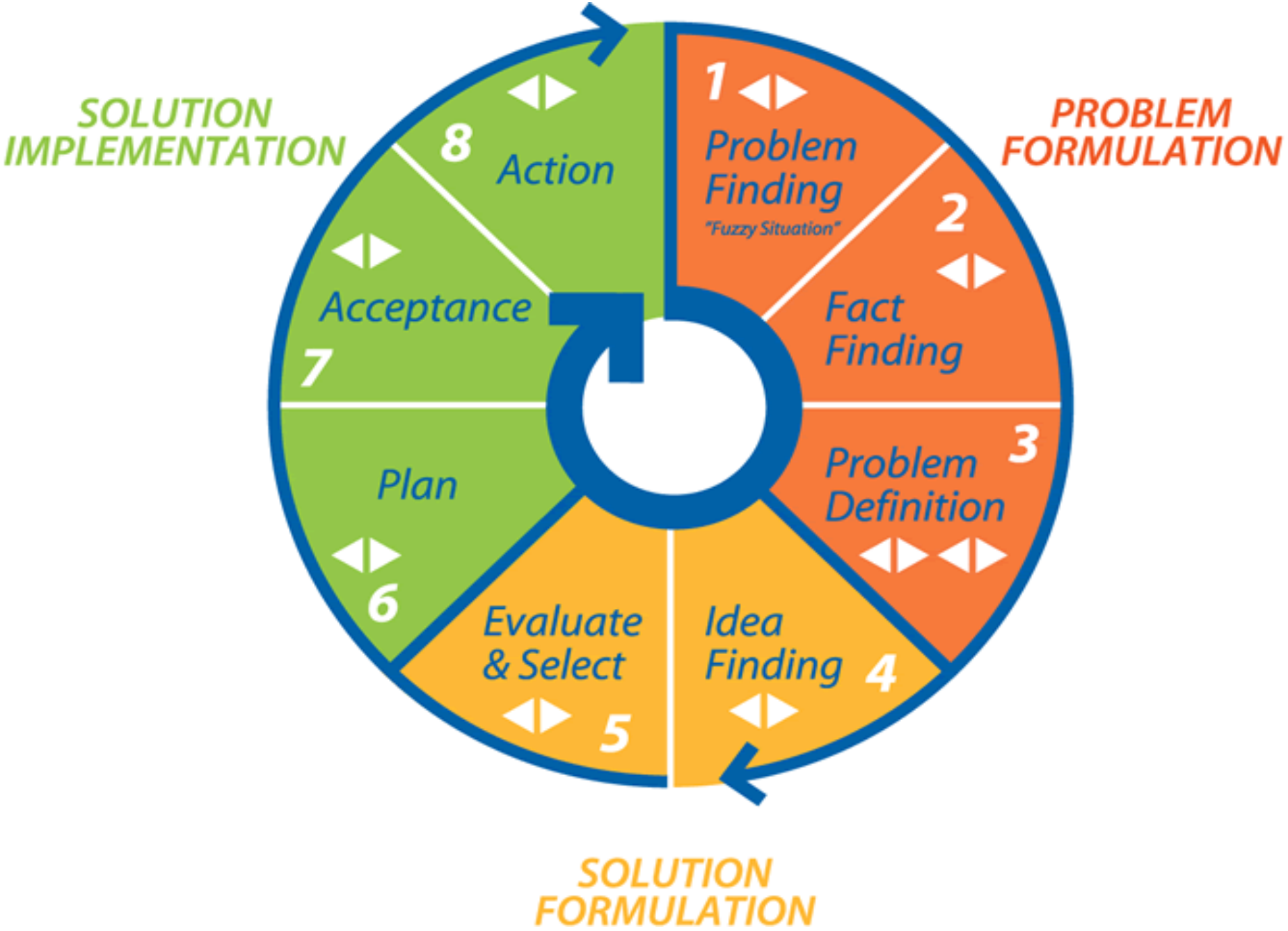
巨量資料與統計分析

政治大學統計系余清祥

2024年11月05日

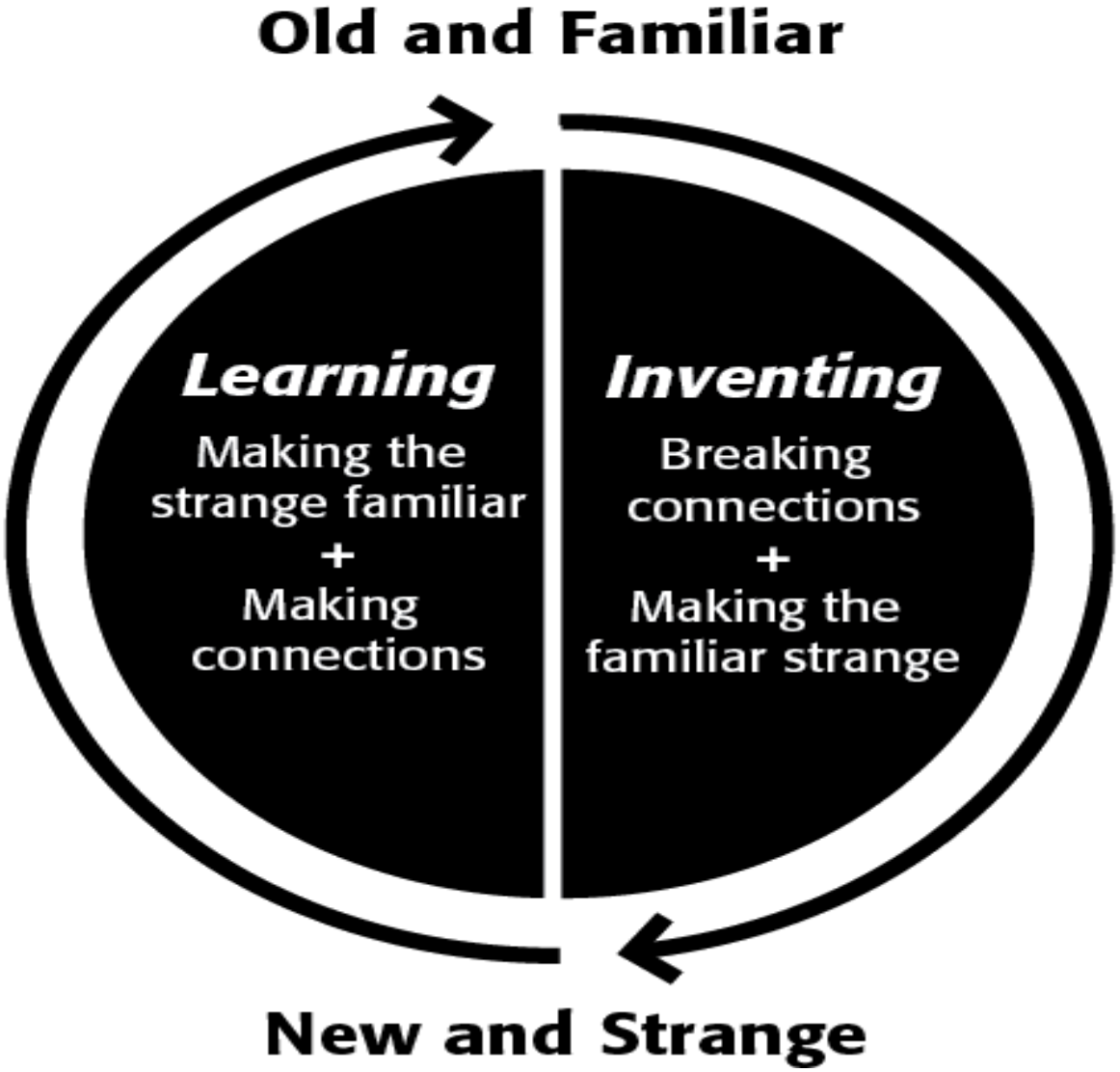
第八週：問題導向

http://csyue.nccu.edu.tw



http://www.basadur.com/insightsresearch/OurThoughtsonCreativityandInnovation/TheProcessofInnovationLearningandInventing/tabid/168/Default.aspx

Two Halves of Simplexity Thinking: A Continuous Process of Inventing and Learning



http://www.basadur.com/insightsresearch/OurThoughtsonCreativityandInnovation/TheProcessofInnovationLearningandInventing/tabid/168/Default.aspx

如何定義決定一切

- Rittel and Webber (1973) suggests that

→The process of formulating the problem and of conceiving a solution..... are identical, since every specification of the problem is a specification of the direction in which a treatment is considered.

註：A problem's definition determines the solution space.

有趣(或殘酷)的範例

- 一位教授及其學生到阿拉斯加探勘，在一望無際的冰原上被北極熊追殺。眼看即將被追到，學生趕緊換上球鞋，教授說：「換上球鞋也跑不過北極熊。」學生卻說：「我不必跑贏北極熊，只要跑贏你就夠了。」

→真正的問題是什麼？

→問題目標vs.解決方案

問題與學習

- 大家必須學習如何自己做決定，這在任何科目都能學到，只要羅列出來的是**問題**，而非必須背誦的**事實**。
- 就像所有哲學問題一樣，沒有**標準或正確的答案**，只有對**議題的探究**，以及**弄清楚自己立場**的挑戰。
- 哲學在意的是**過程**，如何形成**主張**，如何證明結論，而**不在意結果**。

問題與學習（續）

- 哲學裡唯一重要的答案，是那些你自己想出來的答案，因此很難打分數，很難訂立標準，很難找到足夠的師資願意花足夠的時間來上這門課。
- 讀過書的人，都被教導如何理解知識，卻沒被教導**如何行動**。

——摘錄自《你拿什麼定義自己？組織大師韓第的生命故事》，天下文化出版

註：參考「三大步驟：人人都是分析師」

美國精算學會(SOA) Predictive Analytics

2. **Topic: Problem Definition, Exploratory Data Analysis, and Initial Model Selection (15-25%)**

Learning Objectives

The Candidate will be able to identify the business problem, how the available data relates to possible analyses, and use the information to propose models.

Learning Outcomes

The Candidate will be able to:

a) Formulate a business problem in terms that are amenable to an analytic solution.

b) Conduct exploratory data analysis to identify key relationships that inform initial model selection.

c) Select initial models and methods for analyzing the business problem.

問題定義與大數據

大數據分析vs.資料品質

- 大數據的4V，最需注意的是**真實性**與**多樣性**，這也與定義問題關係密切
- 許多人宣稱**資料量多寡**比**資料品質**重要，但如同大家熟知的「凡規則必有例外」，以及學英文常聽到的「It depends!」。
- 資料科學家**分析前應先確認資料來源可信度**，檢查資料品質是否有重大瑕疵

→網路有名的「世界四大不能信」：英國研究、臺灣報導、中國製造、韓國發源。

十個令人傻眼的英國研究

1. 破解千古謎題！英科學家：先有雞才有蛋
2. 英學者：鄭和最早發現美洲
3. 英國研究稱人愈來愈笨
4. 果真要衣裝，超人T恤助考試得佳績
5. 吃早餐，竟不是一天最重要的一餐？
6. 紅橙有助苗條？研究：花青素避免脂肪囤積
7. 手機細菌多，比馬桶手把髒18倍
8. 衛生先進國，花美男當道
9. 血拚23分鐘後理智崩壞，開始亂買東西
10. 哇！四成英國人曾在辦公室做愛

定義錯誤引起的笑話

- 沒有瞭解問題本質，僅憑關鍵詞和搜尋引擎，也會發生令人啼笑皆非的結果。

→前一陣子中國網軍大幅攻擊「台獨份子」，除了周子瑜受害，連蔡正元、邱毅也遭受池魚之殃。

註：蔡正元是蔡英文的弟弟！？

缺乏相關知識的誤判？

- 猜測蔡正元被誤判為蔡英文弟弟，可能來自於先前的「炒地皮請找蔡姊姊」。

→2016年總統大選前，邱毅、蔡正元開記者會指控蔡英文炒15筆土地（獲利1.8億）。

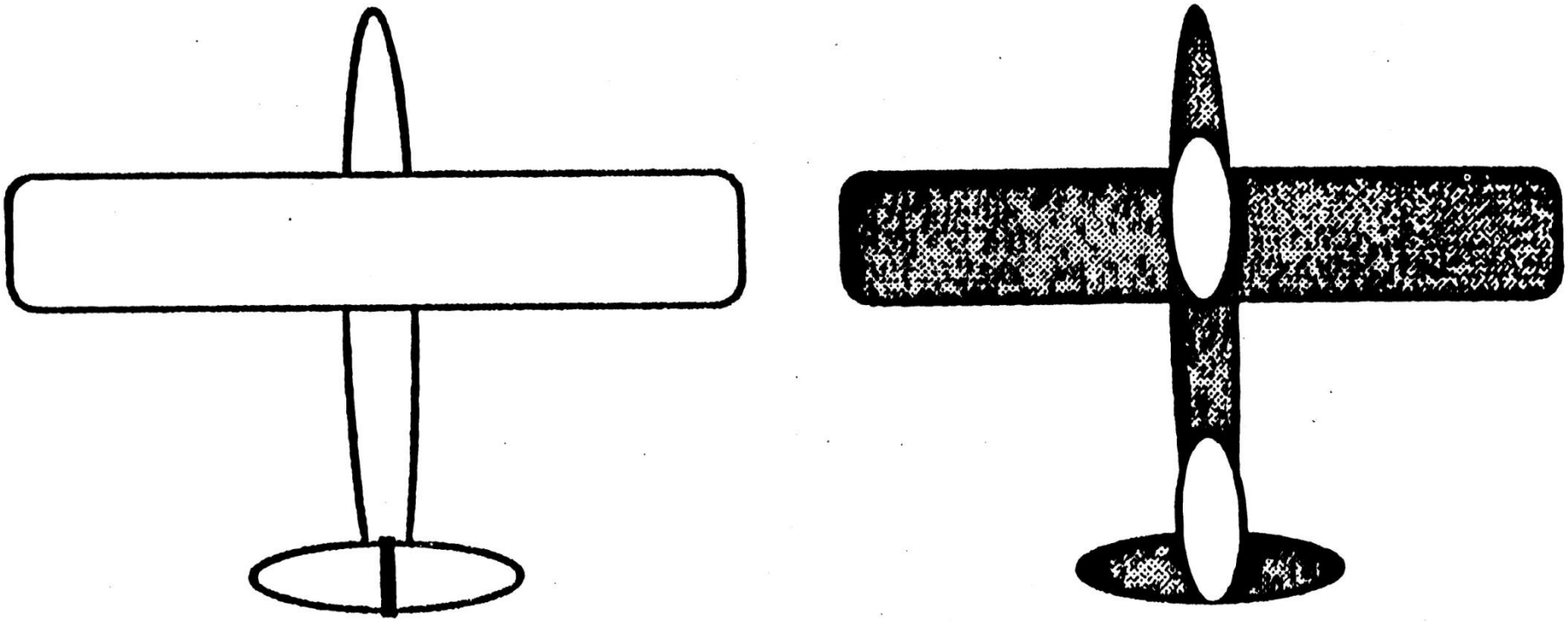
→若斷章取義由「蔡姊姊」判斷兩人關係。。

註：類似笑話也會出現在大數據的分析，可參考《生命中的經濟遊戲》的範例。

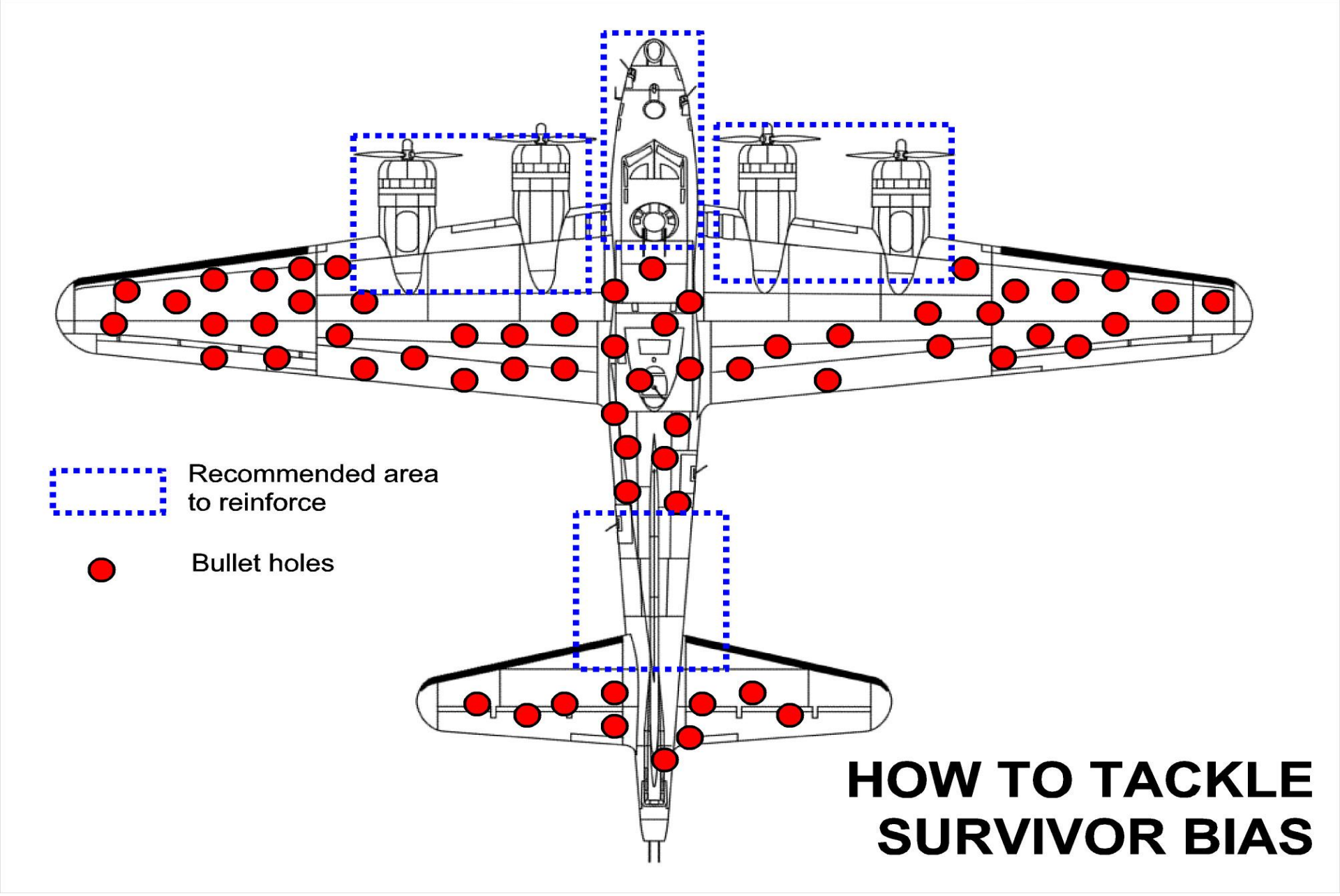
辛普森謬論Simpson's Paradox

- 男性整體錄取率較高
- 但甲、乙兩科系卻是女性錄取率較高？

	女性			男性		
科系	申請者	錄取者	錄取率	申請者	錄取者	錄取率
甲	99	4	4%	1	0	0%
乙	1	1	100%	99	10	10%
加總	100	5	5%	100	10	10%



A graphical depiction of ***Wald's bullethole*** *data.*



倖存者偏差（Survivorship Bias）

資料品質與定義問題

- 除資料品質外，**選取適合資料（包括抽樣）**也是真實性要考慮的範疇。
- 統計是由觀察值反推出發生原因，足夠觀察值可看出母體原貌（「三人成虎」），抽出的樣本必須能反映全體、亦即樣本需能代表母體。
- ***樣本代表性！！！***

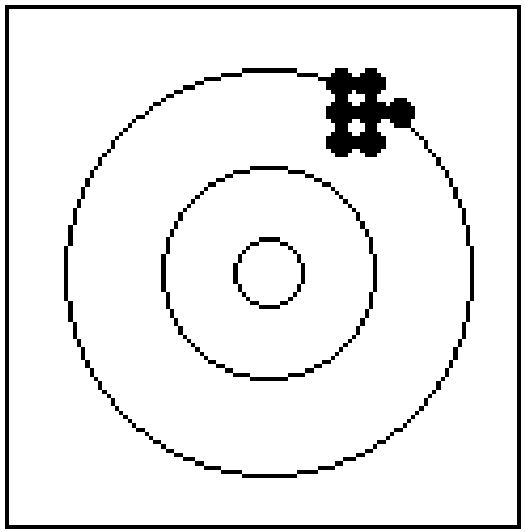
→最忌諱「瞎子摸象」（或「以偏概全」）

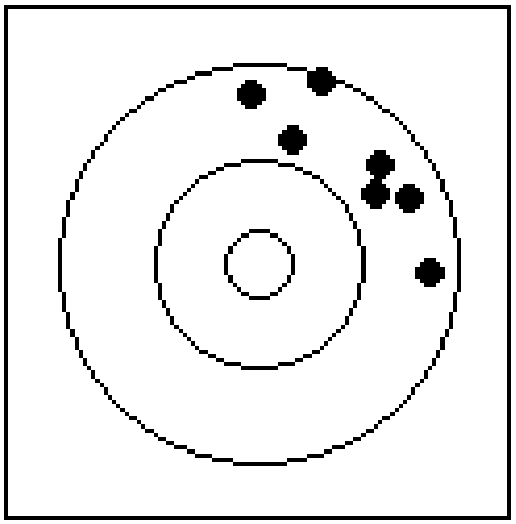
http://www.zou1.com/xinwen/201512/09/33134.html

Precise

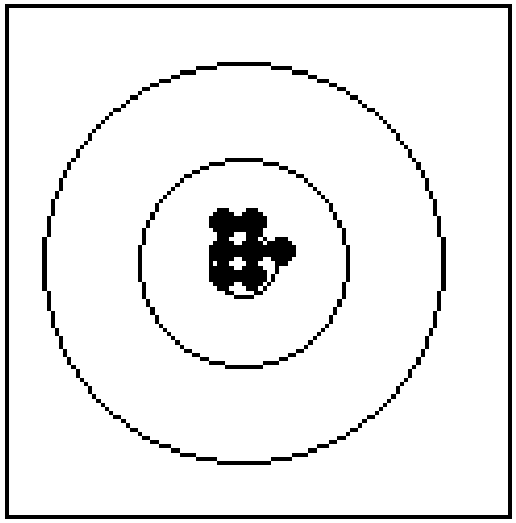
Imprecise

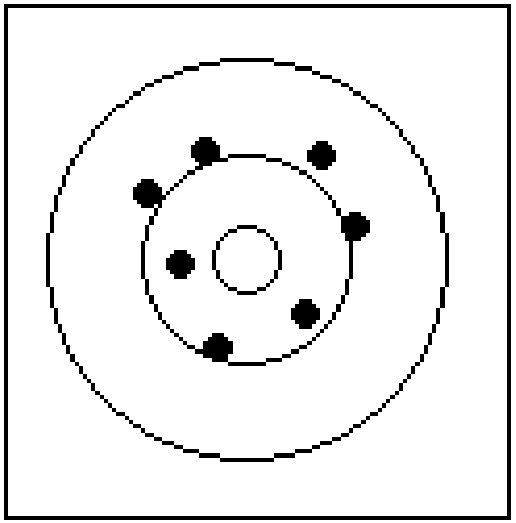
Biased





Unbiased

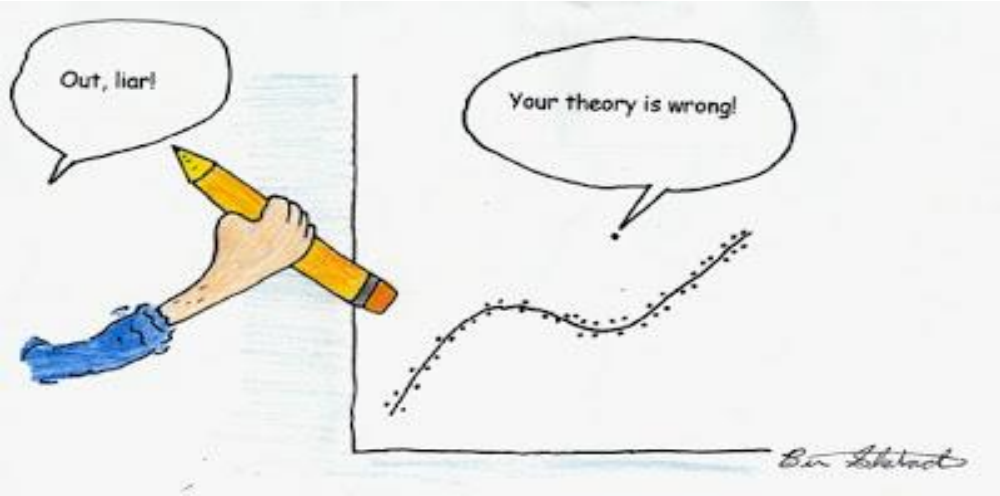




倖存者偏差（Survivorship Bias）

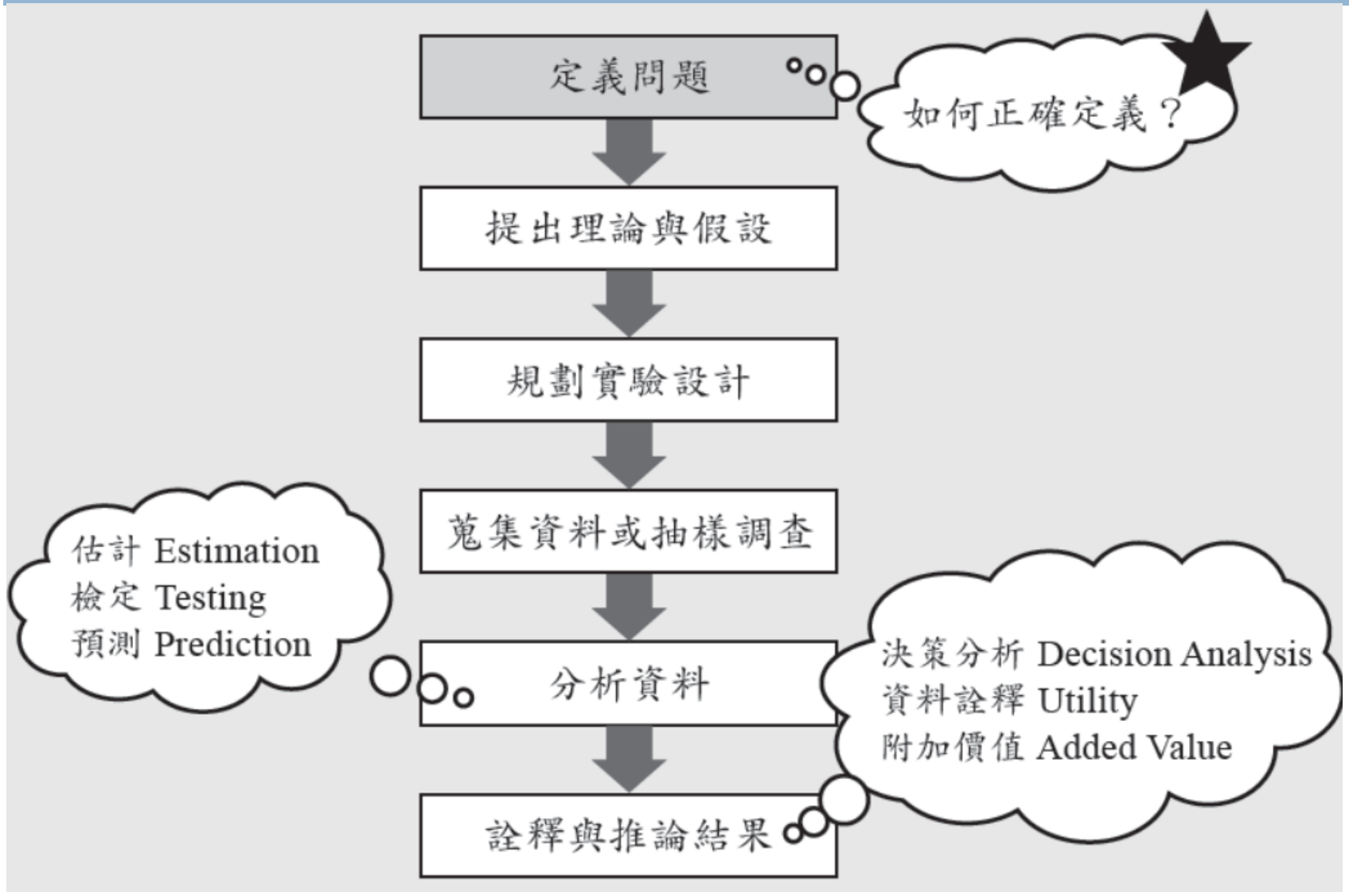
- 若不熟悉資料與問題的相關背景，有時無法判斷蒐集到局部或全體的資料，檢查是否存在倖存者偏差（例如：谷歌流感趨勢預測）。

→不少人為了省事與漂亮結果，移除了離群值！



定義問題為研究首要步驟

解決問題的步驟


定義問題
如何正確定義？
提出理論與假設
規劃實驗設計
蒐集資料或抽樣調查
估計 Estimation
檢定 Testing
預測 Prediction
分析資料
決策分析 Decision Analysis
資料詮釋 Utility
附加價值 Added Value
詮釋與推論結果

- “The formulation of a problem is often more essential than its solution.”

→如果我有一個小時拯救地球，

我會用59分鐘界定問題，然後用一分鐘解決它。

— Albert Einstein

- Exploratory research is often required to help in the formulation of the research problem.

→All research is based on a set of assumptions or factors that are presumed to be true and valid.

(It is called “hypothesis” in science.)

→But be careful not to give too many assumptions.

定義問題與統計諮詢

統計諮詢(Statistical Consultation)是什麼？

→Statistical consulting is the most challenging and most rewarding part of statistics. A consultant uses the art and science of statistics to solve a practical problem. Problems come from many different fields.

統計諮詢的進行步驟


建立良好的合作關係
Establishing rapport
確認研究問題
Identifying the research problem
設定目標
Setting Goals
責任分工
Agreeing on a division of responsibility
總結：提供可行建議
Summing-up stage

統計研究的首要步驟

- 獲取研究問題的相關背景知識
- 確立問題的目標(研究目的)
- 以統計的語言定義問題

→如果與其他人合作，儘量「多發問」！

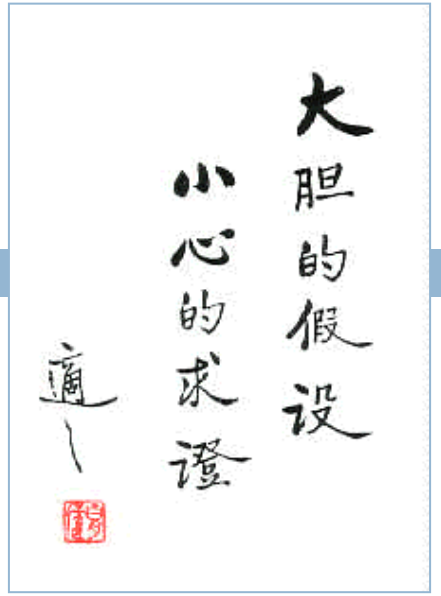
註：上述內容與統計諮詢(Statistical Consulting)吻合。

Statistical Consultant (Wikipedia)

- A statistical consultant provides statistical advice and guidance to clients interested in making decisions through the analysis or collection of data. Clients often need statistical advice to answer questions in business, medicine, biology, genetics, forestry, agriculture, fisheries, wildlife management, psychology, law, industry. The role of the statistical consultant varies from project to project.

定義問題的技巧

學習（統計）的幾個要素



- 解決問題大致有以下幾個要素：

→如何定義、測量？

(e.g. Variable Format, Data Collection)

→如何判斷、取捨

(e.g. Estimation, Prediction, Testing)

→如何詮釋、增加附加價值？

(e.g. Utility, Decision)

如何決定解決方案

- Rittel and Webber (1973) suggests that

→The process of formulating the problem and of conceiving a solution..... are identical, since every specification of the problem is a specification of the direction in which a treatment is considered.

註：A problem's definition determines the solution space.

定義問題(Problem Definition)

「正確問題的近似答案，遠比錯誤問題的精確答案有價值。」

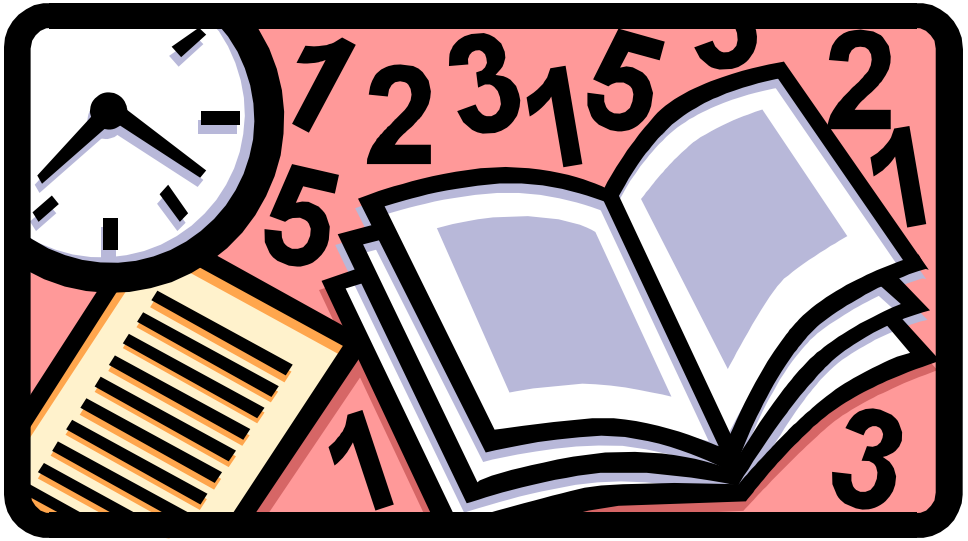
"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong problem."

--- the first golden rule of applied mathematics

統計的第三型誤差

- Type III error (error of the third kind):

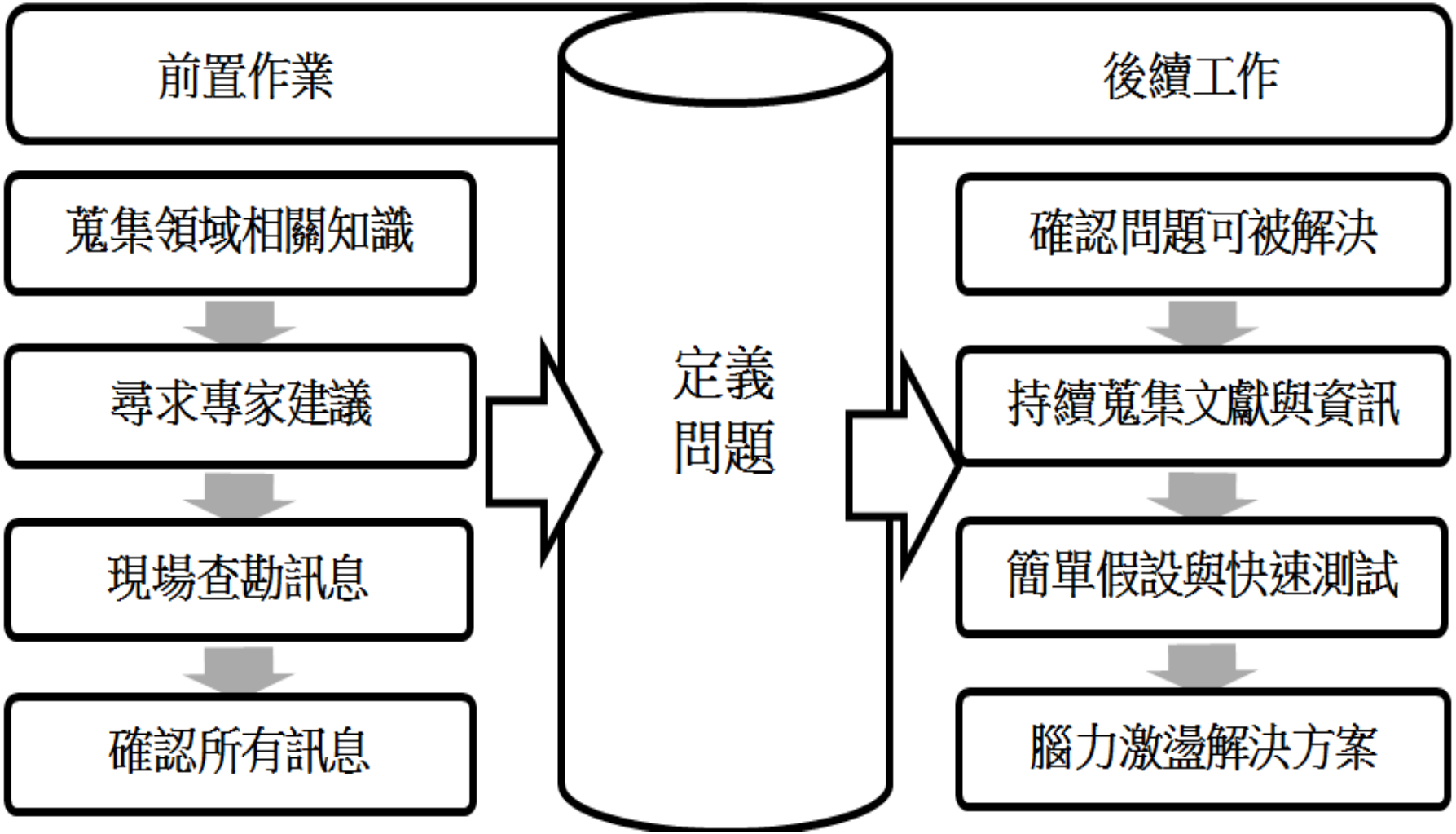
→ Giving the “right” answer to the wrong question (Kimball, 1957)



定義問題

- **A problem is decided by purposes.** E.g., manufacturing managers are usually evaluated with line-operation rate, which is shown as a percentage of operated hours to potential total operation hours. Therefore manufacturing managers sometimes operate lines without orders from their sales division. This operation may produce more than demand and make excessive inventories. The excessive inventories may be a problem for general managers. But for the manufacturing managers, the excessive inventories may not be a problem.
- Therefore, in order to identify a problem, problem solvers must clarify the differences of purposes.

定義問題的技巧


前置作業
後續工作
蒐集領域相關知識
尋求專家建議
現場查勘訊息
確認所有訊息
定義
問題
確認問題可被解決
持續蒐集文獻與資訊
簡單假設與快速測試
腦力激盪解決方案

定義問題的幾個技巧

- Finding out where the problem came from
- Explore the problem
- Present/Desire state technique
- Duncker diagram
- Statement-restatement technique

進一步探索問題

- Recall or learn the fundamental principles related to the problem
- Carry out an order-of-magnitude calculation
- Hypothesize what could be wrong
- Guess the result

定義問題練習案例

《旅館裝修電梯問題》

- 某家旅館重新整修內部，將客房數目擴增為原先的1.25倍，但電梯數維持不變，房客因等待時間增長，而抱怨連連。如果你是旅館經理，請問你該怎麼解決這個問題？增加電梯數、加快電梯速度、或是電梯門口加設電視或鏡子？

定義問題練習案例

《加強轟炸機的防護》

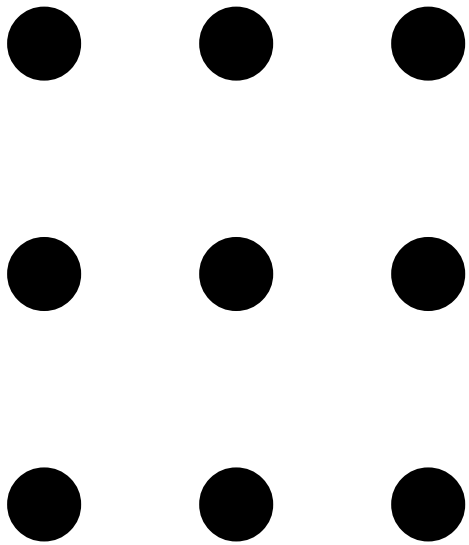
- 二次世界大戰時，知名統計學家Wald被要求研究轟炸機出任務後的存活率，在哪些地方增加裝甲可提高安全返回的機會。Wald蒐集所有飛回基地的轟炸機，統計機身上的彈孔分布（如下圖），發現只有機首、尾翼兩部分幾乎沒有中過彈。如果你是指揮官，會建議在機身上哪些地方優先加厚防護？或是調整轟炸任務的執行方法？

創意思考

愛因斯坦：「陳述一個問題、尋找一個問題，往往比解答一個問題更重要，解答所需要的可能只是數學或實驗的技巧而已，提出新的問題、新的可能性或從新的角度思考舊問題，需要的是創意的想像，才是科學真正的內涵。」

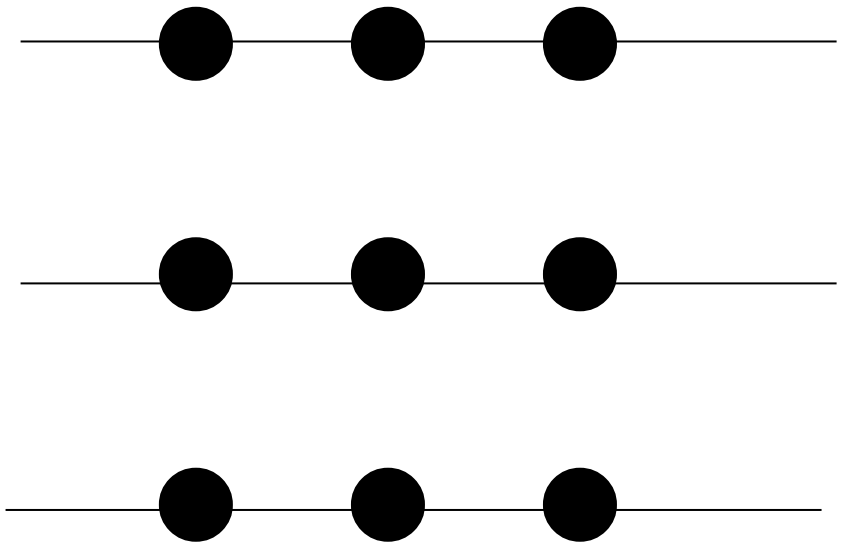
創意思考範例

- 請以三條直線連接以下互相平行的九個點：



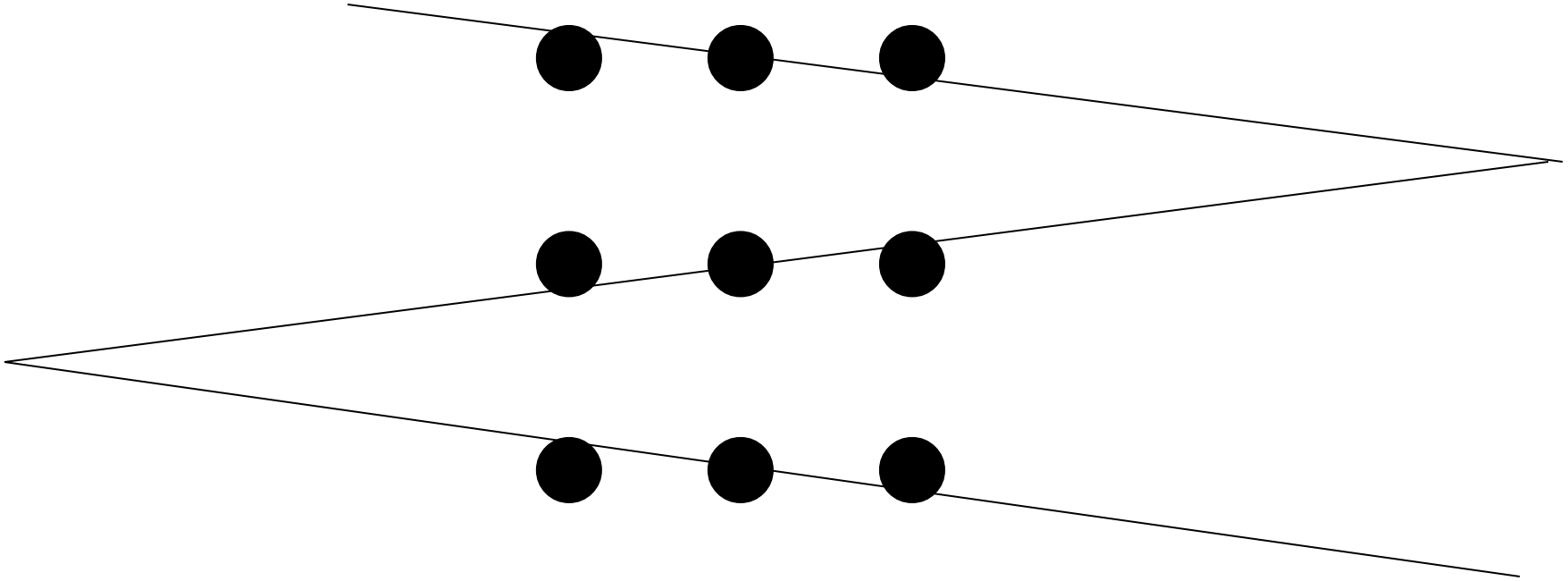
創意思考範例(續)

- 一般的想法可能是：



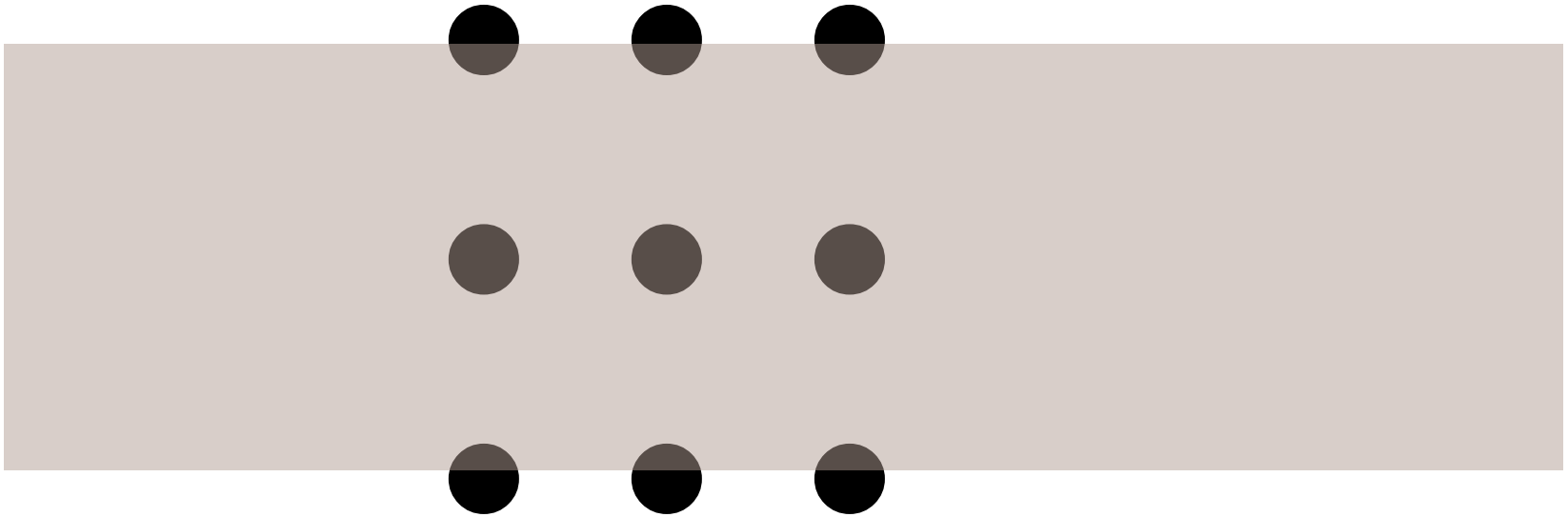
創意思考範例(續)

- 如果點有大小之分：



創意思考範例(續)

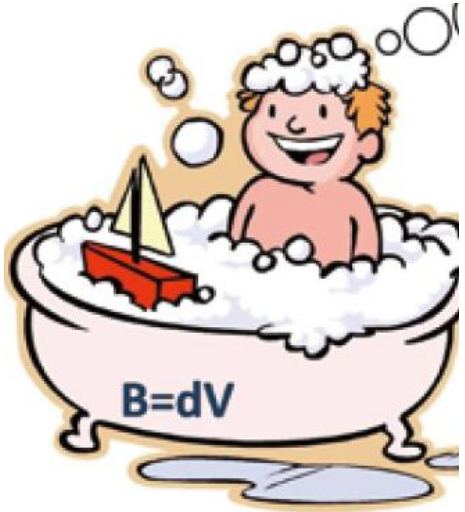
- 如果線也有粗細之分：



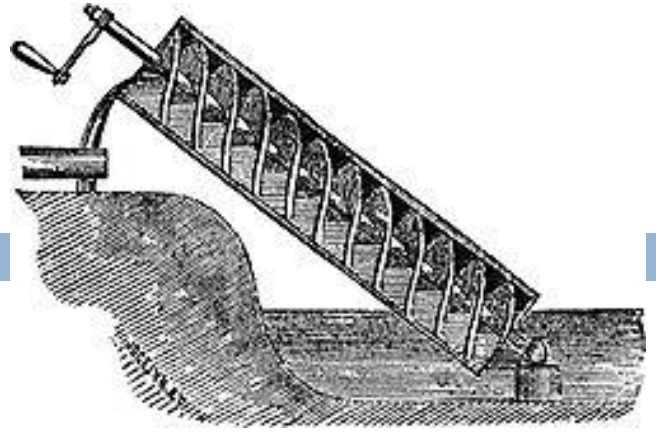
- 一有徵兆顯示可能造成異常現象，或出現有違常理的事件時，有效的決策者總是主動進行測試。他們總會一一寫下，在既有定義之下，他們預期哪些事情會發生（例如：預期交通意外的發生率降為零）？並定期測試是否真會出現自己所預期的情況。

真正的問題在哪裡？

- 有時呈現在表面的因素並非造成問題的實際原因，解決方案需從另一方向或結合專業知識去探索。
- 古今中外的突破與發展，許多藉助於變換思考方向(Paradigm Shift)，換個角度思考有意想不到的驚喜！
- 討論：阿基米德的浮體原理。

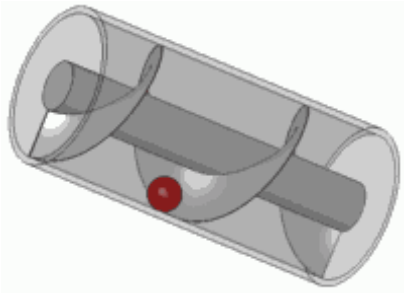


突破性思考的特色



- 長期探索(Long search)
- 沒有明顯的進展(Little apparent progress)
- 突發事件(Precipitating event)
- 靈光一閃(Cognitive snap)
- 轉換(Transformation)

→歷史上幾位著名人物的例子：阿基米德、孟德爾、達爾文、費曼。



過去的習慣不見得有道理

A little girl was watching her mother prepare a fish for dinner. Her mother cut the head and tail off the fish and then placed it into a baking pan. The little girl asked her mother why she cut the head and tail off the fish. Her mother thought for a while and then said, "I've always done it that way – that's how babicka (Czech for grandma) did it."

Not satisfied with the answer, the little girl went to visit her grandma to find out why she cut the head and tail off the fish before baking it. Grandma thought for a while and replied, "I don't know. My mother always did it that way." So the little girl and the grandma went to visit great grandma to find ask if she knew the answer. Great grandma thought for a while and said,

"**Because my baking pan was too small to fit in the whole fish.**"

The sweet old couple (dangers of making assumptions, understand before you intervene)

A little old couple walked into a fast food restaurant. The little old man walked up to the counter, ordered the food, paid, and took the tray back to the table where the little old lady sat. On the tray was a hamburger, a small bag of fries and a drink. Carefully the old man cut the hamburger in two, and divided the fries into two neat piles. He sipped the drink and passed it to the little old lady, who took a sip and passed it back. A young man on a nearby table had watched the old couple and felt sorry for them. He offered to buy them another meal, but the old man politely declined, saying that they were used to sharing everything. The old man began to eat his food, but his wife sat still, not eating. The young continued to watch the couple. He still felt he should be offering to help. As the little old man finished eating, the old lady had still not started on her food.

"Ma'am, why aren't you eating?" asked the young man sympathetically. The old lady looked up and said politely,

"I'm waiting for the teeth..."

http://1.bp.blogspot.com/-HWsJkhuXhp0/UD774ABVi8I/AAAAAAAAEdY/W3n_wM90W0A/s1600/The+sweet+old+couple.jpg

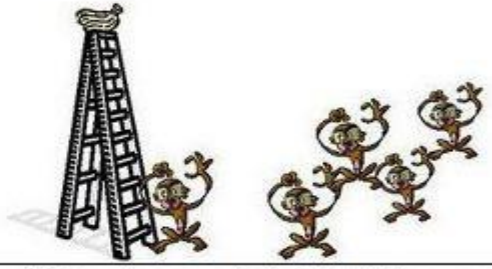
從眾行為(Bandwagon Effect)

http://i.snag.gy/kdu77.jpg

A group of scientists placed 5 monkeys in a cage and in the middle, a ladder with bananas on the top.

Every time a monkey went up the ladder, the scientists soaked the rest of the monkeys with cold water.

After a while, every time a monkey went up the ladder, the others beat up the one on the ladder.

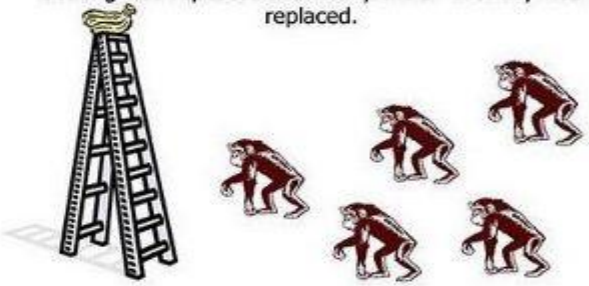


After some time, no monkey dare to go up the ladder regardless of the temptation.

Scientists then decided to substitute one of the monkeys. The 1st thing this new monkey did was to go up the ladder. Immediately the other monkeys beat him up.

After several beatings, the new member learned not to climb the ladder even though never knew why.

A 2nd monkey was substituted and the same occurred. The 1st monkey participated on the beating for the 2nd monkey. A 3rd monkey was changed and the same was repeated (beating). The 4th was substituted and the beating was repeated and finally the 5th monkey was replaced.



What was left was a group of 5 monkeys that even though never received a cold shower, continued to beat up any monkey who attempted to climb the ladder.

If it was possible to ask the monkeys why they would beat up all those who attempted to go up the ladder.....
I bet you the answer would be....

"I don't know – that's how things are done around here"

Does it sounds familiar?

Don't miss the opportunity to share this with others as they might be asking themselves why we continue to do what we are doing if there is a different way out there.

The sweet old couple (dangers of making assumptions, understand before you intervene)

A little old couple walked into a fast food restaurant. The little old man walked up to the counter, ordered the food, paid, and took the tray back to the table where the little old lady sat. On the tray was a hamburger, a small bag of fries and a drink. Carefully the old man cut the hamburger in two, and divided the fries into two neat piles. He sipped the drink and passed it to the little old lady, who took a sip and passed it back. A young man on a nearby table had watched the old couple and felt sorry for them. He offered to buy them another meal, but the old man politely declined, saying that they were used to sharing everything. The old man began to eat his food, but his wife sat still, not eating. The young continued to watch the couple. He still felt he should be offering to help. As the little old man finished eating, the old lady had still not started on her food. "Ma'am, why aren't you eating?" asked the young man sympathetically.

The old lady looked up and said politely,

"I’m waiting for the teeth..“

創意思考練習案例

《科技使我們變笨？》

- 近年搜尋引擎盛行，許多人依賴Google等獲取資訊，這會引起那些後遺症？科技（大數據）正在改變我們的大腦嗎？
- 例如：如果不依賴網路資源(Google Map)，你會如何獲取地圖資訊？

  註：手指靈敏度vs.大腦記憶（常識？）