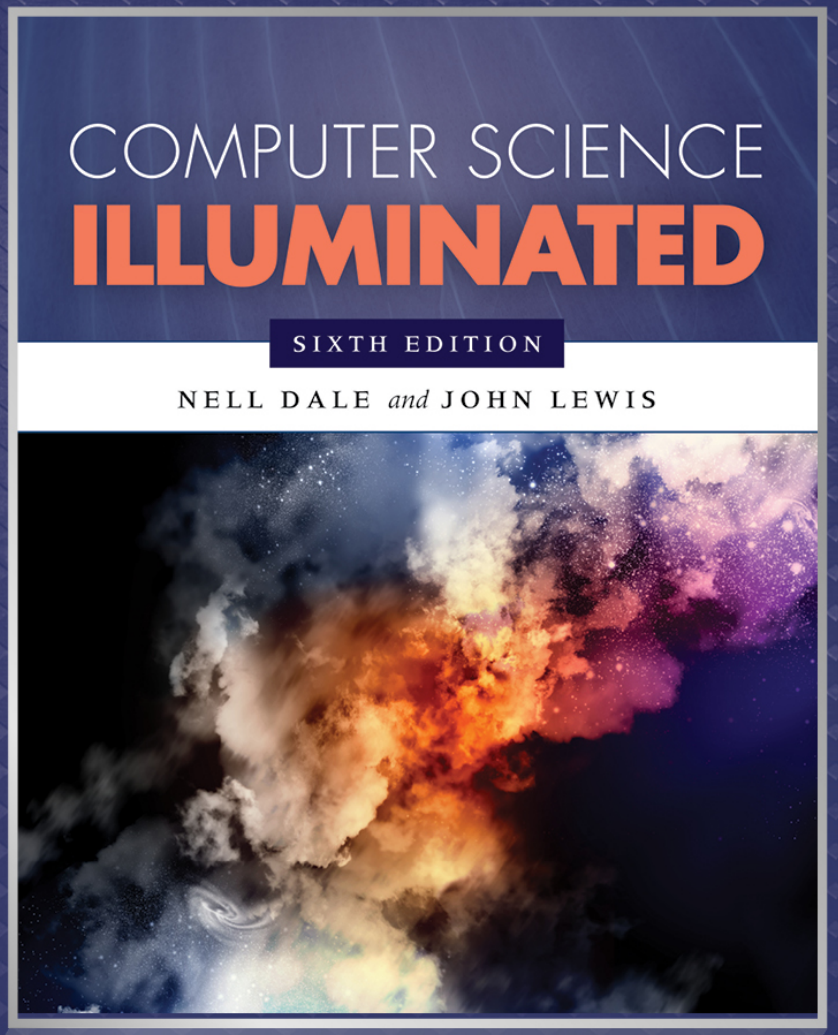
Chapter 3

Data Representation



Chapter Goals

- Distinguish between analog and digital information
- Explain data compression and calculate compression ratios
- Explain the binary formats for negative and floating-point values
- Describe the characteristics of the ASCII and Unicode character sets
- Perform various types of text compression

Chapter Goals

- Explain the nature of sound and its representation
- Explain how RGB values define a color
- Distinguish between raster and vector graphics
- Explain temporal and spatial video compression

Data and Computers

Computers are **multimedia** devices, dealing with a vast array of information categories

Computers store, present, and help us modify

- Numbers
- Text
- Audio
- Images and graphics
- Video

All stored as binary digits (**bits**)

Data and Computers

Data compression

Reduction in the amount of space needed to store a piece of data or the bandwidth to transmit it

Compression ratio

The size of the compressed data divided by the size of the original data

A data compression technique can be

- **lossless**, which means the data can be retrieved without any loss of original information
- **lossy**, which means some information may be lost in the process of compression

Analog and Digital Information

Computers are finite!

How do we represent an infinite world?

We represent enough of the world to satisfy our computational needs and our senses of sight and sound

Analog and Digital Information

Information can be represented in one of two ways: **analog** or **digital**

Analog data

A continuous representation, analogous to the actual information it represents

Digital data

A discrete representation, breaking the information up into separate elements

Analog and Digital Information

A mercury thermometer is an analog device

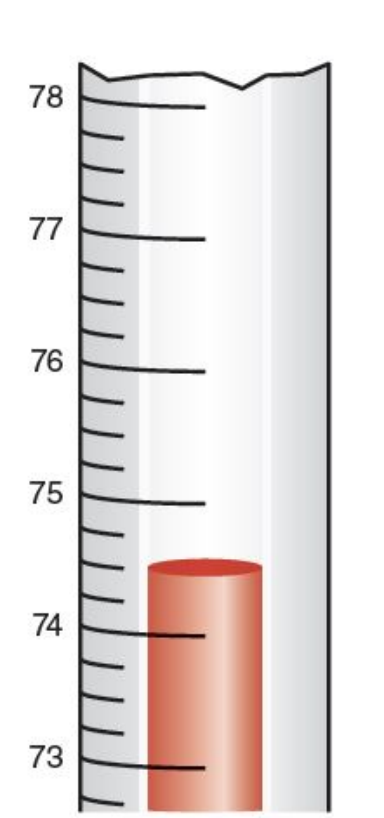


FIGURE 3.1 A mercury thermometer continually rises in direct proportion to the temperature

Analog and Digital Information

Computers cannot work well with analog data, so we digitize the data

Digitize

Breaking data into pieces and representing those pieces separately

Why do we use binary to represent digitized data?

Electronic Signals

Important facts about electronic signals

- An analog signal continually fluctuates in voltage up and down
- A digital signal has only a high or low state, corresponding to the two binary digits
- All electronic signals (both analog and digital) degrade as they move down a line
- The voltage of the signal fluctuates due to environmental effects

Electronic Signals *(Cont'd)*

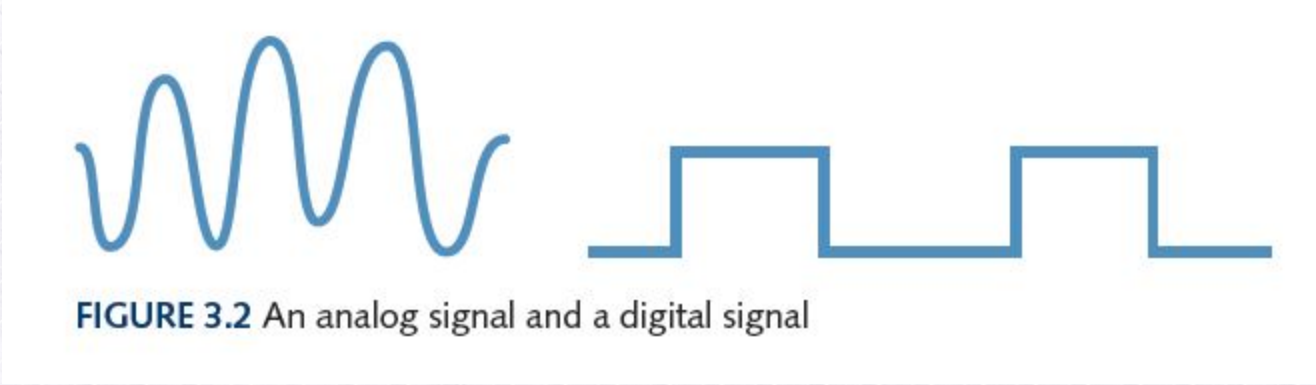


FIGURE 3.2 An analog signal and a digital signal

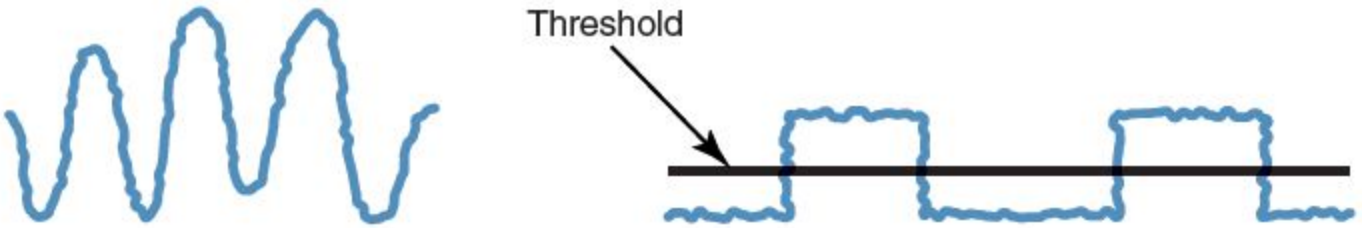


FIGURE 3.3 Degradation of analog and digital signals

Periodically, a digital signal is **reclocked** to regain its original shape

Binary Representations

- Each bit can be either 0 or 1, so it can represent a choice between two possibilities (or “two things”)
- Two bits can represent four things *(Why? Hint:* 00, 01, 10, 11*.)*

How many things can three bits represent?

How many things can four bits represent?

How many things can eight bits represent?

Binary Representations

1 Bit	2 Bits	3 Bits	4 Bits	5 Bits
0	00	000	0000	00000
1	01	001	0001	00001
	10	010	0010	00010
	11	011	0011	00011
		100	0100	00100
		101	0101	00101
		110	0110	00110
		111	0111	00111
			1000	01000
			1001	01001
			1010	01010
			1011	01011
			1100	01100
			1101	01101
			1110	01110
			1111	01111
				10000
				10001
				10010
				10011
				10100
				10101
				10110
				10111
				11000
				11001
				11010
				11011
				11100
				11101
				11110
				11111

FIGURE 3.4 Bit combinations

Binary Representations

How many bits are needed to represent 32 things? One hundred things?

How many things can n bits *represent?*

Why?

What happens every time you increase the number of bits by one?

Representing Natural Numbers

8-bit Binary Representation	Natural Number
01111111	127
01111110	126
…	…
00000011	3
00000010	2
00000001	1
00000000	0

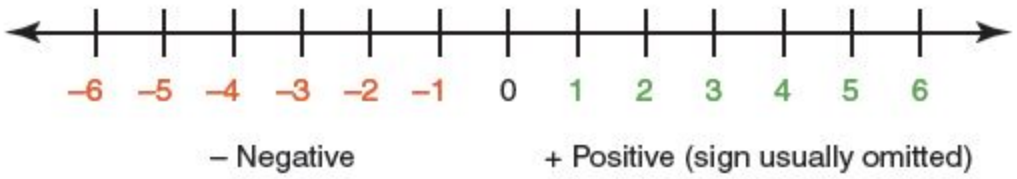
- Easy! Just convert to binary
- Computers store data in fixed-size chunks, so we have leading zeroes

What do the integers include that the natural numbers do not?

Representing Negative Values

Signed-magnitude number representation

- Used by humans
- The sign represents the ordering (the negatives come before the positives in ascending order)
- The digits represent the magnitude (the distance from zero)



Representing Negative Values

Problem: Two zeroes (positive and negative)

No problem for humans, but would cause unnecessary complexity in computers

Solution: Represent integers by associating them with natural numbers

Half the natural numbers will represent themselves

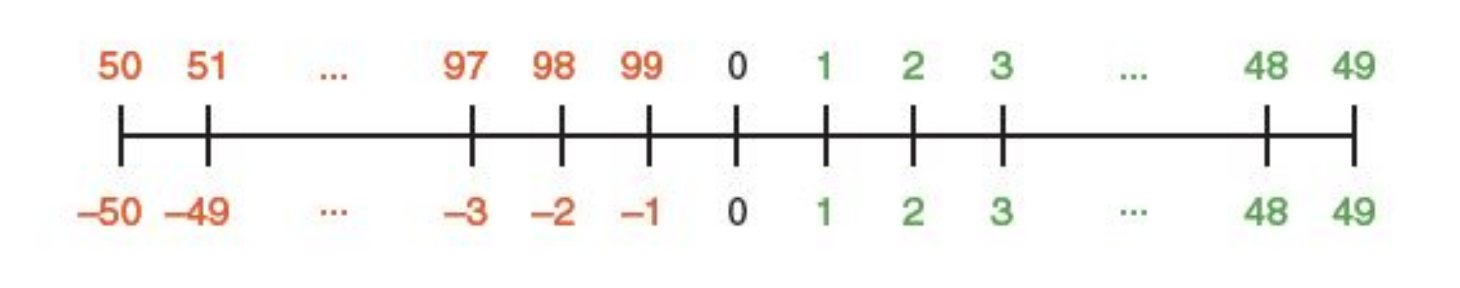
The other half will represent negative integers

Representing Negative Values

Using two decimal digits,

let 0 through 49 represent 0 through 49

let 50 through 99 represent -50 through -1



Representing Negative Values

To perform addition, add the numbers and discard any carry to the hundreds digit

Signed-Magnitude	New Scheme
5 + −6 −1	5 + 94 99
−4 + 6 2	96 + 6 2
−2 + −4 −6	98 + 96 94

Now you try it

48 (signed-magnitude)
- 1
47

How does it work in the new scheme?

Representing Negative Values

To perform subtraction, use A – B = A + (-B)

Add the negative of the second to the first

Signed-Magnitude	New Scheme	Add Negative
– 5 – 3 – 8	95 – 3	95 + 97 92

Try these:

4	*4*	*- 1*
- 3	*- (- 3)*	*- 2*

Representing Negative Values

Called ten's complement representation, because we can use this formula to compute the representation of a negative number

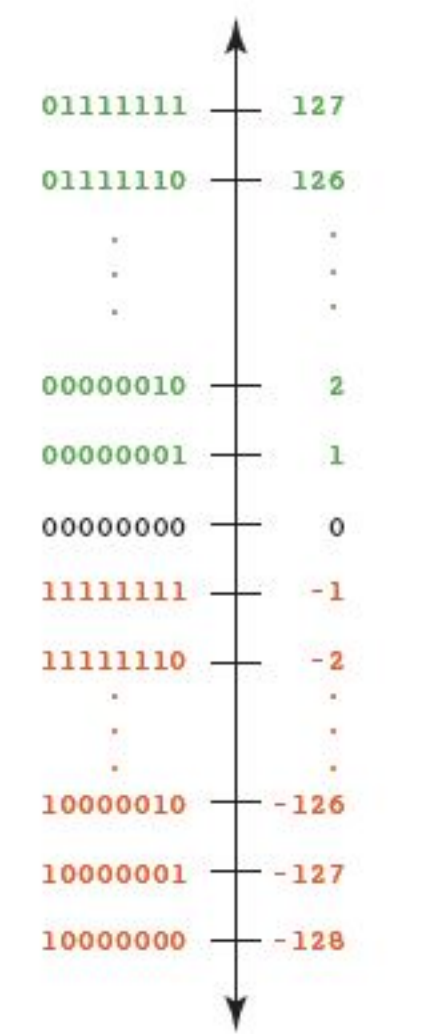
$$\text{Negative}(I) = 10^k - I, \text{ where } k \text{ is the number of digits}$$

For example, -3 is Negative(3), so using two digits, its representation is

Negative(3) = 100 – 3 = 97

What do we get if we try this in binary?

Representing Negative Values



Two's Complement

(The binary number line is easier to read when written vertically)

Remember our table showing how to represent natural numbers?

Do you notice something interesting about the left-most bit?

Representing Negative Values

Addition and subtraction are the same as in ten's complement arithmetic

-127	10000001
+ 1	00000001
-126	10000010

What if the computed value won't fit?

Number Overflow

If each value is stored using 8 bits, then 127 + 3 overflows:

```
    01111111
+   00000011
    10000010
```

Apparently, 127 + 3 is -126. Remember when we said we would always fail in our attempt to map an infinite world onto a finite machine?

Most computers use 32 or 64 bits for integers, but there are always infinitely many that aren't represented

Representing Real Numbers

Real numbers are numbers with a whole part and a fractional part (either of which may be zero)

104.32
0.999999
357.0
3.14159

In decimal, positions to the right of the decimal point are the tenths, hundredths, thousandths, etc.:

10^{-1}, 10^{-2} , 10^{-3} …

Representing Real Numbers

Same rules apply in binary as in decimal

Radix point is general term for “decimal point”

Positions to the right of the radix point in binary:

2^{-1} (halves position),
2^{-2} (quarters position),
2^{-3} (eighths position)
…

Representing Real Numbers

A real value in base 10 can be defined by the following formula where the mantissa is an integer

$$\text{sign} * \text{mantissa} * 10^{\text{exp}}$$

This representation is called **floating point** because the radix point "floats"

In analogy to the fixed number of bits that computers use to represent integers, we'll treat the mantissa as having a fixed number of digits

Representing Real Numbers

TABLE 3.1 Values in decimal notation and floating-point notation (five digits)

Onion: © matka_Wariatka/ShutterStock, Inc.

Real Value	Floating-Point Value
12001.00	$12001 * 10^{0}$
–120.01	$-12001 * 10^{-2}$
0.12000	$12000 * 10^{-5}$
–123.10	$-12310 * 10^{-2}$
155555000.00	$15555 * 10^{4}$

Floating-point in binary:
sign * mantissa * 2^{exp}

Only the base value is different from decimal

Fundamentally, the floating-point used by computers is very similar, but uses complicated tricks to represent more numbers and improve efficiency

Representing Real Numbers

Scientific notation

A form of floating-point representation in which the decimal point is kept to the right of the leftmost digit

12001.32708 is 1.200132708E+4 in scientific notation (E+4 is how computers display $x10^4$)

What is 123.332 in scientific notation?

What is 0.0034 in scientific notation?

Representing Text

What must be provided to represent text?

The number of characters to represent is finite (whew!), so list them all and assign each a binary string

Character set

A list of characters and the codes used to represent each one

Computer manufacturers agreed to standardize

The ASCII Character Set

ASCII stands for American Standard Code for Information Interchange

ASCII originally used seven bits to represent each character, allowing for 128 unique characters

Later extended ASCII evolved so that all eight bits were used

How many characters could be represented?

ASCII Character Set Mapping

Left Digit(s)	Right Digit	ASCII 0	1	2	3	4	5	6	7	8	9
0		NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT
1		LF	VT	FF	CR	SO	SI	DLE	DC1	DC2	DC3
2		DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS
3		RS	US	□	!	"	#	$	%	&	'
4		(	)	*	+	,	-	.	/	0	1
5		2	3	4	5	6	7	8	9	:	;
6		<	=	>	?	@	A	B	C	D	E
7		F	G	H	I	J	K	L	M	N	O
8		P	Q	R	S	T	U	V	W	X	Y
9		Z	[	\	]	^	_	`	a	b	c
10		d	e	f	g	h	i	j	k	l	m
11		n	o	p	q	r	s	t	u	v	w
12		x	y	z	{	|	}	~	DEL		

FIGURE 3.5 The ASCII character set

The ASCII Character Set

The first 32 characters in the ASCII character chart do not have a simple character representation to print to the screen

What do you think they are used for?

The Unicode Character Set

Extended ASCII is not enough for international use

One Unicode mapping uses 16 bits per character

How many characters can this mapping represent?

The first 256 characters correspond exactly to the extended ASCII character set

The Unicode Character Set

Code (Hex)	Character	Source
0041	A	English (Latin)
042F	Я	Russian (Cyrillic)
0E09	ฉ	Thai
13EA	Ꮺ	Cherokee
211E	℞	Letterlike symbols
21CC	⇌	Arrows
282F	⠯	Braille
345F	㑟	Chinese/Japanese/ Korean (common)

FIGURE 3.6 A few characters in the Unicode character set

Text Compression

If storage or bandwidth is scarce, how can we store and transmit data more efficiently?

Compression is most useful for big files (e.g. audio, graphics, video, and scientific data)

Text files are typically pretty small, but as an illustration, can we use less than 16 bits per character without losing information?

Lossless compression techniques include

- Keyword encoding
- Run-length encoding
- Huffman encoding

Keyword Encoding

Replace frequently used patterns of text with a single special character, such as:

WORD	SYMBOL
as	^
the	~
and	+
that	$
must	&
well	%
these	#

Keyword Encoding

Given the following paragraph,

> We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. — That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, — That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness.

Keyword Encoding

The encoded paragraph is

We hold # truths to be self-evident, $ all men are created equal, $ ~y are endowed by ~ir Creator with certain unalienable Rights, $ among # are Life, Liberty + ~ pursuit of Happiness. — $ to secure # rights, Governments are instituted among Men, deriving ~ir just powers from ~ consent of ~ governed, — $ whenever any Form of Government becomes destructive of # ends, it is ~ Right of ~ People to alter or to abolish it, + to institute new Government, laying its foundation on such principles + organizing its powers in such form, ^ to ~m shall seem most likely to effect ~ir Safety + Happiness.

Keyword Encoding

What did we save?

Original paragraph

656 characters

Encoded paragraph

596 characters

Characters saved

60 characters

Compression ratio

596/656 = 0.9085

Could we use this substitution chart for all text?

Run-Length Encoding

In some types of data files, a single value may be repeated over and over again in a long sequence

Replace a repeated sequence with

- a flag
- the repeated value
- the number of repetitions

*n8

- * is the flag
- n is the repeated value
- 8 is the number of times n is repeated

Run-Length Encoding

Original text

bbbbbbbbjjjkllqqqqqq+++++

Encoded text

*b8jjjkll*q6*+5 *(Why isn't J encoded? L?)*

The compression ratio is 15/25 or .6

Encoded text

*x4*p4l*k7

Original text

xxxxpppplkkkkkkk

This type of repetition doesn't occur in English text; can you think of a situation where it might occur?

Huffman Encoding

The characters ‘X’ and ‘z’ occur much less frequently than ‘e’ and the space character, for example.

What if we could use fewer bits for common characters in exchange for using more bits for uncommon characters?

This is the idea behind prefix codes, including Huffman codes

Huffman Encoding

Huffman Code	Character
00	A
01	E
100	L
110	O
111	R
1010	B
1011	D

"ballboard" would be

10100010
01001010
11000111
1011xxxx

compression ratio
4 bytes / 18 bytes = 0.222
assuming 16-bit Unicode

Try "roadbed"

Note: only the part of the code needed to encode "ballboard" and "roadbed" is shown. In the full code, every character would have an encoding, and the most common characters would have the shortest encodings.

Huffman Encoding

Huffman encoding is an example of prefix coding: no character's bit string is the prefix of any other character's bit string

To decode

- Look for match left to right, bit by bit
- Record letter when a match is found
- Begin where you left off, going left to right

Huffman Encoding

Try it!

Huffman Code	Character
00	A
01	E
100	L
110	O
111	R
1010	B
1011	D

Decode

1011111001010

Huffman Encoding

Technique for determining codes guarantees the prefix property of the codes

Two types of codes based on where the frequencies come from

- General, based on use of letters in English (Spanish, ….)
- Specialized, based on text itself or specific types of text

Representing Audio Information

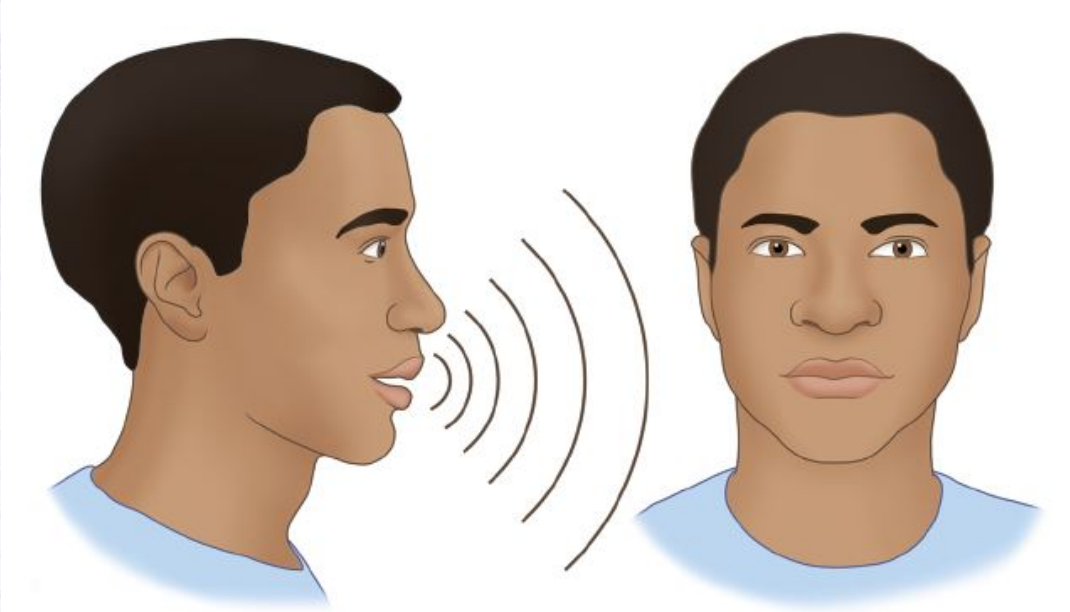


FIGURE 3.7 A sound wave vibrates our eardrums

We perceive sound when a series of air pressure waves vibrate a membrane in our ear, which sends signals to our brain

Representing Audio Information

Your parents may use a "stereo" to listen to music at home. It sends an electrical signal to each speaker, which then vibrates to produce sound. Your MP3 player and ear buds do the same thing.

The signal controls the motion of a membrane in the speaker, which in turn creates the pressure waves that reach our ears

Thus, the signal is an analog representation of the sound wave

Representing Audio Information

Digitize the signal by

- Sampling: periodically measure the voltage
- Quantization: represent the voltage as a number using a finite number of bits

How often should we sample?

A sampling rate of about 40,000 times per second is enough to create a reasonable sound reproduction

Representing Audio Information

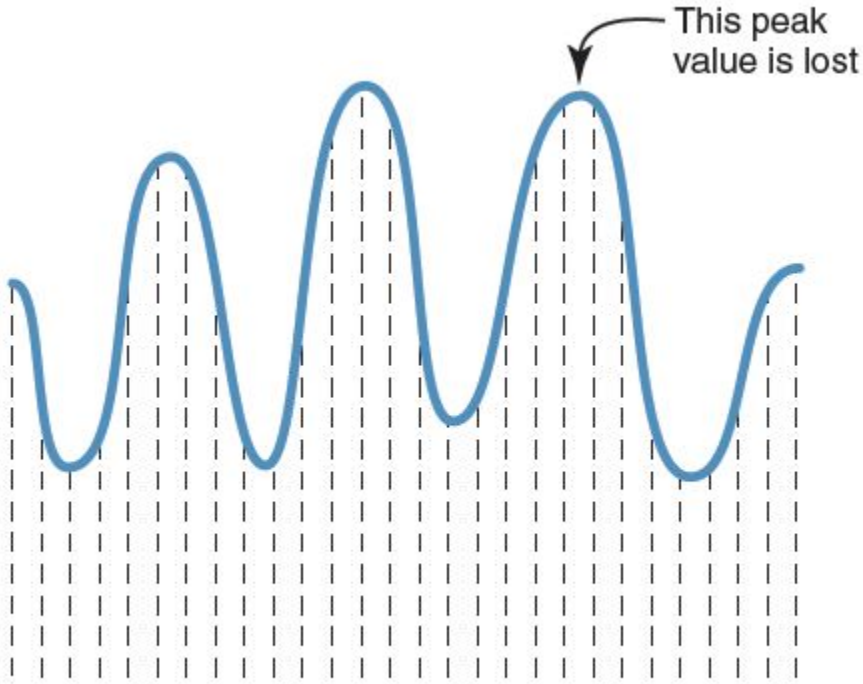


FIGURE 3.8 Sampling an audio signal

Some data is lost, but a reasonable sound is reproduced

Representing Audio Information

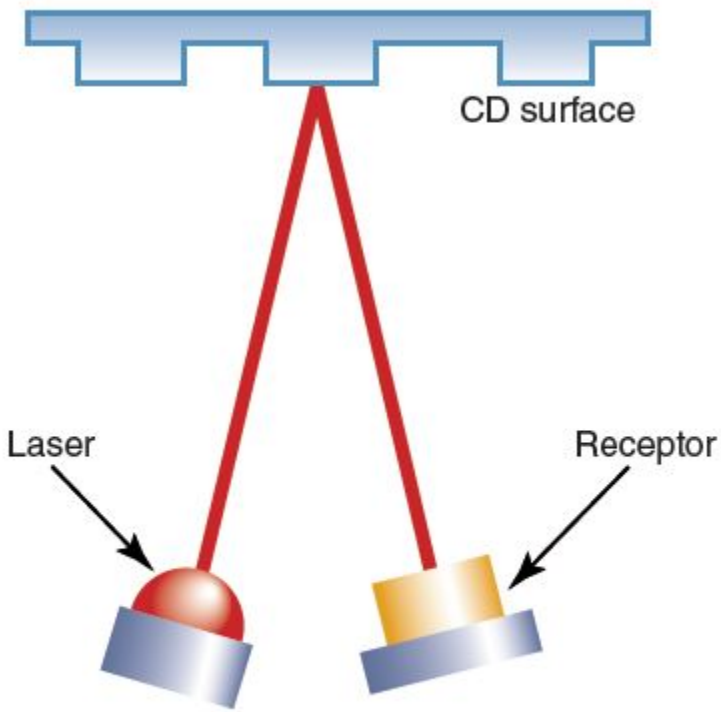


FIGURE 3.9 A CD player reading binary data

- CDs store audio (or other) information digitally
  - Pits (reflect poorly)
  - Lands (reflect well)
- Read by low intensity laser
- Receptor converts reflections into binary digits
- Bit string represents audio signal

Audio Formats

Audio Formats
- WAV, AU, AIFF, VQF, and MP3
- Use various compression techniques

MP3 is dominant
- MPEG-2, audio layer 3 file
- MPEG = Motion Picture Experts Group
- Based on studies of interrelation between ear and brain, discards frequency information that isn't perceived by humans (science!)
- Additional compression by a form of Huffman encoding

Is this a lossy or lossless compression (or both)?

Representing Images and Graphics

Color

- We take it for granted, but what is it really?

Retinas of our eyes have three types of photoreceptor cone cells

- Each type responds to a different set of frequencies of light
- Our brain translates that response into a perception of red, green, or blue

Representing Images and Graphics

Color is expressed as an RGB (red-green-blue) value – three numbers that indicate the relative contribution of each of these three primary colors

An RGB value of (255, 255, 0) maximizes the contribution of red and green, and minimizes the contribution of blue, which results in a bright yellow

Representing Images and Graphics

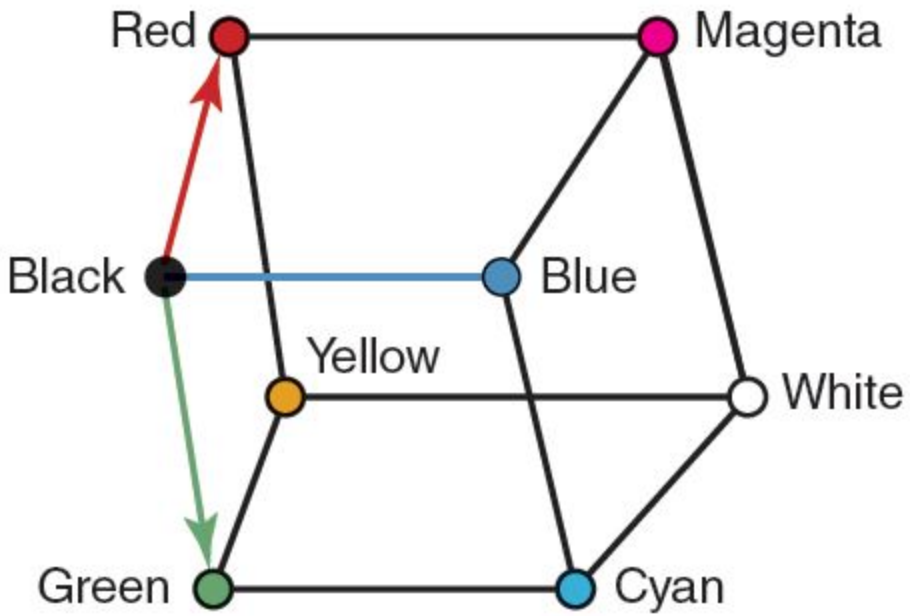


FIGURE 3.10 A three-dimensional color space

Representing Images and Graphics

Color depth

The amount of data that is used to represent a color

HiColor

A 16-bit color depth: five bits used for each number in an RGB value with the extra bit sometimes used to represent transparency

TrueColor

A 24-bit color depth: eight bits used for each number in an RGB value

Representing Images and Graphics

RGB VALUE			
Red	Green	Blue	Color
0	0	0	black
255	255	255	white
255	255	0	yellow
255	130	255	pink
146	81	0	brown
157	95	82	purple
140	0	0	maroon

A few TrueColor RGB values and the colors they represent

Representing Images and Graphics

A color palette is a set of colors, for example

- Colors supported by a monitor
- Web-safe colors for use with Internet browsers
- Colors from which user can choose
- Colors used in an image

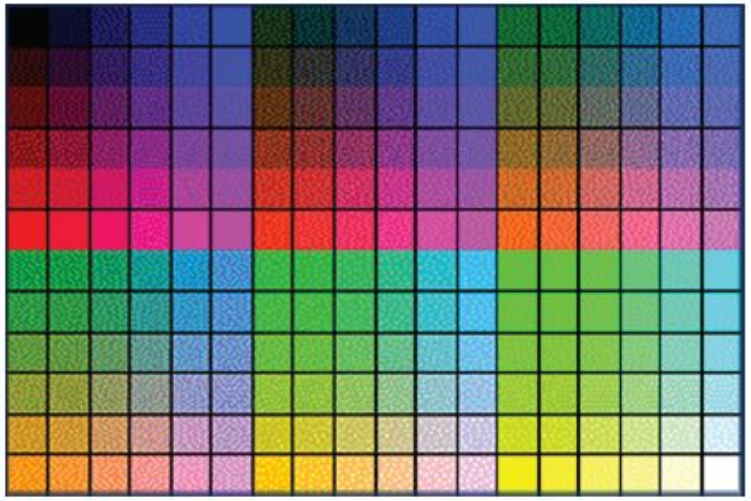


FIGURE 3.11 A restricted color palette

Digitized Images and Graphics

- Pixels (picture elements)
  - Dots of color in image (or display device)
- Resolution
  - Number of pixels in image (or device)
- Raster Graphics
  - Treat image as collection of pixels
  - Most common formats: BMP, GIF, PNG, and JPEG
- Vector Graphics
  - Treat image as collection of geometric objects
  - Most important formats: Flash and SVG

Digitized Images and Graphics

- BMP (bitmap)
  - TrueColor color depth, or less to reduce file size
  - Well suited for compression by run-length encoding
- GIF (indexed color)
  - File explicitly includes palette of 256 or fewer colors
  - Each pixel thus requires only 8 or fewer bits
  - Animated GIFs are short sequences of images
- PNG (Portable Network Graphics)
  - Intended to replace GIFs
  - Greater compression with wider range of color depths
  - No animation

Digitized Images and Graphics

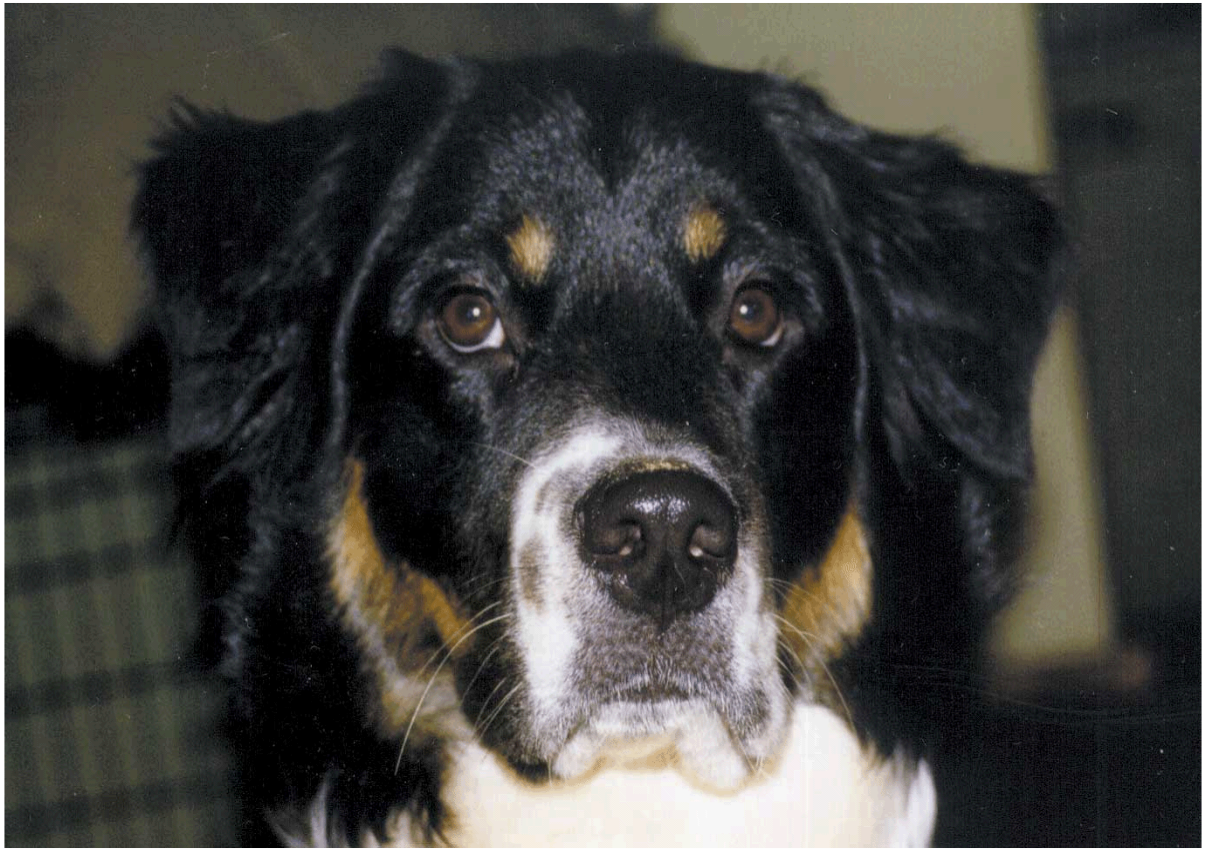
- JPEG (Joint Photographic Experts Group)
  - Averages hues over short distances
    - Why? Human vision tends to blur colors together within small areas (science!)
    - How? Transform from the spatial domain to the frequency domain, then discard high frequency components (math!)
    - Sound familiar? Essentially the same idea used in MP3
  - Adjustable degree of compression

Raster graphics recap: BMP, GIF, PNG, and JPEG

Which use lossless compression? Lossy?

Which would you use for line art? For a color photograph?

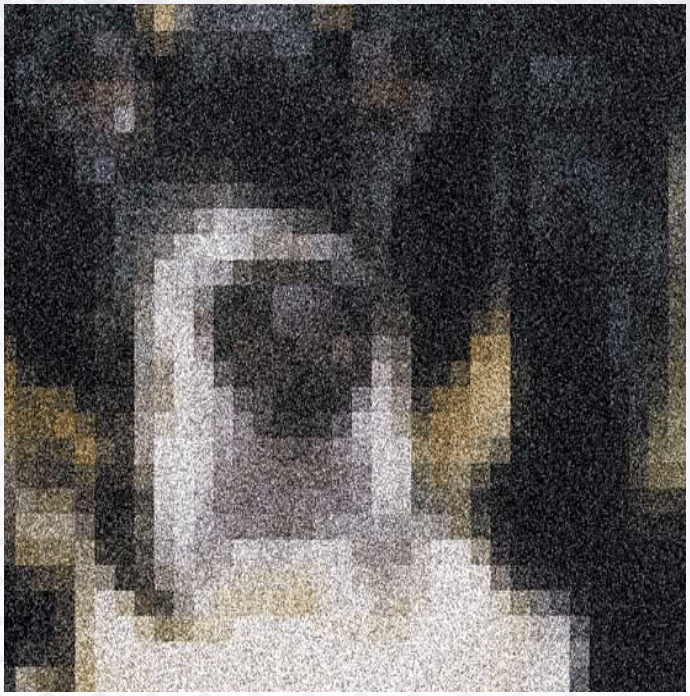
Digitized Images and Graphics



Whole picture

Figure 3.12 A digitized picture composed of many individual pixels

Digitized Images and Graphics



Magnified portion of the picture

See the pixels?

Each pixel of the image now fills a block of screen pixels

Figure 3.12 A digitized picture composed of many individual pixels

Vector Graphics

Vector graphics

- A format that describes an image in terms of lines and geometric shapes
- A vector graphic is a series of commands that describe shapes using mathematical properties (e.g. direction, length, thickness, color)
- For some types of images, the file sizes can be smaller than with raster graphics because not every pixel is described.

Vector Graphics

The good side and the bad side…

Vector graphics can be resized mathematically and changes can be calculated dynamically as needed.

Vector graphics are good for line art (e.g. diagrams) and cartoon-style drawings

Vector graphics are *not* good for representing images of the real-world

Representing Video

Video codec COmpressor/DECompressor

Methods used to shrink the size of a movie to allow it to be played on a computer or over a network

Almost all video codecs use lossy compression to minimize the huge amounts of data associated with video

Representing Video

Temporal compression

A technique based on differences between consecutive frames: If most of an image in two frames has not changed, why should we waste space duplicating information?

Spatial compression

A technique based on removing repetitive information within a frame: This problem is essentially the same as that faced when compressing still images

Ethical Issues

The Fallout from Snowden's Revelations

What government program was revealed by the documents that Edward Snowden leaked?

When was this program first authorized?

What led to President Obama announcing that the program would be scaled back?

Do you think Snowden is a criminal? A hero? A traitor? A patriot?

Who am I?

I was very versatile. Can you name four items on my resume?

Do you know?

How many computer character sets existed in 1960?

What happened between 10/24/02 and 10/26/02 to guests of Holiday Inn, Holiday Inn Express or Crown Plaza?

What criteria are used in Japan's phone answering competition?

Who described the telegraph as a kind of very long cat?