AI and ML
Overview and Applications of AI/ML

Jia-Wei Chang

Department of Computer Science and Information Engineering

National Taichung University of Science and Technology
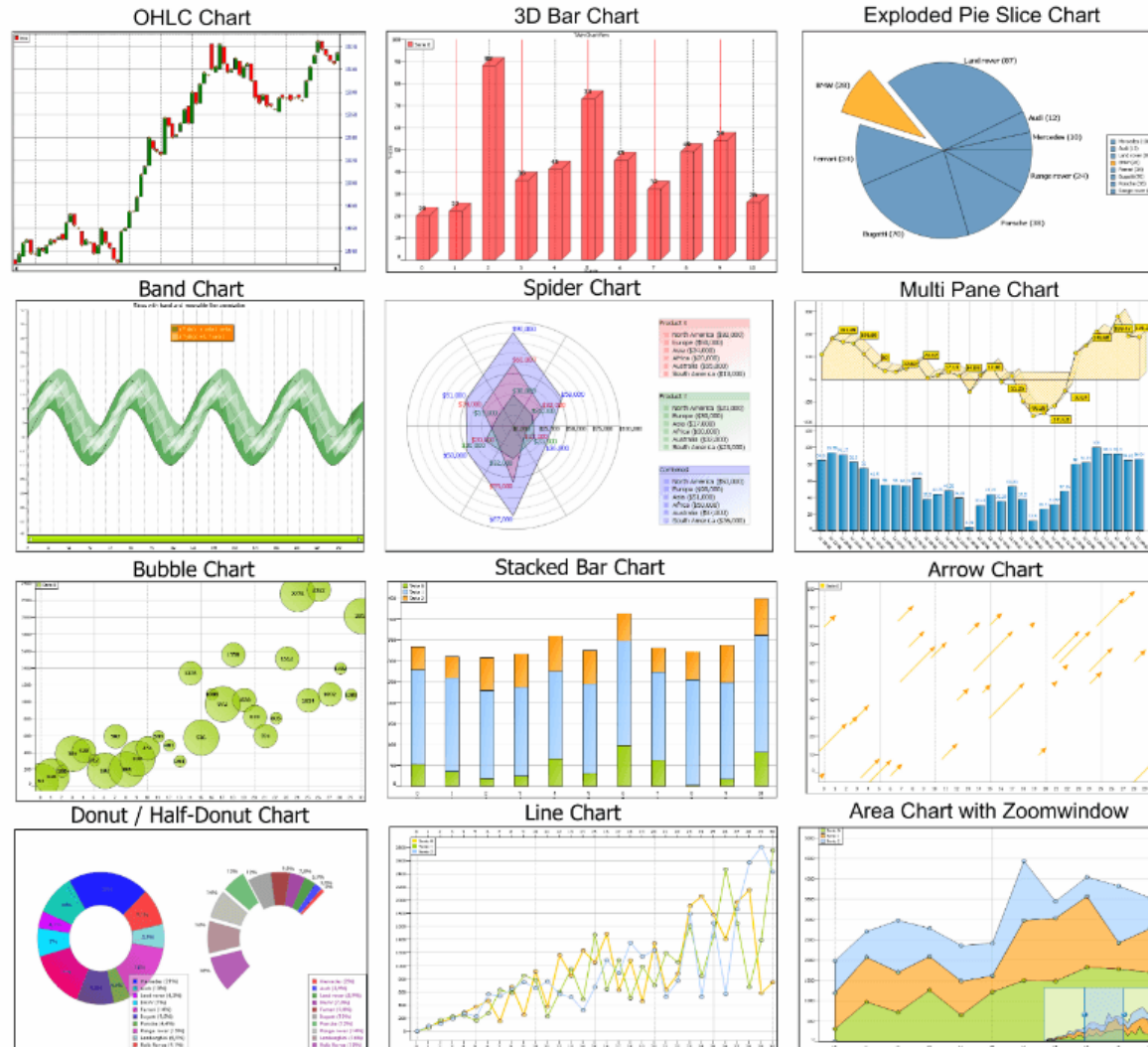
Basic Background of
Data Science

Data?

- 未經過處理的原始記錄。

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/5LGrKvbixIk

IBM Watson

https://youtu.be/gZ9yy58yBKs
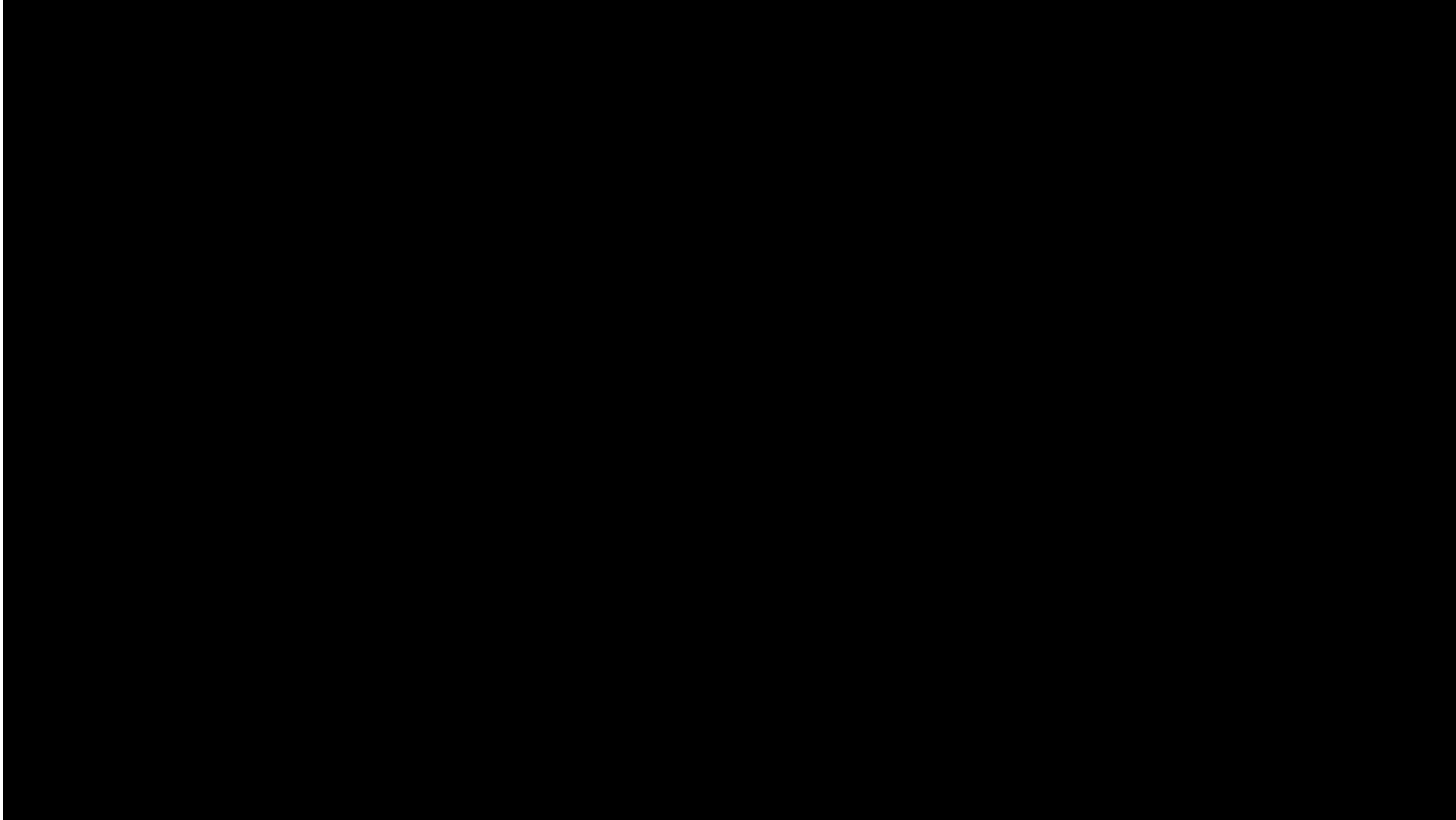
過去

- 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=SUbqykXVx0A

AlphaGo ZERO

AlphaGo 中理解增強式學習

https://youtu.be/WIZH61sODNI
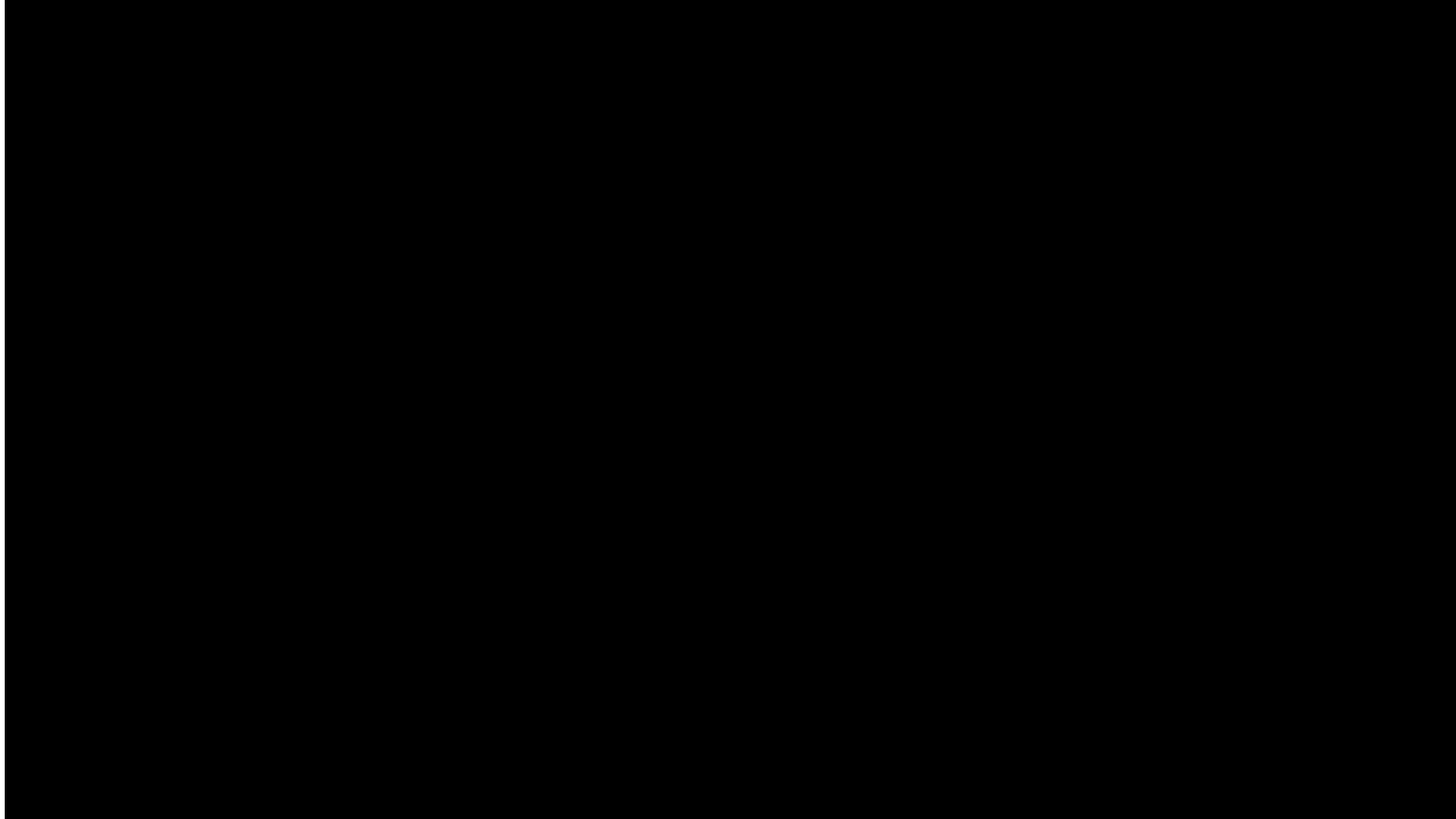
Big Data 的沿革 (1/3)

- 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。

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。
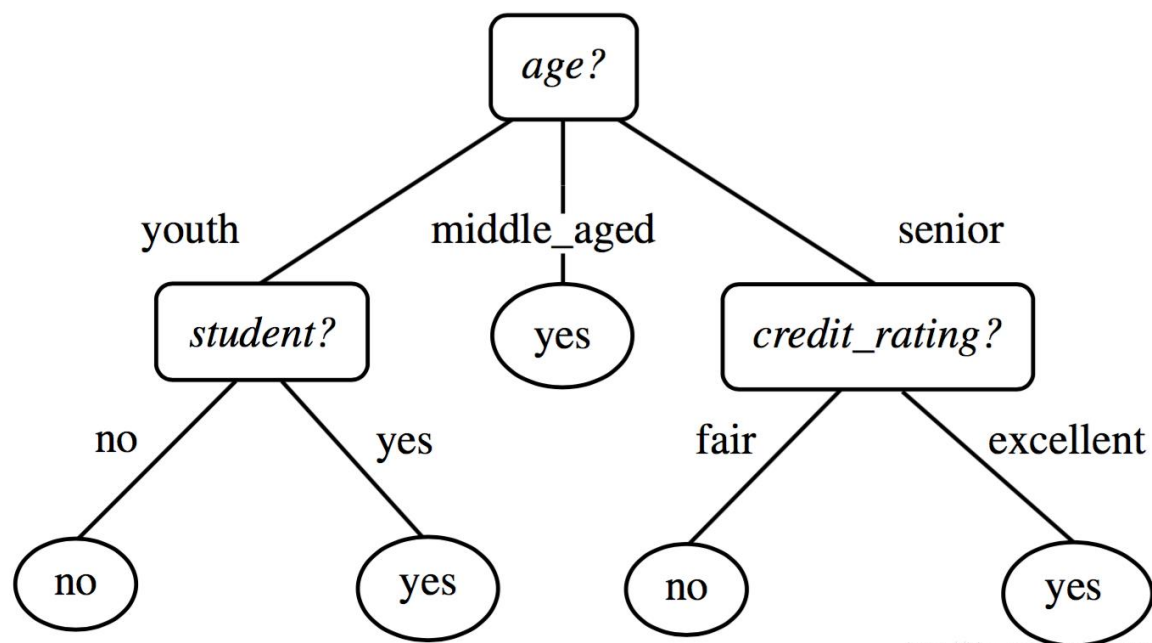
Data Mining

✓關聯式法則
- 有序性 (尿布與啤酒)
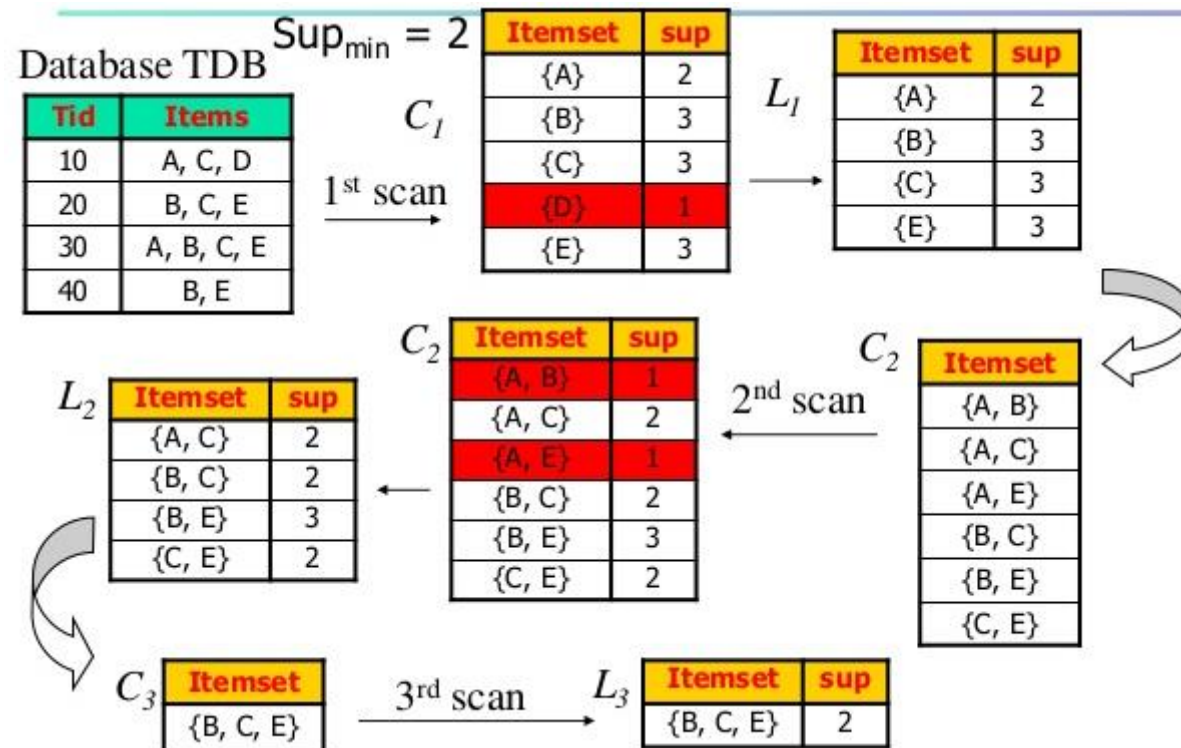
Big Data 的沿革 (2/3)

- 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
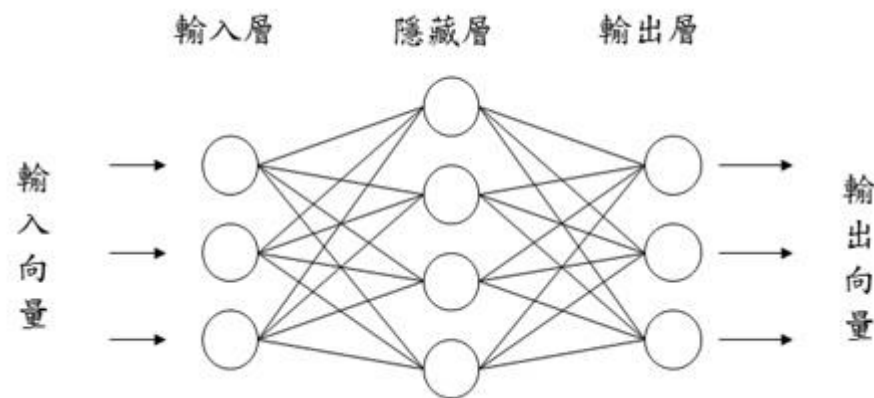
資料分析的基本步驟

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	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.30	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	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)

XY Plot of Words and Titles

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
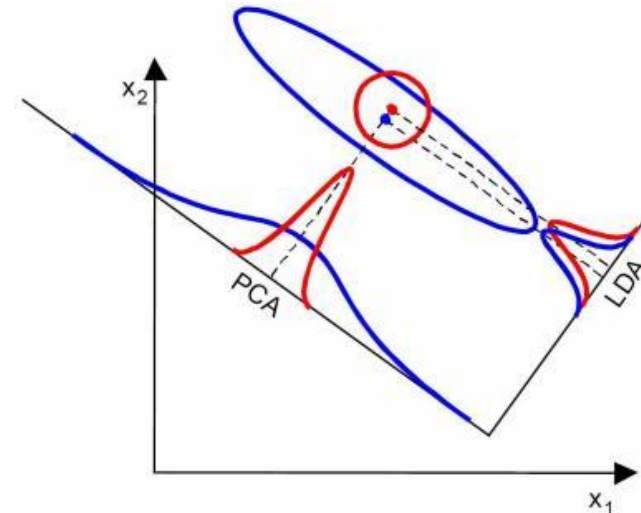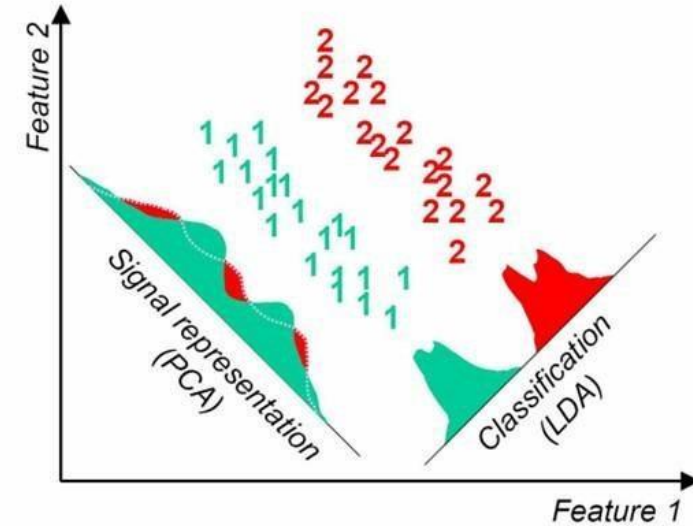
Thinking Time

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

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

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

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

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

$$pmi(x = information, y = data)$$

$$= log \frac{P(x = information, y = data)}{P(x = information) \times P(y = data)}$$

$$= log 1.49$$

$$= 0.57$$

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 。/PU

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/

40

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

- 資料來源
  1. 2K Games
  2. NBA

以NBA的應用為例

- NBA 2K17能力預測
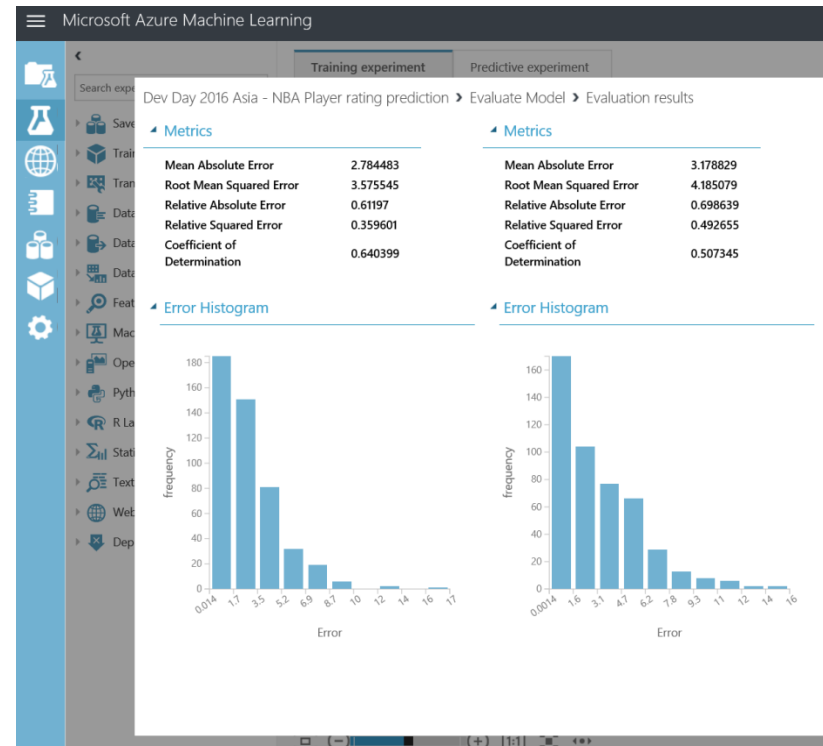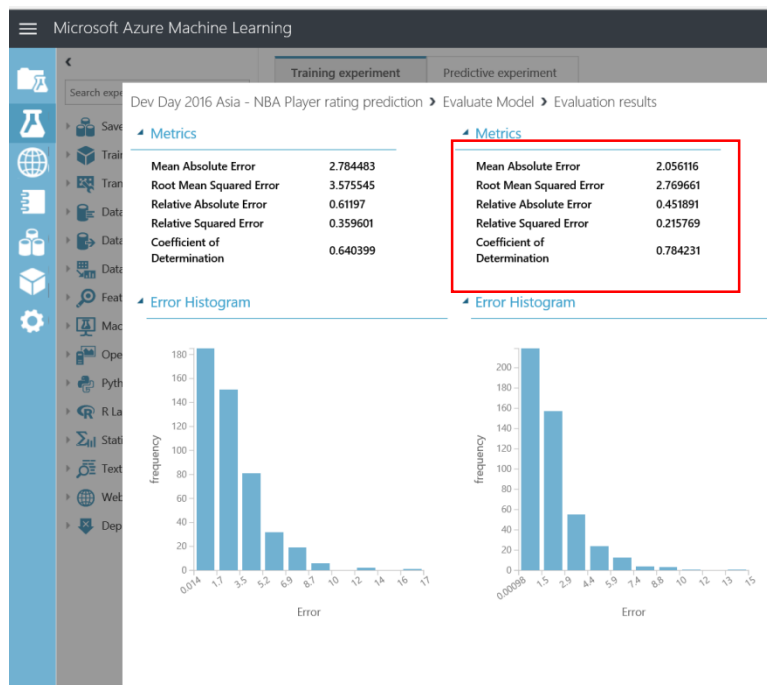
NBA季後賽預測
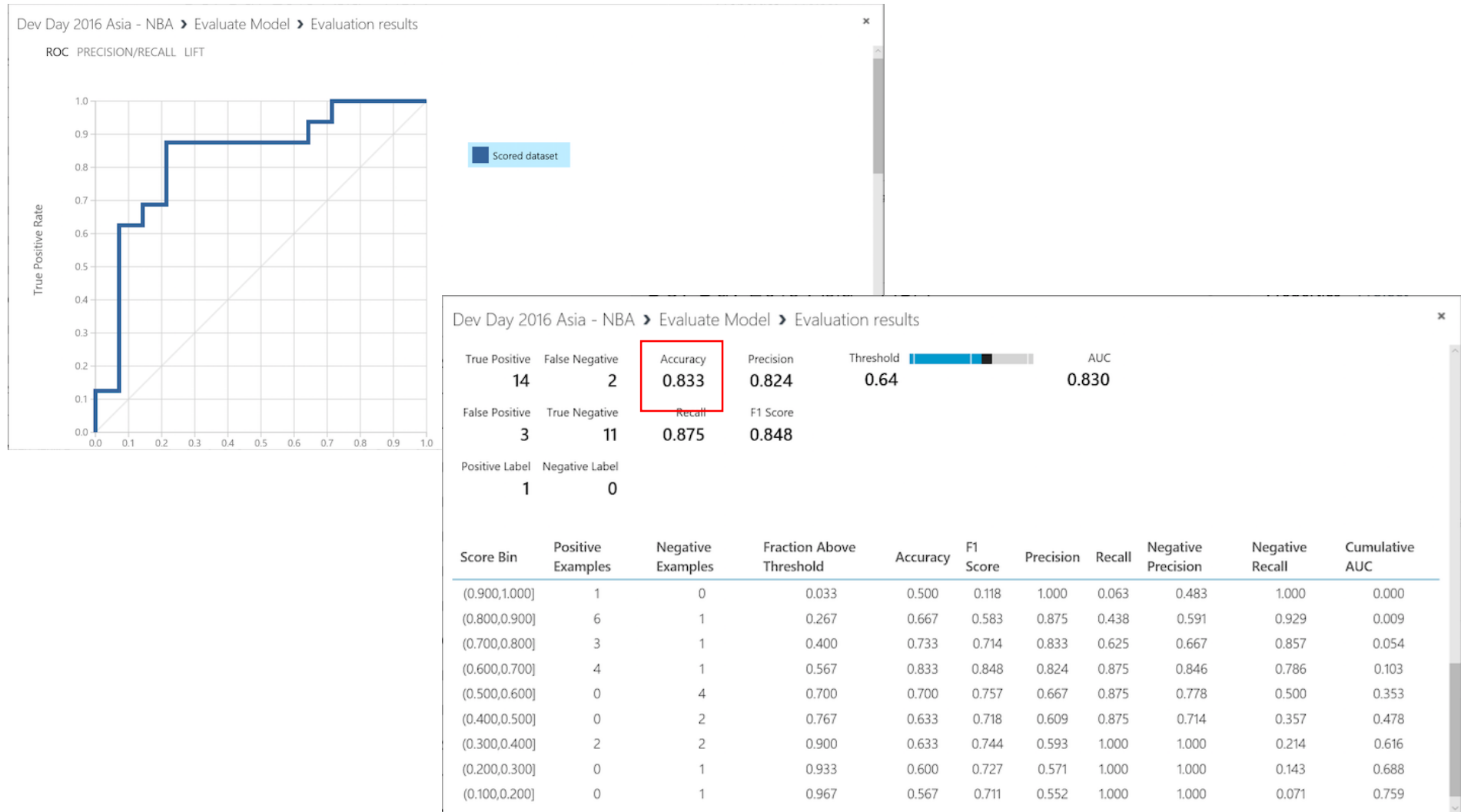
資料驅動創新應用

- 文字、聲音、影像
  - 自然語言處理
  - 語音辨識
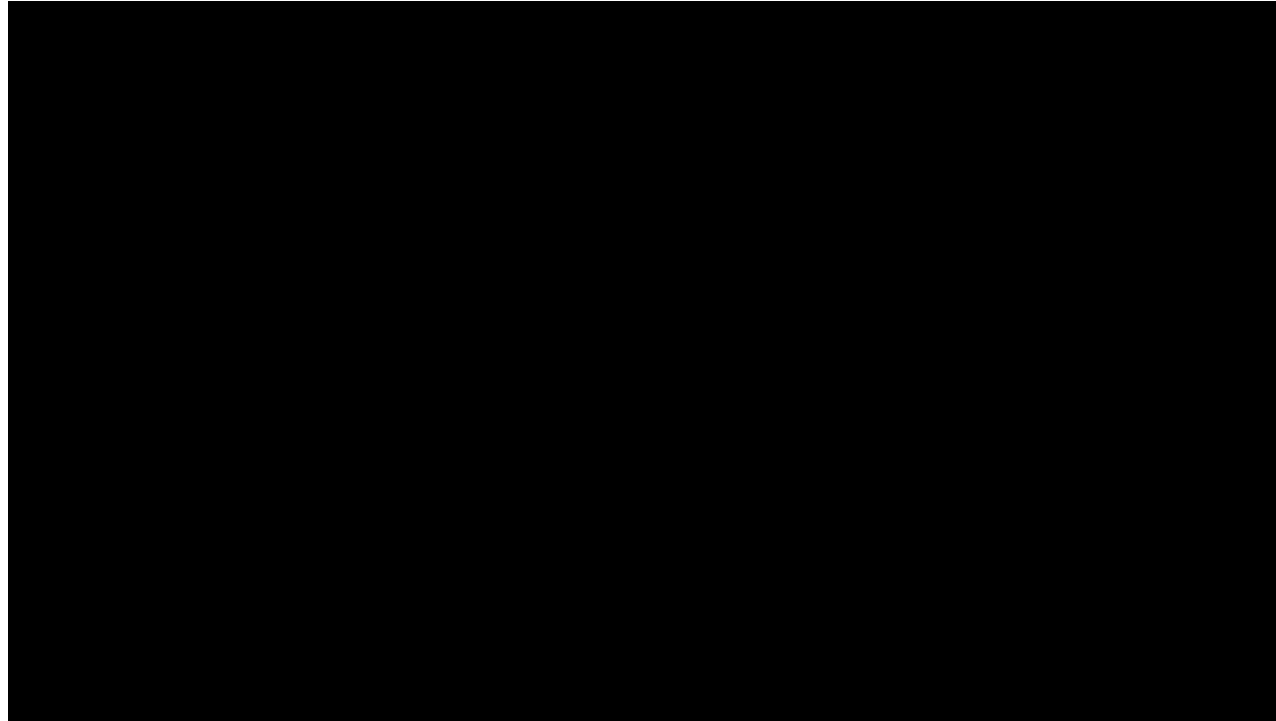 - 影像辨識

- 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

49

工業 4.0

Fintech

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

Open Datasets

UC Irvine Machine Learning Repository

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

Kaggle DataSets

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

臺南市開放資料

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

Thank you