science of information summer school
stanford university

may 30, 2012

(re)reading shannon

sergio verdú
A mathematical theory of communication

by

C. E. Shannon
MIT (1936-1940)
Letter to Vannevar Bush, February 16, 1939

Claude E. Shannon

Dear Dr. Bush,

Off and on I have been working on an analysis of some of the fundamental properties of general systems for the transmission of intelligence, including telephony, radio, television, telegraphy, etc. Practically all systems of communication may be thrown into the following form:

\[ f_1(t) \rightarrow T \rightarrow F(t) \rightarrow R \rightarrow f_2(t) \]
Institute for Advanced Study, Princeton (1940-1941)
Bell Laboratories, New York, 1941-1949
Bell Laboratories, Murray Hill, New Jersey
A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist\(^1\) and Hartley\(^2\) on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are
The Mathematical Theory of Communication

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist\(^1\) and Hartley\(^2\) on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.
The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.
Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The
If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.
The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly bits, a word suggested by J. W. Tukey. A device with two
Fig. 1—Schematic diagram of a general communication system.
suitably idealized from their physical counterparts. We may roughly classify communication systems into three main categories: **discrete, continuous and mixed**. By a discrete system we will mean one in which both the message and the signal are a sequence of discrete symbols. A typical case is telegraphy where the message is a sequence of letters and the signal a sequence of dots, dashes and spaces. A continuous system is one in which the message and signal are both treated as continuous functions, e.g. radio or television. A mixed system is one in which both discrete and continuous variables appear, e.g., PCM transmission of speech.
“mixed”
PART I: DISCRETE NOISELESS SYSTEMS

1. The Discrete Noiseless Channel
mitted from one point to another. Each of the symbols $S_i$ is assumed to have a certain duration in time $t_i$ seconds (not necessarily the same for different $S_i$, for example the dots and dashes in telegraphy). It is not required that all possible sequences of the $S_i$ be capable of transmission on the system; certain sequences only may be allowed. These will be pos-
following definition: The capacity $C$ of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration $T$. 
\[ C = \log X_0 \]

\( X_0 \) is the largest real solution of

\[ X^{-t_1} + X^{-t_2} + \cdots + X^{-t_n} = 1 \]
A very general type of restriction which may be placed on allowed sequences is the following: We imagine a number of possible states $a_1, a_2, \ldots, a_m$. For each state only certain symbols from the set $S_1, \ldots, S_n$ can be transmitted (different subsets for the different states). When one of these has been transmitted the state changes to a new state depending both
Fig. 2—Graphical representation of the constraints on telegraph symbols.
Theorem 1: Let \( b_{ij}^{(s)} \) be the duration of the \( s^{th} \) symbol which is allowable in state \( i \) and leads to state \( j \). Then the channel capacity \( C \) is equal to \( \log W \) where \( W \) is the largest real root of the determinantal equation:

\[
\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0
\]

where \( \delta_{ij} = 1 \) if \( i = j \) and is zero otherwise.
storage
2. **The Discrete Source of Information**

We now consider the information source. **How is an information source to be described mathematically, and how much information in bits per second is produced in a given source?** The main point at issue is the effect of statistical
In the general \( n \)-gram case a set of \( n \)-gram probabilities \( p(i_1, i_2, \ldots, i_n) \) or of transition probabilities \( p_{i_1, i_2, \ldots, i_{n-1}}(i_n) \) is required to specify the statistical structure.
1. Zero-order approximation (symbols independent and equiprobable).

    XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.
2. First-order approximation (symbols independent but with frequencies of English text).

    OCRO HLI RGWR NMIELWIS EU LL NBNE-SEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.
3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.
4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOAC TIONA OF CRE.
5. First-order word approximation. Rather than continue with tetragram, ..., \(n\)-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.
6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHOEVER TOLD THE PROBLEM FOR AN UNEXPECTED.
HR-35, High Speed!
High Performance!!

FEATURES
- Buffer ... 7K bytes
- Color printing
  - Red and black color printing
- Super/sub script
- Auto underlining
- Bold printing
- Carriage skip movement
- Proportional spacing
- Text reprinting
- Clear buffer
- Line spacing 6, 4, 3 positions/inch
- Copy capacity
  - Original (45 kg) + 4 copies (15 kg)
  - 96 characters/wheel, cassette type
- Print wheel
  - 10 million characters min./wheel
- Reliability
  - MTBF 3,000 hrs (25% duty)
- Dimensions
  - 538(W) x 195(H) x 385(D) mm (21.2 x 7.7 x 15.2"
- Weight
  - 14.7 kg (32.4 lbs)

OPTION
- TF-100 tractor feeder
  - Paper width Max. 15 inches (381 mm)
  - Paper feed Bi-directional
- CF-100 cut sheet feeder
  - Paper width Max. 8½ inches (216 mm)
  - Paper capacity 150 sheets

SPECIFICATIONS
- Interface CENTRONICS parallel, RS-232C serial
- Print speed
  - Max. 36 CPS, 32 CPS (Shannon Text)
- Paper width
  - 16.5 inches (419 mm)
  - 13.2 inches (335 mm)
- Carriage motion
  - Bi-directional: 10, 12, 15, PS
- Paper feed
  - 4-inches/second
- Paper feed speed
  - 4-inches/second
- Line spacing
  - Minimum increment 1/120 inch
- Copy capacity
  - 96 characters/wheel, cassette type
- Print wheel
  - 10 million characters min./wheel
- Print wheel life
- Ribbon
  - Cassette type:
    - Single strike ribbon
    - Multi strike ribbon
    - Fabric ribbon
- Noise
  - Less than 65 dB, A scale
- Temperature
  - 10°C to 40°C (50°F to 104°F) (operating)
- Humidity
  - 20% to 80% (operating, non-condensation)
- Power
  - 115, 220, or 240V, 50/60 Hz, 65W
- Reliability
  - MTBF 3,000 hrs (25% duty)
- Dimensions
  - 538(W) x 195(H) x 385(D) mm (21.2 x 7.7 x 15.2"
- Weight
  - 14.7 kg (32.4 lbs)

*Specifications subject to change without notice.
<table>
<thead>
<tr>
<th>Specification</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>CENTRONICS parallel, RS-232C serial</td>
</tr>
<tr>
<td>Print speed</td>
<td>Max. 36 CPS, 32 CPS (Shannon Text)</td>
</tr>
<tr>
<td>Paper width</td>
<td>16.5 inches (419 mm)</td>
</tr>
<tr>
<td>Print width</td>
<td>13.2 inches (335 mm)</td>
</tr>
<tr>
<td>Carriage motion</td>
<td>Bi-directional: 10, 12, 15, PS</td>
</tr>
<tr>
<td></td>
<td>minimum increment 1/120 inch</td>
</tr>
</tbody>
</table>
4. **Graphical Representation of a Markoff Process**

Stochastic processes of the type described above are known mathematically as discrete Markoff processes and have been extensively studied in the literature.⁶ The general case can

THREE MODELS FOR THE DESCRIPTION OF LANGUAGE*

Noam Chomsky
Department of Modern Languages and Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract

We investigate several conceptions of linguistic structure to determine whether or not they can provide simple and "revealing" grammars that generate all of the sentences of English and only these. We find that no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar. Furthermore, the particular subclass of such processes that produce n-order statistical approximations to
5. **Ergodic and Mixed Sources**

Physically the situation represented is this: There are several different sources $L_1, L_2, L_3, \ldots$ which are each of homogeneous statistical structure (i.e., they are ergodic). We do not know *a priori* which is to be used, but once the sequence starts in a given pure component $L_i$, it continues indefinitely according to the statistical structure of that component.
6. Choice, Uncertainty and Entropy

We have represented a discrete information source as a Markov process. Can we define a quantity which will measure, in some sense, how much information is “produced” by such a process, or better, at what rate information is produced?
If there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is reasonable to require of it the following properties:

1. $H$ should be continuous in the $p_i$.

2. If all the $p_i$ are equal, $p_i = \frac{1}{n}$, then $H$ should be a monotonic increasing function of $n$. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original $H$ should be the weighted sum of the individual values of $H$. The meaning of this is illustrated
Theorem 2: The only $H$ satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$
This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.
Quantities of the form $H = -\sum p_i \log p_i$ (the constant $K$ merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty.
\( H \) will be recognized as that of entropy as defined in certain formulations of statistical mechanics\(^8\) where \( p_i \) is the probability of a system being in cell \( i \) of its phase space. \( H \) is then, for example, the \( H \) in Boltzmann’s famous \( H \) theorem.

Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht.

On the relationship between the second main theorem of mechanical heat theory and the probability calculation with respect to the results about the heat equilibrium.

Von dem c. M. Ludwig Boltzmann in Graz
Es ist also das Produkt \( F_1 F_2 F_3 \ldots \), welches Herr Meyer auf Seite 262 entwickelt, so zu verstehen, dass darin jeder Factor mehrmals auftritt, und könnte auch so geschrieben werde

\[
F_1^k F_1 F_2^k F_2 \ldots
\]

und dessen Logarithmus wäre:

\[
k(F_1 \log F_1 + F_2 \log F_2 + \ldots)
\]

\[
\Omega = - \iiint \iiint \iiint \iiint \iiint f(x, y, z, u, v, w) \ell f(x, y, z, u, v, w) \, dx \, dy \, dz \, du \, dv \, dw
\]
1. \( H = 0 \) if and only if all the \( p_i \) but one are zero, this one having the value unity. Thus only when we are certain of the outcome does \( H \) vanish. Otherwise \( H \) is positive.
2. For a given \( n \), \( H \) is a maximum and equal to \( \log n \) when all the \( p_i \) are equal, i.e., \( \frac{1}{n} \). This is also intuitively the most uncertain situation.
It is easily shown that

$$H(x, y) \leq H(x) + H(y)$$

with equality only if the events are independent (i.e., $p(i, j) = \ldots$)
4. Any change toward equalization of the probabilities \( p_1, p_2, \ldots, p_n \) increases \( H \). Thus if \( p_1 < p_2 \) and we increase \( p_1 \), decreasing \( p_2 \) an equal amount so that \( p_1 \) and \( p_2 \) are more nearly equal, then \( H \) increases. More generally, if we perform any “averaging” operation on the \( p_i \) of the form

\[
p_i' = \sum_j a_{ij} p_j
\]

where \( \sum_i a_{ij} = \sum_j a_{ij} = 1 \), and all \( a_{ij} \geq 0 \), then \( H \) increases (except in the special case where this transformation amounts to no more than a permutation of the \( p_j \) with \( H \) of course remaining the same).
We define the *conditional entropy* of \( y \), \( H_x(y) \) as the average of the entropy of \( y \) for each value of \( x \), weighted according to the probability of getting that particular \( x \). That is

\[
H_x(y) = - \sum_{i,j} p(i,j) \log p_i(j).
\]

\[
H(x,y) = H(x) + H_x(y).
\]

\[
H(y) \geq H_x(y).
\]
dancy. The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters, is roughly 50%.

Two extremes of redundancy in English prose are represented by Basic English and by James Joyce’s book *Finnegans Wake*. The Basic English vocabulary is limited to 850 words and the redundancy is very high. This is reflected in the expansion that occurs when a passage is translated into Basic English. Joyce on the other hand enlarges the vocabulary and is alleged to achieve a compression of semantic content.
FINNEGANS WAKE

by

James Joyce
riverrun, past Eve and Adam's, from swerve of shore to bend of bay, brings us by a commodius vicus of recirculation back to Howth Castle and Environs.

Sir Tristram, violer d'amores, fr'over the short sea, had passen-core rearriived from North Armorica on this side the scraggy isthmus of Europe Minor to wielderfight his penisolate war: nor had topsawyer's rocks by the stream Oconee exaggerated themselse to Laurens County's gorgios while they went doublin their mumper all the time: nor avoice from afire bellowsed mishe mishe to tauftauf thuartpeatrick: not yet, though venissoon after, had a kidscad buttended a bland old isaac: not yet, though all's fair in vanessy, were sosie sestthers wroth with twone nathandjoe. Rot a peck of pa's malt had Jhem or Shen brewed by arclight and rory end to the regginbrow was to be seen ringsome on the aquaface.
The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its *relative entropy*. This, as will appear later, is
The redundancy of a language is related to the existence of **crossword puzzles**. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two-dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just possible when the redundancy is 50%. If the redundancy is 33%, three-dimensional crossword puzzles should be possible, etc.
7. The Entropy of an Information Source

\[ H = \sum_{i} P_{i} H_{i} \]

\[ = - \sum_{i,j} P_{i} p_{i}(j) \log p_{i}(j). \]

\[ P_{j} = \sum_{i} P_{i} p_{i}(j). \]
$H$. $F_N$ of course is the conditional entropy of the next symbol when the $(N - 1)$ preceding ones are known, while $G_N$ is the entropy per symbol of blocks of $N$ symbols.

\[
F_N = NG_N - (N - 1)G_{N-1},
\]

\[
G_N = \frac{1}{N} \sum_{1}^{N} F_N,
\]

\[
F_N \leq G_N,
\]

\[
\lim_{N \to \infty} F_N = H.
\]
If successive symbols are independent then $H$ is simply $-\sum p_i \log p_i$ where $p_i$ is the probability of symbol $i$. Suppose in this case we consider a long message of $N$ symbols. It will contain with high probability about $p_1N$ occurrences of the first symbol, $p_2N$ occurrences of the second, etc. Hence the probability of this particular message will be roughly

$$p = p_1^{p_1N} p_2^{p_2N} \cdots p_n^{p_nN}$$

$$H = \frac{\log 1/p}{N}.$$ 

$H$ is thus approximately the logarithm of the reciprocal probability of a typical long sequence divided by the number of symbols in the sequence. The same result holds for any source.
Theorem 3: Given any $\epsilon > 0$ and $\delta > 0$, we can find an $N_0$ such that the sequences of any length $N \geq N_0$ fall into two classes:

1. A set whose total probability is less than $\epsilon$.

2. The remainder, all of whose members have probabilities satisfying the inequality

$$\left| \log \frac{p^{-1}}{N} - H \right| < \delta.$$
We assume the source to be ergodic so that the strong law of large numbers can be applied. Thus the number of times a given path $p_{ij}$ in the network is traversed in a long sequence of length $N$ is about proportional to the probability of being at $i$, say $P_i$, and then choosing this path, $P_i p_{ij} N$. If $N$ is large enough the probability of percentage error $\pm \delta$ in this is less than $\epsilon$ so that for all but a set of small probability the actual numbers lie within the limits

$$ (P_i p_{ij} \pm \delta) N. $$

Hence nearly all sequences have a probability $p$ given by

$$ p = \prod p_{ij} (P_i p_{ij} \pm \delta) N. $$
Consider again the sequences of length $N$ and let them be arranged in order of decreasing probability. We define $n(q)$ to be the number we must take from this set starting with the most probable one in order to accumulate a total probability $q$ for those taken.

*Theorem 4:*

$$\lim_{N \to \infty} \frac{\log n(q)}{N} = H$$

when $q$ does not equal 0 or 1.
Theorem 4 follows immediately from this on calculating upper and lower bounds for $n(q)$ based on the possible range of values of $p$ in Theorem 3.
The author considers an information source which can produce any one of a finite number of symbols. These are fed into a channel for transmission, each symbol having a positive time duration in the channel. The problem becomes statistical with the assumption that if $x_n$ is the $n$th symbol produced by the source the $x_n$ are a stationary stochastic process (Further, brevity of Mathematical terms)

a certain degree of fidelity is defined and discussed. The discussion is suggestive throughout, rather than mathematical, and it is not always clear that the author's mathematical intentions are honorable. The point of view is that stressed by Wiener in his NDRC report [soon to be published as a book] \`\`The Interpolation, Extrapolation, and Smoothing of Stationary Time Series," in which communication is considered as a statistical problem, specifically in its mathematical formulation as the study of stationary stochastic processes, and of the results of various operations performed on them.

Reviewed by J. L. Doob
In the mixed (not ergodic) case if

\[ L = \sum p_i L_i \]

and the entropies of the components are \( H_1 \geq H_2 \geq \cdots \geq H_n \) we have the

**Theorem:** \( \lim_{N \to \infty} \frac{n(q)}{N} = \varphi(q) \) is a decreasing step function,

\[ \varphi(q) = H_s \quad \text{in the interval} \quad \sum_{1}^{s-1} \alpha_i < q < \sum_{1}^{s} \alpha_i. \]
9. **The Fundamental Theorem for a Noiseless Channel**

Theorem 9: Let a source have entropy $H$ (bits per symbol) and a channel have a capacity $C$ (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{C}{H} - \epsilon$ symbols per second over the channel where $\epsilon$ is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

The converse part of the theorem, that $\frac{C}{H}$ cannot be ex-
messages of length $N$ in order of decreasing probability and suppose their probabilities are $p_1 \geq p_2 \geq p_3 \cdots \geq p_n$. Let $P_s = \sum_{i=1}^{s-1} p_i$; that is $P_s$ is the cumulative probability up to, but not including, $p_s$. We first encode into a binary system. The binary code for message $s$ is obtained by expanding $P_s$ as a binary number. The expansion is carried out to $m_s$ places, where $m_s$ is the integer satisfying:

$$\log_2 \frac{1}{p_s} \leq m_s < 1 + \log_2 \frac{1}{p_s}.$$  

$$G_N \leq H_1 < G_N + \frac{1}{N}.$$
12. **Equivocation and Channel Capacity**

Following this idea the rate of actual transmission, $R$, would be obtained by subtracting from the rate of production (i.e., the entropy of the source) the average rate of conditional entropy.

$$R = H(x) - H_y(x)$$

The conditional entropy $H_y(x)$ will, for convenience, be called the **equivocation**.
Fig. 8—Schematic diagram of a correction system.
will be greater than zero and \( H_{yz}(x) > 0 \). But this is the uncertainty of what was sent, knowing both the received signal and the correction signal. If this is greater than zero the frequency of errors cannot be arbitrarily small.
The capacity $C$ of a noisy channel should be the maximum possible rate of transmission, i.e., the rate when the source is properly matched to the channel. **We therefore define the channel capacity by**

$$C = \text{Max}(H(x) - H_y(x))$$

where the maximum is with respect to all possible information sources used as input to the channel. If the channel is
It is possible to send information at the rate $C$ through the channel with as small a frequency of errors or equivocation as desired by proper encoding. This statement is not true for any rate greater than $C$. 
Theorem 11: Let a discrete channel have the capacity \( C \) and a discrete source the entropy per second \( H \). If \( H \leq C \), there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If \( H > C \), it is possible to encode the source so that the equivocation is less than \( H - C + \epsilon \) where \( \epsilon \) is arbitrarily small. There is no method of encoding which gives an equivocation less than \( H - C \).
$N(T, q)$ to be the maximum number of signals we can choose for the subset such that the probability of an incorrect interpretation is less than or equal to $q$.

Theorem 12: $\lim_{T \to \infty} \frac{\log N(T, q)}{T} = C$, where $C$ is the channel capacity, provided that $q$ does not equal 0 or 1.
The method of proving the first part of this theorem is not by exhibiting a coding method having the desired properties, but by showing that such a code must exist in a certain group of codes. In fact we will average the frequency of errors over this group and show that this average can be made less than $\epsilon$. If the average of a set of numbers is less than $\epsilon$ there must exist at least one in the set which is less than $\epsilon$. This will establish the desired result.

possible coding systems. This is the same as calculating the frequency of errors for a random association of the messages and channel inputs of duration $T$. Suppose a particular out-
Fig. 10—Schematic representation of the relations between inputs and outputs in a channel.
Now $R < H(x) - H_y(x)$ so $R - H(x) = -H_y(x) - \eta$ with $\eta$ positive. Consequently

$$P = \left[1 - 2^{-TH_y(x) - T\eta}\right] 2^{TH_y(x)}$$

approaches (as $T \to \infty$)

$$1 - 2^{-T\eta}.$$ 

Hence the probability of an error approaches zero and the first part of the theorem is proved.
Actually more has been proved than was stated in the theorem. If the average of a set of positive numbers is within $\epsilon$ of zero, a fraction of at most $\sqrt{\epsilon}$ can have values greater than $\sqrt{\epsilon}$. Since $\epsilon$ is arbitrarily small we can say that almost all the systems are arbitrarily close to the ideal.
As in the noiseless case, a delay is generally required to approach the ideal encoding. It now has the additional function of allowing a large sample of noise to affect the signal before any judgment is made at the receiving point as to the original message. Increasing the sample size always sharpens the possible statistical assertions.
if the source already has a certain redundancy and no attempt is made to eliminate it in matching to the channel, this redundancy will help combat noise.
17. **An Example of Efficient Coding**

are two channel symbols, 0 and 1, and the noise affects them in blocks of seven symbols. A block of seven is either transmitted without error, or exactly one symbol of the seven is incorrect. These eight possibilities are equally likely. We have

\[
C = \text{Max} \left[ H(y) - H_x(y) \right] \\
= \frac{1}{7} \left[ 7 + \frac{8}{8} \log \frac{1}{8} \right] \\
= \frac{4}{7} \text{ bits/symbol.}
\]
due to R. Hamming):

Let a block of seven symbols be $X_1, X_2, \ldots, X_7$. Of these $X_3, X_5, X_6$ and $X_7$ are message symbols and chosen arbitrarily by the source. The other three are redundant and calculated as follows:

$$X_4 \quad \text{is chosen to make} \quad \alpha = X_4 + X_5 + X_6 + X_7 \quad \text{even}$$
$$X_2 \quad \text{“ “ “ “ “} \quad \beta = X_2 + X_3 + X_6 + X_7 \quad \text{“ “}$$
$$X_1 \quad \text{“ “ “ “ “} \quad \gamma = X_1 + X_3 + X_5 + X_7 \quad \text{“ “}$$

When a block of seven is received $\alpha, \beta$ and $\gamma$ are calculated and if even called zero, if odd called one. The binary number $\alpha \beta \gamma$ then gives the subscript of the $X_i$ that is incorrect (if 0 there was no error).
PART III: CONTINUOUS INFORMATION

We will not attempt, in the continuous case, to obtain our results with the greatest generality, or with the extreme rigor of pure mathematics, since this would involve a great deal of abstract measure theory and would obscure the main thread of the analysis. A preliminary study, however, indicates that
**Theorem 13:** Let $f(t)$ contain no frequencies over $W$. Then

$$f(t) = \sum_{n=-\infty}^{\infty} X_n \frac{\sin \pi (2Wt - n)}{\pi (2Wt - n)}$$

where

$$X_n = f \left( \frac{n}{2W} \right).$$
20. Entropy of a Continuous Distribution

The entropy of a discrete set of probabilities $p_1, \ldots, p_n$ has been defined as:

$$H = - \sum p_i \log p_i.$$  

In an analogous manner we define the entropy of a continuous distribution with the density distribution function $p(x)$ by:

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) \, dx.$$
Similarly the $n$ dimensional Gaussian distribution with associated quadratic form $a_{ij}$ is given by

$$p(x_1, \ldots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum a_{ij}x_ix_j\right)$$

and the entropy can be calculated as

$$H = \log(2\pi e)^{n/2} |a_{ij}|^{-\frac{1}{2}}$$

where $|a_{ij}|$ is the determinant whose elements are $a_{ij}$. 
The following result is derived in Appendix 6

**Theorem 15:** Let the average power of two ensembles be $N_1$ and $N_2$ and let their entropy powers be $\overline{N}_1$ and $\overline{N}_2$. Then the entropy power of the sum, $\overline{N}_3$, is bounded by

$$\overline{N}_1 + \overline{N}_2 \leq \overline{N}_3 \leq N_1 + N_2.$$
The channel capacity is thus expressed as follows:

\[ C = \lim_{T \to \infty} \max_{P(x)} \frac{1}{T} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy. \]

\( x \) and \( y \) are transformed in any one-to-one way. This integral expression for \( C \) is more general than \( H(x) - H_y(x) \). Properly interpreted (see Appendix 7) it will always exist while \( H(x) - H_y(x) \) may assume an indeterminate form \( \infty - \infty \) in
Theorem 17: The capacity of a channel of band $W$ perturbed by white thermal noise power $N$ when the average transmitter power is limited to $P$ is given by

$$C = W \log \frac{P + N}{N}.$$ 

To approximate this limiting rate of transmission the transmitted signals must approximate, in statistical properties, a white noise.
The name “white noise,” already firmly entrenched in the literature, is perhaps somewhat unfortunate. In optics white light means either any continuous spectrum as contrasted with a point spectrum, or a spectrum which is flat with *wavelength* (which is not the same as a spectrum flat with frequency).
Theorem 18: The capacity of a channel of band $W$ perturbed by an arbitrary noise is bounded by the inequalities

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

where

$P = \text{average transmitter power}$

$N = \text{average noise power}$

$N_1 = \text{entropy power of the noise}$.
PART V: THE RATE FOR A CONTINUOUS SOURCE

\[ \iint P(x, y) \rho(x, y) \, dx \, dy. \]

It is interesting to note that, in this system, the noise in the recovered message is actually produced by a kind of general quantizing at the transmitter and not produced by the noise in the channel. It is more or less analogous to the quantizing noise in PCM.
\[ \rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 \, dt. \]

\[ \rho(x, y) = \frac{1}{T} \int_0^T |x(t) - y(t)| \, dt. \]

5. The discrete case can be considered as a specialization in which we have tacitly assumed an evaluation based on the frequency of errors. The function \( \rho(x, y) \) is then defined as the number of symbols in the sequence \( y \) differing from the corresponding symbols in \( x \) divided by the total number of symbols in \( x \).
We define the rate $R_1$ of generating information for a given quality $v_1$ of reproduction to be the minimum of $R$ when we keep $v$ fixed at $v_1$ and vary $P_x(y)$. That is:

$$R_1 = \operatorname{Min}_{P_x(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy$$

subject to the constraint:

$$v_1 = \iint P(x, y) \rho(x, y) \, dx \, dy.$$
Theorem 21: If a source has a rate $R_1$ for a valuation $v_1$ it is possible to encode the output of the source and transmit it over a channel of capacity $C$ with fidelity as near $v_1$ as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$. 
Suppose that $P_1(x,y)$ is the particular system which minimizes the rate and gives $R_1$. We choose from the high probability $y$'s a set at random containing

$$2^{(R_1 + \epsilon)T}$$

members where $\epsilon \to 0$ as $T \to \infty$. 
Theorem 22: The rate for a \textit{white} noise source of power $Q$ and band $W_1$ relative to an R.M.S. measure of fidelity is

$$R = W_1 \log \frac{Q}{N}$$

where $N$ is the allowed mean square error between original and recovered messages.
Acknowledgements

The writer is indebted to his colleagues at the Laboratories, particularly to Dr. H. W. Bode, Dr. J. R. Pierce, Dr. B. McMillan, and Dr. B. M. Oliver for many helpful suggestions and criticisms during the course of this work. Credit should also be given to Professor N. Wiener, whose elegant solution of the problems of filtering and prediction of stationary ensembles has considerably influenced the writer’s thinking in this field.
What is Information Theory?

NORBERT WIENER

INFORMATION THEORY has been identified in the public mind to denote the theory of information by bits, as developed by Claude E. Shannon and myself.
what surprises me the most in Shannon’s paper is...

I have read this paper several times since my undergrad years and every time I read it I discover something new and fascinate that I did not realize previously.

It gives you an idea of “hope”. As poetic as it might sound, its descriptive (though not constructive) theory has established the foundations of what can expect from any communication system giving us an idea of how marvelous a device could be.
what surprises me the most in Shannon’s paper is...

the essence of all that I know in information theory (and a lot more of course) is contained in Shannon's first paper on the subject.
what surprises me the most in Shannon’s paper is...

it is surprisingly easy to read. The examples are very helpful, and in general the theory is presented as if it had been around a long time.
what surprises me the most in Shannon’s paper is...

the clear engineering concept he had what a communication system is composed of, and the simplicity of his ideas and examples.
what surprises me the most in Shannon’s paper is...

how simple it is

Even after over 60 years since he published his work, I feel that the importance/relevance of his work remains the same.
what surprises me the most in Shannon’s paper is...

One does not need to read another paper, or study some special mathematical subject in detail to understand the paper. Today, I feel the need to know much more, read so much on prior works etc., just to understand what a paper claims.
what surprises me the most in Shannon’s paper is...

the clarity in the exposition of a collection of profound concepts. This is seldom (if at all) seen in today's publications where authors seek to obfuscate instead of illuminate the ideas in their work via examples and discussions.
what surprises me the most in Shannon’s paper is...

thinking made clear by examples from real data such as English text enabled the derivation of such a longstanding mathematical theory of communication.
what surprises me the most in Shannon’s paper is...

the phrase "The theorem says that..." several times in the paper where Shannon gives deep insight into what the theorems mean.
what surprises me the most in Shannon’s paper is...

I've long been under the impression that Shannon avoided excessive rigor so as to reach a wider audience. This time around, it struck me that the lack of absolute technical precision may also be indicative of Shannon's thought process. While we as researchers are frequently encouraged to be very precise in our problem statements, we must recognize that research is both a deductive science and a creative art. In the latter domain, "thinking vaguely" about a problem can be a powerful tool, and I don't find it unreasonable to suggest that Shannon had a proclivity for this.
what surprises me the most in Shannon’s paper is...

“These semantic aspects of communication are irrelevant to the engineering problem”

Unfortunately, his little concession has had very far-reaching consequences and is nowadays still often used to diminish and even ridicule the importance of Shannon’s insights: this happens especially by communication scholars from the social science, and especially by those coming from the humanities. These often quote Shannon’s introductory quote as a “proof” of their point that communication is much more than what engineers understand.
what surprises me the most in Shannon’s paper is...

Einstein finds the relationship between mass and energy, while Shannon builds a connection between energy and information.
what surprises me the most in Shannon’s paper is...

The definition of continuous entropy is not as natural as the discrete one (although it's a natural extension of the discrete one, the possibility of being negative makes is not very well clear to me)
what surprises me the most in Shannon’s paper is...

he came up with his innovative theory without being aware of all problems that it can be applied to.
what surprises me the most in Shannon’s paper is...

how quickly I found application of principles and ideas in his paper to some things I have been working on recently.
what surprises me the most in Shannon’s paper is...

reminds me of what professor Toby Berger said during his Information Theory class:

don't try to make the problem fit your tools; get yourself the tools that fit the problem.
what surprises me the most in Shannon’s paper is...

due to the similarity with Sir Isaac Newton's Mathematical Principles of Natural Philosophy. Besides the coincidence of the titles, they both initiated new eras of their fields with elegant mathematics and extraordinary intentions, even if some of their proofs were not rigorously given until many years later. It is the intuition rooted deeply inside two mathematical hearts that makes everything happen.
what surprises me the most in Shannon’s paper is...

re-reading this paper is not about learning new theorems, but is about seeing a vivid movie of how a great scientist found an important problem and how he devoted himself to solve it.
what surprises me the most in Shannon’s paper is... 

the power of imagination in science. On one hand coding is very practical thing, usually in the way of a mapping form or somewhat more complex, while on the other hand, Shannon imagined the code to be infinitely long, even though such code might not exist in practice. In other words, I think the beauty of the work is all around the "limit". We know Shannon's work gives upper bounds of channel capacity, and proves codes satisfying those conditions exist, yet all these based on the concept of "limit". But what is limit? It's imagination.
what surprises me the most in Shannon’s paper is...

For me, the surprise comes when we frame Shannon's 1948 paper in a historical context. Probabilistic arguments of the same type that Shannon used first appeared in the literature only a year before in Erdos' 1947 paper that accomplished the same thing for the probabilistic method as Shannon's paper did for the field of information theory [arguably the probabilistic method was first used by Szele in 1943, but Shannon was probably unaware of this since it was published in Russian].
what surprises me the most in Shannon’s paper is...

inventing a new method
--random coding
-- $H(X) - H(X|Y) = H(Y) - H(Y|X)$

turns out to be much more important than the results themselves or the rigor of the proofs.
what surprises me the most in Shannon’s paper is...

its impact on society. Not unlike a piece of art, the paper inspired its audience to follow and improve his work, yet always praise the foundation.
what surprises me the most in Shannon’s paper is...

how he leverages the idea of typical sets