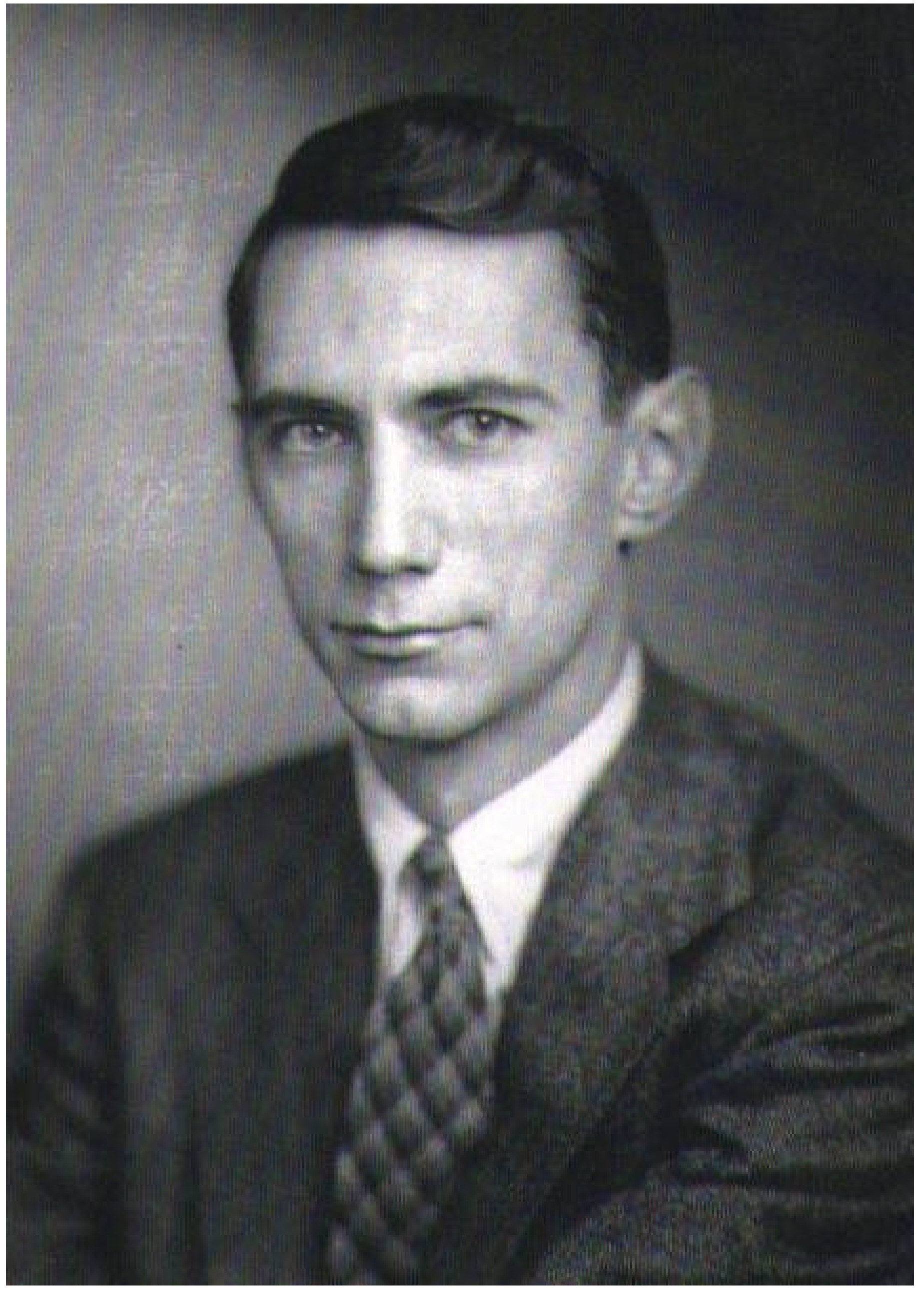
“A Mathematical Theory of Communication”

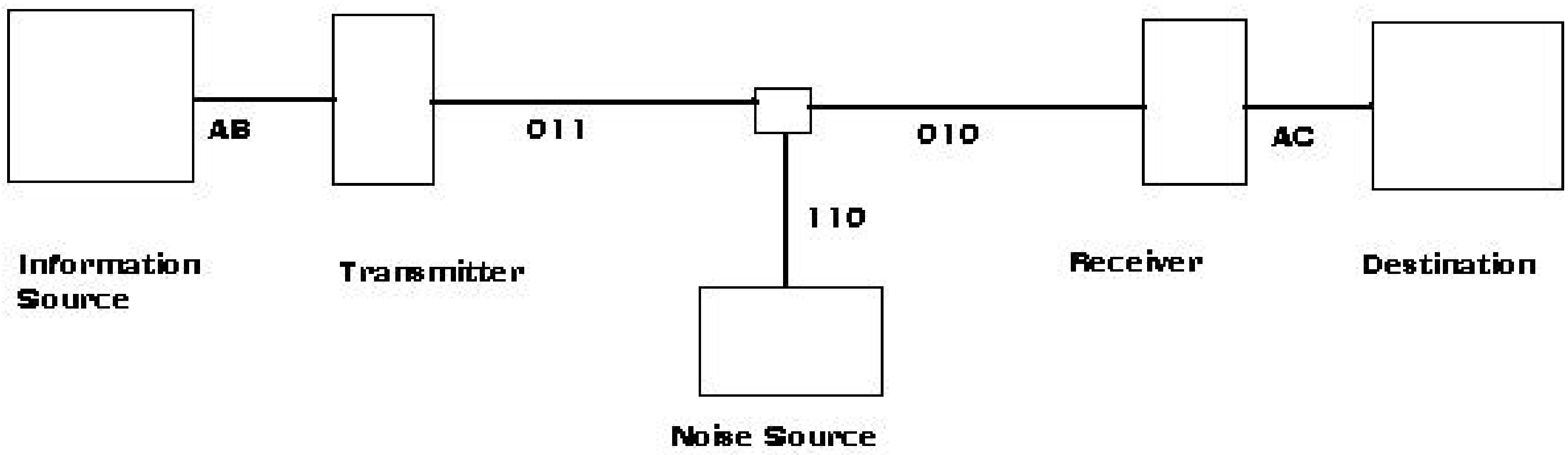
Claude Shannon’s paper

presented by Kate Jenkins

2/19/00

- Published in two parts, July 1948 and October 1948 in the Bell System Technical Journal
- Founding paper of Information Theory
- First person to use a probabilistic model of communication
- Developed around same time as Coding Theory
- Huge Impact:
    - now *the* mathematical theory of communication
    - follow on papers
    - idea that “all information is essentially digital”
    - telecommunications, CD players, computer networks
    - applications to biology, artificial intelligence..





Questions:

• How much information is produced by a source? (info/symbol or info/sec)

•How quickly can information be transmitted through a channel? (info/sec)

•What is best achievable transmission rate (source symbols/sec)?

•If channel has noise, under what conditions can the sent message be reconstructed from the received message?

What is information?

Acquiring information = Reducing uncertainty

Amount of information = Level of “surprise”

$$S = \{s_i : 1 \le i \le n\} \quad \text{set of all possible events}$$

$$p_i = \text{probability that } s_i \text{ occurs}$$

$$\text{Information}(s_i) = \log_2 1/p_i \text{ "bits"}$$

$$\text{Example:} \quad S = \{0,1\} \qquad S_N = \{0,1\}^N$$

$$p_0 = p_1 = 1/2 \qquad \Rightarrow \quad \text{Info}(0) = \text{Info}(1) = 1\,\text{bit}$$

$$s \in S_N \Rightarrow p_s = 1/2^N \Rightarrow \text{Info}(s) = \text{ N bits}$$

$$p_0 = 1/16,\ p_1 = 15/16 \Rightarrow \text{Info}(0) = 4\,\text{bits}$$

$$p_0 = 0,\ p_1 = 1 \qquad \Rightarrow \quad \text{Info}(1) = 0\,\text{bits}$$

Channel capacity measured in bits/sec

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

$N(T) =$ number of allowed signals of duration T

Example: Digital channel

All $\{0,1\}$ sequences allowed, produce r symbols/sec.

$$N(T) = 2^{rT}$$

$$C = r \text{ bits/sec}$$

Allows more complicated channel structures:

- varying time per symbol
- restrictions on allowed sequences of symbols

Define information generated by source (measured in bits/symbol) to be expected amount of information generated per symbol.

Recall,

$$\mathrm{Info}(s_i) = \log 1/p_i, \; s_i \in S$$

So,

$$\mathrm{E(Info)} = \sum_{s_i \in S} p_i \log 1/p_i$$

Call this quantity the “Entropy” of the source. Use the symbol H.

$$\mathrm{H(x)} = -\sum_{s_i \in S} p_i \log p_i$$

Where x is a random variable representing our signal.

Nice properties of Entropy:

$$\mathrm{H} \geq 0$$

$$\mathrm{H} = 0 \text{ only if } p_i = 1 \text{ for some i}$$

$$\text{If } |S| = n, \mathrm{H} \text{ is maximized when } p_i = 1/n \, \forall i$$

$$\text{Suppose x, y two events,}$$

$$\text{then } \mathrm{H(x, y)} = -\sum_{i,j} p(i,j) \log p(i,j) \leq \mathrm{H(x)} + \mathrm{H(y)}$$

$$\mathrm{H(x, y)} = \mathrm{H(x)} + \mathrm{H(y)} \text{ only if x, y independent.}$$

Define Conditional Entropy (uncertainty of y given value of x):

$$\mathrm{H_x(y)} = \sum_i p(i) \mathrm{H}_i(\mathrm{y}) = -\sum_{i,j} p(i) p_i(j) \log p_i(j)$$

$$= -\sum_{i,j} p(i,j) \log p_i(j)$$

$$\text{Then } \mathrm{H(x, y)} = \mathrm{H(x)} + \mathrm{H_x(y)}, \text{ and } \mathrm{H(y)} \geq \mathrm{H_x(y)}$$

Now consider messages of length N, N large.

Suppose source produces each symbol independently at random.

Then with high probability, for a message m

$$\#\text{of occurences of } s_i \text{ in } m \approx p_i N \ \forall i$$

$$\text{so } p_m \approx \prod_i p_i^{p_i N}$$

$$\Rightarrow \log p_m \approx \sum_i N p_i \log p_i = -NH$$

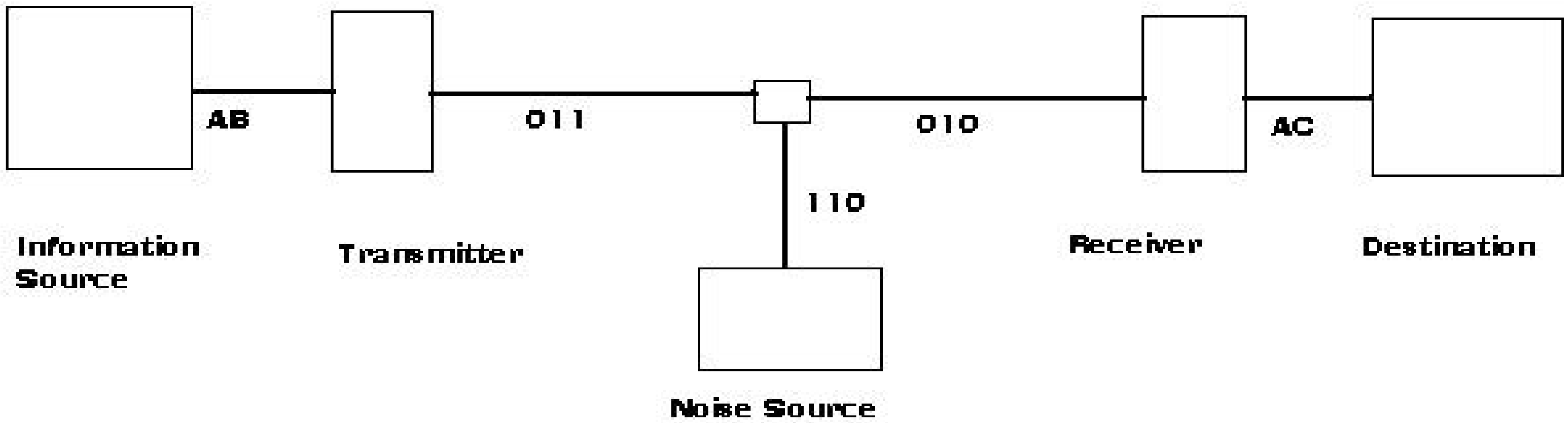
$$\Rightarrow p_m \approx 2^{-HN}$$

$\Rightarrow$ for N large, have $\approx 2^{\mathrm{HN}}$ probable, equally likely messages,

This result also holds for more complicated source models.

For ergodic Markov processes, use entropy $\mathrm{H} = \sum_{\mathrm{i} \in \mathrm{states}} \mathrm{P_i H_i}$

A channel with noise:



Consider two distinct signals

x = signal input into the channel

y = signal received at the other end

$$\text{Equivocation} = H_y(x)$$

$$\text{Rate of actual transmission } R(x) = H(x) - H_y(x) \text{ bits/sec}$$

$$\text{Channel capacity } C = \max_{\text{info sources}} R(x) = \max_{\text{info sources}} (H(x) - H_y(x))$$

The Fundamental Theorem for a Discrete Channel with Noise:

Let a discrete channel have capacity C, and a discrete source have entropy H bits/second. If $H < C$, there exists a coding system such that the output of the source can be transmitted over the channel with arbitrarily small errors.

Proof:

Recall $C = \max_{\text{encodings}} R(x)$

Suppose encoding S attains this maximum (or arbitrarily close).

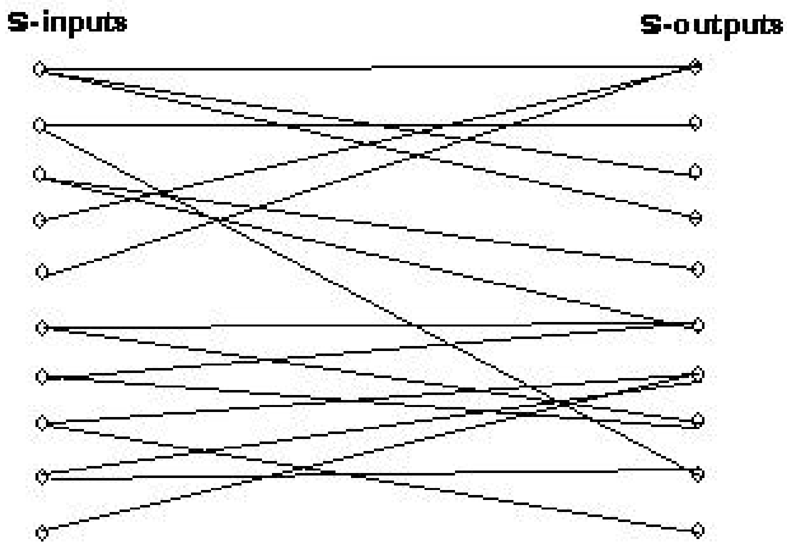
S has input entropy $H^S(x)$, output entropy $H^S(y)$. So there are

$2^{H^S(x)T}$ probable input messages of duration T,

$2^{H^S(y)T}$ probable received messages of duration T, and

$2^{H_x^S(y)T}$ probable inputs for a given output.

Construct a bipartite graph, where each node is a probable input or output message of duration T for source S. Connect nodes A and B by an edge if message A is an input likely to produce output B.



Let R be the source we're interested in, with entropy < C. Encode R by randomly assigning messages of duration T to nodes in the left column of the graph. Given an output message, the probability that it is connected to more than one R-input message is

$$\leq (2^{H^R T} / 2^{H^S T}) 2^{H_x^S T} = 2^{(H^R - (H^S - H_x^S) T} = 2^{(H^R - C)T} \rightarrow 0 \text{ as } T \rightarrow \infty$$

Extensions to Shannon’s work:

- Continuous source/channel (in 2nd part of paper)
- Consider multi-terminal case
- Consider multi-way channels (like telephone lines!)
- Consider more complicated source structures (non-ergodic!) and different memory models for transmitters.
- Kolmogorov applied Shannon’s ideas to solve long-standing problems in ergodic theory.
- Applications to biology:
    - Entropy of DNA to identify binding sites
    - Intra-organism communication

Discussion Topics:

•Any questions?

•Any Shannon anecdotes?

Required reading at NSA

Wrote good article on the mathematics of juggling

Made a maze-learning mouse out of phone-relays

Married a numerical analyst from Bell Labs

•Shannon says (p.413) that no explicit description is known of approximations to the ideal coding for a noisy channel. I understand this is still the case. Comments on what is done in practice?

•Other applications/impact of information theory?

•Any ideas about entropy of English and crossword puzzles? (p.399) How to go about proving such a result?