大數據（巨量資料, Big Data）

大數據的源起[1]

人們每天上傳至雲端的檔案數量，多達一億張相片、十億份文件… 更別提數位影音、交易、生物醫療等等，截至 2012 年，每天全球所創造的資料量高達 2.5 艾位元組（exabyes, 即2.5×10^{18} bytes）。但資料量大就是大數據嗎？

「儲存成本」與「資料取得成本」因科技進步而大幅下降，造就了這個年代大數據的興起。30 年前，1TB 檔案存儲的成本為 16 億美金，如今一個 1TB 的硬碟不到 100 美金。Google 每天要處理超過 24 千兆位元組的資料，Facebook 每天處理 500 億張的上傳相片，每天人們在網站上點擊”讚”（Like）按鈕、或留言次數大約有數十億次。千禧年開始，科學家發現：仰賴於科技的進步（感測器、智慧型手機），資料的取得成本相較於過去開始大幅地下降—過去十多年蒐集的資料，今朝一夕之間即能達成。如何**「取得正確資料」、「儲存」、「挖掘海量數據」，並成功地「溝通」分析結果**，就成為新的瓶頸與研究重點。

大數據的相關報導與研究如雨後春筍的冒出，為了有系統的說明，請參考圖一，我們依序由大數據的取得、儲存、探勘、與溝通這四個流程說明

圖一 解析大數據的流程

一、 大數據資料的取得：

什麼是大數據? [1-3]

大數據（Big Data）又被稱為巨量資料，概念其實就是過去 10 年前廣泛用於企業內部的資料分析、商業智慧（Business Intelligence）和統計應用之大成。但**大數據現在不只是資料處理工具，更是一種企業思維和商業模式**，因為資料量急速成長、儲存設備成本下降、軟體技術進化和雲端環境成熟等種種客觀條件就位，方才讓資料分析從過去的洞悉歷史進化到預測未來，甚至是破舊立新，開創從所未見的商業模式。至於「大」是多大，則各家定義不一，有兆位元組(TB)、千兆位元組(PB)、百萬兆位元組(EB)、甚至更大的規模單位；但其實絕大多數的企業，都不符合這個標準，大企業如 eBay、亞馬遜或 AT&T 或許符合大數據的標準。但其實資料量只是大數據的其中一個面向，大數據揭示的是一種「資料經濟」的精神，而非只是「大」。無論企業規模大小，**我們應注重的不僅是數據量本身，而應將「大數據」作為在科學研究與商業方法的運營心態**；大數據需要全新的處理方式，以新型的儲存運算方法分析數據、產出溝通圖表，並將該分析結果視為一種戰略資產。

大數據的特性?

目前大部份的機構將大數據的特性歸類為「3V」，請參考圖二，分別是資料量（Volume）、資料類型（Variety）與資料傳輸速度（Velocity），另外還有人另外加上 Veracity（真實性）和 Value（價值）兩個 V。

圖二 大數據的特性

1. Volume 資料量

無論是天文學、生物醫療、金融、聯網物間連線、社群互動…每分每秒都正在生成龐大的數據量，如同上述所說的 TB、PB、EB 規模單位。

2. Variety 資料多元性

舉一個規律簡單的例子:

資料類型為｜ 0 ｜ 0 ｜ 1 ｜ 0 ｜ 0 ｜ 1 ｜ 0 ｜ 0 ｜ 1 ｜……

就算上述資料量高達 1 TB，採用傳統統計方法仍能很容易地找到資料規律。因此，真正困難的問題在於分析多樣化的資料，從文字、位置、語音、影像、圖片、交易數據、類比訊號… 等結構化與非結構化包羅萬象的資料，彼此間能進行交互分析、尋找數據間的關聯性。

3. Velocity 資料即時性

大數據亦強調資料的時效性。隨著使用者每秒都在產生大量的數據回饋，過去三五年的資料已毫無用處。一旦資料串流到運算伺服器，企業便須立即進行分析、即時得到結果並立即做出反應修正，才能發揮資料的最大價值。由於進行資料分析的工作時，通常是由資料科學團隊向企業的 IT 部門登入企業伺服器取得資料，一般企業在資料儲存上的量與多樣性已難以達到，在「即時性」這一點上便不符合。唯有企業內部自建即時的資料分析團隊並隨時產出分析反饋，方能稱作大數據分析。

4.Veracity 資料真實性

Veracity 討論的問題包括：資料收集的時候是不是有資料造假、即使是真實資料，是否能夠準確的紀錄、資料中有沒有異常值、有異常值的話該怎麼處理… 等等。大數據的資料特質和傳統資料最大的不同是，資料來源多元、種類繁多，大多是非結構化資料，而且更新速度非常快，導致資料量大增。而要用大數據創造價值，不得不注意數據的真實性。

二、 大數據的儲存：

大數據的發展重點[1,2]

講到大數據，我們便不能不提與之息息相關的軟體技術——「Hadoop」。

圖三 Hadoop 標誌

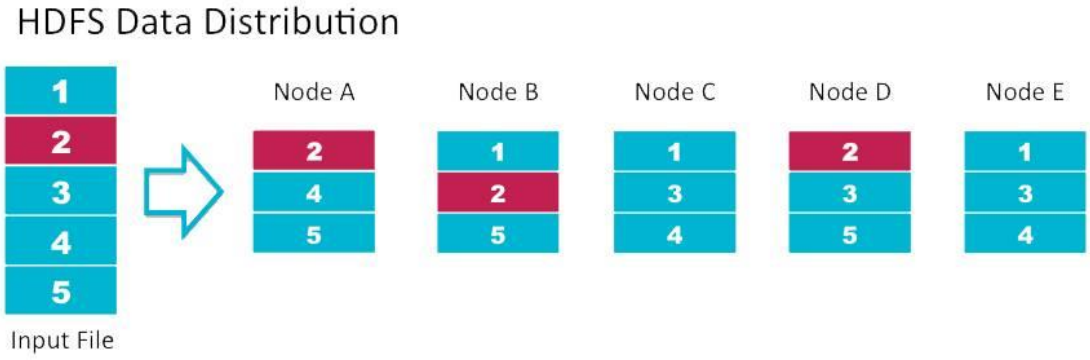
Hadoop 由 Java 語言撰寫，是 Apache 軟體基金會發展的開源軟體框架。不但免費、擴充性高、部屬快速，同時還能自動分散系統負荷，在大數據實作技術上非常受歡迎。

Hadoop 的核心主要由兩個部分所構成：

- 資料儲存:「Hadoop 分散式檔案系統(Hadoop Distributed File System)」
- 資料處理：「Hadoop MapReduce」

1. Hadoop 分散式檔案系統 (Hadoop Distributed File System, HDFS)：

由多達數百萬個叢集(Cluster)所組成，參考圖四，每個叢集有近數千台用來儲存資料的伺服器，被稱為「節點」(Node)。其中包括主伺服器(Master Node)與從伺服器(Slave Node)。每一份大型檔案儲存進來時，都會被切割成一個個的資料塊 (Block)，並同時將每個資料塊複製成多份、放在從伺服器上保管。

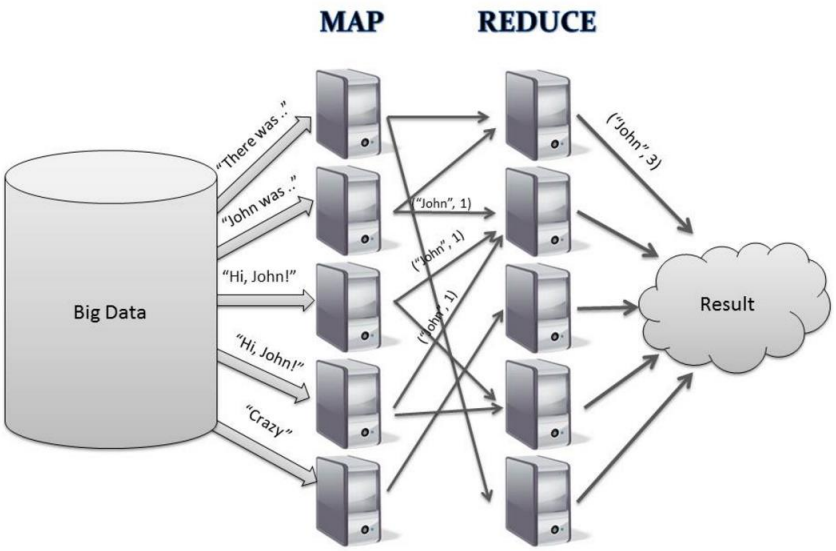


圖四 Hadoop 分散式檔案系統資料儲存方塊圖

當某台伺服器出問題時、導致資料塊遺失或遭破壞時，主伺服器就會在其他從伺服器上尋找副本複製一個新的版本，維持每一個資料塊都備有好幾份的狀態。簡單來說， Hadoop 預設的想法是所有的 Node 都有機會壞掉，所以會用大量備份的方式預防資料發生問題。另一方面，儲存在該系統上的資料雖然相當龐大、又被分散到數個不同的伺服器，但透過特殊技術，當檔案被讀取時，看起來仍會是連續的資料，使用者不會察覺資料是零碎的被切割儲存起來。

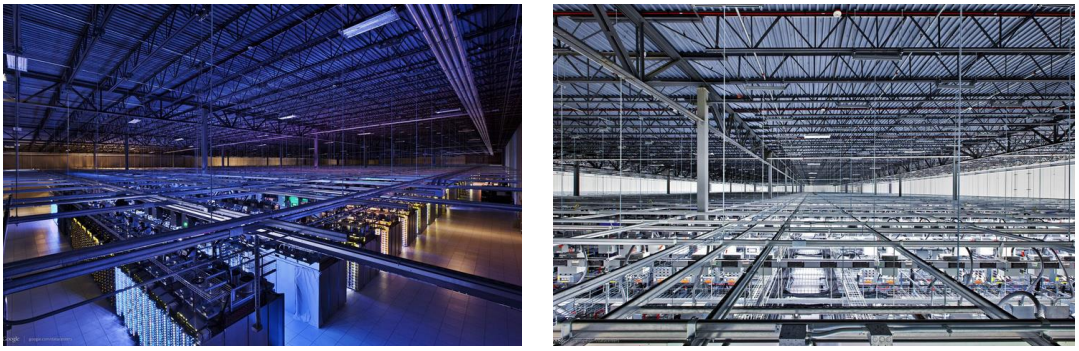
2. Hadoop MapReduce：

MapReduce 是一種計算模型，分為 Map 和 Reduce 兩項功能。「Map」功能會先將大資料拆成小資料，並以 Key-Value 格式備用。比如有數千萬份的資料傳入，Map 會計算每個字出現的次數；比如 computer 這個字出現了一次、便以（computer, 1）這樣的（Key, Value） 格式表示。「Reduce」則是彙整，意即彙整所有相同的 Key 並計算出現的總次數，參考圖五。

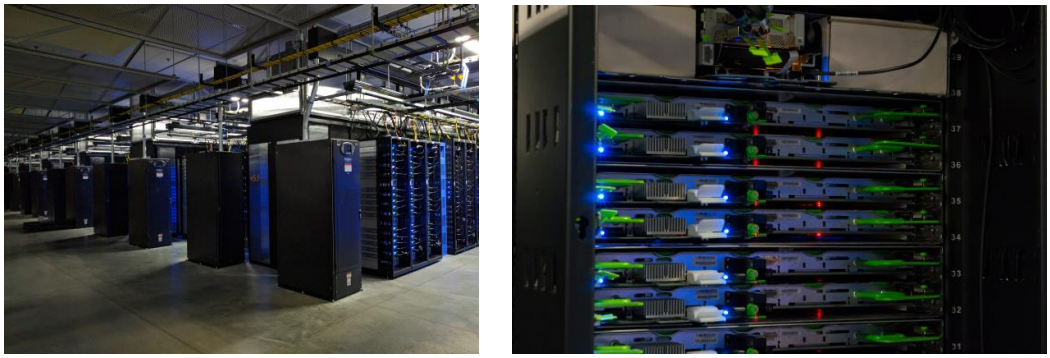


圖五 MapReduce 處理過程

簡單來說，Map 僅是在各節點上計算少量數據，而 Reduce 則是統計各地數據、將結果送回主伺服器進行公布。MapReduce 的好處在於無須將所有資料都搬回中央去運算，而能在各地先簡單的處理完畢後、再回傳數據，如此更有效率。**總而言之，Hadoop 分散式檔案儲存系統（HDFS）是一個超大型的儲存空間，並透過 Hadoop MapReduce 進行運算。**Hadoop 成功解決了檔案存放、檔案備份、資料處理等問題，因而應用廣泛，成為大數據的主流技術。Amazon、Facebook、IBM 和 Yahoo 皆採取 Hadoop 作為大數據的環境，請參考圖六，為 google 數據中心的圖片，圖七為 Facebook 數據中心的照片。



圖六 google 數據中心



圖七 facebook 數據中心

Apache 軟體基金會最近發展的「Spark」隱隱有取代 Hadoop MapReduce 的態勢。在大規模資料的計算、分析上，排序作業的處理時間，一直是個重要的指標。相較於 Hadoop MapReduce 在做運算時需要將中間產生的數據存在硬碟中，因此會有讀寫資料的延遲問題。 Spark 使用了記憶體內運算技術，能在資料尚未寫入硬碟時即在記憶體內分析運算，速度比 Hadoop MapReduce 可以快到 100 倍。許多人誤以為 Spark 將取代 Hadoop。然而，Spark 沒有分散式檔案管理功能，因而必須依賴 Hadoop 的 HDFS 作為解決方案。作為與 Hadoop 相容而且執行速度更快的開源軟體，來勢洶洶的 Spark 想取代的其實是 Hadoop MapReduce。另一方面，Spark 提供了豐富而且易用的 API，更適合讓開發者在實作機器學習演算法。

然而建置大數據架構與環境的確所費不貲，一般中小企業通常無法輕易投入鉅額成本，**但大數據時代的精神在於如何妥善利用既有或非傳統資料，從中挖掘出新商機，因此即使是中小企業甚或者是新創企業，都能在大數據時代用「大數據」**。中研院資訊科學研究所研究員陳昇瑋即指出，在絕大多數情況下，大數據專案其實不需要建置 Hadoop 系統，特別是台灣的社群媒體沒那麼發達，而是直接採用國外的居多，資料都不在自己手上，與其盲目追求技術和工具，不如**先用小量資料去驗證一個概念，是否能將資料轉換成商業機會**，再來決定要不要建置大數據的作業環境。大數據領域權威麥爾苟伯格（Viktor Mayer-Schönberger）在《大數據》一書中便提及[3]，大公司有巨量資料的規模優勢，但小公司有成本及創新上的優勢，小公司因為速度夠快、靈活度高，就算維持小規模，還是能夠蓬勃發展，所以**切勿陷入大數據迷思，與其急著想用數據變現，不如先回頭看看自己企業內部的問題為何，先定義問題，再試圖用數據找解方**。。

三、 大數據的探勘：[4-7]

介紹完了資料儲存的基礎架構後，讓我們來看看資料探勘(data mining)，資料探勘是利用分析技術來發掘資料間未知的關聯性與規則，傳統會用統計的方式來做資料探勘，現在會使用機器學習的方法來處理大量資料。由於機器學習可處理多維度資訊，並發掘多變動因素之間的關聯性，因此非常適用於資料探勘。利用資料探勘的技術並加入機器學習，由大數據中挖掘資料彼此的相關性，以改善科學或商業決策。

1. 資料探勘(data mining)：

在大數據中如何擷取過去未被發掘且有價值的隱藏資訊，並透過分析使潛在資訊呈現而提供未來的決策判斷，這時就必須使用資料探勘的技術；資料探勘是一個跨領域的一門知識，它是用機器學習、統計學和資料庫的交叉方法在相對較大型的資料集(data set)中發現模式(Pattern)的過程。如圖八所示，是資料探勘的三大步驟

圖八 資料探勘的步驟

一般而言，資料探勘功能可包含下列五項[7]：

A. 區隔化(segmentation)或群集化(clustering) ：
將異質母體中區隔為較具同質性之群組(clusters)。同質分組相當於行銷術語中的區隔化(segmentation)，但是，假定事先未對於區隔加以定義，而資料中自然產生區隔。使用的技巧包括 k-means 法及 agglomeration 法。

B. 分類(classification) ：
按照分析對象的屬性分門別類加以定義，建立類組(class)。例如，依信用卡申請者的風險屬性，申請信用卡相同群集內，區分為高度風險申請者，中度風險申請者及低度風險申請者。使用的技巧有決策樹(decision tree)，記憶基礎推理(memory－based reasoning)等。

C. 推估(estimation)：
根據既有連續性數值之相關屬性資料，以獲致某一屬性未知之值；推估是用來猜測現在的未知值。例如按照信用申請者之教育程度、行為別來推估其信用卡消費量。使用的技巧包括統計方法上之相關分析、迴歸分析及類神經網路方法。

D. 預測(prediction)：
根據對象屬性之過去觀察值來推估該屬性未來之值；預測是用來猜測未來的某一未知值。例如由顧客過去之刷卡消費量預測其未來之刷卡消費量。使用的技巧包括迴歸分析、時間數列分析及類神經網路方法，推估與預測在很多時候可以使用相同的演算法。

E. 關聯分組(affinity grouping)：

從所有物件決定那些相關物件應該放在一起。例如超市中相關之盥洗用品(牙刷、牙膏、牙線)，放在同一間貨架上。在客戶行銷系統上，此種功能係用來確認交叉銷售(cross selling)的機會以設計出吸引人的產品群組。

以下細分探勘過程中的 8 個細部步驟[4]：

A. 資訊收集：

根據確定的資料分析物件，萃取出在資料分析中所需要的特徵資訊，然後選擇合適的資訊收集方法，將收集到的資料存入資料庫，對於大數據資料，選擇一個合適的資料儲存和管理方法與工具是很重要的。

B. 資料整合：

把不同來源、格式、特點性質的資料在實體上有效率且安全的集中，進一步為提供企業全面的資料共用。

C. 資料精簡：

如果執行多數資料探勘演算法，即使在少量資料上也需要很長的時間，而做商業營運的資料探勘，其資料量通常很大，資料精簡技術可以用來得到資料集的精簡表示，雖然精簡表示，但仍然盡力保存原資料的完整性，並且精簡後執行資料探勘的結果與精簡前執行結果相同或幾乎相同。

D. 資料清理：

在資料庫中的資料有一些是不完整(有興趣的資料缺少屬性值)、含雜訊(包含錯誤屬性)或不一致(同樣資訊，不同表示法)的，因此需要進行資訊清理，將完整、正確與一致的資料存入資料庫中，不然探勘的結果會引入偏差。

E. 資料轉換：

透過特徵萃取 (Feature Extraction) 與特徵選擇 (Feature Selection)的方式將資料轉成適用於資料探勘的形式

E.1 特徵萃取 (Feature Extraction)

是從資料中挖出可以用的特徵，比如每個會員的性別、年齡、消費金額等；再把特徵量化、如性別可以變成 0 或 1 ，如此以來每個會員都可以變成一個多維度的向量。

E.2 特徵選擇 (Feature Selection)

經過特徵萃取後，根據機器學習模型學習的結果，去看什麼樣的特徵是比較重要的。若是要分析潛在客戶的話，那麼該客戶的消費頻率、歷年消費金額…等可能都是比較重要的特徵，而性別和年齡的影響可能便不會那麼顯著。

步驟(A)－(E)又合稱資料前置處理。

F. 資料探勘的過程：

資料科學家會根據所要解決的問題、擁有的資料類型進行衡量評估，選擇合適的分析工具，如統計方法、事例推論、規則推論、決策樹、模糊集合、類神經網路等等演算法處理資訊，進而得到有用的分析資訊。

G.模式評估：

由商業的角度，請產業專家檢驗資料探勘的正確性。

H.知識表示：

將資料探勘的分析結果以視覺化的方式來呈現給使用者，做為決策判斷的依據。

資料探勘的方法：

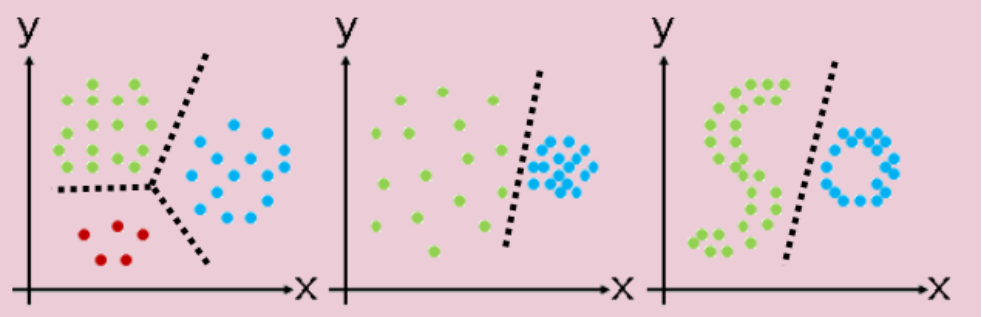
依照資料探勘學術會議(IEEE international conference on Data miming, ICDM)在 2006 選出該領域的十大經典資料探勘演算法，分別是：C 4.5、K-means、SVM、Apriori、EM、PageRank、AdaBoost、KNN、Naïve Bayess、CART 演算法，除了這些之外，我們也補充幾個常見的演算法，我們試著歸納列表中的演算法，有的演算法分類沒有那麼明顯，表中僅供參考，如果需要更深入資訊，請參考學術論文：

功能	代表演算法
群集化(clustering)	K-means、EM、BIRCH
分類(classification)	C 4.5、SVM、Naïve Bayess、CART、AdaBoost、KNN
推估(estimation)	統計方法上之相關分析、迴歸分析及類神經網路方法
預測(prediction)	迴歸分析、時間數列分析及類神經網路
關聯分組(affinity grouping)	Apriori、FP-Growth、PageRank
其他	PrefixSpan 序列、CBA 聚合、Finding reduct 粗糙集、gSpan 圖探勘

以下簡介群集化與分類這兩種資料探勘最重要的方法[8]

A.群集(clustering)：

如圖九所示，將所有數據進行分組，相似數據歸類於同一組，每一筆數據只屬於某特定一組，而分開的組稱作一個「群集 cluster」。

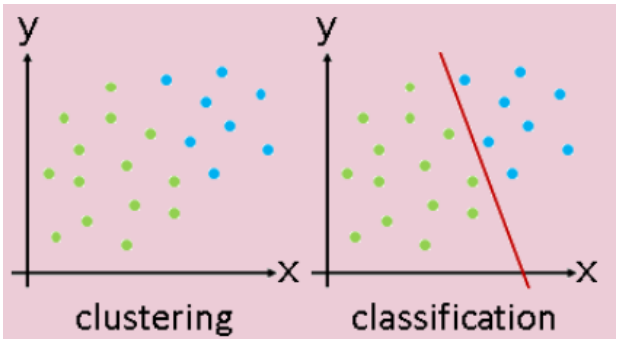


圖九 資料分群結果，分 3 群、2 群、2 群

群集演算法的基本原理有兩個： 1) 第一類是近朱者赤、近墨者黑，距離越近，推定為越相似；2)第二類是不斷切割群集，鄰居越密集，推定為越相似。

B. 分類(classification)：

在已知群集，找到分界線，如圖十所示，同一個群集內，分兩個不同的類別，每筆資料分屬於各自的類別。

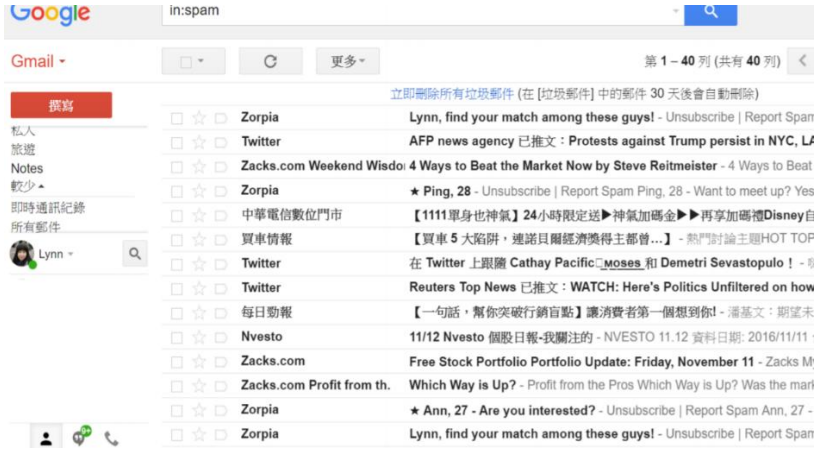


圖十 群集之後繼續接著分類

相似數據群聚，相異數據也漸漸隔離。最後出現了群聚中心，也出現了隔離界線。找到分界線之後，對於一筆新的數據，就利用分界線來決定其類別，這就是分類的主要功能。分類應用十分廣泛，是世上使用最多的演算法之一，也是當前的研究方向之一，應用於商業方面，只要把客戶、人群、地區，商品按照不同屬性區分開來的場景都可以使用分類演算法，應用於機場安全檢查系統方面就屬人臉辨識技術，請參考圖十一，警方只要把犯罪份子的臉部資訊輸入到影像資料庫中，當犯罪份子出現，系統就可以輕易辨識出來，雖然人臉辨識還需要具備其他技術，但是其主要技術還是源於分類演算法。

圖十一 機場人臉辨識系統

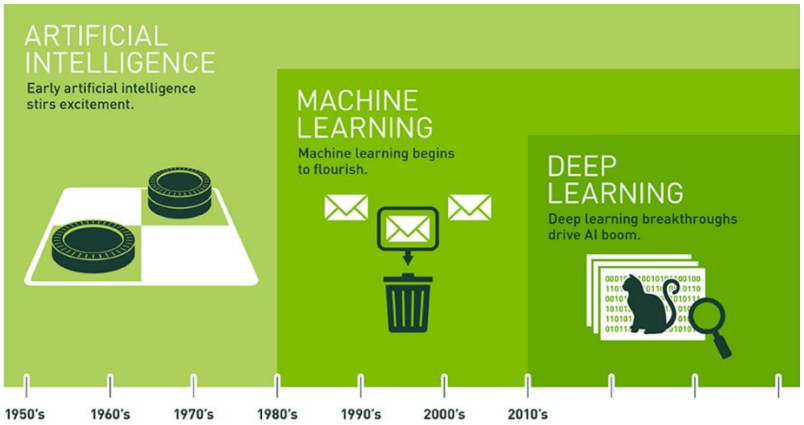
GMAIL 中垃圾信件會自動被丟入垃圾桶[5]



圖十二 信件分類系統

2.機器學習(machine learning)：[5][9][10]

「機器學習」是一門涵蓋電腦科學、統計學、機率論、博弈論等多門領域的學科，從 1980 開始蓬勃興起。機器學習之所以能興起，也歸功於硬體儲存成本下降、運算能力增強（包括本機端與雲端運算），加上大量的數據能做處理。如圖十三所示，機械學習是人工智慧的一個分支，傳統上實現人工智慧的方式需要人們將規則或經驗寫入到系統，機器學習（Machine Learning）則是讓電腦能夠自行從歷史資料中學會一套技能、並能逐步完善精進該項技能。



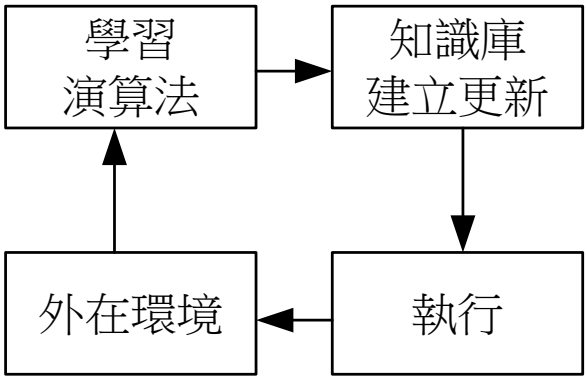
圖十三 人工智慧、機器學習、深度學習關連性

舉例來說，辨識貓咪的技能。人類是如何學會辨識一隻貓的？從短毛貓、摺耳貓、短毛貓、暹羅貓… 等貓咪的外型特徵都不一樣，一般只要父母帶小孩看看貓、或貓咪的圖片，只要看到就告訴孩子這是貓，當小孩把老虎看成貓時進行糾正，久而久之，我們就自然地「學」會辨識一隻貓了。雖然不是原本看過的貓咪，我們仍然知道這是一隻貓。從前讓電腦辨識出貓時，需要工程師將所有貓的特徵用

窮舉法的方式、詳細輸入所有貓的可能條件，比如貓有圓臉、鬍子、肉肉的身體、尖耳朵和一條長尾巴。然而凡事總有例外，若我們在照片中遇到了一隻仰躺只露出肚子的貓？正在奔跑的貓？尖臉短尾貓？也因此誤判的機率很高。

圖十四 各種不同型態的貓

機器學習最基礎的想法是透過使用大量的數據和演算法來「訓練」機器來分析資料、從中學習，以及判斷或預測現實世界裡的某些事或執行任務。 如圖十五為機器學習的主要方塊圖：



圖十五 機器學習方塊圖

機器學習可以分成下面幾種類別，如果需要更深入資訊，請參考學術論文：[10]

A. 監督學習：

從給定的訓練資料集(data set)中學習出一個函式，當新的資料到來時，可以根據這個函式預測結果。監督學習的訓練集要求是包括輸入和輸出，也可以說是特徵和目標。訓練目標是由人標註的。常見的監督學習演算法包括回歸分析和統計分類。

B. 無監督學習：

與監督學習相比，訓練集沒有人為標註。常見的無監督學習演算法有聚類。

C. 半監督學習：介於監督學習與無監督學習之間。

D. 增強學習：

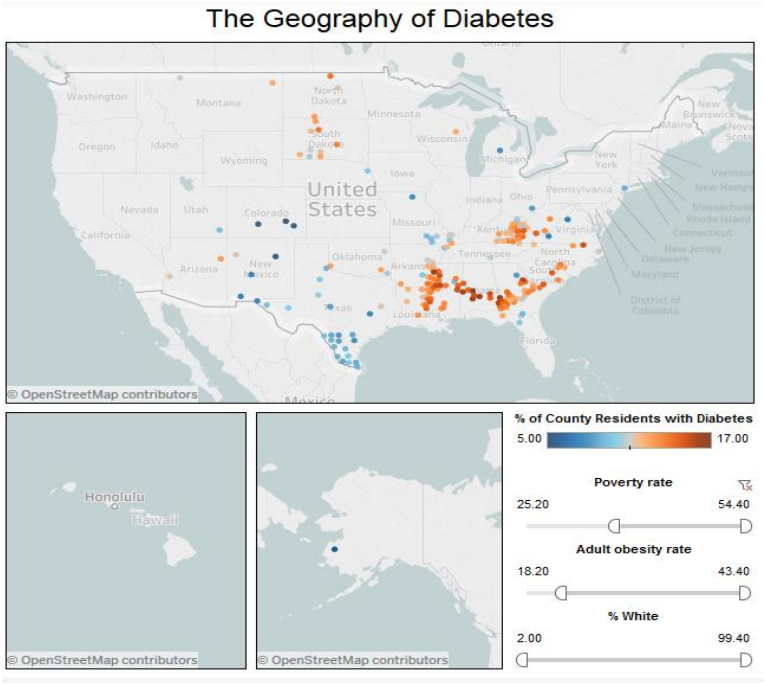
通過觀察來學習做成如何的動作。每個動作都會對環境有所影響，學習物件根據觀察到的周圍環境的反饋來做出判斷。

四、 大數據的溝通：

A. 資料視覺化

隨著「數據導向決策」的時代來臨，資料科學家在分析完數據後，如何成功地將分析結果傳遞出去、使企業接收到該資訊呢？資料視覺化（Data Visualization）的重要性與潛在的龐大商機因此愈發被凸顯出來。人類的大腦在閱讀圖像畫面的速度遠比文字更快。資訊視覺化的優勢在於以一目瞭然的方式呈現資料分析結果，比查閱試算數據或書面報告更有效率。

「Tableau 軟體」和微軟開發的「Power BI」訴求是資料分析後，將自動產生簡潔易懂的資訊圖表，並隨著新增的數據分析結果生成儀錶板（Dashboard），供使用者查詢動態報表、指標管理等服務，圖十六是美國健康局所做的糖尿病分布，利用 Tableau 免費軟體所完成。我們可以調整右下方三個關於肥胖率、窮困率與白人比率開關而觀察資訊的變化。[11]



圖十六 美國糖尿病分布示意圖

B. 新型商業模式的提供:[2] [12]

大數據將全面影響每個人與每家企業。對企業而言，大數據可望提升服務品質、增加管理效率、幫助決策和創造商業模式。大數據的商業模式大概可分成幾種：一、從既有數據變現；二、以數據提升企業競爭力；三、以數據做為服務的基礎與核心，用數據顛覆傳統行業。

模式 1：

數據本身即為產品或根據數據制定行銷策略、改善產品。例如美國運通讓持卡人與自己的 Facebook 帳號連結，持卡人成為美國運通粉絲團粉絲後，美國運通會依據會員在 Facebook 上的活動，提供相應的優惠措施，結合社交數據和會員資料，就是為了提升消費者辦美國運通卡的誘因。

模式 2：

模式二是藉由數據提升競爭力，這類的大數據專案成效較無法直接反映在營收上，而是反映在提升內部工作效率或降低決策成本上。例如許多人都知道 LinkedIn 透過數據精準推薦職場人脈給用戶，卻不知道 LinkedIn 在公司內部推出數百款數據分析產品，幫助內部員工提升工作效率，其中 Voices 就是一款能將 LinkedIn 客服內容，在 1 分鐘內快速生成分析報告的數據分析工具。

無論是模式一還是模式二，其實都有掌握過去、預測未來和防患於未然的共同點，只是一個應用層面是對外，一個對內，這兩種模式常見於既有的企業。

模式 3；

產生以數據做為業務核心的公司，這些公司生來就是要來顛覆傳統行業，它們打從開業的第一天起就把數據當做業務核心，叫車 App Uber 和防詐騙電話 App Whoscall 是最好的例子。

以上為大數據相關的概念與演進，在 Youtube 影音平台上，有許多有關大數據的相關資料，有興趣的同學可以參考[13-15] 。

參考資料:

[1] https://hellolynn.hpd.io/2017/06/09/大數據到底是什麼意思？事實上，它是一種精神！/

[2] https://www.bnext.com.tw/article/35807/bn-2015-03-31-151014-36

[3] 大數據 麥爾荀伯格、庫基耶著 林俊宏譯 天下文化 2013

[4] 大數據挖掘-從巨量資料發現別人看不見的資訊 潭磊著 上奇出版社 2013

[5] https://hellolynn.hpd.io/2017/07/03/不容錯過的人工智慧簡史/

[6] https://en.wikipedia.org/wiki/Data_mining

[7] https://winsys88.wordpress.com/category/big-data-大數據/

[8] http://www.csie.ntnu.edu.tw/~u91029/Classification.html#1

[9] https://hellolynn.hpd.io/2017/07/28/機器是怎麼從資料中「學」到東西的呢/

[10] https://zh.wikipedia.org/wiki//机器学习

[11] https://public.tableau.com/en-us/s/gallery/geography-diabetes

[12] https://www.bnext.com.tw/article/35809/bn-2015-03-31-153046-36

[13] TED 大數據是好數據
https://www.youtube.com/watch?v=DbUuqlPY_Hs

[14] 逢甲通識-翟本喬大數據的理念與應用
https://www.youtube.com/watch?v=OQQO08EX6TM

[15] 數位時代的掏金術-從大數據到人工智慧
https://www.youtube.com/watch?v=X-Q72NiI3SQ

課後選擇測驗

班級：_____ 學號：_____ 姓名：_____

1.大數據的四個流程有哪一個不正確：

A）資料取得、 B）資料儲存、 C）資料探勘、 D）資料刪除

➔_____ （正確答案為 D）

2.何者不是屬於大數據的 3 個 V 的特性之一：

A）資料量（Volumn）、 B）自願地（Voluntarily）、

C）資料傳輸速度（Velocity）、 D）資料類型（Variety）

➔_____ （正確答案為 B）

3.何者是大數據的精神：

A）如何妥善利用既有或非傳統資料，從中挖掘出新商機、
B）不管公司經營面是否可以負擔，逕行設立大數據中心、
C）在中小企業內部組成 50 人以上大數據分析小組，為公司提供建言、
D）把大數據業務外包

➔_____ （正確答案為 A）

4.何者不是資料探勘的功能：

A）群集化（clustering）、 B）分類（Classification）、

C）可用性（usability）、 D）預測（Prediction）

➔_____ （正確答案為 C）

5.大數據的溝通方面，何者描述最佳：

A）利用資料視覺化，傳遞更有效率的溝通結果、 B）無法提供利用大數據資料提升企業競爭力、
C）因為大數據分析結果太抽象，所以企業主管可不採納、
D）可使用「Tableau」和微軟「Power BI」軟體進行資料探勘演算法程式開發、

➔_____ （正確答案為 A）