



COMPUTER CODES

CHAPTER 02

DATA FORMATS

Computers

- Process and store all forms of data in binary format

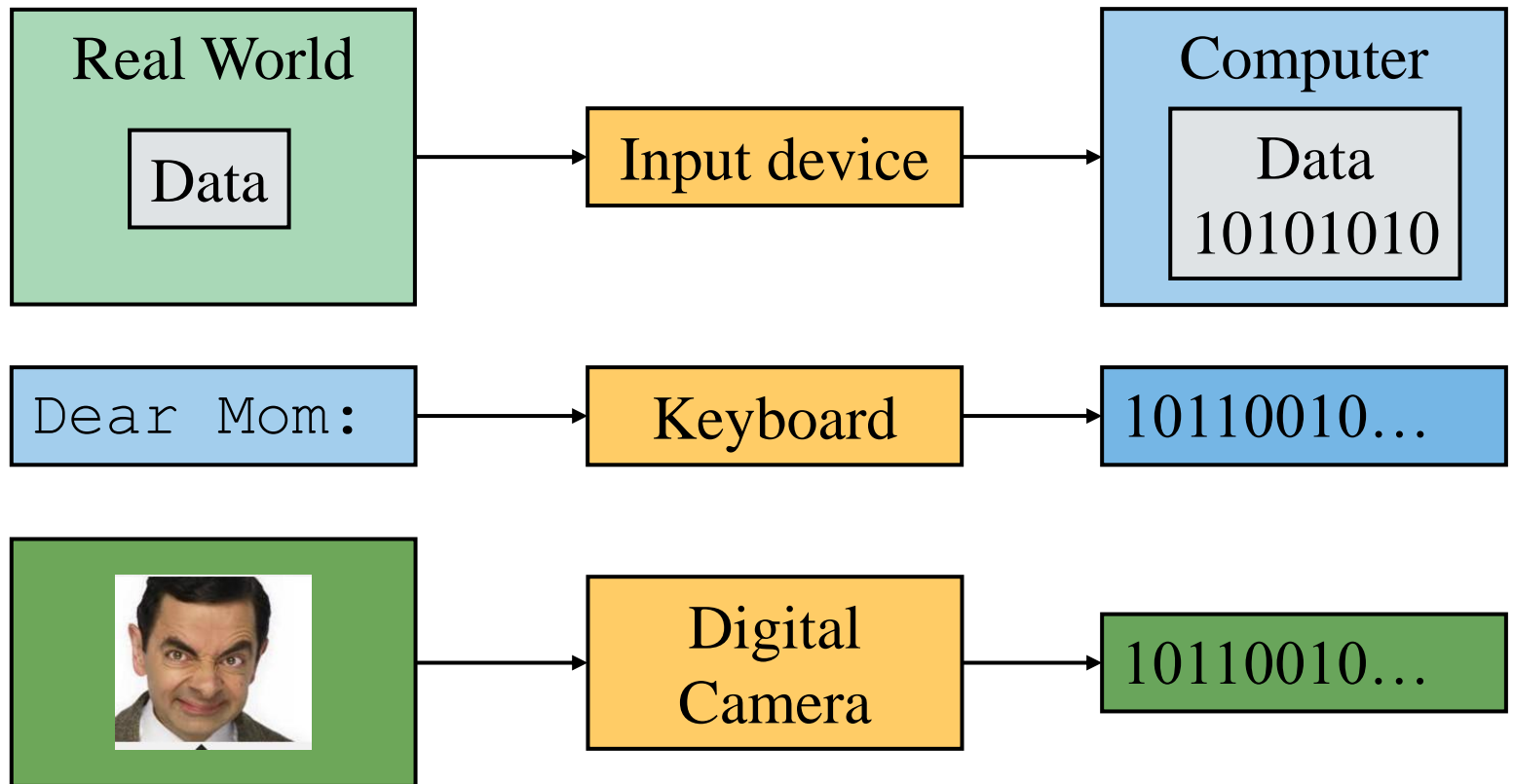
Human communication

- 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

INTRODUCTION



SOURCES OF DATA

Binary 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

A 1+2=3 math



Input
device



Computer

1101000101010101...

STANDARDS ORGANIZATIONS

ISO – International Standards Organization

CSA – Canadian Standards Association

ANSI – American National Standards Institute

IEEE – Institute for Electrical and Electronics Engineers

COMMON DATA REPRESENTATIONS

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC,BCD
Image (bitmapped)	<ul style="list-style-type: none">■ GIF (graphical image format)■ TIF (tagged image file format)■ PNG (portable network graphics)
Sound	WAV, AVI, MP3, MIDI, WMA
Video(Motion Picture)	Quicktime, MPEG-2, RealVideo, WMV

WHY STANDARDS?

Standard are “arbitrary”

They exist because they are

- Convenient
- Efficient
- Flexible
- Appropriate
- Etc.

DATA TYPES

- **Numeric Data** consists of only numbers 0, 1, 2, ..., 9
- **Alphabetic Data** consists of only the letters A, B, C, ..., Z, in both uppercase and lowercase, and blank character
- **Alphanumeric Data** is a string of symbols where a symbol may be one of the letters A, B, C, ..., Z, in either uppercase or lowercase, or one of the digits 0, 1, 2, ..., 9, or a special character, such as + - * / , . () = etc.

DATA TYPES: NUMERIC

Used for mathematical manipulation

- Add, subtract, multiply, divide

Types

- Integer (whole number)
- Real (contains a decimal point)

DATA TYPES: ALPHANUMERIC

Alphanumeric:

- Characters: *b T*
- Number digits: *7 9*
- Punctuation marks: *! ;*
- Special-purpose characters: *\$ &*

Four standards for representing letters (alpha) and numbers

- BCD – Binary-coded decimal
- EBCDIC – Extended binary-coded decimal interchange code
- ASCII – American standard code for information interchange
- Unicode

COMPUTER CODES

- ❑ A computer is a digital system that stores and processes different types of data in the form of 0s and 1s.
- ❑ The different types of data handled by a computer system include numbers, alphabets and some special characters.
- ❑ Therefore, there is a need to change the data entered by the users into a form that the computer system can understand and process.
- ❑ Different types of codes have been developed and used to represent the data entered by the users in the binary format.
- ❑ The binary system represents each type of data in terms of binary digits, 0s and 1s.
- ❑ Since these codes convert the data into the binary form, the computer codes are also referred as binary codes.

BINARY SYSTEM TERMS

The following are some of the technical terms used in binary system:

Bit: It is the smallest unit of information used in a computer system. It can either have the value 0 or 1. Derived from the words *Binary* and *IT*.

Nibble: It is a combination of 4 bits.

Byte: It is a combination of 8 bits.

Word: It is a combination of 16 bits.

Double word: It is a combination of 32 bits.

Kilobyte (KB): It is used to represent the 1024 bytes of information.

Megabyte (MB): It is used to represent the 1024 KBs of information.

Gigabyte (GB): It is used to represent the 1024 MBs of information.



BINARY CODED DECIMAL (BCD) SYSTEMS



4-BIT BINARY CODED DECIMAL (BCD) SYSTEMS

4-BIT BINARY CODED DECIMAL (BCD) SYSTEMS

The BCD system is employed by computer systems to encode the decimal number into its equivalent binary number.

This is generally accomplished by encoding each digit of the decimal number into its equivalent binary sequence.

The main advantage of BCD system is that it is a fast and efficient system to convert the decimal numbers into binary numbers as compared to the pure binary system.


4-BIT BINARY CODED DECIMAL (BCD) SYSTEMS

The 4-bit BCD system is usually employed by the computer systems to represent and process numerical data only. In the 4-bit BCD system, each digit of the decimal number is encoded to its corresponding 4-bit binary sequence. The two most popular 4-bit BCD systems are:

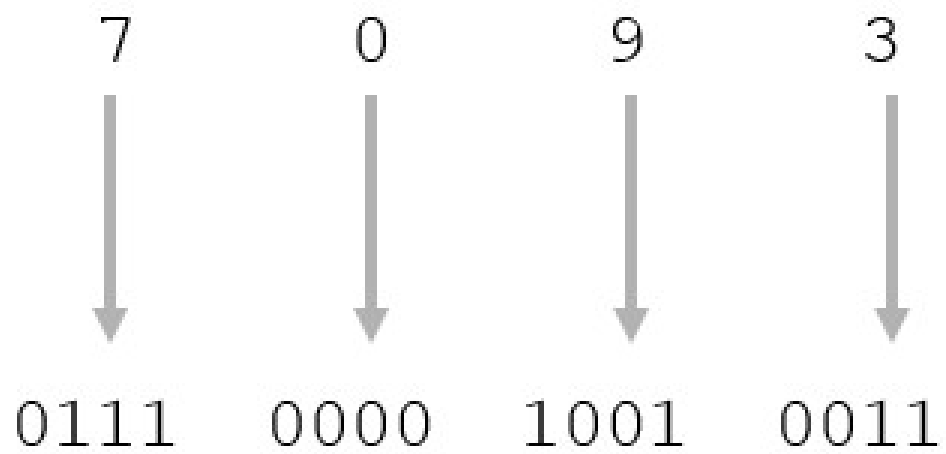
- Weighted 4-bit BCD code
- Excess-3 (XS-3) BCD code

WEIGHTED 4-BIT BCD CODE

- ❑ The weighted 4-bit BCD code is more commonly known as 8421 weighted code.
- ❑ It is called weighted code because it encodes the decimal system into binary system by using the concept of positional weighting into consideration.
- ❑ In this code, each decimal digit is encoded into its 4-bit binary number in which the bits from left to right have the weights 8, 4, 2, and 1, respectively.



$7093_{10} = ? \text{ (in BCD)}$



WEIGHTED 4-BIT BCD CODE

Note: the following
bit patterns are not
used:

1010

1011

1100

1101

1110

1111

Decimal digit	BCD			
	8	4	2	1
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1



8-BIT BINARY CODED DECIMAL (BCD) SYSTEMS

8-BIT BCD SYSTEMS

- ❑ The 6-bit BCD systems can handle numeric as well as non-numeric data but with few special characters.
- ❑ The 8-bit BCD systems were developed to overcome the limitations of 6-bit BCD systems, which can handle numeric as well as nonnumeric data with almost all the special characters such as +, -, *, /, @, \$, etc.
- ❑ Therefore, the various codes under the category of 8-bit BCD systems are also known as *alphanumeric codes*.

8-BIT BCD SYSTEMS

The three most popular 8-bit BCD codes are:

- Extended Binary Coded Decimal Interchange Code (EBCDIC)
- American Standard Code for Information Interchange (ASCII)
- Gray Code

EBCDIC CODE

- ❑ The EBCDIC code is an 8-bit alphanumeric code that was developed by IBM to represent alphabets, decimal digits and special characters, including control characters.
- ❑ The EBCDIC codes are generally the decimal and the hexadecimal representation of different characters.
- ❑ This code is rarely used by non IBM-compatible computer systems.

EBCDIC CONT

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

2nd hex digit

EBCDIC character codes

1st hex digit

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	DS		SP	&	-									0
1	SOH	DC1	SOS				/		a	j			A	J		1
2	STX	DC2	FS	SYN					b	k	s		B	K	S	2
3	ETX	TM							c	l	t		C	L	T	3
4	PF	RES	BYP	PN					d	m	u		D	M	U	4
5	HT	NL	LF	RS					e	n	v		E	N	V	5
6	LC	BS	ETB	UC					f	o	w		F	O	W	6
7	DEL	IL	ESC	EOT					g	p	x		G	P	X	7
8		CAN							h	q	y		H	Q	Y	8
9		EM							i	r	z	`	I	R	Z	9
A	SMM	CC	SM		C CENT	!		:								
B	VT	CU1	CU2	CU3		\$,	#								
C	FF	IFS		DC4	<	*	%	@								
D	CR	IGS	ENQ	NAK	()	_	'								
E	SO	IRS	ACK		+	:	>	=								
F	SI	IUS	BEL	SUB		--	?	"								

Char	EBCDIC Code		Hex
	Digit	Zone	
A	1100	0001	C1
B	1100	0010	C2
C	1100	0011	C3
D	1100	0100	C4
E	1100	0101	C5
F	1100	0110	C6
G	1100	0111	C7
H	1100	1000	C8
I	1100	1001	C9
J	1101	0001	D1
K	1101	0010	D2
L	1101	0011	D3
M	1101	0100	D4

Char	EBCDIC Code		Hex
	Digit	Zone	
N	1101	0101	D5
O	1101	0110	D6
P	1101	0111	D7
Q	1101	1000	D8
R	1101	1001	D9
S	1110	0010	E2
T	1110	0011	E3
U	1110	0100	E4
V	1110	0101	E5
W	1110	0110	E6
X	1110	0111	E7
Y	1110	1000	E8
Z	1110	1001	E9

Character	EBCDIC Code		Hexadecimal Equivalent
	Digit	Zone	
0	1111	0000	F0
1	1111	0001	F1
2	1111	0010	F2
3	1111	0011	F3
4	1111	0100	F4
5	1111	0101	F5
6	1111	0110	F6
7	1111	0111	F7
8	1111	1000	F8
9	1111	1001	F9

ASCII CODE

- ❑ The ASCII code is pronounced as ASKEE and is used for the same purpose for which the EBCDIC code is used. However, this code is more popular than EBCDIC code as unlike the EBCDIC code this code can be implemented by most of the non-IBM computer systems.
- ❑ Initially, this code was developed as a 7-bit BCD code to handle 128 characters but later it was modified to an 8-bit code.

ASCII

- ASCII stands for **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange.
- ASCII is of two types – ASCII-7 and ASCII-8
- ASCII-7 uses 7 bits to represent a symbol and can represent 128 (2^7) different characters
- ASCII-8 uses 8 bits to represent a symbol and can represent 256 (2^8) different characters
- First 128 characters in ASCII-7 and ASCII-8 are same

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

Character	ASCII-7 / ASCII-8		Hexadecimal Equivalent
	Zone	Digit	
0	0011	0000	30
1	0011	0001	31
2	0011	0010	32
3	0011	0011	33
4	0011	0100	34
5	0011	0101	35
6	0011	0110	36
7	0011	0111	37
8	0011	1000	38
9	0011	1001	39

Character	ASCII-7 / ASCII-8		Hexadecimal Equivalent
	Zone	Digit	
A	0100	0001	41
B	0100	0010	42
C	0100	0011	43
D	0100	0100	44
E	0100	0101	45
F	0100	0110	46
G	0100	0111	47
H	0100	1000	48
I	0100	1001	49
J	0100	1010	4A
K	0100	1011	4B
L	0100	1100	4C
M	0100	1101	4D

Character	ASCII-7 / ASCII-8		Hexadecimal Equivalent
	Zone	Digit	
N	0100	1110	4E
O	0100	1111	4F
P	0101	0000	50
Q	0101	0001	51
R	0101	0010	52
S	0101	0011	53
T	0101	0100	54
U	0101	0101	55
V	0101	0110	56
W	0101	0111	57
X	0101	1000	58
Y	0101	1001	59
Z	0101	1010	5A

ASCII REFERENCE TABLE

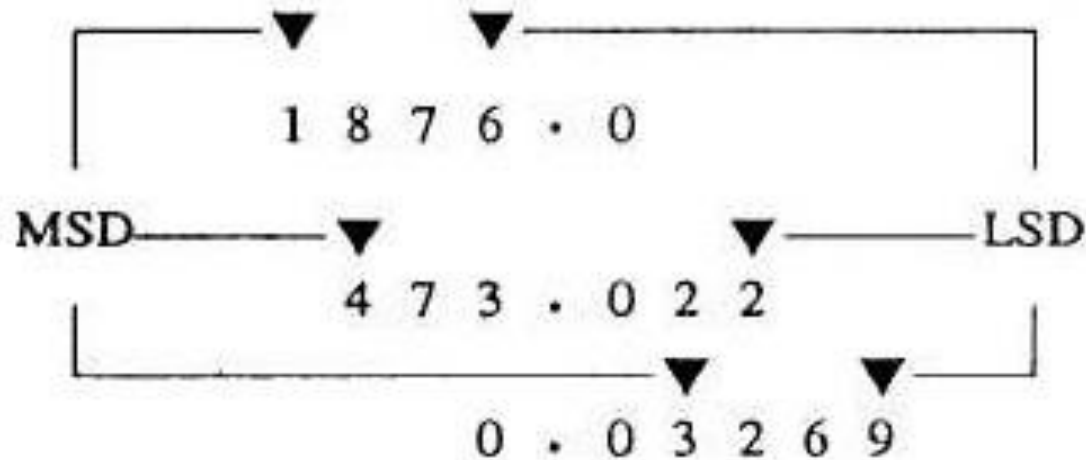
MSD \ LSD	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P		p
1	SOH	DC1	!	1	A	Q	a	W
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	\$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACJ	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(8	H	X	h	x
9	HT	EM)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[k	{
C	FF	FS	,	<	L	\	l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

74₁₆

111 0100

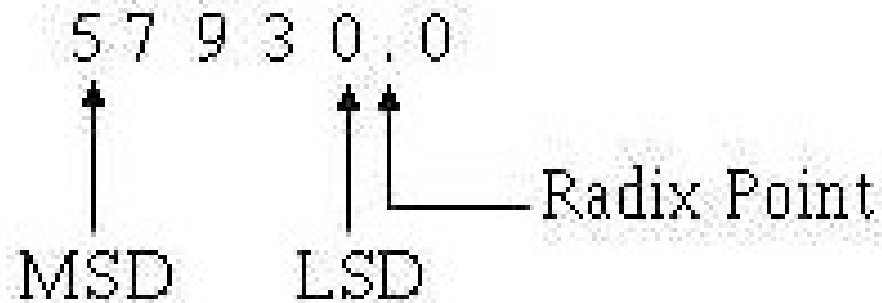
	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1		1	A	Q	a	q
0010	STX	DC2		2	B	R	b	r
0011	ETX	DC3				S	c	s
0100	EDT	DC4				T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
			+	;	K	[k	{
			,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

WHAT IS MSD & LSD?



You can easily see that a change in the MSD will increase or decrease the value of the number the greatest amount. Changes in the LSD will have the smallest effect on the value.

RADIX POINT



In a whole number the LSD will always be the digit immediately to the left of the radix point.

e.g., 'a' = 1100001

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

95 Graphic codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

33 Control codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Alphabetic codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Numeric codes

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Punctuation, etc.

	000	001	010	011	100	101	110	111
0000	NULL	DLE		0	@	P	`	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EDT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	i	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	,	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL



String	H	e	l	l	o
ASCII value	72	101	108	108	111
Binary	01001000	01100101	01101100	01101100	01101111

HOW ?

中文

Chinese

هَيَبَرَعْلَا

Arabic

кириллица

Cyrillic

देवनागरी

Devanagari

UNICODE

Most common 16-bit form represents 65,536 characters

Encoding Forms : UTF-8 , UTF-16 , UTF-32

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

d

Latin Small Letter D

100

あ

Hiragana Letter A

12354

é

Latin Small Letter E with Acute

233

↑
also expressible via
combining modifier
↓

व

Devanagari Letter Va

2357

ŷ

Cyrillic Small Letter Short U

1118



Thumbs Up Sign

128077

你

CJK Unified Ideograph-4F60

20320



Thumbs Up
Sign

128077

Emoji Modifier
+ Fitzpatrick
Type-5

127998

UTF - 8

UTF-8 (8-bit Unicode Transformation Format) is a fixed-length encoding used to encode Unicode code points that uses exactly 8 bits (ONE byte) per code point

UTF - 32

UTF-32 (32-bit Unicode Transformation Format) is a fixed-length encoding used to encode Unicode code points that uses exactly 32 bits (four bytes) per code point

UTF – 32

Hello world!

ASCII 48 65 6C 6C 6F 20 77 6F 72 6C 64 21

UTF-32 00 00 00 48 00 00 00 65 00 00 00 6C
00 00 00 6C 00 00 00 6F 00 00 00 20
00 00 00 77 00 00 00 6F 00 00 00 72
00 00 00 6C 00 00 00 64 00 00 00 21

Parameter	ASCII	UNICODE
Full form	ASCII stands for American Standard Code for Information Interchange.	UNICODE stands for Universal Character Set.
Mutual Relationship	ASCII is a subset of UNICODE encoding scheme.	UNICODE is a superset of ASCII.
Supporting Characters	ASCII supports only 128 characters using 7-bit encoding scheme. It contains codes representing English characters, digits, and standard special symbols.	UNICODE supports a wide range of characters. It supports 154 written scripts.
Bits Per Character	ASCII uses 7-bit or 8-bits (Extended ASCII) to represent different characters.	UNICODE uses mainly four character encoding schemes namely UTF-7 (7-bit), UTF-8 (8-bit), UTF-16 (16-bit), and UTF-32 (32-bit).
Memory Consumption	ASCII consumes less memory.	UNICODE consumes more memory as compared to ASCII.
Characters Represented	ASCII can represent only English letters, digits, certain mathematical symbols, and some grammatical symbols, etc.	UNICODE can represent a large range characters, special symbols, formulae, etc. from different languages such as English, Latin, Greek, etc.
First Edition Release	The first edition of ASCII was released in 1963.	The first edition of UNICODE was released in 1991.

COLLATING SEQUENCE

- Collating sequence defines the assigned ordering among the characters used by a computer
- Collating sequence may vary, depending on the type of computer code used by a particular computer
- In most computers, collating sequences follow the following rules:
 1. Letters are considered in alphabetic order
($A < B < C \dots < Z$)
 2. Digits are considered in numeric order
($0 < 1 < 2 \dots < 9$)

SORTING IN EBCDIC

Suppose a computer uses EBCDIC as its internal representation of characters. In which order will this computer sort the strings 23, A1, 1A?

In EBCDIC, numeric characters are treated to be greater than alphabetic characters. Hence, in the said computer, numeric characters will be placed after alphabetic characters and the given string will be treated as:

$A1 < 1A < 23$

Therefore, the sorted sequence will be: A1, 1A, 23.

SORTING IN ASCII

Suppose a computer uses ASCII for its internal representation of characters. In which order will this computer sort the strings 23, A1, 1A, a2, 2a, aA, and Aa?

In ASCII, numeric characters are treated to be less than alphabetic characters. Hence, in the said computer, numeric characters will be placed before alphabetic characters and the given string will be treated as:

$$1A < 23 < 2a < A1 < Aa < a2 < aA$$

Therefore, the sorted sequence will be: 1A, 23, 2a, A1, Aa, a2, and aA