CS 100: Compression

Chris Kauffman

Week 9-1

Logistics

Feedback

- Midterm overall scores on Blackboard
- Advisory Grades to be posted tomorrow
- Feedback survey results

Today

- HW 5 Overview
- Compression
- Perhaps The Internet

HW 5 Up Now

- Due next week Friday
- Simple Encryption Problems
- Creating your own personal web page
- Individual Work but can help each other

Reading

- "Pattern": Chapter 6
- "Zyante": Ch 5

Midterm Feedback Surveys

- Results posted online
- Review quickly
- Overall good

HW5: Overview

Two Parts

Problems 1-3 Practice a little encryption

Problems 4-5 Get your personal web page made and on the web

Problem 4-5

- Upload stuff to your Mason web space
- Must use some new and interesting tools
  - PuTTY on windows, Terminal on Mac, SFTP on both
- Thursday walk through some of the steps together (bring computers)
- Visit HW 5 section on uploading to `mason` and read

From Last Time

- Caesar and Vigenere Ciphers?
- How are password files stored?
- How are passwords cracked?
- How do you prevent passwords being cracked?

Compression

- Make something smaller
- In computing, make a file smaller
- Lossless Compression: don't lose any information
- Lossy Compression: discard some information to make smaller

Text to Compress

> *The Time Machine by H.G. (Herbert George) Wells*
>
> *The Time Traveller (for so it will be convenient to speak of him) was expounding a recondite matter to us. His grey eyes shone and twinkled, and his usually pale face was flushed and animated. . . .*

Full text available from Project Gutenberg

- `The-Time-Machine.txt`
- 32354 words
- 179377 characters (bytes)
- 176 Kilobytes

Could you make it smaller? By how much? How?

- Discuss

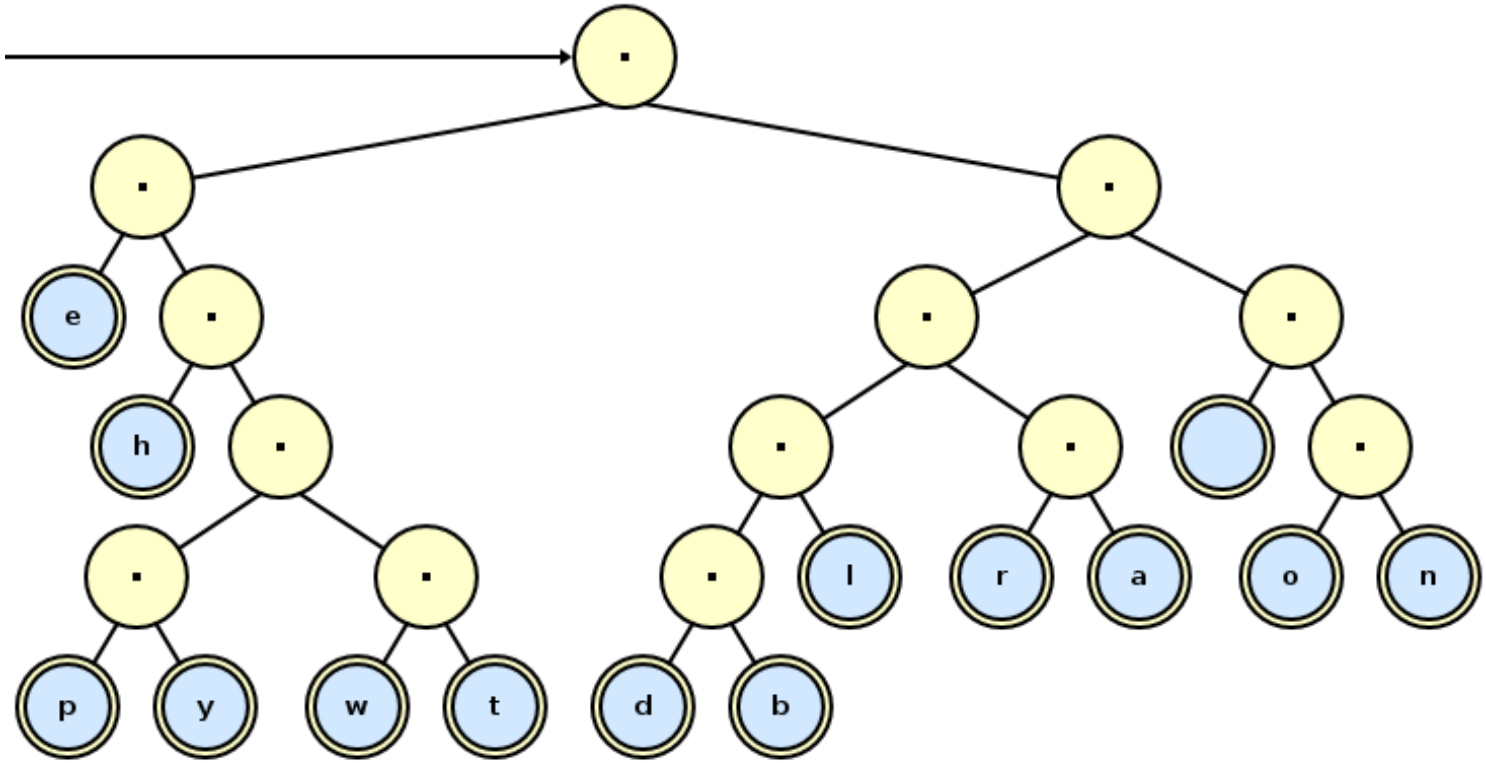
Compress Some Simple Text

Source: Hjoranna on Deviant Art

```
abandon all hope ye who enter here
```

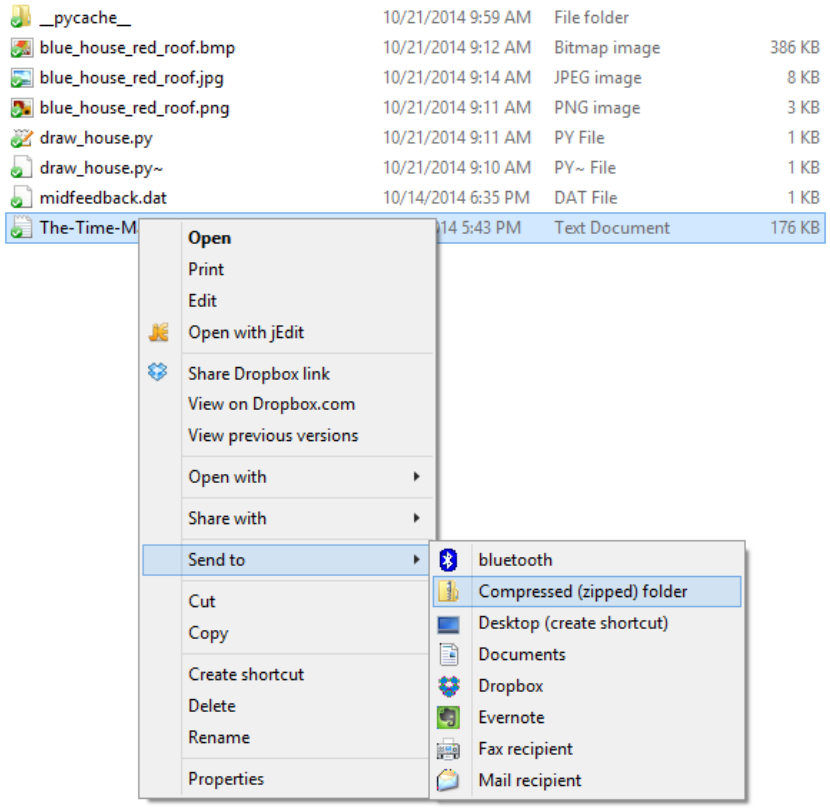
- Which are the most frequent characters?
- How would you make the message shorter

Huffman Encoding



Char	Freq	Code	Char	Freq	Code
e	6	00	p	1	01100
_	6	110	y	1	01101
l	2	1001	w	1	01110
r	2	0100	t	1	01111
a	3	1011	d	1	10000
o	3	1110	b	1	10001
n	3	1111			

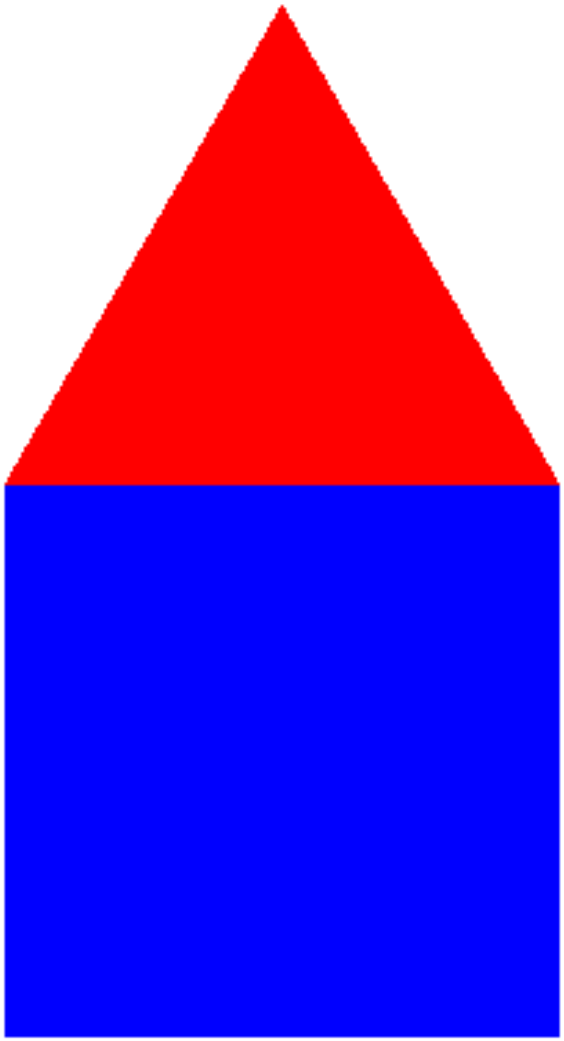
Practical Compression



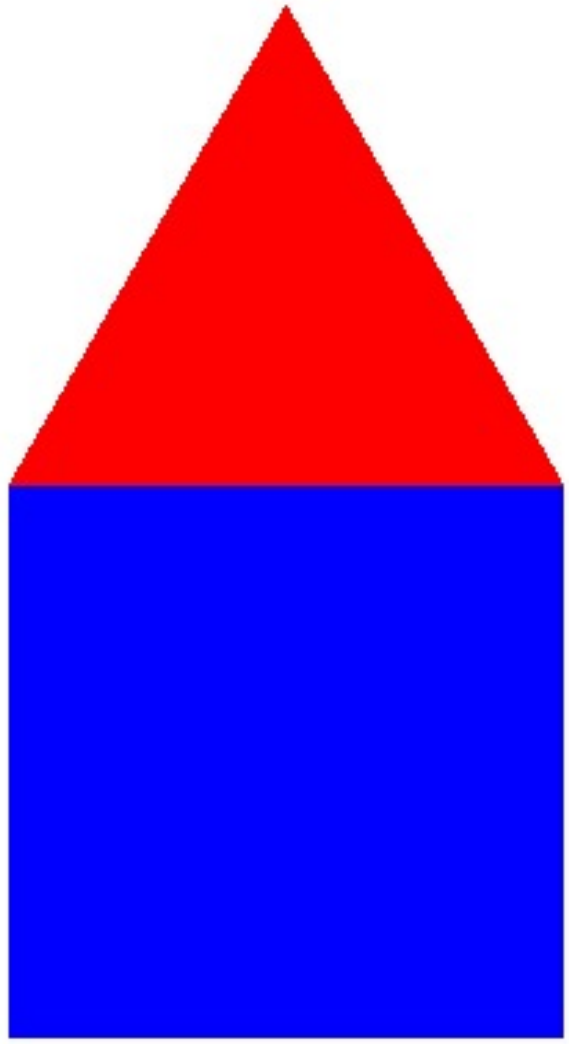
- Lots of programs do it for you
- Most common is `Zip` programs
- The `.zip` extension denotes a compressed
- Usually want to *extract* the contents to work with it
- Zip is a lossless compression technique
- Usually use lossless compression with text

Image Compression: Spot the difference

PNG/BMP



JPEG



Compressing Images: Lossy

10x Compression: 39kb

50x Compression: 9kb

Sound and Video Compression

- Sound is a continuous phenomenom and is always discretized for digital storage
  - Not so for records and magnetic tapes
- `WAV` or `.wav` files are typically raw audio
  - Files a typically very large,
  - 70-minute CD would take about 600 MB of space
- `FLAC` is a lossless audio compression format
  - Can cut down size of files by 20-30%
  - Quality of sound does not suffer
- `MP3` and `AAC` are lossy compression formats
  - Can fit a whole CD into 30-40 MB
  - Quality of sound does change
  - Can choose the level of quality when compressing
  - Higher quality requires more space
- `MP4` is a lossy video format
  - Combines techniques from MP3 and JPEG
  - From one picture to the next, records changes in pixels
  - Often compresses raw video/audio by 100x with little perceptible loss in quality

Information Density

```
176K The-Time-Machine.txt
 70K The-Time-Machine.zip
```

- Which one has more information?
- What is information?

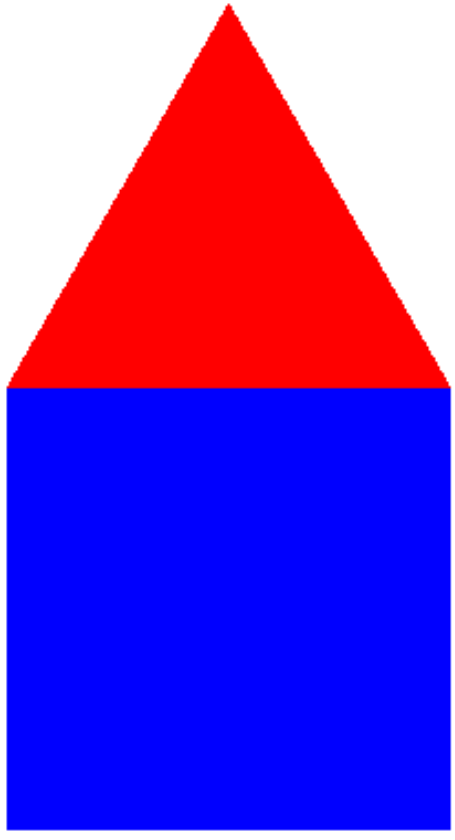
Small Programs that Generate Pictures

Which is smaller?

Python Code

```
from turtle import *
def house(size):
  color("blue")
  begin_fill()
  for i in range(4):
    forward(size)
    right(90)
  end_fill()
  color("red")
  begin_fill()
  right(300)
  for i in range(3):
    forward(size)
    right(120)
  end_fill()
# Draw a big house
house(200)
hideturtle()
```

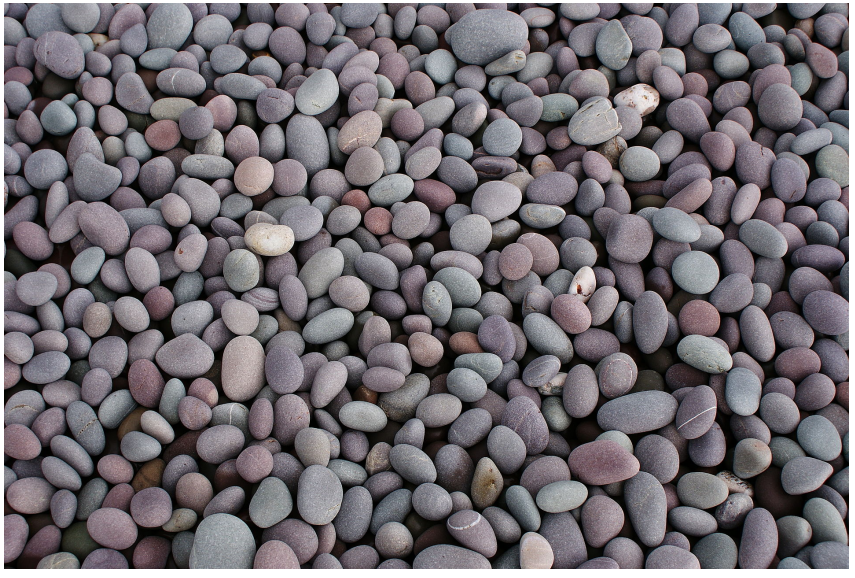
Picture



An Interesting Measure of Information Density

How small is the smallest computer program to re-create the data?

- For *The Time Machine*, would look something like Huffman Encoding
- *Pattern on the Stone*: For pebbles on the beach, things get interesting
- How exact must the recreation be?



Source

How does the Internet work?

- A lot of wires
- An older analogous system
- An interesting set of problems