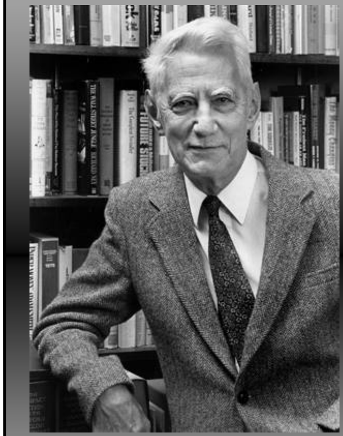
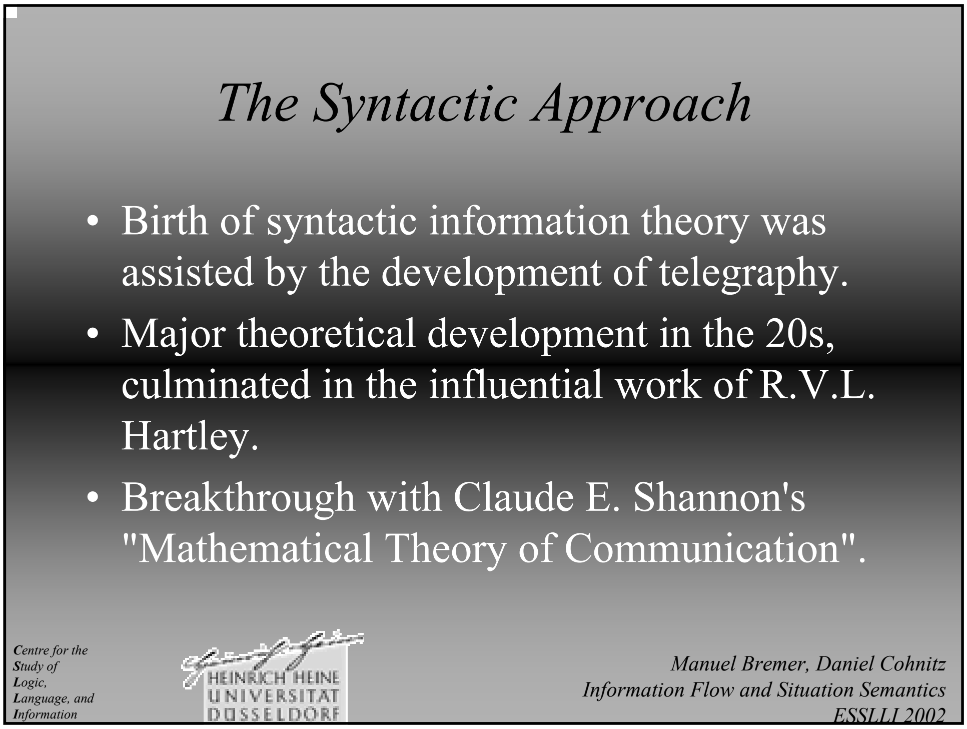
Situation Theory and the Flow of Information

Daniel Cohnitz / Manuel Bremer

CSLLI Düsseldorf

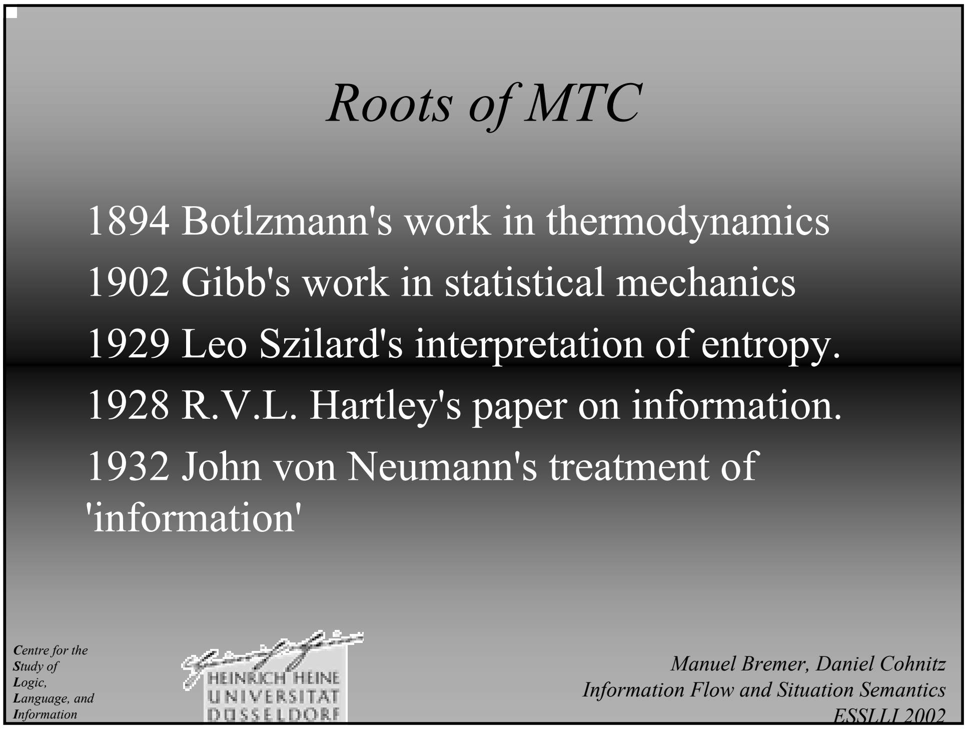


The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at some other point. Frequently the messages have *meaning*; that is they refer to or a correlated with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.



The Syntactic Approach

- Birth of syntactic information theory was assisted by the development of telegraphy.
- Major theoretical development in the 20s, culminated in the influential work of R.V.L. Hartley.
- Breakthrough with Claude E. Shannon's "Mathematical Theory of Communication".



Roots of MTC

1894 Botlzmann's work in thermodynamics

1902 Gibb's work in statistical mechanics

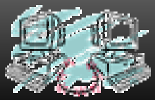
1929 Leo Szilard's interpretation of entropy.

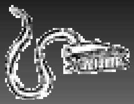
1928 R.V.L. Hartley's paper on information.

1932 John von Neumann's treatment of 'information'

The Shannon/Weaver Communication System I

The information ***source*** selects a desired message.





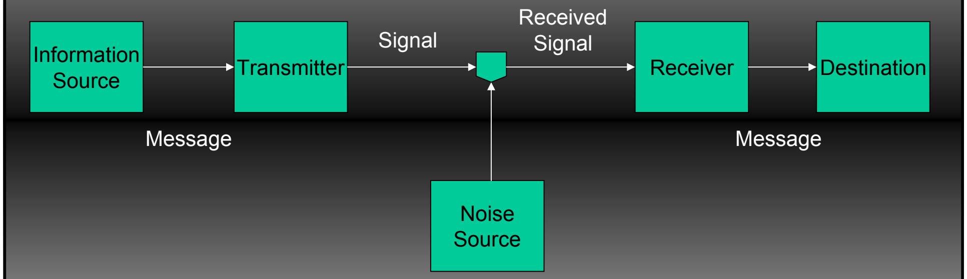
The signal is sent over the communication ***channel*** from the transmitter to the receiver.

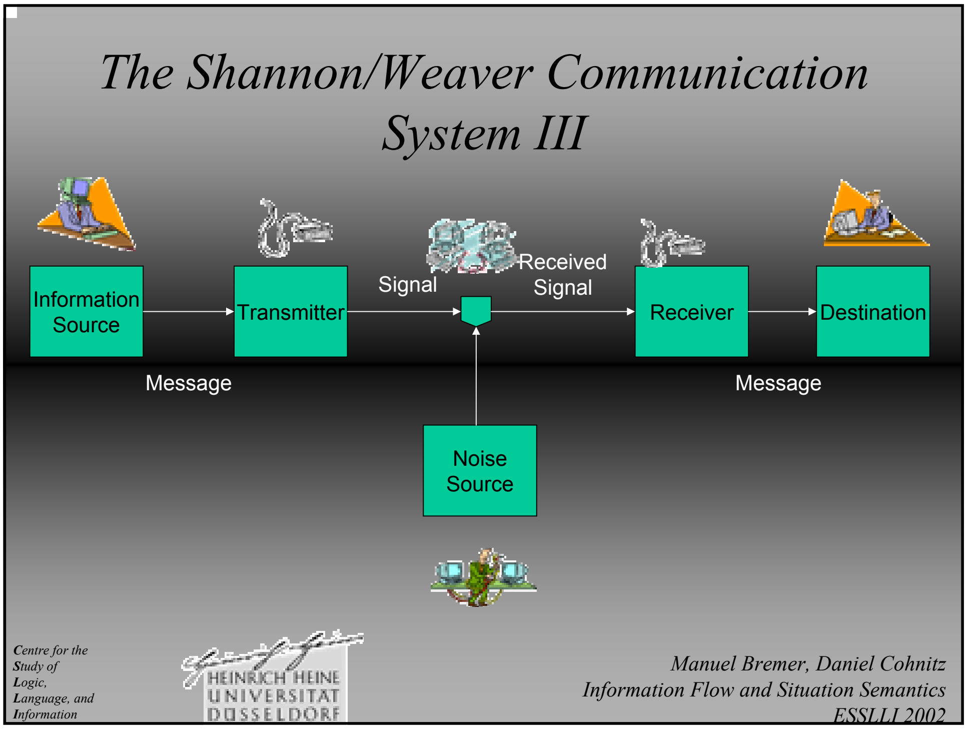
Eventually, the signal is changed back into a message by the receiver and handed on to the ***destination***.

The ***transmitter*** changes the message into the signal.

The distortions in the channel are called ***noise***. They change the transmitted signal.

The Shannon/Weaver Communication System II





The Shannon/Weaver Communication System III

Information Source

Transmitter

Signal

Received Signal

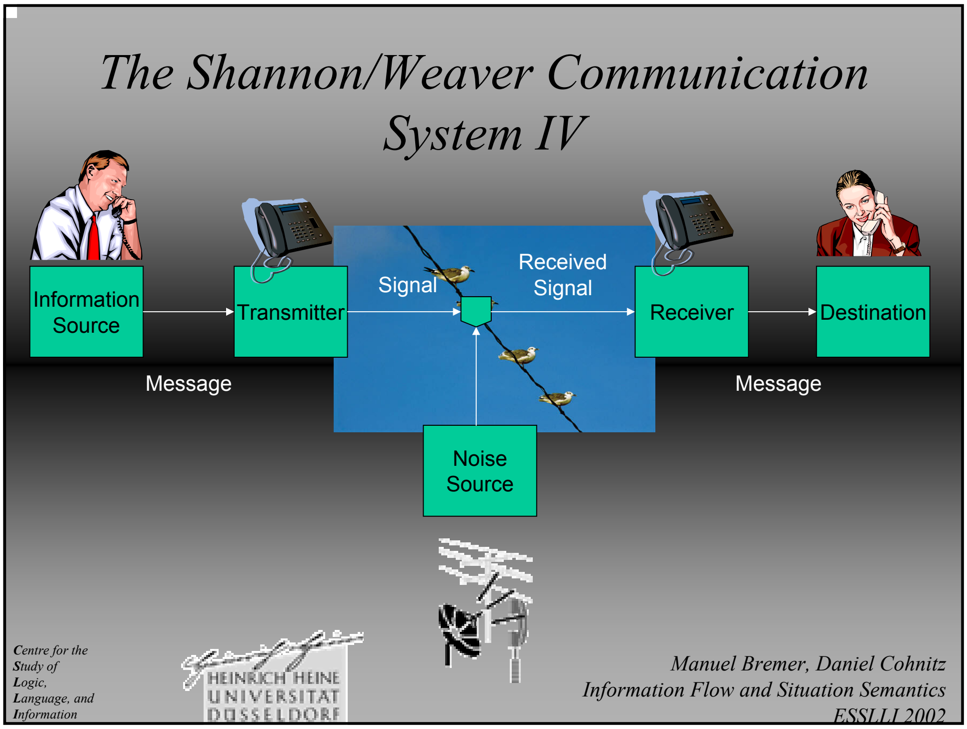
Receiver

Destination

Message

Message

Noise Source



The Shannon/Weaver Communication System IV

Information Source

Transmitter

Signal

Received Signal

Receiver

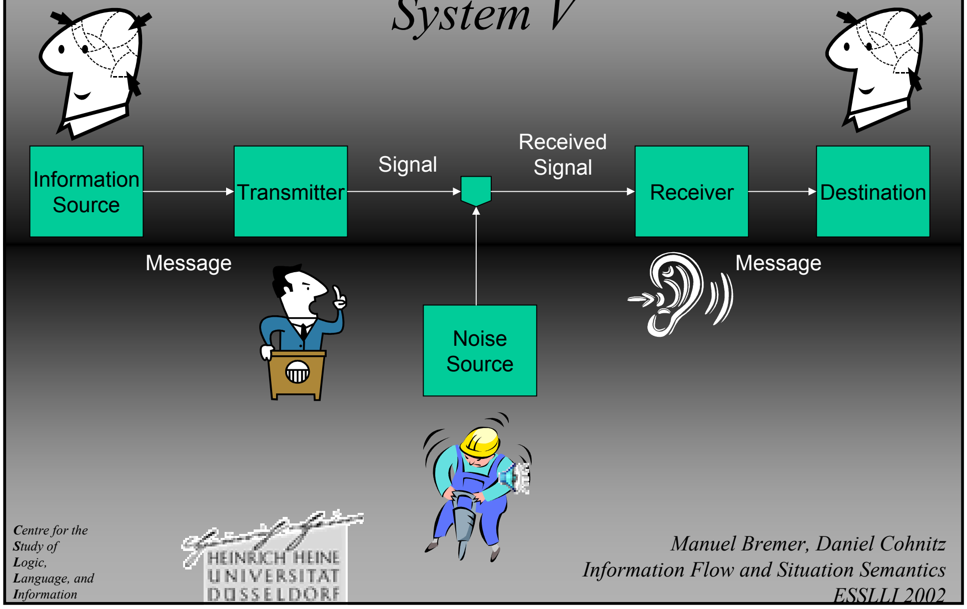
Destination

Message

Message

Noise Source

The Shannon/Weaver Communication System V



The Scope of Communication Theory I

1. Measures the amount of information (of situations as a whole).
2. Measures the capacity of the communication channel.
3. Determines at what rate a channel can convey information (given efficient coding).
4. Determines the eliminability of noise (given efficient coding).
5. Applies to discrete (written speech) and continuous (oral speech) communication.

Intuitions Covered

The mathematical theory of communication covers some of the features of our common sense concept 'information' which are intuitively quantitative:

(1) Information can be *encoded*.

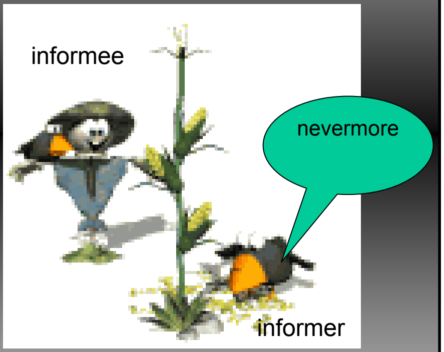
(2) Information is *additive*.

(3) Information is *non-negative*.

(4) Information *decreases uncertainty*.

Borderline cases of information sources

Consider Poe's raven who produces only one symbol, "nevermore". If informer and informee share the same background information about the collection of the usable symbols, it is obvious that a unary device like the raven produces zero amount of information.*



*Examples taken from Floridi forthcoming

Information and Uncertainty

Consider a system which is slightly more complex than our raven. Consider a binary device like a fair coin A, with its two equiprobable symbols {*h*, *t*}.

If we are the receiver, know the source, and wait for a symbol, we are uncertain as to which symbol the source will produce. We are in a state of *data deficite, the* "uncertainty" in Shannon's terms.

Once we receive a symbol, say 'h', our *uncertainty decreases*, and we remark that we have received some *information*. That is the connection between information and uncertainty. Now, how can information be measured?

Information and its Measurement

Once the coin has been tossed, the system produces an amount of *raw information* that is a function of the possible outputs, in this case 2 equiprobable symbols, and equal to the uncertainty it removes.

Let us now consider a more sophisticated source, say a complex system, made of two fair coins *A* and *B*. The *AB* system has the capacity to produce 4 different outputs:

<*h*, *h*> , <*h*, *t*>, <*t*, *h*>, <*t*, *t*>.

This source generates a data deficit of 4 units, each couple counting as a symbol in the source alphabet.

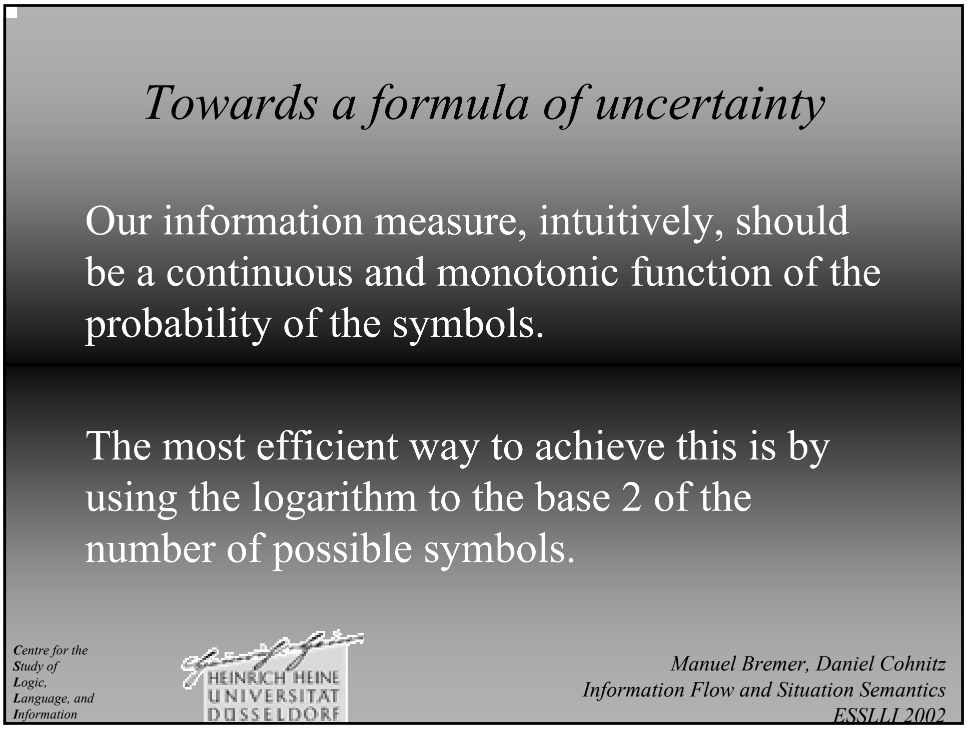
In this more complex system, the occurrence of each symbol contains more raw information than the occurrence in system A did. Adding an extra coin would produce a 8 units of uncertainty, further increasing the amount of information carried by each symbol in the *ABC* system.

Towards a formula of uncertainty

From our simple examples we can start to generalize. Let the number of possible symbols be denoted by '*N*'. For *N* = 1, the amount produced by a unary device (like our raven) is 0.

Towards a formula of uncertainty

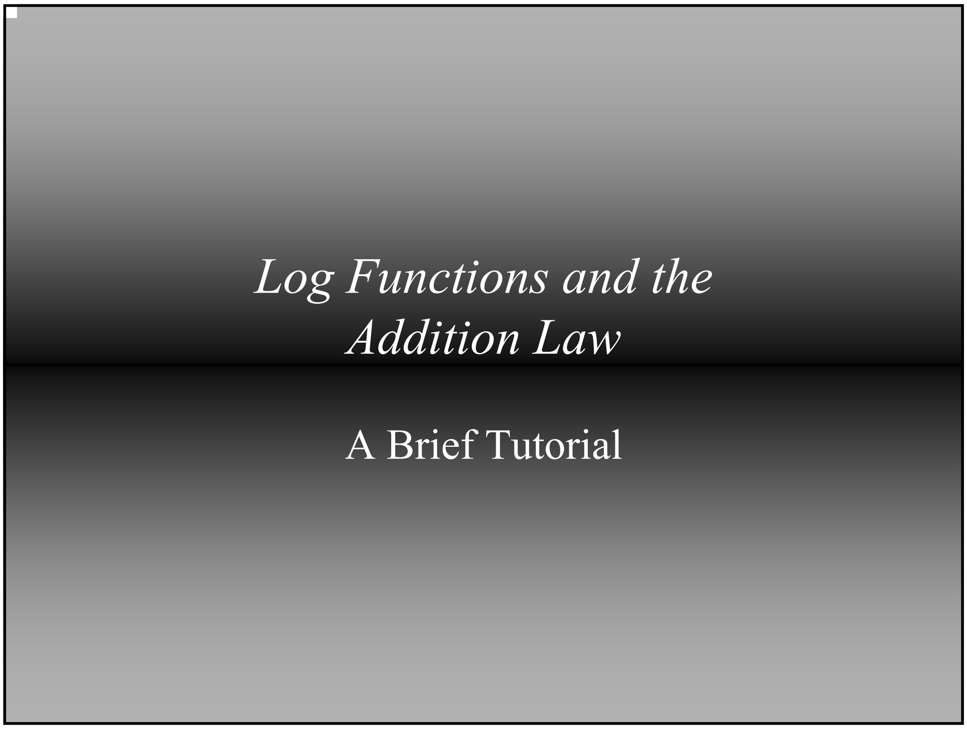
For *N* = 2, by producing an equiprobale symbol (as we assumed with our coin), the device delivers 1 unit of information, for *N* = 4, the device delivers the sum of the amount of information (!) provided by the first coin (A) *plus* the amount of information produced by the second coin (although we arrive at the total number of symbols by *multiplying* A's symbols by B's symbols.



Towards a formula of uncertainty

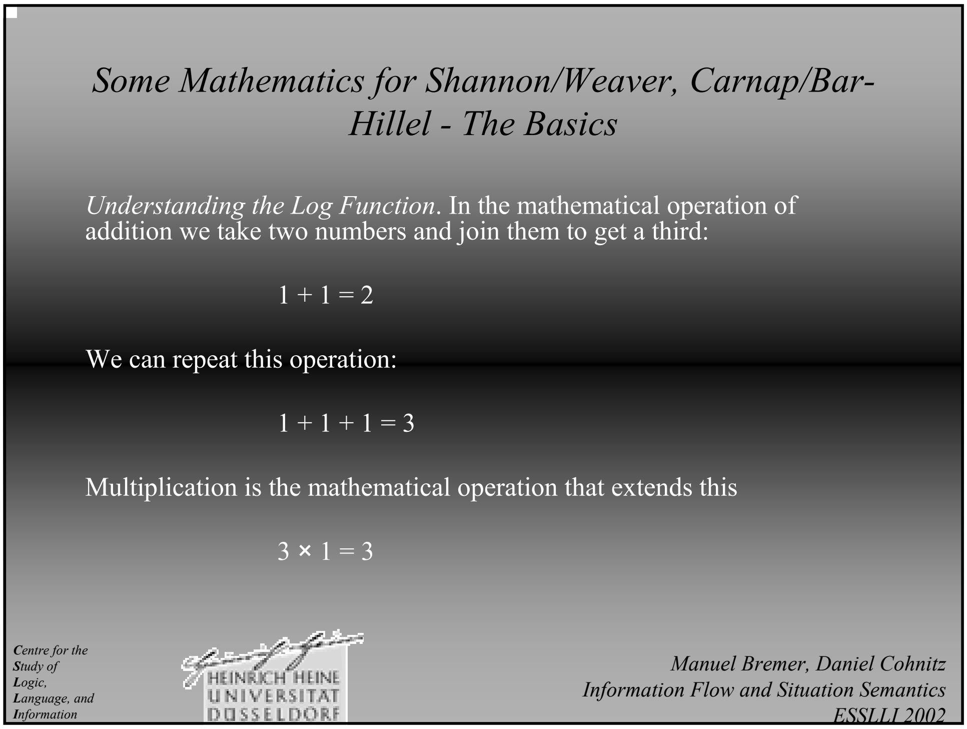
Our information measure, intuitively, should be a continuous and monotonic function of the probability of the symbols.

The most efficient way to achieve this is by using the logarithm to the base 2 of the number of possible symbols.



Log Functions and the Addition Law

A Brief Tutorial



Some Mathematics for Shannon/Weaver, Carnap/Bar-Hillel - The Basics

Understanding the Log Function. In the mathematical operation of addition we take two numbers and join them to get a third:

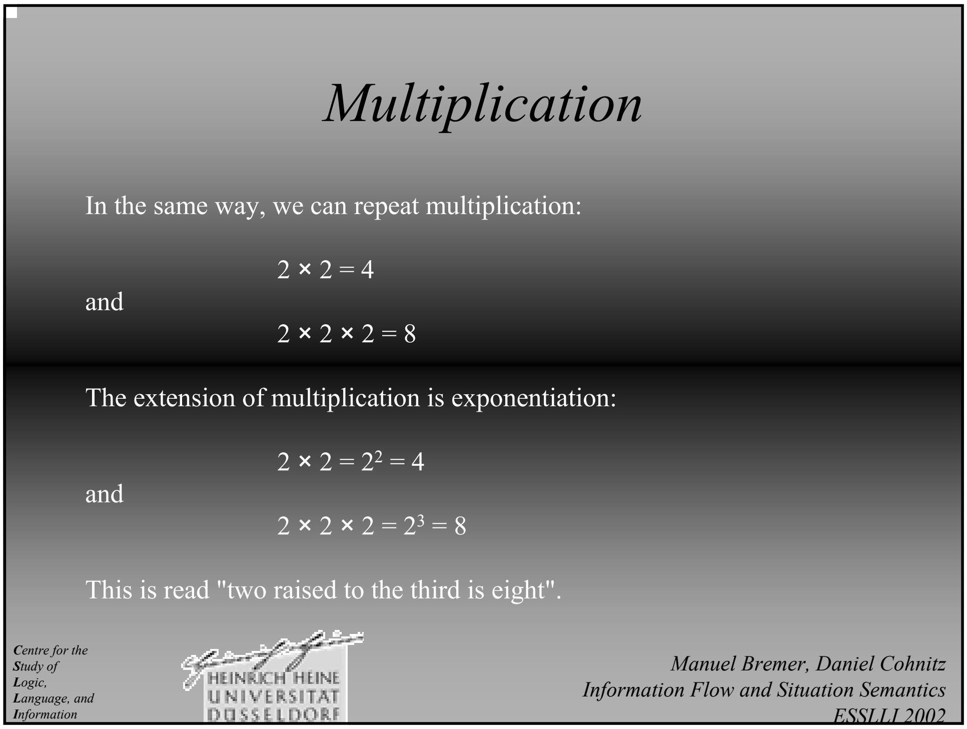
$$1 + 1 = 2$$

We can repeat this operation:

$$1 + 1 + 1 = 3$$

Multiplication is the mathematical operation that extends this

$$3 \times 1 = 3$$



Multiplication

In the same way, we can repeat multiplication:

$$2 \times 2 = 4$$

and

$$2 \times 2 \times 2 = 8$$

The extension of multiplication is exponentiation:

$$2 \times 2 = 2^2 = 4$$

and

$$2 \times 2 \times 2 = 2^3 = 8$$

This is read "two raised to the third is eight".

Exponentiation

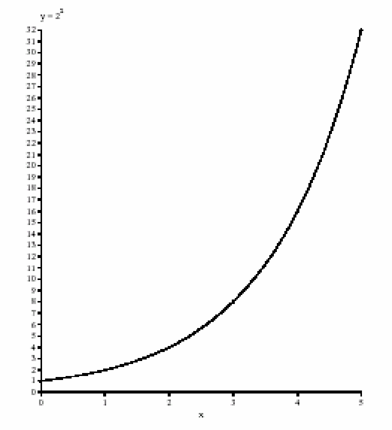
Because exponentiation simply counts the number of multiplications, the exponents add:

$$2^2 \times 2^3 = 2^{2+3} = 2^5$$

The number "2" is called the base of the exponentiation. If we raise the exponent to another exponent, the values multiply:

$$(2^2)^3 = 2^2 \times 2^2 \times 2^2 = 2^{2+2+2} = 2^{2\times3} = 2^6$$

The exponential function $y = 2^x$ is shown in this graph:



Logarithm

Now consider that we have a number and we want to know how many 2's must be multiplied together to get 32? That is, we want to solve this equation:

$$2^B = 32$$

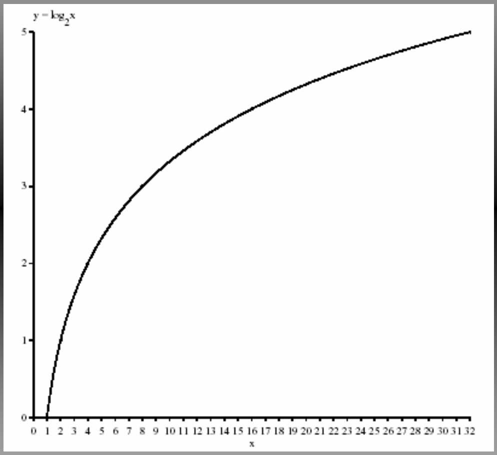
Of course, $2^5 = 32$, so $B = 5$. To be able to get a hold of this, mathematicians made up a new function called the logarithm:

$$\log_2 32 = 5$$

We pronounce this as "the logarithm to the base 2 of 32 is 5". It is the "inverse function" for exponentiation.

$$2^{\log_a a} = a \quad \text{and} \quad \log_2 (2a) = a$$

The logarithmic function $y = \log_2 x$ *is shown in this graph:*



The Addition Law

Consider this equation:

$$2^{a+b} = 2^a \times 2^b$$

Take the logarithm of both sides:

$$\log_2 2^{a+b} = \log_2 (2^a \times 2^b)$$

Since exponentiation and the logarithm are inverse operations, we can collapse the left side:

$$a + b = \log_2 (2^a \times 2^b)$$

The Addition Law

Now we substitute: $\log_2 x = a$ and $\log_2 y = b$:

$$\log_2 x + \log_2 y = \log_2 (2^{\log_2 x} \times 2^{\log_2 y})$$

Again, exponentiation and the logarithm are inverse operations, so we can collapse the two cases on the right side:

$$\log_2 x + \log_2 y = \log_2 (x \times y)$$

This is the additive property, Shannon was interested in.

End of tutorial!

Additivity

Hence, logarithms have the great advantage of turning multiplication of symbols into addition of information units, and by taking the logarithm to the base 2 we have the further advantage of expressing the units in bits.

Information per symbol

Given an alphabet of *N* equiprobable symbols, we can rephrase some examples more precisely by using the following equation:

[1] $\log_2 (N)$ = bits of information per symbol

Information per symbol

Device	Alphabet	Bits of information per symbol
raven (unary)	1 symbol	log(1) = 0
1 coin (binary)	2 symbols	log(2) = 1
2 coins	4 symbols	log(4) = 2
dice	6 symbols	log(6) = 2.58
3 coins	8 symbols	log(8) = 3

Life is unfair

In these examples we assumed that we are dealing with fair coins and an ideal dice. Unfortunately, life isn't fair and coins are always biased.

Now, to calculate how much information a biased source can produce one needs to rely on the probability of the occurrences of symbols in a series of tosses.

Calculating it

Compared to a fair coin, a slightly biased coin must produce less than 1 bit of information, but still more than 0. The raven produced no information at all because the occurrence of a string *S* of “nevermore” was not *informative* (not *surprising*), and that is because the *probability* of the occurrence of “nevermore” was maximum, so overly predictable.

Calculating it

Likewise, the amount of raw information produced by the biased coin depends on the average *informativeness* (or average *surprisal*) of the string *S* of *h* and *t* produced by the coin, and *S*’s average informativeness depends on the *probability* of the occurrence of each symbol.

The higher the frequency of a symbol in *S* the less raw information is being produced by the coin, up to the point when the coin is so biased to produce always the same symbol and stops being informative, behaving like the raven. So, to calculate the average informativeness of *S* we need to know how to calculate *S* and the informativeness of a i^{th} symbol in general, and this requires understanding what the probability of a i^{th} symbol (P_i) to occur is.

Calculating it

The probability P_i of the i^{th} symbol can be "extracted" from equation [1], where it is embedded in log(N), a special case in which the symbols are equiprobable. Using some elementary properties of the logarithmic function we have that:

[2] $\log(N) = -\log(N^{-1}) = -\log(1/N) = -\log(\mathrm{P})$

Calculating it

The value of $1/N = P$ can range from 0 to 1. The probability of "Hail Satan" is 0 if the raven is our source; $P(h) + P(t) = 1$, no matter how biased the coin is.

[3] $$\sum_{i=1}^{N} P_i = 1$$

Calculating it

We can now be precise about the raven: “nevermore” is not informative at all because $P_{nevermore} = 1$. Clearly, the lower the probability of occurrence of a symbol, the higher is the informativeness of an actual occurrence of it. The informativeness *u* of a i^{th} symbol can be expressed by analogy with – log (*P*) in equation [2]:

[4] $u_i = -\log(P_i)$

Calculating it

Next, we need to calculate the length of a general string *S*. Suppose that the biased coin, tossed 10 times, produces the string: <*h*, *h*, *t*, *h*, *h*, *t*, *t*, *h*, *h*, *t*>. The (length of the) string *S* (in our case equal to 10) is equal to the numbers of times the *h* type of symbol occurs added to the numbers of times the *t* type of symbol occurs. Generalizing for *i* types of symbols:

[5] $$S = \sum_{i=1}^{N} S_i$$

Calculating it

Putting together equations [4] and [5] we have that the average informativeness for a string of *S* symbols is the sum of the informativeness of each symbol divided by the sum of all symbols:

[6]
$$\frac{\sum_{i=1}^{N} S_i u_i}{\sum_{i=1}^{N} S_i}$$

Calculating it

Formula [6] can be simplified thus:

[7]
$$\sum_{i=1}^{N} \frac{S_i}{S} u_i$$

Calculating it

Now S_i/S is the frequency with which the ith symbol occurs in S when S is finite. If the length of S is left undetermined, then the frequency of the ith symbol becomes the probability P_i. So, further generalizing formula [7] we have:

[8] $$\sum_{i=1}^{N} P_i u_i$$

The Formula

Finally, by using equation [4] we can substitute for u_i and obtain

$$H = -\sum_{i=1}^{N} P_i \log_2 P_1 \text{ (bits per symbol)}$$

Take a breath!

Interpretation

This function represents the lower limit on the expected number of symbols required to code for the outcome of an event regardless of the method of coding, and is thus the unique measure of the quantity of information. It is the amount of information that would be required to reduce the uncertainty about an event with a set of probable outcomes to a certainty.

Adequacy

As derived by Shannon it is the only measure of information that simultaneously meets the three conditions of being continuous over the probability, of monotonically increasing with the number of equiprobable outcomes, and of being the weighted sum of the same function defined on different partitions of the probable outcomes.

In the discrete and continuous forms, the uncertainty corresponds to the entropy of statistical mechanics and to the entropy of the second law of thermodynamics, and it is the foundation of information theory.

Maximum

Equation [9] indicates that the quantity of raw information produced by a device corresponds to the amount of data deficit erased and it is a function of the average informativeness of the (potentially infinite) string of symbols produced by the device. It is easy to prove that, if symbols are equiprobable, [9] reduces to [1] and that the highest quantity of raw information is produced by a system whose symbols are equiprobable (compare the fair coin to the biased one).

Coding

Consider our *AB* system. Each symbol occurs with 0.25 probability. A simple way of encoding its symbols is to associate each of them with two digits:

<*h*, *h*> = 00

<*h*, *t*> = 01

<*t*, *h*> = 10

<*t*, *t*> = 11

In this Code 1 a message conveys 2 bits of information, as expected. Do not confuse *bits* as *bi*-nary uni*ts* of information (recall that we decided to use log2 only as a matter of convenience) with *bits* as *bi*-nary digi*ts*, which is what a 2-symbols system uses to encode a message.

Coding

Suppose now that the *AB* system is biased, and that the four symbols occur with the following probabilities:

<*h*, *h*> = 0.5

<*h*, *t*> = 0.25

<*t*, *h*> = 0.125

<*t*, *t*> = 0.125

This system produces less information, so by using Code 1 we would be wasting resources.

Coding

A more efficient Code 2 should take into account the symbols’ probabilities, with the following outcomes:

<*h*, *h*> = 0 0.5 × 1 binary digit = .5

<*h*, *t*> = 10 0.25 × 2 binary digits = .5

<*t*, *h*> = 110 0.125 × 3 binary digits = .375

<*t*, *t*> = 111 0.125 × 3 binary digits = .375

In Code 2, known as Fano Code, a message conveys 1.75 bits of information. One can prove that, given that probability distribution, no other coding system will do better than Fano Code.

Redundancy

On the other hand, in real life a good codification is also modestly redundant. *Redundancy* refers to the difference between the physical representation of a message and the mathematical representation of the same message that uses no more bits than necessary. *Compression* procedures work by reducing data redundancy, but redundancy is not always a bad thing, for it can help to counteract *equivocation* (data sent but never received) and *noise* (unwanted data). A message + noise contains more data than the original message by itself. But the aim of a communication process is *fidelity*, the accurate transfer of the original message from sender to receiver, not data increase.

Noise

Noise extends the informee’s freedom of choice in selecting a message, but it is an undesirable freedom and some redundancy can help to limit it. We are more likely to reconstruct a message correctly at the end of the transmission if some degree of redundancy counterbalances the inevitable noise and equivocation introduced by the physical process of communication and the environment. That is why, in a crowded pub, you shout your orders twice and add some gestures.

The Fundamental Theorems

We are now ready to understand Shannon’s two fundamental theorems. Suppose the 2-coins biased system produces the following message:

<*t*, *h*> <*h*, *h*> <*t*, *t*> <*h*, *t*> <*h*, *t*>.

Using Fano Code we obtain: 11001111010. The next step is to send this string through a channel. Channels have different transmission rates (*C*), calculated in terms of bits per second (bps). Shannon’s fundamental theorem of the noiseless channel states that

The Fundamental Theorems

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 of C/H – ε symbols per second over the channel where ε is arbitrarily small. It is not possible to transmit at an average rate greater than C/H.

(Shannon 1998, 59).

The Fundamental Theorems

In other words, if you devise a good code you can transmit symbols over a noiseless channel at an average rate as close to *C*/*H* as one may wish, but, no matter how clever the coding is, that average can never be made exceed *C*/*H*. We have already seen that the task is made more difficult by the inevitable presence of noise. However, the fundamental theorem for a discrete channel with noise comes to our rescue:

The Fundamental Theorems

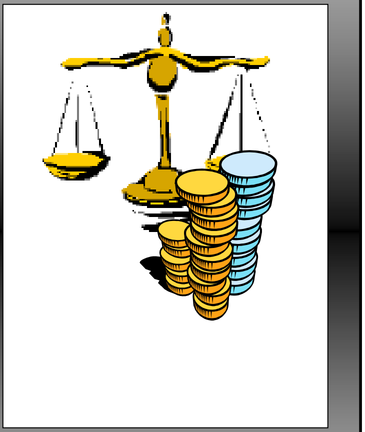
Let a discrete channel have the capacity C and a discrete source the entropy per second H. If H ≤ C there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation) If H > C it is possible to encode the source so that the equivocation is less than H – C + ε where ε is arbitrarily small. There is no method of encoding which gives an equivocation less than H – C.

(Shannon 1998, 71)

Application

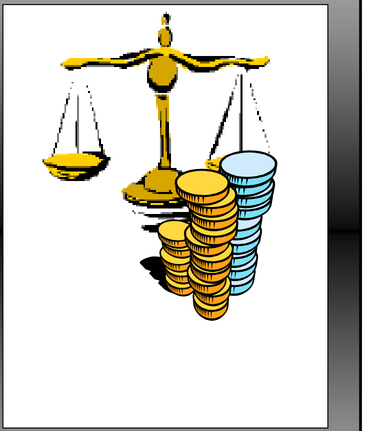
To learn how Information theory can help you in real-world situation, consider the following case you might have encountered frequently already:

Damn Coins



You have a balance and nine coins. Eight of the nine coins are of equal weight. The ninth, however, is of different weight (but it is unbeknownst to you whether it is lighter or heavier than the others.)

Damn Coins



Problem:

Develop a strategy to figure out by weighting only three times which coin differs in weight from the others and whether it is lighter or heavier than the others are.

Possibilities

It seems reasonable to put always an equal number of coins onto the scales. In this case there are three possibilities:

1) the left scale goes down

2) the balance remains in equilibrium

3) the right scale goes down

Hence, the highest amount of information you can receive by weighing once is log 3 = 1.58 bits.

Possibilities

Now weighing three times can possibly create 4.74 bits of information. Being in the dark about i) which is the deviant coin and ii) whether it is lighter or heavier, you are asked to choose one possibility from a set of 18 equiprobable ones.

Solvability

Maybe we should first check whether the problem is solvable at all. For this the information we can receive by weighing three times should be higher or equal to the information that corresponds to the 18 equiprobable outcomes. Luckily this is the case:

log 18 = 4.16 bit < 4.74 bit

Strategy

Unfortunately there is quite a number of ways how to put the coins onto the scales. Now we want to use information theory to develop a strategy.

It seems clever to get always the maximal information out of every single weighting.

Strategy

How much information do we gain from one weighting? Some definitions:

P_l = Probability that the left scale goes down

P_b = Probability that the scales remain in equilibrium

P_r = Probability that the right scale goes down

Application of the formula

Now we can apply our formula and see that the information gained by weighing once is

$$H = -(P_l \log P_l + P_b \log P_b + P_r \log P_r)$$

Now we know, that H is at maximum if the probabilities are all equal. This strategy results in a simple rule: Weigh such that $P_l = P_b = P_r$ for each single case.

The way to solution

If we put n ($1 \leq n \leq 4$) coins onto the left scale and n onto the right, $9 - 2n$ coins will remain unweighted. In probabilities:

$P_b = (9 - 2n)/9$

$P_l = P_r = n/9$

If we want equiprobability, n has to be 3.

The way to solution

Now we mark all coins from 1 to 9. In the first step we put 1, 2, 3 onto the left scale and 4, 5, 6 onto the right. Now, either one of the scales goes down, or not. In case none goes down, we know that the weird coin is among 7, 8, and 9. Now we put 7, and 8 onto the scales and weigh a second time. Easy to see that this leads to a solution.

The way to solution

Assume that after the first weighting the scales weren't in equilibrium. Now we'll use only 4 of the 6 coins we used in the first weighting to keep the probabilities at 1/3. To achieve this we have to move the weird coin with probability 1/2 from one scale to the other.

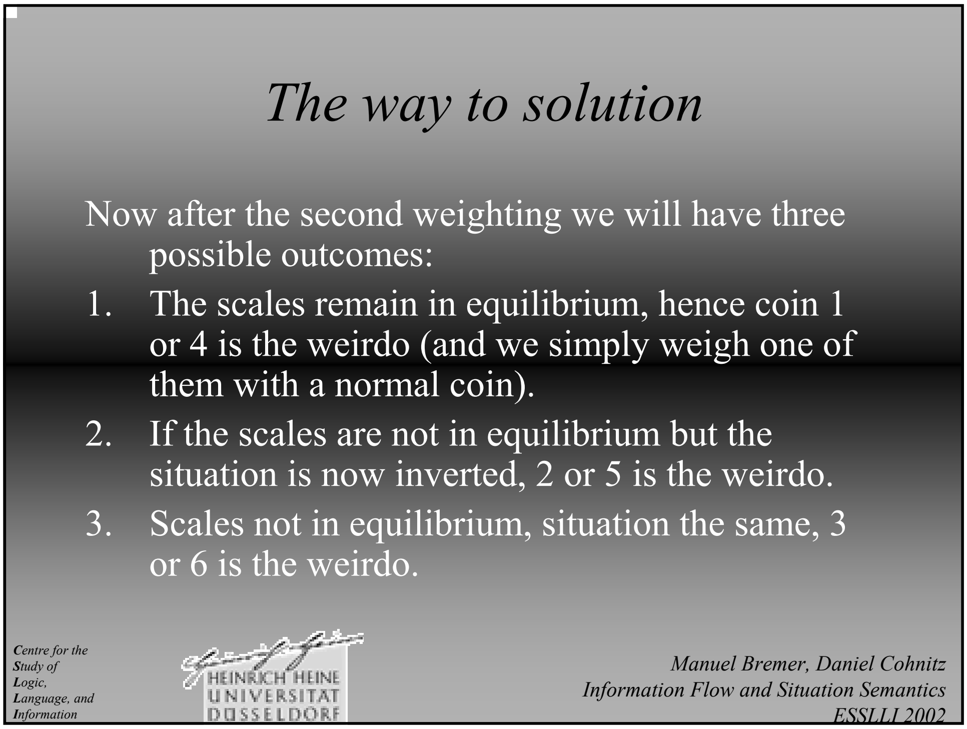
The way to solution

We can do this easily:

Remove 1 and 4 from the scales.

Interchange 2 and 5.

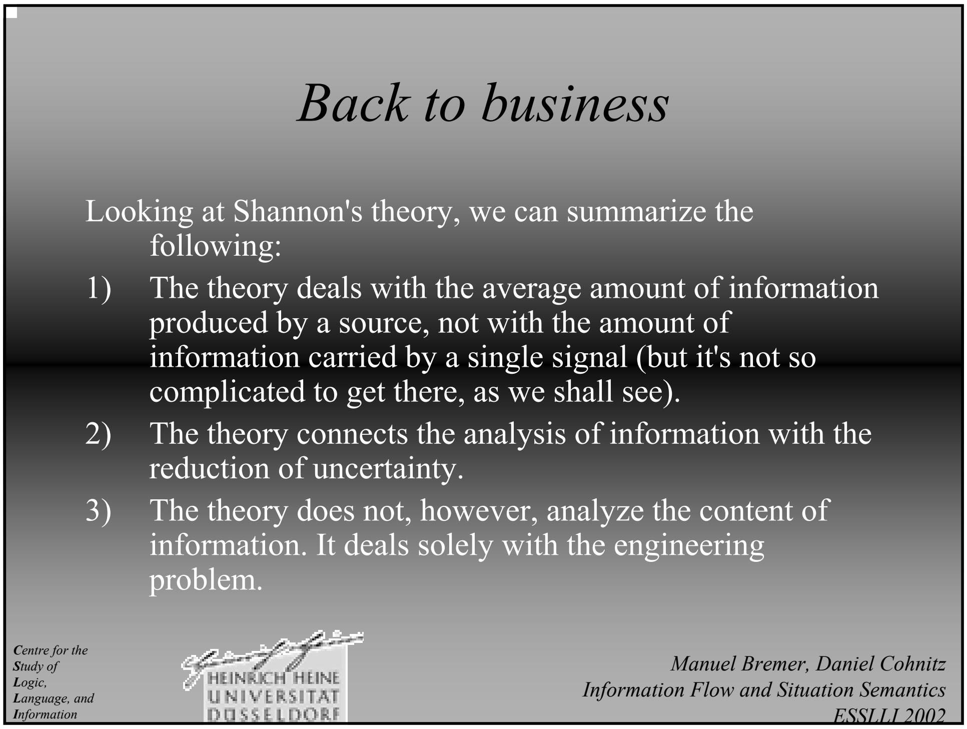
Leave 3 and 6 where they are.



The way to solution

Now after the second weighting we will have three possible outcomes:

1. The scales remain in equilibrium, hence coin 1 or 4 is the weirdo (and we simply weigh one of them with a normal coin).
2. If the scales are not in equilibrium but the situation is now inverted, 2 or 5 is the weirdo.
3. Scales not in equilibrium, situation the same, 3 or 6 is the weirdo.



Back to business

Looking at Shannon's theory, we can summarize the following:

1) The theory deals with the average amount of information produced by a source, not with the amount of information carried by a single signal (but it's not so complicated to get there, as we shall see).

2) The theory connects the analysis of information with the reduction of uncertainty.

3) The theory does not, however, analyze the content of information. It deals solely with the engineering problem.

The Limits of Communication Theory

"Frequently the messages have meaning; that is they refer to or are correlated to some system with certain physical or conceptual entities. These semantic aspects are irrelevant to the engineering problem." (Claude E. Shannon 1948)

"It is important to emphasize at the start that we are not concerned with the meaning or the truth of messages; semantics lies outside the scope of "mathematical information theory"." (E. Colin Cherry 1950)

"Information and uncertainty are technical terms that describe any process that selects one or more objects from a set of objects. We won't be dealing with the meaning or implications of the information since nobody knows how to do that mathematically." (T. Schneider 2000)