Knowledge Discovery and Data Mining

Lab 1 Introduction to Python, Anaconda Jupyter Environment

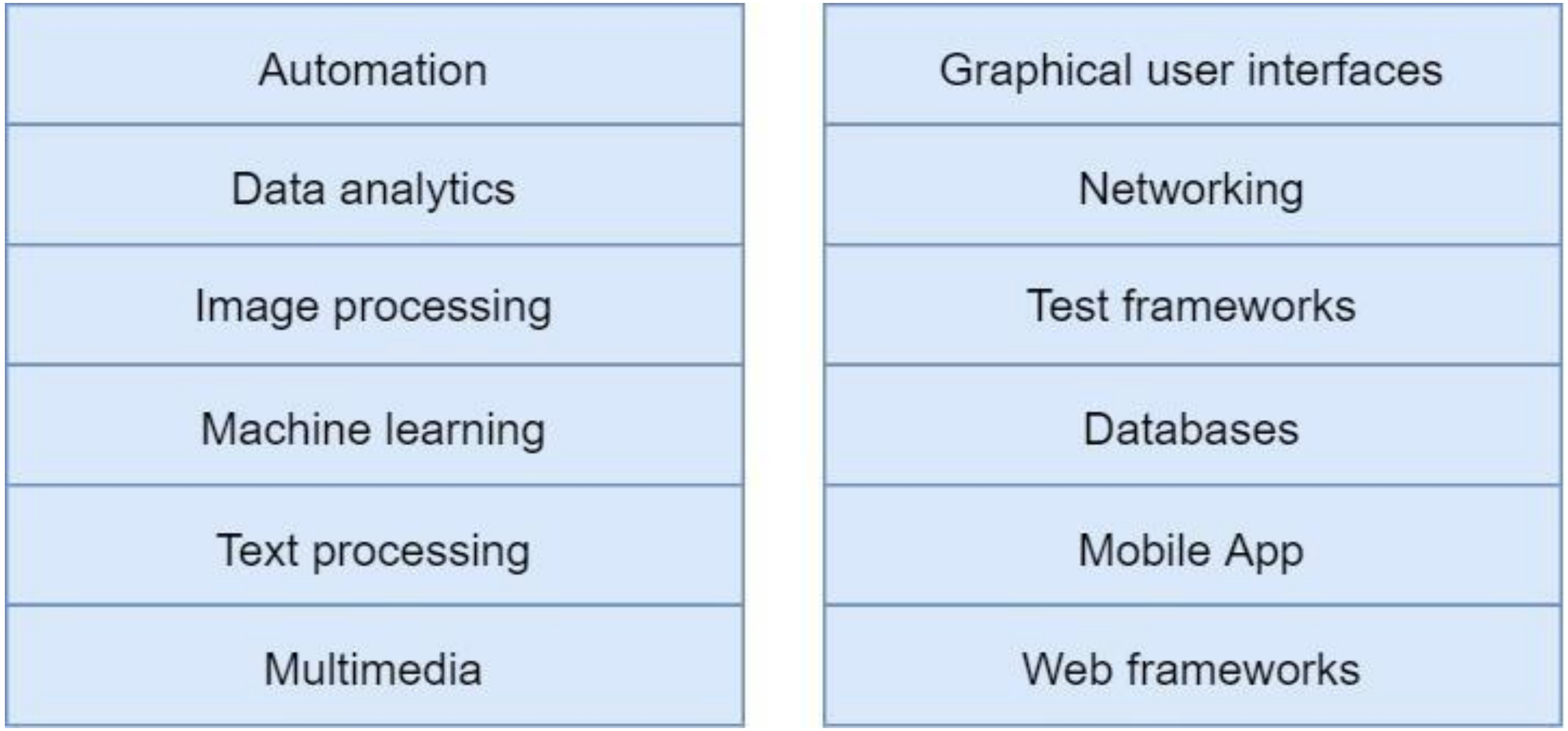
Tianyue Zheng
zhengty@sustech.edu.cn

Python

- Python is an **interpreted**, **high-level** and **general-purpose** programming language.
- Created by Guido van Rossum and first released in 1991.
- Aims to help programmers write clear, logical code for small and large-scale projects.

Why to Learn Python?

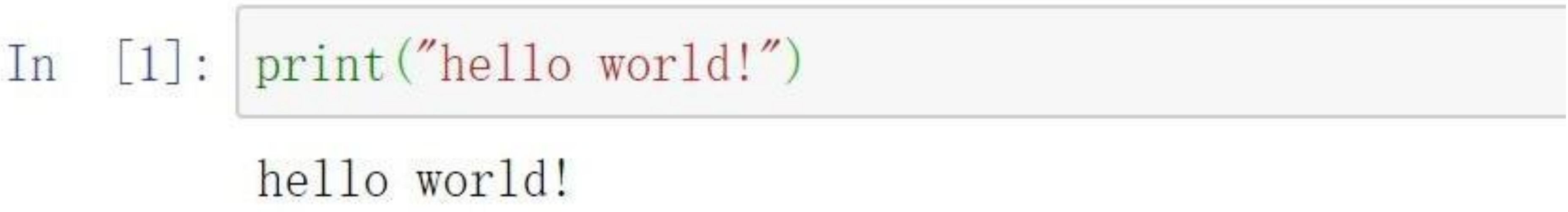
- Easy to learn
- Easy to read
- **Large standard library**



Automation	Graphical user interfaces
Data analytics	Networking
Image processing	Test frameworks
Machine learning	Databases
Text processing	Mobile App
Multimedia	Web frameworks

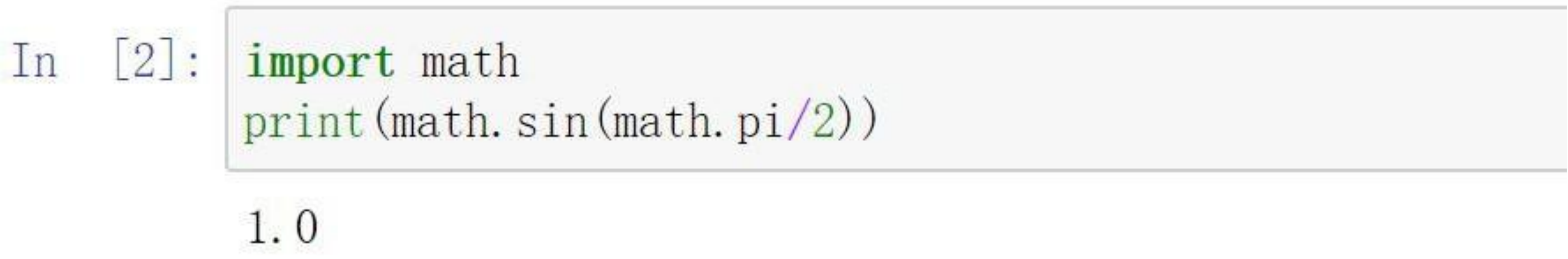
Python Programming Examples

- Example 1



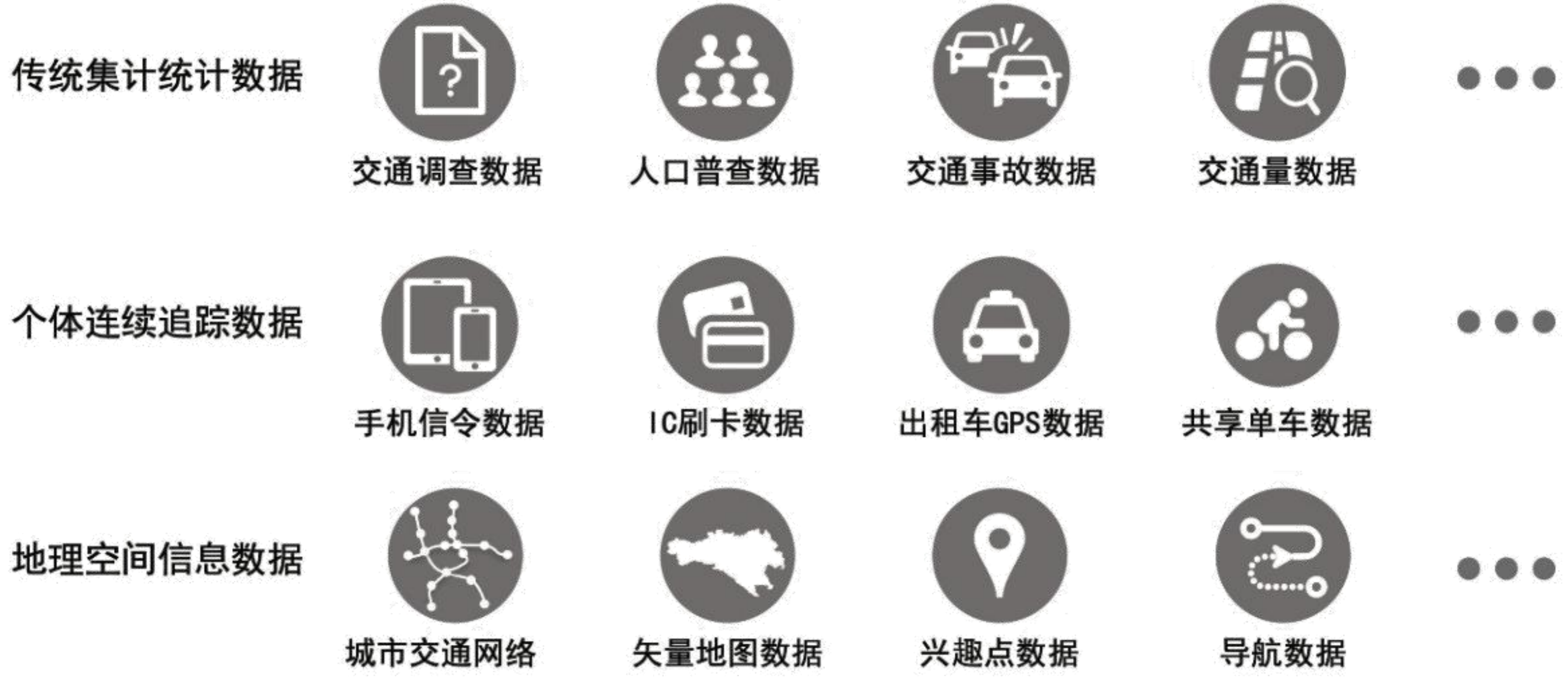
```
In [1]: print(”hello world!”)

hello world!
```

- Example 2



```
In [2]: import math
        print(math.sin(math.pi/2))

1.0
```

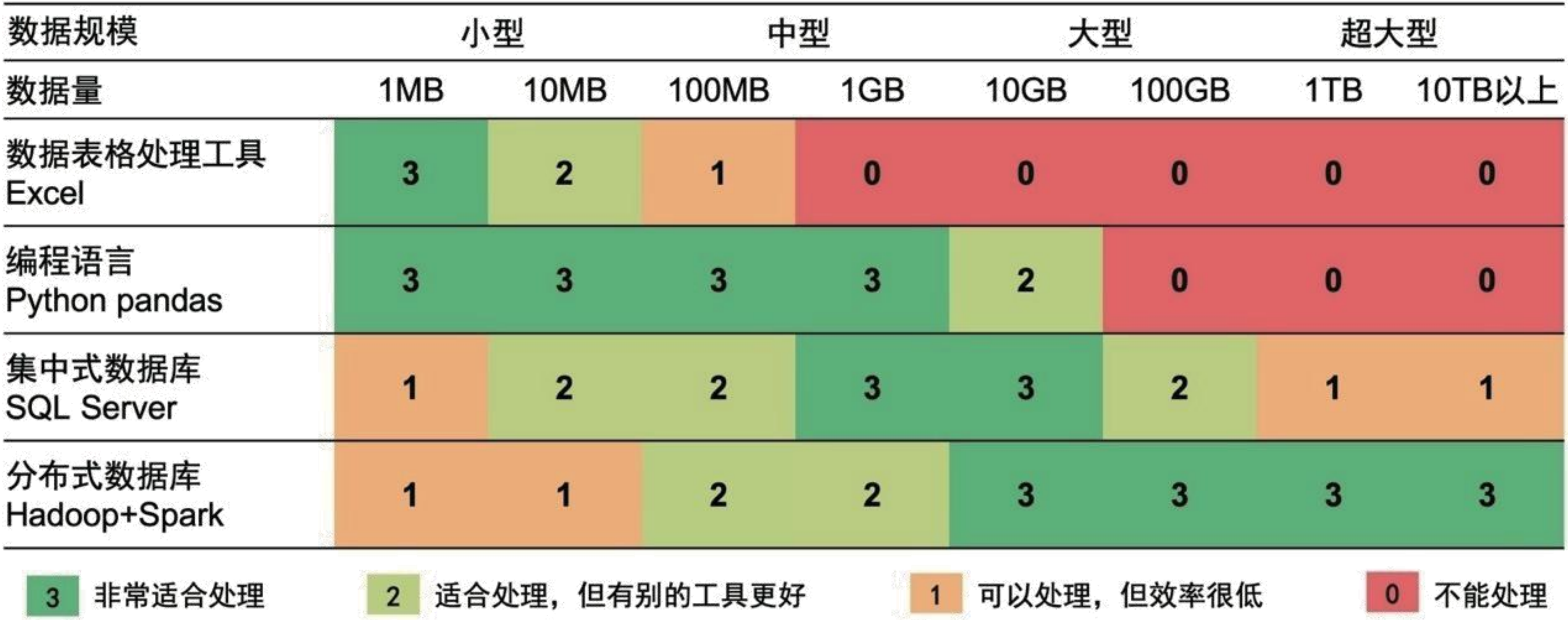
Types of Big Data(example)



传统集计统计数据	交通调查数据	人口普查数据	交通事故数据	交通量数据	•••
个体连续追踪数据	手机信令数据	IC刷卡数据	出租车GPS数据	共享单车数据	•••
地理空间信息数据	城市交通网络	矢量地图数据	兴趣点数据	导航数据	•••

Retrieved from :余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

Data Processing Tools



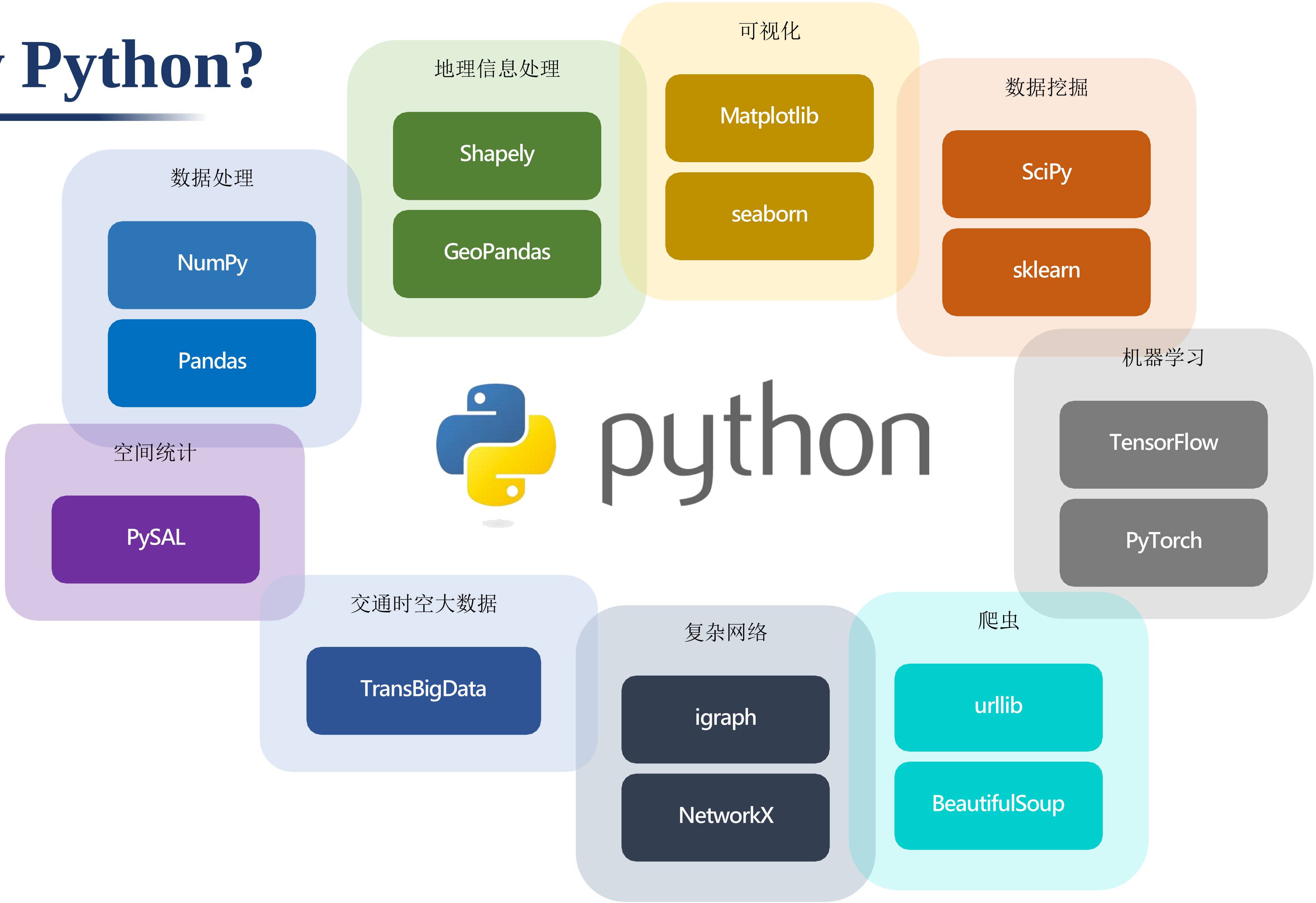
数据规模	小型		中型		大型		超大型	
数据量	1MB	10MB	100MB	1GB	10GB	100GB	1TB	10TB以上
数据表格处理工具 Excel	3	2	1	0	0	0	0	0
编程语言 Python pandas	3	3	3	3	2	0	0	0
集中式数据库 SQL Server	1	2	2	3	3	2	1	1
分布式数据库 Hadoop+Spark	1	1	2	2	3	3	3	3

3 非常适合处理　2 适合处理，但有别的工具更好　1 可以处理，但效率很低　0 不能处理

Excel最大仅支持104万行数据！

Retrieved from :余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

Why Python?



数据处理
- NumPy
- Pandas

地理信息处理
- Shapely
- GeoPandas

可视化
- Matplotlib
- seaborn

数据挖掘
- SciPy
- sklearn

机器学习
- TensorFlow
- PyTorch

空间统计
- PySAL

交通时空大数据
- TransBigData

复杂网络
- igraph
- NetworkX

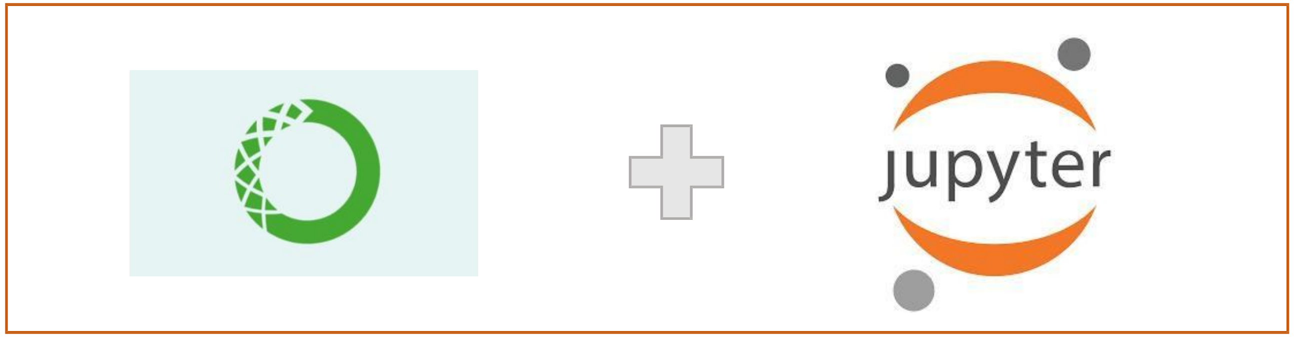
爬虫
- urllib
- BeautifulSoup

python

Retrieved from :余庆，李玮峰《交通时空大数据分析、挖掘与可视化》

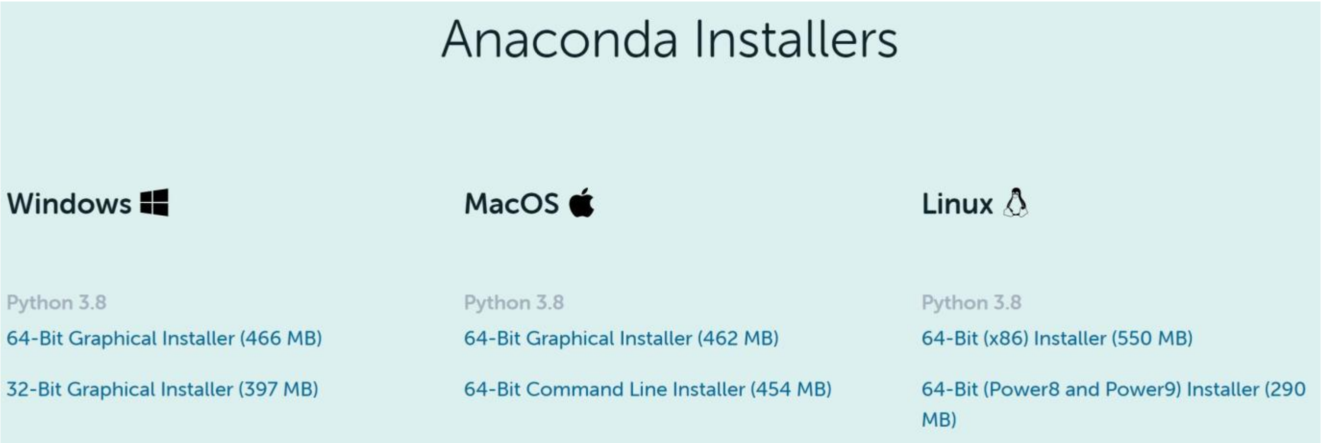
Python Environment

Recommended



Install Anaconda

- Installation of Anaconda



Anaconda Installers

Windows

Python 3.8

64-Bit Graphical Installer (466 MB)

32-Bit Graphical Installer (397 MB)

MacOS

Python 3.8

64-Bit Graphical Installer (462 MB)

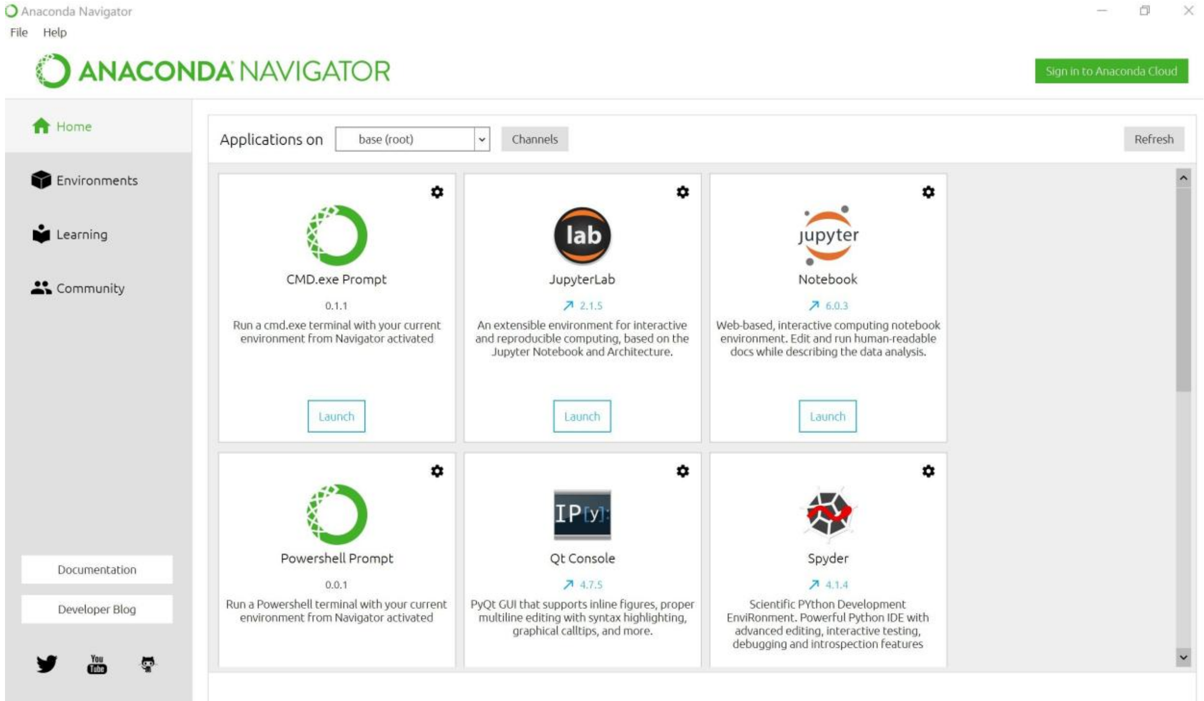
64-Bit Command Line Installer (454 MB)

Linux

Python 3.8

64-Bit (x86) Installer (550 MB)

64-Bit (Power8 and Power9) Installer (290 MB)



CMD.exe Prompt 0.1.1: Run a cmd.exe terminal with your current environment from Navigator activated

JupyterLab 2.1.5: An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook 6.0.3: Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Powershell Prompt 0.0.1: Run a Powershell terminal with your current environment from Navigator activated

Qt Console 4.7.5: PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder 4.1.4: Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

```
conda create -n your_env_name
python=x.x  activate your_env_name
```

Install Jupyter Notebook

- Installation of Jupyter notebook
  - Installing Jupyter using Anaconda and conda
  - Installing Jupyter with pip

If you have any problem to install Jupyter notebook, you can refer to the following websites:

(1) *https://jupyter.readthedocs.io/en/latest/install/notebook-classic.html*

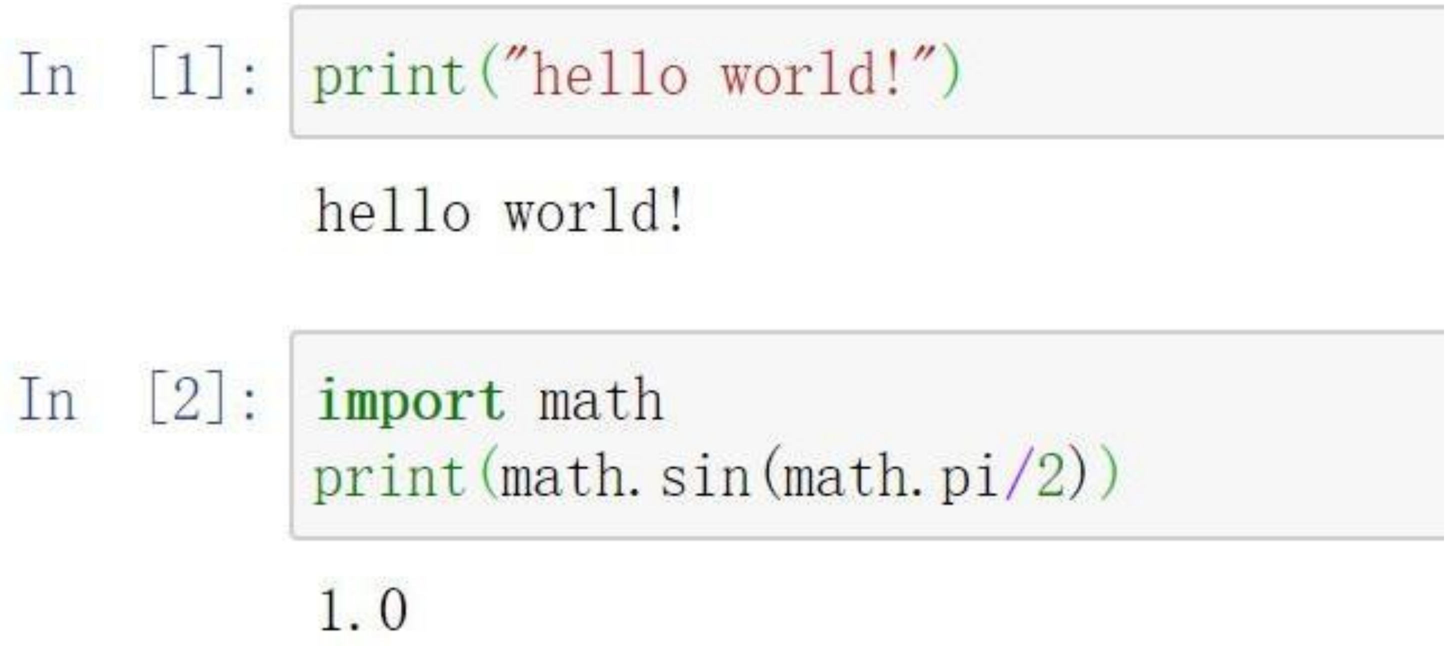
(2) *https://www.jianshu.com/p/91365f343585*

Try to Install Packages

- Install some packages
  - pandas
  - numpy
  - matplotlib
  - scikit-learn

Try to Use Jupyter Notebook

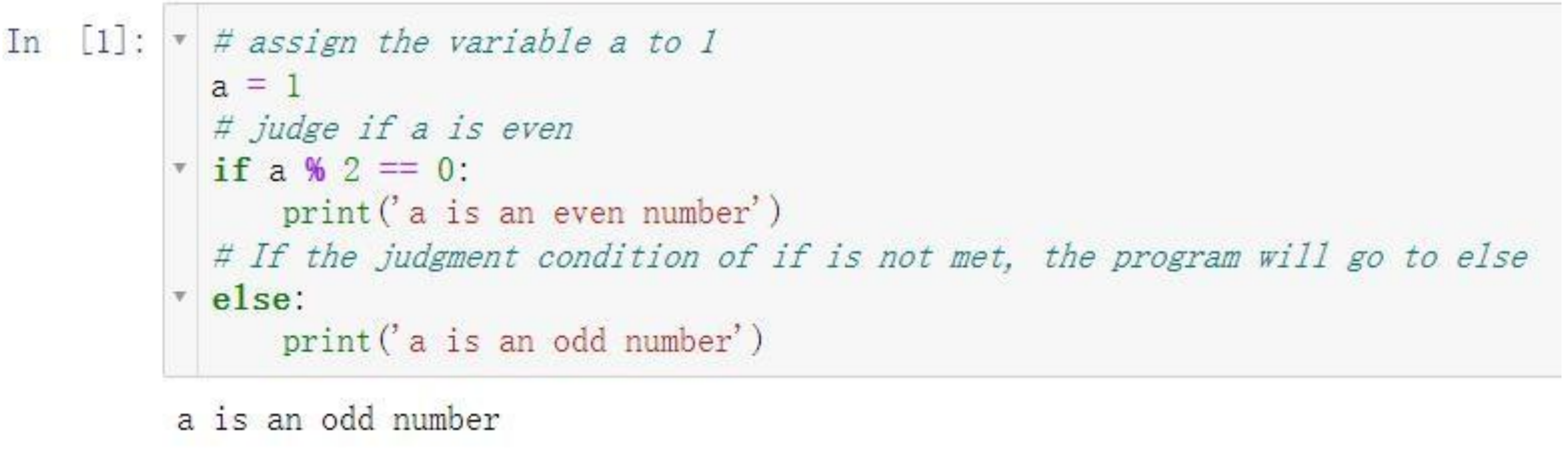
- Implement the sample code mentioned in the previous slides.



```
In  [1]: print(″hello world!″)

         hello world!

In  [2]: import math
         print(math.sin(math.pi/2))

         1.0
```

Python Conditional Statement Examples

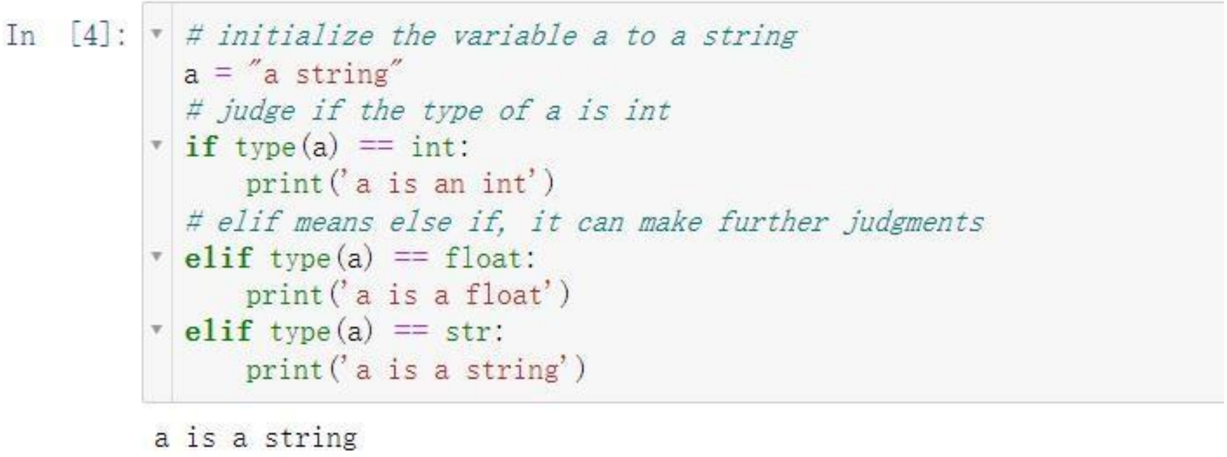
- Example 1



```python
# assign the variable a to 1
a = 1
# judge if a is even
if a % 2 == 0:
    print('a is an even number')
# If the judgment condition of if is not met, the program will go to else
else:
    print('a is an odd number')
```

```
a is an odd number
```

- Example 2

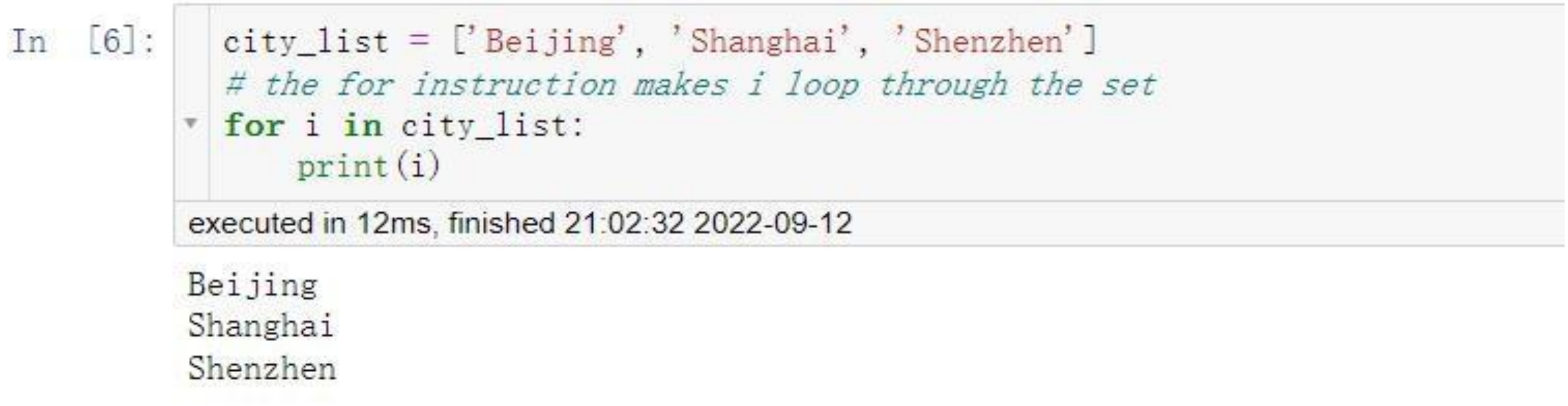


```python
# initialize the variable a to a string
a = "a string"
# judge if the type of a is int
if type(a) == int:
    print('a is an int')
# elif means else if, it can make further judgments
elif type(a) == float:
    print('a is a float')
elif type(a) == str:
    print('a is a string')
```

```
a is a string
```

Python Loop Examples

- Example 1

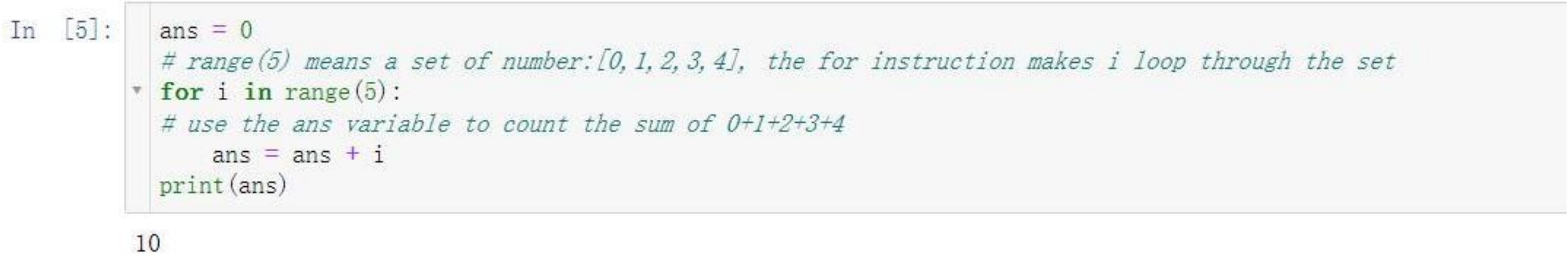


```
In [6]:
city_list = ['Beijing', 'Shanghai', 'Shenzhen']
# the for instruction makes i loop through the set
for i in city_list:
    print(i)
```

executed in 12ms, finished 21:02:32 2022-09-12

```
Beijing
Shanghai
Shenzhen
```

- Example 2

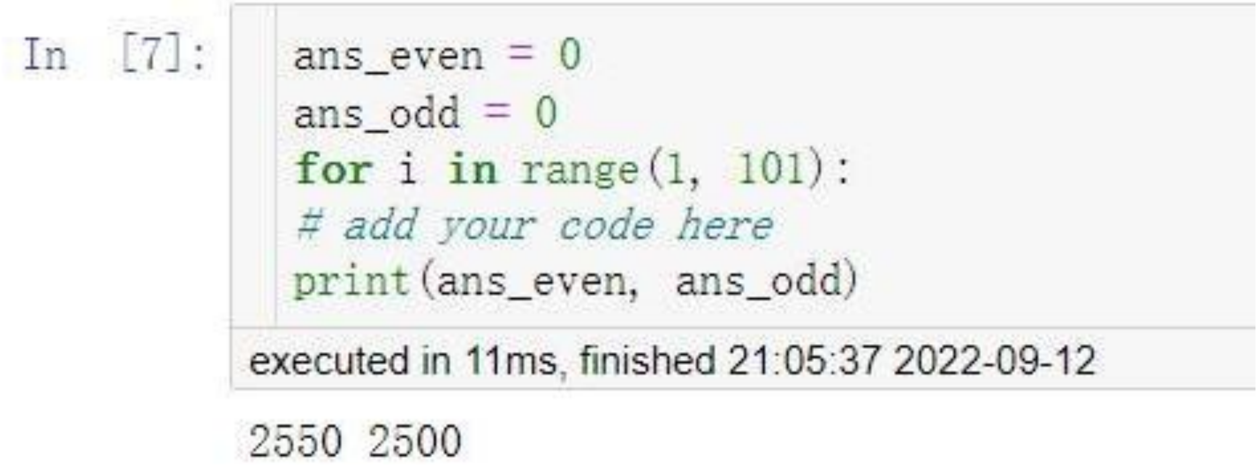


```
In [5]:
ans = 0
# range(5) means a set of number:[0,1,2,3,4], the for instruction makes i loop through the set
for i in range(5):
# use the ans variable to count the sum of 0+1+2+3+4
    ans = ans + i
print(ans)
```

```
10
```

Exercise1

- Calculate the sum of all odd and even numbers from 1 to 100.



In [7]:

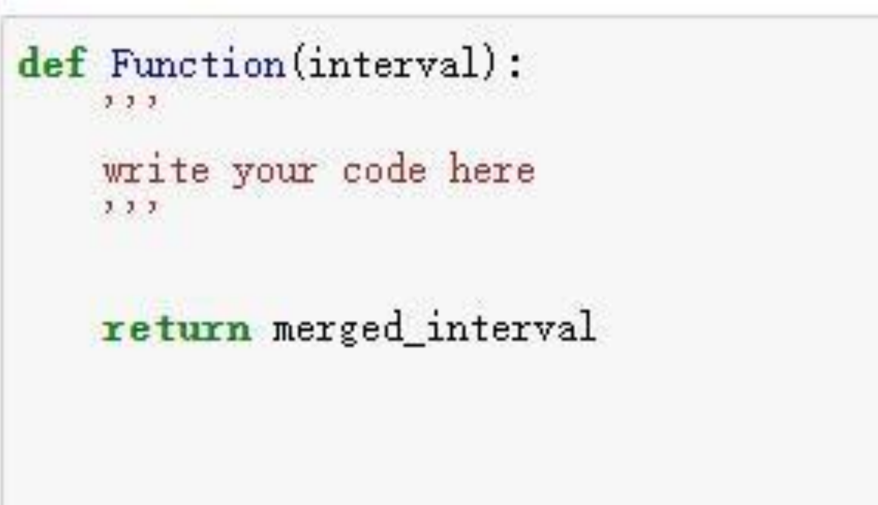
```
ans_even = 0
ans_odd = 0
for i in range(1, 101):
# add your code here
print(ans_even, ans_odd)
```

executed in 11ms, finished 21:05:37 2022-09-12

```
2550 2500
```

Exercise2

- Implement a function in Python that takes a collection of intervals as input and merges all overlapped intervals as output.



```python
def Function(interval):
    '''
    write your code here
    '''


    return merged_interval
```

```
Example1:
    Input: interval =
    [[1,3],[2,6],[8,10],[15,18]]
    Output: [[1,6],[8,10],[15,18]]


Example2:
    Input: interval =
    [[1,4],[4,5]]  Output:
    [[1,5]]
```

Exercise3

- 1. Reading and writing TXT file in jupyter notebook.
- 2. Reading and writing CSV file in jupyter notebook.

Hints:

1. txt file:

https://www.geeksforgeeks.org/reading-writing-text-files-python/
https://pythonexamples.org/python-read-text-file/

2. csv file

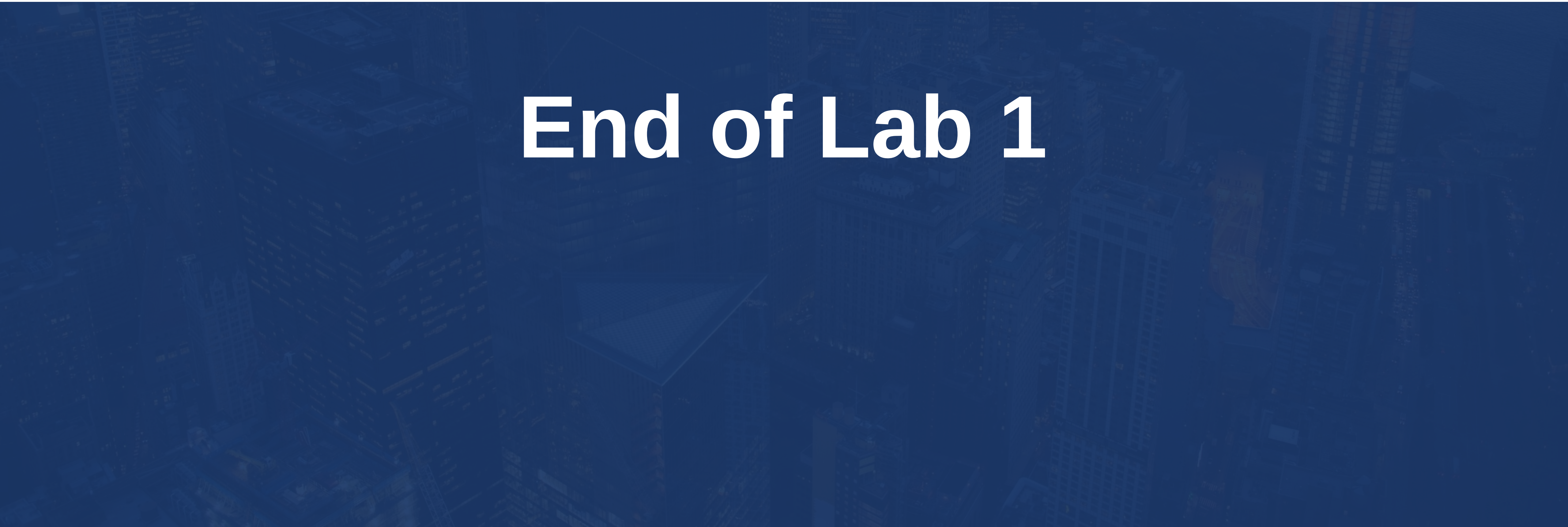
https://realpython.com/python-csv/

Other Resources

- Python:
  - https://www.w3schools.com/python/
  - https://www.runoob.com/python/python-tutorial.html

- Anaconda and Jupyter notebook:
  - https://www.anaconda.com/products/individual/get-started
  - https://blog.csdn.net/zaishuiyifangxym/article/details/83269834
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/
  - https://juejin.im/post/6844903842497167374

- 余庆，李玮峰《交通时空大数据分析、挖掘与可视化》清华大学出版社



End of Lab 1