# The Mathematics of the Information Revolution

Sanjoy K. Mitter
*Department of Electrical Engineering and Computer Science*
*MIT*
*Cambridge, MA 02139 USA*

The popular press keeps reminding us incessantly that we are living in an Information Age. What we are actually living in, is the "data age" in the sense that massive amounts of data on all aspects of modern life is available and accessible. The so-called Information revolution is yet to come. This distinction between data and information is an important one and needs to be made. We need to make this distinction if we truly wish to usher in the Information Age by creating an Information Revolution.

The reason why the distinction between "data" and "information" is not made is that people implicitly identify the linguistic notion of information with the technical notion of information as defined by Shannon in his Mathematical Theory of Communication. The technical notion of information is syntactic and devoid of any notion of meaning and, hence, context independent. In order to attach "meaning" to data, context must be brought into the picture and then any notion of universality will have to be given up. A theory of information which is context dependent and attaches meaning, which does not exist today, needs to be developed. We would then have the hope of attaching "interpretations" to data and thereby make the data relevant for the purpose of taking action to reach desirable goals.

In distant 1948, Claude Shannon, a mathematician, and William Shockley, a physicist, were both at Bell Telephone Laboratories, a research and development laboratory, then part of American Telephone and Telegraph Company. During World War II, Shannon joined an elite group at Bell Telephone Laboratories working on automatic control for anti-aircraft batteries. During this time he also continued to work on Switching Theory, an outgrowth of his work on using Boolean Algebra for the synthesis of Switching Circuits (part of his Master's thesis at MIT). He also started work on Communication Theory which he became interested in while working on the Differential Analyzer, an early analogue computer, being developed under Vannevar Bush. At the same time he became interested in Cryptography and recognized that the science of Cryptography was related to his ideas in Communication Theory. Shannon's work on Cryptography appeared in the open literature in 1949 as "Communication Theory of Secrecy Systems." In many ways, this paper is the foundational paper for what has become the major field of Cryptography. Shannon had been working on his ideas on Communication Theory for about eight years beginning about 1940. His landmark paper, "A Mathematical Theory of Communication," appeared in 1948 in two parts in Volume 27 of the Bell System

Technical Journal, a total of 76 pages. The subject now known as Information Theory, which provides the foundation of Digital Communication, was initiated and essentially completed in this paper. To understand the implication of this statement, suffice it to say that the multi-billion dollar Wireless Communication Industry would not exist without the scientific underpinnings provided by Shannon's seminal paper.

The Mathematical Theory of Communication led to the architecture for the design of digital communication systems which is shown below.

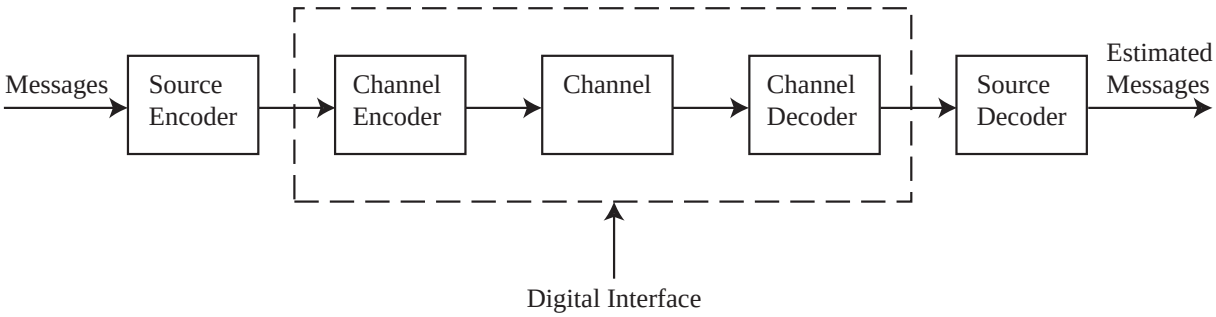


Figure 1:

The fact that Shannon obtained an architecture for design of a Communication System does not mean that is can be built as an Engineering System. What permitted this to happen in the first place, was the invention of the transistor by William Shockley, also in 1948. Shannon's seminal mathematical paper and Shockley's invention has been referred to by the Communication theorist, Andrew Viterbi, as the "double big band' which led to the creation of the modern Communications industry.

It is important to understand that this "abstraction," shown in Figure 1, requires precise mathematical definitions of each entity, namely: messages, source encoder, channel encoder, channel, channel decoder and source decoder. One of the fundamental contributions was to think of messages as choices between alternatives. The set of alternatives may be discrete (that is, finite or countable) or infinite (uncountable). In some sense, this notion of choice between alternatives was present in the earlier work of Hartley, who showed that, in many situations, the number of possible alternatives from a message source over an interval of duration $T$ grows exponentially with $T$, leading to the definition of information as the logarithm of this growth. Shannon extended Hartley's idea by putting a probability measure on the set of alternatives. He pointed out that it is the choice, namely, the probability measure on the set of alternatives that is important, and not the *representation* of the choice as integers, binary code or letters from an alphabet. The representation of interest to the user may be mapped to any convenient representation that is suitable for transmission. This mapping is known ahead of time to both transmitter and receiver and hence, an arbitrarily long sequence can be communicated.

In the sequel, we assume that the reader is familiar with elementary probability theory. A message or source then is a stochastic process $(X_t)_{t\geq 0}$, $t \in Z$ or $\mathbb{R}$, that is a sequence or path consisting of random variables (measurable maps)

$$X_t \,:\, \Omega \to \mathcal{X}$$

where $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathcal{X}$ is a set (a finite set, $\mathbb{R}$, $\mathbb{R}^n$, $C(0, \infty\,; \mathbb{R})$, etc.). In the sequel, we shall take $\mathcal{X}$ to be a finite set and the time set to be $Z_+$. A somewhat general model of a source is an Ergodic Markov Chain taking values in $\mathcal{X}$. Recall that a Markov Chain is defined by

$$P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = P(X_t = x_t | X_{t-1}) \; , \forall t \in Z_+$$

i.e. to specify a Markov Chain one has to remember only the immediate past and not the whole past. An Ergodic Markov Chain is one which has a unique invariant probability measure $\pi$ and even if we start the chain at an arbitrary initial distribution it converges to the unique measure $\pi$.

**Entropy as a Measure of Uncertainty and Information**

Let $X$ be a random variable taking values in $\mathcal{X} = \{1, \ldots, N\}$ and let $p_i = \text{prob}(X = i), i = 1, \ldots, N$. Shannon defined the *entropy* of $X$ as

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i \; .$$

Note that entropy is only a function of the probabilities. The extension of the above definition to the rate of entropy production in a Stationary Markov Chain is straightforward. Essentially, the same formula appears in the work of Boltzmann in Statistical Mechanics, so it is worthwhile understanding how Shannon's work differs from that of Boltzmann. Let $\mathcal{X}$ be a finite set and let $\mu$ be a probability measure on $\mathcal{X}$. In the Boltzmann interpretation, $\mathcal{X}$ represents the set of all possible energy levels of a system of particles and $\mu$ corresponds to a histogram of energies for $n$-trials representing the macrostate of the system. At the microlevel the system is described by a sequence $\omega \in \mathcal{X}^n$. Boltzmann's idea can be stated as a principle:

> The entropy of a macrostate corresponds to the degree of uncertainty about the microstate $\omega$, when $\omega$ is known and can be measured by $\log N_n(\mu) =$ logarithmic number of microstates leading to $\mu$. Now
>
> $$N_n(\mu) = \frac{n!}{\Pi_{x \in \mathcal{X}} (n\mu(x))!} \; .$$
>
> Let $\mu_n$ be a sequence of $n$-particle macrostates such that $\mu_n \to \mu$ and $n\mu_n(x) \in Z$. Then it can be proved that
>
> $$\lim_{n \to \infty} \frac{1}{n} \log N_n(\mu_n)$$
>
> exists and is equal to
>
> $$H(\mu) = -\sum_{x \in \mathcal{X}} \mu(x) \log \mu(x) \; .$$

In the Boltzmann picture $\mu$ corresponds to a histogram resulting from a random phenomenon and $H(\mu)$ corresponds to the observer's uncertainty about the behavior of the system at the microlevel.

Shannon however starts from the measure $\mu$ and generates a random variable taking values in the set $\mathcal{X}$ and has law $\mu$. His approach is to codify the amount of effort, measured in terms of the number of yes/no questions on the average, needed to recover $\mu$, after many independent samplings from $\mu$. But this also measures the degree of information a receiver obtains a posteriori after all yes/no questions have been answered. Hence one obtains:

> The information contained in a random signal with prescribed distribution, equals the number of bits necessary to encode the signal.

We now assume that we are dealing with a memoryless source. Then, the above can be formalized as follows:

> Let $f : \mathcal{X} \to \bigcup_{K\geq 1}\{0,1\}^K$ be an injective map, called the encoding map, and the image the code, and let $\mathcal{N}(f(x))$ be the length of the codeword $f(x)$ and $E_\mu(\mathcal{N}(f(x)))$ be the expected length of the codeword.
>
> Let
> $$I(\mu) = \inf\Big[E_\mu\Big(\mathcal{N}(f(x))\Big)\Big| f \text{ an encoding map}\Big]$$
> Let $\mu^n$ denote the product measure. Then we obtain

**Shannon's Source Coding Theorem**

$$\lim_{n\to\infty}\frac{I(\mu^n)}{n} = -\sum_{x\in\mathcal{X}}\mu(x)\log_2\mu(x) := \frac{1}{\log 2}H(\mu)\ .$$

The proof of this theorem depends on the Asymptotic Equipartition property:
$\forall\epsilon > 0$,

$$\mu^n\Big(\omega\in\mathcal{X}^n\Big|\Big|\frac{1}{n}\log\mu^n(\omega) + H(\omega)\Big| \geq \epsilon\Big) \to 1$$

as $n\to\infty$, which follows from the weak law of Large Numbers.

The interpretation of the above result is that most $\omega$ have probability

$$\exp(-nH(\mu))$$

which is to be compared with Boltzmann's formula

$$S = k\log W\ ,$$

which says that the entropy $S$ of an observed macrostate is equal to the logarithmic probability of its occurrence up to some constant $k$.

**Channels, Mutual Information and the Noisy Channel Coding Theorem**

So far we have discussed the modeling of an information source which generates messages and their encoding. This encoding process may be thought of as compressing the data which amounts to reducing its complexity. This encoded message now needs to be transmitted over a noisy channel (a telephone line, fiber optic cable or via electromagnetic waves) reliably, to the destination where there is a receiver (decoder) which reconstructs the message. The ideas of optimal reconstruction is made precise by saying that the probability of error, namely, the probability of the difference between the message and its reconstruction, should be small. The fundamental question that Shannon's Noisy Channel Coding Theorem addresses is to determine the maximal rate (in bits/sec.) that the data can be transmitted over the noisy channel so that the probability of error can be made arbitrarily small, provided we are willing to send arbitrarily long coded sequences.

Since we have a noisy channel, it is intuitively clear that before sending the encoded source over the channel, one should re-encode it by adding redundancy to mitigate against the effects of noise in the channel so that error correction will be possible at the decoder.

In order to make these ideas concrete, we need a channel and to present a characterization of a channel which permits us to answer the question of determining the maximum rate of transmission so that the probability of error is small.

We will deal with discrete memoryless channels given by a matrix

$$P(Y_k = j|i) \ := p_{ij} \ , \ j = 1, \ldots M \, ; \, i = 1, \ldots N$$

where $Y_k$ is the output of the channel at time $k$, given that the input is $i$. When we connect the source $X_k$ to the channel, we obtain the conditional probability

$$P(Y_k = k|X_k = i) \ := p_{ij}$$

Given an input sequence $\vec{x} = (x_1, \ldots, x_n)$ and output sequence $\vec{y} = (y_1, \ldots, y_n)$

$$P(X = x|Y = y) = \prod_{k=1}^{n} p_{x_k y_k} \tag{1}$$

Shannon defined the transmission rate as the mutual information $I(X \, ; \, Y) = H(X) - H(X|Y)$, where

$$H(X|Y) = \sum_{i,j} p_i p_{ij} \log_2 p_{ij} \tag{2}$$

The interpretation of this rate is that it represents the a priori uncertainty about $X$ less the conditional uncertainty $H(X|Y)$ after the output is observed. We can rewrite (1) as

$$\begin{aligned} I(X \, ; \, Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_k H(Y_k|X_k) \quad \text{from (1)} \end{aligned}$$

Shannon's definition of Channel Capacity in bits/sec. is

$$\sup_{P_x,n} \frac{1}{n}\Big[H(X) - H(X|Y)\Big] .$$

For discrete memoryless channels, the capacity $C$ is achieved by i.i.d. distributions and is given by

$$C = \max_{(p_1,\ldots,p_n)} \left[ \sum_{i,j} p_i p_{ij} \log_2 \frac{p_{ij}}{p \sum_i p_i p_{\ell j}} \right] \tag{3}$$

*The Noisy Channel Coding Theorem* states that any source input to the channel whose entropy per second $H_t$ is less than $C_t$, it is possible to encode the input source to the channel and decode the signal at the output to the channel in such a way that the error rate (in source symbols per second) can be made arbitrarily small, provided we allow code sequences of arbitrarily long block length.

Conversely, if $H_t > C_t$, then reliable transmission in the above sense is not possible.

To give an indication of the proof of the Noisy Channel Coding Theorem, let $p_i^*$, $i = 1, 2, \ldots n$ denote the optimal input distribution obtained from (3) and consider the corresponding i.i.d. input-output pairs $(X_i, Y_i)_{i=1}^n$ and consider the support of this joint distribution as given by the Asymptotic Equipartition Theorem. Then choose $2^{nR}$ codewords randomly for $R < C$. Then define a decoding strategy for each such randomly chosen code. Finally, evaluate an upper $k$ bound on error probability averaged over the ensemble of randomly chosen codes and then show that this bound approaches 0 as $n \to \infty$. There is at least one code for which the above is true.

It can be shown that Maximum Likelihood Decoding provides such a decoding rule.

The Noisy Channel Coding Theorem and its generalization, the Rate Distortion Theorem where reliable transmission is required subject to a constraint between the expected distortion between the input and output sequences (for example, squared error measure), forms the basis for the architecture of Figure 1, where the coding of the source and channel can be separated without losing asymptotic optimality. This means that the source can be encoded into a binary stream and subject to the distortion constraint and then, using Channel Coding, the binary stream can be transmitted over the channel without incurring any penalty in terms of the probability of error. It is worth stating that "Digital Communication" does not mean that the channel is digital nor that the source is discrete, but that the input to the channel encoder is a binary stream, that is, the interface between the source and channel is a binary interface. It is the universality of this binary interface that makes communication over the internet possible.

**Information Transmission and Statistical Mechanics**

There is a surprising connection between the coding-decoding problem and the study of disordered systems. In physics, the study of disordered systems refers to a statistical mechanics model in a random environment. The simplest such models are Ising models whose non-zero

pairwise interactions are i.i.d. random variables. Questions about disordered Ising models are often closely connected to issues concerning related percolation models.

Let $\Lambda$ be a finite subset of $Z^d$. An Ising model on $\Lambda$ is a family $\{S_x | x \in \Lambda\}$ of random variables, called spins taking values in $\{1, -1\}$, whose joint distribution $P_{\Lambda,\beta}$ depends on the parameter $\beta \geq 0$ (inverse temperature) and has the form

$$P_{\Lambda,\beta}(\{s\}) = Z^{-1}_{\Lambda,\beta} \exp(-\beta H_\Lambda(s)) \, . \tag{4}$$

Here $s \in S_\Lambda = \{-1, +1\}^\Lambda$, $H_\Lambda(s)$ is a real-valued function on $S_\Lambda$ and $Z_{\Lambda,\beta}$ is a normalizing constant called the Partition function. An example of a Hamiltonian is

$$H_\Lambda(s) = - \sum_{\substack{x,\, y \,\in\, \Lambda \\ \langle x,\, y \rangle}} -J^\Lambda_{\langle x,y \rangle} s_x s_y \tag{5}$$

where the notation $\langle x, y \rangle$ indicates that the summation is carried out over nearest neighbors. Boundary conditions will usually be imposed on the Hamiltonian. A disordered system is one where the coupling $J^\Lambda_{\langle x,y \rangle}$ is itself random.

The limit as $\beta \to \infty$ (corresponding to zero temperature) of $P_{\Lambda,\beta}$ is the uniform measure of the ground state of $H_\Lambda$. In statistical mechanics, the questions of interest are infinite volume limit $\Lambda \to Z^d$ of ground states and Gibbs distributions.

If we take the random coding argument view of Shannon, as in the proof of the Noisy Channel Coding Theorem, and choose a decoding rule which is Maximum a Posteriori Probability Estimate (equivalent to a ground state calculation), then the coding-decoding problem can be mapped to the study of a disordered system of the form (4). The reader is referred to N. Sourlas, *Nature* **339** 693-694 (1989), where this mapping was first proposed.

There is now considerable interest in quantum computation, quantum communication and quantum control. This has necessitated a reexamination of the problem of quantum measurement since measurements on a quantum system have to be made for the purpose of inferring the state of the quantum system. In view of our previous discussion about the convergence of ideas of Statistical Mechanics and Communication Theory, and interest in Quantum Information Systems, one may speculate that the ideas of Statistical Mechanics and Quantum Physics are intimately related to problems of Statistical Inference and Information Theory. For a discussion of one aspect of these issues, see the author's paper "On the Analogy BetweenMathematical Problems of Non-Linear Filtering and Quantum Mechanics," *Ricerche di Automatica*. Special Issue on System Theory and Physics, 10(2):163-216, December 1979.

**The Structure of Engineering Revolutions**

The fundamental problem of Physics is understanding nature by discovering its laws. These laws, in some sense, universal, have also predictive power. The motion of the Solar system can be predicted from Newton's Laws of Motion. In his book, "The Structure of Scientific Revolutions," Thomas Kuhn makes a distinction between normal Science and a Scientific revolution and argues that Scientific revolutions come about when existing paradigms are unable

to explain certain scientific phenomenon on the basis of existing physical laws. Quantum Mechanics, in the hands of Planck, Heisenberg and Schrödinger, constitutes one such scientific revolution. The Science of Engineering on the other hand is concerned with codifying the fundamental limitations of systems which may exist or which are yet to be synthesized. Shannon's theory of Communication provides the fundamental limitations of communication systems in the sense that it codifies under what conditions a message can be transmitted reliably over a noisy communication channel and when it is possible not to do that. To state the Noisy Channel Coding Theorem, the concepts of source, channel, coding and decoding had to be invented and mathematically defined and only with this new language could the theorem be stated. Shannon's theory constitutes a revolution, as fundamental as that of Quantum Mechanics, but in a sense different from that of Kuhn's scientific revolution. The interplay between an engineering problem, its modeling in mathematics and the subsequent deduction about the fundamental limitations of the engineering system exemplifies how mathematics can play a fundamental role in the synthesis of engineering systems.

**Networks and Communications**

In this lecture, I was unable, due to lack of time, to discuss the science and technology of Communications over Networks. The structure of communications that is emerging today is for the Internet to serve as a universal communications medium with wireless networks and even telephone networks being superimposed on it. Communication today means transmission of voice, data and images over heterogeneous networks. A new Information Theory for Network Communications which uses ideas from Graph Theory, Dynamical Systems, Statistical Physics and Information Theory needs to be developed to provide the scientific foundations of this technological development.

I believe a new revolution as profound as the work of Shannon is needed to make this a reality.

**Final Remarks**

The ability to transmit massive amounts of data (voice, data, images) over long distances almost at the speed of light and the ability to store and process this data using massive computational power has given rise to a global networked economy. These developments in their turn are leading to a "networked society" on a global scale with enormous possibilities for social and economic change, both positive and negative. Social scientists like Manuel Castells (see, for example, his book *The Informational City: Information technology, economic restructuring and the urban-regional process*, Blackwell Publ., Cambridge, MA, 1989.) are trying to come to grips with these issues. It must be realized that these social and economic issues cannot be understood without substantial and deep ideas (ultimately mathematical) underlying the Sciences of Information. It is the responsibility of mathematicians and theoretical engineers to engage their counterparts in Economics and Social Science in a dialogue which would lead to a deeper understanding of the forces which have brought about the Data Revolution. More

importantly, this deeper understanding should help in ushering in a true Information Revolution for the benefit of mankind.

**Notes and References**


(1) For a discussion on Entropy and its role in Physics, Mathematics and Information Theory, see *Entropy*: edited by A. Greveu, G. Keller and G. Warnecke, Princeton University Press, Princeton, NJ, 2003.

(2) For a perspective on Shannon's contributions and his style of doing research see

R.G. Gallager, "Claude E. Shannon: A Retrospective on his life, work and impact," *IEEE Trans. on Info. Theory*, Vol. 47, No. 7, November 2001, pp. 2681-2695.

(3) For Shannon's work, see his Collected Works published by the IEEE Press, 1993.