HOW DO WE COMPRESS INFORMATION?

claude shannon
1916-2016

IEEE
IEEE INFORMATION THEORY SOCIETY it

many thanks to
Kousha Kalantari, Nematollah Iri, Oliver Kosut, and Lalitha Sankar (School of Electrical, Computer, and Energy Engineering / Arizona State University)

design
alesiom.ch

In order to reduce communication and storage costs we often "compress" data. The less "information" data contains, the more it can be compressed. Shannon showed that the amount of information that is present in data can be measured mathematically by what is called the "entropy" and that entropy is also related to the minimum number of trials we would need to "guess" an object.

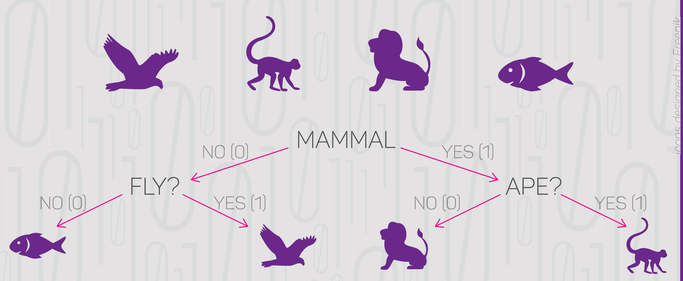
DATA COMPRESSION

Morse code, invented in 1838, is an early example of data compression based on using shorter codewords for letters such as "e" and "t" that are more common in English. Modern work on data compression began in the late 1940s with the development of information theory. In 1949, Claude Shannon and Robert Fano devised a systematic way to assign codewords based on probabilities of blocks. An optimal method for doing this was then found by David Huffman in 1951. Early implementations were typically done in hardware, with specific choices of codewords being made as compromises between compression and error correction.

https://www.wolframscience.com/reference/notes/1069b

SIMPLE GAME

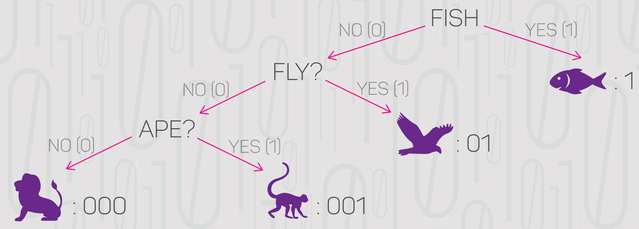
- Alice chooses an animal among these:
- Bob wants to ask yes/no questions from Alice to guess her choice.
- One possible way is with an **equal length code**:



icons designed by Freepik

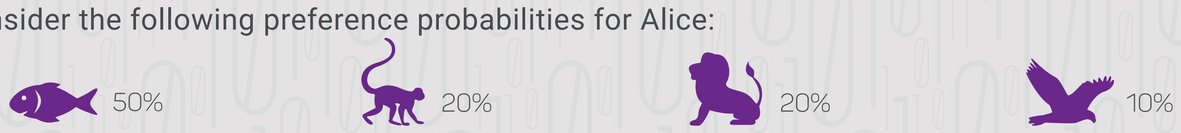
ANOTHER LOOK

- We can call each animal by its binary name.
- What if Bob knows in advance that Alice mostly prefers sea animals? Is there a better way?
- One possible way is with an **unequal length code**:



PERFORMANCE METRIC

- Consider the following preference probabilities for Alice:



- The average codelength for the unequal length code is:

$$\left(\frac{50}{100}\right) * 1 + \left(\frac{20}{100}\right) * 3 + \left(\frac{20}{100}\right) * 3 + \left(\frac{10}{100}\right) * 2 = 1.9\ bits.$$

- Can you come up with a question scheme with lower average length?

HOW MUCH CAN WE COMPRESS

- Minimum average codelength of animals is related to the amount of information Bob obtains from Alice's answers:
  - For equally likely animals, Bob obtains a lot of information.
  - If one animal is more likely than others, Bob obtains less information on average.
- Claude Shannon came up with a measure of information called **entropy** for k data types, with chances $p_1, p_2, \ldots, p_k$:

$$H = p_1 \log\frac{1}{p_1} + p_2 \log\frac{1}{p_2} + \ldots + p_k \log\frac{1}{p_k}$$

- Average codelength of any question scheme is always bigger than Shannon entropy.
- Average codelength of the best question scheme is within one bit of the Shannon entropy.

FUTURE CHALLENGES

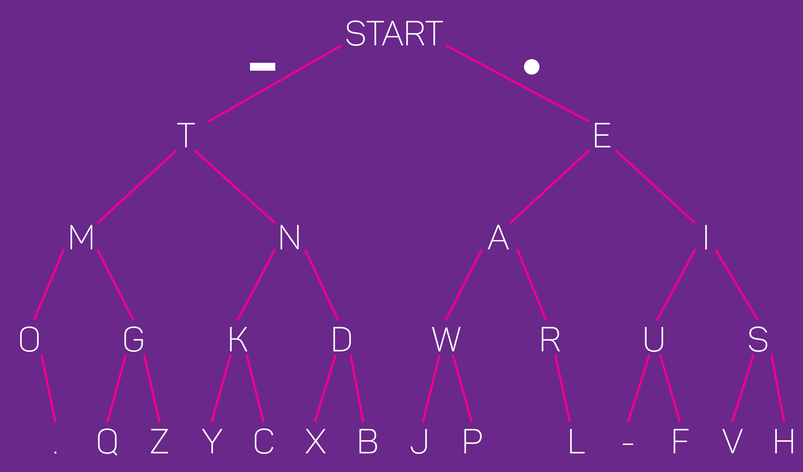
- How to optimally compress data allowing some levels of distortion, when we do not make statistical assumptions on the data and with low complexity? (universal data compression with distortion).
- Compression for new applications.

DAILY APPLICATIONS

Morse code

- Send text over telegraph lines
- Short representations for frequent 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 — — ● ●



JPEG, gzip

- Similar coding methods used to compress images and files.

WHY COMPRESS?

Compression enables you to **optimize storage** and **communication** such that you **reduce the cost per item** you store and/or share.

If you take a photograph with your camera and save it in both bmp (uncompressed) and jpg (compressed - lossy), you will get a file size ratio of 16:1. This means that you can put 16 times more pictures on your memory card if you select the compressed format (jpg). You will also be able to add 16 times more pictures on your dropbox, and upload them 16 times faster on Facebook.

https://www.maketecheasier.com/compression-encryption-and-archive-formats/