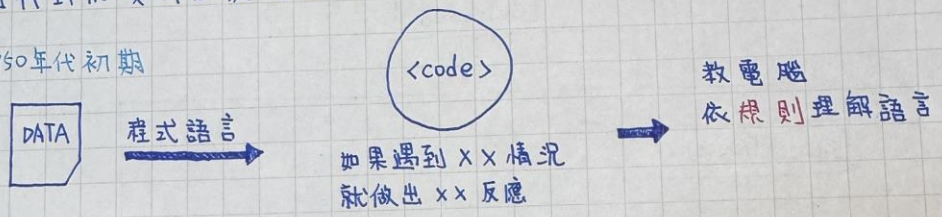
11/11 生成式人工智慧與異質平台整合應用

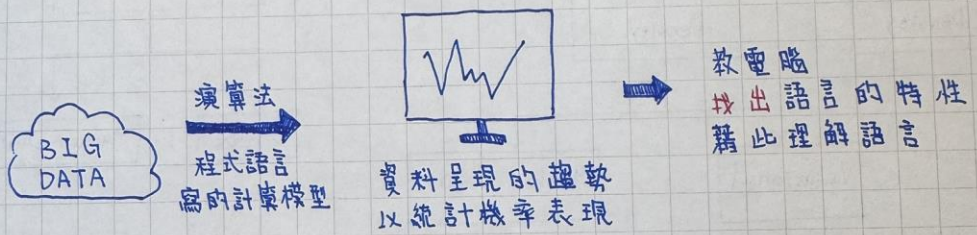
國立勤益科技大學 欉振坤 教授兼系主任

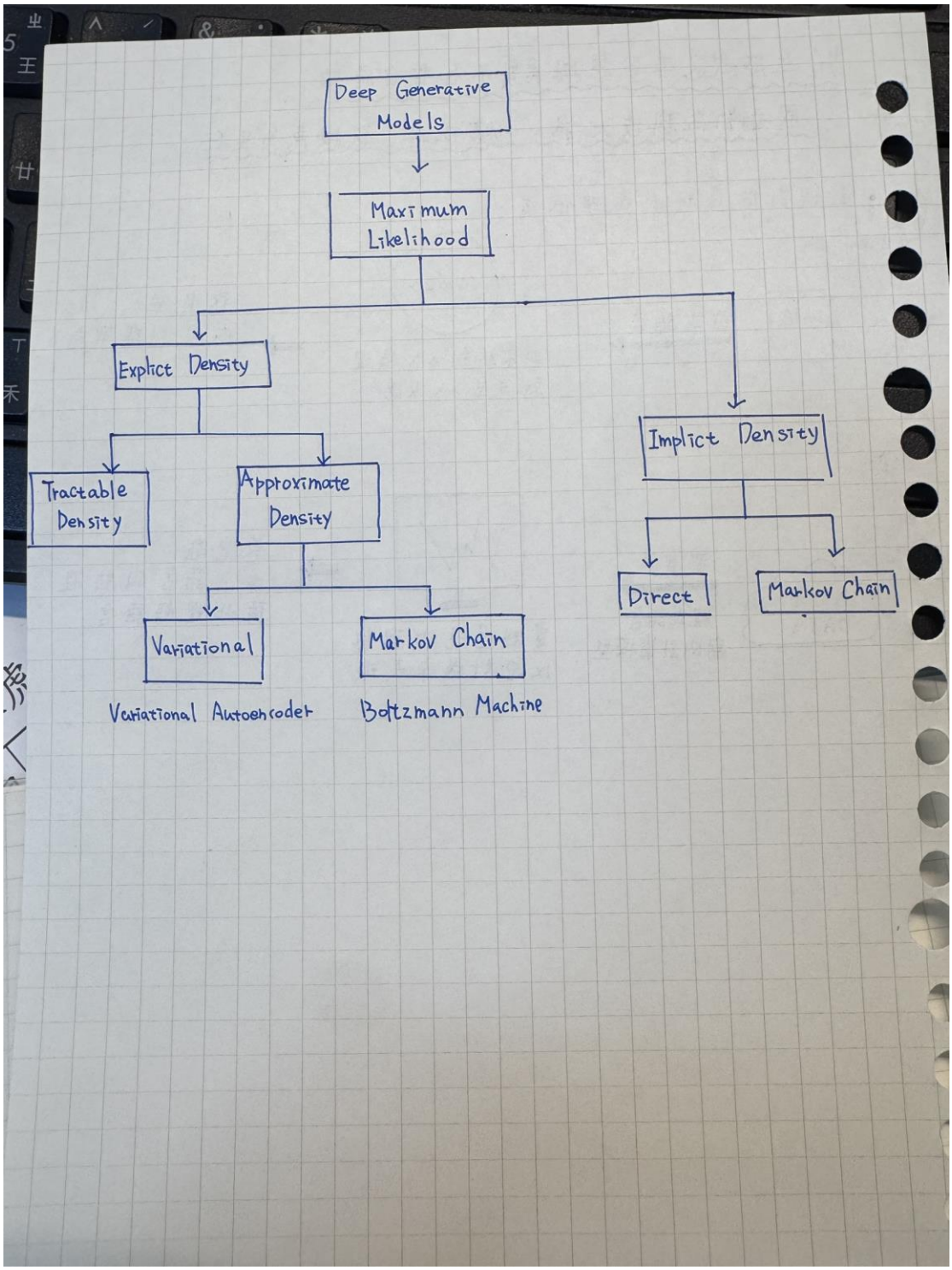
- 近代對於資料的處理做法

1950年代初期



1980年代末期至今




Deep Generative Models
Maximum Likelihood
Explict Density
Implict Density
Tractable Density
Approximate Density
Direct
Markov Chain
Variational
Markov Chain
Variational Autoencoder
Boltzmann Machine

- 生成式AI的主要技術

一. 生成對抗網路（GAN，Generative Adversarial Network）

- 結構：
  - Generator（生成器）：負責產生「假資料」
  - Discriminator（判別器）：負責分辨「真資料」與「假資料」
- 訓練方式：
  兩者互相對抗、不斷改進
  → 生成器會產生越來越逼真的內容。
- 應用範例：
  - 假幣圖像示例（真實 v.s 生成假圖）
  - 可用於圖像生成、影像修復、藝術創作等。

二. Transformer 架構模型

- 代表技術：OpenAI ChatGPT 4.0
- 特點：
  - 基於注意力机制（Attention Mechanism）
  - 能理解上下文語意並生成自然語言。
- 應用範圍：
  對話系統、翻譯、內容生成、輔助寫作等。

三. 異質平台整合應用

- 異質運算平台：指多種硬體（如CPU、GPU、FPGA）共同運作以提升效能
- 將生成式AI與異質運算整合 → 可提升訓練與推論效率

GAN 的優勢：1. 針對已有部份資料來產生不存在的資料
2. 或是擴充原有的功能

How Can GPT communicate well ?

- Step 1 : Collect demonstration data , and train a supervised policy

目標：讓模型學會人類示範的良好行為

過程：1. 從提示（prompt）資料集中抽樣一個問題
例：Explain the moon landing to a 6 year old.

2. 人類標註者（labeler）示範理想的回答方式。
例：Some people went to the moon ....

3. 這些示範資料用來進行監督式微調(SFT)
使 GPT 模型能模仿人類回答風格

- Step 2 : Collect comparison data , and train a reward model.

目標，建立一個能評估模型回答好壞的「獎勵模型（Reward Model）」

過程，1. 從提示集中抽樣一個 prompt，並產生多個模型輸出
例：A、B、C 三種版本回答。
2. 人類標註者對這些回答進行排序，從最好到最差
3. 這些排序資料用來訓練獎勵模型，讓模型能學會
「何種輸出更被人類喜歡」

- Step 3 Optimize a policy against the reward model using reinforcement

目標：透過強化學習，進一步優化模型行為

過程：1. 從資料集中抽樣新的 prompt
2. 模型產生一個回答
3. 獎勵模型計算該回答的「獎勵分數（reward）」
4. 使用 PPO 更新模型，使其在未來產生更高分的回答
5. 不斷重覆這個過程，使模型逐漸「學會更好地與人類溝通」。

Training Cost

① GPT-1 (約4.5G)
使用8個P600運算單元，訓練30天

② GPT-2 (約40G)
未知

③ GPT-3 (約570G)
估計使用1000張A100運算單元
訓練30天 + 60調校

④ GPT-4 (約?G)
使用8192個H100運算單元
訓練90-100天

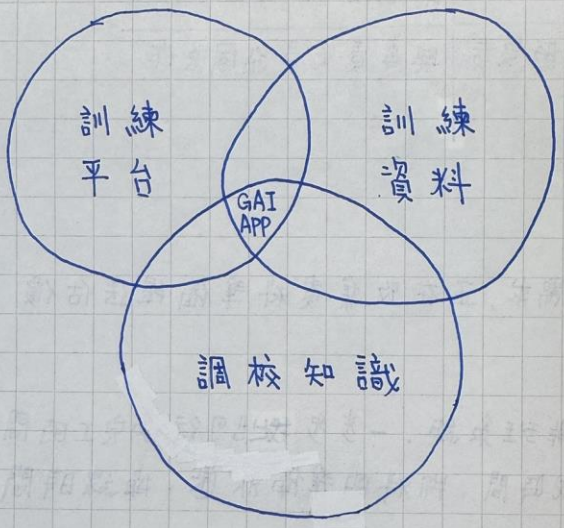
- AI Training Cards - P100:
  - △ Tesla P100 單價1萬
  - △ GPT-1 僅用8萬元設備 + 30天即可完成訓練
- AI Training Cards - A100:
  - △ A100 單價60萬
  - △ 使用1000張A100訓練GPT-3 估計要30天
- AI Training Cards - H100:
  - △ H100單價100萬
  - △ 微軟Azure至少有5萬張H100，Google手上大概有3萬張，Oracle大概有2萬張左右，而特斯拉和亞馬遜手上至少拿有1萬張左右。
  - △ 若使用3584張H100訓練GPT-3，僅需11分鐘

Comparison between different GAI Services

	GPTs	Gemini
面對無法應答問題的處置	偏好於生成該問題的可能的虛擬內容。	僅生成信心值較高的答案。在給定生成範圍時，若信心值較高，可提供生成範圍內的答案；若超出指定範圍，則會生成信心度較高的答案；若無法找到信心值較高的解答，不會生成不確定結果
付費專案	可 Fine-tune；可存取所有 GPT 模型；128000 tokens	可 Fine-tune；可使用所有 Google Assistant 和 2TB Drive；可以任何 Android 設備存取；1000000 tokens.
推論效能	高達 95.3% 的常識推論效能，優於 Gemini 的 87%	大規模多任務語言理解 (MMLU) 能力為 90%，優於 GPT 的 86.4%
網路存取範圍	可存取 Internet 資源，但不確定其可存取範圍	可以正確地存取所有 Google 資源，而 Internet 資源也可以，但不確定結果的正確性

Recall the Necessary

- 軟硬體整合的執行平台
- 設計有限時間的執行演算法



- 需要取得"有用"的資料
- 標註領域知識於資料中

- 對於領域知識有相當豐富的專業知識及經驗
- 掌握 GAI 執行細節

Recall

缺點	潛在的切入點
不是很準確	能夠有些容錯的空間
需要高品質資料	應用於延伸或擴展的情境
需要有人輔助	讓危險的工作由AI代替
無法在意外發生時負責	與專責人員協同合作

生產排程

- 當業務傳來訂單需求，正在收集資料準備提送估價
- As-Is
  - △ 生管人員開啟排班系統，一步步找出可行的完工時間
  - △ 線上工單的完成時間、物料的準備狀態、換線時間...
- To-Be
  - △ 生管人員詢問具有GAI的APS
  - △ APS主動介入協調工單完成時間、物料狀態、換線時間

Summary

- GAI導入在一般化應用的成本過高
  ⇒ 訓練成本、資料成本，系統調校成本
- 使用GAI串聯智慧化成果
  - 扮演資源協調者的角色，彙整智慧化服務以完成任務
  - BOM表生成
  - 生產排程(插單或急單規劃、報價規劃)