

If you can't measure it, you can't manage it. –彼得·杜拉克

資料科學產業應用

Data Science & Business Application

Lecture 1 – Basics of Data Science

郭俊良 博士

逢甲大學-勞動部雲端運算與數位轉型培訓班

MCNU Logistics Big Data Analytics Research Center

About Me



現職

銘傳大學管理學院
風險管理與保險系助理教授

學歷

國立成功大學資訊管理研究所
Syracuse University (美國雪城大學資訊研究所)

博士
碩士

證照

PMI 國際專案管理師 (PMP1763215)

經歷

國防大學運籌管理系
空軍官校電算中心
Syracuse Internet Café Inc. (USA)

助理教授
資訊系統工程師
助理系統分析師

研究領域

智慧型預測維修模式、文本探勘（自然語言處理）
社群網路意見探勘、存管需求預測、金融保險科技

授課

研究所：資料科學、機器學習、深度學習
大學部：人工智慧概論、管理資訊系統、智慧型金融保險



數據資料

企業最誘人的職缺

Data Scientist: The Sexiest Job of the 21st Century

湯瑪斯·戴文波特 Thomas H. Davenport, 帕蒂爾 D.J. Patil

2012年10月號(績效解密) | 2012/10/1

瀏覽人數 : 15745



Preface

- Class Objectives
- Class Arrangement
- Class Materials
- Syllabus



Class Objectives

- Establishing basic knowledge and expertise and lay the foundation for data analytics
- Incorporating the theoretical methods and case-studies while importing programming skills to help students fulfill the real-world industry issues
- Lay the ability to leverage Data Science techniques to distilling Business Intelligence of Industry

Class Arrangement

- **Schedule**

- 1st Stage: DS Basics (17, 18 Dec. 2022)
- 2nd Stage: DS Modeling (13, 14, 15 Jan. 2023)

- **Segmentation**

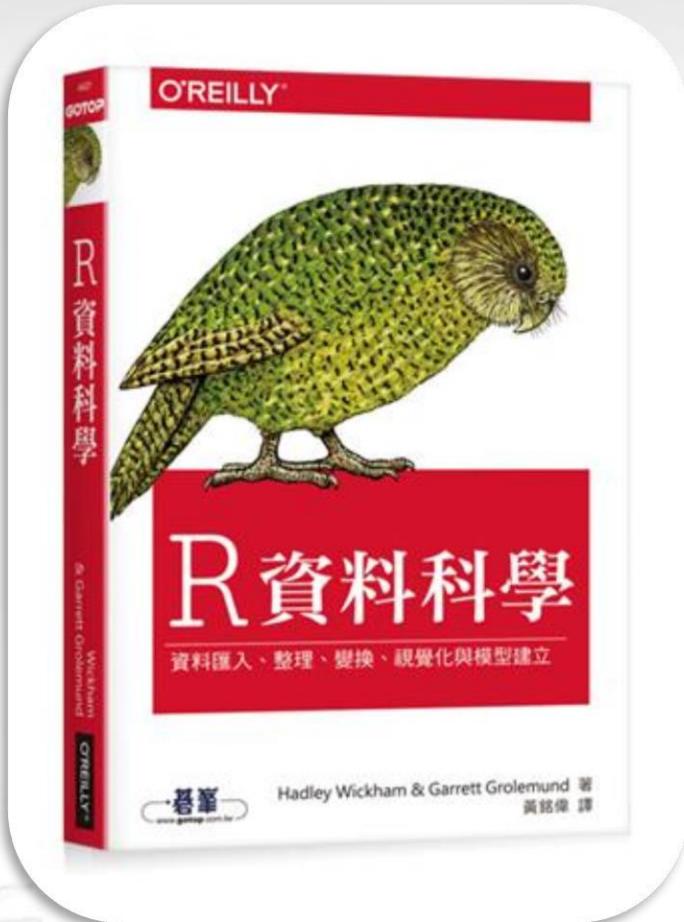
- Lecture (2hrs) + Lab. Practice (1hr)
- Case Study (1hr) + Workshop (2hrs)

- **Hardware Requirement**

- PC/NB (High-end device is recommended)
 - Spec.: 8Gb RAM, 500GB HD
- Internet Connection



Class Materials

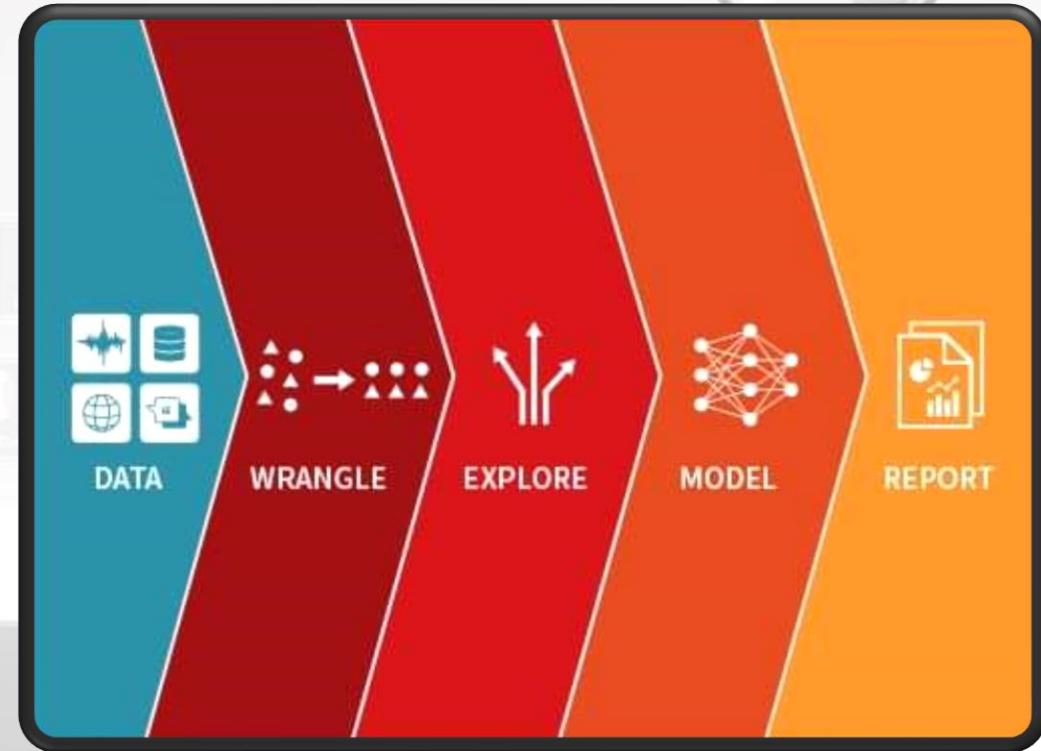


- 雲端硬碟網址
 - <https://reurl.cc/XIAWZ3>
- 課程教材
 - 主要教材
 - ✓ 自訂上課教材(投影片、範例程式)
 - 參考教材
 - ✓ [R資料科學](#)，譯者：黃銘偉，出版社：碁峯(原文：O'REILLY)，ISBN-13: 9789864764808
 - ✓ 線上教材: <https://github.com/hadley/r4ds>
- 使用軟體
 - 程式工具
 - ✓ R Language (Open Source): 線上教材 <https://bookdown.org/jefflinmd38/r4biost/intro.html>
 - ✓ IDE: RStudio (Academic Use Only)

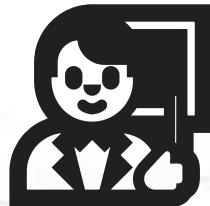
Syllabus

- Day 1 – Basics of Data Science
 - Introduction of ICT Era
 - Introduction of Data Science
- Day 2 – Data Analysis (Wrangle & Explore)
 - Data Understanding and Preparation
 - Exploratory Data Analysis
- Day 3 – Modeling
 - Regression Analysis (Linear Regression)
 - Classification Analysis (Logistic Regression)
- Day 4 – Modeling
 - Classification Analysis (ANN)
 - Classification Analysis (Decision Tree)
- Day 5 – Modeling & System Deployment
 - Clustering Analysis
 - Data Science in Business Application
 - 金融產業應用-FinTech, InsurTech & 程式交易

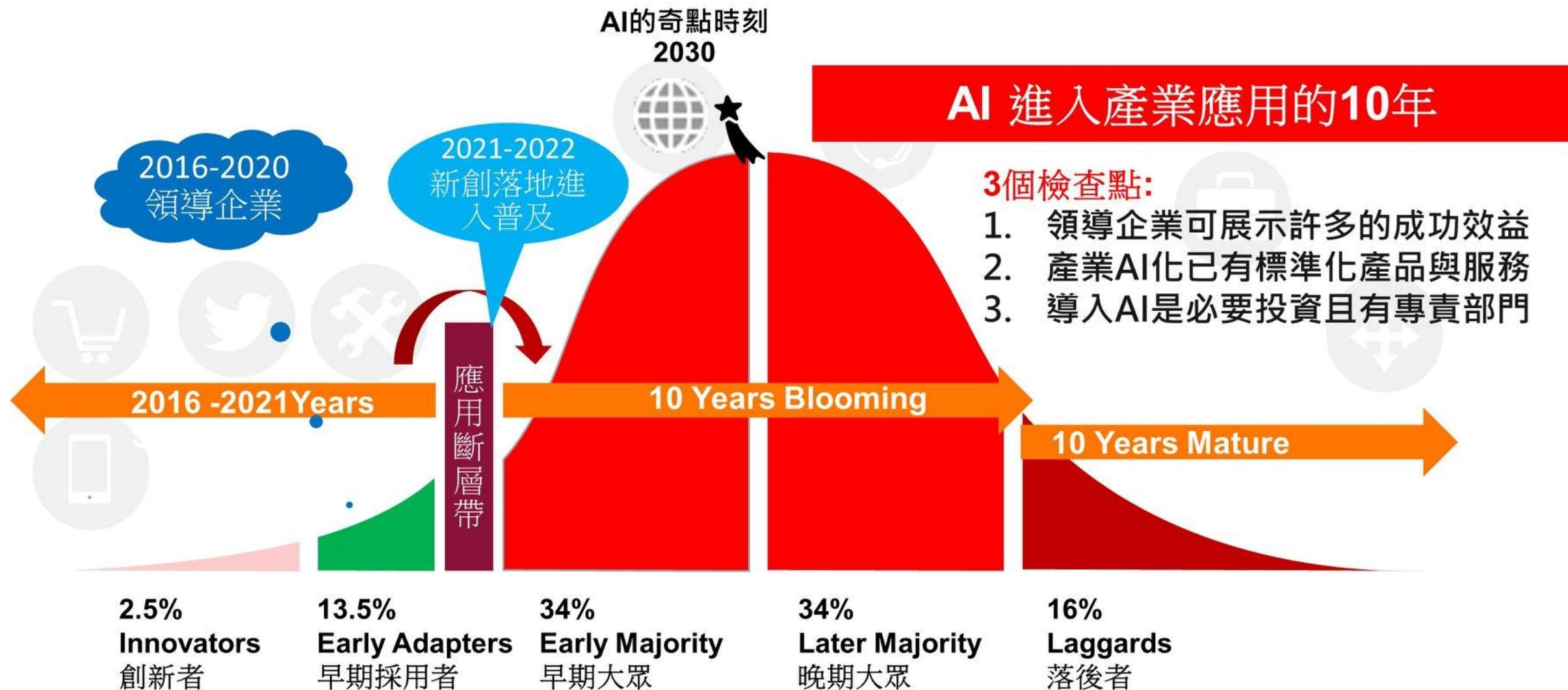
May This Wisdom Guide You to Cross Over Learning Ridge



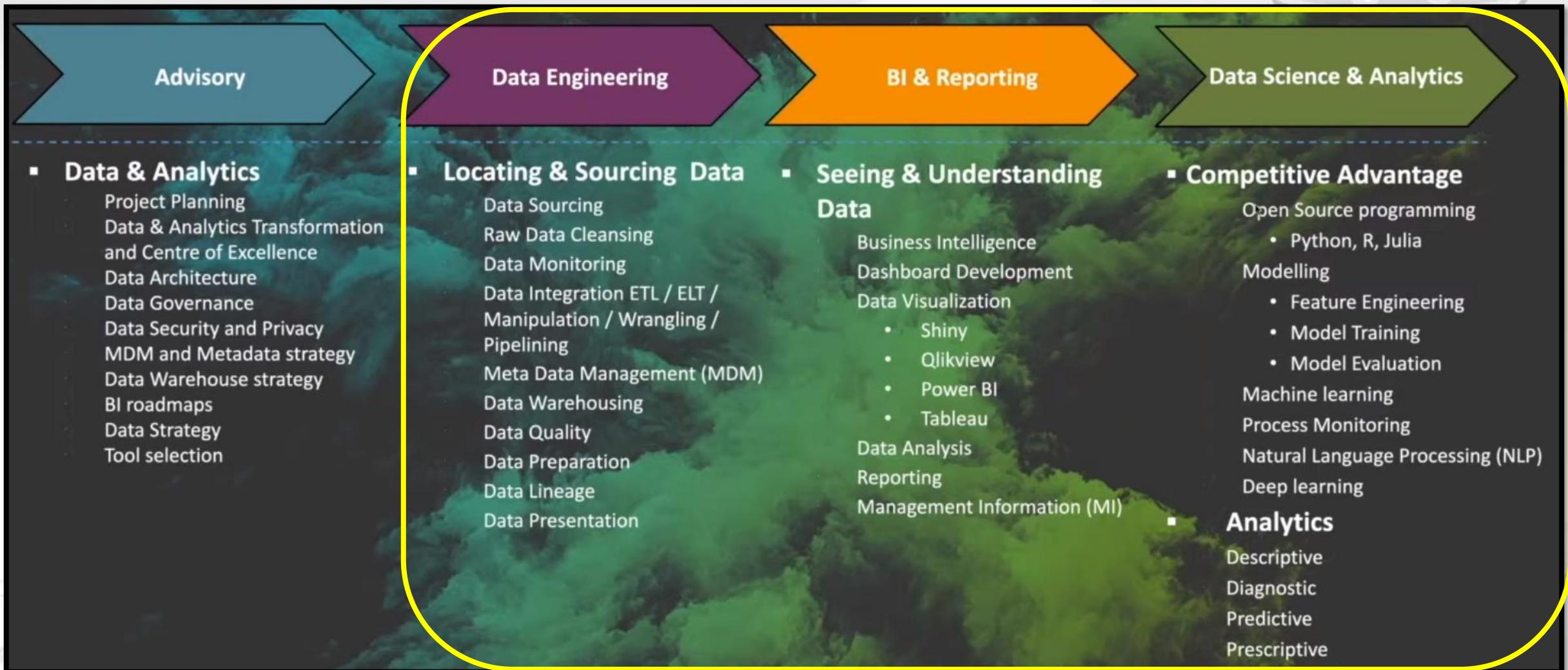
Before We Kick Off.....



產業AI化應用的黃金10年

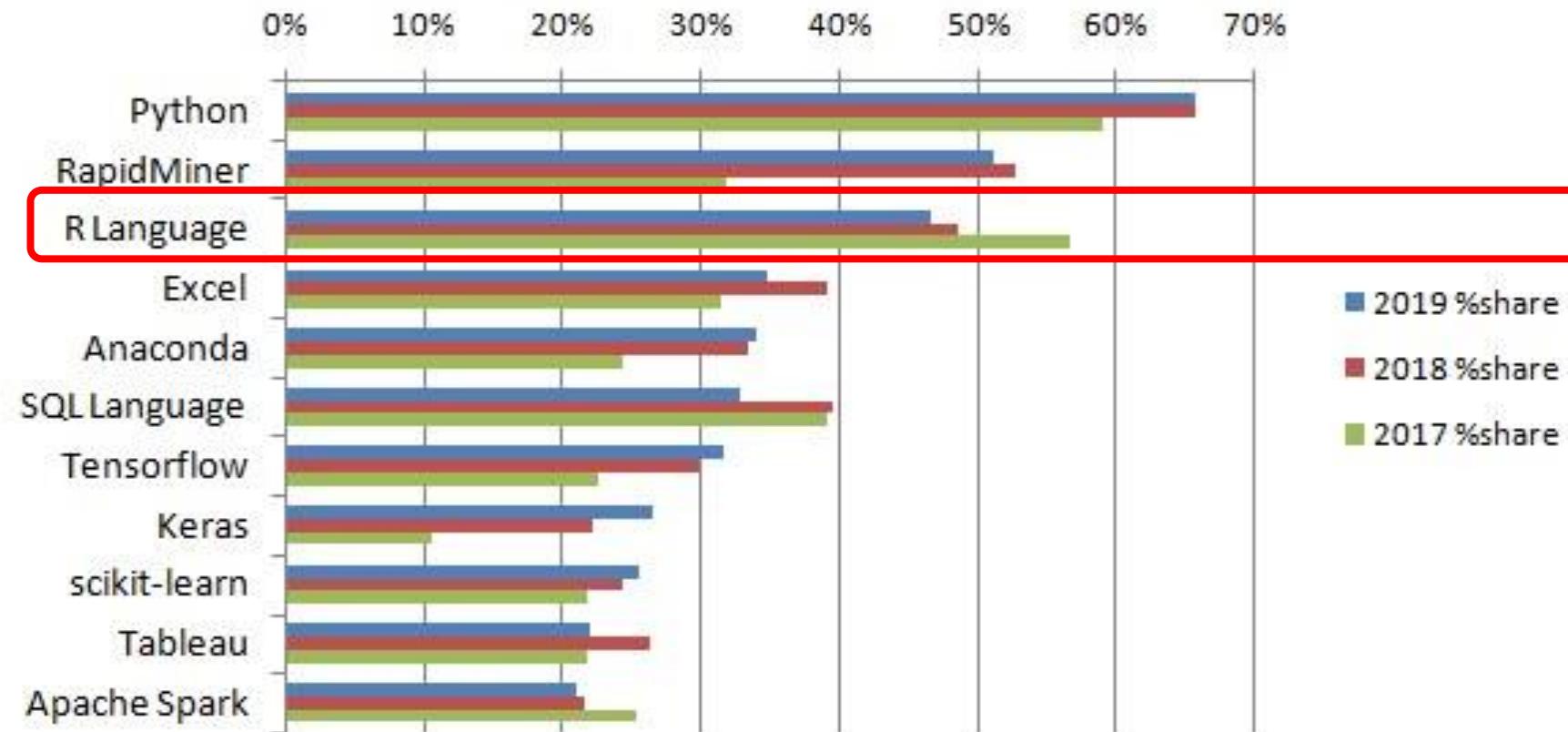


Data Science & Analytics Ecosystem



Something You Should Know

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



The Ten Most Common Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	72%
R	64%
SQL	51%
Hadoop	39%
Java	33%
SAS	30%
Spark	27%
Matlab	20%
Hive	17%
Tableau	14%

Source: Glassdoor Economic Research.

glassdoor

PROGRAMMING LANGUAGE	Data extraction and manipulation	Statistical analysis and data visualizations	Modeling/ Machine learning	Model deployment	Automation
SQL	✓	✗	✗	✗	✗
PYTHON	✓	✓	✓	✓	✓
R	✓	✓	✓	✗	✓
JAVA	✗	✓	✓	✓	✓
JAVASCRIPT	✗	✓	✓	✓	✗
JULIA	✗	✓	✓	✗	✓
C	✗	✗	✓	✗	✗
C++	✗	✗	✓	✗	✗
C#	✗	✗	✗	✓	✓
TYPESCRIPT	✗	✗	✓	✗	✗
HTML/CSS	✗	✗	✗	✓	✗
PHP	✗	✗	✗	✓	✗
RUST	✗	✗	✗	✓	✗
GOLANG	✗	✗	✗	✓	✗
BASH/SHELL	✗	✗	✗	✗	✓

R vs. Python

Python Strengths	R Strengths
Machine Learning	Statistics
Deep Learning	Econometrics
Apps	Statistical Modeling (& Machine Learning)
APIs	Reporting (& Communication)
	Web Apps
	APIs
	Integrates Python

Source: <https://www.business-science.io/careers/2022/03/11/which-data-science-skills-are-important.html>

Something You need to know

基礎

- 迴圈與條件判斷式
- 物件導向程式設計(Class, Function)
- 資料結構：列表、矩陣、向量等



資料處理

- Pandas表格處理
- Matplotlib簡單視覺化
- Numpy簡易資料運算



- Tidyverse表格處理
- ggplot簡單視覺化
- 簡易資料運算

網路爬蟲(Web Crawler)

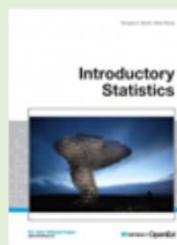
- 看懂HTML與CSS網頁架構
- 學習用BeautifulSoup、httr等套件組合獲取網頁資訊
- 結合資料處理技巧得到資料
- 學習json並串接網站API



練習透過爬蟲抓取每月熱門電影名稱以及票房

統計分析

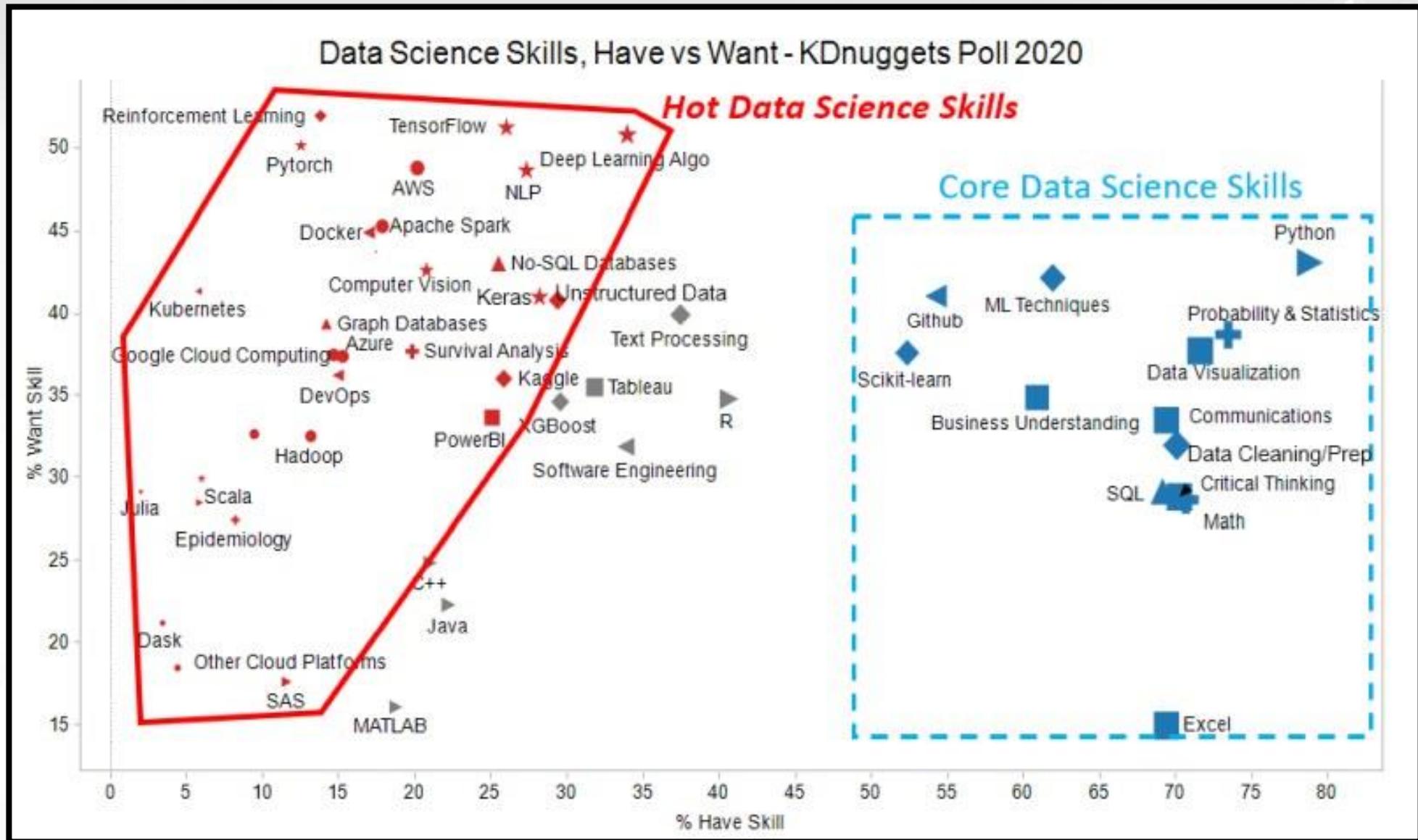
(Statistical Analysis)



用統計課本上的方法處理資料，並練習寫一份簡易的報告



Something You Want to Know



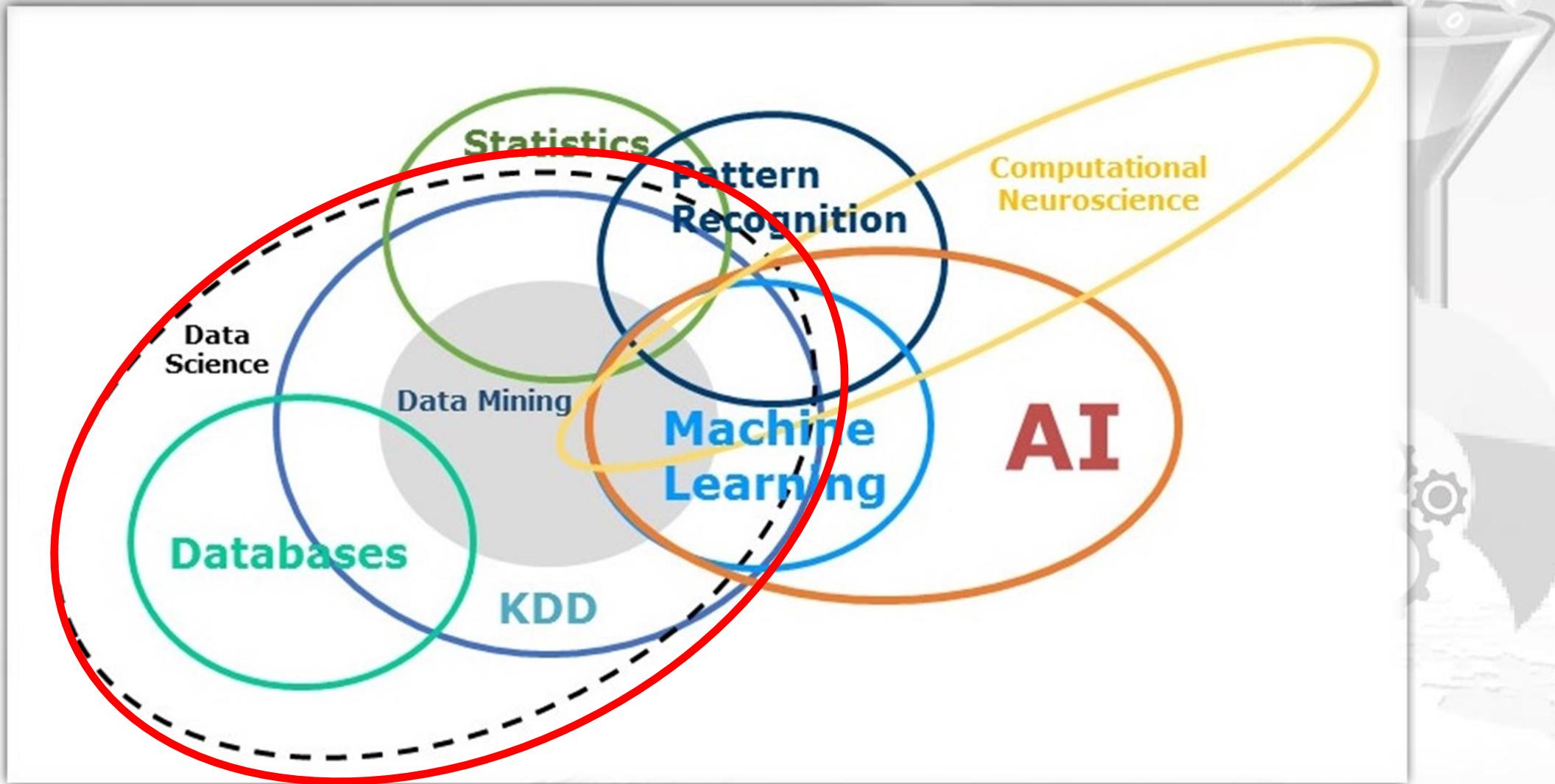
Source: Kdnuggets, <https://www.kdnuggets.com/2020/09/modern-data-science-skills.html>

Skills You Need to Learn

Plan	Skills
Machine Learning	Supervised Classification, Supervised Regression, Unsupervised Clustering, Dimensionality Reduction, Local Interpretable Model Explanation - H2O Automatic Machine Learning, parsnip (XGBoost, SVM, Random Forest, GLM), K-Means, UMAP, recipes, lime
Data Visualization	Interactive and Static Visualizations, ggplot2 and plotly
Data Wrangling & Cleaning	Working with outliers, missing data, reshaping data, aggregation, filtering, selecting, calculating, and many more critical operations, dplyr and tidyr packages
Data Preprocessing & Feature Engineering	Preparing data for machine learning, Engineering Features (dates, text, aggregates), Recipes package
Time Series	Working with date/datetime data, aggregating, transforming, visualizing time series, timetk package
Forecasting	ARIMA, Exponential Smoothing, Prophet, Machine Learning (XGBoost, Random Forest, GLMnet, etc), Deep Learning (GluonTS), Ensembles, Hyperparameter Tuning, Scaling to 1000s of forecasts, Modeltime package
Text	Working with text data, Stringr
NLP	Machine learning, Text Features
Functional Programming	Making reusable functions, sourcing code
Iteration	Loops and Mapping, using Purrr package
Reporting	Rmarkdown, Interactive HTML, Static PDF
Applications	Building Shiny web applications, Flexdashboard, Bootstrap
Deployment	Cloud (AWS, Azure, GCP), Docker, Git
Databases	SQL (for data import), MongoDB (for apps)

Source: <https://www.business-science.io/careers/2022/03/11/which-data-science-skills-are-important.html>

Where does DS Located



Source: <https://jamesmccaffrey.wordpress.com/2016/09/29/machine-learning-data-science-and-statistics/>

Where does DS apply



Introduction of ICT Era

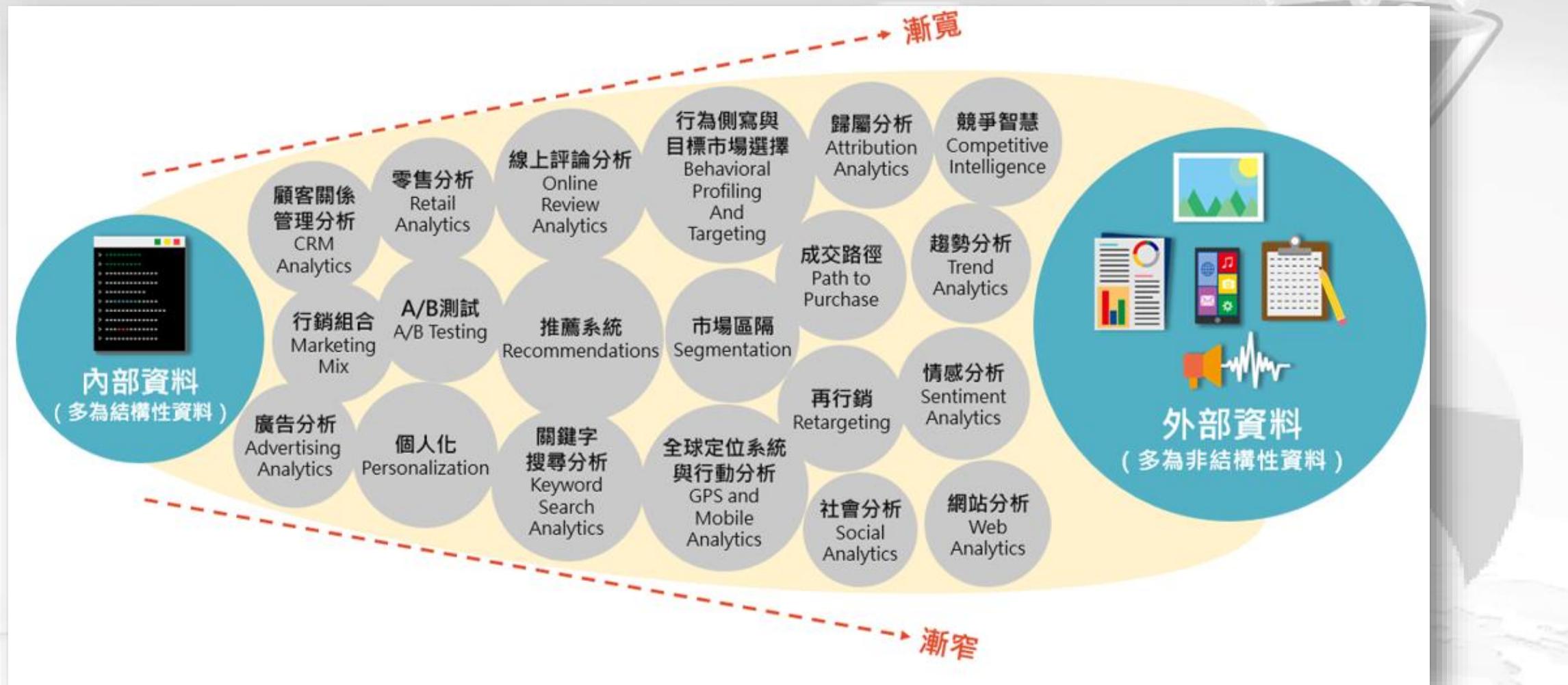
From Data to Knowledge

Interpreting and Mining Intelligence from Data

Era of ICT

- 資訊通訊科技 (Information & Communication Technology, ICT) 開啟人類社會全面進入數位轉型時代的濫觴
- 資訊科技的進步和網路的發達、電腦運算能力的增強，以及資料蒐集與儲存技術持續改進的影響，加速資料的取得與累積，大幅改變資料的應用方式
- 「大數據」(Big Data)或稱「巨量資料」分析(資料科學的特例)可以探索及洞察(Insight)先前未知且潛在的資料樣態 (Pattern)，進而轉化為有價值的資訊或知識
- 資料的「質」與「量」高度影響的資料分析(Data Analytics)技術的選用，技術需聚焦在萃取出有意義樣態(pattern)或規則(rule)的能力
- 促使資料科學的分析技術和應用快速朝人工智慧發展

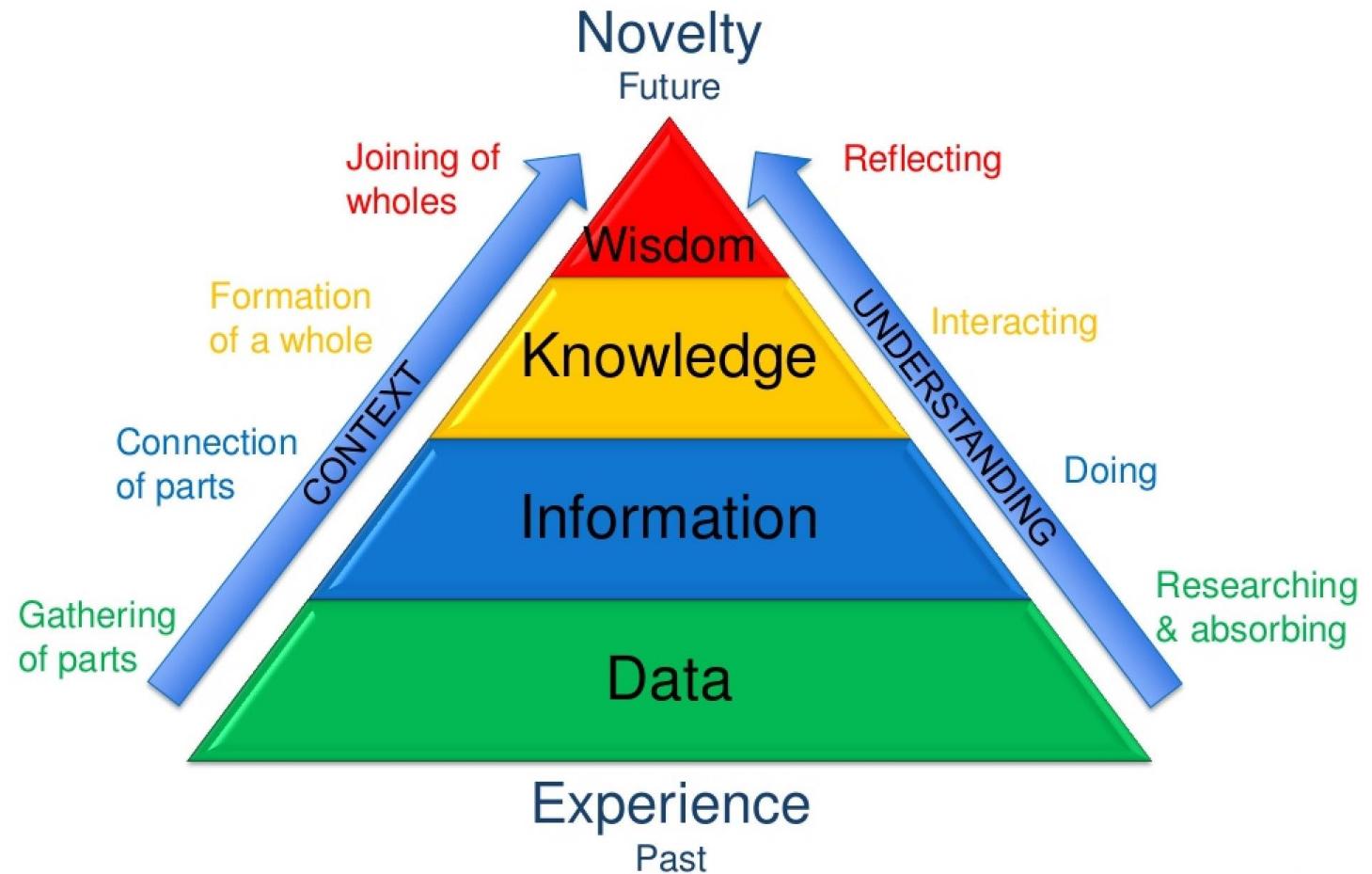
資料類型與資料科學



資料來源:行銷資料科學 <https://www.facebook.com/146686936002012/photos/146693992667973/>

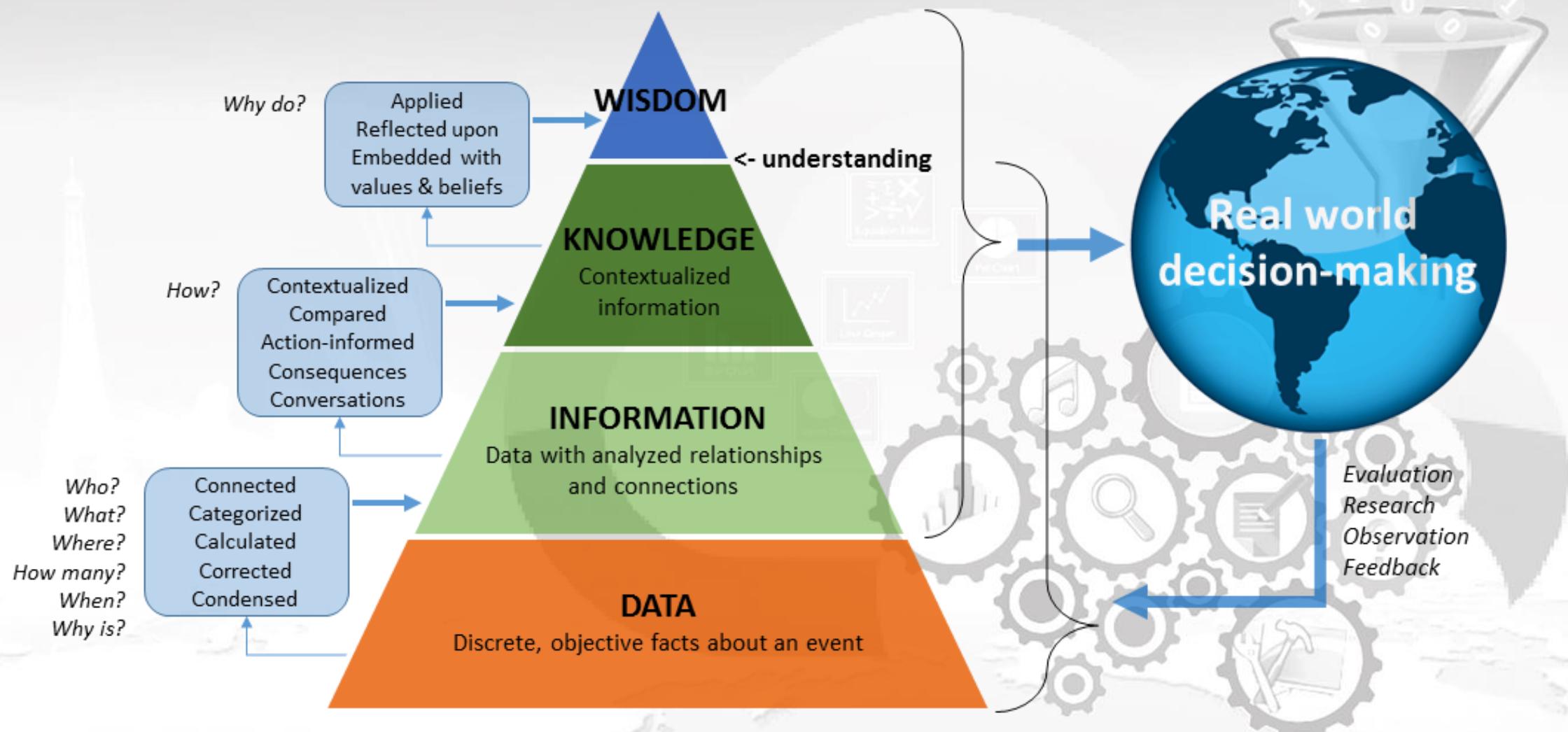
DIKW Pyramid

DIKW Hierarchy



Source: <https://www.slideshare.net/DHA2015/tim-staniland-ixc-dha-masterclass-presentation>

Data-driven Decision-making



Source: <https://www.climate-eval.org/blog/answer-42-data-information-and-knowledge>

結構化資料

固定欄位
固定格式
固定順序

半結構化 資料

具有欄位，但
不一致，如人
力銀行網站上
的職務內容、
各種Log檔

非結構化 資料

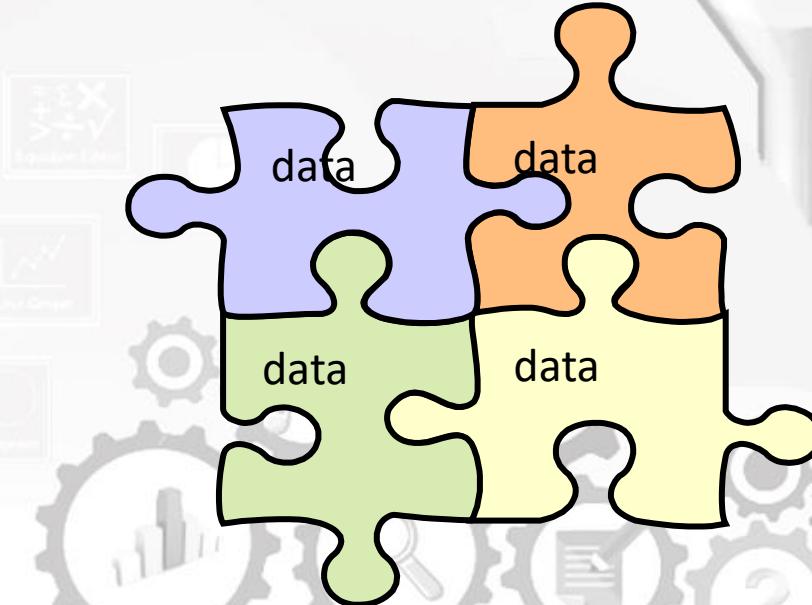
如影像檔、語
音檔、圖檔、
Office檔案、
PDF檔、網頁
、e-mail等

關聯式資料庫

非關聯式資料庫

資訊 (Information)

- 「資料」經過處理與分析並賦予意義，轉換成具潛在價值的「資訊」
 - Data + Data + + Data → Information
 - 例如：
 - ✓ 零售店每月交易金額最高的十項商品
 - ✓ 每日交易的尖峰時段等
 - ✓ 庫存水準的最佳化模式
 - ✓ 裝備維修時機的預測
 - ✓ 賣場商品擺設位置與銷量關聯性
- 資訊可以影響決策者的想法及判斷力，且具有關聯性和目標性
- 每筆資料可能根據不同的衡量尺度 (Scale) 而記錄，在分析整理前，必須先瞭解資料的尺度，必要時先轉換成所需的資料格式



資料到資訊的轉換過程

Data

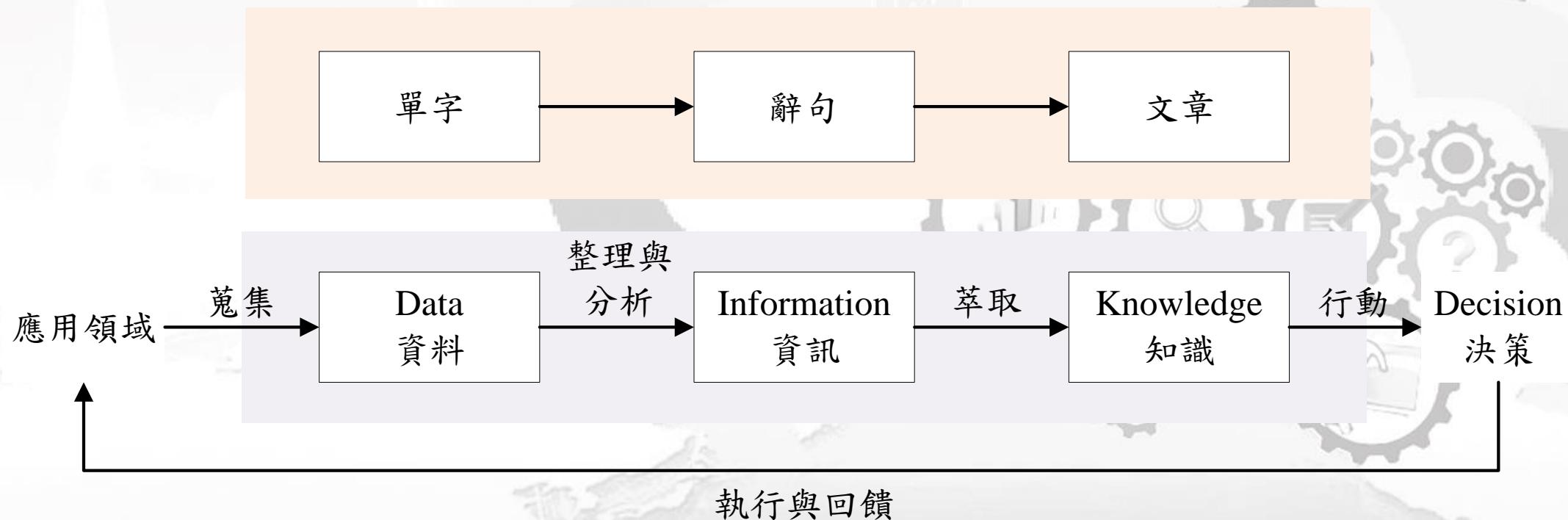
重點在於資料
內容的品質，
不在於資料分
析工具

- 文字化(**Contextualized**)：明白資料蒐集的目的與背景
- 分類(**Categorized**)：瞭解資料分類的項目與分析單位
- 計算(**Calculated**)：透過數學或統計分析來分析資料
- 更正(**Corrected**)：移除資料中的錯誤
- 濃縮(**Condensed**)：將修正後的資料濃縮成更簡單的形式

Information

知識 (Knowledge/Intelligence)

- 從資訊進一步歸納及演譯，但不僅止於資訊所傳遞的訊息。綜合經驗、價值及資訊，並且成為一種接收、評估、整合其他新經驗的架構
- 知識存在於文件和儲存系統中，也遍及在日常工作等規範中
- 資訊轉換到知識的所有環節都需要人的參與



知識的層級

淺

- 認知性的知識 (Cognitive Knowledge; **know-what**)
 - 知與不知之間
- 專業技能 (Advanced Skills; **know-how**)
 - 知其然，亦須知其所以然
- 系統化洞察 (System Understanding; **know-why**)
 - 物有本末，事有始終，知所先後，則近道矣！
- (自發性)創新 (Self-motivated Creativity; Care-why-know why you and/or others want to know why, **how and what**)
 - 舉一反三

深

資訊轉變成知識的過程

Information

Value Added:

- observation
- interpretation
- understanding
- experience
- skills

人



比較(Comparison)：這情形和以前碰過的有什麼不同？

結果(Consequence)：這資訊對決策與行動有什麼啟示？

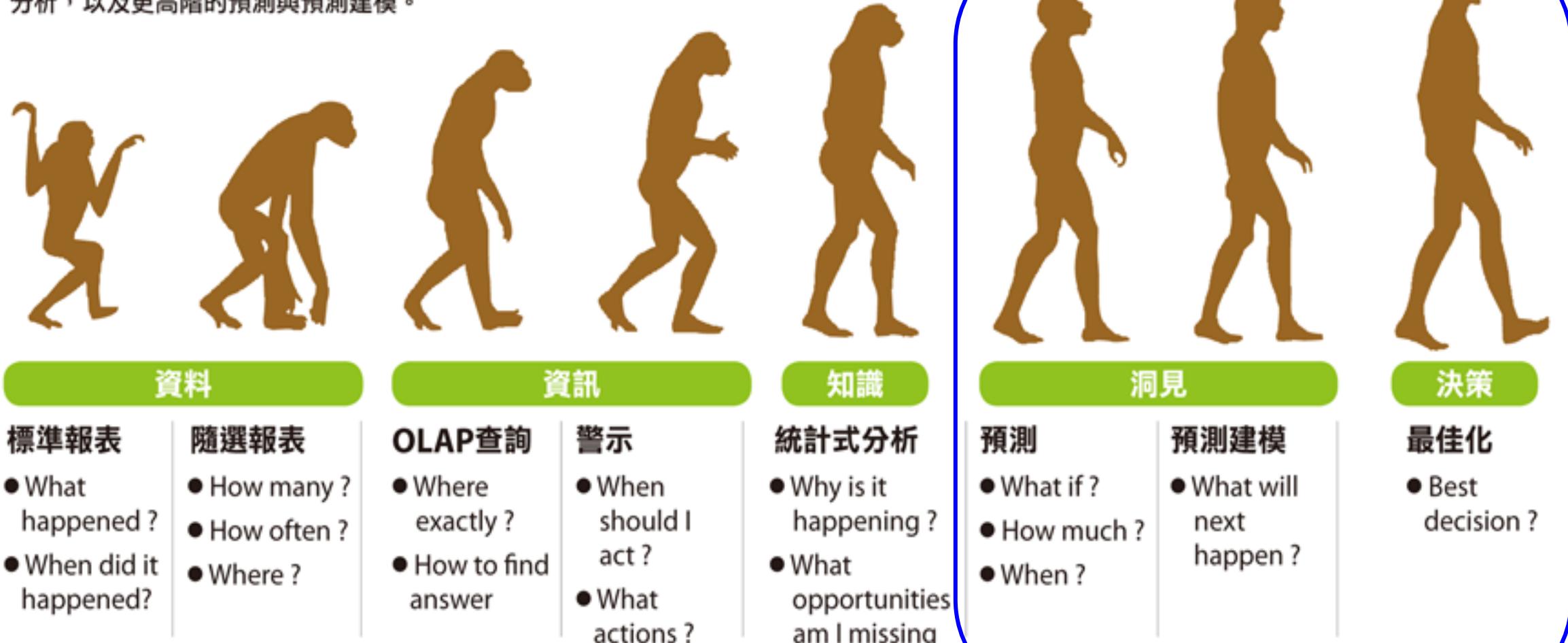
關聯性(Connections)：這些知識和其他部分的知識有什麼關聯？

交談(Conversation)：其他人覺得這資訊怎樣？

Knowledge

商業智慧演進的 8 個階段

目前的商業智慧應用，大多局限在報表的呈現，進階一點可提供隨選報表，甚至是 OLAP 查詢和警示，這些可以提供給決策者的只是資料和資訊，若需要的是知識，需要運用統計分析，以及更高階的預測與預測建模。



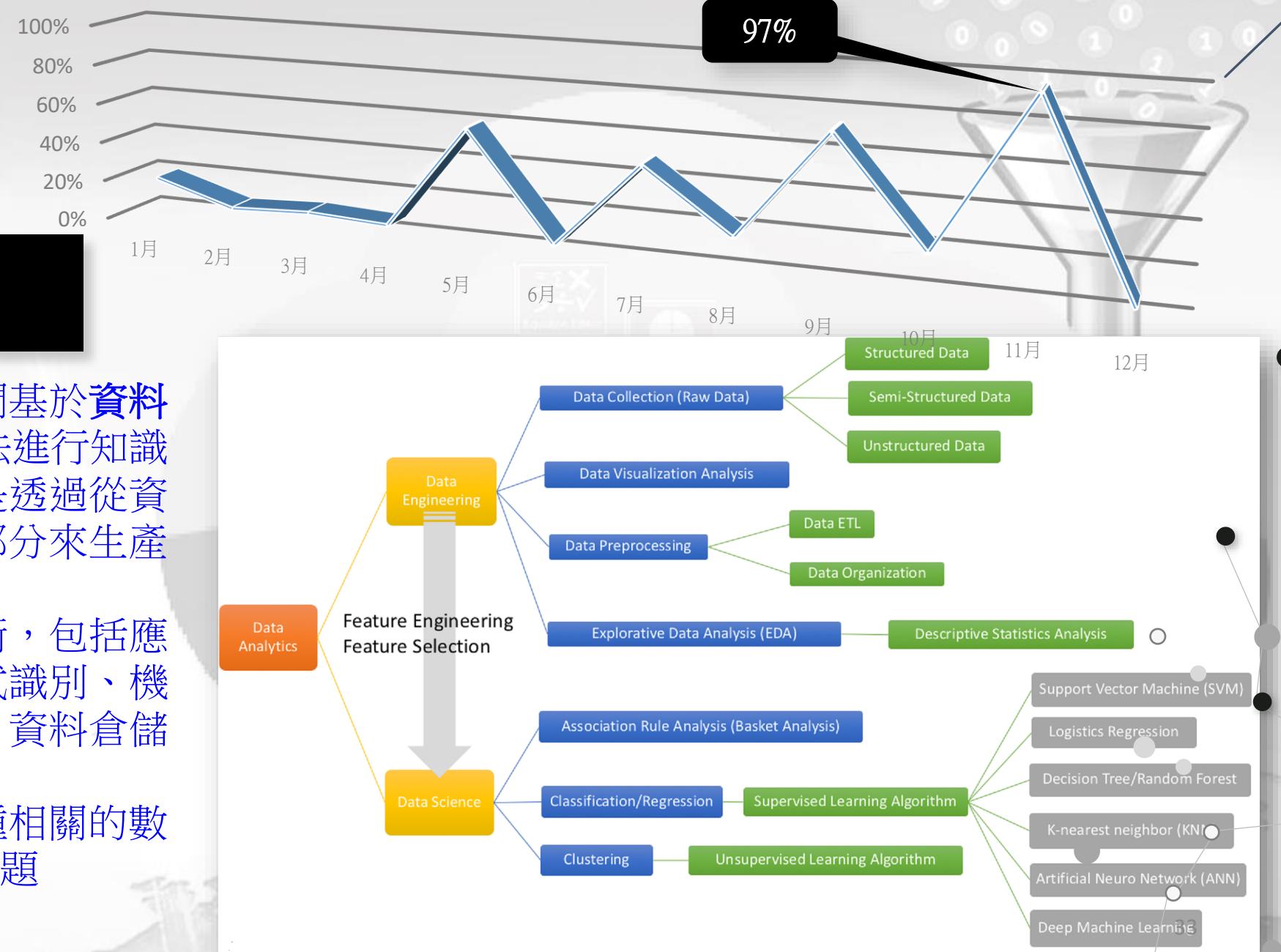
資料來源：SAS，iThome整理，2014年3月

Definition, Essenes, Data Mining, Methods, & Application

Introduction of Data Science

何謂資料科學

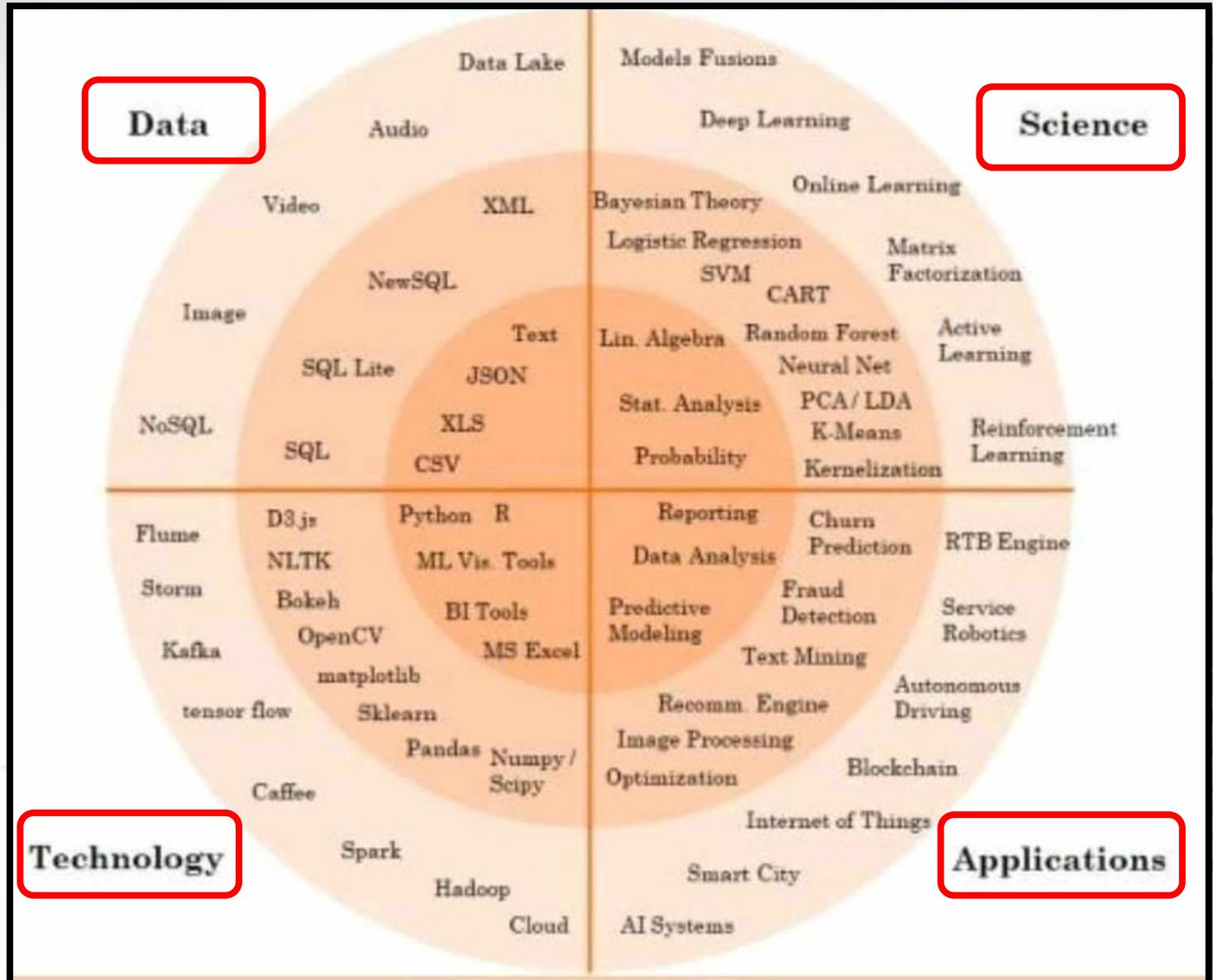
- 又稱**資料科學**，是一門基於**資料驅動 (Data-driven)** 方法進行知識學習的學科，其目標是透過從資料中提取出有價值的部分來生產資訊產品
- 結合跨領域理論和技術，包括應用數學、統計學、模式識別、機器學習、資料視覺化、資料倉儲以及高性能計算
- 資料科學通過運用各種相關的數據來協助理解並解決問題



Other Definitions

- Data Science (a.k.a. **Data Mining**) is about explaining the past and predicting the future by means of data analysis
- An interdisciplinary field about processes and systems
- Extract knowledge or insights from data in various forms, either structured or unstructured
- Consist of various data analysis techniques such as statistics, machine learning, data mining, and predictive analytics
- Similar to Knowledge Discovery in Databases (KDD)

資料科學的4個象限

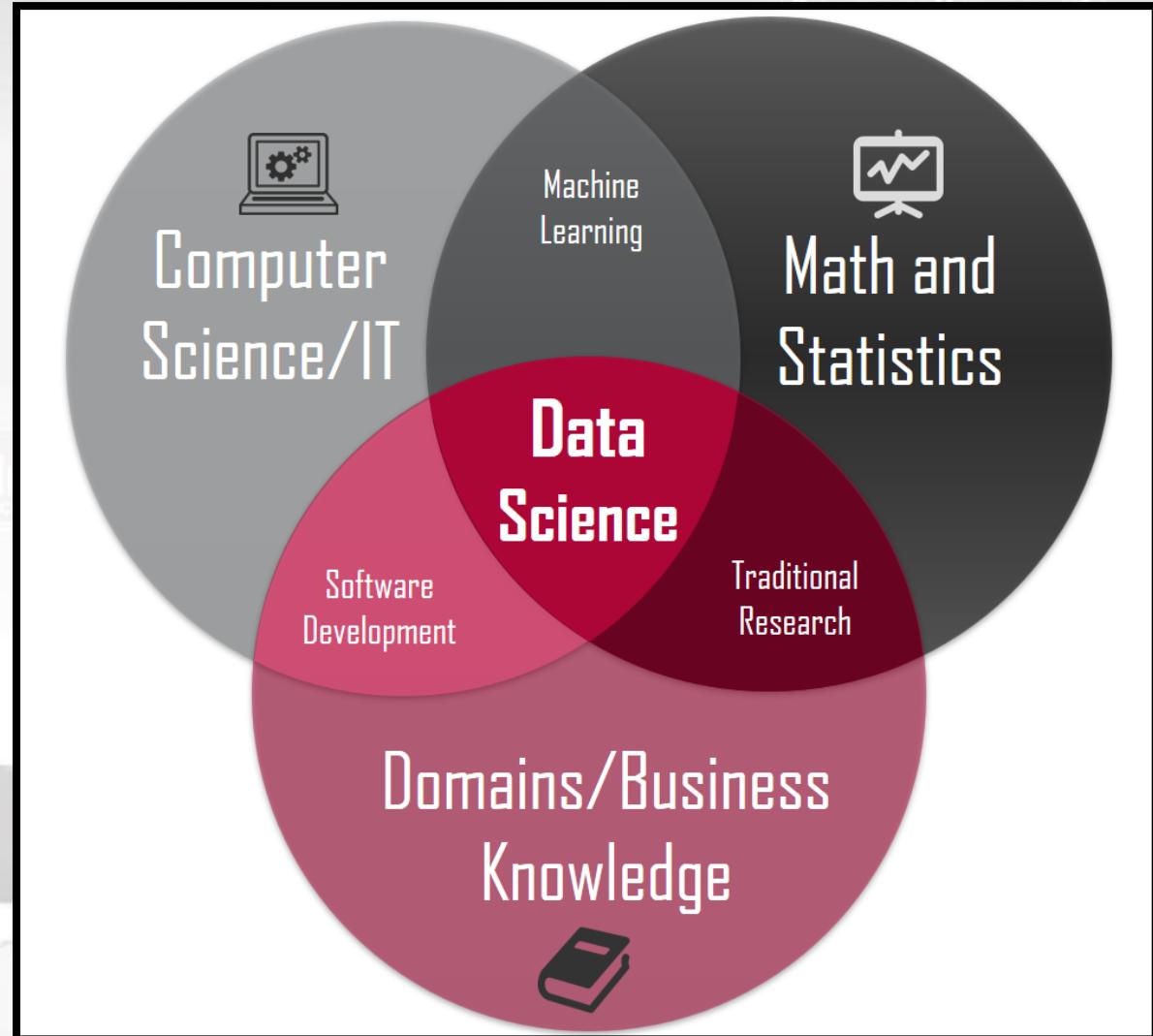
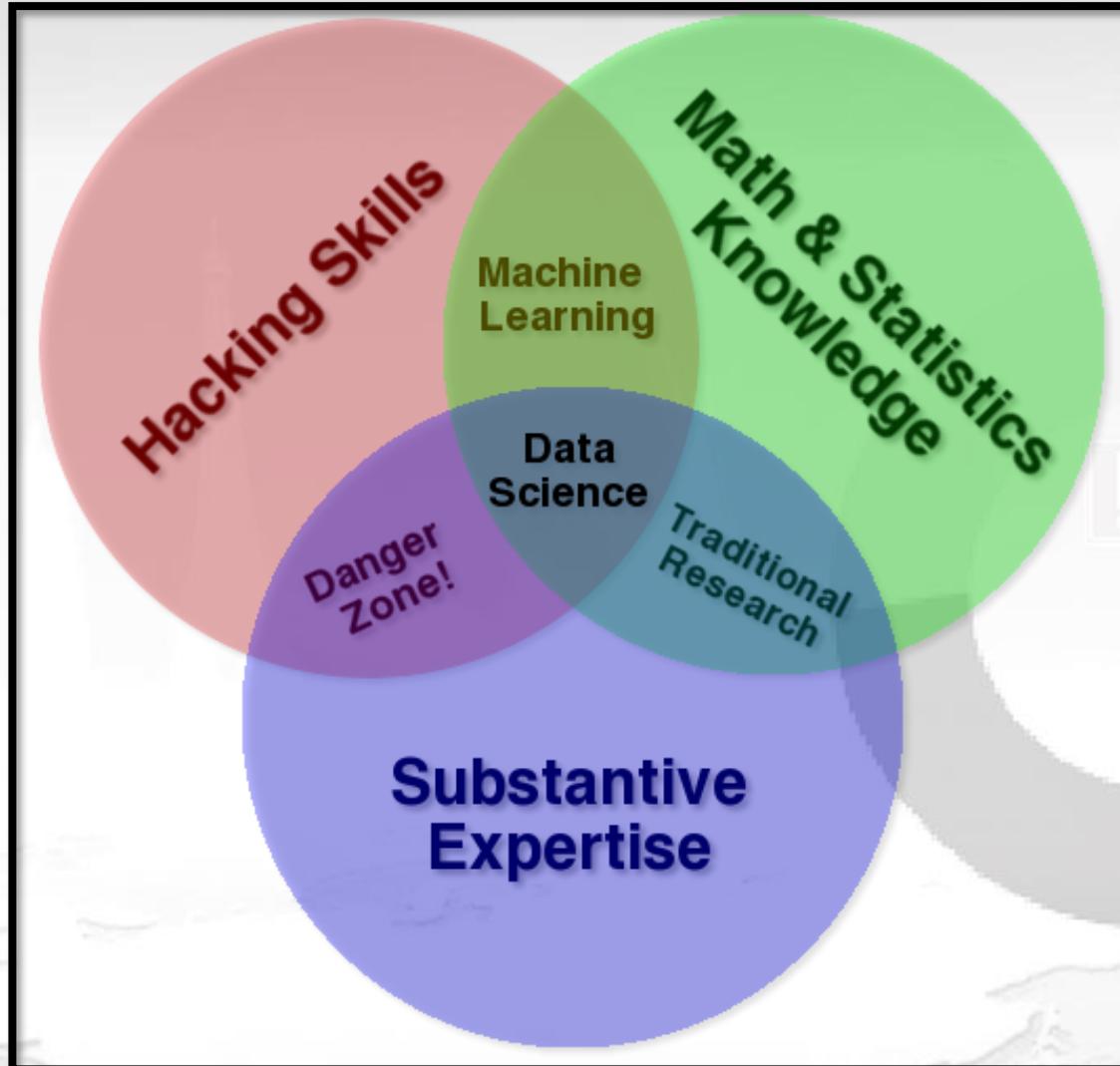


資料科學發展方法論



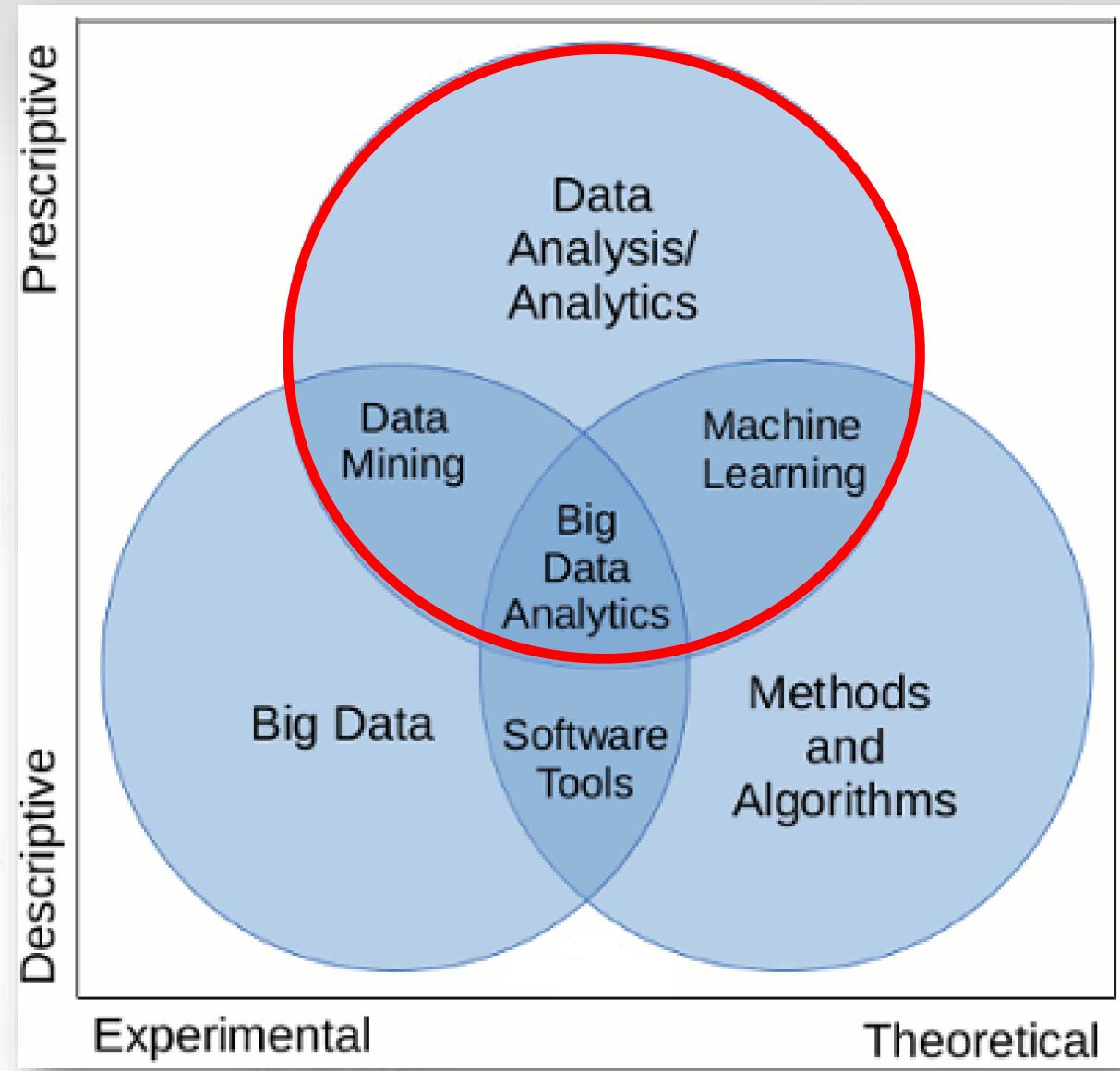
資料來源：獨立評論 <https://opinion.cw.com.tw/blog/profile/52/article/4901>

Venn Diagram of Data Science

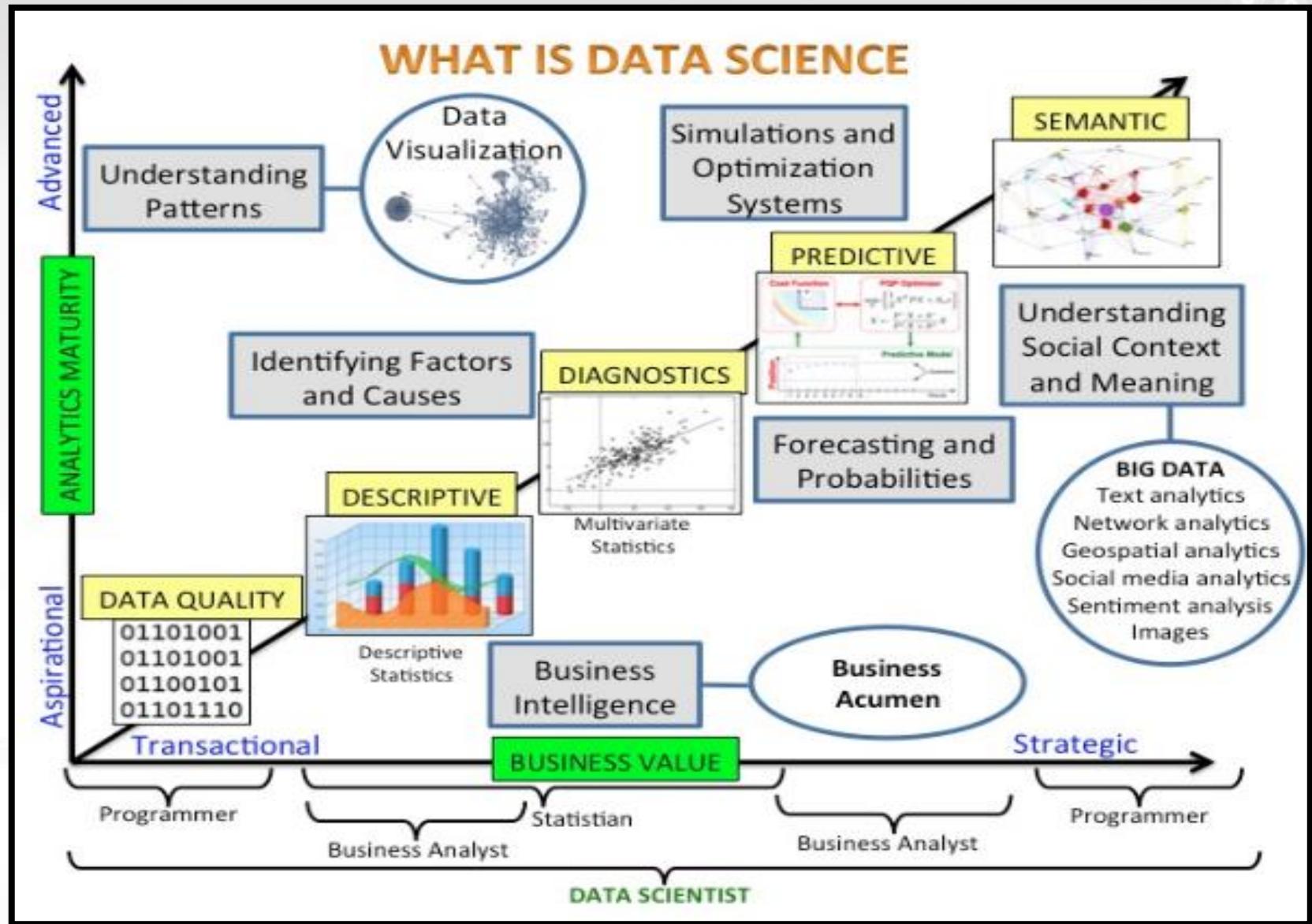


Source: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>, <https://www.kainos.com/my-experience-as-a-data-scientist>

Dimension of Data Science



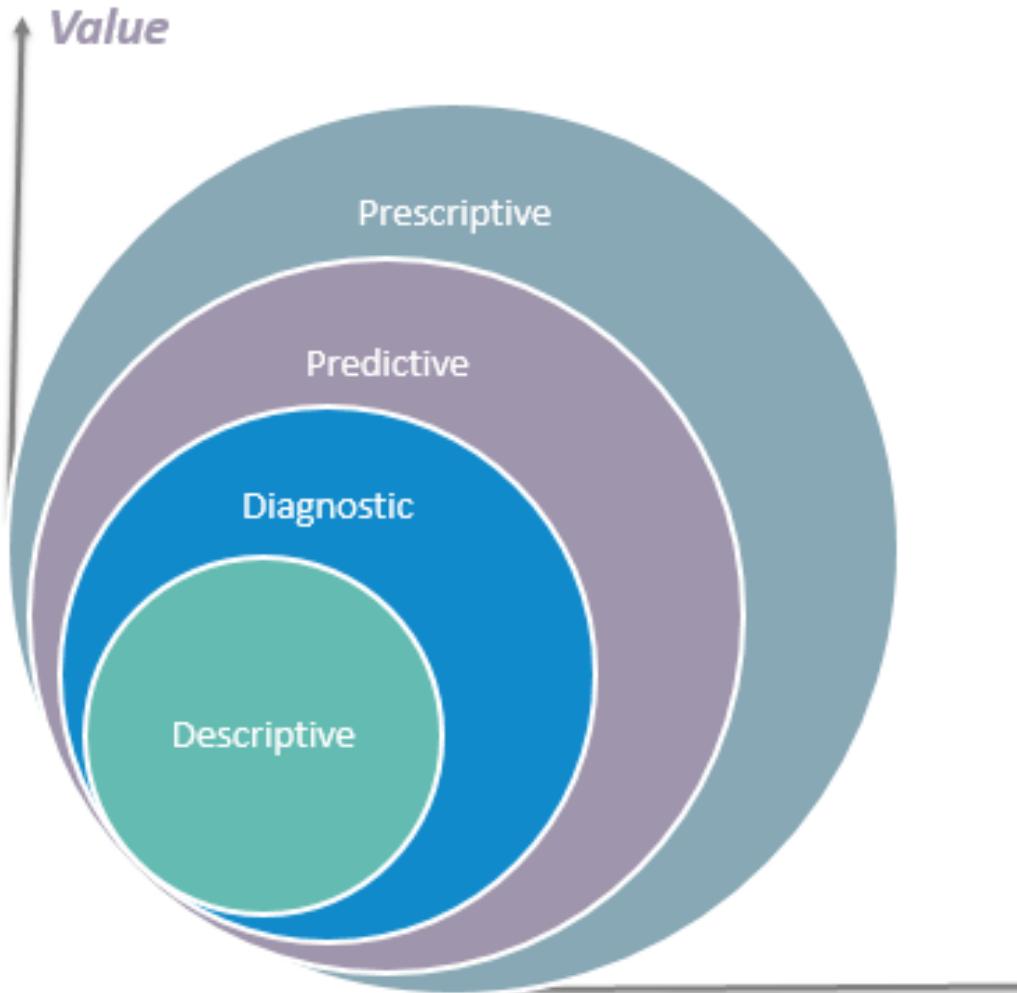
Evolution of Data Science



Source: <https://www.linkedin.com/pulse/business-intelligence-data-science-fuzzy-borders-rubens-zimbres/>

Types of Data Analytics

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

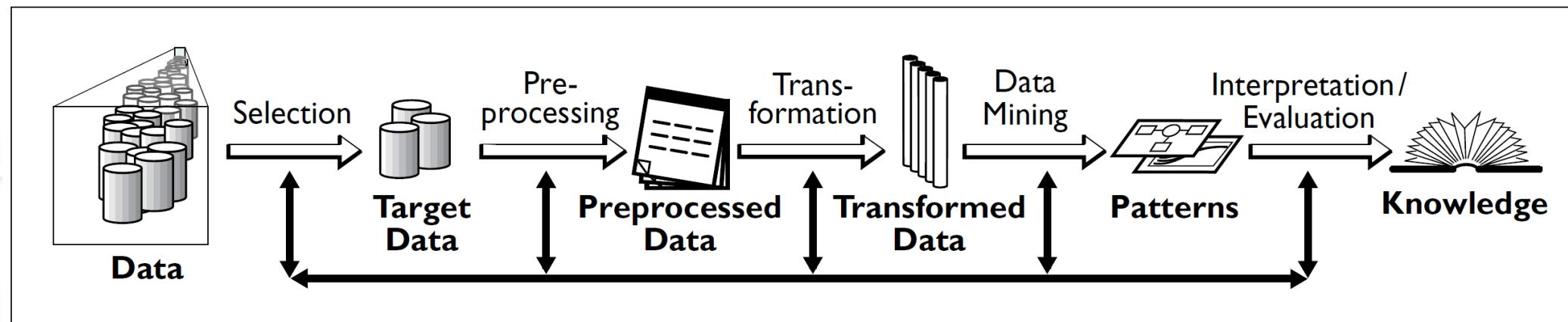
Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

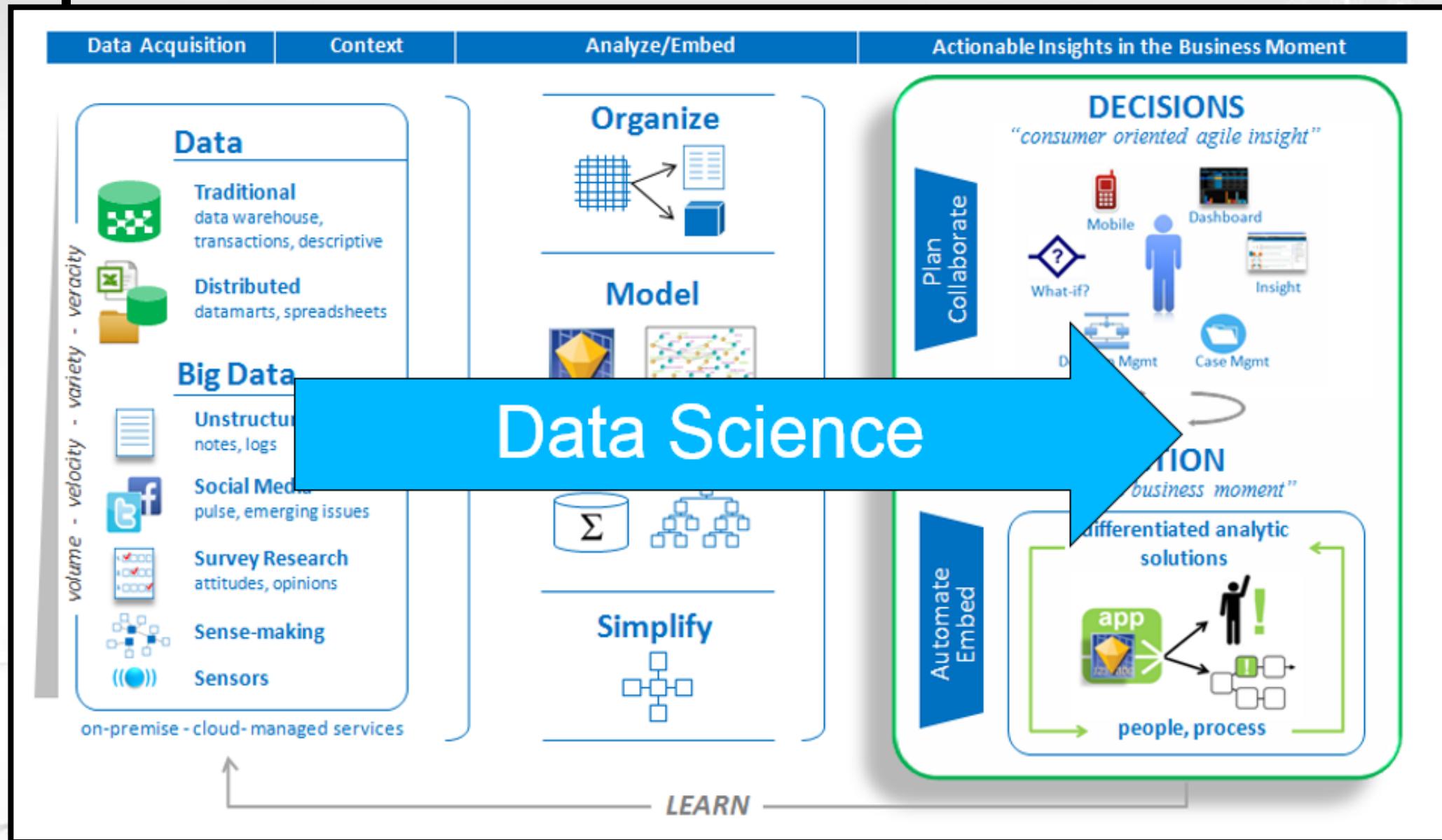
Complexity

Process of Data Analytics

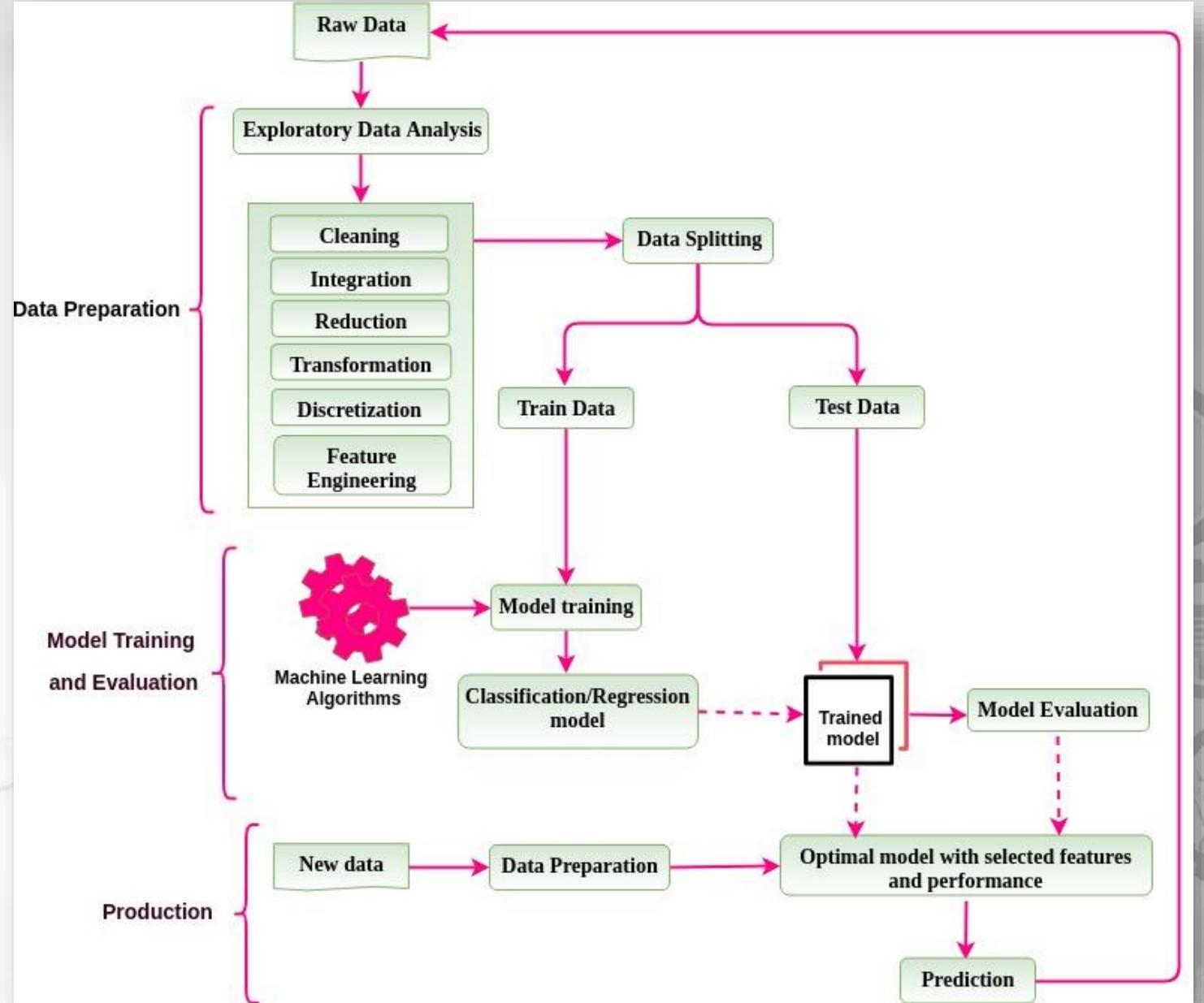
- 資料擷取：從原始資料到決定存放資料庫的過程，一般來說會涉及到資料獲取（data acquisition）、資料爬蟲（data crawler）、資料庫管理（data management）、資料倉儲（data warehouse）等等議題
- 資料前處理：對從資料庫根據規格（API、SQL）取出的資料集，進行資料清理（data cleaning）處理資料中包含的雜訊或錯誤訊息，或是想使用到多個資料集也會在這邊進行整併
- 資料分析：可以分為兩個階段，探索性分析（Exploratory Data Analysis）與資料探勘/機器學習（Data Mining/Machine Learning），可以把探索性分析視為是一種前期的觀察，在經由資料探勘進行近一步地挖掘
- 資料解釋：通常會透過資料視覺化的方式及圖表方式呈現前述的結果，運用一些可能的原因進行解釋，然後把這一整套東西串起來



Pipeline of Data Science



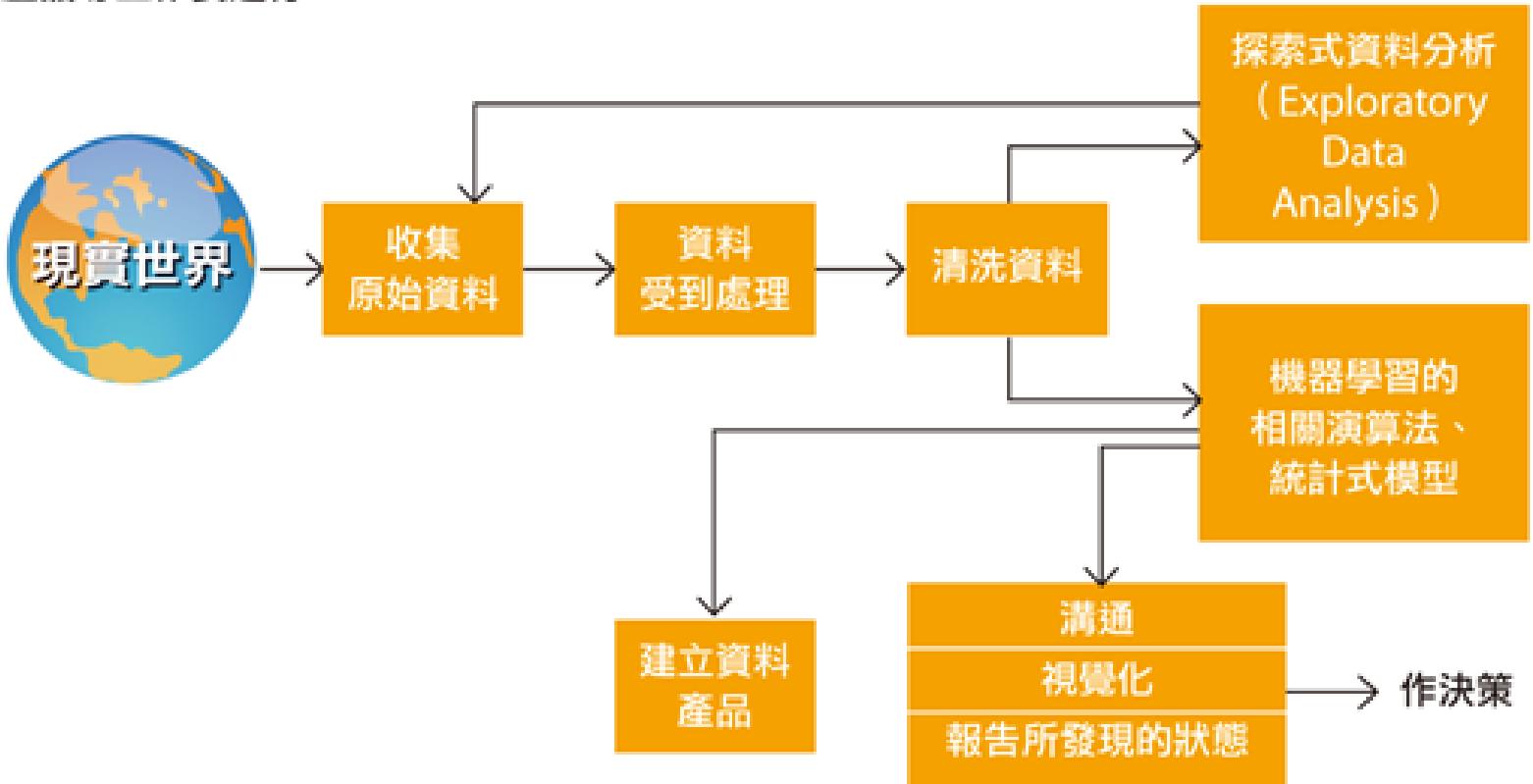
Flow Chart of Data Science



Flow of Data Science in Real World

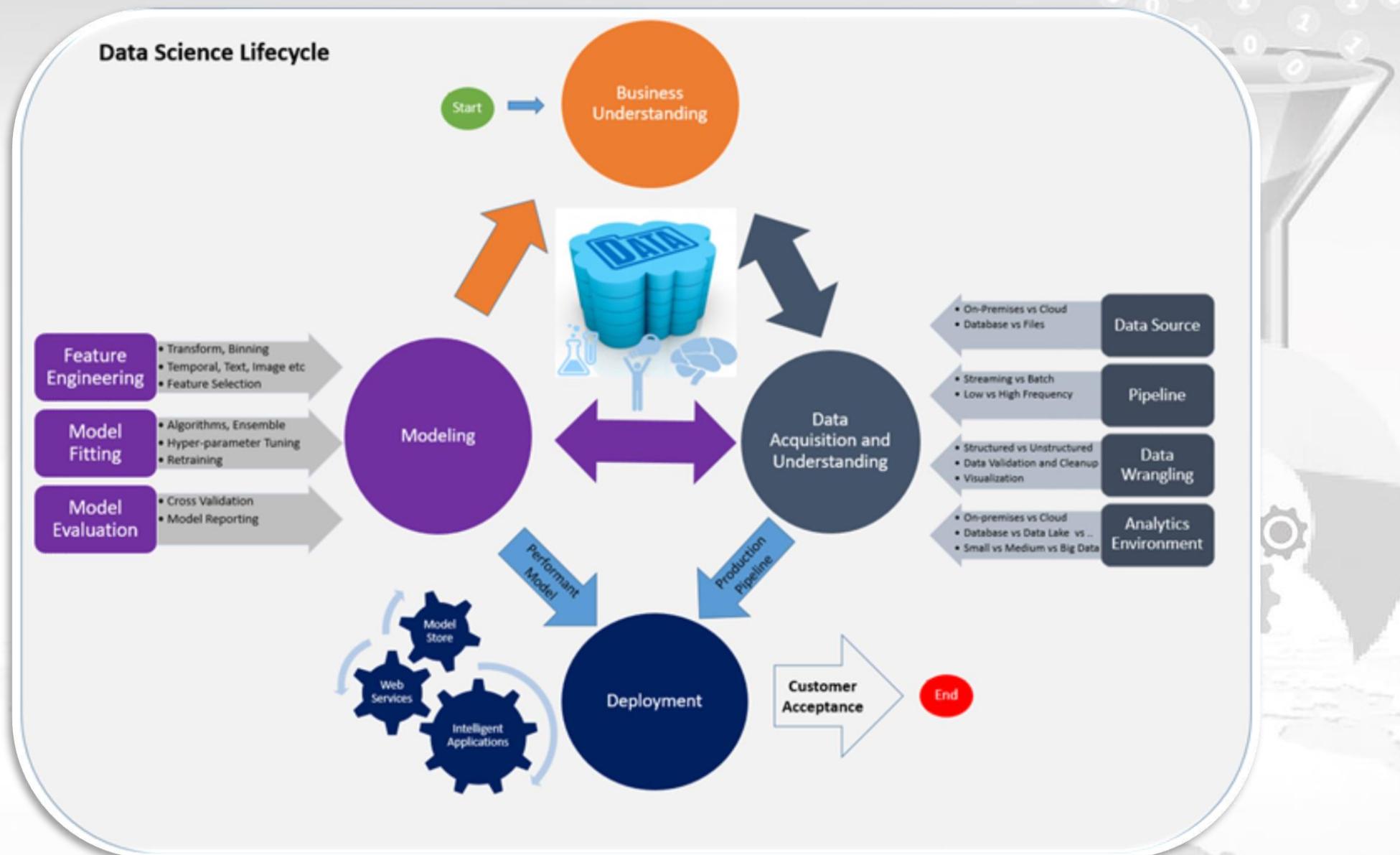
實務的資料科學發展流程

資料科學應用不只是清理資料、建立分析模型，事實上，連清理資料之前的收集資料和處理資料的程序，也包含在內。而資料清理之後，有探索式分析、機器學習、視覺化、建立資料產品等工作要進行。

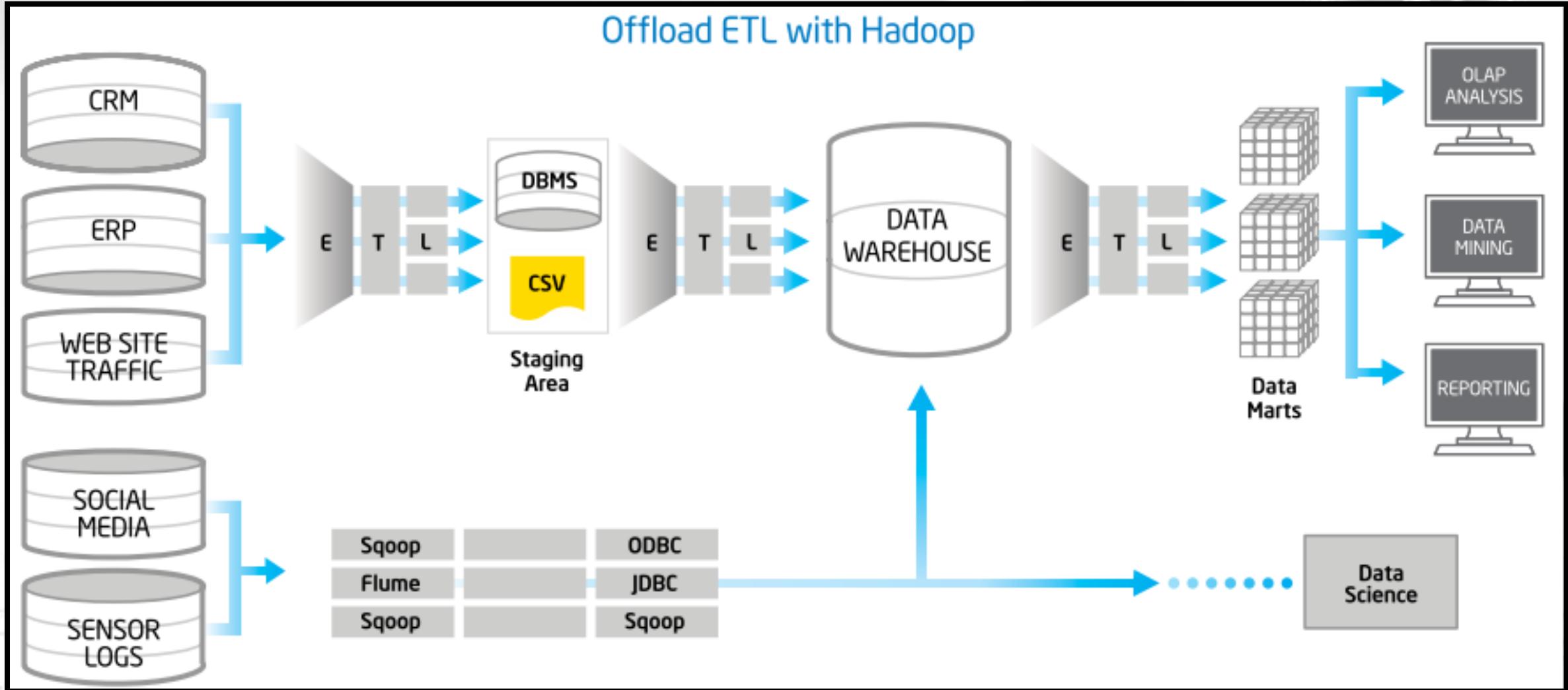


資料來源：Dr. Rachel Schutt 〈Next-Gen Data Scientists Presentation〉，2013年3月

Data Science Lifecycle



Example of Data Analytics Architecture



Source: Intel, Extract, Transform and Load Big Data with Apache Hadoop, <https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf>

Essenes of Data Science

Explanation, Prediction, Estimation / Forecast

資料科學的研究類型

• 解釋 (explanation)

- 因果分析：探索資料間的關聯性並嘗試解釋因果關係
 - 透過資料「歸納」(induce) 出「因變數」與「依變數」之間的關聯
 - 透過資料的「演繹」(deduce) 去預測未來
- 關聯規則分析：用於發現頻繁出現在一起的項目集合，可以幫助企業了解客戶的購物偏好，並制定有效的訂單、促銷、價格策

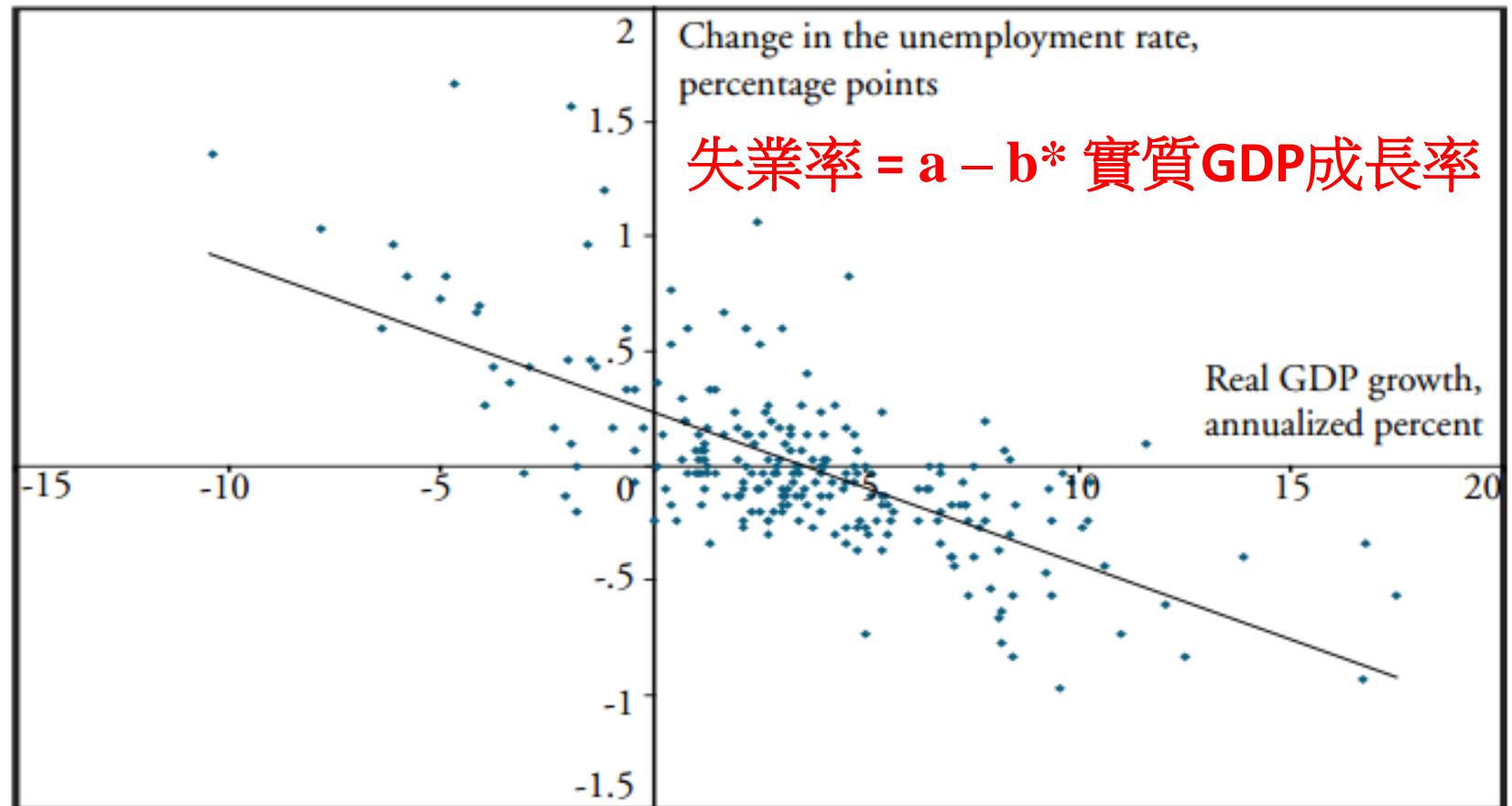
• 預測 (prediction)

- 分類預測：資料類別的辨識與歸類判斷

• 估計 (estimation / forecasting)

- 迴歸分析：鑑往知來的趨勢預測
- 時間序列：實現「千金難買早知道」

解釋 (Explanation)



Note: Data are from the Bureau of Economic Analysis and Bureau of Labor Statistics, from the second quarter of 1948 through the second quarter of 2007.

預測 (Prediction)

資料科學常以「**機器學習**」(machine learning) 的相關演算法進行資料的預測分析，「**機器**」可視為算命師，「**特徵**」(feature) 則是重要的資訊，而算命師不外傳的絕活，就是所謂的「**演算法**」(algorithm)或「**模型**」(model)



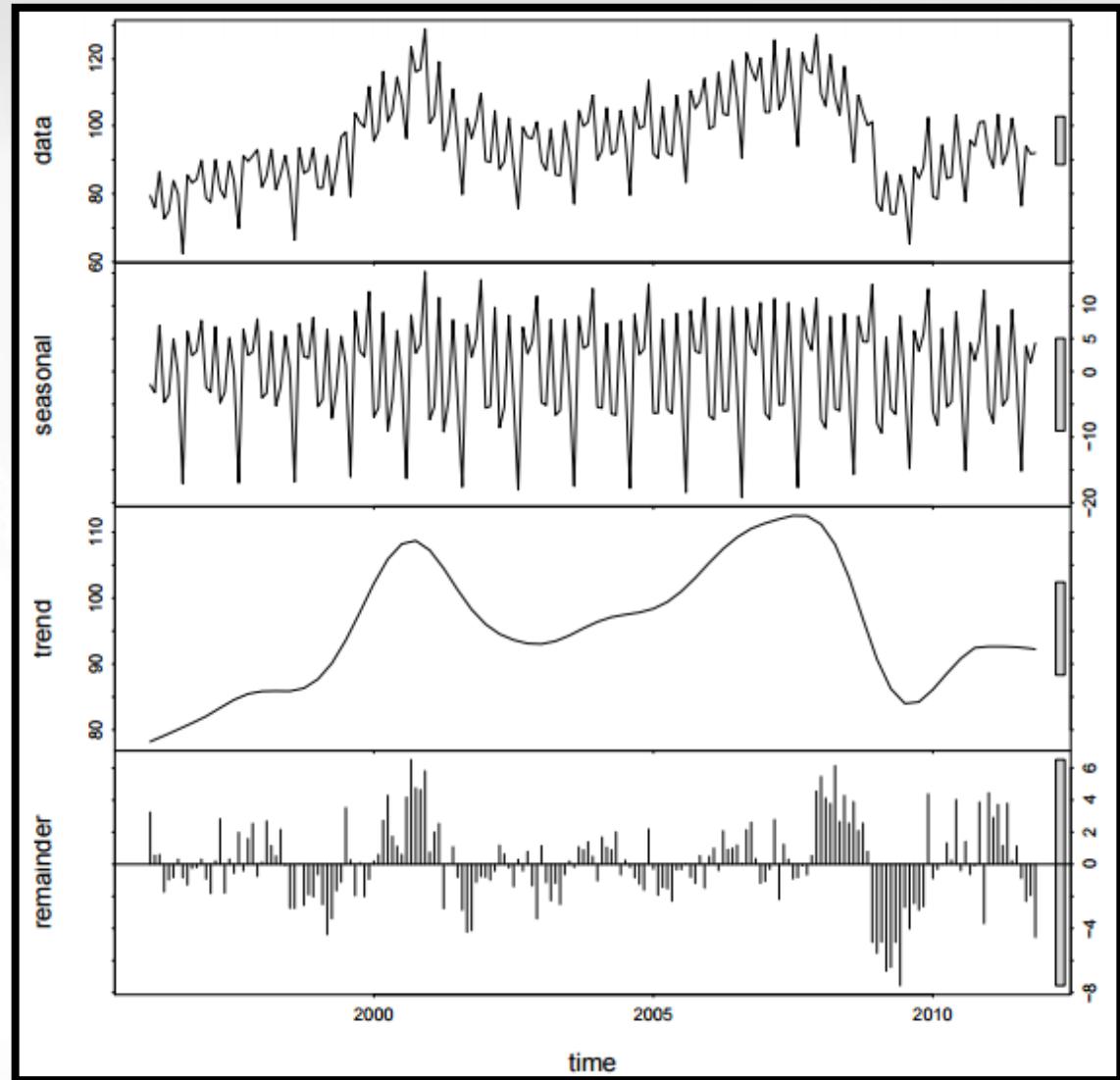
參考資料：[資料科學的三個面向](#)

估計/推測 (Estimation / Forecasting)

- 估計/推測(Estimation / Forecasting) 又稱「趨勢預測」與 預測(Prediction) 都是想要「**預知未來**」
- 預測(Prediction) 想做到的是透過許多不同的資料樣本個體 (**不考慮時間因素**)，來預測下一個新個體的性質
- 但是，推估(Estimation / Forecasting) 則是希望能夠**估算**某觀察的變數**隨時時間變化的發展趨勢**
- 因此，只要有一個變數就能進行「估計推測」，但此變數必需是一個「**時間序列**」(Time Series) 類型資料 - 即隨著連續時間單位變動而改變的變數
 - 例：台積電未來股價的趨勢預測

參考資料：[資料科學的三個面向](#)

趨勢預測 (Forecasting)-範例



參考資料：[資料科學的三個面向](#)

MONDO Logistics Big Data Analytics Research Center

資料科學的前世今生

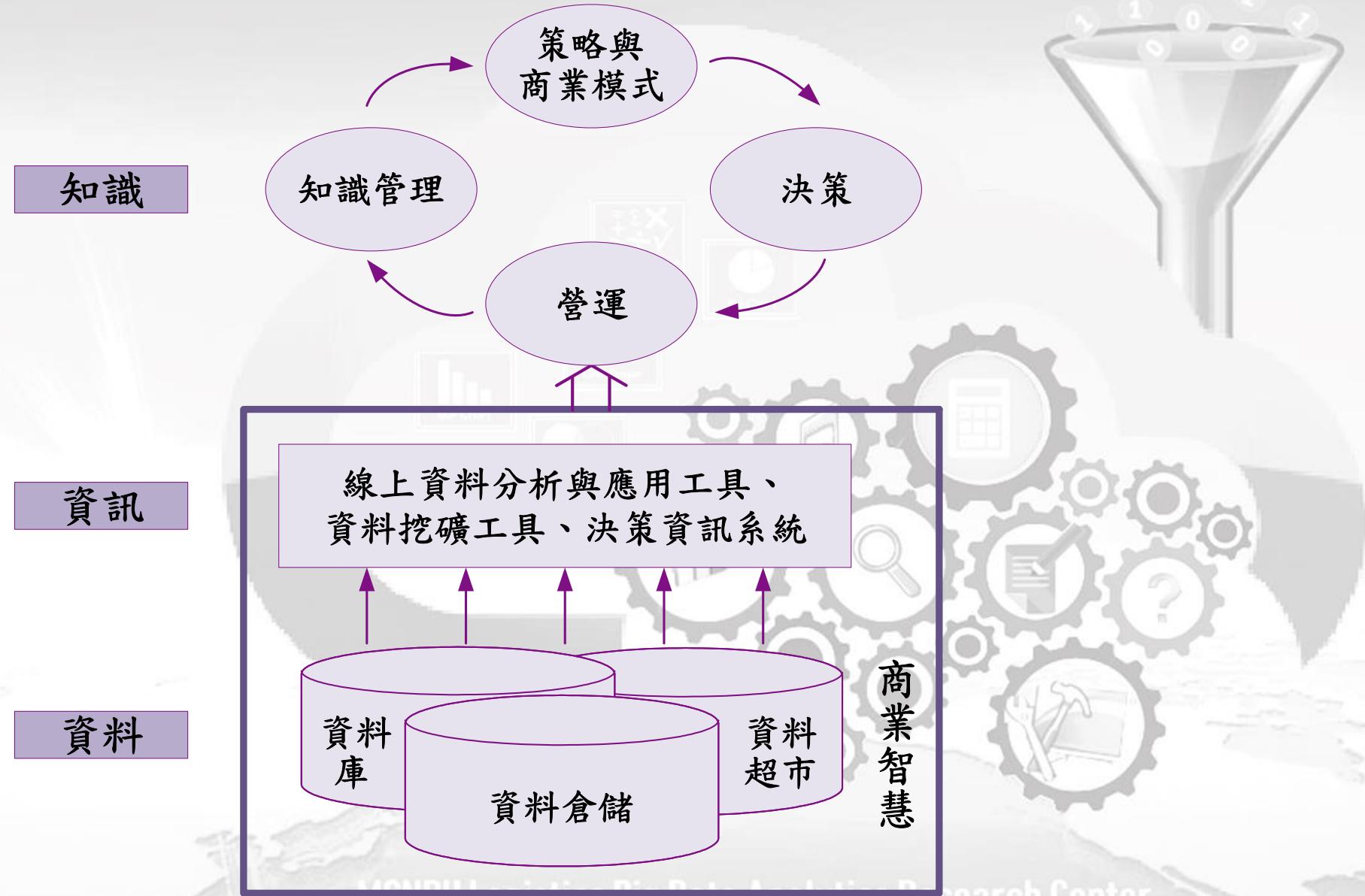
資料探勘 (Data Mining)

Sift the Wheat from the Chaff

資料探勘 (Data Mining)

- 又稱數據挖掘、資料採礦、資料挖掘、資料探礦等
- 如同挖礦的過程從累積如山的資料中，挖掘出如礦石般的**特殊資料樣型**或**規則**，再經由一連串的資料清理、整理與分析的過程，才能獲得**其中最有價值的資訊與知識**
- 資料探勘是從大量自動化蒐集的資料中，找到新奇、過去未知、以及可被解釋的樣型，用以預測未來尚未發生的事件，為不斷循環的資料分析與**決策支援**的過程
- 資料科學的基礎知識

商業智慧、資料探勘與資料科學



資料探勘與數位轉型

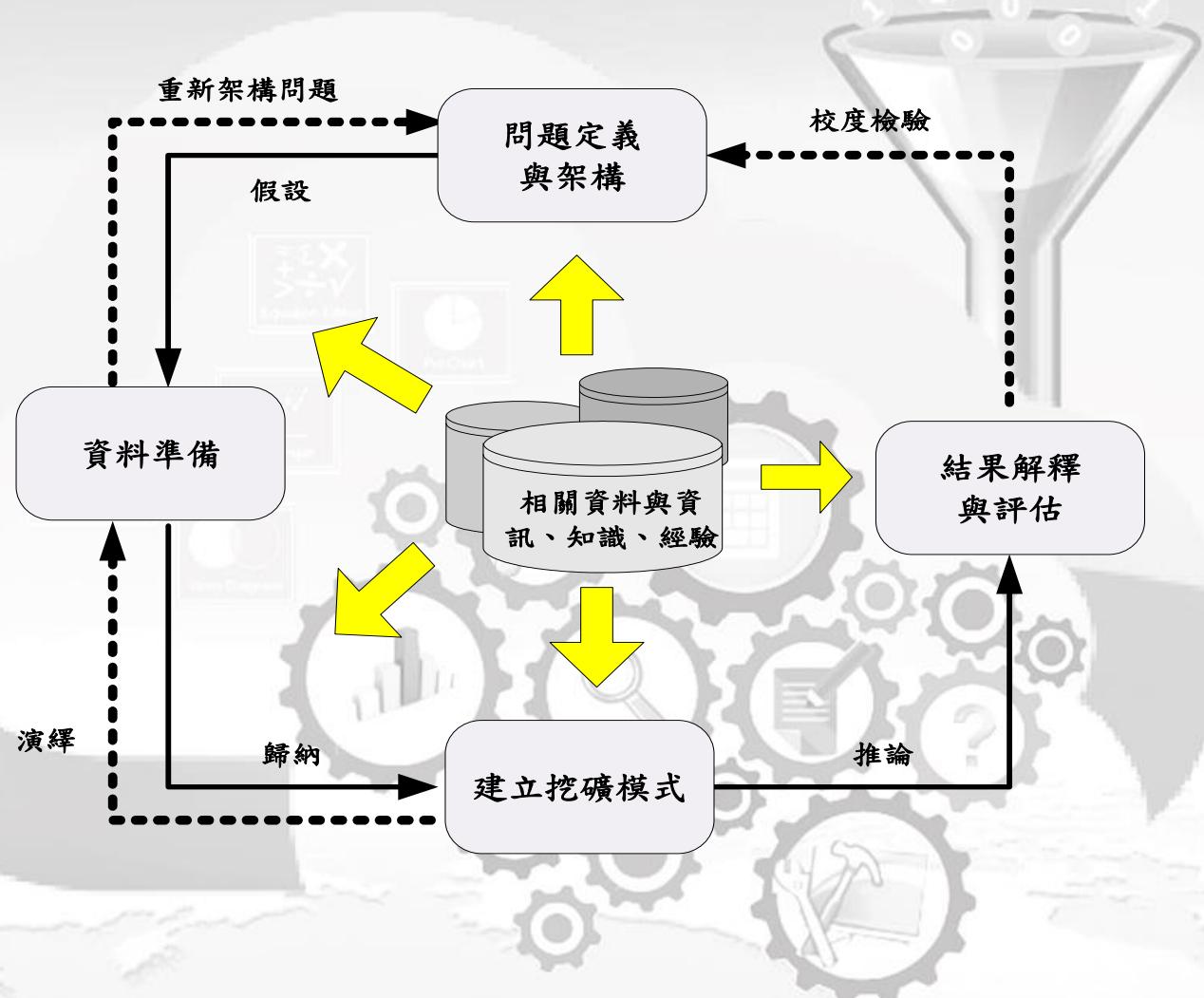
企業決策資訊系統發展的演化過程

	資料	資訊	知識	管理與決策
企業向外延伸與供應鏈管理的層次	供應鏈管理系統與跨公司的溝通與應用軟體	供應鏈的生態與協同	全方位的策略公司定位	管理與決策
企業內部組織整合的層次	企業電子化系統與應用軟體	全企業的溝通和企業整合	全企業的知識管理	企業流程與組織再造
強化工作群組的層次	特殊功能軟體與資料庫系統	資訊整理與工作群組的溝通	工作群組合作與知識分享	流程整合與群體決策
強化個人的層次	資料的創造、存取與使用	資料探勘與資訊萃取	教育訓練與知識累積	流程標準化與專業提升

資料來源：Papows, 1999；簡禎富，2005

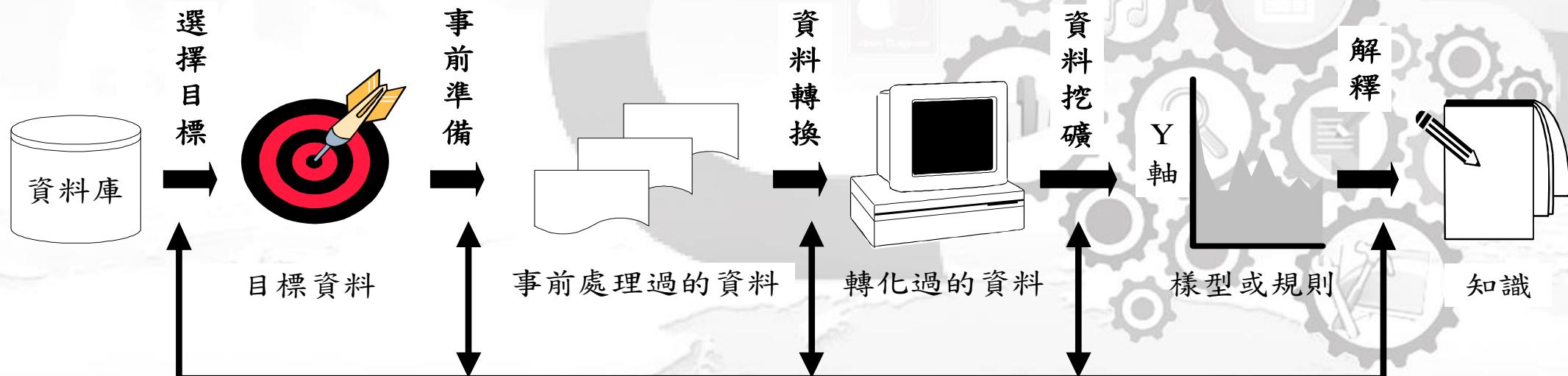
資料探勘實作流程 (Workflow)

- **問題定義與架構**：首先對問題領域與研究目標有清楚的定義與架構
- **資料準備**：整理龐大的資料以利後續分析
- **模式建構**：針對問題本質使用適當工具分析龐大的資料以發現有用的資訊
- **產生方案與評估結果**：根據所得到的資訊擬定適當的行動方案，協助決策；評估挖礦成效，以作為下一個改善循環之依據



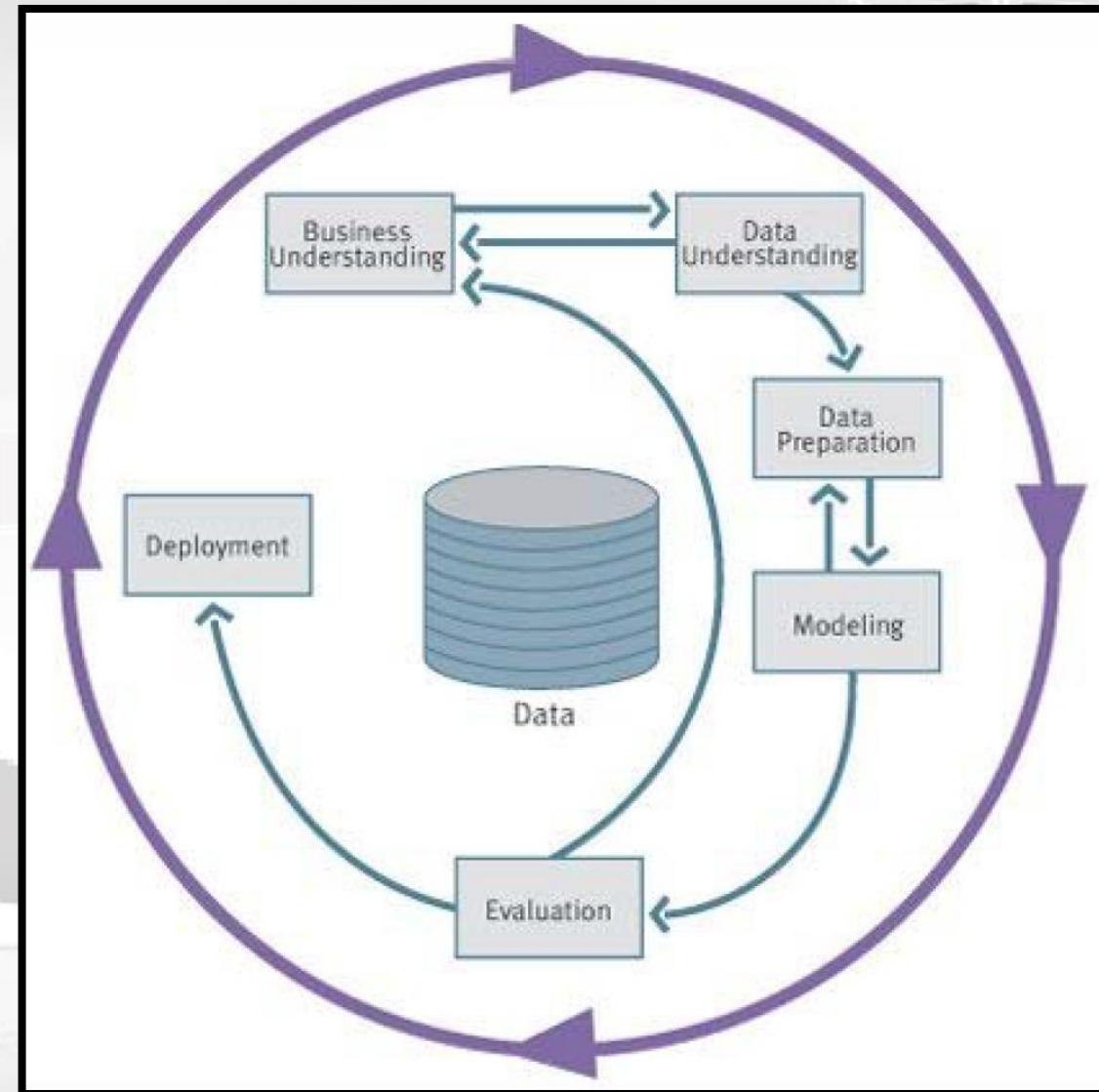
資料庫知識發現 (Knowledge Discovery in Database; KDD)

- 資料探勘為資料庫知識發現過程中的核心步驟，也是**資料科學**的基礎
- 過程：強調從大量資料中，以自動或是半自動的方式**反覆不斷地探索和分析**資料，藉由**資料探勘**技術有效率地擷取出事先未知的隱含資訊與知識，提供決策者參考
- KDD是從資料中建立確定有效的、新奇的、潛在有用的、以及易懂形式的樣型與資訊之過程，且此過程並不顯而易見

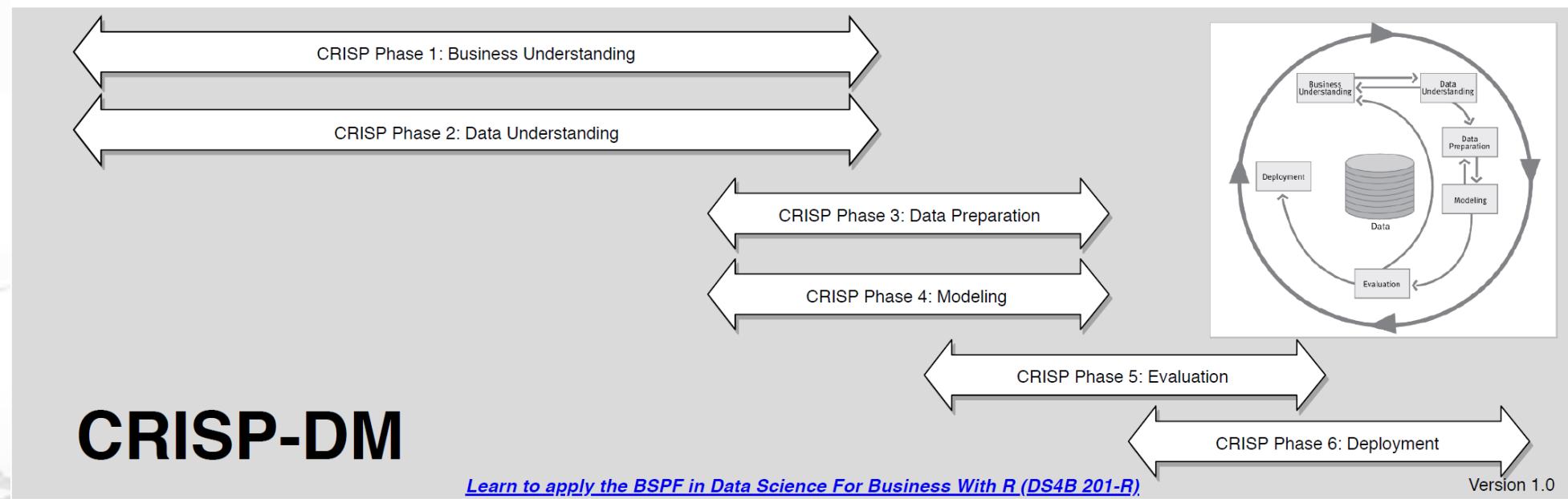
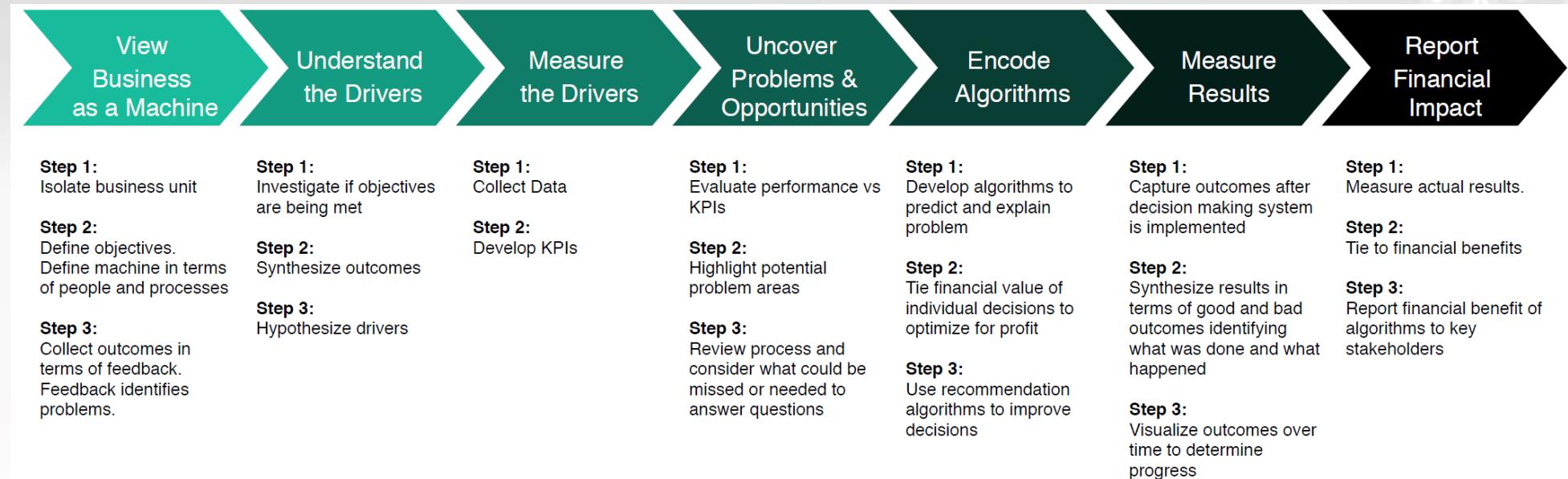


資料探勘建模標準 - CRISP-DM週期

1. 理解商業問題
2. 理解資料問題
3. 資料前置處理
4. 建立模型
5. 評估模型
6. 應用模型

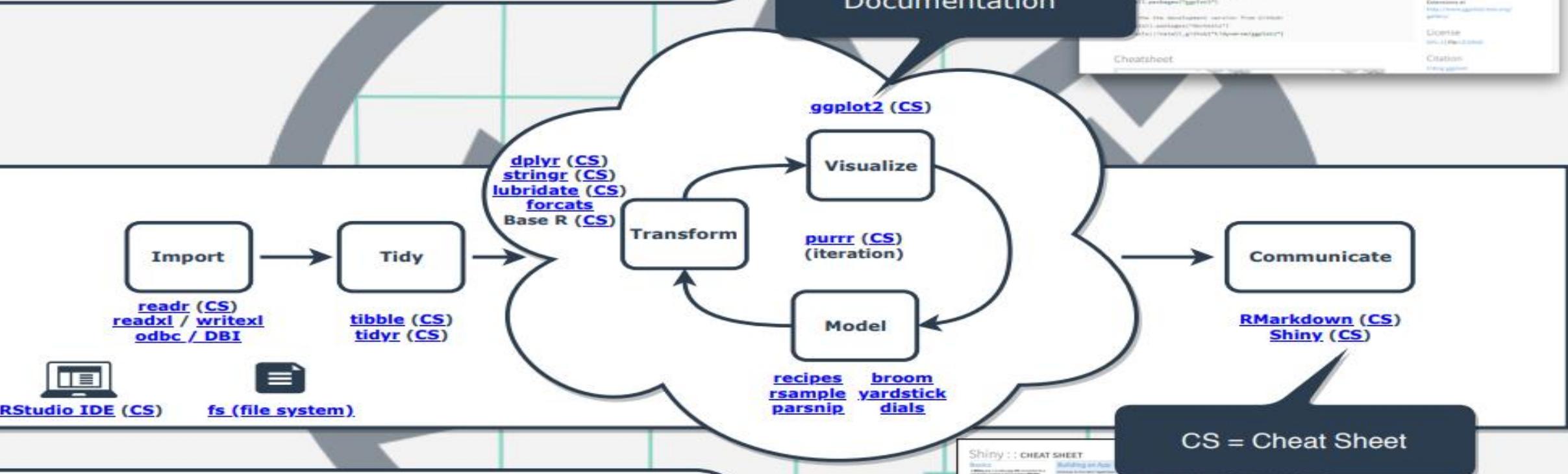


CRISP於商業實務的流程



Data Science with R Workflow

The Data Science With R Workflow is available in the book: [R For Data Science](#). If you want to learn R and this workflow for business, take the [R For Business Analysis \(DS4B 101-R\) course](#) through Business Science University.



Important Resources

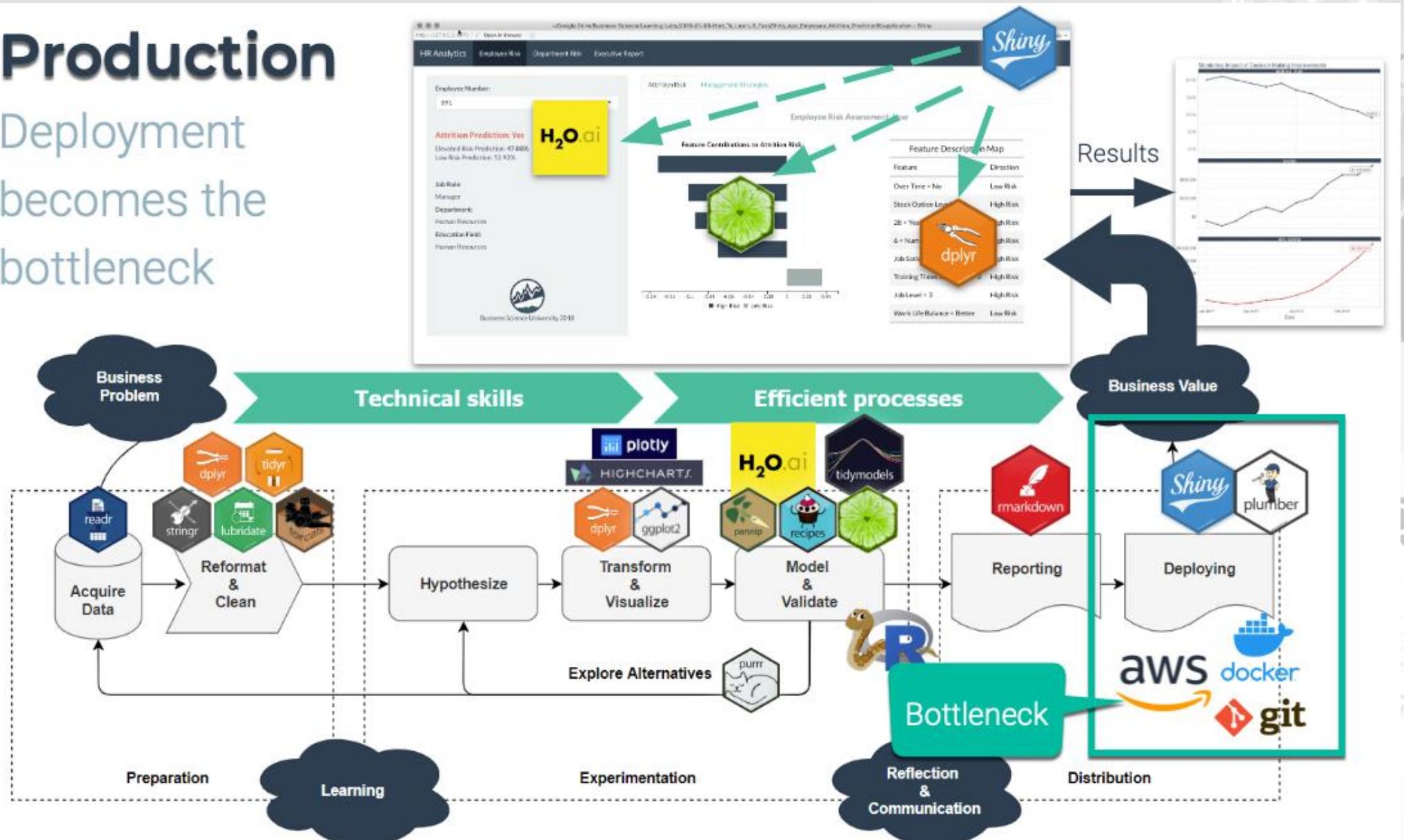
- R For Data Science Book: <http://r4ds.had.co.nz/>
- Rmarkdown Book: <https://bookdown.org/yihui/rmarkdown/>
- Data Visualization Book: <https://rkbacoff.github.io/datavis/>
- More Cheatsheets: <https://www.rstudio.com/resources/cheatsheets/>
- tidyverse packages: <https://www.tidyverse.org/>
- Connecting to databases: <https://db.rstudio.com/>
- RMarkdown website: <https://rmarkdown.rstudio.com/>
- Shiny web applications website: <http://shiny.rstudio.com/>
- Jenny Bryan's purrr tutorial: <https://jennybryan.org/>



CRISP implemented with R & AWS

Production

Deployment becomes the bottleneck



資料探勘的類型

- **描述性 (Descriptive)**

- 從蒐集的資料中用更易瞭解的方式來描述一個隱藏在大量資料背後複雜的現象或狀態，藉由分析資料之間的關聯，找到可能的相關(correlation)、趨勢(trend)、樣型(pattern)或規則(rule)
- 例：根據銷售交易紀錄找出產品間的關聯以決定促銷的產品組合

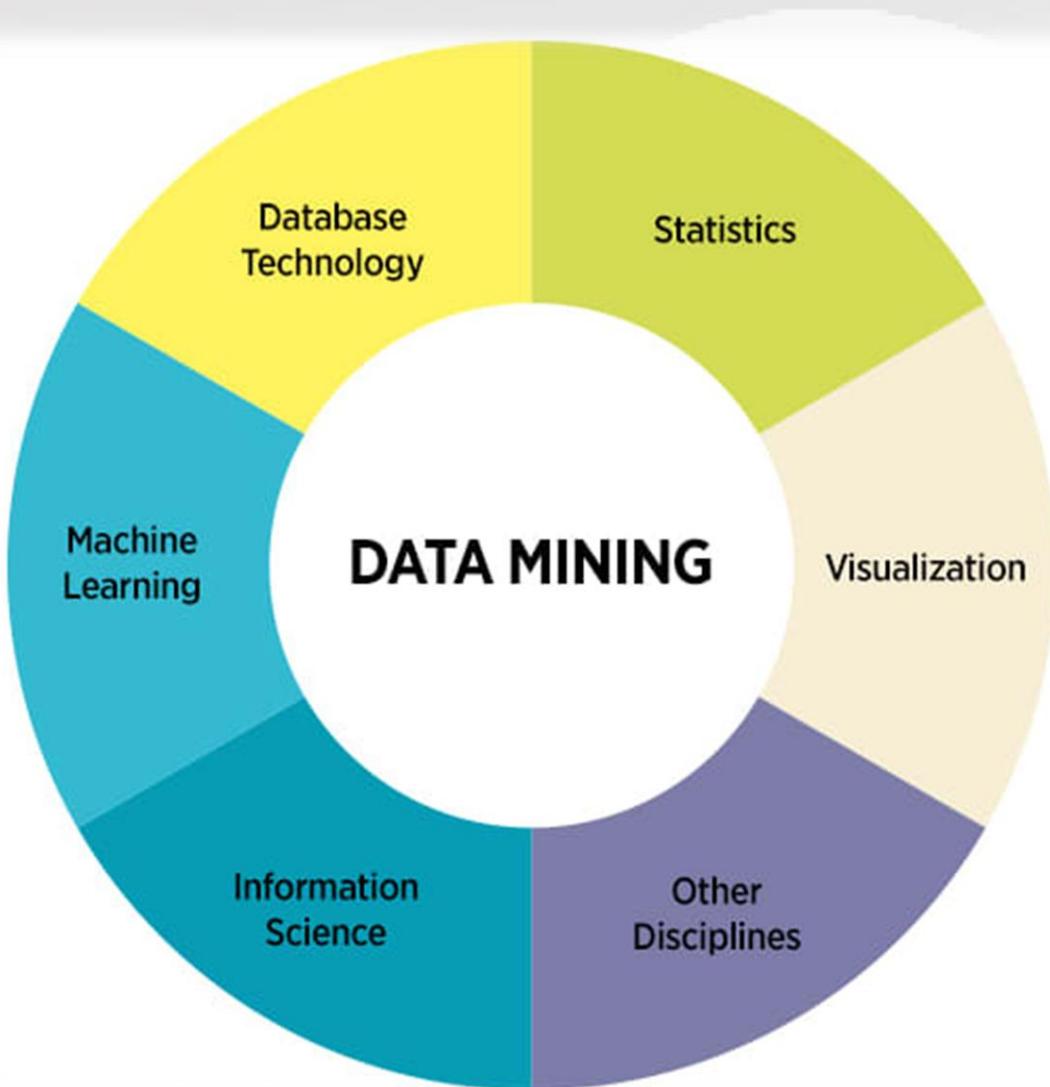
- **預測性 (Predictive)**

- 基於歷史資料的關聯或規律建立模型，作為預測或判別未來的結果
- 例：預估產品未來一季的銷售量、判斷某信用卡客戶是否會有違約風險等

資料探勘的方法

- 監督式(supervised)資料探勘
 - 由上而下(top-down)的方法
 - 利用現有模型建立資料來描述某一特定目標變數與其他變數間的關聯性，例如目標變數可能有信用良好與信用不佳兩種類別
 - 如：分類及預測
- 非監督式(unsupervised)資料探勘
 - 由下而上(bottom-up)的方法
 - 無特別標注特定目標變數，而是嘗試找出所有變數中是否有某種關係存在，發現樣型後再由使用者決定結果的重要與否
 - 如：分群與關聯規則

資料探勘的理論基礎



資料探勘與大數據分析之異同

資料探勘(Data·Mining)·vs·大數據分析(Big·Data·Analytics)之異同		
	資料探勘(Data·Mining)	大數據分析(Big·Data·Analytics)
主要概念及應用	資料庫知識發現(Knowledge-Discovery in Database, KDD)中的一個步驟。	資料量規模巨大到無法透過人工，在合理時間內達到擷取、管理、處理、並整理成為人類所能解讀的資訊。
資料集大小	獨立的小型資料集(TB 以下及 TB~PB 等級)	大型資料集(PB 等級以上)
資料類型(來源)	結構化資料(資料庫、資料倉儲)	半結構、非結構化資料(原生資料(raw data)、文字、聲音、影像、視訊及串流資料(Stream Data))
可預測性	設計模型、獲取預期的結果	探索資料、發現新模型、找出資料關聯性、進而達到精準預測
時間複雜度	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
使用工具	1.統計分析軟體(例如：SPSS、SAS、R) 2.資料探勘軟體(例如：PolyAnalyst、Orange、Weka 等) 3.數學軟體(例·如：MATLAB 等)	無特定套裝軟體工具，但有大量開源碼(Open·Source)供自行開發整合系統 1.雲端運算平台 2.分散式運算(Hadoop、Spark、Flink) 3.平行運算
資料建模	關聯式資料庫及資料倉儲	原生型資料(XML、JSON)·NoSQL 資料庫
分析技術	統計分析、多變量分析、OLAP、線性規劃、機器學習等	統計分析、多變量分析、深度學習(Deep Learning)、增強學習(Reinforcement Learning)

Source: 本課程彙整

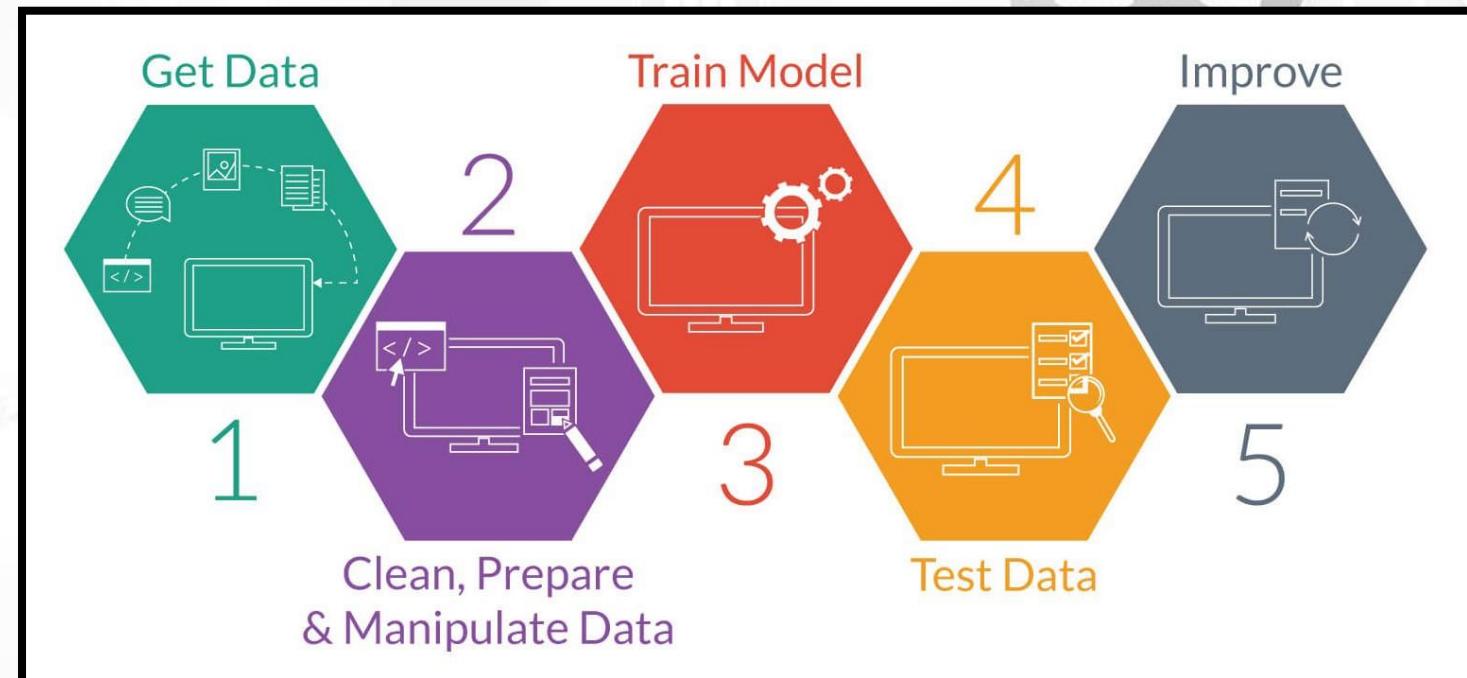
Center

Data Analytics Methods

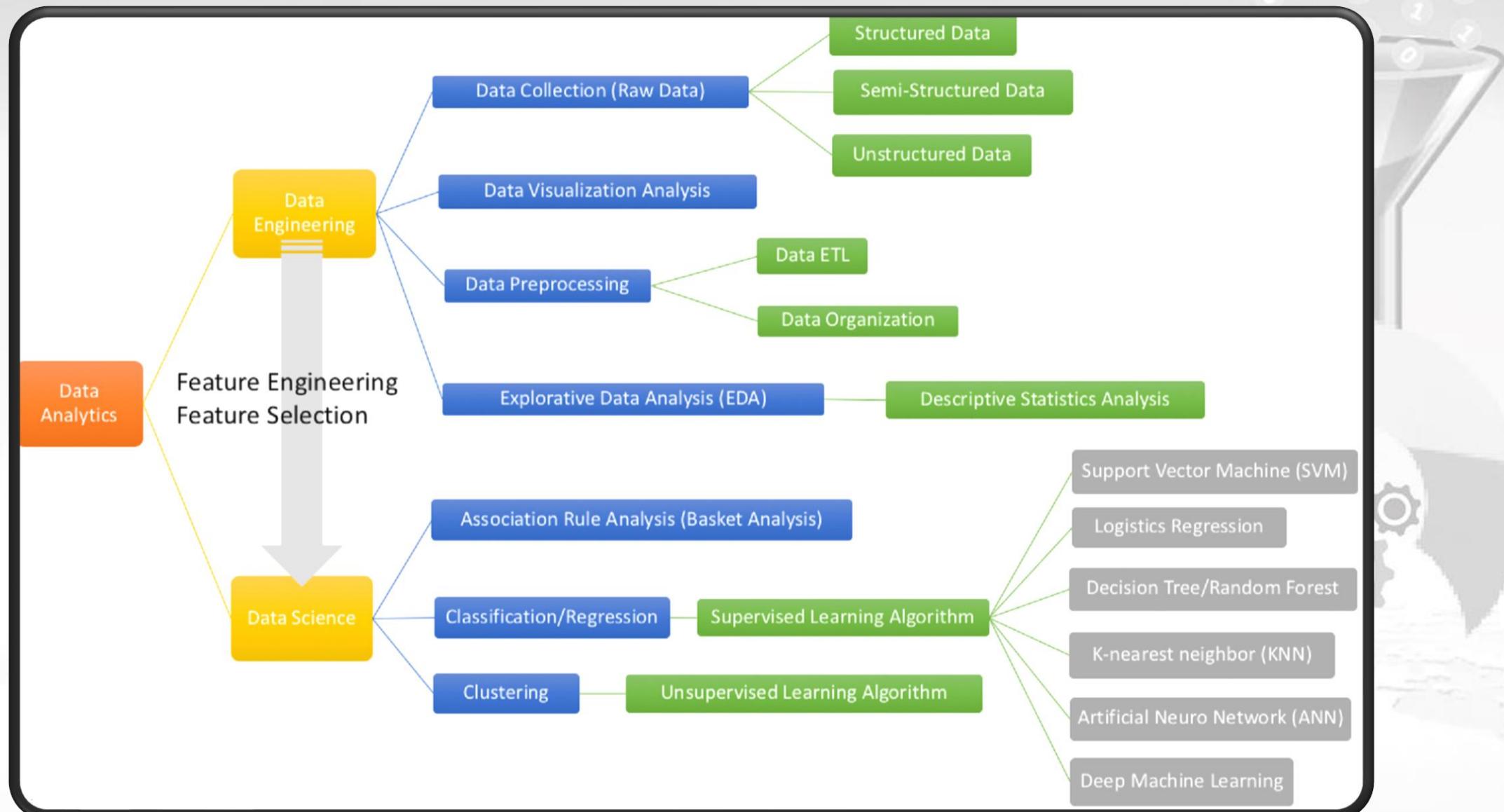
The first step is as good as half over

資料探勘模式

- 資料探勘模式會因欲解決問題的類型與探勘目的的不同而異，且通常不會只使用單一工具
- 不同模型的適配 (Model Fitting) 與否取決於資料的型態與種類、資料與模型應用的假設、資料集合的大小、資料雜訊與資料品質、分析結果的應用目的與方式
- 常用的方法包括統計分析，如迴歸分析、多變量分析，以及關聯規則、決策樹、類神經網路、群聚分析、貝式分類法、約略集合理論、時間序列分析等



Framework of Data Analytics Model



資料科學方法類型

- 方法分類
 - Classification
 - KNN / Decision Tree / SVM / Naïve Bayes / ANN
 - Random Forest / Deep Learning: CNN
 - Association Rule Analysis
 - Apriori(先驗演算法)/FP-Growth(頻繁模式增長法)
 - Clustering Analysis
 - K-Means
 - Regression
 - Forecast
 - Time Series
- 分析的類型
 - 推薦系統/精準運籌/商業智慧/決策支援系統
 - 資料視覺化/系統數位儀表板

Tools & Techniques for DS

- Tools

- Application Suite: SPSS Modeler/SAS/Matlab ...
- Coding Language: C++/JAVA/R/Python/Scala/Julia ...

- Techniques

- Statistics / Multivariate Analysis
- Linear Programming / Calculus
- Distributed Processing (Computing / Data Storage)
- Cloud Computing / Parallel Computing (Data Engineering)
- Machine Learning / Deep Learning (Artificial Intelligence)

Sharp Tools make Good Work

Applications of Data Science

資料科學的應用(1/2)

- 金融服務業
 - 客戶貢獻度分析、信用評分、風險評估、客戶區隔、交叉行銷等
 - 金融科技(FinTech)
- 保險業
 - 顧客貢獻度分析、信用評分、風險評估、客戶區隔、交叉行銷、客戶流失分析和詐欺偵測等
 - 保險科技(InsurTech)
- 電信業
 - 顧客貢獻度分析、信用評分、客戶區隔、交叉行銷、客戶流失分析、銷售預測和詐欺偵測等

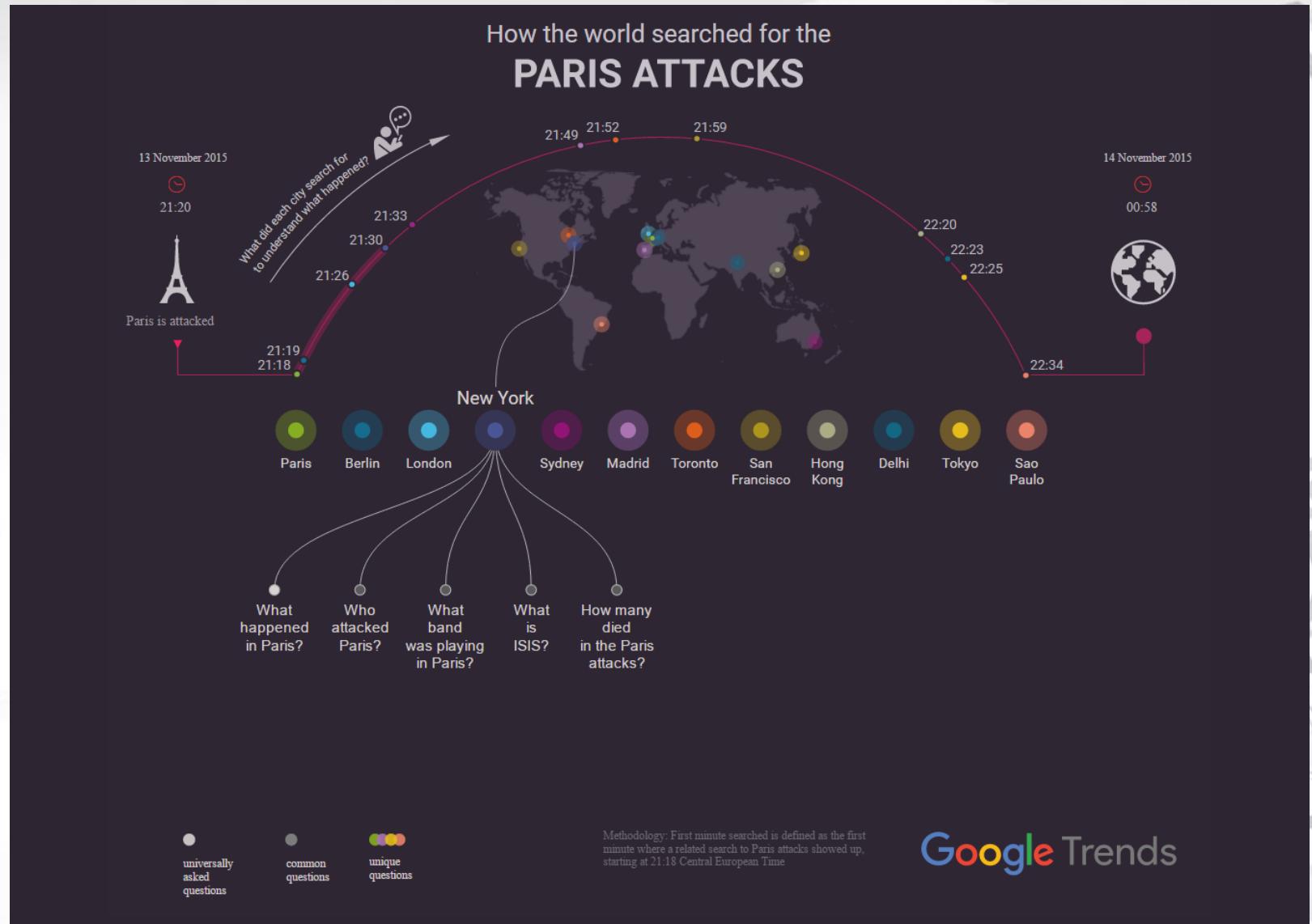
資料科學的應用(2/2)

- 製造業
 - 客戶貢獻度分析、品質管制、行銷績效分析、生產分析和存貨分析等
 - 工業4.0(Industry 4.0)：自動化生產+物聯網(Internet of Thing)+人工智慧(AI)+雲端運算
- 零售業
 - 客戶忠誠度、客戶區隔、購物籃分析、定價分析、交叉行銷和銷售預測等
- 物流業
 - 智慧倉儲、智慧機隊管理
- 犯罪偵防、生物科技、醫療保健、航太空業、環境保護、法律等

Use Case

- 亞馬遜網站透過顧客下單的資料分析，推薦顧客其他他們可能也感興趣的商品
- 美國目標百貨向會員寄送量身訂做的折價卷
- 優比速公司（UPS）利用數據分析指引送貨路線避開左轉等待的時間省下一大筆燃油費順便節能減碳
- NETFLIX利用大數據分析顧客品味，不僅用於推薦影片，甚至更進一步地以分析出的黃金組合開拍叫好又叫座的兩季影集《紙牌屋》（*House of Cards*）

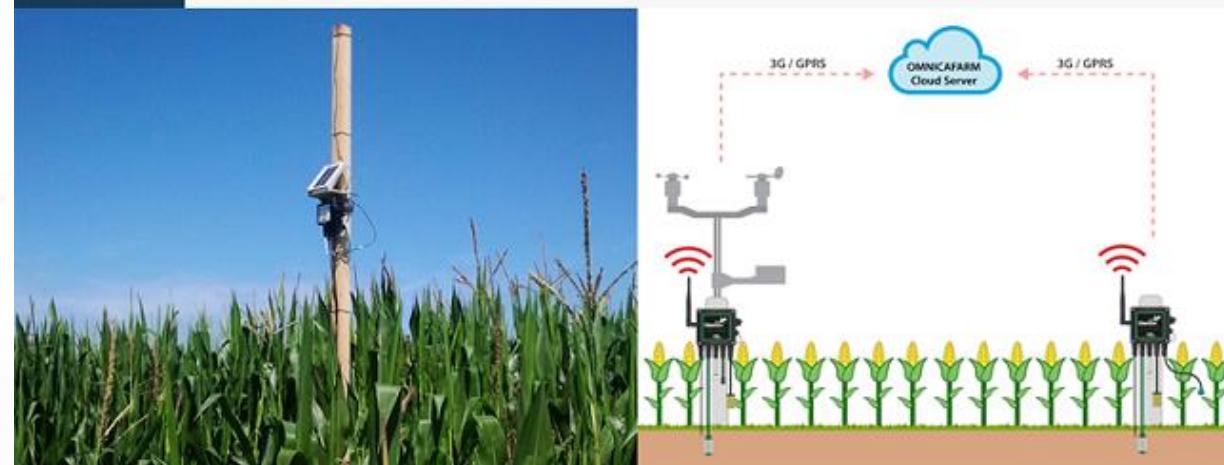
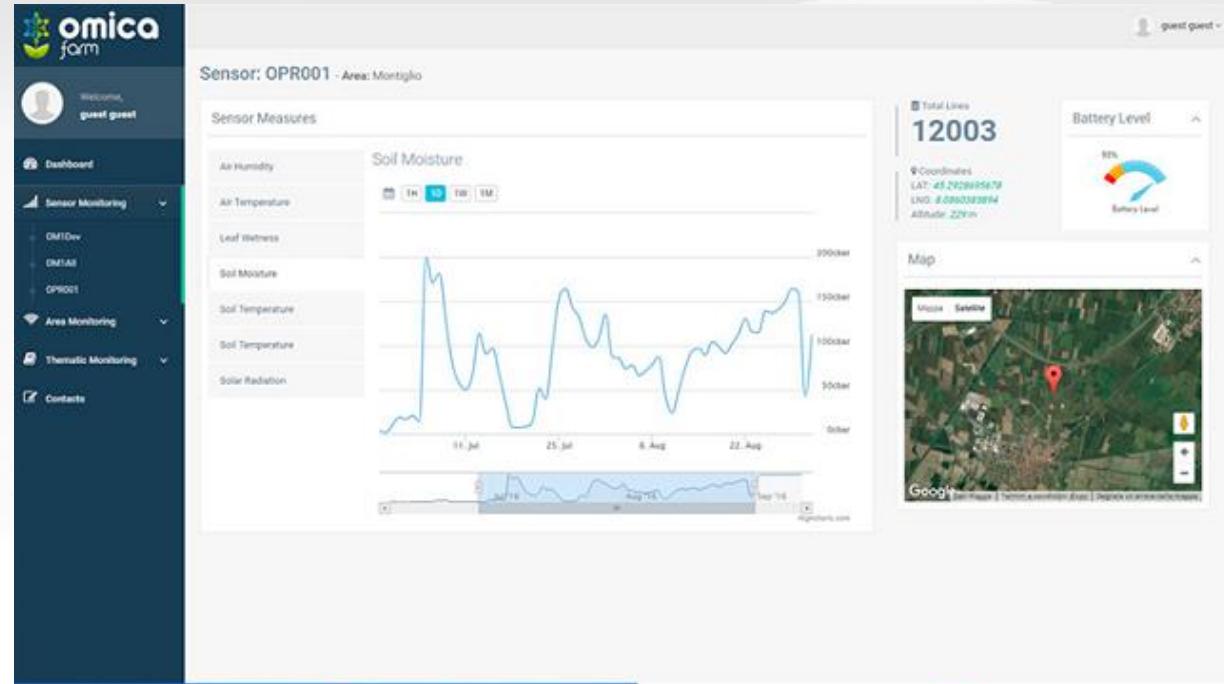
網路新媒體應用：「巴黎恐攻」新聞傳播分析



網路新媒體應用：太陽花學運輿論分析

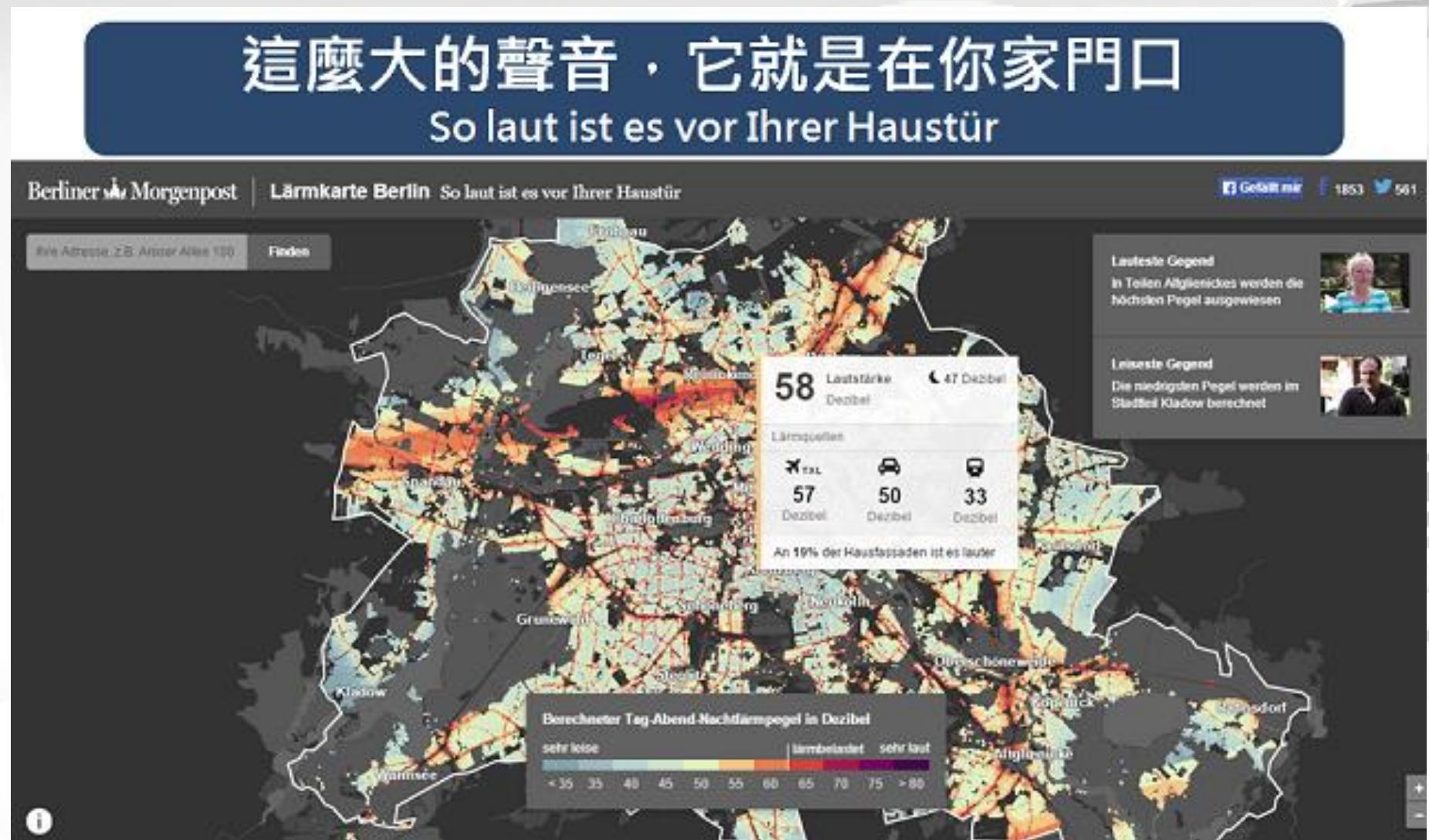


智慧農業：義大利微氣候預測糧食危機

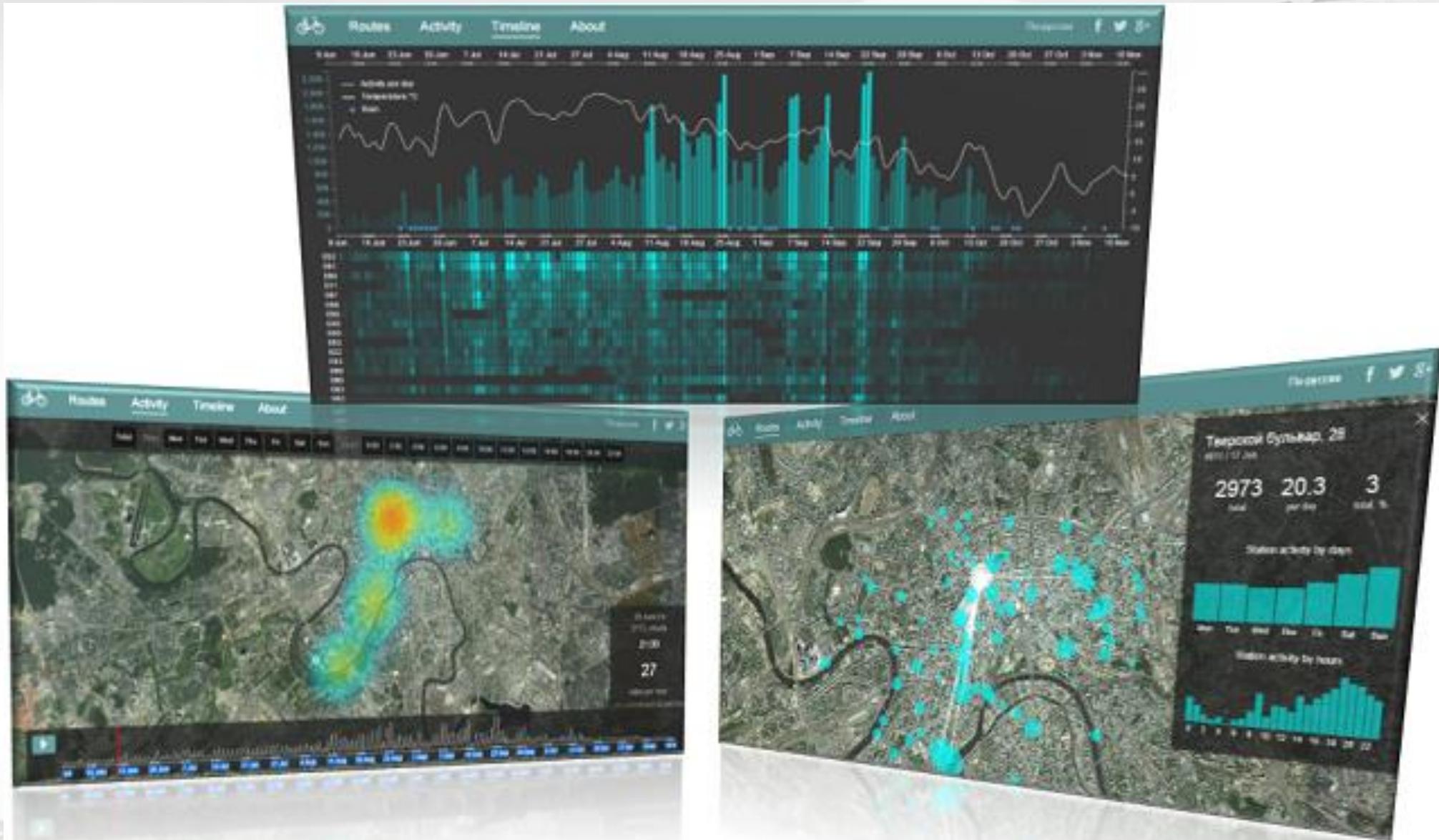


MONDO Logistics big Data Analytics Research Center

智慧環保：德國柏林噪音監控



智慧城市：莫斯科自行車租借



資料科學的前世今生

Meet the Era of Big Data & AI

Sift the Wheat from the Chaff

MCNDU Logistics Big Data Analytics Research Center

Topics

- Big Data in Data Science
 - Evolution of Big Data
 - Origin / Definition / Applications
 - Data Mining in Big Data Analytics
 - Process of Big Data Analytics
 - Architecture of Big Data Analytics

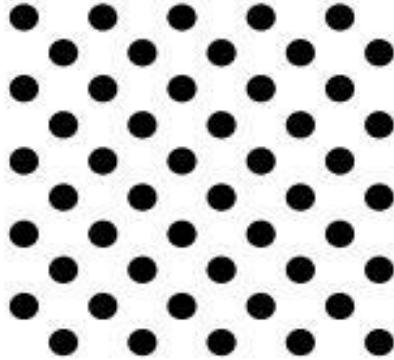
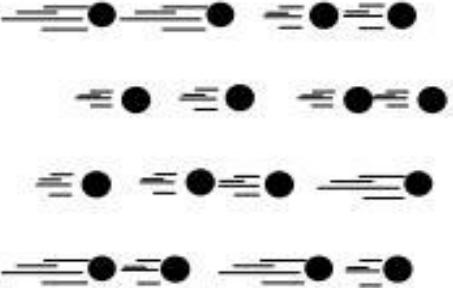
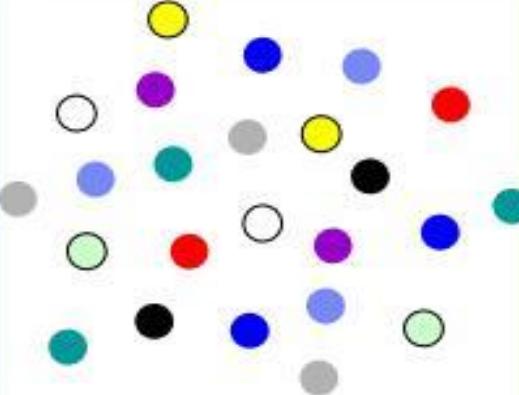
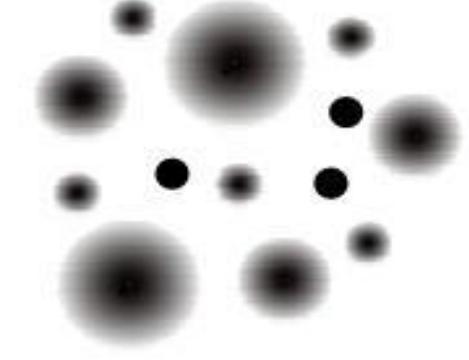


Big Data in Data Science

- 巨量資料分析是**資料科學**研究領域下的**子領域**
- 美國國家標準與技術研究院（NIST）大數據的定義
 - 巨量資料研究計畫分為**巨量資料**（Big Data）和**資料科學**（Data Science）兩個部分
- **巨量資料**
 - 包含**4V**：大量性、迅速性、多樣性及真實性的資料集
 - 需要一個彈性的資訊平台架構，才能有效儲存、處理與分析。
- **資料科學**
 - 強調完整的**資料生命週期過程**，將原始資料轉換為可運用的知識，也就是所謂「**資料驅動**」（Data Driven）的概念

資料來源：獨立評論, <https://opinion.cw.com.tw/blog/profile/52/article/4901>

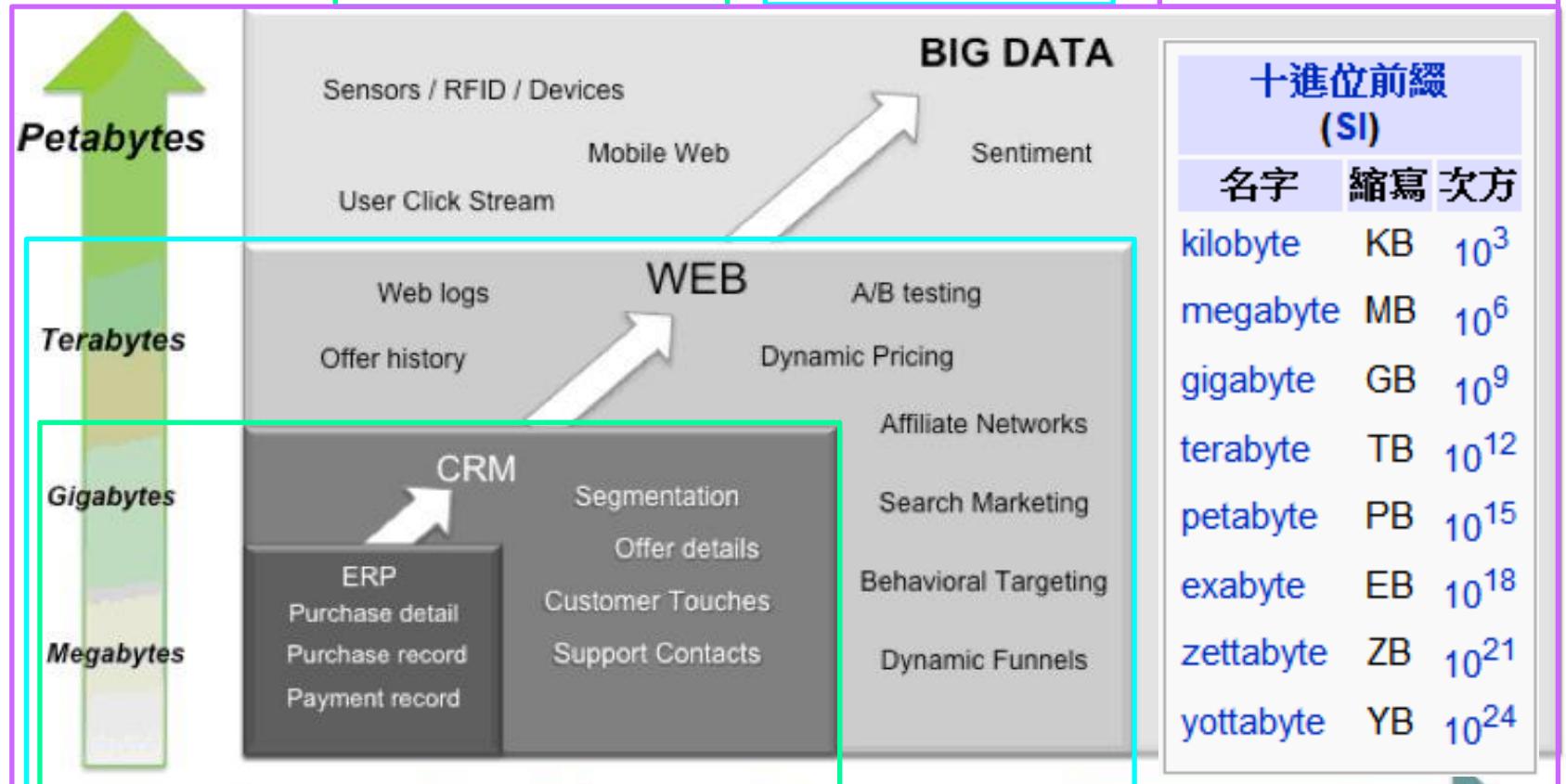
4V of Big Data

Volume	Velocity	Variety	Veracity*
			
Data at Rest <p>Terabytes to exabytes of existing data to process</p>	Data in Motion <p>Streaming data, milliseconds to seconds to respond</p>	Data in Many Forms <p>Structured, unstructured, text, multimedia</p>	Data in Doubt <p>Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations</p>

Source : <http://www.datasciencecentral.com/profiles/blogs/data-veracity>

Rising of Big Data

Big Data = Transactions + Interactions + Observations



十進位前綴
(SI)

名字 縮寫 次方

kilobyte	KB	10^3
megabyte	MB	10^6
gigabyte	GB	10^9
terabyte	TB	10^{12}
petabyte	PB	10^{15}
exabyte	EB	10^{18}
zettabyte	ZB	10^{21}
yottabyte	YB	10^{24}

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

How “Big” is Big Data?

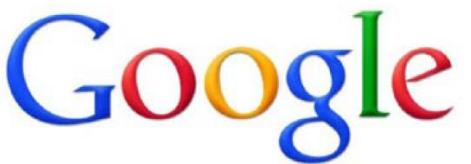
- **175ZB**有多大？(1ZB = 1萬億GB)

有時要理解如此
大的數字會很有難度。
以下例子可以體
會出175ZB的龐大。

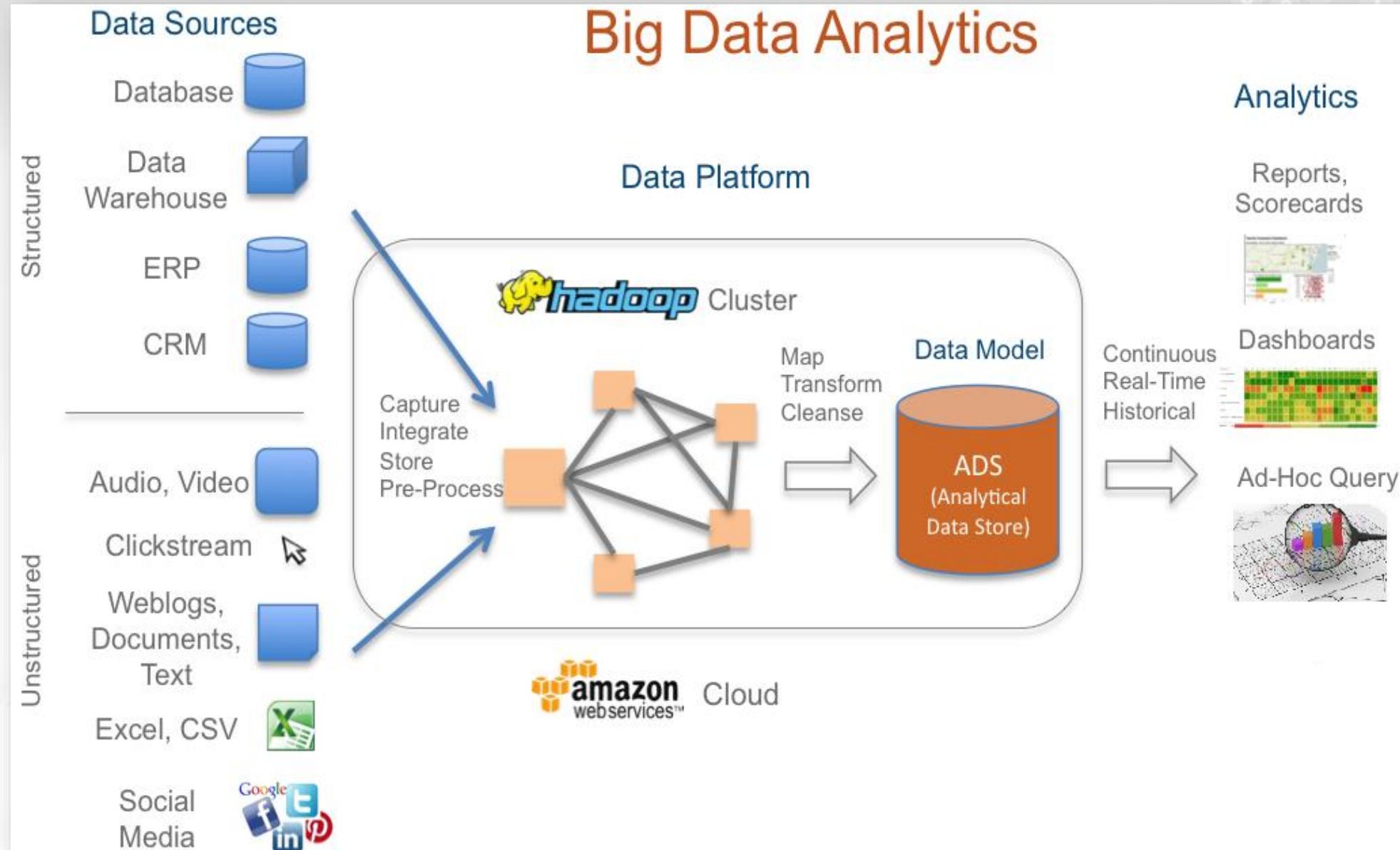
- 儲存在DVD
 - 累積的厚度足以到達月球23次，或環繞地球222圈
- 以平均**25Mb/秒**（全美平均網路速度）速度下載
 - 以一人下載 → 需18億年
 - 以全球人口同時下載 → 需81天

Where are “Big Data”?

- 像Google每天得處理超過24 PB (1 PB=1000 TB)的資料
- Facebook每小時會收到超過1千萬張新照片、30億次的留言
- YouTube用戶每秒上傳的影片總長度超過1小時
- Twitter的訊息量已經突破每天4億則

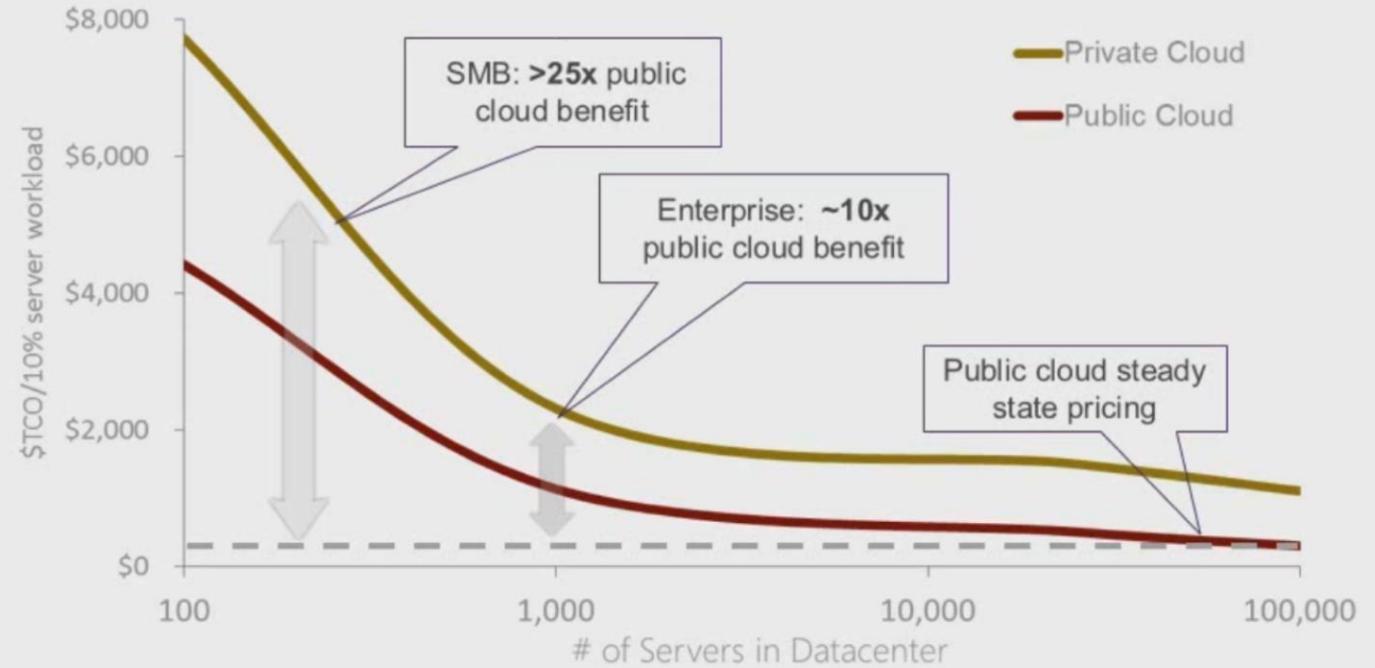


Process of Big Data Analytics



Big Data Analytics with Cloud Computing

Economics of the cloud Private vs. public clouds

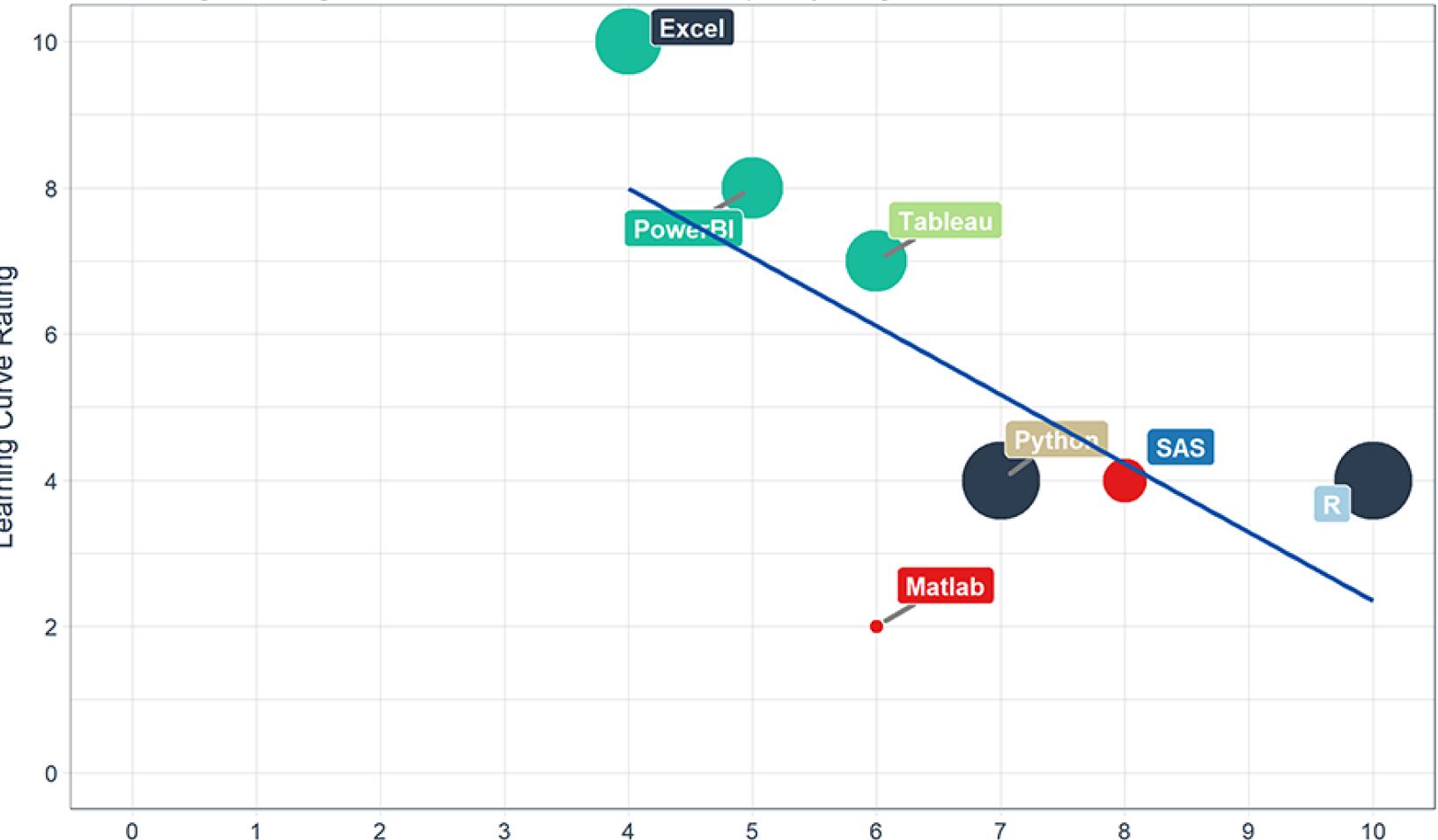


Sum up

- 資料科學的興起與資訊及通訊科技（ICT）的演進息息相關
 - 1960年代—資料庫系統（Database System）
 - 1970~80年代—關聯式資料庫（Relational Database）
 - 1980~2000年代—全球資訊網路的出現、電腦硬體技術的急速成長推動資料庫進階發展與資料倉儲（Data Warehouse）
 - 2000年代以後—雲端運算（Cloud Computing）、人工智慧（Artificial Intelligence）、物聯網（Internet of Thing）及工業4.0
- 未來，資料將成為最寶貴的資產，資料科學將改變企業經營模式，也刺激企業決策者開始思考如何有效運用資料探勘分析技術，從各種資料中淬煉出黃金，以掌握企業競爭優勢

Tools for Data Science

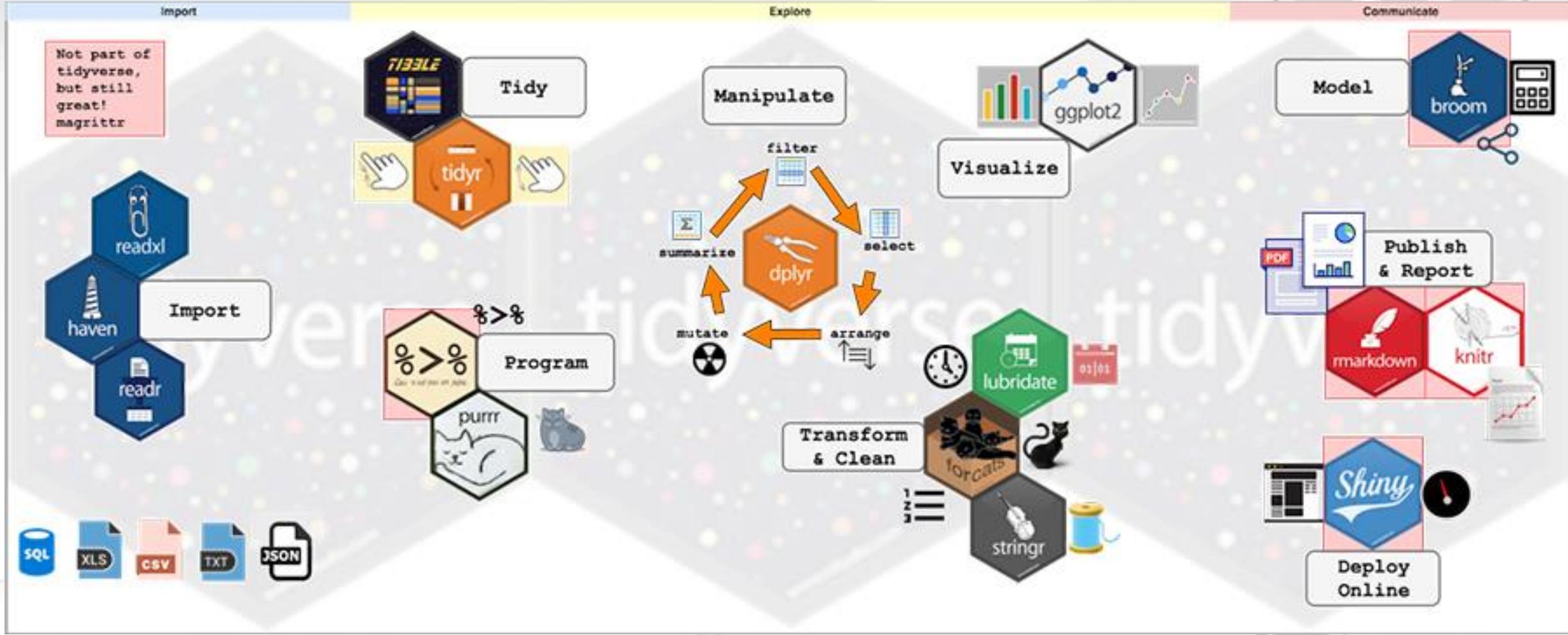
R has a longer learning curve but has a massive business capability rating



Data Science For Business Capability Rating



R Packages for Data Science



Source: <https://www.business-science.io/business/2018/10/08/python-and-r.html>

Workshop – Hands-on Practice

資料科學實作環境安裝與介紹

- R Language & IDE Tool: Rstudio Installation
- 教材：Workshop 1 - R Basics

Q & A