Shannon et la théorie de l'information

Premier Congrès Franco-Marocain De Mathématiques Appliquées

16 avril 2018

Olivier Rioul

<olivier.rioul@telecom-paristech.fr>

Institut
Mines-Télécom

Do you Know Claude Shannon?

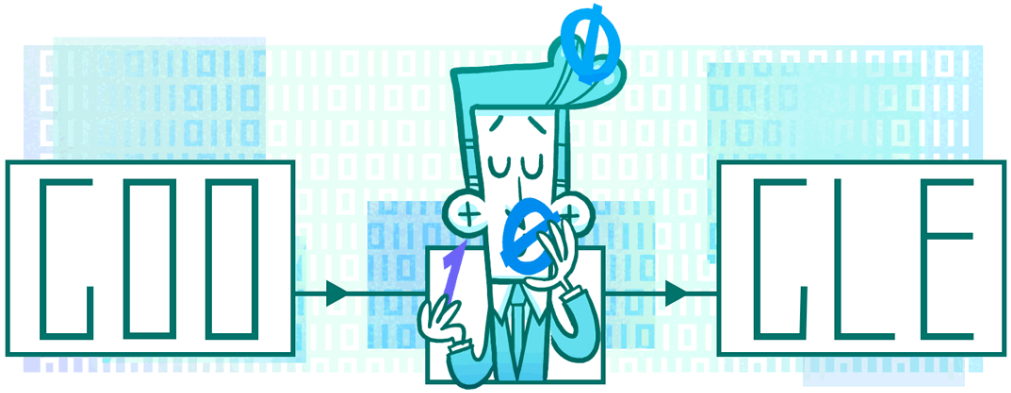
“the most important man... you’ve never heard of”

Claude Shannon (1916–2001)

“father of the information age”

April 30, 1916 Claude Elwood Shannon was born in Petoskey, Michigan, USA

April 30, 2016 centennial day celebrated by Google:



Well-Known Scientific Heroes

Alan Turing (1912–1954)

Well-Known Scientific Heroes

John Nash (1928–2015)

The Quiet and Modest Life of Shannon

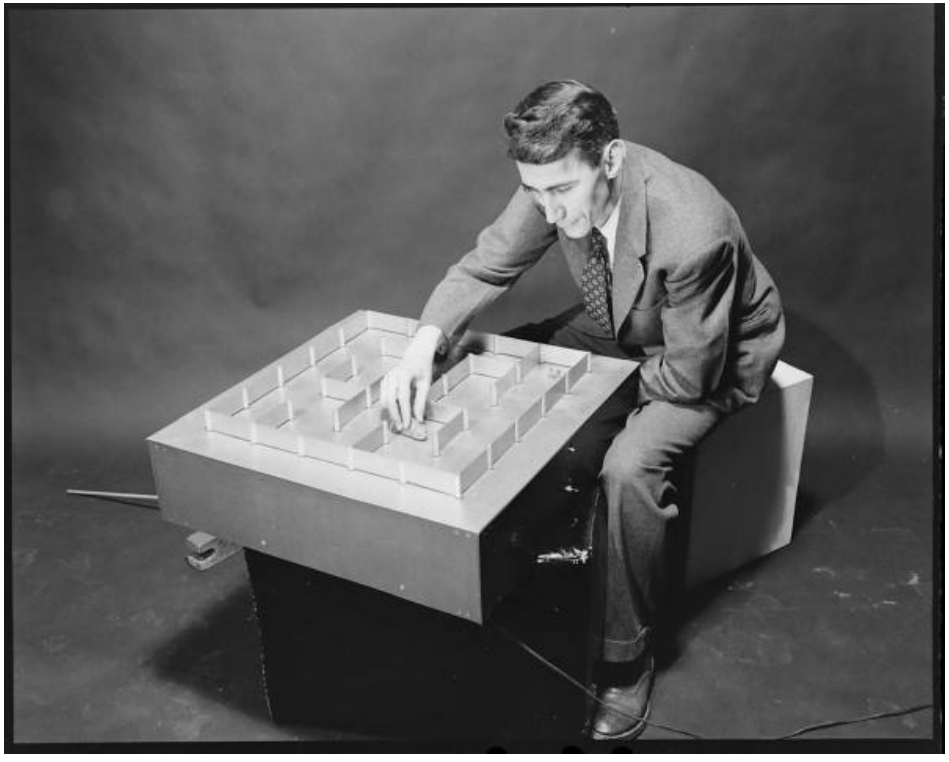
Shannon with Juggling Props

The Quiet and Modest Life of Shannon

Shannon's Toys Room

Shannon is known for riding through the halls of Bell Labs on a unicycle while simultaneously juggling four balls

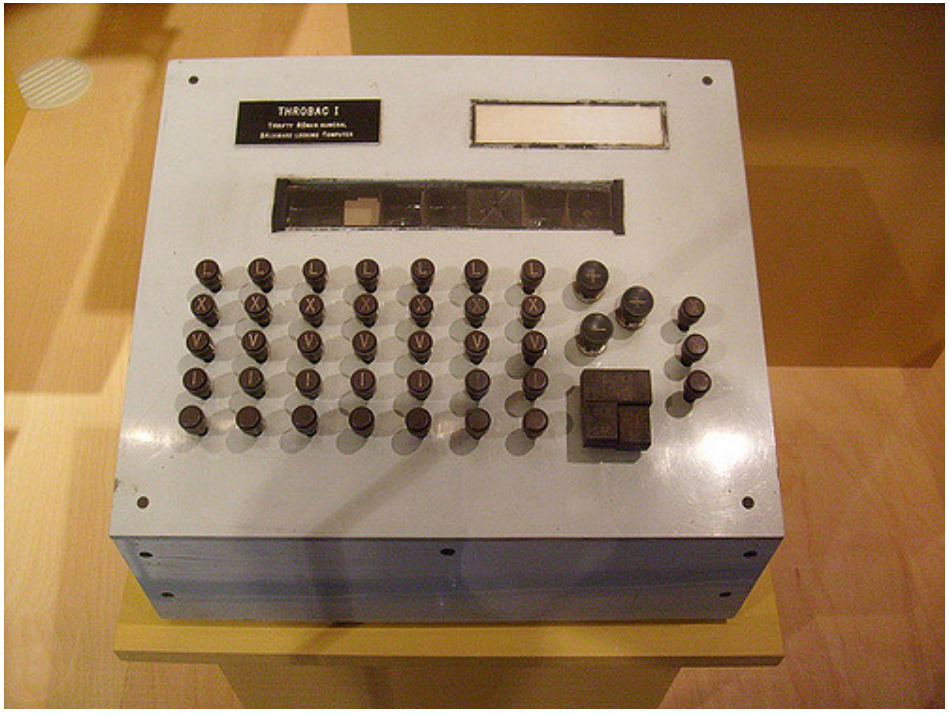
Crazy Machines



Theseus (labyrinth mouse)

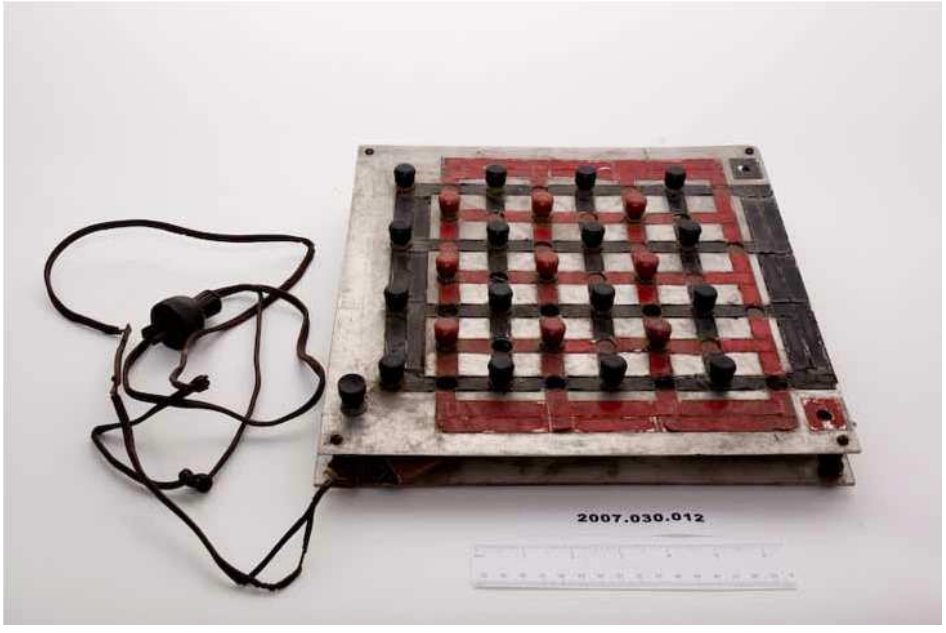
Crazy Machines

Crazy Machines



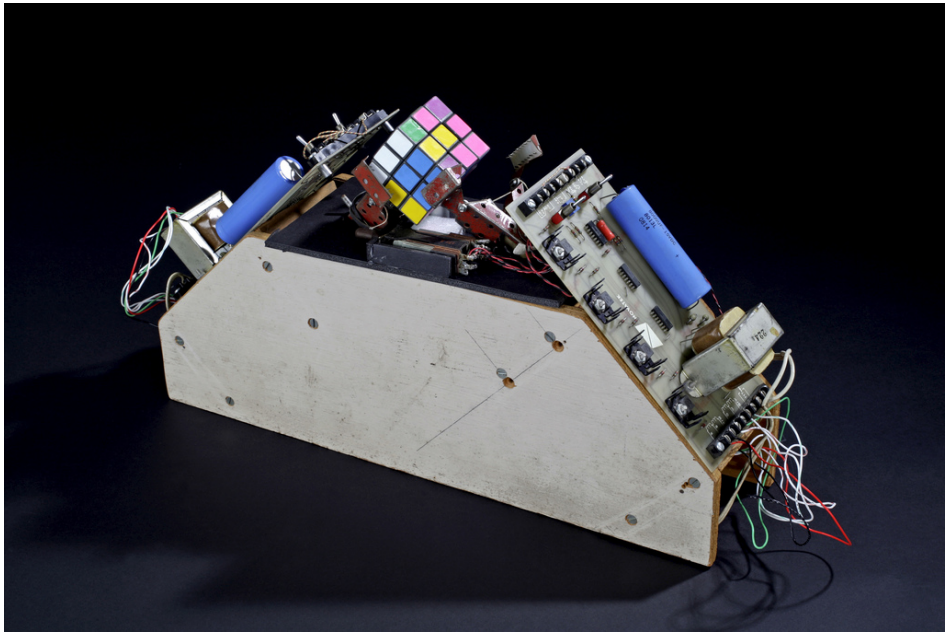
calculator in Roman numerals

Crazy Machines



“Hex” switching game machine

Crazy Machines

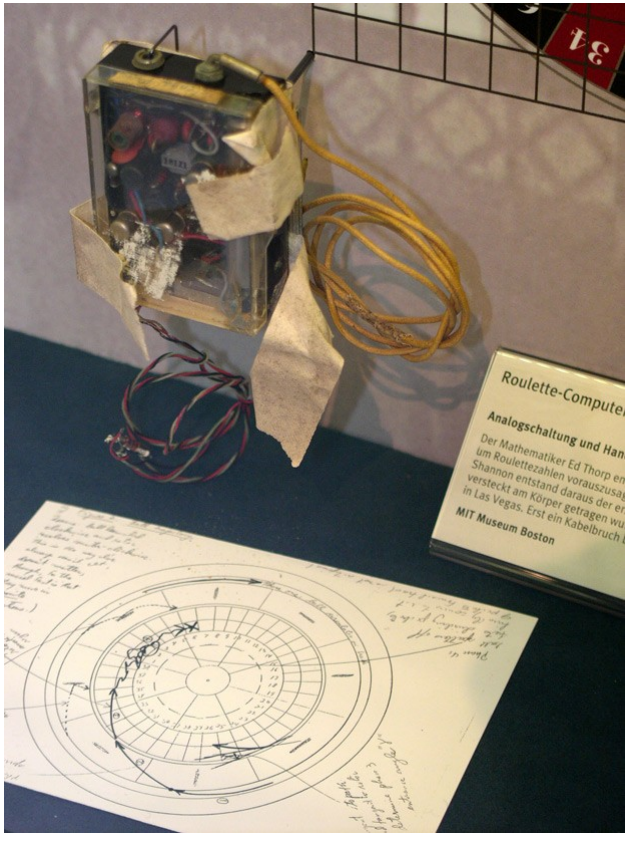


Rubik’s cube solver

Crazy Machines

3-ball juggling machine

Crazy Machines



Wearable computer to predict roulette in casinos
(with Edward Thorp)

Crazy Machines

ultimate useless machine

“Serious” Work

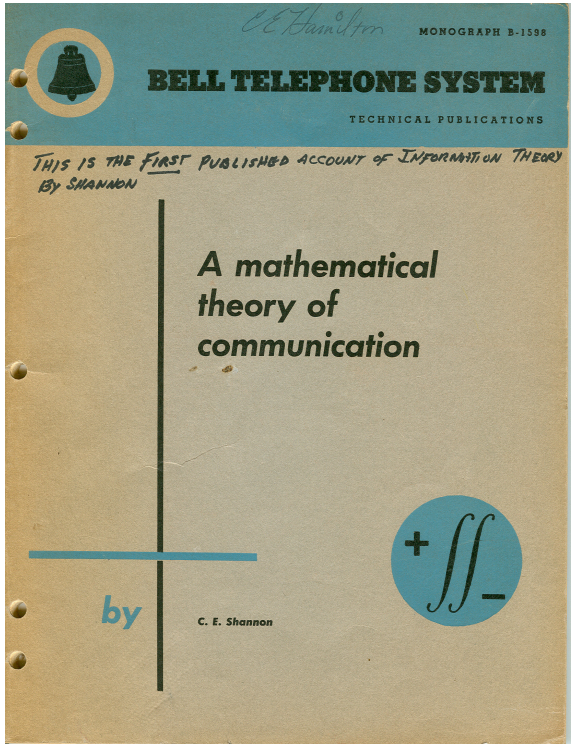
At the same time, Shannon made decisive theoretical advances in ...

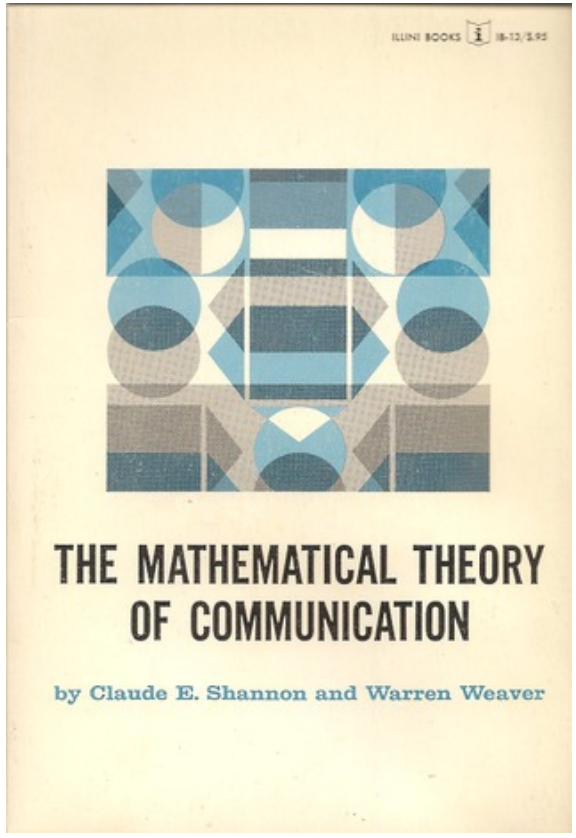
- logic & circuits
- cryptography
- artifical intelligence
- stock investment
- wearable computing

  ⋮

- ...and **information theory**!

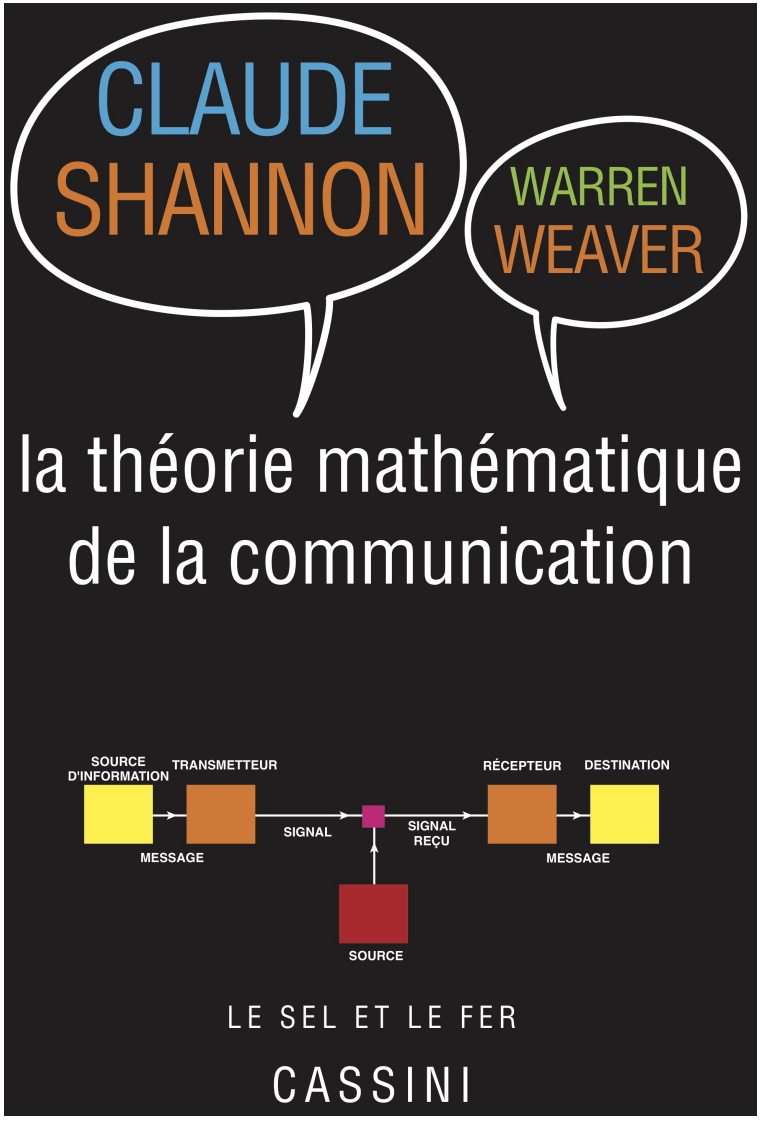
The Mathematical Theory of Communication (BSTJ, 1948)





One article (written 1940–48): **A REVOLUTION !!!!!!**

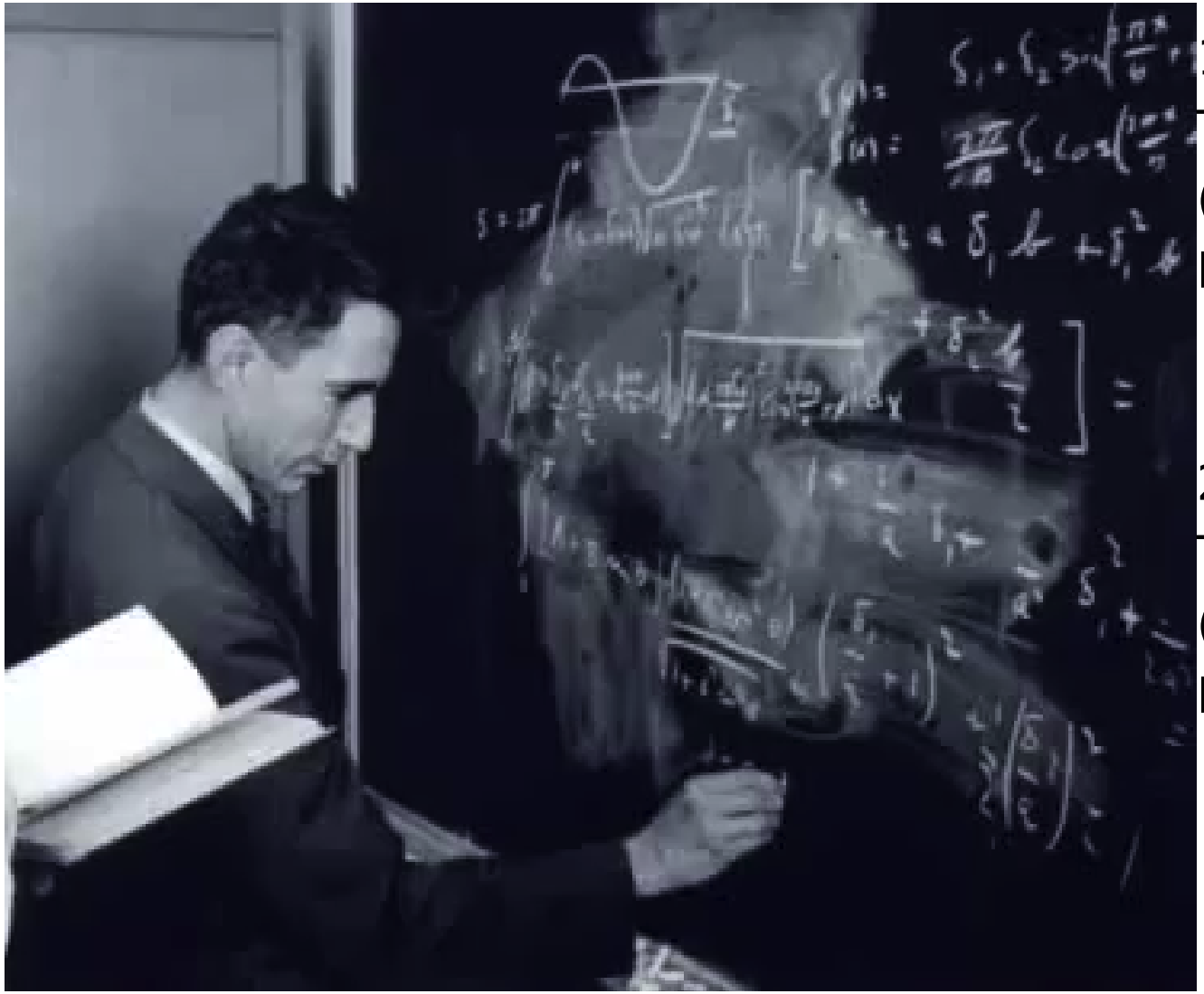
Nouvelle édition française



Without Shannon....

Shannon's Theorems

Yes it's Maths !!



1. Source Coding Theorem (*Compression* of Information)

2. Channel Coding Theorem (*Transmission* of Information)

Shannon's Paradigm

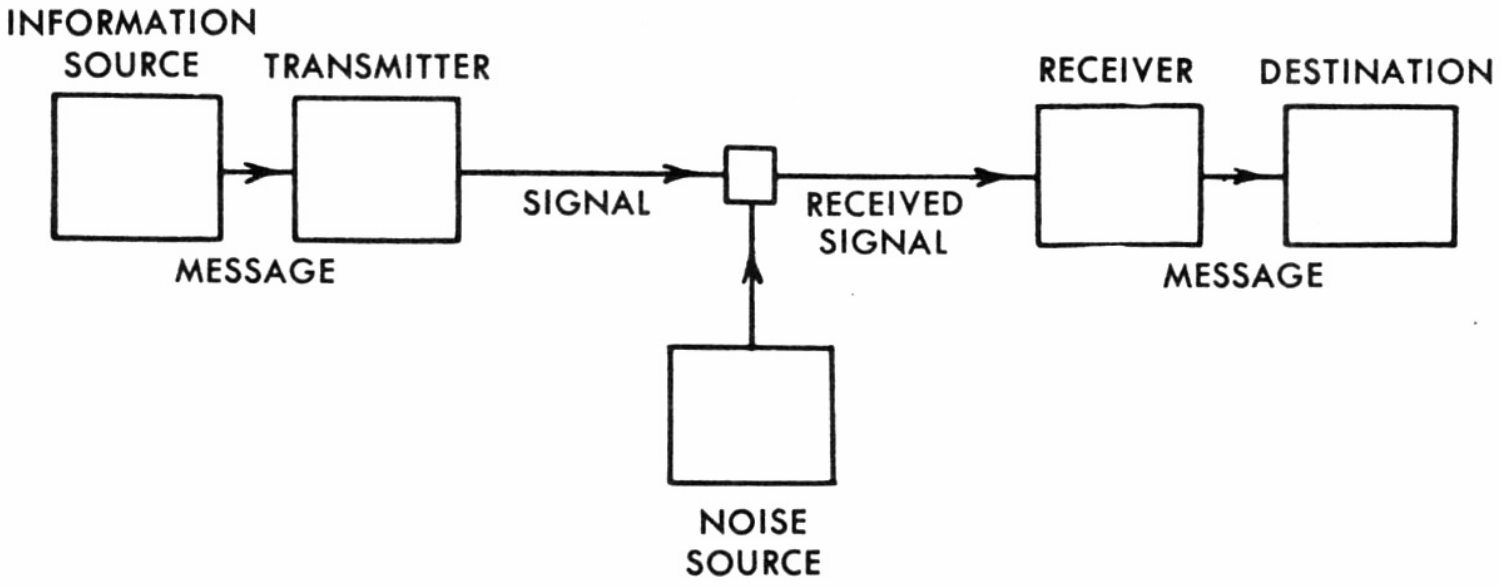
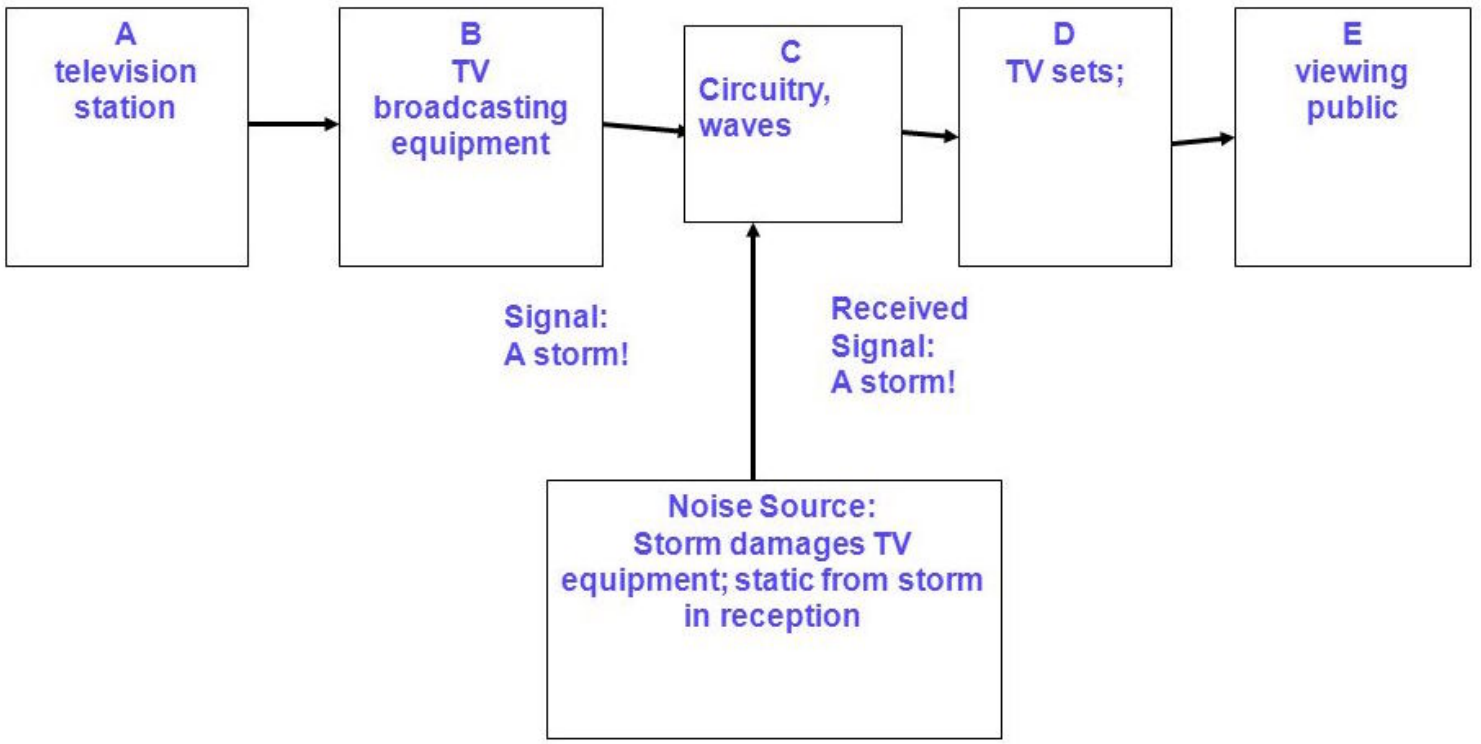


Fig. 1. — Schematic diagram of a general communication system.

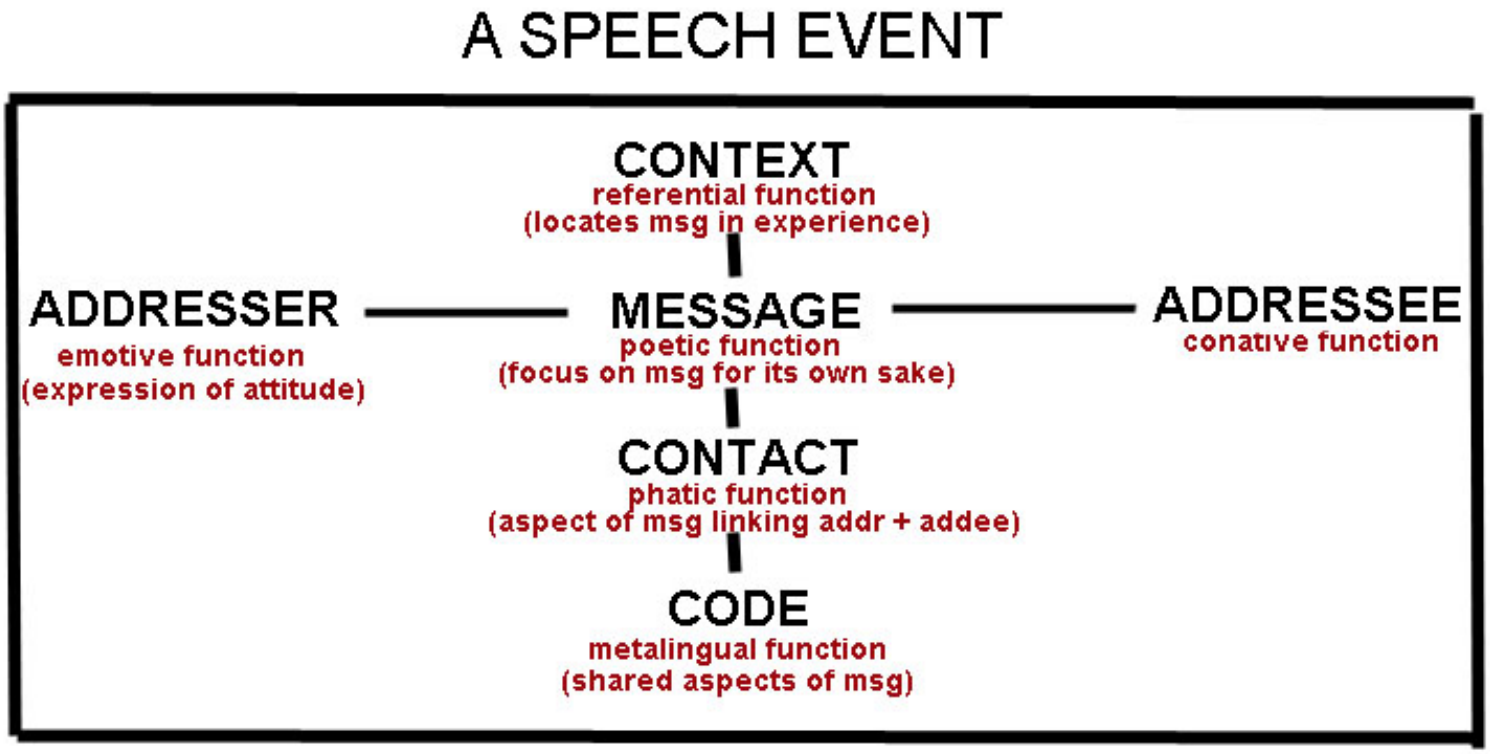
A tremendous impact!

Shannon's Paradigm... in Communication

Example: Broadcast following crisis

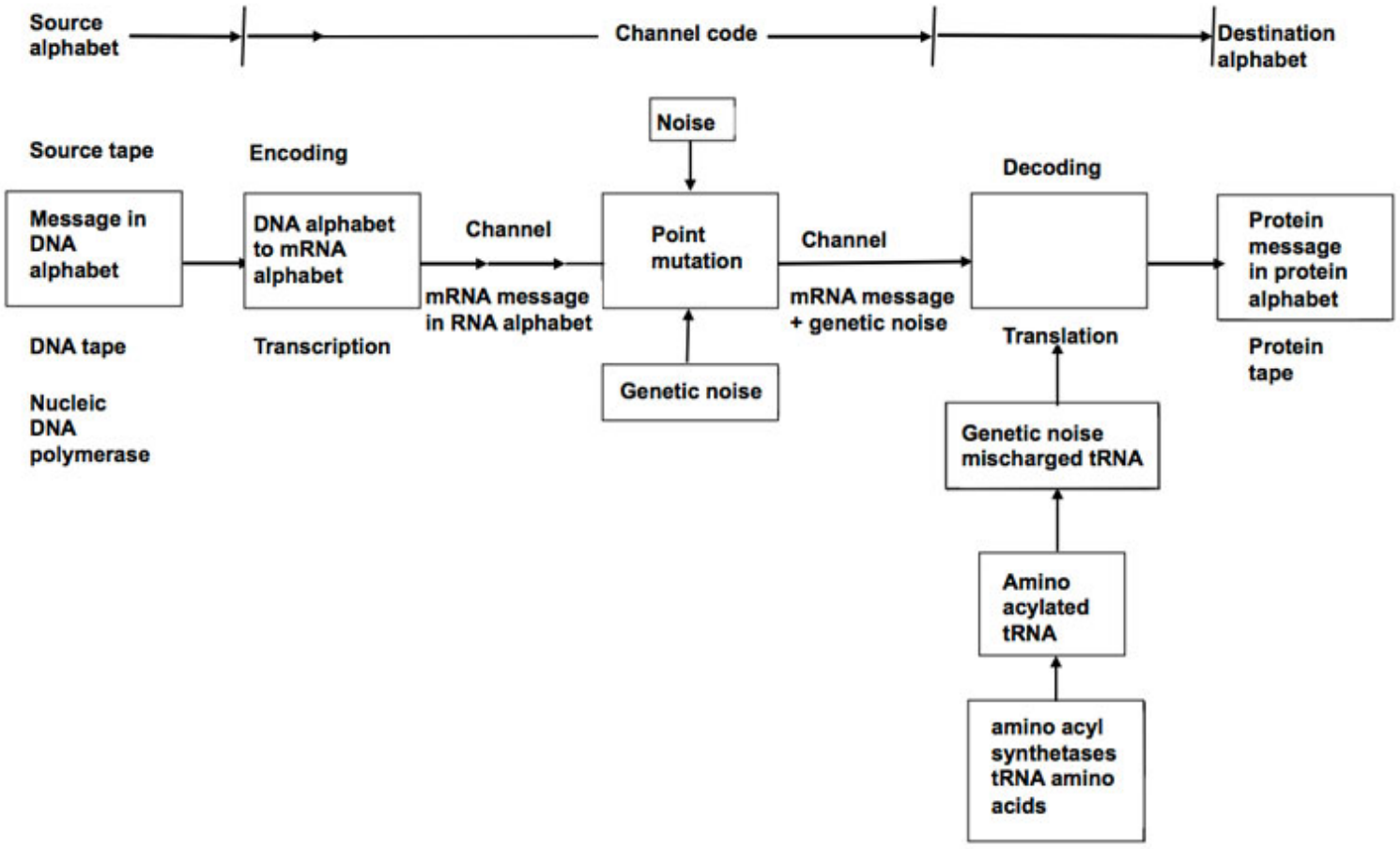


Shannon's Paradigm... in Linguistics

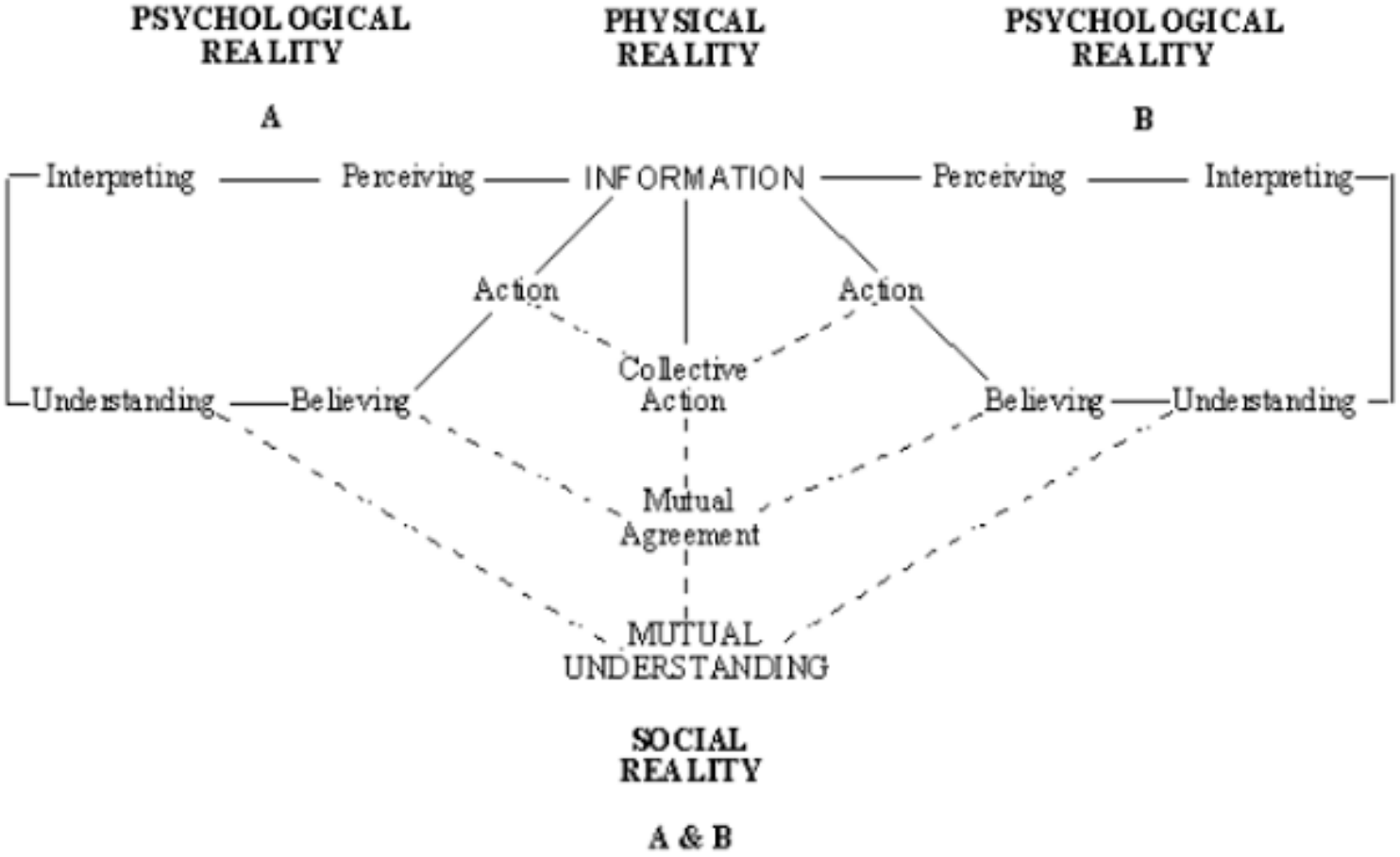


Roman Jakobson's 1960
model of communication

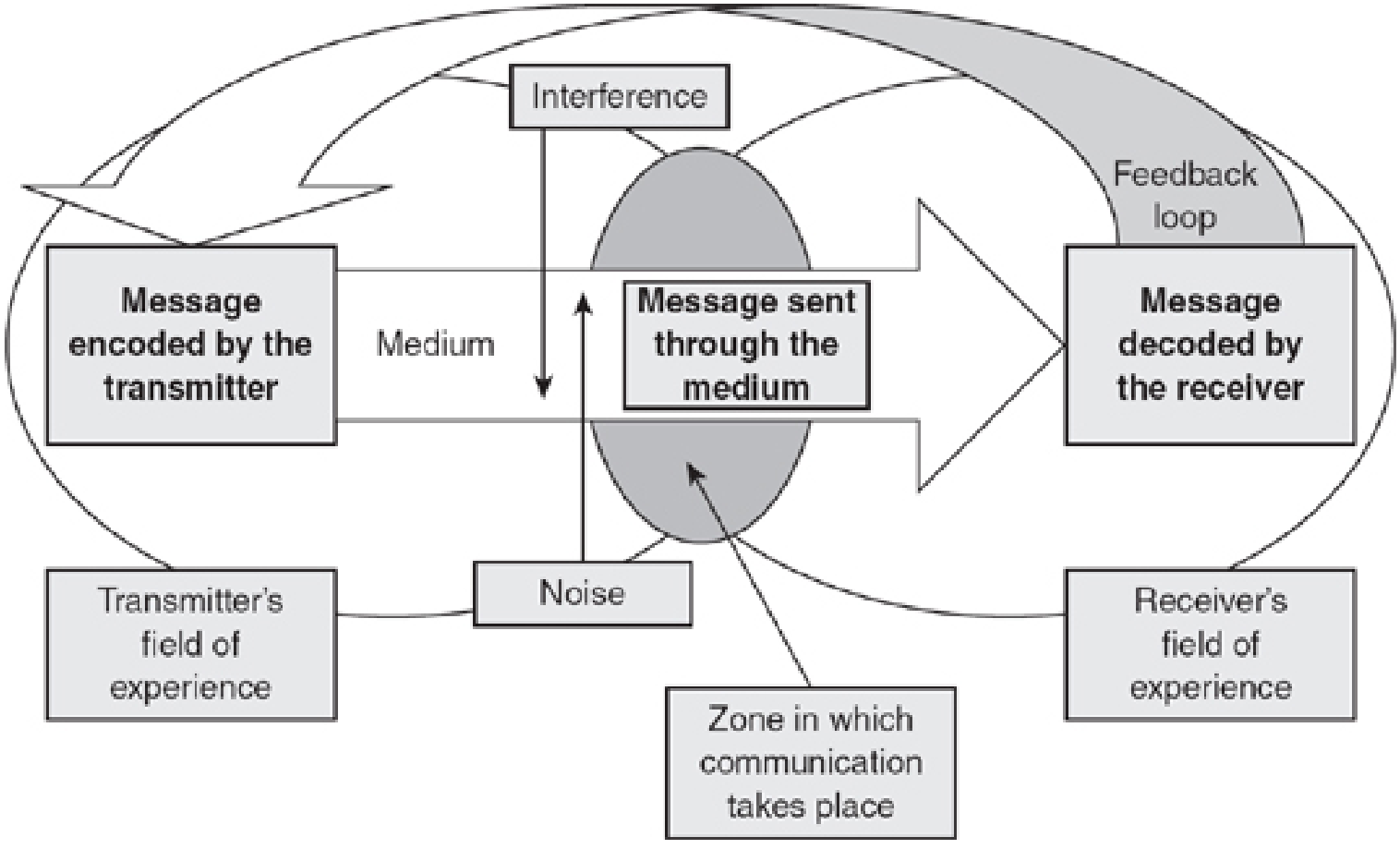
Shannon's Paradigm... in Biology



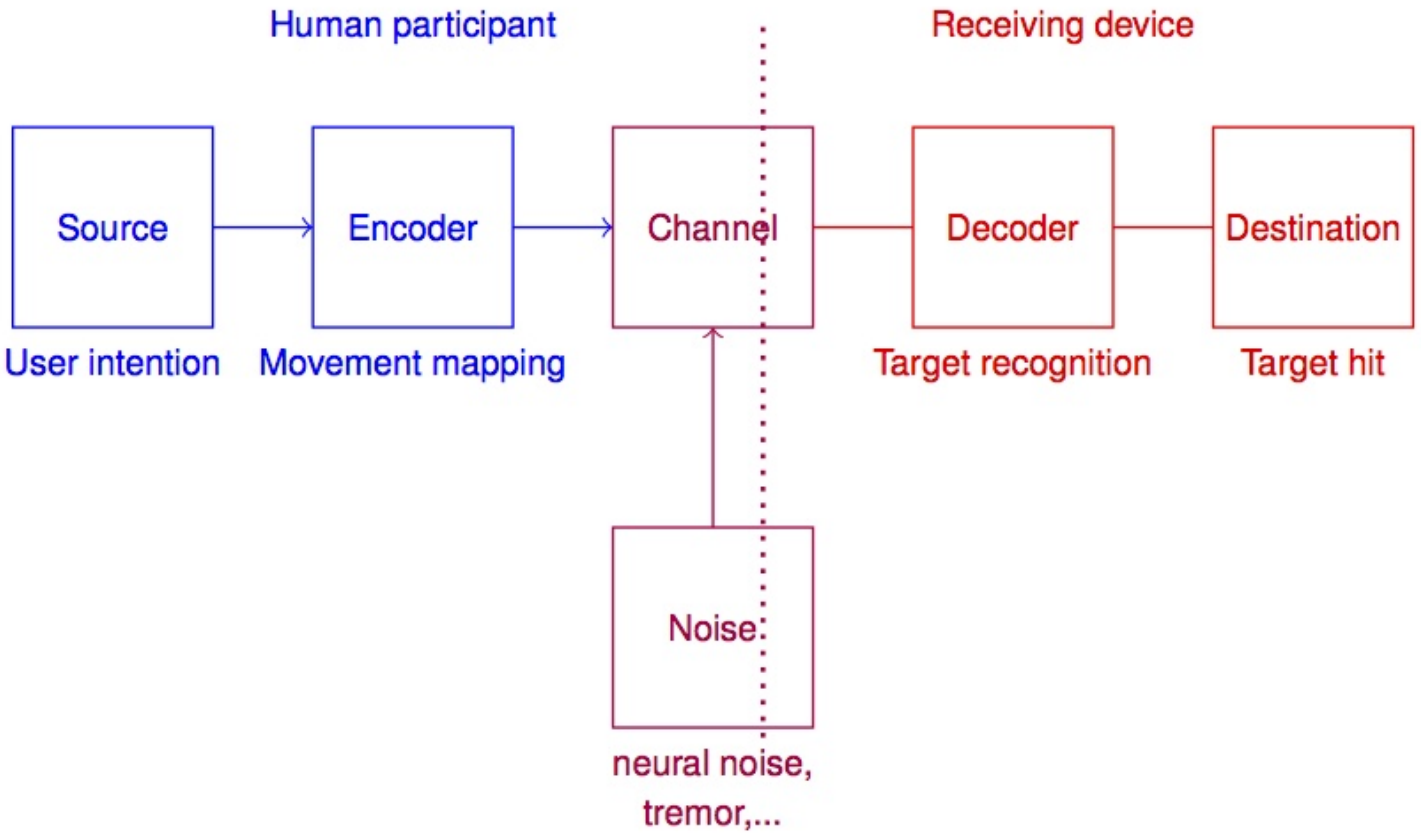
Shannon's Paradigm... in Psychology



Shannon's Paradigm... in Social Sciences



Shannon's Paradigm... in Human-Computer Interaction



Shannon's "Bandwagon" Editorial

The Bandwagon

CLAUDE E. SHANNON

INFORMATION theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, it has perhaps been ballooned to an importance beyond its actual accomplishments. Our fellow scien-

subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications. I personally believe that many of the concepts of information theory will prove useful in these other fields—and, indeed, some results are already quite

TELECOM ParisTech

Shannon's Viewpoint

"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; [...] 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 [...] unknown at the time of design. "

X : a message symbol modeled as a **random variable**

$p(x)$: the **probability** that $X = x$

Kolmogorov's Modern Probability Theory

Andreï Kolmogorov (1903–1987)

- founded modern probability theory in 1933
- a strong early supporter of information theory!

"Information theory must precede probability theory and not be based on it. [...] The concepts of information theory as applied to infinite sequences [...] can acquire a certain value in the investigation of the algorithmic side of mathematics as a whole."

A Logarithmic Measure

- 1 digit represents 10 numbers 0,1,2,3,4,5,6,7,8,9;
- 2 digits represents 100 numbers 00, 01, . . . , 99;
- 3 digits represents 1000 numbers 000, . . . , 999;

  ⋮
- $\log_{10} M$ digits represents M possible outcomes

Ralph Hartley (1888–1970)

"[...] take as our practical measure of information the logarithm of the number of possible symbol sequences"

Transmission of Information, BSTJ, 1928

The Bit

- $\log_{10} M$ digits represents M possible outcomes
- or...
- $\log_2 M$ **bits** represents M possible outcomes

John Tukey (1915–2000)
coined the term “bit” (contraction of “binary digit”)
which was first used by Shannon in his 1948 paper

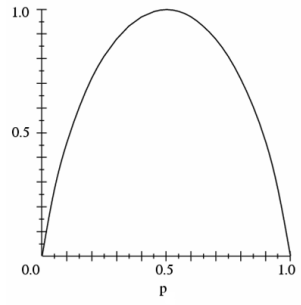
any information can be represented by a sequence of 0’s and 1’s — the Digital Revolution!

The Unit of Information

bit (binary digit, unit of storage) $\neq$ bit (binary unit of information)

- less-likely messages are more informative than more-likely ones
- 1 bit is the information content of one equiprobable bit $(\frac{1}{2}, \frac{1}{2})$

otherwise the information content is < 1 bit:

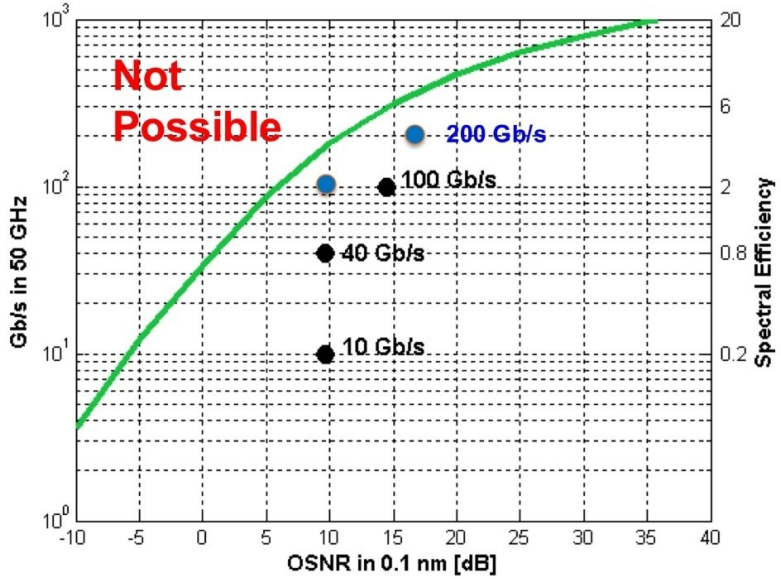


The official name (International standard ISO/IEC 80000-13) for the information unit:

...the **Shannon** (symbol Sh)

Fundamental Limit of Performance

- Shannon does not really give *practical* solutions but solves a *theoretical* problem:
- *No matter what you do*,
  (as long as you have a given amount of ressources)
  you *cannot* go beyond than a certain bit rate limit
  to achieve reliable communication



Fundamental Limit of Performance

- before Shannon: 
  communication technologies did *not* have a landmark
- the limit can be calculated: we know how far we are from it
  and you can be (in theory) *arbitrarily close* to the limit!
- the challenge becomes:
  how can we build practical solutions that are close to the limit?

Asymptotic Results

- to find the limits of performance, Shannon's results are necessarily asymptotic
- a source is modeled as a sequence of random variables

$$X_1, X_2, \ldots, X_n$$

  where the dimension $n \to +\infty$.
- this allows to exploit dependences
  and obtain a geometric "gain"
  using the *law of large numbers*
  where limits are expressed as *expectations* $\mathbb{E}\{\cdot\}$

Asymptotic Results: Example

Consider the source $X_1, X_2, \ldots, X_n$ where each X can take a finite number of possible values, independently of the other symbols.

The probability of message $\underline{x} = (x_1, x_2, \ldots, x_n)$ is the product of the individual probabilities:

$$p(\underline{x}) = p(x_1) \cdot p(x_2) \cdot \cdots \cdot p(x_n).$$

Re-arrange according to the value x taken by each argument:

$$p(\underline{x}) = \prod_x p(x)^{n(x)}$$

where $n(x) =$ number of symbols equal to x.

Asymptotic Results: Example (Cont'd)

By the *law of large numbers*, the empirical probability (frequency)

$$\frac{n(x)}{n} \to p(x) \qquad \text{as } n \to +\infty$$

Therefore, a "typical" message $\underline{x} = (x_1, x_2, \ldots, x_n)$ satisfies

$$p(\underline{x}) = \prod_x p(x)^{n(x)} \approx \prod_x p(x)^{np(x)} = 2^{-n \cdot H}$$

where

$$\boxed{H = \sum_x p(x) \log_2 \frac{1}{p(x)}} = \mathbb{E}\left\{ \log_2 \frac{1}{p(X)} \right\}$$

is a positive quantity called entropy.

Shannon's entropy

$$H = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

- analogy with statistical mechanics

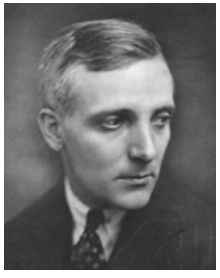
Ludwig Boltzmann (1844–1906)

- suggested by

"You should call it entropy [...] no one really knows what entropy really is, so in a debate you will always have the advantage."

John von Neumann (1903–1957)

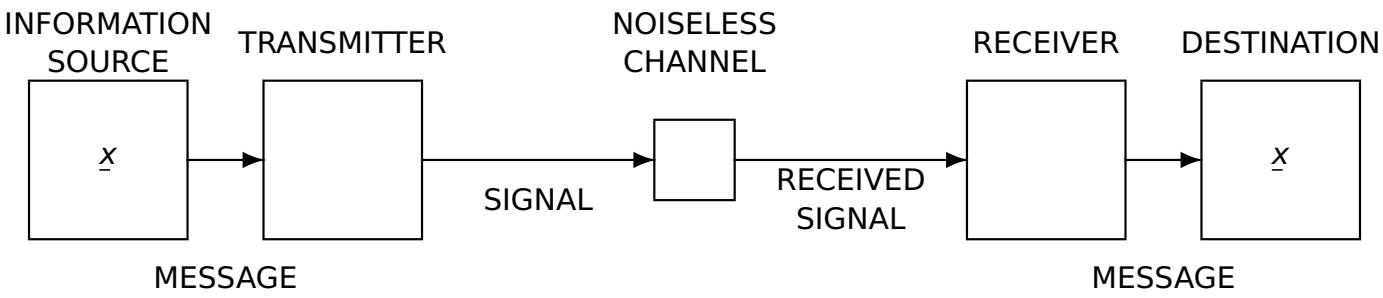
- studied in physics by



Léon Brillouin (1889–1969)

The Source Coding Theorem

Compression problem: noiseless channel, minimize bit rate



A "typical" sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ satisfies $p(\underline{x}) \approx 2^{-nH}$.
Summing over the N typical sequences:

$$1 \approx N\, 2^{-nH}$$

since the probability of $\underline{x}$ being typical is ≈ 1. So $N \approx 2^{nH}$.
It is sufficient to encode only the N typical sequences:

$$\frac{log_2 N}{n} \approx H \quad \text{bits per symbol}$$

The Source Coding Theorem

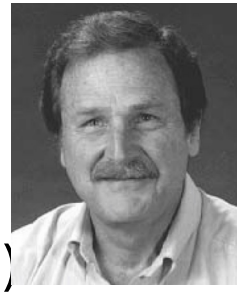
Theorem (Shannon's First Theorem)

Only H bits per symbol suffice to reliably encode an information source.

The entropy H is the bit rate lower bound for reliable compression.

- This is an asymptotic theorem ($n \to +\infty$) not a practical solution.
- Variable length coding solution by Shannon and

Robert Fano (1917–2016)



- Optimal code (1952) by David Huffman (1925-1999)
- Elias, Golomb, Lempel-Ziv, ...

Back to Shannon's Proof

What is the probability that a sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ is q-typical?

$$q(\underline{x}) = q(x_1) \cdot q(x_2) \cdots q(x_n) = \prod_x q(x)^{n(x)} \approx \prod_x q(x)^{np(x)} = 2^{-n \cdot H(p,q)}$$

where $H(p,q) = \sum_x p(x) \log_2 \frac{1}{q(x)}$ is a "cross-entropy".

Thus the probability that the sequence is q-typical is $N \cdot 2^{-n \cdot H(p,q)}$.

Replacing q by p, we would have $N \cdot 2^{-nH(p,p)} = N \cdot 2^{-nH(p)} \leq 1$ (a probability).

Therefore the probability that the sequence is q-typical is bounded by

$$2^{n \cdot (H(p) - H(p,q))} = 2^{-n \cdot D(p,q)}$$

where $D(p,q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$ (relative entropy aka divergence)

Relative Entropy (or Divergence)

$$D(p,q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \geq 0 \text{ with } D(p,q) = 0 \text{ iff } p \equiv q.$$

Bounds of the type $2^{-n \cdot D(p,q)}$ useful in statistics:

- large deviations theory
- asymptotic behavior in hypothesis testing

Chernoff information to classify empirical data

Herman Chernoff (1923–)

- Fisher information for parameter estimation

Ronald Fisher (1890–1962)

Shannon's Mutual Information

Shannon's entropy of a random variable X:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = \mathbb{E}\left\{ \log_2 \frac{1}{p(X)} \right\}$$

Shannon's (mutual) information between two random variables X, Y:

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = \mathbb{E}\left\{ \log_2 \frac{p(X,Y)}{p(X)p(Y)} \right\}$$

This exactly $D(p,q)$ where:

- $p(x,y)$ is the (true) joint distribution;
- $q(x,y) = p(x)p(y)$ is what would have been in the case of *independence*.

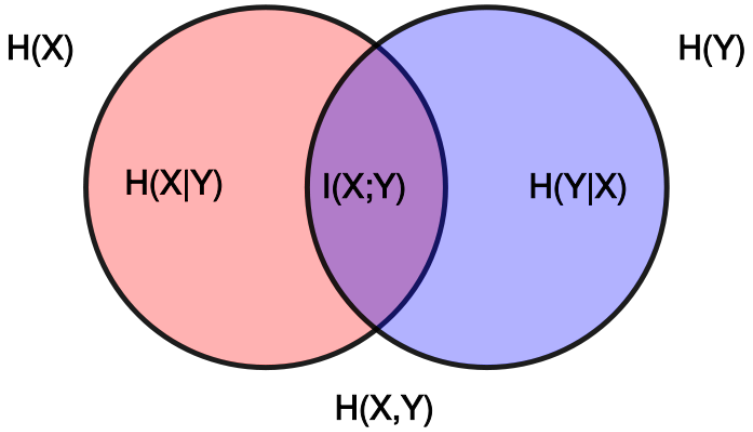
Therefore $I(X;Y) \geq 0$ with $I(X;Y) = 0$ iff X and Y are independent

Shannon's Mutual Information

Shannon writes

$$I(X;Y) = \mathbb{E}\left\{\log_2 \frac{p(X|Y)}{p(X)}\right\} = H(X) - H(X|Y)$$

where $H(X|Y)$ is the conditional entropy of X given Y.



- $H(X|Y) \leq H(X)$: *knowledge decreases uncertainty* by a quantity equal to the information gain $I(X;Y)$.
- intuitive and rigorous!

The Set of All Possible Codes

A *channel* with input $\underline{x} = (x_1, x_2, \ldots, x_n)$ (the channel *code*) and output $\underline{y} = (y_1, y_2, \ldots, y_n)$ is characterized by the conditional distribution $p(\underline{y}|\underline{x}) = p(y_1|x_1) \cdot p(y_2|x_2) \cdot \cdots \cdot p(y_n|x_n)$. (memoryless case).

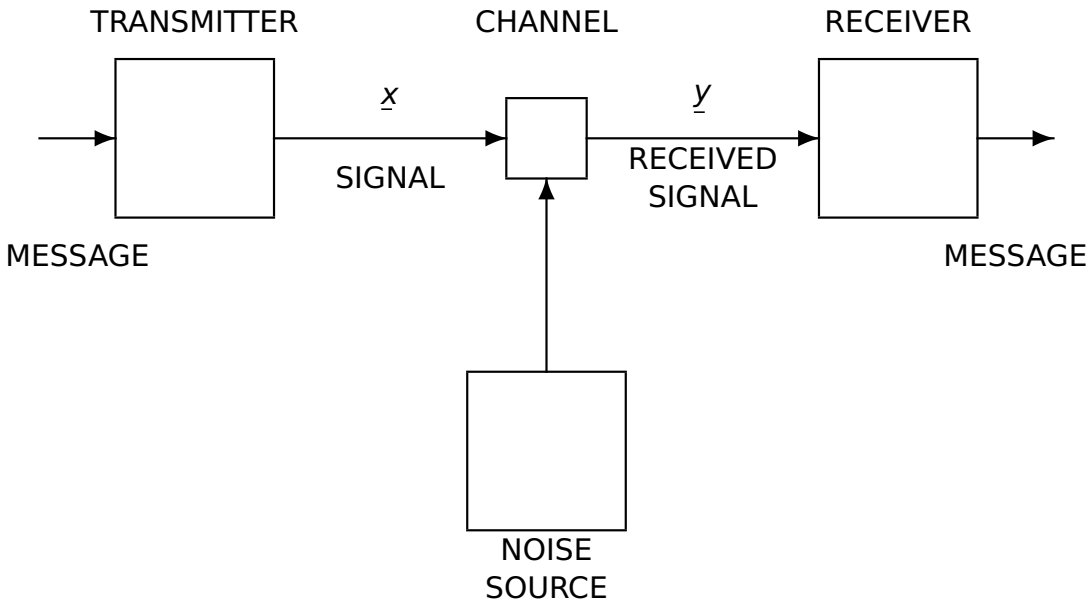
Shannon considers all possible codes **as if** each $\underline{x}$ were chosen according to a probability distribution

$$p(\underline{x}) = p(x_1) \cdot p(x_2) \cdot \cdots \cdot p(x_n). \qquad \textbf{(random coding!)}$$

- $\underline{x}$ is jointly typical with $\underline{y}$ if $p(\underline{x}, \underline{y}) \approx 2^{-n \cdot H(X,Y)}$;
- but another (independent) code has $q(\underline{x}, \underline{y}) = p(\underline{x})p(\underline{y})$;
- thus the probability that it is also jointly typical with $\underline{y}$ is $\leq 2^{-n \cdot D(p,q)} = \boxed{2^{-n \cdot I(X;Y)}}$.

The Channel Coding Theorem

Transmission problem: noisy channel, maximize bit rate for reliable communication



It is sufficient to decode only sequences $\underline{x}$ jointly typical with $\underline{y}$.

The Channel Coding Theorem (Cont'd)

But another code is also jointly typical with $\underline{y}$ with probability bounded by

$$2^{-n \cdot I(X;Y)}.$$

Summing over the N code sequences, the total probability of decoding error is bounded by

$$N \cdot 2^{-n \cdot I(X;Y)}$$

which tends to zero only if the bit rate

$$\frac{\log_2 N}{n} < I(X;Y)$$

Definition (Channel Capacity)

$$C = \max_{p(x)} I(X;Y)$$

The Channel Coding Theorem (Cont'd)

If the bit rate is $< C$, then the error probability, averaged over all possible codes, can be made as small as desired.

Therefore there exists at least one code with arbitrarily small probability of error.

Theorem (Shannon's Second Theorem)

Information can be transmitted reliably provided that the bit rate does not exceed the channel capacity C.

The capacity C is the bit rate upper bound for reliable transmission.

Revolutionary! Transmission noise does not affect quality—it only impacts the bit rate.
This is the theorem that led to the digital revolution!

Shannon's Result is Paradoxical!

- Shannon theorems show that good codes exist, but give no clue on how to build them in practice
- but choosing a code at random would be almost optimal!
- however random coding is impractical (n is large)...
- only 50 years later were found *turbo-codes* (by Claude Berrou & Alain Glavieux) that imitate random coding to approach capacity

Additive White Gaussian Noise Channel

A very common model: $Y = X + Z$ where Z is Gaussian $\mathcal{N}(0, \sigma^2)$.

Shannon finds the exact expression:

$$\boxed{C = W \cdot \log_2\left(1 + \frac{P}{N}\right)} \quad \text{bit/s}$$

where W is the bandwidth and P/N is the signal-to-noise ratio.

- a "concrete" finding of information theory – the most celebrated formula of Shannon!
- to derive this formula, Shannon popularized the Whittaker-Nyquist sampling theorem — "Shannon's Theorem"!

Claude Shannon

Shannon's formula:

$$C = W \log_2 \left(\frac{P+N}{N} \right)$$

"A Mathematical Theory of Communication," *The Bell System Technical Journal*, Vol. 27, pp. 623–656, October, 1948 .

In the end, "The Mathematical Theory of Communication," [1] and the book based on it [25] came as a bomb, and something of a delayed-action bomb.

Note on the Theoretical Efficiency of Information Reception with PPM*

For small P/N ratios, the now classical expression for the information reception ca-

And then there were eight

Quote from Shannon, 1948:

> Formulas similar to $C = W \log \frac{P+N}{N}$ for the white noise case have been developed independently by several other writers, although with somewhat different interpretations. We may mention the work of N. Wiener,[7] W. G. Tuller,[8] and H. Sullivan in this connection.

1. Norbert Wiener, *Cybernetics*, 1948
2. William G. Tuller, PhD Thesis, June 1948
3. Herbert Sullivan (unpublished, 1948)
4. Jacques Laplume, April 1948
5. Charles W. Earp, June 1948
6. André G. Clavier, December 1948
7. Stanford Goldman, May 1948
8. Claude E. Shannon, Oct. 1948

What about the French?

Deux ingénieurs français ont publié la même « formule de Shannon » en 1948:

Clavier & Laplume

André G. Clavier

Evaluation of Transmission Efficiency According to Hartley's Expression of Information Content*

By A. G. CLAVIER

Federal Telecommunication Laboratories, Incorporated, Nutley, New Jersey

small percentage of error due to noise. The total number of distinguishable levels on the ideal line is thus given by

$$\frac{S+\bar{N}\sqrt{2}}{\bar{N}\sqrt{2}}=1+\frac{S}{\bar{N}\sqrt{2}},$$

with a reasonable approximation. It follows that the amount of information transmittible on the ideal line is measured by

$$H_{lm}=k_0\cdot 2f_l\cdot t\cdot\log\left(1+\frac{S_l}{\bar{N}_l\sqrt{2}}\right).$$

* A symposium on "Recent Advance
Communication" was presented at the N
meeting of the New York Section of the
Engineers. Four papers were presented
Federal Telecommunication Laboratori
Hazeltine Electronics Corporat
C. E. Shannon, both of Bell Te

Jacques Laplume

Meanwhile (1948), far away. . .

PHYSIQUE MATHÉMATIQUE. — *Sur le nombre de signaux discernables en présence du bruit erratique dans un système de transmission à bande passante limitée.* Note de M. JACQUES LAPLUME.

Si N et n sont suffisamment grands, on peut former une expression approchée de $\log M$ en utilisant la formule de Stirling limitée aux termes prépondérants. On trouve ainsi

$$\log M \# N \log \frac{N+n}{N} + n \log \frac{N+n}{n}. \qquad (2)$$

Si, de plus, $N \geqslant n$,

$$\log M \# n \log \frac{N}{n} = TW \log \frac{P}{b}. \qquad (3)$$

More on Jacques Laplume...

INSTITUT DE FRANCE
Académie des sciences

Histoire des sciences / Évolution des disciplines et histoire des découvertes — Octobre 2016

Laplume, sous le masque

par Patrick Flandrin (directeur de recherche CNRS à l'École normale supérieure de Lyon, membre de l'Académie des sciences) et Olivier Rioul (professeur à Télécom-ParisTech et professeur chargé de cours à l'École Polytechnique)

Cette note vise à faire sortir de l'oubli un travail original de 1948 de l'ingénieur français Jacques Laplume, relatif au calcul de la capacité d'un canal bruité de bande passante donnée. La publication de sa Note dans les Comptes Rendus de l'Académie des sciences a précédé de peu celle de l'article du mathématicien américain Claude E. Shannon, fondateur de la théorie de l'information, ainsi que celles de plusieurs chercheurs aux U.S.A.

TELECOM ParisTech

Who's formula?

The "Shannon" formula

$$\boxed{C = W \log_2\left(1 + \frac{P}{N}\right)}$$

should actually be the

Shannon-Laplume-Tuller-Wiener-Clavier-Earp-Goldman-Sullivan formula

Derivation: Capacity of the AWGN Channel

For a continuous r.v. X: Differential Entropy

$$h(X) = \mathbb{E}\left(\log \frac{1}{p(X)}\right) = \int p(x) \log \frac{1}{p(x)} \, \mathrm{d}x$$

Lemma (MaxEnt)

$h(X) \leq \frac{1}{2} \log(2\pi e P)$ *with equality iff* $X \sim \mathcal{N}(0, P)$.

Proof.

Information inequality $D(p\|q) \geq 0$ where $q \sim \mathcal{N}(0, P)$. □

Derivation: Capacity of the AWGN Channel

$$\boxed{C = \frac{1}{2}\log_2\left(1+\frac{P}{N}\right)} \quad \text{bits/sample}$$

Proof.

$C = \max I(X;Y)$ where $Y = X + Z$ and Z is Gaussian with power N:

$$\max I(X;Y) = \max h(Y) - h(Z) = \max h(Y) - \frac{1}{2}\log(2\pi eN)$$

$$\max h(Y) = \max h(X+Z) = h(X^* + Z) = \frac{1}{2}\log(2\pi e(P+N))$$

hence $C = \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi eN)$.

Entropy Power

Definition (Entropy Power)

Let X have power P. The *entropy-power* of X is the power P^* of a white Gaussian X^* having the same entropy:

$$h(X) = h(X^*) = \frac{1}{2}\log(2\pi e P^*) P^* = \frac{exp\big(2h(X)\big)}{2\pi e}$$ (which is e to the power 2

By MaxEnt, $P^* \leq P$ with equality iff X is white Gaussian.

Theorem (EPI as stated by Shannon, 1948)

For any independent $X, Y \in L^2$,

$$\boxed{P_X^* + P_Y^* \leq P_{X+Y}^*} \leq P_{X+Y} = P_X + P_Y$$

Application: nonGaussian Capacity

$Y = X + Z$ where Z is of power P with power constraint $\mathbb{E}(X^2) \leq P$

$$\max I(X;Y) = \max h(Y) - h(Z) = \max h(Y) - \frac{1}{2}\log(2\pi e N^*)$$

but for $X^* \sim \mathcal{N}(0,P)$, $\max h(Y) \geq h(X^* + Z) \geq \frac{1}{2}\log(2\pi e(P+N^*))$ (EPI)

Theorem (Shannon lower bound, 1948)

$$C \geq W \log\left(1 + \frac{P}{N^*}\right)$$

with equality iff the channel is Gaussian.

- Gaussian means worst noise / Gaussian means best signal

The Entropy-Power Inequality (EPI)

Differential entropy of a random vector with density p:

$$\boxed{h(X) = \int p(x) \ln\left(\frac{1}{p(x)}\right) \mathrm{x}}$$

For any two X, Y independent continuous random variables,

$$\boxed{P^*_{X+Y} \geq P^*_X + P^*_Y} \quad \boxed{e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}}.$$

Equality holds iff X, Y are *Gaussian*.

The EPI has a Long History

1948 Stated and “proved” by Shannon in his seminal paper
1959 Stam’s proof using Fisher information
1965 Blachman’s exposition of Stam’s proof in IEEE Trans. IT
1978 Lieb’s proof using strengthened Young’s inequality
1991 Dembo-Cover-Thomas’ review of Stam’s & Lieb’s proofs
1991 Carlen-Soffer 1D variation of Stam’s proof
2000 Szarek-Voiculescu variant with Brunn-Minkowski inequality
2006 Guo-Shamai-Verdú proof based on the I-MMSE relation
2007 Rioul’s proof based on Mutual Information
2014 Wang-Madiman strengthening using Rényi entropies
2016 Courtade’s strengthening
2017 Yet another simple proof

A simple change of variables

Lemma (inverse function sampling method)

If U is uniform in $[0,1]$ and X has c.d.f. $F(x) = \mathbb{P}(X \leq x)$, then $F^{-1}(U)$ has the same distribution as X.

Proof.

$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$ □

Corollary (monotonic increasing transport $T = F^{-1} \circ G$)

Let F, G be two c.d.f's. Then $X^* \sim G \implies X = T(X^*) \sim F.$

Proof.

$U = G(X^*) \sim$ uniform; $\quad T(X^*) = F^{-1}\big(G(X^*)\big) = F^{-1}(U) \sim F.$

A simple change of variables: Entropy

Lemma (Change of variable [Shannon'48])

For any continuous X, X^*, monotonic increasing transport $T(X^*) \sim X$,

$$h(X) = \boxed{h\big(T(X^*)\big) = h(X^*) + \mathbb{E}\log T'(X^*)}$$

Proof.

Proof: make the change of variable $x = T(x^*)$ in

$$h(X) = \int f_X(x) \log \frac{1}{f_X(x)} \mathrm{x} = \int \underbrace{f_X\big(T(x^*)\big)T'(x^*)}_{f_{X^*}(x^*)} \log \frac{1}{f_X\big(T(x^*)\big)} \mathrm{x}^*$$

□

- in particular $h(aX) = h(X) + \log|a| \iff P^*_{aX} = a^2 P^*_X$;
- more generally in nD: $h\big(T(X^*)\big) = h(X^*) + \mathbb{E}\log\big|\det T'(X^*)\big|$

A Proof that Shannon Missed

Proceed to prove the EPI:

$$\boxed{P^*_{X+Y} \geq P^*_X + P^*_Y} = P_{X^*} + P_{Y^*} = P_{X^*+Y^*} P^*_{X^*+Y^*}$$

1. Let X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X)$ and $h(Y) = h(Y^*)$, i.e., $P^*_X = P_{X^*}$ and $P^*_Y = P_{Y^*}$.

   One is led to prove $\boxed{P^*_{X+Y} \geq P^*_{X^*+Y^*}}$ $\boxed{h(X+Y) \geq h(X^*+Y^*)}$

2. Scaling $a, b \in \mathbb{R}$: $\boxed{h(aX+bY) \geq h(aX^*+bY^*)}$

3. We may assume $h(X) = h(Y) = h(X^*) = h(Y^*)$
   Otherwise:
   - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
   - apply the above to cX and dY.

   So w.l.o.g. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.

A Proof that Shannon Missed

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$
where X^*, Y^* are i.i.d. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

4. We may always normalize: $\boxed{a^2 + b^2 = 1}$. Otherwise:
   - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
   - the $\log \Delta$ terms cancel.
5. Make the changes of variables $X = T(X^*)$, $Y = U(Y^*)$:
   One is led to prove $\boxed{h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)}$
6. Define $\tilde{X} = aX^* + bY^*$. Complete the rotation: $\tilde{Y} = -bX^* + aY^*$ so that $\boxed{\tilde{X}, \tilde{Y} \text{ are i.i.d.}}$ Gaussian and $\boxed{X^* = a\tilde{X} - b\tilde{Y}}$, $\boxed{Y^* = b\tilde{X} + a\tilde{Y}}$.

A Proof that Shannon Missed

One is led to prove $\boxed{h(aT(X^*)+bU(Y^*)) \geq h(aX^*+bY^*)}$

$\boxed{\tilde{X},\tilde{Y} \text{ are i.i.d.}}$ Gaussian and $\boxed{X^* = a\tilde{X} - b\tilde{Y}}$, $\boxed{Y^* = b\tilde{X} + a\tilde{Y}}$.

7. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*)+bU(Y^*)) &= h(aT(a\tilde{X}-b\tilde{Y})+bU(b\tilde{X}+a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X}-b\tilde{Y})+bU(b\tilde{X}+a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})}|\tilde{Y}) \end{aligned}$$

8. By the change of variable:

$$\begin{aligned} &= h(\tilde{X}|\tilde{Y}) + \mathbb{E}\log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E}\log\big(a^2T'(a\tilde{X}-b\tilde{Y})+b^2U'(b\tilde{X}+a\tilde{Y})\big) \\ &= h(aX^*+bY^*) + \mathbb{E}\log\big(a^2T'(X^*)+b^2U'(Y^*)\big) \end{aligned}$$

9. By concavity of the log:

$$\geq h(aX^*+bY^*) + a^2\underbrace{\mathbb{E}\log T'(X^*)}_{h(X)-h(X^*)=0} + b^2\underbrace{\mathbb{E}\log}_{h(Y)-}$$

Equality Case

For nonzero a, b:

- in log concavity inequality:

$$\mathbb{E}\log\left(a^2T'(X^*)+b^2U'(Y^*)\right) = a^2\mathbb{E}\log T'(X^*) + b^2\mathbb{E}\log U'(Y^*)$$

$\Longrightarrow$ $T'(X^*) = U'(X^*) = c > 0$ constant a.e.

$\Longrightarrow$ T, U are linear: $X = T(X^*) = cX^*$, $Y = U(Y^*) = cY^*$ Gaussian.

$\Longrightarrow$ $c = 1$ since $h(X) = h(X^*)$, $h(Y) = h(Y^*)$.

- in information inequality:

$$h(aT(a\tilde{X}-b\tilde{Y})+bU(b\tilde{X}+a\tilde{Y})) = h(aT(a\tilde{X}-b\tilde{Y})+bU(b\tilde{X}+a\tilde{Y})|\tilde{Y})$$

comes for free since $a(a\tilde{X}-b\tilde{Y})+b(b\tilde{X}+a\tilde{Y}) = \tilde{X}$ is indep of $\tilde{Y}$.

Shannon on Information Theory

“I didn’t think at the first stages that it was going to have a great deal of impact. I enjoyed working on this kind of a problem, as I have enjoyed working on many other problems, without any notion of either financial or gain in the sense of being famous; and I think indeed that most scientists are oriented that way, that they are working because they like the game.”