WORLD STATISTICS DAY
20.10.2020
CONNECTING THE WORLD WITH DATA WE CAN TRUST

數據的真真假假
Data Veracity

楊良河教授
香港教育大學數學與資訊科技學系

主辦機構

香港教育大學 The Education University of Hong Kong

香港統計學會 Hong Kong Statistical Society

政府統計處 Census and Statistics Department

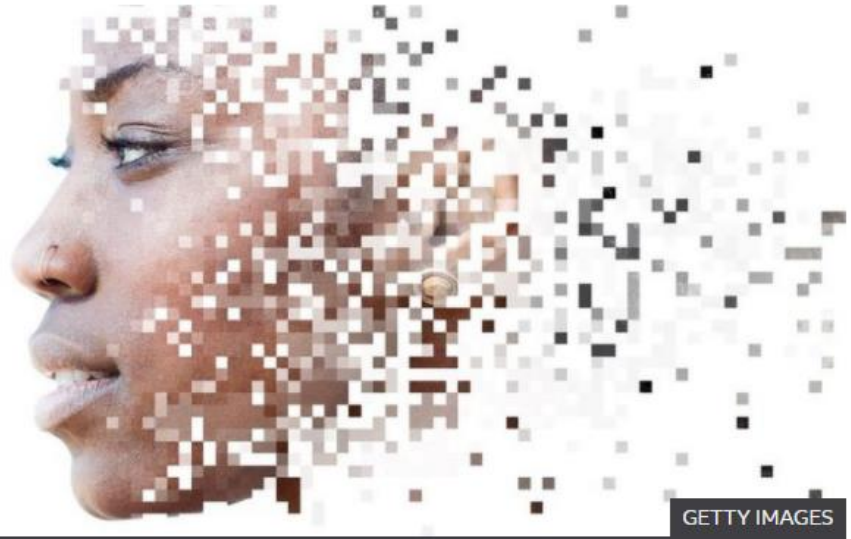
教育局 Education Bureau

https://www.bbc.com/news/technology-49961089

Deepfake videos 'double in nine months'

Rory Cellan-Jones
Technology correspondent
@BBCRoryCJ

7 October 2019



GETTY IMAGES

Deepfake videos typically involve computer-generated images of a subject's face created via analysis of thousands of still images of the person

Everybody Can Make Deepfakes Now!

Go to **www.menti.com** and use the code **27 46 30 8**

Mentimeter

新西蘭(New Zealand)是在澳洲(Australia)的那一面?

東北	東	東南
0	0	0

Result

BIG DATA

THE AGE OF

Big Data Characteristics 大數據特徵

The size of the data

Volume (大量)

Velocity (快速)

The speed at which the data is generated

The different types of data

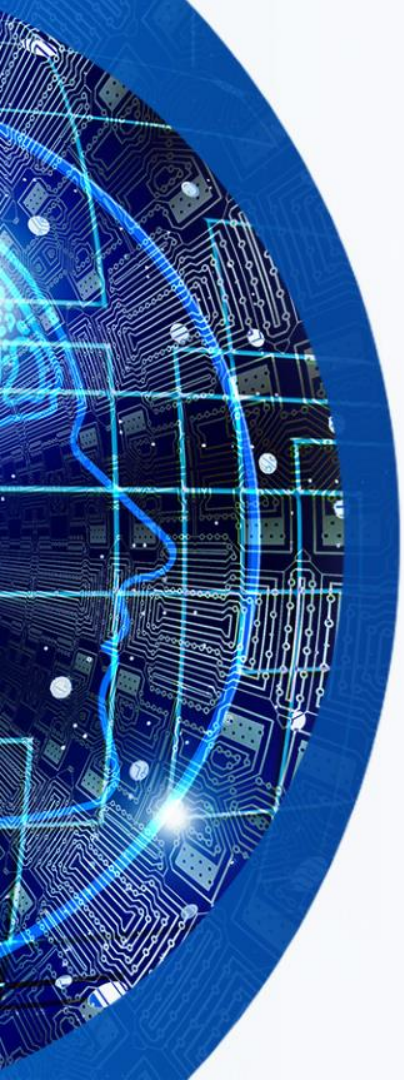
Variety (多樣)

Veracity (真確)

The trustworthiness of the data in terms of accuracy

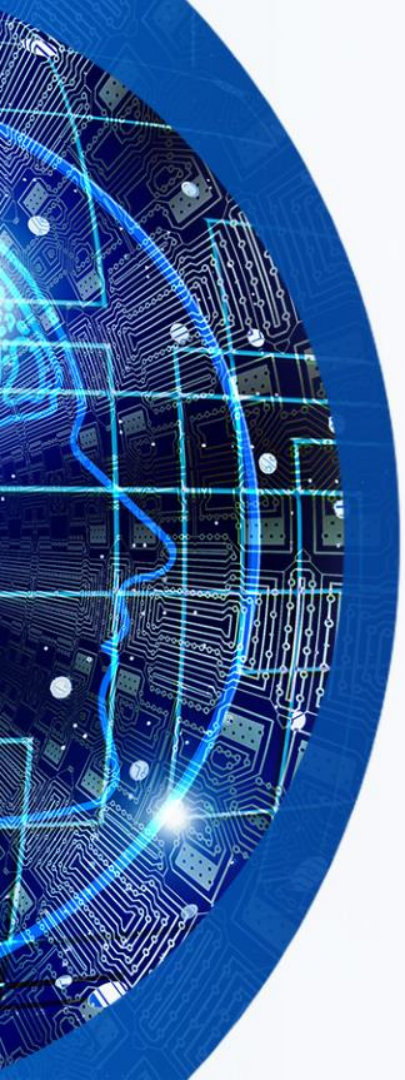
Just having big data is of no use unless we can turn it into value

Value (價值)

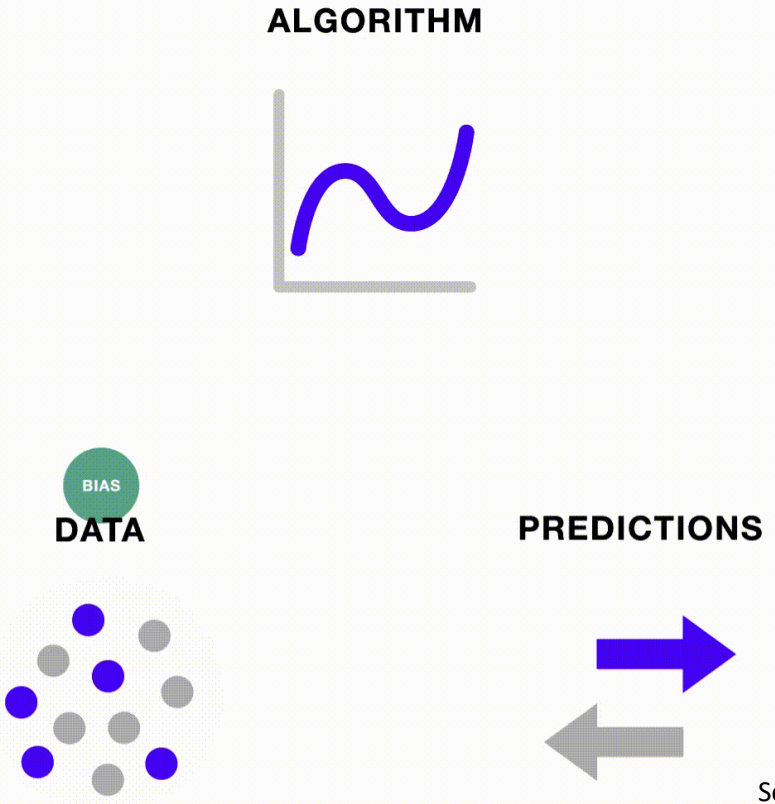


Data Veracity 數據真確性

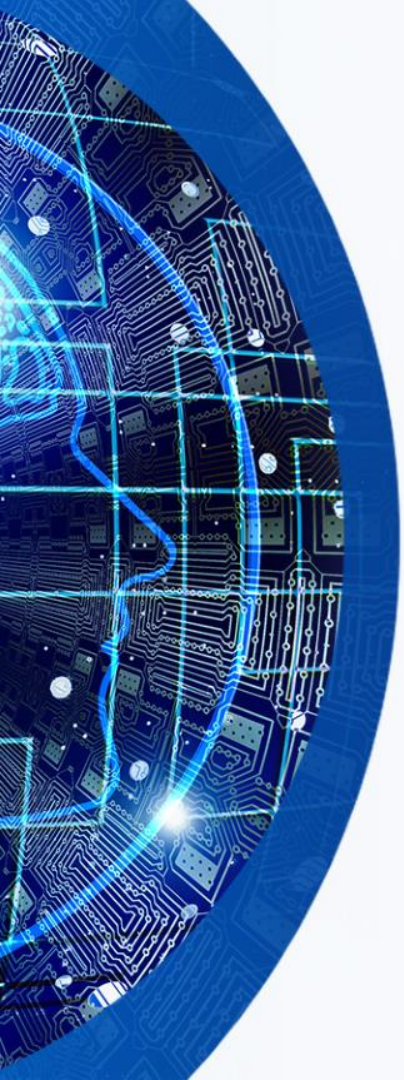
- 數據真確性是指數據的準確性(accuracy)或真實性(truthfulness)的程度。
- 在大數據年代，不僅只重視數據質量(data quality)，並且關注數據的來源、種類和處理是否可信(trustworthy)。
- 人們總是宣稱需要更準確和可靠的數據，但是為了獲得更大量和低廉的數據，人們常常忽略了數據真確性這一需求。



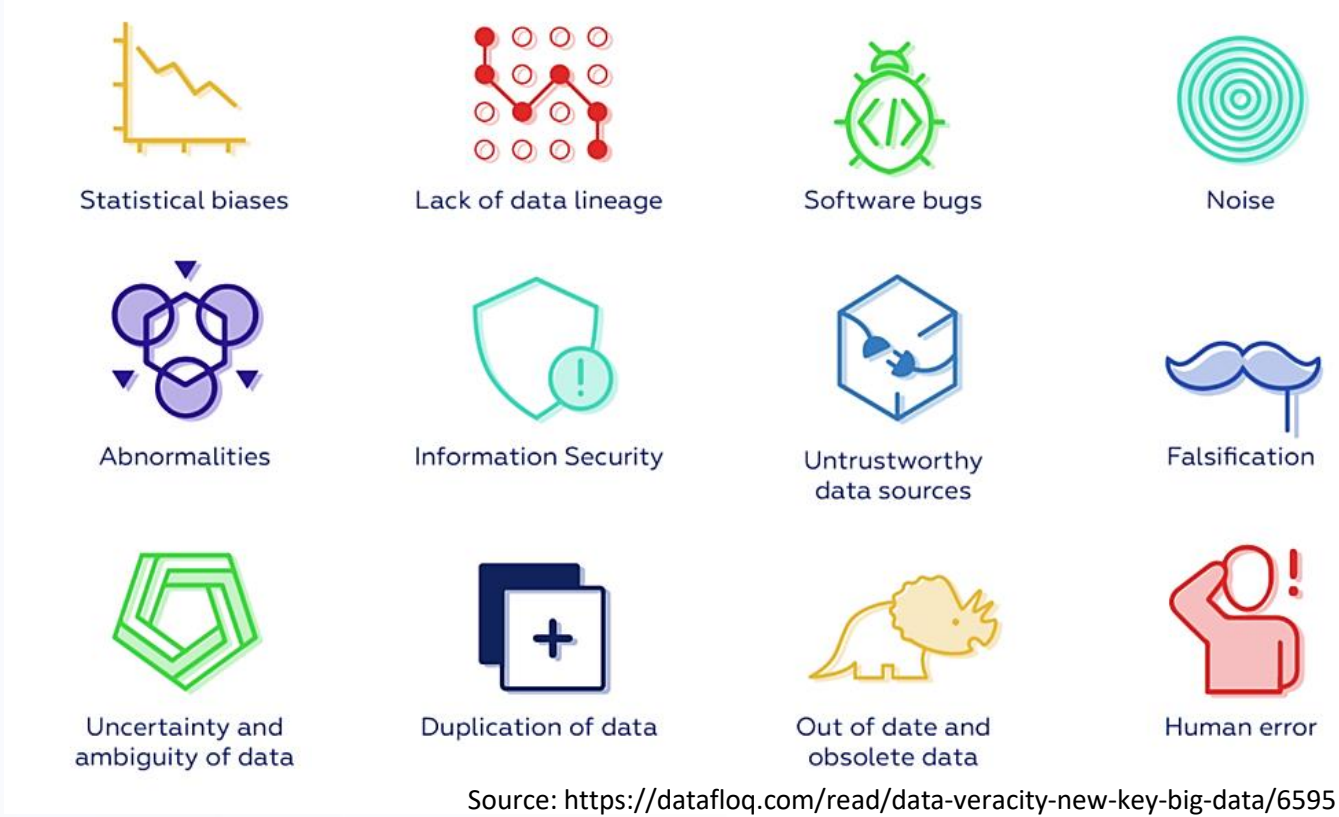
Garbage In, Garbage Out



Source: https://parametric.press/issue-01/the-myth-of-the-impartial-machine/



數據偏差的來源



Source: https://datafloq.com/read/data-veracity-new-key-big-data/6595

Statistical Bias (統計偏差)

Source: http://whatworks.site/statistics-bias-flaws/

煤氣 Towngas

此賬單按照估計讀數計算
請即使用右列報錶方法報讀度數

客戶號碼：

自動轉賬 $

到期日：2020年10月14日

賬單資料　發單日期：2020年10月 3日

上期賬單總數	$ 440.78	$	
繳費:2020年 8月13日	-440.00		上期餘數
上期賬單餘數	0.78	0.78	

由 2020年 8月 2日 至 2020年 9月 2日	1 度 =48 兆焦耳	煤氣用量 (兆焦耳)
4451 估計讀數 4468 估計讀數	17 度 x 48	816
9月2日的估計讀數 4468 是根據8月10日的報錶讀數 4455 而估算。		

報錶方法

- 24小時報錶熱線 ☎2880 5522
- 網站:www.towngas.com
- 手機應用程式「Towngas 煤氣公司」

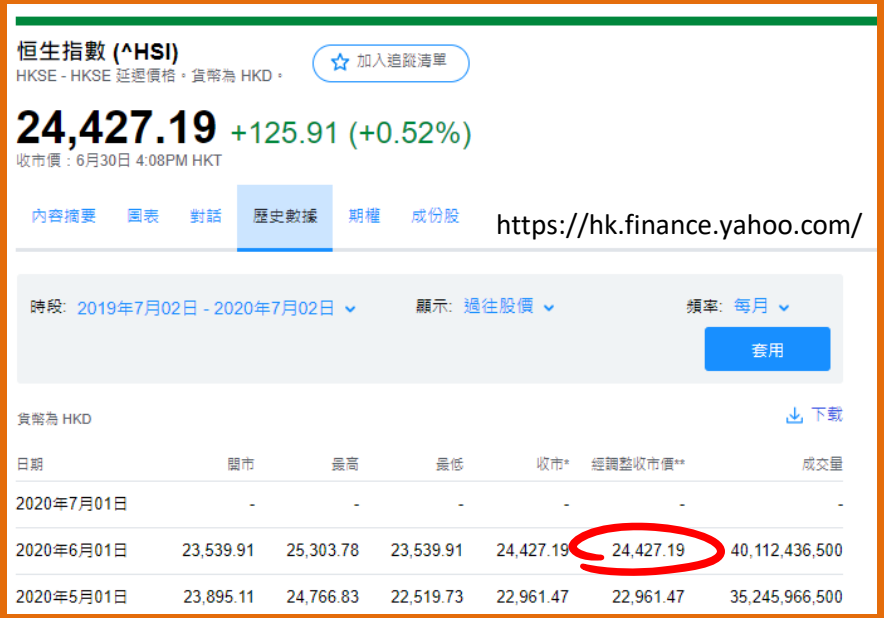
Lack of Data Lineage (缺乏數據沿襲)

數據沿襲即是元數據(Metadata)，用於解釋數據來源及如何產生，是數據管理其中一個重要概念。

例如：

- 確保客戶輸入數據，沒有遺漏並且準確
- 當客戶更新了個人資料，確保數據庫的數據能作出更新，保持客戶資料準確。

Software Bugs (軟件錯誤)



恒生指數 (^HSI)
HKSE - HKSE 延遲價格。貨幣為 HKD。
加入追蹤清單

24,427.19 +125.91 (+0.52%)
收市價：6月30日 4:08PM HKT

內容摘要 圖表 對話 歷史數據 期權 成份股

https://hk.finance.yahoo.com/

時段: 2019年7月02日 - 2020年7月02日 顯示: 過往股價 頻率: 每月
套用

貨幣為 HKD 下載

日期	開市	最高	最低	收市*	經調整收市價**	成交量
2020年7月01日	-	-	-	-	-	-
2020年6月01日	23,539.91	25,303.78	23,539.91	24,427.19	24,427.19	40,112,436,500
2020年5月01日	23,895.11	24,766.83	22,519.73	22,961.47	22,961.47	35,245,966,500

何為每日恆生指數？
何為每月恆生指數？

時段: 2019年7月02日 - 2020年7月02日 顯示: 過往股價 頻率: 每天
套用

貨幣為 HKD 下載

日期	開市	最高	最低	收市*	經調整收市價**	成交量
2020年6月30日	24,468.24	24,598.61	24,271.77	24,427.19	24,427.19	1,892,323,300
2020年6月29日	24,627.64	24,627.64	24,148.62	24,301.28	24,301.28	1,835,855,400
2020年6月26日	24,697.98	24,756.58	24,537.77	24,549.99	24,549.99	1,509,260,100

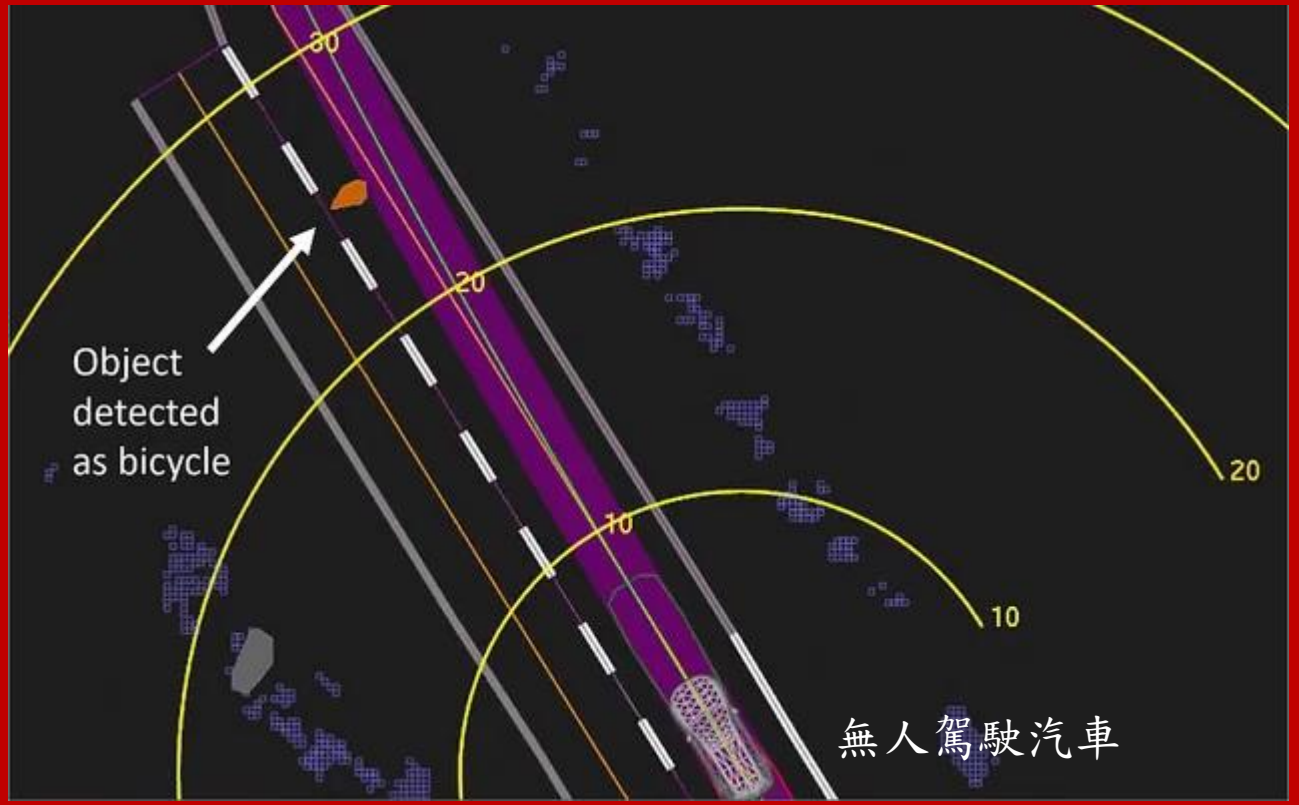
Logic error
(邏輯錯誤)

Testing
(測試)

Open source
software
(開源軟件)

Noise (噪音)

若這是被風吹起的塑膠袋，無人駕駛汽車需要確定這是否危險的障礙物，決定緊急煞車。

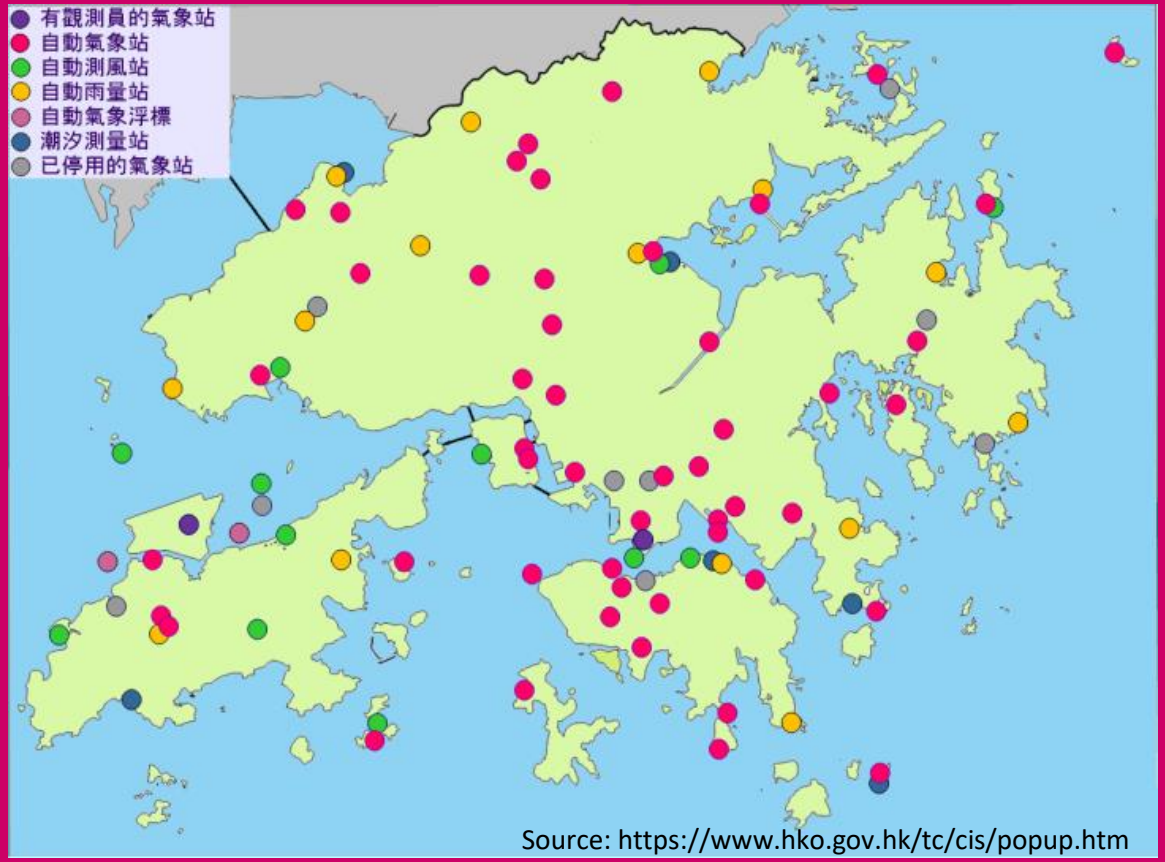


Source: https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/

Noise (噪音)

Source: https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/

Abnormalities (異常現象)



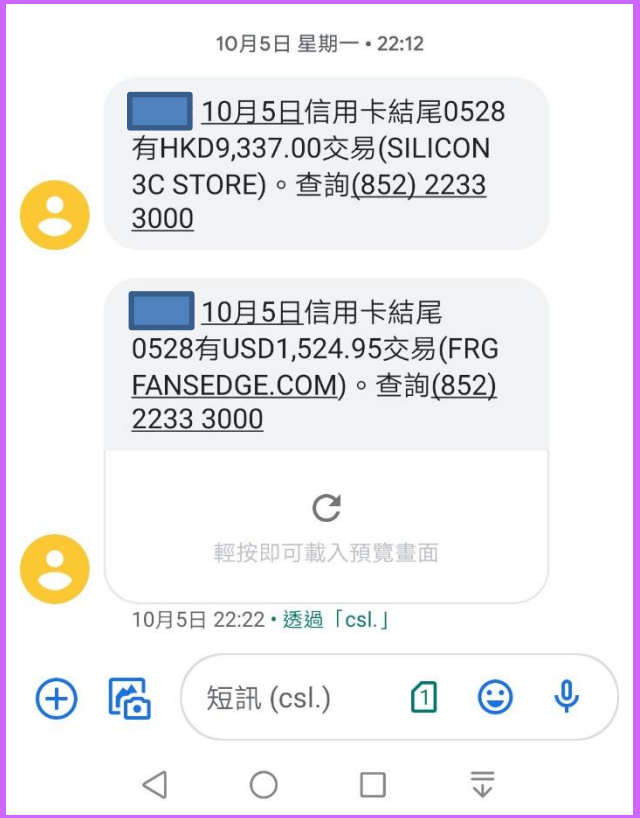
Source: https://www.hko.gov.hk/tc/cis/popup.htm

例如：

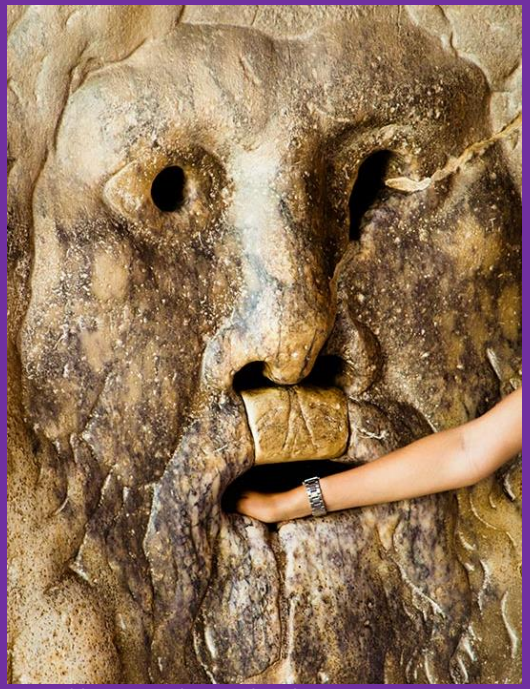
兩個相近的氣象站中，儀器所收集的氣象數據不應截然不同。

Information Security (信息安全)

Source: https://www.afitc-event.com/cyber-security-a-wicked-problem/



Untrustworthy Data Sources
(不完全可靠的數據源)



Source: https://hbr.org/2015/10/can-your-data-be-trusted

2020年DSE預測等級研究結果

科目	預測等級與文憑試實際等級相同（%）	預測等級比文憑試實際等級高 / 低 1 個等級 （%）	兩個等級差距在一個等級範圍內 （%）
英文	70.6	28.6	99.2
中文	56.2	40	96.2
數學必修	54.3	41.4	95.7
通識教育	53.9	42.1	96
數學延伸部分微積分與統計*	30.7	44.6	75.3
數學延伸部分代數與微積分	35.1	45.9	81
中國歷史	41.8	45.9	87.7
歷史	42.5	46	88.5
中國文學 *	39.4	49.1	88.5

資料來源：考評局

* 參與考生人數較少的科目

Falsification (偽造數據)

2008年金融海嘯期間，坊間傳聞東亞銀行財政出現危機及即將倒閉，總行及多間分行一度出現擠提。

東亞銀行當時舉行記者會闢謠，時任財政司長曾俊華、金管局總裁任志剛也分別會見記者，指有關傳聞毫無根據，事件隨後平息。

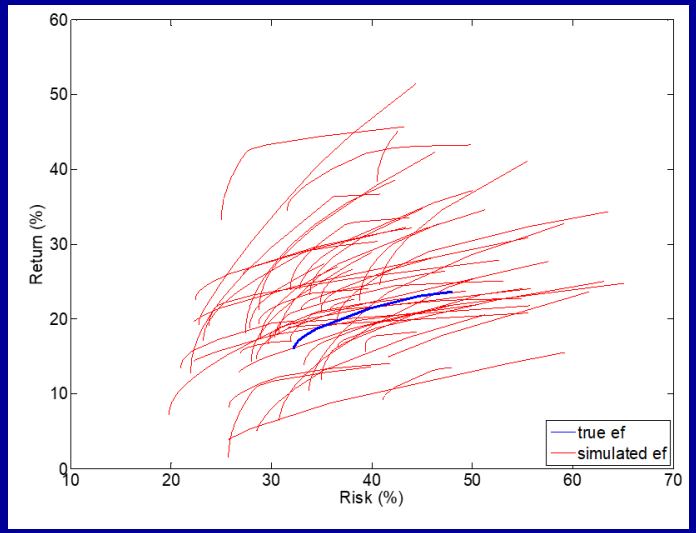
Source: https://businessfocus.io/article/111347

Falsification (偽造數據)

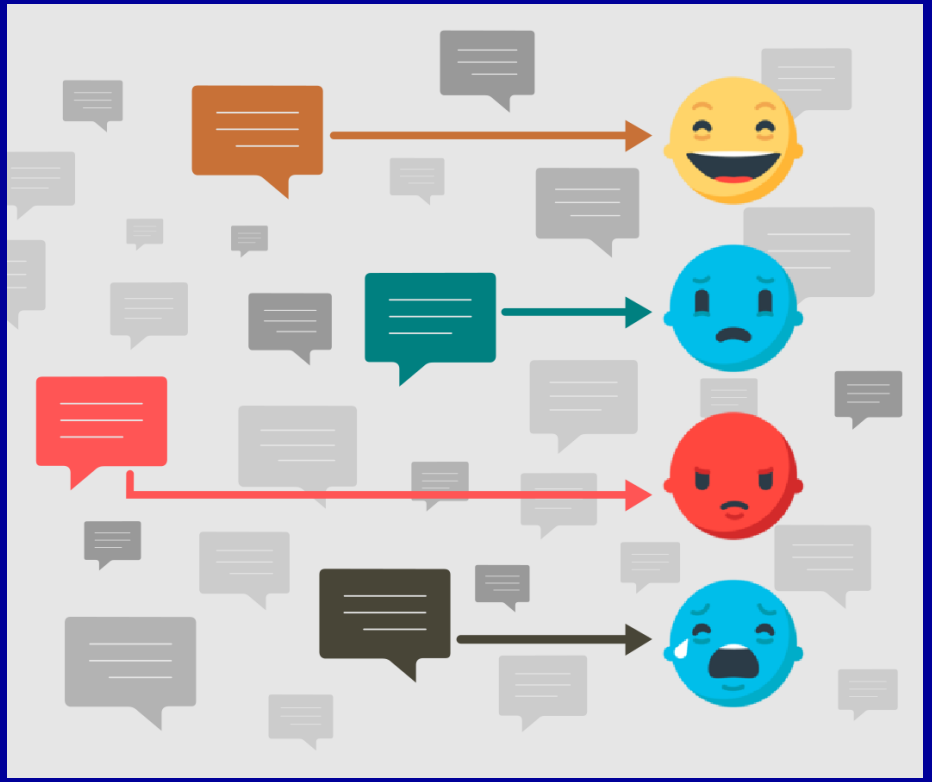
Uncertainty of Data (數據的不確定性)

圖表4.17　利用金磚四國十年數據估計出的投資參數

MSCI國家指數	年度化回報率	風險	相關系數			
			巴西	俄羅斯	印度	中國
巴西	23.6%	48.0%	1	0.60	0.62	0.62
俄羅斯	20.5%	46.7%	0.60	1	0.49	0.49
印度	16.8%	38.2%	0.62	0.49	1	0.63
中國	14.3%	34.9%	0.62	0.49	0.63	1

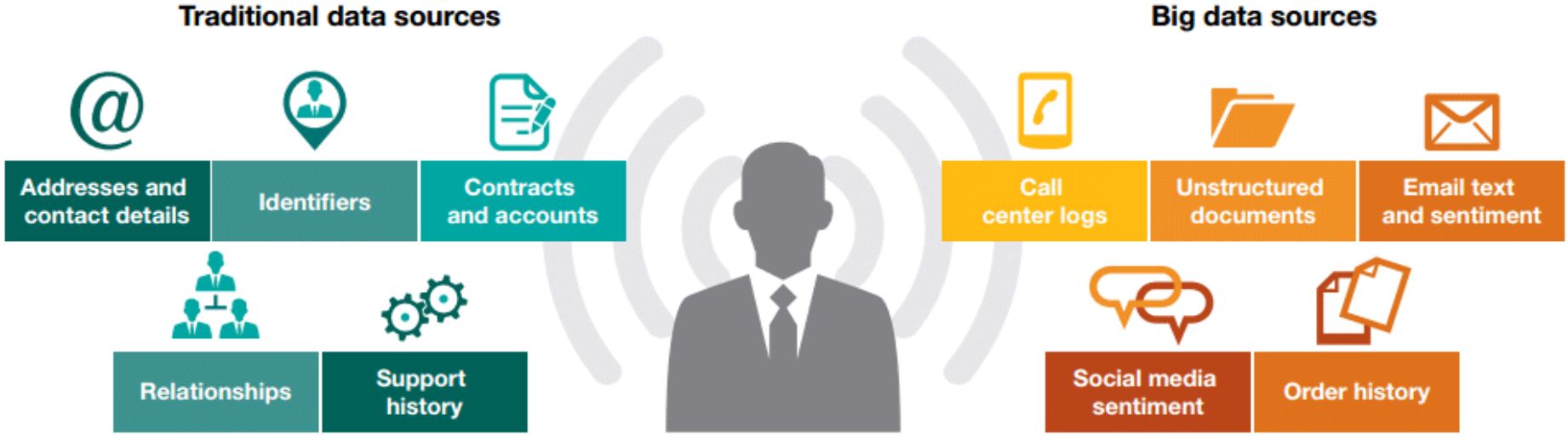


Reference: Chap4.8 尋找金磚四國的投資組合, 林建、楊良河 (2013). 計出你的投資勝算. 天窗出版(信報系列)



Source: https://medium.com/neuronio/from-sentiment-analysis-to-emotion-recognition-a-nlp-story-bcc9d6ff61aecccc

Duplication of Data (重複數據)



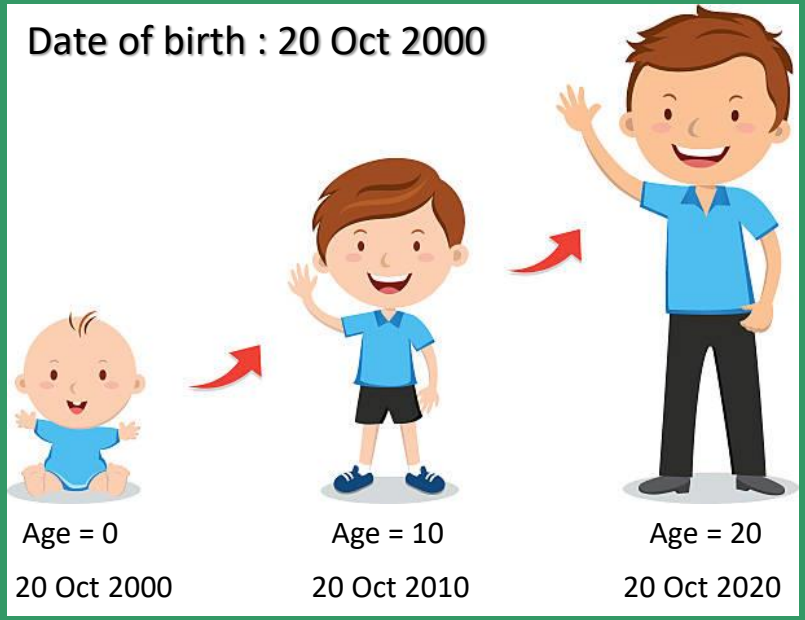
Source: https://www.sparklane-group.com/en/2017/09/13/without-big-data-there-is-no-360-view-of-the-customer/

Philip Yu vs PLH Yu vs 楊良河

News posted online or in blogs

Out-of-Date Data (過時數據)

DoB vs Age



Source: https://www.istockphoto.com/

Data from a questionnaire

Risk Profile Questionnaire 投資風險取向問卷

10) Which level of investment fluctuation per year would you be comfortable with? 閣下每年可承受投資波動的水平?

- ☐ Fluctuates between -5% and +5% (this option indicates that you will only consider funds that expected to have very low volatility such as money market funds) 於-5% 至 +5%之間的波動（此選擇表示閣下僅考慮投資於非常低波幅的基金如貨幣市場基金）
- ☐ Fluctuates between -10% and +10% 於-10% 至 +10%之間的波動
- ☐ Fluctuates between -20% and +20% 於-20% 至 +20%之間的波動
- ☐ Fluctuates between -30% and +30% 於-30% 至 +30%之間的波動
- ☐ Fluctuates below -30% and over +30% 低於-30%及超過+30%的波動

11) How soon would you expect your savings / reserves be used up if you do not sell the investment products held with us? 假如閣下不出售持有的施羅德投資產品，儲蓄/儲備預料在多久會用完？

- ☐ Within 6 months 6 個月以內
- ☐ 7 months – 1 year 7 個月 – 1 年
- ☐ 1 – 2 years 年
- ☐ 2 – 5 years 年
- ☐ More than 5 years 多於 5 年

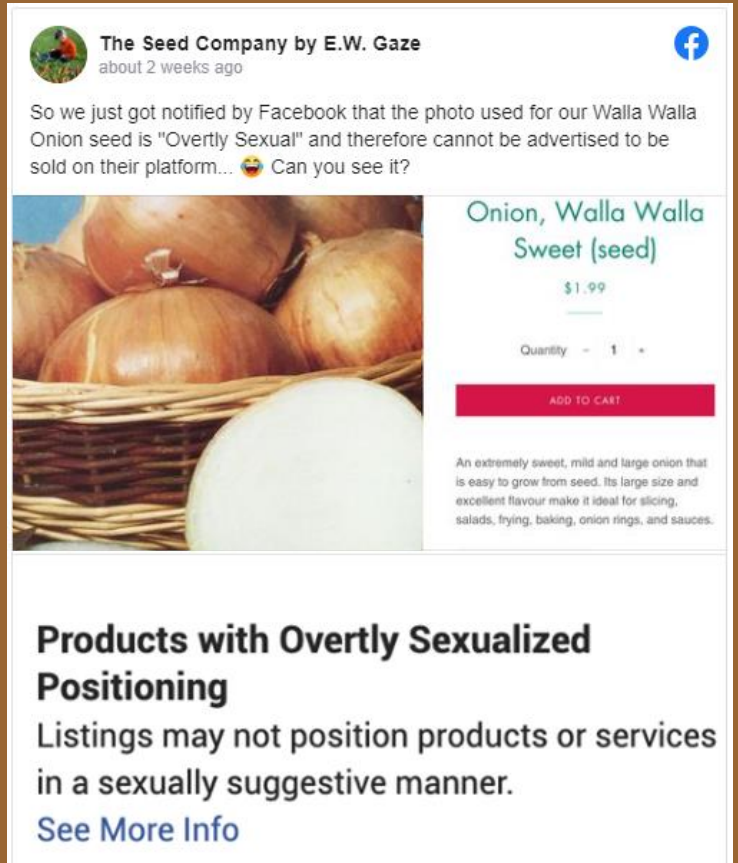
Customer Signature 客戶簽署[1]

_____

Date 日期: _____

Source: https://www.schroders.com/en/sysglobalassets/digital/hong-kong/pdfs/risk-profile-questionnaire.pdf

Human / Model Error (人為/模型錯誤)



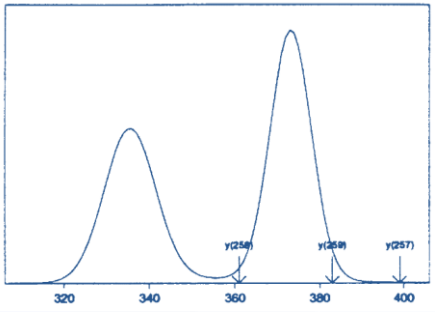
Facebook 一直以來均採用自動 AI 系統來檢測所有用家上載的圖片與影片，並將不符合 Facebook 準則的帖子刪除。

今年九月就有一間在加拿大的園藝中心在 Facebook 上載一張洋蔥照片並下了廣告以宣傳其出售的洋蔥種子，但卻被 Facebook 誤判為成人內容，故刪除該照片。

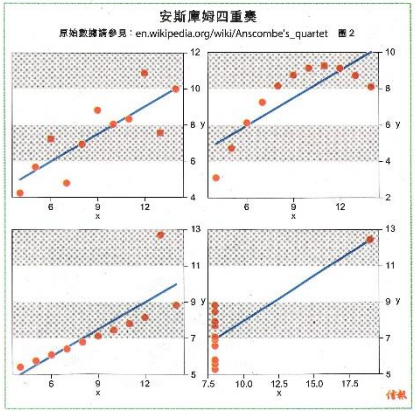
傳統分辨數據真假的方法

- Summarization: 頻率表、統計摘要、假設檢驗
- Visualization: 圖形如直方圖、框線圖、散點圖等，以檢查數據分佈和變量之間的相關性
- Anomaly Detection: 主成分分析、聚類分析等

IBM 某日收市價的預測分佈



Source: Wong, C.S. and Li, W.K. (2000). On a mixture autoregressive model. JRSSB, 62(1), 95-115.



李一鳴、楊良河 2014年6月12日信報《數裏見真章》

先進分辨數據真假的方法

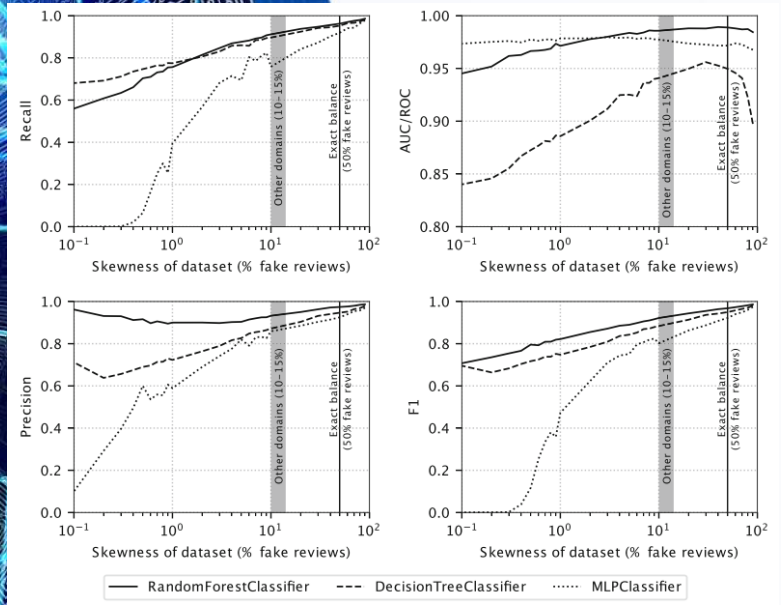
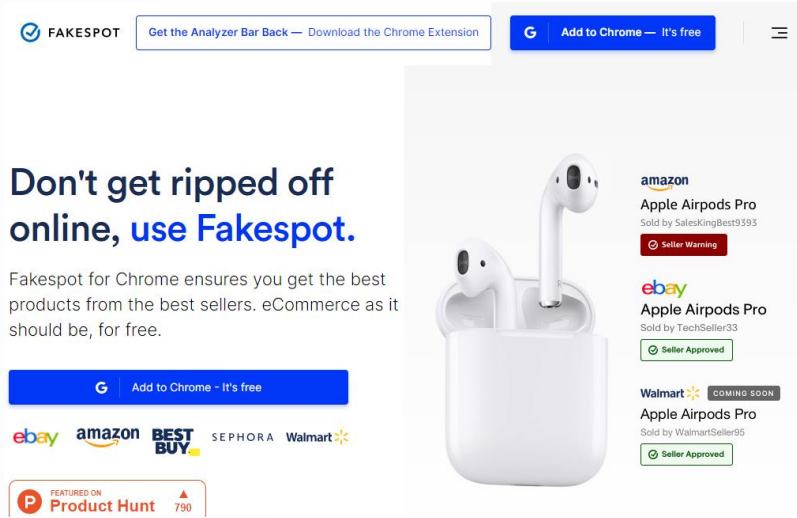


Fig. 17 Classification scores of appropriate machine learning algorithms for datasets with class imbalance, i.e., including 90% to 0.1% fake reviews, plotted on a logarithmic scale

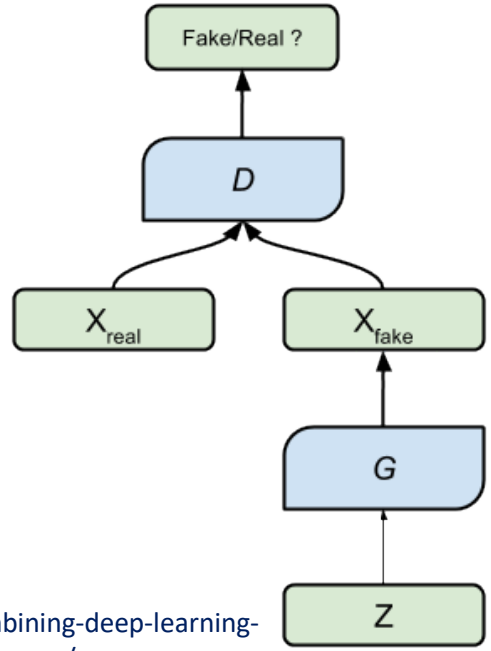
Source: Martens, D. and Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. Empirical Software Engineering, 24, 3316-3355.

- 「知己知彼，百戰百勝」
- 機器/深度學習 (ML/DL)
- 虛假評論檢測 (Fake Review Detection)



先進分辨數據真假的方法

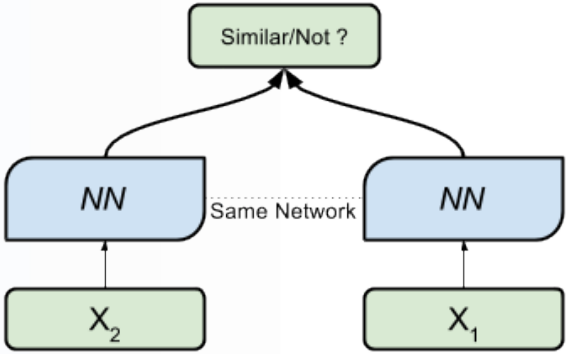
- 生成對抗網絡 (GAN)
  - GAN訓練兩個不同的網絡，一個針對另一個網絡，因此具有對抗性。
  - G網絡：通過真實圖像並對其進行盡可能多的修改來生成圖像（或文本或語音）。
  - D網絡：另一個網絡試圖預測圖像是“假”還是“真實”。



Source: https://aws.amazon.com/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images/

先進分辨數據真假的方法

- Siamese Network (SN)
  - SN是一起工作的兩個網絡。
  - 但是，與GAN在網絡競爭的情況下不同，這兩個網絡是相同的，並且彼此並存。 他們在兩個不同的輸入上比較網絡的輸出，並衡量它們的相似性。
  - 例如，SN可以驗證兩個簽名是由同一個人寫的。



The **Distance Function** decides if the output vectors are close enough to be similar

The **Neural Network** transforms the input into a properties vector

Input Data (image, text, features…)

Source: https://aws.amazon.com/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images/

Other Data Issues

- 缺失數據 (Missing data)
- 私隱問題 (Privacy issues)
- 歧視偏見 (Discrimination bias)

Gender bias of word embedding:
woman + king - man ≈ queen can do a magic but do you like the following?
woman + doctor - man ≈ nurse

Source: https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

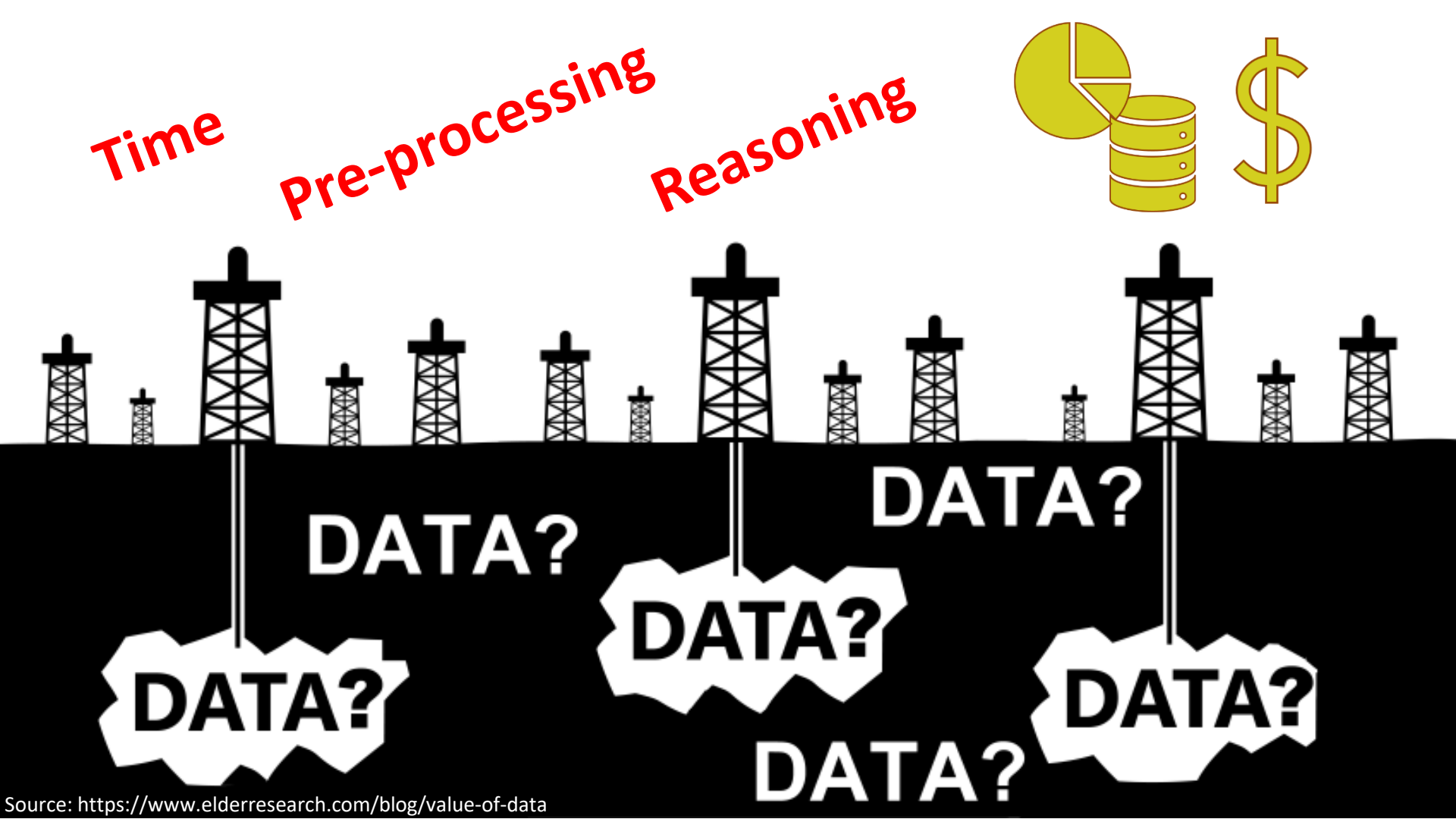
- 特徵提取/選擇 (Feature extraction / selection)

Six Blind Men and An Elephant

- An Indian story
- a group of blind men touch an elephant to learn what it is like.
  - The blind man who feels a leg says the elephant is like a pillar;
  - the one who feels the tail says the elephant is like a rope;
  - the one who feels the trunk says the elephant is like a tree branch;
  - the one who feels the ear says the elephant is like a hand fan;
  - the one who feels the belly says the elephant is like a wall; and
  - the one who feels the tusk says the elephant is like a solid pipe.

- A king explains to them:
  - All of you are right. The reason every one of you is telling it differently is because each one of you touched the different part of the elephant. So, actually the elephant has all the features you mentioned.

Source: http://en.wikipedia.org/wiki/Blind_men_and_an_elephant



Source: https://www.elderresearch.com/blog/value-of-data

Thank You

楊良河 Philip Yu
plhyu@eduhk.hk