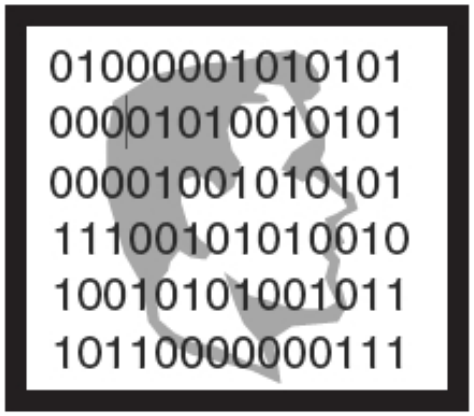
CS Concepts

- document formats
- interpreting bits
- ascii, jpg, mp3, …
- meta data
- representing digital images
- modeling vs rendering
- ocr
- sampling rate
- cloud computing
- data compression
- spatial coherence
- temporal coherence
- TCP/IP
- role of processing power in audio/video
- steganography
- disk format (deleting data)
- persistence of data (good and bad)

Social Issues

- choosing the right file format
- knowing what is REALLY in a document
- risks of digital documents
- power of digital documents
- movies on your phone
- spam
- espionage
- cheating
- forgery
- access to your data years from today
- destroying old records/ data

MODELING

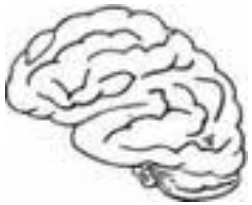


RENDERING

REALITY

REPRESENTATION
OR MODEL

IMAGE



text, image,
video, audio,
building,
imaginary world,
…

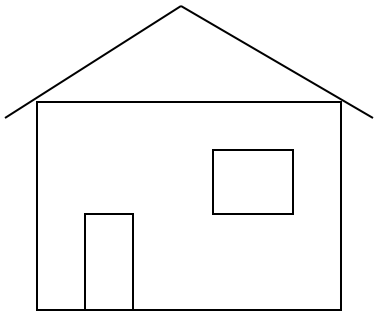
01101010010010
10100101010001
01010010101001
01010101011110
10110101001010

How does a printed document differ from a digital document?

What's in the model?

- omit details
- the number of bits used determines how much of "reality" must be omitted from the model

House



WYSI(not)WYG

- CIA report published by NY Times (2000)
- Washington snipers published by Washington Post (2002)
- case of journalist Gioliana Sgrena (2005)

4. (U) Effectiveness of Attacks

(U) The number of IED detonations from 15 June 2003 through 4 March 2005 (the date of the incident), has steadily increased. [redacted] the overall average number of casualties during that period is nearly one per IED detonation. (Annex 4E).

[redacted]

(U) The number of VBIED detonations from 15 June 2003 through 4 March 2005 has also seen a relatively steady increase. [redacted] there have been spikes for particular VBIED events that have produced large numbers of casualties. (Annex 4E).

[redacted]

WYSI(not)WYG

- pdf blackout
- MS Office track changes
  - UN report on assassination
  - SCO lawsuit
- scanned vs pdf/doc
  - not searchable/searchable
  - not easily “read” by readers for the blind
- use security “features” to prevent unauthorized modification

Document Metaphor?

- More than meets the eye.
- Nearly impossible to retract/destroy once shared (or even when not shared)
- Do you know what is really in the documents you share?

Metadata

- data about data
  - filename
  - file extension
  - timestamp (created, last modified)
  - creator
  - …
- forgeable

Common “Text” Formats

- plain text (.txt)
- rich text format (.rtf)
- MS Office (.doc)
- Adobe PDF (.pdf)

Choosing the right model/format

192 bits in ascii

998,008 bits in jpg

IN PRINCIPIO

ERAT VERBUM

01001001 01001110
00100000 01010000
01010010 01001001
01001110 01000011
01001001 01010000
01001001 01001111
00100000 01000101
01010010 01000001
01010100 **00100000**
01010110 01000101
01010010 01000010
01010101 01001101

Representing Images

- gif, jpg, tiff, png, …
- RGB
- pixel
- compression
- lossy
- lossless

Can You Believe Your Eyes?

Exposing Digital Forgeries in Complex Lighting Environments

M.K. Johnson and H. Farid

IEEE Transactions on Information Forensics and Security, 2007

Need for Compression

- HD TV 1080p is 1920x1080 pixels
- or 2,073,600 pixels
- or 49,766,400 bits using uncompressed 24 bit color (8 bits for each of Red, Green, and Blue). That's ~50Mbits per image/frame.
- 20 pictures would be 1Gbit
- 1 minute of video at 16 frames/sec would be 47,775,744,000 (~50Gbits)

How Long?

Assuming a high speed home Internet connection of 8Mbits/sec, about how long would it take to download 1 minute of uncompressed HDTV video (~50Gbits)?

A. 15 seconds

B. 1.5 minutes

C. 15 minutes

D. 1.5 hours

E. 15 hours

Compression

- sampling rate (not exactly compression)
- lossy or lossless
- run length encoding
- spatial coherence (lots of blue sky)
- temporal coherence (video of static scene)
- compression trades computing time (power) for storage space or bandwidth

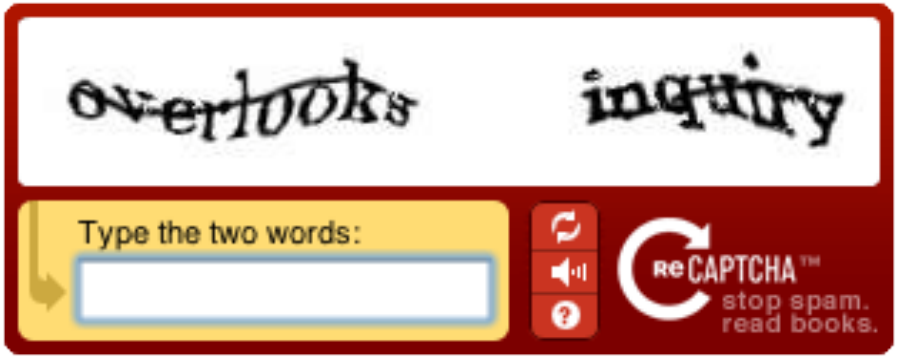
Non-computing examples of trading processing/ preparation time for space or other precious resource?

Proprietary vs Public Data Formats

- MS Office (.doc, .ppt, .xls)
- Open Document Format (.odt, .odp, .ods)
- VHS vs BetaMax
- Blue-Ray vs HD DVD
- TCP/IP (underlying the Internet)

Hiding Information in Images

- Spam Wars (hiding the text from the spam filter)
  - OCR to the rescue
- Captcha



- Steganography (hiding the info from the human viewer/listener)

Combining two images

- Put guest bits into right 2 bits of host

1111 0100 1101 0011 1011 1101

1011 0100 1101 0011 0001 1100

1111 0110 1101 0011 1011 1100

What bits matter most?

If you had to throw away some bits from each pixel of a message which should you throw away so that the remaining image was a close to the original as possible?

A. The rightmost (low order bits).

B. The leftmost (high order bits).

C. Doesn't matter, just how many you throw is all that matters.

Data That Just Won't Go Away

- Disk Format
  - files = magazine article (continues on page x)
  - index/TOC
  - deleting just removes entry from index/TOC
- Deleting data in the Cloud?
- Data on your cell phone?
- Once a file is copied, it is hard to totally eradicate

But Will It Last?

- floppy disks
- home movies and digital photos
- The “Domesday Book” of 1086
  - digital copy made in 1986 was unreadable by 2001
- Are your cds or dvds safe?
- If the media lasts will you be able to read them?

Recap – Ghosts in the Machine

- Meta data – what you see is less than what you get
- Steganography – hiding info in plain sight
- Erased/deleted data may still be around
  - On your disk drive
  - In the cloud
  - On your phone

CS Concepts

- document formats
- interpreting bits
- ascii, jpg, mp3, …
- meta data
- representing digital images
- modeling vs rendering
- ocr
- sampling rate
- cloud computing
- data compression
- spatial coherence
- temporal coherence
- TCP/IP
- role of processing power in audio/video
- steganography
- disk format (deleting data)
- persistence of data (good and bad)

Social Issues

- choosing the right file format
- knowing what is REALLY in a document
- risks of digital documents
- power of digital documents
- movies on your phone
- spam
- espionage
- cheating
- forgery
- access to your data years from today
- destroying old records/ data