Cambridge O level Computer Science 2210. Syllabus for examination in 2018-19

1.1.3 Data storage

- Show understanding that sound (music), pictures, video, text and numbers are stored in different formats
- identify and describe methods of error detection and correction, such as parity checks, check digits, checksums and Automatic Repeat reQuests **(ARQ)**
- show understanding of the concept of Musical Instrument Digital Interface **(MIDI)** files, JPEG files, MP3 and MP4 files
- show understanding of the principles of data compression (lossless and lossy compression algorithms) applied to music/video, photos and text files

Data Storage:

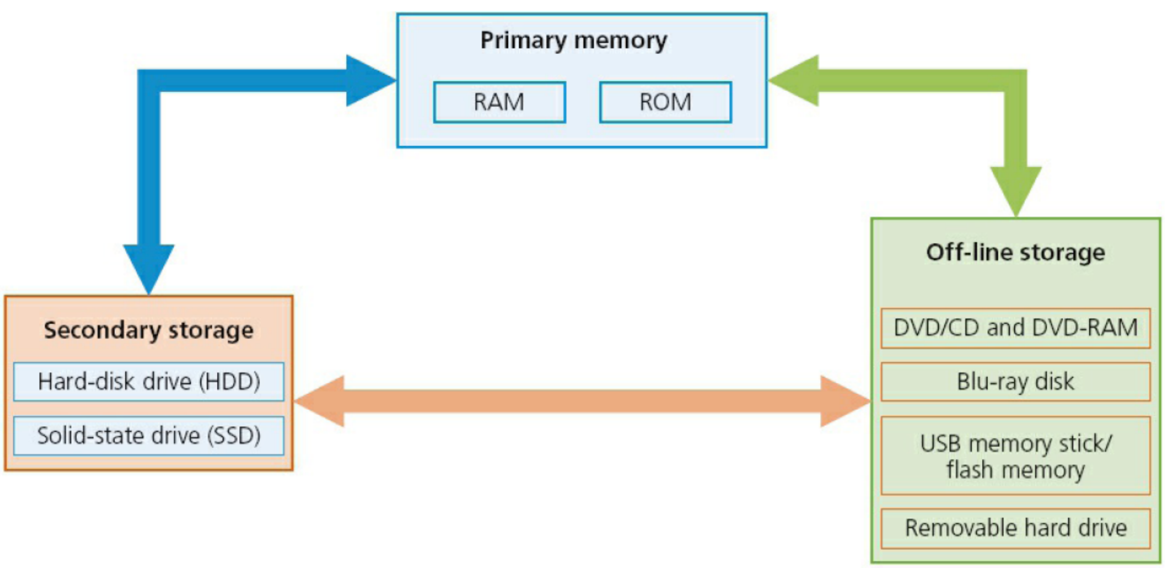
In a computer, storage is the place where data is held in an electromagnetic or optical form for access by a computer processor. There are two general usages:

1) Storage is frequently used to mean the devices and data connected to the computer through input/output operations - that is, hard disk and tape systems and other forms of storage that don't include computer memory and other in-computer storage.

2) In a more formal usage, storage has been divided into:

(a) Primary storage: Holds data in memory (sometimes called random access memory or **RAM, ROM**) and other **"built-in"** devices such as **the processor's L1, L2 cache.** Processor's **MEMORY CACHE** is the high speed portion of the memory; it is effective because most programs access the same data or instructions many times.

(b) Secondary storage: Holds data on hard disks, tapes, and other devices like USB, SD Cards, Zip and Floppy Disks and Optical disks etc.



Primary storage is much faster to access than secondary storage because of the proximity of the storage to the processor or because of the nature of the storage devices. On the other hand, secondary storage can hold much more data than primary storage.

FORMATS FOR STORAGE OF DATA:

The following are the various types of data that you might find:

TEXT

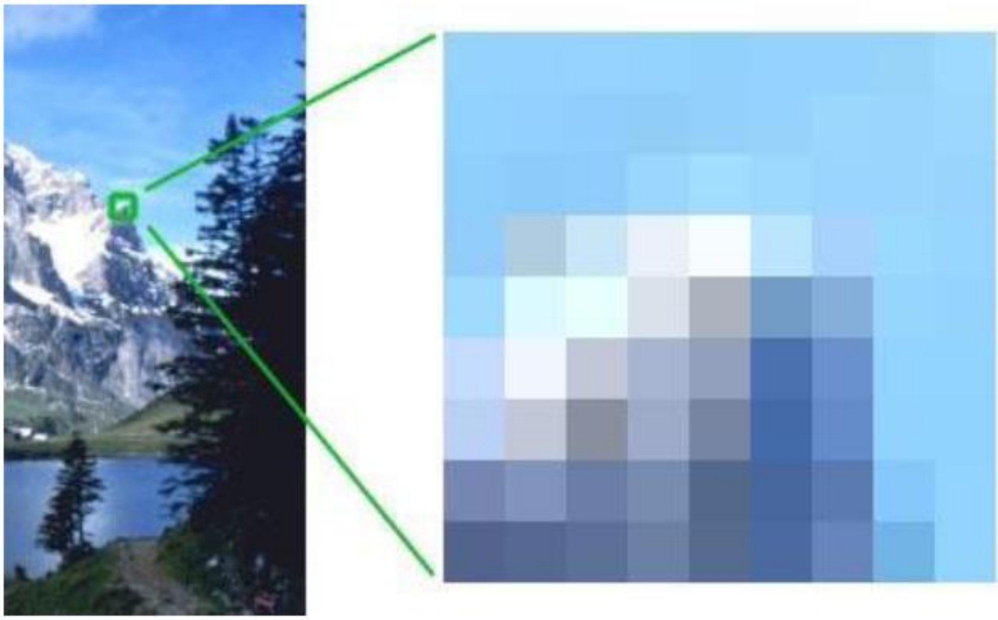
Text can be represented easily by assigning a unique numeric value for each symbol used in the text. For example, the widely used **ASCII code (American Standard Code for Information Interchange)** defines 128 different symbols (all the characters found on a standard keyboard, plus a few extra), and assigns to each a unique numeric code between **0 and 127**. In ASCII, an "A" is 65," B" is 66, "a" is 97, "b" is 98, and so forth. When you save a file as "plain text", it is stored using ASCII. ASCII format uses 1 byte per character 1 byte gives only 256 (128 standard and 128 non-standard) possible characters. The code value for any character can be converted to base 2, so any written message made up of ASCII characters can be converted to a string of 0's and 1's.

Binary code	Hexadecimal equivalent	Character	Description
00000000	00	NUL	Null character
00000001	01	SOH	Start of heading
00000010	02	STX	Start of text
00100000	20		Space
00100001	21	!	Exclamation mark
00100100	24	$	Dollar
00101011	2B	+	Plus
00101111	2F	/	Forward slash
00110000	30	0	Zero
00110001	31	1	One
00110010	32	2	Two
01000001	41	A	Uppercase A
01000010	42	B	Uppercase B
01000011	43	C	Uppercase C
01100001	61	a	Lowercase a
01100010	62	b	Lowercase b
01100011	63	c	Lowercase c

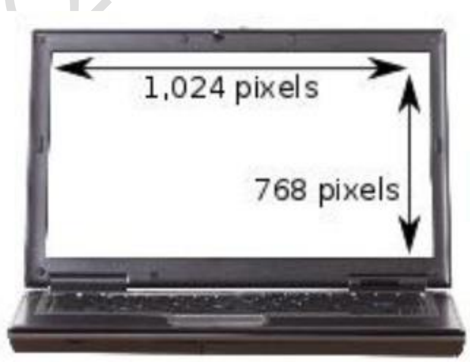
Table 1.03 Some examples of ASCII codes

GRAPHICS

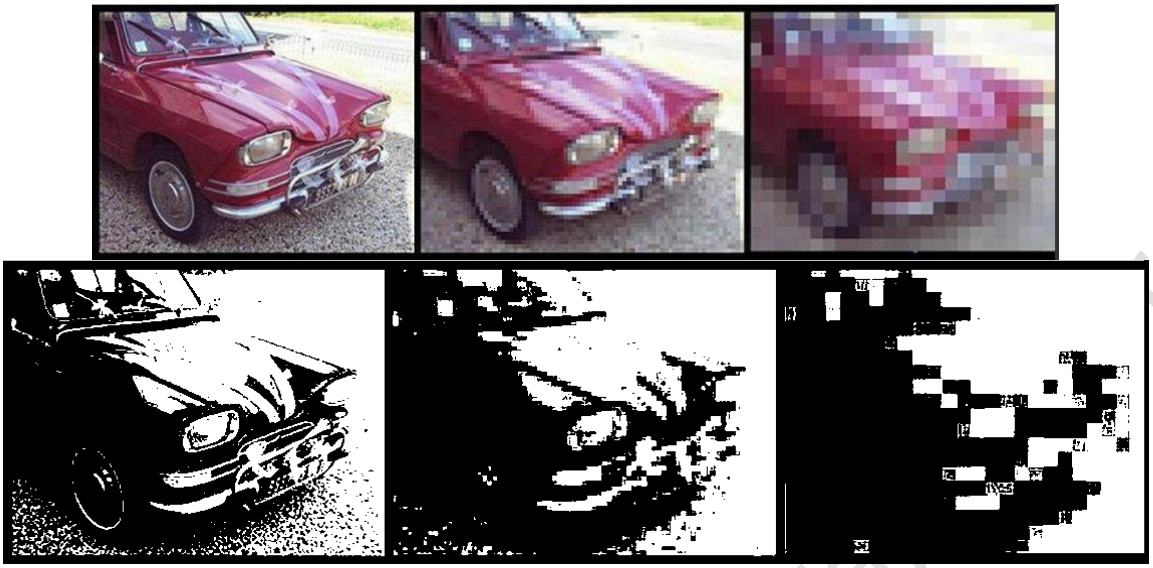
Graphics that are displayed on a computer screen consist of pixels: the tiny "dots" of color that collectively "paint" a graphic image on a computer screen.



The pixels are organized into many rows on the screen. In one common configuration, each row is 640 pixels long, and there are 480 such rows. Another configuration (and the one used on the screens in this lab) is 1024 pixels per row with 786 rows, which is referred to as a "resolution of 1024x786." Each pixel has two properties: its location on the screen and its color.



A graphic image can be represented by a list of pixels. Imagine all the rows of pixels on the screen laid out end to end in one long row. This gives the pixel list, and a pixel's location in the list corresponds to its position on the screen.



A pixel's color is represented by a binary code, and consists of a certain number of bits. In a monochrome (black and white) image, only 1 bit is needed per pixel: 0 for black, 1 for white, for example. A 16 bit color image requires 4 bits per pixel. Modern display hardware allows for 24 bits per pixel, which provides an astounding array of 16.7 million possible colors for each pixel!

ANIMATION

Somewhere between the motionless world of still images and the real-time world of video images lies the flip-book world of computer animation. All of the animated sequences seen in educational programs, motion CAD renderings, and computer games are computer-animated (and in many cases, computer-generated) animation sequences.

Traditional cartoon animation is little more than a series of artwork cells, each containing a slight positional variation of the animated subjects. When a large number of these cells are displayed in sequence and at a fast rate, the animated figures appear to the human eye as if they are moving.

A computer-animated sequence works in exactly the same manner. A series of images are created of a subject; each image contains a slightly different perspective on the animated subject.

When these images are displayed (played back) in the proper sequence and at the proper speed (frame rate), the subject appears to move. Computerized animation is actually a combination of both still and motion imaging. Each frame, or cell, of an animation is a still image that requires compression and storage. An animation file, however, must store the data for hundreds or thousands of animation frames and must also provide the information necessary to play back the frames using the proper display mode and frame rate.

DIGITAL VIDEO:

One step beyond animation is broadcast video. Your television and video tape recorder are a lot more complex than an 8mm home movie projector and your kitchen wall. There are many complex signals and complicated standards that are involved in transmitting those late-night reruns across the airwaves and cable. Only in the last few years has a personal computer been able to work with video data at all.

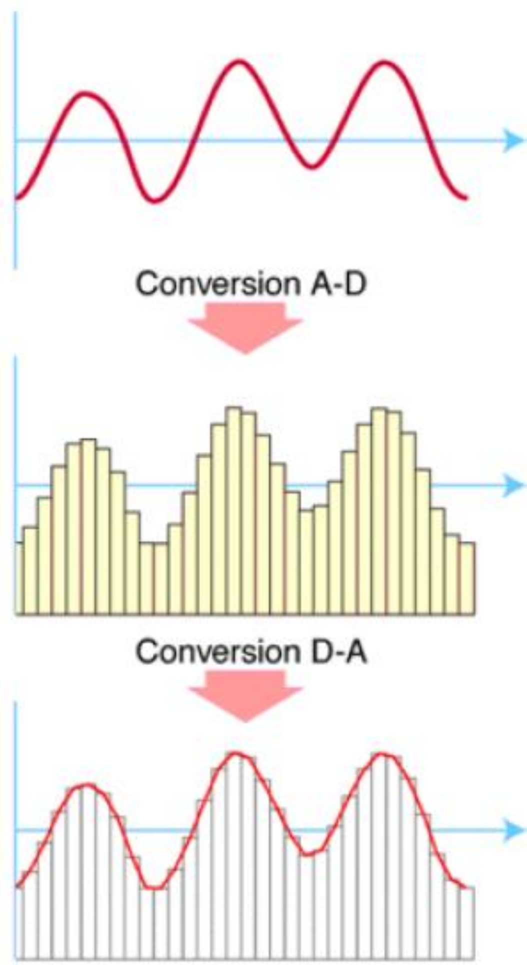
Video data normally occurs as continuous, analog signals. In order for a computer to process this video data, we must convert the analog signals to a non-continuous, digital format. In a digital format, the video data can be stored as a series of bits on a hard disk or in computer memory. The process of converting a video signal to a digital bit stream is called analog-to-digital conversion (A/D conversion), or digitizing. A/D conversion occurs in two steps:

1. **Sampling** captures data from the video stream.
2. **Quantizing** converts each captured sample into a digital format.

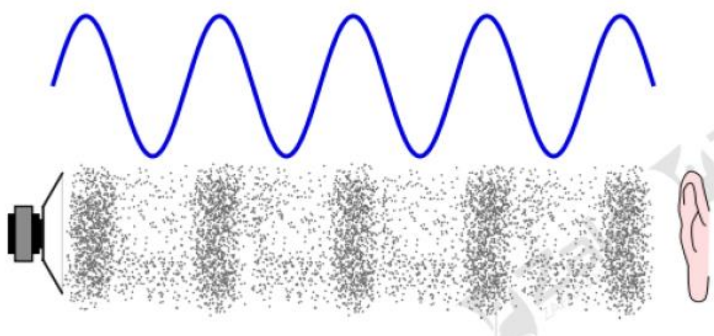
Each sample captured from the video stream is typically stored as a 16-bit integer. The rate at which samples are collected is called the sampling rate. The sampling rate is measured in the number of samples captured per second (samples/second). For digital video, it is necessary to capture millions of samples per second.

<u>DIGITAL AUDIO</u>

All multimedia file formats are capable, by definition, of storing sound information. Sound data, like graphics and video data, has its own special requirements when it is being read, written, interpreted, and compressed. Before looking at how sound is stored in a multimedia format we must look at how sound itself is stored as digital data.

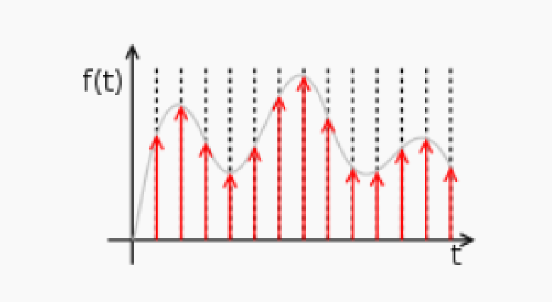
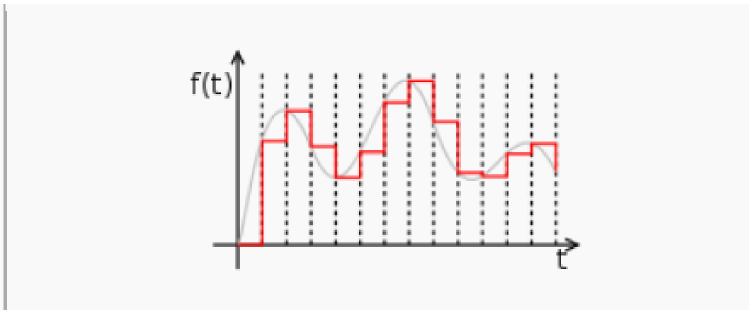


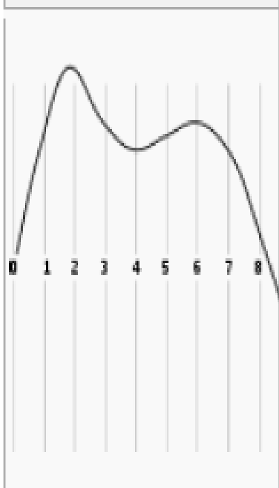
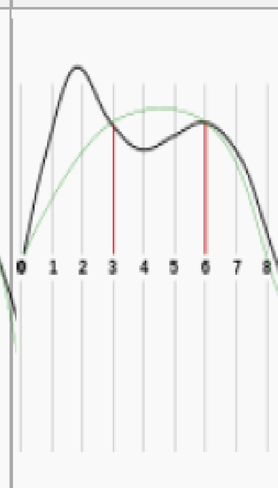
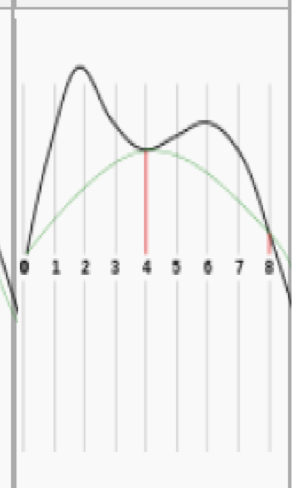
All of the sounds that we hear occur in the form of analog signals.



An analog audio recording system, such as a conventional tape recorder, captures the entire sound wave form and stores it in analog format on a medium such as magnetic tape.

Because computers are now digital devices it is necessary to store sound information in a digitized format that computers can readily use. A digital audio recording system does not record the entire wave form as analog systems do (the exception being Digital Audio Tape [DAT] systems). Instead, a digital recorder captures a wave form at specific intervals, called the sampling rate. Each captured wave-form snapshot is converted to a binary integer value and is then stored on magnetic tape or disk.

	
Sound is a continuous set of data points formed by a wave. Computers sample this sound at discrete points to store a digital approximation	The discrete approximations (in red) can be used to recreate the original sound (grey). However, due to limitations in the number of samples we take we are often unable to truly represent a sound wave, though we can get close enough for the human ear not to notice the difference.

Original Sound	High sample rate	1/2 high sample rate	1/3 high sample rate	1/4 high sample rate
				
original continuous sound wave	digital looks like original	digital loses sharpness	loss of peaks	poor resemblance to original!

DATA COMPRESSION

Storing data in a format that requires less space than usual. Data compression is particularly useful in communications because it enables devices to transmit or store the same amount of data in fewer bits

LOSSY COMPRESSION

Lossy compression refers to discarding irrelevant information. Generally this means compressing images, video, or audio by discarding data that the human perceptual system cannot see or hear.

The state of the art is to apply lossy compression only at a very low level of human sensory modeling, where the model is well understood.

LOSSLESS COMPRESSION

Lossless data compression is a class of **data compression algorithms** that allows the original data to be perfectly reconstructed from the compressed data.
Lossless data compression is used in many applications. For example, it is used in the **ZIP** file format.

Lossless compression is used in cases where it is important that the original and the decompressed data be identical. Typical examples are executable programs, text documents, and source code.

DATA FILE FORMATS:

MIDI STANDARD

Musical Instrument Digital Interface (MIDI) is an industry standard for representing sound in a binary format. MIDI is not an audio format, however. It does not store actual digitally sampled sounds.

Instead, MIDI stores a description of sounds, in much the same way that a vector image format stores a description of an image and not image data itself. Sound in MIDI data is stored as a series of control messages. Each message describes a sound event using terms such as pitch, duration, and volume. When these control messages are sent to a MIDI-compatible device (the

MIDI standard also defines the interconnecting hardware used by MIDI devices and the communications protocol used to interchange the control information) the information in the message is interpreted and reproduced by the device.

MIDI data may be compressed, just like any other binary data, and does not require special compression algorithms in the way that audio data does.

MIDI uses **.mid extension** and takes very less space in memory.

MP3
Filename Extension: .mp3
Format Type: Lossy Compressed

When Internet file-sharing boomed into popularity with Napster and the iPod, the MP3 cornered the market for one reason: it had a small footprint. Without broadband connections, it was impractical at the time to share file sizes larger than the MP3 standard 2 – 3 Megabytes.

And that preference has stuck for some time now even though MP3 does not have nearly the same amount of quality as WAV or AIFF files. But despite this growing base of people using higher quality formats, there are still those who prefer the

So, if you have a slower internet connection or limited hard drive space, MP3 could be your file format of choice. If you're worried about quality loss, don't fret too much about it. While, yes, there is a noticeable drop off in sound quality, MP3 files fall square under the "good enough" umbrella.

JPEG

JPG files, also known as JPEG files, are a common file format for digital photos and other digital graphics. When JPG files are saved, they use "lossy" compression, meaning image quality is lost as file size decreases. JPEG stands for Joint Photographic Experts Group, the committee that created the file type. JPG files have the file extension .jpg or .jpeg.

They are the most common file type for images taken with digital cameras, and widely used for photos and other graphics used on websites.

If file sizes get very low, JPG images will become "muddy." When saving photos and other images as JPG files for the web, email and other uses, you must decide on this compromise between qualities and file size.

MP4

MP4 is an abbreviated term for **MPEG-4 Part 14**. It may also be referred to as **MPEG-4 AVC**, which stands for **Advanced Video Coding**. As the name suggests, this is a format for working with video files and was first introduced in 1998. The **MPEG** refers to Motion Pictures Expert Group who is responsible for setting the industry standards regarding digital audio and video.

The MP4 is a container format, allowing a combination of audio, video, subtitles and still images to be held in the one single file. It also allows for advanced content such as 3D graphics, menus and user interactivity.

Because MP4 was a reliable application that required a relatively low amount of bandwidth, just about everyone could take advantage of using the tool. This was especially true as technology made it possible to create more powerful desktop and laptop systems that had a larger hard drive and could command more power.
The enhancement of the speed of various types of Internet connections also helped to make MP4 more accessible to a greater audience. MP4 works in a similar although much more complex way toMP3s, by compressing the files without losing any of the quality. MP3 technology revolutionized the way in which music and audio files are used and it's looking like

the MP4 format will do the same for the video market.

TIFF is, in principle, a very flexible format that can be lossless or lossy. The details of the image storage algorithm are included as part of the file. In practice, TIFF is used almost exclusively as a lossless image storage format that uses no compression at all. Most graphics programs that use TIFF do not use compression.
Consequently, file sizes are quite big. (Sometimes a lossless compression algorithm called LZW is used, but it is not universally supported.)

PNG is also a lossless storage format. However, in contrast with common TIFF usage, it looks for patterns in the image that it can use to compress file size. The compression is exactly reversible, so the image is recovered exactly.

GIF creates a table of up to 256 colors from a pool of 16 million. If the image has fewer than 256 colors, GIF can render the image exactly. When the image contains many colors, software that creates the GIF uses any of several algorithms to approximate the colors in the image with the limited palette of 256 colors available. Better algorithms search the image to find an optimum set of 256 colors. Sometimes GIF uses the nearest color to represent each pixel, and sometimes it uses "error diffusion" to adjust the color of nearby pixels to correct for the error in each pixel.

GIF achieves compression in two ways. First, it reduces the number of colors of color-rich images, thereby reducing the number of bits needed per pixel, as just described. Second, it replaces commonly occurring patterns (especially large areas of uniform color) with a short abbreviation: instead of storing "white, white, white, white, white," it stores "5 white." Thus, GIF is "lossless" only for images with 256 colors or less. For a rich, true color image, GIF may "lose" 99.998% of the colors.

RAW is an image output option available on some digital cameras. Though lossless, it is a factor of three of four smaller than TIFF files of the same image. The disadvantage is that there is a different RAW format for each manufacturer, and so you may have to use the manufacturer's software to view the images. (Some graphics applications can read some manufacturer's RAW formats.)

BMP is an uncompressed proprietary format invented by Microsoft. There is really no reason to ever use this format.

PSD, PSP, etc., are proprietary formats used by graphics programs. Photoshop's files have the PSD extension, while Paint Shop Pro files use PSP. These are the preferred working formats as you edit images in the software, because only the proprietary formats retain all the editing power of the programs. These packages use layers, for example, to build complex images, and layer information may be lost in the nonproprietary formats such as TIFF and JPG. However, be sure to save your end result as a standard TIFF or JPG, or you may not be able to view it in a few years when your software has changed. Currently, GIF and JPG are the formats used for nearly all web images. PNG is supported by most of the latest generation browsers. TIFF is not widely supported by web browsers, and should be avoided for web use. PNG does everything GIF does, and better, so expect to see PNG replace GIF in the future. PNG will not replace JPG, since JPG is capable of much greater compression of photographic images, even when set for quite minimal loss of quality.

Refrences:

- Computer Science by David Watson & Helen Williams
- https://en.wikipedia.org