Installing Fundamental Software for Data Science Project

Sawitchaya Tippaya

Research Associate, School of Public Health

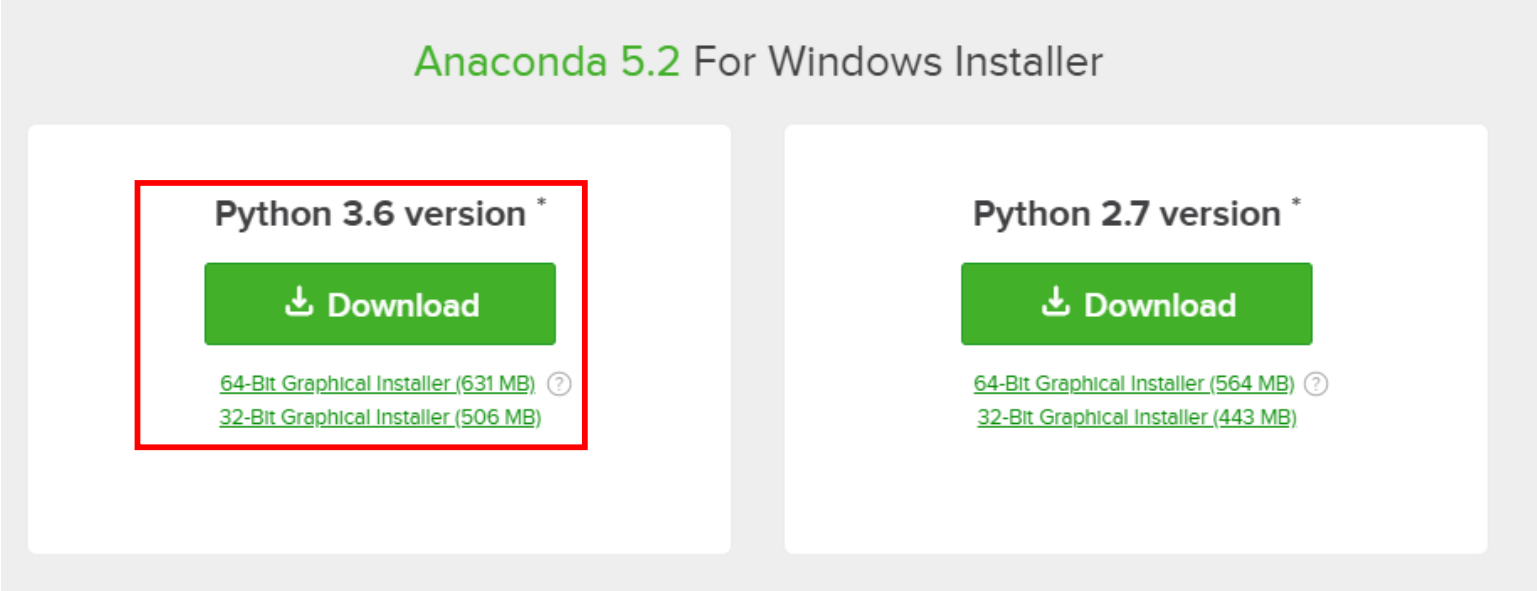
sawitchaya.tippaya@curtin.edu.au

Revision history

Date	Version	Comment
30/7/2018	0.1	Initial release.

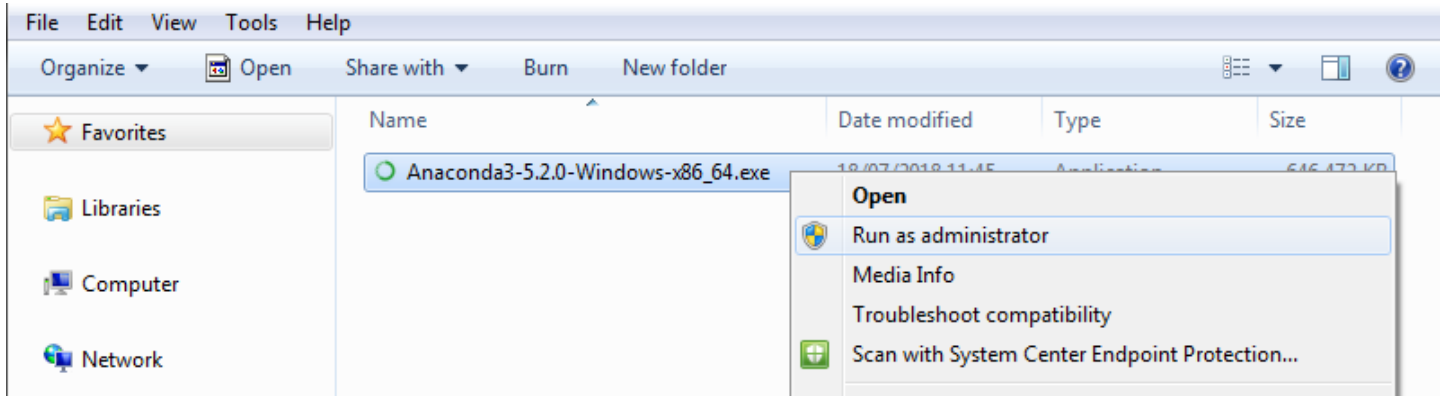
Anaconda Distribution

- Doing Python or R data analytics
- [Recommended] Download the latest version from
  - https://www.anaconda.com/download/
- Choose Python 3.6 version (32-bit or 64-bit depending on your machine)

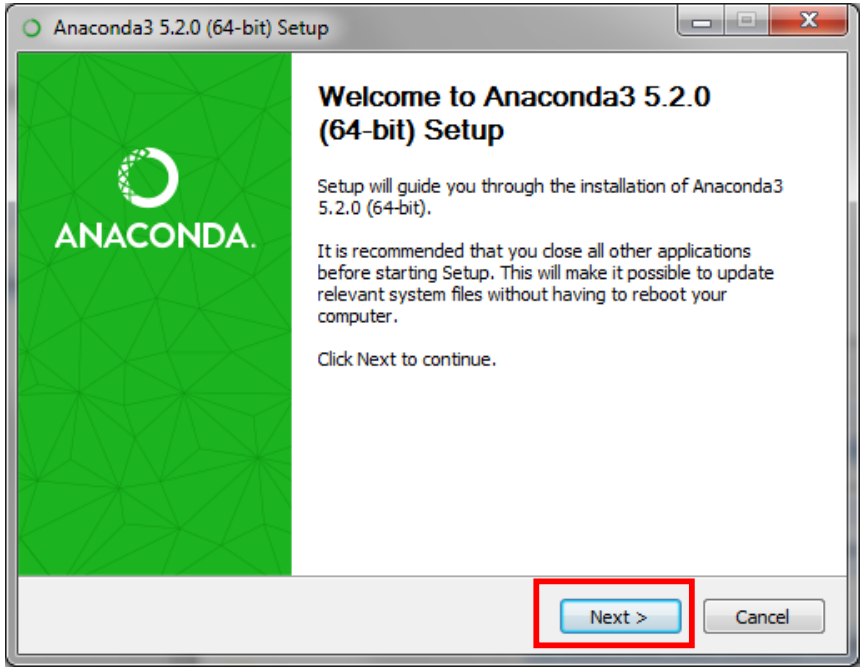


Anaconda Distribution (Cont'd)

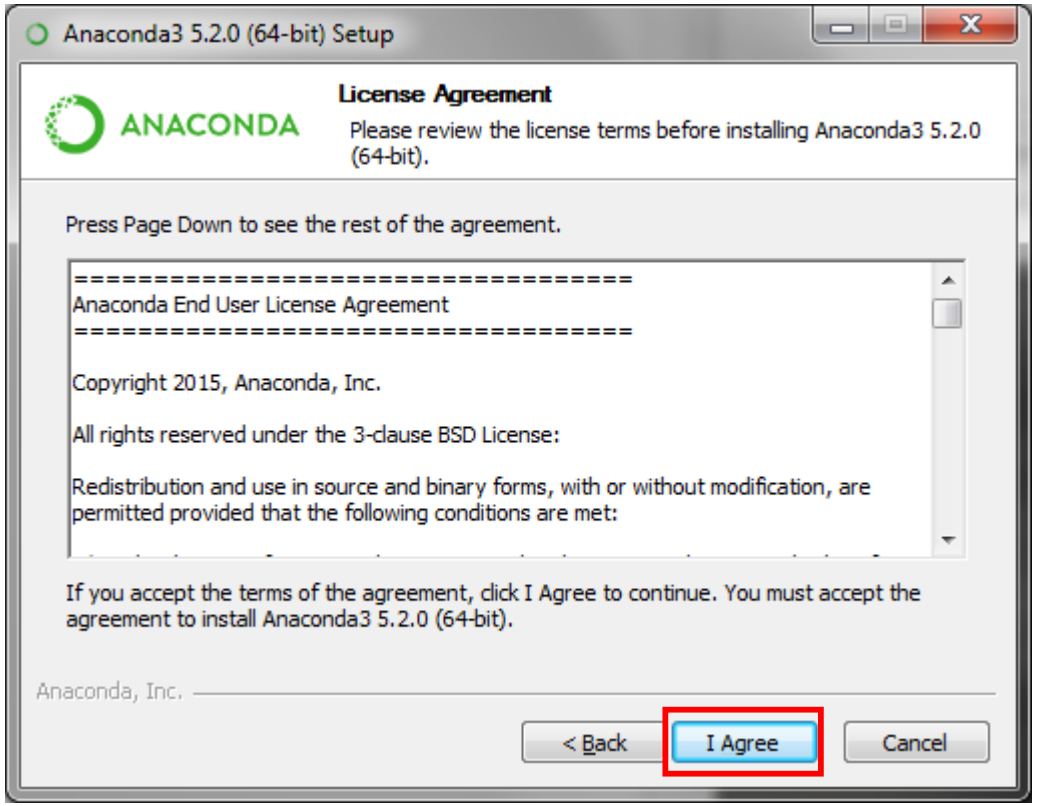
- Install Anaconda (select "Run as admin.")

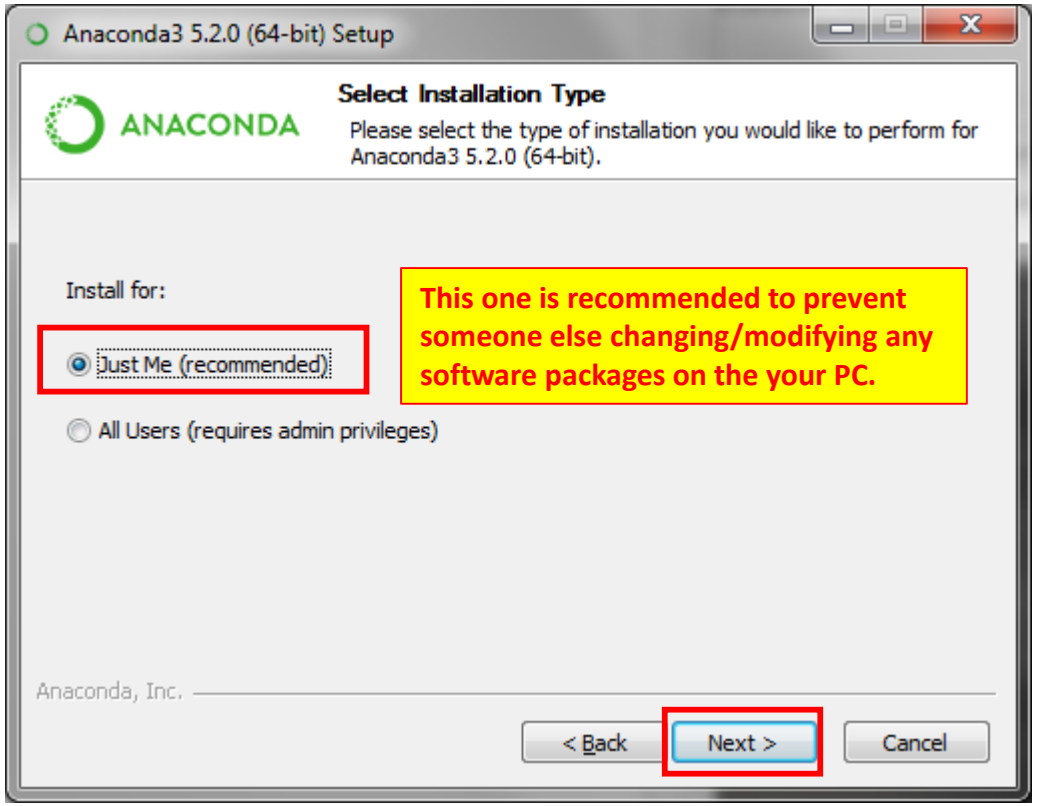


- Proceed the installation steps

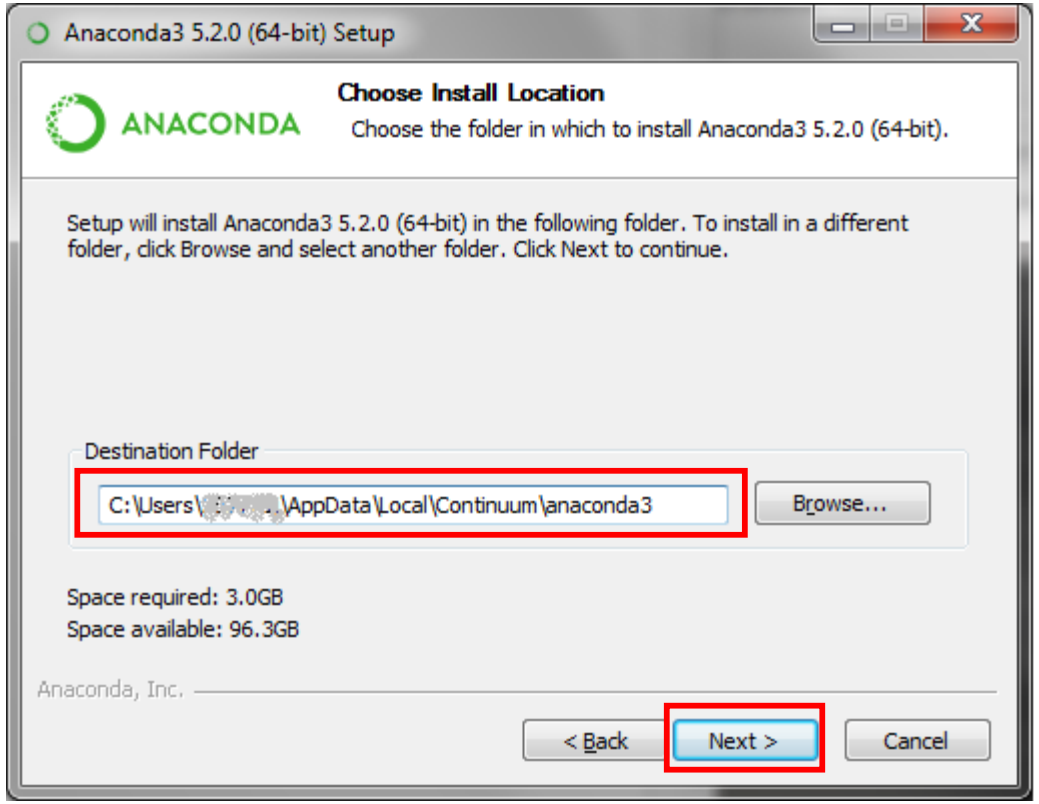


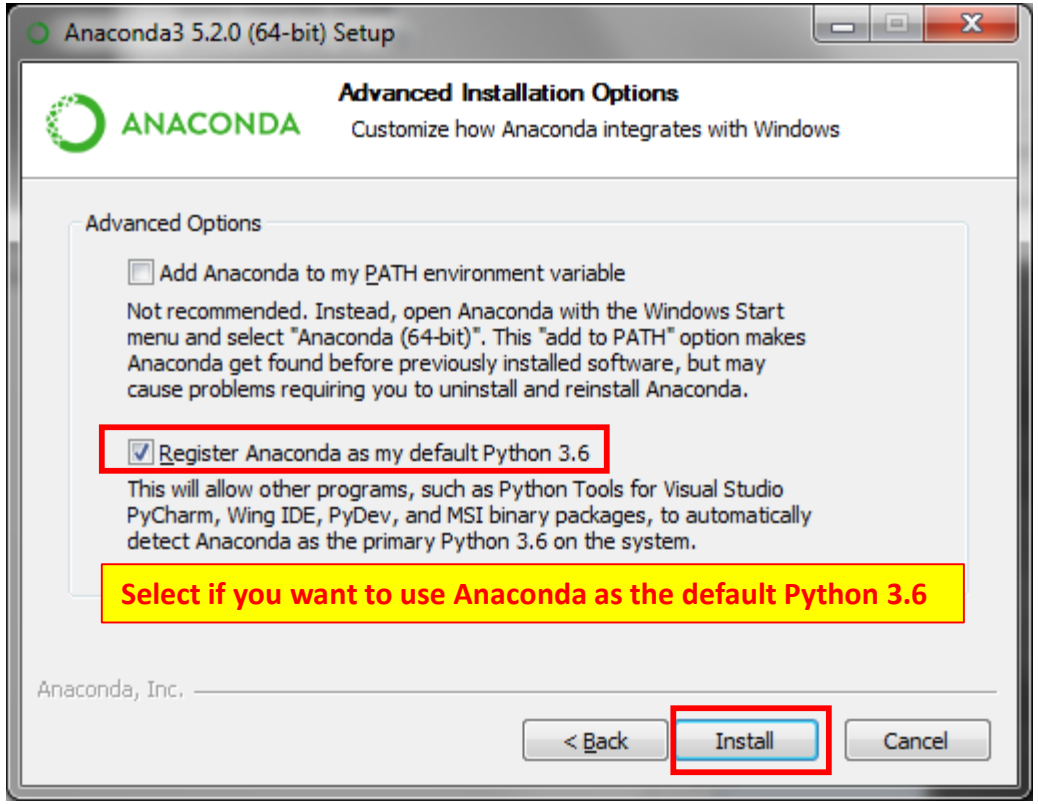
Anaconda Distribution (Cont'd)





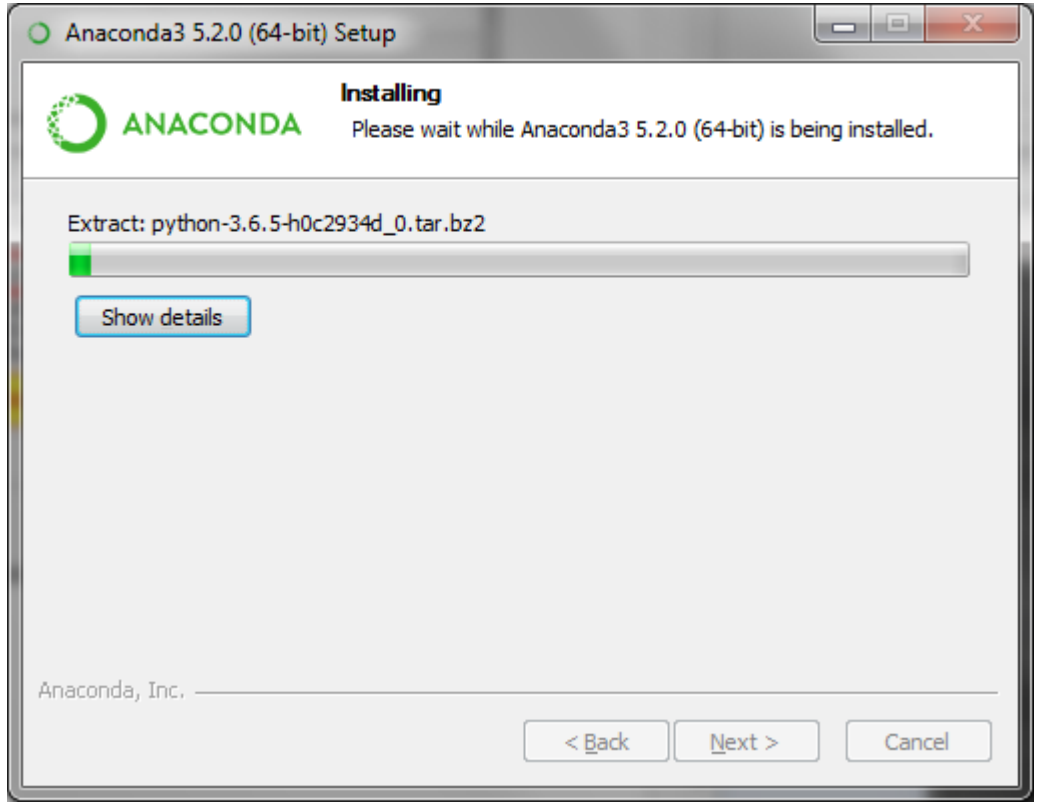
Anaconda Distribution (Cont'd)

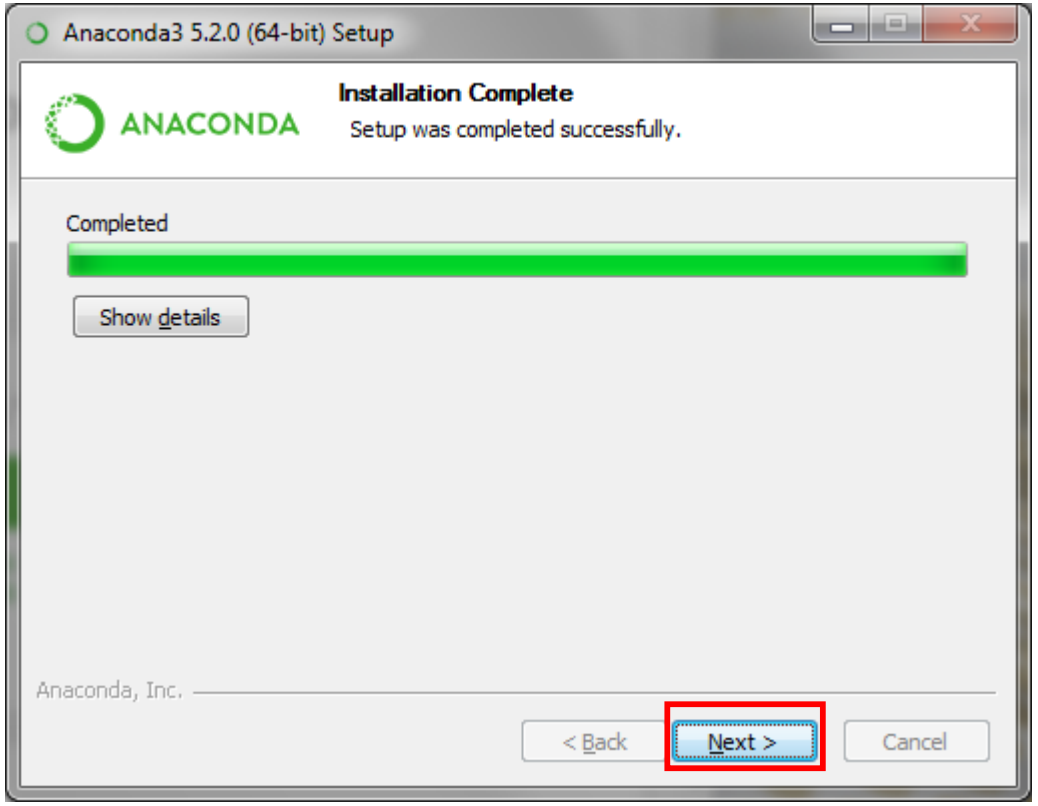




Note: you can use the default path or specify your own directory for Anaconda

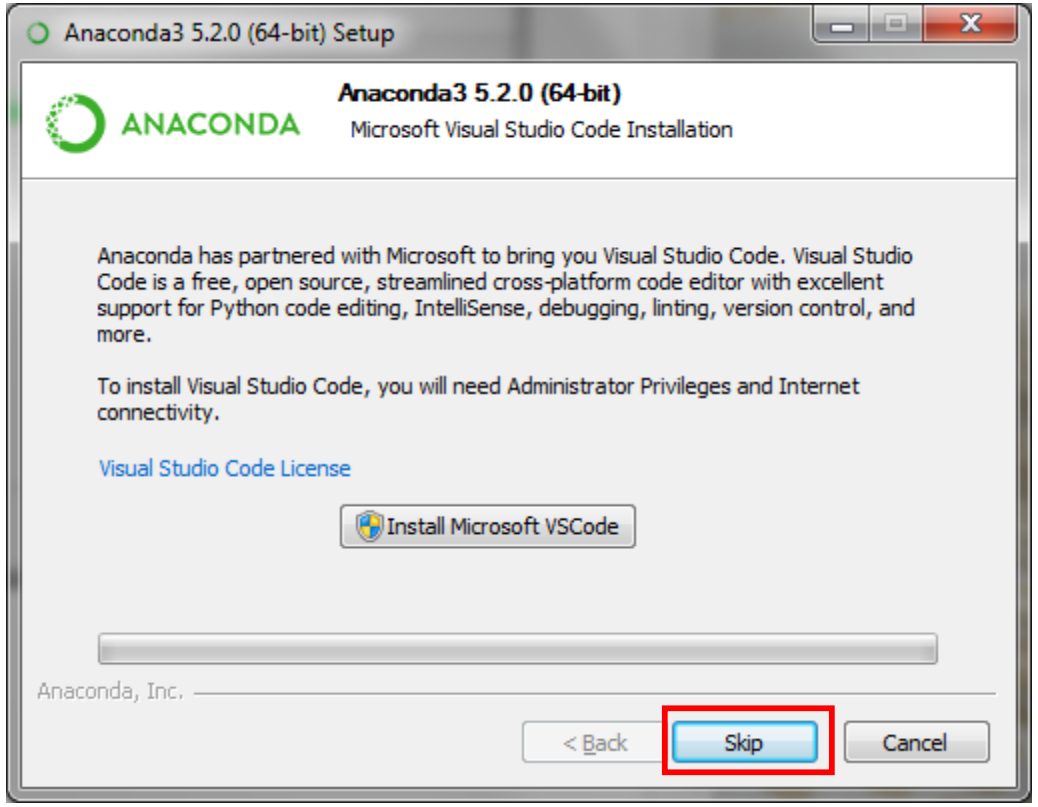
Anaconda Distribution (Cont'd)

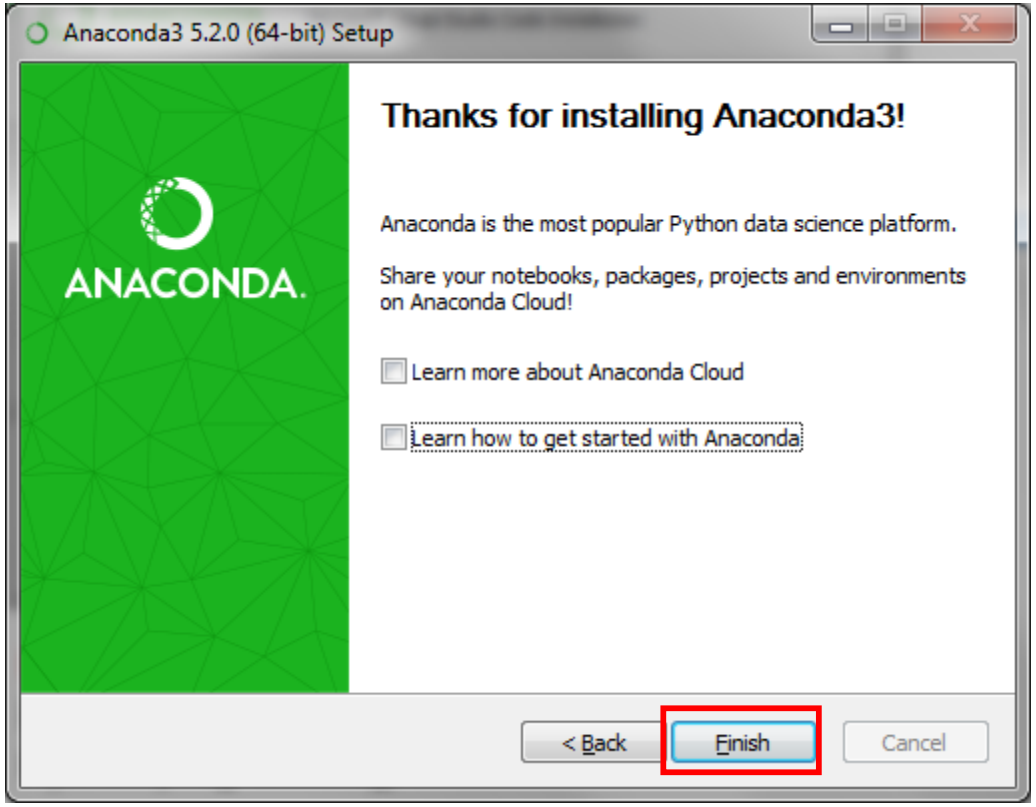




It will take a while … so let's have

Anaconda Distribution (Cont'd)

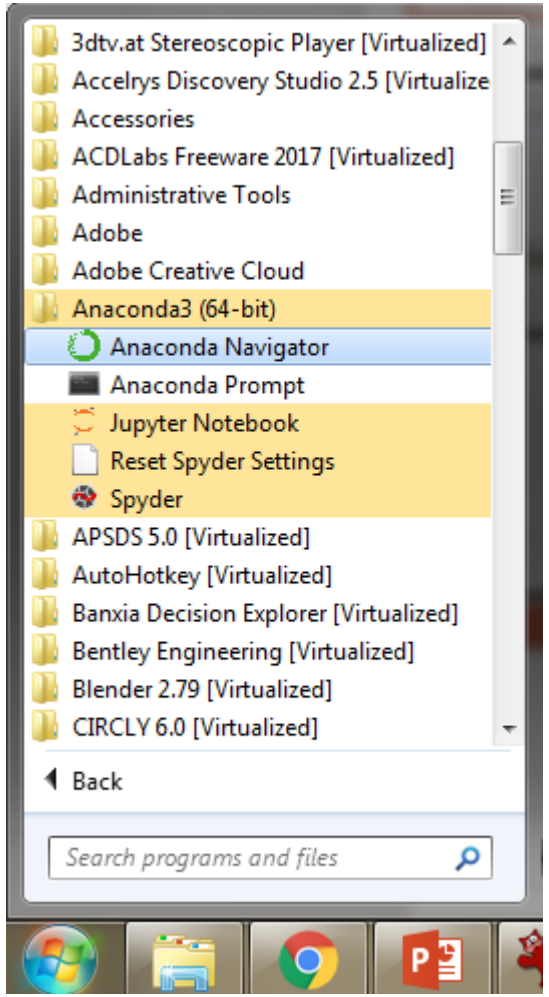


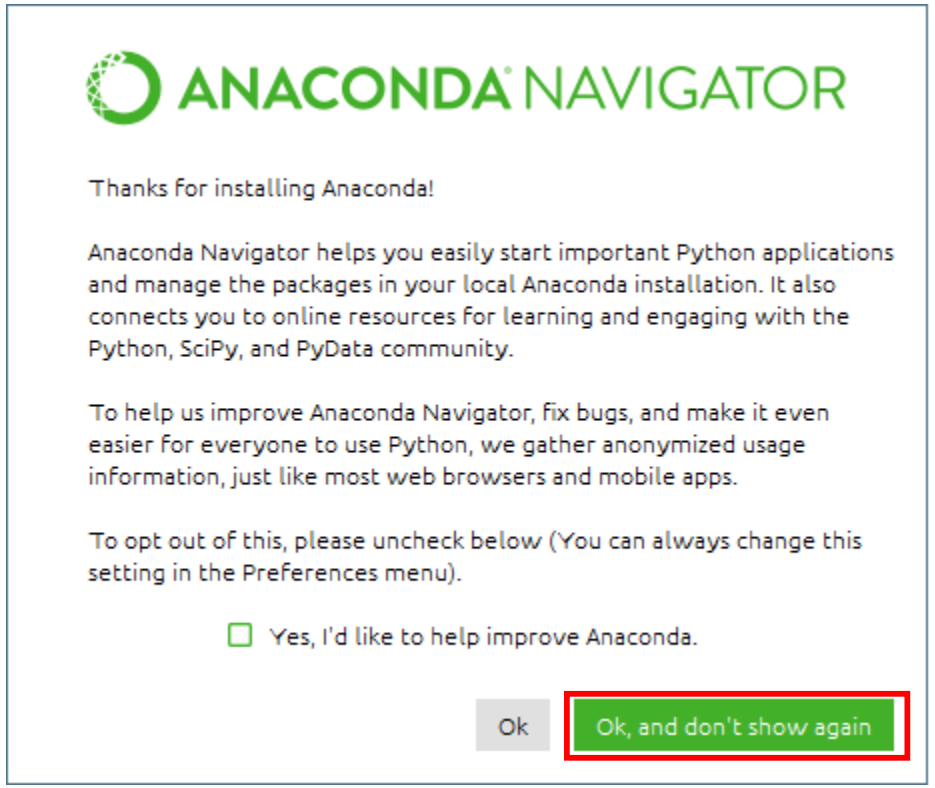


Note: you may install Microsoft VSCode is you are interested in using Visual Studio Code to develop Python programming.

Anaconda Distribution (Cont'd)

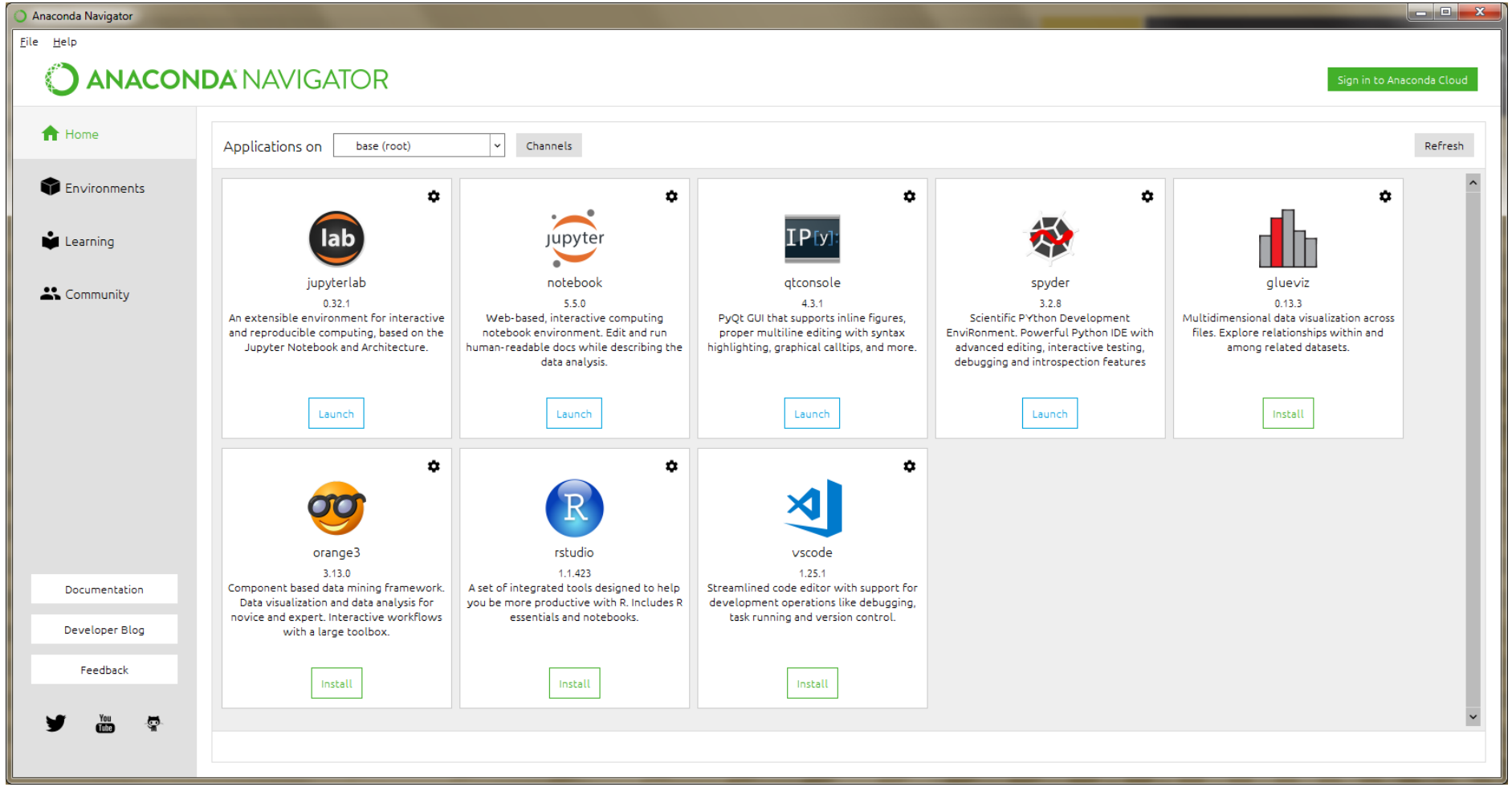
- Run the Anaconda Navigator
  - Start -> All programs -> Anaconda3 (64-bit)





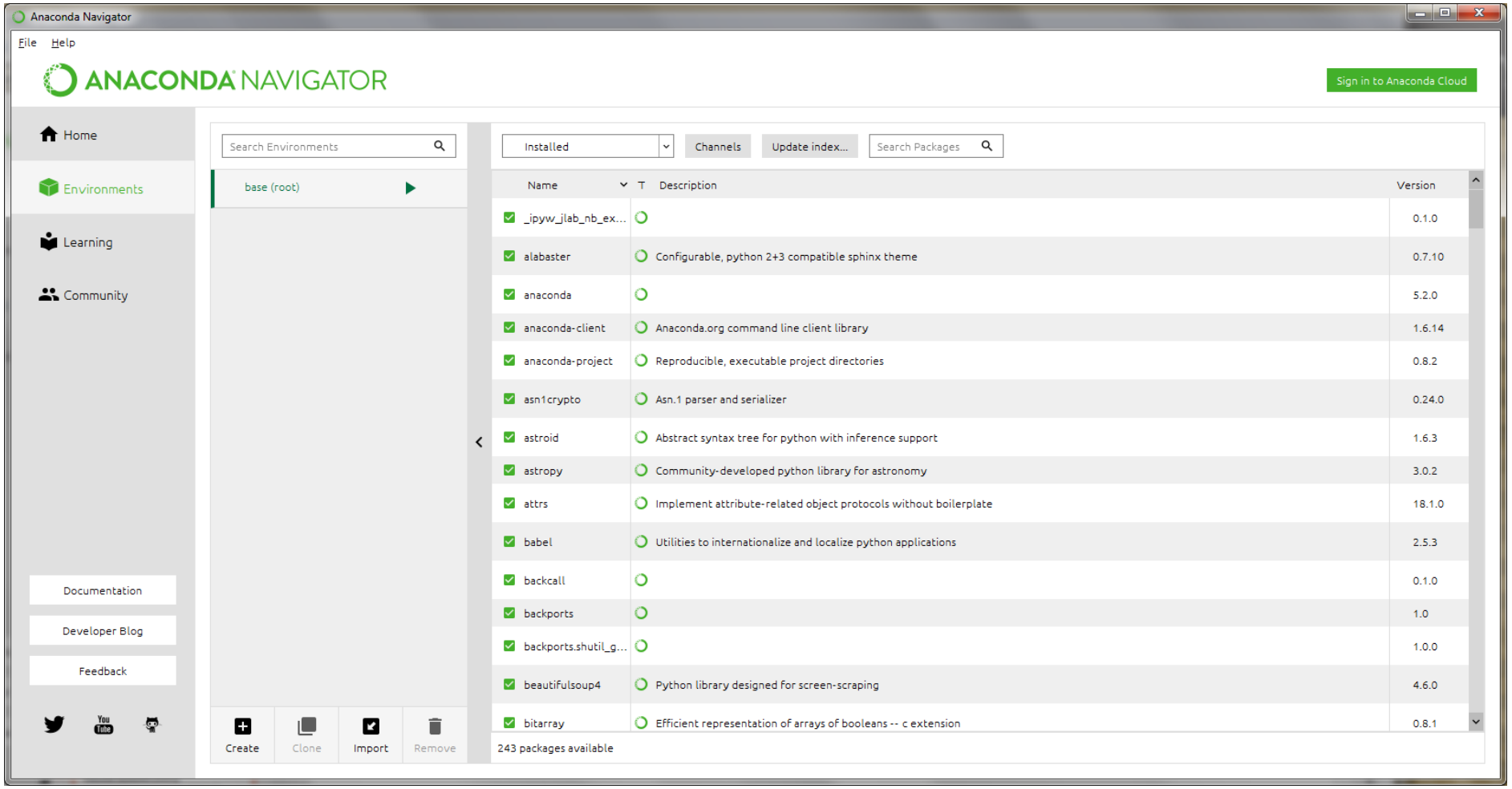
Anaconda Distribution (Cont'd)

- This is the console. You can install the additional packages/tools by clicking install.



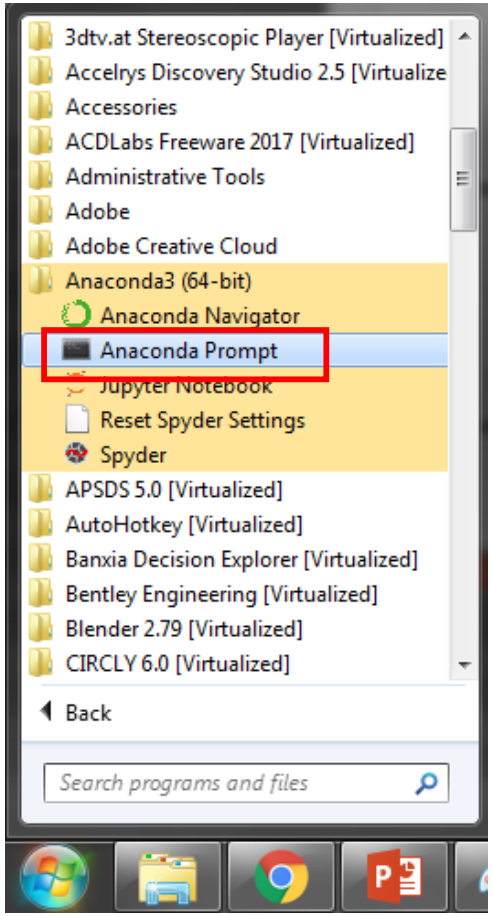
Anaconda Distribution (Cont'd)

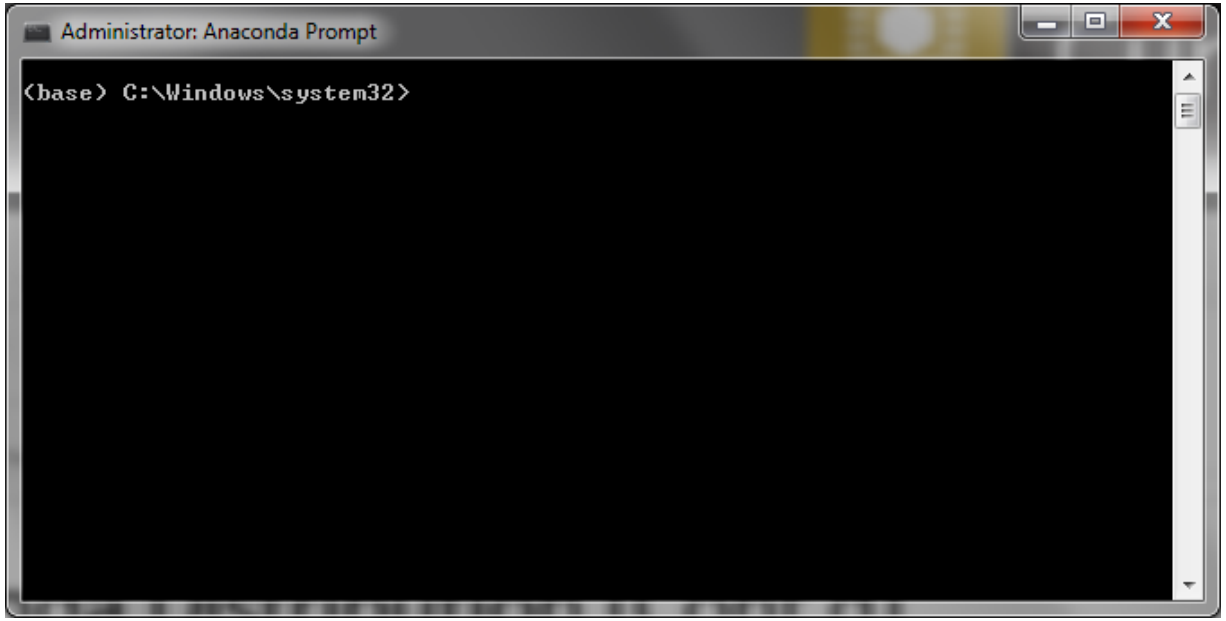
- Or browse the available packages under the Environments tab



Anaconda Distribution (Cont'd)

- Set up Windows PATH variable
  - Open the Anaconda prompt (Start -> All programs -> Anaconda3 (64-bit) -> Anaconda Prompt -> Right Click -> Run as Admin.





Anaconda Distribution (Cont'd)

- Set up Windows PATH variable
  - Type "where python" -> Enter
  - This is the path where python is installed.

```
(base) C:\Windows\system32>where python
C:\Users\[redacted]\AppData\Local\Continuum\anaconda3\python.exe
```

  - Type "where conda" -> Enter
  - This is the path where Anaconda is installed.

```
(base) C:\Windows\system32>where conda
C:\Users\[redacted]\AppData\Local\Continuum\anaconda3\Library\bin\conda.bat
C:\Users\[redacted]\AppData\Local\Continuum\anaconda3\Scripts\conda.exe
```

  - Type environment variables as below and press Enter:

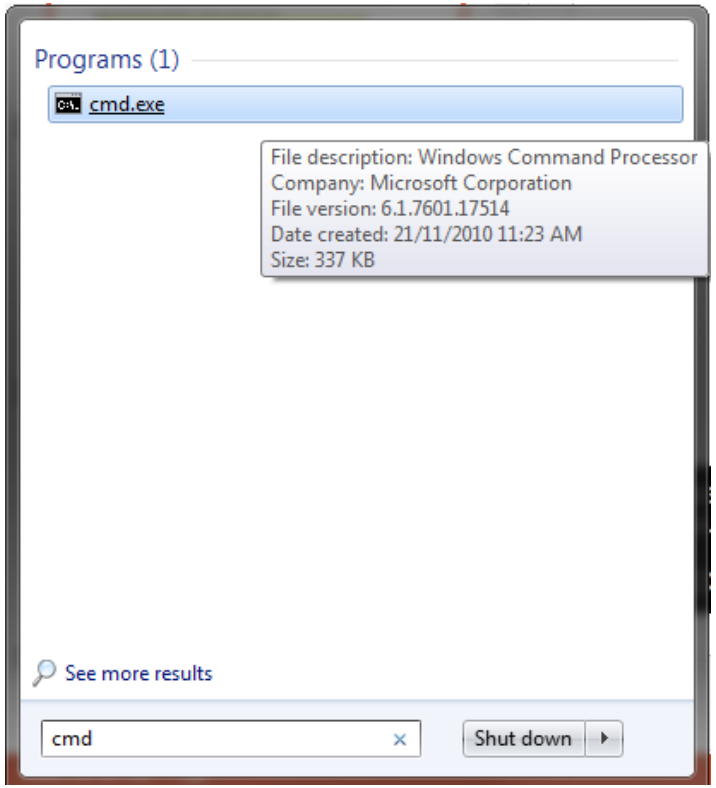
SETX PATH "%PATH%;C:\Users**yourID**\AppData\Local\Continuum\anaconda3\Scripts;
C:\Users**yourID**\AppData\Local\Continuum\anaconda3"

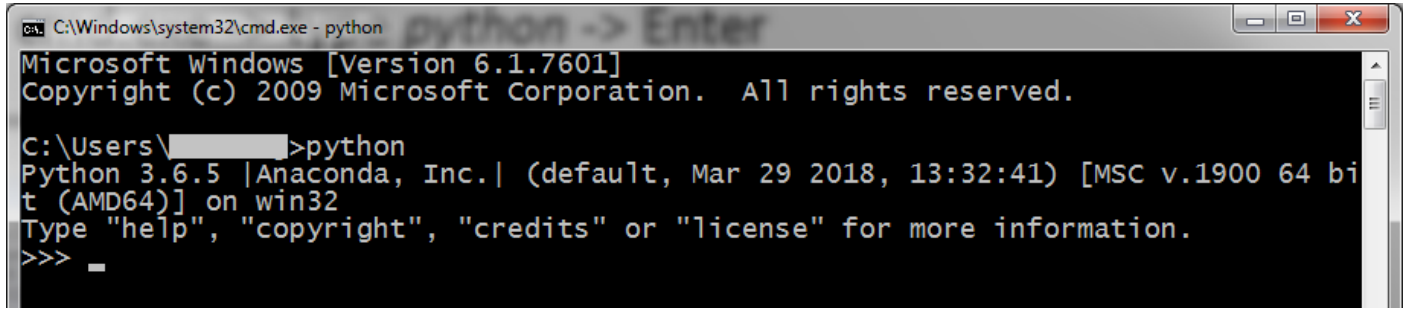
```
(base) C:\Windows\system32>SETX PATH "%PATH%;C:\Users\[redacted]\AppData\Local\Con
inuum\anaconda3\Scripts;C:\Users\[redacted]\AppData\Local\Continuum\anaconda3"

SUCCESS: Specified value was saved.
```

Note: Generally, **yourID** will be your Curtin's staff ID

Anaconda Distribution (Cont'd)

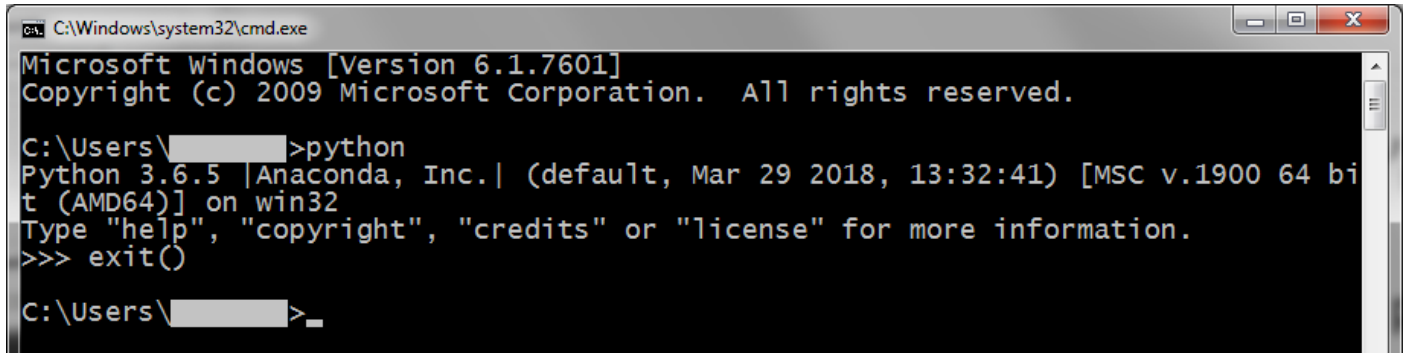
- Testing command via Windows command prompt
- Start -> Run -> cmd.exe -> type *python* -> Enter





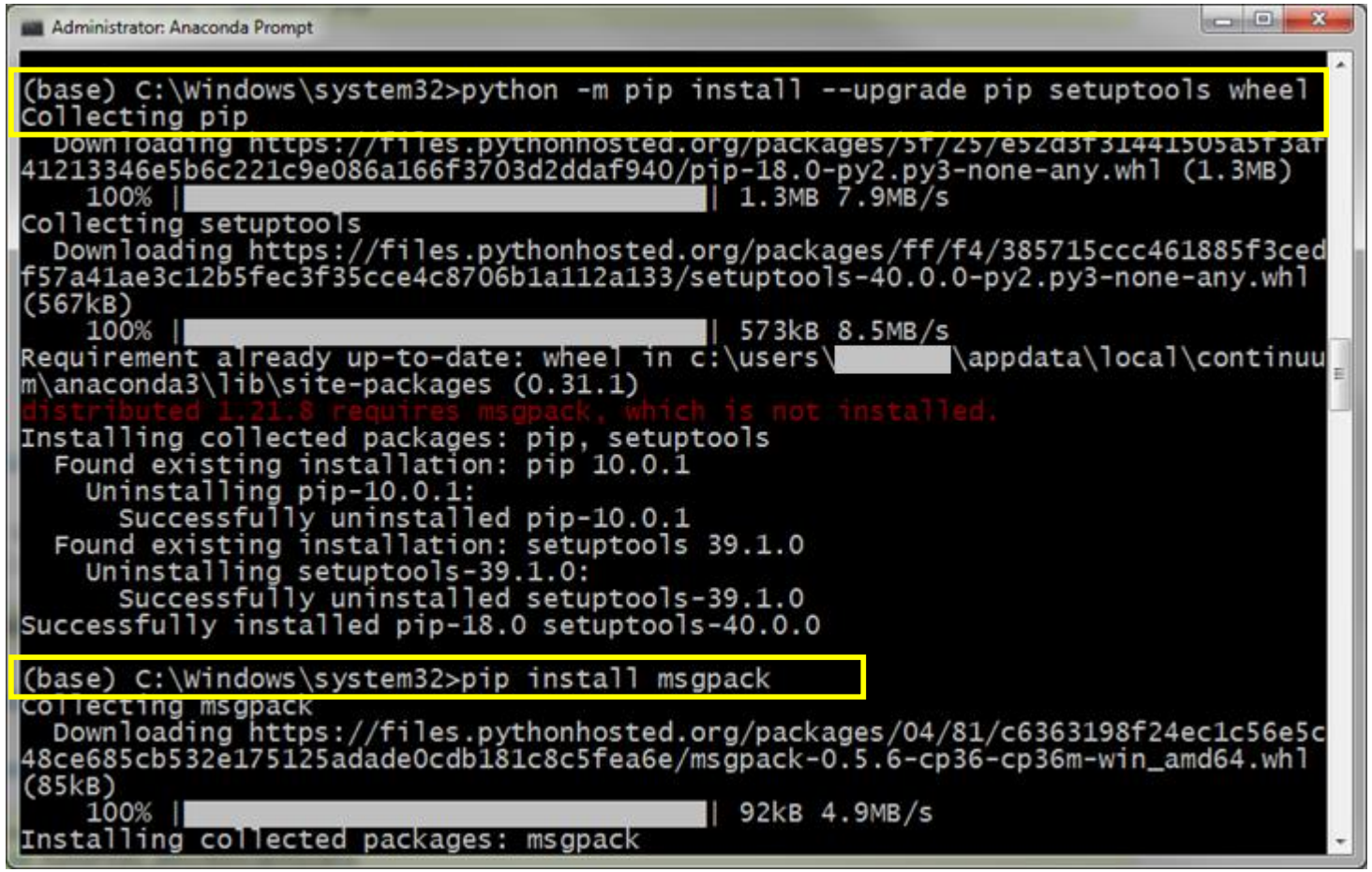
Successfully installed!

To exit from Python mode: type exit() then press enter



Updates pip to the latest version

- Ensure pip, setuptools, and wheel are up to date
  - python -m pip install --upgrade pip setuptools wheel
- Install msgpack
  - pip install msgpack



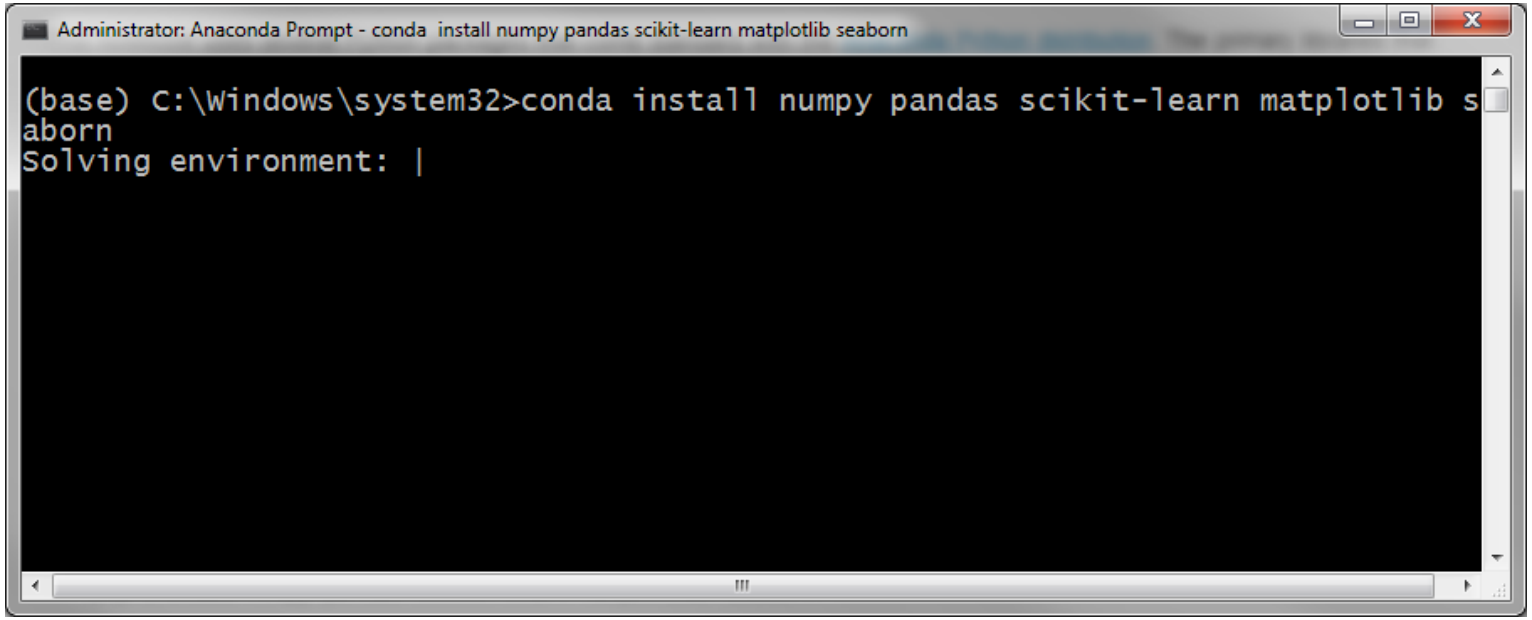
Install Python Packages

- Recommended Python packages:
  - Core Libraries: Numpy, SciPy, Pandas,
  - Visualisation: Matplotlib, Seaborn, Bokeh, Plotly
  - Machine Learning: SciKit-Learn, Deep Learning — Keras / TensorFlow / Theano
  - Data Mining & Statistics: Scrapy, Statsmodels
  - Error Analysis: Lime
  - Misc.: PrettyTable
- Two options to install packages:
  - 1) Using Anaconda Navigator
  - 2) Command line in Anaconda prompt with conda or pip (depending on available options from developers.

Source: https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7

Install Python Packages (Con't)

- Example: To ensure we have Numpy, Pandas, SciKit-Learn, Matplotlib, Seaborn packages using **conda**
  - Start -> All programs -> Anaconda3 (64-bit) -> Anaconda Prompt -> Right Click -> Run as Admin
  - Type conda install numpy pandas scikit-learn matplotlib seaborn
  - Press Enter



Install Python Packages (Con't)

- Sometimes, it also requires to update conda.
- Follow all steps and details…

Administrator: Anaconda Prompt - conda install numpy pandas scikit-learn matplotlib seaborn

```
(base) C:\Windows\system32>conda install numpy pandas scikit-learn matplotlib se
aborn
Solving environment: done

## Package Plan ##

  environment location: C:\Users\[redacted]\AppData\Local\Continuum\anaconda3

  added / updated specs:
    - matplotlib
    - numpy
    - pandas
    - scikit-learn
    - seaborn


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    conda-4.5.8                |           py36_0         1.0 MB

The following packages will be UPDATED:

    conda: 4.5.4-py36_0 --> 4.5.8-py36_0

Proceed ([y]/n)? y_
```

Type y then press enter

Administrator: Anaconda Prompt

```
  environment location: C:\Users\[redacted]\AppData\Local\Continuum\anaconda3

  added / updated specs:
    - matplotlib
    - numpy
    - pandas
    - scikit-learn
    - seaborn


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    conda-4.5.8                |           py36_0         1.0 MB

The following packages will be UPDATED:

    conda: 4.5.4-py36_0 --> 4.5.8-py36_0

Proceed ([y]/n)? y


Downloading and Extracting Packages
conda-4.5.8          |  1.0 MB | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

Conda is updated and Anaconda has python packages that we need.

Install Python Packages (Con't)

- Let's install the new packages such as PrettyTable

Administrator: Anaconda Prompt

```
(base) C:\Windows\system32>pip install PrettyTable
Collecting PrettyTable
  Downloading https://files.pythonhosted.org/packages/ef/30/4b0746848746ed5941f0
52479e7c23d2b56d174b82f4fd34a25e389831f5/prettytable-0.7.2.tar.bz2
Building wheels for collected packages: PrettyTable
  Running setup.py bdist_wheel for PrettyTable ... done
  Stored in directory: C:\Users\[redacted]\AppData\Local\pip\Cache\wheels\80\34\1c\
3967380d9676d162cb59513bd9dc862d0584e045a162095606
Successfully built PrettyTable
Installing collected packages: PrettyTable
Successfully installed PrettyTable-0.7.2

(base) C:\Windows\system32>_
```

All done!

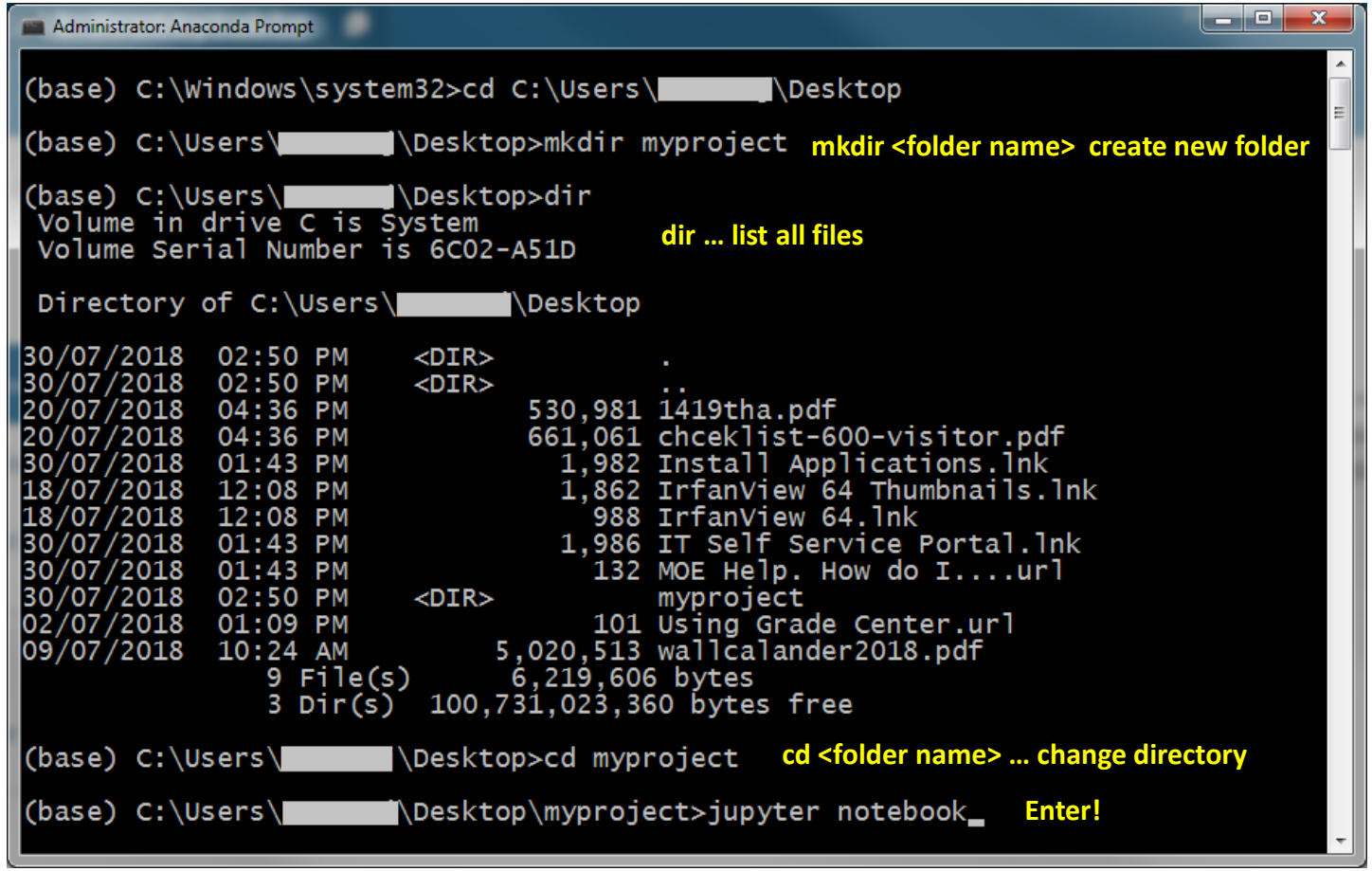
Jupyter Notebook

- An open-source web application to create and share codes and documents.
- Included in Anaconda distribution.
- You can code, run, and visualise all results in one environment.
- It is powerful tool to develop a reproducible research.
- Recommended to install the extensions to enhance the productivity. Run these two commands in Anadaconda prompt
  - pip install jupyter_contrib_nbextensions
  - jupyter contrib nbextension install --user

More useful details: https://www.analyticsvidhya.com/blog/2018/05/starters-guide-jupyter-notebook/

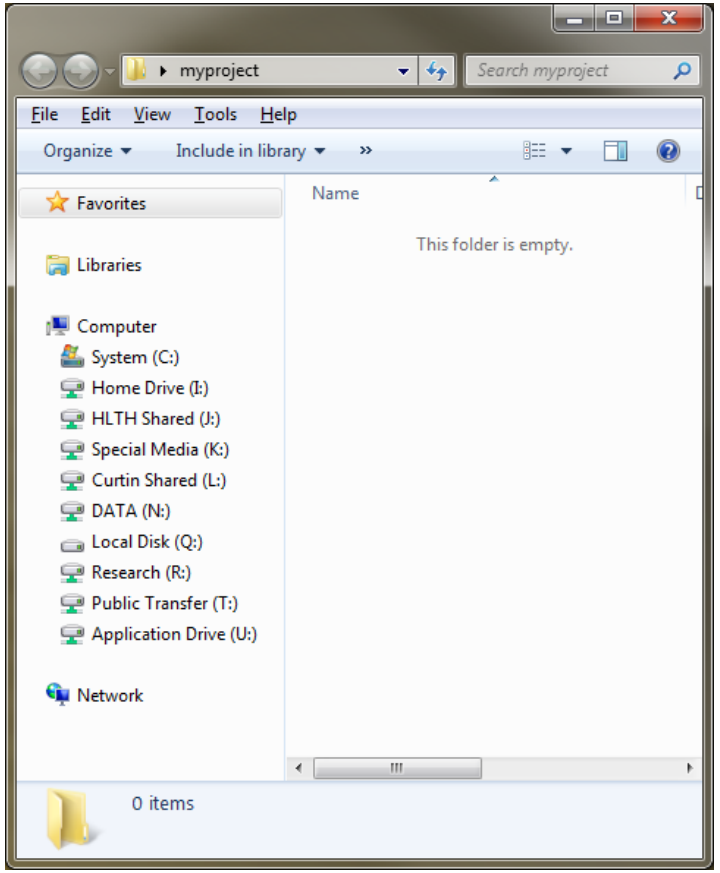
Example Project using Jupyter Notebook

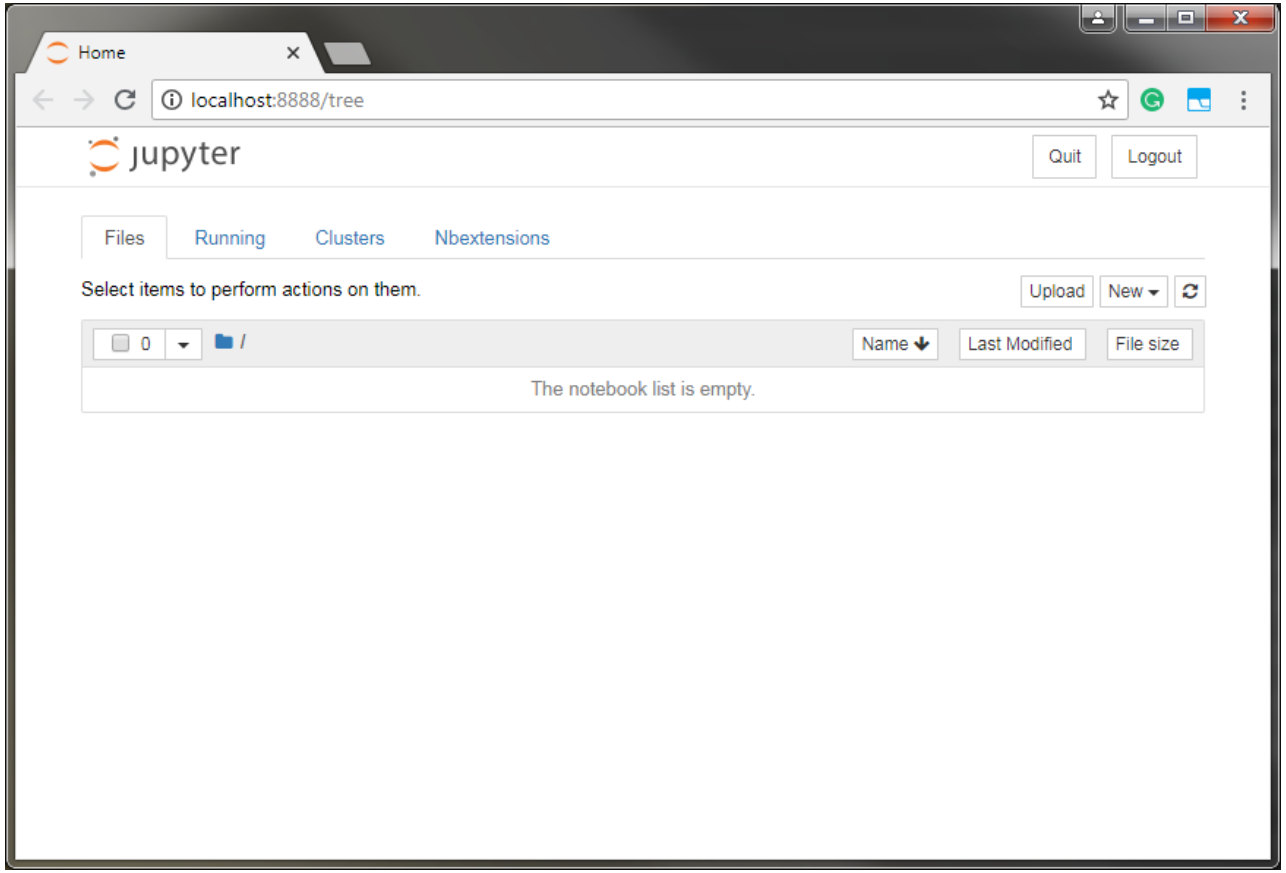
- Start the Anaconda Prompt
- Create new project directory and run Jupyter Notebook from this folder



Example Project using Jupyter Notebook

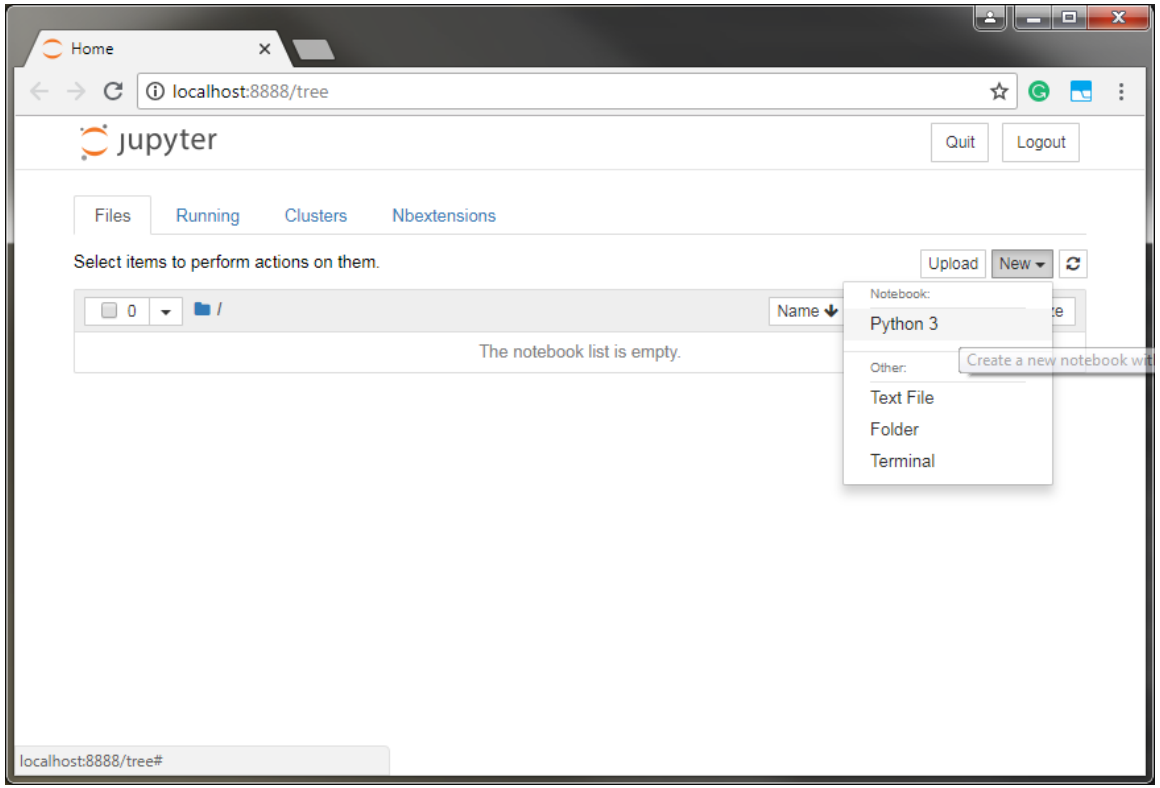
- Now we have created the ‘myproject’ working directory and run the Jupyter notebook from this folder.

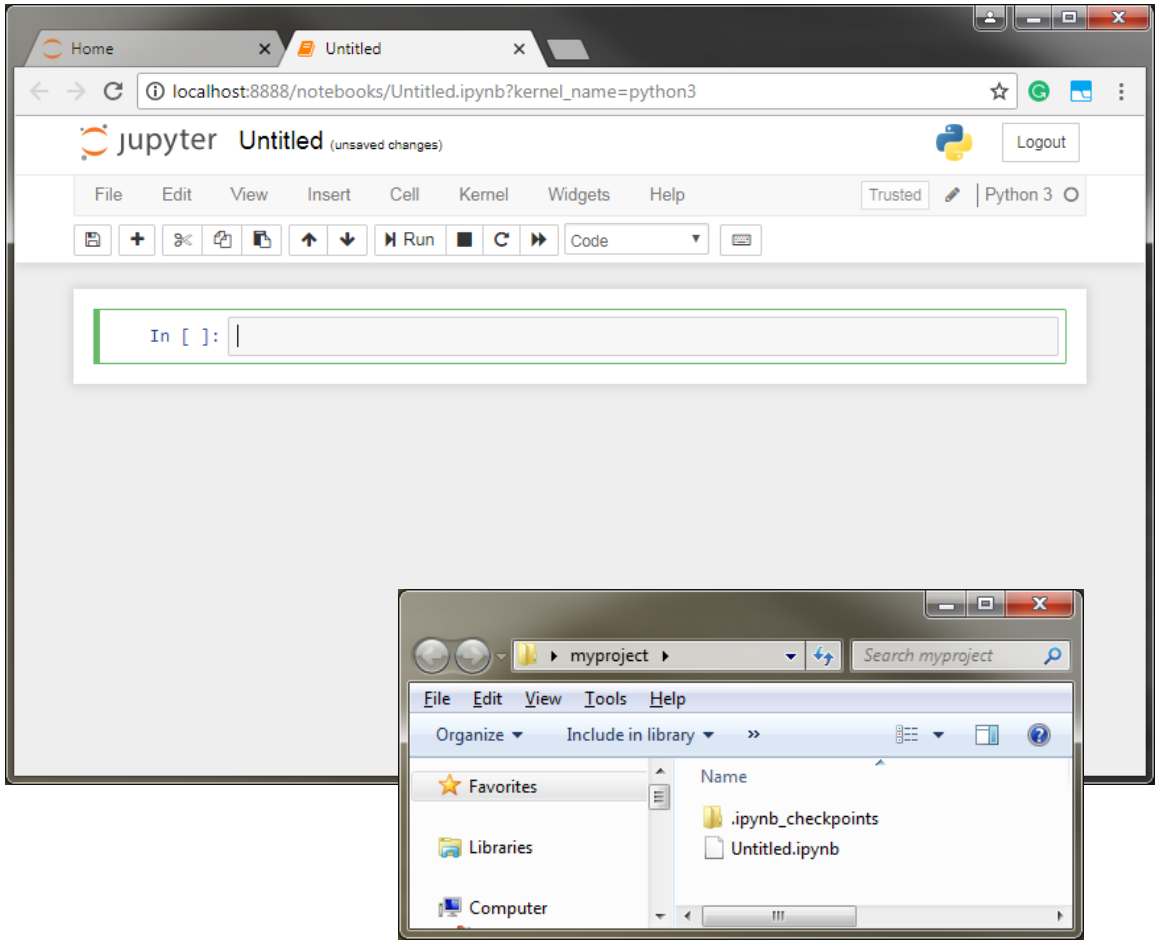




Example Project using Jupyter Notebook

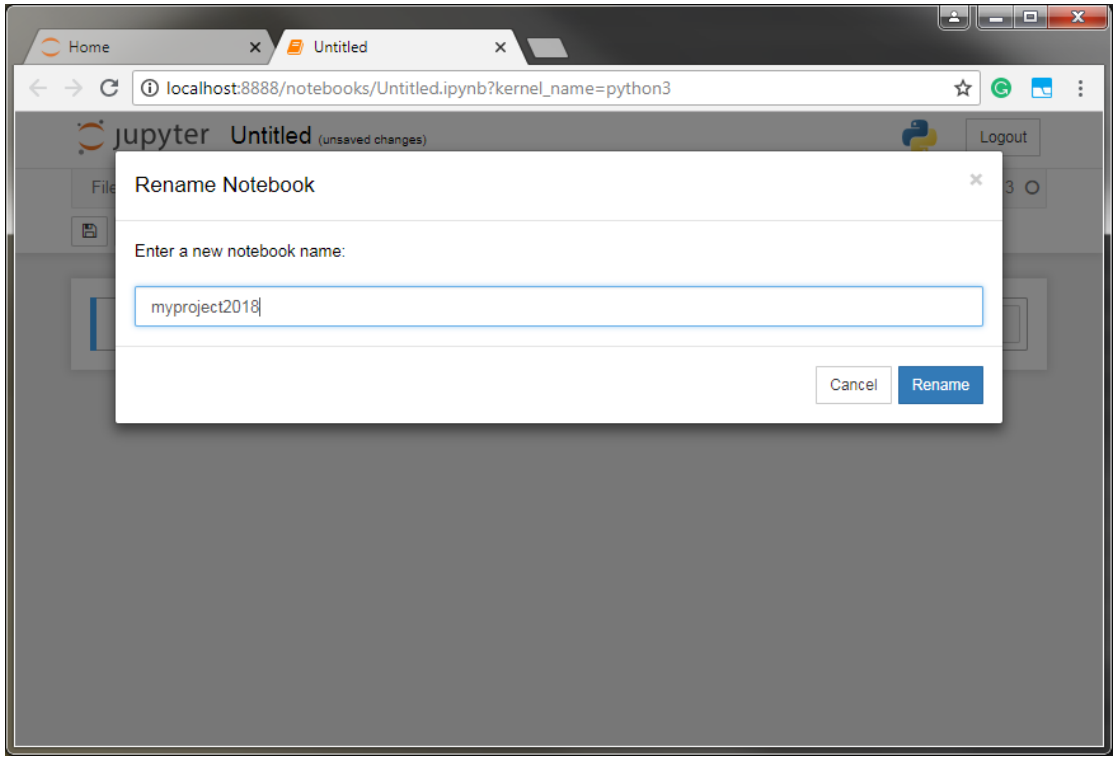
- Create new empty notebook

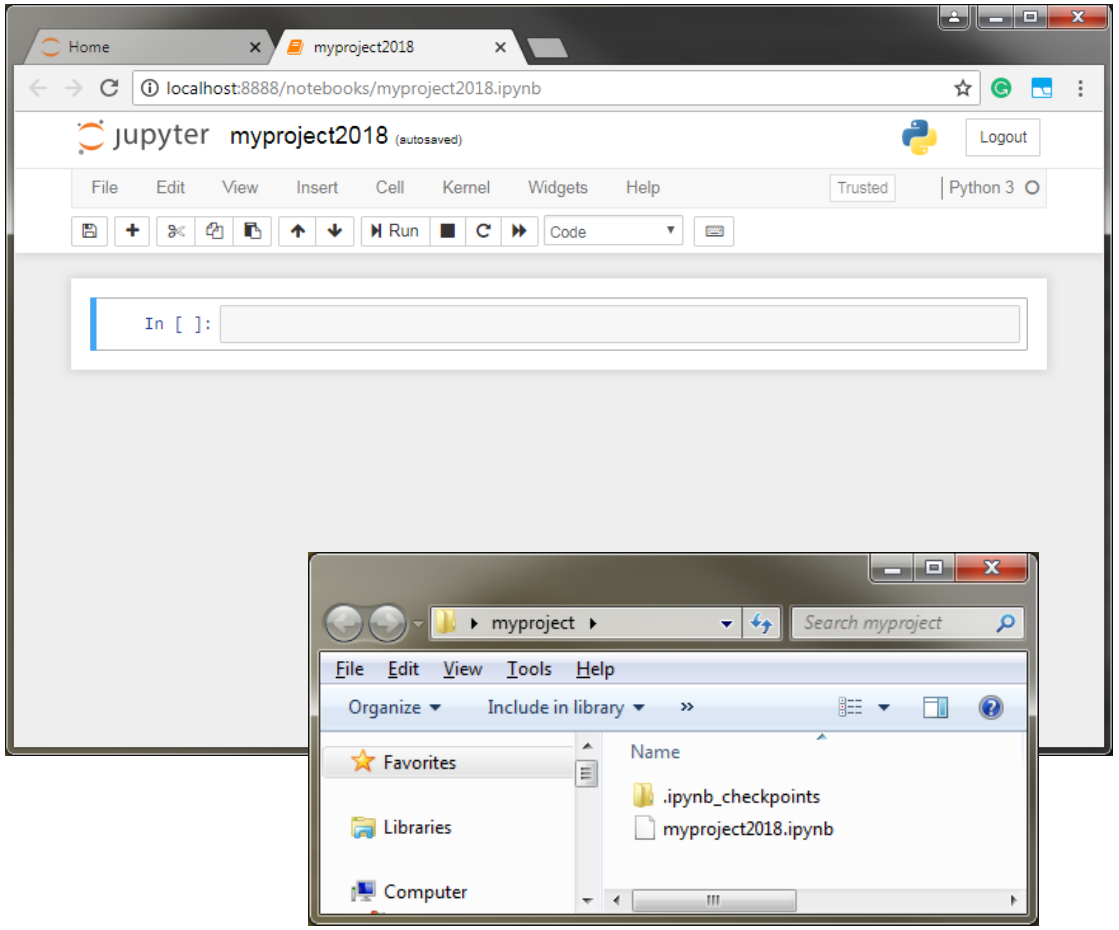




Example Project using Jupyter Notebook

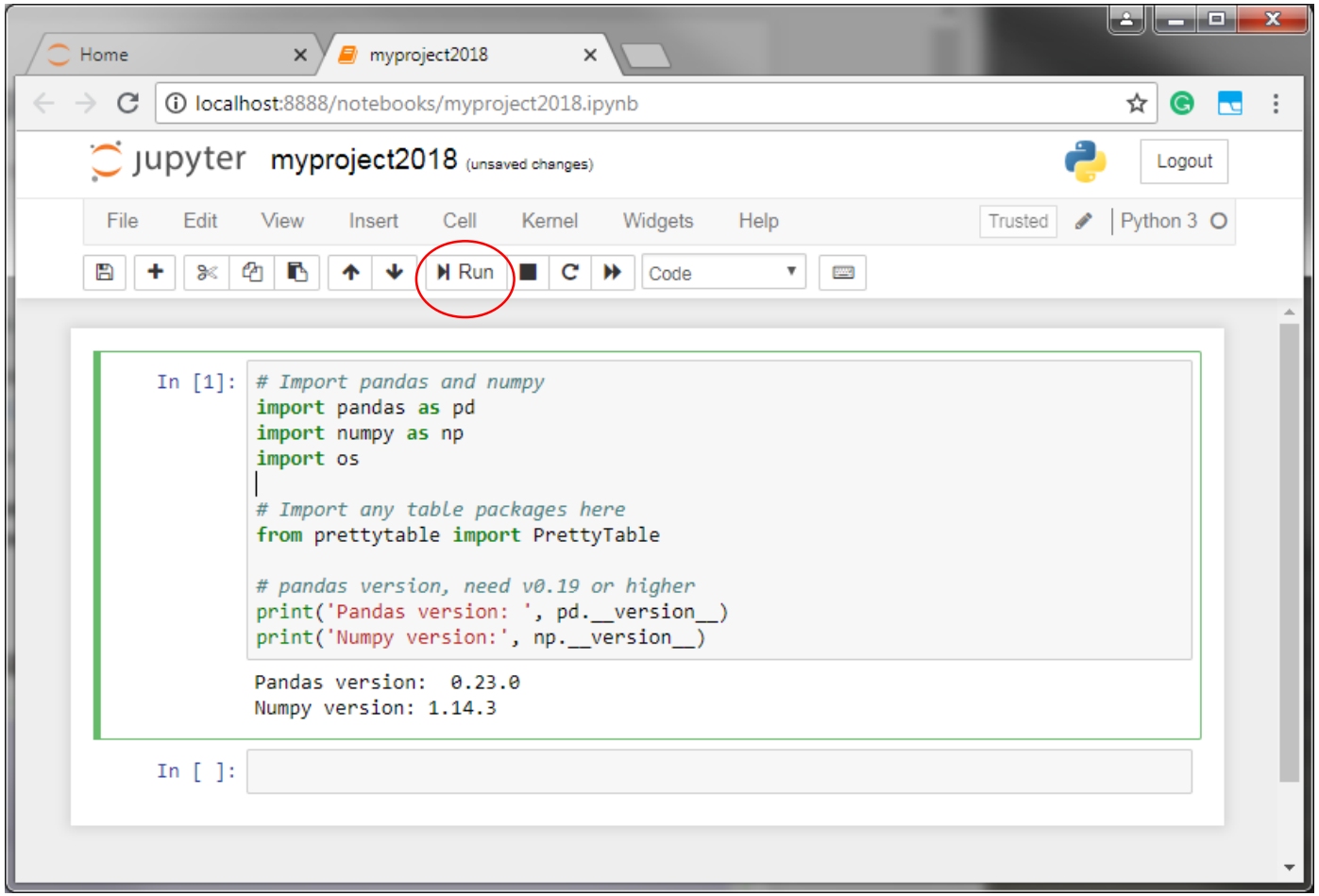
- Rename … click the filename next to Jupyter logo.





Example Project using Jupyter Notebook

- Load packages, print python version, and run the cell.



Basic Python Tutorial

- https://wiki.python.org/moin/BeginnersGuide/Programmers
- https://chrisalbon.com/#python
- Debugging: https://stackoverflow.com/questions/tagged/python

- Workshop by Curtin Institute for Computation (CIC) <- Recommended
  - http://computation.curtin.edu.au/events/training/
  - Software Carpentry and Data Carpentry
  - Learn Python, R, and Git version control

- Hacky Hour by CIC
  - http://computation.curtin.edu.au/events/hacky-hour/

Other recommended software

- Git: a version control system for tracking changes in files and coordinating work on those files among multiple people
- Google Chrome: recommended browser for Jupyter notebook
  - Extension:
- Notepad++ (available on all PCs in Curtin's network)

Learn how to make your first Machine Learning classifier in Scikit-learn (Python)

https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a

Some useful resources for data science project

- Github, Coursera
- http://www.andrewng.org/
- https://www.analyticsvidhya.com
- https://chrisalbon.com/
- http://python-data-science.readthedocs.io
- https://towardsdatascience.com
- CIC: http://computation.curtin.edu.au/
- and definitely… Google!

Cheat Sheet

- Conda: https://conda.io/docs/_downloads/conda-cheatsheet.pdf
- Python [Beginner]: https://github.com/ehmatthes/pcc/releases/download/v1.0.0/beginners_python_cheat_sheet_pcc_all.pdf
- Jupyter Notebook: https://www.datacamp.com/community/blog/jupyter-notebook-cheat-sheet
- Scikit-learn: http://scikit-learn.org/stable/themes/scikit-learn/static/ML_MAPS_README.html