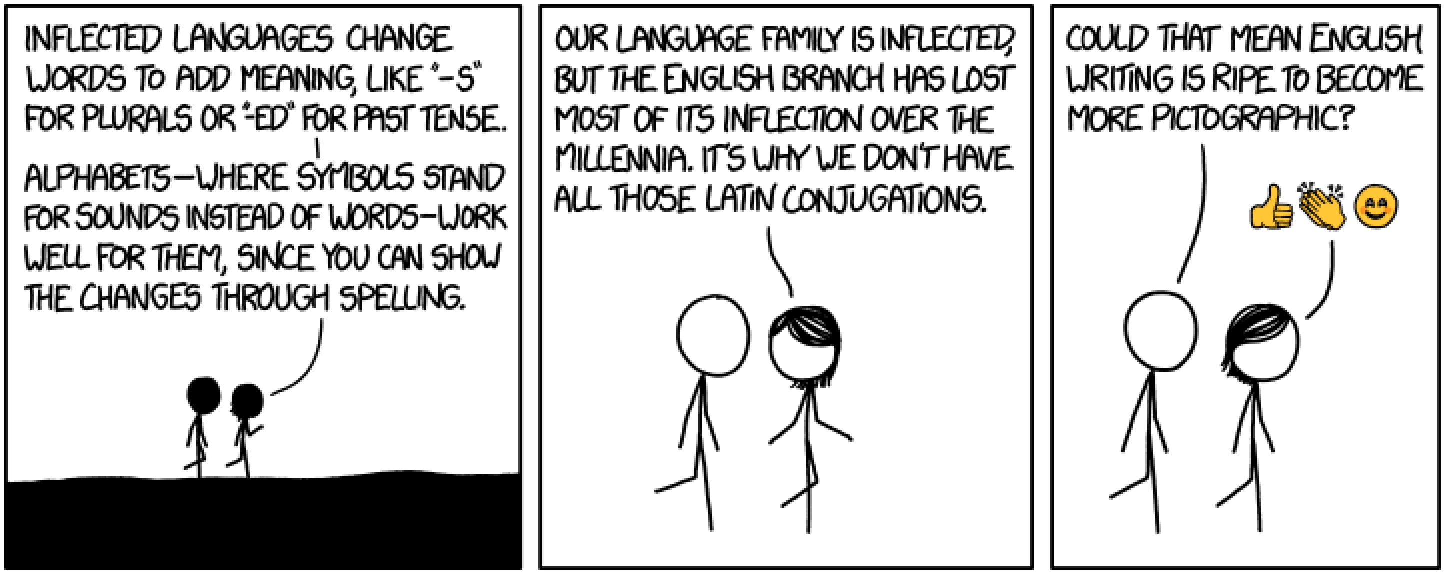
ENGSCI 233 Lecture 2

Structured Data

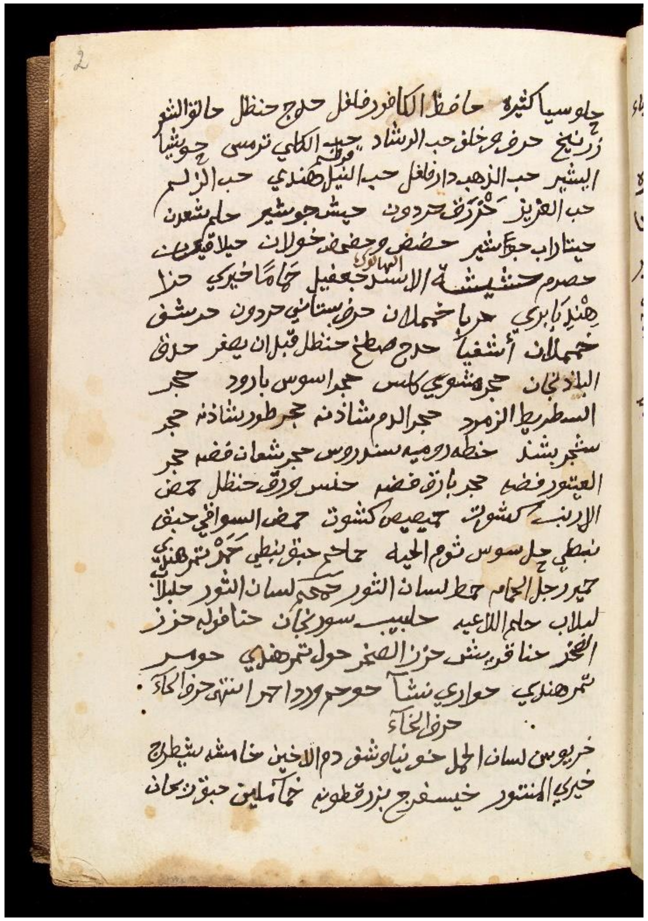

INFLECTED LANGUAGES CHANGE WORDS TO ADD MEANING, LIKE "-S" FOR PLURALS OR "-ED" FOR PAST TENSE.
ALPHABETS—WHERE SYMBOLS STAND FOR SOUNDS INSTEAD OF WORDS—WORK WELL FOR THEM, SINCE YOU CAN SHOW THE CHANGES THROUGH SPELLING.
OUR LANGUAGE FAMILY IS INFLECTED, BUT THE ENGLISH BRANCH HAS LOST MOST OF ITS INFLECTION OVER THE MILLENNIA. IT'S WHY WE DON'T HAVE ALL THOSE LATIN CONJUGATIONS.
COULD THAT MEAN ENGLISH WRITING IS RIPE TO BECOME MORE PICTOGRAPHIC?
👍👏😄

Today’s objectives:

- Understand role of endianness in data storage
- Understand how text data are represented in computers
- Know how to handle situations where other forms of data need to be stored

What happens when we have more than 8 bits?

Which way does a computer read?



2

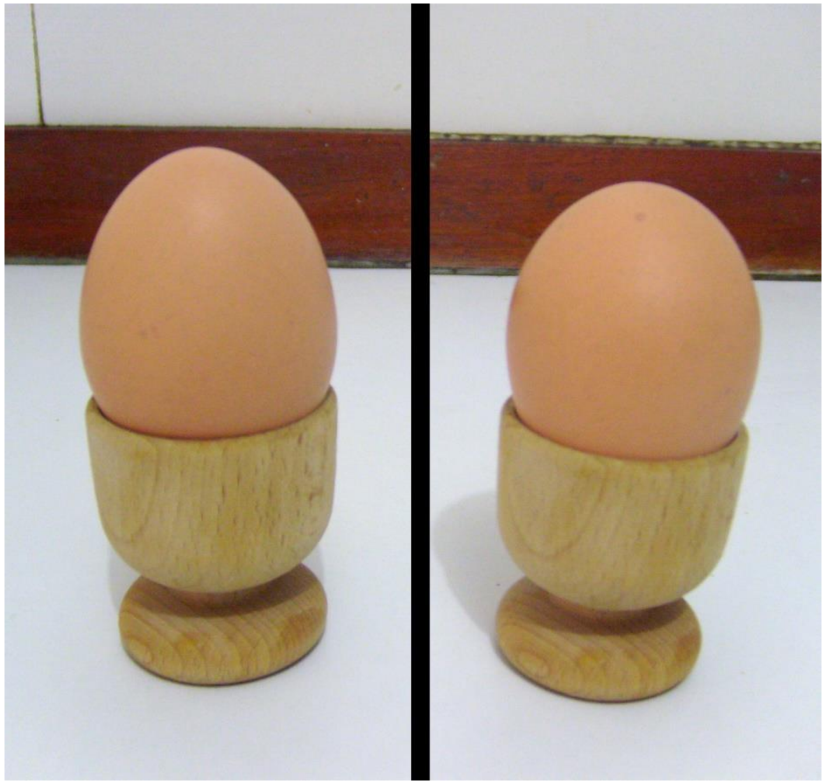
حلوسيا كثيرة حافظ الكافور فلفل حوج حنظل حالق الشعر
زرنيخ حرف وخلو حب الرشاد حب الكلى ترمس حوشيا
البشير حب الزهب دار فلفل حب النيل هندي حب الزلم
حب العزيز خزز حردون حيش جوشير حلم شعرين
حيتاراب حجر اشير حضض وحضض خولان حيلا قيمون
حصرم حشيشة الاسد جعفيل حماما خيري حزا
هندبا بري حربا خملان حرف بستاني حردون حرشف
خملاف أشغيا حرح صطخ حنظل قبل ان يصغر حرف
البادنجان حجر مشوي كلس حجر اسوس بارود حجر
السطريط الزمرد حجر الدم شاذنه حجر طور شاذنه حجر
سنجر بشد حنطة رومية سندروس حجر شعات فضه حجر
العنتور فضه حجر بازه فضه حنس ورق حنظل حمص
الارنب كشوث حميصه كشوث حمض السوافي حبق
نبطي حل سوس ثوم الحية حماحم حبق نبطي حمر تمر هندي
حير رجل الحمام حط لسان الثور حمحم لسان الثور حلبلاب
لبلاب حلم اللاعبه حلبيب سورنجان حنافوله حزز
الصخر حنا قريش حزز الصخر حول تمر هندي حومر
تمر هندي حواري نشا حوحم وردا احمر انتهى حرف الحاء

حرف الخاء

خربوبن لسان الحمل خونياوشو دم الاخين خامشه سيطرج
خيري المنثور خيسفرج بزرقطونه خماملين حبوزبحان

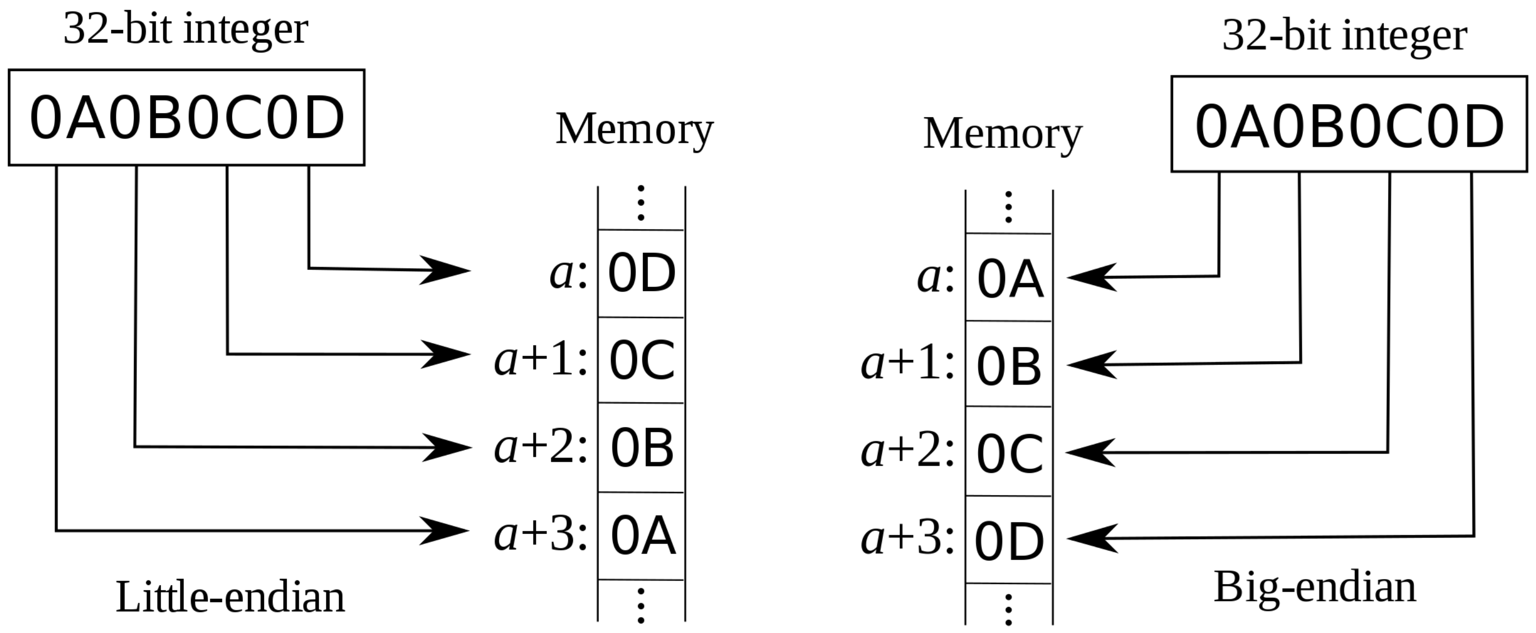
- We have been reading left-to-right
- Not everyone does!
- Computers are no different…

Which way does a computer read?



- Big-endian
  - First byte is highest
  - Used in mainframes and network equipment
- Little-endian
  - First byte is lowest
  - Used on most PCs

What does this look like?



What does this look like?

- 0xFACEFEED

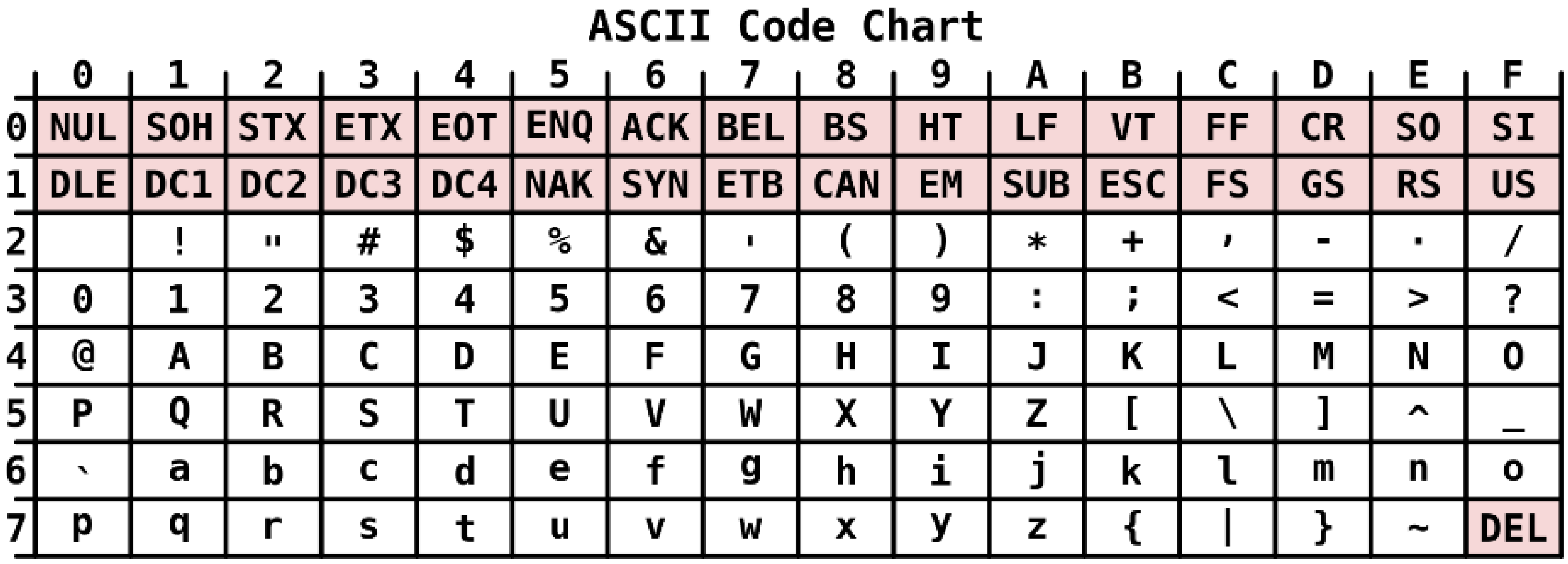
Why does endianness matter?

- Your code normally can’t see it!
  - Programming languages do a good job of hiding it.

- What happens when you store data, though?

What about non-numbers?

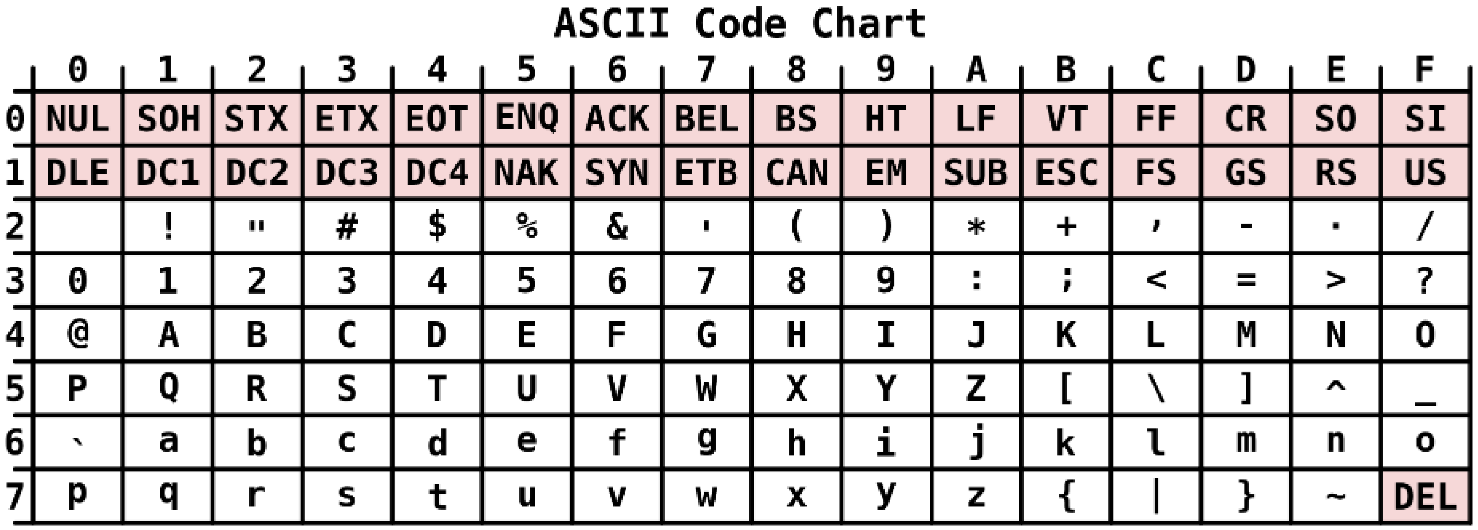
- Encode each character as a number.
- But how?
- English fits in seven bits:



ASCII Code Chart

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

How is ASCII used?

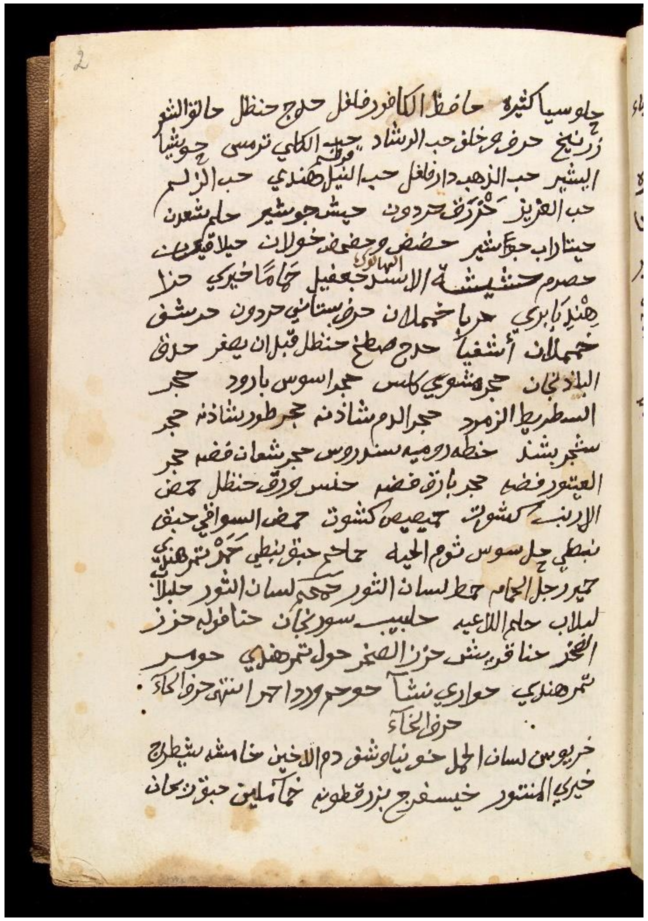


ASCII Code Chart

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

```
Hello, world!
```

The world no longer uses ASCII.



2

جلوسيا كثيره حافظ الكافور فلفل حوج حنظل حالق الشعر
زرنيخ حرف حرخلو حب الرشاد حب الكلي ترمسي جوشيا
البشير حب الزهب دار فلفل حب النيل هندي حب الزلم
حب العزيز خزرف حردون حيشجوشير حلم شعون
حيتاراب حب اشير حضض وحضض خولان حيلاقيعون
حصرم حشيشة الاسر جعفيل حاماخيري حزا
هندبا بري حربا خملان حرف بستاني حردون حرشف
خملاف أشغيا حرج صطخ حنظل قبل ان يصفر حرق
البادنجان حجر مشوي كلس حجر اسوس بارود حجر
السطريط الزمرد حجر الدم شاذنه حجر طور شاذنه حجر
سنجر بشند خنطه روميه سندروس حجر شعان فضه حجر
العيشور فضه حجر بازق فضه حنس ورق حنظل حمض
الارنب كشوث حميصه كشوث حمض السواقي حبق
نبطي جلسوس ثوم الحيه حاحم حبق نبطي حمر تمر هندي
حير رجل الحمام حط لسان الثور حصحم لسان الثور حلبلا
لبلاب حلم اللاعبه حلبيب سورنجان حنافوله حزز
الصخر حنا قريش حزز الصخر حول تمر هندي حومر
تمر هندي حوادي نشا حوحم وردا حمر انتهى حرف الحاء

حرف الخاء

خربوبين لسان الحمل خو نباوشو دم الاخين خامشه شيطرج
خيري المنثور خيسفرج بزرقطونه خماملين حبون نحاف

The world no longer uses ASCII.

How can we represent every language?

- One standard with 137,439 characters
- Room for over 1 million!
- We can’t fit every character in one byte any more…

How do we store this?

- The simple way: UTF-32
- 4 bytes per character

💩	☕	A
U+1F4A9	U+2615	U+41
0x0001F4A9	0x00002615	0x00000041

How do we store this?

- The Windows way: UTF-16
- 2 bytes per character, if possible

		A
U+1F4A9	U+2615	U+41
0xD83D DCA9	0x2615	0x0041

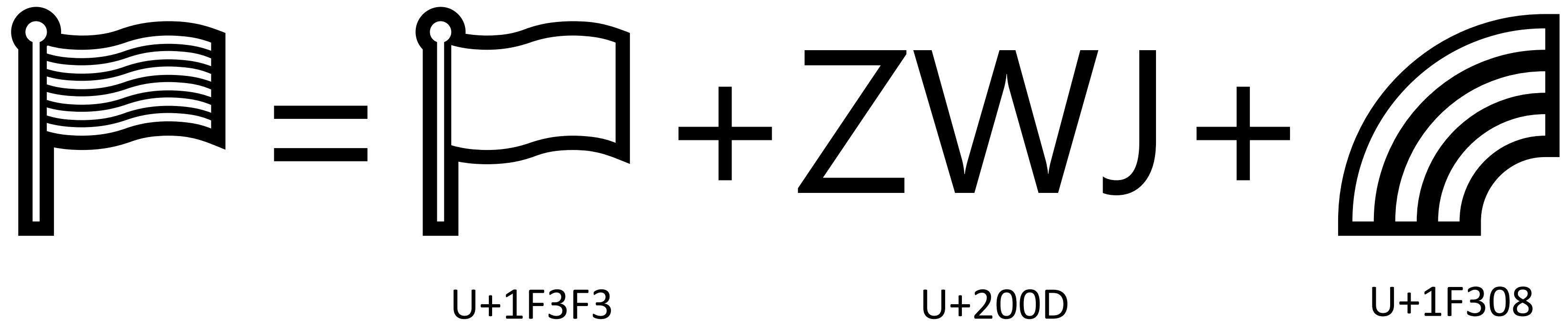
How do we store this?

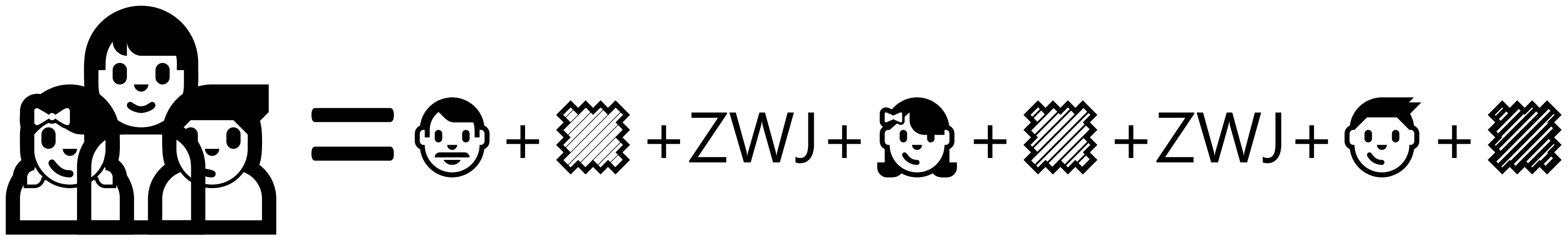
- The standard way: UTF-8
- As few bytes as unambiguously possible

A

U+1F4A9	U+2615	U+41
0xF0 9F 92 A9	0xE2 98 95	0x41

Some entities use more characters.





What does this mean for you?

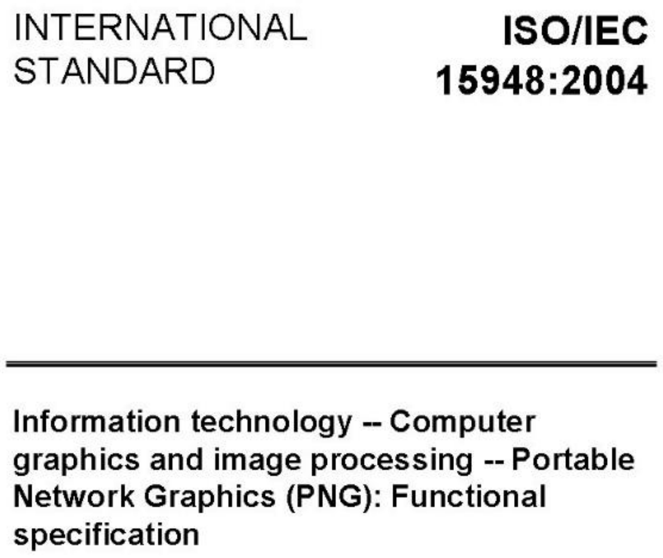
- ASCII strings are always 1 byte per character
- UTF-8 and UTF-16 have a variable character size
  - How do you figure out the length of a string in memory?
- Python cheats and uses the smallest constant-size representation that will work…
- You can’t assume that displayed characters and numbers are interchangeable any more!
- You can’t predict required display space from string length.

What else is there to store?

- Images?
- Formatted text?
- Executable code?

Use a standard!

- Images?
- Formatted text?
- Executable code?



INTERNATIONAL STANDARD

ISO/IEC 15948:2004

Information technology -- Computer graphics and image processing -- Portable Network Graphics (PNG): Functional specification

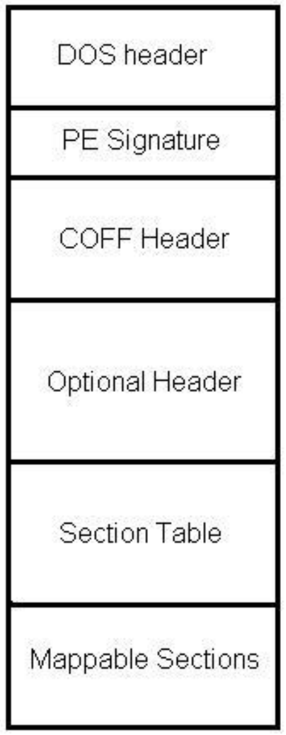
Use a standard!

- Images?
- Formatted text?
- Executable code?

Adobe® PDF

Use a standard!

- Images?
- Formatted text?
- Executable code?



How do standards work?

- Often contain a magic number or signature
  - PNG image has ASCII “PNG” plus bytes to detect mis-processing
  - PDF starts with ASCII “%PDF-”
- Fixed size fields or variable size stored with length
- Metadata stored with data
- Documentation made available to all
  - Not always freely!

What about pointers?

- Don’t keep pointers around for posterity!
- A pointer has the size in memory dictated by the computer architecture.

8-bit

32-bit

16-bit

64-bit

Next time: Computer Hardware

Image References

Slide 1: *Inflection*, by Randall Munroe, from https://xkcd.com/1709/ (CC BY-NC 2.5)
Slides 3 and 19: by Christiaan Colen, from https://www.flickr.com/photos/christiaancolen/20607150556 (CC BY-SA 2.0)
Slides 4 and 11 (left): from the Wellcome Collection, https://wellcomecollection.org/works/vzjsdw4c?query=L0033679 (CC BY 4.0)
Slide 5: by Cirofono, from https://www.flickr.com/photos/ciroduran/2060860569 (CC BY 2.0)
Slide 6 (left): by R. S. Shaw, from https://commons.wikimedia.org/wiki/File:Little-Endian.svg (Public domain)
Slide 6 (right): by R. S. Shaw, from https://commons.wikimedia.org/wiki/File:Big-Endian.svg (Public domain)
Slide 8: by Evan-Amos, from https://commons.wikimedia.org/wiki/File:SanDisk-Cruzer-USB-4GB-ThumbDrive.jpg (Public domain)
Slides 9 and 10: by Anomie, from https://commons.wikimedia.org/wiki/File:ASCII_Code_Chart.svg (Not subject to copyright)
Slide 11 (right): by John Vetterli, from https://www.flickr.com/photos/51824383@N00/935697405 (CC BY-SA 2.0)
Slide 12: from https://techcrunch.com/2017/08/21/meet-android-oreos-all-new-emoji/
Slides 14-16 (left): by Google, from https://commons.wikimedia.org/wiki/File:Emoji_u1f4a9.svg (Apache license)
Slides 14-16 (centre): by Google, from https://commons.wikimedia.org/wiki/File:Emoji_u2615.svg (Apache license)
Slides 14-16 (right): by Eirik1231, from https://commons.wikimedia.org/wiki/File:Latin_alphabet_Aa.svg (Public domain)
Slide 22: by Whiteknight, from https://commons.wikimedia.org/wiki/File:RevEngPEFile.JPG (Public domain)
Slide 24 (top left): by Evan-Amos, from https://commons.wikimedia.org/wiki/File:NES-Console-Set.jpg (Public domain)
Slide 24 (top right): by Evan-Amos, from https://commons.wikimedia.org/wiki/File:PSX-Console-wController.jpg (Public domain)
Slide 24 (bottom left): by Evan-Amos, from https://commons.wikimedia.org/wiki/File:SNES-Mod1-Console-Set.jpg (CC BY-SA 3.0)
Slide 24 (bottom right): by Evan-Amos, from https://commons.wikimedia.org/wiki/File:Nintendo-64-wController-L.jpg (Public domain)