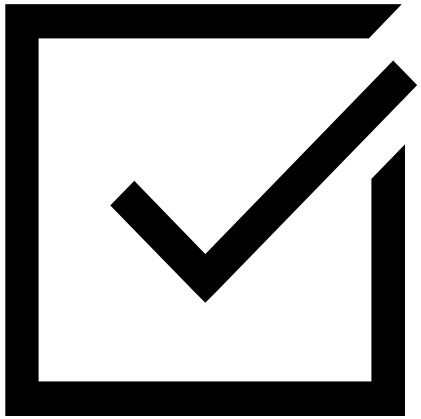
CPR 101

File Compression

Agenda

- Quiz
- News
- Discussion
  - What is file compression
  - Why it is used
  - How compression works
  - Types of Compression formats
  - Loss-less and lossy compression
- Activity

Quiz



News of the Week

What is File Compression?

What is File Compression?

File compression is the process of using an algorithm to reorganize the data structure in a file to reduce the amount of disk space that a file uses.

Why do we use file compression?

Why did we use file compression?

- Historically, data desired exceeds available storage space
- Today, this is not as much of an issue as storage drive space is readily available and considerably cheaper than it used to be

What about data transfer?

- Why is file compression relevant with data transfer?

Case Study

- Let's say you are working on a web application for **30,000** users
- Your total data size is **12,320,982** bytes.
- You decide to add some new content and data, and subsequently, your data size increases to **12,825,603** bytes.

Starting Size:	**12,320,982 bytes**
Current Size:	**12,825,603 bytes**
Difference:	504,621 bytes

Case Study, part 2

- Your application's size is now approximately ½ MB larger

Starting Size:	12,320,982 bytes
Current Size:	12,825,603 bytes
Difference:	**504,621 bytes**

Is this a big deal?

- On your computer, not really, but…

Case Study, part 3

- On your website your application will be delivered to **30,000** users.
- That means, that the total amount of data sent through your network connection could be as much as **15 Terabytes!**

Starting Size:	12,320,982 bytes
Current Size:	12,825,603 bytes
Difference:	504,621 bytes
# of Users:	**30,000**
Total data difference:	**15,138,630,000 bytes**

Case Study, part 4

- If we applied some compression to our update and reduced the data to 80% of it's initial size, that would translate in to a savings of **3 Terabytes** in network traffic**!**

Starting Size:	12,320,982 bytes
Current Size:	12,825,603 bytes
Difference:	504,621 bytes
80% Compressed:	**403,697 bytes**
# of Users:	30,000
Original data difference:	15,138,630,000 bytes
Compressed data difference:	**12,110,910,000 bytes**
Savings through compression	**3,027,720,000 bytes**

How does compression work?

- Here's a paragraph:

  Data compression is particularly useful in communications because it enables devices to transmit or store the same amount of data in fewer bits. There are a variety of data compression techniques, but only a few have been standardized. The CCITT has defined a standard data compression technique for transmitting and a compression standard for data communications through modems. In addition, there are file compression formats, such as ARC and ZIP.

- This quote contains 449 characters.

How Compression Works (cont'd)

- Replace "compression" with ♠. Now, the quote becomes:

  Data ♠ is particularly useful in communications because it enables devices to transmit or store the same amount of data in fewer bits. There are a variety of data ♠ techniques, but only a few have been standardized. The CCITT has defined a standard data ♠ technique for transmitting and a ♠ standard for data communications through modems. In addition, there are file ♠ formats, such as ARC and ZIP.

- Including the dictionary "♠=compression", the total size is now 411 characters.
- Compression algorithms build a token/string dictionary

How Compression Works (cont'd)

- With more pattern matching and a bigger dictionary, our quote becomes (♠=compression , ♣=here are , ♦=communications , ♥=data , ☺=standard, ☼=transmit, ☯=technique)

  ♥♠is particularly useful in ♦because it enables devices to ☼ or store the same amount of ♥in fewer bits. T♣ a variety of ♥♠☯s, but only a few have been ☺ized. The CCITT has defined a ☺ ♥♠☯ for ☼ting faxes and a ♠☺ for ♥♦through modems. In addition, t♣ file ♠formats, such as ARC and ZIP.

- Including the dictionary, the total size is now 361 characters or 80% of original size. The compression advantage increases with the length of the text, i.e. more pattern matches.

Compression file formats

- Do you know any compression file formats?

Compression file formats

ZIP

- The most popular general-purpose compression file format
- Compresses a file or set of files and places them inside another file, with a *.zip* extension
- Included with all modern versions of Windows
- Includes features such as password-protection, …. . . .

Compression file formats

RAR, 7z, TAR, StuffIt

- There are many other general-purpose compression file formats
- These use different algorithms, with various benefits and uses
- Some are designed for different operating systems (StuffIt for Mac, TAR for *nix)

Compression file formats

JPG, PNG, GIF, TIF, TGA

- Compression formats used by the **graphics industry**
- Most of these are lossy formats (TIF offers a ZIP compression option), can not actually be uncompressed to exact source data
- Windows Paint can open and save to some of these formats

Compression file formats

MP3, WMA, MP4, FLAC

- Compression formats used by the **sound engineering and music industry**
- Most of these are lossy formats (although FLAC stands for Fully Lossless Audio Codec), can not actually be uncompressed to exact source data
- Windows Media Player can open many of these formats

Compression file formats

MPG, MP4, DIVX, XVID, MOV, AVI

- Compression formats used by the **video industry**
- Most of these are lossy formats
- Will often mix compression algorithms from audio and image technologies
- Windows Media Player can open many of these formats

Loss of data?

Can we lose data when compressing a file?

There are two basic types of compression:

- Lossy
- Loss-less

Loss-less Compression

- *ZIP*, *TIF*, *FLAC*, and other general file compression routines are considered *lossless* compression.
- The data is always complete; after decompression, no one would be able to determine that the data was compressed, it's all there and useable.

Lossy Compression

- *JPG*, *MPG*, *MP3*, *GIF*, and many other specific end-user formats are all considered *lossy* compression.
- These formats, in order to achieve compression, actually remove data from the source file.
  - GIF images reduce the number of colours in an
  - JPG images effectively delete colour information to achieve compression
  - MP3s simplify the sound waves of audio
- This means that once an image is compressed it can never be returned to a full data uncompressed format
- The amount of loss can usually be adjusted, so a developer can decide how much compression is necessary and what quality of the content is still required.

Lossy vs. Loss-less Compression

Loss-less

Lossy (GIF/JPG)

Lossy vs. Loss-less Compression

Loss-less

Lossy (GIF/JPG)

Backups – Why and What

What is a Backup?

What is a Backup?

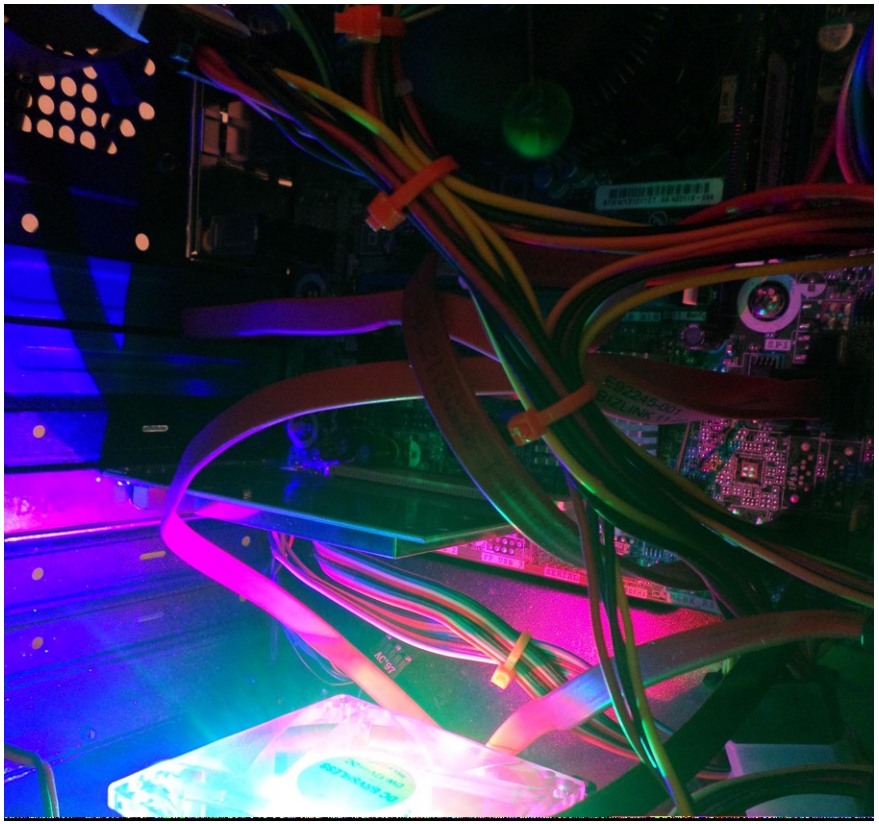
A backup is a copy of content, created as a contingency, in case something should happen to your initial content

What can happen to your work?

- What do we mean by "*created as a contingency*"?

Events that can occur:

- Hardware failure
  - What happens if you hard drive fails, or your motherboard experiences a power surge?



Events that can occur:

- Bus-Raptor incident
  - If a team member gets eaten by a raptor, will you still be able to access their work?

Events that can occur:

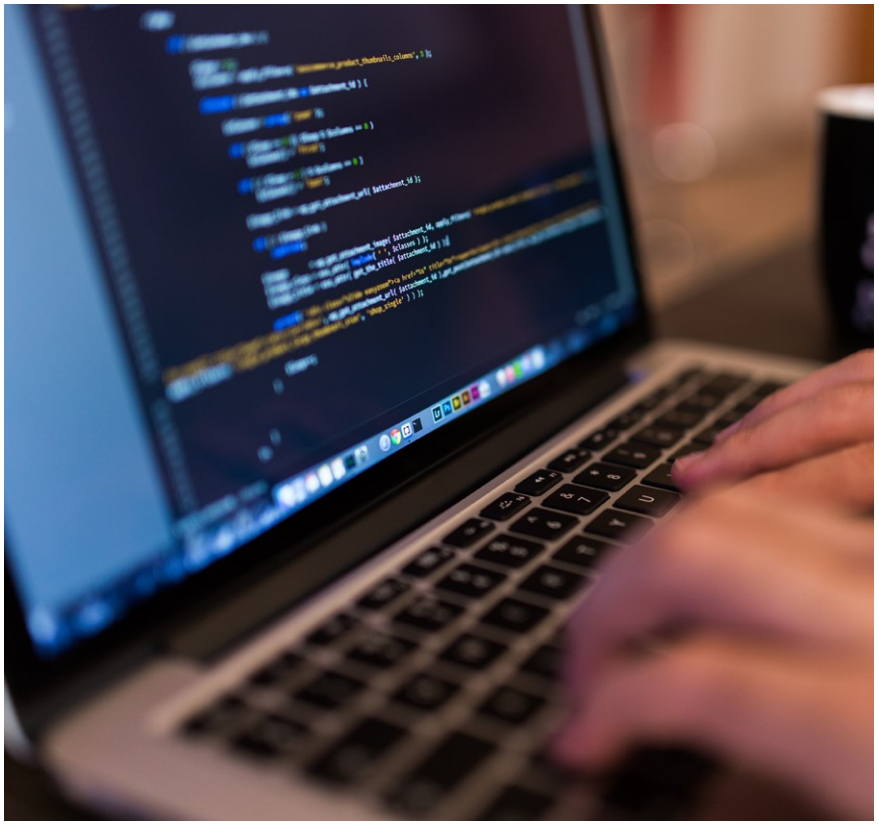
- Missing employee
  - Angry employees might damage equipment or delete data.
  - Employee might be unavailable, how do you get access to the files you need?

Events that can occur:

- Theft
  - This is a reality in business – internal and external theft are common ways to lose data

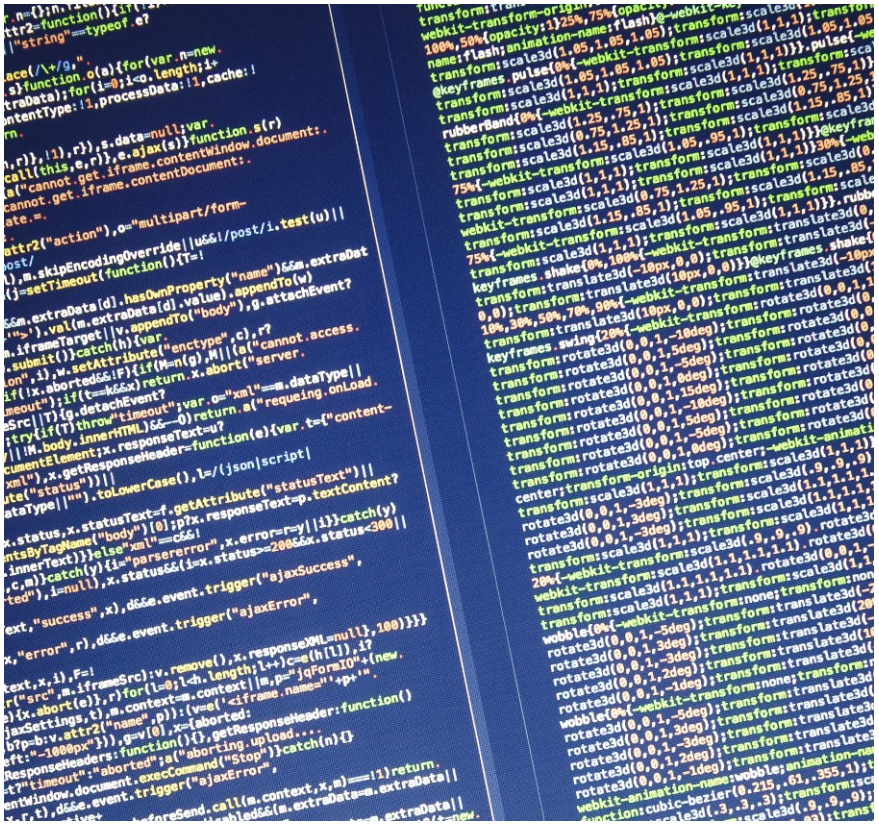
Events that can occur:

- Loss
  - What if you lose your laptop?



Events that can occur:

- Hacking/Cracking
  - High-profile or valuable data is a common target for a hacker – they may delete your data, change it, or hold it for ransom!



Types of Backups

A Simple Solution – Backups!

- Having a backup will allow you to recover from lost, broken or stolen hardware or software, and help you get your system running quickly and effectively.
  - Get it the habit of storing only “Work in Progress” on your laptop, or PC.
  - Once your work is completed: favourite playlist, pictures, school assignment, letters/reports/programs, then copy them to a backup device.
  - This creates **Redundancy**

Types of Backups

Automatic Backup

- Backups are created automatically, by a program running on the computer. You will essentially be saving twice; every time you save, a backup will be made.
- Mirrored RAID system is also a common option here.
- Useful for incremental changes

Types of Backups

Scheduled Backup

- A system operator or software tool performs backups at specific times, such as every night.
- Preferred method when there is a lot of data to backup, that could take many hours

Types of Backups

Manual Backup

- A user performs backups at their own convenience
- Least effective method – if you forget to do it, it doesn't happen!
- Better than no backup at all though!

Backup Media

What backup media can we use?

- What can we store our backups on?

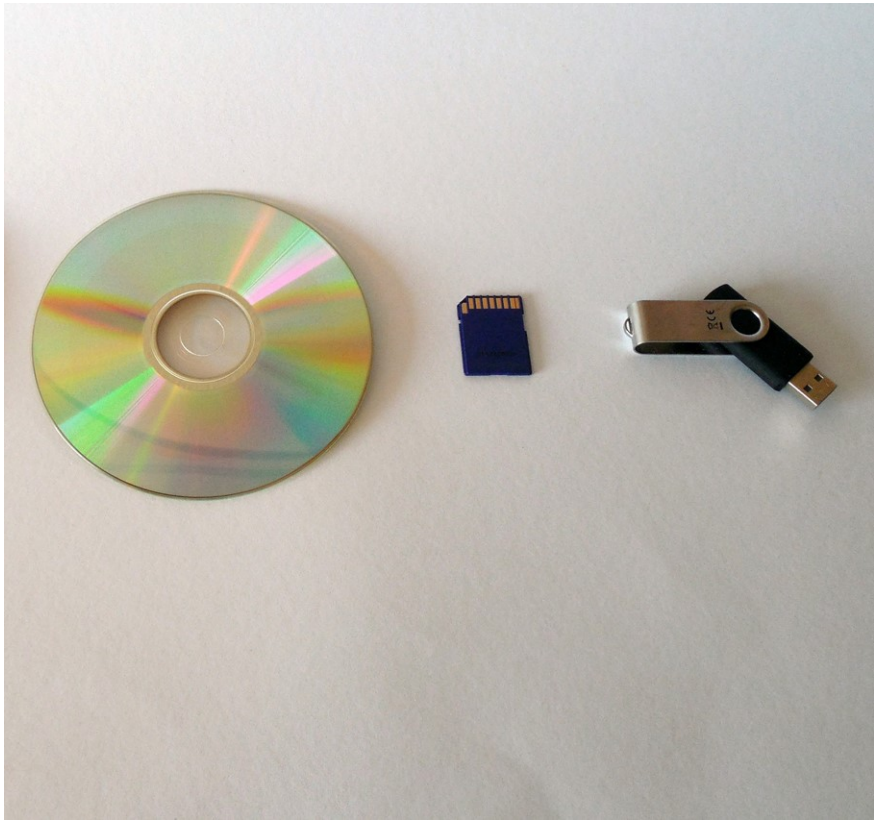
Backup Media – Local

- Stored on a hard drive in use by the system
- RAID Drive
- Fastest method
- Convenient
- If computer is lost, so is the data!

Backup Media - External

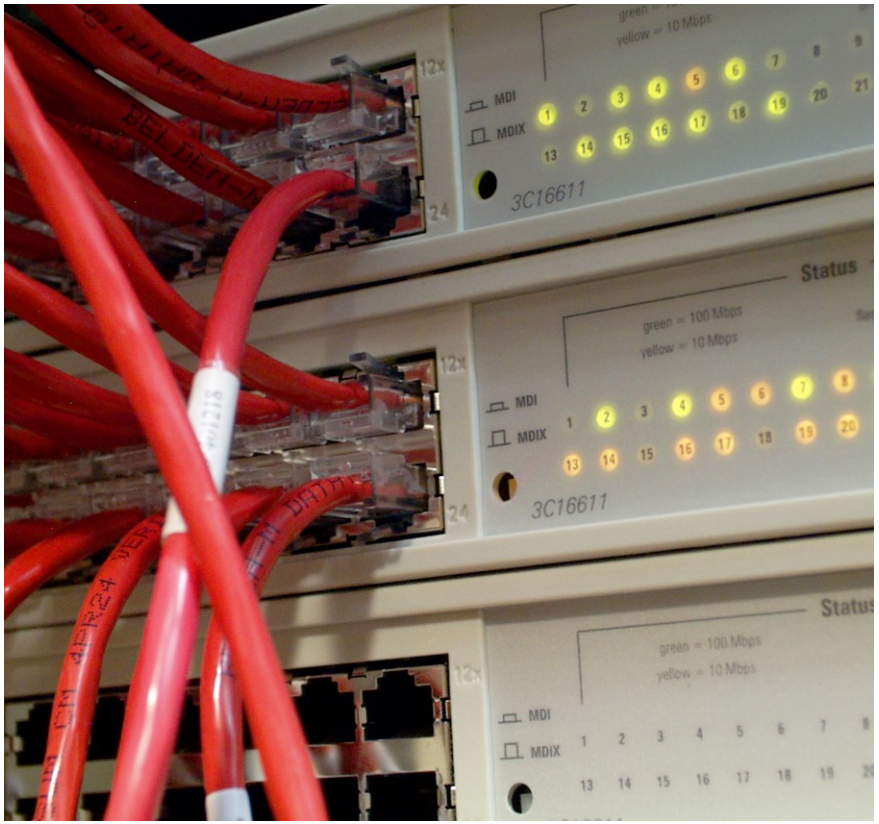
- External Drive / USB Stick
- DVD
- Tape Drive

- Useful if computer stops working
- Can be stored off-site
- More expensive!



Backup Media – Network

- FTP server
- Google Drive / Carbonite / Dropbox / iCloud / others
- Which country is the data stored in?
- Price for service
- Encrypted?
- Slower



Which method is the best?

- Which of these three methods (Local, External, Network) should you use?

Use all three methods!

- Depends on the type of work you're doing
- Using all three methods together is best
  - One local copy that you're working on
  - One external copy on a USB stick or something similar
  - One copy off-site, in case of disaster

3-2-1 Backup Checklist

- **3 copies** *(actions on active file do not change copies)*
  - 1 active file on your machine, 1 local backup, 1 remote backup
- **2 different formats/platforms** *(platform independence)*
  - External drive using File History / Time Machine *
  - One-way backup to cloud (not two-way sync)
- **1 off-site backup** *(geographically separate location)*
  - Cloud storage different from your cloud IaaS, PaaS, SaaS provider
  - rotating external drives from home to office

** File History/Time Machine are not completely platform independent*