computer science
illuminated

Data Representation

Nell Dale & John Lewis
(adaptation by Michael Goldwasser)

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

Analog and Digital Information

- Computers are finite. Computer memory and other hardware devices have only so much room to store and manipulate a certain amount of data. The goal, is to represent enough of the world to satisfy our computational needs and our senses of sight and sound.

Compression

- **Data compression**–reducing the amount of space needed to store a piece of data.
- **Compression ratio**–is 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 losing any of the original information. Or it can be **lossy**, in which case some information is lost in the process of compaction.

Representing Text

- To represent a text document in digital form, we simply need to be able to represent every possible character that may appear.
- There are finite number of characters to represent. So the general approach for representing characters is to list them all and assign each a binary string.
- A **character set** is simply a list of characters and the codes used to represent each one. By agreeing to use a particular character set, computer manufacturers have made the processing of text data easier.

The ASCII Character Set

- ASCII stands for American Standard Code for Information Interchange. The ASCII character set originally used seven bits to represent each character, allowing for 128 unique characters.
- Later ASCII evolved so that all eight bits were used which allows for 256 characters.

The ASCII Character Set *(Cont'd)*

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		

The Unicode Character Set

- The extended version of the ASCII character set is not enough for international use.
- The Unicode character set uses 16 bits per character. Therefore, the Unicode character set can represent $2^{16} = 65536$ distinct characters.
- Unicode was designed to be a superset of ASCII. That is, the first 256 characters in the Unicode character set correspond exactly to the extended ASCII character set.

The Unicode Character Set *(Cont'd)*

Code (Hex)	Character	Source
0041	A	English (Latin)
042F	Я	Russian (Cyrillic)
OE09	ก	Thai
13EA	Ꮺ	Cherokee
211E	℞	Letterlike Symbols
21CC	⇌	Arrows
282F	⠯	Braille
345F	㑯	Chinese/Japanese/ Korean (Common)

Figure 3.6 A few characters in the Unicode character set

Future Character Sets

ISO (International Organization for Standardization)

- Might develop 24-bit patterns to represent symbols
  (17 million of them, potentially)

- or even 32-bit patterns (over 2 billion distinct symbols).

Representing Audio Information

- We perceive sound when a series of air compressions vibrate a membrane in our ear, which sends signals to our brain.
- A stereo sends an electrical signal to a speaker to produce sound. This signal is an analog representation of the sound wave. The voltage in the signal varies in direct proportion to the sound wave.

Representing Audio Information
(Cont'd)

- To digitize the signal we periodically measure the voltage of the signal and record the appropriate numeric value. The process is called *sampling*.

Representing Audio Information
(Cont'd)

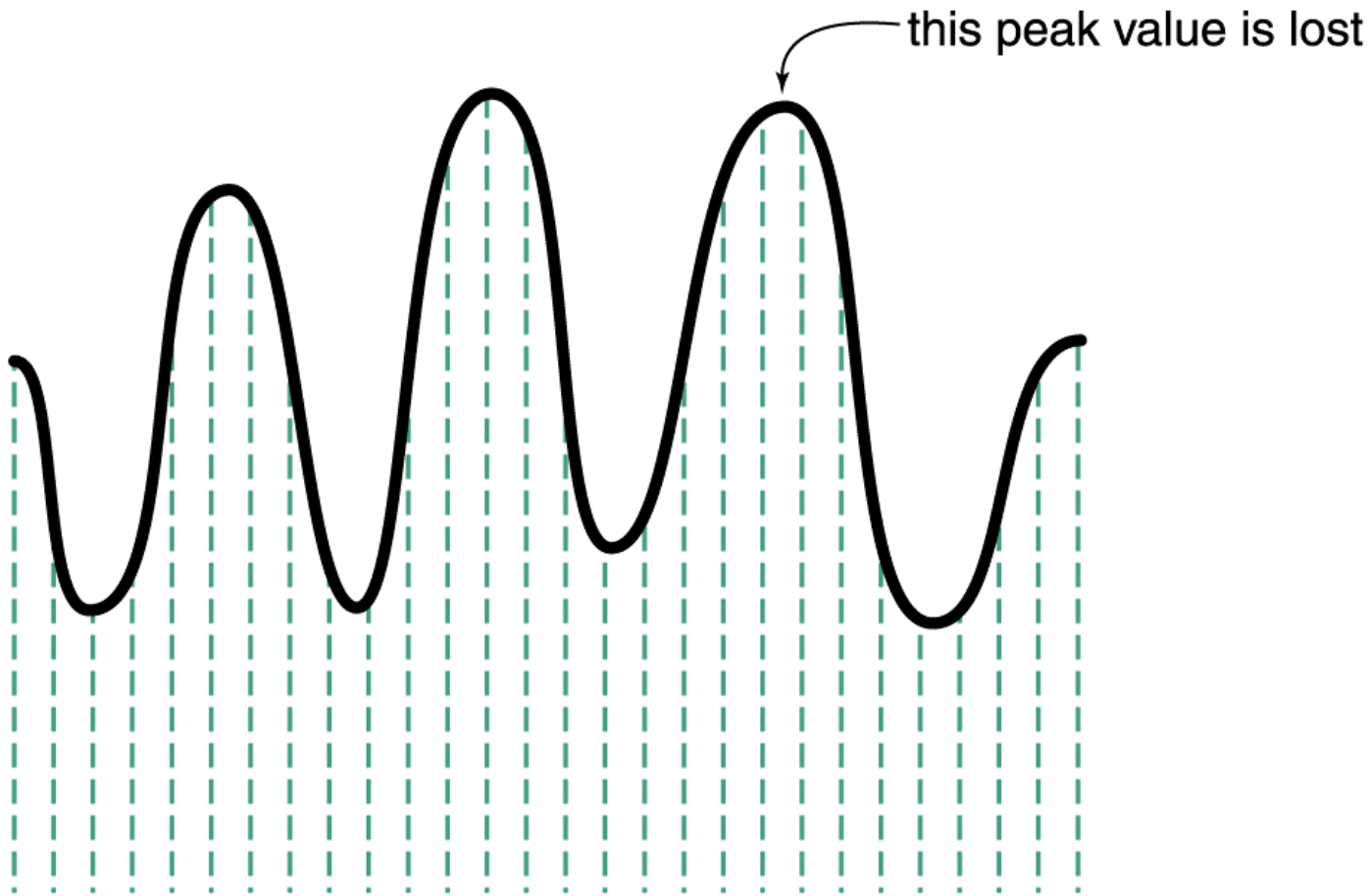

Figure 3.8 Sampling an audio signal

Representing Audio Information
(Cont'd)

To achieve acceptable quality for human perception, there are two issues:

- **Sampling rate** (samples per second)

  Audio CDs: 44100 times per second

- **Amplitudes values** are represented digitally; must decide how many bits to use for range.

  Audio CDs: 16 bits per channel

Audio Formats

- Several popular formats are: WAV, AU, AIFF, VQF, and MP3. Currently, the dominant format for compressing audio data is MP3.
- MP3 is short for MPEG-2, audio layer 3 file.
- Often achieves 12:1 compression ratio

Representing Images and Graphics

- Color is our perception of the various frequencies of light that reach the retinas of our eyes.
- Our retinas have three types of color photoreceptor cone cells that respond to different sets of frequencies. These photoreceptor categories correspond to the colors of red, green, and blue.

Representing Images and Graphics
(Cont'd)

- Color is often expressed in a computer as an RGB (red-green-blue) value, which is actually three numbers that indicate the relative contribution of each of these three primary colors.
- For example, 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
(Cont'd)

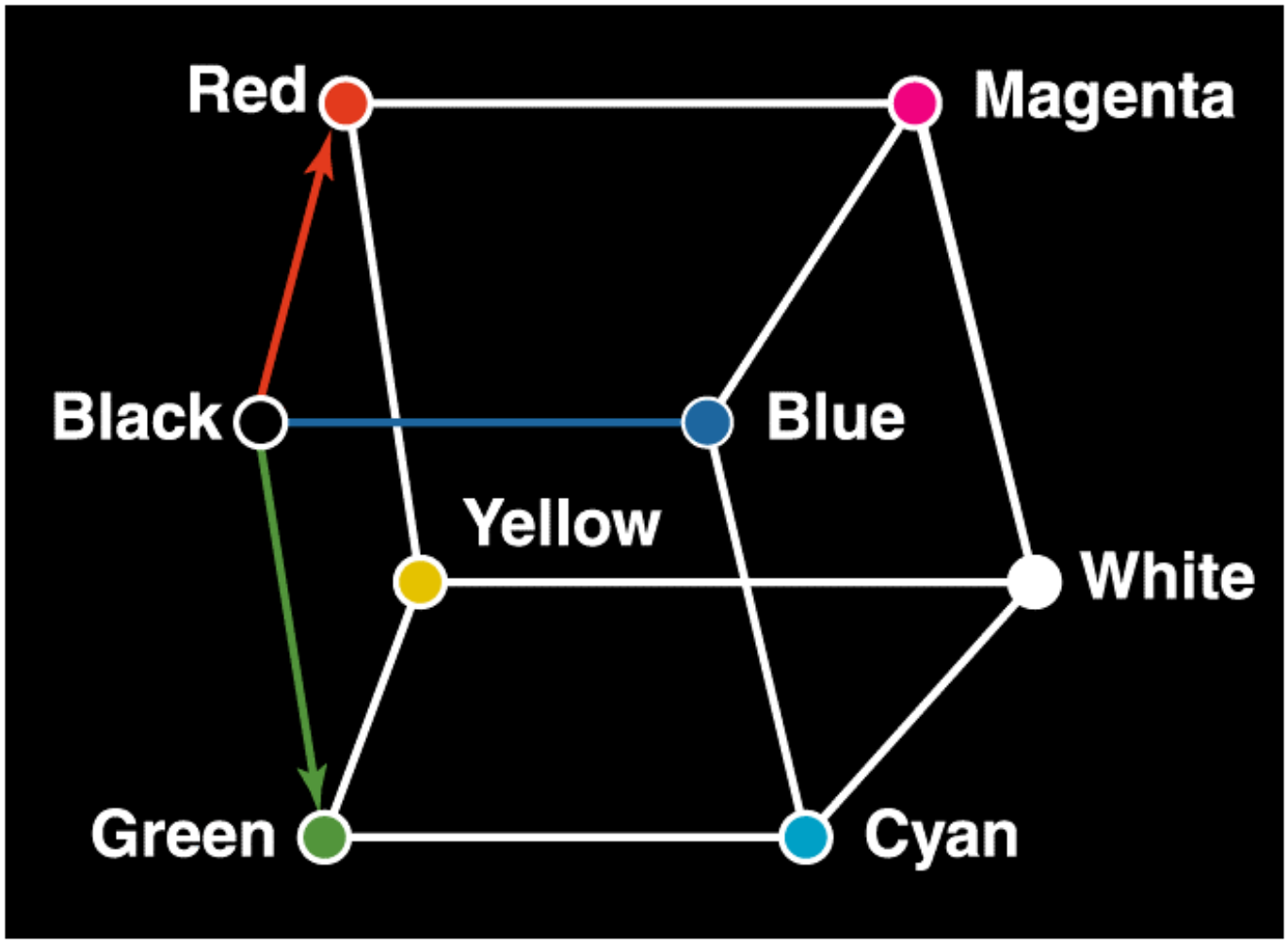

Figure 3.10 Three-dimensional color space

Representing Images and Graphics
(Cont'd)

- The amount of data that is used to represent a color is called the *color depth*.
- *HiColor* is a term that indicates a 16-bit color depth. Five bits are used for each number in an RGB value and the extra bit is sometimes used to represent transparency. *TrueColor* indicates a 24-bit color depth. Therefore, each number in an RGB value gets eight bits.

Representing Images and Graphics
(Cont'd)

RGB Value			Actual Color
Red	Green	Blue	
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

Indexed Color

- A particular application such as a browser may support only a certain number of specific colors, creating a palette from which to choose. For example, the Netscape Navigator's color palette:

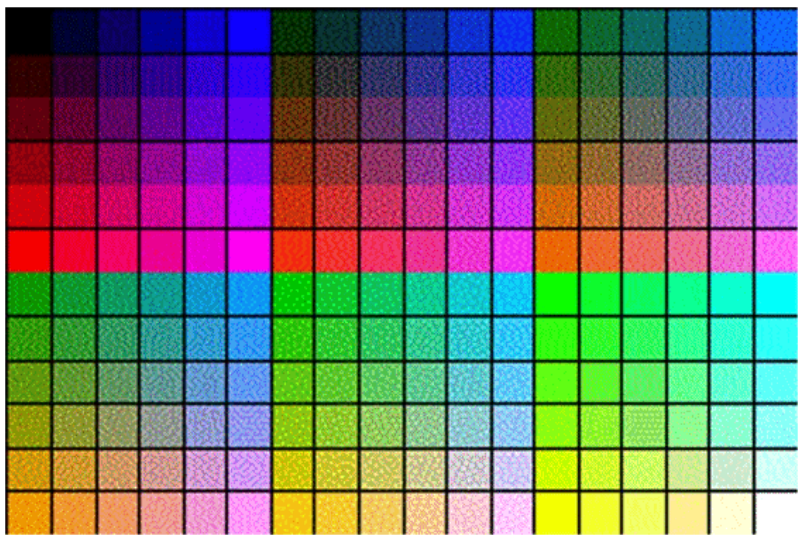

Figure 3.11
The Netscape color palette

Raster-Graphics format

Common method for digitizing a picture

- Represent as collection of individual dots called **pixels**.
- The number of pixels used to represent a picture is called the **resolution**.

Digitized Images and Graphics *(Cont'd)*

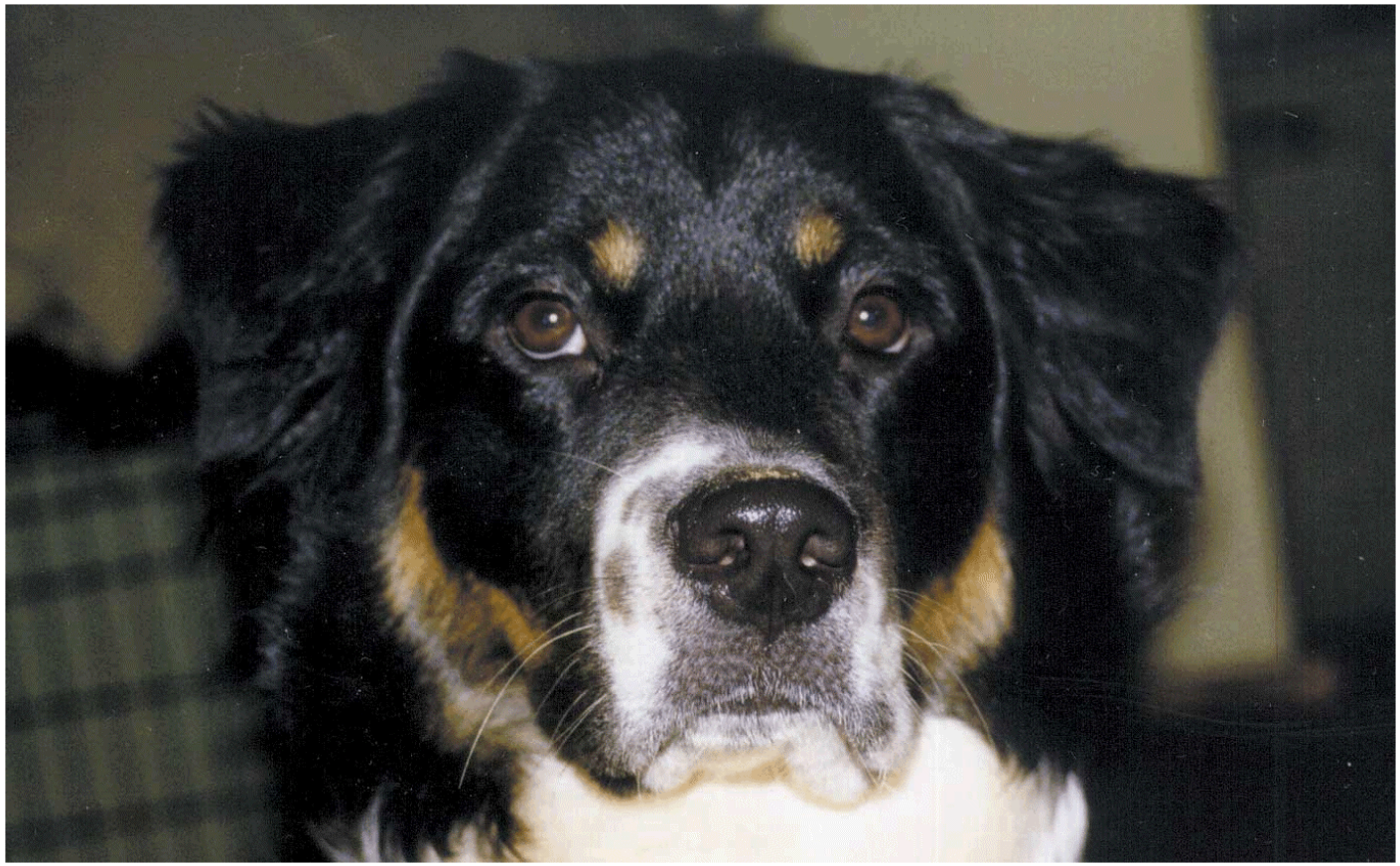

Figure 3.12 A digitized picture composed of many individual pixels

Digitized Images and Graphics *(Cont'd)*

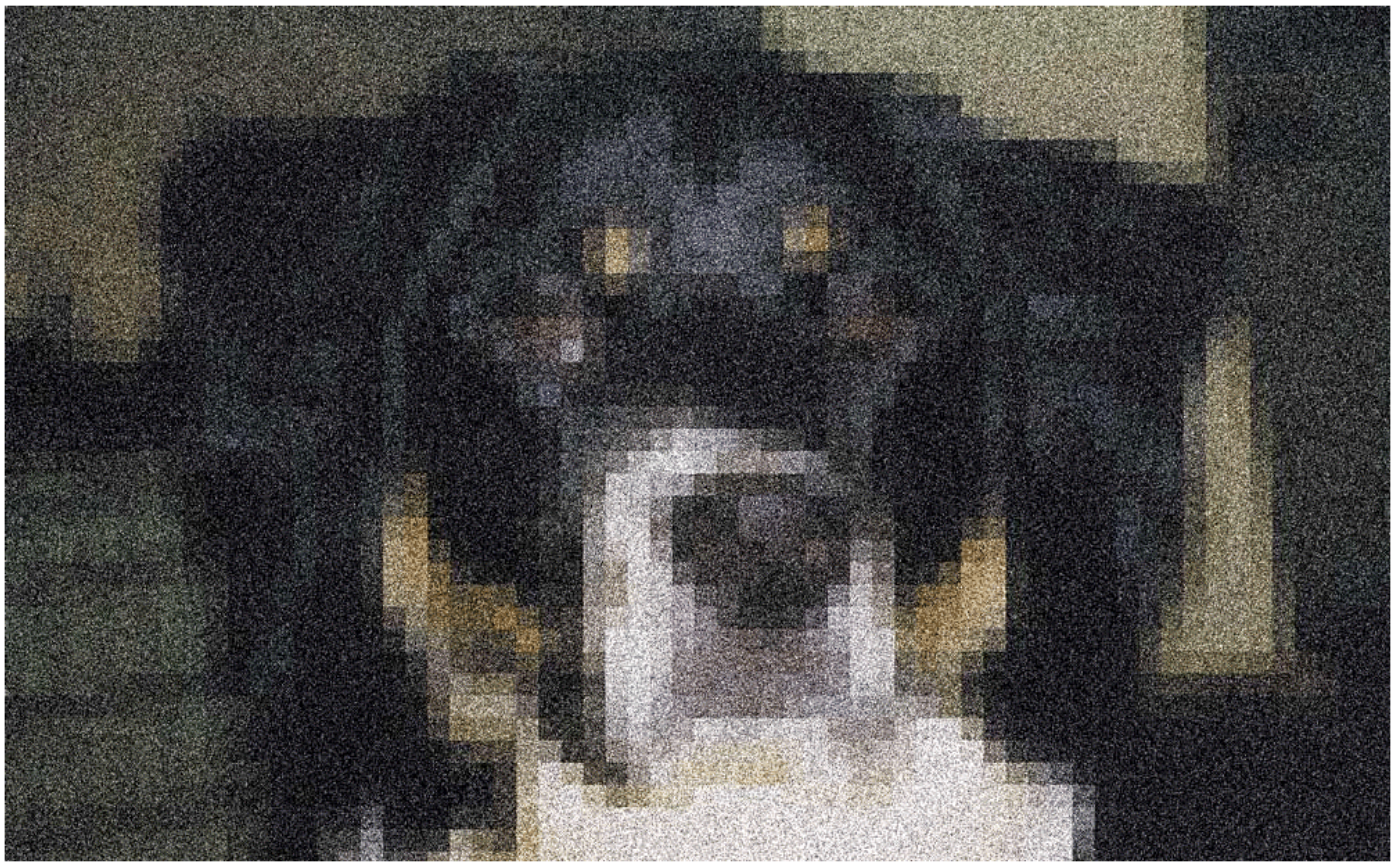

Figure 3.12 A digitized picture composed of many individual pixels

Common Formats

- BMP (primitive format used by Microsoft)
- TIFF (Tagged Image File Format)
- GIF (Graphics Interchange Format)
- JPEG (Joint Photographic Experts Group)

Vector Graphics

- Instead of assigning colors to pixels as we do in raster graphics, a vector-graphics format describe an image in terms of lines and geometric shapes. A vector graphic is a series of commands that describe a line's direction, thickness, and color. The file size for these formats tend to be small because every pixel does not have to be accounted for.

Vector Graphics *(Cont'd)*

- Vector graphics can be resized mathematically, and these changes can be calculated dynamically as needed.
- However, vector graphics is not good for representing real-world images.

- Common Formats:
  - EPS (Encapsulated PostScript)
  - PICT (Macintosh's file format)

Representing Video

- A video codec COmpressor/DECompressor refers to the 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.