Digital Audio/Image/Video Representation

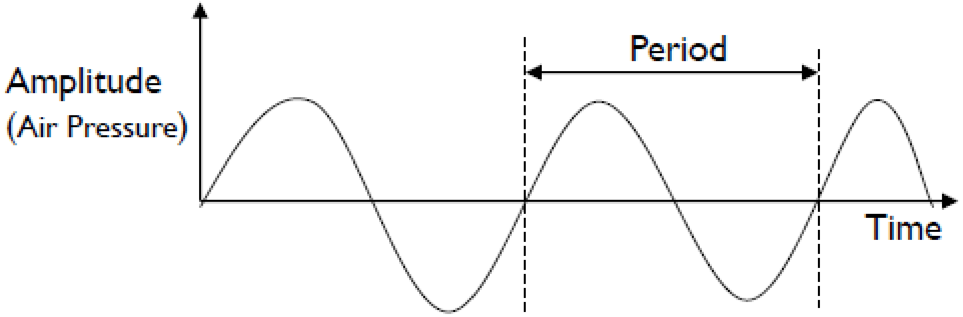
Prof. Dr. Uluğ Bayazıt

Outline

- Digital audio representation
  - Quantization
  - Sampling
- Digital Image Representation
  - Color System
  - Chrominance Subsampling
- Digital Video Representation
- Hardware Requirements

Digital Audio Representation

- Sound
  - due to vibration of matter (i.e., air molecules).
  - continuous wave that travels through air.
    - *Amplitude :* measure of the displacement of air pressure wave from its mean or quiescent state (in decibels (dB))
      - Peak amplitude, peak-to-peak amplitude
    - *Period* is the length of one full cycle
    - *Frequency* represents the number of periods in a second (in hertz, Hz, cycles/second).
      - the reciprocal value of the period

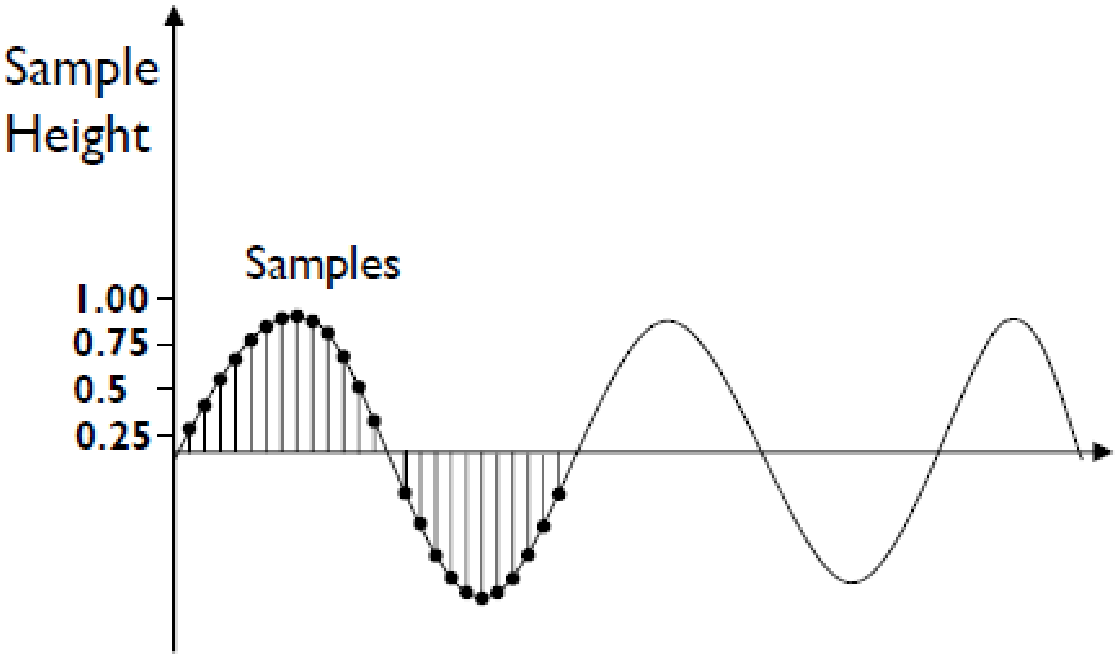


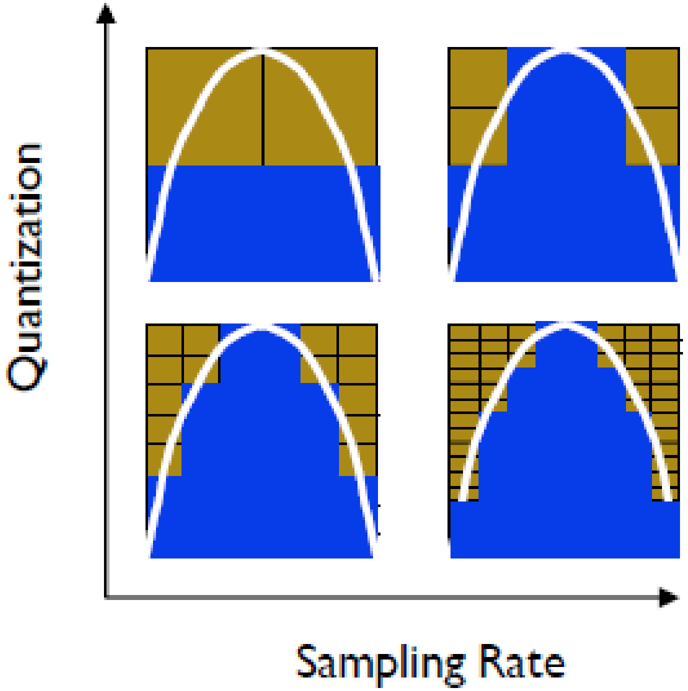
Digital audio representation

Processing steps:

- *Transducer* (inside a microphone) converts pressure to voltage levels.
- *A/D converter* converts analog (voltage, current) signal into a digital stream by discrete sampling.
  - Discretization in time
  - Discretization in amplitude (*quantization*).
- *In a computer,*
  - these values are sampled at intervals to yield a vector of values (samples).

Sampling
Quantization





Sampling Rate &Nyquist Theorem

- Direct relationship between sampling rate, sound quality (fidelity) and storage space.
- Q:How often do you need to sample a signal to avoid losing information?
  - Sampling rate is not playback rate!
- A: It depends on how fast the signal is changing. In reality, more than twice per cycle (a.k.a. *Nyquist sampling theorem*).
  - If a signal *f*(*t*) is sampled at regular intervals of time and at a rate higher than twice the highest significant signal frequency, then the samples contain all the information of the original signal.

- Human hearing
  - Perceptible frequency (audio) range: 20Hz – 20kHz (voice is between 500Hz-2KHz).
  - Discard frequencies above 20KHz (22.05 kHz for CD’s) by low pass filtering.
  - Sample at twice the maximum frequency (44.1kHz for CD’s)

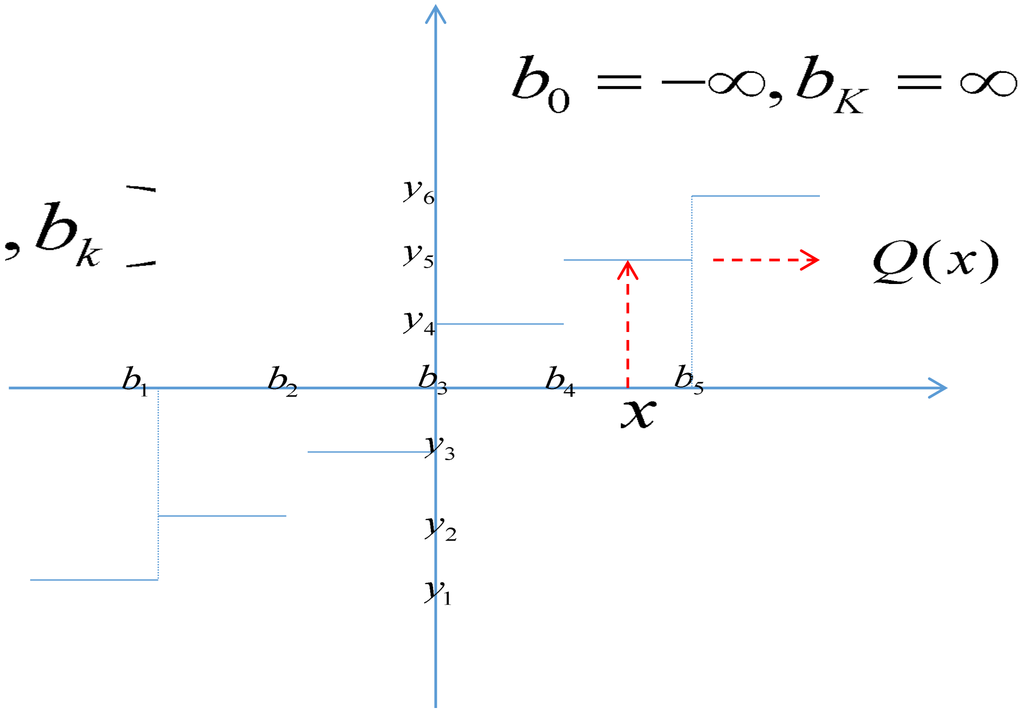
Quantization

- Sample precision - the resolution of a sample value.
- Quantization is an approximation (rounding)
  - Approximation quality depends on the number of bits used to represent the height of the waveform.
- **Sony CD 16 bits**, Philips D/A converter 14 bits
- Audio formats are described by sample rate and quantization
  - Voice quality (Pulse code modulation)- 8 bits quantization, 8,000 Hz mono (64 Kbps)
  - CD quality - 16 bits quantization, 44,100 Hz linear stereo (705.6 Kbps for mono, 1.411 Mbps for stereo (left and right channels))

Scalar quantization

- $Y = Q(X)$ where $Q(.)$ is a staircase function
  - Decision boundaries b_k
  - Reconstruction (quant.) levels y_k

$$Q(x) = y_k \text{ if } x \in B_k = [b_{k-1}, b_k)$$



$$b_0 = -\infty, b_K = \infty$$

Uniform scalar quantization

- All granular bins are of same size

$$b_k - b_{k-1} = \Delta \quad \text{for} \quad k = 2,\ldots, K-1$$

- If $f_X(x) = 0 \;\; \text{for } x > x_{\max}, x < -x_{\max}$

$$\Delta = \frac{2x_{\max}}{K}$$

- Within each bin distribution is approx. uniform
  - Assume $f_X(x \mid x \in B_k) \approx \dfrac{1}{\varDelta}, \qquad (k\text{-}1) < x \le k\Delta$
  - Quantization levels $y_k = \dfrac{b_k + b_{k-1}}{2}$ are then optimal

MSE distortion approximation

- Uniform quantizer MSE distortion

$$D_{Q,MSE} = D_{G,MSE}(B_k) \cong \int_{-\Delta/2}^{\Delta/2} x^2 f_X(x \mid x \in B_k)dx$$

$$= \int_{-\Delta/2}^{\Delta/2} x^2 \frac{1}{\Delta} dx = \frac{\Delta^2}{12}$$

- Assume fixed length coding of quantization indices with n bits

$$K = 2^n, \Delta = 2\frac{x_{\max}}{2^n}$$

$$SNR(dB) = 10\log_{10}\left(\frac{\sigma_X^2}{D_{Q,MSE}}\right) = 10\log_{10}\left(\frac{\sigma_X^2.12}{\Delta^2}\right)$$

$$= 10\log_{10}\left(\frac{\sigma_X^2.12.2^{2n}}{4x_{\max}^2}\right) = 10\log_{10}\left(\frac{\sigma_X^2.12}{4x_{\max}^2}\right) + 20\log_{10}2^n$$

$= C + 6.02ndB$ with every bit SNR increases by 6dB

Signal-to-Noise Ratio

- A measure of the quality of the signal. Let P_{signal} and P_{noise} be the signal power and noise power (variances), respectively
- SNR = 10 log10 (P_{signal} / P_{noise})
- Assuming quantization error is uniform, and the variance of signal is not too large compared to the maximum signal value x_{max}, then each bit adds about 6 dB of resolution!
  - Assume fixed length coding of quantization indices with n bits $K = 2^n, \Delta = 2\frac{x_{\max}}{2^n}$

$$SNR(dB) = 10\log_{10}\left(\frac{\sigma_X^2}{D_{Q,MSE}}\right) = 10\log_{10}\left(\frac{\sigma_X^2.12}{\Delta^2}\right)$$

$$= 10\log_{10}\left(\frac{\sigma_X^2.12.2^{2n}}{4x_{\max}^2}\right) = 10\log_{10}\left(\frac{\sigma_X^2.12}{4x_{\max}^2}\right) + 20\log_{10} 2^n$$

$$= C + 6.02ndB$$

Pulse Code Modulation (PCM)

- The two step process of sampling and quantization is known as *Pulse Code Modulation.*
    - Based on the Nyquist sampling theorem.
    - Used in speech and CD encoding.

Representation of Audio Samples

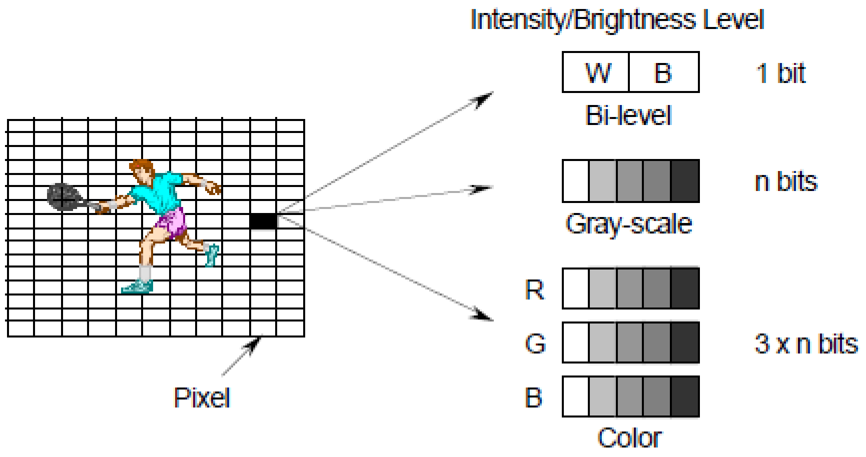
- Audio samples are represented as formats characterized by four parameters:
  - *Sample rate:* Sampling frequency
  - *Precision:* Number of bits used to store audio samples
  - *Encoding:* Audio data representation (compression)
  - *Channel:* Multiple channels of audio may be interleaved at sample boundaries.
- Raw speech data
  - PCM-encoded speech (64 Kbps)
  - Music (1.411 Mbps for stereo)
  - strains the bandwidth of cellular networks/Internet => compression is needed!

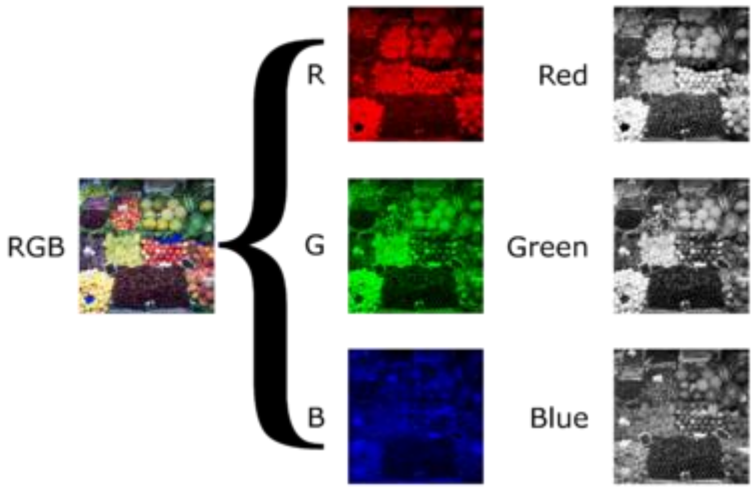
Audio compression basics

- Audio samples are encoded (compressed) based on
  - Non-uniform quantization - humans are more sensitive to changes in “quiet” sounds than “loud” sounds:
    - Companding (compress- uniform quantize – expand) | μ-law and A law companders
  - High correlation between adjacent samples
    - Difference encoding
  - Psychoacoustic Principles - humans do not hear all frequencies the same way due to *Auditory Masking:*
    - Simultaneous masking
    - Temporal masking
- These approaches are used in MPEG-1 Layer 3, *known as MP3.*
  - Reduces bit rate for CD quality music down to 128 or 112 Kbps.

Digital Image Representation

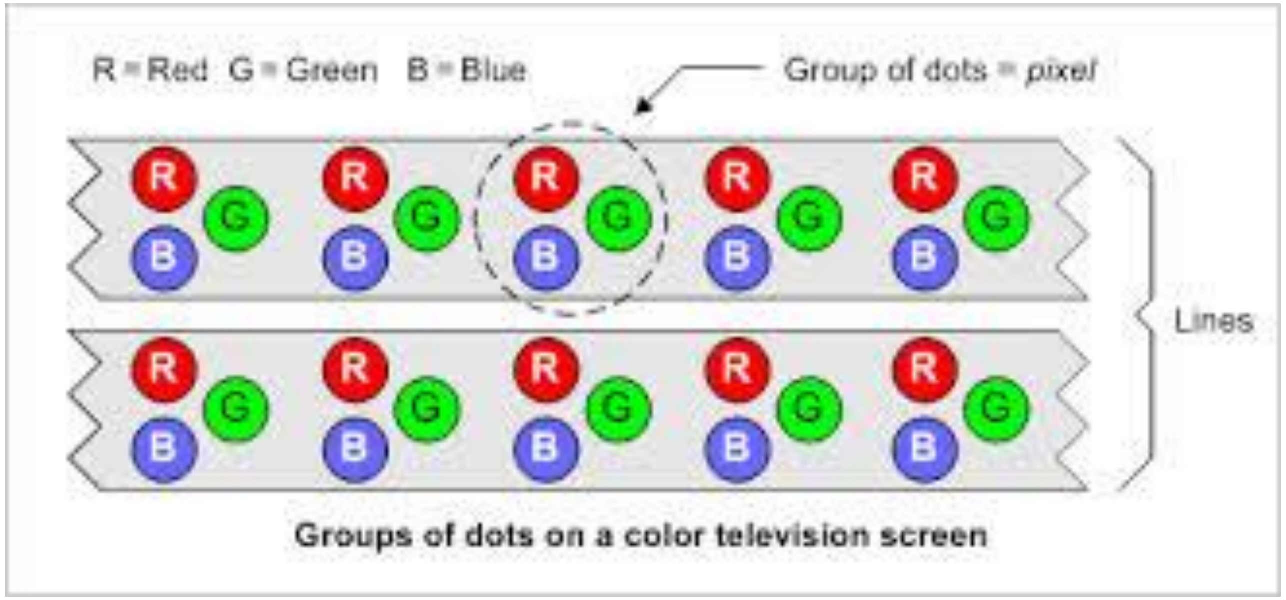
- An image is a collection of *picture elements or pixels* on a nxm grid.
  - Pixel representation can be bi-level, gray-scale, or color.
  - *Resolution specifies the distance between points – akin to sample rate.*





Pixels

- Images are made up of dots called **pixels for picture elements**
  - The number of pixels affects the resolution of the monitor
  - The higher the resolution, the better the image quality
    - at a given viewing distance

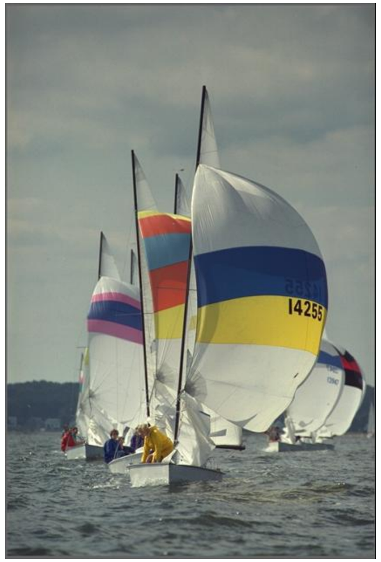


Groups of dots on a color television screen

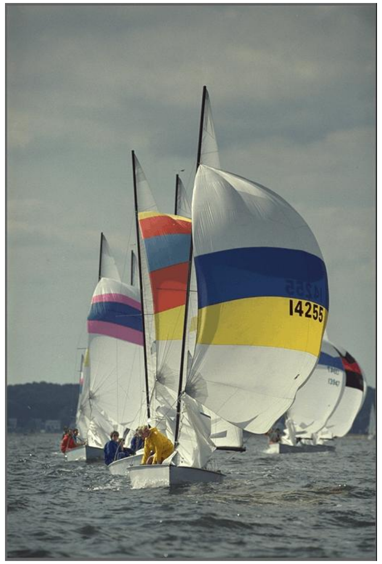
Color

- The amount of information per pixel is known as the *color depth*
    - Monochrome (1 bit per pixel)
    - Gray-scale (8 bits per pixel)
    - Color (8 /16 /18 bits per pixel)
        - 8-bit indexes to a color palette
        - 16 bits
            - 5 bits for each RGB + 1 bit Alpha (16 bits)
            - 4 bits for each RGB ,Alpha
            - 5 bits for each RB+ 6 bits for G
        - 18 bits
            - 6 bits for each RGB (cheap LCD displays)
    - True color (24 or 32 bits per pixel)
        - RGB (sRGB 24 bits) : 2^{24} color variation of which 2/3rd can be discriminated
        - RGB + Alpha (RGBA 32 bits)
    - Deep color (30/36/48 bits per pixel)
        - More recent (HEVC(H.265-High Efficiency Video Coding), HDMI 1.3)
        - More info than can be displayed all at once

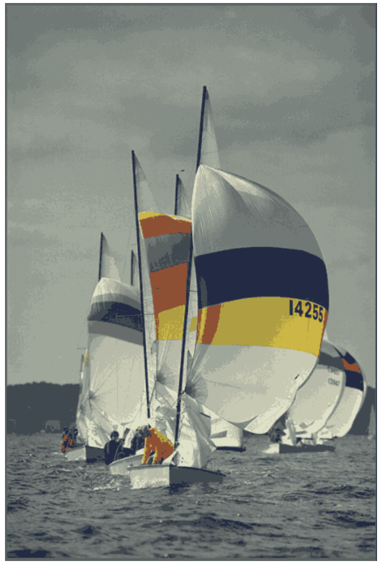
Example of color depth



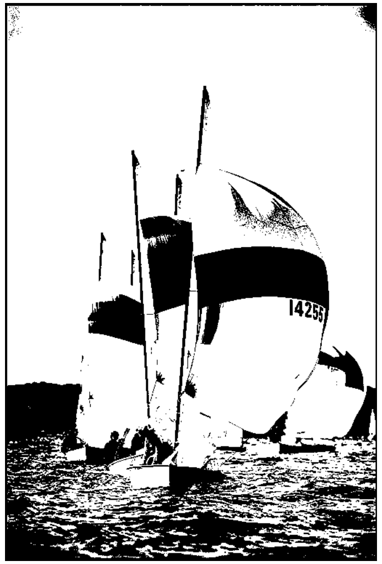
24 bits



8 bits

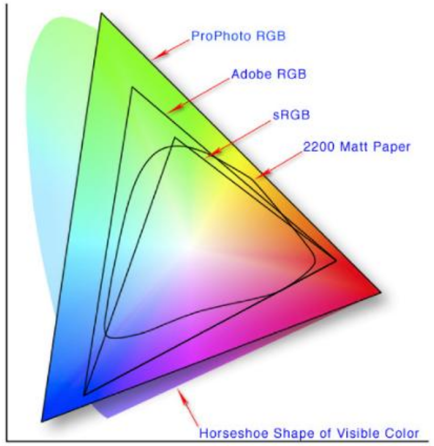


4 bits

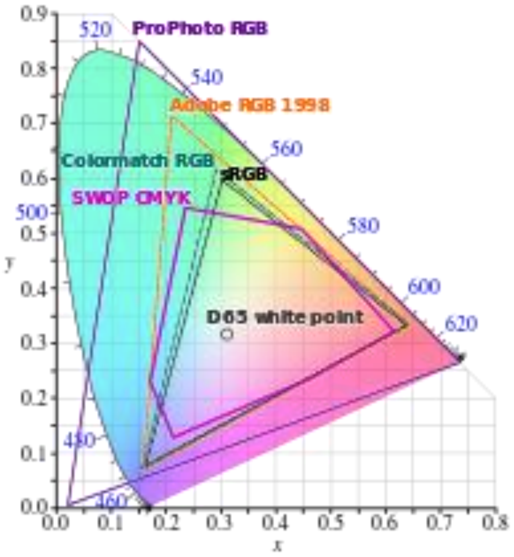


1 bit

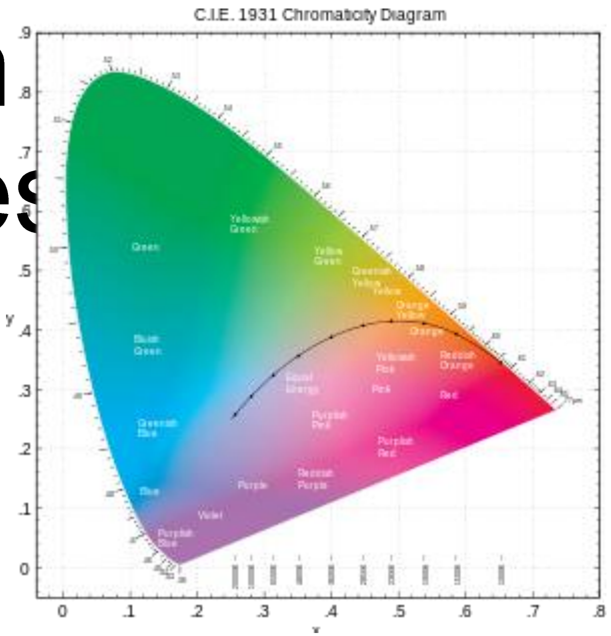
Color spaces



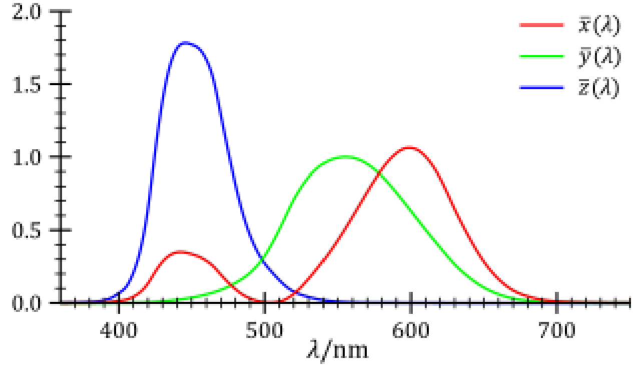
- A method by which we can specify, create, and visualize color.
  - Color model: specifies the abstract model describing the way colors can be represented as tuples of numbers (color components) – specifies coordinate system
  - Mapping function associates to a color space for color interpretation => a gamut
  - Gamut/Color space: abstract three dimensional region within which we can plot points that represent a (visible) color – recently carries the notion of color model as well
- Why more than one color space? Different color spaces are better for different applications.
  - Humans => Hue, Saturation, Lightness or Brightness (HSL or HSB)
  - CRT monitors => Red Green Blue (RGB)
  - Printers => Cyan Magenta Yellow Black (CMYK)
  - Compression => Luminance and Chrominance (YIQ, YUV, YCbCr)



CIE (International Commission on Illumination) XYZ and RGB spaces



- Color matching functions ($M(\lambda)$ spectral power distribution)
  - XYZ

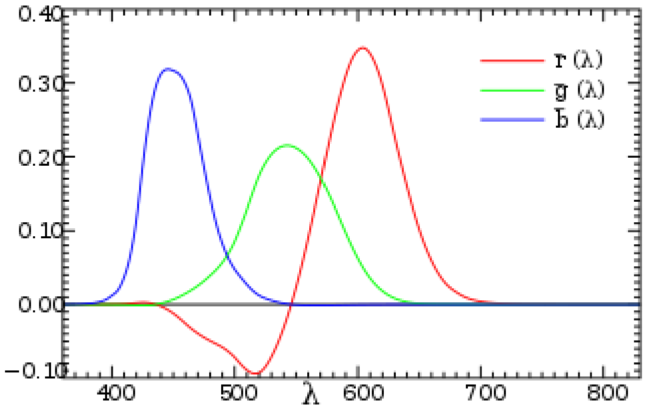


$$X = \int_{380}^{780} M(\lambda)\,\overline{x}(\lambda)\,d\lambda \qquad x = \frac{X}{X+Y+Z}$$

$$Y = \int_{380}^{780} M(\lambda)\,\overline{y}(\lambda)\,d\lambda \qquad y = \frac{Y}{X+Y+Z}$$

$$Z = \int_{380}^{780} M(\lambda)\,\overline{z}(\lambda)\,d\lambda \qquad z = \frac{Z}{X+Y+Z} = 1 - x - y$$

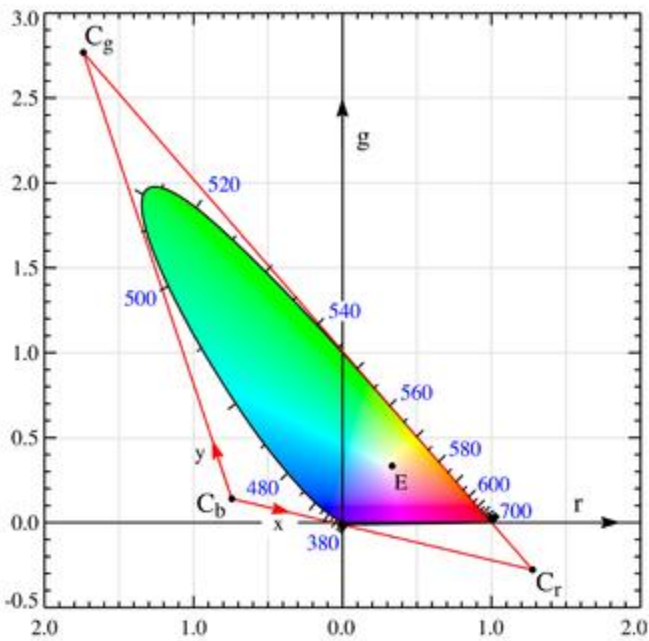
  - RGB



$$R = \int_0^\infty M(\lambda)\,\overline{r}(\lambda)\,d\lambda \qquad r = \frac{R}{R+G+B},$$

$$G = \int_0^\infty M(\lambda)\,\overline{g}(\lambda)\,d\lambda \qquad g = \frac{G}{R+G+B}.$$

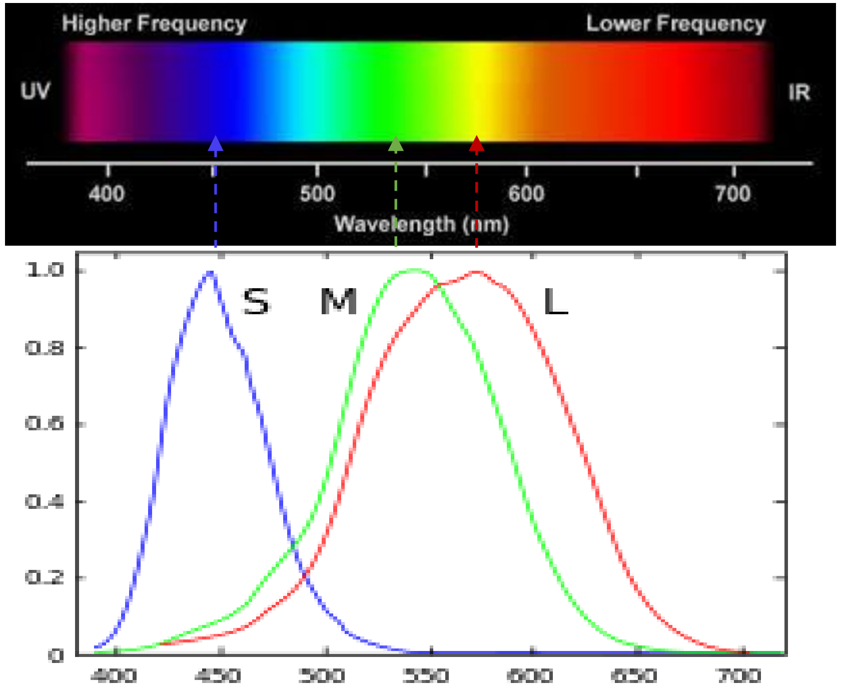
$$B = \int_0^\infty M(\lambda)\,\overline{b}(\lambda)\,d\lambda$$



$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 0.41847 & -0.15866 & -0.082835 \\ -0.091169 & 0.25243 & 0.015708 \\ 0.00092090 & -0.0025498 & 0.17860 \end{bmatrix} \cdot \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

Visible spectrum/Sensitivity

- three kinds of cone cells, which sense light, with spectral sensitivity peaks in
  - short (*S*, 420–440 nm),
  - middle (*M*, 530–540 nm),
  - long (*L*, 560–580 nm) wavelengths
- *three* parameters, corresponding to levels of *stimulus* of the three types of cone cells, can in principle describe any color sensation
  - LMS color space

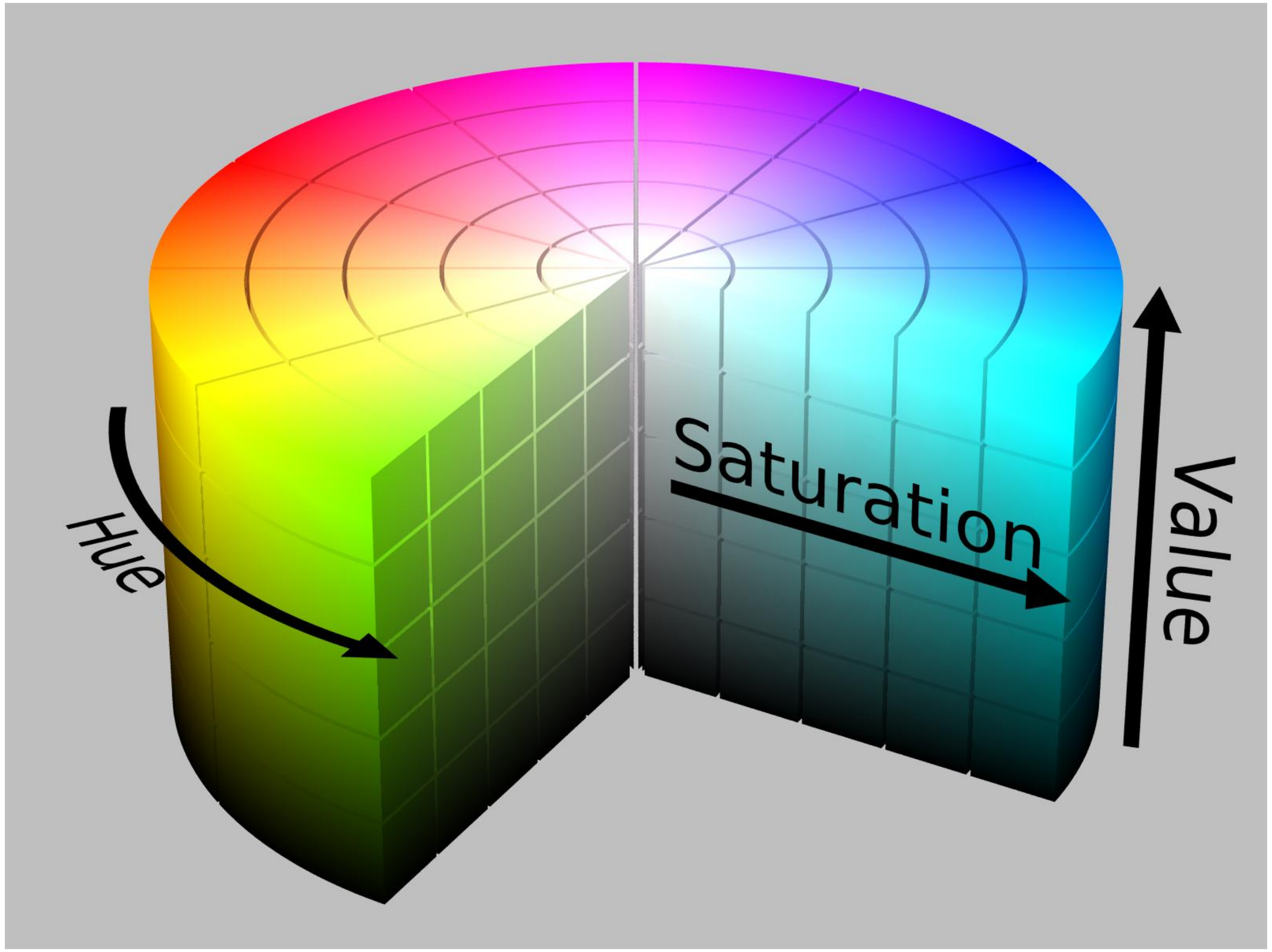


HSB

Hue: Color in pure form

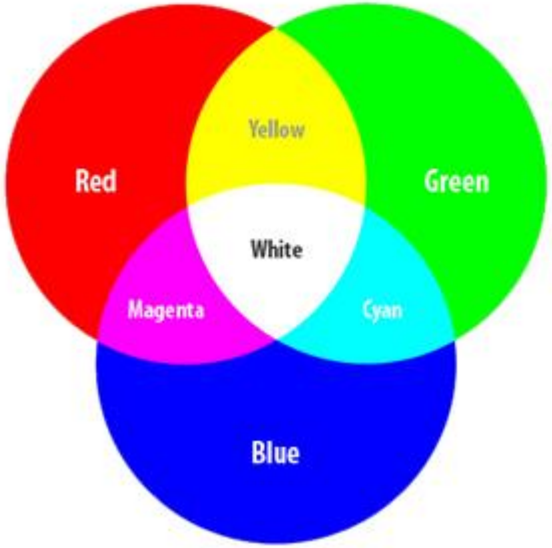
Saturation: Purity – degree to which hue differs from neutral gray with same brightness

Brightness: Level of illumination (intensity of light)



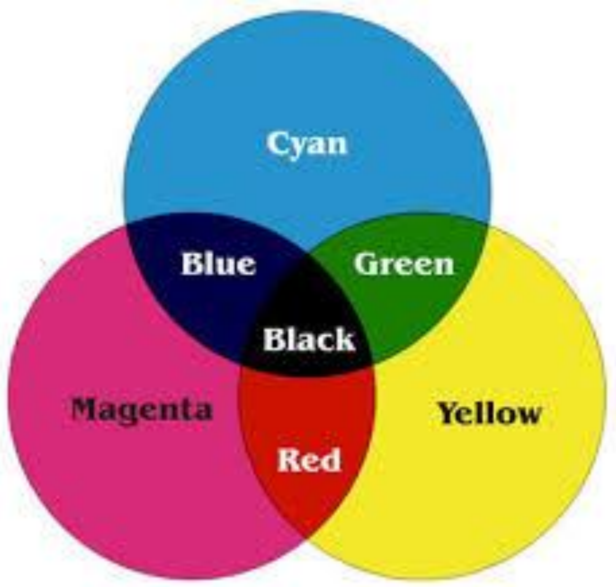
RGB

- RGB (Red-Green-Blue) is the most widely used color system.
- Represents each pixel as a color triplet in the form (R, G, B), e.g., for 24-bit color, each numerical values are 8 bits (varies from 0 to 255).
  - (0, 0, 0) = black
  - (255, 255, 255) = white
  - (255, 0, 0) = red
  - (0, 255, 255) = cyan
  - (65, 65, 65) = a shade of gray
- RGB is an additive model.
  - No beam => no light
  - 3 beams => white



CMYK (Cyan, Magenta, Yellow, Key(black)) color system

- For printing, there is no light source. We see light reflected from the surface of the paper.
- CMYK is a subtractive color model



YUV color system

- PAL (Phase Alternating Line) standard.
  - Humans are more sensitive to luminance (brightness) fidelity than color fidelity.
  - *Luminance* (Y) - Encodes the brightness or intensity.
  - *Chrominance* (U and V) -Encodes the color information.
- Compatible with black/white
- Reduced bandwidth for chrominance components
  - YUV420 uses 1 byte for luminance component, and 4 bits for each chrominance components.
    - Requires only 2/3 of the space (RGB = 24 bits), so better compression! This coding ratio is called 4:2:2 subsampling.
- RGB => YUV

  Y = 0.299R + 0.587G + 0.114B

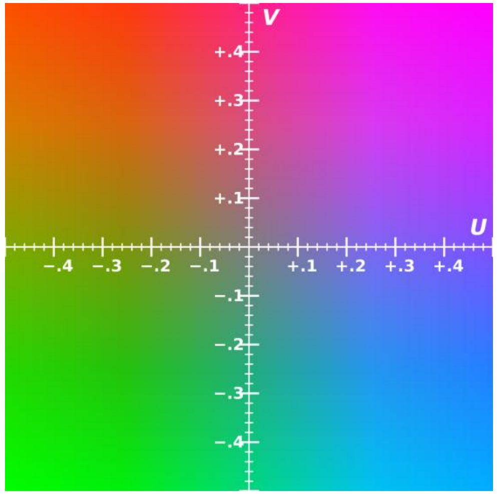
  U = (B-Y) * 0.492=-0.14713R-0.28886G+0.436B

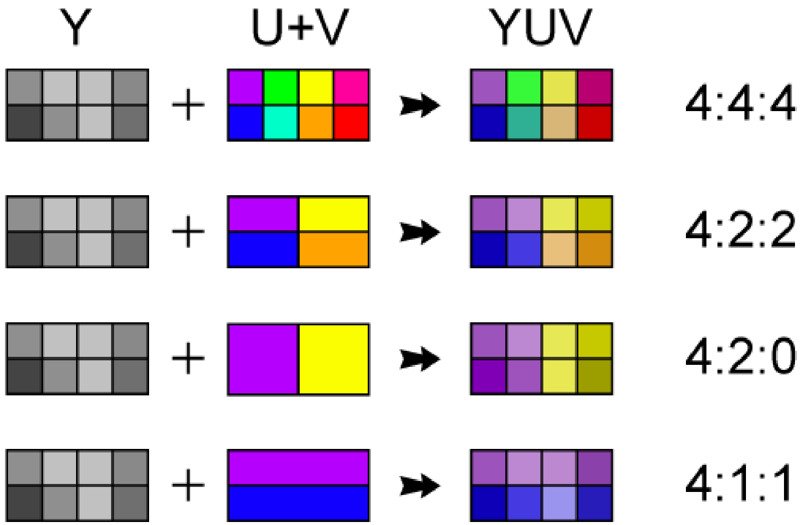
  V = (R-Y) * 0.877=0.615R-0.51499G-0.10001B
- YUV => RBG

  R=Y + 1.14V

  G=Y - 0.395U - 0.581V

  B=Y + 2.033U





YCrCb color system

- Closely related to YUV. It is a scaled and shifted YUV.
  - Cb (blue) and Cr (red) chrominance.
  - Used in JPEG and MPEG.
- YCbCr => RGB

  Y = 0.257R + 0.504G + 0.098B + 16

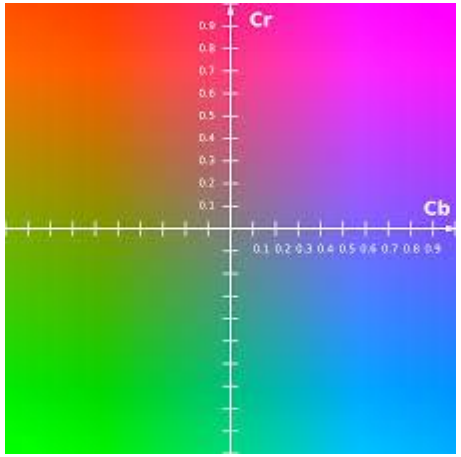
  Cb = - 0.148R - 0.291G + 0.439B + 128

  Cr = 0.439R - 0.368G - 0.071B + 128
- RGB => YCbCr

  R=1.164Y+1.596Cr-222.921

  G=1.164Y-0.392Cb-0.823Cr+135.576

  B=1.164Y+2.017Cb-276.836



$$Y \in [16,235]$$
$$Cb \in [16,240]$$
$$Cr \in [16,240]$$

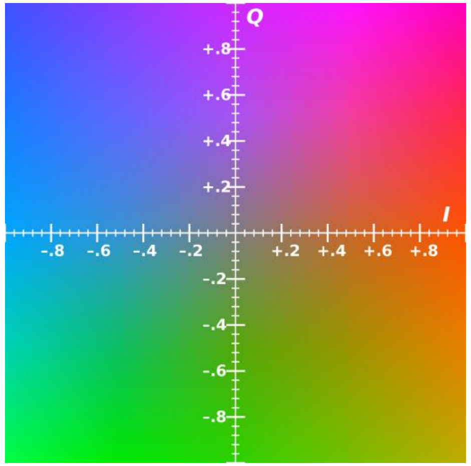
YIQ color system

- Used in NTSC color TV broadcasting. B/W TV if only Y is used.
- YIQ signal
  - similar to YUV

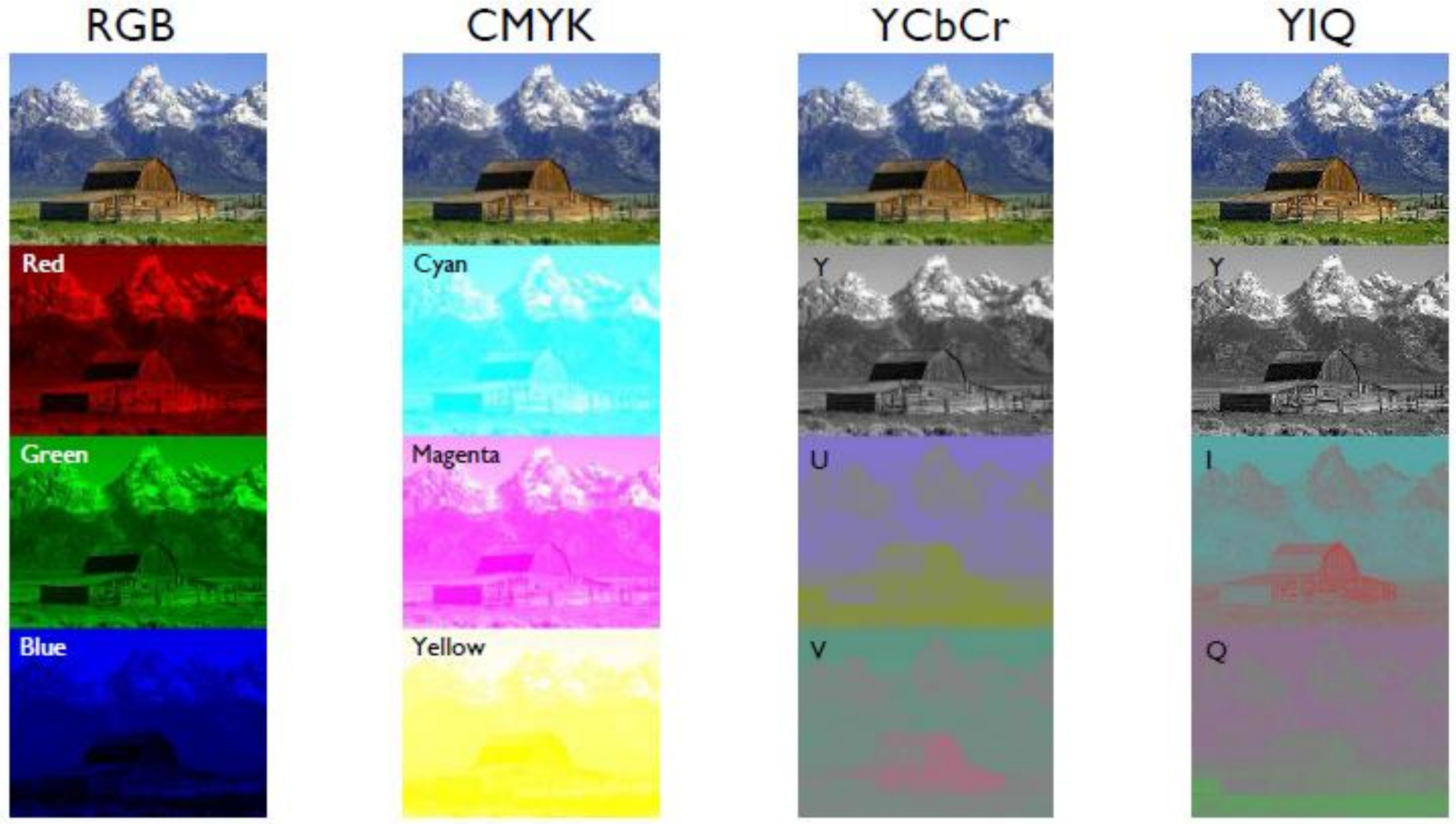
    Y = 0.299R + 0.587G + 0.114B

    I = 0.596R - 0.275G - 0.321B

    Q = 0.212R -0.528G + 0.311B
- Composite signal
  - All information is composed into one signal.
  - To decode, need modulation methods for eliminating interference b/w luminance and chrominance components.



Color decomposition



RGB

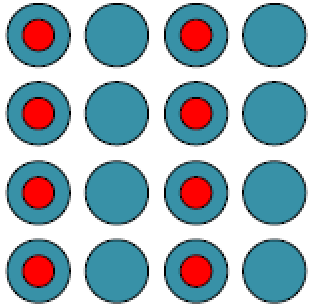
CMYK

YCbCr

YIQ

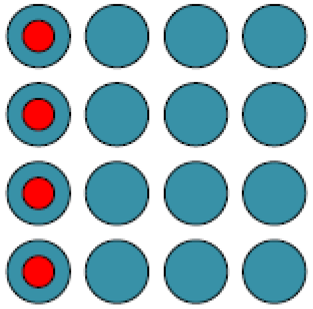
Chrominance Subsampling

- Cut chrominance bandwidth in half/quarter?
  - Use 4-bits/2bits per pixel.
- Human eye less sensitive to variations in color than in brightness.
- Compression achieved with little loss in perceptual quality.
- **4:2:2 Subsampling**
  - For every 4 luminance samples, take 2 chrominance samples (subsampling by 2:1 horizontally only).
  - Chrominance planes just as tall, half as wide.
  - Reduces bandwidth by 1/3
  - Used in professional editing (high-end digital video formats)

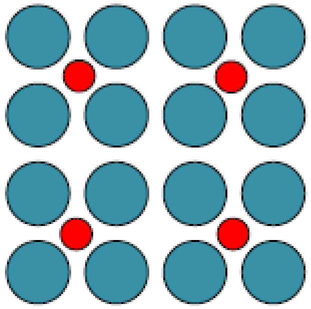


Chrominance subsampling

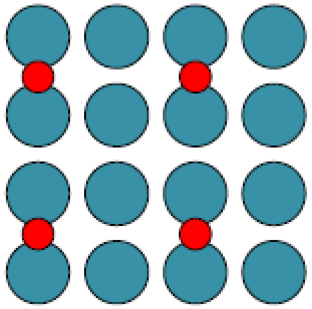
- **4:1:1 Subsampling**
  - For every 4 luminance samples, take 1 chrominance sample (subsampling by 4:1 horizontally only).
  - Used in digital video.



- **4:2:0 Subsampling**
  - For every 4 luminance samples, take 1 chrominance sample (subsampling by 2:1 both horizontally and vertically).
  - Chrominance halved in both directions.
  - Most commonly used.
  - Three varieties:



JPEG, MPEG-1, MJPEG



MPEG-2

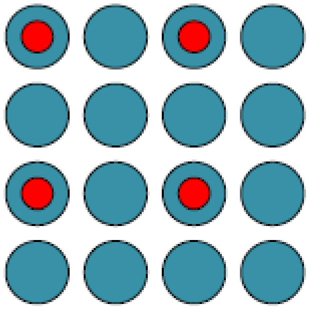


Image compression motivation

- A single digitized image of 1024 pixels x 1024 pixels, 24 bits per pixels requires
  - ~25 Mbits of storage
  - ~7 minutes to send over a 64 Kbps modem!
  - ~3-25 seconds to send over a 1-8 Mbps ADSL!
    - Think of downloading a document with several such images.
  - Some form of compression is needed!

Image compression basics

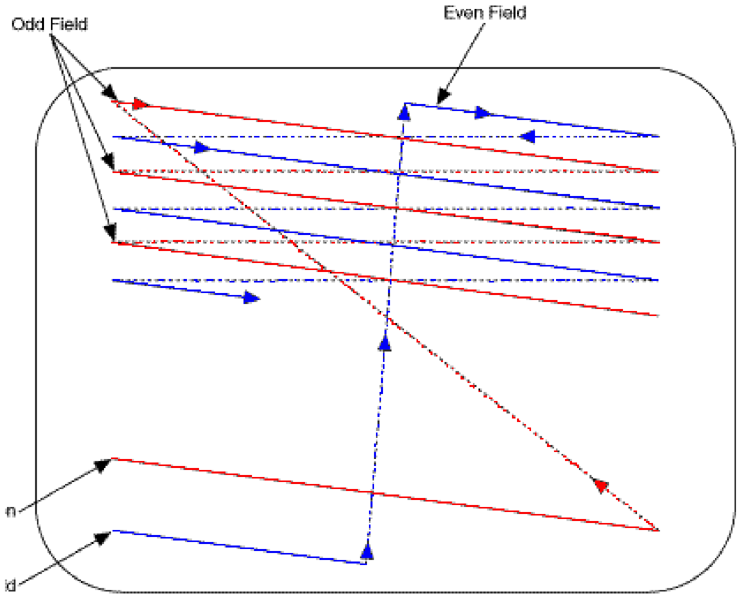
- Lossless - no information is lost:
  - Exploits redundancy / probability distribution
  - Most probable data encoded with fewer bits
- Lossy - approximation of original image
  - Looks for how pixel values change
  - Human eye more sensitive to luminance than chrominance.
  - Human eye less sensitive to subtle feature of the image.
    - Give priority to low-pass image signal wrt high pass image signal
- JPEG uses both techniques

Digital Video Representation

- Can be thought of as a sequence of moving images (or frames).
- Important parameters in video:
  - Frame (image) resolution (e.g., *n*x*m* pixels)
  - Quantization (e.g., *k*-bits per pixel)
  - Frame rate (*p* frames per second, i.e., fps)
- Continuity of motion is achieved at
  - a minimal 15 fps
  - is good at 30 fps
  - HDTV recommends 60 fps!

Standard Video Data Formats

- National Television System Committee (NTSC)
  - Set the standard for transmission of analog color pictures back in 1953!
  - Used in the US and Japan.
  - 525 lines (480 visible) per frame.
  - Resolution? Not digital, but equivalent to the quality produced by a digital image of 720x486 pixels.
  - 30 fps (i.e., delay between frames = 33.3 ms).
  - Video aspect ratio of 4:3 (e.g., 12 in. wide, 9 in. high)
  - Two interlaced fields per frame at 262.5 lines at 60 fields per second
    - Matches 60 Hz power line frequency
    - Increases vertical resolution (to which eye is sensitive)



Standard Video formats

- PAL (Phase Alternating Line):
  - Used in parts of Western Europe.
  - 625 lines (576 visible) per frame.
  - 25 fps (i.e., delay between frames = 40 ms).
  - Two interlaced fields per frame at 312.5 lines at 50 fields per second
    - Matches 50 Hz power line frequency
    - Increases vertical resolution (to which eye is sensitive)
- SECAM: French Standard

SDTV/HDTV/UHDTV

- Technical Societies
  - Advanced Television Systems Committee (ATSC)
  - **MPEG (Motion Pictures Experts Group)**
  - SMPTE (Society of Motion Pictures & Television Engineers)
- 60fps+
- SDTV
  - Resolutions of 720x480, 720x576 pixels
- HDTV
  - Resolutions of 1280x720, 1920x1080 pixels
- Ultra HDTV
  - Resolutions of 3840x2160 (4K), 7680x4320 (8K) pixels
- Video aspect ratio of 16:9 (wide screen)
- MPEG-2/MPEG-4 H.264/H.265 for video compression
- Both interlaced and progressive (except for H.265 which does not support interlaced)
- AC-3 (Dolby Audio Coding-3)/AAC (MPEG-2, MPEG-4) for audio compression