

Data Mining

Overview and Applications of Data Mining

Jia-Wei Chang

Department of Computer Science and Information Engineering
National Taichung University of Science and Technology

Basic Background of Data Science

Data?

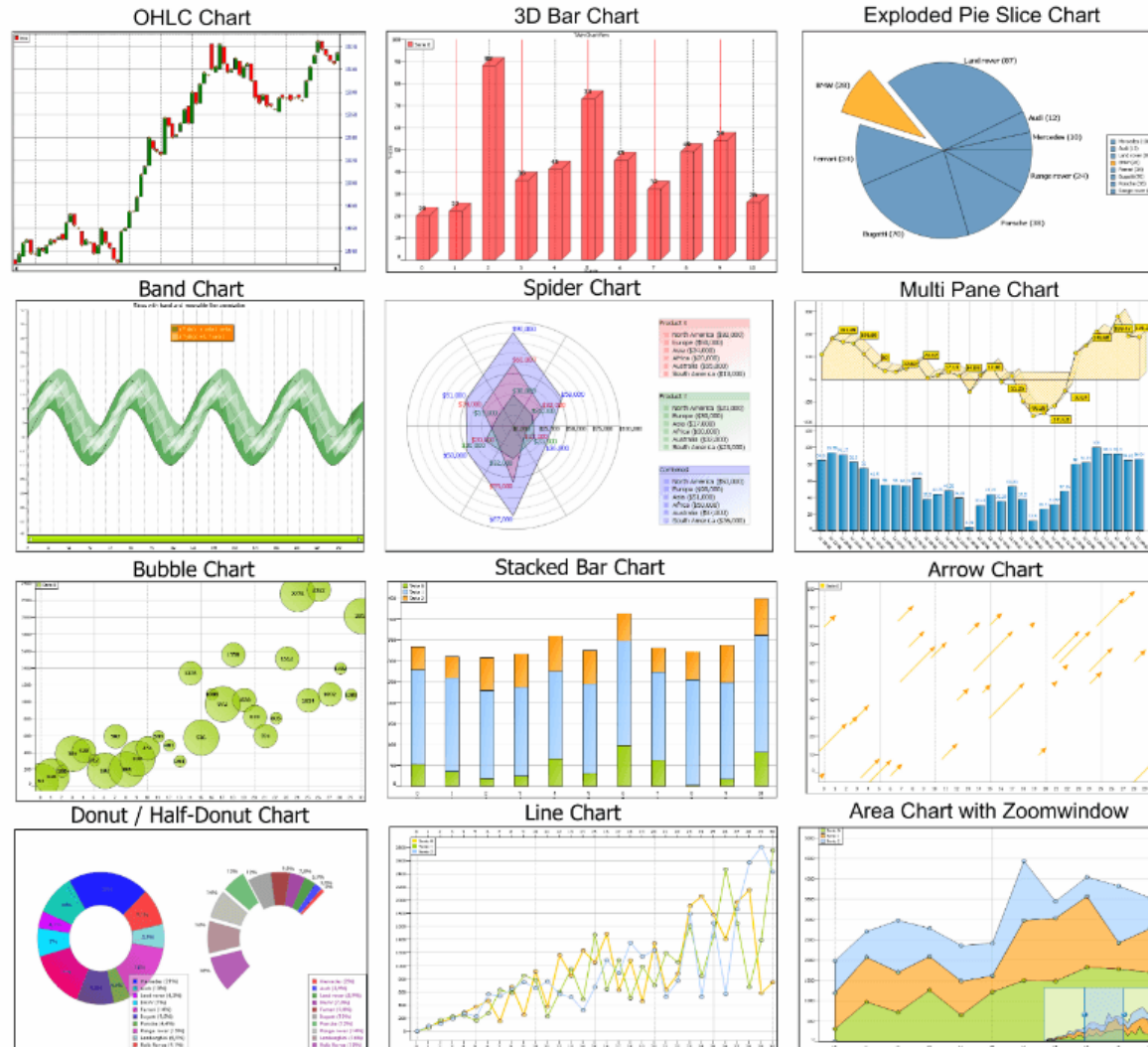
- 未經過處理的原始記錄。

RespondentId	StartDate	CompletedDate	LanguageCode	Question1	Question2	Question3	Question4	Question5	Question6	Question7	Question8
27357	2006.11.27	15:6, 2006.11.27	15:7,en,Denmark,Financial Services,6 - 12 months,26-100,4,4,2,"cvbcvb",2,3,3,1,Opinio,1,0,0,1								
27359	2006.11.27	15:7, 2006.11.27	15:8,en,Italy,Hardware Vendor,1 - 2 years,26-100,3,5,4,,1,3,3,4,Opinio,0,0,0,0,1,0,0,1,0,,0								
27360	2006.11.27	15:8, 2006.11.27	15:8,en,Lithuania,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opini,								
27361	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Opinio,0								
27362	2006.11.27	15:8, 2006.11.27	15:8,en,Djibouti,Manufacturing,6+ years,101-250,0,4,0,"another random text",5,5,5,5,Opinio,1,								
27363	2006.11.27	15:8, 2006.11.27	15:8,en,Tanzania,Retail,1 - 2 years,1001-5000,1,1,1,"123456",2,2,2,2,Opinio,0,1,1,1,1,1,1,1,								
27364	2006.11.27	15:8, 2006.11.27	15:8,en,Vanuatu,Other,1 - 2 years,1001-5000,6,5,6,"123456",6,6,6,6,Opinio,0,0,1,1,1,0,1,1,1,"								
27365	2006.11.27	15:8, 2006.11.27	15:8,en,Angola,Government,1 - 2 years,11-25,4,2,4,"123456",3,3,3,3,Opinio,0,0,1,1,1,1,1,0,,								
27366	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Manufacturing,<6 months,1-5,1,4,1,"hey",5,5,5,5,Opinio,0,1,0,0,0,1,0,0,0,"hey								
27367	2006.11.27	15:8, 2006.11.27	15:8,en,Norway,Education,2 - 5 years,5001-10000,6,0,6,"f6{[]]+&#x' '*-/\\",1,1,1,1,Opinio,1,								
27368	2006.11.27	15:8, 2006.11.27	15:8,en,Bermuda,Software Vendor,1 - 2 years,11-25,0,2,0,"123456",3,3,3,2,Opinio,1,0,1,0,0,1,0								
27369	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Transportation,1 - 2 years,11-25,5,4,5,"123456",5,5,5,5,Opinio,0,1,0,0,0,1,0,0								
27370	2006.11.27	15:8, 2006.11.27	15:8,en,Maldives,Other,6+ years,10001 or more,2,5,2,"another random text",6,6,6,6,Network Pro								
27371	2006.11.27	15:8, 2006.11.27	15:8,en,Kyrgyzstan,Medical,2 - 5 years,26-100,3,5,3,"f6{[]]+&#x' '*-/\\",6,6,6,6,Network Pro								
27372	2006.11.27	15:8, 2006.11.27	15:8,en,Antigua and Barbuda,Government,6 - 12 months,501-1000,6,2,6,"this is a random other t								
27373	2006.11.27	15:8, 2006.11.27	15:8,en,Belarus,Financial Services,6+ years,10001 or more,2,1,2,"another random text",2,2,2,2,								
27374	2006.11.27	15:8, 2006.11.27	15:8,en,Vatican City,Non-profit,1 - 2 years,11-25,0,0,0,"123456",1,1,1,1,Network Probe,1,0,0,								
27375	2006.11.27	15:8, 2006.11.27	15:8,en,Georgia,Financial Services,6+ years,10001 or more,6,1,6,"another random text",2,2,2,2,								
27376	2006.11.27	15:8, 2006.11.27	15:8,en,Tokelau,Transportation,1 - 2 years,11-25,2,4,2,"123456",5,5,5,5,Network Probe,0,1,0,0								
27377	2006.11.27	15:8, 2006.11.27	15:8,en,Chad,Software Vendor,<6 months,1-5,6,2,6,"hey",3,3,3,3,Network Probe,1,1,1,1,1,1,1,1,								
27378	2006.11.27	15:8, 2006.11.27	15:8,en,Turkey,Software Vendor,6 - 12 months,501-1000,1,2,1,"this is a random other text",3,3								
27379	2006.11.27	15:8, 2006.11.27	15:8,en,East Timor,Transportation,<6 months,1-5,0,4,0,"hey",5,5,5,5,Opinio,1,1,0,0,1,0,1,1,0,								
27380	2006.11.27	15:8, 2006.11.27	15:8,en,Nicaragua,Medical,6 - 12 months,6-10,5,5,5,"this is a random other text",6,6,6,6,Opin								
27381	2006.11.27	15:8, 2006.11.27	15:8,en,Equatorial Guinea,Software Vendor,6+ years,101-250,6,2,6,"another random text",3,3,3,								
27382	2006.11.27	15:8, 2006.11.27	15:8,en,Zambia,Retail,<6 months,251-500,1,1,1,"hey",2,2,2,2,Surveyor,0,1,0,0,0,0,0,1,0,"hey"								
27383	2006.11.27	15:8, 2006.11.27	15:8,en,French Southern and Antarctic Lands,Retail,1 - 2 years,1001-5000,2,1,2,"123456",2,2,2								
27384	2006.11.27	15:8, 2006.11.27	15:8,en,Guinea-Bissau,Hardware Vendor,2 - 5 years,26-100,6,3,6,"f6{[]]+&#x' '*-/\\",4,4,4,4,								
27385	2006.11.27	15:8, 2006.11.27	15:8,en,Viet Nam,Medical,2 - 5 years,26-100,4,5,4,"f6{[]]+&#x' '*-/\\",6,6,6,6,Opinio,1,1,1,								
27386	2006.11.27	15:8, 2006.11.27	15:8,en,Reunion,Medical,1 - 2 years,1001-5000,2,5,2,"123456",6,6,6,6,Opinio,1,1,1,1,1,1,1,1,								
27387	2006.11.27	15:8, 2006.11.27	15:8,en,Puerto Rico,Non-profit,<6 months,1-5,0,0,0,"hey",1,1,1,1,1,Opinio,1,1,1,0,1,1,1,0,"h								
27388	2006.11.27	15:8, 2006.11.27	15:8,en,East Timor,Financial Services,6 - 12 months,6-10,1,1,1,"this is a random other text",								
27389	2006.11.27	15:8, 2006.11.27	15:8,en,Northern Mariana Islands,Software Vendor,<6 months,1-5,2,2,2,"hey",3,3,3,3,Opinio,1,0								

Information?

- 資訊是經過處理後的資料。
- 資訊是有用的或有意義的資料。
- 對接受者有意義的資料能使接受者產生資訊。

Information?



Knowledge?

- 知識是資訊、文化脈絡以及經驗的整合。
- 知識是對某個主題確信的認識，並且這些認識擁有潛在的能力為特定目的而使用。
- 藉由專業技能或豐富經驗用以分析資訊的結果。

Intelligence?

- 以知識為基礎，運用個人能力，實踐能力來開創價值。
- 分析、判斷、創造、思考的能力。
- 智慧具有反應能力與價值判斷。

人工智慧？

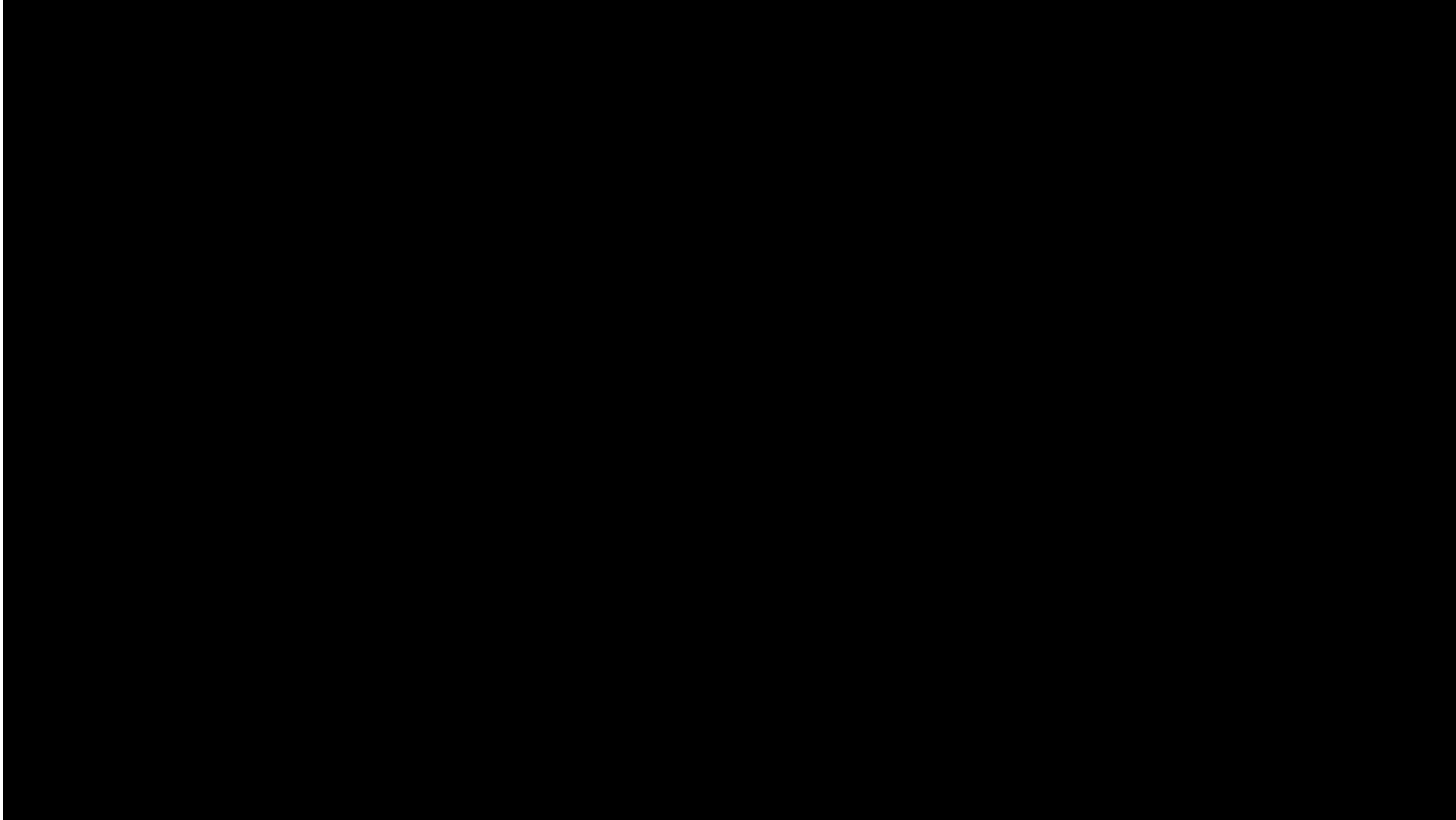
- 指由人製造出來的機器所表現出來的智慧。(Wiki)
- 弱人工智慧 → 專家系統
 - 處理特定的問題
- 強人工智慧
 - 通用人工智慧

IBM Watson 益智節目



<https://youtu.be/5LGrKvbixlk>

IBM Watson



<https://youtu.be/gZ9yy58yBKs>

過去

- Small Data
 - 針對某一個問題，只能獲得小量數據。
 - 數百筆到數萬筆。
 - 花費大量人工編碼。

過去

- Small Data 統計分析
 - 樣本推論母體(抽樣)
 - 在小樣本中，需要發展一系列理論來解釋事物的原理(學說)
 - [啟示] 1936 羅斯福與藍頓 的民調

過去

- 小數據
- Rule-based AI
- 類神經網路 (1980s)

現在與未來

- Big Data
 - 由"母體"來分析數據
 - 數萬筆到幾近無限
 - 雜亂的原始資料

上一世代

- 大數據
- 分類：SVM → 機器學習
- 分群：Kmeans
- 關聯式法則：Apriori

What's difference?

- Small Data vs Big Data
 - 都有目的或待解的問題

But

- 減少假設
- 力求呈現真實世界

What's difference?

- 資料可重組與檢視關聯。
- 接受「數據的雜亂性」，不再追求「精確」的數據。
- 重「相關」而輕「因果」。

現在

- 大數據
- 運算力的提升
- 深度學習 (強AI的可能性)
 - 類神經網路的文藝復興
- 演進趨勢
 - 腦神經科學
 - 認知科學
 - 認知心理學

AlphaGo



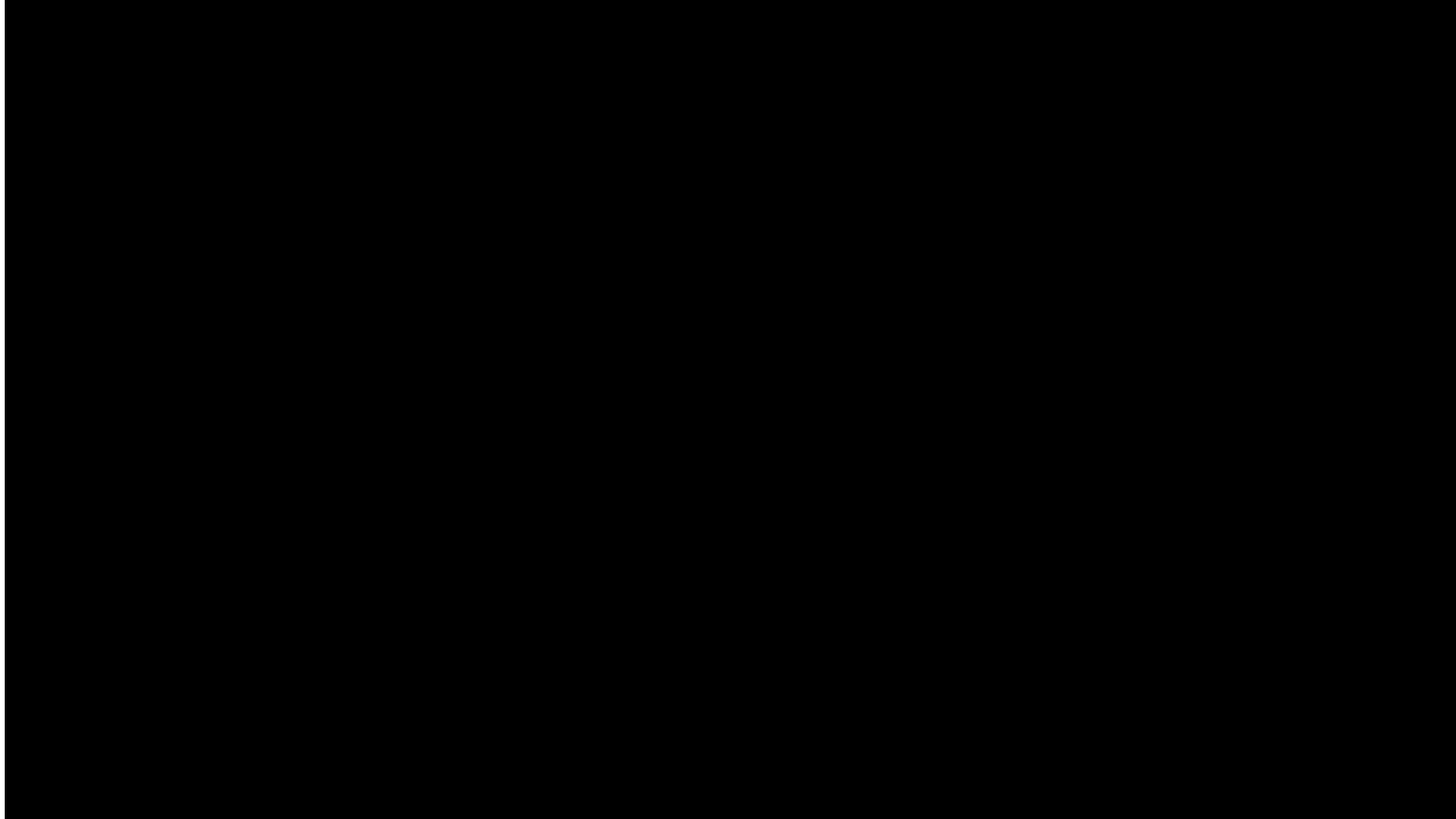
<https://www.youtube.com/watch?v=SUbgqkXVx0A>

AlphaGo ZERO



<https://www.facebook.com/watch/?v=1318229674952254>

AlphaGo 中理解增強式學習



<https://youtu.be/WIZH61sODNI>

Big Data 的沿革 (1/3)

- Data Mining

- 資料探勘是利用分析技術來發掘資料間未知的關聯性與規則。
- 少女未婚懷孕 購物商場比老爸還早知道？！
 - <https://www.nownews.com/news/20120223/42676>

Data Mining

✓分群

- 用於沒有標籤的資料，又通常為非監督式演算法。

✓分類

- 用於有標籤的資料，又通常為監督式演算法。

✓關聯式法則

- 有序性規則的資料

Data Mining

✓分群

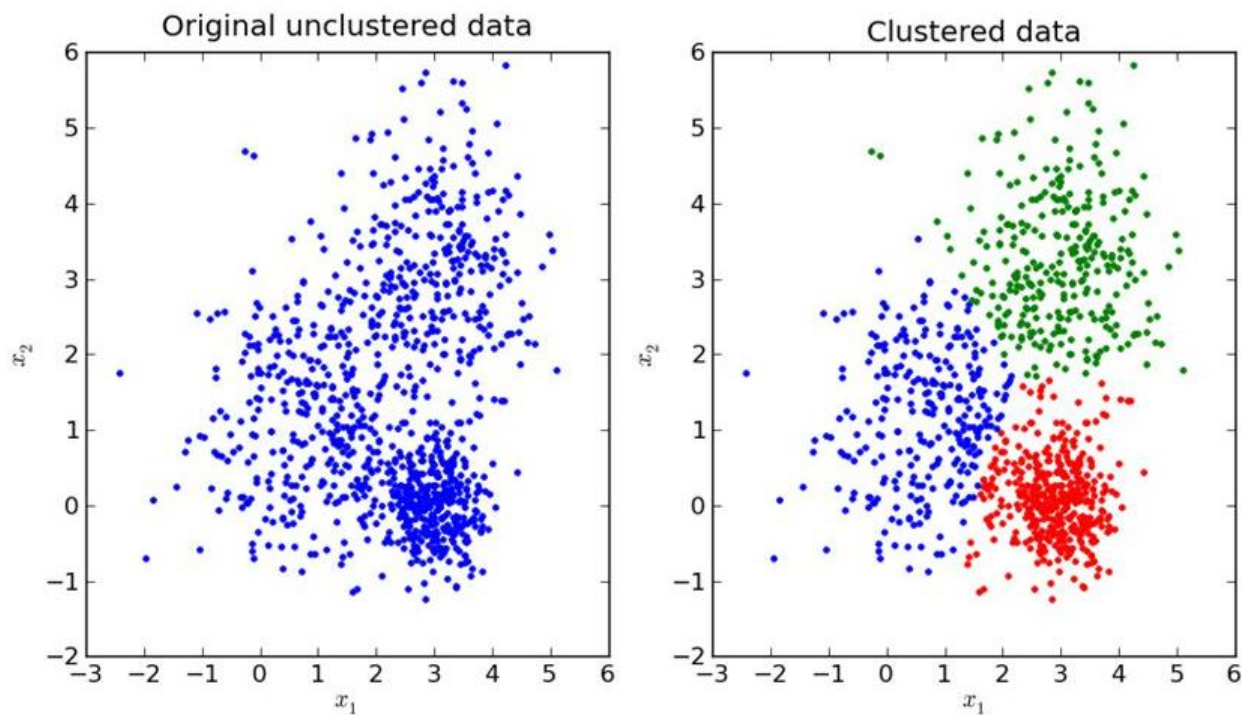
- 用於沒有標籤的資料，又通常為非監督式演算法。

RespondentId	StartDate	CompletedDate	LanguageCode	Question1	Question2	Question3	Question4	Question5	Question6	Question7	Question8
27357	2006.11.27	15:6, 2006.11.27	15:7,en,Denmark,Financial Services,	6 - 12 months,	26-100,4,4,2,"cvbcvcb",	2,3,3,1,Opinio,	1,0,0,1				
27359	2006.11.27	15:7, 2006.11.27	15:8,en,Italy,Hardware Vendor,	1 - 2 years,	26-100,3,5,4,,1,3,3,4,Opinio,	0,0,0,0,1,0,0,1,0,,,	0				
27360	2006.11.27	15:8, 2006.11.27	15:8,en,Lithuania,Retail,	6 - 12 months,	6-10,4,1,4,"this is a random other text",	2,2,2,2,Opini					
27361	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Retail,	6 - 12 months,	6-10,4,1,4,"this is a random other text",	2,2,2,2,Opinio,					
27362	2006.11.27	15:8, 2006.11.27	15:8,en,Djibouti,Manufacturing,	6+ years,	101-250,0,4,0,"another random text",	5,5,5,5,Opinio,	1				
27363	2006.11.27	15:8, 2006.11.27	15:8,en,Tanzania,Retail,	1 - 2 years,	1001-5000,1,1,1,"123456",	2,2,2,2,Opinio,	0,1,1,1,1,1,1,1,				
27364	2006.11.27	15:8, 2006.11.27	15:8,en,Vanuatu,Other,	1 - 2 years,	1001-5000,6,5,6,"123456",	6,6,6,6,Opinio,	0,0,1,1,1,0,1,1,				
27365	2006.11.27	15:8, 2006.11.27	15:8,en,Angola,Government,	1 - 2 years,	11-25,4,2,4,"123456",	3,3,3,3,Opinio,	0,0,1,1,1,1,1,1,0,,				
27366	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Manufacturing,	<6 months,	1-5,1,4,1,"hey",	5,5,5,5,Opinio,	0,1,0,0,0,1,0,0,0,,	"hey"			
27367	2006.11.27	15:8, 2006.11.27	15:8,en,Norway,Education,	2 - 5 years,	5001-10000,6,0,6,"f6{[]]+aæ' '*-/+",	1,1,1,1,Opinio,	1				
27368	2006.11.27	15:8, 2006.11.27	15:8,en,Bermuda,Software Vendor,	1 - 2 years,	11-25,0,2,0,"123456",	3,3,3,3,Opinio,	1,0,1,0,0,1,0				
27369	2006.11.27	15:8, 2006.11.27	15:8,en,Panama,Transportation,	1 - 2 years,	11-25,5,4,5,"123456",	5,5,5,5,Opinio,	0,1,0,0,0,1,0,0				
27370	2006.11.27	15:8, 2006.11.27	15:8,en,Maldives,Other,	6+ years,	10001 or more,	2,5,2,"another random text",	6,6,6,6,Network Pro				
27371	2006.11.27	15:8, 2006.11.27	15:8,en,Kyrgyzstan,Medical,	2 - 5 years,	26-100,3,5,3,"f6{[]]+aæ' '*-/+",	6,6,6,6,Network Pro					
27372	2006.11.27	15:8, 2006.11.27	15:8,en,Antigua and Barbuda,Government,	6 - 12 months,	501-1000,6,2,6,"this is a random other t						
27373	2006.11.27	15:8, 2006.11.27	15:8,en,Belarus,Financial Services,	6+ years,	10001 or more,	2,1,2,"another random-text",	2,2,2,2				
27374	2006.11.27	15:8, 2006.11.27	15:8,en,Vatican City,Non-profit,	1 - 2 years,	11-25,0,0,0,"123456",	1,1,1,1,Network Probe,	1,0,0,				
27375	2006.11.27	15:8, 2006.11.27	15:8,en,Georgia,Financial Services,	6+ years,	10001 or more,	6,1,6,"another random text",	2,2,2,2				
27376	2006.11.27	15:8, 2006.11.27	15:8,en,Tokelau,Transportation,	1 - 2 years,	11-25,2,4,2,"123456",	5,5,5,5,Network Probe,	0,1,0,0				
27377	2006.11.27	15:8, 2006.11.27	15:8,en,Chad,Software Vendor,	<6 months,	1-5,6,2,6,"hey",	3,3,3,3,Network Probe,	1,1,1,1,1,1,1,				
27378	2006.11.27	15:8, 2006.11.27	15:8,en,Turkey,Software Vendor,	6 - 12 months,	501-1000,1,2,1,"this is a random other text",	3,3					
27379	2006.11.27	15:8, 2006.11.27	15:8,en,East Timor,Transportation,	<6 months,	1-5,0,4,0,"hey",	5,5,5,5,Opinio,	1,1,0,0,1,0,1,1,0,				
27380	2006.11.27	15:8, 2006.11.27	15:8,en,Nicaragua,Medical,	6 - 12 months,	6-10,5,5,5,"this is a random other text",	6,6,6,6,Opin					
27381	2006.11.27	15:8, 2006.11.27	15:8,en,Equatorial Guinea,Software Vendor,	6+ years,	101-250,6,2,6,"another random-text",	3,3,3,					
27382	2006.11.27	15:8, 2006.11.27	15:8,en,Zambia,Retail,	<6 months,	251-500,1,1,1,"hey",	2,2,2,2,Surveyor,	0,1,0,0,0,0,0,1,0,,	"hey"			
27383	2006.11.27	15:8, 2006.11.27	15:8,en,French Southern and Antarctic Lands,Retail,	1 - 2 years,	1001-5000,2,1,2,"123456",	2,2,2					
27384	2006.11.27	15:8, 2006.11.27	15:8,en,Guinea-Bissau,Hardware Vendor,	2 - 5 years,	26-100,6,3,6,"f6{[]]+aæ' '*-/+",	4,4,4,4,					
27385	2006.11.27	15:8, 2006.11.27	15:8,en,Viet Nam,Medical,	2 - 5 years,	26-100,4,5,4,"f6{[]]+aæ' '*-/+",	6,6,6,6,Opinio,	1,1,1,				
27386	2006.11.2										

Data Mining

✓分群

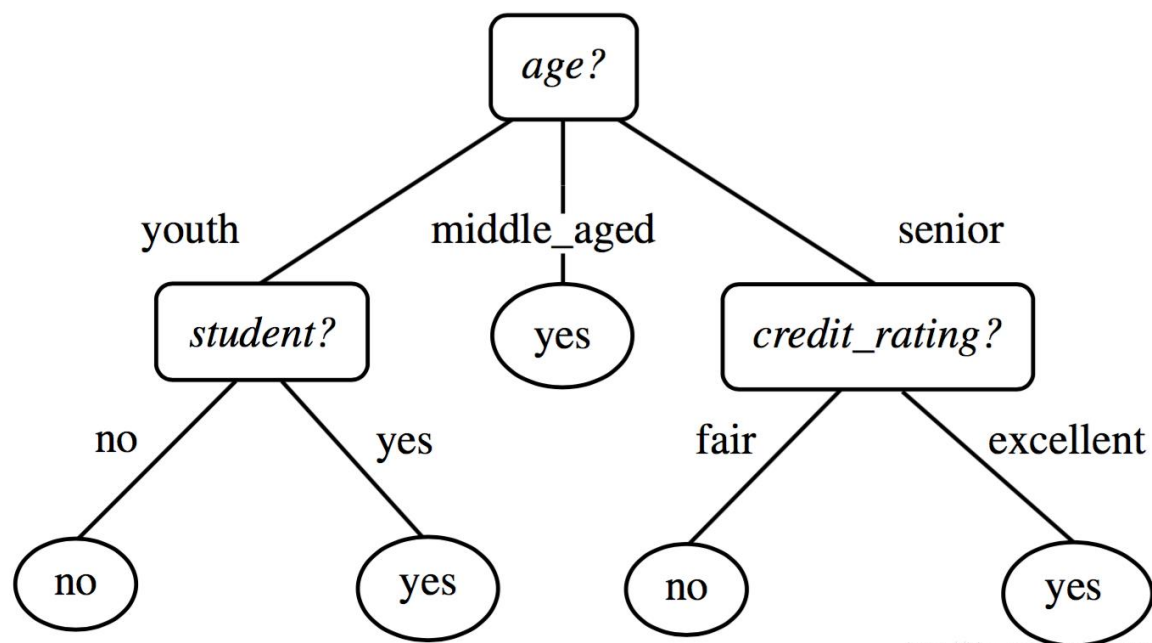
- 用於沒有標籤的資料，又通常為非監督式演算法。



Data Mining

✓分類

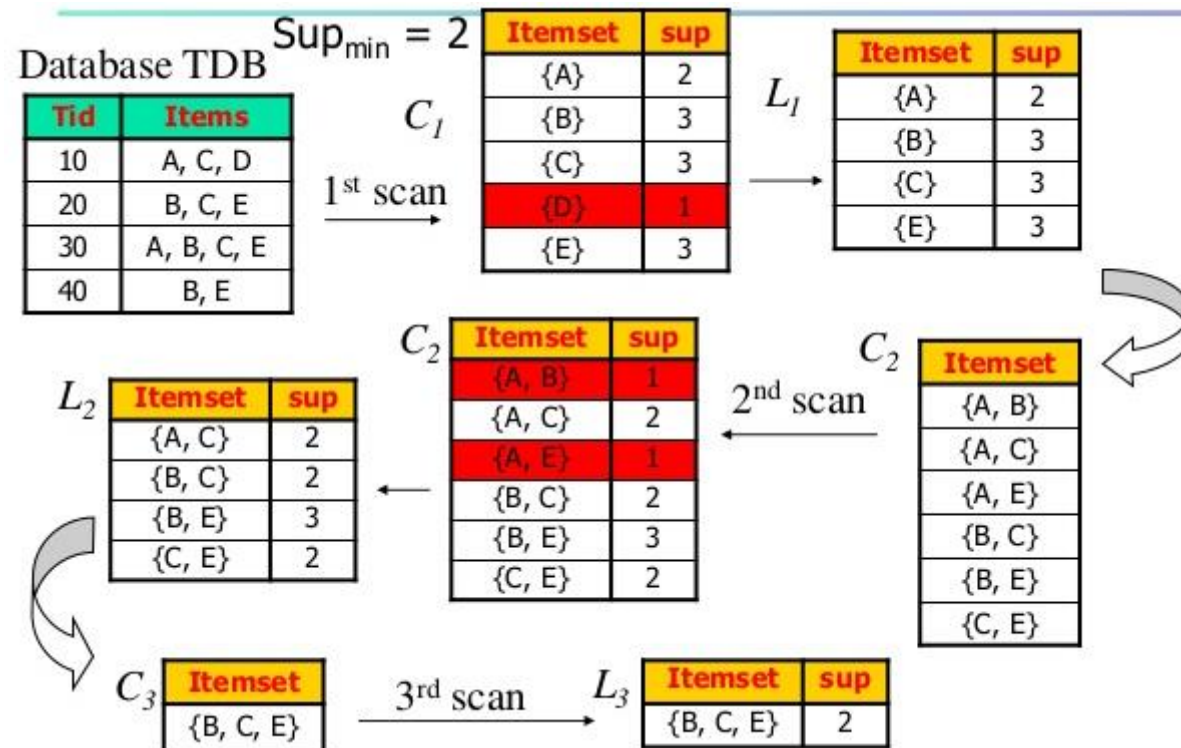
- 用於有標籤的資料，又通常為監督式演算法。



Data Mining

✓ 關聯式法則

- 有序性 (尿布與啤酒)



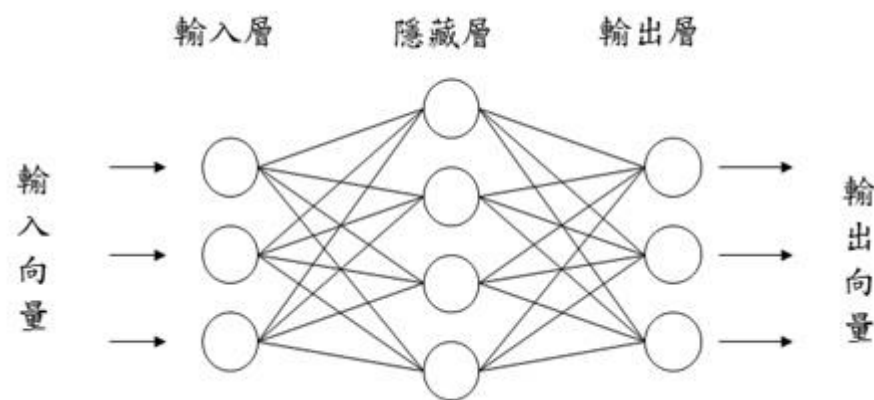
Big Data 的沿革 (2/3)

- Machine Learning
 - 人工智慧的分支，可用於資料探勘。
 - 讓機器可以自動學習、從巨量資料中找到規則，進而有能力做出分類或預測。
 - 判斷出類別
 - 估計出數值

Big Data 的沿革 (3/3)

- Deep Learning

- 是機器學習的分支
- 類神經網路的文藝復興
- 從大規模未標記資料中建立更好的預測模型
- 建立強 AI 的可能性



資料分析的基本步驟

1. 資料清除：去除極端、遺失值資料、不重要的屬性
2. 資料整合：因應用目的或特性，整合不同來源的資料
3. 資料選擇：揀選重要的屬性來逼近目的之最佳成效
4. 資料轉換：基於領域知識進行特徵縮放、數值類別轉換等
5. 資料探勘：選用合適的分析演算法得到目的之結果
6. 樣式評估：評估結果的樣式，是否如預期
7. 知識表示：因應目的將樣式轉換成合適的表達方法

資料分析的演算法重點

- 預處理 (Preprocessing)
- 降維 (Dimensionality Reduction)
- 模型選擇 (Model Selection)
 - 監督式學習 (Supervised learning)
 - 分類 (Classification) : 機器給出一個類別
 - 迴歸 (Regression) : 機器給出一個數值
 - 非監督式學習 (Unsupervised learning)
 - 分群 (Clustering)

降維 (Dimensionality reduction)

- 奇異值分解
 - Singular Value Decomposition (SVD)

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

book

dads

dummies

estate

guide

investing

market

real

rich

stock

value

0.15

-0.27

0.04

0.24

0.38

-0.09

0.13

-0.17

0.07

0.18

0.19

0.45

0.22

0.09

-0.46

0.74

-0.21

0.21

0.18

-0.30

-0.28

0.18

0.19

0.45

0.36

0.59

-0.34

0.25

-0.42

-0.28

0.12

-0.14

0.23

3.91

0

0

0

2.61

0

0

0

2.00

T1

T2

T3

T4

T5

T6

T7

T8

T9

0.35

0.22

0.34

0.26

0.22

0.49

0.28

0.29

0.44

-0.32

-0.15

-0.46

-0.24

-0.14

0.55

0.07

-0.31

0.44

-0.41

0.14

-0.16

0.25

0.22

-0.51

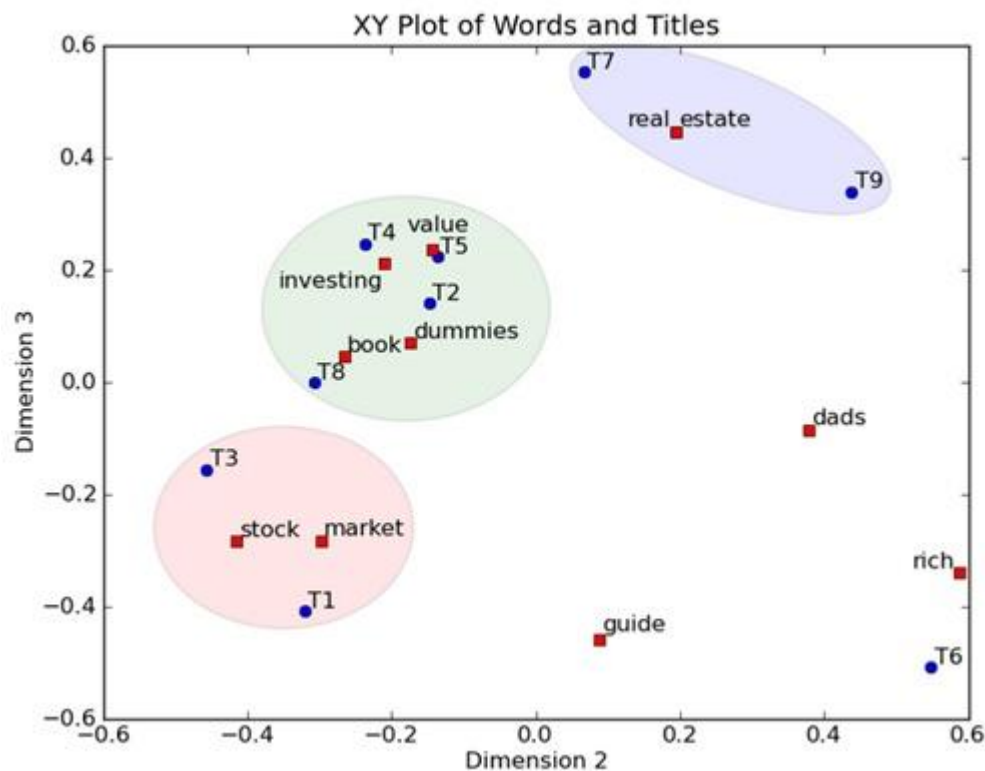
0.55

0.00

0.34

降維 (Dimensionality reduction)

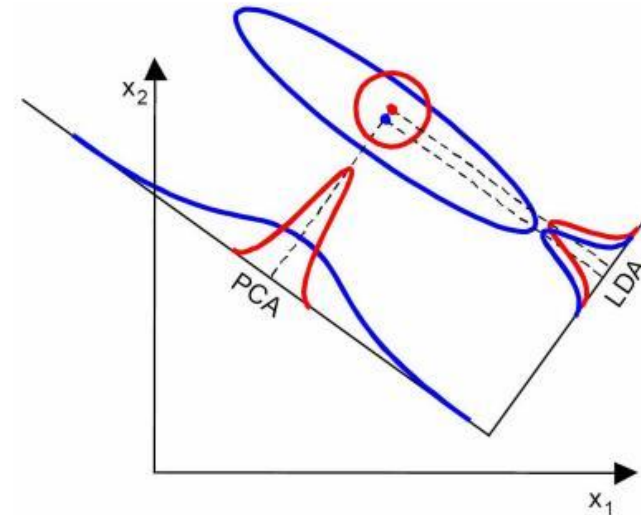
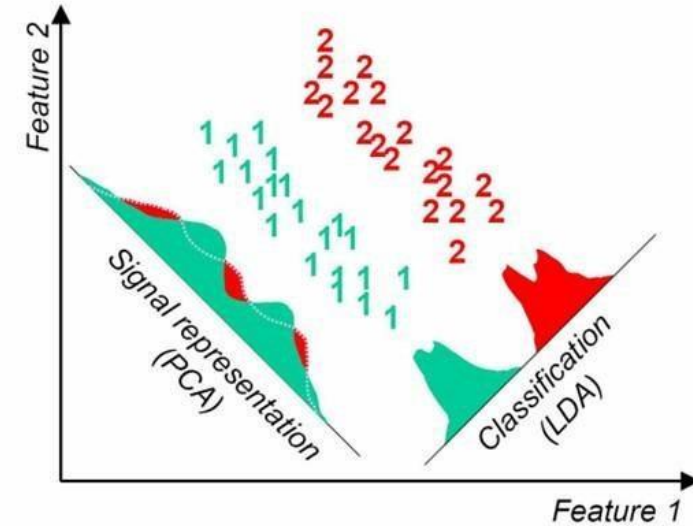
- 奇異值分解
 - Singular Value Decomposition (SVD)



Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

降維 (Dimensionality reduction)

- 主成分分析
 - Principal Component Analysis (PCA)
 - 非監督式
- 線性判別分析
 - Linear Discriminant Analysis (LDA)
 - 監督式



Thinking Time

資料分析的常見角色

- 資料產品經理人：將真實世界的問題轉換成資料可以解決的問題，通常是該問題領域的專業人士
- 資料工程師：蒐集、整理、清理資料，通常是具備程式技術能力的工程師
- 資料分析師：負責資料建模和分析，通常由擅長找出資料關聯的統計人擔當
- 資料視覺化設計師：將報表變得簡明易懂

Applications

以語言學習輔助工具為例

Collocation online suggestion v1.0

英語搭配詞線上檢索系統

[介紹](#) [常用搭配詞查詢](#) [整句搭配詞查詢與推薦](#)

整句搭配詞查詢與推薦

輸入句子: We commonly use a small cell for medical research.

清除

送出

輸入的句子為

We commonly use a small cell for medical research.

副詞修飾(V/Adv/Adj組合)
commonly + V/Adv/Adj

#	collocation	freq(%)
1	commonly use	46.5
2	commonly used	4.7
3	commonly find	4.4
4	commonly know	3.3
5	commonly employ	2.4
6	commonly refer	2.2
7	commonly observe	1.9
8	commonly report	1.9
9	commonly encounter	1.4
10	commonly available	1.3

commonly與use的搭配字同義組合
commonly + 搭配同義字

#	collocation	freq(%)	
1	commonly use	46.5	✍
2	commonly employ	2.4	✍
3	commonly apply	0.5	✍

同義詞搭配詞組搜尋結果
commonly的同義字 + use的同義字

#	collocation	count	
1	commonly use	296	✍
2	often use	140	✍
3	frequently use	68	✍
4	commonly employ	15	✍
5	frequently employ	9	✍
6	often employ	6	✍
7	frequently apply	5	✍
8	repeatedly use	5	✍
9	routinely use	5	✍
10	frequently utilize	4	✍
11	routinely employ	3	✍
12	commonly apply	3	✍

查詢總時間:0.52sec

以語言學習輔助工具為例

	computer	data	pinch	result	sugar
aprocot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

$$P(x = \text{information}, y = \text{data}) = \frac{6}{19} = 0.32$$

$$P(x = \text{information}) = \frac{6 + 4 + 1}{19} = \frac{11}{19} = 0.58$$

$$P(y = \text{data}) = \frac{6 + 1}{19} = \frac{7}{19} = 0.37$$

$$\begin{aligned} & pmi(x = \text{information}, y = \text{data}) \\ &= \log \frac{P(x = \text{information}, y = \text{data})}{P(x = \text{information}) \times P(y = \text{data})} \\ &= \log 1.49 \\ &= 0.57 \end{aligned}$$

以語言學習輔助工具為例

Stanford Parser

Please enter a sentence to be parsed:

My dog also likes eating sausage.

Language: English Sample Sentence

Parse

Your query

My dog also likes eating sausage.

Tagging

My/PRP\$ dog/NN also/RB likes/VBZ eating/VBG sausage/NN ./.

Parse

```
(ROOT
 (S
  (NP (PRP$ My) (NN dog))
  (ADVP (RB also))
  (VP (VBZ likes)
   (S
    (VP (VBG eating)
     (NP (NN sausage)))))
  (. .)))
```

Universal dependencies

```
nmod:poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)
```

Your query

猴子喜欢吃香蕉。

Segmentation

猴子 喜欢 吃 香蕉 。

Tagging

猴子/NN 喜欢/VV 吃/VV 香蕉/NN 。

Parse

```
(ROOT
 (IP
  (NP (NN 猴子))
  (VP (VV 喜欢)
   (IP
    (VP (VV 吃)
     (NP (NN 香蕉)))))
  (PU 。)))
```

Universal dependencies

```
nsubj(喜欢-2, 猴子-1)
root(ROOT-0, 喜欢-2)
ccomp(喜欢-2, 吃-3)
dobj(吃-3, 香蕉-4)
punct(喜欢-2, 。-5)
```

- <http://nlp.stanford.edu:8080/parser/>

以NBA的應用為例

- DevDays Asia 2016
- 入圍前五
 - 24小時黑客松
 - 3人臨時組團
 - 臨時命題
 - 純好玩




NBA公開資料 & Game NBA 2K 能力值

NBA Games Top Stories Video Standings Stats Players Teams Store Tickets

Stats Home / Player / Jeremy Lin

Menu Stats Home Players Teams Scores Schedule Standings SEARCH FOR A PLAYER OR TEAM

 #7 Jeremy Lin
G | BROOKLYN NETS

Compare Player

HT	6-3	WT	200 lbs	PRIOR	Harvard/United States	PTS	14.2	REB	3.7	AST	6.3	P/E	15.1
AGE	28 112d	BORN	8/23/88	DRAFT	Undrafted	EXP	6 yrs						

Profile

Traditional Splits										
BY YEAR	GP	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	F
2016-17	6	25.8	14.2	5.0	11.0	45.5	1.2	3.7	31.8	
2015-16	78	26.3	11.7	3.8	9.3	41.2	1.0	2.9	33.6	
2014-15	74	25.8	11.2	3.7	8.8	42.4	0.9	2.4	36.9	
2013-14	71	28.9	12.5	4.2	9.3	44.6	1.2	3.2	35.8	
2012-13	82	32.2	13.4	4.8	10.9	44.1	1.1	3.1	33.9	
2011-12	35	26.9	14.6	4.9	10.9	44.6	0.7	2.1	32.0	
2010-11	29	9.8	2.6	1.0	2.5	38.9	0.0	0.2	20.0	

JEREMY LIN PG | SG

HEIGHT 75"
WEIGHT 90 KG
AGE 23

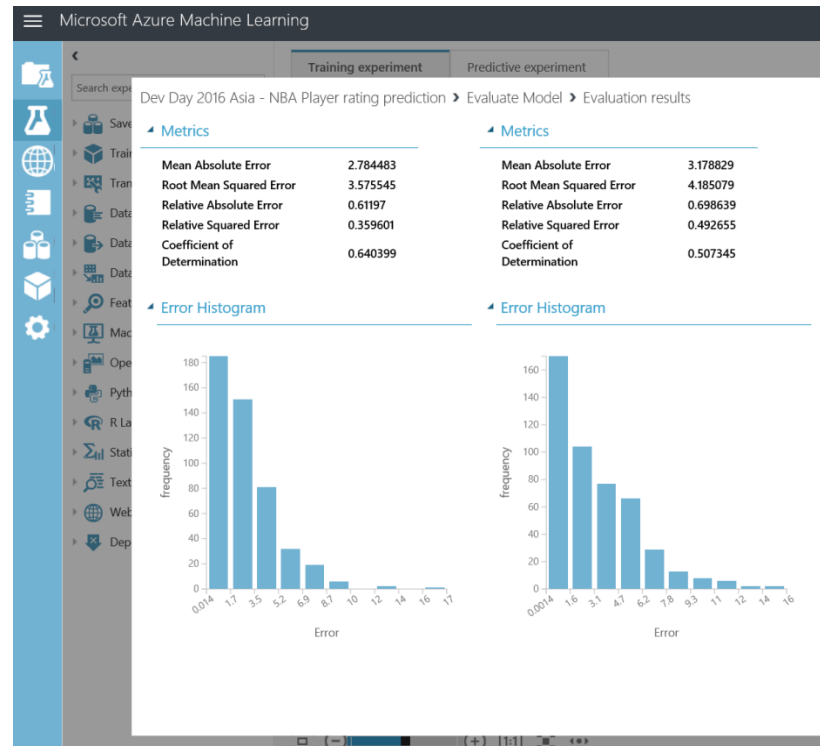
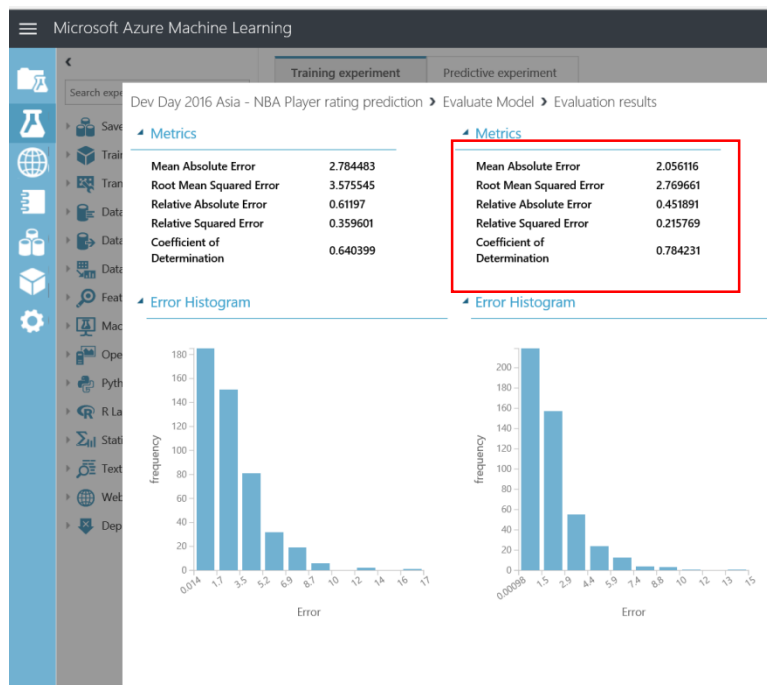
73 OFFENSE
68 DEFENSE
41 REBOUND

LB	ALL POSITIONS			RB	LT	NEW YORK KNICKS			RT
NAME	POS	RTG	IN	OUT	PER D	POST D	HNDR	REB	I.Q.
T.Chandler	C	77	C+	F	C	A	F	A	B+
B.Davis	PG	73	B-	B	B	D	A	D-	B+
I.Shumpert	PG	72	C+	C+	A-	D	B	D-	B+
L.Fields	SG	72	C+	B+	C+	F	C+	C+	B-
T.Douglas	PG	69	C	B+	B+	F	B-	D-	B-
J.Lin	PG	69	C+	C+	B	F	B+	D-	B+

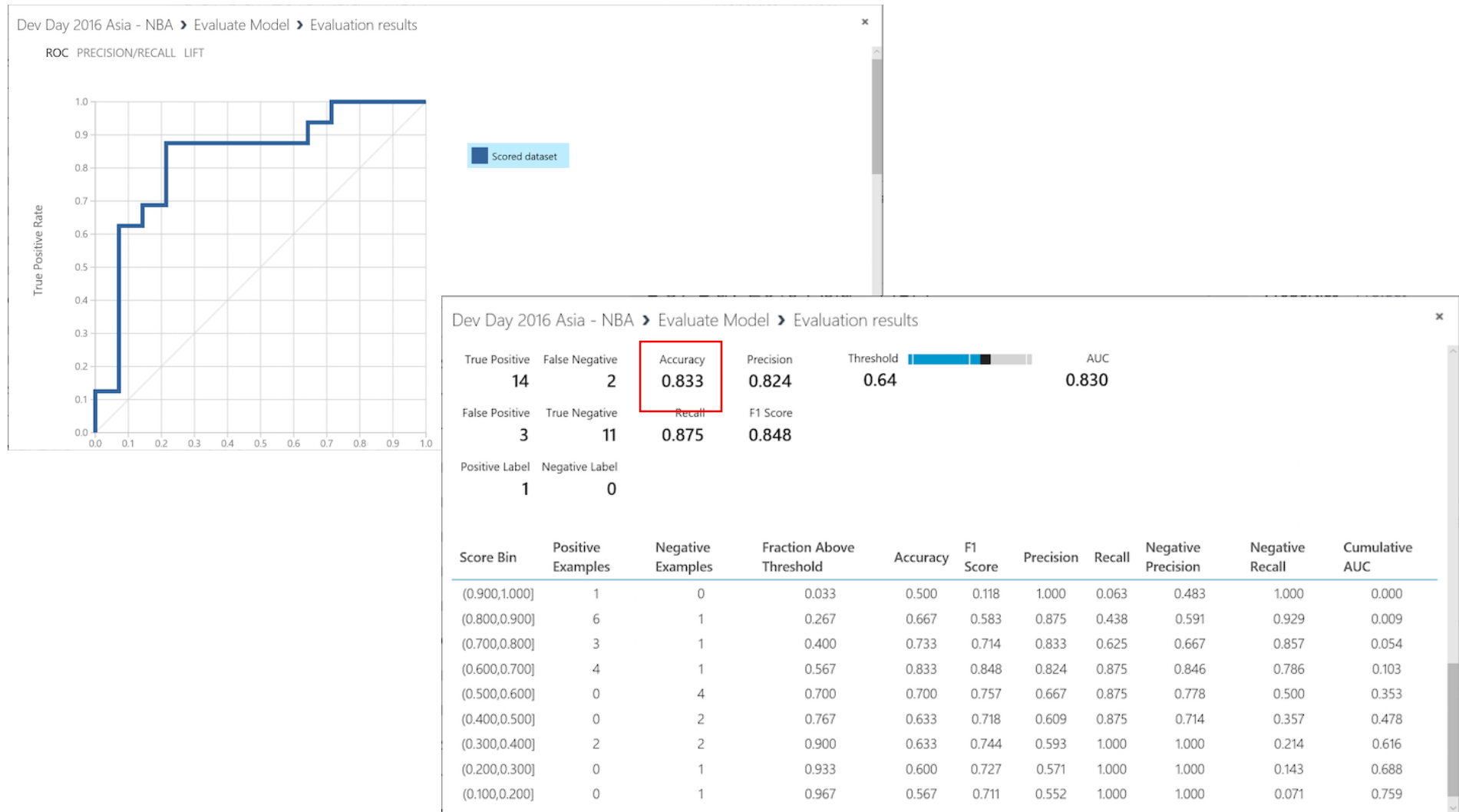
- 資料來源
 - 2K Games
 - NBA

以NBA的應用為例

- NBA 2K17能力預測



NBA季後賽預測



資料驅動創新應用

- 文字、聲音、影像
 - 自然語言處理
 - 語音辨識
 - 影像辨識
- 數值與非數值
 - 連續性
 - 離散性、類別

Thinking Time

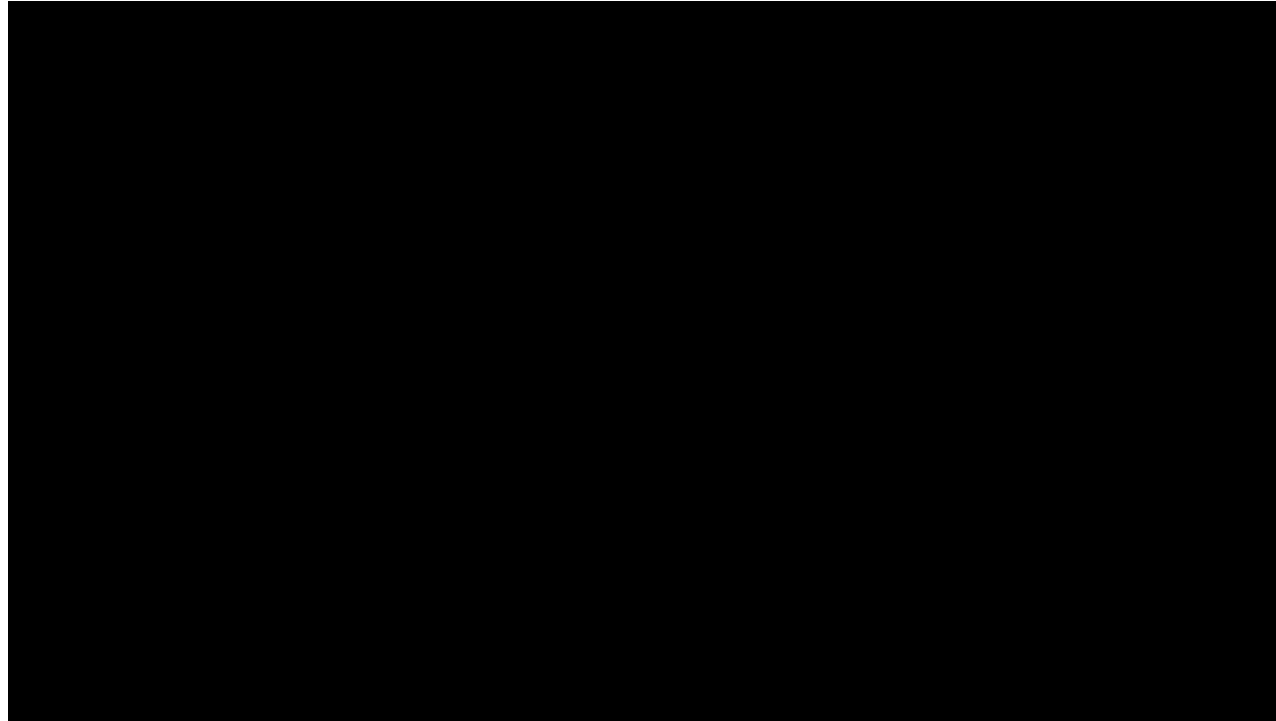
Potential Applications

中文OCR辨識為例



- Lee, M. C., Chiu S. Y., & [Chang, J. W.](#) (2017) A Deep Convolutional Neural Network based Chinese Menu Recognition App. Information Processing Letters, 128, 14-20. <https://doi.org/10.1016/j.ipl.2017.07.010> (SCI, COMPUTER SCIENCE, INFORMATION SYSTEMS)

CNN應用於情緒感知辨識的照護機器人



[link](#)

- Lee, M. C., Yeh, S. C., Chiu, S. Y. & [Chang, J. W.](#) (2017, June). A Deep Convolutional Neural Network Based Virtual Elderly Companion Agent. ACM Multimedia Systems 2017 (MMSYS2017), Taipei, Taiwan. (Accept Rate: 28%) <http://dl.acm.org/citation.cfm?id=3083220>

工業 4.0



<http://www.esta.com.tw/knowledge-info.asp?id=263>

<https://www.youtube.com/watch?v=HexQLQIHrAY>

工業 4.0



Fintech

TradingBot - 程式交易機器人



TradingBot - 程式交易機器人
@tradingbot

首頁
關於

TradingBot 演算法

- 影像辨識訊號處理
- Machine Learning
 - SVM
- 自動交易系統
 - Java

基於模式辨識及支援向量機之期貨當沖交易策略和系統

A Strong Futures Day-Trading Strategy and System Using Pattern Recognition and Support Vector Machine

瞭解詳情 發送訊息

TradingBot - 程式交易機器人

請問想查看哪種商品資訊？

期貨走勢

臺指期038 - 臺指期038 - 臺指現貨



12:31

項目	開盤	最高	最低	收盤	漲跌
臺指期038	11028.00	11080.00	11013.00	11077.00	55.00

項目	開盤	最高	最低	收盤	漲跌
臺指現貨	11028.00	11080.00	11013.00	11077.00	55.00

未平倉 85500 試撮量 -

委買價 38639 委賣價 37578

委買口 95349 委賣口 92083

時間 12:31:20

委買價/委買量 委賣價/委賣量

委買價	委買量	委賣價	委賣量
11076.00	27	11077.00	25
11075.00	45	11078.00	46
11074.00	38	11079.00	72
11073.00	32	11080.00	110
11072.00	54	11081.00	42

最佳衍生一檔價量 11074.00 / 1311078.00 / 9

請選擇下列功能：

交易現況 選擇權策略 金融新聞 商品資訊 未平倉量 到價警示 問答測驗

輸入訊息.....

<https://www.messenger.com/t/tradingbot>

Open Datasets

UC Irvine Machine Learning Repository

- <http://archive.ics.uci.edu/ml/datasets.html>

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Search

任務 資料型態 資料筆數

Browse Through: 426 Data Sets

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (314) Regression (82) Clustering (72) Other (54)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (37) Numerical (273) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (324) Univariate (19) Sequential (44) Time-Series (79) Text (44) Domain-Theory (23) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Area Life Sciences (98) Physical Sciences (47) CS/Engineering (148) Social Sciences (24) Business (26) Game (10) Other (69)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
# Attributes Less than 10 (99) 10 to 100 (195) Greater than 100 (76)	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
# Instances Less than 100 (25) 100 to 1000 (149) Greater than 1000 (220)	 Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Format Type Matrix (292) Non-Matrix (134)	 Audiology (Original)	Multivariate	Classification	Categorical	226		1987
	 Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
	 Auto MPG	Multivariate	Regression	Categorical, Real	398	8	1993
	 Automobile	Multivariate	Regression	Categorical, Integer, Real	205	26	1987
	 Badges	Univariate, Text	Classification		294	1	1994

Table View List View

Kaggle DataSets

- <https://www.kaggle.com/datasets>

The screenshot displays the Kaggle Datasets interface. At the top, there's a 'Public' tab and a 'Sort by' dropdown set to 'Hotness'. Below this, a header bar shows '12,800 Datasets' and filters for 'Sizes', 'File types', 'Licenses', and 'Tags'. A search bar is also present. The main content area lists six datasets, each with a rank, a thumbnail, a title, a description, the creator, the update time, tags, file format, size, license, and view/download counts.

Rank	Thumbnail	Title	Description	Creator	Updated	Tags	Format	Size	License	Views	Downloads
373		Data Science for Good: Kiva Crowdfunding	Use Kernels to assess welfare of Kiva borrowers for \$30k in prizes	Kiva	updated 10 days ago	geography, finance, lending, + 2 more...	CSV	42 MB	CC0	106	45k
295		Trending YouTube Video Statistics	Daily statistics for trending YouTube videos	Mitchell J	updated 12 hours ago	languages, popular culture, statistics, + 2 more...	CSV	64 MB	CC0	27	51k
254		Huge Stock Market Dataset	Historical daily prices and volumes of all U.S. stocks and ETFs	Boris Marjanovic	updated 4 months ago	business, finance, economics, artificial intelligen...	Other	245 MB	CC0	12	28k
154		Bitcoin Blockchain	Complete live historical Bitcoin blockchain data	Google BigQuery	updated a month ago	finance, money, internet, bigquery	BigQuery	821 GB	CC0	430	31k
70		Historical Air Quality	Air Quality Data Collected at Outdoor Monitors Across the US	US Environmental Protection Agency	updated 3 months ago	pollution, bigquery	BigQuery	323 GB	CC0	16	17k
63		HackerRank Developer Survey 2018	Survey of 25,000 professionals and students on the state of developer skills	HackerRank	updated 12 days ago	women, demographics, programming, + 2 more...	CSV	5 MB	CC4	17	6k

臺南市開放資料

- <http://data.tainan.gov.tw/dataset>



The screenshot shows the Tainan Open Data website interface. At the top, there is a navigation bar with the 'DATA.TAINAN' logo and links for '資料建議', '資料集', '群組', '使用說明', '使用規範', '應用展示', and '關於'. A search bar is located on the right. Below the navigation bar, the '資料集' (Dataset) section is active. On the left, a sidebar lists various government departments and their dataset counts: 衛生局 (76), 主計處 (74), 消防局 (73), 地政局 (52), 財政稅務局 (47), 環境保護局 (46), 交通局 (44), 經濟發展局 (39), 農業局 (30), 民政局 (28), and a link to '顯示更多 組織'. Below this, a '群組' (Group) section lists categories like '公共安全/防救災/環保 (94)', '戶籍/地政 (62)', '健康/照護/社福 (62)', '統計數據 (61)', and '其他行政 (44)'. The main content area displays a search bar, the text '找到675個資料集', and a '排序依照' dropdown menu set to '關聯'. Three dataset entries are shown: 1. '臺南市政府市長及副市長公務禮儀受贈品清冊' with a description, contact info, and a CSV download button. 2. '0206地震專區' with a description and CSV, HTML, and KML download buttons. 3. '臺南市地價評議委員會評議案件數量統計表' with a description and a CSV download button. The fourth entry, '臺南市土地徵收補償市價查估及評議作業情形表', is partially visible.

Thank you