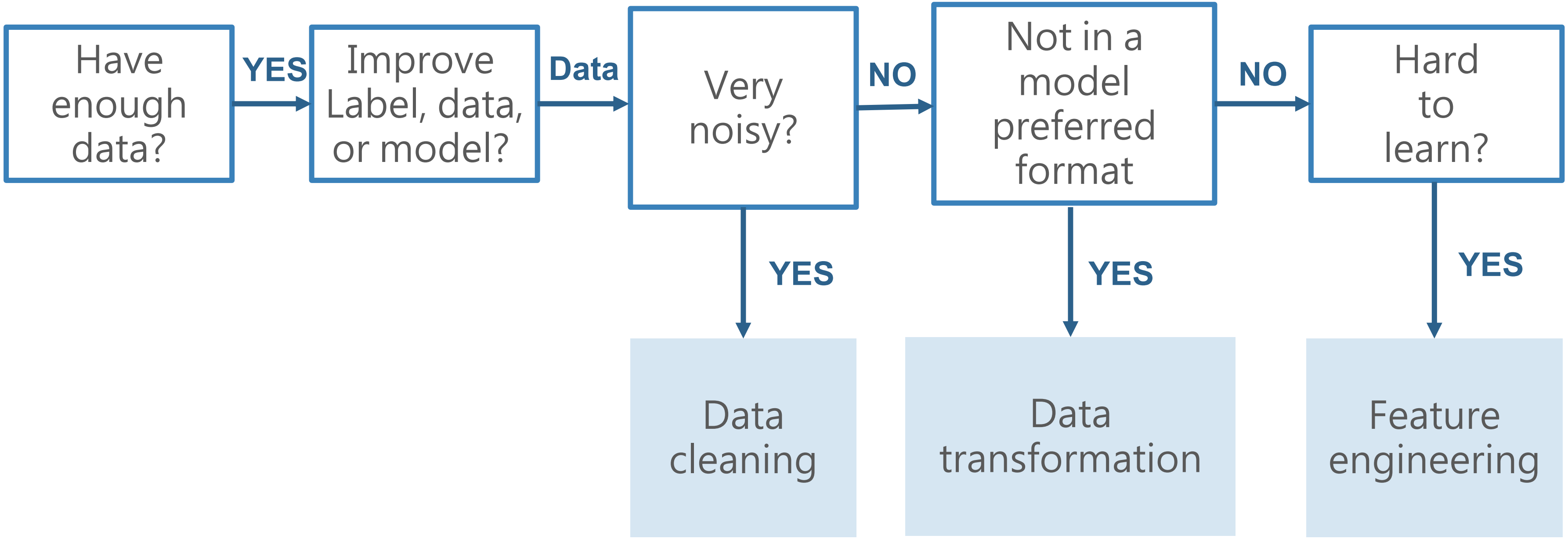
# 0.2 Data Preprocessing

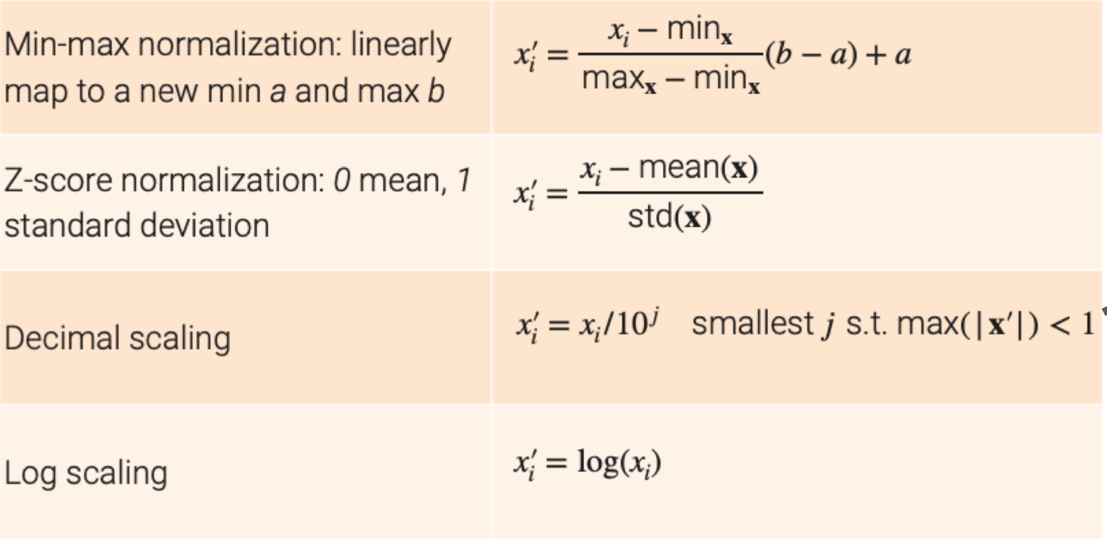
# Data preprocessing flow chart



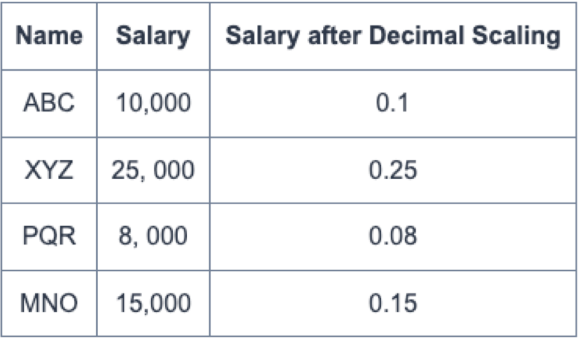
# Data Transformation

# Normalization for Real Value Columns

Min-max：將資料等比例縮放到 [0, 1] 區間中，不能處理outlier，適用在數值比較集中的情況。
z-score：適用於分佈大致對稱的資料，平均數為 0 且標準差為 1 的常態分佈，可處理outlier。



| | |
|---|---|
| Min-max normalization: linearly map to a new min *a* and max *b* | $x_i' = \frac{x_i - \min_{\mathbf{x}}}{\max_{\mathbf{x}} - \min_{\mathbf{x}}}(b-a) + a$ |
| Z-score normalization: *0* mean, *1* standard deviation | $x_i' = \frac{x_i - \text{mean}(\mathbf{x})}{\text{std}(\mathbf{x})}$ |
| Decimal scaling | $x_i' = x_i/10^j$ smallest $j$ s.t. $\max(\lvert\mathbf{x}'\rvert) < 1$ |
| Log scaling | $x_i' = \log(x_i)$ |



| Name | Salary | Salary after Decimal Scaling |
|---|---|---|
| ABC | 10,000 | 0.1 |
| XYZ | 25, 000 | 0.25 |
| PQR | 8, 000 | 0.08 |
| MNO | 15,000 | 0.15 |

Decimal scaling can tone down big numbers into easy to understand smaller decimal values.

Stanford CS 329P (2021 Fall) - https://c.d2l.ai/stanford-cs32

# Feature Engineering

- Feature Engineering -> Machine Learning

- Feature Learning -> Deep Learning (images/videos/audio/text)
  Train deep neural networks to extract features.

# 表格資料

**Tabular data are in the form of a table,
feature columns of numeric / categorical / string type**

1. Int / float: directly use or bin to unique int values

2. Categorical data: one-hot encoding
   - Map rare categories into “Unknown”

$$\begin{array}{ccccc} \text{fish} & \text{cat} & \text{mouse} & \text{dog} & \text{others} \\ [\ 0, & 1, & 0, & 0, & 0\ ] \\ [\ 0, & 0, & 0, & 1, & 0\ ] \end{array}$$

3. Date-time: a feature list such as
   - [year, month, day, day_of_year, week_of_year, day_of_week]

4. Feature combination: Cartesian product of two feature groups
   - [cat, dog] x [male, female] -> [(cat, male), (cat, female), (dog, male), (dog, female)]

# THANKS