台科大
since 1997
PF522

最新 人工智慧概論

CHAPTER 07 大數據思維

最新人工智慧概論

- 7-1 什麼是大數據
- 7-2 大數據與人工智慧應用案例簡介

Chapter 07 大數據思維

【影片案例】大數據的故事

自2017 年後，人工智慧與大數據(big data) 觀念、技術與應用在全球掀起了火熱的關注。大數據與過去傳統資料庫相比較，無論在觀念上、技術上都有很大的差異。大數據技術群已經成為人工(智慧) 大腦的重要組成部分。請觀看建議的「大數據」教學影片案例，並請思考，其中包含了哪些AI 技術的應用？

問題思考

看到這些關於報導 AI 與 Big data 大數據發展及人才需求的影片，對大數據的發展給了你什麼樣的啟發？它包含了哪些 AI 應用？它是屬於專用人工智慧還是通用人工智慧？

大數據當道！數據科學家年薪500萬

p159, 3:24 20150306

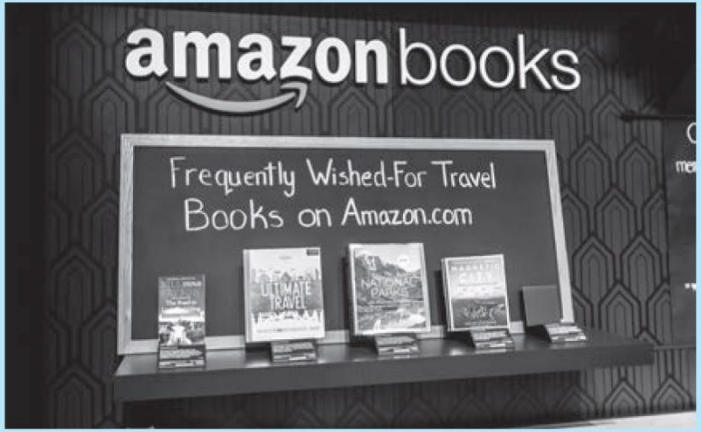
把資料變有用資訊 企業急挖大數據人才

p159, 7:01 20170519

Chapter 07 大數據思維

【影片案例】亞馬遜推薦系統

據說亞馬遜（Amazon）的三分之一銷售額都來自於它的個性化推薦系統(personalization recommendation)。有了它，亞馬遜不僅使眾多大型書店和音樂唱片商店歇業，而且當地數百個自認為有自己風格的書商也難免受轉型之風的影響。



• 書屋內推薦亞馬遜產品的攤位
（圖片來源：123RF 圖庫）

YouTube 推薦影片的運作方式

p160, 1:09 20170830

人工智慧在視頻網站推薦系統中的應用

p160, 17:18 20170818

Netflix 紀錄片揭露大數據操控選舉黑幕

p160, 09:34 20190817

Chapter 07 大數據思維

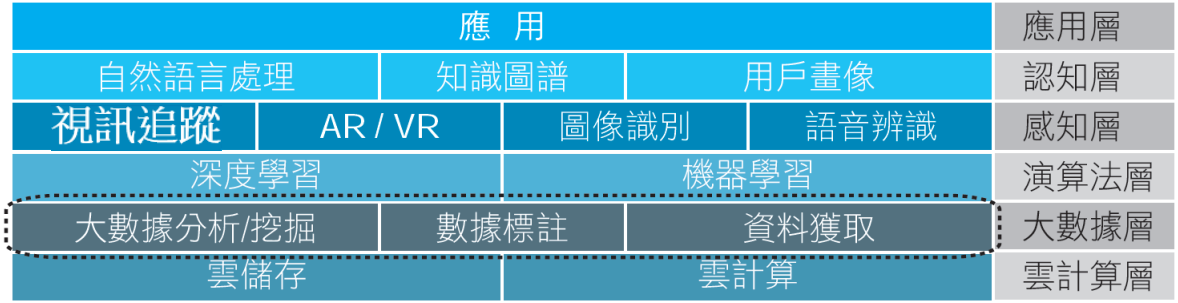
知道「是什麼」可以創造點擊率，這種洞察力足以重塑電子商務，以及其他很多行業。所有行業中的銷售人員早就被告知，他們需要瞭解「是什麼」讓客戶做出了選擇，要把握客戶做決定背後的真正原因，因此專業技能和多年的經驗受到高度重視。大數據(big data) 卻顯示，還有另外一個在「兩件事物間的相對現象或是關聯性的發現等方面」更有用的方法。亞馬遜的推薦系統梳理出了有趣的相關關係，但不必知道背後的原因——知道是什麼(What) 就夠了，沒必要知道為什麼(Why)。

問題思考

看到這個運用大數據做的「個性化推薦系統」的案例，給了你什麼樣的啟發？你會想用類似的大數據分析做些什麼事情？

7-1 什麼是大數據 p. 161

在第一章中曾經列出的人工智慧大腦框架圖，重複呈現在圖 7-1 中，但本章重點是在「大數據層」的大數據分析/ 挖掘、數據標註、資料獲取等，作大數據基本觀念與技術的說明，並配合案例做介紹。



• 圖 7-1 人工智慧的基本框架——以百度大腦為例
（圖片來源：百度 (2017)）

7-1 什麼是大數據

當互聯網(Internet) 開始進一步向外延伸、並與世上的很多物品連結之後，這些物體開始不停地將即時變化的各類數據，透過互聯網回傳到資料庫或資料處理系統，產生物與人、物與物之間的互動，形成了物聯網IoT (the Internet of Things)。物聯網是個大奇蹟，被認為可能是繼互聯網之後人類最偉大的技術革命。物聯網IoT 誕生也造成各行業資料的收集需求更為快速、龐大，間接也催生了大數據的環境與應用需求，如圖7-2。

7-1 什麼是大數據

● 圖 7-2　物聯網 IoT 誕生也造成各行業資料的收集需求更為快速、龐大（圖片來源：123RF 圖庫）

7-1 什麼是大數據

如今，即便是一件物品幾天內被感知到的各種動態資料，都可能足以和古代一位王國一年所收集的各類資料相匹敵。在物聯網上數以萬計、億計的物品，所感知到的資料太龐大了，於是產生大數據(big data) 了。

如此浩如雲海的資料數據，如何分類提取、有效處理呢？這個需要強大的技術設計與運算能力，於是有了「雲計算」；其中的「設計技術」屬於演算法(algorithm)。

7-1 什麼是大數據

「雲計算」需要從大量資料中挖掘有用的資訊，於是產生了資料採擷(data mining) 的數據分析或挖掘技術。這些被挖掘出來的有用資訊去服務城市就叫做「智慧城市」，去服務交通就叫做「智慧交通」，去服務家庭就叫做「智慧家居」，去服務於醫院就叫做「智慧醫院」……於是，「智慧社會」便產生了。不過，智慧社會要有序、有效地運行，中間必須依託一個「橋樑」和借助於某個工具，那就是AI「人工智慧」。

7-1 什麼是大數據

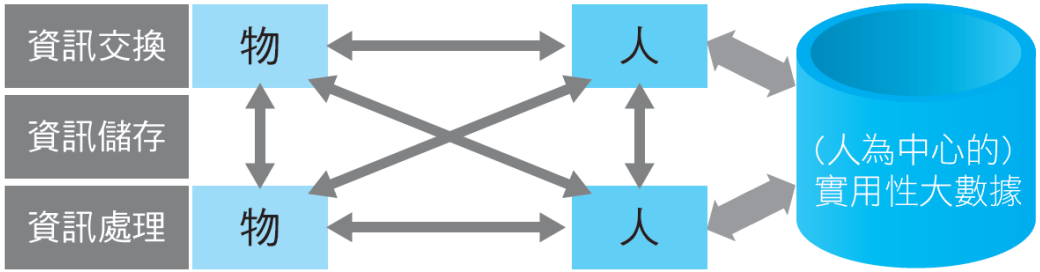
這就是為什麼近幾年時間內，諸如「人工智慧」、「物聯網」、「大數據」、「雲計算」、「量子計算(quantum computing)」、「演算法」、「資料採擷」和「智慧XX」這些新興科技或概念突然紛紛冒出來的理由，原來它們都是具有相關性的技術或應用！

• 圖 7-3　人工智慧與物聯網、大數據、雲計算等等都有發展技術或應用相關性

7-1 什麼是大數據

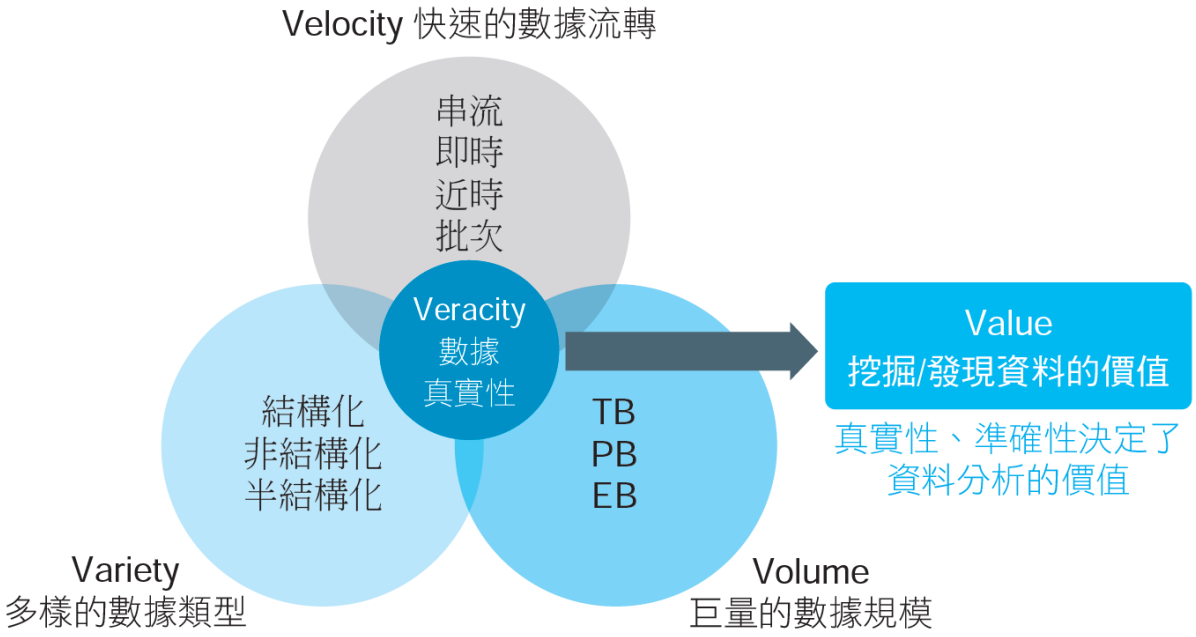
萬物的大數據主要包括人與人、人與物、物與物三者相互作用所產生（製造）的大數據。對於人與人、人與物之間被感知到的那部分的資料，雖然相對於萬物釋放的量來說還是非常小，但是絕對量卻非常大。在2000 年後，因為人類資訊交換、資訊儲存、資訊處理三方面能力的大幅增長而產生巨量的資料，如圖7-4。這個就是我們日常所聽到的「大數據」概念，是以人為中心的狹義大數據，又稱為實用性（商業、監控或發展等使用）大數據。資訊儲存、處理等能力的增強，讓我們利用大數據時提供了近乎無限的想像空間。

7-1 什麼是大數據



- 圖 7-4　以人為中心的狹義大數據——實用性（商業、監控或發展等使用）大數據

7-1 什麼是大數據



• 圖 7-5　大數據的 5V (=3V+2V) 特性

7-1 什麼是大數據

在數位化時代，資料處理變得更加容易、更加快速，人們能夠在瞬間處理成千上萬的資料。而「大數據」全在於發現和理解資訊內容及資訊與資訊之間的關係。如圖7-5，大數據的特點可用5V（3V 是指下列的前面三項特性）來代表：

1 Volume （容積）：巨量的資料規模，目前是以TB（terabyte = 1024GB（gigabyte ）為基本單位，資料量很龐大。

2 Velocity （速度）：快速的資料流轉，即時性、隨時隨地的變化。包括串流(stream)、即時(real time)、近時(near time)、批次(batch)的變化與資料流轉。

7-1 什麼是大數據

3 Variety （種類）：多樣的資料類型，資料格式繁多。包括：
結構化(structured)——例如學生成績資料庫資料
非結構化(unstructured)——例如會議的發言、留言等
半結構化(semi- structured)——例如前面兩者的混合格式，如一般公文等

4 Veracity （真實性）：資料再多、變化再快若不具有真實性，則這些資料將無法採用，是沒有任何價值的。

5 Value （價值）：發現資訊內容、理解資訊與資訊之間的關係，才能發現資料的價值。

7-1 什麼是大數據

延伸學習

大數據未來很快地就會以PB（petabyte =1,024 TB = 百萬GB）或EB（exabyte =1,024 PB = 百萬TB）更驚人的資料量為處理單位。其關係可以整理如下，或參考表7-1

1 EB = 1,024 PB = 1,024 X 1,024 TB = 1,024 X 1,024 X 1,024 GB

= 1,024 X 1,024 X 1,024 X 1,024 MB = 1,024 X 1,024 X 1,024 X 1,024 X 1,024 KB

7-1 什麼是大數據

◆ 表 7-1　儲存單位的關係

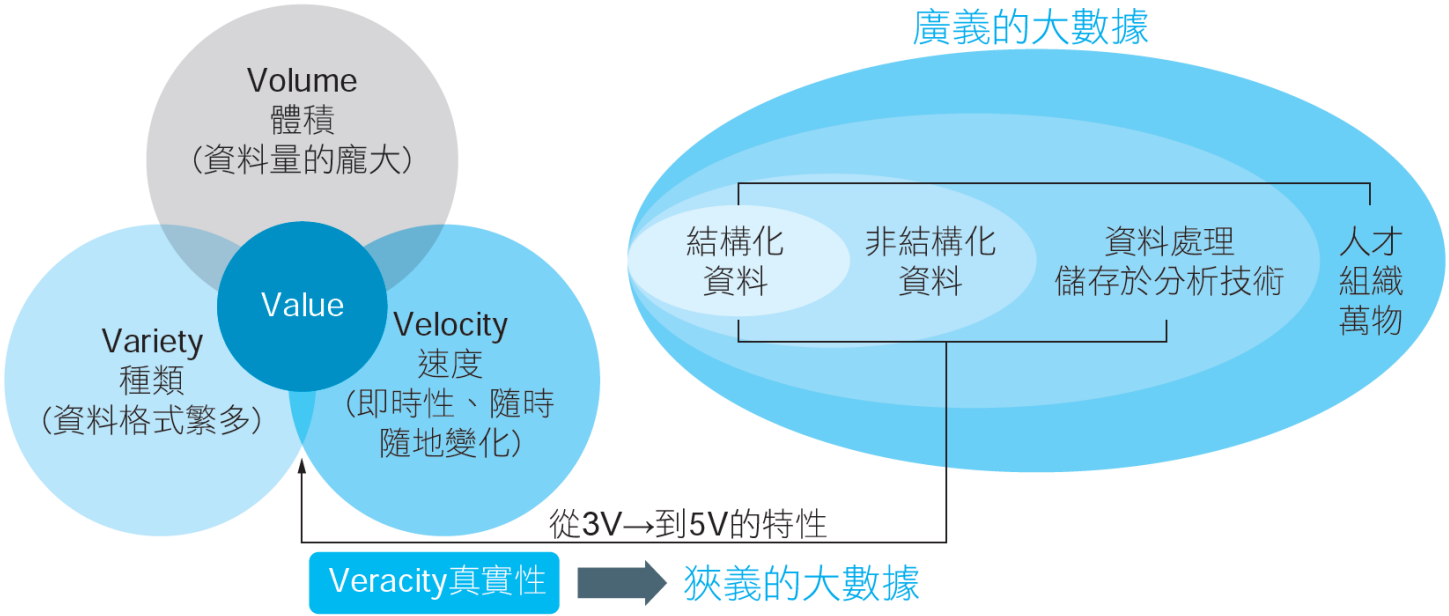
儲存的單位與簡稱	說明	中文稱呼
1 byte (B)	8 bit	位元組
1 kilobyte (K/KB)	2^{10} byte= 1024 byte (B)	千位元組
1 megabyte (M/MB)	2^{20} byte= 1024 KB	百萬位元組
1 gigabyte (G/GB)	2^{30} byte= 1024 MB	十億位元組；十億位元組
1 terabyte (T/TB)	2^{40} byte= 1024 GB	萬億位元組；太位元組
1 petabyte (P/PB)	2^{50} byte= 1024 TB	
1 exabyte (E/EB)	2^{60} byte= 1024 PB	
1 zettabyte (Z/ZB)	2^{70} byte= 1024 EB	
1 yottabyte (Y/YB)	2^{80} byte= 1024 ZB	

7-1 什麼是大數據

狹義大數據與廣義大數據

狹義大數據是指多樣的資料類型（結構化、非結構化等），結合資料處理、分析技術等；廣義的大數據除了狹義大數據的內涵外，再把人才、組織、萬物等都考慮包含進來，請參考圖7-6。

7-1 什麼是大數據



- 圖 7-6　狹義大數據與廣義大數據

7-1 什麼是大數據

延伸學習

關於大數據常用的技術參考框架，請參考圖7-1。

一般企業針對大數據的處理技術框架主要可分成六個階段，包括：

(1) 資料獲取 (data collection)：資料獲取包括歷史資料的採集和當前市場資料的採集，是進行科學化數據分析的基礎。資料獲取準確性決定了數據分析的價值。

(2) 資料儲存(data storage)：是指儲存框架的方式，有非常多的類型。常見的有使用傳統的關係型資料庫，如：Oracle、MySQL；另外，新興方式的技術，如NoSQL：HBase、Cassandra、Redis；全文檢索框架技術，例如：ES (Elastic Search)、Solr 等。

7-1 什麼是大數據

(3) 資料管理(data management)：是指利用電腦硬體和軟體技術對資料進行有效的收集、儲存、處理和應用的過程。資料管理目的在於充分有效地發揮資料的效用。實現資料有效管理的關鍵是資料組織。

(4) 資料分析(data analysis)：是指用適當的統計分析方法對收集來的大量資料進行分析，提煉出有用的資訊、形成結論，並對資料做詳細研究和概括總結的處理過程。資料分析可幫助我們作判斷、決策，以便採取適當行動。

(5) 演算法(algorithm)：對解題方案作準確而完整的描述，是一系列解決問題的明確指令，演算法代表用系統的方法描述解決問題的策略與機制。

7-1 什麼是大數據

(6) 資料視覺化(data visualization)：資料視覺化是有關於將資料視覺表現形式的科學技術研究，也是商務推動上的基本呈現技術與要求的能力。

在表7-2 列出了小米採用的大數據技術框架，裡面的工具或技術名稱，會因為時間或新產品的問世，而被替換掉。在此僅做為參考，在需要瞭解技術框架時，可作為查詢或對照之用。

7-1 什麼是大數據

◆ 表 7-2　大數據處理階段與參考技術框架—以小米採用的工具或技術為例

處理階段	使用工具或技術示例					
6. 數據視覺化	E-Chart	JavaScript	H5/App			
5. 演算法	機器學習	自然語言	資料採擷	統計分析		
4. 數據分析	MapReduce	Spark	Storm	Hive	Impala	Druid
3. 資料管理	Hue	Kerbros				Zookeeper
2. 資料儲存	HDFS	Hbase	Kudu	Kafka		
1. 資料獲取	ETL	scribe				

（資料來源：小米 (2019)）

7-1 什麼是大數據

問題思考

多樣的資料類型、資料格式繁多是大數據的 Variety（**種類**）特性，請根據下面表格的結構化、非結構化、半結構化或其他類型，找一下你在學校或是外面活動時，你發現的實際使用各種資料類型的案例。

大數據多樣的資料類型	本書舉的例子	你發現的實際使用案例
結構化	學生成績資料庫資料	
非結構化	會議的發言	
半結構化	一般公文	
其他		

7-2 大數據與人工智慧應用案例簡介 P.168

長久以來，因為紀錄、儲存和分析資料的工具不夠好或儲存空間、網路速度等的限制，當面臨大量數據時，為了讓分析工作變得簡單，通常都依賴於採樣分析(sampling analysis)。但是採樣分析是資訊缺乏時代，和資訊流通受限制的類比資料時代的產物。現今資訊技術的條件已經有了轉型且提高，可以處理的數據量與速度已經大幅度地增加。

在大數據時代資料處理技術已經發生了翻天覆地的改變，我們的思維和方法必須跟上這種改變。其中，對大數據有三項重要的思維轉變：

樣本 = 總體、接受資料的混雜性與不精確性、找出資料間的相關關係

7-2 大數據與人工智慧應用案例簡介

(1) 大數據時代第一個轉變，是要分析與某事物相關的所有資料，而不是依靠分析少量的資料樣本。也就是說，大數據分析是全數據模式：樣本=總體。

(2) 大數據時代的第二個轉變，是我們樂於接受資料的紛繁複雜，而不再一味追求其精確性。

(3) 第三個轉變我們嘗試著不再探求難以捉摸的因果關係，轉而關注事物的相關關係。

有部分人們依然沒有意識到，自己擁有能夠收集和處理更大規模資料的能力，還是在資訊匱乏的假設下做很多工作。人們甚至發展了一些使用盡可能少的資訊技術，例如，統計學(statistics)的目的之一就是，用盡可能少的資料來證實可能的重大發現。

7-2 大數據與人工智慧應用案例簡介

一、小數據時代的隨機採樣案例簡介

人口普查(census)是一項耗資且費時的事情，當時收集的資訊也只是一個大概情況，實施人口普查的人也知道他們不可能準確紀錄下每個人的資訊。實際上，「人口普查」這個詞來源於拉丁語的「censere」，本意就是指推測、估算。

7-2 大數據與人工智慧應用案例簡介

示例7-1 使用新技術來克服數據收集與整理問題

美國在1880 年進行的人口普查，耗時8 年才完成資料匯總工作。實際上，他們獲得的很多資料都是過時的。1890 年進行的人口普查，預計要花費13 年的時間做資料匯總。然而，因為稅收分攤和國會代表人數確定都是建立在人口的基礎上的，必須獲得正確且及時的資料，所以就需要使用新技術來克服此問題。

後來美國發明家赫爾曼・霍爾瑞斯（Herman Hollerith）後人稱為現代自動計算之父，他發明的穿孔卡片製表機成功地在1 年時間內完成了人口普查的資料匯總工作。這簡直就是一個奇蹟，它代表著自動處理資料的開端，也為後來IBM 公司的成立奠定了基礎。

7-2 大數據與人工智慧應用案例簡介

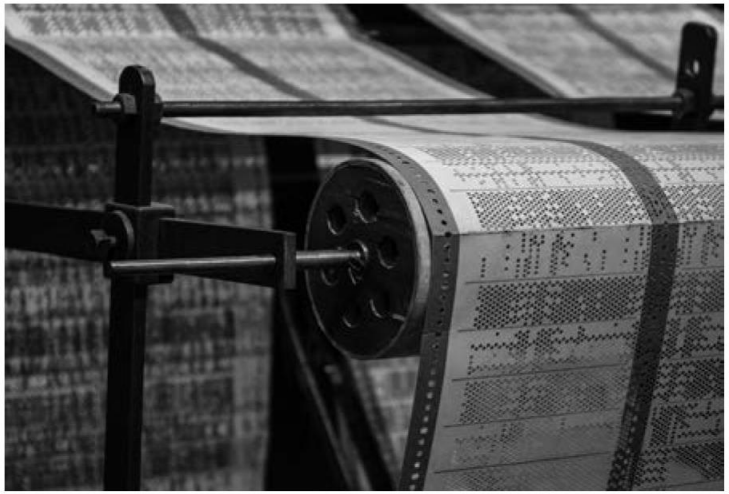


圖 7-7　霍爾瑞斯普查機的穿孔卡片製表機制（圖片來源：123RF 圖庫）

7-2 大數據與人工智慧應用案例簡介

在進行資料調查主要考慮的問題，是到底要利用所有的資料？還是僅僅採用一部分呢？有人提出有目的地選擇最具代表性的樣本是最恰當的方法，後來統計學家們證明，問題的關鍵是選擇樣本時的隨機性。

> 採樣分析的精確性隨著採樣隨機性的增加而大幅提高，但與樣本數量的增加關係不大。

在商業領域，隨機採樣被用來監管商品品質，使得監管商品品質和提升商品品質變得更容易，花費也更少。本質上來說，隨機採樣讓大數據問題變得更加切實可行。同理，它將客戶調查引進了零售行業，將焦點討論引進了政治界，也將許多人文問題變成了社會科學問題。

7-2 大數據與人工智慧應用案例簡介

隨機採樣取得了巨大的成功，成為現代社會、現代測量領域的主軸。但這只是一條捷徑，是在不可收集和分析全部資料的情況下的選擇，它的成功依賴於採樣的絕對隨機性，但是實現採樣的隨機性非常困難。一旦採樣過程中存在任何偏見，分析結果就會相去甚遠。

7-2 大數據與人工智慧應用案例簡介

二、 大數據分析思維的第一個轉變——全數據模式（樣本=總體）應用案例簡介

採樣(sampling) 的目的是用最少的資料得到更多的資訊，而當我們可以處理海量資料的時候，採樣就沒有什麼意義了。現今的計算和製表已經不再困難，感應器、手機導航、網站點擊和微信等被動地收集了大量資料，而電腦可以輕易地對這些資料進行處理。

假如用採樣的方法分析結果的正確率可達97%。對於某些事物來說，3%的錯誤率是可以接受的，但是我們可能會失去對某些特定子類別進行進一步研究的機會。

7-2 大數據與人工智慧應用案例簡介

透過使用與分析所有的資料，我們可以發現，如果不對所有資料檢視，則少數重要的異常資料會在大量資料中被淹沒掉。例如，信用卡詐騙是透過觀察異常情況來識別的，只有掌握了所有的資料才能做到這一點。異常值是最有用的資訊，你可以把它與正常交易情況進行對比。因為交易是即時的，所以你的數據分析也應該是即時的。

因為大數據是建立在掌握所有資料，至少是盡可能多資料的基礎上，所以我們就可以正確地考察細節並進行新的分析。在任何細微的層面，我們可以用大數據去論證新的假設。

7-2 大數據與人工智慧應用案例簡介

示例7-2 使用全數據模式（ 樣本 = 總體 ）做流感趨勢分析與預測

谷歌流感趨勢預測不是依賴於隨機樣本，而是分析了全美國幾十億條互聯網檢索紀錄。透過分析整個資料庫，而不是僅對一個小樣本進行分析，這樣才能提高微觀層面分析的準確性，甚至能夠推測出某個特定城市的流感狀況。

問題思考

可否分別舉出使用樣本分析法做資料調查分析與使用大數據調查分析的例子？

7-2 大數據與人工智慧應用案例簡介

三、大數據分析思維的第二個轉變——允許不精確案例簡介

對「小數據」而言，最基本、最重要的要求就是減少錯誤，保證品質。因為收集的資訊量比較少，所以我們必須確保紀錄下來的資料儘量精確。在採樣的時候，對精確度的要求更苛刻了。因為收集資訊的有限，意味著細微的錯誤會被放大，甚至有可能影響整個結果的準確性。

在不斷湧現的新情況裡，允許不精確的出現已經成為大數據的一個亮點，而非缺點。因為放鬆了容錯的標準，人們掌握的資料也多了起來，還可以利用這些資料做更多新的事情。

7-2 大數據與人工智慧應用案例簡介

大數據時代的第二個轉變，是我們樂於接受資料的紛繁複雜，而不再一味追求其精確性。

在大數據時代我們需要與各種各樣的混亂資料做鬥爭。混亂(chaos)，簡單地說就是隨著資料的增加，錯誤率也會相應增加。例如，如果橋樑的壓力資料量增加1000 倍的話，其中的部分讀數就可能是錯誤的，而且隨著讀數量的增加，錯誤率可能也會繼續增加。在整合來源不同的各類資訊的時候，因為它們通常不完全一致，所以也會加大混亂程度。

7-2 大數據與人工智慧應用案例簡介

混亂也包含了格式的不一致性，因為要達到格式一致，就需要在進行資料處理(data processing) 之前仔細地清洗資料(data cleansing)，而這在大數據背景下很難做到。當然，在萃取或處理資料的時候，混亂也會發生。因為在進行資料轉化的時候，我們是在把它變成另外的事物。

在很多情況下，與致力於避免錯誤的耗費時間、成本相比較，對錯誤的包容會反而帶給我們更多好處。

7-2 大數據與人工智慧應用案例簡介

示例7-3 用大數據做訓練可提升演算法的效能

大數據在多大程度上優於演算法，這個問題在自然語言處理上表現得很明顯。

2000 年，微軟研究中心的蜜雪兒·班科（Michelle Banco）和埃裡克·布里爾（Eric Brill）一直在尋求改進Word 程式中語法檢查(在「校閱」清單下的「拼字及文法檢查」功能）的方法。他們不能確定是努力改進現有的演算法？或研發新的方法？還是添加一些更細膩精緻的特徵點更有效。在實施這些措施之前，他們決定從現有的演算法中添加更多的資料，看看會有什麼不同的變化。很多對電腦學習演算法的研究都建立在百萬字左右的語料庫（corpus）基礎上。最後，他們決定對4 種常見的演算法中逐步增添資料，先是一千萬字、再到一億字、最後到十億。

7-2 大數據與人工智慧應用案例簡介

結果有點令人吃驚。他們發現，隨著資料的增多，4 種演算法的表現都大幅提高了。

當資料只有500 萬的時候，有一種簡單的演算法表現得很差；當資料達10 億的時候，它變成了表現最好的，準確率從原來的75%提高到了95%以上。

相反地，在少量資料情況下運行得最好的演算法，當加入更多的資料時，也會像其他的演算法一樣準確率有所提高，但是卻變成了在大量資料條件下運行得最不好的。它的準確率會從86% 提高到94%，請參閱表7-3。

7-2 大數據與人工智慧應用案例簡介

◆ 表 7-3　隨著資料的增多 4 種演算法的準確率表現都大幅提高

演算法 / 準確率	語料庫大小			
	500 萬	1000 萬	1 億	10 億
演算法 1	75%	80.5%	89%	95%
演算法 2	78%	82%	89.4%	94.5%
演算法 3	86%	88%	92%	94%
演算法 4	81%	86%	91.2%	94.8%

後來，班科和布里爾在他們發表的研究論文中寫到：「事實證明，我們得重新衡量一下，要將更多的人力物力應該用在演算法發展上，還是用在語料庫發展上。」

7-2 大數據與人工智慧應用案例簡介

四、 大數據時代思維第二個轉變——紛繁的資料越多越好的案例簡介

通常傳統的統計學家都很難容忍錯誤資料的存在，在收集樣本的時候，他們會用一整套的策略來減少錯誤發生的概率。但是，即使只是少量的資料，這些規避錯誤的策略實施起來還是耗費巨大的時間與費用。尤其是當我們收集所有資料的時候，還因為在大規模的基礎上保持資料收集標準的一致性，變得吃力而不討好。

生活在資訊時代，掌握的資料庫越來越全面，它包括了與這些現象相關的大量甚至全部資料。我們不再需要那麼擔心某個資料點對整套分析的不利影響。我們要做的就是接受這些紛繁的資料並從中受益，而不是以高昂的代價消除所有的不確定性。

7-2 大數據與人工智慧應用案例簡介

示例7-4 適當接受數據的不精確和不完美，大數據會帶來意想不到的利益

美國華盛頓州布萊恩市的英國石油公司（BP）切裡波因特煉油廠裡，無線感應器遍佈於整個工廠，形成無形的網路，能夠產生大量即時資料，參考圖7-8。

在這裡，酷熱的惡劣環境和電氣設備的存在，有時會對感應器讀數有所影響，形成錯誤的資料。但是資料生成的數量之多可以彌補這些小錯誤。隨時監測管道的承受壓力，使得BP 公司能瞭解到，有些種類的原油比其他種類更具有腐蝕性。以前，這些都是無法發現、也無法防止的。

7-2 大數據與人工智慧應用案例簡介

接受大數據的不精確和不完美，除了在一開始會與我們的直覺相矛盾之外，反而讓我們能夠更好地進行預測，也能夠更好地理解這個世界。

擁有更大數據量所能帶來的商業利益遠遠超過增加一點精確性，所以通常我們不會再花大力氣去提升資料的精確性。這又是一個關注焦點的轉變，正如以前，統計學家們總是把他們的興趣放在提高樣本的隨機性而不是數量上。大數據給我們帶來的利益，讓我們能夠接受不精確的存在了。

• 圖 7-8 切裡波因特煉油廠——無線感應器遍佈於整個工廠，產生大量即時資料

（資料來源：https://www.mi.com/watch?v=FHVnSF-Ogls）

（圖片來源：123RF 圖庫）

7-2 大數據與人工智慧應用案例簡介

五、 大數據時代應用混雜性是標準途徑的案例簡介

長期以來，人們一直用分類法和索引法來儲存和檢索資料，在「小數據」範圍內這些方法就很有效。但是分類法和索引法的分級系統在大數據時通常都不完善，一旦資料規模增加好幾個數量等級後，原來一切預設好、且都已各就各位的分類系統就會崩潰。

7-2 大數據與人工智慧應用案例簡介

示例 7-5 使用者自創照片標籤，適當的混亂卻更靈活實用容易查到

一家加拿大的相片分享網站 Flickr 在2011 年就已經擁有來自大概1 億用戶的60億張照片，如圖7-9。

● 圖 7-9 相片分享網站 Flickr 在 2011 年度最受歡迎的照片之一

（圖片來源：Flickr, https://www.flickr.com
Google, https://ai.googleblog.com/2017/11/feature-visualization.html）

7-2 大數據與人工智慧應用案例簡介

根據預先設定好的分類來標註每張照片就沒有意義了。恰恰相反，清楚的分類被更混亂卻更靈活的機制所取代了，這些機制才能適應改變著的世界。

當使用者自行上傳照片到Flickr 網站的時候，自己會給照片添加標籤(tag)，也就是使用一組文本標籤來編組和搜索這些資源。

人們用自己的方式創造和使用標籤，所以它是沒有標準、沒有預先設定的排列和分類，也沒有我們所必須遵守的類別規定。任何人都可以輸入新的標籤，標籤內容事實上就成為了網路資源的分類標準。

7-2 大數據與人工智慧應用案例簡介

標籤被廣泛地應用於微信、臉書、博客等社交網路上。因為它們的存在，互聯網上的資源變得更加容易找到，特別是像圖片、影片和音樂這些無法用關鍵字搜索的非文本類資源。要想獲得大規模數據帶來的好處，混亂應該是一種標準途徑，而不應該是竭力避免的。

7-2 大數據與人工智慧應用案例簡介

六、接受混亂——5% 的結構化資料與95% 的非結構化資料

據估計，只有5%的數位資料是結構化的，且能適用于傳統資料庫。如果不接受混亂，剩下95%的非結構化資料都無法被利用，比如網頁和影片資源。透過接受不精確性，我們打開了一個從未涉足的新天地。

我們怎麼看待使用所有資料和使用部分資料的差別？

我們怎樣選擇放鬆要求，並取代嚴格的精確性？

將會對我們與世界的溝通產生深刻的影響。隨著大數據技術成為日常生活中的一部分，我們應該開始從一個比以前更大更全面的角度來理解事物，也就是說應該將「樣本= 總體」植入我們的思維中。

7-2 大數據與人工智慧應用案例簡介

只要我們能夠得到一個事物更完整的概念，
我們就能接受**模糊 (fuzzy**[15]**)** 和**不確定 (uncertain**[16]**)** 的存在。

相比依賴於小數據和精確性的時代，大數據更強調資料的完整性和混雜性，幫助我們進一步接近事實的真相。

當我們的視野侷限在我們可以分析和能夠確定的資料上時，我們對世界的整體理解就可能產生偏差和錯誤。侷限於狹隘的小數據中，我們可以自豪於對精確性的追求，但是就算我們可以分析得到細節中的細節，依然會錯過事物的全貌，也失去了從各個不同角度來觀察事物的權利。

7-2 大數據與人工智慧應用案例簡介

問題思考

相對於傳統小數據的思維，大數據強調資料的完整性和混雜性，請比較小數據與大數據兩者的優缺點？

7-2 大數據與人工智慧應用案例簡介

七、大數據思維第三個轉變——找出關聯物才是預測的關鍵

尋找因果關係是人類長久以來的習慣，即使確定因果關係很困難而且用途不大，人類還是習慣性地尋找緣由。在大數據時代，我們無須再緊盯事物之間的因果關係(causal relationship)，而應該尋找事物之間的相關關係(correlationship)，這會給我們提供非常新穎且有價值的觀點。相關關係也許不能準確地告知我們某件事情為何會發生，但是它會提醒我們這件事情正在發生。在許多情況下，這種提醒的幫助已經足夠大了。

7-2 大數據與人工智慧應用案例簡介

示例 7-6 運用大數據知道「是什麼」而不一定要知道「為什麼」

如果數百萬條電子醫療紀錄都顯示橙汁和阿斯匹林的特定組合可以治療癌症，那麼找出具體的藥理機制就沒有這種治療方法本身來得重要。同樣的，只要我們知道什麼時候是買機票的最佳時機，不需知道機票價格瘋狂變動的原因。

大數據告訴我們「是什麼」而不是「為什麼」。在大數據時代，我們不必知道現象背後的原因，只要讓資料自己發聲。我們不再需要在還沒有收集資料之前，就把分析建立在早已設立的少量假設的基礎之上。讓資料發聲(數字會說話)，我們會注意到很多以前從來沒有意識到的聯繫存在。

7-2 大數據與人工智慧應用案例簡介

在傳統觀念下，人們總是習慣致力於找到一切事情發生背後的原因。然而在很多時候，尋找資料間的關聯、並利用這種關聯就足夠了。這些思想上的重大轉變導致了第三個變革。

我們嘗試著不再探求難以捉摸的因果關係，轉而關注事物的相關關係。

雖然在小數據世界中相關關係也是有用的，但在大數據的背景下，相關關係大放異彩。透過應用相關關係，我們可以比以前更容易、更快捷、更清楚地分析事物。

所謂相關關係，其核心是指量化兩個資料值之間的數理關係。相關關係強是指當一個資料值增加時，另一個資料值很有可能也會隨之增加。我們已經看到過這種很強的相關關係成功的應用案例。例如，買嬰兒尿布的男性，通常會順便買啤酒等關係的實務應用。

7-2 大數據與人工智慧應用案例簡介

示例7-7 谷歌流感趨勢預測：應用大數據找出關聯物。

在一個特定的地理位置，越多的人透過谷歌搜索特定的流感相關詞條，表示該地區就有更多的人患了流感。

在大數據時代，擁有如此多的資料，這麼好的機器計算能力，不再需要人工選擇一個關聯物或者一小部分相似資料來逐一分析了。複雜的機器分析有助於我們做出準確的判斷，就像在谷歌流感趨勢中，電腦把5 億個檢索詞條在數學模型上進行測試之後，準確地找出了哪些是與流感傳播最相關的詞條。我們理解世界不再需要建立在假設的基礎上，這些假設是指針對現象建立的有關其產生機制和內在機理的假設。

7-2 大數據與人工智慧應用案例簡介

相反的，如果相關關係弱就意味著當一個資料值增加時，另一個資料值幾乎不會發生變化。例如，我們可以尋找關於個人的鞋碼和幸福的相關關係，但會發現它們幾乎扯不上什麼關係。

相關關係透過識別有用的關聯物來幫助我們分析一個現象，而不是透過揭示其內部的運作機制。即使是很強的相關關係也不一定能解釋每一種情況，比如兩個事物看上去行為相似，但很有可能只是巧合。相關關係沒有絕對，只有可能性。

7-2 大數據與人工智慧應用案例簡介

• 圖 7-10　運用大數據找出人與人、人與物、物與物的相關關係
（圖片來源：123RF 圖庫）

7-2 大數據與人工智慧應用案例簡介

示例7-8 相關關係分析的預測或推廣是大數據的重要應用

並不是亞馬遜推薦(recommend) 的每本書都是顧客想買的書，如果相關關係強，一個相關連結成功的概率是很高的，亞馬遜的書架上有很多書都是因推薦而購買的。

透過找到一個現象的良好關聯物，相關關係可以幫助我們捕捉現在和預測未來。如果A 和B 經常一起發生，我們只需要注意到B 發生了就可以預測A 也發生了。這有助於我們捕捉可能和A 一起發生的事情，即使我們不能直接測量或觀察到A，卻可以幫助我們預測未來可能發生什麼。相關關係是無法預知未來的，它只能預測可能發生的事情，但已經極其珍貴了。

7-2 大數據與人工智慧應用案例簡介

除了僅僅依靠相關關係，專家們還會使用一些建立在理論基礎上的假想，來指導自己選擇適當的關聯物。這些理論就是關於事物是怎樣運作的一些抽象觀點，然後收集與關聯物相關的資料，來進行相關關係分析，以證明這個關聯物是否真的合適。如果不合適，人們通常會固執地再次嘗試，因為擔心可能是資料收集的錯誤，而最終卻不得不承認一開始的假想、甚至假想建立的基礎，都是有缺陷和必須修改的。這種對假想的反復試驗促進了學科的發展，但這種發展非常緩慢，因為個人以及團體的偏見會蒙蔽我們的雙眼，導致我們在設立假想、應用假想和選擇關聯物的過程中犯錯誤。

建立在相關關係分析法(correlation analysis) 基礎上的預測，是大數據的核心應用。這種預測發生的頻率非常高，經常會有它的創新性，而且應用會越來越多。

7-2 大數據與人工智慧應用案例簡介

示例7-9 用大數據找關聯物並監控它做好故障預測

一個東西或設備要出故障，不會是瞬間的，而是慢慢地出問題的。找出一個關聯物並監控它，我們就能預測未來。透過收集所有的資料，我們可以預先捕捉到設備要出故障的信號。

比方說發動機的嗡嗡聲、引擎過熱都說明它們可能要出故障了。系統把這些異常情況與正常情況進行對比，就會知道什麼地方出了毛病。透過儘早地發現異常，系統可以提醒我們在故障之前，更換零件或者修復問題。

7-2 大數據與人工智慧應用案例簡介

八、 大數據思維第三個轉變——「是什麼」，而不是「為什麼」的案例簡介

在小數據時代，相關關係分析和因果分析不容易運用且耗費巨大，都要從建立假設(assumption) 開始，然後進行實驗——這個假設要麼被證實或被推翻。由於小數據的相關關係分析和因果分析兩者都始於假設，這些分析就都有受偏見影響的可能，極易導致錯誤。此外，用來做相關關係分析的資料很難得到。

另一方面，在小數據時代，由於電腦能力的不足，大部分相關關係分析僅限於尋求線性關係(linear relationship)。而事實上，實際情況遠比我們所想像的要複雜。經過複雜的分析，我們會經常發現有些資料具有非線性關係 (nonlinear relationship)

7-2 大數據與人工智慧應用案例簡介

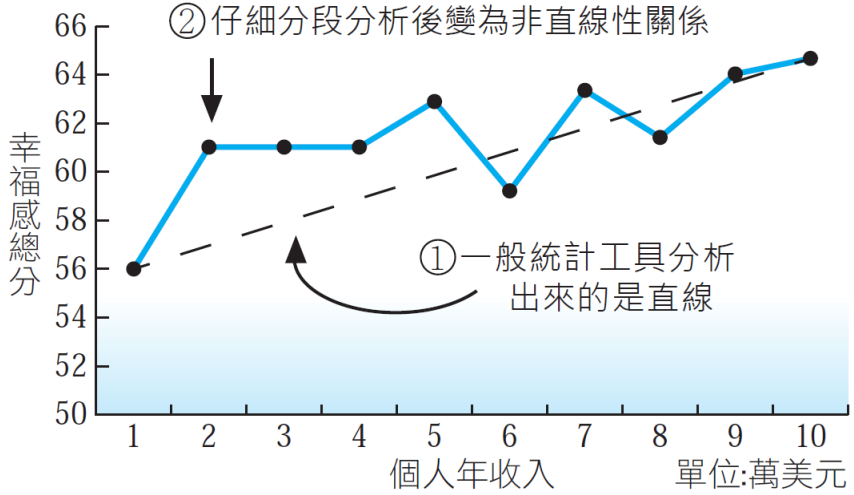
示例7-10 低收入人群的收入水準提高和幸福感是成正比的

多年來，經濟學家和政治家一直認為收入水準和幸福感是成正比的。從資料圖表上可以看到，雖然統計工具呈現的是一種線性關係，但事實上，它們之間卻存在一種更複雜的動態關係。例如，對於年收入水準在2 萬美元以下的人來說，一旦收入增加，幸福感會隨之提升；但對於年收入水準在2 萬美元以上的人來說，幸福感並不會隨著收入水準提高而提升。如果能發現這層關係，我們看到的就應該是一條曲線(curve)，而不是統計工具分析出來的直線。這個發現對決策者來說非常重要。

7-2 大數據與人工智慧應用案例簡介

如果只看到主軸的線性關係的話，那麼政策重心應完全放在增加收入上，因為這樣才能增加全民的幸福感。而一旦察覺到這種非線性關係，策略的重心就會變成提高低收入人群的收入水準，因為這樣的效果明顯、更具針對性。當相關關係變得更複雜時，感覺一切跟過去不同思考、較混亂，但是卻更貼切地、務實地做出正確的決策了。

7-2 大數據與人工智慧應用案例簡介



• 圖 7-11　幸福感與收入之間的關係圖

(資料重繪參考：邢占軍 (2011), 我國城市居民收入如何影響幸福感？
http://www.chinareform.org.cn/society/income/forward/201103/t20110316_63016_1.htm

7-2 大數據與人工智慧應用案例簡介

大數據時代，專家們正在研發能發現並對比分析非線性關係的技術工具。一系列飛速發展的新技術和新軟體也從多方面提高了相關關係分析工具的數量與功能，進而幫我們較容易發現非因果關係(non-causal relationship) 的能力。這些新的分析工具和思路為我們展現了一系列新的視野、並找出有用的預測，我們看到了很多以前不曾注意到的聯繫關係，還掌握了以前無法理解的複雜技術和社會動態。最重要的是，透過去探求「是什麼」而不是「為什麼」，相關關係幫助我們更好地瞭解了這個世界。

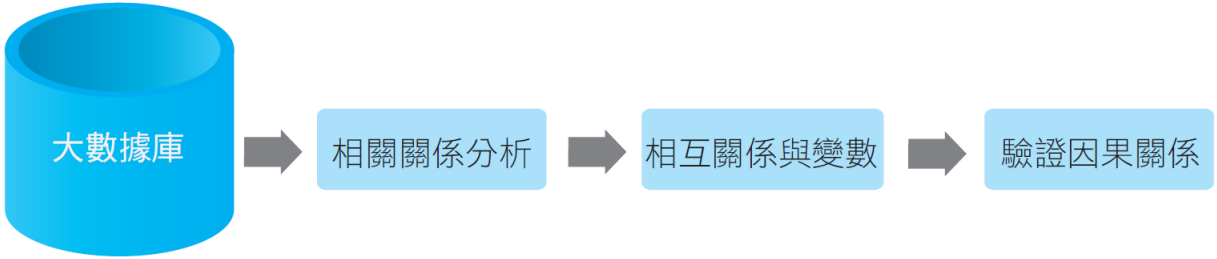
7-2 大數據與人工智慧應用案例簡介

九、 大數據思維第三個轉變——透過相關關係瞭解世界的案例簡介

運用大數據來證明相關關係的實驗耗資，比因果關係相對少很多、費時也少。分析相關關係，既有數學方法(mathematics)，也有統計學方法；同時，許多軟體工具已能幫我們準確地找出相關關係。

相關關係分析本身也可為研究因果關係奠定了基礎。透過找出可能相關的事物，有時候反而可以在此基礎上，進行進一步的因果關係分析。如果存在因果關係的話，我們再進一步找出原因。這種便捷的資料處理機制，透過實驗證實可降低了因果分析的成本。我們也可以從相互聯繫中找到一些重要的變數(variance)這些變數可以用到驗證因果關係的實驗中去。

7-2 大數據與人工智慧應用案例簡介



• 圖 7-12　大數據透過相互關係分析可降低因果分析成本

快速便捷的大數據處理機制，透過「相互關係」與「變數」實驗證實降低了因果分析成本。

相關關係很有用，它能為我們提供新的視角，而且提供的視角都很清晰。如果我們一開始就把因果關係考慮進來，這些視角就有可能被蒙蔽掉。

7-2 大數據與人工智慧應用案例簡介

示例7-11 用相關關係分析哪些車品質有問題——是什麼車有問題？

Kaggle 是一家為所有人提供資料採擷競賽平臺的公司，舉辦了關於二手車的品質競賽。經銷商將二手車資料提供給參賽者，統計學家們用這些資料建立一個演算法系統來預測經銷商拍賣的哪些車有可能出現品質問題。經過相關關係分析表明，橙色的車有品質問題的可能性只有其他顏色車的一半。

• 圖 7-13 二手車橙色汽車似乎品質比其他顏色好兩倍（圖片來源：123RF 圖庫）

7-2 大數據與人工智慧應用案例簡介

當讀到這裡的時候，不禁也會思考其中的原因。難道是因為橙色車的車主更愛車，所以車被保護得更好嗎？或是這種顏色的車子在製造方面更精良些嗎？還是因為橙色的車更顯眼、出車禍的概率更小，所以轉手的時候，各方面的性能保持得更好？

馬上，我們就陷入了各種謎一樣的假設中。若要找出相關關係，我們可以用數學方法，但如果是因果關係的話，這卻是行不通的。所以，我們沒必要一定要找出相關關係背後的原因，當我們知道了「是什麼(what)」的時候，「為什麼(why)」其實沒那麼重要了，否則就會催生一些滑稽的想法。比方說我們是不是應該建議車主把車漆成橙色呢？

7-2 大數據與人工智慧應用案例簡介

快速清晰的相關關係分析，比慢速的因果分析更有用和更有效。慢速的因果分析集中體現為透過嚴格控制的實驗來驗證的因果關係，而這必然是非常耗時、耗力的工作。

近年來，科學家一直在試圖減少這些實驗的花費，比如說，做成類似實驗(quasiexperiment)或稱準實驗。讓因果關係的調查成本降低，但還是很難與相關關係體現的優越性相抗衡。在大多數情況下，一旦我們完成了對大數據的相關關係分析，而又不再滿足於僅僅知道「是什麼」時，我們就會繼續向更深層次研究因果關係，找出背後的「為什麼」。

因果關係還是有用的，但是它將不再被看成是意義來源的基礎。因果關係只是一種特殊的相關關係。

7-2 大數據與人工智慧應用案例簡介

問題思考

從前面三節的說明，可否整理大數據思維轉變的三個方向？

	思維轉變的方向
思維轉變之一	樣本
思維轉變之二	接受數據的
思維轉變之三	找出數據的

小組討論：網路搜索與討論：大數據思維變革

1 實訓目的 在開始本實訓之前，請認真閱讀課程的相關內容。

(1) 熟悉大數據思維變革的基本概念和主要內容；

(2) 熟悉大數據時代人們思維變革的第一個轉變，全數據模式（樣本 = 總體 ）即 「分析更多資料而不再是只依賴於隨機採樣」

(3) 熟悉大數據時代人們思維變革的第二個轉變，接受數據的混亂允許不精確數 據，即「不再熱衷於追求精確度」。

(4) 熟悉大數據時代人們思維變革的第三個轉變，找出關聯物（數據的相關）關 係，即「找出相關關係、不再熱衷於尋找因果關係」。

MIGO 功典【零售業】

知名運動品牌大數據應用案例

P183,4:15 20161017

智慧零售：用大數據打造「最懂人」的服務

P.183, 4:07 20160808

實訓內容與步驟

請帶著問題分別進行網路搜索，尋求和思考答案，參加小組討論
(1) 尋找「小數據時代隨機採樣」的案例並進行簡單分析
(2) 尋找「有關產品推薦」的案例並進行簡單分析，探索其中大數據分析的因素 與作用。
(3) 請簡述，在大數據時代，為什麼「我們樂於接受資料的紛繁複雜，而不再一 味追求其精確性」？
(4) 請簡述，在大數據時代，為什麼「我們不再探求難以捉摸的因果關係，轉而 關注事物的相關關係」？

Chapter 7 作 業

一、重要關鍵字練習：根據英文關鍵字，把適當的中文編碼寫至對應的空格中

A. 統計學；統計；統計資料
B. 等於1,024GB（簡稱）；萬億位元組
C. 串流
D. 迅速；快速；速度
E. 容積；表示海量的資料規模
F. 種類；表示多樣的資料類型
G. 真實性；真實；誠實
H. 價值；估價；評價
I. 結構；構造；組織；構造體
J. 半結構化

題號	英文	中文	題號	英文	中文
1	stream	C	6	variety	F
2	semi-structured	J	7	structure	I
3	value	H	8	TB	B
4	volume	E	9	velocity	D
5	veracity	G	10	statistics	A

Chapter 7 作 業

一、重要關鍵字練習：根據英文關鍵字，把適當的中文編碼寫至對應的空格中

K. 相當於1,024TB；相當於百萬GB
L. 相互關係；關聯
M. 因果關係
N. 相關關係
O. 模糊不清的；有絨毛的
P. 線性關係
Q. 假定；設想
R. 曲線
S. 變數；變化；變動；變異
T. 採樣分析

題號	英文	中文	題號	英文	中文
11	correlationship	N	16	fuzzy	O
12	linear relationship	P	17	sampling analysis	T
13	variance	S	18	petabyte	K
14	curve	R	19	assumption	Q
15	correlation	L	20	causal relationship	M

Chapter 7 作 業

二、是非題

(✗) 1. 谷歌曾應用其大數據資料庫，進行流感趨勢預測。透過分析整個資料庫，分析了全美國幾十億條互聯網檢索紀錄。這樣雖然提高了微觀層面分析的準確性，但是無法推測出美國某特定城市的流感狀況。

(✗) 2. 大數據是完全禁止有混亂(chaos) 資料的，因為它會隨著資料的增加，錯誤率也會相應增加。在整合來源不同的各類資訊的時候，也禁止讓大數據加大混亂程度。

(✗) 3. 大數據處理時，數據間的相關關係無法預知未來的，所以並不是大數據分析師所需要的處理方式。

(○) 4. 允許不精確的數據出現已經成為大數據的一個優點。

Chapter 7 作 業

三、選擇題（單複選）

(C) 1. 19 世紀以來，當面臨大量資料時，社會都依賴於採樣分析。但是採樣分析是_____時代的產物。

(A) 電腦 (B)青銅器 (C)類比資料 (D) 雲

(C) 2. 長期以來，人們已經發展了一些使用盡可能少的資訊技術。例如，傳統統計學的一個目的就是_____

(A) 用盡可能多的資料來驗證一般的發現

(B) 同盡可能少的資料來驗證盡可能簡單的發現

(C) 用盡可能少的資料來證實盡可能重大的發現

(D) 用盡可能少的資料來驗證一般的發現。

Chapter 7 作 業

(A) 3. 因為大數據是建立在_____，所以我們才可以正確地考察細節並進行新的分析。

(A) 掌握所有資料，至少是盡可能多資料的基礎上

(B) 在掌握少量精確資料的基礎上，盡可能多地收集其他資料

(C) 掌握少量資料，至少是盡可能精確的資料的基礎上

(D) 盡可能掌握精確資料的基礎上

Chapter 7 作 業

(A) 4. 直到今天，我們的數位技術依然建立在精準的基礎上，這種思維方式適用於掌握_____的情況。

(A) 小數據量 (B) 大數據量 (C) 無數據 (D)多數據

(A) 5. 當人們擁有海量即時資料時，絕對的精準不再是人們追求的主要目標。當然，_____。

(A) 我們應該適當地放棄精確度，不再沉迷於此

(B) 我們不能放棄精確度，需要努力追求精確度

(C) 我們應完全放棄了精確度，且不再沉迷於此

(D) 我們是確保精確度的前提下，適當尋求更多資料

Chapter 7 作 業

(B) 6. 為了獲得更廣泛的資料而犧牲了精確性，也因此看到了很多無法被關注到的細節。_____。

(A) 在很多情況下，與致力於避免錯誤相比，對錯誤的包容會帶給我們更多問題

(B) 在很多情況下，與致力於避免錯誤相比，對錯誤的包容會帶給我們更多好處

(C) 無論什麼情況，我們都不能容忍錯誤的存在

(D) 無論什麼情況，我們都可以包容錯誤

Chapter 7 作 業

(A) 7. 過去，統計學家們總是把他們的興趣放在提高樣本的隨機性而不是數量上。這是因為_____。

(A) 提高樣本隨機性可以減少對資料量的需求

(B) 樣本隨機性優於對大數據的分析

(C) 可以獲取的資料少，提高樣本隨機性可以提高分析準確率

(D) 提高樣本隨機性是為了減少統計分析的工作量

Chapter 7 作 業

(A) 8. 研究表明，在少量資料情況下運行得最好的演算法，當加入更多的資料時，_____。

(A) 也會像其他的演算法一樣有所提高，但是有可能卻變成了在大量資料條件下運行得較不好的

(B) 與其他的演算法一樣有所提高，仍然在大量資料條件下運行最好的

(C) 與其他的演算法一樣有提高，在大量資料條件下運行得比較好的

(D) 雖然沒有提高，還是在大量資料條件下運行得最好的

Chapter 7 作 業

(D) 9. 在大數據時代，要想獲得大規模資料帶來的好處，混亂應該是一種_____。

(A) 不正確途徑，需要竭力避免的

(B) 非標準途徑，應該儘量避免的

(C) 非標準途徑，但可以勉強接受的

(D) 標準途徑，而不應該是竭力避免的

(C) 10. 研究表明，只有_____的數位資料是結構化的且能適用於傳統資料庫。如果不接受混亂，剩下_____的非結構化資料都無法被利用。

(A) 95%，5% (B) 30%，70% (C) 5%，95% (D) 70%，30%

Chapter 7 作 業

(B) 11. 尋找_____是人類長久以來的習慣，即使確定這樣的關係很困難而且用途不大，人類還是習慣性地尋找緣由。

(A) 相關關係 (B) 因果關係 (C) 資訊關係 (D) 組織關係

(A) 12. 在大數據時代，我們無需再緊盯事物之間的_____，而應該尋找事物之間的_____，這會給我們提供非常新穎且有價值的觀點。

(A) 因果關係，相關關係 (B)相關關係，因果關係

(C) 複雜關係，簡單關係 (D)簡單關係，複雜關係

Chapter 7 作 業

(C) 13. 所謂相關關係，其核心是指量化兩個資料值之間的數理關係。相關關係強是指當一個資料值增加時，另一個資料值很有可能會隨之_____。

(A) 減少 (B)顯現 (C)增加 (D)隱藏

(D) 14. 透過找到一個現象的_____，相關關係可幫助捕捉現在和預測未來。

(A) 出現原因 (B) 隱藏原因

(C) 一般的關聯物 (D) 良好的關聯物

Chapter 7 作 業

(A) 15. 大數據時代，專家們正在研發能發現並對比分析非線性關係的技術工具。透過_____，相關關係幫助我們更好地瞭解了這個世界。

(A) 探求「是什麼」而不是「為什麼」

(B) 探求「為什麼」而不是「是什麼」

(C) 探求「原因」而不是「結果」

(D) 探求「結果」而不是「原因」

Chapter 7 作 業

(A) 16. 有一個使用大數據的飛機保養與修護系統，系統可提醒飛機在故障之前更換零件或者修復問題。這是使用_____分析方法。

(A) 相關關係 (B) 採樣 (C)分類 (D)神算

(B) 17. 大數據的3V 特點中，除了Volume（容積）外，還有可用5V 或來代表，_____就是3V 的代表。（請選三項以上）

D

(A) Veracity（真實性） (B) Velocity（速度）

(C) Value（價值） (D) Variety（種類）

Chapter 7 作 業

(C) 18. 大數據的混亂(chaos) 資料，是指_____。（請選兩項）

D (A) 只要取樣得宜，在萃取或處理資料的時候，就不會發生資料混亂

(B) 隨著資料的增加，錯誤率相應增加，其中的部分資料必須是正確的

(C) 在整合來源不同的各類資訊時，因格式不完全一致，會加大混亂程度

(D) 在大數據背景下很難做到清洗資料(data cleansing)

Chapter 7 作 業

(A B) 19. 關於資料結構化與非結構化的觀點，下列說法不正確的是_____。（請選兩項以上）

(A) 據估計，有95%的數位資料是屬於結構化資料

(B) 非結構化資料才能適用於傳統資料庫

(C) 網頁和影像（影片）資源大多屬於非結構化資料

(D) 非結構化資料可能包括很多模糊(fuzzy）和不確定(uncertain）的數據