9/19

圖靈測試—人工智慧定義

如果人類與機器進行非面對面的對話，人類無法判對出對方是機器那麼這台機器就是有人工智能

1980 john searle劃分三種

弱人工智慧

強記、博聞(快速掃描、儲存大量資料)

運用數學模型、演算法模擬人類學習、分析邏輯思考表現

左腦功能(邏輯、理性)

強人工智慧

與人類具有相同邏輯思考能力，並能及時做出適當反映

具有創新能力、美感

兼具左腦與右腦(感性、藝術)的功能

AI(大領域代表)🡪機器學習(監督、非監督)🡪深度學習(類神經)

AI發展史

1次巔峰:很多算法 but領域狹隘

2次巔峰:語言辨識、機器翻譯、類神經網絡、專家系統 but技術不夠好、無法

達到人類期待

3次巔峰:演算法更新、電腦功能增強、大數據、深度學習

遊戲是測試AI的最佳平台 but是完全訊息

完全訊息VS部分訊息 Game(星海、dota)

泛人工智慧

外在行為像人類，但不能表達自我意識

腦袋在幼兒階段

IBM Watson

很難對單一領域外的事情發展，所以都只能深入單一領域

AI三大領域 AI三大技術

語音辨識 遺傳演算法(擇優傳下)

影像辨識 專家系統

自然語言處理 類神經網路

全域搜尋法

若我要一個有小排到大的數列 ex:234

那我將所有可能排出來找到符合的為止🡪全域搜尋法

But 運氣好是第一個數列 運氣差就是最後一個數列才對

搜尋n!次 n=50 =>50!種組合 50x…x1 > 10^40

龐大的演算量很沒效率

超級電腦1秒處理一兆數列(10^12)=>10^28秒=>1年小於10^8秒

能否達到對的答案取決於你的運氣(隨機)

選擇排序法

隨便找一個數列random出來，找出最小的那個進行位置交換

一個一個找數字最小放最左邊

回合數

搜尋次數:n+(n-1)+(n-2)…+1 < n^2

若50個數 n^2=2500 用超級電腦只要0.0000…25秒

平行運算—分群計算

區塊鏈、量子計算、增強分析、人工智慧

+互聯網🡪傳統交易可利用網路交易

互聯網+🡪互聯網+傳統行業 ex:智慧城市 推動技術進步、效率提升

ex: google地圖查景點 可以直接訂房買票

互聯網+交通=>Uber 用數據分析找網路的需求，定價浮動

互聯網+醫療、互聯網+旅行預定=>房價資訊

互聯網+金融支付=>支付寶…

互聯網+零售=>用你的歷史資料預估你什麼東西快用完了，通知你要不要買

4V🡪多樣variety、真實veracity、大量volume、快速velocity

9/26

手機的出現—>改變了資訊蒐集的數量、型式

4V🡪多樣variety、真實veracity、大量volume、快速velocity

快速velocity:累積速度很快(24hr即時更新)瞬間爆量

大量volume:大量資料儲存(硬體競賽) but能否使用(?) 垃圾資料

多樣variety:影音、word…等等 以前是娛樂用，今天變成訓練機器用

如何把資料變得有用 but 如何融合? 一起分析

真實veracity:確認資料的真偽、確認可靠性(如何蒐集?)

資料失真:爬取open data時，若是透過第三方一定要確認資料，不然對方改了

一點點可能都會導致模型無效。

最多時間會花在資料清理(耗時，功夫) 過程、既是別人看不到(只看結果)

所以Value(價值)是大家想看的

有用的資料多才是好的!!

數據多樣性

每天12+TB的推文 (會有條款要搜你的cookies、資料軌跡，可合法進你後端抓

資料，不簽就不能用只能上網完清掉你的cookies)

Cookies🡪包含你帳密的資料

TB🡪把兩艘貨櫃輪船的裝滿

ZB🡪太平洋

YB🡪一顆地球 but儲存裝置越來越小 累積速度目前全球大概4ZB的 60%資料是近五年累積的，預計明年資料會長10倍

資料複雜程度才是挑戰🡪處理資料到可用狀態 結構化和非結構化的資料都有

結構化資料(如:性別、年齡等欄位明顯的) 現今資料複雜程度高

OLTP交易型資料(資料庫管理系統)

OLAP分析型資料(資料倉儲)

RTAP即時分析資料(資料架構和科技)

理論的改變

行銷理論從4P到4D

4P理論 產品、價格、通路、促銷 如何透過數據分析達成促銷??

4C理論 消費者、成本、便利、溝通

4D理論 需求、數據、傳遞、動態 (建立在數據分析的強大後可直接分析)

需求:數據分析，了解用戶需要，提供符合用戶需求的產品(追蹤消費者使用者經

驗)CX,UX

數據:透過使用者蒐集，充分消費者痕跡(足跡)從是我的客戶變成不再是客戶的過程去重建軌跡、分析行為數據。 「要拿怎樣的資料去做怎樣的分析?」

Ex:去健身房的目的蒐集🡪減肥、健身的人消費會有差異

Ex:用紅外線來看客人走的動線、碰的東西來看消費者的心理🡪擺架方式

傳遞:將產品價值便利的傳給消費者

動態:即時性因應

全新的4P

人、成效、步驟、預測🡪個人化的行銷1對1的

過往不做是因為困難且成本高

人:異質性、變動性🡪消費者分群

同質高的比較好預測誤差小所以要做異質性切割分群(NES族群)

得出消費族群分三種:新顧客、主力顧客、半夢半醒的顧客(如何喚醒or可

否喚醒?)

N新顧客

E0主力顧客

S1~S3半夢半醒顧客

E0個人週期X2倍時間都會回來購買的人

S1瞌睡顧客比一班人X2時段未回來

S2半瞌睡比一般人x2.5時段未回來

所以🡪如何使瞌睡者變成主力，主力不變瞌睡是重點 而不只是增加新客戶

經數據分析決定政策🡪保留客人或是增加新客人 每一間店的行銷不同

1.顧客數下降、2.客單數下降、3.活躍度降低 策略不同

1.新客戶增加、留住舊客戶

2.新客戶促銷、舊客戶優惠:但降價是個爛方法，看到降價一次買大量但留

不住客人，因為之後不會再買。失去永遠比賺

得多，促銷是毒藥

做什麼活動，對象是誰、優先度要去分析。預測新客戶何時會變主力，主

力何時變瞌睡，要讓新客人變主力的區間變長，堵塞客戶流失，了解誰要

改變，去喚醒🡪一對一行銷手段

不是要看個別商品的上升下降，重點是人。

3.客戶定期關懷、購買時事先提醒

資料蒐集方式改變

敘述性偏好:模擬未存在環境進行需求預測

顯示性偏好:假設消費者是理性，會清楚知道自己的偏好

過往都是透過問卷蒐集 BUT可操弄

Amazon mechanical turk

在網路的行為是真實的(沒有人知道你是誰)用網路資訊蒐集

常收到B2C的廣告🡪打開廣告信就會啟動連結推相關的標題

現在是先去找喜愛產品的特徵，把擁有該特徵的人找出來在投放廣告

但只是特徵相似而已，目的性不同

導致廣告沒興趣、價位太高、預期有折扣、或根本不需要🡪包含雜訊

不夠精準

透過大數據找到對的客戶、對的價格、對的地點、對的時間(有興趣去的時

候)、對的產品(不是類別而是商品，例如你愛的品牌)

這五項的精準度都要高，要建模。假如這五個條件都有90%那整體呢?

0.9^5=0.59 等於亂槍打鳥(0.5) 缺陷就是提升不多

利用hit rate來檢測5R程度

希望「精準」的「預知未來」

如何從大批資料裡面拿出有用的預知未來，生活就是數據

但模型無法將不理性的事情納入🡪我們都在簡化模型

把可以用得先拿出來用

發生什麼? 為何發生? 未來是否發生? 如何讓它發生? 要做什麼?如何?為什麼?

先做敘事型分析，在做診斷型分析最後才做預測型分析(什麼因素對目標有影響)去複製「旺季」，了解什麼影響消費者購買去行動🡪指示型分析(最後)

認知型分析🡪理解數據背後的「為什麼?」🡪AI

經濟學:預測來源🡪需求與供給 (預測的起源) 假設人是自利、理性的

假設完全信息的狀態

行為分析才是最重要的，把人類行為放到數據分析裡面考量

行為經濟學，結合了心理學🡪將人類設為有限理性(心理帳戶)

損失意識(買賣價格要一致)ex:買賣家預期的價差 社會偏好 缺乏自我控制

傳統經濟模型的預測缺乏太多因素 but如何把這些因素放進去??-🡪認知分析

想法、感覺、行為(認知行為理論)FMRI🡪強人工智慧要追求的就是這個

假如這些因素可以把握，就能更精確的預測行為。

10/3

創造性🡪資料分析出的價值

EX:

床亂的人比床整潔的人，創造力比平均高50% 🡪資料分析圖片找床亂的(指標)

=>這些現象(行為)所代表的是什麼? (key point)

遲到的實驗

A型:爭強好勝、野心 =>感覺一分鐘，實際平均58秒

B型:有創造力、喜歡沉思及探索 =>感覺一分鐘，實際平均77秒

運用:例如用在排隊等待上的評估，從間接訊息挖出有用的資料

標價100:重點不是定值，而是知覺價格=>貴or不貴

啤酒&尿布的行銷分析

1992年teradata對大賣場的發票作分析

根據分析，這兩個常放在同一個購物籃=>一起結帳

傍晚5點到7點的相關性極強=>下班時間爸爸買啤酒順便買尿布

結果:尿布放在啤酒旁邊 => 你會順手去拿「相關」的商品一起買

But 應該是啤酒+魷魚絲(下酒菜)吧??

關聯分析=>觀察各種變數的關係存在 but是否為虛假關係?

先拿尿布再拿啤酒 or 先拿啤酒再拿尿布?? =>心理狀態不一樣

只做關聯不夠=>沒有因果關係 所以這是否為虛假關係要再檢驗

啤酒&尿布&派

這裡還沒討論品牌、價格

派是現做的，而且狀況奇怪，一下子鹹的好一下子甜的好

美國:買完啤酒、尿布順便買派回家吃 but沒辦法繼續挖

派跟電影院波的電影有差=>正向情緒下傾向健康食物

去挖出電影放的型態=>得出今天要準備鹹派or甜派

巧克力&諾貝爾獎

一個國家的巧克力吃得越多，諾貝爾獎得獎越多

巧克力有刺激腦的激素激發學習能力

But不是相關係數高就代表真的有關係=>即便高相關也不代表真的有關聯

預測車禍的分析系統

美國田納西州

用各種模型分析大概只到50% but用其他的外部資料就會增加準確率

Ex:足球賽的時間、販賣酒的位置

盒虛圖適合用在資料分析=>可以看異常值

利用空間、時間的資料蒐集得出uber需求的分析=>告訴司機哪裡有需求

動態定價=>需求高的可以抬高一點價格

空間大數據分析

把熱門上下車地點標出來，並計算車資+小費，算出哪裡在客人收益最高

But不會全部人都去=>會搶爆 可能要新增預估競爭者

懷孕預測

從25項產品售出預測你有沒有懷孕 Target

恐怖攻擊預測 predate

挖掘資料=>萃取找出相關訊息

AI判斷性向(DNN) =>整形 眼睛的軸距沒法改、有一個特殊的三角形角度動不了

可用虹膜分析 but 成本、困難

特徵分析

傳統

變數的輸入與輸出間的關係

分析前提條件改變就不適用

相似比對

無需考慮輸出變數

考慮客戶的相似性與結構性=>把相似結構的人找出來

EX

洗髮精效果=>長頭髮、烏黑亮麗

傳統方法能找到人

相似性對比=>小狗長髮(類型)

特徵分析=>狗通常長得像主人=>建資料庫辨識就不用打晶片了

分群:相似性 分類:有固定型態如男女

行為分析建置客戶型態

心理:網路社群發言

男性:A比B還好、對比式發言、行動電話上下擺動

女性:A和B相似、類化式發言、行動電話左右擺動

=>去分析發言形式的含量，對於後面的行為影響(內在影響)

因為男性手機放口袋，女生通常放包包=>從擺動來估計男女性數量

具名發言VS不具名發言

不具名比較能反映真實狀況

用行為分析來判別性別

很多的會員資料六成是錯的:沒有更新

不可能用單一訊息去分析資料

精準行銷基礎:實體數據🡪實體賣場中的人的行為

網站數據🡪在逛什麼網站

從發票資料估出哪些人想去健身房

消費動機不同:健身or減肥

企業的適性測驗=>從量表去看得出如何操縱型態

AI百年疑題=>是鴨子還是兔子? AI理解力下=>鴨子

But再試一次=>又判斷是兔子(將圖片旋轉)

AI仍有其偏誤、不穩定

用社群網路的使用來判斷一個人的型態，比量表準確率還高

消費大數據

目的:精準預測未來

Amazon專利:預測式發貨🡪判斷你的瀏覽行為，把商品往你家附近發貨

用行為、其他因素判斷(願望清單、商品關注、滑鼠流動狀態)判斷購物機率

🡪發貨(預測需求)

用模糊填寫用戶地址方式將特定商品送往購買者附近

他也會推薦商品，將特定商品也送至潛在購買者附近

加上無人機配送=>目的在於3分鐘內商品到你手上

文本分析、文字探勘

把文字精簡化、卻又不遺失意涵去萃取出話語的情緒

中文斷字、排停用字、語意解析 =>但不同領域的分析語句不同

如何拆解詞彙來解析語意

透過電腦來讀文字

語意分析臆測心理產生判斷結果=>從推特上的發言來猜心臟病發機率

要了解心臟疾病的發生原因=>心理 因此有效

醫學檢查中心的方式=>有偏誤 空腹不能吃早餐並且要保持愉快心情(???)

杜鵑在呼喚

用文本分析找作者 羅伯特.加爾布雷斯 => JK羅琳

AI寫詩🡪小冰

意象抽取🡺圖片 後面都是根據第一句話產生的(前面有固定風格)

意象抽取->靈感激發->風格->試寫第一句->讓第一句去迭代->完成

社群分析

你在網路上案讚的社團=>分辨男女、胖瘦、學歷、收入

怎麼利用大數據分析

應用端的加值；創新

預判、預警、部署、應變、檢證、進步

加值:不同類型業務資料的整併

關鍵在於「價值」=>呈現與改變

把沒有價值的東西變得有價值or價值在我們認為沒價值的東西上(羊毛出在豬身上)

從蛛絲馬跡辨識商機 but蘋果、臉書、google、amazon壟斷資料

從非結構化數據中洞悉顧客需求 =>雜亂數據的結構化

社群網站試理解客戶的寶藏=>很好分析和滲透

從網路中得到資料再去應用(爬蟲技術)

10/17

11/21,11/28 社群分析專題

Day1 概念與應用

Day2 實際操作🡪有作業(報告成績、期末database)

大數據的創新

消費者標籤數據=>ex 音樂、時尚、科技、文字…等等

畫像數據=>基本維度、興趣維度、商業維度

推估生活型態，再推導心理層面

相同資料重建完整or挖到更多有用資料=>價值up

每個人的屬性和資料型態有很多差異

物以類聚，根據被推薦客戶相似的那些人購買項目來推薦

大數據影響解決問題的基本思維改變

傳統=>如何用數據支持或反駁

大數據=>我能從這堆數據找出怎樣的命題

羊毛出在豬身上，狗買單=>去哪兒網

穿山甲計畫(用數據分析和模型做顧客分析和價格分析)和保險公司合作買單

攜程去哪兒市占75%

移動技術到大數據分析

用新的渠道接觸現有和潛在客戶

建立品牌忠誠度和顧客參與度，帶來客戶成長與收入

大數據的迅速變化也意味著總是有更多空間

改善時間和空間

1.把現在沒有用的閒置資產有效利用，以及現在很缺錢的人如何把未來的錢拿來用

2.讓金錢從A到B更便利=>第三方支付、虛擬貨幣

3.人工智慧ex:智惠飛、Freee日本財務分析

不畏懼AI

機器學習是分析師的助手=>決策速度緩慢

ML可在效率和開拓(以前沒碰過的資料)上協助分析師

AI對於主觀數據判斷表現不好

語言藝術的影響

報表、投影片=>要用精彩的方式來講解數據，講故事的能力，如何利用數據講故事

一個好的報告=>不能回頭看投影片超過五次，時限誤差上下一分鐘，正裝

NLP自然語言處理很流行=>要產品化(機器人)

搜尋、指令要用到NLP 諮詢、辦理手續

多重雲AWS

雲端的問題:備份(被攻擊後怎麼辦)

CDO

資料首席官

群眾外包是數據控制的未來

自助式分析顛覆了商業知能的世界

又稱作傭兵=>企業外包給群眾幫忙分析給予賞金

但問題是數據品質的控管、可信度問題

數據保險網路安全保險的年度增長率為30%

但如何估算風險、數據價值多少

=>財務模型重視數據的成本價值、經濟、市場價值

=>非財務模型重視數據內在價值、業務、績效價值

數據洩漏的平均總成本估計為362萬美元但美國只有15%的公司有數據保單

數據工程師的角色

需要獨特的技能組合，了解數據中隱含的訊息

解讀能力、洞察

物位網

Location of thing推動物聯網創新

現在是逐漸成為分析過程中的一項輸入指標(一定要配備位置資料)

學術研究投資

大學院校與公司的產學合作案

數據科學分析已經成為必備技能

Data university

企業內部自己培養各種人才

2019年趨勢

可被解釋的AI =>我們不知道他為何精準or不精準

NLP讓數據人性化

可操作的分析數據隨手可得open data

數據協作放大社會良好的影響力

道德準則追趕數據=>為何你一定要去辨識個人資料

BI商業智慧平台(ex:數據台灣)提供資料呈現、產業發展

數據講故事

企業分析導入方面變得更聰明

雲數據的遷移、網路應用變得普遍

數據民主提升數據科學家(以前有用的資料很少，但現在資料開放程度變高了)

AI元年

要發展無監督式的學習，因為監督式學習需要大量資料

人工智慧發展站在大數據的基礎上

Open data

政府資料開放平台Data.gov.tw

內政部資料開放平台

台北市政府資料開放平台

交通資料庫(很完整推薦)=>tisvcloud.freeway.gov.tw

Open data nyc

Kaggle(可以練程式不適合分析)因為資料都是乾淨的

網路數據

網站數據=>近來此網站的行為

地理位置=>到網站瀏覽的地理位置

轉介來源=>從哪個網站來，之後去哪個網站

搜尋流量=>連進網站以後相關資料

社群網站=>從哪個社群網站來的

網站分析的追蹤碼=>知道你的動向 可以封鎖

Similar web google analytics ghostery=>都是數據蒐集的好工具

網站追蹤最常用GA(google analytics)

大數據不一定要花很多錢

Ikea 每個商品都有「名字」=>買關鍵字，輸入後第一個呈現

蓋掉對手的網站

Ikea改商品名🡪 she doesn’t want to cuddle

他把在網路上搜尋有關人際關係問題的關鍵字找出來=>當產品名稱

他不花錢買關鍵字直接利用關鍵字邏輯

大數據🡪機械學習🡪人工智慧

三者不可分大數據是資料來源，機械學習是處理方法萃取複雜的規則，再給AI模擬=>餵給AI的數據要夠好成效才會好

維修數據分析預測

列車緊急剎車

號誌天線異常

預測性維護=>為了安全

資訊視覺化分析=>設備視覺化呈現

設備優化=>設備數據關聯分析、設備數據分群分析

異常預測=>異常偵測、預測狀態

設門檻值—當測量超過或低於可接受值會產生警示

但有在可接受範圍內仍然出現故障=>找出每個故障的點

即使值在可接受範圍內，也能檢測到的異常移動

利用過去資料，識別在故障前後會有的狀況

監測故障紀錄數據

透過資料視覺化提升故障紀錄數據的萃取效率進行狀態檢測

智慧維修決策

資料科學

C、C++不適合作資料科學

R、Python適合

程式語言要看你需要再學=>R容易上手

R

用來數據探索、統計分析、作圖的程式語言

R的優勢

1.語法簡單

2.packages很多

R的缺點

跟python比起來在某些情況下效能比較差

R的計算是內存式進行=>要把資料讀進去R

與Python一樣的缺點:packages的原程式碼你沒辦法看=>可能兩種語言跑出來結果不同

但R很自由大家可以自己寫套件，也能用來開發AI

數據對象類型

讀取資料檔

資料可能是變數或是資料表，要去賦予名字

例如:x <- 8 or a <- “city” 可以是數值型or非數值型

變數

類別變數🡪性別、品牌、系級….(根據資料特性或屬性區分資料)

數值變數🡪間斷型變數(計數EX:1個人、2個人…)、連續型變數(常態分配)

包含小數點由測量得來的例如平均數(不會有1.5個人)。但都可以加減

乘除(依據數據尺度分，有scale)間斷型和連續型的圖表表達意義和圖形顯示方式有差異

量測尺度

名目、次序🡺類別變數

等距、比率🡺數值變數

名目:無法排序

次序:可排序

等距:有順序的數值型=>有排序和倍數關係，不能判斷量，原點(0)沒有意義

比率:有順序尺度，可以表達量的大小、順序、有原點(0)

量:可以比較大小差異的

補充:Tableau資料處理軟體

|  |  |  |  |
| --- | --- | --- | --- |
| Q | A | Q | A |
| 結帳時間 | 比率 | 溫度 | 等距 |
| 職業 | 名目 | 年級 | 次序 |
| 電話費 | 比率 | 身高 | 比率 |
| 產品滿意度 | 等距 | 營運績效 | 等距 |
| 畢業學校 | 名目 | 統計學分數 | 等距 |

11/14

STEP1資料預處理

1.缺失值=>是否補值?或空值?

2.異常值=>刪除 but 很多資料的價值來自這個🡪可能只是抽樣誤差

或其實是重要訊息

如何檢測?=>標準化(6 sigma 99.7%資料會在這個range，超出🡪異常)

盒虛圖

資料不符合常態🡺切皮雪夫(對任何分布的資料都能用)

3.不一致值

Ex:測女性是否有小孩

1.全抓出來一個一個問

2.沒結婚🡪沒小孩(這並非正確邏輯)

3.看小孩欄位是否為空(但這欄有錯誤資訊)

4.用年齡可以判斷(但不一定全對)🡺例如15歲結婚有小孩

4.有特殊符號的數據or重複數據

「資料很髒」就是指這種

也有全型字or半型字的問題

缺失值的原因

1.有些訊息暫時無法獲取or取得信息代價太大

2.遺漏🡺採數據的機器壞掉、人為因素(忘記、漏看、私人、理解錯誤…)

或追蹤類資料(第一季~第四季的電話訪問)可能人會不答or消失

一定要拿「事件簿」

例如:銷售量曲線🡪可能有產品出問題or促銷導致下降甚至是補貨不及

🡺沒有事件簿紀錄，後面分析會有問題(你預測他會週期性下降)

有些屬性值不存在

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 年齡 | 性別 | 男 | 女 |
| A | 15 | 男 | 是 | 否 |
| B | 25 | 女 | 否 | 是 |
| C | 30 | 其他 | 否 | 否 |

如果你要用語言判斷🡺年齡性別即可

但如果要跑回歸🡺要用男女的方式(dummy)

處理之前先找出unique鍵來當作串接點=>否則資料會膨脹

缺失值的影響🡺bias，但不一定會有影響

統計缺失值的個數

判斷邏輯性問題🡺ex只有123的選項卻出現5

統計每個變量的未缺失數

統計遺失值的比例

CV變異係數🡺可以看出離散狀況

不能用標準差，可能平均數很大但標準差很小🡺離散小

當CV相等時，才能用標準差判斷

異常值處理

平均數、標準差、偏態、峰態

6 sigma

箱型圖🡪很快就能知道差距多少 R:boxplot(x)

異常值的好壞要判斷、要判斷來源

一致性分析

單一或不同table的方式不同

政府單位資料:先對地址🡪搬家不會即時更新

STEP2數據特徵分析

Ex:百分比、年齡分布、性別…..

1.分布分析

2.對比分析

3.統計量分析

4.週期性分析

5.貢獻度分析

6.相關性分析

更進一步了解資料

社群大數據發展趨勢及應用實例 (意藍資訊)

擅長:語意分析技術、搜尋引擎

以前:要自己買授權、自己電腦裝

今天:amazon雲端訂閱(租)

Opview社群內容資料庫:追蹤儀器

Datamarket.tw

如何令數據變得有價值🡪資料分析師

1.資料有何特性、價值 了解資料性質

2.工具方法應用

3.實際案例

社群大數據

1.網路社群普及、表達意見快、方便

2.影響更多民眾、指標、風向球

3.可聆聽民眾真實的聲音

商業價值:消費者研究、調查 例如mastercard

台灣有七成還在用傳統調查工具

網路口碑影響力提升🡪更換的成本高的商品夠明顯:如電信業、保險、汽車

每個人資源有限🡪為了提升效率慎選商品

或是因為好玩

品牌創造

負面影響>正面影響

口碑數量、一致性 正面口碑+一致性高 = 可信度高

同質性群體發言多🡪可信度高

傳播者的專業程度

網站本身可信度也會影響

個人信任傾向

正面口碑-->影響購買行為(只會增加意願不會立即影響)、增加與客戶良好相處

負面口碑🡪阻斷立即購買行為(焦慮、懷疑)，任何人都能產生殺傷力(無視專業度)，也會影響商家信任度。但要建立需要「一致性」，破壞其一致性就可以

不同管道傳播效果不同、影響族群也不同

6大來源:新聞、討論區、FB、blog、BBS、其他社群網站(youtube、line…)

FB🡪實名制導致講話會有些顧忌，導致沒辦法討論🡪但匿名制同溫層言論真實

部落格🡪FB文章的儲存處 SEO搜尋引擎可以找到

Gogoro來台🡪掀起討論區討論🡪FB關注🡪降價🡪新聞關注

FB、新聞出現的議題很容易一天以內就冷掉🡪他們不會有置頂功能延燒時間短

PTT、討論區的延燒時間長

產生的附加價值🡪搜尋引擎會找的到但FB的更迭太快不好找到

醫生的評價可以找

負面評價要先看

1.聲量

2.出現在四大平台PTT、mobile01、巴哈姆特、卡迪諾 就可能變成全國議題

FB的重要度開始提升

社團變多

粉絲團開始沒落，變成行銷平台

政府單位粉絲團人變少，很好搶

Google map搜尋美食(提供即時、打卡)🡪影響力強

個人網紅影響力變強

歷久不衰的:排行榜、傳播遊戲(ex:冰桶)、開箱評測

平均貼文案讚數:網紅變多，而非官方粉絲團

社群大數據🡪商業用途:行銷、服務、預測

特定議題意見蒐集:例如電子支付🡪安全問題，綁定方式複雜

客群、消費市場研究、公司品牌商品評價、市場警示趨勢預測、產業競爭者分析、行銷成效評估及追蹤

行銷、廣告、公關、與企業營運結合有更多用途

風險管理、投資關係評估、人事

消費品牌尤其需要這種服務

人事部門:高科技產業，找不到人。

離職工程師的po文、不愉快的面試經驗(意見信箱是花瓶)

想知道員工滿意度🡪網路上找特別有用🡪對人才招募有殺傷力

社群大數據分析方法

統計方法不是新的，但應用是新的

用內容調查法🡪焦點團體大題綱自由表達🡪抽取想法得出分析結果

但整理起來麻煩，龐大(時間成本高)

問卷只是因為成本低而已，而且回答意願是個問題，還有問卷有效性

電話問卷還有打電話的問題

網路興起後大量內容可以使用

OPview 國內最大雲端網路輿論觀測中心

資料處理、垃圾文章

前台介面分析、語意分析技術(機器自動分類標記) OPview觀測平台

關注議題聲量大小、網路輿論的詞彙、討論的地方、作者是誰、性別?、立場?

🡺議題、來源、發言者

量化 關注度

正負評比 好感度(負、正、中立)

頻道數 廣度(同溫層or公眾)

回應比 擴散動能(回應多寡、行銷)

作者 意見領袖

族群 族群樣貌

關鍵字 重點

議題 詳細

頻道數:摩斯、肯德基、麥當勞🡺廣度肯德基高，但聲量麥當勞大所以肯德基擴散力強

麥當勞的回應來源以FB為主 eney討論區。投資在影響力最大的來源

作者:提供影響力、表達意見。看作者目的評估

族群:子女、上班族、結婚與否

可以從字裡行間的關鍵字去分析一個人的狀態。

關鍵字:例如電視:基本安裝、hdmi…

議題:討論家電:噪音、安裝費、省電or臭味……等等

危機事件警示及追蹤、案例分享

華航機師罷工:為何要選過年時

八月份有發布、一月分又發布預告、快放假又報

到了放假後:直接吵起來🡪罷工最佳時機

很多記者都掛在PTT或FB等消息🡺推波助瀾

Line today的影響力也漸漸升高

分析:嚴重程度、該不該出手🡺要看誰才是主要影響人

Ex:交通問題統計🡪交通部🡪總統府….勞動部排名很後面

根據統計，網友的點名會影響到是否需要踹共

在不對的時間講不對的話導致問題被放大

一場危機事件壓過形象廣告的效果

停飛事件的子議題

1.濫用迫降

2.董座發言導致第二波聲量

3.員工求去引發更多人的追究

網友前後態度大不同

1.傳出空姐網路控訴🡺網路傾向罵空姐颱風天上班大有人在

2.董座可能因為這篇文發言🡺網路傾向開始同情員工

網路效應:欺強扶弱🡪空姐發文(相對其他工作強勢)🡪老闆發文(公司高層員工相對弱)

評估行銷成效

這個活動在網路有多大的討論是一個客觀指標

屏東燈會:在數據上、媒體上是成功的

有多成功是可以量化的

討論與關注比去年增加一倍以上

擴散度:作者數、頻道數也高於嘉義燈會🡺關注的人多、且廣

燈會結束後的討論聲量持續多久

P/N也是指標之一 =>看各面向的好壞占比

行銷活動是一種手段而非目的

例如屏東燈會🡺目的是帶動觀光 🡺評估有沒有帶動當地觀光討論

彩妝代言人行銷效益分析

品牌 VS 代言人聲量及好感度變化

代言人能對於品牌帶來的聲量不高，好感度也升的不高

能帶來多少新的頻道，有沒有新的族群近來討論

引發新的族群討論效果很有限(傳統操作方式幫助不大)

網紅帶動品牌效果更好

結果是操作方式的問題(行銷的創意表現)

產業研究、競爭分析

連鎖眼鏡

先看他是熱絡的產業or慢慢消退

成長部分都來自社群網站🡪FB為主IG、youtube 🡪可看出他的行銷方式還是傳統

重點在:好感度、價格、鏡片機能、清潔

傳統眼鏡公司:小林、寶島、台灣眼鏡

快時尚:Lohas jins🡺講究快速，不用比價

聲量表現:lohas、寶島聲量最大🡪但來源最多都是官方粉絲團

真實來自外部的只有owndays 更多來自非官方粉絲團來自IG占比大

傳播策略:寶島用網紅為主優惠方法 小林偏向用官方粉絲 其他都是自然推薦

德恩堂也以粉絲團為主、lohas也是粉絲團也有找網紅操作傳統

Jins以PTT討論區為主，比較少行銷自然發展

Owndays利用行銷活動、網紅合作、業配貼文(他號召IG網紅超過一萬，IG粉絲量可以折抵)馬上配眼鏡馬上分享，網紅帶來的按讚數不一定

不能只看粉絲數多寡，也要看他的成長速度以及網紅活動力(要看近期指標)還有擅長的商品、主題領域

Social lab 訂閱電子報可以看到很多案例

進階應用

利用社群資料來預測Ex:總統候選

票房預測、股票預測，還在實驗中但評估效率很高

能和真實行為有很高的連動性，但沒辦法驗證，之後公司賺多少不會跟你說

銷售量與曝光度是有高相關度的

信用卡發卡數與聲量有相關，頻道數、作者數為最主要相關(擴散度高、討論度高)🡺容易發卡

高麗菜價是否會跌

預測未來七天走勢

菜價有季節性效應，但最大變因來自心理預期🡺大家覺得會跌他就一定會跌

因為大家都不買所以必須跌價

只要有人網路上先講就有影響，找出27個變因

徵信

可從網路上看到勞資的和諧度、資安風險、企業營運體質好壞

預知災難

將社群當作災害的觀測來源

網路做災情蒐集也是有困難度🡪要15分鐘之內

災難也有很多雜訊🡪去年「淹水」or開玩笑的導致句子意義有雜訊

透過人事時地物、地標去判斷災情位置

Social-lab、OPview

11/28

設定關鍵字、社群資料概論

IG、youtube新的傳播方式

傳統調查特色、限制

依照設定的問題與選項回答🡪固定

但無法涵蓋母體且問問題方式會有偏誤

拒訪者、受訪者問題

金錢與範圍人力問題

網路網友的回覆不受限制且有即時性、內容形式多元

缺點是資料雜亂，要重新編碼，不同情況有不同意義(情緒判斷準確度)

Opview🡪把非結構資料變成結構資料

無法接觸使用資料人群、以及潛水的使用者

如何判斷影響力? 🡪案讚數、分享數、留言數 判斷效益好壞

新聞🡪瀏覽數

客戶常見問題

來源/網站(PTT、巴哈姆特…)/頻道(xx版)

討論區、社群網站、問答網站(知識+)、部落格、新聞(專欄文章)

台灣地區公開資料都有收入

24HR爬文、只收文字資料

Opview🡪布林邏輯(用符號串接詞語之間意義)🡪但要學習

Google🡪自然語言查詢(他自己猜你的意思)好上手 但不精準(字組要同時符合或不用?)

想要的資料

or🡪| ex:50嵐|五十嵐

and 🡪 &

not 🡪 !

優先運算🡪() ex 抹茶&(50嵐|五十嵐)

? 🡪考慮錯字、別字 特定結構用詞 一般任意文字可適用 ex:酷?沙

\* 🡪 可塞0~5個任意字元

布林邏輯要用半形才有用

英文用法

空格也是獨立字元 happygo|happy go

大小寫無差異

符號有全半形差別:全形才是為字元 半形是布林邏輯

\*可當字首或字尾

Icash\*點數 🡪 icash開頭的單字都會搜

Icash \*點數 🡪 空一格代表單字已結束

高雄捷運主題如何下關鍵字

高雄捷運(一定要連在一起)最嚴謹

高雄&捷運(最不嚴謹，不管多長有出現這兩字就撈)

高雄\*捷運(包含高雄捷運但有一些雜訊)ex:高雄來台北坐捷運有打折

如何排除雜訊?

用!

觀測🡪文章列表🡪排除文章

觀測🡪媒介🡪網站/頻道🡪排除網站/頻道

觀測🡪領袖🡪關鍵領袖🡪排除作者

1.主題是啥

2.是否有其他說法別稱

3.是否需聚焦特定概念

4.雜訊處理

一次只能查三個月資料

主文:主動發文方

回文:對主文的回覆

🡪可分辨正面詞彙、負面詞彙

可做頻道網站篩選

最相關🡪熱門度

擴散🡪看出事件的嚴重性與傳播來源

產業研究

根據市場行為、結構分析給予企業發展方向

Michael Porter 五力分析模型

同業競爭者:哪個品牌最受關注、因怎樣的話題引發關注、好感度、哪個品牌與自己的關係最強烈

消費者:討論重點、在意面向、哪些品牌最常被提及

替代品:比較聲量、比較產品間的討論都在哪些頻道、都討論些什麼

缺失值

資料無法獲取、蒐集過程被遺漏、獲取成本高、根本不存在

用補值、推估或忽略 估計出來的風險值有多高

遺失資訊導致模型偏誤

敘述性統計相互比較🡪可得缺失數量、缺失率(資料品質能不能接受)

異常值

判斷好壞

簡單統計量分析、3sigma、箱型圖、偏態峰態(是否為常態分配)

區間推估🡪99.7%會在3sigma以內 若在外就是異常值

檢視異常值的問題在哪裡、從箱型圖來看峰態、分布

不一致值

數據的一致性，提升數據品質(補值的準確性、一致性檢測)

發現不一致的時候如何處理

數據特徵分析:分布分析、對比分析、統計量分析、周期性分析、貢獻度分析、

相關性分析

週期性分析:觀察數據有沒有週期性

分布分析:最直觀用圖形表達🡪算數量

對數據分組、組距多少、分幾組要考慮

求極差(全距)🡪得出組距要多少宇組數🡪決定分點🡪列出頻率分布(落點、

占比)🡪繪製頻率直方圖(附上累積頻率，帕雷托圖)

各組之間需要周延互斥、組寬相等、各組包含所有數據

帕雷托法則🡪又稱二八定律80%的結果取決於20%的原因

圖形的目的性(例如直方圖沒有間隔🡪數值型資料)

頻率、次數為何要互相轉換?(頻率🡪標準化)

對比分析(了解差異):相對數比較、絕對數比較

相對數比較

1.結構相對數:可以表達比例關係，例如某產品所佔的消費支出總額跟其他

產品比較，是看總體的不是單項。比較兩人消費結構差異(就會包含食衣住行各項兩人的比例)

2.比例相對數:比較各種部份的比例關係，例如人口的性別比，數值型比例

相對比較

3.比較相對數:同一個時期兩個性質相同，相同指標。但空間區隔來比較。

例如不同地區的產品價格對比

4.強度相對數:強度(密度)，例如人均、每平方米有多少的農作。熱力圖呈

現。

5.計畫完成度相對數:某一時間實際完成數

6.動態相對數:同一現象在不同時期的指標數值對比

對比分析

Lines(x,col=1)

統計量分析

集中趨勢:均值、中位數、眾數

離散趨勢:全距

平均值:最常用的，但容易受極端值影響

中位數:排除極值影響，但會犧牲掉離散的影響

眾數:不受極值影響，但可能會有很多眾數或是沒有眾數的情況

全距:最大-最小的差，對極值非常敏感

標準差:最常見的變異測量方式(大家對於平均數的距離)

變異係數:CV

週期性分析:

趨勢如何?上或下

有沒有週期?可預測可解釋，或是不可預測不可解釋

🡪為了要做預測🡪量化、質化預測方法 質化:主觀判斷、定性方法，沒有

歷史資料 量化:用歷史資料預測，尤其用時間序列，找出變動規則

定量分析法:

因果分析法(例如身高對體重關係)、回歸分析法

平滑技術法:把不規則變動抹平(把長短互補)減少誤差的風險

時間序列分析法:要去了解資料特性，才能用正確的時間序列模型預測

由x的觀察值找出過去一段時間的行徑圖案(Pattern)

再用圖案進行預測。優點是只要用過去資料就能對

變數做描述與預測。缺點是你必須要有基本假設，

未來是過去的延伸，要假設「圖案」趨勢會重複出現🡪過去的特性極為重

要，也找出哪些經驗是要到下一個階段去增加準確性

只有時間序列的資料，才會允許其相關性

連續型時間數列:例如地震儀，連續型序列

離散行時間數列:例如每天中午的電力需求，間斷型

時間序列型資料:例如每年降雨量、每年出生率、自殺率犯罪率…都可還原成時

間序列資料

但不是每種時間序列資料都適合用時間序列分析🡪有沒有規律

性(Pattern)

每一個時間都有觀察值，可以有各種維度

時間序列圖:折線圖類，用二維圖形表示時間序列資料，垂直是目標變數，水平

軸是時間變化

平穩型數列:

落在某個區間去做變動的資料🡪容易預測也可以判斷outlier

但這種類型的數列不多🡪好處理但價值不高

非平穩型數列:

很常見，甚至看起來像隨機的。也可能是指數型變化的

很挑戰

季節性時間數列🡪有固定的擺動方式

動態回歸模式🡪指數型型態🡪每一期可能都受前面幾期的影響

時間序列

趨勢變動:長期變化🡪長期的增加或減少(整體方向)，觀察長期資料(不都是

線性的)是函數型態的

季節變動:短期變化🡪不超過一年的擺動，是規律性的變動，也會看月份趨

勢

例如可以拿機票的來做分析，把全部票價爬出來做整理

預估星期幾適合買(最便宜)

有規律的波動、周期固定、每年重覆出現、各年變化幅度略相同

用歷史資料去預測需求高低🡪動態定價 ex:Airbnb

但機票沒有動態定價🡪分析各種座位會有10%的價格落在哪些區間

為何會有你剛買完價格變低的狀況(台灣有旅行社包票的問題)

成長是加速度(熱門航班)、成長慢(看空位、預測便宜的機票)

循環變動:

介於長和短之間，較長時間的規律性變動🡪但時間區間不一定相

同，高點到高點或低古到低谷出現的時機

不規則變動:

無法預測、隨機的波動

造成的原因可能是自然的、意外或偶然(需要事件簿來去除雜訊)

時間序列的雜訊

從圖形去猜特性

例如電量分析:工廠修工就會掉到低谷但用電會達到高峰

貢獻度分析

帕雷托圖表達又稱二八定律

看每個項目的貢獻值，算頻率和次數🡪要做排序，最多人用排序在最左邊

折線圖有幾項加起來就有八十趴(累積百分比)

累積百分比的用途就在這🡪可以判斷出他的資料集中狀態

我有80%的服務使用集中在60%的服務項目中🡪代表沒有過度集中在某些

服務

最常拿來測貧富差距

相關性分析

X、Y一起變動的量的大小🡪變動的強度

若是數值型變數🡪回歸

畫散點圖矩陣(每個圖相對地散布情況)、散點圖或是算相關係數

把input 放入有很明確產出🡪函數關係 若沒有🡪統計型關係(統計預估結

果)

散布圖就是要看他的關係🡪看他非線性的狀況

相關只有低度或高度(-1~+1)，很難找到完全無相關

皮爾森相關 r(幾乎都是這個)

什麼時候用? 假設X、Y都是常態分佈，透過這樣的形態找

但通常不會知道資料分布狀態，而且檢查不太出來

spearman相關

不知道X、Y的型態的時候用

設 X身高 Y 體重🡪同一人 配對型資料(pannel)

可以算身高與體重的相關係數

先將資料排序，然後看相關性

看常態分佈🡪經驗法則 or 切皮雪夫

Kendal 相關

把相關數值轉換成級距(順序尺度)

要挑變數的相關程度作分析🡪不能用錯相關分析

貢獻度分析

帕雷托圖，右邊的尺度是反映折線的尺度🡪要累積百分比

作了主要項目排序、用二八定律看資料的集中程度

期望值

作預測🡪一般性(大家最一般化的狀況)

平均🡪常態分布 高度代表機率🡪平均數出現的機率最高 🡪長期以來=0

在一條平線上去波動🡺平滑技術法(要平滑幾期是重點)🡪變成直線

這要看各期的隨機性來判斷

數據預處理

認識你的資料(60%~80%)的時間在處理資料

將無關、重複的數據處理掉🡪很多不是唯一值作串接

平滑雜訊(outlier)

處理遺失值

遺失值填補

近似法:用眾數、中位數、平均數來補

Ex:薪資要用中位數來呈現(平均數會受極端值影響)

排除法:

把有遺漏值的資料排除(可能會失去珍貴資料)

回歸方法:

找到一個大家通用的方程式，再把有缺失的項目數據丟進去模型，就

能得到缺失值

但是資料間相關性要高，但要獨立

插值法:

內插法:轉換成z值再去折算

缺失值分析

Na值

is.na(x)

anyNA(x,recursive=FALSE)

na.omit(x)

complete.cases(x)

Lagrange插值法

內插法假設🡪 每一筆數據都是線性，用近似值去求

但數據不會都是線性🡪用這個方式不準確

所以lagrange跟牛頓就是將這個問題解決🡪變成曲線

拉格朗日求出方程式後🡪把缺失的帶進去就可以得到

拉格朗日例題

(4,10)(5,5)(6,9)🡺求方程式

10[(x-5)(x-6)/(4-5)(4-6)]+5[(x-4)(x-6)/(5-4)(5-6)]+9[(x-4)(x-5)/(6-4)(6-5)]

🡺 10[(x^2-11x+30)/2]-5[(x^2-10x+24)]+9[(x^2-9x+20)/2]

🡺 5x^2-55x+150-5x^2-50x-120+(9x^2-81x+180)/2

🡺 -105x+30+4.5x^2-40.5x+90

🡺 4.5x^2-145.5x+120

牛頓插值法

不斷重新設計方程式值到符合所有點

異常值

1.刪除異常值

2.當作缺失值

3.平均值修正(快速解決)

4.不處理(異常值分析後覺得沒影響🡪根據異常值數量來判斷)3~5%

若到了30~40%就會有很大誤差🡪補值或不用

補值:用回歸(但是變數之間必須獨立、分布要常態、標準差要同值)

數據變換

正規化

1.把最大值-最小值做為比較基準點

2.轉換成Z值的型態

下禮拜題目

1.週期

2.貢獻

資料類型辨別

人工智慧含意、類型

分析類型(敘述型分析…)

巨量資料的特性+例子

計算題(內插法、拉格朗日)

時間序列相關邏輯概念

資料的特徵特性