

CSCI 15: AN INTRODUCTION TO THE MODERN INTERNET

Lecture 2: File formats

FILE EXTENSIONS

- .txt .bmp .jpg .mpeg .exe
- What do these do?
- Are they required?

MEASURING NUMBER OF BITS

- Byte, KB, MB, GB, TB
- Byte is 8 bits (why?)

Kilobyte (KB)	1000 Bytes
Megabyte (MB)	1000 KB = 1,000,000 Bytes
Gigabyte (GB)	1000 MB = 1 billion Bytes
Terabyte (TB)	1000 GB = 1 trillion Bytes
Petabyte (PB)	1000 TB = 1 quadrillion Bytes
Exabyte, Zetabyte	1000 or 1,000,000 PB

EVERYTHING IS BITS

- How can we represent text using bits?
- First: what is text made of?
- Then: how do you represent it?

ASCII

- One way to encode text
- A bit out of date, but:
 - Simple
 - Still frequently in use

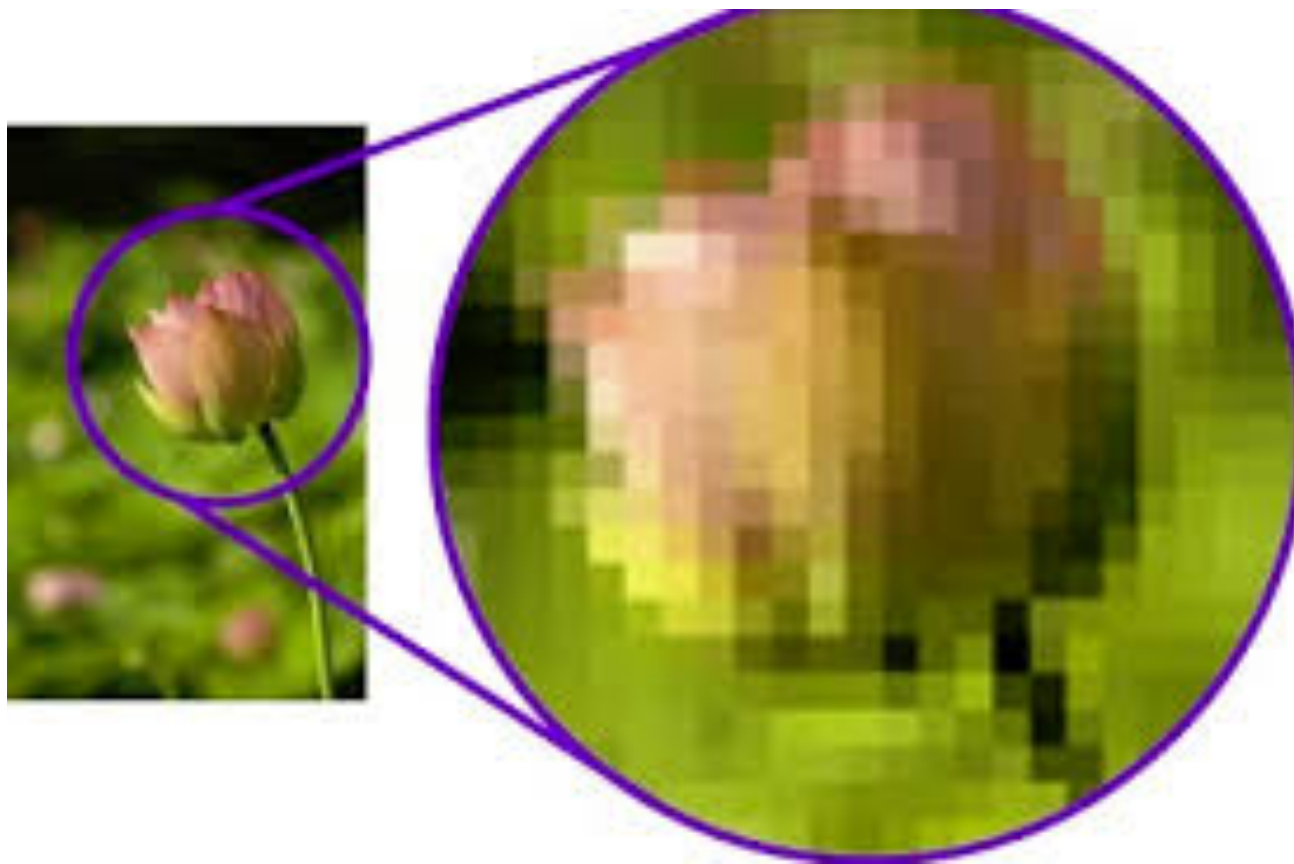
UNICODE

- Much more flexibility!
- The ASCII code for a character is the same as its Unicode code
- <http://www.unicode.org/charts/>

PICTURES?

- Last week we said that colors can be represented as numbers
- How is a picture “made of” colors?

PIXELS



BIT MAP

- How do we store these colors?
- One by one, each row one at a time
- This is called a .bmp file (bitmap)
 - (A .bmp is slightly more complicated in reality)

TRY IT OUT!

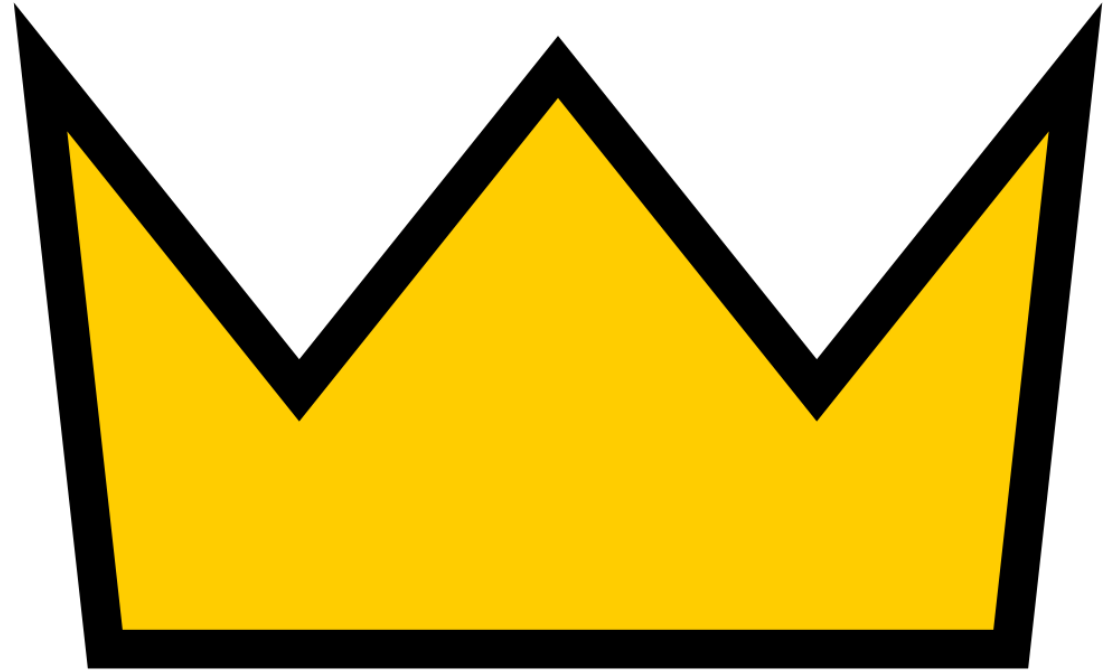
- Download hello.txt and feep.pgm
- Open hello.txt. What should its bits look like?
- Open feep.pgm. How should it be stored?
 - Hint: larger numbers are brighter

HOW TO CHECK

- Right click on folder, click “open in terminal”
- `xxd -b hello.txt`
- `nano feep.pgm`

WHAT'S WRONG WITH THIS?

- This image is 1024x1024 px
- must store 1 million pixels
- That seems excessive!



COMPRESSION

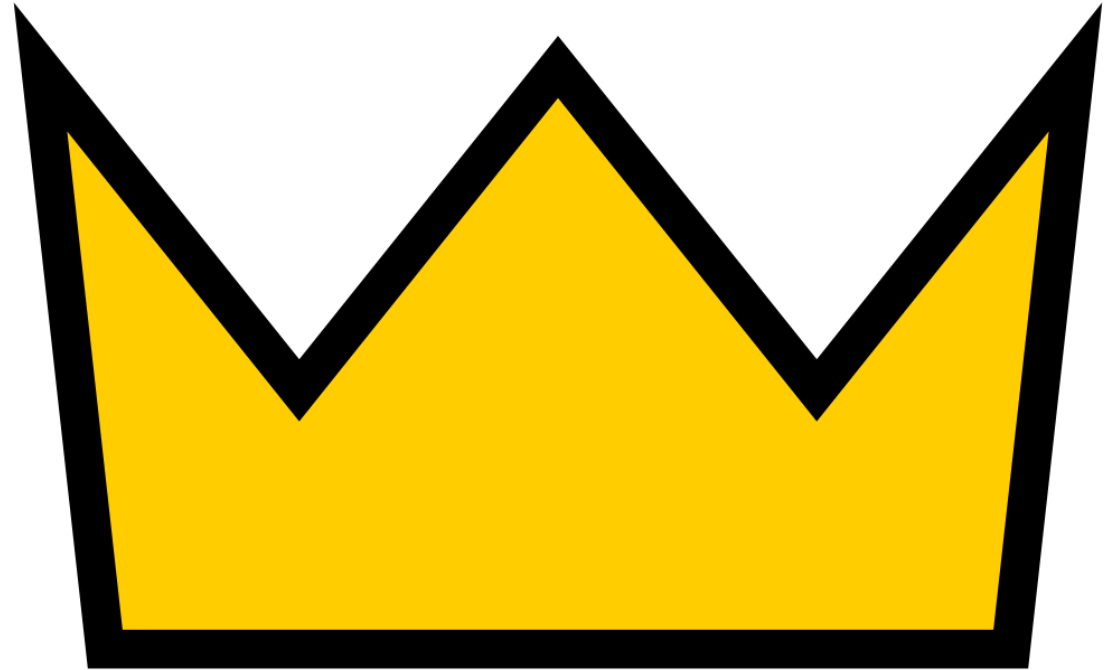
- Making files smaller
- That crown is only 25Kb (~25,000 bytes) rather than 1,000,000 bytes
- The magic of compression!

TWO KINDS OF COMPRESSION

- **Lossless: don't lose information**
 - Can get the exact picture back
- **Lossy: do lose information**
 - Called a “compression artifact”

LOSSLESS COMPRESSION

- How could you store this picture efficiently?
 - Store lines instead of pixels, or
 - Take advantage of repetition



LOSSLESS COMPRESSION

- Lots of modern ways to do this
- Zip files, for example
- Let's compress some text losslessly!
 - gzip/gunzip
 - mobydick.txt

LOSSY COMPRESSION

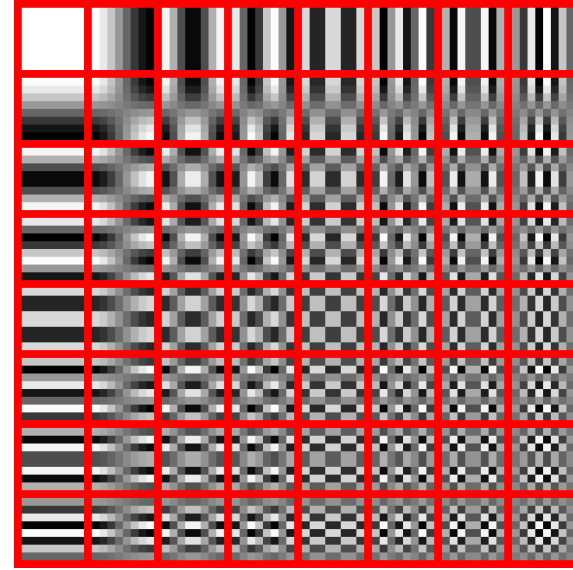
- Why would we WANT to lose stuff???
- Files can get MUCH smaller
- We usually won't miss it



JPEG

- Lossy method for image compression
- Makes files really really small.
- that's 19KB!

HOW DOES JPEG WORK?



- Divides picture into small “blocks”
- Makes the blocks out of a combination of waves
- Looks for “structure” in the blocks
 - Repeated blocks are definitely good structure
 - Or blocks that are the same except a different shade
 - Or similar blocks in less-obvious ways
- Tunable – how lossy do you want to be?

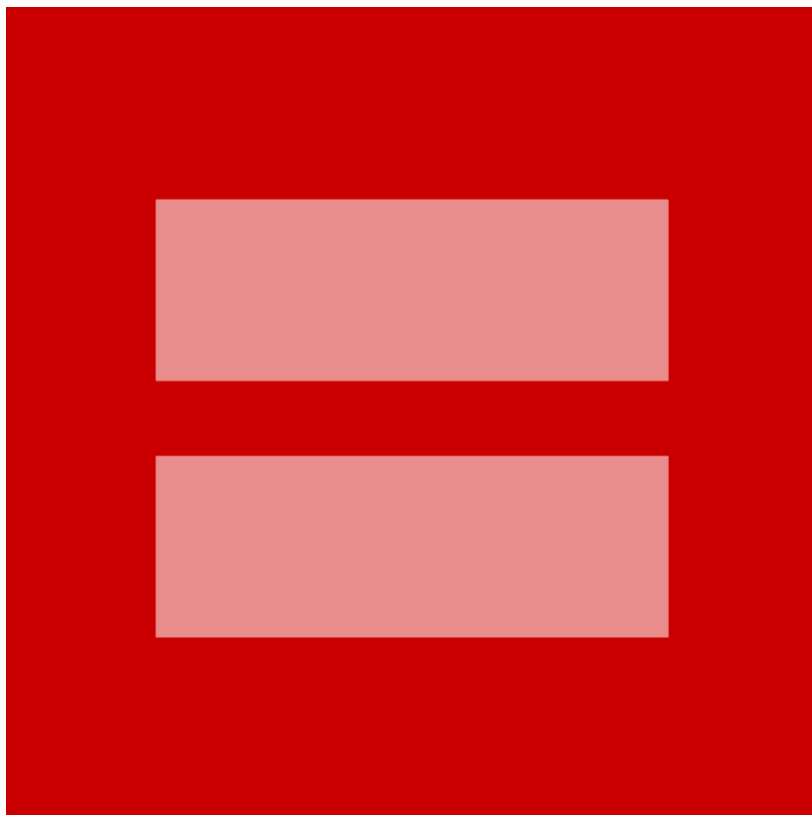
LOSS IN JPEG

- So what do we lose then?
- JPEG tries to get blocks to repeat
 - Get “blocky” artifacts
 - Does badly on borders (why?)

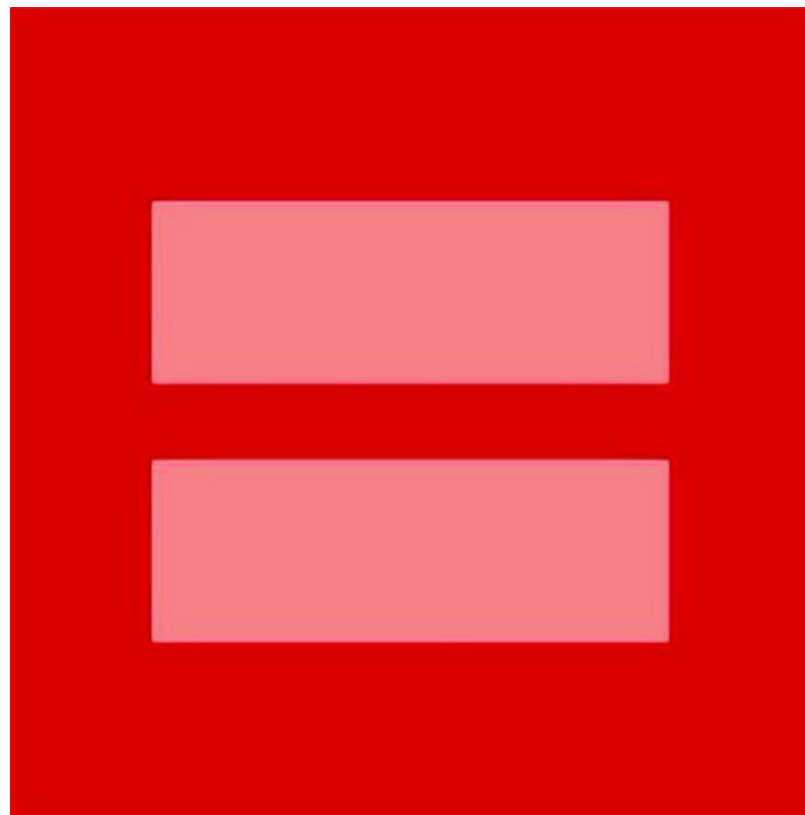
EXAMPLE OF JPEG COMPRESSION



CAN GET REALLY BAD



Original



Compressed

A screenshot of a Facebook post. At the top right, there is a close button (X). The post is by **George Takei**, who has liked this page and posted 8 hours ago. Below the name are options to Like, Comment, and Share. The post has 6,608 likes and 11,655 shares. There are 6 of 422 comments visible. The comments include:

- Allie Bish: "I'm catholic but I still think same sex marriage should be legal. 😊 Everyone was created equal and everyone should have the right to get married no matter what gender. 😊" (14 minutes ago, 9 likes)
- Elaine Brockmeier: "Love is Love!" (13 minutes ago, 1 like)
- Riley D Gin: "what does it mean" (7 minutes ago, 0 likes)
- James Robbin Cochrane: "Equality, Riley!" (3 minutes ago via mobile, 0 likes)
- Scott Murfitt: "is this in support of red velvet cake?" (3 minutes ago via mobile, 1 like)
- CJ Johnston: "Marriage", in every state, establishes, between 2 people, a WHOLE BUNCH of rights and responsibilities and creates a legal entity that sets up rules about how the government must treat that "community"..... That is all it does and there is no reason L... See More a few seconds ago - Like

At the bottom, there is a text input field for writing a comment.

CAN GET REALLY BAD (WHY THIS EXAMPLE?)



OTHER MEDIA

- How do you store audio? A video?
- Compression gets VERY important for video
 - Uncompressed HD video takes about 1GB/second
 - Can't stream that... (your connection is about 1MB/second)
- Same basic idea as JPEG to compress video
- 10min HD Youtube video: 250MB to 300MB

I WANT TO EMPHASIZE

- There is no modern internet without compression

METADATA

- What is metadata?
- Why does it matter?
- When is metadata stored?

METADATA

- What does this photo say about me?



STENOGRAPHY

- What is it?
- We'll talk about it a little more when we get to cryptography.

(IF WE HAVE TIME) BITS ON THE INTERNET

- Bits we send on the internet need a format too!
- What do we need to include?
 - Actual data (payload)
 - Format we're using (there's a bunch!)
 - Where it's going
 - Where it's from
 - Length of the payload
 - Checksum (don't want errors in where it's going!)

HOW CAN WE SHARE A MEDIUM?

- Let's say many devices are all using the same connection
- Only one can broadcast successfully at a time
- How can you make sure things get sent?
- Don't know anything about the network (except that we'll all agree on one protocol)