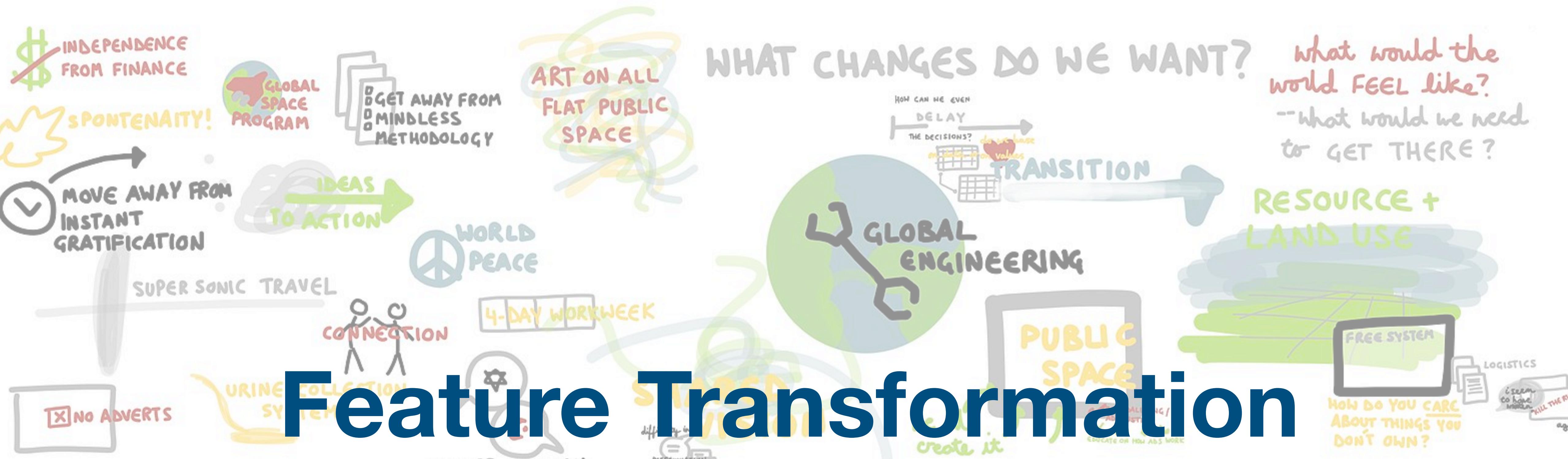


Feature Transformation

Feature Engineering Part 1

Miles

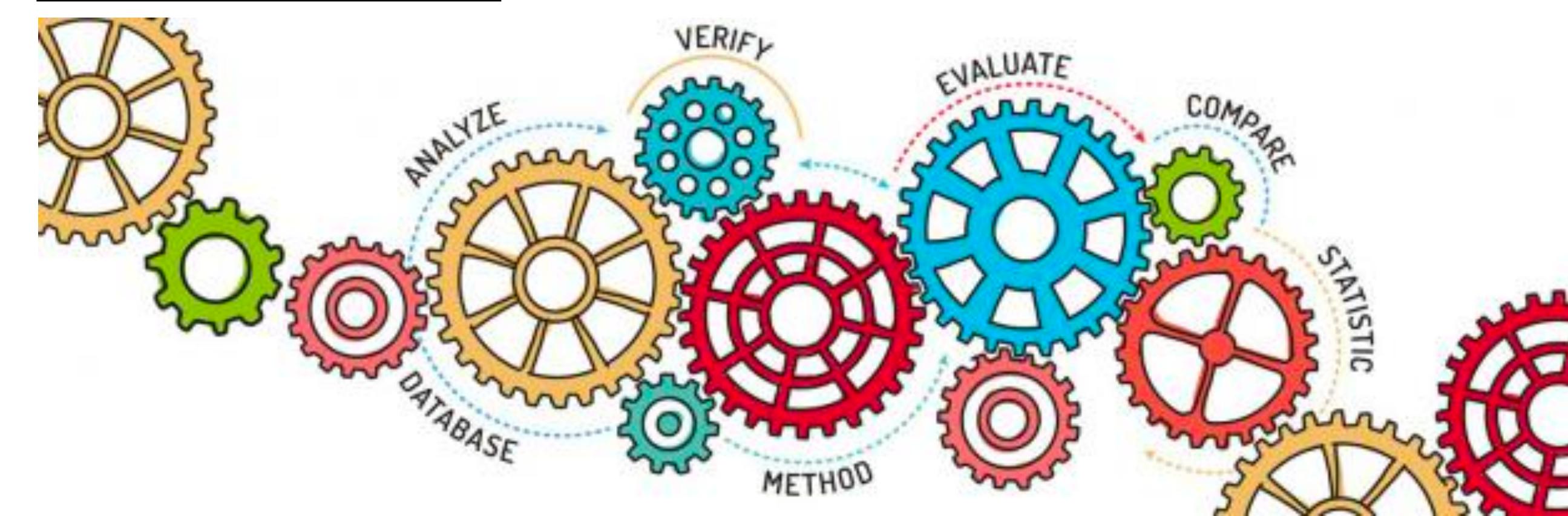


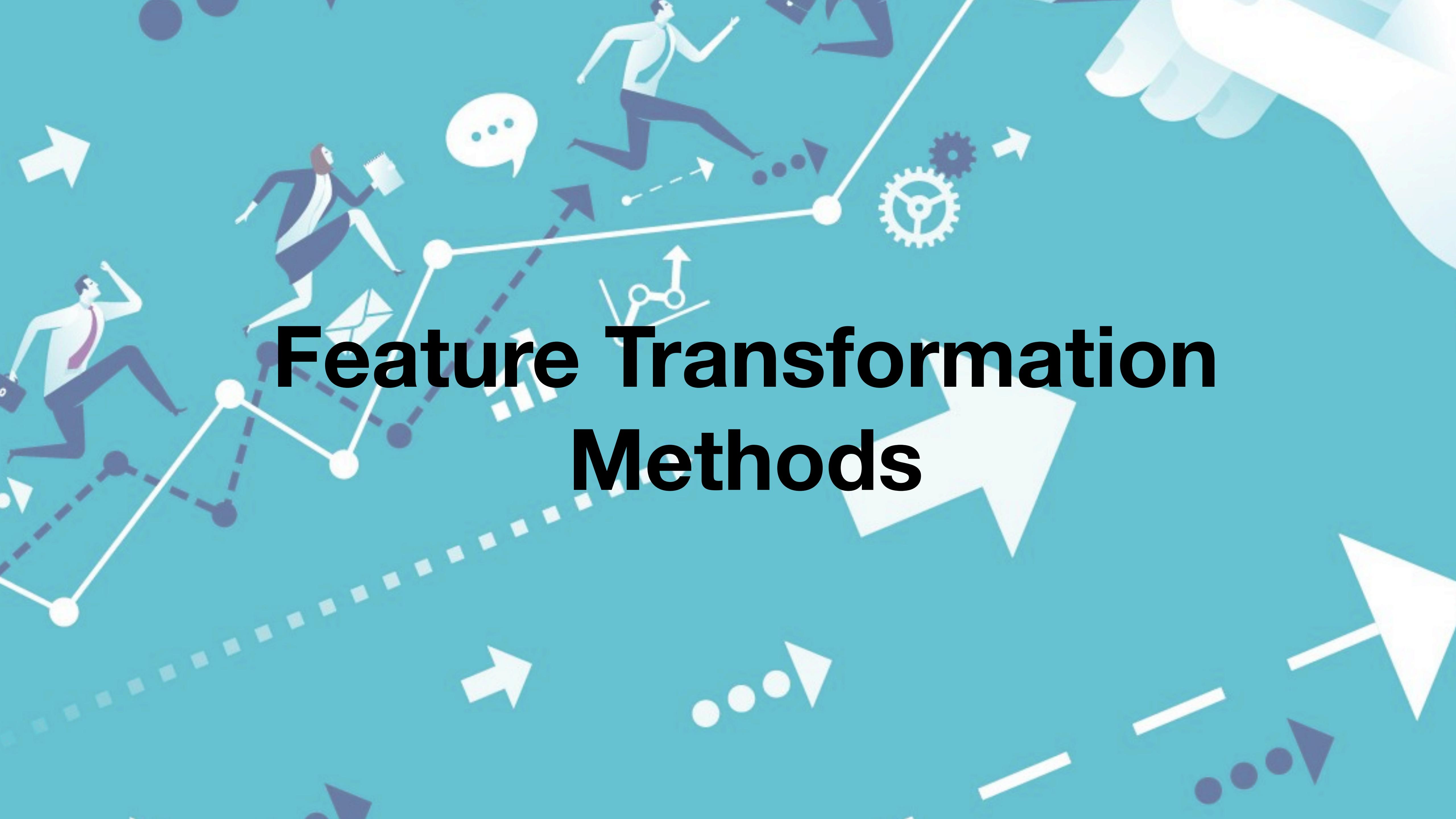
Why Do We Need Feature Transformation?



Feature Transformation

- 更方便觀察數據間的關係
- Feature數值分佈太廣，減少Outlier對模型的影響
- Reducing skewness
- Feature Scale差距過大
- 使用特定統計模型時，假設資料必須符合某種分配





Feature Transformation Methods

Binarization

將特徵映射成0或1兩個數值，適用於數量無法表達層級的情況

- 是否成年
- 是否為銀行VIP
- 是否持有信用卡

Discretization

將特徵進行離散化處理，從某些層面來說比連續更容易理解不同階層的特性

- 中小學生 (5-18)
- 大學生 (19-23)
- 工作新鮮人 (24-29)
- 成家立業 (30-40)
- 中年人 (40-60)

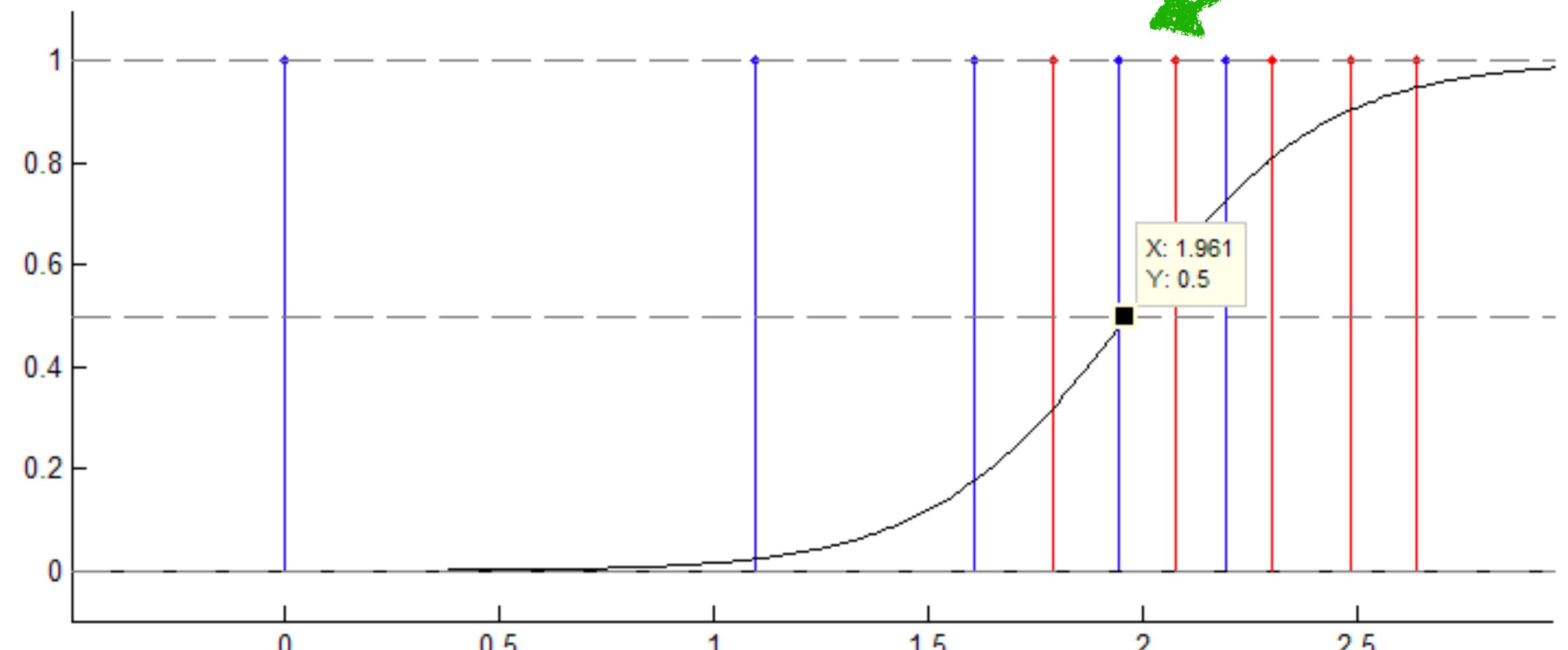
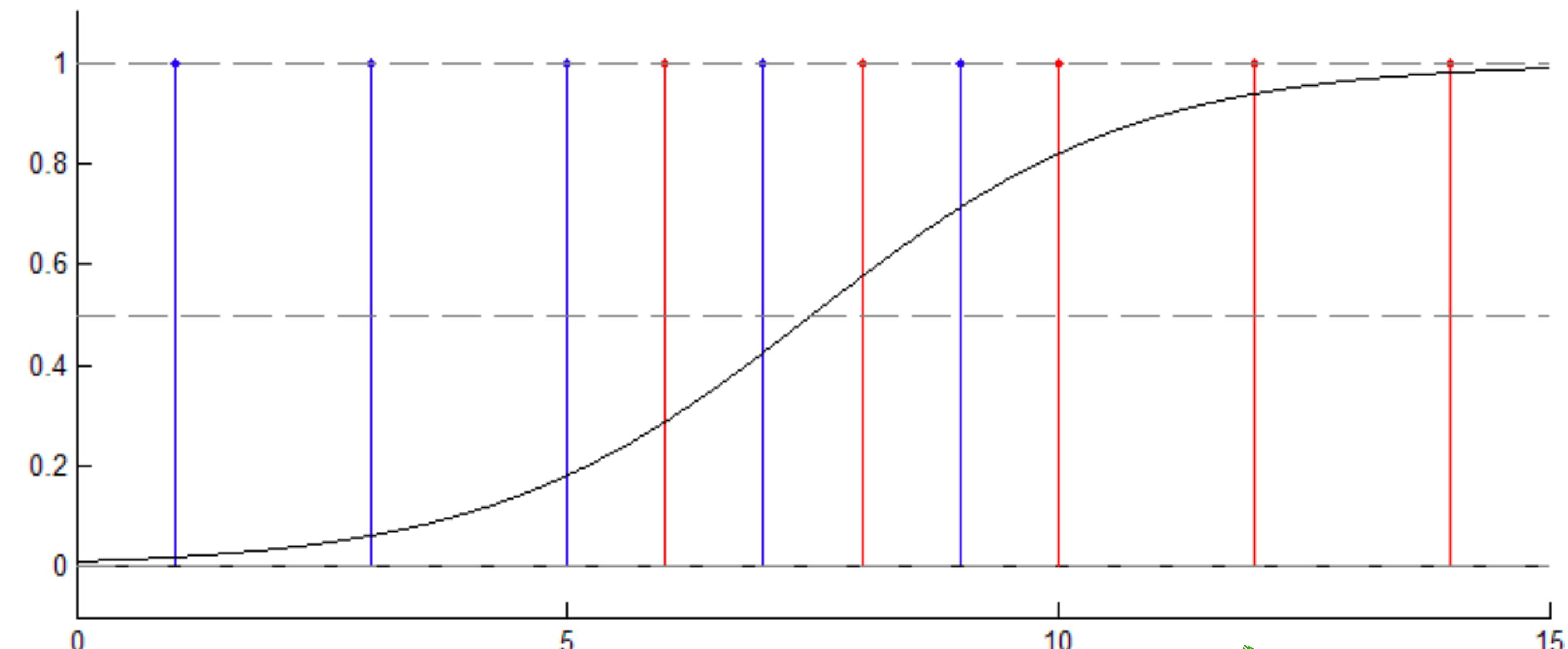
典型的離散化步驟：對特徵做排序 -> 選擇合適的分割點 -> 作出對整體的分割
-> 查看分割是否合理

Log Transformation

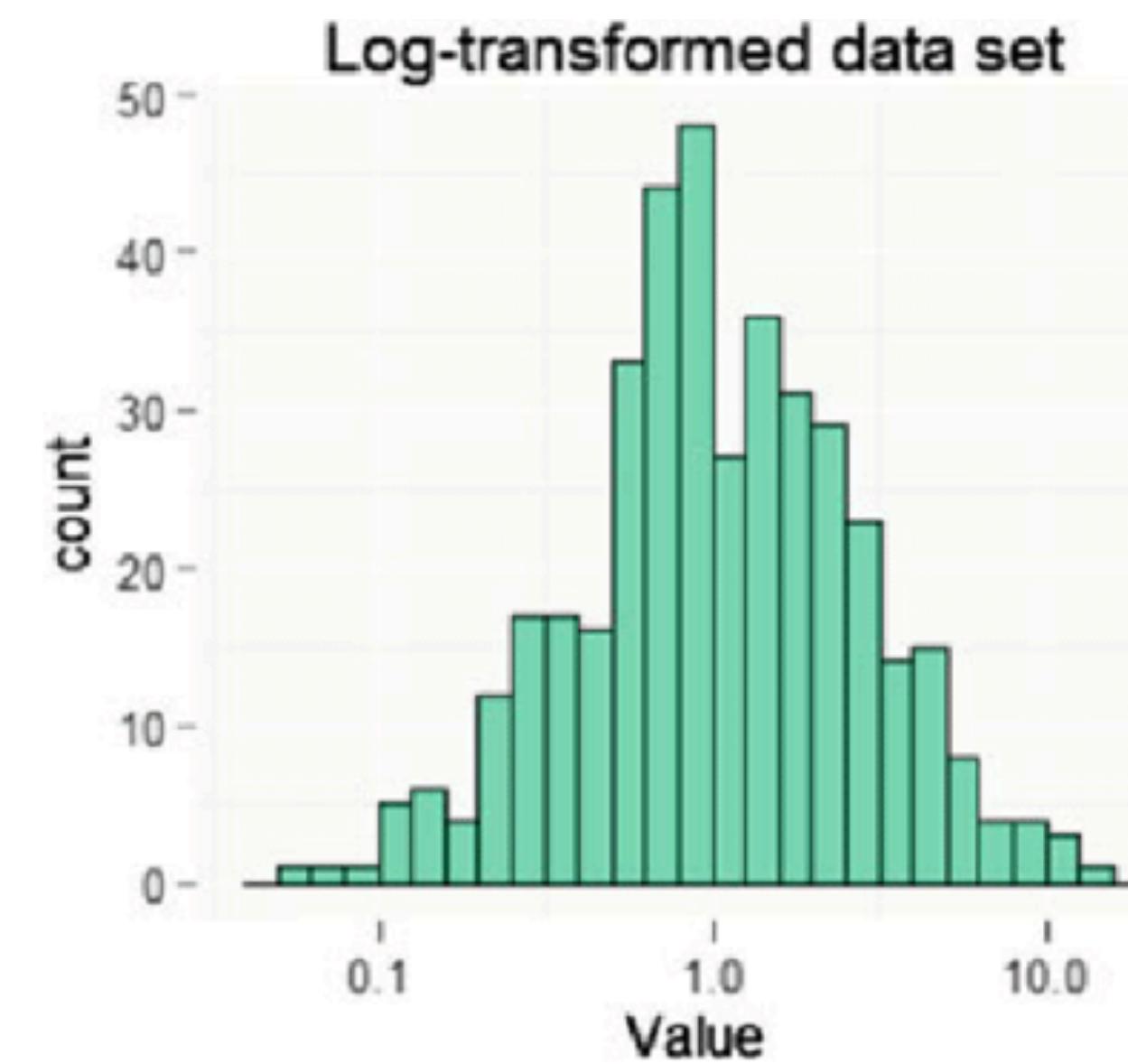
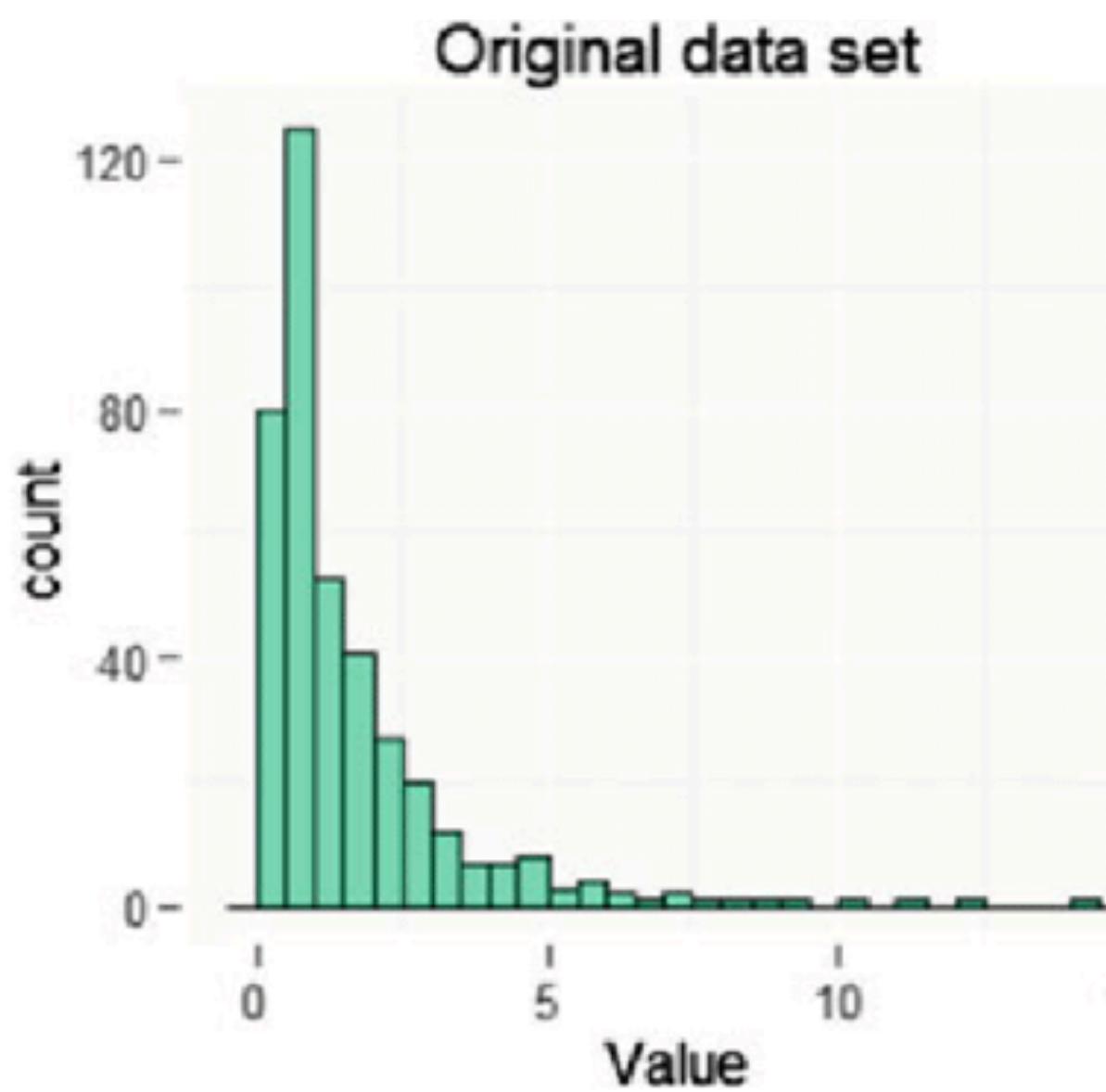
對特徵取 Log，
將會對座標軸進行不均勻的伸縮

值小的特徵分布更分散

值大的特徵分布更集中

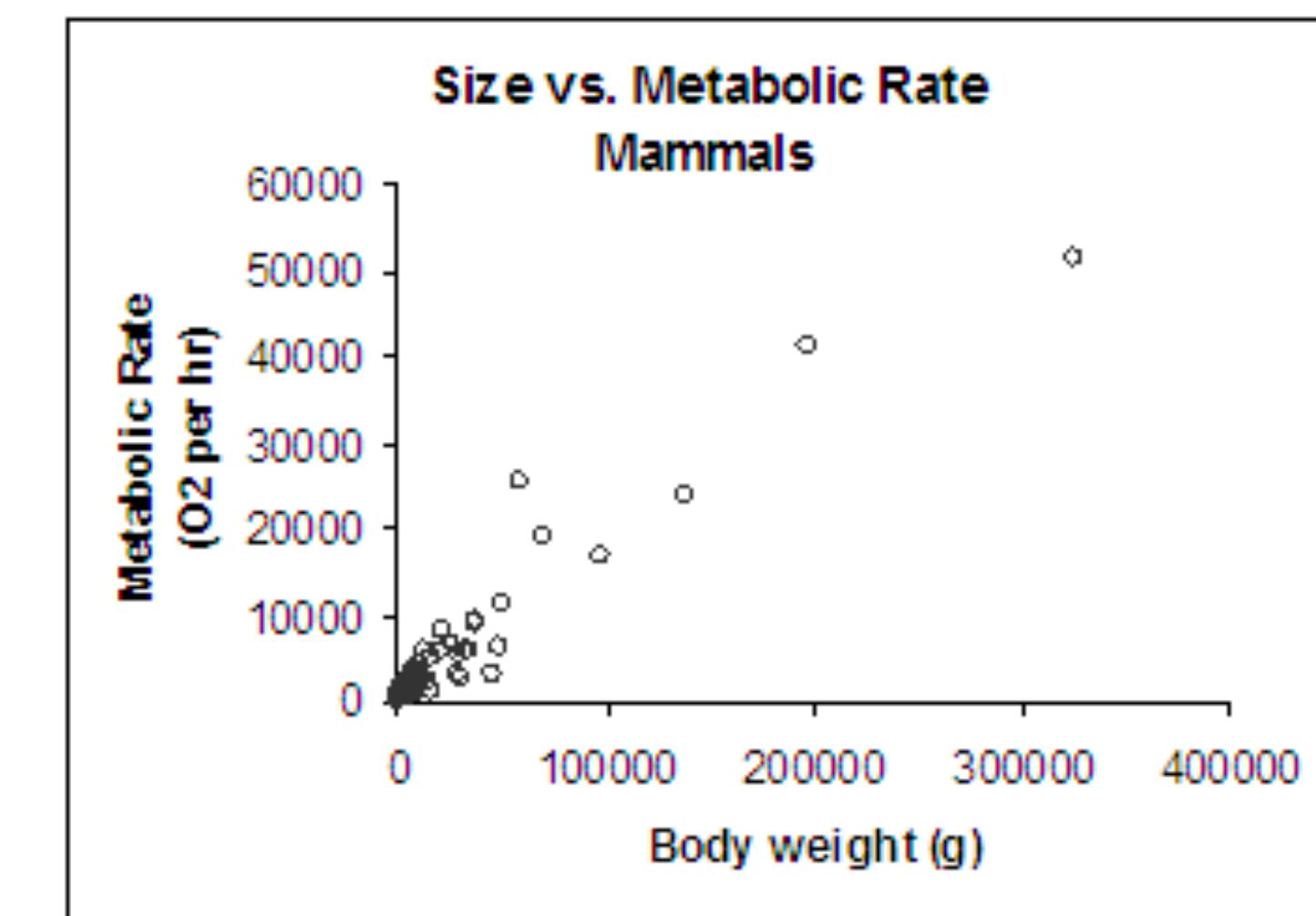


Log Transformation

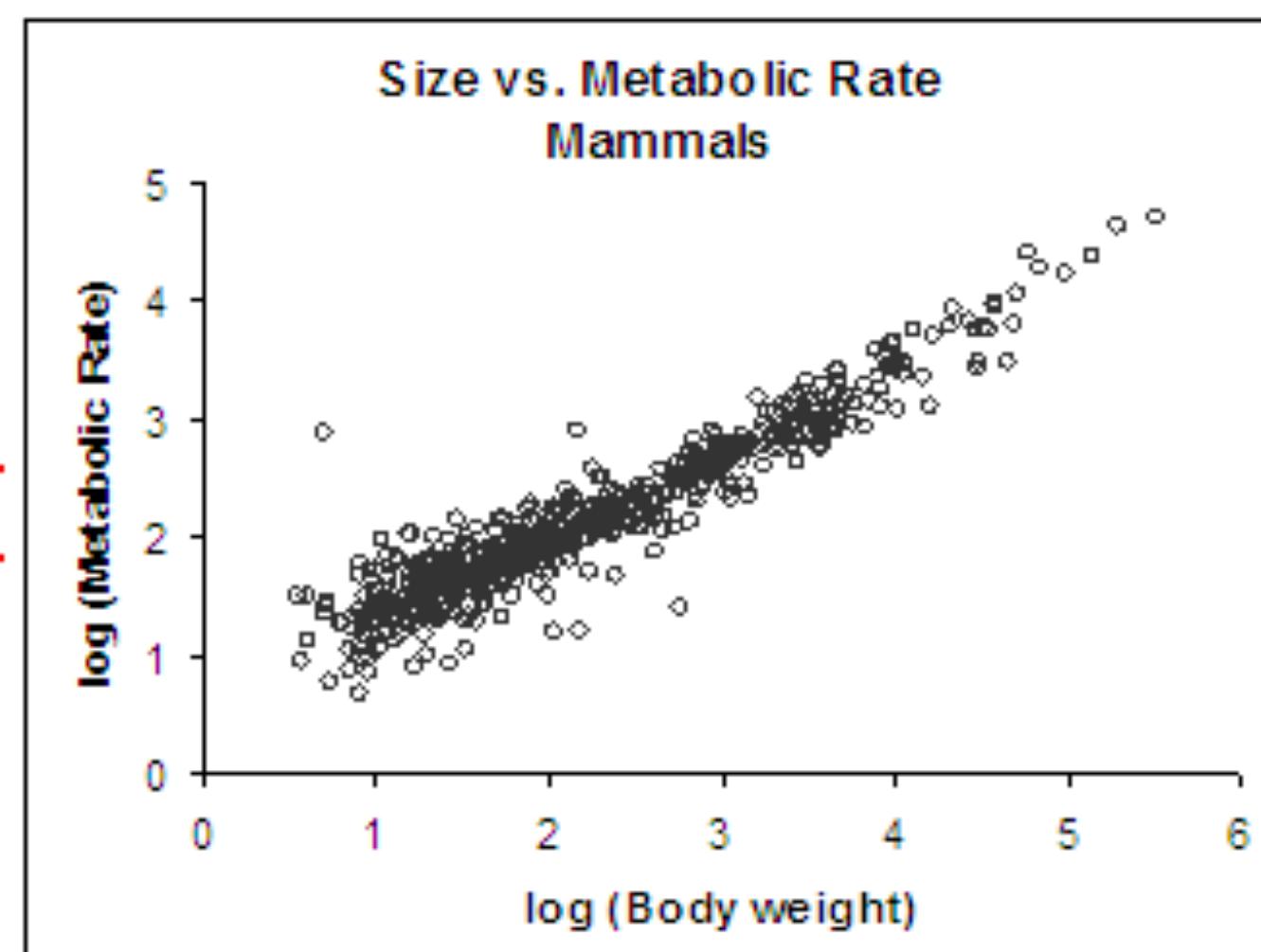


右偏的資料取Log後，會更趨近常態分佈

轉換後資料明顯貼近線性關係
更加配適線性迴歸

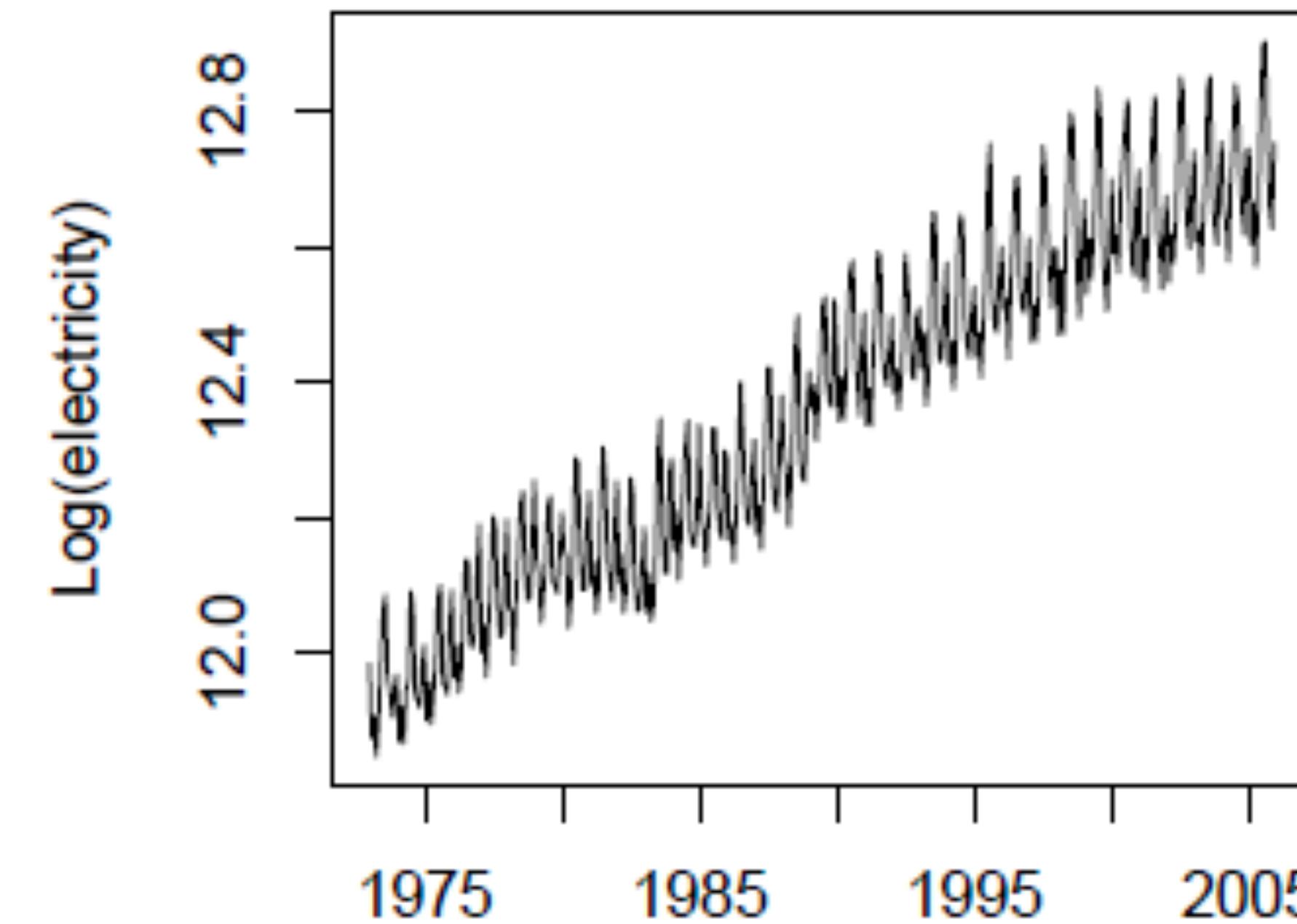
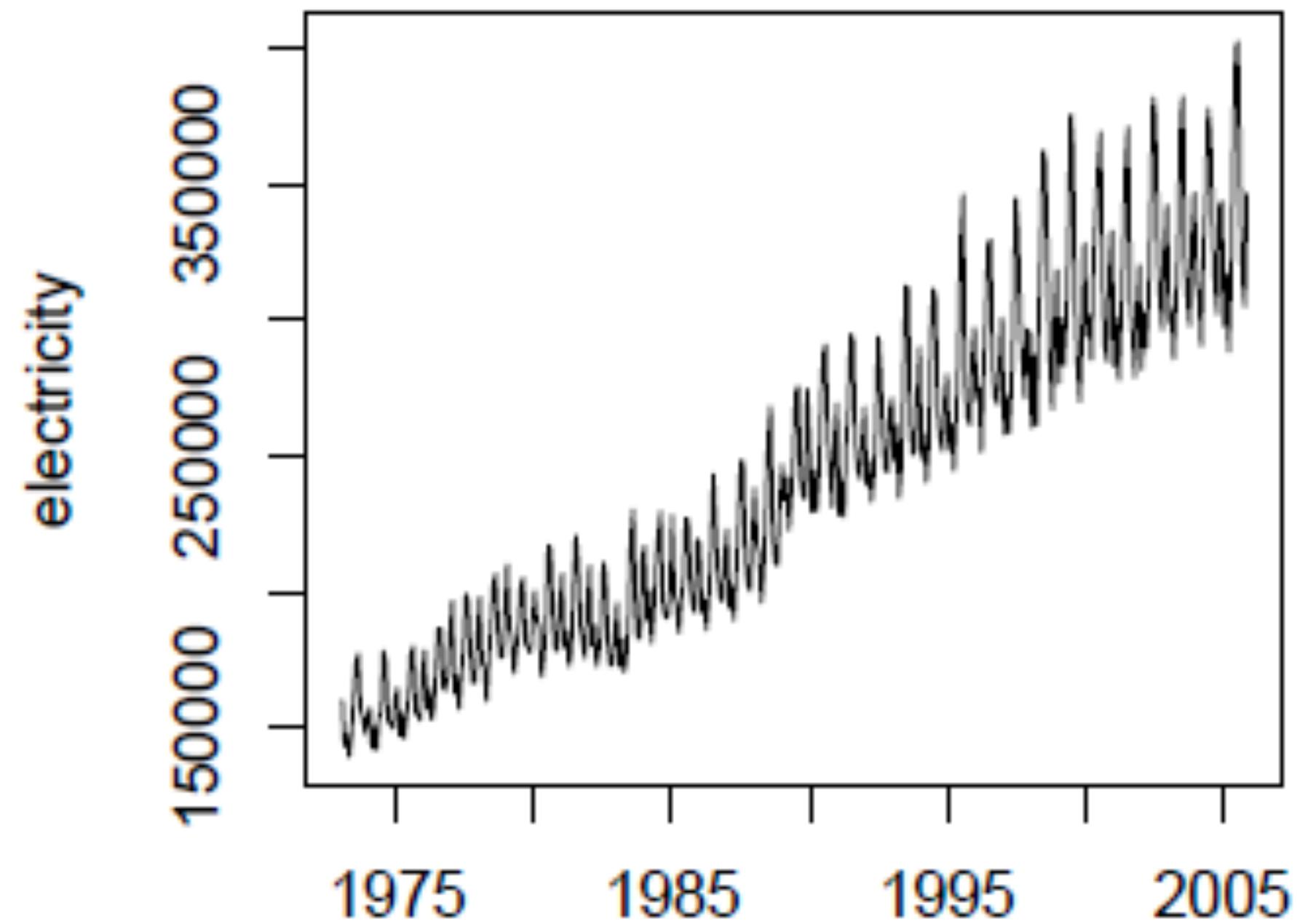


Log
Trans-
form



Log Transformation

以Time Series來說，序列需要符合Stationarity，數據分佈有時會因數值大小而變動過大，但其實趨勢是相同的，取Log後反而可以看出數據的趨勢。



需注意Log Transformation須確保資料大於0，若否須位移後再轉換。

Box-Cox Transformation

使用時機？

許多連續型統計方法的前提假設資料必須要服從常態分佈，故在進行統計前，應先檢測是否符合此前提假設。如果不是，則不可以使用這個統計方法

特色？

此轉換方式可讓資料近似於常態分佈

限制？

所有輸入值必須為正且不等於0。若不滿足要求，可將數據位移致讓所有數值為正

Box-Cox Transformation

Transform Function

$$X = x_0, x_1, x_2, \dots, x_{N-1}$$

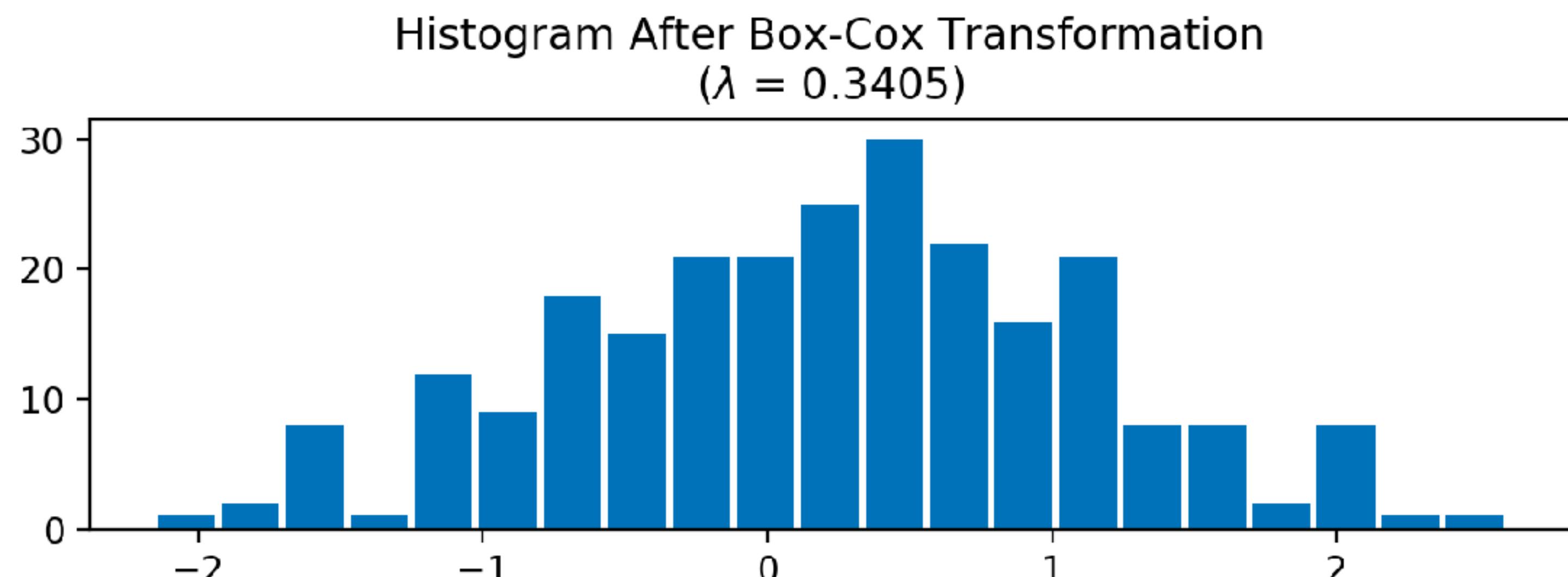
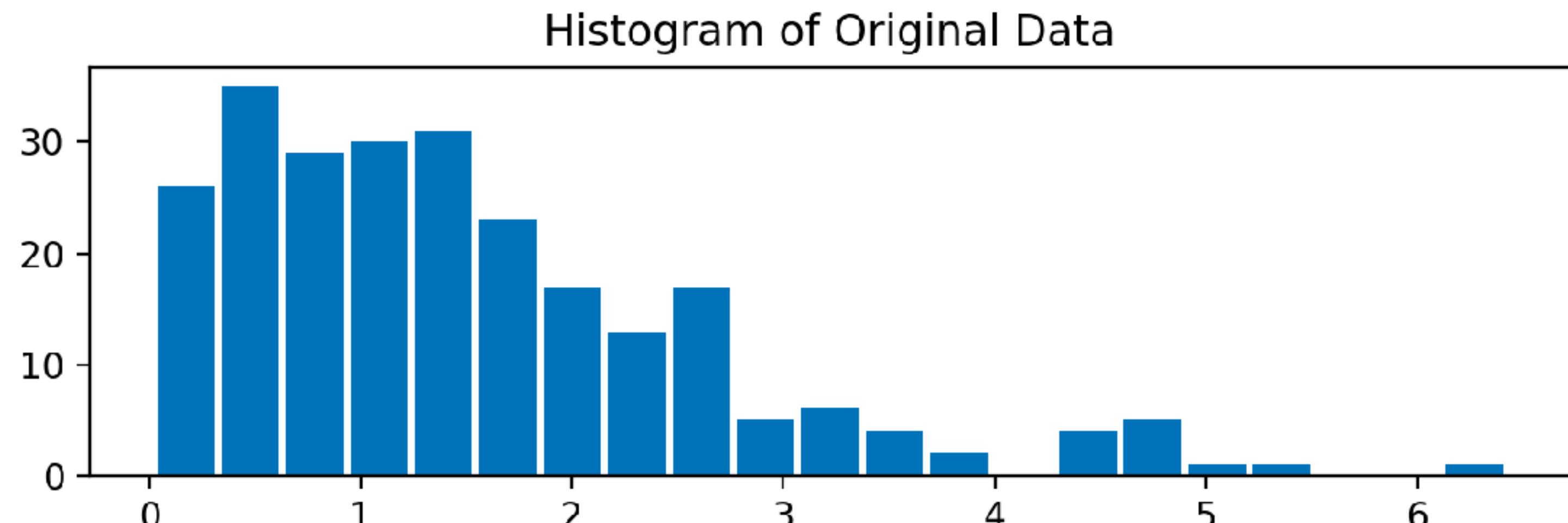
$$x_i(\lambda) = \begin{cases} \frac{x_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x_i), & \lambda = 0 \end{cases} \quad i = 0, 1, 2, \dots, N-1$$

Maximum Likelihood Estimate

$$f(x, \lambda) = -\frac{N}{2} \ln \left[\sum_{i=0}^{N-1} \frac{(x_i(\lambda) - \bar{x}(\lambda))^2}{N} \right] + (\lambda - 1) \sum_{i=0}^{N-1} \ln(x_i)$$

Box-Cox Transformation

用最大概似估計得出
Lambda值後，代入Box-Cox Transform Function
轉換後近似於常態分佈



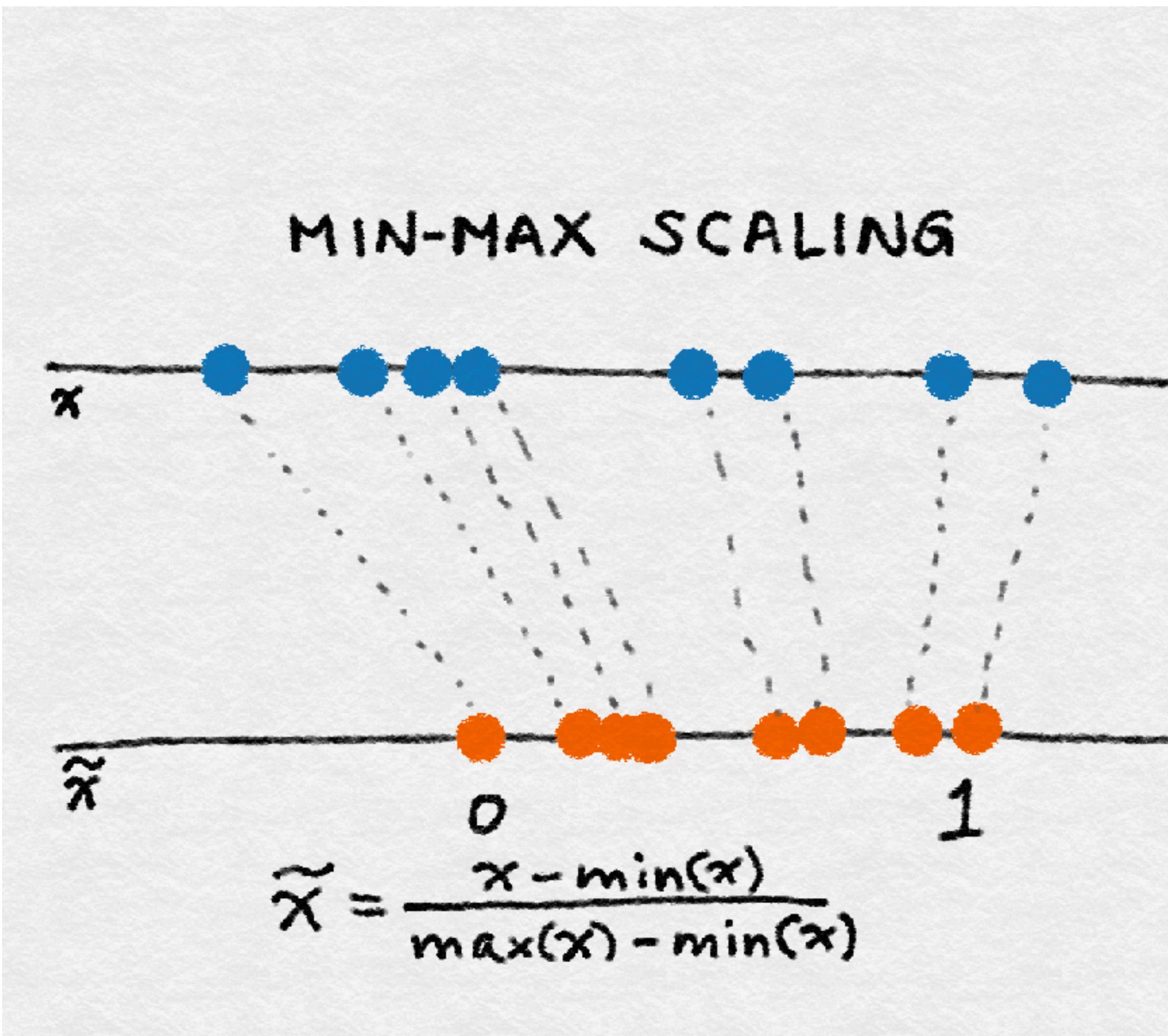


機器學習模型和統計模型

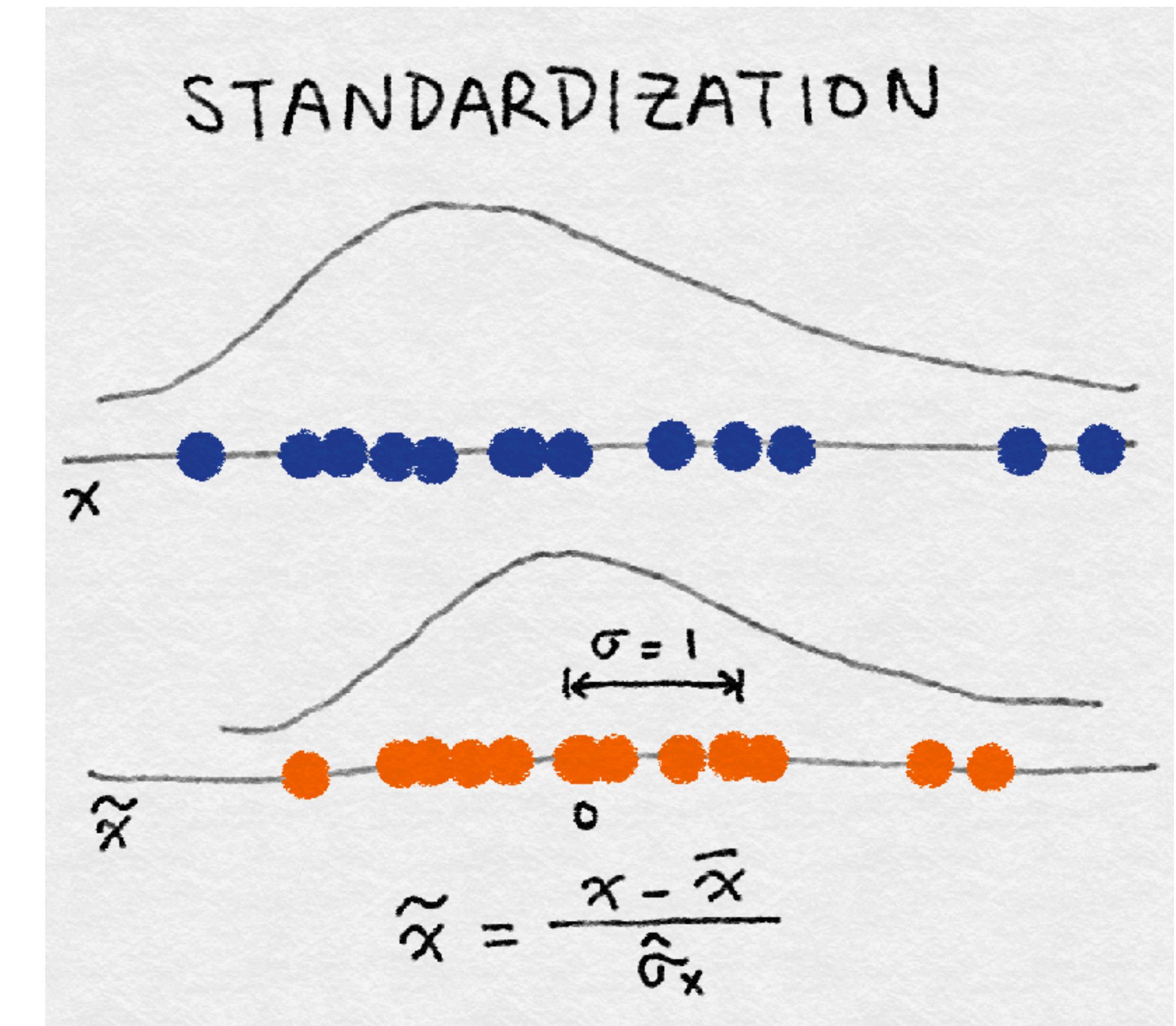
差別在哪？

Feature Scaling

Min-Max Scaling



Standardization



Feature Scaling

使用時機？

許多機器學習模型在不同範圍特徵的數據中呈現不同的學習效果。例如，SVM在沒有標準化調整過的數據中表現很差，所以在丟入模型前將**Feature Scale一致**

特色？

把單位的概念去除，且轉換方式會將數據限制在某個區間中

限制？

離散型資料較不適合使用

Feature Transformation

是必須的嗎？

Feature Transformation前先問問自己：

要使用的是哪種模型？

模型有沒有基本假設？

Feature要用哪種方法轉換？

Feature Transformation前先問問自己：

要使用的是哪種模型？ 樹型的模型對於Scale較不敏感，可不用轉換，
但可能要注意outlier問題。

模型有沒有基本假設？ 統計模型通常有基本假設，
須確認Feature有沒有符合假設。

Feature要用哪種方法轉換？ 機器學習模型通常要Scaling，
統計模型需確認假設後再轉換

Let's Try Some Transformation

