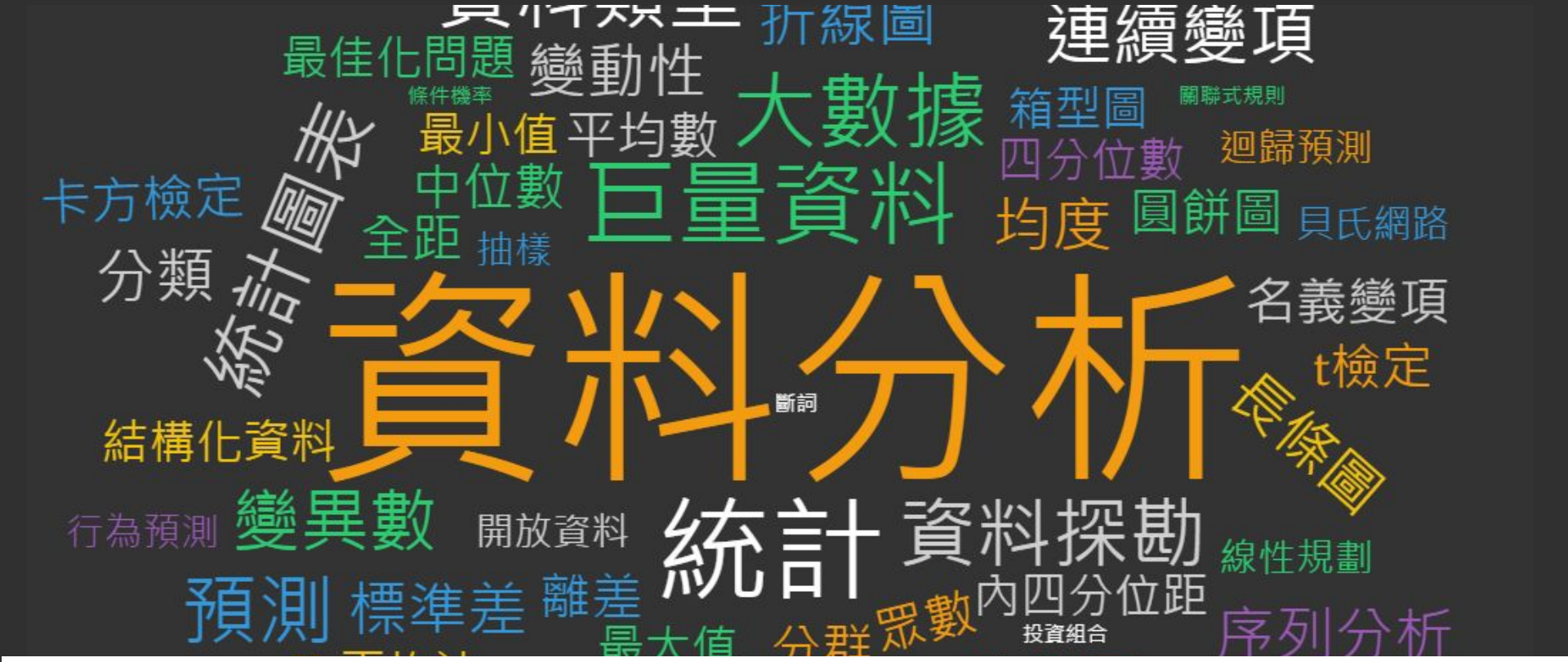


巨量資料探勘與統計應用 W01

課程介紹

布丁布丁吃布丁

http://blog.pulipuli.info/

教師簡介

學歷

- 輔大圖資系
- 政大圖檔所
- 政大圖檔所博士候選人

專長

- 程式開發
  (網頁網站應用)
- 伺服器管理
- 資料分析

授課經驗

- 國立空中大學
  - 影像處理
  - 文書處理
  - 資料結構
  - 作業系統

布丁布丁吃什麼？
http://blog.pulipuli.info/

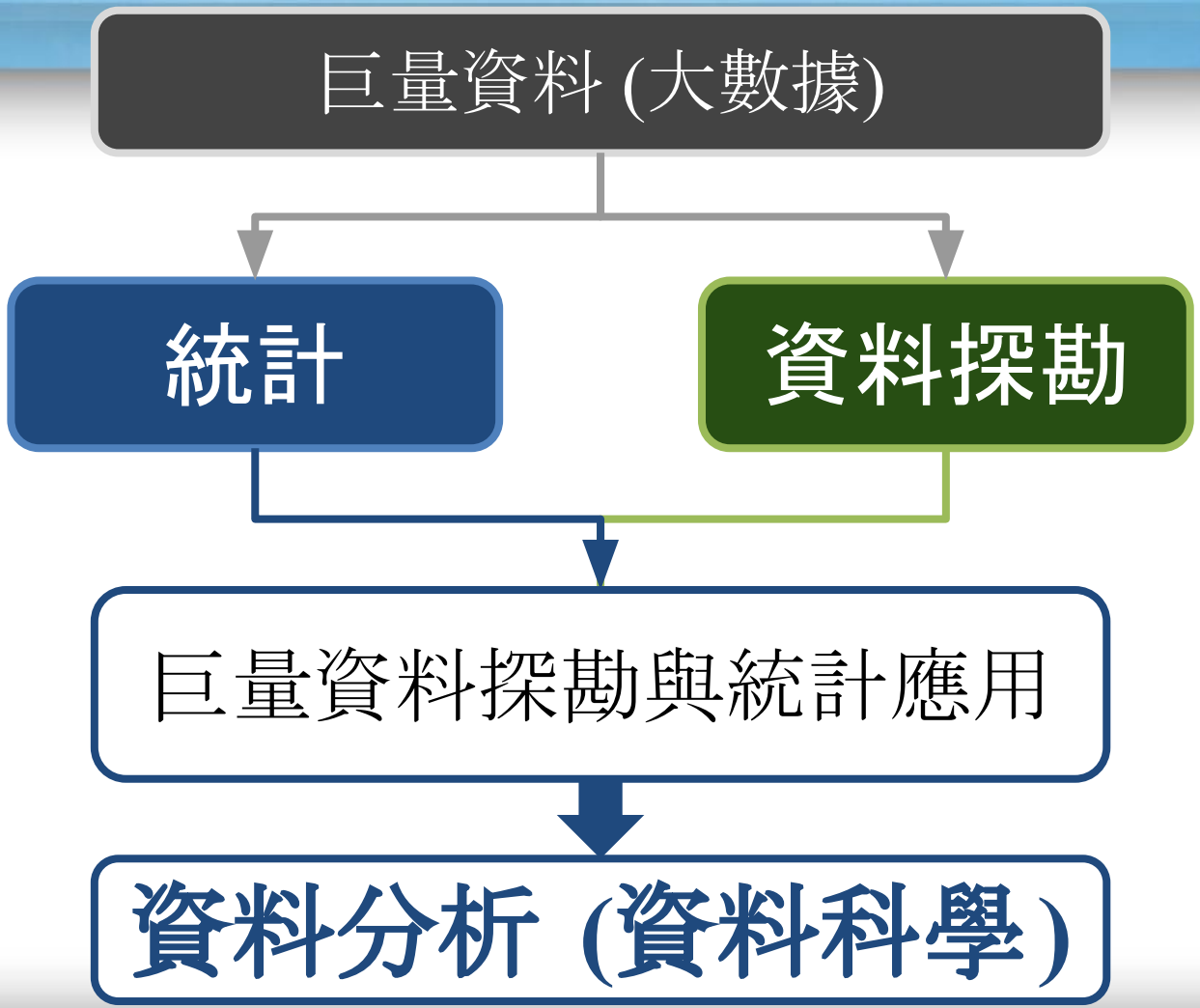
本週課程大綱

1. **為什麼要修這門課？**
2. 本學期課程概論
3. 課程部分單元簡介
4. 修課規定:作業與考試

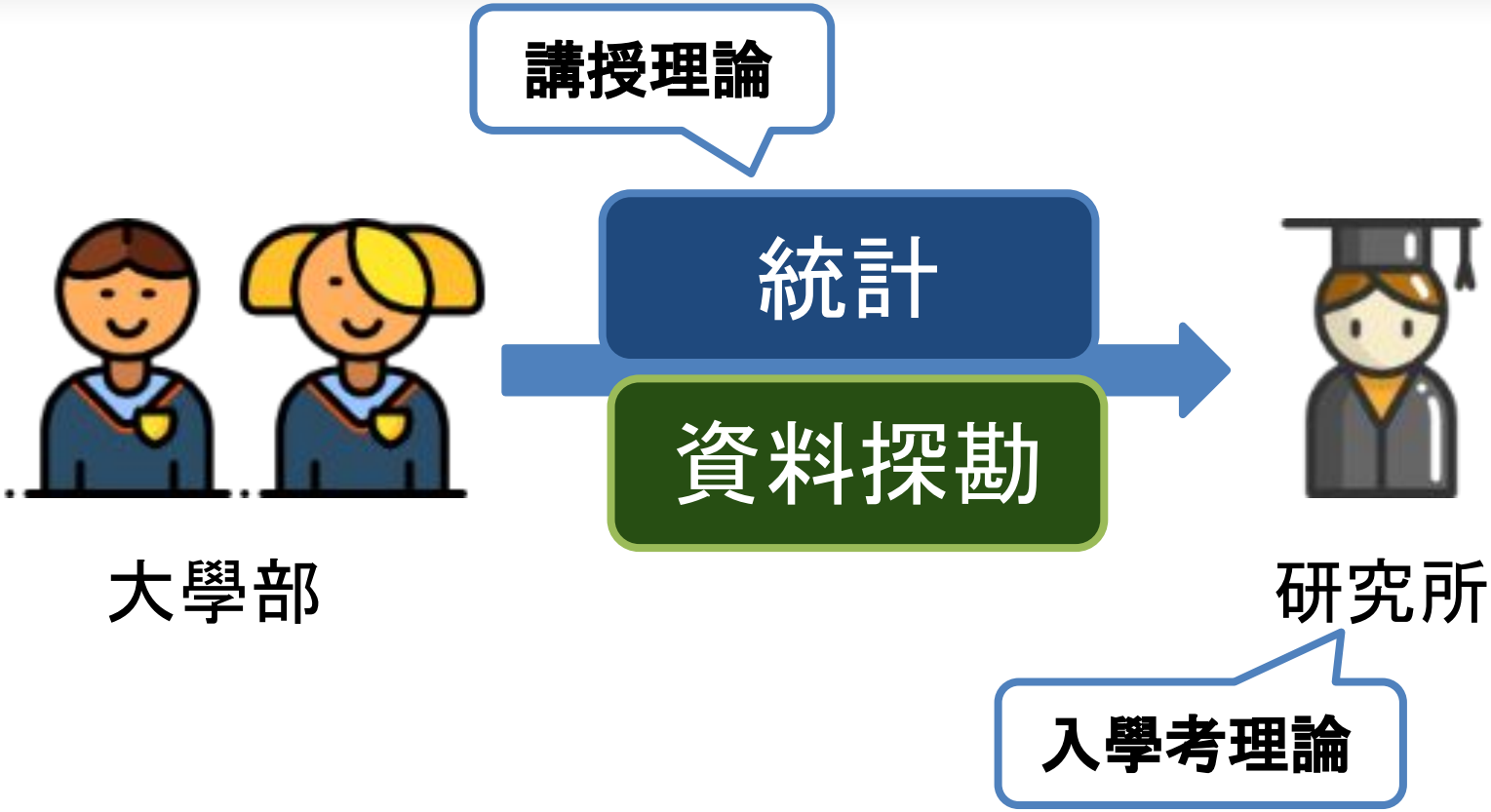
Part 1

為什麼要修這門課？

本課程的涵蓋領域

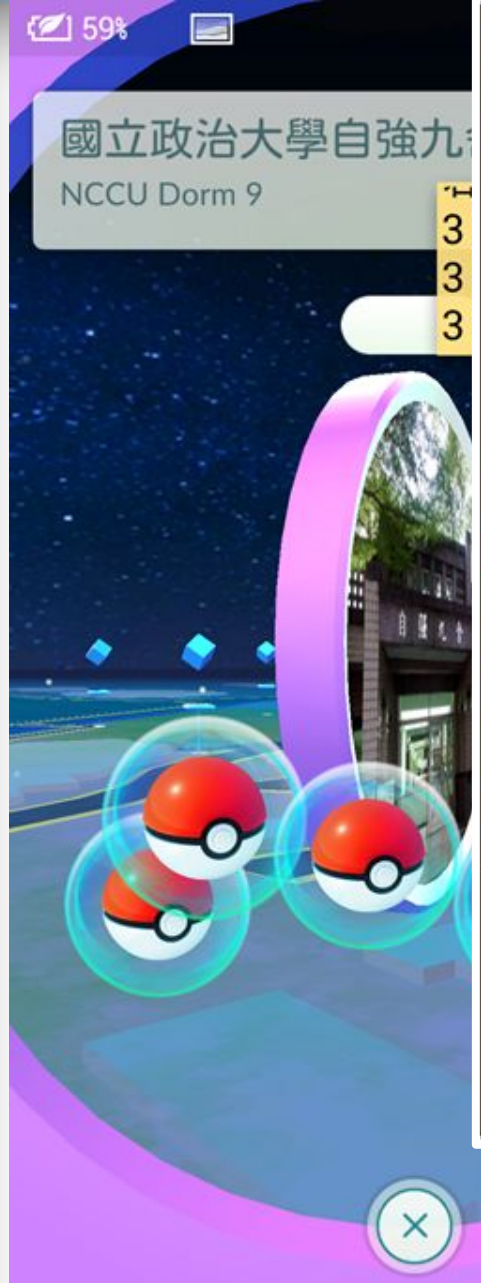


從大學的授課角度來 說...



即戦力


59%
19:42
國立政治大學自強九舍
NCCU Dorm 9
3 3 1 2 1
3 2 3 4 3
3 3 3 1 1 2



PokeStop 左轉右轉道具

www.e-zone.com.hk/channelnews.php?id

PokeStop 左轉右轉道具有分別？

【傳聞】不想要 Potion 請左轉

Pokemon Go 訓練員其中一大煩惱，就是背包空間不夠，想要儲存更多精靈球、寶貝球？唯有拋棄超佔位置的 3 色 potion。可惜，永遠到 PokeStop 轉道具的時候，都會掉落一堆用不完的 potion，要經常清掉都是一件挺煩惱的事。不過，近日一位日本網民的發現，可能會令大家免卻這個煩惱！

【相關報道】第二代精靈寶貝進化指南
【精選消息】Pokemon GO 第二世代精靈寶貝圖鑑
【精選消息】【附表】Pokemon GO 精靈寶貝巢穴大更新
【精選消息】Pokemon Go 道館對決有必勝法
【精選消息】【附列表】Pokemon Go 看精靈寶貝評語即知精靈寶貝 IV 值？

日網友發現，原來在 Pokestop 左轉或是右轉地標，得到的道具機率是不同的。據網友測試，多數人都是右轉地標（手指從屏幕左邊，滑到屏幕右邊），得到 Potion、Razz Berry 的機率會更大；而左轉地標（手指從屏幕右邊，滑到屏幕左邊），得到精靈球的機率就會更

變數名稱 Variable	t檢定統計量 t-statistics	自由度 d.f.	臨界值 t(d.f.,1-α)	p-值III p-value	平均數的差異 Difference between sample and null means	的 95% 信賴區間 95% C.I. for difference	
						下界 Lower	上界 Upper
NEW	-0.3787	62	1.6698	0.64691	-0.0968	-0.5235	Inf

I：分組變數為_GROUP_

II：根據雙樣本變異數檢定結果，假設兩母體具有相同變異數進行雙樣本平均數差異t檢定

III：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- **分析結果建議**：由於檢定結果P-值(0.64691) > 顯著水準0.05，因此無法拒絕虛無假設。


國立政治大學自強九舍
NCCU Dorm 9
33121
32343
333112
道具有分別？
不想要 Potion 請左轉
假的！
哎呀！我的眼睛業障重啊！


8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
1700
1450
1750
1800
12%
10%
HRT
FEW
BGY
RDW
FRT division
Share of market activity
Distribution of the securities market key players

註:本課程沒有變成小學生的風險


註:本課程不會教您
如何變得像阿湯哥這
麼帥

這堂課就是為了您！

Meta Science

學科之神

Marcia J. Bates

UCLA 美國加州大學洛杉磯分校
資訊科學系榮譽教授

這世界上，有三門學科，
專門掌握著全人類的知識。
它們被稱之為...

Meta Science

Bates, M. J. (1999). The Invisible Substrate of Information Science. Journal of the American Society for Information Science, 50(12), 1043-1050.

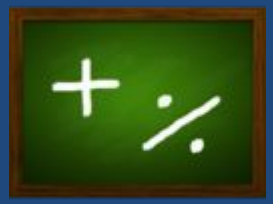
Meta Science
學科之神

Information Science
圖書資訊學

資料的儲存與檢索

Journalism
新聞學

新聞的發掘與傳播

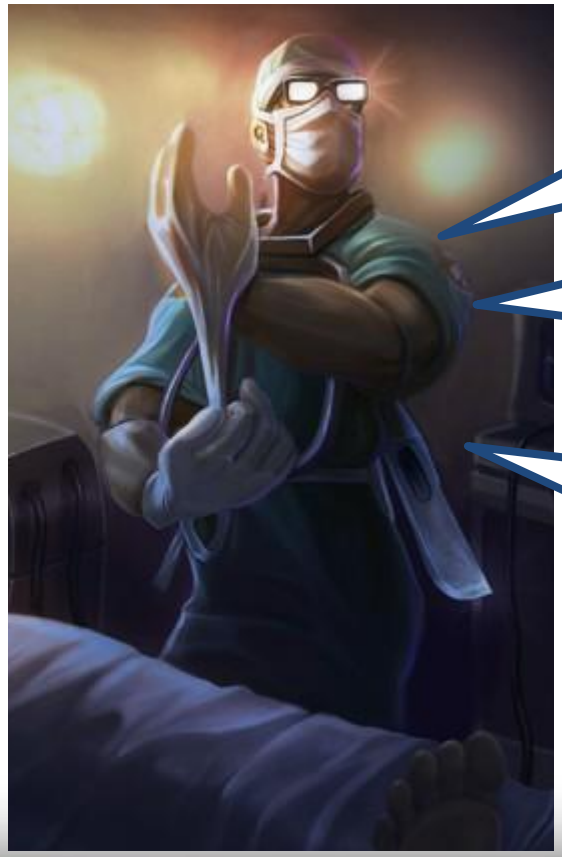


Education
教育學

知識的教導與學習

為什麼圖書資訊學是學科之神？

醫學



病歷資料

最新手術

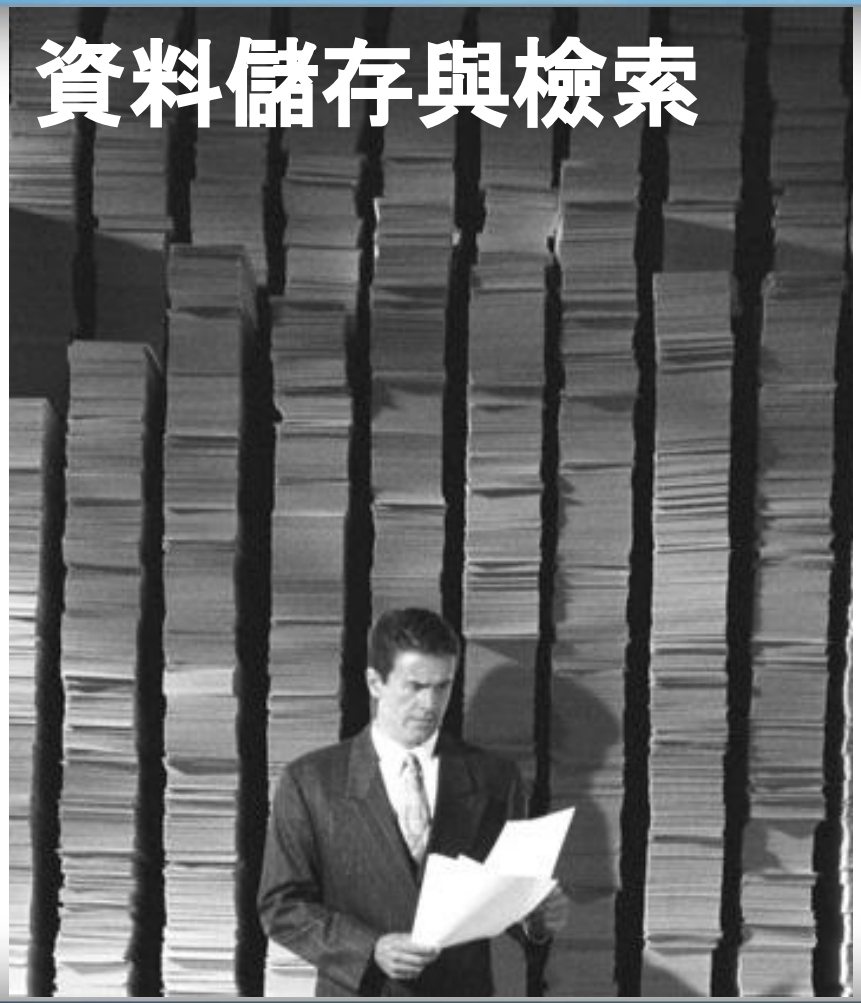
醫學知識

圖書資訊學

全部交給我管理吧

圖書資訊學的定位大概等於 ...

新時代的圖書資訊學



21世紀最性感的工作
資料科學家

- 統計學
- 資料探勘 (機器學習)
- 多變量微積分、
  線性代數
- 資料處理 (清理數據)
- 資料視覺化與溝通
- 軟體工程

https://www.inside.com.tw/2015/03/27/8-skills-you-need-to-be-a-data-scientist

2015年美國最佳職業排行榜

排名	職業	年薪(美元)	就業增長率(2012~2022)
1	精算師	$94,209	25.09%
2	聽覺矯正師	$71,133	33.33%
3	數學家	$102,182	25.91%
4	統計學家	$79,191	25.91%
5	生物醫學工程師	$89,165	26.65%
6	**資料科學家**	**$124,149**	**14.97%**
7	齒科醫師	$71,102	31.02%
8	軟體工程師	$93,113	21.13%
9	物理治療師	$77,114	29.14%
10	電腦系統分析師	$81,150	23.50%

http://www.epochtimes.com/b5/15/4/15/n4412000.htm

2016年美國最佳職業排行榜

排名	職業	年薪(美元)	就業增長率(2012~2022)
1	**資料科學家**	**$128,240**	**16.02%**
2	統計學家	$80,295	36.53%
3	資安系統分析師	$89,280	17.83%
4	聽覺矯正師	$73,231	29.11%
5	超音波醫療診斷師	$68,200	27.75%
6	數學家	$104,285	23.20%
7	軟體工程師	$93,233	17.78%
8	電腦系統分析師	$83,255	20.57%
9	語言治療師	$72,247	22.19%
10	精算師	$97,362	21.60%

http://www.careercast.com/jobs-rated/jobs-rated-report-2016-ranking-200-jobs

舊時代的神

圖書館員

新時代的神

資料科學家

本課程與其他課程的關係

資料處理

- 計算機概論
- 資料庫系統概論
- 研究方法*

原理細節

- 統計學
- 資料探勘*
- 資料檢索*

整合應用

- 系統分析
- 資訊系統概論
- 系統分析

* 研究所課程

本課程適合的對象

1. 比起被動聽枯燥乏味的理論，更喜歡動手實作的您
2. 想成為主管層級的您
3. 想成為資料科學家的您
4. 想成為新世界圖書館員的您
5. 想進修研究所的您
6. ~~想用神奇的分析技術揭露事實真相，表現出高格調的您~~
7. ~~玩遊戲想要有效率提升勝率的您~~

(系主任表示: 不可在課堂教學生裝B玩遊戲)

本課程不適合的對象

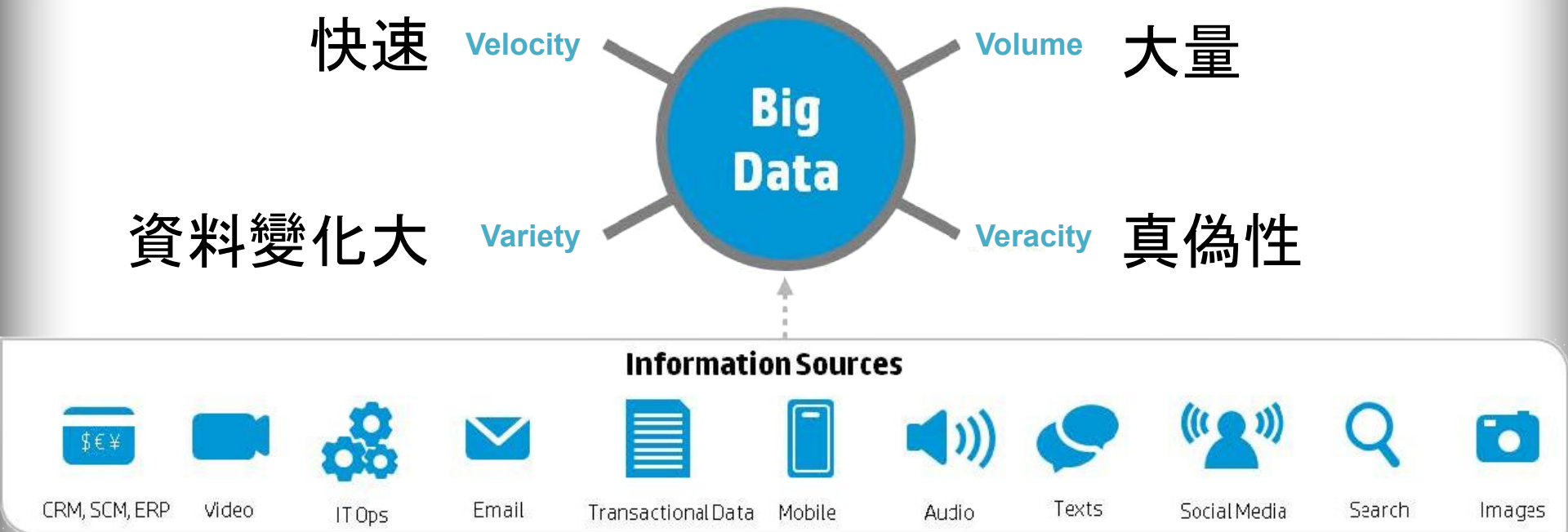
1. 喜歡傳統被動聽課，不喜歡動手操作，更討厭老師胡言亂語的人
2. 一定要用手算分析，排斥用電腦計算的人
3. 比起分析得到的結果，更在意分析技術細節(演算法)的人


什麼時候才要講
這門課要教什麼...？
我在考慮是否要退選呢...

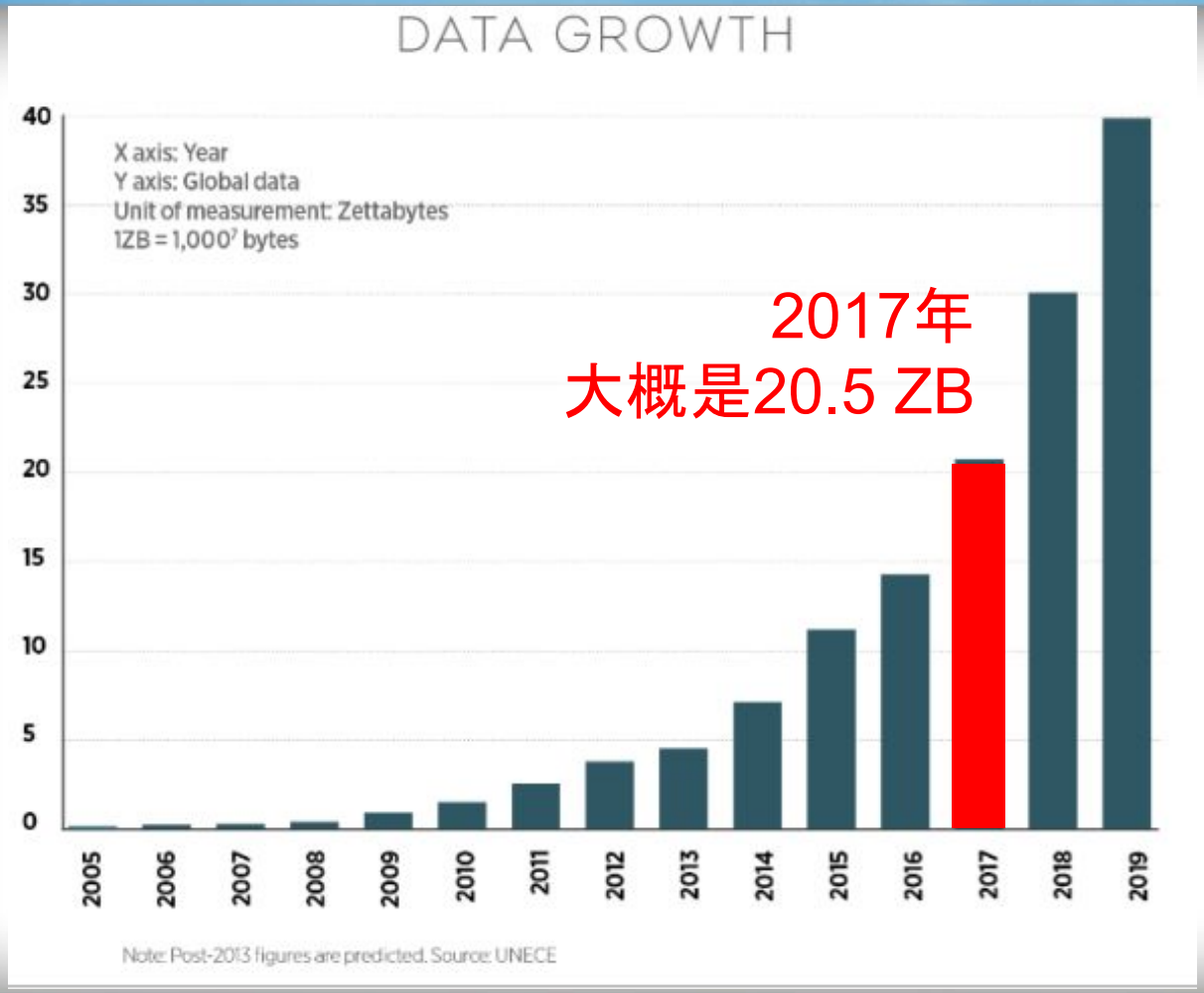
Part 2.

本學期課程介紹

巨量資料=大數據=
BIG DATA：大,快,雜,疑



資料的層級



ZB層級

EB層級

PB層級

TB層級

GB層級

MB層級

KB層級

http://www.nojitter.com/post/240170228/the-network-impact-of-big-data

資料分析的層級

- 昂貴的硬體設備
- 複雜的軟體架構

國家單位
大型公司

- 個人電腦即可
- 開源的技術

中小型公司
個人

ZB層級 

EB層級 

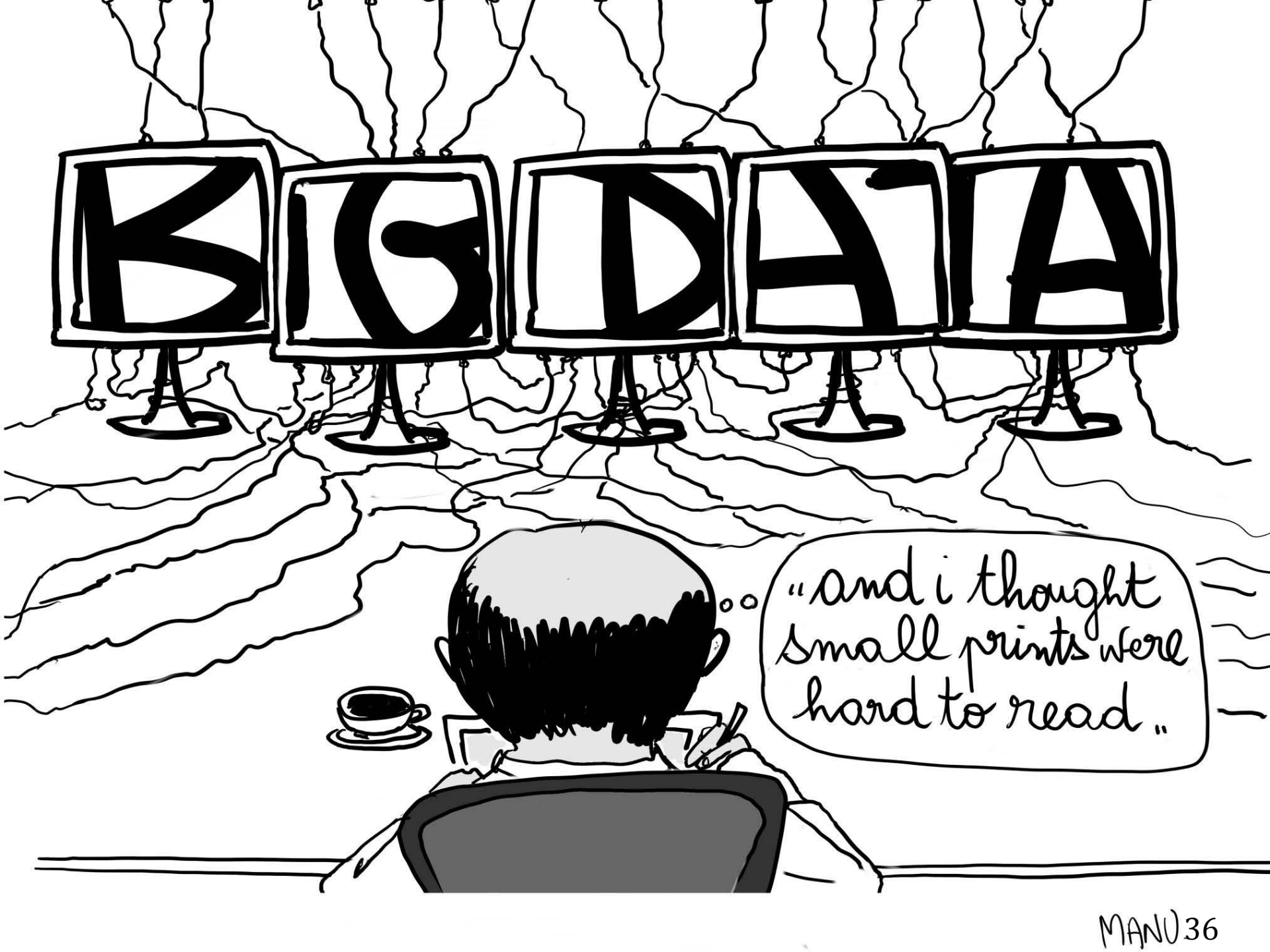
PB層級 

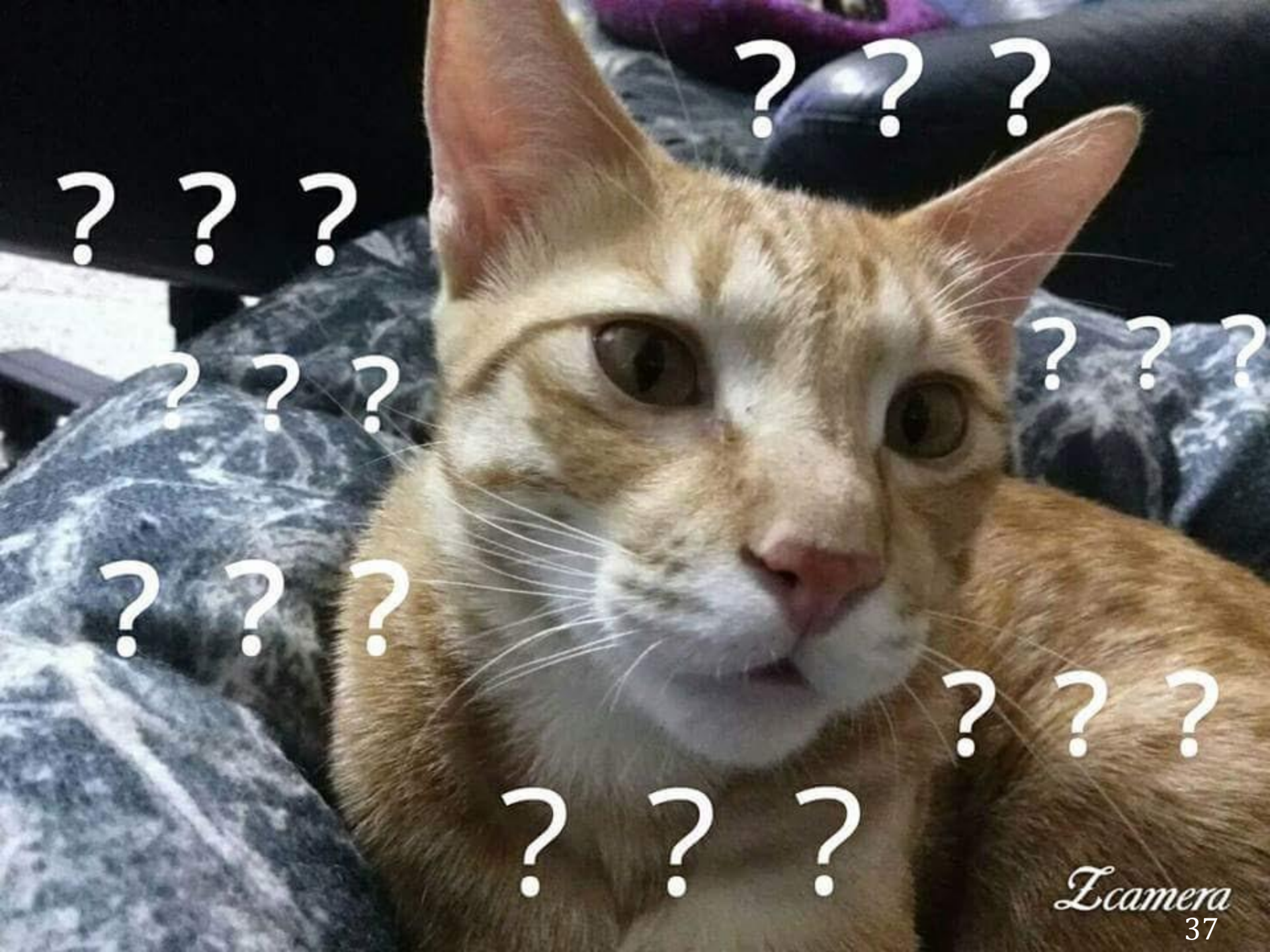
TB層級 

GB層級 

MB層級 

KB層級 


BIG DATA
"and i thought small prints were hard to read."
MANU36

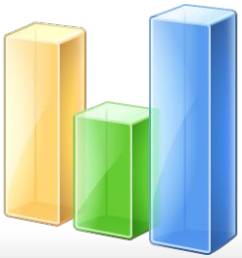

? ? ?
? ? ?
? ? ?
? ? ?
? ? ?
? ? ?
? ? ?
Icamera

資料分析的流程

問題層級

1. 資料敘述級

- 用簡單易懂的方式來**呈現**大量的資料
- 找出能夠代表大量資料的指標

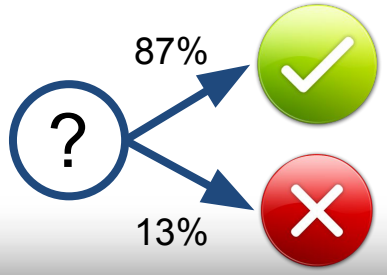


2. 資料檢定級

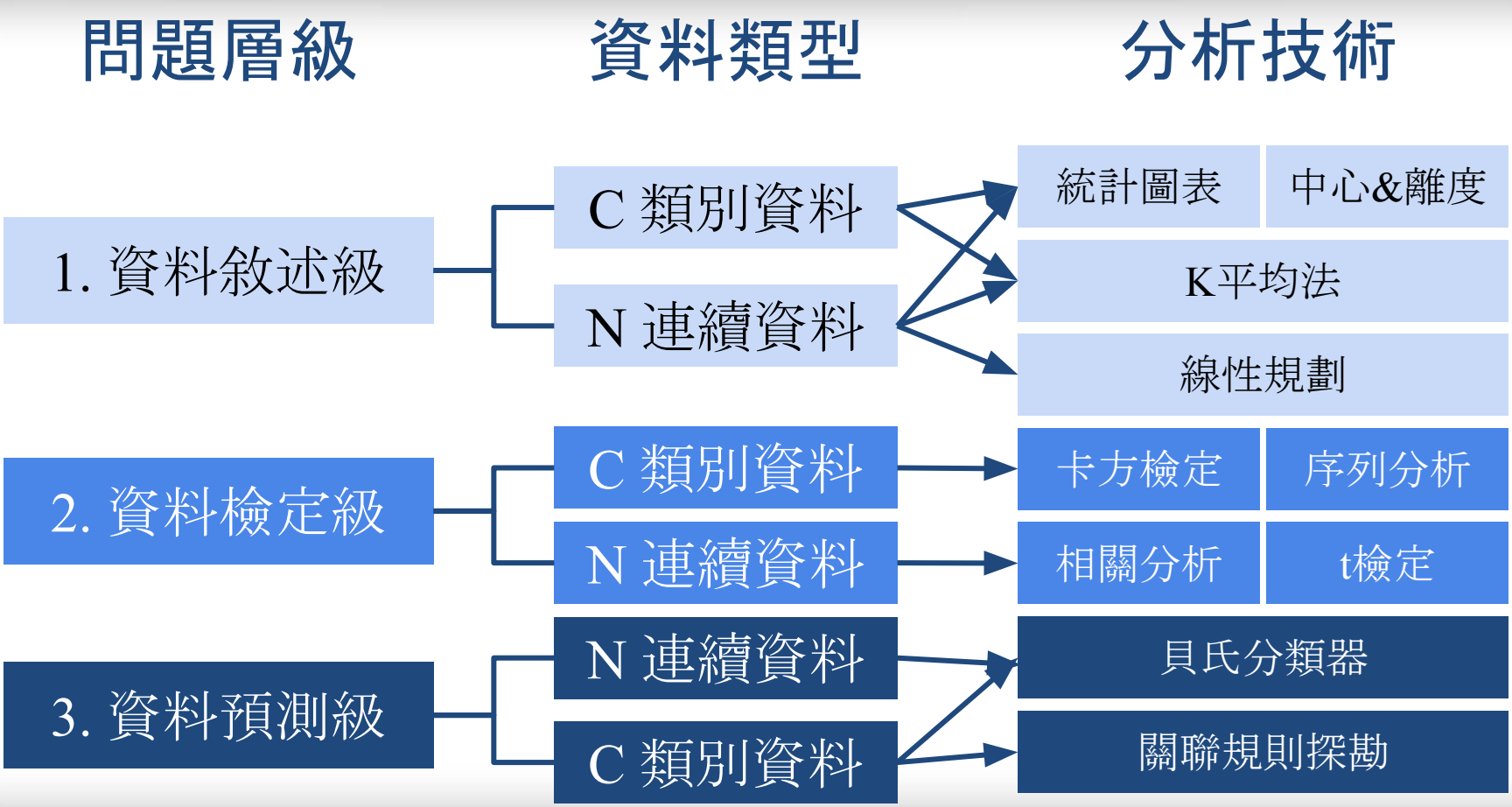
- 比較不同的資料
- 分析資料之間**是否**有有別一般的關係或有差異

3. 資料預測級

- 分析隱含在資料中的模式(模型)
- 利用模型來預測**未知的資料**，以獲得可能的**建議**



問題、資料與分析技術


問題層級
資料類型
分析技術
1. 資料敘述級
C 類別資料
N 連續資料
統計圖表
中心&離度
K平均法
線性規劃
2. 資料檢定級
C 類別資料
N 連續資料
卡方檢定
序列分析
相關分析
t檢定
3. 資料預測級
N 連續資料
C 類別資料
貝氏分類器
關聯規則探勘

單元課程設計

問題層級

敘述 / 檢定 / 預測

資料類型

N 連續 / C 類別

分析技術

？

本學期課程表:期中考範圍

Part 1.
課程導論
資料處理

1. 課程簡介
3. 資料來源與類型

Part 2.
資料敘述級

4. 資訊視覺化:統計圖表
5. 資料的中心與離度
6. 資料群集分析
8. 最佳化問題:線性規劃
9. 期中考

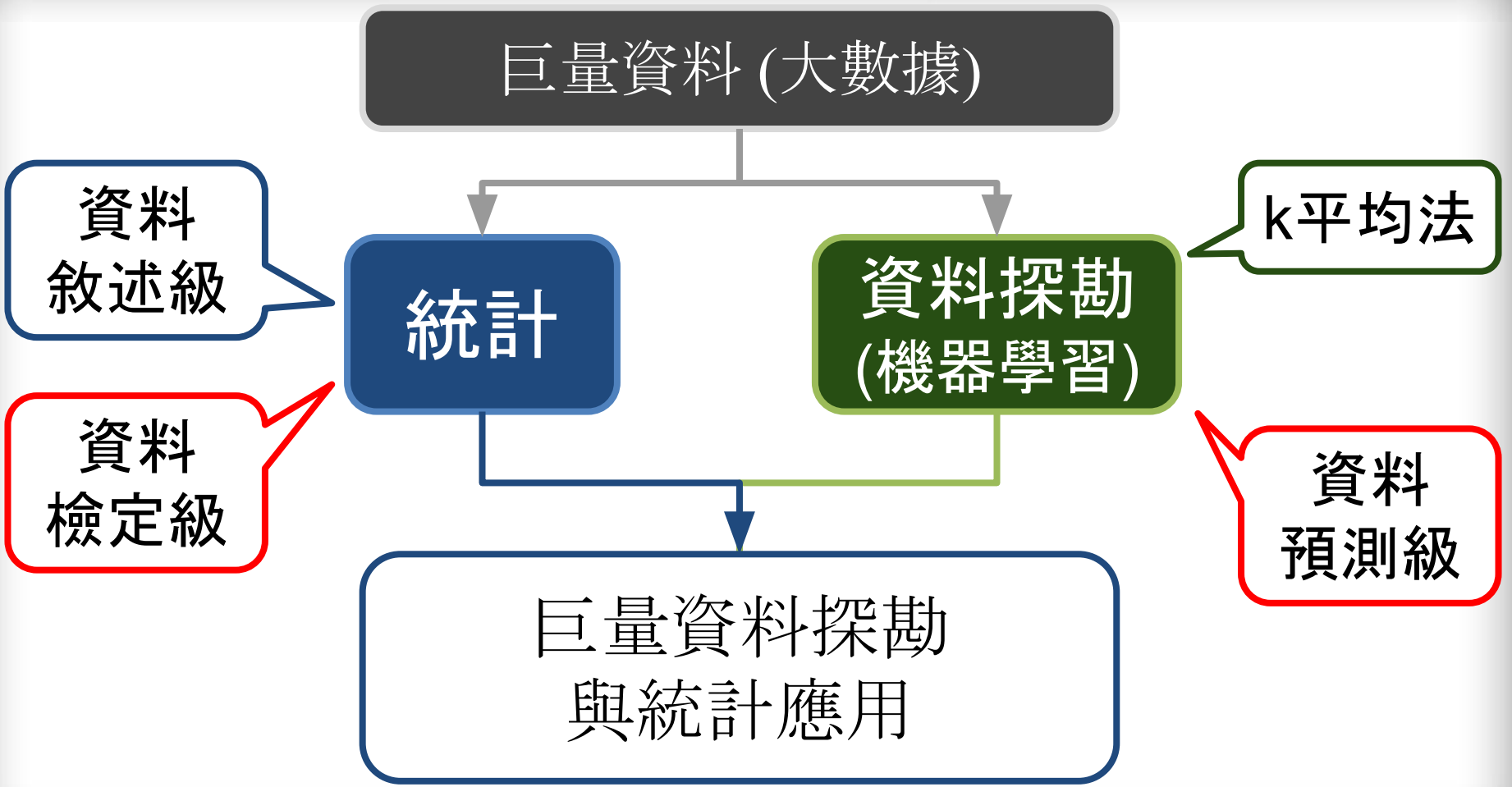
本學期課程表:期末考範圍

Part 3. 資料檢定級	10. 連續資料的相關檢定 11. 類別資料的相關檢定 12. 序列行為模式檢定 13. 連續資料的差異檢定
Part 4. 資料預測級	14. 資料的分類與預測 16. 非結構文字資料的分類 18. 期末考

Part 2.

本課程與大數據的關係

本課程認為的大數據



統計跟大數據有什麼關係？

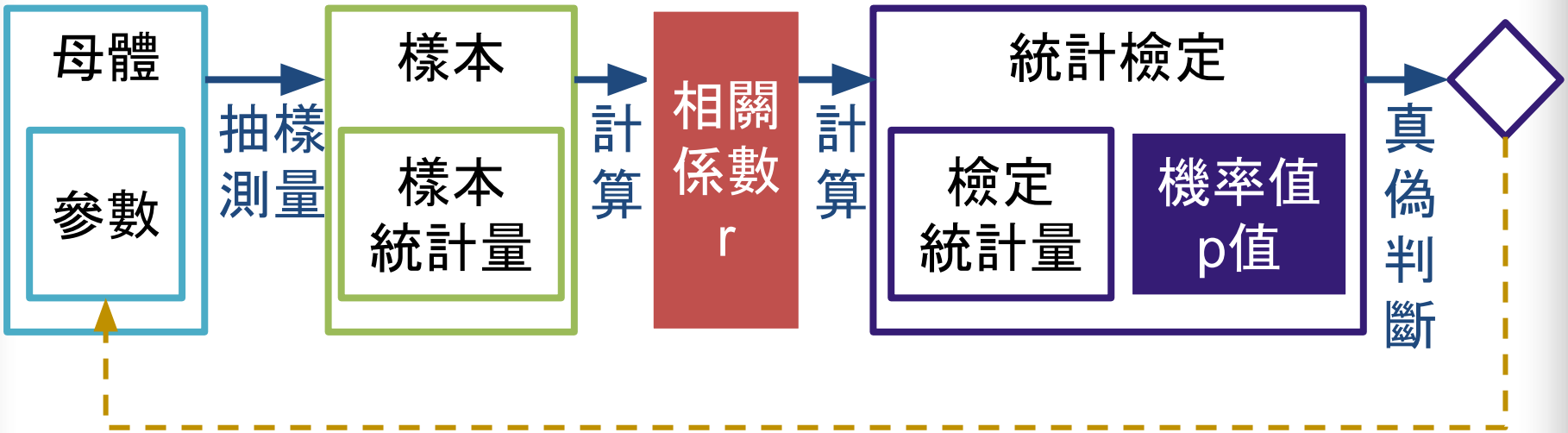
資料敘述級

- 統計圖表
- 中心與離度
- 群集分析 (資料探勘)
- 規劃求解 (最佳化問題)

資料檢定級

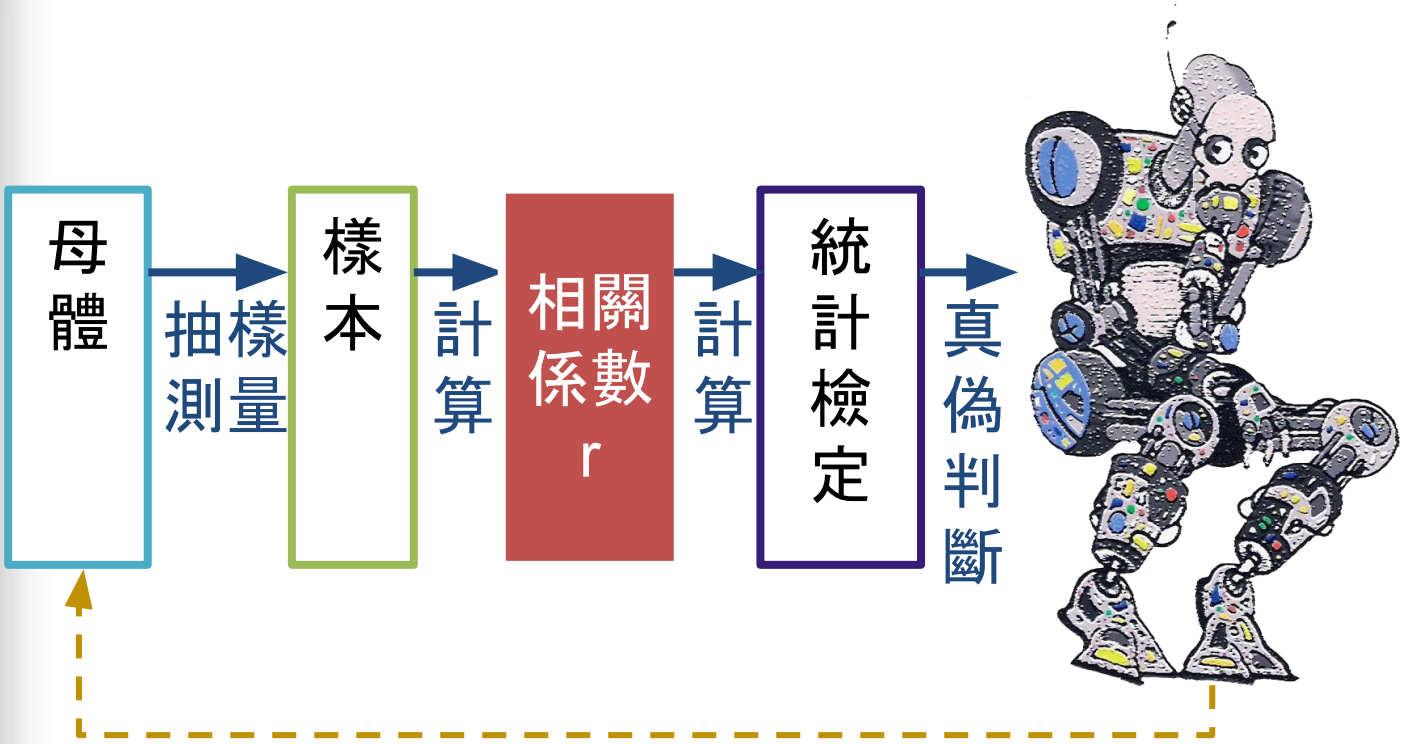
- 獨立樣本t檢定
- 積差相關分析
- 卡方獨立性檢定
- 滯後序列分析

相關分析的流程



人工智慧

相關分析搭配應用的流程



簡易型人工智慧專家系統

簡易型人工智慧專家系統

科學競賽活動參賽者資料 - data.csv

相關分析結果顯示，

數理能力與比賽總成績二者具有顯著的高度正相關，表示數理能力越高者，比賽總成績也會越高。

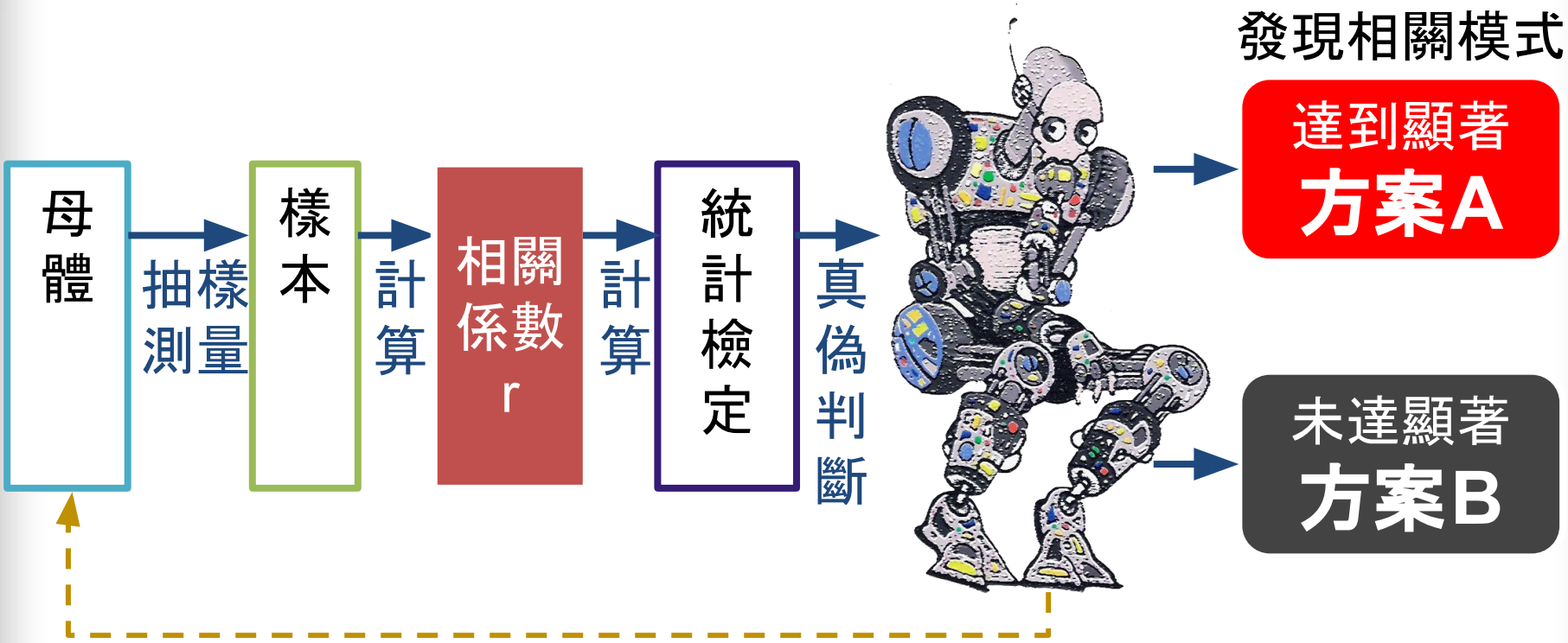
此外，

數理能力與美術能力、美術能力與比賽總成績之間的線性關係皆不明顯。

相關分析到此結束。

人工智慧

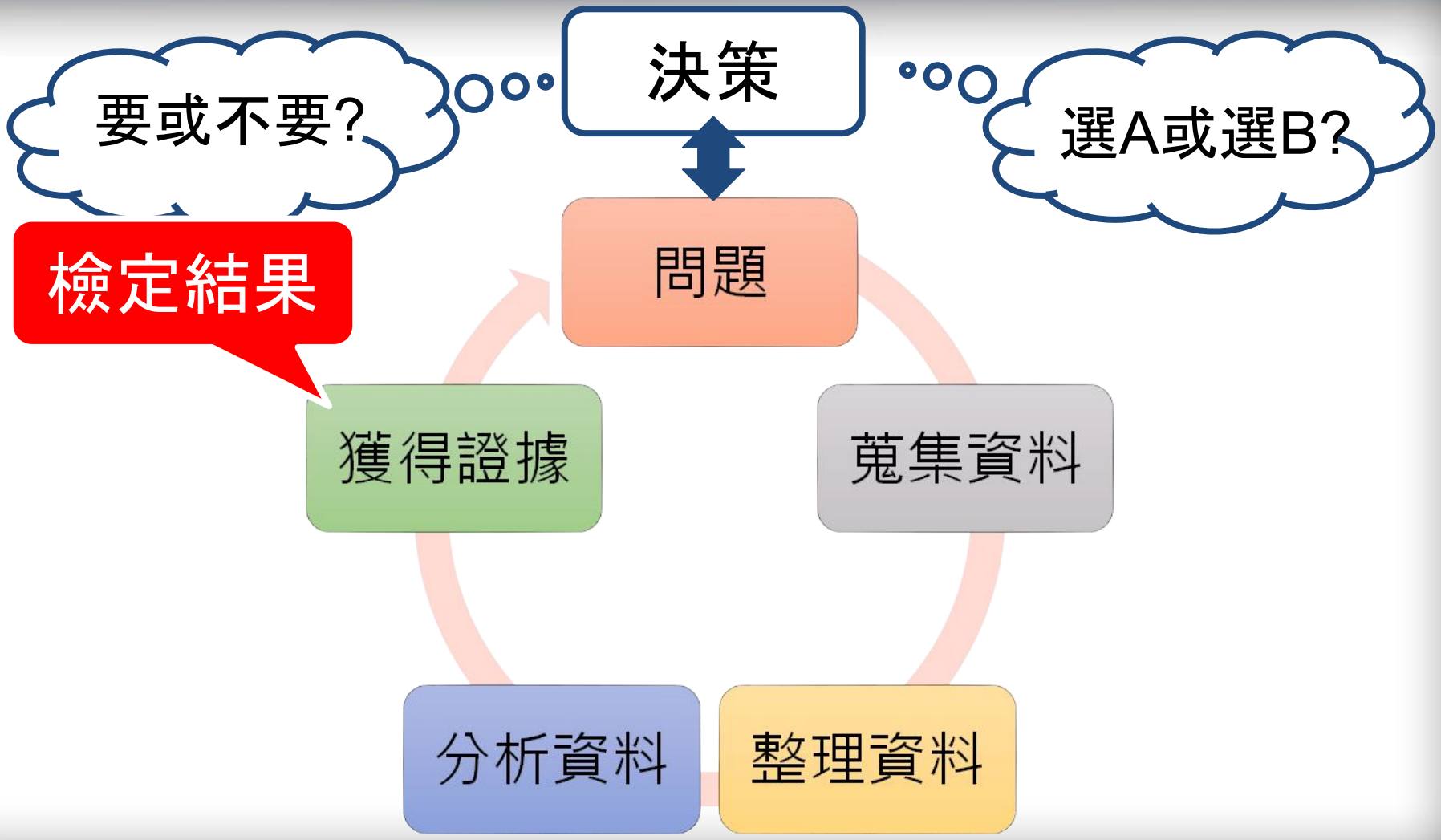
相關分析搭配應用的流程



人類

相關分析搭配應用的流程

資料分析的流程



Part 3.

課程部分單元介紹

資料處理:資料的來源與類型

零時政府

http://g0v.tw/
irc:freenode#g0v.tw

去中心化
前瞻性
雲端
群眾參與
政府體制研究暨開放與
海量資料應用
改善國家計畫

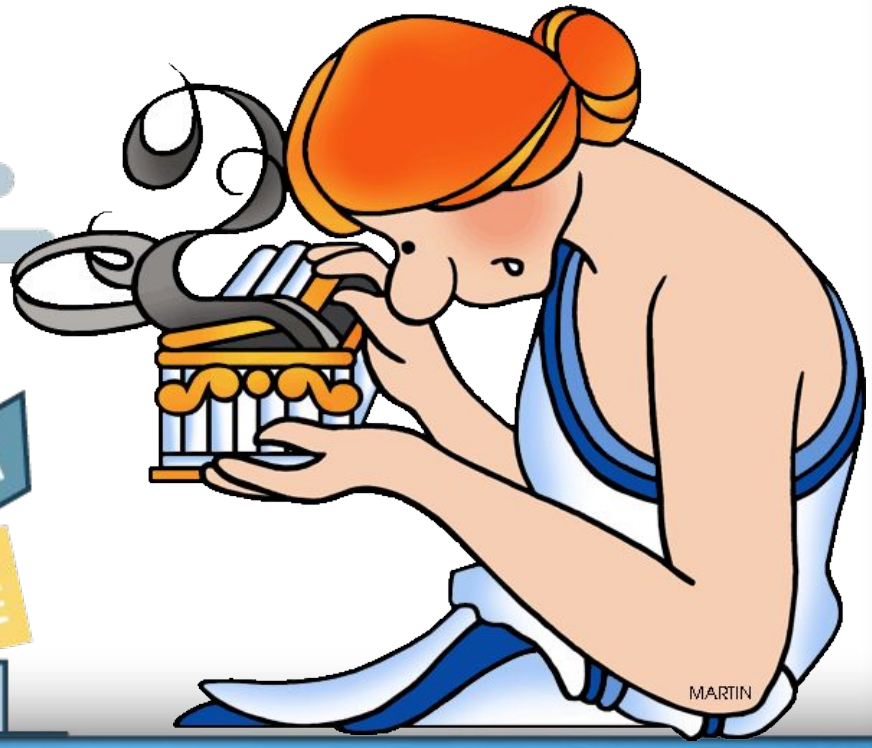
群眾參與
政府體制研究暨開放與
海量資料應用
改善國家計畫

零時政府

http://g0v.tw/
irc:freenode#g0v.tw

去中心化
前瞻性
雲端
群眾參與
政府體制研究暨開放與
海量資料應用
改善國家計畫

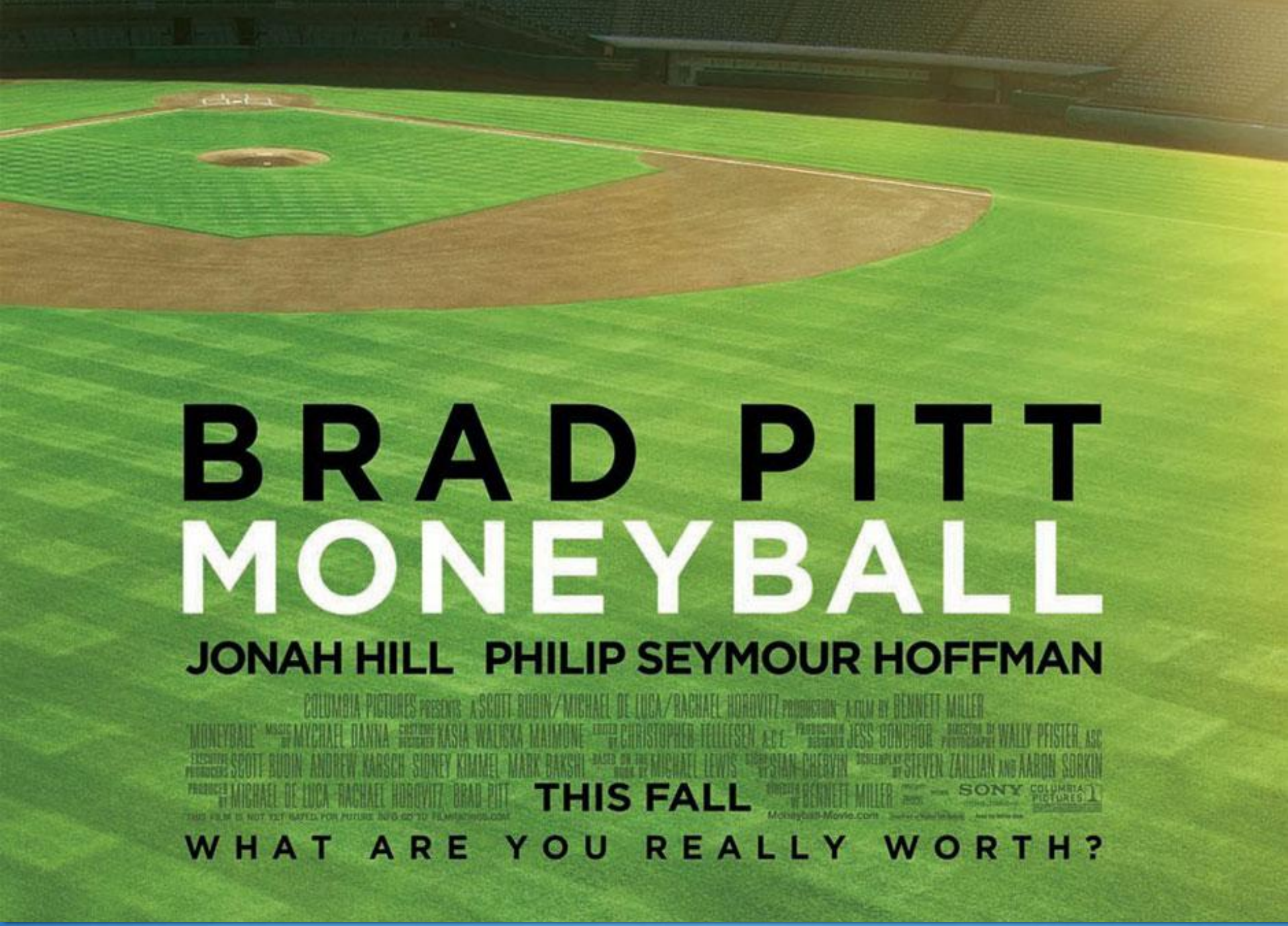
裡面有什麼資料？

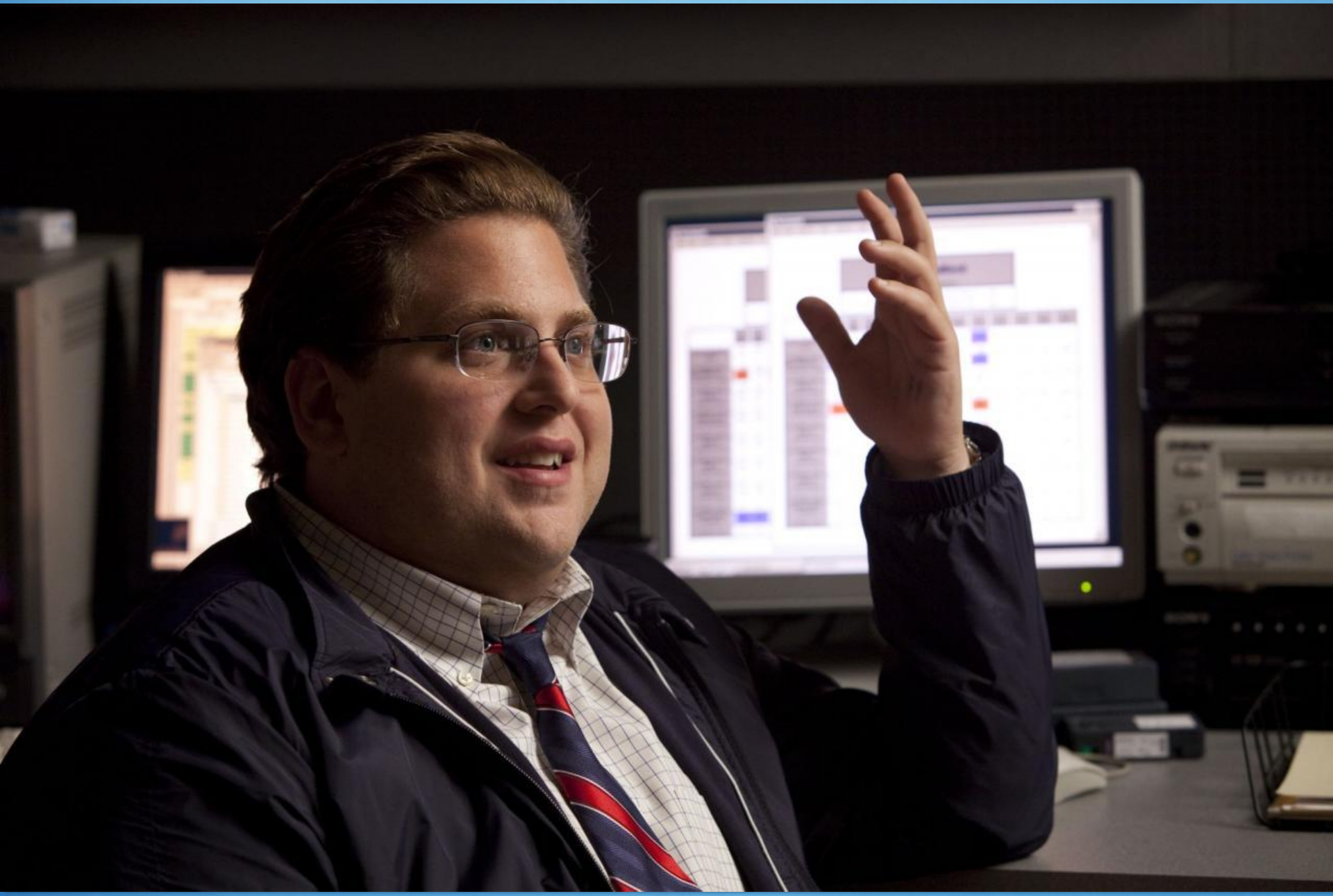


3/7

資料來源與類型

資料敘述級:最佳化問題


BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
COLUMBIA PICTURES PRESENTS A SCOTT RUDIN/MICHAEL DE LUCA/RACHAEL HOROVITZ PRODUCTION A FILM BY BENNETT MILLER
"MONEYBALL" MUSIC BY MYCHAEL DANNA PRODUCTION DESIGNER KASIA WALICKA MAIMONE EDITED BY CHRISTOPHER TELLEFSEN A.C.E. PRODUCTION DESIGNER JESS GONCHOR DIRECTOR OF PHOTOGRAPHY WALLY PFISTER ASC
EXECUTIVE PRODUCERS SCOTT RUDIN ANDREW KARSCH SIDNEY KIMMEL MARK BAKSHI BASED ON THE BOOK BY MICHAEL LEWIS STORY BY STAN CHERVIN SCREENPLAY BY STEVEN ZAILLIAN AND AARON SORKIN
PRODUCED BY MICHAEL DE LUCA RACHAEL HOROVITZ BRAD PITT
THIS FALL
DIRECTED BY BENNETT MILLER
SONY COLUMBIA PICTURES
THIS FILM IS NOT YET RATED. FOR FUTURE INFO GO TO FILMRATINGS.COM
MoneyballMovie.com
WHAT ARE YOU REALLY WORTH?




這是你要的球員分析報告

最少成本

最大~~戰力~~效益

資料敘述級

N 連續資料

線性規劃

4/11 給我答案

最佳化問題

最少成本

最大~~戰力~~效益

資料檢定級:類別資料的相關檢定

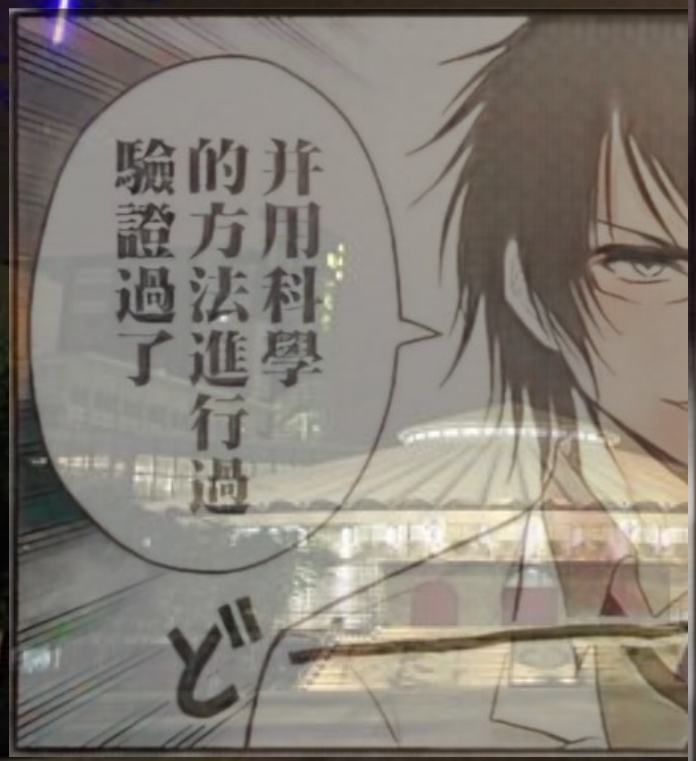

傳說...
在輔大聖誕樹下告白
就一定會成功...

傳說...

在輔大聖誕樹下告白

就一定會成功...

註：這雖然是漫畫，但可不是鬧著玩的



資料檢定級

C 類別資料

卡方檢定

5/2 不是偶然

類別資料的相關檢定

資料預測級:資料的分類與預測

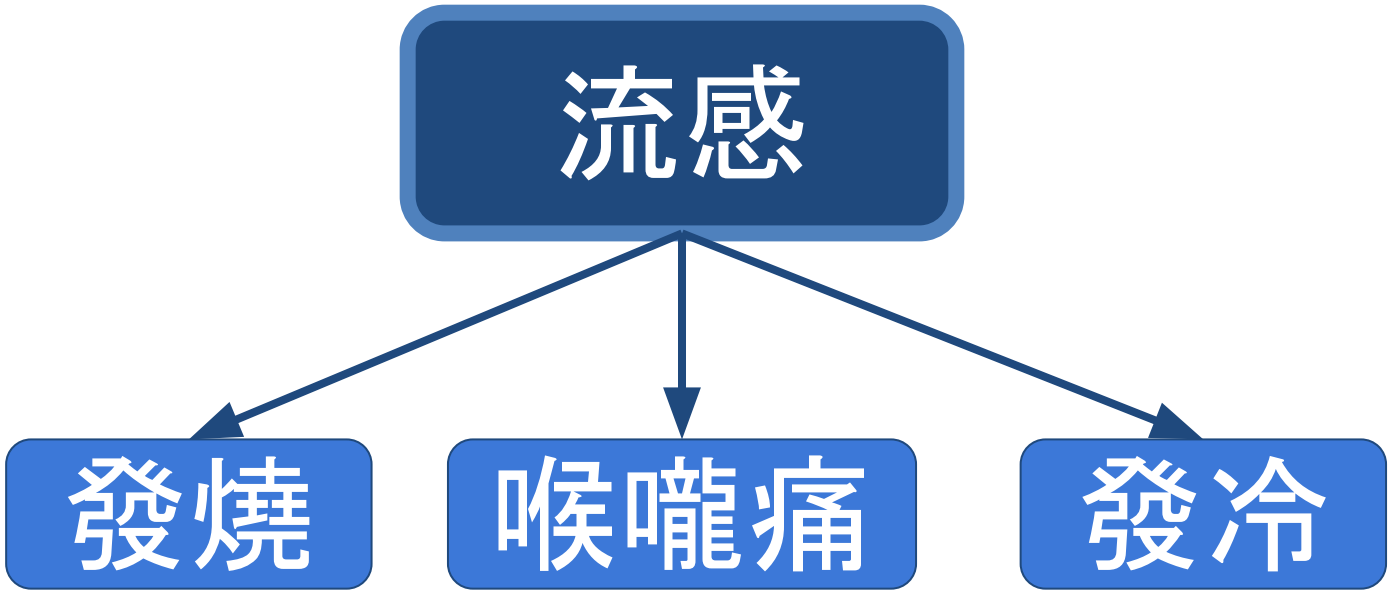

被害人和凶手姓名刻在木球上

以統計分析與預測技術降低犯罪率：靠著有效佈署警力，達到預防犯罪

http://blogger.gtwang.org/2013/07/memphis-police-department-reduces-crime-rates-with-ibm-predictive-analytics-software.html

美國政府要求入境外國人填寫社群媒體帳號

http://technews.tw/2016/12/23/us-ask-foreigner-to-write-down-their-personal-social-media-accounts/


流感
發燒
喉嚨痛
發冷

機率上升

流感

- 發燒
- 喉嚨痛
- 發冷

機率上升

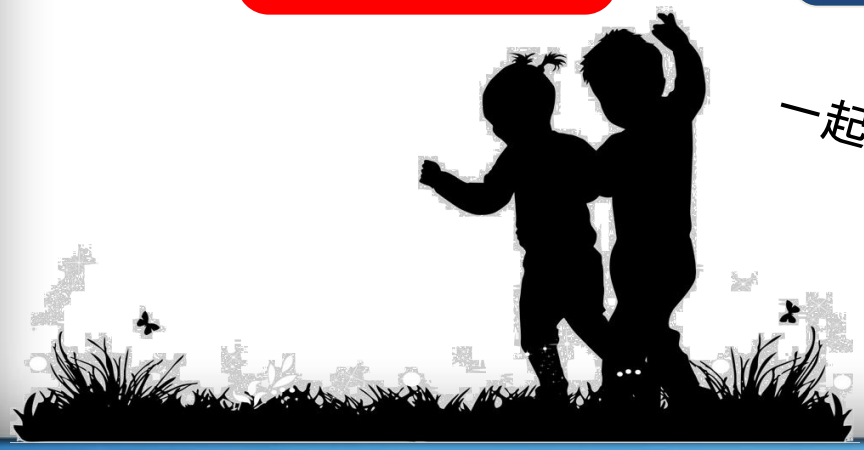
交往

我已經
看到結局了！

出遊

?

?



一起粗乃丸！

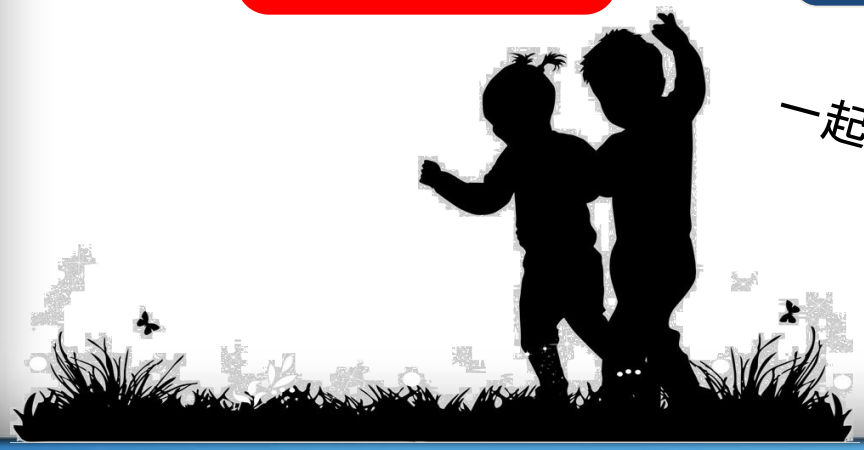
機率上升

交往

出遊

?

一起粗乃



資料預測級

連續&類別資料

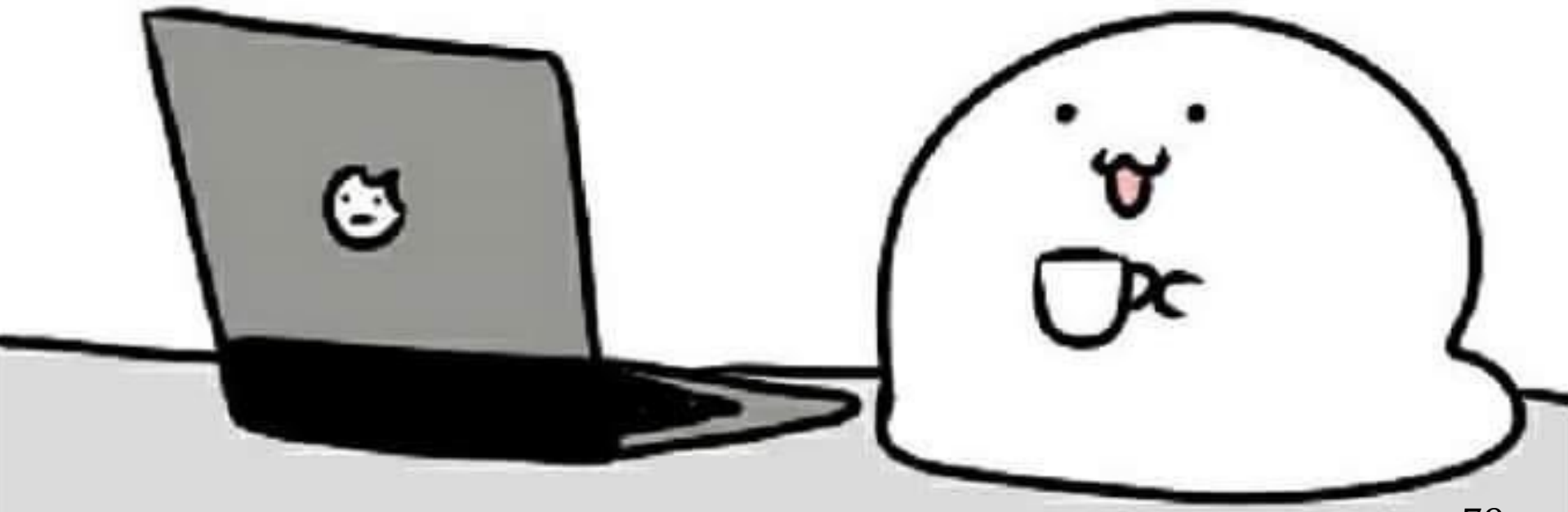
貝氏網路

5/23 機率占卜

資料的分類與預測

哇~

我到底看了什麼？



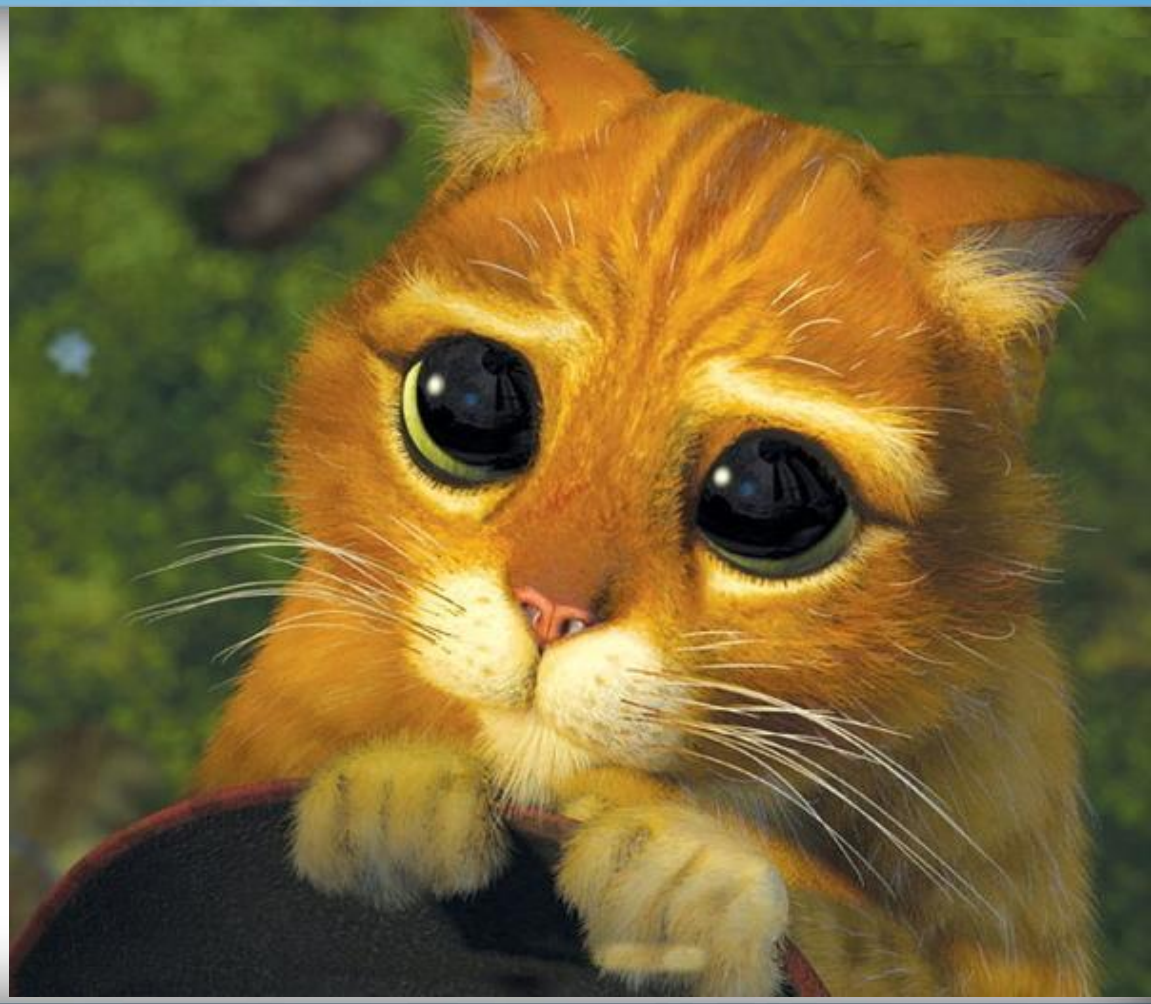
Part 4.

修課規定

修課要求

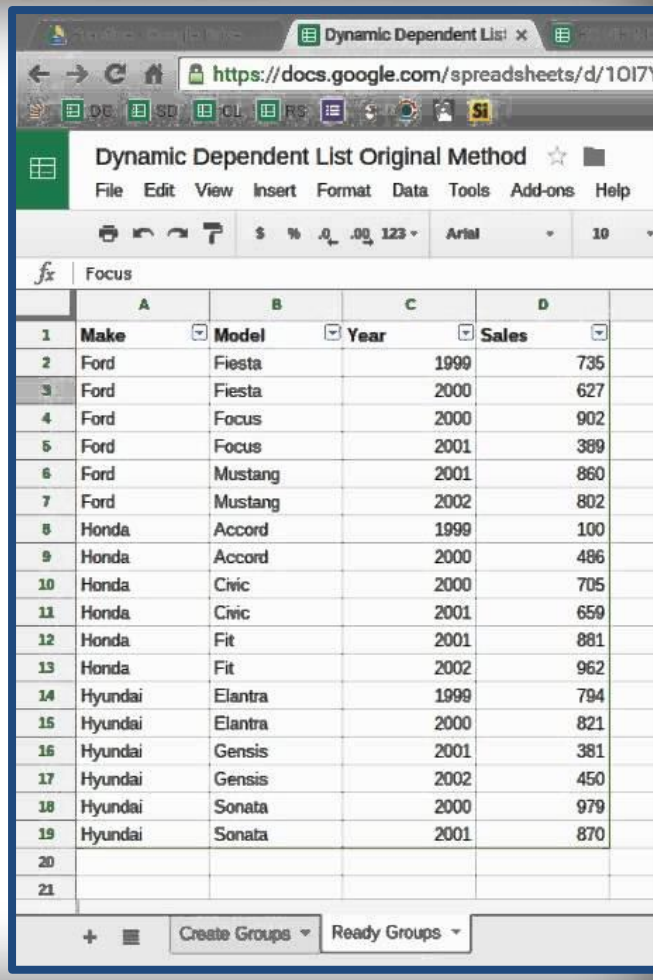
這堂課聽起來很難

我可以修嗎？



1. 看到大量資料不會逃跑

我得了一種看到數字就會逃跑的病



	A	B	C	D
1	Make	Model	Year	Sales
2	Ford	Fiesta	1999	735
3	Ford	Fiesta	2000	627
4	Ford	Focus	2000	902
5	Ford	Focus	2001	389
6	Ford	Mustang	2001	860
7	Ford	Mustang	2002	802
8	Honda	Accord	1999	100
9	Honda	Accord	2000	486
10	Honda	Civic	2000	705
11	Honda	Civic	2001	659
12	Honda	Fit	2001	881
13	Honda	Fit	2002	962
14	Hyundai	Elantra	1999	794
15	Hyundai	Elantra	2000	821
16	Hyundai	Gensis	2001	381
17	Hyundai	Gensis	2002	450
18	Hyundai	Sonata	2000	979
19	Hyundai	Sonata	2001	870
20				
21				

2. 會點滑鼠

3. 會打數字

你真的會打數字嗎？

4. 會輸入中英文

相關技能 (非必備)

1. 用過試算表工具：Excel

   會用函式, 例如=SUM(A1:A100)

   不會也沒關係, 本課程會教

2. 有點程式概念：JavaScript (HTML)、R

   但本課程不寫程式

3. 用過SPSS、SAS、SQL Server等統計與資料探勘工具

   不過本課程不使用這些工具

課程使用的工具

Google 試算表
https://docs.google.com/spreadsheets/

R-Web
http://www.r-web.com.tw/

Weka
http://www.cs.waikato.ac.nz/ml/weka/

RStudio
https://www.rstudio.com/

上課方式

- 教師:課程單元介紹
  - 本單元的位置
  - 說明處理問題、資料類型
  - 分析技術的原理
  - 課程展示實作
- 學生:實作 (計分)
  - 重複課程展示實作
  - 課程練習

期中考與期末考規定

1. 上機實作
2. 題目類型
   a. 名詞配對 : 專有名詞 - 說明定義
   b. 該範圍各單元的實作學習單
   (操作方式類似課堂練習)
3. 考試規定
   a. OpenBook形式 : 可上網看投影片講義
   b. 不可交談、實作不可假手他人, 請當場親自完成

歡迎發問

Thank you for your attention