DATA MINING

Python

Python

- In the last few years there is an increasing community that creates Data Mining tools in Python
  - There are also tools in other languages but we will use Python whenever we can for a common point of reference.
- We will use Iron Python that interfaces with .NET and we can run Notebooks in a browser.
  - You can also use any editor and compile and run from a terminal

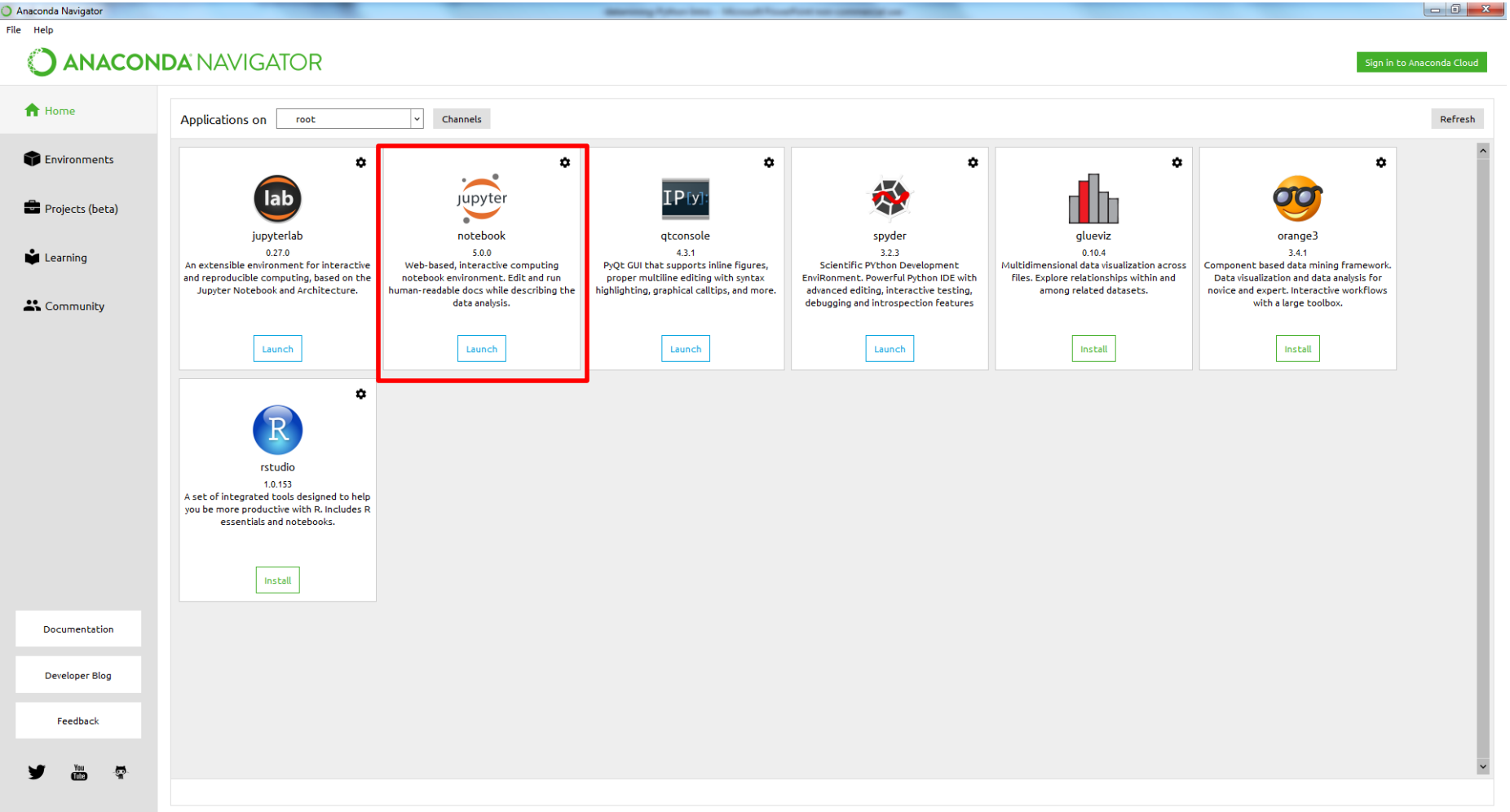
Installing Python

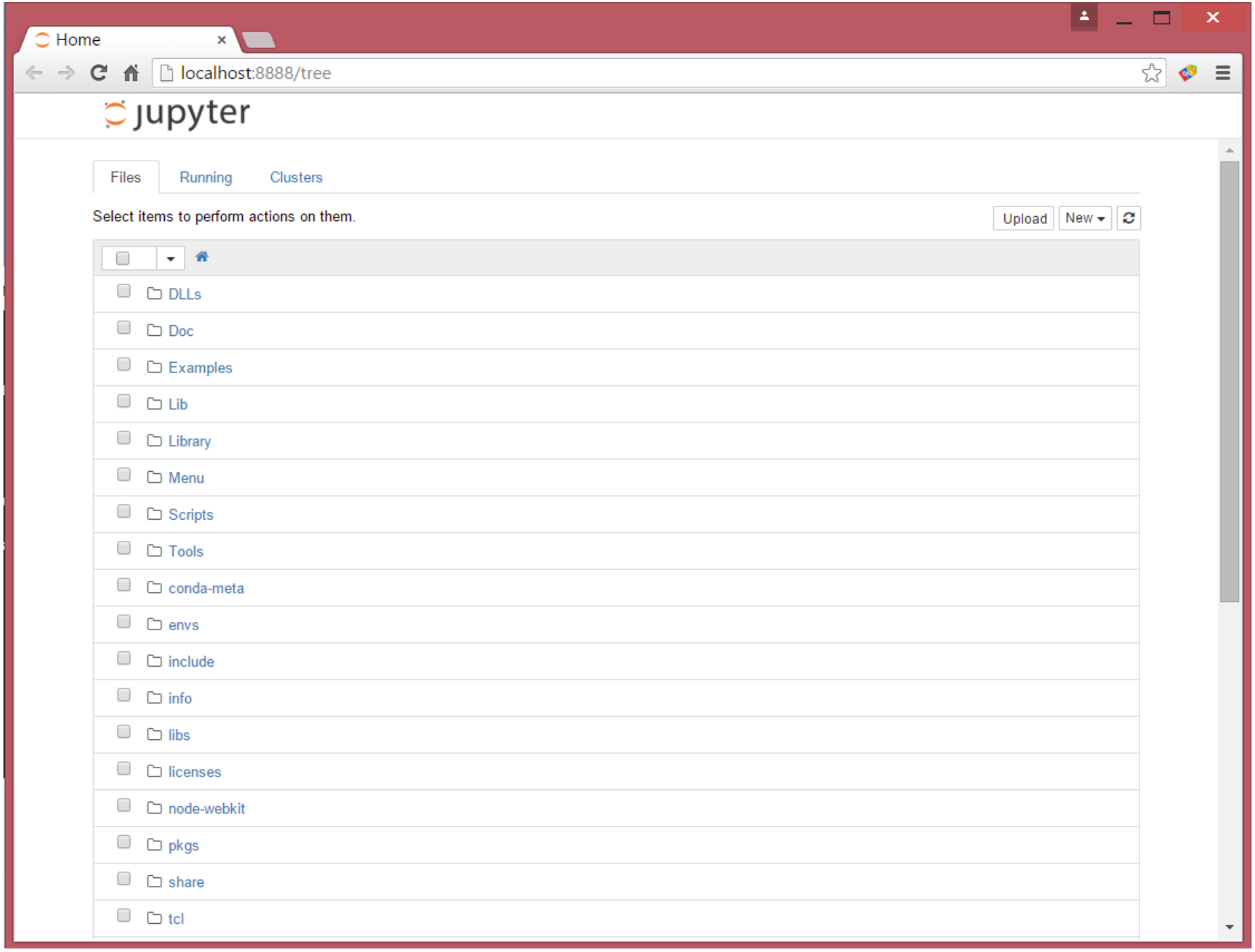
- Installing libraries in Python is complex, so you should download the Anaconda Scientific Python distribution which will install most of the libraries that we will use.
  - There are two versions, Python 2.7 and Python 3.0 and they are not compatible. We will use Python 3.0

Resources

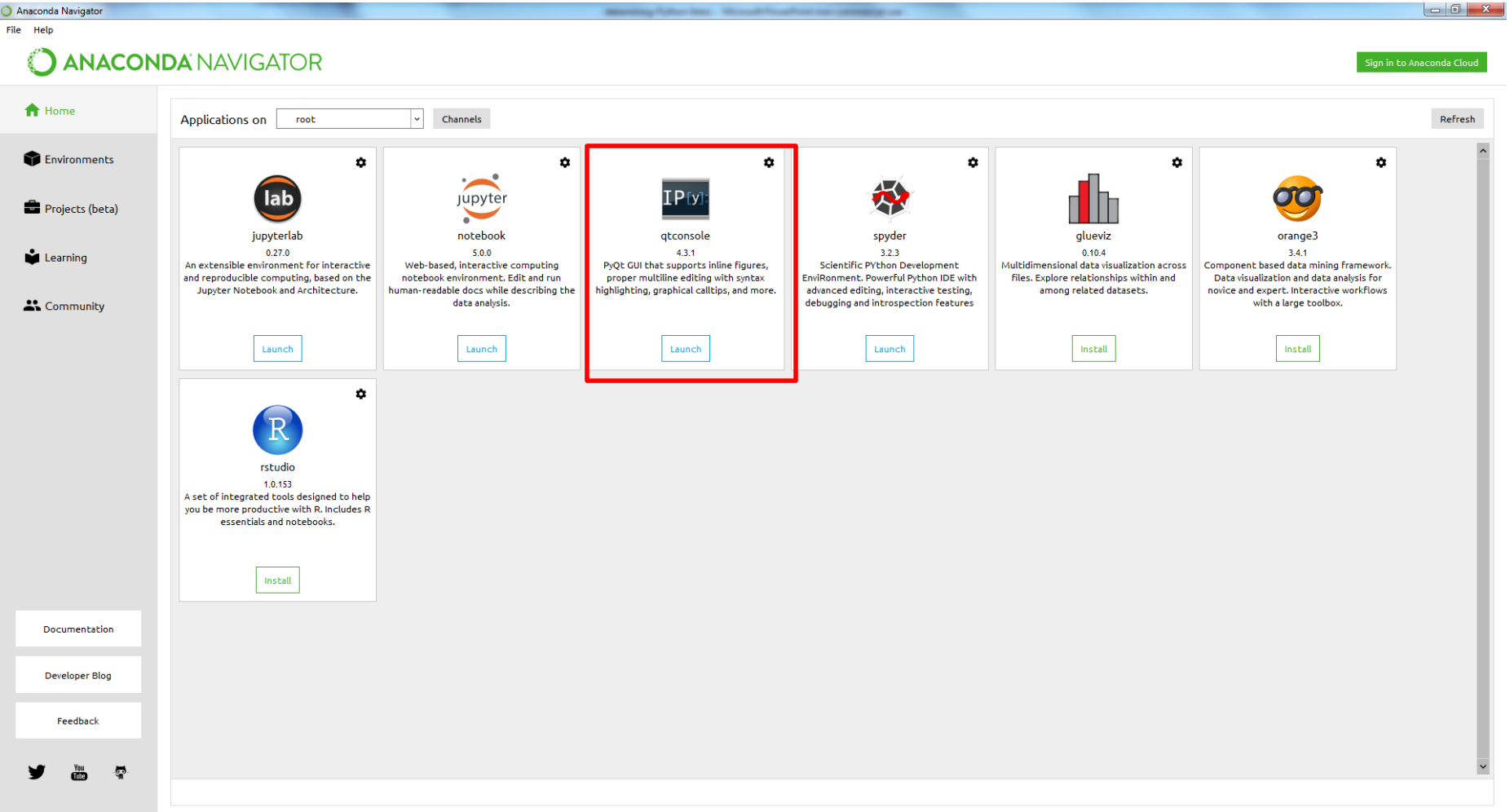
- There are tons of resources online for Python.
- For an introduction you can also look at the slides of the Introduction to Programming course by prof. N. Mamoulis

The Anaconda Navigator

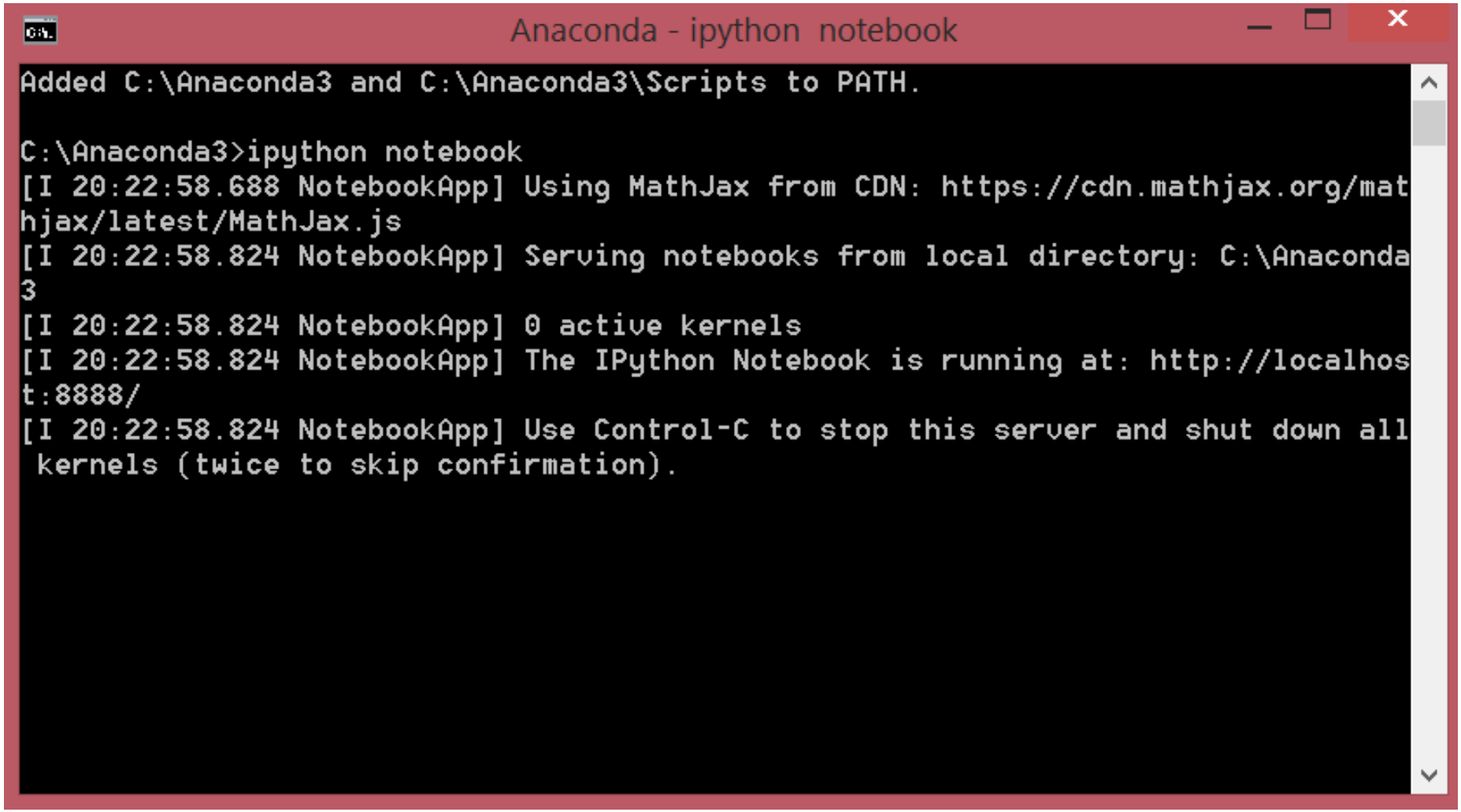



Home
localhost:8888/tree
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
DLLs
Doc
Examples
Lib
Library
Menu
Scripts
Tools
conda-meta
envs
include
info
libs
licenses
node-webkit
pkgs
share
tcl

The Anaconda Navigator



Starting iPython notebook



Changing the notebook default directory

- From the Anaconda terminal type the command:
  - `jupyter notebook --generate-config`

- This will generate `.jupiter/jupyter_notebook_config.py` file under your home directory.

- Find, un-comment and modify the line

`# c.NotebookApp.notebook_dir = ''` in the config file to point to the desired directory

Installing Packages

- You can install packages from the Anaconda terminal using the command:
  - `conda install <name of package>`
- For example, Seaborn is a package for Statistical Data Visualization.
  - `conda install seaborn`
- panda-datareader is a package for loading online datasets.
  - `conda install panda_datareader`