

Theorizing Information



Concepts of Information i218
Geoff Nunberg

Jan. 29, 2015

Grocis pondenome!



Itinerary 1/29

Background: what is the "information" of "information theory"

Basic concepts: measuring information

Probability and redundancy

Applications

"The Bandwagon"



Information as "Informativeness": Two Theories

Two "producers" of theories of information

"Information theory" ("mathematical theory of communication").

Some applications of MTC

Philosophical accounts of (semantic) information

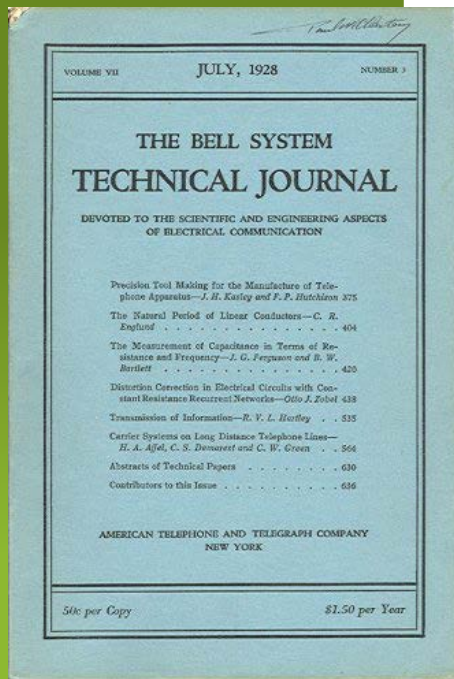


Predecessors

information is a very elastic term, and it will be necessary to set up a more specific meaning ...

There must be a group of symbols...the sender mentally selects a symbol and causes the attention of the receiver to be directed ... At each selection there are eliminated all of the other symbols which might have been chosen ... more and more possible symbols sequences are eliminated ... the information becomes more precise ... *Inasmuch as the precision of the information depends upon what other symbol sequences might have been chosen ...* reasonable to hope to find in the number of these sequences the desired quantitative measure... [we desire to] to eliminate the psychological factors involved and to establish a measure of information in terms of purely physical quantities.

R. V. Hartley "The Measurement of Information," 1928





Narrowing the problem down

"The word communication will be used here in a very broad sense to include all of the procedures by which one mind may affect another. This, of course, involves not only written and oral speech, but also music, the pictorial arts, the theater, the ballet, and in fact all human behavior."
Weaver, *The Math. Theory of Communication*



Narrowing the problem down

"Relative to the broad subject of communication, there seem to be problems at three levels. Thus it seems reasonable to ask, serially:

LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem.)

LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)

LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Cf syntax/semantics/pragmatics; form/meaning/use...



The Mathematical Theory of Communication



Rarely does it happen in mathematics that a new discipline achieves the character of a mature developed scientific theory in the first investigation devoted to it... So it was with information theory after the work of Shannon.

A. I. Khintchin, 1956



Elements of the Theory

"The fundamental problem of communication is that of reproducing at one point a message selected at another point."

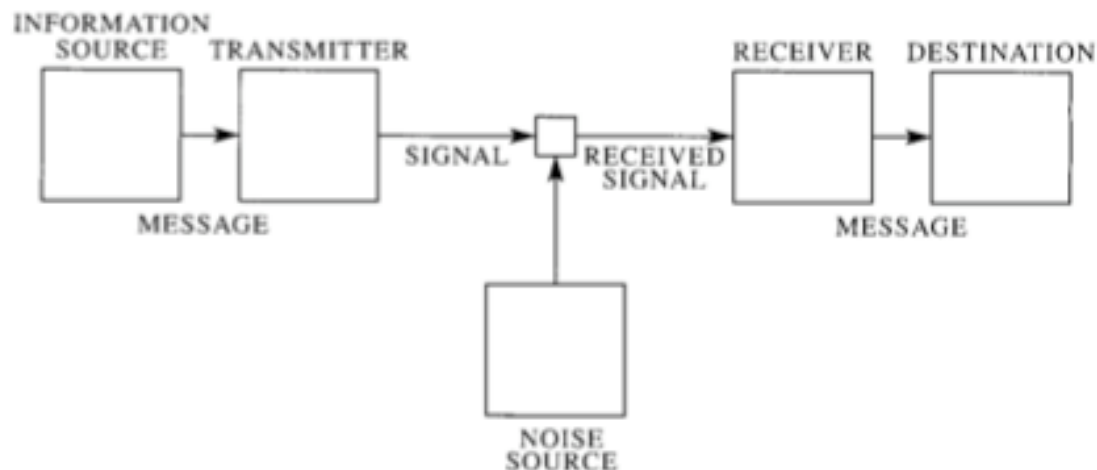


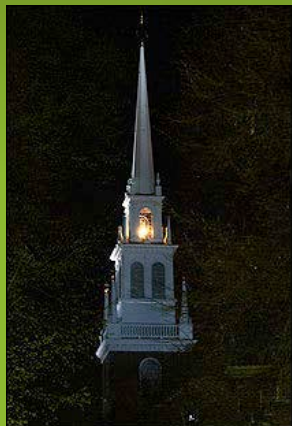
Fig. 1—Schematic diagram of a general communication system.



Varieties of Signals

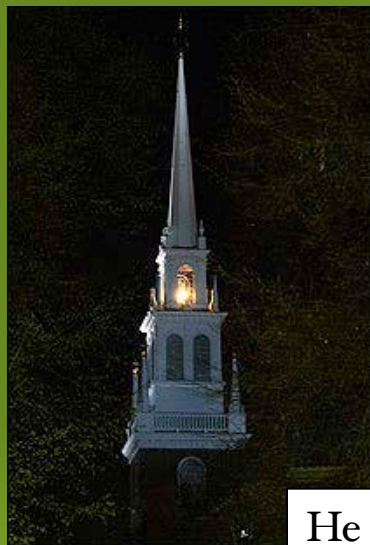


Form of signal and channel is irrelevant; matters only if signal is discrete or continuous.





Varieties of Signals



The British are coming to take your guns away!



He said to his friend, "If the British march
By land or sea from the town to-night,
Hang a lantern aloft in the belfry arch
Of the North Church tower as a signal light,
--One if by land, and two if by sea;
And I on the opposite shore will be,
Ready to ride and spread the alarm
Through every Middlesex village and farm...



Reducing Uncertainty

Information is that which reduces uncertainty in a choice situation.

Weaver: "... in this new theory the word information relates not so much to what you *do* say as to what you *could* say. That is, information is a matter of your freedom of choice when you select a message."



Anti-semantics

"The fundamental problem of communication is that of reproducing at one point 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." Shannon, 1948

I.e., "Communication" ends when it is determined **which** message was sent.



Anti-semantics

First off, we have to be clear about the rather strange way in which, in this theory, the word "information" is used; for it has a special sense which... must not be confused at all with meaning. It is surprising but true that, from the present viewpoint, two messages, one heavily loaded with meaning and the other pure nonsense, can be equivalent as regards information.

Warren Weaver, 1949



A misnomer?



I didn't like the term Information Theory. Claude didn't like it either. You see, the term 'information theory' suggests that it is a theory about information – but it's not. It's the transmission of information, not information. Lots of people just didn't understand this... I coined the term 'mutual information' to avoid such nonsense: making the point that information is always about something. It is information provided by something, about something.

Roberto Fano, 2001



Reducing Uncertainty

Hartley: Information content is proportional to the number of symbols, size of the set of possibilities.

$$H = n \log s$$

Where H = amount of information, n = number of symbols transmitted, and s = size of the "alphabet" (dot-dash, alphabet, Chinese logographs, etc.)

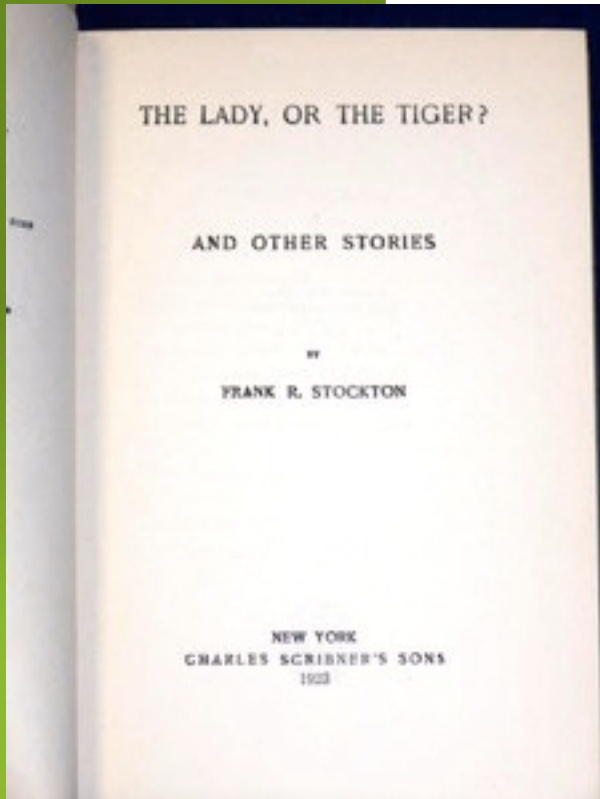
E.g., Information in all possible 3-letter strings =
 $3 * \log_2 (26) = 3*4.7 = 14.1$ bits

Why " $\log s$ "?



A simple instance

In the very olden time there lived a semi-barbaric king, whose ideas, though somewhat polished and sharpened by the progressiveness of distant Latin neighbors, were still large, florid, and untrammelled, as became the half of him which was barbaric...
"The Lady or the Tiger," Frank Stockton, 1882

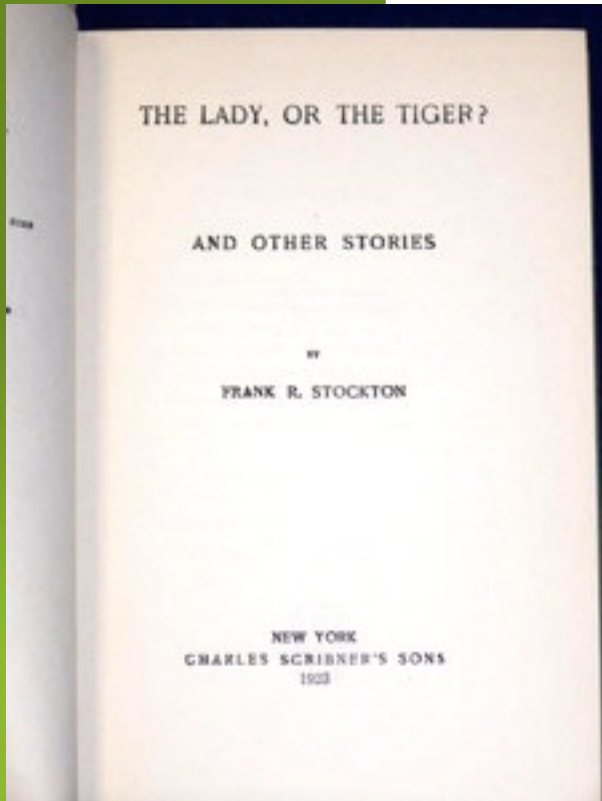




A simple instance

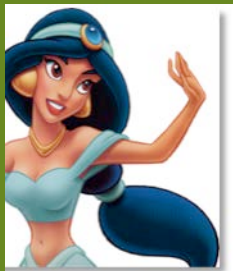
When a subject was accused of a crime of sufficient importance to interest the king, public notice was given that on an appointed day the fate of the accused person would be decided in the king's arena...When all the people had assembled in the galleries [the king] gave a signal, a door beneath him opened, and the accused subject stepped out into the amphitheater. Directly opposite him, on the other side of the enclosed space, were two doors, exactly alike and side by side. It was the duty and the privilege of the person on trial to walk directly to these doors and open one of them... If he opened the one, there came out of it a hungry tiger, the fiercest and most cruel that could be procured, which immediately sprang upon him and tore him to pieces as a punishment for his guilt.... But, if the accused person opened the other door, there came forth from it a lady, the most suitable to his years and station that his majesty could select among his fair subjects, and to this lady he was immediately married, as a reward of his innocence....

The institution was a very popular one.





A simple instance



"Did the tiger come out of that door, or did the lady?"
"The Lady and the Tiger," Frank Stockton, 1882

Arm up → right door
Arm down → left door





A simple instance



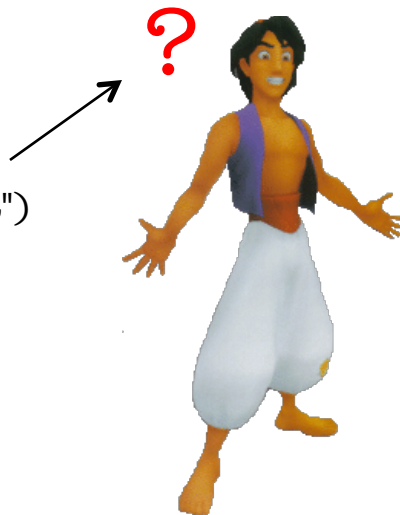
Arm up → right door
Arm down → left door

...the amount of information is defined, in the simplest cases, to be measured by the logarithm of the number of available choices. It being convenient to use logarithms to the base 2... This unit of information is called a 'bit' ... a condensation of 'binary digit'.

$$\text{Log}_2 2 = 1 \text{ (i.e., } 2^1 = 2\text{)}$$

Signal contains 1 bit of information ("informativeness," "surprisal")

Uncertainty
("data deficit")





Reducing Uncertainty

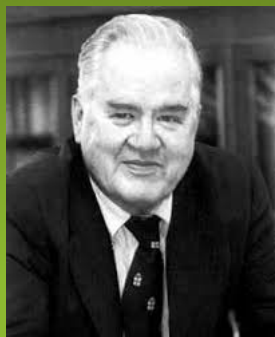


...the amount of information is defined, in the simplest cases, to be measured by the logarithm of the number of available choices. It being convenient to use logarithms to the base 2... This unit of information is called a 'bit' ... a condensation of 'binary digit'.

$$\text{Log}_2 2 = 1 \text{ (i.e., } 2^1 = 2)$$

Signal contains 1 bit of information

Arm up → right door
Arm down → left door



John Tukey

Less certainty





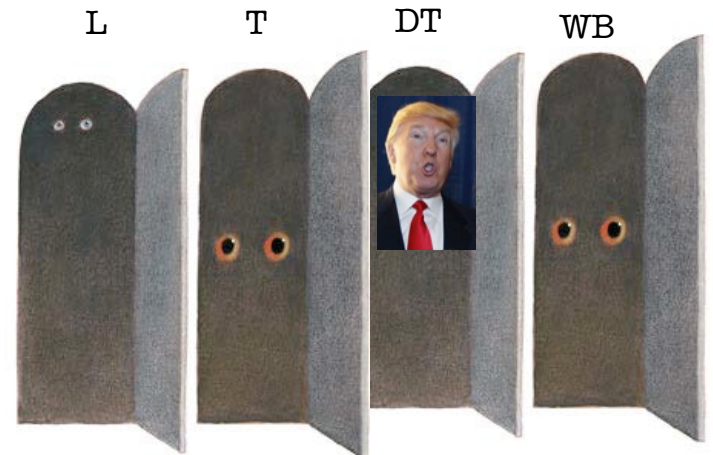
Reducing Uncertainty



"Did the tiger come out of that door, or did the lady, or did the wild boar, or did Donald Trump?"

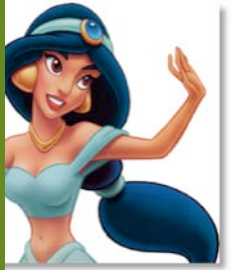
$\log_2 4 = 2 \rightarrow 2$ bits of information (i.e., $2^2 = 4$)

L down R up \rightarrow door 1
L up R down \rightarrow door 2
L down R down \rightarrow door 3
L up R up \rightarrow door 4





It's not about semantics



"The significant aspect is that the actual message is one selected from a set of possible messages."

Arm up → one signal
Arm down → another signal

Is her arm up or down?





The effects of probability

Signals are not equiprobable...



The effects of probability

Signals are not equiprobable...

_ I _ _ _ _ _ I _ S



The effects of probability

Signals are not equiprobable...

K _ X _ _ _ _ K _ _ _



The effects of probability

Signals are not equiprobable...

Freq. of initial letters of English words

T	15.2%
A	11.4%
H	8.5 %
W	7.0%
F	3.5 %

Knowing that a message begins with T reduces the set of possible messages by 84.8%

Knowing that a message begins with F reduces the set of possible messages by 96.5%

→ #F__ is more informative (“contains more information”) than #T_____



The effects of probability

What every "Wheel of Fortune" viewer knows:

_o_e_ _ a_e _e_ _i_ _o_ _a_ i_ e _ _a_
 o _o_ a_ _ _



The effects of probability

What every "Wheel of Fortune" viewer knows:

_o_e_ _ a_e _e_ _i_ _o_ _a_ i_e _ _a_

o _o_ a_ _ _

V_w_ls _r_ l_ss _nf_rm_t_v_ th_n

c_ns_n_nts



The effects of probability

What every "Wheel of Fortune" viewer knows:

_o_e_ _ a_e _e_ _i_ _o_ _a_i_e _ _a_

o _o_a_ _ _

V_w_ls _r_ l_ss _nf_rm_t_v_ th_n

c_ns_n_nts

How long would it take to finish a game if the answers were Polish surnames? Licence-plate numbers?



The effects of probability



Morse(?) - Vail Code: (roughly) fewer bits for most common letters:

A: .- B: -... C: -.-. D: -.. E: .
F: ..-. G: --. H: I: .. J: ---.
K: -.- L: -.-. M: -- N: -. O: ---
P: -.-. Q: --.- R: ._. S: ... T: -
U: .._ V: ...- W: .-- X: -.- Y: -.-. Z: ---.



<u>1-BIT</u>	<u>2-BITS</u>	<u>3-BITS</u>	<u>4-BITS</u>
E,T	I,A,N,M	S,U,R,W,D,O,G,K	H,V,F,L,P,J,B,X,C,Z,Q,Y



The effects of state-dependence (transitional probabilities)

Signals are prior-state-dependent (markoff process)

Overall frequency of $h = .06$

Overall frequency of $s = .06$

But probabilities are not the same after an initial t :

$$P(h) / \#t_ _ > P(s) / \#t_ _$$

i.e., a signal string beginning with $ts\dots$ is more informative than one beginning with $th\dots$

Which is more informative?

Let the cat out of the b____

Let the cat out of the p____



The effects of probability

Signals are not equiprobable...

...if we are concerned with English speech, and if the last symbol chosen is “the,” then the probability that the next word be an article, or a verb form other than a verbal, is very small. This probabilistic influence stretches over more than two words, in fact. Weaver, *Math. Theory of Comm.*



The effects of probability

How to accommodate probability?

Hartley's formulation for equiprobable symbols:

“What we have done is to take as our practical measure of information the logarithm of the number of possible symbol sequences.”

$$H = n \log s$$

Where $s = \#$ of possible symbols, $n = \#$ of symbols in transmission



Contrast info in letters & numbers)



The effects of probability

Shannon's formulation for stochastic processes:

$$H = -\sum p_i \log_2 p_i$$

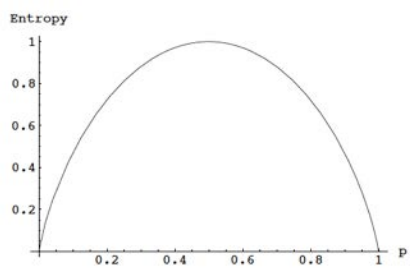
Uncertainty = the weighted sum of the (log of) improbabilities of messages.*

If average $p_i = 1$, $H = 0$

If average $p_i = .5$, H reaches maximum of 1 ($\log_2 .5 = -1$)

If the set of symbols/messages is larger, H increases.

Compare "20 questions"





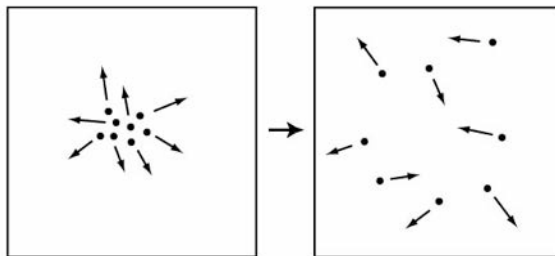
The entropy connection

The informativeness of a message varies inversely as its probability:

$$H = -\sum p_i \log_2 p_i$$

H is the "entropy" of the message.

The quantity which uniquely meets the natural requirements that one sets up for "information" turns out to be exactly that which is known in thermodynamics as *entropy*. [. . .] Thus when one meets the concept of entropy in communication theory, he has a right to be rather excited—a right to suspect that one has hold of something that may turn out to be basic and important. Weaver

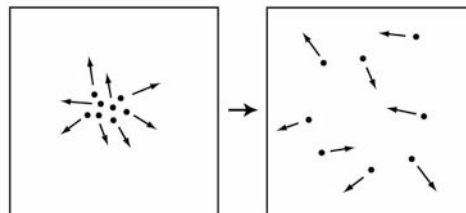




The entropy connection

Why the connection? Information and disorder

$$H = -\sum p_i \log_2 p_i$$



An informative desk





Consequences of low entropy

Low entropy = more predictability

Redundancy permits compression

abc _ _ _

dabchick



Consequences of low entropy



Redundancy permits compression

dabchick

These are the times that try men's souls.

Thes ar th time tha tr men' soul

The ar th tim th tr men' sou



Consequences of low entropy

Redundancy permits compression

dabchick





Consequences of low entropy

Redundancy permits compression

dabchick



These are the times that try men's souls.

Thes ar th time tha tr men' soul

The ar th tim th tr men' sou



Redundancy & pattern permits compression

Cf Dictionary compression: "small bits"

chromate

chromatic

chromatin

chromatogram

chromatograph

chromatography

chrome

chromic

chromium

chromosome (= 86 characters)



Redundancy & pattern permits compression

Cf Dictionary compression: "small bits"

chromate

chromatic

chromatin

chromatogram

chromatograph

chromatography

chrome

chromic

chromium

chromosome (86 characters)

→ chromate F7ic F8n F7ogram F1lph F1ly F5e F5ic F6um

F5osome (29 characters)



Consequences of low entropy

Redundancy facilitates error detection and signal recovery in noisy channels

Redundancy facilitates processing

I know that the boy will come.

I know the boy will come.

I know that she will come.

I know she will come.



Applications



Applications of Information Theory

An example: "Optimizing information density through syntactic reduction" (Roger Levy & Florian Jaeger)

Assumption (per Shannon): speakers structure utterances to minimize information density (amount of information per utterance unit). I.e. speakers try to spread out the surprisal.

"speakers structure their utterances in ways that buy them time to prepare difficult words and phrases."

Cf effects on contraction, speech rate, etc.

Phonetics: speakers lengthen syllables of less familiar words. *antimetabole* vs *antidepressant*...



Applications of Information Theory

An example: Optimizing information density through syntactic reduction (Roger Levy & Florian Jaeger)

Syntactic reduction:

(I) How big is [NP the family_i [RC (*that*) you cook for i]]?

Assume information density is higher when relativizer is omitted. (because then the 1st word of the rel. clause does double work.)

Then "full forms (overt relativizers) should be used more often when the information density of the RC *would be high if the relativizer were omitted.*"

1. I believe (that) that drug makes you sleepy.
2. I believe (that) this drug makes you sleepy.



Applications of Information Theory

An example: Optimizing information density through syntactic reduction (Roger Levy & Florian Jaeger)

Syntactic reduction:

(1) How big is [NP the family_i [RC (*that*) you cook for i]]?

Assume information density is higher when relativizer is omitted.

Then "full forms (overt relativizers) should be used more often when the information density of the RC *would be high if the relativizer were omitted.*"

1. I believe (that) that drug makes you sleepy.
2. I believe (that) this drug makes you sleepy.

Table 1 Rate of optional that (% OPT) before pronoun that and this in CCs

Subject	WSJ		BC		SWBD		Total	
	% OPT	Total	% OPT	Total	% OPT	Total	% OPT	Total
Pronoun <u>this</u>	26%	39	58%	19	18%	50	28%	108
Pronoun <u>that</u>	2%	41	33%	6	9%	675	9%	722
Fisher's Exact	p<0.01		n.s.		p<0.05		p<0.001	



Applications of Information Theory

Assume information density is higher when relativizer is omitted. Then "full forms (overt relativizers) should be used more often when the information density of the RC *would be high if the relativizer were omitted.*"

Contrast the effects of pronoun case...

I believe (that) she took the test.

I believe (that) the girl took the test.

... and NP complexity:

I believe (that) the student who failed the test has dropped the course.

I believe (that) the student has dropped the course.



Applications of Information Theory

Assume information density is higher when relativizer is omitted. Then "full forms (overt relativizers) should be used more often when the information density of the RC *would be high if the relativizer were omitted.*"

Contrast the effects of pronoun case...

✓ I believe (that) she took the test.

I believe (that) the girl took the test.

... and NP complexity:

I believe (that) the student who failed the test has dropped the course.

✓ I believe (that) the student has dropped the course.



Predicting word-length choices

"Speakers actively choose shorter word in predictable contexts" Mahowald et al. 2012

supportive-context: Susan was very bad at algebra, so she hated... 1. math 2. mathematics (67%)

neutral-context: Susan introduced herself to me as someone who loved... 1. math 2. mathematics (56%).



Development of MTC

Development of information theory in mathematics, engineering, computer science

e.g., work on compression, IR, etc.

Influence/cross-fertilization w/ biology & physics ("It from Bit")



The "information" explosion

The 50's moment: cybernetics, game theory, generative grammar..

Information theory also has influence in economics, psychology, linguistics, sociology and anthropology, sometimes lasting, sometimes temporary...

Information theory is alive and well in biology, engineering, physics, and statistics, although one rarely sees Shannon's information theory in contemporary psychology articles except to the extent of the late John W. Tukey's term *bit*, which is now a permanent word of our vocabulary. Duncan Luce, 2001

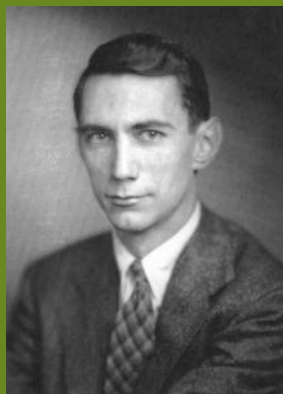
The mathematical theory of information ... is irrelevant [to computation] although computer programs are often said to be information-processing mechanisms. Aaron Sloman



"Information Please,"
NBC Radio



"The Bandwagon"

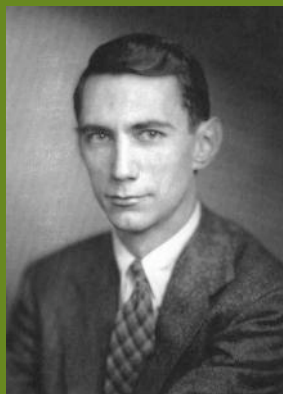


Information theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such 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 balloned to an importance beyond its actual accomplishments. ... Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. In short, information theory is currently partaking of a somewhat heady draught of general popularity...

Claude Shannon, 1956



"The Bandwagon"



Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. . . . It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like information, entropy, redundancy, do not solve all our problems.

Claude Shannon, 1956

The reaction:

“If you need to count higher than one, you’ve made a mistake.”
An MIT AI professor, 1972