Principal Component Analysis

Dimensionality Reduction

Dimensionality Reduction

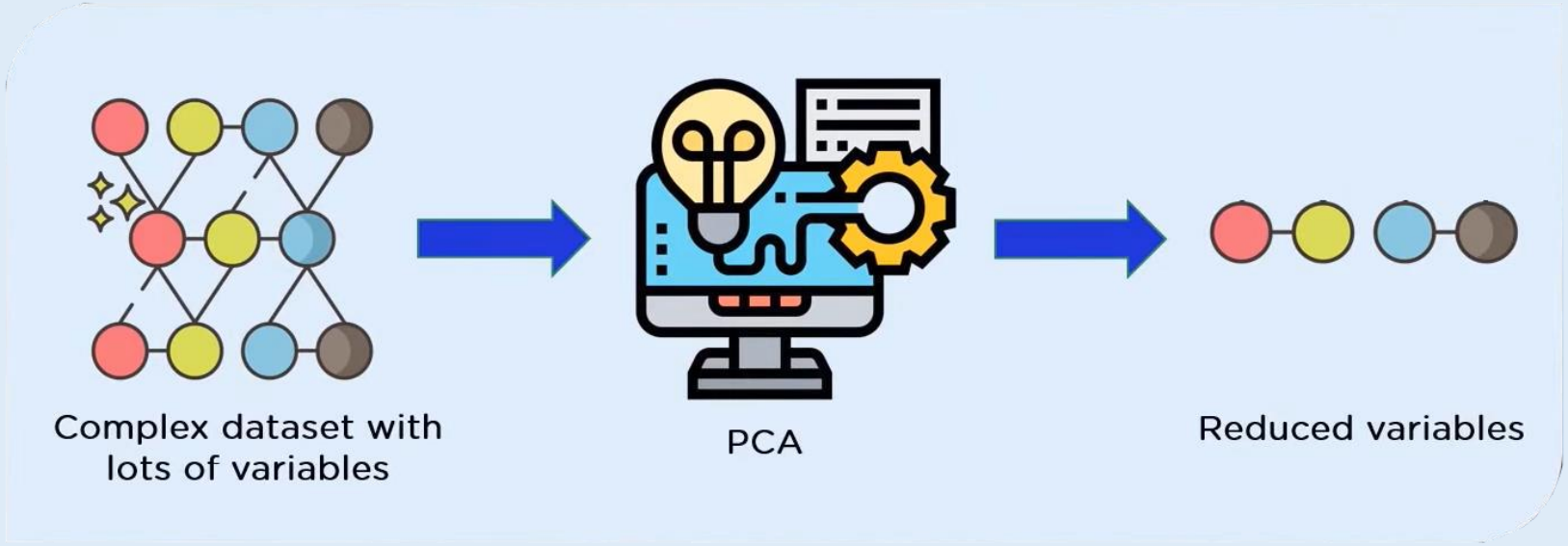
Dimensionality reduction refers to the technique that reduce the number of input variables in a dataset

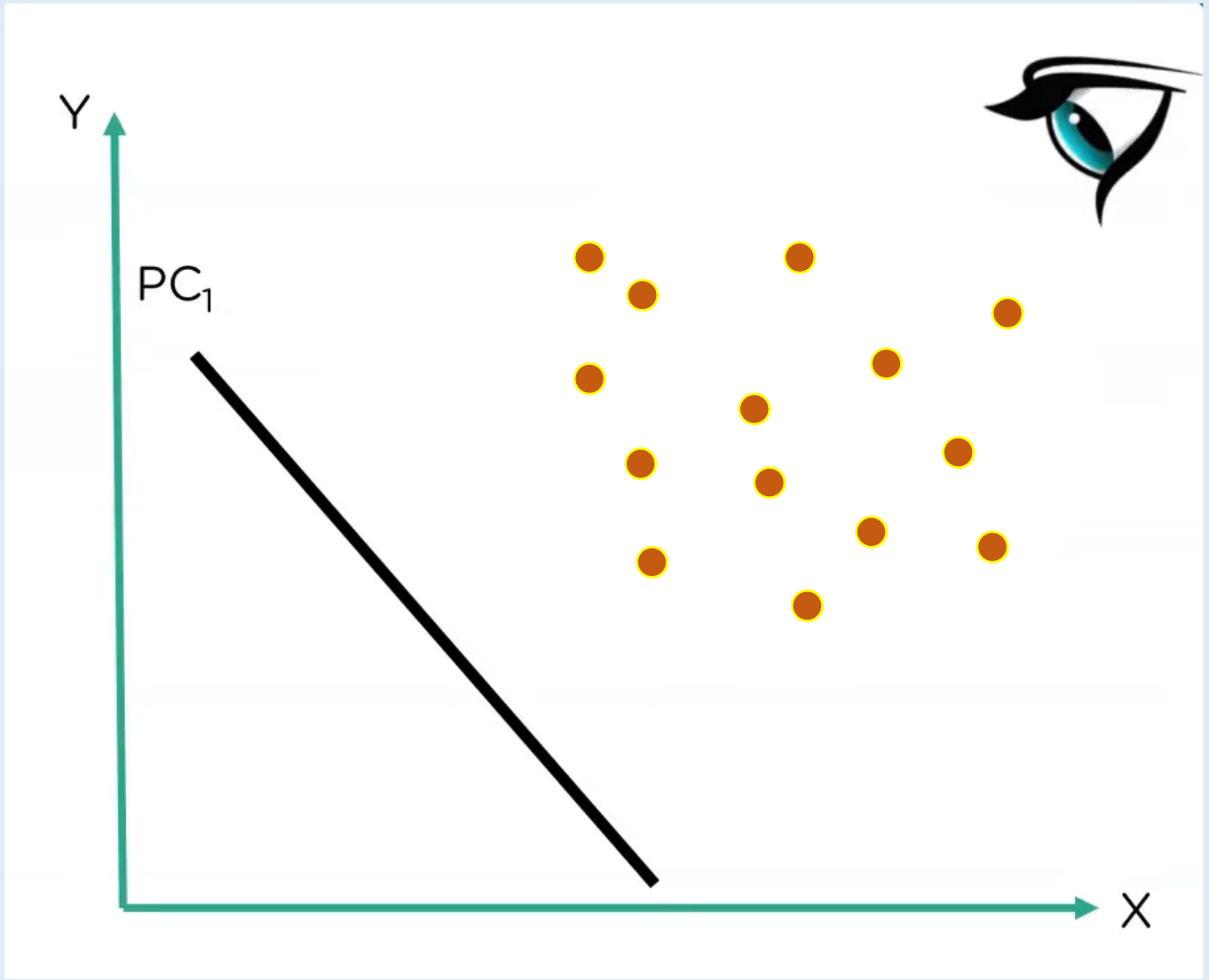
Why to use dimensionality reduction?

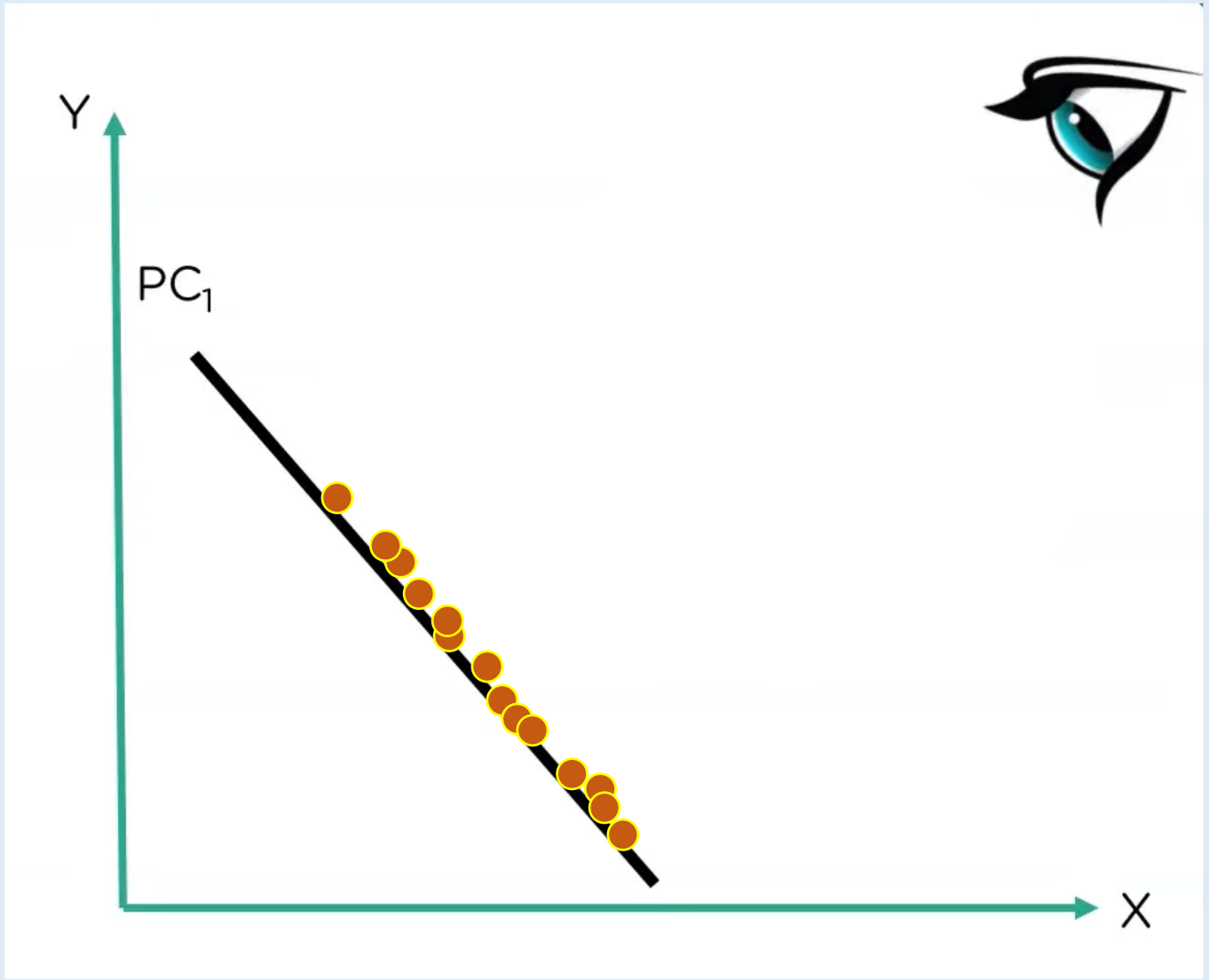
- Less dimensions (Lesser number of feature) leads to less computation and training time
- Redundancy is removed after removing similar entries from the dataset
- To reduce the amount of space required to store the data
- One can plot the data on 2D or 3D plots
- Helps to find out most significant features and skip the rest
- Leads to better human interpretation

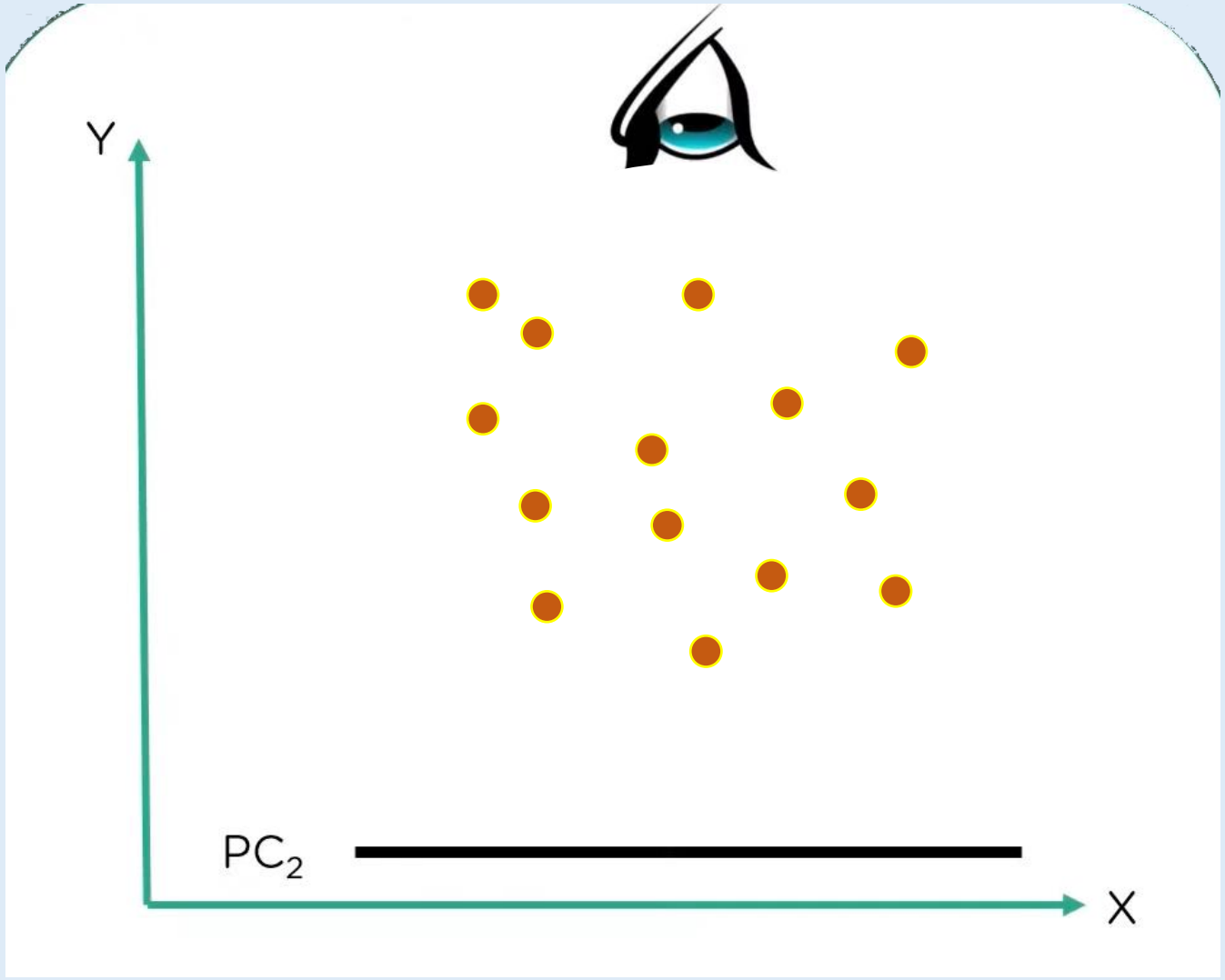
Principal Component Analysis (PCA)

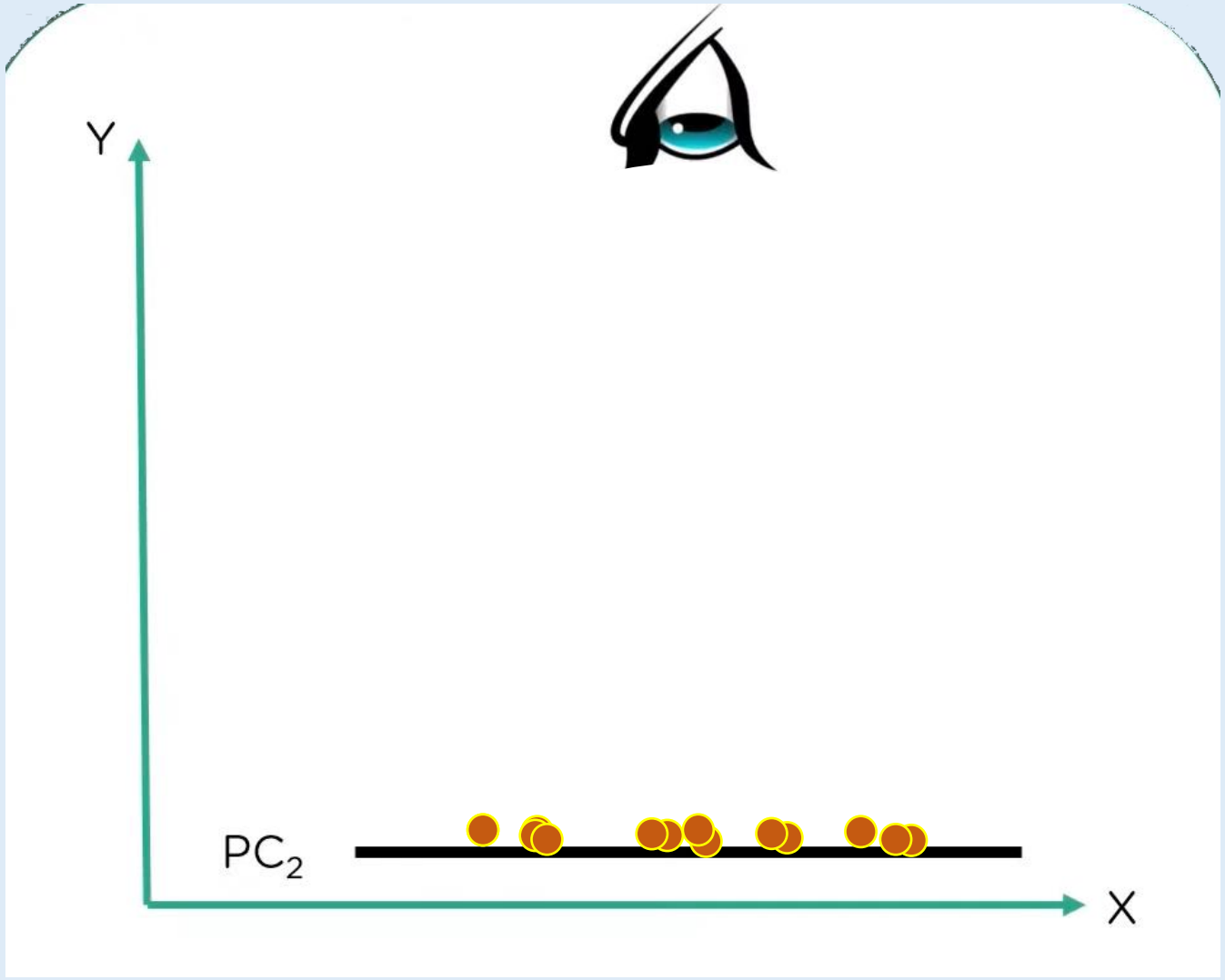
- Principal Component Analysis is a technique for reducing the dimensionality of the datasets, increasing interpretability but at the same time minimizing information loss

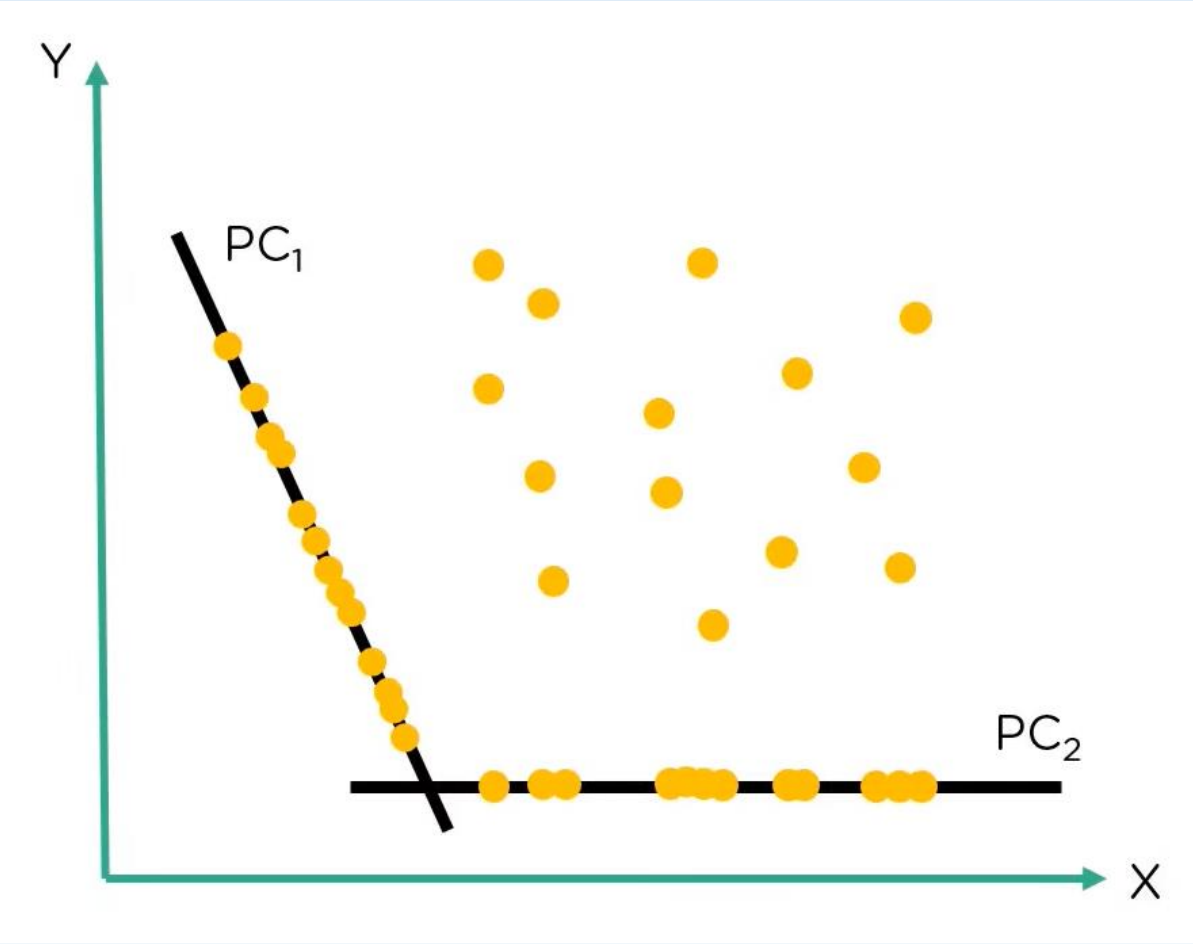


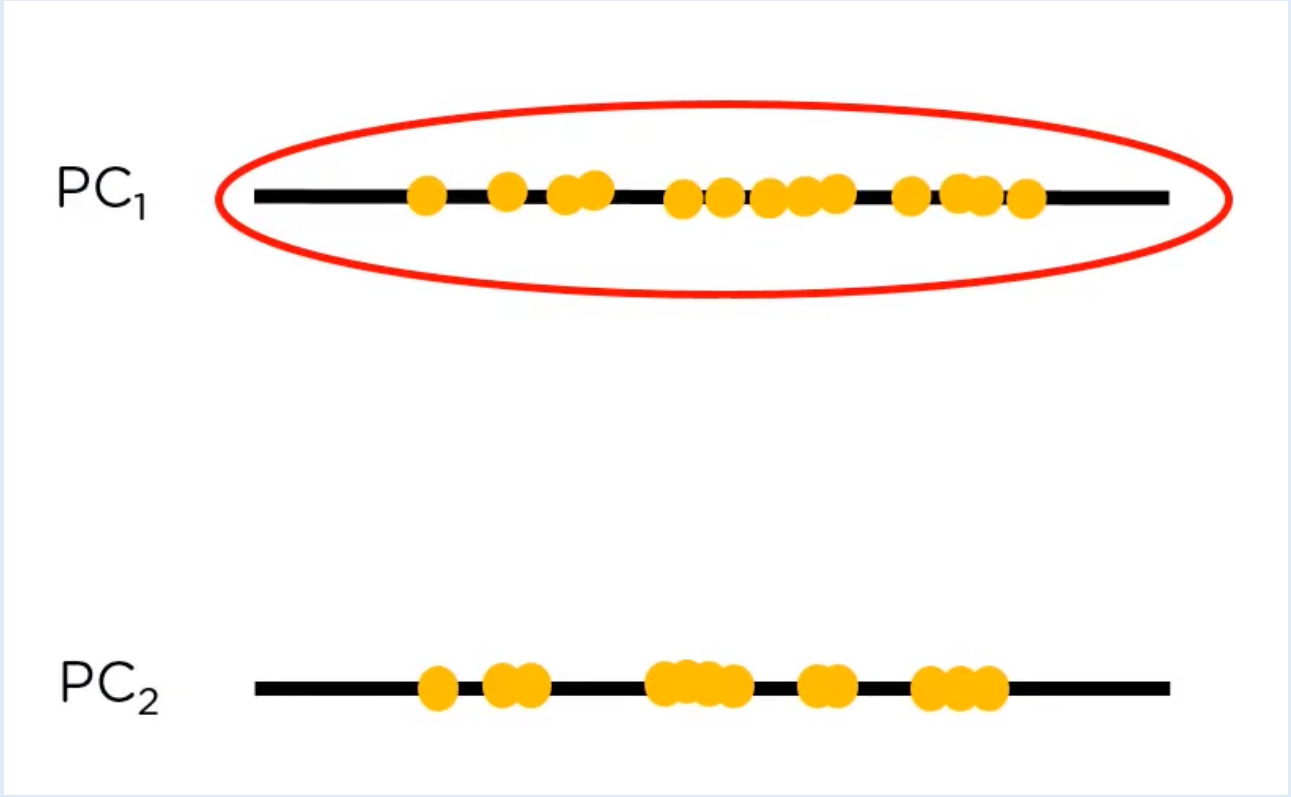

Y
PC1
X


Y
PC1
X


Y
PC2
X


Y
PC2
X


Y
PC1
PC2
X

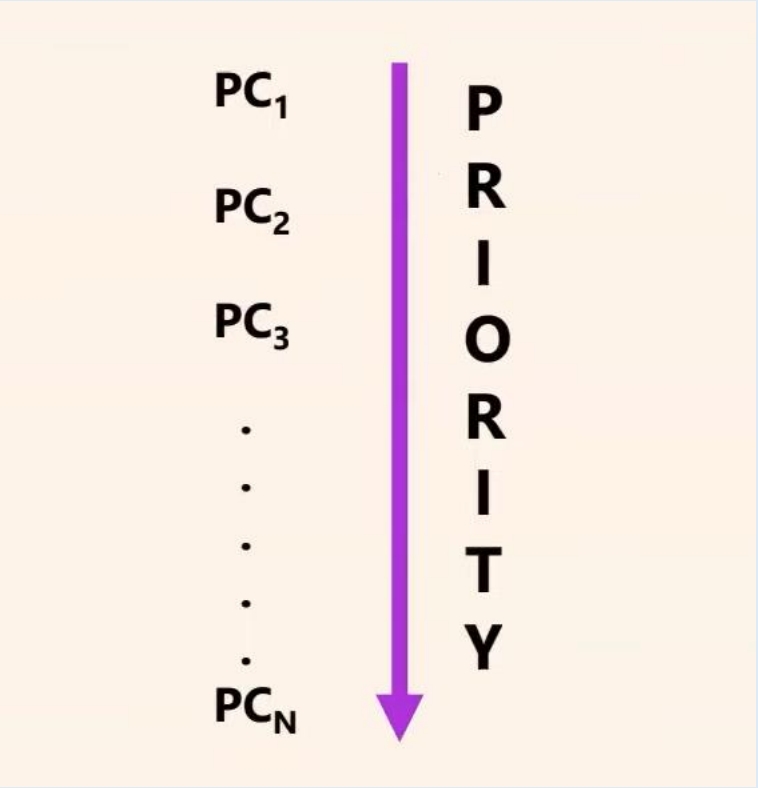

PC1
PC2

Important Terms in PCA

- View: Perspective through which the data points are observed
- Dimension : Number of columns in a dataset
- Principal Component : New variable that are constructed as linear combination or mixture of initial variable
- Projection : The perpendicular distance between the principal component and data points

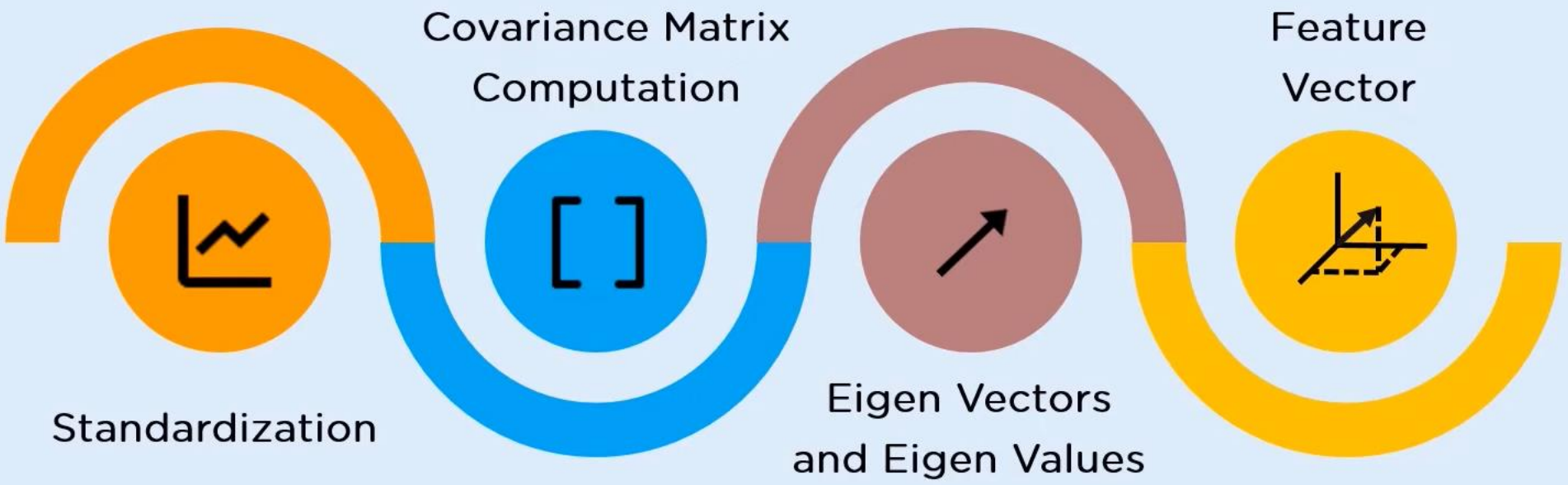
Properties of Principal Component

- Number of principal component always less than or equal to the number of attributes (Features)
- Principal Components are orthogonal
- Priority of Principal component decreases as their number increases



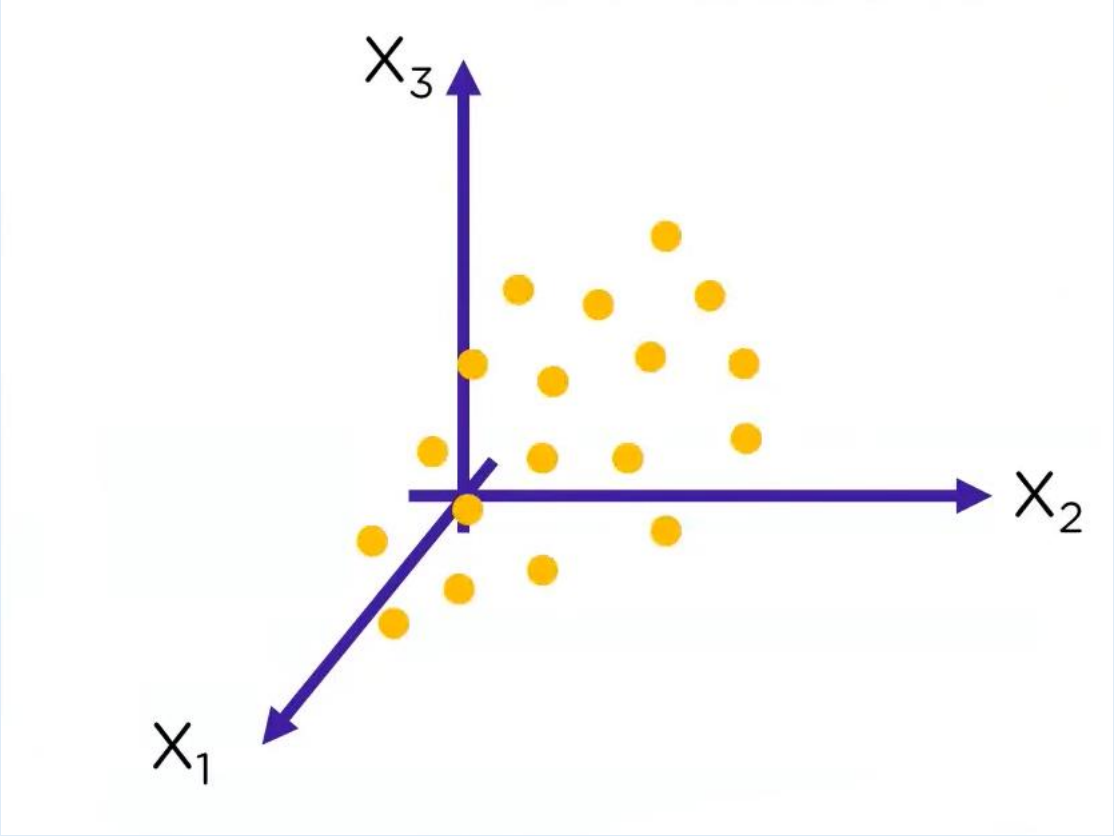
How does the PCA work?

PCA performs following operations on the data so as to obtain the principal components of the data…



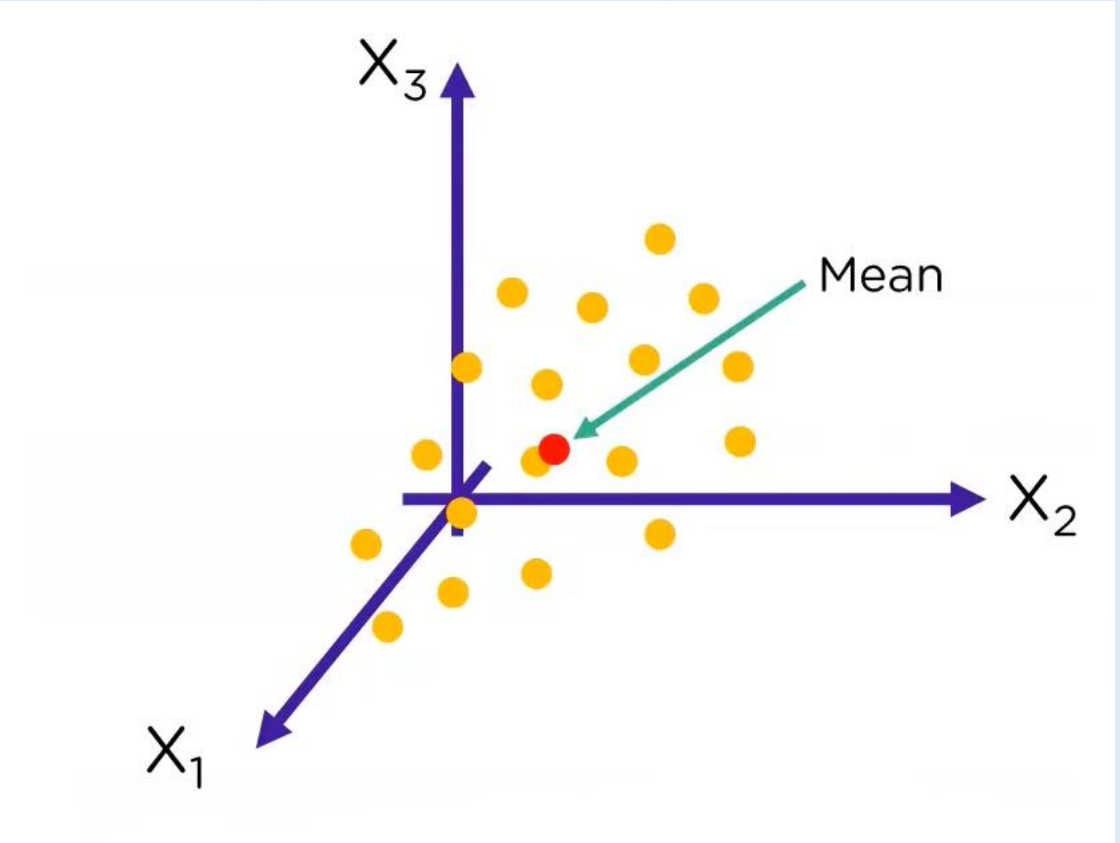
PCA Example

Consider a dataset with three features (Dimensions)



PCA Example

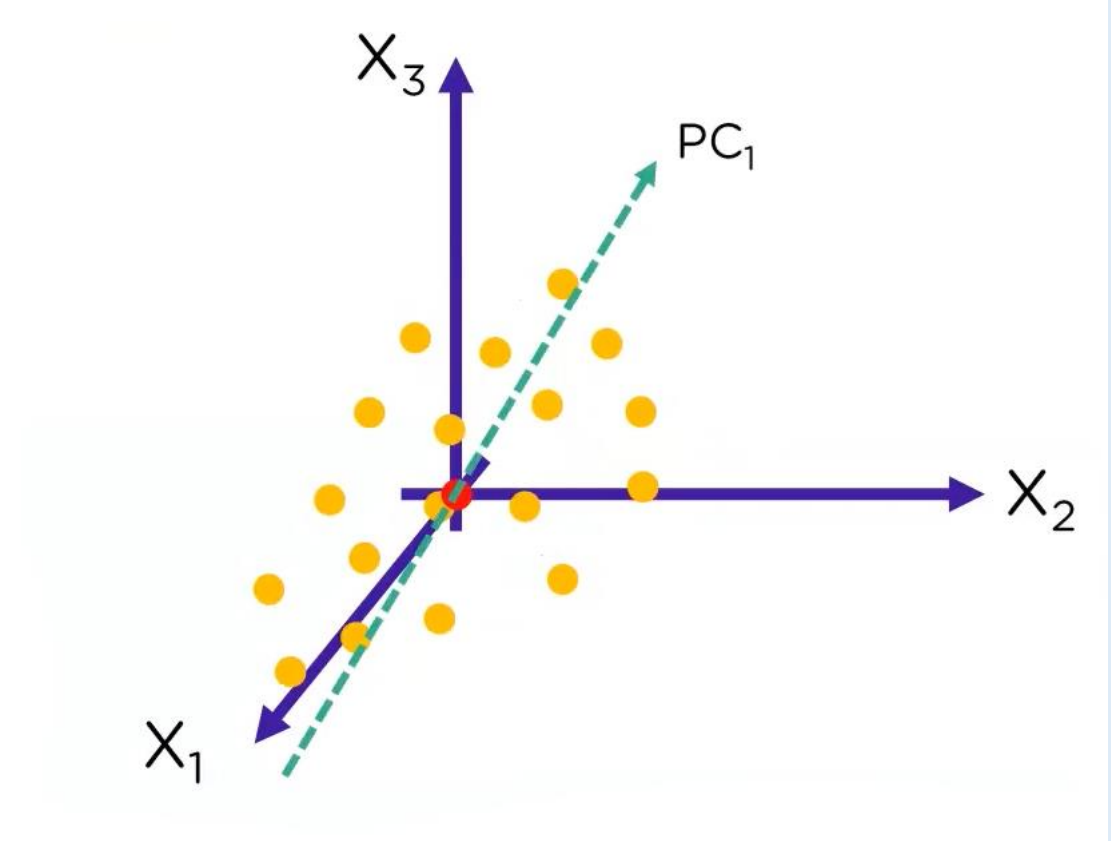
Take mean of all the observation in 3D space



PCA Example

With mathematical operations Principal component one (PC1) is found. It is a line passing through the mean of observation and has direction along maximum variance of the observation

Each observation is projected onto this line in order to get coordinate value along the PC1, This value is known as score.



PCA Example

Similarly, other line passing through mean of observation and orthogonal to PC1 is drawn. PC2 also has direction along best variance direction perpendicular to PC1