

# 機器學習--自然語言處理 Natural language Processing



數數看這一頁有多少AI技術  
learning by doing

von anwendeng



產生AI假圖不稀奇，產生AI短影片就要龐大的計算能力





當然，以下是一張維根斯坦的圖像，這張圖像來自維基共享資源，因此不違反著作權法：



維根斯坦在劍橋大學

這張圖像是維根斯坦在劍橋大學的照片。維根斯坦是一位奧地利出生的哲學家，他對語言哲學有著深遠的影響。

- In den „Philosophischen Untersuchungen“ sagt Wittgenstein: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.“ (§43) Das heisst, dass, wenn wir ein Wort verstehen wollen, wir die Regeln lernen müssen, wie wir ein Wort anzuwenden haben.



請用台灣用語解釋這段維根斯坦的話，並包含德文精確翻譯

• In den „Philosophischen Untersuchungen“ sagt Wittgenstein: „Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.“ (§43) Das heisst, dass, wenn wir ein Wort verstehen wollen, wir die Regeln lernen müssen, wie wir ein Wort anzuwenden haben.



好的，以下是維根斯坦這段話的台灣用語解釋，以及德文精確翻譯：

德文原文：

In den "Philosophischen Untersuchungen" sagt Wittgenstein: "Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache." (843) Das heisst, dass, wenn wir ein Wort verstehen wollen, wir die Regeln lernen müssen, wie wir ein Wort anzuwenden haben.

台灣用語解釋：

維根斯坦在《邏輯哲學論》中說：「一個字的意思就是它在語言中的用法。」這句話的意思是，要理解一個字的意思，我們必須先學會這個字在語言中的使用規則。

舉例說明：

例如，我們都知道「桌子」這個字，但如果我們不知道桌子是什麼東西，那麼我們就無法理解這個字的意思。只有當我們知道桌子是一種用來放東西的傢俱，而且它通常有四條腿和一個桌面，我們才能真正理解「桌子」這個字的意思。

## chatGPT說：



路德維希·維特根斯坦 ( Ludwig Wittgenstein ) 是奧地利出生的猶太人哲學家，生於1889年。他是家中九個子女之一，父親是奧匈帝國鋼鐵工廠的主管。他曾就讀工程學和數學，後轉向哲學。維根斯坦於1913年出版了《邏輯哲學論》 ( Tractatus Logico-Philosophicus ) ，該書深刻影響了哲學和語言哲學。第一次世界大戰期間，他在軍中服役。

1920年代初，他回到奧地利，開始教授小學，並追求哲學研究。他的第二個重要著作是《哲學研究》 ( Philosophical Investigations ) ，於1953年出版。維根斯坦強調語言和語言遊戲的重要性，並對語言哲學提出了新的看法。

他的生平充滿著思考和思索，對哲學和語言哲學做出了深遠貢獻。維根斯坦於1951年去世，留下了一個精彩而有影響力的哲學遺產。

# 傳統機器學習演算

- 監督式學習
  - 回歸
    - 線性回歸
    - 非線性回歸
  - 分類
    - 邏輯回歸
    - 支持向量機
    - 決策樹
    - K-近鄰
    - 朴素貝葉斯
- 無監督式學習
  - 聚類
  - 降維
  - 異常偵測
- 非監督式學習

# 特徵演算

- 數值特徵演算
  - 標準化
  - 正規化
  - 歸一化
  - 異常值處理
  - 缺失值處理
- 文字特徵演算
  - 詞頻統計
  - 向量化
  - 降維
  - 特徵選擇
- 影像特徵演算
  - 特徵提取
  - 特徵描述
  - 特徵選擇

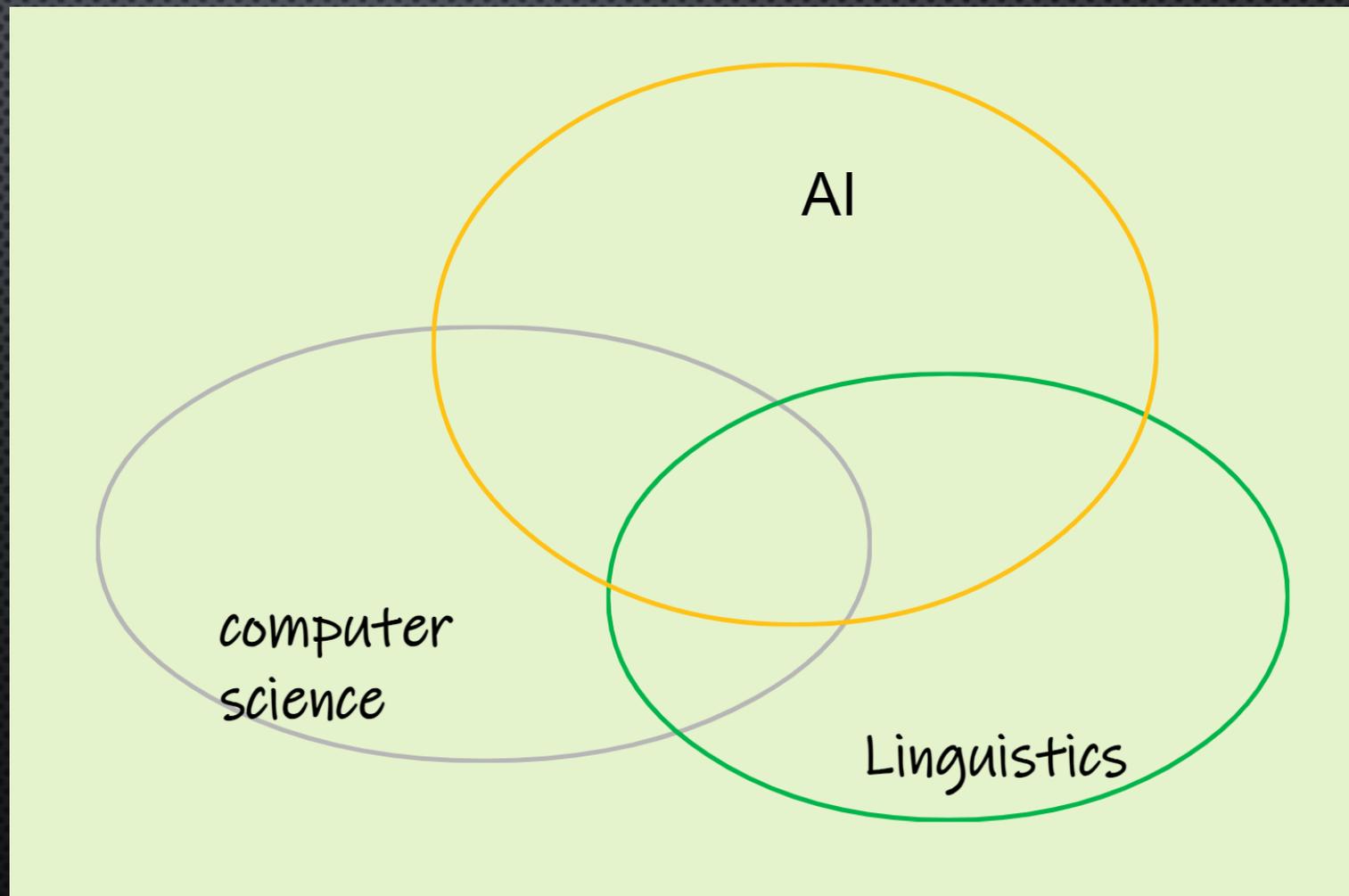
# 機器學習

傳統機器學習

特徵學習

神經網路  
/深度學習

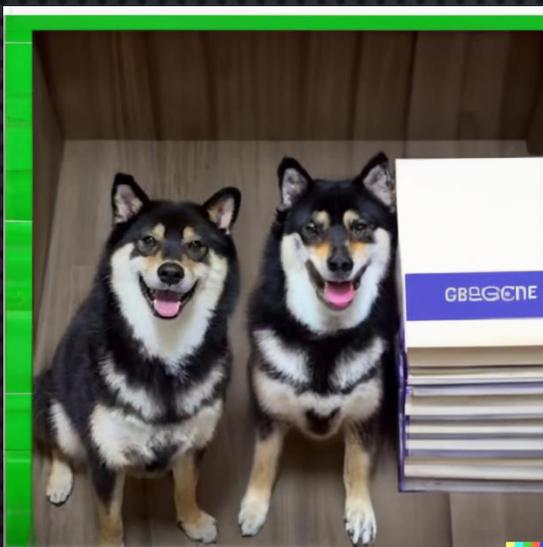
# 自然語言處理



## NLP 的應用

- 文件分類
- 機器翻譯
- 搜尋引擎
- 語音辨識
- 聊天機器人

# 文字怎麼量化？



# 語言量化

## 字詞的 One-Hot 編碼

將語句中出現的詞  
一一列在各「行」

the bat sat on the cat.

words

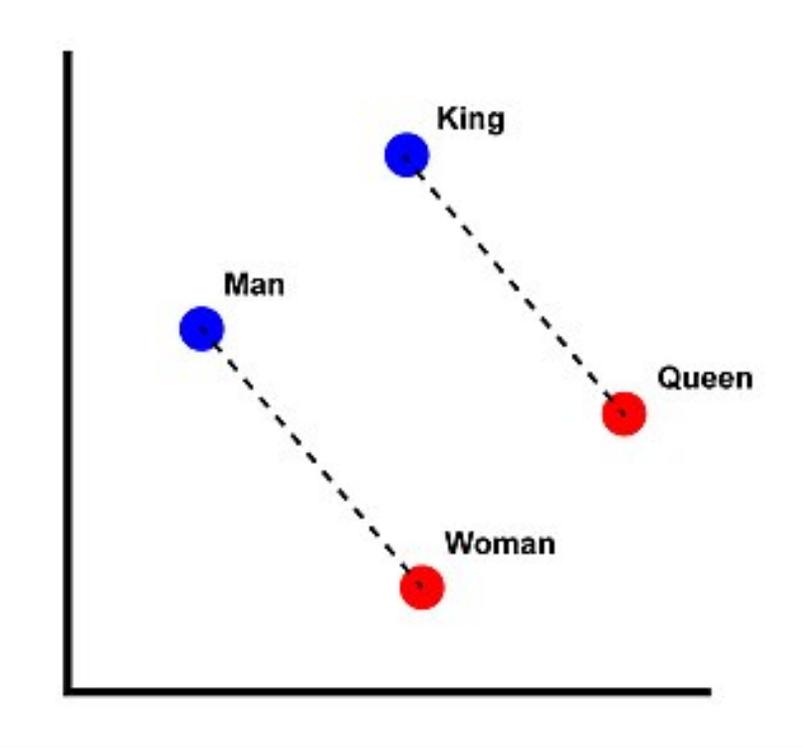
|     |   |   |   |   |   |   |
|-----|---|---|---|---|---|---|
| the | 1 | 0 | 0 | 0 | 1 | 0 |
| bat | 0 | 1 | 0 | 0 | 0 | 0 |
| on  | 0 | 0 | 0 | 1 | 0 | 0 |
| ⋮   |   |   |   |   |   |   |



▲ 圖 2.4: One-Hot 傳統編碼方式

# John Rupert Firth

- You shall know a word by the company it keeps
- context words/target word



# X之於Y如同A之於B

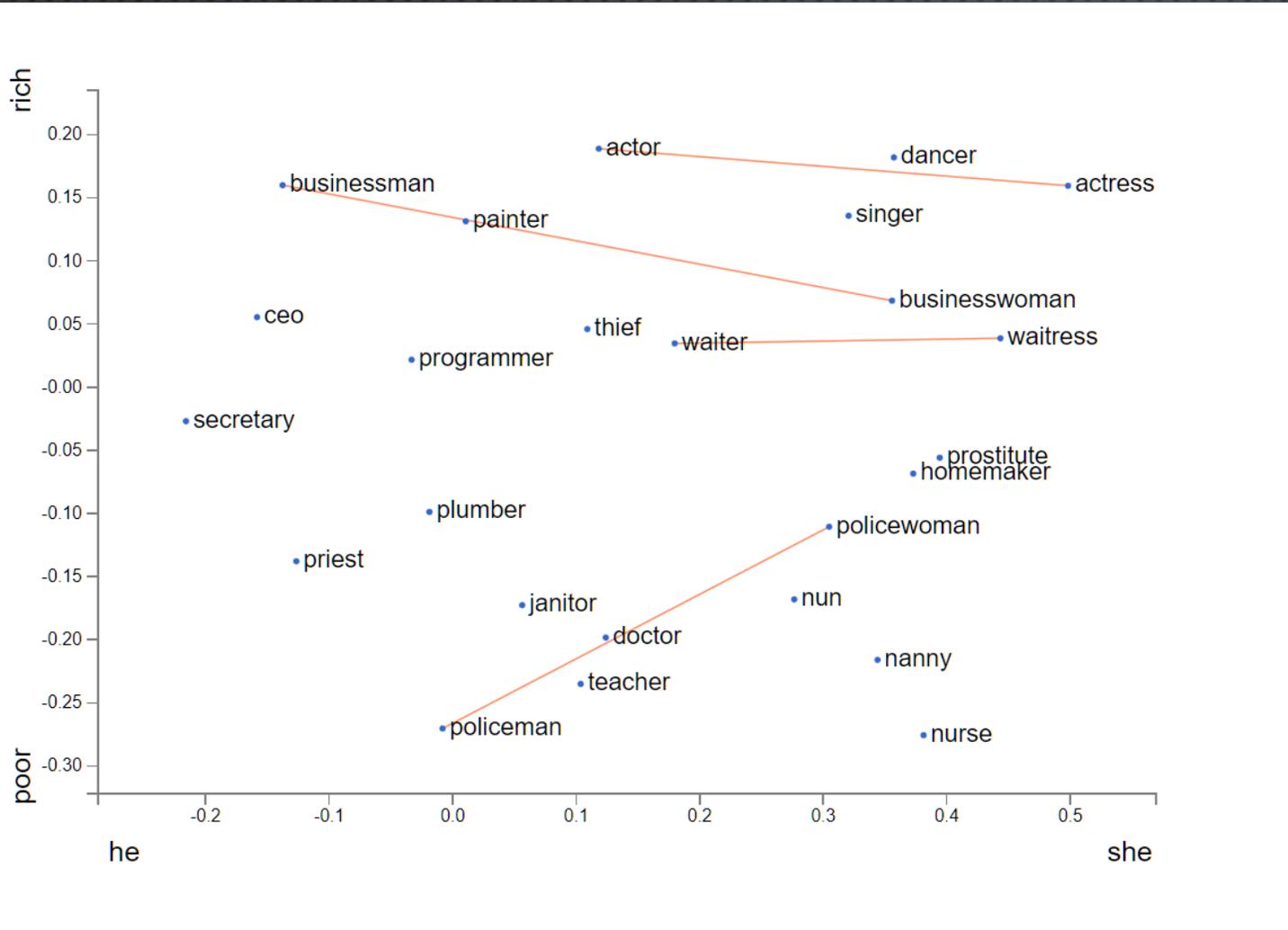
$$\begin{array}{l} \mathbf{v}_{\text{king}} - \mathbf{v}_{\text{man}} + \mathbf{v}_{\text{woman}} = \mathbf{v}_{\text{queen}} \\ \mathbf{v}_{\text{bezos}} - \mathbf{v}_{\text{amazon}} + \mathbf{v}_{\text{tesla}} = \mathbf{v}_{\text{musk}} \\ \mathbf{v}_{\text{windows}} - \mathbf{v}_{\text{microsoft}} + \mathbf{v}_{\text{google}} = \mathbf{v}_{\text{android}} \end{array}$$

▲ 圖 2.7: 詞向量的運算

## 詞向量 (word vector) 詞嵌入 (word embeddings) 或向量空間嵌入 (vector-space embeddings)

- 它會同時紀錄單字的位置與意義，是開發者較偏愛的編碼方式。詞向量空間的概念是將語料庫中的每個字詞都與一個多維度向量空間中的位置對應起來，這是模型訓練前須先完成的前置工作。初始時各字詞所對應的向量空間位置是隨機決定的，在訓練中深度學習模型會分析哪些字較常與特定字詞一起使用，就會逐漸移動到適合的新位置上（物以類聚），這樣語意相近的字詞就會靠得比較近

<https://lamyiwce.github.io/word2viz/>



## Explore word analogies

What do you want to see?

Jobs

Modify words

Type a new word...

Add

Type a new word...

Type a new word...

Add pair

X axis:

he

she

Y axis:

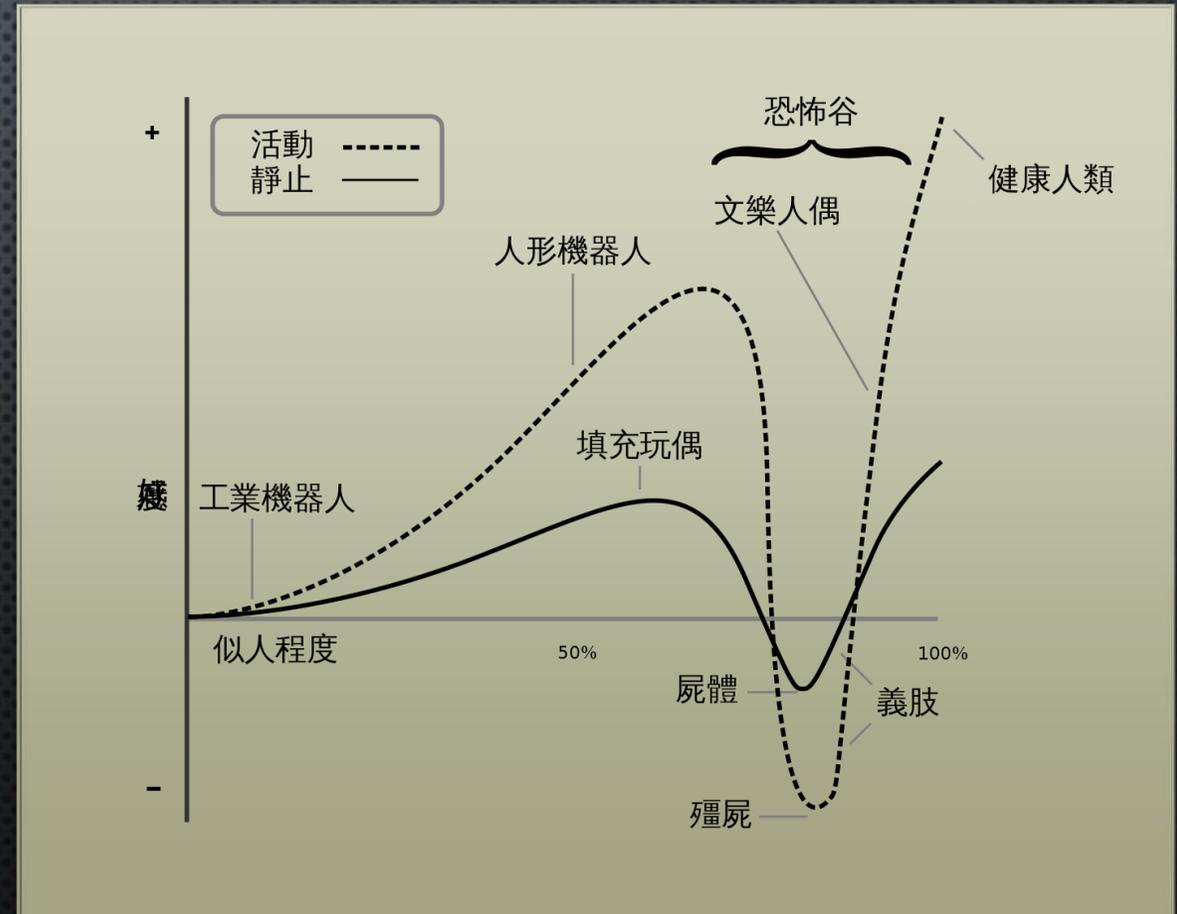
poor

rich

Change axes labels

# 恐怖谷, 不気味の谷現象, Uncanny Valley

- 1970年，森政弘Mori Masahiro發表了一篇題為《恐怖谷》的文章。這篇文章假設機器人會越來越像人，假如它們與人類的相似程度達到一定的程度的話它們外形上的微小的不完善會在人類的心理中引起不安。
- 1978年，森政弘在東京工業大學工作時通過實驗發現人類可以對看上去機器似的機器人產生好感，但假如同樣的機器人看上去幾乎像人時人類這個好感會變成厭惡。他將這個現象稱為恐怖谷理論。森政弘因此相信機器人生產商不應該嘗試讓他們的產品在外表和行動上過分像人。



- 深度學習能自動從資料中萃取關鍵特徵的優勢，將字詞轉成詞向量，嵌入詞向量空間的概念，記錄字詞的脈絡資訊，提供給NLP 模型使用

# 感謝觀賞

Herzlichen Dank für die  
Aufmerksamkeit

von anwendeng