資料科學概論 – 資料分析與商業智慧

Dr. Tun-Wen Pai

1) 商業智慧
2) 辨識模式
3) 資料處理鏈
4) 資料視覺化

Feb. 22, 2023

商業挑戰:
盲人騎瞎馬
夜半臨深池
《世說新語-排調》

- 盲人 :企業缺乏具資訊能力的專業經理、無法有效經營管理及提供正確決策
- 騎瞎馬：缺乏具資訊能力的員工、無適當之工具及技術方法
- 夜半：COVID19 經濟不景氣的時機
- 臨深池 :參與激烈的全球化市場競爭

企業內部資料應用問題

Volume
Variety
Value
Velocity
Veracity

資料的保存及安全性

資料一致性、可整合性
個人目標或部門目標與公司目標不同

資料的價值性
那些資料可以推估經營績效

資料的有效性
資訊的關連性和即時處理的能力

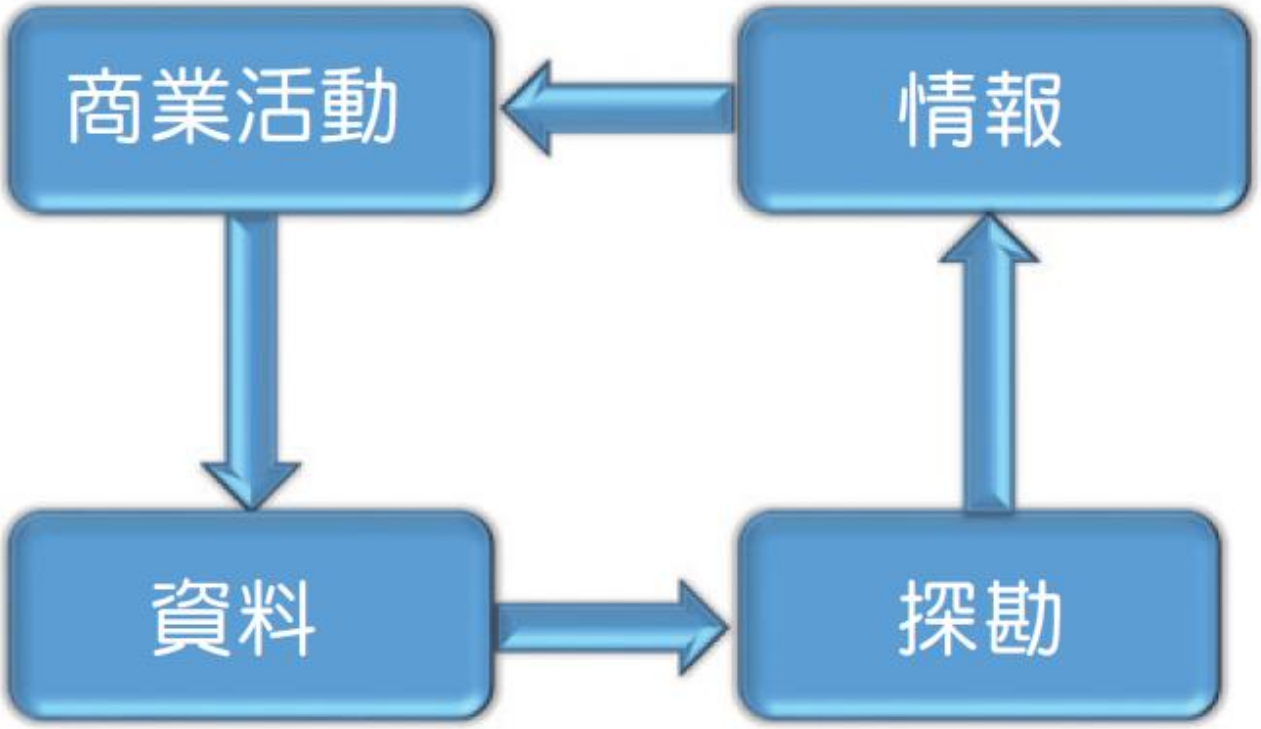
資料的正確性
如何從資料中保證其真實性及可信度

經營管理及數據導向決策問題

- 決策時缺乏合適、即時有效的資訊參考
- 報表不夠齊全，格式與決策層需要的不同
- 花費大批人物力，製作使用率低卻無效的報表，浪費生產力
- 報表和管理規範結合困難，經常發生異常狀況
- 無法從報表分析產生新管理觀念，找出成功關鍵因素及核心競爭力

商業智慧與資料探勘的循環

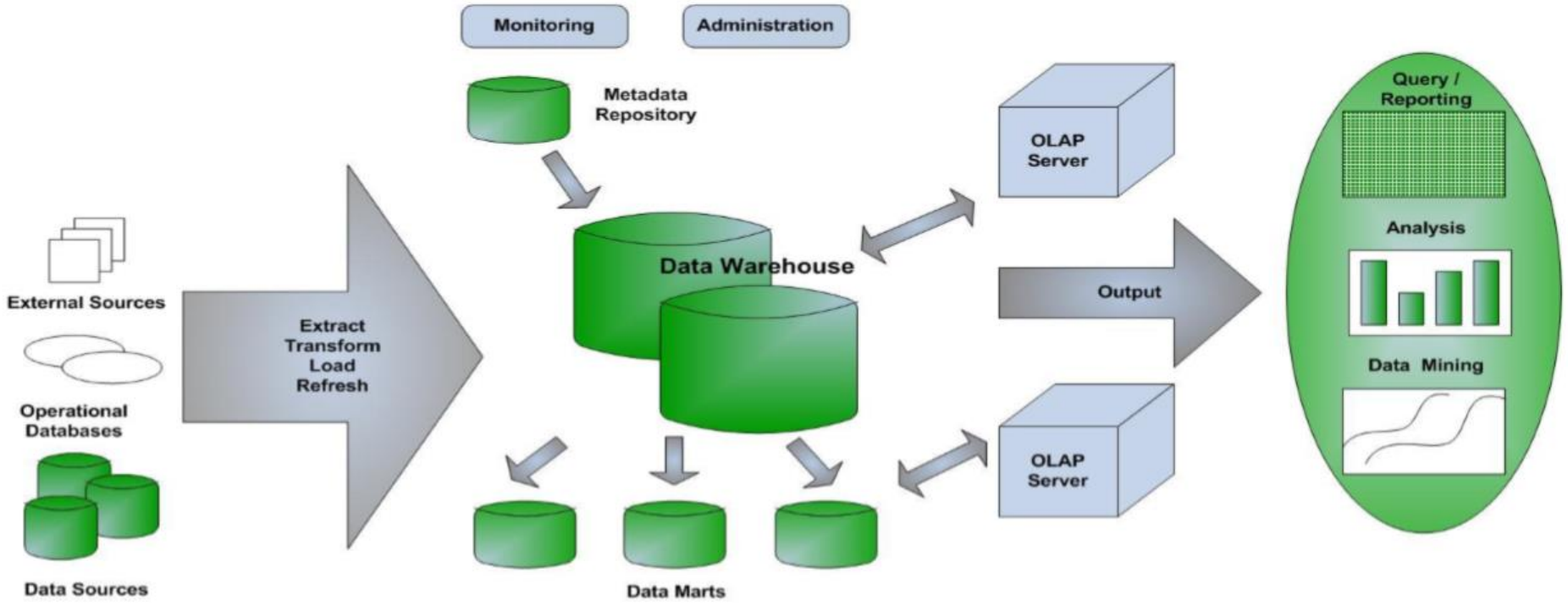
- Business Intelligence Development Model



▲圖 1.1：商業智慧與資料探勘（BIDM）循環

Business Intelligence Development Model

- Predefined reporting → Data marts → Enterprise-wide data warehouse → Predictive analytics → Operational BI → Business performance management (BPM)



Data Warehouse Architecture (Chaudhuri & Dayal, 1997)

資料倉儲Data Warehouse

- **資料倉儲**之父**比爾·恩門**（Bill Inmon）在1991年出版的“Building the Data Warehouse”一書中所提出的定義被廣泛接受，資料倉儲是一個主題式的（Subject Oriented）、整合的（Integrate）、穩定的（Non-Volatile）、時變性（Time Variant）的數據集合，用於支持管理決策。

- **資料倉儲(Data Warehouse)**

資料倉儲是過程而不是一個項目；資料倉儲是環境，而不是一件產品。資料倉儲提供用戶用於決策支援的當前和歷史數據，這些數據在傳統的操作型態資料庫很難或不能得到。資料倉儲技術是為了有效的把原始數據整合到統一的環境中以提供決策型數據的各種技術的總稱。所做的一切都是為了讓用戶更快更方便查詢所需要的資訊，提供決策支援。

- **資料超市**（Data Marts）

為**特定的應用目的或應用範圍**，而從資料倉儲中獨立出來的一部分數據，也可稱為**部門數據**或**主題數據集**。在資料倉儲的實施過程中往往可以從一個部門的資料超市著手，以後再用幾個資料超市組成一個完整的資料倉儲。需要注意的是在實施不同的資料超市時，同一定義的欄位一定要相容，之後才能實施新的資料倉儲。

https://wiki.mbalib.com/zh-tw/数据仓库

辨識模式

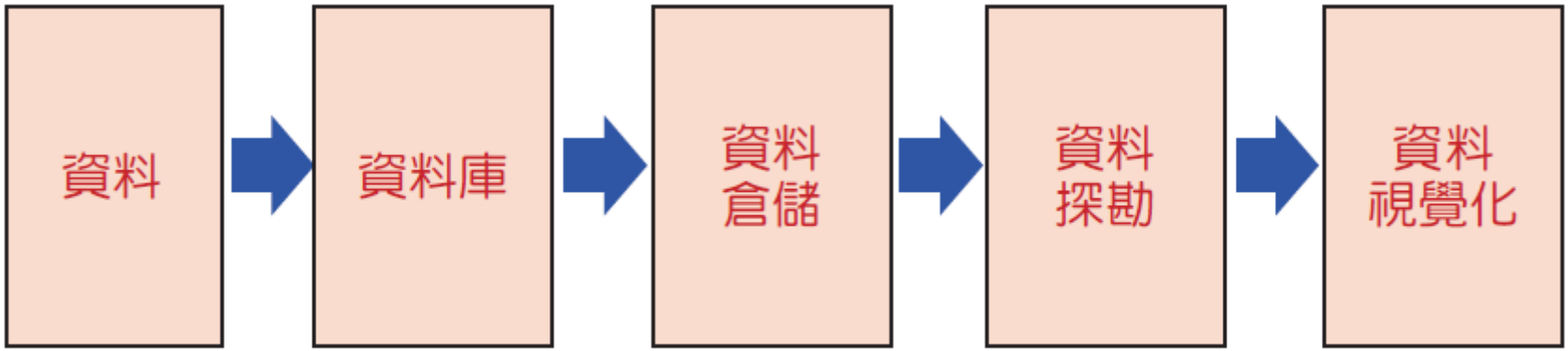
- 模式Pattern有助於解析複雜事物、展露趨勢。
- 模式的種類– 時間、空間、功能
  - Parkinson's Law(帕金森定理):

    在工作能夠完成的時限內，工作量會一直增加，直到所有可用時間都被填充為止。帕金森定理被當成一個數學等式，用來描述官僚組織隨著時間而擴大的速率。帕金森觀察到，一個官僚組織中的雇員總數，通常以每年5-7%的速度增加。他認為，有兩股力量造成了這個增長：(1) 一個官員希望他的下屬增加，但不希望解僱造成敵人增加；以及(2) 官員會製造工作給彼此。帕金森定律因此也可以表示，大型企業組織因雇用太多冗餘人員，造成企業規模膨脹、組織效率低下。
  - Pareto Principle (80/20 法則、關鍵少數法則、八二法則):

    約僅有20%的變因操縱著80%的局面。也就是說：所有變量中，最重要的僅有20%，雖然剩餘的80%占了多數，控制的範圍卻遠低於「關鍵的少數」。此概念起源於義大利經濟學家帕雷托 (Vilfredo Pareto) 在洛桑大學注意到了80/20的聯繫，於他的第一篇文章《政治經濟學》中說明了該現象，例如：義大利約有80%的土地由20%的人口所有、80%的豌豆產量來自20%的植株、 80%的銷售額來自20%的客戶、北科的研究計畫經費及論文80%是來自20%的老師、有80%的客服電話都與20%的產品有關等等。
  - 功能性模式指執行某事而導致特定效果，例如與應試技巧相關範例，有些學生對論述問題表現佳、有些則對多重選擇擅長，有些學生喜歡實作專案或是口頭簡報，若能認知學生存在不同模式，有助於設計對大家公評的測驗機制。

Examples: 學生離校、員工對雇主態度、行動電話在高速公路的移動位置、員工手機定位系統、Google查詢關鍵字、船隻定位導航系統

資料處理鏈 (Data processing chain)



▲ 圖 1.2：資料處理鏈

- 資料是新自然資源 (natural resources)
- 資料是商業智慧的核心!
- 資料 → 資料庫 → 報表分析→ 資料倉儲 → 整合資料→ 探勘技術→ 產生新見解→ 視覺化呈現

資料與變數

- **資料 (Data)** 是指各種數據分析應用所觀察或測量的集合。

樣本資料集包含每個樣本個別的觀測變數值，研究盡量使用每一個樣本所收集到的測量值，但有些實際狀況是無法取得到每一個樣本體的測量值。

例如在醫學或公衛應用領域要比較不同國家在COVID19疫情下的死亡率及危險因子，對任一個國家所提供的資料而言，通常無法取得真正因為COVID死亡樣本的真實有關資料，例如感染人數、真正死亡人數、死亡時間及死亡原因、日期或年紀，也無法得到每一位死亡樣本的其他相關生醫共病資訊。

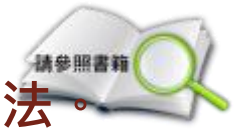
資料與變數

- **變數 (Variable)** 是指每一個樣本可以測量或是描述的特徵(**獨立變數/相依變數**)。

每一個樣本定義將使用的變數(特徵)，每個變數有相對收集到的特徵數據。例如醫學研究常使用的特徵有年齡、身高、體重、有無吸菸與喝酒的習慣、是否罹患肺癌、感染COVID之後是否死亡等特徵。

觀測值(Observed value)是指某一特定樣本所觀察或測量的數值。例如一位COVID患者感染後的臨床用藥試驗，研究資料使用的變數包括年齡、性別、身高、體重、罹患不同慢性病歷史紀錄、施打疫苗種類等資料。

通常經過資料分析或預測的結果變數會有一個或多個(包含主要結果或次要結果)的測量，這些分析結果可以稱為反應變數、相依變數、結果變數。其他用來解釋結果變化的變數則稱為獨立變數、共變數、解釋變數、干擾因子、風險因子或是預後因子等不同辨識名稱。



分析資料前必須設定每個變數的分類與屬性，以選擇適當的分析方法。根據不同觀點及應用情境，變數的分類方式有所不同，例如使用科學測量尺度變數可分為類別尺度 (categorical)或數量尺度(quantitative, numerical)。 還有其他分類方式如統計觀點或是資料管理的觀點等。

科學測量尺度的資料 (類型)

名目資料 (nominal data, unordered collection)

次序資料 (ordinal data, ordered data)

區間資料 (interval data, equal intervals)

比值資料 (ratio data, any fraction data)

二進位大型物件BLOB(Binary Large Objects)

名目資料與次序資料

- **類別尺度**定義的特定幾種類別水準是否有大小差距又可分類成名目尺度 (nominal scale)與次序尺度(ordinal scale)。名目尺度是最簡單的測量，將變數值分成互斥的類別水準，同一變數內的類別水準並無量化大小的差別.
- 名目尺度的特別情形是該變數只有2個類別水準，如存活或死亡、感染或無感染(統計習慣會分別標記為0或1)，這類變數常稱做二元變數(binary variable/dichotomous variable)。
- 順序尺度(Ordinal scale)是指同一變數的類別水準有輕重、大小、強弱、好壞等級順序的資料。例如教學評量問卷、罹患癌症期數 I, II, III, IV 等 4 期、 COVID19的無症狀、輕症及重症患者的感染等級。雖然順序尺度用數字表示或標記，但是數字本身通常不能用來做運算，只能比較相對大小或高低次序，順序之間的實際差異並無法從標記的數字差異得知.

區間資料與比值資料

- 數量尺度(Quantitative/Numerical)是指變數觀察值在數字之間的差異是有意義的。例如身高、體重、溫度等。可分成區間尺度(Interval scale)與比值尺度(Ratio scale)。
- 區間尺度(Interval scale)測量值不限於離散的整數，可以測量數字間的距離，但沒有一個真正零點(True zero)，沒有絕對零值 (Absolute zero)，因此可以有負值。例如溫度測量的-2度等。區間尺度無法真正表示成倍數，例如室內目前溫度15度C、室外溫度5度C，我們不會描述現在室內比外面熱三倍或是室外的冷度是室內的三倍。室外如果剛好零度或是-1度...比值??
- 比值尺度 (Ratio scale) 是指變數測量值有一個真正零點，或稱真實零值(True zero)，測量值不可以有負值，例如年齡、身高、體重、血壓等變數。比值尺度可以表示或計算倍數，例如 80 公斤為 40 公斤的 2 倍重。

Summary

In summary, **nominal** variables are used to "*name*," or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a "true zero" can be defined.

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Summary of data types and scale measures

測量尺度(scale of measure)

水平	名稱	*又稱*	可用的邏輯與數學運算方式	舉例	集中趨勢的計算	離散趨勢的計算	定性或定量
1	名目	*名義、類別*	等於、不等於	二元名目：性別（男、女） 二元名目：出席狀況（出席、缺席） 二元名目：真實性（真、假） 多元名目：語言（中、英、日、法、德文等） 多元名目：上市公司（蘋果、美孚、中國石油、沃爾瑪、雀巢等）	眾數	*無*	定性
2	次序	*順序、序列、等級*	等於、不等於 大於、小於	多元次序：服務評等（傑出、好、欠佳） 多元次序：教育程度（小學、初中、高中、學士、碩士、博士等）	眾數、中位數	分位數	定性
3	等距	*間隔、間距、區間*	等於、不等於 大於、小於 加、減	溫度、年份、緯度等	眾數、中位數、算術平均數	分位數、全距	定量
4	等比	*比率、比例*	等於、不等於 大於、小於 加、減 乘、除	價格、年齡、高度、絕對溫度、絕大多數物理量	眾數、中位數、算術平均數、幾何平均數、調和平均數等	分位數、全距、標準差、變異係數等	定量

https://zh.wikipedia.org/wiki/%E6%B8%AC%E9%87%8F%E5%B0%BA%E5%BA%A6

分位數 (Quantile)

- **中位數（即二分位數）、四分位數（quartile）、十分位數（decile）、百分位數等。q-quantile是指將有限值集分為q個接近相同尺寸的子集**
- **四分位數（Quartile）是統計學中分位數的一種，即把所有數值由小到大置換並分成四等份，處於三個分割點位置的數值就是四分位數。**
- 第一四分位數 (Q1) 又稱較小四分位數，等於該樣本中所有數值由小到大置換後第25%的數字。
- 第二四分位數 (Q2) 又稱中位數，等於該樣本中所有數值由小到大置換後第50%的數字。
- 第三四分位數 (Q3)，又稱較大四分位數，等於該樣本中所有數值由小到大置換後第75%的數字。
- 第Q3-Q1的差距又稱四分位距（**InterQuartile Range, IQR**）。

例1

數據總量：$6, 47, 49, 15, 42, 41, 7, 39, 43, 40, 36$

由小到大置換的結果：$6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49$

$$\begin{cases} Q_1 = 15 \\ Q_2 = 40 \\ Q_3 = 43 \end{cases}$$

例2

數據總量：$7, 15, 36, 39, 40, 41$

$$\begin{cases} Q_1 = 15 \\ Q_2 = 37.5 \\ Q_3 = 40 \end{cases}$$

例3

數據總量：$1, 2, 3, 4$

$$\begin{cases} Q_1 = 1.5 \\ Q_2 = 2.5 \\ Q_3 = 3.5 \end{cases}$$

Interquartile range (IQR= Q3 – Q1) is the width of the box in the box-and-whisker plot.

Outliers: If a data point is below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$

Extreme value: If a data point is below $Q_1 - 3 \times IQR$ or above $Q_3 + 3 \times IQR$

箱形圖 (Box plot or box-and-whisker plot)

Example: find the outliers and the extreme values, if any, for the following data set:

29, 23, 25, 24, 21, 49, 33

Sorting: 21, 23, 24, 25, 29, 33, 49

$Q_1 = 23$, $Q_2 = 25$, $Q_3 = 33$

IQR = 33 – 23 = 10

Outlier values:

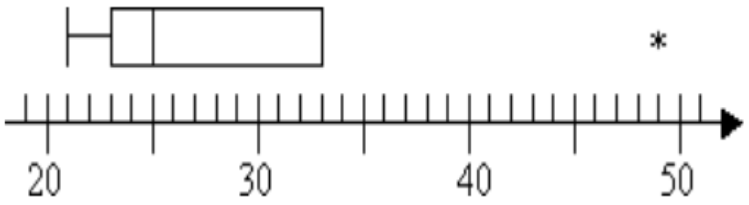
23 – 1.5×10 = 23 – 15 = 8

33 + 1.5×10 = 33 + 15 = 48

Extreme values:

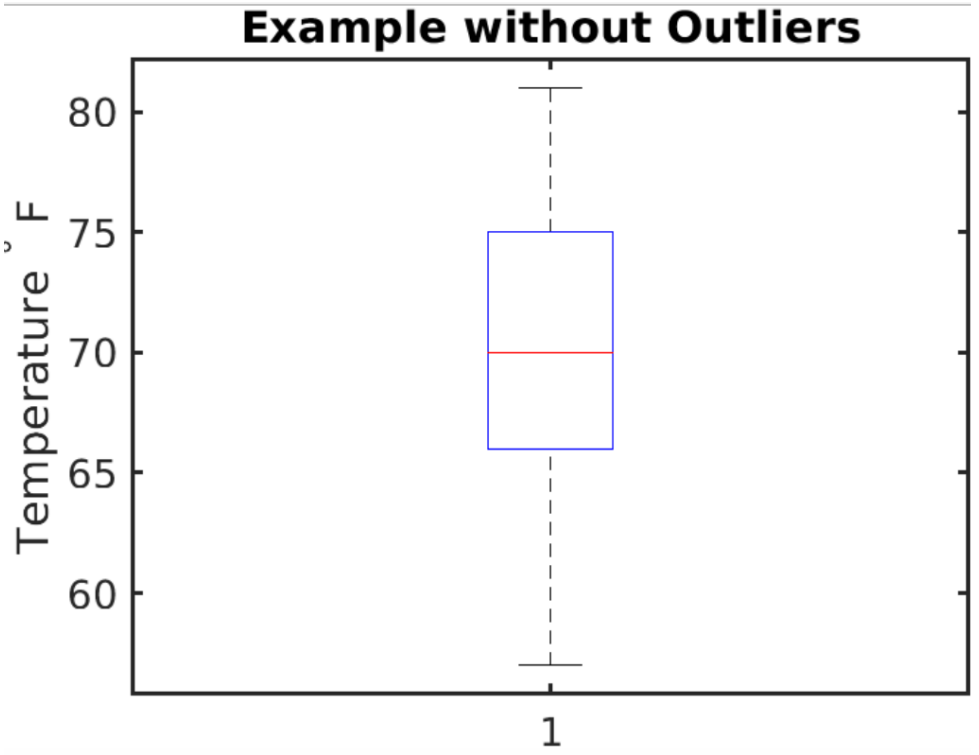
23 – 3×10 = 23 – 30 = –7

33 + 3×10 = 33 + 30 = 63



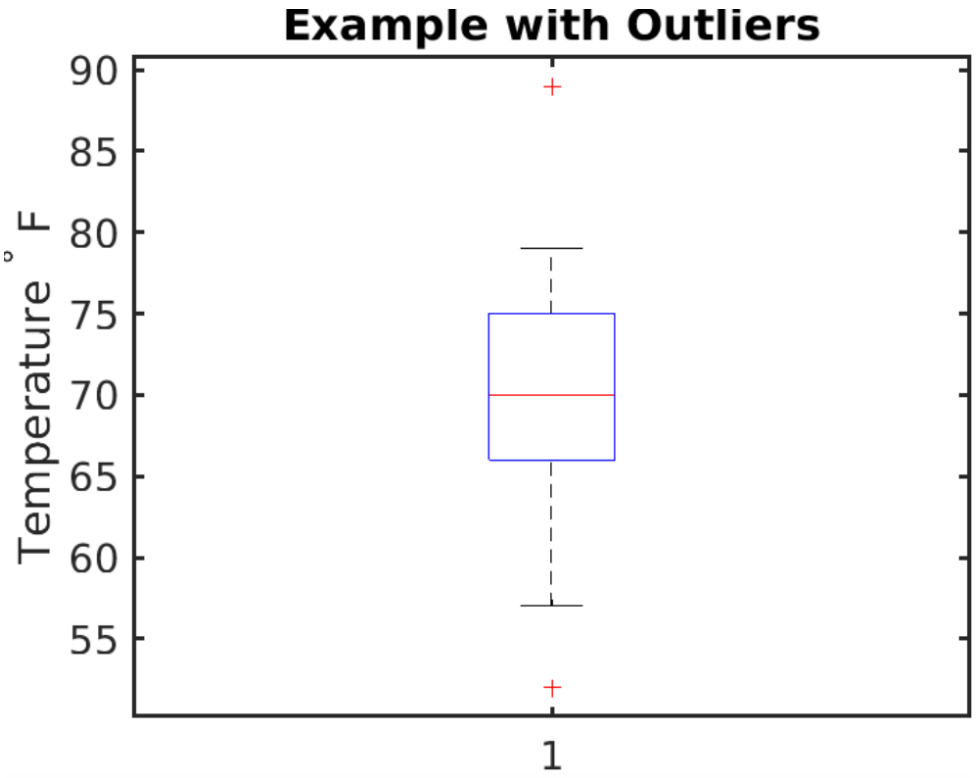
Example without outliers

- The recorded temperature values are listed in order as follows (°F): 57, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 81.
- IQR=Q3-Q1=75°F-66°F=9°F
- 1.5xIQR=1.5x9°F=13.5°F
- Q3+1.5 IQR=88.5°F
- Q1-1.5 IQR=52.5°F

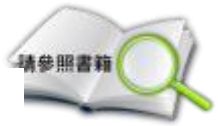


Example with outliers

- The recorded temperature values are listed in order as follows (°F): 52, 57, 57, 58, 63, 66, 66, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 89.
- IQR=Q3-Q1=75°F-66°F=9°F
- 1.5xIQR=1.5x9°F=13.5°F
- Q3+1.5 IQR=88.5°F
- Q1-1.5 IQR=52.5°F



Box Plot in Jupyter Notebook

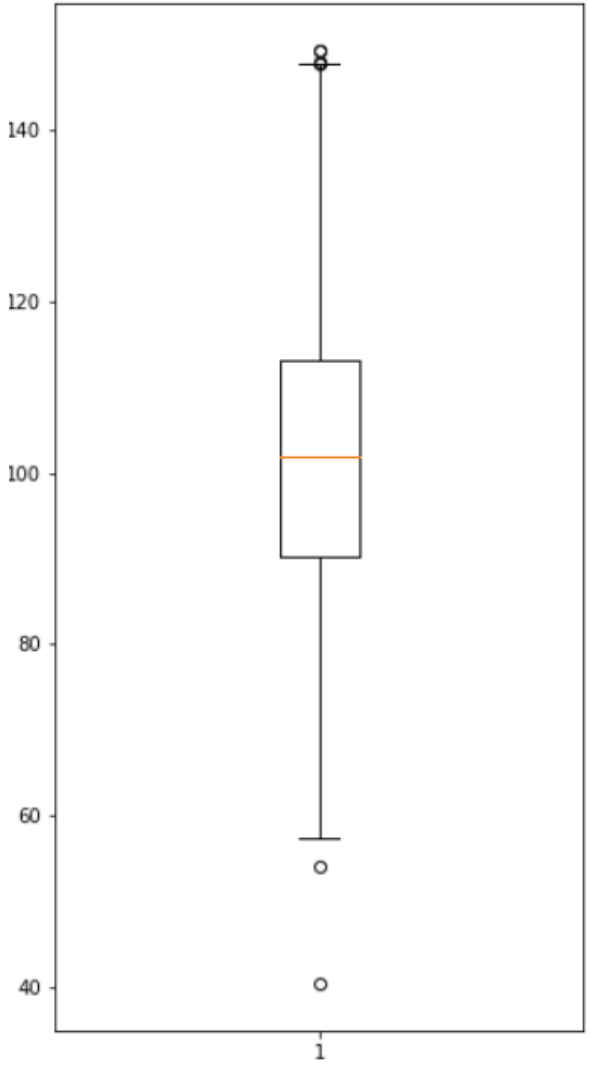



```
# Import libraries
import matplotlib.pyplot as plt
import numpy as np

# Creating dataset
np.random.seed(10)
mu=100
sigma=20
data = np.random.normal(mu, sigma, 200)
Print(data)
fig = plt.figure(figsize =(5, 10))

# Creating plot
plt.boxplot(data)

# show plot
plt.show()
```



HW1-Problem 3 (Due 3/15/2023)

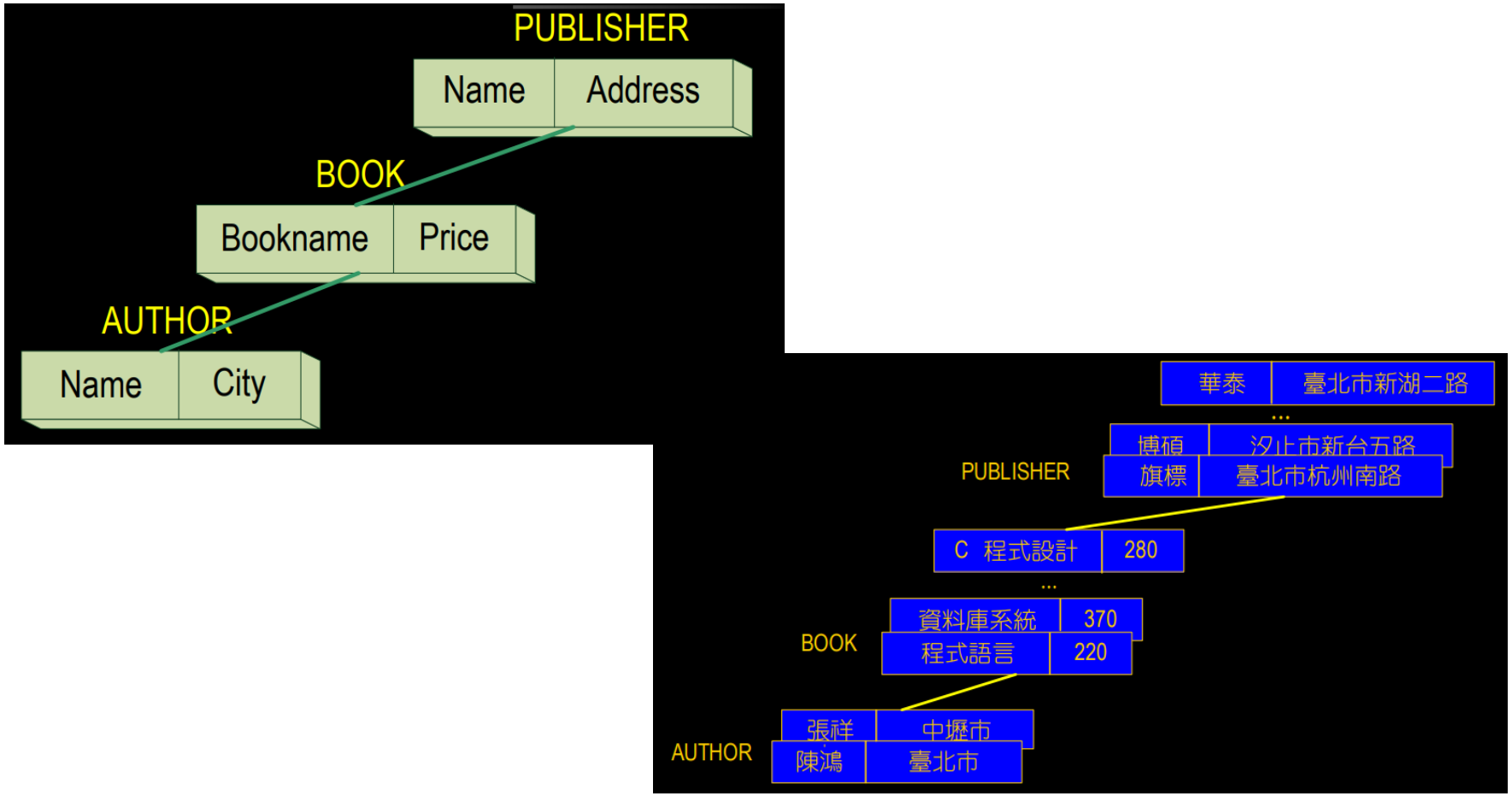
- Draw the box-and-whisker plot of the following data **in Jupyter Notebook**.
- Describe the outlier/extreme values, if any, for the following data set:
- 14.7, 14.4, 8.2, 10.7, 14.6, 14.1, 14.4, 14.4, 18.2, 14.5, 14.5, 14.7, 14.9, 15.1, 15.9, 25.0

資料處理 -- 資料模型(Data Model)

- 資料庫技術開始於 20 世紀 60 年代末期，其主要目的是有效地管理和存取大量的資料。而為了能有組織有效率的將我們需求的資料儲存於資料庫系統中，我們需要一種能將資料適當表示的方式，稱之為**資料模型 (data model)**。將現實世界的資料特徵抽象化，用於描述一組資料的概念和定義。

- **階層式模式**主要是將一筆一筆的紀錄 (record) 以樹狀結構的方式組織起來，由於**樹狀結構**的特性，這樣的方式非常適用於一對多的資料組成關係，樹狀結構並不能描述紀錄與紀錄之間的網路連結關係，所以網路式資料模式也被提出討論(多對多的資料組成)。**關聯式資料模型 (relational data model)** 革新了整個資料庫領域，並大幅地取代了這些早期的模型。由於大量非結構化資料的數據世代來臨，**NoSQL 資料庫**因應產生，可以為特定資料模型而建立，並具有構建新型應用程式的彈性結構描述。NoSQL 資料庫在開發的容易性、功能性和大規模效能方面廣受肯定。

階層式模式範例


PUBLISHER
Name
Address
BOOK
Bookname
Price
AUTHOR
Name
City
華泰
臺北市新湖二路
博碩
汐止市新台五路
PUBLISHER
旗標
臺北市杭州南路
C 程式設計
280
資料庫系統
370
BOOK
程式語言
220
張祥
中壢市
AUTHOR
陳鴻
臺北市

關聯式資料庫

- 關聯式資料庫即是藉由分析欲儲存的實體 (entity) 彼此間的關係 (relation)，進而建構儲存資料模型的方式。在建構資料模型時，常常會利用**實體關聯圖 (Entity Relation Diagram, ER Diagram)** 的方式作為輔助設計的依據。
- 一個關聯的表達方式是一個表格 (Table)。換句話說，關聯式資料庫是由許多表格 (Table) 所組成之資料的集合。
- 一個關聯 (或稱表格) 是由許多列 (Row) 和行 (Column) 所組成。列(row)表示一個關聯的列組 (Tuple)，行(column)的欄位表示關聯的屬性 (Attribute)。
- 一個關聯表的列組 (Tuple) 經常被稱為一個紀錄 (Record)。

- 以設計簡單的選課系統為例子，來說明實體關聯圖以及關聯式資料庫的觀念。

關聯式資料庫 (續)

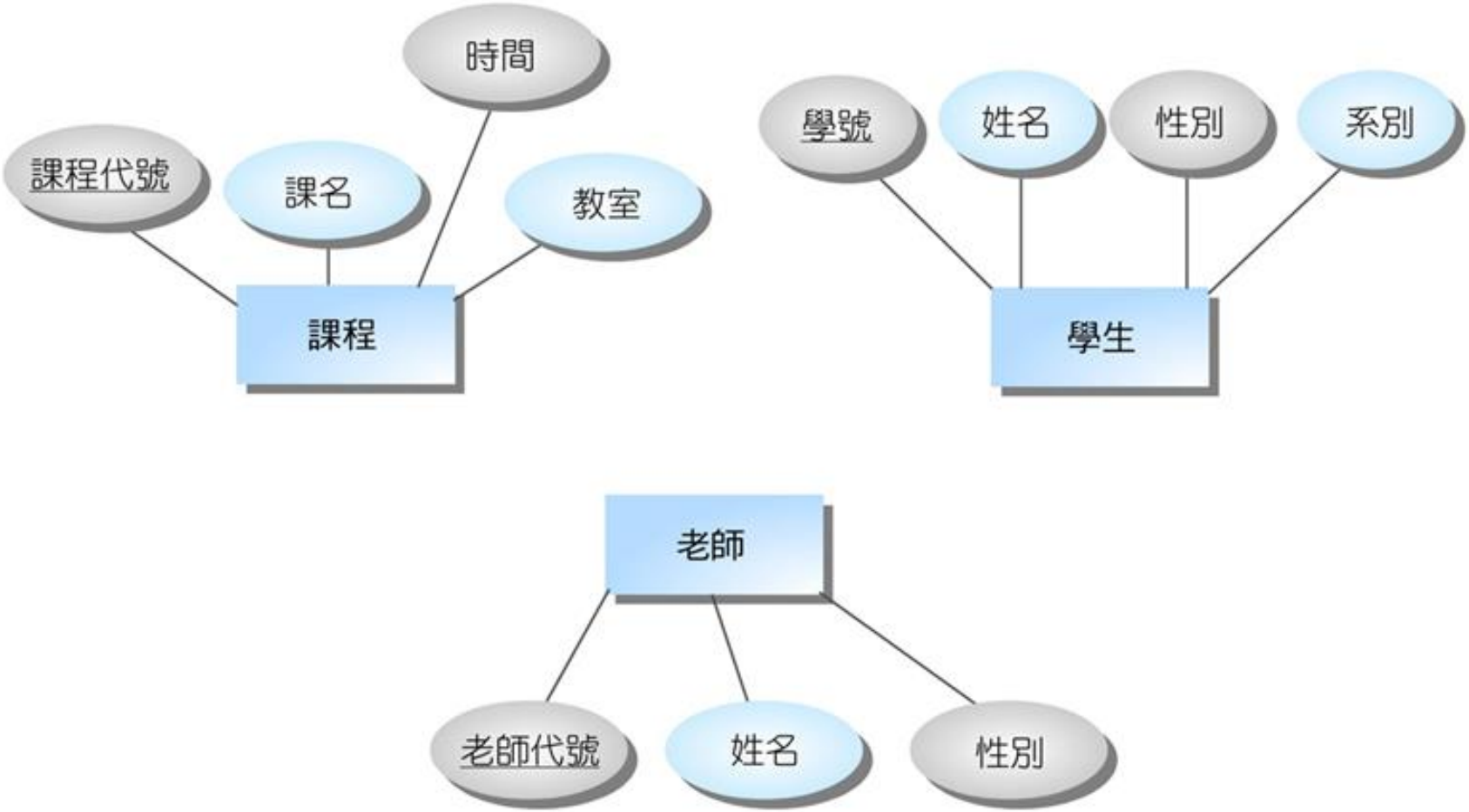
- 一個選課系統應該有開授的課程，修課的學生以及開課的老師，因此裡面應該包含「課程」、「學生」及「老師」三個角色，在實體關聯圖的設計中，這些角色稱之為實體，可以想成是同一類型且具有相同特性的物件集合。

實體表示方式

關聯式資料庫 (續)

- 每一個實體，我們將感興趣的特性一一列下，對於學生而言，學號、姓名、性別與系別等是該物件感興趣的特性，而我們會稱這些特性為學生實體的**屬性 (attribute)**。屬性在實體關聯圖中是以連在對應實體上的橢圓形表示。
- 實體的**主鍵 (primary key)** 是由一個或一個以上的屬性所組成，其值組具有**唯一性**，亦即主鍵的值是不能重複的。例如學號可以當成學生的主鍵。
- 在實體關聯圖中選為主鍵的屬性下方會加上一條橫線作為標示。

關聯式資料庫 (續)



屬性以及主鍵的表示方式

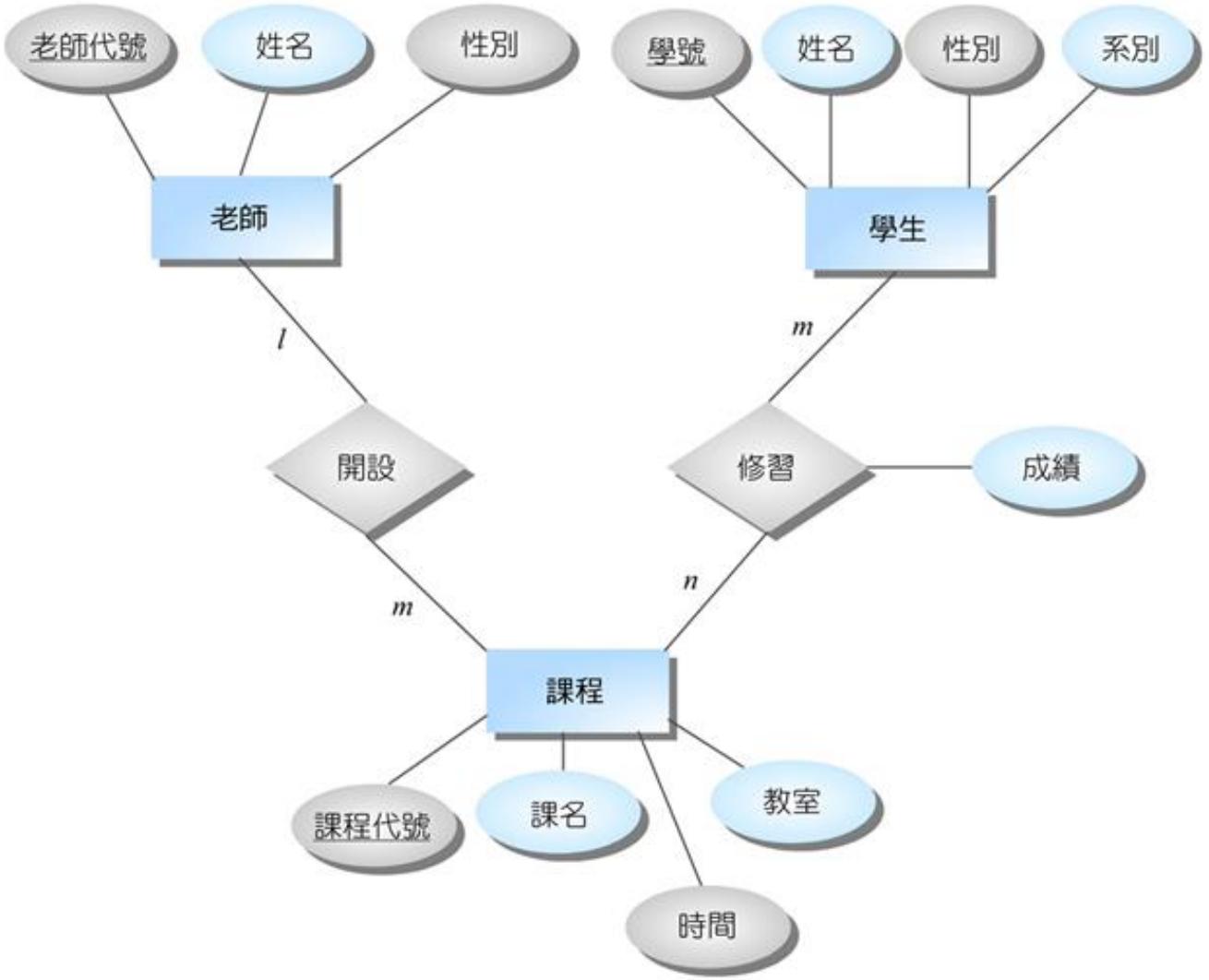
關聯式資料庫 (續)

- 建構完實體之後，接著就要考慮實體間的關係。

在選課系統中，學生與課程之間有修習的關係，而且一個學生可以修習多門課程，而一個課程也可以有多名學生修習，此為**多對多 (many to many)** 的關係。而老師與課程有開設的關係，假設一個老師可以開設多門課程，但一門課程只能由一個老師開設，那麼老師與課程間的關係即為**一對多 (one to many)** 的關係。

- 學生修習課程後會有學習成績，學習成績是因為修課關係的成立而擁有的屬性。
- 在實體關聯圖中，**關係是以菱形表示**，並會利用數字 1、*m* 與 *n* 代表實體間的對應關係。

關聯式資料庫 (續)



選課資料庫的實體關聯圖

關聯式資料庫 (續)

- 只須要根據實體關聯圖，將實體與關係利用資料結構儲存就可以。在關聯式資料庫中，儲存這些資料的結構稱之為**關聯 (relation)**，可以將**關聯想成是表格，而表格的欄位就是屬性**。

根據實體所建置的表格

老師

老師代號	姓名	性別
T1	趙依	女
T2	錢二	男
T3	孫參	男
T4	李似	女
…	…	…

學生

學號	姓名	性別	系別
S1	王曉明	男	資工
S2	林小花	女	資工
S3	陳依依	女	電機
S4	蘇怡君	女	數學
…	…	…	…

課程

課程代號	課名	時間	教室
C1	計算機概論	M2~M4	R101
C2	資料庫系統	T6~T8	R321
C3	微積分	F5~F8	R101
C4	個體經濟	F5~F8	R202
…	…	…	…

關聯式資料庫 (續)

根據關係所建置的表格

開設

老師代號	課程代號
T1	C1
T1	C2
T1	C3
T2	C4
…	…

修習

學號	課程代號	成績
S1	C1	A
S1	C2	B
S2	C1	A+
S2	C3	C–
…	…	…

- 透過上述的五個表格，就可以回答各式各樣的問題

關聯式資料庫 (續)

- 資料庫內用的表格越多，在查詢處理時，要參考的表格也會變多。如何在不影響資料表示能力的情形下，減少表格的個數，老師與課程的關係是一對多的關係，其中的課程代號是不會重複出現的，我們只要將課程表格多加一個老師代號的欄位，即可表示該門課是由哪位老師開課，可以省去開設的表格！這樣的方式只有在關係是**一對多或是一對一** 的情況下適用。
- 課程表格的前四個欄位都是屬於課程本身的屬性，而老師代號則是老師實體的主鍵，對於這樣外來的屬性，我們稱之為**外來鍵 (foreign key)**。

關聯式資料庫 (續)

選課資料庫表格

老師

老師代號	姓名	性別
T1	趙依	女
T2	錢二	男
T3	孫參	男
T4	李似	女
…	…	…

學生

學號	姓名	性別	系別
S1	王曉明	男	資工
S2	林小花	女	資工
S3	陳依依	女	電機
S4	蘇怡君	女	數學
…	…	…	…

外來鍵

課程

課程代號	課名	時間	教室	老師代號
C1	計算機概論	M2~M4	R101	T1
C2	資料庫系統	T6~T8	R321	T1
C3	微積分	F5~F8	R101	T1
C4	個體經濟	F5~F8	R202	T2
…	…	…	…	

修習

學號	課程代號	成績
S1	C1	A
S1	C2	B
S2	C1	A+
S2	C3	C–
…	…	…

關聯式資料庫 (續)

- 通常為了簡化描述的方式，我們會以**關聯表綱要 (relation schema)** 的方式表示在資料庫中有哪些關聯。

老師	<老師代號，姓名，性別>
學生	<學號，姓名，性別，系別>
課程	<課程代號，課名，時間，教室，老師代號>
修習	<學號，課程代號，成績>

關聯表綱要

查詢語言 SQL

- 標準的關聯式資料庫使用的查詢語言稱為**結構化查詢語言 (Structured Query Language, SQL)**。

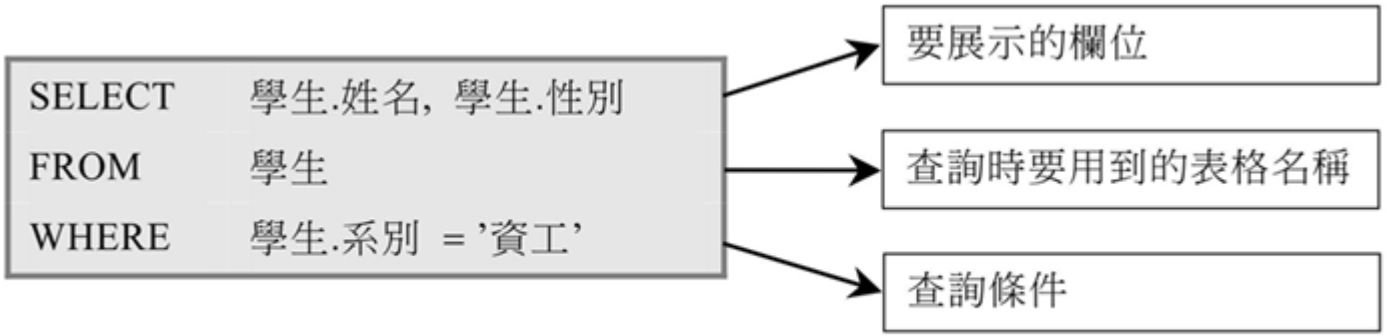
SELECT	要顯示的結果屬性
FROM	查詢處理時會用到的表格（關聯）
WHERE	查詢條件

SQL 基本語法

查詢語言 SQL (續)

範例一

- 在選課資料庫中，我想要找出所有資工系學生的姓名與性別，查詢語言要怎麼寫呢？
- 根據查詢的需求，我們要展示的屬性是學生的姓名與性別，而查詢條件是跟系別有關；因為姓名、性別以及系別這三個屬性都記錄在學生的表格中，因此在查詢處理時，只需參考學生表格即可，最後，再把查詢條件「系別 = 資工」放入 WHERE 子句中，即完成查詢



範例一的 SQL

查詢語言 SQL (續)

```
SELECT    姓名, 性別
FROM      學生
WHERE     系別 = '資工'
```

範例一簡化後的 SQL

範例一的查詢結果

姓名	性別
王曉明	男
李小花	女
…	…

查詢語言 SQL (續)

範例二

- 請找出趙依老師開設的課程名稱，SQL 該怎麼寫呢？
- 在這個查詢中，我們會用到老師的姓名，以及課程名稱兩個欄位，而這些欄位分別屬於老師表格與課程表格。又因為查詢需求是趙依老師開設的課程名稱，因此查詢條件除了老師姓名要等於趙依之外，我們還須要加上「老師.老師代號 = 課程.老師代號」條件 (在 SQL 中，AND 代表兩個條件都要成立)，透過這樣的關係條件，就可以限制在課程表格中只找趙依老師開設的課程

```
SELECT    課名
FROM      老師, 課程
WHERE     姓名 = '趙依' AND
          老師.老師代號 = 課程.老師代號
```

範例二的 SQL

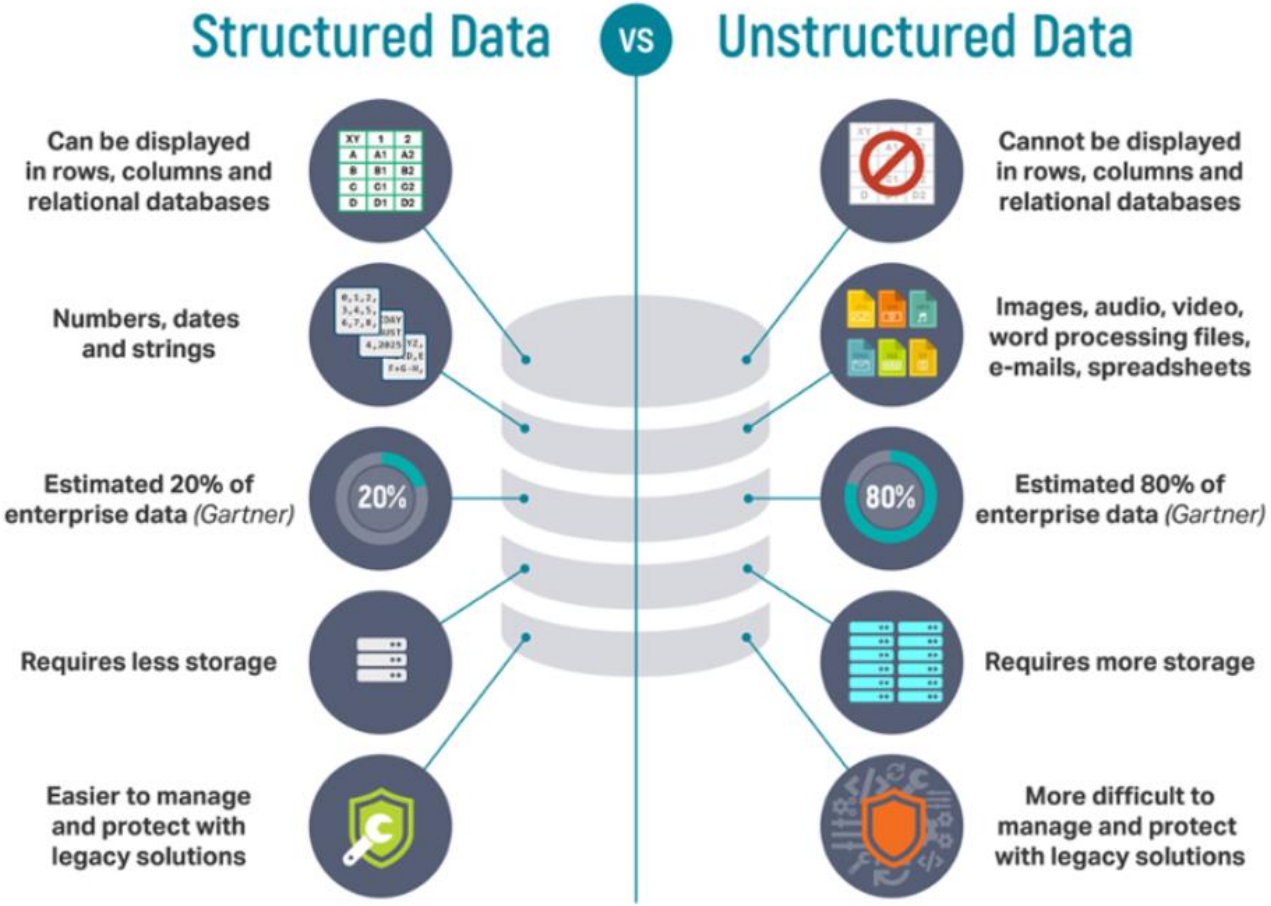
NoSQL (非關聯式) 資料庫

- NoSQL 資料庫是為特定資料模型而建立，具有彈性結構描述、容易性、功能性和大規模效能的特色
- NoSQL 資料庫使用多種資料模型來存取及管理資料。資料庫透過放寬傳統關聯式資料庫的一些資料一致性限制，特別針對需要大量資料、低延遲和彈性資料模型的應用程式進行優化
- 關聯式資料庫--資料庫需進行正規化以減少資料冗餘，同時減少異常情形並保護資料庫的完整性。RDBMS資料庫以結構化的方式組織可保有的優勢包括**單元性、一致性、隔離性、持續性**（Atomicity、Consistency、Isolation、Durability，ACID）之標準符合性。 RDBMS的主要問題是隨著資料庫的增長而需要面臨擴展資料庫的挑戰。
- NoSQL 資料庫是處理大量**非結構化資料**，或者當資料需求在一開始就不明確時的最佳選擇。

結構化 vs 非結構化資料

- **結構化資料**：資料被呈現在資料庫表格的row(列)、column(行)。一列資料代表一筆紀錄(record)，統計的術語稱為一次觀測(observation)。每一行或欄位(column)則稱為特徵(characteristics)或變數(variable)。使用統計術語描述，表格資料的每一筆紀錄代表一次觀測，而每一個觀測的每一個欄位都是該觀測的特徵值。
- **非結構化資料**：形式自由且不遵循標準的格式規範，一組沒有組織的數據。非結構化數據的示例包括影像資料，語音資料，視訊資料，表格和文字處理檔等，實質上是存儲為文件的資料架構。非結構化數據往往比結構化數據更大，會佔用更多存儲空間。

結構化 vs 非結構化資料



NoSQL 資料庫

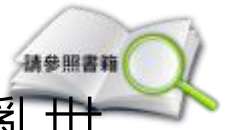
優勢

- 可以使用沒有結構的資料庫儲存大量資料(CouchDB、MongoDB、Cassandra 和 HBase)
- NoSQL 資料庫容易擴展多個資料中心

弱點

- NoSQL 社群缺乏 MySQL 使用者群的成熟度
- 缺乏效能測試和分析的報告工具
- 缺乏標準化，各自傾向使用自己的語法，不同語法可能不易掌握，且通常與關聯式資料庫所使用的 SQL 不相容
- NoSQL 資料庫可能會犧牲 ACID 標準符合性來提高處理速度和靈活性

資料庫另一簡單範例



Monty Python(聖杯傳奇,1975)、Gone With the Wind(亂世佳人,1939)、Matrix(駭客任務,1999)

電影交易資料庫				
訂單編號	銷售日期	產品名稱	地點	金額
1	2015 年 4 月	聖杯傳奇	US	$9
2	2015 年 5 月	亂世佳人	US	$15
3	2015 年 6 月	聖杯傳奇	India	$9
4	2015 年 6 月	聖杯傳奇	UK	$12
5	2015 年 7 月	駭客任務	US	$12
6	2015 年 7 月	聖杯傳奇	US	$12
7	2015 年 7 月	亂世佳人	US	$15
8	2015 年 8 月	駭客任務	US	$12
9	2015 年 9 月	駭客任務	India	$12
10	2015 年 9 月	聖杯傳奇	US	$9
11	2015 年 9 月	亂世佳人	US	$15
12	2015 年 9 月	聖杯傳奇	India	$9
13	2015 年 11 月	亂世佳人	US	$15
14	2015 年 12 月	聖杯傳奇	US	$9
15	2015 年 12 月	聖杯傳奇	US	$9

資料倉儲

電影銷售資料倉儲			
橫列 #	銷售量	產品名稱	金額
1	Q2	亂世佳人	$15
2	Q2	聖杯傳奇	$30
3	Q3	亂世佳人	$30
4	Q3	駭客任務	$36
5	Q3	聖杯傳奇	$30
6	Q4	亂世佳人	$15
7	Q4	聖杯傳奇	$18

▼ 表 1.1：資料庫系統與資料倉儲的比較

層面	資料庫	資料倉儲
目的	儲存在資料庫中的資料可用於許多用途，包括日常作業	儲存在 DW（資料倉儲）中的資料乃是淨化過的資料，適用於報表與分析
詳盡程度	包括所有活動與交易細節的高詳盡度資料	低詳盡度的資料；向上彙整至特定重點興趣層面
複雜度	高度複雜，存在數十或數百個以共同資料欄位互相連結的資料檔案	通常組織成一個大型事實資料表，以及許多查找表
大小	資料隨著活動與交易量的成長而增加。舊的完成交易會被刪除以降低大小。	隨著每日從作業資料庫而來的資料向上彙整並附加進來而成長，資料會保留下來供長期趨勢分析
架構選擇	關聯式、物件導向的資料庫	星型架構（Star schema），或雪花型架構（Snowflake schema）
資料存取機制	主要透過 SQL 這類高階語言。傳統程式存取 DB 是透過開放資料庫互連（Open DataBase Connectivity，ODBC）介面	透過 SQL 存取；SQL 輸出會轉送至報表工具以及資料視覺化工具

資料探勘(Data Mining)

Data Mining is the art and science of discovering **useful innovative patterns** from data.

季度電影銷量－交叉表列				
數量／產品	亂世佳人	駭客任務	聖杯傳奇	合計銷售額
Q2	$15	0	$30	$45
Q3	$30	$36	$30	$96
Q4	$15	0	$18	$33
合計銷售額	$60	$36	$78	$174

What is the best-selling movie by revenue? – Monty Python.
What is the best quarter by revenue this year? – Q3
Any other patterns? – Matrix movie sells only in Q3 (seasonal item).

- **Different patterns can be noticed by analyzing data in different ways**

- Q: What is the best-selling geography? – US
- Q: What is the worst selling geography? – UK
- Any other patterns? – Monty Python sells globally, while Gone with the Wind sells only in the US.

- Selecting Data Mining Projects

Data mining should be done to solve **high-priority, high-value problems**. Much effort is required to gather data, clean and organize it, mine it with many techniques, interpret the results, and find the right insight. It is important that there be a large expected payoff from finding the insight.

One should **select the right data** (and ignore the rest), organize it into a nice and imaginative framework that brings relevant data together, and then apply data mining techniques to deduce the right insight.

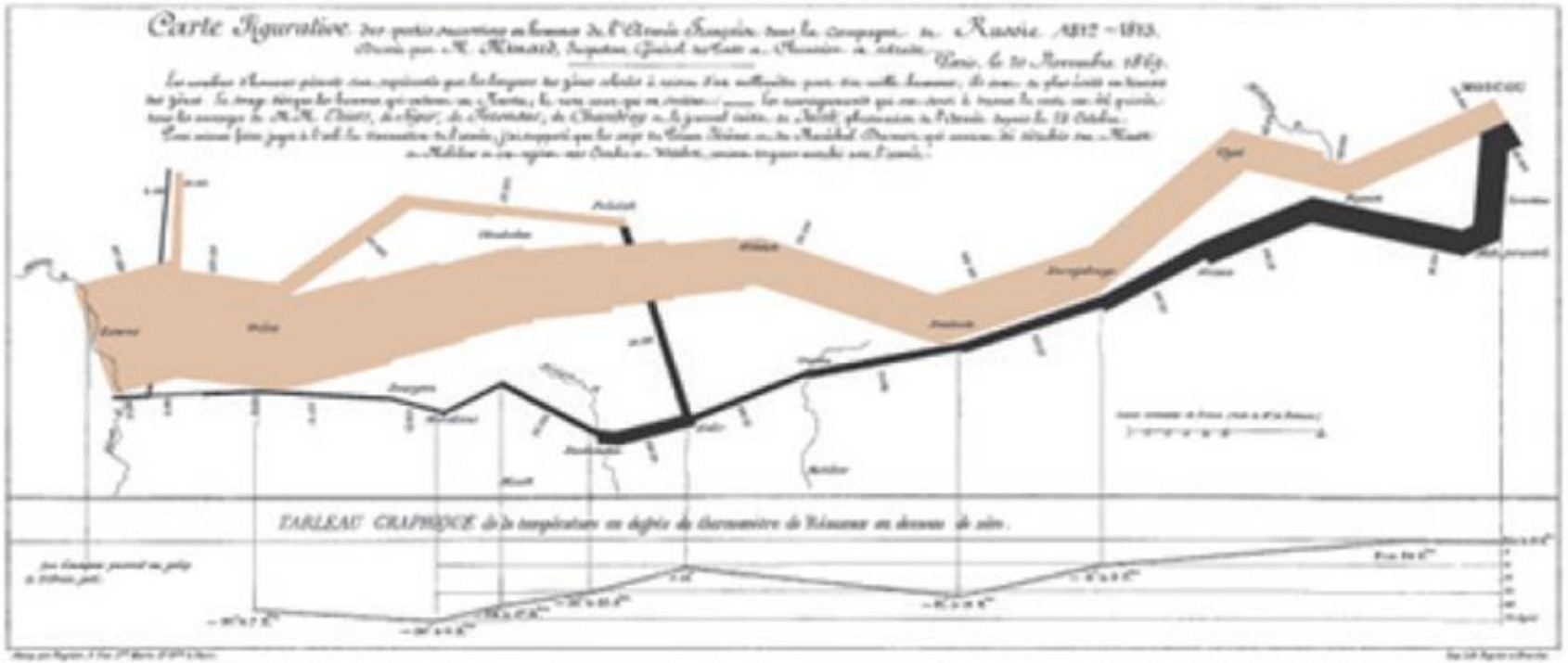
Data Mining Techniques

- **Decision Trees:** It is said that 70% of all data mining work is about classification solutions; and that 70% of all classification work uses decision trees. The most popular and important data mining technique.
- **Regression:** This is a well-understood technique from the field of statistics. The goal is to find a best fitting curve through the many data points. The best fitting curve is that which minimizes the (error) distance between the actual data points and the values predicted by the curve. Regression models can be projected into the future for prediction and forecasting purposes.
- **Artificial Neural Networks:** Originating in the field of artificial intelligence and machine learning, ANNs are multi-layer non-linear information processing models that learn from past data and predict future values. Neural networks are opaque like a black-box. These systems also require a large amount of past data to adequate train the system.
- **Cluster analysis:** This is an important data mining technique for dividing and conquering large data sets. The data set is divided into a certain number of clusters, by discerning similarities and dissimilarities within the data. There is no one right answer for the number of clusters in the data. This is most commonly used for market segmentation. Unlike decision trees and regression, there is no one right answer for cluster analysis.
- **Association Rule Mining:** Also called Market Basket Analysis when used in retail industry, these techniques look for associations between data values. An analysis of items frequently found together in a market basket can help cross-sell products, and also create product bundles.

資料視覺化 Data Visualization

▲ 圖 1.3：管理者儀表板範例

- The figure(1812) covers about **six dimensions**. Time is on horizontal axis. The geographical coordinates and rivers are mapped in. The thickness of the bar shows the number of troops at any point of time that is mapped. One color is used for the onward march and another for the retreat. The weather temperature at each time is shown in the line graph at the bottom.



Napoleon's March to Moscow The War of 1812

Charles Joseph Minard

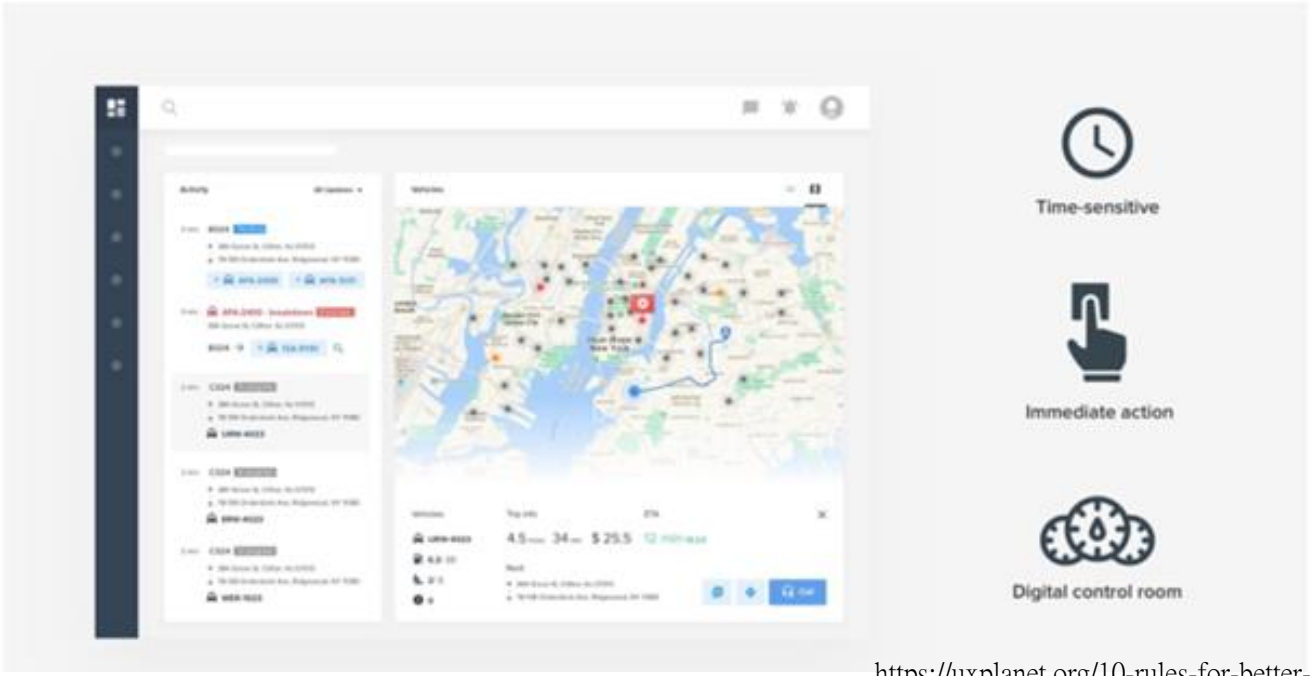
Dashboard design
10 rules for better design by Taras Bakusevych

1. Define the purpose of the dashboard

Operational dashboard

Operational dashboards aim to impart critical information quickly to users as they are engaged in **time-sensitive tasks**. Main goals of the operational dashboard are to present data deviations to the user quickly and clearly, show current resources, their status. View support actions, it's a digital control room designed to help users be quick, proactive, and efficient.

Operational dashboard key qualities : Time-sensitive and Immediate action

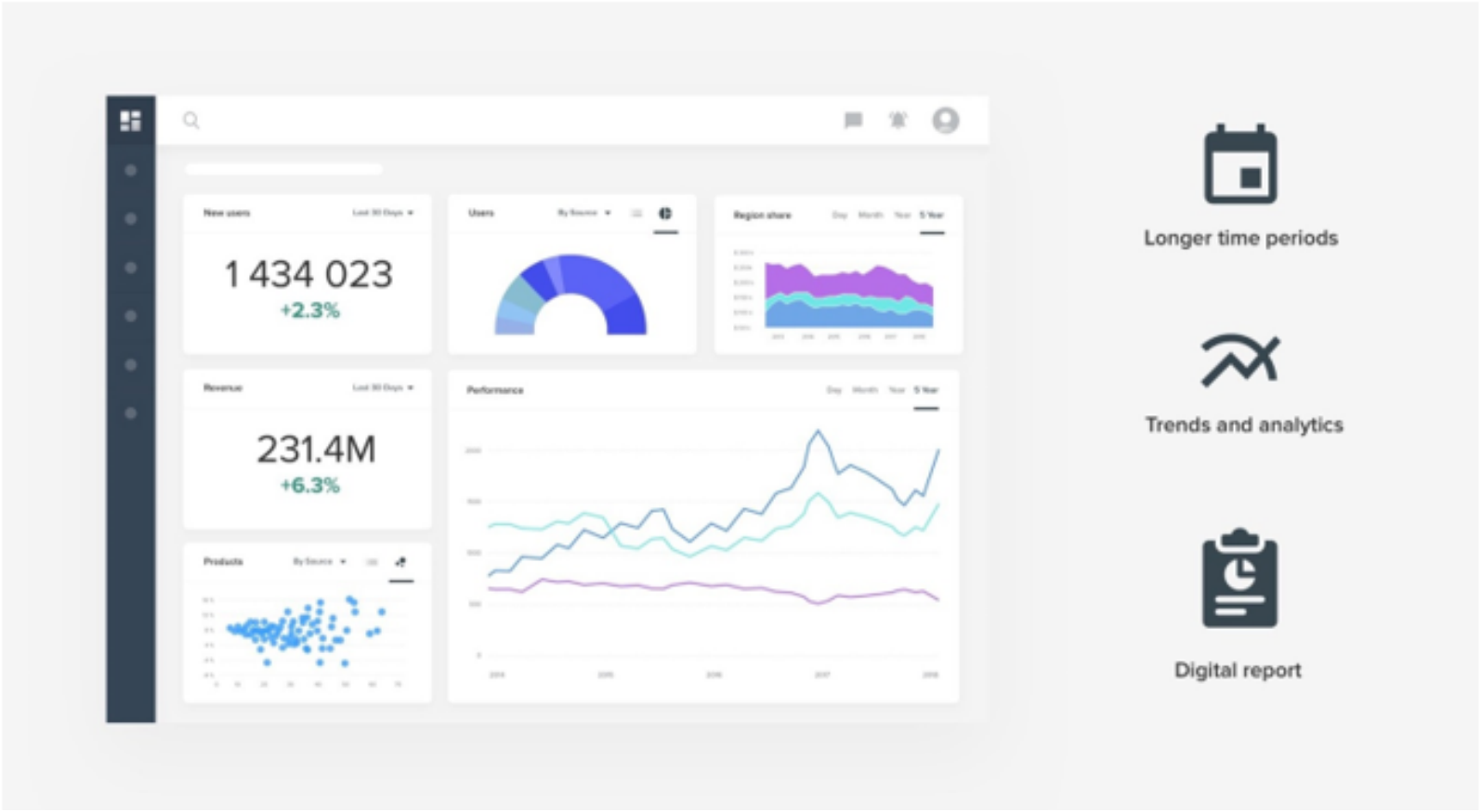


https://uxplanet.org/10-rules-for-better-dashboard-design-ef68189d734c

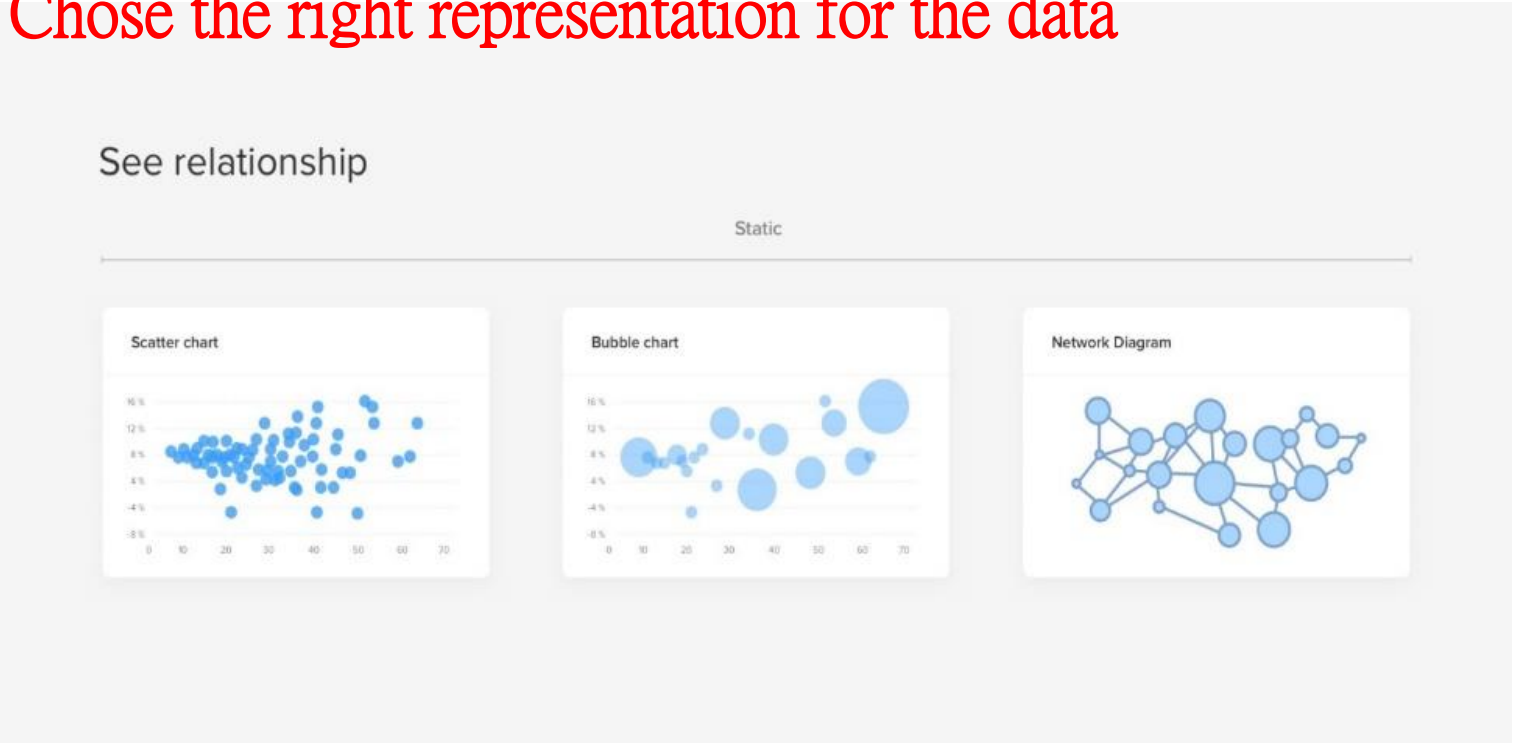
Analytical dashboard

In contrast to Operational, Analytical dashboards provide the user with at-a-glance information used for analysis and decision making, and are less time sensitive and not focused on immediate action. A primary goal is to help users make the best sense of the data, analyze trends and drive decision making.

Analytical dashboard key qualities: Longtime Periods and Trends/analytics

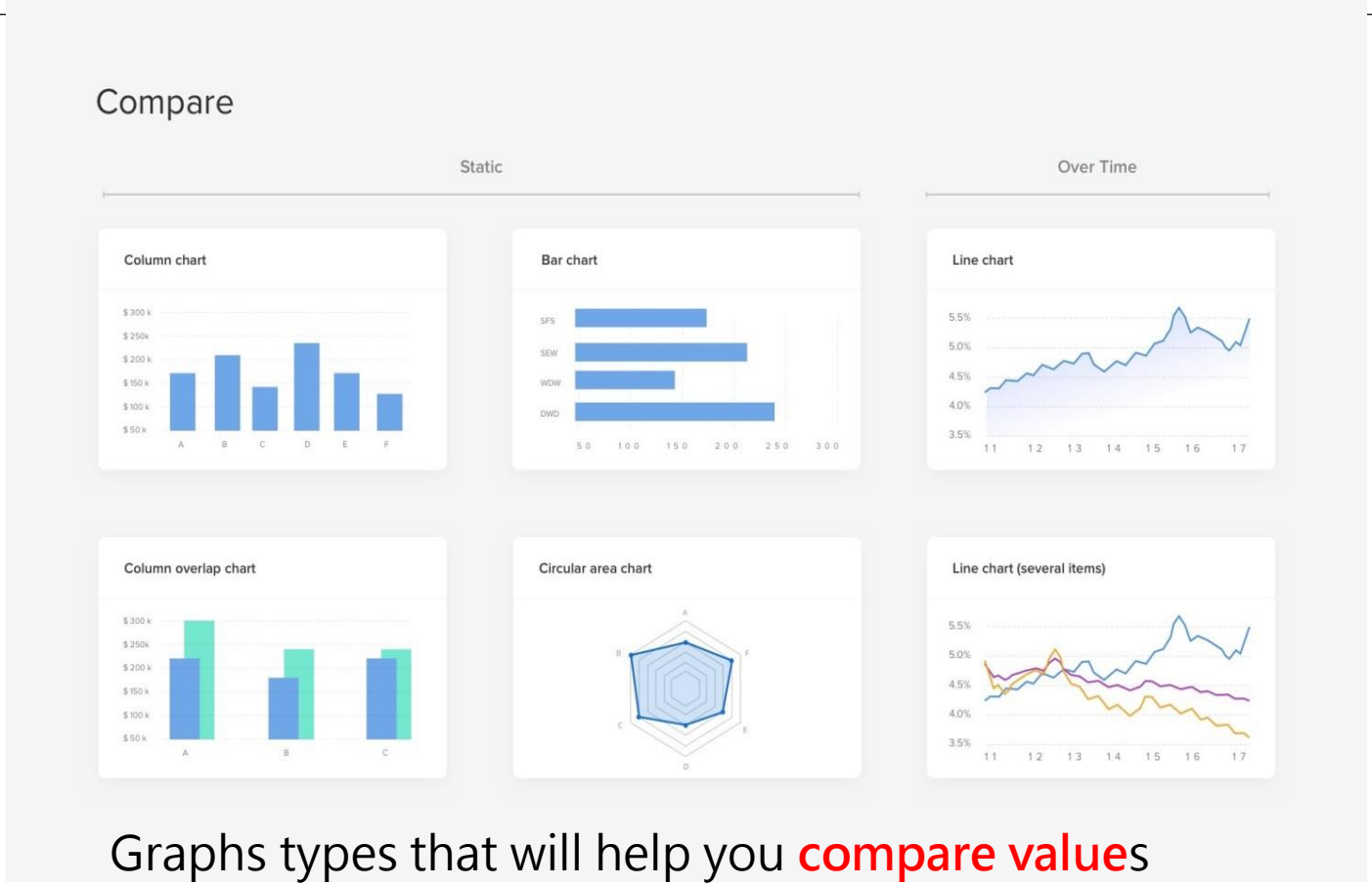


2. Chose the right representation for the data



Graphs types that will help you **see relationship** in data

Scatter charts are primarily used for correlation and distribution analysis. **Bubble chart** helps introduce the third dimension into the chart. **Nnetwork diagram** is handy when even the most connections between data points are very important.

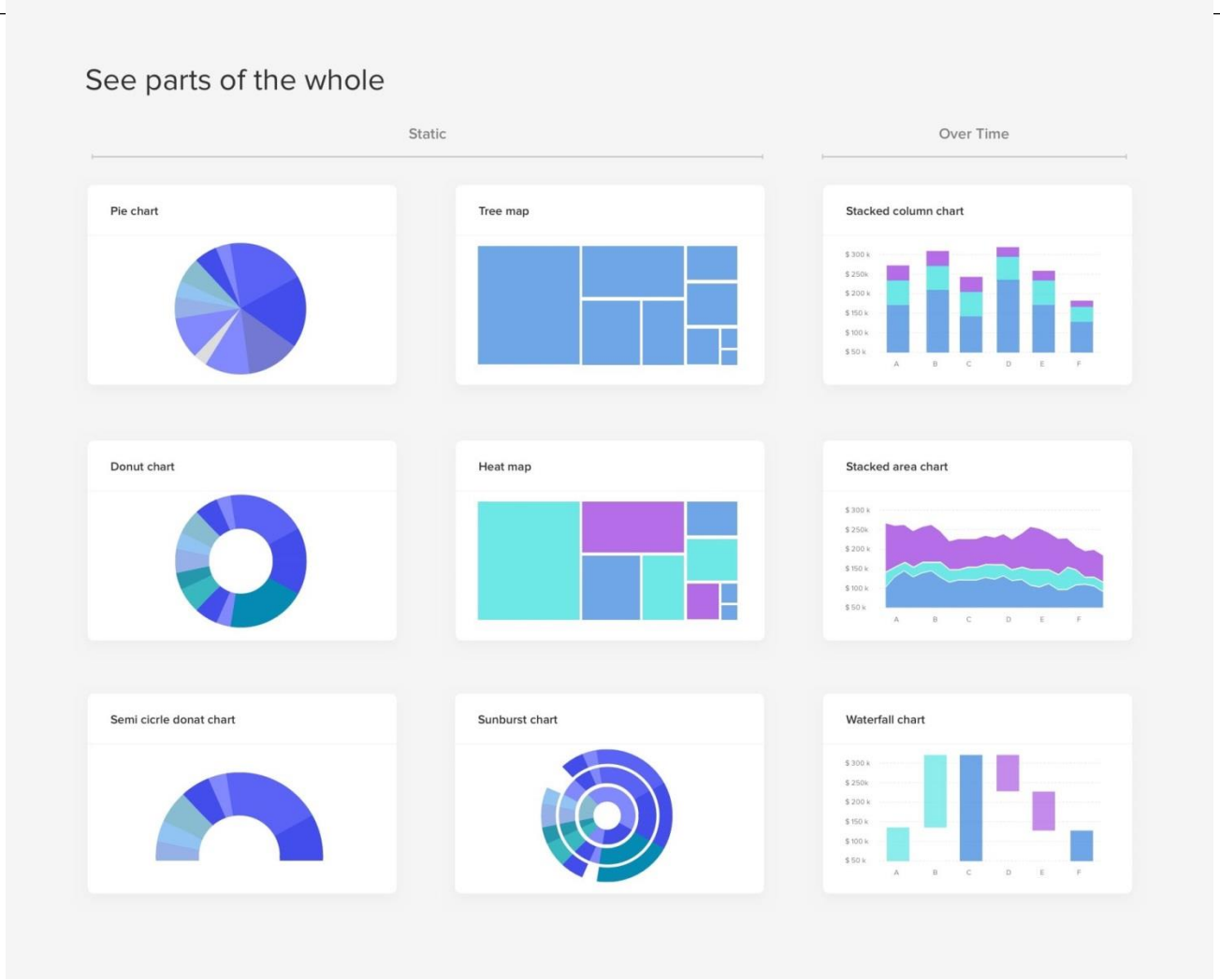


Graphs types that will help you compare values

Using visualization to **compare** one or many values sets is so much easier than looking at numbers in the grid. Column and line charts are probably the most used ones.

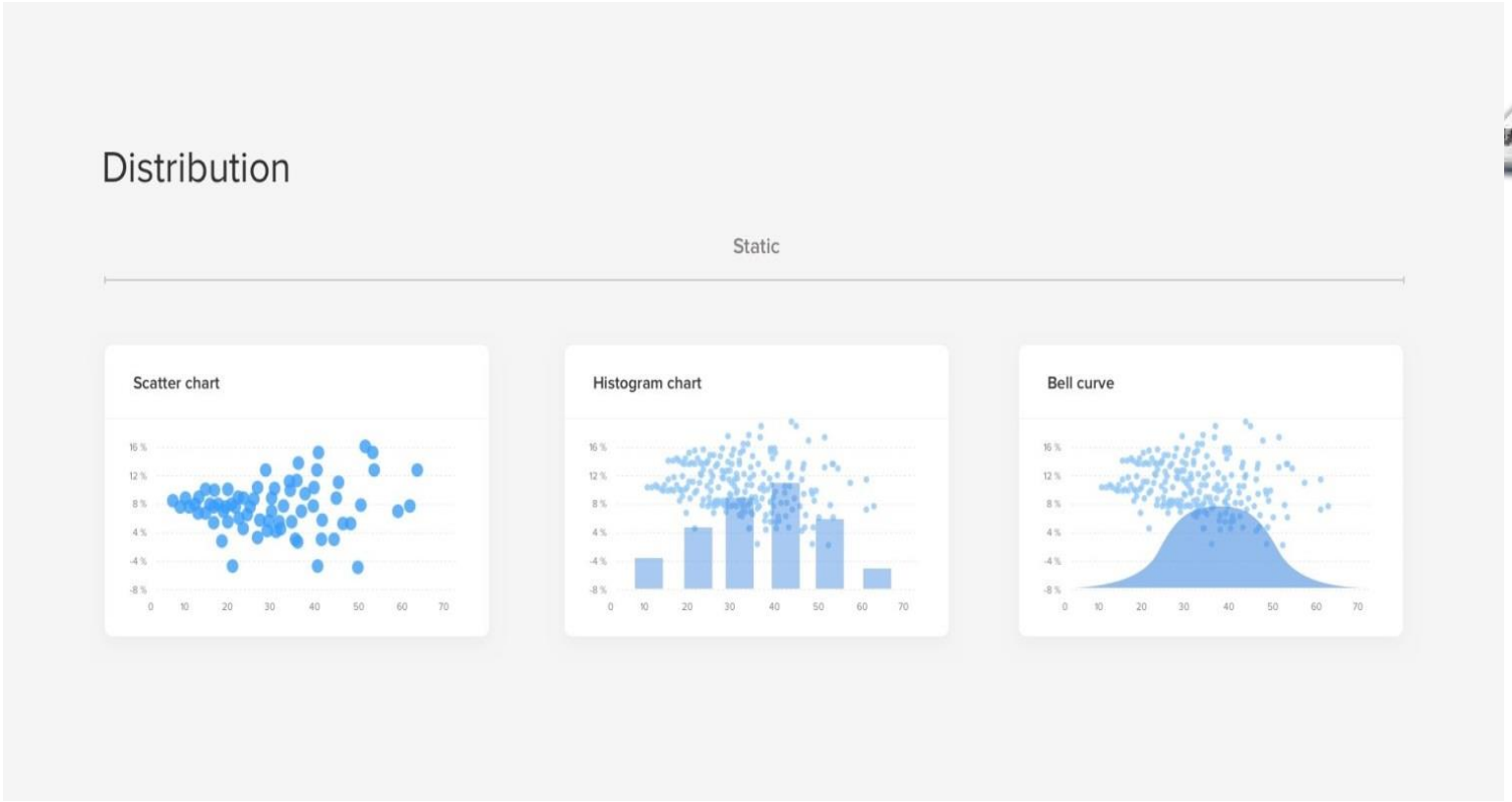
Some recommendations: When one of your dimensions is a time it should always be an axis X, time in charts flows from left to right

- When using a horizontal or vertical bar chart, try to sort column by biggest value not randomly
- With the line, charts don’ t show more than 5 values and with bar charts, it’ s not recommendable to show more than 7



Graphs types that will help you see **composition**

Pie and Donut charts have a bad reputation for data visualization. These charts are among the most frequently used and also misused charts. They are quite bad to read when there are too many components, very similar values. It' s hard for humans to differentiate values when it comes to angles and areas.

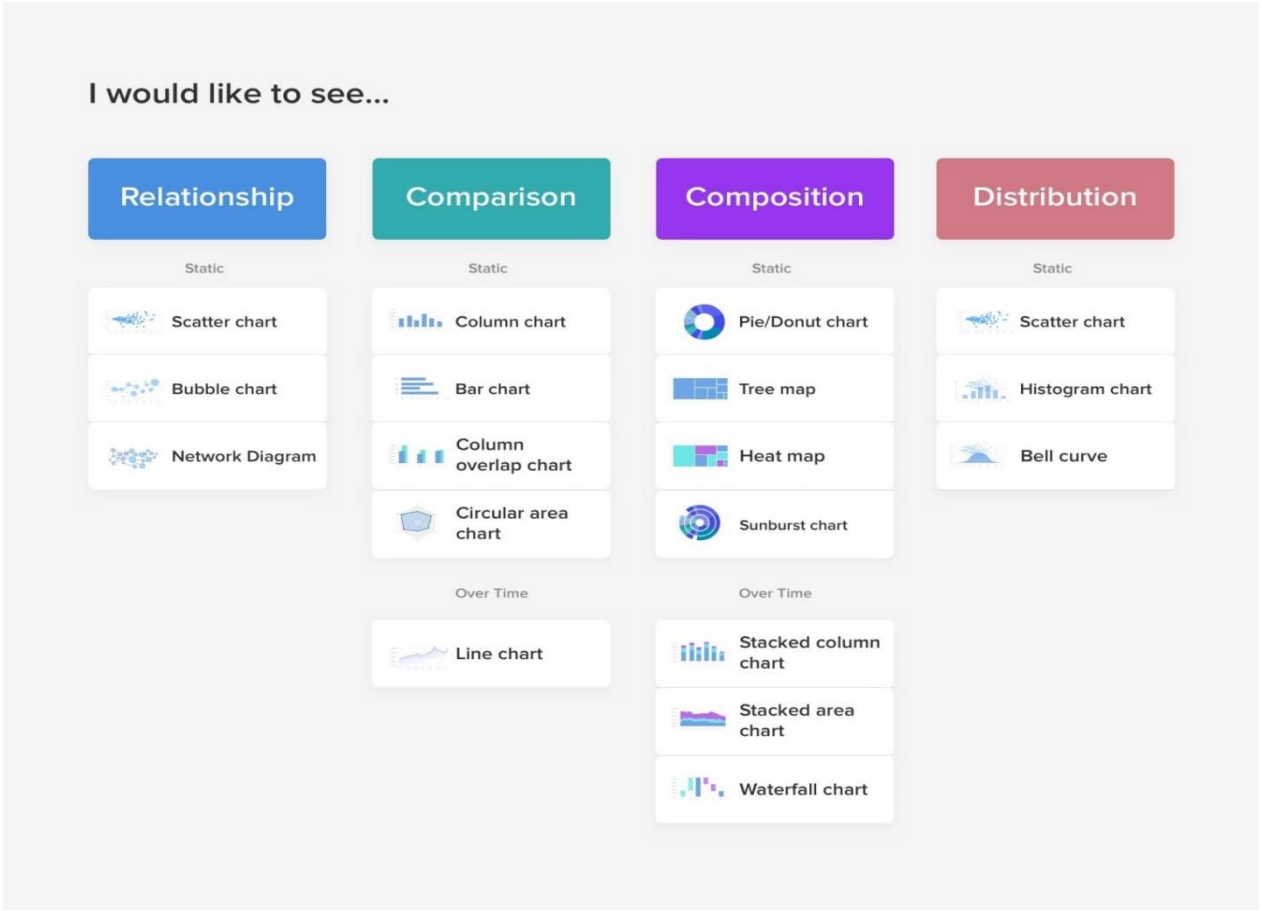


Graphs types that will help you **see distribution**

Distribution charts help you to understand outliers, the normal tendency, and the range of information in your values.

Charts types to **avoid**

But certain chart types you need to avoid at all. Gauge was a big deal for dashboards in the past, trying to replicate physical object digitally is a bad idea. 3D charts and over styled charts have lover readability, distract the viewer from data and even harder to develop, so little reason to go there.



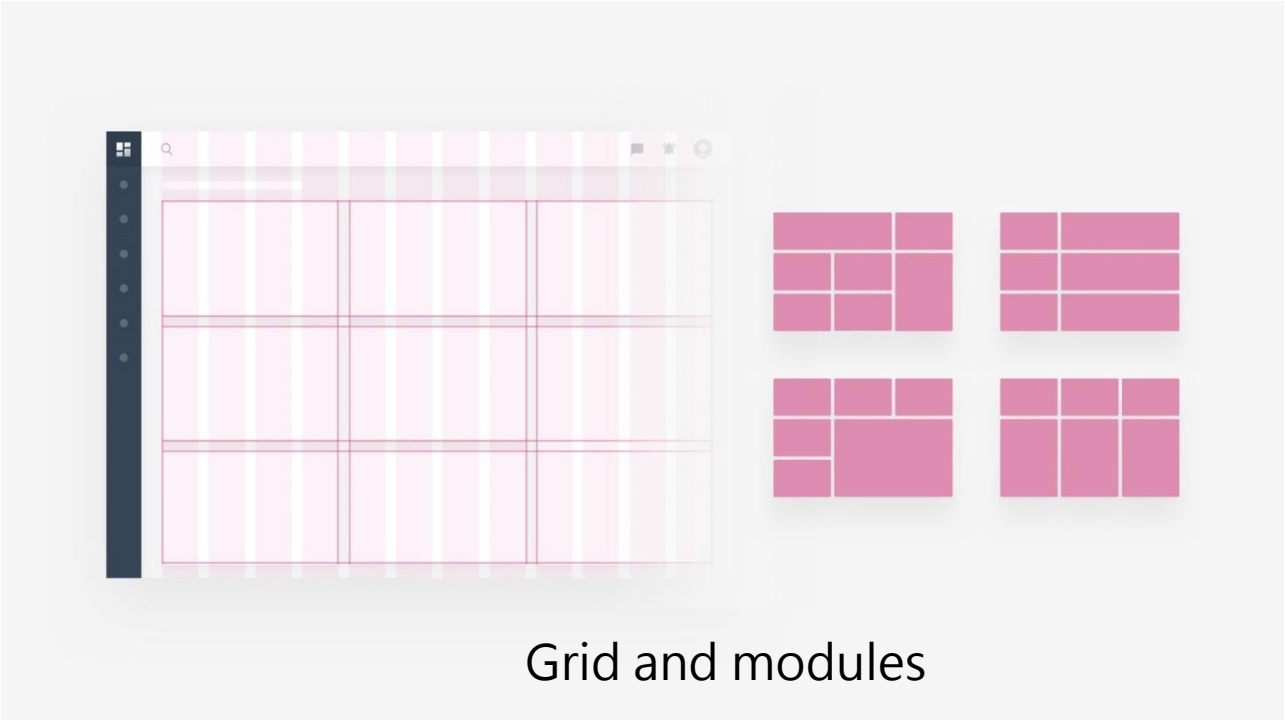
When to use various graph types
To help you choose the right representation type for the chart.
Ask yourself this questions:
- How many variables do you want to show in a single chart?
- Will you display values over a period of time, or among items or groups?
- How many data points we need to display for each variable?

3. Clear and consistent naming convention, Consistent date formatting, truncate large values

As the main goal of the dashboard is to get the message at a glance, every little thing counts. The biggest benefit of using a clear framework is data consistency. If your data is named the same way in each tool, it’ s easier for you to use those tools.

4. Define the layout.

Grids help you to achieve effective alignment and consistency with little effort, create a basic structure, a skeleton for your design. They consist of “invisible” lines upon which your design elements can be placed. Doing so ties them together in an overall “system” and supports your composition—rationally. That is crucial for dashboard design as you will need to organize a ton of information in a seamless way.



Grid and modules

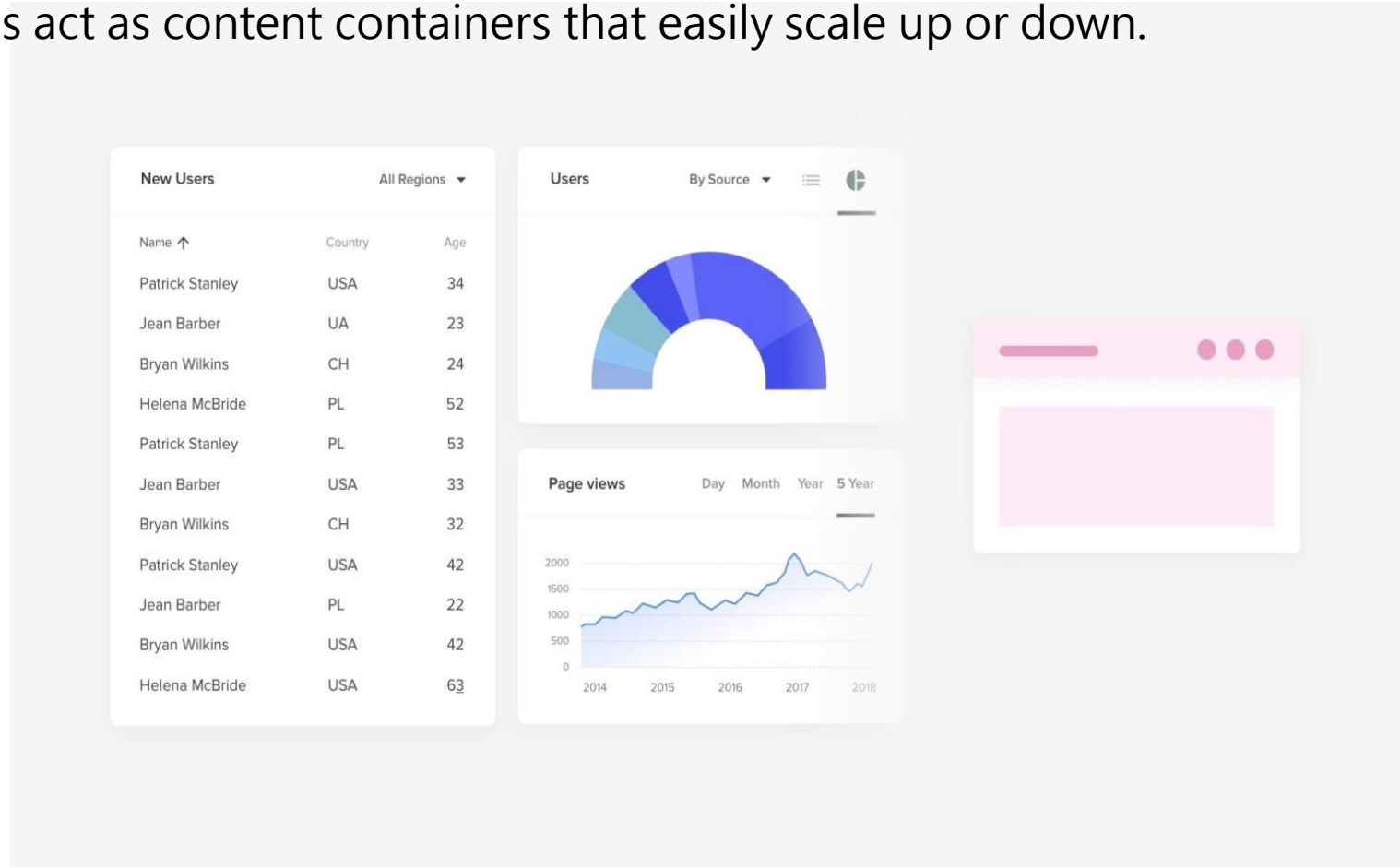
When making decisions on what information should go where, keep this in mind:

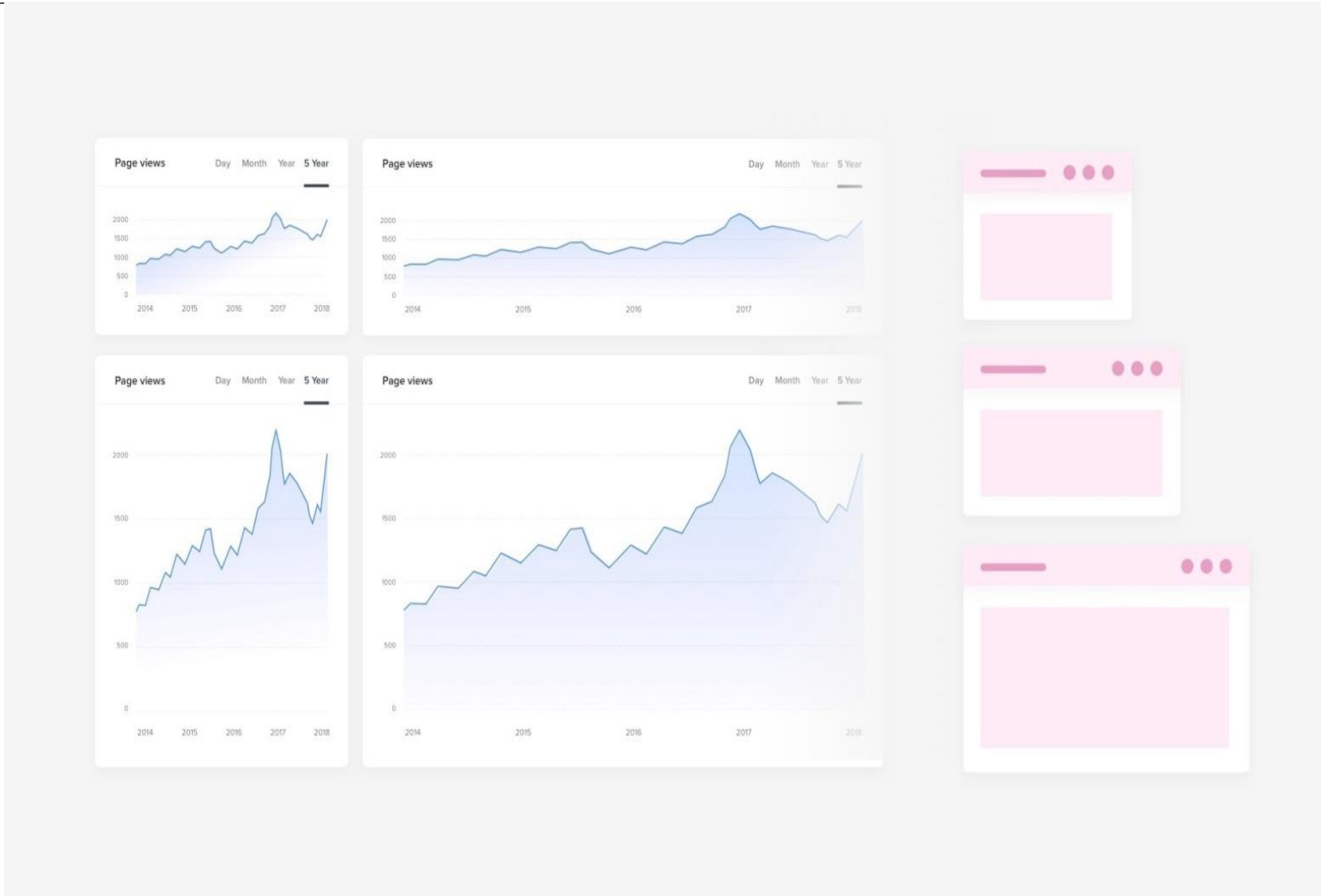
- Naturally top left corner of the screen will get more attention so try to position key info from left to right (this may be the opposite way, based on the region you making a design for) this is based on the way we read information. And when they will finish with the first row, they move down to the next one.
- If there are dependencies that will affect decisions making on one group of information from based on info from another, create a layout in a way so users do need to go back and forth, create a continuous flow for easy scanning

5. Use building blocks with consistent structure

After we defined the grid, we can start work with multiple "widgets" that will hold the info, charts, and controls. Cards sweet and tasty, easy to arrange.

The most important thing about cards is that they are almost infinitely manipulatable. They are a good choice for responsive design since cards act as content containers that easily scale up or down.

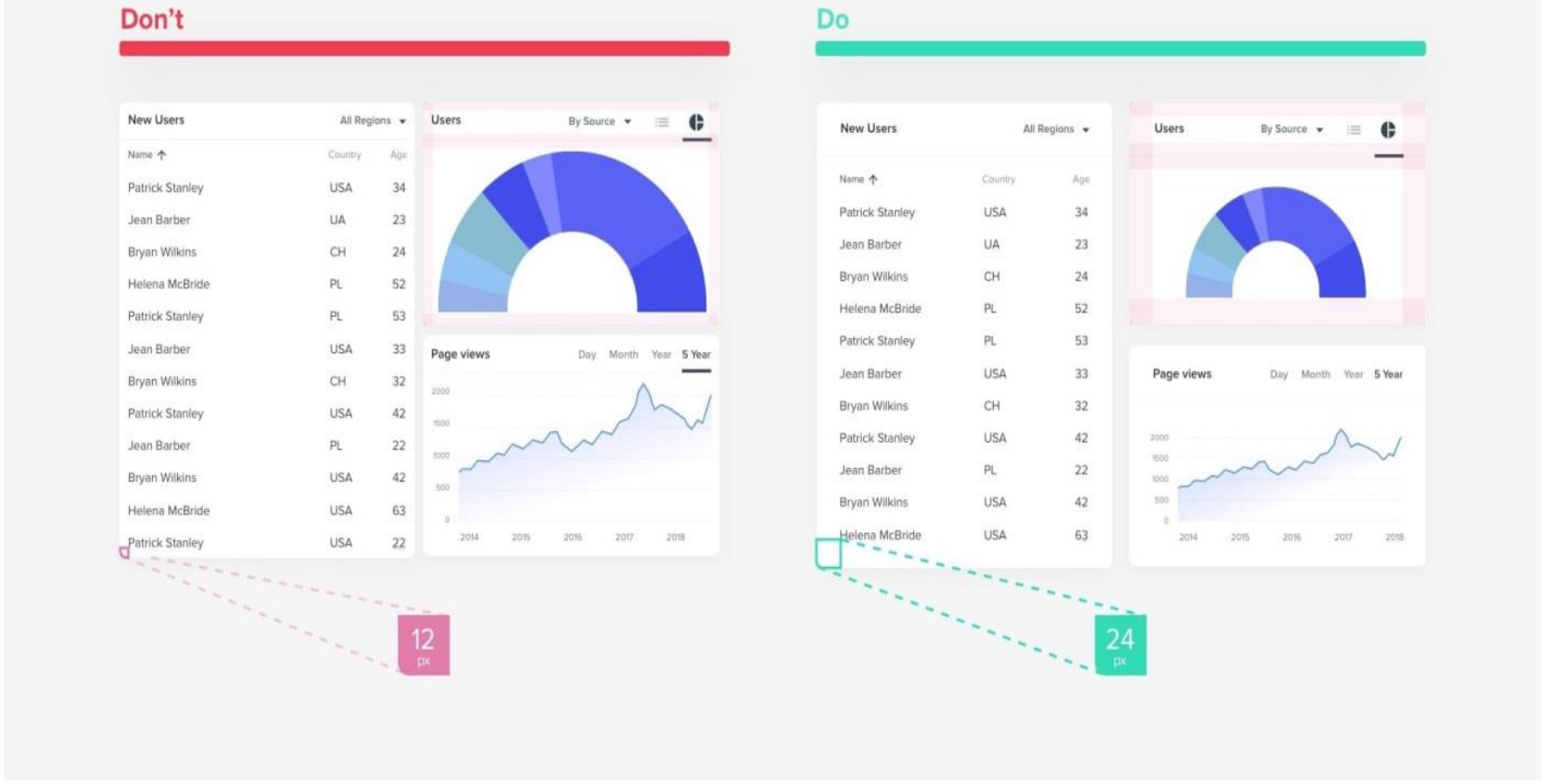




An important characteristic of cards is the consistent layout of controls and data inside. Put the name in the top left corner, align view controls or actions to in the top right corner of the card, leave the rest for the content. When all have a consistent structure, it' s easier for the user to work with the interface, he finds everything where he expects it.

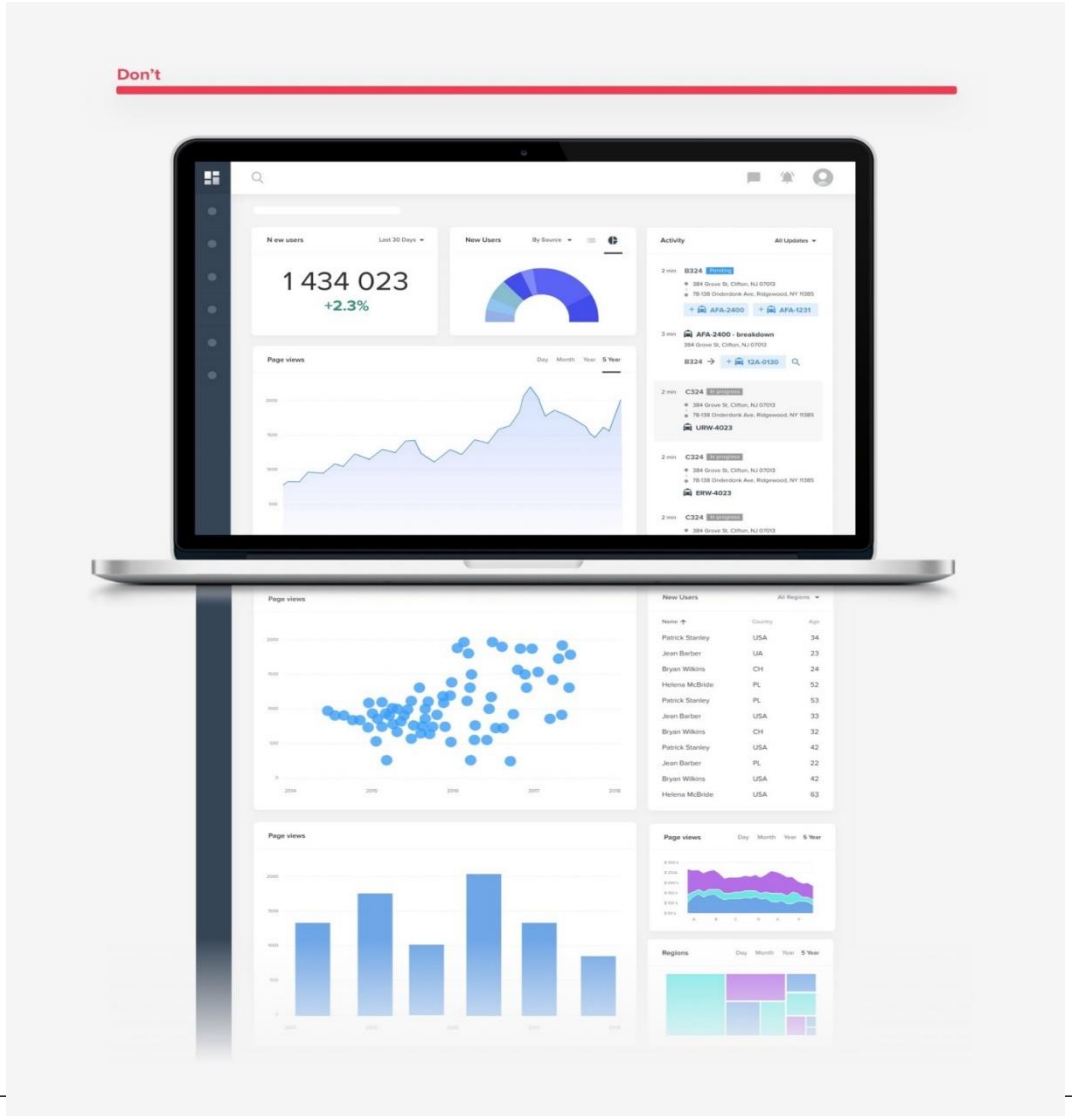
6. Double your margins

White space, also known as negative space, is the area between elements in a design composition. Readers aren’ t usually aware of the great role of the space, but designers pay a lot of attention to it. In case the white space is not balanced, a copy will be hard to read. That’ s why negative space matters as much as any other typography element.



7. Don’ t hide information, or rely on interactions too much

As one of the primary goals of the dashboard is to surface information at a glance, relying on scrolling or many interactions dilutes the whole purpose.

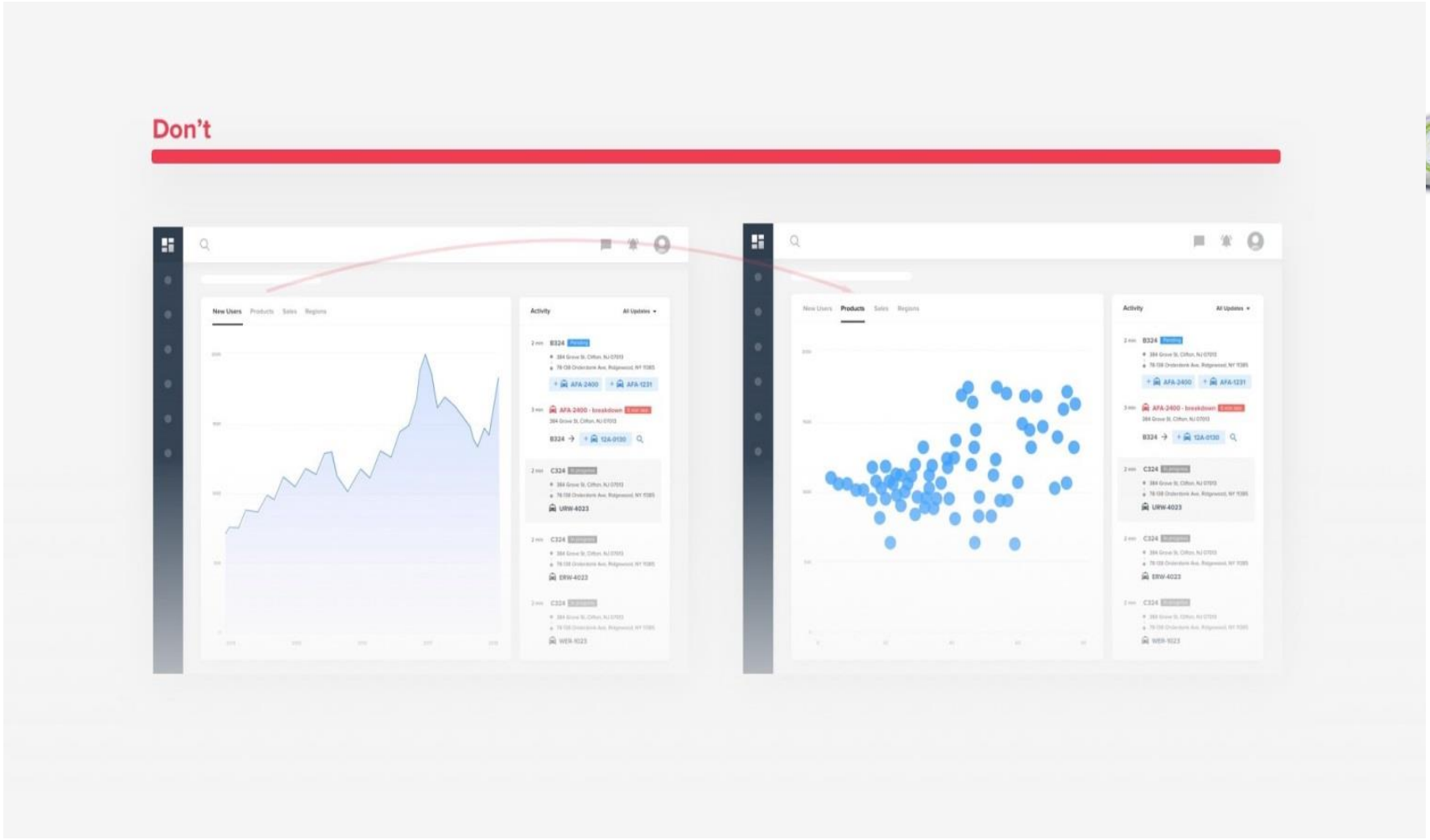


Designing long scrollable dashboards is one of the most frequent mistakes designers make.

They try to display more information in a clear way, positioning it one under another to not overwhelm the user. This leads to only information visible above the screen fold to be discovered.

Everything below gets little attention from users. So what’ s the point? The solution is prioritization, after doing more research and interviews you should be able to identify core information, work only with space above the fold to display it.

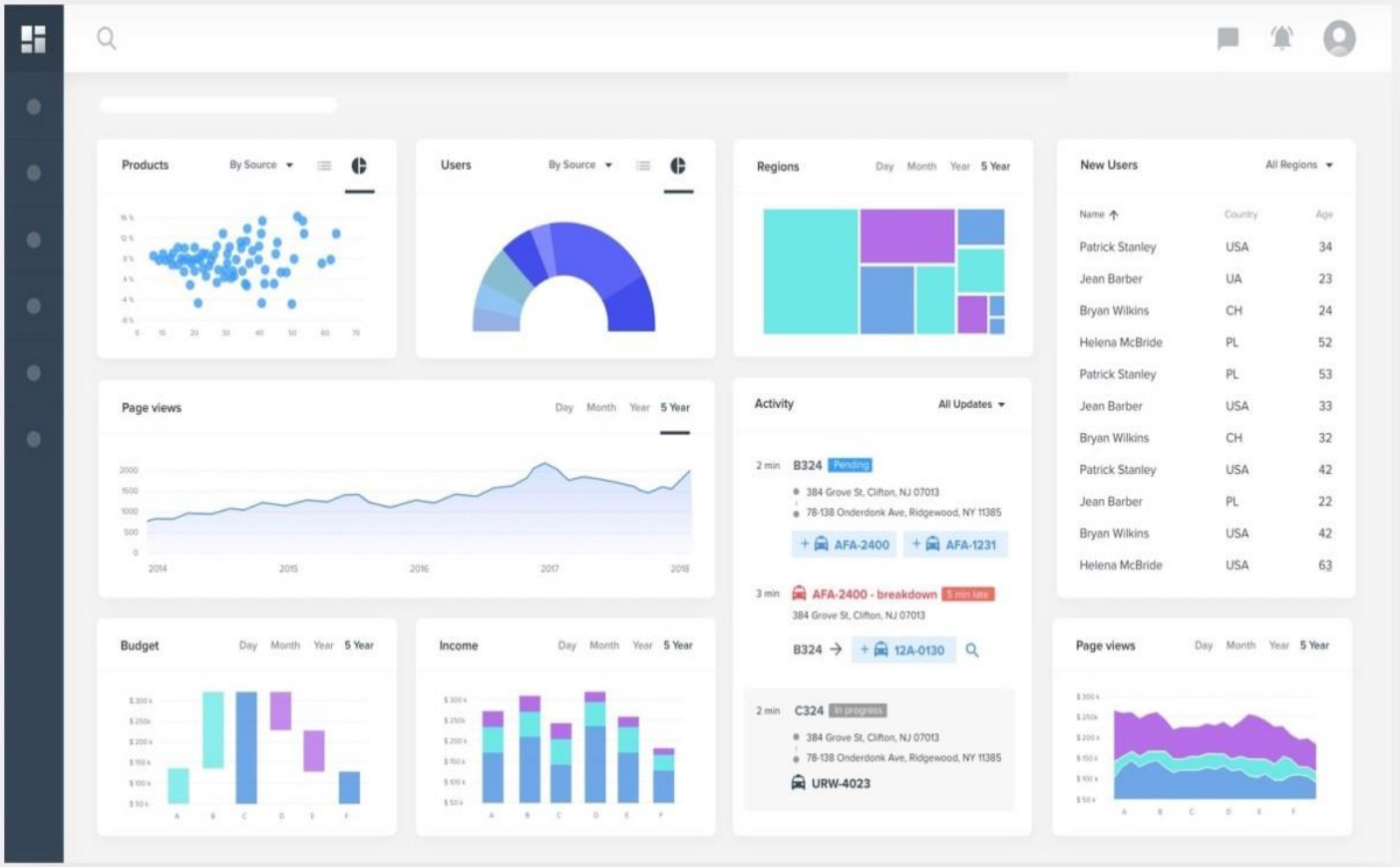
Don’ t tell the full story, instead summarize, surface only key info. You can use additional interactions as a way to fit more content, and not overwhelm the user with data.



Don' t rely on too many interactions to surface information

Interactions help surface secondary information. Fully relying on them as the main way to work with the dashboard is a big mistake. In the example above we see how a user will have to painfully switch between multiple tabs to get the full picture. This renders information on all other tabs hidden from user same way like content below the fold.

Don't



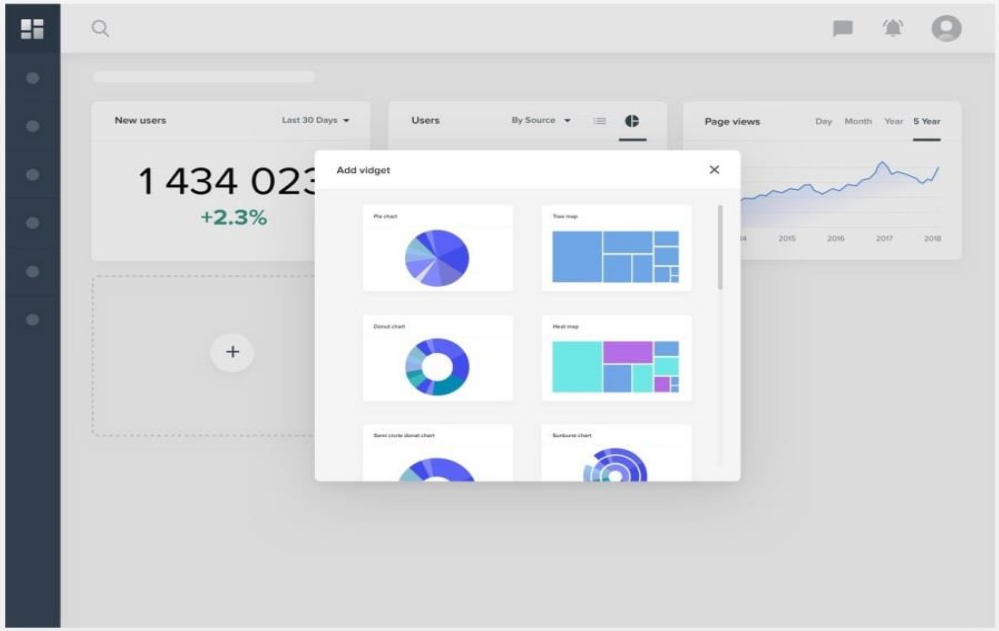
Data overloaded dashboard example

Use max 5–7 different widgets to create a view. Otherwise, it will be hard for a user to focus and get a clear overview.

8. Personalization rather than Customization

Users expect that the content they see will be relevant to their individual needs. Personalization and customization are techniques that can help you ensure that users see what matters to them.

Personalization is done by the system itself. System is set to identify users and deliver to them the content, experience, or functionality that matches their role. Customization is done by the user. A system may enable users to customize or make changes to the experience to meet their specific needs by configuring layout, content, or system functionality.



9. When integrating data tables or list, make sure they are interactive and data is aligned correctly

A data table is a great solution when you need to show a lot of information for a big number of items. For example list of clients with their ID, status, contacts, last activity etc… would be best displayed as a data table.

There are many other benefits, like a great use of space, easy scalability, easy in development, user comfortable working with grids (majority already working with Excel for many years), an easy way to find and change something.

10. Design the dashboard in the end

As the dashboard is one of the most visually exciting views, it’ s often a first thing that is being designed. I would recommend the opposite.

A dashboard is a summary view of everything else, display a key info from various part of the application, it’ s just more practical to design it the end. Otherwise, you will need to constantly go back and update your dashboard designs while you are working on all other pages. Also once a majority of the views is designed, you have a ton of components to work with when putting together a dashboard.

Student-level

PROFILES

ACHIEVEMENTS

FOCUS

ACTIVITY REPORT

宏觀層面(班級) → 微觀層面 (個人)

BI tools for better decisions

▲ 圖 2.2：管理者儀表板範例

IBM

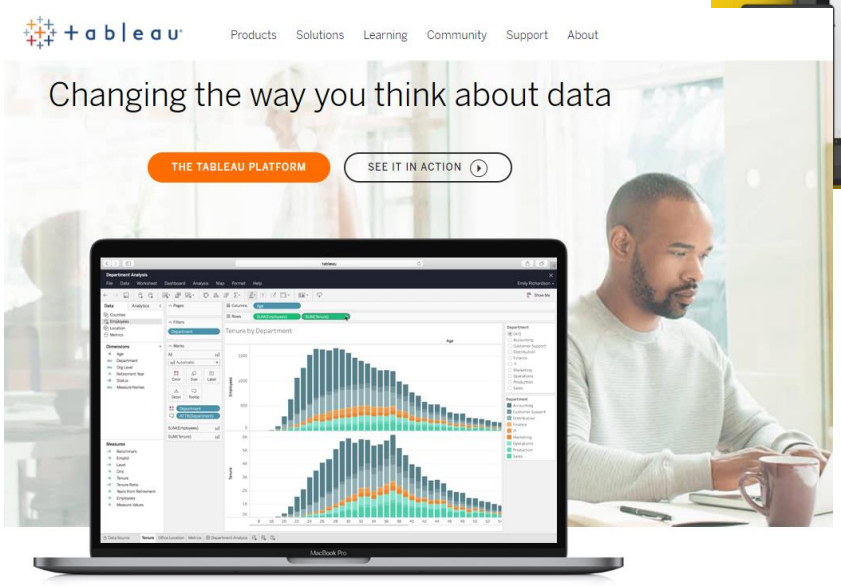
IBM Analytics 行業 技術 商業

Cognos Analytics

新一代商業智慧軟體

圖像解析，智取商機

立即試用 操作 demo (01:38)



- 商業智慧 (BI) 工具

應用程式軟體的各式類型，收集及處理大量的**非結構化資料**，範圍包括書籍、報章雜誌、文件、健康記錄、影像、檔案、電子郵件、視訊及其他商務來源。

BI工具可協助準備資料進行分析，建立報表、儀表板及資料視覺效果。成果可讓員工及經理都能改善決策並加快其制定速度、提升運作效率、直接點出新的收益潛力、找出市場趨勢、回報真實的 KPI，並找出新的商機。

商業智慧工具通常使用更為直接的商務資料查詢及報告，並結合一組廣泛的資料分析應用程式，內含隨選分析及查詢、企業報表、線上分析處理 (OLAP)、行動 BI、即時 BI、作業 BI、雲端及軟體即服務 BI、開放原始碼 BI、共同作業 BI 及位置智慧。它還可以內含設計圖表用的資料視覺效果軟體，以及用來建置 BI 儀表板及績效計分卡的工具，進而顯示商務計量及 KPI，將公司資料以易於理解的視覺效果融入生活。

BI 協助做出更好決策

目的就是為了做出有效的決策，並且降低風險。企業依據廣泛的事實與見解來計算風險並制定決策。關於未來的可靠知識，將有助於管理者在低風險下做出正確的決策。

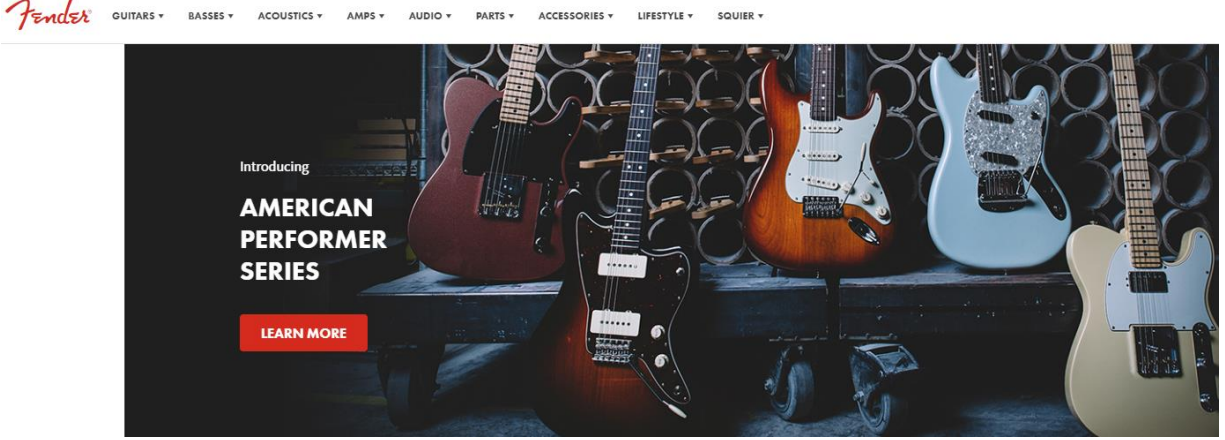
BI決策類型

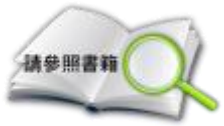
- **策略決策(strategic decision):** impact the direction of the company, how to reach out to a new customer set.
- **作業決策(operational decision):** routine and tactical decision, focused on developing greater efficiency. Such as updating old website with new features.

BI 應用

- 客戶關係管理

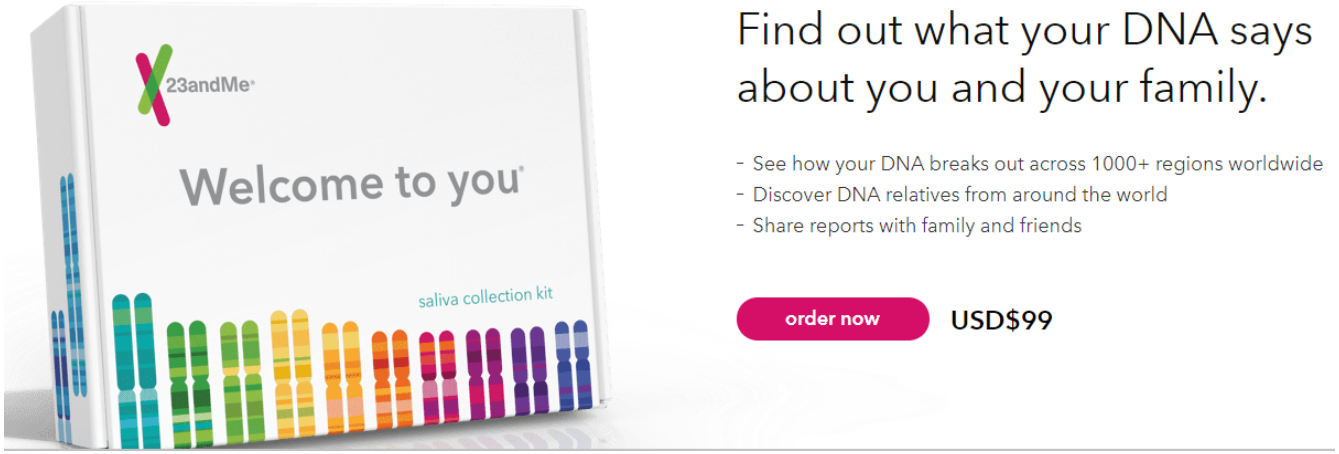
1. 極大化行銷活動的回報
2. 增進客戶留存率（流失分析）
3. 極大化客戶價值（交叉、追加銷售）
4. 找出高價值客戶，並滿足他們
5. 管理品牌形象





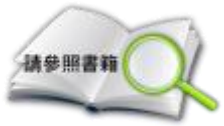
醫療保健與健康

1. 診斷疾病
2. 治療有效性
3. 健康管理
4. 濫用管理
5. 公共衛生管理

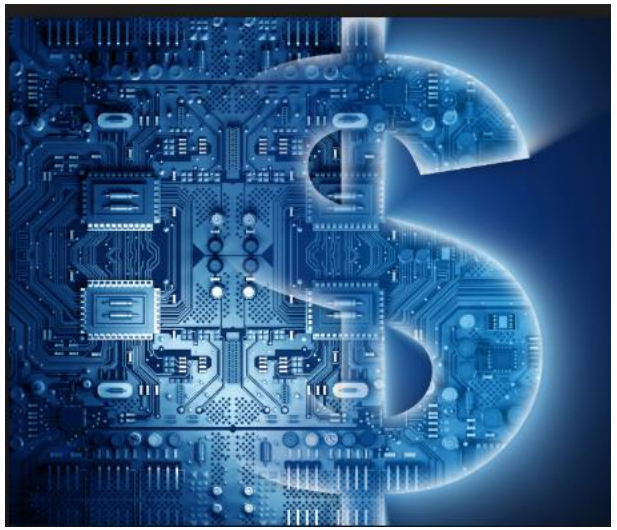


- 教育
  1. 學生註冊人數（招募與保留）
  2. 課程提供
  3. 從校友及其他捐贈者處募款

- 零售業
  1. 最佳化不同區域的庫存水準
  2. 改善商店陳列與銷售宣傳
  3. 為季節效應安排最佳物流
  4. 減少因有限賞味期的損失



- 銀行業
  1. 自動化借貸申請流程
  2. 偵測詐欺交易
  3. 極大化客戶價值（交叉、追加銷售）
  4. 運用預估做出最佳現金準備



- 金融服務
  1. 預測債券與股票價格的變動
  2. 評估事件影響對市場造成的移動
  3. 辨識與避免交易中的舞弊行為

- 保險業
  1. 預估索賠成本以利更好的商業規劃
  2. 決定最佳利率計劃
  3. 對特定客戶進行最佳行銷
  4. 發現並避免詐領行為

- 製造業
  1. 發掘新穎模型以增進產品品質
  2. 預測 / 避免機械故障

- 電信業
  1. 客戶流失管理
  2. 行銷與產品規劃
  3. 網路故障管理
  4. 詐騙管理

- 公共區域
  1. 執法
  2. 科學研究

好的資料探勘專案都是從解決有趣的問題開始!!

選擇正確的資料探勘問題是一項重要的技能!!

It takes a lot of time and energy to gather, organize, cleanse, and prepare the data for mining and other analysis.

The data miner needs to persist with the exploration of patterns in the data.

The skill level has to be deep enough to engage with the data and make it yield new useful insights.