Types of Compression

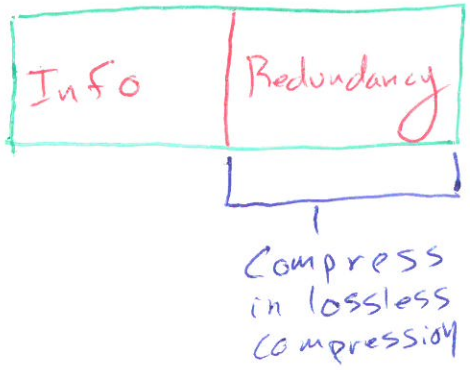
Lossy v. Lossless

Lossy means we can lose information
↳ media (images, video, sound)

Lossless means you get back exactly what you started with.

Theorem #1 There is no "perfect" compression algorithm that can take any data and make it smaller (always).

Information Theory:



Compression as a pipeline

Plain Data
↓
Domain Specific Redundancy Removal
↓
General Techniques (LZ a.k.a zip)
↓
Compressed data
↓
General Decompress
↓
Domain Specific Decompress
↓
Plain Data

Huffman I

A
B
C
D
.
.
.
Z

A-Z 5 bits

An improvement.

E 0
T 10
M 110
A 1110
R 11110
S 11111 0

AT MARS
1110 10 110 11101111011110

BAD Way To assign few bytes to Frequent letters & many bytes to infrequent letters.

ATMARS

E 0
T 1
M 01
A 10
R 11
S 00

BAD ENCODING: 10101101100

7?
First half of A
First half of R?

Huffman II

AAAAAAAAABBBBBBCCCDD 2 bits each

Histogram
A 9 · 2 = 18
B 5 · 2 = 10
C 3 · 2 = 6
D 2 · 2 = 4
38

A: 1
B: 00
C: 010
D: 011

9 · 1	9
5 · 2	10
3 · 3	9
2 · 3	6
	34

1. Find the two least frequent letters & assign them to 0 & 1.

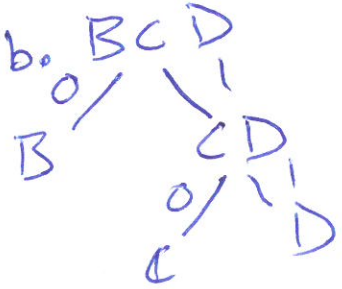
1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 1 1
A B B B B B C C C D D

Compression Ration

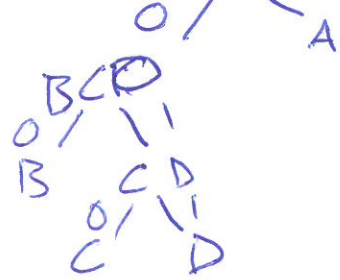
$\frac{34}{38} = \frac{17}{19}$

2. Reinsert combined letters into histogram
   A 9
   B 5
   CD 5

3. Repeat.

1b.


2b. BCD 10
A 9

1c.


2c. Done

LZ77, LZ78, (ZIP), etc.

first binary called PKZIP
Successor gzip

ABABACABA

Triples
(-Look back
-Read forward
-Next Byte)

(0, 0, A) A
(0, 0, B) AB
(2, B, C) ABA, ABAC
(4, 2, -) ABABACABA