

DATA SCIENCE TOOLS

Adam Jones, PhD

Data Scientist @ Critical Juncture

DATA SCIENCE TOOLS

LEARNING OBJECTIVES

- Identify the data science toolkit
- Navigate Git and the Command Line
- Describe Probability vs Odds

COURSE

PRE-WORK

PRE-WORK REVIEW

- Explain the difference between *variance* and *bias*
- Use **descriptive stats** to understand your data

OPENING

DATA SCIENCE TOOLS

LET'S DISCUSS THE CURRENT LESSON OBEJCTIVES

- Identify the data science toolkit
- Navigate Git and the Command Line
- Describe Probability vs. Odds

INTRODUCTION

TOOLS OF THE TRADE

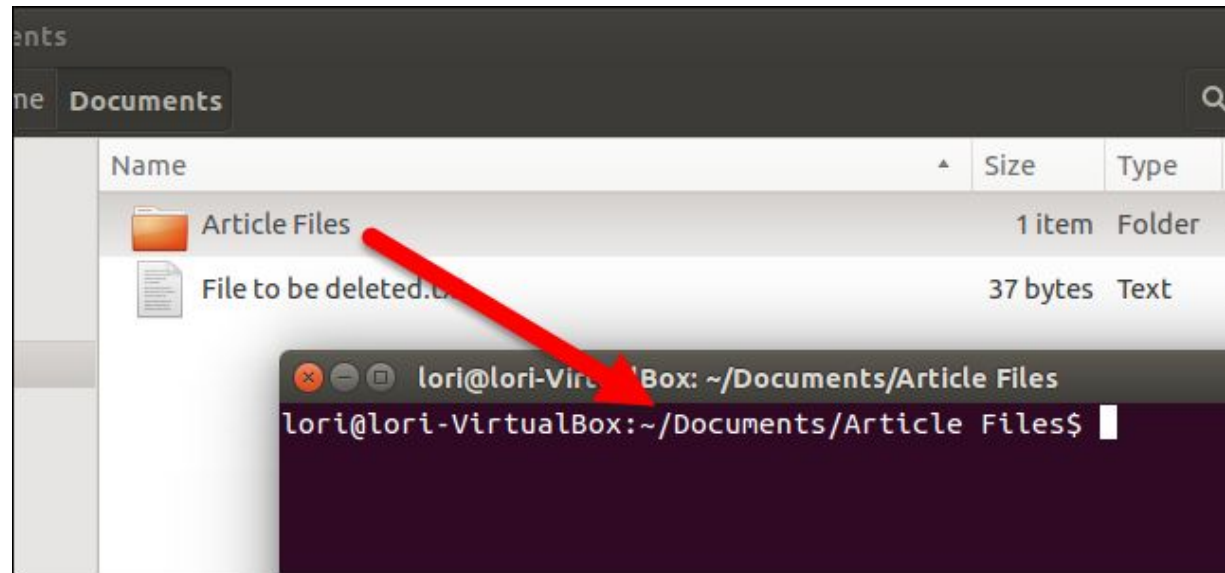
TOOLS OF THE TRADE

- 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



- ▶ 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

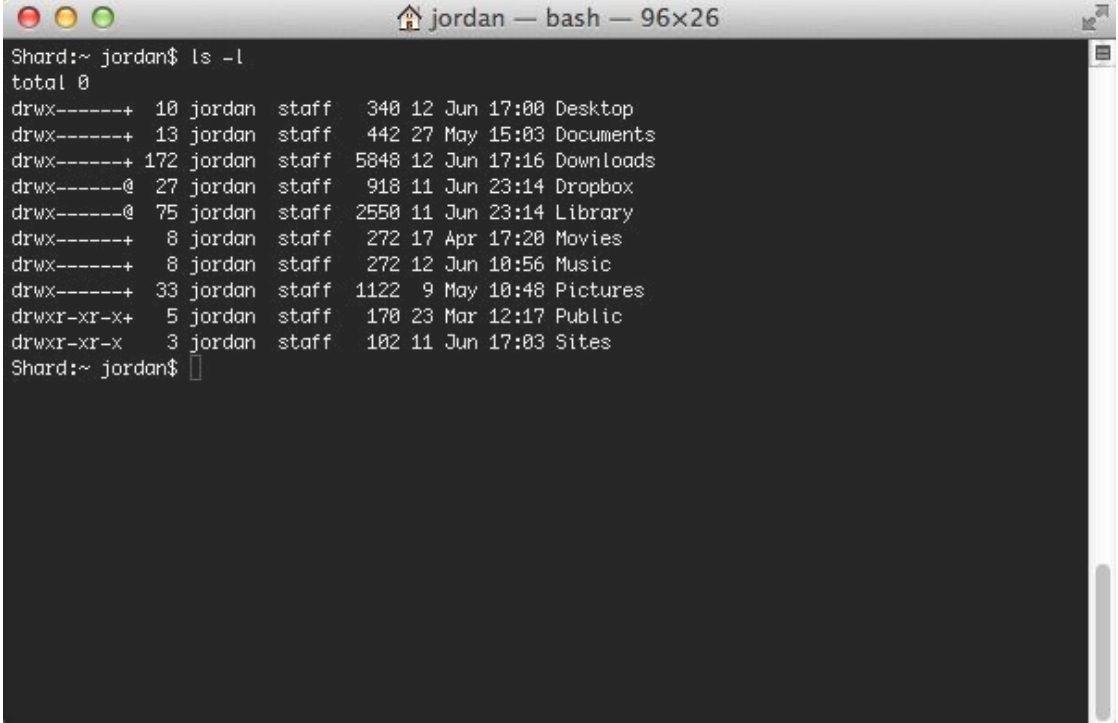
COMMAND LINE

▸ Let's walk through a few commands

- cd
- pwd
- \$home
- mkdir
- open

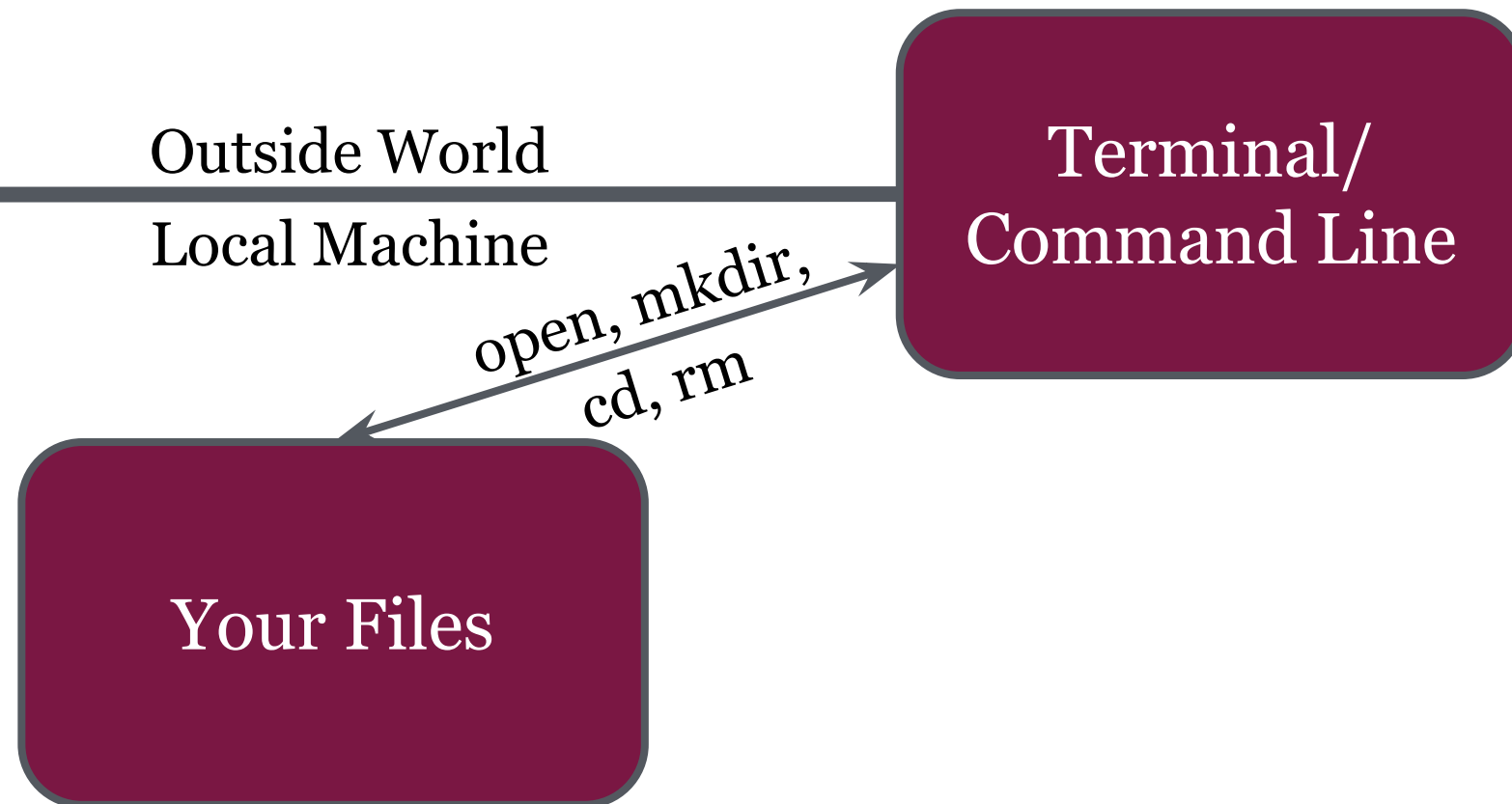
We can access many tools with the terminal

- Let's walk through a few...



```
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$
```

DATA SCIENCE TOOLS

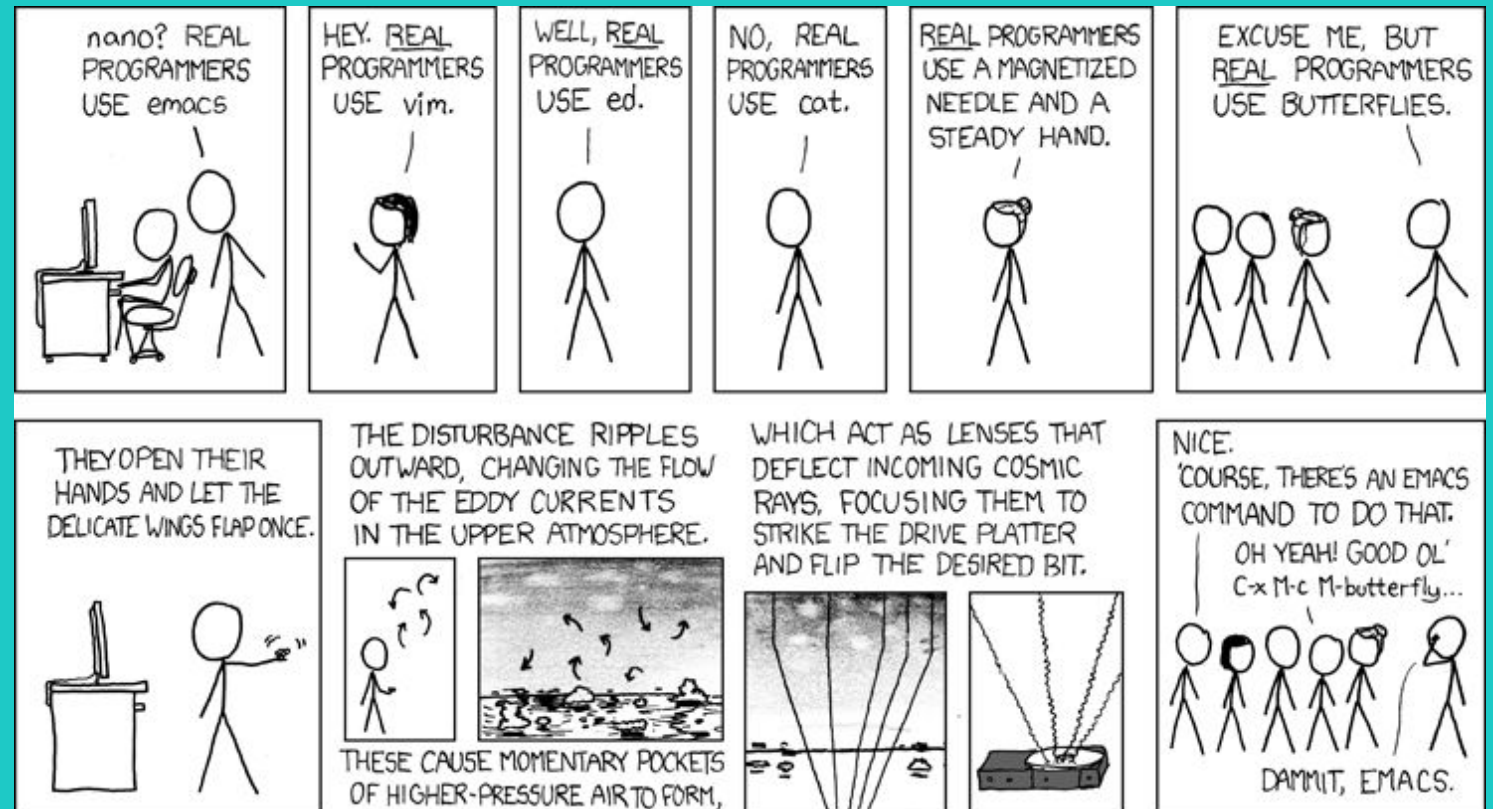


INTRODUCTION

TEXT EDITORS

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

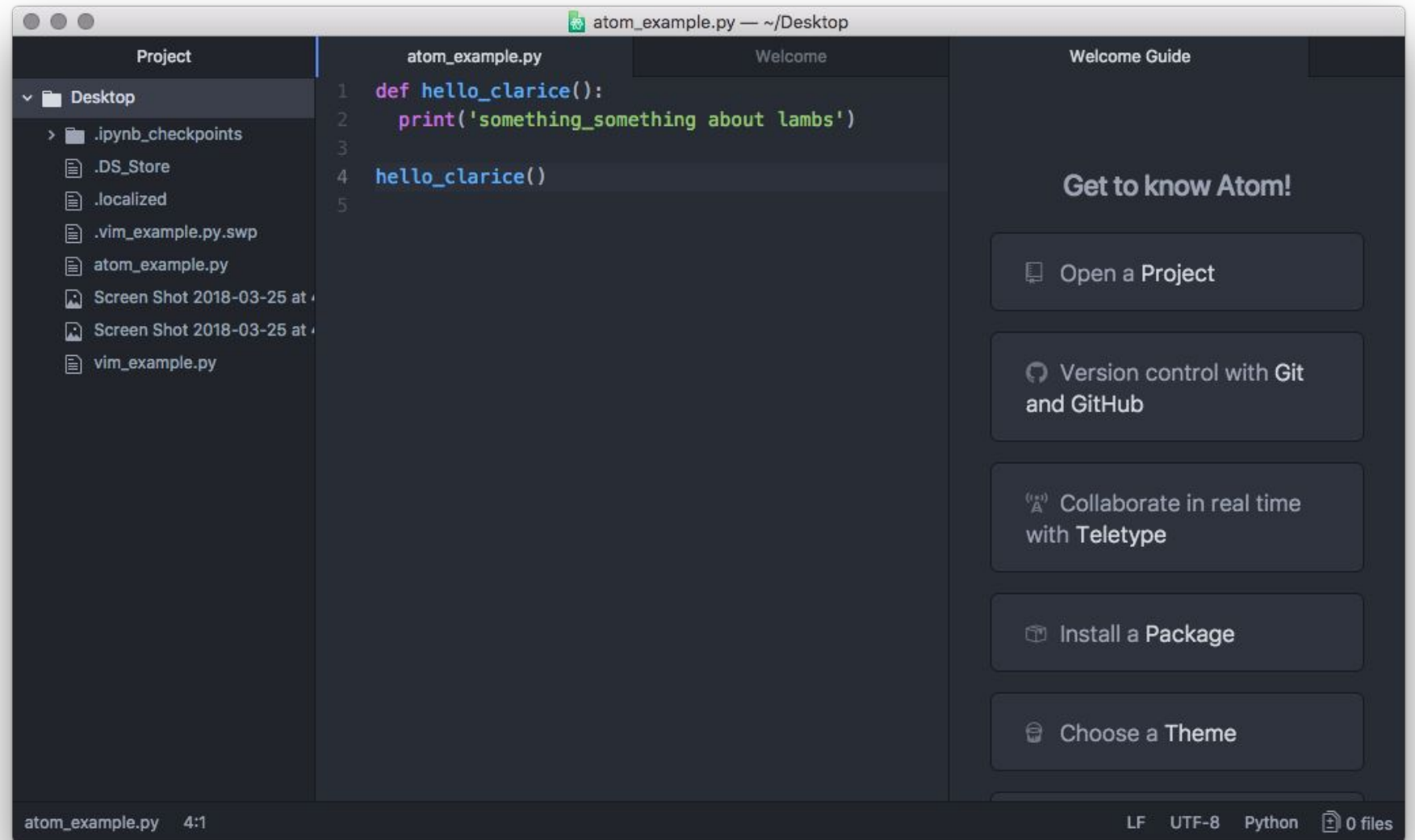
- Vim

- Sublime Text



- Let's look at a few examples...

TEXT EDITORS



TEXT EDITORS



```
Desktop — vim vim_example.py — 80x24
def hello_clarice():
    print('something_something about lambs')

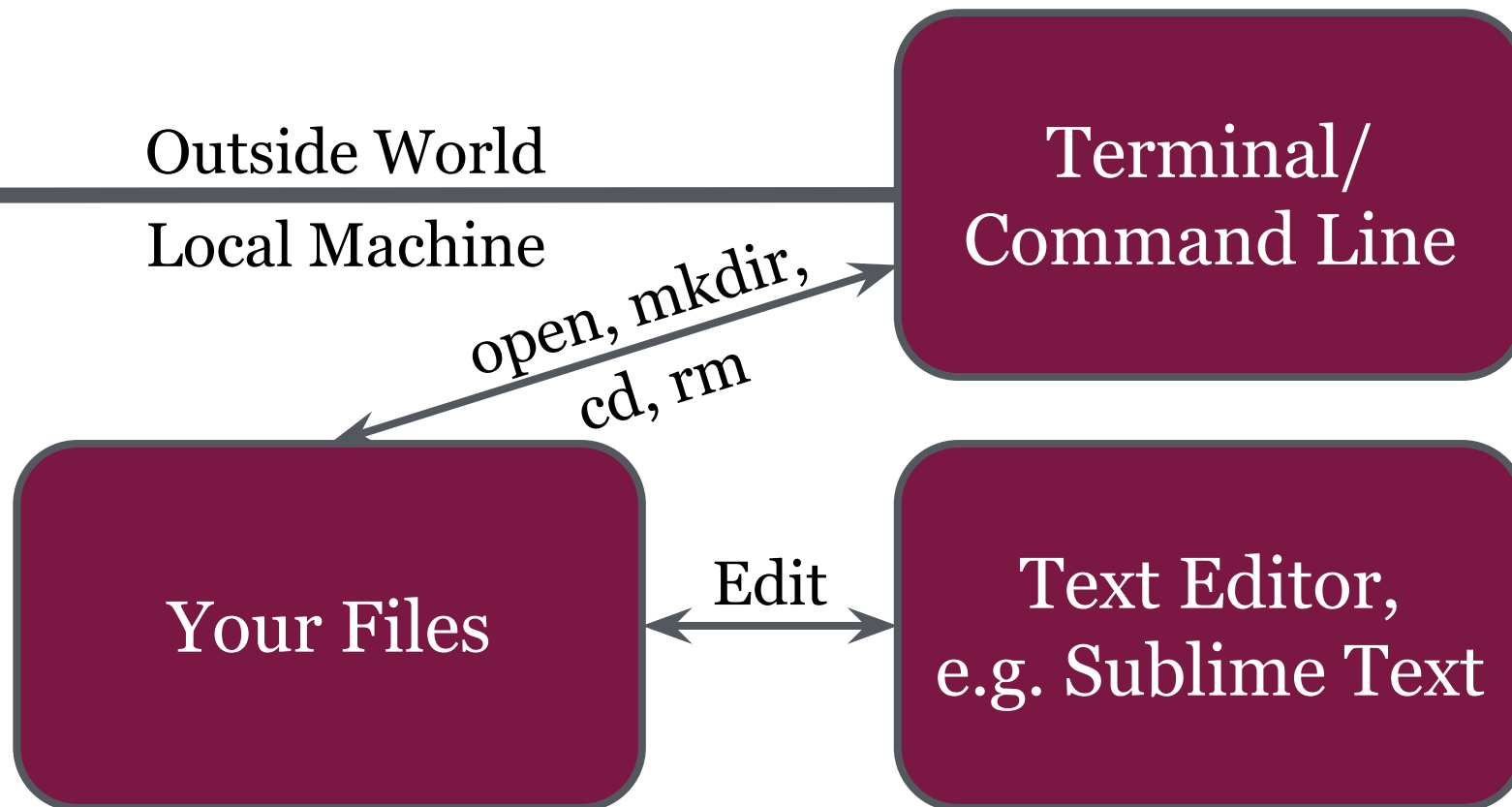
hello_clarice()

"vim_example.py" 4L, 81C
```

TEXT EDITORS

- Open “say-hi.py”, in the lesson-05 folder of the class repo, in a text editor

DATA SCIENCE TOOLS



ACTIVITY: KNOWLEDGE CHECK



EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is a text editor?
2. Can you name any other examples?

DELIVERABLE

Answers to the above questions

INTRODUCTION

JUPYTER NOTEBOOK

JUPYTER NOTEBOOK

- Where does Jupyter Notebook fit in?
- The notebook combines the console, web apps, and markdown to capture the whole computation process
- Jupyter/iPython notebooks combine two components:
 1. A web application
 2. Notebook documents

JUPYTER NOTEBOOK

- Where does Jupyter Notebook fit in?
- The notebook combines the console, web apps, and markdown to capture the whole computation process
- Jupyter/iPython notebooks combine two components:
 - A web application
 - Notebook documents

INTRODUCTION

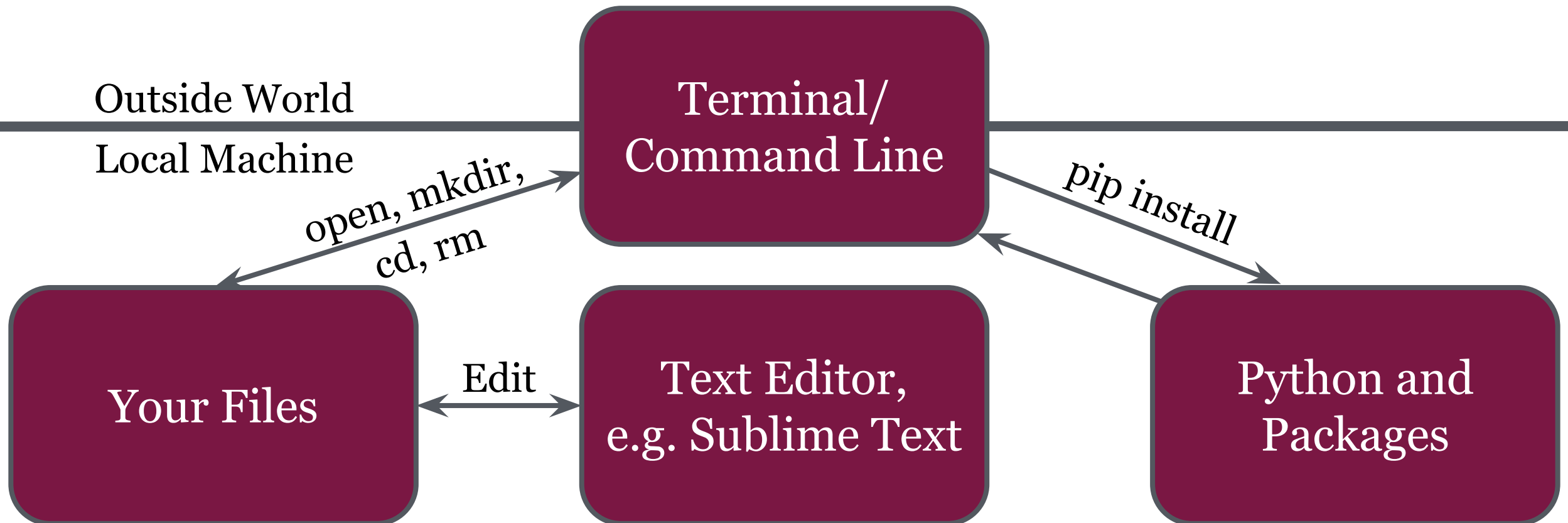
PYTHON PACKAGES

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

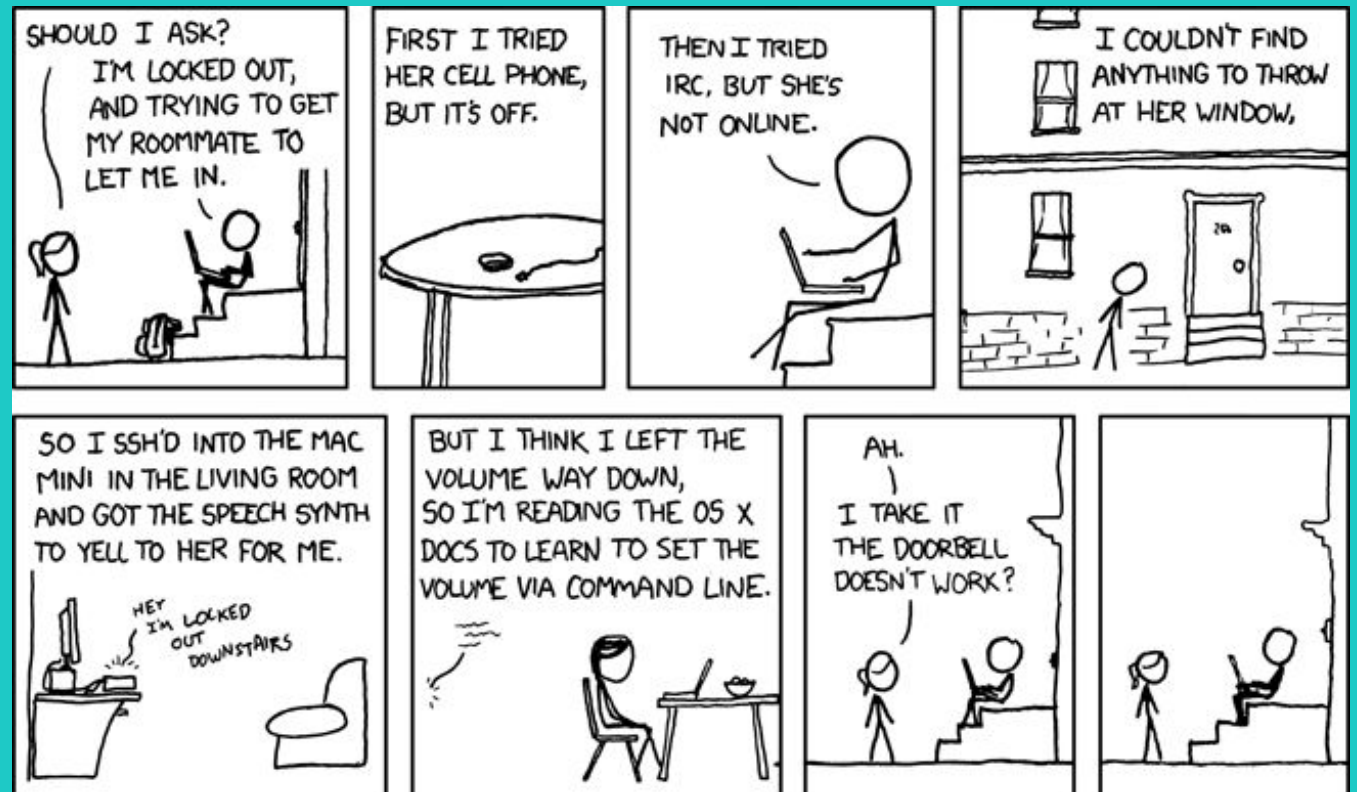


INTRODUCTION

THE OUTSIDE WORLD

INTRODUCTION

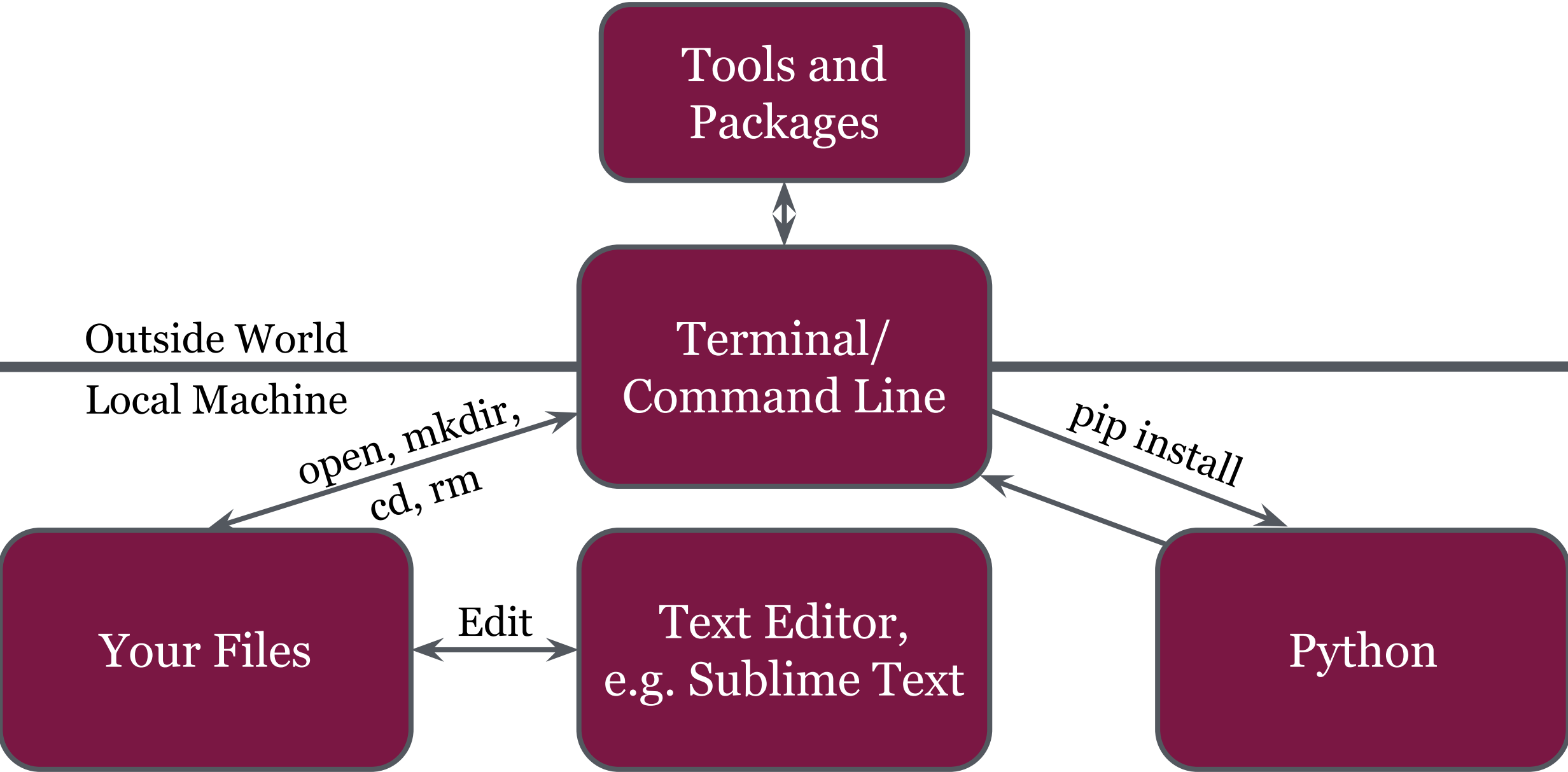
THE OUTSIDE WORLD



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

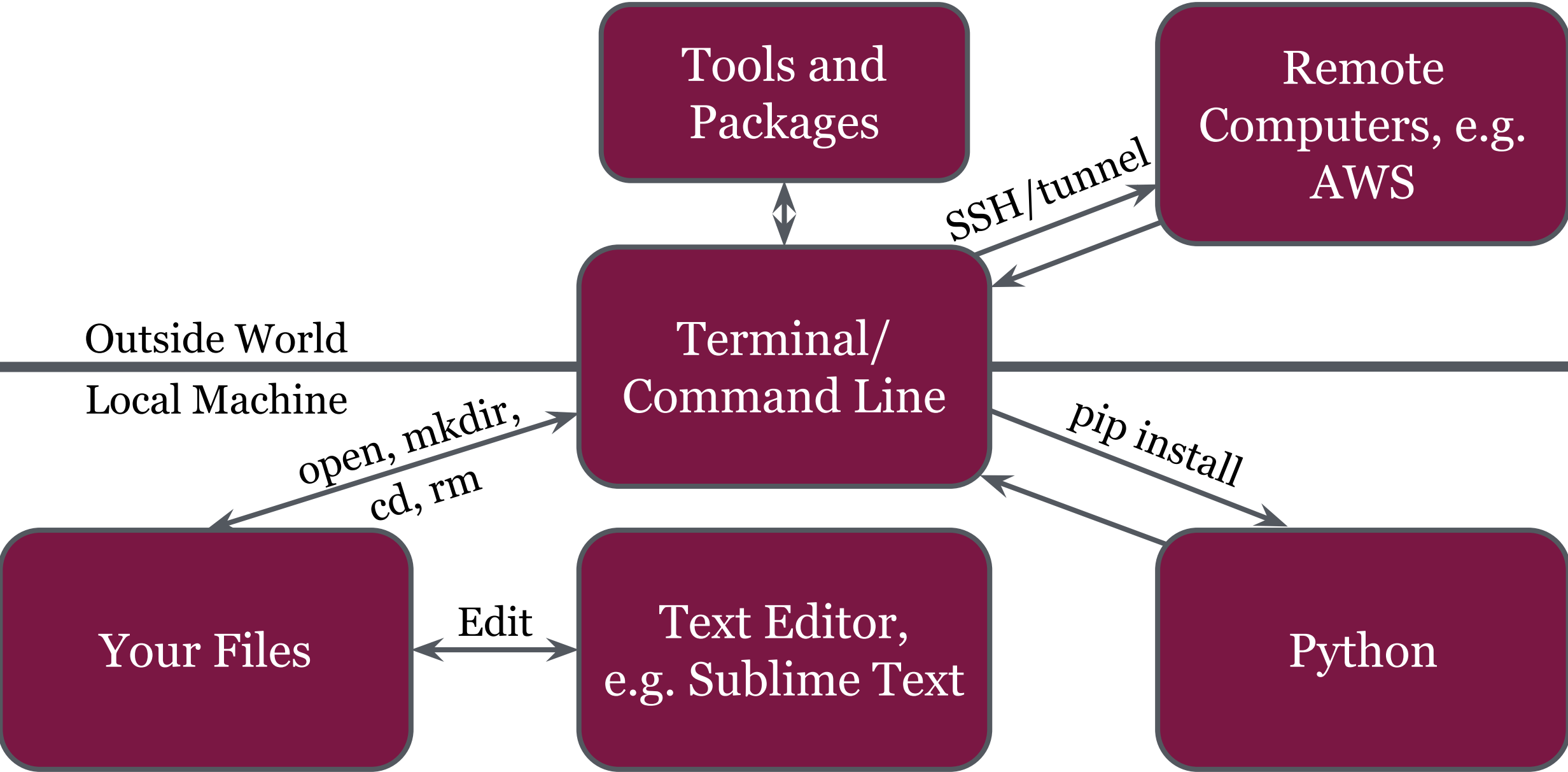
DATA SCIENCE TOOLS



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 demo!

DATA SCIENCE TOOLS



INTRODUCTION

GIT

INTRODUCTION

GIT

	COMMENT	DATE
○	CREATED MAIN LOOP & TIMING CONTROL	14 HOURS AGO
○	ENABLED CONFIG FILE PARSING	9 HOURS AGO
○	MISC BUGFIXES	5 HOURS AGO
○	CODE ADDITIONS/EDITS	4 HOURS AGO
○	MORE CODE	4 HOURS AGO
○	HERE HAVE CODE	4 HOURS AGO
○	AAAAAAAAA	3 HOURS AGO
○	ADKFJSLKDFJSDKLFJ	3 HOURS AGO
○	MY HANDS ARE TYPING WORDS	2 HOURS AGO
○	HAAAAAAAAAANDS	2 HOURS AGO

AS A PROJECT DRAGS ON, MY GIT COMMIT MESSAGES GET LESS AND LESS INFORMATIVE.

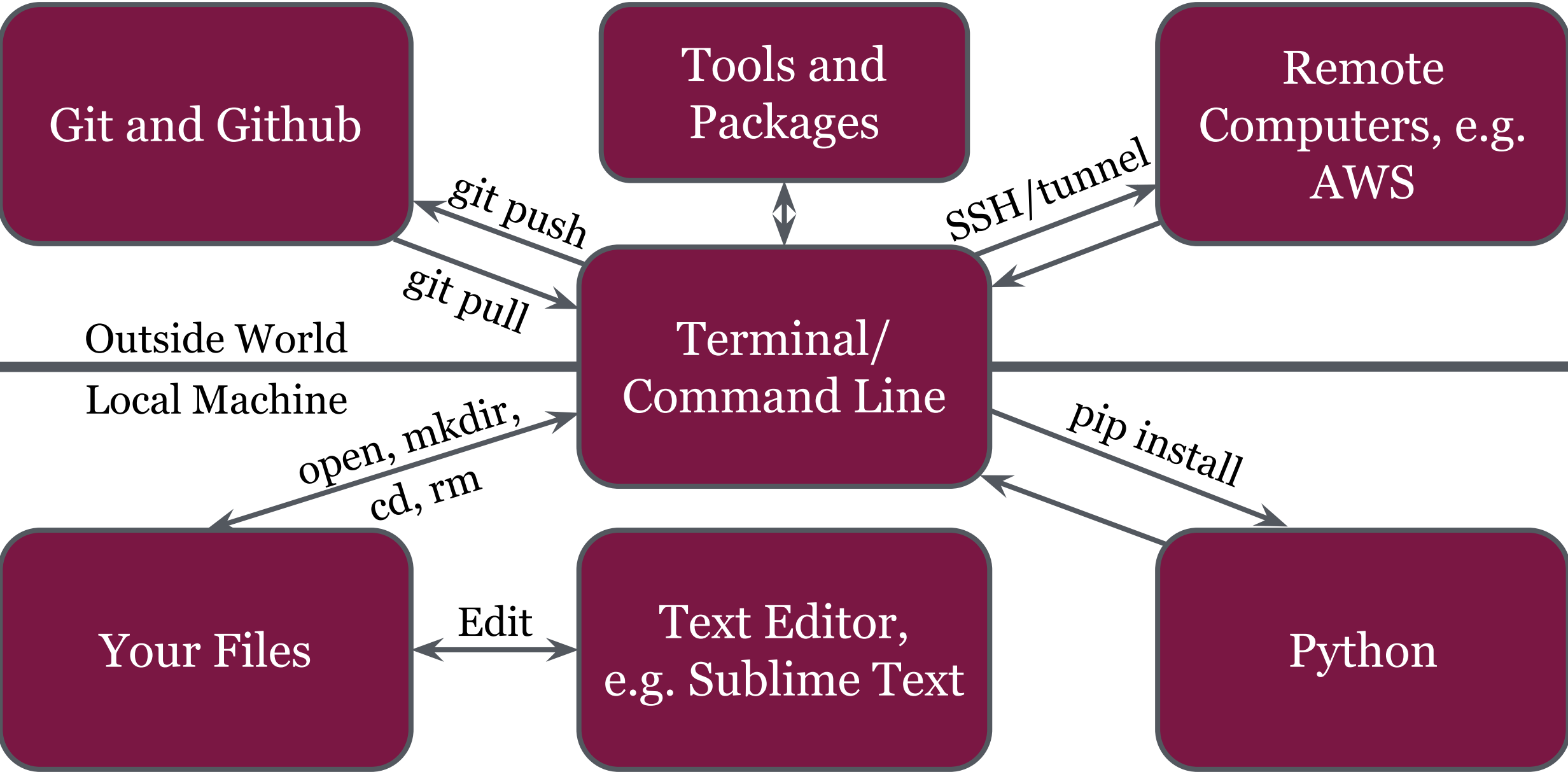
GIT

- 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`

DATA SCIENCE TOOLS



ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS



EXERCISE

1. What is a GUI?
2. What is the command line?
3. What are the big advantages of using the command line over a GUI?

DELIVERABLE

Answers to the above questions

GUIDED PRACTICE

GIT AND COMMAND LINE

ACTIVITY: GIT AND COMMAND LINE



EXERCISE

DIRECTIONS (20 minutes)

1. Review the exercises from **try.github.io**
2. Are there any questions?

DELIVERABLE

Questions

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-05 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?

COURSE

UPCOMING WORK

BEFORE NEXT CLASS

DUE DATE

- Final Project Pt. 1: Thurs (4/12)

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET