Digital Representation

INFO/CSE 100, Spring 2005

Fluency in Information Technology

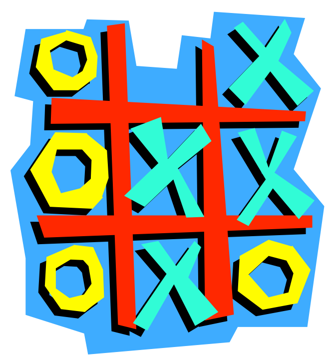
http://www.cs.washington.edu/100

Readings and References

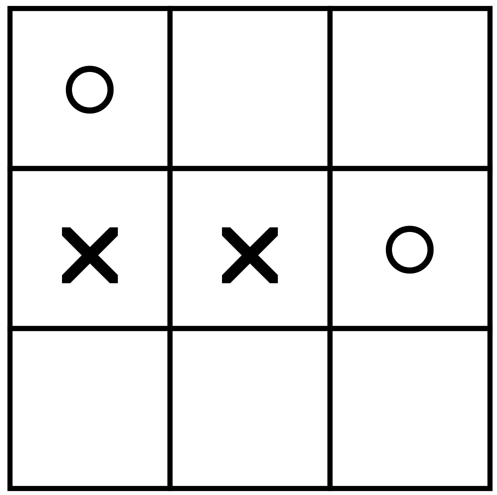
- Reading
  - » *Fluency with Information Technology*
    - Chapter 8, Bits and the "Why" of Bytes

Info Representation

- Digitization: representing information by **any** fixed set of symbols
  - » decide how many different items of information you want to represent
    - Tic Tac Toe: 2 items - player 1 or player 2
  - » decide how many "digits" or positions you want to use
    - Tic Tac Toe: 1 position - a board square
  - » decide on a set of symbols
    - player 1: ✕
    - player 2: O



Are two symbols enough?



We can represent each player's move this way, but what about representing the whole game?

Empty position: ⊗

use this set of symbols

- empty cell: ⊗
- player 1: ✕
- player 2: ○

○	⊗	⊗
✕	✕	○
⊗	⊗	⊗

- Now we can represent this game as one 9-digit length string:

  ○ ⊗ ⊗ ✕ ✕ ○ ⊗ ⊗ ⊗

- How many possible game states are there?
  - » $3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 = 3^9 = 19683$

Another encoding

use a different set of symbols

- empty cell: 0
- player 1: **1**
- player 2: **2**

2	0	0
1	**1**	**2**
0	0	0

- Now we can represent this game as one 9-digit number:

  2 0 0 1 1 2 0 0 0

- How many possible game states are there?

  » $3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 \times 3 = 3^9 = 19683$

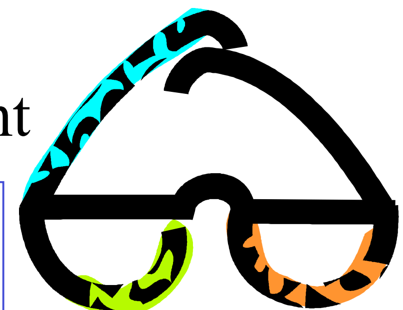
Info in the Physical World

- Physical world:
  - » The most fundamental representation of information is presence/absence of a phenomenon
    - matter, light, magnetism, flow, charge, …

The PandA representation

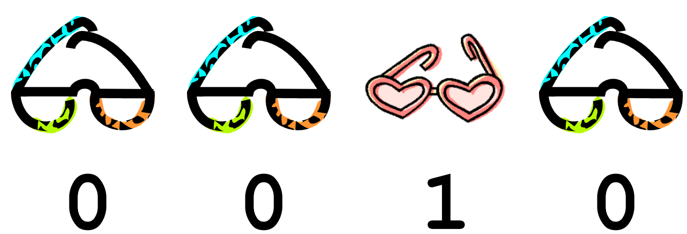
- detect: "Is the phenomenon present?"
- set: make phenomenon present or absent

Any controllable phenomenon works: define it right



Info in the Logical World

- Logical World:
  - » Information, reasoning, computation are formulated by true/false and logic
    - All men are mortal
    - Aristotle is a man
    - Aristotle is mortal
- True and false can be the patterns for encoding information



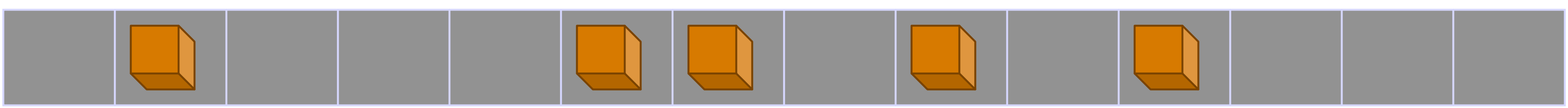
Connect Physical/Logical

- The power of IT comes from the fact that physical and logical worlds can be connected

Present represents true / Absent represents false

-- or maybe vice versa --

Pavement Memory



false true false false false true true false true false true false false false

0 1 0 0 0 1 1 0 1 0 1 0 0 0

Bits

- PandA is a *binary representation* because it uses 2 patterns
- The word "bit"
  - » is a contraction for "binary digit"
  - » represents a position in space/time capable of being set and detected in 2 patterns

Sherlock Holmes's *Mystery of Silver Blaze* -- a popular example where "absent" gives information … the dog didn't bark, that is the phenomenon wasn't detected

Possible Interpretations of Bit Patterns

Present	Absent
True	False
1	0
On	Off
Yes	No
+	-
Black	White
For	Against
Yang	Ying
…	…

Assigning Symbols for Characters

26 uppercase and 26 lowercase letters

10 digits

20 basic punctuation characters

= 95 distinct characters

Representing this many characters in binary takes 7 bits!

2^6 (6 bits) gives 64 symbols

2^7 (7 bits) gives 128 symbols

7-bit code for characters is ASCII

(American Standard Code for Information Interchange)

8-bit ASCII

ASCII	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111
0000	NU	SH	SX	EX	ET	EQ	AK	BL	BS	HT	LF	VT	FF	CR	SO	SI
0001	DL	D1	D2	D3	D4	NK	SY	EB	CN	EM	SB	EC	FS	GS	RS	US
0010		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
0011	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0100	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0101	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0110	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0111	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DT
1000	80	81	82	83	IN	NL	SS	ES	HS	HJ	VS	PD	PU	RI	S2	S3
1001	DC	P1	P2	SE	CC	MW	SP	EP	98	99	9A	CS	ST	OS	PM	AP
1010	A0	¡	¢	£	¤	¥	¦	§	¨	©	♀	«	¬	-	®	¯
1011	°	±	²	³	´	µ	¶	·	¸	¹	♂	»	¼	½	¾	¿
1100	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
1101	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
1110	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
1111	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

0100 0110
0100 1001
0101 0100

Bytes

- A byte is eight bits treated as a unit
  - » Adopted by IBM in 1960s
  - » A standard measure until very recently
  - » Bytes encode the Latin alphabet using ASCII -- the American Standard Code for Information Interchange

0100 0110
0100 1001
0101 0100

How many bytes?!?

Unicode

- Although 8-bit ASCII is widely used, there is a problem!!!
  - » Doesn’t can’t support more than 256 characters
  - » This eliminates more than half of the world’s language from the character set
- Unicode is a 16-bit representation
  - » Supports 65,536 symbols
  - » Can handle all languages

0100 0110 0000 1001

Escape Codes

- Escape codes solve the problem of creating more symbols
- Put one symbol aside to be the esc symbol.
- Add esc symbol in front of another to create a new symbol
  - » Ctrl-N
- HTML uses 7-bit ASCII when transmitting data over the web
  - » HTML uses two special characters < > symbols
  - » What happens if you want those symbols to appear in the content?
    - < >

Hexadecimal Representation

- Computers can very fluently read the binary representations
  - » 0100001010101110101011110101010001010
- Hex digits (base-16) numbers are used instead
  - » 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F
  - » Easily represent 4-bit sequences
  - » 0010 1011 1010 1101 = 2BAD
  - » 0001 1011 0100 0000 = 1B40
- Examples of hex in use: HTML color codes
  - » red = #FF0000

Encoding Information

- Bits and bytes encode the information, but that's not all
  - » Tags encode format and some structure in word processors
  - » Tags encode format and some structure in HTML
  - » In the Oxford English Dictionary tags encode structure and some formatting

Summary

- IT joins physical & logical domains so physical devices do our logical work
  - » Symbols represent things 1-to-1
  - » Create symbols by grouping patterns
  - » PandA representation is fundamental
    - presence and absence
  - » Bit, a place where 2 patterns set/detect
  - » ASCII is a byte encoding of Latin alphabet
  - » In addition to content, encode structure