



Day 5-1 資料清理數據前處理

如何新建一個 dataframe ?



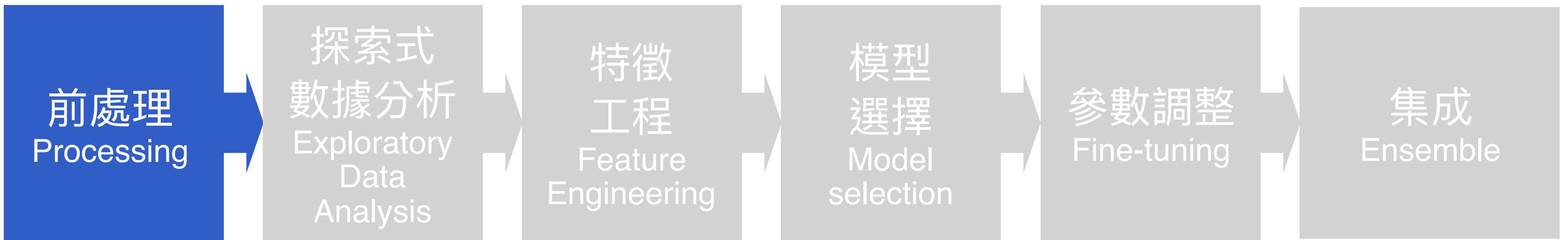
游為翔 / 杜靖愷

出題教練

知識地圖 機器學習前處理 讀取各式資料

機器學習概論 Introduction of Machine Learning

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



前處理 Processing



本日知識點目標

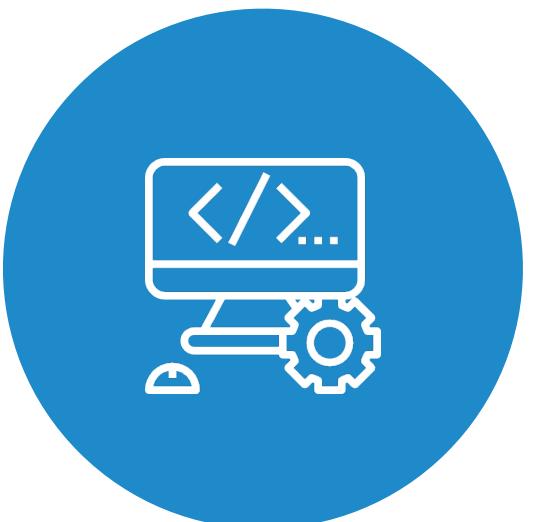
了解如何快速驗證 dataframe 操作的程式碼

為什麼新建一個 dataframe 重要？



需要把分析過程中所產生的數據或者結果儲存為結構化的資料

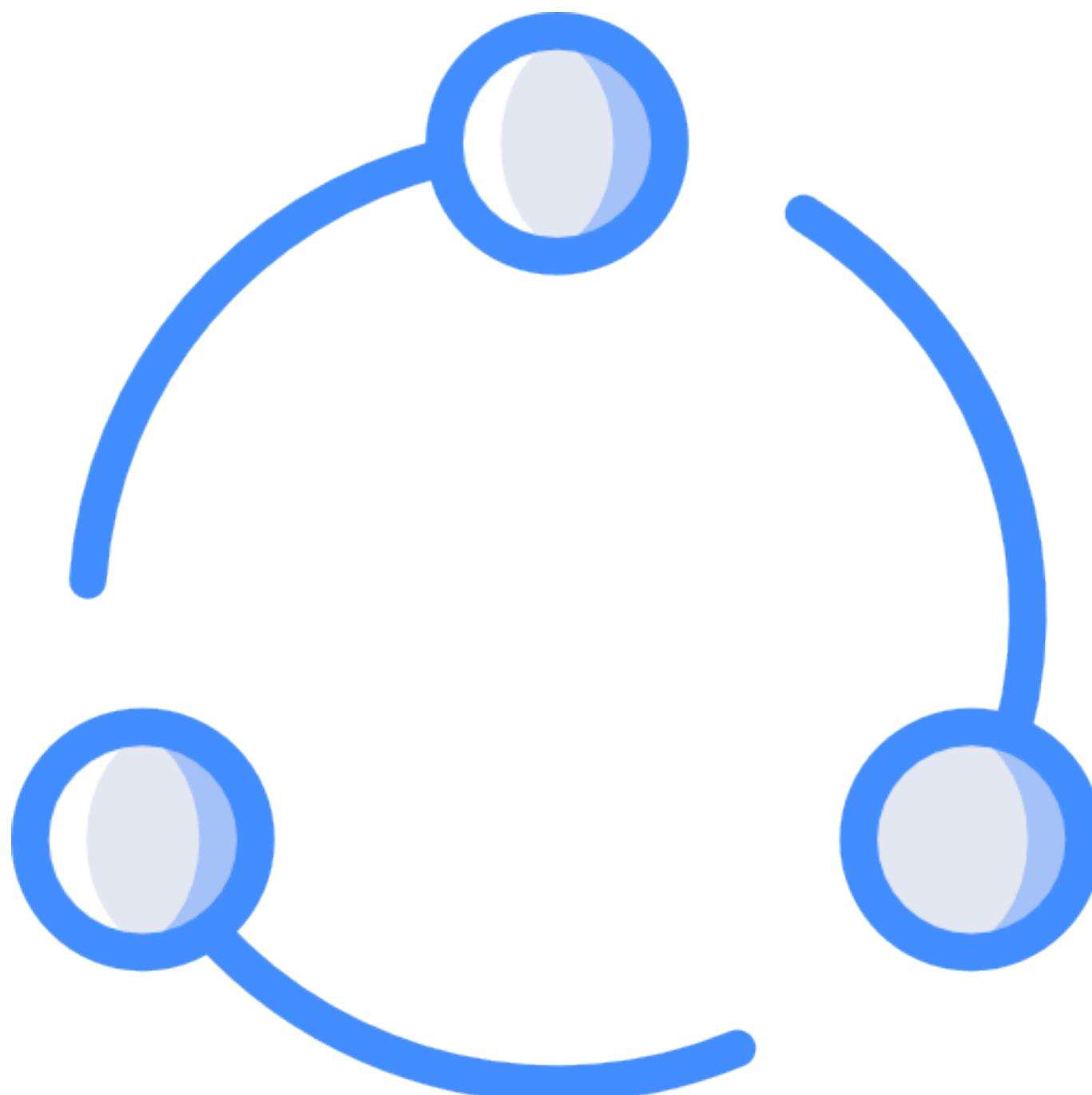
- Ex 1: 將每筆交易資料匯總計算平均值、標準差等統計數值
- Ex 2: Kaggle 比賽要上傳的結果



測試程式碼

- 有時候原始資料太大了，有些資料的操作很費時，先在具有同樣結構的資料上測試程式碼是否能夠得到理想中的結果。
- 不確定視覺化程式碼中所需要的資料結構，用新建立的 dataframe 結構來去了解，而不是急著在原始資料上操作。

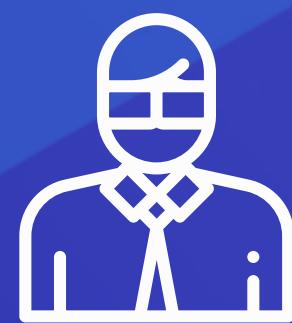
重要知識點複習



- 在資料量很大時，可以先在和資料具有同樣結構的小樣本驗證程式碼執行的結果是否符合預期
- 用 `pd.DataFrame` 來創建一個 `dataframe`
- 用 `np.random.randint` 來產生隨機數值

Day 5-2 資料清理數據前處理

如何讀取其他資料？ (非CSV的資料)



游為翔 / 杜靖愷

出題教練

本日知識點目標

學會如何讀取其他資料格式：txt / jpg / png / json /
mat / npy / pkl ...

讀取其他非csv資料格式？

檔案格式

文本 (txt)

Json

矩陣檔 (mat)

讀取範例

```
with open('example.txt', 'r') as f:  
    data = f.readlines()  
print(data)
```

```
import json  
with open('example.json', 'r') as f:  
    data = json.load(f)  
print(data)
```

```
import scipy.io as sio  
data = sio.loadmat('example.mat')
```

讀取其他非csv資料格式？

檔案格式

圖像檔 (PNG / JPG ...)

Python npy

Pickle (pkl)

讀取範例

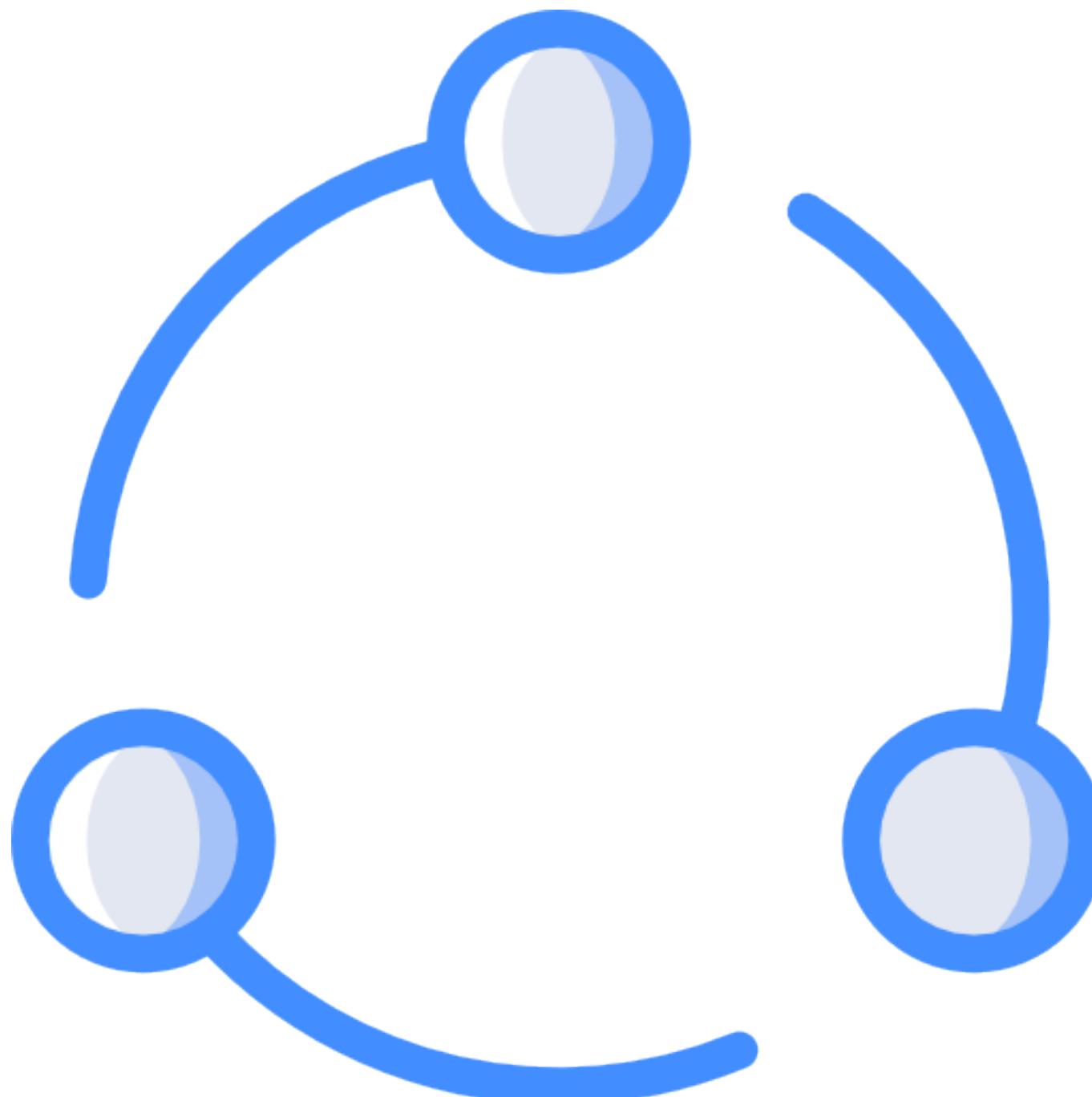
```
Import cv2  
image = cv2.imread(...) # 注意 cv2 會以 BGR 讀入  
image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
```

```
from PIL import Image  
image = Image.read(...)  
import skimage.io as skio  
image = skio.imread(...)
```

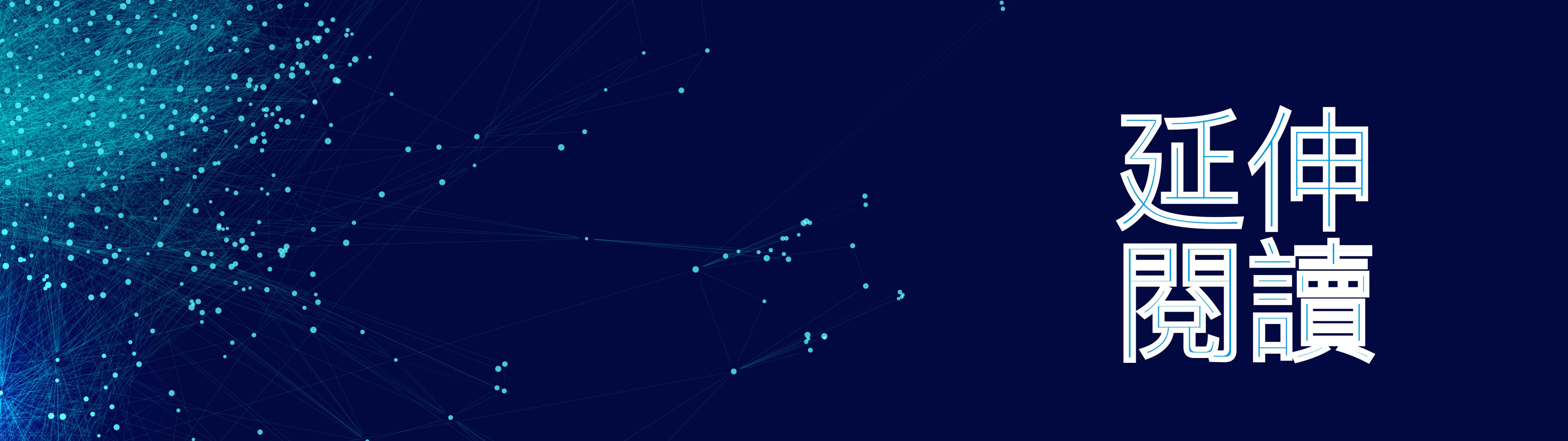
```
import numpy as np  
arr = np.load(example.npy)
```

```
import pickle  
with open('example.pkl', 'rb') as f:  
    arr = pickle.load(f)
```

重要知識點複習



- 不同的資料有不同讀取方式
 - 文字格式通常可以用 `with open()`
 - 圖像格式可以使用 PIL, Skimage 或是 CV2
 - CV2 的速度較快，但須注意讀入的格式為 BGR
 - 其他形式如 npy / pickle 可以儲存經過處理後的資料



延伸 閱讀

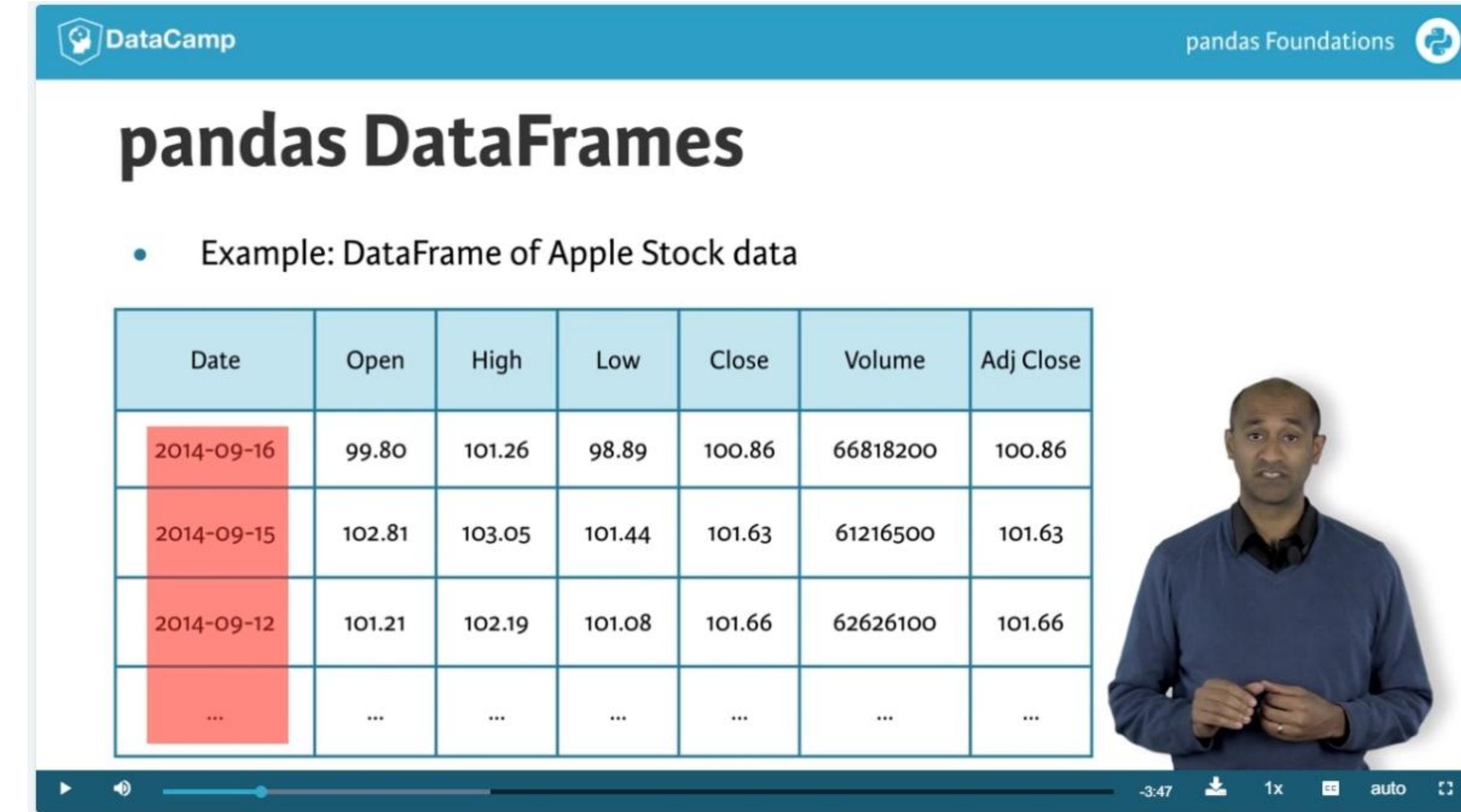
除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有
多餘時間，可再補充延伸閱讀文章內容。

推薦延伸閱讀

Pandas Foundations

網頁連結

第一個 chapter 是免費的，建議可用來預習 pandas，如果覺得英文聽不懂也沒關係，可以按部就班跟著我們後面的課程，也可以學到相關的內容。



The screenshot shows a DataCamp video player interface. At the top, the DataCamp logo is on the left and the course title "pandas Foundations" is on the right. The main content area features the title "pandas DataFrames" in large bold letters. Below the title is a bulleted list: "• Example: DataFrame of Apple Stock data". To the right of the list is a video frame showing a man in a blue sweater speaking. To the left of the video frame is a table titled "Apple Stock Data" with columns: Date, Open, High, Low, Close, Volume, and Adj Close. The first three rows of the table are highlighted in red. The table data is as follows:

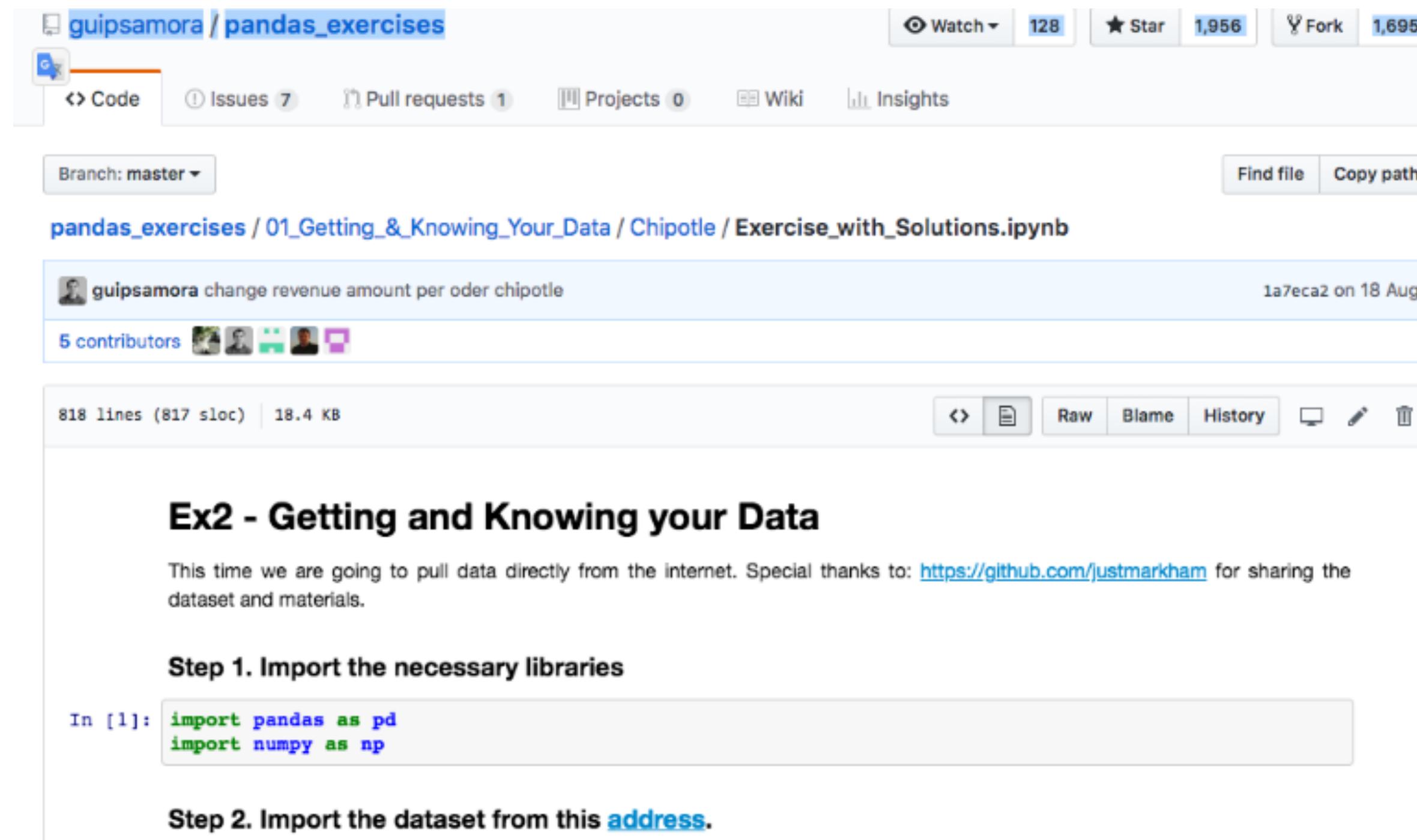
Date	Open	High	Low	Close	Volume	Adj Close
2014-09-16	99.80	101.26	98.89	100.86	66818200	100.86
2014-09-15	102.81	103.05	101.44	101.63	61216500	101.63
2014-09-12	101.21	102.19	101.08	101.66	62626100	101.66
...

At the bottom of the video player are standard controls: a play button, volume control, progress bar, and a set of icons for download, 1x speed, closed captions, auto, and full screen.

推薦延伸閱讀

推薦 github repo

馬拉松之後也會有 pandas 操作相關的練習，但若你迫不及待想要更精進自己 pandas 技能，可以到這個 [github repo](#) 挑戰！



The screenshot shows a GitHub repository page for `guipsamora / pandas_exercises`. The repository has 128 issues, 1 pull request, 0 projects, and 1,956 stars. It also has 1,695 forks. The current branch is `master`. A specific file, `Exercise_with_Solutions.ipynb`, is highlighted. The commit `1a7eca2 on 18 Aug` by `guipsamora` changes revenue amount per order chipotle. There are 5 contributors. The notebook contains 818 lines (817 sloc) and 18.4 KB. The content of the notebook starts with:

```
In [1]: import pandas as pd  
import numpy as np
```

Ex2 - Getting and Knowing your Data

This time we are going to pull data directly from the internet. Special thanks to: <https://github.com/justmarkham> for sharing the dataset and materials.

Step 1. Import the necessary libraries

```
In [1]: import pandas as pd  
import numpy as np
```

Step 2. Import the dataset from this [address](#).



解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

