兩位跨域者的深度學習之路

游為翔 & 楊証琨
中研院資訊所資料洞察實驗室
@ 台灣人工智慧年會 – 2017/11/09

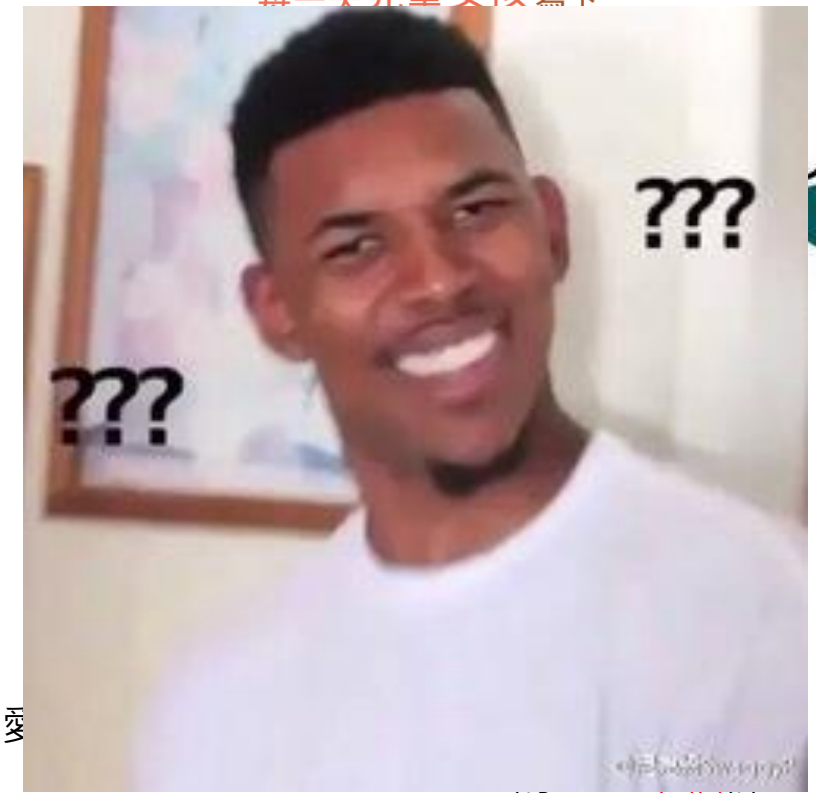
講者

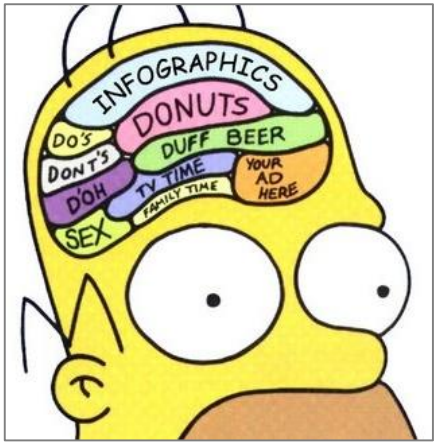
游為翔 Sean Yu

楊証琨 Jimmy Yang

	游為翔 Sean Yu	楊証琨 Jimmy Yang
背景	心理學 / 神經科學	土木工程 / 交通
專長	✓消費及使用者行為分析 ✓深度學習於影像辨識研究	✓ 社群媒體資料處理分析 ✓ 製造工業影像資料分析

心理系到底在學什麼





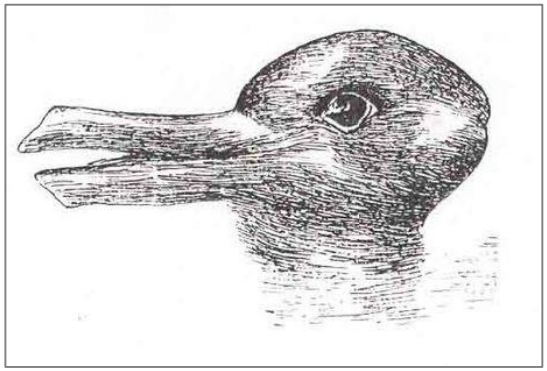
把妹達人
正妹心理學
心理學家的專業把妹術
搭訕聖經
正妹沒告訴你的事

真愛

貼心的女人，幸福無敵：改變男人的賀爾蒙，從貼心做起

各種行為觀察與實驗

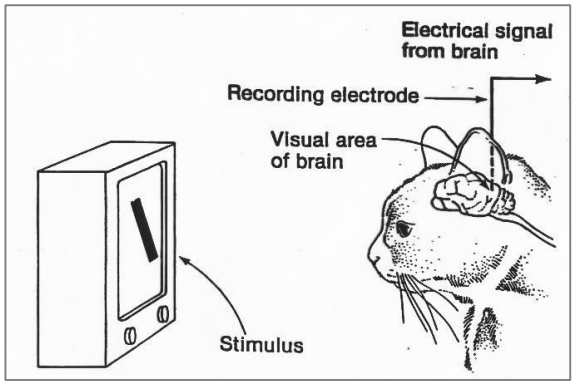
Bandura, 1961



Shifting Gestalt
Jastow, 1899

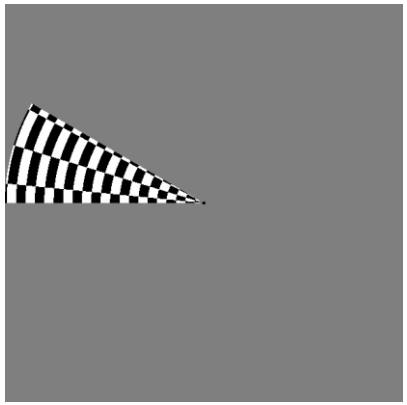
Do yuo fnid tihs smilpe to raed? Bceuase of the phaonmneal pweor of the hmuan mnid, msot plepoe do.

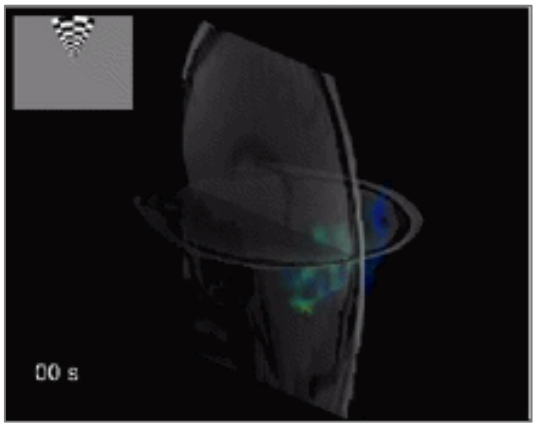
Jumbled letter
Warrington, 1980

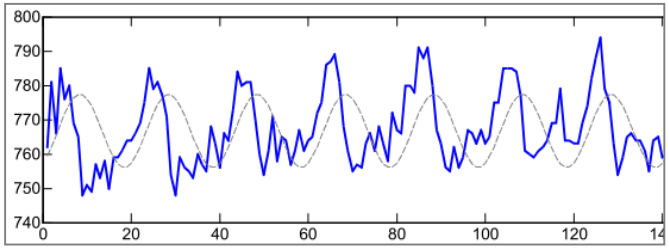


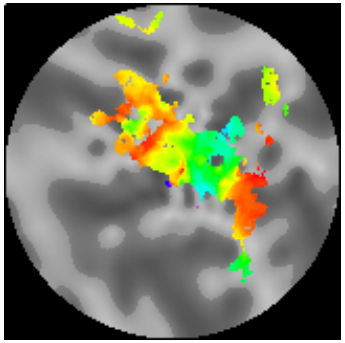
Receptive fields
Hubel & Wiesel, 1959

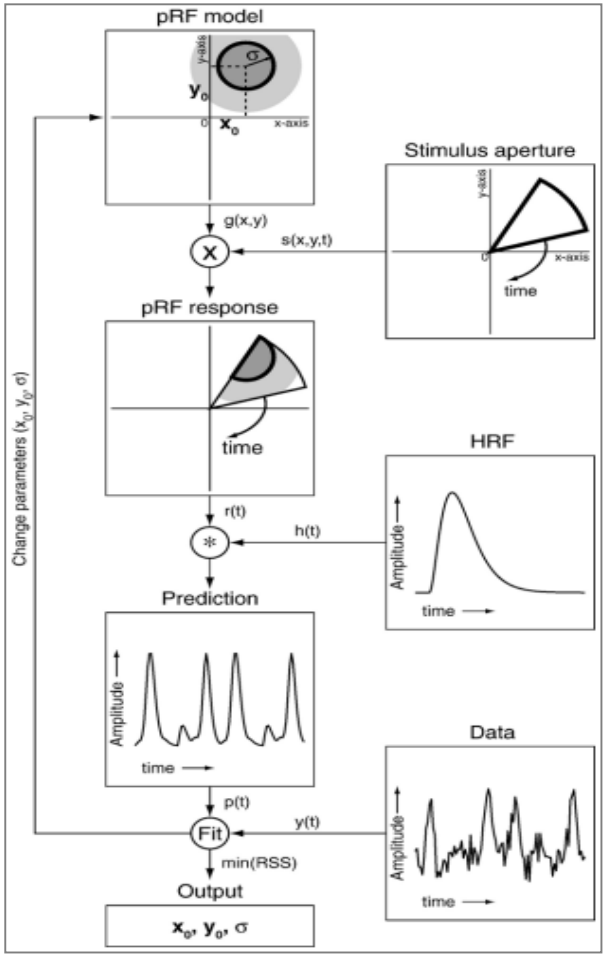
實際上，我過去的實驗室日常 ...











為何來到這裡?

Bridge the Gap between Human Brain and Artificial Intelligence

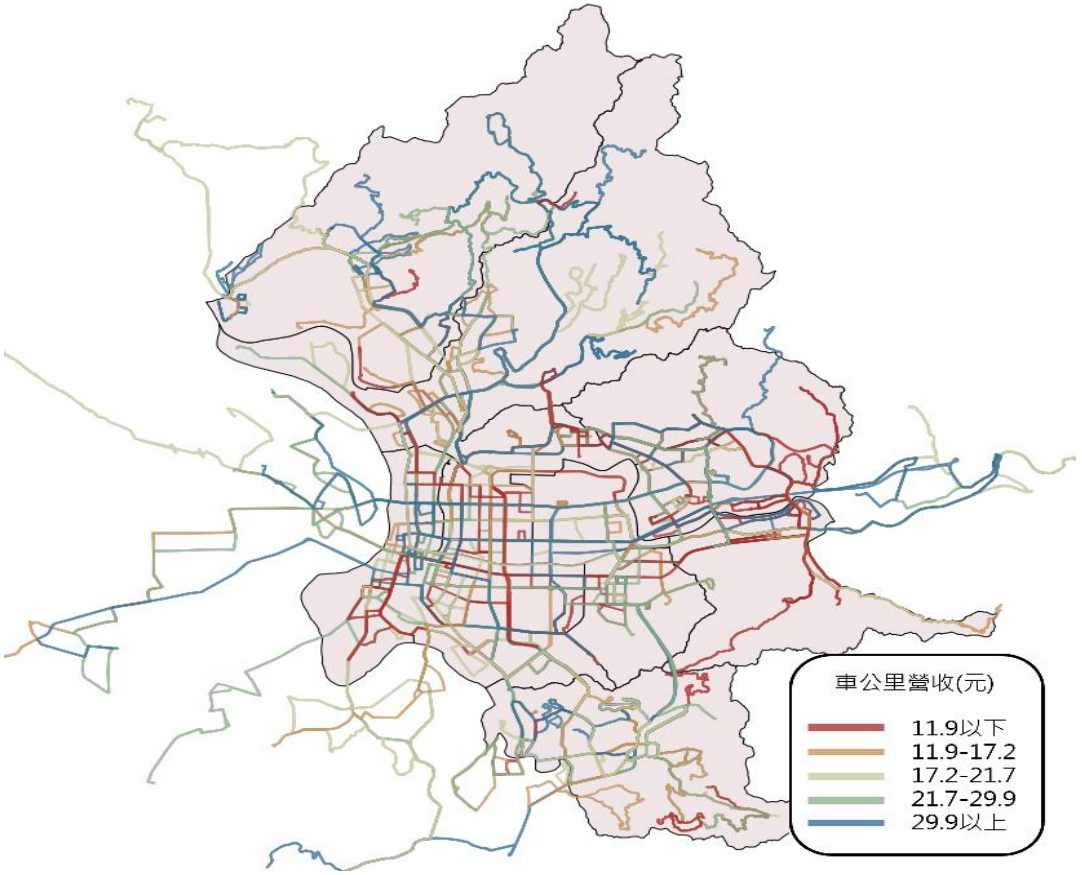
另一個故事

過去在做甚麼

- 背景

交大運管系　　台大土木所

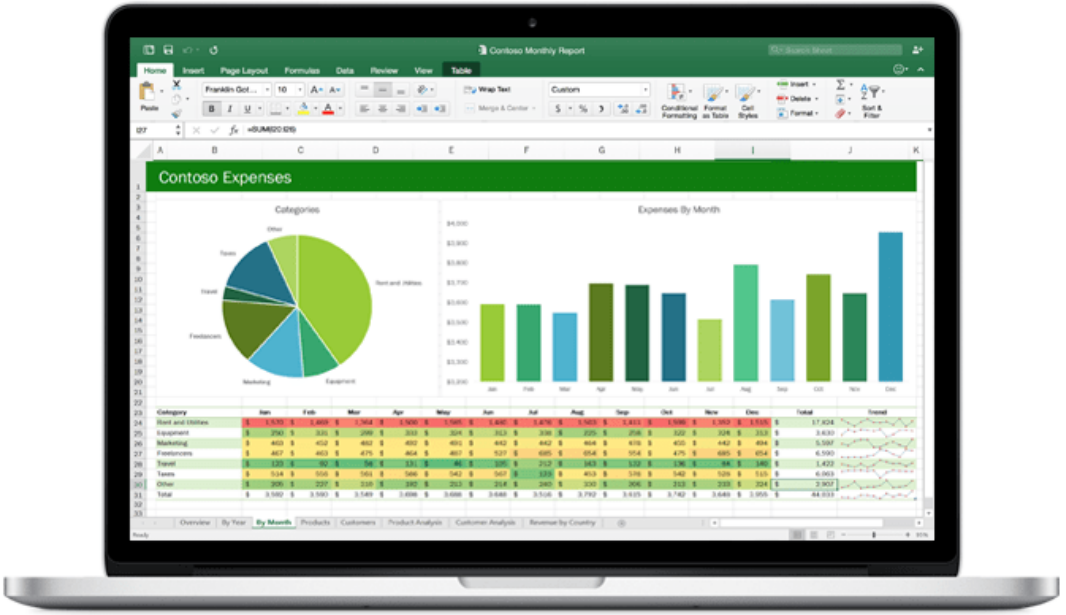
- 台北市虧損公車路線營運分析
  - 真的虧錢?
  - 為什麼虧錢?



技術能力

- Coding 經驗
  - 大一計算機概論 → 6 學分

- 資料分析技術
  - 基礎統計分析
  - ~~Machine learning~~
  - ~~Deep learning~~



FAQ

Why 為什麼想要轉換跑道?

How 非本科系的要怎麼做?

What 想跨入這領域該做甚麼?

Why 為什麼想要轉換跑道?

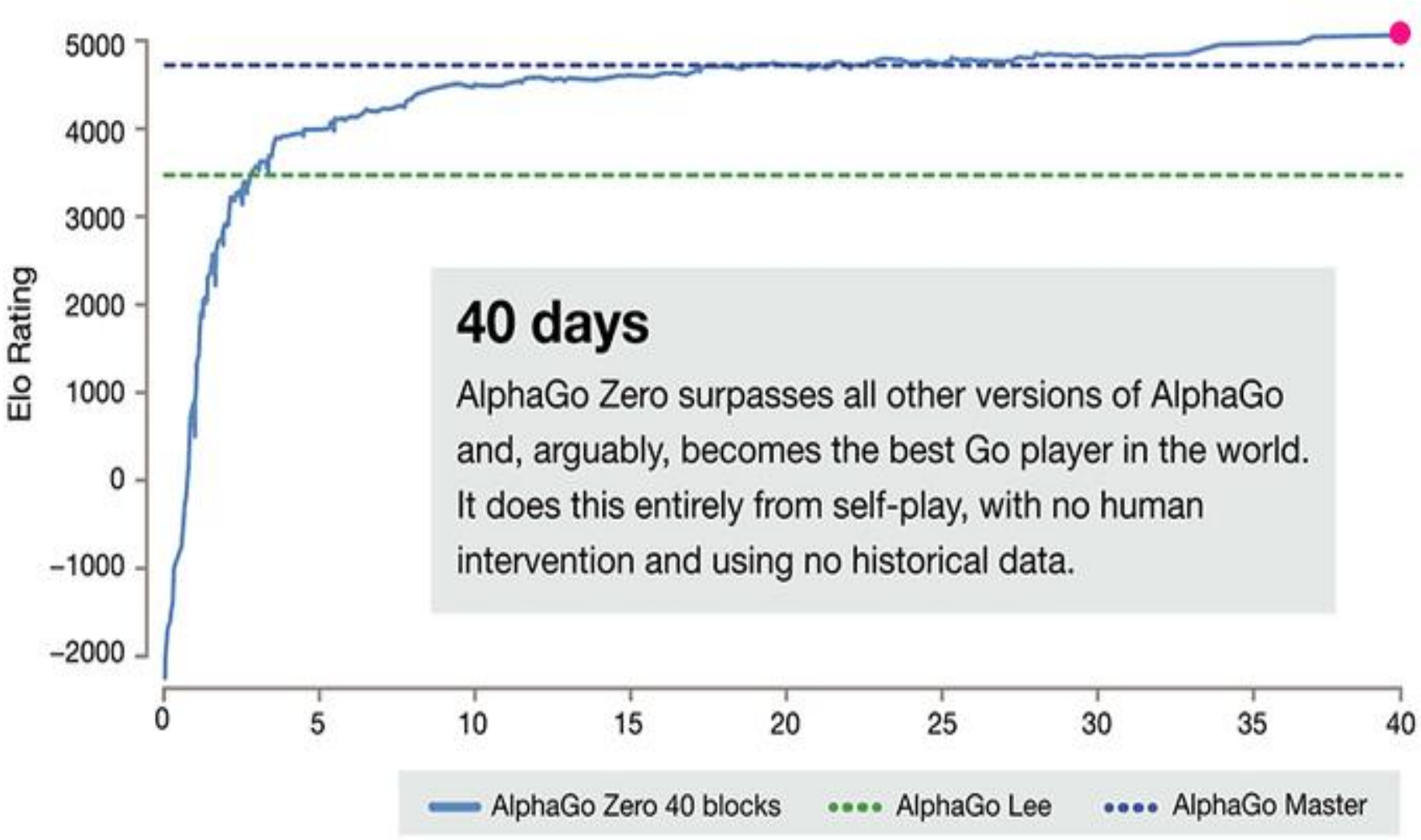
How 非本科系的要怎麼做?

What 想跨入這領域該做甚麼?

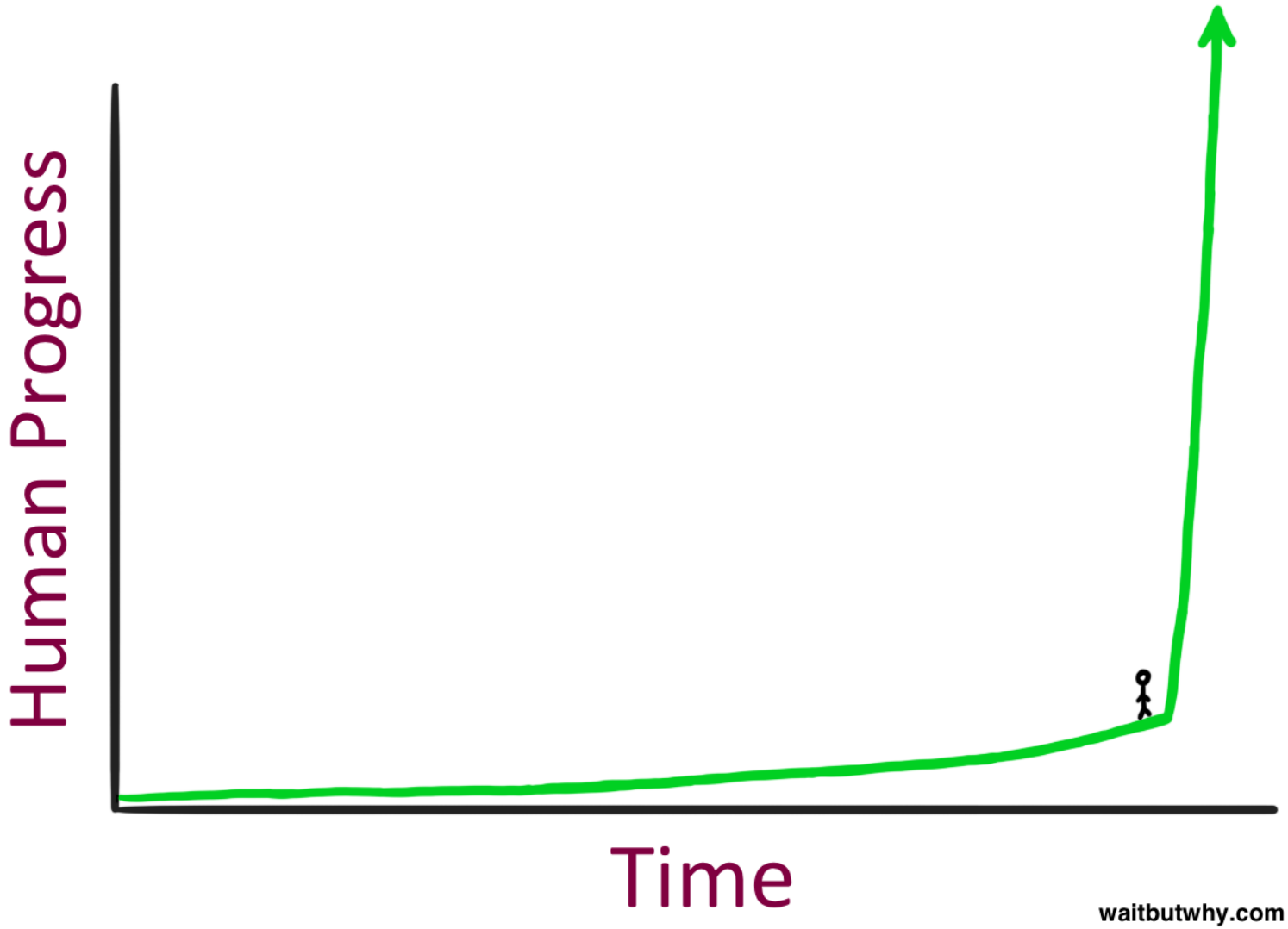
why data science?

- *卷積神經網路 (Convolutional Neural Network, CNN) ?*
- *深度強化學習 (Deep Q-Learning, DQN) ?*
- *蒙地卡羅樹狀搜尋演算法 (Monte Carlo Tree Search) ?*

Now? AlphaGo Zero!



科技進展



https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

Why 為什麼想要轉換跑道?

How 非本科系的要怎麼做?

What 想跨入這領域該做甚麼?

MOOC!

Code

DATAQUEST

codecademy

DataCamp

Project

coursera

UDACITY

edX

該選擇甚麼語言?

- 世紀之爭: R vs. Python

R 的優點

- 資料整理與清洗非常簡潔

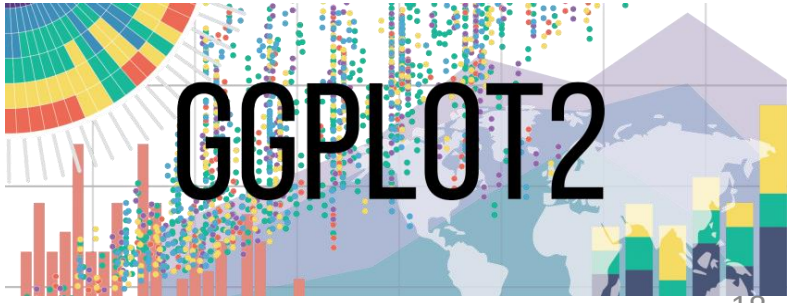
- 快速的統計建模

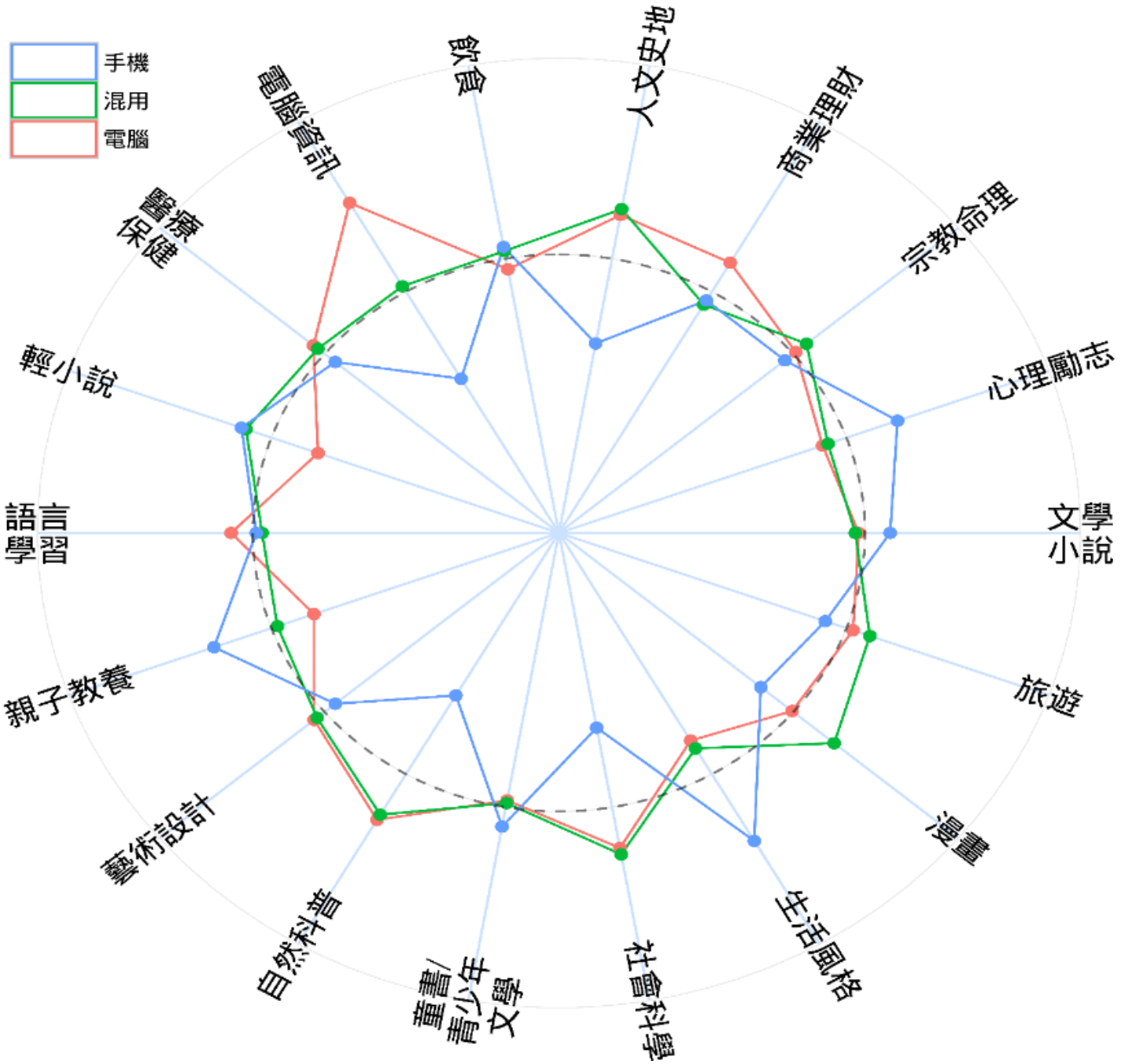
Linear Regression in R!!!

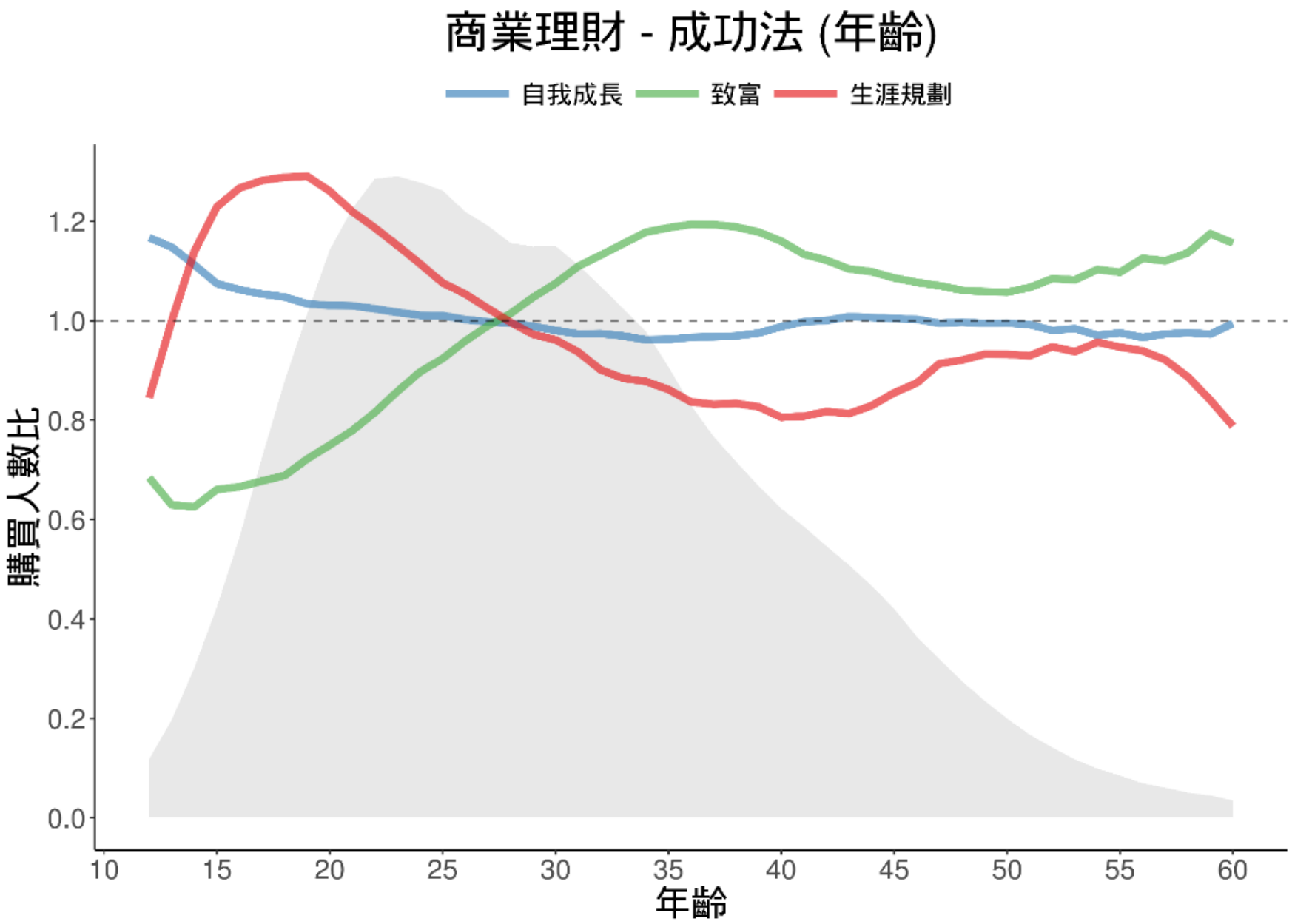
```
Call:
lm(formula = size ~ weight, data = mouse.data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5482 -0.8037  0.1186  0.6186  1.8852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5813     0.9647   0.603   0.5658
weight        0.7778     0.2334   3.332   0.0126 *
```

So Easy!!!!

- 強大的 ggplot2




手機
混用
電腦
文學小說
心理勵志
宗教命理
商業理財
人文史地
勵志
電腦資訊
醫療保健
輕小說
語言學習
親子教養
藝術設計
自然科普
青/青少年文學
社會科學
生活風格
漫畫
旅遊


商業理財 - 成功法 (年齢)
自我成長
致富
生涯規劃
購買人數比
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10
15
20
25
30
35
40
45
50
55
60
年齢

百搭
日記
初學
彩繪
甜美
造型
縫紉
天然
手帳
禮物
超可愛
手工皂
美腿
繽紛
毛氈
生活
瘦身
創意
我愛
美麗
巴黎
一個
曲線
簡單
雜貨
立體
針法
串珠
編織
小物
花樣
寶貝
愛上
鉤織
光療
吊飾
鉤針
黏土
設計
手作
女孩
優雅
織布
幸福
法式
實用
美甲
飾品
可愛
圖案
毛線
彩妝
ok
拼布包
甜點
玩偶
韓國
女王
提袋
娃娃
刺繡
拼布
學者
蕾絲
動物
布作
人氣
手縫
卡片
提案
時尚
布包
開心
最愛
學會
時光
OK
手感
公斤
包包


Feature importance in each reaction
Reaction
V4 5.45 60.31 46.81 24.92 5.34 3.56 0.31
V12 2.09 18.32 78.64 40.63 1.15 7.33 0.1
V9 3.98 13.51 100 1.68
V2 35.71 60.73 16.02 0.31
V16 2.3 7.33 75.5 0.31
V8 2.3 9.84 4.19 34.35 0.73 2.62 0.42
V6 2.09 14.03 16.13 17.91 0.94 2.41 0.21
V1 1.99 22.51 16.13 7.85 1.68 3.66 0.1
V13 1.68 13.82 20.94 10.05 0.63 5.13 0.21
V5 1.68 12.57 16.13 7.02 0.73 2.2 0.1
V19 6.28 6.91 2.83 5.86 0.73 1.15 3.35
V17 2.93 7.96 2.72 6.81 1.15 1.47 0.73
V10 2.3 6.49 6.28 5.86 0.63 1.57 0.42
V14 2.93 5.76 2.51 5.55 0.63 1.47 0.84
V15 2.83 5.55 2.83 4.82 0.52 1.57 0.84
V20 2.2 6.91 1.99 3.66 2.41 0.94 0.31
V11 2.41 4.71 2.3 4.5 0.73 1.26 0.73
V3 2.51 4.92 2.2 4.19 0.63 1.15 0.52
V7 2.41 5.03 1.88 3.87 0.73 0.94 0.42
intra.Kcat.Vmax
intra.Km.Kcat
intra.Km.Km
intra.Km.Vmax
Kcat
Km
Vmax
Effect
Positive
None
Negative

Python 的優點

- 許多套件支援連接資料庫

- 成熟的網路爬蟲套件

- 深度學習最常用的語言

如何選擇?

- 常需做探索式資料分析 (EDA)、資料視覺化與統計建模

- 網路爬蟲、串接資料庫或是做 deep learning

常被問的問題

Why 為什麼想要轉換跑道?

How 非本科系的要怎麼做?

What 想跨入這領域該做甚麼?

想跨入這領域該做甚麼?

- 做 Project!

- Scikit-learn toy datasets

鳶尾花資料集 IRIS dataset	鐵達尼號資料集 titanic dataset
波士頓房價資料集 BOSTON dataset	手寫數字資料集 MNIST dataset

嫌資料集太簡單?

- Kaggle 是一個 data science 競賽平台。企業和研究者可在其上發布資料集，讓全球專家可在其上進行競賽以產生最好的模型
  - Heritage Health Prize，獎金高達 300 萬美金

- **11/12(日) 11:20-12:05 我在 Kaggle 數海獅**
  - 使用深度學習方法協助 NOAA 生物學家監控阿留申群島的海獅數量
  - 勇奪世界第一!

My project in
資料洞察實驗室

線上交友詐騙帳號分析

- 線上交友平台充斥詐騙帳號，如何透過機器學習技術，揪出這些斂財騙子，拯救悲情男女的荷包?

用戶資訊

- ✓ 性別
- ✓ 年齡
- ✓ 學歷
- ✓ 婚姻態度
- ✓ 年收入
- ✓ 帳號資訊

FB 資訊

- ✓ 粉絲團
- ✓ 大頭貼建立時間
- ✓ 好友

建立 features

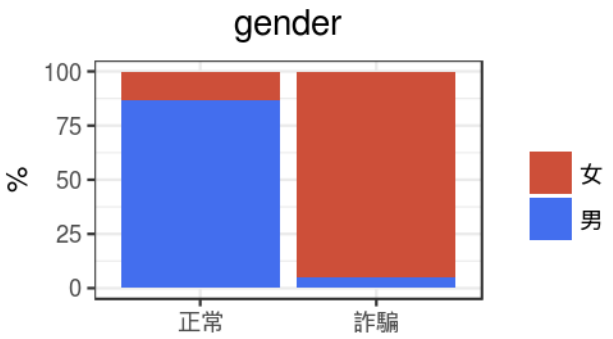
- 觀察詐騙用戶與正常用戶在我們建立的 features 上是否會有所不同。

正常用戶

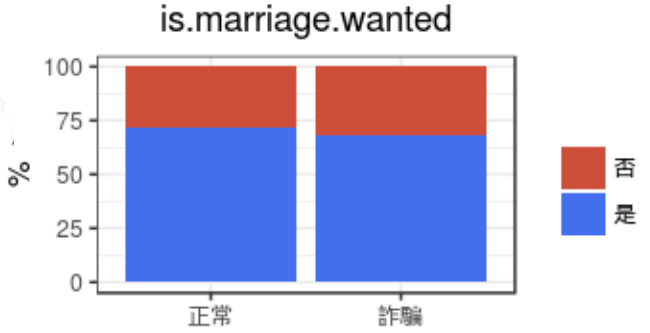
詐騙用戶

詐騙用戶與正常用戶之差異 – 用戶資訊

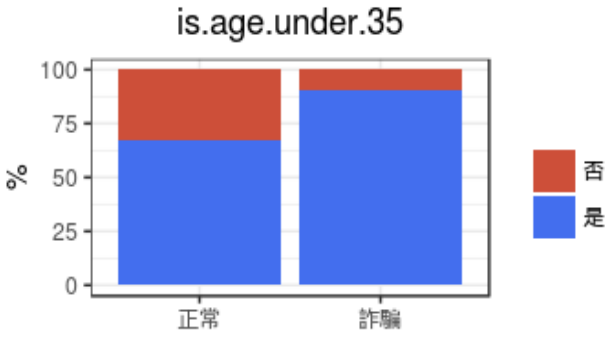
性別



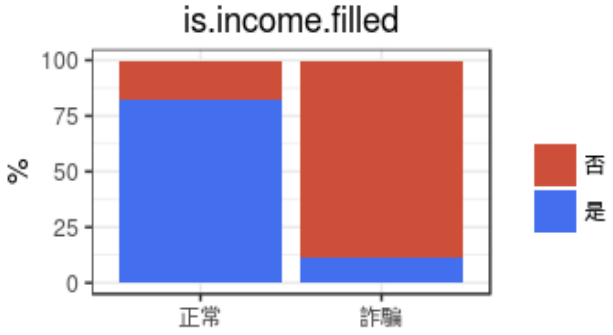
婚姻態度



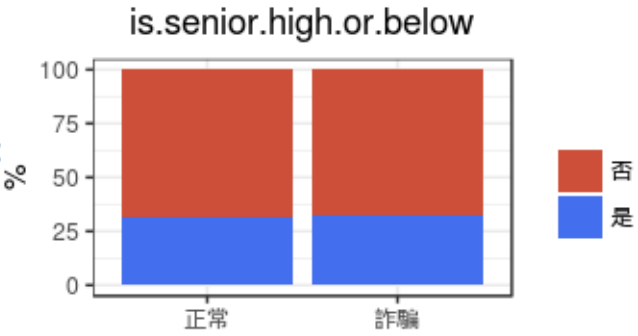
年齡



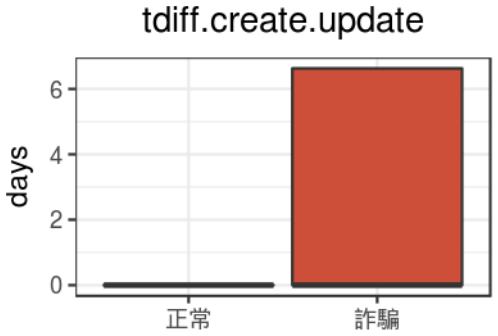
年收入



學歷

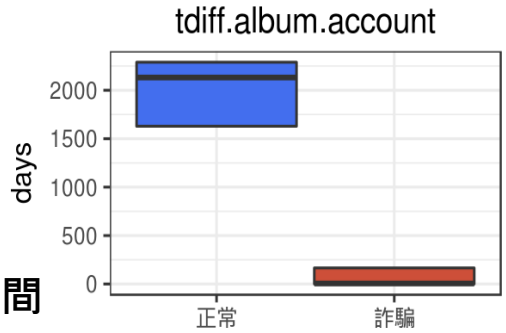


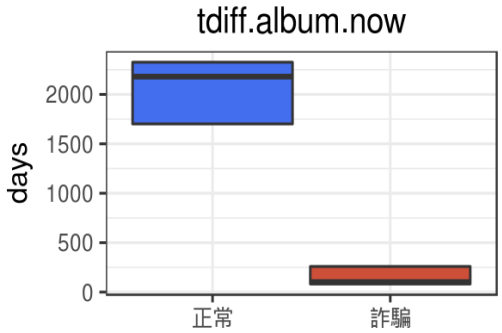
帳號資訊



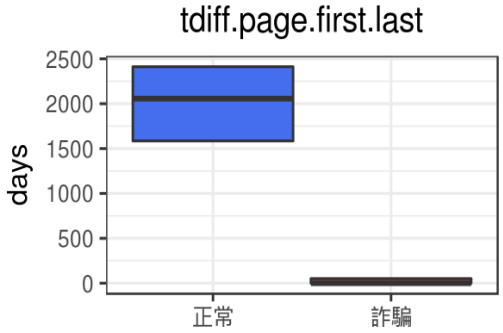
詐騙用戶與正常用戶之差異 – FB 資訊

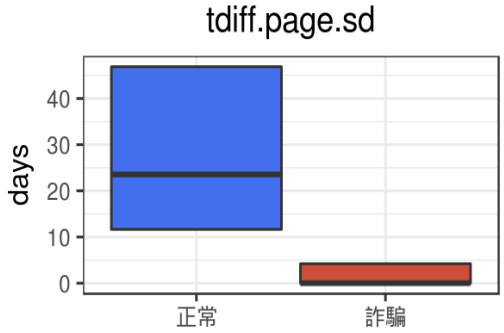
大頭貼建立時間



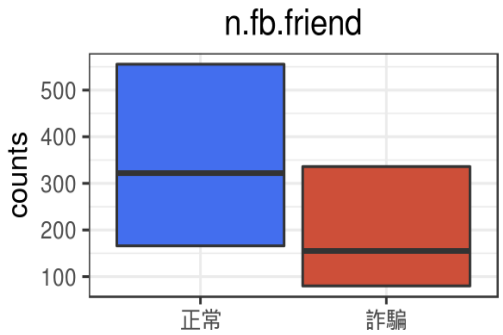


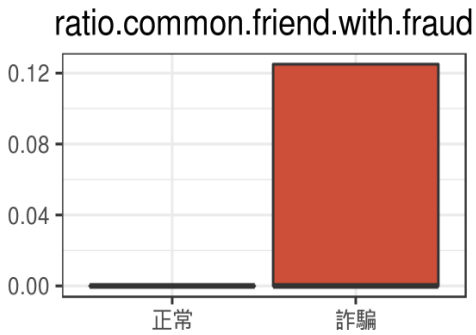
粉絲團





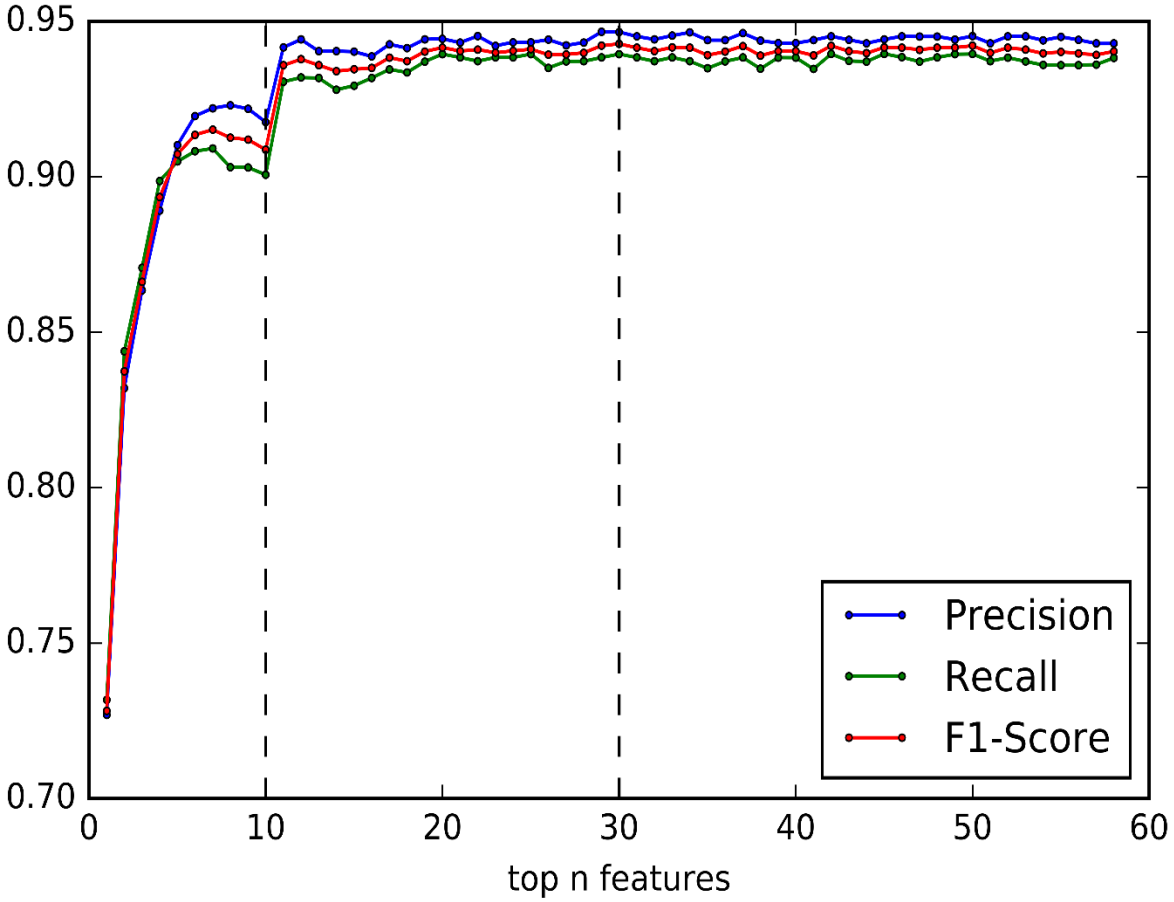
好友





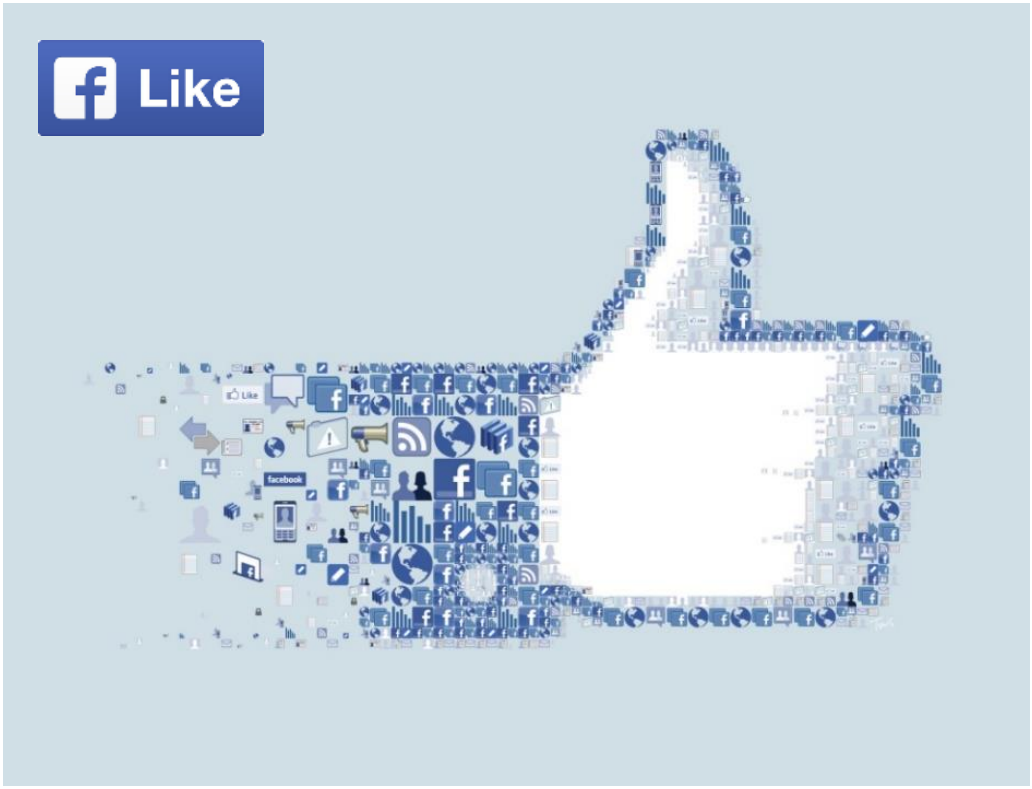
modeling

- 用有效的 features，f1-score 達到 0.94



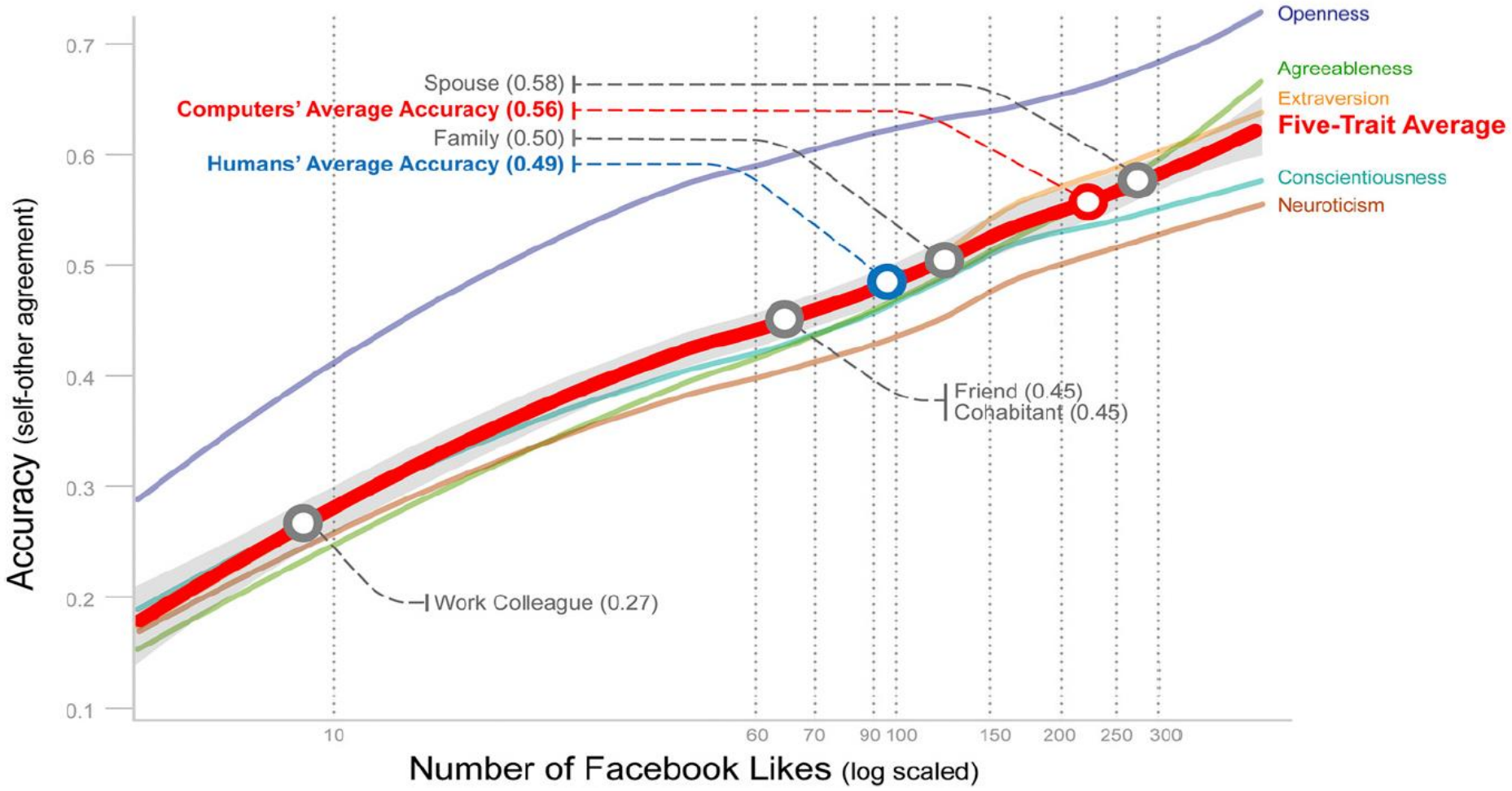
Random Forest	
Precision	0.947
Recall	0.940
F1-Score	0.943


You are what you like!


Like

Computer vs. Humans

- 用你按過讚的粉絲團來猜你的個性



Computer-based personality judgments are more accurate than those made by humans.
Proceedings of the National Academy of Sciences

誰按的粉絲團?

??	??
小茉莉-陳瑀希	Catworld小舖
Garena《英雄聯盟 LOL》	EYESCREAM Inc.
波多野結衣Hatano Yui	LOVFEE
解婕翎	OB嚴選
豆花妹 蔡黃汝	grace gift
Nono_辜莞允	BEVY C.
張景嵐	Joyceshopstyle
FHM Taiwan 男人幫國際中文版	QUEEN FASHION SHOP
潮物blog - 街頭潮流男著	SweeSa水莎
張小筑 Ya Chu	Lulus

性別

男	女
小茉莉-陳瑀希	Catworld小舖
Garena《英雄聯盟 LOL》	EYESCREAM Inc.
波多野結衣Hatano Yui	LOVFEE
解婕翎	OB嚴選
豆花妹 蔡黃汝	grace gift
Nono_辜莞允	BEVY C.
張景嵐	Joyceshopstyle
FHM Taiwan 男人幫國際中文版	QUEEN FASHION SHOP
潮物blog - 街頭潮流男著	SweeSa水莎
張小筑 Ya Chu	Lulus

誰按的粉絲團?

??	??
一休陪你一起愛瘦身	《HITO 本舖》
iFit 愛瘦身	潮物部落格
小甜甜 張可昀	PAZZO
FB減肥達人 輕鬆教你瘦	Image
美樂蒂 Melody	《 OneBoy 》
BEMAX	UNO STORE
Woma	RockSteady
OB嚴選	SweeSa水莎
鍾欣凌	高高-流行服飾Store
杜詩梅 Tu Shih Mei	Maxy

體型

胖	瘦
一休陪你一起愛瘦身	《HITO 本舖》
iFit 愛瘦身	潮物部落格
小甜甜 張可昀	PAZZO
FB減肥達人 輕鬆教你瘦	Image
美樂蒂 Melody	《 OneBoy 》
BEMAX	UNO STORE
Woma	RockSteady
OB嚴選	SweeSa水莎
鍾欣凌	高高-流行服飾Store
杜詩梅 Tu Shih Mei	Maxy

誰按的粉絲團?

高	低
TED	在不瘋狂就等死
Technews 科技新報	羅志祥 SHOW
批踢踢實業坊 (Ptt.cc)	Mimi Dancing Club
PanSci 科學新聞網	小A辣
商業周刊（商周.com）	小三魔法棒 (小三美日)
TEDxTaipei	真愛談戀愛×真愛橋
The News Lens 關鍵評論網	宛宛兒
背包客棧	連靜雯joanne lien
VoiceTube 看影片學英語	BY2
Cheers：快樂工作人	爆笑禁區

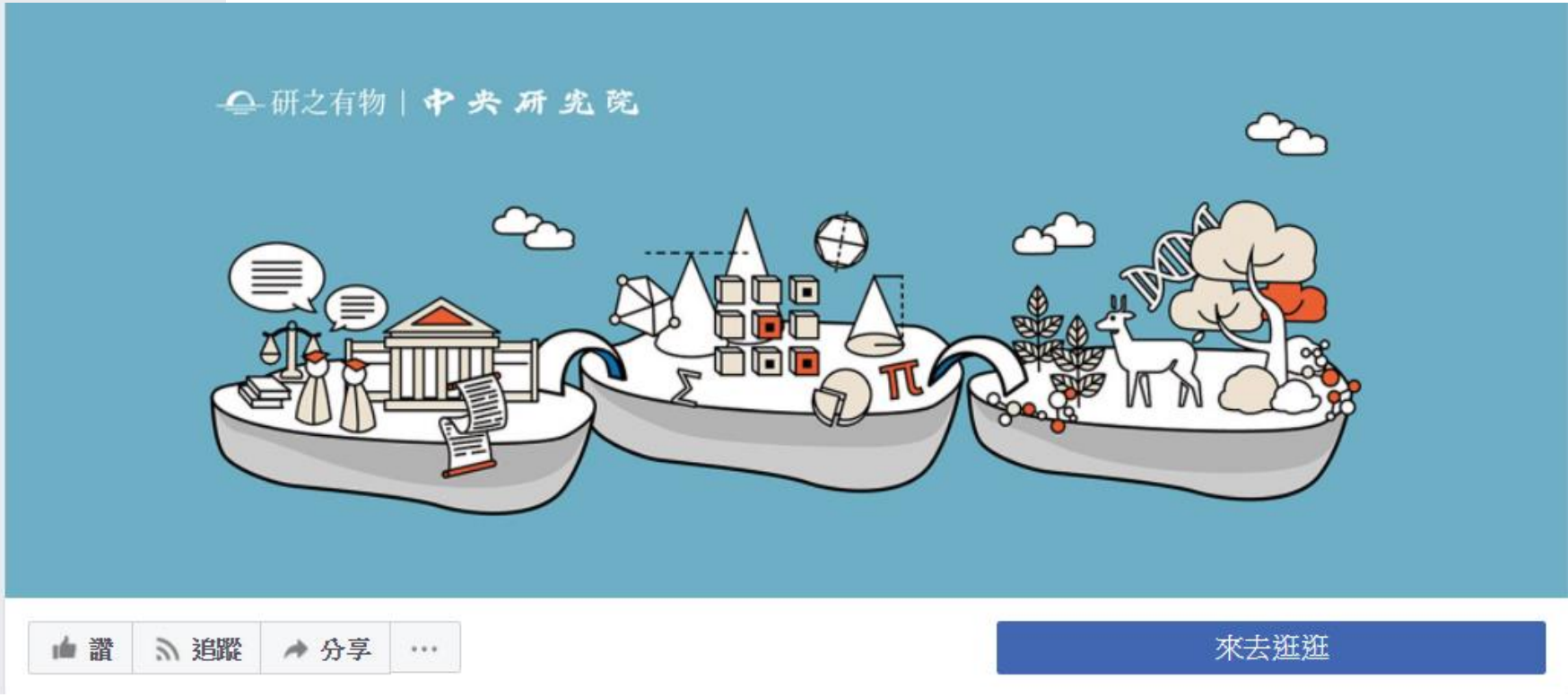
學歷

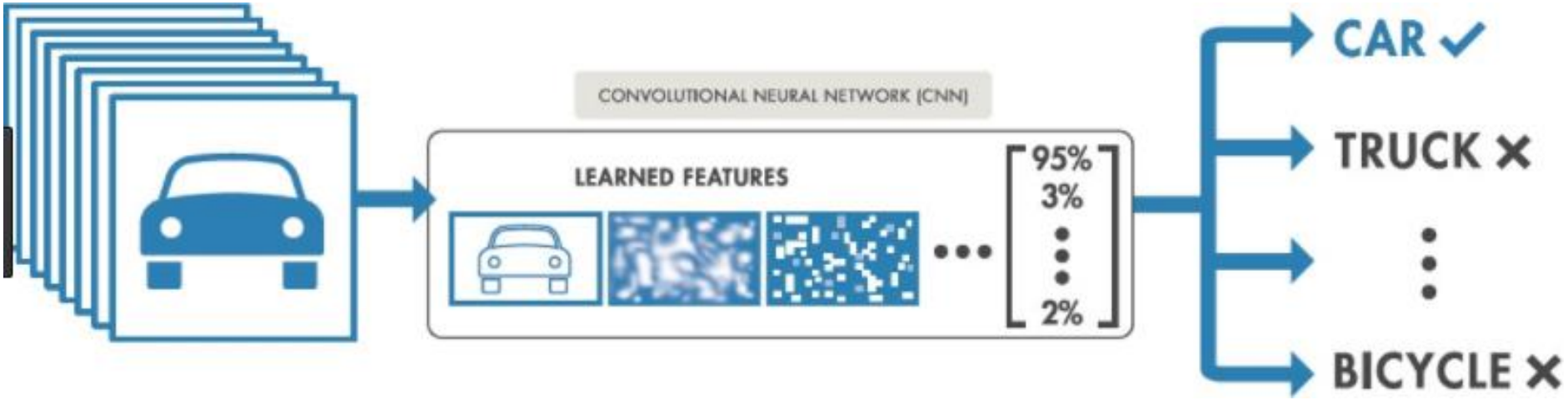
高	低
TED	在不瘋狂就等死
Technews 科技新報	羅志祥 SHOW
批踢踢實業坊 (Ptt.cc)	Mimi Dancing Club
PanSci 科學新聞網	小A辣
商業周刊（商周.com）	小三魔法棒 (小三美日)
TEDxTaipei	真愛談戀愛×真愛橋
The News Lens 關鍵評論網	宛宛兒
背包客棧	連靜雯joanne lien
VoiceTube 看影片學英語	BY2
Cheers：快樂工作人	爆笑禁區

研之有物

研之有物
@research.sinica

本站盼以具體的研究案例、真實的研究員生活，帶您前往數理科學、生命科學、人文社會三大領域研究現場，揭開中央研究院神秘的面紗

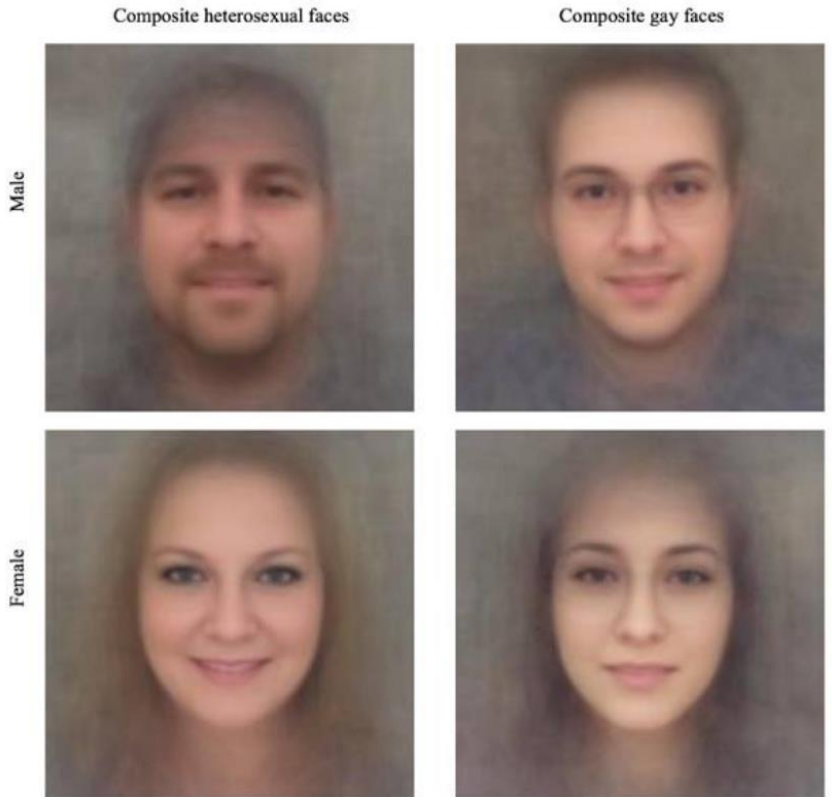




卷積神經網路
Convolutional Neural Network

AI 判斷同性戀?

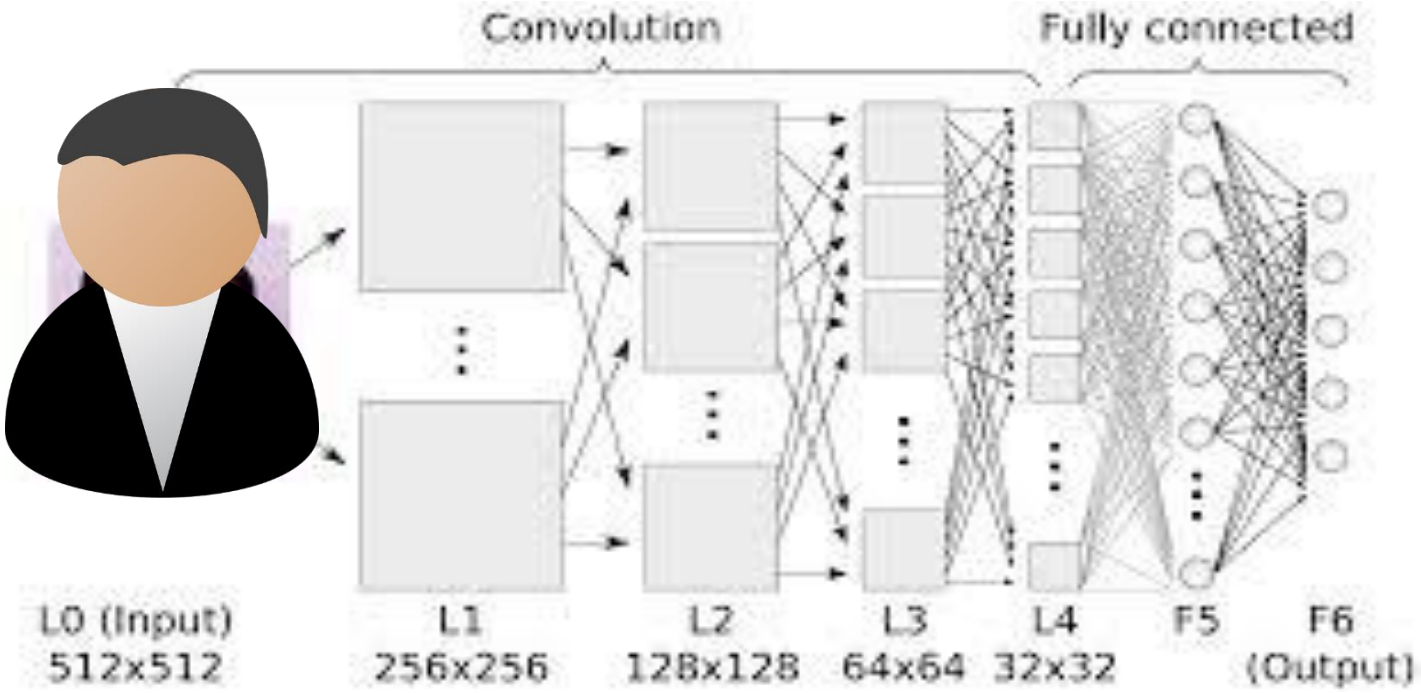
- 史丹佛大學透過美國約會網站 3 萬 5000 多名男女會員的臉部照片為樣本，分辨出同男和異男的準確度可達81%



Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.

用照片預測基本資料

- 運用 CNN 提取圖片特徵，進行分類

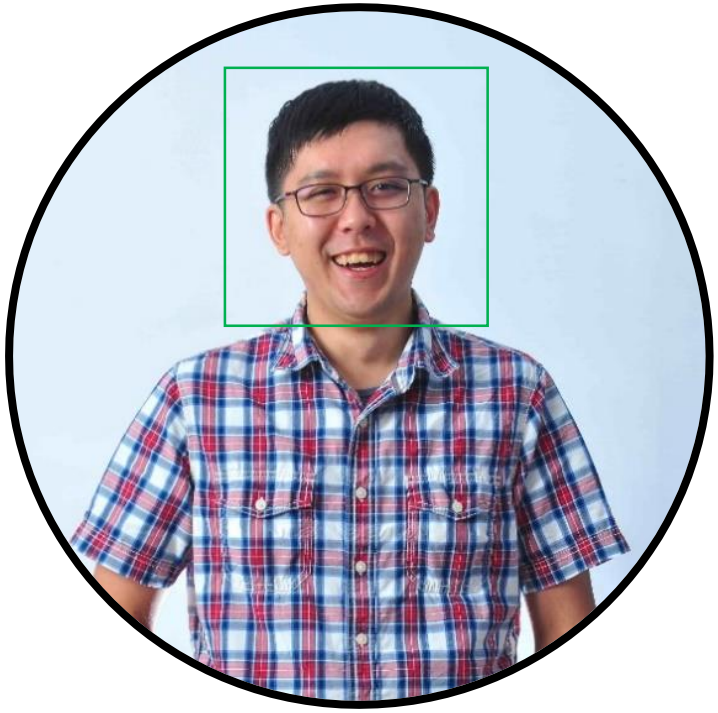


預測結果

- 純粹用照片來預測

會不會被喜歡 (女生)	0.9
會不會被喜歡 (男生)	0.78
FB 朋友數	0.67
是否抽菸喝酒	0.65

預測誰是基督徒?



基督教徒的機率: 0.9999

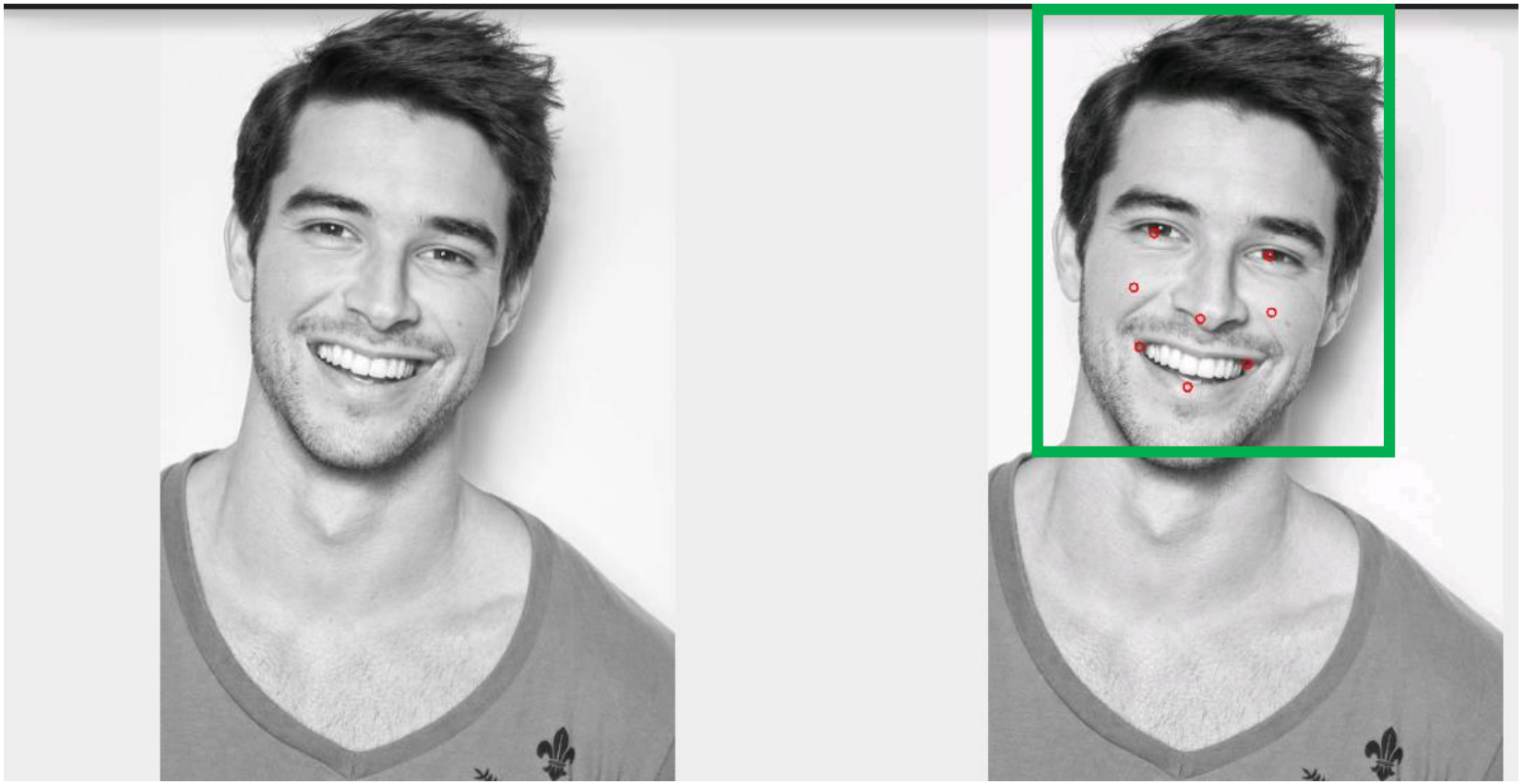
基督教徒的機率: 0

三冠王模型的故事

- 預測是否為基督徒
  - Accuracy: 1.00
  - AUC: 1.00
  - F1-score: 1.00
- 確定是在 testing data 上面的表現

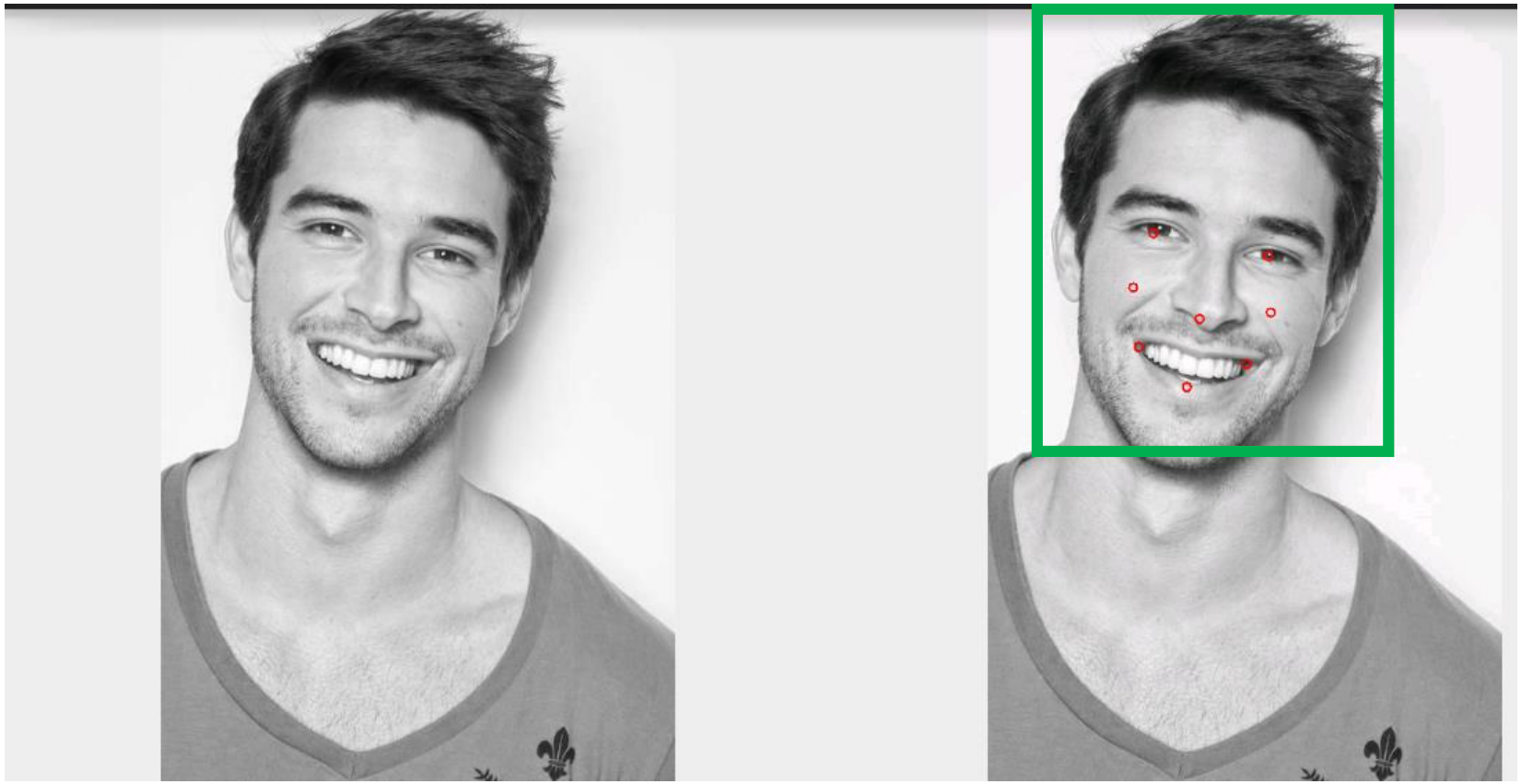
三冠王模型的故事

- 資料前處理 → 用 dlib 確保每張圖片都有包含臉，且全臉佔圖的比例不能太小

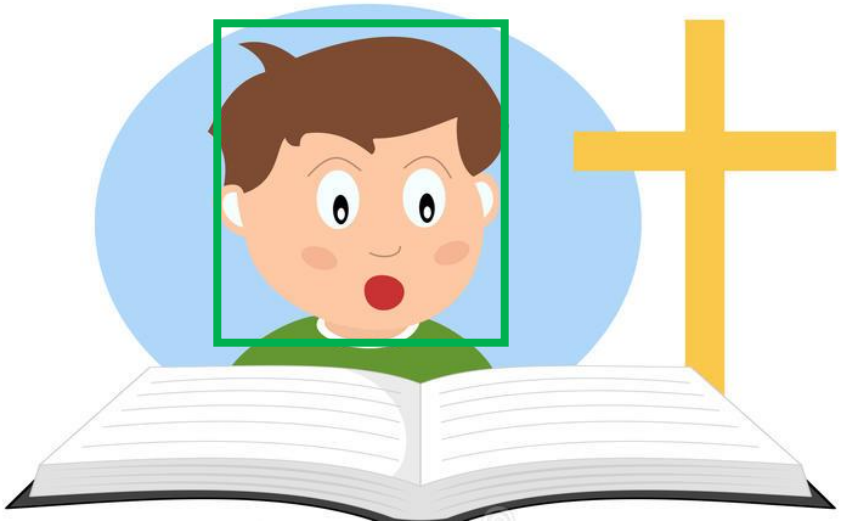


三冠王模型的故事

- 忘記把所有基督徒照片的綠框刪除



資料 Recheck

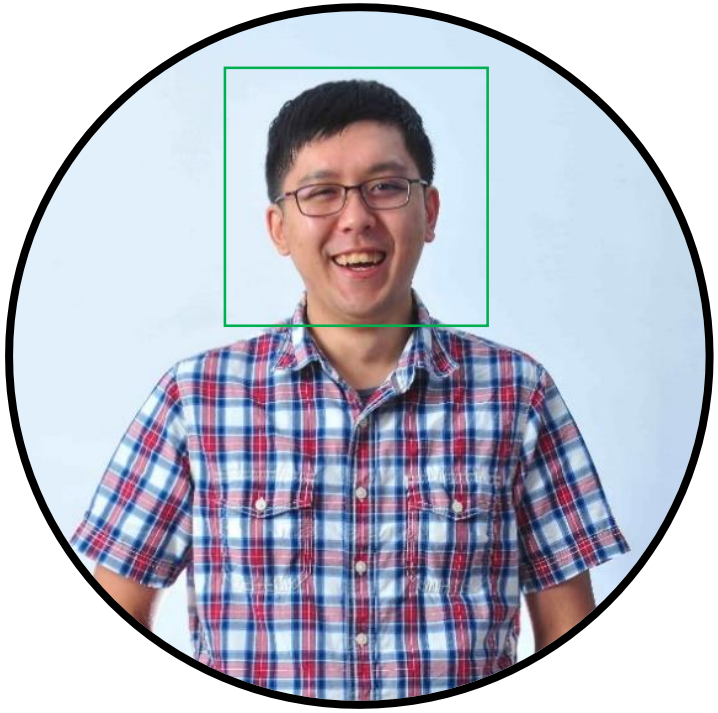


基督徒

非基督徒

- 我訓練好的模型不是一個基督徒分類器
- 而是一個**“綠框”**分類器!

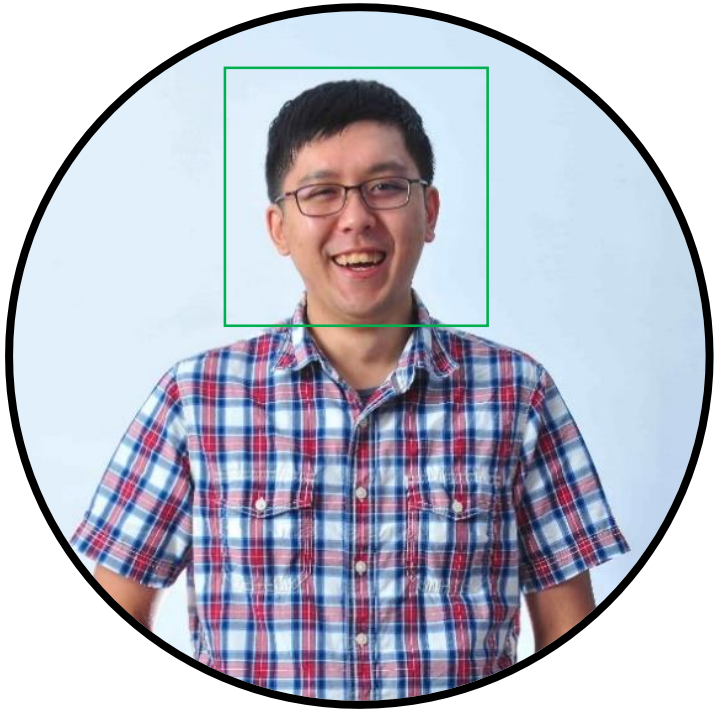
預測誰是基督徒?



基督教徒的機率: 0.9999

基督教徒的機率: 0

預測誰是基督徒?哪張照片有綠框?



基督教徒的機率: 0.9999

有綠框的機率: 0.9999

基督教徒的機率: 0

有綠框的機率: 0

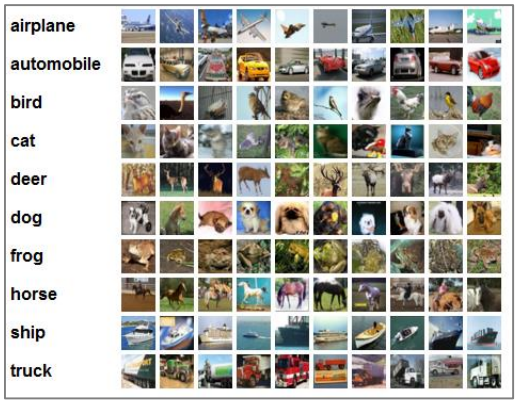
深度學習在現實世界應用中的各種趣味陷阱

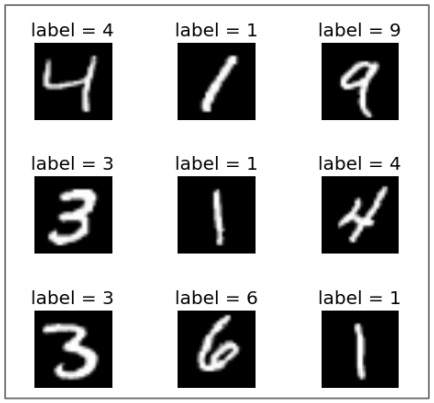
深度學習的美好想像與迷思

- 萬物皆可 Train
  - FizzBuzz 的故事

深度學習的美好想像與迷思

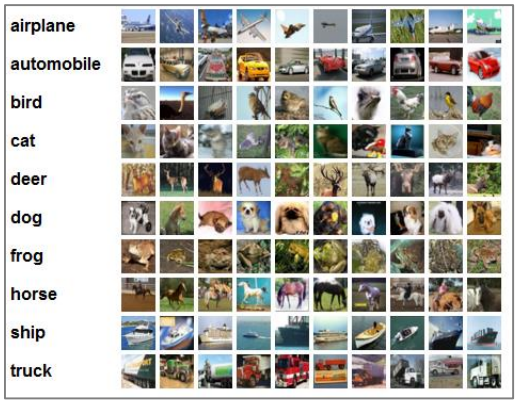
- 萬物皆可 Train
  - FizzBuzz 的故事
- 網路上的資源很多, 教學也很多, 直接套不難吧?

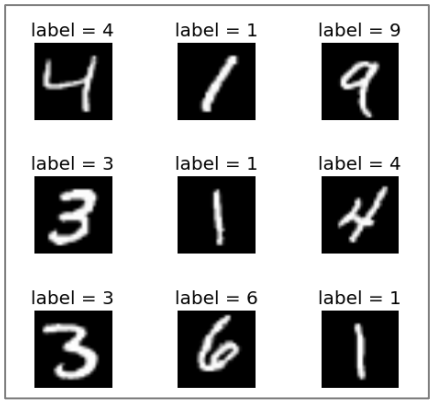




深度學習的美好想像與迷思

- 萬物皆可 Train
  - FizzBuzz 的故事
- 網路上的資源很多, 教學也很多, 直接套不難吧?

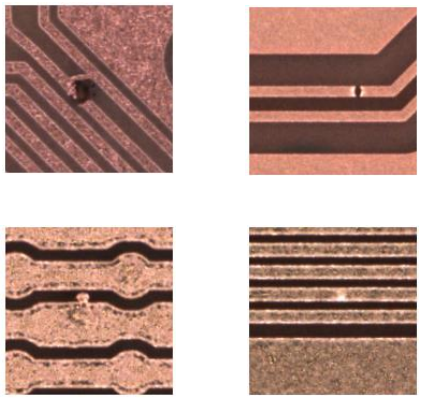




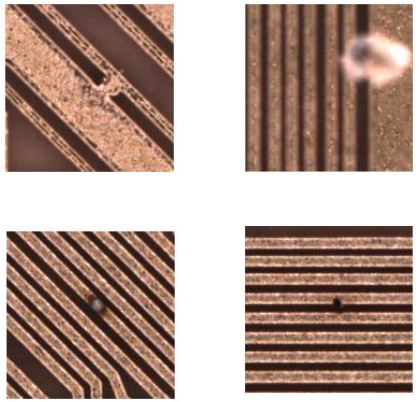
- 深度學習一定強!
  - 很多時候傳統的演算法反而更好
  - Occam's razor


ForGIFs.com

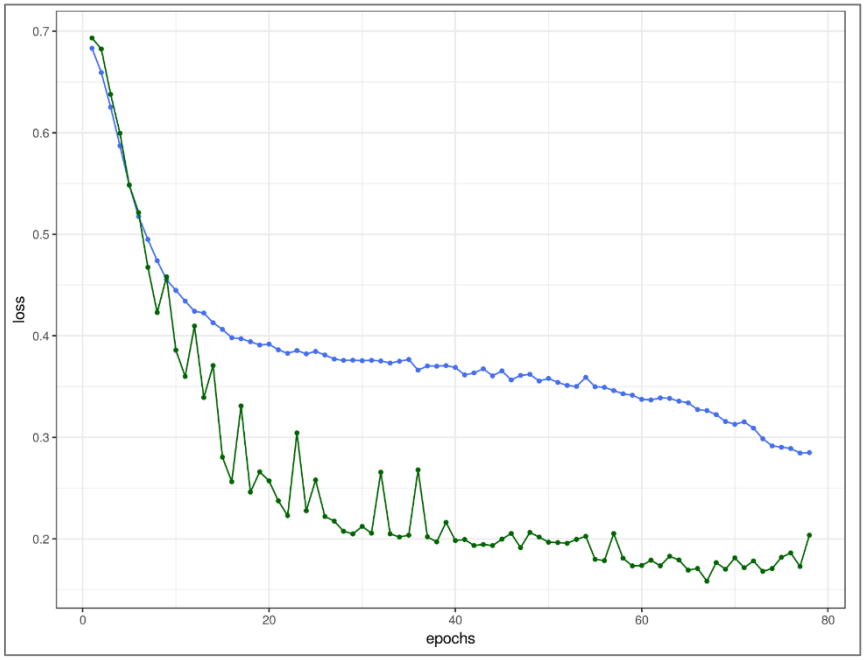
1. 異常分類 - 不就是個分類問題嗎?



Fail



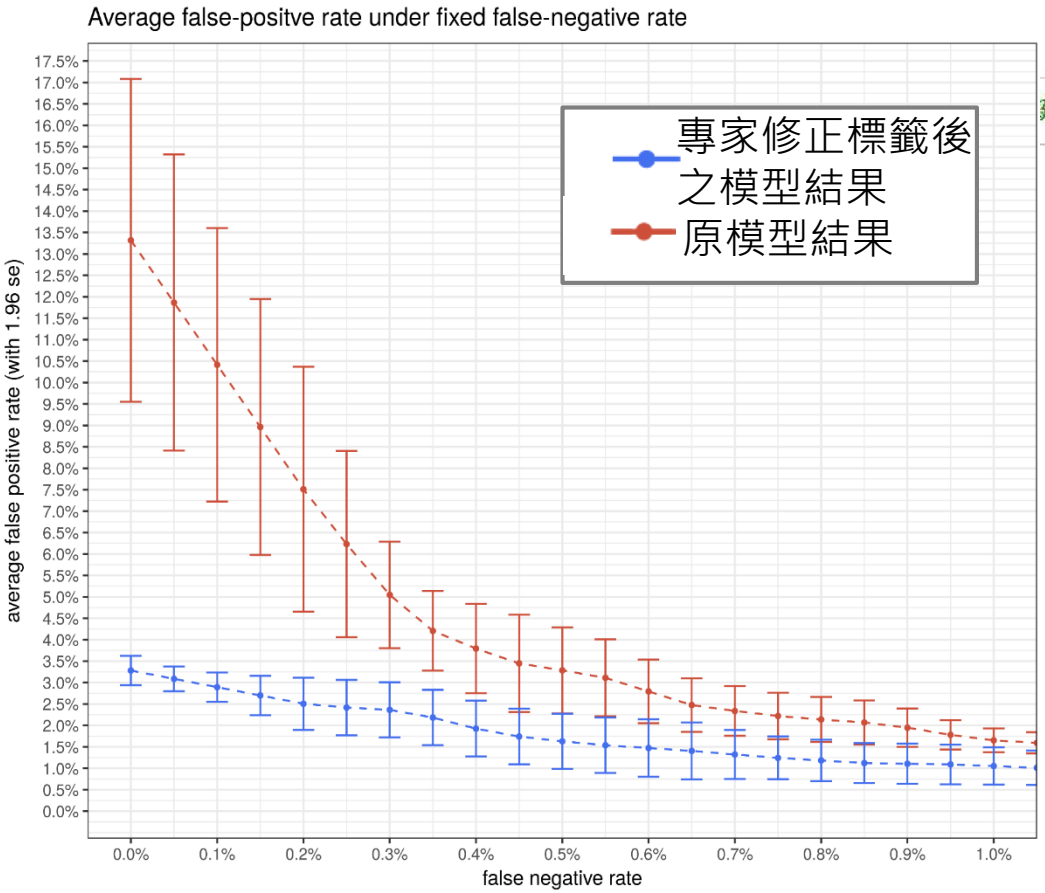
Pass



異常分類 - 不就是個分類問題嗎?

原本人工標記的 label 真的是乾淨的?

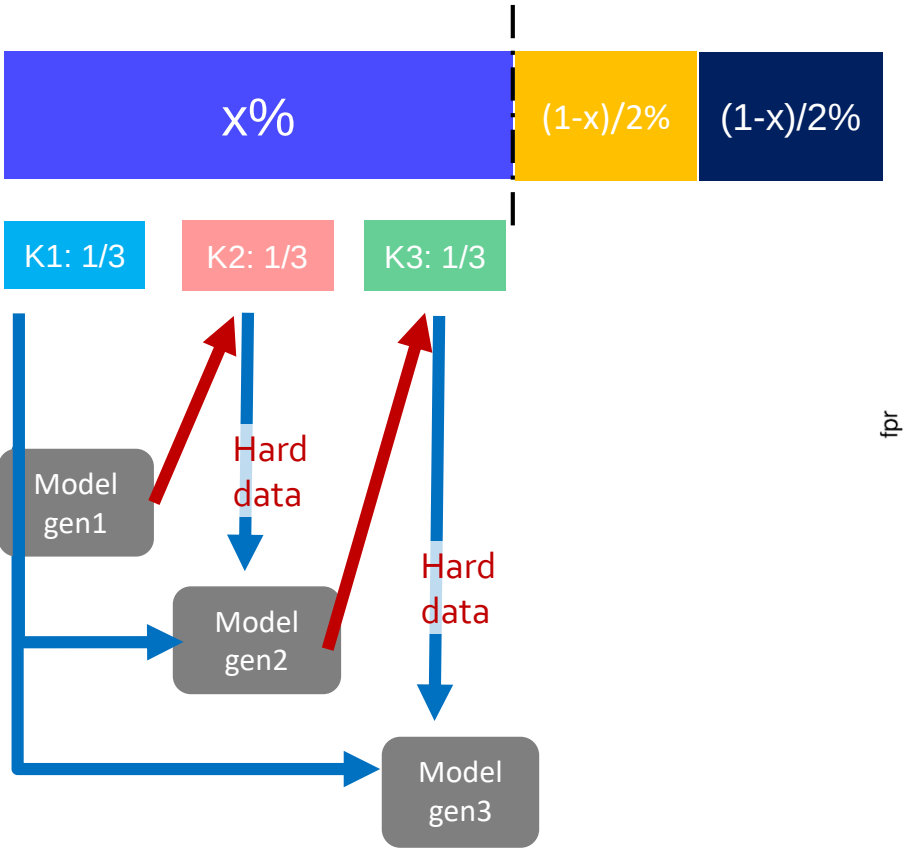
- 專家複判 Web UI

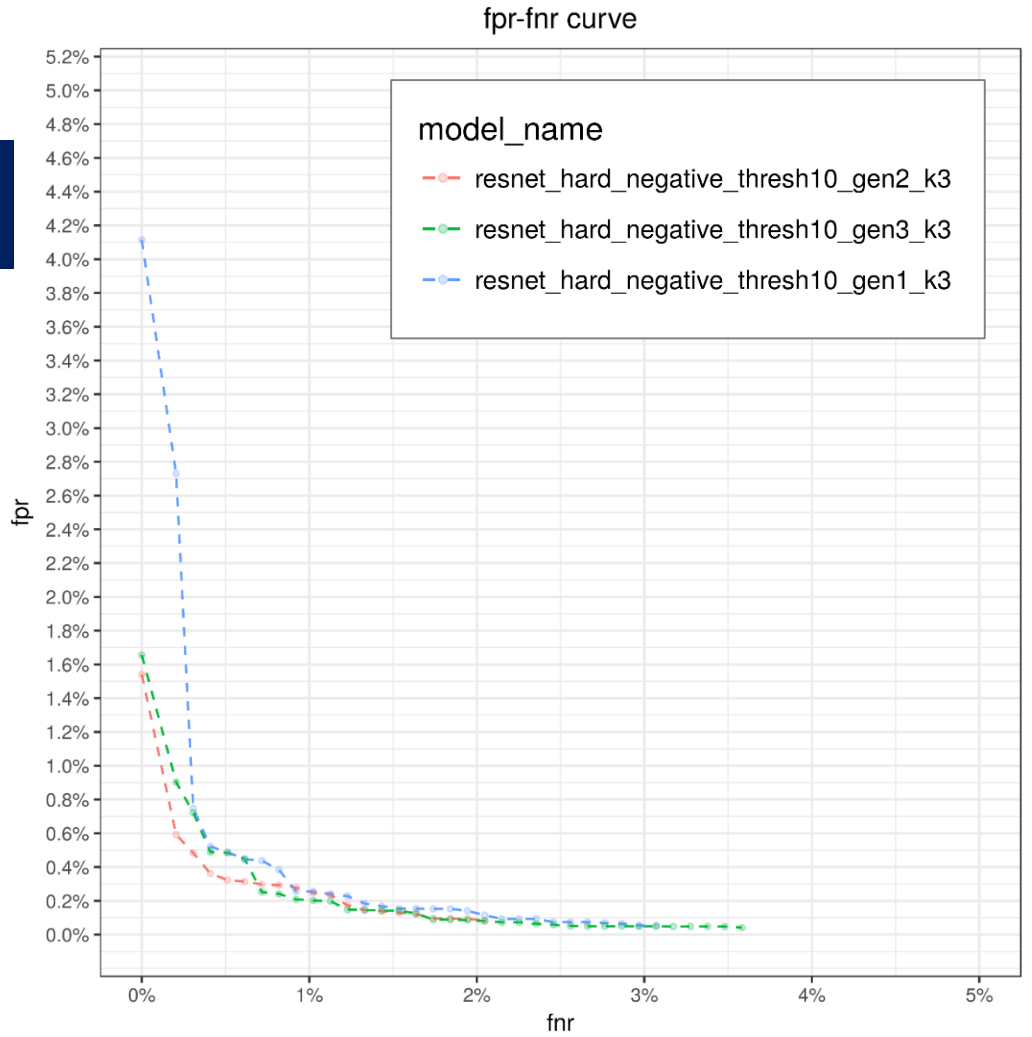


異常分類 - 不就是個分類問題嗎?

訓練一次不好, 你有試過訓練兩次嗎?

- Data boosting





2. 原料配色 – 神秘的那道光?

顏色差異（英語：**Color difference**），亦稱**距離**，是色彩學上的一個關注點。它量化了一個概念。在未量化之前，人們只能用形容詞來大概描述這個概念，這使得對顏色要求嚴格的工作者們很不方便。顏色差異可以通過色彩空間內的歐氏距離簡單計算得出，也可以使用CIE較為複雜、均勻的人類知覺公式計算。

2. 原料配色 – 神秘的那道光?

$$\Delta E = \sqrt{\left(\frac{\Delta L}{l S_L}\right)^2 + \left(\frac{\Delta C}{c S_C}\right)^2 + \left(\frac{\Delta H}{S_H}\right)^2}$$

where

$$\Delta C = C_1 - C_2$$

$$C_1 = \sqrt{{a_1}^2 + {b_1}^2}$$

$$C_2 = \sqrt{{a_2}^2 + {b_2}^2}$$

$$\Delta H = \sqrt{\Delta a^2 + \Delta b^2 - \Delta C^2}$$

$$\Delta L = L_1 - L_2$$

$$\Delta a = a_1 - a_2$$

$$\Delta b = b_1 - b_2$$

$$S_L = \begin{cases} 0.511 & \text{if } L_1 < 16 \\ \frac{0.040975 L_1}{1 + 0.01765 L_1} & \text{if } L_1 \geq 16 \end{cases}$$

$$S_C = \frac{0.0638 C_1}{1 + 0.0131 C_1} + 0.638$$

$$S_H = S_C(FT + 1 - F)$$

$$T = \begin{cases} 0.56 + |0.2 \cos(H_1 + 168^\circ)| & \text{if } 164^\circ \leq H_1 \leq 345^\circ \\ 0.36 + |0.4 \cos(H_1 + 35^\circ)| & \text{otherwise} \end{cases}$$

$$F = \sqrt{\frac{C_1^4}{C_1^4 + 1900}}$$

$$H = \arctan\left(\frac{b_1}{a_1}\right)$$

$$H_1 = \begin{cases} H & \text{if } H \geq 0 \\ H + 360^\circ & \text{otherwise} \end{cases}$$

Ref: http://www.brucelindbloom.com/index.html?Eqn_DeltaE_CMC.html

2. 原料配色 – 神秘的那道光?

$$\Delta E = \sqrt{\left(\frac{\Delta L}{lS_L}\right)^2 + \left(\frac{\Delta C}{\quad}\right)^2 + \left(\frac{\Delta H}{\quad}\right)^2}$$

where

$$\Delta C = C_1 - C_2$$

$$C_1 = \sqrt{{a_1}^2 + {b_1}^2}$$

$$C_2 = \sqrt{{a_2}^2 + {b_2}^2}$$

$$\Delta H = \sqrt{\Delta a^2 + \Delta}$$

$$\Delta L = L_1 - L_2$$

$$\Delta a = a_1 - a_2$$

$$\Delta b = b_1 - b_2$$

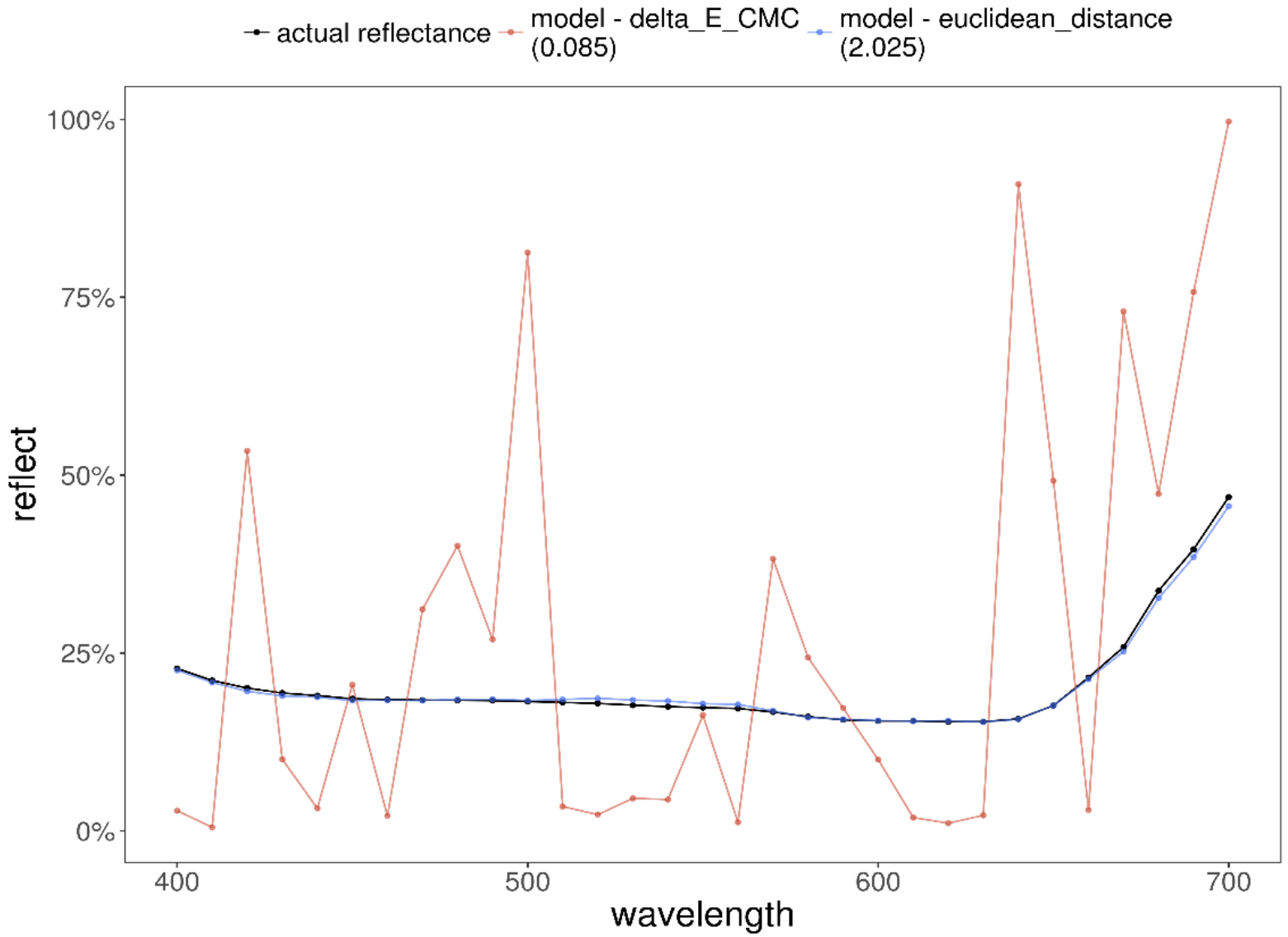
if $164^\circ \leq H_1 \leq 345^\circ$
otherwise

$$H_1 = \begin{cases} H & \text{if } H \geq 0 \\ H + 360^\circ & \text{otherwise} \end{cases}$$

Ref: http://www.brucelindbloom.com/index.html?Eqn_DeltaE_CMC.html

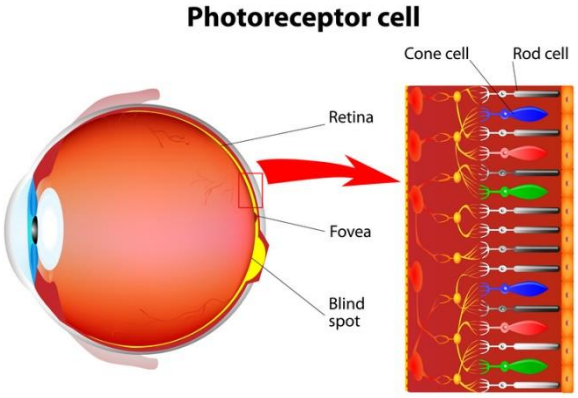
原料配色 – 神秘的那道光?

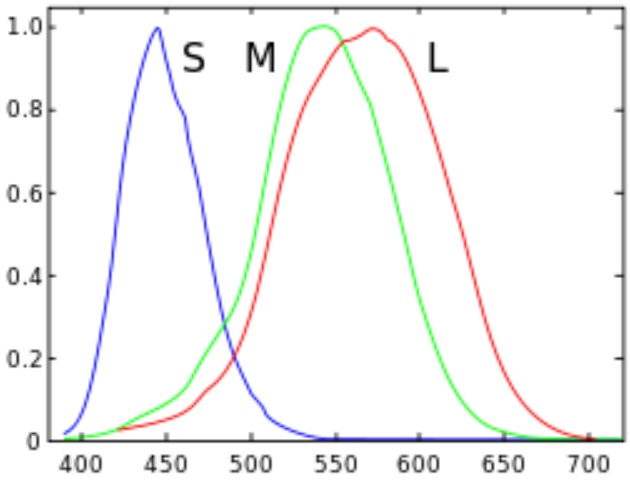
直觀的 Loss function = 詭異的 fitting result? : dECMC vs. eud.loss

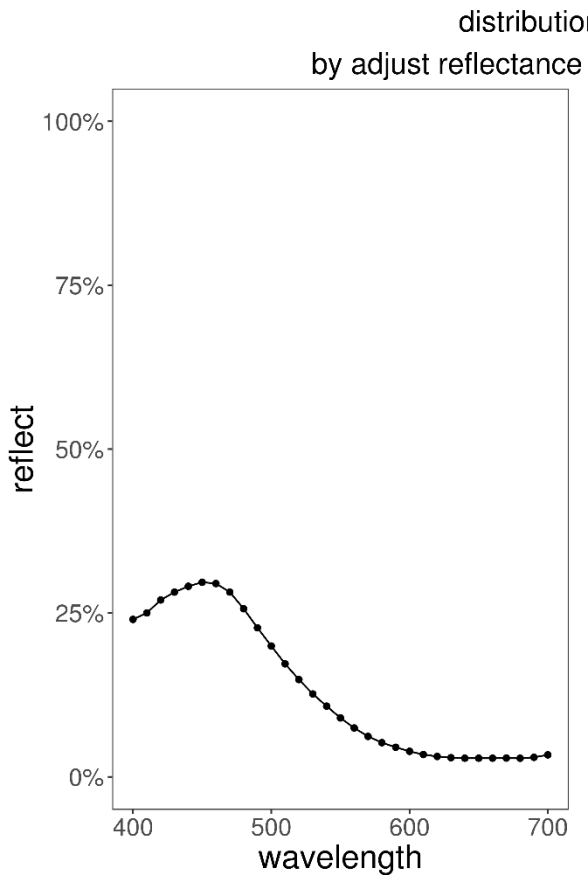


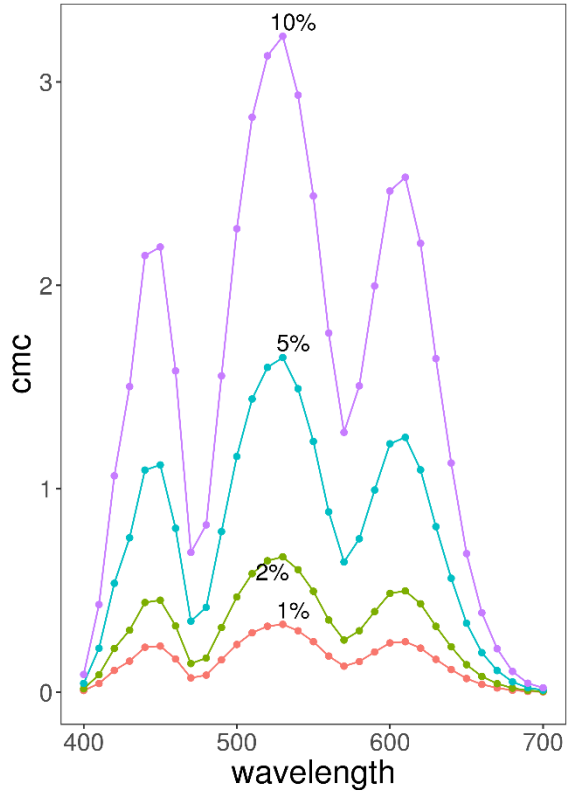
原料配色 – 神秘的那道光?

為何出現詭異的結果?



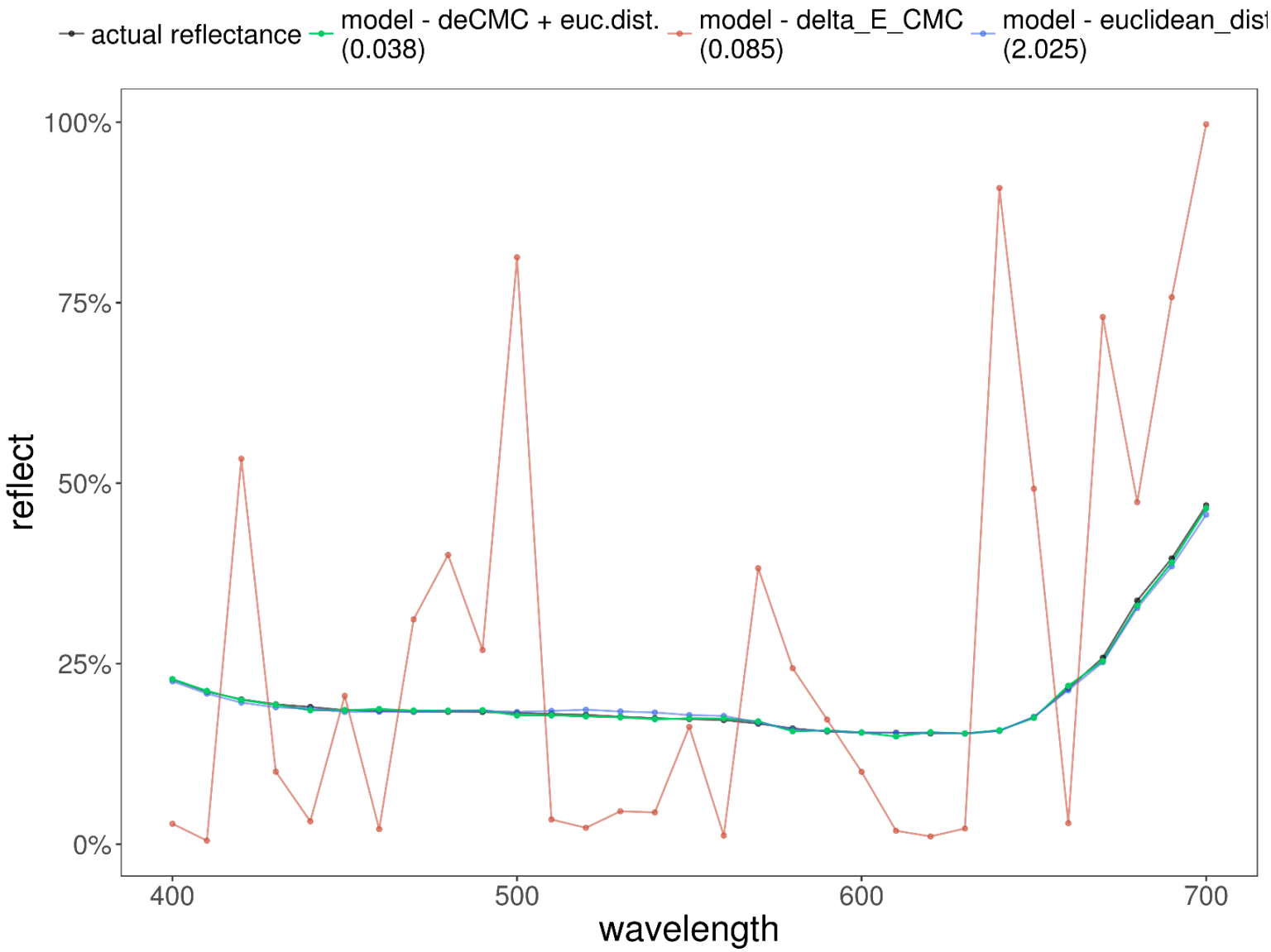






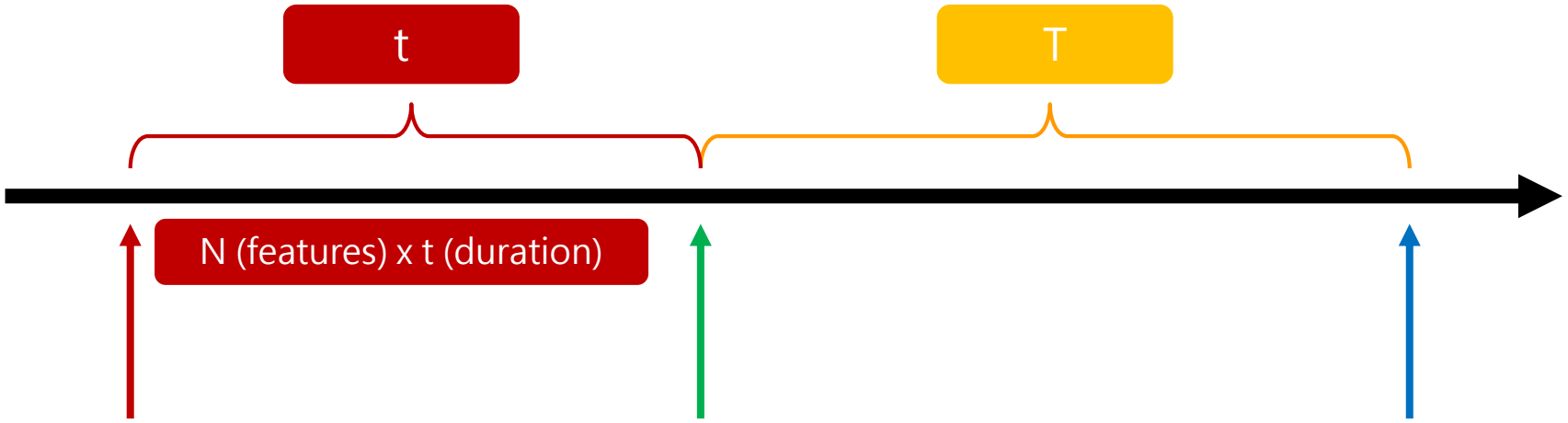
原料配色 – 神秘的那道光?

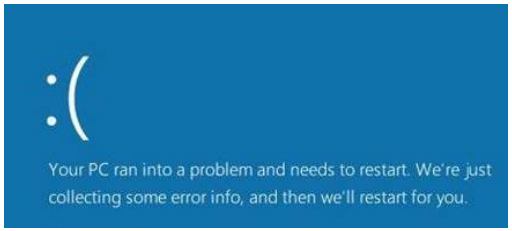
改成兩階段的 model fitting



3. 預測性維護 – 又是分類問題?

- 標準問題定義

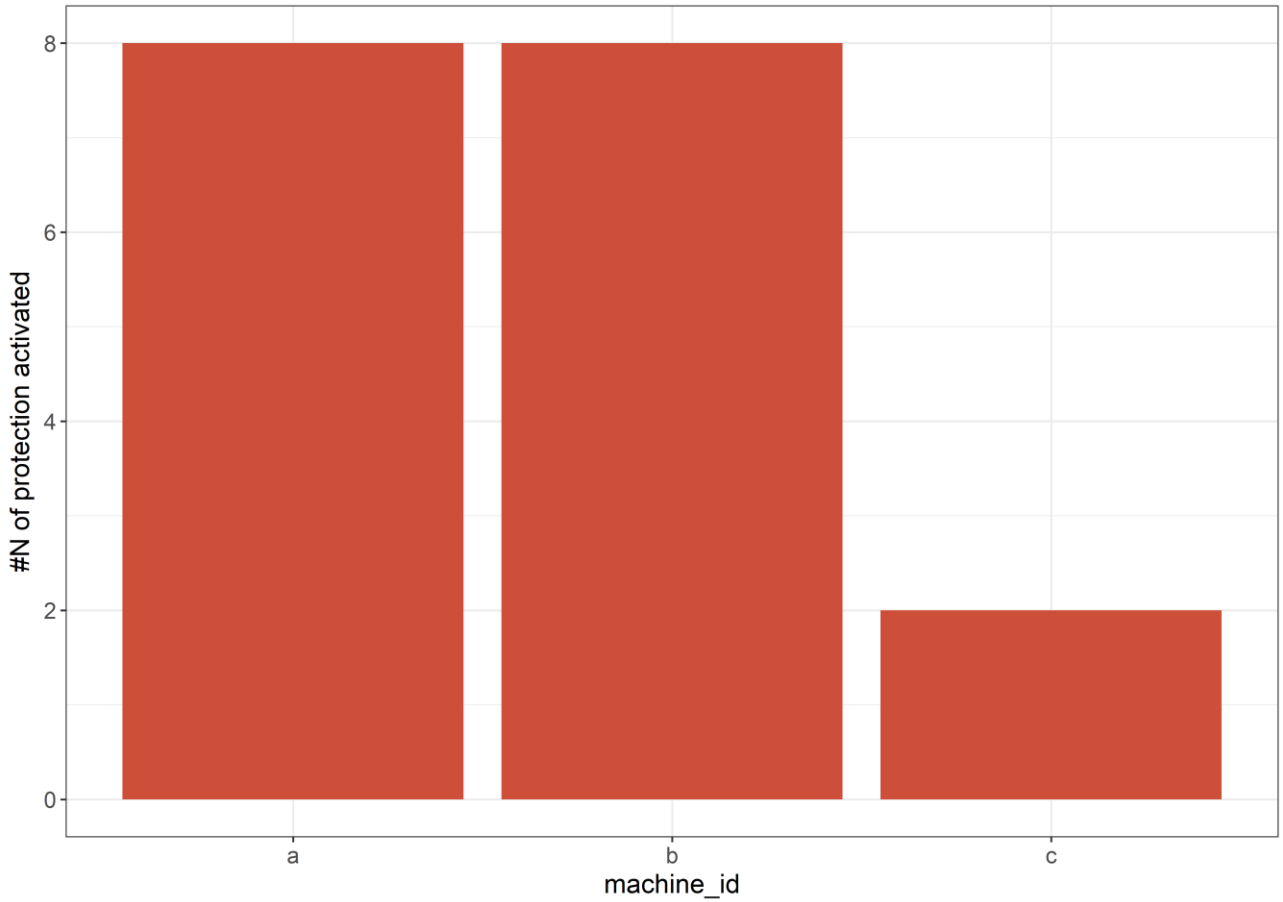




預測性維護 – 又是分類問題?

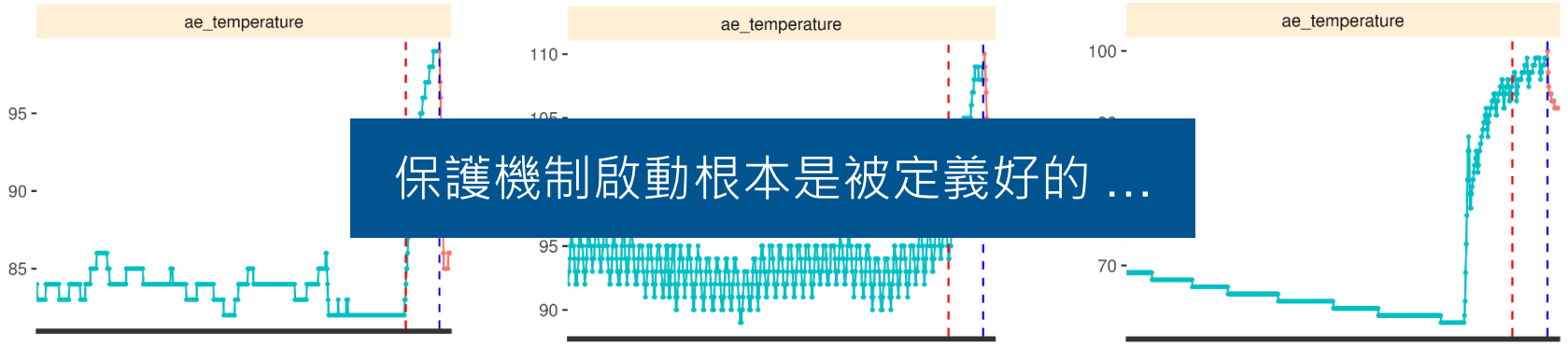
案情並不單純

- 八個月來... 最多也只出現八次異常?!
  - 每 5 秒一筆資料, 八個月就是 ...五百多萬筆 ...
- “異常” 不是損壞, 只是保護機制 ...



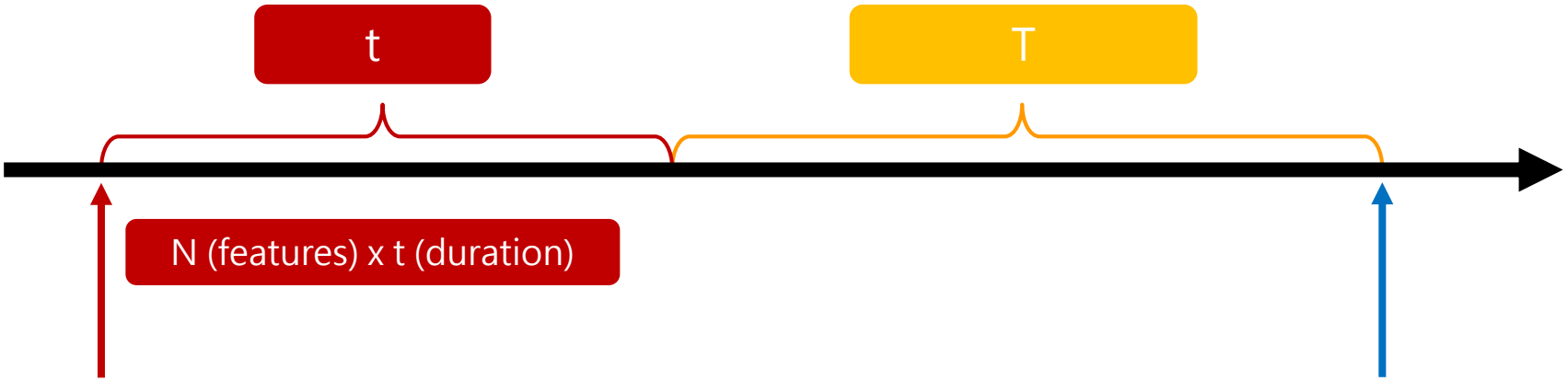
預測性維護 – 又是分類問題?

重新定義問題



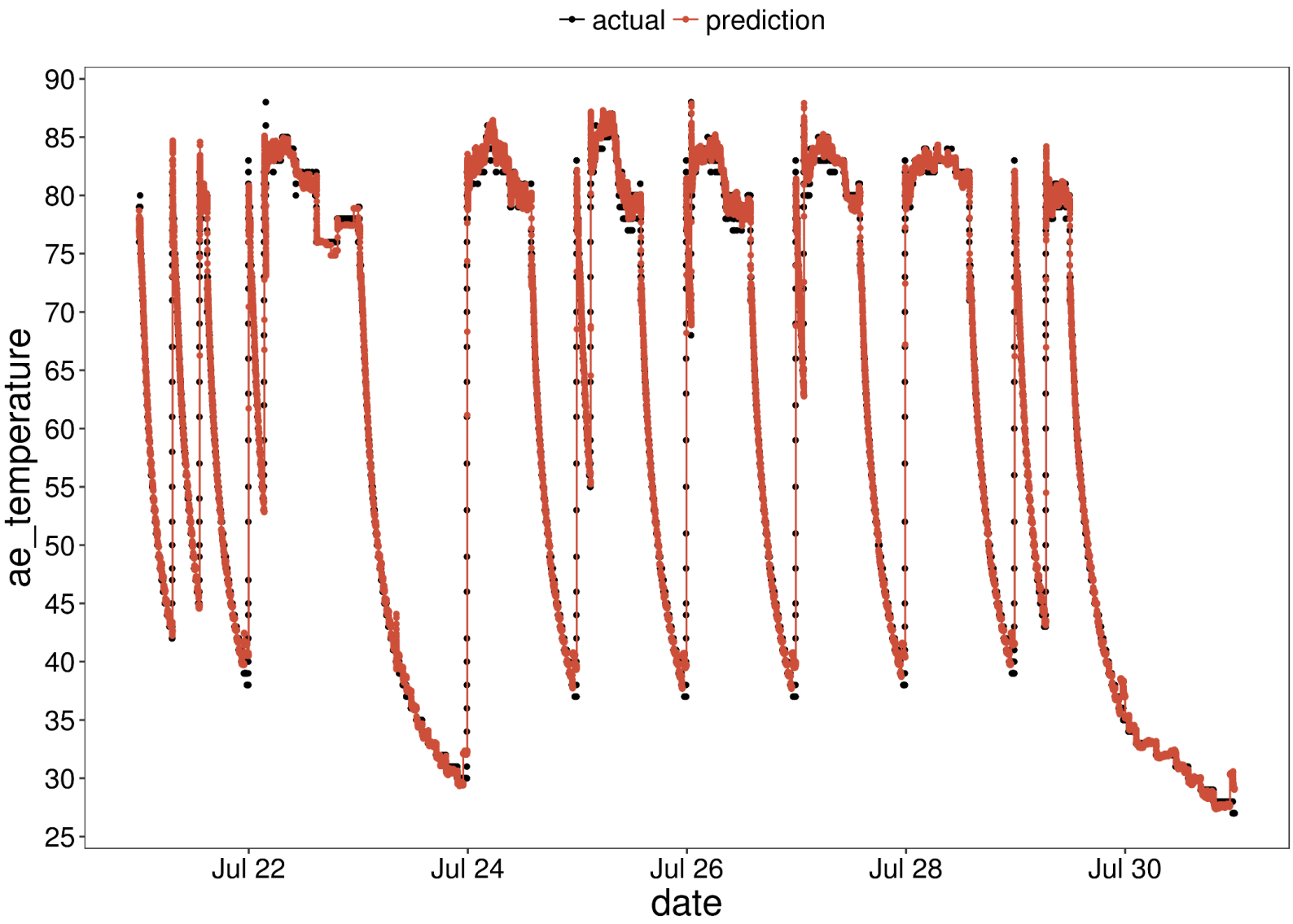
保護機制啟動根本是被定義好的 ...

壓死駱駝的**最後一根稻草**什麼時候會出現?



預測性維護 – 又是分類問題?

轉換目標：變成迴歸問題



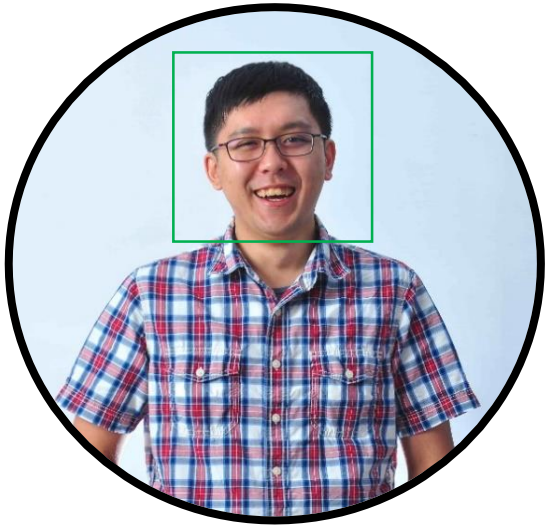
借模型很容易, 但是 …
問對問題、**了解的模型限制**、**山不轉路轉**
才能讓你在陷阱重重的現實世界存活

Take home message

- 對於 **”跨領域”**
  - 資料分析不單是資工、電機的領域
    - Lab: 會計、應用數學、風險管理、土木、醫學、光電, …
  - 能帶來更多不同的思維，激盪出新的火花
  - 要踏入之前，可以先多上點線上課程或多聽演講，確認你真的有興趣
    - 資料分析中 90% 的時間是很枯燥乏味的，但當成功找到有意義的結果時，會覺得努力都是值得的

Take home message

- **“資料分析”** 跟 **“深度學習”** 這件事
  - 別衝動, 好好先檢查資料
  - 思考問題的本質是什麼
  - 了解模型的限制、改變模型的架構
    - 窮則變，變則通
  - **結果太好的時候，不要高興得太早!**
    - 莫忘綠框



Thank You!