CS2842 Computer Systems – Lecture IV

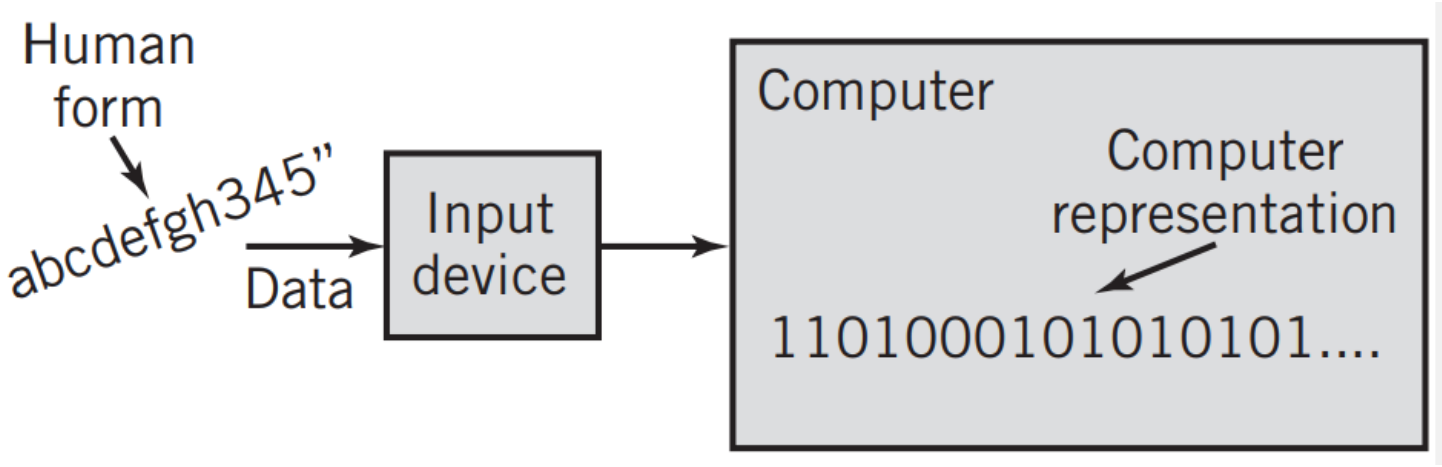
Data Formats

Dr. Sapumal Ahangama
Department of Computer Science and Engineering

DATA FORMATS

- Computers
  - Process and store all forms of data in binary format

- Human communication
  - Includes language, images and sounds



DATA FORMATS

- Specifications for converting data into computer usable form
- Define the different ways human data may be represented, stored and processed by a computer
- The data must have the ability to be moved between computers
  - Metadata: information that describes or interprets the meaning of the data

DATA FORMATS

- Proprietary formats
  - Individual programs can store and process data in any format that they want

- Standard data representations
  - to be used as interfaces between different programs,
  - between a program and the I/O devices used by the program,
  - between interconnected hardware,
  - between systems that share data

COMMON DATA REPRESENTATIONS

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC
Image (bitmapped)	▪GIF (graphical image format) ▪TIF (tagged image file format) ▪PNG (portable network graphics)
Image (object)	PostScript, SWF (Macromedia Flash), SVG
Outline graphics and fonts	PostScript, TrueType
Sound	WAV, AVI, MP3, MIDI, WMA
Page description	PDF (Adobe Portable Document Format), HTML, XML
Video	Quicktime, MPEG-2, RealVideo, WMV

ALPHANUMERIC DATA

- Much of the data that will be used in a computer are originally provided in human-readable form,
  - Letters of the alphabet, numbers, and punctuation,
  - English or some other language

- Alphanumeric data are a combination of alphabetical and numerical characters

- Since alphanumeric data must be stored and processed within the computer in binary form, each character must be translated to a binary representation

ALPHANUMERIC DATA

- Three alphanumeric codes are in common use,
  - ASCII (American Standard Code for Information Interchange)
  - EBCDIC (Extended Binary Coded Decimal Interchange Code)
  - Unicode

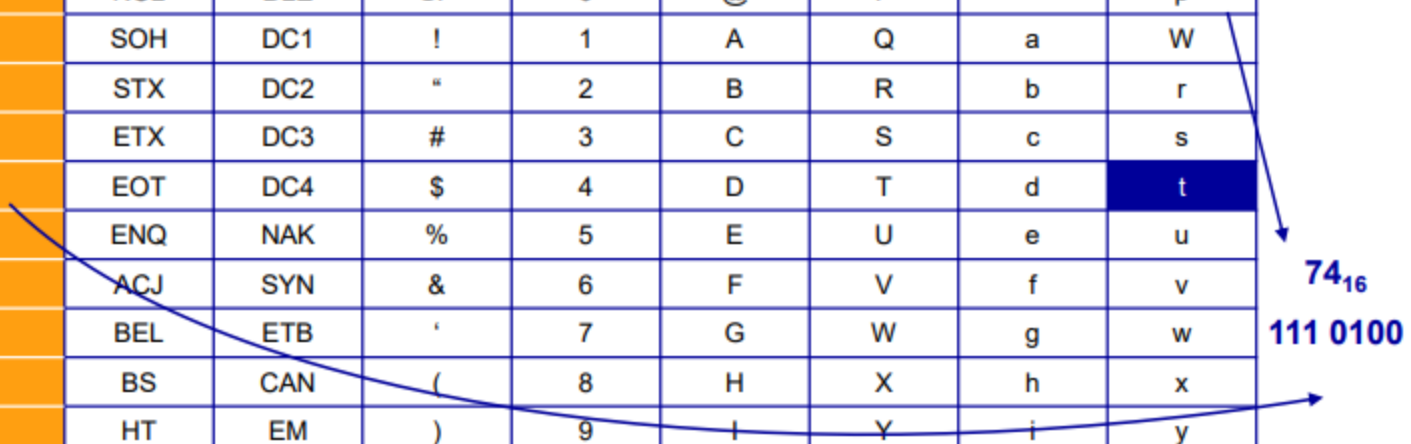
- Nearly every system today uses Unicode or ASCII

ASCII

- Each character represented with a 7 bit code
  - 128 characters
- Consists of,
  - digits 0 to 9,
  - lowercase letters a to z,
  - uppercase letters A to Z,
  - punctuation symbols,
  - 33 non-printing control codes
- Extended to 8 bit code – Latin-1

ASCII

MSD / LSD	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P		p
1	SOH	DC1	!	1	A	Q	a	W
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACJ	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



UNICODE

- ASCII and EBCDIC have limitations
  - 8-bit word limit the number of possible characters
  - Other major languages?
  - Omitted characters [,], ^, {, }, ~

- These issues led to a 16 bit standard – Unicode or UTF-16
  - 65,536 characters
  - 49,000 are defined to represent the world's most used characters
  - 6,400 16-bit codes are reserved for private use
  - Each character can be stored in 2 bytes

UNICODE

Sinhala[1][2]

Official Unicode Consortium code chart (PDF)

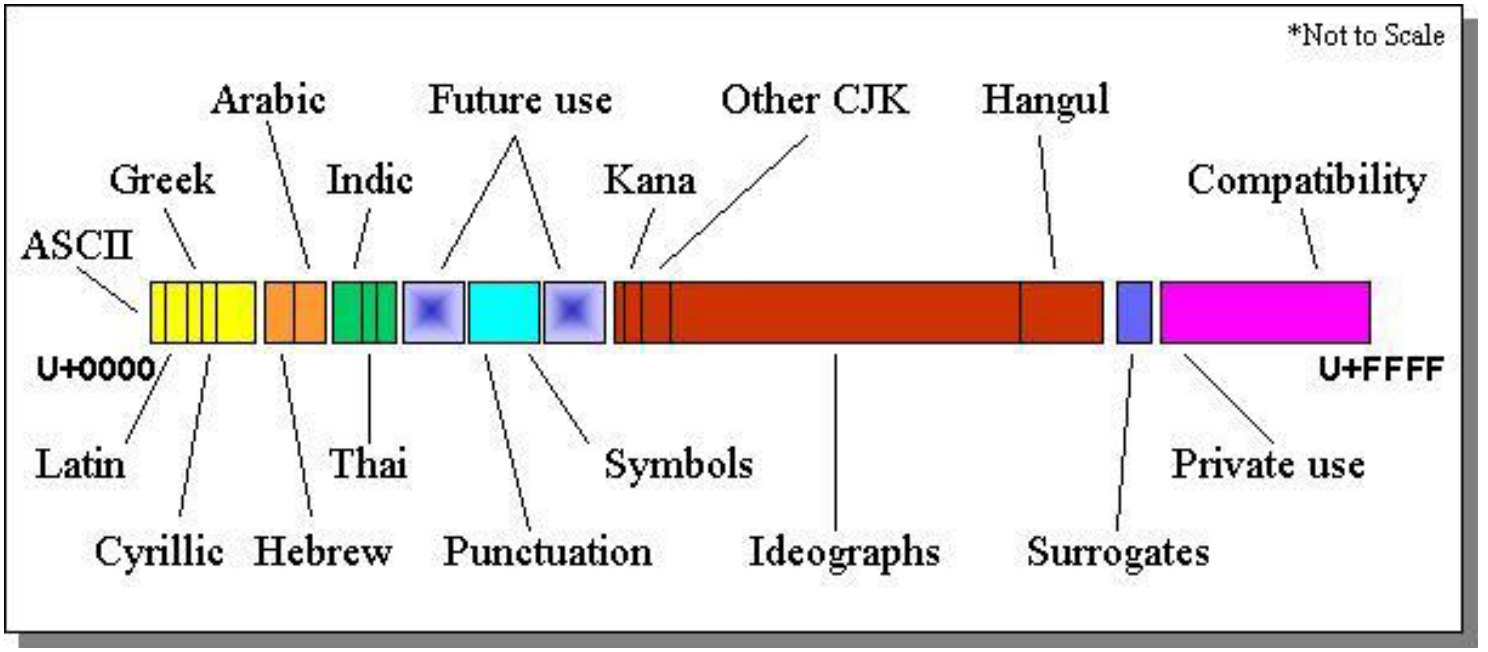
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0D8x			◌ං	◌ඃ		අ	ආ	ඇ	ඈ	ඉ	ඊ	උ	ඌ	ඍ	ඎ	ඏ
U+0D9x	ඐ	එ	ඒ	ඓ	ඔ	ඕ	ඖ				ක	ඛ	ග	ඝ	ඞ	ඟ
U+0DAx	ච	ඡ	ජ	ඣ	ඤ	ඥ	ඦ	ට	ඨ	ඩ	ඪ	ණ	ඬ	ත	ථ	ද
U+0DBx	ධ	න		ඳ	ප	ඵ	බ	භ	ම	ඹ	ය	ර		ල		
U+0DCx	ව	ශ	ෂ	ස	හ	ළ	ෆ				◌්					◌ා
U+0DDx	◌ැ	◌ෑ	◌ි	◌ී	◌ු		◌ූ		◌ෘ	◌ෙ	◌ේ	◌ෛ	◌ො	◌ෝ	◌ෞ	◌ෟ
U+0DEx							෦	෧	෨	෩	෪	෫	෬	෭	෮	෯
U+0DFx			◌ෲ	◌ෳ												

Notes

1.^ As of Unicode version 12.0

2.^ Grey areas indicate non-assigned code points

UNICODE


*Not to Scale
Arabic
Future use
Other CJK
Hangul
Greek
Indic
Kana
Compatibility
ASCII
U+0000
U+FFFF
Latin
Thai
Symbols
Private use
Cyrillic
Hebrew
Punctuation
Ideographs
Surrogates

2 CLASSES OF CODE

- Printing characters
  - Produced on the screen or printer

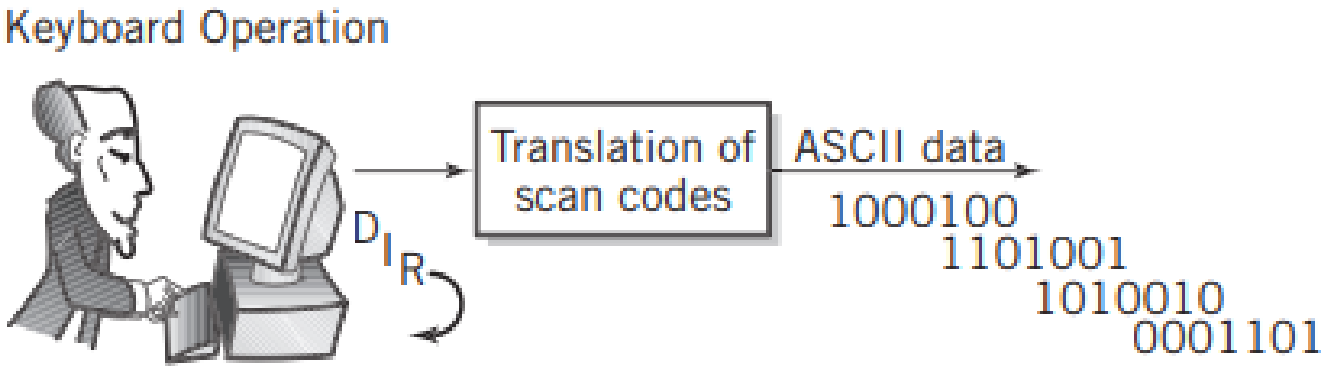
- Control characters

NUL	(Null) No character; used to fill space
SOH	(Start of Heading) Indicates start of a header used during transmission
STX	(Start of Text) Indicates start of text during transmission
ETX	(End of Text) Similar to above
EOT	(End of Transmission)
ENQ	(Enquiry) A request for response from a remote station; the response is usually an identification
ACK	(Acknowledge) A character sent by a receiving device as an affirmative response to a query by a sender
BEL	(Bell) Rings a bell
BS	(Backspace)
HT	(Horizontal Tab)
LF	(Line Feed)
VT	(Vertical Tab)
FF	(Form Feed) Moves cursor to the starting position of the next page, form, or screen
CR	(Carriage return)
SO	(Shift Out) Shift to an alternative character set until SI is encountered
SI	(Shift In) see above

DLE	(Data Link Escape) Similar to escape, but used to change meaning of data control characters; used to permit sending of data characters with any bit combination
DC1,DC2, DC3, DC4	(Device Controls) Used for the control of devices or special terminal features
NAK	(Negative Acknowledgment) Opposite of ACK
SYN	(Synchronous) Used to synchronize a synchronous transmission system
STB	(End of Transmission Block) Indicates end of a block of transmitted data
CAN	(Cancel) Cancel previous data
EM	(End of Medium) Indicates the physical end of a medium such as tape
SUB	(Substitute) Substitute a character for one sent in error
ESC	(Escape) Provides extensions to the code by changing the meaning of a specified number of contiguous following characters
FS, GS, RS, US	(File, group, record, and united separators) Used in optional way by systems to provide separations within a data set
DEL	(Delete) Delete current character

KEYBOARD INPUT

- Scan code
  - When a key is struck on the keyboard, the circuitry in the keyboard generates a binary code



KEYBOARD INPUT

- Other alphanumeric inputs:
  - OCR
  - Barcode
  - Magnetic Strip Reader
  - RFID

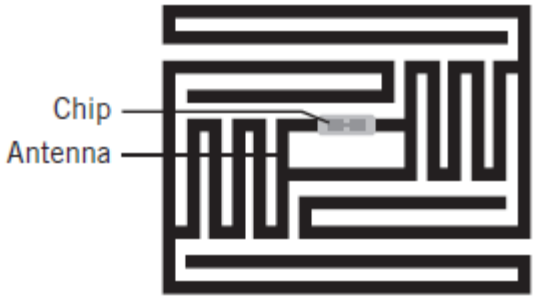


IMAGE DATA

- Images come in many different shapes, sizes, textures, colors, and shadings
  - Different requirements require different forms for image data
    - Quality of the image
    - Storage space required
    - Time to transmit
    - Ease of modification
  - Make it difficult to define a single universal format

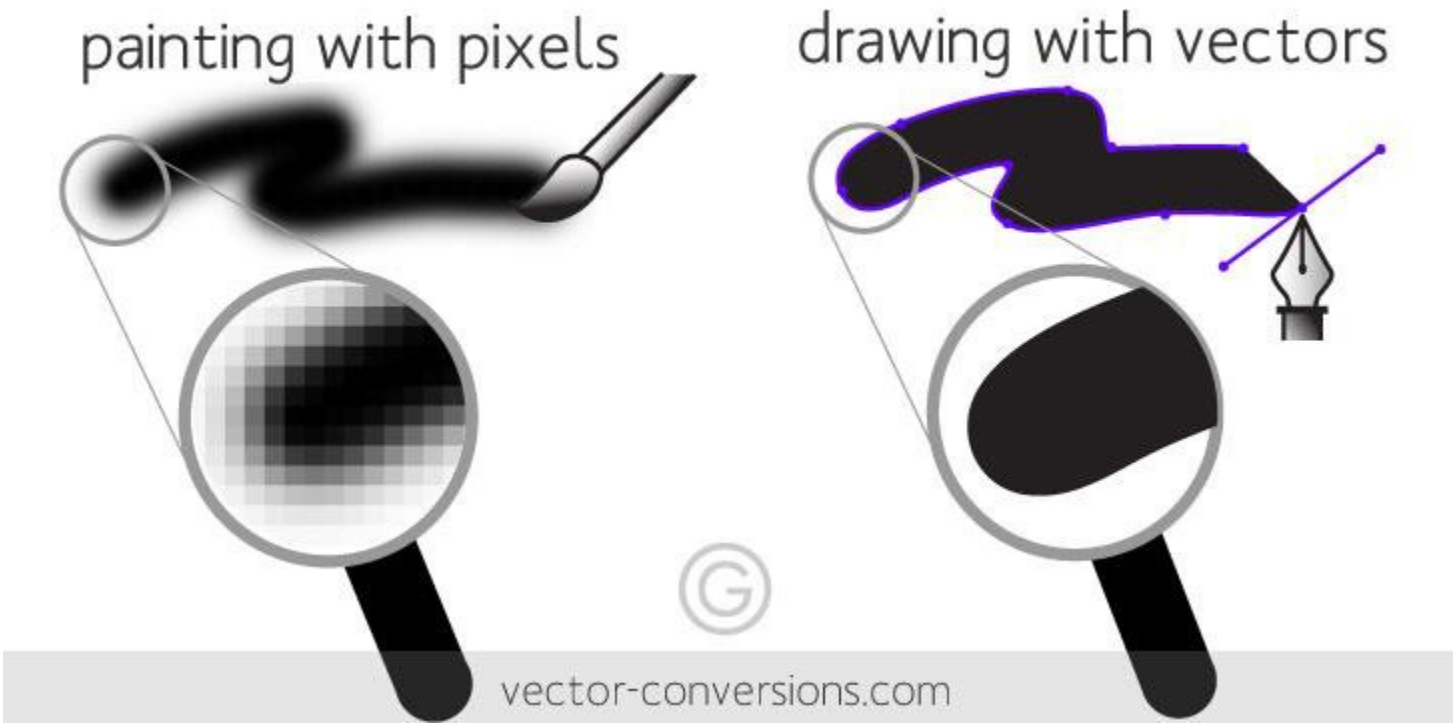
IMAGE DATA

- Two distinct categories
  - Bitmap or raster images
    - Characterized by continuous variations in shading, color, shape, and texture
    - JPEG, GIF
  - Graphical objects
    - Made up of graphical shapes such as lines and curves that can be defined geometrically
- The nature of display technology make it much more convenient and cost effective to display and print most images as bitmaps

IMAGE DATA

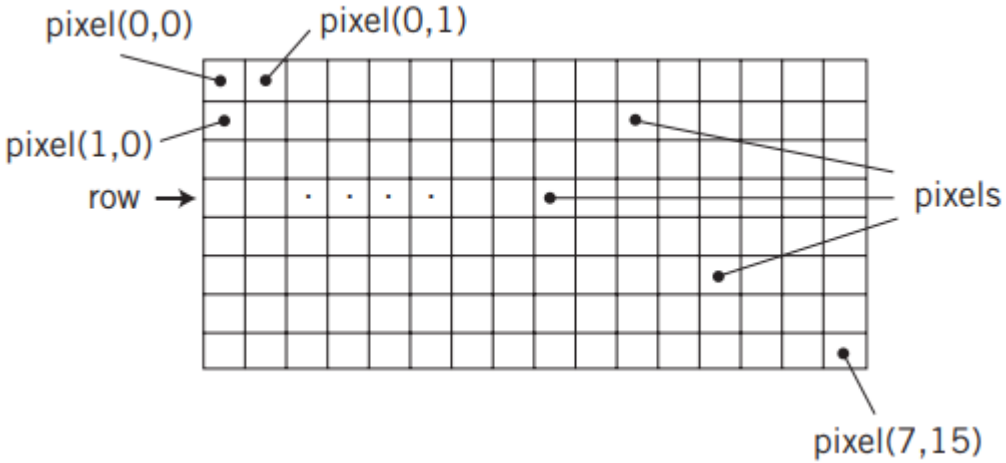
- Two distinct categories
  - Bitmap or raster images
    - Characterized by continuous variations in shading, color, shape, and texture
    - JPEG, GIF
  - Graphical objects
    - Made up of graphical shapes such as lines and curves that can be defined geometrically
- The nature of display technology make it much more convenient and cost effective to display and print most images as bitmaps

IMAGE DATA


painting with pixels
drawing with vectors
vector-conversions.com

BITMAP IMAGES

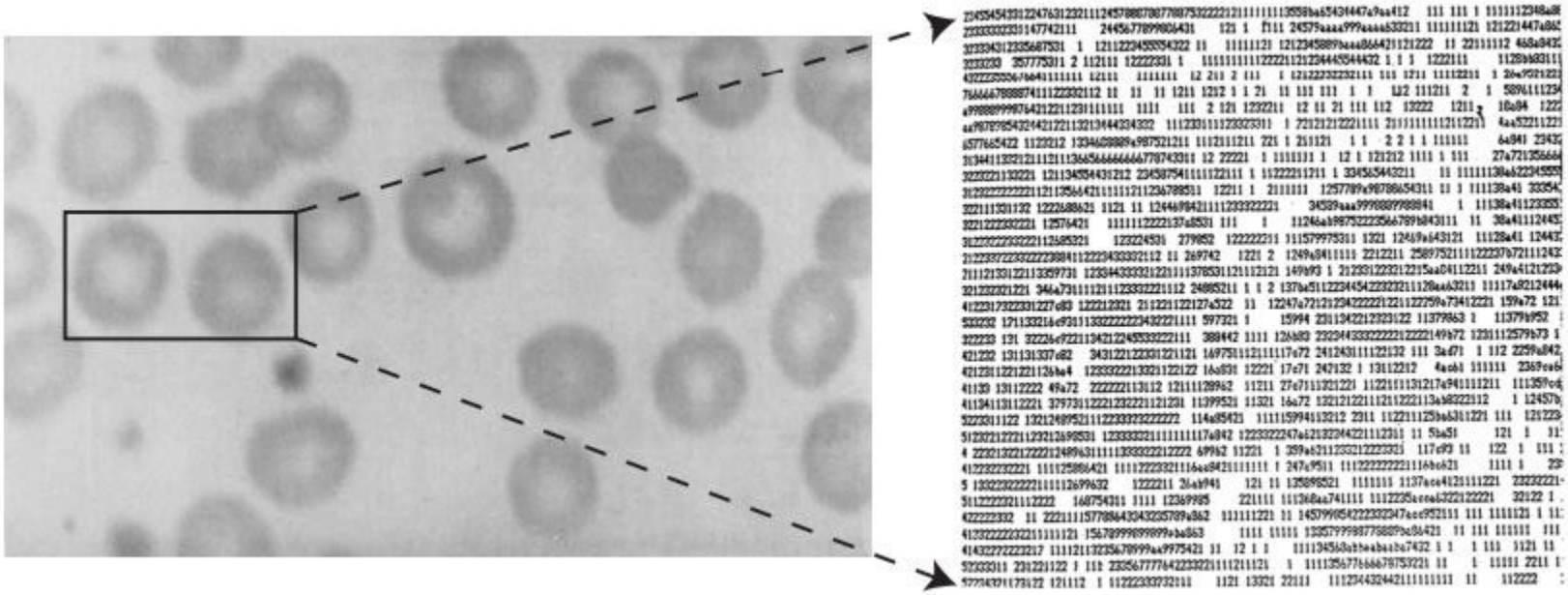
- Bitmap image format
  - A rectangular image is divided into rows and columns
  - The junction of each row and column is a point known as a pixel
  - Pixel is a set of one or more binary numerical values that define the visual characteristics



- Preferred when image contains large amount of detail and processing requirements are fairly simple

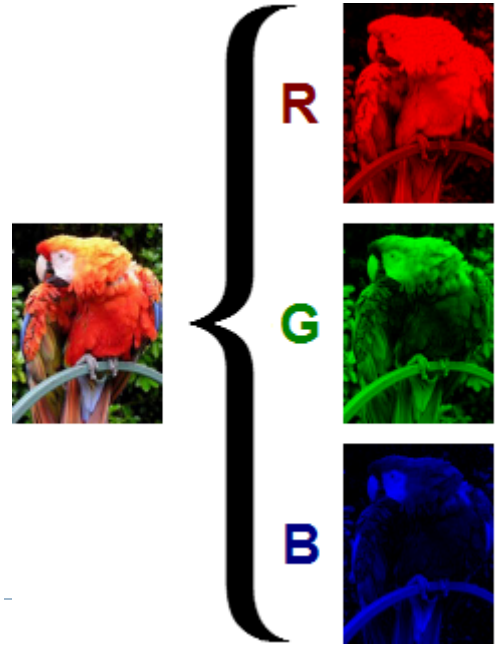
BITMAP IMAGES

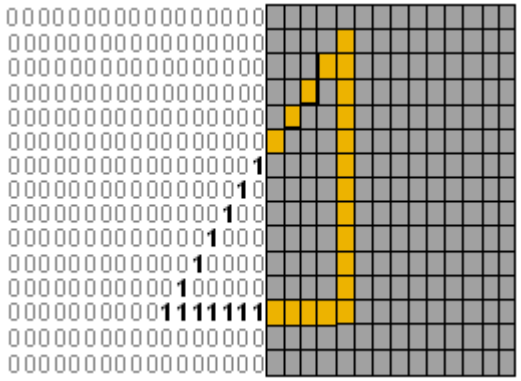
- Example each point below represented by a 4 bit code corresponding to 1 of 16 shades
- Meta data
- Pixel data
  - Stored from top to bottom one row at a time



BITMAP IMAGES

- Data value representing a pixel
  - Could be as simple as one bit
  - For color image, might consist of many bytes
    - RGB
    - Additional bytes for other characteristics such as transparency and color correction.

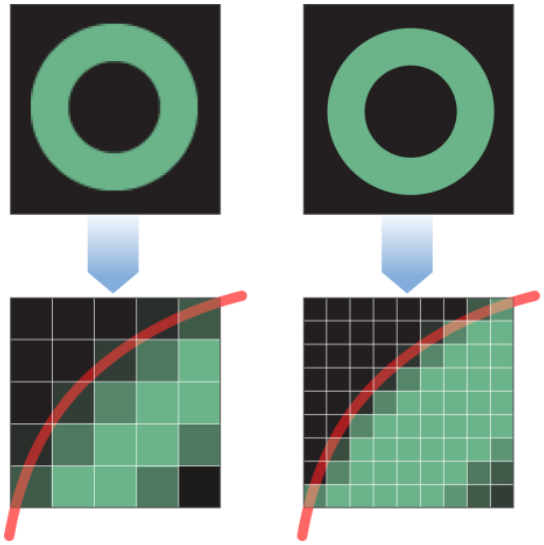




BITMAP IMAGES

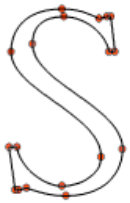
- File size affected by
  - Resolution
    - Reducing the size of a pixel to improve details
  - Levels: number of bits to represent each pixel

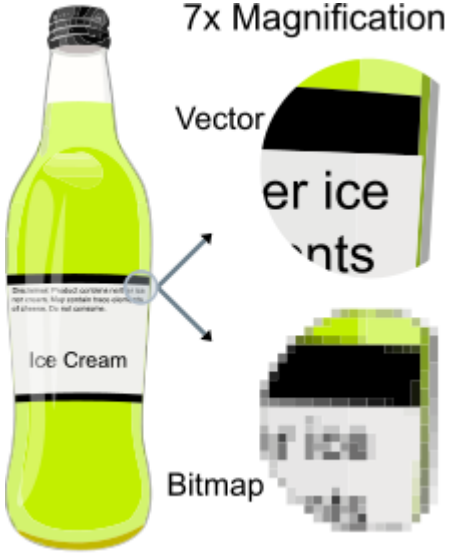
- Image formats
  - GIF (Graphics Interchange Format)
  - JPEG (Joint Photographers Expert Group)
  - PNG (Portable Network Graphic)



OBJECT IMAGES

- Object images are made up of simple elements like straight or curved lines, circles and arcs etc.
  - Each element can defined mathematically by parameters
  - Circle requires 3 parameters, Cartesian coordinates + radius
  - Straight line needs the coordinates of its end points





OBJECT IMAGES

- Advantages
  - Require less storage space
  - Can be manipulated easily

- Photographs as object images?

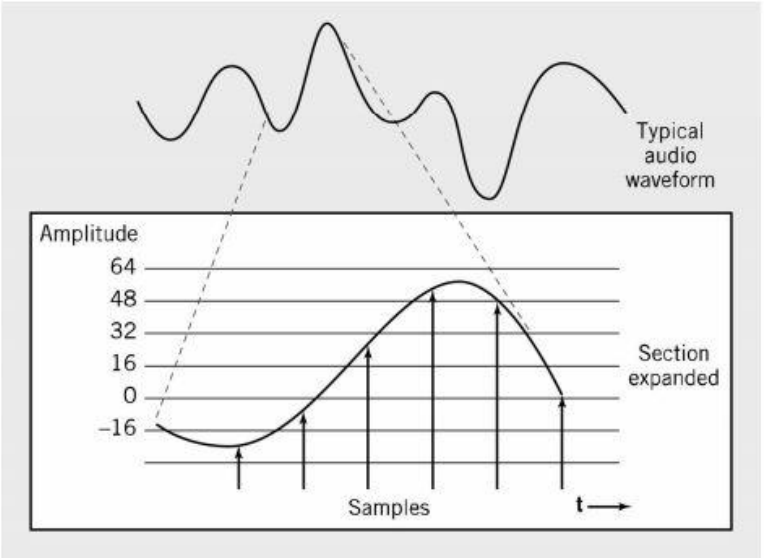
VIDEO DATA

- Requires a large amount of data
  - 1024 × 768 pixel true-color images at a frame rate of 30 frames per second?
  - 70.8 megabytes of data per second!
  - 4.25 gigabytes per minute
- How to reduce video size?

AUDIO DATA

- Sound is naturally an analog wave that needs to be digitized

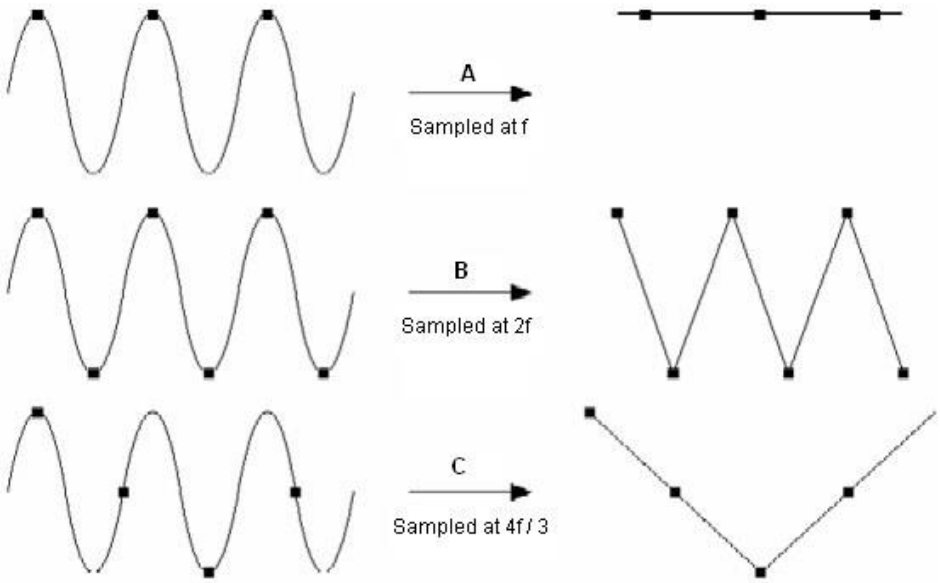
- Sampling
  - 1000 samples per second = 1 KHz (kilohertz)
  - Example : Audio CD sampling rate = 44.1KHz



Sampling rate normally 50KHz

AUDIO DATA

- Sampling Rate



- Height of each sample saved as,
  - 8 bit number for radio quality recordings
  - 16 bit number for high fidelity recordings
  - 2 x 16 bits for stereo sound

DATA COMPRESSION

- Compression: reducing data so that it requires fewer bytes of storage space
- Compression ratio: the amount of file shrunk

- Lossless Compression
  - Inverse algorithm restores data to exact original form
  - Examples GIF, PCX, TIFF

  - 0 5 5 7 3 2 0 0 0 0 1 4 7 3 2 9 1 0 0 0 0 0 6 6 8 2 7 3 2 7 3 2
  - 0 1 5 5 7 3 2 0 4 1 4 7 3 2 9 1 0 5 6 6 8 2 7 3 2 7 3 2
  - 0 1 5 5 Z 0 3 1 4 Z 9 1 0 5 6 6 8 2 Z Z

DATA COMPRESSION

- Lossy Compression
  - Trades off data degradation for file size and download speed
  - Much higher compression ratios, often 10 to 1
  - JPEG

Original Lena Image (12KB size)

Lena Image, Compressed (85% less information, 1.8KB)

Lena Image, Highly Compressed (96% less information, 0.56KB)

- MPEG-2?

THANK YOU