Chapter 2 Python for Data Science

CSS 341 Introduction to Data Science

Chukiat Worasucheep

KMUTT

Important Notice

การเรียนการสอนหัวข้อนี้ ผ่านทางสื่อออนไลน์ (Online meeting)

และมีการบันทึกภาพและเสียงเพื่อประโยชน์ทางการศึกษาต่อไปในอนาคต.

หากท่านไม่ยินยอมให้มีการเผยแพร่การบันทึกดังกล่าว ขอให้แจ้งให้ผู้สอนทราบภายใน 36 ชั่วโมง.

Learning objectives

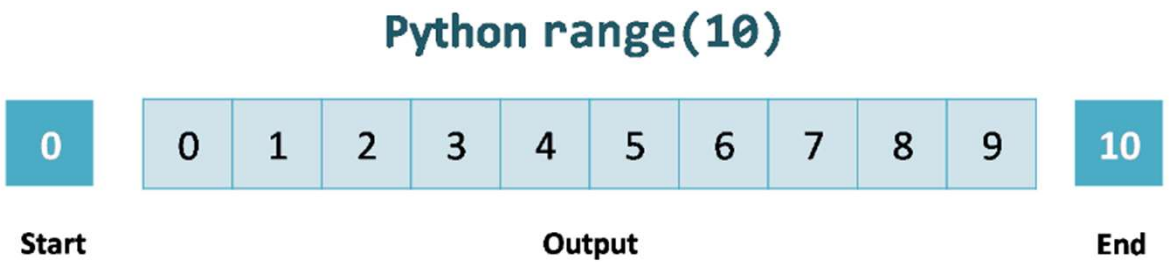
- เข้าใจการใช้งาน Python เพื่องาน data science
- หัดทดลองใช้ Jupyter notebook เพื่องาน data science
- ทบทวน data structures เช่น tuple, list, dictionary สำหรับงาน data science
- แนะนำ numpy และ pandas เบื้องต้น

Outline

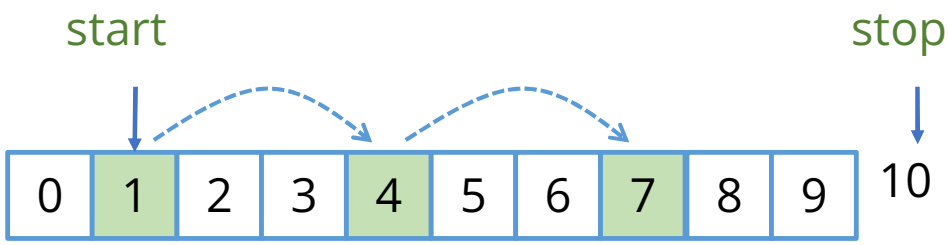
- Review basic Python
- Tools esp. Jupyter notebook
- Introduce numpy and pandas libraries

range()

```
range(start, stop, step_size)
```

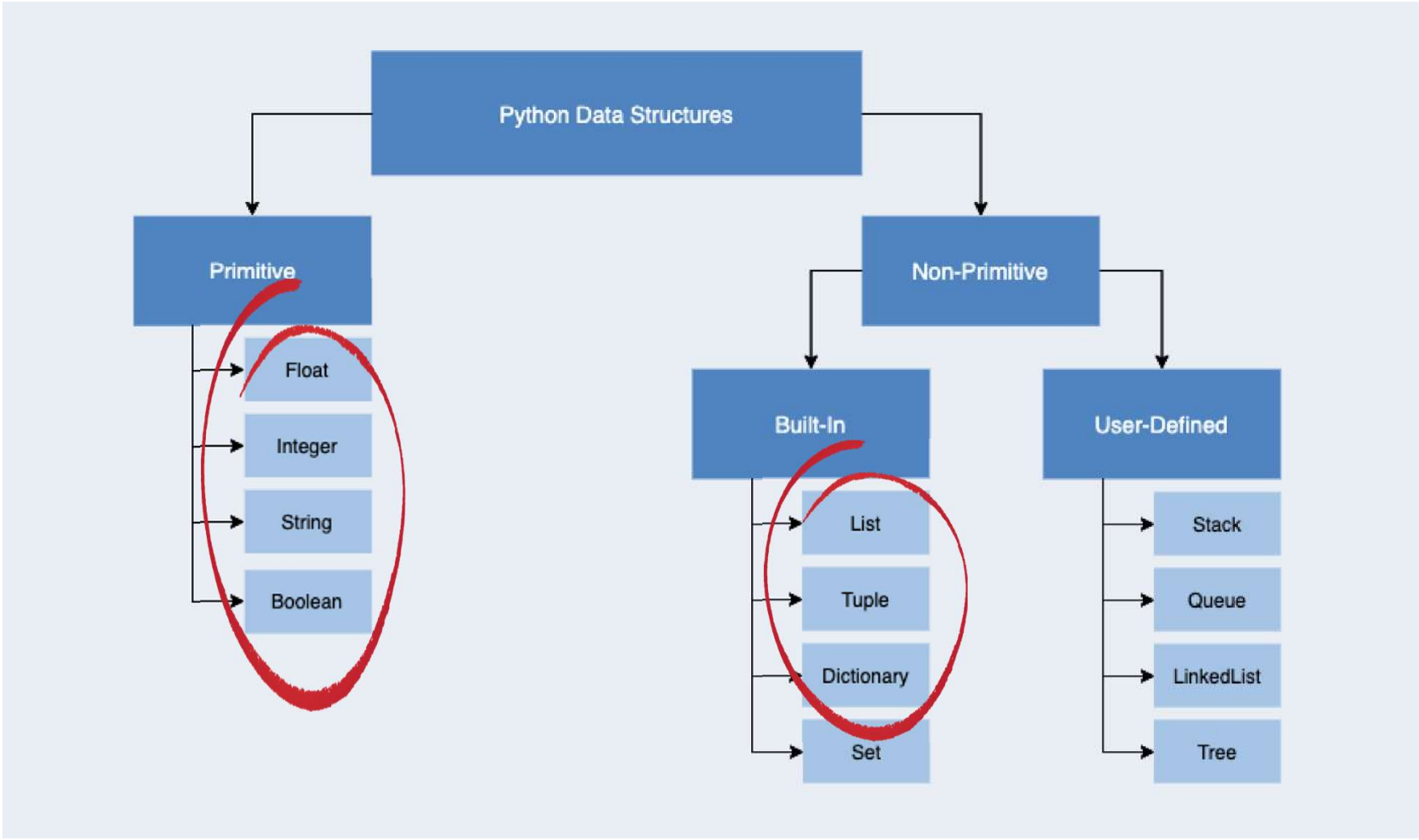


range(1, 10, 3)

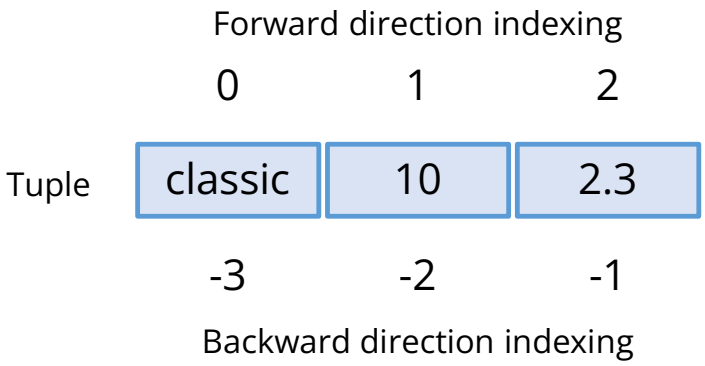


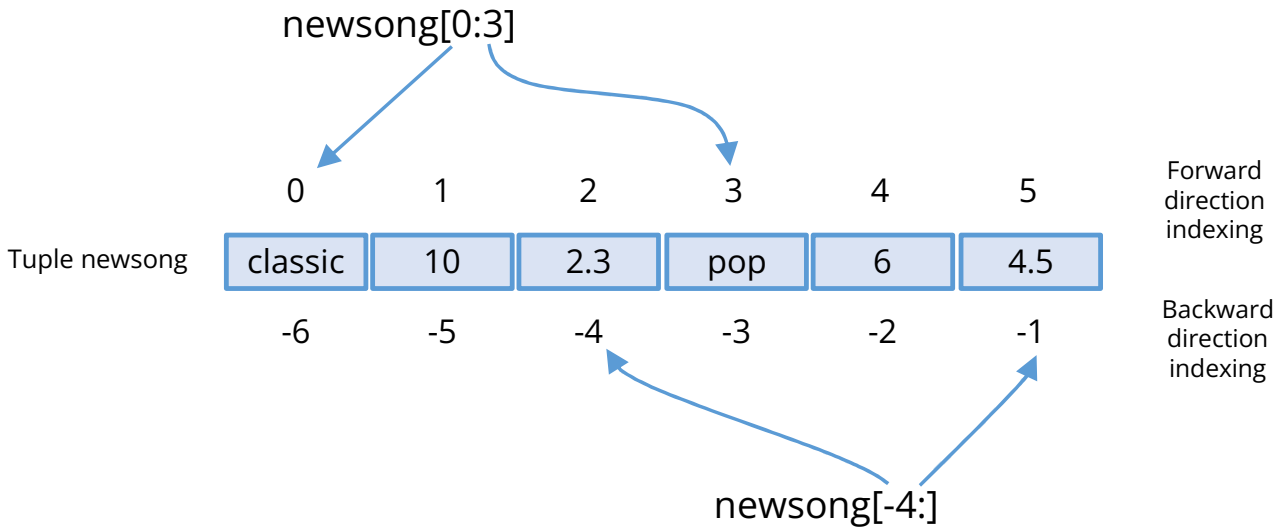
1, 4, 7

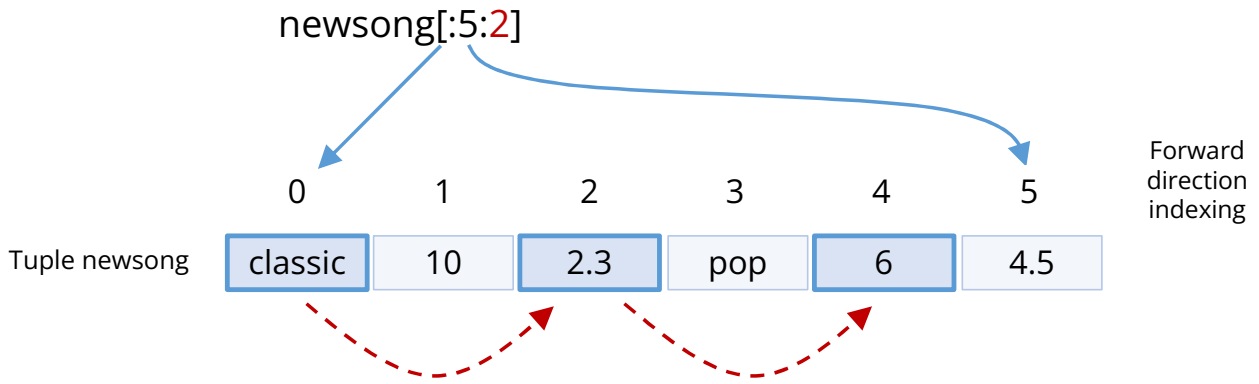
Python data structures



Tuple


Forward direction indexing
0
1
2
Tuple
classic
10
2.3
-3
-2
-1
Backward direction indexing


newsong[0:3]
0
1
2
3
4
5
Tuple newsong
classic
10
2.3
pop
6
4.5
-6
-5
-4
-3
-2
-1
Forward direction indexing
Backward direction indexing
newsong[-4:]


newsong[:5:2]
0
1
2
3
4
5
Tuple newsong
classic
10
2.3
pop
6
4.5
Forward direction indexing

Outline

- Review basic Python
- Tools esp. Jupyter notebook
- Introduce numpy and pandas libraries

Tools for developing Python programs

PyCharm
Community

VS Code
with Python Extension

Jupyter
Notebook

Install program jupyter

d:\>pip install jupyter

Command Prompt for Notebooks

```
D:\notebooks>pip install jupyter
Defaulting to user installation because normal site-packages is not writeable
Collecting jupyter
  Using cached jupyter-1.0.0-py2.py3-none-any.whl (2.7 kB)
Requirement already satisfied: qtconsole in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (5.1.1)
Requirement already satisfied: notebook in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (6.4.10)
Requirement already satisfied: ipykernel in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (6.2.0)
Requirement already satisfied: nbconvert in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (6.4.4)
Requirement already satisfied: jupyter-console in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (6.4.0)
Requirement already satisfied: ipywidgets in c:\users\chuki\appdata\roaming\python\python38\site-packages (from jupyter) (7.6.3)
Requirement already satisfied: jupyter-client<8.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (7.0.1)
Requirement already satisfied: matplotlib-inline<0.2.0,>=0.1.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (0.1.2)
Requirement already satisfied: tornado<7.0,>=4.2 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (6.1)
Requirement already satisfied: traitlets<6.0,>=4.1.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (5.0.5)
Requirement already satisfied: debugpy<2.0,>=1.0.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (1.4.1)
Requirement already satisfied: ipython<8.0,>=7.23.1 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipykernel->jupyter) (7.26.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipywidgets->jupyter) (3.5.1)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipywidgets->jupyter) (1.0.0)
Requirement already satisfied: nbformat>=4.2.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from ipywidgets->jupyter) (5.1.3)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from packaging->bleach->nbconvert->jupyter) (2.4.7)

Installing collected packages: jupyter
Successfully installed jupyter-1.0.0
WARNING: You are using pip version 22.0.3; however, version 22.2.2 is available.
You should consider upgrading via the 'c:\program files\python38\python.exe -m pip install --upgrade pip' command.

D:\notebooks>
```

Install program Jupyter lab

d:\>pip install jupyterlab

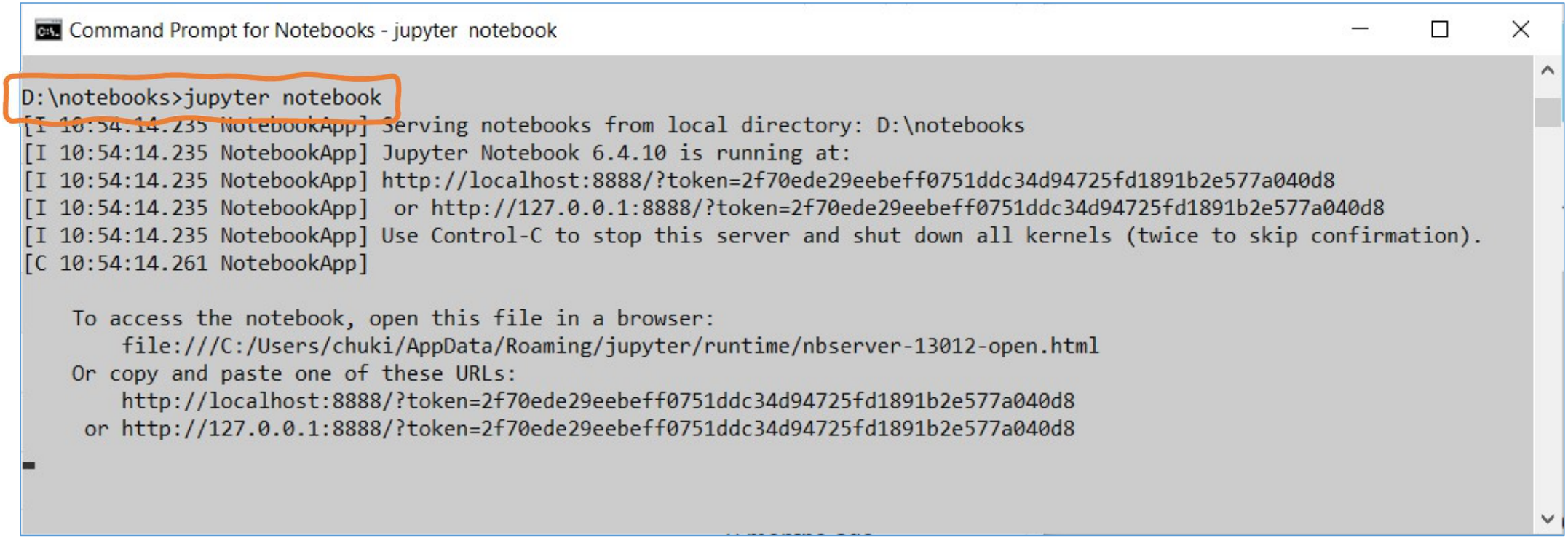
```
Administrator: Command Prompt

C:\Windows\system32>pip install jupyterlab
Requirement already satisfied: jupyterlab in c:\users\chuki\appdata\roaming\python\python38\s
Requirement already satisfied: packaging in c:\users\chuki\appdata\roaming\python\python38\s
Requirement already satisfied: jupyterlab-server~=2.3 in c:\users\chuki\appdata\roaming\pytho
) (2.7.1)
Requirement already satisfied: tornado>=6.1.0 in c:\users\chuki\appdata\roaming\python\pytho
Requirement already satisfied: jinja2>=2.1 in c:\users\chuki\appdata\roaming\python\python38
Requirement already satisfied: nbclassic~=0.2 in c:\users\chuki\appdata\roaming\python\pytho
)
Requirement already satisfied: jupyter-core in c:\users\chuki\appdata\roaming\python\python3
Requirement already satisfied: jupyter-server~=1.4 in c:\users\chuki\appdata\roaming\python\
1.10.2)
Requirement already satisfied: ipython in c:\users\chuki\appdata\roaming\python\python38\site
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\chuki\appdata\roaming\python\pytho
terlab) (2.0.1)
Requirement already satisfied: ipython-genutils in c:\users\chuki\appdata\roaming\python\pyth
1.4->jupyterlab) (0.2.0)
```

Install commonly used packages for data science works

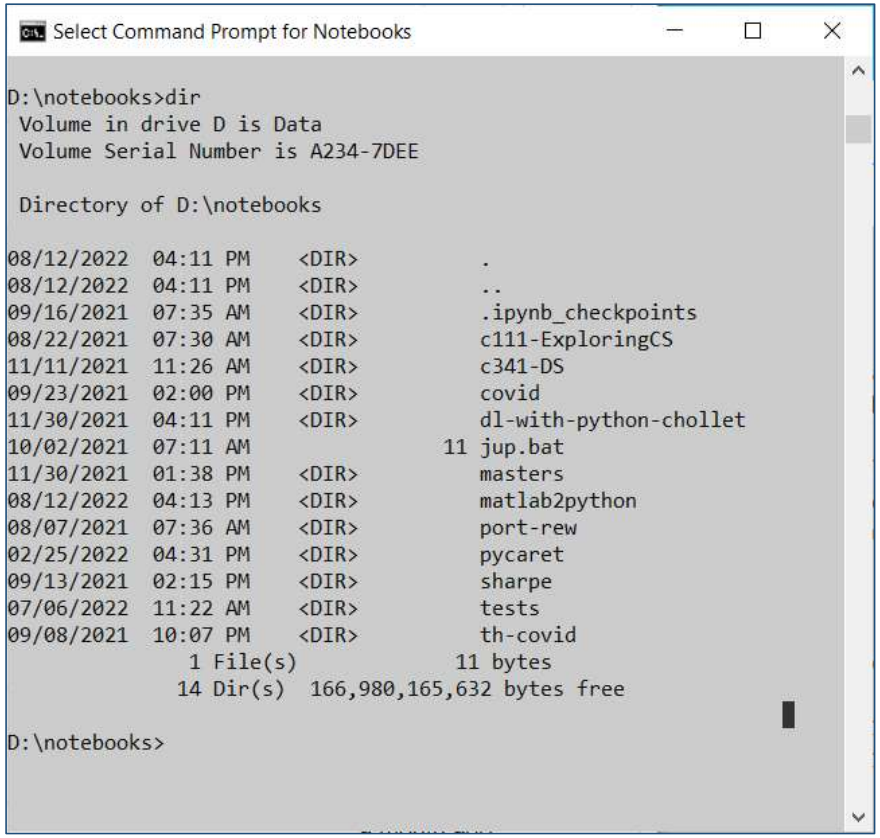
d:\>pip install numpy pandas matplotlib sklearn scipy

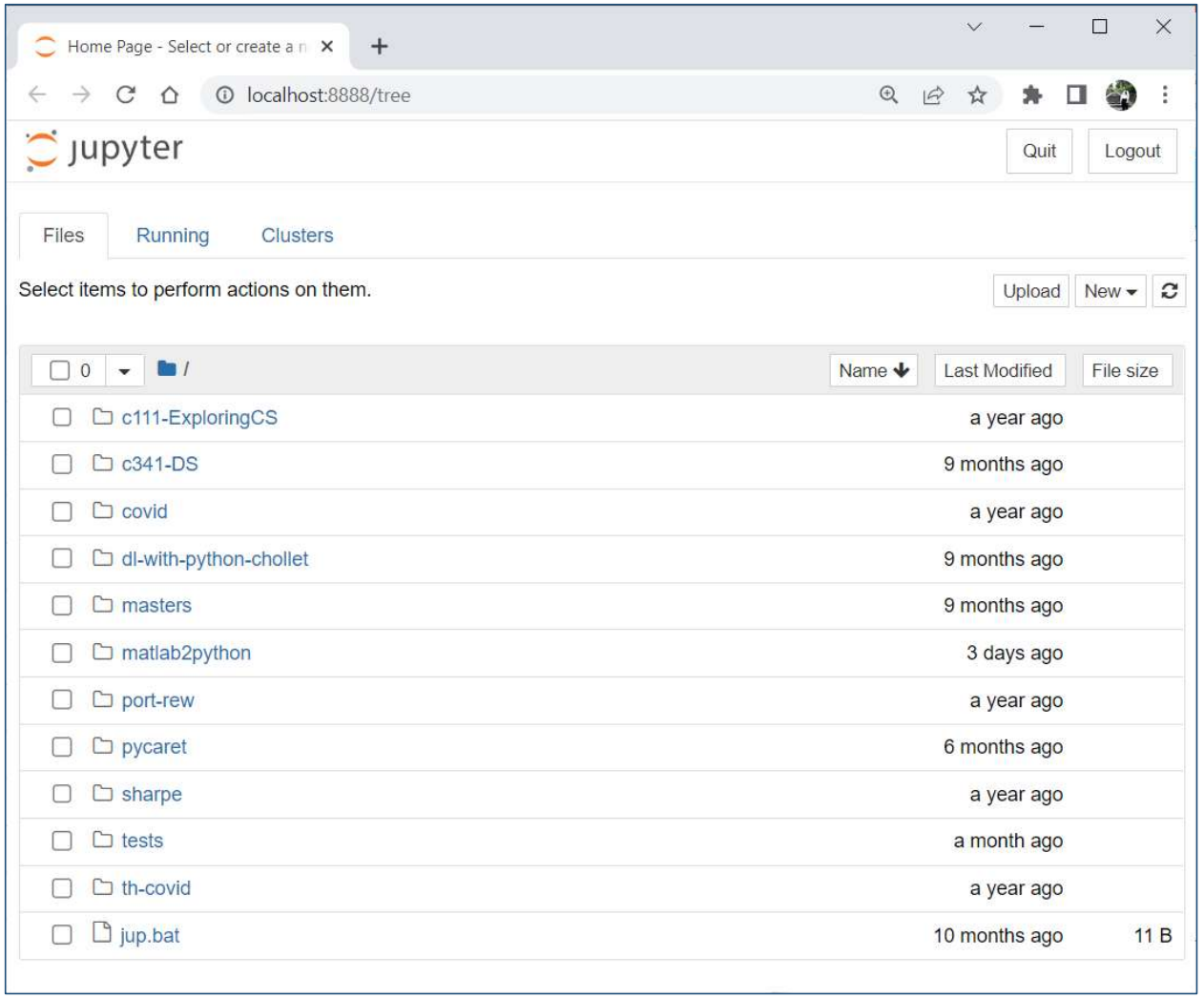
```
Administrator: Command Prompt
(c) Microsoft Corporation. All rights reserved.

C:\Windows\system32>pip install numpy pandas matplotlib sklearn scipy
Requirement already satisfied: numpy in c:\users\chuki\appdata\roaming\python\python38\site-packages (1.19.5)
Requirement already satisfied: pandas in c:\users\chuki\appdata\roaming\python\python38\site-packages (1.3.1)
Requirement already satisfied: matplotlib in c:\users\chuki\appdata\roaming\python\python38\site-packages (3.4.2)
Requirement already satisfied: sklearn in c:\users\chuki\appdata\roaming\python\python38\site-packages (0.0)
Requirement already satisfied: scipy in c:\users\chuki\appdata\roaming\python\python38\site-packages (1.5.4)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\chuki\appdata\roaming\python\python38\site-packa
Requirement already satisfied: pytz>=2017.3 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\chuki\appdata\roaming\python\python38\site-packages (
Requirement already satisfied: pillow>=6.2.0 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from
Requirement already satisfied: cycler>=0.10 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\chuki\appdata\roaming\python\python38\site-packages (f
Requirement already satisfied: scikit-learn in c:\users\chuki\appdata\roaming\python\python38\site-packages (from
Requirement already satisfied: six in c:\program files\python38\lib\site-packages (from cycler>=0.10->matplotlib)
Requirement already satisfied: joblib>=0.11 in c:\users\chuki\appdata\roaming\python\python38\site-packages (from
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\chuki\appdata\roaming\python\python38\site-package
WARNING: You are using pip version 22.0.3; however, version 22.0.4 is available.
You should consider upgrading via the 'c:\program files\python38\python.exe -m pip install --upgrade pip' command.

C:\Windows\system32>
```

Run Jupyter notebook

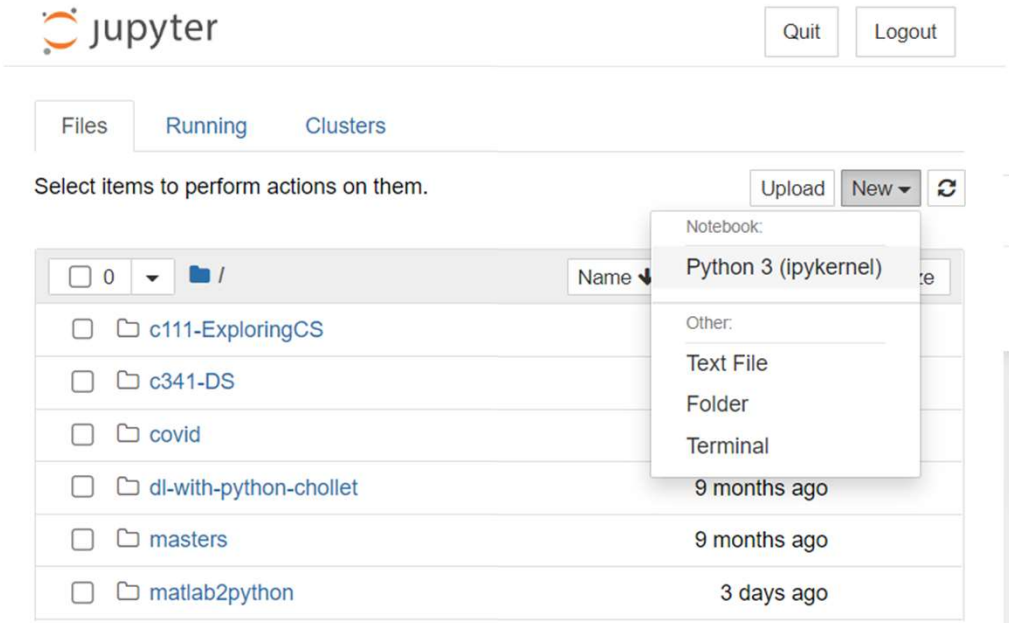


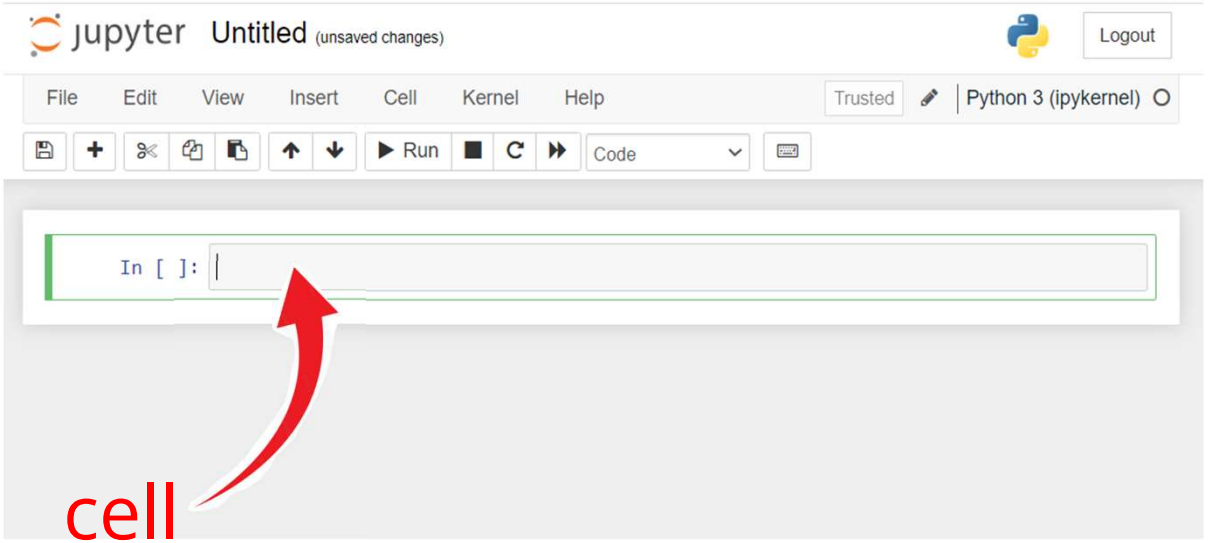
Jupyter notebook home





Create a new Python *notebook* file

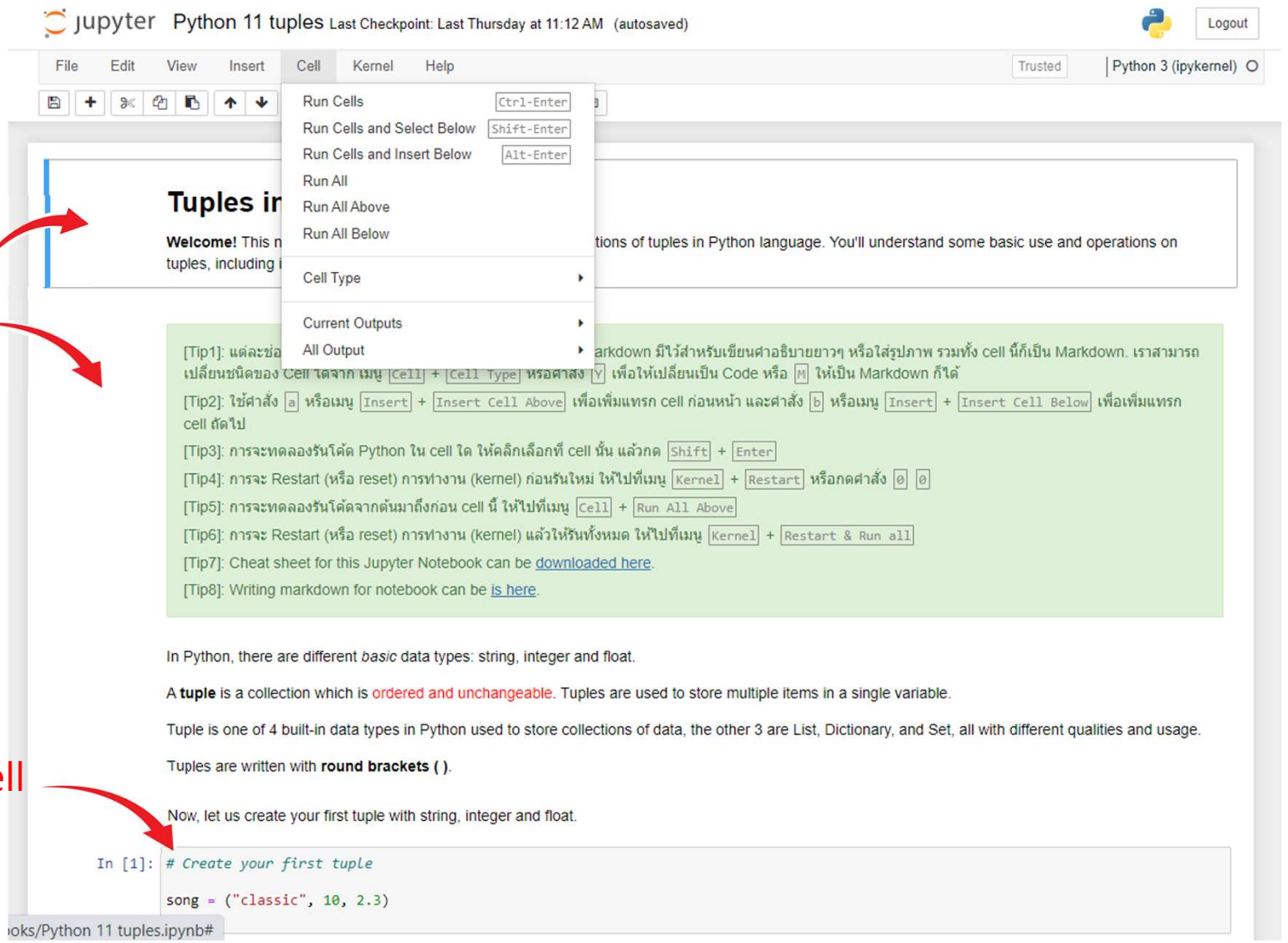




Code cells and markdown cells

Markdown cell

Code cell



Notebook files for learning

Name	Date modified	Type	Size
.ipynb_checkpoints	11/8/2565 10:08	File folder	
Python 04 strings.ipynb	28/6/2565 9:23	IPYNB File	33 KB
Python 11 tuples.ipynb	11/8/2565 11:12	IPYNB File	83 KB
Python 12 lists.ipynb	28/6/2565 9:25	IPYNB File	203 KB
Python 13 dictionaries.ipynb	28/6/2565 9:26	IPYNB File	98 KB
Python 21 conditions.ipynb	28/6/2565 10:14	IPYNB File	1,346 KB
Python 22 loops.ipynb	28/6/2565 10:15	IPYNB File	167 KB
Python 23 functions.ipynb	28/6/2565 10:16	IPYNB File	122 KB
Python 24 classes.ipynb	28/6/2565 10:20	IPYNB File	25 KB
Python 31 numpy1D.ipynb	28/6/2565 10:22	IPYNB File	74 KB
Python 32 numpy2D.ipynb	28/6/2565 10:26	IPYNB File	26 KB

Outline

- Review basic Python
- Tools esp. Jupyter notebook
- Introduce numpy and pandas libraries

numpy

- A Python open-source library, created in 2005 by Travis Oliphant, used for working with arrays.
- NumPy stands for Numerical Python.
- Widely used in domain of linear algebra, matrices, Fourier transform, image processing.
- Mainly used for n-dimensional arrays.
- Implemented in C, and thus is up to 50x faster than traditional Python lists.
- Great place for further study: https://www.w3schools.com/python/numpy/numpy_intro.asp
- http://mscbio2025.csb.pitt.edu/notes/numpy.slides.html#/
- https://cloudxlab.com/blog/numpy-pandas-introduction/
- https://www.codecademy.com/article/introduction-to-numpy-and-pandas

pandas

- pandas is a fast, powerful, flexible and easy to use open-source data analysis and manipulation tool.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- *pandas* stands for "*Python Data Analysis*" and was created by Wes McKinney in 2008.
- The source code for Pandas: https://github.com/pandas-dev/pandas

- Great place for further study: https://www.w3schools.com/python/pandas/default.asp
- https://pandas.pydata.org/

What we've learnt today

- Review basic Python
- Tools esp. Jupyter notebook
- Introduce numpy and pandas libraries