Unix Shell

MAS Departmental Disclaimers:

For students trying to take or audit from outside the MAS program.

Taking or auditing 400 courses is simply is not permitted because this is a self-supporting program. Sorry, unfortunately, you will NOT be able to take any of the 400 level Stats courses.

There are NO exceptions that can be made by the department. These classes were designed specifically for students who applied directly to the program.

The students of this program are also not allowed to audit or enroll in classes outside of the program as it was created for working professionals.

If you would like to apply for the program, you are welcome to do so: https://master.stat.ucla.edu/admissions/

Information is found here: https://master.stat.ucla.edu/ And here: https://master.stat.ucla.edu/faq/

The shell: why is it important??

to introduce the shell means to have a discussion about the structure of the computer, operating systems, file systems and history

the shell offers programmatic access to a computer's underlying parts, providing the ability to "do" data analysis on directories, on processes, and on their networks

as there are many flavors of the shell, it is the first time we come to a decision about choosing 'tools', about evaluating which shell is best for you and the shifting terrain of software development

The shell: why is it important?

as a practical matter, shell tools are an indispensable part for my own data science practice, data ‘cleaning’, data processing, exploratory analysis, automation; by design they let us deal with data on a scale that can be difficult from with R or Python

it also provides a low level link to other data platforms, data sources, and remote environments

The shell: how we'll learn

everyone needs access to the shell. everyone with a Mac and Linux already have one. those with Windows machines need something else..

there are instructions on the site for Windows users to get a version of it, but..

as we all have docker installed we can use an image that provides us all with the same environment for learning and exploration. not all functionality works in the image but overall will be useful for us (also we've seem a version in our docker Rstudio environment)

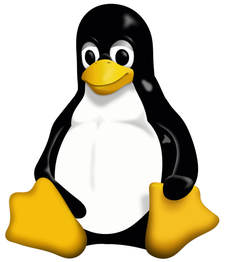
Operating Systems

most devices that contain a computer of some kind will have an OS, they tend to emerge when the appliance will have to deal with new applications, complex user-input and possibly changing requirements on its function

Operating Systems

An operating system is a piece of software that organizes and controls hardware and other software so your computer behaves in a flexible but predictable way

maybe obviously to us all, Window, MacOS and Linux are those most commonly used for home computers. iOS and Android are most common on mobile devices



A history

In 1964, Bell Labs partnered with MIT and GE to create Multics (Multiplexed Information and Computing Service)

“Such systems must run continuously and reliably 7 days a week, 24 hours a day in a way similar to telephone or power systems, and must be capable of meeting wide service demand from multiple man-machine interaction to the sequential processing of absentee-user jobs, from the use of the system with dedicated languages and subsystems to the programming of the stem itself”

A history

Bell Labs pulled out of the Multics project in 1969, a group of researchers at Bell Labs started work on Unica (uniplexed information and computing system) because initially it could only support one user; as the system matured it was renamed Unix, which isn't an acronym for anything

Ritchie simply says that Unix is a 'somewhat treacherous pun on Multics'

The Unix filesystem

Multics brought about the first notion of a hierarchical file system; files were arranged in a tree structure allowing users to have control of they on areas

Unix began (more or less) as a file system and then an interactive shell emerged to let you examine its contest and perform basic operations

The kernel and the shell

the Unix kernel is the part of the operating system that carries out basic functions like accessing files, handing communication, and others

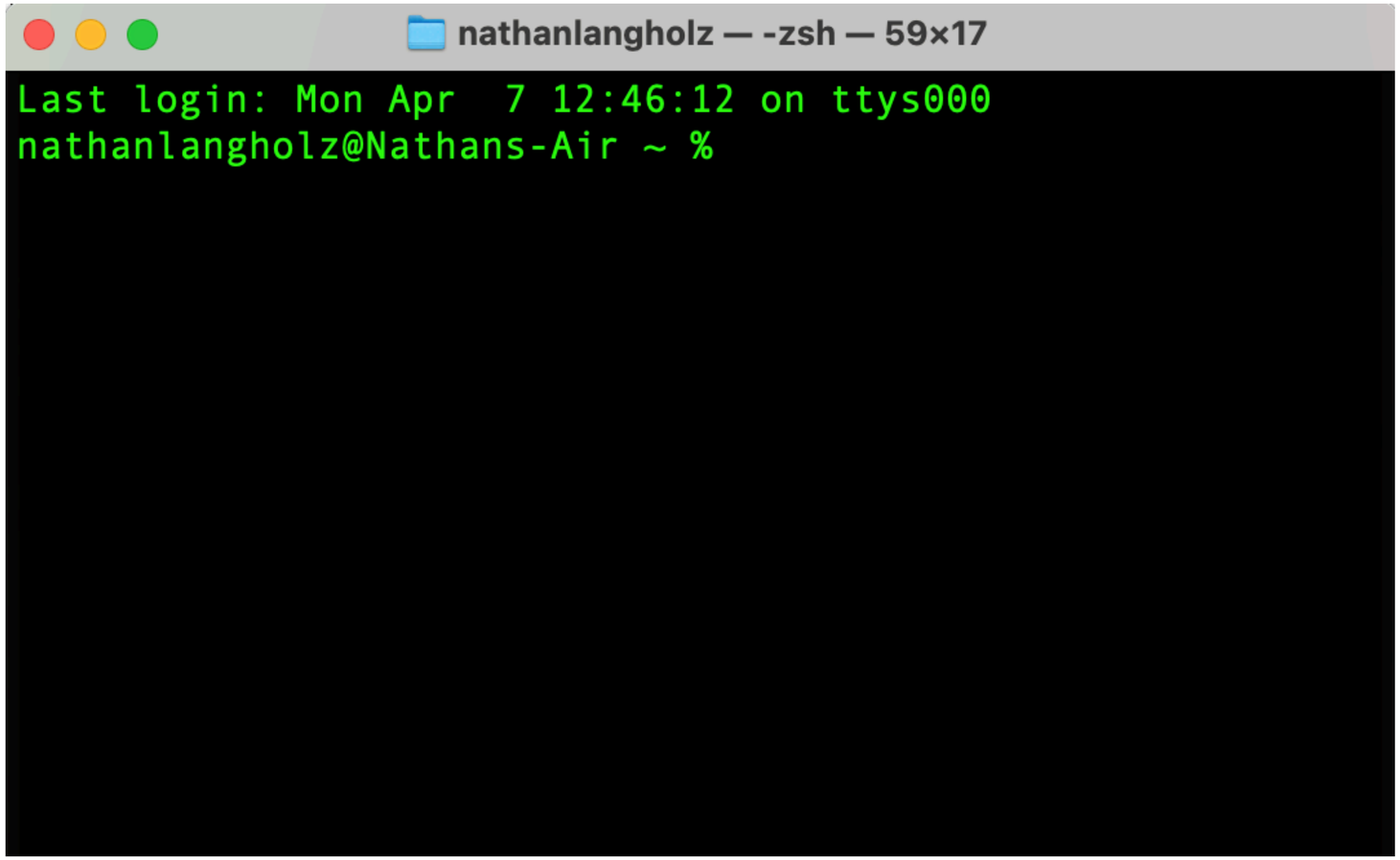
the Unix shell is a user interface to the kernel (keep in mind that Unix was designed for computer scientists and the interface is not optimized for novices)

Unix shells

A shell is a type of program called an interpreter, think of it as a text-based interface to the kernel

It operates in simple loop: it accepts a command, interprets it, executes the command and waits for another

The shell displays a prompt to tell you that it is ready to accept a command



Not much going on but lots of potential!

The % is the prompt

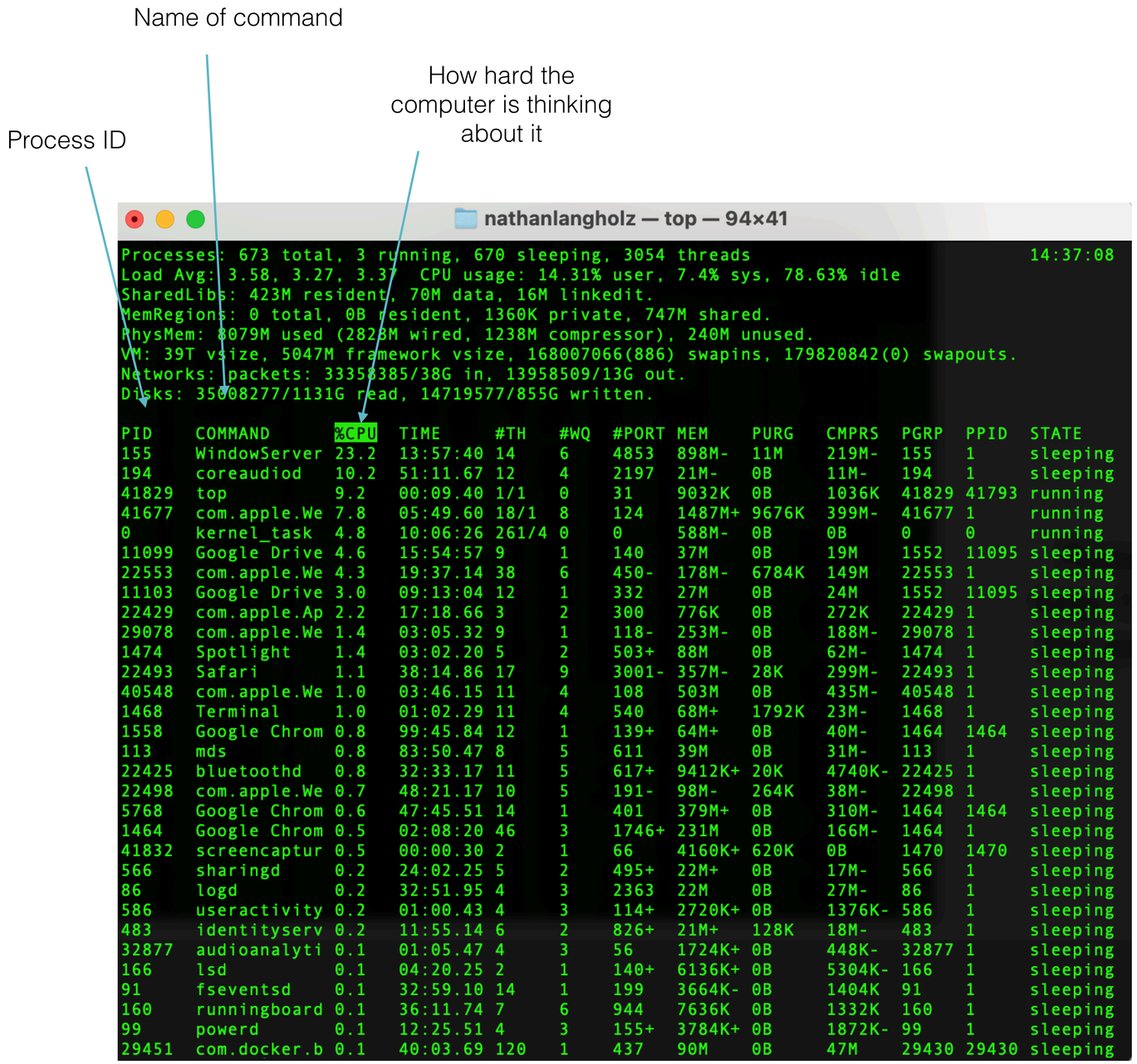
Unix shells

The shell is itself a program the the Unix operating system runs for you (a program is referred to as a process when its running)

The kernel manages many processes at once, many of which are the result of user commands (others provide services that keep the computer running)

Some commands are built into the shell, others have been added by users

Either way, the shell waits until the command is executed



Result of the command **`top`**; this is a printout of all the processes running on your computer

Operating Systems

Process Management

Schedules jobs(formally referred to as processes) to be executed by the computer

Memory and storage management

Allocate space required for each running process in main memory (RAM) or in some other temporary location if space is tight and supervise the storage of data onto disk

Operating Systems

Device Management

A program called a driver translates data (files from the filesystem) into signals that

Application Programming Interface

An API (application programming interface) let programmers use functions of the computer and the operating system without having to know *how* something is done

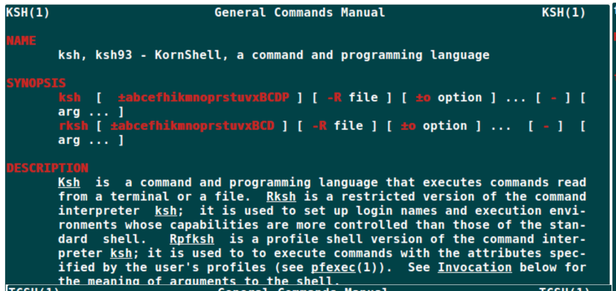
User Interface

Finally, the operating system turns and looks at you; the UI is a program that defines how users interact with the computer; some are graphical (Windows is a GUI) and some are text-based (your Unix shell)

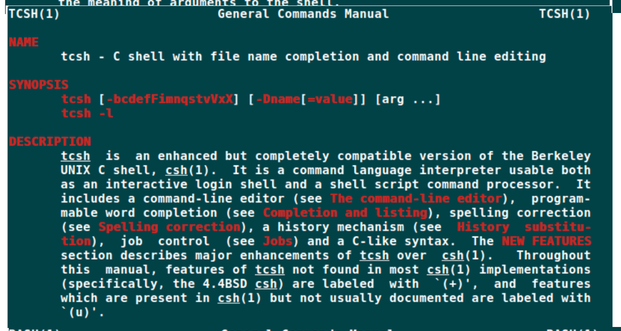
Unix Shell(s)

There are, in fact, many different kinds of Unix shells

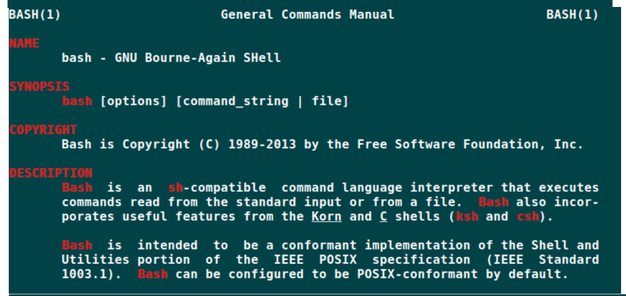
The table on the right lists a few of the most popular



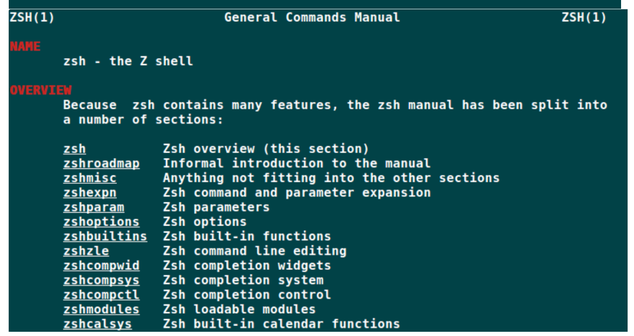
```
KSH(1)                    General Commands Manual                    KSH(1)

NAME
       ksh, ksh93 - KornShell, a command and programming language

SYNOPSIS
       ksh  [  ±abcefhikmnoprstuvxBCDP ] [ -R file ] [ ±o option ] ... [ - ] [
       arg ... ]
       rksh [ ±abcefhikmnoprstuvxBCD ] [ -R file ] [ ±o option ] ...  [ - ]  [
       arg ... ]

DESCRIPTION
       Ksh  is  a command and programming language that executes commands read
       from a terminal or a file.  Rksh is a restricted version of the command
       interpreter  ksh;  it is used to set up login names and execution envi-
       ronments whose capabilities are more controlled than those of the stan-
       dard  shell.   Rpfksh  is a profile shell version of the command inter-
       preter ksh; it is used to to execute commands with the attributes spec-
       ified by the user's profiles (see pfexec(1)).  See Invocation below for
       the meaning of arguments to the shell.
```



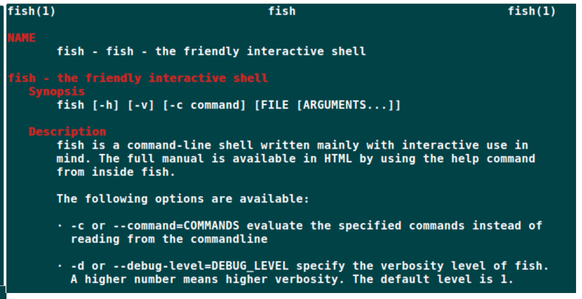
```
TCSH(1)                   General Commands Manual                    TCSH(1)

NAME
       tcsh - C shell with file name completion and command line editing

SYNOPSIS
       tcsh [-bcdefFimnqstvVxX] [-Dname[=value]] [arg ...]
       tcsh -l

DESCRIPTION
       tcsh  is  an enhanced but completely compatible version of the Berkeley
       UNIX C shell, csh(1).  It is a command language interpreter usable both
       as an interactive login shell and a shell script command processor.  It
       includes a command-line editor (see The command-line editor),  program-
       mable word completion (see Completion and listing), spelling correction
       (see Spelling correction), a history mechanism (see  History  substitu-
       tion),  job  control  (see Jobs) and a C-like syntax.  The NEW FEATURES
       section describes major enhancements of tcsh over  csh(1).   Throughout
       this  manual, features of tcsh not found in most csh(1) implementations
       (specifically, the 4.4BSD csh) are labeled  with  `(+)',  and  features
       which are present in csh(1) but not usually documented are labeled with
       `(u)'.
```



```
BASH(1)                   General Commands Manual                    BASH(1)

NAME
       bash - GNU Bourne-Again SHell

SYNOPSIS
       bash [options] [command_string | file]

COPYRIGHT
       Bash is Copyright (C) 1989-2013 by the Free Software Foundation, Inc.

DESCRIPTION
       Bash  is  an  sh-compatible  command language interpreter that executes
       commands read from the standard input or from a file.  Bash also incor-
       porates useful features from the Korn and C shells (ksh and csh).

       Bash  is  intended  to  be a conformant implementation of the Shell and
       Utilities portion  of  the  IEEE  POSIX  specification  (IEEE  Standard
       1003.1).  Bash can be configured to be POSIX-conformant by default.
```



```
ZSH(1)                    General Commands Manual                    ZSH(1)

NAME
       zsh - the Z shell

OVERVIEW
       Because  zsh contains many features, the zsh manual has been split into
       a number of sections:

       zsh           Zsh overview (this section)
       zshroadmap    Informal introduction to the manual
       zshmisc       Anything not fitting into the other sections
       zshexpn       Zsh command and parameter expansion
       zshparam      Zsh parameters
       zshoptions    Zsh options
       zshbuiltins   Zsh built-in functions
       zshzle        Zsh command line editing
       zshcompwid    Zsh completion widgets
       zshcompsys    Zsh completion system
       zshcompctl    Zsh completion control
       zshmodules    Zsh loadable modules
       zshcalsys     Zsh built-in calendar functions
```



```
fish(1)                          fish                          fish(1)

NAME
       fish - fish - the friendly interactive shell

fish - the friendly interactive shell
  Synopsis
       fish [-h] [-v] [-c command] [FILE [ARGUMENTS...]]

  Description
       fish is a command-line shell written mainly with interactive use in
       mind. The full manual is available in HTML by using the help command
       from inside fish.

       The following options are available:

       · -c or --command=COMMANDS evaluate the specified commands instead of
         reading from the commandline

       · -d or --debug-level=DEBUG_LEVEL specify the verbosity level of fish.
         A higher number means higher verbosity. The default level is 1.
```

Why the choices?

A shell program was originally meant to take commands, interpret them and then execute some operation

Inevitably, one wants to collect a number of these operations into programs that execute compound tasks at the the same time you want to make interaction on the command line as easy as possible (a history mechanism, editing capabilities and so on)

The original Bourne shell is ideal for programming; the C-shell and its variants are good for interactive use; the Korn shell is a combination of both

Steve Bourne, created of sh, 2005 (source wikipedia)

And while were at it

unix itself comes in different flavors; the 1980s saw an incredible proliferation of Unix versions, somewhere around 100 (System V, AIX, Berkely BSD, SunOS, Linux, …)

vendors provided (diverging) version of Unix, optimized for their own computer architectures and supporting different features

despite the diversity it was still easier to “port” applications between versions of Unix than it was between different proprietary OS

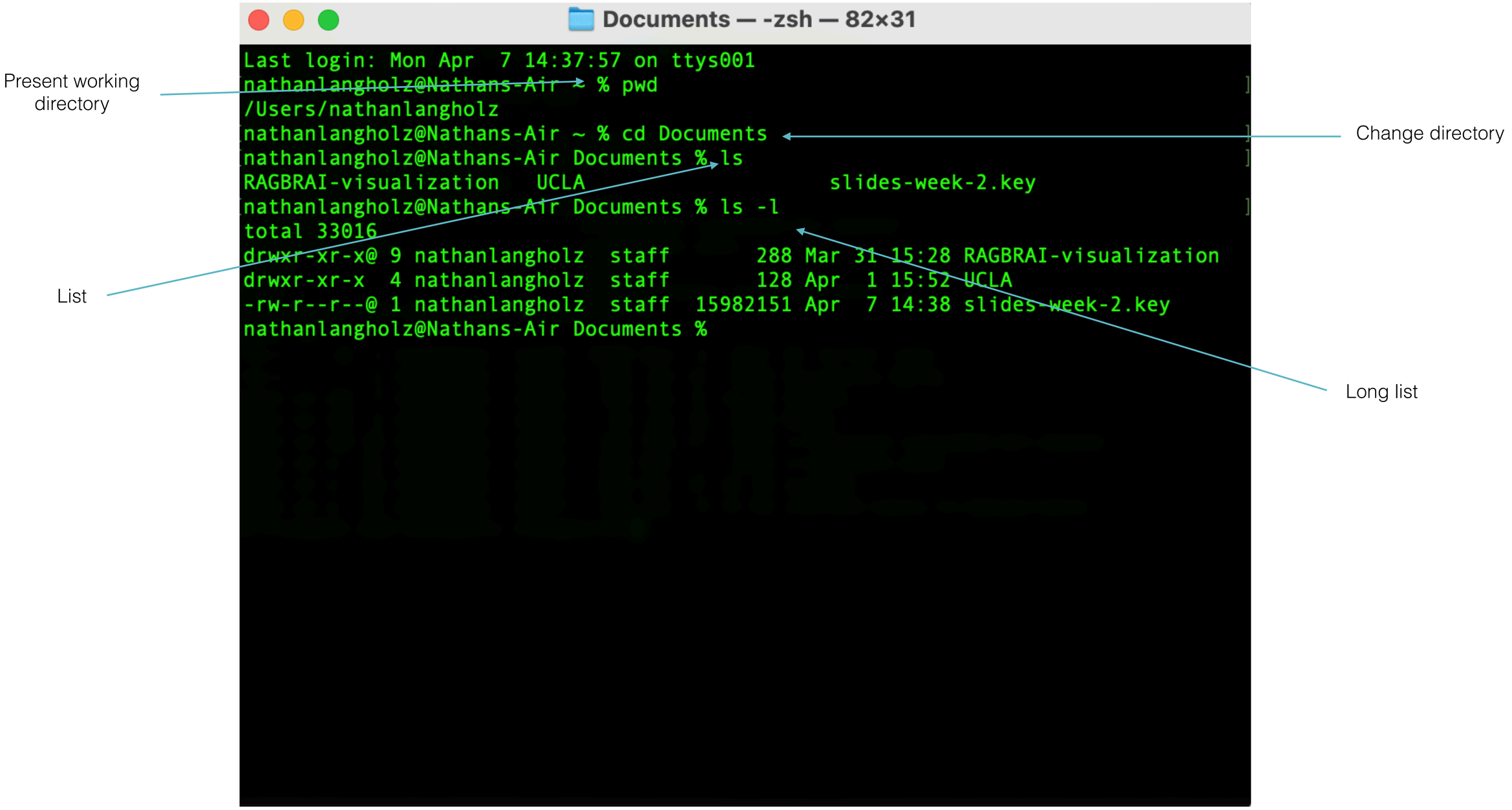
A few common commands

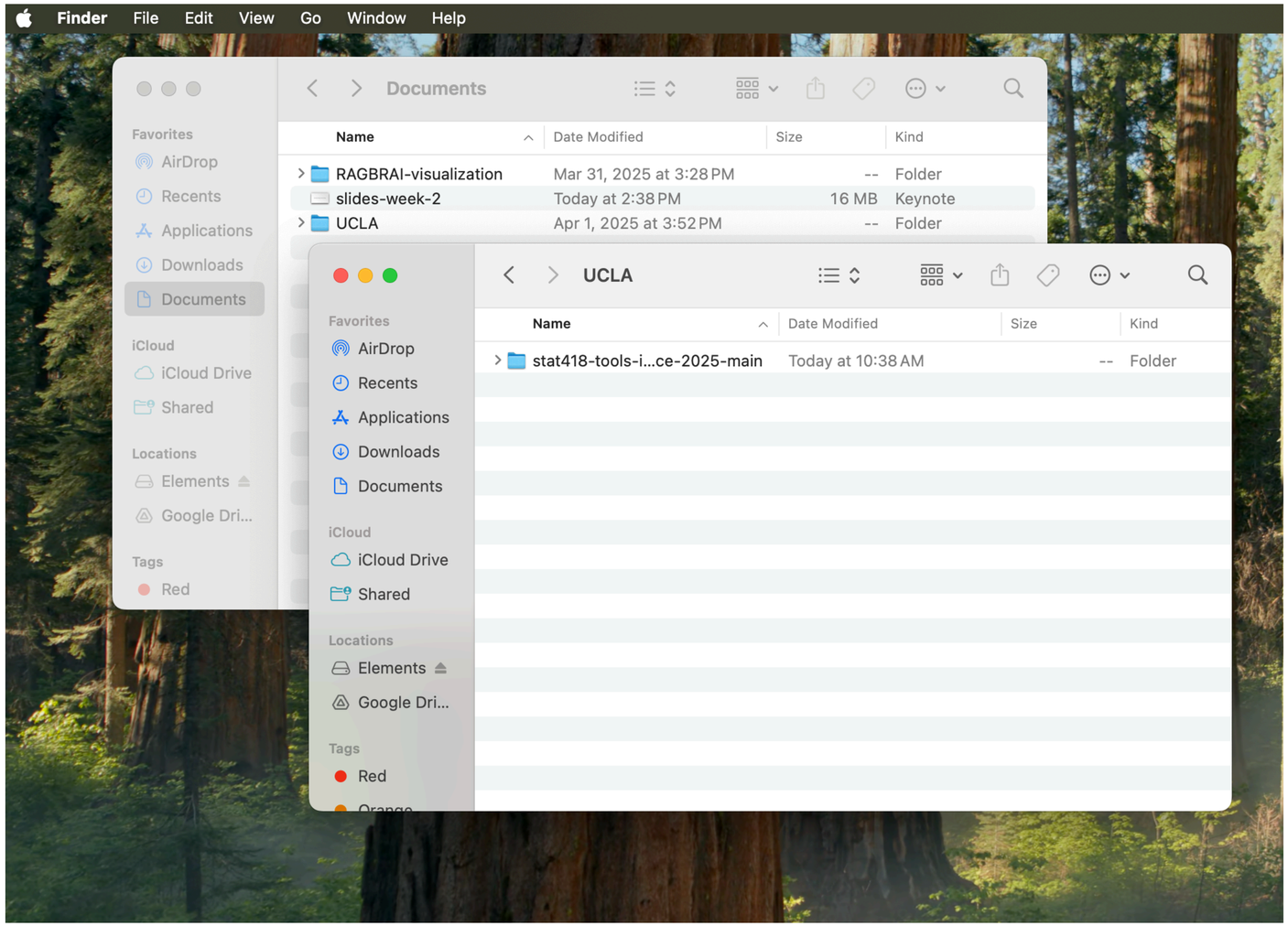
FIrst, commands to explore your file system, walk through directories and list flies

`pwd, ls, cd`

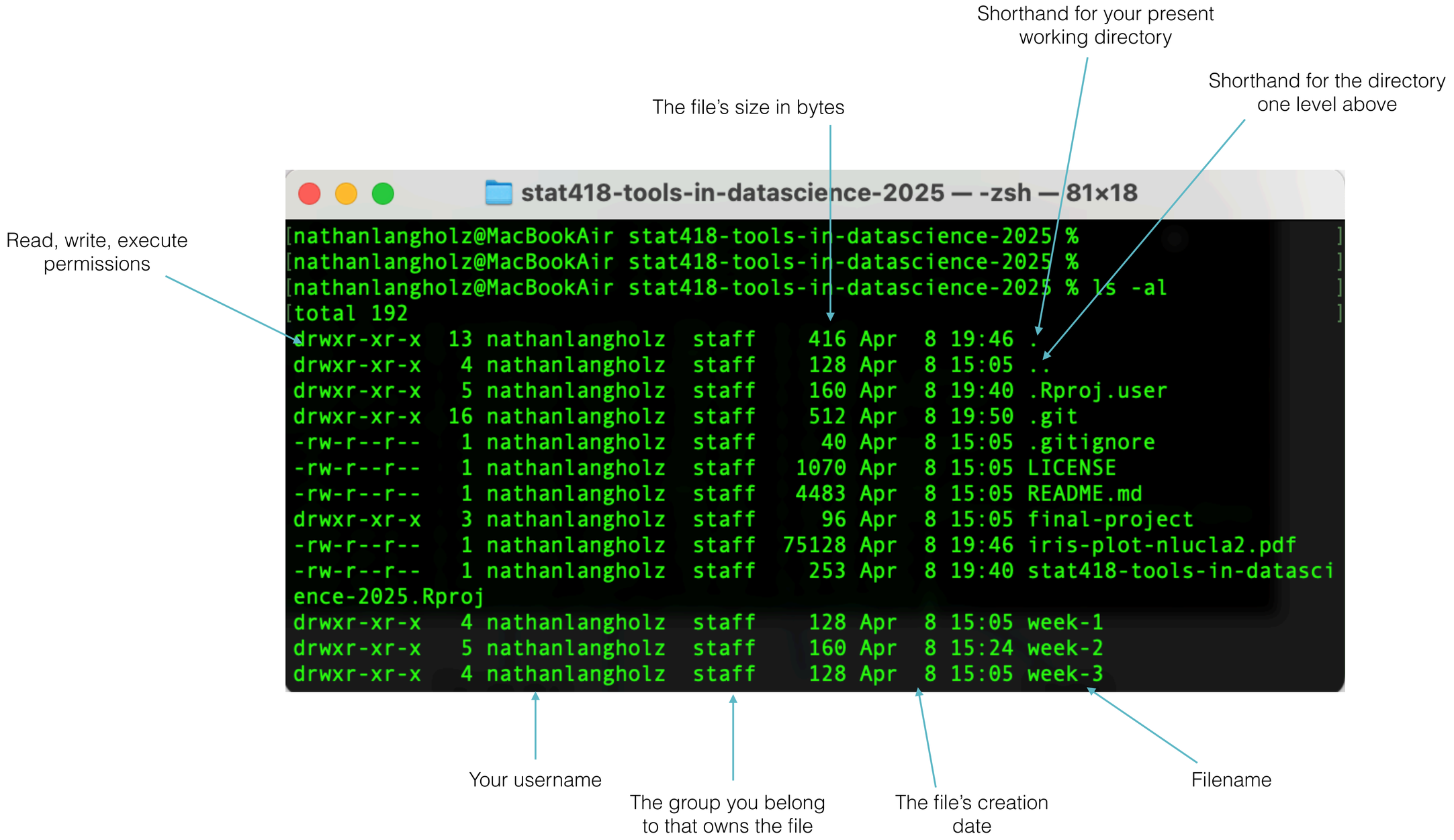
`mkdir, rmdir`

`cp, mv, rm`


Present working directory
List
Change directory
Long list
Documents — -zsh — 82×31
Last login: Mon Apr 7 14:37:57 on ttys001
nathanlangholz@Nathans-Air ~ % pwd
/Users/nathanlangholz
nathanlangholz@Nathans-Air ~ % cd Documents
nathanlangholz@Nathans-Air Documents % ls
RAGBRAI-visualization UCLA slides-week-2.key
nathanlangholz@Nathans-Air Documents % ls -l
total 33016
drwxr-xr-x@ 9 nathanlangholz staff 288 Mar 31 15:28 RAGBRAI-visualization
drwxr-xr-x 4 nathanlangholz staff 128 Apr 1 15:52 UCLA
-rw-r--r--@ 1 nathanlangholz staff 15982151 Apr 7 14:38 slides-week-2.key
nathanlangholz@Nathans-Air Documents %



Another view of the filesystem; here your Mac will display directories as folders and of course you navigate by clicking rather than typing commands


Shorthand for your present working directory
Shorthand for the directory one level above
The file's size in bytes
Read, write, execute permissions
stat418-tools-in-datascience-2025 — -zsh — 81×18
[nathanlangholz@MacBookAir stat418-tools-in-datascience-2025 %
[nathanlangholz@MacBookAir stat418-tools-in-datascience-2025 %
[nathanlangholz@MacBookAir stat418-tools-in-datascience-2025 % ls -al
[total 192
drwxr-xr-x 13 nathanlangholz staff   416 Apr 8 19:46 .
drwxr-xr-x   4 nathanlangholz staff   128 Apr 8 15:05 ..
drwxr-xr-x   5 nathanlangholz staff   160 Apr 8 19:40 .Rproj.user
drwxr-xr-x 16 nathanlangholz staff   512 Apr 8 19:50 .git
-rw-r--r--   1 nathanlangholz staff   40 Apr 8 15:05 .gitignore
-rw-r--r--   1 nathanlangholz staff   1070 Apr 8 15:05 LICENSE
-rw-r--r--   1 nathanlangholz staff   4483 Apr 8 15:05 README.md
drwxr-xr-x   3 nathanlangholz staff   96 Apr 8 15:05 final-project
-rw-r--r--   1 nathanlangholz staff 75128 Apr 8 19:46 iris-plot-nlucla2.pdf
-rw-r--r--   1 nathanlangholz staff   253 Apr 8 19:40 stat418-tools-in-datasci
ence-2025.Rproj
drwxr-xr-x   4 nathanlangholz staff   128 Apr 8 15:05 week-1
drwxr-xr-x   5 nathanlangholz staff   160 Apr 8 15:24 week-2
drwxr-xr-x   4 nathanlangholz staff   128 Apr 8 15:05 week-3
Your username
The group you belong to that owns the file
The file's creation date
Filename

Kinds of Files

What you'll notice right away is that there are different types fo files having different permissions

Unix filesystem conventions places (shared, commonly used) executable files in places like /usr/bin or /usr/local/bin

Different files are opened by different kinds of programs, in OSX, there is a command called open that decides which program to use

Permission bits

Unix can support many users on a single system and each user can belong to one or more groups

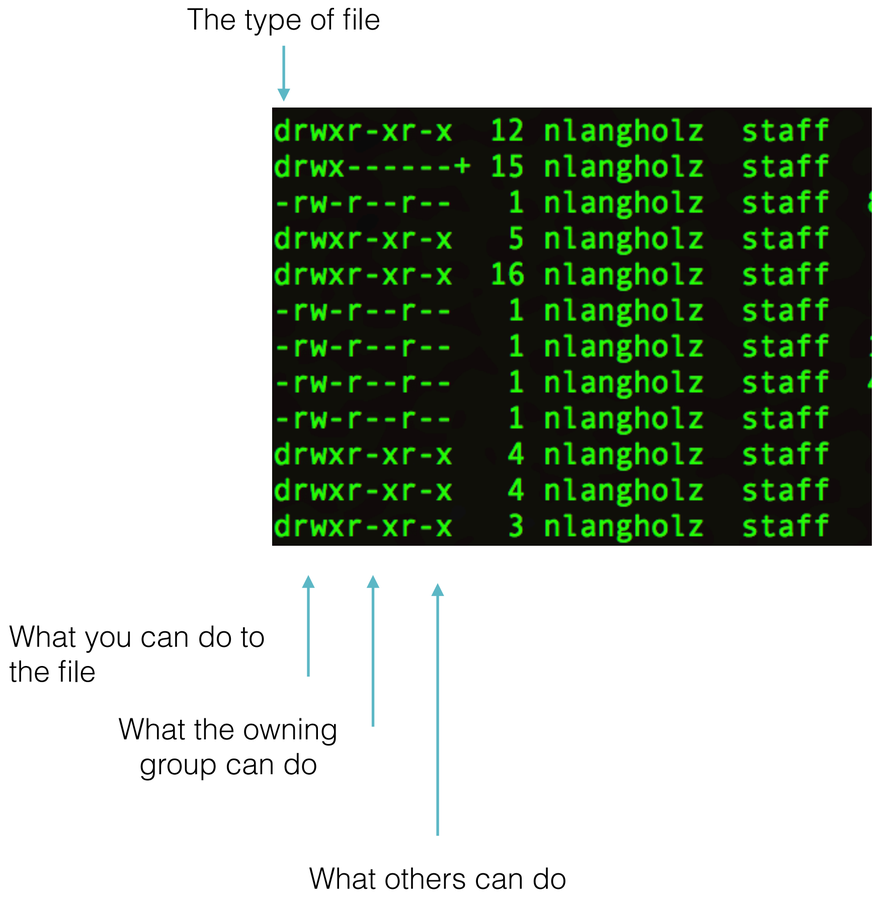
every file in a Unix filesystem is owned by some user and one of that user's groups; each file also has a set of permissions specifying which users can

`r: read` or **`w: write (modify)`** or **`x: execute`**

the file; these are specified with three 'bits' and we need three sets of bits to define what the user can do, what their group (that owns the file) can do and what others can do

the command **`chmod`** changes the permissions on a file but we will leave that for you to discover on your own

Permission bits



Although we saw a terminal environment in our Rstudio docker container, now let's get another docker image that has some nice data science specific code.

we will use the docker image from the book Data Science at the Command Line by Jeroen Janssens

https://www.datascienceatthecommandline.com/index.html

First, we need to pull the image using the following

```
docker pull jeroenjanssens/data-science-at-the-command-line
```

Data Science at the Command Line

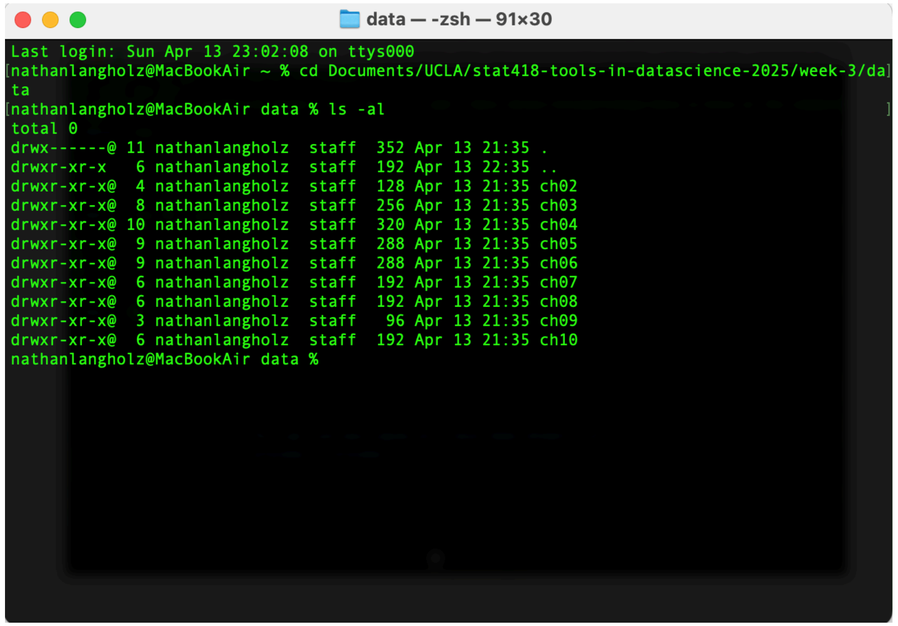
To run the image

```
docker run --rm -it jeroenjanssens/data-science-at-the-command-line
```

To leave the environment simply type **`exit`**

again we didn't mount a local volume so we have no connection to our computer

Before mounting to local directory go to the directory where you have downloaded the data from the book site.



Download and unzip from here
https://www.datascienceatthecommandline.com/2e/data.zip

Now to mount our local directories use

on a Mac

```
docker run --rm -it -v `pwd`:/data jeroenjanssens/data-science-at-the-command-line
```

Or

```
docker run --rm -it -v "$(pwd)":/data jeroenjanssens/data-science-at-the-command-line
```

in Windows command line

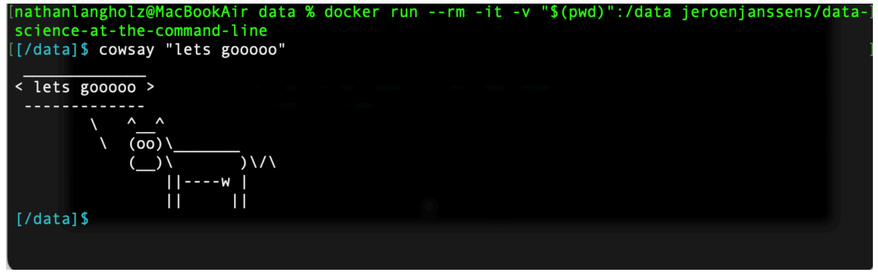
```
C:\> docker run --rm -it -v "%cd%":/data jeroenjanssens/data-science-at-the-command-line
```

in Windows powershell

```
PS C:\> docker run --rm -it -v ${PWD}:/data jeroenjanssens/data-science-at-the-command-line
```

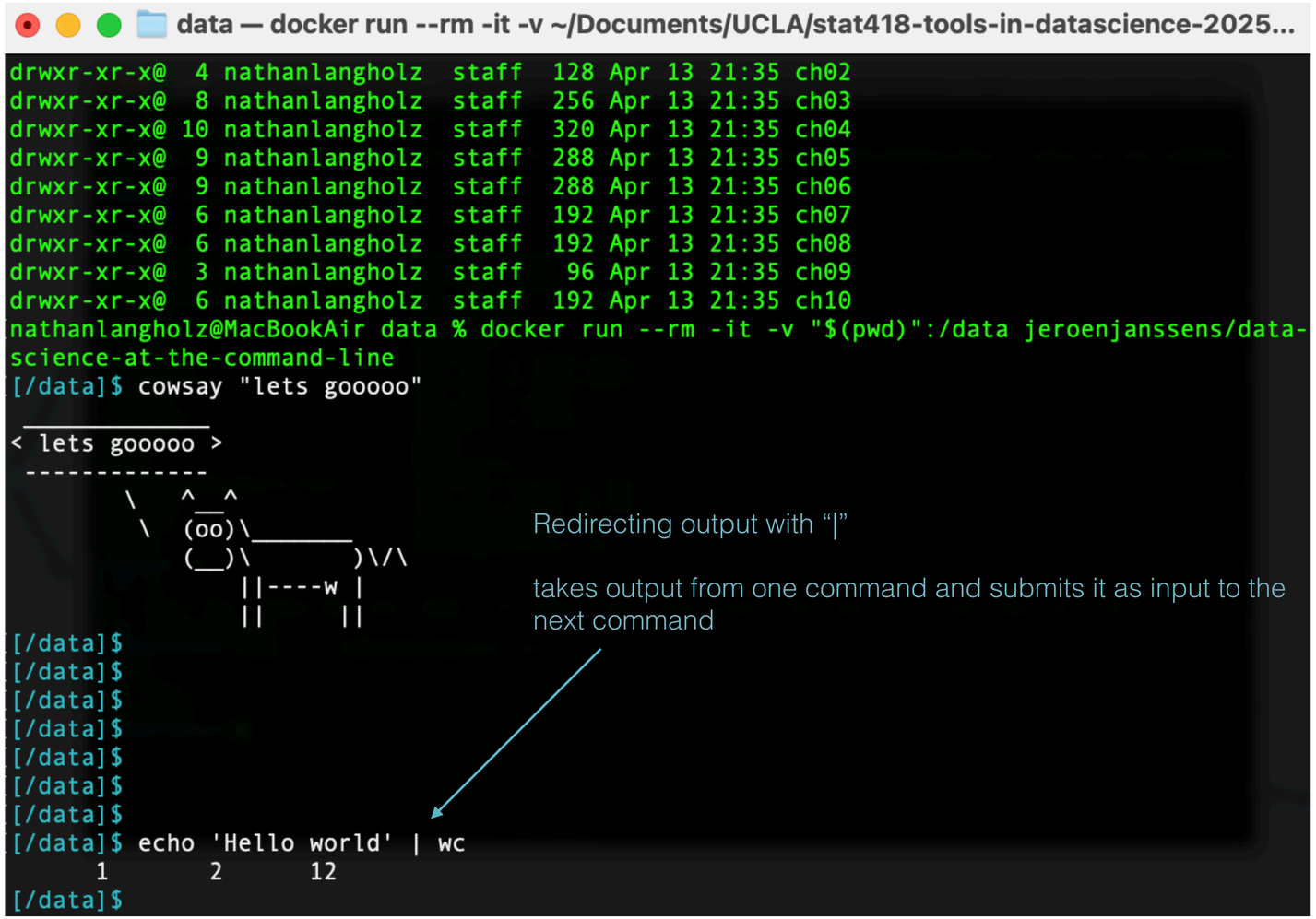
To make sure everything is running correctly, type

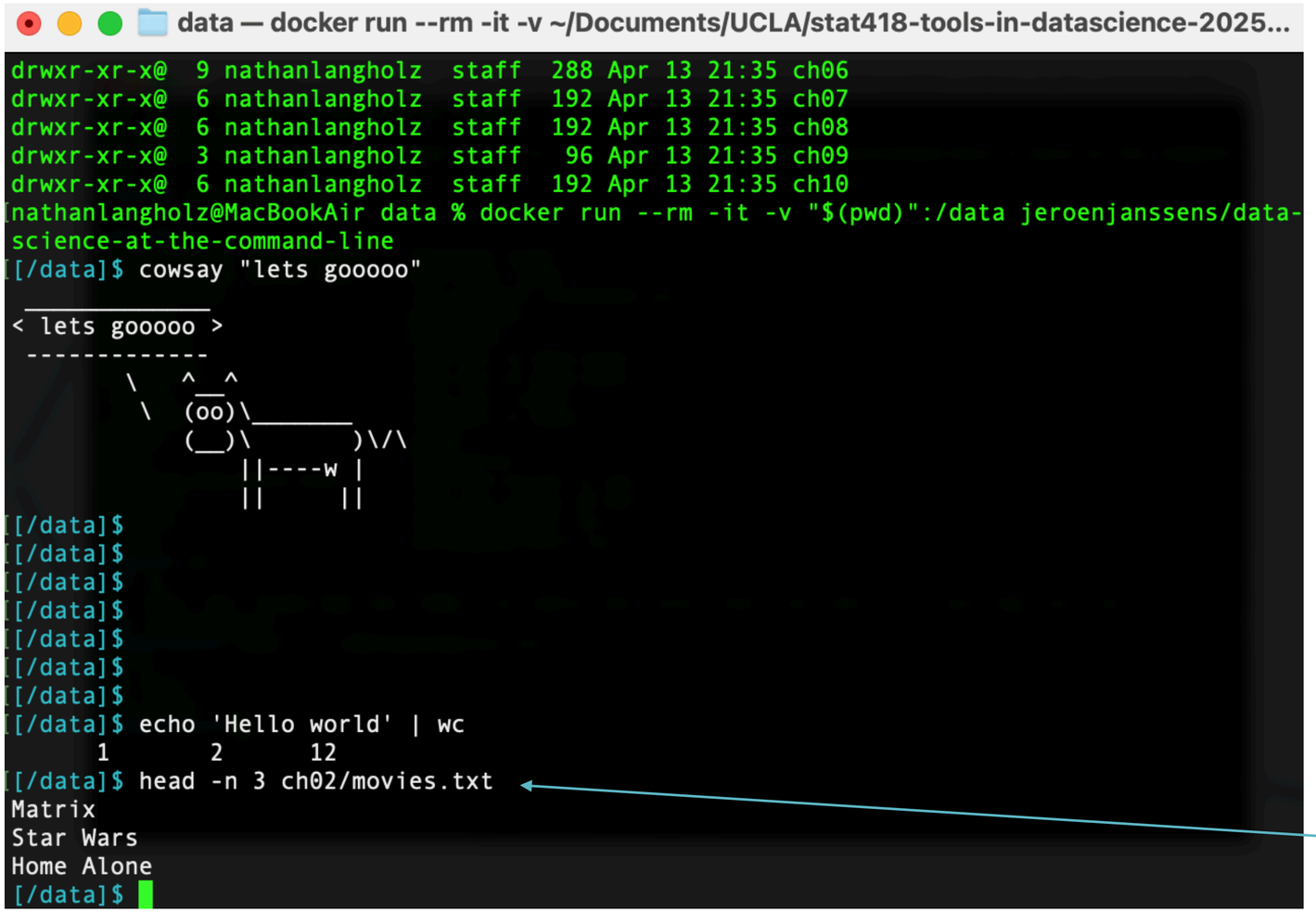
```
cowsay “ready to go!”
```



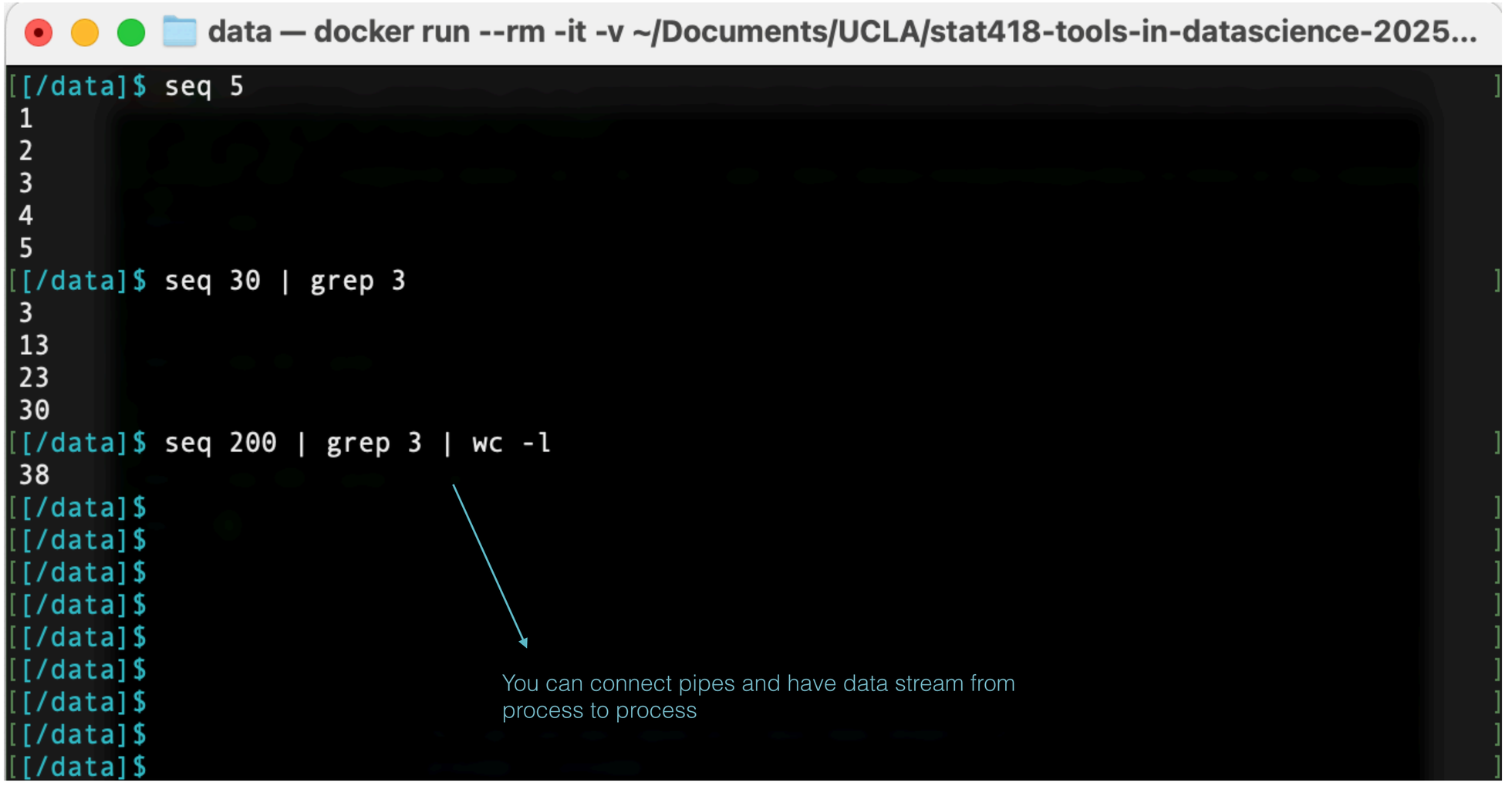
This environment we now have is a bash shell running on a Linux operating system.

Note: this only works in this environment not in all terminal windows


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025...
drwxr-xr-x@ 4 nathanlangholz staff 128 Apr 13 21:35 ch02
drwxr-xr-x@ 8 nathanlangholz staff 256 Apr 13 21:35 ch03
drwxr-xr-x@ 10 nathanlangholz staff 320 Apr 13 21:35 ch04
drwxr-xr-x@ 9 nathanlangholz staff 288 Apr 13 21:35 ch05
drwxr-xr-x@ 9 nathanlangholz staff 288 Apr 13 21:35 ch06
drwxr-xr-x@ 6 nathanlangholz staff 192 Apr 13 21:35 ch07
drwxr-xr-x@ 6 nathanlangholz staff 192 Apr 13 21:35 ch08
drwxr-xr-x@ 3 nathanlangholz staff 96 Apr 13 21:35 ch09
drwxr-xr-x@ 6 nathanlangholz staff 192 Apr 13 21:35 ch10
nathanlangholz@MacBookAir data % docker run --rm -it -v "$(pwd)":/data jeroenjanssens/data-science-at-the-command-line
[/data]$ cowsay "lets gooooo"
< lets gooooo >
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$ echo 'Hello world' | wc
1 2 12
[/data]$
Redirecting output with "|"
takes output from one command and submits it as input to the next command



Listing the first 3 movies from the file movies.txt


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025...
[[/data]$ seq 5
1
2
3
4
5
[[/data]$ seq 30 | grep 3
3
13
23
30
[[/data]$ seq 200 | grep 3 | wc -l
38
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
[[/data]$
You can connect pipes and have data stream from process to process

“Grep was invented for me. I was making a program to read text aloud through a voice synthesizer. As I invented phonetic rules I would check Webster's dictionary for words on which they might fail. For example, how do you cope with the digraph `ui', which is pronounced many different ways: `fruit', `guile', `guilty', `anguish', `intuit', `beguine'? I would break the dictionary up into pieces that fit in ed's limited buffer and use a global command to select a list. I would whittle this list down by repeated scannings with ed to see how each proposed rule worked.”

The process was tedious, and terribly wasteful, since the dictionary had to be split (one couldn't afford to leave a split copy on line). Then ed copied each part into /tmp, scanned it twice to accomplish the g command, and finally threw it away, which takes time too.”

- Doug McIlroy, Adjunct Professor of Computer Science Dartmouth college

Chapter 9, On the Early History and Impact of Unix Tools to Build the Tools for a New Millennium
http://www.columbia.edu/~rh120/ch001j.c11

“One afternoon I asked Ken Thompson if he could lift the regular expression recognizer out of the editor and make a one-pass program to do it. He said yes. The next morning I found a note in my mail announcing a program named grep. It worked like a charm. When asked what that funny name meant, Ken said it was obvious. It stood for the editor command that it simulated, g/re/p (global regular expression print).”

- Doug McIlroy, Adjunct Professor of Computer Science Dartmouth college

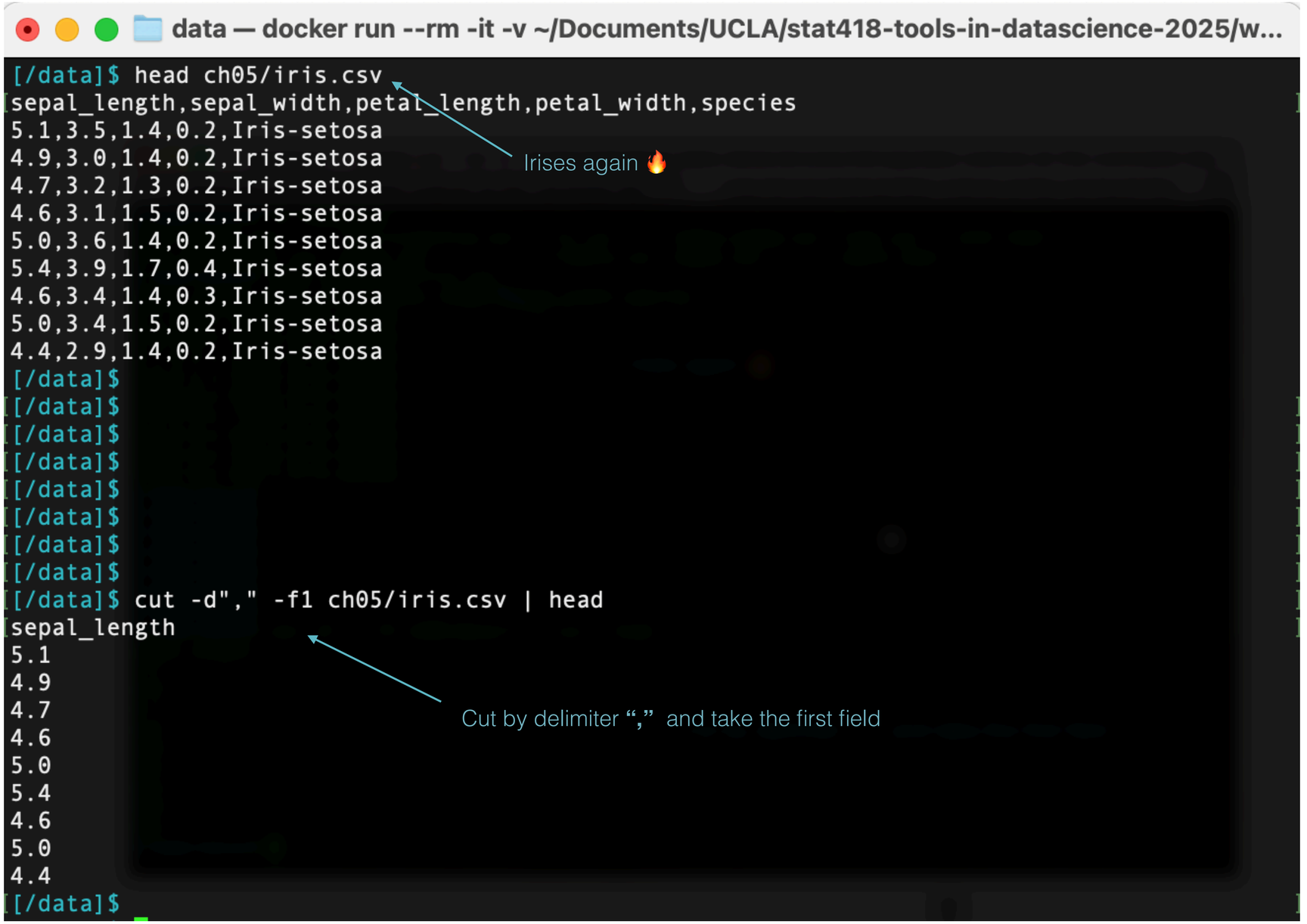
Chapter 9, On the Early History and Impact of Unix Tools to Build the Tools for a New Millennium
http://www.columbia.edu/~rh120/ch001j.c11

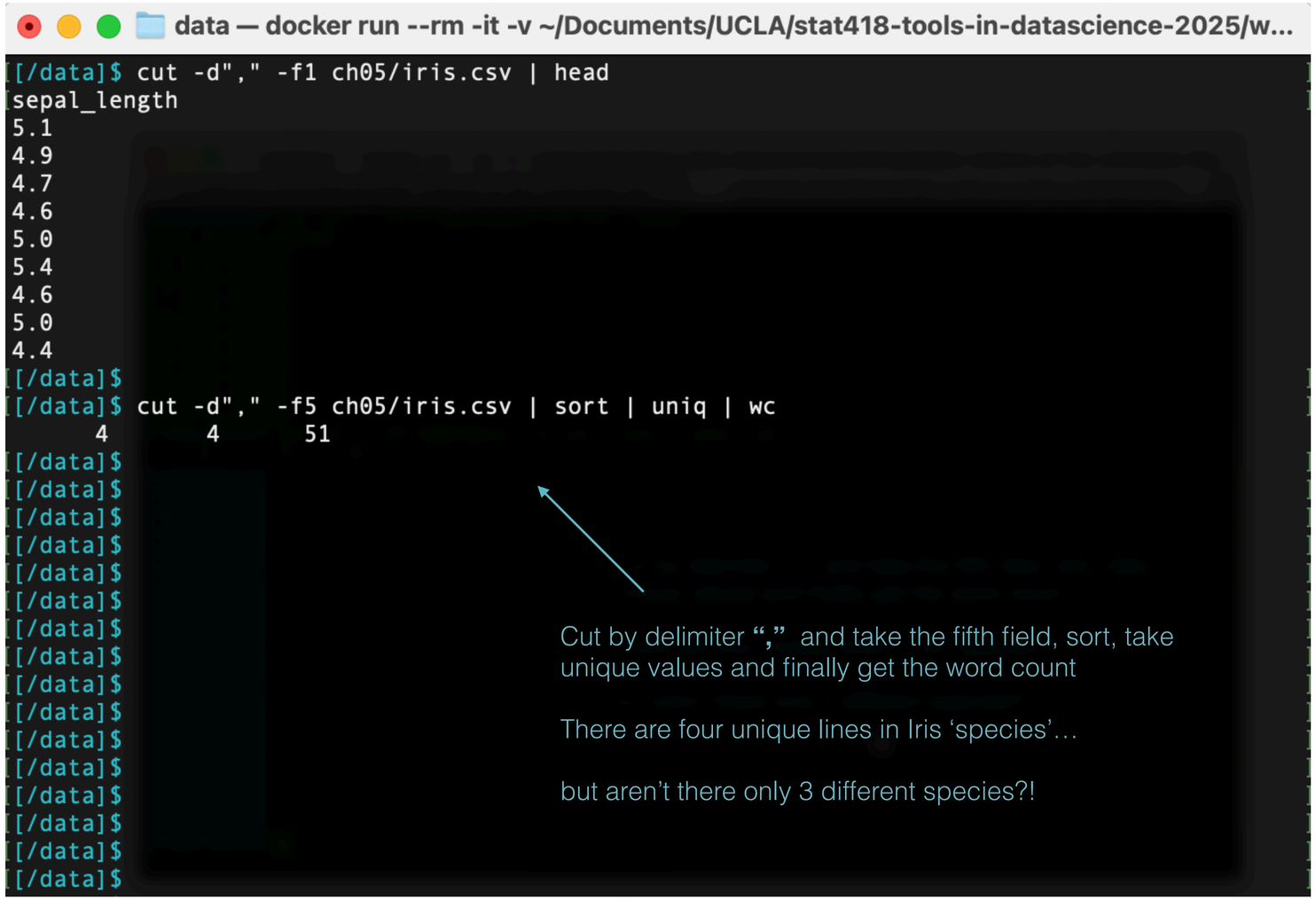
Documents — docker run --rm -it -v ~/Documents:/data jeroenjanssens/data-science-at-the-command-line — 132×41

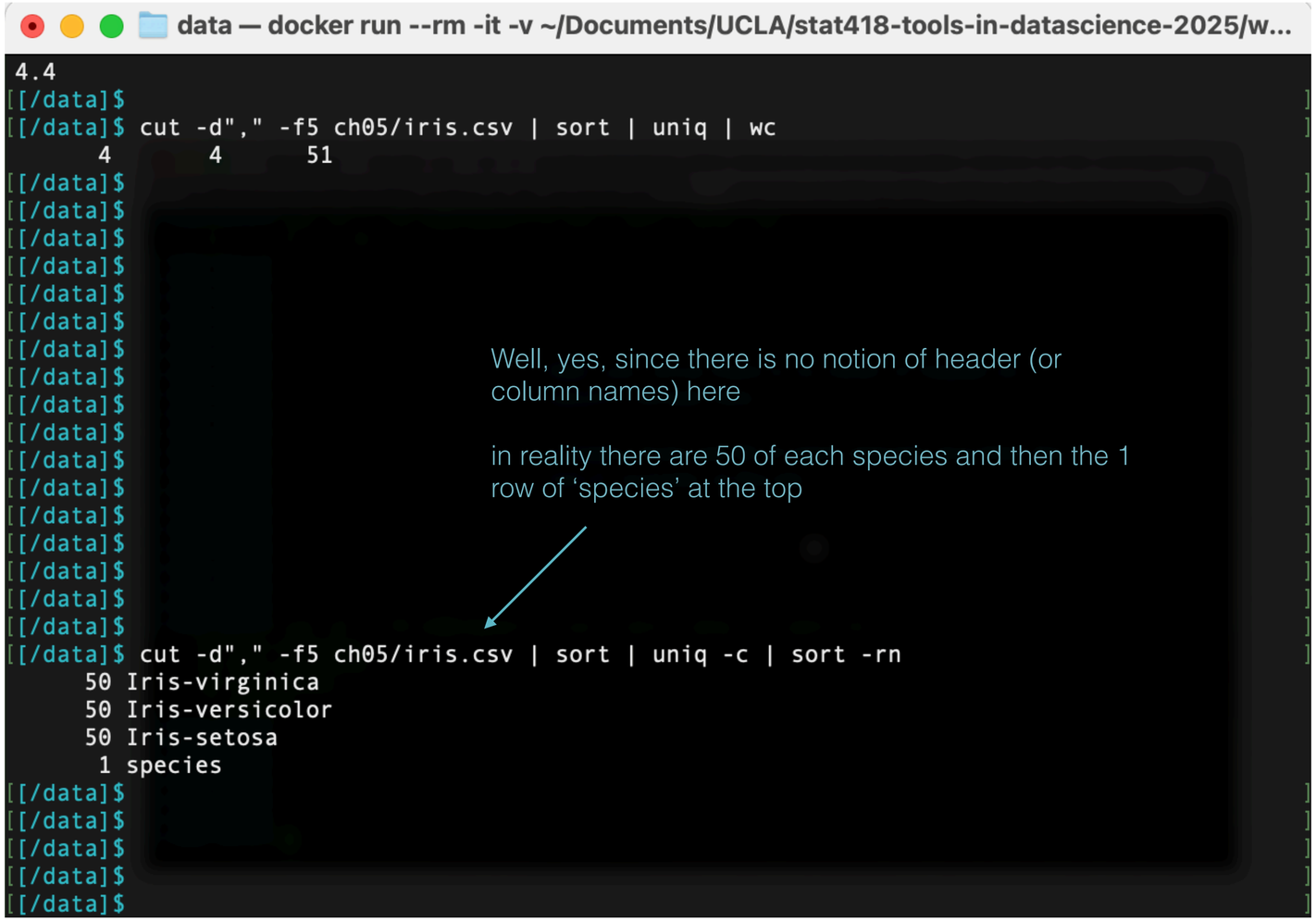
```
nathanlangholz@MacBookAir Documents % docker run --rm -it -v "$(pwd)":/data jeroenjanssens/data-science-at-the-command-line
[/data]$ ls -al
total 1884
drwx------ 7 root root     224 Apr 14 04:37 .
drwxr-xr-x 1 root root    4096 Apr 14 22:55 ..
-rw-r--r-- 1 root root   10244 Apr 14 21:22 .DS_Store
-rw-r--r-- 1 root root       0 Jul 27  2023 .localized
-rw-r--r-- 1 root root 1909170 Apr 12 23:49 R-project-github-sync.pdf
drwxr-xr-x 9 root root     288 Mar 31 22:28 RAGBRAI-visualization
drwxr-xr-x 4 root root     128 Apr 14 05:35 UCLA
[/data]$ cd UCLA/stat418-tools-in-datascience-2025/week-3
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$ ls
R  README.md  data  sh-scraper-file  slides-week-3.pdf
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$ wc data/ch03/finn.txt
 12346 114126 622513 data/ch03/finn.txt
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$ head data/ch03/finn.txt
The Project Gutenberg eBook of Adventures of Huckleberry Finn

This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at www.gutenberg.org. If you are not located in the United States,
you will have to check the laws of the country where you are located
before using this eBook.

[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$ tail data/ch03/finn.txt

Most people start at our website which has the main PG search
facility: www.gutenberg.org.

This website includes information about Project Gutenberg™,
including how to make donations to the Project Gutenberg Literary
Archive Foundation, how to help produce our new eBooks, and how to
subscribe to our email newsletter to hear about new eBooks.


[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$
[/data/UCLA/stat418-tools-in-datascience-2025/week-3]$
```

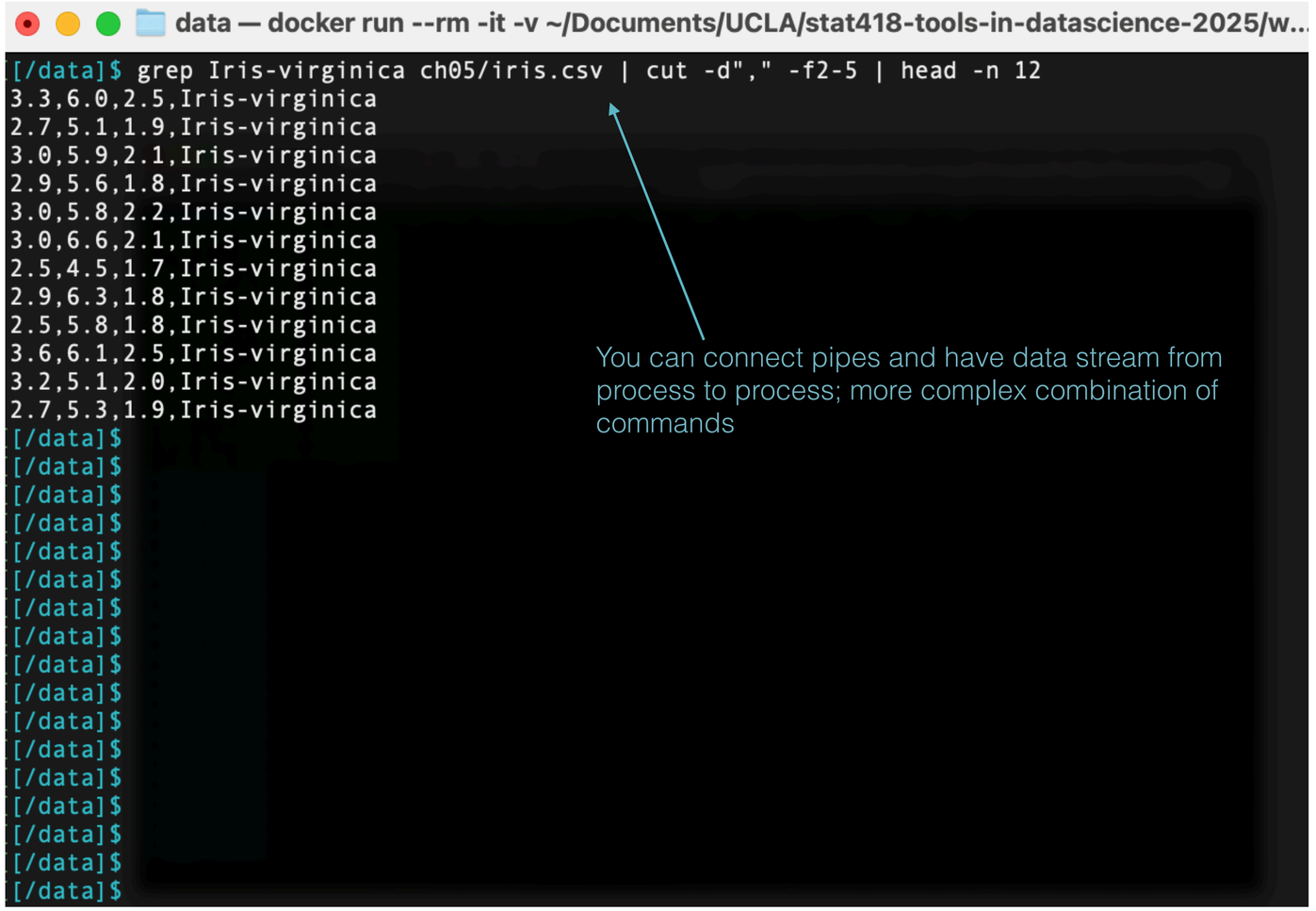
Similar to head we have ‘tail’ end of the document

Finn.txt doesn’t exist in the data folder, this will be added in slide 49, but can try on any of the other text files in the directory


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025/w...
[/data]$ head ch05/iris.csv
sepal_length,sepal_width,petal_length,petal_width,species
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$ cut -d"," -f1 ch05/iris.csv | head
sepal_length
5.1
4.9
4.7
4.6
5.0
5.4
4.6
5.0
4.4
[/data]$
Irises again 🔥
Cut by delimiter “,” and take the first field


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025/w...
[/data]$ cut -d"," -f1 ch05/iris.csv | head
sepal_length
5.1
4.9
4.7
4.6
5.0
5.4
4.6
5.0
4.4
[/data]$
[/data]$ cut -d"," -f5 ch05/iris.csv | sort | uniq | wc
4 4 51
[/data]$
Cut by delimiter “,” and take the fifth field, sort, take unique values and finally get the word count
There are four unique lines in Iris ‘species’…
but aren’t there only 3 different species?!


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025/w...
4.4
[/data]$
[/data]$ cut -d"," -f5 ch05/iris.csv | sort | uniq | wc
4 4 51
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$ cut -d"," -f5 ch05/iris.csv | sort | uniq -c | sort -rn
50 Iris-virginica
50 Iris-versicolor
50 Iris-setosa
1 species
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
Well, yes, since there is no notion of header (or column names) here
in reality there are 50 of each species and then the 1 row of ‘species’ at the top


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025/w...
[/data]$ grep Iris-virginica ch05/iris.csv | cut -d"," -f2-5 | head -n 12
3.3,6.0,2.5,Iris-virginica
2.7,5.1,1.9,Iris-virginica
3.0,5.9,2.1,Iris-virginica
2.9,5.6,1.8,Iris-virginica
3.0,5.8,2.2,Iris-virginica
3.0,6.6,2.1,Iris-virginica
2.5,4.5,1.7,Iris-virginica
2.9,6.3,1.8,Iris-virginica
2.5,5.8,1.8,Iris-virginica
3.6,6.1,2.5,Iris-virginica
3.2,5.1,2.0,Iris-virginica
2.7,5.3,1.9,Iris-virginica
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
You can connect pipes and have data stream from process to process; more complex combination of commands

Sending output to a file with “>”

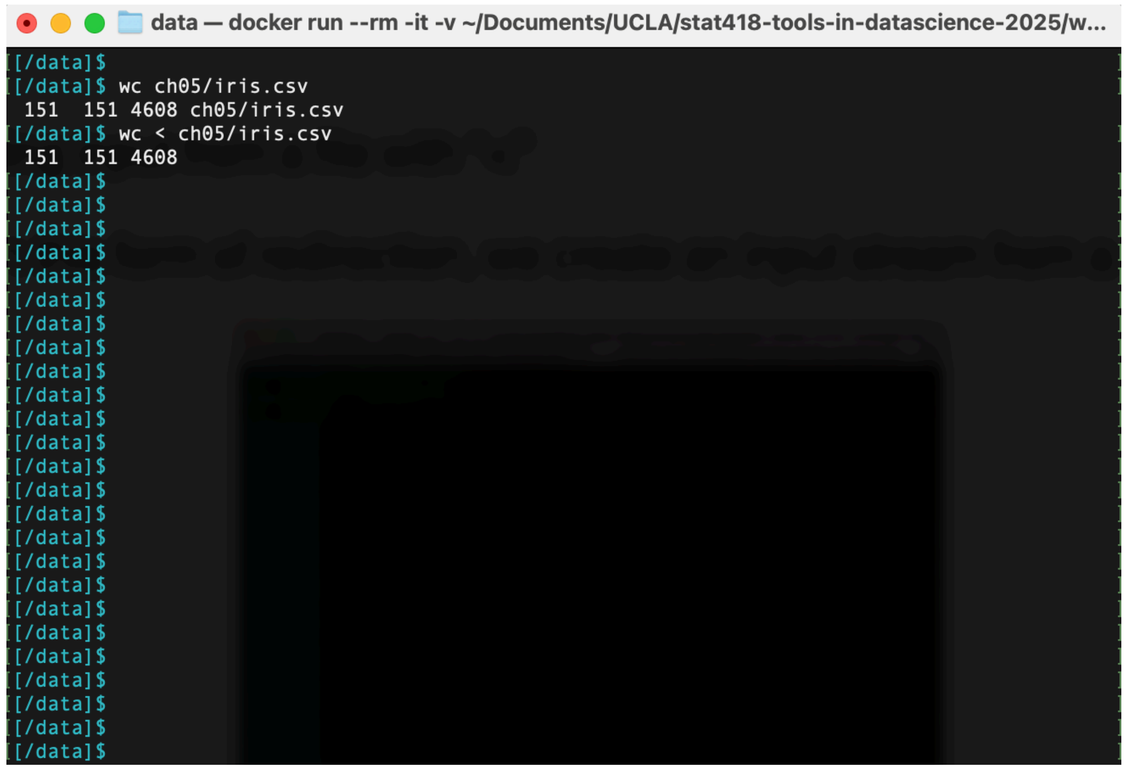
With this form of redirection, we take a stream of processed data and store it in a file

Example

```
cut -d”,” -f5 ch05/iris.csv > species.txt
```

Taking input from a file with “<”

With this form of redirection, we create an input stream from a file



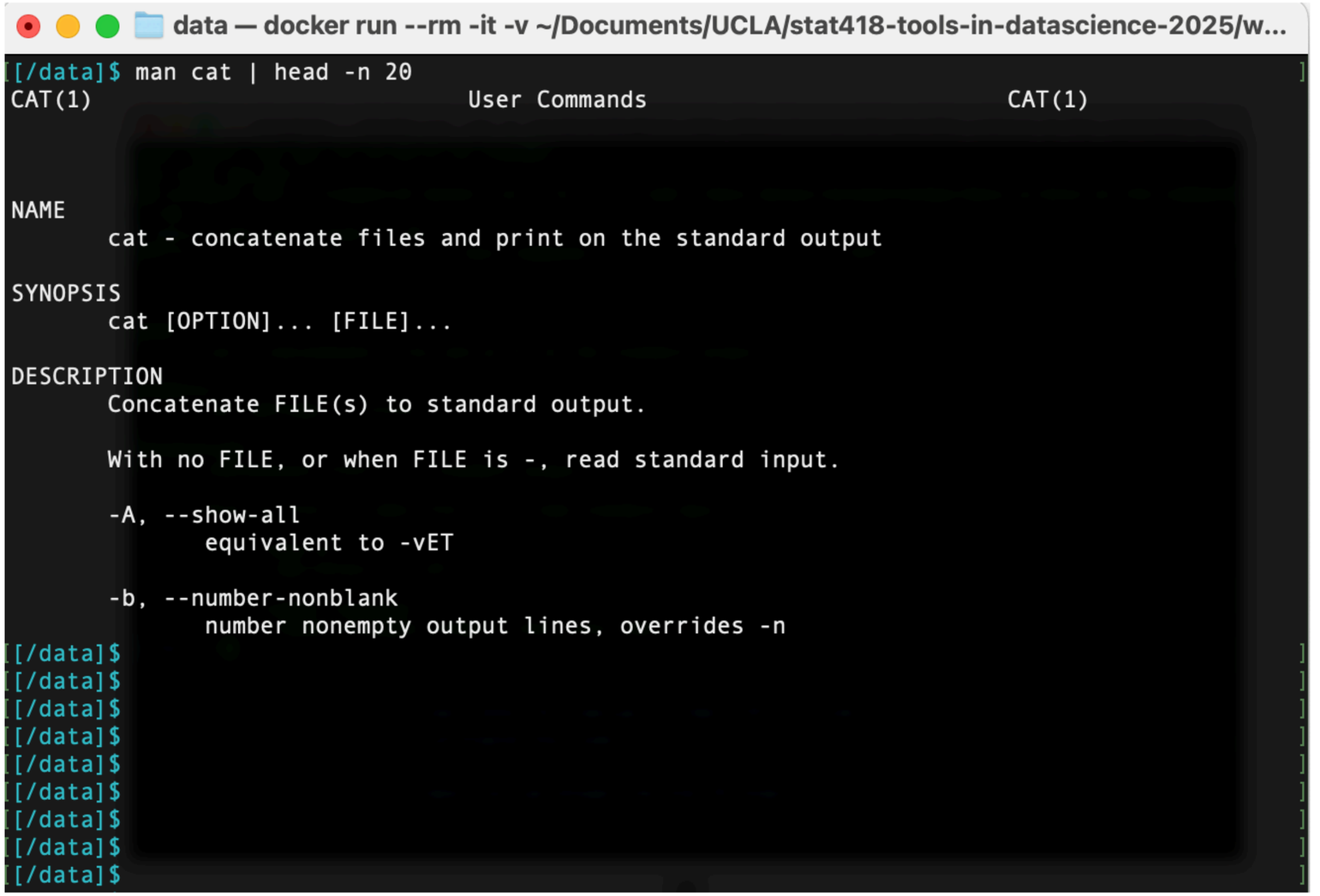
Read the man and help pages!

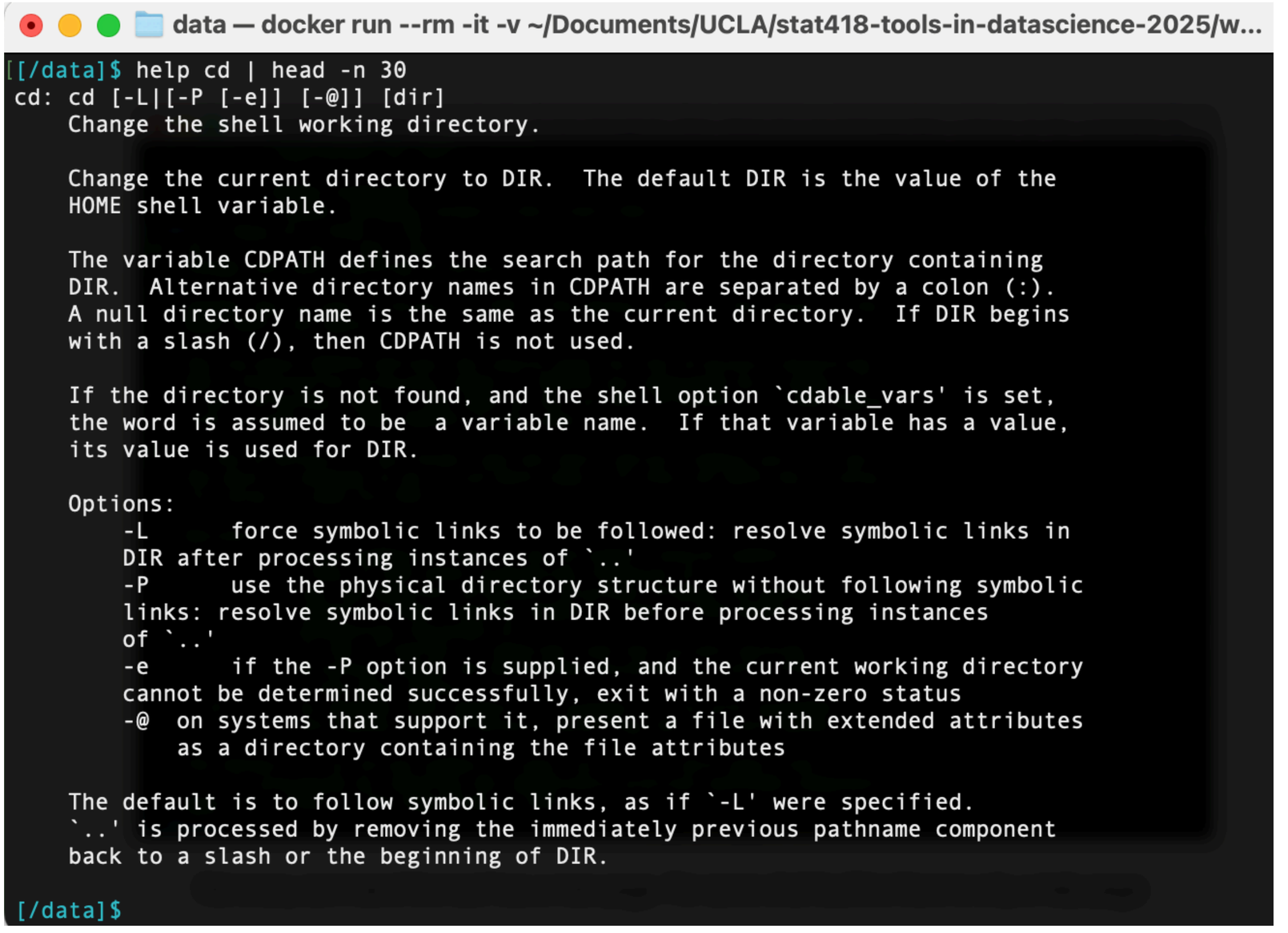
If the command uniq is unfamiliar you can look up its usage

```
man uniq
```

or

```
uniq --help
```


data — docker run --rm -it -v ~/Documents/UCLA/stat418-tools-in-datascience-2025/w...
[/data]$ man cat | head -n 20
CAT(1) User Commands CAT(1)
NAME
cat - concatenate files and print on the standard output
SYNOPSIS
cat [OPTION]... [FILE]...
DESCRIPTION
Concatenate FILE(s) to standard output.
With no FILE, or when FILE is -, read standard input.
-A, --show-all
equivalent to -vET
-b, --number-nonblank
number nonempty output lines, overrides -n
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$
[/data]$



Not every command-line tool has a man page. For some shell builtins there is only a help page

again I've shown only the top 30 lines

Accessing a URL

When accessing a URL (uniform resource locator) through your browser the data that is being downloaded can be interpreted; html as a website, MP3 as played music, pdf automatically downloaded

When cURL is used to access a URL the data is downloaded and is printed to standard output. As we have seen we can simply specify as an argument

```
$ curl -s https://www.gutenberg.org/cache/epub/76/pg76.txt | head
-n 10
```

This downloads the book Adventures of Huckleberry Finn by Mark Twain from Project Gutenberg that we saw before. The **`-s`** command stands for *silent* so a progress meter

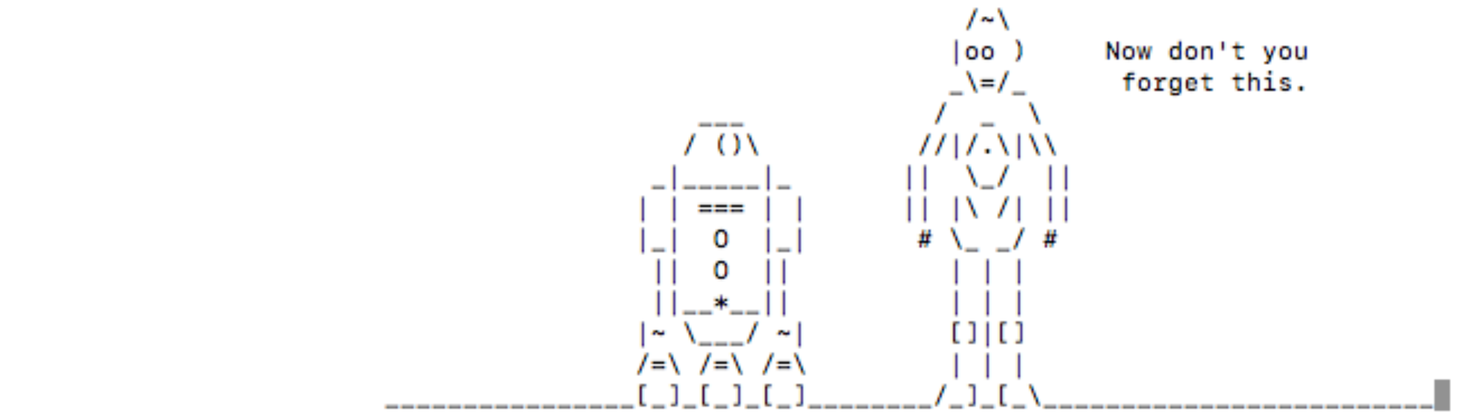
Saving a URL

Or course we can continue to combine our commands to save a file

```
$ curl -s https://www.gutenberg.org/cache/epub/76/pg76.txt > data/ch03/
finn.txt
```

Don’t need to specify **`-s`** in this case

Star Wars: Fun, but… useless?



try

`$ nc towel.blinkenlights.nl 23` [Unfortunately this doesn't work anymore]

It does exist at https://telehack.com and run starwars, but not the same as doing it directly from cl

Data Acquisition (…still in the shell)

Accessing the New York Times

The New York Times is one of the most well-known, reputable, widely circulated news outlets worldwide. The have the complete archive of all their articles available dating back to 1851. Obviously this is a huge wealth of information that we can access through their developers API.

Why does NYT (or anyone) provide an API? In their words:

3. Why are you offering APIs?

Like many organizations, we hope to encourage innovation through collaboration. When you build applications, create mashups and otherwise reveal the potential of our data, we learn more about what our readers want and gain insight into how news and information can be reimagined. We're hoping you'll show us what's next for The Times.

But we also have a simpler, more compelling reason: journalism. To inform the public or tell a story, we use articles, photos, videos, interactive graphics, slideshows and more. Data has always been the primary force behind those features, and now it can become a feature in its own right. Our APIs help us fulfill the newspaper's journalistic mission by putting more information in the hands of the public — and they also expand that mission by giving users the ability to find and tell their own stories.

nytimes.com

Monday, April 14, 2025
Today's Paper

The New York Times

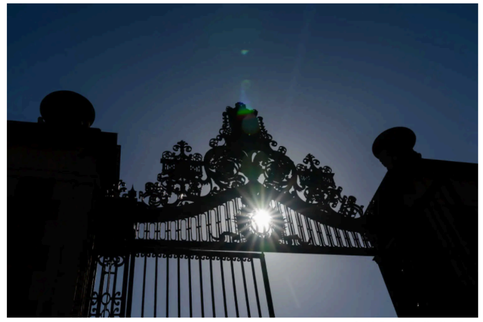
Nasdaq -0.05% ↓

U.S. World Business Arts Lifestyle Opinion | Audio Games Cooking Wirecutter The Athletic

Inside Trump's Pressure Campaign on Universities

The opaque process, part of a strategy by President Trump and conservatives to realign the liberal tilt of elite universities, has upended higher education.

9 MIN READ



Sophie Park for The New York Times

Justin T. Gellerson for The New York Times

She's Young, Trump-Friendly, and Has a White House Press Pass

Natalie Winters, a protégée of Stephen Bannon, belongs to a newly high-profile contingent of West Wing reporters.

6 MIN READ

Tariffs What to Know Fact Check Timeline Inside Trump's Reversal

Trump Signals Semiconductor Tariffs Are Coming 'Very Soon'

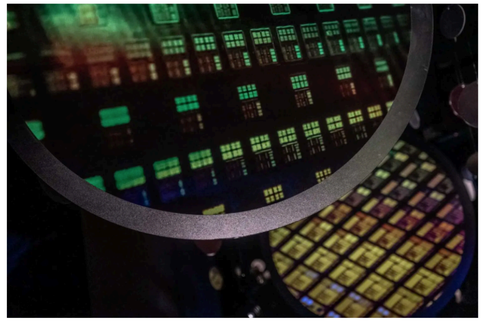
President Trump has suggested that he will move forward with a national security investigation this week that is likely to result in tariffs on chips.

4 MIN READ

Europe Fears U.S. Tariffs Will Set Off a Flood of Cheap Goods From China

President Trump's tariffs on China could lead to a hazardous scenario for Europe: artificially cheap products that could undermine local industries.

5 MIN READ



Lam Yik Fei for The New York Times

What a Wirecutter Editor Learned After Joining the Cult of Quince

FROM WIRECUTTER

'S.N.L.': Laughing All the Way to Financial Chaos

4 MIN READ

Opinion

James Carville: How to Turn Trump's Economic Chaos Against Him

5 MIN READ

China Halts Critical Exports as Trade War Intensifies

6 MIN READ

When Elected Leaders Pursue Risky Policies, What Can Stop Them?

6 MIN READ

Stocks Jump After More Tariff Whiplash

3 MIN READ

MICHELLE COTTLE

Matt Gaetz Hit the Skids. These Days, It's Not Disqualifying.

5 MIN READ

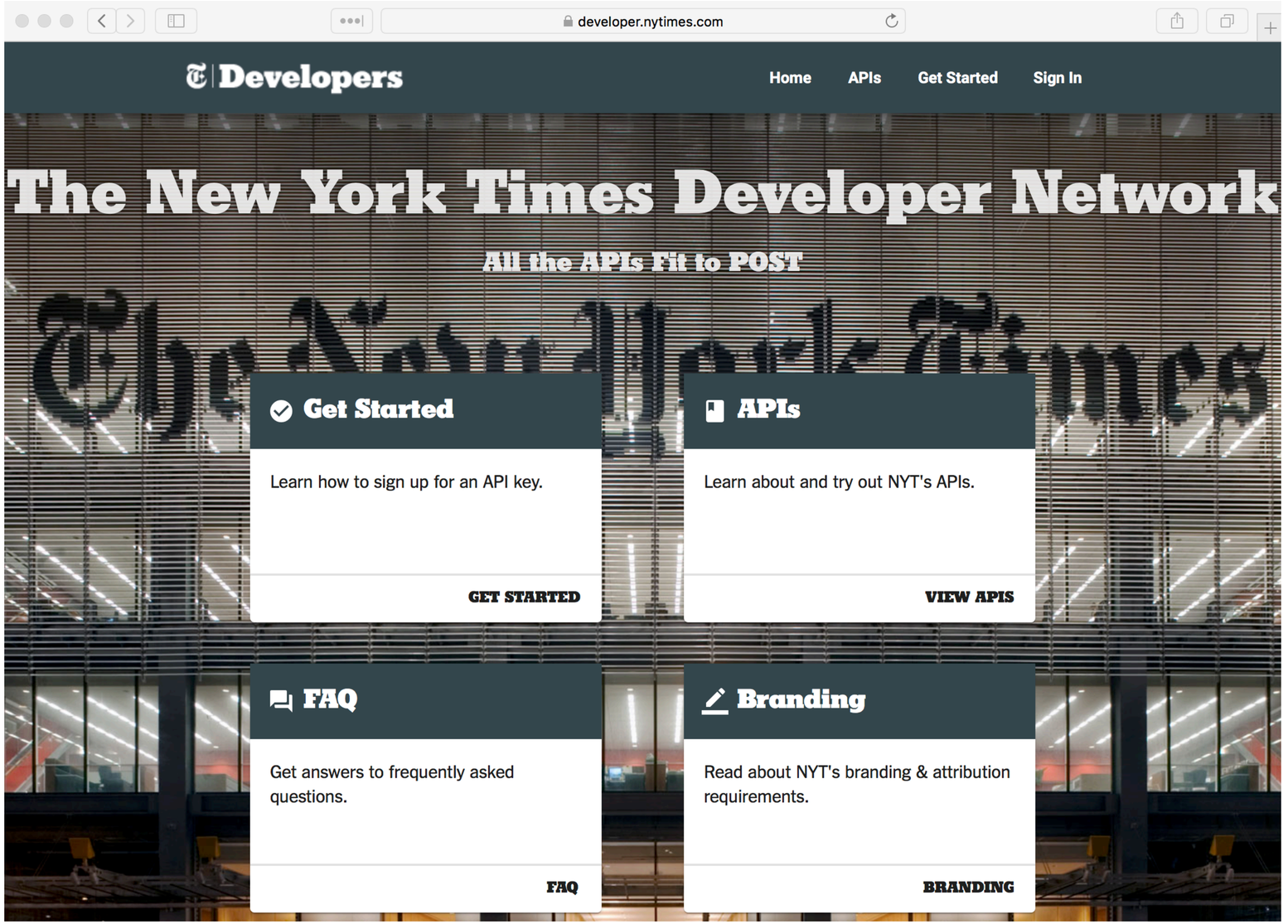
Trump Administration Major Moves Approval Ratings Lawsuits Tracker Tracking the Layoffs

With the ongoing news of tariffs since President Trump came into office, the NYT front page is full of articles about the about them on a daily basis..

I was curious as to how many articles had been posted in the previous five years prior to his election and inagauration.

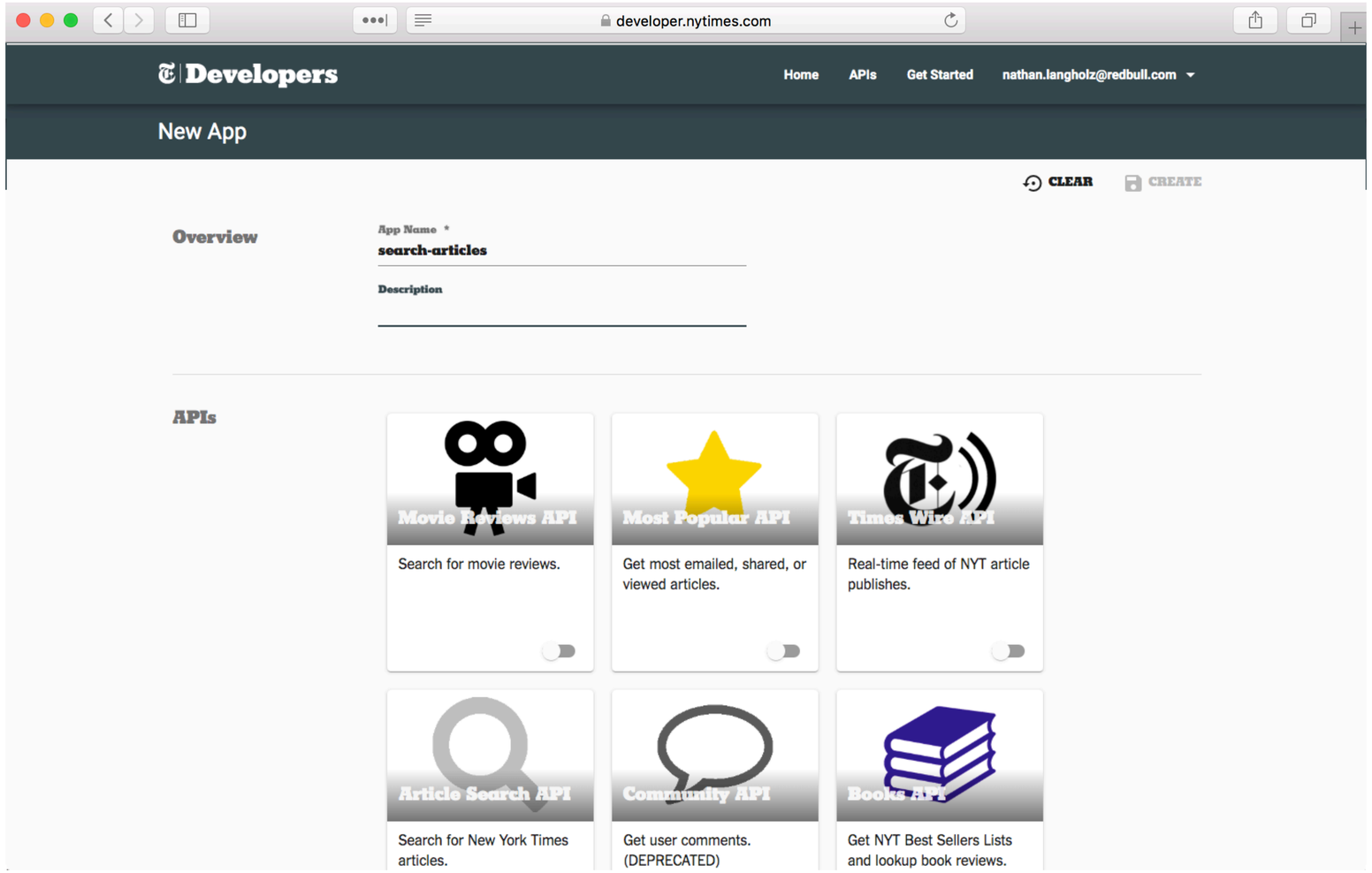
This is the New York Times Developer Network

If you want to follow along you will need to make an account to get your own API key



You will need to create an app; name doesn't matter

We will be using the Article Search API, but it is a similar process to use any of the others



FAQ

1. What are the Times APIs?
Our APIs (Application Programming Interfaces) allow you to programmatically access New York Times data for use in your own applications. Our goal is to facilitate a wide range of uses, from custom link lists to complex visualizations. Why just read the news when you can hack it?

NYT currently has ten public APIs: Archive, Article Search, Books, Most Popular, Semantic, Times Newswire, TimesTags, and Top Stories.

2. Who is the intended audience for the Times APIs?
We've designed our APIs for the web developer community, but all non-commercial users are welcome. See our Terms of Use for more information.

3. Why are you offering APIs?
Like many organizations, we hope to encourage innovation through collaboration. When you build applications, create mashups and otherwise reveal the potential of our data, we learn more about what our readers want and gain insight into how news and information can be reimagined. We're hoping you'll show us what's next for The Times.

But we also have a simpler, more compelling reason: journalism. To inform the public or tell a story, we use articles, photos, videos, interactive graphics, slideshows and more. Data has always been the primary force behind those features, and now it can become a feature in its own right. Our APIs help us fulfill the newspaper's journalistic mission by putting more information in the hands of the public — and they also expand that mission by giving users the ability to find and tell their own stories.

4. What kinds of data can I access with Times APIs?
Please see our APIs page for the current list of available APIs.

5. How do I use the Times APIs?
Our APIs use a RESTful style and a resource-oriented architecture. Calls are made via HTTPS requests. Your request URIs should be patterned after the examples in the API documentation, and you should always include your API key in a query string. See the documentation for each API for more details on request parameters and URI structure.

6. Why can't I access all NYTimes.com content?
For each set of data we open up to the developer community, we have to consider a host of issues, ranging from load balancing to copyright law. If you have ideas for our APIs or are hoping to get access to a specific kind of data, contact us at code@nytimes.com.

7. Why do I have to register for a key to use your APIs?
Key registration allows us to monitor usage levels and ensure that developers are complying with our Terms of Use. We respect your privacy and will not share your registration information with third parties.

8. Do I have to agree to your Terms of Use just to test an API?
Our Terms of Use apply to all uses of our APIs. Please agree to the terms before testing or using an API.

9. Can I use your APIs commercially?
Our APIs are for non-commercial use only. For details, see our Terms of Use. If you have a commercial use case, please contact us at nytlg-sales@nytimes.com or visit https://nytlicensing.com/. For Text & Data Mining options, please visit https://nytlicensing.com/data-solutions/.

10. What do you mean by "commercial purposes"?
Here are a few examples of what we consider commercial purposes:

- Selling New York Times content or data in any application.
- Charging a subscription fee or selling advertising for any service or application that includes New York Times content or data.
- Selling any application built with one of our APIs.
- Using our content inside an application that is paid or has a paid tier currently or in the future.

If you aren't sure whether your plans constitute "commercial purposes," please contact us at nytlg-sales@nytimes.com. For commercial use, please visit https://nytlicensing.com/ and for Text & Data Mining, please visit https://nytlicensing.com/data-solutions/.

11. Is there an API call limit?
Yes, there are two rate limits per API: 500 requests per day and 5 requests per minute. You should sleep 12 seconds between calls to avoid hitting the per minute rate limit. If you need a higher rate limit, please contact us at code@nytimes.com.

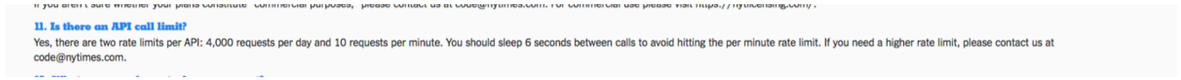
12. What response formats do you support?
Data is returned as JSON. Specific APIs may also return other formats. See the documentation for each API for more details.

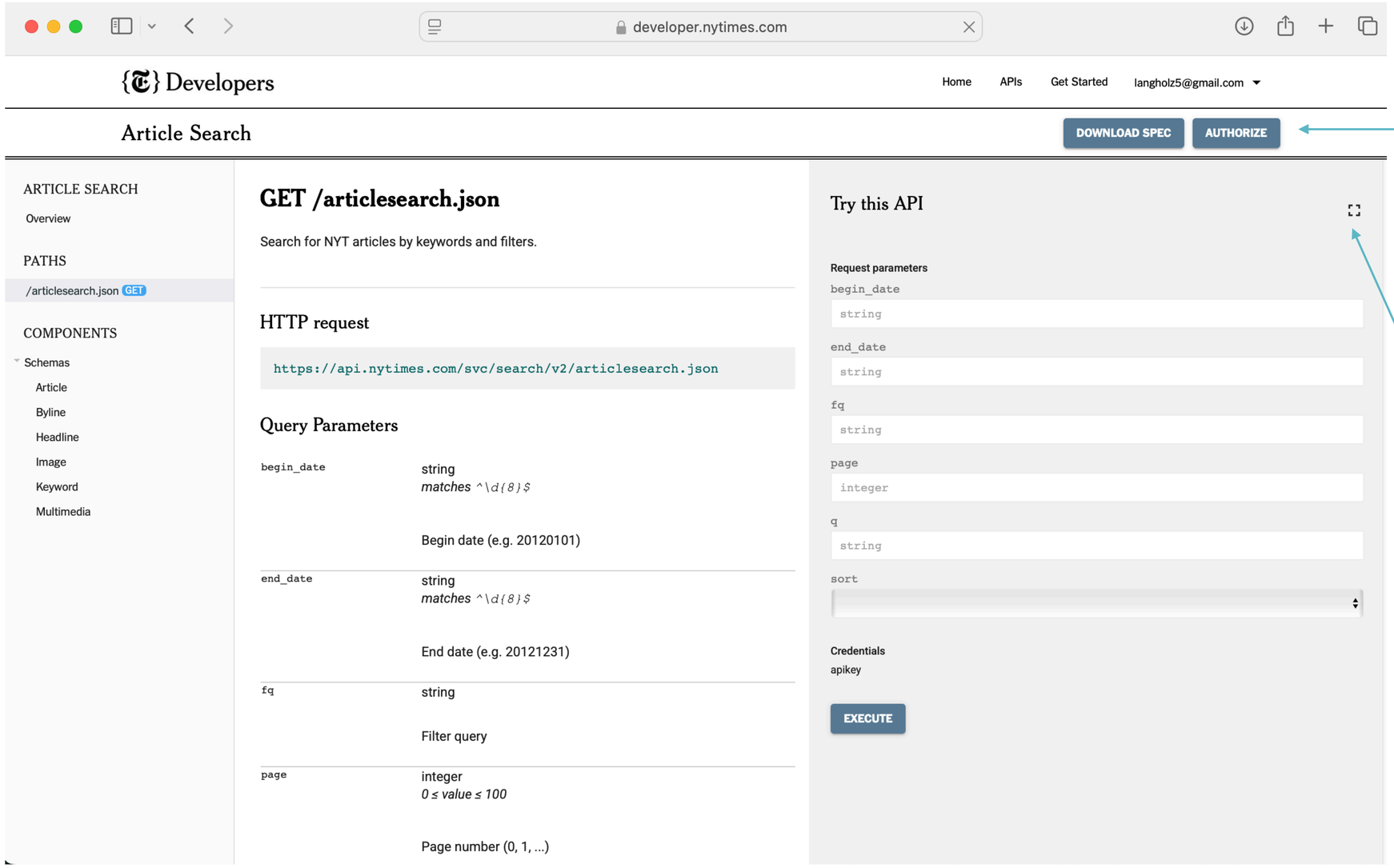
13. What happened to the Community API?
The Times is no longer providing reader comments for non-commercial use cases via the Developer Portal. We apologize for any inconvenience that this might cause. We expect to have details regarding a commercial offering in 2023 and for further information, please contact nytlg-sales@nytimes.com.

14. Can I make suggestions for future development?
Yes, please! You can email us at code@nytimes.com.

This is important to note

This is what it was in 2019





You will need to authorize your credentials

You can also test query and get the resulting url to test in a browser

Running a shell script

There are two ways to run a shell script; you can either execute it within a new shell or make the file executable. Executing in a new shell using either:

```
$ bash <filename.sh>
```

```
$ ./<filename.sh>
```

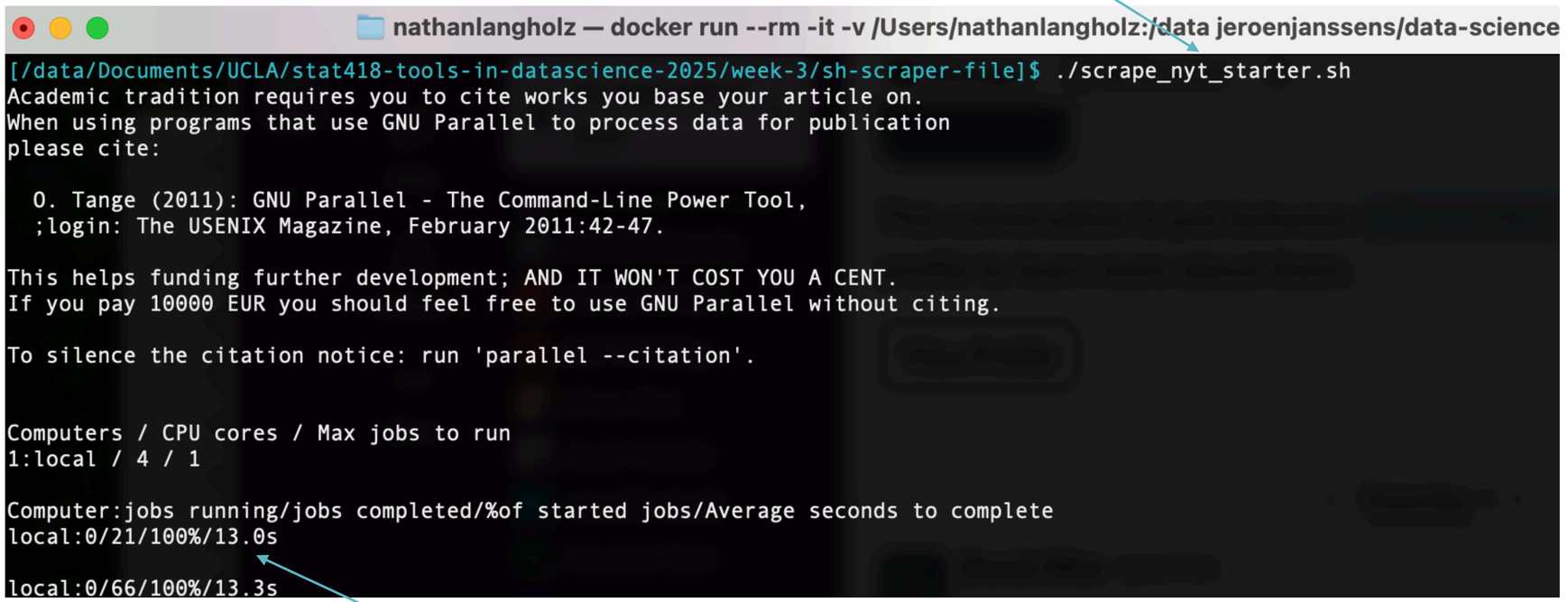
This should explain the .sh suffix; this naming convention will help you (and others) recognize this as a shell script

Making a file executable means making it become just like any command Unix knows. This requires changing file permissions which we saw briefly earlier but we won't cover

Executing .sh file

Lets see how this works in the shell

First, I've gone to my class directory to make a scraper file directory



Average amount of time to complete request

scrape_nyt_starter.sh

```
#!/bin/bash

#API_KEY = $1
#echo "API Key: $API_KEY"

# this pulls data from the NY Times developer api through the 'Data Science at the Command Line' book linux environment
#script adapted from 'Data Science at the Command Line' book

#https://api.nytimes.com/svc/search/v2/articlesearch.json?q=tariffs&begin_date=20200101&end_date=20251231&page=1&api-key='$API_KEY'

#takes some time as there is a 13 second delay
parallel -j1 --progress --delay 13 --results results "curl -sL "\
"'http://api.nytimes.com/svc/search/v2/articlesearch.json?q=tariffs&'"\
"'begin_date={1}0101&end_date={1}1231&page={2}&sort=newest&api-key='"\
"'{your-api-key}'" ::: {2021..2025} ::: {0..20} > /dev/null

cat results/1/*/2/*/stdout | head -n 5

tree results | head

cat results/1/*/2/*/stdout | jq -c '.response.docs[] | {date: .pub_date, type: .document_type, '\
'title: .headline.main }' > articles

jq -r '[.date, .type, .title] | @csv' articles > tariff-articles.csv

rm articles
rm -r results

wc -l tariff-articles.csv

#< tariff-articles.csv cols -c date cut -dT -f1 | head | csvlook
```

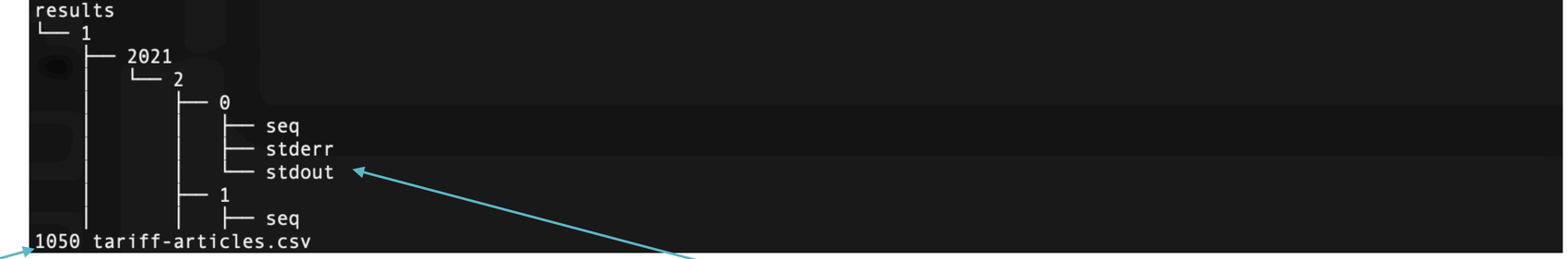
There is a lot going on in this command. I have delayed each request by 13 seconds, and am searching for articles from 2021 to 2025 with 20 pages of results

Would need to pass put your api-key or pass it in the execution command

Format results from json to csv from these two lines. Also where columns are selected

Snippet of data returned from API request


```
nk":9},{"name":"Location","value":"Sunderland (England)","rank":10}],"multimedia":{"caption":"Ashwani Gupta, Nissan's chief operatin
g officer, in Sunderland, England, on Thursday.","credit":"Phil Noble/Reuters","default":{"url":"https://static01.nyt.com/images/202
1/07/01/business/01economy-briefing-nissan/merlin_190183539_efee7418-a627-4e98-995d-8010a75a9c54-articleLarge.jpg","height":399,"wid
th":600},"thumbnail":{"url":"https://static01.nyt.com/images/2021/07/01/business/01economy-briefing-nissan/01economy-briefing-nissan
-thumbStandard.jpg","height":75,"width":75}},"news_desk":"Business","print_page":"3","print_section":"B","pub_date":"2021-07-01T09:3
5:29Z","section_name":"Business","snippet":"The Japanese carmaker will also make a new electric vehicle in England as part of a £1 b
illion investment, partly supported by the government.","source":"The New York Times","subsection_name":"","type_of_material":"News"
,"uri":"nyt://article/c22f8bf4-2a49-5b5a-a60a-2760d1b892f4","web_url":"https://www.nytimes.com/2021/07/01/business/nissan-battery-fa
ctory-britain.html","word_count":613},{"abstract":"Forced to pay more for wood, home builders want the Biden administration to settl
e a long-running dispute over Canadian imports. It won't be easy.","byline":{"original":"By Thomas Kaplan"},"document_type":"article
","headline":{"main":"High Lumber Prices Add Urgency to a Decades-Old Trade Fight","kicker":"","print_headline":"Lumber Prices Raise
 Stakes In an Old Feud With Canada"},"_id":"nyt://article/c763b808-86fa-5bb8-9a83-435e66b841a0","keywords":[{"name":"Subject","value
":"Wood and Wood Products","rank":1},{"name":"Subject","value":"International Trade and World Market","rank":2},{"name":"Subject","v
alue":"Customs (Tariff)","rank":3},{"name":"Subject","value":"Prices (Fares, Fees and Rates)","rank":4},{"name":"Organization","valu
e":"National Assn of Home Builders","rank":5},{"name":"Subject","value":"Real Estate and Housing (Residential)","rank":6},{"name":"S
ubject","value":"United States International Relations","rank":7},{"name":"Subject","value":"United States Politics and Government",
"rank":8}],"multimedia":{"caption":"A house being built in Vestavia Hills, Ala. A trade dispute between the United States and Canada
 is adding a layer of diplomatic intrigue to the scramble for in-demand building materials.","credit":"Wes Frazer for The New York T
imes","default":{"url":"https://static01.nyt.com/images/2021/06/25/us/00dc-lumberfight01/merlin_189846987_a6129ce0-4bde-4d72-93f2-9a
2c6073bb9d-articleLarge.jpg","height":400,"width":600},"thumbnail":{"url":"https://static01.nyt.com/images/2021/06/25/us/00dc-lumber
fight01/00dc-lumberfight01-thumbStandard.jpg","height":75,"width":75}},"news_desk":"Business","print_page":"1","print_section":"B","
pub_date":"2021-06-28T19:50:05Z","section_name":"Business","snippet":"Forced to pay more for wood, home builders want the Biden admi
nistration to settle a long-running dispute over Canadian imports. It won't be easy.","source":"The New York Times","subsection_name
":"Economy","type_of_material":"News","uri":"nyt://article/c763b808-86fa-5bb8-9a83-435e66b841a0","web_url":"https://www.nytimes.com/
2021/06/28/business/economy/lumber-prices-canadian-trade.html","word_count":1518},{"abstract":"It lifts all boats, but not by equal
amounts.","byline":{"original":"By Thomas B. Edsall"},"document_type":"article","headline":{"main":"Is Education No Longer the 'Grea
t Equalizer'?","kicker":"Guest Essay","print_headline":""},"_id":"nyt://article/bd5fb540-4ac3-5281-9bee-44cb33639e7a","keywords":[{"
name":"Subject","value":"United States Economy","rank":1},{"name":"Subject","value":"Education","rank":2},{"name":"Subject","value":
"Income Inequality","rank":3},{"name":"Subject","value":"Parenting","rank":4},{"name":"Location","value":"United States","rank":5}],
"multimedia":{"caption":"","credit":"Audra Melton for The New York Times","default":{"url":"https://static01.nyt.com/images/2021/06/
23/opinion/23edsall-new/merlin_174339084_85264678-082c-4a8d-95e7-de6b906d4ef5-articleLarge.jpg","height":400,"width":600},"thumbnail
":{"url":"https://static01.nyt.com/images/2021/06/23/opinion/23edsall-new/23edsall-new-thumbStandard.jpg","height":75,"width":75}},"
news_desk":"OpEd","print_page":"","print_section":"","pub_date":"2021-06-23T09:00:14Z","section_name":"Opinion","snippet":"It lifts
all boats, but not by equal amounts.","source":"The New York Times","subsection_name":"","type_of_material":"Op-Ed","uri":"nyt://art
icle/bd5fb540-4ac3-5281-9bee-44cb33639e7a","web_url":"https://www.nytimes.com/2021/06/23/opinion/education-poverty-intervention.html
","word_count":2664}],"metadata":{"hits":10000,"offset":120,"time":138}}}
```

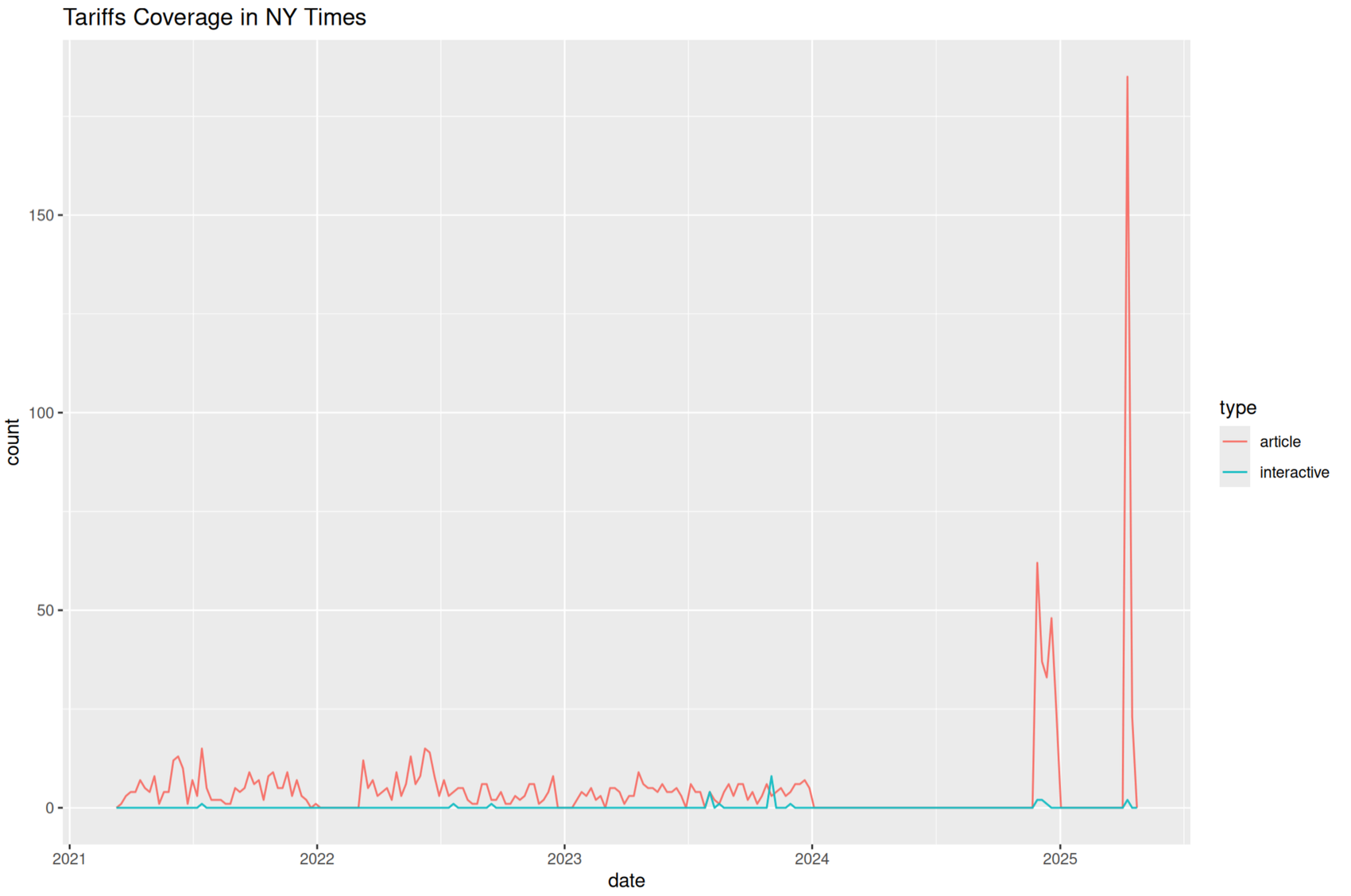



Number of articles

Results tree

The results

A .sh script exists in the GitHub repo to recreate this data acquisition



R and the Shell

There is interactivity between R and the shell in way of:

- executing shell commands in R (`system2`)
- executing shell script commands in Rstudio (`ctrl + enter` sends to terminal)
- executing shell commands in RMarkdown
- execute R commands and scripts in the shell (assuming you have R on your machine)

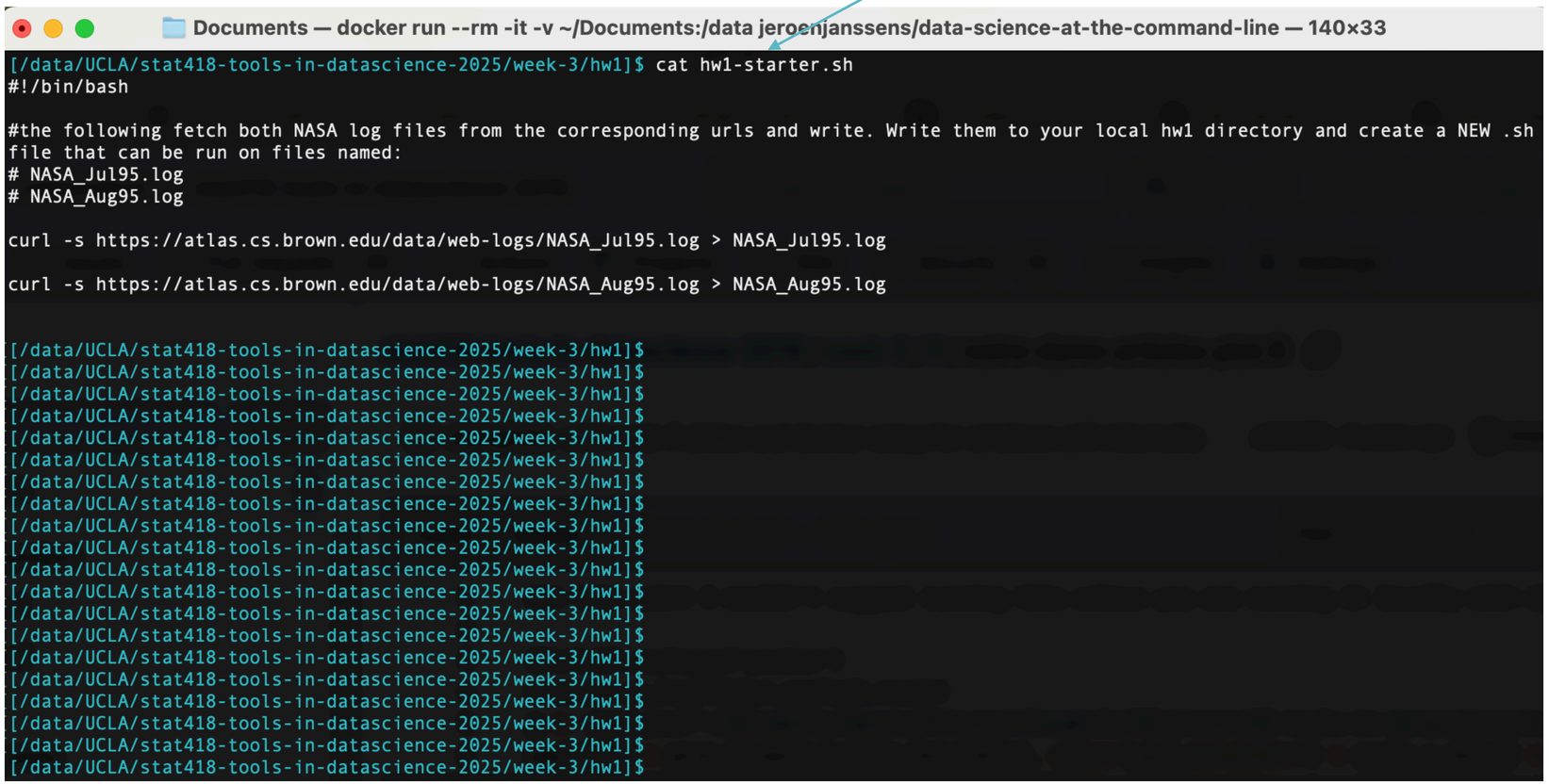
```
$ R -e "head(mtcars); tail(mtcars)”
```

```
$ Rscript -e “head(mtcars)”
```

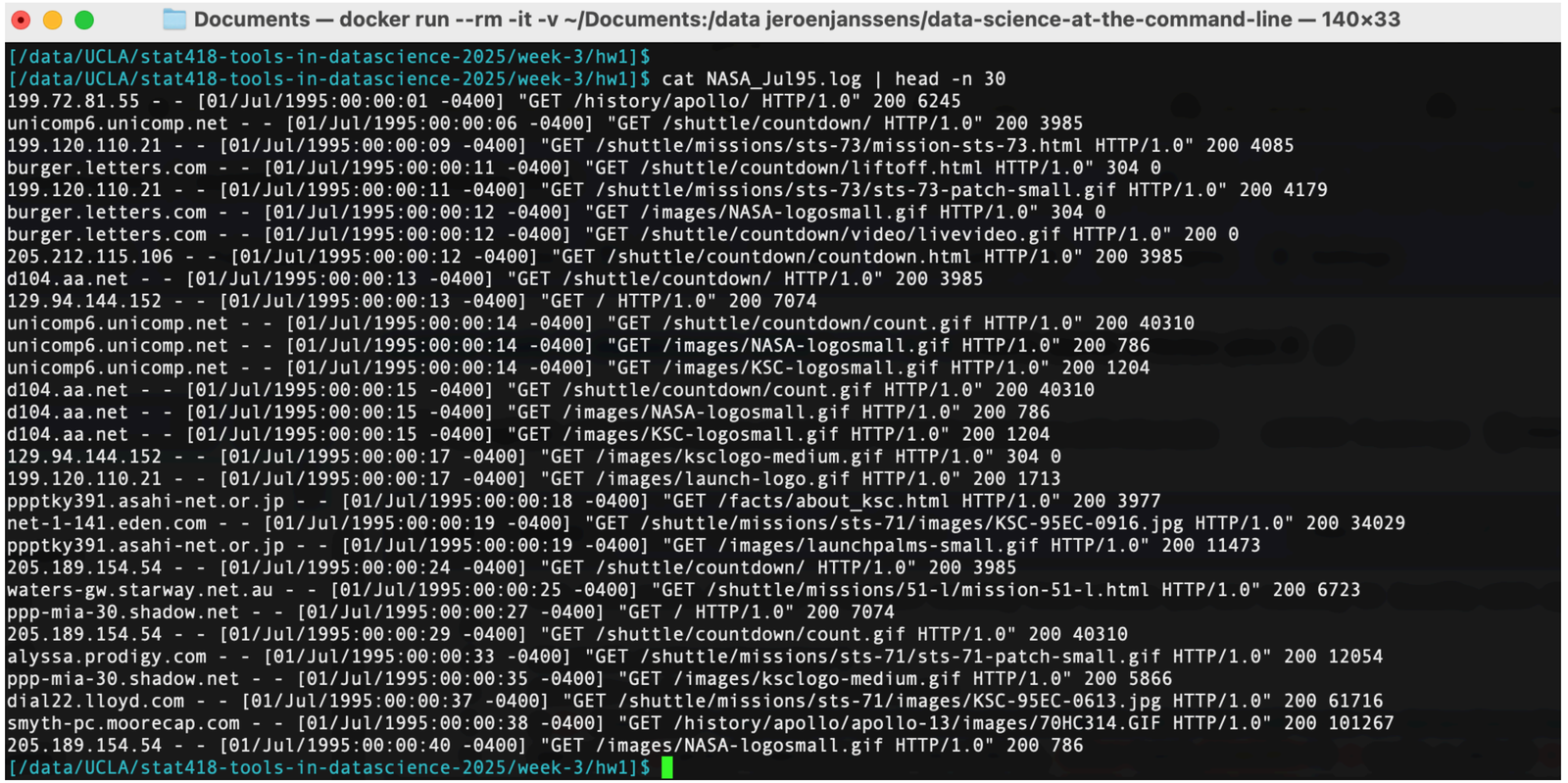
```
$ Rscript <Rscript.R>
```

cat returns the entirety of the file in the terminal

curl goes to the specified url to download log file and then write it



What the July web log actually looks like…



Log format (for homework) - The logs are ASCII with one line per request with the following comments

host - making the request. A hostname when possible otherwise, the IP address

timestamp - in the format “DAY MON DD HH:MM:SS YYYY”

request - given in quotes

HTTP - reply code

bytes - in the reply