

CIS 399: Science of Data Ethics

Spring 2019

Getting Started: Jupyter Setup Tutorial

Due January 29, 2019 by 3pm

For this tutorial, our primary goals are to get you set to run the infrastructure we'll need for this class. That software infrastructure includes **Jupyter**, **Python**, and assorted data science libraries. To do this, we'll be using **Docker**, which is a “container” manager that enables you to pull down and run different software components. Docker will manage your development tools and environment.

This tutorial is largely borrowed from CIS545 - Big Data Analytics. We thank the CIS545 teaching staff for putting this resource together.

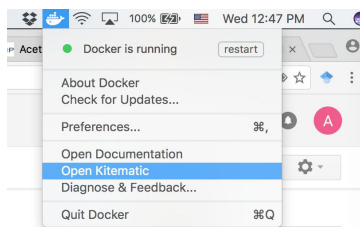
1 Getting the Necessary Software

1.1 Installing Docker

Your first task will be to install Docker itself. Please see the setup instructions below, which depend on your operating system. If, during download of the installer from Docker, you have an option to choose between the “Stable” and “Beta” versions, please stick with Stable!

- [Mac](#).

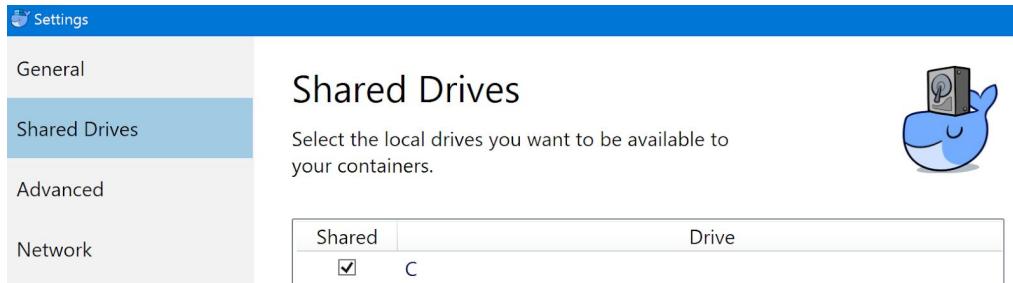
Once Docker is installed, go to the Docker “whale” icon in your menu bar and choose “Open Kitematic”:



When the dialog box pops up, click on the link that says “You can download it [here](#).” Download Kitematic and install in Applications.

- [Windows 10 Pro/Education or better](#).

Once Docker is installed, you should see the little whale icon for Docker on your Task Bar. Right-click, then choose Settings, then “Shared Drives.” Add a check for the C Drive.



- [Windows 8/Windows 10 Home](#). We recommend that you upgrade to Windows 10 Education or Pro if possible, as the Docker integration is much better. However, if you have Windows 8 or Windows 10 Home, you can install the “legacy” product called [Docker Toolbox](#). Before installing Docker Toolbox, check if you have [Oracle VirtualBox](#) installed. If so, please update to the latest build else check the option to install virtualbox.

Once Docker is installed, run Docker Quickstart Terminal. Approve the various requests. Ultimately you should get a bash terminal.

- [Linux \(Ubuntu\)](#). Install [Docker CE](#) according to the basic instructions.

As you follow the instructions, note that you don’t need to validate with `docker-compose` (which we won’t be using for this class).

1.2 Installing Jupyter on Docker

Initially Launching Jupyter and Sharing a Directory

Launch your operating system command-line: in Mac OS and Linux, this is “Terminal” and on Windows it’s “Command Prompt”. Then type in the following two lines. In this document, we’ll use *userid* to refer to your user login ID on your local machine.

For Windows 10 Pro/Education (where `\Users\{userid}` is in `%USERPROFILE%`):

```
mkdir %USERPROFILE%\Jupyter
```

```
echo "Test" > %USERPROFILE%\Jupyter\test.txt
```

```
docker run -v %USERPROFILE%\Jupyter:/home/jovyan/work -it -p 8888:8888 jupyter/all-spark-notebook
```

For Windows 10 Home you’ll need to first open the Command Prompt and run:

```
mkdir %USERPROFILE%\Jupyter
```

```
vboxmanage sharedfolder add default --name "%USERPROFILE%\Jupyter" --hostpath "%USERPROFILE%\Jupyter" --automount
```

```
echo "Test" > %USERPROFILE%\Jupyter\test.txt
```

```
docker run -v %USERPROFILE%\Jupyter:/home/jovyan/work -it -p 8888:8888 jupyter/all-spark-notebook
```

NOTE : If you get ‘vboxmanage is not a recognized..’ error while running vboxmanage command, you will have to add vboxmanage into your system path. For windows, virtual box is by default stored at `C:\Program Files\Oracle\VirtualBox`. This location needs to be added to your PATH variable. To add to your PATH variable, go to Start > Search for ‘Edit the system environment variables’ > Environment variables > Select ‘Path’ under User variables > Edit > New > Paste the location of Virtual Box folder

For Mac/Linux, or Windows 10 Home under Docker Quickstart Terminal (where `/Users/{userid}` is in `$HOME`):

```
mkdir ~/Jupyter
```

```
echo "Test" > $HOME/Jupyter/test.txt
```

```
docker run -v $HOME/Jupyter:/home/jovyan/work -it -p 8888:8888 jupyter/all-spark-notebook
```

These two commands will (1) create a directory for your Jupyter environment and files in `/Users/{userid}`, (2) download and install a Docker image containing Python and Jupyter (formerly iPython) as well as a local version of Spark. (If you want to dive deeply into the Docker-Jupyter environment, you can get details [here](#).) For Linux, you may need to add the parameter “`--net=host`” to the command line.

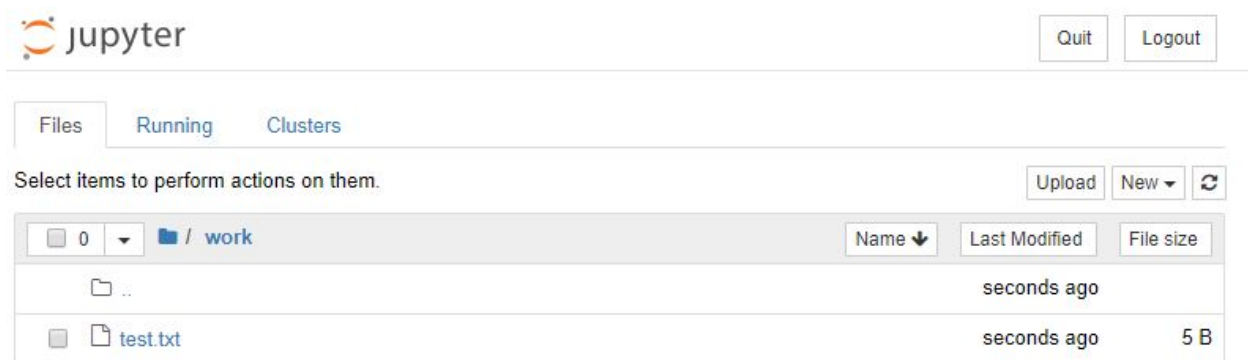
Once Docker says it is ready, it should give a message like:

```
Copy/paste this URL into your browser when you connect for the first time,  
to login with a token:  
http://localhost:8888/?token=9ed4c2dad760cbde0215a0ee7784adf1d416c1ff4d9068eb
```

Select the URL (starting with `http://`) and copy it. Open up your Web browser and paste it into the URL bar. You should see a screen like the one below. Click on the “work” directory.



Verify that **test.txt** exists. This file was created in your \$HOME/Jupyter directory on your host machine, and **it needs to be there** to confirm that your Jupyter instance can share files with the host.



If You Need a Password or Token

Follow the instructions as above.

If It Didn't Work

Check the URL you were given. If it says something like “http://(eabacdef or 127.0.0.1):8888/...” replace the item in the parentheses with **127.0.0.1** or **localhost**.

If you are on Windows 10 Home or otherwise using Docker Toolbox, you may need to replace **127.0.0.1:8888** with **192.168.99.100:8888**. If that still doesn't work, **docker machine ip default** might tell you a different address to use.

1.3 Connecting to Jupyter after a Reboot

At times you'll need to stop your Docker instance, e.g., after rebooting. If you reboot and repeat the steps in Section 1.2, you'll end up creating **another** container with Jupyter, which can be very wasteful. Instead you can relaunch and reconnect to your existing container via Kitematic.

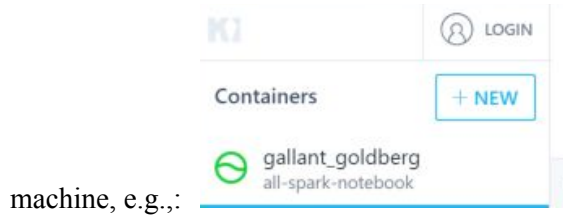
Mac OS X: **Run Kitematic**. You can skip the registration with Docker Hub.

Windows 10: **Run Kitematic**.

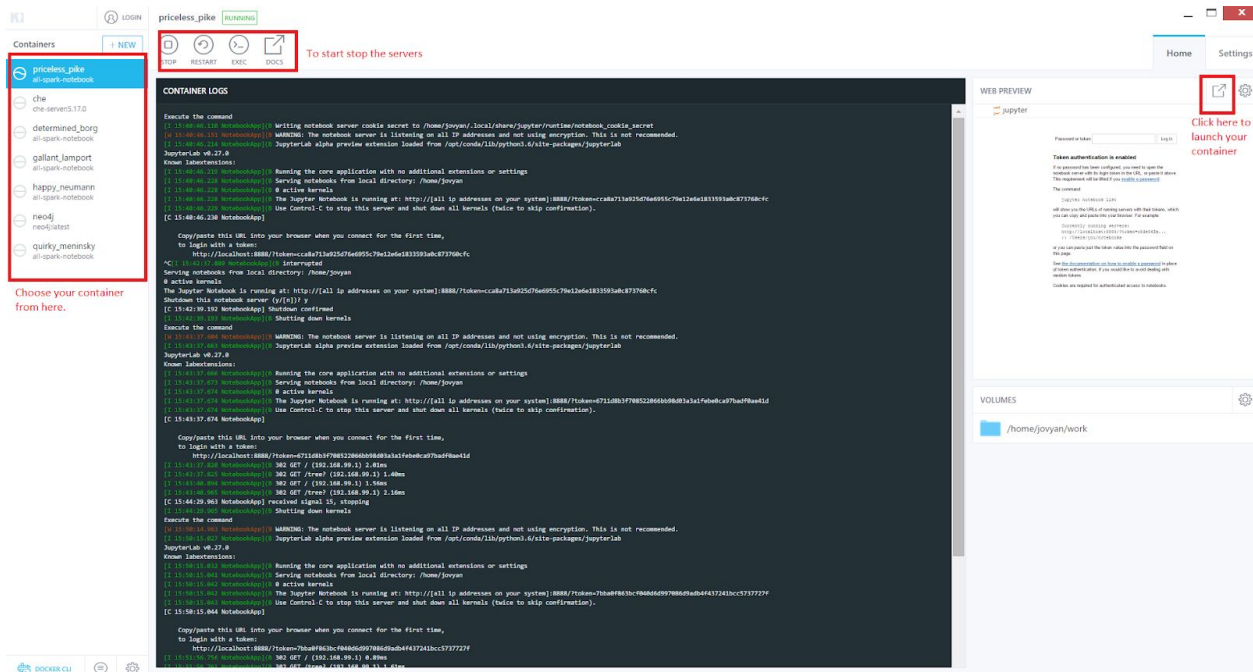
- On Windows 10 Home you will need to do this via the Start Menu.
- On Windows 10 Pro or Education, there is a “whale” icon in the Task Bar that will let you access Kitematic.

If Kitematic complains about an ENOENT error, click on “Use Virtualbox.” You can skip the registration with Docker Hub, since you won't be publishing any containers.

1. Once you are at the main page, look on the left side, where you'll see a list of containers on your



2. Click on the container and click on the **Start** button. This will start your Jupyter Notebook.

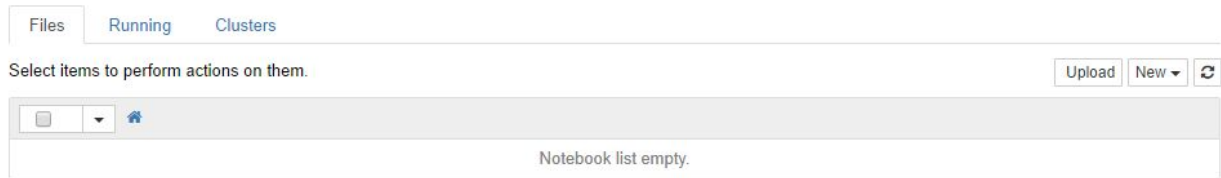


Inside the log window you will eventually see something like:

Copy/paste this URL into your browser when you connect for the first time,
to login with a token:

<http://localhost:8888/?token=9ed4c2dad760cbde0215a0ee7784adf1d416c1ff4d9068eb>

Click on the Web Preview icon on the right to launch it on a browser. If you see a screen like the one below, then you are ready to go, and you can skip to the “Your Data” section below! If not, you should follow the instructions below.



If You Need a Password or Token

If, instead, you see something like:

Password or token:  Log in

Token authentication is enabled. You need to open the notebook server with its first-time login token in the URL, or enable a password in order to gain access. The command:

```
jupyter notebook list
```

will show you the URLs of running servers with their tokens, which you can copy and paste into your browser. For example:

```
Currently running servers:  
http://localhost:8888/?token=c8de56fa... :: /Users/you/notebooks
```

Or you can paste just the token value into the password field on this page.
Cookies are required for authenticated access to notebooks.

Use the token value generated in the Kitematic container terminal.

eg: <http://localhost:8888/?token=9ed4c2dad760cbde0215a0ee7784adf1d416c1ff4d9068eb>

Your Data

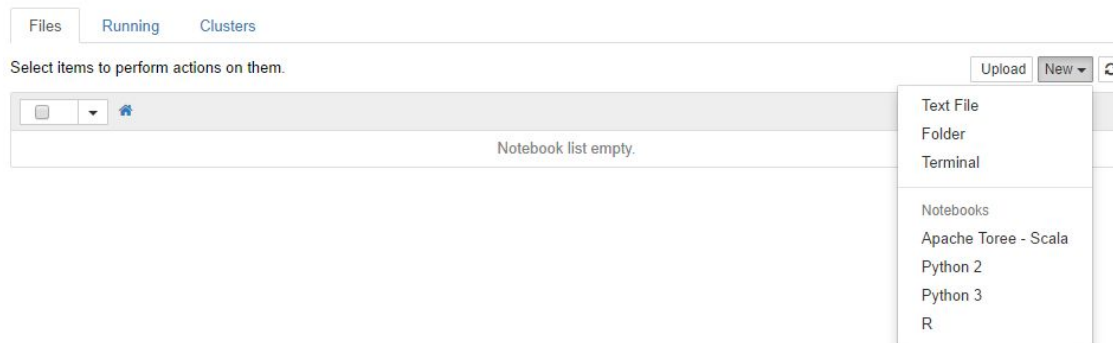
Please consistently work in the “work” directory in Jupyter -- this corresponds to the Jupyter directory on your host machine. Your files should be saved there, and you’ll be able to back up and retrieve them even if Docker crashes.

2 Creating and Visualizing Data

2.1 A First, Really Simple Program

Make sure you are in the **work** directory.


From the browser's view of Jupyter, click on "New":



Click on Terminal. You'll see that by default you get a user called `jovyan`, which astute readers will note is a bad pun on Jupyter. Then type the following:

```
cd work
git clone https://github.com/CIS399-DataEthics/hw0.git
```

This will download the files that you need to complete this homework. Once finished, type `exit`, hit [Enter], and close the Terminal browser window. You should now have a "hw0" folder inside your "work" directory. Click on it and then click on "Jupyter Tutorial.ipynb", which is a Jupyter notebook. Note that the file extension is ".ipynb" for (I)nteractive (Py)thon (N)ote(b)ook. These notebooks are divided into Cells. Some Cells are text (written in Markdown). You won't need to edit these. The other Cells are executable code and will have `In []` to the left of them. After running one of these Cells, a number will appear inside the brackets, indicating the order in which the Cells were run.

Repeatedly click on the "Run" button (it looks like ) or type [Shift]-[Enter] until `In [1]` appears. In a few moments you should see a scatter plot. Now edit the code as instructed in the notebook. Click on the first cell again and press "Run" until `In [4]` appears.

3 Submission

When you have completed the above steps, please click "File -> Download as... -> PDF via LaTeX (.pdf)". This should save a PDF to your Downloads with the contents of the notebook. Please upload this file to Gradescope.

If you do not yet have access to the Gradescope page for the course, you can join using the course entry code “**9YRVVE**” at [gradescope.com](https://www.gradescope.com).