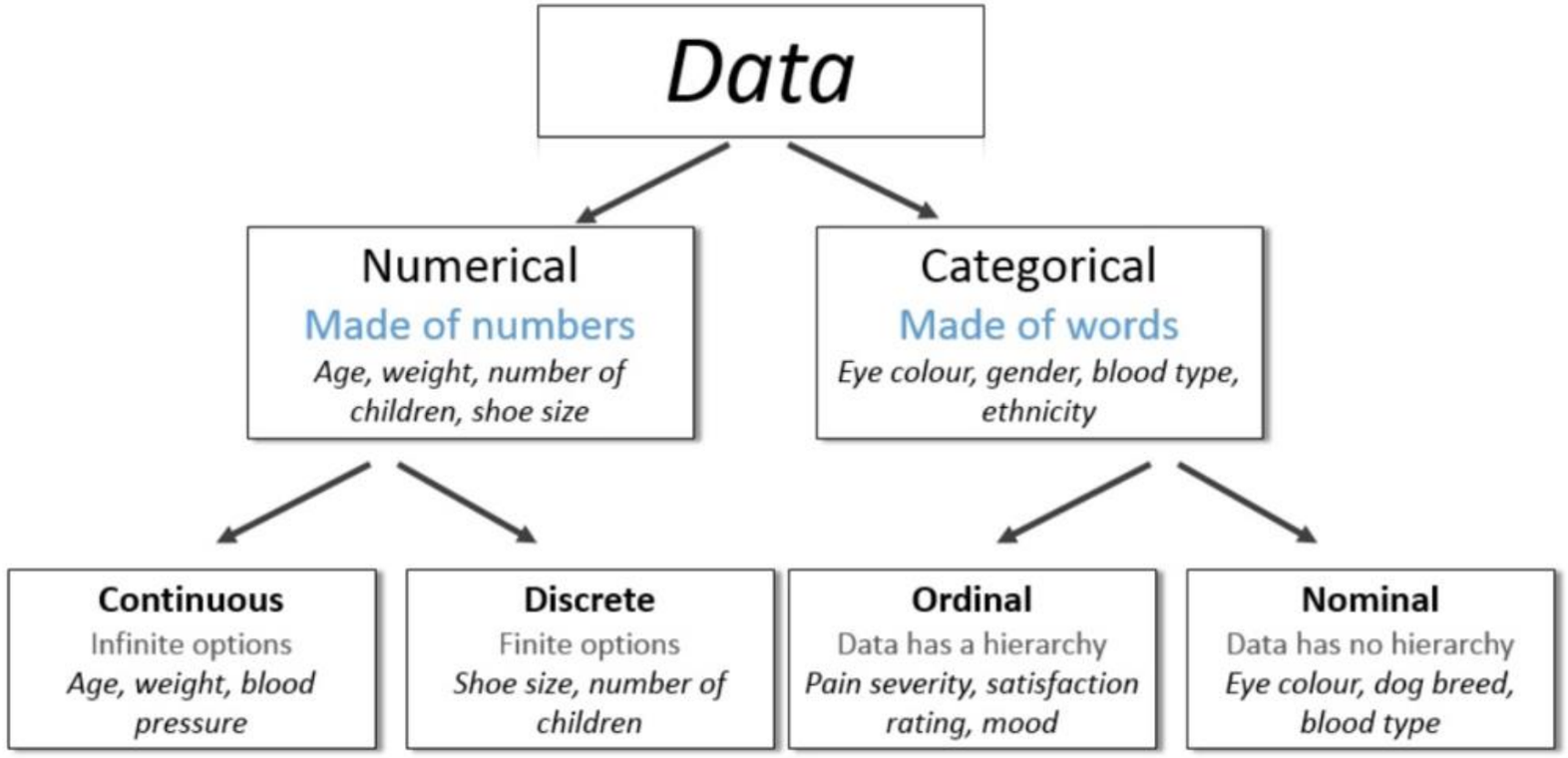
# 資料科學複習教材

授課講師

教材編寫　陳少君



# Data 的種類




Quizlet.com

# CSV/JSON/XML都是資料格式

CSV: Comma Separate Values
XML: eXtended Markup Language
JSON:Javascript Object Notation

## CSV

```
name,age
James Kirk,40
Jean-Luc Picard, 45
WesleyCrusher,27
```

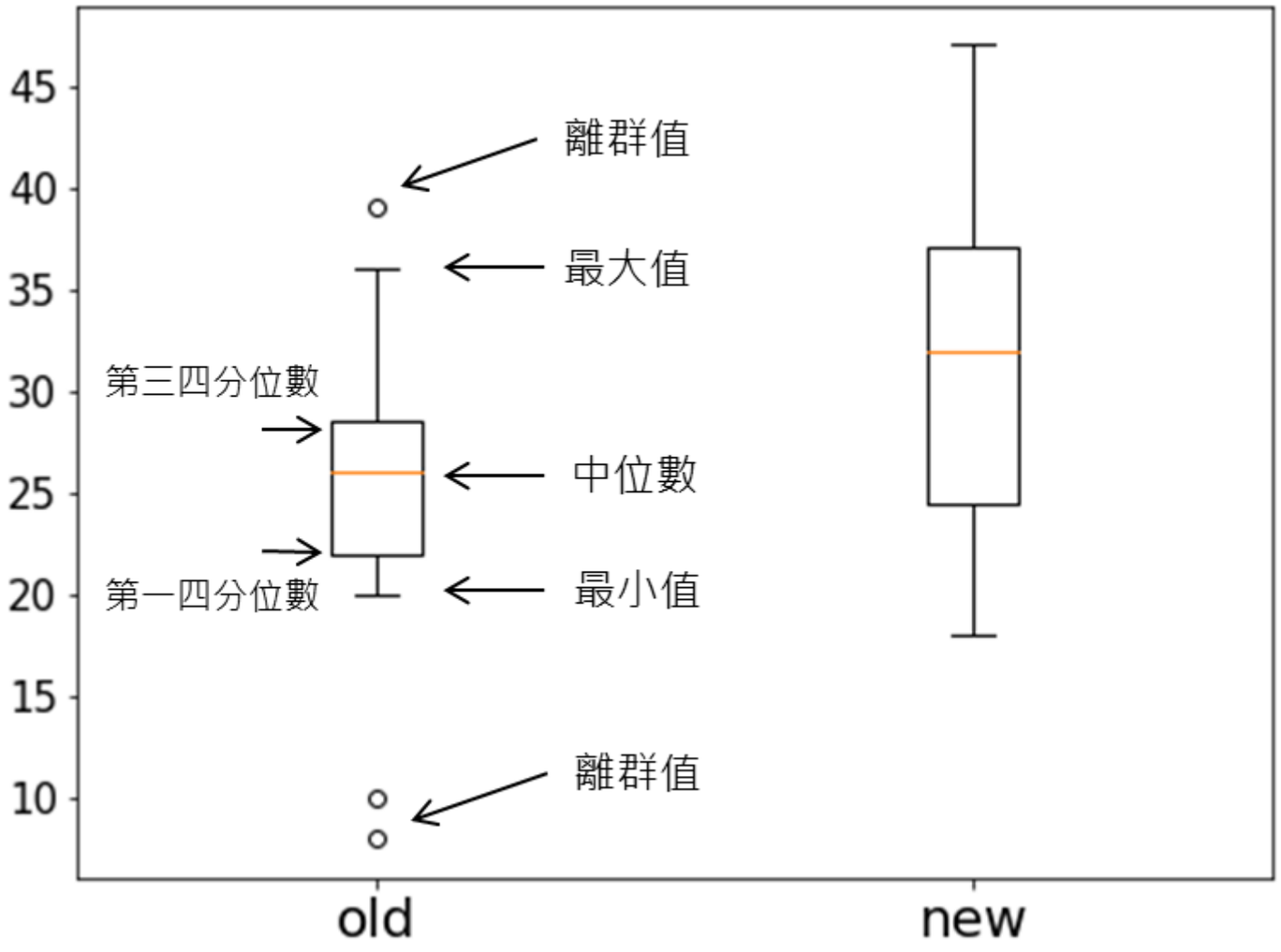
## XML

```
<empinfo>
    <employees>
        <employee>
            <name>James Kirk</name>
            <age>40></age>
        </employee>
        <employee>
            <name>Jean-Luc Picard</name>
            <age>45</age>
        </employee>
        <employee>
            <name>Wesley Crusher</name>
            <age>27</age>
        </employee>
    </employees>
</empinfo>
```

## JSON

```
{   "empinfo" :
        {
            "employees" :   [
            {
                "name" : "James Kirk",
                "age" : 40,
            },
            {
                "name" : "Jean-Luc Picard",
                "age" : 45,
            },
            {
                "name" : "Wesley Crusher",
                "age" : 27,
            }
                        ]
        }
    }
```

# 盒鬚圖(Boxplot)




Ithome.com

# 矩陣乘法

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 10 & 11 \\ 20 & 21 \\ 30 & 31 \end{bmatrix}$$

$$= \begin{bmatrix} 1x10 + 2x20 + 3x30 & 1x11 + 2x21 + 3x31 \\ 4x10 + 5x20 + 6x30 & 4x11 + 5x21 + 6x31 \end{bmatrix}$$

$$= \begin{bmatrix} 10+40+90 & 11+42+93 \\ 40+100+180 & 44+105+186 \end{bmatrix} = \begin{bmatrix} 140 & 146 \\ 320 & 335 \end{bmatrix}$$

https://towarddatascience.com

# 轉置矩陣和反矩陣

用途：求線性迴歸(Linear Regression)係數(截距，斜率)，特徵值降維(PCA)，加速矩陣運算等

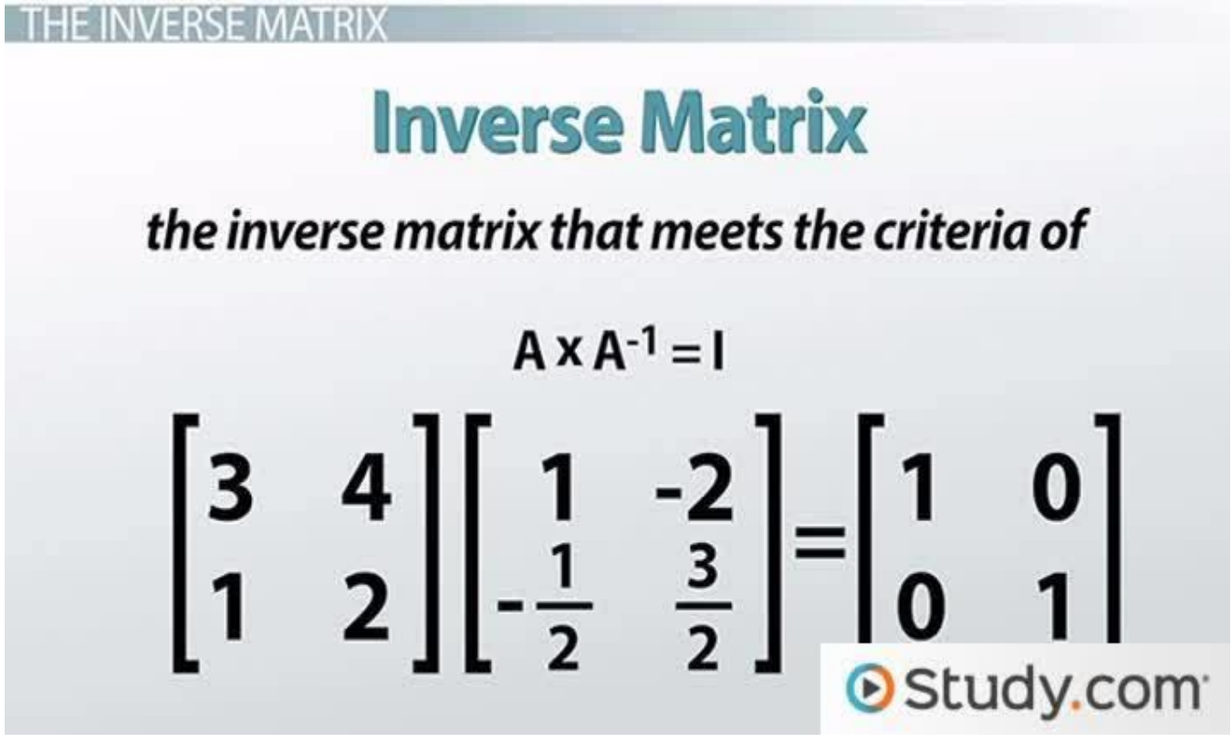
**轉置矩陣**

Transposing a 2x3 matrix to create a 3x2 matrix

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}^T = \begin{bmatrix} 6 & 1 \\ 4 & -9 \\ 24 & 8 \end{bmatrix}$$

Java67.com

**反矩陣**



# 陣列索引

```
import numpy as np
x = np.arange(10)
print(x)
[0,1,2,3,4,5,6,7,8,9]
print (x[2:5])
[2,3,4]
print (x[:-7])
[0,1,2]
print(x[1:7:2])
[1,3,5]

x = np.arange(10,1,-1)
print (x)
[10,9,8,7,6,5,4,3,2]
print(x[np.array([3,3,-3,8])])
[7,7,4,2]
print(x[np.array([[1,1],[2,3]])])
[[9,9],
 [8,7]]

y = np.arange(35).reshape(5,7)
print(y)
[[0,1,2,3,4,5,6],
 [7,8,9,10,11,12,13],
[14,15,16,17,18,19,20],
[21,22,23,24,25,26,27],
[28,29,30,31,32,33,34]]
print(y[1,5,2::3])
[[7,10,13],
 [21,24,27]]
```

# 協方差(Covariance)與相關係數(Correlation Coef.)

## Covariance Formula

**For Population**

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

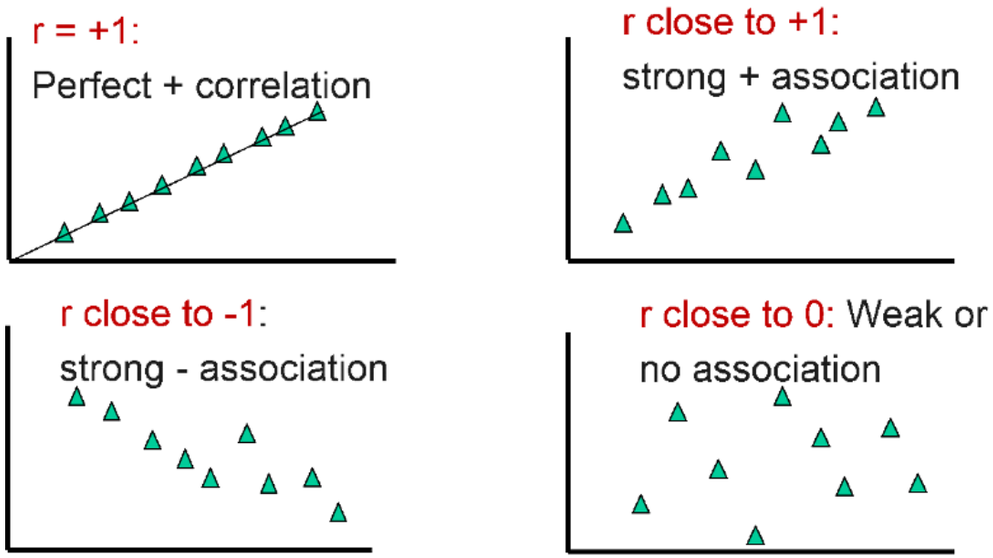
**For Sample**

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

**Correlation Coefficient Formula** =

$$\frac{\Sigma[(X - X_m) * (Y - Y_m)]}{\sqrt{[\Sigma (X - X_m)^2 * \Sigma (Y - Y_m)^2]}}$$

Educba.com




Sphweb.bumc.bu.edu

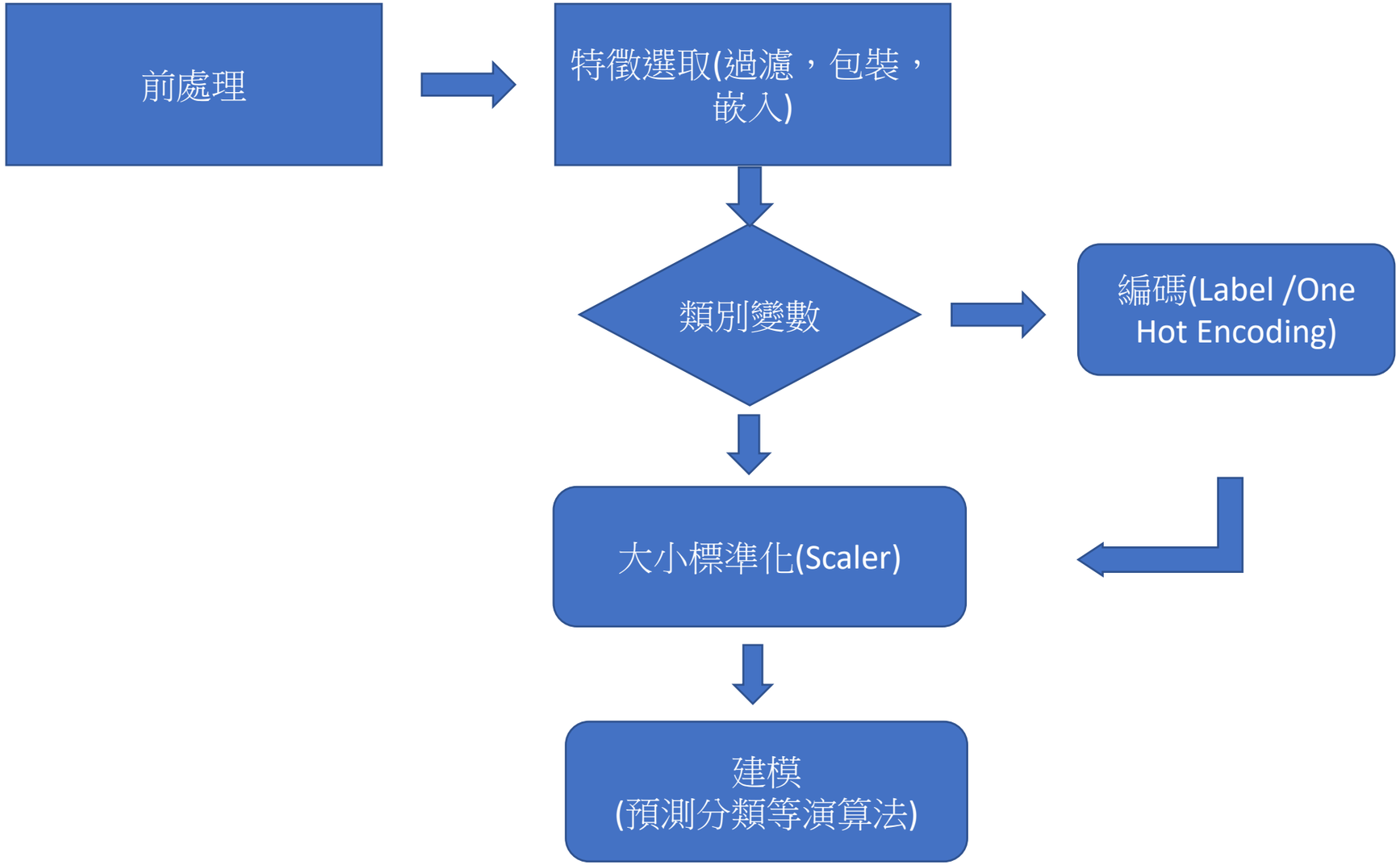
# EDA 特徵選擇 (features selection)

**Filter(過濾法)**：按照發散性或**相關性**對各個特徵進行評分，設定閾值或者待選擇特徵的個數進行篩選 (SelectKBest)

**Wrapper(包裝法)**：根據目標函數（往往是預測效果評分），每次選擇若干特徵，或者排除若干特徵 (RFE)

**Embedded(嵌入法)**：先使用某些機器學習的模型進行訓練，得到各個特徵的權值係數，根據係數從大到小選擇特徵（類似於Filter，只不過係數是通過訓練得來的）(SelectFromModel)

# 從特徵到建模



# EDA:房地產價格分析

Train.csv

detailed-exploratory-data-analysis-with-python.ipynb
Data: train.csv

去除30%以上為NaN的欄位 → 選取數值型欄位 df_num → 數值型欄位直方圖 → df_num與SalesPrice的相關係數，找出top 10 → Pairplot 劃出和SalesPrice的關係 → 拿掉0值再做一次Correlation → Heatmap 找出特殊相關 → Quantitative to Quantitative Regplot for regression model → Categorical to quantitative 使用 Boxplot → CtoQ Count Plot

# 與Apriori 同dataset，FP-growth 演算法

## FP-growth 的算法，假設 min. support threshold = 3

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

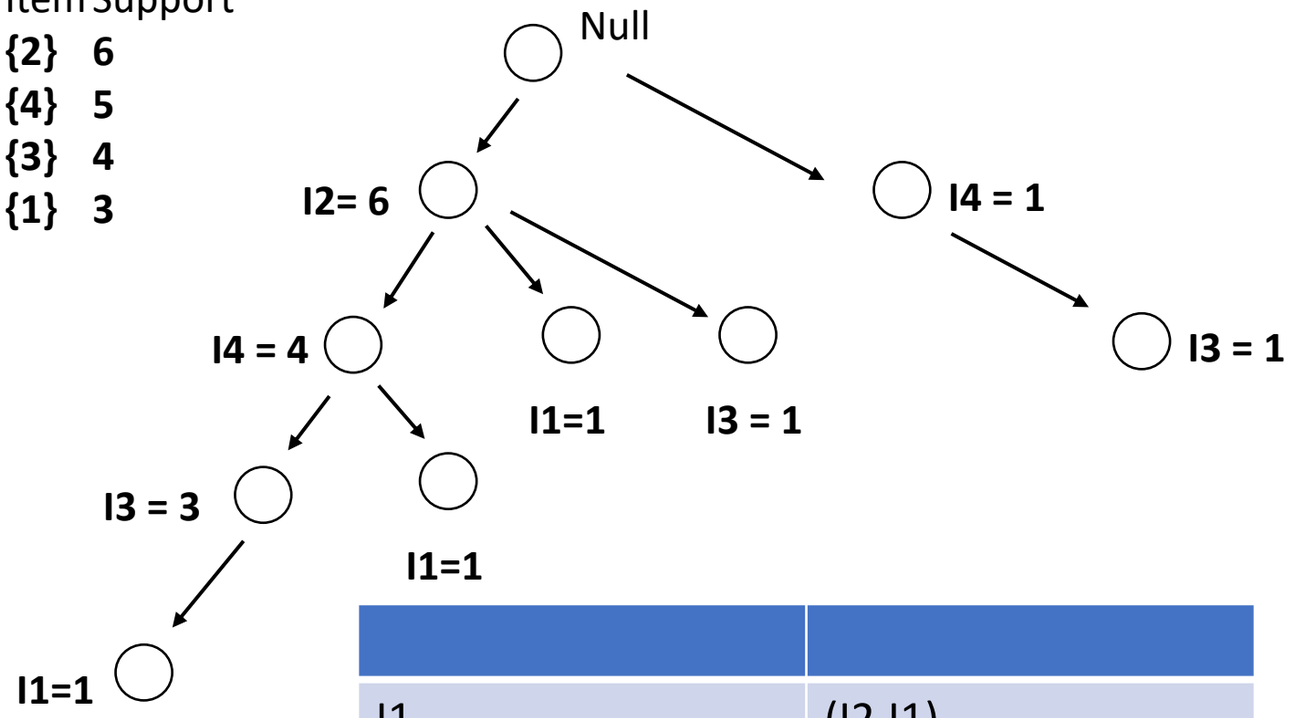
Item Support
**{1} 3**
**{2} 6**
**{3} 4**
**{4} 5**

Sorted
Item Support
**{2} 6**
**{4} 5**
**{3} 4**
**{1} 3**

Item Support
**{1,2}3**
{1,3}1
{1,4}2
**{2,3}3**
**{2,4}4**
**{3,4}3**

Item Support
{2,3,4} 2

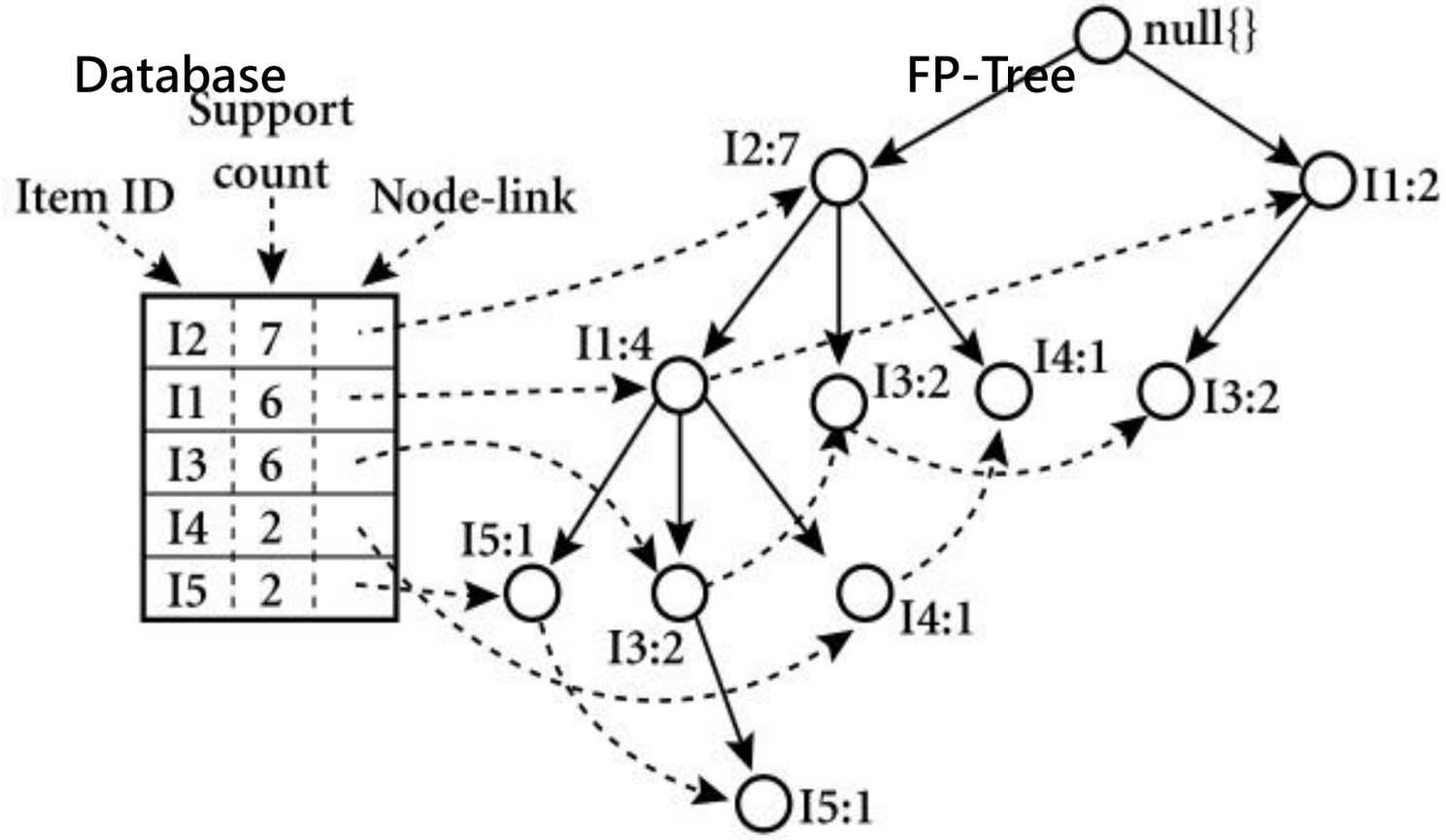
(Not Qualified)




| | |
|---|---|
| I1 | (I2.I1) |
| I3 | (I2,I3)(I4,I3) |
| I4 | (I2,I4) |

# 投影片FP-growth 演算法

## FP-growth 演算法資料結構解答

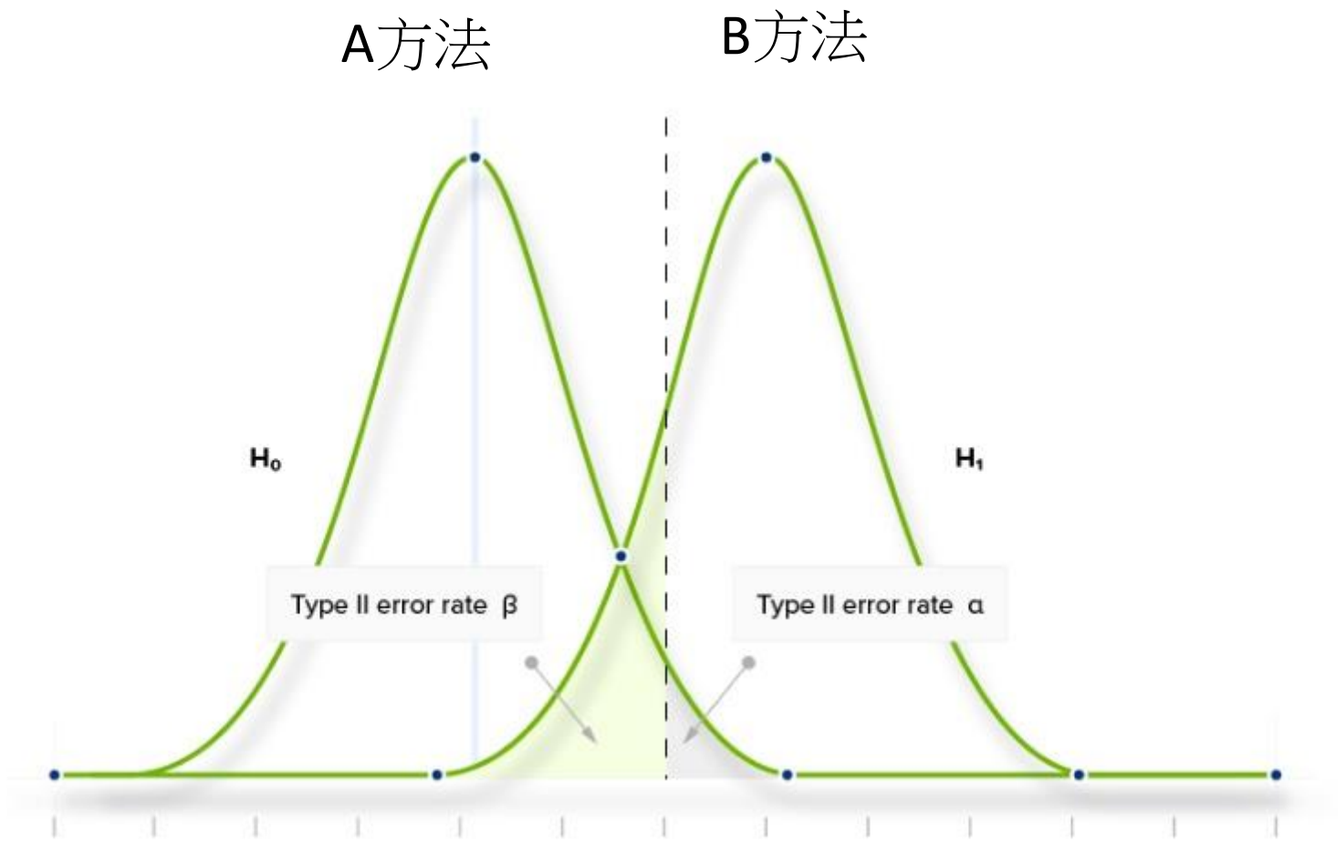



解答：本例FP-growth 演算法一步步詳解-

# A/B - Test

**情境**：希望能透過網頁設計的改變，提高conversion rate，進而增加收入提升利潤。

右圖是conversion rate的樣本統計分佈。雖是二項式分佈，但因中央極限定理(CLT)，大樣本可用常態分佈計算。




Control Group
對照組

Treatment Group
實驗組

convertsize.com

# A/B-test 的步驟

**目的：**以統計分析方式檢定改變(新網頁設計，新藥，新教學方法)是否有效

**步驟：**

1. 分為實驗組(treatment group)與對照組(control group)
2. 決定樣本數大小，一般1:1
3. 隨機取樣(AB樣本需互斥)
4. 定位為**雙樣本比例之平均值是否相等**之假設檢定。虛無假設：相等。
5. Alpha = 0.05 (type 1 error = 0.05)，Power = 0.8 (type 2 error = 1 – 0.8 = 0.2)
6. 求取z-score(z-分數)與p-value (如顯著需要 |z-score| > 1.96，$p < 0.05$)
7. 求取95%信賴區間 (confidence interval)
8. 判斷改變是否顯著

# A/B-Test 計算樣本數

```
import statsmodels.stats.api as sms
effect_size = sms.proportion_effectsize(0.13, 0.
15)
required_n = sms.NormalIndPower().solve_power(
    effect_size,
    power=0.8,
    alpha=0.05,
    ratio=1
    )                                    # Calcula
ting sample size needed
required_n = ceil(required_n)
```

# A/B Test 計算統計值

```
z_stat, pval = proportions_ztest(successes, nobs=no
bs)
(lower_con, lower_treat), (upper_con, upper_treat)
= proportion_confint(successes, nobs=nobs, alpha=0.
05)
```

```
z statistic: -0.34
p-value: 0.732 #不顯著
ci 95% for control group: [0.114, 0.133] (0.13在裏頭，
但達不到0.15)
ci 95% for treatment group: [0.116, 0.135]
```