STORYTELLING WITH DATA

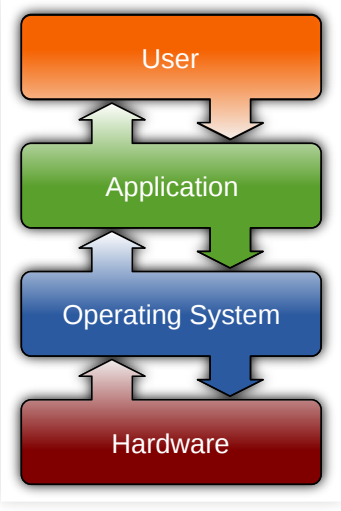
JANUARY 22, 2018

GETTING ACQUAINTED WITH OUR ANALYTIC ENVIRONMENT

WHAT IS AN

OPERATING SYSTEM (OS)?

A BRIDGE BETWEEN APPLICATIONS AND HARDWARE



The most common OSs are Windows and macOS.

The OS should not be confused with the user interface, be it a terminal or GUI.

GRAPHICAL USER INTERFACES ARE QUITE POPULAR

AN OS GUI IS JUST A SPECIAL TYPE OF SHELL

Text-based shells (e.g. the Windows command prompt or Bash) can present as intimidating, but they are your best friend. Graphical shells are just wrappers around some of the commands that shells can execute.

WHAT IS A VIRTUAL MACHINE?

APPLICATIONS THAT EMULATE SOFTWARE SYSTEMS

- VMs allow you to try another OS without a full install
- VMs can enable the use of applications that cannot be run on some systems
- VMs can help us ensure that all of our programs run the same way

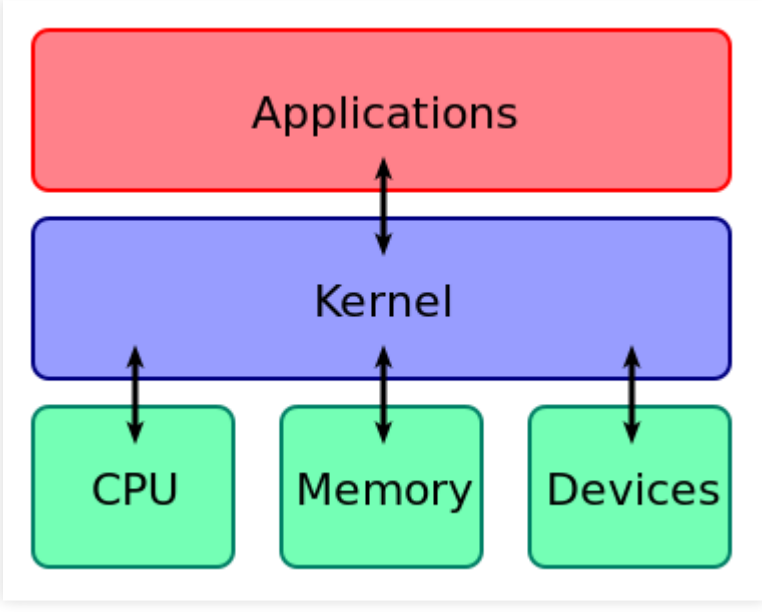
WHY ARE WE USING VIRTUAL MACHINES?

- When software is developed, it assumes a given method of locating the resources it needs
- Typically, software is developed in a single environment, and then adapted to work elsewhere
- Adaptions can create complications in application behavior

WHAT IS LINUX?

LINUX IS NOT AN OS. IT'S A KERNEL

Developed by Linus Torvaldis, Linux is a catchall name for an open-source kernel.



A KERNEL IS ONLY ONE PART OF THE OS...

...but it is the most important part. All in, an OS provides a variety of functions:

- Loads and manages processes
- Provides interface to hardware
- Provides a file system
- Provides a user interface

The kernel essentially coordinates these tasks.

WHY DO WE CARE AGAIN?

- Linux distributions tend to be free for personal use
- Linux distributions tend to be open source
- Linux distributions tend to come packaged with many useful development tools and libraries
- For all these reasons, analytic packages are often designed and tested for use on Linux systems

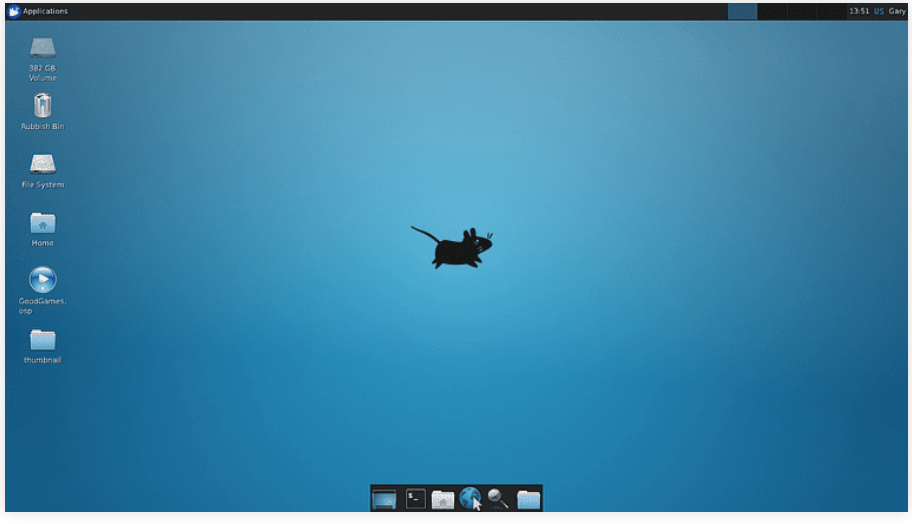
ENTER UBUNTU...

UBUNTU IS ONE OF THE MOST POPULAR LINUX DISTRIBUTIONS BECAUSE...

- ...the GUI is intuitive and easy to use;
- ...it provides full shell functionality
- ...it makes it easy to install and run programs

UBUNTU IS AN OS, NOT A GUI

Multiple command line and graphical shells are compatible. We will be using xfce, but gnome is another popular choice, and unity used to ship with the default Ubuntu install.



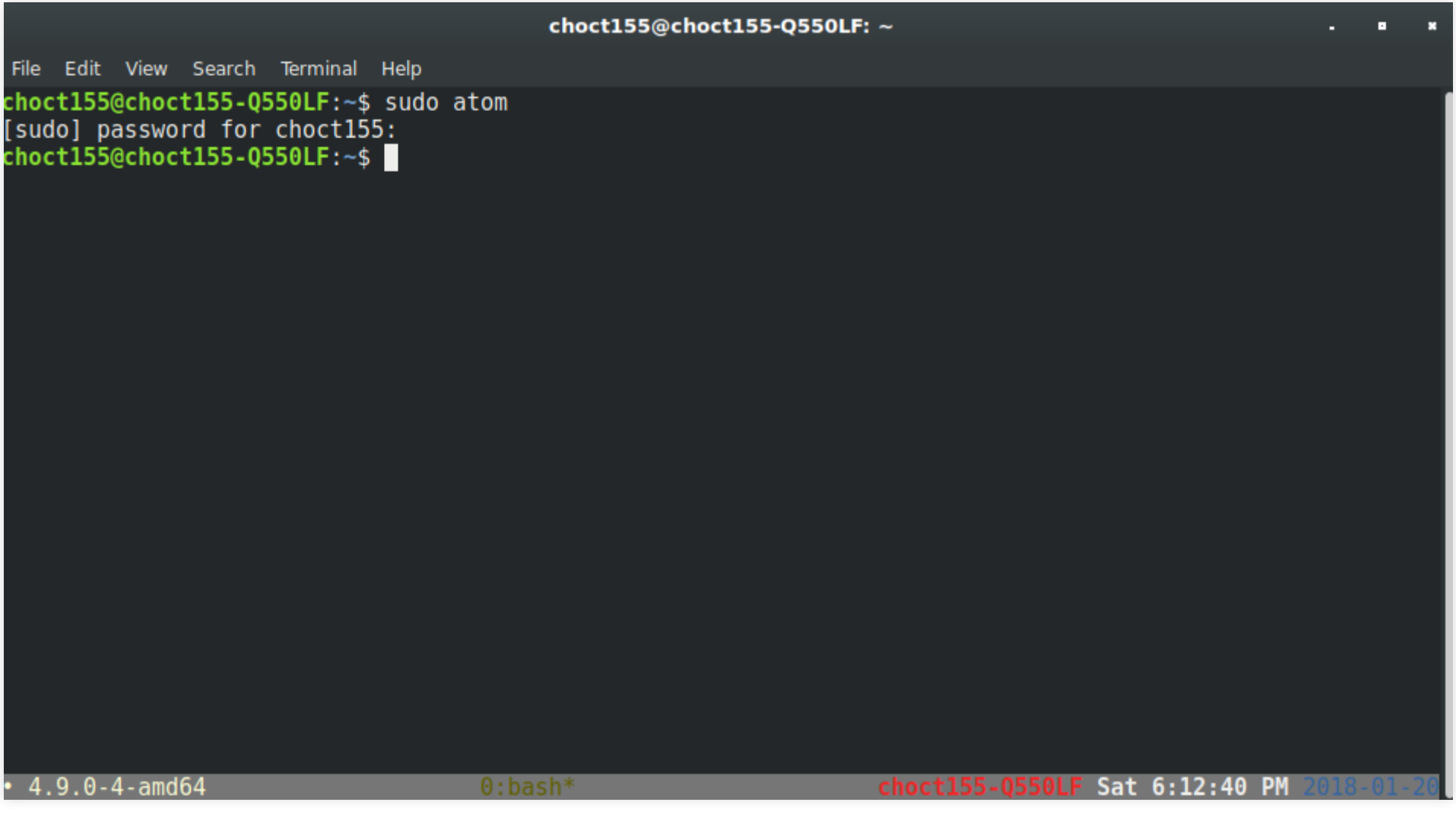
HOW DO I USE A LINUX DISTRIBUTION LIKE UBUNTU?

(SPOILER: YOU CAN USE IT JUST LIKE WINDOWS OR MACOS)

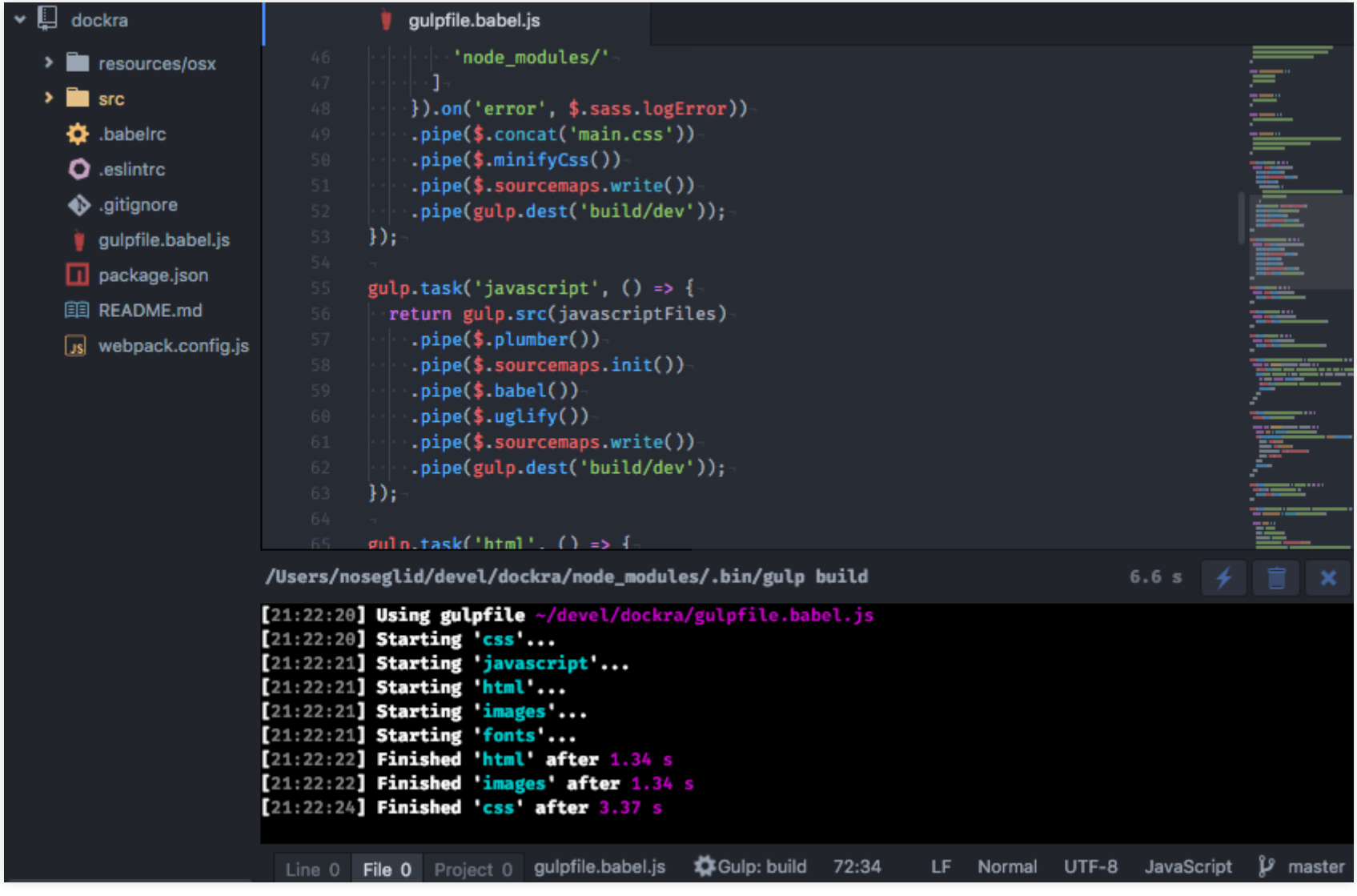
SUPPOSE I WANT TO OPEN AN EDITOR LIKE ATOM

You can navigate to the Atom Editor icon and simply double-click on it...

... OR YOU CAN USE YOUR FRIEND, THE SHELL



THEN YOU CAN CODE TO YOUR HEART'S CONTENT.


dockra
resources/osx
src
.babelrc
.eslintrc
.gitignore
gulpfile.babel.js
package.json
README.md
webpack.config.js
gulpfile.babel.js
'node_modules/'
]
}).on('error', $.sass.logError))
.pipe($.concat('main.css'))
.pipe($.minifyCss())
.pipe($.sourcemaps.write())
.pipe(gulp.dest('build/dev'));
});
gulp.task('javascript', () => {
return gulp.src(javascriptFiles)
.pipe($.plumber())
.pipe($.sourcemaps.init())
.pipe($.babel())
.pipe($.uglify())
.pipe($.sourcemaps.write())
.pipe(gulp.dest('build/dev'));
});
/Users/noseglid/devel/dockra/node_modules/.bin/gulp build
6.6 s
[21:22:20] Using gulpfile ~/devel/dockra/gulpfile.babel.js
[21:22:20] Starting 'css'...
[21:22:21] Starting 'javascript'...
[21:22:21] Starting 'html'...
[21:22:21] Starting 'images'...
[21:22:21] Starting 'fonts'...
[21:22:22] Finished 'html' after 1.34 s
[21:22:22] Finished 'images' after 1.34 s
[21:22:24] Finished 'css' after 3.37 s
Line 0
File 0
Project 0
gulpfile.babel.js
Gulp: build
72:34
LF
Normal
UTF-8
JavaScript
master

BASH IS PROBABLY THE MOST UBQUITOUS SHELL IN COMMON USAGE ON UNIX SYSTEMS

THE BEST WAY TO LEARN ABOUT IT IS BY USING IT

Bash is a low-level interface that allows the user to issue commands to the computer at the command line or by executing a script. Features include but are not limited to ...

- ...potentially complex file management
- ...permissions control
- ...process management
- ...application installation and control

You will be using Bash to, at the very least, start up your VM and version control your projects

Linux Bash Shell Cheat Sheet

Basic Commands

Basic Terminal Shortcuts

CTRL L = Clear the terminal
CTRL D = Logout
SHIFT Page Up/Down = Go up/down the terminal
CTRL A = Cursor to start of line
CTRL E = Cursor the end of line
CTRL U = Delete left of the cursor
CTRL K = Delete right of the cursor
CTRL W = Delete word on the left
CTRL Y = Paste (after CTRL U,K or W)
TAB = auto completion of file or command
CTRL R = reverse search history
!! = repeat last command
CTRL Z = stops the current command (resume with fg in foreground or bg in background)

Basic Terminal Navigation

ls -a = list all files and folders
ls <folderName> = list files in folder
ls -lh = Detailed list, Human readable
ls -l *.jpg = list jpeg files only
ls -lh <fileName> = Result for file only

cd <folderName> = change directory
if folder name has spaces use “ “
cd / = go to root
cd .. = go up one folder, tip: ../../../

du -h: Disk usage of folders, human readable
du -ah: “ “ “ files & folders, Human readable
du -sh: only show disc usage of folders

pwd = print working directory

man <command> = shows manual (RTFM)

Basic file manipulation

cat <fileName> = show content of file
(less, more)
head = from the top
-n <#oflines> <fileName>

tail = from the bottom
-n <#oflines> <fileName>

mkdir = create new folder
mkdir myStuff ..
mkdir myStuff/pictures/ ..

cp image.jpg newimage.jpg = copy and rename a file
cp image.jpg <folderName>/ = copy to folder
cp image.jpg folder/sameImageNewName.jpg
cp -R stuff otherStuff = copy and rename a folder
cp *.txt stuff/ = copy all of *<file type> to folder

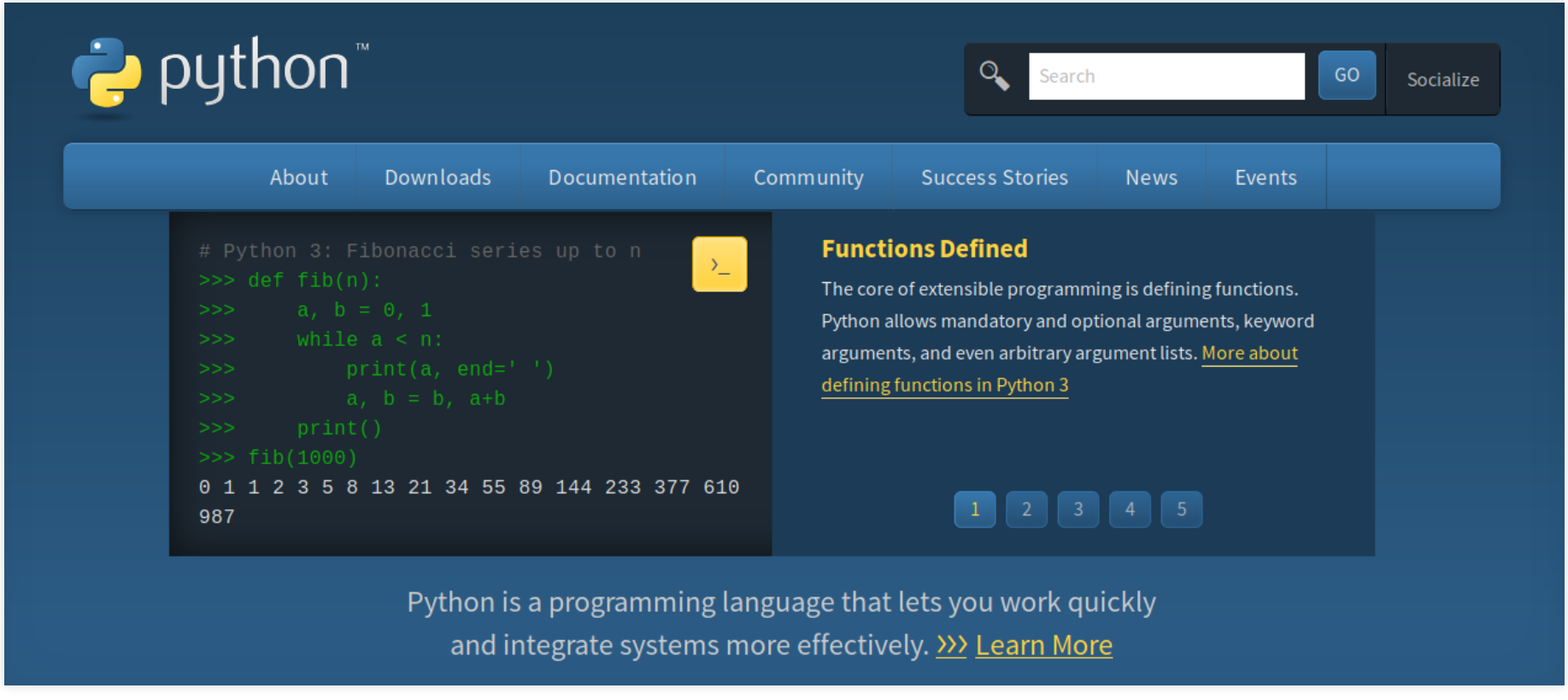
mv file.txt Documents/ = move file to a folder
mv <folderName> <folderName2> = move folder in folder
mv filename.txt filename2.txt = rename file
mv <fileName> stuff/newfileName
mv <folderName>/ .. = move folder up in hierarchy

rm <fileName> .. = delete file (s)
rm -i <fileName> .. = ask for confirmation each file
rm -f <fileName> = force deletion of a file
rm -r <foldername>/ = delete folder

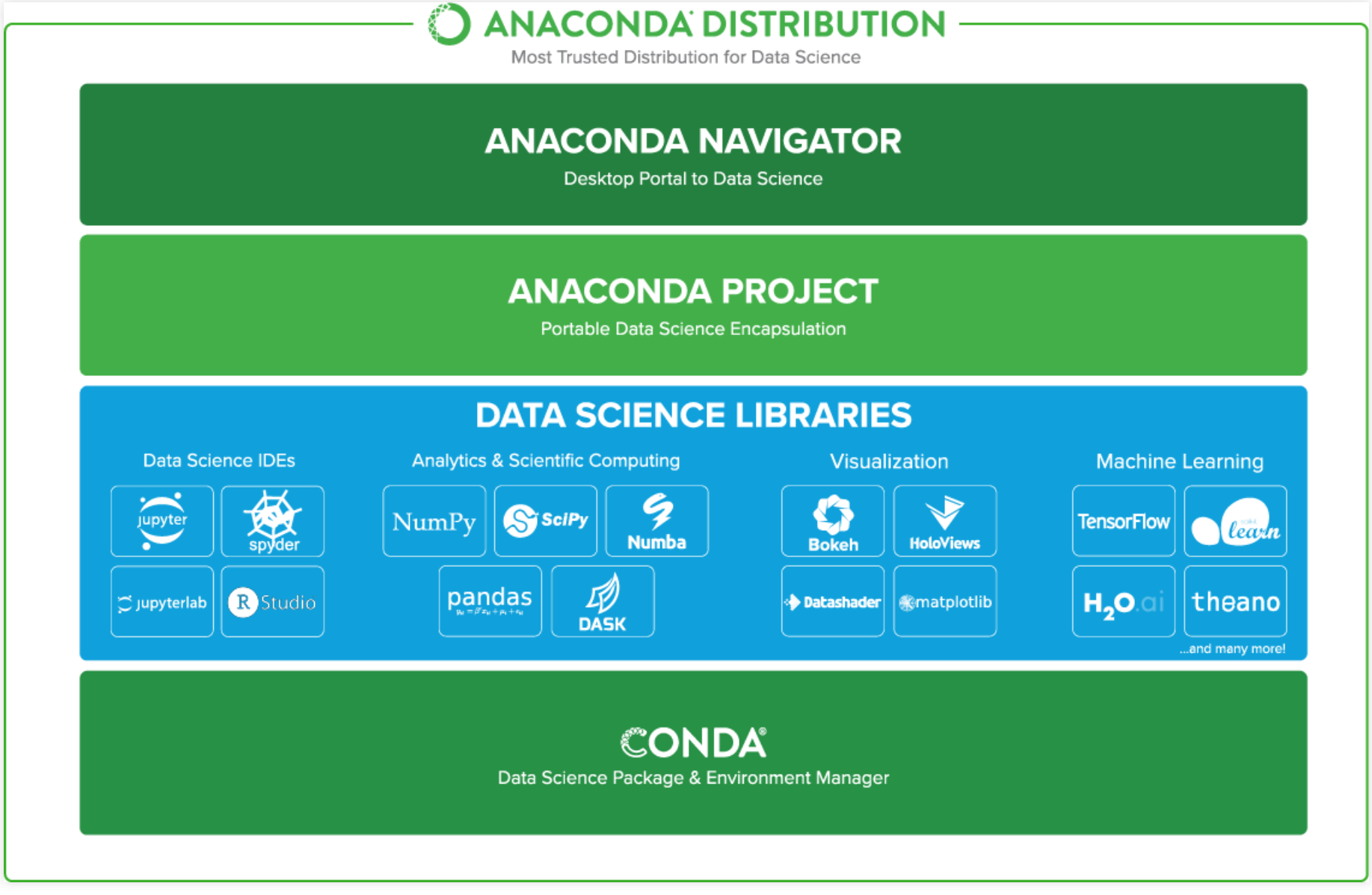
touch <fileName> = create or update a file

ln file1 file2 = physical link
ln -s file1 file2 = symbolic link

PYTHON IS A GENERAL PURPOSE COMPUTING LANGUAGE



WE WILL LEAN HEAVILY ON THE ANACONDA DISTRIBUTION


ANACONDA DISTRIBUTION
Most Trusted Distribution for Data Science
ANACONDA NAVIGATOR
Desktop Portal to Data Science
ANACONDA PROJECT
Portable Data Science Encapsulation
DATA SCIENCE LIBRARIES
Data Science IDEs
Analytics & Scientific Computing
Visualization
Machine Learning
jupyter
spyder
jupyterlab
RStudio
NumPy
SciPy
Numba
pandas
DASK
Bokeh
HoloViews
Datashader
matplotlib
TensorFlow
scikit-learn
H2O.ai
theano
...and many more!
CONDA
Data Science Package & Environment Manager

THE BIGGEST ADVANTAGE OF ANACONDA IS THE CONDA PACKAGE MANAGER

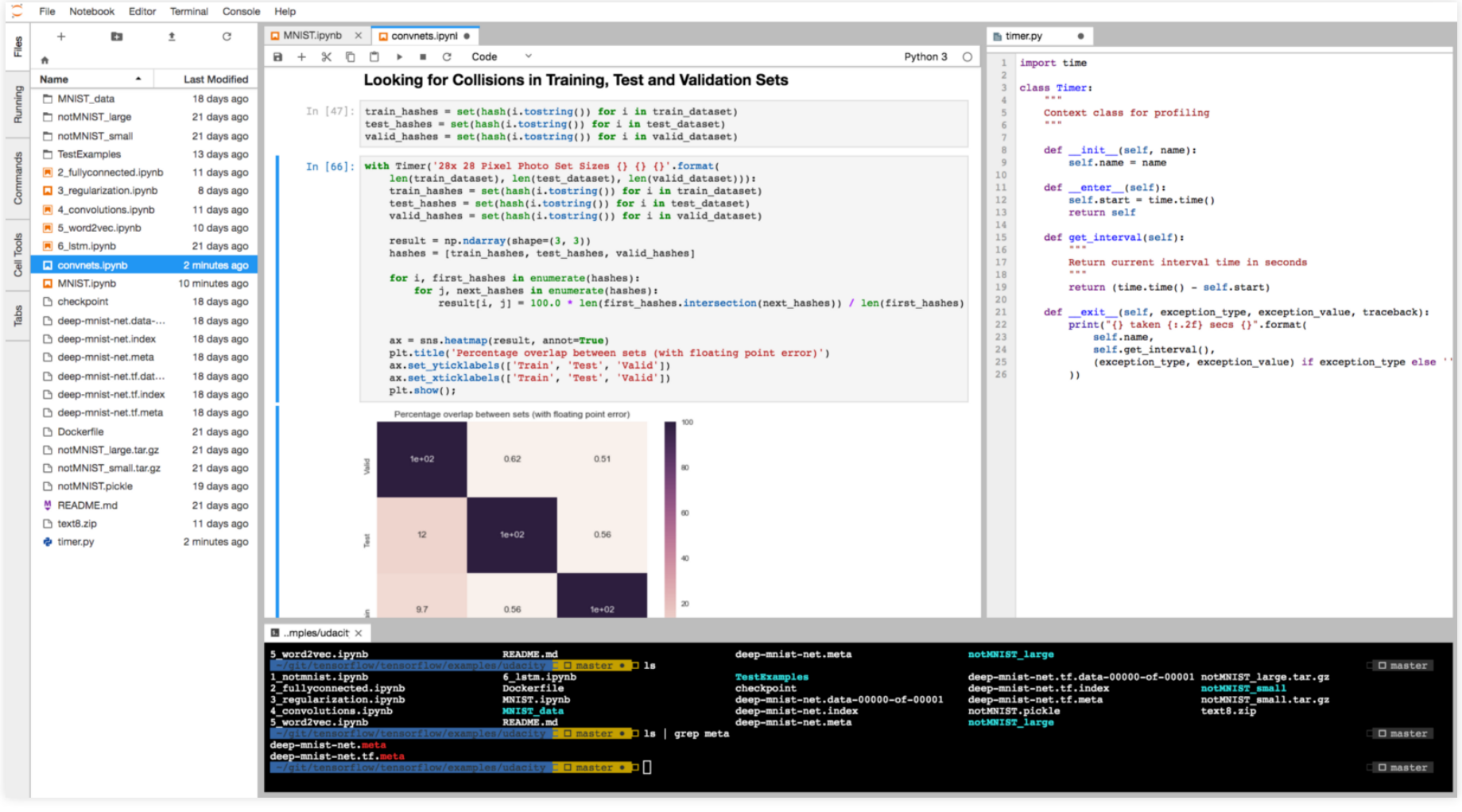
- conda runs on Windows, macOS, and Linux
- It helps find and install packages
- It manages dependencies across packages
- It makes it easy to manage package environments

THE ANACONDA PROJECT IS ALSO COUPLED TIGHTLY WITH THE JUPYTER PROJECT

Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.

OUR CODING ENVIRONMENT:

JUPYTER LAB

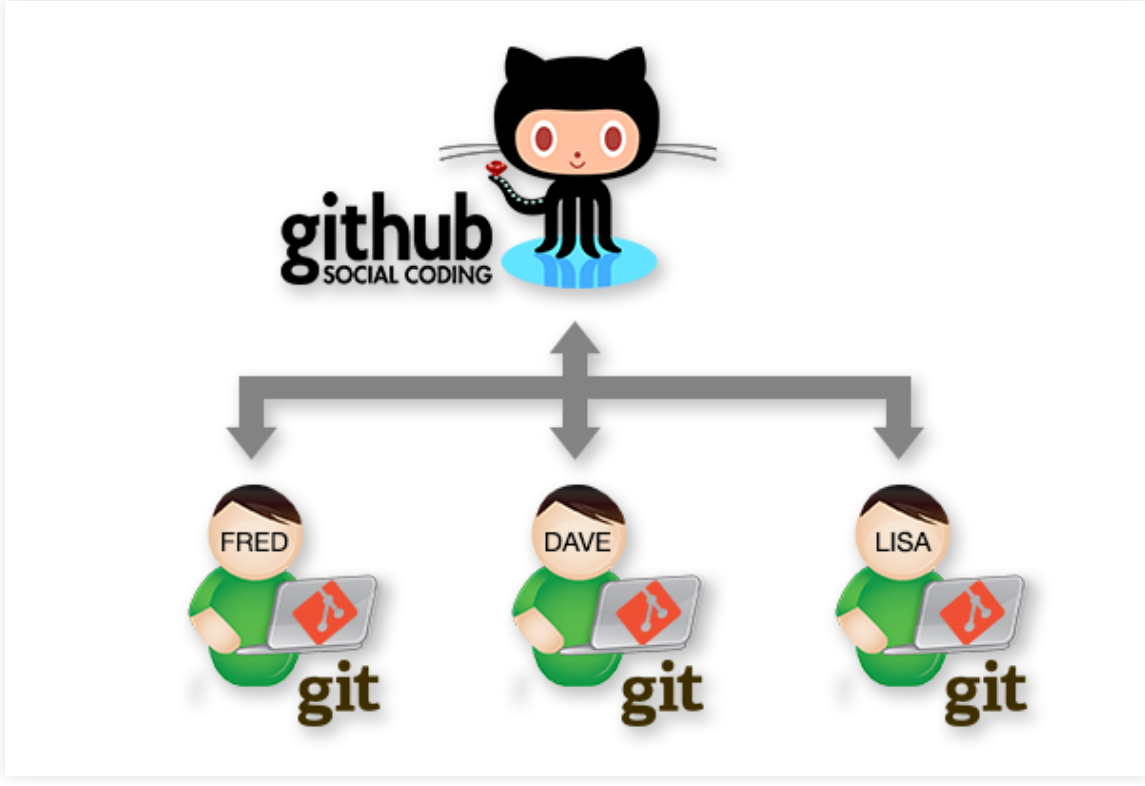


JUPYTER LAB OFFERS MULTIPLE WAYS TO USE PYTHON

A Python interpreter can always be started from the command line with `python` or `ipython`. Jupyter Lab provides more options:

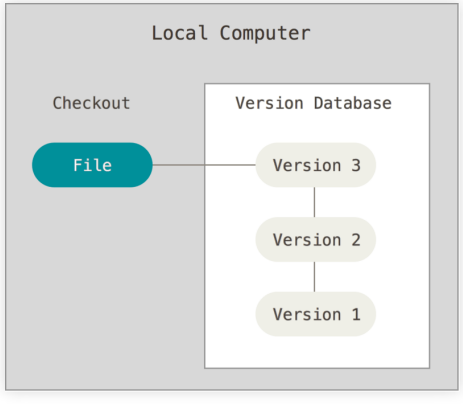
- Qt Console is an enhanced interpreter that allows for inline charting
- Notebook is a web-based application that allows inline charting, embedding of links and other resources, and the flexibility of web development
- Since a shell is available inside of JLab, you can also execute Python scripts that you have written in an editor tab.

TO CODE IS GOOD, TO REPLICATE DIVINE


github
SOCIAL CODING
FRED
DAVE
LISA
git
git
git

VERSION CONTROL IS NOT A SILVER BULLET...

...but it is a necessary component of reproducible research



It enables you to code without fear of mistake, because you can always go back to an earlier version

GITHUB WILL BE OUR COORDINATING MECHANISM

- You will need to create a Github account
- Github has a great collection of guides for common tasks
- A solid cheat sheet of common commands is available
- For more than you ever wanted to know, check out Pro Git

DATA FILES CAN COME IN A WIDE VARIETY OF FORMATS

THEY DO NOT ALWAYS LOOK LIKE TABLES!

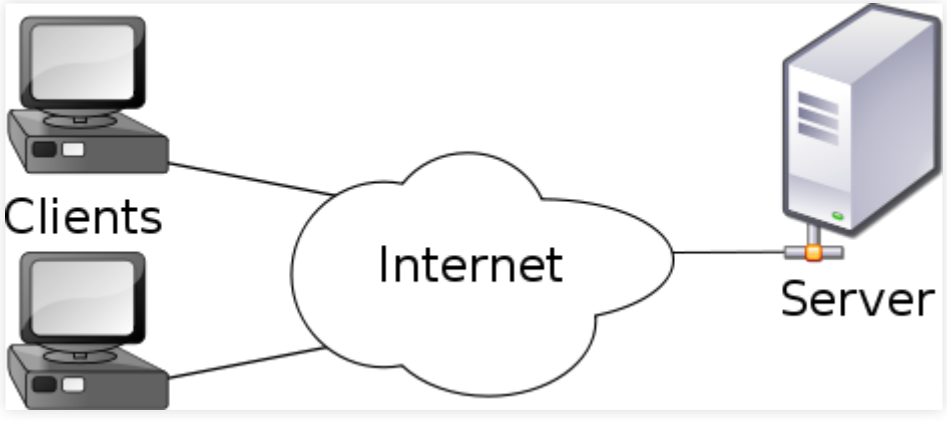
There is a big world out there beyond Excel:

- Tabular data (e.g. CSV, TSV, fixed width)
- Non-tabular data (e.g. JSON)
- Spatial data (e.g. SHP, GeoJSON)
- Relational database data (e.g. DB)

SOMETIMES DATA WILL NOT COME FROM LOCALLY STORED FILES

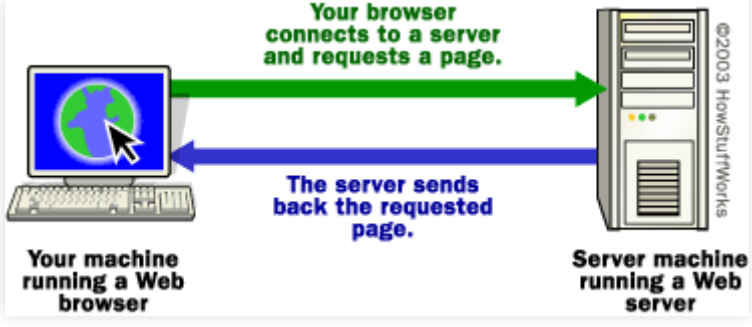
- We can simulate data via random draws from a statistical distribution
- We can hard code data
- We can access data from remote server

WHAT'S ALL THIS ABOUT REMOTE SERVERS??



ANY TIME YOU VISIT A WEBSITE, YOU ARE USING A REMOTE SERVER

A server is just another computer that you have connected to that holds the content you want.



It turns out data can be accessed in a very similar way with the right Uniform Resource Locator (URL)

ENTER DATA APPLICATION PROGRAMMING INTERFACES...

An API, in general, is just a clearly defined set of methods that allow software components to communicate with each other. In the data realm, APIs help our machine communicate with a remote server via URL, so we can retrieve data from it.