巨匠線上真人

Python 資料科學應用開發

# 第七堂：機器學習概念簡介

# （Introduction to Machine Learning）

# 同學，歡迎你參加本課程

- 請關閉你的FB 、Line等溝通工具，以免影響你上課。
- 考量頻寬、雜音，請預設關閉攝影機、麥克風，若有需要再打開。
- 隨時準備好，老師會呼叫你的名字進行互動，鼓勵用麥克風提問。
- 如果有緊急事情，你必需離開線上教室，請用聊天室私訊給老師，以免老師癡癡呼喚你的名字。
- 軟體安裝請在上課前安裝完成，未完成的同學，請盡快進行安裝。

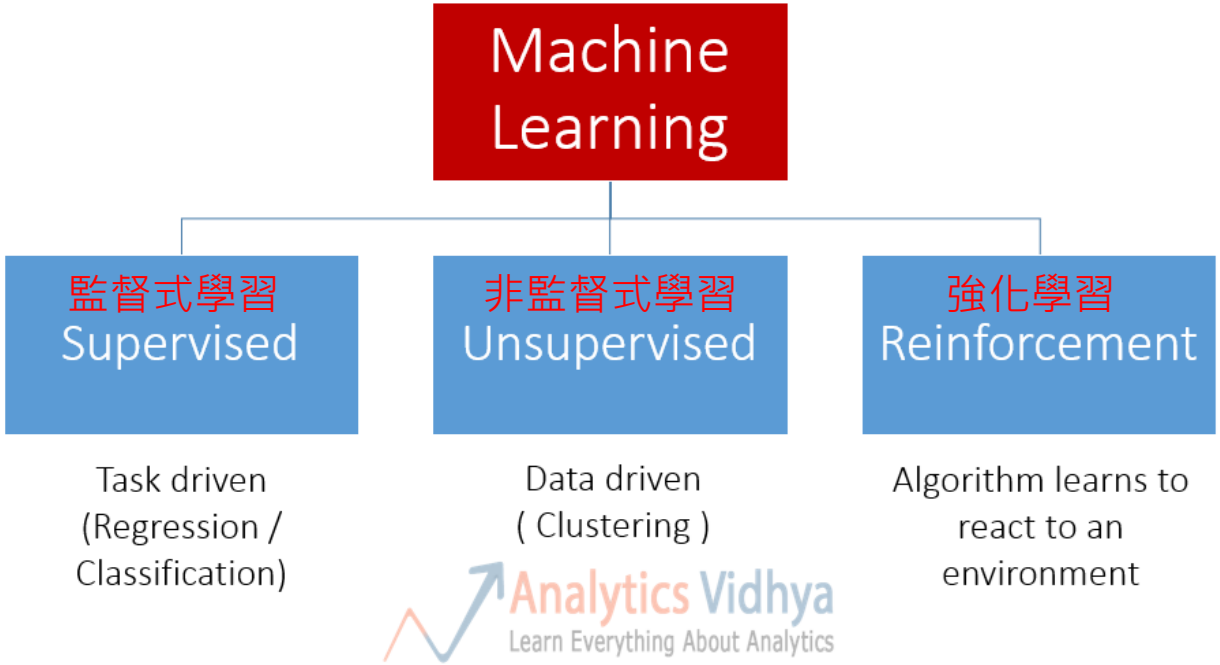
# 課程檔案下載


巨匠電腦線上真人
開課查詢
免費體驗專區
課程總覽
專業師資
學員專區
講師專區
最新消息
您好!
登出
點數卡產品兌換
APCS檢測專區
公告專區
我的課表
IT真人課程劃位
電腦分校課程劃位
外語真人課程劃
美語分校課程劃
取消劃位
課程檔案下載
上課權益查詢
教學平台測試
學習諮詢
常見問題
個資維護
忘記密碼
登出
課程檔案下載
程式語言好難學?
那是因為
你還沒學過Python!
線上老師LIVE直播教學．搶先看
巨匠電腦真人課程

# ZOOM 學員操作說明

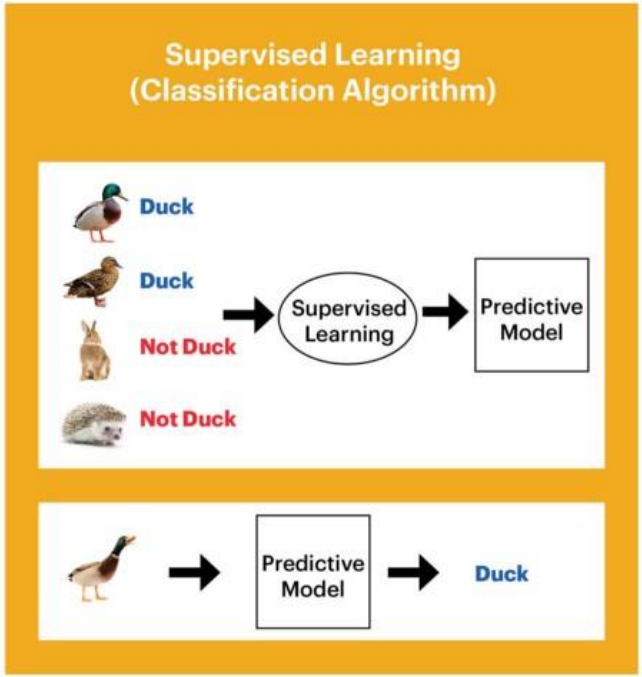

www.pcschool.com.tw
您正在觀看蘇世宗的螢幕
查看選項
原始大小
請求遠端控制
共同註記
退出全螢幕
5 查看選項/共同註記/筆（連連看）
滑鼠
文字
筆
箭頭
擦除
格式
撤銷
重做
清除
與會者 (15)
查找參會者
張齡月 (我)
婷婷
解除靜音
舉手
2 共享螢幕（指導演練；點評作品）
老師須先停止共享螢幕
才能請學生共享螢幕
1 聊天
解除靜音
啟動視訊
邀請
與會者
共享螢幕
聊天
錄影
4 解除靜音
3 與會者/舉手

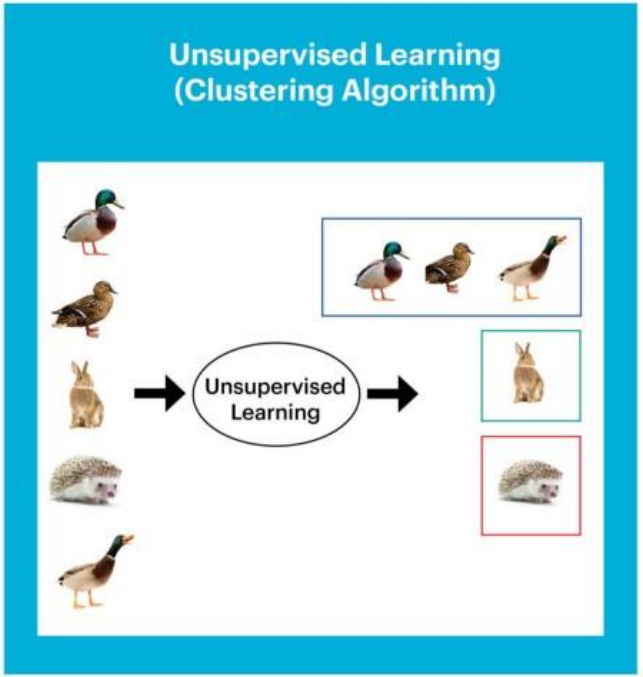
# Types of Machine Learning




http://en.proft.me/2015/12/24/types-machine-learning-algorithms/

# 監督式學習 vs.非監督式學習



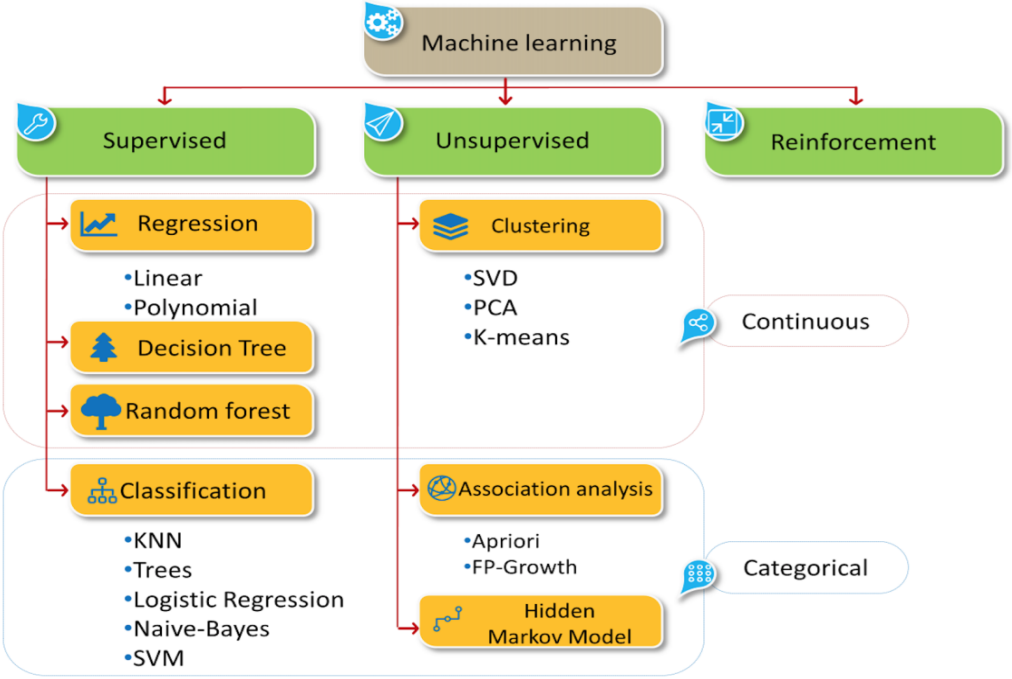




Western Digital.

# 機器學習的區別

| | |
|---|---|
| Supervised Learning | Labeled data<br>Direct feedback<br>Predict outcome/future |
| Unsupervised Learning | No labels/targets<br>No feedback<br>Find hidden structure in data |
| Reinforcement Learning | Decision process<br>Reward system<br>Learn series of actions |

# 機器學習演算法



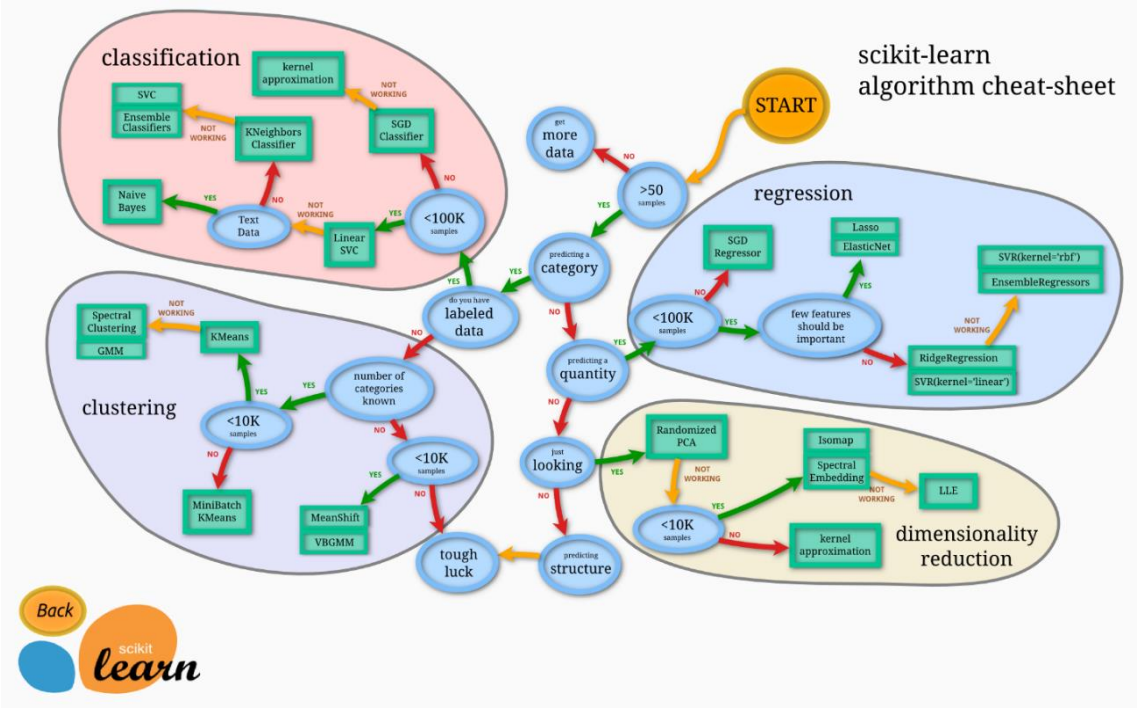

https://yourfreetemplates.com/free-machine-learning-diagram/

# 十大熱門方法（資料探勘 Data Mining）

1. C4.5 (decision tree)
2. k-means (clustering)
3. Support vector machines (next to C4.5, a classifier to try out first)
4. Apriori (association rule learning --> recommendation engine)
5. EM (i.e. expectation-maximization for clustering)
6. PageRank (network analysis; think of the PageRank in Google's search engine)
7. AdaBoost (boosting, and thus an ensemble learning algorithm; taking in and combining multiple learning algorithm)
8. kNN (aka k-Nearest Neighbors, thus classification)
9. Naive Bayes (family of classification algorithms assuming that all features is independent of each other)
10. CART (aka classification and regression trees, thus a classifier)

http://oliviaklose.azurewebsites.net/machine-learning-11-algorithms-explained/

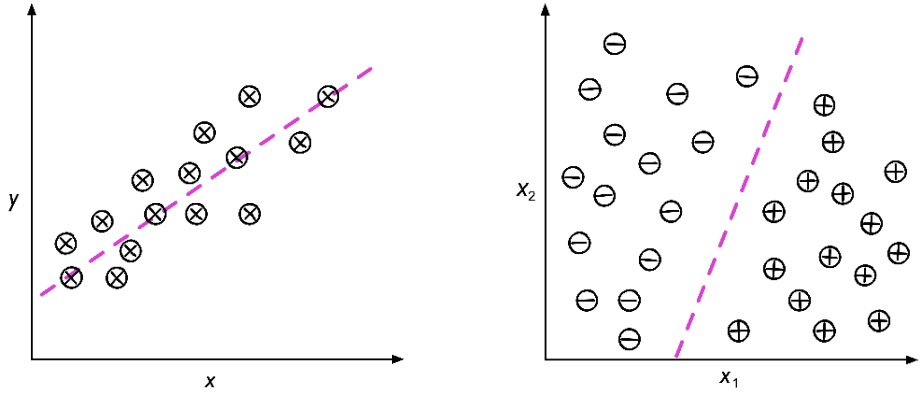
# 機器學習的分類與演算法選擇




http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

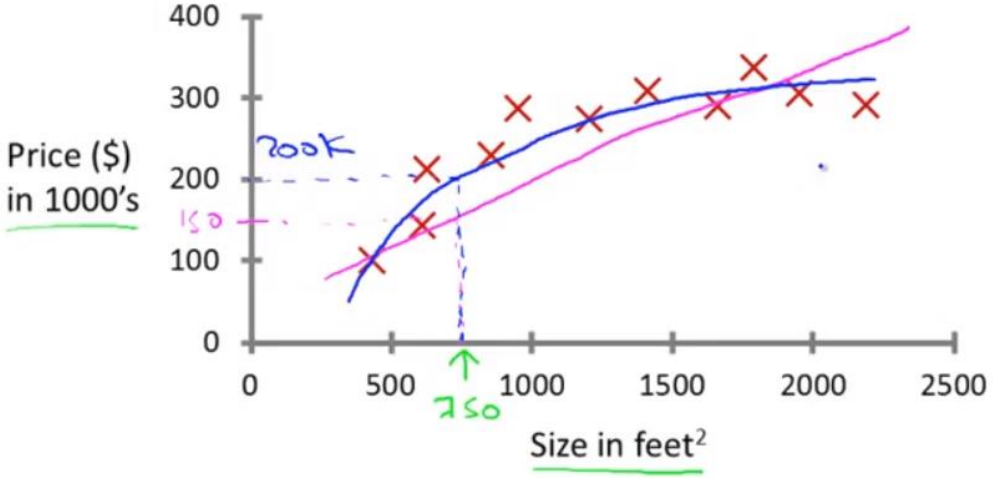
# 監督式學習

- 監督式學習（Supervised learning）分為兩類：
  - regression：predict results within a continuous output, meaning that we are trying to map input variables to some continuous function.
  - classification：predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.



# 迴歸（Regression）

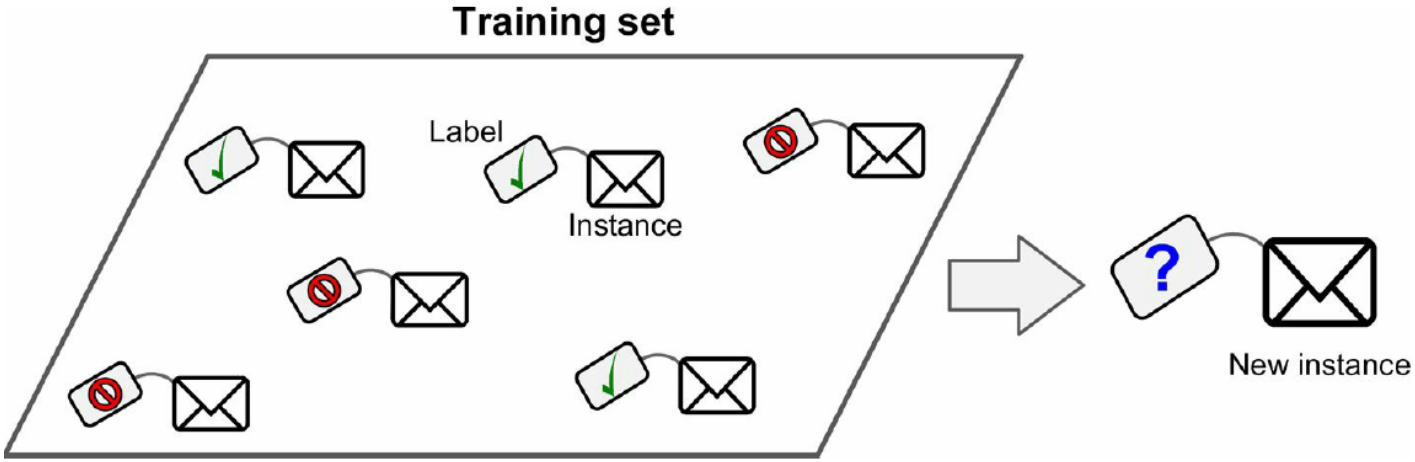
Housing price prediction.



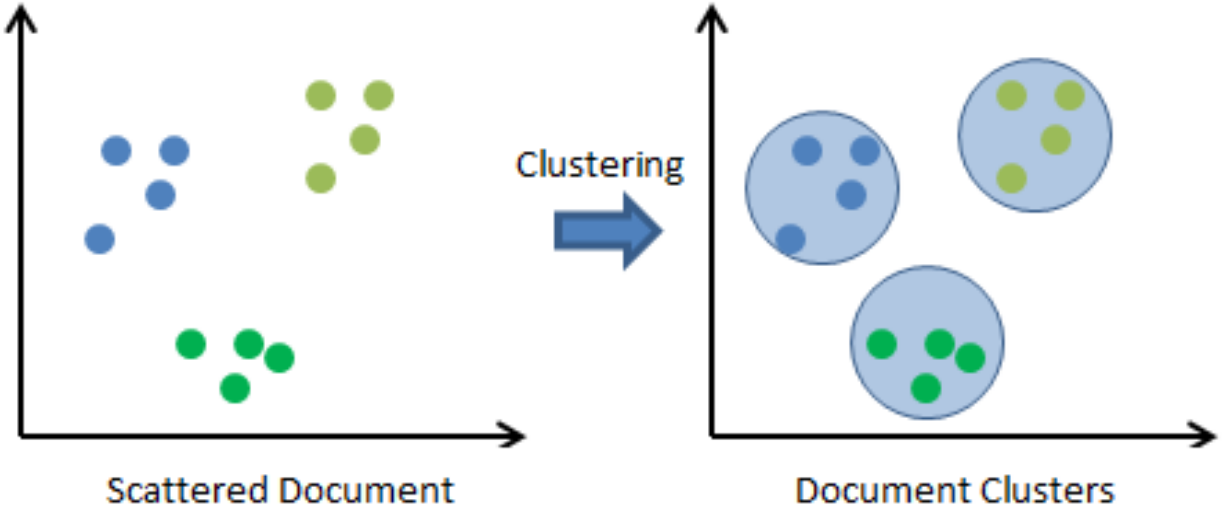

Supervised Learning

"right answers" given

# 分類（Classification）

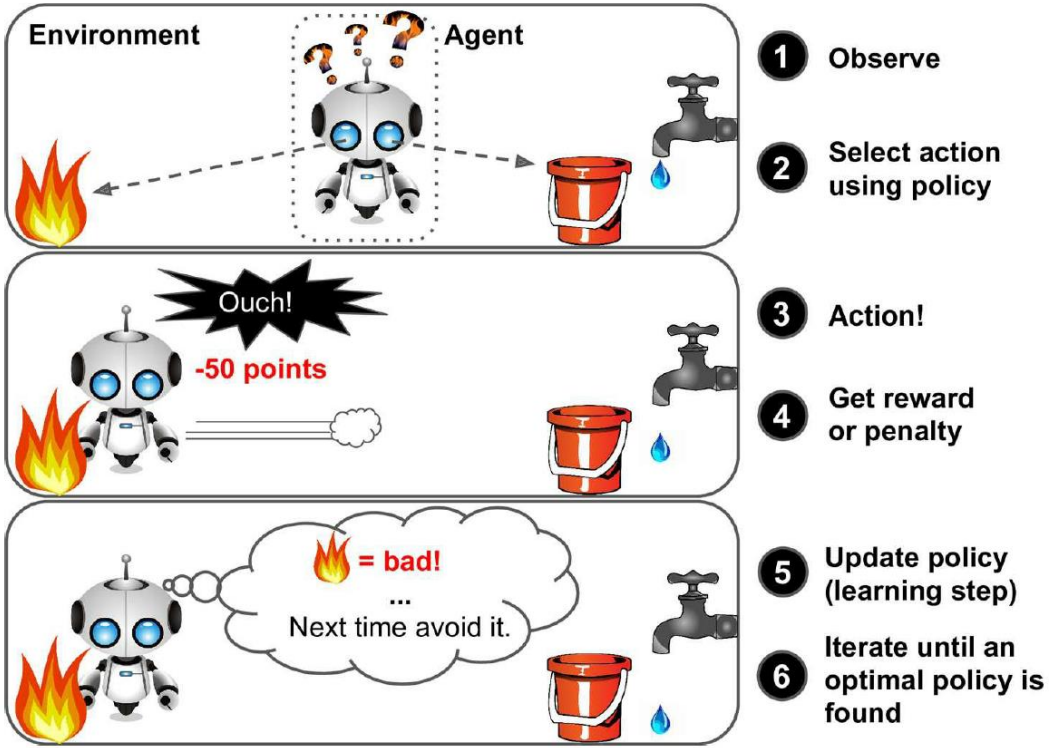


# 集群（Clustering）

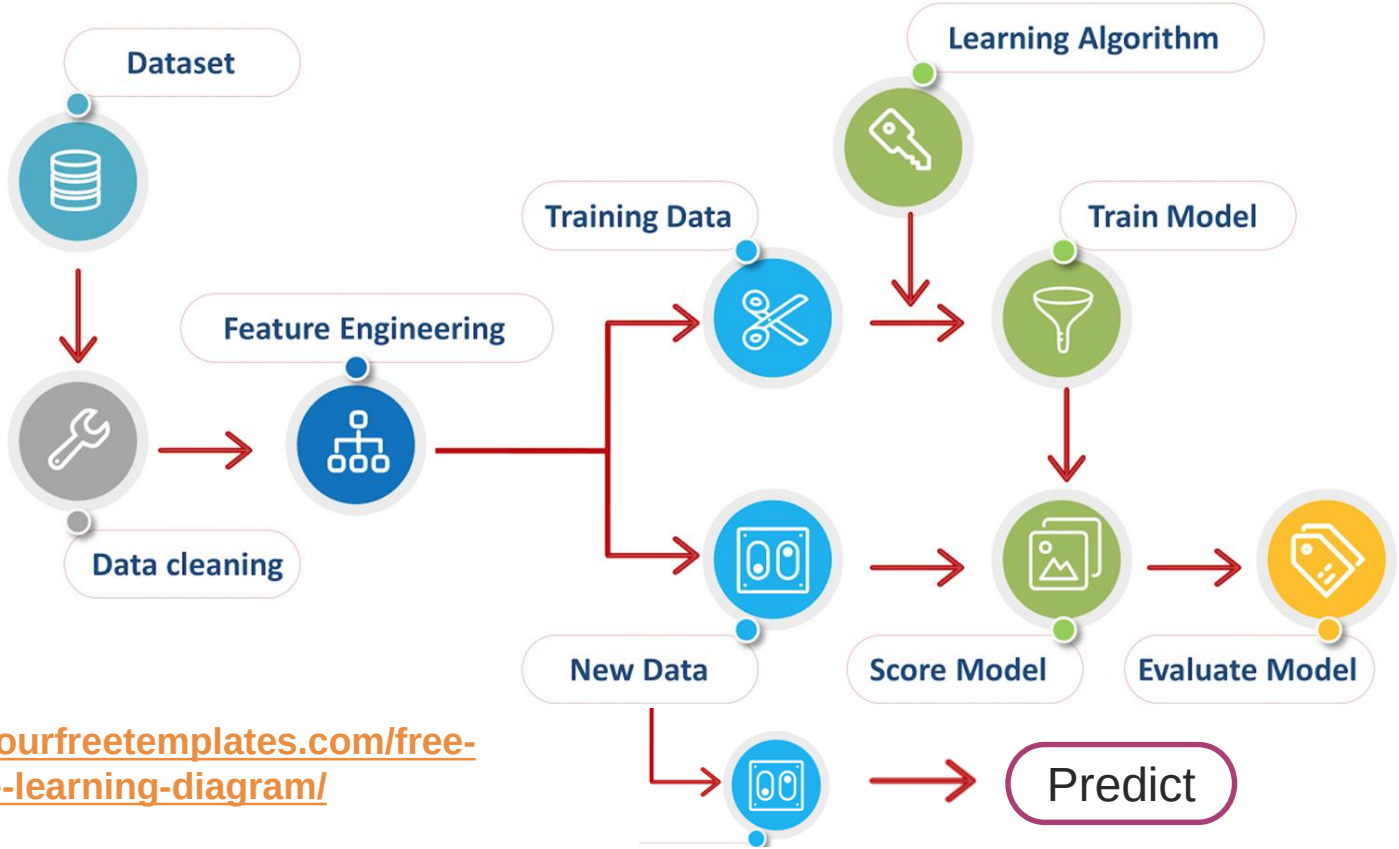



http://en.proft.me/2015/12/24/types-machine-learning-algorithms/

# 強化學習（Reinforcement Learning）

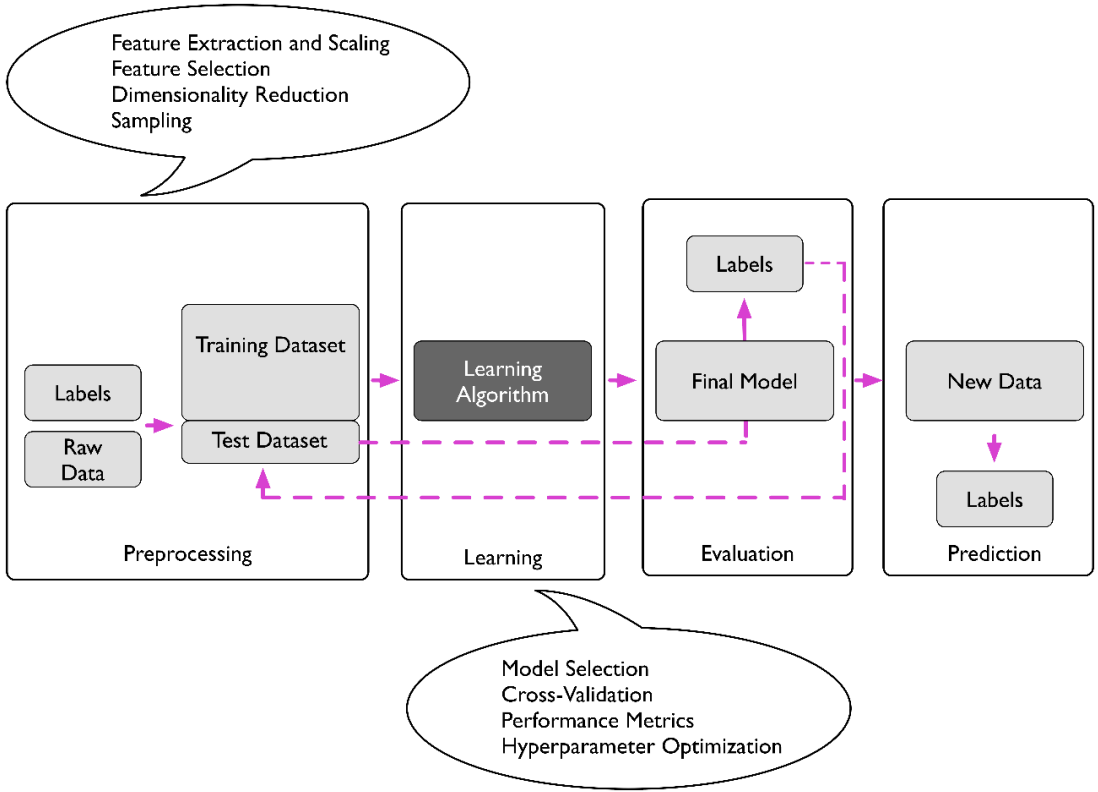


# 機器學習流程

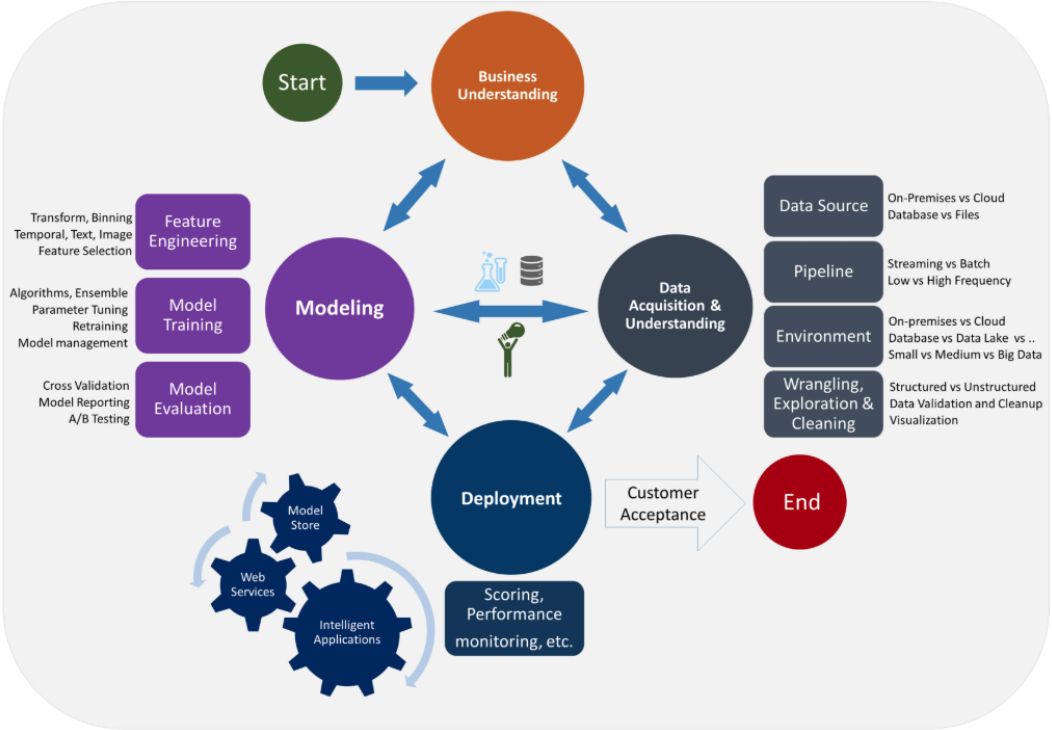



https://yourfreetemplates.com/free-machine-learning-diagram/

# 機器學習流程（另一個角度）

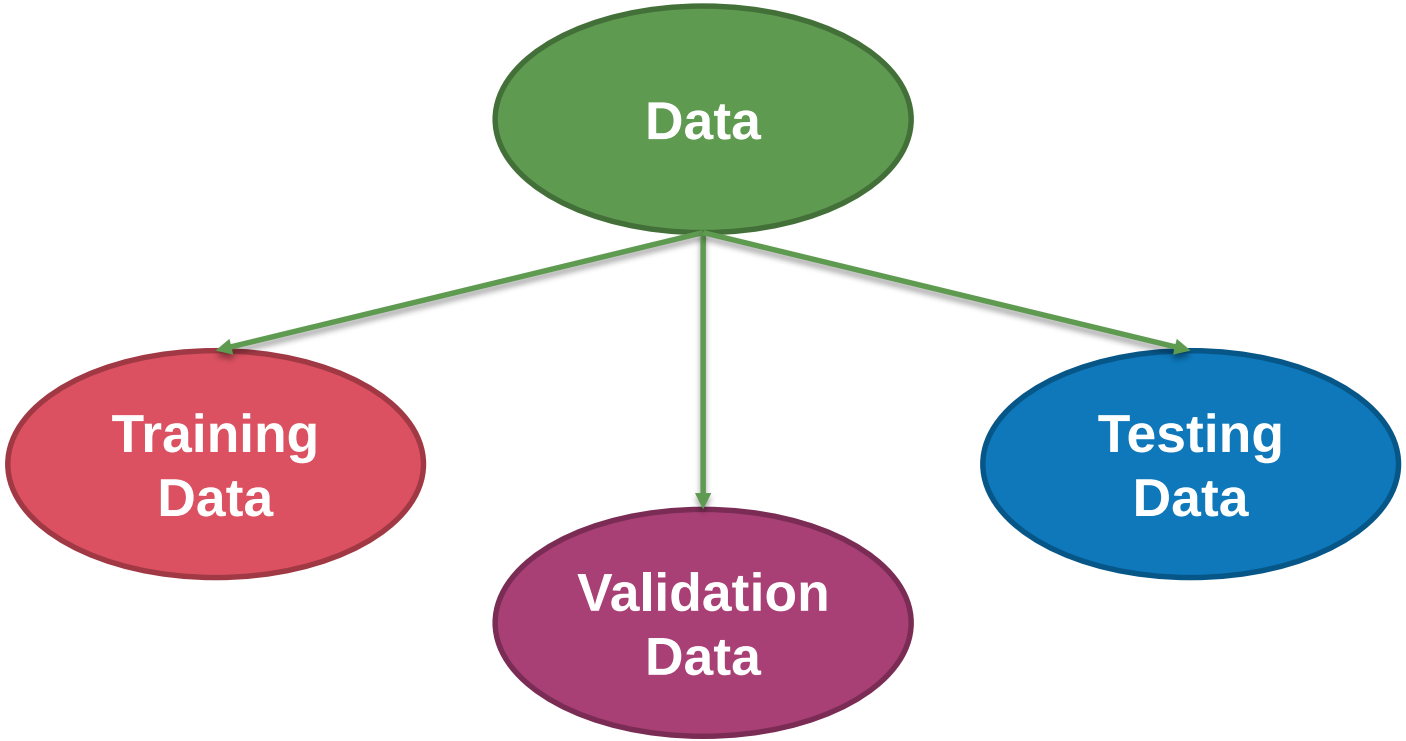


# 資料科學流程（拉遠一點看）




Team Data Science Process (TDSP)

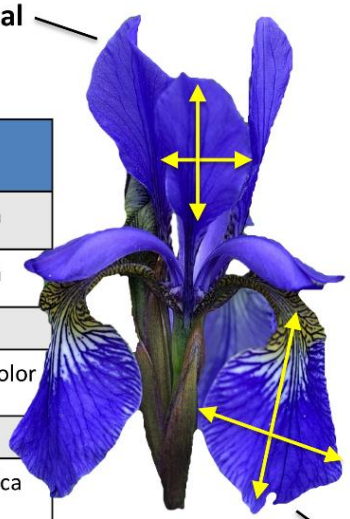
# 資料的配置（Data Split）



# 專有名詞（Terminology）

**Samples**
(instances, observations)

| | Sepal length | Sepal width | Petal length | Petal width | Class label |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | Setosa |
| ... | | | | | |
| 50 | 6.4 | 3.5 | 4.5 | 1.2 | Versicolor |
| ... | | | | | |
| 150 | 5.9 | 3.0 | 5.0 | 1.8 | Virginica |




**Features**
(attributes, measurements, dimensions)

**Class labels**
(targets)

- Samples：樣本
- Observations：觀察值
- Features：特徵
- Labels：標籤
- Classes：類別

# 個案研究 (1)
# 線性迴歸（Linear Regression）

# 線性迴歸-評估業績

- 由氣溫與營業額預估當日的業績

| 氣溫 | 29 | 28 | 34 | 31 | 25 | 29 | 32 | 31 |
|---|---|---|---|---|---|---|---|---|
| 營業額 | 7.7 | 6.2 | 9.3 | 8.4 | 5.9 | 6.4 | 8.0 | 7.5 |

| 氣溫 | 24 | 33 | 25 | 31 | 26 | 30 |
|---|---|---|---|---|---|---|
| 營業額 | 5.8 | 9.1 | 5.1 | 7.3 | 6.5 | 8.4 |

# 評估業績：請留意輸出結果

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
temperatures = np.array([29, 28, 34, 31,25, 29, 32, 31,24, 33, 25, 31,26, 30])
drink_sales = np.array([7.7, 6.2, 9.3, 8.4,5.9, 6.4, 8.0, 7.5,5.8, 9.1, 5.1, 7.3,6.5, 8.4])
X = pd.DataFrame(temperatures, columns=["Temperature"])
target = pd.DataFrame(drink_sales, columns=["Drink_Sales"])
y = target["Drink_Sales"]
lm = LinearRegression( )
```

★操作檔案
「linear01.py」

```
lm.fit(X, y)
print("迴歸係數:", lm.coef_)
print("截距:", lm.intercept_ )
```

# 評估業績：請留意輸出結果

```
import linear01
import  matplotlib.pyplot as plt
new_temperatures = linear01.pd.DataFrame(linear01.np.array([26, 30]))
predicted_sales = linear01.lm.predict(new_temperatures)
print(predicted_sales)
plt.scatter(linear01.temperatures, linear01.drink_sales)  # 繪點
regression_sales = linear01.lm.predict(linear01.X)
plt.plot(linear01.temperatures, regression_sales, color="blue")
plt.plot(new_temperatures, predicted_sales,color="red", marker="o", markersize=10)
plt.show( )
```

★操作檔案「linear02.py」

# 線性迴歸-預測體重

- 預測體重

| | | | | | |
|---|---|---|---|---|---|
| **身高** | **147.9** | **163.5** | **159.8** | **155.1** | **163.3** |
| 體重 | 41.7 | 60.2 | 47.0 | 53.2 | 48.3 |

| | | | | | |
|---|---|---|---|---|---|
| **身高** | **158.7** | **172.0** | **161.2** | **153.9** | **161.6** |
| 體重 | 55.2 | 58.5 | 49.0 | 46.7 | 52.5 |

# 複迴歸

- 線性迴歸是一個原因一個結果
- 複迴歸是多個原因一個結果

# 複迴歸-使用身高與腰圍預測體重

- 預測體重

| **腰圍** | **67** | **68** | **70** | **65** | **80** |
|---|---|---|---|---|---|
| 身高 | 160 | 165 | 167 | 170 | 165 |
| 體重 | 50 | 60 | 65 | 65 | 70 |

| **腰圍** | **85** | **78** | **79** | **95** | **89** |
|---|---|---|---|---|---|
| 身高 | 167 | 178 | 182 | 175 | 172 |
| 體重 | 75 | 80 | 85 | 90 | 81 |

# 評估體重：請留意輸出結果

★操作檔案「liner03.py」

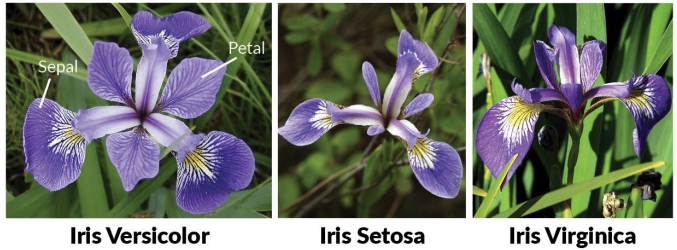
```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
waist_heights = np.array([[67,160], [68,165], [70,167],
                          [65,170], [80,165], [85,167],
                          [78,178], [79,182], [95,175],
                          [89,172]])
weights = np.array([50, 60, 65, 65,70, 75, 80, 85,90, 81])
X = pd.DataFrame(waist_heights, columns=["Waist", "Height"])
target = pd.DataFrame(weights, columns=["Weight"])
y = target["Weight"]
lm = LinearRegression( )
```

```
lm.fit(X, y)
print("迴歸係數:", lm.coef_)
print("截距:", lm.intercept_ )
new_waist_heights = pd.DataFrame(np.array([[66, 164], [82, 172]]))
predicted_weights = lm.predict(new_waist_heights)
print(predicted_weights)
```

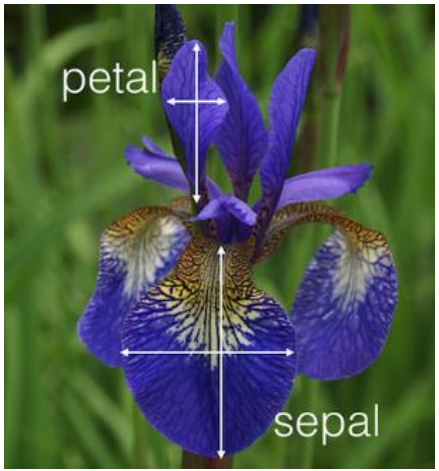
# 個案研究 (2)
# 分類（Classification）

# 鳶尾花（Iris）資料集

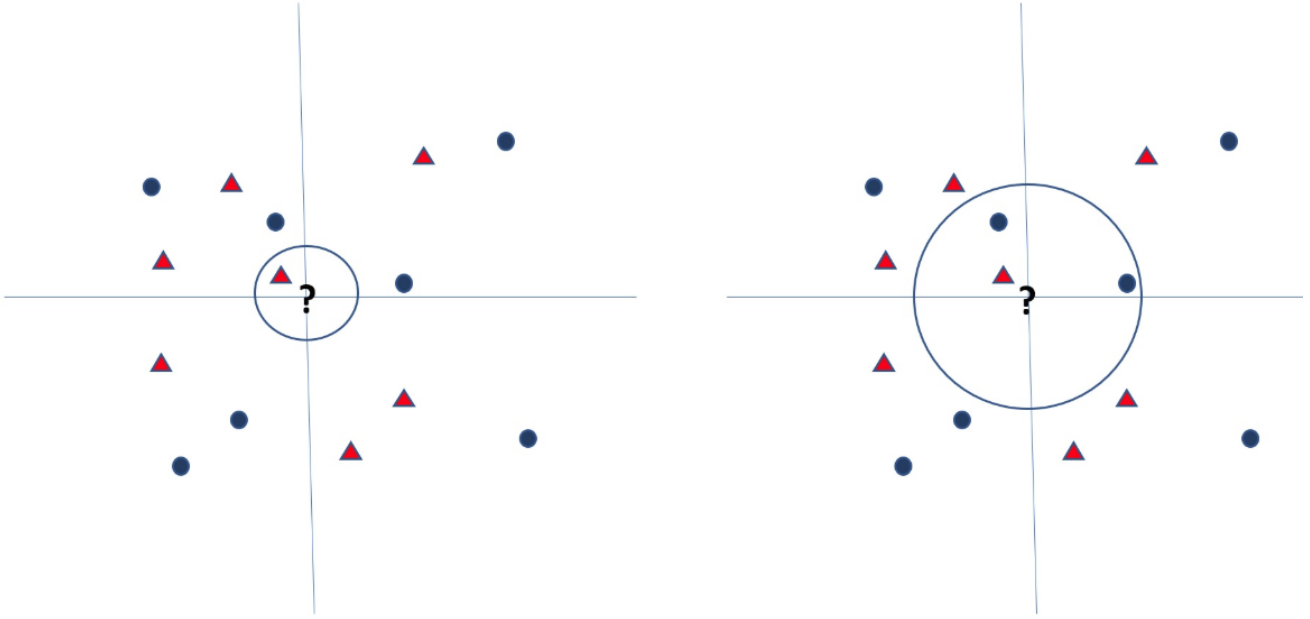
- 資料集：鳶尾花（Iris），https://archive.ics.uci.edu/ml/datasets/iris
  - 三個品種：Setosa、Versicolour、Virginica
  - 自變數
    - 花萼（sepal）長度
    - 花萼（sepal）寬度
    - 花瓣（petal）長度
    - 花瓣（petal）寬度
  - 150個樣本



Iris Versicolor　Iris Setosa　Iris Virginica



# K-Nearest-Neighbor（KNN）




K = 1, NN

K = 3, NN

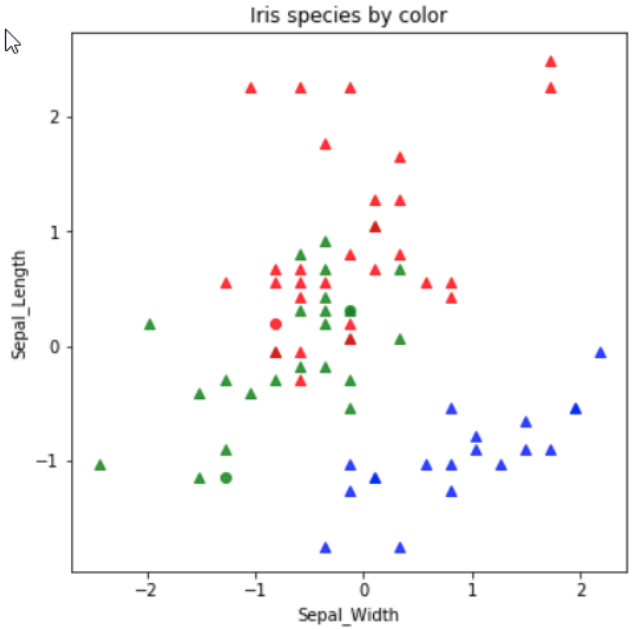
KNN for k = 1 and k = 3

# 目標

- 以 KNN 演算法預測『鳶尾花』（Iris）分類
- Lab (DAT275x)
  - Module1-275 / IntroductionToMachineLearning.ipynb

# 處理流程

- Examine the data set：載入資料集
- Prepare the data set
  - Species 轉換為(0,1,2)
  - 資料切割為訓練及測試資料
- Train and evaluate the KNN model
  - 訓練模型（Model Fitting）
  - 預測（Predict ）
  - 繪圖顯示正確（三角形）/ 錯誤（圓形）結果



- From the plot, which species are more separated than the others?

Setosa

Versicolor

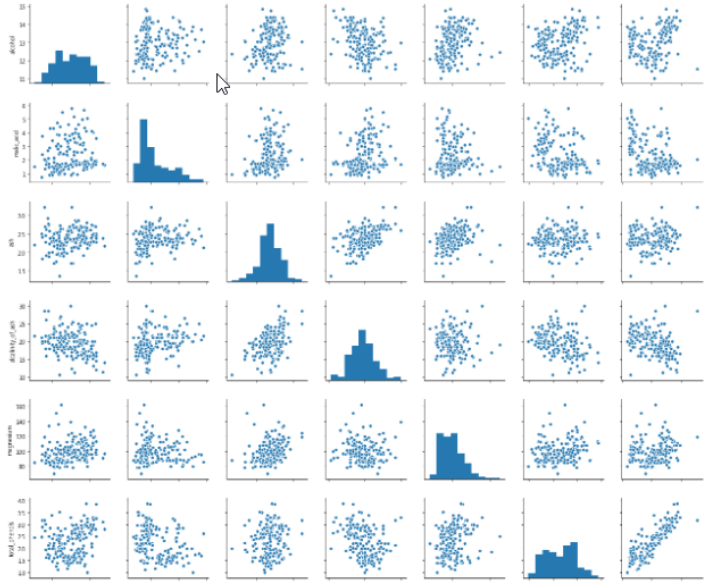
Virginica

- What is the accuracy printed?

- How many cases are mis-classified?

# 作品：酒類預測

- 試用本次課程內容分析『酒類資料集』 展示
- data = datasets.load_wine( )

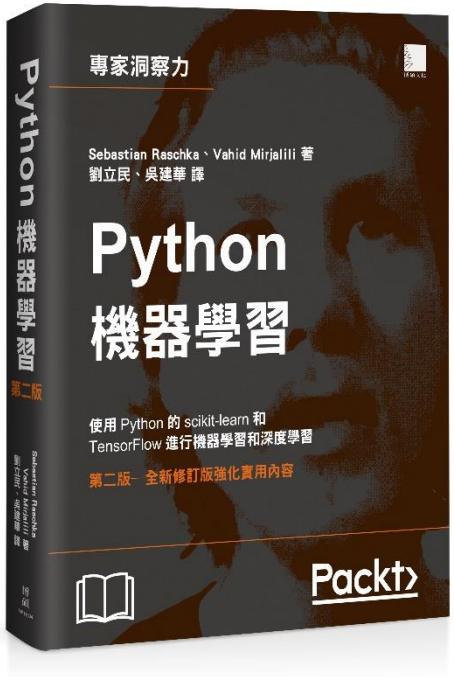


# Fall In Love

## WITH THE PROBLEM

## AND NOT WITH THE SOLUTION

# 參考用書




- 書名：Python機器學習（第二版）

  http://www.drmaster.com.tw/bookinfo.asp?BookID=MP11804
- 作者：Sebastian Raschka, Vahid Mirjalili ISBN
- 譯者：劉立民、吳建華
- 出版社：博碩

# 問卷

http://www.pcschoolonline.com.tw


www.pcschoolonline.com.tw