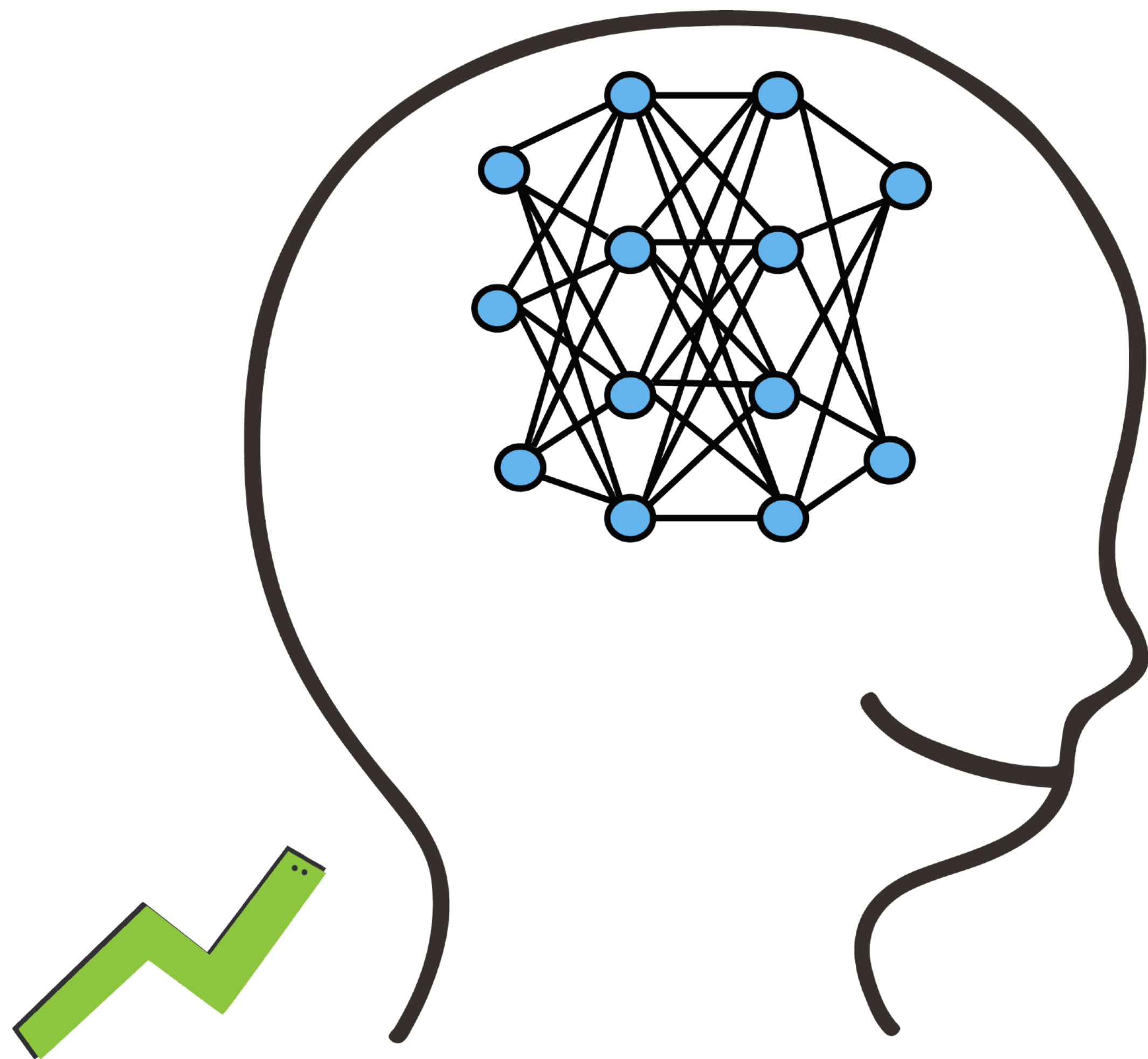


基礎機器學習預測

Prediction Using Basic Machine Learning



機器學習導論

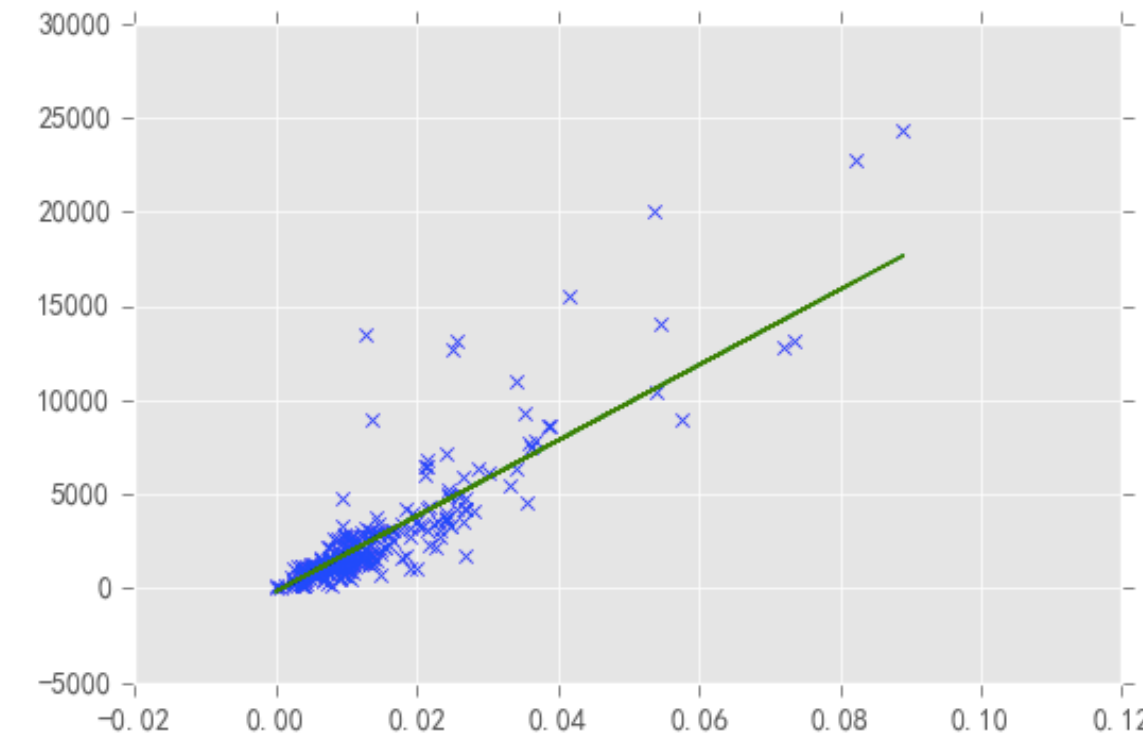
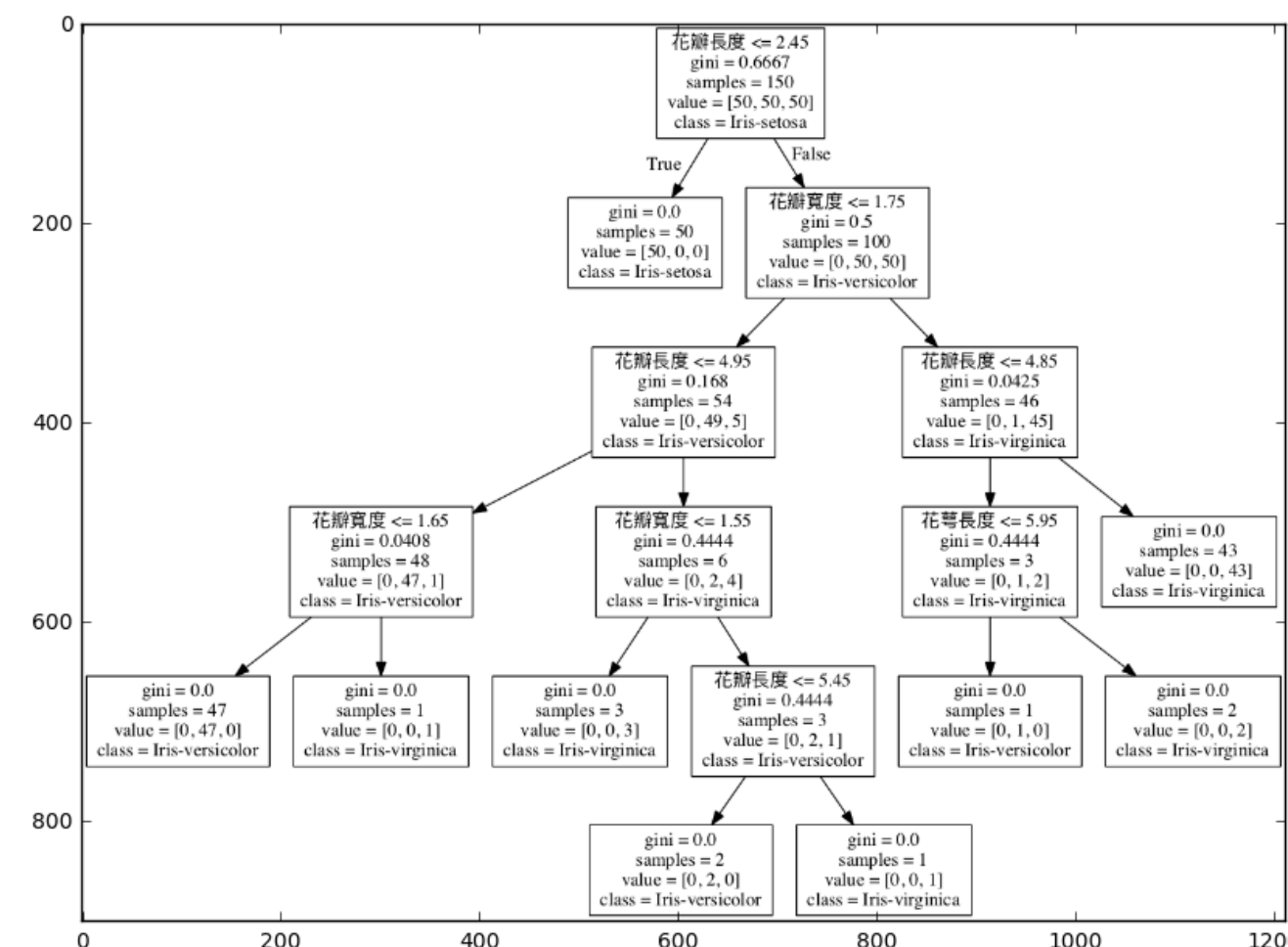
Introduction to Machine Learning

什麼是機器學習？

- 機器可以學什麼？怎麼學？
- Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).

if...
elif...
else...

Modeling with Machine Learning



Data-Driven

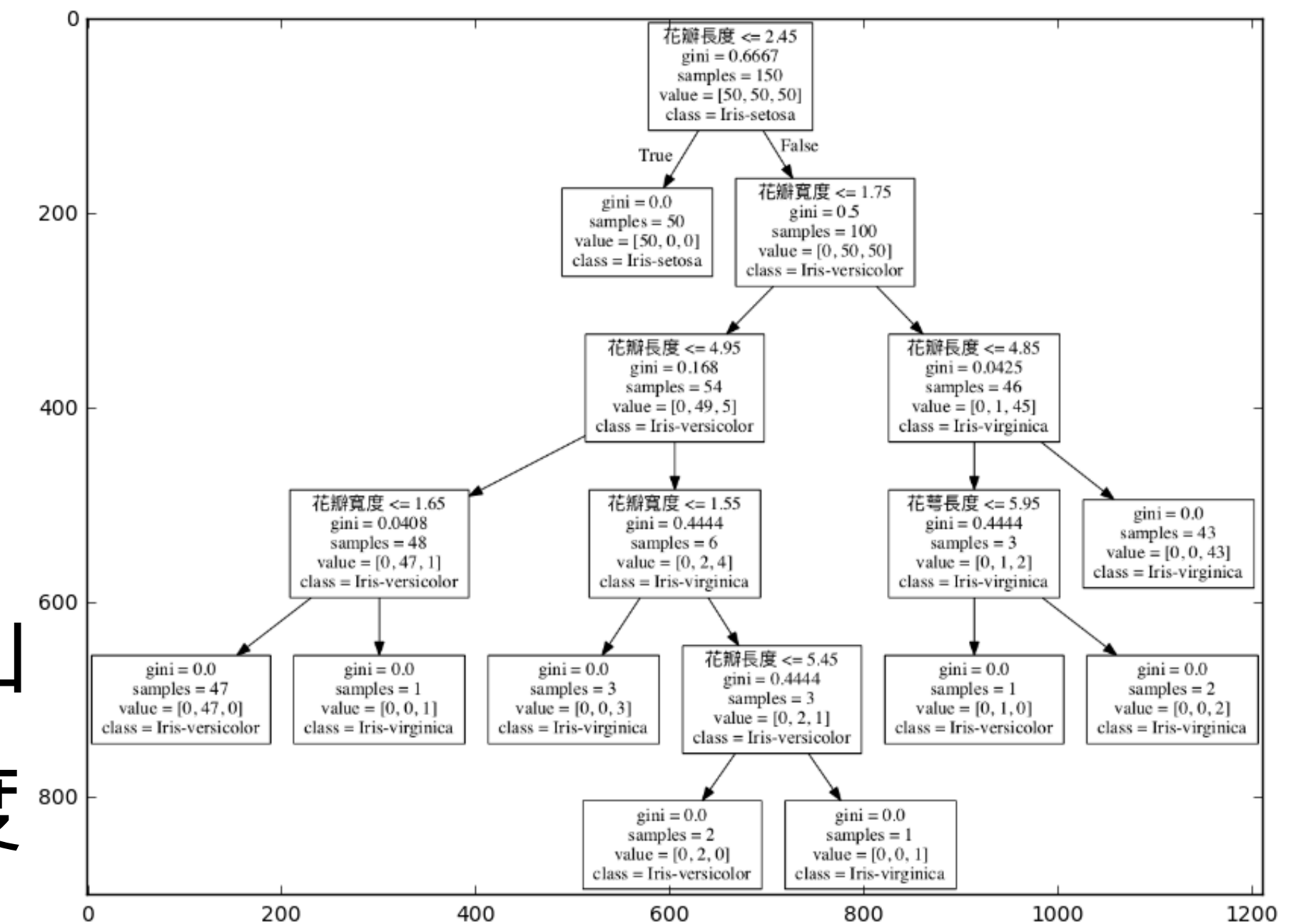


機器學習的種類

- 監督式學習 (Supervised Learning)
 - 從答案 (標籤/labels) 中學習規則 (模型/model)
- 非監督式學習 (Unsupervised Learning)
 - 沒答案 (標籤/labels) 自己找到規則
- 半監督式學習 (Semi-supervised Learning)
 - 從”部分”的答案 (標籤/labels) 中學習規則推測無答案的項目
- 增強式學習 (Reinforcement Learning)
 - 藉由外在的訊號 (獎勵/懲罰) 不斷改進自己
- 線上學習 (Online Learning)
 - 以即時的資料流訓練

監督式學習 - 分類 (Classification)

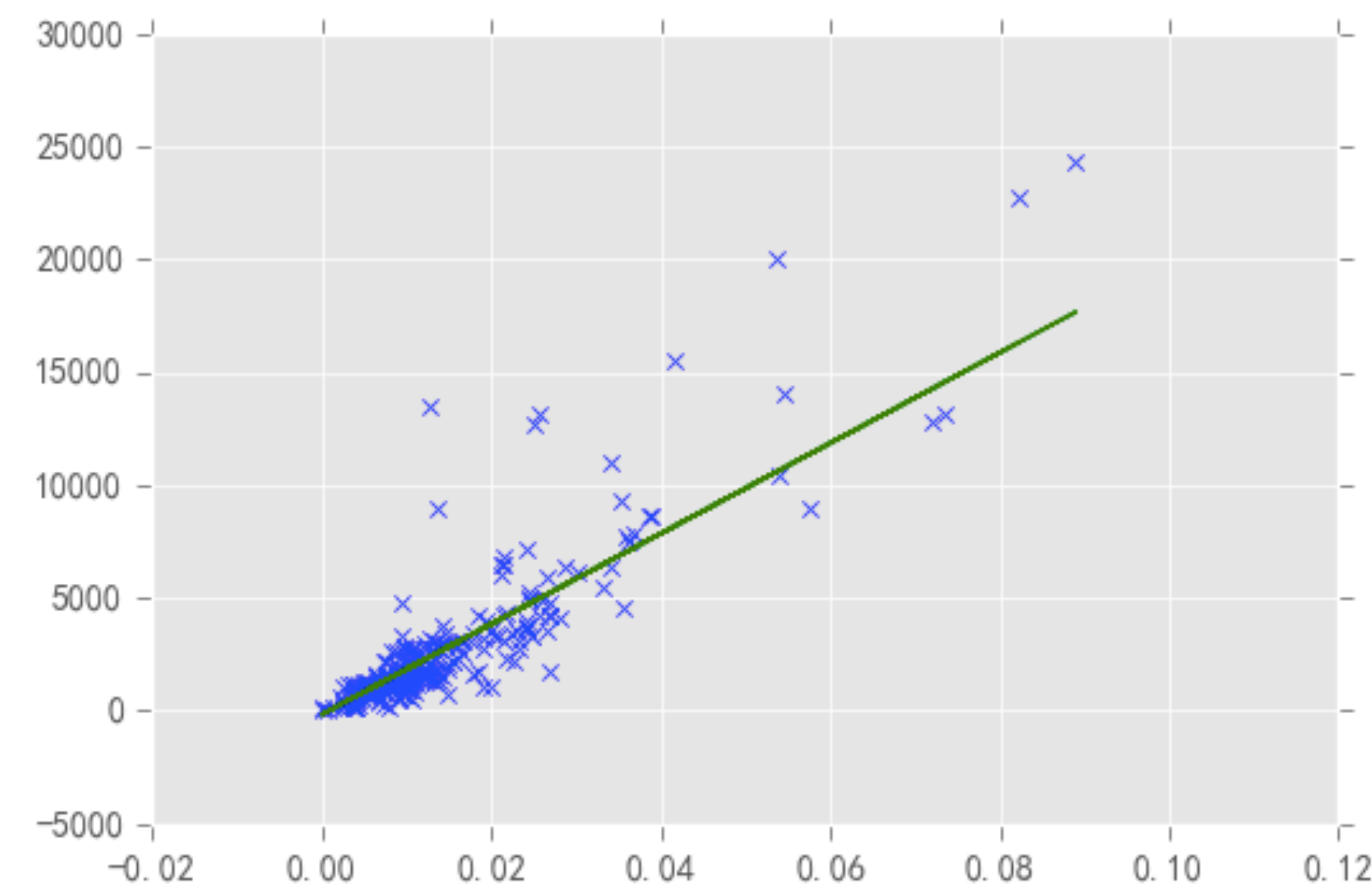
- 是否是垃圾信件？ (True/False)
- 是否罹患疾病？ (True/False)
- 生物品種分類 (e.g. 鳶尾花有很多品種，如山 (Iris Versicolor)等，可以根據花瓣和花萼寬度
- 硬幣分類 (e.g. 1元、5元、10元、50元)
- Size (e.g. 2L、XL、L、M、S、XS)





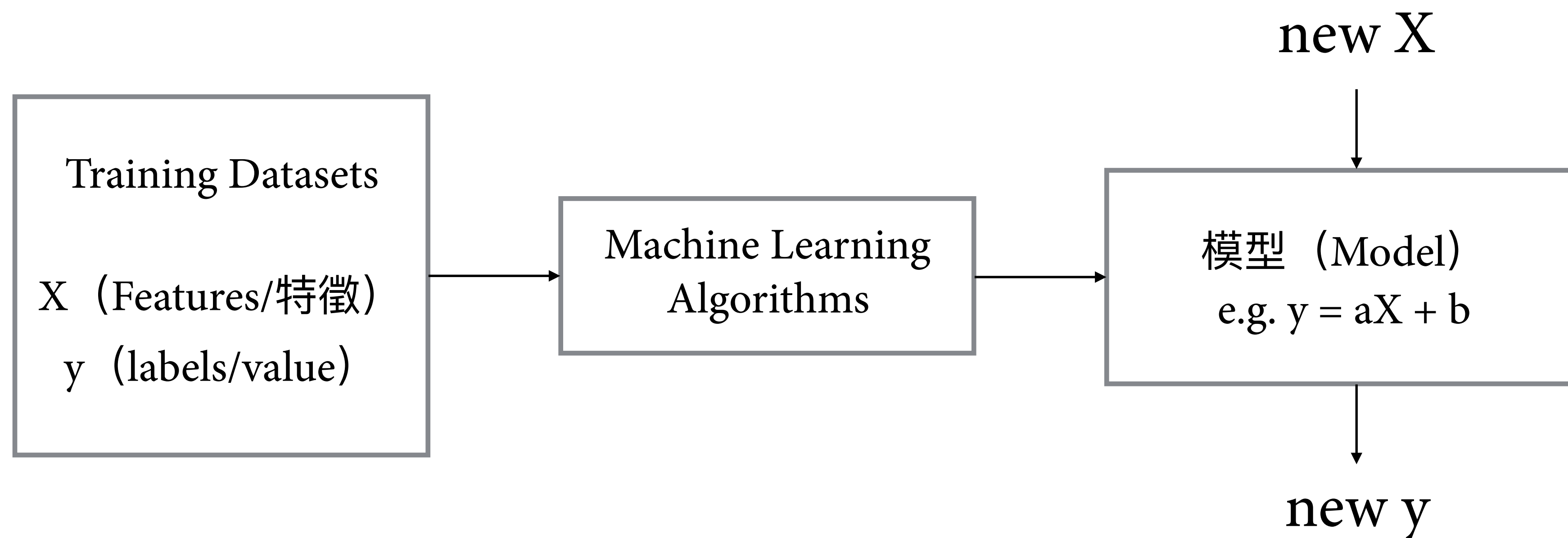
監督式學習 - 迴歸 (Regression)

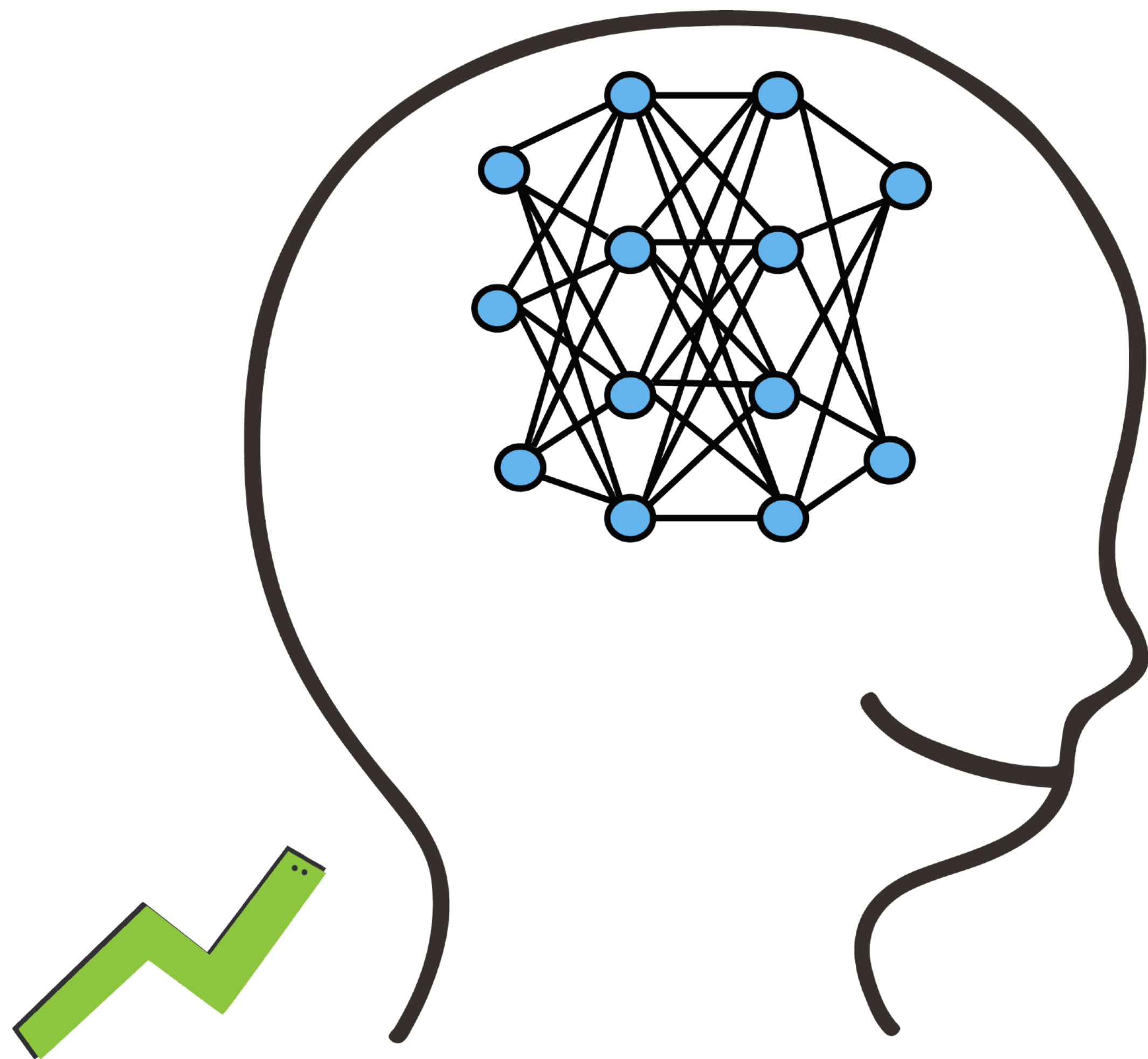
- 預測一個數值
 - 股市
 - 價格
 - 成績
 - 降雨機率
 -





怎麼學？





訓練資料預處理

Training Dataset Preprocessing



切分資料

- 資料集一般會切分為訓練資料集 (Training Dataset) 和測試資料集 (Testing Dataset)
- 切分參考比例：訓練資料集 vs. 測試資料集
 - 70% 、 30%
 - 75% 、 25%
 - 99% 、 1% (大數據)
- `sklearn.cross_validation.train_test_split()`



標準化 (Normalize)

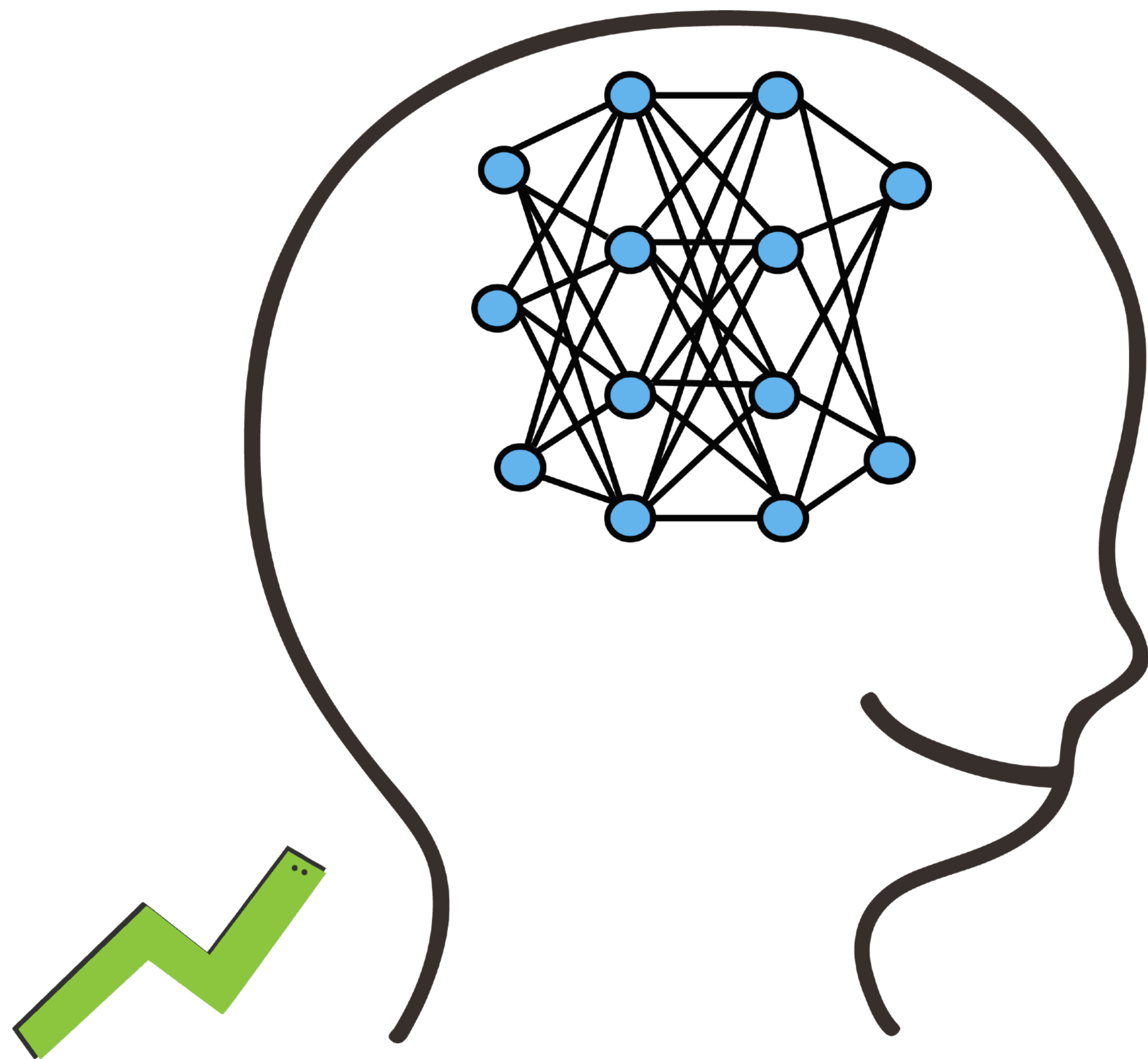
- 特徵的數值大小容易影響建模過程
 - e.g. 2 features
 - 坪數（平方公尺）：十、百、千
 - 房間數：個位數
 - 訓練時會對坪數過於敏感
- 標準化： $(\text{原值} - \text{平均}) / \text{標準差}$
 - `sklearn.preprocessing.StandardScaler()`



編碼

- 順序量尺：定義一個轉換成數值的方式
 - e.g. S -> 1, M -> 2, L -> 3
- 名義量尺：不可以直接對應成數字，因為沒有大小順序。
- One-hot Encoding

	顏色		紅色	藍色	綠色
0	紅色	0	1	0	0
1	藍色	1	0	1	0
2	綠色	2	0	0	1



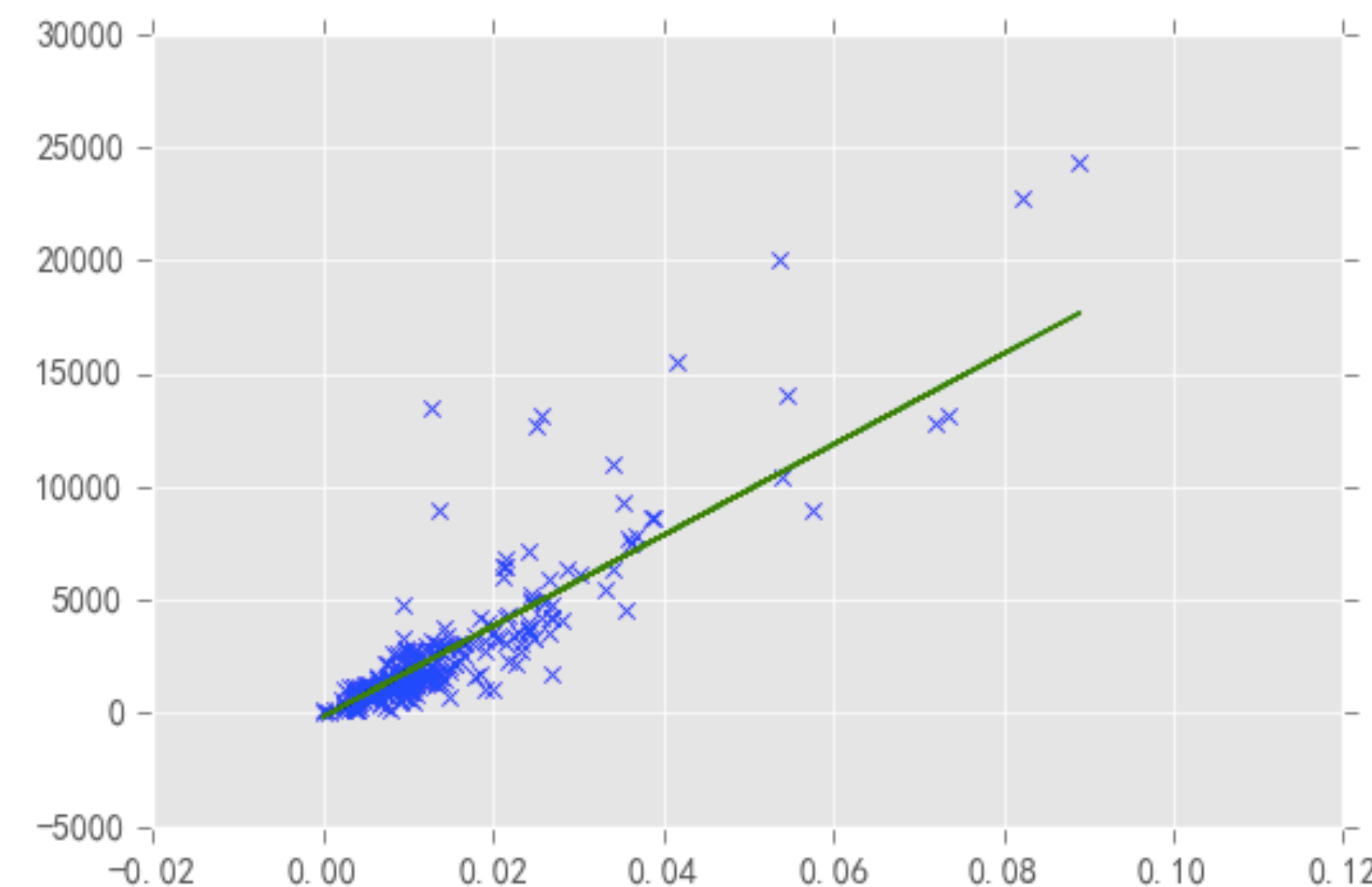
基礎監督式機器學習

Basic Supervised Machine Learning



線性迴歸 (Linear Regression)

- 線性迴歸
 - 簡單線性迴歸
 - e.g. $y = ax + b$
 - 多變項線性迴歸
 - e.g. $y = ax_1 + bx_2 + c$
- 非線性迴歸：多項式 (Polynomial)
 - e.g. $y = ax + bx^2 + cx^3$
- `sklearn.linear_model.LinearRegression()`





線性迴歸 (Linear Regression)

- 解參數值方法

- 正規方程 (Normal Equation)

- $(X^T X)^{-1} X^T y$

- 梯度下降 (Gradient descent)

- Hypothesis (假設) : $h_{\theta}(x) = \theta_0 + \theta_1 x$

- Parameters : θ_0 、 θ_1 (所求)

- Cost Function : $J(\theta_0, \theta_1) =$

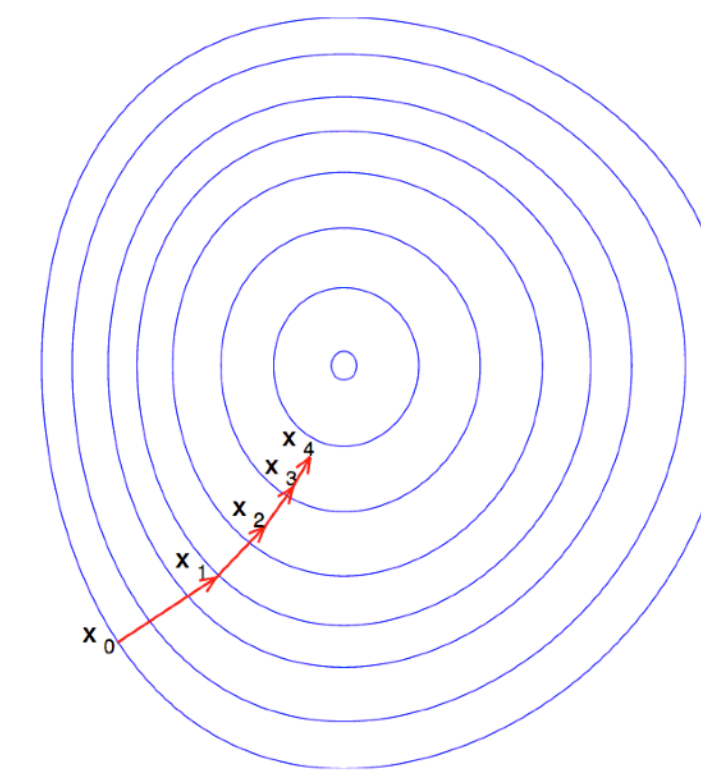
- 目標：最小化Cost Function $\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

- repeat : $\theta = \theta - \alpha * (\text{derivative/導函數 of cost function})$ 直到到達local optimum (區域最佳解)

- α : 學習速率

Notes

► Scikit-Learn解線性迴歸是使用最佳化過的正規方程



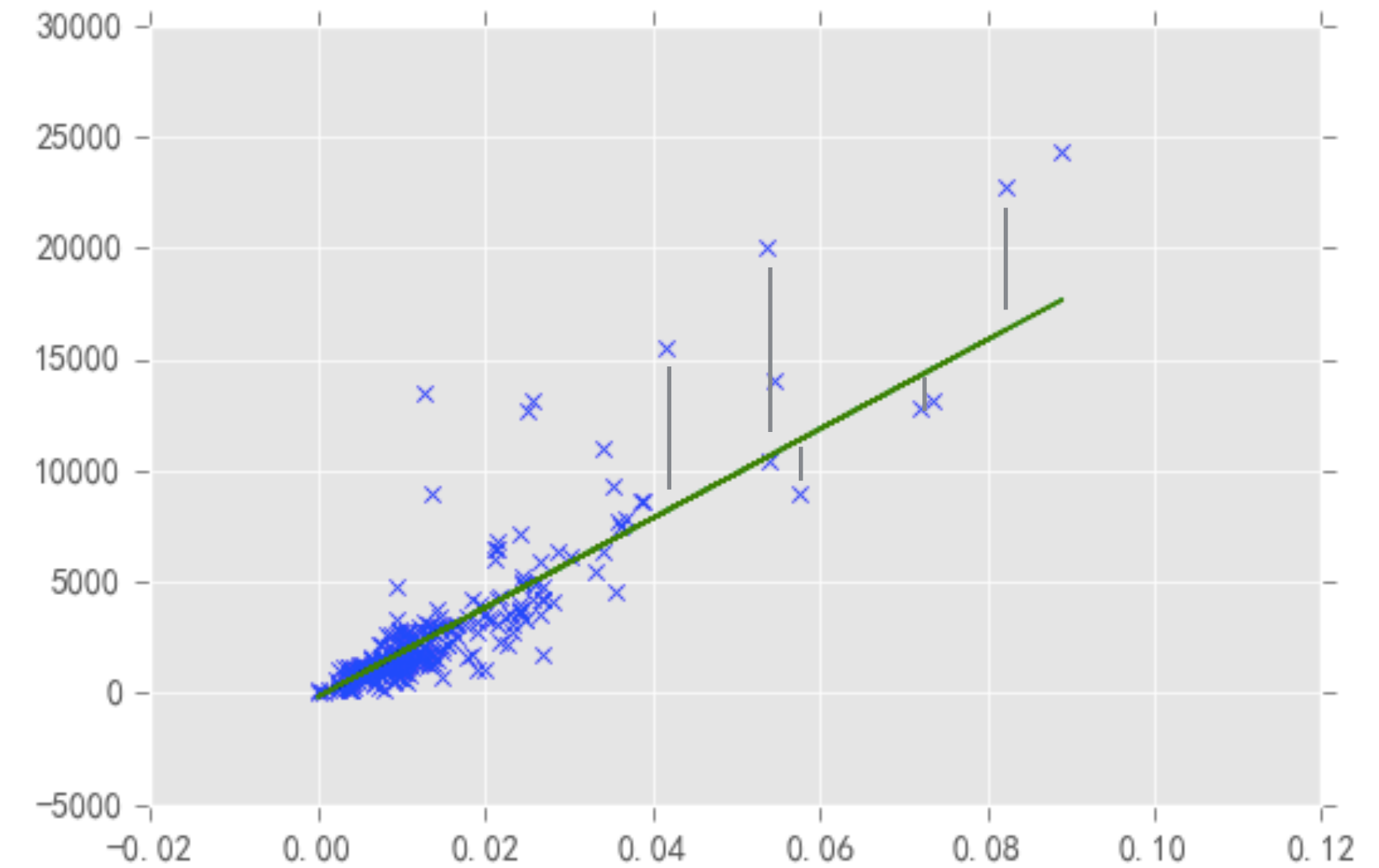
(from wiki)

線性迴歸效果評估

- 均方誤差 (Mean Square Error, MSE)

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

↑
↑
 預測y 實際y



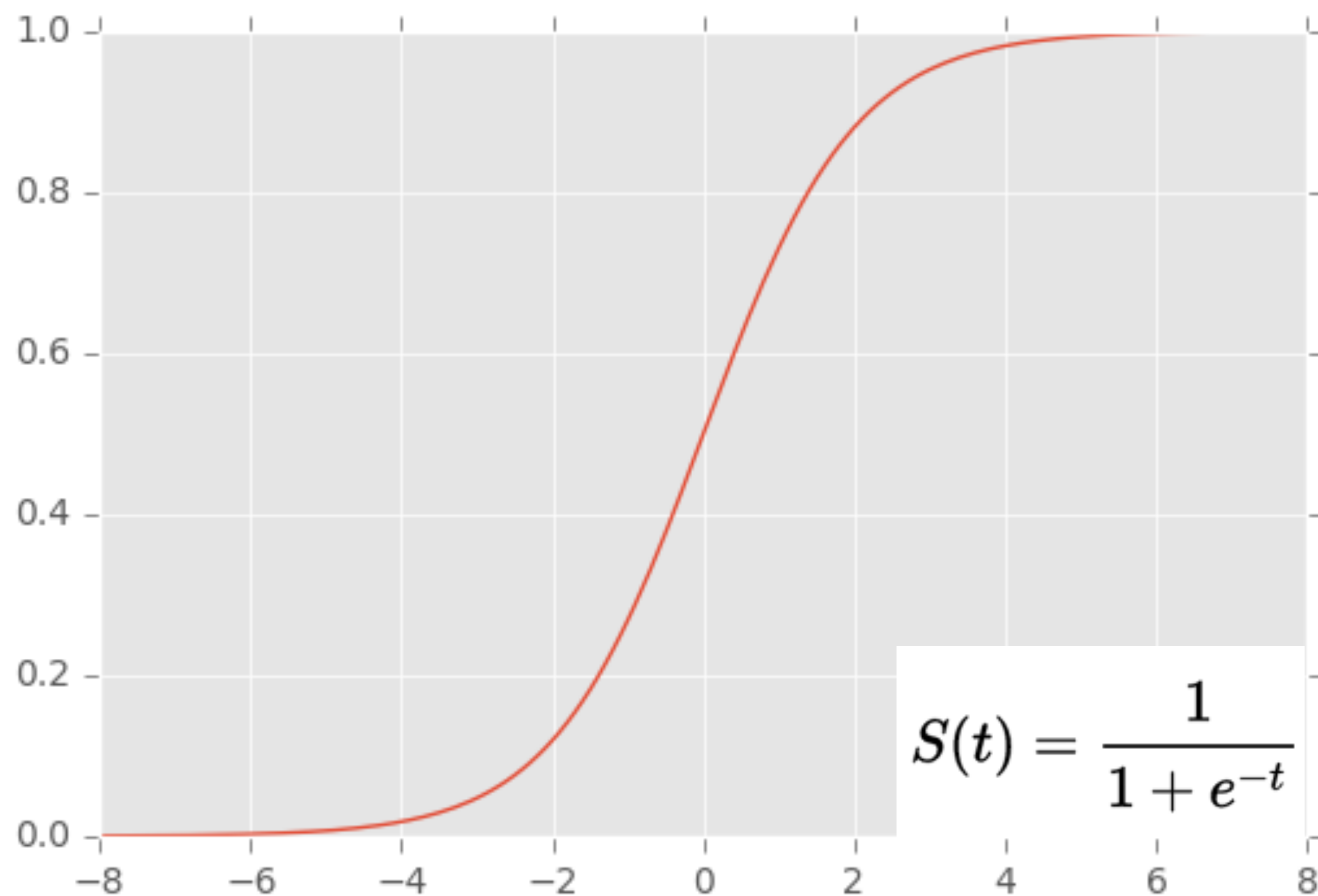
Notes

- ▶ 其他評估方法：(詳見補充講義)
 - ▶ SSE (Sum of Squared Errors)
 - ▶ R^2 (R Square) : 0最差、1最好
 - ▶ ...



羅吉斯迴歸 (Logistic Regression)

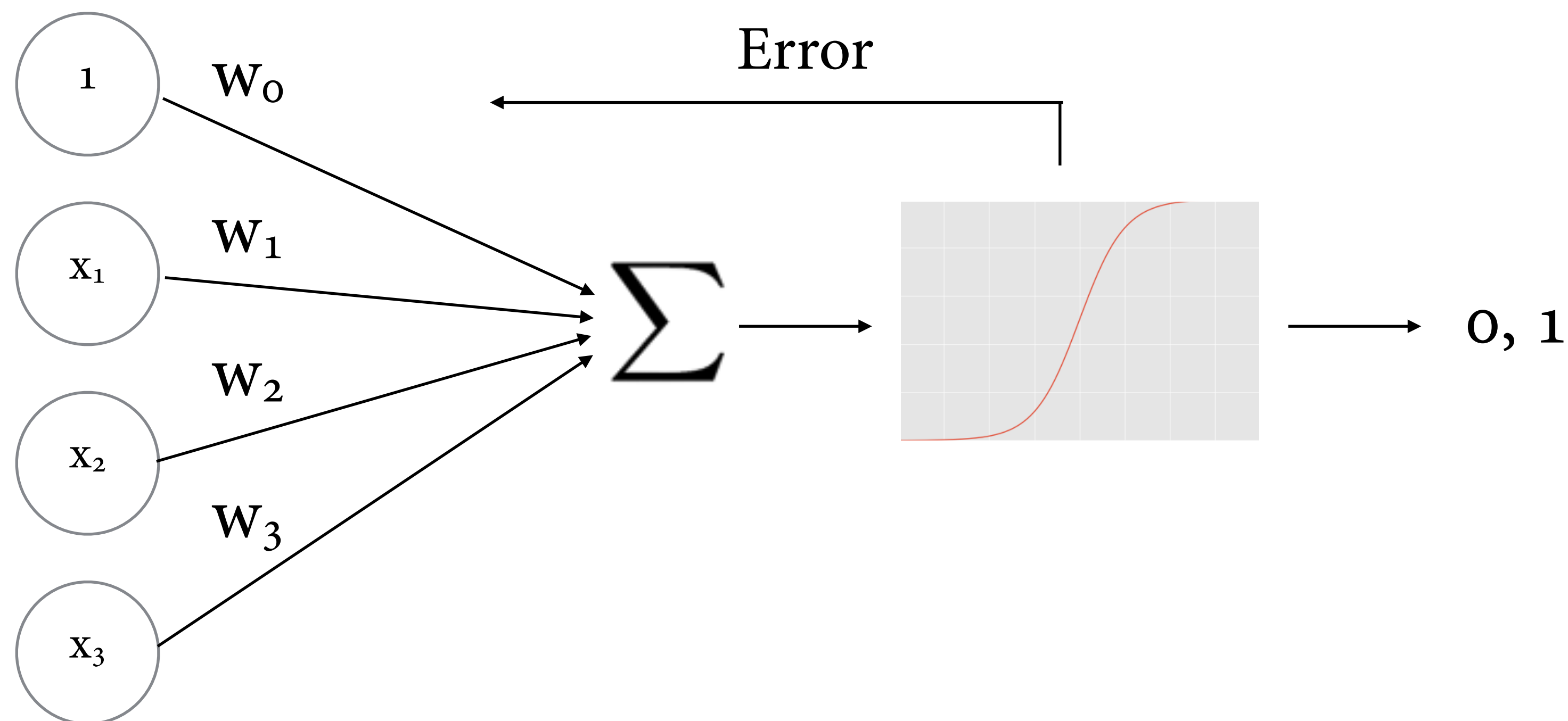
- Logistic function / Sigmoid function





羅吉斯迴歸 (Logistic Regression)

- 雖然名為迴歸，但常用於分類（二元或多類別）



羅吉斯迴歸 (Logistic Regression)

- 多類別分類，使用One-vs-Rest (OvR)
 - e.g. A, B, C三類，分別計算是A的機率、是B的機率、是C的機率

```
array([[ 0.009,  0.401,  0.59 ],  
       [ 0.008,  0.436,  0.555],  
       [ 0.009,  0.585,  0.406],  
       [ 0.76  ,  0.137,  0.103],  
       [ 0.007,  0.505,  0.488],  
       [ 0.    ,  0.399,  0.601],  
       [ 0.018,  0.496,  0.487],  
       [ 0.004,  0.419,  0.577],  
       [ 0.864,  0.088,  0.048],
```



分類效果評估

- 混淆矩陣 (Confusion Matrix)

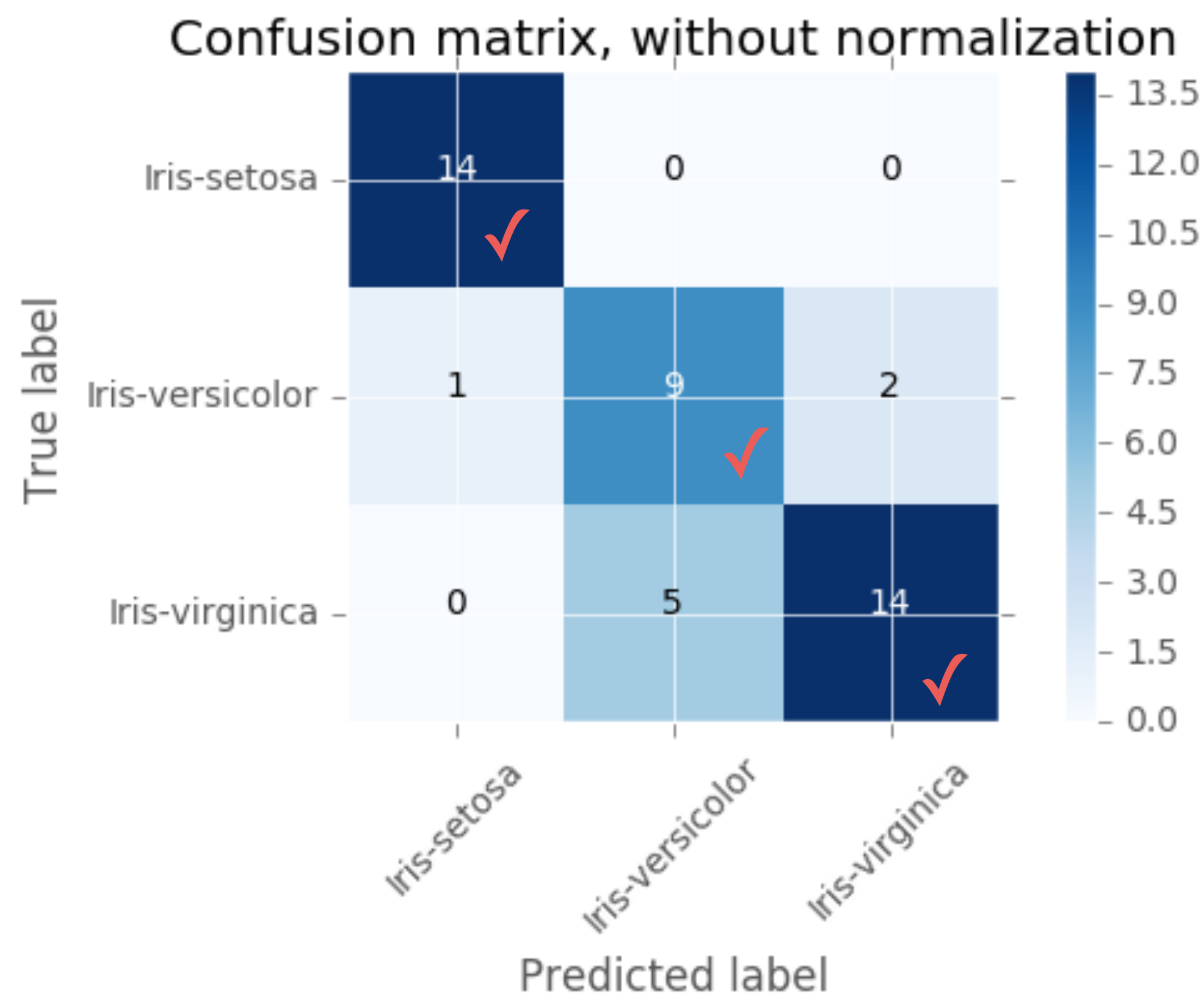
	Predicted 0	Predicted 1
True 0	True negatives (TN)	False positives (FP)
True 1	False negatives (FN)	True positives (TP)

- 正確率(Accuracy): $A = (TN + TP) / (TN + FN + FP + TP)$
- 精確率(Precision): $P = TP / (TP + FP)$
- 召回率(Recall): $R = TP / (TP + FN)$
- F1 score = $2PR / (P+R)$ (P, R的調和平均)



分類效果評估

- 混淆矩陣 (Confusion Matrix) - 多類別



	precision	recall	f1-score
Iris-setosa	0.93	1.00	0.97
Iris-versicolor	0.64	0.75	0.69
Iris-virginica	0.88	0.74	0.80

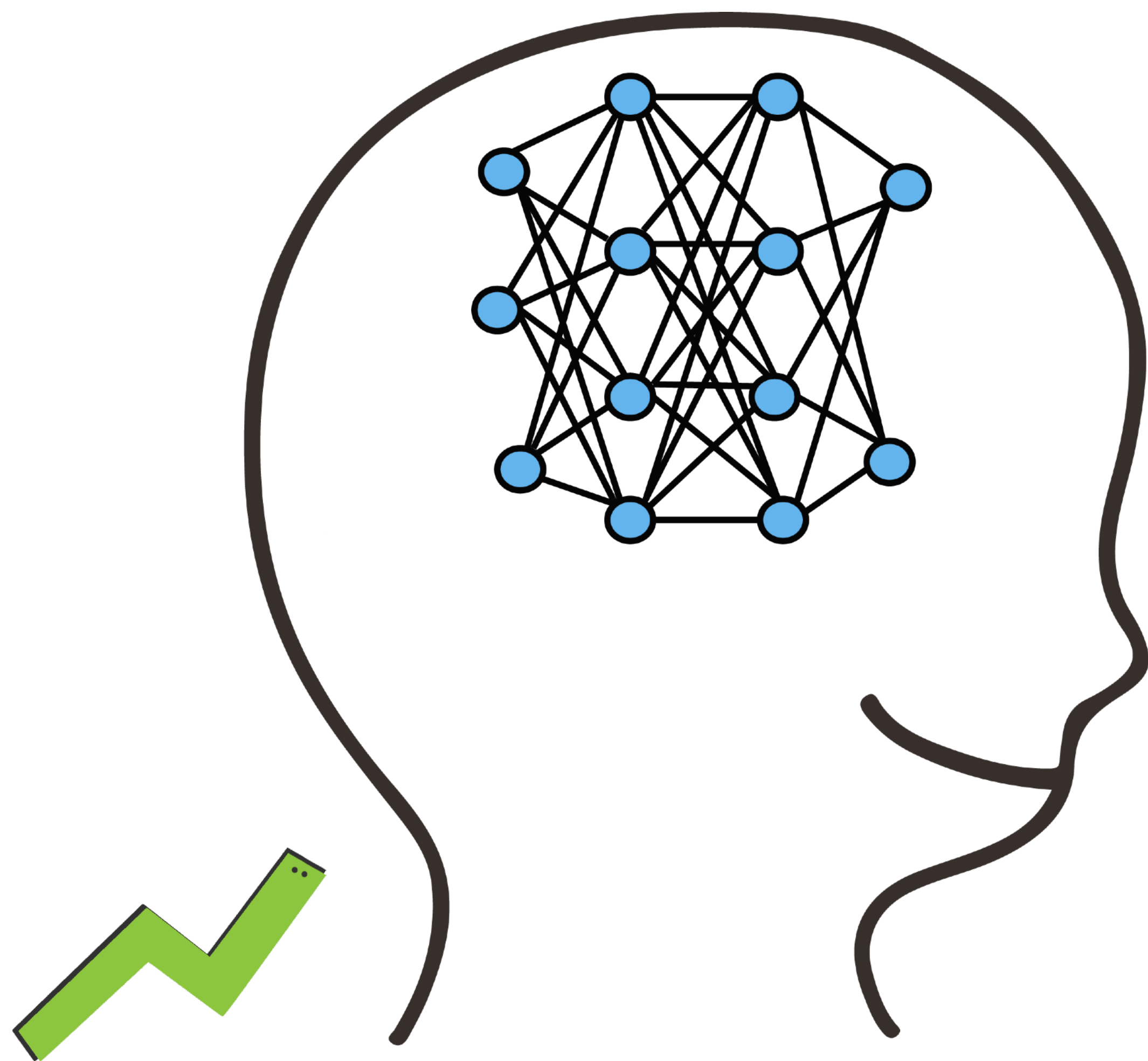
e.g. Iris-versicolor (變色鳶尾花)

- ▶ Precision = $9/(9+5) = 0.64$
- ▶ 預測14個變色鳶尾花，9個命中
- ▶ Recall = $9/(1+9+2) = 0.75$
- ▶ 有12個變色鳶尾花，找回了9個



其他常用監督式學習演算法

- 支持向量機 (Support Vector Machine, SVM)
- 決策樹 (Decision Tree)
- 人工神經網路 (Artificial Neural Network, ANN)
- K最近鄰 (k-NN)
- 樸素貝式分類器 (Naive Bayes classifiers)
- ...



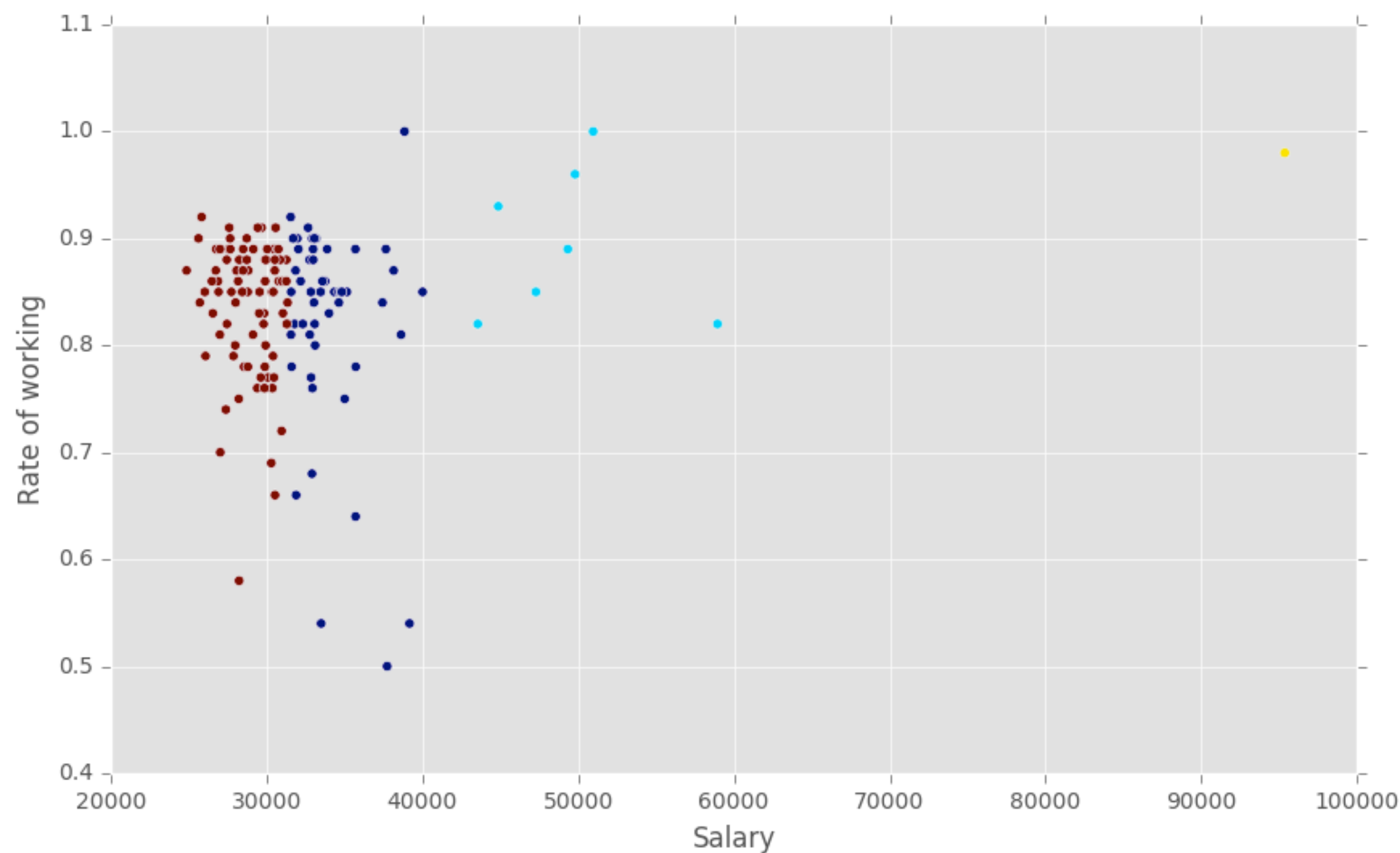
基礎非監督式機器學習

Basic Unsupervised Machine Learning



分群 (Clustering)

- 將靠近的資料分成不同群體





K-means

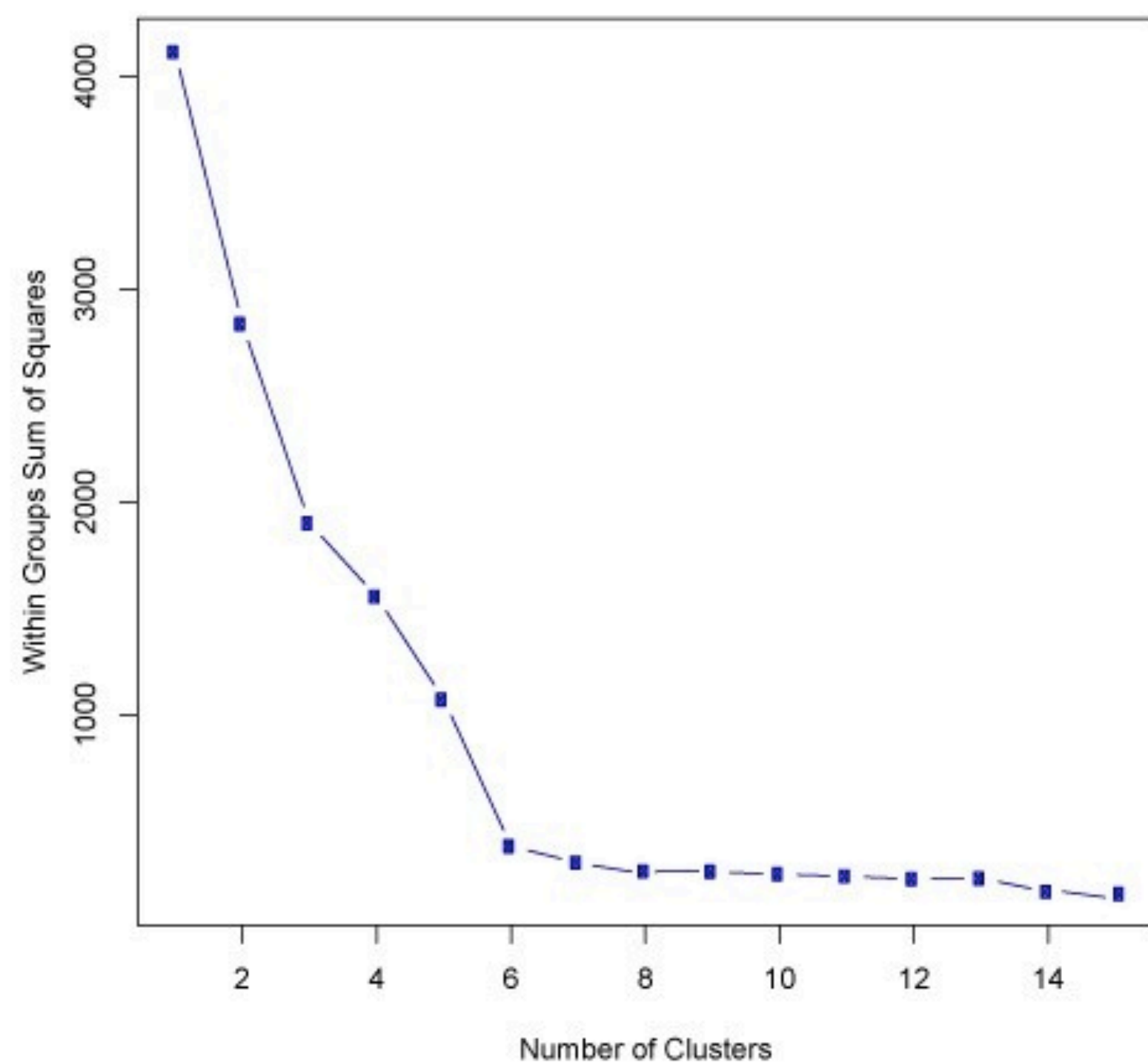
- 設 $K = 3$ ，起始隨機挑選3個點作為集群中心點
- repeat：將附近的點根據與這3中心點的距離分配到這三群，並重新計算中心點，直到收斂為止
- 收斂：得到與各集群中心點距離和最小值

$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



如何選好的k值？

- 指定k：e.g. Size 分為L、M、S
- 不指定k：使用不同的k，計算點和中心的距離總和



(Source from: <https://www.quora.com/How-can-we-choose-a-good-K-for-K-means-clustering>)



機器學習線上學習資源

- 台灣大學 林軒田教授
 - 機器學習基石：<https://www.youtube.com/playlist?list=PLXVfgk9fNX2I7tB6oIINGBmW5orrmFTqf>
 - 機器學習技法：<https://www.youtube.com/playlist?list=PLXVfgk9fNX2IQOYPmqjqWsNUFl2kpk1U2>
- Stanford Andrew Ng
 - Machine Learning：<https://zh-tw.coursera.org/learn/machine-learning>
- ShareCourse 系列進階課程「Python 機器學習與深度學習實作」