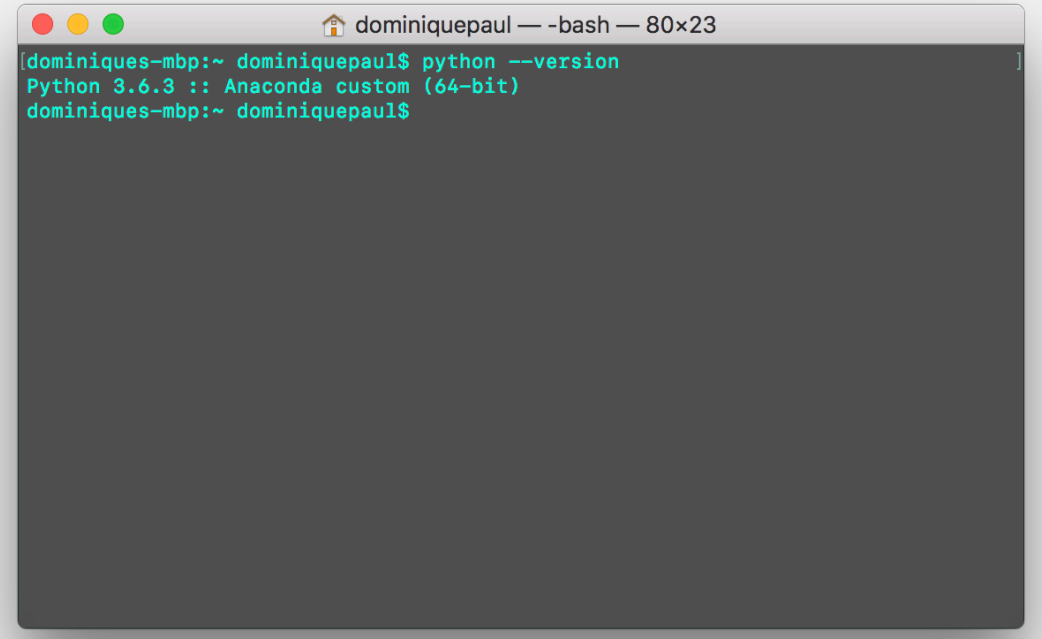
INSTALLING PYSPARK

This presentation will explain how to install PySpark on MacOS
and automatically connect it to Jupyter Notebooks

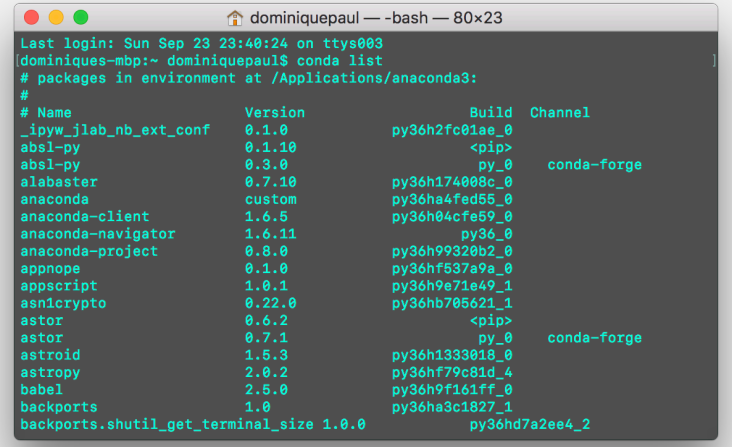
Created by Dominique Paul for Prof. Johannes Binswanger (SIAW – University of St. Gallen)

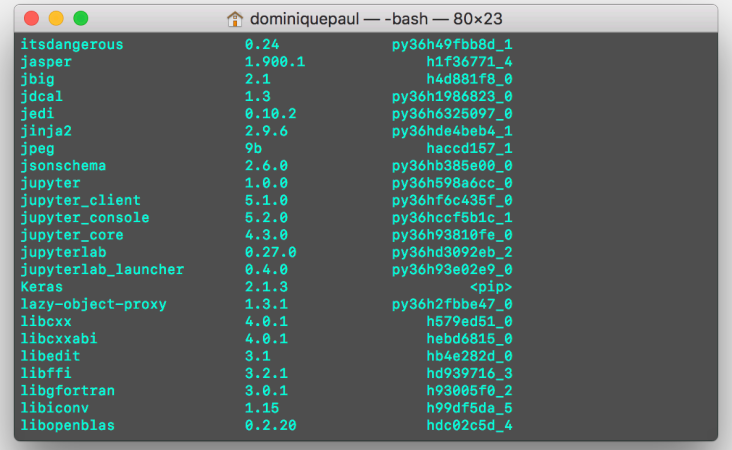


1. PYTHON

- We assume that you have python installed
- If you are not sure, open Terminal and enter:
  python –version
- If nothing shows up, install python from here:

  https://www.python.org/downloads/



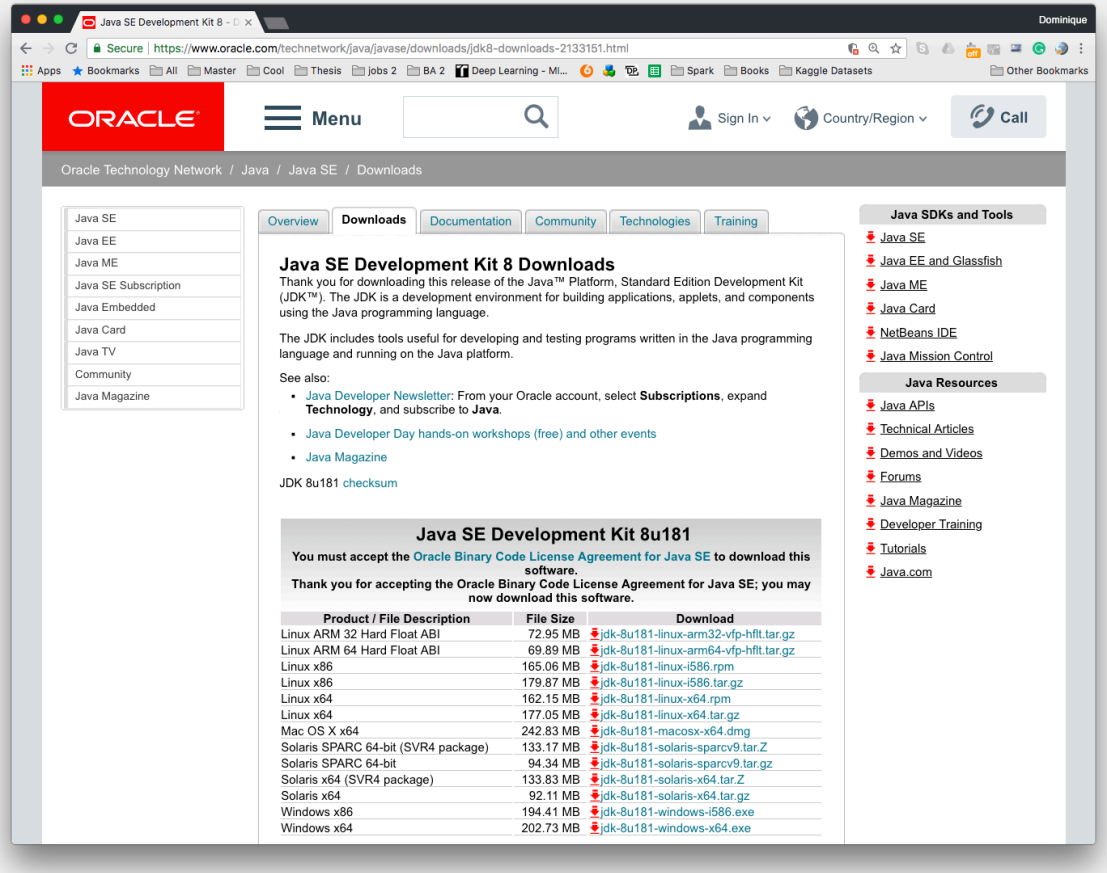


2. JUPYTER

- We also assume that you have Jupyter Notebooks installed
- If you are not sure, open Terminal and enter one of the following:

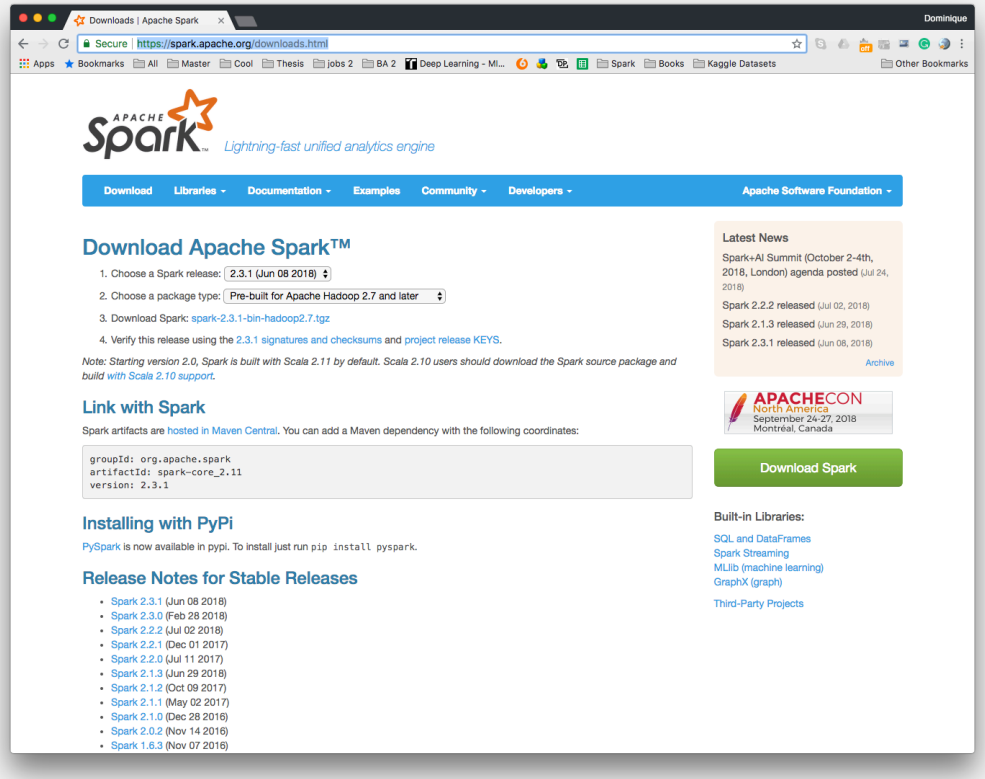
  conda list
  pip list
- Jupyter should be one of the packages listed
- Anaconda comes with Jupyter Notebooks pre-installed
- If you are working with pip, you can install Jupyter Notebooks with

  pip install jupyter



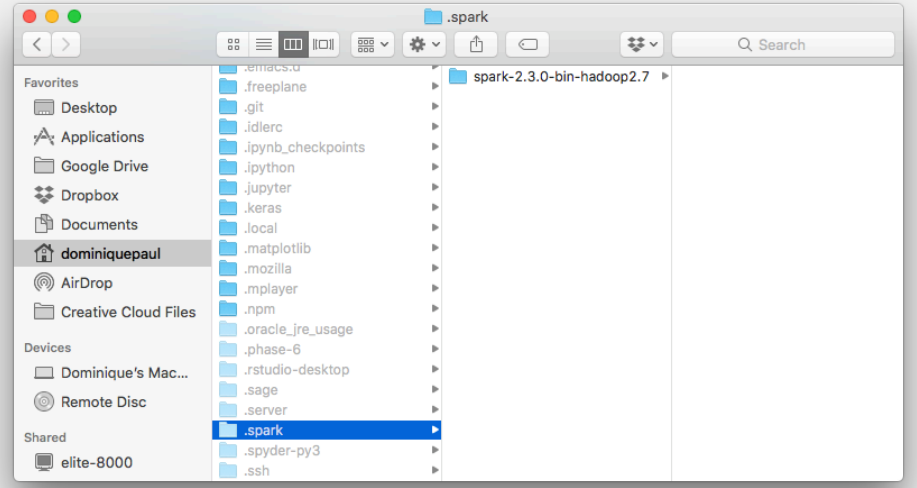
3. JAVA INSTALLATION

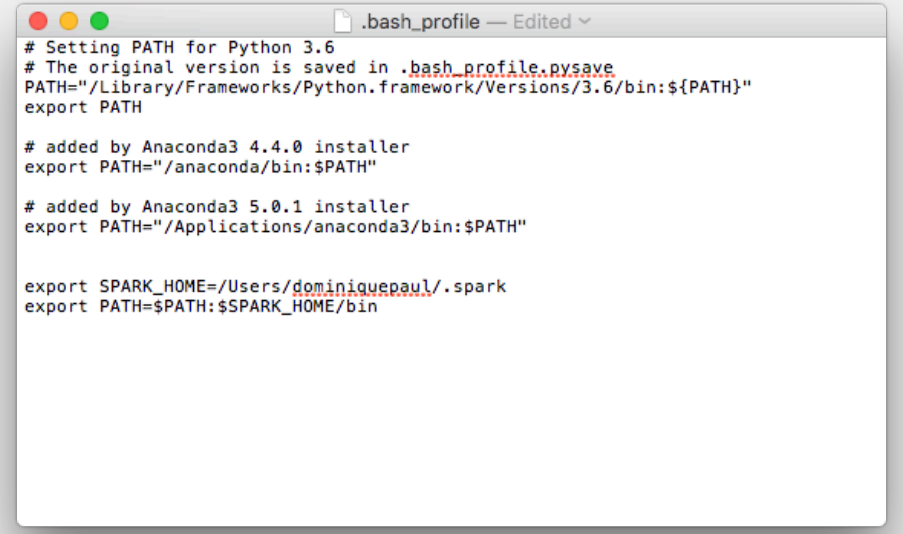
- To use PySpark, we will need the Java Developer Kit
- **Make sure that you are using version 8.** PySpark may not work with other versions such as 10.0
- Go to https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html and download the version for your system
- Once the download has completed, open the installer and follow the instructions



4. DOWNLOAD SPARK

- Go to https://spark.apache.org/downloads.html
- Download the latest stable version of Spark based on a pre-built version for Apache Hadoop





4. INSTALLING SPARK

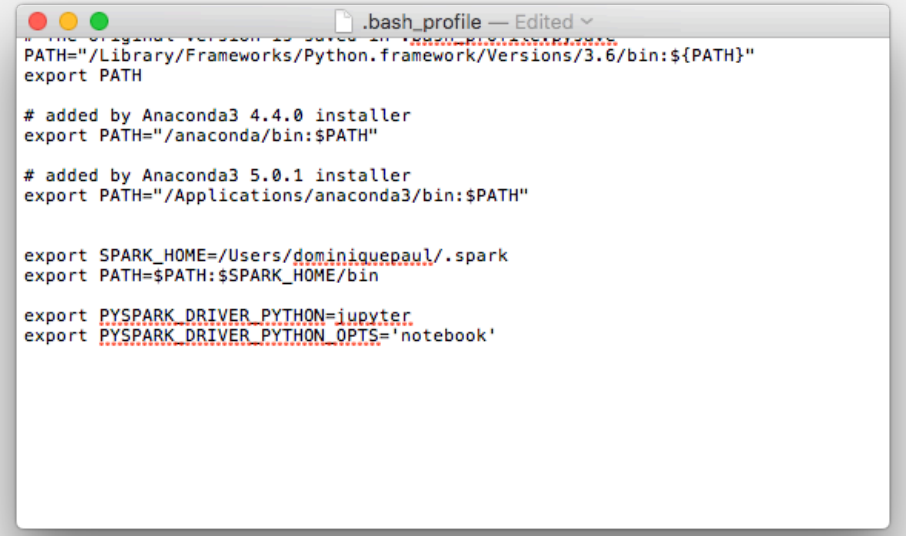
1. In your finder home directory, press (command + shift + .) This will reveal all hidden folders and files. You can reverse this by using the same keyboard shortcut again.
2. create a folder called ".spark" and move the unzipped Spark folder into it.
3. Right click (ctrl + click) onto the spark folder and click on "copy spark as pathname)
4. Go back to your homedirectory and open the file called ".bash_profile". This will open the file in textedit.
5. In the file, add the following lines, thereby replacing

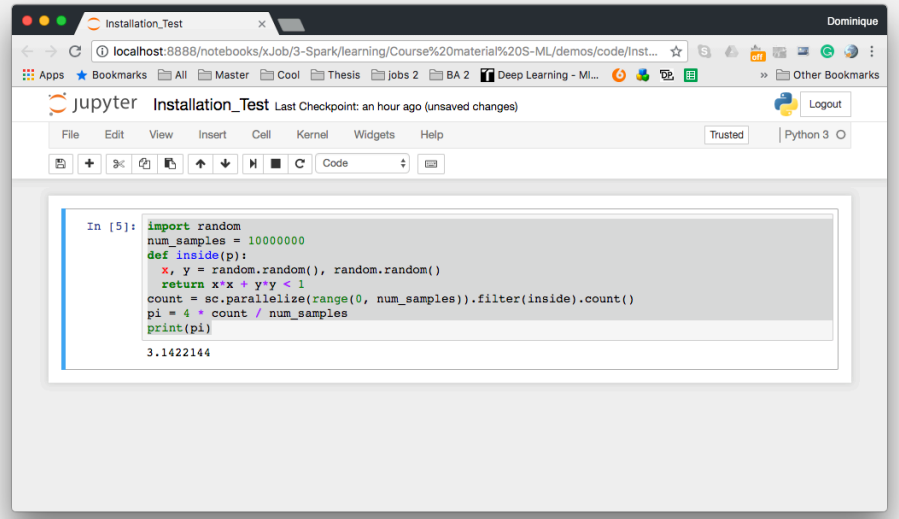
   "YOUR_FILEPATH" with the filepath of the spark folder you just

   copied:

   *export SPARK_HOME=**YOUR_FILEPATH***

   export PATH=$PATH:$SPARK_HOME/bin





4. MAKING THE JUPYTER CONNECTION

- In the same .bash_profile file, add the following two lines:

  export PYSPARK_DRIVER_PYTHON=jupyter

  export PYSPARK_DRIVER_PYTHON_OPTS='notebook'

- Save the .bash_profile file and open terminal. If you had it open already, quit it (command + Q) and re-open it so it can re-load the .bash_profile file
- Enter *pyspark* in the terminal, this opens a new Jupyter Notebook with a spark connection.
- Test whether the installation is successful by running the code on the next slide

CODE SNIPPET

```
import random
num_samples = 10000000
def inside(p):
        x, y = random.random(), random.random()
        return x*x + y*y < 1
count = sc.parallelize(range(0, num_samples)).filter(inside).count()
pi = 4 * count / num_samples
print(pi)
```