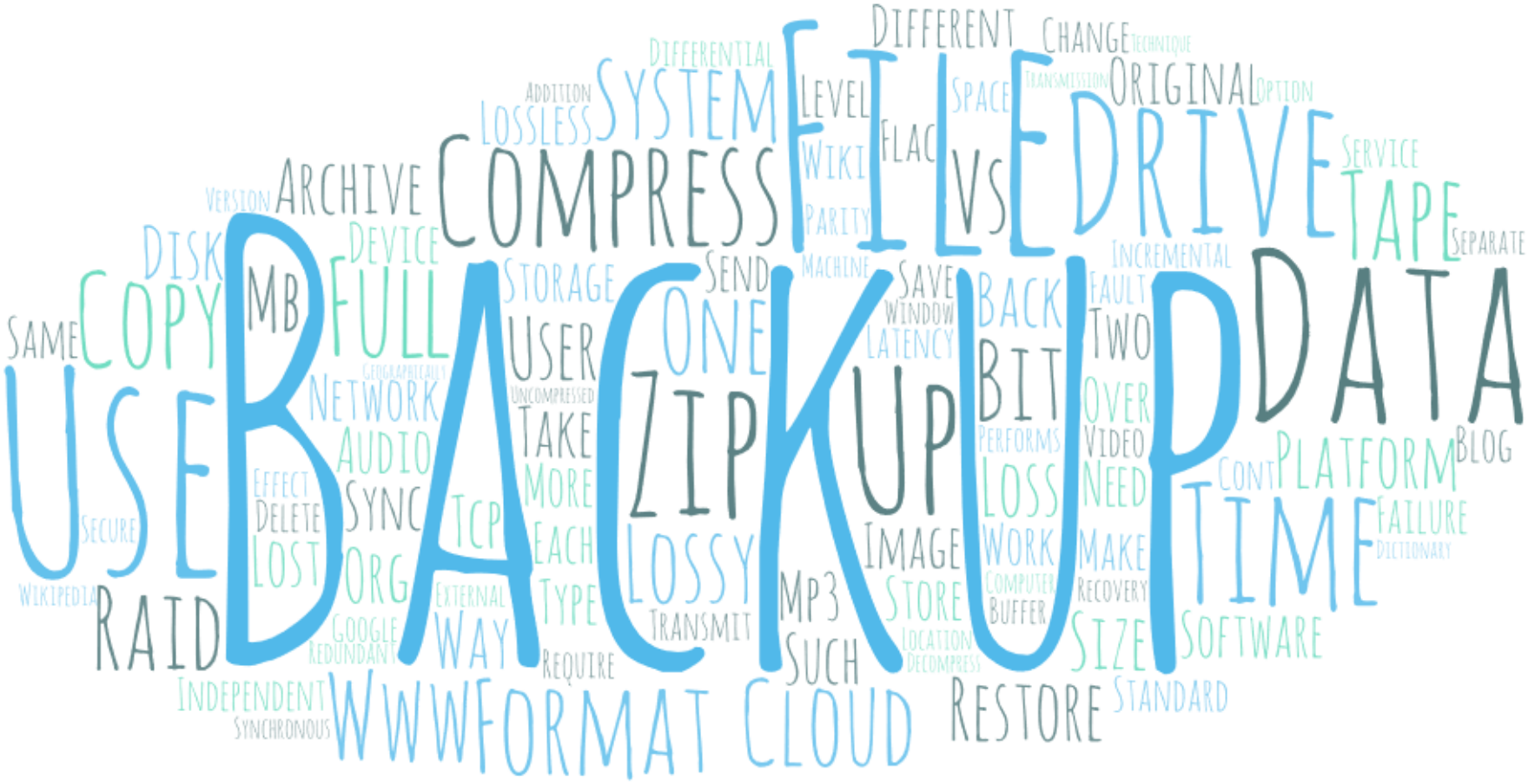
COMPUTER PRINCIPLES FOR PROGRAMMERS

File Compression, 3-2-1 Backup with 4QFeybBPWkeiOg==

News of the Week


BACKUP
USE
DATA
FILE
DRIVE
TAPE
TIME
COMPRESS
SYSTEM
ZIP
ONE
FULL
COPY
DISK
ARCHIVE
RAID
WWW
FORMAT
CLOUD
RESTORE
SOFTWARE
PLATFORM
DIFFERENT
CHANGE
ORIGINAL
LOSSLESS
LOSSY
NETWORK
AUDIO
SYNC
TCP
WAY
ORG
LOST
DELETE
EFFECT
SECURE
WIKIPEDIA
INDEPENDENT
SYNCHRONOUS
REDUNDANT
GOOGLE
EXTERNAL
REQUIRE
TYPE
EACH
MORE
TAKE
USER
STORAGE
DEVICE
VERSION
SAME
MB
DIFFERENTIAL
ADDITION
LEVEL
WIKI
PARITY
FLAC
VS
SPACE
TRANSMISSION
OPTION
SERVICE
SEPARATE
INCREMENTAL
FAULT
TWO
BACK
BIT
OVER
VIDEO
PERFORMS
LOSS
NEED
CONT
BLOG
FAILURE
DICTIONARY
MAKE
WORK
IMAGE
STORE
LOCATION
DECOMPRESS
COMPUTER
BUFFER
RECOVERY
SIZE
STANDARD
SUCH
MP3
TRANSMIT
SEND
MACHINE
SAVE
WINDOW
LATENCY
UNCOMPRESSED
GEOGRAPHICALLY

Agenda

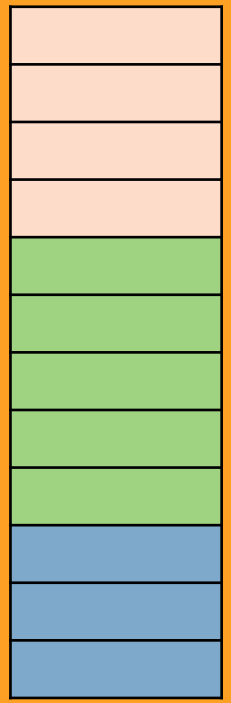
Lecture:

1. What, Why, and How of “File Compression”
   … effect on data transfer
   … drawbacks
   … overview of formats
   … Lossless vs Lossy
2. What, Why, and How of “Backup”
   … types of backups
   … backup media

Agenda (Cont'd)

- Activity:
  1. Explore File Compression
  2. Compress various native file formats within a ZIP archive and compare the compression factors
  3. Upload files to Blackboard to demonstrate a network backup

What is “File Compression?”



4 pink
5 green
3 blue

What is “File Compression?”

- storing a file’s data in “less space” by “minimizing redundancy” in the content
- An **archive** is a collection of folders and files stored in one file, e.g. *filename*.ZIP
  - Files are usually compressed
  - Files can be encrypted
  - Cross platform exchange
- OS options to compress / encrypt local files

Why use File Compression?

- Writing / Sending data takes bandwidth and I/O time
  - Tape – not as obsolete as you might think with Linear Tape-Open
  - NAS – Network Attached Storage (local or remote)
  - USB – removable drives (ad hoc, user level)
  - FTP or Cloud Storage AWS Glacier, Google Nearline
- Encrypt off-site data for security
  - compression software has a password encrypt option
- SSH (Secure SHell) can compress data: asynch.
- VoIP must do this in real-time, synchronously.
- Streaming sends compressed, receiver decompresses.

Effect of File Compression on Data Transfer

Assumptions:

- 1MB plain text file, unique for each of 30,000 users
- Network throughput is 2 seconds per file
- text compressed to 35% of original, throughput 1 sec/file

	Data	Time	Size
Original 1MB plain text	8.24 Mb	16.6 hours	241.4Gb
Compressed to 35%	2.88 Mb	8.3 hours	84.5Gb

How File Compression works

Compression combines:

- **matching and replacement of duplicate strings with pointers.**
  - Lempel–Ziv–Welch (LZW) compression (1984)
- **replacing symbols with new, weighted symbols based on frequency of use.**
  - Huffman coding (1952)
  - David A. Huffman was a Ph.D student at MIT

Measuring Repetitive Lyrics With Compression

Are Pop Lyrics Getting More Repetitive?

Lempel-Zi-Welch compression measures repetitive lyric sequences.

```
baby I don't need dollar bills to have fun tonight
I love cheap thrills!
baby I don't need dollar bills to have fun tonight
I love cheap thrills!
I don't need no money
as long as I can feel the beat
I don't need no money
as long as I keep dancing
```

0.0% Size Reduction

Original size: 247 bytes/characters

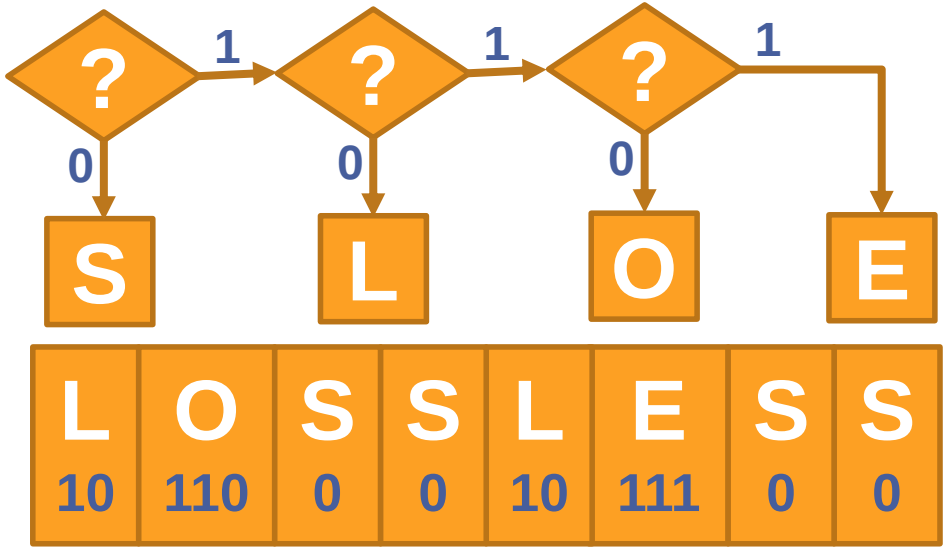
Compressed size: 247 bytes (247 characters + 0 × ●)

How Huffman decoding works

Encoded:

Decode Logic:

? = next encoded bit



8 chars × 8 bits = 64 bits compressed to 14 bits (22%)

Lossless vs. Lossy Compression

Lossless: contains all original data with redundancies removed.

- Data archives, PNG/TIFF images, FLAC/ALAC audio

Lossy: sacrifices quality for smaller file size by dropping fine-grained, subtle details from the original data

- GIF/JPG images, most audio, all video
- for end-use only, not for modification/editing

Lossy vs. Lossless compression

Lossy (GIF), 45.4kb	**Lossless**, 941kb	Lossy (JPG), 25.4kb

MP3/AAC	CD	HiRes
320 kbps	1,411 kbps	9,216 kbps
22.7%	100%	653%

Overview of some Compression File Formats

Data

ZIP 7z RAR
.docx .pptx .xlsx
StuffIt .tar.gz

Images

GIF JPEG
RAW PNG TIFF

ZIP

The Standard

Music

MP3 AAC MQA
WAV ogg **FLAC**

Video

MPG MP4 DIVX
XVID MOV AVI

BOLDed formats are Lossless.

Drawbacks to Compression

- **Time**: compression needs CPU and primary storage resources
  - PCs have lots of both and only one user. Servers on the other hand…
- **Space**: archived files must be uncompressed before use, extra space needed for both compression & decompression
- **Integrity**: any data corruption can cause loss of entire archive
  - Solid or multi-volume archives can be lost with even minor data corruption. Archive repair is possible but not probable.
  - *Test your archives to confirm integrity.*
- **Recoverability**: the Lossy sacrifice is reduced quality
  - Some data is lost after Lossy compression (hope you don't notice)

Why do we need backups?

Accidental deletion by users including IT people

Hardware failure: *all* storage devices eventually fail

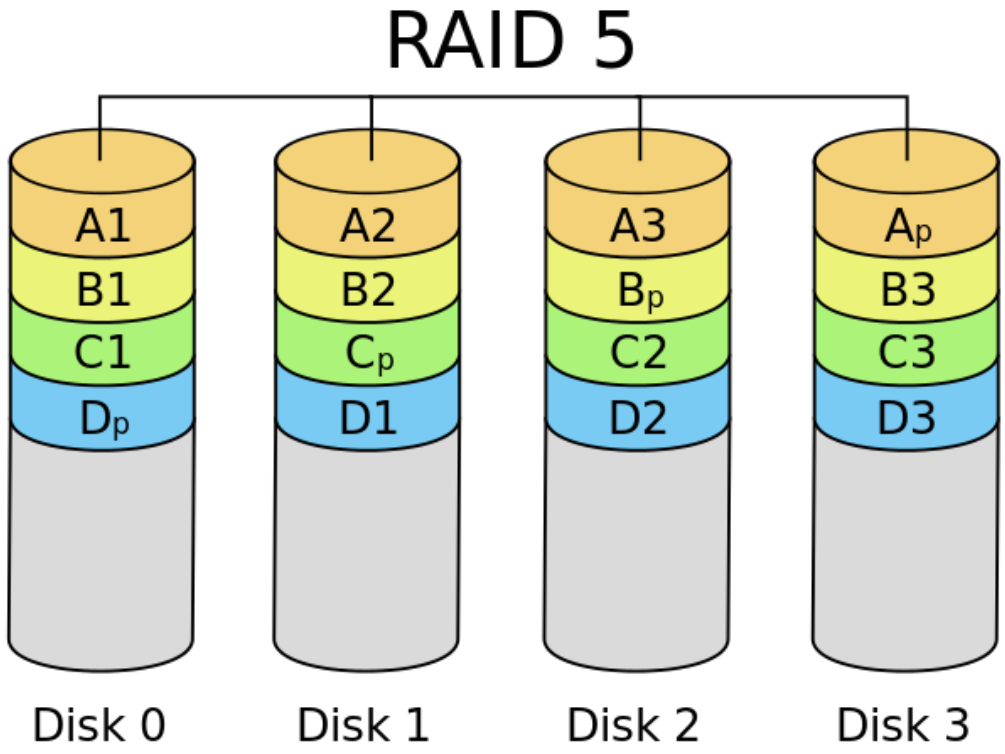
2/3 to 3/4 of all data loss

Less Frequent Causes

- Ransomware infection
- Catastrophe: fire, flood, theft, loss
- SQL injection attacks
- cloud provider's business failure or account closed on cloud system.

Continuous Data Integrity with RAID

- **Redundant Array of Independent Disks**
- RAID 5 or 6 tolerates failure of 1 or 2 drives
- RAID 5 needs +1 drive for parity, RAID 6 needs +2 drives
- RAID appears as one logical drive to OS.
- Provides increased read/write performance with multiple spindles doing parallel I/O



User Level File Recovery ... *is not backup*

Windows File History, macOS Time Machine are not backup

- Automatic copying of files to external or network drive
- Historical versions of user files maintained. Easy to restore.
- Must configure and test to ensure copying of all user folders.
  - *If drive is always connected*, it is not a backup, just a copy; it is likely **not geographically separate** and certainly **not platform independent.**

Windows Recycle Bin, macOS Trash can are not backup

- Only good for *oops!* and short-term recovery.
  - The bin/trash is not a reliable copy much less a backup.

Two-way synchronization is not backup

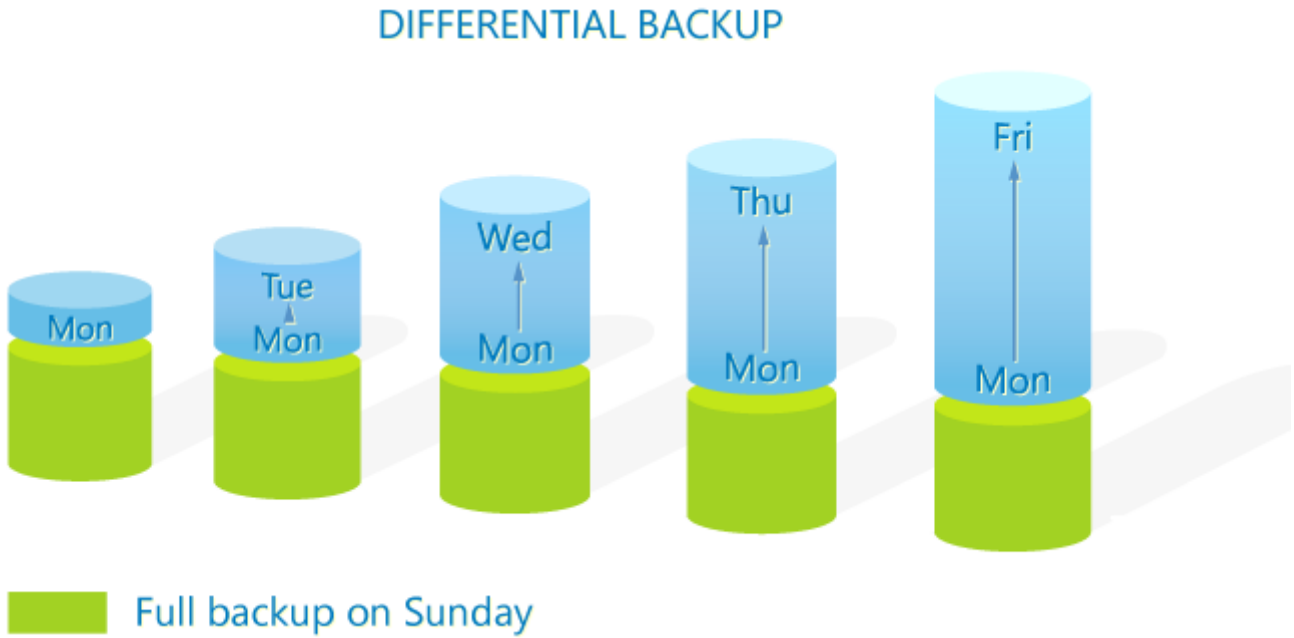
- Synchronization is platform interdependent, not independent.
- A file on one system does not have a "copy" on other systems, the same file **co-exists** on all synchronized systems.

Three characteristics define a Backup

A **copy** in a **geographically separate location** that is **platform independent**.

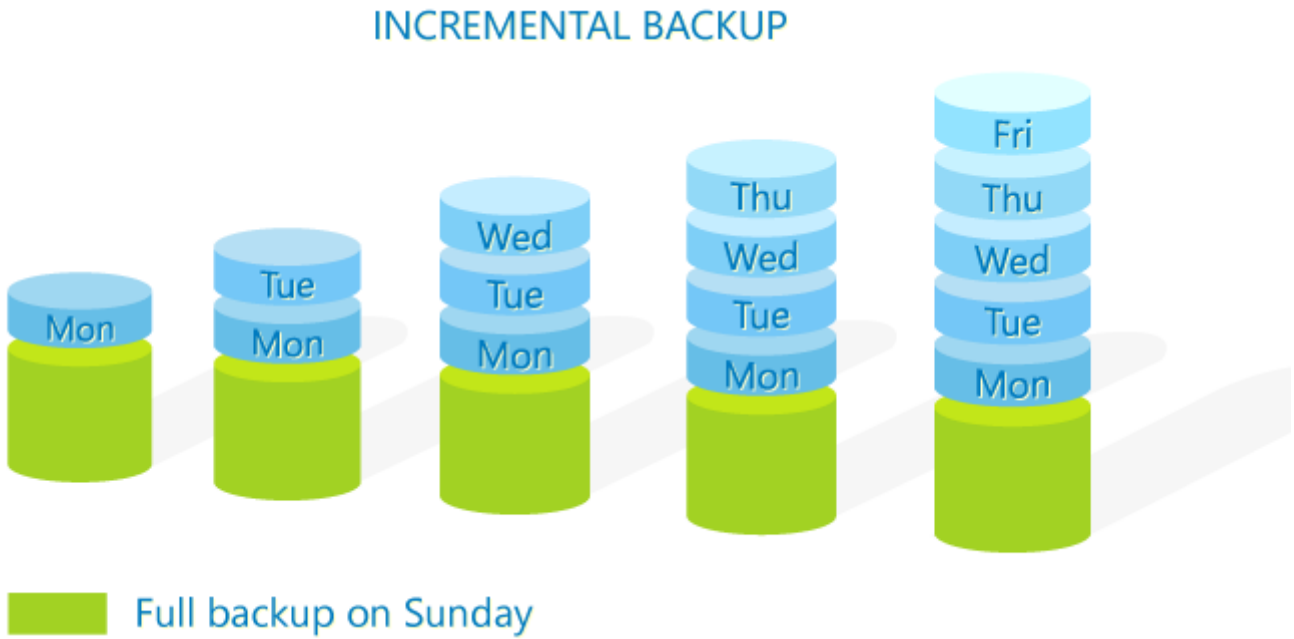
Classic File Backup Types/Strategy

Full + **Differential** (only files changed since last Full backup)
Full backup is slow, Differential backup is faster but gets slower.
Restore from Full + Differential is fastest.



Classic File Backup Types/Strategy

Full + **Incremental** (only files changed since last backup)
Full backup is slow, Incremental backups are much faster than Diff.
Restore from Full + Incremental sets is slowest.



Enterprise backup

- Backup software manages Full, Differential, Incremental strategies
  - Options for file versioning (retaining file generations)
  - Journal snapshot, DB duplication for continuously running systems.
  - Checkpoint processing with save-while-active on IBM Power
- LTO tape or Optical Disc libraries as nearline tertiary storage
- AWS Glacier, Google Nearline, Sync cloud cold storage
  - Inexpensive backup storage but expensive and slow restore
- Recovery and Restoration speed is highly variable.
  - Depends on data transfer rates from backup to server and complexity of rebuilding the relational aspect of data base objects
- Data deduplication and Single-instance optimize storage
  - eliminate duplicate copies of data within and across systems

3-2-1 Backup Checklist

- **3 copies** *(change only the active file, not the backups)*
  - 1 active, 1 local backup, 1 remote backup
- **2 different formats/platforms** *(platform independence)*
  - External drive is platform independent **only when not plugged in**
  - LTO tape or optical disc. Initially local, optionally moved offsite.
  - One-way backup to cloud cold storage (*not* two-way cloud sync)
- **1 off-site backup** *(geographically separate location)*
  - Cloud storage different from your cloud IaaS, PaaS, SaaS provider
  - tape/optical media – rotate Full, Diff, Incr offsite storage services
- The near loss of Toy Story 2

The final word on backups...

Backups do not matter.

Only RESTORE matters.

NOTES

…not on the quiz but here for further information and explanation.

Compression file formats

JPG, PNG, GIF, TIF / TIFF, TGA

- Compression formats used by the **graphics industry**

MP3, WMA, MP4, MQA, WAV, FLAC

- Compression formats used by the **sound engineering and music industry**

MPG, MP4, DIVX, XVID, MOV, AVI

- Compression formats used by the **video industry**

How Compression Works

- Here is an old quote from Vangie Beal:

  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 "♠". The text 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.

- With dictionary “♠compression␣” and 5 replacements, total size is 406 characters, 90.4% of 449.
- algorithm builds a token/string dictionary

How Compression Works (cont'd)

- With more pattern matching and a bigger dictionary…

  ♠compression␣ ☻transmit ♣here are␣ ♥data␣
  ♦communications␣ ☺standard␣ ♪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 dictionary, total size is 363 characters or 81% of original. The more pattern matches there are for each dictionary item, the higher the compression.

Two types of Compression Formats

- Loss-Less
  - *ZIP*, *TIF*, *FLAC*, and other general file compression routines are considered *lossless* compression. The original data is completely encoded; compression reduces redundant data (e.g. a large blank space in a TIFF image or a long noiseless passage in FLAC audio).
  - 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*
  - *JPG*, *MPG*, *MP3*, *GIF*, and other end-user formats use *lossy* compression. Data is removed from the original in order to achieve compression. You can not return to the original uncompressed file.
    - GIF images reduce the number of colours in an image
    - JPG images effectively delete colour information to achieve compression
    - MP3s simplify the sound waves of audio
  - Developers can adjust the amount of compression and quality content

What is “Lossless” Vs. “Lossy” Compression? (Cont’d)

- *JPG*, *MPG*, *MP3*, *GIF*, and other formats use *lossy* compression.
- These formats, in order to achieve compression, remove data from the source file:
  - *GIF* images *limit the number of colours* in an image to 256 per pixel but preserve detail with lossless data compression similar to ZIP files. Useful for sharp-edged line art.
  - *JPG* images *delete colour and fine detail information* to achieve compression with acceptable reproduction of photographs.
  - *MP3s simplify the sound waves of audio.*
- A file with Lossy compression *can never be returned to its original, complete state.*
- *The amount of loss or compression can be adjusted;* a developer decides *how much compression* can be applied while retaining enough useful *quality* of the content for the intended purpose. E.g. high compression for JPG thumbnail images but little compression for large, zoomable images.

Overview of some Compression File Formats

- ZIP:
  - The most popular general-purpose compression archive.
  - Supported on virtually all platforms from mainframes to PCs.
  - Includes features such as encryption using password protection.
- RAR, 7z, TAR, StuffIt:
  - Proprietary general-purpose compression file formats with incremental improvements over Zip but with the loss of standardized support.
  - use different algorithms, with various benefits and uses.
  - Some are designed for different operating systems (StuffIt for Mac, TAR for *nix--**T**ape**AR**chive).

Overview of some Compression File Formats (Cont'd)

- GIF, JPG, PNG, TIFF:
  - These are compression formats used by the **graphics industry.**
  - GIF and JPG are lossy formats and cannot be uncompressed to the original source data.
- MP3, MP4, OGG, FLAC, MQA:
  - These are compression formats used by the **sound engineering and music industry.**
  - These are lossy formats – except FLAC - Fully Lossless Audio Codec – and cannot be uncompressed to original source data.

Overview of some Compression File Formats (Cont'd)

- MPG, MP4, DIVX, XVID, MOV, AVI:
  - These compression formats are used by the **video industry.**
  - These are all lossy formats.
  - These will often mix compression algorithms from audio and image. technologies.

What is a “Backup” and why do we need backups? (Cont’d)

- Backup is the “procedure for making extra copies” of data “in case the original is lost or damaged and must be restored.” The procedure includes storing the copy in a geographically separate location which is platform independent from the original file and host system.
- Having a backup will allow you to recover from lost, broken or stolen hardware, and from your own accidental deletions.
- You should be in the habit of backing up user created files on your laptop or PC. OS and apps can be restored from their original software providers or a system Restore Point but user created data can only be restored from backups.

When & How to run your Backup

- Automatically:
  - Performed by continuously running backup software that constantly monitors for file changes. Used for Full and Incremental strategy.
- Scheduled:
  - system operator or backup software runs a backup at specific times, such as overnight, when it has the least impact on business operations or at critical business times such as at accounting month/year end. Used for Full, Differential, and Incremental strategy.
- Manual Backup:
  - a user performs backups at their own convenience. Not a *strategy*.
  - It is the least effective method (what if you forget to do it?), but it's better than no backup at all!

Locations of Backup Media

- Local:
  - copies files to a drive in use by the system.
  - fastest and most convenient, but if the computer is lost or malfunctions, so goes the data! Just having a **copy** *is not* a backup.
  - Local copies may be made to reduce downtime. The copies are then moved to External media or transmitted which is a slower process.
- External:
  - Copies files to External/Portable/Flash Drive.
    i.e. a device which can be disconnected from the computer.
  - It is a backup when the platform independent device is taken off-site.

Locations of Backup Media (Cont'd)

- Network:
  - Back up files to the cloud (Google Drive, OneDrive, Dropbox, iCloud)
  - It is a slower option for large backup. Cost effective communications bandwidth has significantly less throughput than writing data to a directly attached device.
- The best location depends on the type of work you're doing, the volatility of the data (how quickly it changes), the volume of data, the backup window (available downtime), security considerations, and the speed/availability of restoration.