著作權等聲明公告

本資料所含內容與相關附屬文件均為經濟部產業發展署（下稱本署）及所屬人員職務上所完成之著作，本署依法擁有著作權及其他相關智慧財產權，而受著作權法及相關法規保護。業經申請並經本署授權同意使用之個人、法人等，於使用時敬請註明出處，並僅限非商業用途之使用。

謹提醒，倘未取得同意或授權，而逕自重製、改作、公開傳輸或有任何侵害本署著作權之行為者，本署將視違法情節逕行依相關法律追訴。另提醒，如有違法情事，則依不同情節，除行為人個人應負賠償責任以外其所屬單位、法人亦可能應負連帶責任。

經濟部產業發展署
產業AI三日班公版教材

單元四 機器學習技術理論與案例

經濟部產業發展署
Industrial Development Administration
Ministry of Economic Affairs

目錄

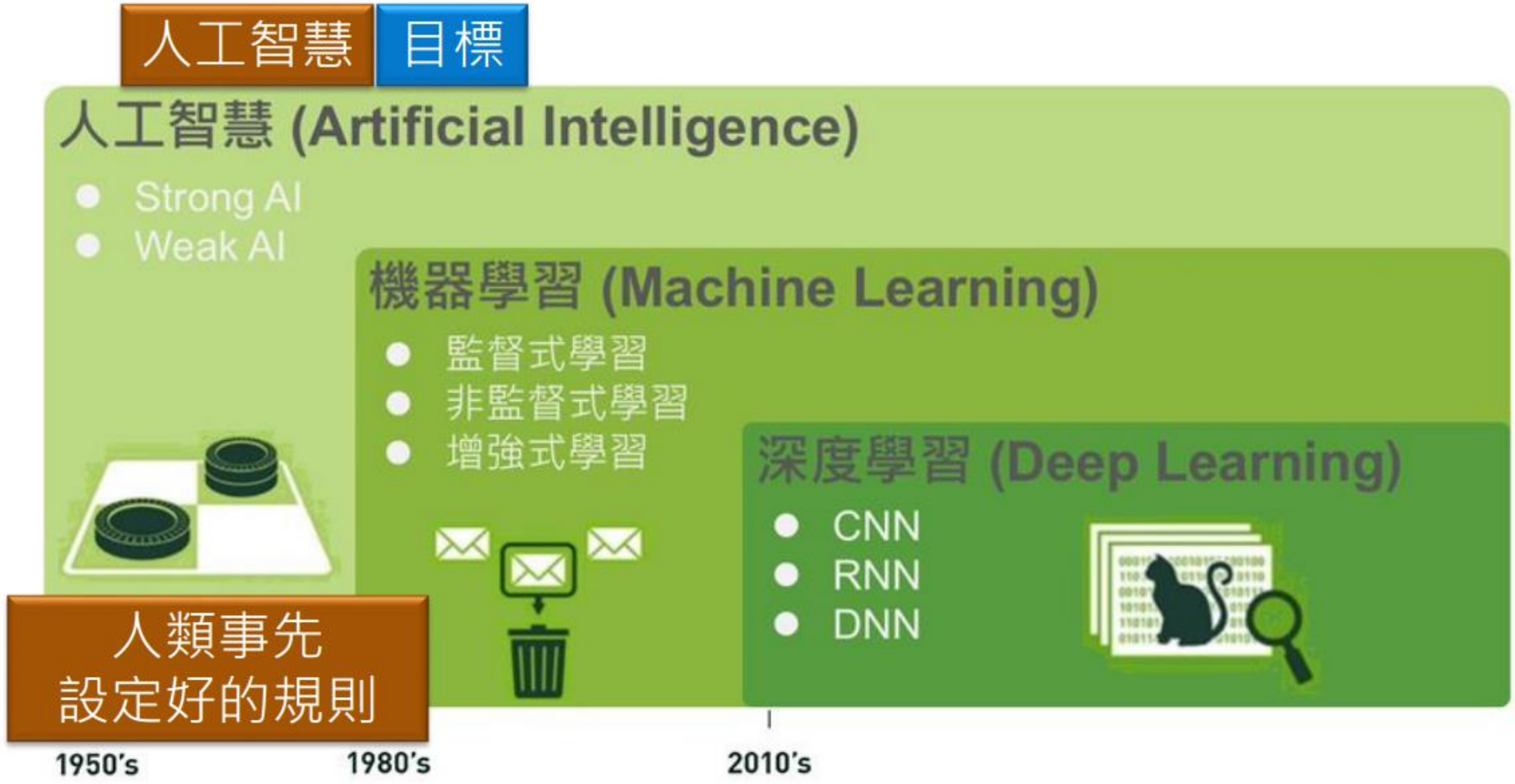
1. 課程目標與先備知識

- 在課程之前，建議宜具備的知識與經驗
  - ➢計算機概論
  - ➢使用網路服務與查找資料的經驗

- 教學目標：
  - ➢瞭解機器學習的重要性
  - ➢瞭解機器學習的方法原理
  - ➢瞭解機器學習的使用方法
  - ➢培養具有建構學習模型的能力

機器學習的基本原則—

了解「人工智慧、機器學習、深度學習」的關連性

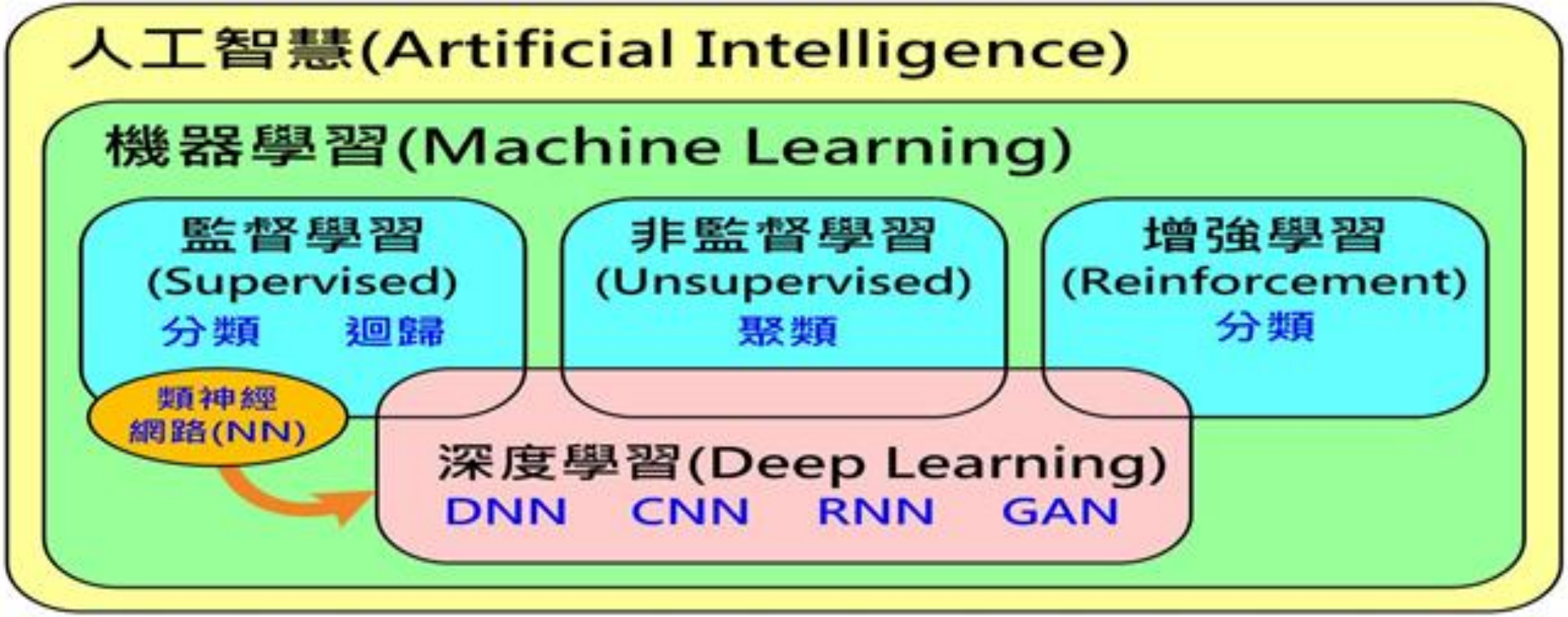


Source of image: https://blogs.nvidia.com.tw/2016/07/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

機器學習的基本原則—

了解「人工智慧、機器學習、深度學習」的關連性

（圖片來源：OmniXRI 整理繪製）



機器學習的基本原則—

關於人工智慧

人工智慧(Artificial Intelligence) 是指讓電腦具有人類的知識與行為，有下列特性：

- **學習能力**
- **判斷能力**
- **記憶能力**
- **語言能力**

機器學習的基本原則—

關於Strong AI(強人工智慧)

- 相當接近人類智慧，可以像人類一樣學習、理解和創造新知識的人工智慧並完成人類所能做的大部分工作的資訊系統。

- 目前，強人工智慧還未真正出現，但是有一些研究者和科學家正在尋找如何開發強人工智慧。像是能夠進行自我學習的AlphaGo就是一個擁有相對強大能力的人工智慧，但就強人工智慧角度來看，AlphaGo的能力相比下仍然有一大段距離。

機器學習的基本原則—

關於Weak AI(弱人工智慧)

- 弱人工智慧（applied AI，narrow AI，weak AI, artificial narrow intelligence, ANI）只處理特定的問題。弱人工智慧不需要具有人類完整的認知能力，甚至是完全不具有人類所擁有的感官認知能力，只要設計得看起來像有智慧型就可以了。

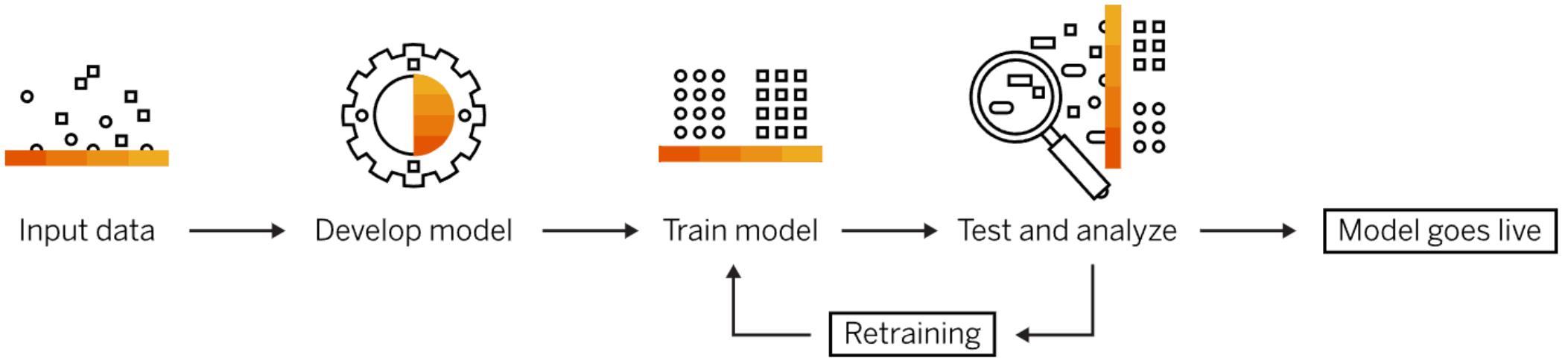
- 只能在特定的領域中進行工作。目前，大多數人工智慧的應用都屬於弱人工智慧。例如，Siri、Alexa、Google 語音助手，它們可以聽懂你的問題，並回答你的問題。

機器學習的基本原則

- 機器學習(Machine Learning) 是人工智慧的一個部份，就是透過特殊演算法，讓電腦能經由訓練從一大堆數據中找出規律性並產生模型，然後利用訓練後產生的模型進行預測，輸入的數據越來越多，機器也會自動學習做出更精準的分析。

- 機器學習演算法主要用於
  - **分類事物**
  - **辨識模式**
  - **預測結果**
  - **做出判斷**

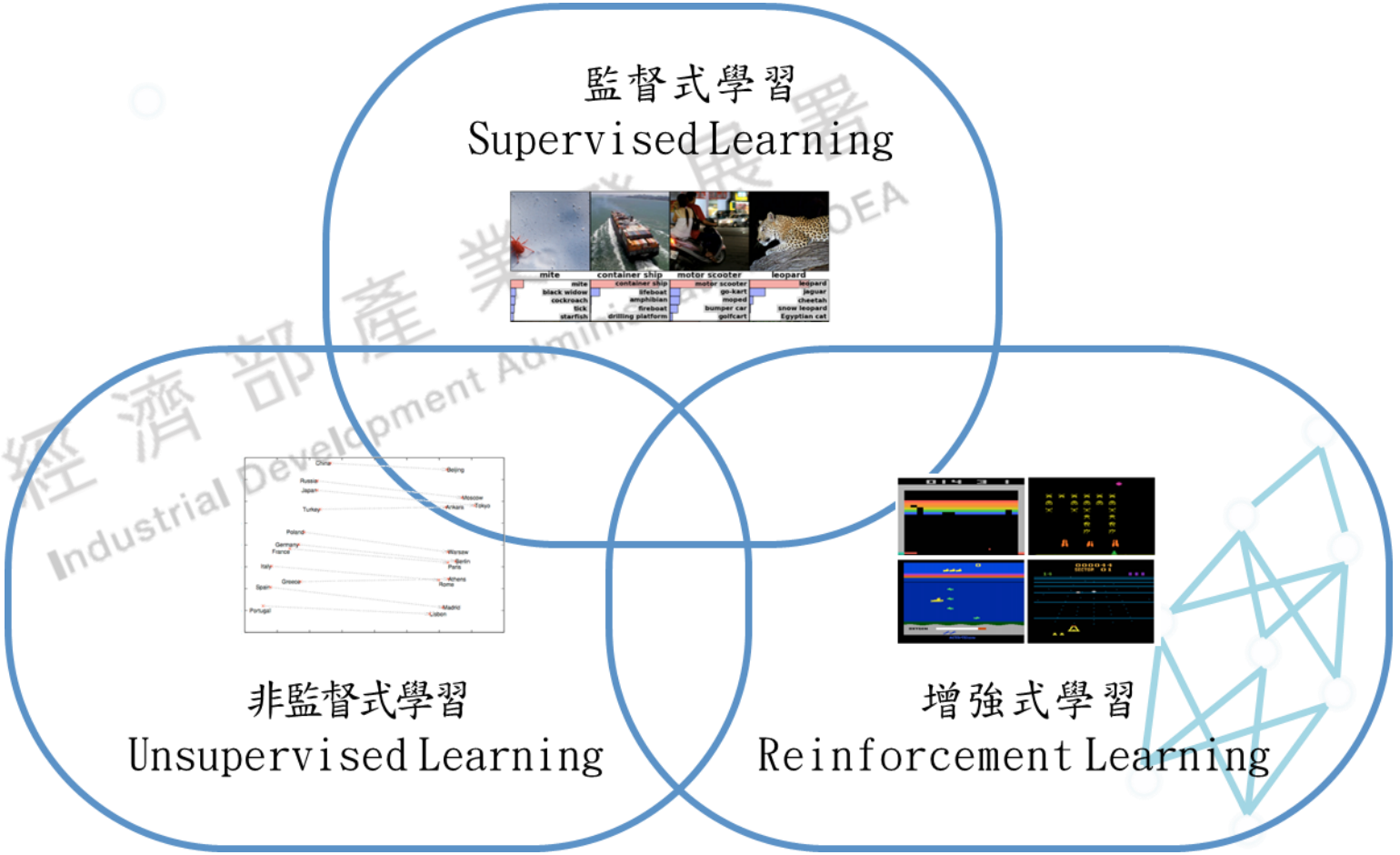
機器學習的運作模式—運作流程



資料來源：引用自https://www.sap.com/taiwan/products/artificial-intelligence/what-is-machine-learning.html

機器學習的運作模式—學習類型

- 監督式學習
- 非監督式學習
- 增強式學習



資料來源：引用自https://www.slideshare.net/slideshow/tensorflow-61523042/61523042

機器學習的運作模式—監督式學習

- **特性**
  - ➢使用標籤化訓練資料，建構出人工智慧模型。可以針對未曾見過的輸入來預測輸出。
  - ➢可以進行知識萃取，所得規則易於理解。有助於建構可解釋的AI模型。

- **應用範例**
  - ➢可用於規則未涵蓋的異常值檢測： 例如：新欺詐樣態辨識、信用卡盜刷。

機器學習的運作模式—監督式學習(範例)

監督式學習演算法會以範例訓練機器，學習模式包含「輸入」和「輸出」資料配對，其中輸出會標示期望值

資料來源：引用自https://www.slideshare.net/slideshow/tensorflow-61523042/61523042

機器學習的運作模式—非監督式學習

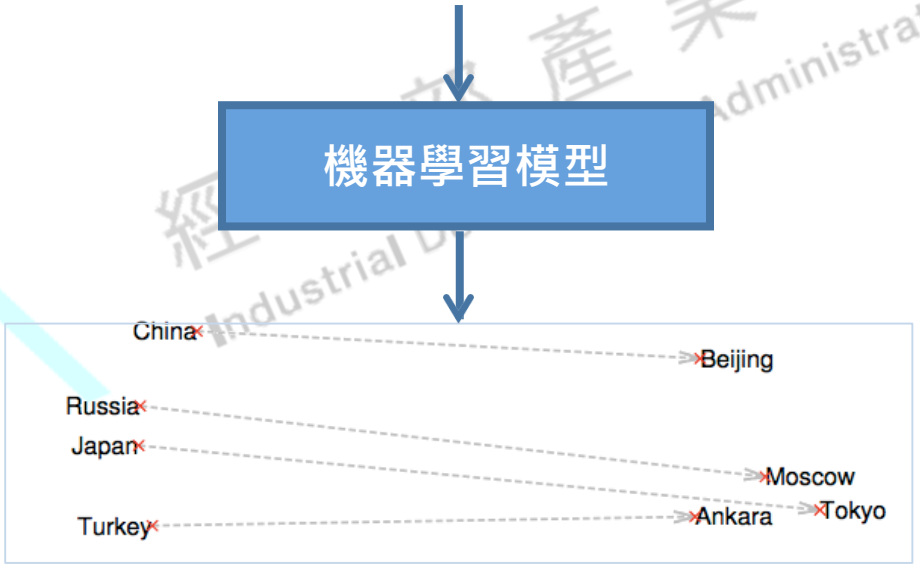
- **特性**
  - ➢使用未標籤化訓練資料，建構出函式，針對未曾見過的輸入，預測輸出。
  - ➢學習“通常發生什麼”。
  - ➢無輸出。
  - ➢聚類：分類相似的實例。

- **應用範例**
  - ➢CRM 中的客戶細分。
  - ➢圖像壓縮：色彩量化。
  - ➢生物資訊學：蛋白質結構。

機器學習的運作模式—非監督式學習(範例)

資料

Beijing is the capital of China.
As China's capital, Beijing is a large and vibrant city. Tokyo is the capital of Japan.
As Japan’ s capital, Tokyo is a large and vibrant city.
.....



結果

非監督式學習是在模仿人類如何觀察世界，運用直覺和經驗將事情分類

資料來源：引用自https://www.slideshare.net/slideshow/tensorflow-61523042/61523042

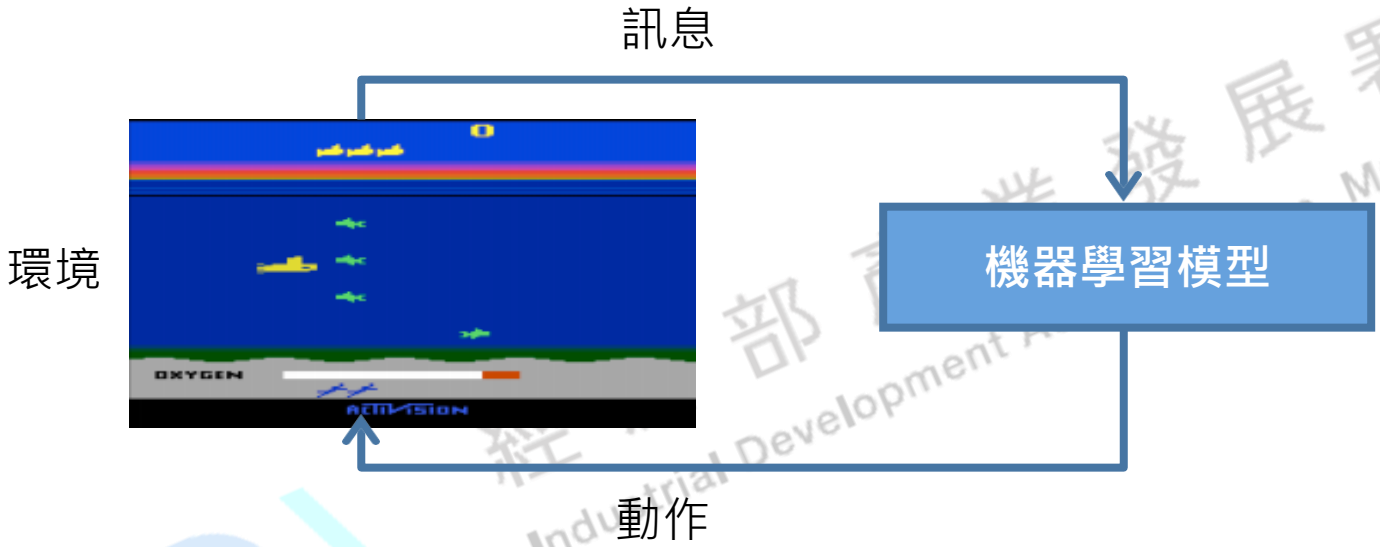
機器學習的運作模式—增強式學習

增強式學習(Reinforcement Learning)概念

- **特性**
  - 學習策略：從未曾學過的資料中找出輸出序列。
  - 沒有監督輸出可以參考，但延遲報酬。

- **應用範例**
  - 信用分配問題。
  - 玩遊戲。
  - 機器人迷宮。

機器學習的運作模式—增強式學習(範例)



增強式學習模式不包含參考答案，而是輸入一系列允許的動作、規則和潛在結束狀態

資料來源：引用自https://www.slideshare.net/slideshow/tensorflow-61523042/61523042

機器學習的運作模式—應用類型

- **分類Classification/Categorization**
  - 垃圾信過濾、物體辨識、片語辨識。
- **分群 Clustering**
  - 非監督式。
  - 找出自然群組：文件、搜尋結果、人。
- **迴歸分析**
  - 監督式學習方法 。
  - 迴歸分析是尋找資料間關係，可能是線性關係也可能非線性關係。
  - 迴歸演算法的最終目標是在數據之間繪製最佳擬合線或曲線。
- **資訊粹取**：學習評分

機器學習的運作模式—數據分類

序號	年齡	收入	健康狀況	買保險
1	55	40000	差	有
2	42	28000	佳	無
3	22	45000	差	有
4	25	50000	差	有
5	32	30000	佳	無

訓練

序號	年齡	收入	健康狀況	買保險
1	23	31000	佳	?
2	55	52000	差	?

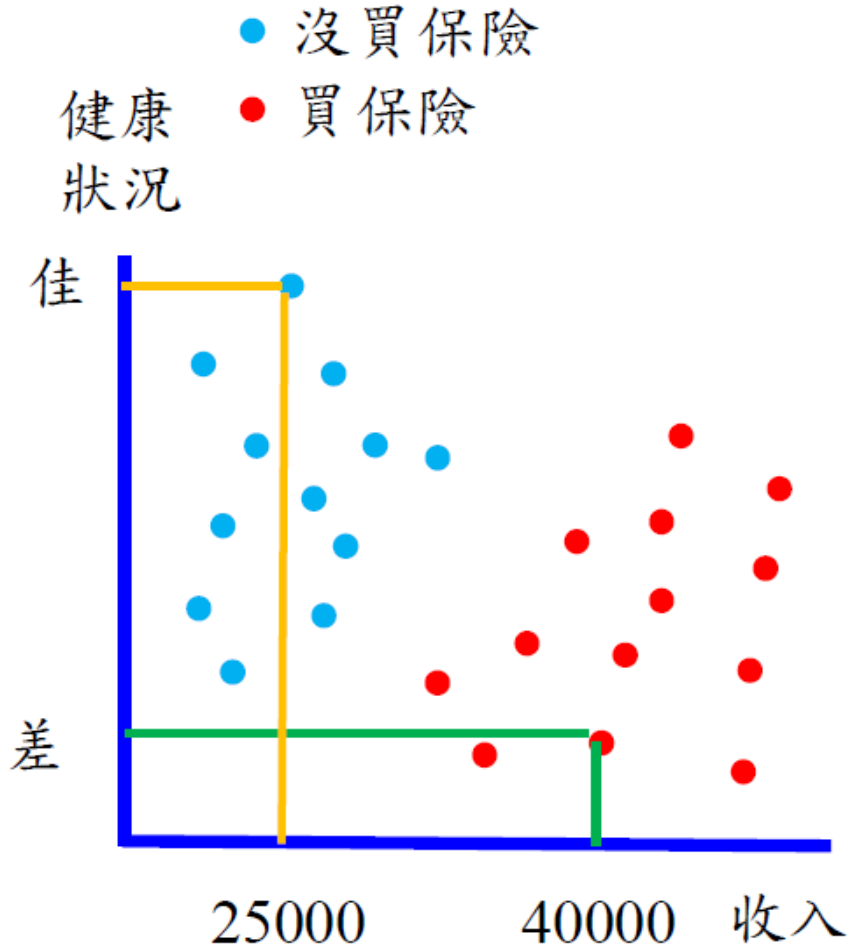
模型

分類
預測

資料來源：引用自淡江大學張志勇教授

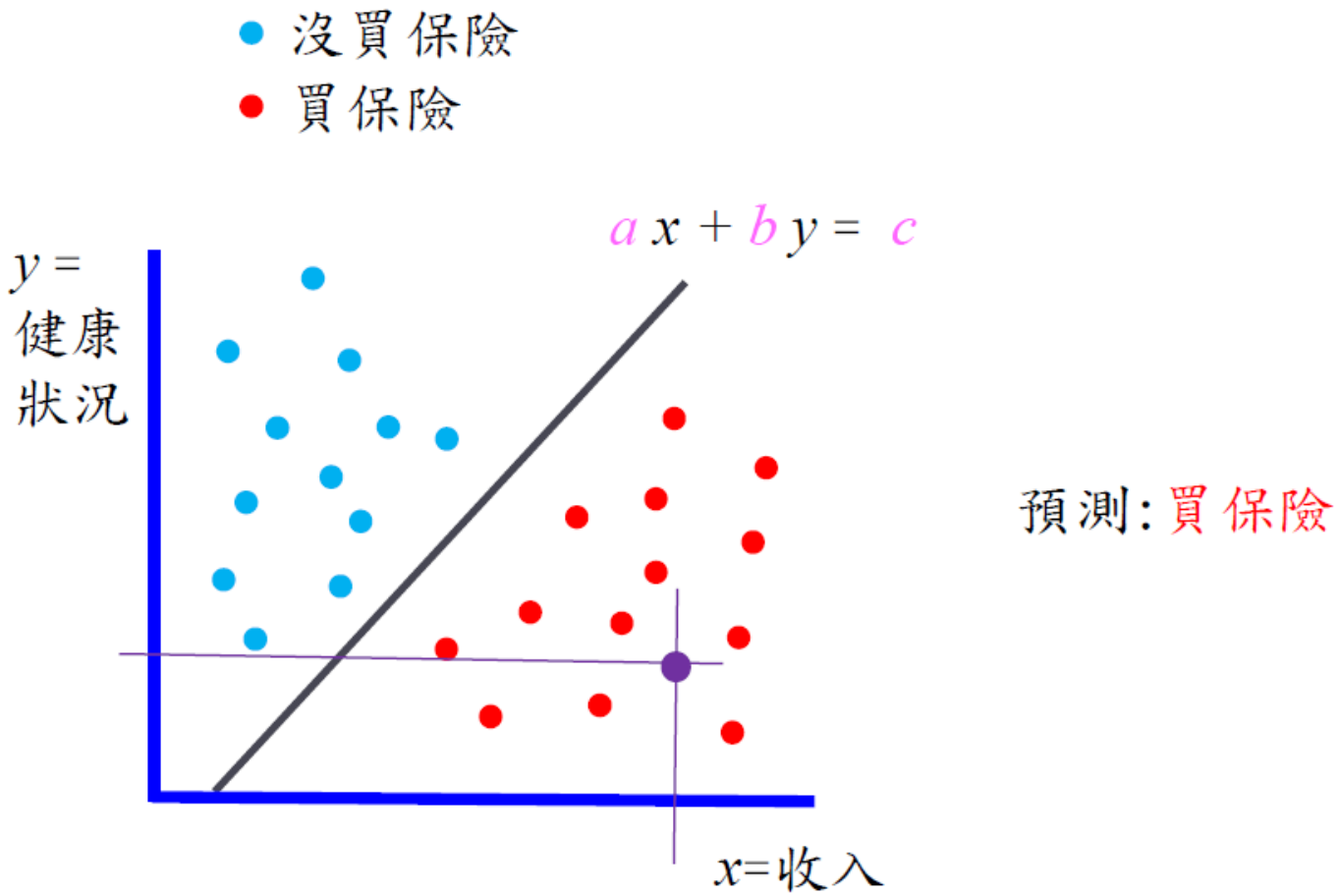
機器學習的運作模式—**數據分類**

收入	健康狀況	買保險
40000	差	有
60000	中	有
25000	佳	無



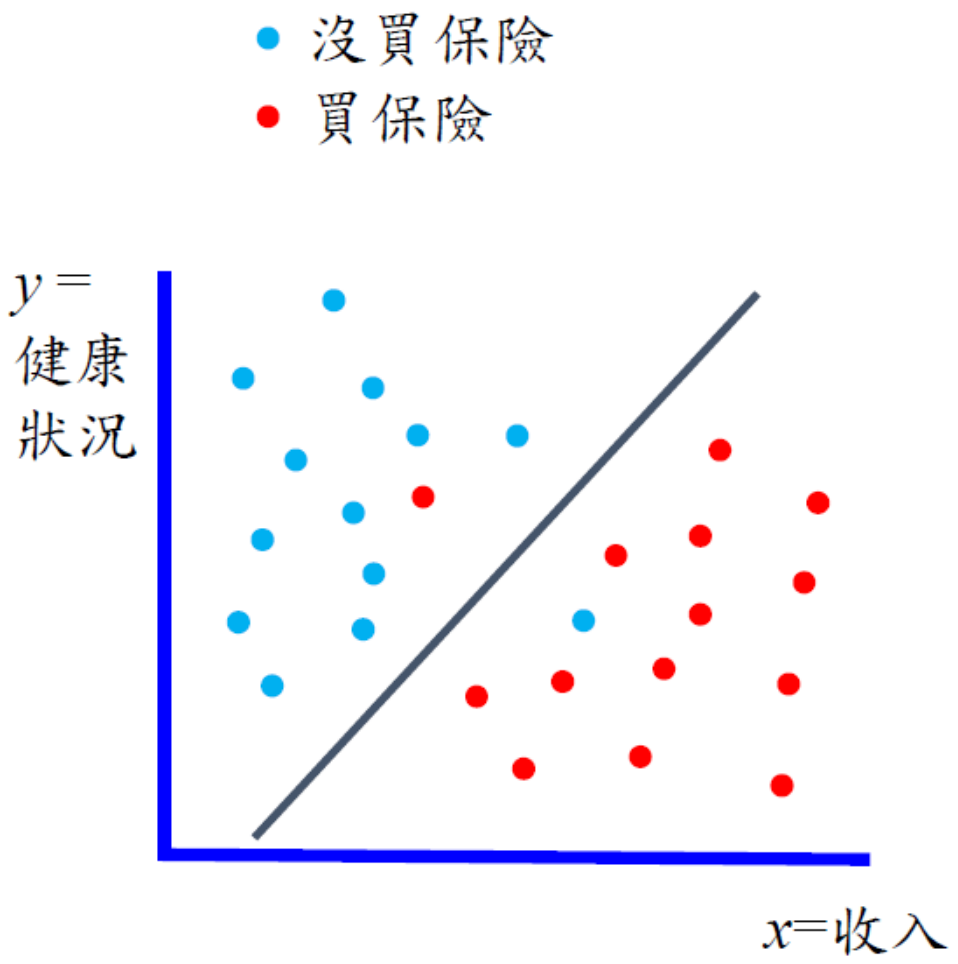
資料來源：引用自淡江大學張志勇教授

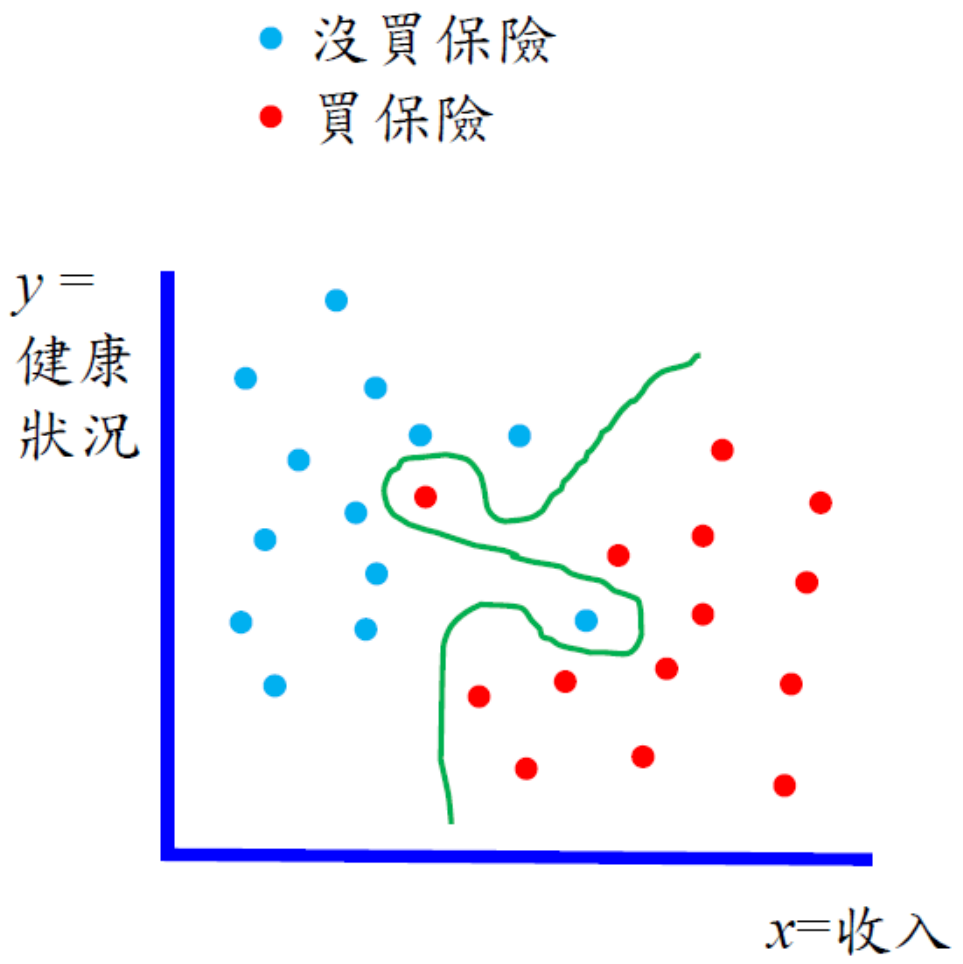
機器學習的運作模式—**數據分析與預測**



資料來源：引用自淡江大學張志勇教授

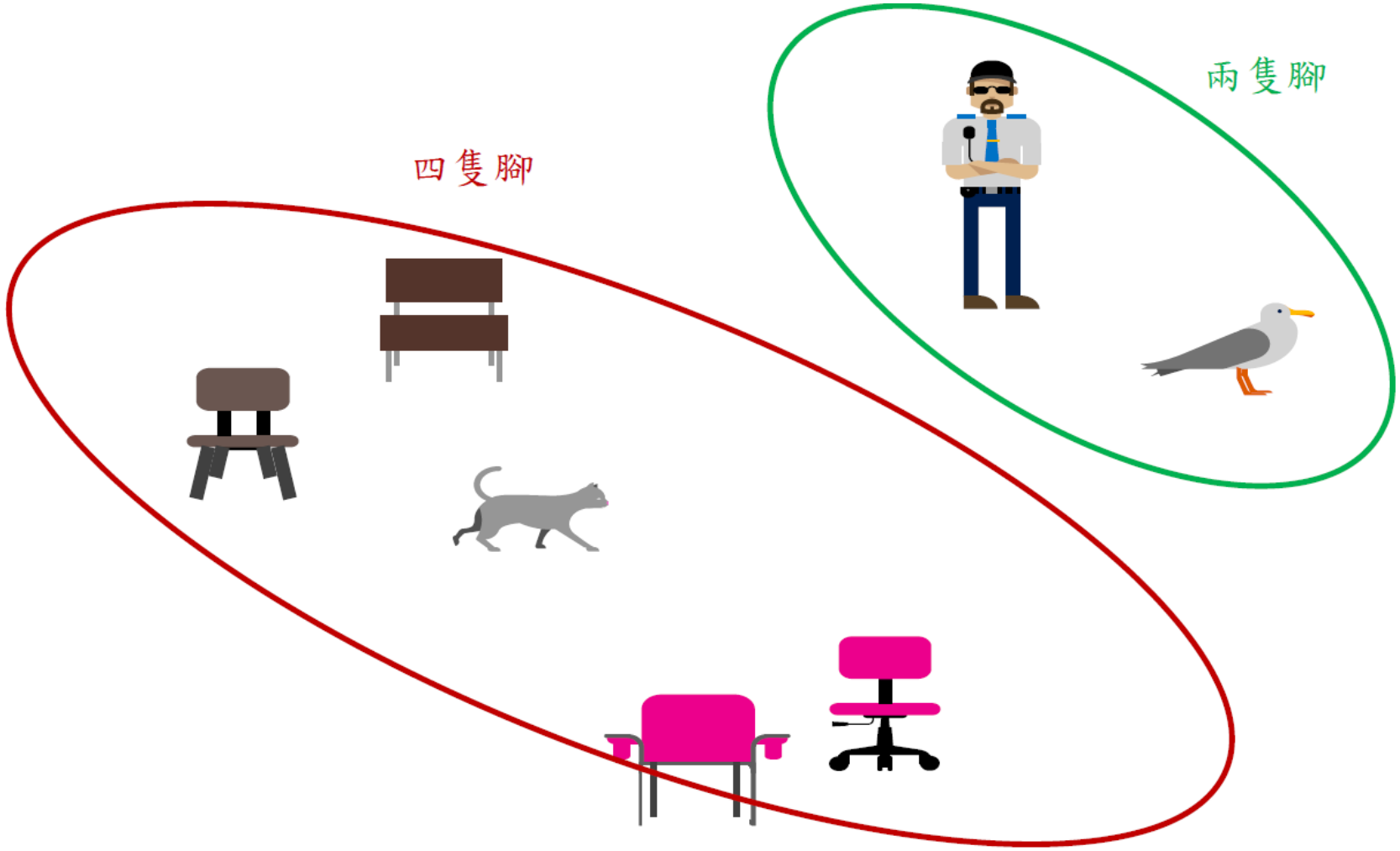
機器學習的運作模式—**過擬合現象**





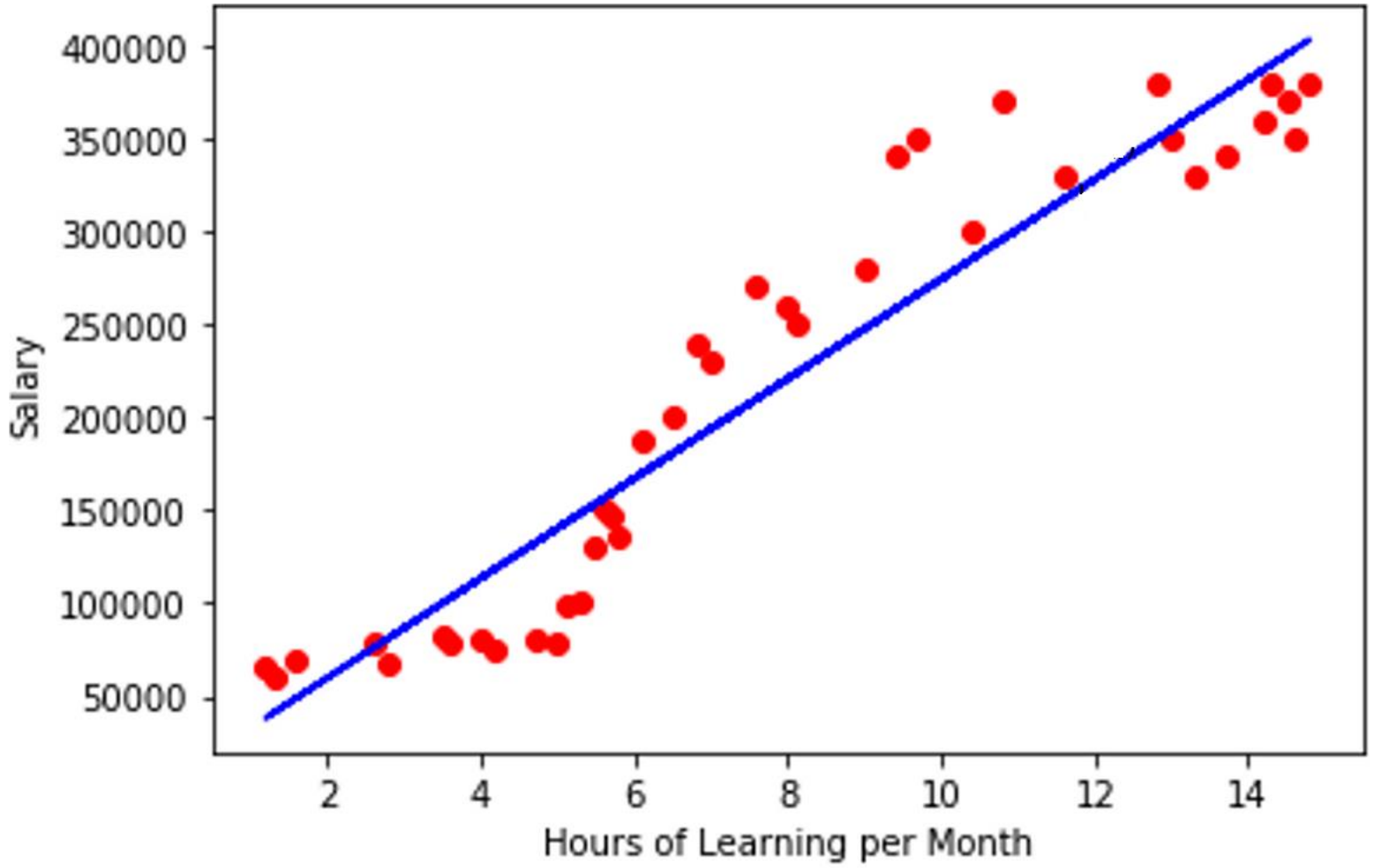
資料來源：引用自淡江大學張志勇教授

機器學習的運作模式—分群



資料來源：引用自淡江大學張志勇教授

機器學習的運作模式—迴歸分析



資料來源: https://www.bing.com/images/create

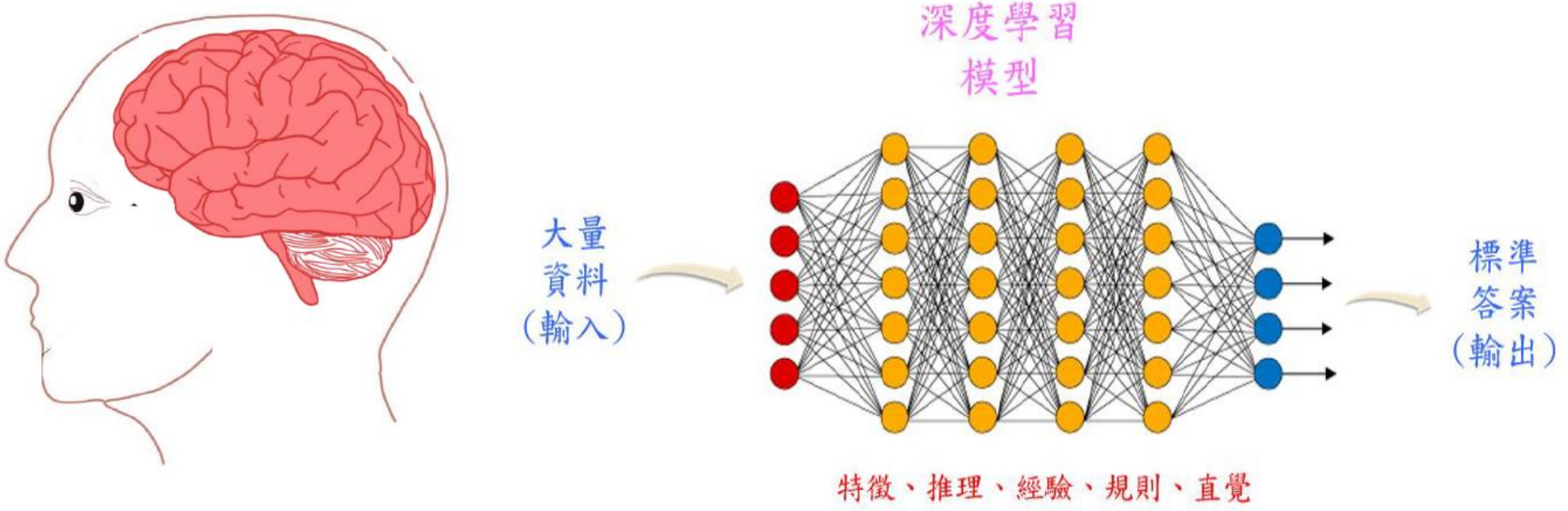
深度學習運作架構

- 深度學習是人工智慧中的一種方法，可指導電腦以受人腦啟發的方式來處理資料。深度學習模型可識別圖片、文字、聲音和其他資料的複雜模式，藉此產生更準確的預測。可以使用深度學習方法將通常需要人類智慧的任務自動化，例如描述影像或將聲音檔案轉錄為文字。
- 深度學習日常產品中使用，例如：
  - 數位助理
  - 聲控電視遙控器
  - 詐騙偵測
  - 自動臉部辨識
  - 自動駕駛

深度學習運作架構

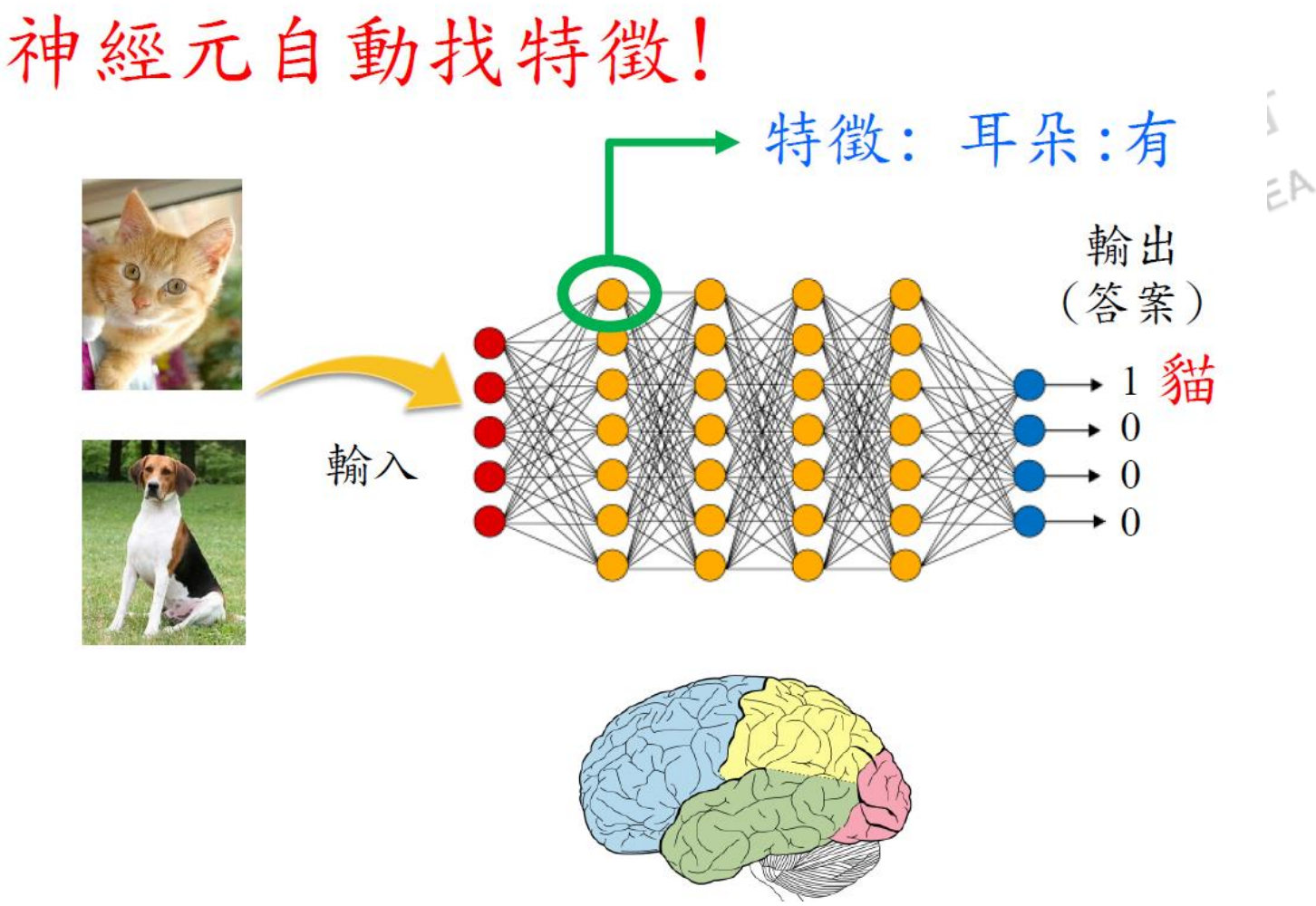
- 深度學習演算法是以人類大腦為模型的神經網路。例如，人腦包含數百萬個互連的神經元，這些神經元會協同合作以學習和處理資訊。同樣地，深度學習神經網路或人工神經網路，也包含了許多人工神經元層，其可在電腦內部協同運作。
- 人工神經元是稱為節點的軟體模組，會使用數學計算來處理資料。人工神經網路屬於深度學習演算法，可使用這些節點來解決複雜問題。

深度學習運作架構—**深度學習模仿人腦**



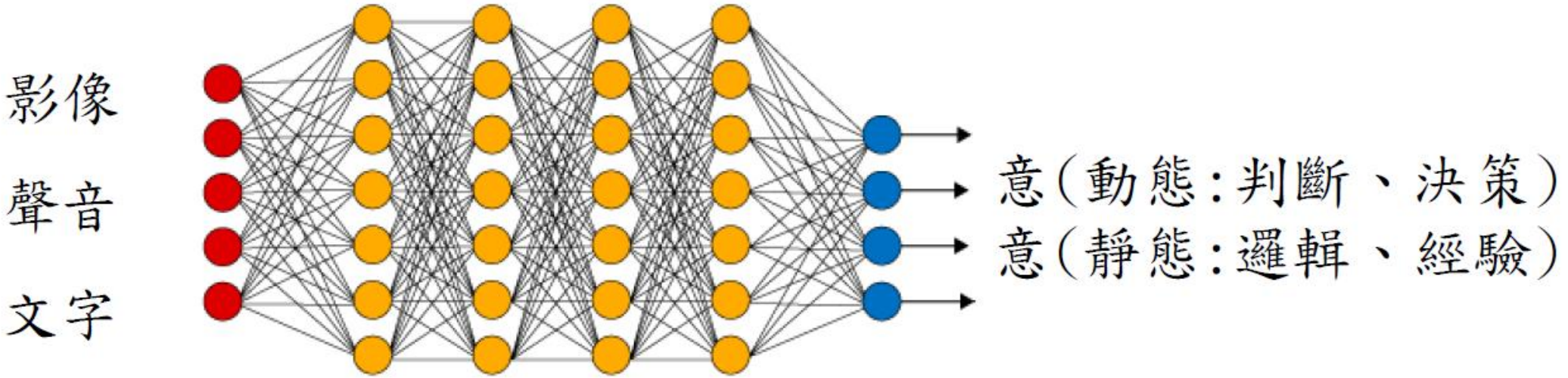
資料來源：引用自淡江大學張志勇教授

深度學習運作架構—輸入圖片，輸出答案(貓)



資料來源：引用自淡江大學張志勇教授

深度學習運作架構—輸入資料多樣化

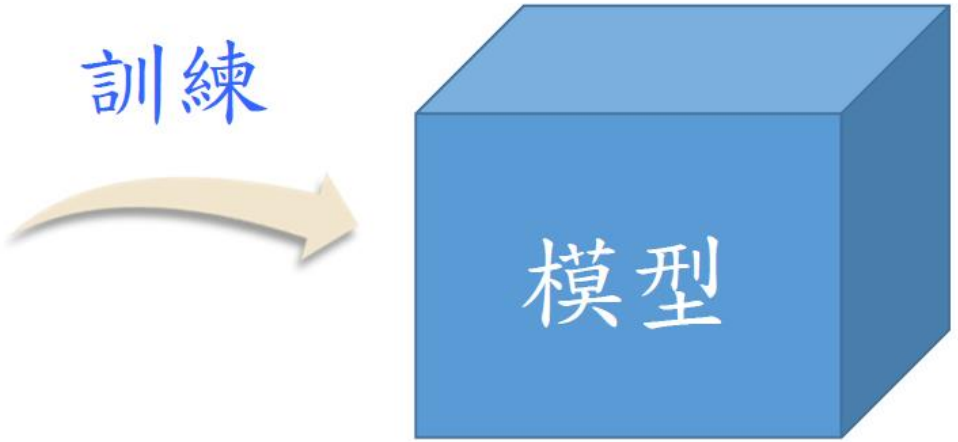


資料來源：引用自淡江大學張志勇教授

深度學習運作架構—**訓練模型**

過去的數據

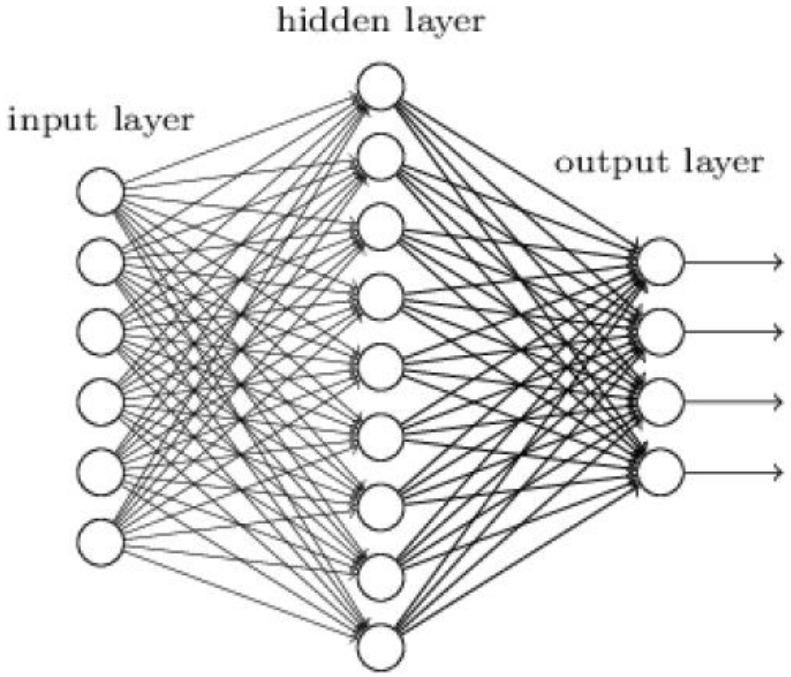
訓練



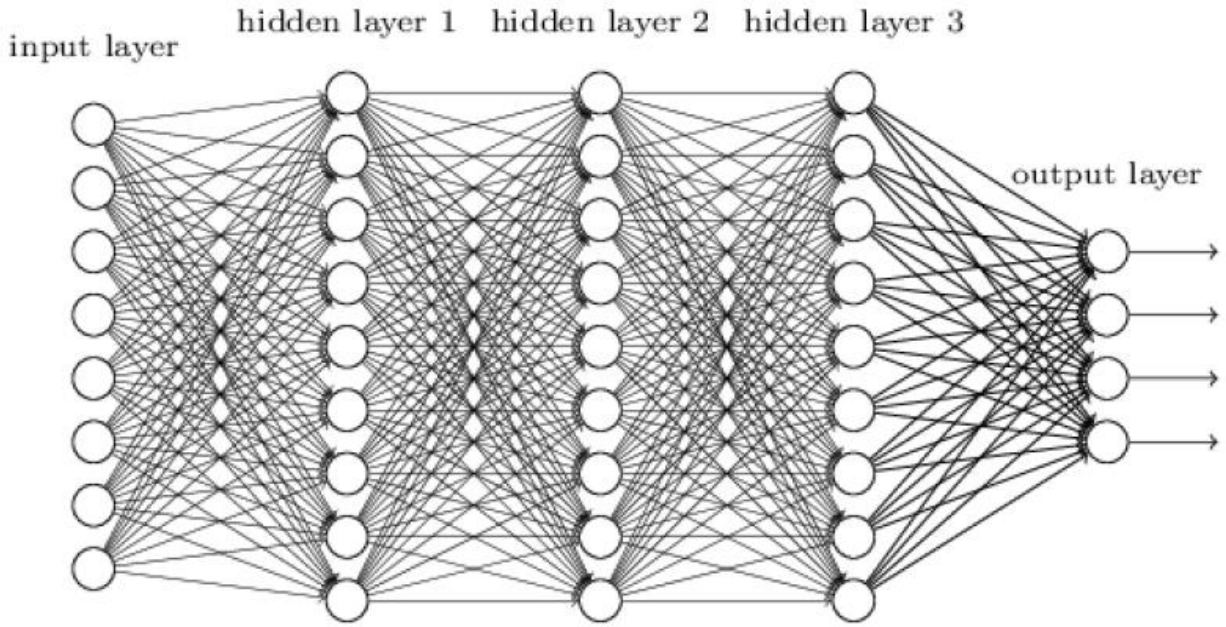
資料來源：引用自淡江大學張志勇教授

深度學習運作架構—深度學習神經網路

傳統神經網路
NN

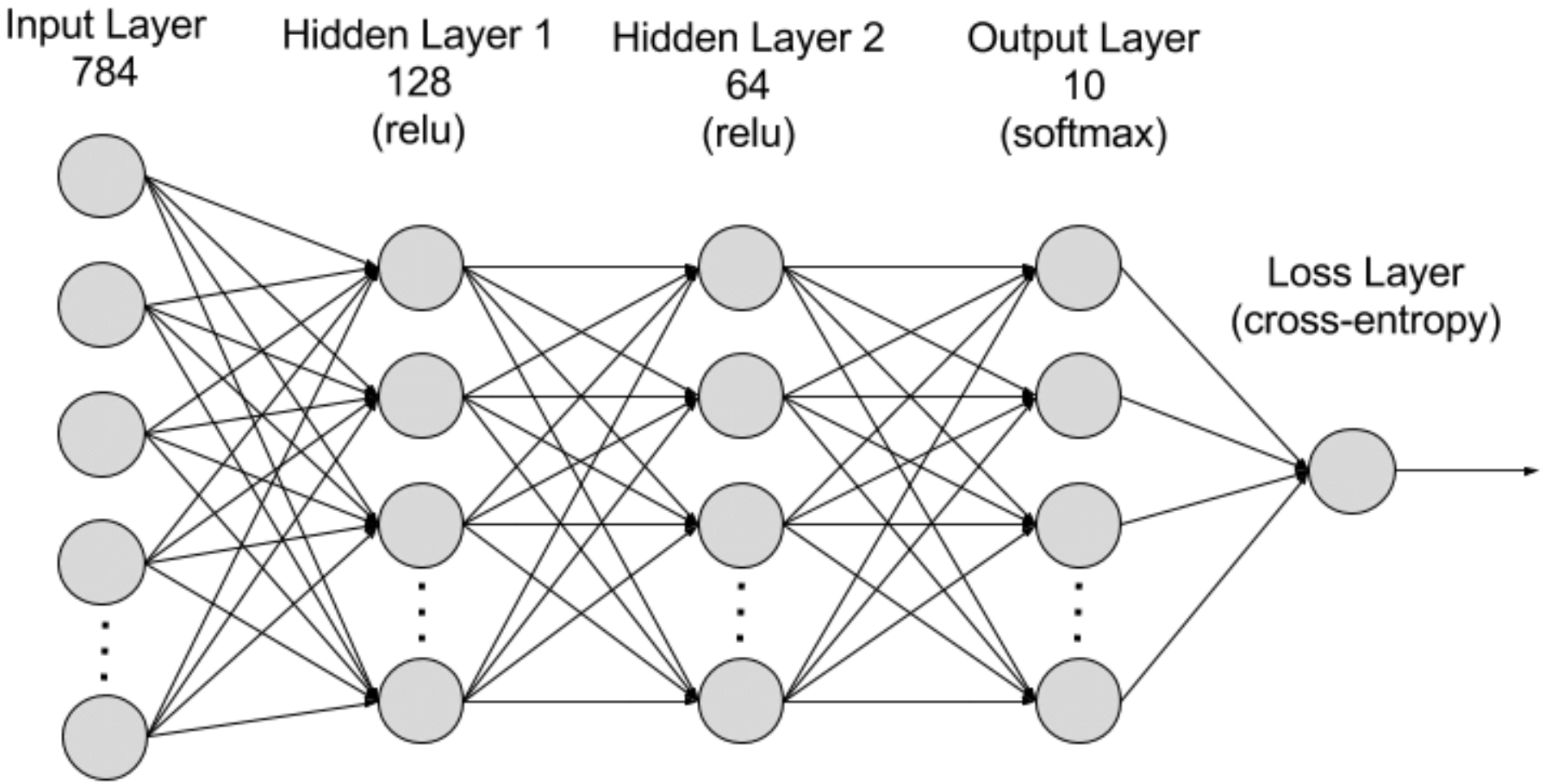


深度神經網路
DNN



資料來源：https://blog.openaimp.com/2020/09/deep-neural-network-dnn.html

深度學習運作架構



資料來源：引用自 https://aws.amazon.com/tw/what-is/deep-learning/

深度學習運作架構—深度學習網路元件

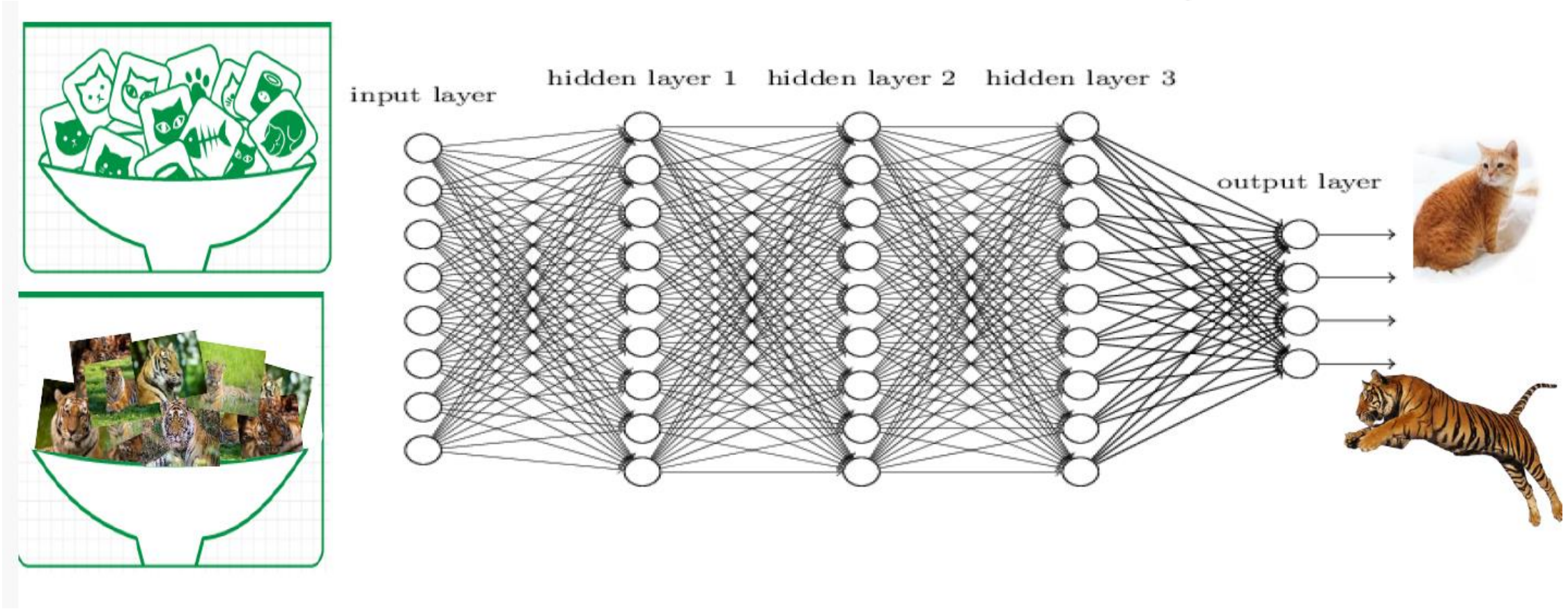
- **輸入層**
  - ➢可以有數個節點，將資料輸入到其中。

- **隱藏層**
  - ➢隱藏層會在不同層級處理資訊，在接收新資訊時調整它們的行為。深度學習網路有數百個隱藏圖層，可用來從幾個不同的角度分析問題。
  - ➢例如，如果您獲得了一張未知動物的影像，必須對其進行分類的，則可以將其與已經認識的動物進行比較。例如，您可以觀察它的眼睛和耳朵的形狀、大小、腿的數量以及其毛皮圖案。
  - ➢深度神經網路中的隱藏層會以相同的方式運作。每個隱藏層都會處理動物的不同特徵，並嘗試對其進行準確的分類。

深度學習運作架構—深度學習網路元件

- **輸出層**
  - ➢輸出層由輸出資料的節點組成。
  - ➢若只有「是」或「否」答案的深度學習模型，在輸出層中只有兩個節點。
  - ➢輸出範圍更廣的答案則具有更多節點。

深度學習運作架構—
找特徵（前幾層），建立分類規則（後幾層)



資料來源：引用自淡江大學張志勇教授

深度學習演算法介紹

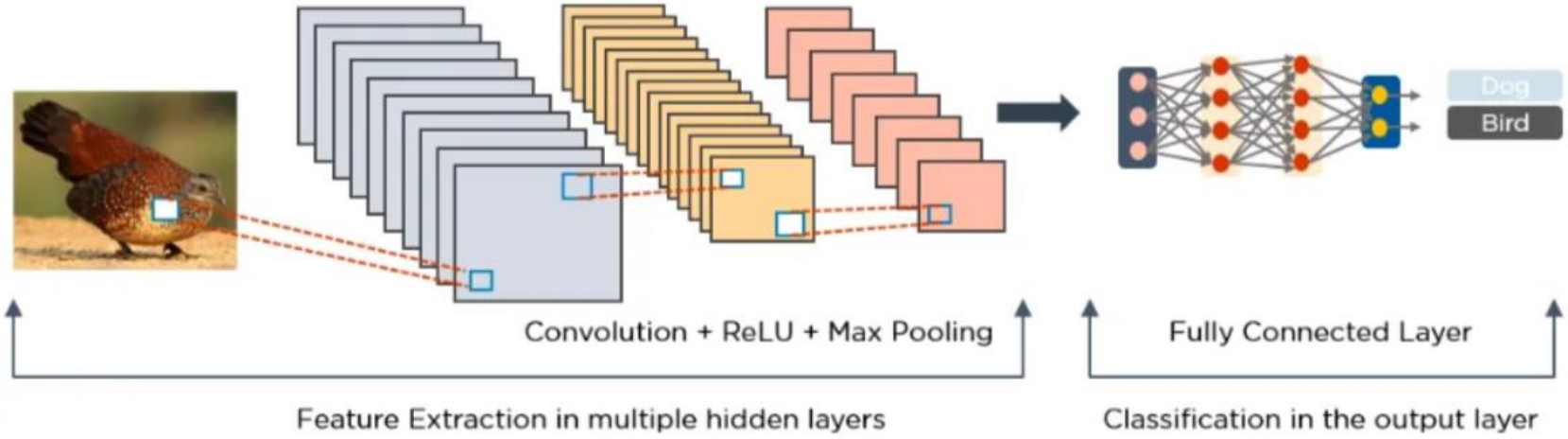
- 深度類神經網路(DNN)及相關知識(Activation Function、Loss Function等)
- 卷積類神經網路(CNN) 及相關知識
- 遞迴式類神經網路(RNN、LSTM)及相關知識
- 生成對抗網路(GAN)及相關知識

深度學習演算法介紹—類神經網路(DNN)

- **生物神經網路**
  - **兩種專家研究大腦功能**
    - ✓由下而上方式：生物神經學家(neurobiologists)，對單一神經細胞的刺激與反應，研究神經網路；
    - ✓由上而下：心理學家 (psychologists)從知覺與行為反應瞭解。
  - **人類神經系統**

深度學習演算法介紹—CNN

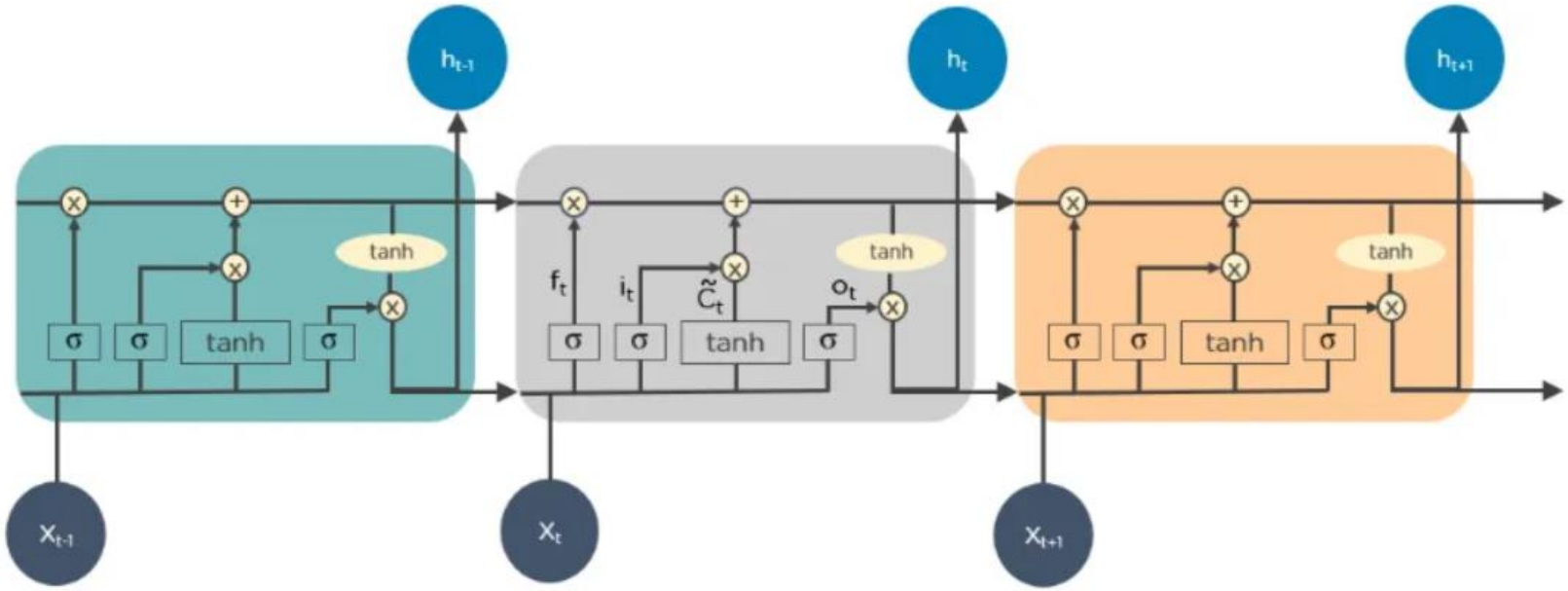
- Convolutional neural network(CNN)，由多層組成，主要用於圖像處理和物體檢測。
- CNN無法找連續圖片的特徵。



資料來源：引用自https://kilong31442.medium.com/top-10-%https://top-10-您應該要學會的深度學習演算法-fundamental-review-series-d8c69897e010

深度學習演算法介紹—長短期記憶網路 (LSTM)

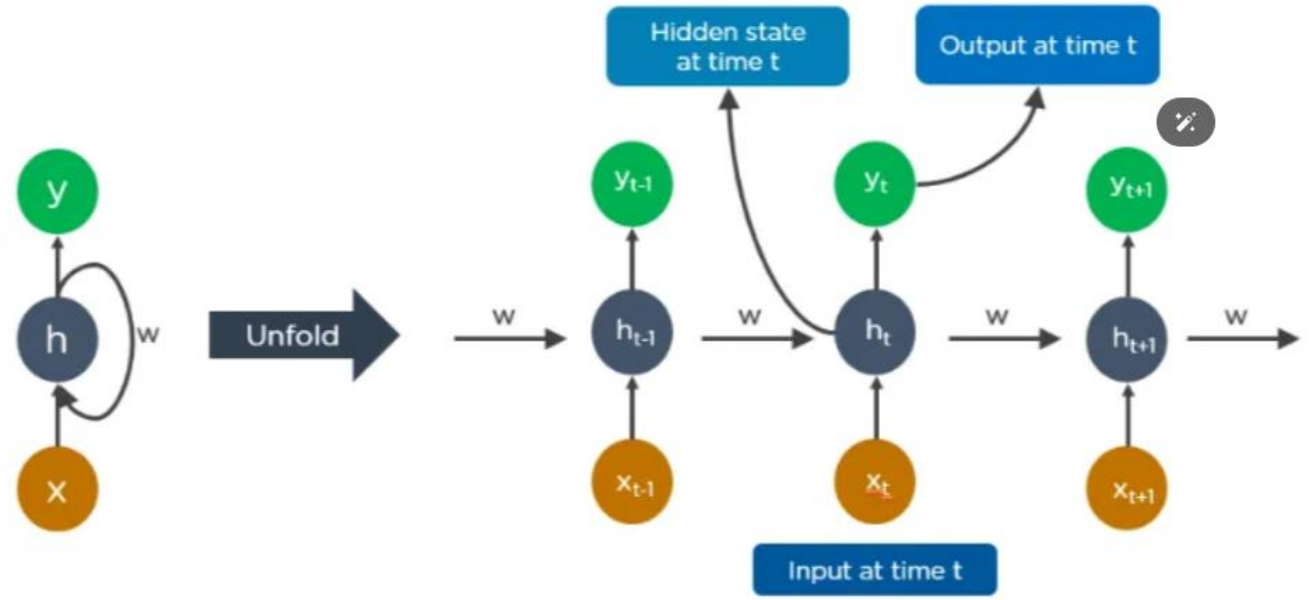
- LSTM是一種遞迴神經網路（RNN），可以學習和記憶長期依賴關係。長時間回憶過去的資訊是預設行為。



資料來源：引用自https://kilong31442.medium.com/top-10-%https://top-10-您應該要學會的深度學習演算法-fundamental-review-series-d8c69897e010

深度學習演算法介紹—遞迴神經網路(RNN)

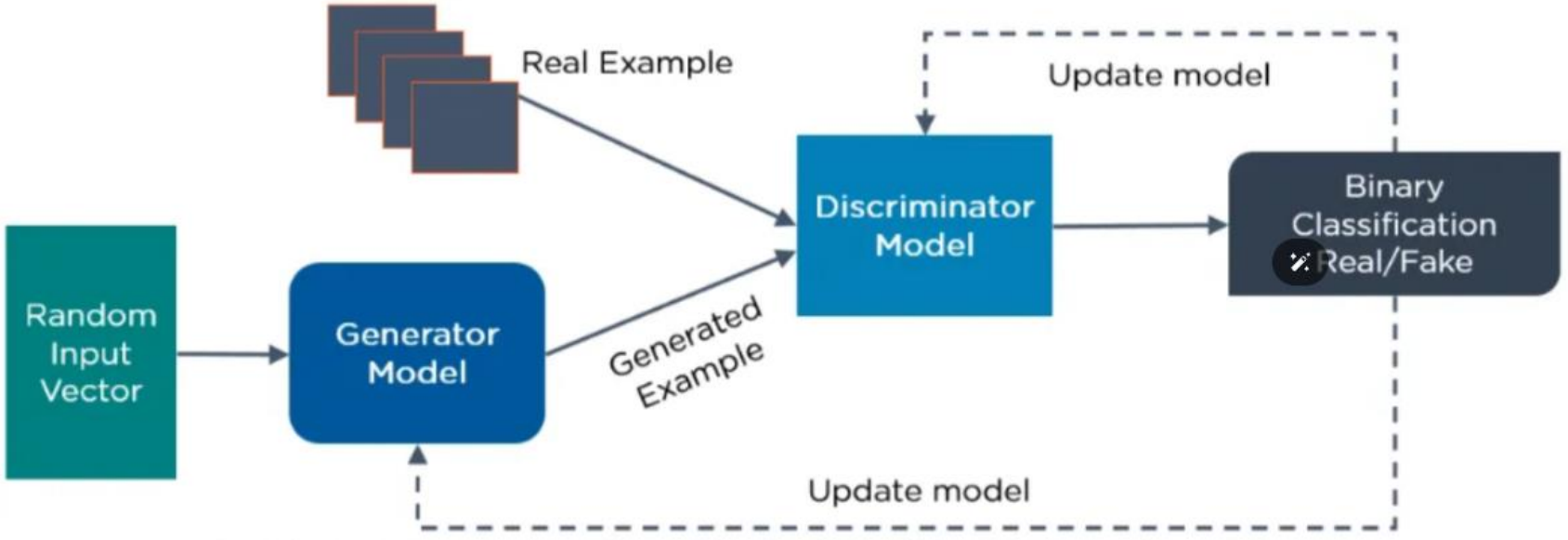
- RNN具有形成定向迴圈的連接，允許 LSTM 的輸出作為輸入饋送到當前相位。



資料來源：引用自https://kilong31442.medium.com/top-10-%https://top-10-您應該要學會的深度學習演算法-fundamental-review-series-d8c69897e010

深度學習演算法介紹—生成對抗網路 (GAN)

- GAN 是生成式深度學習演算法，用於創建類似於訓練數據的新資料實例。
- GAN有兩個元件：一個生成器，它學習生成假數據，另一個鑑別器，它從虛假信息中學習。



資料來源：引用自https://kilong31442.medium.com/top-10-%https://top-10-您應該要學會的深度學習演算法-fundamental-review-series-d8c69897e010

深度學習演算法介紹—TensorFlow

- TensorFlow 是一個由 Google Brain 團隊內部開發的第一個深度學習系統，而後逐漸演變成為一個開源的機器學習框架。

- Tensorflow是一個框架(framework)，主要是用來做深度學習的各種演算法。深度學習是以矩陣運算的方式來模擬神經訊息的傳送。而所謂的深度學習，指的就是神經網路，機器學習的一種，深度學習演算法等同於機器學習的演算法。

- 提供了豐富的工具和庫，支持各種機器學習算法和模型的實現，包括深度學習模型。

- 特點包括靈活性、可擴展性和支持分佈式計算，使其成為研究人員和開發人員在機器學習和人工智慧領域中的首選框架之一。

深度學習演算法介紹—TensorFlow發展歷程

- 2011年：Google Brain 團隊開發了第一個內部深度學習框架 DistBelief，用於大規模分布式訓練神經網絡。
- 2015年11月：Google 在2015年11月推出了 TensorFlow 的第一個公開版本，作為 Google Brain 的後繼者，並將其開源。這使得更多的研究人員和開發者可以使用和貢獻 TensorFlow。
- 2016年2月：TensorFlow 1.0 版本正式發布，標誌著它成為一個成熟且穩定的機器學習框架。
- 2017年9月：Google 在 TensorFlow Dev Summit 上宣布推出 TensorFlow 1.4 版本，引入了一些新的功能和改進，包括對 TensorFlow Lite 的支持，使其能夠在移動和嵌入式設備上運行。
- 2019年3月：TensorFlow 2.0 正式發布。TensorFlow 2.0 的目標是改進用戶體驗，簡化 API，提高執行效率，並更好地支持 Pythonic 編程風格。
- 2020年5月：TensorFlow Extended (TFX) 正式成為 TensorFlow 生態系統的一部分，用於構建端到端的機器學習管道。
- TensorFlow 在持續發展和改進中，積極擴展其功能，並增強對不同硬體和平台的支持，如GPU、TPU等，以滿足不斷增長的機器學習和深度學習應用需求。

深度學習演算法介紹—TensorFlow 2.0

- **簡化的 API**
  - 引入了 Keras 作為 TensorFlow 的高階 API，使得定義、訓練和評估神經網絡模型更加簡單和直觀。Keras 提供了易於使用的模型構建塊，使得用戶可以輕鬆地創建各種類型的神經網絡，包括序貫模型和功能 API 模型。
- **動態圖**
  - TensorFlow 2.0 引入了即時執行模式（eager execution），使得模型開發和調試過程更加直觀和互動式。這樣的設計使得用戶可以像使用 NumPy 一樣運行操作，即時查看結果。
- **Keras 支持**：
  - TensorFlow 2.0 將 Keras 完全整合為其主要的模型構建和訓練 API。這意味著所有 TensorFlow 2.0 的高階 API 都基於 Keras，從而統一了模型的定義和訓練流程。

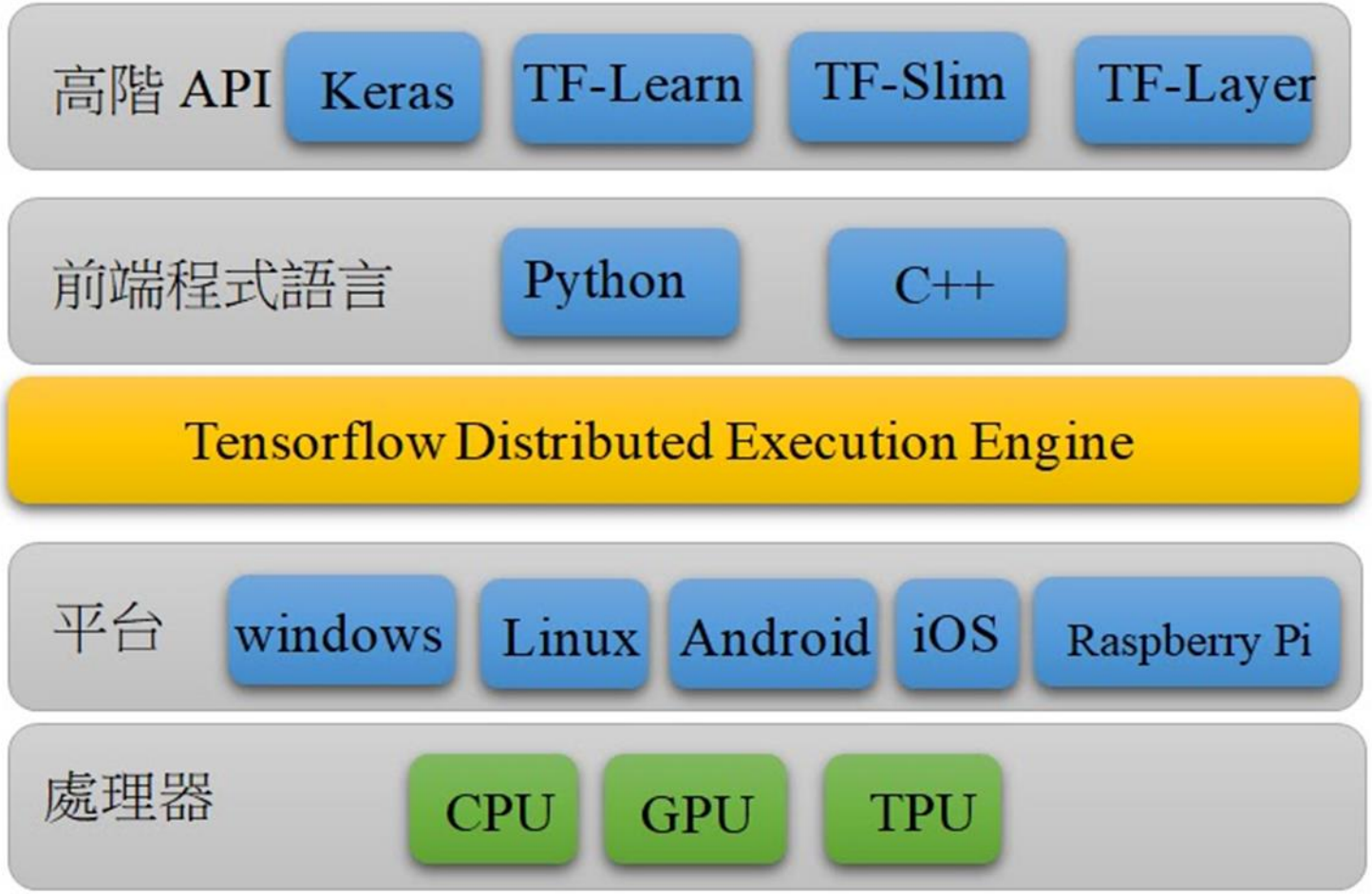
深度學習演算法介紹—TensorFlow 2.0

- **性能提升**：
  - TensorFlow 2.0 通過諸如自動圖優化、GPU 和 TPU 加速、改進的分佈式訓練等功能，大幅提升了模型執行的性能和效率。
- **兼容性**：
  - TensorFlow 2.0 支持舊版本 TensorFlow 1.x 的程式，並提供了向後兼容性。這使得現有的 TensorFlow 1.x 用戶可以升級到新版本，同時還能利用新的功能。
- **模型部署**：
  - TensorFlow 2.0 引入了 TensorFlow Serving 和 TensorFlow Lite 等工具，用於簡化模型的部署和在不同平台上的運行，如雲端、移動設備等。
- 加強了與 Python 生態系統的集成，使得開發者能夠更輕鬆地建立和訓練機器學習模型。

深度學習演算法介紹—Keras

- Keras是一個用Python撰寫的OpenSouce神經網路Library，高階整合的Deeplearning-toolkit，其整合TensorFlow與Theano tool kit，讓使用者能透過Keras API來達到活用。
- Kera是一個model-level的深度學習程式庫，Keras處理模型(model)的建立(create)、訓練(prediction)等功能。
- 主要的開發人員為Google工程師，以MIT開放原始碼授權。

深度學習演算法介紹—TensorFlow架構圖



資料來源：引用自 Tensorflow+keras深度學習人工智慧實務

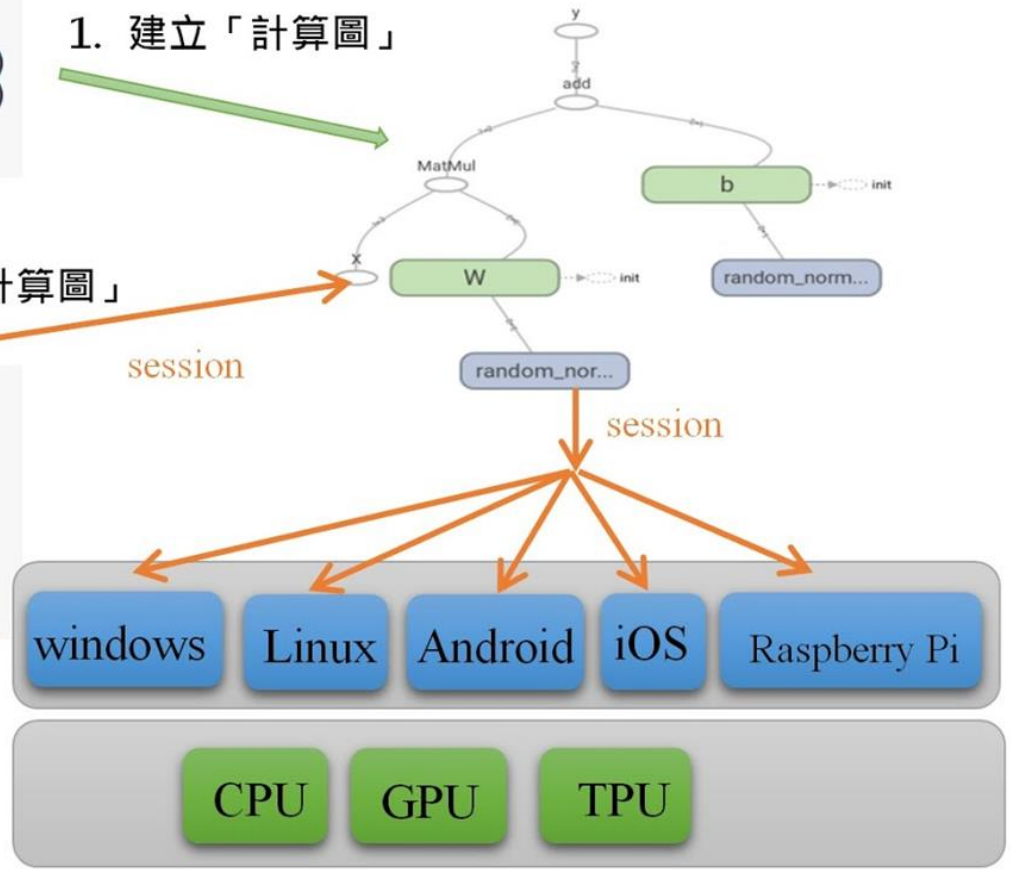
深度學習演算法介紹—TensorFlow模式

```python
import tensorflow as tf
import numpy as np
W = tf.Variable(tf.random_normal([3, 2]),name='W')
b = tf.Variable(tf.random_normal([1, 2]),name='b')
X = tf.placeholder("float", [None,3],name='X')
y=tf.nn.sigmoid(tf.matmul(X,W)+b,'y')
```

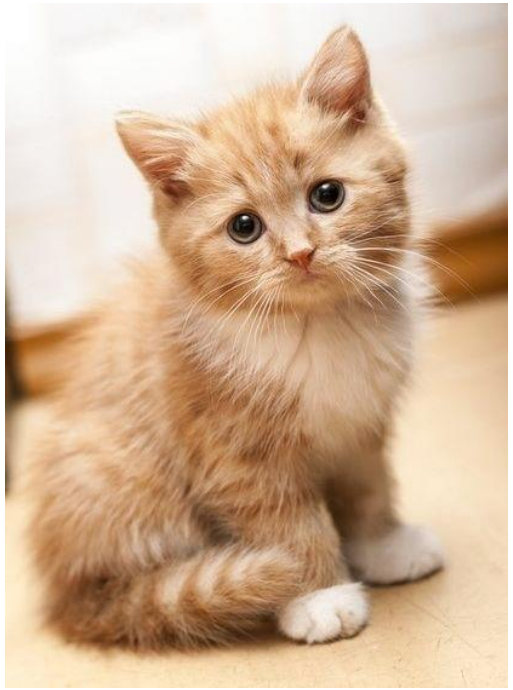
1. 建立「計算圖」

2.執行「計算圖」

```python
with tf.Session() as sess:
    init = tf.global_variables_initializer()
    sess.run(init)
    X_array = np.array([[0.4,0.2 ,0.4],
                        [0.3,0.4 ,0.5],
                        [0.3,-0.4,0.5]])
    (_b,_W,_X,_y)=sess.run((b,W,X,y),
                           feed_dict={X:X_array})
```



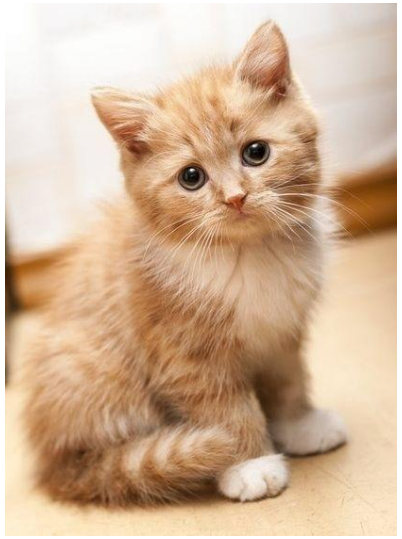
深度學習演算法介紹—TensorFlow示意



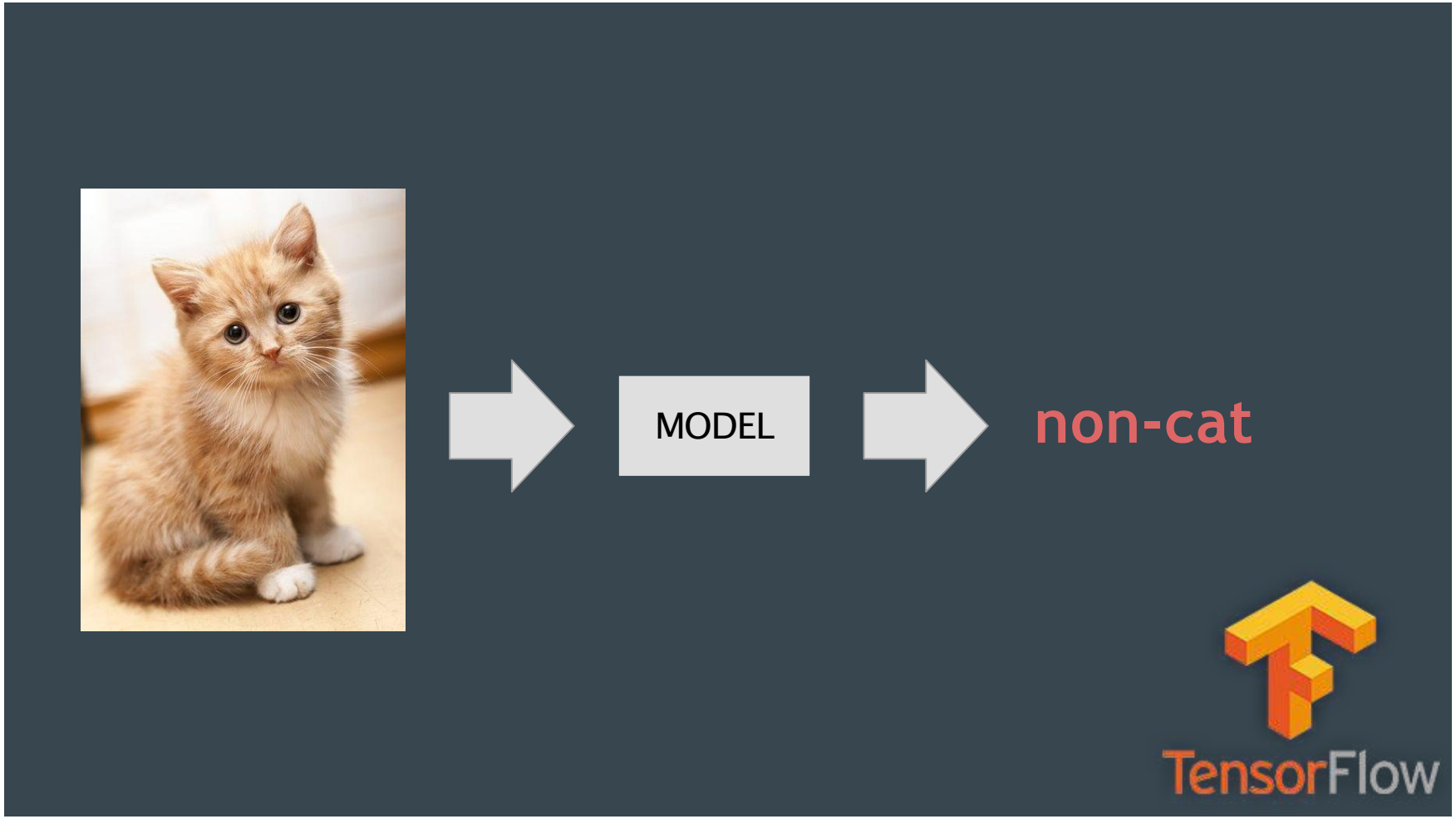
cat

TensorFlow

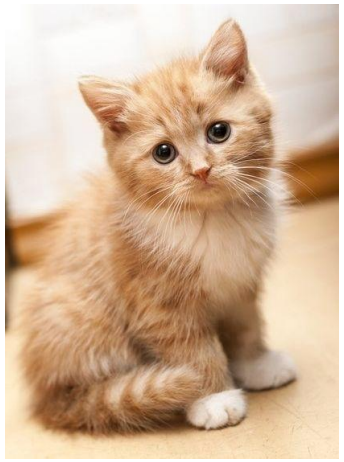
深度學習演算法介紹—TensorFlow示意



深度學習演算法介紹—TensorFlow示意



深度學習演算法介紹—TensorFlow示意



MODEL

prediction

non-cat

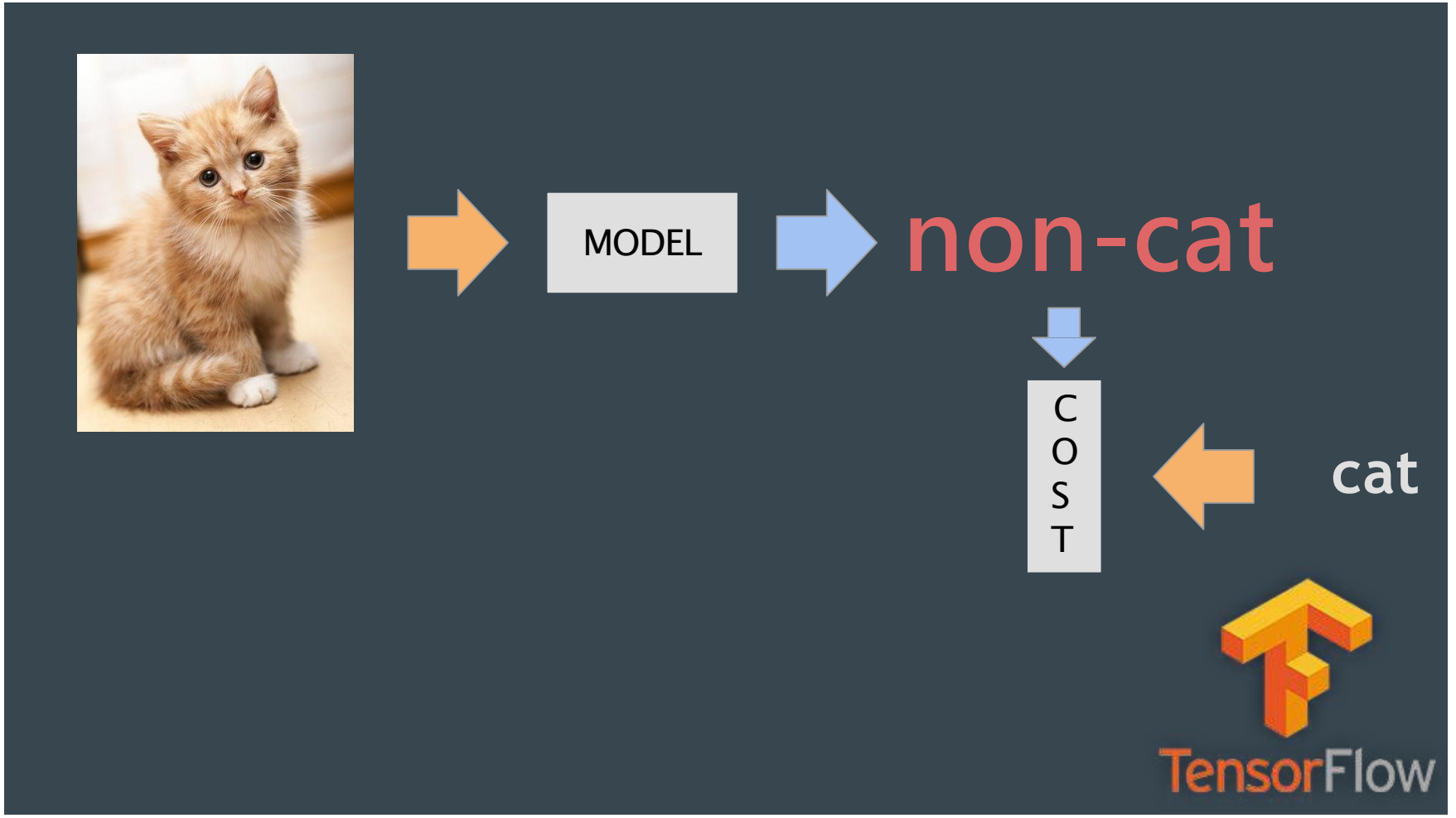
target

cat

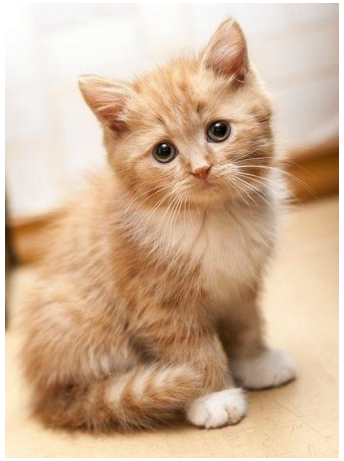
input

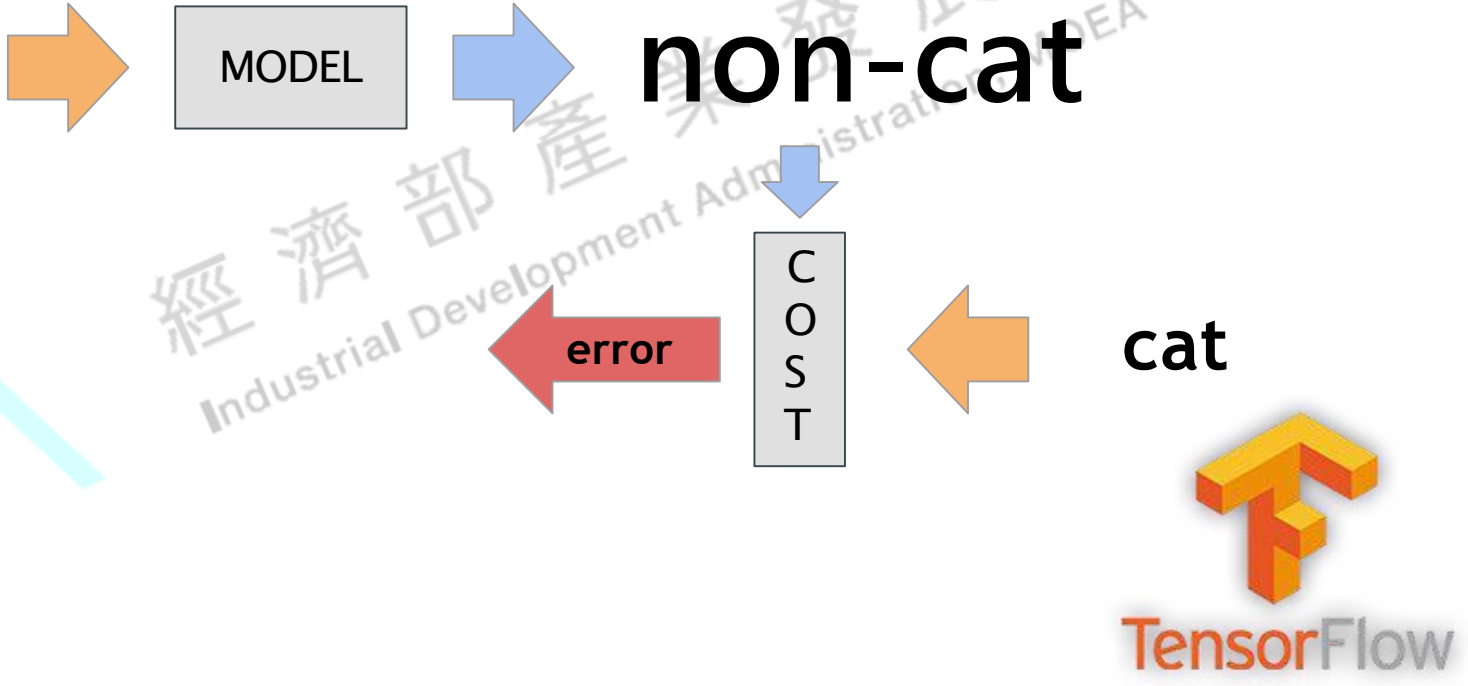
TensorFlow

深度學習演算法介紹—TensorFlow示意

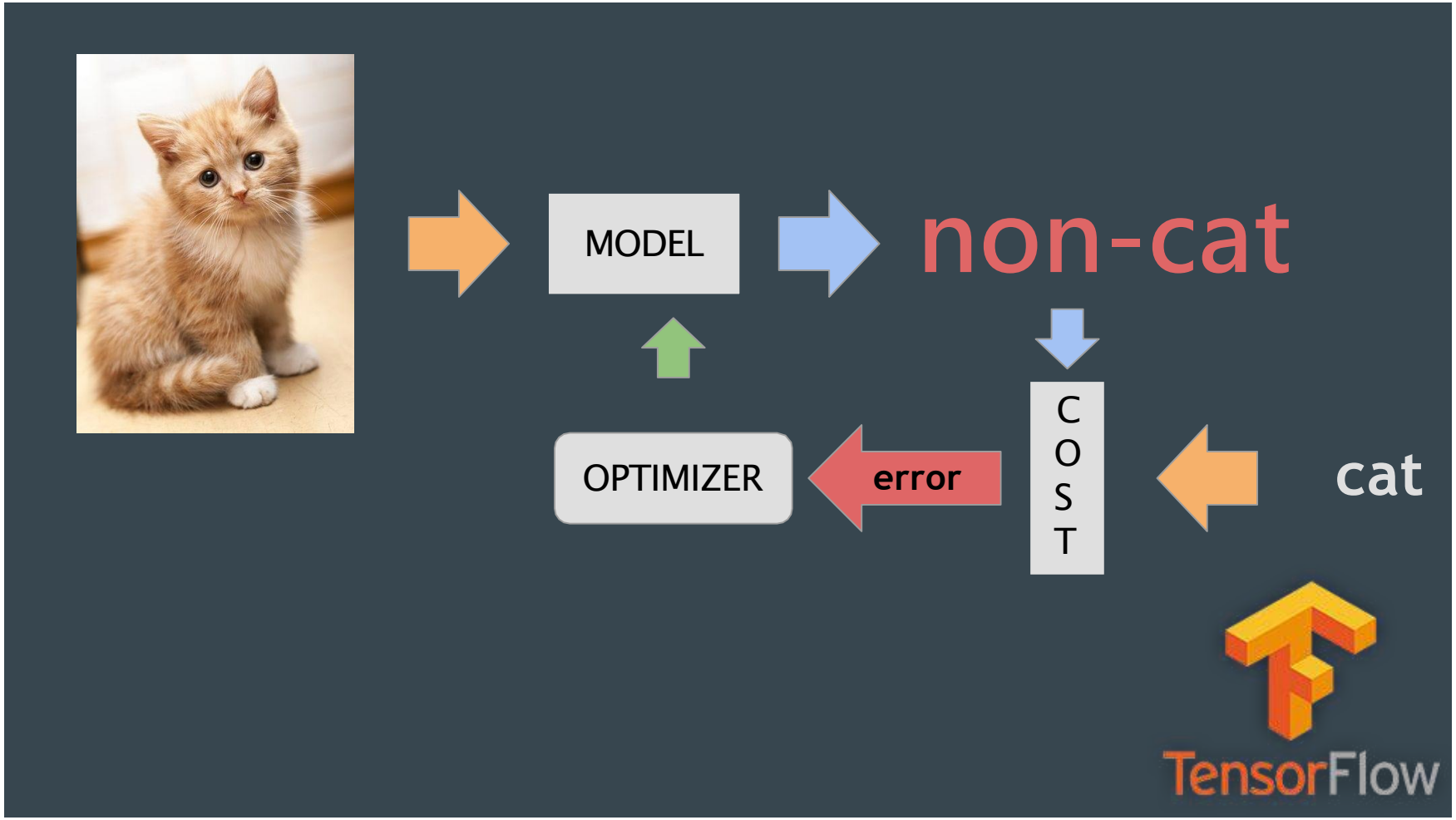


深度學習演算法介紹—TensorFlow示意

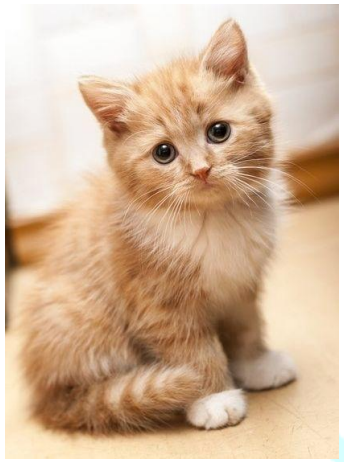


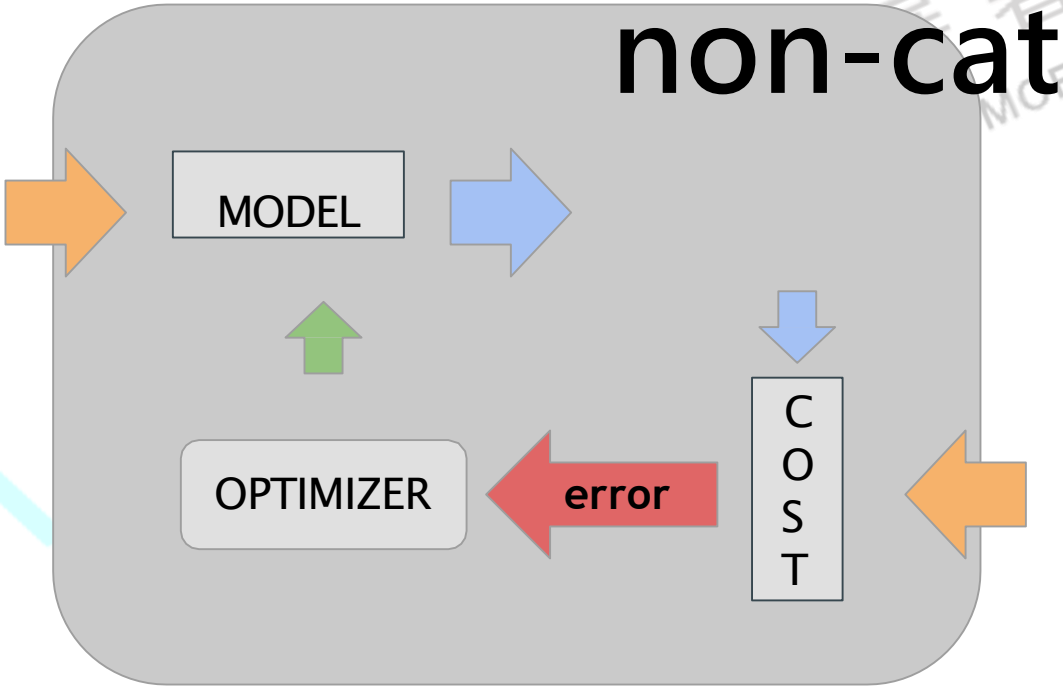


深度學習演算法介紹—TensorFlow示意



深度學習演算法介紹—TensorFlow示意

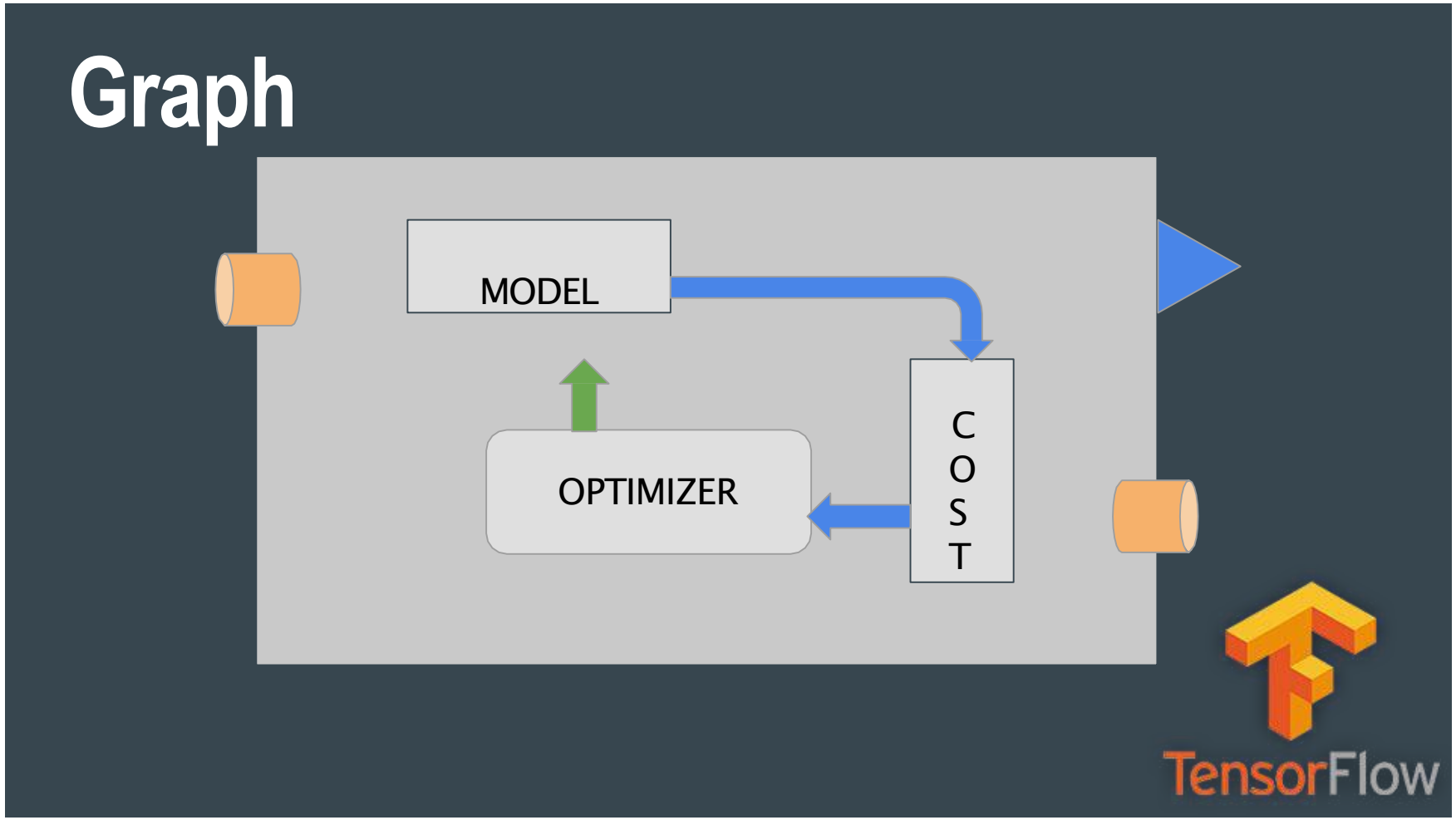




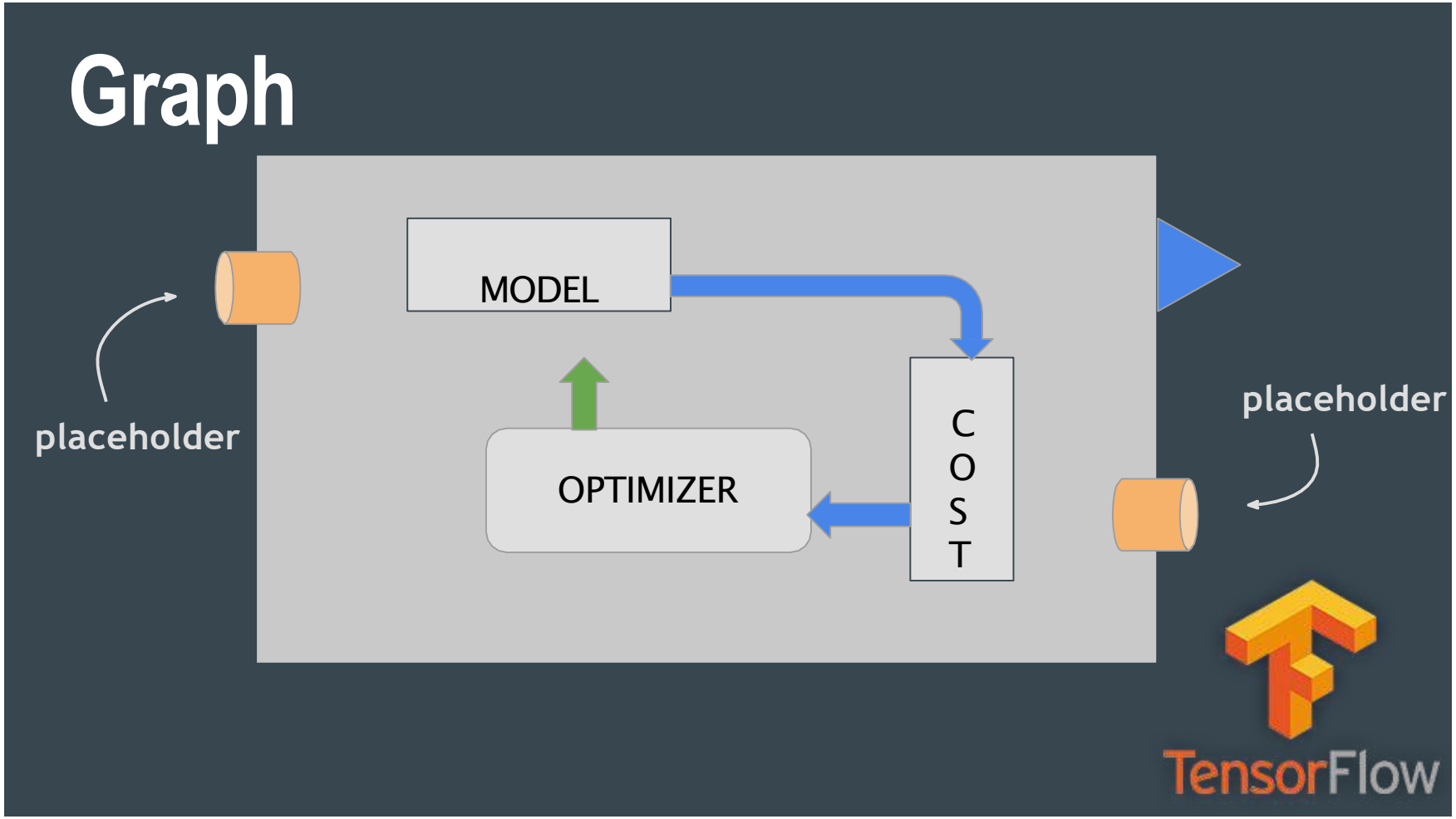
Graph

TensorFlow

深度學習演算法介紹—TensorFlow示意

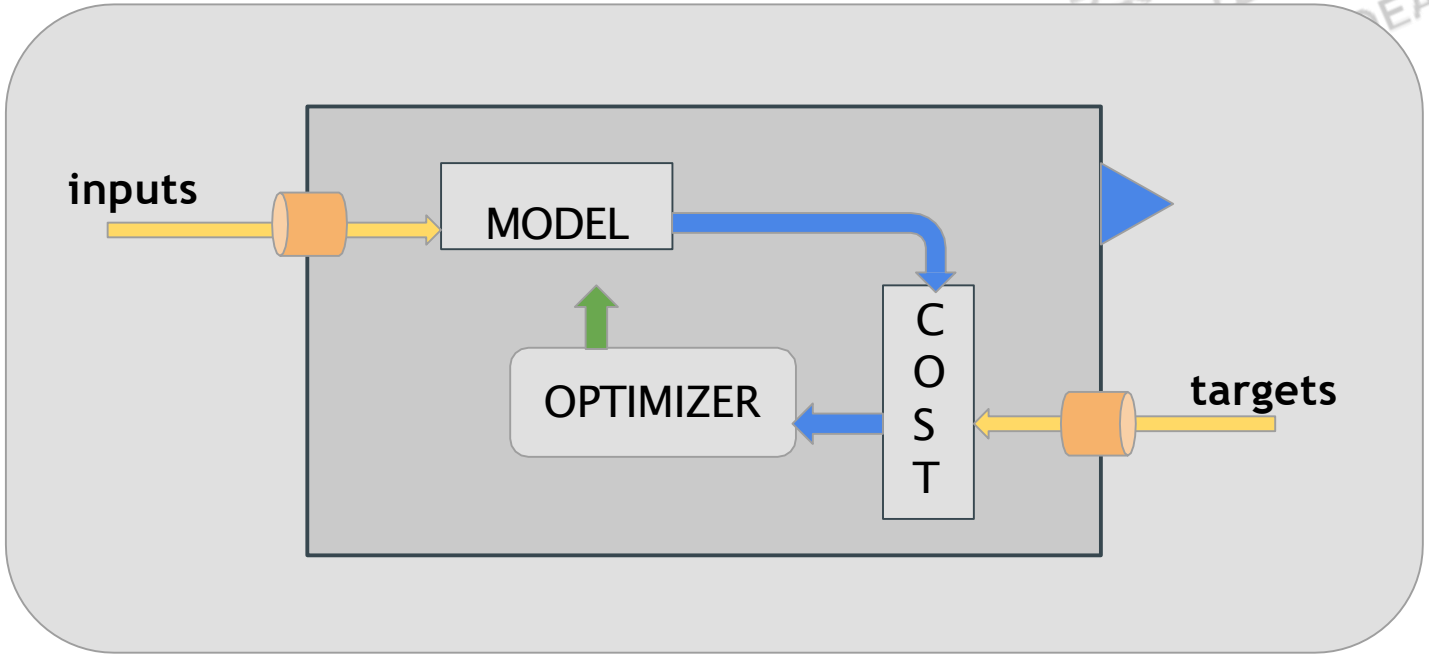


深度學習演算法介紹—TensorFlow示意



深度學習演算法介紹—TensorFlow示意

Session: Graph + Data



TensorFlow

深度學習演算法介紹—TensorFlow示意

用 Tensorflow 打造一個神經網路！

3 .延伸閱讀與思維創新

- **延伸閱讀**
  - Introduction to Machine Learning, second edition, Ethem ALPAYDIN，MIT Press，出版日期：2010，ISBN：978026201243-0。
  - 吳作樂、吳秉翰 (2020). 圖解機器學習、人工智慧與人類未來，五南圖書，出版日期 2022-4-1， ISBN：978957763903-5。

- **思維創新**
  - 工作與生活環境中，有那些機器學習方法的應用案例?
  - 說說你曾經用過哪機器學習的方法，用在那些情境?
  - 想想未來有哪些情境可以使用機器學習來解決問題?

※機器學習技術理論與應用演練I
—解析機器學習核心概念

目錄

1. 課程目標與先備知識

- 在課程之前，建議宜具備的知識與經驗
  - ➢計算機概論
  - ➢人工智慧概論

- 教學目標：
  - ➢瞭解資料標籤的的重要性
  - ➢瞭解資料規劃原則
  - ➢瞭解資料預處理方法
  - ➢培養具有訓練模型所需資料的收集、處理、標記與規劃的能力

掌握機器學習資料集中的特徵和標籤

- **特徵的重要性**

  在機器學習領域中，特徵是作為系統輸入的獨立變量。模型在進行預測時會利用這些特徵。通過特徵工程的過程，還可以從舊特徵中獲得新特徵。簡單來說，資料集中的一列可以視為一個特徵，也稱為"變量"或"屬性"。特徵的數量越多，維度就越高。根據分析目的的不同，資料集中包含的特徵也會有很大差異。

- **特徵定義**

  - 特徵是機器學習模型的輸入變量，代表資料的屬性或特性。
  - 特徵是資料集代表資料的基本性質，反映原始資料的關鍵資訊。
  - 選擇合適特徵需要深入理解具體的所欲解決問題內涵及目標。

掌握機器學習資料集中的特徵和標籤—特徵工程

- 特徵工程是利用領域知識建構所需特徵的過程。
- 目的是使機器學習模型能夠更好地訓練與工作。
- 如果特徵選取得當，有助於模型有效萃取隱藏在資料中的規律，進而提高模型的預測能力。
- 特徵選取過程需要深入理解問題，是機器學習方法中最具挑戰性和創造性的部分之一。

掌握機器學習資料集中的特徵和標籤—特徵選取

- 特徵的選取可依照下列四原則
  - ➢理解資料的內容
  - ➢選擇部分資料或是可衍生資料作為資料特徵
  - ➢評估特徵對模型性能的影響
  - ➢根據評估結果不斷改進特徵

掌握機器學習資料集中的特徵和標籤
—特徵對機器學習的影響

- **好的特徵**
  - ➢提高模型準確性
  - ➢減少訓練時間
  - ➢增強模型解釋性

- **特徵選擇的重要性**
  - ➢識別最相關特徵
  - ➢降低模型複雜度
  - ➢提高計算效率

掌握機器學習資料集中的特徵和標籤—如何建立學習模型

- **收集資料**
  - ➢從各種來源獲取原始資料，確保資料的完整性和代表性。
- **清理資料**
  - ➢處理缺失值、異常值，確保資料質量和一致性。
- **特徵選取**
  - ➢選取現有資料作為特徵或是轉換現有特徵來衍生新的特徵。
- **定義模型**
  - ➢選擇適合問題的機器學習算法和定義模型架構。
- **訓練與測試**
  - ➢使用訓練集資料訓練模型，並在測試集上評估性能。

掌握機器學習資料集中的特徵和標籤
—特徵選取的技巧與方法

技術	描述	適用場景
特徵縮放	將特徵值調整到相同範圍	距離基礎的算法
特徵編碼	將分類變量轉換為數值	處理非數值特徵
特徵組合	創建新的複合特徵	捕捉特徵間交互
特徵分箱	將連續變量分為離散區間	處理非線性關係

掌握機器學習資料集中的特徵和標籤—特徵選取的注意事項

- **從問題內涵出發**
  - ➢將特徵選取與具體問題內涵和解題目標聯繫起來，確保所選取的特徵具有實際意義和價值也保有AI可解釋性。
- **持續實證與修正**
  - ➢採用持續修正的方式，不斷測試和改進特徵。並透過交叉驗證來實證所選特徵的效果，同時避免過擬合。
- **保持簡單性**
  - ➢優先考慮簡單、可解釋的特徵。複雜的特徵可能導致模型難以理解和維護。
- **版本控制**
  - ➢ 詳細記錄特徵工程過程，包括嘗試過的方法和結果。使用版本控制來跟踪特徵的演變。

掌握機器學習資料集中的特徵和標籤—特徵與標籤

- 在機器學習領域中，"特徵"和"標籤"是監督式學習模型的基礎概念
- 特徵用於監督式學習與非監督式學習方法，作為分類或預測的依據，
  - ➢特徵具有可測量性，是可以量化和記錄的屬性。
  - ➢特徵可以是數值或文本，是模型用來進行預測的輸入變量。
- 標籤只見於監督式學習方法中，是模型產出的結果
  - ➢在監督式學習中，標籤是已知的結果，模型在訓練過程中學習將其與輸入特徵相關聯。
- 特徵和標籤之間的關係是模型訓練的核心

掌握機器學習資料集中的特徵和標籤—標籤範例

- **房價預測模型**
  - ➢標籤將是房屋的實際價格。
- **垃圾郵件分類器**
  - ➢標籤將是郵件是否為垃圾郵件。
- **預測疾病結果**
  - ➢標籤是患者是否有糖尿病。
- **金融信用評分**
  - ➢標籤是信用評級或信用度。
- **文本情感分析**
  - ➢文本中的情感極性（正面、負面或中性）。

機器學習中使用訓練和驗證資料

- 資料集對於機器學習系統至為重要。
- 資料品質直接影響模型訓練的成效。
- 資料集是一組二維陣列，由屬性欄位與資料實例所構成。

機器學習中使用訓練和驗證資料—資料集分割與模型訓練

- **訓練數據集**
  - ➢最大的部分,用於訓練模型。通常佔總數據的70-80%。
  - ➢機器學習中使用訓練和驗證資料
- **測試數據集**
  - ➢用於評估模型的無偏性能。模型預測與實際標籤進行比較。通常佔15%。
- **驗證數據集**
  - ➢用於調整和優化最終模型。根據驗證集的性能指標調整模型的超參數。通常佔10-15%。

機器學習中使用訓練和驗證資料—資料集分割的注意事項

- **原始數據分割**
  - ➢分割操作應在原始數據上進行,後續處理應分別應用於每個子集。
- **取樣隨機化**
  - ➢在分割前對數據進行隨機取樣,確保數據在各個子集中的隨機分布。
- **分布一致性**
  - ➢驗證集和測試集應具有相似的分布,以確保評估的一致性。
- **重複交叉驗證**
  - ➢重複隨機執行上述的步驟，觀察所得結果的變異程度。

機器學習中使用訓練和驗證資料—資料收集的注意事項 (1/2)

- **數據取得**
  - ➢是否已有必要的數據?數據是否已存在或需要收集?
- **法律和道德考慮**
  - ➢數據是否受版權保護?是否存在道德問題?是否符合個人數據法律或需要匿名化?
- **可擴展性**
  - ➢能多頻繁地生成新數據?模型需要多少數據才能產生有意義的結果?數據集的最佳大小是多少?
- **可用性**
  - ➢數據品質如何?是否存在缺值、異常值或不一致?

機器學習中使用訓練和驗證資料—資料收集的注意事項 (2/2)

- **多樣性**
  - ➢確保數據集包含廣泛的多樣性。
- **代表性**
  - ➢數據集內容是否充分代表相關類別。例如：面部識別系統應包括各種性別的照片。
- **偏見和公平性**
  - ➢注意數據集中的潛在偏見。確保數據集平衡,不偏向任何特定群體。解決偏見和公平性問題對於建立道德和無偏見的AI系統至關重要。

機器學習中使用訓練和驗證資料—數據收集的挑戰

- **成本**
  - ➢收集標記數據通常成本高昂，而獲取未標記數據也有其挑戰。標記未標記數據增加了額外的複雜性和成本。
- **質量問題**
  - ➢數據可能存在品質問題，影像失真或標記錯誤。這些問題可能源於設備限制或人為疏忽。
- **雜訊影響**
  - ➢模糊的圖像、扭曲的文本、背景雜訊干擾的聲音記錄等都可能影響數據品質。

3 .延伸閱讀與思維創新

- **延伸閱讀**
  - Joel Grus, Data Science from Scratch中文版（第二版）：用Python學資料科學, 歐萊禮，2019-22-25， ISBN：9789865023195 。
  - 吳作樂、吳秉翰 (2020). 圖解機器學習、人工智慧與人類未來，五南圖書，出版日期 。

- **思維創新**
  - 資料是機器學習的核心，請說說你工作與生活環境會有產出哪些資料？
  - 請舉例說明你的工作環境中會有哪些型態的資料？
  - 如何收集這些資料？
  - 想想哪些資料具有隱私性需要保護？