AI Tech 第39次 meetup

小資料科學家的生存之道

Cicilia Lee

2020/07/26

Outline

- 自我介紹
- Data Science 是什麼
- Data Science 相關的工作內容與職缺
- Startup 遇到的各種神奇客戶
- 實例分享：線上教學的分班問題
- 實例分享：網路輿情分析
- Q & A

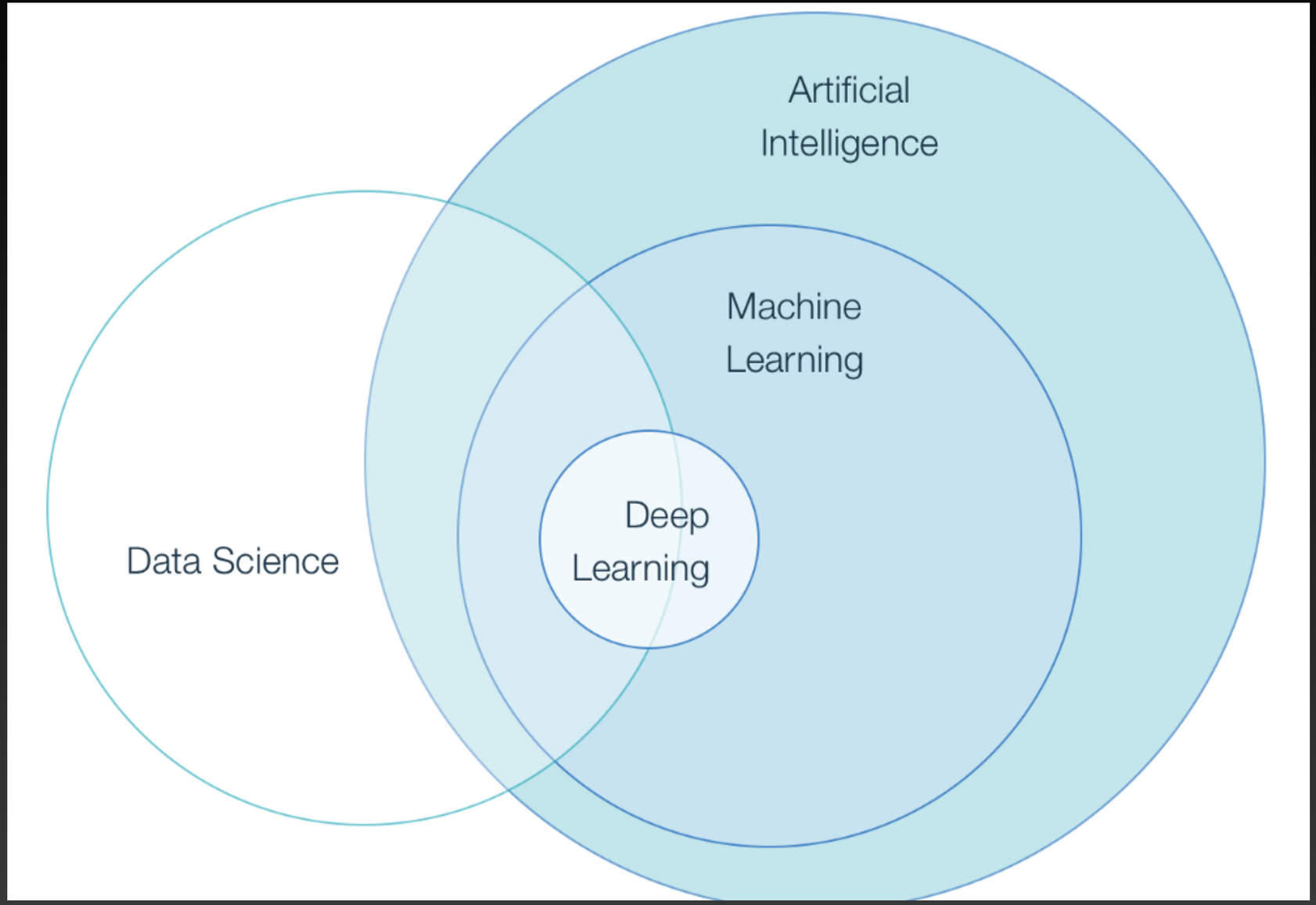
Cicilia Lee -1/2

- 台大資工所畢業，主修自然語言處理
- 修過人工智慧，機器學習，資訊檢索，網路資料探勘等課程
- 2009 畢業遇到金融海嘯，去當了軟體工程師
- 軟體工程師 + 機器學習 => 資料科學家！？

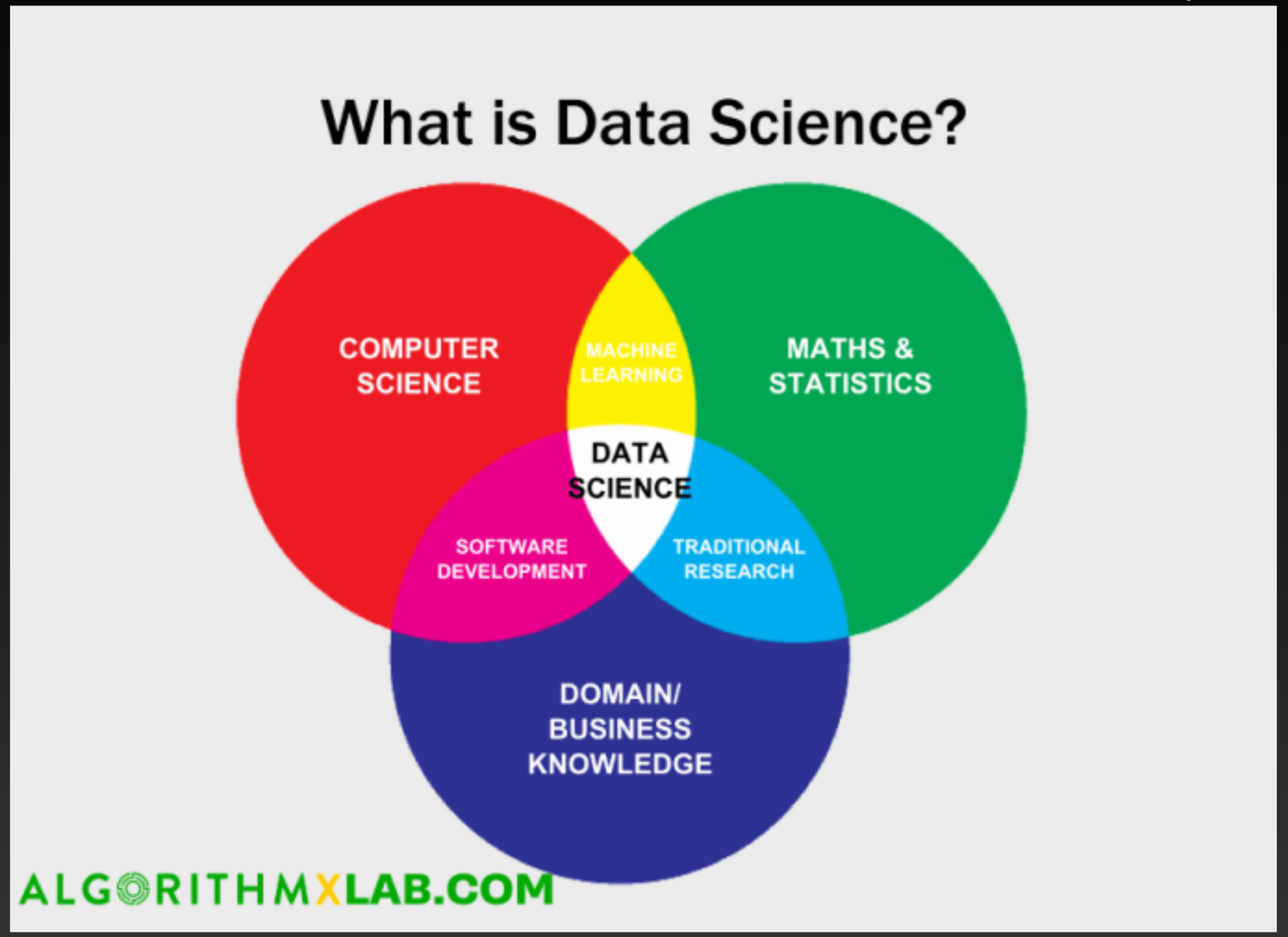
Cicilia Lee 遇到的挑戰

- 2012 Deep Learning 爆紅，以前在學校學的 Support Vector Machine 等傳統機器學習方法跟 feature engineering 相形失色
- Machine Learning 變化快，要新學CNN, RNN, BERT, word embedding等等
- 後來發現在小資料上面，傳統機器學習方法跟 feature engineering 還是有不錯的發揮

Data Science - 1/2



Data Science - 2/2



黃色是Machine Learning

Data Science 子領域

- Data Clean (資料清洗)
- Feature Engineering(特徵工程)
- Machine Learning(機器學習)
- Data Warehouse(數據倉庫)
- Data Visualisation(數據視覺化)
- …

Data Science 相關職位

- Data Analyst
- Business Intelligence Analyst
- Data Engineer
- Data Scientist
- Machine Learning Engineer
- Backend Engineer

Startup 遇到的各種神奇客戶

- 客戶說他有大資料，但他的大資料是200筆
  - => 跑一個簡單的統計分析
- 客戶說他有很多資料，但卻不知道這些資料能幹嘛
  - 要花很多時間瞭解客戶的痛點
  - 也要告訴客戶資料要標記跟清理後才能用

工人智慧為人工智慧之母

- 客戶的問題能不能人工做？人工作後能不能一步一步自動化？

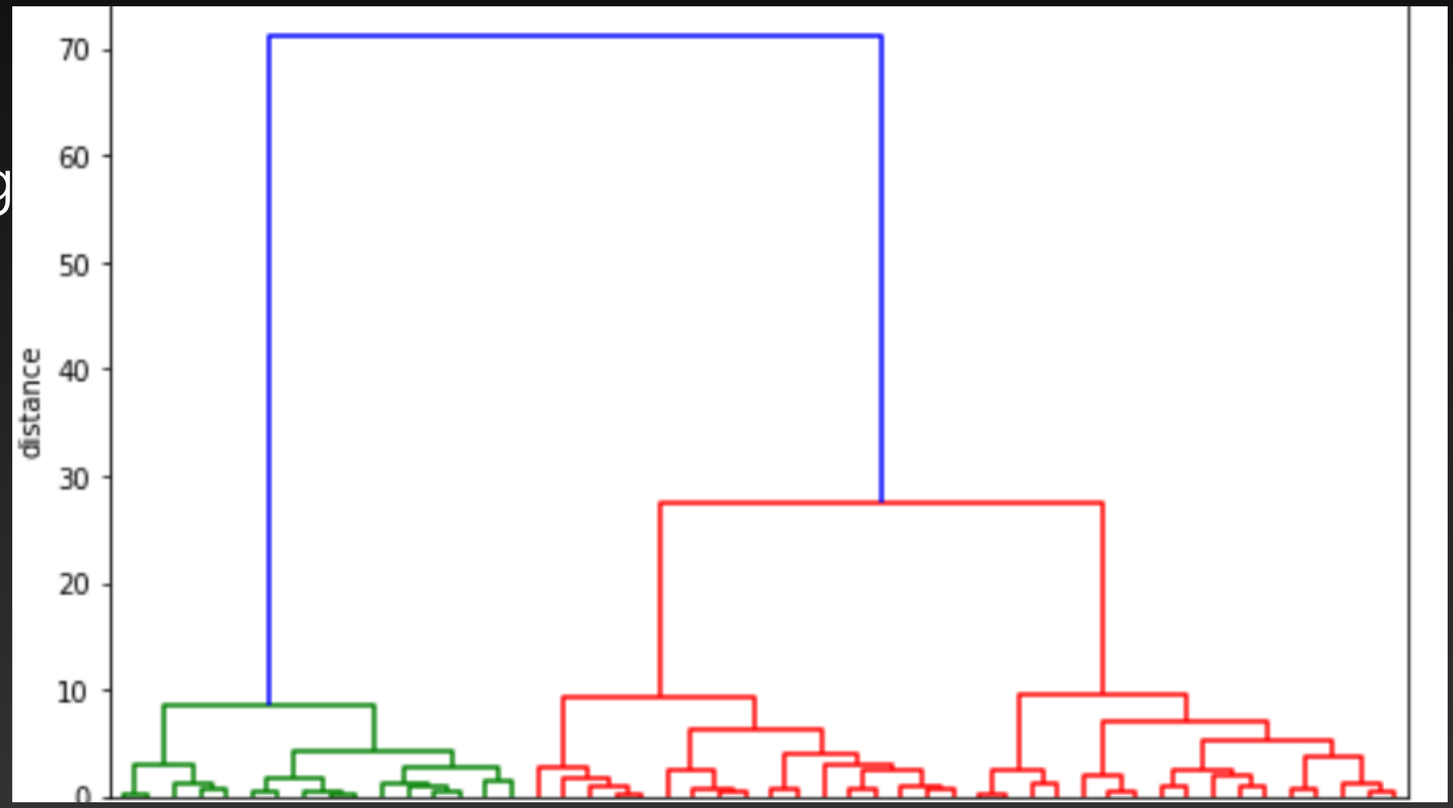
- 髮型自動標記
  1. 前處理，圖片品質，去雜訊，去紅眼等等
  2. 髮型資料人工標記
  3. 訓練工讀生知道各種髮型
  4. 研究相關的議題，人臉標記，圖片自動產生文字描述
  5. 嘗試用各種影像處理的pattern, windows 來當他的feature
  6. 看人工的標記品質夠不夠好？

線上教學的分班問題 - 1

- 每個小時有一群學生會需要線上上課，小班制。學生，老師，教材有不同的屬性。
- 客戶要求：
  - 演算法有可解釋性
  - 處理有時間限制
  - 最佳化學生滿意度

線上教學的分班問題 - 2

- Big data / Clustering
- 使用 Hierarchical Clustering



線上教學的分班問題 - 3

- 訪問客服：
  - 學生不喜歡某類老師
    - 後處理用規則濾掉
  - 某地方的學生跟另一個地方的學生上課網路會太慢
    - 前處理分開
  - 某學生不喜歡某學生
    - 例外處理

網路輿情分析

- Opinion or Truth
- Relevant to the topic or not
- Positive, Neutral or Negative
- Opinion holder
- Opinion Aspect
  - Aspect Word vs Sentiment word

https://ckip.iis.sinica.edu.tw/service/opinion/

•Opinion Aspect

以旅館為例：

Aspect 可能有：服務，食物，CP值，地點，房間，其他

例句：

“**房間廁所空間**及**廁所馬桶**偏小，但服務不錯，**遊客**品質不錯，好多日本人"

房間 廁所 空間 及 **廁所 馬桶** 偏小 ，但服務不錯，**遊客**品質不錯，好多日本人

[**廁所空間;廁所馬桶**], [偏小]

Aspect: 空間

Aspect Word: **廁所空間，廁所馬桶**

Sentiment Word: 偏小

Aspect: 服務

Aspect Word: 服務

Sentiment Word: 不錯

- Opinion Aspect

- Opinion Aspect Classification
- Opinion Aspect Word Extraction (IE)
- Opinion Sentiment Word Extraction (IE)
- NLP Tool, BERT, dialogflow, LUIS
  斷詞，取原型，NER

Q & A

- 分類方法 intent => key information
- 1. 訂餐廳意圖：
- 2. Entity 餐廳類型，地點，價位，人數
- Named entity recognition
- Intent 意圖，對話 Scenario
- Entity 實體
- Diagflow, Luis, BERT, ALBERT specific ch

Q & A

- Thank you
- https://www.linkedin.com/in/cicilialee