Computer Science in Bioinformatics

DAVID L. TABB, PH.D.

Overview

- Metadata: context for interpretation
- Bits are mapped to information
- Compression: efficient space usage
- Scaling: more data, more time

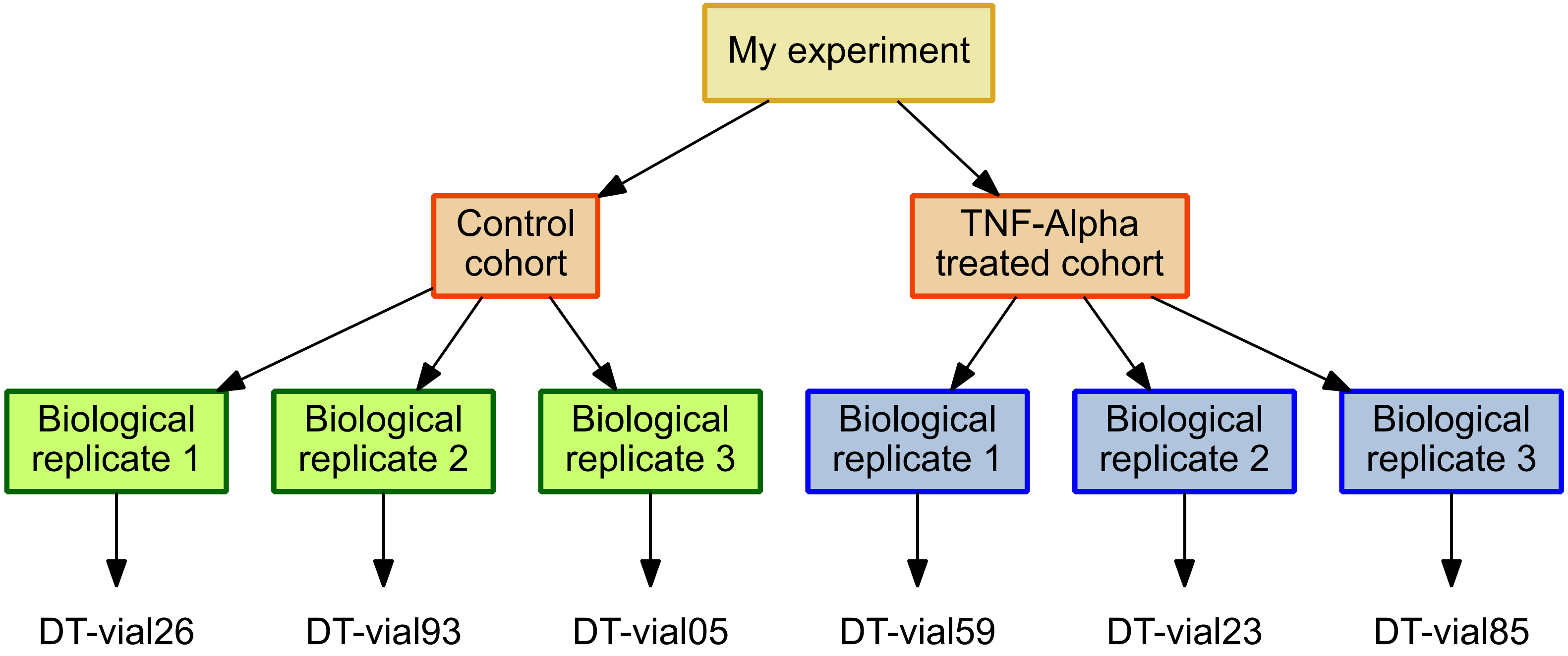
Metadata: describing data

- Metadata are experimental descriptions.
- They contextualize results, making them understandable.
- Metadata are essential for replication of the experiment and the analysis of its outputs.

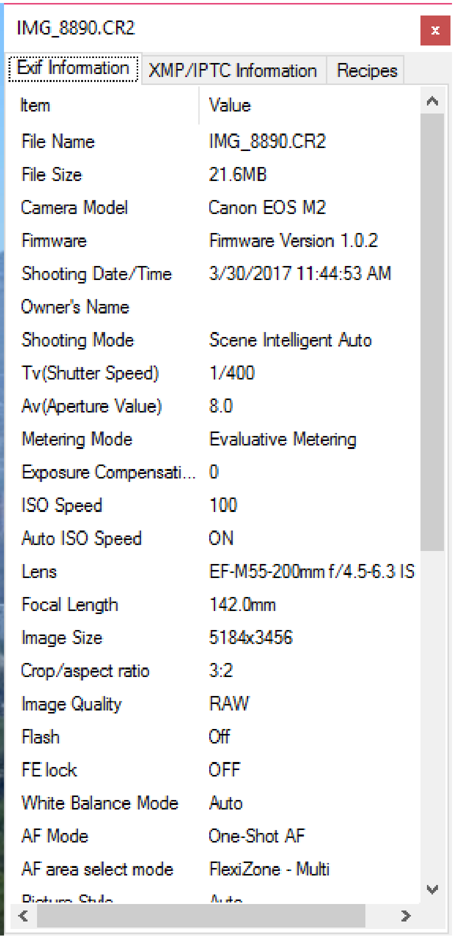
An experiment without its metadata is wasted effort!

https://www.ncbi.nlm.nih.gov/books/NBK53999/

Output files do not tell the full story.



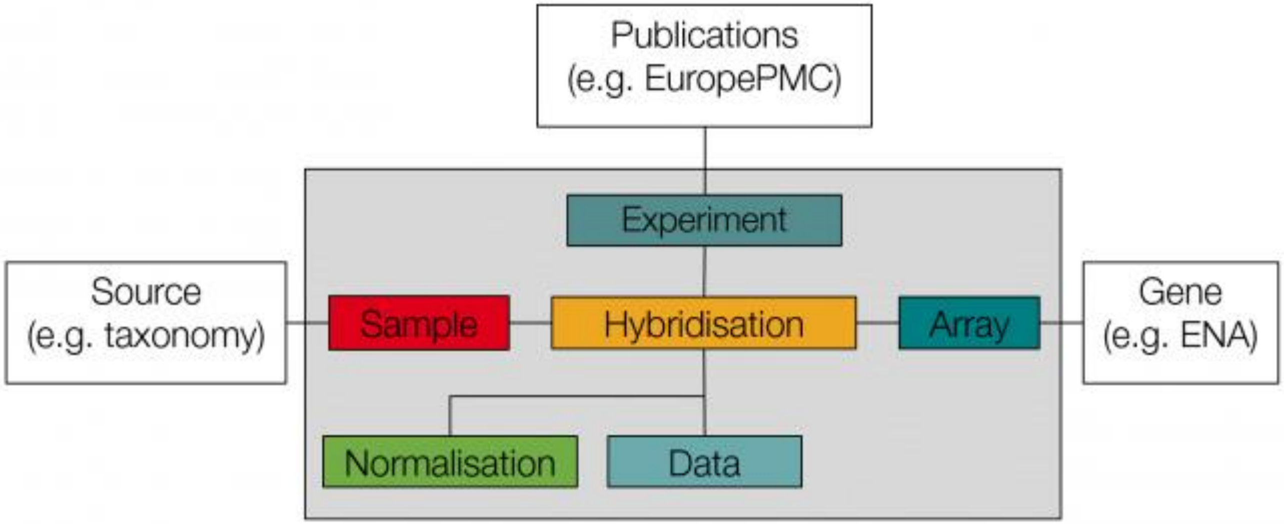
Exif: metadata for photos, sound



Item	Value
File Name	IMG_8890.CR2
File Size	21.6MB
Camera Model	Canon EOS M2
Firmware	Firmware Version 1.0.2
Shooting Date/Time	3/30/2017 11:44:53 AM
Owner's Name	
Shooting Mode	Scene Intelligent Auto
Tv(Shutter Speed)	1/400
Av(Aperture Value)	8.0
Metering Mode	Evaluative Metering
Exposure Compensati...	0
ISO Speed	100
Auto ISO Speed	ON
Lens	EF-M55-200mm f/4.5-6.3 IS
Focal Length	142.0mm
Image Size	5184x3456
Crop/aspect ratio	3:2
Image Quality	RAW
Flash	Off
FE lock	OFF
White Balance Mode	Auto
AF Mode	One-Shot AF
AF area select mode	FlexiZone - Multi
Picture Style	Auto

http://www.geofflawrence.com/photography_tutorial_exif_data.html

MIAME: minimum information standards



“Minimum information standards are sets of guidelines and formats for reporting data derived by specific high-throughput methods. Their purpose is to ensure the data generated by these methods can be easily verified, analysed and interpreted by the wider scientific community”

https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified/minimum-information-standards

Representing numbers in binary

SUMMING POWERS OF TWO

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0	∑
128	64	32	16	8	4	2	1	
1	1	1	1	1	1	1	1	255
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	2
0	0	0	0	0	0	1	1	3
0	0	0	0	0	1	0	0	4
0	0	0	0	0	1	0	1	5
0	0	0	0	0	1	1	0	6
0	0	0	0	0	1	1	1	7
0	0	0	0	1	0	0	0	8

SIGNED? INTEGER? PRECISION?

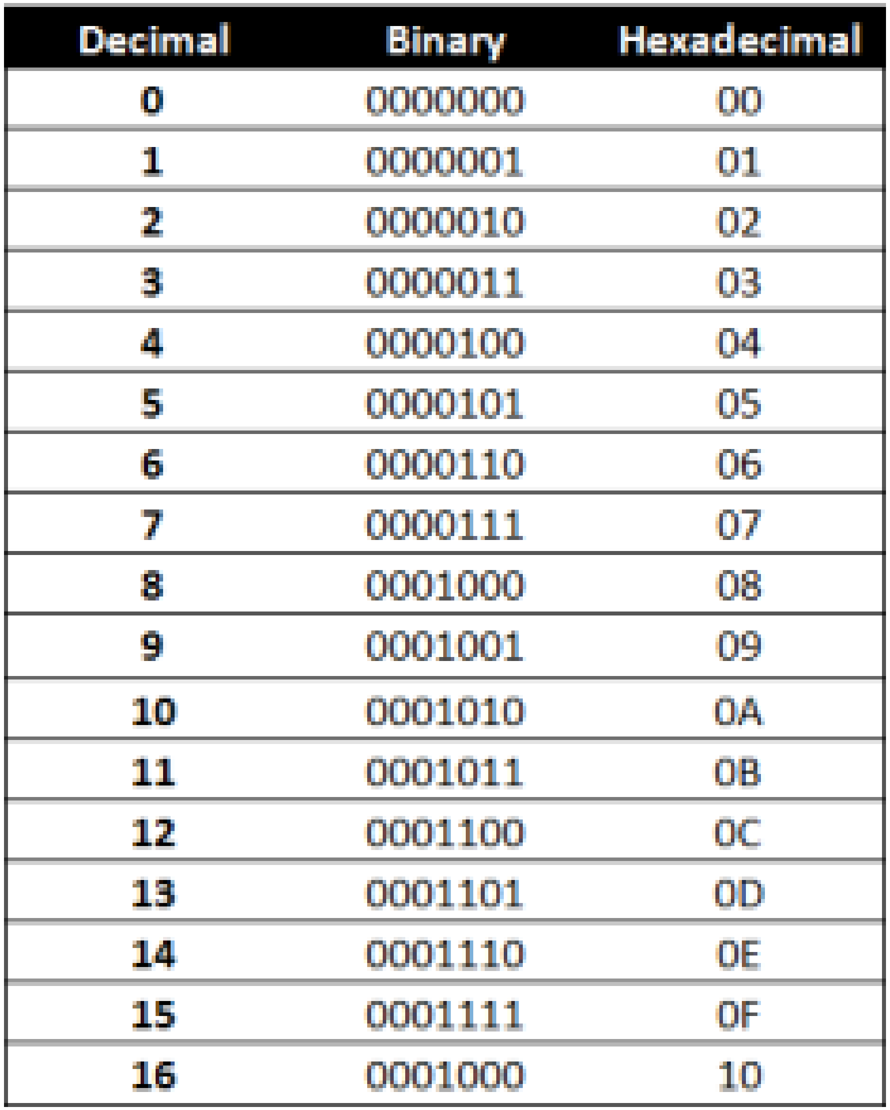
Different answers imply different storage options.

char	1 byte
short int	2 bytes
long int	4 bytes
float	1+8+23 bits
double float	1+11+52 bits

https://www3.ntu.edu.sg/home/ehchua/programming/java/datarepresentation.html

Base what?

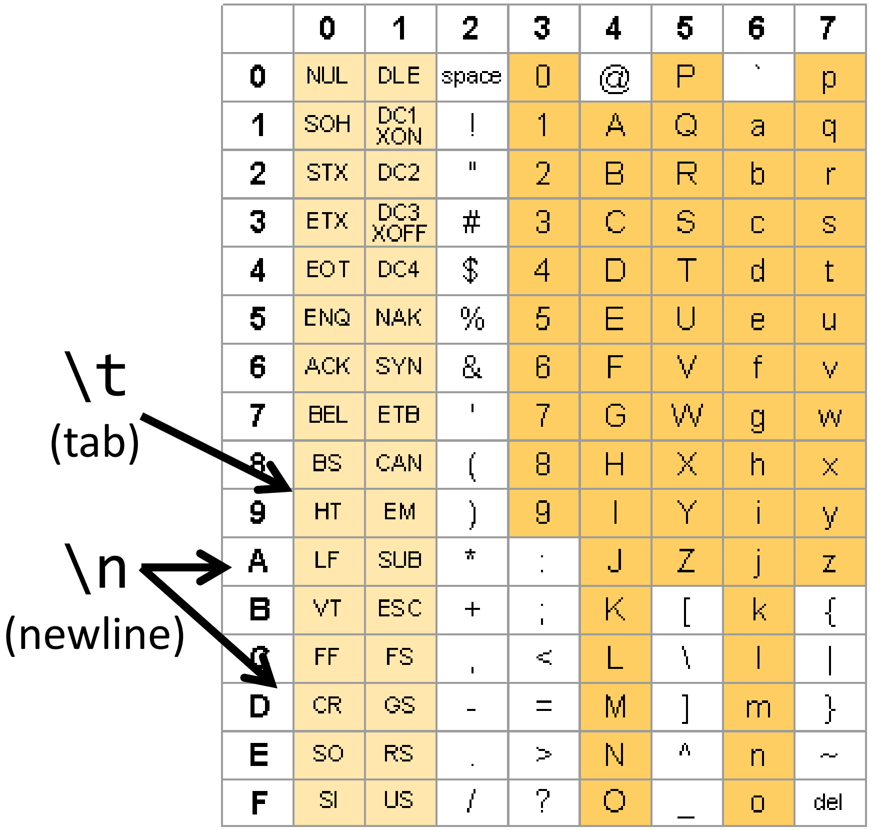
- The base of numbers defines how many symbols are usable in each place.
- The base also defines exponential increase in weight from right to left.



Decimal	Binary	Hexadecimal
0	0000000	00
1	0000001	01
2	0000010	02
3	0000011	03
4	0000100	04
5	0000101	05
6	0000110	06
7	0000111	07
8	0001000	08
9	0001001	09
10	0001010	0A
11	0001011	0B
12	0001100	0C
13	0001101	0D
14	0001110	0E
15	0001111	0F
16	0001000	10

http://www.wirebiters.com/binary-hexadecimal-number-systems/

ASCII: How we map bytes to characters



	0	1	2	3	4	5	6	7
0	NUL	DLE	space	0	@	P	`	p
1	SOH	DC1 XON	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3 XOFF	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	del

\t (tab)

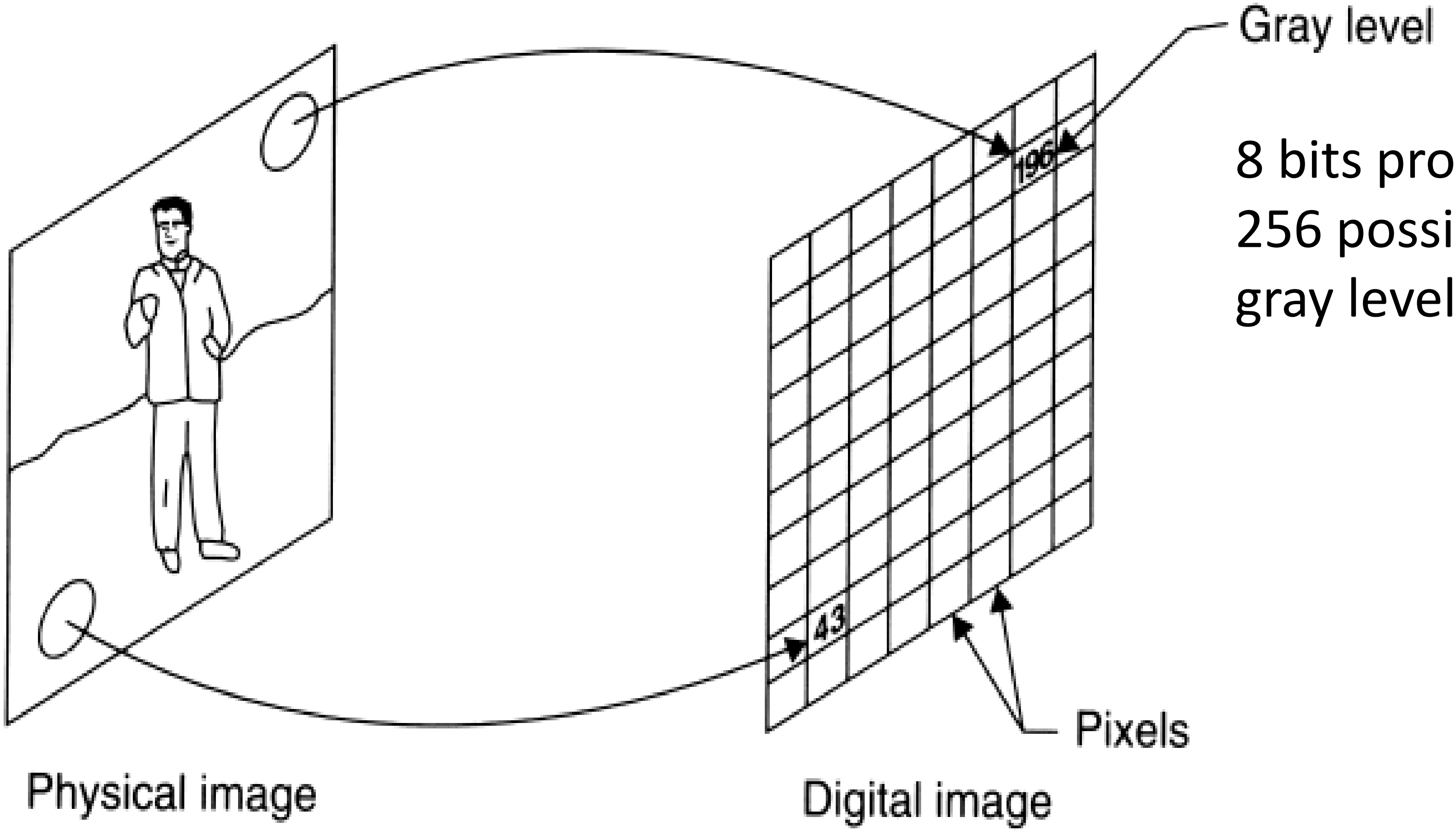
\n (newline)

- In hexadecimal (base 16), A-F follow 9.
- The character ‘A’ is in the fourth column, so it is 4*16+1=65

Column Row

http://ascii-table.com/

Intensity at each *pixel* is recorded in a number of bits. In a typical 24-bit image, 8 bits are recorded for Red, for Green, and for Blue values.

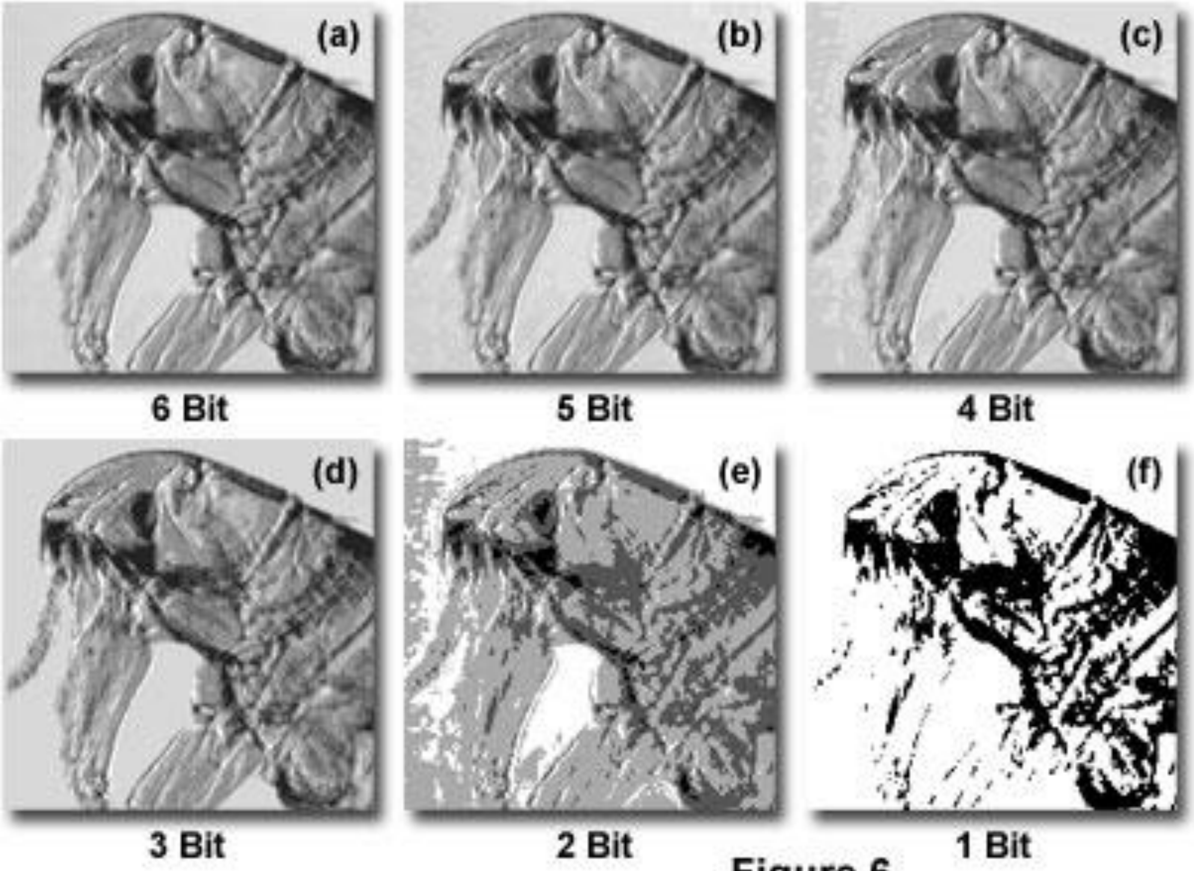


8 bits provide 256 possible gray levels

Figure 1–1 A physical image and a corresponding digital image

Intensity fidelity and bit depth

Grayscale Resolution and Digital Image Appearance



6 Bit 5 Bit 4 Bit

3 Bit 2 Bit 1 Bit

Figure 6

Compression: efficient storage

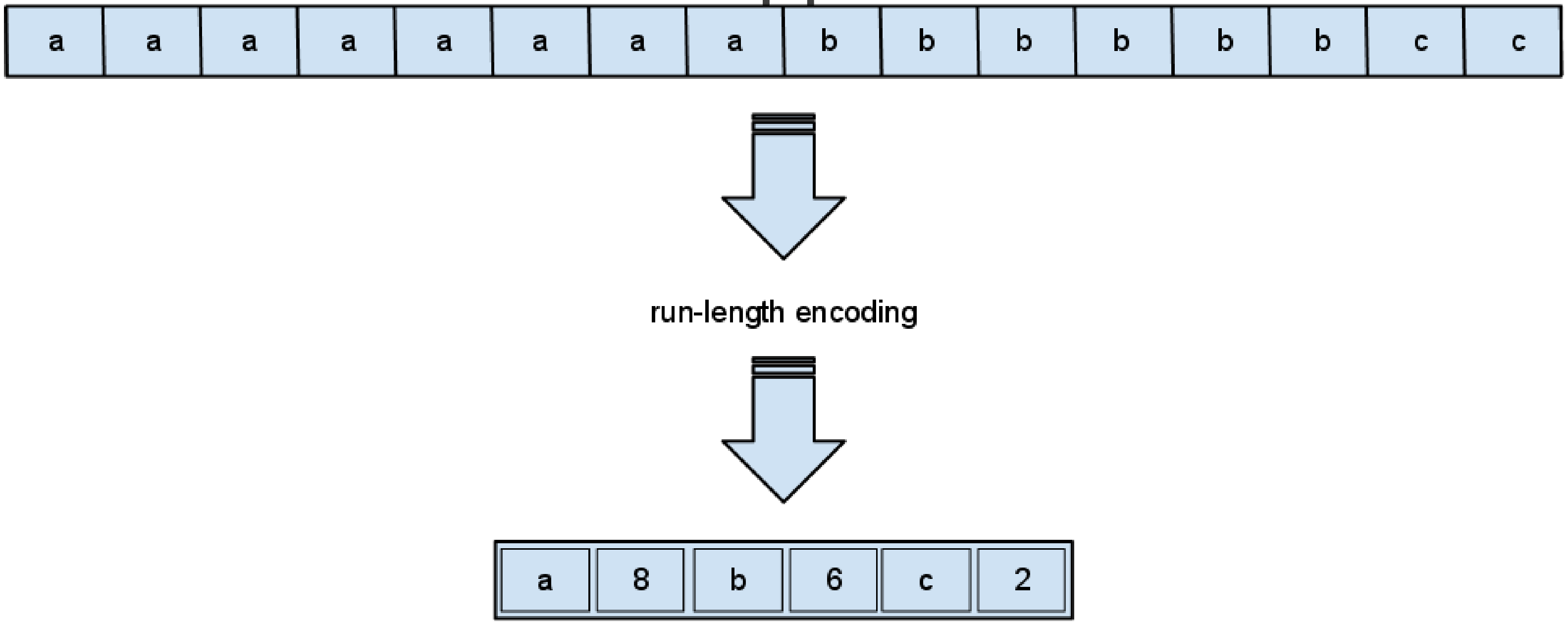
- Biological data are very large. As a result, we frequently use compression to fit data into as small a file size as possible.
- Compression seeks the smallest number of bits or bytes to represent the original content.
- Formats may allow for compression internally.
- Recognize these formats: .zip, .gz, .bz2, .rar

.tar is not compression; instead, it combines many files into one file.

http://www.7-zip.org/

Run-Length Encoding

In data with multiple repeats, one can simply give the symbol and then state how many times in a row it will appear.



a	a	a	a	a	a	a	a	b	b	b	b	b	b	c	c

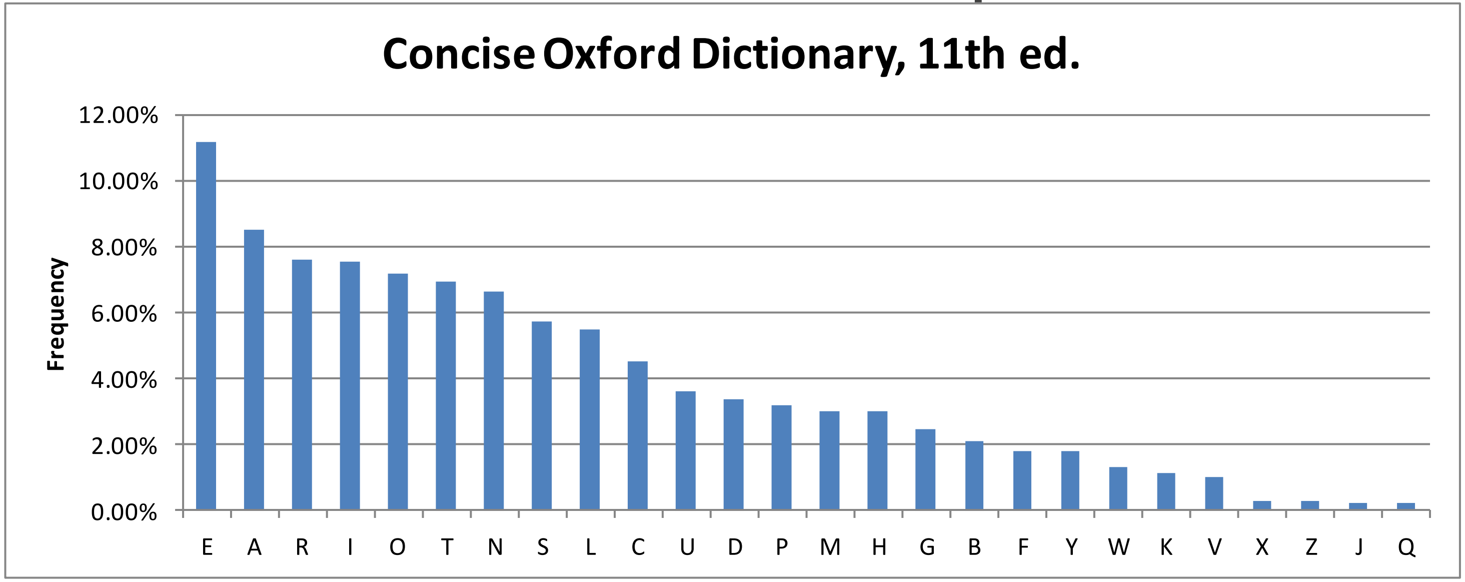
run-length encoding

a	8	b	6	c	2

http://www.stoimen.com/blog/2012/01/09/computer-algorithms-data-compression-with-run-length-encoding/

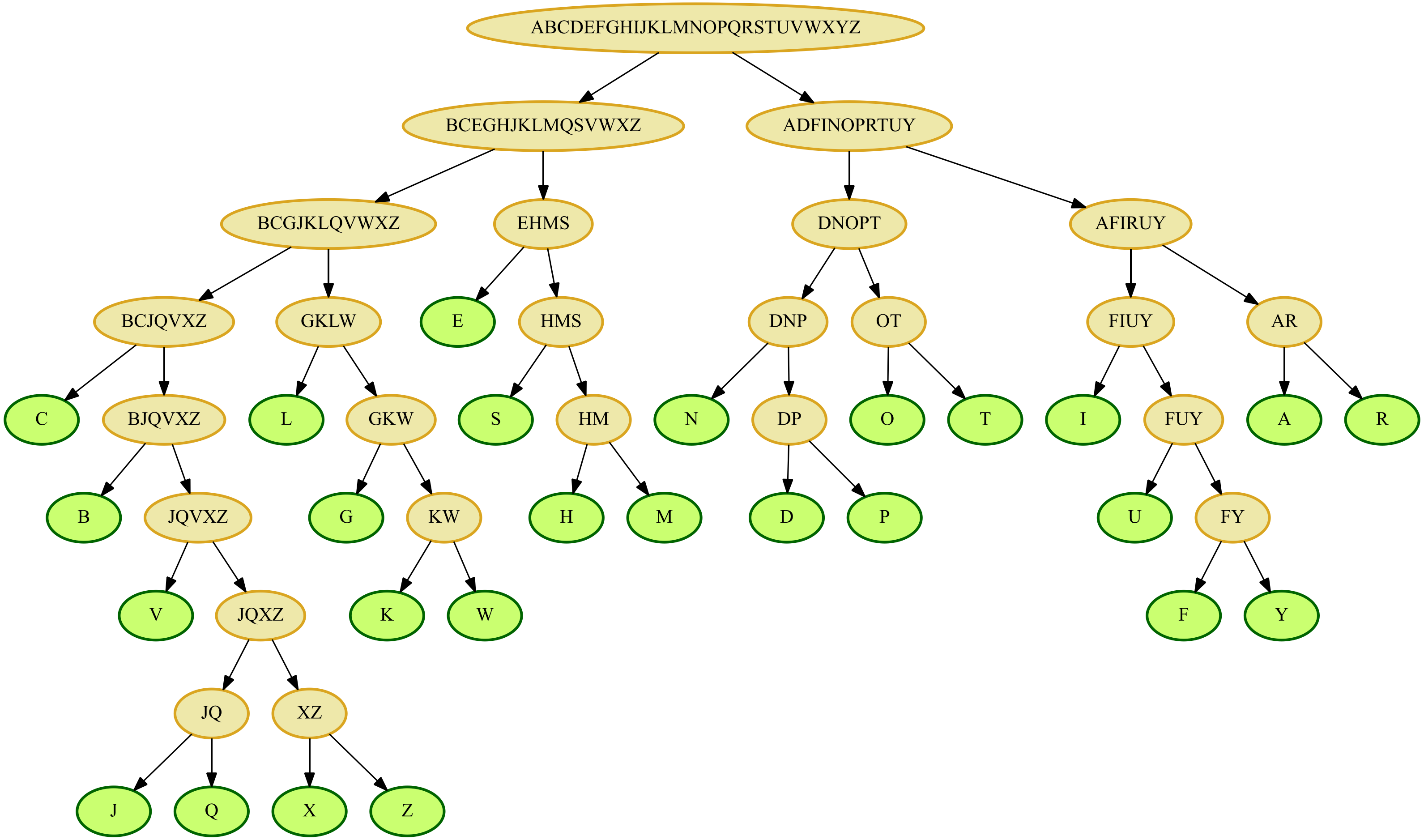
Huffman Codes

In English text, some letters appear more frequently than others; use the smallest number of bits for most frequent letters.



https://en.oxforddictionaries.com/explore/which-letters-are-used-most

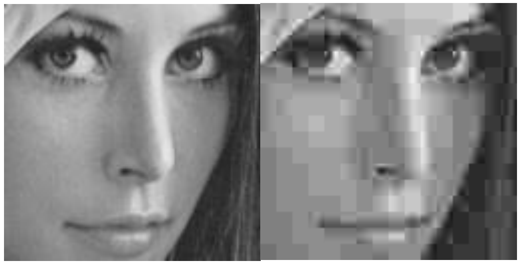
From frequencies to a binary tree: greedy algorithm



More steps away from root → more bits per character

JPG, MPEG: ***Lossy*** compression and artifacts

- Shifting *chrominance* is less perceivable by eye than *luminance* alterations.
- Each save to JPG re-approximates image.
- Image is split to 8x8 blocks, with each subjected to discrete cosine transform (DCT).
- DCT quantizing results in “blocking,” “ringing,” and “mosquitos.”



Lossy media formats: .mp3, .jpg, .mov, .avi

http://www.utdallas.edu/~aria/mcl/post/

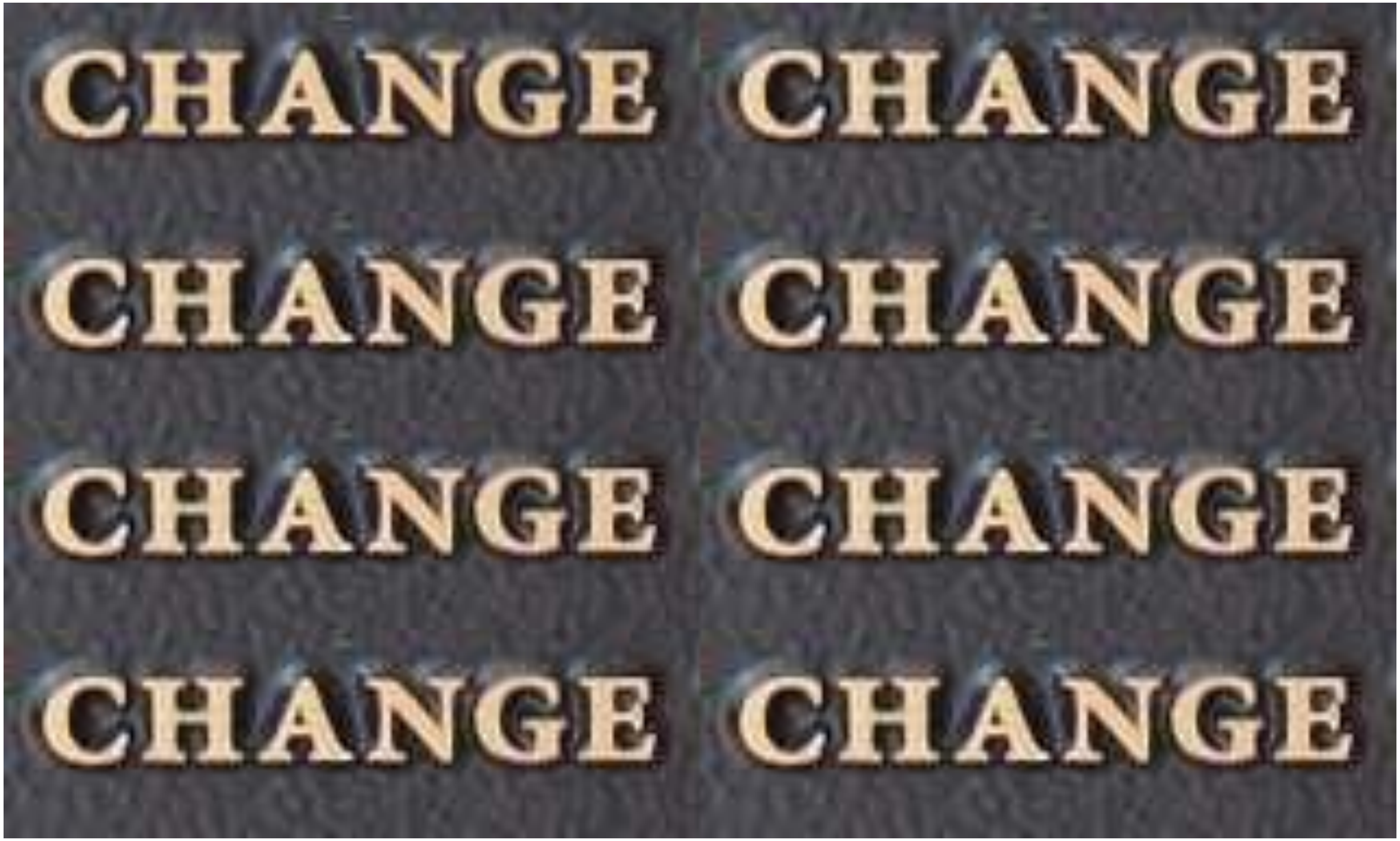
Re-saving JPGs has a cost

Original TIFF

First JPEG

Second JPEG

…



Big O notation: worst-case algorithmic efficiency

- $O(\log_2 n)$: run time scales with log of data size
  - Binary search for a value in a binary search tree
- O(n): run time scales linearly with data size
  - Add 2 to every item in this vector
- O(mn): run time is multiple of two data sizes
  - Multiply each in A with each in B
- $O(2^n)$: run time scales exponentially with data
  - What is shortest possible trip through cities?

http://www.daveperrett.com/articles/2010/12/07/comp-sci-101-big-o-notation/

Closing thoughts

- Systems biology employs bioinformatics for initial pre-processing, follow-on summarization, established knowledge representation, and visualization.
- Bioinformatics draws upon diverse computer science fundamentals to accomplish its aims.
- Biologists need to understand some computer science. Computers scientists need to learn some biology!