

5.5 DATA COMPRESSION



- ▶ basics
- ▶ run-length coding
- ▶ Huffman compression
- ▶ LZW compression

- ▶ **basics**
- ▶ run-length coding
- ▶ Huffman compression
- ▶ LZW compression

Data compression

Compression reduces the size of a file:

- To save **space** when storing it.
- To save **time** when transmitting it.
- Most files have lots of redundancy.

Who needs compression?

- Moore's law: # transistors on a chip doubles every 18-24 months.
- Parkinson's law: data expands to fill space available.
- Text, images, sound, video, ...

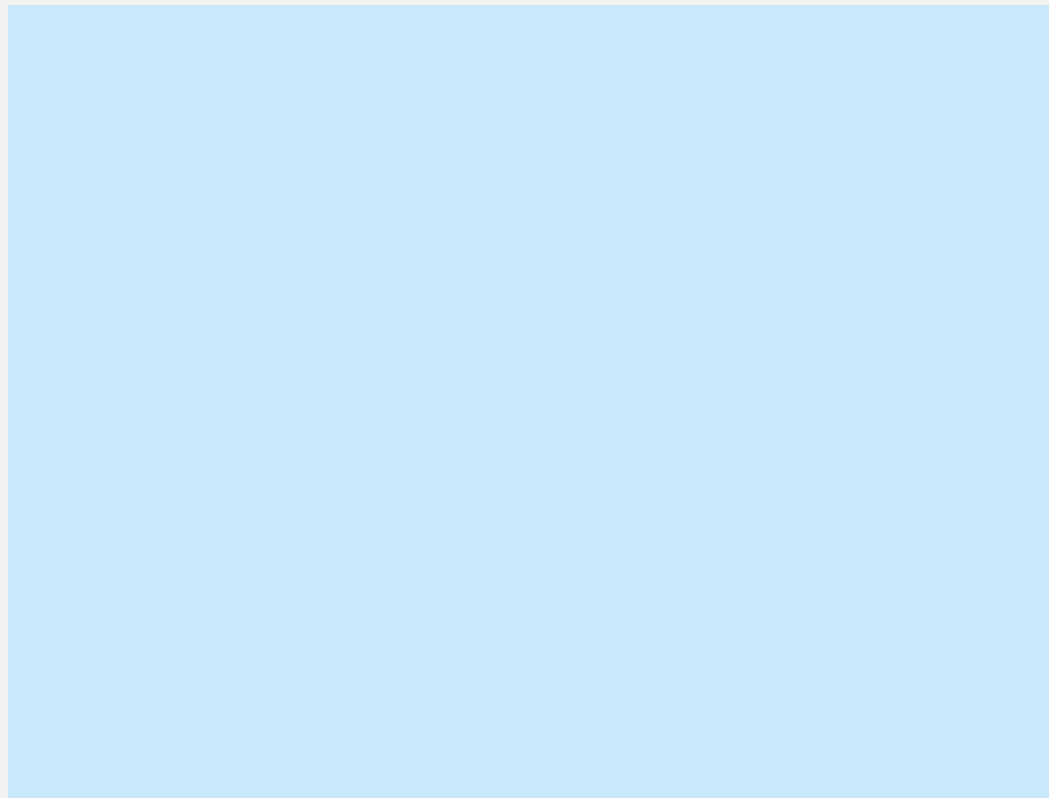
Lossy vs Lossless compression

- Maybe we can sacrifice some of the non-essential detail...

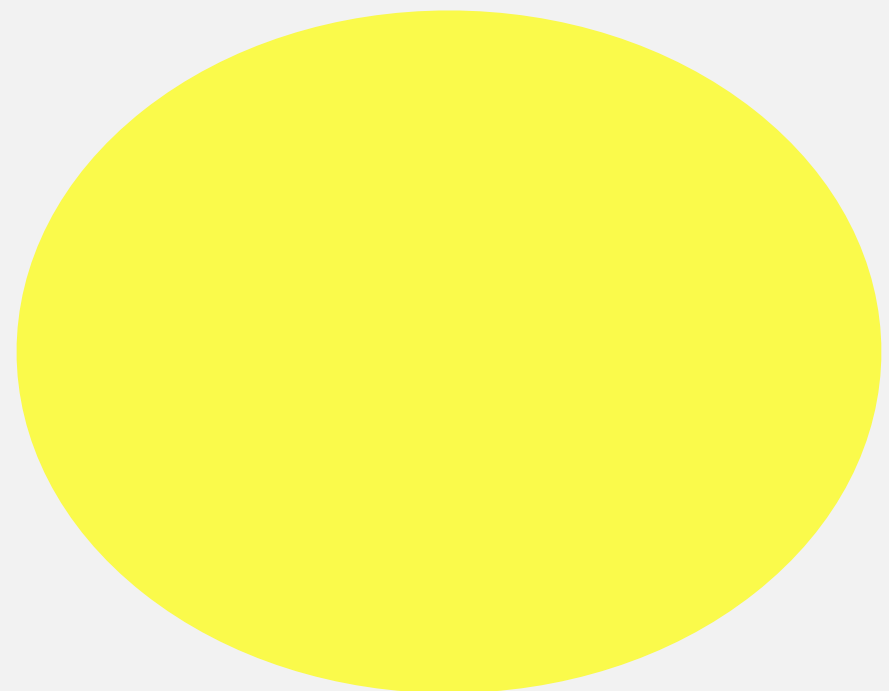
Lossy compression for images

Exploit what we know about human vision

- Human eye tends to “blend” nearby colors
- Visual acuity varies markedly across features
 - Discontinuities easily seen, absolutes less crucial
- Color perception is relative



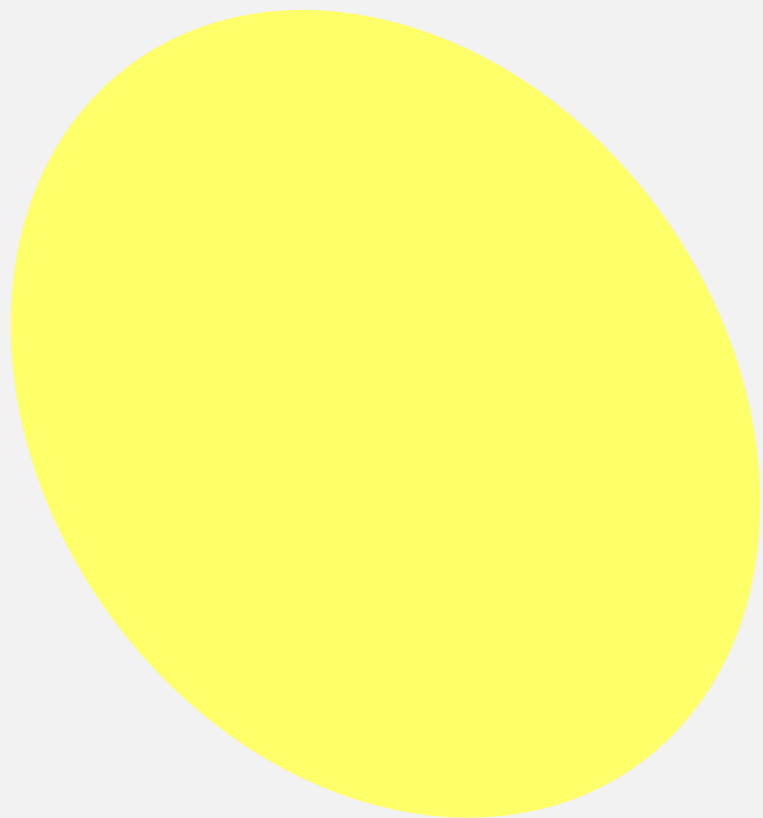
Do these colors look the same as ...



Lossy compression for images

Exploit what we know about human vision

- Human eye tends to “blend” nearby colors
- Visual acuity varies markedly across features
 - Discontinuities easily seen, absolutes less crucial
- Color perception is relative



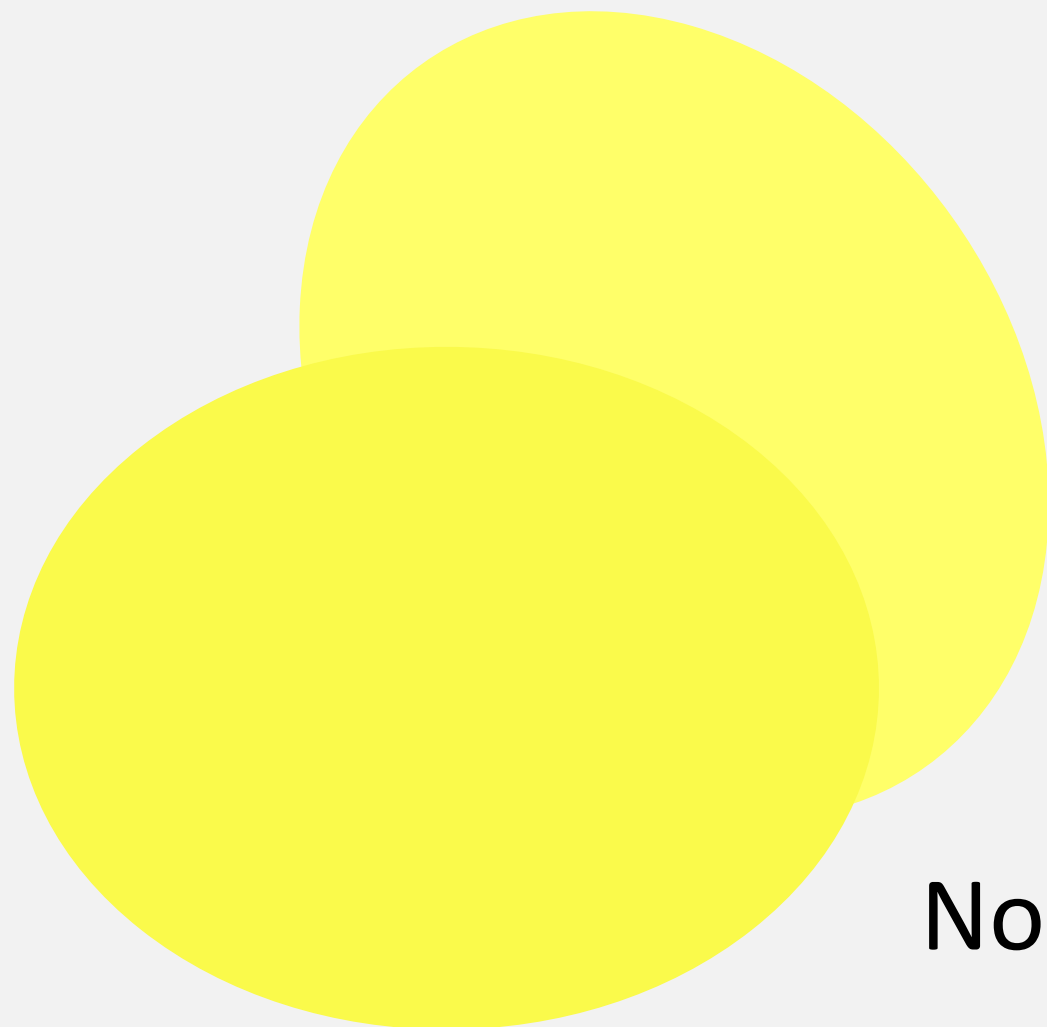
these?



Lossy compression for images

Exploit what we know about human vision

- Human eye tends to “blend” nearby colors
- Visual acuity varies markedly across features
 - Discontinuities easily seen, absolutes less crucial
- Color perception is relative



Not quite

Applications of compression

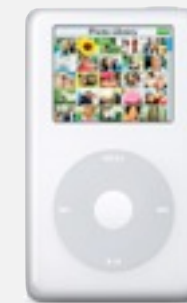
Generic file compression.

- Files: GZIP, BZIP, 7z.
- Archivers: PKZIP.
- File systems: NTFS, HFS+, ZFS.



Multimedia.

- Images: GIF, JPEG.
- Sound: MP3.
- Video: MPEG, DivX™, HDTV.



Communication.

- ITU-T T4 Group 3 Fax.
- V.42bis modem.
- Skype.



Databases. Google, Facebook,



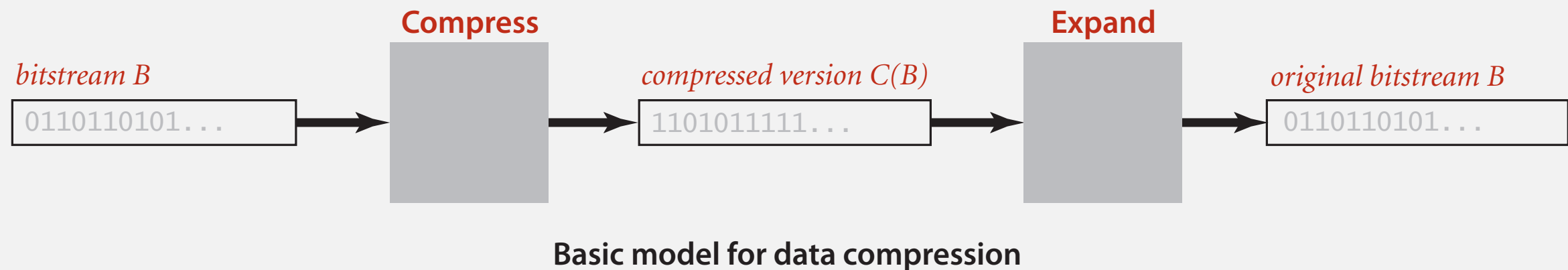
Lossless compression and expansion

Message. Binary data B we want to compress.

Compress. Generates a "compressed" representation $C(B)$.

Expand. Reconstructs original bitstream B .

uses fewer bits (you hope)



Compression ratio. Bits in $C(B)$ / bits in B .

Ex. 50-75% or better compression ratio for natural language.

Data representation: genomic code

Genome. String over the alphabet { A, C, T, G }.

Goal. Encode an N -character genome: **ATAGATGCATAG...**

Standard ASCII encoding.

- 8 bits per char.
- $8 N$ bits.

char	hex	binary
A	41	01000001
C	43	01000011
T	54	01010100
G	47	01000111

Two-bit encoding.

- 2 bits per char.
- $2 N$ bits.

char	binary
A	00
C	01
T	10
G	11

Fixed-length code. k -bit code supports alphabet of size 2^k .

Amazing but true. Initial genomic databases in 1990s did not use such a code!

Reading and writing binary data

Binary standard input and standard output. Libraries to read and write **bits** from standard input and to standard output.

```
public class BinaryStdIn
```

```
boolean readBoolean()      read 1 bit of data and return as a boolean value
```

```
char readChar()           read 8 bits of data and return as a char value
```

```
char readChar(int r)      read r bits of data and return as a char value
```

[similar methods for byte (8 bits); short (16 bits); int (32 bits); long and double (64 bits)]

```
boolean isEmpty()         is the bitstream empty?
```

```
void close()             close the bitstream
```

```
public class BinaryStdOut
```

```
void write(boolean b)     write the specified bit
```

```
void write(char c)        write the specified 8-bit char
```

```
void write(char c, int r) write the r least significant bits of the specified char
```

[similar methods for byte (8 bits); short (16 bits); int (32 bits); long and double (64 bits)]

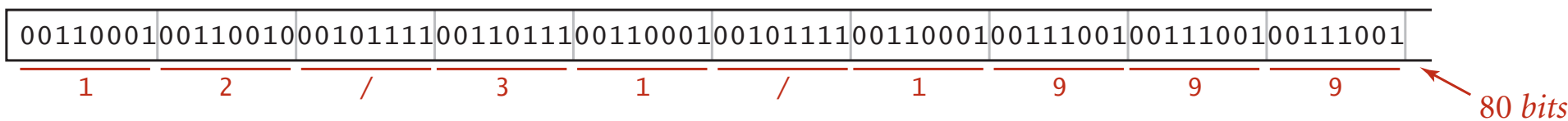
```
void close()             close the bitstream
```

Writing binary data

Date representation. Three different ways to represent 12/31/1999.

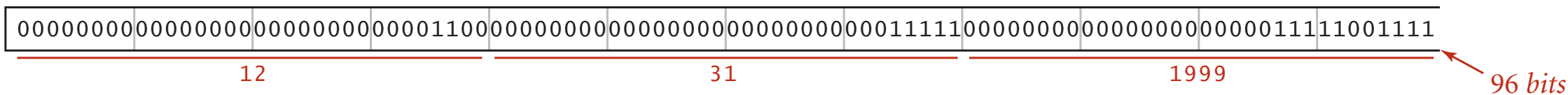
A character stream (StdOut)

```
StdOut.print(month + "/" + day + "/" + year);
```



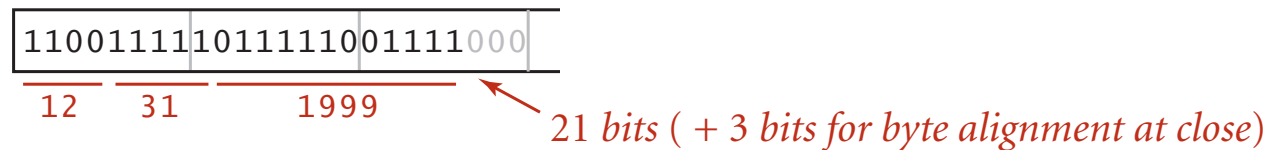
Three ints (BinaryStdOut)

```
BinaryStdOut.write(month);  
BinaryStdOut.write(day);  
BinaryStdOut.write(year);
```



A 4-bit field, a 5-bit field, and a 12-bit field (BinaryStdOut)

```
BinaryStdOut.write(month, 4);  
BinaryStdOut.write(day, 5);  
BinaryStdOut.write(year, 12);
```



Binary dumps

Q. How to examine the contents of a bitstream?

Standard character stream

```
% more abra.txt
ABRACADABRA!
```

Bitstream represented as 0 and 1 characters

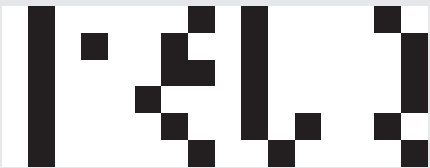
```
% java BinaryDump 16 < abra.txt
0100000101000010
0101001001000001
0100001101000001
0100010001000001
0100001001010010
0100000100100001
96 bits
```

Bitstream represented with hex digits

```
% java HexDump 4 < abra.txt
41 42 52 41
43 41 44 41
42 52 41 21
12 bytes
```

Bitstream represented as pixels in a Picture

```
% java PictureDump 16 6 < abra.txt
```



← 16-by-6 pixel window, magnified

96 bits

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	“	#	\$	%	&	‘	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

Hexadecimal to ASCII conversion table

Universal data compression

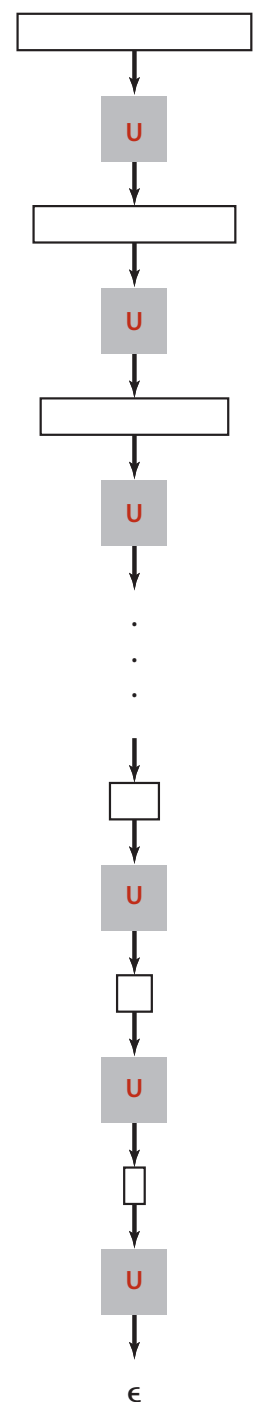
Proposition. No algorithm can compress every bitstring.

Pf 1. [by contradiction]

- Suppose you have a universal data compression algorithm U that can compress every bitstream.
- Given bitstring B_0 , compress it to get smaller bitstring B_1 .
- Compress B_1 to get a smaller bitstring B_2 .
- Continue until reaching bitstring of size 0.
- Implication: all bitstrings can be compressed to 0 bits!

Pf 2. [by counting]

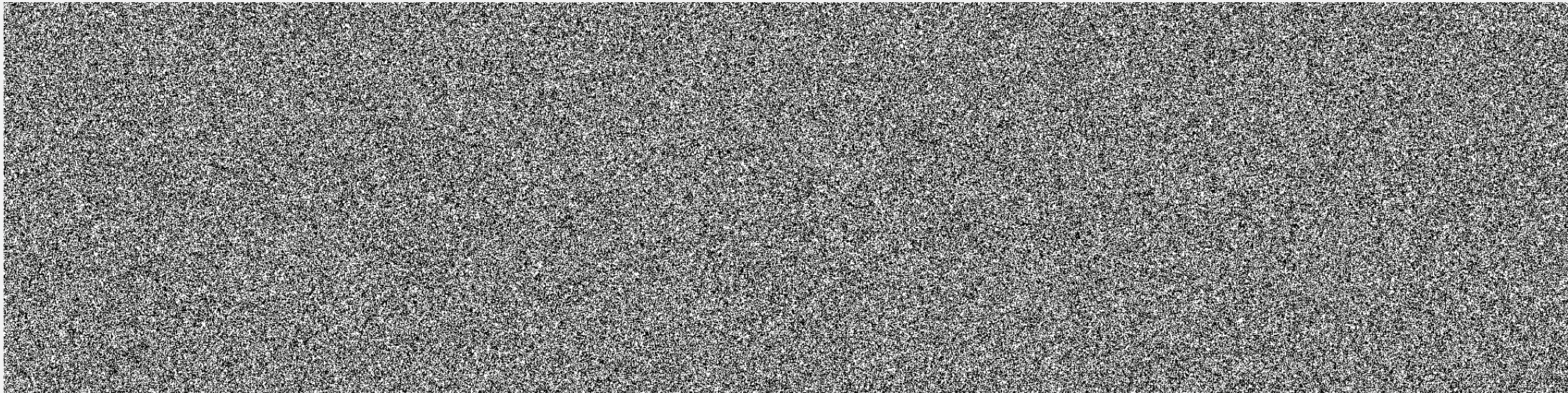
- Suppose your algorithm that can compress all 1,000-bit strings.
- 2^{1000} possible bitstrings with 1,000 bits.
- Only $1 + 2 + 4 + \dots + 2^{998} + 2^{999}$ can be encoded with ≤ 999 bits.



Universal
data compression?

Undecidability

```
% java RandomBits | java PictureDump 2000 500
```



1000000 bits

A difficult file to compress: one million (pseudo-) random bits

```
public class RandomBits
{
    public static void main(String[] args)
    {
        int x = 11111;
        for (int i = 0; i < 1000000; i++)
        {
            x = x * 314159 + 218281;
            BinaryStdOut.write(x > 0);
        }
        BinaryStdOut.close();
    }
}
```

Redundancy in English Language

Q. How much redundancy is in the English language?

“ ... randomising letters in the middle of words [has] little or no effect on the ability of skilled readers to understand the text. This is easy to demonstrate. In a publication of New Scientist you could randomise all the letters, keeping the first two and last two the same, and readability would hardly be affected. My analysis did not come to much because the theory at the time was for shape and sentence recognition. Saberi's work suggests we may have some powerful parallel processors at work. The reason for this is surely that identifying content by parallel processing speeds up recognition. We only need the first and last two letters to spot changes in meaning. ” — *Graham Rawlinson*

A. Quite a bit.

- ▶ basics
- ▶ **run-length coding**
- ▶ Huffman compression
- ▶ LZW compression

Run-length encoding

Simple type of redundancy in a bitstream. Long runs of repeated bits.

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1

Representation. Use 4-bit counts to represent alternating runs of 0s and 1s:
15 0s, then 7 1s, then 7 0s, then 11 1s.

1 1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 ← 16 bits (instead of 40)
15 7 7 11

- Q. How many bits to store the counts?
- A. We'll use 8 (but 4 in the example above).
- Q. What to do when run length exceeds max count?
- A. If longer than 255, intersperse runs of length 0.

Applications. JPEG, ITU-T T4 Group 3 Fax, ...

An application: compress a bitmap

Typical black-and-white-scanned image.

- 300 pixels/inch.
- 8.5-by-11 inches.
- $300 \times 8.5 \times 300 \times 11 = 8.415$ million bits.

Observation. Bits are mostly white.

Typical amount of text on a page.

40 lines \times 75 chars per line = 3,000 chars.

[illegible]

A typical bitmap, with run lengths for each row

- ▶ basics
- ▶ run-length coding
- ▶ **Huffman compression**
- ▶ LZW compression



David Huffman

Variable-length codes

Use different number of bits to encode different chars.

Ex. Morse code: ••• — — — •••

Issue. Ambiguity.

SOS ?
V7 ?
IAMIE ?
EEWNI ?

In practice. Use a medium gap to separate codewords.

codeword for S is a prefix
of codeword for V

Letters		Numbers	
A	• —	1	• — — — —
B	— • • •	2	• • — — —
C	— • — •	3	• • • — —
D	— • •	4	• • • • —
E	•	5	• • • • •
F	• • — •	6	— • • • •
G	— — •	7	— — • • •
H	• • • •	8	— — — • •
I	• •	9	— — — — •
J	• — — —	0	— — — — —
K	— • —		
L	• — • •		
M	— —		
N	— •		
O	— — —		
P	• — — •		
Q	— — • —		
R	• — •		
S	• • •		
T	—		
U	• • —		
V	• • • —		
W	• — —		
X	— • • —		
Y	— • — —		
Z	— — • •		

Variable-length codes

Q. How do we avoid ambiguity?

A. Ensure that no codeword is a **prefix** of another.

Ex 1. Fixed-length code.

Ex 2. Append special stop char to each codeword.

Ex 3. General prefix-free code.

Codeword table

key	value
!	101
A	0
B	1111
C	110
D	100
R	1110

Compressed bitstring

011111110011001000111111100101 ← 30 bits
A B RA CA DA B RA !

Codeword table

key	value
!	101
A	11
B	00
C	010
D	100
R	011

Compressed bitstring

11000111101011100110001111101 ← 29 bits
A B R A C A D A B R A !

Prefix-free codes: trie representation

Q. How to represent the prefix-free code?

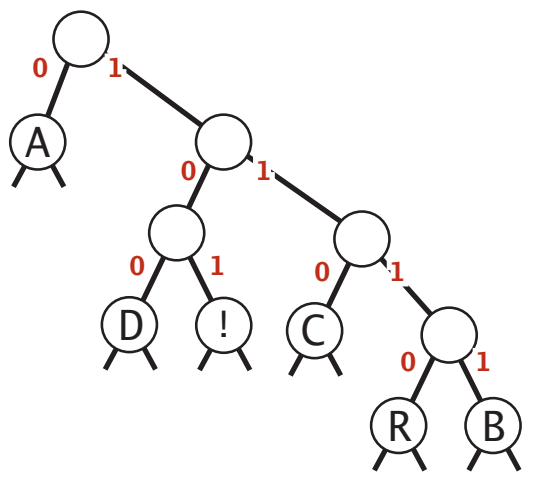
A. A binary trie!

- Chars in leaves.
- Codeword is path from root to leaf.

Codeword table

key	value
!	101
A	0
B	1111
C	110
D	100
R	1110

Trie representation



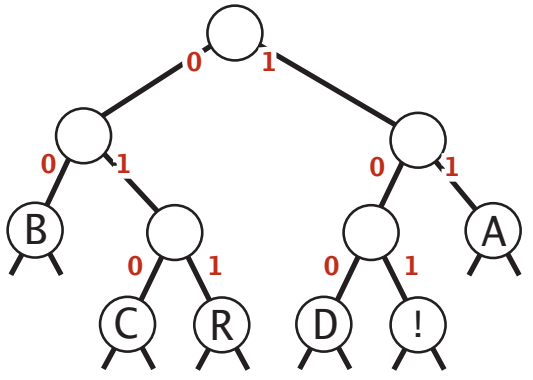
Compressed bitstring

011111110011001000111111100101 ← 30 bits
A B RA CA DA B RA !

Codeword table

key	value
!	101
A	11
B	00
C	010
D	100
R	011

Trie representation



Compressed bitstring

11000111101011100110001111101 ← 29 bits
A B R A C A D A B R A !

Prefix-free codes: compression and expansion

Compression.

- Method 1: start at leaf; follow path up to the root; print bits in reverse.
- Method 2: create ST of key-value pairs.

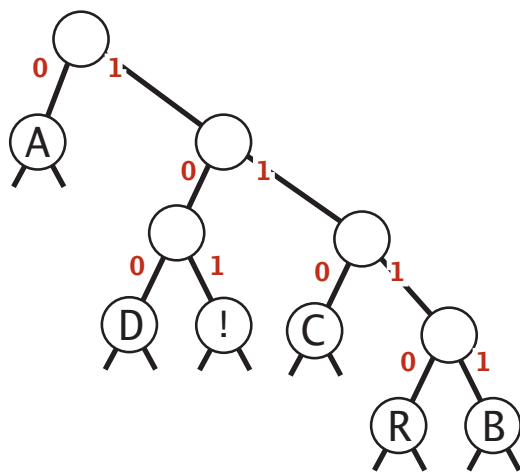
Expansion.

- Start at root.
- Go left if bit is 0; go right if 1.
- If leaf node, print char and return to root.

Codeword table

key	value
!	101
A	0
B	1111
C	110
D	100
R	1110

Trie representation



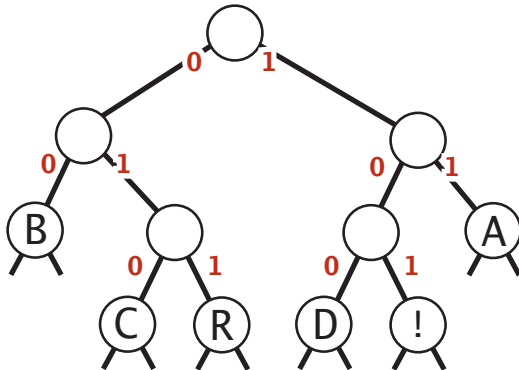
Compressed bitstring

011111110011001000111111100101 ← 30 bits
A B RA CA DA B RA !

Codeword table

key	value
!	101
A	11
B	00
C	010
D	100
R	011

Trie representation



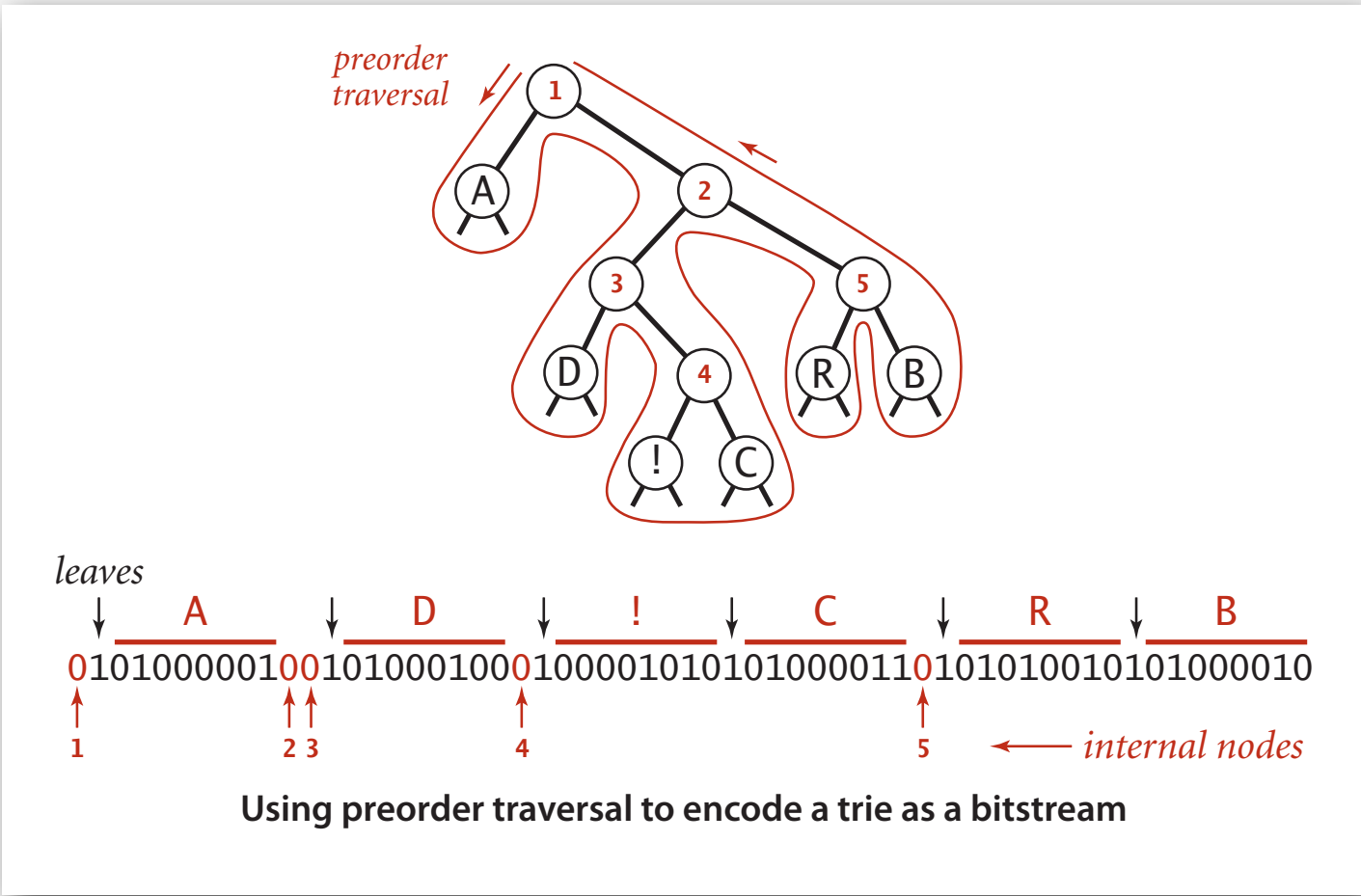
Compressed bitstring

11000111101011100110001111101 ← 29 bits
A B R A C A D A B R A !

Prefix-free codes: how to transmit

Q. How to transmit the trie?

A. Write preorder traversal of trie; mark leaf and internal nodes with a bit (can then reverse process to reconstruct).



Note. If message is long, overhead of transmitting trie is small.

Huffman codes

Q. How to find best prefix-free code?

Huffman algorithm:

- Count frequency $\text{freq}[i]$ for each char i in input.
- Start with one node corresponding to each char i (with weight $\text{freq}[i]$).
- Repeat until single trie formed:
 - select two tries with min weight $\text{freq}[i]$ and $\text{freq}[j]$
 - merge into single trie with weight $\text{freq}[i] + \text{freq}[j]$

Applications:



Huffman encoding summary

Proposition. [Huffman 1950s] Huffman algorithm produces an optimal prefix-free code.

Pf. See textbook.

↑
no prefix-free code uses fewer bits

Implementation.

- Pass 1: tabulate char frequencies and build trie.
- Pass 2: encode file by traversing trie or lookup table.

Running time. Using a binary heap $\Rightarrow N + R \log R$.

↑ ↑
input alphabet
size size

Lossless data compression benchmarks

year	scheme	bits / char
1967	ASCII	7
1950	Huffman	4.7
1977	LZ77	3.94
1984	LZMW	3.32
1987	LZH	3.3
1987	move-to-front	3.24
1987	LZB	3.18
1987	gzip	2.71
1988	PPMC	2.48
1994	SAKDC	2.47
1994	PPM	2.34
1995	Burrows-Wheeler	2.29
1997	BOA	1.99
1999	RK	1.89

data compression using Calgary corpus