**Dimensionality Reduction** – Unsupervised Learning

Why Dimensionality Reduction?

* When many separate group working on a single project, Will have many redundant factors.
* Redundant data set - different units for same attribute-A same feature expressed in different unit will also be a redundant factor and needs to be reduced.

Data Compression

* Speeds up the algorithms
* Reduces space utilized by them

Visualization

It’s difficult to visualize high dimensional data. Dimensionality reduction can explain the Data more intuitively. Data is best understood when it can be visualized. Suppose there are 50 features explaining about a country, this data is reduced to 2 dimensions and hence in this way could be best explained.

**Principal Component Analysis:**

* Most common approach used for Dimensionality Reduction.
* It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, PCA is a powerful tool for **analyzing data**.
* The other main advantage of PCA is that once you have found these patterns in the data, and you **compress** the data, ie. by reducing the number of dimensions, without much loss of information.

**PCA Algorithm**

**Step1:** Before applying PCA must do data preprocessing

1.Mean Normalization

Replace each xji with xj - μj,

2.Feature Scaling (depending on data)

Here 3D data is given , We perform mean Normalization of the data

**Step 2:** Calculate Covariance Matrix

**Step 3**: Calculate Eigen Vector and Eigen Values

**Step 4:** Choosing Principal Components and forming a new feature .

**Step 5:** Finding a new Data Set of reduced dimensionality here, 2D