GENERAL ASSEMBLY

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

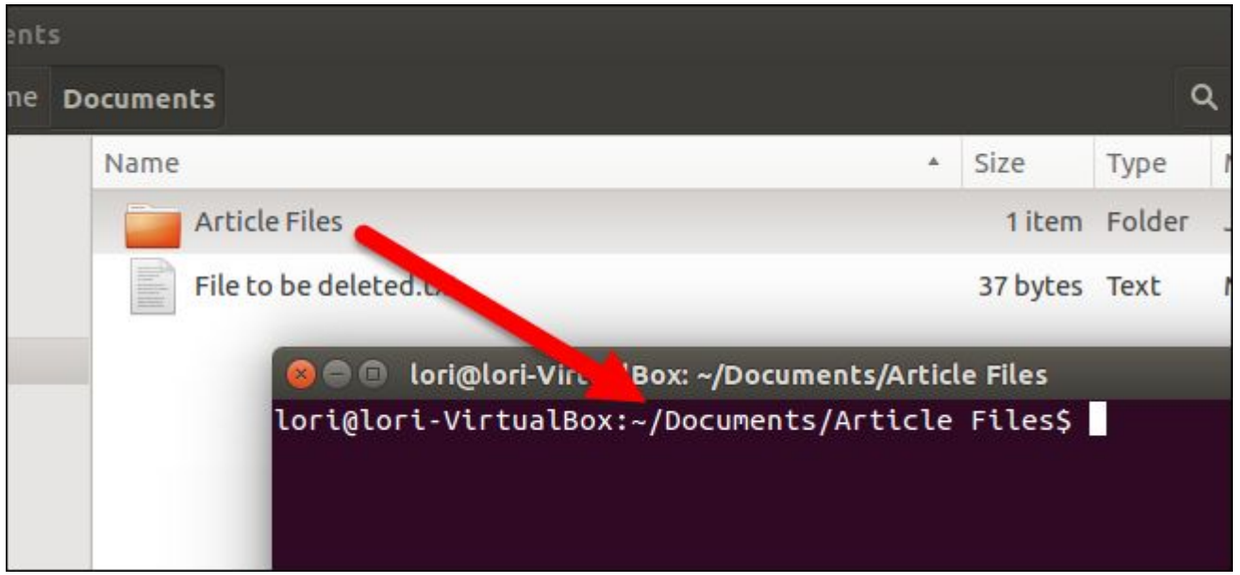
TOOLS OF THE TRADE

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

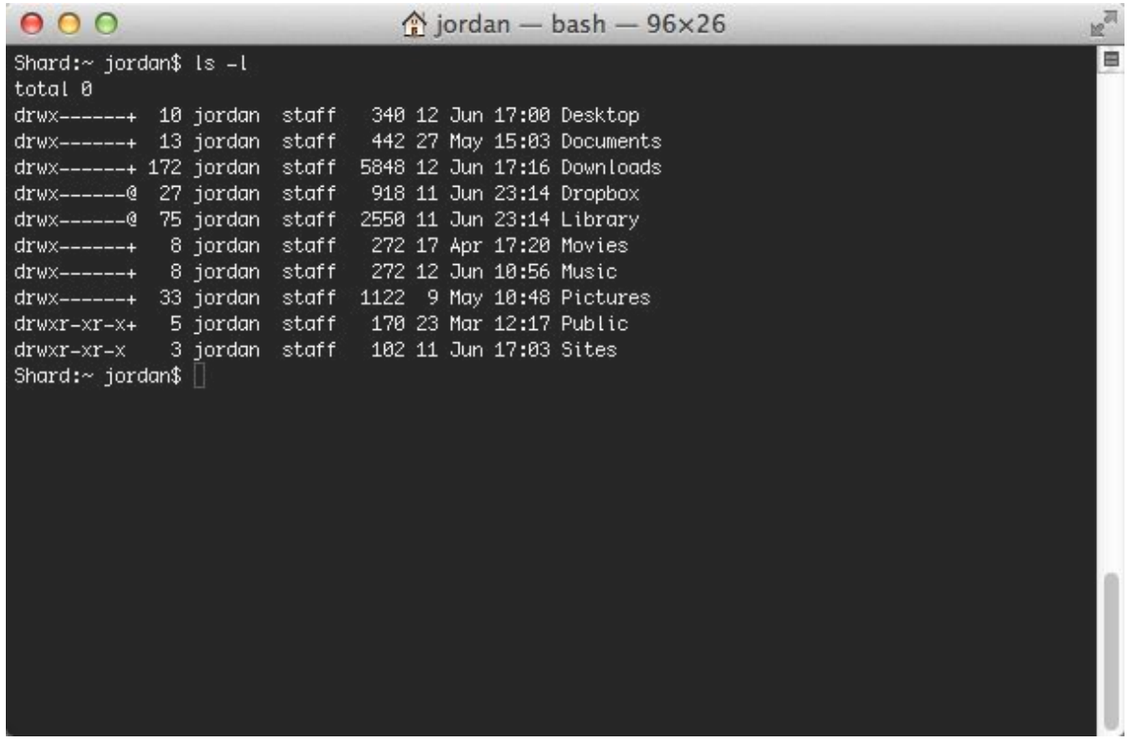
Terminal/
Command Line

DEMO

COMMAND LINE

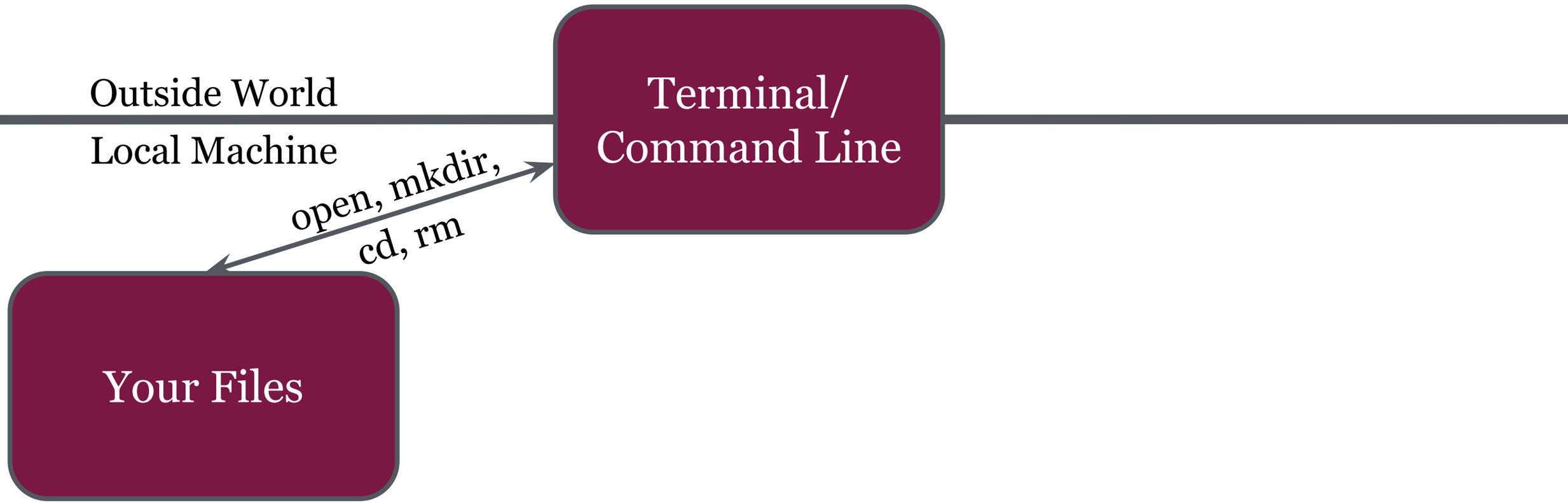
COMMAND LINE

- 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.

DATA SCIENCE TOOLS



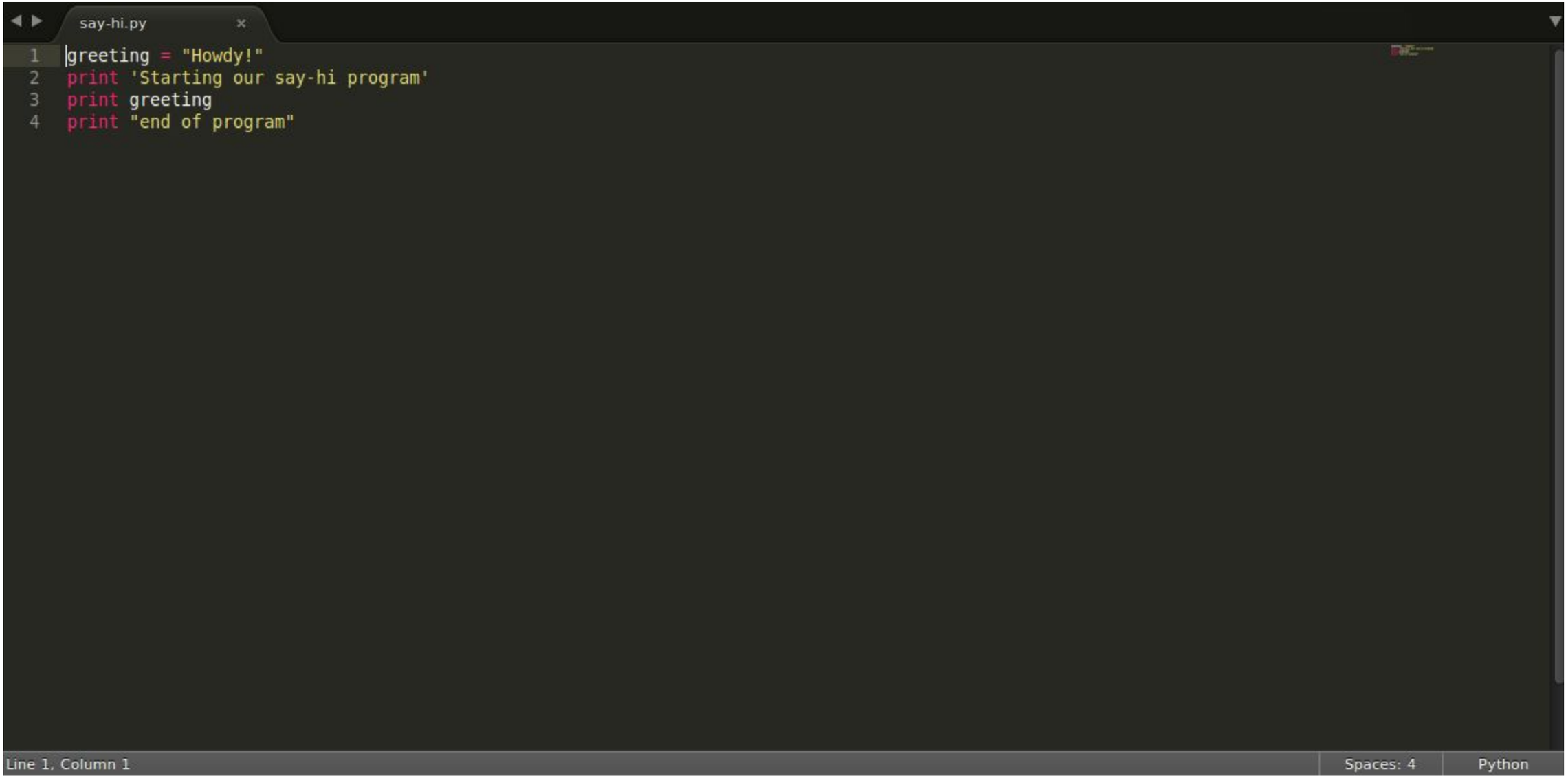
TEXT EDITORS

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 see what Sublime Text look like with Python.

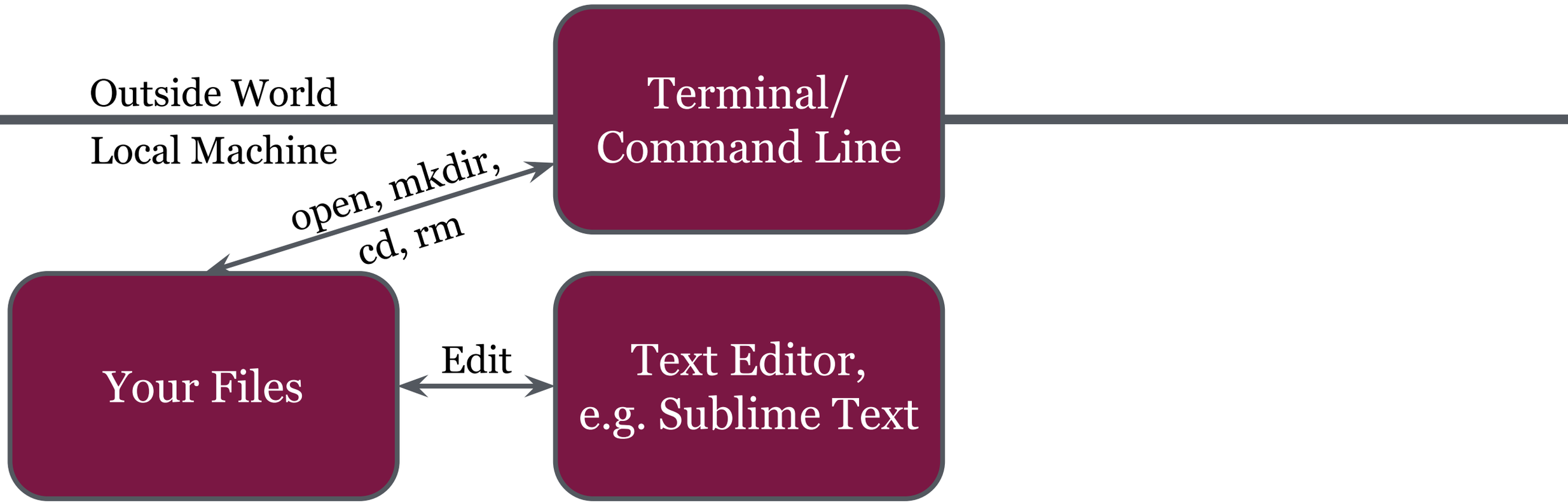
TEXT EDITORS



TEXT EDITORS

- Open “say-hi.py”, found in the lesson-05 folder of the class repo, in Sublime Text to see it for yourself.

DATA SCIENCE TOOLS



ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS

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

DELIVERABLE

Answers to the above questions

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 two components:
  - A web application
  - Notebook documents

INTRODUCTION

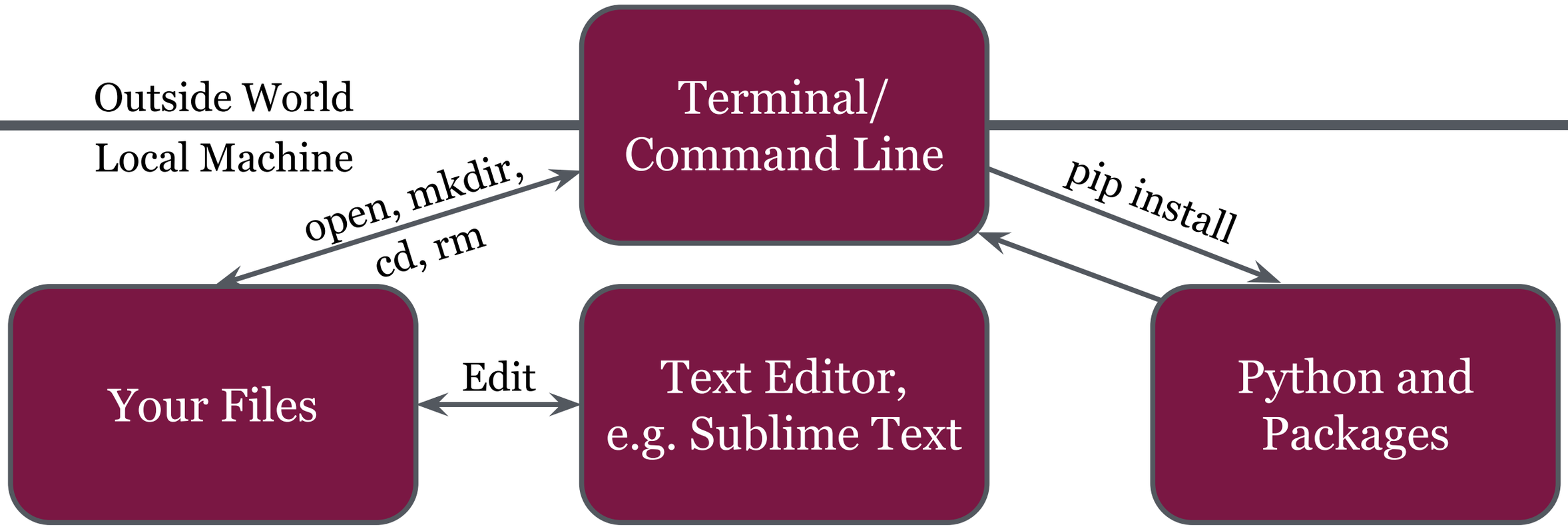
PYTHON PACKAGES

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

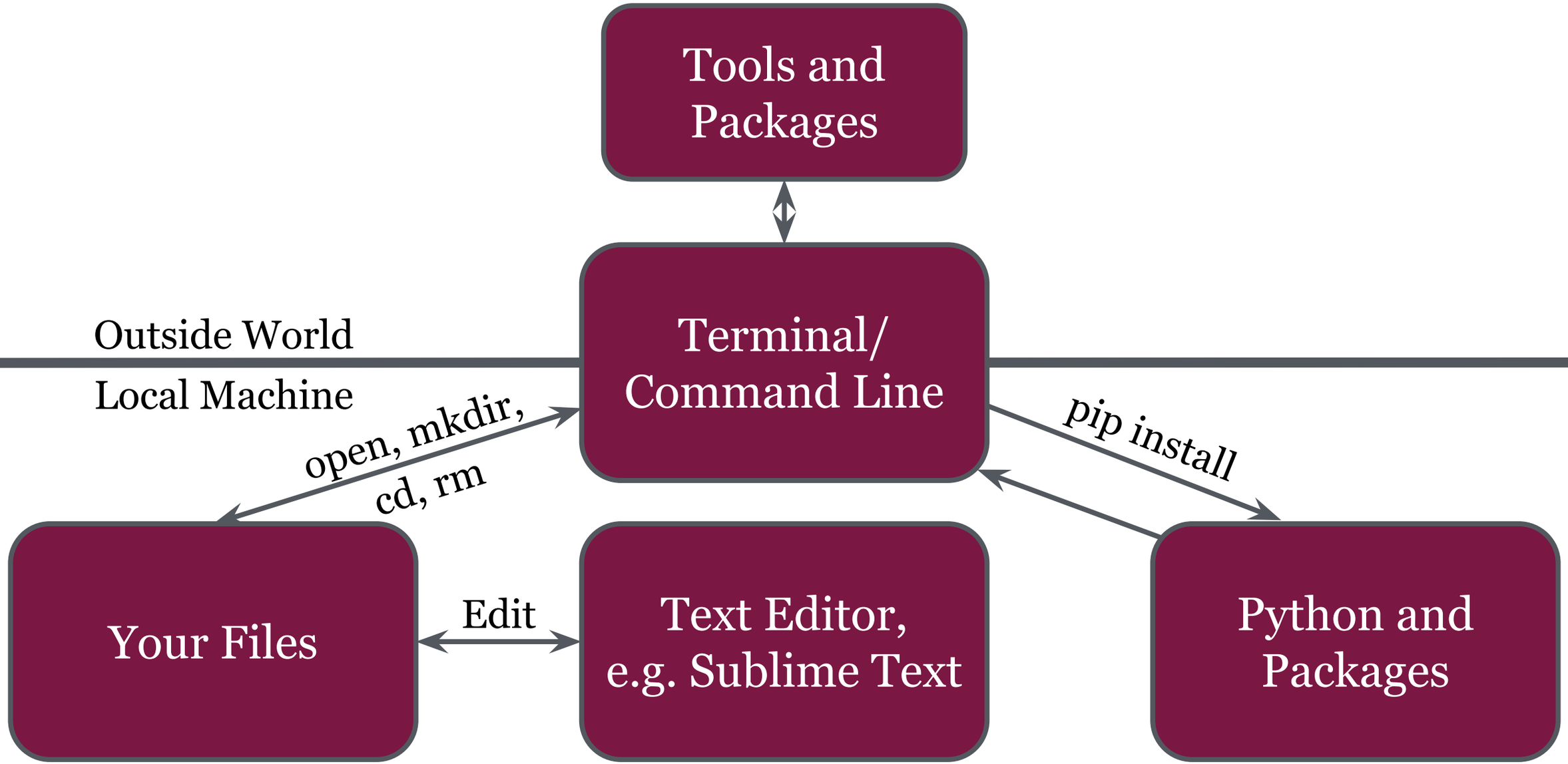


THE OUTSIDE WORLD

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.

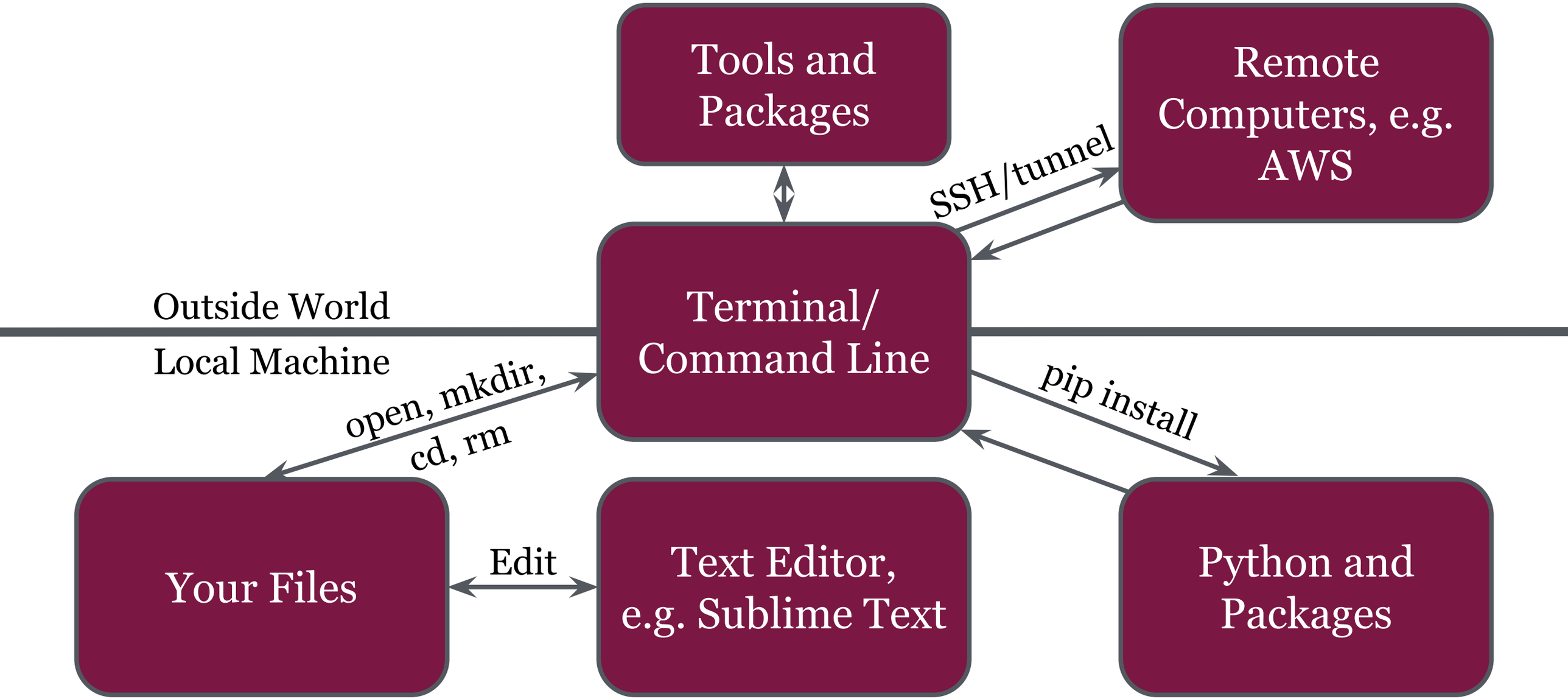
DATA SCIENCE TOOLS



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 demonstration of this.

DATA SCIENCE TOOLS



GIT

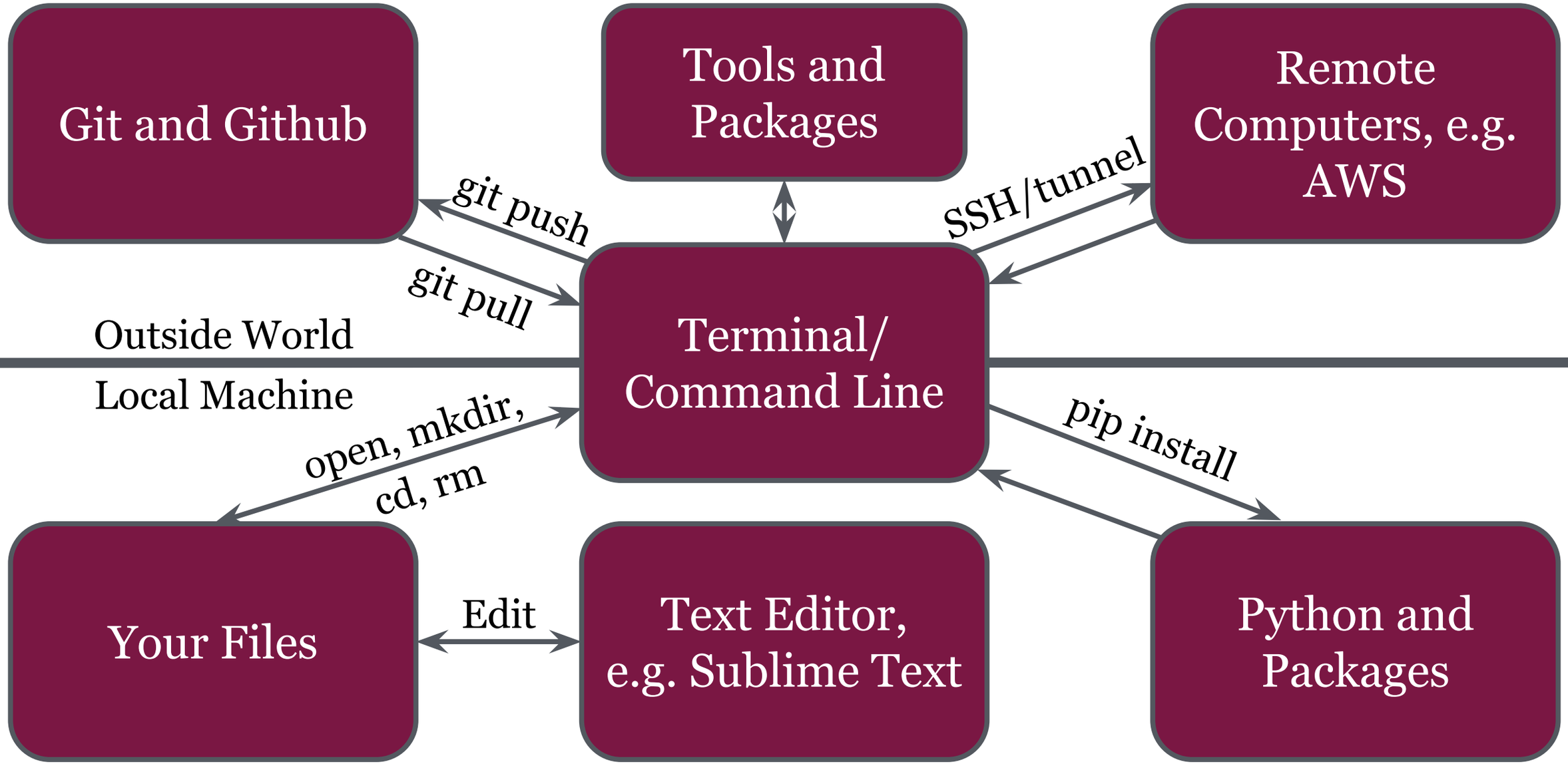
GIT

- 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`.

DATA SCIENCE TOOLS



ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS

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

DIRECTIONS (20 minutes)

1. Let's review the exercises from Codecademy Python.
2. Let's review the exercises from the GA command line tutorial.
3. Are there any questions?

DELIVERABLE

Questions

GUIDED PRACTICE

ODDS AND PROBABILITY

ACTIVITY: ODDS & PROBABILITY

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?