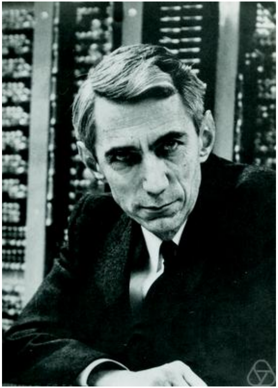
How much does this matchbox know?

(measuring information and entropy)

Alex Vlasiuk

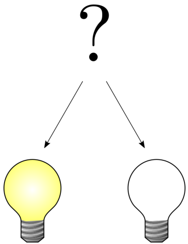
Claude Elwood Shannon

- "the father of information theory"
- ideas that have left a powerful imprint on how we think about information
- his 1948 paper is very accessible
- it is also very deep: for example, all Pierce's 1961 300-page book does is recycling Shannon's ideas



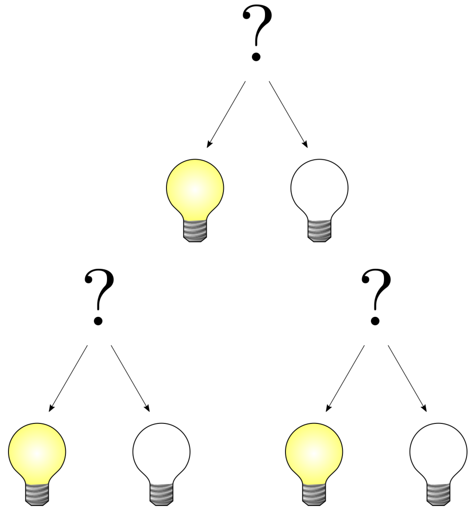
Shannon, Claude Elwood. "A mathematical theory of communication." ACM SIGMOBILE Mobile Computing and Communications Review 5.1 (2001): 3-55.

Light bulbs



ON/OFF

Light bulbs



ON/OFF

ON-ON	ON-OFF
OFF-ON	OFF-OFF

Uncertainty of the light bulbs

For n light bulbs:

$$H = \log_2 \{\text{number of states}\} = \log_2 2^n = n.$$

These light bulbs are independent. Units of uncertainty: **bits**. The number of binary choices needed to identify the state of a system.

Causing uncertainty is opposite to having structure! Example: you can predict the next letter in the word "Floccinaucinihili

† estimating something as worthless

Uncertainty of the light bulbs

For n light bulbs:

$$H = \log_2 \{\text{number of states}\} = \log_2 2^n = n.$$

These light bulbs are independent. Units of uncertainty: **bits**. The number of binary choices needed to identify the state of a system.

Causing uncertainty is opposite to having structure! Example: you can predict the next letter in the word "Floccinaucinihili**p**ilification",[†] because of the structure imposed by the dictionary.

[†] estimating something as worthless

Example: approximations to English

Using 27 (26+space) alphabet.

- symbols independent and equally probable:
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

Example: approximations to English

Using 27 (26+space) alphabet.

- symbols independent and equally probable:
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
- symbols independent but with frequencies of English text:
  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

Example: approximations to English

Using 27 (26+space) alphabet.

- symbols independent and equally probable:
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
- symbols independent but with frequencies of English text:
  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
- digram structure as in English:
  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE

Example: approximations to English

Using 27 (26+space) alphabet.

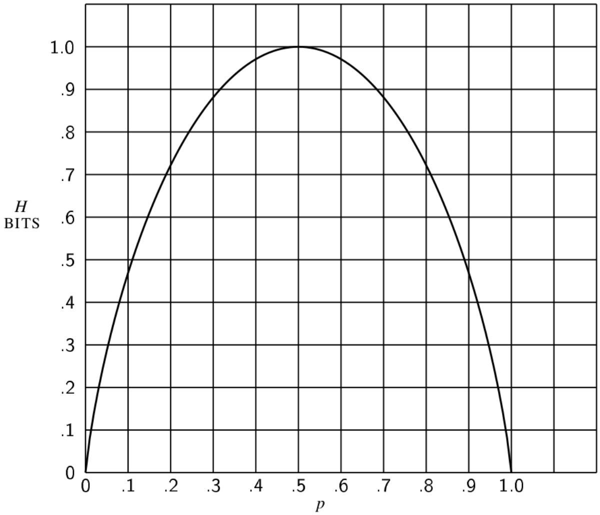
- symbols independent and equally probable:
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
- symbols independent but with frequencies of English text:
  OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL
- digram structure as in English:
  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE
- trigram structure as in English:
  IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

- first-order word approximation:
  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

- first-order word approximation:
  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
- Second-order word approximation:
  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 WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Coin toss

Probabilities of the outcomes: p and $1 - p$



Entropy

For the "coin" with probabilities p and $q = 1 - p$, the measure of uncertainty can be

$$H = -(p \log_2 p + q \log_2 q)$$

For arbitrary set of possible events with probabilities

$$p_1, p_2, \ldots, p_k,$$

the measure of uncertainty in the outcome, when sampling one event:

$$\boldsymbol{H = -\sum_{i=1}^{k} p_i \log_2 p_i}$$

A motivation

Every step selects one of k symbols in a finite "alphabet". Suppose successive symbols are independent.

For a long message of N symbols:
around $p_1 N$ of the first, $p_2 N$ of the second, etc, so probability of this particular long message (independence!) is about

$$p = p_1^{p_1 N} p_2^{p_2 N} \dots p_k^{p_k N}$$

$$p = p_1^{p_1 N} p_2^{p_2 N} \ldots p_k^{p_k N}$$

$$\begin{aligned}\log_2 p &= p_1 N \log_2 p_1 + \ldots + p_k N \log_2 p_k \\ &= N \sum_{i=1}^{k} p_i \log_2 p_i \\ &= -NH\end{aligned}$$

so

$$H = -\frac{\log_2 p}{N}$$

— average number of binary choices needed to specify one symbol in a long message.

Redundancy

{ Maximal entropy of N letters } – { Entropy of an N-letter message }

- Low redundancy (good password):
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
- High redundancy (bad password):
  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 WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

Redundancy $\Longleftrightarrow$ word puzzles!

- Redundancy $= 0 \implies$ everything is a word! Any two-dimensional array of letters forms a crossword puzzle. If the redundancy is too high though, too many constraints for a crossword.
- Redundancy allows statistical attacks against encryption: see E. A. Poe "The Gold-Bug".
- Such attacks actually only work against naive cyphers, such as the one in Gold-Bug.
- Redundancy makes it possible to read your scratched CDs.

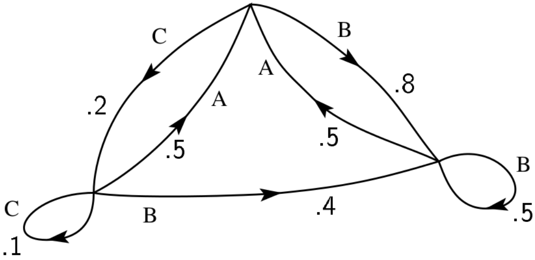
Sources with dependent symbols

Markov processes

Source is a stochastic (random) process.

Example. Alphabet: A, B, C. Transition probabilities:

$p_i(j)$		j		
		A	B	C
	A	0	$\frac{4}{5}$	$\frac{1}{5}$
i	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$



Entropy of a source

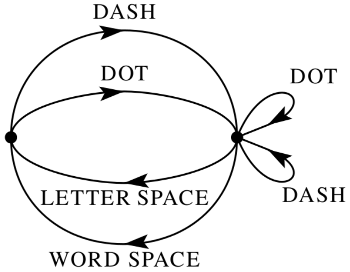
A source has states with entropies H_i and probabilities P_i, then

$$H = \sum_i P_i H_i = -\sum_{i,j} P_i\, p_i(j) \log_2 p_i(j)$$

As before, for N large,

$$H = -\frac{\log_2 p}{N}$$

p – probability of a typical sequence of length N



Entropies of natural languages

- another paper by Shannon: entropy of the English language is 11.82 bits per word, and since the average word has 4.5 letters, entropy is about 2.62 bits per letter
- size of `char` on a standard system is 8 bits
- Fabrice Bellard's compression of English: `http://textsynth.org/sms.html`, about 15% ratio (number of output bits divided by the number of input bits)
- Kolmogorov mentions some studies of Russian with entropy around 2 bits/letter for random text with correct grammar, and 1 bit/letter for "War and piece".
- For a matchbox...

Capacity and states of a channel

Symbols: $S_1, \ldots, S_k$ with certain durations $t_1, \ldots, t_k$. Allowed combinations of symbols are signals.

Capacity of a channel:

$$C = \frac{\log_2 N(T)}{T},$$

with T large, where $N(T)$ is the number of allowed signals of duration T.

Units: bits per second.

Noiseless channel

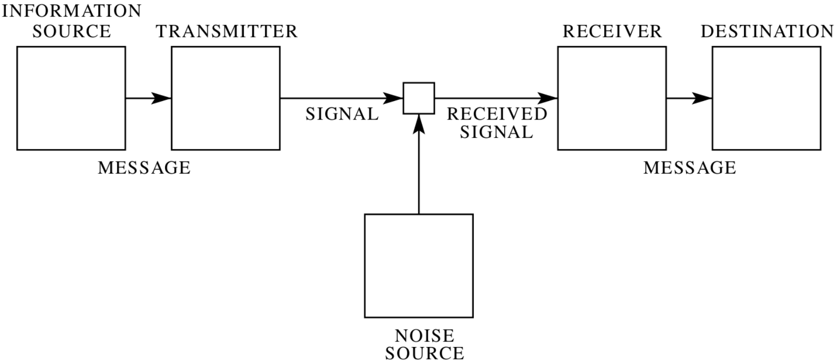
Theorem (the fundamental theorem for a noiseless channel)

Suppose a source has entropy $\boldsymbol{H}$ bits/symbol and a channel has capacity $\boldsymbol{C}$ bits/second. Then it is possible to encode the output of the source to transmit at the average rate $\boldsymbol{\frac{C}{H} - \epsilon}$ symbols/second over the channel where ϵ is arbitrarily small.
It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

$$\frac{[C]}{[H]} = \frac{\text{bits/s}}{\text{bits/sym}} = \text{sym/s}$$

The proof involves constructing an explicit code that achieves the required rate: Shannon-Fano coding.

Overall setting



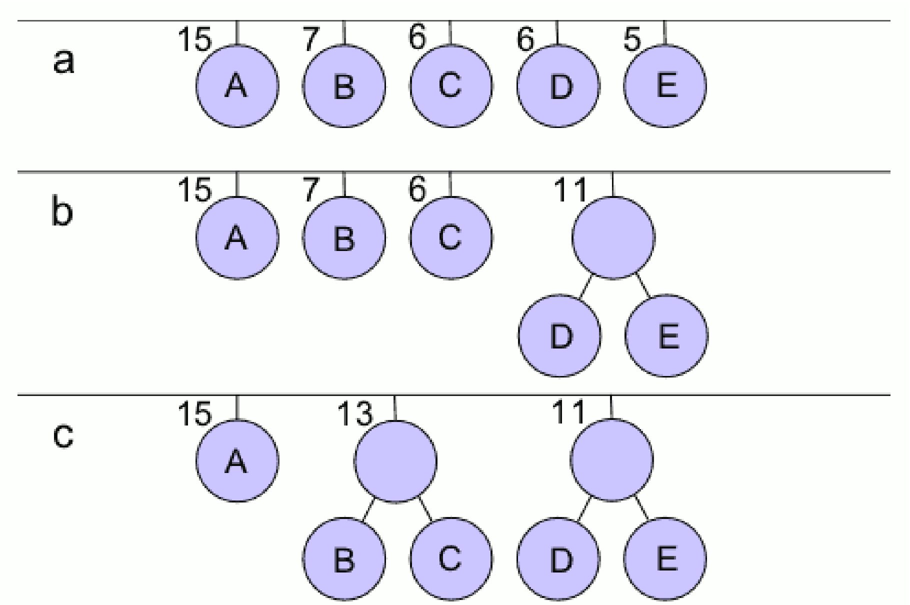
Huffman coding

Algorithm

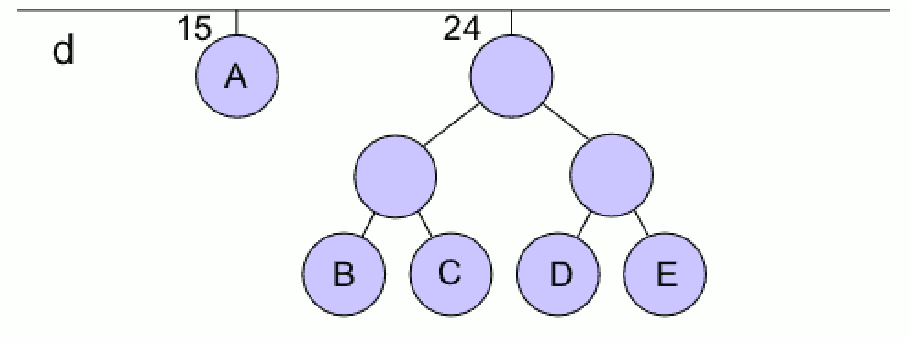
1. Create a leaf node for each symbol and add it to a **priority queue**, by decreasing frequency.
2. While there is more than one node in the queue:
   - Remove the two nodes of lowest probability or frequency from the queue
   - Prepend 0 and 1 respectively to any code already assigned to these nodes
   - Create a new internal node with these two nodes as children and with probability equal to the sum of the two nodes' probabilities.
   - Add the new node to the queue.

The remaining node is the root node and the tree is complete.

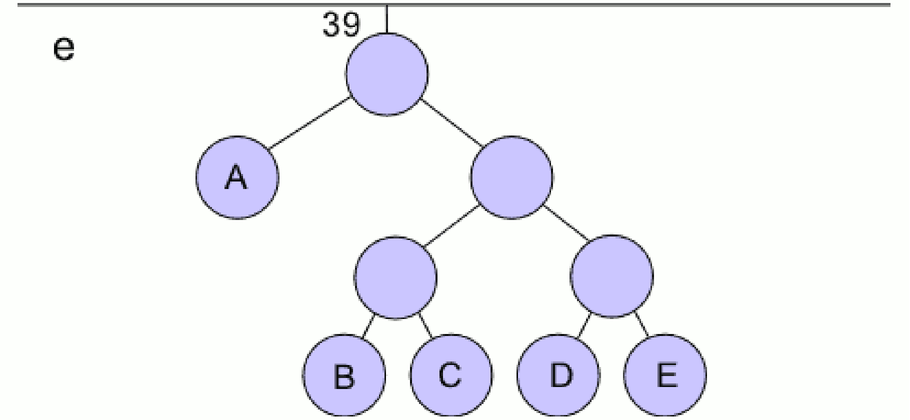
Huffman coding example



ⒸImage: Wikimedia Commons



d
15
24
A
B
C
D
E



Codewords:

A	0	C	101	E	111
B	100	D	110		

$$\frac{1\,\text{bit} \cdot 15 + 3\,\text{bits} \cdot (7+6+6+5)}{39\,\text{symbols}} \approx 2.23\,\text{bits per letter.}$$

©Image: Wikimedia Commons

Conclusions

- Entropy $H = -\sum_i p_i \log_2 p_i$ is a measure of uncertainty/information
- Allows to determine optimal compression
- Can be used to introduce redundancy, increasing reliability of the encoding

Thanks!