Ghosts in the Machine
CS Concepts

- Document formats
- Interpreting bits
- ASCII, JPG, MP3, ...
- Metadata
- Representing digital images
- Modeling versus rendering
- OCR
- Sampling rate
- Cloud computing

- Data compression
- Spatial coherence
- Temporal coherence
- TCP/IP
- Processing power in audio/video
- Steganography
- Formatting disks
- Data persistence
Social Issues

- Choosing the right format
- What's really in a document?
- Risks of digital documents
- Power of digital documents
- Kiddie porn
- Handheld movies
- Spam
- Espionage
- Cheating
- Forgery
- Access to data years from now
- Destroying old records/data
Model of Reality

(text, image, video, audio, imaginary worlds)

(Do we live in a digital simulation???)
How does a printed document differ from a digital document?
What's in the model?

- Models omit details
- # of bits determines how much “reality” must be omitted
WYSI$(not)$WYG

- What you see is $not$ what you get
- Journalist Gioliana Sgrena (2005)
- UK Ministry of Defence nuclear submarine (2011)
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).

(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).
WYSI(\textit{not})WYG

- PDF blackout
- Microsoft Office track changes
  - UN report on assassination
  - SCO lawsuit
  - countless resumes
- Scanned vs. PDF/doc
  - not searchable/indexable
  - harder to “read” by the blind
- Security “features” to prevent unauthorized modification, copying
● Thinking of digital data as a “document”
  - This is a metaphor
  - More going on than might be obvious
  - Do you know what is in the data?
● Retract/destroy data
  - Nearly impossible once shared
  - Streisand effect
    • 2003, aerial photographs of Barbra Streisand's Malibu residence
    • Part of coastal erosion research
    • Unsuccessful lawsuit to suppress photos
    • More attention to the data
• Data about the data
  – filename
  – file extension
  – time created
  – time last modified
  – creator
  – length
  – endpoints (for communications)
• Can be forged
Common “text” Formats

- Plain text (.txt)
- Rich Text Format (.rtf)
- Microsoft Office (.doc .docx)
- Adobe PDF (.pdf)
Choosing the model/format

192 bits in ASCII

IN PRINCIPIO
ERAT VERBUM
01001001 00100000
01010010 01001110
01001001 01001001
00100000
01010010
01010100 01010110
01010010 01010101
01001110 01010000
01001001 01000011
01010000 01001111
01000101 01000001
00100000
01000101
01000010 01001101

998,008 bits in JPG

192 bits in ASCII

IN PRINCIPIO
ERAT VERBUM
01001001 00100000
01010010 01001110
01001001 01001001
00100000
01010010
01010100 01010110
01010010 01010101
01001110 01010000
01001001 01000011
01010000 01001111
01000101 01000001
00100000
01000101
01000010 01001101

998,008 bits in JPG
Representing Pictures

```
0 1 1 1 1 1 1 0
1 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0
0 1 1 1 1 1 0 0
0 0 0 0 0 0 1 0
0 0 0 0 0 0 1 0
1 1 1 1 1 1 0 0
0 0 0 0 0 0 0 0
```
Can you believe what you see?
This is not Magaritte's pipe
Unreality

- Computers let us make images of things that don't exist.
- “Kiddie porn” is illegal, other forms are not.
- What about artificial “kiddie porn”?
  - PROTECT Act of 2003
- Second Life
  - Banned “ageplay” in 2007

This cat does not exist
Why we need compression

- HD TV 1080p is 1920x1080 pixels
  - One image is 1920 x 1080 = 2,073,600 pixels
  - At 8 bits for R, 8 bits for G, 8 bits for B...
  - 49,766,400 bits per image, ~50 Mbits per image
  - 24 frames is ~1.2 Gbits (1 second at 24 frames/sec)
  - 1 minute is 60 seconds, ~72 Gbits or ~9 Gbytes
Suppose you have a high-speed home internet connection of 8Mbit/s, about how long would it take to download 1 minute of uncompressed HDTV video at 24 fps (~72Gbits)?

A. 9 seconds
B. 15 seconds
C. 2.5 minutes
D. 2.5 hours
E. 12 hours
Compression

- Sampling rate
  - e.g. time lapse photography
- lossy or lessless
- run length encoding
- spatial conherence (solid white background)
- temporal coherence (nothing moving)
- compression trades:
  - computing time vs. storage space/bandwidth
Other examples...

What are non-computing examples of trading processing/preparation time for space or other precious resources?
Formats: Proprietary vs. Open

- Microsoft Office (.doc .docx .ppt .xls)
- Open Document Format (.odt .odp .ods)
  - LibreOffice
- VHS versus BetaMax
- Blu-Ray vs. HD-DVD
- TCP/IP
Hiding Information

- Spam Wars
  - hiding from the filter
  - “Amaz!ng 0ffer”
- CAPTCHA
- Steganography
  - Hiding information inside other data
Encoding (Example)

000001 = 1
00010 = 2
00100 = 4
01000 = 8
10000 = 16

11001 = ?
<table>
<thead>
<tr>
<th>Binary</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>000001</td>
<td>a</td>
</tr>
<tr>
<td>000100</td>
<td>c</td>
</tr>
<tr>
<td>001000</td>
<td>d</td>
</tr>
<tr>
<td>010000</td>
<td>e</td>
</tr>
<tr>
<td>100000</td>
<td>f</td>
</tr>
<tr>
<td>110011</td>
<td>g</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binary</th>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>00010</td>
<td>b</td>
</tr>
<tr>
<td>00100</td>
<td>c</td>
</tr>
<tr>
<td>01000</td>
<td>d</td>
</tr>
<tr>
<td>10000</td>
<td>e</td>
</tr>
<tr>
<td>11001</td>
<td>f</td>
</tr>
</tbody>
</table>

- 00001 = 1  
- 00010 = 2  
- 00100 = 4  
- 01000 = 8  
- 10000 = 16 
- 11001 = 16+8+1 = 25

Message: 01101 10101 10011 01001 00011
What was the message?

A. easy
B. abuzz
C. decoy
D. music
E. stone
Secret Messages

- http://youtu.be/L-v4Awj_p7g
Hiding an image inside an image

00110000 01000000 00101010

00110010 01000101 00010011

001011...
If you had to throw away some bits from each pixel in an image, which should you throw away so that the remaining image is as close to the original as possible?

A. The rightmost (low order bits).
B. The leftmost (high order bits).
C. It doesn't matter, only how many bits you throw away.
Data Just Won't Go Away

- Format a disk
  - “Warning!! This operation will erase all information on the disk.”
  - Actually...
    - Disk contains chunks of content indexed with tables
    - Format just erases tables (data is still there)
- Delete data in the cloud?
- Delete data in your phone?
The Cloud

• What is the cloud?
  – Google Docs
  – Google Drive, Dropbox
  – Microsoft Office
  – How are these different?
Will it Last?

- Floppy disks
  - Wear out in a year or two
- Home movies, digital photographs
- Domesday Book from 1086
  - A digital copy made in 1986 was unreadable by 2001
- Digital archives can be actively copied
  - If the equipment still exists...
Recap

- Ghosts in the machine
- Metadata
  - What you see is less than what you get
- Steganography
  - Hiding information in plain sight
- Erased data
  - Might still be around
    - On your hard drive
    - In the cloud
    - On devices
CS Concepts

- Document formats
- Interpreting bits
- ASCII, JPG, MP3, ...
- Metadata
- Representing digital images
- Modeling versus rendering
- OCR
- Sampling rate
- Cloud computing
- Data compression
- Spatial conherence
- Temporal coherence
- TCP/IP
- Processing power in audio/video
- Steganography
- Formatting disks
- Data persistence
Social Issues

- Choosing the right format
- What's really in a document?
- Risks of digital documents
- Power of digital documents
- Kiddie porn
- Handheld movies

- Spam
- Espionage
- Cheating
- Forgery
- Access to data years from now
- Destroying old records/data
Essay

- Due by 5pm Friday
- Turn in at:
  http://www.crowdgrader.org/crowdgrader/venues/view_venue/239
Peer Grading

- Your work is graded by your peers
- You grade others' work (5 others)

Your grade is

75% evaluations of your work

25% how good your evaluations were

sophisticated algorithms

4.3 Helpfulness grade

To provide an incentive for students to write helpful reviews, CrowdGrader allows students to rate and leave feedback on each review they receive. The feedback provided on reviews goes some of the way towards providing an incentive to write helpful reviews: all students naturally like to receive praise for their work, rather than having sloppiness or imprecisions pointed out to them. Furthermore, students can rate the reviews they receive, assigning them integer ratings ranging from -2 (incorrect, completely unhelpful) to +2 (very helpful), with 0 being the neutral rating. These review ratings are used to compute the helpfulness grade $h_j \in [0, 1]$ of each reviewer $j \in U$, as follows. Let $L_j$ be the list of review feedbacks received by $j$ (this list may be shorter than the number of reviews performed by $j$, as some reviews might not have received any feedback). We drop from $L_j$ one of the lowest feedbacks $f$ with $f = \min L$, obtaining $L'$. For a feedback $f \in \{-2, \ldots, +2\}$, we let the weight of feedback $f$ be $w(f) = 2$ if $f < 0$, and $w(f) = 1$ if $f \geq 0$, so that negative feedback weights twice as much as positive feedback. We then compute the helpfulness grade of student $j \in U$ via:

$$h_j = \max \left[ 0, \min \left( 1 + \frac{\sum_{f \in L'} w(f)f}{2 \sum_{f \in L'} w(f)} \right) \right],$$
End