GENERAL ASSEMBLY

DATA SCIENCE TOOLS

Angelo Klin

Katra Analytics

LEARNING OBJECTIVES

- Identify the Data Science toolkit
- Navigate Git and the Command Line
- Describe Probability vs Odds

DATA SCIENCE

PRE-WORK

PRE-WORK REVIEW

- Use descriptive statistics to understand your data

OPENING

DATA SCIENCE TOOLS

LET’S DISCUSS THE CURRENT LESSON OBJECTIVES

- Identify the Data Science toolkit
- Navigate Git and the Command Line
- Describe Probability vs Odds

INTRODUCTION

TOOLS OF THE TRADE

LOCAL MACHINE

- On your local computer, you have a variety of tools at your disposal.
  - Text Editor
  - Programs, packages and tools
  - Your files

- All of these can be accessed through a **Terminal** or through a **GUI** (Graphical User Interface)

- You can navigate your files through the Terminal or through Finder/ Explorer

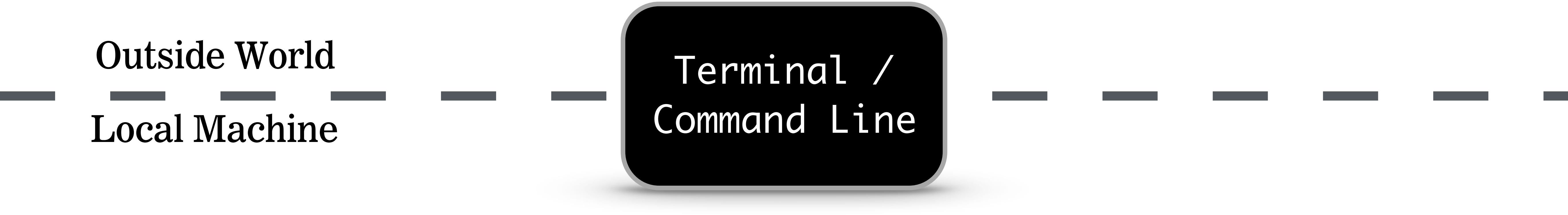
TOOLS OF THE TRADE

- Today we are going to review **some** of the tools we use in Data Science

- We will see how they fit into the wider environment

- We will start with the Command Line
  - This is your portal to your computer and the outside world

DATA SCIENCE TOOLS



Outside World

Local Machine

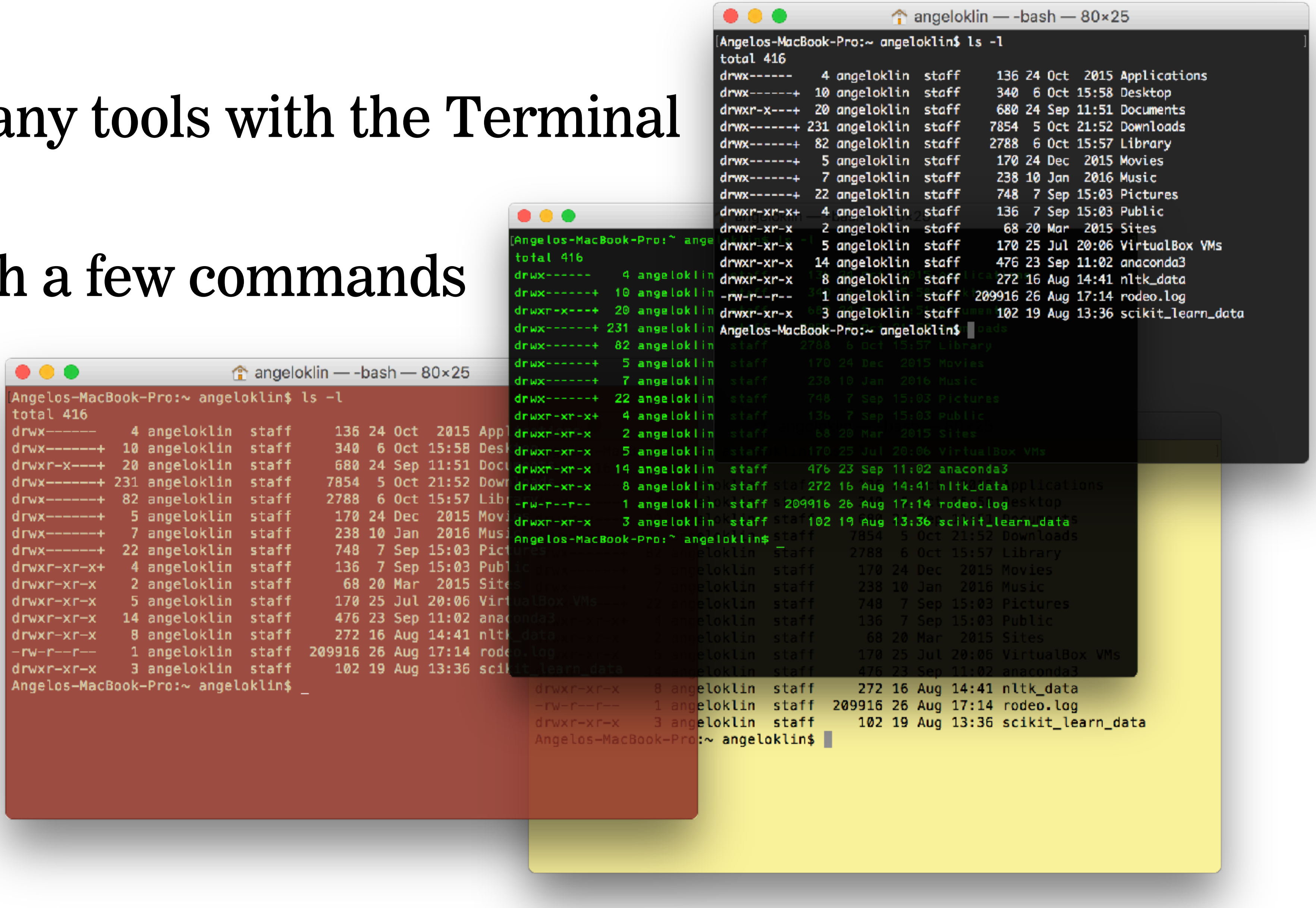
Terminal / Command Line

DEMONSTRATION

COMMAND LINE

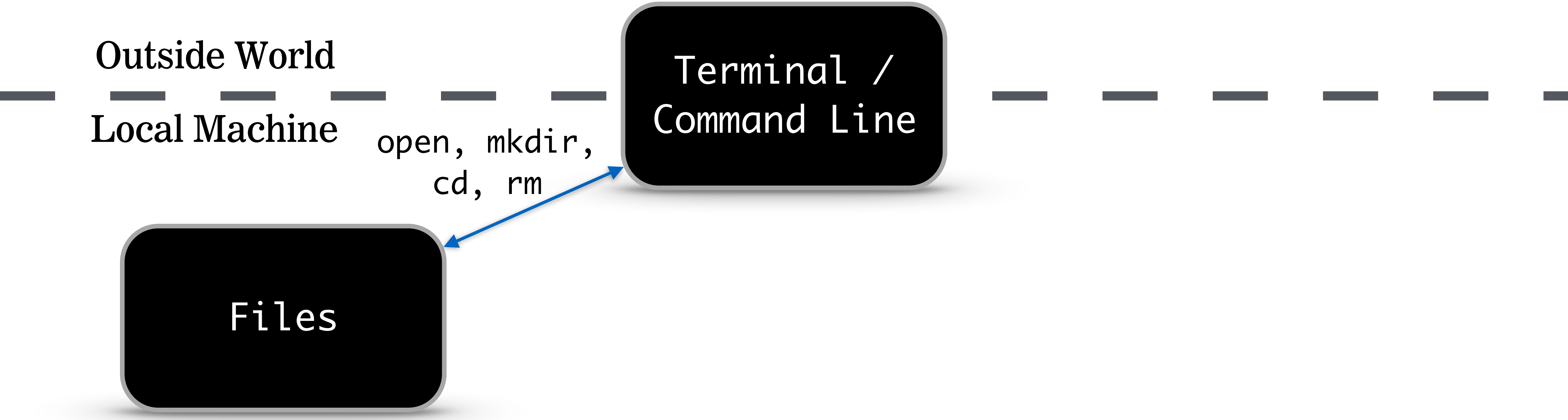
COMMAND LINE

- We can access many tools with the Terminal
- Let's walk through a few commands
  - cd
  - pwd
  - mkdir
  - open



```
Angelos-MacBook-Pro:~ angeloklin$ ls -l
total 416
drwx------    4 angeloklin  staff     136 24 Oct  2015 Applications
drwx------+  10 angeloklin  staff     340  6 Oct 15:58 Desktop
drwxr-x---+  20 angeloklin  staff     680 24 Sep 11:51 Documents
drwx------+ 231 angeloklin  staff    7854  5 Oct 21:52 Downloads
drwx------+  82 angeloklin  staff    2788  6 Oct 15:57 Library
drwx------+   5 angeloklin  staff     170 24 Dec  2015 Movies
drwx------+   7 angeloklin  staff     238 10 Jan  2016 Music
drwx------+  22 angeloklin  staff     748  7 Sep 15:03 Pictures
drwxr-xr-x+   4 angeloklin  staff     136  7 Sep 15:03 Public
drwxr-xr-x    2 angeloklin  staff      68 20 Mar  2015 Sites
drwxr-xr-x    5 angeloklin  staff     170 25 Jul 20:06 VirtualBox VMs
drwxr-xr-x   14 angeloklin  staff     476 23 Sep 11:02 anaconda3
drwxr-xr-x    8 angeloklin  staff     272 16 Aug 14:41 nltk_data
-rw-r--r--    1 angeloklin  staff  209916 26 Aug 17:14 rodeo.log
drwxr-xr-x    3 angeloklin  staff     102 19 Aug 13:36 scikit_learn_data
Angelos-MacBook-Pro:~ angeloklin$
```

DATA SCIENCE TOOLS



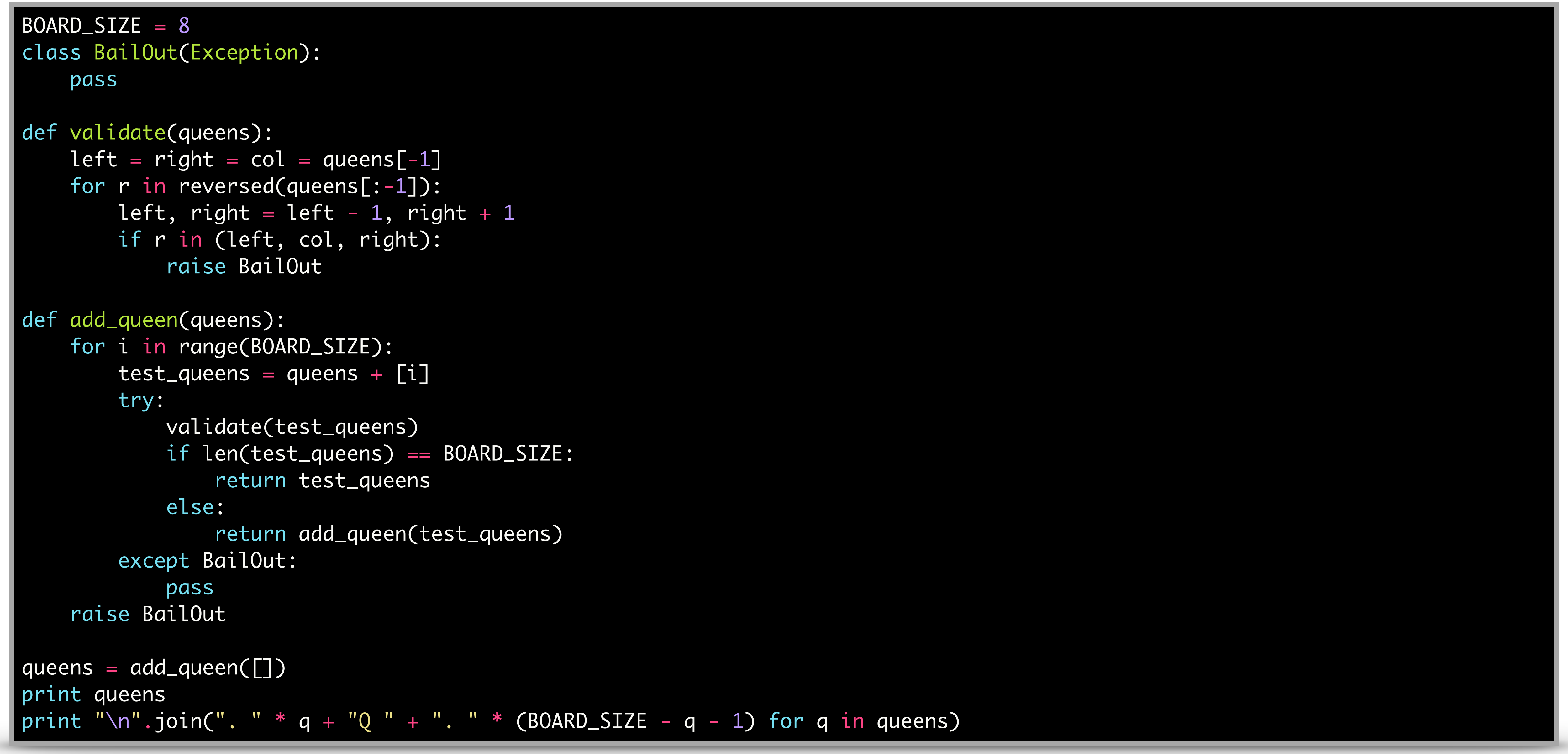
INTRODUCTION

TEXT EDITORS

TEXT EDITORS

- So far, we have used Jupyter Notebooks in place of a text editor
- However, there are many options available
  - Vim
  - Sublime Text
  - Atom
- Let’s see how Python looks with Syntax Highlight

TEXT EDITORS

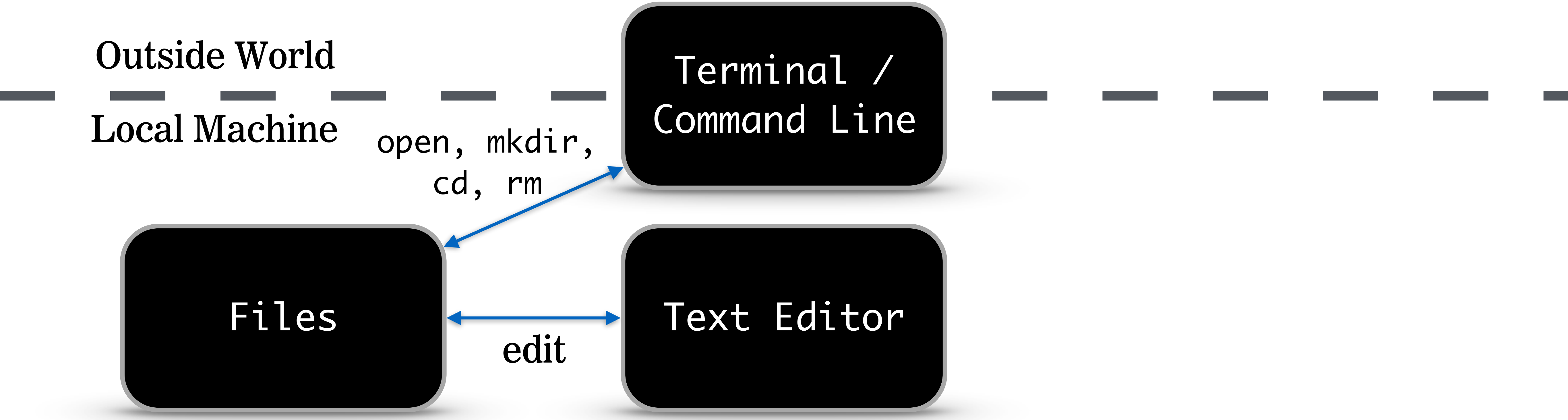


```
BOARD_SIZE = 8
class BailOut(Exception):
    pass

def validate(queens):
    left = right = col = queens[-1]
    for r in reversed(queens[:-1]):
        left, right = left - 1, right + 1
        if r in (left, col, right):
            raise BailOut

def add_queen(queens):
    for i in range(BOARD_SIZE):
        test_queens = queens + [i]
        try:
            validate(test_queens)
            if len(test_queens) == BOARD_SIZE:
                return test_queens
            else:
                return add_queen(test_queens)
        except BailOut:
            pass
    raise BailOut

queens = add_queen([])
print queens
print "\n".join(". " * q + "Q " + ". " * (BOARD_SIZE - q - 1) for q in queens)
```

TEXT EDITORS

- Open the lesson 05 folder of the class repository and open the files
  - ~/lessons/lesson-05/code/say-hi.py
  - ~/lessons/lesson-05/code/eight-queens.py
- **NOTE**: These are Python source code, NOT Jupyter Notebooks!

DATA SCIENCE TOOLS



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS

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

INTRODUCTION

JUPYTER NOTEBOOK

JUPITER NOTEBOOK

- Where does Jupyter Notebook fit in?
  - “The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualisations and explanatory text.”

- Jupyter notebooks combine
  - The console
  - Web application
  - Markdown to capture the whole computation process

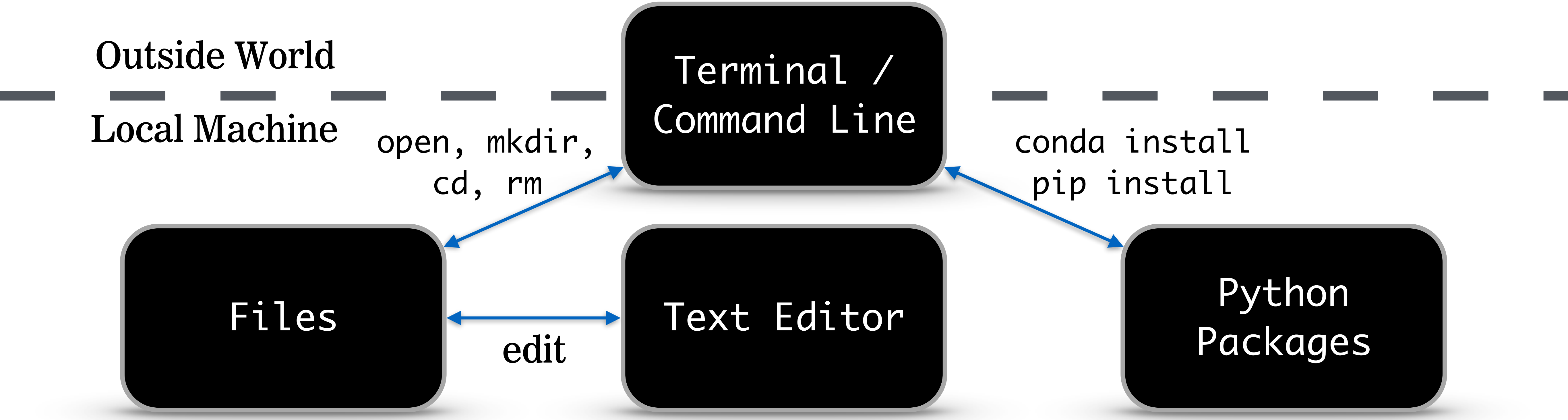
INTRODUCTION

PYTHON PACKAGES

PYTHON PACKAGES

- We can add programs and packages as needed
- To add Python packages, we use tools like `conda` and `pip`
- To install Beautiful Soup, a HTML/XML parsing package
  - `conda install beautifulsoup4`
  - `pip install beautifulsoup4`

DATA SCIENCE TOOLS



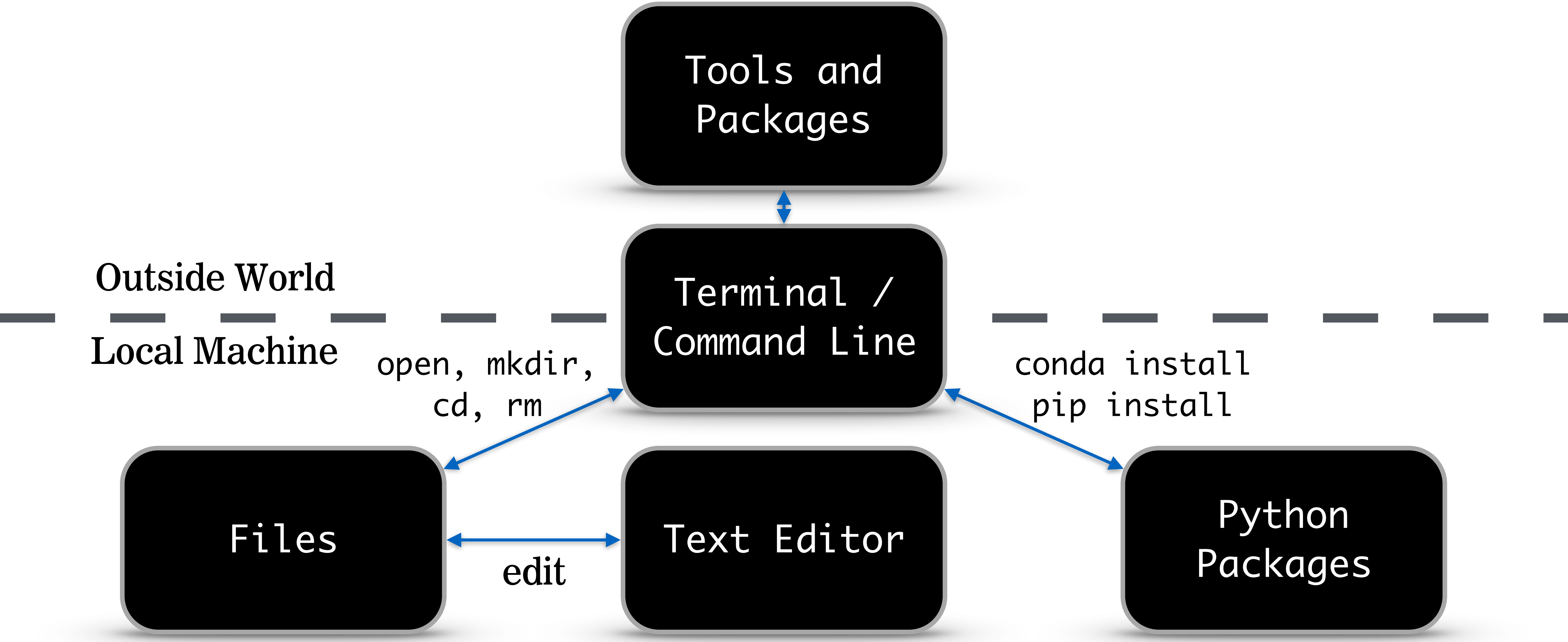
INTRODUCTION

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

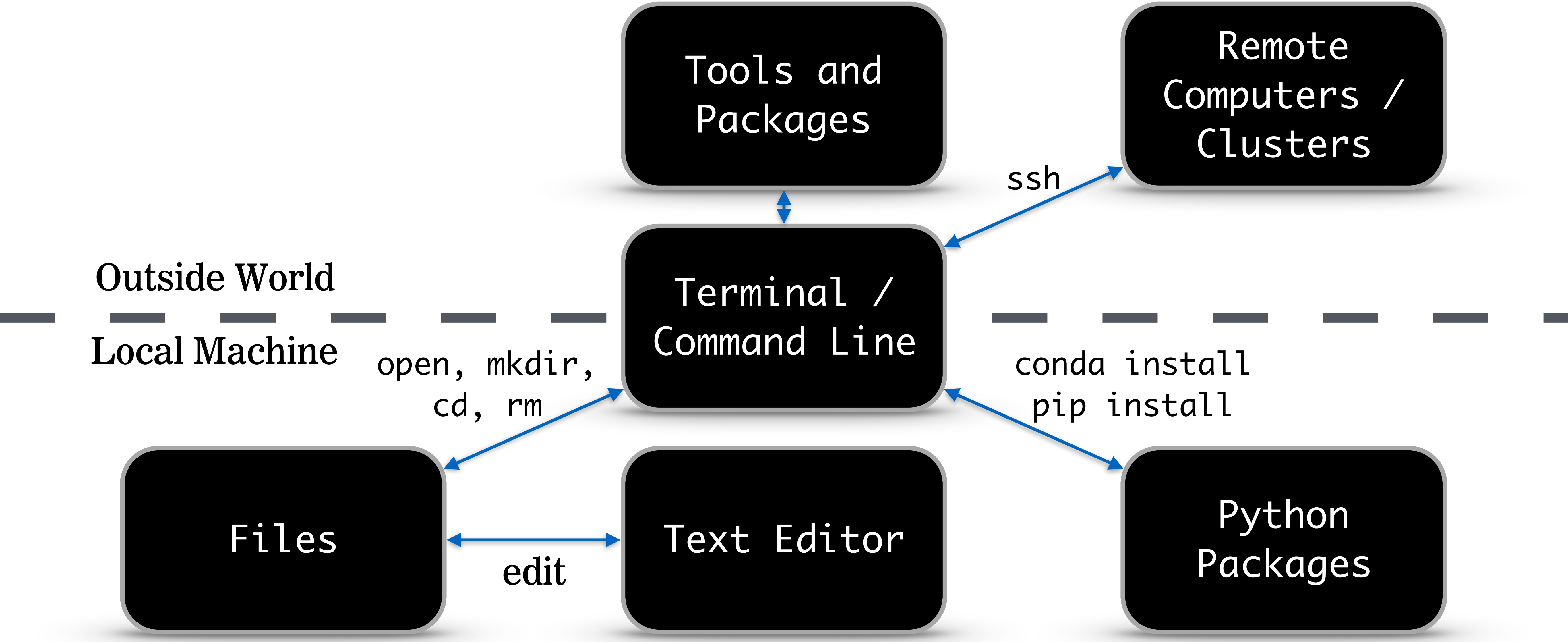
- As we saw with conda/pip/git, the Command Line can connect us to the outside world
  - This becomes more important for data

- We may have HIPAA protected data
  - This means we can’t leave this sensitive data on our local machine

- 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



INTRODUCTION

GIT

GIT

- Version control is necessary when working on complex projects
- Git is a way of tracking changes we have 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

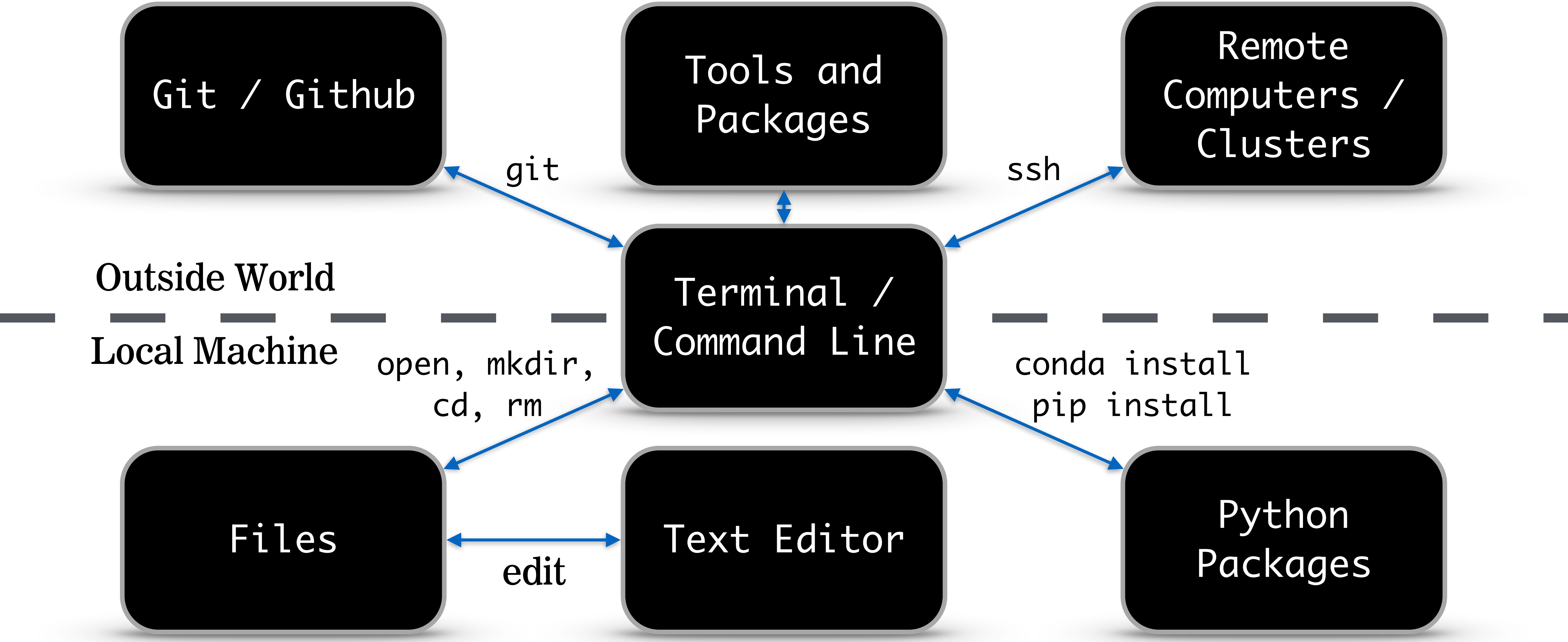
GIT

- Let’s see an example of using Git and Github

- There are three primary commands we will 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

EXERCISE

DIRECTIONS: 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?

GUIDED PRACTICE

GIT AND THE COMMAND LINE

ACTIVITY: GIT AND THE COMMAND LINE

EXERCISE

DIRECTIONS (35 MINUTES)

1. Let’s review the exercises from Codecademy Python
2. Let’s review the exercises from the GA’s Command Line Tutorial
3. Are there any questions?

GUIDED PRACTICE

ODDS AND PROBABILITY

ACTIVITY: ODDS AND PROBABILITY

EXERCISE

DIRECTIONS (20 MINUTES)

1. Some of you may already be familiar with odds and probability.
2. We will use the starter code in lesson 05 of the class repository to review the concepts of odds and probability.
    a. ~/lessons/lesson-05/code/starter/starter-5.ipynb

CONCLUSION

TOPIC REVIEW

TOPIC REVIEW

- What are some common Data Science tools?
- Why are these tools useful?
- Any other questions?

BEFORE NEXT CLASS

DUE DATE

- Project
  - Unit Project 2

Q & A