

AUDIOVISUAL REPRESENTATION BASICS



Fernando Pereira
Instituto Superior Técnico



A Multimedia World !



Multimedia regards content and technologies dealing with a combination of different content forms/media/modalities, not only including text, audio (speech, sound and music), and visual (image, video, and graphics) ...

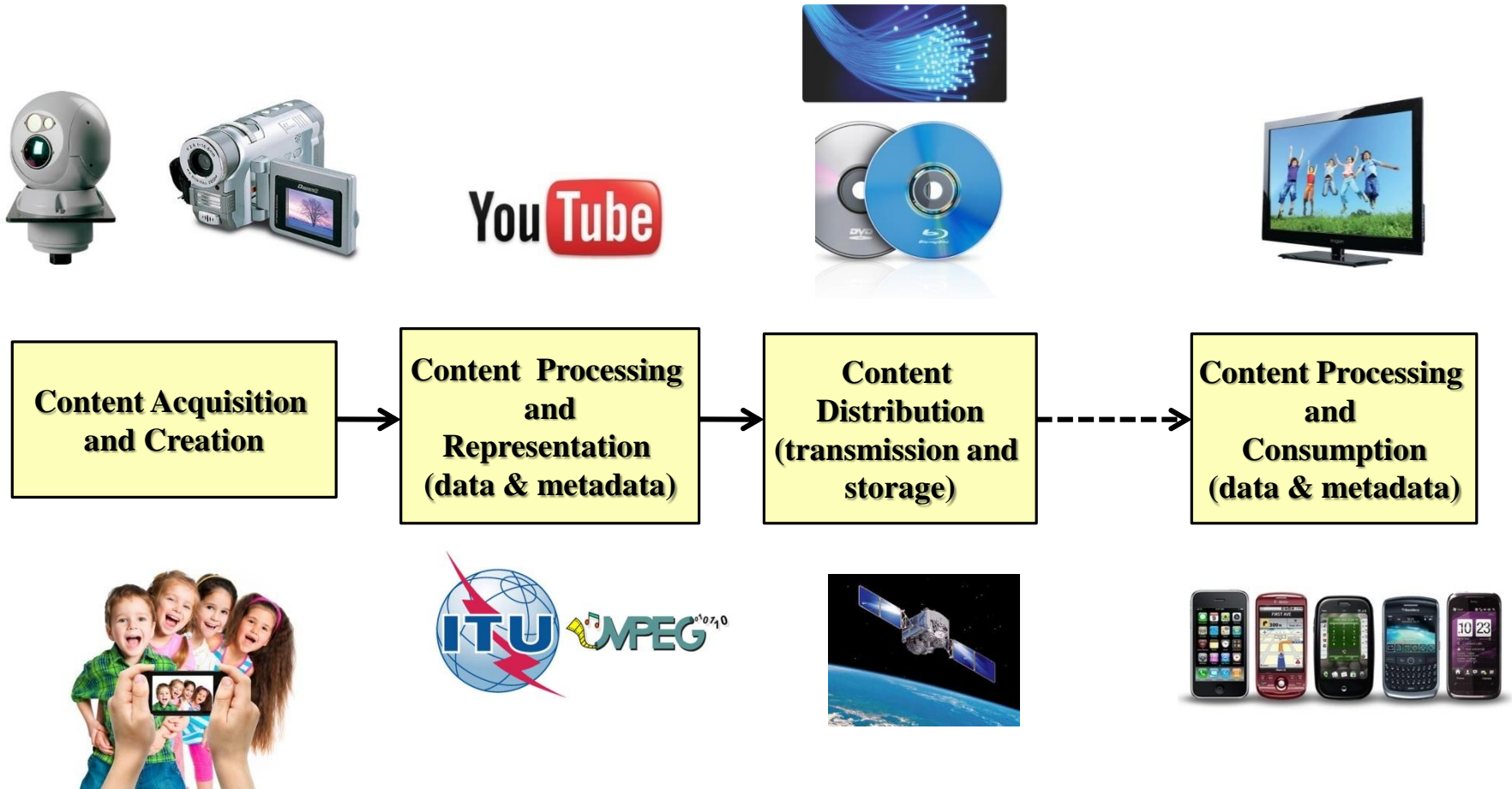
but also other sensors capturing information in novel contexts of mobile, game, health, biomedical, environment, and many others.

What do the Users Want ?

- ★ Entertainment
- ★ Communication
- ★ Information
- ★ Games
- ★ Surveillance
- ★ Education
- ★ Shopping
- ★ ...

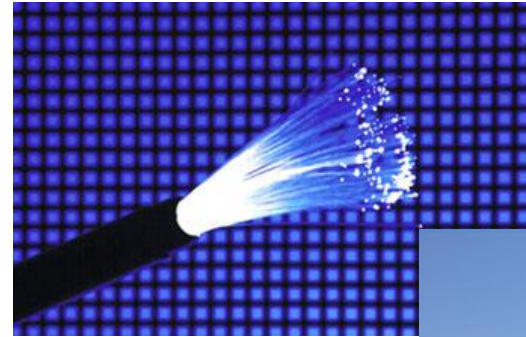


Visual Content Chain ...



There are limitations and constraints all along the content chain !

Communications: the Skeleton ...



The Importance of the Users ...



All similar ...
All different ...



How Shall a Multimedia Experience Be ?

Depending on the specific application, a multimedia experience may have to be

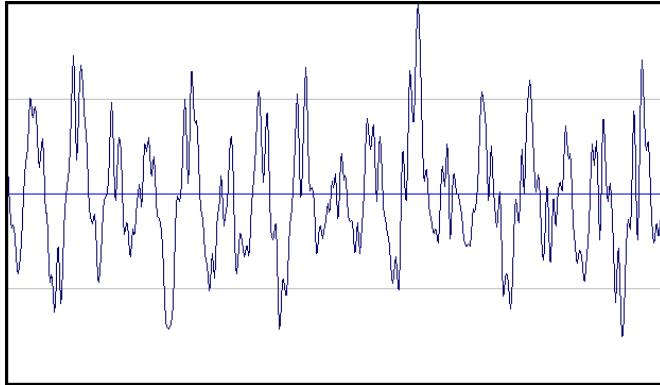
- ★ *Faithful - accuracy*
- ★ *Truthful – realistic if relevant, synchronization*
- ★ *Immersive – natural, multimodal consistency*
- ★ *Individual – emotional*
- ★ *Contextual - adaptive*
- ★ *Engaging – fun, intense*
- ★ *Effective – fast, recognition*
- ★ *Useful – task performing*
- ★ *Interactive – natural, short delay*
- ★ *Intuitive, Easy – interfaces*
- ★ ...



The Analogue World: Signals



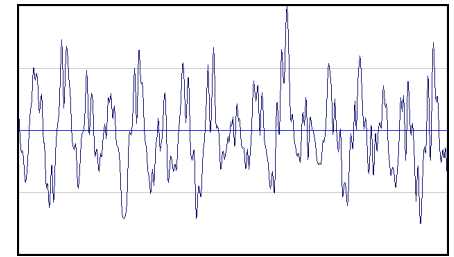
An Analogue World ...



An analog/analogue signal is any variable signal, continuous in both time and amplitude.

- ★ **Any information may be conveyed by an analogue signal; often such a signal is a measured response to changes in physical phenomena, such as sound or light, and is obtained using a transducer, e.g. camera or microphone.**
- ★ **A disadvantage of analogue representation is that any system has noise—that is, random variations—in it; as the signal is transmitted over long distances, these random variations may become dominant.**

Analogue Signal Bandwidth



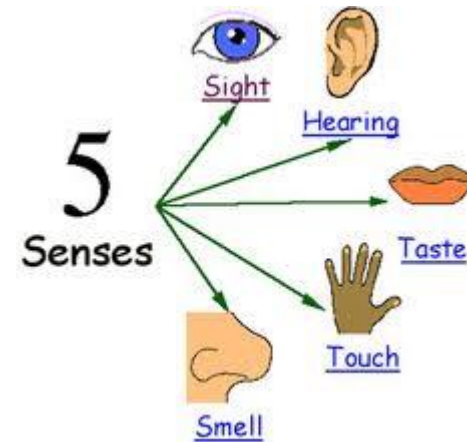
- ★ The bandwidth of a signal defines its set of frequencies this means the frequencies needed to reproduce the signal
- ★ A signal with a larger baseband bandwidth (starting from 0 Hz) may vary more quickly in time ... Because it includes higher frequencies ...
- ★ A signal with higher bandwidth is a richer signal ... and more complex and expensive ...
- ★ Typical signal bandwidths: speech – 4 kHz, music – 22 kHz, analogue TV – 5 MHz



Signal Types and Sources

In modern multimedia, there are many types of relevant signals, also called *media* or *modalities*, used to produce sensory effects, and richer user experiences, notably

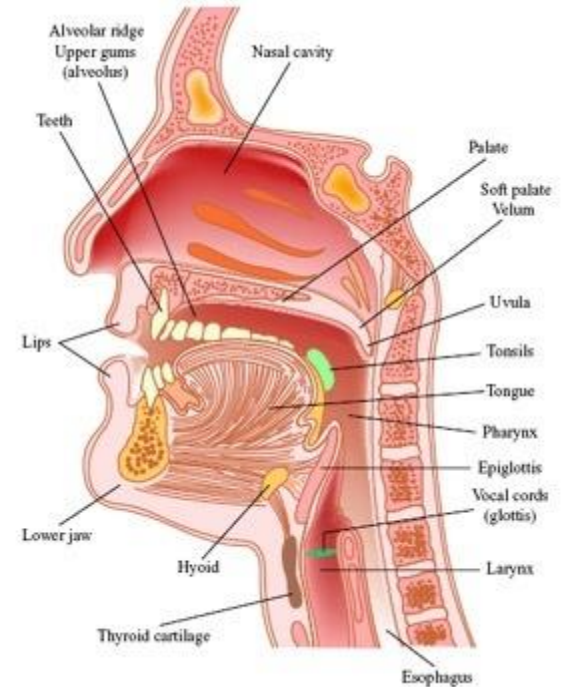
- ★ Text
- ★ Speech
- ★ Audio (includes music)
- ★ Monochromatic and colour imaging
- ★ Monochromatic and colour video
- ★ 3D image/video and 3D synthetic models
- ★ Olfactory data
- ★ Haptic data
- ★ ...





Speech Signals ...

- ★ **The human voice consists of sound made by a human being using the vocal folds for talking, singing, laughing, crying, screaming, etc. with frequency ranging from about 60 Hz to 7 kHz.**
- ★ **The human voice is specifically that part of human sound production in which the vocal cords are the primary sound source.**
- ★ **Generally speaking, the mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx, and the articulators, e.g. tongue, palate, cheek, lips.**
- ★ **In telephony, the usable voice frequency band ranges from approximately 300 Hz to 3.4 kHz. The bandwidth allocated for a single voice-frequency transmission channel is usually 4 kHz, including guard bands.**





TÉCNICO
LISBOA

Audio Signals ...

- ★ An audio signal is a representation of sound, typically as an electrical voltage.
- ★ Audio signals have frequencies in the audio frequency range roughly from 20 Hz to 20-22 kHz (the limits of the human auditory system).
- ★ Audio signals may be synthesized directly or may originate at a transducer such as a microphone.
- ★ Loudspeakers or headphones convert an electrical audio signal into sound.
- ★ An audio signal may be characterized by parameters such as their bandwidth (in Hz) and power level in decibels (dB).





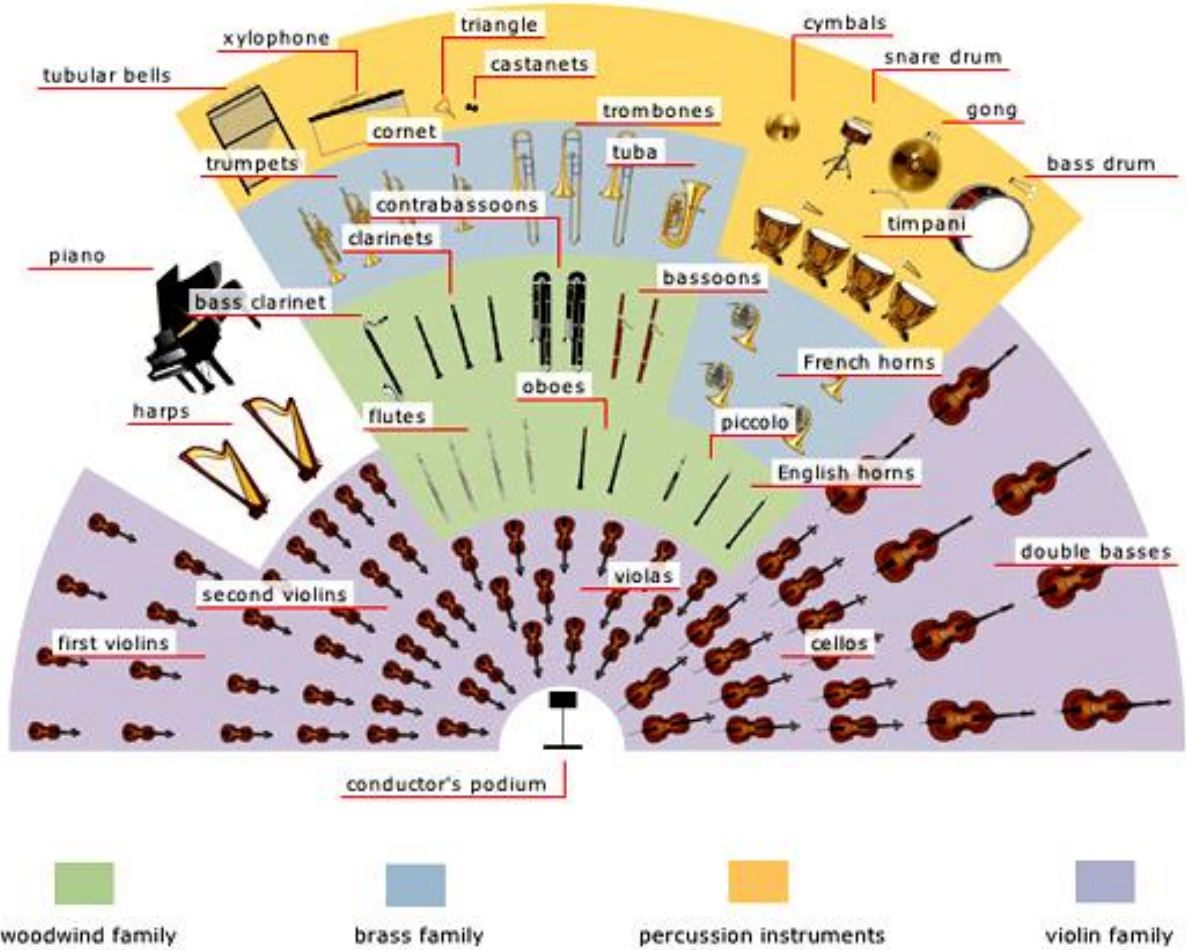
TÉCNICO
LISBOA

Music Signals ...



- ★ **Music is an art form whose medium is sound/audio ... and silence.**
- ★ **The creation, performance, significance, and even the definition of music vary according to culture and social context.**
- ★ **Music can be divided into genres and subgenres, although the dividing lines and relationships between music genres are often subtle, sometimes open to individual interpretation, and occasionally controversial.**
- ★ **The music bandwidth regards the range of audio frequencies which directly influence the fidelity of the music. The higher the audio bandwidth, the better the sound fidelity. The highest practical frequency which the human ear can normally hear is about 20 kHz.**
- ★ **Naturally, music is a very relevant type of audio signal as it is associated to extremely important applications and businesses.**

Musical Instruments for all Tastes ...





Audio Transducers

A transducer is a device (commonly implies the use of a sensor/detector) that converts one form of energy to another. Energy types include (but are not limited to) electrical, mechanical, electromagnetic (including light), chemical, acoustic or thermal energy.



A microphone is an acoustic-to-electric transducer that converts sound into an electrical signal.

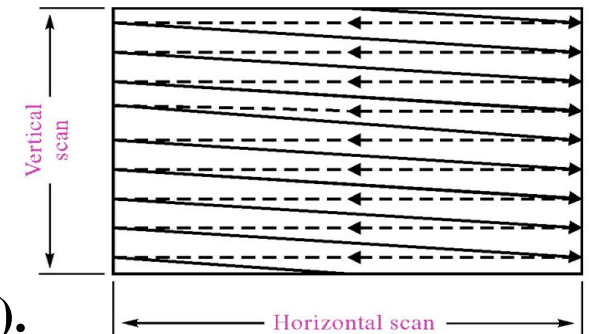


A loudspeaker is an electroacoustic transducer that produces sound in response to an electrical audio signal input.

Image and Video Signals ...



- ★ An image/video signal is a representation of light, typically as an electrical voltage.
- ★ Video corresponds to a succession of images at some temporal rate, typically 25 Hz in Europe and 30 Hz in US (due to different electrical network frequencies).
- ★ In analogue video, each image/frame is represented as a discrete number of lines, with each line represented by a time-continuous waveform.
- ★ Analogue TV video signals have frequencies in the range of roughly 0 to 5 MHz with this value depending on the image/frame rate and number of lines per image (temporal and spatial resolutions).
- ★ Video signals may be synthesized directly or may originate at a transducer such as a camera. Displays convert an electrical video signal into light.



A transducer is a device (commonly implies the use of a sensor/detector) that converts one form of energy to another. Energy types include (but are not limited to) electrical, mechanical, electromagnetic (including light), chemical, acoustic or thermal energy.



A video camera is an light-to-electric transducer used for image acquisition, initially developed by the television industry but now common in many other applications.



A display is an electric-to-light transducer that produces images in response to an electrical video signal.

Text Signals ...

- ★ Text is the representation of written language which is the representation of a language by means of a writing system.
- ★ Text is another form of media corresponding to a sequence of characters that may have to be coded.

قرآن مجید
(عربی متن بمعہ اردو ترجمہ)
از مولانا محمد علی

लड़का: किसी वीरान और सुनसान जगह
चलते हैं
गलफ्रेंड: कुछ ऐसा वैसा तो नहीं करोगे
लड़का: नहीं बिलकुल नहीं
गलफ्रेंड: फिर रहने दो, कल चलेंगे

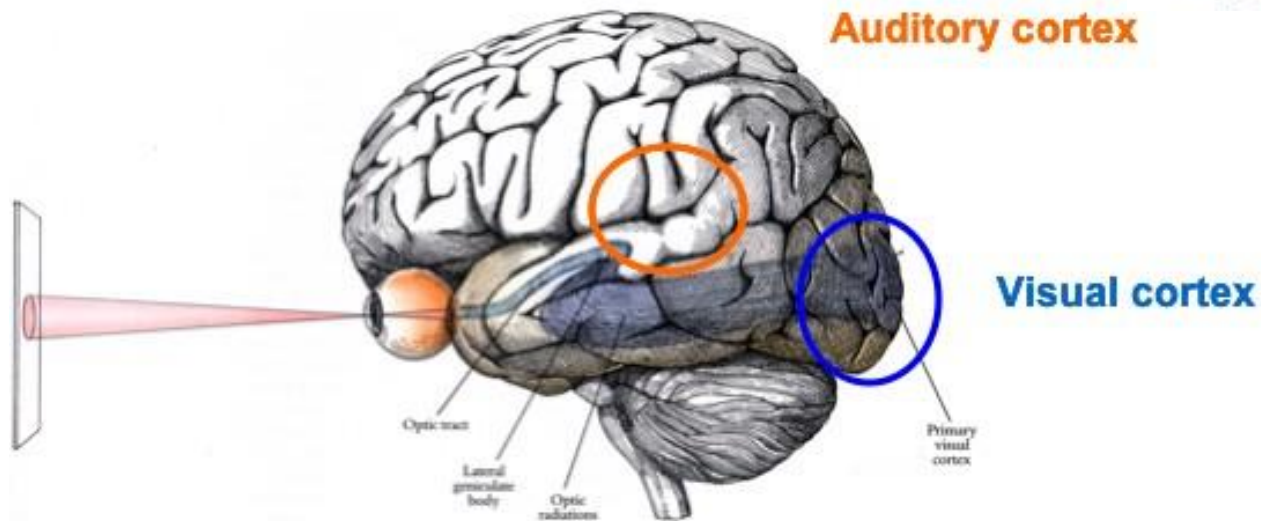


Basics on Human Perception



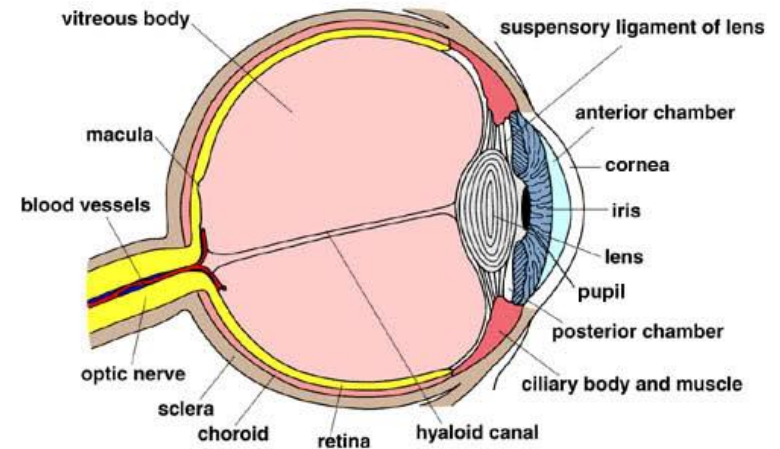
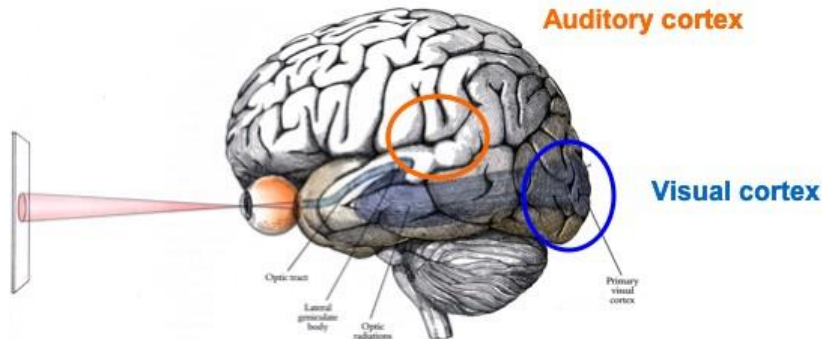
phillipmartin.com

We, the Users ...

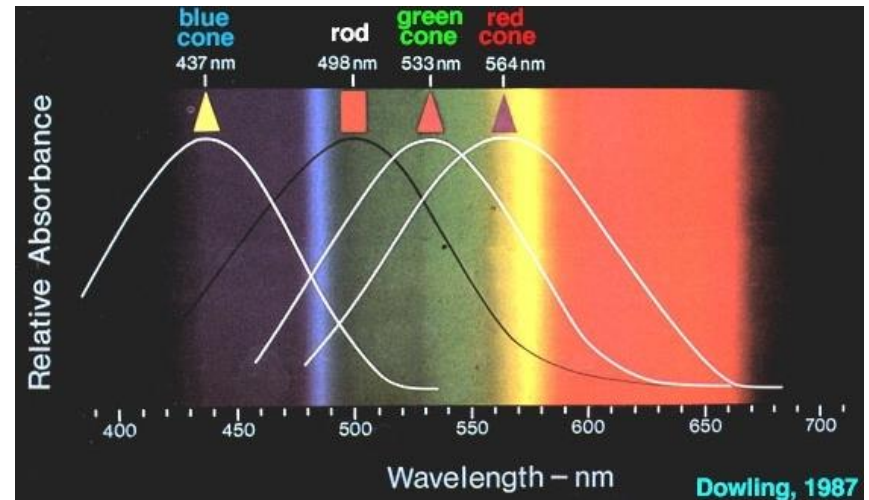
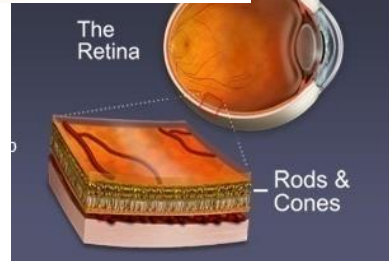
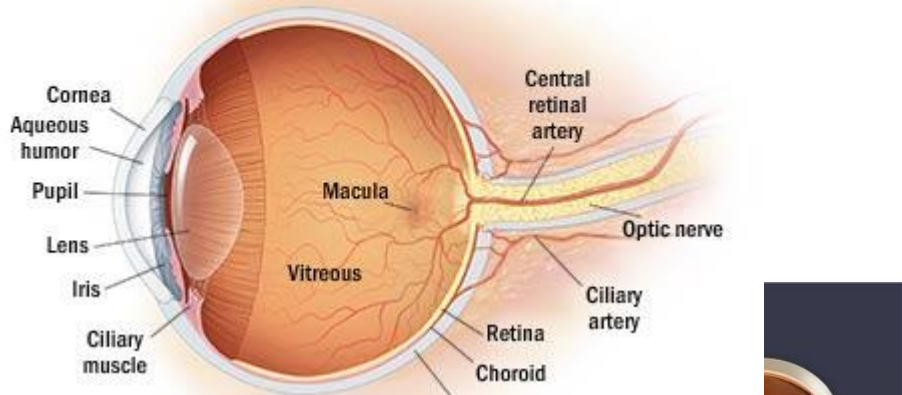
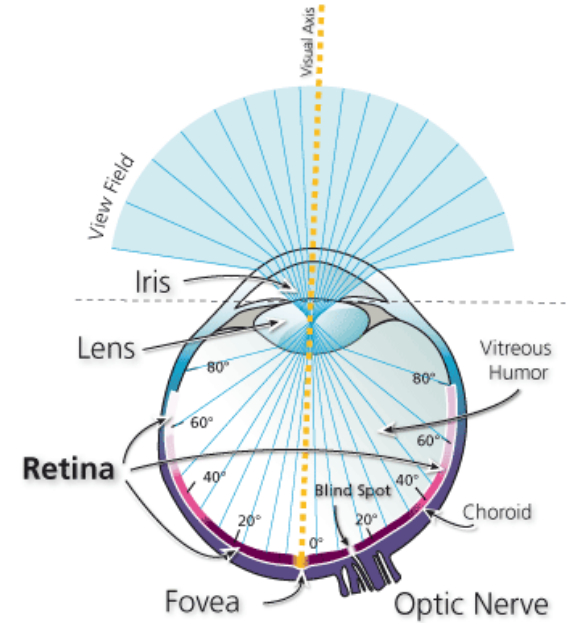
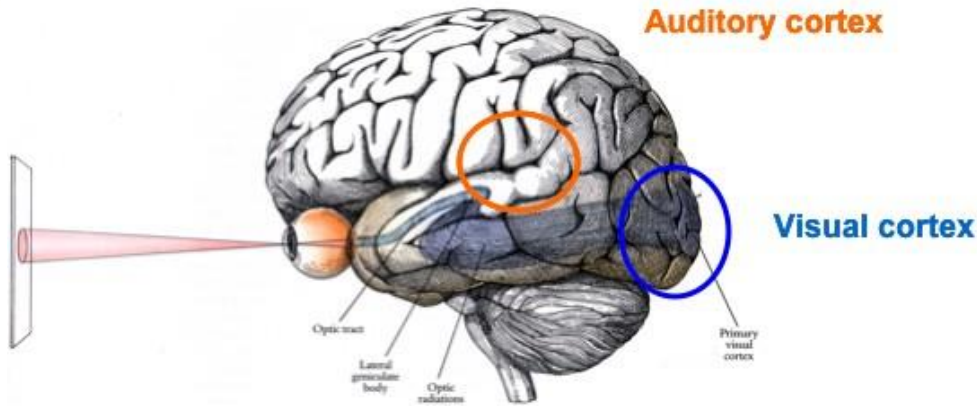


Audiovisual communication services must, above all, satisfy the final user needs, maximizing the quality of the user experience for the available resources !

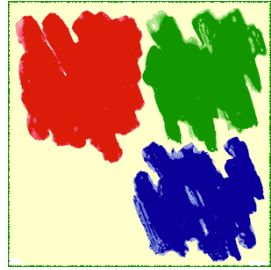
Human Visual System



- ★ **The visual system is the part of the central nervous system which enables organisms to process visual detail. It interprets information from visible light to build a representation of the surrounding world.**
- ★ **The visual system accomplishes a number of complex tasks, including**
 - i) reception of light and the formation of monocular representations;
 - ii) construction of a binocular perception from a pair of 2D projections;
 - iii) identification and categorization of visual objects;
 - iv) assessing distances to and between objects; and
 - v) guiding body movements in relation to visual objects.



Human Visual System: Rods and Cones



Rods (*bastonetes*)

- ★ Photoreceptor cells (about 90 million) in the eye retina that can function in less intense light than the other type of photoreceptor, the cone cells.
- ★ Named for their cylindrical shape, rods are concentrated at the outer edges of the retina and are used in peripheral vision.
- ★ More sensitive than cone cells (100 times more), rod cells are sensitive to luminance and are almost entirely responsible for night vision.

Cones

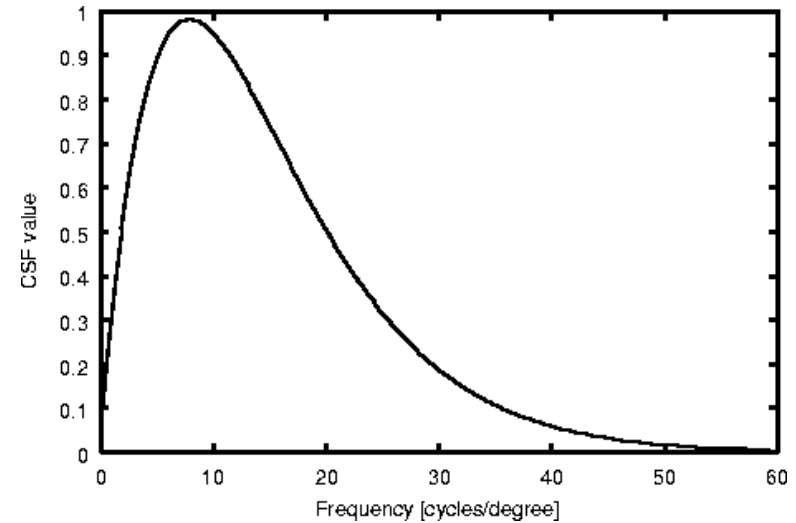
- ★ Less sensitive to light than the rod cells in the retina (which support vision at low light levels), but allow the perception of color.
- ★ The cone cells gradually become sparser towards the periphery of the retina (there are about 4-6 million in the human eye).
- ★ They are also able to perceive finer detail and more rapid changes in images, because their response times to stimuli are faster than those of rods.
- ★ Because humans usually have three kinds of cones with different response curves and, thus, respond to variation in color in different ways, they have trichromatic vision.

Low-Level Vision Modeling

- ★ **Spatial vision** – Characterization of the human visual system in terms of processing spatial data
 - *Human contrast sensitivity function (CSF)*
 - *Masking effects, notably noise, contrast and entropy masking*
 - *Weber's law: the just noticeable variation in luminance against a uniform image is linearly proportional to the background luminance level*
- ★ **Temporal vision** - Characterization of the human visual system in terms of processing temporal data
 - *Adds time to the spatial CSF*
- ★ **Color vision** - Characterization of the human visual system in terms of processing color data
- ★ **Foveation** - describes the non-uniform sensitivity across the field of view resulting from the unequal density of cones in the retina

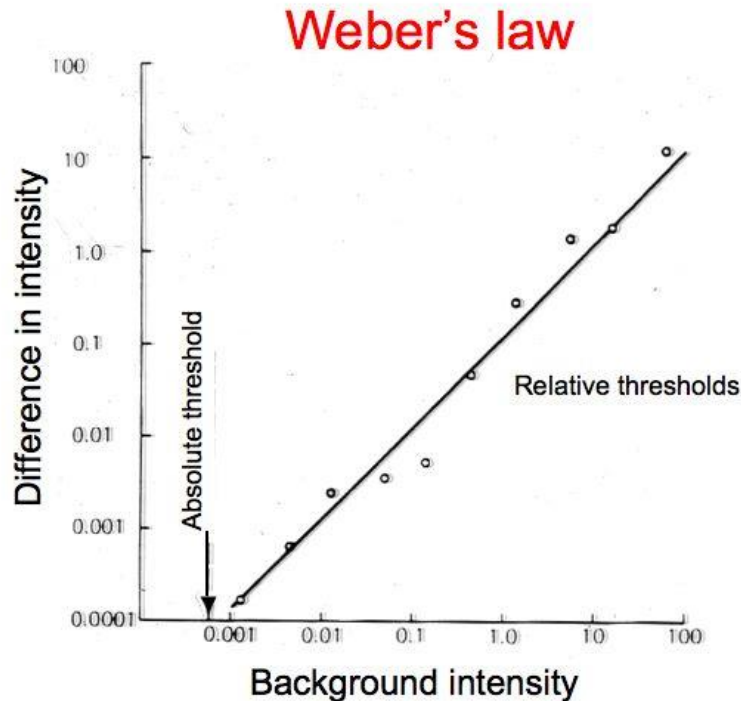
Contrast Sensitivity Function

- ★ **The human Contrast Sensitivity Function (CSF) describes spatial frequency perception.**
- ★ **It is effectively the spatial frequency response of the HVS, i.e., contrast sensitivity versus spatial frequency in units of cycles/degree of visual angle.**
- ★ **The CSF tells how sensitive the HVS is to the various frequencies of visual stimuli.**
- ★ **If the frequency of visual stimuli is too high, the HVS will not be able to recognize the stimuli pattern anymore.**
- ★ **Temporal vision can be characterized by a spatio-temporal CSF, which adds the dimension of frequency (in time) to the spatial CSF.**



For medium frequency, you need less contrast than for high or low frequency to detect the sinusoidal fluctuation

Weber's Law ...



$$\frac{\Delta I}{I} = K$$

← The Weber Fraction

Weber's Law:



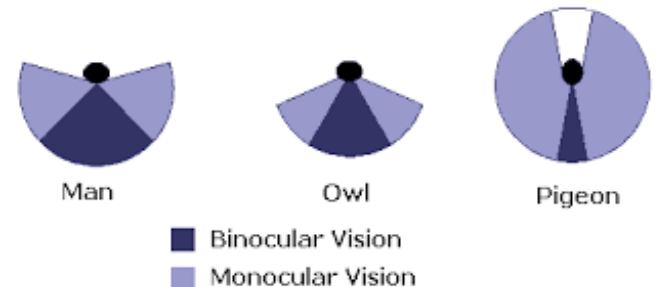
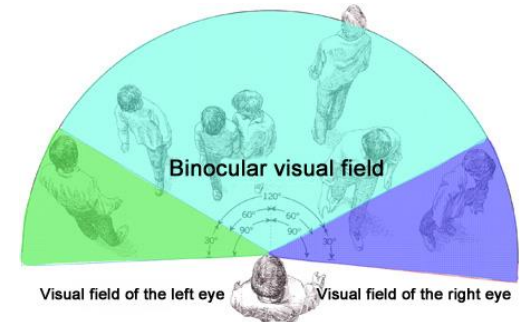
- Example:** When you are in a noisy environment you must shout to be heard while a whisper works in a quiet room.

I don't think this guy understands Weber's Law!

Weber's law states that the just-noticeable difference between two stimuli is proportional to the magnitude of the stimuli.

Binocular Vision

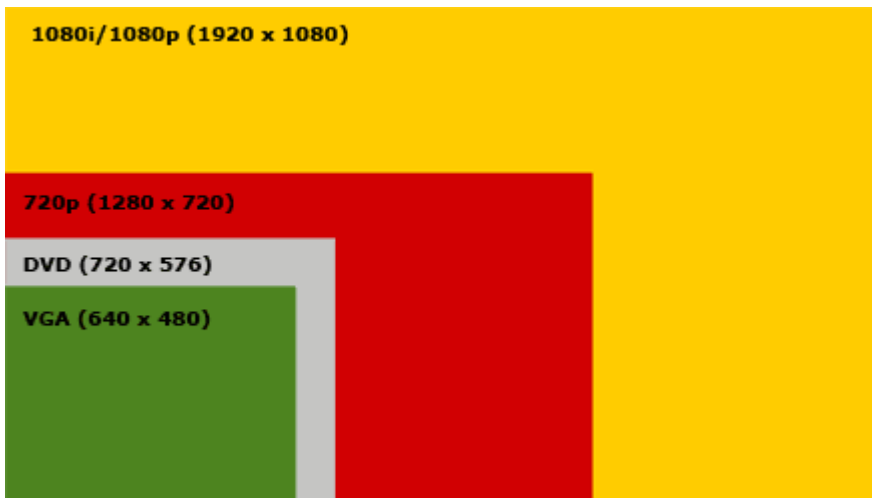
- ★ **Binocular vision is vision in which both eyes are used together.**
- ★ **Having two eyes confers at least four advantages over having one:**
 1. *Gives a creature a spare eye in case one is damaged ...*
 2. *Gives a wider field of view.*
 3. *Gives binocular summation in which the ability to detect faint objects is enhanced (the detection threshold for a stimulus is lower with two eyes than with one).*
 4. *Gives stereopsis in which parallax provided by the two eyes' different positions on the head give precise depth perception.*



Human Visual System: Technical Impacts ...

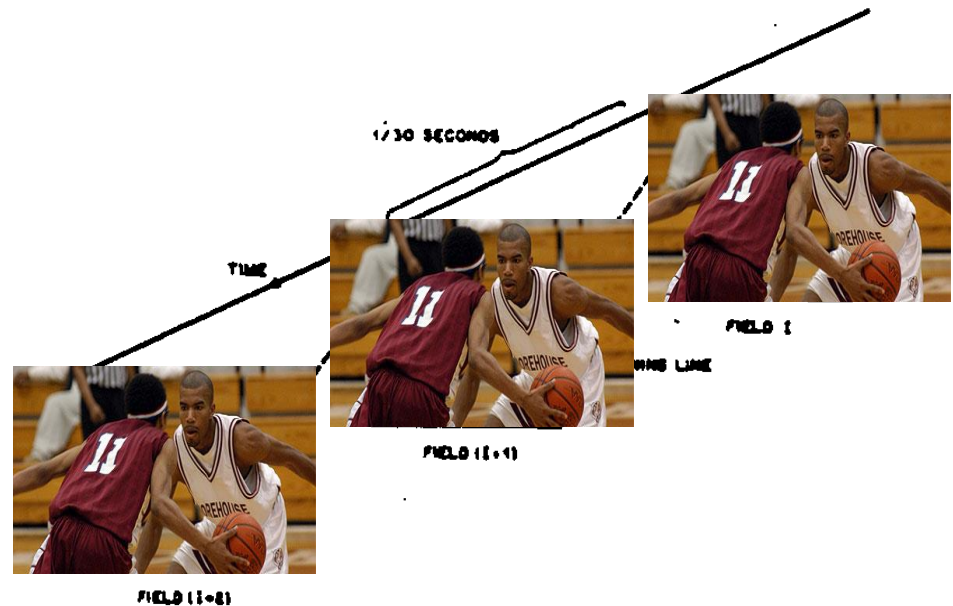
While designing a video system, it is essential to account for:

- ★ The limited human capacity to see spatial detail
- ★ The conditions under which the human visual system reaches the ‘illusion of motion’
- ★ The lower sensibility to color in comparison with luminance/brightness



Illusion of Motion: Temporal Resolution

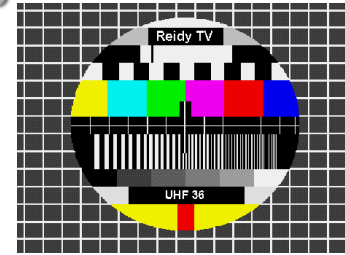
- ★ Video information corresponds to a time varying 2D signal which has to be transformed into a time varying 1D signal to be transmitted using the available channels.
- ★ At the reception, the information is visualized in a 2D display corresponding to the projection (during acquisition) into the camera plane.
- ★ The 2D signal is sampled in time at a rate that guarantees the illusion of motion; this illusion improves with the image rate.



Experience shows that it is possible to get a good illusion of motion up from 16-18 image/s, depending on the image content.

For TV, the frame rate is 25 Hz (Europe) and 30 Hz (US and Japan) due to the electromagnetic interference with the electric network at 50/60 Hz for the old CRT (cathode ray tube) displays.

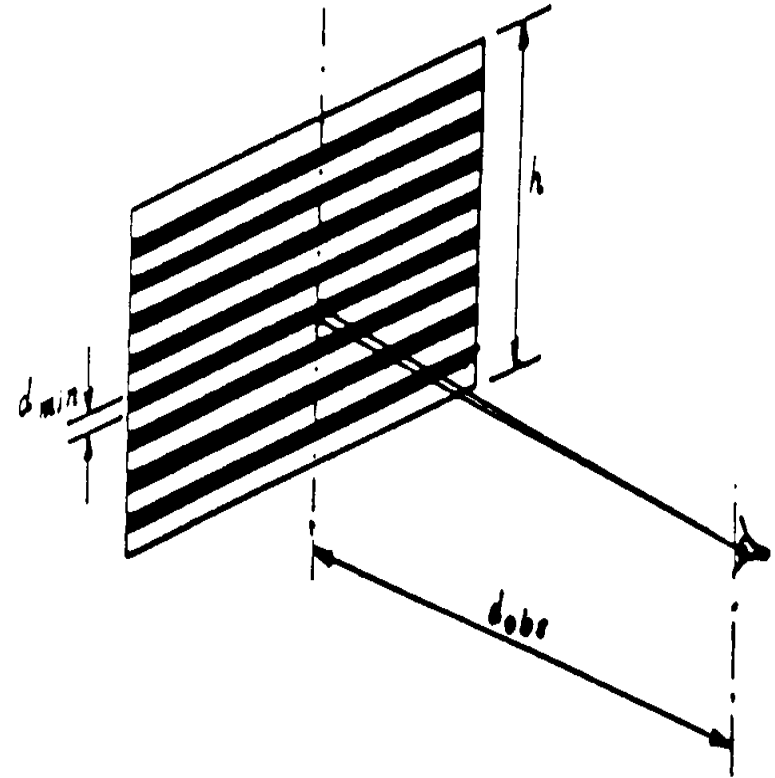
Visual Acuity versus Number of Lines



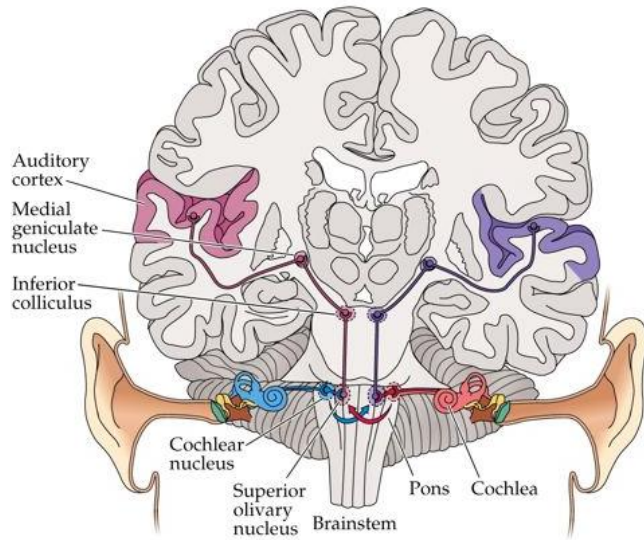
- ★ Visual acuity regards the eye capability of distinguishing (resolving) spatial detail; it is measured with the help of special test images called *Foucault bars images*.
- ★ The visual acuity determines the minimum number of lines in the image in order the user located at a certain distance does not ‘see’ the lines and gains the sensation of spatial continuity.
- ★ The maximum number of lines that the Human Visual System manages to distinguish in a Foucault bars image is given by

$$N_{\max} \sim 3400 h / d_{\text{obs}}$$

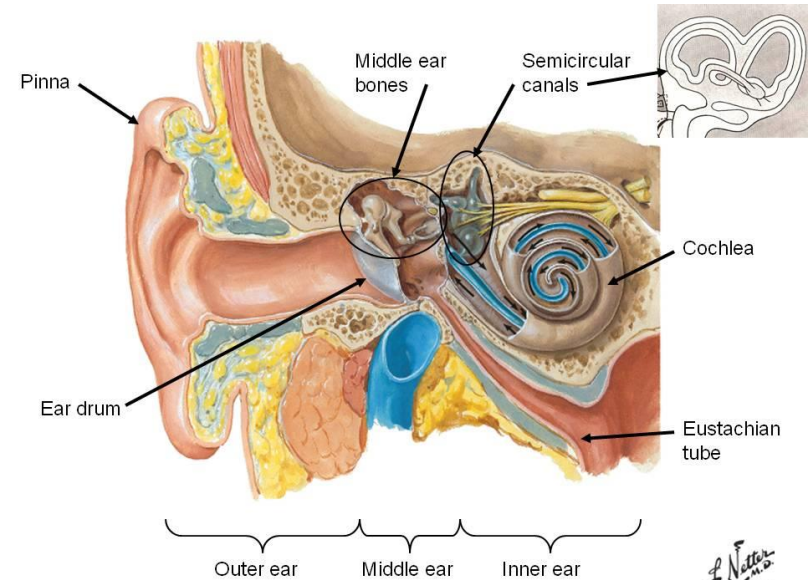
for $d_{\text{obs}}/h \sim 8$, $N_{\max} \sim 425$ lines; $d_{\text{obs}}/h \sim 3$, $N_{\max} \sim 1150$ lines.



Human Auditory System



© 2001 Sinauer Associates, Inc.



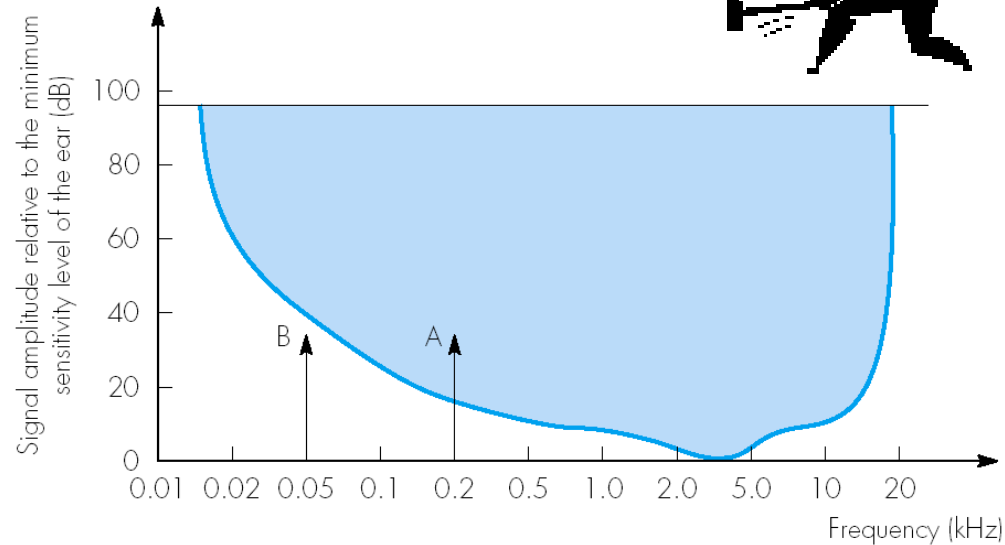
F. Netter M.D.

- ★ **The sensory system for the sense of hearing is the auditory system.**
- ★ **The ability to hear is not found as widely in the animal kingdom as other senses like touch, taste and smell. It is restricted mainly to vertebrates and insects. Within these, mammals and birds have the most highly developed sense of hearing.**

<i>Humans</i>	<i>20-20000 Hz</i>
<i>Whales</i>	<i>20-100000 Hz</i>
<i>Bats</i>	<i>1500-100000 Hz</i>
<i>Fish</i>	<i>20-3000 Hz</i>

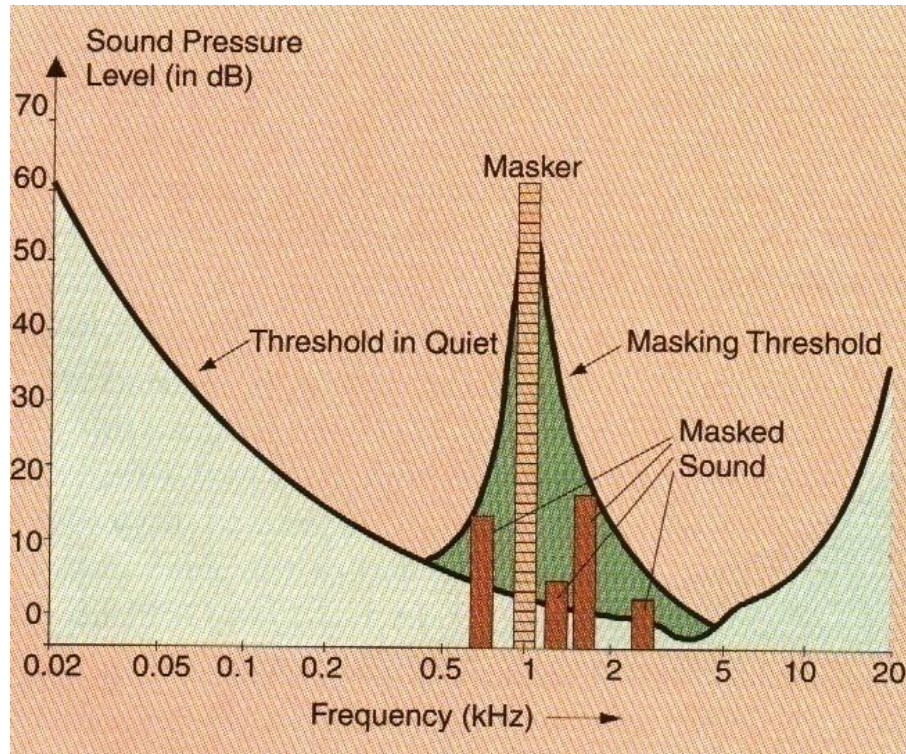
Physiological Effects: the Thresholds

- ★ **Threshold of Hearing** – Defines the minimum sound intensity which may be perceived; this threshold varies along the audio band.
- ★ **Threshold of Feeling or Pain** – Defines the sound intensity above which the sounds may cause pain and provoke hearing damages.



Typically, the threshold of pain is about 120 to 140 dB; sound intensity is measured in terms of Sound Pressure Level relatively to a reference intensity with 10^{-16} W/cm² at 1 kHz.

Audio Frequency Masking



Auditory masking occurs when the perception of one sound is affected by the presence of another sound.

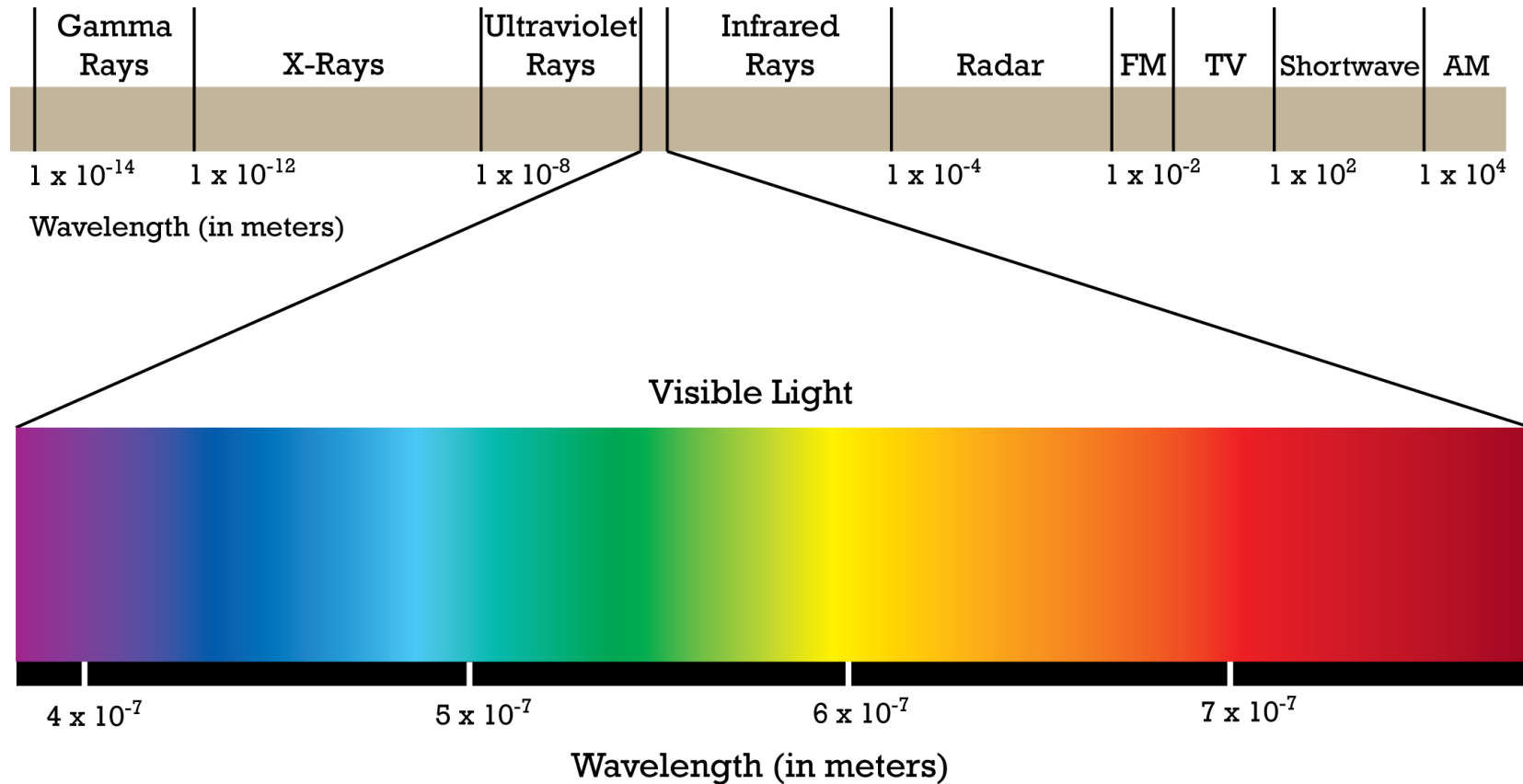
Auditory masking in the frequency domain is known as simultaneous masking, frequency masking or spectral masking.

Visual Signal Representation

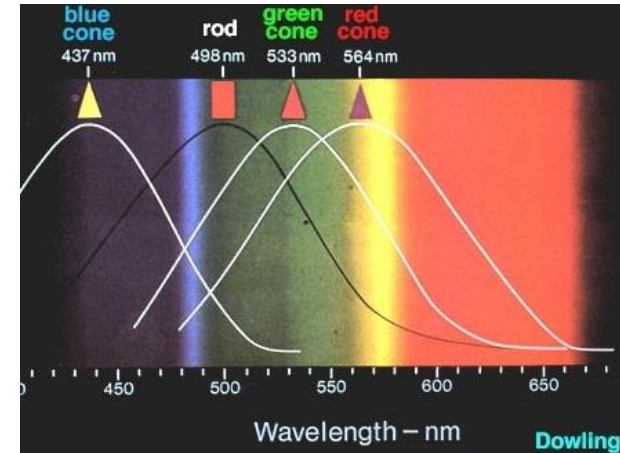
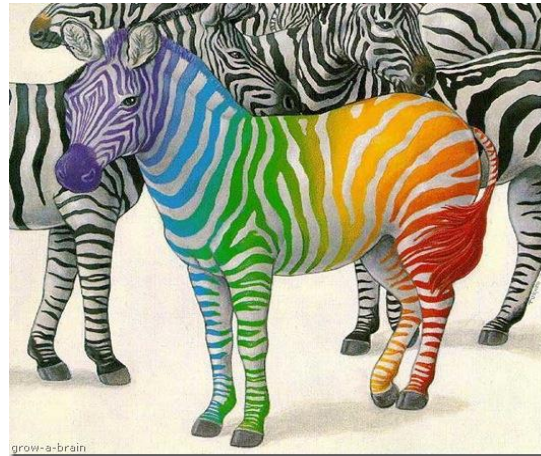




Light Spectrum



Black and White versus Colour



- ★ **Black and white (monochrome) imaging** requires the representation of a single signal called *luminance* which indicates *how much luminous power will be detected by an eye looking at the surface from a particular angle of view*. Luminance is thus an indicator of how bright the surface will appear.
- ★ For **colour imaging** visually acceptable results, it is necessary (and almost sufficient) to use three color signals, which are interpreted as coordinates in some *color space*. The RGB color space is commonly used in cameras and displays, but other spaces such as YCbCr and HSV are often used in other contexts.

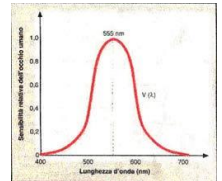
Monochrome Video: Luminance Signal

Luminance is a photometric measure of the luminous intensity per unit area of light travelling in a given direction. It describes the amount of light that passes through or is emitted from a particular area, and falls within a given solid angle.

- ★ The **luminous flux** radiated by a luminous source with a power spectrum $G(\lambda)$ is given by:

$$\Phi = k \int G(\lambda) y(\lambda) d\lambda \quad [\text{lm or lumen}] \quad \text{with } k=680 \text{ lm/W}$$

where $y(\lambda)$ is the average sensibility function of the human eye

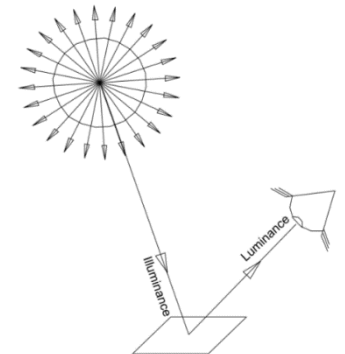


- ★ The way the radiated power is distributed by the various directions is given by the **luminous intensity**:

$$J_L = d\Phi / d\Omega \quad [\text{lm/sr or vela (cd)}]$$

- ★ For video systems, the relevant quantity is the **luminance** of a surface element dS when it is observed with an angle θ such that the surface orthogonal to the observation direction is dS_n

$$Y = dJ_L / dS_n \quad [\text{lm/sr/m}^2]$$



Luminance is what gets into your eye

which corresponds to **the luminous flux, per solid angle, per unit of area.**





A Bit of Colorimetry ...

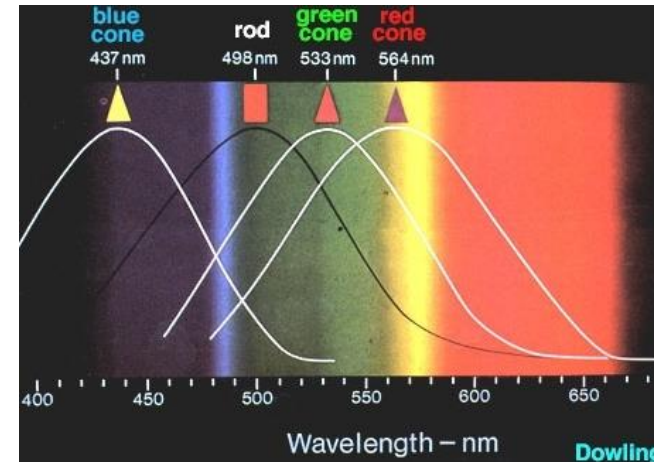
- ★ *“Colour is a property of the mind and not of the objects in the world; it results from the interaction of a light source, an object, and the visual system.” Newton*
- ★ Colorimetry studies show that it is possible to reproduce a high number of colours through the addition of only 3 (carefully chosen) primary colours.
- ★ The primary colours used in most cameras and displays to generate most of the other colours are

- Vermelho (RED)
- Verde (Green)
- Azul (Blue)

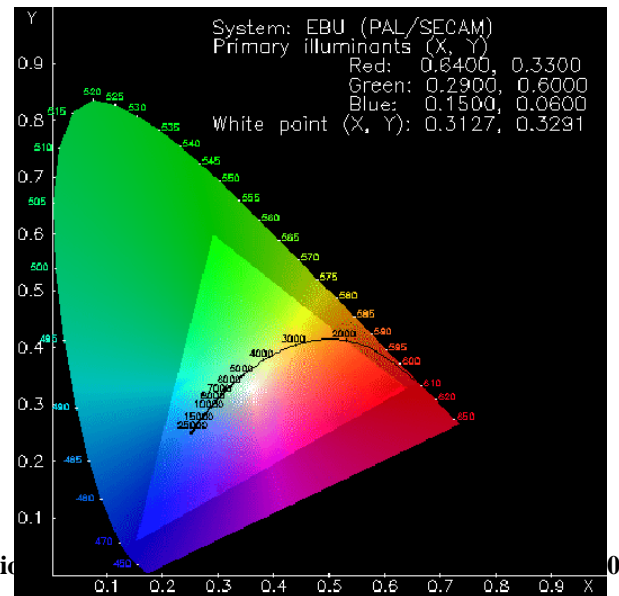
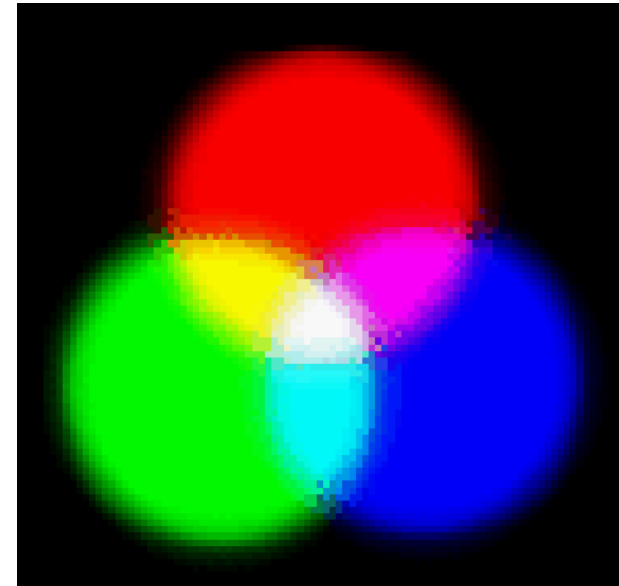
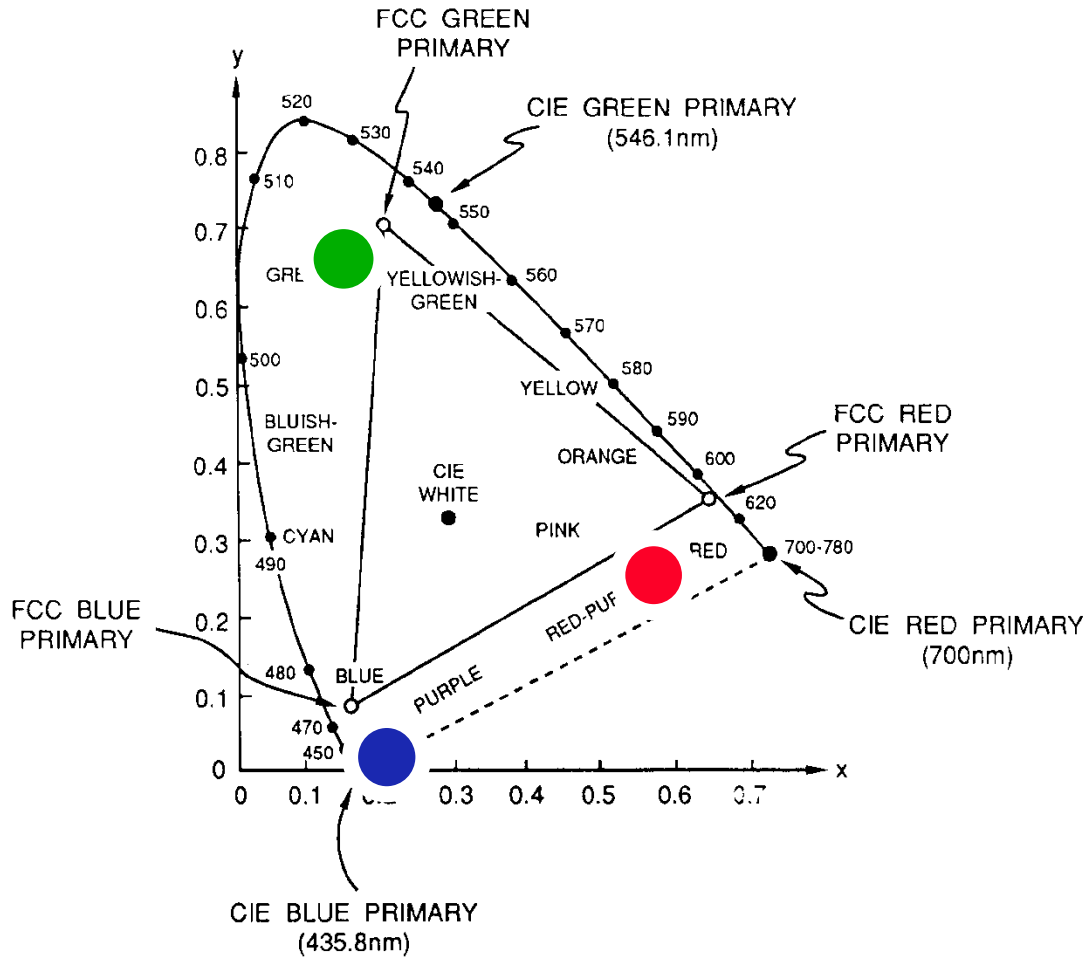


- ★ Luminance, Y , may still be obtained from the primary colours as

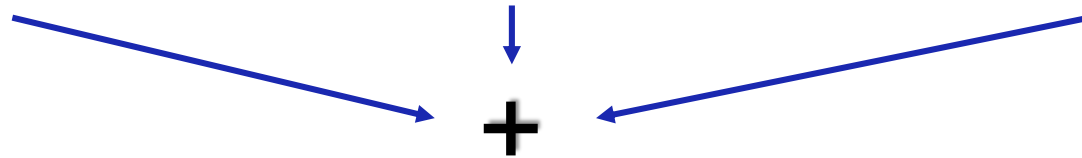
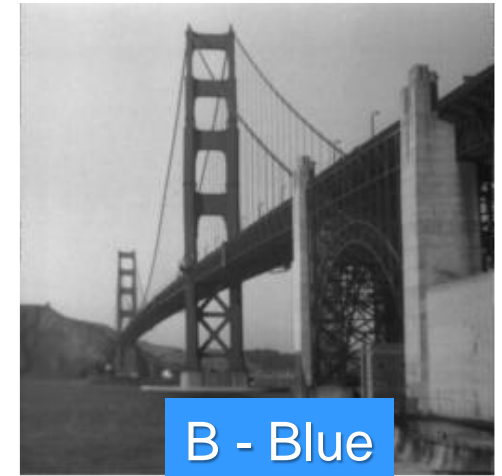
$$Y = 0.3 R + 0.59 G + 0.11 B$$



Chromaticity Diagram and Colour Gamut

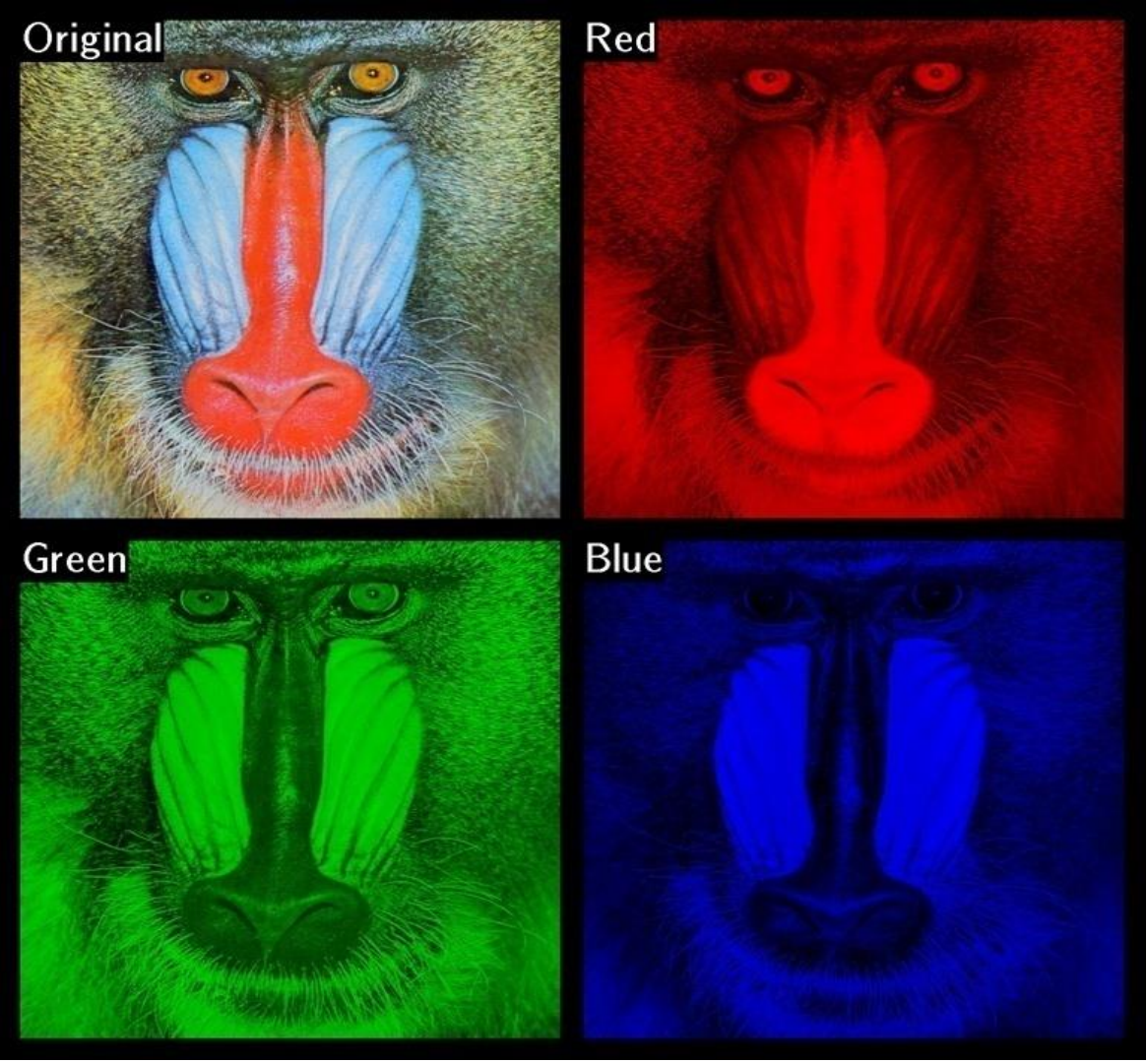
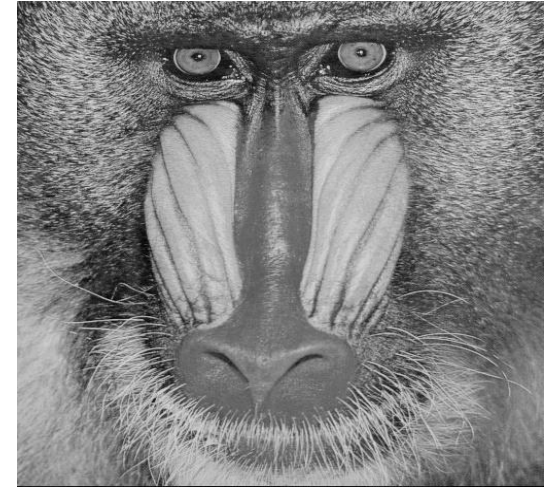


Chromaticity is an objective specification of a color regardless of its luminance, that is, as determined by its hue and saturation.



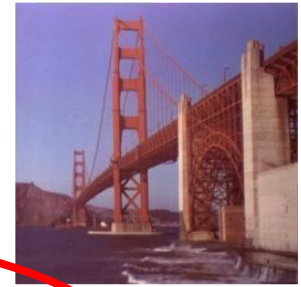
$$Y = 0.3 R + 0.59 G + 0.11 B$$





$$Y = 0.3 R + 0.59 G + 0.11 B$$

Luminance and 2 Chrominances ...



$Y = 0.30R + 0.59G + 0.11B \sim 5 \text{ MHz}$
 $B - Y = U \sim 1-2 \text{ MHz}$
 $R - Y = V \sim 1-2 \text{ MHz}$



Y - Luminance

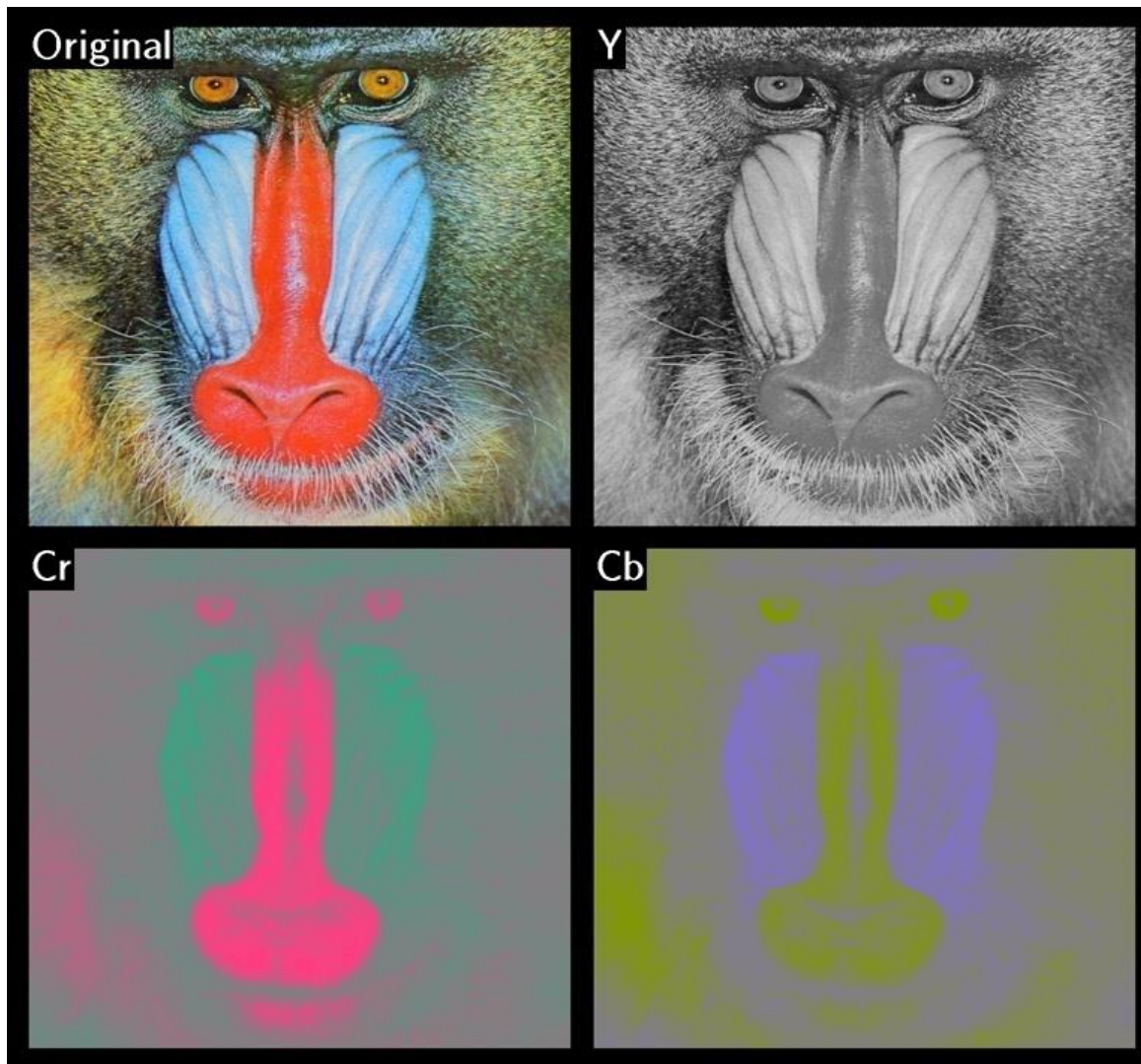


B - Y = U



R - Y = V

$$Y = 0.30R + 0.59G + 0.11B$$



$$B - Y = U$$
$$R - Y = V$$

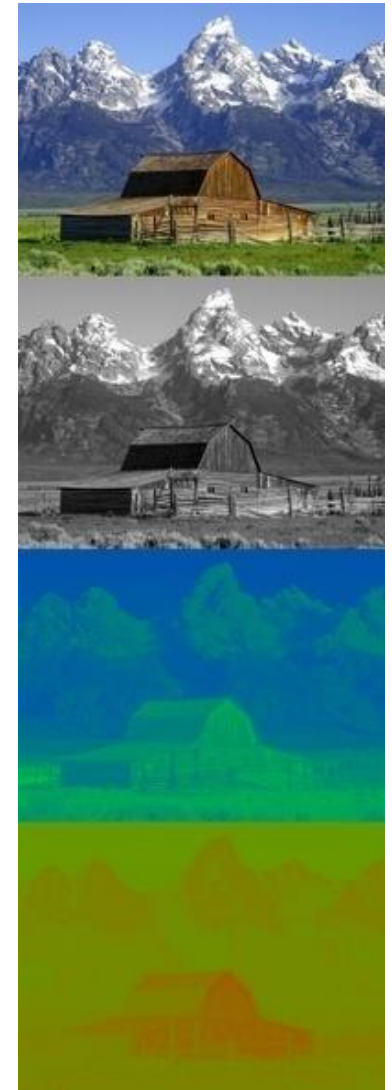
Why YUV and not RGB for Transmission?

YUV is another color space (beyond RGB) for representing a color image

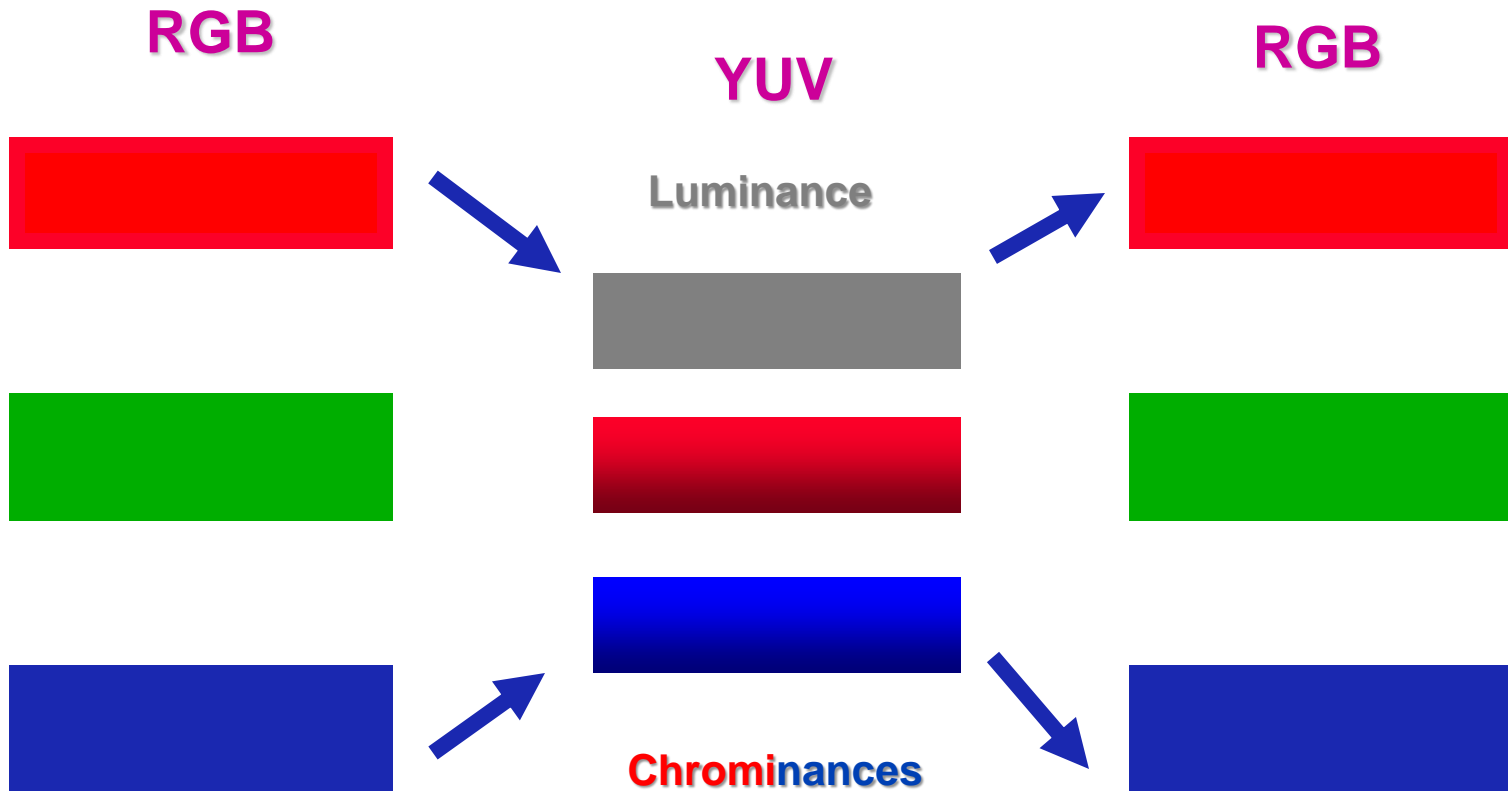
1. *Taking human perception into account to allow reduced bandwidth (this means compression) for chrominance components*
2. *Typically enabling transmission errors or compression artifacts to be more efficiently masked by the human perception than using a 'direct' RGB representation.*

While other color spaces have similar properties, an additional reason to adopt YUV would be for better interfacing analog and digital television and also photographic equipment that conform to certain YUV standards.

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$



Acquisition, Transmission and Synthesis Signals ...



The Analogue World: Systems





Main Analogue AV Systems



- ★ **Telephone** - The telephone is a telecommunications device that transmits and receives sounds, usually the *human voice*. Telephones are a point-to-point communication system whose most basic function is to allow two people separated by large distances to talk to each other. **±1880**
- ★ **Radio** - Radio broadcasting is a *one-way* wireless transmission of audio (notably music) signals over radio waves intended to reach a wide audience. Stations can be linked in radio networks to broadcast a common radio format, either in broadcast syndication or simulcast or both. **±1905**
- ★ **Television** - Television (TV) is a telecommunication medium for transmitting and receiving moving images that can be monochrome (black-and-white) or colored, with accompanying sound. "Television" may also refer specifically to a television set, television programming, or television transmission. **±1920**



Analogue TV Systems

- ★ **Monochrome** – Only the luminance signal is transmitted; systems with a different number of lines per frame have existed.
- ★ **Colour** – Three signals – *luminance plus two chrominance signals* – are transmitted; systems with a different number of lines per frame exist.



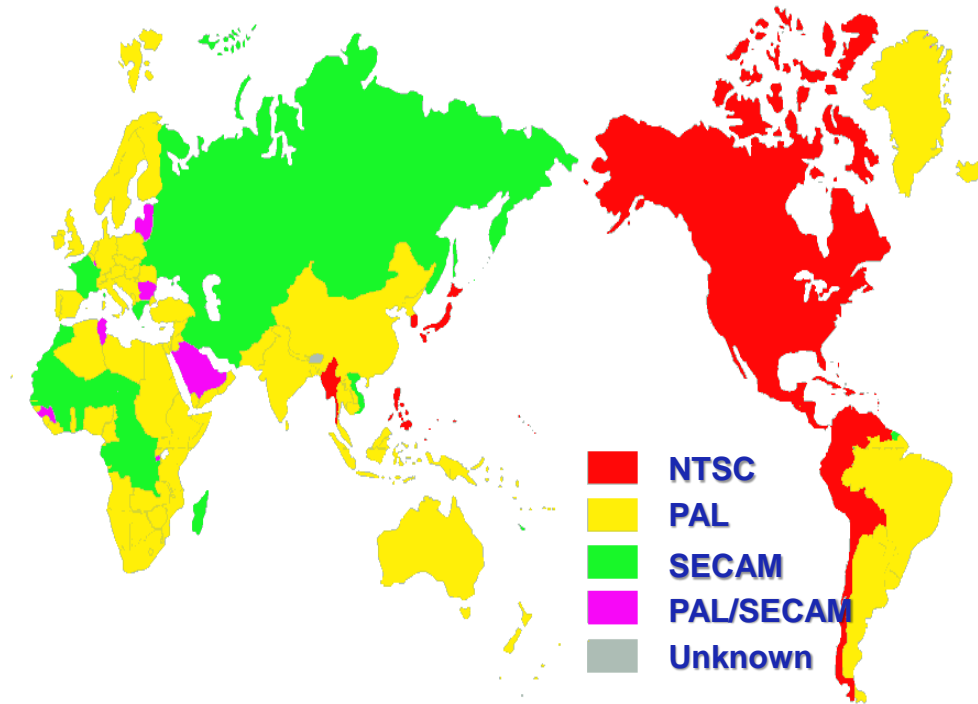
- *National Television System Committee (NTSC)*



- *Phase Alternate Line (PAL)*



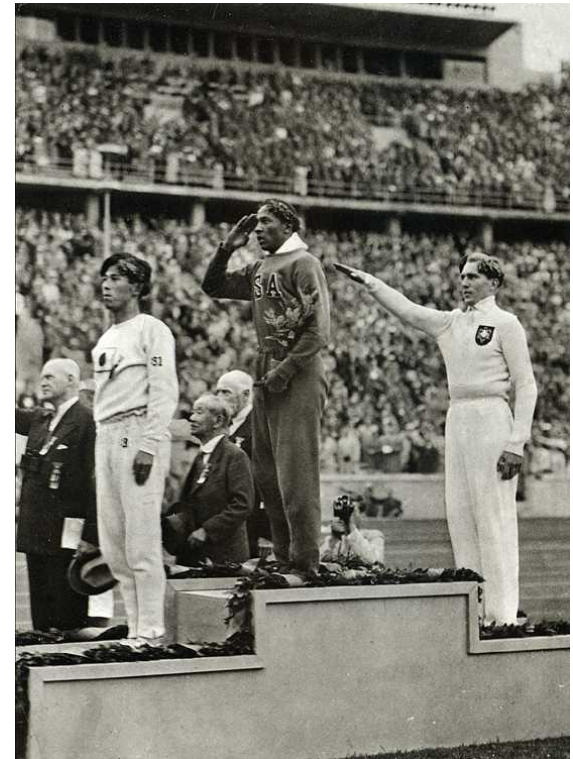
- *Séquentiel couleur à mémoire (SECAM)*



The Starting of Analogue TV ...



1936 Olympic Issue of the German Magazine „Die Woche“ (The Week)



Portuguese TV Milestones

- ★ **1957 – Start of black and white emission with one RTP channel.**
- ★ **1968 – Start of the emissions for the second channel, RTP2.**
- ★ **1972 – Start of RTP Madeira.**
- ★ **1975 – Start of RTP Açores.**
- ★ **1980 – Start of regular colour TV emissions.**
- ★ **1992 – Start of SIC emissions, the first private TV channel.**
- ★ **1993 – Start of TVI emissions, the second private TV channel.**
- ★ **1994 – Start of cable TV.**
- ★ **2012 – Switch off of the analogue emissions and start of digital TV emissions with DVB-T.**



From Analogue to Digital



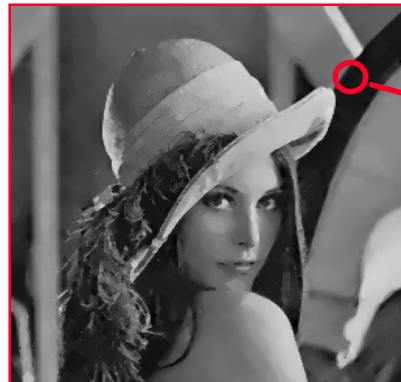


Process of expressing analogue data in digital form.

Analogue data implies ‘continuity’ while digital data is concerned with discrete states, e.g. symbols, digits.

Advantages of digitization:

- ★ Easier to process
- ★ Easier to compress
- ★ Easier to multiplex
- ★ Easier to protect
- ★ Lower powers
- ★ ...



134 135 132 12 15...
133 134 133 133 11...
130 133 132 16 12...
137 135 13 14 13...
140 135 134 14 12...

Sampling or Time Discretization

Sampling is the process of obtaining a periodic sequence of samples to represent an analogue signal.

Sampling is governed by the Sampling Theorem which states that:

An analog signal may be fully reconstructed from a periodic sequence of samples if the sampling frequency is, at least, twice the maximum frequency present in the signal.

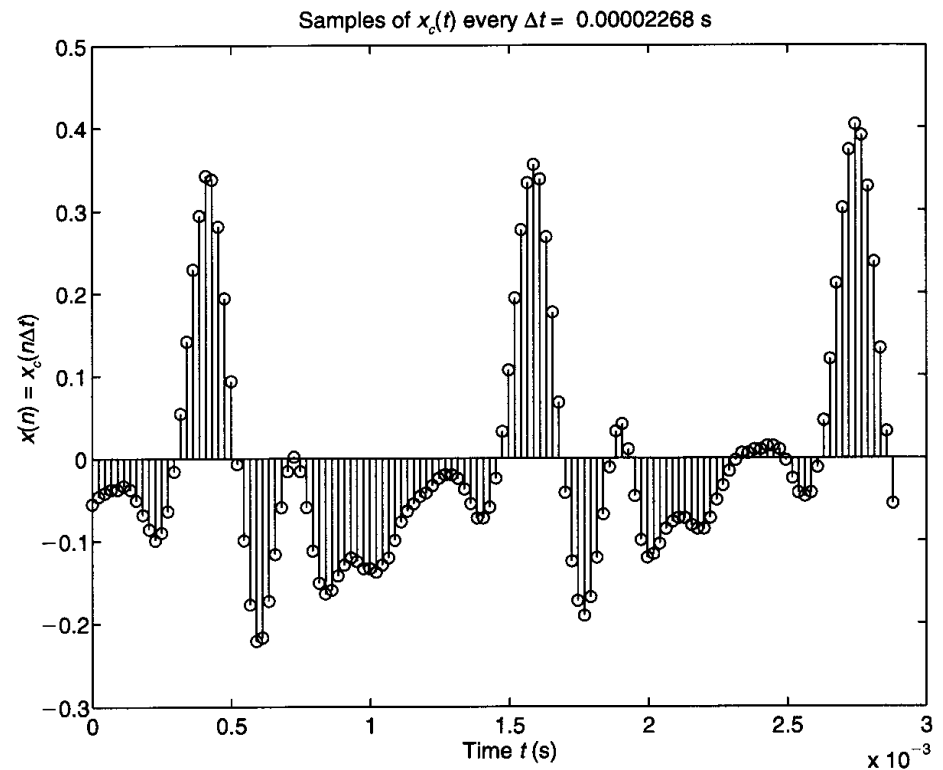
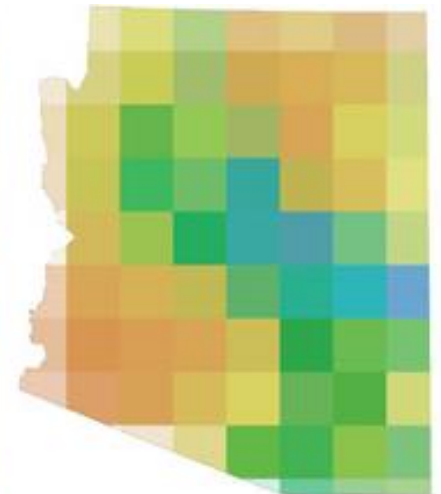
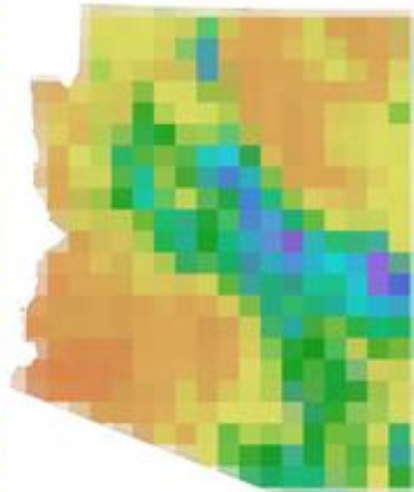
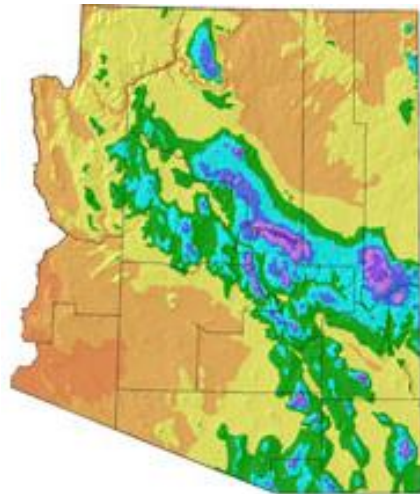
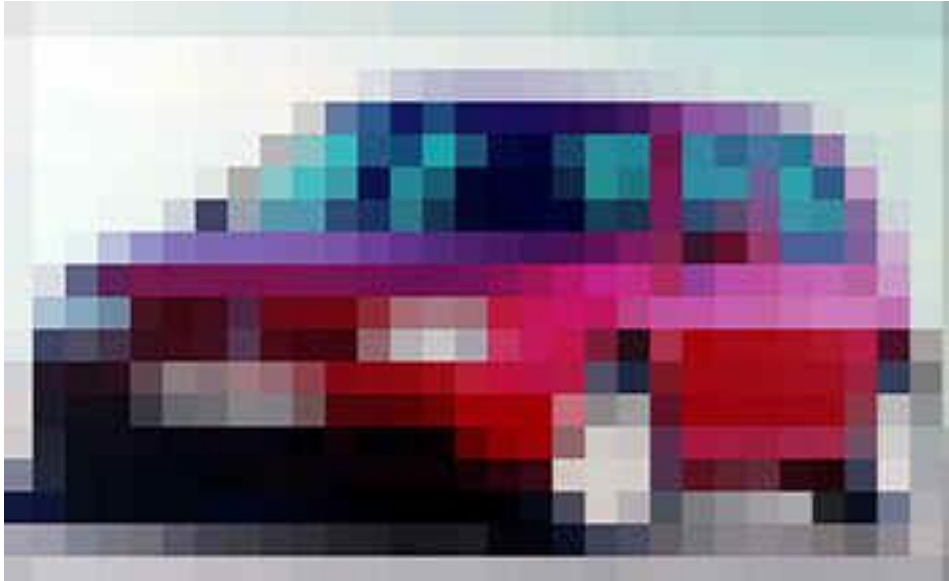


Image Sampling

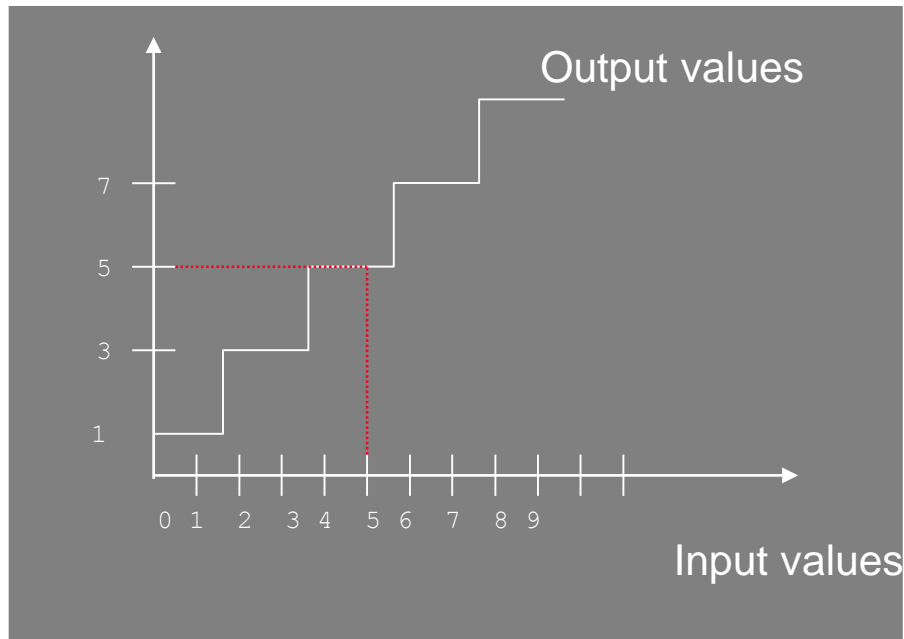


The number of samples (resolution) of an image is very important to determine the ‘final fidelity/quality’.

The required resolution must take into account at least the content, the human visual system and the display conditions.

Quantization or Amplitude Discretization

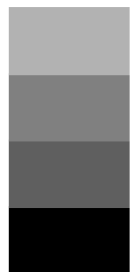
Quantization is the process in which the continuous range of values of a sampled input analogue signal is divided into non-overlapping subranges; to each subrange, a discrete value of the output is uniquely assigned.



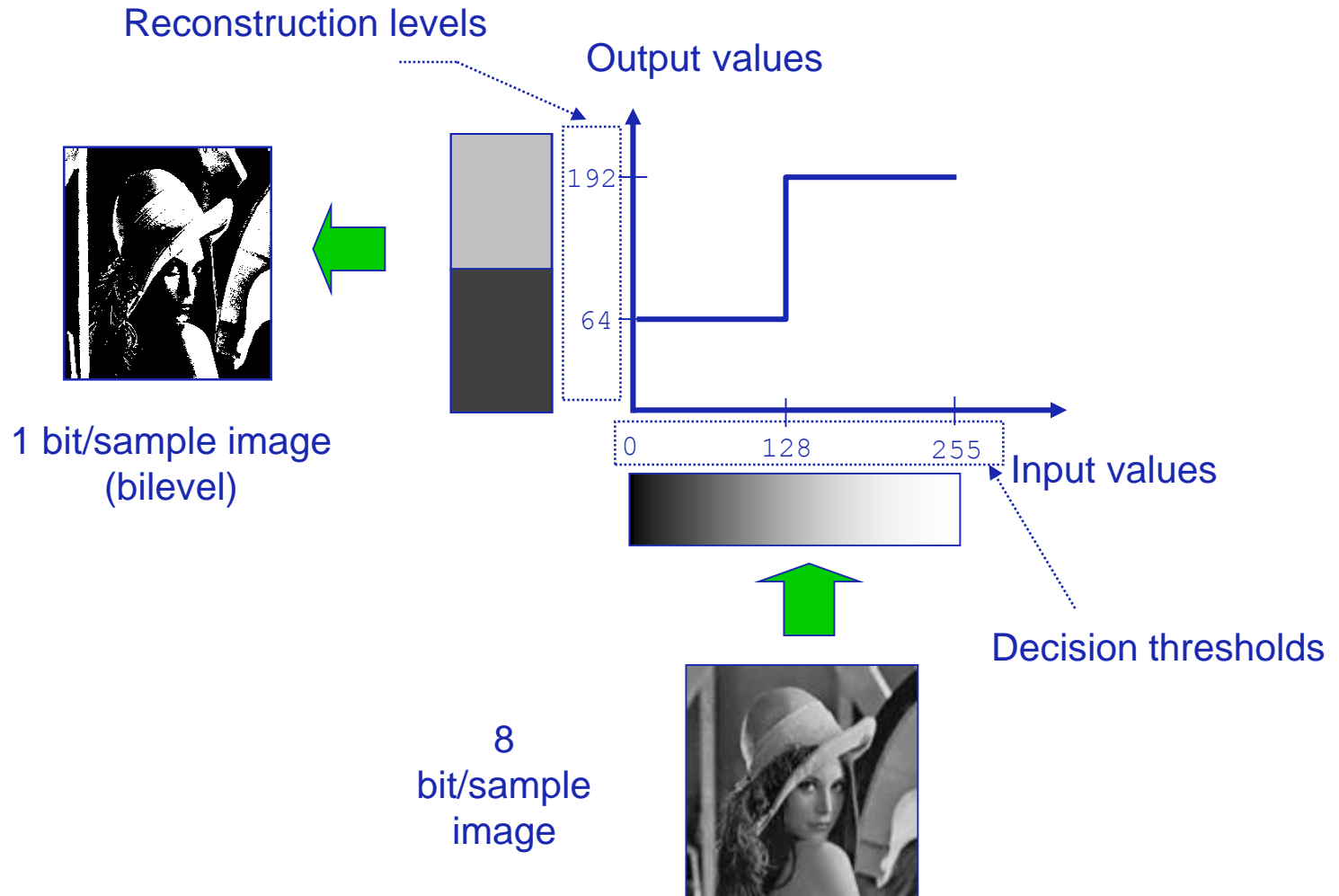
Continuous input



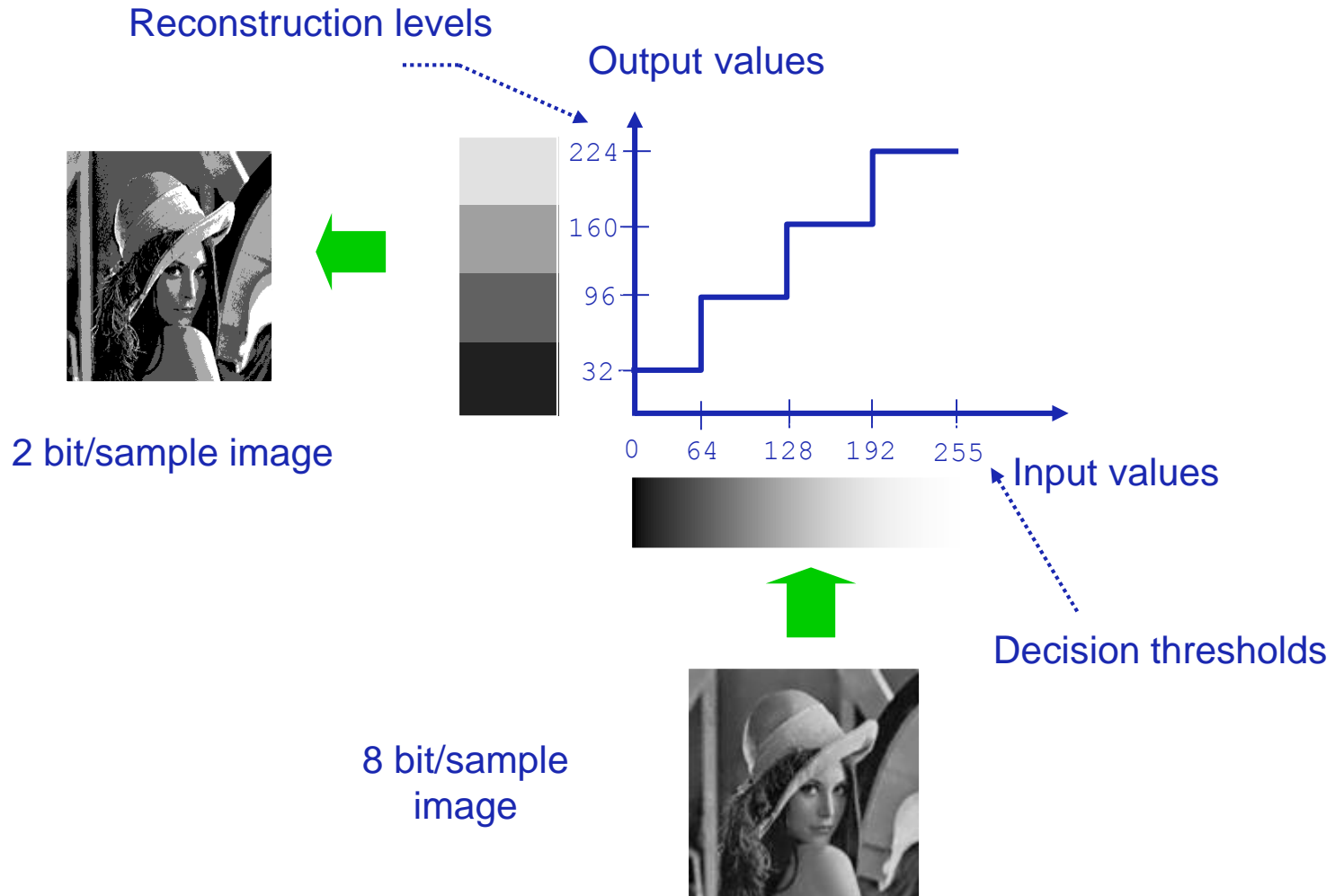
Discrete output



2 Levels Quantization



4 Levels Quantization





TÉCNICO
LISBOA

Uniform Quantization



4 bit/sample
0000, 0001,
0010, 0011, ...



2 bit/sample
00, 01, 10, 11

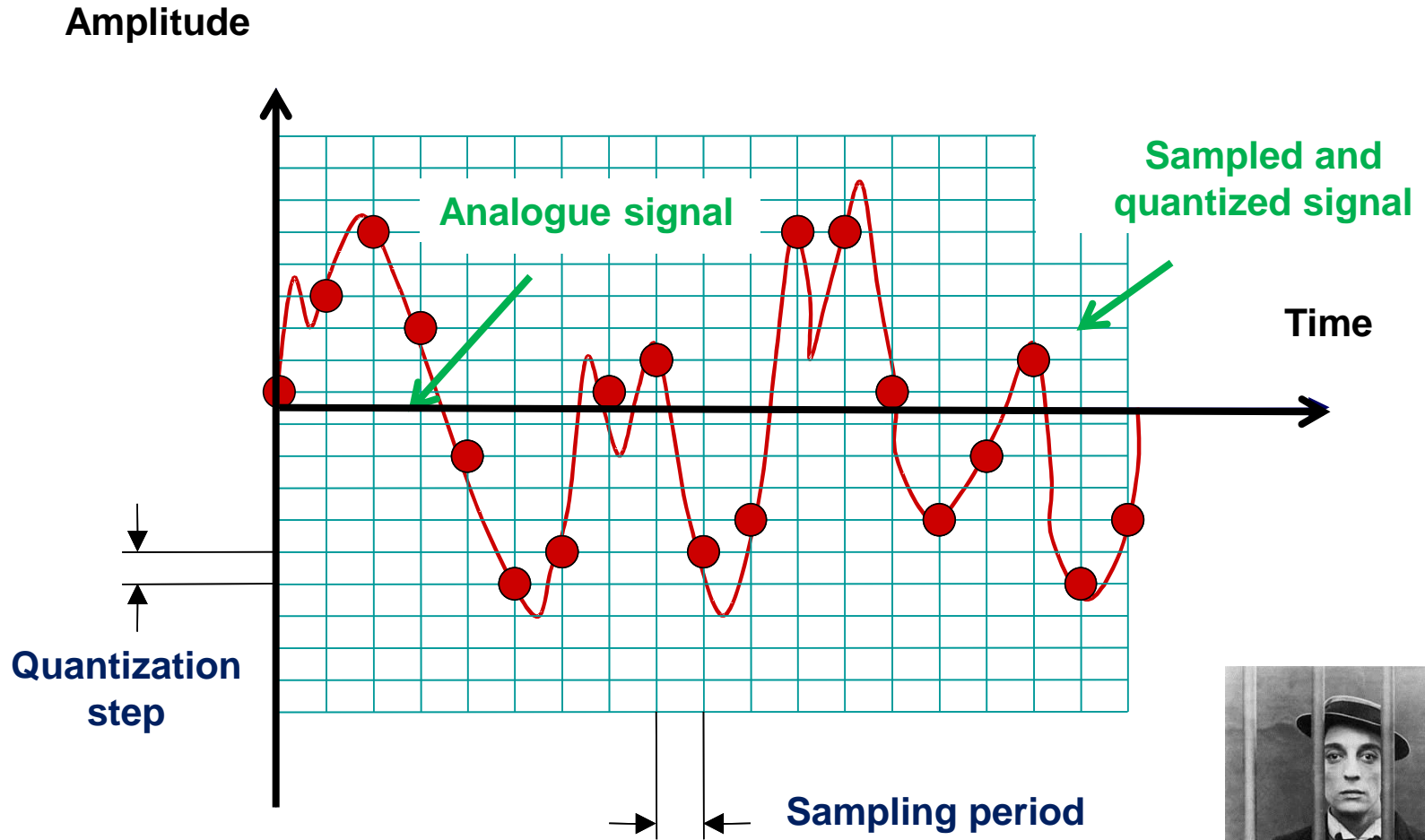


3 bit/sample
000, 001, 010,
011, 100, 101,
110, 111

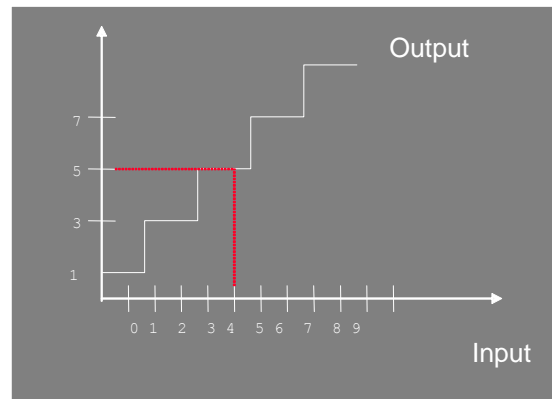
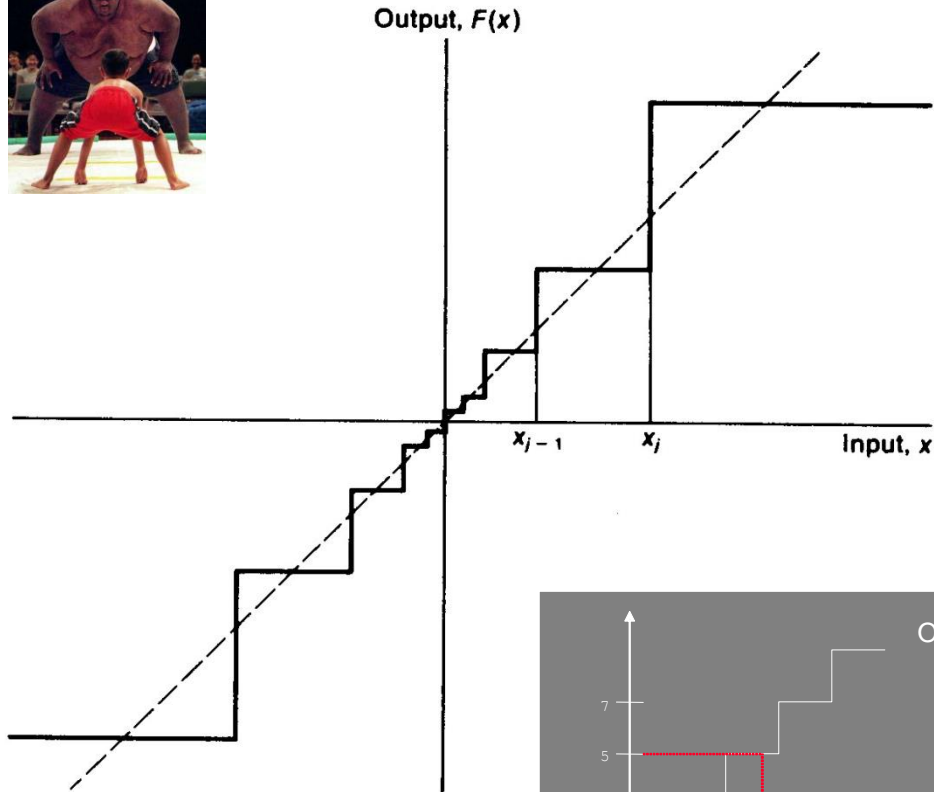


1 bit/sample
0, 1

Digitization: the Signal 'Behind the Bars' ...



Non-Uniform Quantization

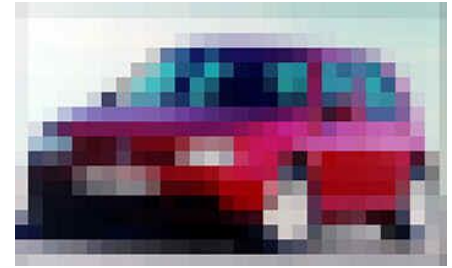


For many signals, e.g. speech, uniform or linear quantization is not a good solution in terms of minimizing the mean square error (and thus the Signal to Quantization noise Ratio, SQR) due to the non-uniform statistics of the signal.

Also to get a certain SQR, lower quantization steps have to be used for lower signal amplitudes and vice-versa.

Pulse Code Modulation (PCM)

The Raw/Original Format

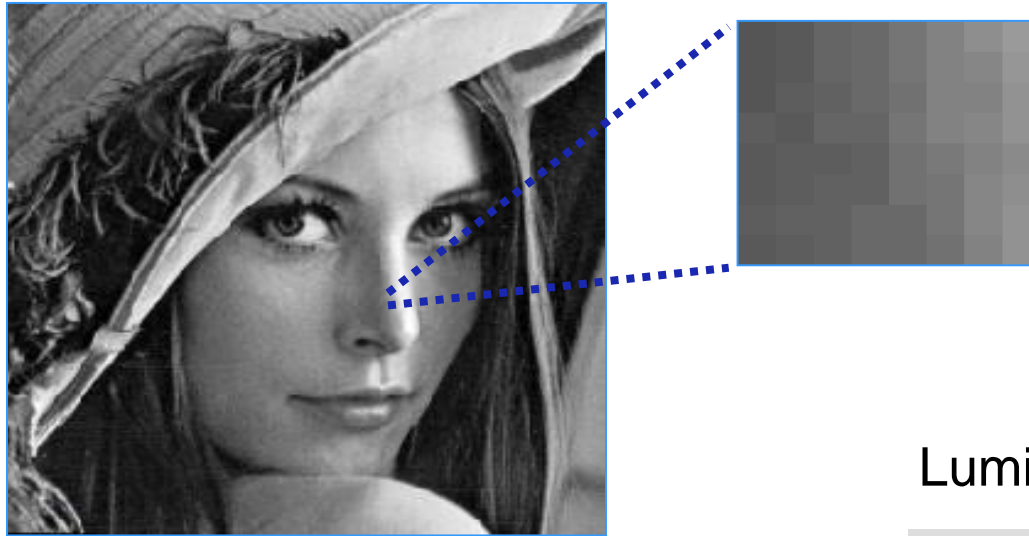


PCM is the simplest form of digital source representation/coding where each sample is independently represented with the same number of bits.

- ★ **Example 1: Image with 200×100 samples at 8 bit/sample takes $200 \times 100 \times 8 = 160000$ bits with PCM coding**
- ★ **Example 2: 11 kHz bandwidth audio at 8 bit/sample takes $11000 \times 2 \times 8 = 176$ kbit/s with PCM coding**

Being the simplest form of coding, as well as the least efficient, PCM is typically taken as the reference/benchmark coding method to evaluate the performance of more powerful (source) coding/compression algorithms.

Image, Samples and Bits ...



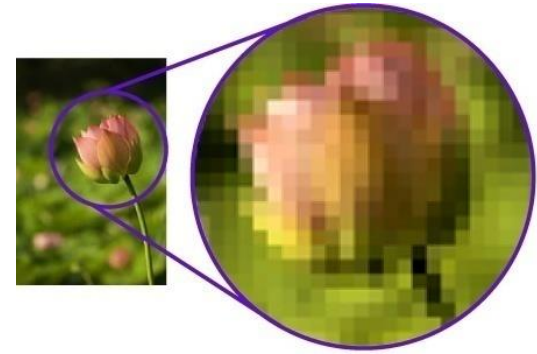
Binary representation
 8 bit/sample -> 256 (2^8) levels

87 = **0101 0111**
 130 = **1000 0010**

Luminance =

87	89	101	106	118	130	142	155
85	91	101	105	116	129	135	149
86	92	96	105	112	128	131	144
92	88	102	101	116	129	135	147
88	94	94	98	113	122	130	139
88	95	98	97	113	119	133	141
92	99	98	106	107	118	135	145
89	95	98	107	104	112	130	144

Samples versus Pixels ...

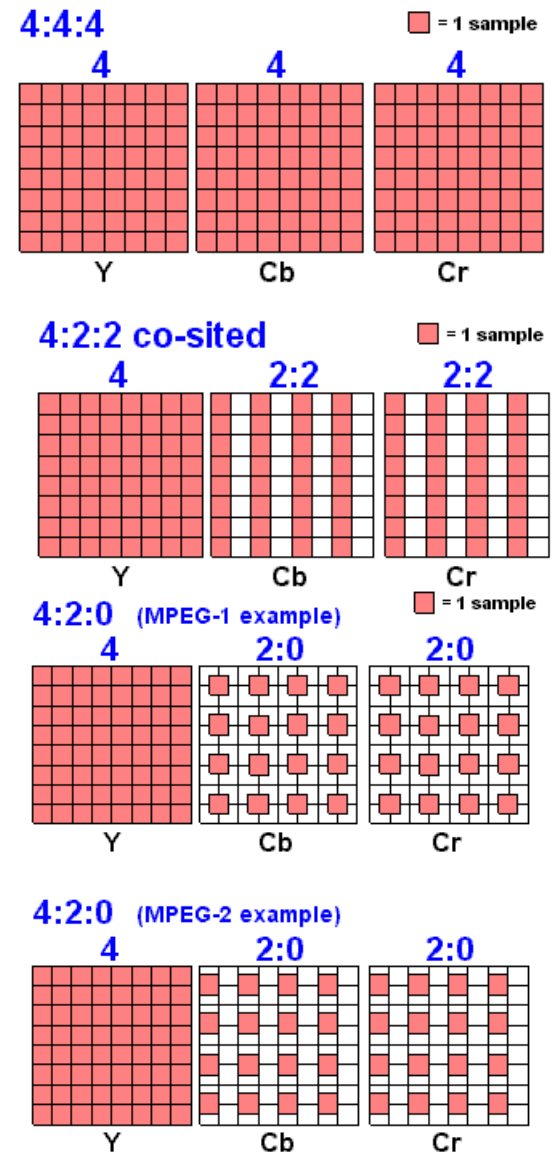


- ★ **Sample** - A sample refers to a component value at a point in time and/or space. A sampler is a subsystem or operation that extracts samples from a continuous signal. In video, there are R,G,B or luminance and chrominance samples, most of the times not with the same density/size.

- ★ **Pixel** - A pixel is generally thought of as the smallest relevant element of a digital image including all components. The more pixels are used to represent an image, the closer the final result can resemble the original. The number of pixels in an image is sometimes called the *spatial resolution*.
 - If all the image components have the same resolution, the number of pixels in the image is the number of samples of each component.
 - However, if the various components have different resolutions, than the number of pixels corresponds to the number of samples of the component with the highest resolution, typically the luminance.

Colour Subsampling Solutions

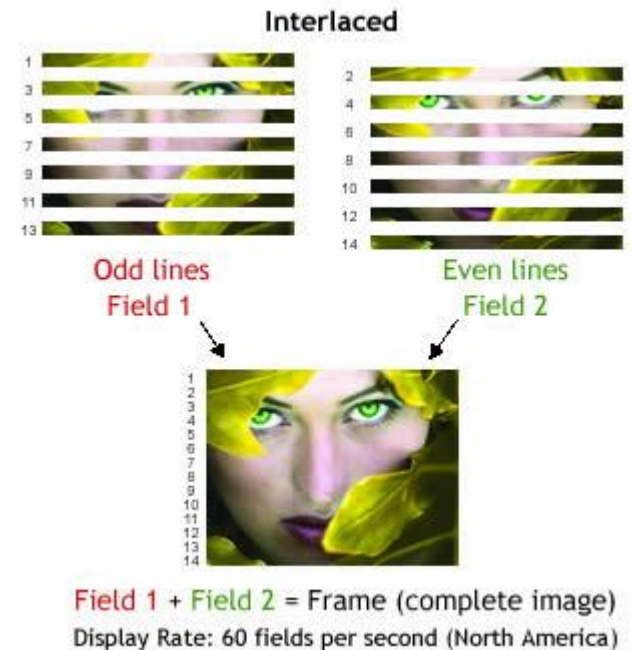
- ★ **4:4:4** – Luminance and each chrominance with the same number of samples; **targets very high quality, professional applications, studios, etc.**
- ★ **4:2:2** – Luminance with twice the samples of each chrominance (chrominances with same number of lines but half the samples per line); **still targets rather high quality applications such as HDTV.**
- ★ **4:2:0** – Luminance with 4 times the samples of each chrominance (chrominances with half the number of lines and half the samples per line); **targets medium and lower quality applications, notably digital TV, mobile video streaming, and Internet video.**



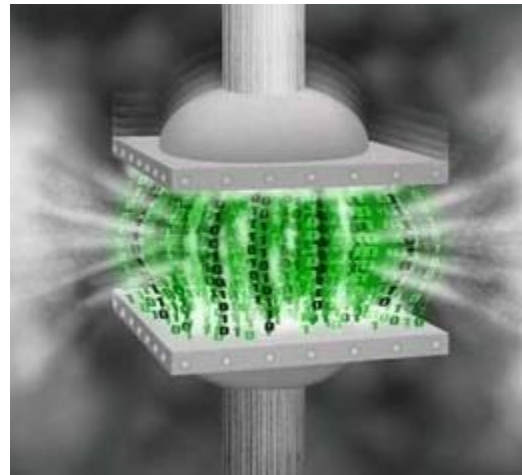
Progressive versus Interlaced Formats

- ★ **Progressive format** - Progressive scan differs from interlaced scan in that the image is displayed on a screen by scanning each line (or row of pixels) in a sequential order rather than an alternate order, as done with interlaced scanning.
- ★ **Interlaced format** - Interlacing divides the lines in a single frame into odd and even lines and then alternately refreshes them at 25/30 frames per second, leading to the so-called *odd an even fields*.

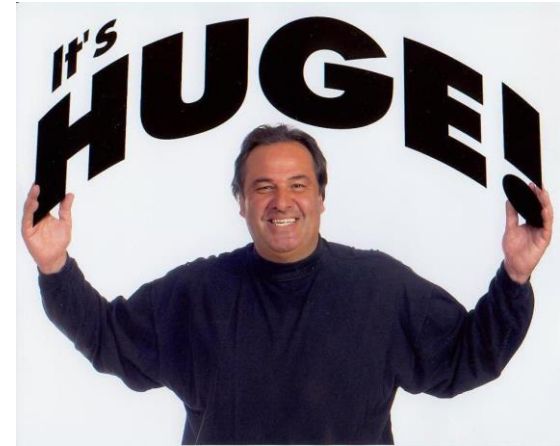
In other words, in progressive scan, the image lines (or pixel rows) are scanned in ‘regular’ numerical order (1,2,3) down the screen from top to bottom, instead of in an alternate order (lines or rows 1,3,5, etc... followed by lines or rows 2,4,6).



Digital Compression

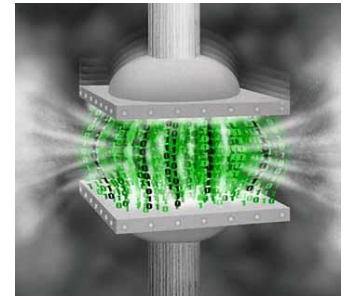


Why Compressing ?



- ★ **Speech** – e.g. 2×4000 samples/s with 8 bit/sample – **64000 bit/s = 64 kbit/s**
- ★ **Music** – e.g. 2×22000 samples/s with 16 bit/sample – **704000 bit/s = 704 kbit/s**
- ★ **Standard Video** – e.g. $(576 \times 720 + 2 \times 576 \times 360) \times 25$ (20736000) samples/s with 8 bit/sample – **166000000 bit/s = 166 Mbit/s**
- ★ **Full HD 1080p** - $(1080 \times 1920 + 2 \times 1080 \times 960) \times 25$ (103680000) samples/s with 8 bit/sample – **829440000 bit/s = 830 Mbit/s**

How Much is Enough ?



- ★ **Recommendation ITU-R 601: 25 images/s with 720×576 luminance samples and 360×576 samples for each chrominance with 8 bit/sample**

$$[(720 \times 576) + 2 \times (360 \times 576)] \times 8 \times 25 = 166 \text{ Mbit/s}$$

- ★ **Acceptable rate, p.e. using H.264/AVC: 2 Mbit/s**

=> Compression Factor: $166/2 \approx 80$

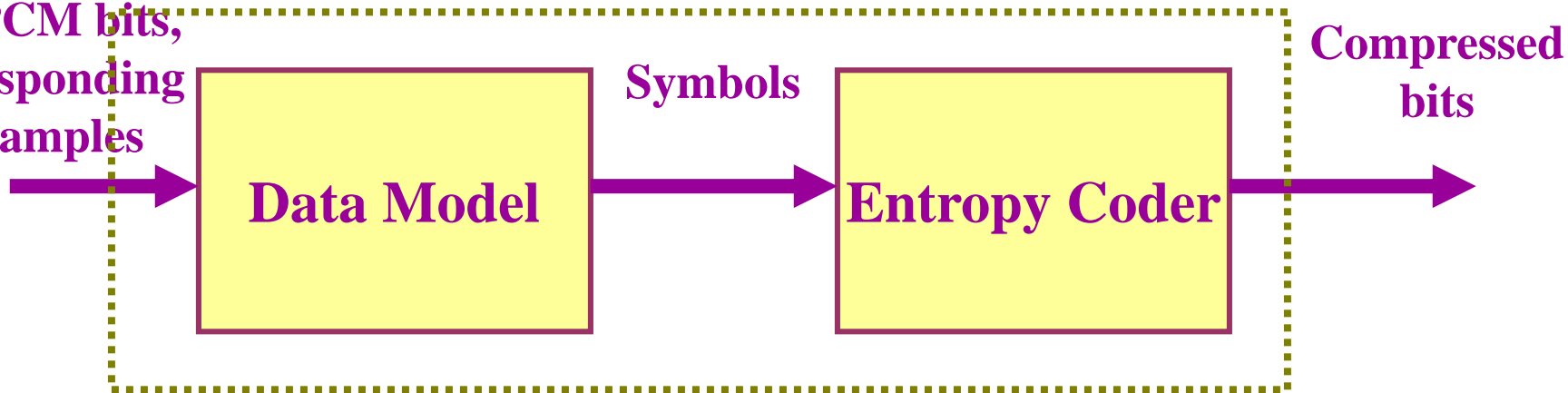
The difference between the resources requested by compressed and non-compressed formats may lead to the emergence or not of new industries, e.g., DVD, digital TV.

Source Coding: Original Data, Symbols and Bits



Original data,
e.g. PCM bits,
corresponding
to samples

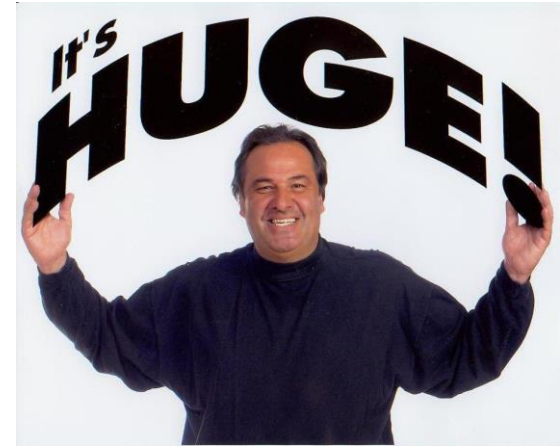
Encoder



Source Coding implies two main steps:

- ★ **Data modeling** – Adopting a more powerful data representation model than the raw acquisition model, notably exploiting spatial and temporal redundancies as well as irrelevancy, targeting the relevant representation requirements
- ★ **Entropy coding** - Exploiting the statistical characteristics of the symbols produced by the data modeling process

Why Compressing ?

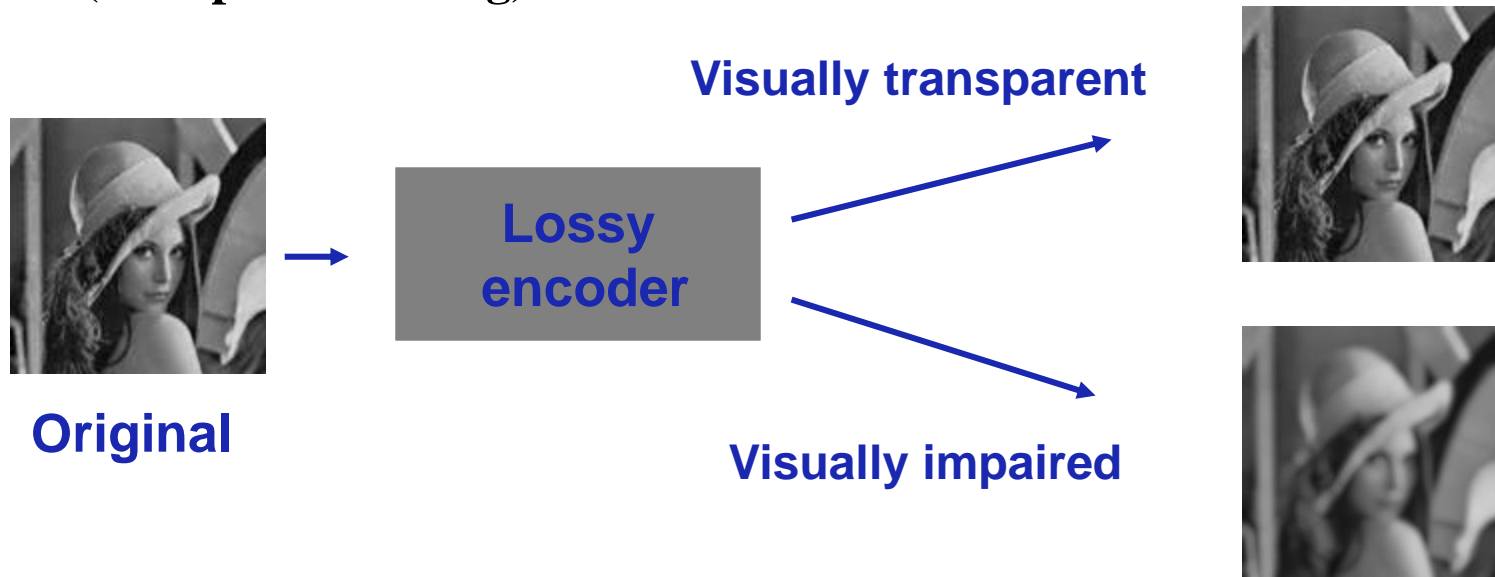


- ★ **Speech** – e.g. 2×4000 samples/s with 8 bit/sample – **64000 bit/s = 64 kbit/s**
- ★ **Music** – e.g. 2×22000 samples/s with 16 bit/sample – **704000 bit/s = 704 kbit/s**
- ★ **Standard Video** – e.g. $(576 \times 720 + 2 \times 576 \times 360) \times 25$ (20736000) samples/s with 8 bit/sample – **166000000 bit/s = 166 Mbit/s**
- ★ **Full HD 1080p** - $(1080 \times 1920 + 2 \times 1080 \times 960) \times 25$ (103680000) samples/s with 8 bit/sample – **829440000 bit/s = 830 Mbit/s**



Digital Coding: Main Types

- ★ **LOSSLESS (exact) CODING** – The content is coded preserving all the information present; this means the original and decoded contents are mathematically the same.
- ★ **LOSSY CODING** – The content is coded without preserving all the information present; this means the original and decoded contents are mathematically different although they may still look/sound subjectively the same (transparent coding).



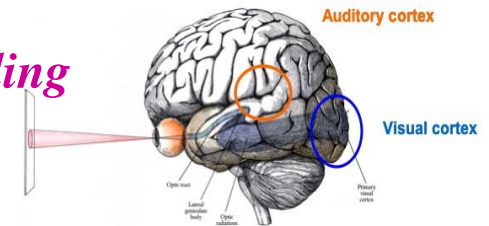
Where does Compression come from ?

★ **REDUNDANCY** – Regards the similarities, correlation and predictability of samples and symbols corresponding to the image/audio/video data.

-> **redundancy reduction does not involve any information loss this means it is a reversible process -> *lossless coding***

★ **IRRELEVANCY** – Regards the part of the information which is imperceptible for the visual or auditory human systems.

-> **irrelevancy reduction is an irreversible process -> *lossy coding***



Source coding exploits these two concepts: for that, it is necessary to know the source statistics and the human visual/auditory systems characteristics.

The Importance of (Open) Standards

- ★ **Media technologies, notably representation technologies, are used in many audiovisual applications for which interoperability is a major requirement.**
- ★ **The interoperability requirement is solved by specifying standards.**
- ★ **To allow evolution and competition, standards shall provide interoperability by specifying the minimum possible set of elements, for example the bitstream syntax and the decoder (*not the encoder*) for a coding format.**

Standards are also repositories of the best technology and thus an excellent place to check technology evolution and trends !

Standards are Good for Users ! And for Many Companies ...

The Impact of Interoperability ...

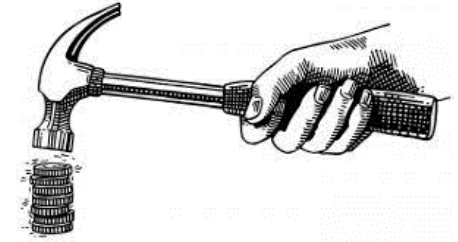


Performance Assessment



Low Rate and High Quality: Too Many Dreams ...





$$\text{Compression Factor} = \frac{\text{Number of bits for the original PCM data}}{\text{Number of bits for the coded data}} \gg 1$$

(for any media)

$$\text{Bit/pixel} = \frac{\text{Number of bits for the coded image}}{\text{Number of pixels (typically Y samples)}}$$

(for images)

The number of pixels in an image corresponds to the number of samples of its component with the highest resolution, typically the luminance.

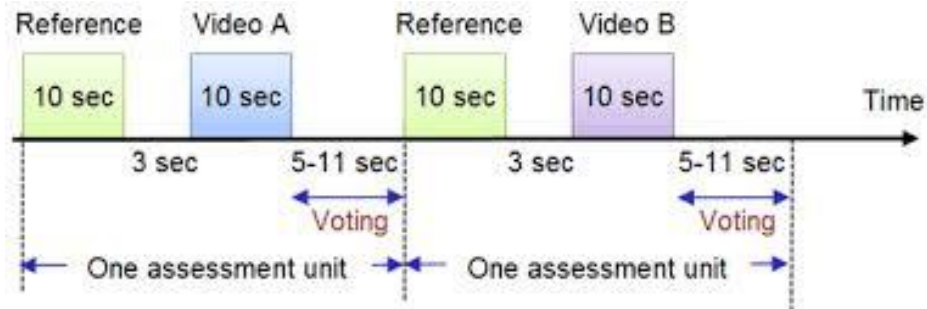


Subjective Quality Assessment

- ★ **Subjective video quality is a subjective characteristic of video quality concerned with how video is perceived by a viewer and designates his/her opinion on a particular video sequence.**
- ★ **Subjective video quality tests are quite expensive in terms of time (preparation and running) and human resources.**
- ★ **There are many of ways of showing video/audio sequences to experts and to record their opinions. A few of them have been standardized, e.g. in ITU-R BT.500 :**
 - **Degradation Category Rating (DCR) or Double Stimulus Impairment Scale (DSIS)** - the viewer sees an unimpaired reference video, then the same video impaired, and after that they are asked to vote on the second video using a so-called impairment scale.
 - **Pair Comparison (PC)** - instead of comparing an unimpaired and impaired sequence, different impairment types are compared with the viewer stating which is better (no unimpaired reference is available).
 - **Single Stimulus Absolute Category Rating (ACR)** - each impaired video is rated individually using an absolute scale, e.g. "bad", "poor", "fair", "good", and "excellent".

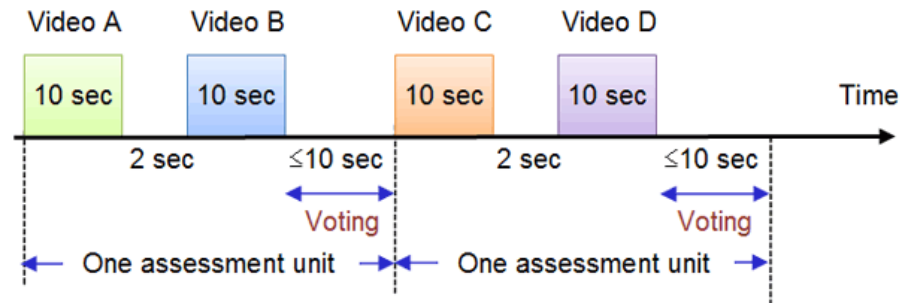


Subjective Quality Assessment



Double Stimulus Impairment Scale (DSIS)

Pair Comparison (PC)



Objective Quality Assessment

Objective video evaluation techniques are mathematical models that approximate results of subjective quality assessment, but are based on criteria and metrics that can be measured objectively and automatically evaluated by a computer program.

- ★ **Full Reference Methods (FR)** – compare the processed/decoded to the original videos/audios (*require original content !*)

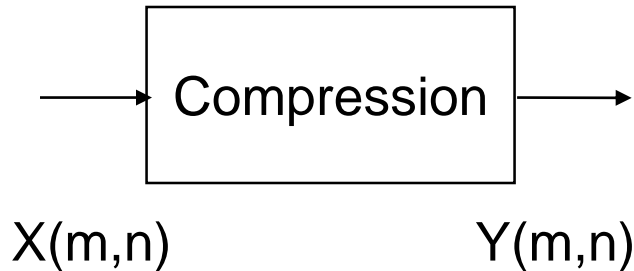


- ★ **Reduced Reference Methods (RR)** - extract and compare some features from the distorted/decoded videos/audios to derive a quality score (*require original features !*)
- ★ **No-Reference Methods (NR)** - assess the quality of a distorted/decoded video/audio without any reference to the original video (*no originals are needed*).

Objective Quality Metrics

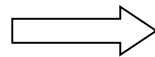


Original/reference



Decoded

Objective evaluation



x and y are the original and decoded data

$$PSNR(dB) = 10 \log_{10} \frac{255^2}{MSE}$$

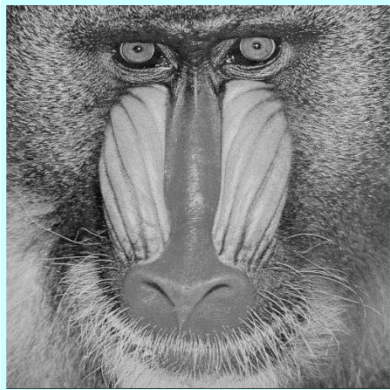
$$MSE = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - x_{ij})^2$$

There are other objective quality metrics !

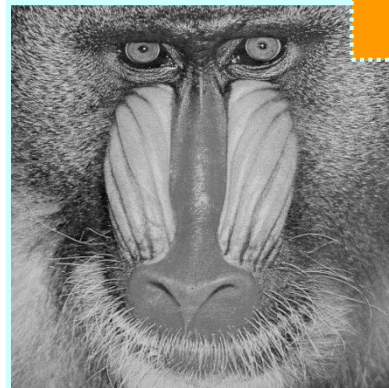
How Does PSNR Fail ...

$$\text{PSNR(dB)} = 10 \log_{10} \frac{255^2}{\text{MSE}}$$

$$\text{MSE} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (y_{ij} - x_{ij})^2$$



Original

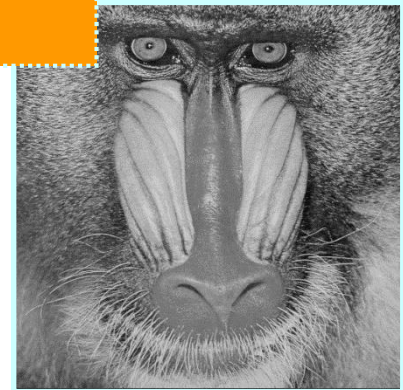


Degraded

PSNR: 50.98 dB

Subjective quality: High

Horizontally mirrored!

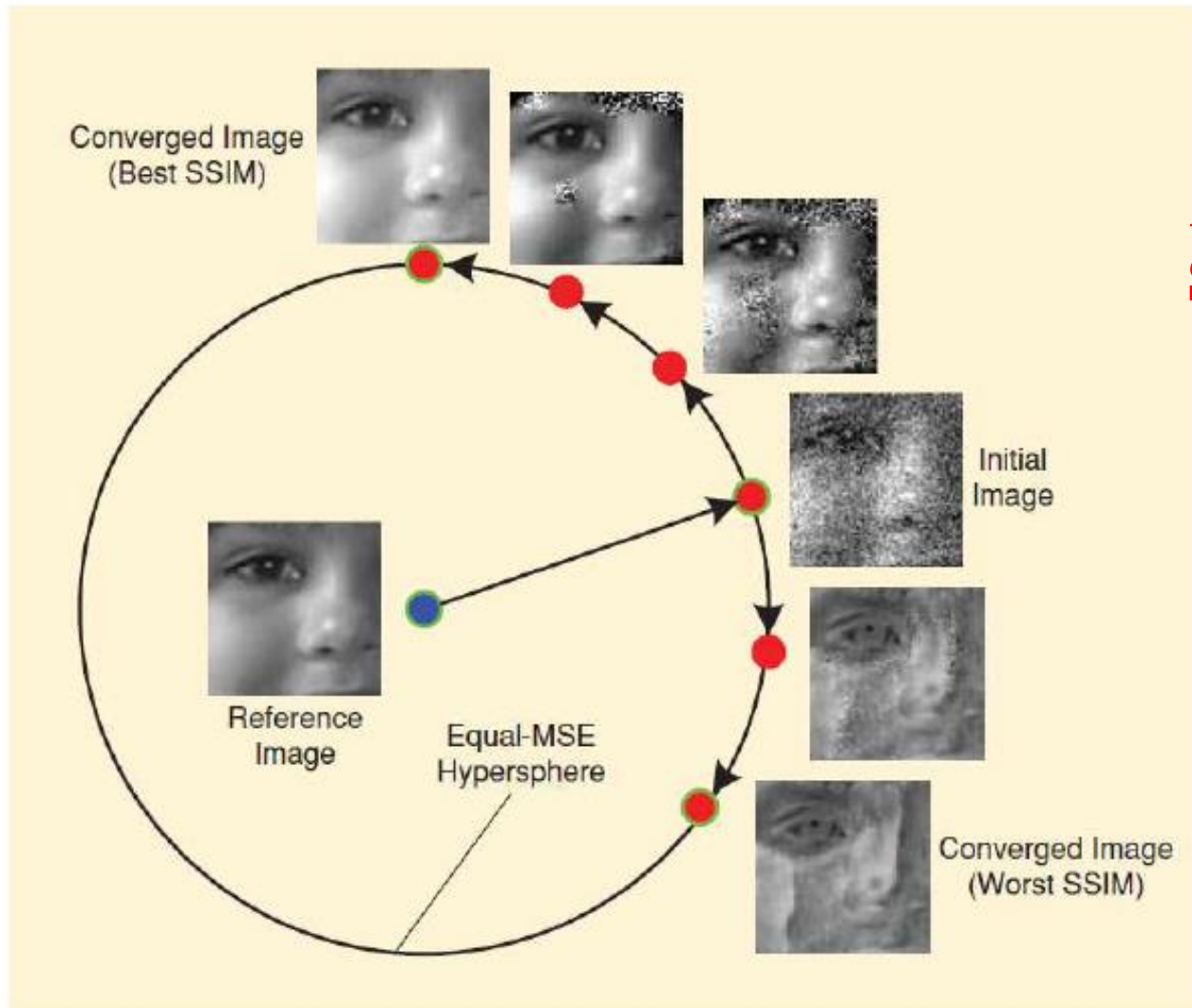


Degraded

PSNR: 14.59 dB

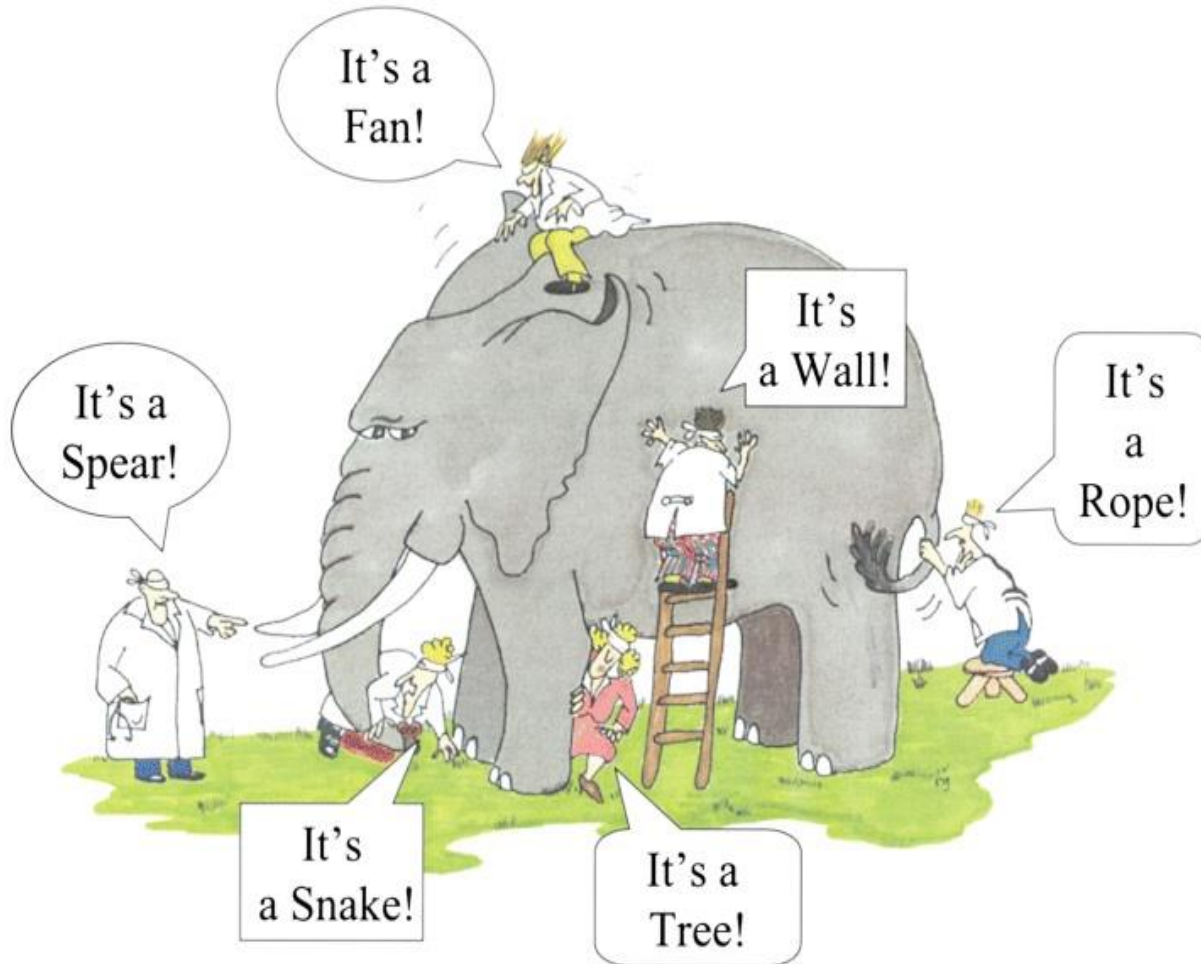
Subjective quality: High ?

MSE: Getting Tricked by MSE ...



ATTENTION:
Same error for
very different
subjective
qualities !

Quality is like an Elephant ...



The blind men and the elephant: Poem by John Godfrey Saxe

Quality of Service versus Quality of Experience

- ★ **Quality of Service** - Value of the average user's service richness estimated by a service/product/content provider, e.g. bandwidth, error rate, delay
- ★ **Quality of Experience** - Value (estimated or actually measured) of a specific user's experience richness

Quality of Experience is the dual (and extended) view of Quality of Service

QoS=provider-centric

QoE=user-centric



Metadata: Data about the Data



Seeing is Believing ! But ...

Although replication for visualization/auralization is a major target, there are other tasks where the visual representation does not need, or even should not be, made at pixel level:

- ★ **Searching**
- ★ **Filtering**
- ★ **Understanding**
- ★ **Control**
- ★ **...**



In fact, automatic processing tasks do not typically need a pixel-based representation as relevant information is limited ...

Content, Content, and More Content ...

How to Get what is Needed ?

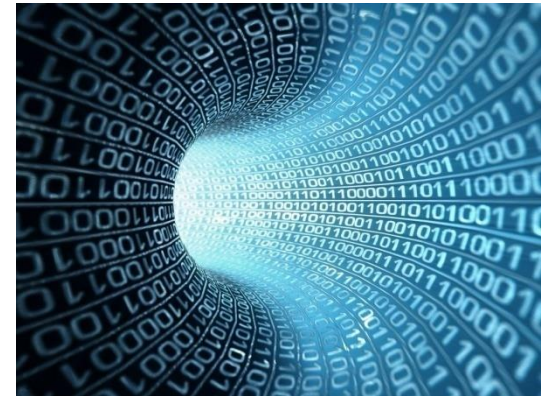


- ★ **Increasing availability of multimedia information**
- ★ **Difficult to find, select, filter, manage AV content**
- ★ **Because the value of content depends on how easy it is to find, select, manage and use it !**
- ★ **More and more situations where it is necessary to have ‘information about the content’**

Metadata: Data about the Data

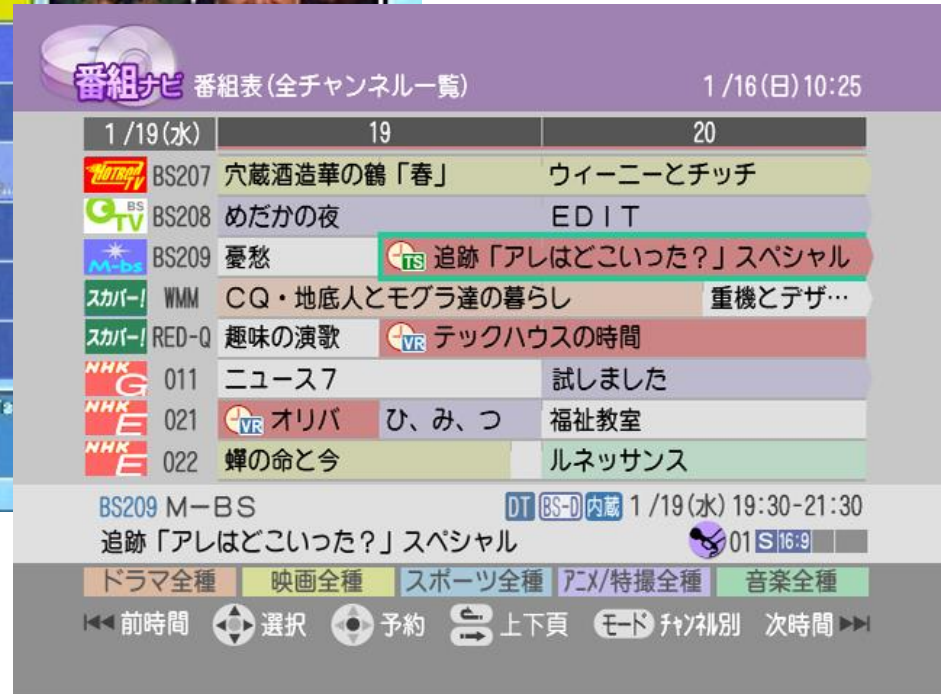
- ★ **Content description or metadata regards all types of data features which may be relevant for a more efficient searching, filtering, adaptation, management and, in general, consumption of data.**

- ★ **Metadata or "data about the data" may:**
 - *Describe the data/content itself, e.g. genre*
 - *Describe the data/content coding format, coded quality, etc.*
 - *Describe conditions about the data/content, e.g. licensing*
 - ...

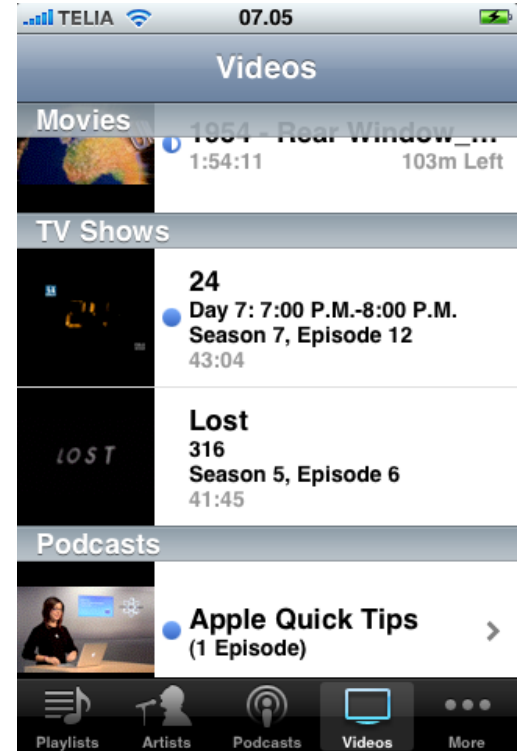
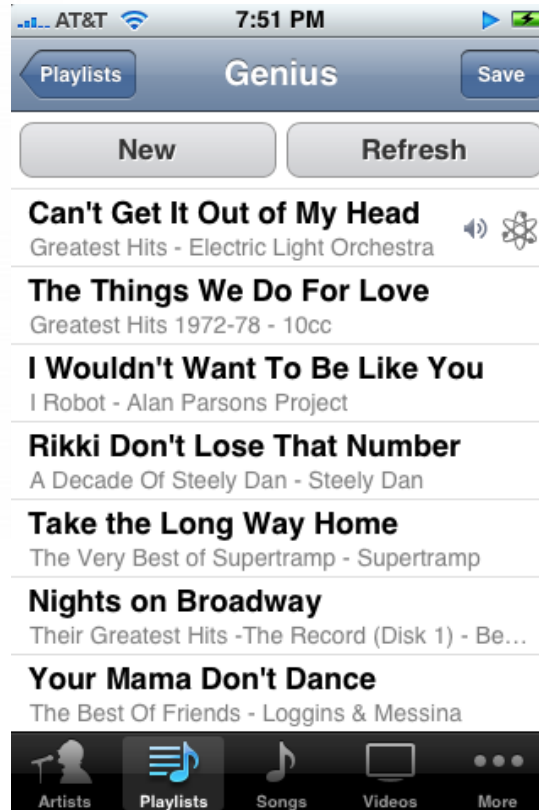


The more it is known about the data (metadata), the better the data can be processed, filtered, segmented, coded, adapted, ...

Filtering TV ...



Managing iPods Data ...



YouTube: Metadata, Searching ...

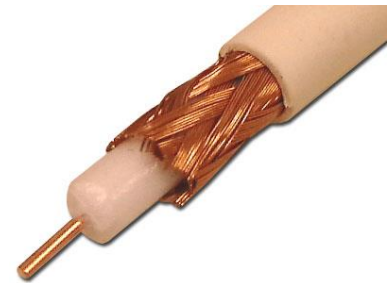
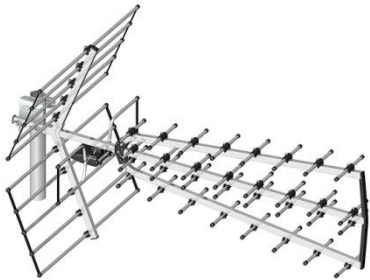
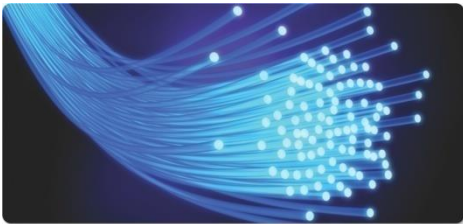


YouTube considers metadata fields such as

- ★ **Title**
- ★ **Description**
- ★ **Category**
 - *Autos & Vehicles, Comedy, Education, Entertainment, Film & Animation, Gaming, Howto & Style, Music, News & Politics, People & Blogs, Pets & Animals, Science & Technology, Sports, Travel & Events, ...*
- ★ **Date of upload**
- ★ **Number of views**
- ★ **Scores**
- ★ ...



And, finally, Transmission ...





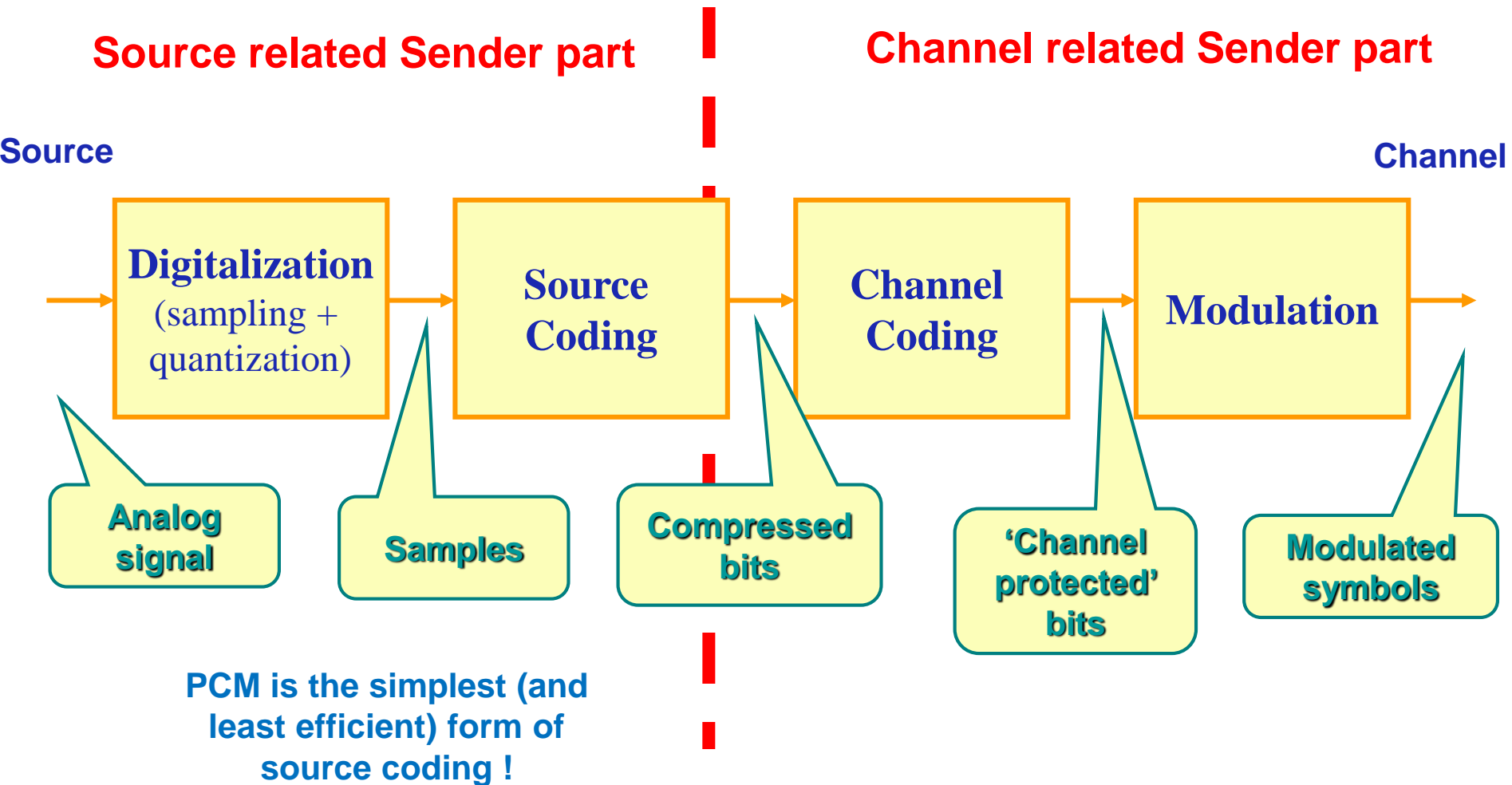


Channel Types



- ★ Data transmission, digital transmission, or digital communications is the physical transfer of data (a digital bit stream) over a point-to-point or point-to-multipoint communication channel.
- ★ There are so-called ‘**guided**’ channels and ‘**atmospheric**’ channels depending if some form of cable or the atmosphere are used for the transmission. Examples of such channels are copper wires, optical fibres, wireless communication channels, and storage media.
- ★ The data are represented as an electromagnetic signal, such as an electrical voltage, radiowave, microwave, or infrared signal.
- ★ **While analog transmission is the transfer of a continuously varying analog signal, digital communications is the transfer of discrete messages.**

Typical Digital Sender Chain ...

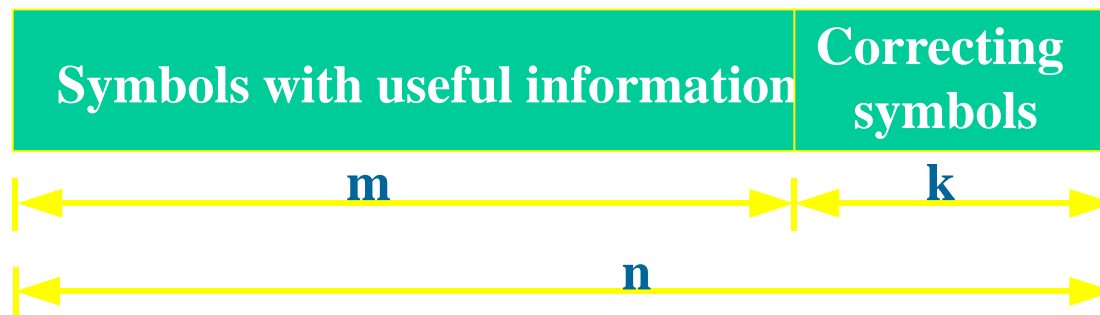


Channel Coding or the *Bodyguards* ...



Channel coding is the process applied to the bits produced by the source encoder to increase its robustness against channel or storage errors.

- ★ **At the sender, redundancy is added to the source compressed signal in order to allow the channel decoder to detect and correct channel errors.**
- ★ **The introduction of redundancy results in an increase of the amount of data (bits) to transmit. The selection of the channel coding solution must consider the type of channel, and thus the error characteristics, and the modulation.**



Block Codes

$$R = m/n = 1 - k/n$$



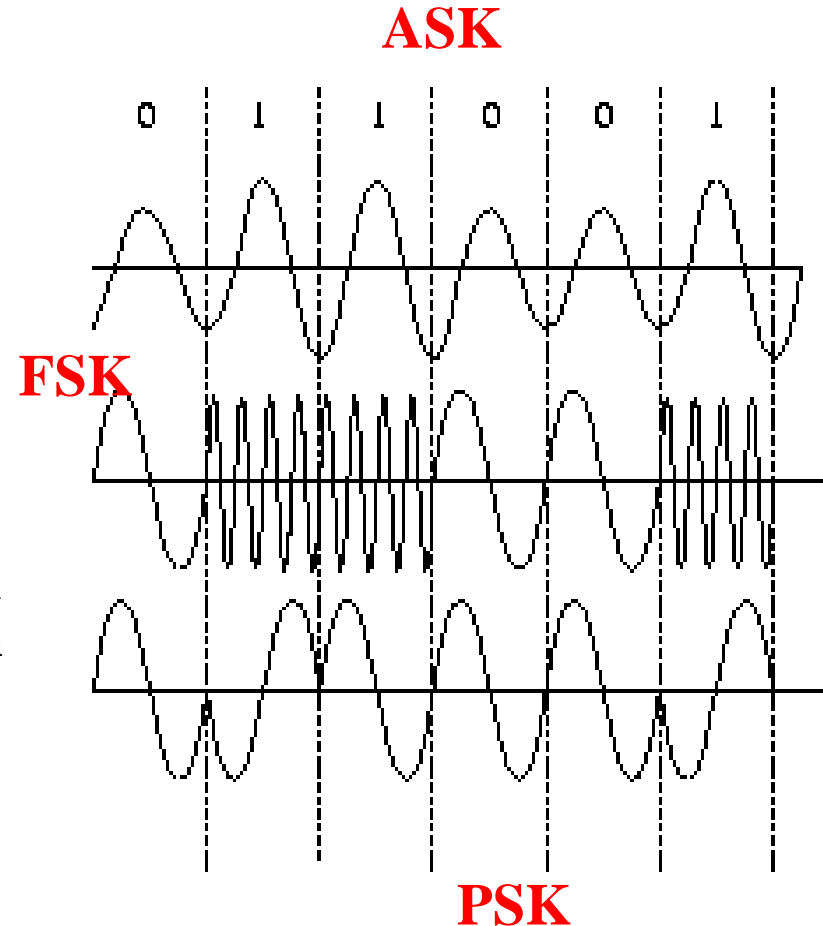
**Matching the
signal to the
channel !**

Modulation is the process through which one or more properties of a carrier (amplitude, frequency or phase) vary as a function of the modulating signal (the signal to be transmitted). If digital, the modulation is performed in a discrete way with a sequence of symbols.

Any of these properties can be modified in accordance with a baseband signal to obtain the modulated signal.

The selection of an adequate modulation is essential for the efficient usage of the available bandwidth and for the quality of the communication.

Together, (source and channel) coding and modulation determine the bandwidth necessary for the transmission of a certain signal.



Selecting a Modulation ... or Matching the Signal to the Channel ...

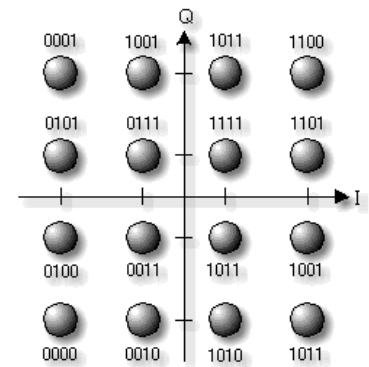
★ Factors to consider in selecting a modulation:

- *Channel characteristics*
- *Spectrum efficiency*
- *Resilience to channel distortions*
- *Resilience to transmitter and receiver imperfections*
- *Minimization of protection requirements against interferences*

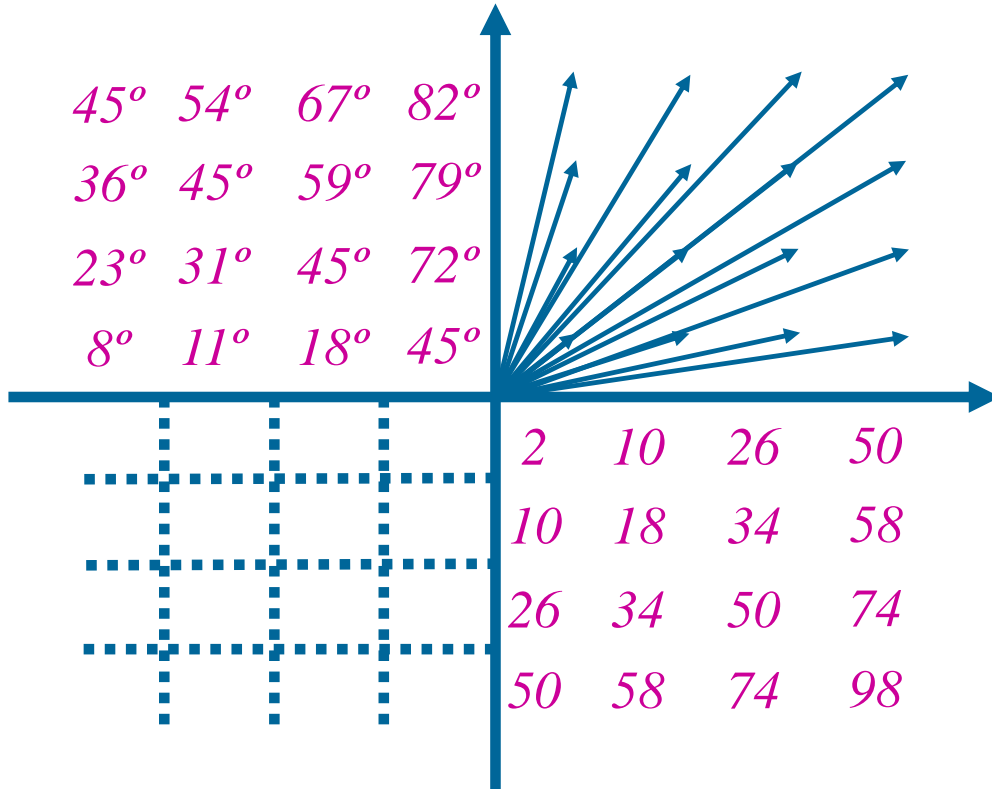


★ Basic digital modulation techniques:

- *Amplitude modulation (ASK)*
- *Frequency modulation (FSK)*
- *Phase modulation (PSK)*
- *Mix of phase and amplitude modulation (QAM)*



64-QAM Modulation Constellation



For 64-QAM, only 64 modulated symbols are possible !

More symbols would make the system more bandwidth efficient but also less error resilient

...

Trade-offs are critical !

Wireline head-end network equipment outputs a high-order modulated (e.g. 256QAM greater) signal.

Broadband head-end router



Original High Order Modulated Signal



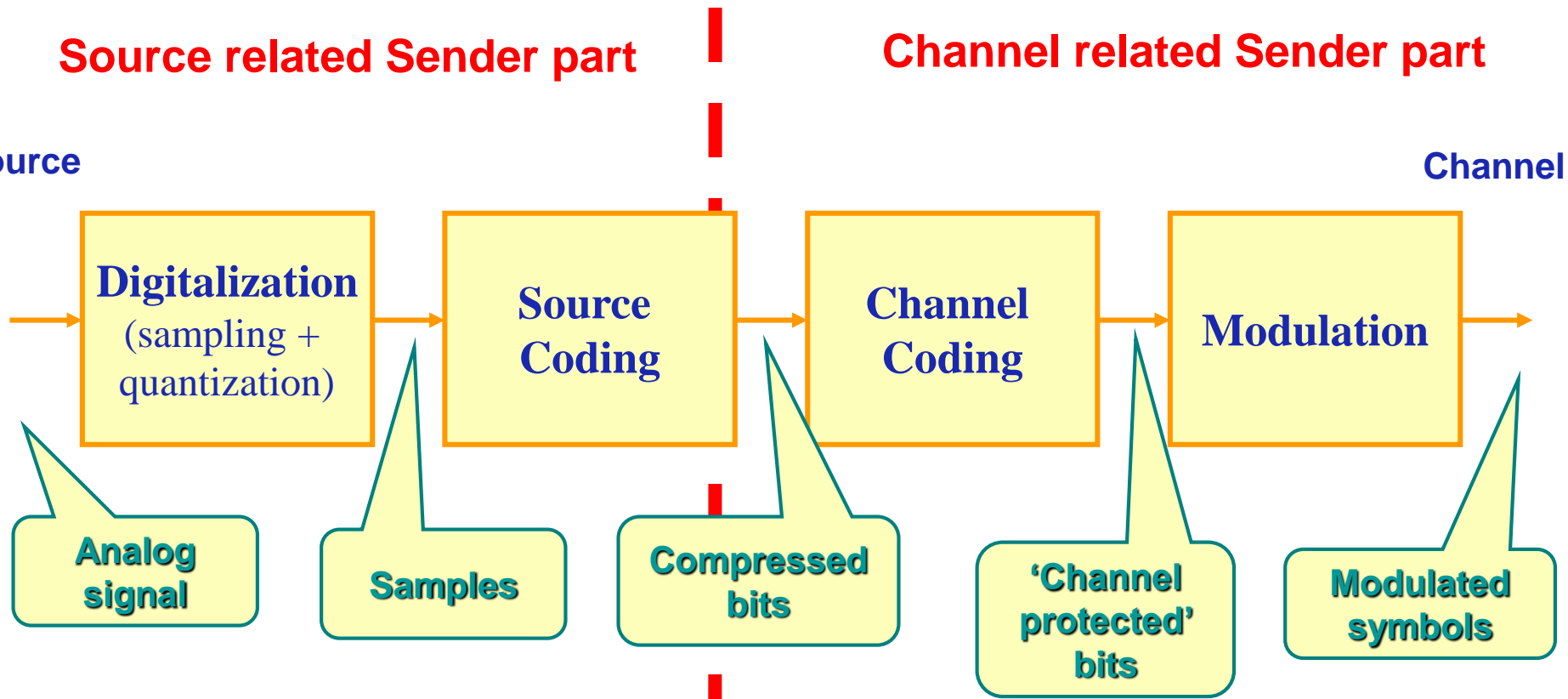
Typical Digital Sender Chain ...

Source related Sender part

Channel related Sender part

Source

Channel



PCM is the simplest (and least efficient) form of source coding !

Digital TV: a Full Example



- ★ **ITU-R 601 Recommendation: 25 images/s with 720×576 luminance samples and 360×576 samples for each chrominance with 8 bit/sample**
$$[(720 \times 576) + 2 \times (360 \times 576)] \times 8 \times 25 = \mathbf{166 \text{ Mbit/s}}$$
- ★ **Acceptable rate after source coding/compression, p.e. using H.264/AVC:**
 $\mathbf{2 \text{ Mbit/s}}$
- ★ **Rate after 10% of channel coding 2 Mbit/s + 200 kbit/s = 2.2 Mbit/s**
- ★ **Bandwidth for video information in a digital TV channel, e.g. with 64-PSK or 64-QAM: 2.2 Mbit/s / $\log_2 64 \approx \mathbf{370 \text{ kHz}}$**
- ★ **Number of digital TV channels / analogue TV RF slot: 8 MHz / 400 kHz $\approx \mathbf{20 \text{ channels}}$**

Patents and Copyright





Patent: What is it ?



- ★ **Patents is a form of intellectual property intended to facilitate and encourage disclosure of innovations into the public domain for the common good.**
- ★ **Patenting can be viewed as contributing to open knowledge after an embargo period (usually of 20 years).**
- ★ **If inventors did not have the legal protection of patents, in many cases, they might prefer or tend to keep their inventions secret.**
- ★ **A patent is a set of exclusive rights granted by a sovereign state to an inventor or assignee for a limited period of time in exchange for detailed public disclosure of an invention.**
- ★ **The costs of preparing and filing a patent application, prosecuting it until grant, and maintaining the patent vary from one jurisdiction to another.**
- ★ **Patents can generally only be enforced through civil lawsuits ...**

Copyright: What is it ?



- ★ **Copyright is a form of intellectual property, applicable to certain forms of creative work.**
- ★ **Copyright is a legal right created by the law of a country that grants the creator of an original work exclusive rights for its use and distribution.**
- ★ **This is usually only for a limited time, e.g. 50 years.**
- ★ **Copyright is often shared among multiple authors, each of whom holds a set of rights to use or license the work, and who are commonly referred to as rights holders.**
- ★ **Copyrights are considered territorial rights, which means that they do not extend beyond the territory of a specific jurisdiction.**
- ★ **Typically, the duration of a copyright spans the author's life plus 50 to 100 years, depending on the jurisdiction.**

Licensing: What is it ?



- ★ **Licensing is the selling of intellectual property to a person or business that wishes to produce it for a profit.**
- ★ **The intellectual property could be a patent or copyright.**
- ★ **Reasonable and non-discriminatory terms (RAND) denote a voluntary licensing commitment that standards organizations often request from the owner of an intellectual property right (usually a patent) that is, or may become, essential to practice a technical standard.**
- ★ **A patent becomes standard-essential when a standard-setting organization sets a standard that adopts the technology that the patent covers.**
- ★ **Licensing is a very tricky issue in these days !**

Image and Video/Audio Coding: Different Licensing Traditions

- ★ **Standardization organizations cannot impose licensing conditions to the patent owners.**
- ★ **For sure, they CANNOT guarantee any royalty free conditions... they can just ‘wish’ ...**
- ★ **Traditionally, standard image codecs are not burdened by royalties.**
- ★ **Traditionally, standard video and audio codecs are strongly burdened by royalties.**
- ★ **Some big companies ‘offer’ royalty free video codecs ...**



Copyright Conditions Examples

Codec	Licensing Organization	Per-Device Royalties		Per-Title Royalties		Subscription-Based Royalties		Free/Public Over-the-Air Broadcast	Internet Broadcast	Per-Organization, Yearly Cap for all Royalties
		Royalty	Yearly Cap	Royalty	Yearly Cap	Royalty	Yearly Cap			
MPEG-2 Video	MPEG-LA	Originally, \$2.50/unit. Now, \$0.50/unit or less.	None	None	None	None	None	None	None	None
H.264/AVC)	MPEG-LA	\$0.10-\$0.20/unit depending on volume. No royalty for volumes less than 100,000 units.	2011-2015, \$6.5M. As of 2016, \$9.75M	Lesser of \$0.02 or 2% of title sale value	None	\$25,000-\$100,000 per year depending on number of subscribers. No royalty for organizations with less than 100,000 subscribers.	None	One-time fee of \$2500/encoder OR \$2500-\$10,000 per year based on number households in broadcast area	None	2011-2015, \$6.5M. As of 2016, \$9.75M
HEVC	MPEG-LA	\$0.20/unit. No royalty for volumes less than 100,000 units	\$25M	None	None	None	None	None	None	None
	HEVCAdvance	\$0.40-\$1.20/unit depending on volume. Additional \$0.10-\$0.75/unit for higher HEVC profiles.	\$20M-\$30M (Mobile) \$20M (Other Devices) \$20M (4K UHD+ TVs)	\$0.025/unit	\$2.5M	\$0.005 per subscriber, increasing to \$0.025 by 2020	\$2.5M	None	None	\$5M (Per-Title and Subscription-Based) \$40M (Devices)

Bibliography

- ★ **Comunicações Audiovisuais: Tecnologias, Normas e Aplicações”, chapter 5, edited by F.Pereira, IST Press, Julho 2009.**

- ★ **Fundamentals of Digital Image Processing, Anil K. Jain, Prentice Hall, 1989**

- ★ **Digital Video Processing, A. Murat Tekalp, Prentice Hall, 1995**