

Installing Fundamental Software for Data Science Project

Sawitchaya Tippaya

Research Associate, School of Public Health

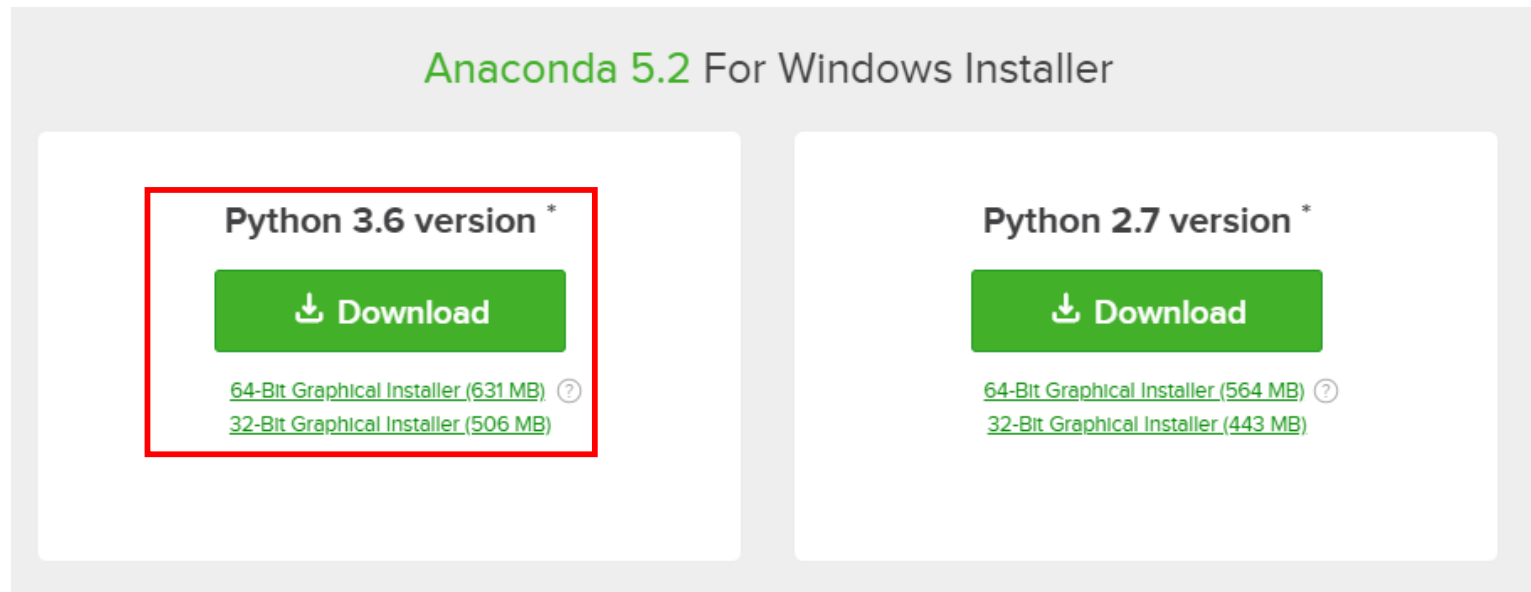
sawitchaya.tippaya@curtin.edu.au

Revision history

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

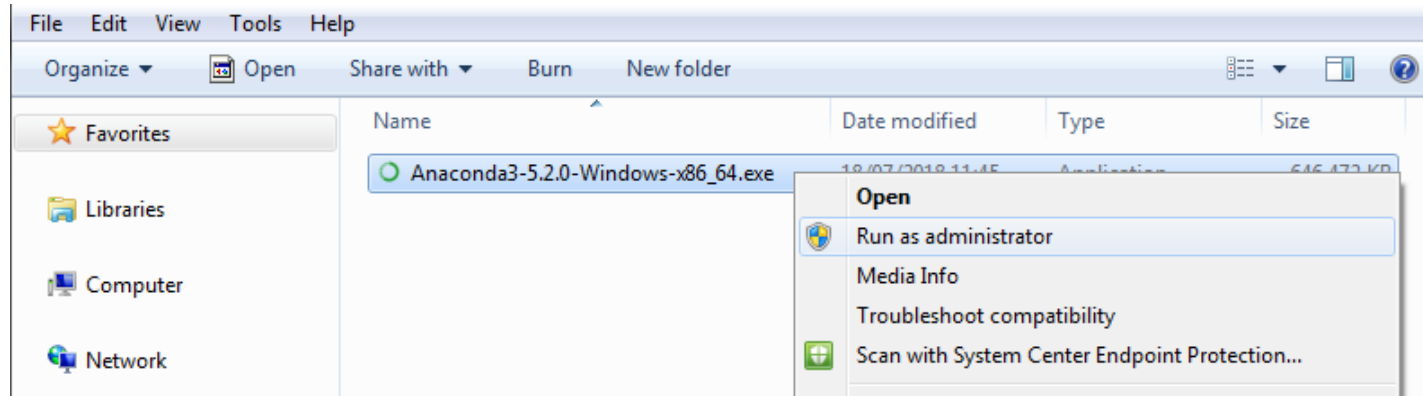
Anaconda Distribution

- 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)

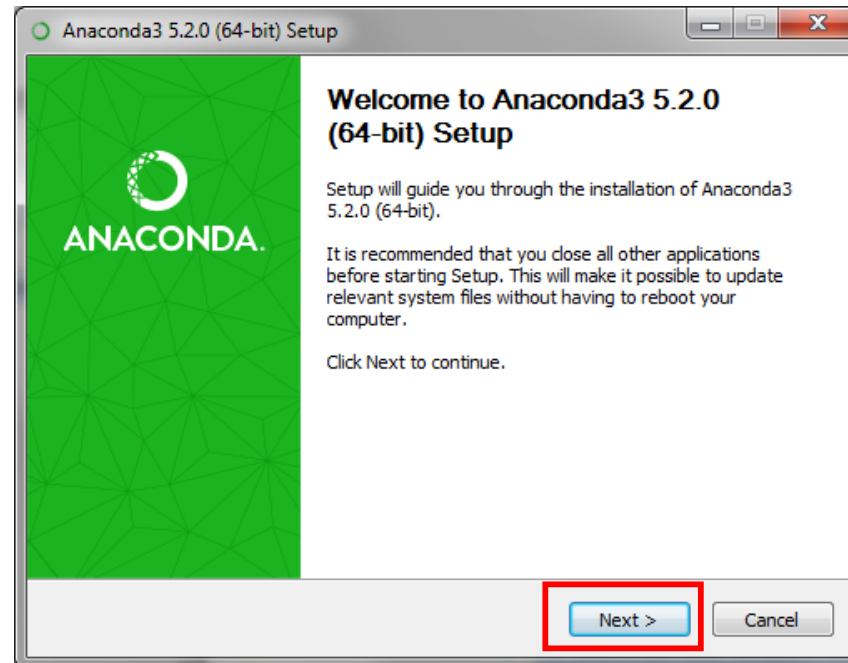


Anaconda Distribution (Cont'd)

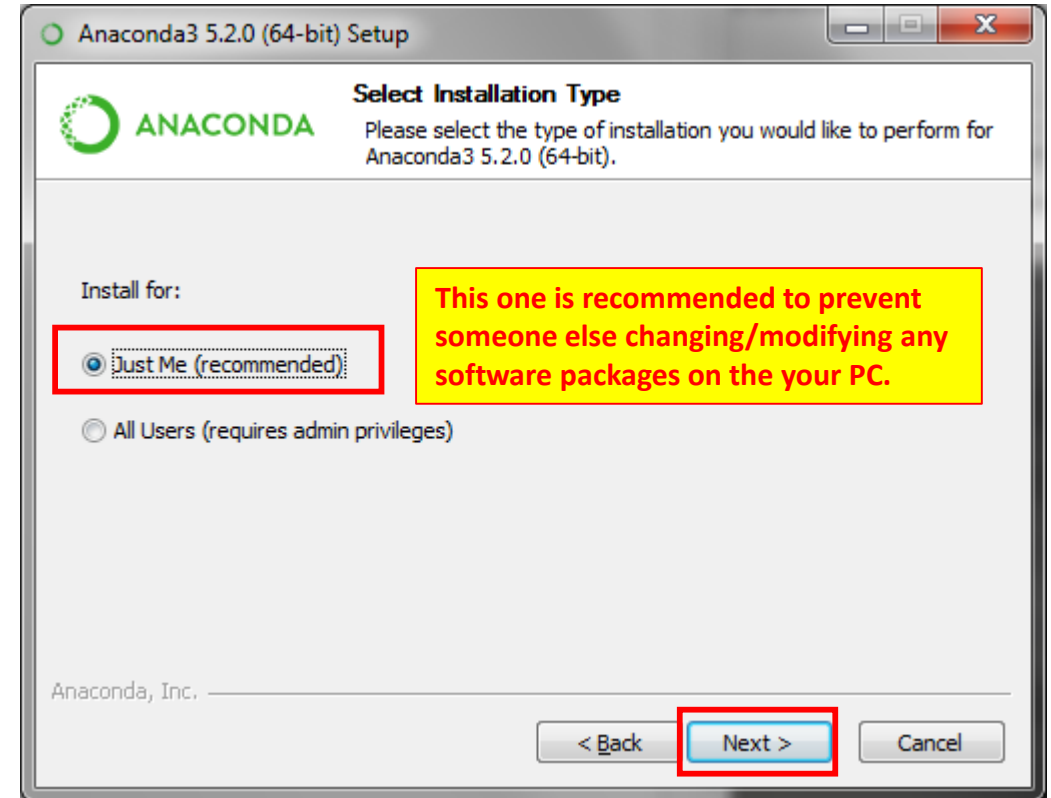
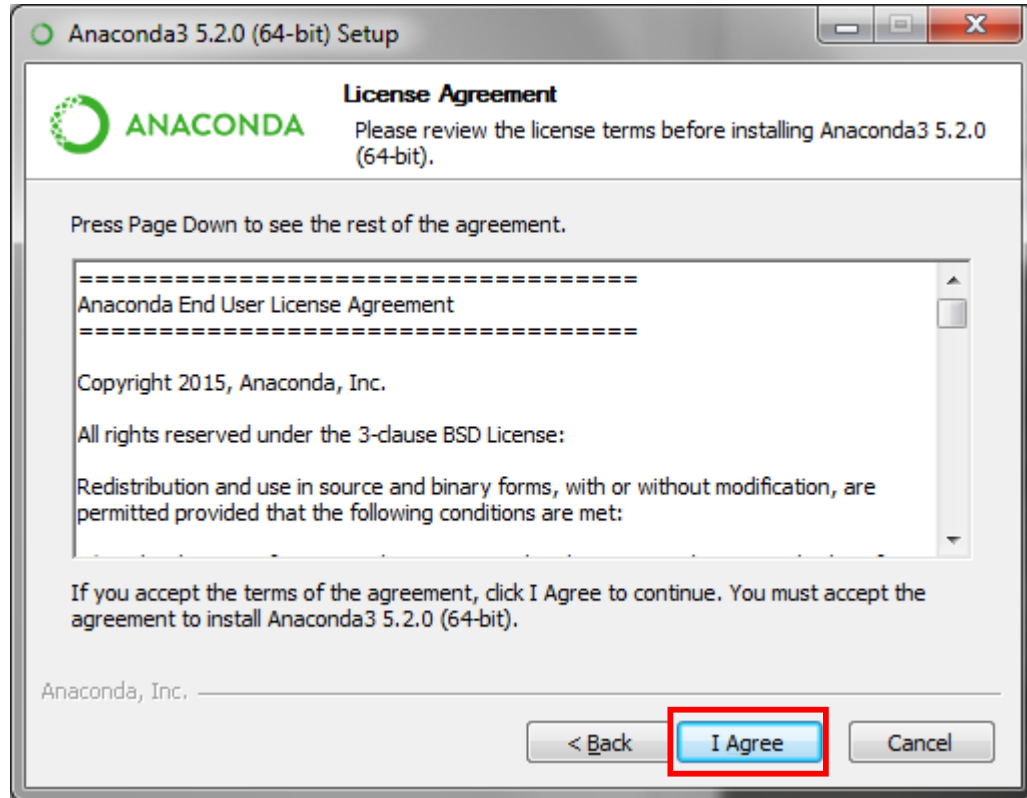
- Install Anaconda (select “Run as admin.”)



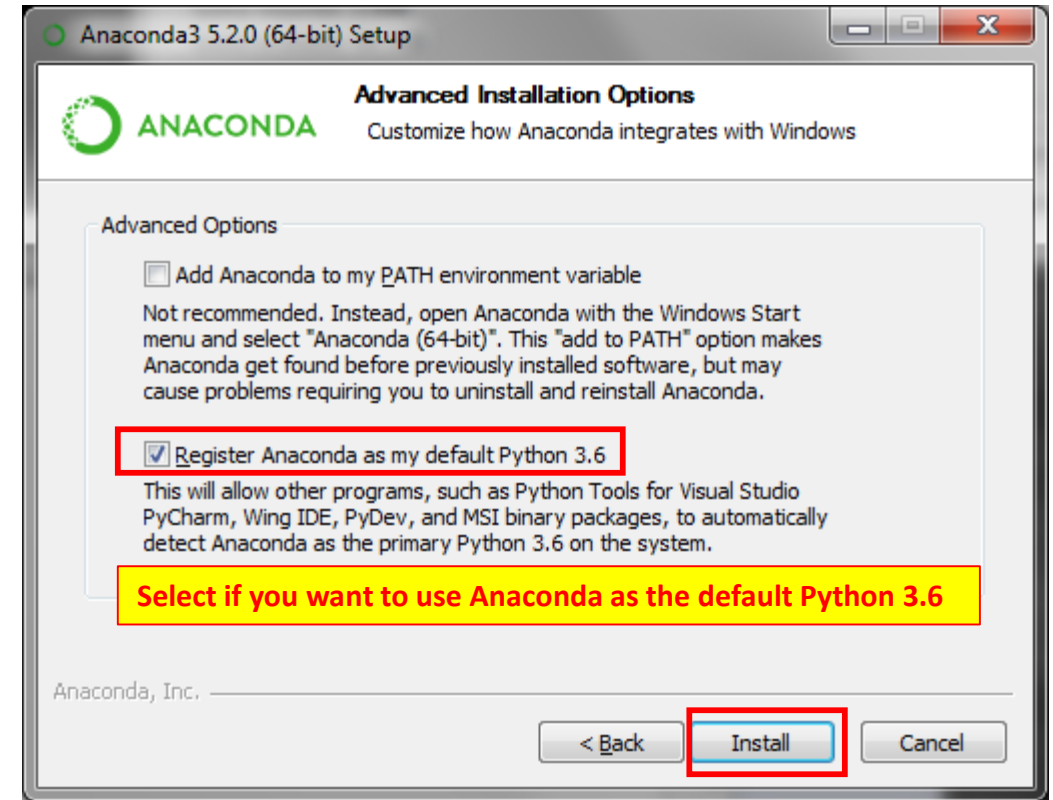
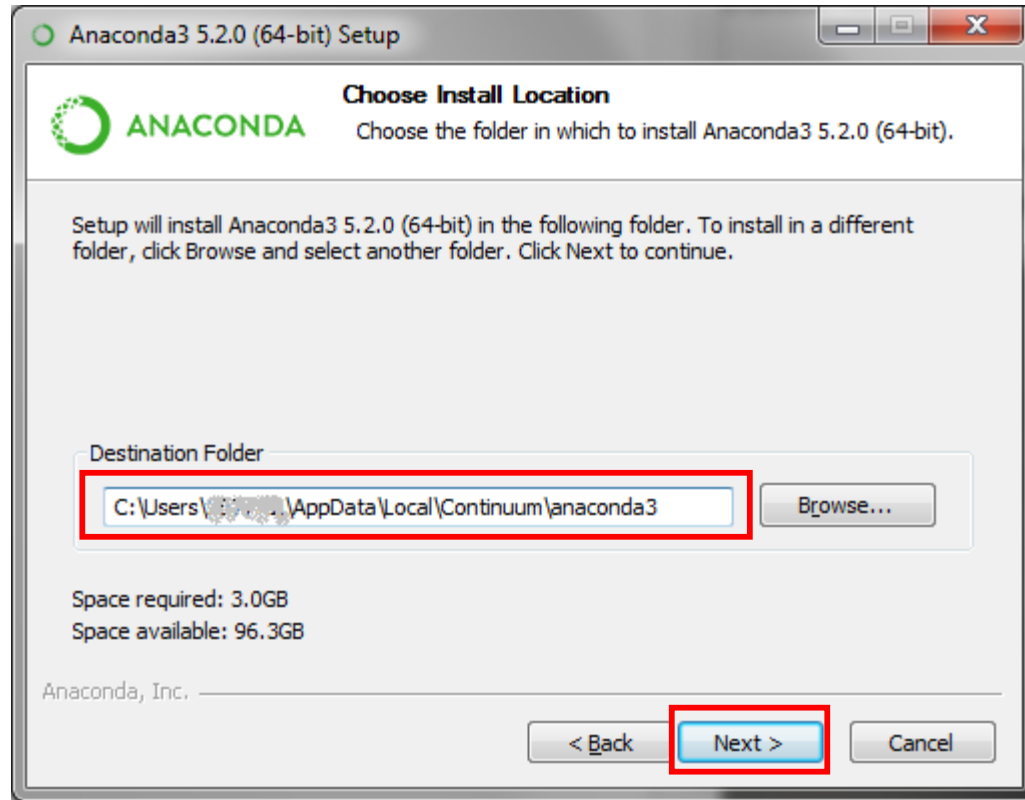
- Proceed the installation steps



Anaconda Distribution (Cont'd)

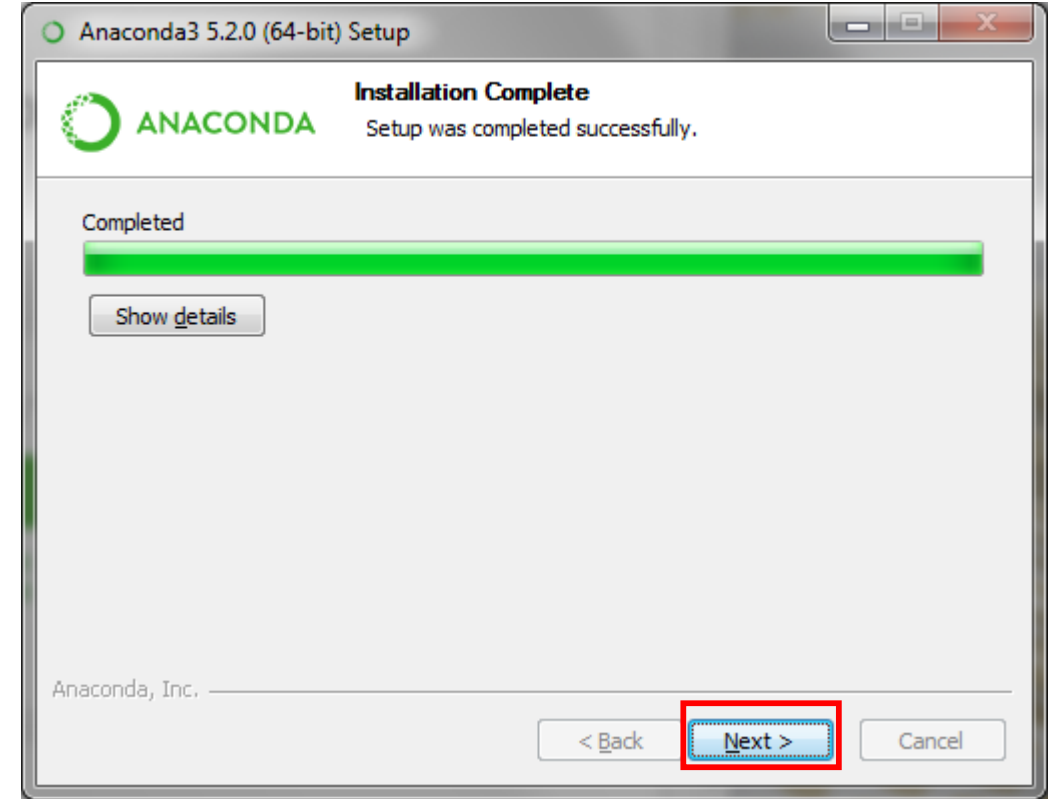
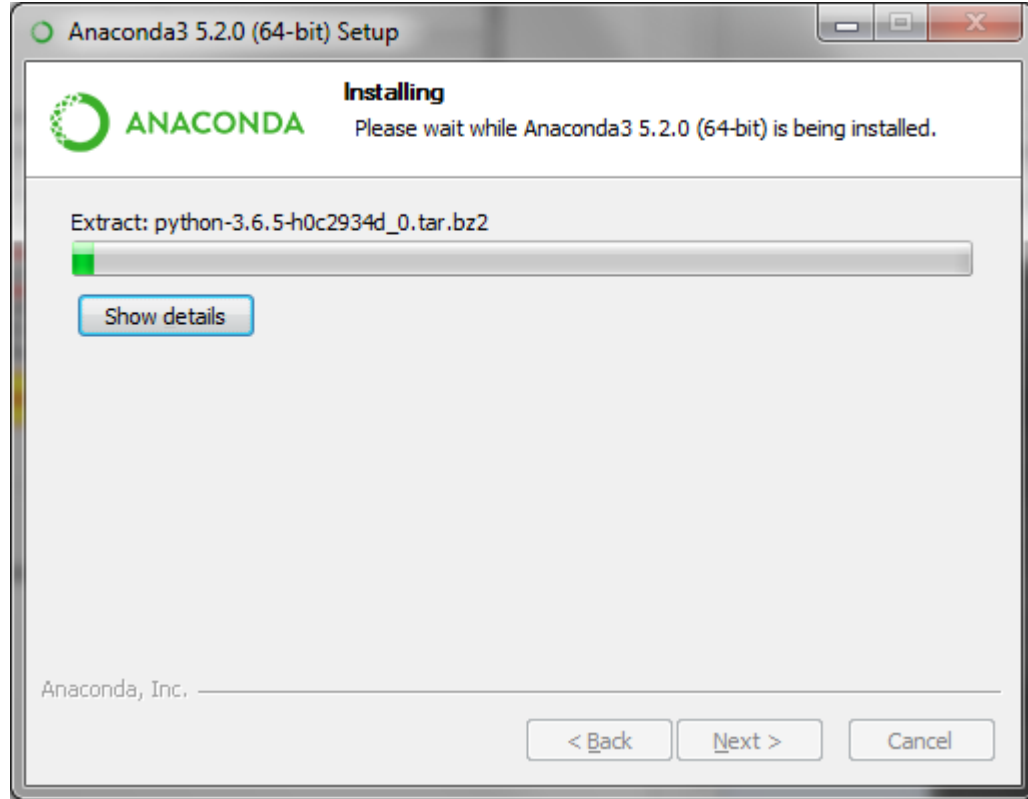


Anaconda Distribution (Cont'd)



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

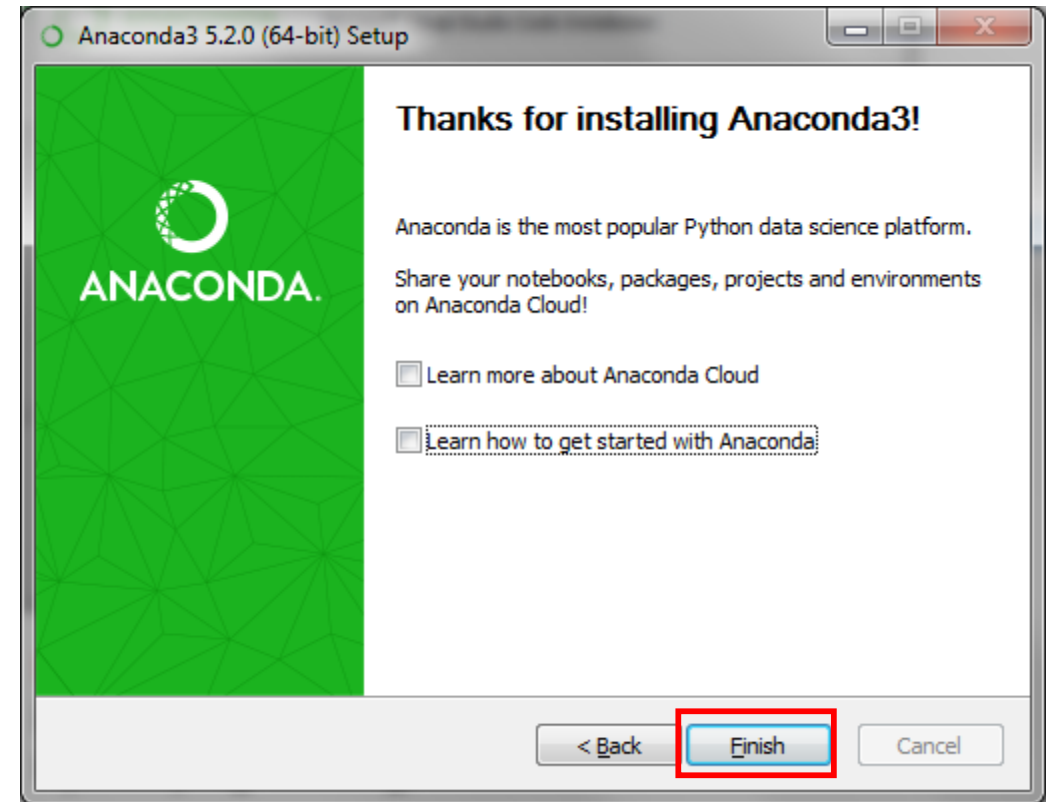
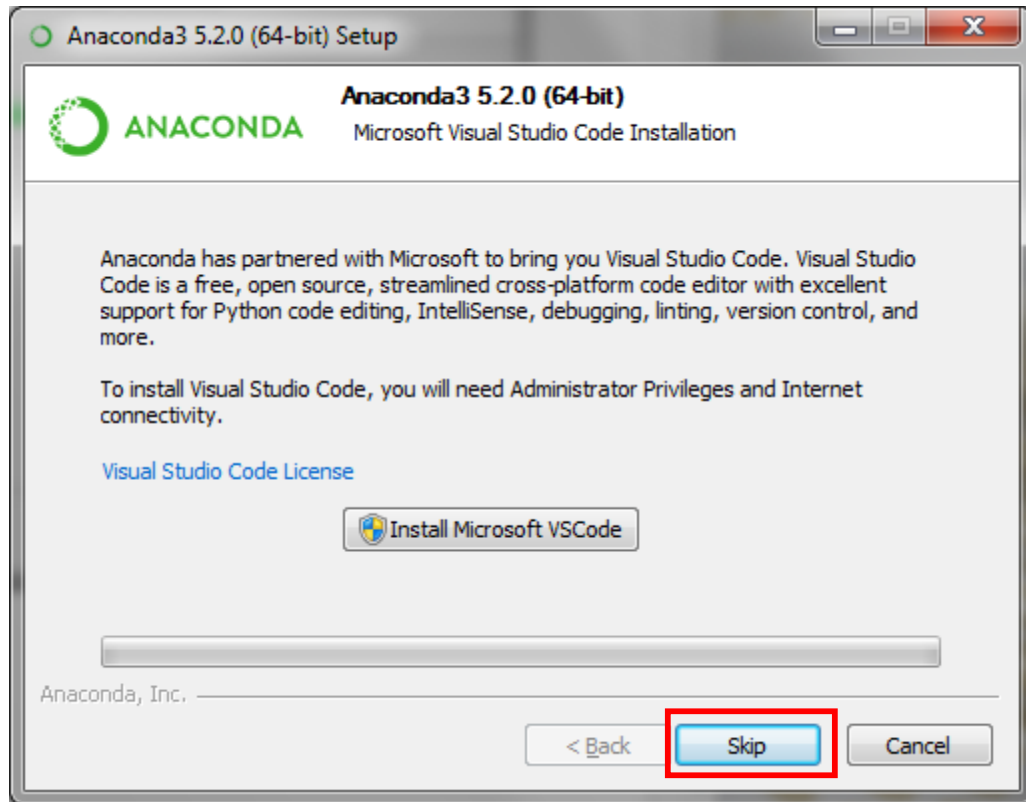
Anaconda Distribution (Cont'd)



It will take a while ... so let's have



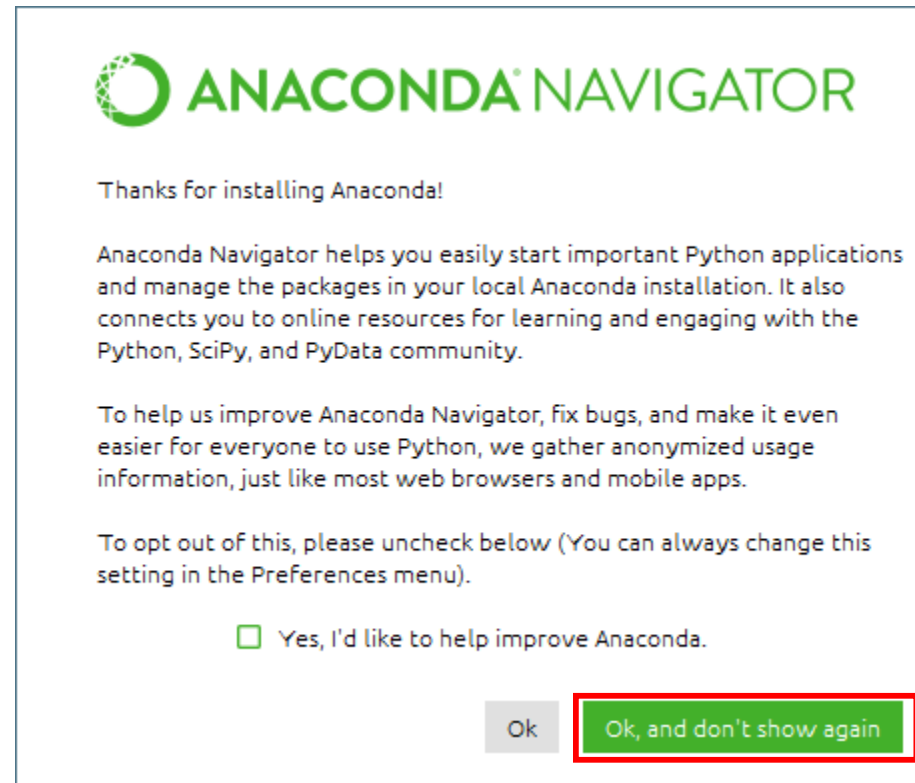
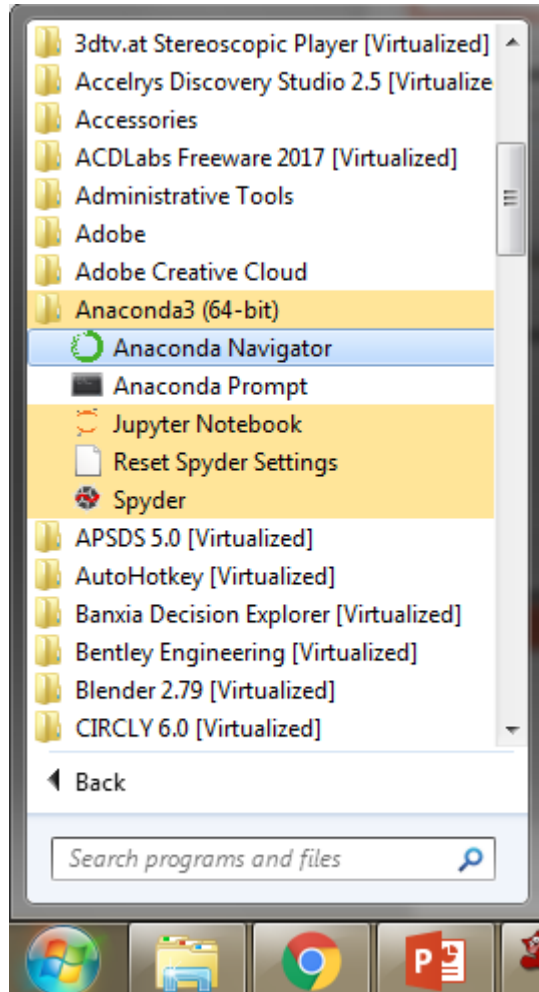
Anaconda Distribution (Cont'd)



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

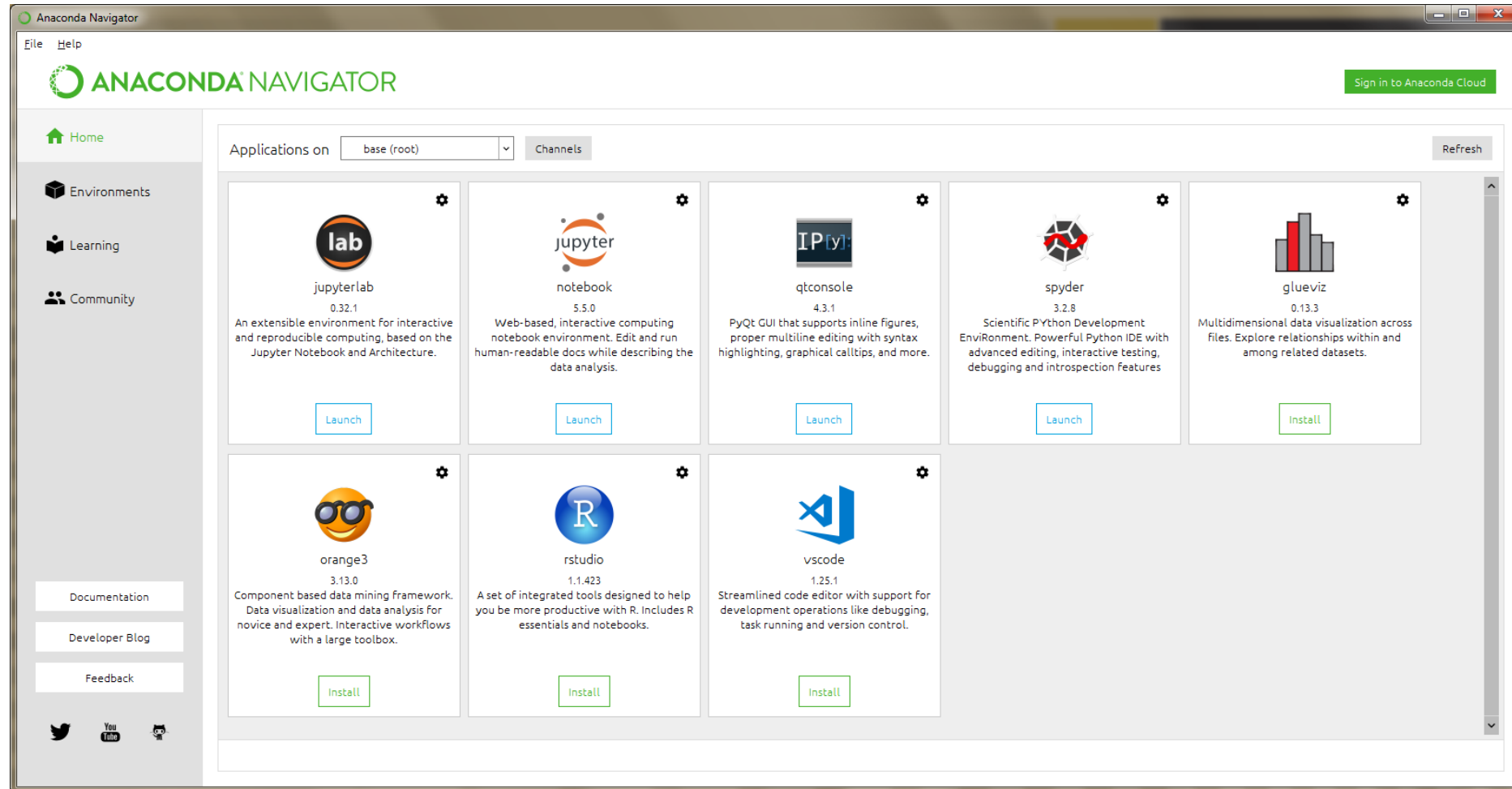
Anaconda Distribution (Cont'd)

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



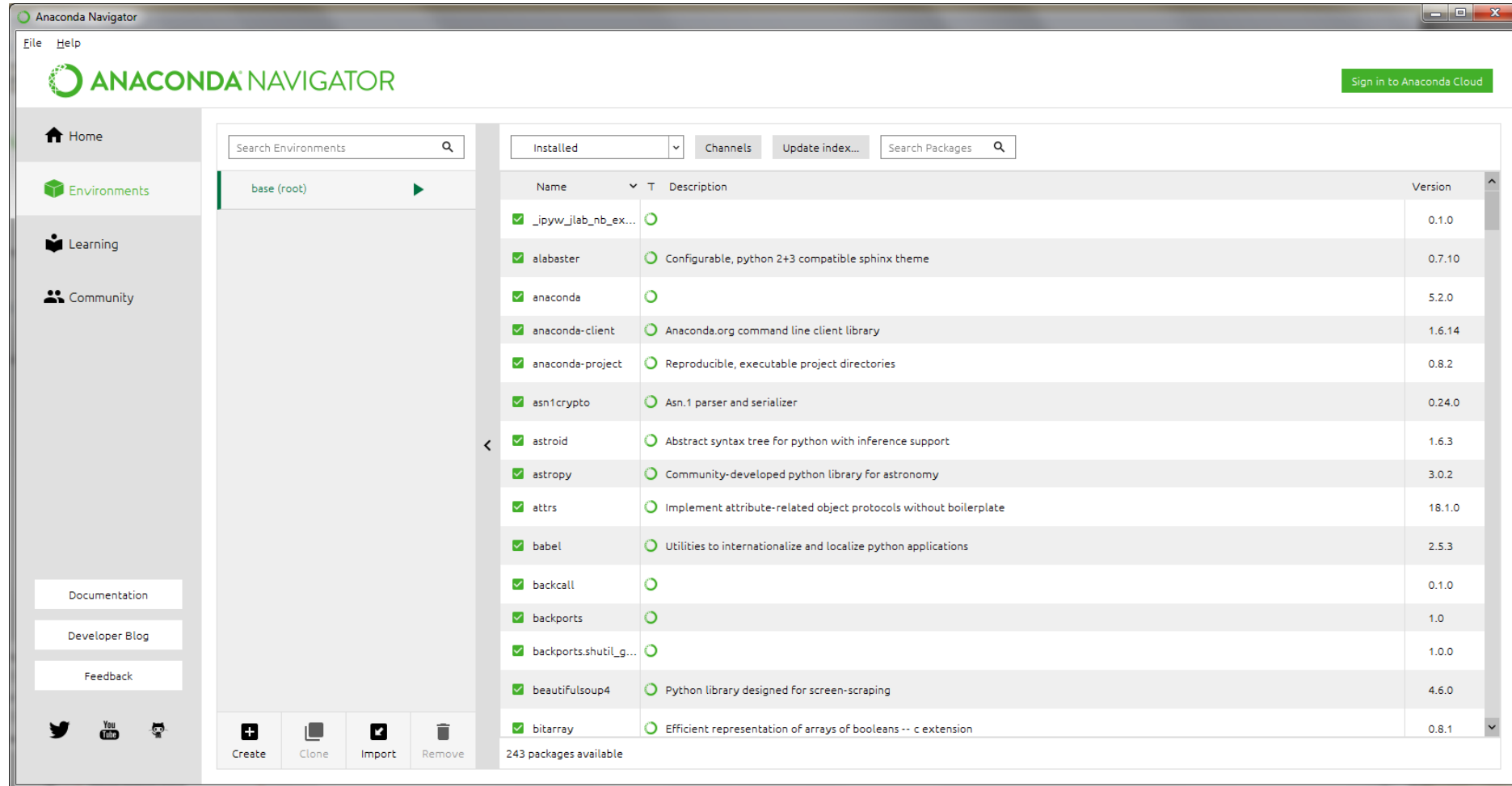
Anaconda Distribution (Cont'd)

- 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:

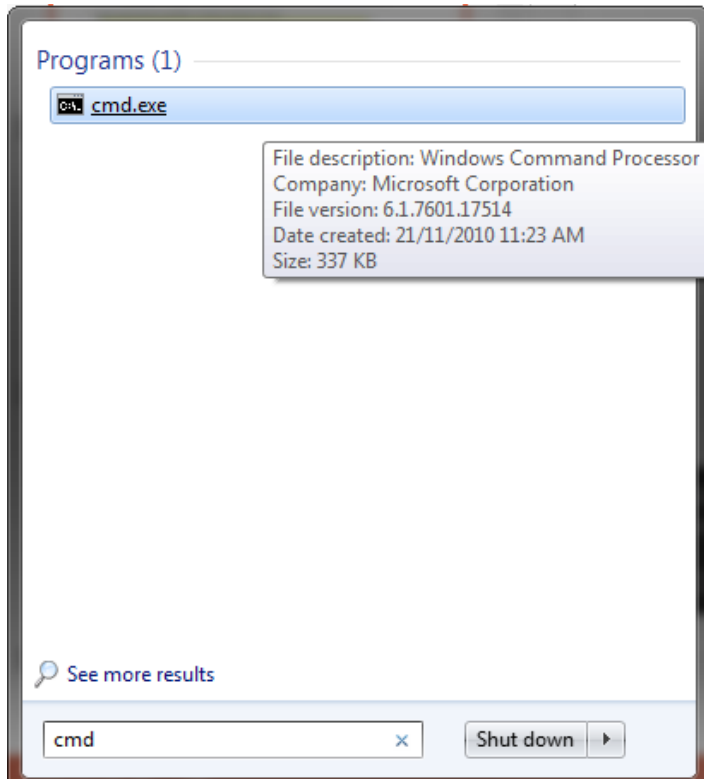
```
SETX PATH "%PATH%;C:\Users\yourID\AppData\Local\Continuum\anaconda3\Scripts;
C:\Users\yourID\AppData\Local\Continuum\anaconda3"
```

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



```
C:\Windows\system32\cmd.exe - python
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\[redacted]>python
Python 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:32:41) [MSC v.1900 64 bi
t (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> _
```

Successfully installed!


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

```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\[redacted]>python
Python 3.6.5 |Anaconda, Inc.| (default, Mar 29 2018, 13:32:41) [MSC v.1900 64 bi
t (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()

C:\Users\[redacted]>_
```

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`
- 
- A screenshot of an Anaconda Prompt terminal window. The title bar reads "Administrator: Anaconda Prompt". The command prompt shows the current directory as "C:\Windows\system32" and the active environment as "(base)". The command "python -m pi" is partially visible and highlighted with a yellow background.

```
Administrator: Anaconda Prompt
(base) C:\Windows\system32>python -m pip install --upgrade pip setuptools wheel
Collecting pip
  Downloading https://files.pythonhosted.org/packages/5f/25/e52d3f31441505a5f3a741213346e5b6c221c9e086a166f3703d2ddaf940/pip-18.0-py2.py3-none-any.whl (1.3MB)
    100% |████████████████████████████████████████| 1.3MB 7.9MB/s
Collecting setuptools
  Downloading https://files.pythonhosted.org/packages/ff/f4/385715ccc461885f3cedf57a41ae3c12b5fec3f35cce4c8706b1a112a133/setuptools-40.0.0-py2.py3-none-any.whl (567kB)
    100% |████████████████████████████████████████| 573kB 8.5MB/s
Requirement already up-to-date: wheel in c:\users\██████████\appdata\local\continuum\anaconda3\lib\site-packages (0.31.1)
distributed 1.21.8 requires msgpack, which is not installed.
Installing collected packages: pip, setuptools
  Found existing installation: pip 10.0.1
  Uninstalling pip-10.0.1:
    Successfully uninstalled pip-10.0.1
  Found existing installation: setuptools 39.1.0
  Uninstalling setuptools-39.1.0:
    Successfully uninstalled setuptools-39.1.0
Successfully installed pip-18.0 setuptools-40.0.0

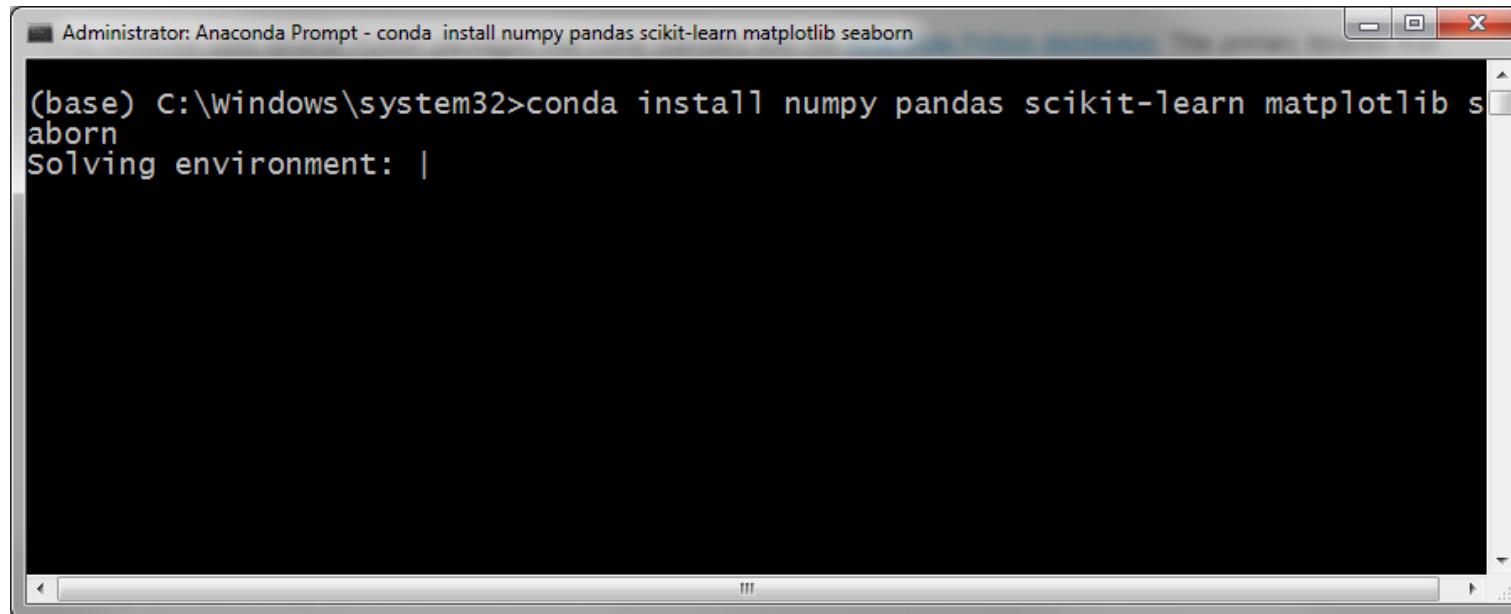
(base) C:\Windows\system32>pip install msgpack
Collecting msgpack
  Downloading https://files.pythonhosted.org/packages/04/81/c6363198f24ec1c56e5c48ce685cb532e175125adade0cdb181c8c5fea6e/msgpack-0.5.6-cp36-cp36m-win_amd64.whl (85kB)
    100% |████████████████████████████████████████| 92kB 4.9MB/s
Installing collected packages: 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: Scrappy, 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).

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



```
Administrator: Anaconda Prompt - conda install numpy pandas scikit-learn matplotlib seaborn  
(base) C:\Windows\system32>conda install numpy pandas scikit-learn matplotlib seaborn  
Solving environment: |
```

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 seaborn
Solving environment: done

## Package Plan ##

  environment location: C:\Users\████████\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\████████\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

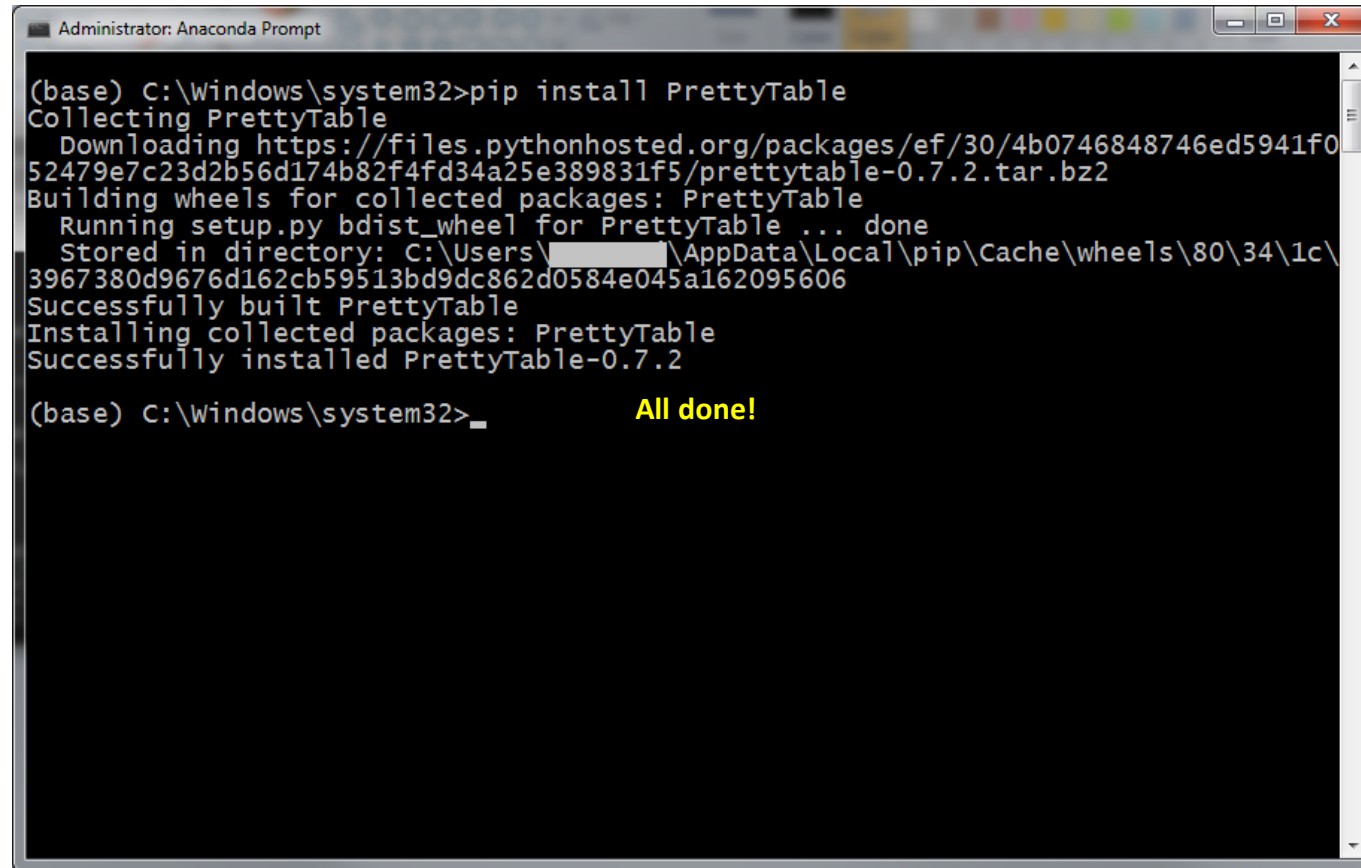
Proceed ([y]/n)? y

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/30/4b0746848746ed5941f052479e7c23d2b56d174b82f4fd34a25e389831f5/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\...\AppData\Local\pip\Cache\wheels\80\34\1c\3967380d9676d162cb59513bd9dc862d0584e045a162095606
Successfully built PrettyTable
Installing collected packages: PrettyTable
Successfully installed PrettyTable-0.7.2

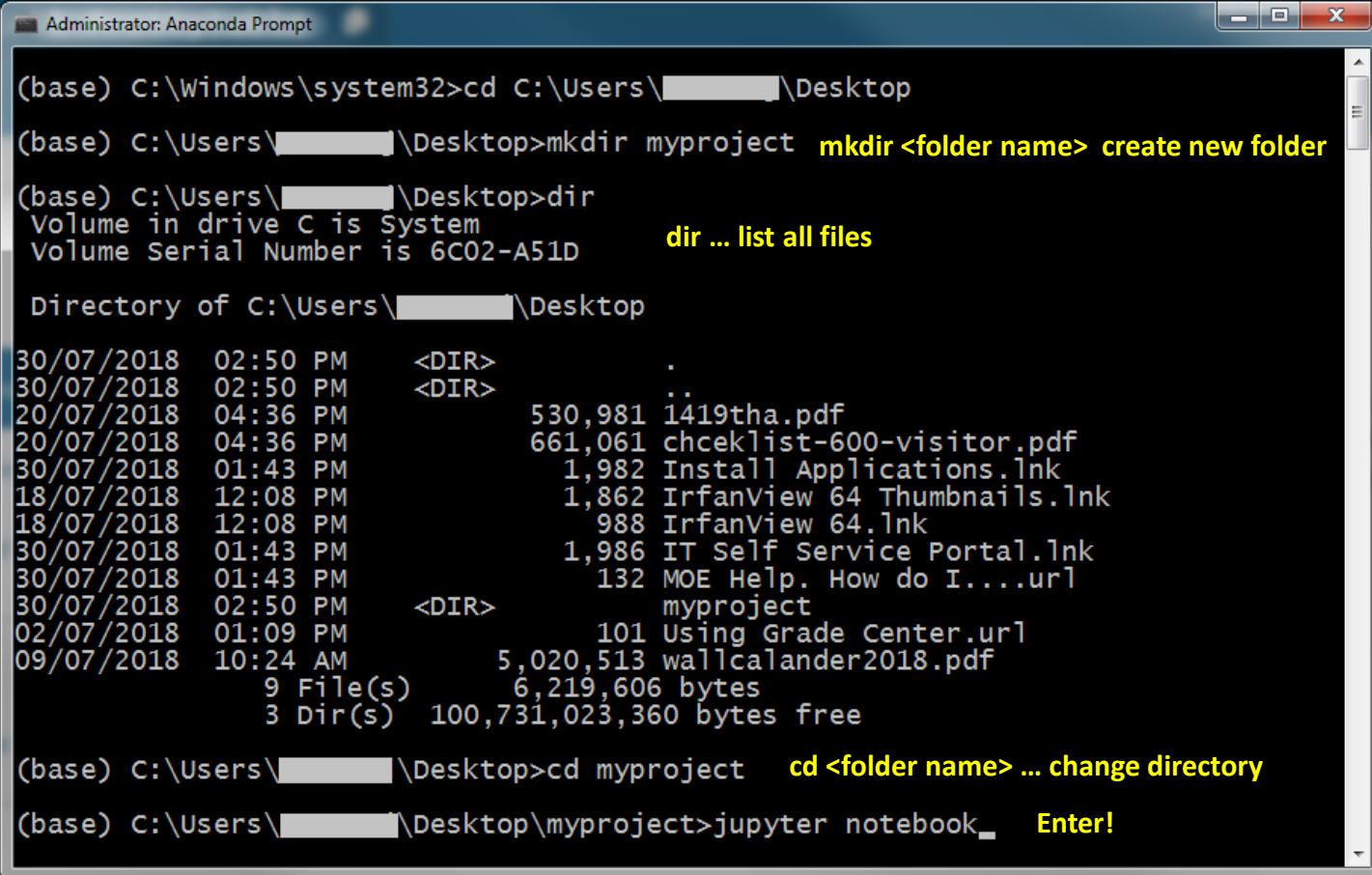
(base) C:\Windows\system32>_           All done!
```

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 Anaconda prompt
 - `pip install jupyter_contrib_nbextensions`
 - `jupyter contrib nbextension install --user`

Example Project using Jupyter Notebook

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



```
Administrator: Anaconda Prompt

(base) C:\Windows\system32>cd C:\Users\██████\Desktop

(base) C:\Users\██████\Desktop>mkdir myproject  mkdir <folder name> create new folder

(base) C:\Users\██████\Desktop>dir
Volume in drive C is System
Volume Serial Number is 6C02-A51D
dir ... list all files

Directory of C:\Users\██████\Desktop

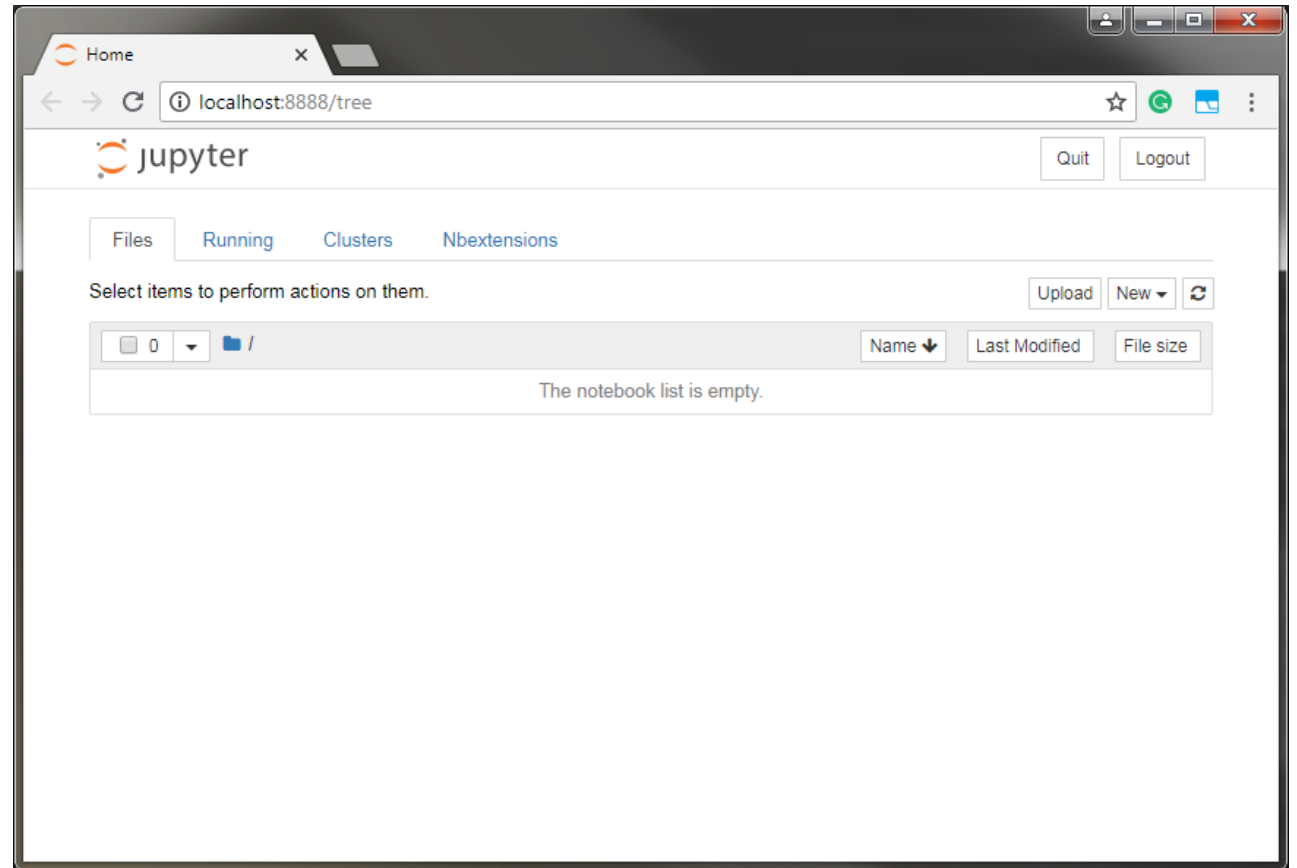
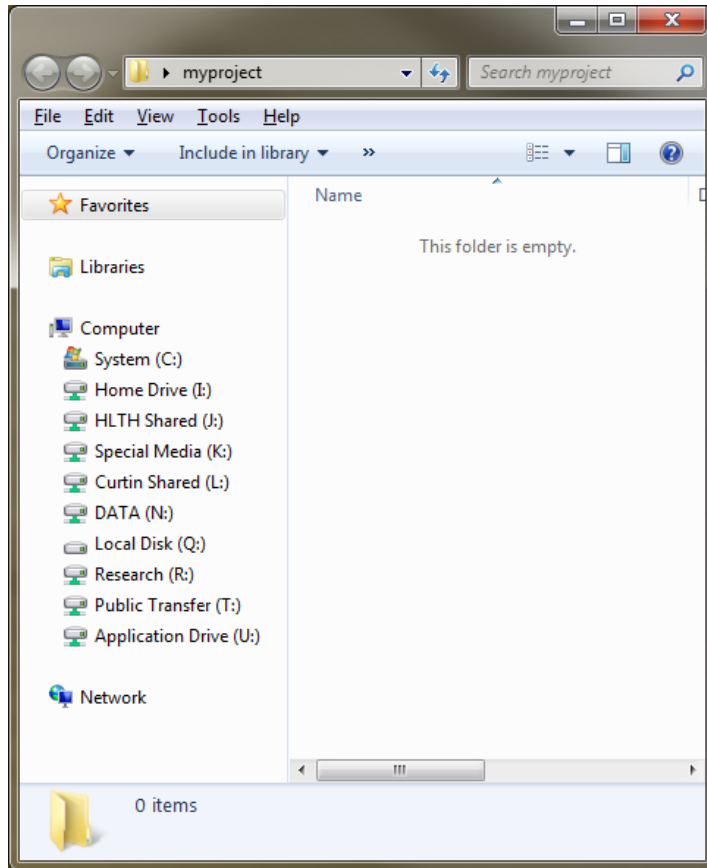
30/07/2018  02:50 PM    <DIR>          .
30/07/2018  02:50 PM    <DIR>          ..
20/07/2018  04:36 PM             530,981  1419tha.pdf
20/07/2018  04:36 PM             661,061  chceklist-600-visitor.pdf
30/07/2018  01:43 PM              1,982  Install Applications.lnk
18/07/2018  12:08 PM             1,862  IrfanView 64 Thumbnails.lnk
18/07/2018  12:08 PM              988  IrfanView 64.lnk
30/07/2018  01:43 PM             1,986  IT Self Service Portal.lnk
30/07/2018  01:43 PM              132  MOE Help. How do I....url
30/07/2018  02:50 PM    <DIR>          myproject
02/07/2018  01:09 PM              101  Using Grade Center.url
09/07/2018  10:24 AM             5,020,513 wallcalander2018.pdf
               9 File(s)              6,219,606 bytes
               3 Dir(s)  100,731,023,360 bytes free

(base) C:\Users\██████\Desktop>cd myproject  cd <folder name> ... change directory

(base) C:\Users\██████\Desktop\myproject>jupyter notebook_  Enter!
```

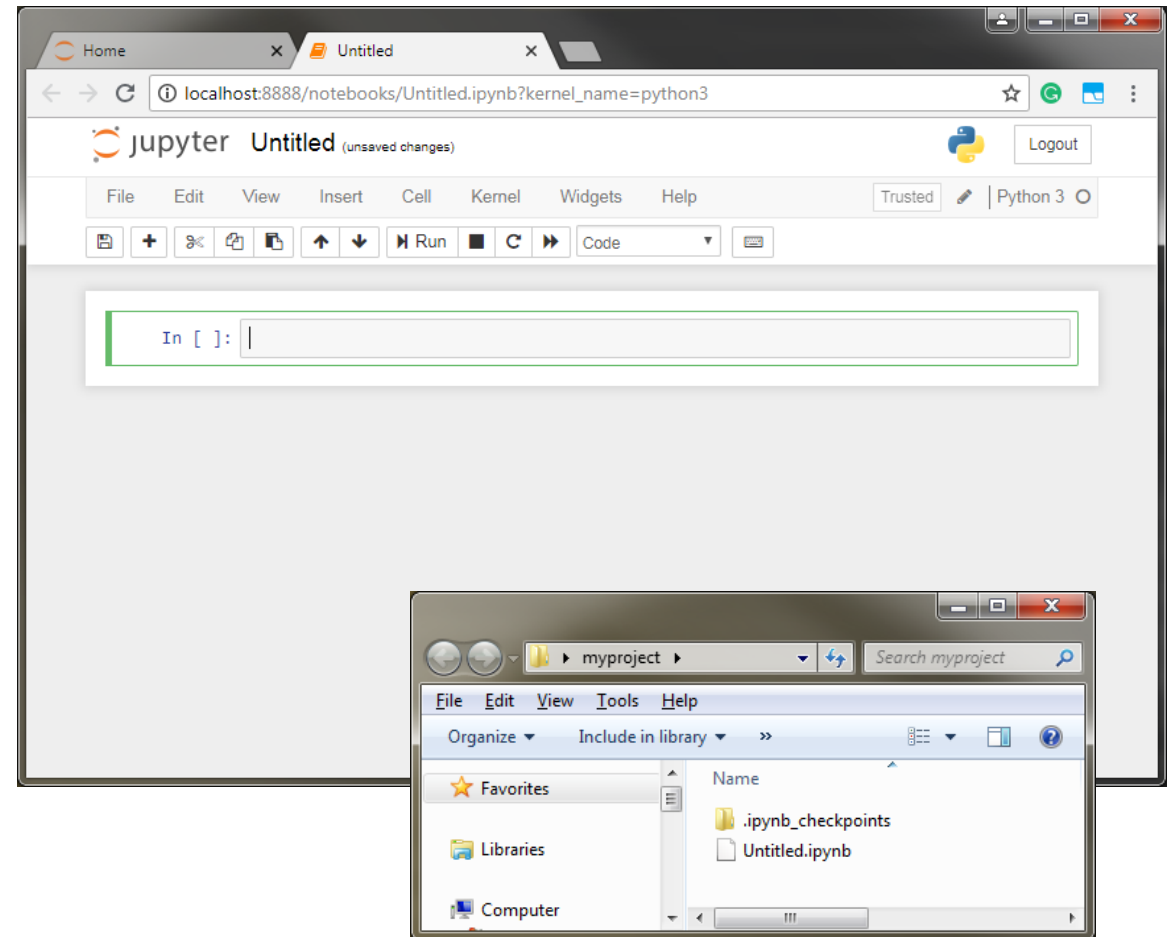
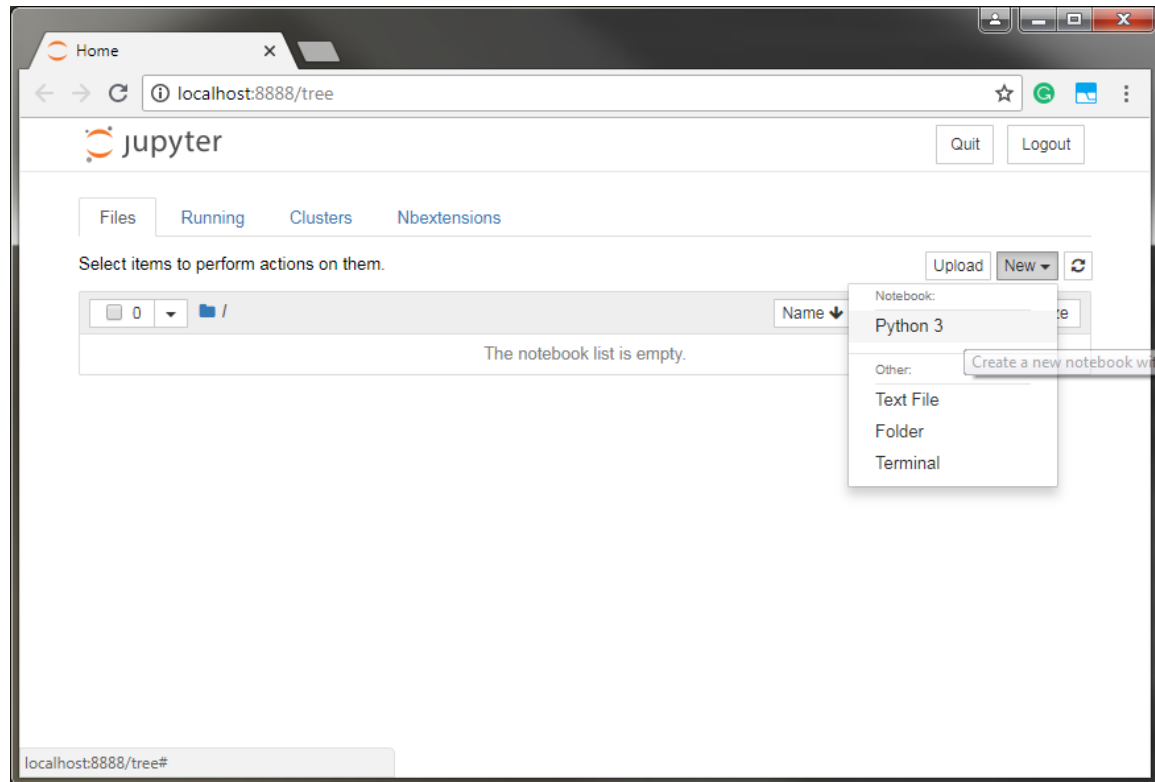
Example Project using Jupyter Notebook

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



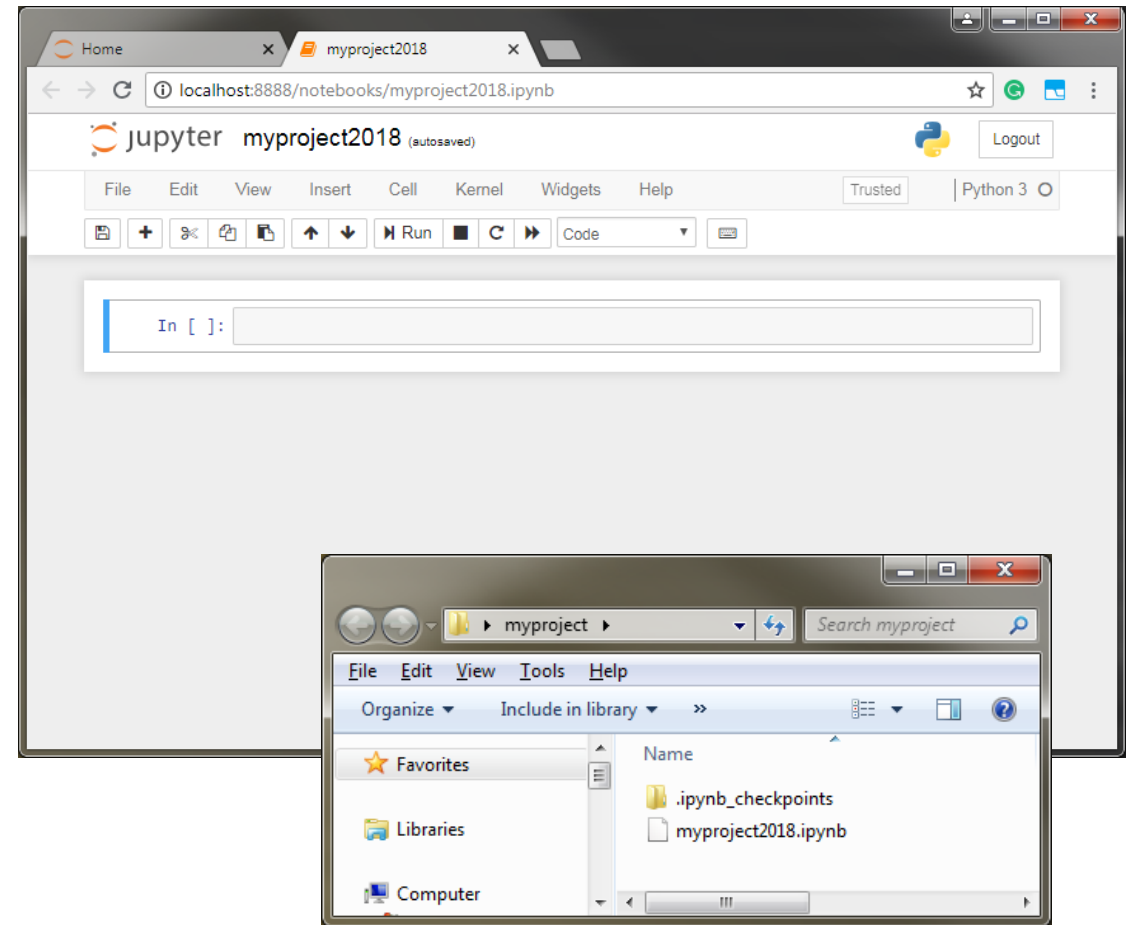
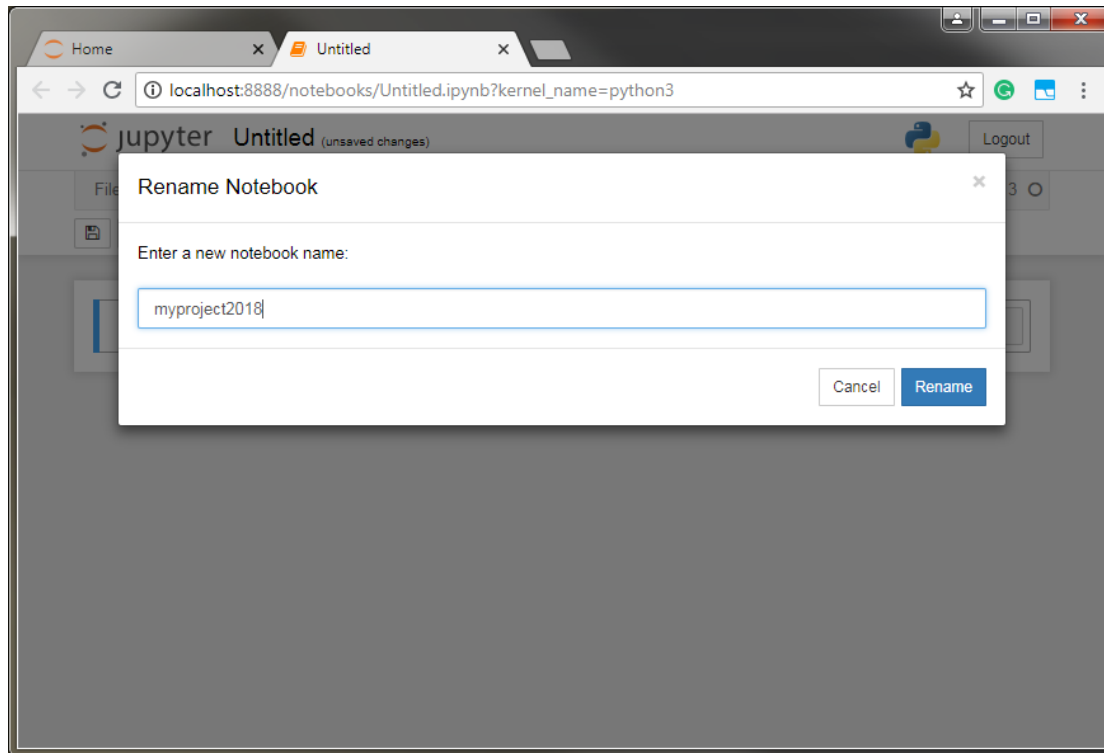
Example Project using Jupyter Notebook

- Create new empty notebook



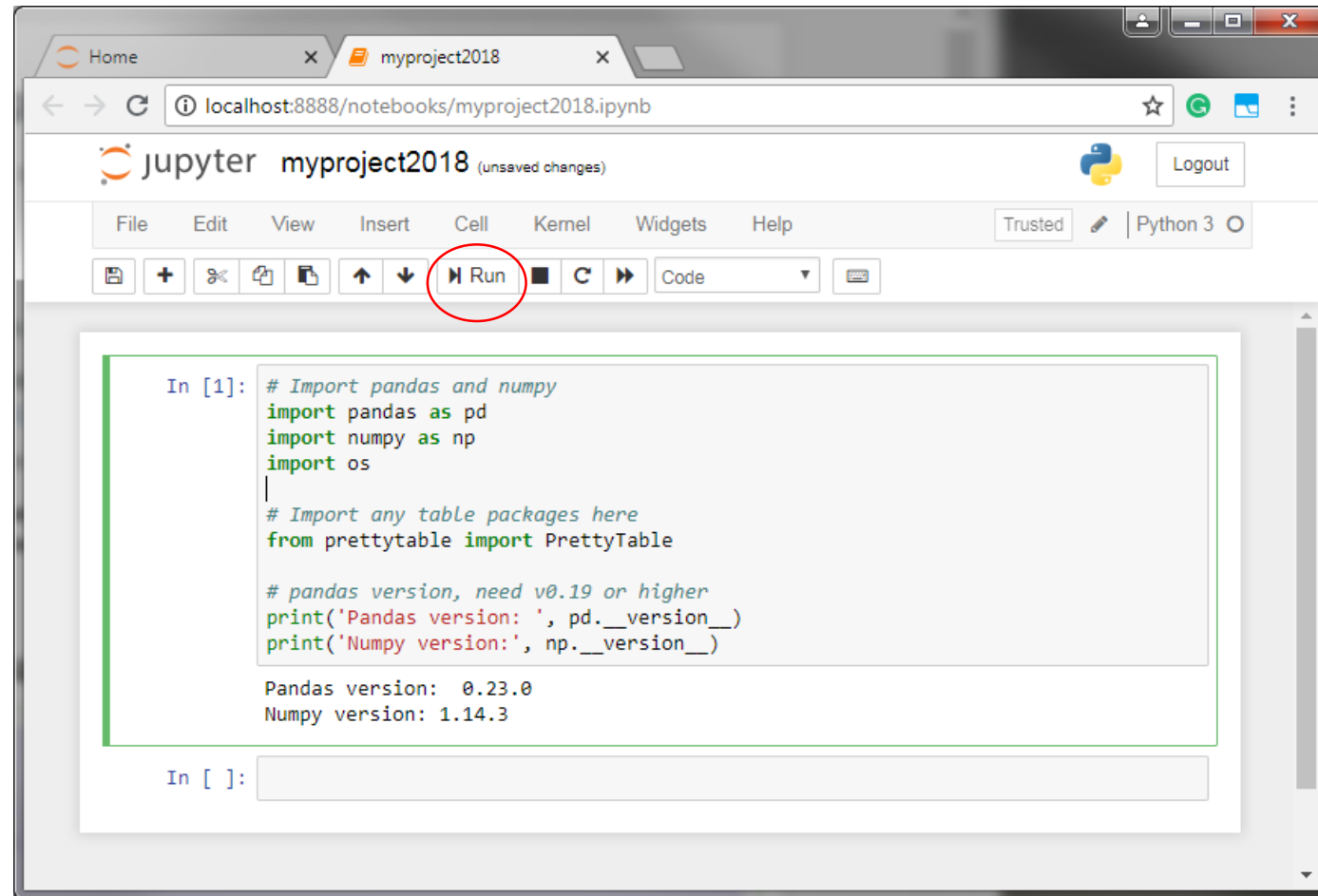
Example Project using Jupyter Notebook

- 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