GENERAL ASSEMBLY

DATA SCIENCE TOOLS

Katherine Bueno

Data Analyst/ Data Scientist at AKQA

LET'S DISCUSS THE CURRENT LESSON OBEJCTIVES

- Identify the data science toolkit we'll be using in class
- Learn how to navigate Git and the Command Line
- Download the course Git repository and practice some git commands
- Make a probability and odds IPython notebook (time permitting)
- Slides will be available on the course repository for your reference.

INTRO

- Name
- Familiarity with programming, languages used
- Familiarity with UNIX/command line tools
- What you do for work/why you’re taking data science

Please post your github username in the slack channel so I can add you to our demo repository!

INTRODUCTION

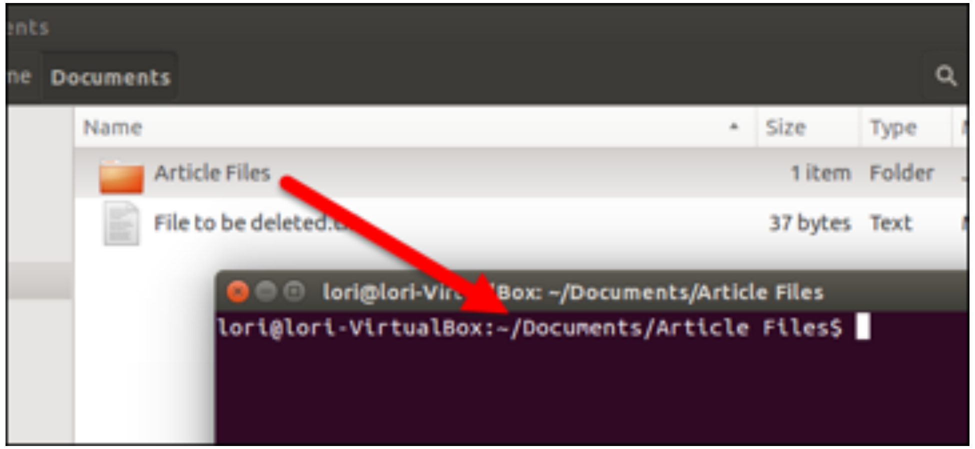
TOOLS OF THE TRADE

TOOLS OF THE TRADE

- Today we are going to review some of the tools we use in data science.
- We’ll see how they fit into the wider programming environment.
- We’ll start with the command line. This is your portal to your computer and the outside world.

LOCAL MACHINE

- On your local computer, you have a variety of tools at your disposal.
  - Text editor
  - Programs/tools
  - Your files

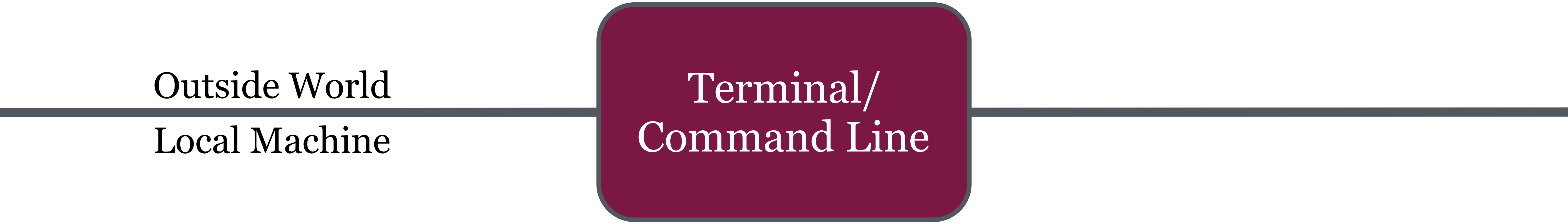


Name	Size	Type
Article Files	1 item	Folder
File to be deleted.	37 bytes	Text

```
lori@lori-VirtualBox:~/Documents/Article Files$
```

- All of these can be accessed through the terminal or through a GUI (Graphical User Interface).
- You can navigate your files through the terminal or through Finder.

DATA SCIENCE TOOLS



Outside World

Local Machine

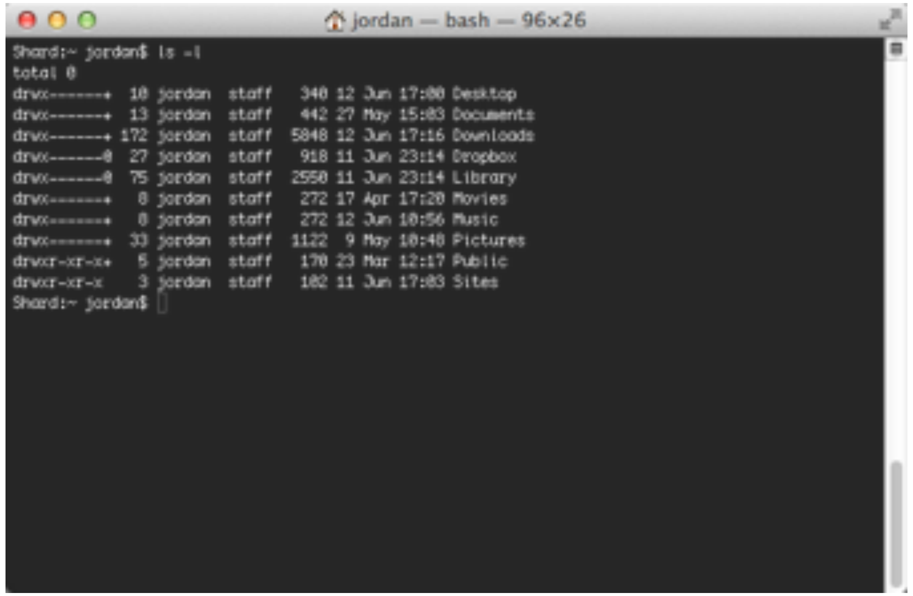
Terminal/
Command Line

DEMO

COMMAND LINE

COMMAND LINE

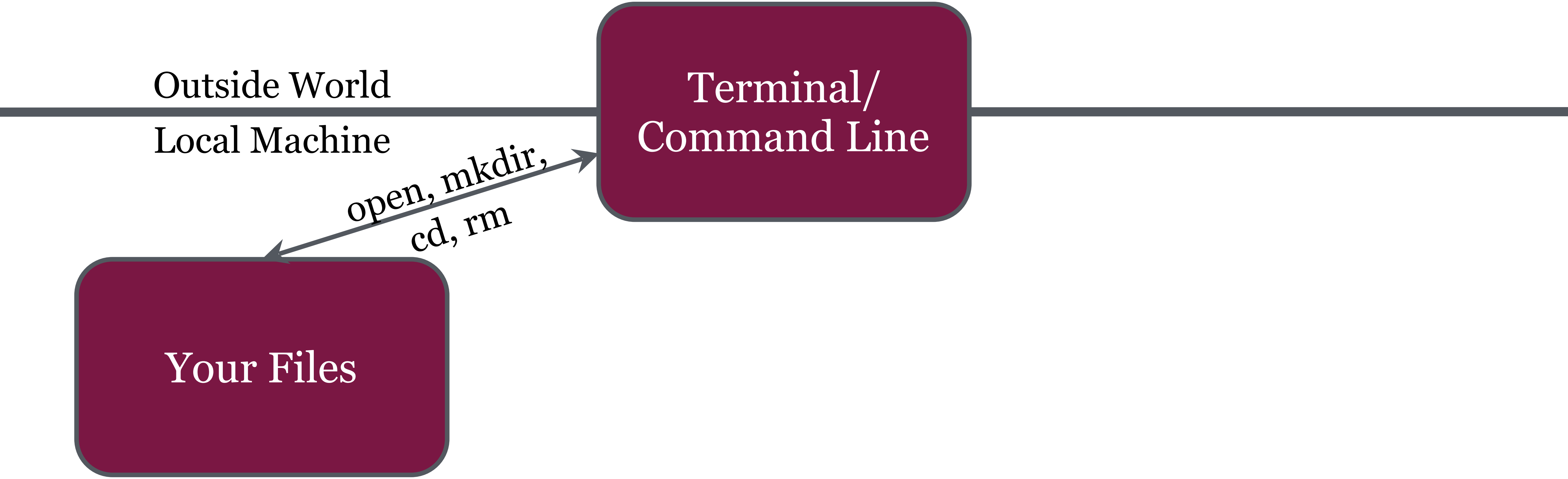
- Let’s walk through a few very basic commands.
  - `cd`
  - `pwd`
  - `$home`
  - `mkdir`
  - `open`



```
Shard:~ jordan$ ls -l
total 0
drwx------+  10 jordan  staff   340 12 Jun 17:00 Desktop
drwx------+  13 jordan  staff   442 27 May 15:03 Documents
drwx------+ 172 jordan  staff  5848 12 Jun 17:16 Downloads
drwx------@  27 jordan  staff   918 11 Jun 23:14 Dropbox
drwx------@  75 jordan  staff  2550 11 Jun 23:14 Library
drwx------+   8 jordan  staff   272 17 Apr 17:20 Movies
drwx------+   8 jordan  staff   272 12 Jun 10:56 Music
drwx------+  33 jordan  staff  1122  9 May 10:48 Pictures
drwxr-xr-x+   5 jordan  staff   170 23 Mar 12:17 Public
drwxr-xr-x    3 jordan  staff   102 11 Jun 17:03 Sites
Shard:~ jordan$
```

- We can access many tools with the terminal. Let’s walk through a few.

DATA SCIENCE TOOLS



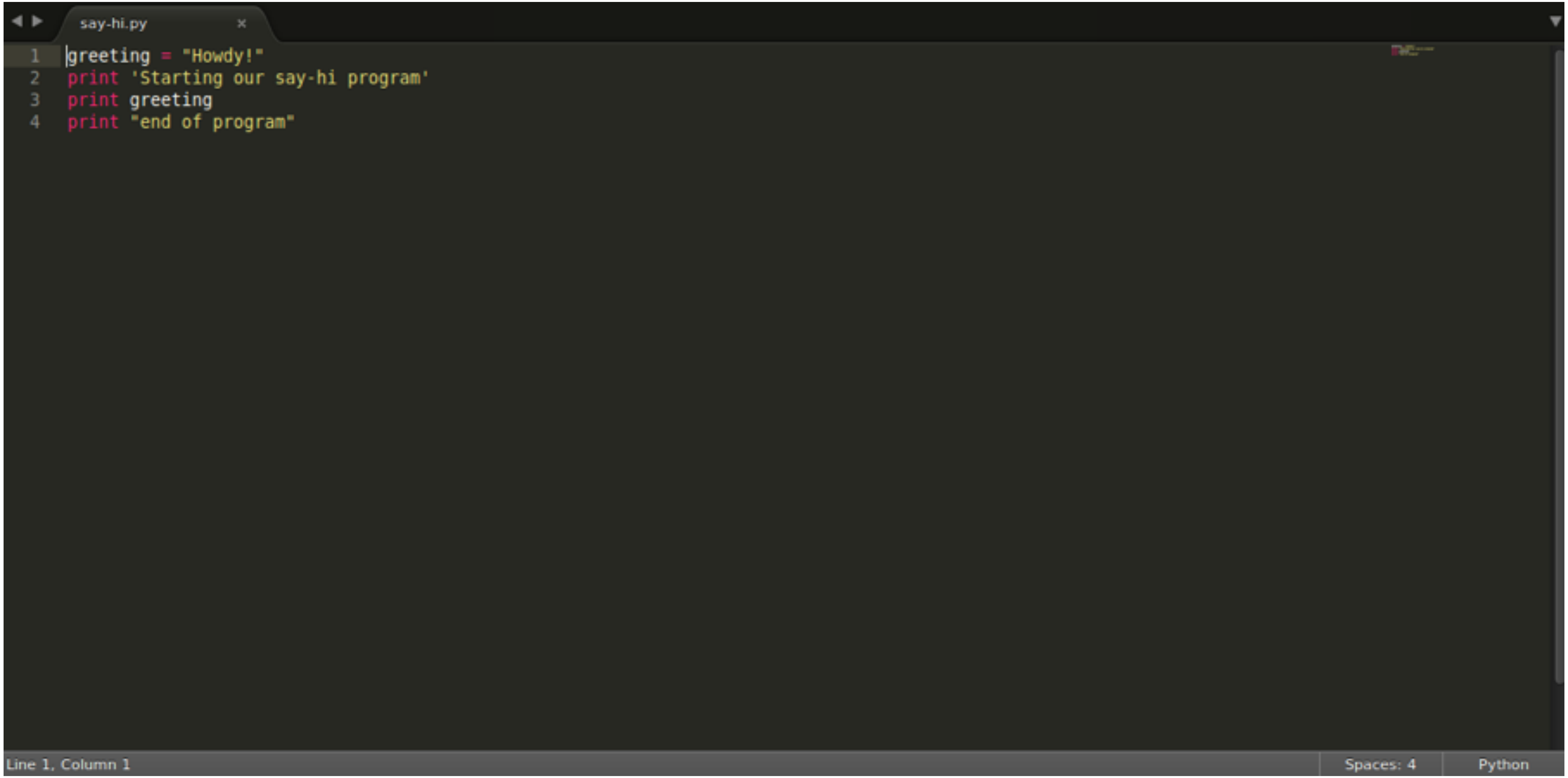
INTRODUCTION

TEXT EDITORS

TEXT EDITORS

- So far, we've used iPython Notebooks in place of a text editor.
- However, there are many options available
    - eMacs
    - Vim
    - Sublime Text
    - IDE (Integrated Development Environments) such as PyCharm or Eclipse
    - Many programmers love to argue at length about which is the 'best' one.
- Let's see what Sublime Text looks like with Python!

TEXT EDITORS



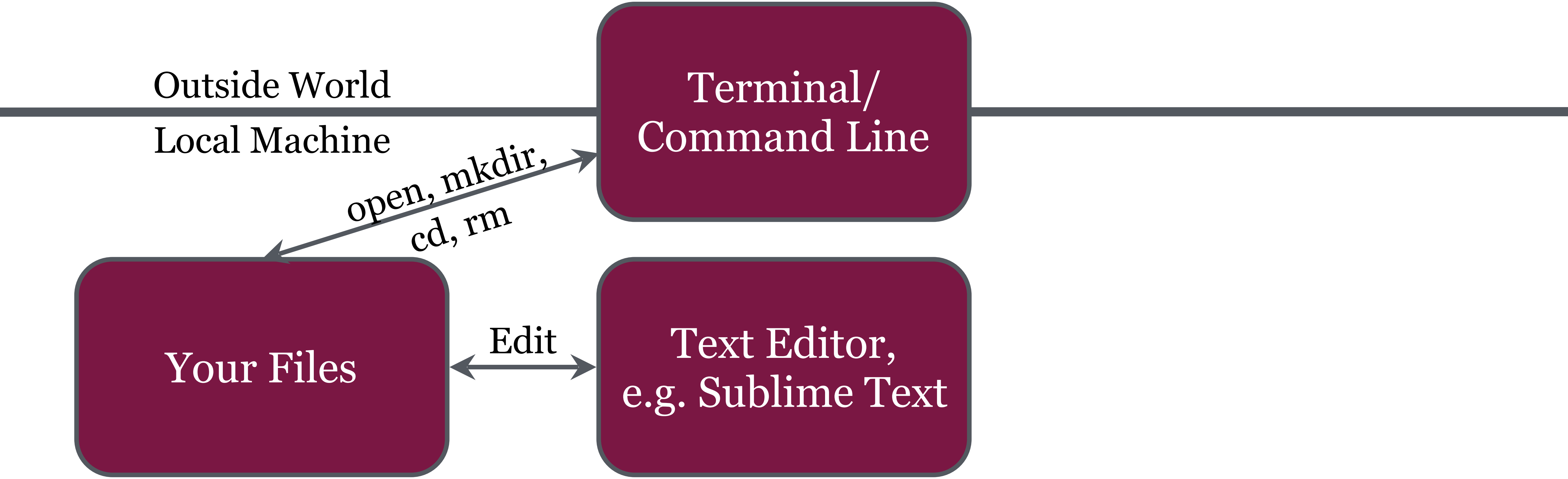
```
greeting = "Howdy!"
print 'Starting our say-hi program'
print greeting
print "end of program"
```

TEXT EDITORS

- Open “say-hi.py”, found in the lesson-02 folder of the class repo, in Sublime Text to see it for yourself.

- Check out the difference between saving a text file in MS Word and using sublime text

DATA SCIENCE TOOLS



Outside World

Local Machine

Terminal/ Command Line

open, mkdir, cd, rm

Your Files

Edit

Text Editor, e.g. Sublime Text

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. Who has a text editor they like that I didn’t talk about? Why do you like it?
2. What are some good reasons to use a programming text editor instead of Microsoft Word or Google Docs?

DELIVERABLE

Answers to the above questions

INTRODUCTION

IPYTHON NOTEBOOK

IPYTHON NOTEBOOK

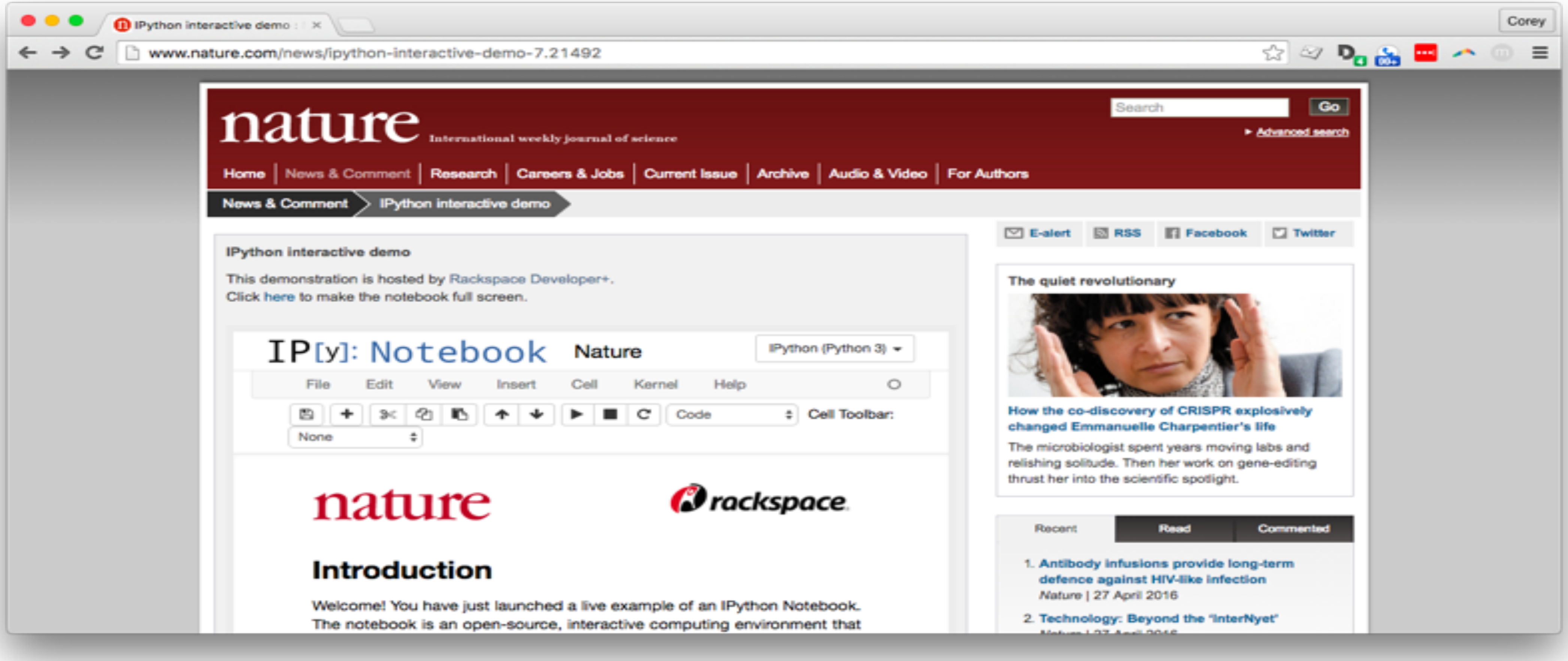
- Where does iPython Notebook fit in?

- We can refer to the iPython Notebook docs to get a better idea: the notebook combines the console, web apps, and markdown to capture the whole computation process.

- IPython notebooks combine three components in sequential ‘cells’ that can be edited and run in any order:
    - Text (Markdown/HTML/Plain text)
    - Code
    - Output (plots, command line output, etc)

IPYTHON NOTEBOOK

- IPython notebook demo:
  https://nature.tmpnb.org



IPython interactive demo

This demonstration is hosted by Rackspace Developer+.
Click here to make the notebook full screen.

Introduction

Welcome! You have just launched a live example of an IPython Notebook.
The notebook is an open-source, interactive computing environment that

The quiet revolutionary

How the co-discovery of CRISPR explosively changed Emmanuelle Charpentier's life

The microbiologist spent years moving labs and relishing solitude. Then her work on gene-editing thrust her into the scientific spotlight.

1. Antibody infusions provide long-term defence against HIV-like infection
   Nature | 27 April 2016
2. Technology: Beyond the 'InterNyet'

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What are the three components of IPython notebooks?

DELIVERABLE

Answers to the above questions

INTRODUCTION

PYTHON PACKAGES

PYTHON PACKAGES

- The terminal allows us to run programs and reach out to the outside world.

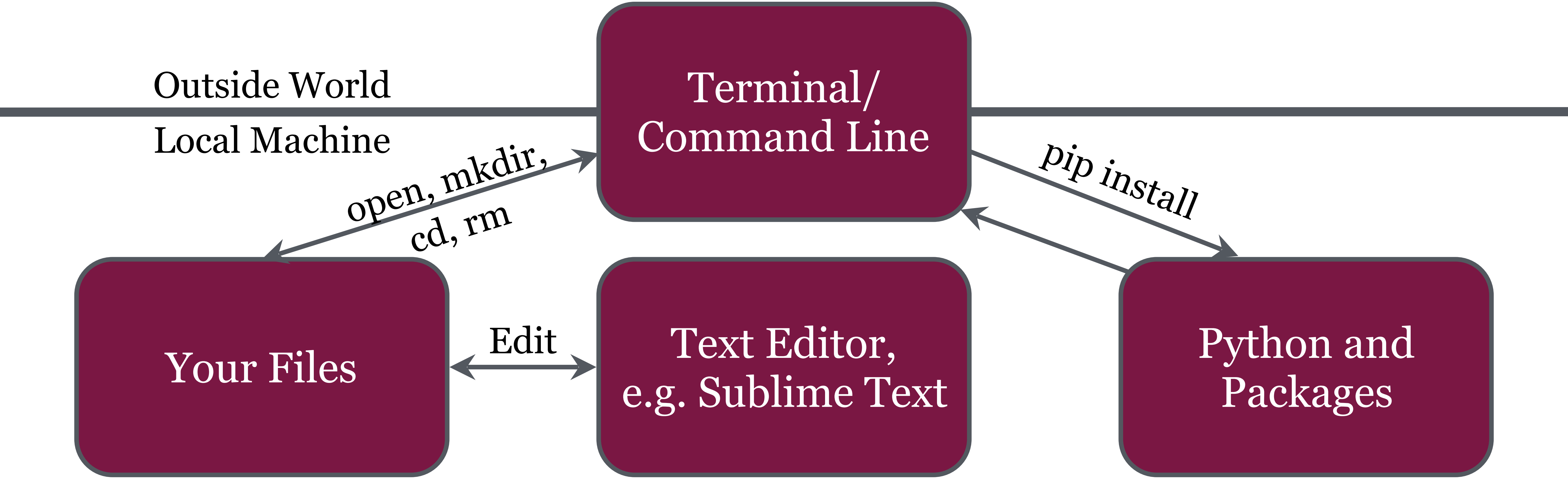
- We can add programs and packages as needed.

- To add Python packages, we use a tool called *pip*.

- Let's `pip install` a package with the command line. We'll install Beautiful Soup, a HTML/XML parsing package.

```
pip install beautifulsoup4
```

DATA SCIENCE TOOLS



Outside World

Local Machine

Terminal/ Command Line

open, mkdir, cd, rm

pip install

Your Files

Edit

Text Editor, e.g. Sublime Text

Python and Packages

INTRODUCTION

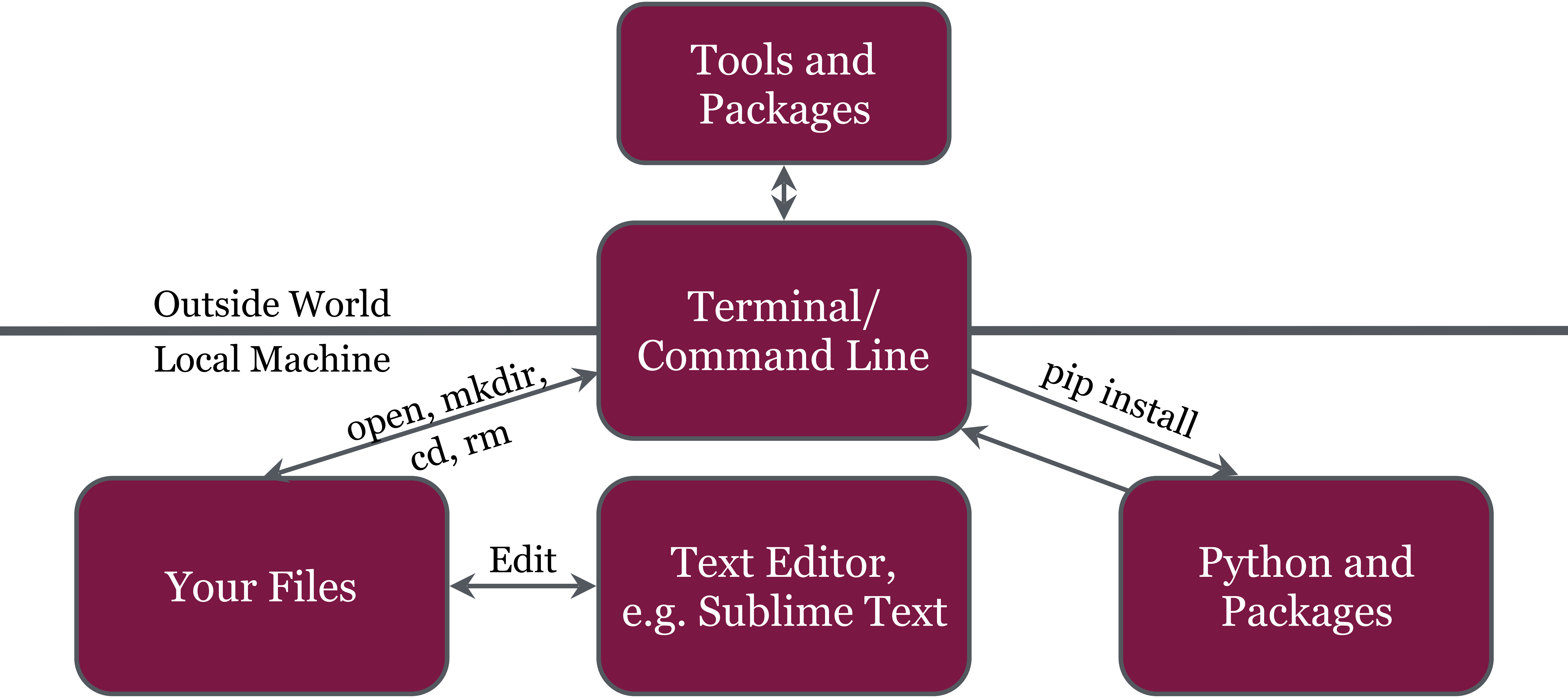
THE OUTSIDE WORLD

THE OUTSIDE WORLD

- The command line also allows you to download and use other tools and packages.

- There are many tools for different purposes available in the outside world.

DATA SCIENCE TOOLS



Tools and Packages

Terminal/ Command Line

Outside World

Local Machine

open, mkdir, cd, rm

pip install

Your Files

Edit

Text Editor, e.g. Sublime Text

Python and Packages

THE OUTSIDE WORLD

- As we saw with pip, the command line can connect us to the outside world. This becomes more important for data.

- We may have HIPAA protected data. This means we can’t leave this sensitive data on our *local* machine (i.e. laptop).

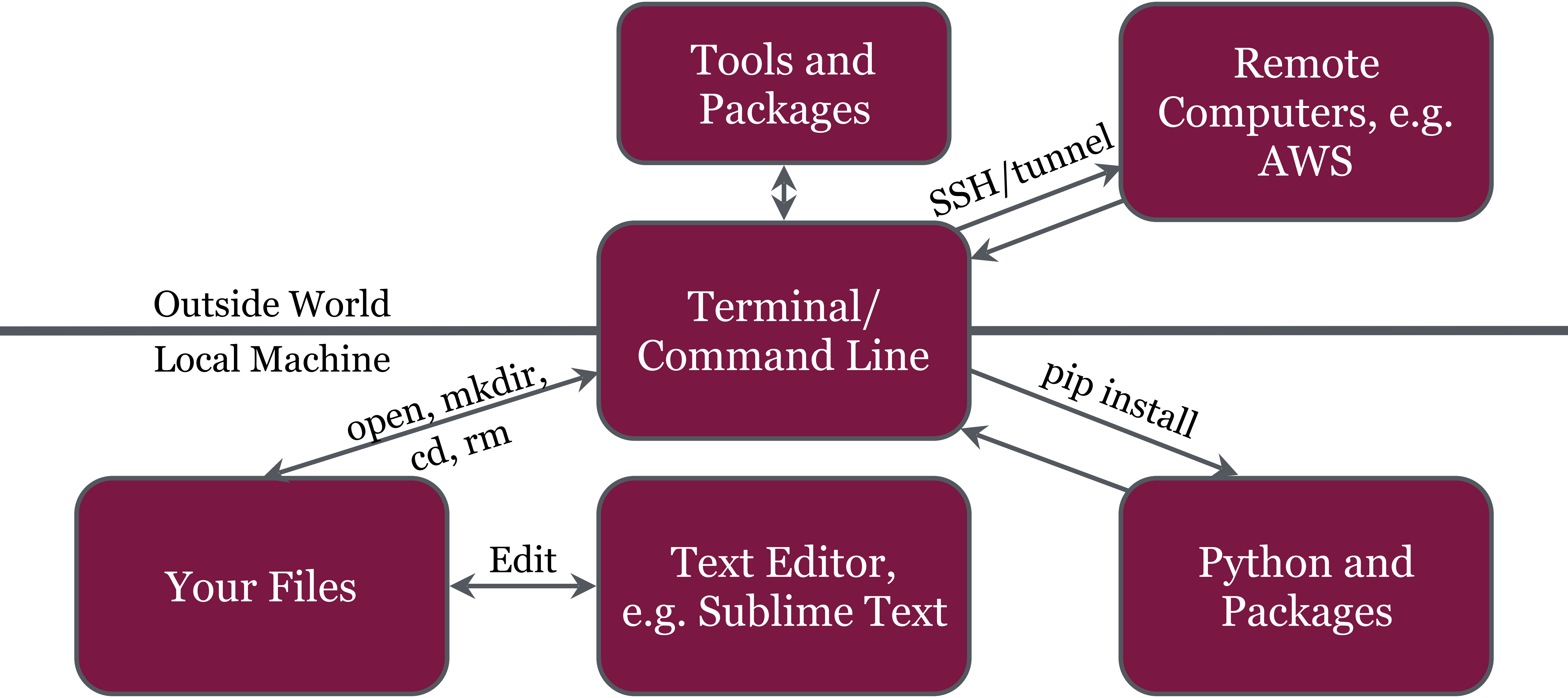
- We need to communicate with a *remote* machine (i.e. server) to access the data via command line.

- Let’s see a demonstration of this.

THE OUTSIDE WORLD

- AWS EC2 demo
  - SSH
  - Local database
  - Text files
- S3 Demo

DATA SCIENCE TOOLS



Tools and Packages

Remote Computers, e.g. AWS

SSH/tunnel

Outside World

Terminal/ Command Line

Local Machine

open, mkdir, cd, rm

pip install

Your Files

Edit

Text Editor, e.g. Sublime Text

Python and Packages

BREAK

10 mins

INTRODUCTION

GIT

GIT

- Version control is necessary when working on complex projects.

- Git is a way of tracking changes we’ve made to our programs that allows us to go back in time to fix errors.

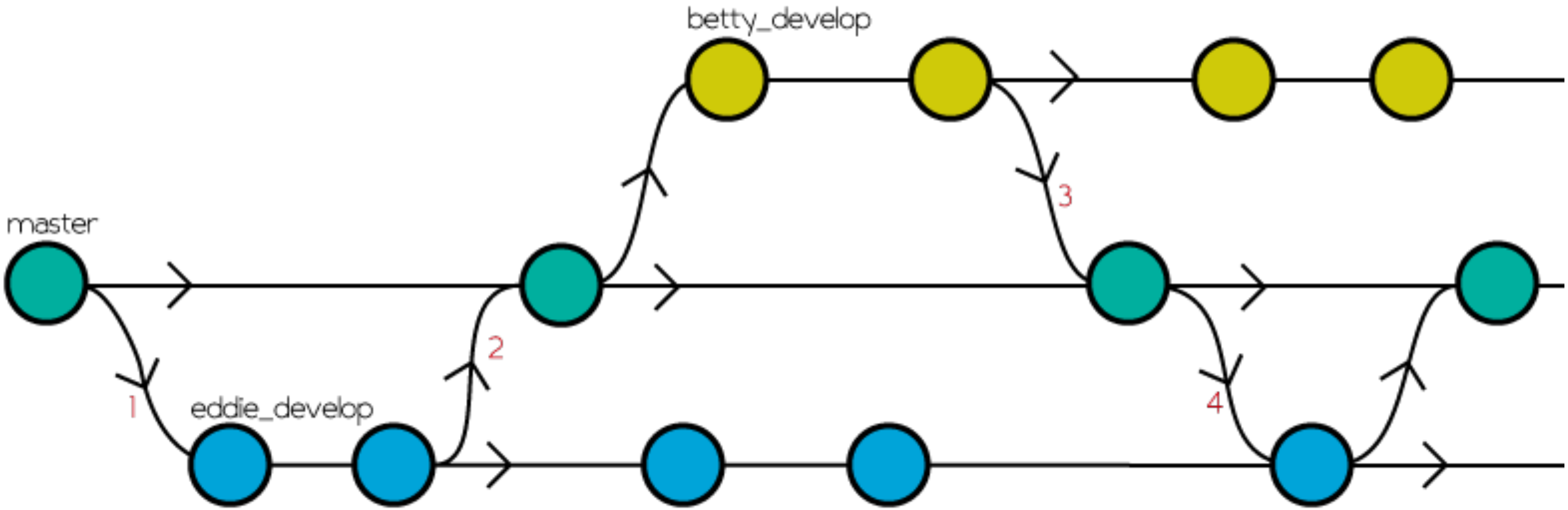
- Combined with Github, Git is a powerful tool for collaborating with colleagues. You can work on different aspects of projects simultaneously and merge the changes together seamlessly.

- There are many different ways to use these tools.

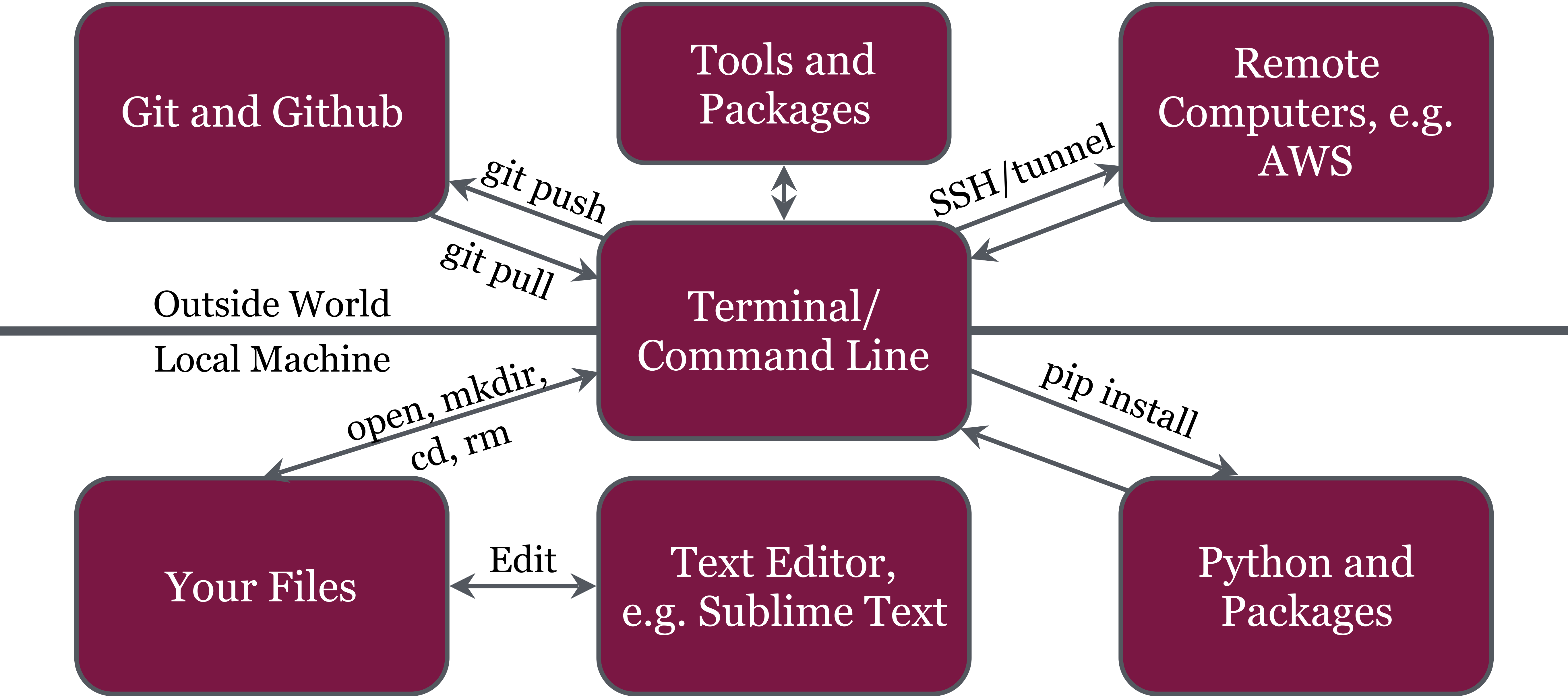
GIT

- Let's see an example of using Git and Github.
- There are three primary commands we'll use.
    - `git add`
    - `git commit`
    - `git push`
- When a colleague wants to implement our change, we may use the command `git pull`.

GIT



DATA SCIENCE TOOLS



Git and Github

Tools and Packages

Remote Computers, e.g. AWS

git push

git pull

SSH/tunnel

Outside World

Local Machine

Terminal/ Command Line

open, mkdir, cd, rm

pip install

Your Files

Edit

Text Editor, e.g. Sublime Text

Python and Packages

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. How are Python software modules shared?
2. What tools do you use to find and get new Python software?

DELIVERABLE

Answers to the above questions

GUIDED PRACTICE

GIT AND COMMAND LINE

ACTIVITY: GIT AND COMMAND LINE

EXERCISE

DIRECTIONS (30 minutes)

1. https://guides.github.com/activities/hello-world/
2. https://try.github.io/
3. Don't forget to paste your github username in the slack channel if you haven't done so already

DELIVERABLE

Github username

LESSON

Q & A

BREAK

10 mins

BRINGING IT ALL TOGETHER

GIT EXERCISE

GIT

- **Goal:** to fork and clone our own repositories on github, make sure we can edit a trivial file, and push it back up.
- **Secondary goal:** become comfortable wrestling the Git monster.

When the office git expert has to come fix everything

source: http://wheningit.tumblr.com/post...

BRINGING IT ALL TOGETHER

EXERCISE

Activity (20 mins)

1. Break into groups of 4
2. Fork and clone the class repository at https://github.com/DAT-44-...
3. Create a text file in students/firstnamelastname.txt with your first name, last name, favorite food, and first place you would go if someone gave you a free plane ticket anywhere in the world.
4. ‘Git add’ the text file to your local repository
5. ‘Git commit -m “type a commit message here”’
6. ‘Git push origin’

DELIVERABLE

A link to your text file on github in the slack channel

GUIDED PRACTICE

ODDS AND PROBABILITY

ACTIVITY: ODDS & PROBABILITY

EXERCISE

DIRECTIONS (20 minutes)

Some of you may already be familiar with odds and probability.

1. We will use the starter code in lesson-02 of the class repo to review the concepts of odds and probability.

DELIVERABLE

Answer the questions in the notebook

CONCLUSION

TOPIC REVIEW

REVIEW

- What are some common data science tools?
- Why are these tools useful?
- Any other questions?

BEFORE NEXT CLASS:

Practice Git, review today’s lesson, and schedule office hours with me, if needed.

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

=Will send link in slack