Definition

Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.

- `2D` & `3D` → Scatter Plot and other related plots
- `4D`, `5D`, & `6D` → Pair Plots
- `nD` → Dimensionality Reduction
  - PCA (Principal Component Analysis)
  - t-SNE (t-distributed Stochastic Neighborhood Embedding)

In [ ]:

Row and Column Vector

- A row vector is a row of entires. It has `1` row and `n` columns.

$$a = [a_1, a_2, a_3, \ldots, a_n]$$

- A column vector is a column of entries. It has `1` column and `n` rows.

```
x = [[x1],
     [x2],
     [x3],
     ...,
     [xn]]
```

Note

- By default, when someone says a vector, it means that it is a column vector.
- The transpose of column vector is called a row vector.
- Please refer to wiki article.

In [ ]:

Dataset representation

A dataset is represented as $D = \{x_i, y_i\}$ where $X = x_i$ (independant variables or features) and $Y = y_i$ (target variable or dependant variable).

```
For example
-----------
```

Iris data $\implies$

```
[[PL],
 [PW],
 [SL],
 [SW]]
```

which are features and

```
[species]
```

represents target variable.

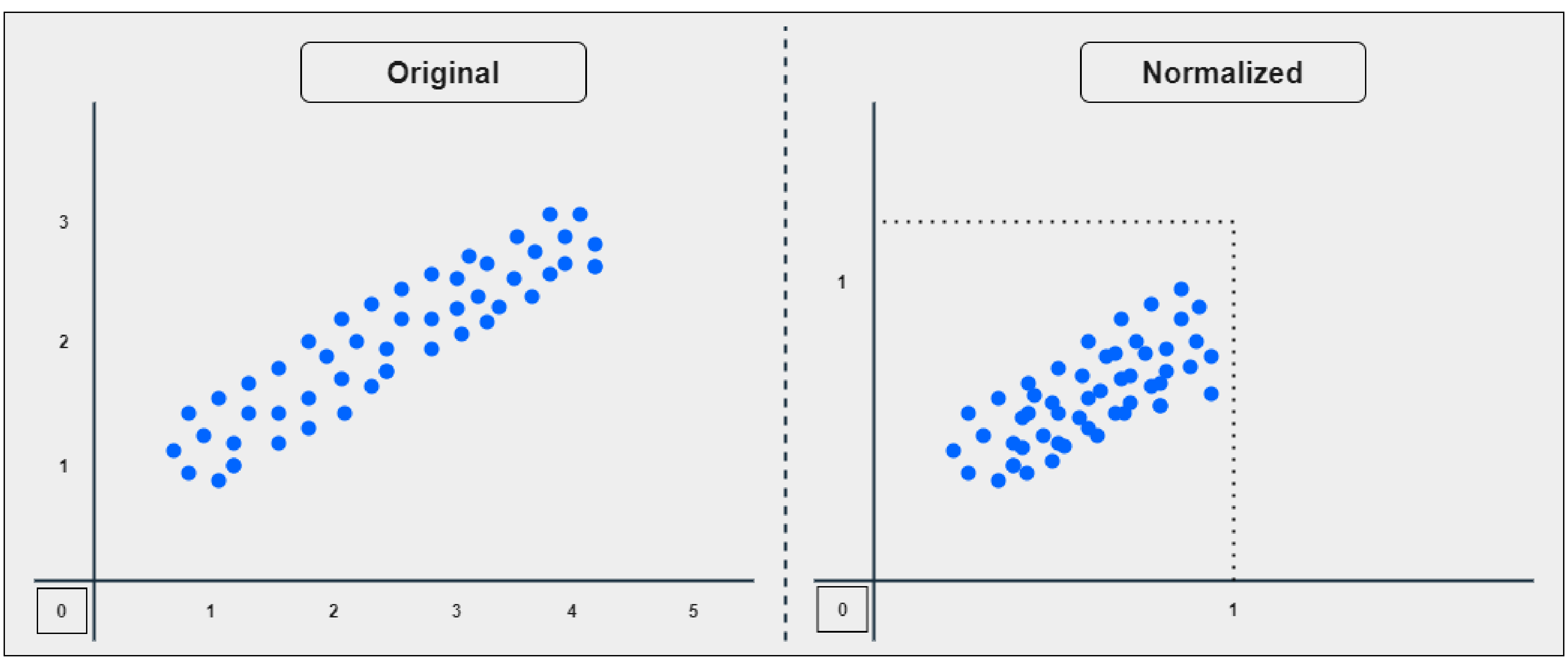
In [ ]:

Data Preprocessing

- **Column Normalization**
  - Consider each col from the dataset and find out col_{min} and col_{max}.
  - Compute

$$col_i^1 = \frac{(col_i - col_{min})}{(col_{max} - col_{min})}$$

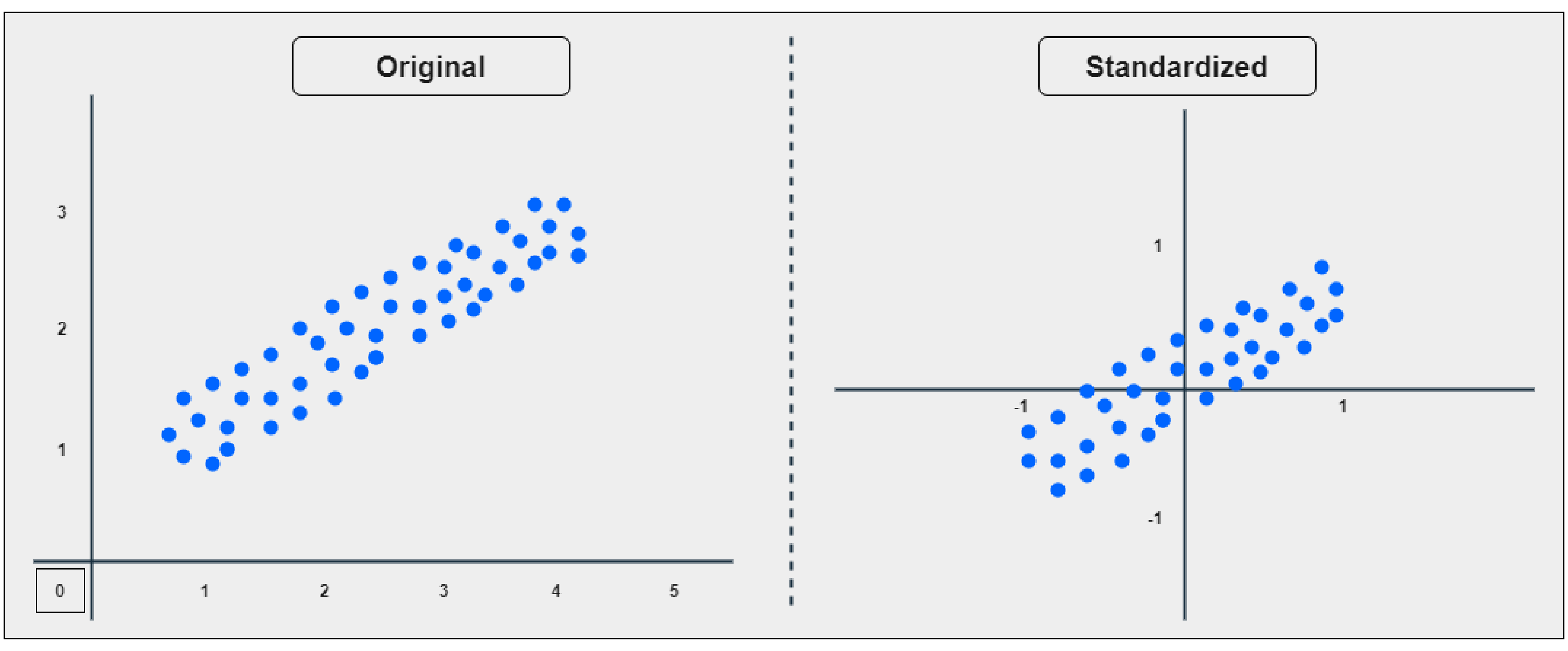
  - All the value of col will be in the range of 0 and 1, $col_i^1 \in [0, 1]$
  - This helps to get rid off scale measurement.
  - Squishes the data into one unit measurement.



- **Column Standardization**
  - Consider each col from the dataset and find out μ_{col} and σ_{col}.
  - Compute standard normal variate for the col such as

$$col_z = \frac{(col_i - \mu_{col})}{\sigma_{col}}$$

  - The mean of col_z is 0 and standard deviation is 1.
  - It is also called as mean centering, i.e., mean is at origin and scaling is done by standard deviation (1).



In [ ]:

Co-Variance Data Matrix (Symmetric Matrix)

Let A be a matrix where $A_{ij} = A_{ji}$ then A is known as symmetric matrix.

```
A = [[2, 1, 2],
     [1, 1, 5],
     [2, 5, 3]]
```

Co-Variance Data Matrix is a symmetric matrix.

- $\mathrm{Cov(X, Y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$
- $\mathrm{Cov(X, X)} = Var(X)$
- $\mathrm{Cov(X, Y)} = \mathrm{Cov(Y, X)}$

Let f_1 and f_2 are two features which are `column standardized`. The $\mathrm{Cov}(f_1, f_2)$ is written as -

$$\implies \frac{1}{(n-1)}\sum_{i=1}^{n} f_1 f_2$$

$$\implies \frac{f_1^T f_2}{(n-1)}$$

Note - We consider $(n-1)$ so as to make sure we get an unbiased estimator.

If X is a dataset irrespective of target variable. Assuming X is `column standardized`, we get covariance matrix as -

$$S_{ij} = \frac{X^T X}{(n-1)} = \frac{f_i^T f_j}{(n-1)}$$

where

- f_i and f_j are features.

In [ ]:

Processing math: 100%