1] of ordinary English is roughly 50%. Furthermore, he elaborated that the existence of multi-dimensional crossword puzzles is closely related to the redundancy of a language and argued that two-dimensional crossword puzzles are just possible if the redundancy is 50%. Moreover, drawing examples from the English prose, he argues that Basic English, whose vocabulary is limited to 850 words, has very high redundancy, whereas the language in James Joyce's book "Finnegans Wake", in which Joyce enlarges the vocabulary has smaller redundancy than that of ordinary English.

[1] *Relative entropy of a source* is the ratio of the entropy of the source to the maximum value it could have while still restricted to the same symbols and *redundancy of a source* is one minus the relative entropy of the source.

2.5 The Channel Coding Theorem

The channel coding theorem is probably the climax of the paper. The theorem shows the possibility of reliable communication over a noisy channel. Given a noisy channel, the theorem says that we can assign a number C to the channel, called its capacity, with the following properties:

- Information can be transmitted on the channel at rates arbitrarily close to the capacity of the channel and can be recovered with high probability, in fact with a failure probability exponentially small in the length of the message.
- Information transmitted at a rate larger than the capacity cannot be recovered with high probability.

The theorem was quite unexpected at that time, since the general consensus before 1948 was that in order to make the failure probability arbitrarily small, we have to decrease the transmission rate as well. However, Shannon showed that by a proper encoding of the messages that adds redundancy in the transmitted symbols and in which the number of transmitted symbols depending on each message symbol is large, we can obtain negligible error probabilities.

Shannon's theorem also gives a natural interpretation for the capacity C of a channel. Consider the amount of information that a symbol received from the channel provides about the symbol originally sent over the channel, or in other words, the difference between the entropy of the sent symbol before and after receiving it from the channel. The channel coding theorem asserts that the capacity C is equal to the maximum value of this difference over all channel input distributions:

$$C = \max(H(X) - H(X|Y)).$$

For example, this shows that the capacity of the binary symmetric channel with cross-over probability p is equal to

$$1 - H(p) = 1 + p\log_2(p) + (1-p)\log_2(1-p).$$

2.6 The Probabilistic Method

In this section we give a rough sketch of Shannon's proof of the channel coding theorem. As we will see, the crucial argument in the proof is a probabilistic argument.

Assume that we want to send a message of length T and rate R. There are approximately 2^{TR} possible typical messages. Consider the input distribution for the channel that attains the capacity. There are $2^{H(X)T}$ typical transmittable signals and $2^{H(Y)T}$ typical receivable signals. For each received signal, there are $2^{H(X|Y)T}$ different major candidates as transmitted signals. Similarly, for each transmitted signal, there are $2^{H(Y|X)T}$ typical noisy received signals. (See Figure 2.)

Consider encoding each of the typical 2^{TR} messages to a random element of the $2^{H(X)T}$ typical input signal. Let's see how we can decode a noisy received signal. After receiving the signal, we look at all the $2^{H(X|Y)T}$ input signals that can be a candidate for the received noisy signal, and we check if any of those input signals is the image of a message. If one such input signal, say M, is the image of a message, we return an arbitrary element of the pre-image of M. The probability of decoding failure is at most the probability that a message different from the original message is encoded to the set of candidate signals. But this happens with probability at most $2^{RT}2^{H(X|Y)}/2^{H(X)}$, which is small when $R < C$. This way, we showed that each signal is decoded with high probability if the code is chosen at random as above. Therefore, there exists at least one code that decodes signals with high probability, where the probability is over choices of input messages and channel noises.

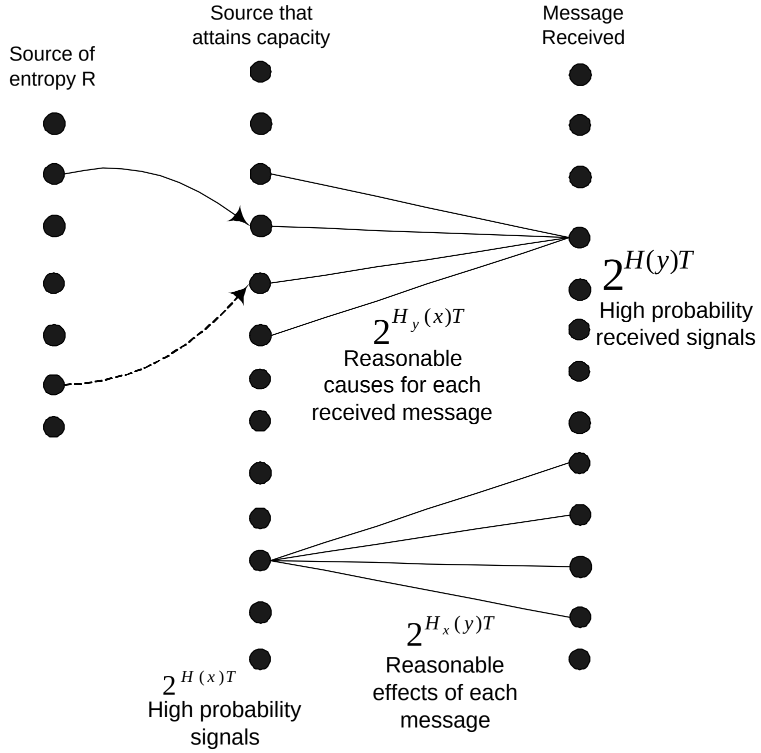

Source of entropy R
Source that attains capacity
Message Received
$2^{H(y)T}$
High probability received signals
$2^{H_y(x)T}$
Reasonable causes for each received message
$2^{H_x(y)T}$
Reasonable effects of each message
$2^{H(x)T}$
High probability signals

Figure 2: Illustration of channel coding theorem.

As we saw, Shannon proved the existence of good codes by showing that a random code is good with positive probability. (In fact, his argument implies the stronger fact that random codes are almost always good.) This kind of argument is now known as the probabilistic method.

Shannon's argument is considered as one of the first proofs that use the probabilistic method. The method is sometimes attributed to Paul Erdos for his 1947 paper that gave a probabilistic argument to lower-bound Ramsey numbers, since this is probably the most typical early application which formed the direction in which the method developed.

Interestingly, in the same vein, Shannon showed in a 1949 paper that the computation of a random Boolean function from n bits to 1 bit requires large circuits (exponentially large in n) with high probability. One major goal of complexity theory is to show that a specific function (corresponding to a natural computational problem) satisfies the exponential circuit lower bound. This goal has not yet been realized in complexity theory. However, as we will see in the next section, people have been able to find specific codes that certify Shannon's channel coding theorem.

2.7 Coding Theory

Shannon's paper does not address the computational efficiency of communication. Hamming, at around the same time as Shannon (1950) and also in Bell Labs, was working on explicit error-correcting codes with efficient encoding and decoding algorithms. He used the minimum distance between codewords as a measure of resilience to error. This way he analyzed codes in terms of the worst amount of noise they can handle — Hamming adapted an adversarial model in contrast to Shannon's probabilistic model. Although the rate of

Hamming's codes were much less than what Shannon had expected, eventually this line of research lead to a theory of coding that shows the existence of polynomial time algorithms for encoding and decoding at rates arbitrarily close to capacity.

2.8 Algorithmic Information Theory

In the late 1960s, Kolmogorov, Solomonoff, and Chaitin came up with a new notion of randomness, which is quite different from Shannon's statistical entropy. The idea is to define the complexity (Kolmogorov complexity) of a string S as the size of the smallest program that generates it. This way, a string is considered to have a high amount of randomness if its complexity is high. This notion of randomness is the counterpart of Shannon's notion of entropy. However, we can now talk about the randomness of the string S with no need to refer to the statistical properties of the source of S. It is easy to see that Kolmogorov complexity and Shannon entropy are very well related: Using Shannon's source coding theorem, almost all strings of length N generated from a random source with entropy per symbol H have Kolmogorov complexity approximately equal to HN.

3 Paper Discussion

Shannon's 1948 paper is a unique example where one paper can be traced back as the beginning of an entire field of research in modern times. As rare that is, it also points out to the significance of its place in modern scientific history. This paper is also considered to be one of the first papers to have used probabilistic method to prove the theoretical result along with Erdos' paper on Ramsey numbers.

In the discussion during the class, it was felt that this paper made significant contributions on a very broad level, but left the specifics open - for others to solve them later on. We think that this is a general trait for ground-breaking papers - that they make convincing argument about the points they are proving, but they don't solve them completely - this combination creates a burning interest in research community to pursue the path laid out by the paper and create a lasting body of future research along those lines.

3.1 Digital Representation and Quantification of Information

Building on the point made by Hartley's paper, Shannon's paper showed how to quantify information absolutely and in doing so, unified the notion of information across several different modalities of communication such as speech (telephone), symbols (telegraph, Morse code), text (facsimile), pictures (television) and so on. This unifying theme is one of the fundamental contributions made by Shannon's 1948 paper.

This paper pointed out that the content of the message was irrelevant to its transmission: it did not matter what the message represented. It could be text, sound, image, or video, but it was all 0's and 1's to the channel. In a follow-up paper, Shannon also pointed out that once data was represented digitally, it could be regenerated and transmitted without error. This was a radical idea to engineers who were used to thinking of transmitting information as an electromagnetic waveform over a wire. Before Shannon, communication engineers worked on their own distinct fields, each with its own distinct techniques: telegraphy, telephony, audio and data transmission all had nothing to do with each other. Shannon's vision unified all of communication engineering, establishing that text, telephone signals, images and film - all modes of communication - could be encoded in bits, a term that was first used in print in his article. This digital representation is the fundamental basis of all we have today.

3.2 Understanding of Information Content

Shannon's paper expressed the capacity of a channel: defining the amount of information that can be sent down a noisy channel in terms of transmit power and bandwidth. In doing so, Shannon showed that engineers could choose to send a given amount of information using high power and low bandwidth, or high bandwidth and low power.

The traditional solution was to use narrow-band radios, which would focus all their power into a small range of frequencies. This is along the lines of thinking that if you want to be heard, yell.

> ”I remember John Pierce at Bell Labs. He was Shannon's boss. He was playing down the importance of the noisy channel theorem, saying: 'just use more bandwidth, more power'. there was no limitation then - you could do whatever you needed in terms of reliable communication without long encoding. And besides, even if you wanted to, you were very badly limited by equipment complexity and cost..” [Interview with Fano, R. 2001]

The problem was that as the number of users increased, the number of channels began to be used up. Additionally, such radios were highly susceptible to interference: so much power was confined to a small portion of the spectrum that a single interfering signal in the frequency range could disrupt communication Shannon offered a solution to this problem by redefining the relationship between information, noise and power. Shannon quantified the amount of information in a signal, stating that is the amount of unexpected data the message contains. He called this information content of a message 'entropy'. In digital communication a stream of unexpected bits is just random noise. Shannon showed that the more a transmission resembles random noise, the more information it can hold, as long as it is modulated to an appropriate carrier: one needs a low entropy carrier to carry a high entropy message. Thus Shannon stated that an alternative to narrow-band radios was sending a message with low power, spread over a wide bandwidth.

3.3 Formal Architecture of Communication Systems

Almost every single modern textbook after Shannon's paper uses this block diagram as a way of representing the general communication system.

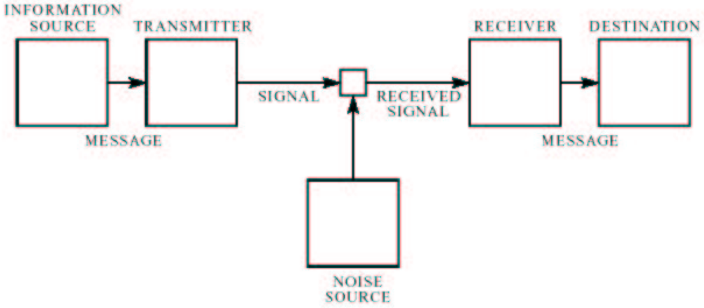


Figure 3: Schematic diagram of a general communication system.

This figure is one of the great contributions of the 1948 Shannon paper. It demonstrates that any communication system can be separated into components, which can be treated independently as distinct mathematical models. Thus, it is possible to completely separate the design of the source from the design of the channel. Shannon himself, realized that his model had ”applications not only in communication theory, but also in the theory of computing machines, the design of telephone exchanges and other fields.

Shannon's block diagram allowed researchers to look at different components of the communication system separately and analyze them in a systematic manner. All of today's communication systems are essentially based on this model - it is truly 'a blueprint for the digital age'.

3.4 Notion of Channel Capacity

One of the most un-intuitive and eminent results from Shannon's results was the idea that every communication channel has a limit, measured by the amount of binary digits that can be sent through the channel per second: the famous Shannon Limit. This limit inspired generations of engineers and researchers to push forward the current state of communication technology in an attempt to bridge the gap between existing communication capacity and the Shannon Limit. In a philosophical sense, this limit that Shannon's paper proposed served as the dangling carrot that drove significant advances in communication theory.

The outcome of Shannon Limit is that it is mathematically impossible to get error free communication above the limit. No matter how sophisticated an error correction scheme you use, no matter how much you can compress the data, you can not make the channel transmit faster than the limit without losing some information. On the other hand, it is possible to transmit information with zero error if the system operated below the Shannon Limit. Shannon mathematically proved that there were ways of encoding information that would allow one to get up to the limit without any errors: regardless of the amount of noise or static, or how faint the signal was. One might need to encode the information with more and more bits, so that most of them would get through and those lost could be regenerated from the others. The increased complexity and length of the message would make communication slower and slower, but essentially, below the limit, you could make the probability of error as low as you wanted.

> "To make the chance of error as small as you wish? Nobody had ever thought of that. How he got that insight, how he even came to believe such a thing, I don't know. But almost all modern communication engineering is based on that work." [Fano, R. Quoted in Technology Review, Jul 2001]

4 More Biographical Background

While Shannon largely worked alone, he took interest in other people's work, and others often consulted him to get a new perspective on their problems. He was a great asset because he was good at stripping away the complexity from other people's problems to find a simple and fundamental insight. Part of his approach was to find a "toy problem" which made the fundamentals of a problem easier to examine. He was novel in this respect; Gallager poses that "the use of simple toy models to study real situations appears not to have been common in engineering and science before Shannon's work" [1].

At Bell Labs and later at MIT, he was able to pursue wherever his interests led him, and Shannon followed his instincts as a researcher to make lasting contributions in many fields, including switching, computing, artificial intelligence and games. Shannon's work and play came together, and he seems to have often been motivated by the challenge of a puzzle. He had a love of juggling; he built the first known machine able to juggle which he named W. C. Fields (Figure 4). It was able to bounce-juggle, a simplification of the problem that made it possible to build a machine that could do the task without feedback. He also formulated a juggling theorem, which set forth the relation between the position of the balls and the action of the hands. Big Blue and other chess programs that are now playing against humans follow from Shannon's pioneering work in games [1], and he built a mechanical mouse named Theseus (see Figure 5) that was one of the earliest examples of machine learning. Theseus was able to learn the solution to a maze, and when the walls or the goal was changed, it was able to notice the change and learn the new solution. He worked on an

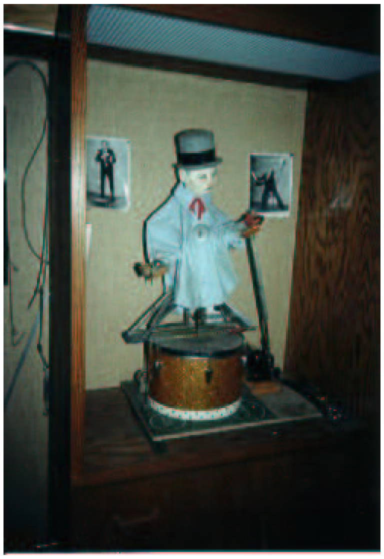


Figure 4: W. C. Fields, a machine Shannon built that was able to bounce-juggle.

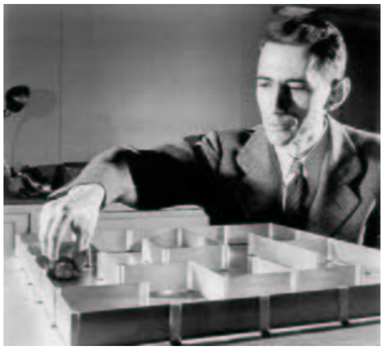


Figure 5: Shannon with Theseus, a mechanical mouse that was some of the earliest work in machine learning.

analog computer that could win at roulette by exploiting the irregularities in the table [4] and on modeling investments and the stock market, though he wasn't seeking to get rich off either.

It was at Bell Labs that Shannon met his wife, Mary Elizabeth (Betsy) Moore, where she worked as a numerical analyst. Gallager describes them as having "shared a good natured intellectual sense of humor and a no-nonsense but easy-going style of life" [1]. They had three children together, and he shared with his family his love of toys, building many for his children, including unicycles, looms, chess sets, erector sets, musical instruments, and the mechanical mouse Theseus (see pictures of the toy room in Figure 6). He was known to ride his unicycle through the halls of Bell Labs after hours.

Shannon remained in the math group at Bell Labs until 1956, with a constant stream of new and interesting results. In 1954, he published [10], showing that a Turing machine could be constructed using only two internal states, as well work with Edward Moore showing how reliable computing could be accomplished with unreliable components [5].

He spent 1956 visiting MIT and 1957 at the Center for the Study of Behavioral Sciences in Palo Alto. In 1958, he accepted a permanent appointment at MIT as a professor in Electrical Engineering and Mathematics, and

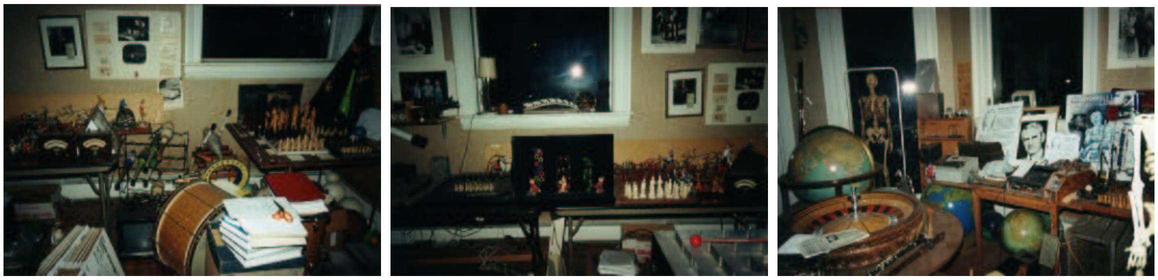
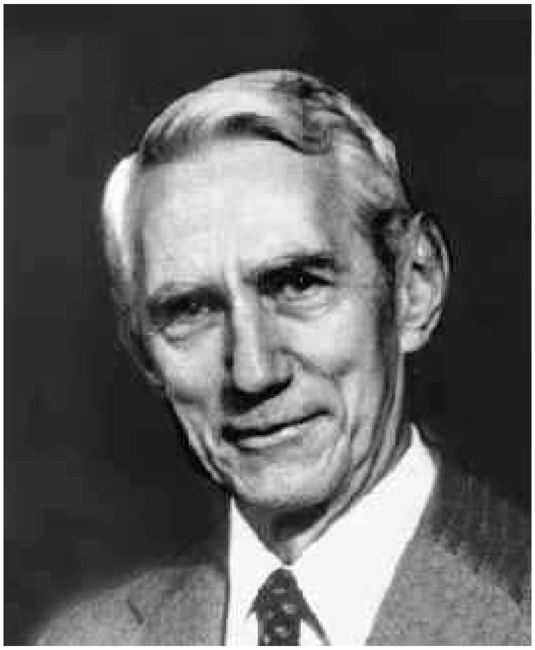


Figure 6: Pictures of Shannon's toys at home

continued to be prolific in the field of information theory; between 1956 and 1967, he wrote six important papers extending information theory. He also gave frequent seminars at MIT, often on research ideas he was working on at the time. Once he gave an entire semi-weekly seminar course with new research results at each lecture.

By the 1980's, Shannon began having problems with his memory and he was later diagnosed with Alzheimer's disease. In his final years he was "good-natured as usual" and enjoyed daily visits with his wife, Betsy. Eventually his body failed and he passed away in February 2001. Over the course of his life, he was remarkably prolific, producing 127 published and unpublished documents. Kolmogorov described Shannon's great talent [2]:

> In our age, when human knowledge is becoming more and more specialized, Claude Shannon is an exceptional example of a scientist who combines deep abstract mathematical thought with a broad and at the same time very concrete understanding of vital problems of technology. He can be considered equally well as one of the greatest mathematicians and as one of the greatest engineers of the last few decades.



References

[1] Robert G. Gallager. Claude E. Shannon: A Retrospective on His Life, Work, and Impact. *IEEE Transactions on Information Theory*, 47(7):2681–2695, November 2001.

[2] Robert G. Gallager. Claude Elwood Shannon. *Proc. of the American Philosophical Society*, 147(2), June 2003.

[3] R. V. L. Hartley. Transmission of information. *Bell Syst. Tech. J.*, 7:53511, 1928.

[4] Robert E. Kahn. A Tribute to Claude E. Shannon (1916-2001). *IEEE Communications Magazine*, July 2001.

[5] Edward F. Moore and Claude E. Shannon. Reliable circuits using crummy relays. Technical Report Memo. 54-114-42, Bell Labs, 1954.

[6] Robert Price. A Conversation with Claude Shannon: One man's approach to problem-solving. *IEEE Communications Magazine*, 22(5), May 1984.

[7] Claude E. Shannon. A symbolic analysis of relay and switching circuits. *Trans. AIEE*, 57:713–723, 1938.

[8] Claude E. Shannon. A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27:379–423, 623–656, July, October 1948.

[9] Claude E. Shannon. Communication theory of secrecy systems. *Bell Syst. Tech. J.*, 28:656–715, Oct 1949.

[10] Claude E. Shannon. A universal Turing machine with two states. Technical Report Memo. 54-114-38, Bell Labs, 1954.