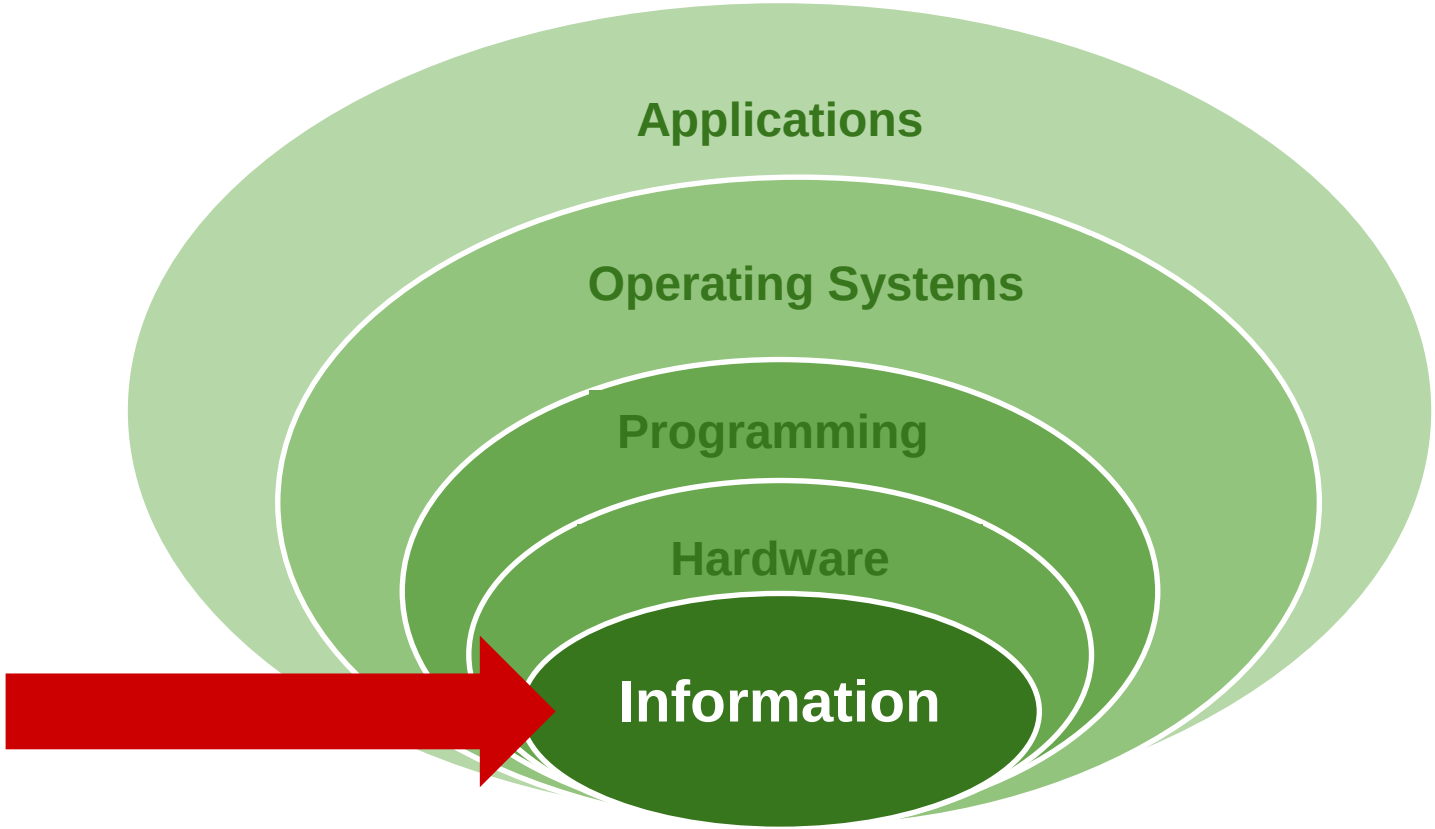
Computer Science Fundamentals

Lecture 4 Data representation

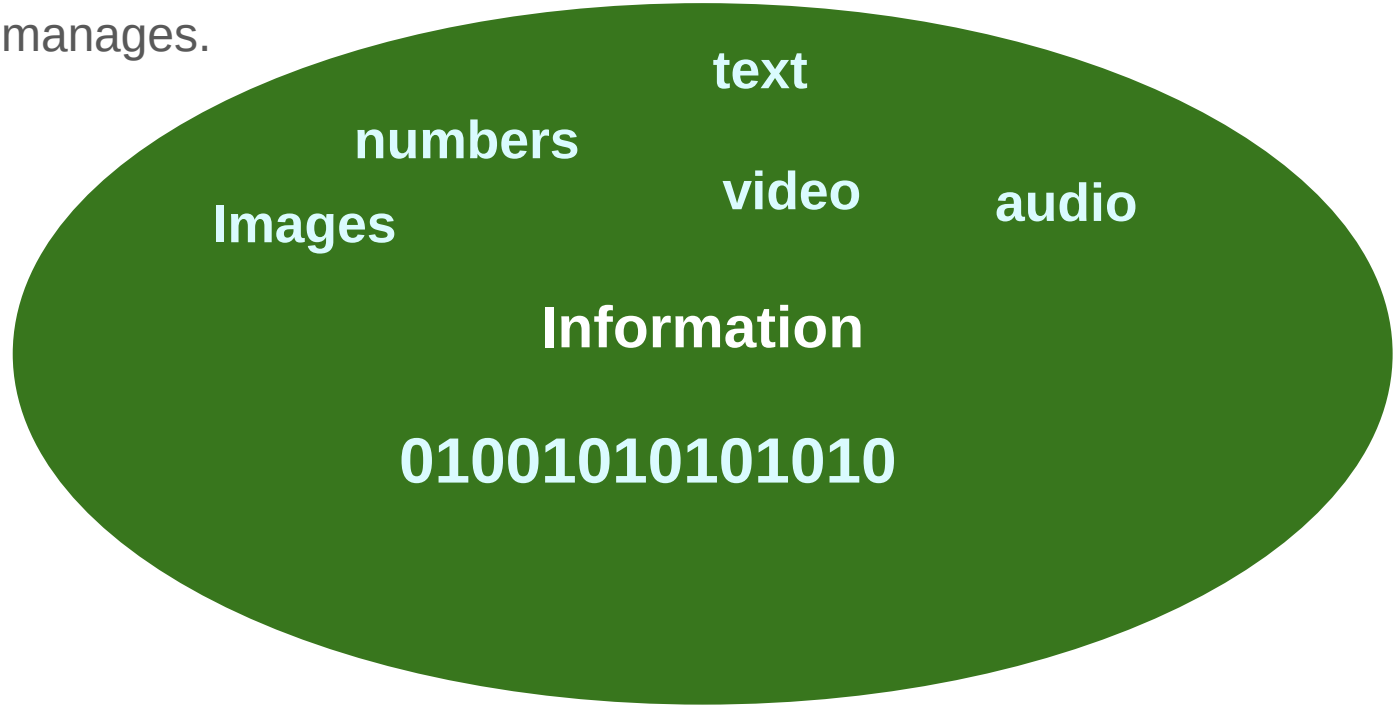
Part A

Today’s focus



Today’s focus

Understand how to **represent** and **store** the various kinds of information a computer manages.



Agenda

- Analog vs Digital information
- Representation of different data types on a computer
  - Numbers
  - Text
  - Audio
  - Images and graphics
  - Video
- Data compression

Some background: analog vs digital

Background

- In the not-so-distant past, computers dealt almost exclusively with numeric and textual data
- But now computers are truly multimedia devices. They can store, present, and help us modify many different types of data:
  - Numbers
  - Text
  - Audio
  - Images and graphics
  - Video
- All of this data is stored as binary digits, i.e. strings of **0**s and **1**s.

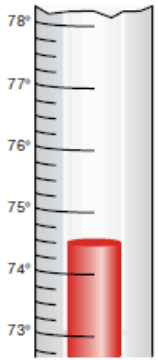
1001010101010010010101010010010101010 [1]

[1] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Analog and Digital Information

Information can be represented in one of two ways:

Analog data is a continuous representation, analogous to the actual information it represents. [2]



Digital data is a discrete representation, breaking the information up into separate elements. [2]

[2] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Analog and Digital Information

- Computers cannot work well with analog information. So instead, we **digitize** information by breaking it into pieces and representing those pieces separately.
- Electronic signals are far easier to maintain if they transfer only binary data.
  - An analog signal continually fluctuates in voltage up and down.
  - Digital signal has only a high or low state, corresponding to the two binary digits. [3]

[3] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Binary representations

Intro

Khan Academy and Code.org | Binary & Data [4]

[4] Khan Academy and Code.org | Binary & Data. 2022. [online] Available at: <https://www.youtube.com/watch?v=ewokFOSxabs> [Accessed 19 April 2022].

Binary representations

- One bit can be either 0 or 1. There are no other possibilities. Therefore, one bit can represent only two things.
  - For example, if we wanted to classify a food as being either sweet or sour, we would need only 1 bit
- What if we need to represent gear of a car (park, drive, reverse, or neutral)?
  - 2 bits for four different states

    00 - park

    01 - drive

    10 - reverse

    11 - neutral [5]

[5] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Bit combinations

1 bit = 2^1 = 2 combinations

2 bit = 2^2 = 4 combinations

3 bit = 2^3 = 8 combinations

4 bit = 2^4 = 16 combinations

And etc.

==>

n bits can represent 2^n things because 2^n combinations of 0 and 1 can be made from n bits [6]

1 Bit	2 Bits	3 Bits	4 Bits	5 Bits
0	00	000	0000	00000
1	01	001	0001	00001
	10	010	0010	00010
	11	011	0011	00011
		100	0100	00100
		101	0101	00101
		110	0110	00110
		111	0111	00111
			1000	01000
			1001	01001
			1010	01010
			1011	01011
			1100	01100
			1101	01101
			1110	01110
			1111	01111
				10000
				10001
				10010
				10011
				10100
				10101
				10110
				10111
				11000
				11001
				11010
				11011
				11100
				11101
				11110
				11111

[6] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Bit combinations

- How many bits are needed to represent 25 unique states?

- How many bits are needed to represent “regions” of Uzbekistan, a.k.a. “tumans” or “viloyats”?

Bit combinations

Keep in mind that even though we may technically need only a certain minimum number of bits to represent a set of items, we may allocate more than that for the storage of data.

There is a minimum number of bits that a computer architecture can address and move around at one time, and it is usually a power of 2, such as 8, 16, or 32 bits.

==> the minimum amount of storage given to any type of data is allocated in multiples of that value.

Representing text

Representing text

- Did you ever get an email with the subject line “???? ?????? ??? ????”?

  => caused by incorrect encoding settings

Character encoding

..is a mechanism which tells the computer **how to interpret** raw zeroes and ones into real characters.

How the computer does it?

- By **pairing** numbers with characters

Decimal - Binary - Octal - Hex – ASCII Conversion Chart

Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII	Decimal	Binary	Octal	Hex	ASCII
0	00000000	000	00	NUL	32	00100000	040	20	SP	64	01000000	100	40	@	96	01100000	140	60	`
1	00000001	001	01	SOH	33	00100001	041	21	!	65	01000001	101	41	A	97	01100001	141	61	a
2	00000010	002	02	STX	34	00100010	042	22	"	66	01000010	102	42	B	98	01100010	142	62	b
3	00000011	003	03	ETX	35	00100011	043	23	#	67	01000011	103	43	C	99	01100011	143	63	c
4	00000100	004	04	EOT	36	00100100	044	24	$	68	01000100	104	44	D	100	01100100	144	64	d
5	00000101	005	05	ENQ	37	00100101	045	25	%	69	01000101	105	45	E	101	01100101	145	65	e
6	00000110	006	06	ACK	38	00100110	046	26	&	70	01000110	106	46	F	102	01100110	146	66	f
7	00000111	007	07	BEL	39	00100111	047	27	'	71	01000111	107	47	G	103	01100111	147	67	g
8	00001000	010	08	BS	40	00101000	050	28	(	72	01001000	110	48	H	104	01101000	150	68	h
9	00001001	011	09	HT	41	00101001	051	29	)	73	01001001	111	49	I	105	01101001	151	69	i
10	00001010	012	0A	LF	42	00101010	052	2A	*	74	01001010	112	4A	J	106	01101010	152	6A	j
11	00001011	013	0B	VT	43	00101011	053	2B	+	75	01001011	113	4B	K	107	01101011	153	6B	k
12	00001100	014	0C	FF	44	00101100	054	2C	,	76	01001100	114	4C	L	108	01101100	154	6C	l
13	00001101	015	0D	CR	45	00101101	055	2D	-	77	01001101	115	4D	M	109	01101101	155	6D	m
14	00001110	016	0E	SO	46	00101110	056	2E	.	78	01001110	116	4E	N	110	01101110	156	6E	n
15	00001111	017	0F	SI	47	00101111	057	2F	/	79	01001111	117	4F	O	111	01101111	157	6F	o
16	00010000	020	10	DLE	48	00110000	060	30	0	80	01010000	120	50	P	112	01110000	160	70	p
17	00010001	021	11	DC1	49	00110001	061	31	1	81	01010001	121	51	Q	113	01110001	161	71	q
18	00010010	022	12	DC2	50	00110010	062	32	2	82	01010010	122	52	R	114	01110010	162	72	r
19	00010011	023	13	DC3	51	00110011	063	33	3	83	01010011	123	53	S	115	01110011	163	73	s
20	00010100	024	14	DC4	52	00110100	064	34	4	84	01010100	124	54	T	116	01110100	164	74	t
21	00010101	025	15	NAK	53	00110101	065	35	5	85	01010101	125	55	U	117	01110101	165	75	u
22	00010110	026	16	SYN	54	00110110	066	36	6	86	01010110	126	56	V	118	01110110	166	76	v
23	00010111	027	17	ETB	55	00110111	067	37	7	87	01010111	127	57	W	119	01110111	167	77	w
24	00011000	030	18	CAN	56	00111000	070	38	8	88	01011000	130	58	X	120	01111000	170	78	x
25	00011001	031	19	EM	57	00111001	071	39	9	89	01011001	131	59	Y	121	01111001	171	79	y
26	00011010	032	1A	SUB	58	00111010	072	3A	:	90	01011010	132	5A	Z	122	01111010	172	7A	z
27	00011011	033	1B	ESC	59	00111011	073	3B	;	91	01011011	133	5B	[	123	01111011	173	7B	{
28	00011100	034	1C	FS	60	00111100	074	3C	<	92	01011100	134	5C	\	124	01111100	174	7C	|
29	00011101	035	1D	GS	61	00111101	075	3D	=	93	01011101	135	5D	]	125	01111101	175	7D	}
30	00011110	036	1E	RS	62	00111110	076	3E	>	94	01011110	136	5E	^	126	01111110	176	7E	~
31	00011111	037	1F	US	63	00111111	077	3F	?	95	01011111	137	5F	_	127	01111111	177	7F	DEL

[7] Internet Archive. 2022. ASCII Conversion Chart : Decimal - Binary - Octal - Hex – ASCII Conversion Chart : Free Download, Borrow, and Streaming : Internet Archive. [online] Available at: <https://archive.org/details/ASCIIConversionChart/mode/1up> [Accessed 19 April 2022].

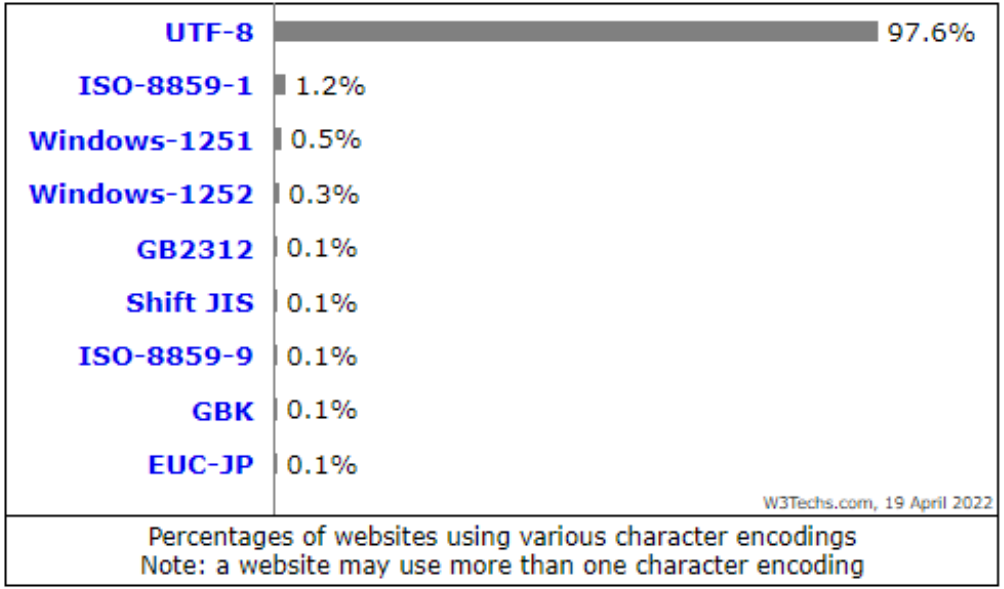
Character set

- **Character set** – a list of characters and the codes to represent each one
- **ASCII** – American Standard Code for Information Interchange
  - 8 bit per character
  - Limited to 128 unique characters (8 bits >> 2^7)
- **Unicode Standard**
  - Superset of ASCII (256 characters in unicode correspond to ASCII character set)
  - Represent every character in every language used in the entire world and special scientific symbols
  - **UTF-8 - one of most popular encodings defined by the Unicode Standard** [8]

[8] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Usage Statistics and Market Share of Character

UTF-8 is used by 97.6% of all the websites whose character encoding we know. [9]



[9] W3techs.com. 2022. *Usage Statistics and Market Share of Character Encodings for Websites, April 2022*. [online] Available at: <https://w3techs.com/technologies/overview/character_encoding> [Accessed 19

Unicode character set

- **UTF-8** – character encoding defined by Unicode
- **Goal**: represent every character in every language used in entire world and scientific symbols
- Unicode is a **superset** of ASCII
  - 256 characters in the Unicode character set correspond exactly to the extended ASCII character set [10]

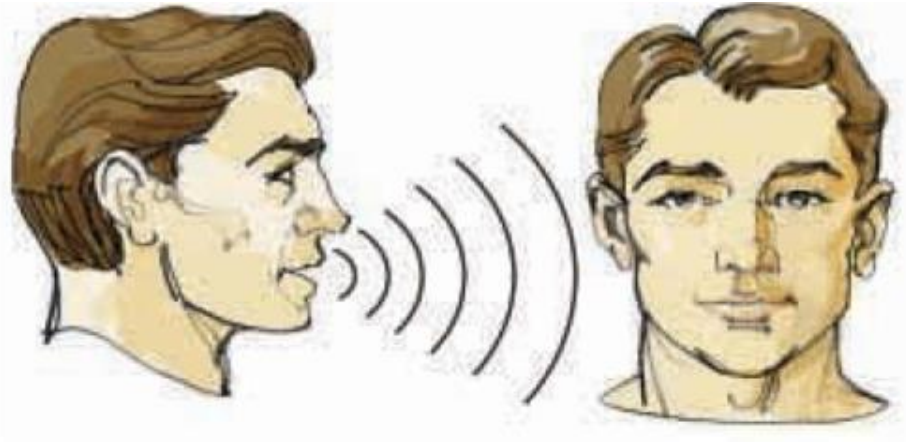
Code (Hex)	Character	Source
0041	A	English (Latin)
042F	Я	Russian (Cyrillic)
OE09	ก	Thai
13EA	Ꮹ	Cherokee
211E	℞	Letterlike Symbols
21CC	⇌	Arrows
282F	⠯	Braille
345F	㑯	Chinese/Japanese/ Korean (Common)

[10] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing audio data

Representing audio data

- A series of air compressions vibrate a membrane in our ear, which sends signals to our brain. Thus a sound is defined in nature by the wave of air that interacts with our eardrum. [11]



[10] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

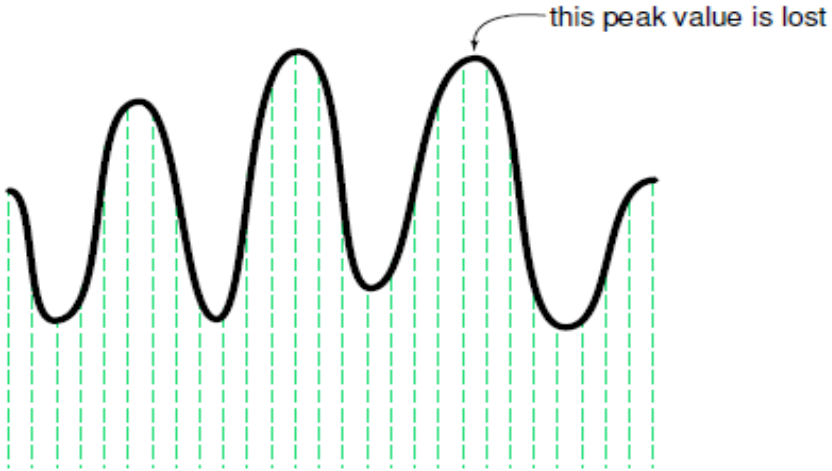
Representing audio data

- To represent audio information on a computer, we must digitize the sound wave, somehow breaking it into discrete, manageable pieces.
- To digitize the signal we periodically measure the voltage of the signal and record the appropriate numeric value - *sampling*.
- sampling rate of around 40,000 times per second is enough to create a reasonable sound reproduction. [11]

[11] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing audio data

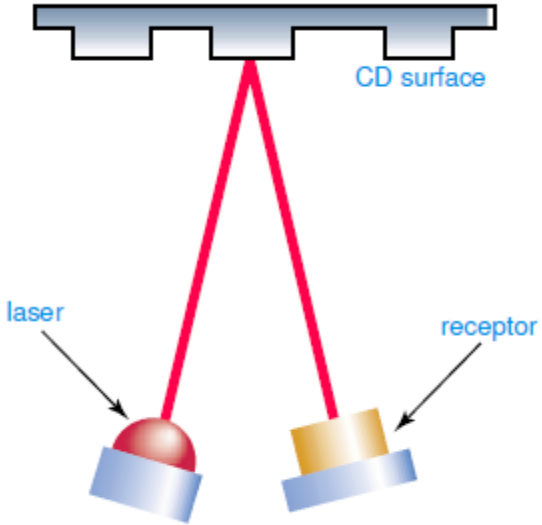
- Sampling an audio signal [12]



[12] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing audio data

- CD surface contains microscopic pits that represent binary digits.
- A low intensity laser is pointed at the disk.
  - If surface smooth = laser reflects strongly
  - If surface pitted = laser reflects poorly
- A receptor analyzes the reflection and produces appropriate string of binary data
- The signal is reproduced and sent to the speaker [13]



[13] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing images and graphics

Representing images and graphics

- Our retinas have three types of color photoreceptor cone cells that respond to different sets of frequencies (red, green, and blue)
- On computer colour represented as an RGB
  - Three numbers that indicate the relative contribution of each of these three primary colours

  0 no contribution

  255 full contribution [14]

RGB Value			Actual Color
Red	Green	Blue	
0	0	0	black
255	255	255	white
255	255	0	yellow
255	130	255	pink
146	81	0	brown
157	95	82	purple
140	0	0	maroon

[14] Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Digitized images and graphics

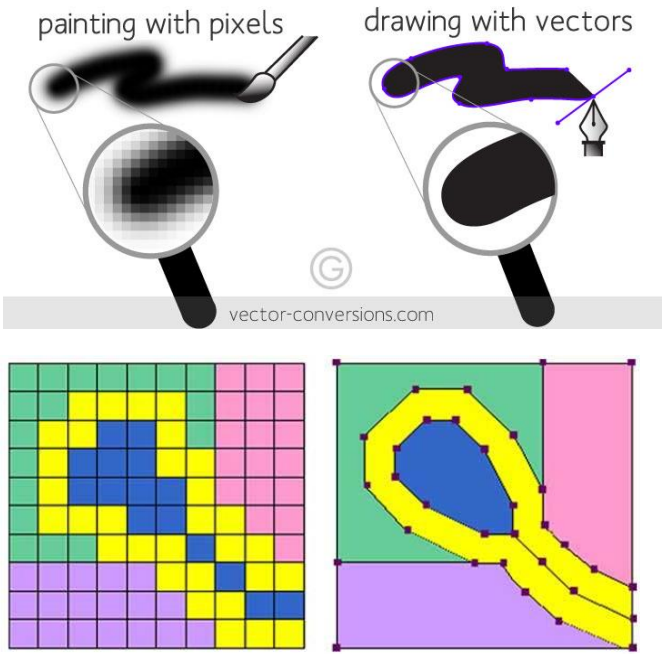
- **Pixels** - Individual dot used to represent a picture; stands for picture elements
  - Each pixel composed of a single colour
- **Resolution** - The number of pixels used to represent a picture [15]

[15]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Graphics formats

- **Raster-graphics format**
  - Storing image information pixel by pixel
  - E.g. BMP, GIF, JPEG
  - Photoshop – famous raster graphics editor
- **Vector graphics format**
  - Representation of an image in terms of lines and shapes
  - Use formula to calculate the shape and represent image
  - E.g. PDF, SVG
  - Corel draw – famous vector graphics editor

Raster vs Vector



[16] Vector-conversions.com. 2022. *Raster (Bitmap) vs Vector*. [online] Available at: <https://vector-conversions.com/vectorizing/raster_vs_vector.html> [Accessed 19 April 2022].

SVG file

An **SVG file** is a graphics **file** that uses a two-dimensional vector graphic **format** created by the World Wide Web Consortium (W3C).

images are described using a text **format** that is based on XML [16]

[16]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing video

Representing video

- Video information is one of the most complex types of information to capture and compress to get a result that makes sense to the human eye.

- Video clips contain the equivalent of many **still** images, each of which must be **compressed**. [17]

[17]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Representing video

- **Video codec** refers to the methods used to **shrink** the size of a movie to allow it to be played on a computer or over a network.
- Almost all video codecs use lossy compression to minimize the huge amounts of data associated with video.
- The goal therefore is not to lose information that affects the viewer's senses. [18]

[18]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Video summary

Read Ch 3, and watch the following video to understand more about representing numbers

Representing Numbers and Letters with Binary: Crash Course Computer Science #4

Data compression

Data compression

- **Data compression** - reducing the amount of space needed to store a piece of data.
- **Compression ratio** - The size of the compressed data divided by the size of the uncompressed data. The ratio should result in a number between 0 and 1. The closer the ratio is to zero, the tighter the compression.
- **Lossy compression** - A technique in which there is loss of information
- **Lossless compression -** A technique in which there is no loss of information
- **Bandwidth restrictions** - the maximum number of bits or bytes that can be transmitted from one place to another in a fixed amount of time [19]

[19]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Text compression

- Keyword encoding
  - Replacing a frequently used word (or part of the word) with a single character
- Run-length encoding
  - Replacing a long series of a repeated character with a count of the repetition
- Huffman encoding
  - Using a variable-length binary string to represent a character so that frequently used characters have short codes [20]

[20]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Keyword encoding

"The human body is composed of many independent systems, such as the circulatory system, the respiratory system, and the reproductive system. Not only must all systems work independently, they must interact and cooperate as well. Overall health is a function of the well-being of separate systems, as well as how these separate systems work in concert."

TOTAL: 352 characters

Encoded version

"The human body is composed of many independent systems, such ^ ~ circulatory system, ~ respiratory system, + ~ reproductive system. Not only & each system work independently, they & interact + cooperate ^ %. Overall health is a function of ~ %-being of separate systems, ^ % ^ how # separate systems work in concert." [21]

TOTAL: 317 characters

Compression rate: 317/352=0.9

Word	Symbol
as	^
the	~
and	+
that	$
must	&
well	%
those	#

[21]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Run-length encoding

- A.k.a. recurrence coding
- Sequence of repeated characters is replaced by flag character, followed by the repeated character, followed by a single digit that indicates how many times the character is repeated [22]
- E.g. * is a flag
  - AAAAAAA -> *A7

[22]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Run-length encoding

*n5*x9ccc*h6 some other text *k8eee

TOTAL: 35 characters

Decoded as:

nnnnnxxxxxxxxxccchhhhhh some other text kkkkkkkkeee

TOTAL: 51 characters

compression ratio 35/51 = 0.68.

Run-length encoding for bitmap [23]

- In raw format the first three rows would be

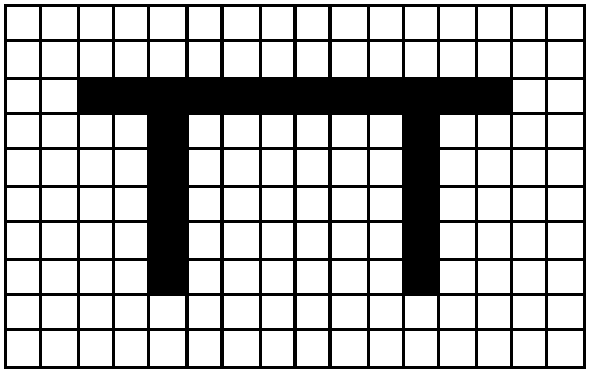
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

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

- Using run length encoding the first three rows would be

16 0 16 0 2 0 12 1 2 0



[23]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Huffman encoding

- Use **variable-length** bit strings to represent each character
- Use only few bits to represent frequently used characters
- Use longer bit strings for rarely used characters [24]
- E.g.

DOORBELL -> 10111101101111010011001100

Huffman Code	Character
00	A
01	E
100	L
110	O
111	R
1010	B
1011	D

[24]Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Ch3

Decoding Huffman code

- Since the variable length encoding is used, won’t we get confused when trying to decode a string?
  - we do not know how many bits we should include for each character!!
  - Read the book and get the answer ready for the tutorial! ☺

Part B:

A development plan (one of several approaches)

A development plan is a process for writing programs.

1. Start by writing a small program with no function definitions.

2. Once you get the program working, identify a coherent piece of it, encapsulate the piece in a function and give it a name.

3. Generalize the function by adding appropriate parameters.

4. Repeat steps 1–3 until you have a set of working functions. Copy and paste working code to avoid retyping (and re-debugging).

5. Look for opportunities to improve the program by refactoring. For example, if you have similar code in several places, consider factoring it into an appropriately general function.

docstring

A docstring is a string at the beginning of a function that explains the interface (“doc” is short for “documentation”). Here is an example:

```
def square(n):

    '''Takes in a number n, returns  the square of n'''

    return n**2

print(square.__doc__)
```

By convention, all docstrings are triple-quoted strings, also known as multiline strings because the triple quotes allow the string to span more than one line.

It explains what effect each parameter has on the behavior of the function and what type each parameter should be

Homework

- Dale, Computer Science Illuminated, Ch3
  - End of the chapter questions and exercises
- Representing Numbers and Letters with Binary: Crash Course Computer Science #4
- Character Encoding in Depth
- Downey, A. (2015). Think Python, How to think like a computer scientist. Chapter 3, 4

References

- Dale, N., & Lewis, J. (2020). Computer Science Illuminated (7th ed.). Jones & Bartlett Learning, Chapter 2
- Downey, A. B., Elkner, J., & Meyers, C. (2015). Learning with python: How to think like a computer scientist. Green Tea Press., Chapter 1
- Khan Academy and Code.org | Binary & Data. 2022. [online] Available at: <https://www.youtube.com/watch?v=ewokFOSxabs> [Accessed 19 April 2022].
- Internet Archive. 2022. ASCII Conversion Chart : Decimal - Binary - Octal - Hex – ASCII Conversion Chart : Free Download, Borrow, and Streaming : Internet Archive. [online] Available at: <https://archive.org/details/ASCIIConversionChart/mode/1up> [Accessed 19 April 2022].

References

- Vector-conversions.com. 2022. Raster (Bitmap) vs Vector. [online] Available at: <https://vector-conversions.com/vectorizing/raster_vs_vector.html> [Accessed 19 April 2022].
- W3techs.com. 2022. Usage Statistics and Market Share of Character Encodings for Websites, April 2022. [online] Available at: <https://w3techs.com/technologies/overview/character_encoding> [Accessed 19 April 2022].
- Youtube.com. 2022. Representing Numbers and Letters with Binary: Crash Course Computer Science. [online] Available at: <https://www.youtube.com/watch?v=1GSjbWt0c9M> [Accessed 19 April 2022].
- Medium. 2022. Character Encoding in Depth. [online] Available at: <https://medium.com/tech-tajawal/https-medium-com-tech-tajawal-character-encoding-in-depth-6f1df87888d8> [Accessed 19 April 2022].