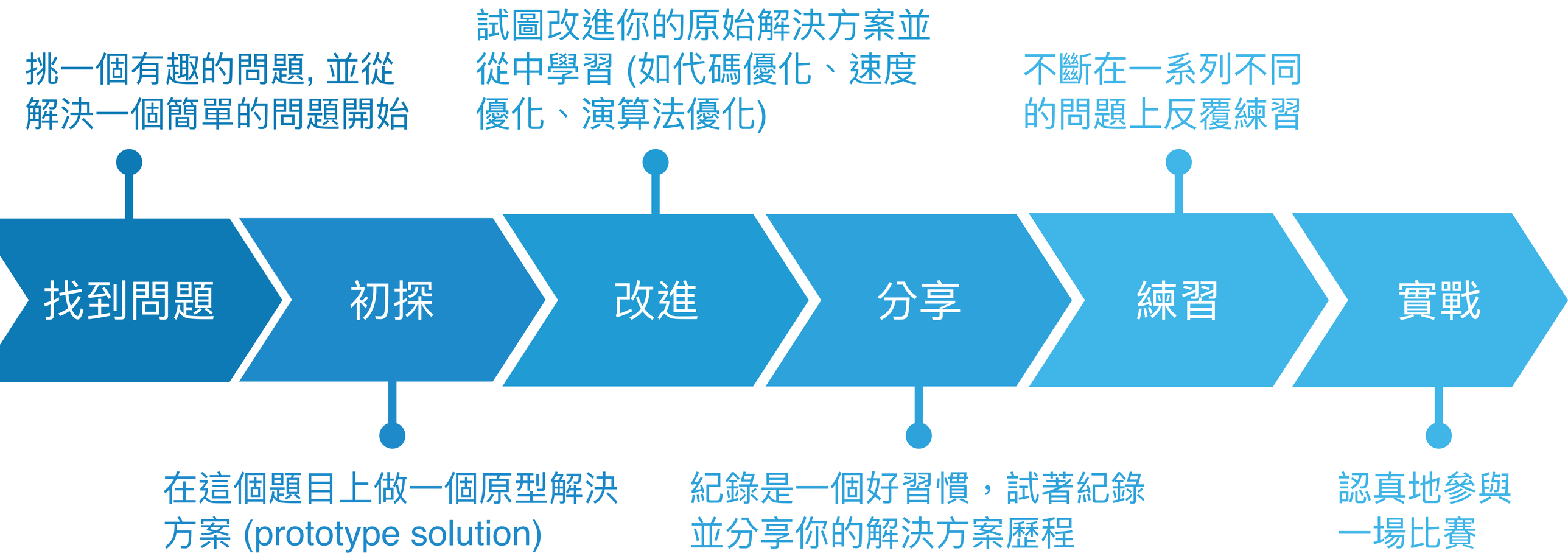
Day 1

資料清理數據前處理

資料介紹與評估指標

學習路徑


挑一個有趣的問題, 並從解決一個簡單的問題開始
找到問題
在這個題目上做一個原型解決方案 (prototype solution)
初探
試圖改進你的原始解決方案並從中學習 (如代碼優化、速度優化、演算法優化)
改進
紀錄是一個好習慣，試著紀錄並分享你的解決方案歷程
分享
不斷在一系列不同的問題上反覆練習
練習
認真地參與一場比賽
實戰

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)	• 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)

Day1 Homework 作業

- 請上 Kaggle, 在 Competitions 或 Dataset 中找一組競賽或資料並寫下
  1. 你選的這組資料為何重要
  2. 資料從何而來 (tips: 如提供者是誰、以什麼方式蒐集)
  3. 蒐集而來的資料型態為何
  4. 這組資料想解決的問題如何評估

Day1 Homework 作業

- 想像你經營一個自由載客車隊，你希望能透過數據分析以提升業績，請你思考並描述你如何規劃整體的分析/解決方案
  1. 核心問題為何 (tips: 如何定義 提升業績 & 你的假設)
  2. 資料從何而來 (tips: 哪些資料可能會對你想問的問題產生影響 & 資料如何蒐集)
  3. 蒐集而來的資料型態為何
  4. 你要回答的問題，其如何評估 (tips: 你的假設如何驗證)

- 請依照 Day_001_example_of_metrics.ipynb 完成 Mean Squared Error 的函式

補充資料 推薦閱讀文章

1. Data Scientist or Data Engineer?
    - Data Scientist or Data Engineer? (續 - 看看中國網友的討論)
2. R or Python for Data Science?
3. Why Data Scientist Must Focus on Developing Product Sense
4. Think twice before getting into data science (原文：Why so many data scientist leaving their jobs)