

File System Review

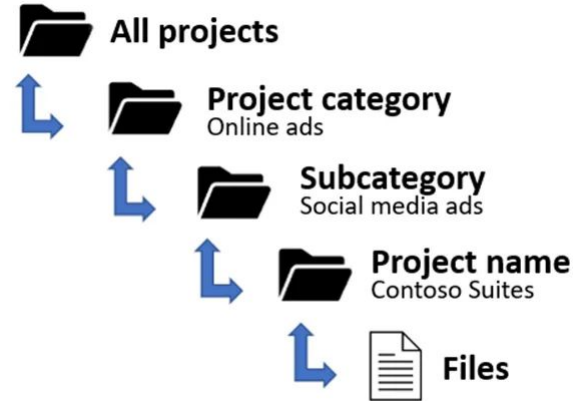
<https://www.suitefiles.com/folder-structures-guide/>

Navigating your computer is required for data science

Files are stored in folders

Folders may be stored in other folders

hierarchical organization

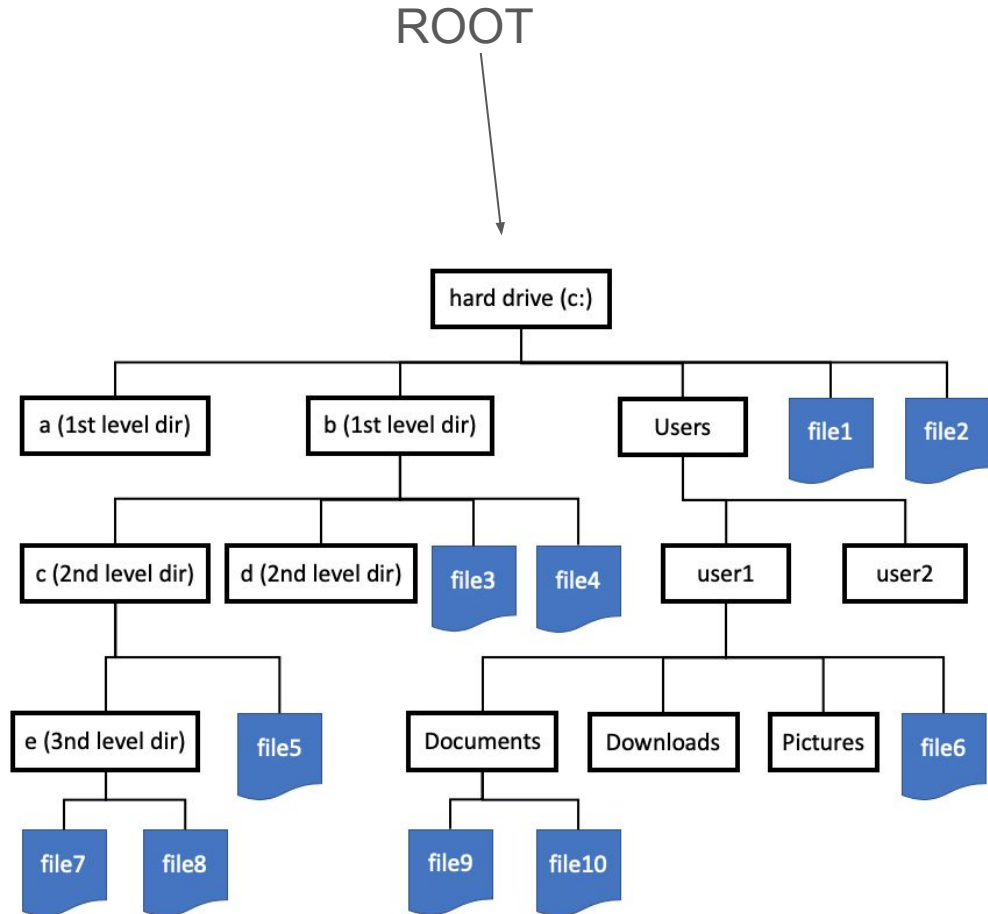


Hierarchy from root

Folders have parents and children

Folders may contain folders and files

Only root has no parent



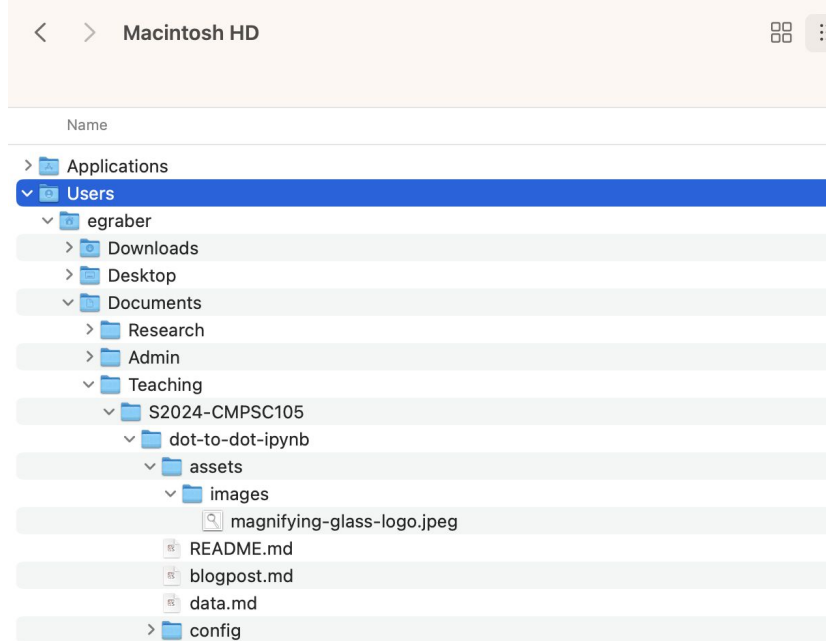
Terminology

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

The main window displays the file list for 'S2024-CMPSC105' with the following columns: Name, Date Modified, Size, and Kind.

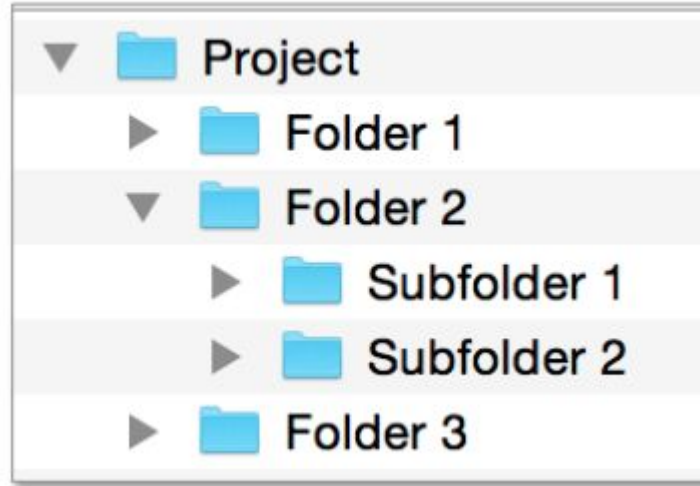
Name	Date Modified	Size	Kind
> 02-IntroductionVisualization	Jan 24, 2024 at 7:47 PM	--	Folder
> BVKER-Foley.zip	Jan 23, 2024 at 4:06 PM	569.9 MB	ZIP archive
> classDocs	Jan 10, 2024 at 5:04 PM	--	Folder
> course-materials	Jan 29, 2024 at 11:04 PM	--	Folder
> 202140126-CMPSC105-Syllabus.pdf	Jan 28, 2024 at 1:34 PM	218 KB	PDF Document
> datasets	Jan 23, 2024 at 9:18 PM	--	Folder
> transformation-data.csv	Jan 23, 2024 at 9:21 PM	199 bytes	CSV Document
> notes	Jan 31, 2024 at 12:59 PM	--	Folder
> README.md	Jan 28, 2024 at 1:34 PM	11 KB	Text File
> Schedule.md	Jan 31, 2024 at 6:18 PM	2 KB	Text File
> Data_Analytics_with_Tableau.pdf	Oct 19, 2023 at 3:27 PM	190 KB	PDF Document
> data-exploration-blog-emgraber	Jan 18, 2024 at 10:54 PM	--	Folder
> dot-to-dot-data-organization-transformation	Yesterday at 7:57 PM	--	Folder
> dot-to-dot-emgraber-main	Jan 31, 2024 at 11:28 AM	--	Folder
> dot-to-dot-emgraber-main.zip	Jan 26, 2024 at 3:17 PM	21 KB	ZIP archive

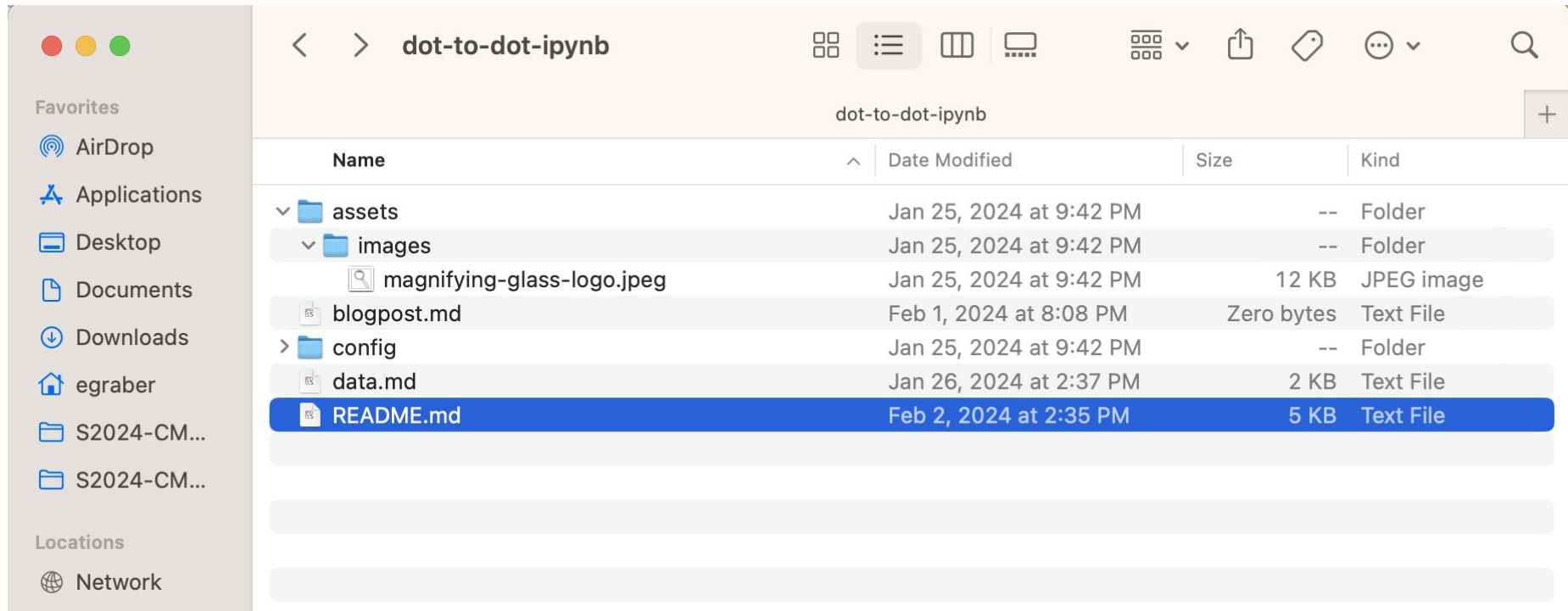
The overlaid windows show the following steps:

- Clicking the 'Kind' column header to sort by Kind.
- Clicking the 'Size' column header to sort by Size.
- Clicking the 'Kind' column header again to sort by Kind.

A large blue arrow points from the bottom left towards the bottom right, indicating the sequence of steps.

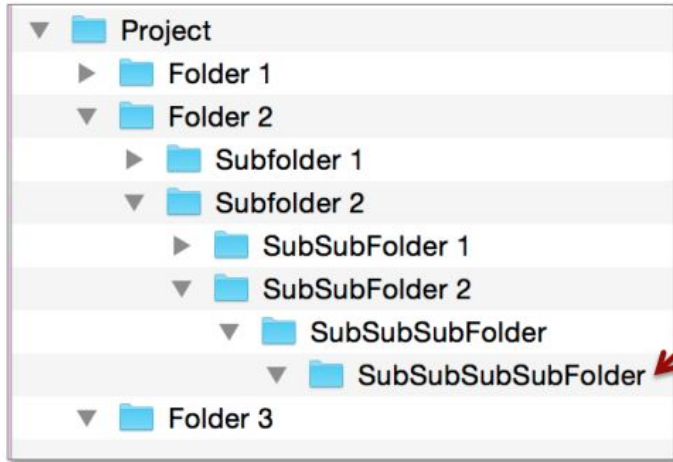
If my current location is Project,
what is the path to Subfolder 2?



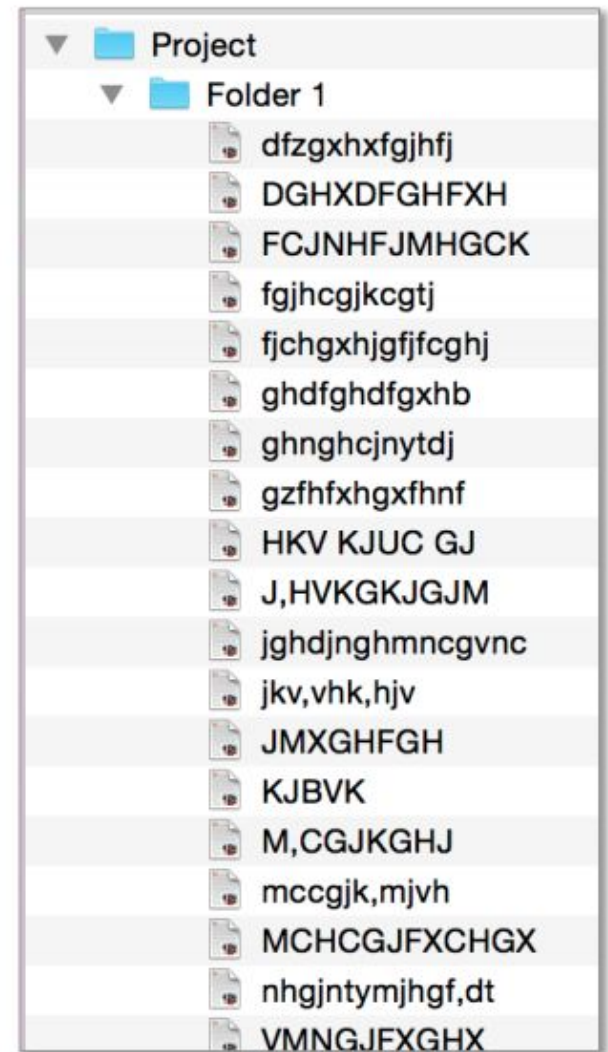


what is the path to the .jpeg file from the top-level directory in the current view?

Too much vs too little



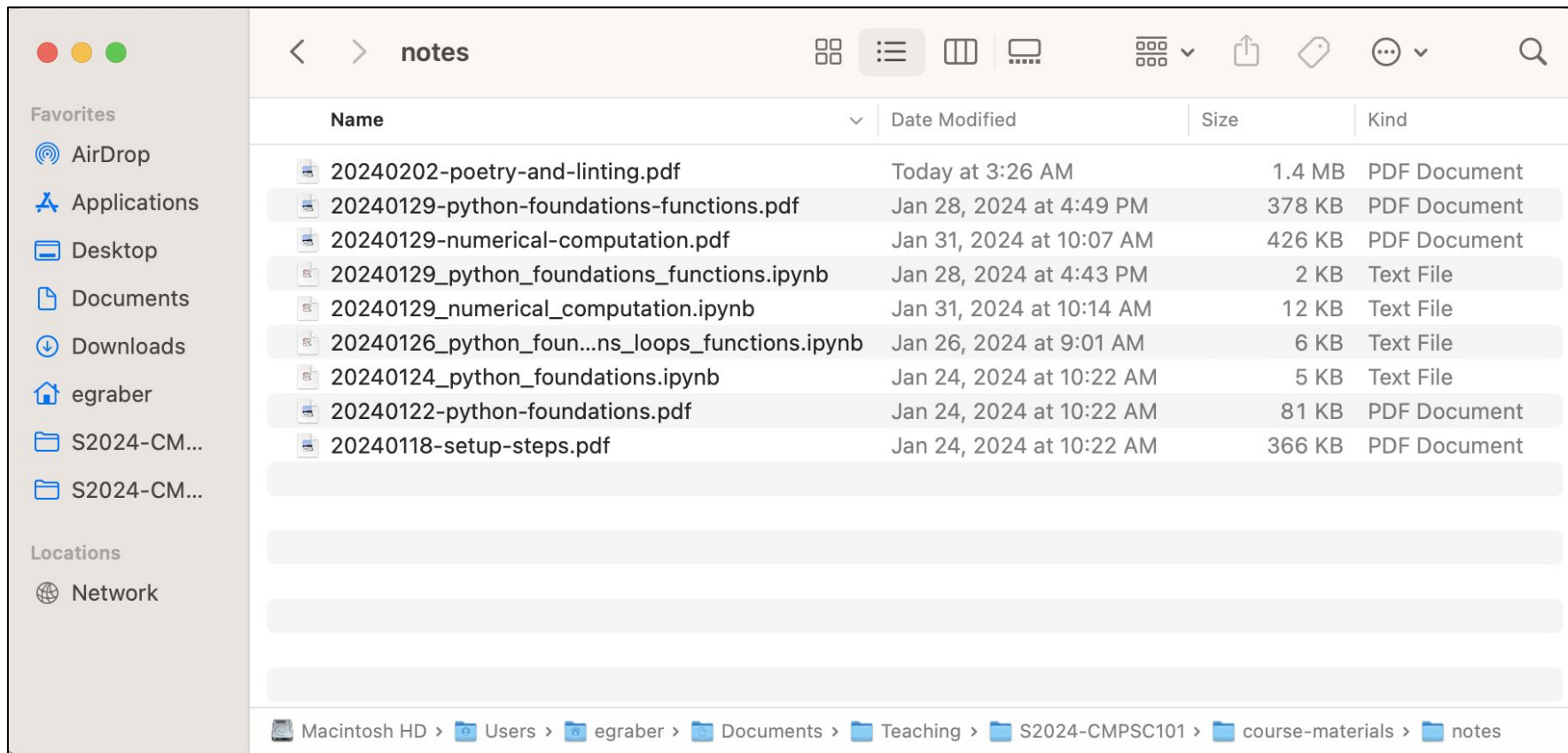
How many
clicks does it
take to get
there?



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.

Real Example of file names



The screenshot shows a macOS Finder window titled 'notes'. The left sidebar contains 'Favorites' (AirDrop, Applications, Desktop, Documents, Downloads, egraber, S2024-CM..., S2024-CM...) and 'Locations' (Network). The main pane displays a table of files in the 'notes' directory.

Name	Date Modified	Size	Kind
20240202-poetry-and-linting.pdf	Today at 3:26 AM	1.4 MB	PDF Document
20240129-python-foundations-functions.pdf	Jan 28, 2024 at 4:49 PM	378 KB	PDF Document
20240129-numerical-computation.pdf	Jan 31, 2024 at 10:07 AM	426 KB	PDF Document
20240129_python_foundations_functions.ipynb	Jan 28, 2024 at 4:43 PM	2 KB	Text File
20240129_numerical_computation.ipynb	Jan 31, 2024 at 10:14 AM	12 KB	Text File
20240126_python_foun...ns_loops_functions.ipynb	Jan 26, 2024 at 9:01 AM	6 KB	Text File
20240124_python_foundations.ipynb	Jan 24, 2024 at 10:22 AM	5 KB	Text File
20240122-python-foundations.pdf	Jan 24, 2024 at 10:22 AM	81 KB	PDF Document
20240118-setup-steps.pdf	Jan 24, 2024 at 10:22 AM	366 KB	PDF Document

The breadcrumb path at the bottom is: Macintosh HD > Users > egraber > Documents > Teaching > S2024-CMPSC101 > course-materials > notes

"FINAL".doc



FINAL.doc!



FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.#@\$%WHYDID
ICOMETOGRADSCHOOL????.doc

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

Cannot be viewed in text editor



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.

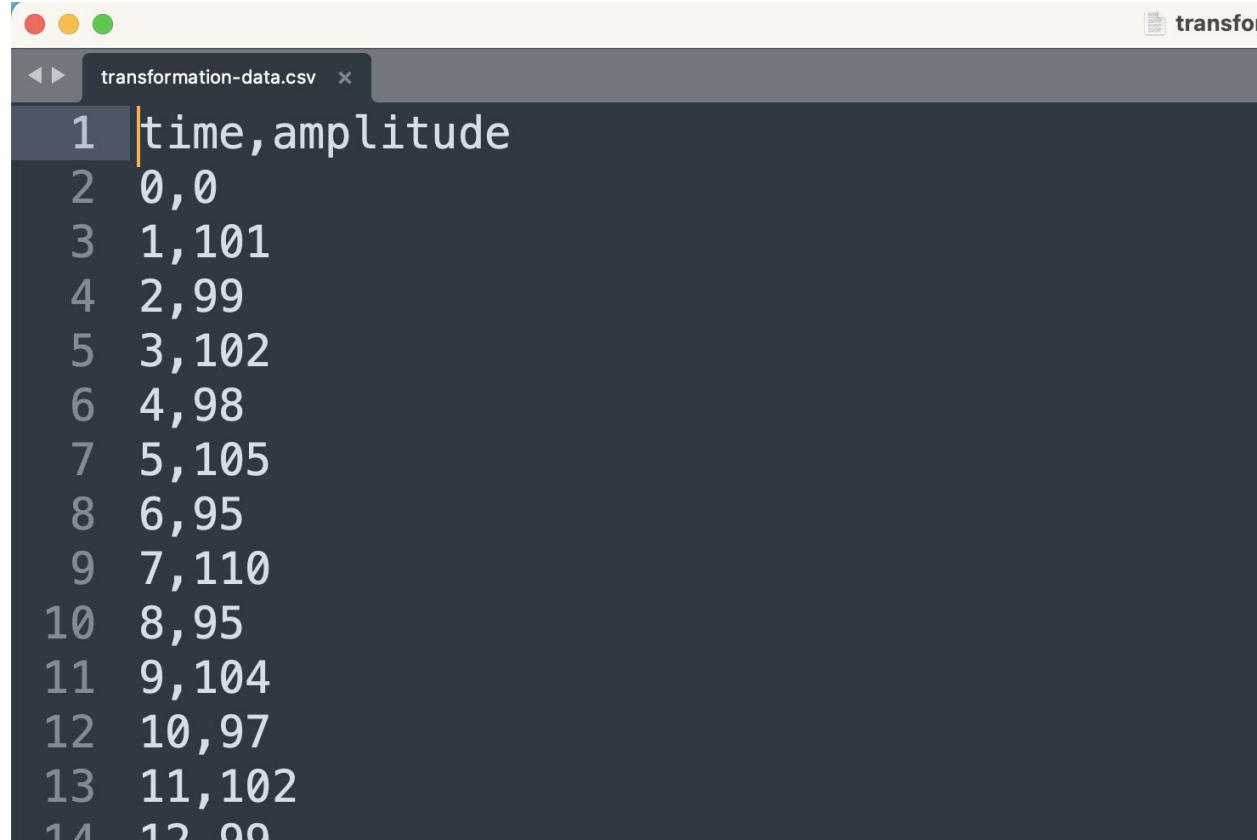
Binary or Text?

what is the extension?

```
geogebra-export (1).pdf x
21
22 stream
23 xæíšA<0x8f>Û6<0x10>...ÿŠŽí!<0x1b>Q"(ê~<0x02>r
<0x0e>ùû™÷8|-x<0x0c>àl<0x90>•<0x02>{<0x17>ö
<0x08>ó<vwo>»â"Wk><0x7f>qÿïë¿úîõ;:}÷ÿ+<0x17>
]÷<0x8f>÷<0x1a>Đkà¿õZÄfz}Ûýñ$%¿×ýö7ó4,c<0x0
f<0x0f>N<0x1f>Æ>X'á[( ' <0x0f>Û©'-øû5á°...§
<0x16>gð<0x1c>Dw<<0x07>±8Ñs<0x10>Ýñ<0x1c>Ää
<0x1c>~<0x18>μ=-¶...ÒžJÛBiİ¥mi ' <0x81>`m<0x0b>
-š<0x1c>‡@tÇ1<0x10>Ýq<0x10>Dw<0x1c><0x05>Ñ-
<0x07>±ŽH<0x1d><0x12>9ÎfèŽó °ă<^î8<0x0f>¢;
DçAtÇy<0x10>Ýq<0x1e>Äæ<0x19>º>«º·1óIÉ<0x04>
<0x00>EG)PìăĐ%çD9Pt ' <0x15><0x14>æQ9Pt"<0x0
@ÁIÊ<0x81>¢f<0x1c>(<0x1e>•r<0xa0>ö;f<0x18>
24 @<0x8f>ÿ+İÇ1Căă†aa1x¶ ' ú6ÁýăÚi) ) <0x07><
```

Binary of Text?

What is the extension?

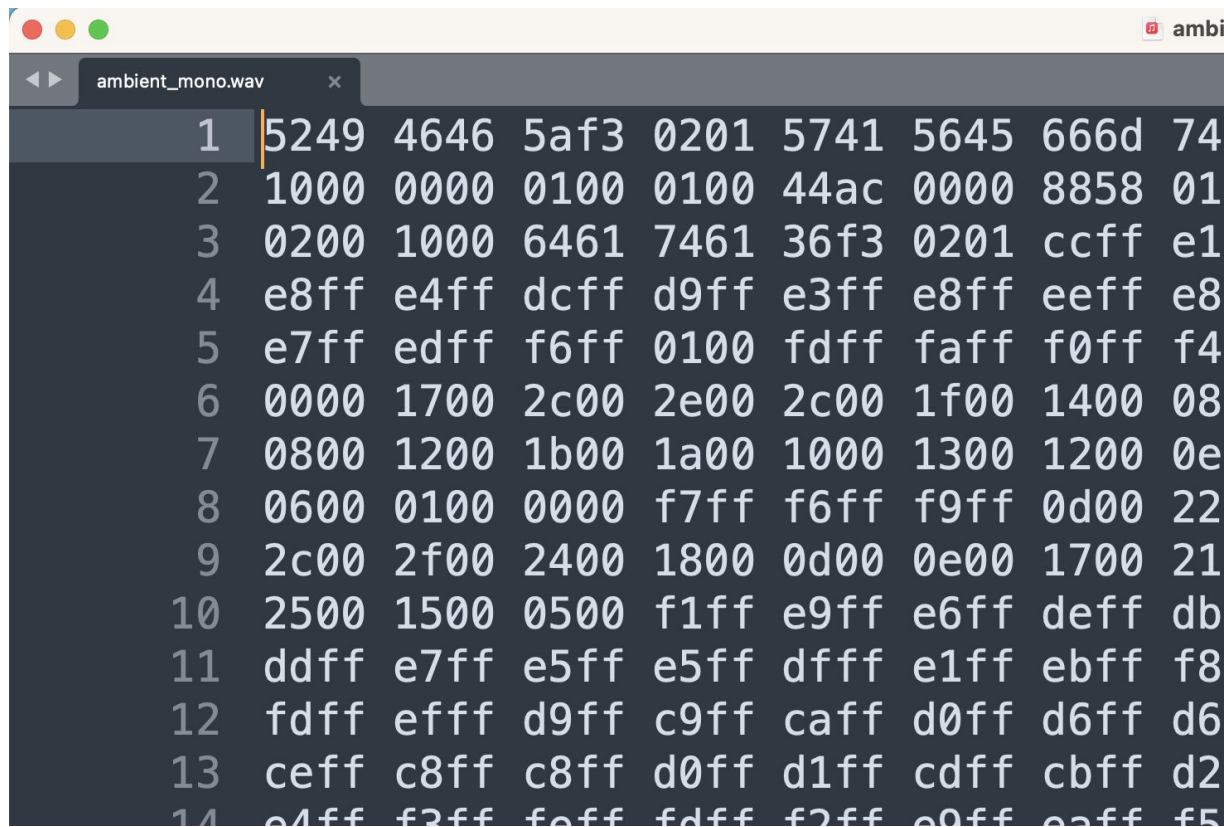


A screenshot of a text editor window titled 'transformation-data.csv'. The window displays a CSV file with 14 lines of data. The first line is a header: 'time,amplitude'. The subsequent lines contain numerical values for time and amplitude, separated by commas. The text is displayed in a dark-themed editor with a light-colored cursor at the end of the first line.

Line	time	amplitude
1	time,amplitude	
2	0,0	
3	1,101	
4	2,99	
5	3,102	
6	4,98	
7	5,105	
8	6,95	
9	7,110	
10	8,95	
11	9,104	
12	10,97	
13	11,102	
14	12,99	

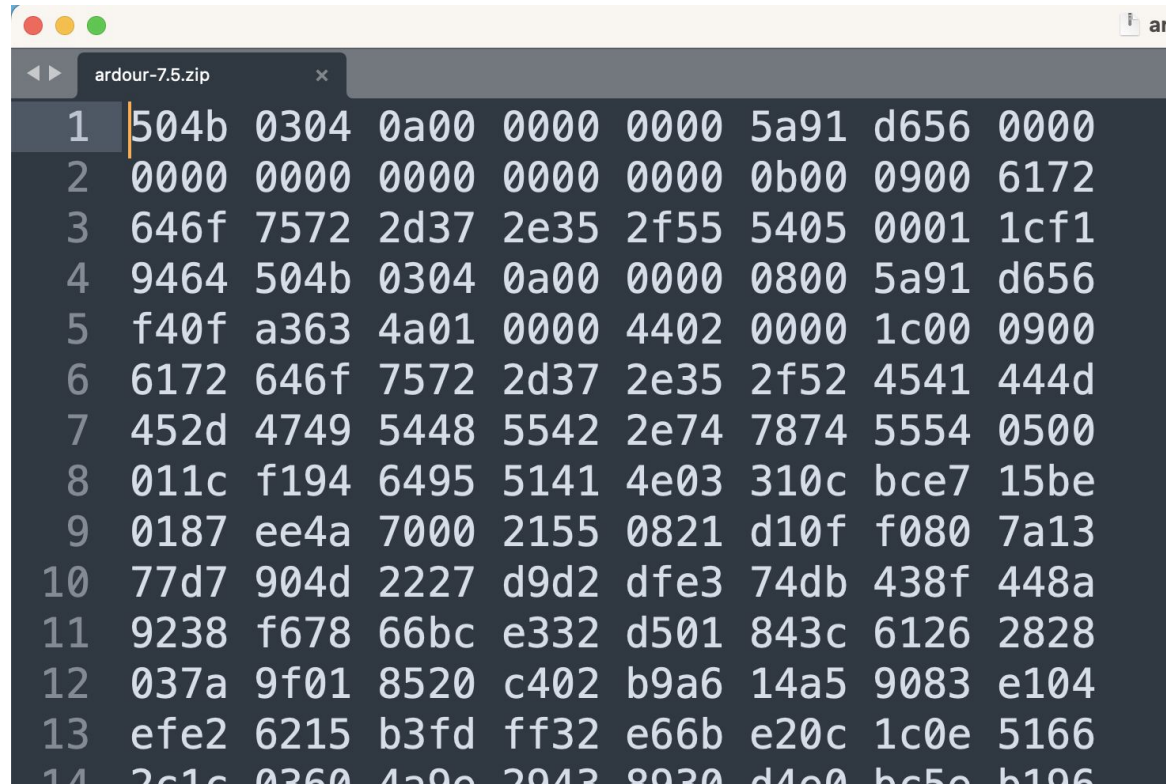
Binary of Text?

What is the extension?



Binary of Text?

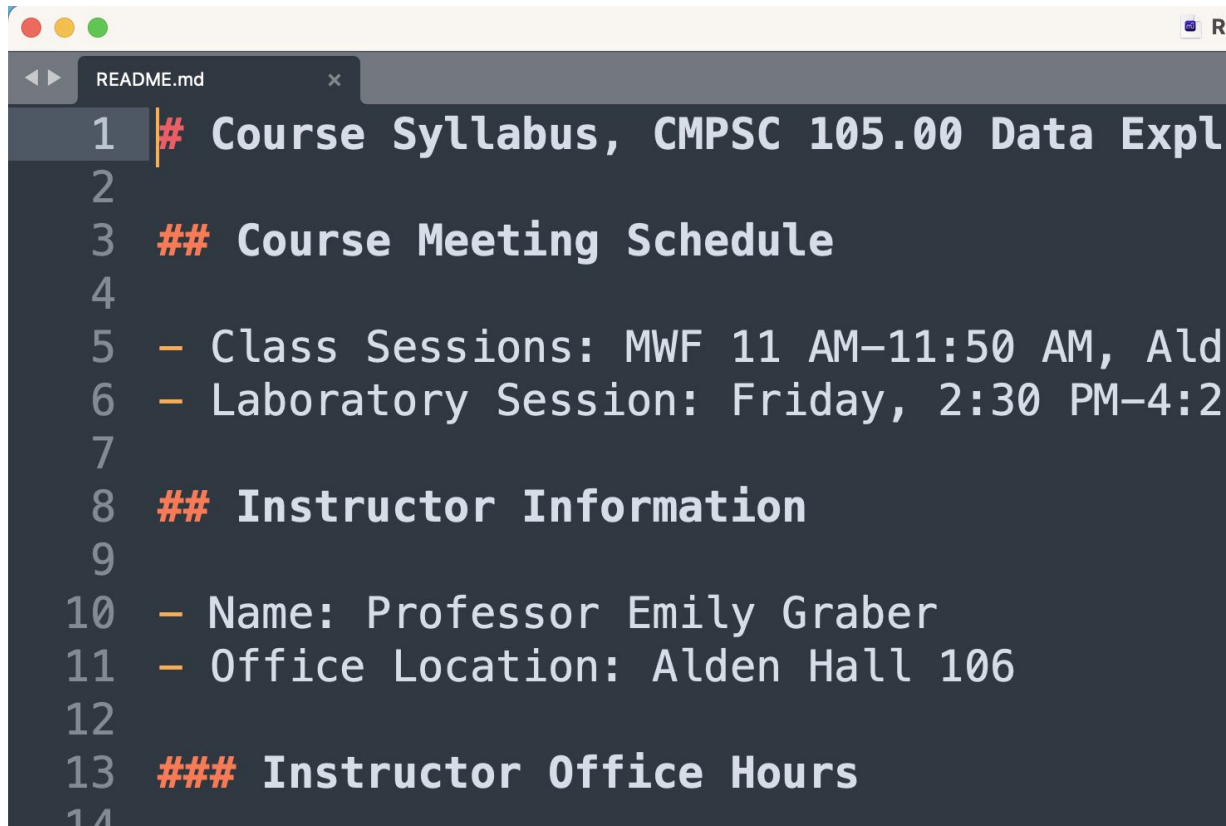
What is the extension?



```
ardour-7.5.zip
1 504b 0304 0a00 0000 0000 5a91 d656 0000
2 0000 0000 0000 0000 0000 0b00 0900 6172
3 646f 7572 2d37 2e35 2f55 5405 0001 1cf1
4 9464 504b 0304 0a00 0000 0800 5a91 d656
5 f40f a363 4a01 0000 4402 0000 1c00 0900
6 6172 646f 7572 2d37 2e35 2f52 4541 444d
7 452d 4749 5448 5542 2e74 7874 5554 0500
8 011c f194 6495 5141 4e03 310c bce7 15be
9 0187 ee4a 7000 2155 0821 d10f f080 7a13
10 77d7 904d 2227 d9d2 dfe3 74db 438f 448a
11 9238 f678 66bc e332 d501 843c 6126 2828
12 037a 9f01 8520 c402 b9a6 14a5 9083 e104
13 efe2 6215 b3fd ff32 e66b e20c 1c0e 5166
14 2c1c 0360 4a00 2043 8030 d4e0 bc5e b106
```

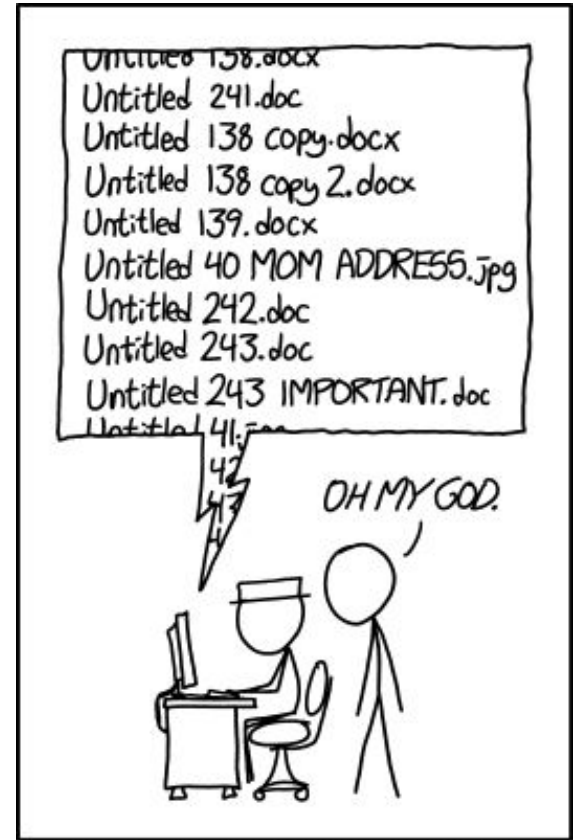
Binary of Text?

What is the extension?

A screenshot of a code editor window with a dark theme. The window has a title bar with three colored circles (red, yellow, green) on the left and a small icon on the right. The tab is labeled 'README.md'. The code is as follows:

```
1 # Course Syllabus, CMPSC 105.00 Data Expl
2
3 ## Course Meeting Schedule
4
5 - Class Sessions: MWF 11 AM-11:50 AM, Ald
6 - Laboratory Session: Friday, 2:30 PM-4:2
7
8 ## Instructor Information
9
10 - Name: Professor Emily Graber
11 - Office Location: Alden Hall 106
12
13 ### Instructor Office Hours
14
```

Things to avoid



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

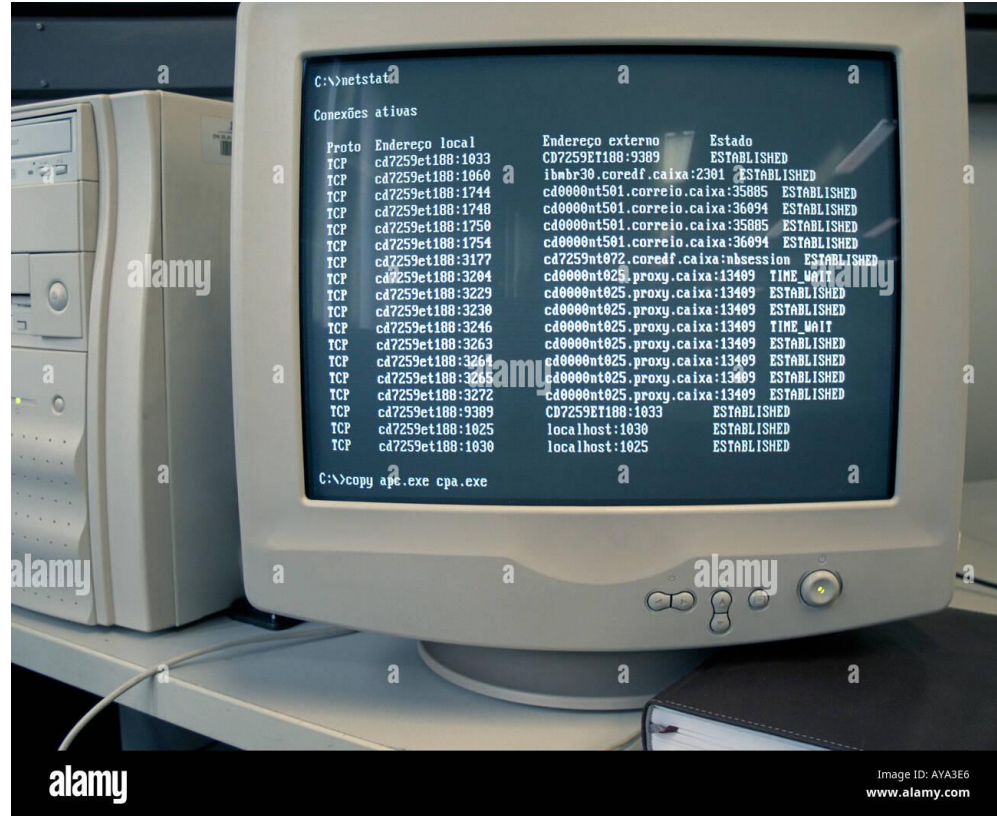
Things to avoid

Not knowing where your files are!

Using "recent" files instead of the hierarchy

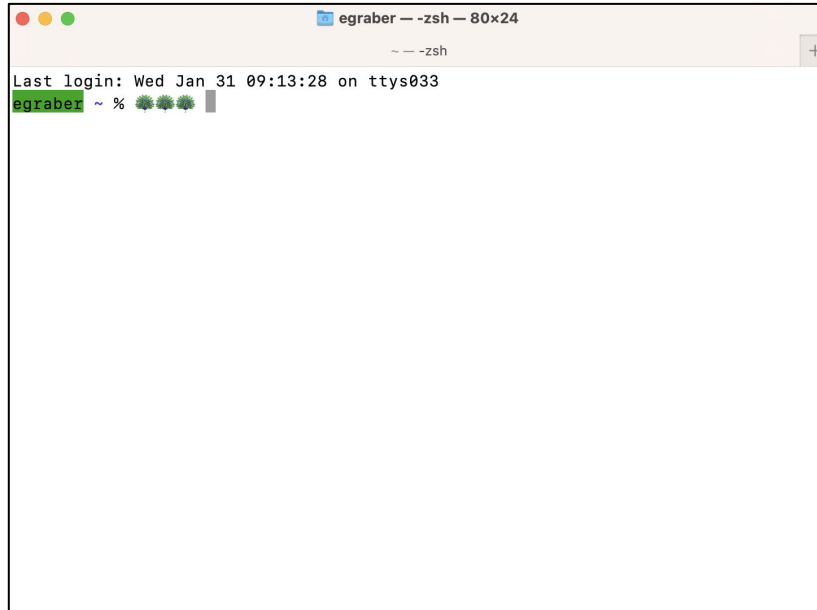
Using spaces in filenames

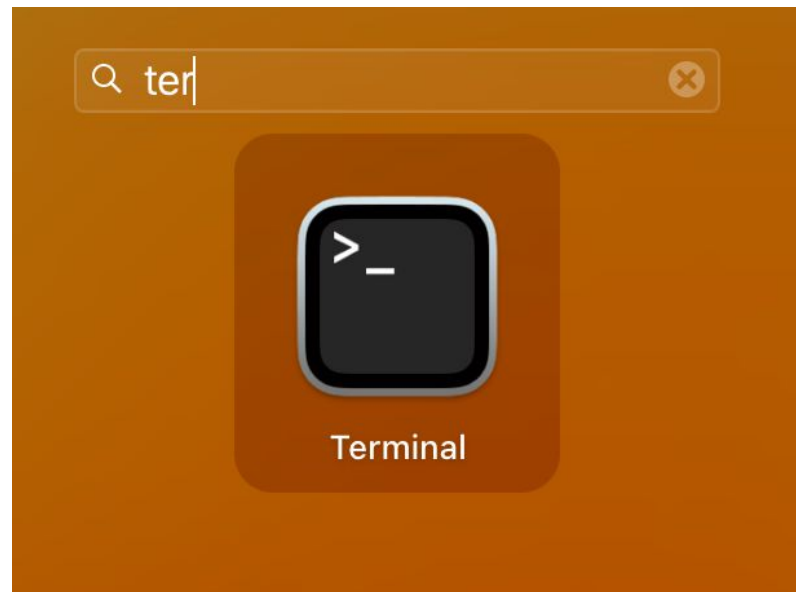
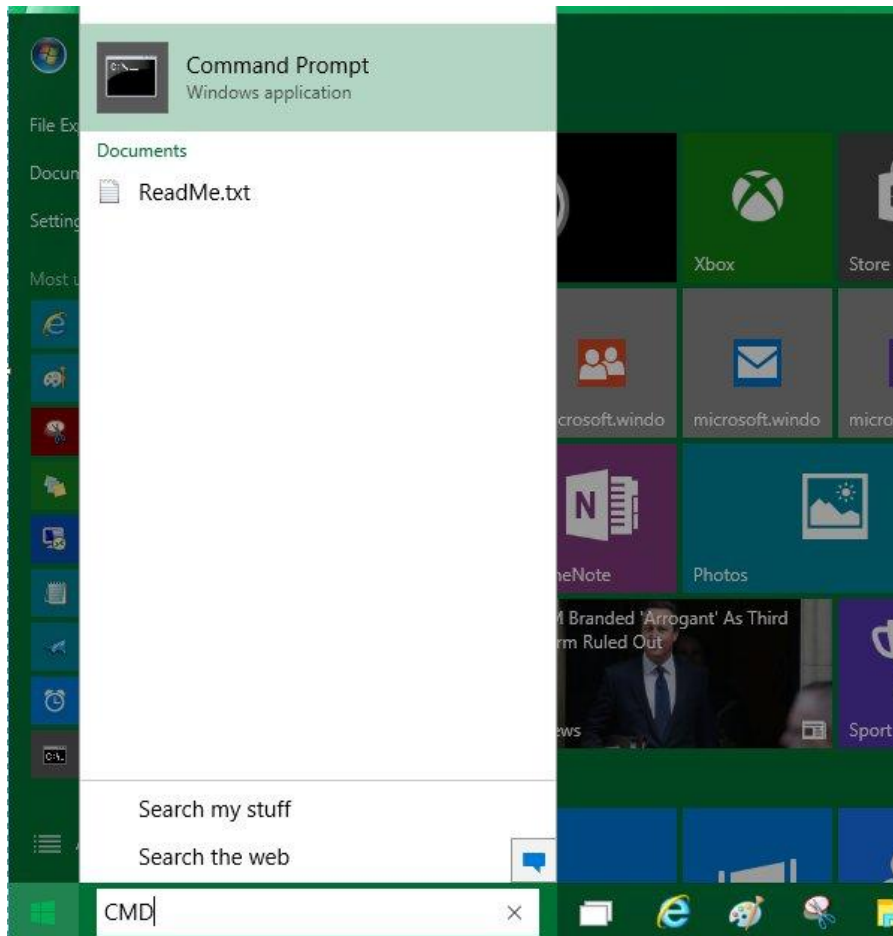
Before there were GUIs, there was the terminal



Terminal

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





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

