File System Review

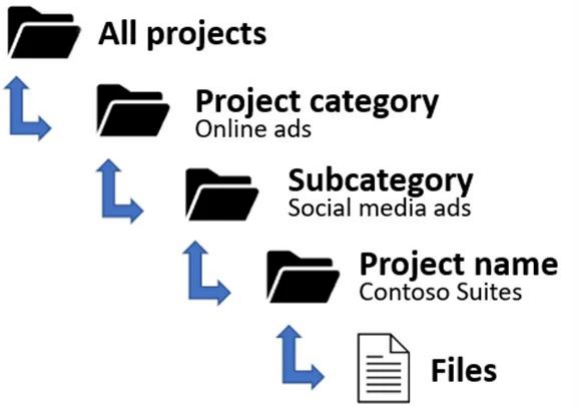
https://www.suitefiles.com/folder-structures-guide/

Navigating your computer is required for data science

Files are stored in folders

Folders may be stored in other folders

hierarchical organization

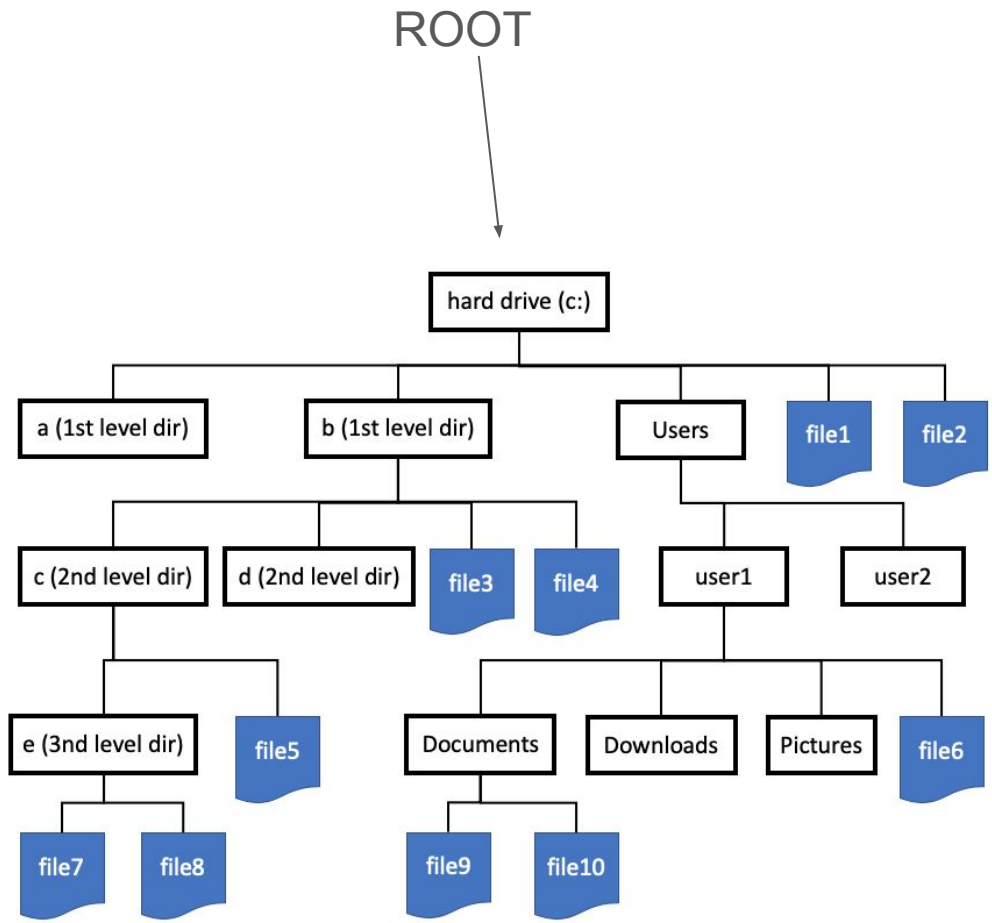


Hierarchy from root

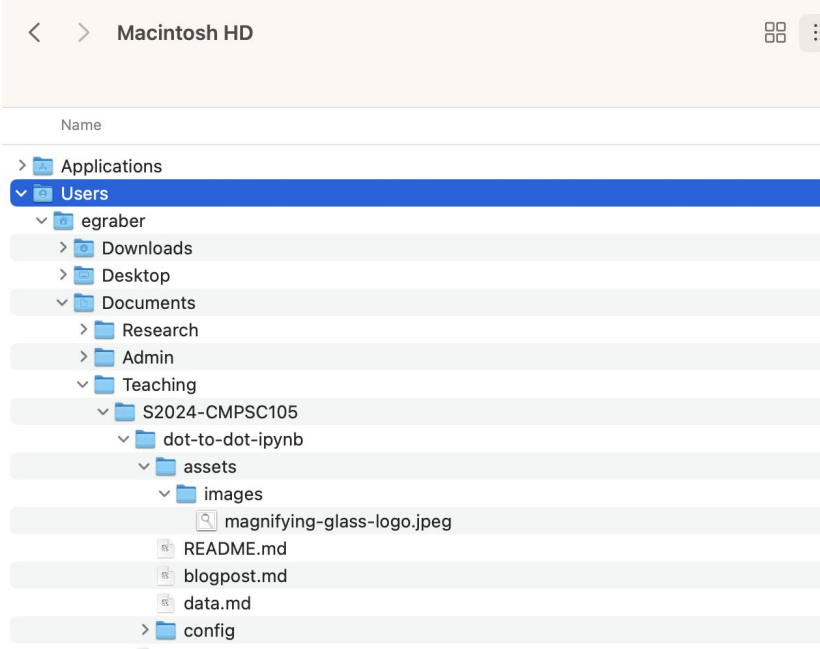
Folders have parents and children

Folders may contain folders and files

Only root has no parent



Terminology



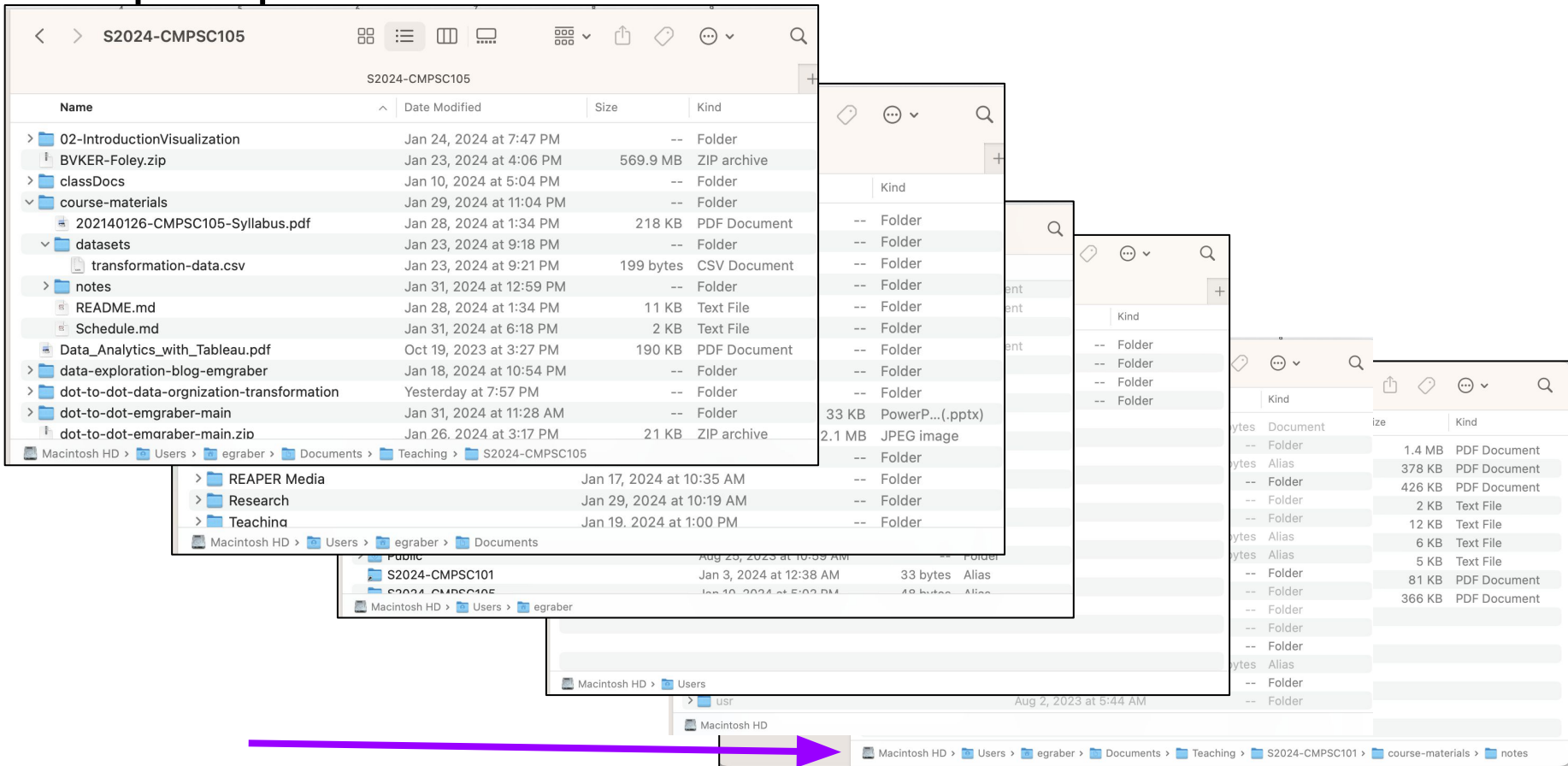
Directory == folder

Root == highest directory

Path

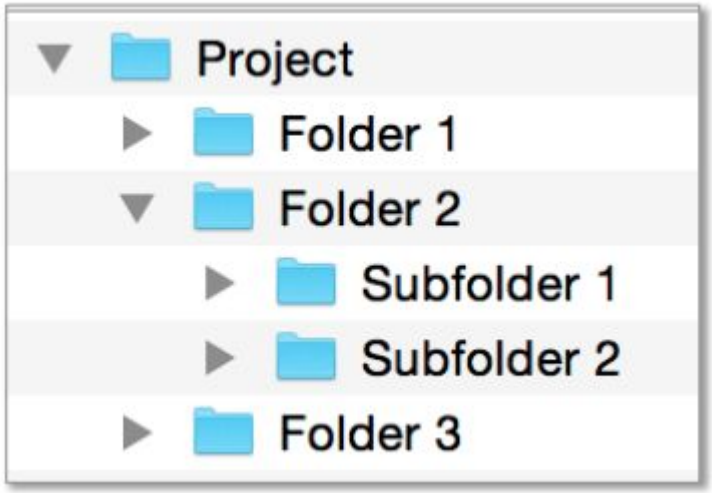
- a path gives directions to a location on your computer
- separated by / (mac) \ (windows)
- /Users/egraber/Documents/Teaching/S2024-CMPSC105/dot-to-dot-ipynb/assets/images

Set up helpful views

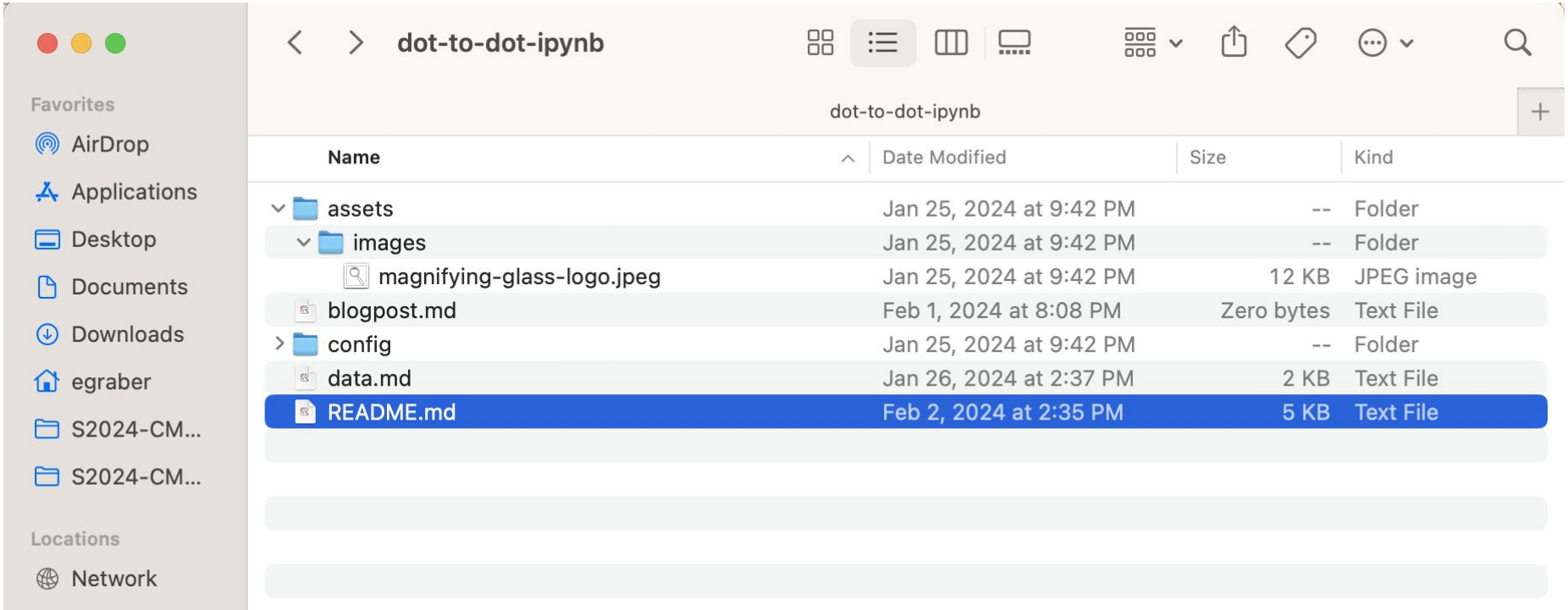


If my current location is Project,

what is the path to Subfolder 2?

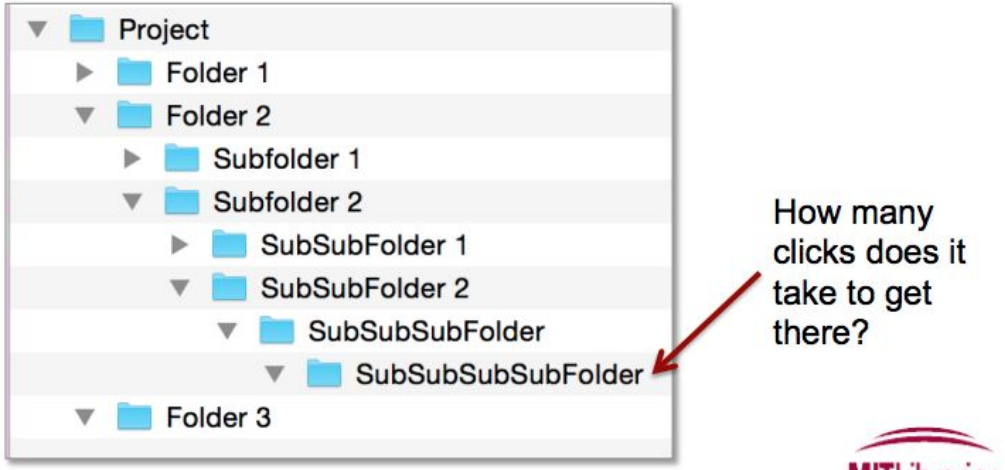


https://library.ucmerced.edu/node/66751



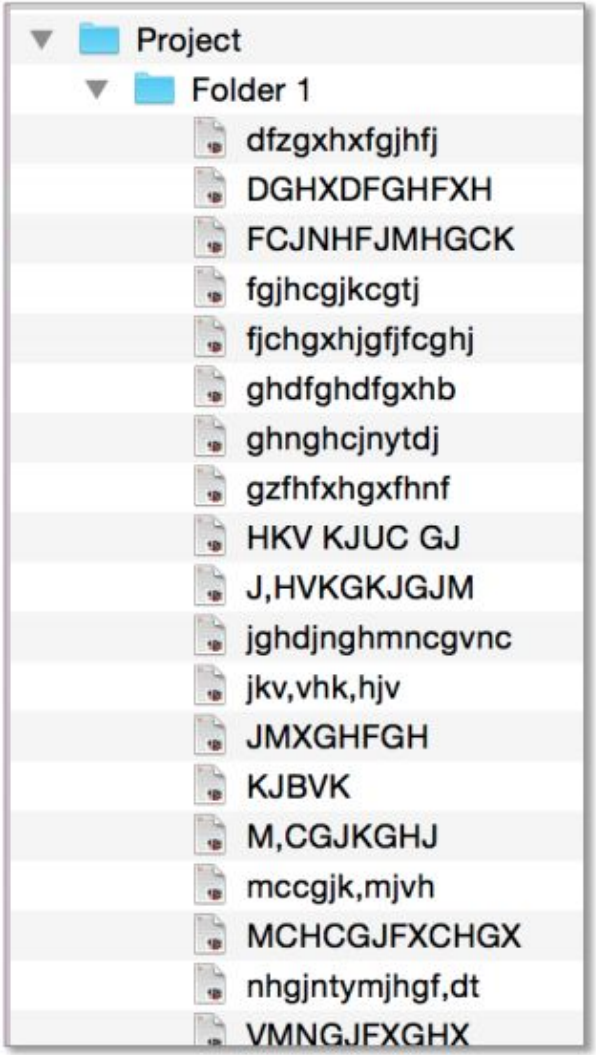
what is the path to the .jpeg file from the top-level directory in the current view?

Too much vs too little



How many clicks does it take to get there?

MITLibraries



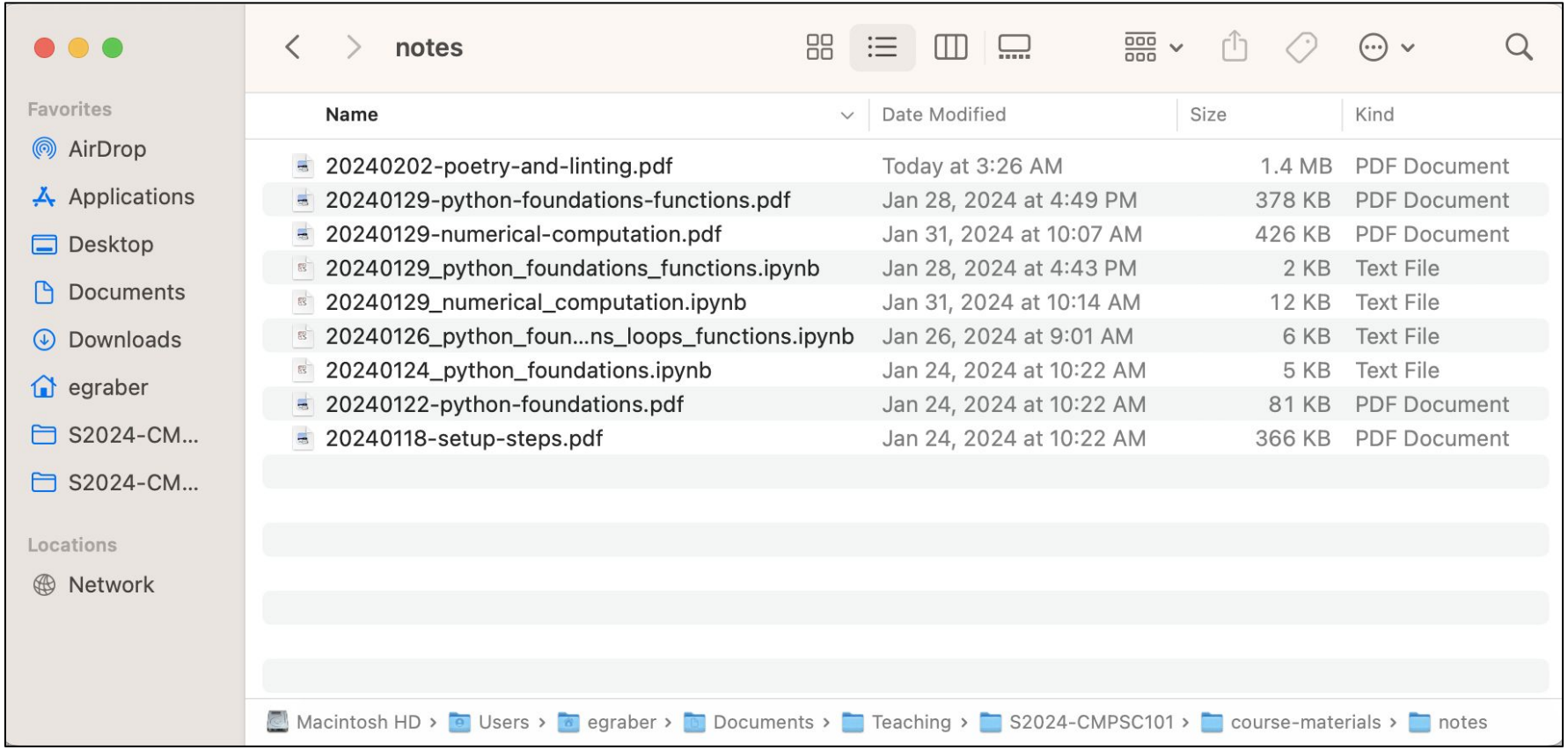
https://library.ucmerced.edu/node/66751

Good File Naming Habits

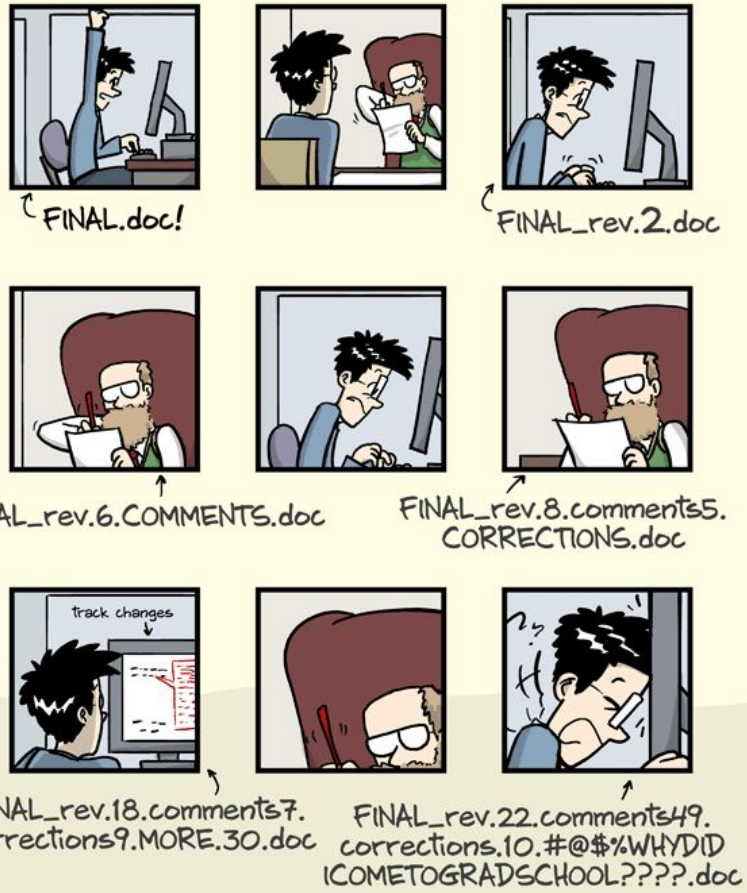
- Be consistent!
- Determine a file naming convention before you gather data.
- Limit file names to 32 characters or less (usually less). Keep it short, but make sure to provide all necessary information. If you use abbreviations, define them in a README file (and keep the README file linked to the files it describes).
- With sequential numbering (e.g., 1, 2, 3, etc.), use leading zeros to accommodate multi-digit versions. For example, use 01-10 for 1-10, 001-100 for 1-100, and so on.
- Avoid special characters like & , * % # ; ( ) ! @ $ ^ ~ ' { } [ ] ? < >
- Use underscores _ rather than spaces!
- Use descriptive names that document the important aspects of your project. These can differ across projects. Put the most important information first.
- Keep names easy to read (and consider case sensitivity).
- Use a consistent date and time convention. For dates, YYYYMMDD will result in your files being sorted chronologically.

https://library.ucmerced.edu/node/66751

Real Example of file names



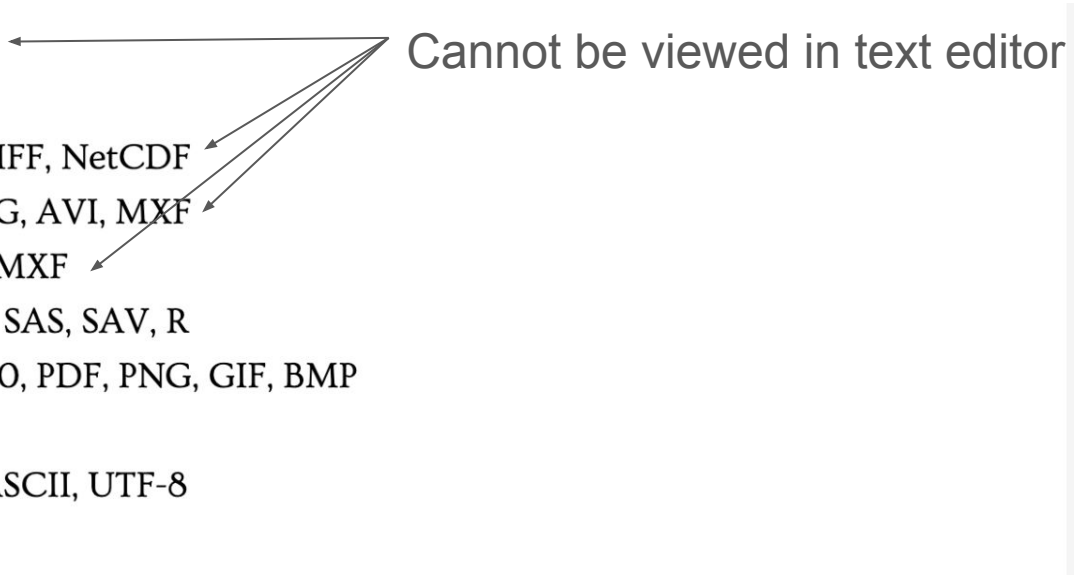
"FINAL".doc



WWW.PHDCOMICS.COM

File extensions

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV, R
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC



Binary vs Text

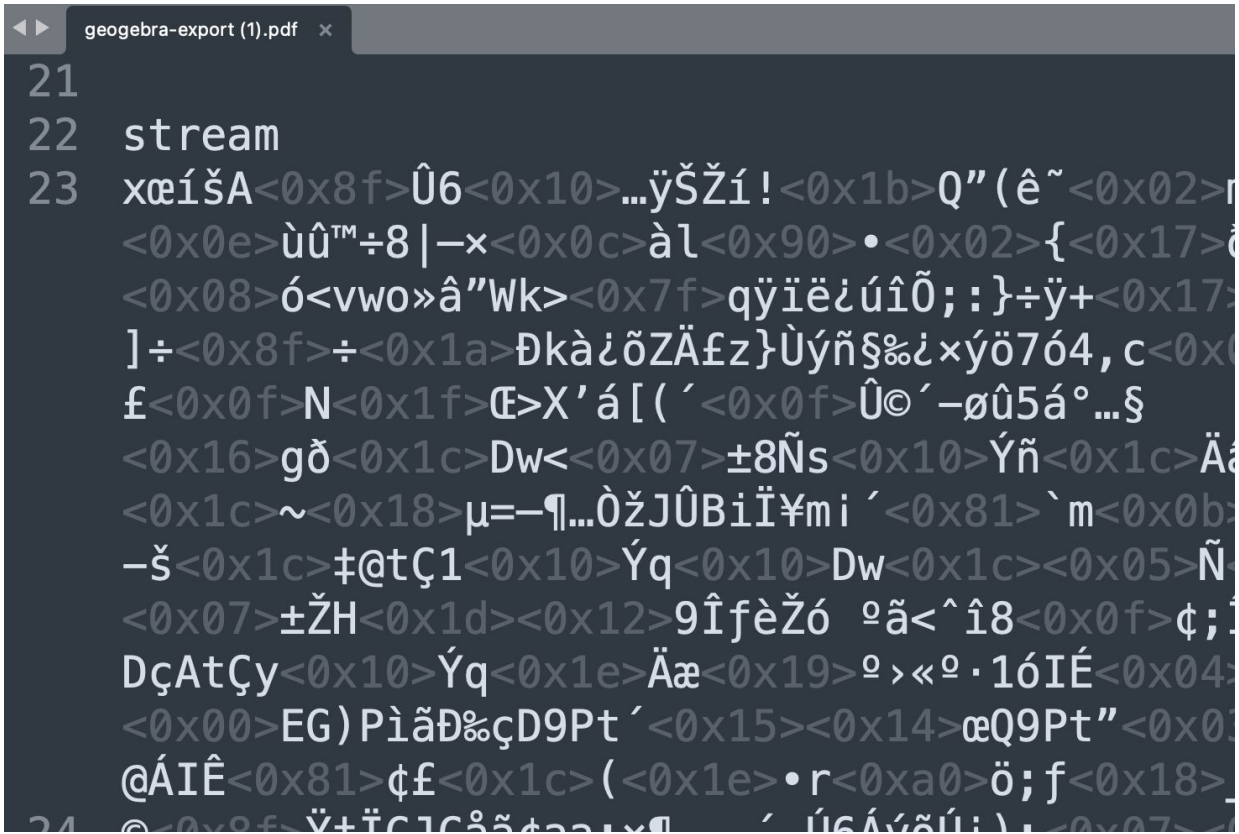
Binary files contain information in 8-bit format. But when you open a binary file (.png), you will see a massive load of garbage (accented characters with weird-looking characters here and there). These are chunks of bytes (8 bits) instead of bit information. If 1's and 0's were used instead, it would contain a gazillion characters inside a comparatively small-sized file. Hence, chunks are used. However, the data is stored as 0's and 1's, which is readable by computers.

Text files are also binary files but the files are readable by the human eye as they are formatted according to ASCII (American Standard Code for Information Interchange). ASCII maps each character to a unique decimal number, which can be converted to binary to make computers know what character to display or not. For example, the letter 'A' has a unique decimal ASCII value of 65, which, when converted to binary, is represented by 01000001. So, the computer prints A when we type A on our keyboard. So, characters are stored as 7- or 8-bit bytes in a text file, and even numbers are read as characters — ASCII reads '0' as decimal 48, '1' as 49, and '9' as 57, 'A' as 65, 'a' as 98, and so on.

https://levelup.gitconnected.com/binary-files-vs-ascii-files-106a0e07617d?gi=6caa0e1ef906

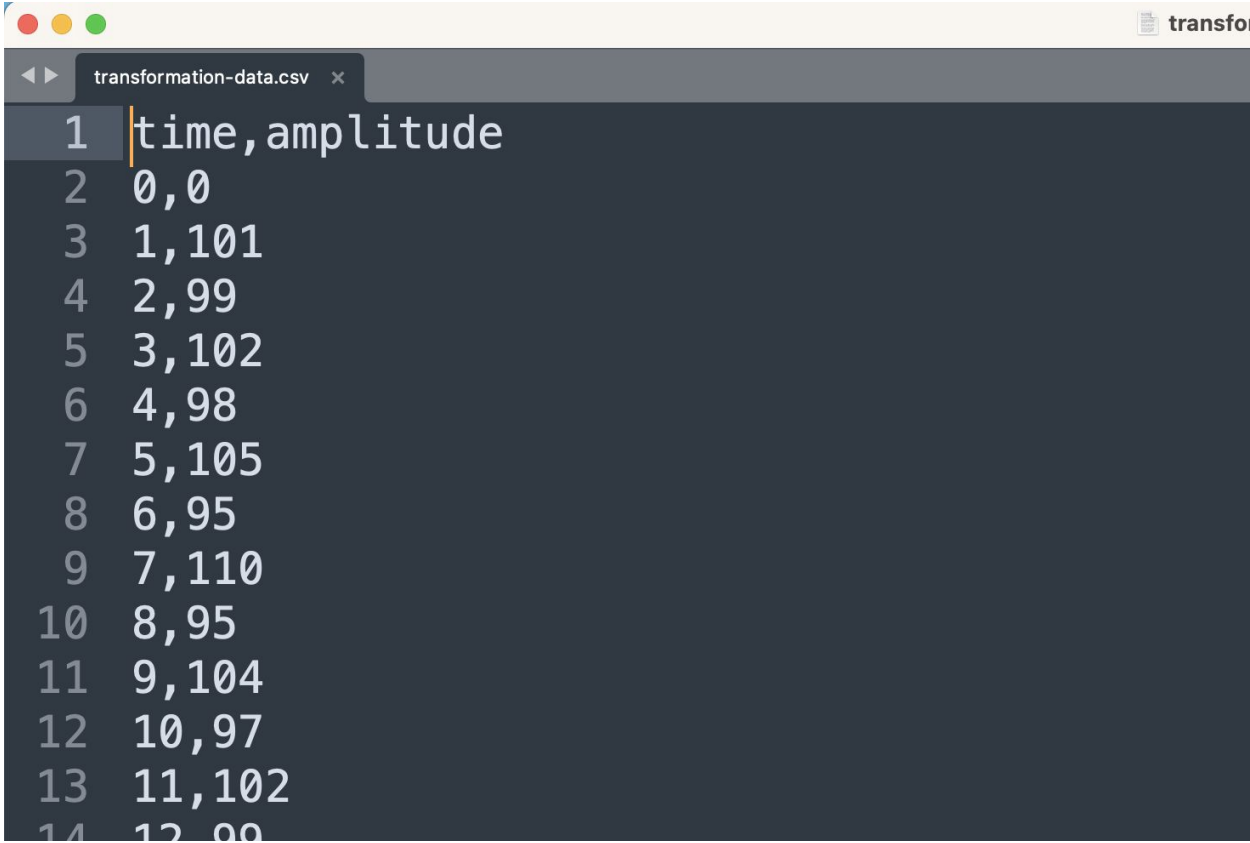
Binary or Text?

what is the extension?



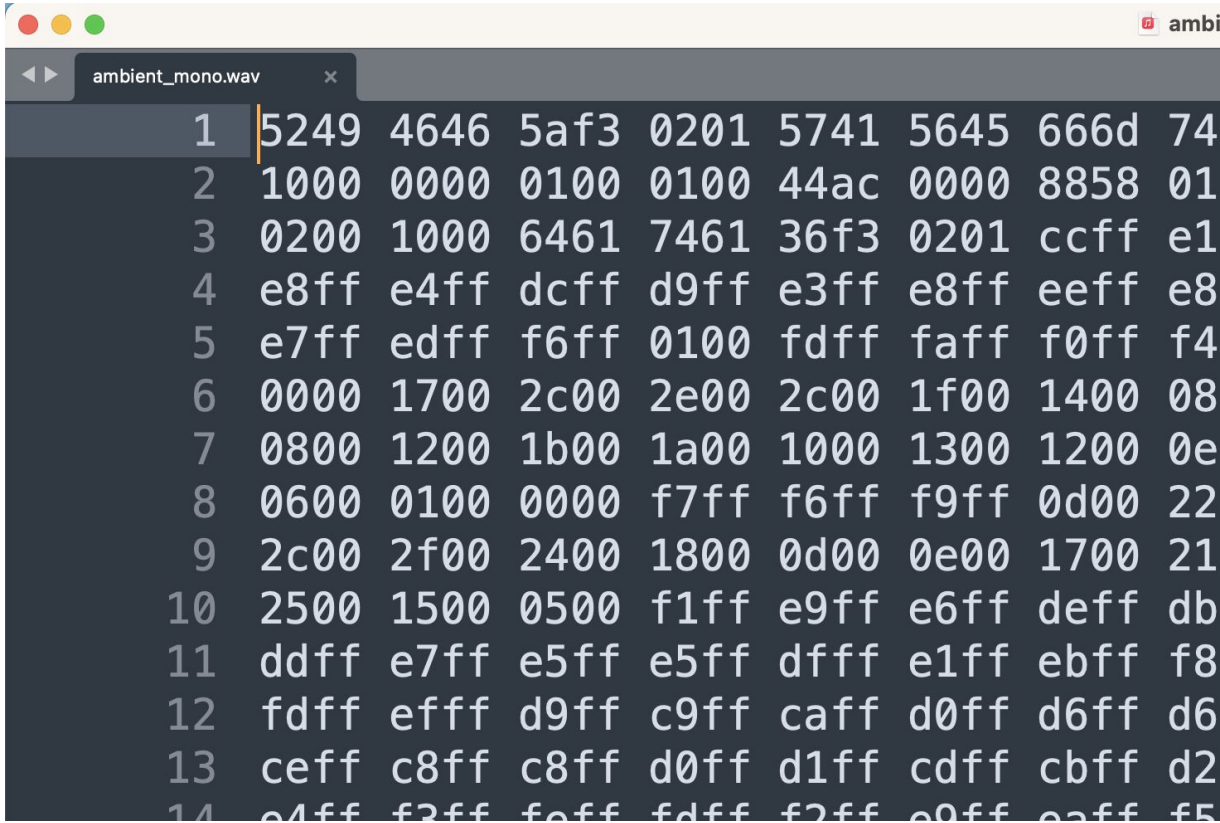
Binary of Text?

What is the extension?



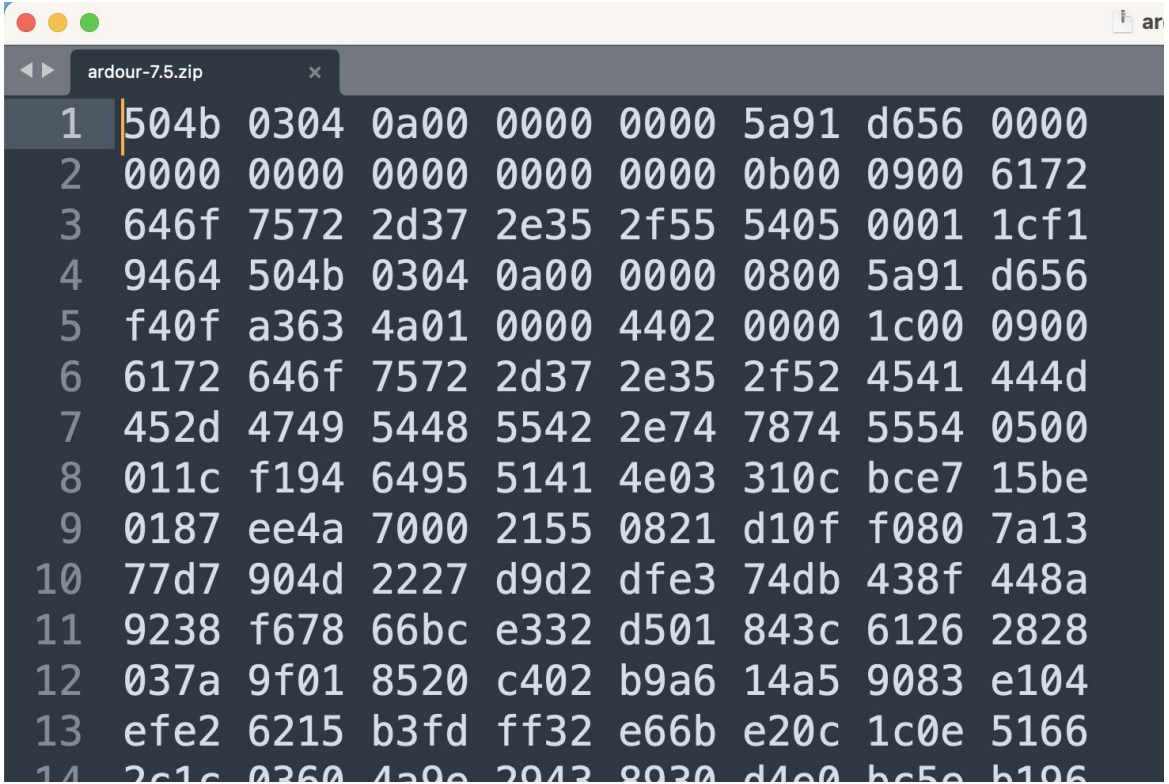
Binary of Text?

What is the extension?



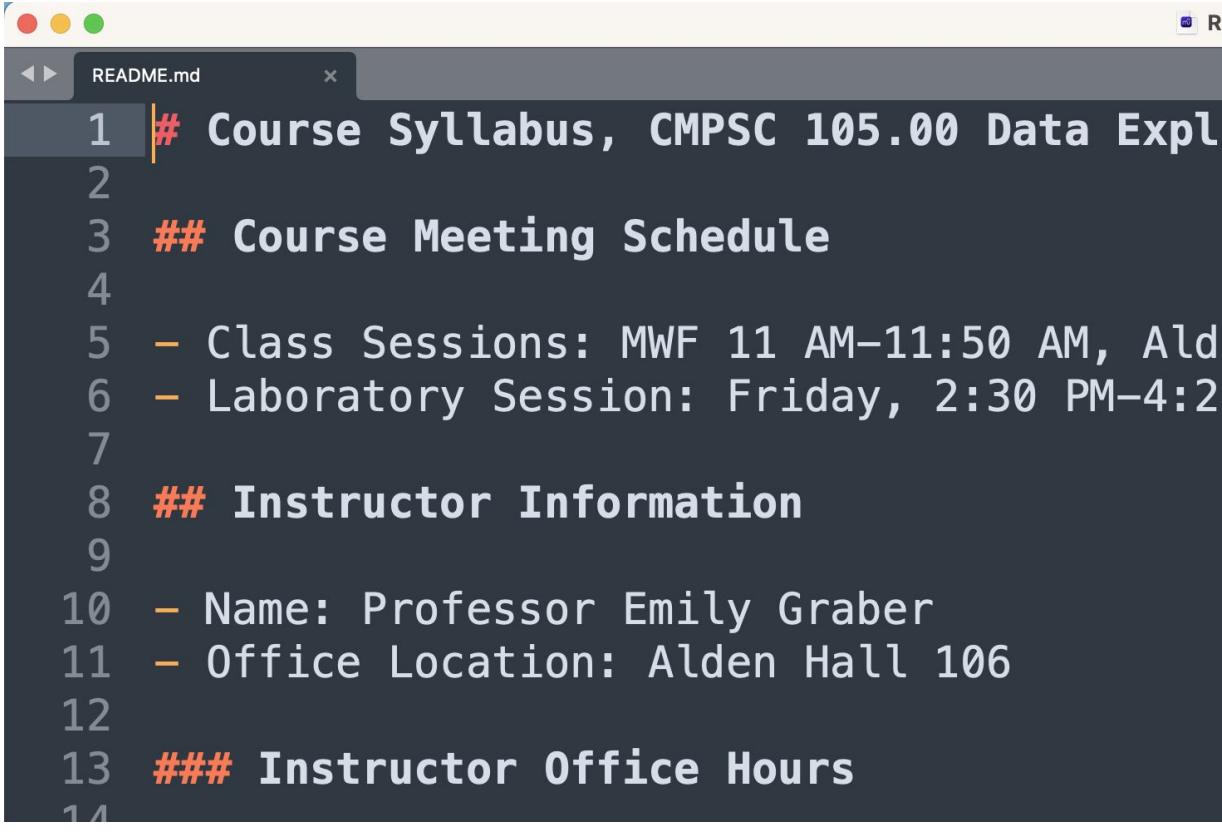
Binary of Text?

What is the extension?



Binary of Text?

What is the extension?



Things to avoid

Things to avoid

Not knowing where your files are!

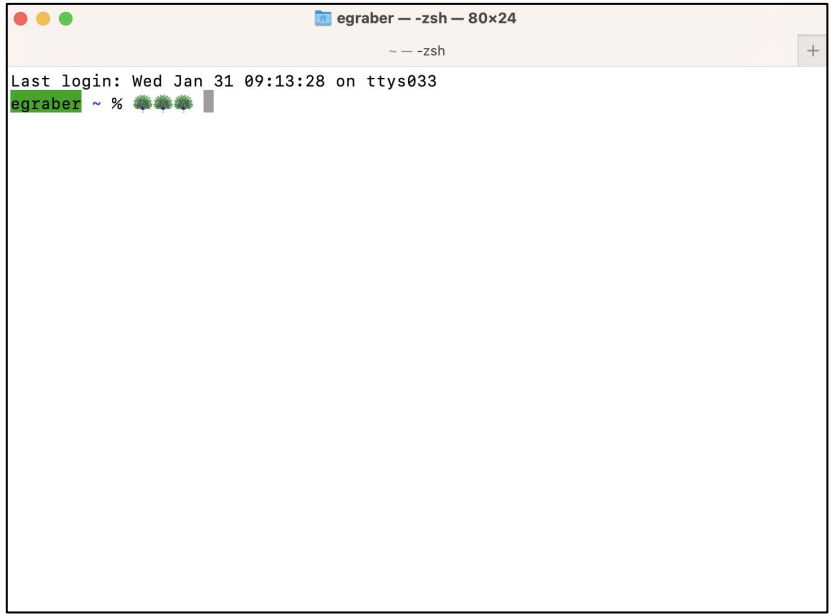
Using "recent" files instead of the hierarchy

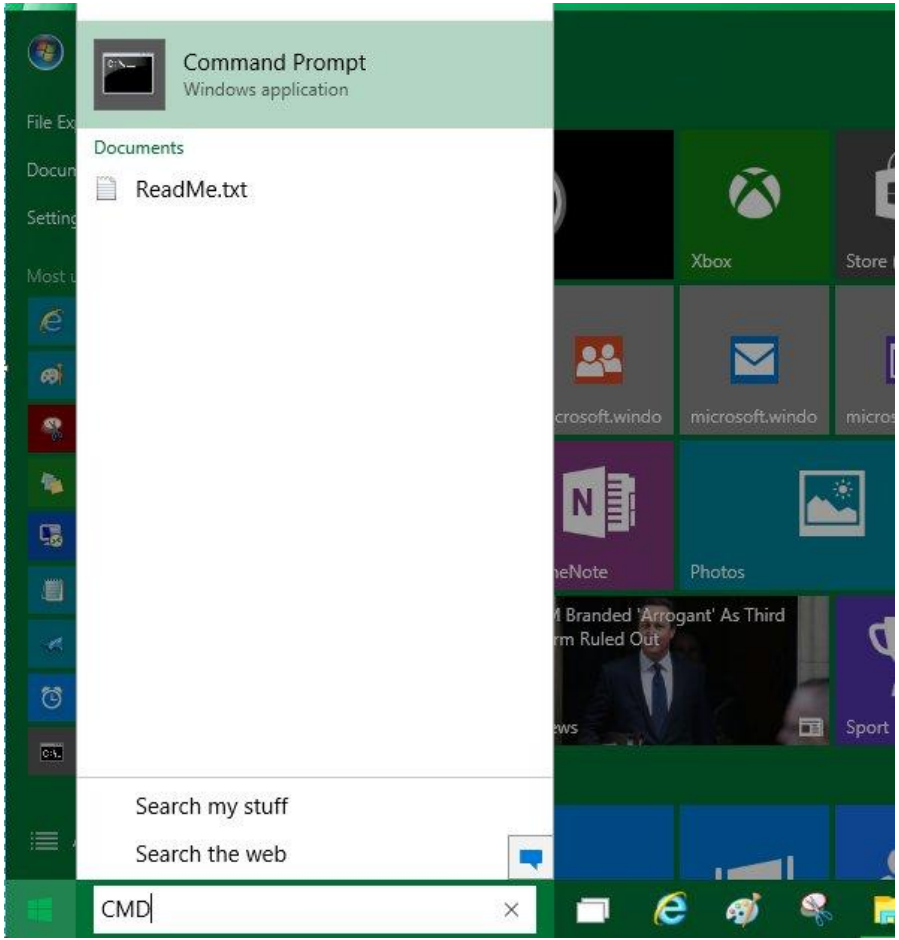
Using spaces in filenames

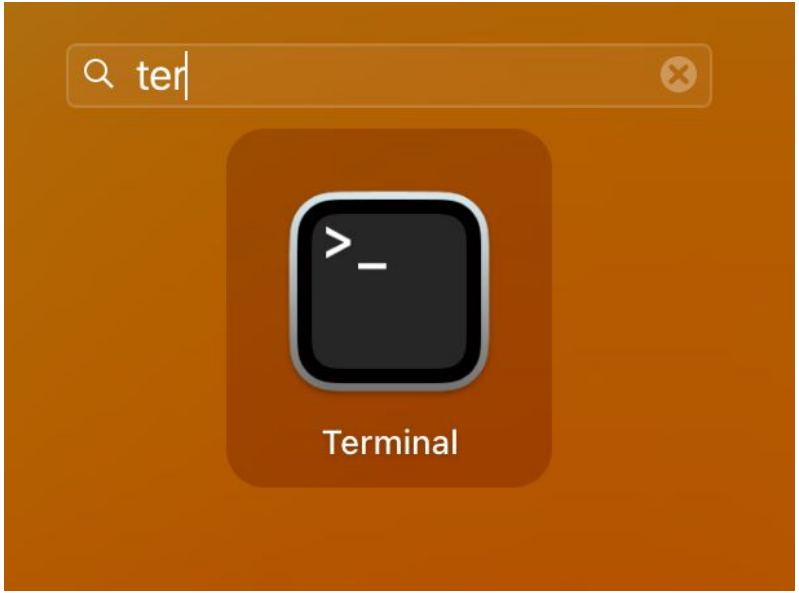
Before there were GUIs, there was the terminal

Terminal

- terminal is a program that allows you to access all the files, folders, and programs on your computer without a GUI




Command Prompt
Windows application
Documents
ReadMe.txt
Search my stuff
Search the web
CMD
Xbox
Store
Photos
Sport


ter
Terminal

Terminal

In order to use the terminal

- need to know where your files are
  - "path"
  - a path is directions to a location on your computer
  - separated by / (mac) \ (windows)
  - ~/Documents/courses/cmpsc105/course-materials/README.md
- need to know the names of the files
  - "filename"
- need to call programs from the terminal
  - example programs - next slide

Terminal Programs

- change directory: cd
  - `cd path/hierarchy/structure/`
  - `cd ..`
  - windows: `cd path\hierarchy\structure\`
- list out everything in current location:
  - macos / linux / bash shells: `ls`
  - windows: `dir`
- Go to home directory
  - `cd`
  - or macos / linux: `cd ~/`
  - windows cd ~\

Please Install Spyder

https://www.spyder-ide.org/

For laptops with windows, linux, macos