Data Representation

Prof. Shahenda Sarhan
Prof @ computer Science dept.

Data representation units

- Bit
- Byte
- Nibble
- Word

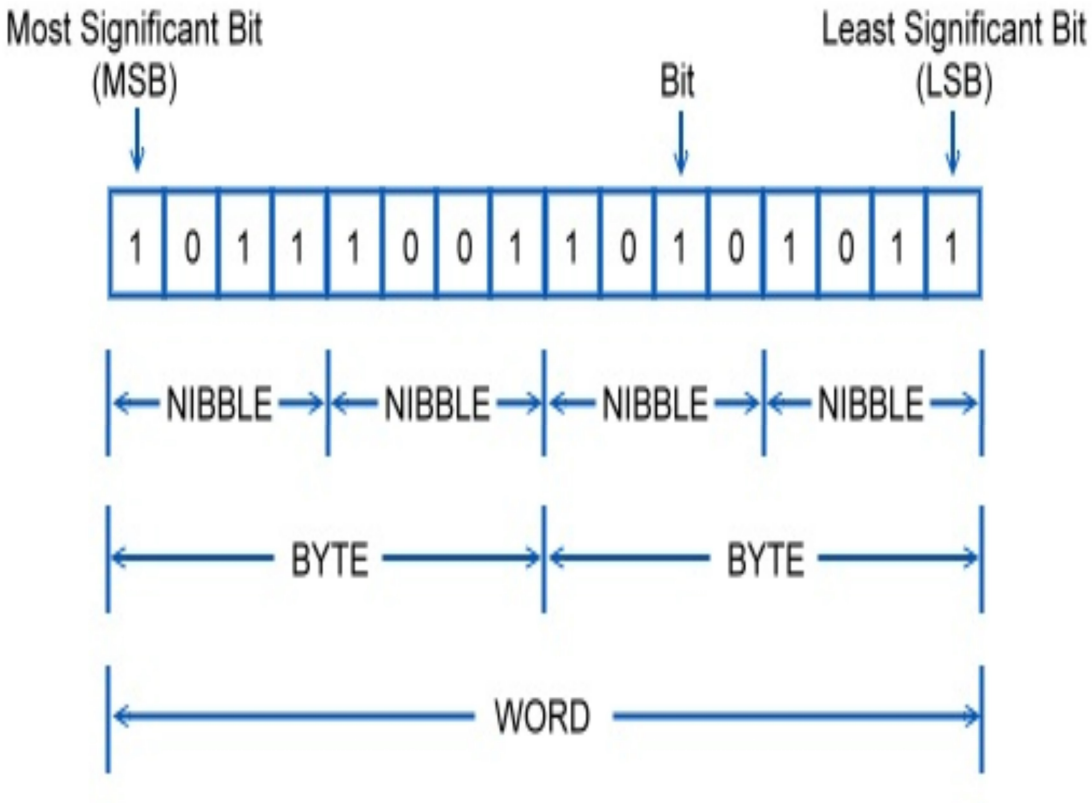
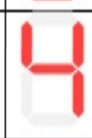


Image Representation

- Depends on what the output device can display.
- ex: an image on the seven segments can be represented by 7 bits.

No	Img	Repre.	3	Img	1111001	7	Img	1100000
0		1111110	4		0110011	8		1111111
1		0110000	5		1011011	9		1111011
2		1101101	6		1011111	A		1110111

Raster Images

- pixel could be represented by different sizes depending on the image type.
- Image types
  - Binary image: pixel represented by one bit such an image is also called a 1-bit monochrome image since it contains no color.

Monochrome 1-bit Lena image.

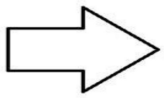
Raster Images

- Gray image: pixel represented by 8 bits where, each pixel has a gray-value between 0 and 255,
  - for example a dark pixel might have a value of 10, and a bright one might be 230.
  - As each pixel is usually stored as a byte, so a 640 ×480 grayscale image requires 300 KB of storage.

Grayscale image of Lena.

Raster Images

- Color image: pixel represented by 24 bits.
  - Each pixel is composed by three values Red, Green, Blue.
  - Supports 256 × 256 × 256 possible combined colors, or a total of 16,777,216 possible colors.
  - Storage Penalty: A 640×480 24bit color image would require 921.6 KB of storage without any compression.



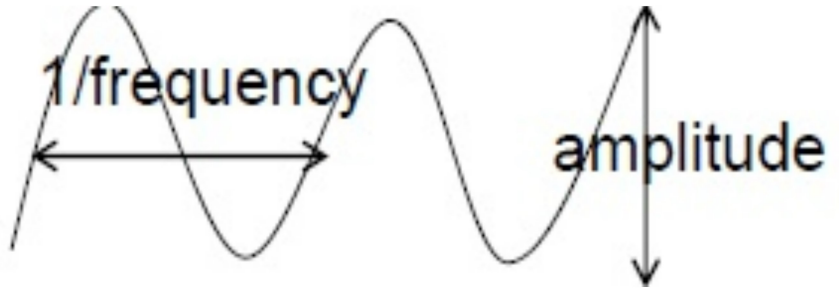
Red

Green

Blue

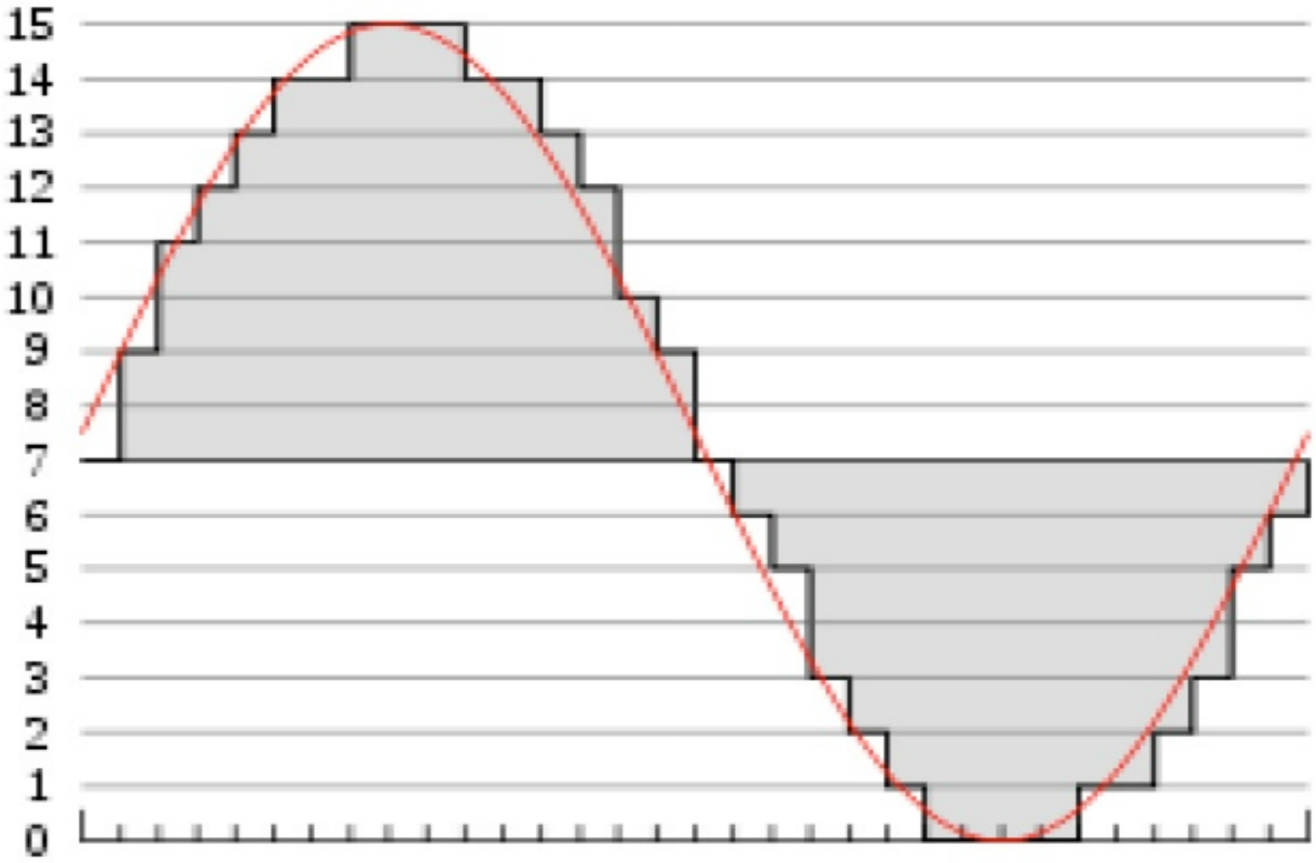
Sound Representation

- Sound is an analog acoustic wave
- A simple wave can be characterized by amplitude and frequency.
- The larger the amplitude the louder the sound
- The higher the frequency the higher the tone.



Sound Representation

- In order to store the sound wave digitally:
  - Sampling : the voltage is sampled at frequent intervals typically 48kHz per second.
  - Decoding : Stored as a binary code typically 16 or 32 bits per sample.



Sampling

Video Representation

- Composed by a sequence of continuous images and synchronized sound tracks.
- Each image is called a frame.
- Each frame is flashed on the screen for a short time faster than 50/sec
- makes the impression of continuous movie.

Data representation standards

Data representation standards

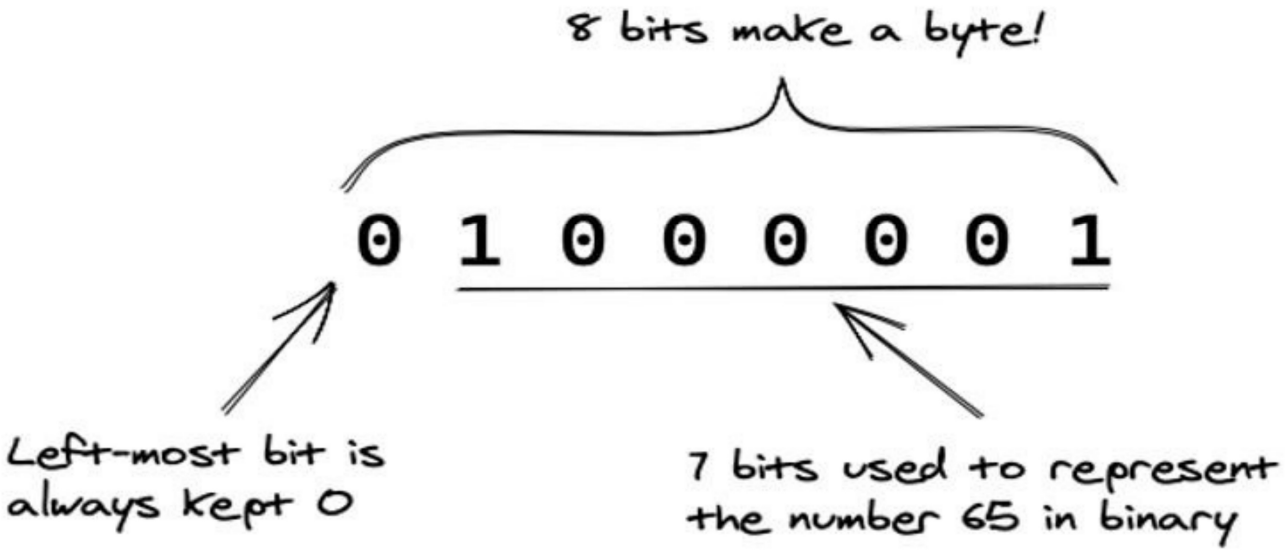
- A standard is needed to insure that the code that's used on your computer is the same as the code that is used on any computer.
- Two standard codes
  - **ASCII :** American Standard Code for Information Interchange
  - **EBCDIC :** Extended Binary Coded Decimal Interchange Code (used in the past by IBM mainframe s)
  - limited to 256 possible combinations,
  - certain character sets, such as Chinese, Arabic, Japanese, Klingon and others cannot be represented using these codes

ASCII CODE

Hex	Dec	Char		Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL	null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH	Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX	Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX	End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT	End of transmission	0x24	36	$	0x44	68	D	0x64	100	d
0x05	5	ENQ	Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK	Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL	Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS	Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB	Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF	New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT	Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF	Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR	Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO	Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI	Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE	Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1	Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2	Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3	Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4	Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK	Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN	Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB	End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN	Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM	End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB	Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	FSC	Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS	File separator	0x3C	60	<	0x5C	92	\	0x7C	124	|
0x1D	29	GS	Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS	Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US	Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL

ASCII

For example, the character "A" has a value of 65 (in decimal). This means it would be stored in memory as: 01000001.



EBCDIC CODE

Bit Positions 4,5,6,7		BIT POSITIONS 0, 1, 2, 3															
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
	Hex	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	0	NUL	DLE	DS		SP	&	-						{	}	\	0
0001	1	SOH	DC1	SOS						a	j	~		A	J		1
0010	2	STX	DC2	FS	SYN					b	k	s		B	K	S	2
0011	3	ETX	DC3							c	l	t		C	L	T	3
0100	4	PF	RES	BYP	PN					d	m	u		D	M	U	4
0101	5	HT	NL	LF	RS					e	n	v		E	N	V	5
0110	6	LC	BS	EOB / ETB	UC					f	o	w		F	O	W	6
0111	7	DEL	IL	PRE / ESC	EOT					g	p	x		G	P	X	7
1000	8		CAN							h	q	y		H	Q	Y	8
1001	9	RLF	EM						\	i	r	z		I	R	Z	9
1010	A	SMM	CC	SM		¢	!	¦	:								
1011	B	VT				.	$	,	#								
1100	C	FF	IFS		DC4	<	*	%	@								
1101	D	CR	IGS	ENQ	NAK	(	)	_	'								
1110	E	SO	IRS	ACK		+	;	>	=								
1111	F	SI	IUS	BEL	SUB	!	¬	?	"								

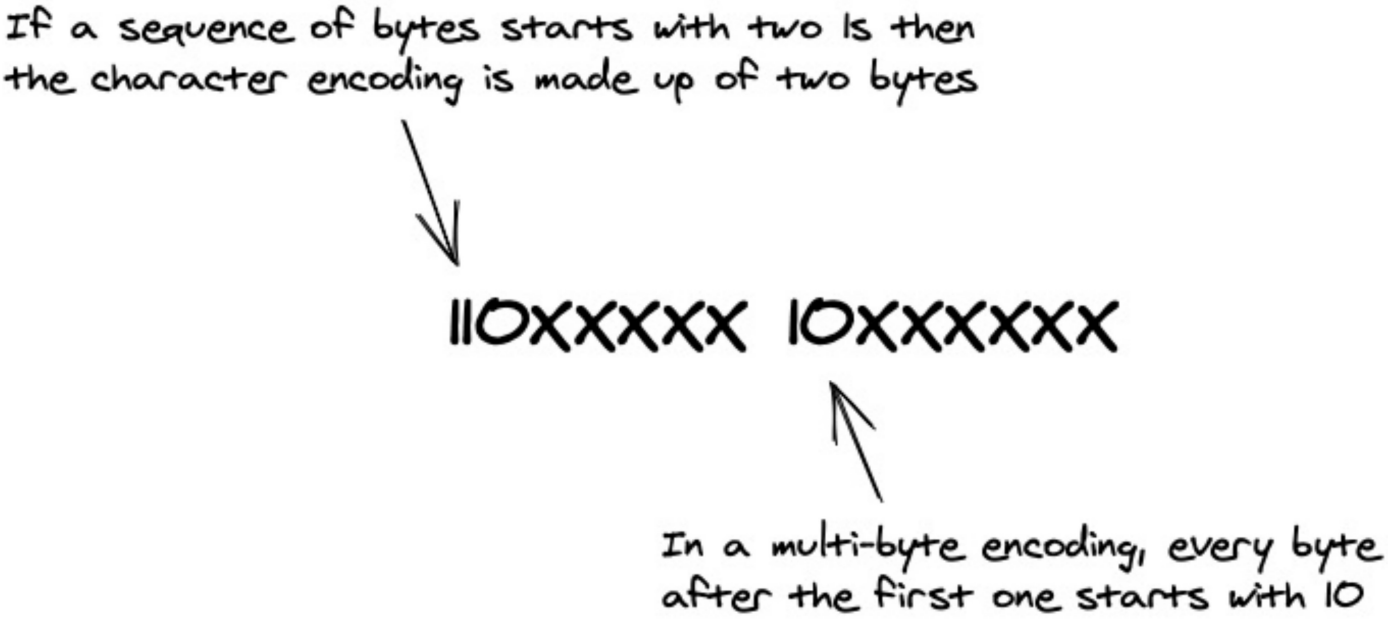
Unicode

- uses 2 bytes for each character.
- Advantages
  - allows 2^{16} different symbols to be represented, a total of 65,536.
  - flexibility
- Disadvantages
  - takes twice as much space to store the same number of characters.

UNICODE

0020	0 0030	@ 0040	P 0050	` 0060	p 0070	00A0	° 00B0	À 00C0	Ð 00D0	à 00E0	ð 00F0
! 0021	1 0031	A 0041	Q 0051	a 0061	q 0071	¡ 00A1	± 00B1	Á 00C1	Ñ 00D1	á 00E1	ñ 00F1
" 0022	2 0032	B 0042	R 0052	b 0062	r 0072	¢ 00A2	² 00B2	Â 00C2	Ò 00D2	â 00E2	ò 00F2
# 0023	3 0033	C 0043	S 0053	c 0063	s 0073	£ 00A3	³ 00B3	Ã 00C3	Ó 00D3	ã 00E3	ó 00F3
$ 0024	4 0034	D 0044	T 0054	d 0064	t 0074	¤ 00A4	´ 00B4	Ä 00C4	Ô 00D4	ä 00E4	ô 00F4
% 0025	5 0035	E 0045	U 0055	e 0065	u 0075	¥ 00A5	µ 00B5	Å 00C5	Õ 00D5	å 00E5	õ 00F5
& 0026	6 0036	F 0046	V 0056	f 0066	v 0076	¦ 00A6	¶ 00B6	Æ 00C6	Ö 00D6	æ 00E6	ö 00F6
' 0027	7 0037	G 0047	W 0057	g 0067	w 0077	§ 00A7	· 00B7	Ç 00C7	× 00D7	ç 00E7	÷ 00F7
(0028	8 0038	H 0048	X 0058	h 0068	x 0078	¨ 00A8	¸ 00B8	È 00C8	Ø 00D8	è 00E8	ø 00F8
) 0029	9 0039	I 0049	Y 0059	i 0069	y 0079	© 00A9	¹ 00B9	É 00C9	Ù 00D9	é 00E9	ù 00F9
* 002A	: 003A	J 004A	Z 005A	j 006A	z 007A	ª 00AA	º 00BA	Ê 00CA	Ú 00DA	ê 00EA	ú 00FA
+ 002B	; 003B	K 004B	[005B	k 006B	{ 007B	« 00AB	» 00BB	Ë 00CB	Û 00DB	ë 00EB	û 00FB
, 002C	< 003C	L 004C	\ 005C	l 006C	| 007C	¬ 00AC	¼ 00BC	Ì 00CC	Ü 00DC	ì 00EC	ü 00FC
- 002D	= 003D	M 004D	] 005D	m 006D	} 007D	00AD	½ 00BD	Í 00CD	Ý 00DD	í 00ED	ý 00FD
. 002E	> 003E	N 004E	^ 005E	n 006E	~ 007E	® 00AE	¾ 00BE	Î 00CE	Þ 00DE	î 00EE	þ 00FE
/ 002F	? 003F	O 004F	_ 005F	o 006F	007F	¯ 00AF	¿ 00BF	Ï 00CF	ß 00DF	ï 00EF	ÿ 00FF

UTF-8


If a sequence of bytes starts with two 1s then the character encoding is made up of two bytes
110XXXXX 10XXXXXX
In a multi-byte encoding, every byte after the first one starts with 10

UTF-8

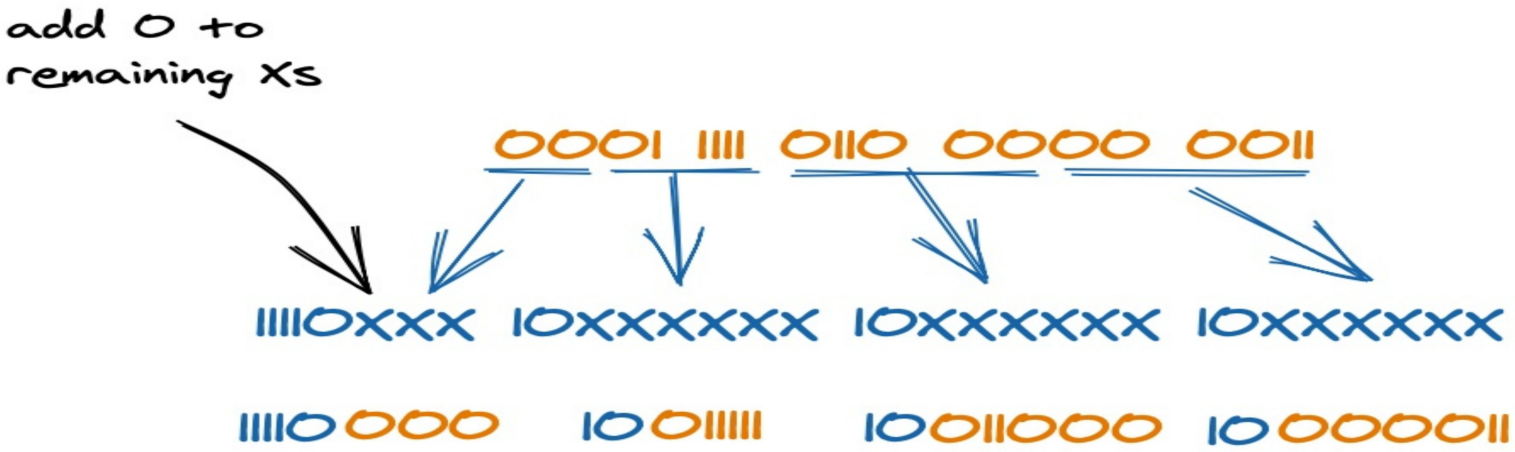
Counting

Number of bytes	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
1	U+0000	U+007F	0xxxxxxx	same as ASCII				
2	U+0080	U+07FF	110xxxxx	10xxxxxx	Data			
3	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx			
4	U+10000	U+1FFFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
5	U+200000	U+3FFFFFF	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
6	U+4000000	U+7FFFFFFF	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

UTF-8 example

Encode the following emoji using UTF-8: 😄

- We first need to find the code point value of this emoji on the internet and you will find that the emoji has a **Unicode number** of U+1F603.
- This just means that the emoji has a code point value of 1F603 (in hexadecimal).
- Converting 1F603 to binary, we get: 0001 1111 0110 0000 0011.
- Now we can start filling in these bytes starting from the right-most bit



UTF-8 Questions

Q1: Represent the following code in UTF-8:

1110 0111 01 011011 101011

In the result, separate the bytes with a space

- Answer : 11111000 10001110 10011101 10011011 10101011

Q2: Is this a correct UTF-8 code: 11000000 10000000 10000000?

- Answer : No