

Day 3

資料清理數據前處理

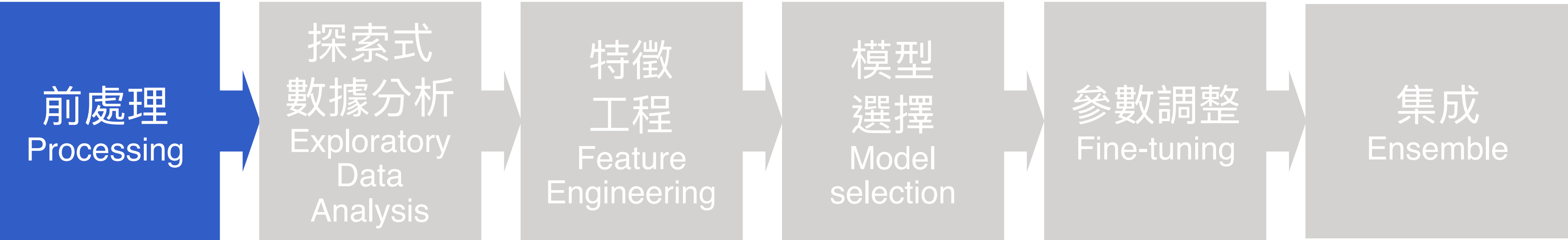
如何新建一個  
dataframe ?



# 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



# 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。



# Day 3-2 資料清理數據前處理

## 如何讀取其他資料？ (非CSV的資料)



# 讀取其他非csv資料格式？

---

## 檔案格式

## 讀取範例

文本 (txt)

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

Json

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

矩陣檔 (mat)

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

# 讀取其他非csv資料格式？

## 檔案格式

## 讀取範例

圖像檔 (PNG / JPG ...)

```
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(...)
```

Python npy

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

Pickle (pkl)

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



# 解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業  
開始解題

