# 資料預處理（Data Preprocessing）

> Reference：
>
> 1. Data Preprocessing by Dr. Tun-Wen Pai

## 資料預處理

對於一個資料集來說，若數值偏大或偏小的情況下，我們很難去做資料視覺化，甚至資料過大時會讓自然指數過大導致 overflow。

故我們希望可以把資料集的值域縮小，有以下幾種方法：

1. Z-score normalization
2. MinMaxScaler
3. MaxAbsScaler
4. Robust Scaling
5. QuantileTrasformation

### *Z-score normalization*

對於每個資料，我們利用常態分佈將他標準化（Standardization），定義為

$$x' = \frac{x - \overline{x}}{\sigma}$$

用這樣的方式就能夠將其值域限縮至 $[-1, 1]$ 之間。

### *MinMaxScaler*

對於每個資料，我們使用最大與最小值的區間來限縮值域，定義如下：

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

用這樣的方式就能夠將其值域限縮至 $[0, 1]$ 之間。

### *MaxAbsScaler*

對於每個資料，我們單純使用最大與最小值的區間來限縮值域，定義如下：

$$x' = \frac{x}{\max\{|x|\}}$$

用這樣的方式就能夠將其值域限縮至 $[-1, 1]$ 之間。

# *Robust Scaling*

Robust Scaling 是一種非線性的限縮方式，使用第三分位與第一分位來進行限縮。

前面的方法，以最大值來說，若最大值過大則限縮資料會變小。

Robust Scaling 優化了這個部分，他沒有特定的值域限縮範圍，但可以對極值有良好的抗噪性，定義如下：

$$x' = \frac{x - Q_2}{(Q_3 - Q_1)}$$