

統計諮詢簡介(1)

統計學、數據科學、人工智慧



吳漢銘 副教授

國立政治大學 統計學系

wuhm@g.nccu.edu.tw

<http://www.hmwu.idv.tw/>



大綱 (本簡報會持續更新)

統計諮詢簡介 (1)

- 統計學 (Statistics) **過去**
- 數據科學 (Data Science) **現在**
 - 第五波科技浪潮: Big Data
 - 數據科學家
- 人工智慧 (Artificial Intelligence) **未來 已來**

目的:

了解統計諮詢的範疇
以及我們擅長的領域

統計諮詢簡介 (2)

- 待續

本系「統計諮詢」課程

余清祥老師個人網站



通訊方式

研究室
商學院館十樓 261026室

email: csyue@nccu.edu.tw

聯絡電話

(02)2938-7695

(校內分機81026)

傳真號碼: (02)2939-8024

<http://csyue.nccu.edu.tw/ch/index.htm>

核心主題

數據科學

統計諮詢

統計方法

工具

112-1 統計諮詢

09/12/2023	第一週	課程大綱	「統計諮詢」投影片	美國2022年統計出路	統計諮詢作些什麼
09/14/2023		上課教室：9/19~10/17在商學院六樓寶來科技廳、10/24及10/31在六樓炎成創新廳、11月起在201教室。			
09/19/2023	第二週	「統計諮詢的角色」投影片	統計在大數據時代的角色	統計：消失或繼續演化	
09/26/2023	第三週	「定義問題」投影片	第一次作業題目	SOA PA資訊	統計是未來的熱門行業
10/03/2023	第四週	統計教學的省思	定義問題參考資料	從統計到資料科學家	三大步驟：人人都是分析師
10/10/2023	第五週	國慶日，停課一次。			
10/17/2023	第六週	Clayton諮詢講義	統計諮詢錄影帶	第二次作業題目	挑戰者太空梭講義
10/24/2023	第七週	統計諮詢角色參考講義	統計諮詢投影片	歷史教育的核心	
10/31/2023	第八週	樣本代表性投影片	直昇機製作講義	理查費曼的生平簡介	
11/07/2023	第九週	諮詢的專業建議	統計諮詢的考量	諮詢經驗分享	第三次作業題目
11/14/2023	第十週	探索性資料分析投影片	資料整理講義	圖形使用講義	第三次作業資料
11/21/2023	第十一週	資料分析的陷阱投影片	統計分析的限制(參考資料)	EDA	網路講義
11/28/2023	第十二週	統計諮詢期中報告(2023/12/8繳交)			
12/05/2023	第十三週	廣義線性模式投影片	過去50年最重要的統計學思想		
12/12/2023	第十四週	期末報告投影片	Purdue大學寫作講義	Wisconsin大學寫作講義	第四次作業題目
12/19/2023	第十五週	論文寫作講義	寫作講義	主題句寫作講義	摘要書寫參考資料
12/26/2023	第十六週	統計職業道德投影片	製作簡報講義	口頭報告參考資料	口頭報告講義
		說話的藝術			

時代變了!

余清祥老師「112-1統計諮詢」作業#1

Statistical Consulting Fall 2023
Assignment #1 10/12/2023 due

1. 大數據時代為數位化相關產業帶來新氣象，對資料分析的需求高於先前資料採礦(Data Mining)時代，各領域都需要統計專業的協助，但其中存在統計發展的潛在威脅。請各組參考書籍、期刊及各國統計學會網站，整理相關研究論文後提出建議，說明現代統計學家（或資料科學家）需要哪些訓練，以及如何區隔統計、資訊科學扮演的角色。（註：請詳細註明參考文獻及其出處。）

現代統計學家（或資料科學家）需要的訓練：

2. 上課 1. **數據倫理與隱私保護**：隨著數據收集和分析能力的增強，如何在保護個人隱私的同時進行數據分析變得極其重要。統計學家需要了解相關的倫理規範和法律法規。
2. **機器學習與人工智慧**：除了傳統的統計方法，現代的數據分析還涉及到機器學習和人工智慧技術。掌握這些技術可以幫助統計學家從大數據中提取更深入的見解。
3. **大數據技術**：了解和掌握處理大規模數據集的技術，如Hadoop和Spark等，對於在大數據環境下進行有效的數據分析至關重要。
4. **跨領域知識**：統計學家應該具備一定的跨領域知識，包括業務知識和其他相關領域的知識，以便更好地理解數據背後的實際問題。

統計學與資訊科學的角色區隔：

- 統計學：更專注於模型的建立和假設檢驗。
- 資訊科學：涵蓋數據採集、數據清洗、數據分析、數據可視化等方法進行預測和分類。

參考文獻：

由於這是一個示例回答，沒有直接引用具體的書籍、期刊或網站。在實際撰寫報告時，請參考以下來源：

- 相關學術期刊，如《統計學報》、《機器學習研究》等。
- 各國統計學會網站，如美國統計協會（ASA）、國際統計學會（ISI）等。
- 專業書籍，特別是那些關於大數據、數據倫理、機器學習和人工智慧的最新出版物。

在報告中提出建議時，應詳細註明每一條參考文獻的出處，包括作者、出版年份、書籍名稱、卷期和頁碼等信息。這樣不僅增強了報告的可信度，也方便讀者查找和驗證信息。

5. 以統計的角度說明虛無假設的設定，應該是「被告有罪」或「被告無罪」。（註：不考慮「無罪推定論」之類非統計思維的理由。）



您

大數據時代為數位化相關產業帶來新氣象，對資料分析的需求高於先前資料採礦(Data Mining)時代，各領域都需要統計專業的協助，但其中存在統計發展的潛在威脅。請各組參考書籍、期刊及各國統計學會網站，整理相關研究論文後提出建議，說明現代統計學家（或資料科學家）需要哪些訓練，以及如何區隔統計、資訊科學扮演的角色。（註：請詳細註明參考文獻及其出處。）



ChatGPT

在大數據時代，數位化相關產業的蓬勃發展對資料分析提出了前所未有的需求，這不僅推動了資料採礦(Data Mining)的進步，也促使統計學和資訊科學等領域的專業知識變得更加重要。然而，這樣的發展同時帶來了對統計學發展的潛在威脅，包括資料隱私、數據安全、誤用統計方法等問題。為了應對這些挑戰，現代統計學家（或資料科學家）需要接受特定的訓練，並清晰地界定統計學與資訊科學在數據分析中的角色。



您

以統計的角度說明虛無假設的設定，應該是「被告有罪」或「被告無罪」。（註：不考慮「無罪推定論」之類非統計思維的理由。）



ChatGPT

在統計學中，虛無假設（Null Hypothesis，通常表示為 H_0 ）是一種假設，用於表示沒有效果、沒有差異或沒有關聯的情況。在進行假設檢定時，虛無假設通常作為一種基準，與之對立的是替代假設（Alternative Hypothesis，表示為 H_1 或 H_a ），後者表示存在效果、差異或關聯。

在討論「被告有罪」或「被告無罪」的情境下，從統計的角度出發，虛無假設的設定應遵循能夠被量化檢驗的原則，且傾向於假設「無效果」或「無差異」的情況。因此，虛無假設應該設定為「被告無罪」（ H_0 ：被告無罪），這表示在沒有充分證據之前，我們假設被告的行為與罪行無關。

相對地，替代假設則會是「被告有罪」（ H_1 ：被告有罪），這代表如果數據分析或證據足以拒絕虛無假設，我們有理由相信被告與罪行有關聯。

這種設定方式符合統計學中假設檢定的一般原則，即先設定一個「無效果」的虛無假設，然後通過收集和分析數據來檢驗這一假設是否成立。如果數據顯示虛無假設下的情況極不可能發生，則我們有理由拒絕虛無假設，從而支持替代假設。在法律上，這相當於在證據足夠強大以推翻「無罪」的假設前，被告應被假定為無罪。

什麼是統計?

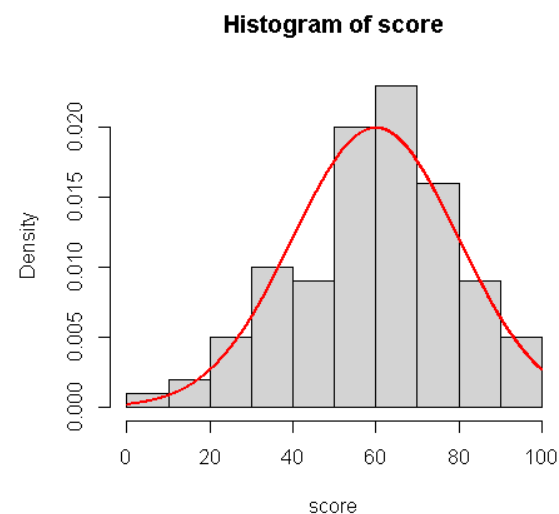
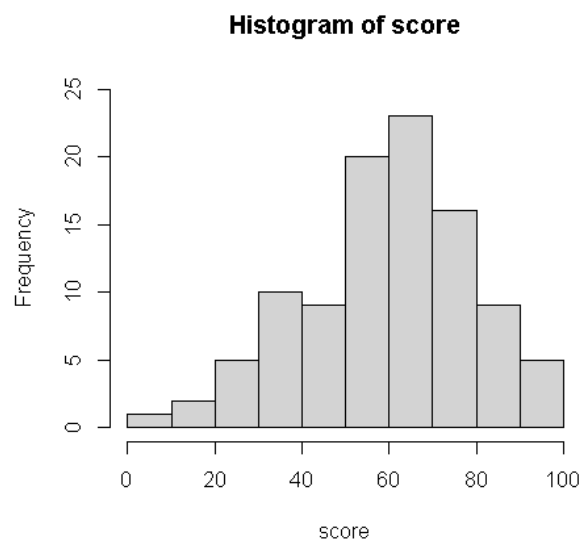


- **Merriam-Webster dictionary** defines statistics as "a branch of mathematics dealing with the **collection, analysis, interpretation, and presentation** of masses of numerical data."

- 傳統統計(歷史源自17世紀), 分兩類:
 - **敘述統計**: 對所收集到樣本的摘要結果。
 - **推論統計**: 考慮隨機性之下, 根據樣本的特性去推論母體的參數(例如: 估計母體平均數、推論母體的分佈)。

- 統計研究領域的分類: 數理統計、工業統計、商用統計、生物統計、社會統計、貝氏統計、空間統計等等。

<http://www.theurus.de/blog/some-truth-about-big-data/>



統計改變了世界 (1)



□十九世紀初:

「機械式宇宙」的哲學觀

□只要有一些**數學公式**

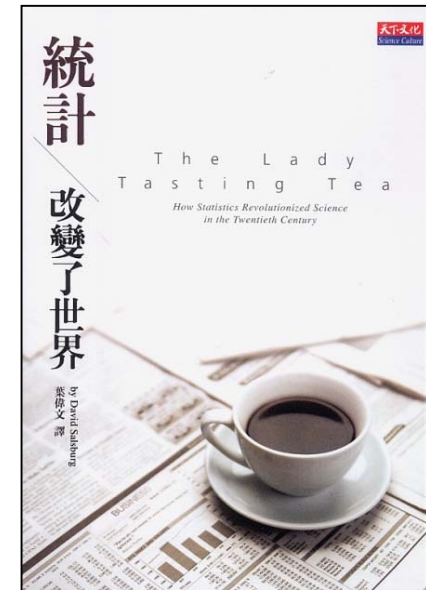
(e.g., 牛頓運動定律、波以耳定律)

就能了解世界所有**物理現象**並**預測**即將發生的事件。

(1) 拉普拉斯與拿破崙的對話。

(2) 1840年代，有人以牛頓的數學定律預測行星。

□預測與觀測：**誤差理論**(error theory)為統計的先驅。



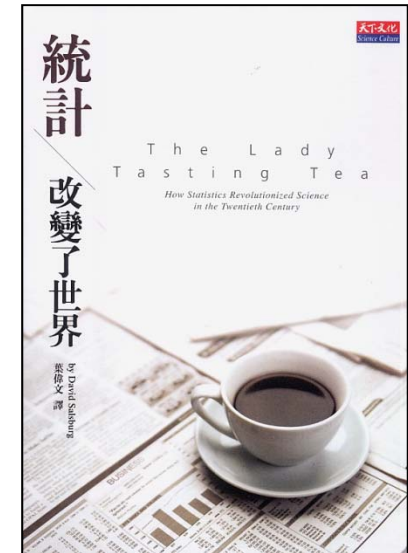
統計改變了世界 (2)



- 十九世紀末：
真實世界充滿了不確定性
 - 量測愈精確，反而產生更多誤差。
 - 尋找**生物學**定律，**社會學**定律徒勞無功。物理化學定律僅是粗略逼近，機械式宇宙觀開始動搖。
 - 科學家漸開始以新的典範來運作：現實世界的**統計模型**。

- 二十世紀：
科學界的**統計革命**。

- 二十一世紀：
幾乎所有的科學已經轉而運用**統計模式**了。



統計學的起源



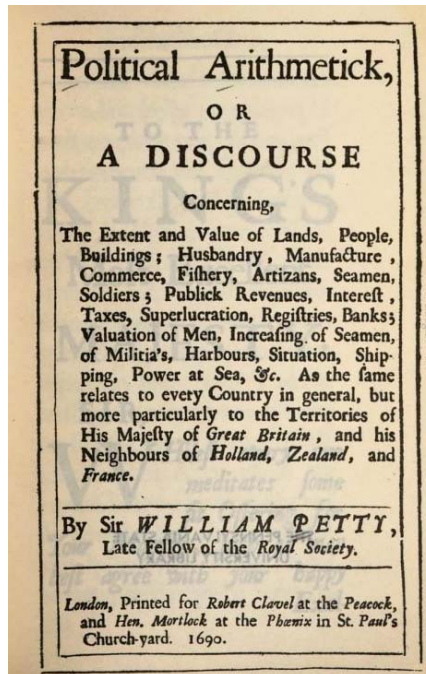
- **統計學** (statistics) 作為一門學科始於**19世紀**，但統計工作自古就有。
- **古埃及時期(公元前3千年)**，國王為修建金字塔徵收稅款，就曾對全國人口與財產做過**調查統計**。
- **古希臘(公元前8世紀)及羅馬時期(公元前27年)**，許多國家用**統計方法**進行人口調查和財產登記，並且從各國統計數字的差別研究各國的**政治經濟**情況，稱之為「**政治算術**」。
- **中國**在四千多年前的夏朝，《尚書》〈禹貢〉篇一文記載了當時的中國九州的各地物產、交通、植物特徵等統計資料，以及田地及貢賦分為九等的作法。

Statistics字源(1)



- **16世紀拉丁文 ”status”**：意指狀態或國會(statisticum collegium)，又代表對國家(state)各種現象或基本情況進行簡單的估量。
- **義大利文 ”stato”**：由status演變而來，為“國家概念，國家機構和國力的總稱”。
- **1749年德國” statistika ”**：哥廷根大學教授Gottfried Achenwall以“stato”為基礎，在「**歐洲各國國勢學綱要**」一書中首次使用德文“statistika”，正式命名為“**統計學**”，意思是“**國家應該注意的事實學問**”，包括國家的組織、人口、軍隊、國民職業和地上地下資源等。但當時“統計”多用文字表述，極少用數字。

Statistics字源(2)



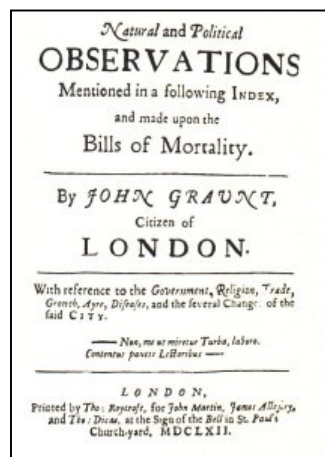
統計學在歷史上的雛形就是為統治者掌握國家的整體實力特別是經濟實力服務的。

- 英國數學家威廉配第(William Petty, 1623-1687) ，他在《政治算術》一書中，第一次提出運用統計和數學方法來研究政治、經濟、軍事現象。William Petty被馬克思譽為統計學的創始人。
- 1787年，英國齊默爾曼 (Dr.E.A.Zimmerman) 把德語statistika譯為**statistics**， “統計學” 才代替 “政治算術” 一詞。
- 1791年，英國一位長老教會牧師Sir John Sinclair編撰了廿一卷的《蘇格蘭統計報告》「 Statistical Account of Scotland」，將Statistika翻譯成英文**Statistics**，用以代表收集和分析數字的科學。

John Graunt (1620~1674)

最早的統計分析:

- 1662年英國經濟學家格蘭特John Graunt (24 April 1620 -18 April 1674)發表論文**分析死亡**的一些原因及其趨勢。編製了世界上第一個死亡表(即現在**生命表**的基礎)。是首次根據大量原始資料所作的實證研究，有「**現代統計學之父**」之稱。



(194)
The Number of the Weddings, Christenings, and Burials, that were in the Parish of Cranbrook, from March 26, 1560 to March 24, 1649; (as appears by the Register) only in the years 1574 and 1565 the Christenings are wholly omitted, because the Register is very imperfect for the greater part of those years.

Years	Weddings		Christened			Buried		
	M.	F.	M.	F.	Both	M.	F.	Both
1560	30	36	55	69	29	21	50	
61	34	46	53	79	23	32	55	
62	31	32	26	58	40	31	71	
63	15	23	21	45	19	24	43	
64	23	29	29	58	10	8	18	
65	19	44	59	73	27	34	61	
66	25	39	26	65	29	35	64	
67	23	43	41	83	36	24	60	
68	23	35	44	82	31	31	62	
69	22	36	35	71	25	19	44	
	231	370	517	667	319	248	567	

(195)
The Table of the Parish of Cranbrook.

Years	Weddings		Christened			Buried		
	M.	F.	M.	F.	Both	M.	F.	Both
1570	18	17	24	34	14	14	28	
71	14	11	27	35	14	16	30	
72	21	11	34	39	16	19	35	
73	29	28	41	51	20	21	41	
74	28				18	14	32	
75	25				18	12	30	
76	30	30	48	54	21	22	43	
77	18	16	28	34	14	16	30	
78	14	13	29	31	14	16	30	
79	14	17	22	31	13	18	31	
	115	128	198	241	111	125	236	

威廉配第 William Petty (1623-1687): 統計學之父 (政治經濟學之父)

凱特勒 Lambert Adolphe Jacques Quetelet (1796-1874): 近代統計學之父

高頓 Francis Galton (1822-1911): 心理統計學之父

皮爾森 Karl Pearson (1857-1936): 統計學之父

費雪 R.A. Fisher (1890-1962): 現代統計學之父



Karl Pearson (1857~1936)

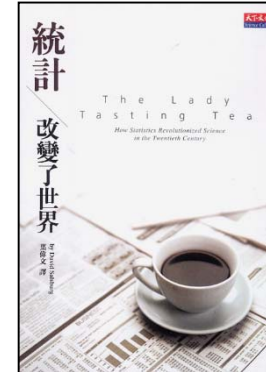


- 卡爾皮爾森：英國數學家、生物學家和統計學家。公認為統計學之父。
- 1879年畢業於劍橋大學數學系；曾參與激進的政治活動。出版幾本文學作品，並且作了三年的律師實習。

機率分佈在統計學中的重要性

統計革命的起點

- Karl Pearson (1857-1936)，發表一系列和**相關性** (correlation) 有關的論文，涉及動差、相關係數、標準差、卡方適合度檢定，**奠定了現代統計學的基礎**。
- 引入了**統計模型**的觀念：如果能夠決定所觀察現象的**機率分佈的參數**，就可以了解所觀察現象的本質。

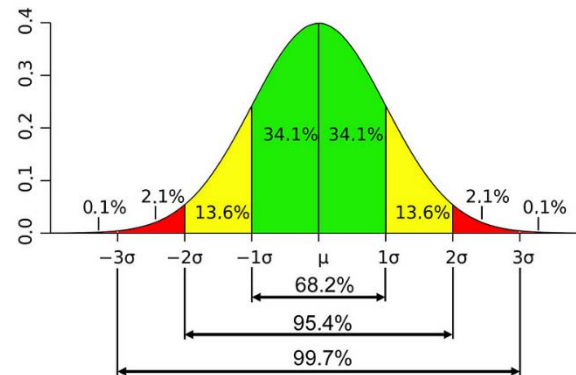


樣本變異數與樣本標準差

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

母體變異數與母體標準差

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$



機率在機器學習的分類模型中的重要性：

機率是非常重要的基礎，因為**分類模型**就是藉由**激活函數**將運算結果用**機率**呈現，才得以依據輸入資料預測最後應該分到哪一類。

Schweizer, B. (1984), **Distributions Are the Numbers of the Future**, in Proceedings of The Mathematics of Fuzzy Systems Meeting, eds. A. di Nola and A. Ventre, Naples, Italy: University of Naples, 137–149. (The present is that future.)

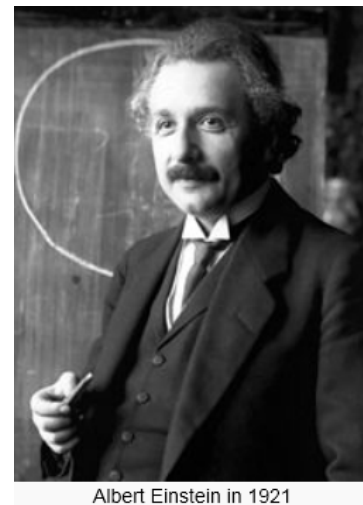
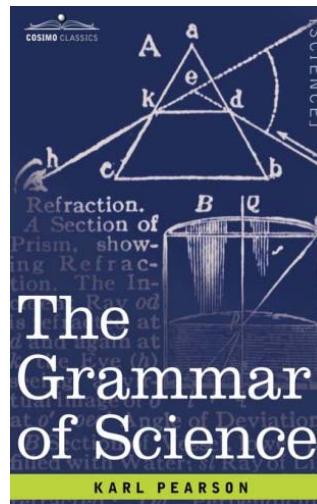


Karl Pearson對後世的貢獻

- 1890年於格里辛學院 (Gresham College) 開授《現代科學的範圍與概念》(the Scope and Concepts of Modern Science)，開始注意統計的原理及對科學研究的影響，開啟了現代統計學之門。此後統計學被應用於許多科學方面的研究。
- Pearson 的主觀強，經常對他本人認為有「爭議」的文章，刪改或退稿，並因此與英國本世紀最有才華的統計學家 Fisher 結下樑子。
- 1901年，創辦《**Biometrika**雜誌 (生物量測學)》，使數理統計有了自己的陣地。是數學在進入20世紀時最重大收穫之一。



(1900)



Albert Einstein in 1921

(1879-1955)

When the 23
year-old Albert
Einstein

實證精神的貫徹

Karl Pearson 小時候的故事



- 有人問皮爾森他所記得**最早的事**，他說「我不記得那時是幾歲，但是我記得是坐在高椅子上**吸吮著大拇指**，有人告訴我最好停止吮它，不然被吮的大拇指會變小」。
- 我把兩手的大拇指並排看了很久，它們似乎是一樣的，我對自己說：我看不出被吸吮的大拇指比另一個小，我懷疑她是否在騙我」。
- 「**不盲信權威**，**要求實證**，對於自己對觀測數據的意義的解釋深具信心，和懷疑與他的判斷不同的人態度是否公平」這些就是皮爾森一生獨具的特徵。

實證精神的貫徹

Karl Pearson丟銅板的故事

請問用手丟一個公正的銅板，

得到正面的機率是多少？得反面的機率又是多少？

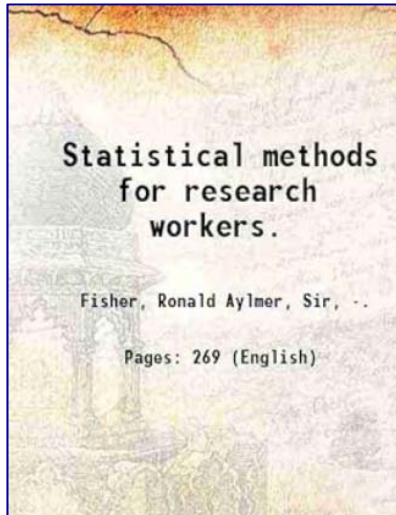
- 十八世紀法國布豐伯爵(Count Buffon): 丟過4040次銅板，有2048次是正面。
- 第二次世界大戰時，英國數學家柯瑞屈(John Kerrich)曾當過德軍的戰俘，他在集中營丟了10000次銅板，5067次是正面。
- 近代英國的統計學家皮爾森，更是擲了24000次銅板，得到12012次正面。

你問他們銅板的機率問題，那他們的答案可能是：

- 布豐伯爵說：「我確定機率是0.5069，不是0.5，我可能是千百年來第一個就此問題作實驗的第一人，以前的學者都不作實驗！」
- 柯瑞屈會說：「機率應該是0.5067，我作的實驗次數比一百多年前的布豐伯爵多兩倍以上，而且也比較接近理論值！」
- 皮爾森則說：「正確機率應該是0.5005，我花了好多的時間才完成這24000次實驗，是數學家柯瑞屈的兩倍，更是布豐伯爵的四倍以上。我們三個人實驗的結果都不是0.5，令人不禁懷疑古典機率理論是不是有錯！」

統計學者最少可分為理論及實驗這兩種學派，而將「銅板出現正面的機」說成二分之一的人，是在教入門班。親自做過實驗，並願意帶領學生做實驗親證的才是完整的進階學習。

R.A. Fisher (1890~1962)



- 生於英國, 為20世紀初期最重要的統計學家及近代**數理統計**開創者。
- 1925年, 著作《**研究者的統計方法**》(Statistical Methods For Research Workers, 1925), 對很多世代的科學家而言, 是一本劃時代的著作。
- 1935年, 著作《**實驗設計法**》(The Design of Experiments, 1935): 考量科學問題的第一步, 就是**實驗數據**。
- 1940年, 費雪檢驗了孟德爾提出的豌豆培育數據實例, 發現這些數據完美的不像是真的, 根本沒有展現應有的**隨機程度**。
- 發展適用於**小樣本的方法**(統計量的精確分配), 推導一些樣本統計量的確實分配(exact distribution), 奠定檢定假設 (testing hypothesis) 的邏輯化原則公式。
- 發展出實驗設計、變異數分析、最大概似法等, 將統計奠基在嚴密的**數學理論**上, 不僅使統計學的理论粲然大備, 也使之成為一個強而有力的科學工具。

淑女與下午茶

(The Lady Tasting Tea)

- 時間: 1920年代後期的一個夏日午後
- 地點: 英國劍橋戶外餐桌旁
- 人物: 一群大學研究員、女眷、訪客
- 事情經過: 某女士在下午茶時說，
把茶加進牛奶裡，和把牛奶加進茶裡
兩者喝起來完全不同!



我們來檢驗這
個命題!

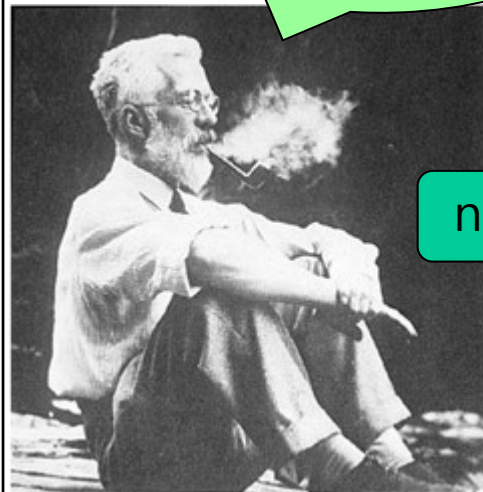
無法理解兩種混
合結果的化學成
份會有什麼差異?

nonsense!

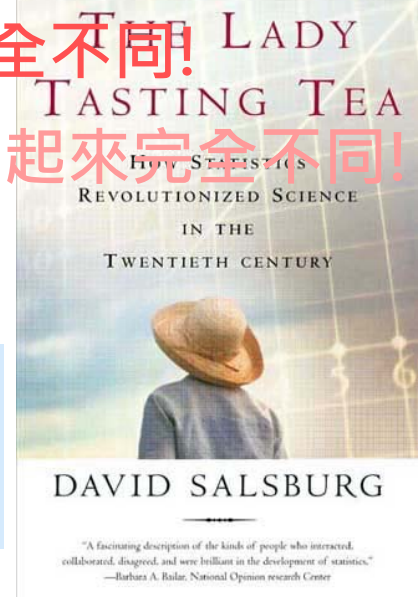
兩者喝起來完全不同!

能不能分辨不同
茶的風味有什麼
關係?這小問題沒
有科學價值啦!

不務正業，
浪費時間。



兩者喝起來完全不同!





費雪精確檢定

(Fisher's Exact Test)

Fisher先生的試驗

- Fisher給了這位淑女8杯茶，其中有4杯先放奶，讓她在不知情的情況下嘗奶茶，猜這杯是先加奶還是先加茶。
- 虛無假設(H0): 放奶和茶的順序與該位淑女的猜測獨立。
- 對立假設(H1): 放奶和茶的順序與該位淑女的猜測有關。

		倒入的順序		
		茶先倒	牛奶先倒	
女士猜測	茶先倒	a	b	4
	牛奶先倒	c	d	4
		4	4	8

Fisher's Exact Test

H0: 列變數與行變數為統計獨立

H1: 它們是統計相關的。

$a+b$ 及 $c+d$ 稱為列邊際次數。

$a+c$ 與 $b+d$ 稱為行邊際次數。

		行變數		
		1	2	
列變數	1	a	b	$a+b$
	2	c	d	$c+d$
		$a+c$	$b+d$	n

費雪精確檢定: 先備知識

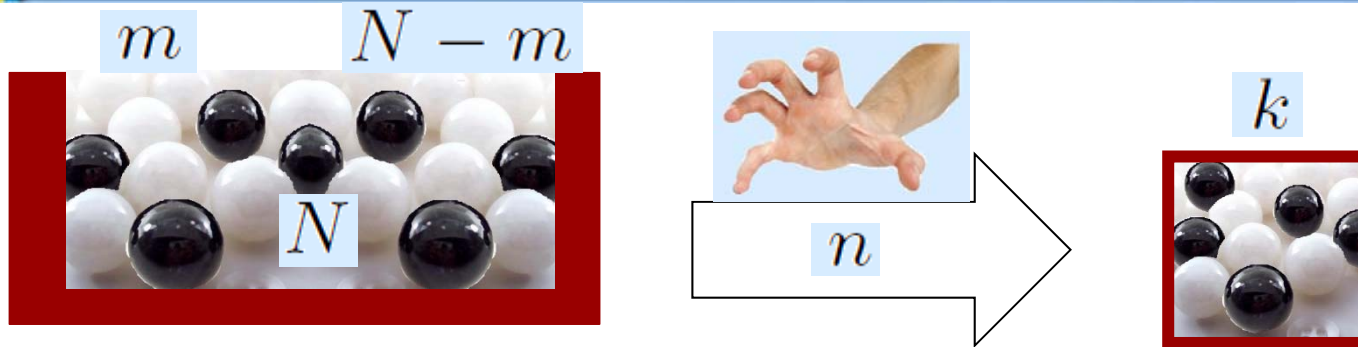


Table: Hypergeometric problem as a contingency table.

marbles	sampling without replacement		Total
	drawn	not drawn	
white	k	$m - k$	m
black	$n - k$	$N + k - n - m$	$N - m$
Total	n	$N - n$	N

Hypergeometric Distribution

$$P(k; N, m, n) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

Table: Hypergeometric problem as a contingency table.

marbles	sampling without replacement		Total
	drawn	not drawn	
white	n_{11}	$n_{1+} - n_{11}$	n_{1+}
black	$n_{+1} - n_{11}$	$n - n_{+1}$	$n - n_{+1}$
Total	n_{+1}	$n - n_{+1}$	n

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}} \binom{n - n_{1+}}{n - n_{+1} - n_{11}}}{\binom{n}{n_{+1}}}$$



費雪精確檢定：可能的結果

		倒入的順序		
		茶先倒	牛奶先倒	
女仕 猜測	茶先倒	3	1	4
	牛奶先倒	1	3	4
		4	4	8

		倒入的順序		
		茶先倒	牛奶先倒	
女仕 猜測	茶先倒	4	0	4
	牛奶先倒	0	4	4
		4	4	8

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = 0.229$$

p-value for $H_a : \theta > 1$ equals the right-tail probability that n_{11} is at least as large as observed:

$$P = P(3) + P(4) = 0.243 : \text{ not much evidence against } H_0.$$

n_{11}	Probability	p-value	X^2
0	0.014	1.000	8.0
1	0.229	0.986	2.0
2	0.514	0.757	0.0
3	0.229	0.243	2.0
4	0.014	0.014	8.0

one-sided alternative, $H_a : \theta > 1$.

費雪精確檢定： 數學為實驗之本



- 《實驗設計法》(The Design of Experiments, 1935)中費雪幾個實驗設計的絕佳範例，所提的方法牽涉複雜的數學。
- 科學界以費雪的研究，發展出不同的實驗設計法並應用在醫學，化學及工業上的品質管制。
- <<Lady Tasting Tea>>
 - 最少該喝多少杯檢定的效果才會出來？喝每一杯的順序要安排嗎？如何判別女士是猜對的還是喝得出不同？
 - 究竟是氣味的香醇，還是色澤的優美，抑或是滋味的甜甘讓這位淑女能區分不同的次序。



John Tukey (1915~2000):

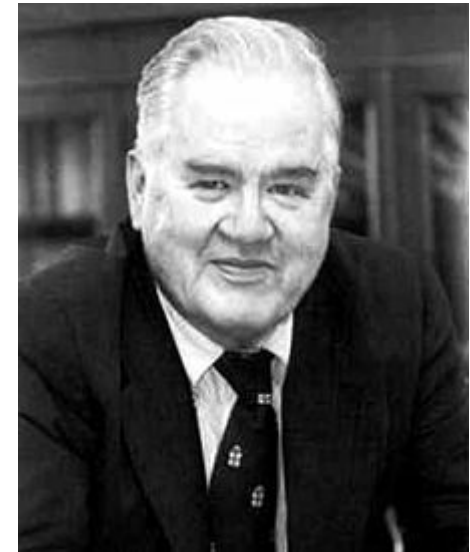
統計學界的畢卡索

23/118

生平

- 布朗大學**化學**學士及碩士。
- 1939年: 普林斯頓大學**數學**博士。(數理統計)
- 二次大戰加入火砲控制研究室，以及後來加入**AT&T**貝爾實驗室(**創立統計組**)，接觸統計上的實際問題。

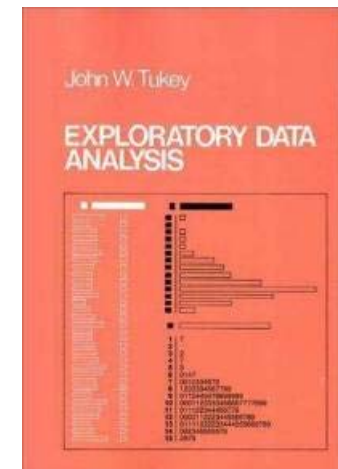
「**對正確的問題有個近似的答案，**
勝過對錯的問題有精確的答案。」



對後世的貢獻

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

- 發明快速傅立葉轉換(FFT)。
- 創造bit (位元)及 software(軟體)。
- 探索性的資料分析 (Exploratory Data Analysis, EDA, 1977)





「統計應該是科學，而非數學！」



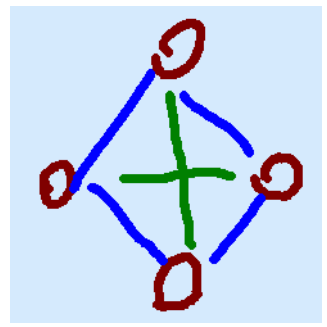
Stanford Linear Accelerator (1973)

他曾挑戰當時主流的數理統計學家，堅持 data analysis 是統計分析中不可忽視的步驟，**數學的假設需要 data 加以驗證才可行**。Tukey 說過統計應該是科學，而非數學！

數學思維 vs 統計思維
證明在哪裏? vs 數據在哪裏?

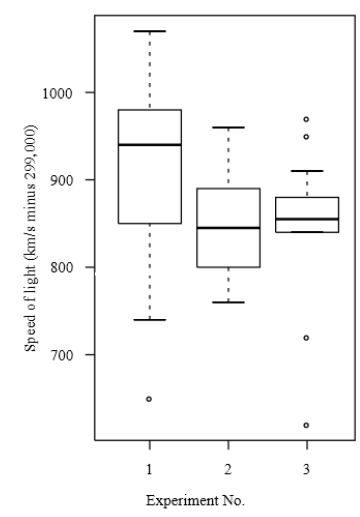
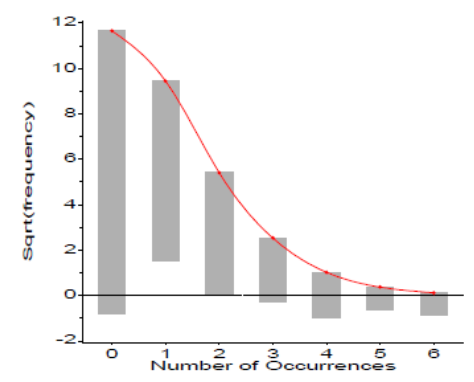


Stem and Leaf Plot



42		0
44		0000
46		000000
48		0000000000
50		00000000000000000000
52		000000
54		00000000000000
56		0000000000000000
58		0000000000
60		00000000000000
62		00000000000000
64		00000000000000
66		0000000000
68		00000000
70		00
72		0000
74		0
76		00000
78		0

Box-and-whisker plot



吳建福 院士



國立清華大學
工學院及理學院合聘榮譽講座
講座教授
吳建福 教授 (Prof. Chien-Fu Jeff Wu)
美國喬治亞理工學院 可口可樂講座教授

美國國家工程學院院士
中央研究院院士
中國科學院榮譽教授
加拿大 Waterloo 大學榮譽博士學位
印度統計學院 Mahalanobis 紀念講座
潘文淵傑出研究講座
國立清華大學 台積電講座

COPSS 會長獎
美國品管學會 Shewhart Medal、
Jack Youden 獎、
Frank Wilcoxon 獎、
Brumbaugh 獎

美國統計學會、國際數理統計學會、美國
品管學會 Fellow

AMSTATNEWS

The Membership Magazine of the American Statistical Association

HOME ABOUT EDITORIAL CALENDAR SUBMISSION INSTRUCTIONS PDF ARCHIVES

ADVERTISE

Home » Departments, Meetings

Jeff Wu to Deliver COPSS Fisher Lecture

1 MAY 2011 558 VIEWS NO COMMENT

Prepared by Michael Newton, COPSS Fisher Lecture Committee Chair, and Hugh Chipman, Acadia University

The 2011 [Committee of Presidents of Statistical Societies \(COPSS\)](#) asked C. F. Jeff Wu, professor and Coca Cola Chair in Engineering Statistics at Georgia Institute of Technology, to deliver the Fisher Lecture at the [Joint Statistical Meetings](#) in Miami Beach, Florida, on August 3 at 4:00 p.m.

Wu's research contributions span the full range of statistics, from theory to application, and touch many applied domains, from sample surveys to nanotechnology. They are notable for their combination of novelty, technical strength, and far-reaching vision. He has made especially significant contributions to experimental design. As one of his supporters wrote, "In view of Professor Wu's contribution to design of experiments, it is particularly fitting for him to deliver a lecture that honors R. A. Fisher, commonly regarded as father of the modern theory of experimental design."

Wu was born in Taiwan and earned his BS degree in mathematics from National Taiwan University in 1971. After two years of compulsory military service, he came to the United States to study at the University of California, Berkeley and earned his PhD in statistics in 1976. He spent his formative years at the University of Wisconsin-Madison, from 1977–1988, then moved to the University of Waterloo as the first holder of the GM/NSERC Chair in Quality and Productivity from 1988–1993. He spent 10 years at the University of Michigan before joining Georgia Institute of Technology.

吳建福 院士：統計學大家



English 繁體 RSS 網站地圖 收藏 郵箱



今天是 2011年10月1日 星期一

您現在的位置

科学 民主 爱国 奉献

吳建福認為，這個問題可以從3個境界來看待：Inspiration，靈感；Aspiration，抱負；Ambition，野心，或者企圖。他以統計學界“大家”的經歷對此進行解讀。

1919年，歷史上最偉大的統計學家R.A.Fisher從劍橋畢業，他的第一份工作是在一個農業試驗站開始的。Fisher本科讀數學，也讀生物，但對農業沒有任何接觸。然而試驗站有很多農業育種栽培的實驗，產生很多數據，Fisher在幾年內發明了兩個大成果——方差分析和試驗設計，當然，還有後來的似然函數。

吳建福心目中有兩位統計學天才，另一位是John W.Tukey。“他博士論文研究的是拓撲學，但他對統計學以及信號處理、電機、計算機科學等領域都有基本貢獻。”Tukey有個對統計學發展影響至深至遠的成果從未發表，只是在一本期刊上做過摘要，僅一頁。吳建福笑稱：“他這個文章的影響力頁數比（impact pages ratio）太大了。”

“要能成為大家，首先要有靈感。但要把一個領域做大，還一定要有志向、有抱負，要有能力把它做到底：有體力，能說能寫，有門生，能夠推廣，能夠應用。這不只是一兩把刷子了。”吳建福說。

統計學應用於農業和生物學的奠基者Karl Pearson，是Fisher出現之前最重要的統計學家。在學術地位受到Fisher挑戰後，Pearson“暴怒”了，他甚至放言：“Fisher就像堂吉訶德拿著長矛要去打風車一樣，他會把自己毀掉，或者他的整個理論會被毀掉。”

對聲望過於渴求的Pearson幾乎成了“學霸”代言人，此後卻在Fisher、Neyman等人的映襯下日漸狹隘了。




統計界的諾貝爾獎： COPSS Awards

WINNERS OF THE COPSS PRESIDENTS' AWARD

- 1979: Peter J. Bickel, University of California, Berkeley
- 1982: Stephen E. Fienberg, Carnegie Mellon University
- 1983: Tze Leung La, Stanford University 黎子良, 中研院院士/美國史丹佛大學教授
- 1984: David V. Hinkley, University of California, Santa Barbara
- 1985: James O. Berger, Duke University
- 1986: Ross L. Prentice, Fred Hutchinson Cancer Research Center
- 1987: Chien-Fu Jeff Wu, Georgia Institute of Technology 吳建福, 中研院院士 美國喬治亞理工學院教授
- 1988: Raymond J. Carroll, Texas A&M University
- 1989: Peter Hall, Australian National University
- 1990: Peter McCullagh, University of Chicago
- 1991: Bernard W. Silverman, University of Bristol
- 1992: Nancy Reid, University of Toronto
- 1993: Wing-Hung Wong, Stanford University 王永雄, 中研院院士 美國史丹佛大學教授
- 1994: David L. Donoho, Stanford University
- 1995: Iain M. Johnstone, Stanford University
- 1996: Robert J. Tibshirani, Stanford University
- 1997: Kathryn Roeder, Carnegie Mellon University
- 1998: Pascal Massart, Université de Paris-Sud
- 1999: Larry A. Wasserman, Carnegie Mellon University
- 2000: Jiangjing Fan, Princeton University 范劍青, 普林斯頓大學統計系教授
- 2001: Xiao-Li Meng, Harvard University 孟曉犁, 美國哈佛大學統計系教授
- 2002: Jun Liu, Harvard University 劉軍, 美國哈佛大學統計系教授
- 2003: Andrew Gelman, Columbia University
- 2004: Michael A. Newton, University of Wisconsin
- 2005: Mark J. van der Laan, University of California, Berkeley
- 2006: Xihong Lin, Harvard University 林希虹, 美國哈佛大學生物統計系教授
- 2007: Jeffrey S. Rosenthal, University of Toronto
- 2008: T. Tony Cai, University of Pennsylvania 蔡天文, 賓州大學統計系教授
- 2009: Rafael Irizarry, Johns Hopkins University
- 2010: David B. Dunson, Duke University

of Mathematical Statistics



[Awards & Publications](#)
[Awards & Honors](#)
[Meetings](#)
[Jobs](#)
[IMS & Officials](#)
[Contacts](#)
[Home](#)

COPSS Awards

The Committee of Presidents of Statistical Societies (COPSS) Awards are jointly sponsored by the American Statistical Association, the Institute of Mathematical Statistics, the International Biometric Society ENAR, the International Biometric Society WNAR, and the Statistical Society of Canada. They are: the **R.A. Fisher Lectureship**; the **Florence Nightingale David Award**; the **Elizabeth L. Scott Award**; the **George W. Snedecor Award**; and the **Presidents' Award**. Each award consists of a plaque, a citation, and a cash honorarium. Additional information for all the awards and a list of past COPSS awards recipients can be found at the COPSS web site: <http://www.niss.org/copss>. A list of COPSS members and current committee members is also available at the COPSS website.

- 2011: Nilanjan Chatterjee, Johns Hopkins University
- 2012: Samuel Kou (寇星昌), Harvard University
- 2013: Marc A. Suchard, UCLA
- 2014: Martin J. Wainwright, University of California, Berkeley
- 2015: John D. Storey, Princeton University
- 2016: Nicolai Meinshausen, ETH Zurich

https://en.wikipedia.org/wiki/COPSS_Presidents%27_Award

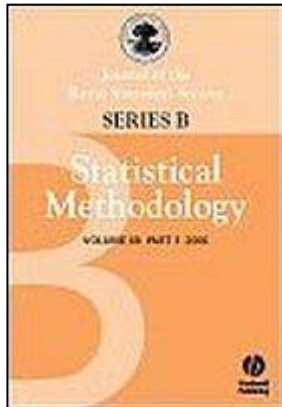
統計期刊四大天王



The **Annals** of Statistics
(統計年鑒)



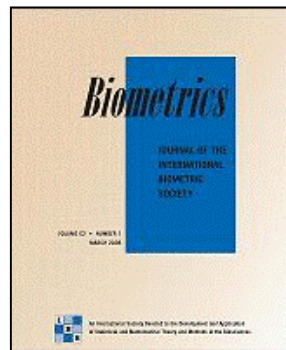
Journal of the American Statistical Association
(**JASA**) (美國統計協會會刊)



Journal of the Royal Statistical Society: Series B, Statistical Methodology
(**JRSSB**) (英國皇家統計協會會刊理論分冊)



Biometrika
(生物測量學)



Biometrics
(生物測量)

<https://www.scimagojr.com/journalrank.php?category=2613>

SJR Scimago Journal & Country Rank

Title	Type	↓ SJR
1 Annals of Mathematics	journal	6.760 Q1
2 Journal of Business and Economic Statistics	journal	6.150 Q1
3 Annals of Statistics	journal	5.219 Q1
4 Journal of the Royal Statistical Society. Series B: Statistical Methodology	journal	3.976 Q1
5 Annual Review of Statistics and Its Application	journal	3.896 Q1
6 Biometrika	journal	3.566 Q1
7 Journal of the American Statistical Association	journal	3.397 Q1
8 Annals of Probability	journal	3.099 Q1
9 Bioinformatics	journal	2.940 Q1
10 Vital and health statistics. Series 2, Data evaluation and methods research	book series	2.723 Q1
11 Journal of Statistical Software	journal	2.722 Q1

1 - 50 of 258



數據科學 Data Science



趙民德
台灣

九歌文學網

趙民德，國立台灣大學數學系畢業、美國加州大學柏克萊分校統計博士。在美國求學及工作多年後，1982年回台灣籌設中央研究院統計學研究所，該所於1987年正式成立，並正名為統計科學研究所。國內統計學有今日的發展，以及能在世界佔一席之地，功不可沒。

在文學成就上，名家王鼎鈞以「詩的精緻，劇的張力，散文的鋪陳」肯定其業餘小說家的地位。

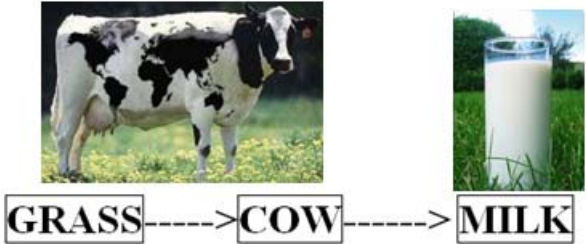
「現代的數據，已多得到了可怕的程度。老鼠若是掉到了米倉裡還會餓死，

當然是老鼠的錯。

統計這個行業在現代這個充滿數據的社會裡為何還要和一些不食人間煙火的數學家爭那區區的一塊小餅？我想，我們不要把自己這個行業看得太小了。」

...統計是以**數據**為食物的動物。
統計的本業是消化數據，並產生有營養的結果。

**它的本質，
和母牛相差不多。**



統計學是工具。工具必需要有使用的空間及場合，才能見到真正的效益。

引自趙民德，1999，「統計已死，統計萬歲！」第八屆南區統計研討會演說稿



常用機率分佈的應用

- **Discrete uniform distribution (離散型均勻分佈)**, for a finite set of values (e.g. **the outcome of a fair die**) ($X \sim Unif(\{a, \dots, b\})$)
- **Binomial distribution (二項式分佈)**, for the number of "positive occurrences" (e.g. **successes, yes votes, etc.**) given a fixed total number of independent occurrences. ($X \sim B(n, p)$)
- **Normal distribution (常態分佈)**, for a single real-valued quantity that grow linearly (e.g. **errors, offsets**) ($X \sim N(\mu, \sigma^2)$)
- **Log-normal distribution (對數常態分佈)**, for a single positive real-valued quantity that grow exponentially (e.g. **prices, incomes, populations**) ($\log(X) \sim N(\mu, \sigma^2)$)
- **Chi-squared distribution (卡方分佈)**, the distribution of a sum of squared standard normal variables; useful e.g. for **inference** regarding the **sample variance** of normally distributed samples. ($X \sim \chi^2_{(d)}$)

現代統計時期



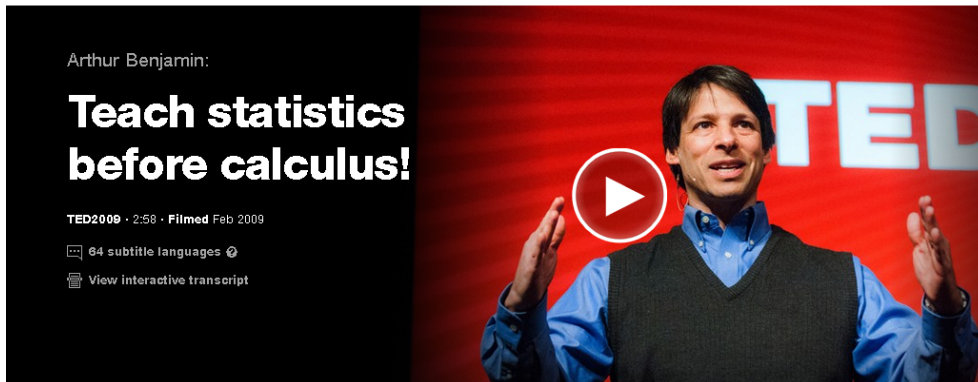
- 1980年代個人電腦的發展與相關的統計軟體的開發, 例如: SAS、SPSS、BMD, 促進了統計的發展與應用。今天人們日常生活已經不知不覺的在使用統計學或應用統計學的觀念。



TED Ideas worth spreading

WATCH DISCOVER ATTEND PARTICIPATE

SAS 解決方案 產業類別 專業服務 成功案例 事業夥伴 社群資源 關於 SAS



Share this idea

f in t e c 1,906,711 Total views

http://www.ted.com/talks/arthur_benjamin_s_formula_for_changing_math_education

Analytics software & solutions

Request Pricing



敏捷靈活

從簡單的桌面解決方案到高性能、分散式處理, 我們提供滿足您各種需要且快速提供洞察的方案。了解為何SAS進階分析市場占有率高達33%



統計 – 21世紀的明星產業

32/118

大數據(big data)似乎無所不在，也無所不能，但其中也存在許多迷思。請舉例說明有哪些迷思？



您

大數據(big data)似乎無所不在，也無所不能，但其中也存在許多迷思。請舉例說明有哪些迷思？



ChatGPT

大數據在當今社會確實無處不在，從商業決策、醫療保健到政府政策等各個領域都有其身影。然而，圍繞大數據的討論中也存在許多迷思和誤解。以下是一些常見的大數據迷思：

1. 更多數據等於更多洞察

許多人認為擁有更多數據自然能導出更多有價值的洞察和更準確的預測。然而，數據量的增加並不總是導致更好結果的關鍵因素。數據質量、相關性和處理方法才是獲得有用洞察的關鍵。

2. 大數據意味著完全自動化

存在一種觀點認為，有了大數據和先進的分析工具，所有的數據分析過程可以完全自動化，無需人工干預。實際上，數據分析需要專業知識來設定正確的分析框架、解釋結果並做出明智的決策。

3. 大數據保證客觀性

許多人認為大數據分析是完全客觀的，因為它基於數據和算法。但實際上，數據收集、處理和分析過程中的主觀決策可能導致偏見和錯誤。數據來源的選擇、數據清洗的方法以及算法的設計都可能引入偏見。

4. 大數據技術只適用於大公司

有一種迷思認為只有大公司才能利用大數據技術，因為需要昂貴的技術和專業人才。實際上，隨著技術的發展和成本的降低，中小企業也能夠利用大數據技術來改進業務和決策過程。

5. 大數據可以解決所有問題

另一個常見的迷思是大數據可以為所有問題提供解決方案。雖然大數據分析能夠提供深刻的洞察和有價值的信息，但它不是萬能的。成功利用大數據需要明確的目標、適當的方法和對結果的正確解讀。

總之，雖然大數據提供了前所未有的機會，但理解其局限性和挑戰對於有效利用這些技術至關重要。正确地识别和应对这些迷思有助于更有效地利用大数据的潜力。



(註: 此問題是參照余清祥教師「112-1統計諮詢簡介」講義中的提問)



統計 – 21世紀的明星產業

33/118

大數據時代，統計為什麼應運而起，
它憑藉哪些特殊本領？它可以創造哪些價值？



您

大數據時代，統計為什麼應運而起，它憑藉哪些特殊本領？它可以創造哪些價值？



ChatGPT

在大數據時代，統計學的重要性日益凸顯，這主要得益於其在數據分析和解釋方面的獨特能力。統計學憑藉以下幾個特殊本領，在大數據背景下應運而起，創造出巨大價值：

特殊本領

- 數據處理與分析能力：**統計學提供了一套成熟的方法論，用於處理和分析大量數據，包括數據清洗、數據轉換和數據摘要等。
- 模型建立與預測：**統計學能夠通過建立數學模型來揭示數據之間的關係，進而進行預測和推斷。這對於預測市場趨勢、用戶行為等至關重要。
- 不確定性的量化與管理：**統計學專注於量化不確定性和風險，提供了評估和管理不確定性的工具，如置信區間、假設檢定等。
- 決策支持：**通過數據分析和模型預測，統計學為決策提供了科學依據，幫助企業和組織在不確定性中做出更加明智的選擇。

創造的價值

- 提高決策質量：**統計學通過提供基於數據的見解和預測，幫助決策者理解市場和客戶行為，從而提高決策的質量和效率。
- 優化業務流程：**統計分析可以揭示業務流程中的瓶頸和效率損失點，指導業務優化和成本控制。
- 創新和產品開發：**統計學可以幫助企業分析用戶需求和市場趨勢，指導產品開發和創新，從而保持競爭優勢。
- 風險管理：**在金融、保險等領域，統計學的風險評估工具對於識別、量化和**管理**風險至關重要。
- 增強客戶洞察：**統計學能夠幫助企業從大量的客戶數據中提取有價值的信息，深化對客戶行為和偏好的理解，從而提升客戶體驗和滿意度。

總之，統計學在大數據時代扮演著不可或缺的角色，它不僅能夠幫助我們理解和利用數據，還能夠在不確定性中尋找到明確的方向，為各行各業的發展提供強大的數據支持和決策依據。

(註: 此問題是參照余清祥教師「112-1統計諮詢簡介」講義中的提問)



統計 – 21世紀的明星產業

34/118

大數據時代，統計專業人員必需培養哪些技能與知識？
如何因應大數據發展過程的震盪？



您

大數據時代，統計專業人員必需培養哪些技能與知識？如何因應大數據發展過程的震盪？



ChatGPT

在大數據時代，統計專業人員面臨著前所未有的機遇與挑戰。為了充分利用大數據的潛力並有效應對其發展過程中可能出現的震盪，統計專業人員需要培養以下技能與知識：

必需的技能與知識

1. **數據處理與管理能力**：熟悉數據預處理、清洗和轉換技術，以及數據庫管理系統，如SQL、NoSQL。
2. **統計分析與建模**：深入理解統計理論，包括推斷統計、預測模型、迴歸分析等，並能夠應用這些知識解決實際問題。
3. **機器學習與人工智慧**：掌握基礎的機器學習算法和深度學習框架，了解它們在大數據分析中的應用。
4. **編程與軟件工具**：熟練使用至少一種數據分析相關的編程語言，如Python或R，以及數據分析和可視化工具。
5. **大數據技術**：了解並能夠應用大數據處理框架和工具，如Hadoop、Spark等。

6. **溝通與解釋能力**：能夠清晰地溝通分析結果，並將複雜的統計概念解釋給非專業人士聽。
7. **批判性思維**：培養批判性思維能力，對數據源和分析結果進行質疑，確保分析的準確性和可靠性。

如何因應大數據發展過程的震盪

1. **持續學習**：大數據技術和工具正在快速發展，統計專業人員應該持續學習新技術、新方法，以保持自己的競爭力。
2. **靈活適應**：對於大數據發展中出現的新問題和挑戰，保持開放和靈活的態度，適應變化，尋找創新的解決方案。
3. **跨領域合作**：與其他領域的專家合作，如資訊科學家、業務分析師等，共同開發解決方案，以應對複雜的大數據問題。
4. **強化倫理意識**：在數據分析過程中，嚴格遵守數據隱私和倫理規範，保護個人隱私和數據安全。
5. **風險管理**：學習如何識別和管理與大數據相關的風險，包括數據質量問題、偏見和錯誤的風險。

通過培養這些技能和知識，統計專業人員不僅能夠有效利用大數據的潛力，還能夠在大數據發展過程中的各種震盪中保持穩定，為決策提供堅實的數據支持。

(註: 此問題是參照余清祥教師「112-1統計諮詢簡介」講義中的提問)

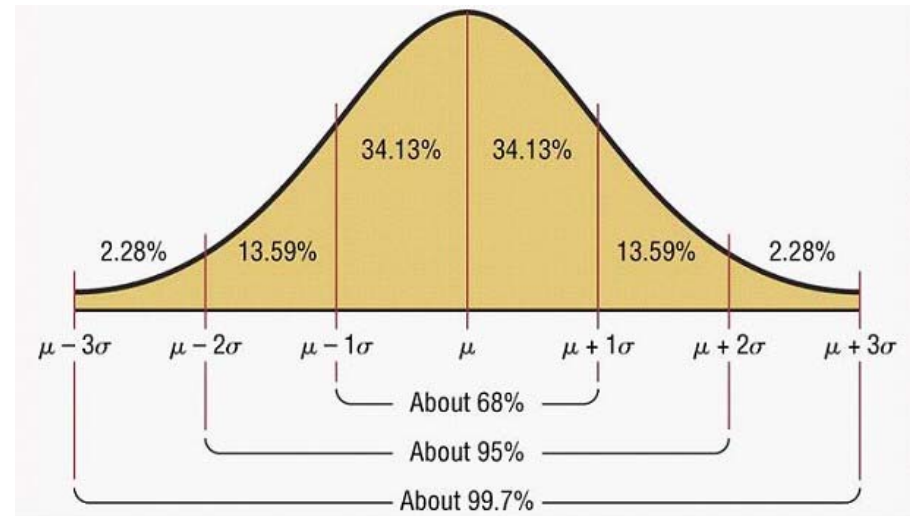


統計應用實例

35/118

六標準差管理 (Six Sigma, 6σ)

- 標準差是用來描述母體的「變異性」(variation)或「不一致程度」(inconsistency)。
- 過去習慣用「平均數」作為其績效表現，例如平均成本、平均產能、平均交貨時間、平均工資等，但卻忽略了變異程度。現今用「標準差」來衡量產品之品質分佈的變異狀況。
- 六標準差管理是1986年Motorola發展出來的管理方法。符合 6σ 就代表每生產出一百萬個產品，其不良品必須低於3.4個(良率99.99966%)。轉換成一般服務業的用語，就是每一百萬人次的客人，其中不滿意服務的不可以超過四人。



Source: <https://www.usastock88.com/2013/04/BBand.html>

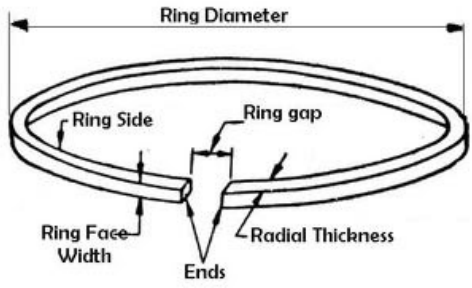
- 在美國而言(Harry 「1978」)，平均99%的品質水準相當於 (假設資料呈現常態分佈)：
(1)每小時有2萬件郵件遺失，這相當於1%的誤投率。(2)每天供應的自來水有15分鐘是不適合飲用的。(3)每星期有5千例外科誤診。(4)每個月有7小時停電。
- 對良率的要求需更嚴格的場景：
(1) 某航空公司宣稱其飛安率可高達99.73%。(每起落一百萬個架次，約有2,700次失事)
(2) 某醫院聲明其婦產科接生新生嬰兒的平均成功率是99.9%。
- 六標準差是一個利用統計、問題排除和問題預防等工具，將客戶滿意度提升至99.999%完美層級的品質改善商業策略。

中華六標準差管理學會
<https://www.6sigmai.org/>

統計應用實例: 製造業 (管制圖, Control Chart) 統計製程管制 (Statistical Process Control, SPC)



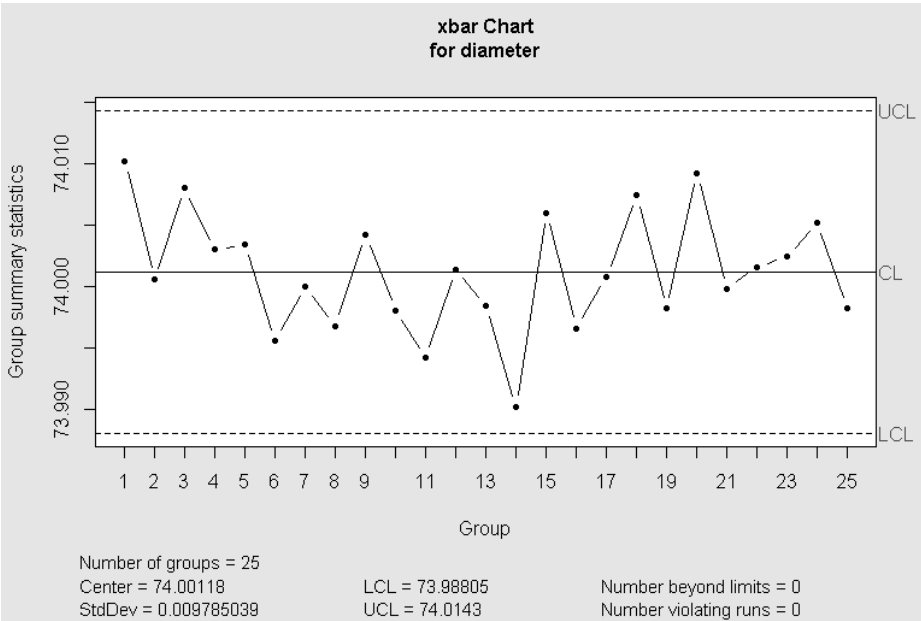
活塞環 (Piston Rings)



Piston rings (活塞環)					
Sample	inside diameter				
1	74.030	74.002	74.019	73.992	74.008
2	73.995	73.992	74.001	74.011	74.004
3	73.988	74.024	74.021	74.005	74.002
4	74.002	73.996	73.993	74.015	74.009
•	•	•	•	•	•
23	74.010	73.989	73.990	74.009	74.014
24	74.015	74.008	73.993	74.000	74.010
25	73.982	73.984	73.995	74.017	74.013

- 管制圖: 一種圖形表示工具, 縱軸代表產品品質特性(例如: 長度、重量、強度), 橫軸代表產品的編號或製造日期, 用來監測產品製造過程是否呈穩定狀態, 藉以作為品質改進之資訊。
- 發明: Walter A. Shewhart (1924) (蕭華特於貝爾實驗室)。
- 意義: 將測量數據, 計算位置統計量(例如: 平均數、中位數)及變異統計量(例如: 標準差、全距)後, 依中央極限定理及常態分配原理, 求得製程之平均及變異之中心線 (Centerline, CL)與上、下管制界線(Upper Control limit, UCL)(Lower Control limit, LCL), 連同數據繪成線圖, 藉以判斷品質隨時間之變化狀況。
- 判別: 樣本點出現在UCL或CLC之外, 表示為異常狀況。

<https://www.automobilehut.in/2023/01/what-is-piston-rings-types-of-piston.html>



管制圖種類



計量值管制圖 (通常假設製程資料來自常態分配，且資料間彼此相互獨立)

- 管制目標值: 平均數管制圖 (X-bar Chart)

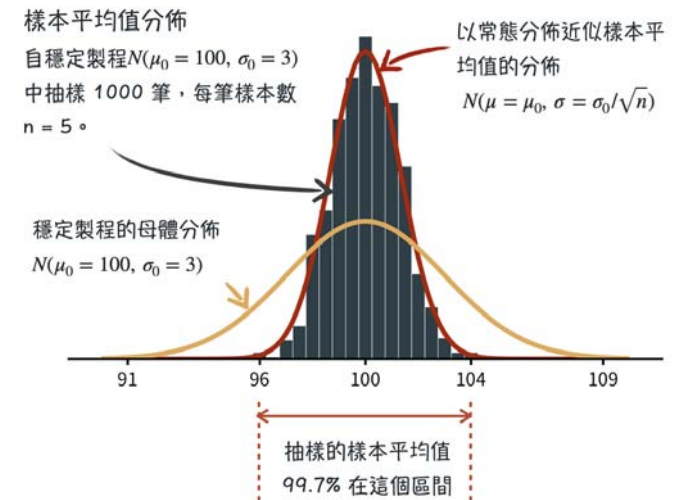
設 y 為一品質特性之樣本統計量， y 的平均數 μ_y ，標準差為 σ_y ，則管制圖中UCL/LCL通式:

$$UCL = \mu_y + k\sigma_y$$

$$CL = \mu_y$$

$$LCL = \mu_y - k\sigma_y$$

- 管制變異: 全距管制圖 (R Chart)、標準差管制圖 (S Chart)



Source: https://www.sightingdata.com/post/build_and_operate_control_chart/

計數值管制圖

- 不合格率管制圖 (p Chart)
- 不合格數管制圖 (pn Chart)
- 缺點數管制圖 (C Chart)
- 單位缺點數管制圖 (u Chart)

運用二項分配(Bernoulli distribution)的特徵值來建構其上下界限。

運用卜瓦松分配(Poisson distribution)的特徵值來建構其上下界限。

自穩定製程中抽樣 30 筆的樣本數據, 每筆樣本數為 5。

抽樣編號	\bar{x}					樣本平均值	樣本標準差
	樣本 1	樣本 2	樣本 3	樣本 4	樣本 5		
1	99.02	101.66	97.98	100.64	100.93	100.0	1.51
2	103.52	101.86	99.66	102.75	99.33	101.4	1.86
3	101.58	97.62	104.28	95.6	99.29	99.7	3.38
4	99.42	97.45	100.18	97.55	93.85	97.7	2.45
5	99.51	102.13	99.2	95.61	102.23	99.7	2.71
...
26	97.69	97.74	99.1	104.99	96.27	99.2	3.41
27	97.65	102.73	99.57	95.17	102.85	99.2	2.93
28	94.54	105.35	105.66	104.47	98.86	101.8	4.90
29	97.27	98.99	95.76	100.65	102.01	98.9	2.51
30	99.14	101.41	98.59	99.28	98.66	99.4	1.15

取平均值 $\bar{\bar{x}}$ 和 樣本標準差 \bar{s}
 樣本總平均值 $\bar{\bar{x}} = 99.85$
 樣本標準差平均值 $\bar{s} = 2.82$

管制圖之統計原理

Source: 鄭春生(2010), 品質管理：現代化觀念與實務應用(第四版), 全華圖書出版社。

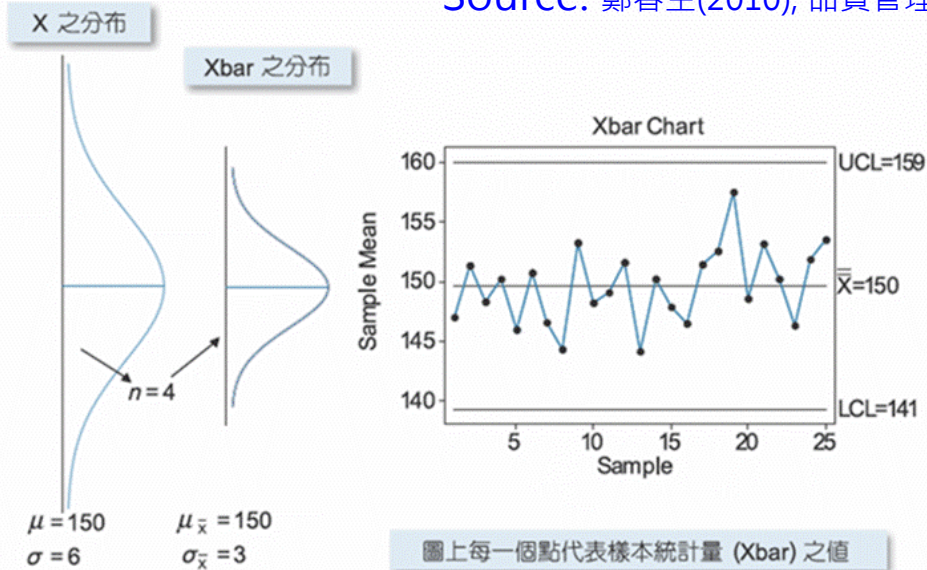


圖 6-5 管制圖之原理

H_0 : 製程處於穩定狀態 (the process is in control)

H_1 : 製程處於失控狀態 (the process is out of control)

決策	穩定狀態	失控狀態
製程調整	型I誤差 (Type I)	決策正確
不調整	決策正確	型II誤差 (Type II)

- 管制圖可說是統計假設檢定之圖形表示法。
 - 將樣本統計量描繪在管制圖上，可視為執行一次假設檢定。
 - 一點落在管制界限內：不能拒絕製程是在管制內之虛無假設。
 - 一點落在管制界限外：拒絕接受製程為管制內之假設。
- 管制圖之型I誤差：當製程實際為管制內時，卻誤判為管制外之機率。
- 管制圖之型II誤差：製程實際為管制外，卻誤判為統計管制內之機率。

統計與生物醫學



專題報導

生物醫學研究上的統計

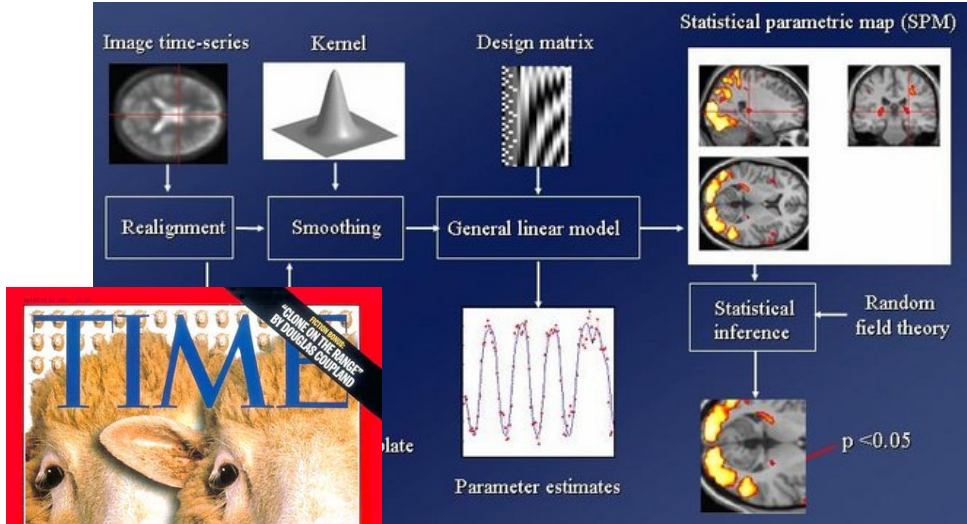
■ 石瑜 · 蘇佩芳

隨著21世紀生物科技資訊時代的來臨，
生物統計在醫學上扮演著舉足輕重的角色，
但統計在生物醫學研究領域中究竟扮演怎樣的角色？
對於生物醫學未來的發展又會產生怎樣的衝擊呢？

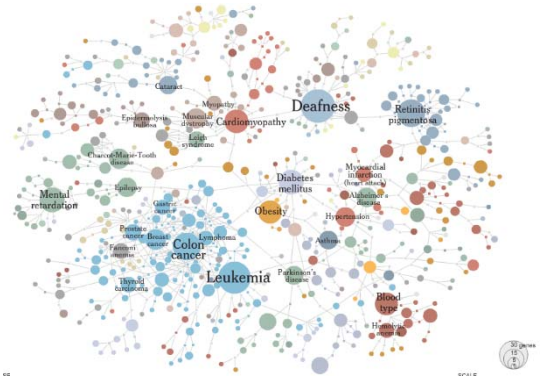
2010年初美國一項針對200種職業的調查，統計師的工作被評選為全美前十名最佳工作之一。另外，針對工作環境佳及體力要求低這兩項指標，在200個職業中，統計師都排名第一。隨著21世紀生物科技資訊時代的來臨，生物統計在醫學上的確扮演

排名	工作名稱	起薪	平均收入
1	精算師	\$49,000	\$85,000
2	軟體工程師	\$54,000	\$85,000
3	電腦系統分析師	\$45,000	\$76,000
4	生物學家	\$39,000	\$71,000
5	歷史學家	\$34,000	\$62,000
6	數學家	\$54,000	\$95,000
7	律師助理	\$29,000	\$46,000
8	統計學家	\$40,000	\$73,000
9	會計師	\$37,000	\$59,000
10	牙科技師	\$44,000	\$67,000

2010年全美最佳工作前十名 (資料來源：華爾街日報 (01/05/2010))



(Dolly, 1996~2003)
蘇格蘭羅斯林研究



在生物醫學的領域中，統計學家扮演著從事
實驗設計、資料的量測和蒐集、資料的歸納
和分析等重要工作的角色。



1953: James Watson and Francis Crick

40/118

Three dimensional double helix structure of DNA.

No. 4356 April 25, 1953

NATURE

737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

- ¹ Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
- ² Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
- ³ Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (5) (1956).
- ⁴ Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey¹. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β -D-deoxy-ribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis.

is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphate atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally^{2,4} that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data^{3,4} on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

738

NATU

King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. WATSON
F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.

¹ Pauling, L., and Corey, R. B., *Nature*, **171**, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, **39**, 84 (1953).

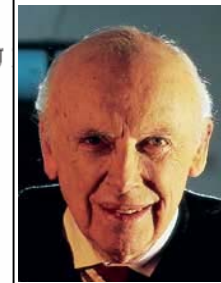
² Furberg, S., *Acta Chem. Scand.*, **6**, 634 (1952).

³ Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, **9**, 402 (1952).

⁴ Wyatt, G. B., *J. Gen. Physiol.*, **36**, 201 (1952).

⁵ Astbury, W. T., *Symp. Soc. Exp. Biol.*, **1**, Nucleic Acid, 66 (Camb. Univ. Press, 1947).

⁶ Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, **10**, 192 (1953).



1928~

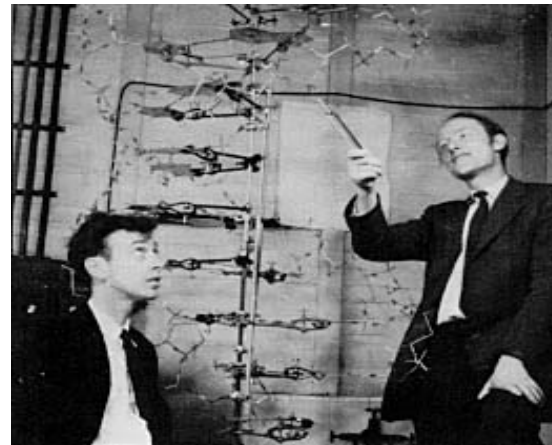
1916~2004

source: 科學人

羅莎琳·富蘭克林



Nobel Prize (1962)



出生

1920年7月25日

英國倫敦諾丁山

逝世

1958年4月16日 (37歲)

英國倫敦切爾西

卯巢樓



2001:人類基因體計畫 (Human Genome Project)





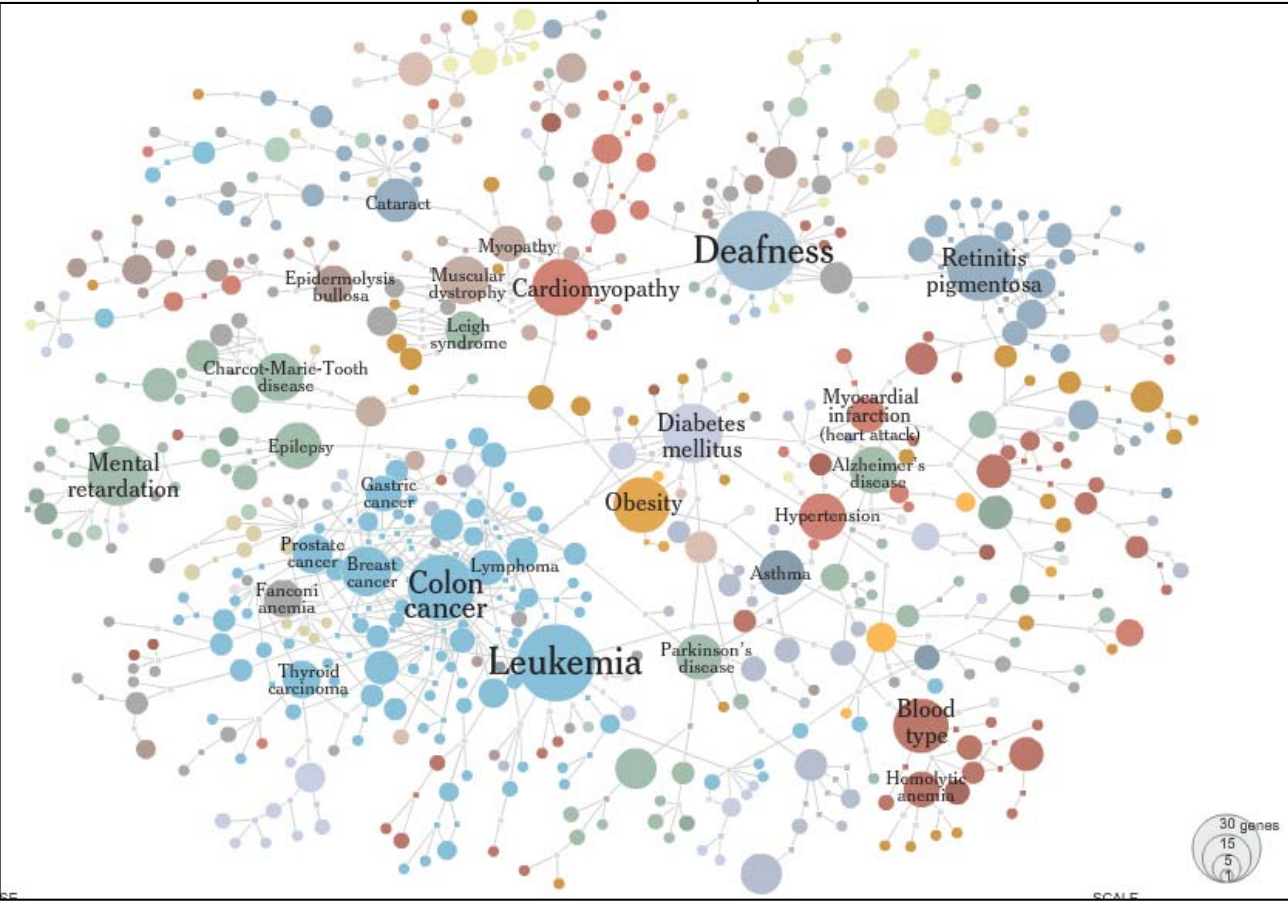
生物資訊學 Bioinformatics

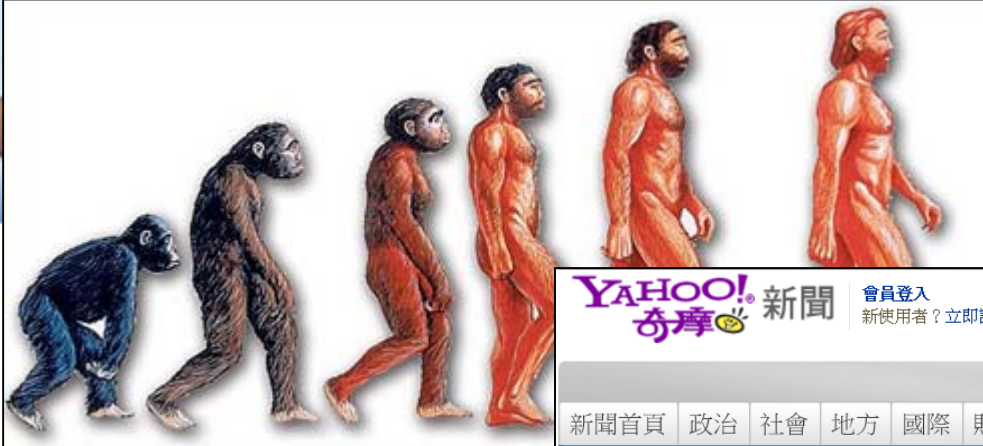
生物學 + 資訊學 + 統計學

```

000001 gctgctggaaggggagctggccgggtgggcccattggccggctgcaggetctgggttccgctgctgc
000065 tggggcggggttggcttgccttggccacggcaactgtggccgtggccccagtagatccaaaccta
000129 ccaccggcgtacaccctgtaccccaacaacttccagttccggtagcatgtcagttcggccgog
000193 caggcgggtgcgtcgtcctcgcagaggccttccagcgtaccgtaacctgctcttcggttcog
000257 gctcttggccccgac
000321 ctccgtcgtcacagc
000385 accattaatgatgac
000449 agaacttcagtcagc
000513 taagaacttctccg
000577 ttgtctagcatcctg
000641 acttgggtggacgact
000705 gtcctcaacctgt
000769 aggcttcggggtatc
000833 caggtgccccgggt
000897 ggtgaaccccagctc
000961 gtcttcccggacttt
001025 accccaacatccagg
001089 ctacatccagacgct
001153 gtatttgataataaa
001217 cagttagatcatgt
001281 tcctggtagctgaa
001345 ctggggttccatggt
001409 gagagtatgtggacc
001473 gagactgtggagcag
001537 cgttgtgagctggtg
001601 agtttgagcagactt
001665 gcgctgagcaccacg
001729 actgcctcccggccc
001793 ctggtgctggcactg

```





YAHOO! 新聞 會員登入 新使用者? 立即註冊

新聞首頁 政治 社會 地方 國際 財經 科技 運動 健康 教育 藝文 資訊3C 科學發展 自然環境 照片故事 專輯 民調中心 雜誌 愛蘋果 優等3C

新聞首頁 > 科技 > 科學發展 > 法新社 寄給朋友 | 友善列印 | 字級設定: 小 中 大 巨

研究：人類基因與海綿最接近

更新日期: 2010/08/06 23:35 張曉雯 (法新社雪梨6日電) 人類可能是猩猩的後代; 但澳洲科學家發現人類與海底生物更相近的證據。一項研究指出, 海綿動物有7成基因與人類相同。

世界遺產大堡礁的海綿基因序列顯示, 這種古老的海洋動物有許多基因跟人類相同, 包括與疾病與癌症相關的基因。

昆士蘭大學 (University of Queensland) 領導研究的迪克南 (Bernard Degnan) 說, 研究結果「可解釋一連串差異」, 並可為癌症與幹細胞研究的突破奠定基礎。他並說: 「這可能會影響我們對幹細胞的看法, 以及如何將海綿運用在未來醫學。」

研究本周發表於「自然」(Nature) 期刊, 是國際科學家團隊5年以上的研究成果。(譯者: 中央社張曉雯)

NATIONAL GEOGRAPHIC

- HOME
- ANIMALS
- DAILY NEWS
- ENVIRONMENT
- GAMES
- GREEN GUIDE
- HISTORY
- KIDS
- MAPS
- MUSIC
- PHOTOGRAPHY
- SCIENCE & SPACE
- TRAVEL & CULTURES
- VIDEO

MAIN ANIMAL

Chimps, Gene Stu

Stefan Lovgren for National Geog August 31, 2005

Scientists have that humans ar

"Darwin wasn't just provocative in saying that we descend from the apes—he didn't go far enough," said Frans de Waal, a primate scientist at Emory University in



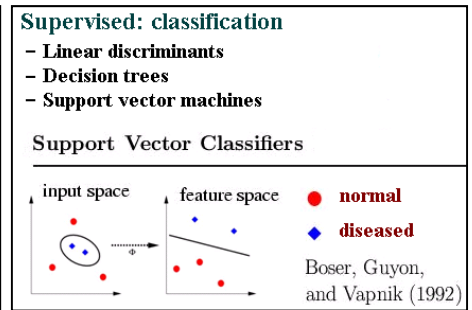
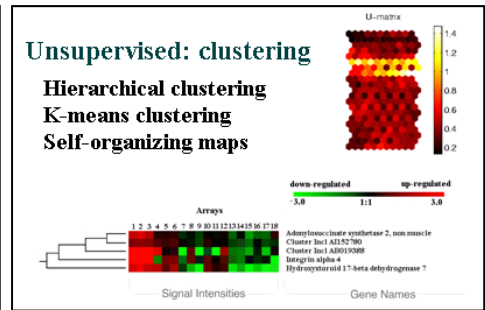
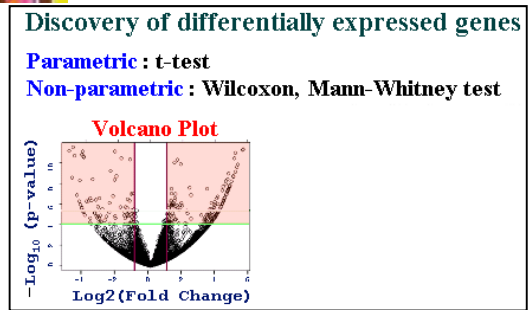
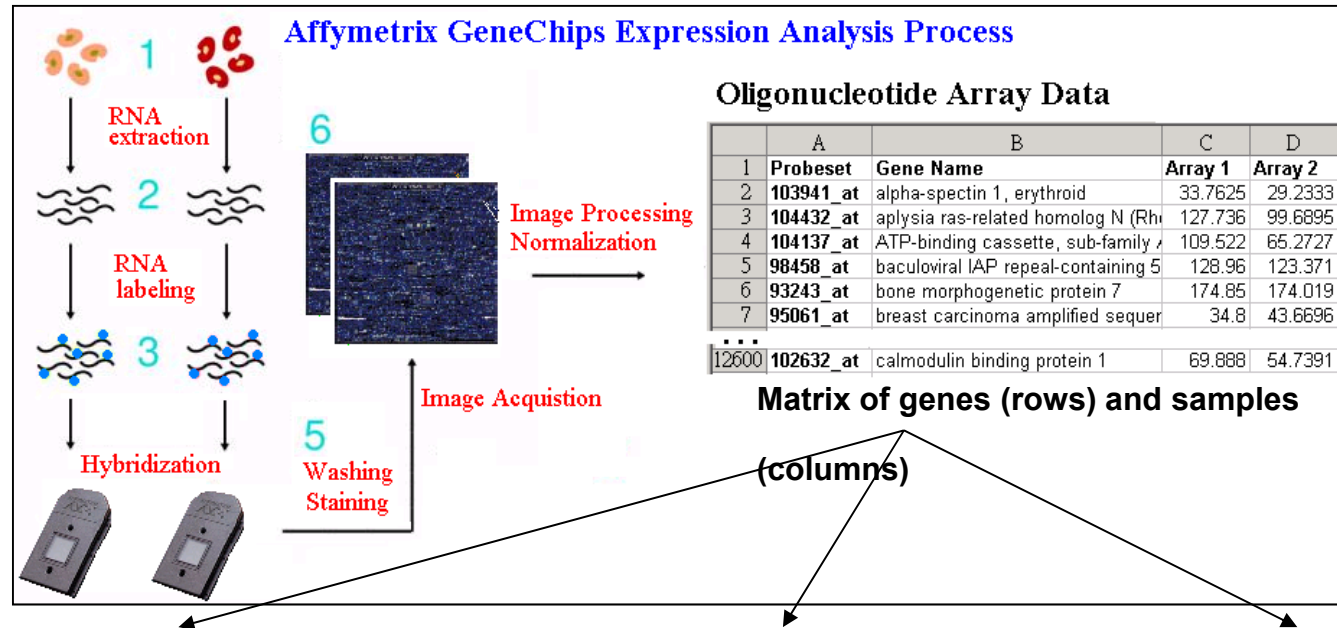
SPACE PHOTO Galaxy, Sun's Ir

LATEST VIDEO Stuck Mars Rov

Most Viewed News

ADVERTISEMENT

基因晶片資料統計分析



Biological Relevance

Software Development



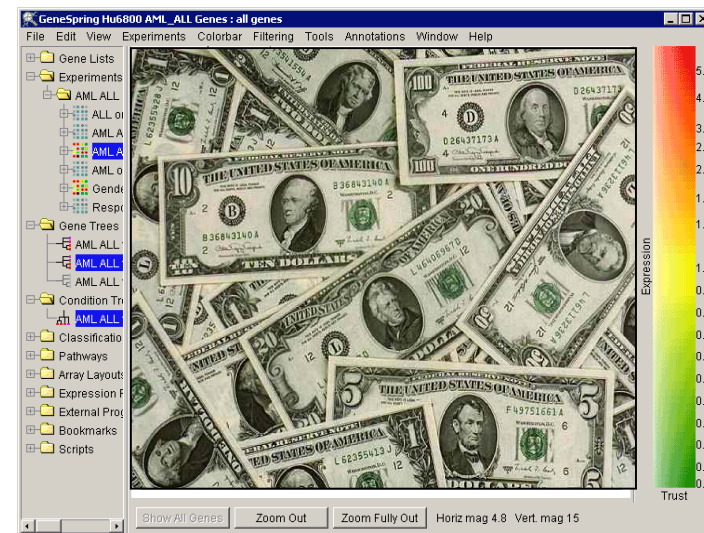
Generalized Association Plots



DNA-Chip Analyzer



GenePix Pro



GeneSpring GX



第五波科技浪潮: Big Data

大數據、海量資料、巨量資料



2013.05.04

遇見未來城市 / 21世紀原油
大數據時代的台灣機會

海量資料開始備受關注，是全球知名諮詢公司麥肯錫於2011年，發布了《**海量資料：創新、競爭和生產力的下一個新領域**》報告，宣布「大數據時代」來臨。

McKinsey&Company

Report | McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

May 2011 | by James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers

Download [Executive Summary](#) PDF-922KB [Full Report](#) PDF-6MB [Kindle](#) MOBI-4MB [eBook](#) EPUB-3MB



The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with

McKinsey Global
Institute

第一波是大型電腦，第二波是個人電腦，第三波是網路，第四波則是社群媒體。

相關報導



Master 60 60分鐘與全球大學齊讀 大師輕鬆讀 NO.506

大數據

看龐大資料如何靠分析顛覆一切

8項特點
 1. 處理全部資料，不僅是樣本
 2. 降低對精確度的要求
 3. 不受原因、只在模式
 4. 如何過濾信息？
 5. 我們可以認為人們的傾向與行為嗎？
 6. 我們會成為資料的犧牲品嗎？

8項啟發
 1. 數據化非世界
 2. 資料將成為關鍵資產
 3. 分析將改變商業知識

Big Data
A Revolution That Will Transform How We Live, Work, and Think

原著：維克多·麥爾奇伯格 Viktor Mayer-Schönberger、肯尼希·漢基那 Kenneth Cukier

定價：NT120

大師輕鬆讀 Master 60 449

大數據時代 創造大業績

120位頂尖銷售領袖證實有效的成長全攻略

Sales Growth
Five Proven Strategies From the World's Sales Leaders

原著：湯馬斯·鮑加納 Thomas Baumgartner / 賀馬米·哈格米 Homayoun Hatami / 羅·范德·亞克 Jon Vander Ark

Big Data 10億客戶的超級武器 專訪日本網天創人三木谷浩史：敵人的敵人 就是我的朋友！

數位時代

Big Data

數字煉金

2012 台灣 Web 100 迎接 大商務元年!

威名百貨 (Walmart) 為何全球最大零售業者的關鍵是什麼？歐洲流行服飾品牌Zara為什麼能夠一躍而出兩次新品？Google 如何管理龐大流量，從數據中找到什麼？台灣 E-commerce 轉賣及不斷創新管理能力的從何而來？所有這些的背後，都在企業所擁有的數據沉澱，即無限無限的大資料中，創造出倍數的數據價值，你準備好知道了嗎？

3 定價：NT\$150

天下文化 www.Bookzone.com.tw 如您無法看到完整訊息，請點選

大數據 BIG DATA

每十年，總是有極少數的書能改變你看待一切的方式

Big Data 時代來臨 誰能在龐雜的訊息中掌握關鍵 誰就能取得解決問題的先機

PNN 公視新聞議題中心 PTS NEWS NETWORK

首頁 P 頭條 P 今選 P 專題 P 評 P 部落 P 卡秋 P 全球

P 今選 他們想知道你的一切~巨量資料時代的省思 2013/04/29

TechOrange 科技報橘 techorange.com

Blog > 創新與應用 > 今天不談商機，我們必須對 BigData 將顛覆人類思維與生活方式有這些感知！

今天不談商機，我們必須對 **BigData** 將顛覆人類思維與生活方式有這些感知！

[創新與應用] 於 2013-07-11 11:38:14 發布 Tags: Big Data

Blog > 創新與應用 > 我完全明白要做 Big Data，但是到底該怎麼開始？

我完全明白要做 **Big Data**，但是到底該怎麼開始？

[創新與應用] 於 2013-06-18 14:59:54 發布 Tags: Big Data

Business Next 數位時代

創業小聚 網路 科技 人物 品牌與設計 綠能

熱門關鍵字: 12/12創業之星Demo Show 智慧眼鏡 iPad 數位工具術 行動支付來了！

目前位置: 首頁 > Topics and Links新聞精選 > 【MMDays】Big Data 大數據 大商機 大未來

【MMDays】Big Data 大數據 大商機 大未來

數位時代網站 | 撰文者: Ms. Valentine 發表日期: 2012-12-24



(1) 啤酒、尿布、星期五

(2) 颶風、啤酒、草莓夾心酥

48/118



- ❑ **購物車分析**: 現在也普遍的存在於一般電子商務中。這個尿布配啤酒也成了經典的 data mining 實例，經常被用來說明最基礎且直觀的**關聯法則**(Association rule)演算法。
- ❑ Wal-Mart在全球超過200萬名員工，總共有110個超大型配送中心，每天處理的資料量超過**10億筆**。
- ❑ Wal-Mart使用的Big Data模式，已經從「**挖掘**」顧客需求進展到要能夠「**創造**」消費需求。



Chih-Cheng Liang [Follow](#)
chihchengliang@gmail.com
2 days ago · 2 min read

49/118

莫再提了！啤酒尿布是都市傳說



台灣某知名超市，啤酒旁邊很直覺的放了零食。那尿布旁放

資料分析領域流傳這麼一則故事：

某家超市（大多時候是 Walmart）使用某種分析方
析）分析資料庫資料，發現每週五晚上啤酒與尿布銷
原因是星期五晚上年輕父親會幫家裡買尿布，並順便為

這個故事常被引用來強調資料分析的價值。但事實上源
Google 也只會告訴你各種版本的都市傳說。



答案揭曉：放洗手乳

Big Data有多巨大？

2013年1月號《遠見雜誌》3分鐘輕鬆了解Big Data海量資料

- 「**過去3年**裡全世界產生的資料量，比以往**4萬年的總和**還要多。」
- 2009年地球上所有的資料總共**0.8ZB**，2011年全球創造與複製的資料量已經暴增到**1.8ZB**，所有的資料，得用575億部32GB的iPad才能裝完。這也相當於全台灣所有的人連續36萬246年，每人每分鐘在噗浪發表3篇文章。這等於一個人每天24小時不停觀看現今世上超過2千億部的2小時HD電影，必須花上4700萬年才能看完。

Kilobyte(KB)=1024B 相當於一則短篇故事的內容。

Megabyte(MB)=1024KB 相當於一則短篇小說的文字內容。

Gigabyte(GB)=1024MB 相當於貝多芬第五樂章交響曲的樂譜內容。

Terabyte(TB)=1024GB 相當於一家大型醫院中所有的X光圖片資訊量。

Petabyte(PB)=1024TB 相當於50%的全美學術研究圖書館藏書資訊內容。

Exabyte (EB)=1024PB；5EB 相當於至今全世界人類所講過的話語。

Zettabyte(ZB)=1024EB 如同全世界海灘上的沙子數量總和。

Yottabyte(YB)=1024ZB 相當於7000位人類體內的微細胞總和。

Big Data從哪裡來？

2013年1月號《遠見雜誌》3分鐘輕鬆了解Big Data海量資料

我們在使用這些**服務**的同時，
也隨時在**創造**新的數據。

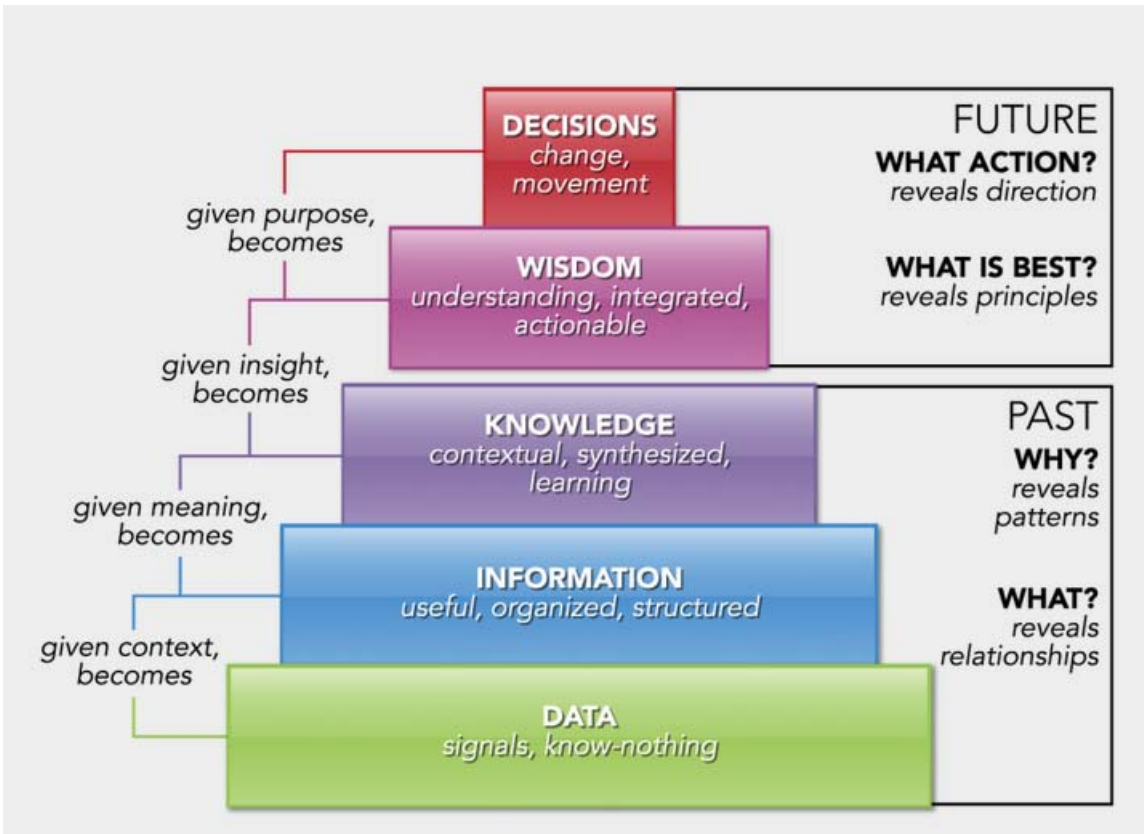


- 海量資料不單指文字及數字等「**結構性資料**」，也包含了照片、影音，座標GPS、心電圖、X光片等「**非結構化資料**」。
- 中國電子商務龍頭**淘寶**，在2010年時便已經擁有3.7億位會員，8.8億件商品，每日交易達數千萬筆，單日便可產生超過50TB的資料。
- 每天流向**Facebook**與**Twitter**的資料量，就多達3億張照片、25億則發文、27億按讚數。



為何「大數據」重要？

- 資料在以前被視為垃圾，現在卻被視為資產，這是大數據時代的第一個變革。2012年初的瑞士達沃斯論壇上，一份題為《大數據，大影響》(Big Data Big Impact) 的報告宣稱，數據已經成為一種新的**經濟資產類別**，就像貨幣或黃金一樣。



<http://www.pursuant.com/blog/tag/dikw-model/>



NVIDIA悄然壟斷算力：人工智能背後的新帝國

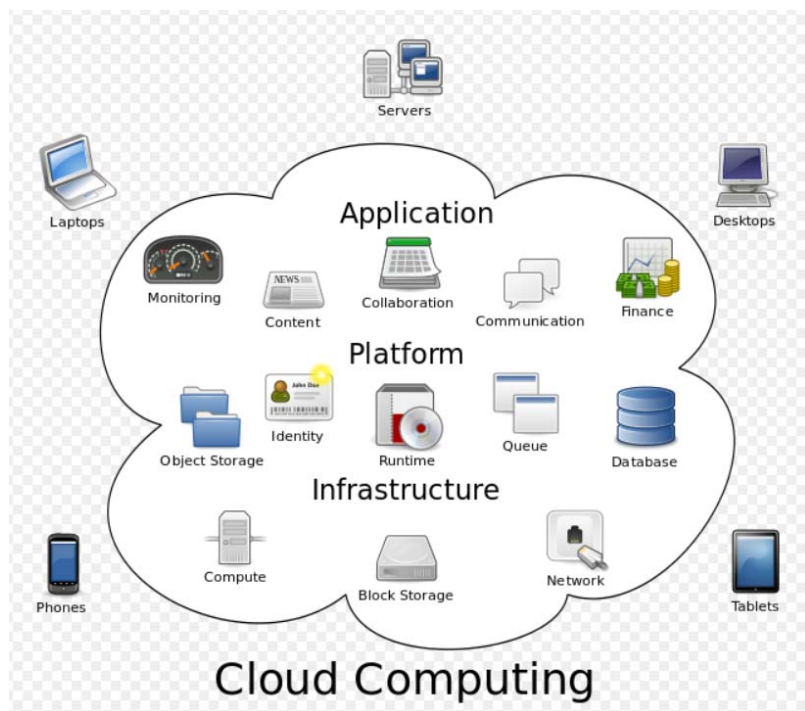
<https://hao.cnyes.com/post/37888>

全球前十大企業市值排行		
排序	企業名稱	市值金額
1	微軟	3.04兆美元
2	蘋果	2.84兆美元
3	沙烏地阿拉伯國家石油公司	2.06兆美元
4	輝達	1.83兆美元
5	Alphabet	1.82兆美元
6	亞馬遜	1.78兆美元
7	Meta	1.21兆美元
8	波克夏公司	8,652億美元
9	禮來藥廠	7,189億美元
10	特斯拉	6,010億美元

資料來源：截至2月14日，彭博資訊顯示的全球前十大企業市值排行與其市值金額。編譯葉學均 / 製表 經濟日報

Big Data 特性

- Big Data有三種特色 (3V) :
 - 巨量 (Volume) 。
 - 即時性(Velocity) : 串流至企業便須立即使用，方能發揮其最大價值。
 - 多樣性 (Variety): 資料包含各類非結構化的資料：諸如文字、音訊、視訊、點擊串流 (click stream)、日誌檔等等。
- 其它: 價值 (Value)、變化性 (Variability)、複雜度 (Complexity).

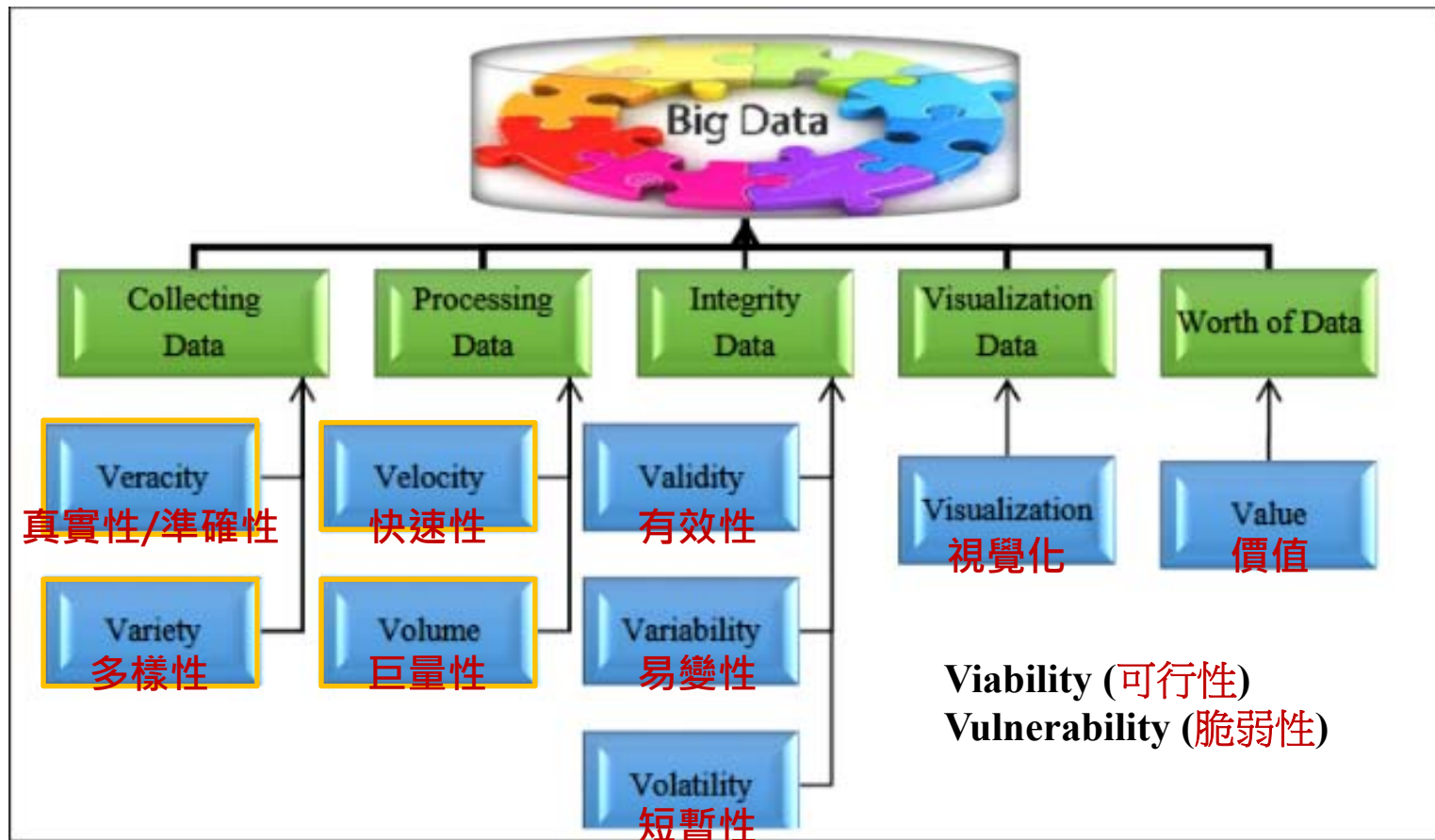


- 大數據成為熱門題材且被視為下個金礦，就是因為**雲端運算**這個趨勢已經被普遍接受。
- 而雲端運算又是解決大數據特性的最佳工具。
- 沒有雲端運算，無法容納及處理大量多元化的資訊，大數據大不了，也快不了，自然無法提供精準的即時分析。

Big Data: The Era of 9 Vs

Visualization:

- Visualization will be key to making big data an integral part of decision making.
- Visualization will be the only way to make big data accessible to a large audience.
- Visualization will be essential to the analysis of big data so it can be of highest value.



Categorization of Big Data V's

<http://blogs.systweak.com/2017/03/big-data-vs-represents-characteristics-or-challenges-of-big-data/>

案例 (1): 醫療

- 相信每個人都有過相同經驗，拖著不舒服的身子，在醫院等了將近一小時，好不容易輪到你時，醫生只問了幾句話，就立刻開好處方籤叫你回家好好休息。有時你不禁會懷疑，醫生真的可以在這麼短的時間內，診斷出我的病因嗎？
- 在美國，每年有**超過150萬件醫療糾紛**，絕大多數都因為醫師的診斷沒有實證基礎。



- IBM推出華生(Watson) 機器人。華生運用**大量的臨床病例**，可在短時間內分析可能的結果，並協助醫生做出「治療建議」，大大減少醫生疏忽的機會。
- 華生不僅可以即時讓醫師參考診斷與治療方式，針對可能的疾病做深入的問診，更可以有效減少醫療糾紛，縮小判斷誤差。

Watson配備了 90 台 IBM Power 750 伺服器，能在 3 秒內處理完等同於約 100 萬冊圖書的數據量(圖片來源：cnet)



案例 (2): 2012年美國總統大選

TechOrange
科技報橘
techorange.com

Blog > 創新與應用 > 連歐巴馬都靠資料分析贏了，你還要小看 Big Data ?

連歐巴馬都靠資料分析贏了，你還要小看 Big Data ?

[創新與應用] 於 2012-11-14 11:28:08 發布 Tags : Big Data、美國大選



關於作者

米蘭達



歐巴馬團隊能精準針對不同族群進行**客製化募款**並進行**針對性的發言**。同時，歐巴馬團隊可對主要選區每晚模擬超過六萬次選舉了解選情變動，投票前就對勝負有所把握。

隨後，這項技術也被用來**預測選情**，針對各州勝出的可能性，分配適當的資源，最終歐巴馬獲得**連任**。

- 在2012年，美國總統歐巴馬尋求連任，他的競選團隊花了**兩年**的時間，蒐集分析資料：
- 競選團隊發現到，影星喬治克隆尼(George Clooney)對美西40到49歲的女性，有非常巨大的吸引力，且他們是最有可能為了與喬治克隆尼和歐巴馬**共進晚餐**，而**自掏腰包**的族群。
- 同樣的在其他區域，也有發現某些族群喜歡**名人**、喜歡**聚餐**，重要的是，他們願意掏出腰包來與他們的偶像共進晚餐。
- 競選團隊舉辦了多場類似的**募款餐會**，為歐巴馬籌集到**10億美元**的競選資金。



歐巴馬競選團隊的工作情形。(圖片來源：《TIME》)

案例 (3): 預測女生是否懷孕?



- 美國的Target連鎖超市**分析大量的顧客消費紀錄**，佐以其他個資，能推算哪些女客很可能懷孕了，然後針對她們做個人化行銷。
 - 有次，一位中年男子怒氣沖沖跑到店裡，責問超市竟然寄嬰兒用品優惠券給他的**未成年女兒**。
 - 但幾天後，本來要求超市道歉的他，卻反過來道歉了，因為和女兒懇談後，他才知女兒真的懷孕了——超市比他**早知道**。
-
- Target搜尋孕婦，可不只是為了賣尿片奶瓶這樣單純，而是行銷人員知道，從懷孕中期開始，胎兒父母親原有的**購物習慣會鬆動**，易受行銷手法左右，如果這時抓住他們，他們在未來幾年都會是絕佳主顧。
 - 資料的處理系統可能會自作聰明，錯誤解讀，而不利於當事人；又或者系統抓選某些資料種類，只是因為它們較方便，而不是它們較重要，易犯見樹不見林的毛病。巨量資料分析還在初期階段，將來的進展尚未可知，要冷靜**省思科技迷信**。

案例 (4): 文字與語意分析



TechOrange
科技報橘
techorange.com

Blog > 創新與應用 > 電影公司投資 10 部，賣 10 部的成功關鍵：Big Data !

電影公司投資 10 部，賣 10 部的成功關鍵：Big Data !

[創新與應用] 於 2013-04-24 16:00:32 發布 Tags: Big Data、Hollywood、大數據、好萊塢、海量數據、電影



關於作者

Wired.big.data



美國超級盃美式足球賽的廣告費，每 30 秒要價高達 400 萬美金。但是對企業主來說，如何讓天價般的廣告支出發揮**最大效益**？

IBM 為片商提供的語意分析工具，在 2012 年超級盃舉行前一個月，就開始**監測社交媒體對特定電影的討論**，以每一分鐘為單位，蒐集到計 11 億條推特推文、570 萬篇部落格以及論壇貼文、350 萬條相關訊息，擷取其中關鍵訊息，分析主題，判斷網友意向，歸結出網友對電影預告片的看法與評價。

這些文字被擷取下來之後，首先會進行分類、統計，再與觀眾性別、職業、是否為常上電影院或愛看漫畫等**個人特質交叉分析**。分析結果顯示，18% 的網友對《復仇者聯盟》正向情緒，預告播出後甚至達到 35% 的高峰，一段時間以後也一直維持在 18% 的水準，顯示這將近兩成的人有意願觀賞這部電影。此外，《復仇者聯盟》相關討論多半集中在預告片上，且壓倒性地給予正面評價，這意味著預告片拍得好，廣告的播出十分有效。

果然，《復仇者聯盟》2012 年 5 月上片以後，美國本土首週票房就達到兩億美金，幾乎追平製作成本；截至同年 9 月為止，總票房達到 6 億 2 千多萬美金，成為**世界電影史票房排行榜第三名**，僅次於《阿凡達》、《鐵達尼號》。

其它應用範例

- **臺灣「e化天眼計劃」系統**，在2013年八里雙屍案偵辦時，調閱警方與民政系統共305具監視器影像資料時間長達1個月，將謝依涵、呂炳宏、張翠萍、陳進福等人的車輛軌跡與通聯紀錄輸入後，將零星、片段資料，分析影像、比對車號進行交叉比對。鎖定死者失蹤前、屍體被發現後，失蹤當晚究竟是赴誰的約會？警方表示，「人證會有記憶上的遺漏或誤差」，但科學辦案完全以證據說話。
- **日本「N系統」(自動車ナンバー自動読取装置)**，可以全年無休對道路上行駛的車輛牌照拍照存檔，記錄下行駛的路線與時間。同時在資料庫中比對被通緝中嫌犯或竊賊車的車牌號碼，如果發現符合，該系統立即通知在外巡邏的警員(配合攜帶式接收裝置)，及時對該車輛進行攔截圍捕。另外、東京都警視廳也配合「3D臉部自動辨識系統」辨識人貌鑑定，警察如需調查案情之時，就能夠調出資料庫資料，進行交叉比對，篩選出可疑的犯罪目標。
- **預測爆發流感**: 2009年Google工程師依據搜尋引擎的查詢資料，預測H1N1在美國冬天即將爆發流感，資訊速度比起聯邦政府疾病管制局(CDC)提前。並且在《自然》期刊上發表「Detecting influenza epidemics using search engine query data (使用搜尋引擎查詢資料檢測流感疫情)」的文章。
- **找出有致命危險的人孔蓋**: 許多老舊的人孔蓋是不定時炸彈，但其中有些更容易闖禍，快速有效地尋找出它們悠關公共安全；
- **預測機票的價格**: 這在美國已成了一個網站Farecast.com，讓使用者可以依據機票浮動的機率決定是否下單，Farecast.com後來被微軟併購；
- **規劃快遞的送貨路線**: 尋找最少停停走走的路線，以節約能量；
- **防盜系統**: 能從駕駛人的臀部形態判斷是否為車主的防盜系統，甚至還能判斷車主有沒有酒駕或打瞌睡.....
- **更多...**



行政院國家科學委員會
National Science Council

您在這裡: 首頁 >>> 最新消息

國科會企劃處公開徵求巨量資料(Big Data)推動規劃先導計畫

一、計畫說明:

國科會企劃考核處(以下簡稱本處)有鑑於網路世界蓬勃發展,ICT應用已快速產生並累積大量資料,從大量資料萃取價值的需求,已帶來知識與技術創新的機會,因此依據「補助產業前瞻技術計畫作業要點」之規定,徵求以巨量資料(Big Data)為主題之先導計畫。

(二)執行期間:102年12月1日至103年11月30日止。



Google Trend



維京人酒吧 Viking Bar

VIKING BAR 商業科技 經濟金融 大眾消費 議題廣場 ABOUT US COM

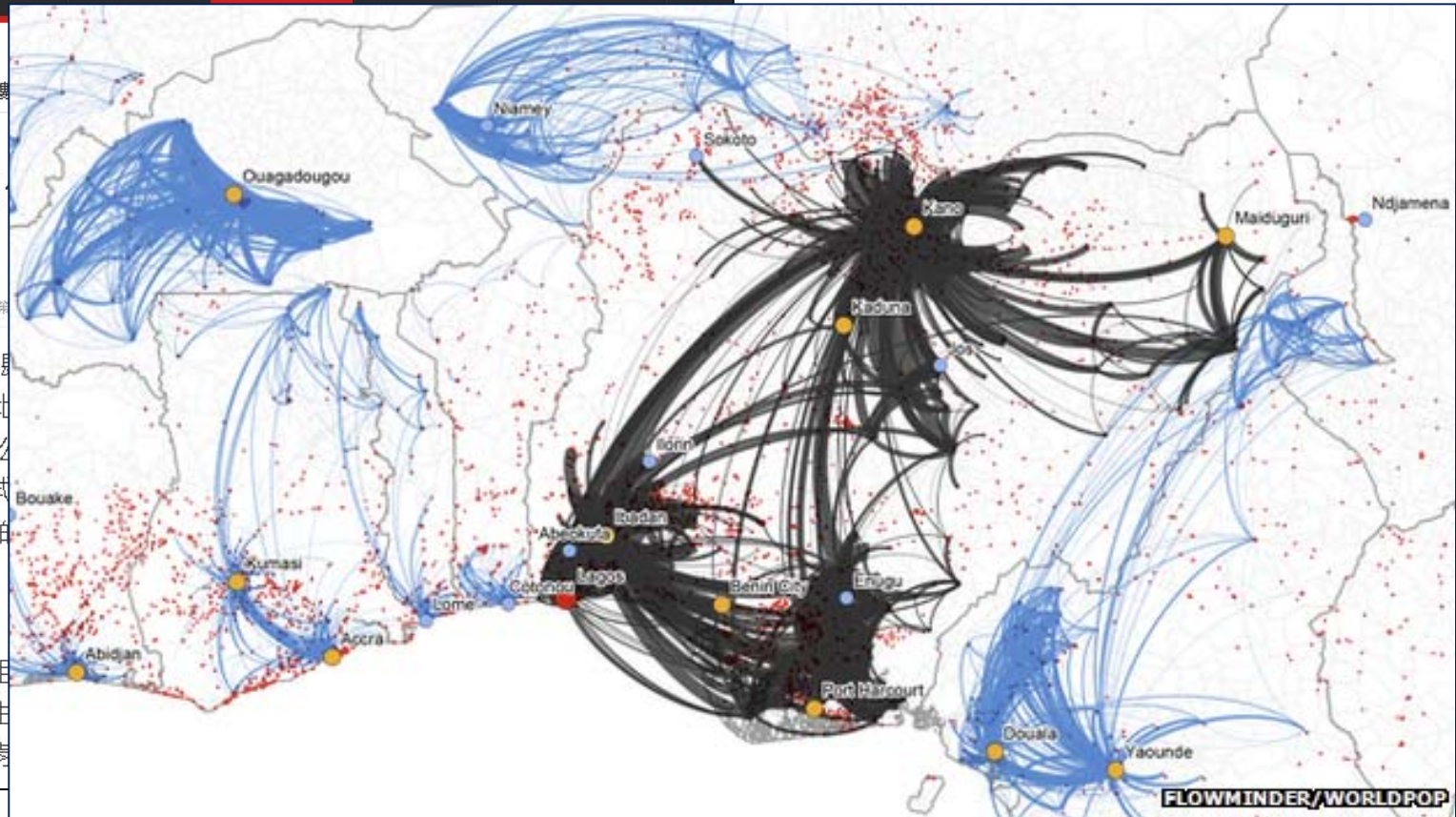
Home / 大眾消費 / 用大數據

用大數據畫出 做疾病預測

Posted by: Alison Yeh in 大眾消費, 第

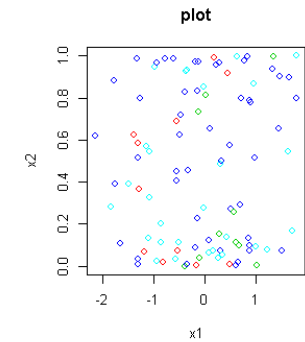
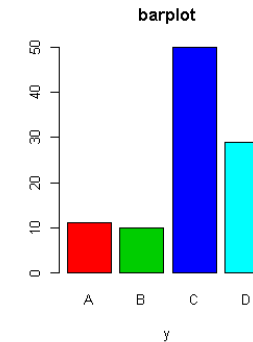
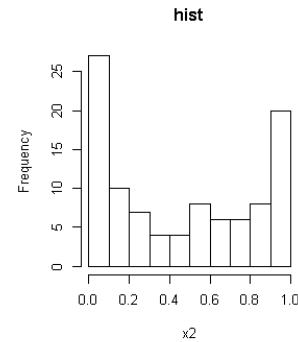
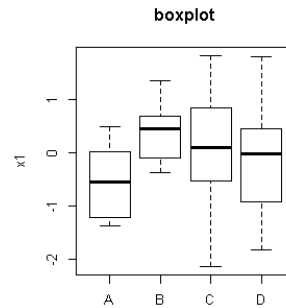
伊波拉病毒 (Ebola) 是
比瑞亞及奈及利亞，此
界衛生組織 (WHO) 公
命。此次疫情已被正式
僅發佈兩次，2009年的

無數援助機構、國際組
試圖控制疫情，除了生
越來越多科學家表示肯

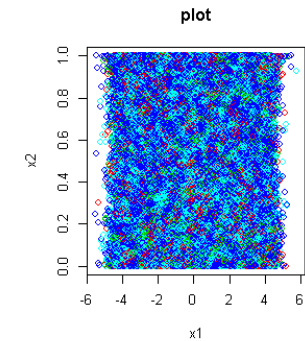
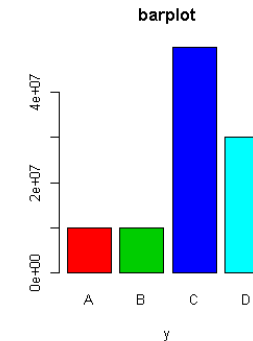
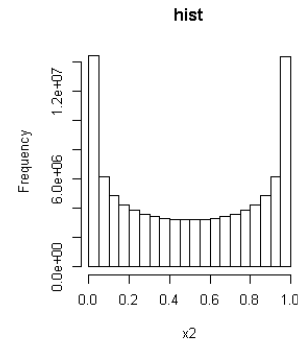
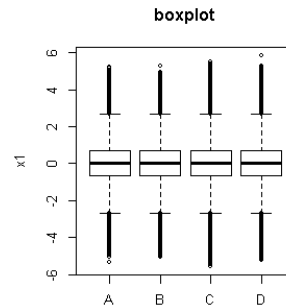


The Challenge of Visualizing Big Data

```
> n <- 1e+02
```



```
> n <- 1e+08
```



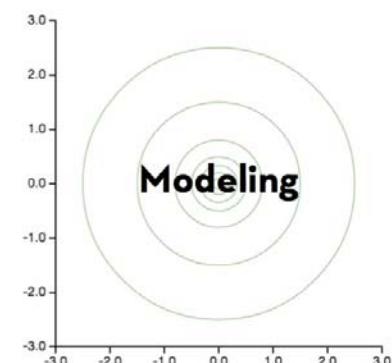
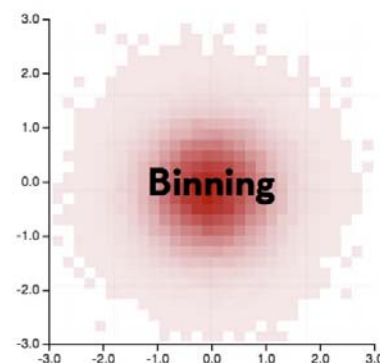
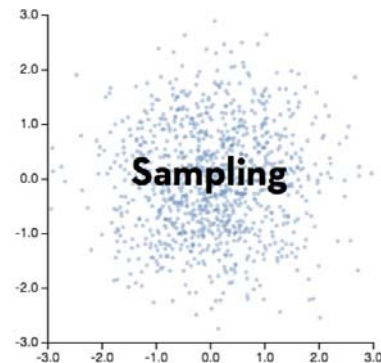
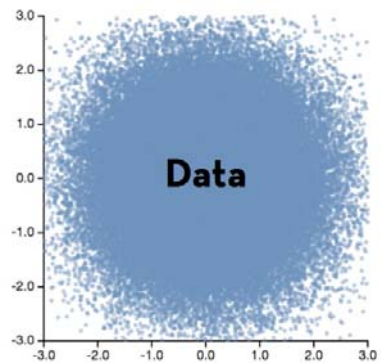
a large p?

**Two principles:
Look at Less Data;
or Look at Data Faster**

How Can We Visualize and Interact with Billion+ Record Databases in Real-time?



- ❑ Two central challenges:
 - ❑ need to keep visualizations **perceptually effective** regardless of the number of input data points.
 - ❑ need to support **real-time interaction** to enable rapid and iterative exploratory analysis.
- ❑ Perceptual and interactive scalability should be limited by the chosen **resolution of the visualized data**, not the number of records.





ggplot2.SparkR: Rebooting ggplot2 for Scalable Big Data Visualization

ggplot2.SparkR
Welcome to ggplot2.SparkR

[View project on GitHub](#)

Overview
ggplot2.SparkR is an R package for scalable visualization of big data represented in Spark DataFrame.
It is an extension to the original ggplot2 package and can seamlessly handle both R data.frame and Spark DataFrame with no modifications to the original API.

Installation

SparkR Installation

Build Spark
Build Spark with **Maven** and include the `-PsparkR` profile to build the R package. For example to use the default Hadoop versions you can run

```
build/mvn -DskipTests -PsparkR package
```

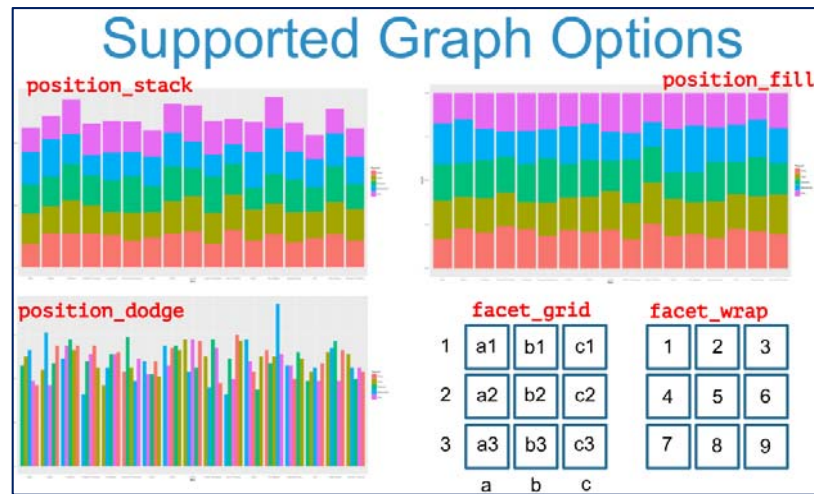
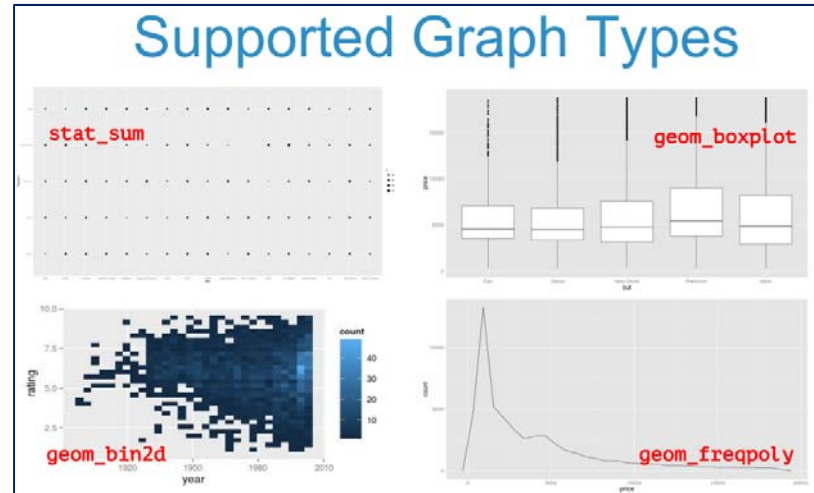
Using SparkR from RStudio
If you wish to use SparkR from RStudio or other R frontends you will

Download zip file
Download tar.gz file

is maintained by SKKU-SKT.

This page was generated by GitHub Pages using the Architect theme by Jason Long.

- Home
- Hello ggplot2.SparkR
- Supported Plot Types



<http://skku-skt.github.io/ggplot2.SparkR/>

數學統計人才



HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR TIMES TOPICS

The New York Times

Technology

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

Search Technology

Inside Technology

Bits
Blog

Internet Start-Ups Business Computing Companies

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to “computer and math stuff” that was part of

Enlarge This Image



Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

Multimedia



PH.D. in computer science with focus on artificial intelligence and text analytics, M.I.T.
CAREER He is a research scientist at I.B.M. who uses computing and modeling to extract knowledge patterns from text, video and audio data.

EDUCATION Bachelor's degree in computer science, Cornell; Ph.D. in computer science, M.I.T., a MacArthur Fellow.
CAREER He is a professor of computer science at Cornell who mines huge data sets on the Web to analyze social behavior. Research code available at <http://www.cornell.edu/~jkleinberg/>.

Graphic
Data Sleuths in an Internet Age

“People that Indiana Jo really do is Now Ms. C of digging where she mounds of search eng Ms. Grime who are ch dronish nu increasing

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

The rising stature of statisticians, who can earn \$125,000 at top companies in their first year after getting a

Hal Varian explains why statisticians will be the sexiest job in the next decade (1



spencer11111 於 2009-09-15 上傳
I keep saying that the sexy job in the next 10 years will be statisticians, said Hal Varian, chief economist at Google. And I'm not kidding. People think I'm joking, but who would've guessed that computer engineers would've been the

是新用戶嗎？請註冊 | 登入 | 服務說明

設Yahoo!奇摩為首頁

YAHOO! 新聞
奇摩

Q 搜尋

新聞首頁 政治 財經 社會 地方 影劇 運動 國際 生活 文教 健康

資訊3C 科學發展 自然環境 科技熱門

雲端崛起 數學統計人才搶手

中時電子報 作者：記者何英煒/台北報導 | 中時電子報 - 2012年10月3日 上午5:30

字 大字

工商時報【記者何英煒/台北報導】

趨勢科技創辦人張明正昨(2)日表示，能源和科技是人類社會200年來進步的源頭，現今的「資料」(Data)正是當年的「石油」，誰能夠掌握，就能稱霸一方。

張明正從趨勢科技退休後，在台大及輔大捐助不少雲端課程，面對近來各界討論台灣人才缺乏的議題，張明正表示，雲端及海量資料處理是未來趨勢，他預言未來3個領域的達人將成職場當紅炸子雞。

首先，目前全球的資料呈現爆炸式成長，企業繼需要CEO(執行長)、CIO(資訊長)之外，還迫切需要懂得資料分析的科學家(Data Scientist)。他表示，資料分析科學家不僅要懂得關聯式資料，更要了解如何分析非結構的資料，所謂的非關聯式資料，指的是在網路上的瀏覽行為、上傳的評論，以及社群網站所顯示的社交關係等資料。

張明正說，國外有企業特地招募數學系博士生，派他們到哈佛商學院上2年MBA課程，為的就是借重這些人的長才，為其分析資料、建立商業模式。

輔大數學系畢業的張明正表示，過去數學系、物理系都是冷門科系，但未來企業需要能夠分析資料、做出模型的專才，數學系、物理系及統計系都將躋身熱門科系。

其次，張明正認為，台灣過去一向擅長做硬體，但不可諱言，硬體OEM的毛利很低。但這也未嘗不是個優勢，未來市場上需要整合軟硬體的人才，尤其是懂得個人裝置加上雲端運算的整合的人才。

第三種人才，張明正認為，可補足Hadoop工具及現有資料的人才。換言之，過去資料分析的人才都需要熟悉SAP等傳統軟體才能夠找到工作。但未來熟悉海量資料、平行處理、非結構性資料等運算的Hadoop平台工程師，將成為市場搶手貨。尤其是可為企業同時處理關聯性及非結構性資料的工程師。

二十一世紀最性感的職業



THE MAGAZINE

BLOGS

VIDEO

BOOKS

CASES

WEBINARS

COURSES

Guest

Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE

October 2012



ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here to register for FREE access** »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (87)



RELATED

Executive Summary

ALSO AVAILABLE

- Buy PDF

- 《哈佛商業評論》宣布，「Data Scientist，數據科學家」是二十一世紀最性感的職業。
- 從計算機科學衍生而成，關鍵在於成功管理數據。
- 從學術邁向新職業，支撐海量數據研究與應用。
- 所謂數據科學家就是：**運用數據和科學，創造新東西的人**
- **統計知識 + 科學經驗 + 工程經驗**，成就性感數據科學家

<http://techorange.com/2012/10/05/data-scientists-the-definition-of-sexy/>

數據科學家



Microsoft

搜尋結果

Microsoft Press 台灣微軟新聞室

2013 年新聞

海量資料帶動下一波科技人才需求 「資料科學家」正夯！

微軟攜手中華資料採礦協會 產學合作打造新科技人才黃金搖籃

...

(2013 年 10 月 09 日, 台北) 「究竟誰會是 2013 年金馬獎最佳男女主角?」、「黃金、基金市場行情起起伏伏, 何時才是最佳進場時機?」爆炸性成長的資料是「海量資料」時代的特性, 但要如何從資料巨浪中分析出市場趨勢、未來商機, 除了可靠的資料庫平台之外, 「資料科學家」專業的數據分析更是關鍵, 下一波熱門職缺「資料科學家」也就因應而生。《哈佛商業評論》也曾提到, 21 世紀最誘人的職業, 就以資料科學家莫屬! 台灣微軟瞄準未來企業對海量資料人才的需求, 與中華資料採礦協會攜手合作, 結合產業和學界的師資與資源, 共同成立「微軟資料科學研究院」, 培育跨領域及跨專業的「資料科學家」。

這是個數據說話的時代! 懂得與資料對話就等於掌握了黃金礦脈

台灣微軟營運暨行銷事業群總經理康容說: 「資料是企業創造金脈的來源, 如何將 Big Data 海量資料轉化成企業可以做為營運依據的 Big Decision 大決策, 考驗著資料科學家的實力; 當企業身處資訊爆炸的時代, 如何從漫天的資料黃沙中挖掘出最具價值的黃金, 就是資料科學家被賦予的使命! 」

<http://www.ithome.com.tw/news/109308>

Business Next 數位時代

創業小聚

網路

科技

人物

品牌與設計

綠能

熱門關鍵字: 12/12創業之星Demo Show 智慧眼鏡 iOS 7 數位工具術 行動支付來了!

目前位置: 首頁 > 軟體 > 微軟在台開設資料科學研究院, 培育Big Data人才

微軟在台開設資料科學研究院, 培育Big Data人才

數位時代網站 | 撰文者: 趙那竹 | 發表日期: 2013-10-09

36 1 9 19
讚 推 文 推 文 Share

微軟和中華資料採礦協會合作, 將在台開設免費的資料科學研究系列課程, 鼓勵IT或非IT背景的工作者都能參加, 一堂課將招募80至100人, 預計十堂課可幫助台灣培育800名大資料 (Big Data) 人才。

台灣微軟營運暨行銷事業群總經理康容指出, 大資料是下世代最重要的四大趨勢之一, 「資料科學家」更是因此趨勢而產生的新機會。哈佛商業評論也曾提到, 21世紀最誘人的職業, 就是資料科學家。台灣微軟瞄準此需求, 和中華資料採礦協會合作, 提供微軟的分析工具, 並結合學界專家師資, 成立「微軟資料科學研究院」, 希望及早在台培育更多資料科學人才。

數據科學家



- 想學機器學習應該準備具備哪些數學知識？

Posted on 2017/11/09

<https://buzzorange.com/techorange/2017/11/09/how-to-learn-machine-learning/>

- 數學不學好也想當數據科學家？想轉型，你會需要這份攻略

Posted on 2018/02/11

<https://buzzorange.com/techorange/2018/02/11/data-scientists-have-to-improve-math/>

- 不学好数学也想当数据科学家？不存在的

原创：文摘菌 大数据文摘 2月6日

<https://mp.weixin.qq.com/s/3d5UL3Hajl2-0Z6QA6kNiA>

- 數據科學家們的無知困境：只會 coding，卻忘記用更大的格局思考問題

Posted on 2017/11/14

<https://buzzorange.com/techorange/2017/11/14/you-should-think-bigger/>

- 數據科學家必讀的五本書：重要的不是會打 Code，而是背後的資料邏輯思維

Posted on 2018/09/11

<https://buzzorange.com/techorange/2018/09/11/these-non-code-book-worth-reading-for-programmers/>

資料科學家的專業能力 與人格特質

專業能力

1. IT素養
2. 統計知識
3. 產業知識

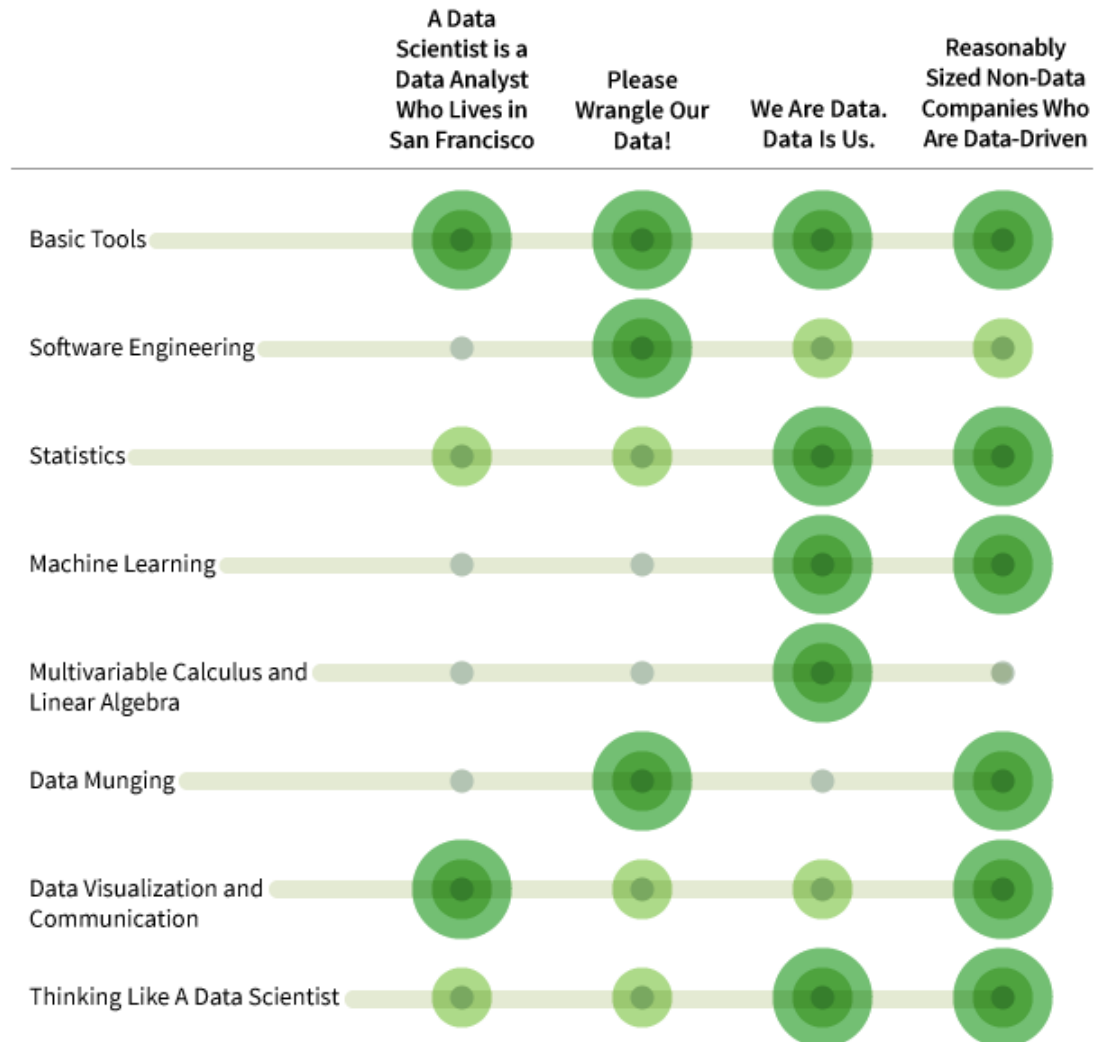
人格特質

1. 喜歡問問題及解決問題
2. 追求作業流程的改善
3. 不被過往經驗所限制

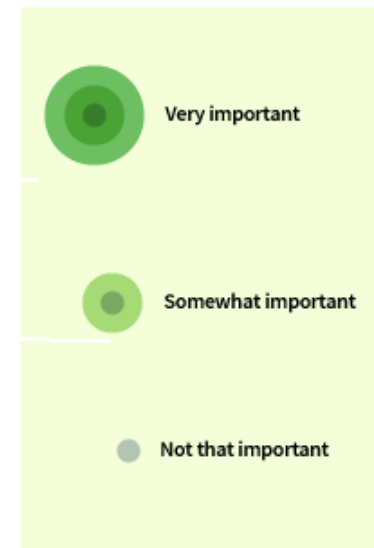
iThome



8 Data Skills to Get You Hired



4 Types of Data Science Jobs

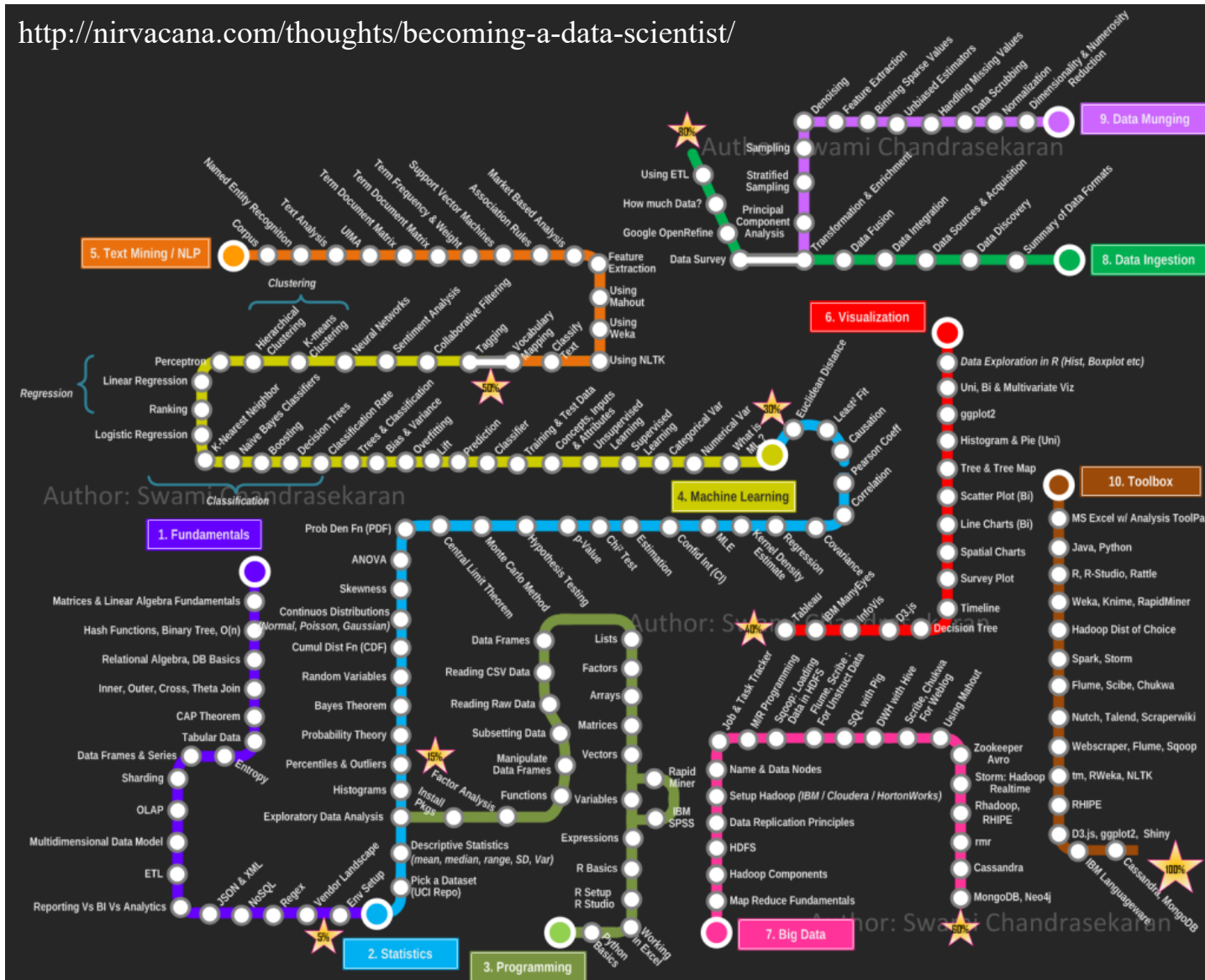


UDACITY

Becoming a Data Scientist – Curriculum via Metromap

69/118

<http://nirvacana.com/thoughts/becoming-a-data-scientist/>



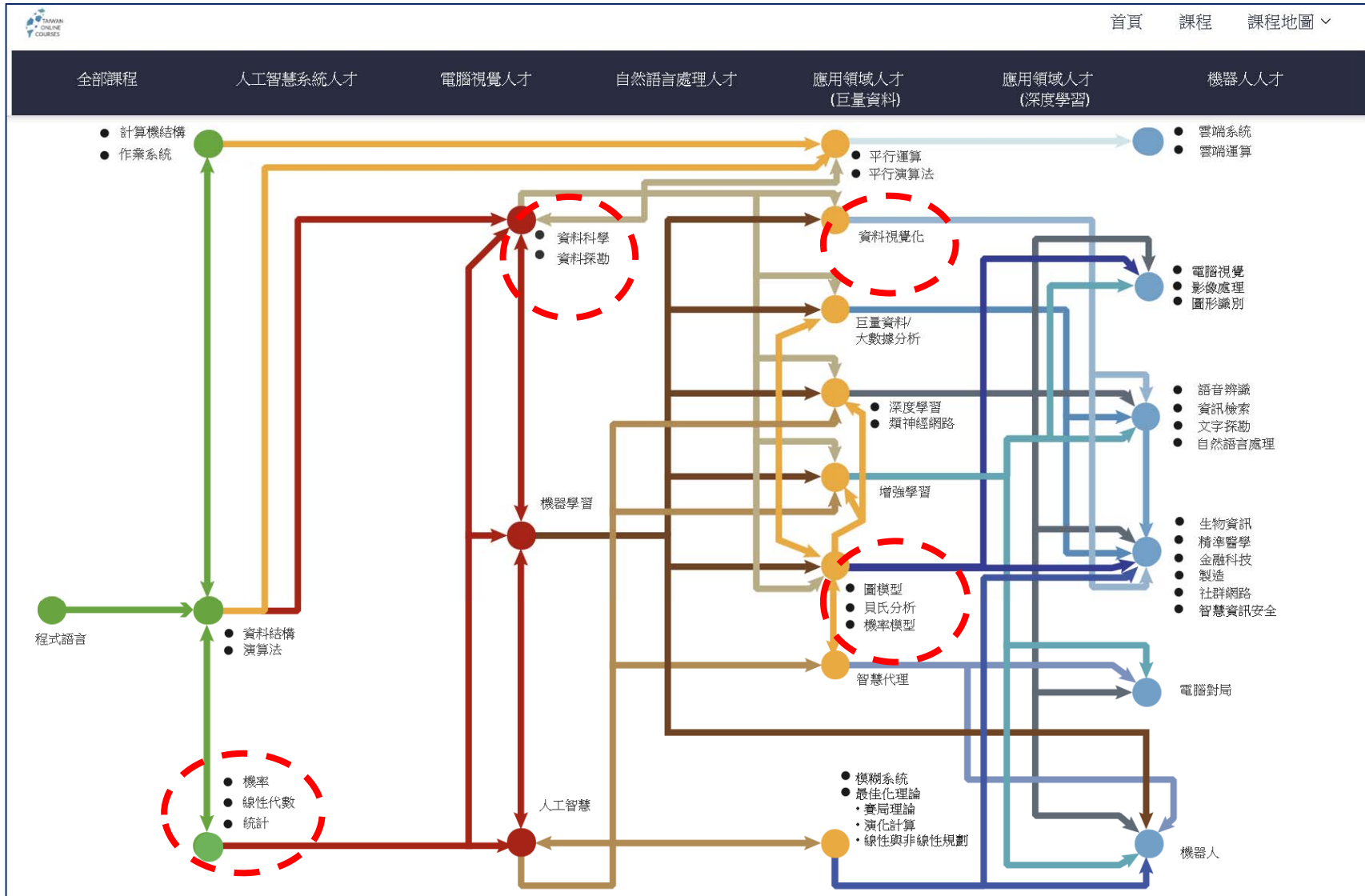
1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning
5. Text Mining / Natural Language Processing
6. Data Visualization
7. Big Data
8. Data Ingestion
9. Data Munging
10. Toolbox



教育部人工智慧技術及應用人才培育計畫: 台灣人工智慧教育平台

70/118

<https://idea.cs.nthu.edu.tw/~Alcoursemap/home/roadmap-course-series-chinese.html>



台灣資料科學系所及學程

對外招生之資料科學學程

學校	系 / 院 / 所	學程名稱	獲得文
台灣大學	電資學院	資料科學學位學程	碩士、博
清華大學	資應所	資訊系統與應用研究所碩士班	碩士
交通大學	資訊工程學系	數據科學與工程研究所碩士班	碩士
中山大學	應用數學系	應用數學系碩士班 丁組 (數據科學組)	碩士
臺北醫學大學	管理學院	學士後大數據科技及管理學士學位學程	學士
臺北醫學大學	管理學院	大數據科技及管理研究所	碩士
東吳大學	巨量資料管理學院	學士學位學程	學士
東吳大學	巨量資料管理學院	碩士學位學程	碩士
東吳大學	巨量資料管理學院	碩士在職學位學程	碩士在
中原大學	商學院	學士後商業巨量資料管理學士學位學程	學士
高雄醫學大學	醫學院	學士後醫學檢驗數據整合判讀學士學位學程	學士

僅校內招生之資料科學學程

學校	系 / 院 / 所	學程名稱
清華大學	統計學研究所	數據科學學分學程
交通大學	資訊工程學系 統計學研究所	巨量資料分析學分學程
成功大學	工業與資訊管理學系 暨資訊管理研究所	巨量資料分析學分學程
政治大學	商學院統計系	巨量資料分析學程
中央大學	統計研究所	巨量資料分析學程
中山大學	資訊管理學系	巨量資料分析學程
臺北大學	商學院	巨量資料探勘學士學分學程
臺灣師範大學	教育心理與輔導學系	大數據資料與數位評量學分學程
高雄應用科技大學	資訊管理系	大數據學程
東吳大學	巨量資料管理學院	巨量資料分析學分學程
銘傳大學	管理學院應用統計資訊學系	大數據統計學分學程
長庚大學	資訊管理學系	大數據資料科學與產業應用學程
元智大學	資訊學院	數據科學學程
元智大學	資訊管理所	大數據商業智慧分析學程
世新大學	管理學院	財經數據學程
淡江大學	統計學系	資料科學學分學程
輔仁大學	管理學院	大數據產業智慧學分學程
東海大學	工學院	大數據碩士學程
逢甲大學	統計學系	巨量金融碩士學分學程
開南大學	資訊管理學系	資料管理與分析學分學程

http://foundation.datasci.tw/academy_tw/

大數據風潮會消失嗎？



數位時代
BUSINESS NEXT

新聞 ▾ 觀點 專題 PX酷品 活動 ▾ 雜誌 創業小聚 數位行銷學院 未來商務展

科技 大數據

大數據會消失，資料科學不會！你該知道的資料科學第一堂課

by  曾彥 2016.07.15



(圖說：第三年舉辦的台灣資料科學愛好者年會於7月14~17日在中央研究院盛大展開，今年邀請來自各領域的50多位資料科學家，進行共43場演講，吸引超過1700人報名。照片來源：台灣資料科學愛好者年會。)

 **臺灣資料科學協會**
TDSF Taiwan Data Science Foundation

理事長 陳昇瑋
中央研究院資訊科學研究所 研究員

陳昇瑋認為，這幾年所有媒體都在講大數據，台灣資料科學協會也在三年前開始積極推廣資料科學 (Data Science) 概念。為什麼當初年會名稱不選擇正當紅的「大數據」一詞？「我們相信大數據這個詞最終會消失，就像現在沒有人在講彩色電視機，因為已經幾乎沒有黑白的電視機；我們現在講手機，也不強調是『智慧型』手機。未來如果所有的工具都是大數據，你就不會特別提，但 Data Science 是不會消失的。」



近年 AI 的快速發展， 奠基於累積半個世紀的 8 大統計學思想！

Journal of the American Statistical Association >
Volume 116, 2021 - Issue 536

[Submit an article](#) [Journal homepage](#)

57,483 Views

2 CrossRef citations to date

Altmetric

[Listen](#)

Reviews

What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman & Aki Vehtari

Pages 2087-2097 | Received 30 Nov 2020, Accepted 23 May 2021, Accepted author version posted online: 02 Jun 2021, Published online: 08 Jul 2021

TechOrange 科技報橘

最新文章 2022 智慧大工廠論壇 PODCAST 雲端運算 未來生活 通訊科技 供應鏈 WEB 3.0

科技動態

近年 AI 的快速發展，奠基於累積半個世紀的 8 大統計學思想！

TO TO 精選觀點 2021-07-09

<https://buzzorange.com/techorange/2021/07/09/8-statistics-theory-develop-ai>

1. 反事實因果推斷 (counterfactual causal inference) (反事實因果推論)
2. 自舉法和基於模擬的推斷 (bootstrapping and simulation-based inference) (自助抽樣法、拔靴法)
3. 超參數化模型和正則化 (overparameterized models and regularization)
4. 貝葉斯多級模型 (Bayesian multilevel models) (貝氏多層次模型)
5. 通用計算演算法 (generic computation algorithms) (泛型計算演算法)
6. 自適應決策分析 (adaptive decision analysis)
7. 魯棒性推斷 (robust inference) (穩健性統計)
8. 探索性數據分析 (exploratory data analysis) (探索式資料分析)

我為什麼要使用R做為資料分析工具?



[Home]

Download

CRAN

R Project

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It covers a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please check the [CRAN mirror](#).

- R is a high-quality, cross-platform, flexible, widely used open source, free language for statistics, graphics, mathematics, and data science.
- R contains more than 5,000 algorithms (>10,000 packages) and millions of users with domain knowledge worldwide.

<http://www.r-project.org>

<https://www.rstudio.com/>

寫程式是資料分析的必要技能

<https://medium.com/datainpoint/9ee15b58cc>

Python or R, what should you learn first?

<https://read01.com/0ePnyD.html#.Wu66C3--kZY>

Why I use R for Data Science – An Ode to R

<https://www.r-bloggers.com/why-i-use-r-for-data-science-an-ode-to-r-2/>

選擇R開發數據分析平台的 4 個不錯的理由

<https://read01.com/660M4g.html>

做數據分析必須學R語言的4個理由

<https://read01.com/yyREB2.html>

Hadley Wickham：一個改變了R的人

<https://read01.com/Mmy64J.html>

Hadley Wickham: "R is ... tailored to the problems of data science"

COPSS Presidents' Award: Hadley Wickham

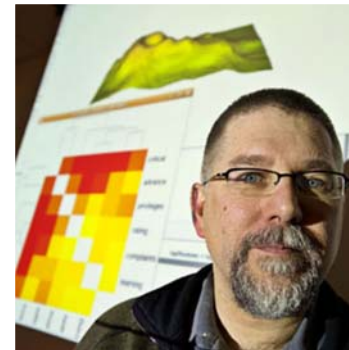
SEPTEMBER 2, 2019

Hadley Wickham wins the prestigious 2019 COPSS Presidents' Award



Robert C. Gentleman

Department of Biostatistical Sciences,
Dana Farber Cancer Institute



Ross Ihaka,
University of Auckland, 2017





TIOBE 全球程式語言排名

The screenshot shows the TIOBE Index website header with the logo and the text "TIOBE Index for February 2024". Below the title, it states "February Headline: The Go programming language enters the top 10".

<http://www.tiobe.com/tiobe-index/>

281種程式語言:

http://www.tiobe.com/tiobe-index/programminglanguages_definition

Feb 2024	Feb 2023	Change	Programming Language	Ratings	Change
1	1		Python	15.16%	-0.32%
2	2		C	10.97%	-4.41%
3	3		C++	10.53%	-3.40%
4	4		Java	8.88%	-4.33%
5	5		C#	7.53%	+1.15%
6	7	▲	JavaScript	3.17%	+0.64%
7	8	▲	SQL	1.82%	-0.30%
8	11	▲	Go	1.73%	+0.61%
9	6	▼	Visual Basic	1.52%	-2.62%
10	10		PHP	1.51%	+0.21%
11	24	▲	Fortran	1.40%	+0.82%
12	14	▲	Delphi/Object Pascal	1.40%	+0.45%
13	13		MATLAB	1.26%	+0.27%
14	9	▼	Assembly language	1.19%	-0.19%
15	18	▲	Scratch	1.18%	+0.42%
16	15	▼	Swift	1.16%	+0.23%
17	33	▲	Kotlin	1.07%	+0.76%
18	20	▲	Rust	1.05%	+0.35%
19	30	▲	COBOL	1.01%	+0.60%
20	16	▼	Ruby	0.99%	+0.17%

Position	Programming Language
21	R
22	SAS

Low Code / No Code AI



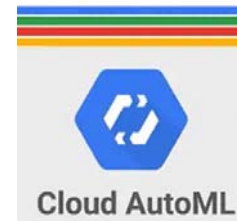
WEKA

The workbench for machine learning

Weka 3: Machine Learning Software in Java

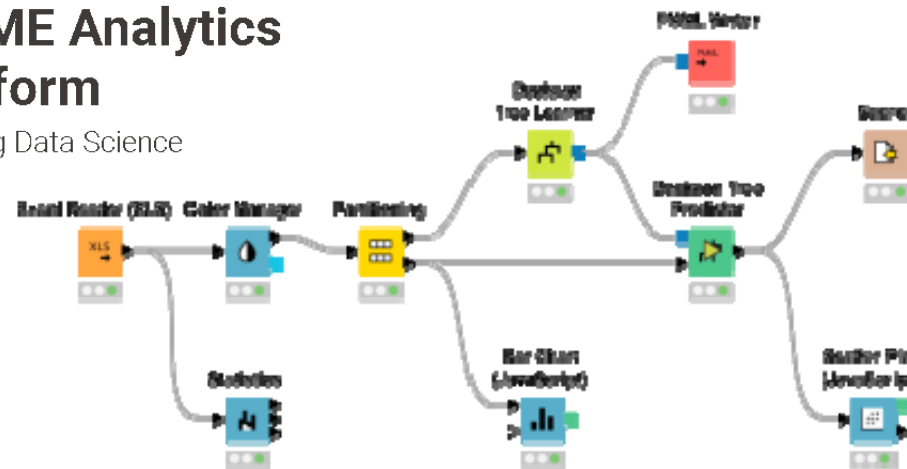
<https://www.cs.waikato.ac.nz/ml/weka/>
<https://waikato.github.io/weka-wiki/>

Google 機器學習三大服務：
AutoML, Cloud ML Engine, ML API



KNIME Analytics Platform

Creating Data Science



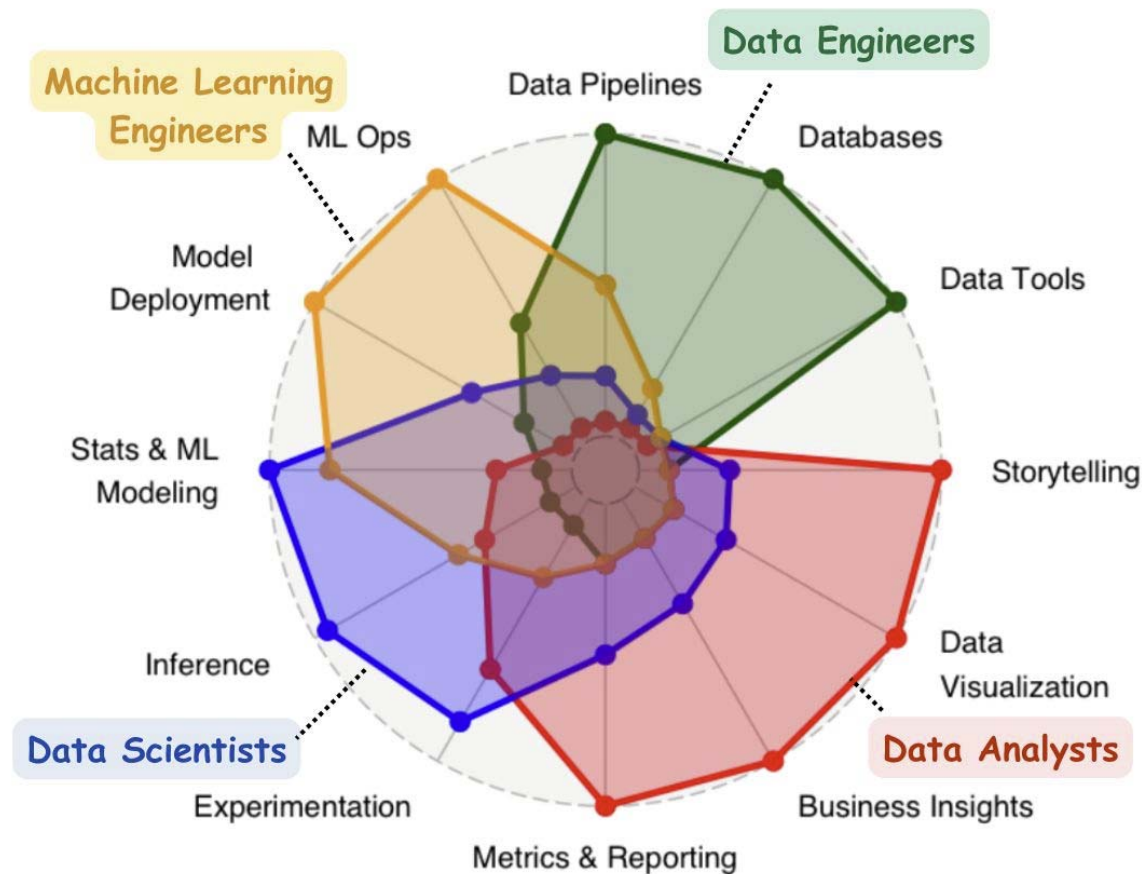
<https://blog.superannotate.com/guide-to-low-code-no-code-ai-platforms/>



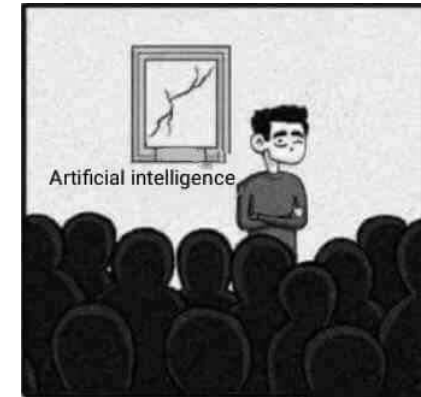
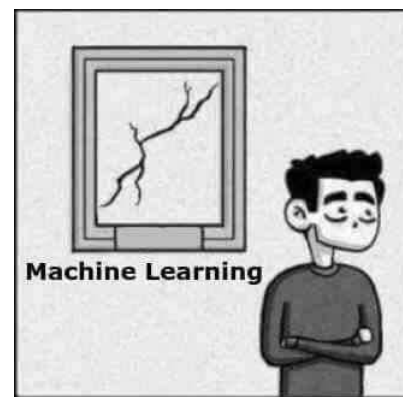
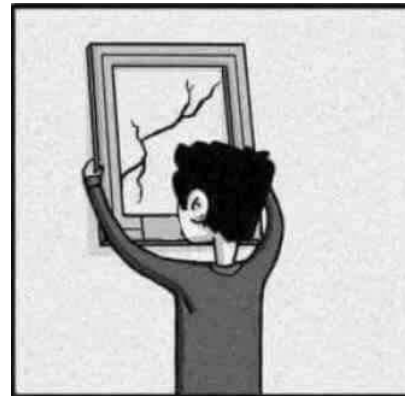
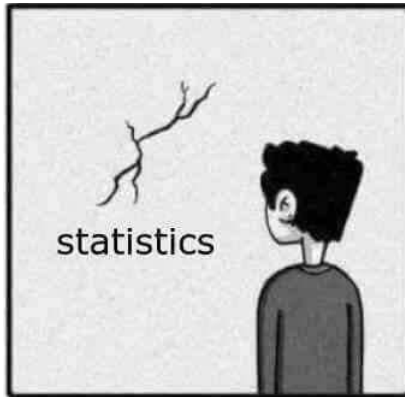
A Guide to Data Roles

<https://www.datacaptains.com/blog/guide-to-data-roles>

Types of Data Roles - Where are you?



統計、資料探勘、機器學習 新瓶舊酒？



不！機器學習才不只是統計學的美化！

<https://kknews.cc/tech/n3yrpyq.html>

機率統計

機率統計方法在其他領域通常很快就成為主流方法，但是在 AI 領域卻經過了很久都沒受到重視，直到最近由於隱馬可夫模型 (HMM) 逐漸在語音辨識領域嶄露頭角，才開始有越來越興盛的趨勢。最近，統計方法在機器翻譯上有越來越強的趨勢，像是貝氏網路 (Bayesian Network)、期望最佳化學習法 (Expectation-Maximization, EM)、蒙地卡羅馬可夫學習法 (Montecarlo Markov Chain, MCMC) 等，都開始展露其優勢，相信在未来的幾年，機率統計法將會在 AI 領域大展身手，推進整個 AI 科學技術的進展。

人工智慧 (陳鍾誠 於 金門大學) <https://cckmit.github.io/aibook/htm/basic.html>

統計 vs 機器學習



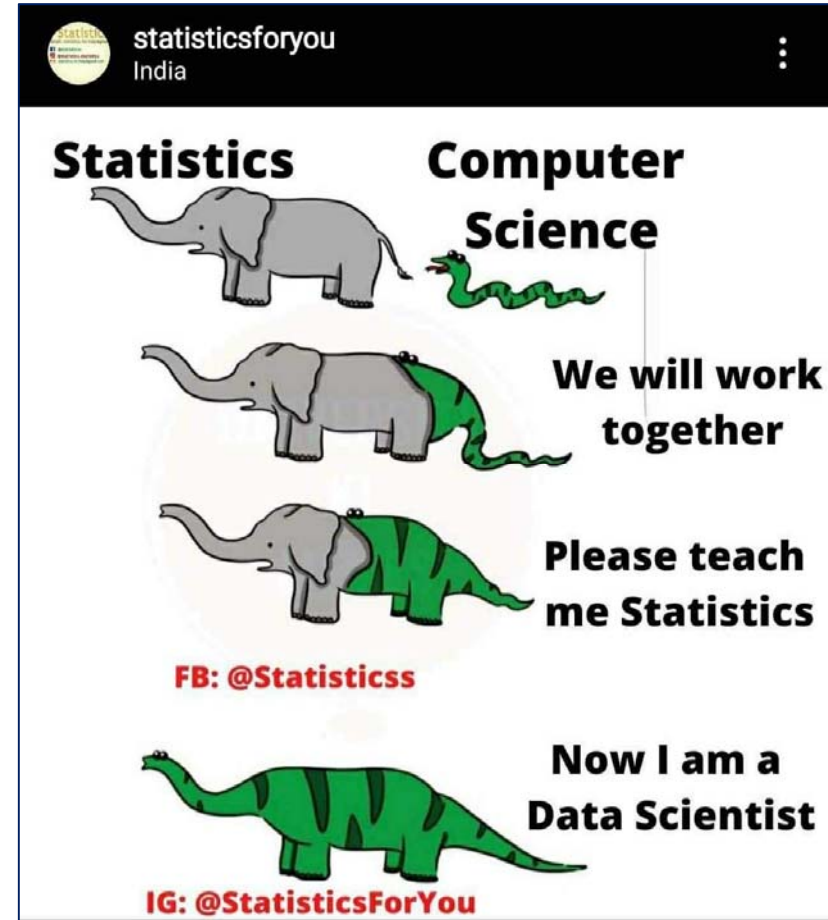
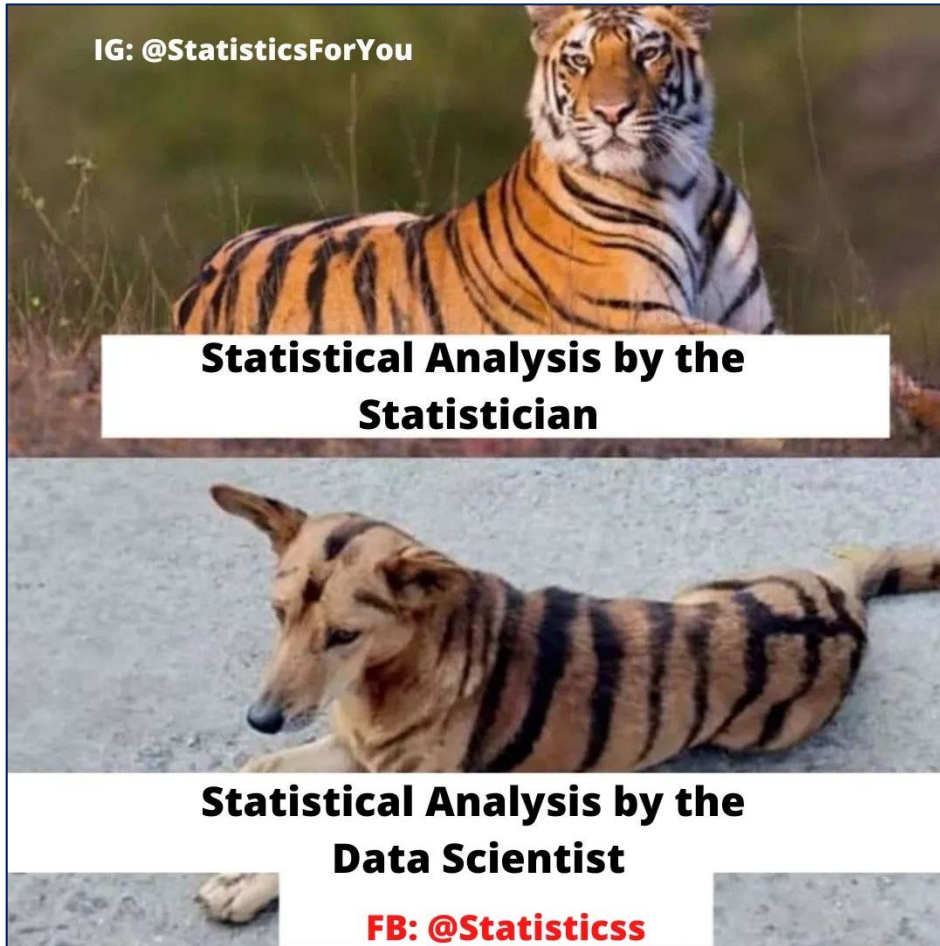
- 統計vs機器學習，數據領域的「少林和武當」！ <https://read01.com/O3dPexn.html>
- 數據科學「內戰」：統計vs.機器學習
<https://read01.com/ePRGMz7.html>
- 機器學習VS統計模型
<https://kknews.cc/zh-tw/tech/gz22r3y.html>
- 運算思維：一張圖看懂機器翻譯（人工智慧）的原理
https://web.ntnu.edu.tw/~samtseng/present/CT_STM.html
[運用電腦來做自動翻譯: 機率、貝氏定理]
- 臉書人工智慧研究主管、紐約大學教授
揚·勒丘恩 (Yann LeCun) :
「人工智慧完全是數學。」
- 人工智慧浪潮下的數學教育 <https://www.ettoday.net/news/20180508/1161306.htm>
- 人工智慧大商機 https://www.hbrtaiwan.com/article_content_AR0007381.html
- 統計學和機器學習到底有什麼區別？ <http://bangqu.com/iw4cp6.html>
- 不要只關心怎麼優化模型，這不是機器學習的全部 <http://bangqu.com/niYN6Z.html>



勒丘恩：「人工智慧完全是數學。」 (Wikipedia)



統計學 VS 機器學習



統計模型、資料探勘、機器學習

- Machine Learning is an algorithm that can learn from data without relying on rules-based programming.
- Statistical Modelling is the formalization of relationships between variables in the form of mathematical equations.

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation/ clustering

機器學習和統計模型的差異

<http://vvar.pixnet.net/blog/post/242048881>

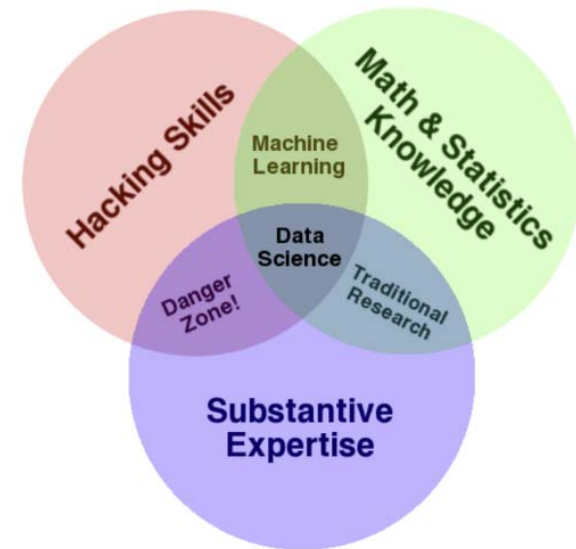
為什麼統計學家、機器學習專家解決同一問題的方法差別那麼大?

<https://read01.com/EBPPK7.html>

機器學習與統計學是互補的嗎?

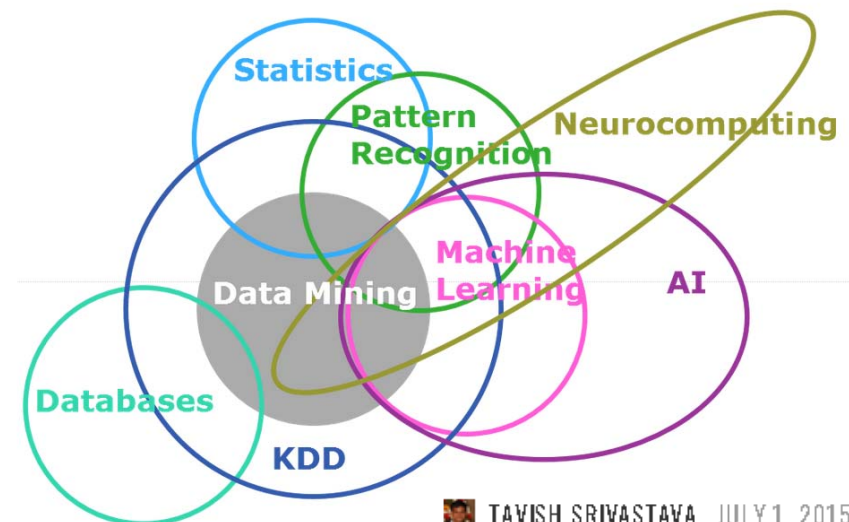
<https://read01.com/ezQ3K.html>

<https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/>



The Data Science Venn Diagram

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>



TAVISH SRIVASTAVA, JULY 1, 2015



統計學(Statistics)、資料探勘(Data Mining)、資料科學(Data Science)、機器學習(Machine Learning)和人工智慧(AI)的差別是什麼?



蔡明順

2023年12月9日 · 🌐

10個最常被詢問產業AI問題

這幾年下來在AI研討會演講、座談與EMBA授課，面對產業經理人與社會人士，關於人工智慧 (AI) 的應用，我累積了一些對話經驗，列出最常見被詢問的10個問題，分享給大家在做2024的年度策略計畫、個人學習計畫或是授課時參考：

1. AI機器學習與統計機率或資料科學的差別? (通用性技術方法的定位問題。AI與統計兩者都是通用性技術，在各種學科都可以結合運用產生結果與論述。機器學習是從數據中學習並做出預測或決策的演算方法，更重視預測效能與精準率；統計學則更關注數據分析、推斷、解釋，重視資料的因果性與可推斷)

統計學 (Statistics)

統計學是數據分析的基礎，專注於數據的收集、處理、分析、解釋和呈現。它使用數學理論來進行數據分析，以推斷群體的特性。統計學方法被廣泛應用於各種領域中，包括社會科學、生物科學和商業分析。

資料探勘 (Data Mining)

資料探勘是從大量數據中識別模式和關聯性的過程。它是資料科學的一部分，結合了統計學、機器學習和數據庫技術，用於預測分析和信息發現。資料探勘關注於從數據集中提取有用信息和知識。

資料科學 (Data Science)

資料科學是一個跨學科領域，使用統計學、資料探勘、機器學習和其他方法來分析和解釋複雜數據。資料科學的目的是從數據中提取知識和見解，並應用這些見解於決策過程中。



您

統計學(Statistics)、資料探勘(Data Mining)、資料科學(Data Science)、機器學習(Machine Learning)和人工智慧(AI)的差別是什麼?



ChatGPT

統計學、資料探勘、資料科學、機器學習和人工智慧各自在資料分析和處理的領域中扮演著獨特且相互關聯的角色。以下是它們之間主要的差異：

機器學習 (Machine Learning)

機器學習是人工智慧的一個分支，專注於開發算法和統計模型，使計算機系統能夠從數據中學習並進行預測或決策，而無需明確編程。機器學習利用數據來改善其性能和預測能力。

人工智慧 (Artificial Intelligence, AI)

人工智慧是指使機器模擬人類智慧過程的技術，包括學習、推理和自我修正。AI的應用範圍廣泛，從自動駕駛汽車到語音識別系統。機器學習是實現AI的一種方法，而深度學習（一種特定類型的機器學習）則推動了近年來AI領域的許多進展。

總的來說，這些領域雖然有所重疊，但各自專注於數據分析和處理的不同方面，從基礎的統計分析到通過機器學習和人工智慧技術實現的複雜預測和決策過程。

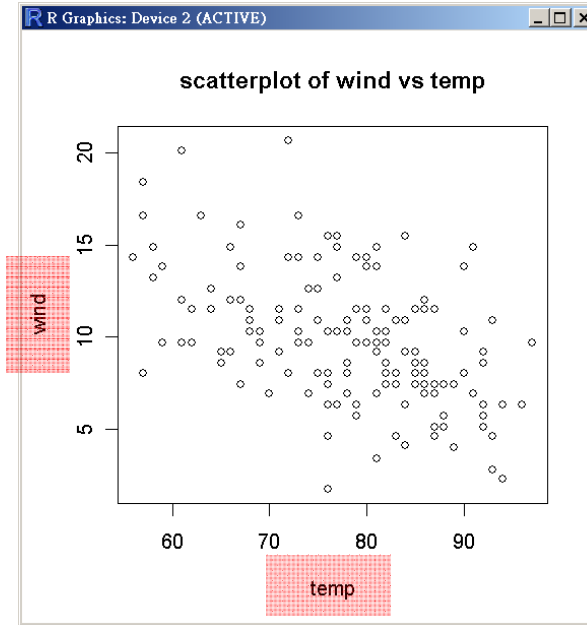


以迴歸分析來說明演算法和統計方法之不同

簡單線性迴歸 (Simple Linear Regression)

```
> dim(airquality)
[1] 153 6
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41    190  7.4  67     5    1
2    36    118  8.0  72     5    2
3    12    149 12.6  74     5    3
4    18    313 11.5  62     5    4
5     NA     NA 14.3  56     5    5
6    28     NA 14.9  66     5    6
```

空氣品質
資料



抽10個觀察值
出來看, 比較清楚

數學模型 $y = \beta_0 + \beta_1 x$

$(y_1, x_1), \dots, (y_n, x_n)$

參數估計: 最小平方法

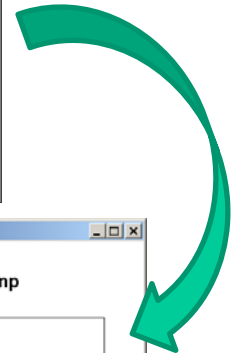
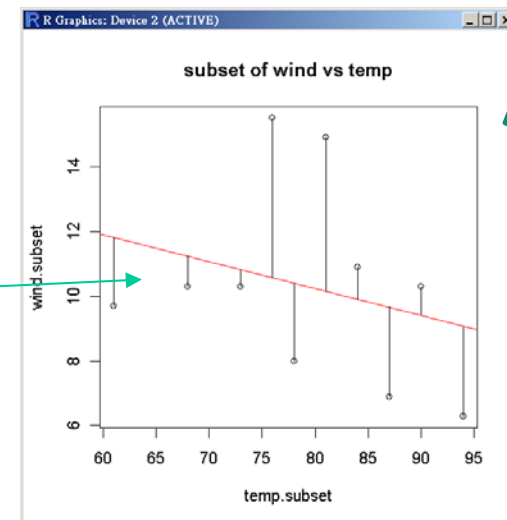
$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

可當成評估指標

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$





簡單線性迴歸 (Simple Linear Regression)

統計模型

$$y = \beta_0 + \beta_1 x + \epsilon \quad E(\epsilon) = 0 \quad Var(\epsilon) = \sigma^2$$

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I) \quad y \sim N(X\beta, \sigma^2 I)$$

參數估計: 最大概似法

$$\prod_{i=1}^n p(y_i | x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}} \quad \text{likelihood}$$

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \log \prod_{i=1}^n p(y_i | x_i; \beta_0, \beta_1, \sigma^2) \\ &= \sum_{i=1}^n \log p(y_i | x_i; \beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned} \quad \begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \end{aligned}$$

統計推論: 信賴區間、假設檢定、近似理論

Testing just one predictor $H_0: \beta_i = 0$.

Test of all predictors $H_0: \beta_1 = \dots = \beta_{p-1} = 0$

統計模型檢測 (Model Checking): 殘差分析(Residual Analysis)



"Artificial Intelligence is All About Math"

科學月刊
SCIENCE MONTHLY

關於科月 | 購買科月 | 訂閱科月 | 投稿須知 | 各期

封面故事 News Focus 專訪 專欄 科技報導 評論 精選文章與其他 活動訊息

2018年3月27日

人工智慧浪潮下的數學教育

魏澤人／任教於國立東華大學，創立花蓮-py社群及實做數學粉專。

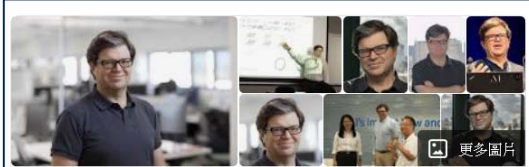
我們曾學過的數學，究竟對人生有什麼幫助？

第一個尷尬點是，數學的實用性變得十分明顯，明顯到令人尷尬。學數學的人，常會聽到人問：「學數學有什麼用？」、「我高中數學都忘光了，還不是活得好好的？」、「工作上好像完全沒用到。」、「我寫程式這麼久了，也沒用到什麼數學。」

身，微

當然，內行人都知道數學在科學、技術、工程中，應用十分廣泛。特別在網路時代，網路加密、電腦運算、影像壓縮，甚至網頁和應用程式的自動排版都得用到數學。即使連照片

編修這種屬電腦工程師們重新拿起統計、微積分及線性代數課本，想要了解現代的人工智慧在玩什麼把戲。現代人工智慧的領軍人物之一勒丘恩 (Yann LeCun) 說「人工智慧就是數學 (artificial intelligence is all about math)」，他給想從深入人工智慧領域大學生的建議是：「如果在『iOS程式設計』及『量子力學』中要選一門課來修的話，選量子力學，且一定要選修微積分一、微積分二、微積分三.....、線性代數、機率與統計，和盡可能的多選物理課程。即便如此，最重要的還是要會寫程式。」



楊立昆

電腦科學家

楊立昆，法國籍計算機科學家，他在機器學習、計算機視覺、移動機器人和計算神經科學等領域都有很多貢獻。他最著名的工作是在光學字符識別和計算機視覺上使用卷積神經網絡，他也被稱為卷積網絡之父。他同Léon Bottou和Patrick Haffner等人創建了DJVu圖像壓縮技術。維基百科

出生資訊：1960年7月8日（60歲），法國蘇瓦西蘇蒙特莫朗西

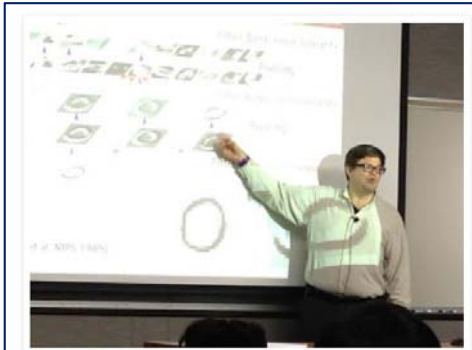
獲頒獎項：圖靈獎 (2018); AAAI Fellow (2019); 法國榮譽軍團勳章 (2020)

論文：Modèles connexionnistes de l'apprentissage (1987)

模範生：沃伊切赫·薩倫巴

獲獎記錄：圖靈獎

學歷：巴黎第六大學 (1983年-1987年)，巴黎高等電子工程師學校，索邦大學



勒丘恩：「人工智慧完全是數學。」 (Wikipedia)



人工智慧只是統計學的延伸

每日頭條

- 首頁
- 健康
- 娛樂
- 時尚
- 遊戲
- 3C
- 親子
- 文化
- 歷史
- 動

AI（人工智慧）就是統計學？

2018-10-26 由 麥思生涯規劃 發表于科技

諾貝爾經濟學獎托馬斯·薩金特在《財經》世界科技創新論壇上的演講中說過一句話：人工智慧首先是一些很華麗的辭藻。人工智慧其實就是統計學，只不過用了一個很華麗的辭藻，其實就是統計學。



<https://kknews.cc/zh-tw/tech/5z36z28.html>

AI 不過是統計學

Thomas J. Sargent：人工智慧只是統計學的延伸

2011年諾貝爾經濟學獎取得者Thomas J. Sargent在題為“共享全球智慧 引領未來科技”的世界科技創新論壇上表示：

計算機是非常擅長計算，它們可以非常快速地完成計算人算不了的東西，但最終必須由人來組織和分析這些計算。你可以看一些非常成功的人工智慧應用，它不僅是機器在「思考」，也是科學家在思考。像 AlphaGo 的演算法看上去是第一次出現，但其實有很多非常聰明的數學，並且是由人設定教學內容。人工智慧是由機器和人分飾兩角的，非常有趣。

AI時代的中層支柱：統計學: <https://www.mdeditor.tw/pl/2nBY/zh-tw>

但是為什麼不說搞統計學呢？很簡單，因為不如人工智慧說法高大上，為什麼要高大上？因為高大上有人投錢。人工智慧代表了最新科技、最熱行業，能吸引投資，你要說是做統計學的誰投錢？誰買單？實際上這些工作很早就有人研究，只是那時候都歸類於統計學領域。

就像顯示系統縱橫位置指示器就是滑鼠；人體表皮污垢學就是搓灰；人體表皮死細胞分離器就是搓澡巾；智能高端數字通訊設備表面高分子化合物線性處理就是手機貼膜.....

所以現在大家都學聰明了，那怕是老生常談也得包裝一個好聽的名稱，你說搞機器學習、深度學習就有人投錢，有人出大價錢挖你，你要說搞統計學大家立馬就不感興趣了，其實做的還是一回事。當然資本也是知道這些道理的，那為什麼還要投錢，因為資本是逐利的，投錢是為了掙更多錢，高大上的外衣就是掙錢的保障之一。

人工智慧和統計學不能完全劃等號

人工智慧和統計學存在莫大的關係，或者說統計學是人工智慧的最重要的理論基礎，但統計學和人工智慧依然有著很大不同，更不是一回事。

科技領域都需要數學



聯合新聞網

[願景](#)
[時尚](#)
[汽車](#)
[NBA台灣](#)
[運動筆記](#)
[遊戲](#)
[國際](#)
[嗶人堂](#)
[新鮮事](#)
[部落格](#)
[全部](#)

[即時](#)
[要聞](#)
[娛樂](#)
[運動](#)
[全球](#)
[社會](#)
[專題](#)
[產經](#)
[股市](#)
[房市](#)
[健康](#)
[生活](#)
[文教](#)
[評論](#)
[地方](#)
[兩岸](#)
[數位](#)
[旅遊](#)
[閱讀](#)
[雜誌](#)
[購物](#)

udn / 生活 / 職場觀測

相關新聞

郭董指數學很重要 台師大教授：數學人才出路好的時代來了

f 分享

LINE 分享

留言

列印

存新聞

A-

A+

2019-03-01 21:28 聯合報 記者陳智華／即時報導 [讚 7,878](#) [分享](#)

在美國，數學家的出路非常好，台灣這樣的時代也要來了嗎？

台師大電機工師系助理教授、數學專欄作家賴以威今在臉書貼文指出，鴻海集團董事長郭台銘說：「數學很重要！」他因之前受邀鴻海跟高階主管演講時，親耳聽郭董對總部3、4百位高階主管、十幾個遠端連線的各地分公司主管員工這麼說。

賴以威聽到郭董跟員這樣說時，才知道郭董找他去演講是為了推廣數學，推廣成長型數學思維。

郭董指出，工業物聯網、人工智慧、資料分析，這些鴻海現今著重的科技領域，背後都需要數學。因此，讓集團意識到數學的重要性，知道如何學好數學的心態是非常重要的。

他表示，演講後的隔天，郭董跟他聊到，鴻海很歡迎電機系和數學系的人才加入，與相關的產學合作。他這兩天認識許多裡面的同事、主管也是相關科系。

賴以威說，美國就業網站 CareerCast 曾經統計過，在美國數學人才有非常好的工作機會。在台灣，看來這樣的時代也要來臨了。

但賴以威強調，數學很重要，這指的不是程序性的計算，而是懂得活用的數感。

賴以威說：「郭台銘董事長都這樣說了，你不覺得嗎？」



<https://udn.com/news/story/7266/3672385>

「無用之用，是為大用」



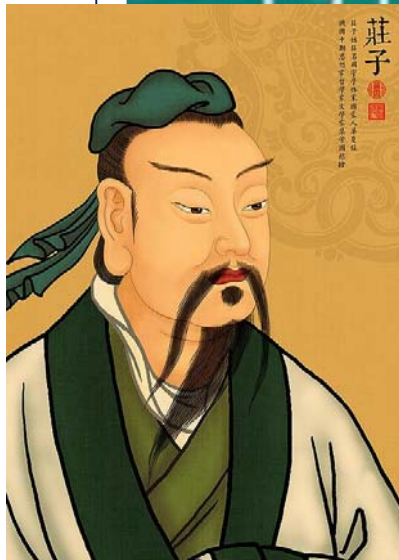

訂閱天下
Q
User icon

天下財經週報 ▶ 美國挺台，但這件事可能不挺

教育 > 教育趨勢

【蘇文鈺專欄】「用不到、不用學」？三角函數可以不用上？

生活中用不到就不用學？教育界愈來愈流行這樣的說法，作者提醒，如果大家都抱持這樣的想法看「學問」，今天世界上許多偉大的發明，例如影片或數位音樂都不會出現。

<https://www.cw.com.tw/article/5096186>

莊子說過：
「無用之用，是為大用」



大數據、機器學習及人工智慧的必讀書目：統計學

每日頭條

作者：吳喜之

<https://kknews.cc/tech/vzj4vrq.htm>

首頁 健康 娛樂 時尚 遊戲 3C 親子 文化 歷史 動漫

入行AI，首先要了解機器學習機率圖模型

2017-12-21 由 異步社區 發表于科學

大數據機器學習及人工智慧必讀書目——《統計學：從數據到結論》

2017-10-19 由 愛編程愛統計 發表于科技

大數據、機器學習及人工智慧必讀書目——《統計學：從數據到結論》

我們已經進入了全新的數據時代，大數據、雲計算、物聯網、機器學習、人工智慧等等一系列技術紛至沓來，數據的管理和應用已經滲透到每一個行業和業務領域，成為當今以及未來商業運作的基礎資產。可以說，只有掌握數據並善於運用數據的人，才會在競爭日益激烈的環境中尋得先機。那麼我們該怎麼樣學習大數據分析、機器學習以及人工智慧？作者認為，學習大數據、機器學習和人工智慧，所需的知識分為五個層次，一是數學知識；二是統計學知識；三是算法知識；四是工具知識；五是哲學思想知識。所謂工具知識，就是我們需要藉助計算機軟體來完成相關的分析 and 運算，目前大數據和機器學習領域熱門的語言就是 R 和 Python。我們會分別介紹這五個層次所需要看的書，希望對大家有用。

<https://kknews.cc/tech/4v4ymkg.htm>

機器學習是關於從數據和觀察中學習和適應的算法研究，並實現推理和藉助學到的模型和算法來執行任務。由於我們生活的世界本身就是不確定的，從這個意義上講，即便是最簡單的觀察，例如天空的顏色也不可能絕對的確定。我們需要一套理論來解決這些不確定性。最自然的方法是機率論，它也是本文的數學基礎。

<https://kknews.cc/science/bk58y3m.html>

從最大似然估計開始，你需要打下的機器學習基石

2018-01-09 由 機器之心 發表于科技

選自Medium

作者：Jonny Brooks-Bartlett

機器之心編譯

機率論是機器學習與深度學習的基礎知識，很多形式化的分析都是以機率的形式進行討論。而這些討論或多或少都離不開最大似然估計，因為它是參數估計的基礎之一，也是構建模型的基石。在本文中，我們從最大似然估計到貝葉斯推理詳細地討論了機器學習的機率論基石，並希望能為讀者的預習與複習提供優秀的參考資源。



「了解AI背後的原理和數學是必要的」

YC Chen AI.ML 剖析深度學習

- (1)：為什麼Normal Distribution這麼好用？
https://www.ycc.idv.tw/deep-dl_1.html
- (3)：MLE、MAP差在哪？談機器學習裡的兩大統計觀點
https://www.ycc.idv.tw/deep-dl_3.html
- Sigmoid, Softmax怎麼來？為什麼要用MSE和Cross Entropy？談廣義線性模型
https://www.ycc.idv.tw/deep-dl_4.html

趙民德，1999，

「統計已死，統計萬歲！」

第八屆南區統計研討會演說稿



趙民德
台灣

趙民德，國立台灣大學數學系畢業、美國加州大學柏克萊分校統計博士。在美國求學工作多年後，1999年回台灣擔任中央研究院

“統計有沒有死？會不會萬歲？只要有米倉，就會有老鼠；只要有數據，就會發展處理數據的方法。但是不是叫做統計學、或者叫做computer science的data mining，就要看這一代的統計人如何因應變局。”

剖析深度學習 (1)：為什麼Normal Distribution這麼好用？

YC Chen 2020-02-18 AI.ML 剖析深度學習

如果你已經學了好一陣子的機器學習或深度學習，應該對於Normal Distribution不陌生，但是你真的懂Normal Distribution嗎？本講會詳細的探討Normal Distribution，並且引入中央極限定理（Central Limit Theorem）來解釋為何自然界的隨機誤差大都呈現Normal Distribution，再來介紹Entropy，並且利用Entropy揭示Normal Distribution具有最少先驗知識（Prior Knowledge）的特性。

深度學習發展至今已經有相當多好用的套件，使得進入的門檻大大的降低，因此如果想要快速的實作一些深度學習或機器學習，通常是幾行程式碼可以解決的事。但是，如果想要將深度學習或機器學習當作一份工作，深入了解它背後的原理和數學是必要的，才有可能因地制宜的靈活運用，YC準備在這一系列當中帶大家深入剖析深度學習。



AIA蔡明順校務長：「當知識生產不再是永久優勢時，想像力、提問力、深度思考會讓你與眾不同，內在驅動將成為最有價值的關鍵。」



資料分析的基本流程



到底要用哪一種處理/分析方法?

統計方法? 大數據? 機器學習? 深度學習? AI?

Source: 資料科學的統計實務，溫政堯譯，旗標出版社，2021/11/01。



今日新聞 NOWNEWS | 60.9k 人追蹤 ☆ 追蹤

善用房產科技！靠數位工具突破限制

生活消費中心/綜合報導
2022年11月22日

房地產科技PropTech (Property Technology)，結合房地產銷售結合AI人工智慧、大數據、機器學習等數位科技，藉此突破時間和空間限制，提升銷售效率，包括增加消費者和物件的媒合效率，國外已有許多房地產業者逐漸向科技靠攏，而國內房仲龍頭信義房屋近年也致力於PropTech發展，以科技賦能輔助，希望為買賣雙方建構出更多元、貼心的服務。

<http://shorturl.at/kzOU7>

3.1k 人追蹤 ☆ 追蹤 yahoo! 新聞

大數據分析交通違規事件 違停居首

【記者張淑珠 / 台中報導】
2023年2月7日 週二 下午8:47

在重大違規部分，黃士哲說，依據無照駕駛違規態樣分析，實際未領有駕照有59%，其中未滿18歲占14%、滿18歲占45%；進一步分析，未領有駕照、且未滿18歲所駕駛的車種有95%是機車。對此，交通事件裁決處自110年6月起，針對未滿18歲違規駕駛案，每月挑檔提供清冊、交由警察局進行後續個案關懷協助，警察局另將學籍在台中市的無照駕駛列管案件，每月通報教育局轉各校，並啟動校園關懷輔導。

此外，針對酒後駕車，經大數據分析顯示，男性為主要的酒駕違規者，且在各年齡層的每十萬人違規人數中，以成年男性 (25-64歲) 酒駕情形最為嚴重，年輕男性 (18-24歲) 次之。交通事件裁決處表示，將強力配合法務部行政執行署執行滯欠酒

<http://shorturl.at/oGY04>

統計與實務問題



- 學校的統計課程多半專注方法論，應用時卻需考慮實際需求、跨領域知識的整合。
- 統計諮詢目的在於「解決問題」(Problem Solving)，藉由統計思維找出關鍵，洞察問題比建立完美方法論重要。

在學校我們學習教學的方式：

- Theory and Methods > Analyzing Messy Data > Interacting with Data > (Data Management) > (Empirical Values)

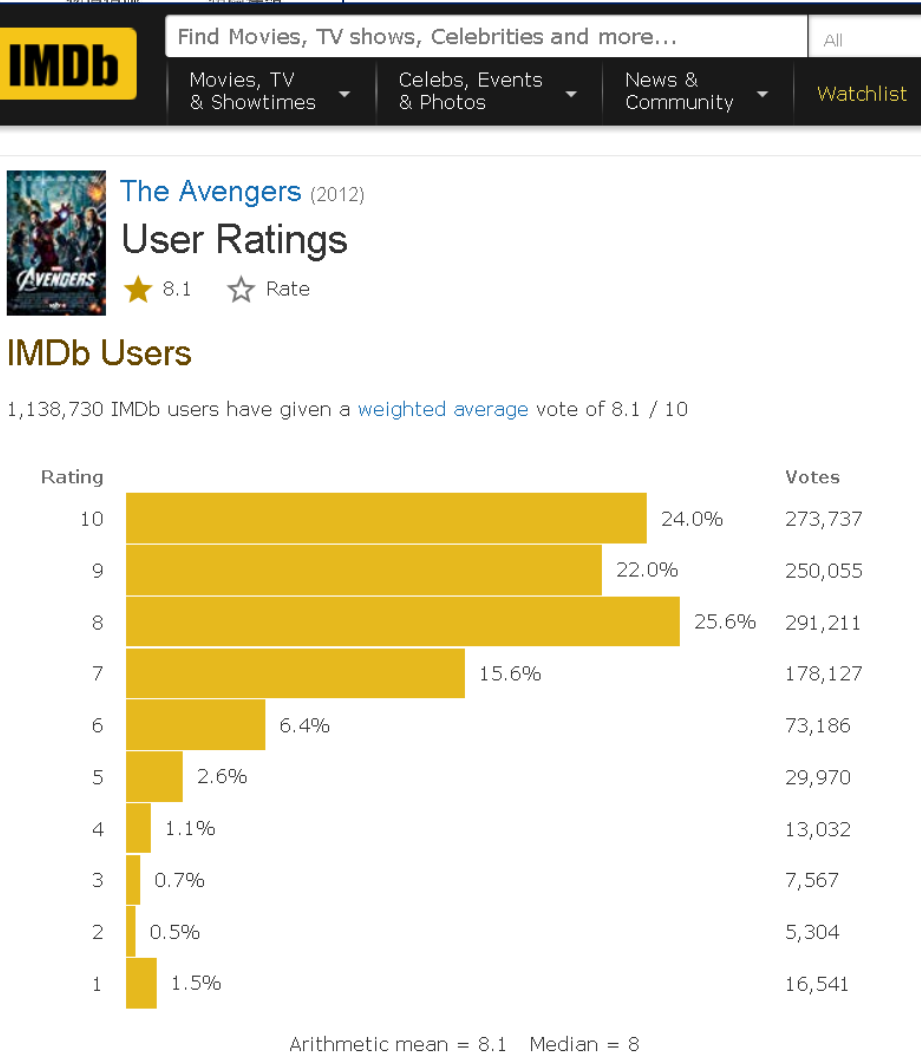
實際解決問題時的順序：

- Interacting with Clients > Data Management > Analyzing Messy Data > Methods > Theory

象徵型資料 (Symbolic Data)



(intervals)



象徵型資料分析 (Symbolic Data Analysis) (Billard and Diday, JASA 2003)






Standard data table

Players	Y_1	Y_j		
ind_1				
ind_i		Y_{ij}		
ind_n				

A number
(Messi age)
or a
Category
(Messi
nationality)

Symbolic data table

	Y'_1	Y'_j		
Cl_1				
Cl_i				
Cl_k				

A symbolic
data
describing
Messi team

Age
interval

Weight
histogram

Nationalities
barshart

資料無邊界！統計教學

Statistician



- 統計教學裡的範例幾乎都是結構性的數據。
- 大數據時代，80%的資料是非結構性的，統計課程如何面對？

Image source: <http://marketbusinessnews.com/financial-glossary/what-is-a-statistician>



☰ Data Science Education

資料科學時代的統計教學

Statistics Education in the Data Science Era

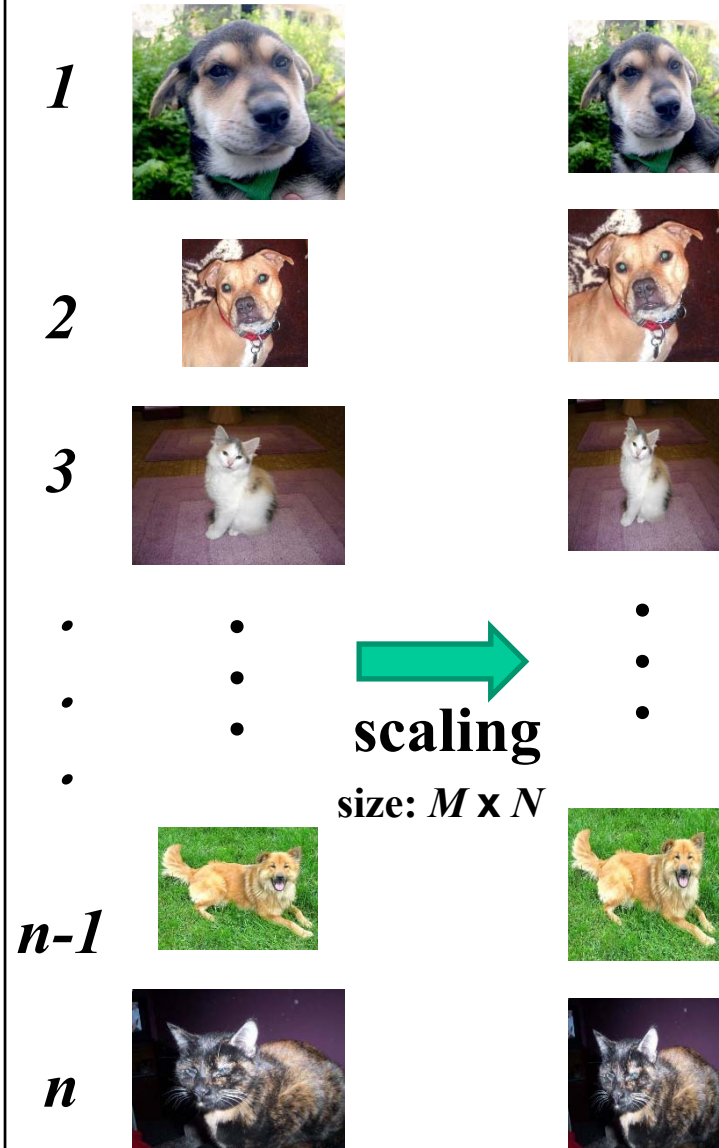
2018.3.9下午4點起至2018.3.10中午12:30 於

台南國立成功大學光復校區統計學系

2018統計教學工作坊

18~19 May, 2018
Lounge (2F),
Institute of Statistical Science, AS

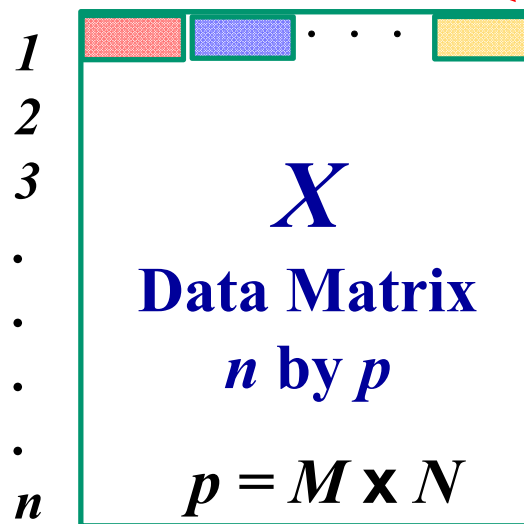
非結構化資料: 圖像



pixel
示意圖



貓: 1
狗: 0



feature
extraction,
learning
algorithms

非結構化資料: 文字



iRobot Roomba j7+ 掃地機器人試用 以頂尖智慧搭配可靠性能的地板清

潔管家



討論群組

1

最喜歡的是有自動閃避的功能, 家裡的東西太多, 很需要這種功能的掃地機器人
 連接App的功能也能開啟定時清掃, 很適合懶人

2022-03-14 10:46 #10

2

有掃地機器人很棒
 不過現在家庭越來越小 掃地機器人運行的聲音頗大聲 需要注意 雖然一下子就走完全部的
 範圍了

2022-03-14 10:53 #11

.

.

.

.

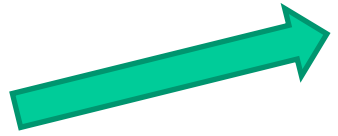
n

最喜歡的除了自動集塵外 其實更愛掃完後運動拖地的功能
 雖然現在市面上很多掃拖一體的機器 但總覺得會有還沒掃乾淨就拖過去的疑慮
 所以可以運動真的滿方便的 如果抽到m6 就去買一台j7來用

2022-03-14 14:04 #41

因為家裡有總來在, 每次掃地機器人跑過之後, 就必須手動清理集塵盒, 但因為毛髮實在太多, 常常打開來之後, 會有幾根飄出來, 所以自動集塵的那個功能看了蠻喜歡的 (所以我有在看小米的那款) .
 避開糞便的功能則是, 家裡養的上一條狗 (後來回天國了), 會在廁所外面門框處便溺, 因此當時就沒有買掃地機器人 (因為掃地機+便便=家中恐怖災難), 雖然現在總來會定點上廁所, 可是萬一哪天有啥狀況, 又造成災難, 所以有這功能的感覺還是蠻不錯的。

2022-03-15 14:37 #92



自然語言處理(NLP)
 (斷詞、理解詞、語法、語義)
 Bag of Words, Word Vector, Word Embedding, ...



正評: + 1
 負評: - 1

方便 躁音 . . . 節能標章

使用者
 資料
 &
 LOG

1
 2
 3
 ⋮
 n

5	10	0	...
X			
Data Matrix			
<i>n by p</i>			

y
 Clas
 s
 n
 by
 1

餘弦相似性 (Cosine similarity)
https://en.wikipedia.org/wiki/Cosine_similarity



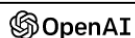
feature extraction, learning algorithms



2022/11/30 ChatGPT 橫空出世

98/118

OpenAI 推出了生成式 AI 對話機器人 ChatGPT，
立下了人工智慧技術的一個重要里程碑



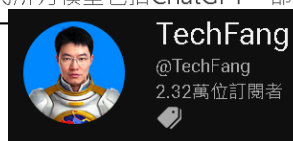
ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT到底是什麼？OpenAI這樣說

ChatGPT是OpenAI訓練的對話式大規模語言模型，以對話互動，和另一款模型InstructGPT屬於同級模型，代表「GPT 3.5」這代。之前微軟和OpenAI簽訂戰略合作計劃，GPT 3.5代所有模型包括ChatGPT，都是用Azure AI超級計算集群訓練。

<https://openai.com/blog/chatgpt/>



2023 AI 工具大集合: 50+ 個人工智慧軟體應用整理:

<https://rb.gy/yteei4>

TechNews 科技新報

airasia Super App The New Super App Unlimited free sea

搶先看 半導體 零組件 網路 AI 人工智慧 尖端科技 環境科學 能源科技 系列專題 研討會活動

行走的程式碼產生器，chatGPT 會讓程式設計師失業？

作者 品玩 | 發布日期 2022 年 12 月 07 日 8:30 | 分類 人力資源, 機器人, 網路

Source: <https://technews.tw/2022/12/07/chatgpt-openai>



Source: <https://www.cw.com.tw/article/5123853>

「學生將使用AI來編寫作業，教師將使用AI對他們評分。沒有人學習，沒有人獲得。」夏普爾斯直言。



作者：
理查·伍爾曼
出版社：時報
文化出版企業
股份有限公司
出版日期：
1994-08-15

AI Tools



<https://yuvaraj.gs/upgrade-your-business-with-the-latest-ai-tools-of-2023-plus-coupons-to-save-big/>

AI Tools	
Text-to-Image	
Text-to-Video	
Text-to-Audio	
Text-to-Text	
Text-to-motion	
Text-to-Code	
Audio-to-Text	
Text-to-music	


Writing	ChatGPT ChatGPT is the world's most advanced AI text generation tool.	Jasper Instantly create content for your blog, social media, website, and more.	CopyAI An AI-powered content generator for all your marketing needs.
Video	Fliki Turn text into videos with AI voices.	SuperCreator Create short form videos 10x faster using AI.	Runway Turn any image, video clip or text prompt into a compelling piece of video.
Audio	Krisp The best AI tool to eliminate background noise during calls.	Altered An AI tool to record and edit your voice recordings everywhere.	Voicemod Real-time voice changing and custom sound effects for every game and app.
Coding	GitHub Copilot Suggests code and entire functions in real-time, right from your editor.	AutoRegEx Converting plain English to RegEx with Natural Language Processing.	Lightning AI Fast and minimal libraries to train and deploy AI models.
Productivity	Albus An AI assistant to get answers quickly and easily.	Slides AI Create compelling presentations with AI in seconds.	Notion AI Write better in Notion with the help of AI. Create better content 10x faster.
Images	Midjourney Midjourney uses AI to create images from text descriptions.	RocketAI Generate creative images that perfectly capture your unique brand style.	BRIA BRIA allows you to generate high-quality images with one click.

Sam Altman, CEO and cofounder of OpenAI.

https://superibosk.live/product_details/47450774.html

Synthesia: 152 Best AI Tools Of 2024 (Fresh Update!)

<https://www.synthesia.io/post/ai-tools>



Google 2023年底推出Gemini： 一個更強大可以處理多模態的LLMs

Introducing Gemini: our largest and most capable AI model

Dec 06, 2023

<https://blog.google/technology/ai/google-gemini-ai/>



Gemini 是一種可以處理文字、圖像、程式碼和影片的新型人工智慧模型，它可能是繼PageRank之後 Google 歷史上最重要的演算法，它讓搜尋引擎進入了公眾的視野，並創造了一個企業巨頭。



親自試用 Gemini：與多模態 AI 互動

- 0:00 簡介
- 0:19 多模態對話
- 1:32 多語言
- 2:04 遊戲製作
- 2:31 視覺謎題
- 3:17 建立聯繫
- 3:39 圖片與文字產生
- 4:06 邏輯與空間推理
- 4:55 視覺翻譯
- 5:27 文化理解

親自試用 Gemini：與多模態 AI 互動



訂閱

6.2萬



分享

下載

剪輯片段

儲存

觀看次數：2,870,730次 2023年12月6日

Gemini 是我們原生的多模態 AI 模型，能跨文字、圖片、音訊、影片和程式碼進行推理。本影片重點介紹我們最喜歡與 Gemini 進行的一些互動。了解更多並試用模型：<https://deepmind.google/gemini>

在這裡探索我們的提示方法：<https://goo.gle/how-its-made-gemini>

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



2024年2月OpenAI推出 Sora: text-to-video

101/118

OpenAI 降維打擊！推出 60 秒影片生成模型 Sora

閱讀時間：6 分鐘



Introducing Sora — OpenAI's text-to-video model

近期 AI 最重大的事件來了！OpenAI 推出了能夠生成 60 秒影片的 AI 模型 Sora，而且影片的真實度幾乎已經無法跟真實影片做出區別。

雖然目前只有 Demo 影片，但與先前 Runway、Pika 釋出的 Demo 影片品質相比，完全是降維打擊！推薦你一定要看看影片！

而且 Sora 可能不只是一個影片生成模型，它之所以能夠生成出那麼真實的畫面，是因為 Sora 能理解、模擬真實世界的 3D 運動、物體行為和複雜互動。

OpenAI 甚至將 Sora 定位成世界模擬器 (World Simulators) 🤖！

https://www.youtube.com/watch?v=HK6y8DAPN_0

OpenAI: Introducing Sora — OpenAI's text-to-video model



AI 繪圖工具箱

1天

上新片啦~ 火速弄了部講解 Sora 的影片。

相信大家這兩天已經各種被 OpenAI 的這個新模型洗版了吧。它誇張的品質和能力直接讓一堆新創就地解散，也讓一堆在做同題目的研究員懷疑人生。除了厲害和好看之外，它對於未來可能會有什麼革命性的影響呢？它的技術原理大致又是如何呢？(劇透一下，至少就公開情報看來，sora 本質上使用的技術全部都是現有的)

這部影片會涵蓋

- 👉 sora 為什麼重要
- 👉 跟圖片生成模型比起來，sora 能帶來什麼根本性的顛覆
- 👉 sora 使用的技術原理為何
- 👉 下一步是什麼

大家有興趣來聽聽吧~

🔥 影片: <https://youtu.be/LtrVwgxBwPE>

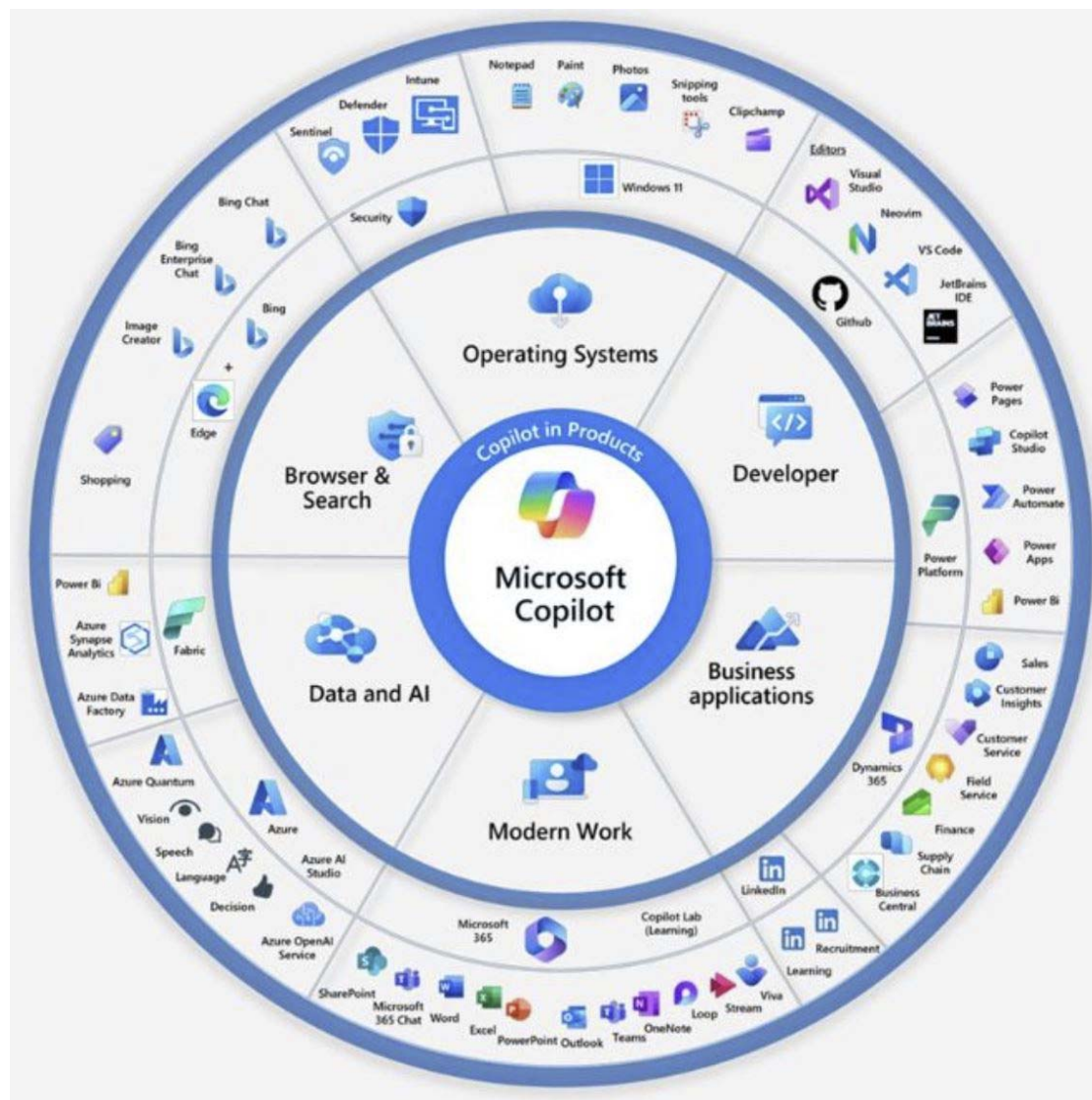
影片: <https://youtu.be/LtrVwgxBwPE>

Microsoft Copilot的生態系統與運作

1. Microsoft Ignite 2023: AI 轉型與推動變革的科技的說明

2. Microsoft 2023年的《工作趨勢指數》報告

https://assets-c4akfrf5b4d3f4b7.z01.azurefd.net/asset/s/2023/09/a81fcdeb-860a-44f2-aaeb-0525d38358ae-2022_Work_Trend_Index_Pulse_Report_Sep-3697v2.pdf





資料科學的角色類型與GenAI帶來的影響



蔡明順

2023年12月30日 · 🌐



資料科學的角色類型與GenAI帶來的影響

認識不少前來AIA進修的工程師校友與準備深造的碩博士生，都志於進入資料科學及人工智慧領域，希望成為一名專業人士，在資料科學領域一般可以分成四種角色: 資料工程師、資料科學家、資料分析師、機器學習工程師。除了機器學習工程師之外，其他三個角色的職責可能會根據專案的不同角色，界線上變得模糊。

I. 資料科學角色的深化

1. 資料工程師：負責建立與維護數據管道，確保資料的品質與可用性。他們使用大數據技術，如Hadoop和Spark，以及資料庫管理系統，以支持資料科學家和分析師的工作。
2. 資料科學家：專注於使用統計學、機器學習和預測建模來解釋數據，發現洞見，並對業務問題提供解決方案。他們經常需要具備跨學科的知識，包括領域專業知識。
3. 資料分析師：專注於解釋資料，以幫助決策者理解資料背後的趨勢和模式。他們通常使用資料可視化工具，如Tableau或Power BI，來幫助解釋複雜的資料集。
4. 機器學習工程師：著重於設計、實施和部署機器學習算法和模型。這些專業人士需要有深厚的程式設計技能和對演算法的深入理解。

II. 技能發展階段的擴展

1. 基礎階段（Coding）：除了基本的數據提取、分析和視覺化，還應包括資料清理和預處理的技能，以及對各種程式語言（如Python、R）的熟練使用。
2. 中級階段（故事講述）：進一步發展資料解釋能力，將資料背後故事講述與業務決策緊密結合，並能夠使用高級分析方法如時間序列分析、預測模型等。
3. 高級階段（業務理解）：在這個階段，專業人員應能夠不僅理解並解釋數據，還能夠領導跨部門的資料專案，並具備影響高層決策的能力。此外，能夠運用先進技術如深度學習、自然語言處理來解決複雜問題。



資料科學的角色類型與GenAI帶來的影響



蔡明順

2023年12月30日

資料科學的角色類型與GenAI帶來的影響

III. 生成式AI技術對資料科學的影響與新機會

1. 自動化數據分析和優化資料處理：AI技術能夠自動化許多資料處理和分析任務，從而提高效率並允許資料科學家專注於更複雜的問題，LLM可以協助資料清理和預處理步驟，減少人工錯誤，提高資料處理的效率和準確性。
2. 資料增強與模擬：在資料稀缺的領域，生成式AI能夠創建合成資料來擴充訓練集，用於訓練模型從而提高模型的性能和泛化能力。LLM可以模擬各種場景，幫助研究人員和決策者理解可能的結果和風險。
3. 新的內容生成：生成式AI能夠生成新的文件、圖像、程式碼和影音，為市場行銷、媒體產業和設計領域帶來革命性的變化，提供更多分析所需的資料，基於本地端資料加上RAG，可以為使用者提供個性化的內容和建議，提高客戶體驗和參與度。
4. 道德和隱私問題：生成式AI的能力也引發了關於資料隱私、假訊息、倫理、道德、偏見的新問題，需要有判斷力、查證與提升新技能和知識，才能發揮生成式AI的潛力。
5. 技能需求的轉變：資料科學家需要適應新的工具和技術，並在分析和建模時考慮生成式AI的潛力和限制，例如：提示詞工程(Prompt Engineering)、模型微調(Fine-Tuning)、檢索增強生成 (Retrieval-Augmented Generation, RAG)、向量資料庫(Vector Database)，都是接下來2-3年企業急需的技術。

IV. 產業界與學術界的差異

在產業界，資料分析通常用於支持業務決策、增強產品或服務、驅動創新。這要求資料專業人員不僅要有技術技能，還要有業務洞察力和戰略思考能力。

在學術界，資料分析更多被用於探索性研究，挑戰現有理論或開發新理論。這要求更強的研究技能，對方法論的深入理解，以及與其他學科的合作能力。

資料科學是一個多面向且不斷發展的領域，涉及各種技能和專業知識。為了成為這個領域的專業人士，重要的是要持續學習和適應新的技術和方法。生成式AI技術為資料科學帶來了新的變革機會，無論是在產業還是學術界，未來幾年都另外開闢一條新的道路。

最高薪職業/最佳職業



排名	職業	年薪中位數
1	Nurse Practitioner 執業護士	\$121,610
2	Financial Manager 財務經理	\$139,790
3	Software Developer 軟件開發人員	\$127,260
4	IT Manager IT經理	\$164,070
5	Physician Assistant 助理醫師	\$126,010
6	Medical and Health Services Manager 醫療衛生服務經理	\$104,830
7	Information Security Analyst 信息安全分 析師	\$112,000
8	Data Scientist 數據科學家	\$103,500
9	Actuary 精算師	\$113,990
10	Speech-Language Pathologist 語言病理 學家	\$84,140
11	Marketing Manager 市場經理	\$140,040
12	Statistician 統計師	\$98,920
13	Management Analyst 管理分析師	\$95,290
14	Genetic Counselor 遺傳顧問	\$89,990
15	Operations Research Analyst 運籌學分 析師	\$85,720



U.S. News: Best-Paying Jobs (2024)

<https://money.usnews.com/careers/best-jobs/rankings/best-paying-jobs>

U.S. News Best Jobs Rankings (2024)

<https://money.usnews.com/careers/best-jobs/rankings>

Top 10 Highest Paying Jobs in 2024

November 2, 2023 by Aditya Gupta

<https://worldwide10.com/highest-paying-jobs/>

1. Medical Specialists
2. Pharmacists
3. Dentists
- 4. Data Scientists**
5. Software Developers
6. Aerospace Engineers
7. Financial Managers
8. Petroleum Engineers
9. Pharmaceutical Sales Representatives
10. Artificial Intelligence and Machine Learning Specialists



2024年美國

106/118

最好和最有保障的10種工作

《大紀元》2024年美國最好和最有保障的10種工作

<https://www.epochtimes.com/b5/24/1/11/n14155760.htm>

總體排名：2024年十個最好的工作

1. 護理醫師 (Nurse practitioner)
2. 財務經理 (Financial manager)
3. 軟件開發人員 (Software developer)
4. IT經理 (IT manager)
5. 醫師助理 (Physician assistant)
6. 醫療衛生服務經理 (Medical and health services manager)
7. 資訊安全分析師 (Information security analyst)
- 8. 數據科學家 (Data scientist)**
9. 精算師 (Actuary)
10. 言語病理學家 (Speech-language pathologist)

總體排名：2024年最安全的工作

1. 護理醫師 (Nurse practitioner)
2. 統計員 (Statistician)
3. 獸醫技術員和技師 (Veterinary technologist and technician)
4. 軟件開發人員 (Software developer)
5. 資訊安全分析師 (Information security analyst)
6. 醫療衛生服務經理 (Medical and health services manager)
7. 物理治療師助理 (Physical therapist assistant)
8. 職業治療助理 (Occupational therapy assistant)
9. 醫師助理 (Physician assistant)
- 10. 數據科學家 (Data scientist)**



16 high-paying jobs for people who don't work well with others

16 high-paying jobs for people who don't work well with others (2017)

Source: Occupational Information Network (O*NET)

<https://finance.yahoo.com/news/16-high-paying-jobs-people-130400499.html>

16 HIGH-PAYING JOBS FOR PEOPLE WHO HATE PEOPLE

RANK	JOB	SOCIABILITY SCORE	AVERAGE SALARY
1	Computer-hardware engineer	64.5	\$115,080
2	Political scientist	63	\$114,290
3	Mathematician	58	\$105,810
4	Astronomer	60.5	\$104,740
5	Economist	60.5	\$101,050
6	Environmental economist	58.5	\$101,050
7	Geoscientist	64.5	\$89,780
8	Web administrator	64.5	\$86,510
9	Statistician	64	\$80,500
10	Molecular and cellular biologist	60.5	\$74,790
11	Financial quantitative analyst	65	\$69,470
12	Photonics technician	64	\$62,330
13	Poet, lyricist, or creative writer	44.5	\$61,240
14	Aircraft mechanic or service technician	64	\$60,170
15	Geological sample-test technician	56	\$56,470
16	Mine shuttle car operator	64	\$56,450

SOURCE: Occupational Information Network (O*NET)

BUSINESS INSIDER

勞動部職類別薪資調查動態查詢

<https://pswst.mol.gov.tw/psdn/>

職類別薪資調查動態查詢

薪資內涵及資料使用說明 全時受僱員工人數及薪資

各業全時受僱員工人數、薪資

步驟1：資料時期
 年度： 111年

步驟2：顯示項目
 全選 7月底受僱人數 7月經常性薪資 過去一年總薪資

步驟3：統計項目分類
 行業別編號： 依第10次修訂之行業標準分類

行業編號(全選)
 全選(部門別)
 (sub1)工業及服務業

職類別編號： 職類說明

職類編號(全選)
 全選(各職類)
 (100000)主管及監督人員
 全選(各職類)
 (112000)高階主管(總經理及總執行長)
 (120090)中階主管(經理)
 (312190)監督人員(含現場及辦公室主任、科/股長)
 (200000)專業人員
 全選(各職類)
 (211090)數理工程及地球科學研究人員
 (212001)精算師(取得正式資格者)
 (212091)統計及精算專業人員
 (213190)生命科學及生態環境研究人員
 (213200)農、林、漁、牧業專業人員
 (214100)品管、工管及生管工程師

下載(ODS檔) 重選 回動態查詢首頁 回勞動部

年8月至111年7月。

2022年我國十大高薪職業

	每月經常性薪資(元)	年薪(萬元)
航空駕駛員	278,954	401.1
精算師	181,921	325.1
醫師	156,176	230.2
船舶監管人員(含引水人員)	126,776	187.5
職業運動員	114,009	139.6
律師	94,802	161.3
電信工程師	83,939	162.2
統計及精算專業人員	75,892	131.5
航空機械工程師	75,888	120.5
地質探礦鑽探冶金工程師	72,348	115.5

來源：經濟日報(勞動部統計)

<https://www.cw.com.tw/article/5127359>

學統計，超簡單

至於學不學的會，又是另外一個故事了



博客來 最後2天 精選香氛 最高現折\$1000

 四大品牌 指定儲物罐 7/30 限定 單件9折/任3件88折

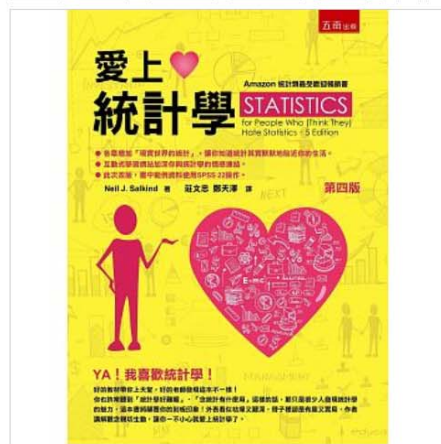
上半年暢銷 博客來選書 每日一句 今日66折 天天BUY

全站分類 旗艦店 全部

中文書 2018兒童暑期閱讀 新書 預購 排行榜 選書 即將出版 特價書 香港出版 讀者書評 出版社專區 分類總覽

保持心是癢癢的 **向——偷——一點時間** 2018詩人節69折起 6/15-7/31

博客來 > 中文書 > 商業理財 > 會計/統計 > 統計學 > 商品介紹



愛上統計學(4版)
 STATISTICS for People Who (think they) hate statistics · (5Edition)

作者：尼爾·J·薩爾金德 遺囑作者

原文作者：Neil J. Salkind

譯者：莊文忠 鄭天澤

出版社：五座 訂閱出版社新書快訊

出版日期：2017/03/25

語言：繁體中文

定價：680元

優惠價：95折55

本商品單次購買10

抵用購物金最

運送方式：

中午前訂 可明天拿

庫存 > 10

放入購物車

直接結帳

加入下次再買清單



誰都看得懂的

機率統計科普書籍



機率思考：大數據時代，不犯錯的決斷武器

CHANCING IT The Laws of Chance and How They Can Work for You

原文作者：Robert Matthews

譯者：高英哲

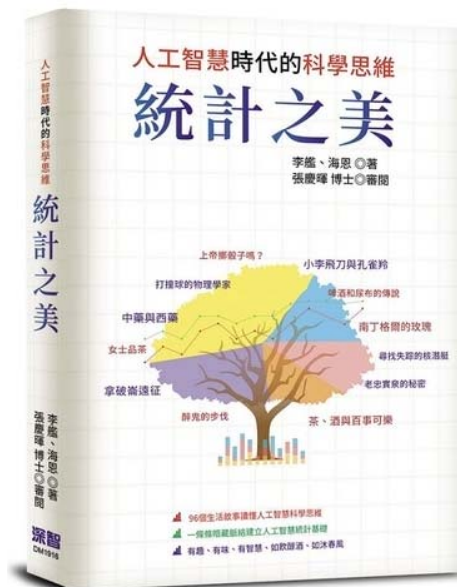
出版社：大牌出版

出版日期：2017/02/22

語言：繁體中文

本書的目的之一，就是透過揭示機率定理及其應用，使讀者理解日常生活中的機率事件。你會讀到如何運用這些定理預測巧合，幫助你在商場上與生活中做出更佳決策，並更能明智解讀從醫療診斷到投資建議等種種事務。

不懂機率，就等著當理盲的傻瓜



統計之美：人工智慧時代的科學思維

作者：李艦, 海恩

出版社：深智數位

出版日期：2019/09/15

語言：繁體中文

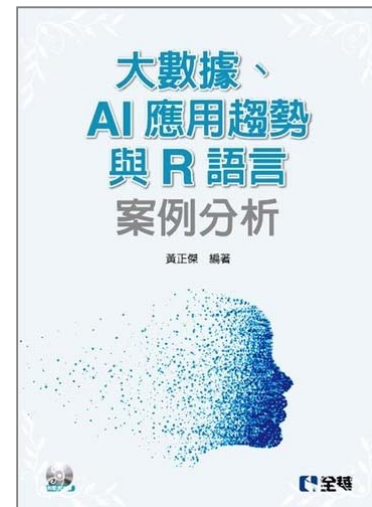
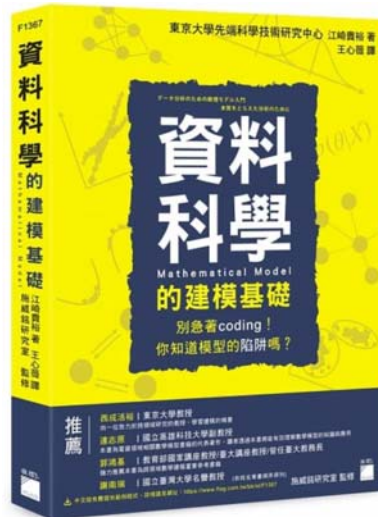
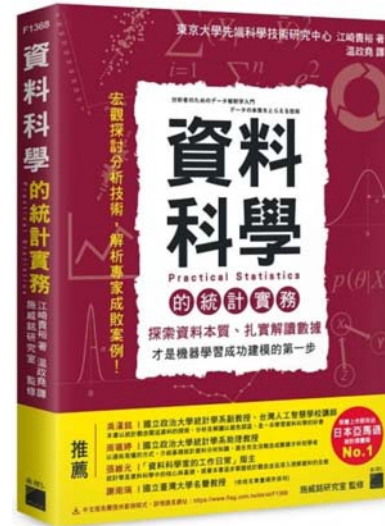
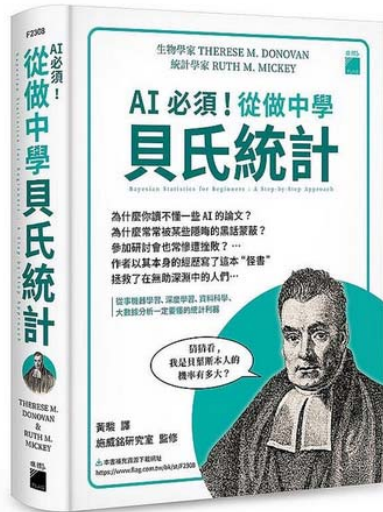
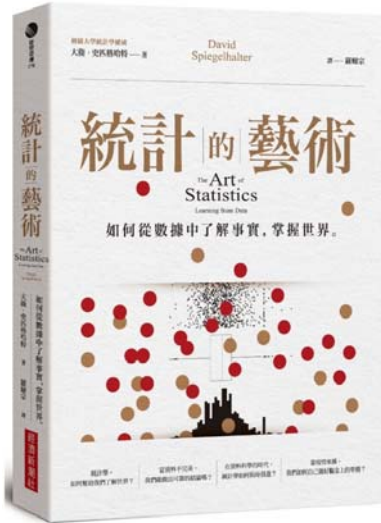
序言 Preface

統計學可真是一個尷尬的存在，常常處在各種鄙視鏈的低端。從數學的角度看，統計學中的數學原理太膚淺，撐死也就一個大數法則，一個中央極限定理，這也能算數學？從應用學科的角度看（例如，電腦、管理學），統計學太數學，一點也不應用。分析資料就好好分析資料，還要證什麼大數法則，什麼中央極限定理，這也能算應用？作為一名統計學的工作者，對這樣的觀點雖然並不苟同，但確實很慚愧。常常為此，反省自問：問題到底出在哪裡？我輩應該如何作為？

王漢生

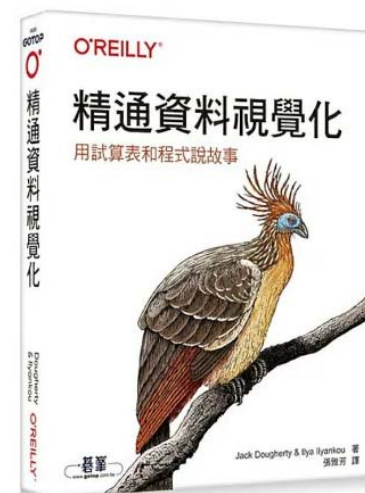
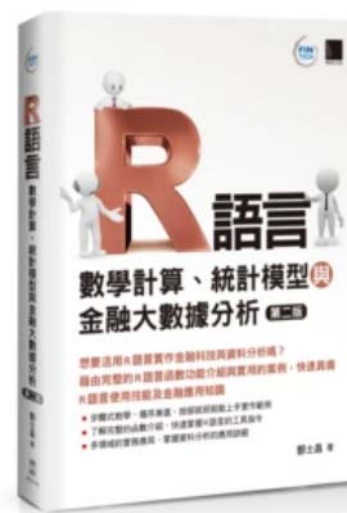
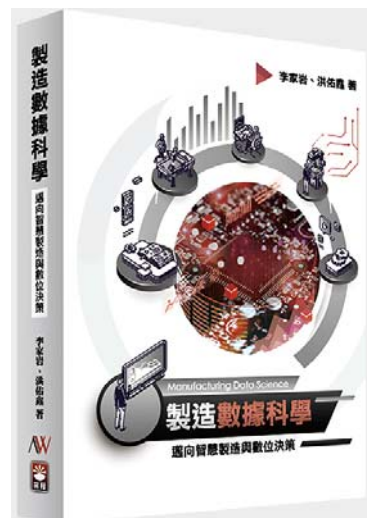
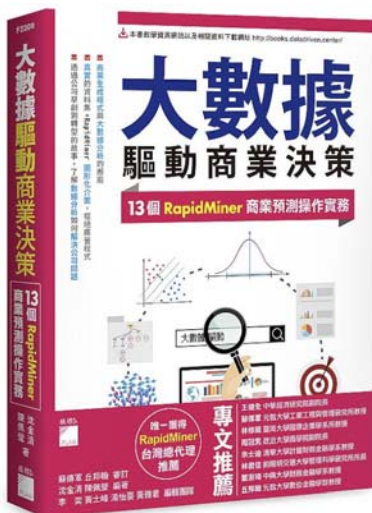
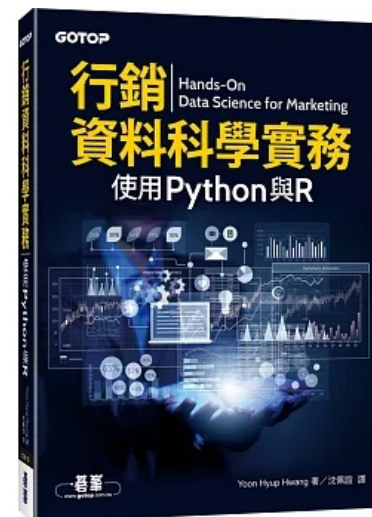
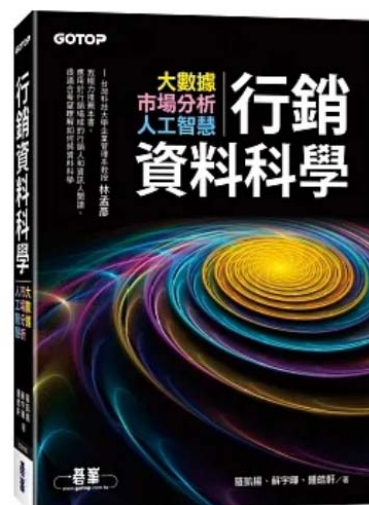
北京大學光華管理學院教授

統計、資料科學



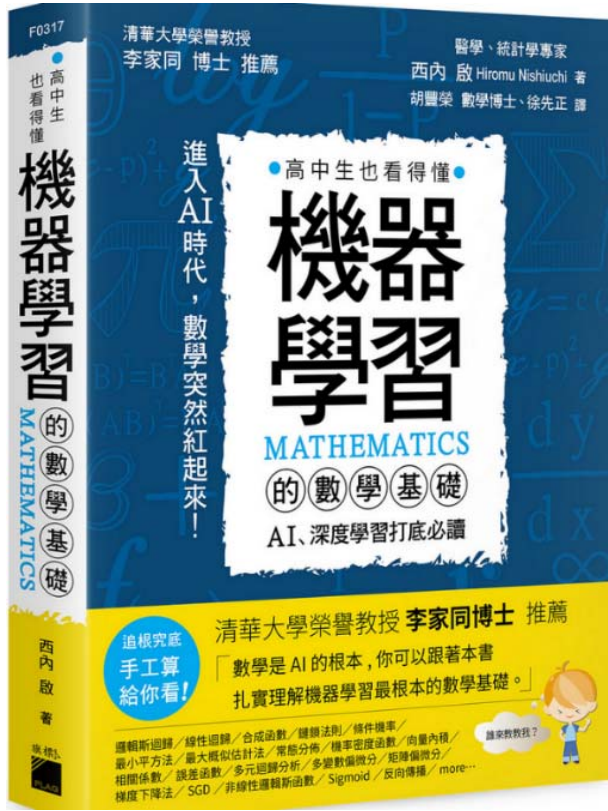
統計、資料科學於各領域的實務應用

112/118



機器學習的數學基礎

AI、深度學習打底必讀



機器學習的數學基礎：
AI、深度學習打底必讀
 醫學統計學專家 西內啟 著
 胡豐榮博士, 徐先正 合譯 ·
 出版商: 旗標科技出版
 (2020-01-31)

目錄：

序篇 AI、機器學習需要什麼樣的數學能力
 單元01 21世紀每個人都需要具備數學能力
 單元02 數學金字塔

第1篇 機器學習的數學基礎
 單元03 將事物用數字來表現
 單元04 將數字用字母符號代替
 單元05 減法是負數的加法, 除法是倒數的乘法
 單元06 機率先修班：集合
 單元07 機率先修班：命題的邏輯推理
 單元08 機率先修班：機率、條件機率與貝氏定理

第4篇 機械學習需要的 Σ 、向量、矩陣
 單元25 整合大量數據的 Σ 運算規則
 單元26 向量基本運算規則
 單元27 向量的內積
 單元28 向量內積在計算相關係數的應用
 單元29 向量、矩陣與多元線性迴歸
 單元30 矩陣的運算規則
 單元31 轉置矩陣求解迴歸係數

第2篇 機器學習需要的一次函數與二次函數
 單元09 座標圖與函數
 單元10 聯立方程式求解與找出直線的斜率與截距
 單元11 用聯立不等式做線性規劃
 單元12 從線性函數進入二次函數
 單元13 利用二次函數標準式求出最大值與最小值
 單元14 找出二次函數最適當的解
 單元15 用最小平方方法找出誤差最小的直線

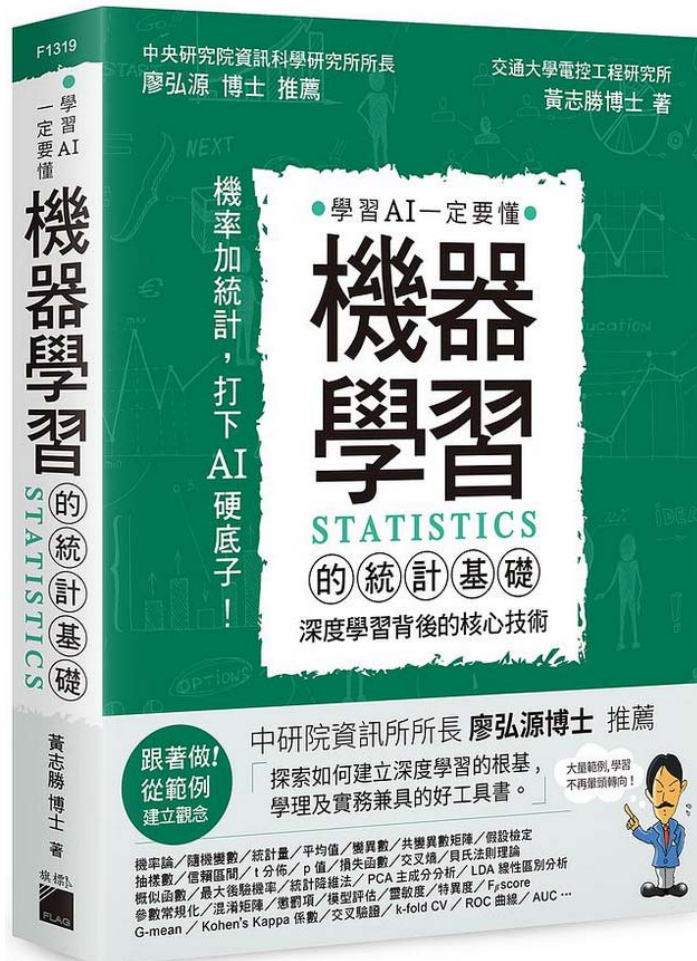
第5篇 機器學習需要的微分與積分
 單元32 函數微分找出極大值或極小值的位置
 單元33 n次函數的微分
 單元34 積分基礎－從幾何學角度瞭解連續型機率密度函數
 單元35 積分基礎－用積分計算機率密度函數
 單元36 合成函數微分、連鎖法則與代換積分
 單元37 指數函數、對數函數的微分積分
 單元38 概似函數與最大概似估計法
 單元39 常態分佈的機率密度函數
 單元40 多變數積分－雙重積分算機率密度函數係數

第3篇 機械學習需要的二項式定理、對數、三角函數
 單元16 二項式定理與二項式係數
 單元17 利用二項分布計算重複事件發生的機率
 單元18 指數運算規則與指數函數圖形
 單元19 用對數的觀念處理大數字
 單元20 對數的性質與運算規則
 單元21 尤拉數 e 與邏輯斯迴歸
 單元22 畢氏定理計算兩點距離
 單元23 三角函數的基本觀念
 單元24 三角函數的弧度制與單位圓

第6篇 深度學習需要的數學能力
 單元41 多變數的偏微分－對誤差平方和的參數做偏微分
 單元42 矩陣型式的偏微分運算
 單元43 多元迴歸分析的最大概似估計法與梯度下降
 單元44 由線性迴歸瞭解深度學習的多層關係
 單元45 多變數邏輯斯迴歸與梯度下降法
 單元46 神經網路的基礎－用非線性邏輯斯函數組合出近似函數
 單元47 神經網路的數學表示法
 單元48 反向傳播－利用隨機梯度下降法與偏微分連鎖法則

機器學習的統計基礎

深度學習背後的核心技術



機器學習的統計基礎：深度學習背後的核心技術

作者：黃志勝

出版社：旗標

出版日期：2021/11/15

第 1 章 機器與深度學習常用的數學基礎

向量/矩陣/矩陣分解/特徵分解/奇異值分解

第 2 章 機器學習相關機率論

隨機試驗/條件機率與貝氏定理/隨機變數/機率分布與機率密度函數

第 3 章 機器學習常用的統計學 (一)

期望值/相關係數與共變異數

第 4 章 機器學習常用的統計學 (二)

母體與樣本估計/信賴區間/假設檢定/計算p值

第 5 章 機器學習常用的資料處理方式

資料標準化/Min-max 正規化/非線性轉換

第 6 章 機器與深度學習常用到的基礎理論

最大概似估計/貝氏法則理論與最大後驗機率/距離和相似度計/損失函數

第 7 章 迴歸分析 Regression

簡單線性迴歸分析/最小平方方法/最大概似函數估計法

第 8 章 分類 Classification

第 9 章 統計降維法 Dimension Reduction

第 10 章 類神經網路 Artificial Neural Network

第 11 章 梯度下降法 Gradient Descent

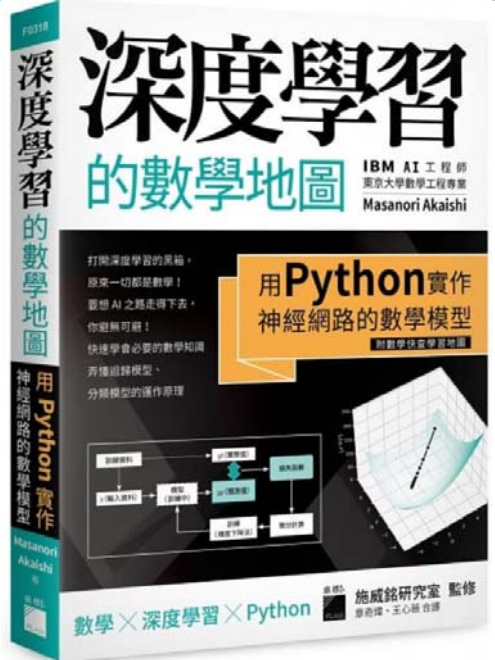
第 12 章 倒傳遞學習法 Backpropagation

第 13 章 參數常規化 Parameter Regularization

第 14 章 模型評估 Model Validation

深度學習的數學地圖

用Python實作神經網路的數學模型



深度學習的數學地圖：
用 Python 實作神經網路的數學模型
 作者：Masanori Akaishi
 譯者：章奇煒, 王心薇
 出版社：旗標
 出版日期：2020/05/28

第6章 機率、統計

- 6.1 隨機變數與機率分佈
- 6.2 機率密度函數與累積分佈函數
專欄 Sigmoid 函數的機率密度函數
- 6.3 概似函數與最大概似估計法
專欄 為何概似函數的極值是求最大值，而不是最小值？

目錄

【導入篇 機器學習快速指引】

第1章 機器學習入門

- 1.1 何謂機器學習
 - 1.1.1 何謂機器學習模型
 - 1.1.2 機器學習的訓練方法
 - 1.1.3 監督式學習的迴歸、分類模型
 - 1.1.4 訓練階段與預測階段
 - 1.1.5 損失函數與梯度下降法
- 1.2 第一個機器學習模型：簡單線性迴歸模型
- 1.3 本書討論的機器學習模型
- 1.4 數學是深度學習的核心
- 1.5 本書架構

第4章 多變數函數的微分

- 4.1 多變數函數
- 4.2 偏微分
- 4.3 全微分
- 4.4 全微分與合成函數
- 4.5 梯度下降法 (GD)
專欄 梯度下降法與局部最佳解

【實踐篇 機器學習、深度學習實作】

- 第7章 線性迴歸模型 (迴歸)
- 第8章 邏輯斯迴歸模型 (二元分類)
- 第9章 邏輯斯迴歸模型 (多類別分類)
- 第10章 深度學習

【發展篇 實務上的解決方法】

第11章 以實用的深度學習為目標

【理論篇 數學速學課程】

第2章 微分、積分

- 2.1 函數
 - 2.1.1 函數運作行為
 - 2.1.2 函數的圖形
- 2.2 合成函數與反函數
 - 2.2.1 合成函數
專欄 合成函數的表示法
 - 2.2.2 反函數
- 2.3 微分與極限
 - 2.3.1 微分的定義
 - 2.3.2 函數值增量與微分的關係
 - 2.3.3 切線方程式
專欄 切線方程式與訓練階段、預測階段的關係
- 2.4 極大值與極小值
- 2.5 多項式的微分
 - 2.5.1 x^n 的微分 (n 是正整數)
 - 2.5.2 微分計算的線性關係與多項式的微分
 - 2.5.3 x^r 的微分 (r 是實數)

第5章 指數函數、對數函數

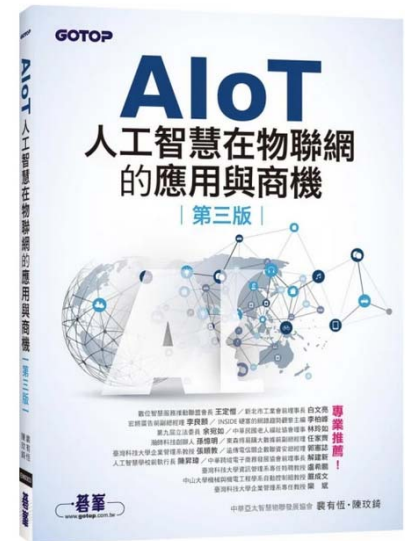
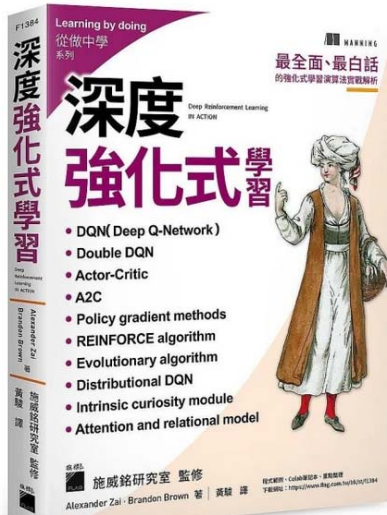
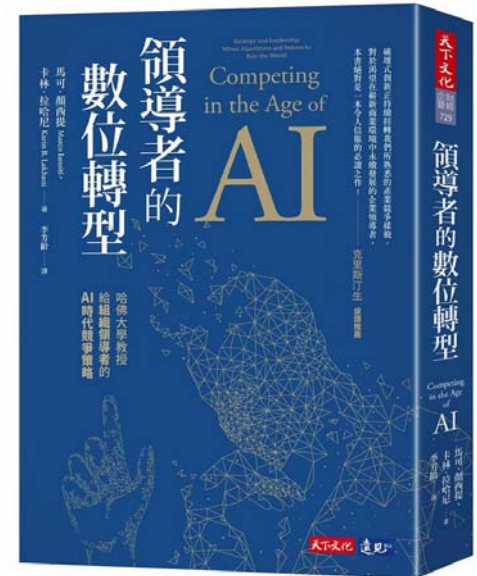
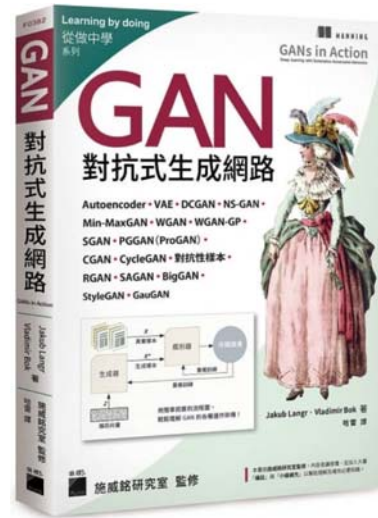
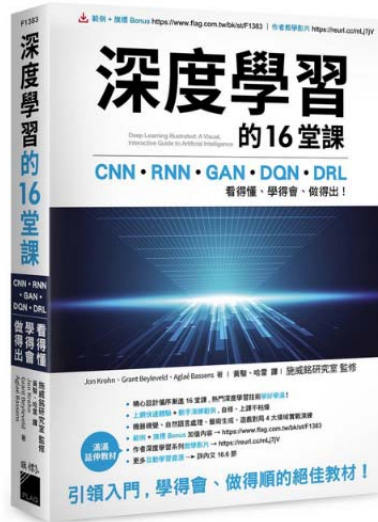
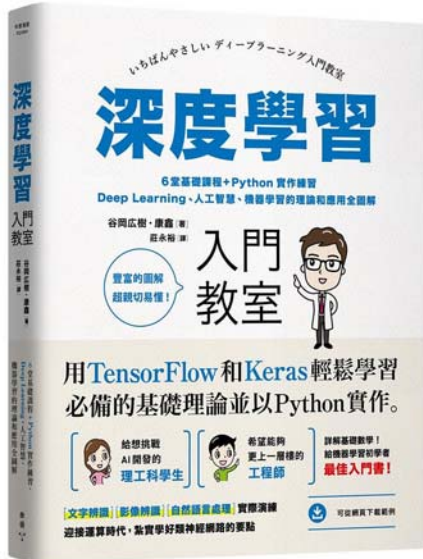
- 5.1 指數函數
 - 5.1.1 連乘的定義與公式
 - 5.1.2 連乘觀念的推廣
 - 5.1.3 將連乘寫成指數函數形式
- 5.2 對數函數
專欄 對數函數的意義
- 5.3 對數函數的微分
專欄 用 Python 來計算尤拉數 e
- 5.4 指數函數的微分
專欄 以 e 為底的指數函數也可用 \exp 表示
- 5.5 Sigmoid 函數
- 5.6 Softmax 函數
專欄 Sigmoid 和 Softmax 函數的關係

第3章 向量、矩陣

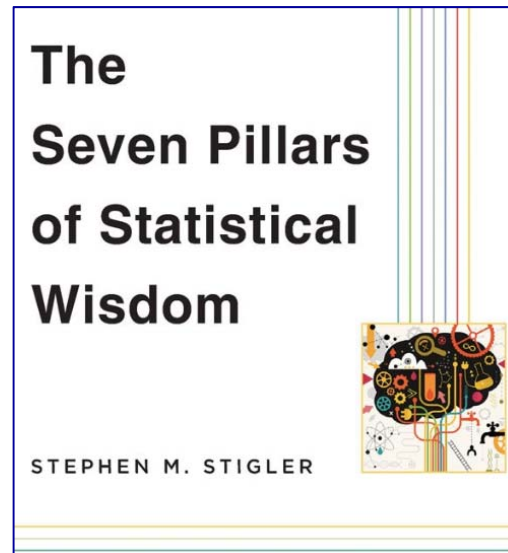
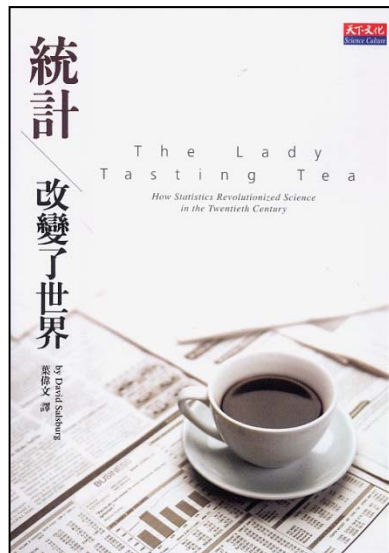
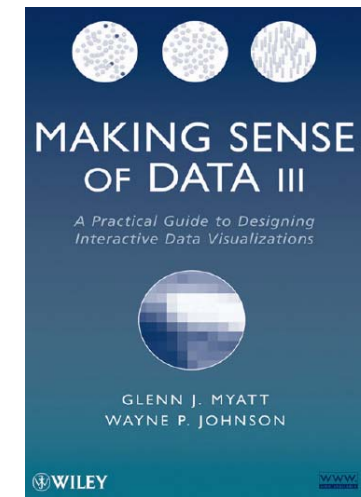
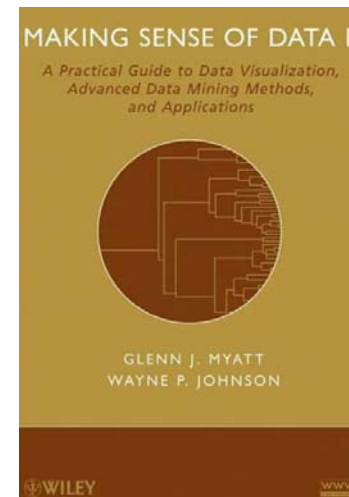
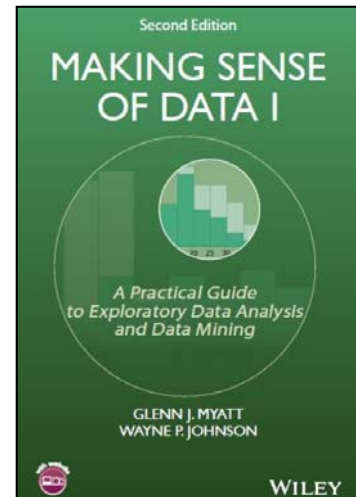
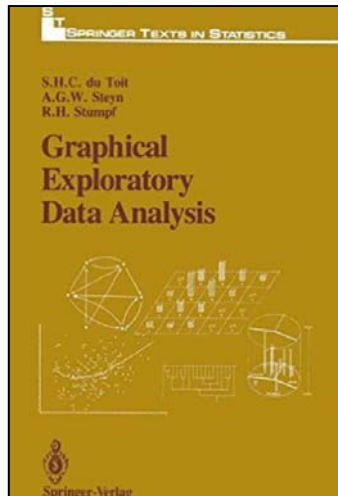
- 3.1 向量入門
 - 3.1.1 何謂向量
 - 3.1.2 向量的標記法
 - 3.1.3 向量的分量
 - 3.1.4 往多維擴展
 - 3.1.5 分量的符號
- 3.2 向量和、向量差、純量乘積
 - 3.2.1 向量和
 - 3.2.2 向量差
 - 3.2.3 向量與純量的乘積
- 3.3 向量的長度 (絕對值) 與距離
 - 3.3.1 向量的長度 (絕對值)
 - 3.3.2 Σ 可整合冗長的加法算式
 - 3.3.3 向量間的距離
- 3.4 三角函數
 - 3.4.1 三角比：三角函數的基本定義
 - 3.4.2 單位圓上的座標
 - 3.4.3 三角函數的圖形
 - 3.4.4 用三角函數表示直角三角形的邊長
- 3.5 向量內積
 - 3.5.1 向量內積的幾何定義
 - 3.5.2 用分量來表示內積公式
- 3.6 餘弦相似性
 - 3.6.1 兩個二維向量的夾角
 - 3.6.2 n 維向量的餘弦相似性
專欄 餘弦相似性的應用範例
- 3.7 矩陣運算
 - 3.7.1 一個輸出節點的內積表示法
 - 3.7.2 三個輸出節點的矩陣相乘



根據自己需求找一些書來學習



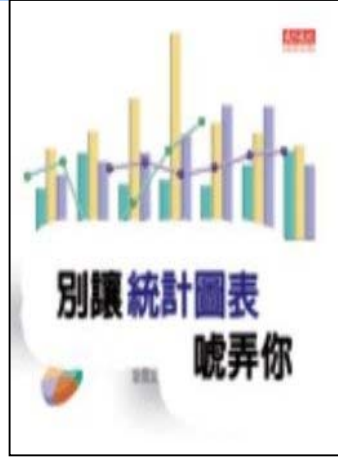
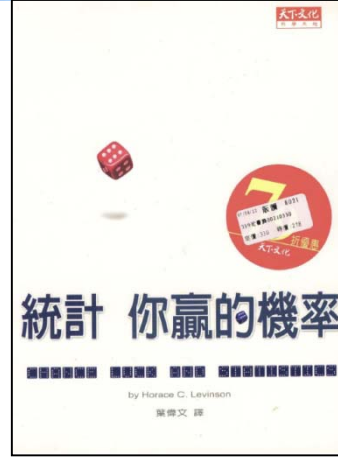
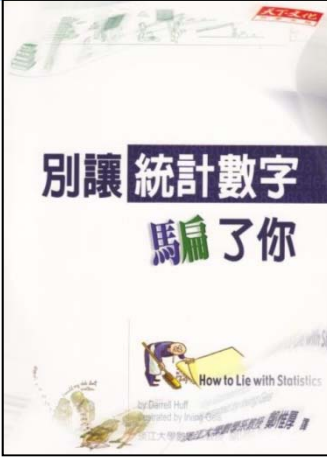
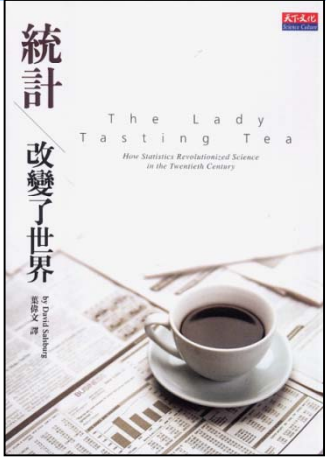
給統計背景的推薦書目



- 1 **AGGREGATION** From Tables and Means to Least Squares
- 2 **INFORMATION** Its Measurement and Rate of Change
- 3 **LIKELIHOOD** Calibration on a Probability Scale
- 4 **INTERCOMPARISON** Within-Sample Variation as a Standard
- 5 **REGRESSION** Multivariate Analysis, Bayesian Inference, and Causal Inference
- 6 **DESIGN** Experimental Planning and the Role of Randomization
- 7 **RESIDUAL** Scientific Logic, Model Comparison, and Diagnostic Display



期待下一場的科學革命



**"只要有米倉，就會有老鼠；
只要有數據，就會發展處理數據的方法。"** (趙民德)

