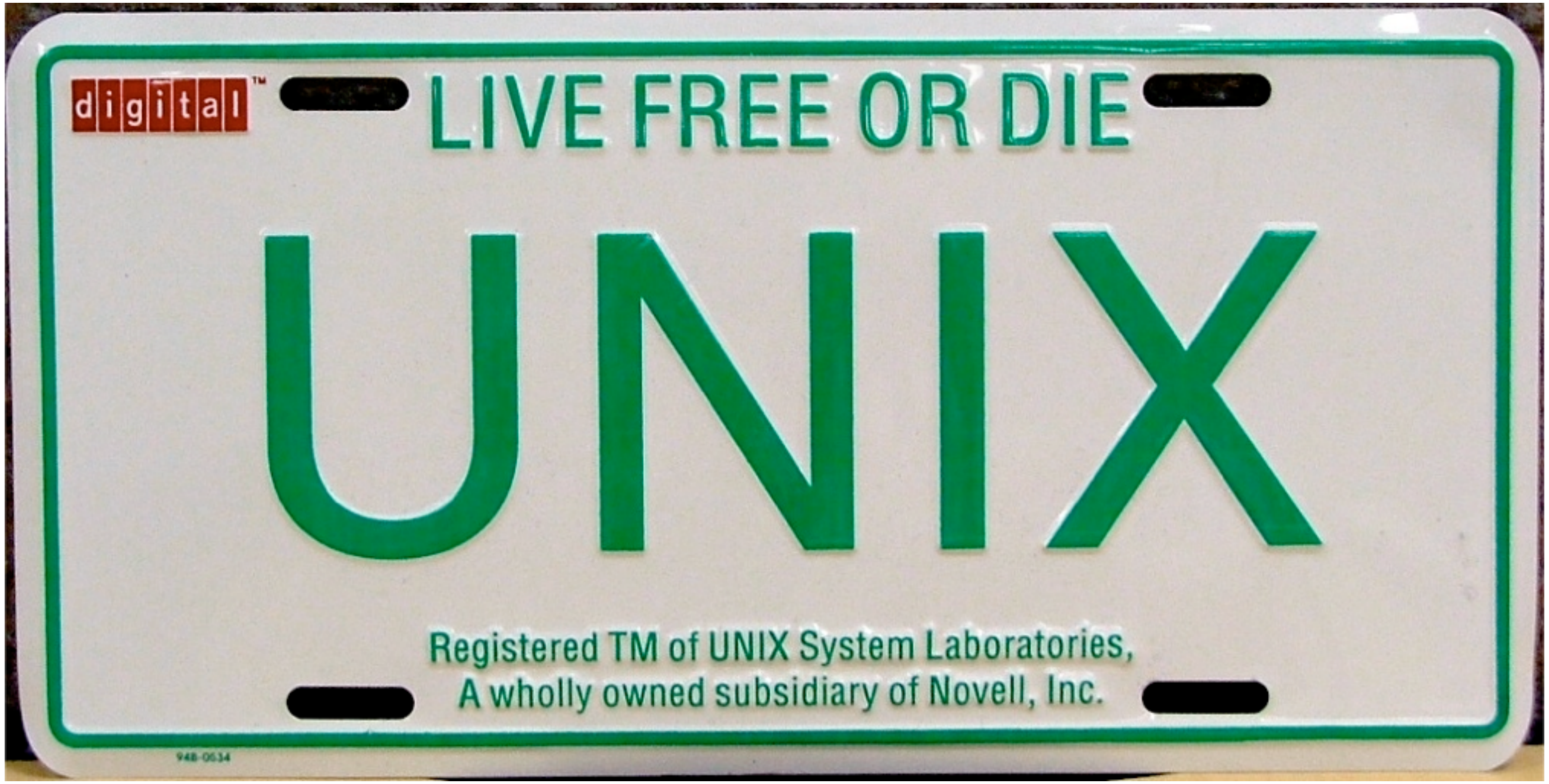
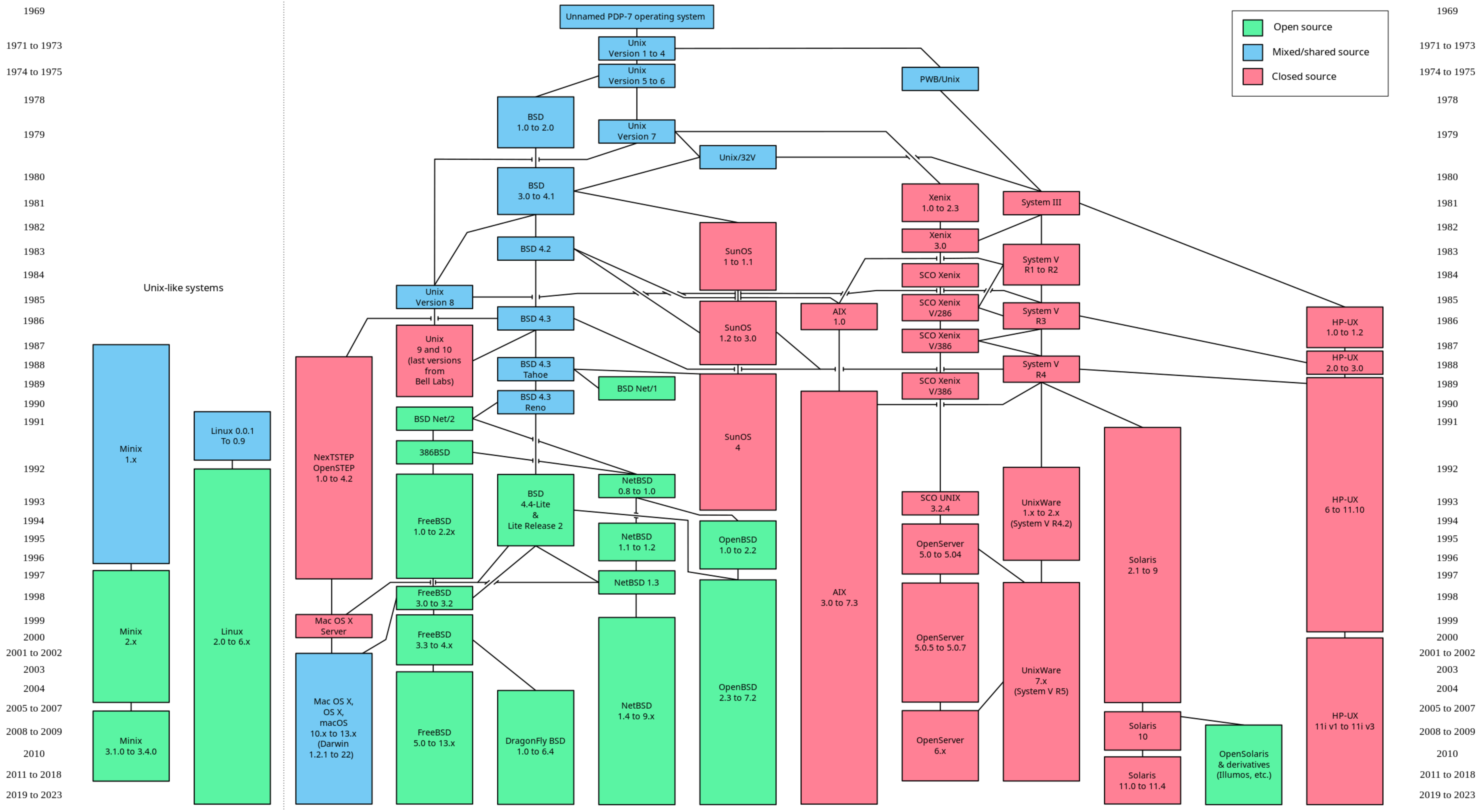
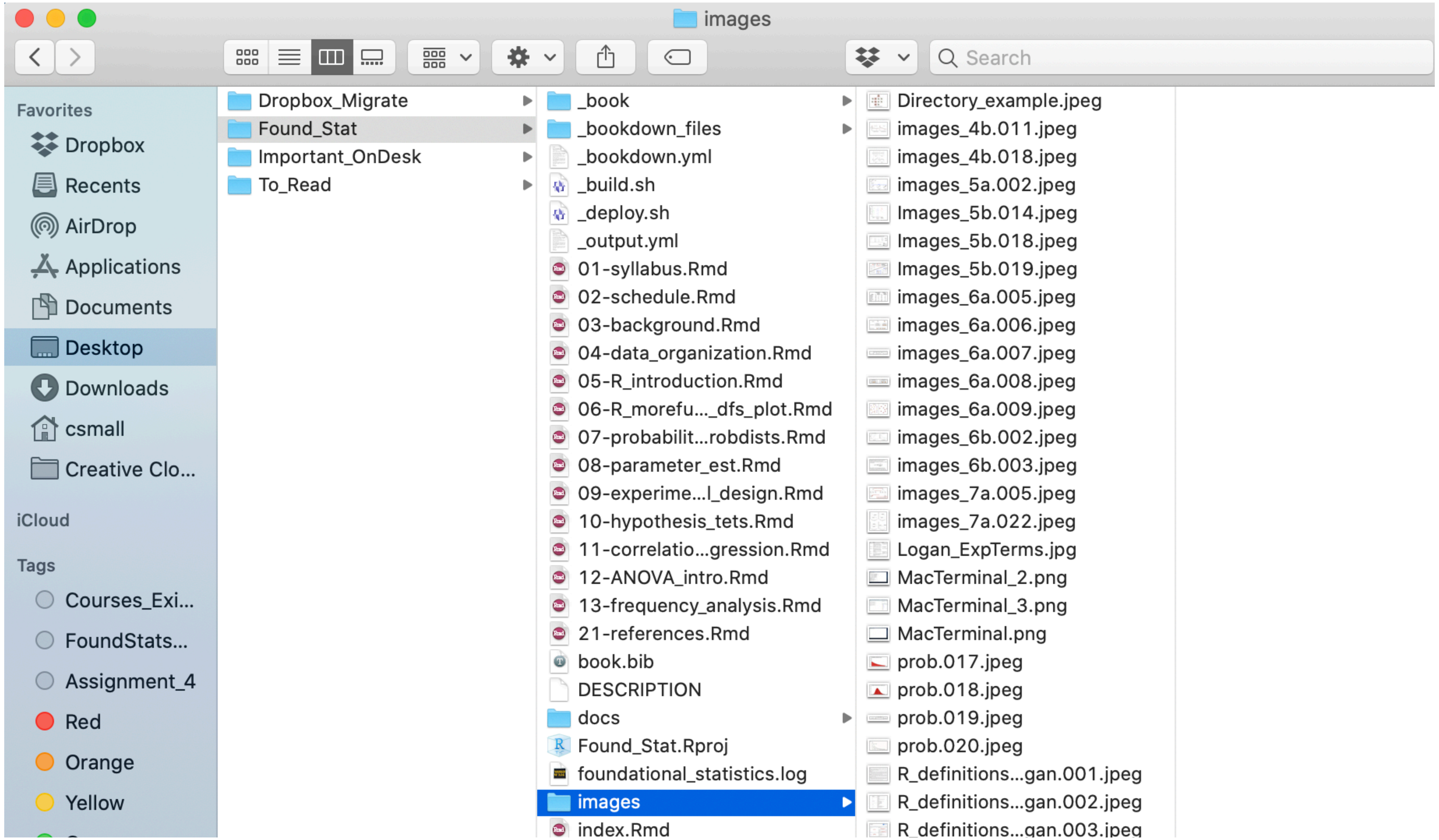
Foundational Statistics for Data Science

Data File Management and Manipulation


digital™
LIVE FREE OR DIE
UNIX
Registered TM of UNIX System Laboratories,
A wholly owned subsidiary of Novell, Inc.
948-0534


Open source
Mixed/shared source
Closed source
1969
1971 to 1973
1974 to 1975
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001 to 2002
2003
2004
2005 to 2007
2008 to 2009
2010
2011 to 2018
2019 to 2023
Unix-like systems
Minix 1.x
Minix 2.x
Minix 3.1.0 to 3.4.0
Linux 0.0.1 To 0.9
Linux 2.0 to 6.x
Unnamed PDP-7 operating system
Unix Version 1 to 4
Unix Version 5 to 6
PWB/Unix
BSD 1.0 to 2.0
Unix Version 7
Unix/32V
BSD 3.0 to 4.1
Xenix 1.0 to 2.3
System III
Xenix 3.0
SunOS 1 to 1.1
BSD 4.2
System V R1 to R2
SCO Xenix
Unix Version 8
AIX 1.0
SCO Xenix V/286
System V R3
HP-UX 1.0 to 1.2
BSD 4.3
SunOS 1.2 to 3.0
SCO Xenix V/386
NexTSTEP OpenSTEP 1.0 to 4.2
Unix 9 and 10 (last versions from Bell Labs)
BSD 4.3 Tahoe
System V R4
HP-UX 2.0 to 3.0
BSD Net/1
SCO Xenix V/386
BSD 4.3 Reno
BSD Net/2
SunOS 4
386BSD
NetBSD 0.8 to 1.0
SCO UNIX 3.2.4
UnixWare 1.x to 2.x (System V R4.2)
HP-UX 6 to 11.10
FreeBSD 1.0 to 2.2x
BSD 4.4-Lite & Lite Release 2
NetBSD 1.1 to 1.2
OpenBSD 1.0 to 2.2
OpenServer 5.0 to 5.04
Solaris 2.1 to 9
NetBSD 1.3
FreeBSD 3.0 to 3.2
AIX 3.0 to 7.3
Mac OS X Server
FreeBSD 3.3 to 4.x
OpenServer 5.0.5 to 5.0.7
UnixWare 7.x (System V R5)
Mac OS X, OS X, macOS 10.x to 13.x (Darwin 1.2.1 to 22)
FreeBSD 5.0 to 13.x
DragonFly BSD 1.0 to 6.4
NetBSD 1.4 to 9.x
OpenBSD 2.3 to 7.2
OpenServer 6.x
Solaris 10
OpenSolaris & derivatives (Illumos, etc.)
Solaris 11.0 to 11.4
HP-UX 11i v1 to 11i v3

1. File and File System Organization and Navigation



Dropbox_Migrate
Found_Stat
Important_OnDesk
To_Read

_book
_bookdown_files
_bookdown.yml
_build.sh
_deploy.sh
_output.yml
01-syllabus.Rmd
02-schedule.Rmd
03-background.Rmd
04-data_organization.Rmd
05-R_introduction.Rmd
06-R_morefu..._dfs_plot.Rmd
07-probabilit...robdists.Rmd
08-parameter_est.Rmd
09-experime...l_design.Rmd
10-hypothesis_tets.Rmd
11-correlatio...gression.Rmd
12-ANOVA_intro.Rmd
13-frequency_analysis.Rmd
21-references.Rmd
book.bib
DESCRIPTION
docs
Found_Stat.Rproj
foundational_statistics.log
images
index.Rmd

Directory_example.jpeg
images_4b.011.jpeg
images_4b.018.jpeg
images_5a.002.jpeg
Images_5b.014.jpeg
Images_5b.018.jpeg
Images_5b.019.jpeg
images_6a.005.jpeg
images_6a.006.jpeg
images_6a.007.jpeg
images_6a.008.jpeg
images_6a.009.jpeg
images_6b.002.jpeg
images_6b.003.jpeg
images_7a.005.jpeg
images_7a.022.jpeg
Logan_ExpTerms.jpg
MacTerminal_2.png
MacTerminal_3.png
MacTerminal.png
prob.017.jpeg
prob.018.jpeg
prob.019.jpeg
prob.020.jpeg
R_definitions...gan.001.jpeg
R_definitions...gan.002.jpeg
R_definitions...gan.003.jpeg

Hierarchical Structure

```
Root -> Users -> csmall -> Desktop -> Found_Stat -> images
```

File and File System Organization and Navigation

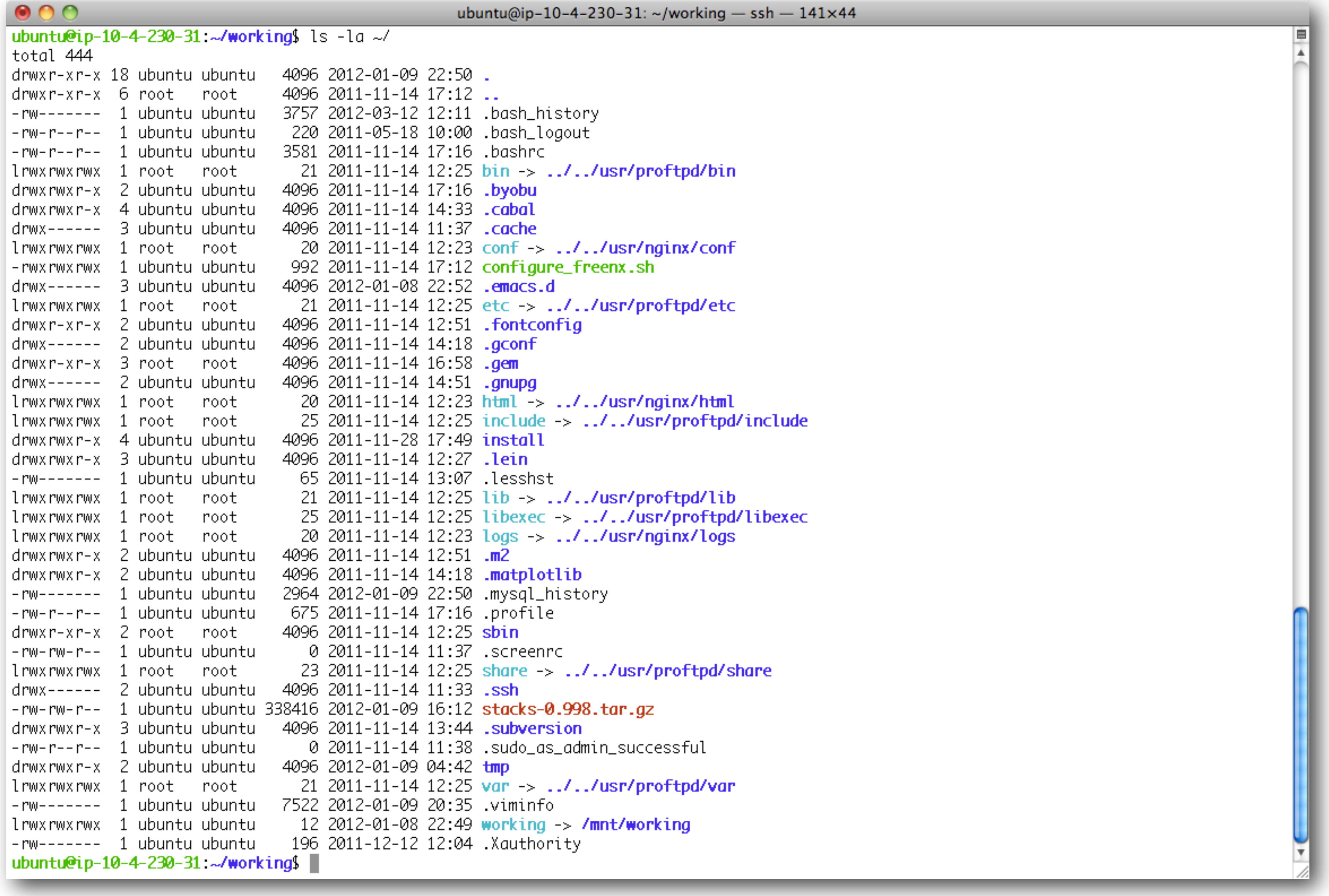
Navigation from the command line is a useful data management skill!

Why?

- Some software can only be run from the terminal (no GUI)
- Working remotely on shared computing resources
- Speed and efficiency!

Unix and Unix-like environments:

Convenient handling and manipulation of data files from a “terminal window”

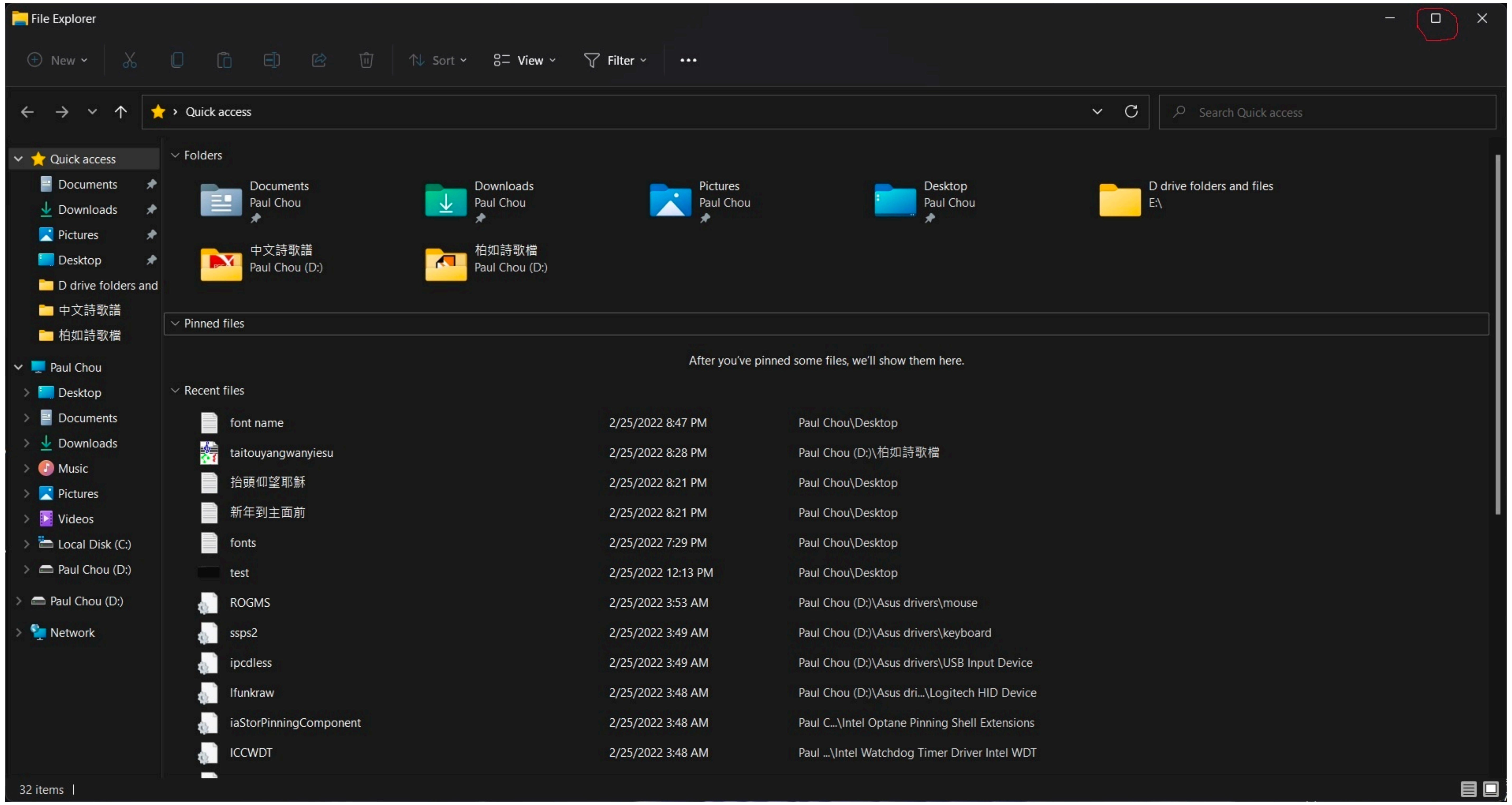


```
ubuntu@ip-10-4-230-31:~/working$ ls -la ~/
total 444
drwxr-xr-x 18 ubuntu ubuntu   4096 2012-01-09 22:50 .
drwxr-xr-x  6 root   root     4096 2011-11-14 17:12 ..
-rw-------  1 ubuntu ubuntu   3757 2012-03-12 12:11 .bash_history
-rw-r--r--  1 ubuntu ubuntu    220 2011-05-18 10:00 .bash_logout
-rw-r--r--  1 ubuntu ubuntu   3581 2011-11-14 17:16 .bashrc
lrwxrwxrwx  1 root   root       21 2011-11-14 12:25 bin -> ../../usr/proftpd/bin
drwxrwxr-x  2 ubuntu ubuntu   4096 2011-11-14 17:16 .byobu
drwxrwxr-x  4 ubuntu ubuntu   4096 2011-11-14 14:33 .cabal
drwx------  3 ubuntu ubuntu   4096 2011-11-14 11:37 .cache
lrwxrwxrwx  1 root   root       20 2011-11-14 12:23 conf -> ../../usr/nginx/conf
-rwxrwxrwx  1 ubuntu ubuntu    992 2011-11-14 17:12 configure_freenx.sh
drwx------  3 ubuntu ubuntu   4096 2012-01-08 22:52 .emacs.d
lrwxrwxrwx  1 root   root       21 2011-11-14 12:25 etc -> ../../usr/proftpd/etc
drwxr-xr-x  2 ubuntu ubuntu   4096 2011-11-14 12:51 .fontconfig
drwx------  2 ubuntu ubuntu   4096 2011-11-14 14:18 .gconf
drwxr-xr-x  3 root   root     4096 2011-11-14 16:58 .gem
drwx------  2 ubuntu ubuntu   4096 2011-11-14 14:51 .gnupg
lrwxrwxrwx  1 root   root       20 2011-11-14 12:23 html -> ../../usr/nginx/html
lrwxrwxrwx  1 root   root       25 2011-11-14 12:25 include -> ../../usr/proftpd/include
drwxrwxr-x  4 ubuntu ubuntu   4096 2011-11-28 17:49 install
drwxrwxr-x  3 ubuntu ubuntu   4096 2011-11-14 12:27 .lein
-rw-------  1 ubuntu ubuntu     65 2011-11-14 13:07 .lesshst
lrwxrwxrwx  1 root   root       21 2011-11-14 12:25 lib -> ../../usr/proftpd/lib
lrwxrwxrwx  1 root   root       25 2011-11-14 12:25 libexec -> ../../usr/proftpd/libexec
lrwxrwxrwx  1 root   root       20 2011-11-14 12:23 logs -> ../../usr/nginx/logs
drwxrwxr-x  2 ubuntu ubuntu   4096 2011-11-14 12:51 .m2
drwxrwxr-x  2 ubuntu ubuntu   4096 2011-11-14 14:18 .matplotlib
-rw-------  1 ubuntu ubuntu   2964 2012-01-09 22:50 .mysql_history
-rw-r--r--  1 ubuntu ubuntu    675 2011-11-14 17:16 .profile
drwxr-xr-x  2 root   root     4096 2011-11-14 12:25 sbin
-rw-rw-r--  1 ubuntu ubuntu      0 2011-11-14 11:37 .screenrc
lrwxrwxrwx  1 root   root       23 2011-11-14 12:25 share -> ../../usr/proftpd/share
drwx------  2 ubuntu ubuntu   4096 2011-11-14 11:33 .ssh
-rw-rw-r--  1 ubuntu ubuntu 338416 2012-01-09 16:12 stacks-0.998.tar.gz
drwxrwxr-x  3 ubuntu ubuntu   4096 2011-11-14 13:44 .subversion
-rw-r--r--  1 ubuntu ubuntu      0 2011-11-14 11:38 .sudo_as_admin_successful
drwxrwxr-x  2 ubuntu ubuntu   4096 2012-01-09 04:42 tmp
lrwxrwxrwx  1 root   root       21 2011-11-14 12:25 var -> ../../usr/proftpd/var
-rw-------  1 ubuntu ubuntu   7522 2012-01-09 20:35 .viminfo
lrwxrwxrwx  1 ubuntu ubuntu     12 2012-01-08 22:49 working -> /mnt/working
-rw-------  1 ubuntu ubuntu    196 2011-12-12 12:04 .Xauthority
ubuntu@ip-10-4-230-31:~/working$
```

- •Apple OS X Macs
- •Linux workstations and servers (and HPCs)
- •Virtual Machines
- •Google’s Android phones

Aside (for folks running Windows Subsystem for Linux):

If you want to access / navigate to the files on your primary Windows partition



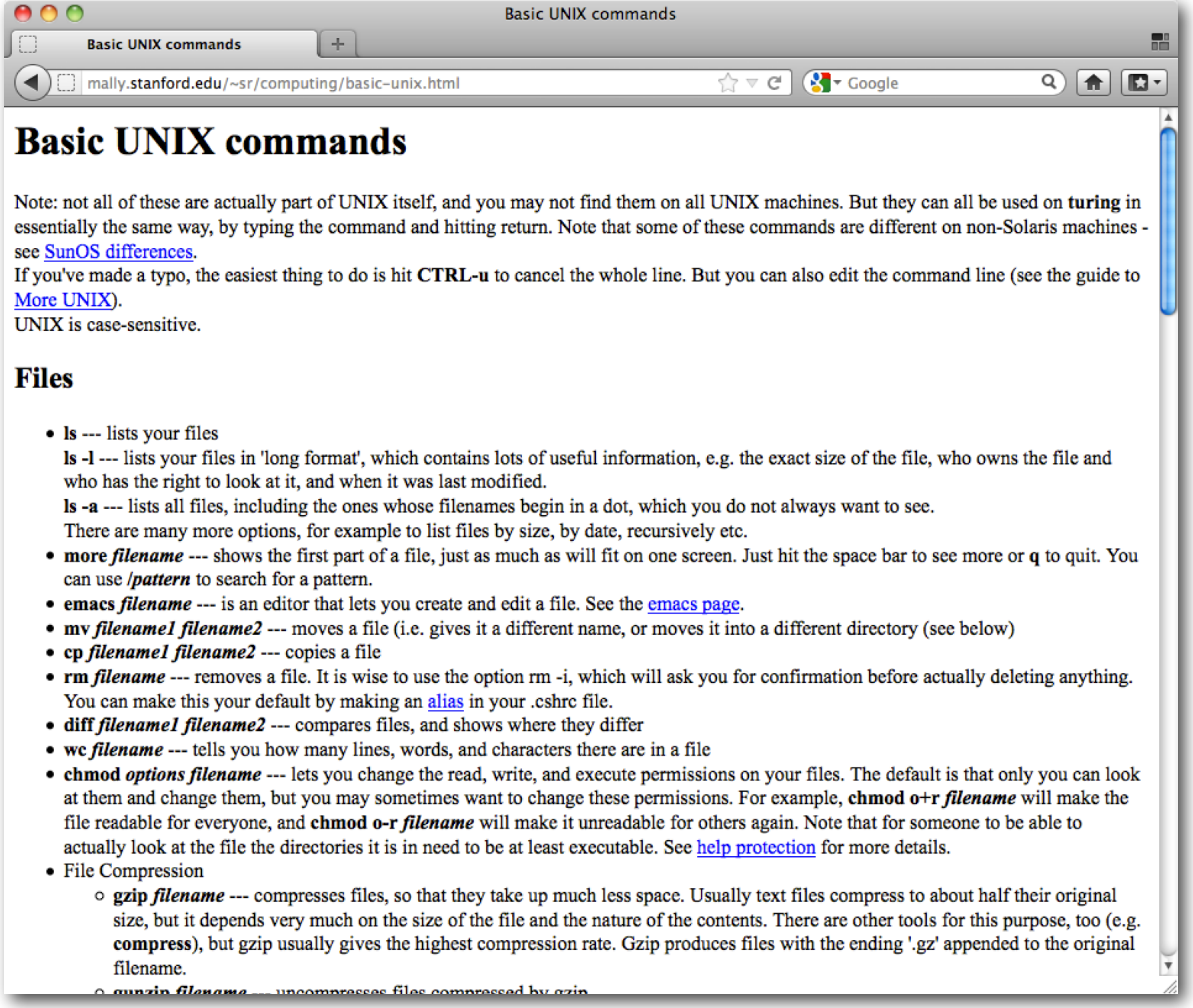
After you've pinned some files, we'll show them here.

font name	2/25/2022 8:47 PM	Paul Chou\Desktop
taitouyangwanyiesu	2/25/2022 8:28 PM	Paul Chou (D:)\柏如詩歌檔
抬頭仰望耶穌	2/25/2022 8:21 PM	Paul Chou\Desktop
新年到主面前	2/25/2022 8:21 PM	Paul Chou\Desktop
fonts	2/25/2022 7:29 PM	Paul Chou\Desktop
test	2/25/2022 12:13 PM	Paul Chou\Desktop
ROGMS	2/25/2022 3:53 AM	Paul Chou (D:)\Asus drivers\mouse
ssps2	2/25/2022 3:49 AM	Paul Chou (D:)\Asus drivers\keyboard
ipcdless	2/25/2022 3:49 AM	Paul Chou (D:)\Asus drivers\USB Input Device
lfunkraw	2/25/2022 3:48 AM	Paul Chou (D:)\Asus dri...\Logitech HID Device
iaStorPinningComponent	2/25/2022 3:48 AM	Paul C...\Intel Optane Pinning Shell Extensions
ICCWDT	2/25/2022 3:48 AM	Paul ...\Intel Watchdog Timer Driver Intel WDT

```
cd /mnt/c/Users/<your_Windows_username>
```

UNIX commands help us navigate and edit

google “unix commands” to find your own “cheat sheet”



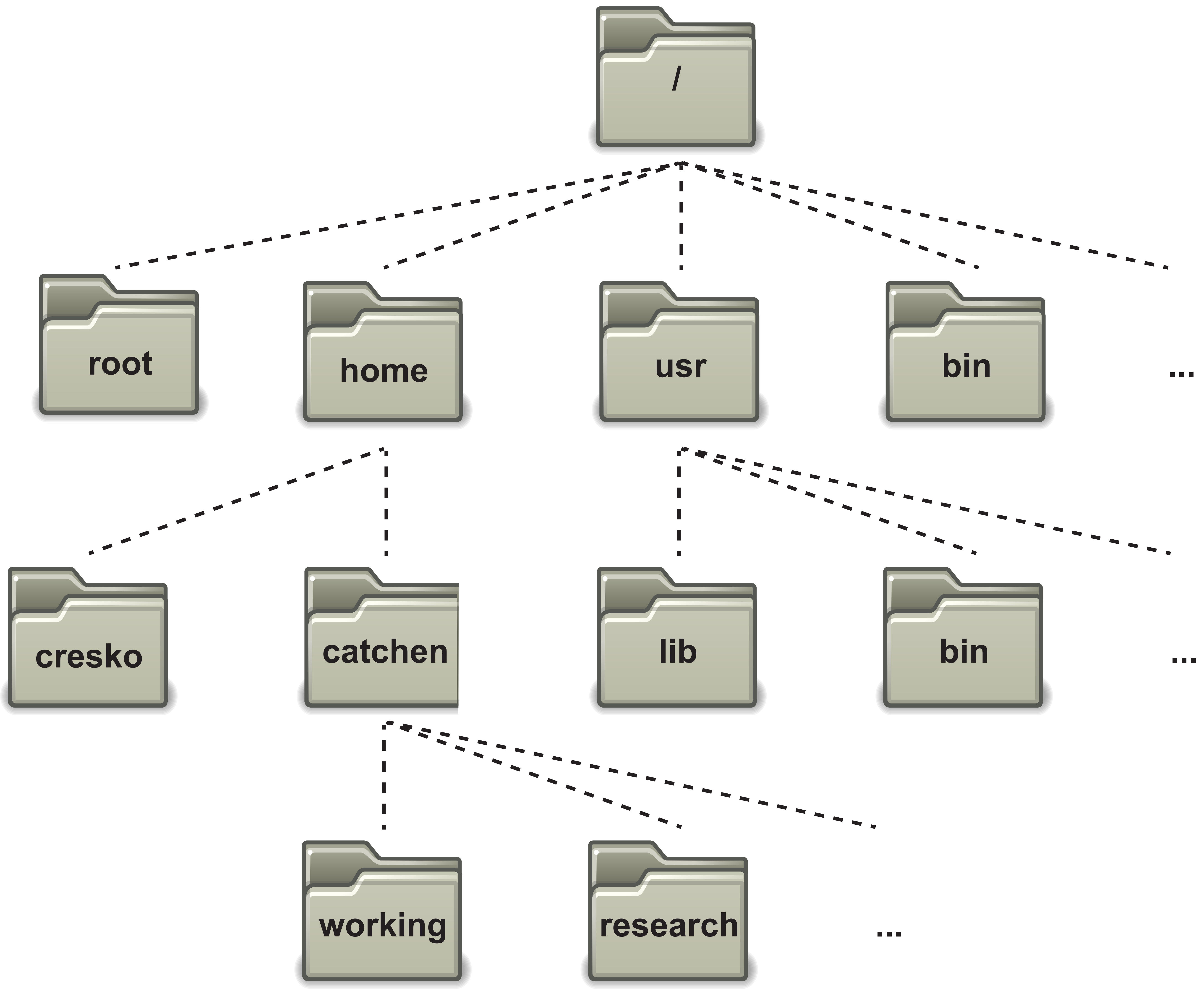
Basic UNIX commands

Note: not all of these are actually part of UNIX itself, and you may not find them on all UNIX machines. But they can all be used on **turing** in essentially the same way, by typing the command and hitting return. Note that some of these commands are different on non-Solaris machines - see SunOS differences.
If you've made a typo, the easiest thing to do is hit **CTRL-u** to cancel the whole line. But you can also edit the command line (see the guide to More UNIX).
UNIX is case-sensitive.

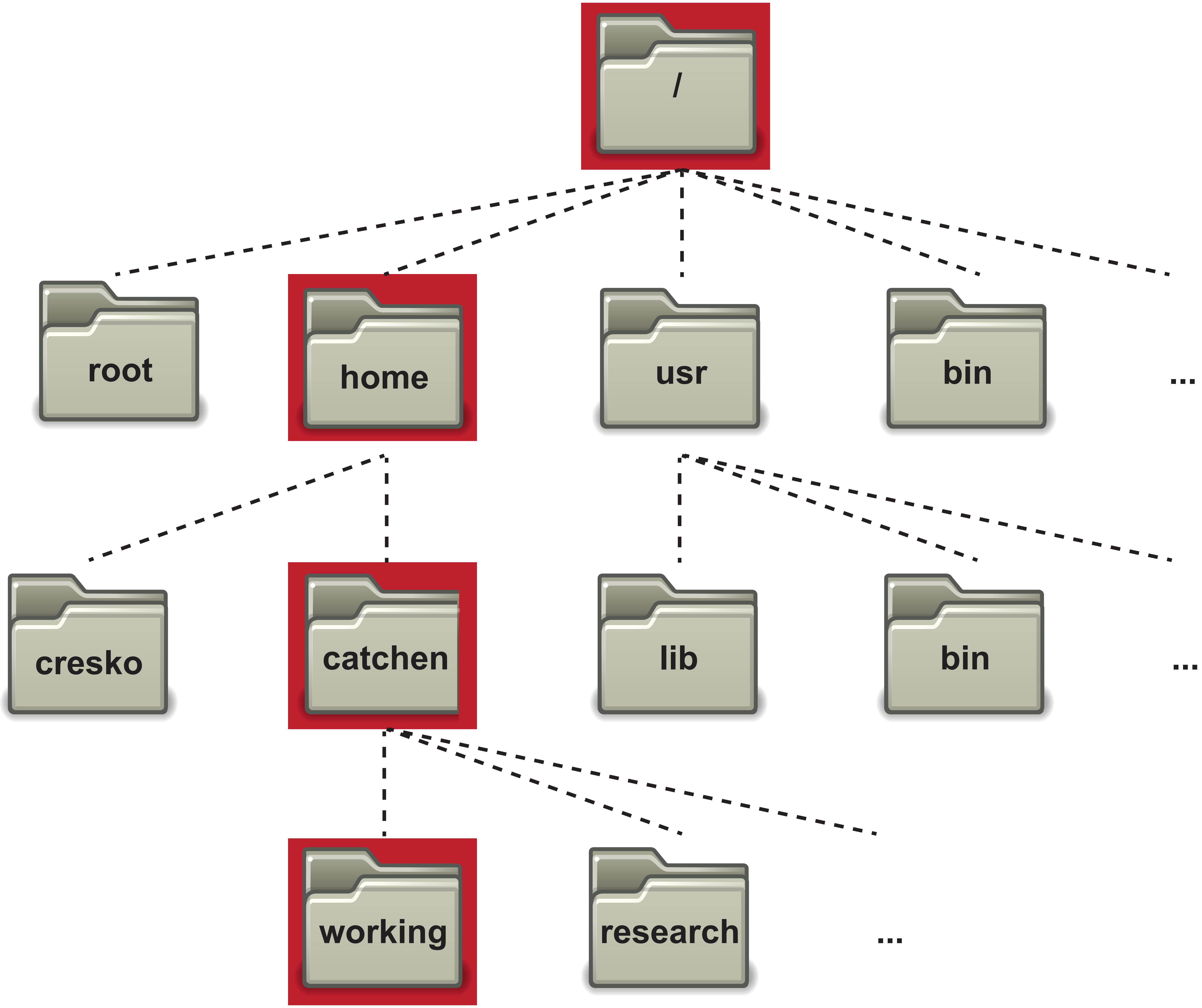
Files

- **ls** --- lists your files
  ls -l --- lists your files in 'long format', which contains lots of useful information, e.g. the exact size of the file, who owns the file and who has the right to look at it, and when it was last modified.
  ls -a --- lists all files, including the ones whose filenames begin in a dot, which you do not always want to see.
  There are many more options, for example to list files by size, by date, recursively etc.
- **more *filename*** --- shows the first part of a file, just as much as will fit on one screen. Just hit the space bar to see more or **q** to quit. You can use /***pattern*** to search for a pattern.
- **emacs *filename*** --- is an editor that lets you create and edit a file. See the emacs page.
- **mv *filename1 filename2*** --- moves a file (i.e. gives it a different name, or moves it into a different directory (see below)
- **cp *filename1 filename2*** --- copies a file
- **rm *filename*** --- removes a file. It is wise to use the option rm -i, which will ask you for confirmation before actually deleting anything. You can make this your default by making an alias in your .cshrc file.
- **diff *filename1 filename2*** --- compares files, and shows where they differ
- **wc *filename*** --- tells you how many lines, words, and characters there are in a file
- **chmod** *options* ***filename*** --- lets you change the read, write, and execute permissions on your files. The default is that only you can look at them and change them, but you may sometimes want to change these permissions. For example, **chmod o+r *filename*** will make the file readable for everyone, and **chmod o-r *filename*** will make it unreadable for others again. Note that for someone to be able to actually look at the file the directories it is in need to be at least executable. See help protection for more details.
- File Compression
  - **gzip *filename*** --- compresses files, so that they take up much less space. Usually text files compress to about half their original size, but it depends very much on the size of the file and the nature of the contents. There are other tools for this purpose, too (e.g. **compress**), but gzip usually gives the highest compression rate. Gzip produces files with the ending '.gz' appended to the original filename.
  - **gunzip *filename*** --- uncompresses files compressed by gzip.

In UNIX everything is a file organized in a hierarchy

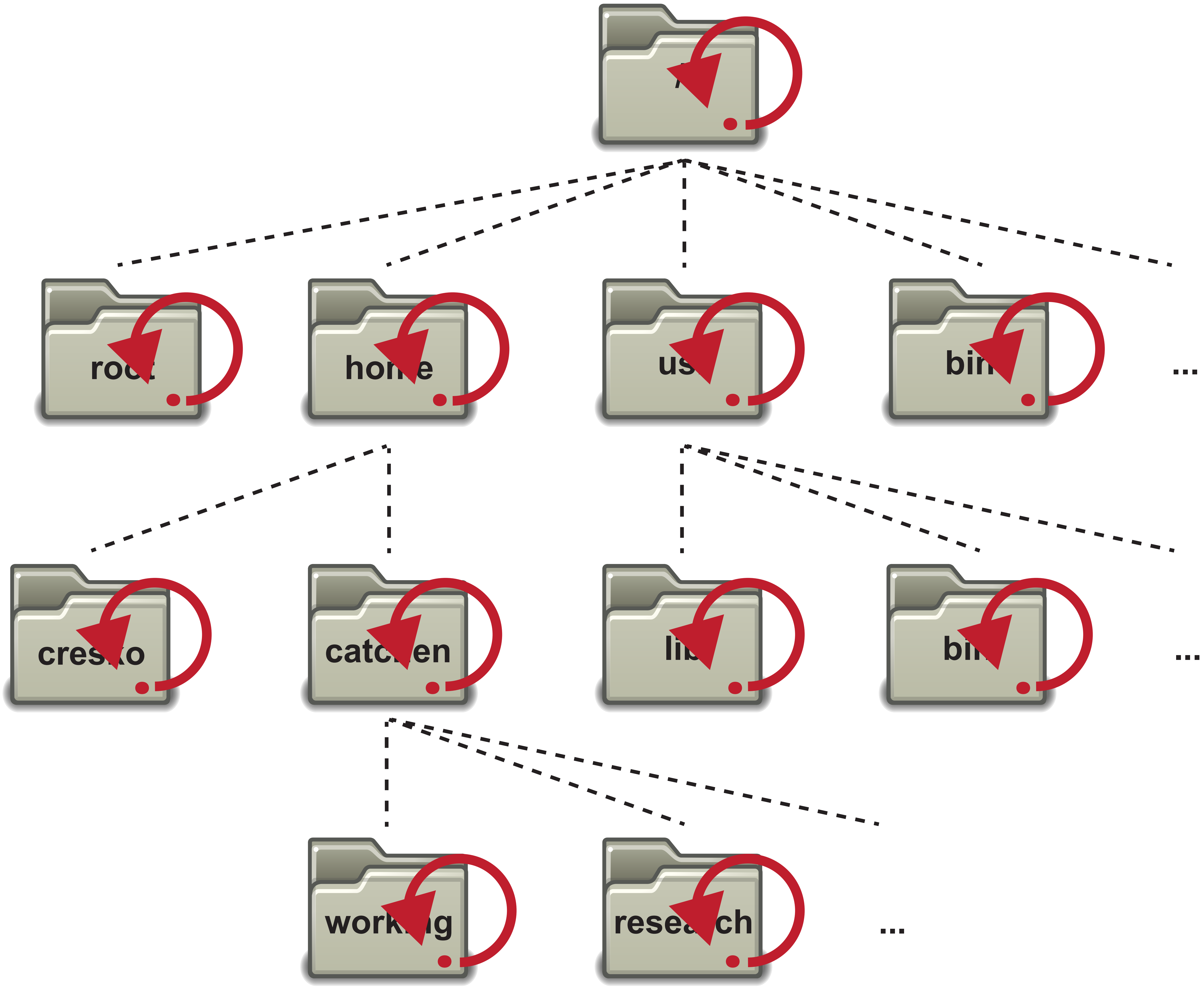


Paths



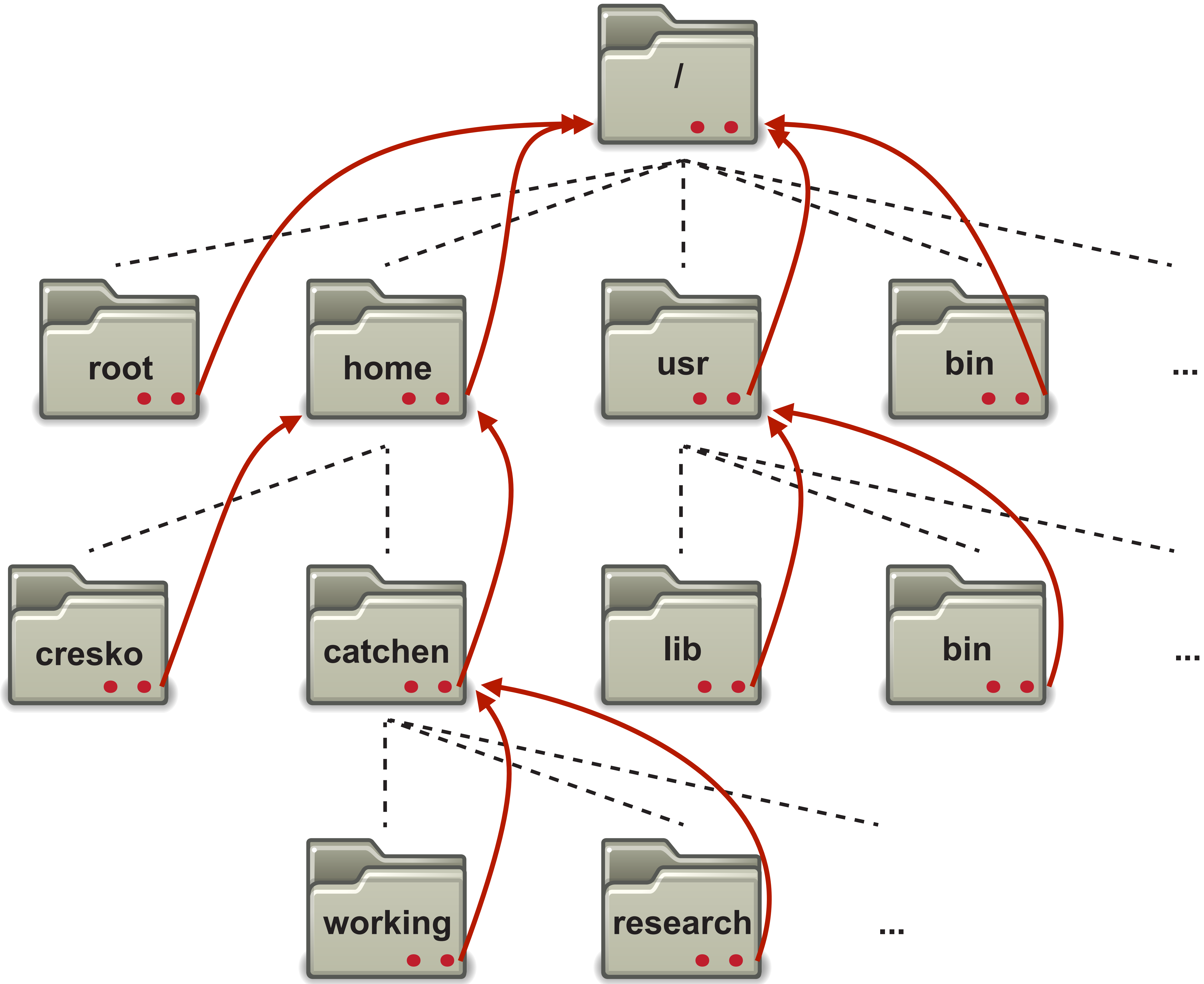
`/home/catchen/working`

Special files -- ‘*dot*’



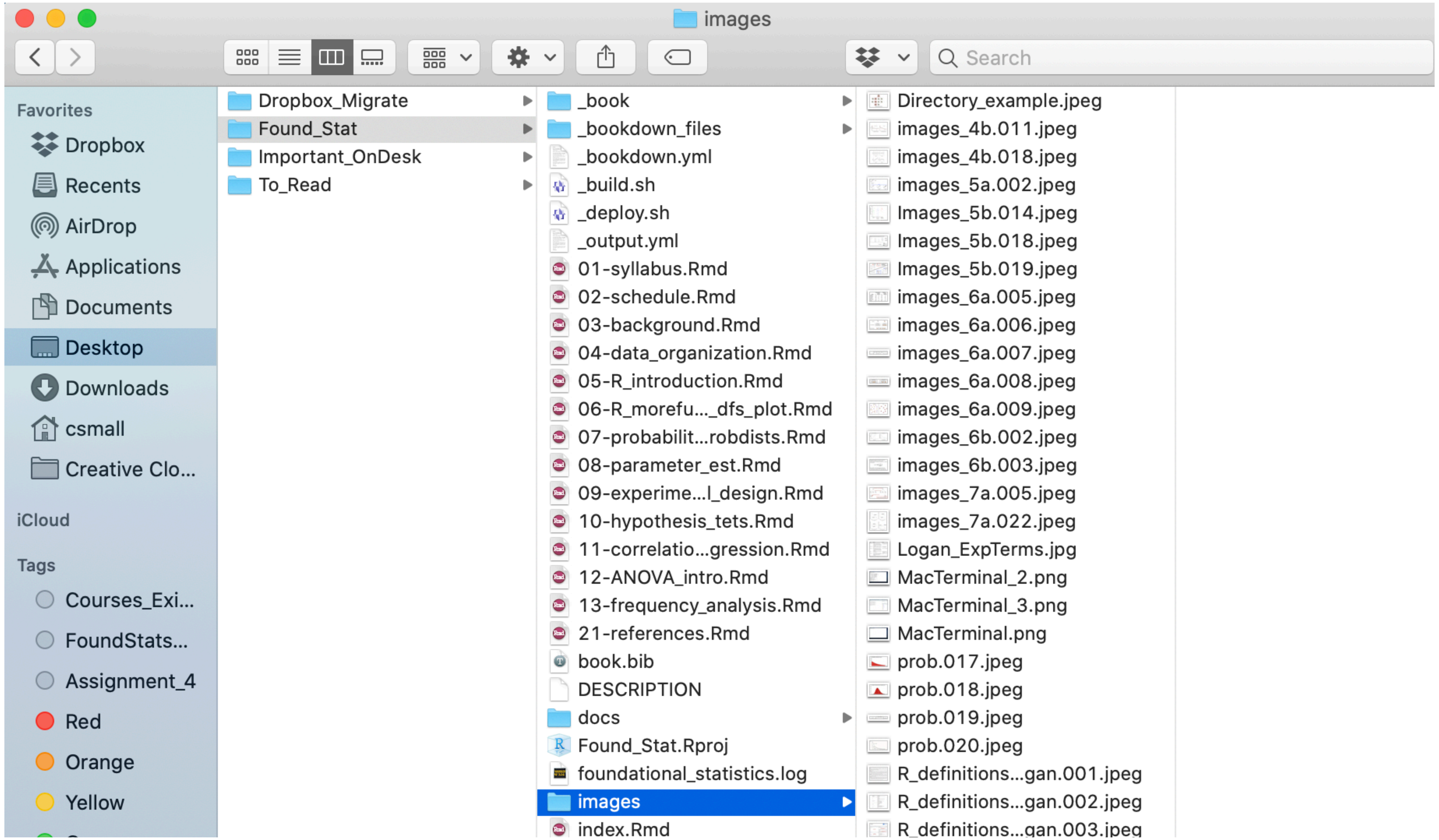
Is a representation for “current directory”

Special files -- ‘*dot dot*’



Is a representation for “the directory above”

“absolute paths” include the full path, from root to endpoint



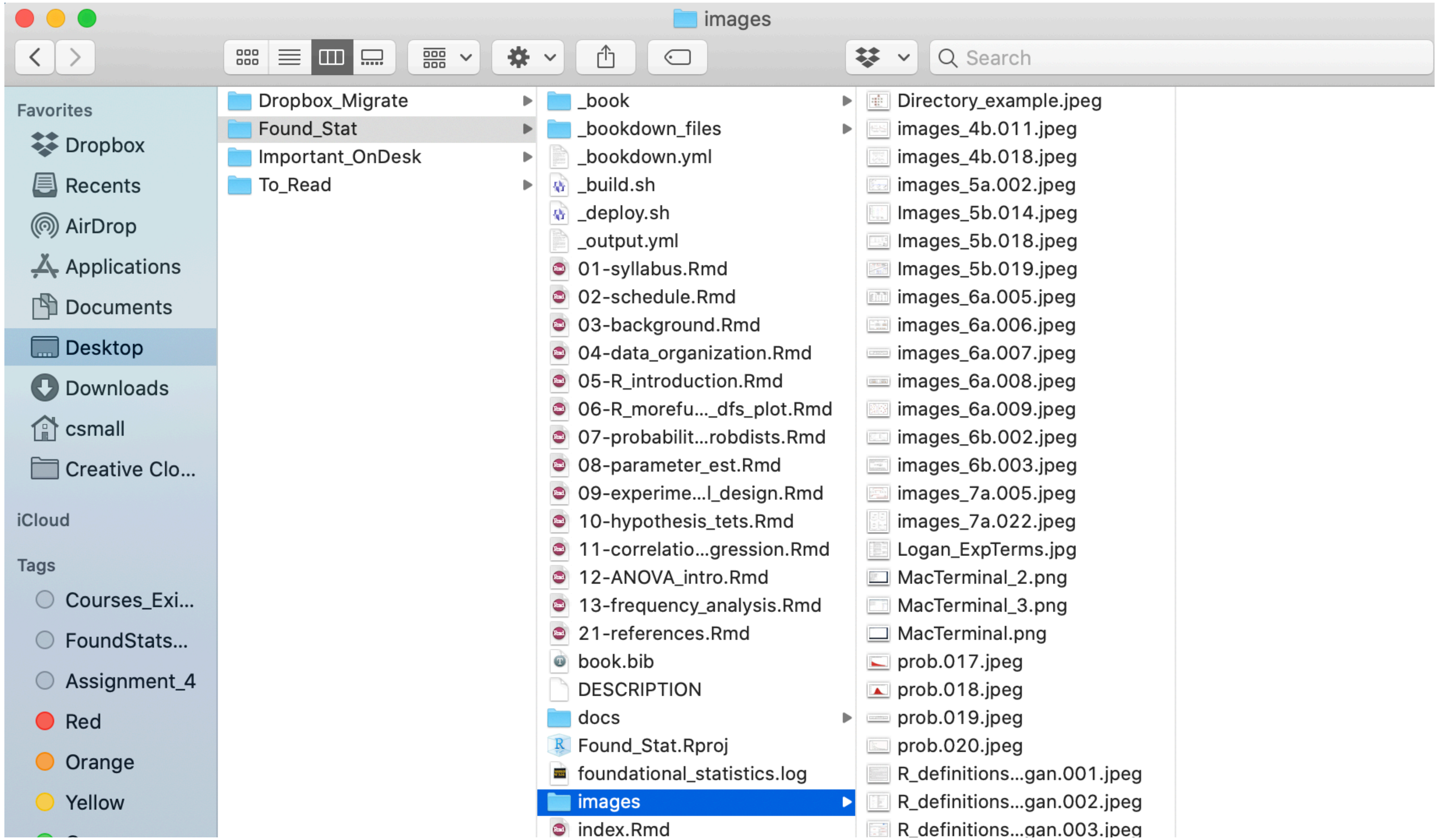
Hierarchical Structure

```
Root -> Users -> csmall -> Desktop -> Found_Stat -> images
```

Absolute Path:

```
/Users/csmall/Desktop/Found_Stat/images
```

"relative paths" use your current location as a reference point



Dropbox_Migrate
Found_Stat
Important_OnDesk
To_Read

_book
_bookdown_files
_bookdown.yml
_build.sh
_deploy.sh
_output.yml
01-syllabus.Rmd
02-schedule.Rmd
03-background.Rmd
04-data_organization.Rmd
05-R_introduction.Rmd
06-R_morefu..._dfs_plot.Rmd
07-probabilit...robdists.Rmd
08-parameter_est.Rmd
09-experime...l_design.Rmd
10-hypothesis_tets.Rmd
11-correlatio...gression.Rmd
12-ANOVA_intro.Rmd
13-frequency_analysis.Rmd
21-references.Rmd
book.bib
DESCRIPTION
docs
Found_Stat.Rproj
foundational_statistics.log
images
index.Rmd

Directory_example.jpeg
images_4b.011.jpeg
images_4b.018.jpeg
images_5a.002.jpeg
Images_5b.014.jpeg
Images_5b.018.jpeg
Images_5b.019.jpeg
images_6a.005.jpeg
images_6a.006.jpeg
images_6a.007.jpeg
images_6a.008.jpeg
images_6a.009.jpeg
images_6b.002.jpeg
images_6b.003.jpeg
images_7a.005.jpeg
images_7a.022.jpeg
Logan_ExpTerms.jpg
MacTerminal_2.png
MacTerminal_3.png
MacTerminal.png
prob.017.jpeg
prob.018.jpeg
prob.019.jpeg
prob.020.jpeg
R_definitions...gan.001.jpeg
R_definitions...gan.002.jpeg
R_definitions...gan.003.jpeg

We are in the `images` directory, but we want to reference a file called `doc_1.Rmd` in `docs`

Relative Path:

```
../docs/doc_1.Rmd
```

Important UNIX navigation and organization commands

`pwd` - prints working directory

`ls` - lists directory contents

`cd` - changes your current directory

`mkdir` - makes a new directory

`touch` - creates a new, empty file

`rmdir` - deletes an empty directory

`rm` - deletes a file, or (with the -r flag) a directory and its contents

`cp` - copies a file or directory

`mv` - moves or renames a file or directory

Structure of a command

```
$ command -options arguments
```

```
$ ls -lh /Users/csmall/Desktop
```

Take some time to practice using one or more of these commands

`pwd` - prints working directory

`ls` - lists directory contents

`cd` - changes your current directory

`mkdir` - makes a new directory

`touch` - creates a new, empty file

`rmdir` - deletes an empty directory

`rm` - deletes a file, or (with the -r flag) a directory and its contents

`cp` - copies a file or directory

`mv` - moves or renames a file or directory

2. File summary, formatting, and manipulation

The command line is also very useful for “looking at” a file’s contents, summarizing it, or reformatting it!

- Avoids having to load large files into spreadsheet software
- Don’t always have spreadsheet “GUI” software available
- Very fast and efficient (“heavy lifting” for big files)
- NOTE: Can also do some of this (esp. formatting) effectively within R

Many ways to view a file

`more`	`less`	`head`	`tail`	`cat`
View text file one screen at a time	Same as more, but allows backwards movement	View the first 10 lines of a file	View the last 10 lines of a file	Spit out the whole file at once
Space-bar: scroll q: quit	Arrow keys: scroll Space-bar: end of file q: quit	-n num Controls the number of lines	-n num Controls the number of lines	

Many ways to summarize or “subset” a file

`wc`	`cut`	`sort`	`grep`
“word count”	Isolates “columns” of a file if field-delimited (e.g. csv)	Sort file based on a given column	Find lines in a file with specific patterns
`-l` (counts lines)	`-f` (column number) `-d` (column delimiter)	`-k` (column key) `-r` (reverse) `-n` (numeric) `-u` (unique)	`-c` (counts line matches) `-v` (“reverse”: finds lines without the pattern)

`sed` - a powerful “search and replace” tool

- `s/query/replacement/flag`
  - The pattern to find: `(query)`
  - The text you want to swap: `(replacement)`
  - Options for frequency of replacements: `(flag)`

```
$ sed ’s/sample1/sampleA/g’ file.txt
```

(the “global” `g` flag specifies ALL replacements)

awk - the UNIX “utility knife”

It’s actually a stream-based programming language!

- Works effectively in a field (column)-wise manner
  `$0` - entire line
  `$1` - column 1
  `$2` - column 2 , etc.

- Complex, with multiple built-in functions (e.g. `print`)

- Patterns can be logical evaluations

`$3 > 0`	if column 3 is greater than 0
`$1 == 32`	if column 1 equals 32
`$1 == “consensus”`	if column 1 is the string “consensus”

```
$ awk -F, ‘$1 > 1000 {print $1,$2}’ ./file.txt
```

“pipes” and “carrots”- controlling information flow

| (“pipe”) - passes output from one command as input to another

```
cat ./file.txt | grep ‘data’ | wc -l
```

< (STDIN) - an input stream going into a command

```
wc -l < ./file.txt
wc -l <(cat ./file.txt ./file2.txt)
```

> (STDOUT) - writes stream output to a specified file name (will overwrite)

```
wc -l ./file.txt > ./wc.txt
```

>> will append to a file instead of overwriting

Take some time to practice subsetting and reformatting text files

Do these for data files and data sets

- Store a copy of data in nonproprietary software and hardware formats, such as plain ASCII text (aka a flat file). Good options are tab- or comma-separated
- Leave an uncorrected file when doing analyses
- Use **descriptive** names for your data files and variables
- Include a **header line** with descriptive variable names
- Maintain effective **metadata** about the data
- When you add new **observations** to a file or spreadsheet, add rows, not columns
- When you add new **variables** to a database, add columns, not rows
- Use a scripted program like **R** for analysis, and **RMarkdown** for documentation and clear presentation

Do not do this for data files and data sets

- Don’t include multiple data types in the same column

- Don’t use non-alphanumeric characters in file or directory names (_ is okay)

- Don’t use characters that are common field delimiters in individual data entries (e.g. “medium,2”)

- Don’t copy and paste data from rich text-formatted files (like Microsoft Word, .pdfs, etc.) into primary data files.