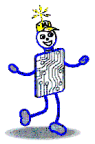


CHAPTER 4: Data Formats

The Architecture of Computer Hardware, Systems Software & Networking:

An Information Technology Approach

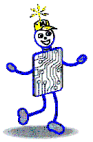
4th Edition, Irv Englander

John Wiley and Sons ©2010

PowerPoint slides authored by Wilson Wong, Bentley University

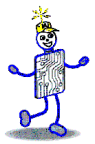
PowerPoint slides for the 3rd edition were co-authored with Lynne Senne, Bentley University

PowerPower slides modified by Gianluca Amato, Univ. di Chieti-Pescara



Data Formats

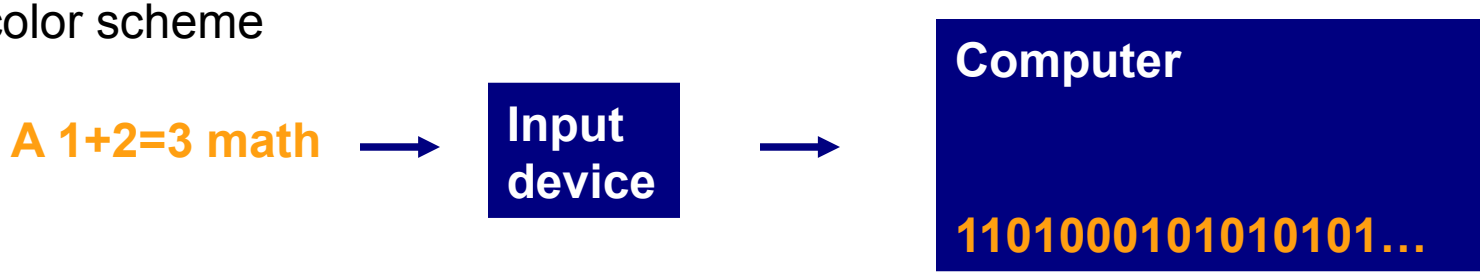
- Computers
  - Process and store all forms of data in binary format
- Human communication
  - Includes language, images and sounds
- Data formats:
  - Specifications for converting data into computer-usable form
  - Define the different ways human data may be represented, stored and processed by a computer

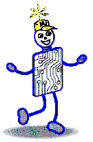


Sources of Data

- Binary/Digital input
  - Begins as discrete input
  - Example: keyboard input such as ***A 1+2=3 math***
  - Keyboard generates a binary number code for each key
- Analog
  - Continuous data such as sound or images
  - Requires hardware to convert data into binary numbers

Figure 3.1 with this color scheme



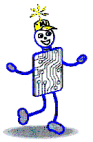


Int. / ext. data representation

- Internal representation
  - Used internally by the program
  - Optimized for easy manipulation
- External representation
  - Used for storage and transmission
    - *Compressio*n represents data in a more compact form
    - *Metadata:* data that describes or interprets the meaning of data
  - Standardization
    - *Proprietary formats* for storing and processing data (WordPerfect vs. Word)
    - De facto standards: proprietary standards based on general user acceptance (PostScript)
    - Official standards

Some External Representations

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC
Image (bitmapped)	▪GIF (graphical image format) ▪TIF (tagged image file format) ▪PNG (portable network graphics)
Image (object)	PostScript, JPEG, SWF (Macromedia Flash), SVG
Outline graphics and fonts	PostScript, TrueType
Sound	WAV, AVI, MP3, MIDI, WMA
Page description	PDF (Adobe Portable Document Format), HTML, XML
Video	Quicktime, MPEG-2, RealVideo, WMV

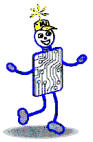


Internal Data Representation

- Reflects the
  - Complexity of input source
  - Type of processing required
- Trade-offs
  - Accuracy and resolution
    - Simple photo vs. painting in an art book
  - Compactness (storage and transmission)
    - More data required for improved accuracy and resolution
    - *Compressio*n represents data in a more compact form
    - *Metadata:* data that describes or interprets the meaning of data
  - Ease of manipulation:
    - Processing simple audio vs. high-fidelity sound
  - Standardization
    - *Proprietary formats* for storing and processing data (WordPerfect vs. Word)
    - De facto standards: proprietary standards based on general user acceptance (PostScript)

Data Types: Numeric

- Used for mathematical manipulation
  - Add, subtract, multiply, divide
- Types
  - Integer (whole number)
  - Real (contains a decimal point)
- Covered in Chapters 4 and 5



Data Types: Alphanumeric

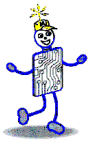
- Alphanumeric:
  - Characters: *b T*
  - Number digits: *7 9*
  - Punctuation marks: *! ;*
  - Special-purpose characters: *$ &*
- Numeric characters vs. numbers
  - Both entered as ordinary characters
  - Computer converts into numbers for calculation
    - Examples: Variables declared as numbers by the programmer
  - Treated as characters if processed as text
    - Examples: Phone numbers, ZIP codes

Alphanumeric Codes

- Arbitrary choice of bits to represent characters
  - Consistency: input and output device must recognize same code
  - Value of binary number representing character corresponds to placement in the alphabet
    - Facilitates sorting and searching

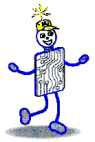
Representing Characters

- ASCII: most widely used coding scheme
- EBCDIC: IBM mainframe (legacy)
- Unicode: developed for worldwide use



ASCII

- Developed by ANSI (American National Standards Institute)
- Represents
  - Latin alphabet, Arabic numerals, standard punctuation characters
  - Plus small set of accents and other European special characters
- ASCII
  - 7-bit code: 128 characters

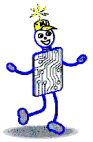


ASCII Reference Table

MSD / LSD	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P		p
1	SOH	DC1	!	1	A	Q	a	W
2	STX	DC2	“	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACJ	SYN	&	6	F	V	f	v
7	BEL	ETB	‘	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	DEL

74_{16}

111 0100



EBCDIC

- Extended Binary Coded Decimal Interchange Code developed by IBM
  - Restricted mainly to IBM or IBM compatible mainframes
  - Conversion software to/from ASCII available
  - Common in archival data
  - Character codes differ from ASCII

	ASCII	EBCDIC
Space	20_{16}	40_{16}
A	41_{16}	$C1_{16}$
b	62_{16}	82_{16}

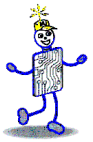
Collating Sequence

- Alphabetic sorting if software handles mixed upper- and lowercase codes
- In ASCII, numbers collate first; in EBCDIC, last
- ASCII collating sequence for string of characters

Letters								Numeric Characters				
Adam	A	d	a	m				1	011	0001		
Adamian	A	d	a	m	i	a	n	12	011	0001	011	0010
Adams	A	d	a	m	s			2	011	0010		

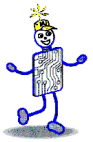
2 Classes of Codes

- *Printing* characters
  - Produced on the screen or printer
- *Control* characters
  - Control position of output on screen or printer
    - VT: vertical tab
    - LF: Line feed
  - Cause action to occur
    - BEL: bell rings
    - DEL: delete current character
  - Communicate status between computer and I/O device
    - ESC: provides extensions by changing the meaning of a specified number of contiguous following characters



Unicode

- 32-bit code
- ASCII Latin-I subset of Unicode
  - Values 0 to 255 in Unicode table
- Multilingual: defines codes for
  - Nearly every character-based alphabet
  - Large set of ideographs for Chinese, Japanese and Korean
  - Composite characters for vowels and syllabic clusters required by some languages
- Allows software modifications for local-languages



Unicode

- Several representations
  - UCS-4
  - UTF-16
  - UTF-8

Keyboard Input

- *Scan code*
  - Two different scan codes on keyboard
    - One generated when key is struck and another when key is released
  - Converted to Unicode, ASCII or EBCDIC by software in terminal or PC
- Advantage
  - Easily adapted to different languages or keyboard layout
  - Separate scan codes for key press/release for multiple key combinations
    - Examples: shift and control keys

Other Alphanumeric Input

- *OCR* (optical character reader)
  - Scans text and inputs it as character data
  - Used to read specially encoded characters
    - Example: magnetically printed check numbers
- *Bar Code Readers*
  - Used in applications that require fast, accurate and repetitive input with minimal employee training
  - Examples: supermarket checkout counters and inventory control
- *Magnetic stripe reader:* alphanumeric data from credit cards
- *RFID:* store and transmit data between RFID tags and computers
- *Voice*
  - Digitized audio recording common but conversion to alphanumeric data difficult
  - Requires knowledge of sound patterns in a language (*phonemes*) plus rules for pronunciation, grammar, and syntax

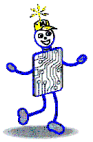
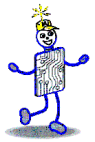


Image Data

- Photographs, figures, icons, drawings, charts and graphs
- Two approaches:
  - *Bitmap* or *raster images* of photos and paintings with continuous variation
  - *Object* or *vector images* composed of *graphical objects* like lines and curves defined geometrically
- Differences include:
  - Quality of the image
  - Storage space required
  - Time to transmit
  - Ease of modification

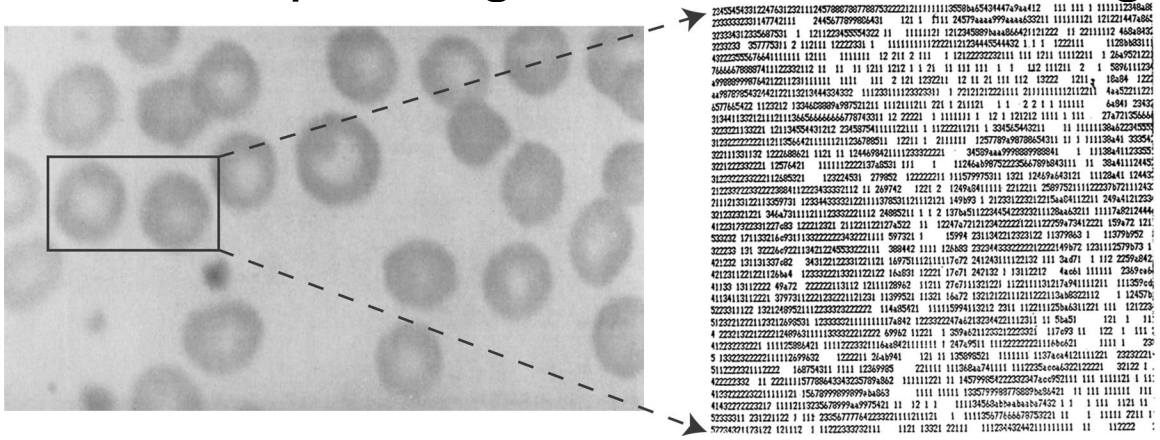
Bitmap Images

- Used for realistic images with continuous variations in shading, color, shape and texture
  - Examples:
    - Scanned photos
    - Clip art generated by a *paint* program
- Preferred when image contains large amount of detail and processing requirements are fairly simple
- Input devices:
  - Scanners
  - Digital cameras and video capture devices
  - Graphical input devices like mice and pens
- Managed by *photo editing software* or *paint software*
  - Editing tools to make tedious bit by bit process easier



Bitmap Images

- Each individual *pixel* (*pi(x)*cture *el*ement) in a graphic stored as a binary number
  - Pixel: A small area with associated coordinate location
  - Example: each point below represented by a 4-bit code corresponding to 1 of 16 shades of gray



Bitmap Display

- Monochrome: black or white
  - 1 bit per pixel
- Gray scale: black, white or 254 shades of gray
  - 1 byte per pixel
- Color graphics: 16 colors, 256 colors, or 24-bit true color (16.7 million colors)
  - 4, 8, and 24 bits respectively

Storing Bitmap Images

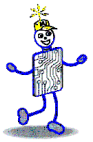
- Frequently large files
  - Example: 600 rows of 800 pixels with 1 byte for each of 3 colors ➡ ~1.5MB file
- File size affected by
  - *Resolution* (the number of pixels per inch)
    - Amount of detail affecting clarity and sharpness of an image
  - Levels: number of bits for displaying shades of gray or multiple colors
    - *Palette*: color translation table that uses a code for each pixel rather than actual color value
  - Data compression

Data Compression

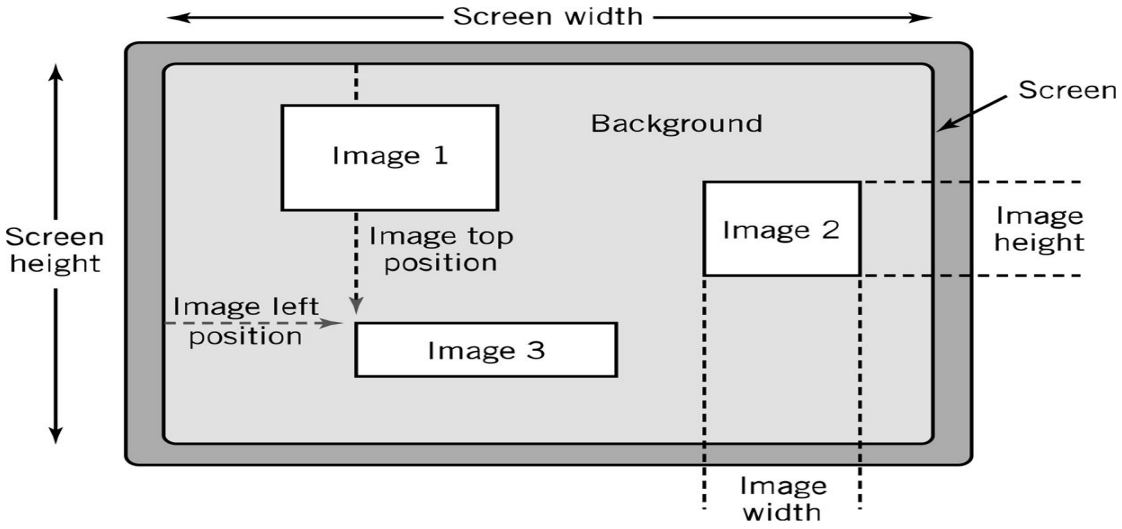
- *Compression*: recoding data so that it requires fewer bytes of storage space.
- *Compression ratio*: the amount file is shrunk
- *Lossless*: inverse algorithm restores data to exact original form
  - Examples: GIF, PCX, TIFF
- *Lossy*: trades off data degradation for file size and download speed
  - Much higher compression ratios, often 10 to 1
  - Example: JPEG
  - Common in multimedia
- MPEG-2: uses both forms for ratios of 100:1

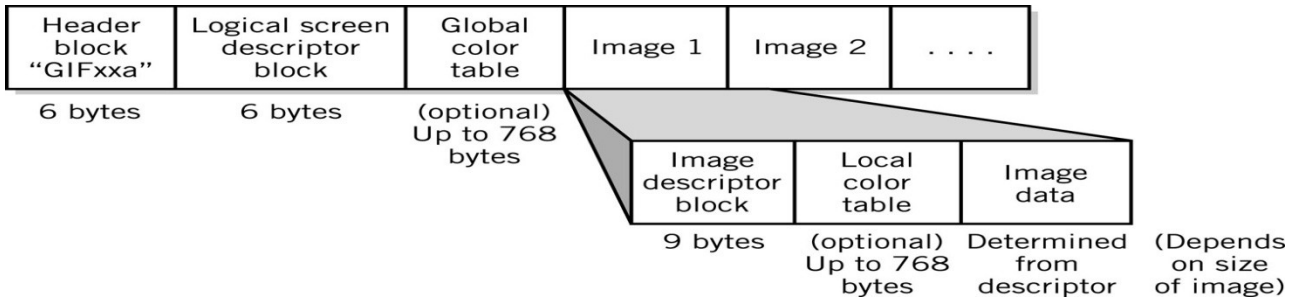
GIF (Graphics Interchange Format)

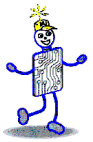
- First developed by CompuServe in 1987
- GIF89a enabled animated images
  - allows images to be displayed sequentially at fixed time sequences
- Color limitation: 256
- Image compressed by LZW (Lempel-Zif-Welch) algorithm
- Preferred for line drawings, clip art and pictures with large blocks of solid color
- *Lossless compression*



GIF (Graphics Interchange Format)





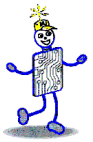


JPEG (Joint Photographers Expert Group)

- Allows more than 16 million colors
- Suitable for highly detailed photographs and paintings
- Employs *lossy compression* algorithm that
  - Discards data to decreases file size and transmission speed
  - May reduce image resolution, tends to distort sharp lines

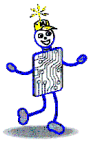
Object Images

- Created by *drawing* packages or output from spreadsheet data graphs
- Composed of lines and shapes in various colors
- Computer translates geometric formulas to create the graphic
- Storage space depends on image complexity
  - number of instructions to create lines, shapes, fill patterns
- Movies *Shrek* and *Toy Story* use object images



Object Images

- Based on mathematical formulas
  - Easy to move, scale and rotate without losing shape and identity as bitmap images may
- Require less storage space than bitmap images
- Cannot represent photos or paintings
- Cannot be displayed or printed directly
  - Must be converted to bitmap since output devices except plotters are bitmap



SVG

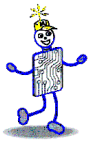
- SVG (Scalable Vector Graphics) is a W3C standard for object images
  - XML based
  - Animations
  - Supported by web browsers

Bitmap vs. Object Images

Bitmap (Raster)	Object (Vector)
Pixel map	Geometrically defined shapes
Photographic quality	Complex drawings
Paint software	Drawing software
Larger storage requirements	Higher computational requirements
Enlarging images produces jagged edges	Objects scale smoothly
Resolution of output limited by resolution of image	Resolution of output limited by output device

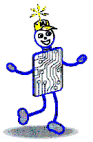
Video Images

- Require massive amount of data
  - Video camera producing full screen 640 x 480 pixel true color image at 30 frames/sec → 27.65 MB of data/sec
  - 1-minute film clip → 1.6 GB storage
- Options for reducing file size: decrease size of image, limit number of colors, reduce frame rate
- Method depends on how video delivered to users
  - *Streaming video*: video displayed as it is downloaded from the Web server
  - Local data (file on DVD or downloaded onto system) for higher quality
    - MPEG-2: movie quality images with high compression require substantial processing capability

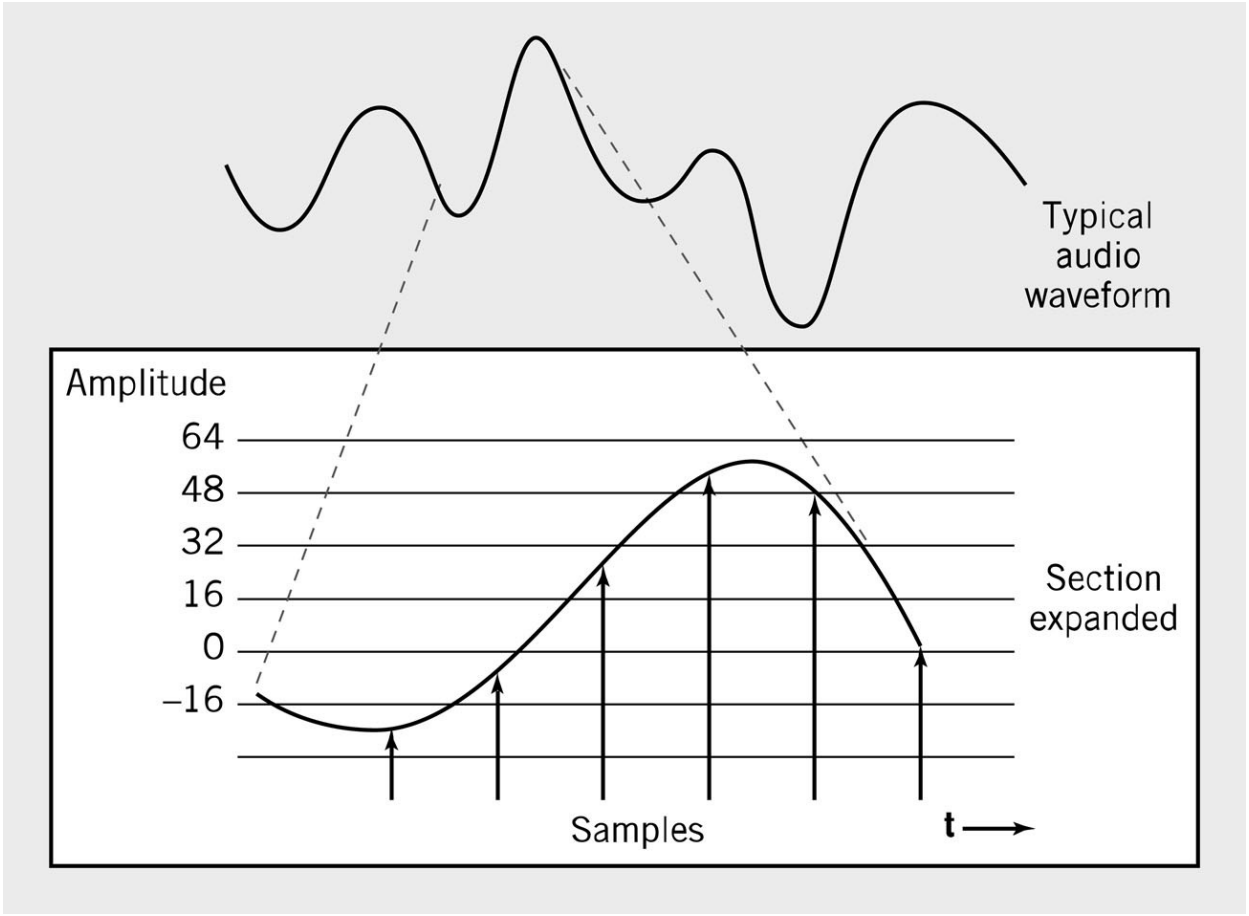


Audio Data

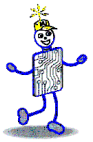
- Transmission and processing requirements less demanding than those for video
- *Waveform audio*: digital representation of sound
- *MIDI* (Musical Instrument Digital Interface): instructions to recreate or synthesize sounds
- Analog sound converted to digital values by *A-to-D converter*



Waveform Audio

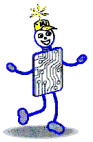


Sampling rate normally 50KHz



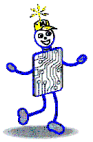
Sampling Rate

- Number of times per second that sound is measured during the recording process.
  - 1000 samples per second = 1 KHz (kilohertz)
  - Example: Audio CD sampling rate = 44.1KHz
- Height of each sample saved as:
  - 8-bit number for radio-quality recordings
  - 16-bit number for high-fidelity recordings
  - 2 x 16-bits for stereo

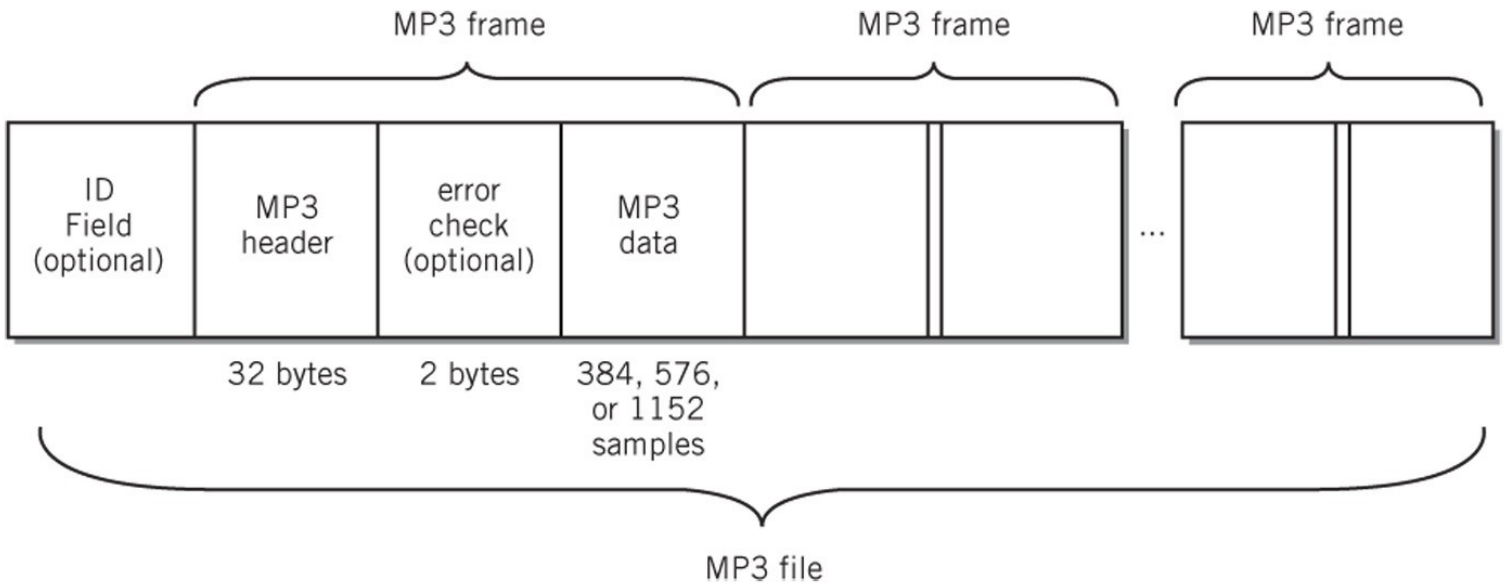


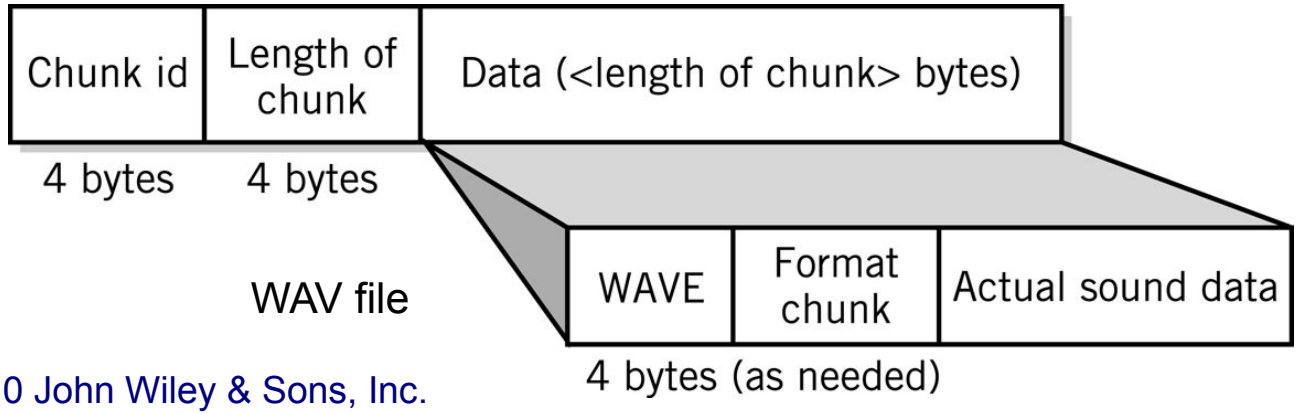
Audio Formats

- *MP3*
  - Derivative of MPEG-2 (ISO *M*oving *P*icture *E*xperts *G*roup)
  - Uses psychoacoustic compression techniques to reduce storage requirements
- *WAV*
  - Developed by Microsoft as part of its multimedia specification
  - General-purpose format for storing and reproducing small snippets of sound



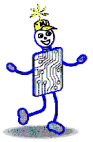
Audio Data Formats





Page Description Languages

- Describe layout of objects on a displayed or printed page
- Objects may include text, object images, bitmap images, multimedia objects, and other data formats
- Examples
  - HTML, XHTML, XML
  - PDF
  - Postscript



PostScript

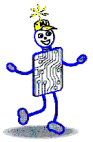
- *Page description language*: list of procedures and statements that describe each of the objects to be printed on a page
  - Stored in ASCII or Unicode text file
  - Interpreter program in computer or output device reads PostScript to generate image
- Scalable font support
  - Font outline objects specified like other objects

Internal Computer Data Format

- All data stored as binary numbers
- Interpreted based on
  - Operations computer can perform
  - Data types supported by programming language used to create application

5 Simple Data Types

- Boolean: 2-valued variables or constants with values of true or false
- Char: Variable or constant that holds alphanumeric character
- Enumerated
  - User-defined data types with possible values listed in definition
    - Type DayOfWeek = Mon, Tues, Wed, Thurs, Fri, Sat, Sun
- Integer: positive or negative whole numbers
- Real
  - Numbers with a decimal point
  - Numbers whose magnitude, large or small, exceeds computer's capability to store as an integer



Copyright 2010 John Wiley & Sons