

# AI基礎程式設計

## 大數據分析

- R/Python/Julia/SQL 程式設計與應用  
(R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)



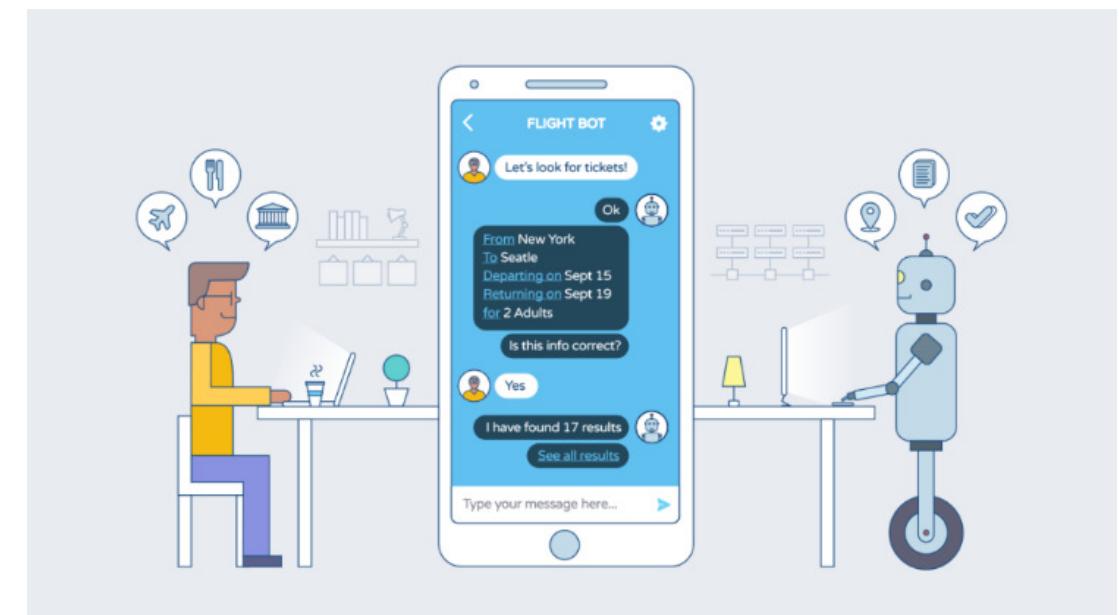
李明昌博士

alan9956@gmail.com

<http://rwepa.blogspot.com/>

# 大綱

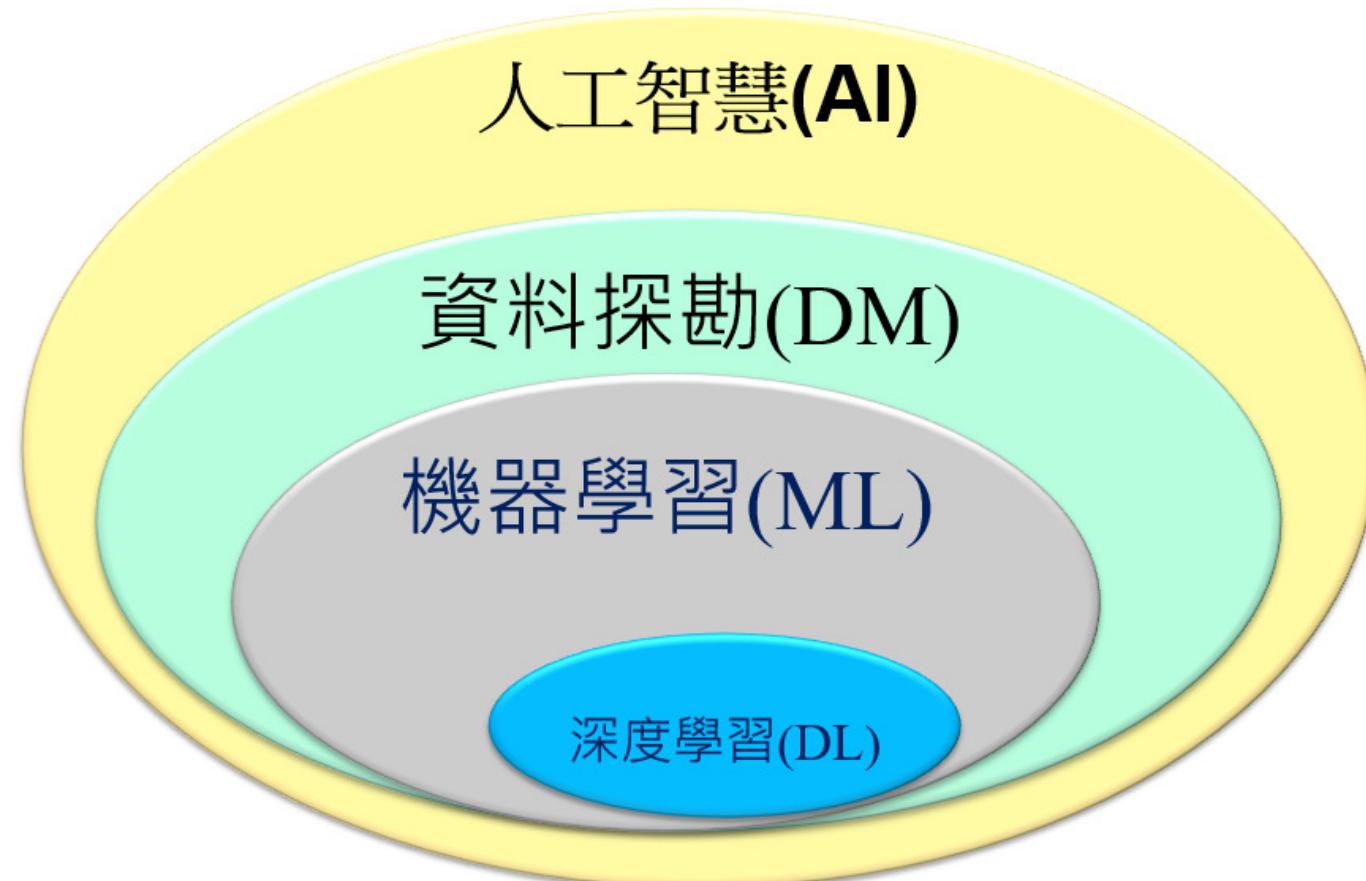
- 1. 人工智慧簡介
- 2. 人工智慧技術
  - 2.1 機器學習
  - 2.2 深度學習
  - 2.3 影像辨識
  - 2.4 聯天機器人
  - 2.5 感測器
  - 2.6 大數據分析
- 3. 程式設計(大數據分析工具)



# 1.人工智慧簡介

---

# 人工智慧簡介



# 人工智慧發展史



- 1943年：美國數學家 **Walter Pitts** 和心理學家 **Warren McCulloch** 提出人工神經元。
- 1957年：美國心理學家 **Frank Rosenblatt** 提出了感知器(Perceptron)。
- 1980年：多層類神經網路失敗，淺層機器學習方法(**SVM**等)興起。
- 2006年：**Geoffrey Hinton** 成功訓練多層神經網路(限制玻爾茲曼機, RBM)，命名為**深度學習**。
- 2012年：**ImageNet** 比賽讓深度學習重回學界視野，開啟 **NVIDIA GPU** 為重要運算硬體。

## 2. 人工智慧技術

---

# 人工智慧技術

- 人工智慧 (artificial intelligence, AI) 亦稱人工智能、機器智慧，指由人類製造出來的機器所表現出來的智慧。
- 通常人工智慧是指透過普通電腦程式來呈現人類智慧的技術。該詞也指出研究這樣的智慧系統是否能夠實現，以及如何實現。同時，透過醫學、神經科學、機器人學及統計學等的進步，常態預測則認為人類的很多職業也逐漸被其取代。

參考: [https://en.wikipedia.org/wiki/Artificial\\_intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence)

# AI 第一次潮流 (1/3)

- 第一次潮流在1956-1987，「人工智慧」名詞第一次出現在1956年的美國達特矛斯（Dartmouth）會議上。當時由約翰·麥卡錫(John McCarthy)等十位學者出席會議並且秉持著「人類般思考的機器稱為人工智慧」的信念，針對初期的人工智慧程式和各種相關的理論進行討論。
- 成果：
  - 感知機 (深度學習的雛形) 被提出
  - 用機器證明「數學原理」定義，也提出歸結原理
  - 發展出模型識別程序，並編製可分辨積木構造的程序
  - 編制通用的問題解決程序 (General Problem Solver，GPS)
  - 研製成功專家系統 DENDRAL
  - 研發人工智慧語言 (List Processing，LISP) ←

# 約翰麥卡錫 (John McCarthy) 1967

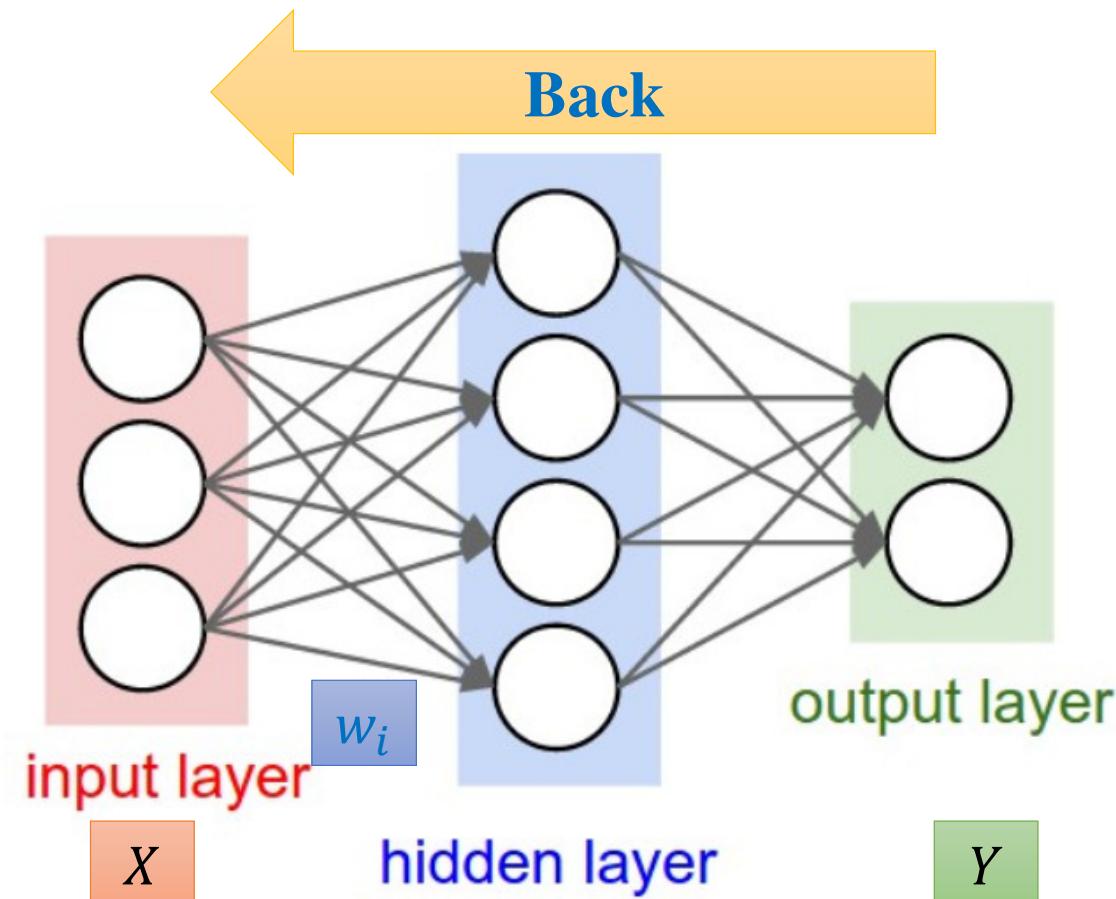
Professor John McCarthy - Stanford University ~1967



# AI 第二次潮流 (2/3)

- 第二次潮流在1987-1993，以規則庫建立的專家系統。
- 成果：
  - 專家系統的誕生
  - 發現「智慧」需要建立在分類的知識多種處理方式
  - 反向傳播演算法（Backpropagation，BP演算法）實現神經網路訓練
  - 研究人員首次提出：機器為了獲得真正的智慧，它必要有感知、生存、與世界**交互的能力**，對事物的推理能力比抽象能力更重要，這也推動未來自然語言、機器視覺的發展。

# Backpropagation , BP演算法



# AI 第三次潮流 (3/3)

- 第三次潮流在1993-2022，深度學習概念和急遽發展的時代。
- 成果：
  - 1997年，IBM的計算機系統Deep Blue戰勝國際象棋世界冠軍Garry Kasparov，這也成為重要的里程碑。
  - 2005年，Stanford開發的機器人在沙漠上自動行駛210公里，贏得DARPA挑戰大賽頭獎。
  - 2006年，Geoffrey Hinton提出多層神經網絡的深度學習算法、Eric Schmidt在搜索引擎大會提出「雲端計算」概念。
  - 2014年，微軟亞洲研究院發佈人工智慧聊天機器人和語音助手Cortana。
  - 2016年，Google公司的人工智慧程式「AlphaGO」，與韓國棋手李世乭在圍棋上正面交鋒。AlphaGo 所使用的深度學習技術引起全球關注。
  - 2017年，AlphaGo團隊提出AlphaGo Zero 以自我學習方式，在中國烏鎮圍棋峰會，挑戰排名世界第一的圍棋冠軍柯潔，以3比0獲勝。

# AlphaGo Zero



參考: <https://qbi.uq.edu.au/blog/2017/10/google-alphago-zero-masters-game-three-days>

## 2.1 機器學習

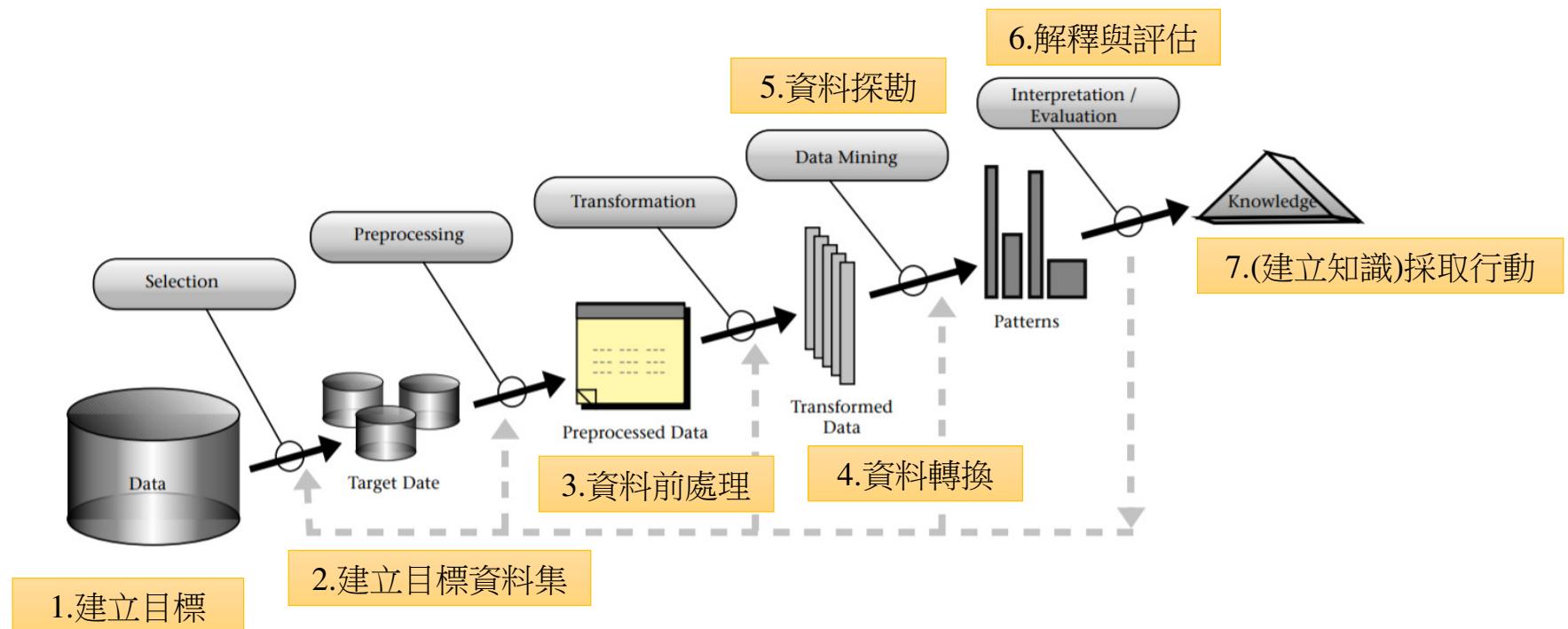
---

KDD → 機器學習

# 資料庫中的知識發掘

## (Knowledge Discovery in Database, KDD, 1996)

- Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth , *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, August 1996 Pages 82 - 88. <https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>



# 機器學習 Machine learning

- 非監督式學習 (Unsupervised learning)
  - No label or target value given for the data
- 監督式學習 (Supervised learning)
  - Telling the algorithm what to predict
- 半監督學習 (Semi-supervised learning)
  - 具有少量標記資料
- 強化學習 (Reinforcement learning)
  - 為了達成目標，隨著環境的變動，而逐步調整其行為，並評估每一個行動之後所到的回饋是正向的或負向的。
- 深度學習 (Deep learning)



# 監督式學習 vs. 非監督式學習

- 非監督式學習 Unsupervised learning
  - 集群法 Clustering
  - 關聯規則 Association rule
  - 主成分分析 Principal Component Analysis
- 監督式學習 Supervised learning (執行 X --> 預測 --> Y ): 分類與數值預測
  - 迴歸分析 Regression analysis
  - 廣義線性模型 General linear model (GLM)
  - 天真貝氏法 Naïve-Bayes
  - K近鄰法 k-nearest neighbors (KNN)
  - 決策樹 Decision tree
  - 支持向量機 Support vector machine (SVM)
  - 類神經網路 Neural network (NN)
  - 集成學習 Ensemble learning: 使用多種學習算法來獲得比單獨使用演算法更好預測結果

# 機器學習方法論

## CRISP-DM 簡介

# 資料探勘生命週期 – CRISP-DM

- 跨產業資料探勘標準作業流程  
(CRoss Industry Standard Process for Data Mining)
- CRISP-DM是於1990年起，由SPSS以及NCR兩大廠商在合作戴姆克萊斯勒-賓士(Daimler Benz)的資料倉儲以及資料探勘過程中發展出來的。

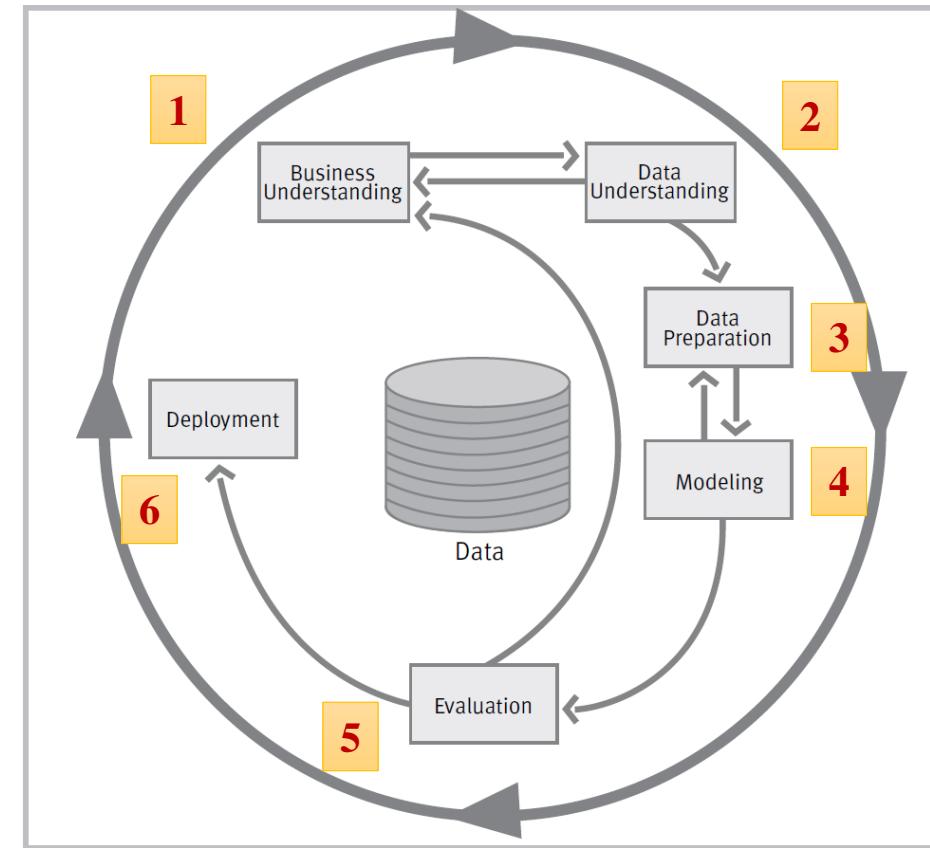
# CRISP-DM 資料探勘流程(續)

- 步驟 1：商業理解
- 步驟 2：資料理解
- 步驟 3：資料準備
- 步驟 4：模式建立
- 步驟 5：評估與測試
- 步驟 6：佈署應用

佔整專案時間的~**80%**

- 訓練資料70%
- 測試資料30%

# CRISP-DM 資料探勘流程(續)



參考 [https://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

# 數值模型績效指標

- 不可直接使用誤差的算術平均!

$$\cancel{\text{Total error}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$

- 均方誤差 (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 均方根誤差 (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- 平均絕對誤差 (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

# 類別模型績效指標 - 混淆矩陣

- <http://rwepa.blogspot.com/2013/01/rocr-roc-curve.html>

```
#           | 真實P類別 真實N類別
# ****|*****|*****|*****
# 預測P類別 | TP真陽數 FP假陽數
# 預測N類別 | FN假陰數 TN真陰數
# ****|*****|*****|*****
#           | P          N

# 1.TPR(True positive rate) 真陽性率，愈大愈好 -----
# =TP/ (TP+FN)
# =TP/P
# =Sensitivity 積敏度
# =Recall 召回率
# =Probability of detection
# =Power
# 實際為陽性的樣本中，判斷為陽性的比例。
# 例如真正有生病的人中，被醫院判斷為有生病者的比例。
```

# 集群法

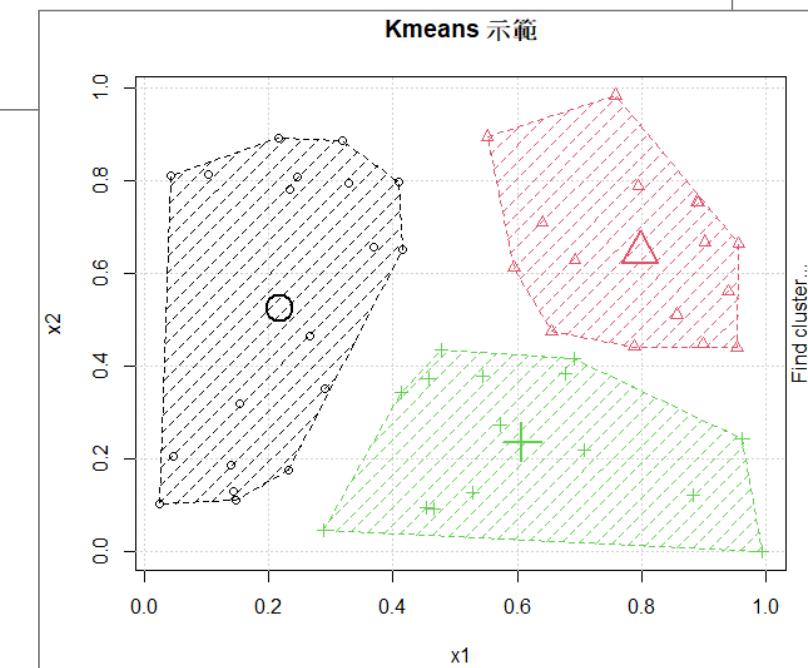
---

# 集群法 (Clustering)

- 集群法或稱為聚類分析,集群分析(Cluster analysis),叢集分析:是一種物以類聚方法.
- 每個集群的相似性是以資料間的距離來判斷.
- 分組後在同一集群組內的樣本點具有高度的相似性.
- 不同群組間的樣本點則具有高度的異質性.
- 集群法屬於非監督式學習法(Unsupervised learning),即資料沒有標籤(unlabeled data).
- 無法藉由的反應變數(Response variable, Y)來做分類之訓練.
- 因為資料沒有標籤,與監督式學習法不同,非監督式學習法較無法衡量演算法的正確率.

# R - animation 套件

```
1 library(animation)
2 set.seed(123)
3 kmeans.ani(x = cbind(x1 = runif(50), x2 = runif(50)),
4             centers = 3,
5             hints = c("Move centers!", "Find cluster..."),
6             pch = 1:3,
7             col = 1:3)
8 title(main = "Kmeans 示範")
```



## 2.2 深度學習

---

# 卷積神經網路

- 卷積神經網路（Convolutional Neural Network, CNN）是一種前饋式 (feed-forward) 神經網路，它的人工神經元對於大型圖像處理有出色表現。
- 卷積神經網絡由輸入層、隱藏層和輸出層組成。
- 卷積神經網路由一個或多個卷積層和頂端的全連通層（對應經典的神經網路）組成，同時也包括關聯權重和池化層（pooling layer）。

# 卷積神經網路 (續)

- 在前饋神經網絡中，任何中間層都被稱為隱藏層，因為它們的輸入和輸出被激活函數和最終卷積屏蔽。
- 在卷積神經網絡中，隱藏層包括執行卷積的層。
- 這一結構使得卷積神經網路能夠利用輸入資料的二維結構。與其他深度學習結構相比，卷積神經網路在圖像和語音辨識方面能夠給出更好的結果。這一模型也可以使用反向傳播演算法進行訓練。相比較其他深度、前饋神經網路，卷積神經網路需要考量的參數更少，使之成為一種頗具吸引力的深度學習結構。

# 卷積 (滑動+內積)

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Image

1	0	1
0	1	0
1	0	1

3x3矩陣

$$\begin{aligned}1 \times 1 + 1 \times 0 + 1 \times 1 &= 1 + 0 + 1 = 2 \\0 \times 0 + 1 \times 1 + 1 \times 0 &= 0 + 1 + 0 = 1 \\0 \times 1 + 0 \times 0 + 1 \times 1 &= 0 + 0 + 1 = 1 \\2 + 1 + 1 &= 4\end{aligned}$$

# 卷積 (續)

1 <small>x1</small>	1 <small>x0</small>	1 <small>x1</small>	0	0
0 <small>x0</small>	1 <small>x1</small>	1 <small>x0</small>	1	0
0 <small>x1</small>	0 <small>x0</small>	1 <small>x1</small>	1	1
0	0	1	1	0
0	1	1	0	0

Image

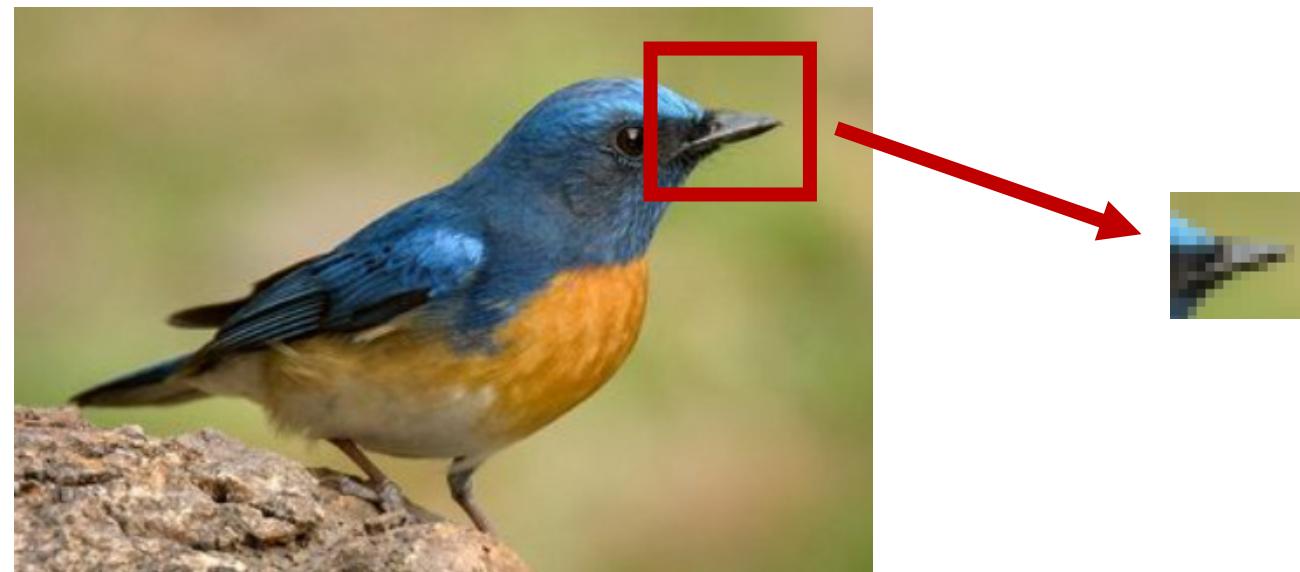
4		

Convolved  
Feature

目標是找出最佳的  
Convolved Feature  
(卷積特徵)- filter

# 卷積層的特色

- 保留圖片中的空間結構，並從此結構中萃取出特徵



# 池化層 (Pooling Layer)

- 池化層的主要概念是，當我們在做圖片的特徵萃取的過程中，圖形的縮放應該不會影響到我們的目的，經由這樣的 scaling 我們也可以再一次的減少神經網路的參數。

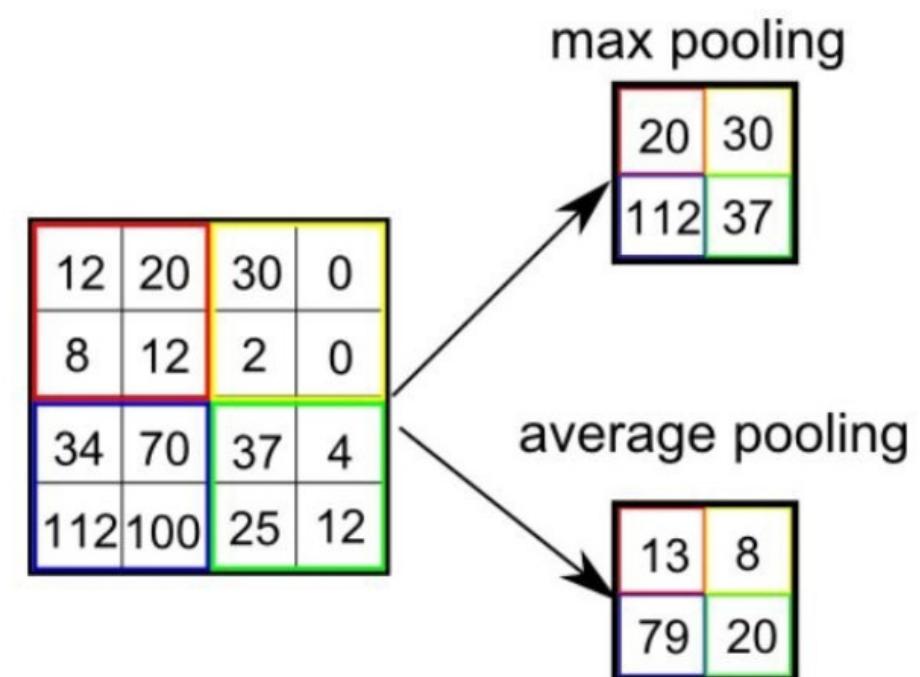


subsampling



# Pooling Layer 池化層

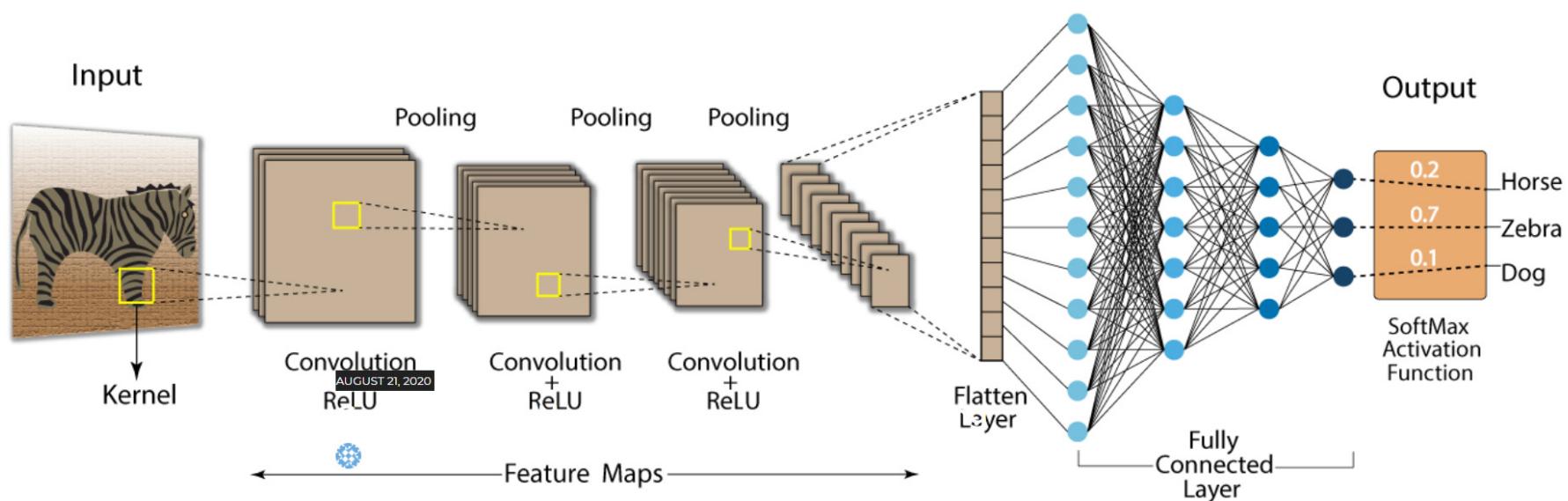
- 圖片的特徵萃取的過程中，圖形的縮放應該不會影響到我們的目的，經由這樣的 scaling 我們也可以減少神經網路的參數。
- 常用的 pooling 方式有 Max pooling 與 Average pooling
- 目前主要使用 Max pooling



# 全連接層

- 這邊的全連接層跟我們進行手寫辨識的方式一樣，說穿了就是一個分類器，把我們經過數個卷積、池化後的結果進行分類。

**Convolution Neural Network (CNN)**



<https://developersbreach.com/convolution-neural-network-deep-learning/>

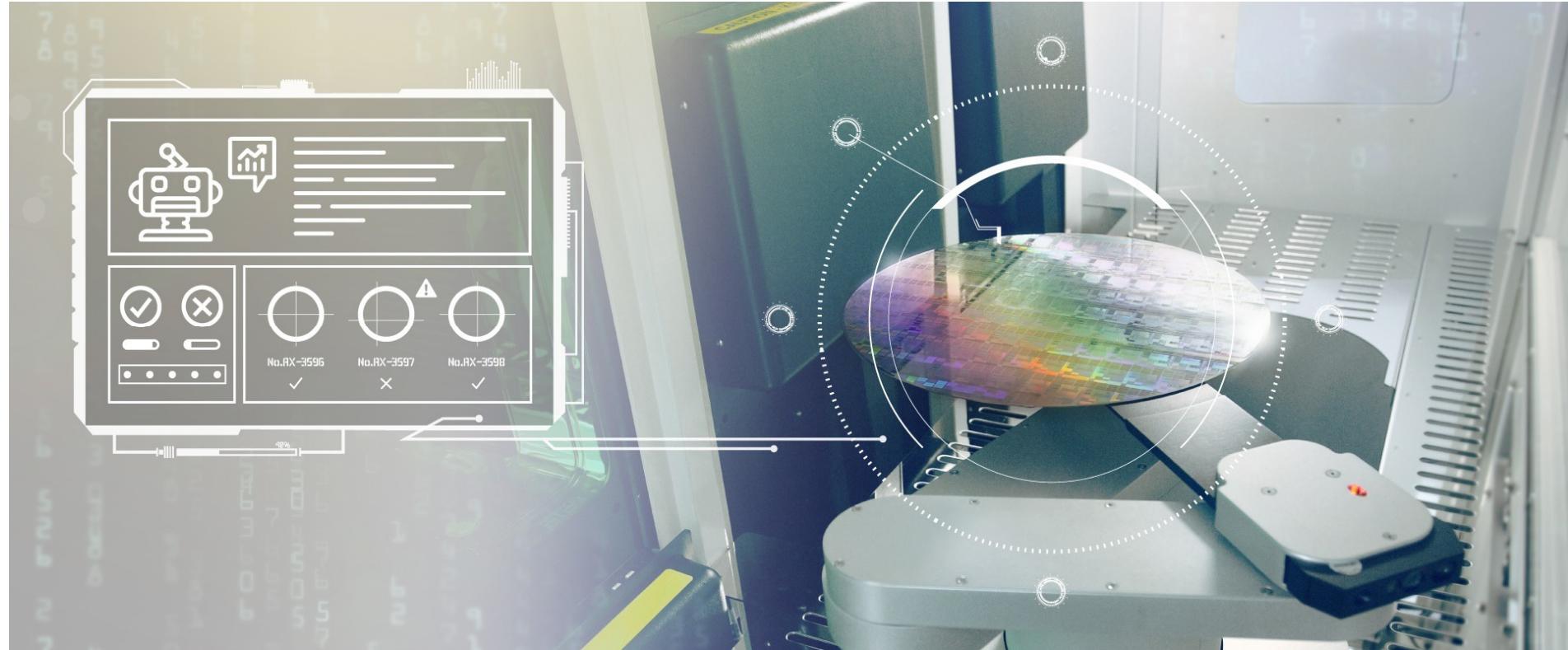
## 2.3 影像辨識

(Image recognition)

# 影像辨識原理

- 人類的腦神經是由許多神經元所組成，雖然每個神經元都只能接收簡單的訊號，但神經元會將其接收到的訊號傳遞給其他神經元，而後其他神經元接連被傳遞的訊號觸發，再結合訊號轉發給其他神經元，形成一個龐大、可處理複雜訊號的訊息處理網路；也因此，人類的視神經具有強大的圖像辨識能力。
- 影像辨識的原理就在於模仿人類視神經，先透過對**邊界的認識強化**，再逐步**組合**出圖像識別的運作方式。
- 套用在深度學習的影像辨識作法就會變成先將圖片分解成許多小像素，做為第一層的輸入資料，接著再經過多層次的演算法處理，從個別像素**擷取特徵**、**組合特徵**，再到最後的輸出層結果來完成影像辨識。

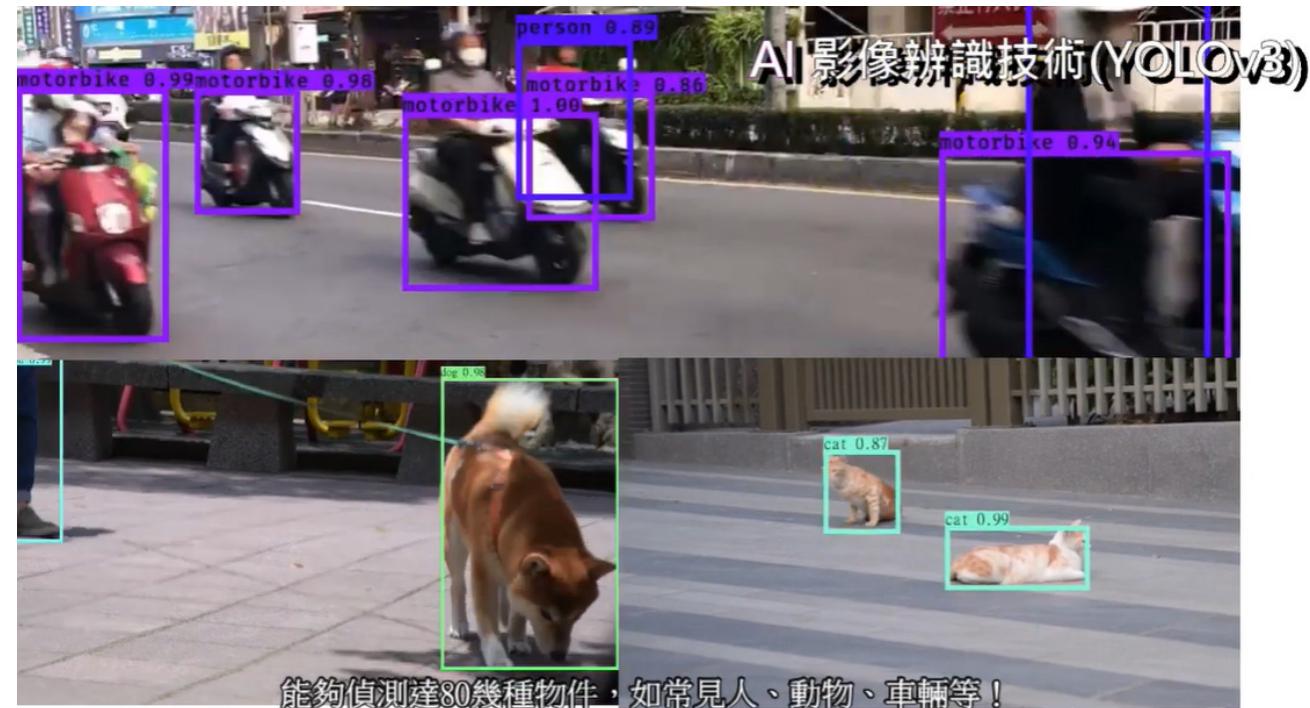
# 台塑勝高科技率先導入AI影像辨識在產線整合應用



參考: <https://www.viatech.com/tw/2019/11/via-provides-advanced-ai-technology-to-formosa-sumco-technology-fst-tw/>

<https://youtu.be/oi4npf1tC-k>

- YOLO 演算法 (You Only Look Once), 2016



## 2.4 聊天機器人

(ChatBot)

# 聊天機器人

- 聊天機器人是經由對話或文字進行交談的電腦程式。能夠模擬人類對話，通過圖靈測試(Turing test, 1950)。
- 聊天機器人可用於實用的目的，如客戶服務或資訊獲取。
- 有些聊天機器人會搭載自然語言處理系統，但大多簡單的系統只會擷取輸入的關鍵字，再從語料庫中找尋最合適的應答句。
- 目前聊天機器人是虛擬助理（如Google智能助理）的一部分，可以與許多組織的應用程式，網站以及即時訊息平台(Facebook Messenger)連接。非助理應用程式包括娛樂目的的聊天室，研究和特定產品促銷，社交機器人。
- 聊天機器人是以 AI、自動化規則、自然語言處理 (NLP) 和機器學習 (ML) 推動，負責處理資料以回應各種類型的要求。

參考: <https://zh.wikipedia.org/wiki/聊天機器人>

# 聊天機器人二大類型

- 任務導向型：以工作為導向（宣告式）的聊天機器人，是專注於執行一項功能的單一用途程式。
- 非任務導向型：以資料驅動和預測性（對話式）聊天機器人為主，是虛擬助理或數位助理，且比以工作為導向的機器人更為精密許多、互動性更高，也更加個人化。

參考：

- <https://www.oracle.com/tw/chatbots/what-is-a-chatbot/>
- AI 大局：鳥瞰人工智慧技術全貌，重塑 AI 時代的領導力，  
古明地正俊、長谷佳明著, 旗標 <https://www.flag.com.tw/books/product/F0322>

# 任務導向型聊天機器人

- 是專注於執行一項功能的單一用途程式。
- 它們會以規則、NLP 和極少的 ML 產生自動化的對話式回應，以答覆使用者的查詢。
- 與這些聊天機器人的互動十分**具體並結構化**，且最適用於支援和服務功能。
- 以工作為導向的聊天機器人可以處理常見的問題，或如營業時間查詢或不涉及各種變數的簡單交易。
- 儘管聊天機器人確實會使用 NLP，讓終端使用者享有聊天機器人提供的對話式體驗，但其功能卻是**相當基本**，這些是目前最常用的聊天機器人。

# 非任務導向型聊天機器人

- 依照不同環境動作，並運用自然語言理解 (Natural-language understanding, NLU)、NLP 和 ML 等技術。
- 它們應用**預測性智慧和分析**，根據使用者個人資料和過去的使用者行為實現個人化。
- 數位助理可以隨著時間學習使用者的偏好、提供建議，甚至**預測需求**。
- 除了監控資料和意向之外，它們還可以展開對話功能。
- Apple 的 Siri 和 Amazon 的 Alexa 即是以消費者為導向、資料驅動的預測性聊天機器人。
- 進階數位數理能夠在同一平台下連線到數台單一用途聊天機器人、從各聊天機器人提取不同資訊，然後合併此資訊，以執行工作，同時仍掌控全局。

參考: <https://www.oracle.com/tw/chatbots/what-is-a-chatbot/>

## 2.5 感測器

---

# 感測器

- 感測器（英語：Sensor）是用於偵測環境中所生事件或變化，並將此訊息傳送出至其他電子裝置（如中央處理器）的裝置，通常由感測元件和轉換元件組成。
- 人類會基於視覺、聽覺、嗅覺、觸覺獲得的資訊進行行動，設備也一樣，根據感測器獲得的資訊進行控制或處理。
- 感測器收集轉換的訊號（物理量）有溫度、光、顏色、氣壓、磁力、速度、加速度等。這些利用了半導體的物質變化，除此之外，還有利用酶和微生物等生物物質的**生物感測器**。



# 自駕車的感測器

[https://youtu.be/Oy0d3v\\_afIQ](https://youtu.be/Oy0d3v_afIQ)



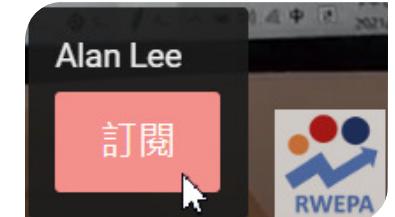
# 自動駕駛車5大等級

<https://youtu.be/CsvOip-Je7I>

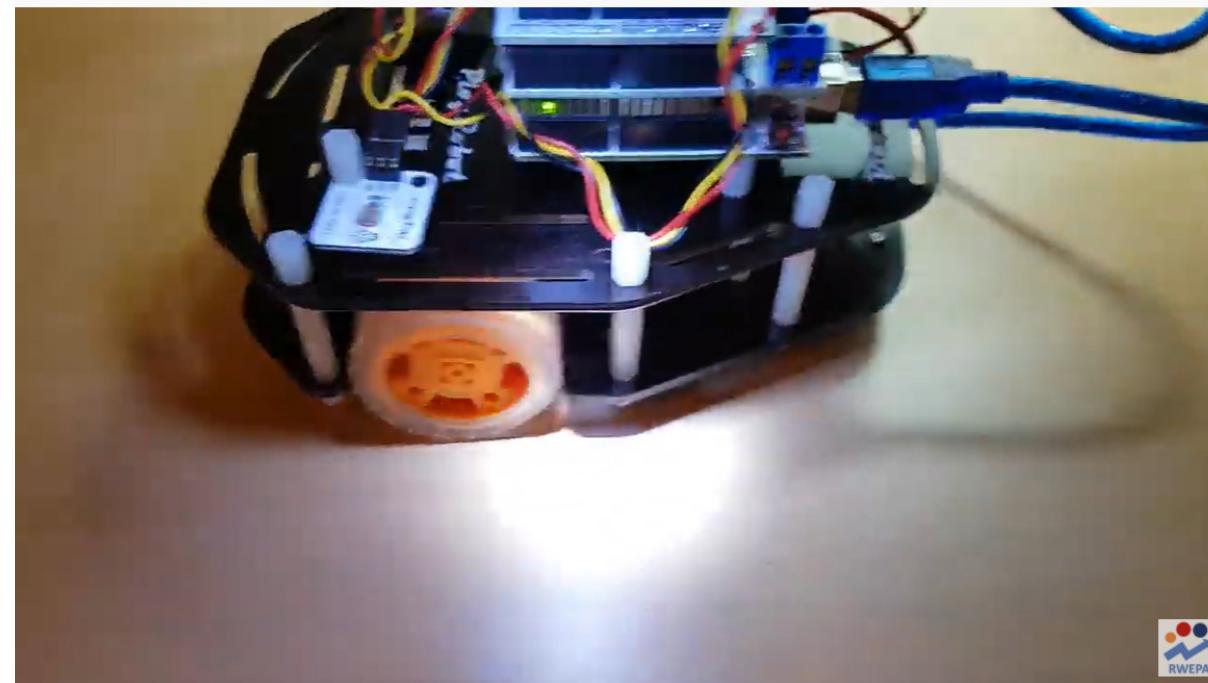


# 跟著那道光 - Arduino Car

按訂閱

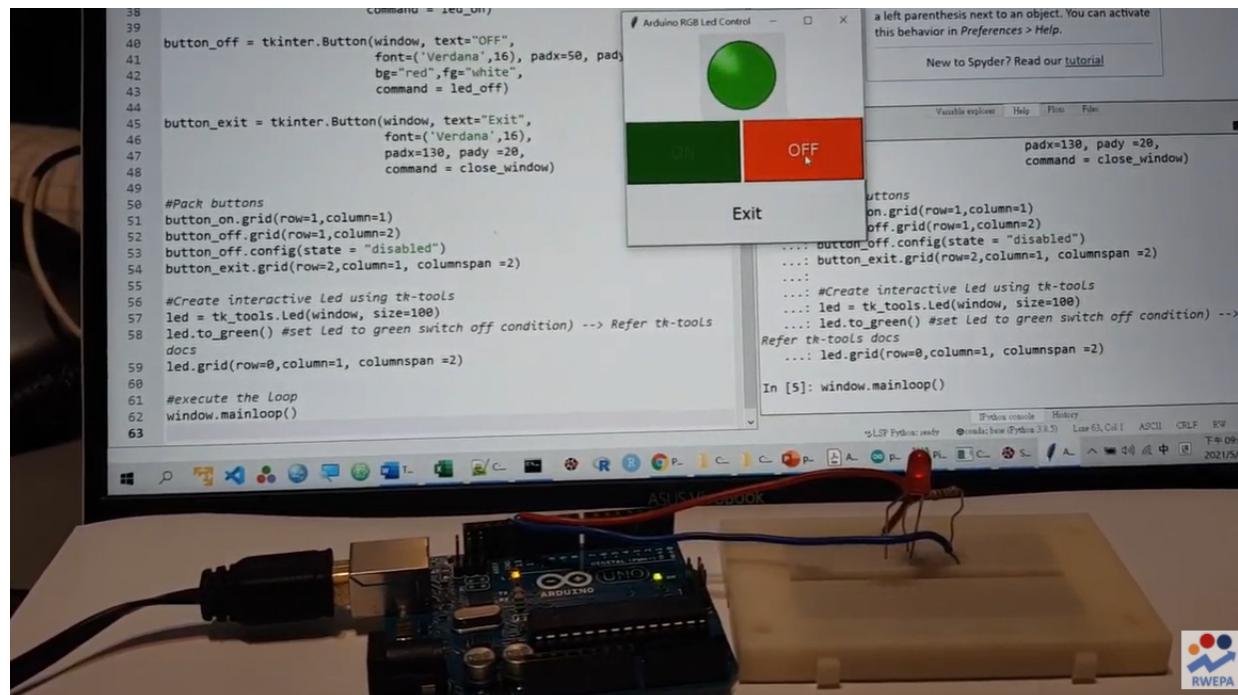


- <https://youtu.be/SXyq6urlTQo>
- <http://rwepa.blogspot.com/2021/05/arduino-car.html>

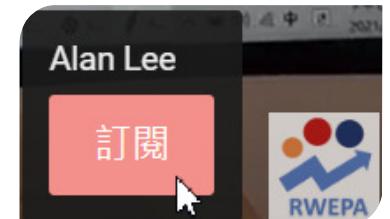


# Arduino + Python tkinter 套件控制LED應用

- <https://youtu.be/LjgFIm1S7tw>
- <http://rwepa.blogspot.com/2021/05/arduino-python-tkinter-led.html>



按訂閱



## 2.6大數據分析

---

# 何謂大數據

- 大數據(巨量資料, Big Data) 指的是所涉及的資料量規模巨大到無法透過人工，在合理時間內達到擷取、管理、處理、並整理成為人類所能解讀的形式的資訊

```
1546107 World,WLD,"Presence·of·peace·keepers  
1546108 Samoa,WSM,"Presence·of·peace·keepers  
1546109 "Yemen,·Rep.",YEM,"Presence·of·peace  
1546110 South·Africa,ZAF,"Presence·of·peace·  
1546111 "Congo,·Dem.·Rep.",ZAR,"Presence·of·  
1546112 Zambia,ZMB,"Presence·of·peace·keeper
```

維基: [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data)

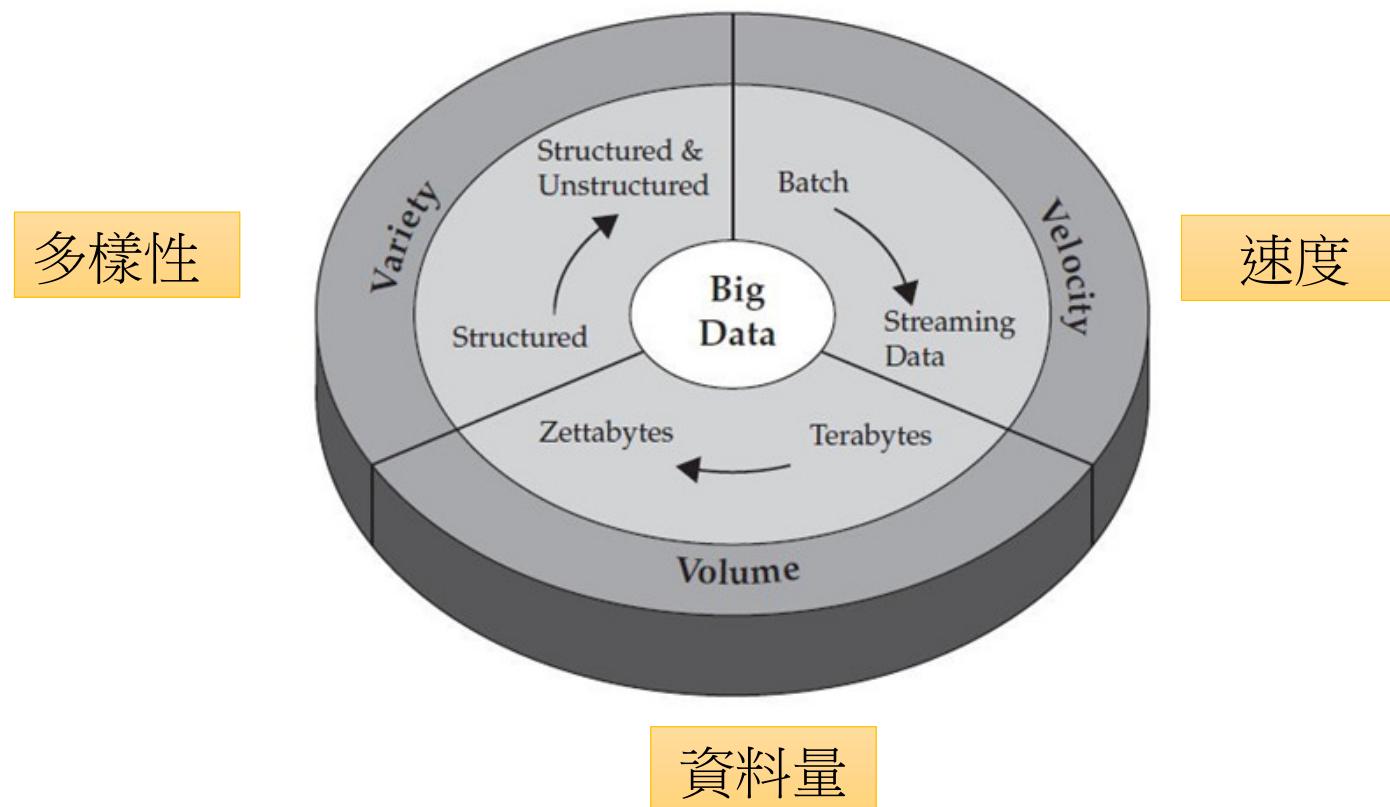
154萬筆資料

# 巨量資料的3V定義

- 2001年Gartner 公司的分析師 Douglas Laney 提出3V  
(三個關鍵挑戰)
  - 資料量 (*Volume*) : 資料總量很大
  - 速度 (*Velocity*) : 資料產生的速度快
  - 多樣性 (*Variety*) : 資料種類繁多

參考: Lancy, D., 3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner, 2001.

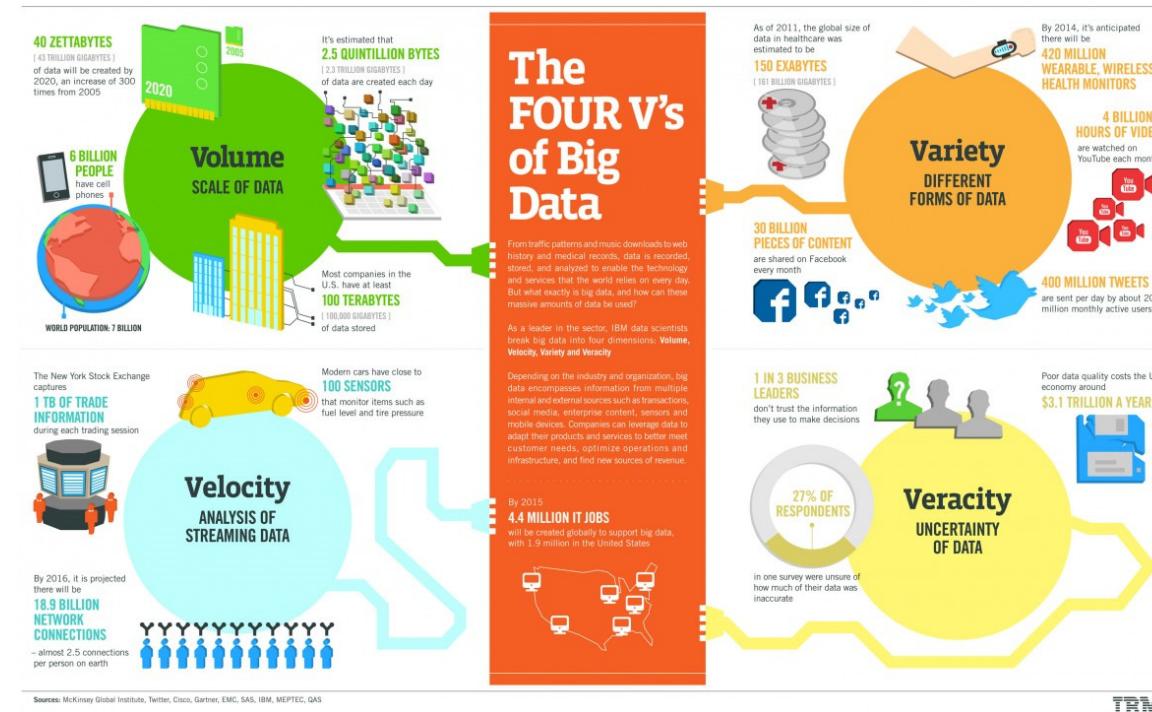
## IBM-3V



參考: IBM, Understanding Big Data, 2012.

# IBM的4V定義

- IBM 的 Big Data 4V: 新增 Veracity (真實性)



參考: <https://opensistemas.com/en/the-four-vs-of-big-data/>

# IBM - Veracity

**1 IN 3 BUSINESS  
LEADERS**

don't trust the information  
they use to make decisions



**27% OF  
RESPONDENTS**

in one survey were unsure of  
how much of their data was  
inaccurate

**Veracity**  
**UNCERTAINTY  
OF DATA**

Poor data quality costs the US  
economy around  
**\$3.1 TRILLION A YEAR**



# IBM 的 Big Data 5V

- Volume 資料量
- Velocity 速度
- Variety 多樣性
- Variability (Veracity) 可變性 (真實性)
- 新增 Value 價值

參考 <https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/>

# 科技驅動與資料加值

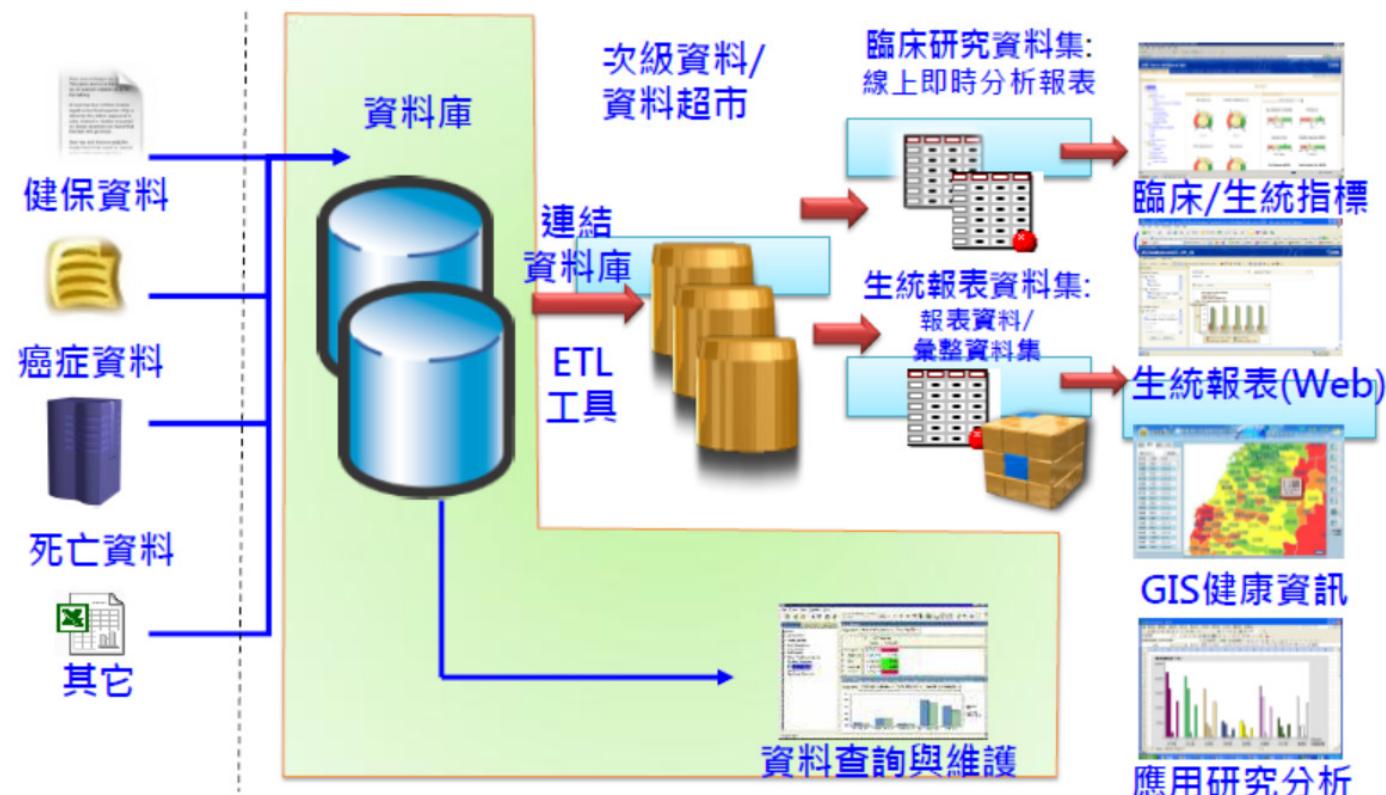


# 加值服務趨勢

- 加值服務(Value-Added Service, VAS)
  - 將某項非核心技術、產品或服務利用新方式加以修正改善，以創造更高的**價值**。
  - VAS廣泛應用於各產業及研究領域，以通訊及網路產業為首的例子分別為Web2.0應用及行動上網等(維基百科)。
- 行政院科技會報辦公室於101年1月18日
  - 政府資料加值(open data)推動策略會議
    - GIS、健康醫療、行動服務

# 健康資料加值應用平台

INPUT → PROCESS → OUTPUT



- 健康資料研究與加值應用, 2013. [LINK]
- 衛生福利部資料加值應用雲端化服務平台簡介, 2018.

<https://www.bas-association.org.tw//catalog/arts/010703088.pdf>

### 3. 程式設計(大數據分析工具)

---

# 程式設計(大數據分析工具)

- Python (免費程式語言)
- R (免費程式語言)
- Julia (免費程式語言)
- Java
- C++
- .NET
- SPSS
- SAS
- Excel

# 參考資料

- RWEPA
  - <https://rwepa.blogspot.com/>
- Python 程式設計-李明昌 <免費電子書>
  - <https://rwepa.blogspot.com/2020/02/pythonprogramminglee.html>
- R入門資料分析與視覺化應用教學(付費)
  - <https://mastertalks.tw/products/r?ref=MCLEE>
- R商業預測與應用(付費)
  - <https://mastertalks.tw/products/r-2?ref=MCLEE>

# 謝謝您的聆聽

## Q & A



李明昌

EMAIL: [alan9956@gmail.com](mailto:alan9956@gmail.com)

WEB: <http://rwepa.blogspot.com/>