

Day 1

資料清理數據前處理

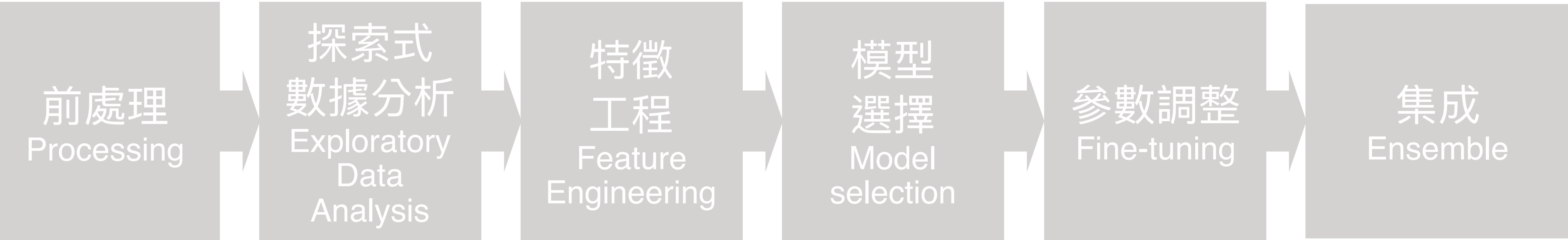
資料介紹與評估指標



知識地圖 機器學習概論 資料介紹與評估資料

機器學習概論 Introduction of Machine Learning

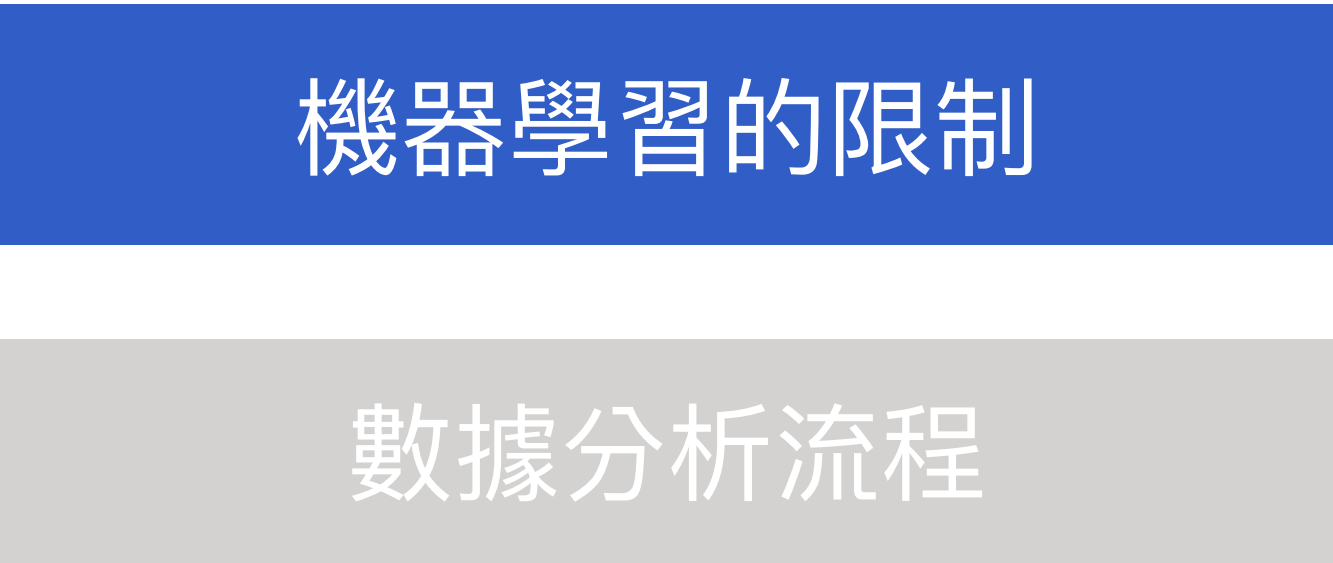
監督式學習 Supervised Learning



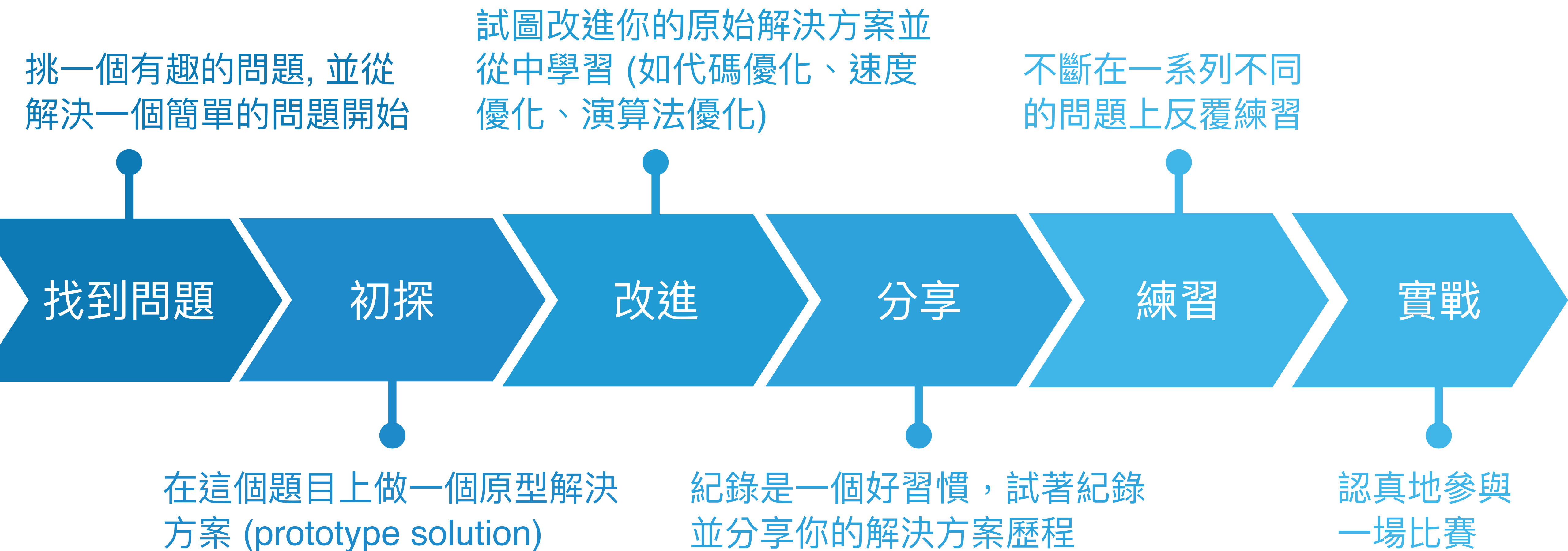
非監督式學習 Unsupervised Learning



機器學習概論 Introduction of Machine learning



學習路徑



首次面對資料，我們應該思考哪些問題？

Questions	Explanation	Examples
為什麼這個問題重要？ (Why it is important)	A. 好玩 B. 企業的核心問題 C. 公眾利益 / 影響政策方向 D. 對世界很有貢獻	A. 預測生存 (吃雞) 遊戲誰可以活得久, PUBG B. 用戶廣告投放, ADPC C. 停車方針 , 計程車載客優化 D. 肺炎偵測
資料從何而來？ (Where do data come from)	<ul style="list-style-type: none">來源與品質息息相關根據不同資料源，我們可以合理的推測/懷疑異常資料異常的理由與頻率	資料來源如： 網站流量、購物車紀錄、網路爬蟲、格式化表單、 Crowdsourcing 、紙本轉電子檔
資料的型態是什麼？ (What are they)	A. 結構化資料需要檢視欄位意義以及名稱 B. 非結構化資料需要思考資料轉換與標準化方式	A. 結構化：數值, 表格, ...etc B. 非結構化：圖像、影片、文字、音訊, ... etc
我們可以回答什麼問題？ 問題：指標 (What is our goal)	每個問題都應該要可以被驗證 → 有一個可供衡量的數學評估指標 (Evaluation Metrics)	常見的衡量指標如： 分類問題：正確率, AUC, MAP, ...etc 迴歸問題：MAE, RMSE, ...etc 補充資料： 衡量指標

範例一：我們應該要 / 可以回答什麼問題？

生存 (吃雞) 遊戲

- 玩家排名：平均絕對誤差 (Mean Absolute Error, MAE)
- 怎麼樣的人通常活得久/不久 (如加入遊戲的時間、開始地點、單位時間內取得的資源量, ...) → 玩家在一場遊戲中的存活時間：迴歸 (Mean Squared Error, MSE)



範例二：我們應該要 / 可以回答什麼問題？

廣告投放

- 不同時間點的客群樣貌如何 → 廣告點擊預測 → 預測哪些受眾會點擊或行動：Accuracy / Receiver Operating Curve, ROC
- 哪些素材很好/不好 → 廣告點擊預測 → 預測在版面上的哪個廣告會被點擊：ROC / MAP@N (eg. MAP@5, MAP@12)



解題時間 It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

