



數據科學思考與實證研究 (Thinking like a Data Scientist)

Chia-Yen Lee, Ph.D. (李家岩 博士)



Department of Information Management (資訊管理學系)
National Taiwan University (國立台灣大學)

□ Education

- Ph.D, 工業與系統工程, Texas A&M University, USA (Major: Operations Research 作業研究/運籌學)
- M.S., 工業工程與工程管理, 國立清華大學
- B.S. & B.B.A., 應用數學暨資訊管理, 國立政治大學



□ Experience

- 教授, 國立台灣大學資訊管理學系
- 教授兼所長, 國立成功大學資訊工程學系暨製造資訊與系統研究所
- 副編輯, IEEE Transactions on Automation Science and Engineering (SCI)
- 半導體廠科技顧問、台積電工業工程師、陸軍少尉資訊官

□ Award

- 呂鳳章先生紀念獎(2019)、美光教師Micron Teacher Award (2018)
- 李國鼎科技與人文講座研究獎 (2018)、
- 科技部吳大猷先生紀念獎 (2017)
- 優秀青年工業工程師獎 (2016)

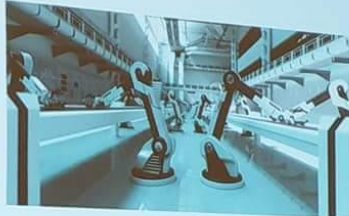
□ Research Interest

- 製造數據科學、智慧型製造系統、生產力與效率分析、多目標決策

智慧工廠

Intelligent Factory is a **decision-oriented** system which has the computational intelligence and self-learning ability to optimize the manufacturing process.

- 計算智慧 → Based on Data (資料處理與分析)
- 自我學習 → Real-time Feedback Control (回饋控制)

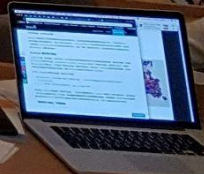


生產力最佳化實驗室@NCKU

製造資料科學?

李家蔚 (成大資訊系統製造所)

A large backdrop for an event featuring logos of various sponsors and partners. The logos include: 玉山金控 (E.SUN FHC), TREND MICRO 趨勢科技, 國泰金控 (Cathay Financial Holdings), funP, intel, KKBOX, MEDIATEK, QNAP, 春暉電算 (Chunhui Computing), 台灣大哥大 卓越 (Taiwan Mobile Excellent), 玉山金控, 趨勢科技, 國泰金控, funP, intel, KKBOX, MEDIATEK, QNAP, 春暉電算, 台灣大哥大 卓越. Below the logos, there are several smaller logos and text, including '主辦單位' (Organized by) and '協辦單位' (Co-organized by) with their respective logos and names.







數據科學思考

如果你參加賽跑，追過第二名，請問你現在第幾名？



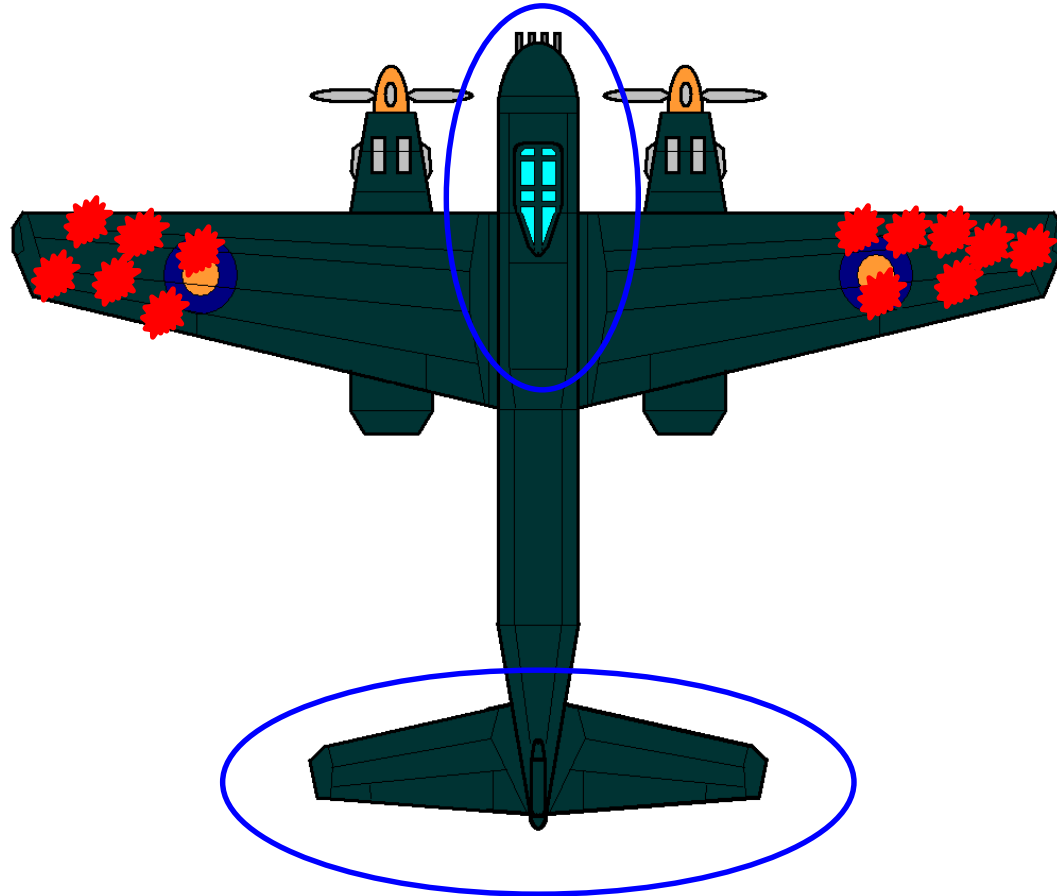
有研究指出，90%的人都死在床上...



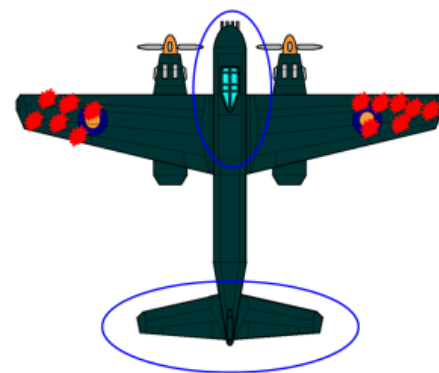
- 1941年，第二次世界大戰正打得如火如荼。
- 英國皇家空軍的作戰指揮官，拜訪了美國哥倫比亞大學著名的統計學家沃德(Abraham Wald)教授。
- 他說：「沃德教授，每次飛行員出發去執行轟炸任務，我們最怕聽到的回報是：『**呼叫總部，我中彈了**』。請協助我們改善這個攸關飛行員生死的難題吧！」



□ 蒐集資料 => 什麼樣的資料？



□ 在之後的會議上，有兩派說法爭執不下：



英國皇家空軍指揮官：機翼加裝板甲

VS

沃德：尾翼(發動機) + 座艙加裝板甲

-看不見的彈痕最致命-

- 諾貝爾經濟學得主卡曼尼 (Daniel Kahneman, 1934年-) 和他的研究夥伴特佛斯基 (Amos Tversky, 1937年-1996年)做了個實驗

假設有位女性名叫琳達，31歲，單身，坦率而且十分聰明。她大學時主修哲學，在學生時代十分關心歧視和社會正義等議題，還參加過反核示威遊行...

- 實驗描述：你覺得下面敘述是對的請舉手 (可重複舉)...

敘述	票數
琳達活躍於婦女運動	19
琳達是精神病院裡的社工	9
琳達在書店工作，並定期上瑜珈課	13
琳達是銀行出納員且活躍於婦女運動	43
琳達是小學老師	33
琳達是某婦權聯盟的一員	8
琳達是銀行出納員	26
琳達是保險業務員	28

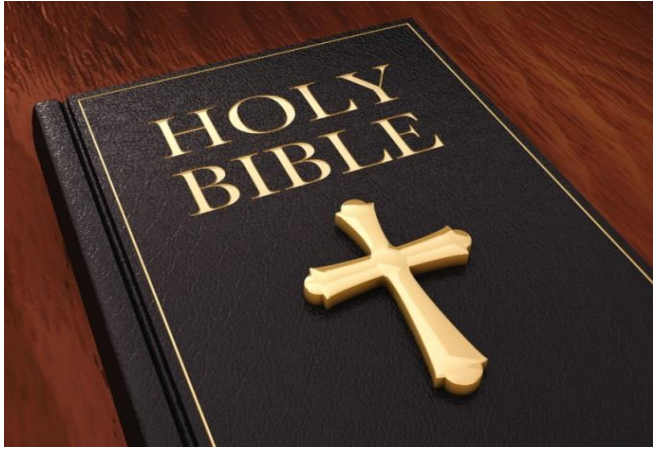
你真的相信你的直覺嗎？

好的決策來自於...

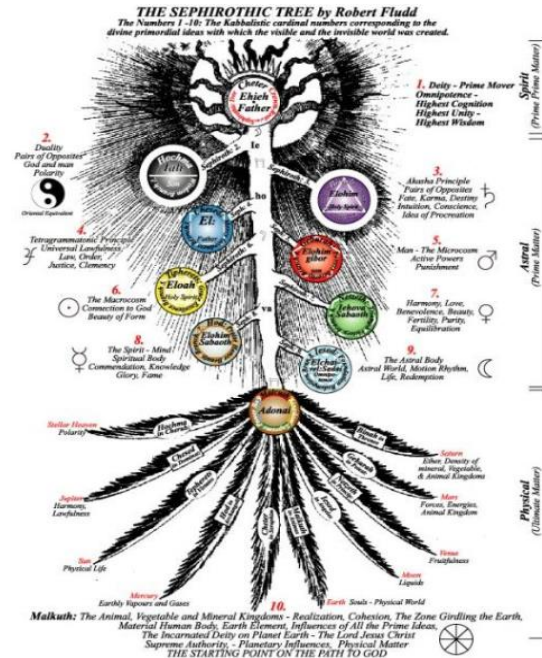
「智慧」!?

智慧來自於...

智慧



智慧是...



□ 聖經故事

□ 某天兩個女人（A和B）先後生下一名男嬰...

- A宣稱B的小孩死了，指責B卻趁A睡著時偷偷將小孩掉包；而B則宣稱A的小孩死了，他的小孩還活著。

□ 最後所羅門王必須判斷孩子是誰的，他最後決定把小孩子剖成兩半，一人拿一半。

□ 這時A卻哀求國王別傷害孩子，自願把孩子讓給B；而B卻接受所羅門王的作法，於是答案呼之欲出，A才是這孩子的生母。

似乎...直覺(或智慧)跟邏輯
有些關聯...

什麼是邏輯？

神存不存在?
prove it!

神唯不唯一?
prove it!

給我證據，其餘免談！！

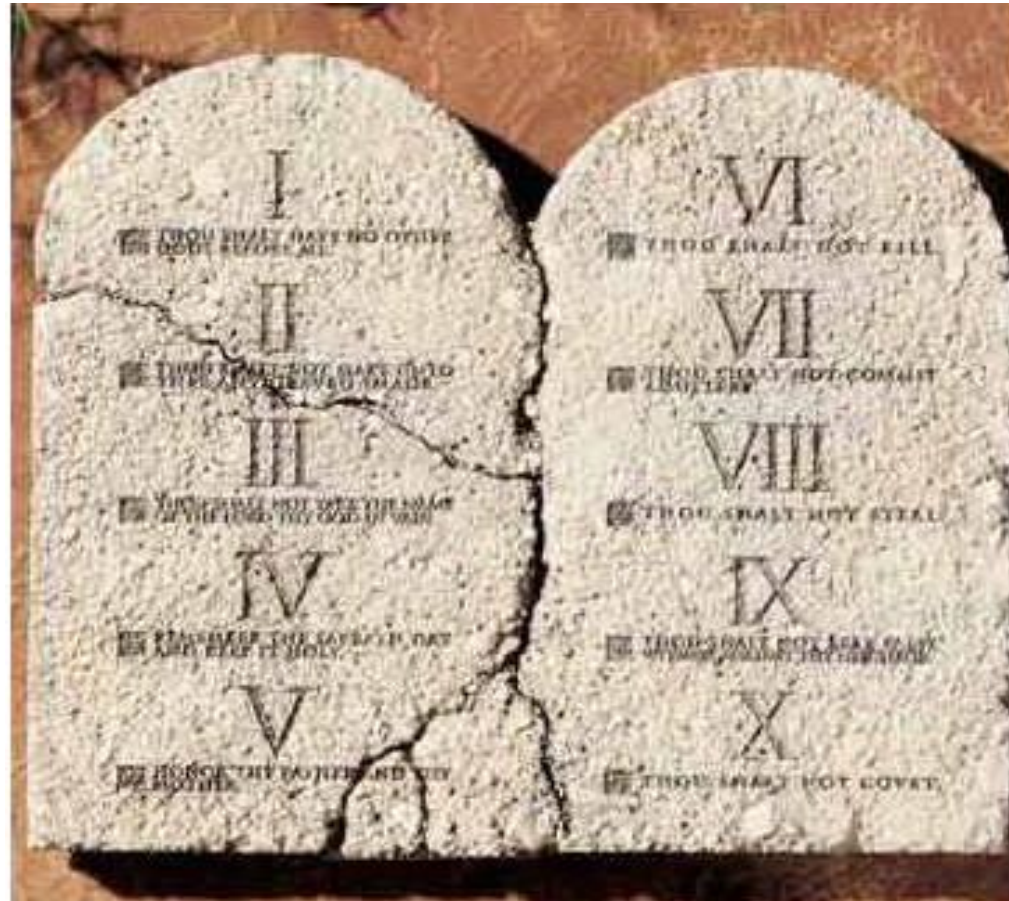
邏輯 與 真理
之間的關係…

人類產生的疑惑…

神與人爭論的結果...

神「妥協」了...

THE TEN COMMANDMENTS



邏輯就是與神爭論的技術

by 小室直樹



(2002)

邏輯就是
使人相信我說詞的一種表達過程

再想一次…邏輯是…

邏輯就是
探索真理的一種思考過程

舉證來證明自己的邏輯是對的!!

換言之，用數據說話!!

為什麼要用數據？

風能吹起大象嗎？



1876年，曾經在印度新德里南方150km的薩沙盧郊外的某個小村落裡，發生了每小時風速超過200公里的龍捲風，村落裡的大象被颶風捲了起來，在空中停留了15秒。。

我是...隨意編出來的... ^O^;;

1876年，曾經在印度新德里南方150km的薩沙盧郊外的某個小村落裡，發生了每小時風速超過200公里的龍捲風，村落裡的大象被颶風捲了起來，在空中停留了15秒。

邏輯就是
使人相信我說詞的一種表達過程

而用數據舉證
是使他人能夠相信我說詞的一種方法

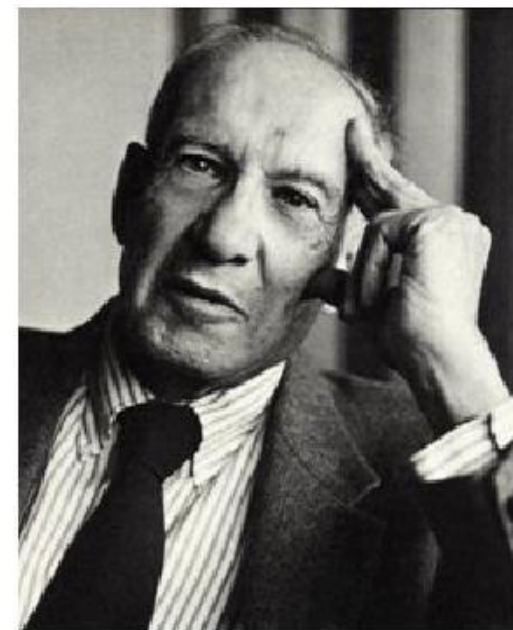
老師養不起家...

老師養不起家...因為有**10**個孩子



“In God we trust, all others must bring data”
- *Edward Deming (1900-1993)*

“What gets measured, gets managed”
- *Peter Drucker (1909-2005)*



- 資料科學提供一個**非主流**的思考方式，同時用了很多的**故事**及龐大的研究**數據**去輔佐說明：「現在主流的**思考問題**的解決方法，其實某種程度上造成更多的**問題**」。
 - 誘因是現代生活的基石
 - 知道該測量什麼、該如何測量，可以讓複雜的世界大為簡化
 - 傳統看法往往是錯誤的
 - 「**相關**」不等同「**因果關係**」

- 然後用許多不同的故事，或是過去前作提到的**案例**，去闡述這些概念的**價值**。

資料科學的題目

天下雨 → 地濕 → 路滑 → 交通事故增加 (X)

天下雨 → 地濕 → 路滑 → 交通事故降低 (O)

背後的故事 **story**

□ 誘因 (Incentive)

「道德不會改變人的行為，價格才會！」

——歐巴馬總統經濟顧問 Austan Goolsbee

□ 例如

- 樂透累積很多期的獎金後，大家一定會忍不住去買
- 學校提供獎學金
- 下星期汽油油價上漲，這星期很多車主會先去加油

□ 非裔美國人罹患心血管病的機率為何比較高？

□ Why?

- 非裔美國人得高血壓的機率較白人高50%
- 明顯的刺激因子：飲食、抽菸、貧窮等都無法解釋
- 加勒比海的非裔種族罹患高血壓亦較高，但現居非洲的種族，統計上患病的機率則跟美洲種族無異。

- 哈佛經濟學者Roland Fryer的觀察，在偶然機會中，看到一本黑奴販賣的交易書...



- 非裔美國人罹患心血管病的機率為何比較高？
- 昔日奴隸貿易的篩選，可能是非裔美國人心血管疾病罹患率較高的根本原因
 - 黑奴從非洲運送至美洲飄洋過海途中，常因長途旅程中脫水死亡
 - 「鹽耐受性」較好的人，較不容易脫水，體質能留住鹽分，就能留住更多的水分
 - 商人(舔臉)找出鹽耐受性較高的黑奴，降低風險
 - 此種鹽耐受性體質是心血管疾病的高度遺傳特徵
 - 此種“人擇”的結果，造成目前非裔美國人心血管病的機率較高

要不是有利可圖，誰會無
聊去舔別人的臉!?

□ 眼鏡蛇效應

□ 印度被殖民時期，因為許多居民被眼鏡蛇咬，受傷致死

□ 為減少當地眼鏡蛇的數量，英國政府懸賞殺眼鏡蛇換獎金

□ 結果...

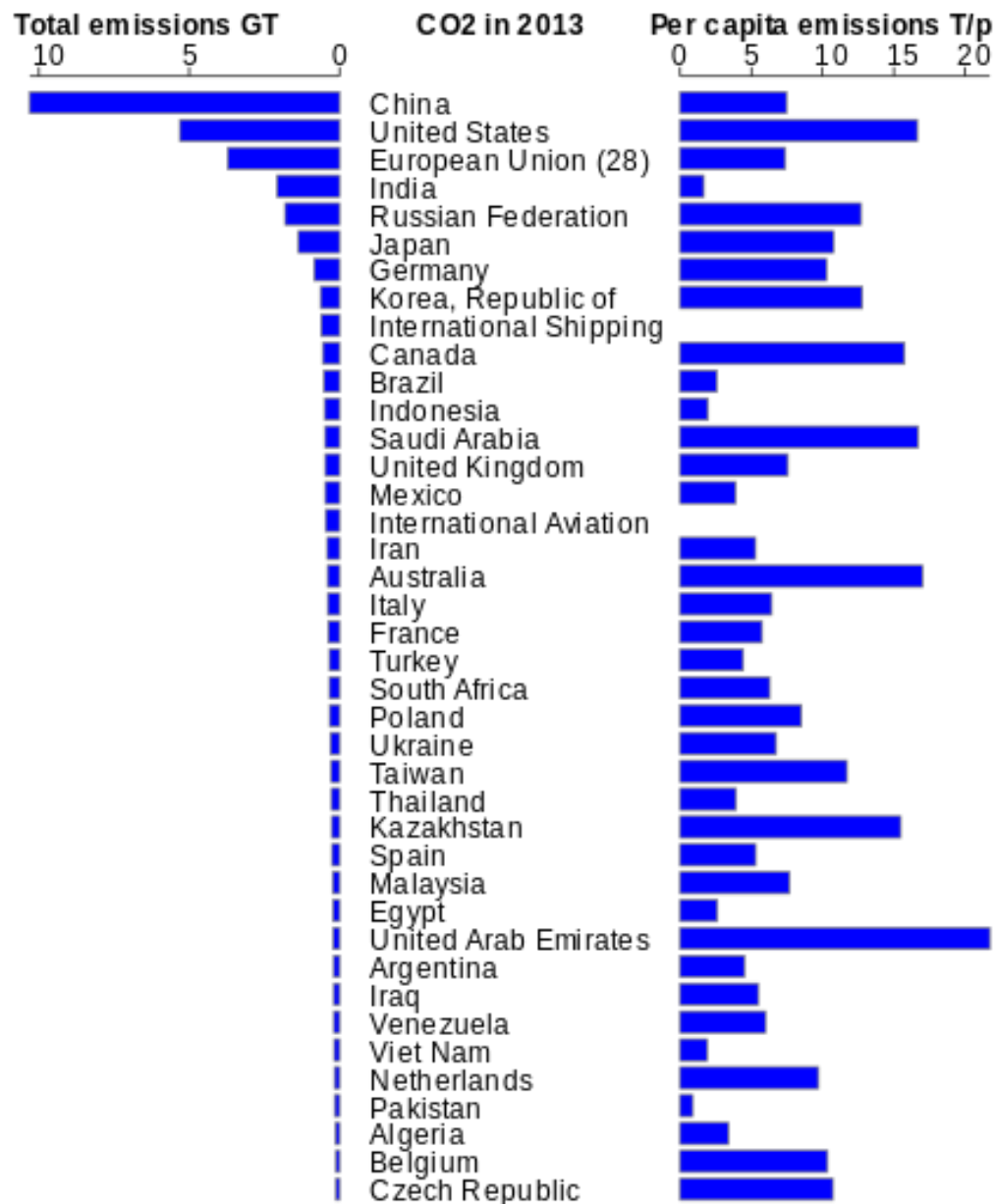




在永續環境的議題裡，經常檢討
“哪一個國家排放較多的二氧化碳？”

如何收集資料呢？

知道該測量什麼、該如何測量



□ 父母對子女成績的影響？

□ 幼兒長期研究計畫

- 美國1990年代晚期，全國各地選出共2萬名以上學童詳細調查背景資料，並測量由幼稚園到五年級的學業進步情形

□ 家中藏書豐富，是否讓小孩在學校表現優良？

- 迴歸分析(Regression)
- 家中藏書豐富的小孩，成績優於沒書的小孩
- 但家中藏書或許只反應家長所得高低，成績高低可能有其他因素影響

Levitt and Dubner, 2009. Freakonomics (蘋果橘子經濟學).

□ 父母對子女成績的影響？

- 哪些是與考試成績高度相關的家庭因素？
 - 父母教育程度高
 - 家庭關係親密
 - 父母社經地位高
 - 最近搬到較好的社區
 - 母親生第一胎時 30 歲以上
 - 小孩出生時體重偏低
 - 小孩參加過學前輔導
 - 母親在小孩出生後到上幼稚園前沒有上班
 - 父母在家中說英語
 - 父母會定期帶小孩上博物館
 - 小孩為領養
 - 小孩常挨打
 - 父母參與學校家長會
 - 小孩常看電視
 - 家裡有很多書
 - 父母幾乎天天唸書給小孩聽

釐清真相：傳統看法往往是錯誤的

□ 父母對子女成績的影響？

- Who you are matters much more than what you do !
- 重要是家長「是」什麼樣的人，而非家長「做」了什麼

家長「是誰」高度相關	家長「做什麼」低度相關
教育程度高	家庭關係親密
社經地位高	最近搬到較好的社區
母親生第一胎時30歲以上	母親在小孩出生後到上幼稚園前沒有上班
小孩出生時體重偏低	小孩參加學前輔導
在家中說英語	定期帶小孩上博物館
小孩為領養	小孩常挨打
參與學校家長會	小孩常看電視
家裡有很多書	幾乎天天唸書給小孩聽

Levitt and Dubner, 2009. Freakonomics (蘋果橘子經濟學)

Mark Twain once said, “There are three kinds of lies: Lies, Damned lies, and **Statistics.**”

馬克吐溫曾說過，謊言有三種：謊言、該死的謊言、**統計數字**。

- 因此，單純透過簡單的相關性分析，不可輕易下結論

- 因果必相關，相關不一定有因果
 - 天下雨則地濕
 - 但地濕則“不一定”天下雨

- 解決問題的基本步驟：瞭解特定情境下，所有相關人士的誘因

- 勿聽其言，而要觀其行(observation)
 - 眼見為憑，而非道聽塗說，無的放矢。

如何透過因果關係，找**根本原因**？

請常問 WHY?

WHY?

WHY???

□ 探索頻道 (discovery channel) - 為什麼忽必烈攻打不下日本?

- 十二世紀末十三世紀初，在元太祖成吉思汗的領軍下為元朝建立了良好的基業，蒙古軍驍勇善戰，不但入主中原而且遠征歐洲。清朝史學家魏源《元史新編》寫道：「帝深沉有大略，用兵如神，故能滅國四十，遂平夏克金，有中原三分之二。」。接著在其孫元世祖忽必烈登基後，統一天下，元朝疆域空前遼闊，是元朝的鼎盛時期。
- 但在東征西討的過程，元朝曾要求日本臣服且接受朝貢，但遭拒絕，因此出兵攻打日本。



- 1274年發動第一次侵日戰爭，在征東都元帥忻都領軍下，以三萬兩千餘人的軍隊遠征日本。
- 在歷經兩個月的征戰，由於後援不足，多數將領主張撤退，結果遭到颱風侵襲，回到中國時只剩下約一萬三千五百人。在戰爭失敗後，元朝令欽差大臣杜世忠等人出使日本，要求日本稱臣，結果遭斬首。
- 忽必烈一氣之下，於1281年發動第二次侵日戰爭，除了蒙古軍與漢族士兵外，更統整高麗軍三萬人，總共募軍十萬餘人，乘戰船千艘，進軍日本。
- 不幸地是，元軍依然遭到颱風侵襲，船艦大部分沉沒，溺死近半，最後大敗，僅不到十分之一生還回國。

- 擁有絕對優勢的蒙古軍，為何兩次在日本大敗而歸，成為了令人匪夷所思的謎題。
- 第一次WHY!
- 「為什麼蒙古軍會輸？WHY？」
- 為了解開謎團，日本科學家展開了辯證之旅。他首先言就兵器，發現蒙古軍弓箭強且騎術與戰鬥技巧都遠勝日本武士，更有火炮與炸彈，沒理由吃敗仗。結果從一幅日本古畫中發現到線索，其畫中描述日軍偷襲蒙古軍大獲全勝。因此科學家揣測日軍趁蒙古軍遠征日本水土不服之際，乘夜「偷襲」，挫敗蒙古軍。

□ 第二次WHY!

□ 「為什麼蒙古軍會被偷襲？WHY？」

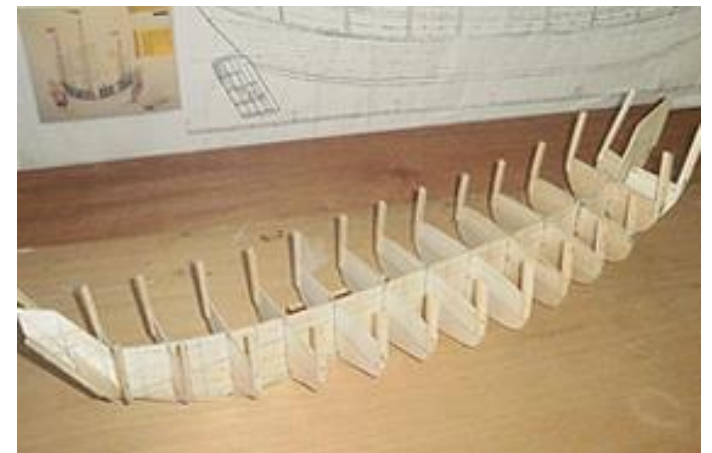
□ 日本科學家接著探討，發現蒙古軍兵力之大是不可能被輕易偷襲的，況且驚動一條船，其他船隻會馬上支援，不可能全軍覆沒。結果他從史料中發現，當時日本武士頌揚「神風（颱風）」之助，而大敗蒙古軍。



□ 第三次WHY!

□ 「為什麼蒙古船隻會被颱風輕易摧毀？WHY？」

□ 日本科學家進一步探討，發現存活的都是將領與尉官，士兵幾乎全滅，WHY？他開始調查船隻，把軍官組與士兵組的船艦作一比較，結果發現士兵組船隻吃水較淺，而且船身的「**龍骨（底盤結構）**」都毀壞了。龍骨當初設計就有問題，用如此設計不良的龍骨在大海中航行，能夠撐到日本已是奇蹟，即便勝利班師回朝，能否安全回到中國都是問題，因此有無颱風蒙古軍都會全軍覆沒。

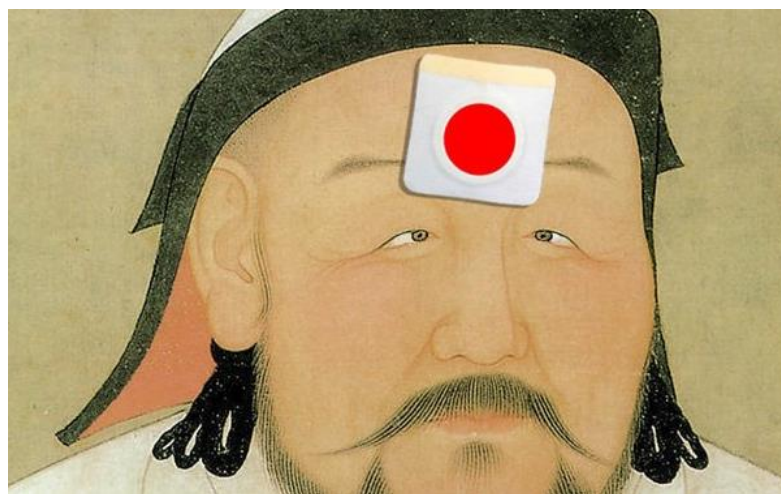


□ 第四次WHY!

□ 「為什麼蒙古船隻龍骨都不合安全標準？WHY？」

□ 為了找出真相，日本科學家遠赴中國考察，結果發現了驚人的事實，當時蒙古士兵所乘的船艦更本不是戰船，而都是「漁船」。當時忽必烈因求勝心切下令一年內迅速造艦，否則嚴懲官員。以精良匠工的技術造一百艘船艦，也需五年時間，可是當時官員為了保命，強徵民工強占民船，這些工匠不知民船行於河川的建造標準，是完全不適用於海洋航行。為求活命，龍骨建造材料與製造工藝十分粗糙，蒙古軍當然大敗。

- 在這一系列的分析、深入探討後，日本科學家下了一個出人意料的結論，我想你應該也猜出來誰是根本問題了。
- 對！讓十萬大軍斷送生命死於異鄉的，不是日本武士的偷襲，不是颱風，更不是龍骨，而是「忽必烈」本人。
- 這個案例告訴我們，當你的WHY問到第四次第五次時，問題也隨之真相大白。



觀察現場，找出問題的本質，
透過資料收集舉證，
經由思辨的過程，
找出真相，說故事並解釋因果，
提出建設性的誘因設計，
改善現狀。

簡言之，就是...探索真理!!

研究方法思考

俗話說的好：「天下沒有白吃的午餐」，夢幻職業看似美好，背後卻是無數辛酸血淚換來啊！

外表光鮮亮麗 辛酸不為人知的職業

排名	夢幻職業	職場現實辛酸	網路聲量
1	教師	孩子頑皮、家長恐龍	80,159
2	創業家	高收入負擔高風險	80,142
3	記者	沒有尊嚴	47,107
4	立委	政治路一步錯步步錯	35,852
5	醫生	高壓血汗	29,929
6	總統	需要有被討厭的勇氣	24,137
7	明星	容易被時代淘汰	20,196
8	總裁	扛起公司重擔	16,419
9	工程師	勞累傷神傷身	11,190
10	導演	苦撐等成名那天	11,121

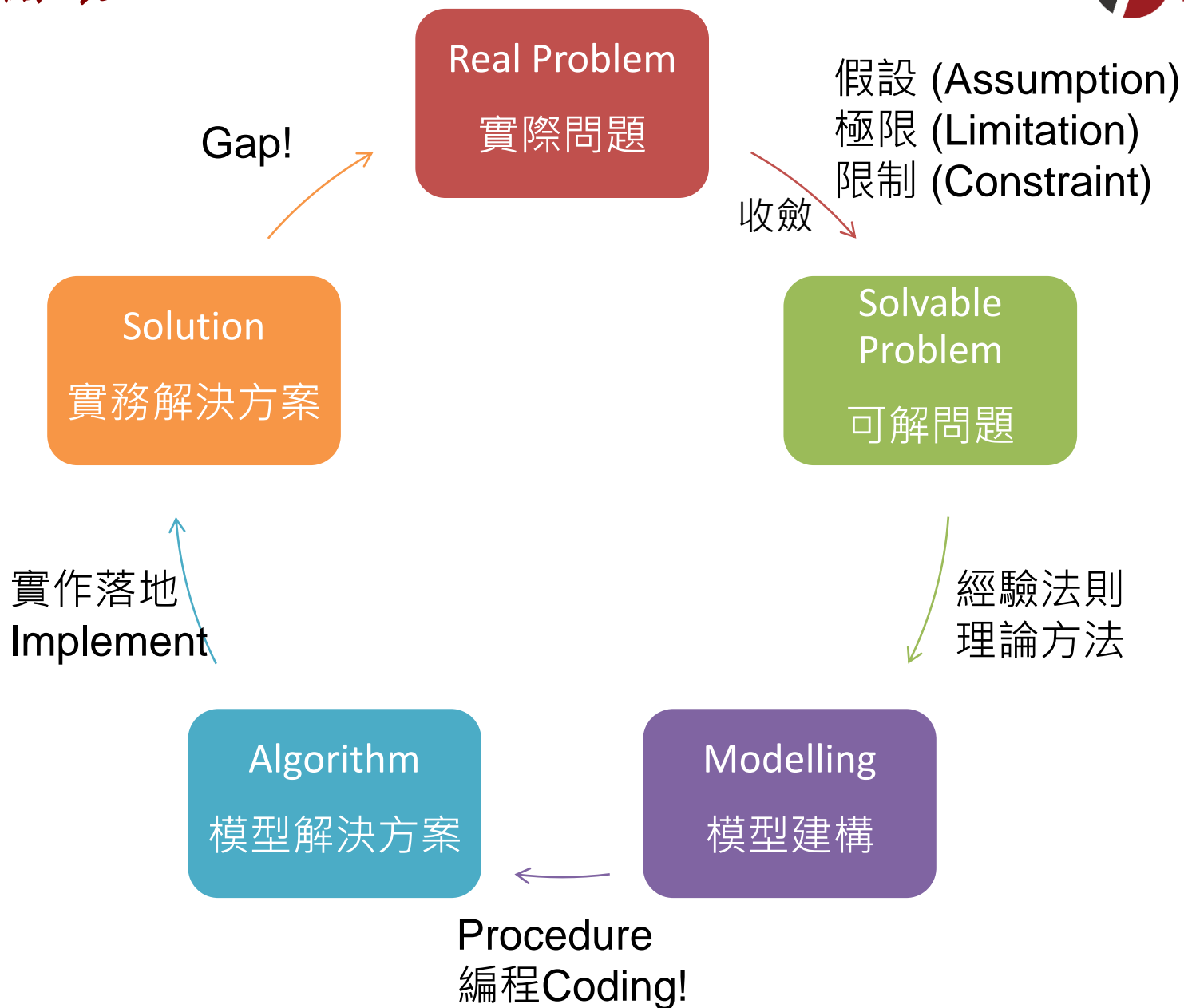
· 資料分析：DailyView網路溫度計 透過 KEYPO大數據關鍵引擎 (keypo.tw) ·
 以國際級的語意分析架構、先進的機器學習技術與人工智慧推論引擎，
 感知網友語意脈絡與情緒，分析時事網路大數據。
 · 分析期間: 2016/09/19~2017/09/18



https://www.jobforum.tw/discussTopic.asp?cat=NCKU&id=126520&agent=out_fan_ETTC_ncku_university_1710111

研究：探究問題本質的思考邏輯過程

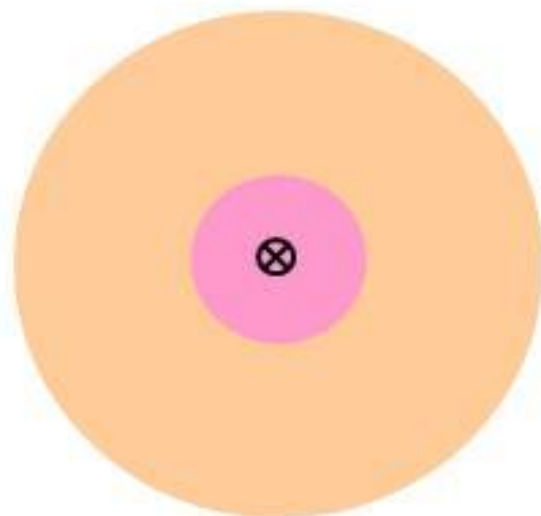
研究方法循環



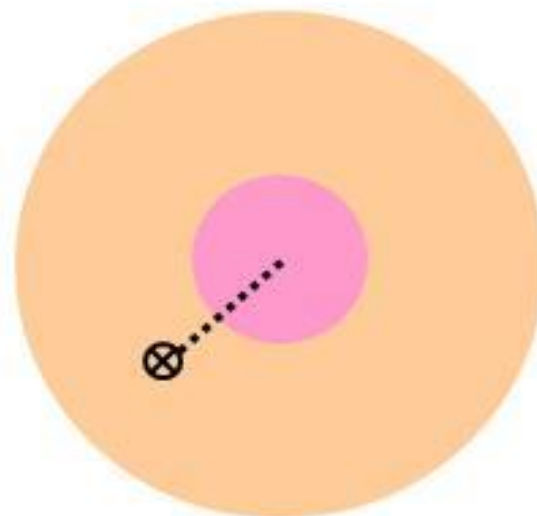
Question vs. Problem

落差 = 預期狀況 - 實際發生
(Gap = Expected - Actual)

沒有落差



有落差
→ “問題”發生

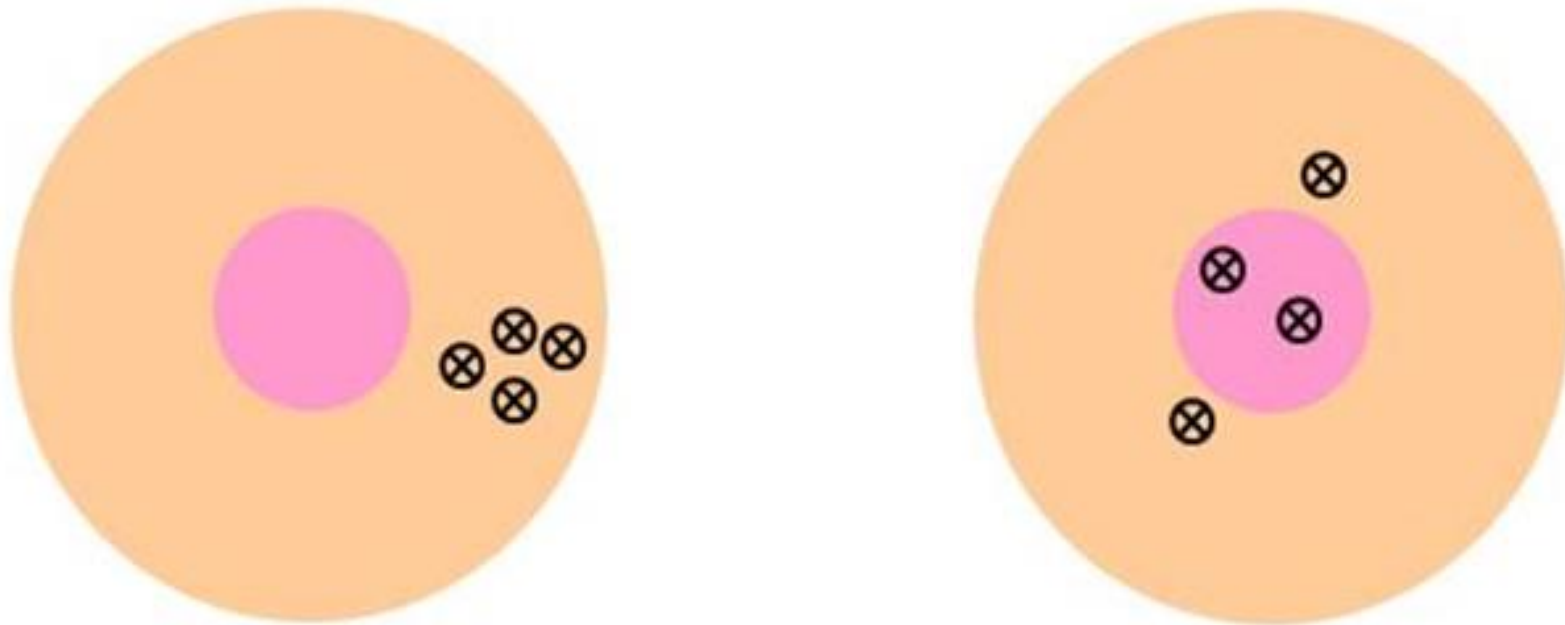


⊗ : 現狀

● : 預期景象

..... : 落差

Shooting Problem



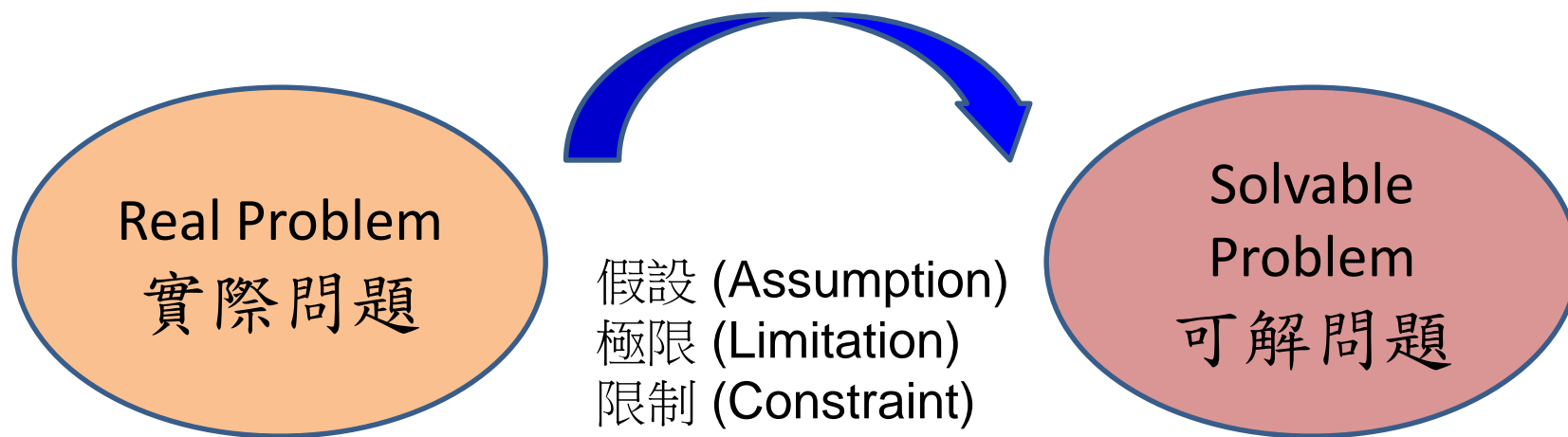
對！

釐清兩者間的差異（問題本質）

就是你們自身的邏輯思維

所以，研究生是…？

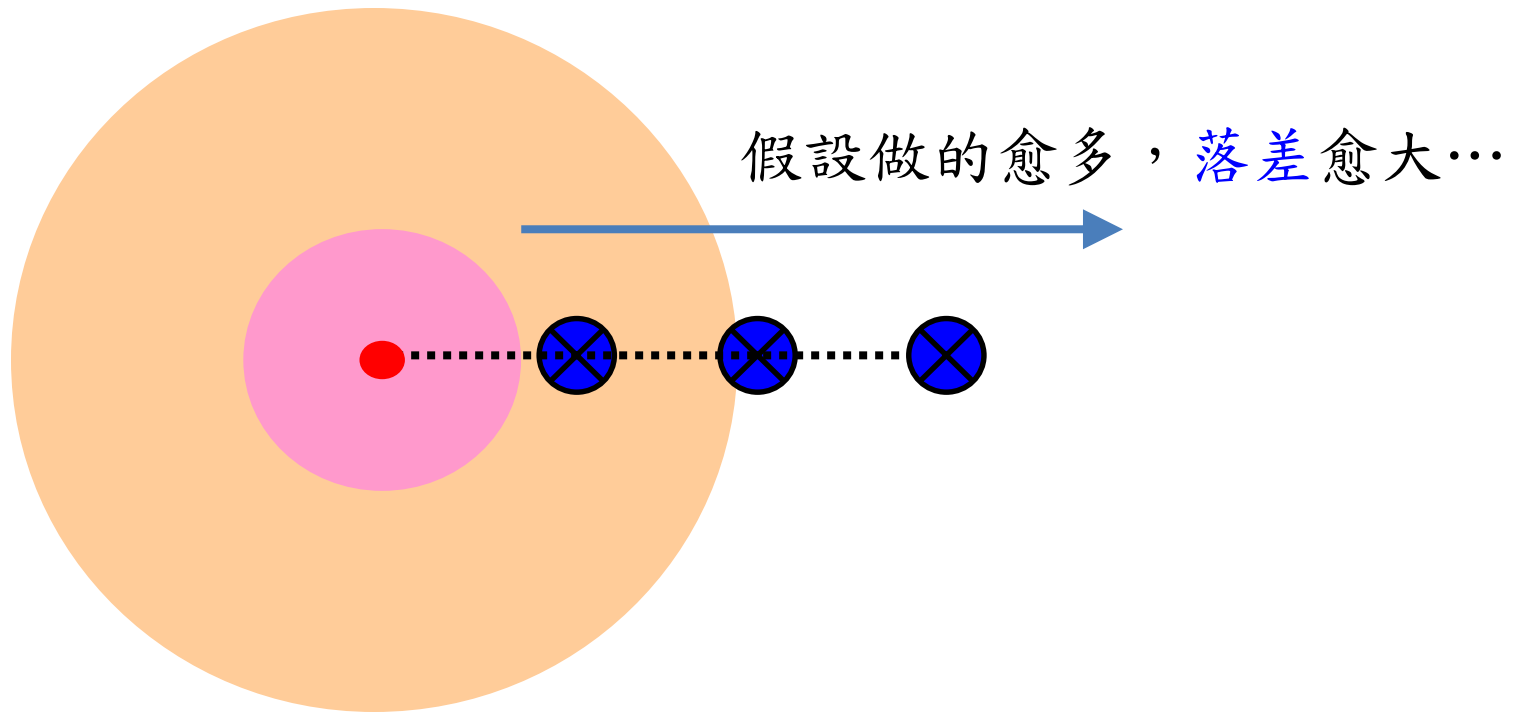
以思考邏輯來探究問題本質之人



假設：為了簡化問題

人心是善良的？人心是抵擋不了誘惑的？

人是懶惰的？人是努力上進的？



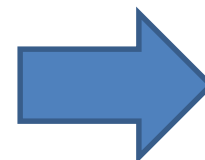
本來是為了解決問題而做的研究
後來卻什麼都沒有解決...

其實，假設是用來被打破的…!!?

顛覆假設!?

馬戲團 (Circus)

	假設
地點	室外大帳篷
價格	低廉
主角	動物
叫賣	大聲叫賣招徠顧客
節目安排	好幾個節目同時進行
音樂	好玩的音樂
攝影	可



顛覆假設
室內飯店
昂貴
人
沒有人大聲叫賣
同一時間只演出一個節目
精緻的音樂
禁止



做研究 ? 做決策

如何選取好的研究題目!!!??

Literature review go first

VS.

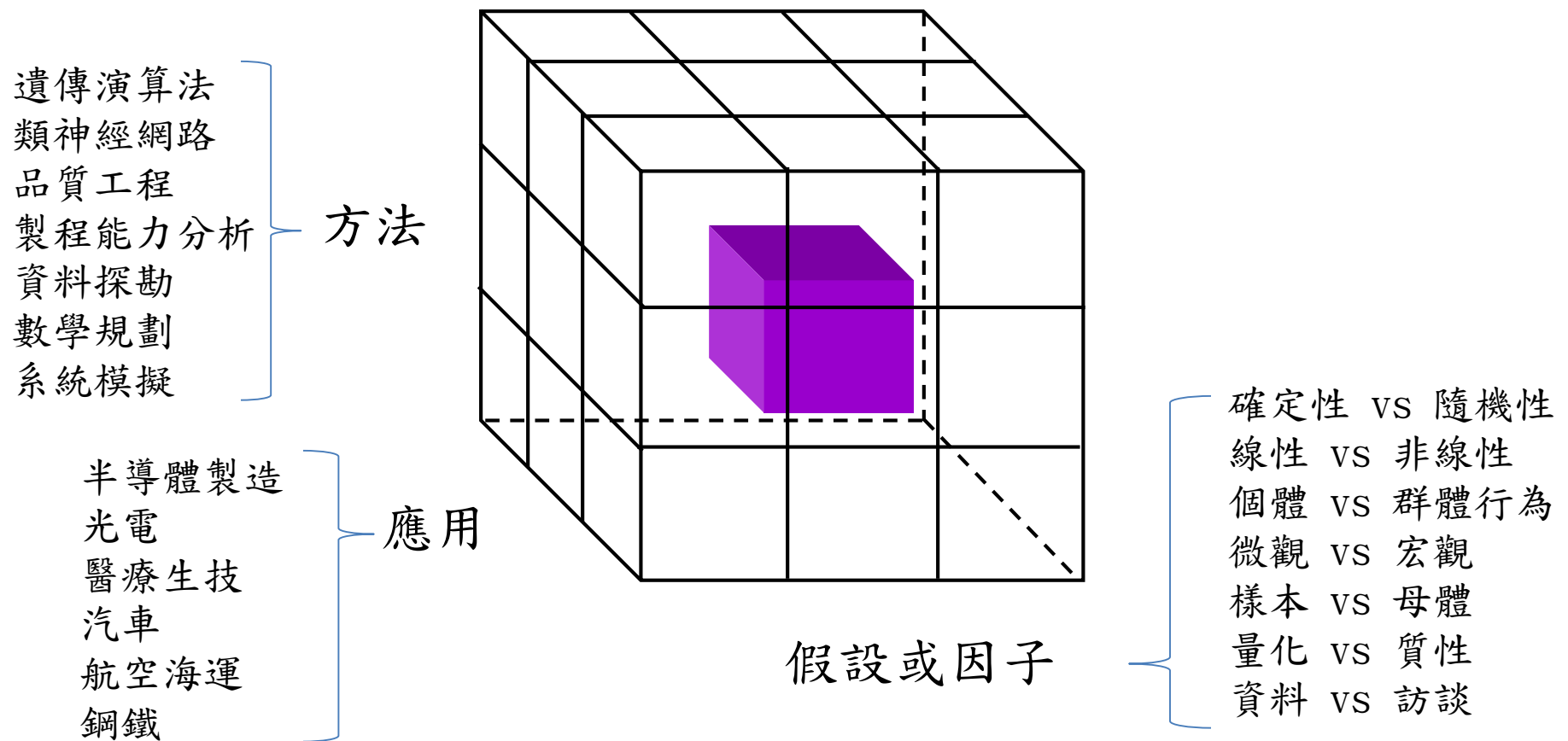
Just do it!

then add a few literatures...

Akerlof (2001) vs. McFadden (2000)

一個好的研究題目..
必然存著高風險..
可能做不出來..或失敗..

高風險、高報酬，是必然的道理



文獻回顧 → 為問題/研究定位
→ 界定利基(niche)

思考是研究的過程…
決策是研究的目的…

研究結果提供了資訊來輔助決策的執行…

Research



構析(Analytics)五階段



U

N

I

S

O

N

Lee, C.-Y., and Chien, C.-F., 2020. Pitfalls and Protocols of Data Science in Manufacturing Practice. Journal of Intelligent Manufacturing.

結語

台灣五大資料經濟人才需求趨勢

2015年從業人數 2018年需求人數



2017全美最棒的工作

排名	職務名稱	年薪中位數 (美元)
1	Data Scientist	\$ 110,000
2	DevOps Engineer	\$ 110,000
3	Data Engineer	\$ 106,000
4	Tax Manager	\$ 110,000
5	Analytics Manager	\$ 112,000
6	HR Manager	\$ 85,000
7	Database Administrator	\$ 93,000
8	Strategy Manager	\$ 130,000
9	UX Designer	\$ 92,500
10	Solutions Architect	\$ 125,000

資料來源：glassdoor

Glassdoor, 2018. https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm



Operations Research Analyst Overview

Overall Score 7.2 / 10

#4 in Best Business Jobs | #10 in Best STEM Jobs | #20 in 100 Best Jobs

Overview Salary Reviews and Advice Job Listings

What is an Operations Research Analyst?

Operations research analysts are high-level problem-solvers who use advanced techniques, such as optimization, data mining, statistical analysis and mathematical modeling, to develop solutions that help businesses and organizations operate more efficiently and cost-effectively. For example, UPS uses operations research to chart the flow of packages, provide real-time route guidance to drivers and help plan and manage distribution. In the health care field, the Memorial Sloan-Kettering Cancer Center in New York used operations research to design a radiation treatment plan for prostate patients using sophisticated modeling and computation techniques.



MEDIAN SALARY

\$83,390

UNEMPLOYMENT RATE

2.1%

USNews, 2020. Best Business Jobs. <https://money.usnews.com/careers/best-jobs/operations-research-analyst>

□ 挑戰自己

- 不要活在別人的思想裡
- 不要怕犯錯，每條路都可以嘗試看看...
- 賈伯斯在2005年史丹佛畢業典禮的演說...
- 「既然你終究會死，就不必擔心冒險會讓你輸得精光。反正遲早，你都會輸得精光。」

□ 研究

- 從課堂、讀paper、產學/專案找到相關**延伸性**的題目
- 從**點**上出發，改善**文獻中的既有方法**或**特定假設下的應用**
- 從**面**上探討研究，長期短期或微觀綜觀所帶來的**impact**可能不同
- 學術 vs. 產學 (可能會忙碌...但產學有時候會碰到**很有趣**的題目...)
- 切入**社群**找到**合作夥伴**
- 實戰實學!!!!!!!

你可能可以思考的是...

學兩個專長(跨領域)

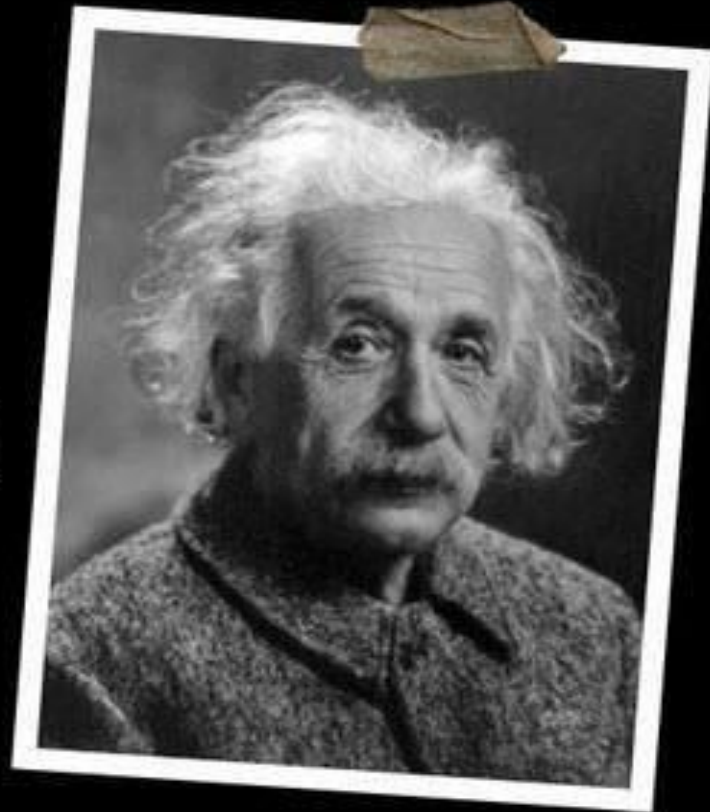
然後激盪出屬於你自己的火花

一位日本小學校長的標語



積少成多..聚沙成塔..

**"Education is not
the learning of
facts, but the
training of the mind
to think."
-Albert Einstein**



More @ [QuotesDump.com](https://www.QuotesDump.com)

感謝大家的支持跟參與 還請多多指教



Contact Information:

name: 李家岩 (Chia-Yen Lee)

phone: 02-33661206

email: chiayenlee@ntu.edu.tw

web: <https://polab.im.ntu.edu.tw/>

台灣人工智慧學校

智慧製造與生產線上的資料科學

http://polab.im.ntu.edu.tw/Talk/Data_Science_in_Manufacturing.pdf

- 蕭瑞麟教授《不用數字的研究》
- 蘋果橘子思考術
- 蘋果橘子經濟學
- 王伯達，預見未來
- 陳昇瑋(2015)，如何培養資料科學團隊，中研院。
- 2012.從統計看世界-魔球 Moneyball.
<http://hbzstat.blogspot.tw/2012/03/moneyball.html>