Representing Real Numbers

A real value in base 10 can be defined by the following formula

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

The representation is called **floating point** because the number of digits is fixed but the radix point floats

Representing Real Numbers

5 digit mantissa examples

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	$15556 * 10^{4}$

A binary floating-point value is defined by the formula

sign * mantissa * 2^{exp}

Representing Text

What must be provided to represent text?

There is a finite number of characters to represent, 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…

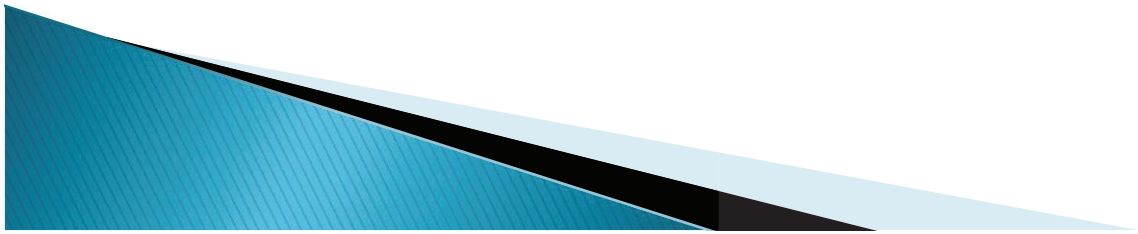
The ASCII Character Set

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?



The ASCII Character Set

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 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

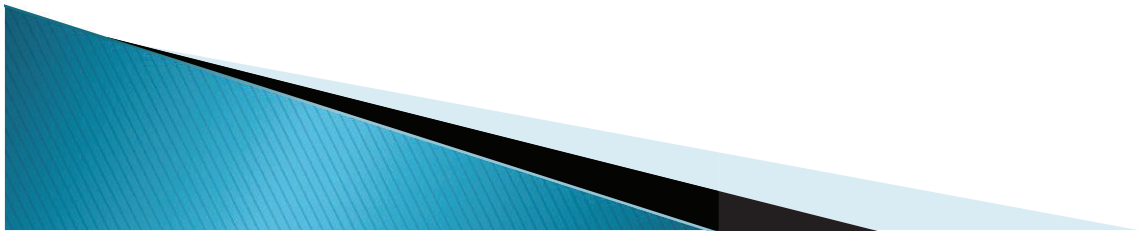
Extended ASCII is not enough for international use

Unicode uses 16 bits per character

How many characters can UNICODE represent?

Unicode is a superset of ASCII

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

Compression

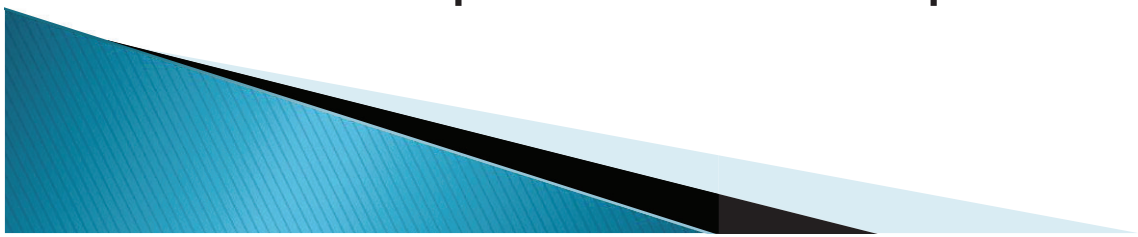
- **Data compression**

Reduction in the amount of space needed to store a piece of data

- **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 the original information
  - **lossy**, which means some information may be lost in the process of compaction



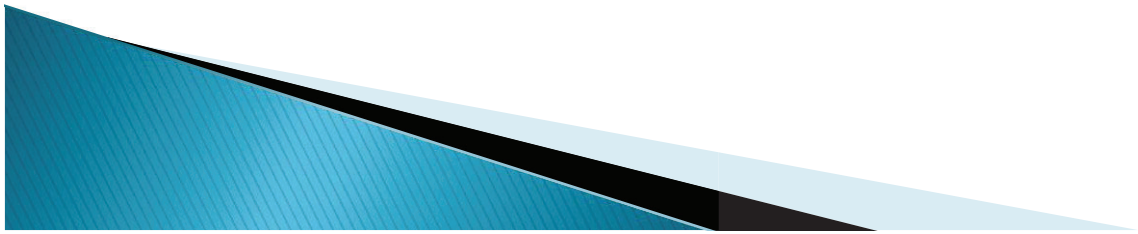
Text Compression

Assigning 16 bits to each character in a document uses too much file space

We need ways to store and transmit text efficiently

Text compression techniques

- keyword encoding
- run-length encoding
- Huffman encoding



Run-Length Encoding

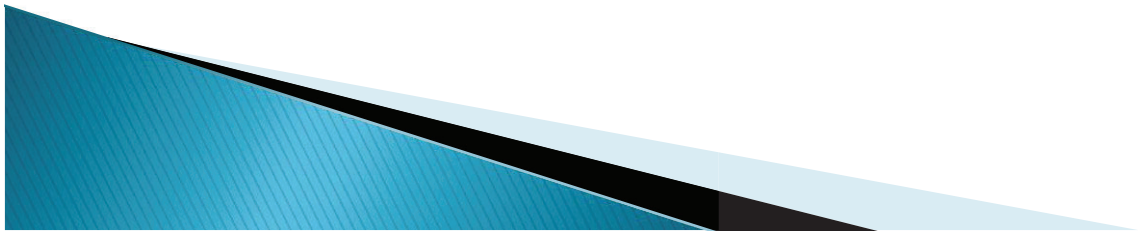
A single character may be repeated over and over again in a long sequence

Replace a repeated sequence with

- a flag character
- repeated character
- number of repetitions

*x8

- * is the flag character
- x is the repeated character
- 8 is the number of times x is repeated



Run-Length Encoding

Original text

bbbbbbbbjjjkllqqqqqq+++++

Encoded text

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

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?