大數據是什麼

- 大數據（Big Data）又被稱為**巨量資料**，其概念其實就是過去10年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**。

一次搞懂大數據

大數據是什麼

- 「**Big Data**」這詞最早由 IBM 提出，2010 年才真正開始受到注目，並成為專業用語登上維基百科 ，算是「大數據」的正式問世。
- 2012 年時，《紐約時報》的專欄文章「**The Age of Big Data** 」更是宣告了「大數據時代」的來臨。
- 大數據不是新興的概念，歐洲粒子物理研究中心（CERN）的科學家已面對巨量資料的問題好幾十年了，處理著每秒上看 PB（Peta Bytes，註：PB = 1,024 TB）的資料量

Big Data - Tim Smith

大數據是什麼

Gartner 公司的分析師 Doug Laney

2001 年

3D Data Management: Controlling Data **Volume**, **Velocity**, and **Variety**.

- 資料量、速度、多樣性

2012 年

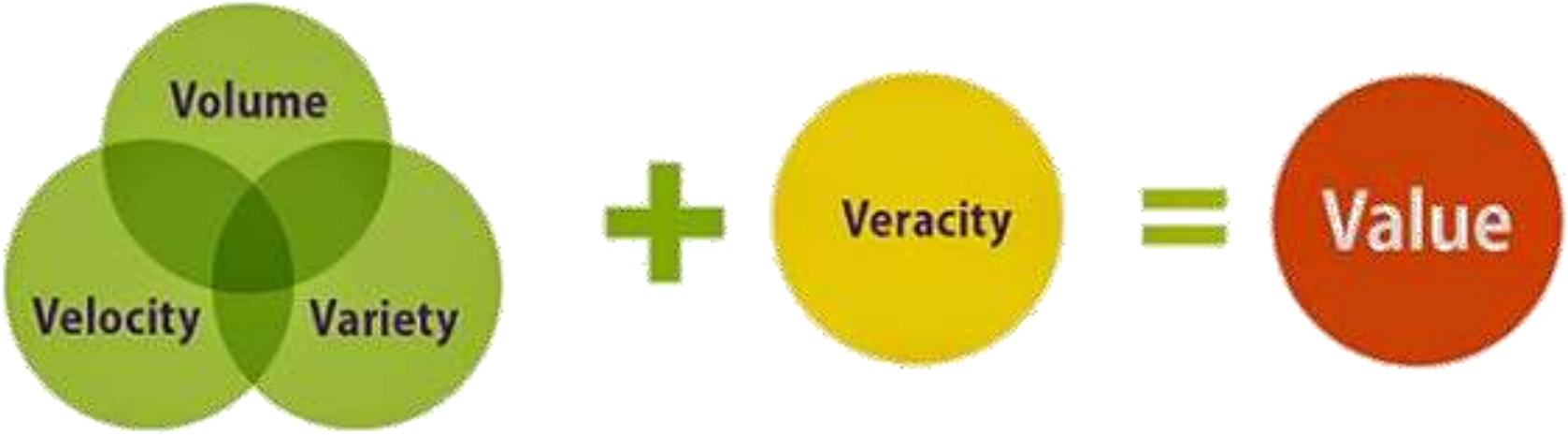
Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

- 大數據是大量、高速、及/或類型多變的資訊資產，它需要全新的處理方式，去促成更強的決策能力、洞察力與最佳化處理。

Big Data: Expanding on 3 fronts at an increasing rate.

大數據是什麼

- 大數據的定義
  - Volume容量、Velocity速度 和 Variety多樣性
  - 有人另加上**Veracity真實性和Value價值**兩個V。
- 大數據的資料特質和傳統資料最大的不同
  - **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**。


Volume
Velocity
Variety
Veracity
Value

Slide

大數據和商業分析之異同

	傳統商業分析	大數據分析
資料分析方法	抽樣分析	使用大數據的原始資料（raw data）進行統計分析、視覺化、產出報表
可預測性	制定模型以跑出期望中的數據結果	大數據最重要的，在於探索資料、發現新模型、找出資料間的關聯性，進而達到精準預測
所需耗費時間	透過資料庫正規化、視覺化，費力耗時、學習門檻高	可即時支援業務單位進行市場策略調整與配合
使用技術	資料庫、正規化、結構化基礎的分析	原始資料、大數據、非結構化和結構化分析、分散式儲存、分散式運算

資料來源：Splunk

為什麼需要大數據

- Ovia Fertility的App
  - 精準計算排卵期，提高懷孕的機率
- Workday推出一套軟體
  - 預測員工的薪水漲幅和可能跳槽時間，幫助企業決定每名員工的加薪幅度、時間點和轉職時機。
- 微眾銀行
  - 結合辨識人臉和公安部門資料，決定借貸者的信用等級。
- 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
- 對一般民眾而言，大數據是另一個自我，它可能比本人更了解本人

大數據從哪來

- 任何地方
  - 隨著物聯網興起，任何以前不可能產生資料的東西或地方都可能「資料化」
- 天睿資訊（Teradata）首席技術長寶立明認為大數據的發展可以分成三階段
  - 第一階段.com時期
  - 第二階段社交網站
  - 第三階段物聯網

大數據從哪來

- **第一階段.com時期**
  - 人們研究log資料，蒐集人們的Cookie和搜尋行為等等，
  - 我們不只知道使用者買了什麼東西而已，而是更深層地去分析行為，一筆交易只說明了價值，但沒有說明顧客體驗，大數據想要去分析的是顧客體驗。
- **第二階段社交網站**
  - 在正在經歷的階段，分析Facebook、Twitter、部落格文章...等等等，這可以幫助我們進一步了解顧客行為。

大數據從哪來

– 第三階段物聯網

- 無論是機器還是人都開始被數據解構，數據可能來自手錶、鞋墊甚至皮帶，這些物聯網數據將是接下來重要的數據分析對象。
- 例如汽車每半年就要進廠維修，就跟人每年都要去做健康檢查一樣，是非常過時的想法，一旦用感測器去蒐集引擎、汽車和生理數據，就可以精確知道何時需要進廠維修或做健康檢查，這就叫做預測性維修（condition based maintenance），這個概念對於促進顧客體驗、效益和健康保險等領域非常重要。

商業模式

- 大數據的商業模式大概可分成幾種：

1. 從既有數據變現
2. 以數據提升企業競爭力
3. 以數據做為服務的基礎與核心，用數據顛覆傳統行業。

商業模式

- 模式一，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。

商業模式

- 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。

商業模式

- 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。
- 但模式三，也就是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。

Volume 資料量

Data volume: amount of data

- 以前人們「手動」在**表格**中記錄、累積出數據
- 現在數據是由**機器、網路、人與人之間的社群互動**來生成。
  - 你現在正在點擊的滑鼠、來電、簡訊、網路搜尋、線上交易... 都正在生成累積成龐大的數據，因此資料量很容易就能達到數 TB（Tera Bytes，兆位元組），甚至上看PB（Peta Bytes，千兆位元組）或 EB（Exabytes，百萬兆位元組）的等級。

Velocity 資料輸入輸出速度

Data velocity: speed of data in and out

- 資料的傳輸流動（data streaming）是連續且快速的，隨著越來越多的機器、網路使用者，社群網站、搜尋結果每秒都在成長，每天都在輸出更多的內容。公司跟機構要處理龐大的資訊大潮向他們襲來，而回應、反應這些資料的速度也成為他們最大的挑戰，許多資料要能即時得到結果才能發揮最大的價值，因此也有人會將 Velocity 認為是「時效性」。

Variety 資料類型

Data variety: range of data types and sources

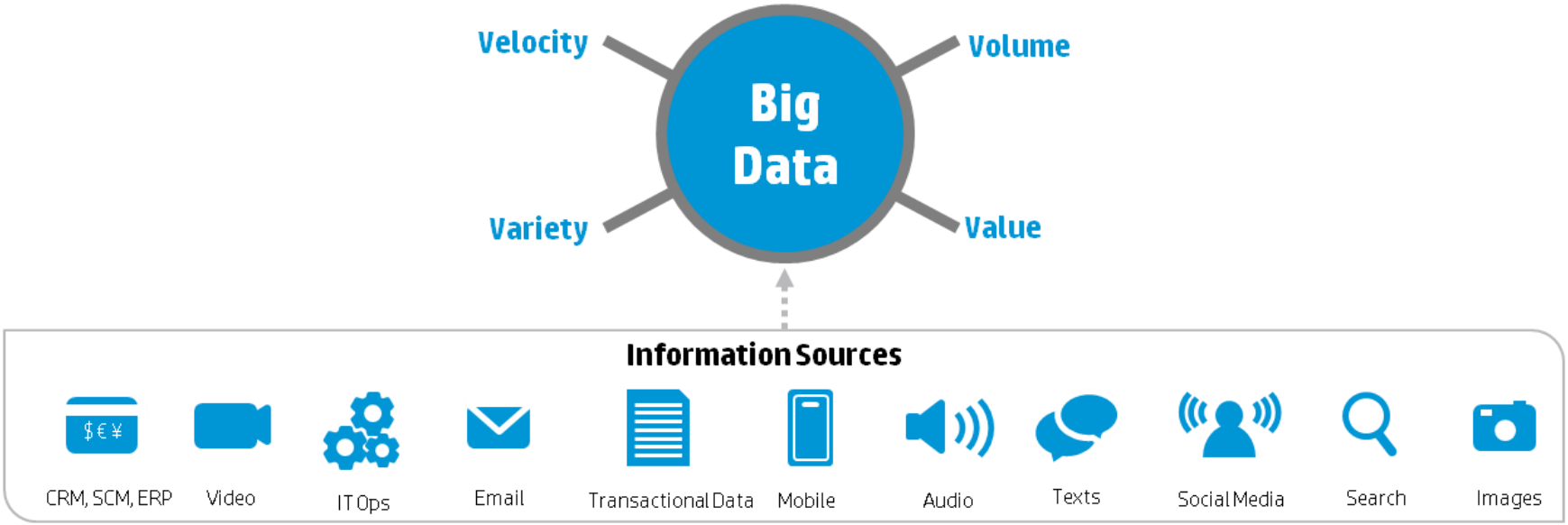
- 大數據的來源種類包羅萬象，十分多樣化，如果一定要把資料分類的話，最簡單的方法是分兩類，結構化與非結構化。
- 早期的非結構化資料主要是文字，隨著網路的發展，又擴展到電子郵件、網頁、社交媒體、視訊，音樂、圖片等等，這些非結構化的資料造成儲存（storage）、探勘（mining）、分析（analyzing）上的困難。

Veracity 真實性

Data veracity: uncertainty of data

- 這個詞由在 Express Scripts 擔任首席數據官（Chief Data Officer, CDO）的 Inderpal Bhandar 在波士頓 大數據創新高峰會（Big Data Innovation Summit） 的演講中提出，認為大數據分析中應該加入這點做考慮，分析並過濾資料有偏差、偽造、異常的部分，防止這些「dirty data」損害到資料系統的完整跟正確性，進而影響決策。

大數據特性，謹記四字箴言：「大、快、雜、疑」



大數據資料量龐「大」（Volume）、變化飛「快」（Velocity），種類繁「雜」（Variety），以及真偽存「疑」（Veracity）。尤其在這資訊大爆炸時代，這些資料變得又多、又快、又雜、又真偽難分。

大數據即科技 Big Data as Technology

- 現今要處理的資料量更龐大、資料產生跟處理速度更驚人、資料來源更多樣，於是處理、儲存大量資料的新技術跟工具快速發展，像是開源軟體 **Hadoop 跟** NoSQL 資料庫。
- 大數據不只是指資料，也指這些用來分析、處理巨量資料的新興科技。

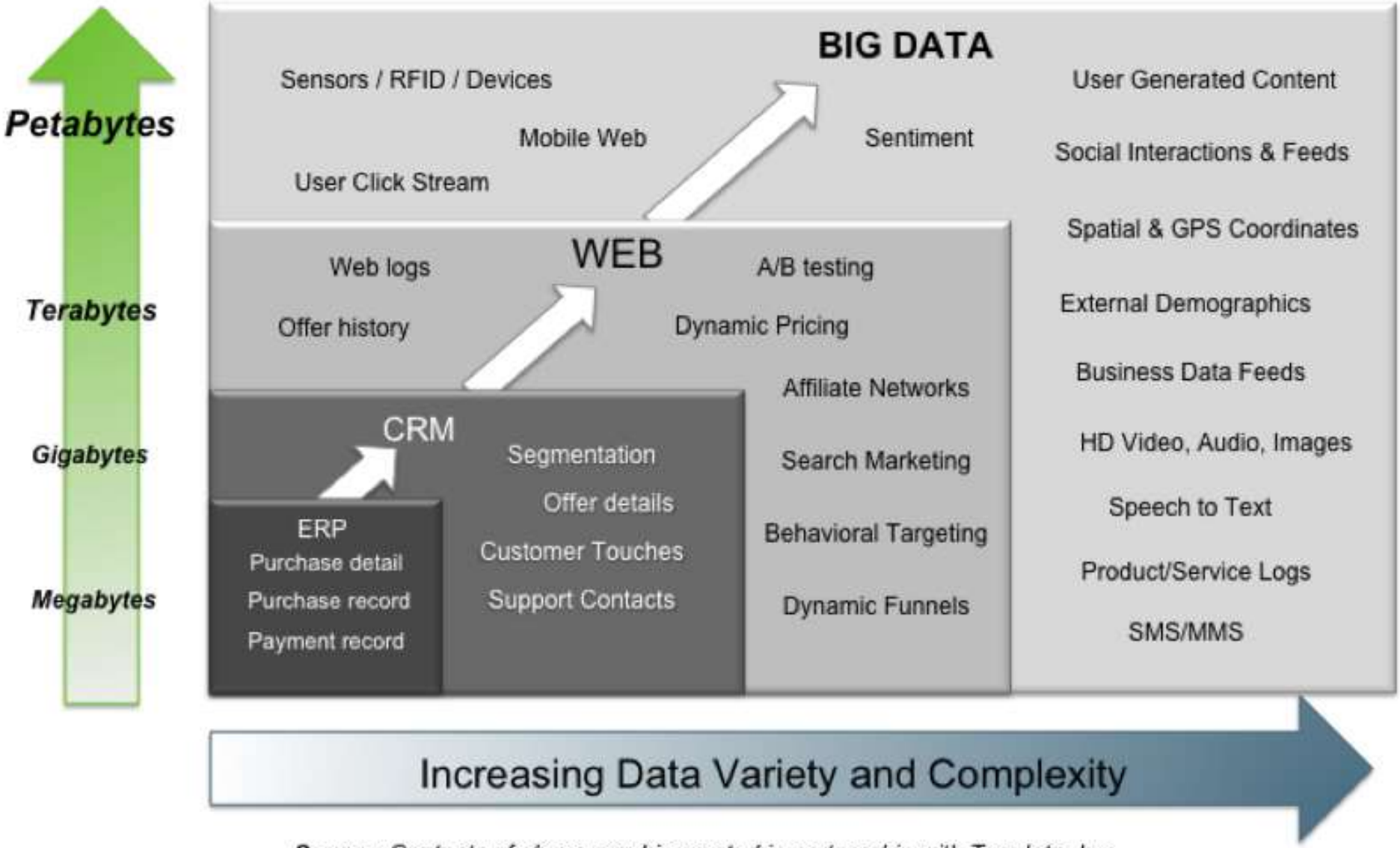
"Big Data is the new tools helping us find relevant data and analyze its implications."

大數據即不同的資料類型
Big Data as Data Distinctions

- Hortonworks 公司戰略副總裁 Shaun Connolly 定義大數據是由交易、互動、觀察資料所組成的資料型態。

"Big Data = Transactions + Interactions + Observations"

Big Data = Transactions + Interactions + Observations



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

參考資料

- **「大數據」到底與我有什麼關聯？ 5 張圖，一次弄懂商業界的熱門關鍵字！**
- 巨量資料的時代，用「大、快、雜、疑」四字箴言帶你認識大數據
- 7 個你不可不知的大數據定義
- 美國Top 4 技術長寶立明：大數據即將在五年內消失