PRINCIPAL COMPONENT ANALYSIS

Khin Nyo Nyo Theint
Aung Khant Myat
Tun Ye Minn

Knowledge Engineering

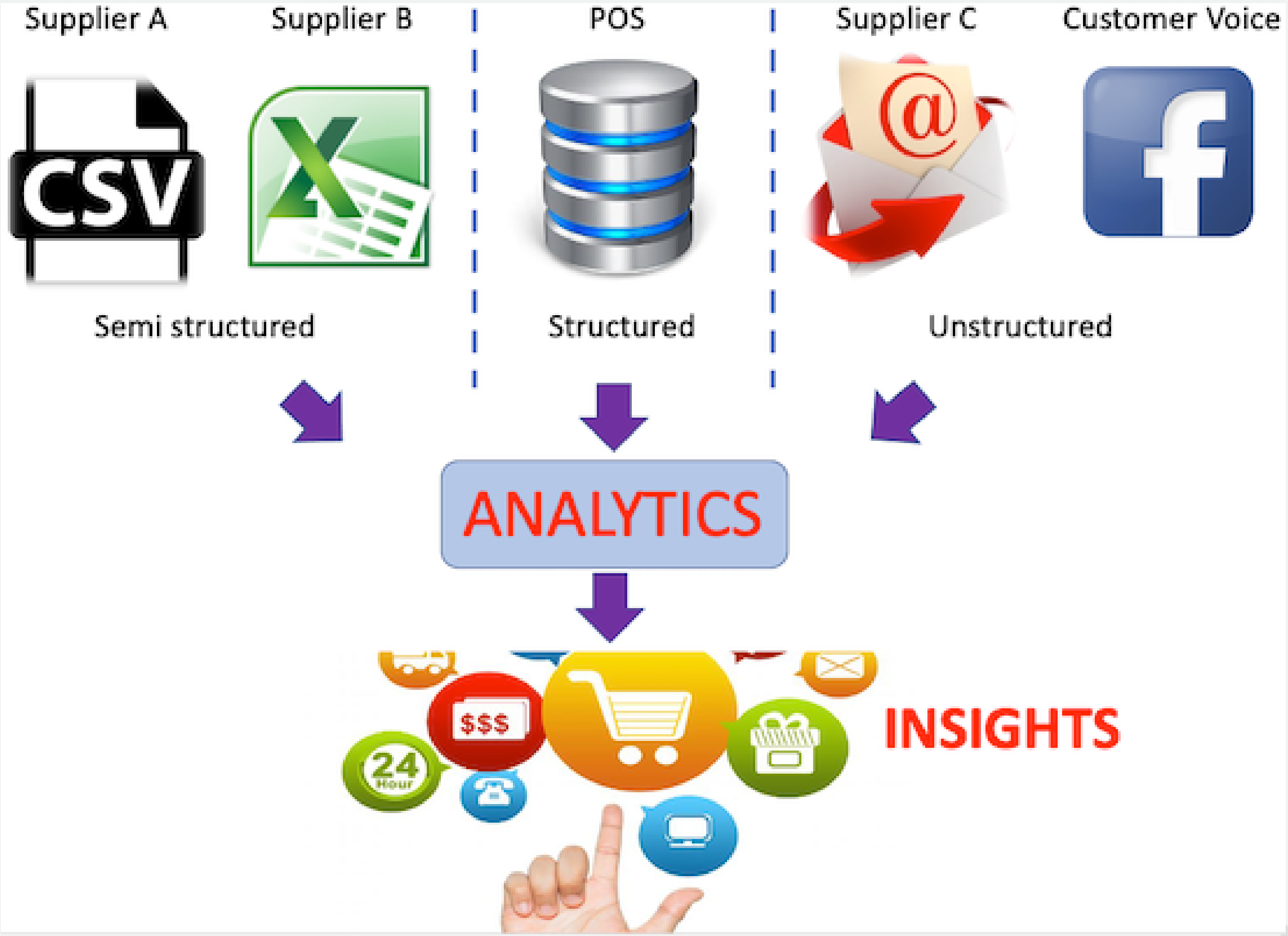
Fourth Year

CONTENTS

WHAT IS DATA REDUCTION?

Motivation - As datasets are growing larger and more complex, mining huge amounts of data can take a long time.

Solution - Data reduction techniques reduce the volume of data yet maintain the integrity of the data.



DATA REDUCTION TECHNIQUES

DIMENSIONALITY REDUCTION

- eliminates the attributes from the data set under consideration thereby reducing the volume of original data.

Methods : wavelet transforms, principal components analysis, Attribute subset selection

NUMEROSITY REDUCTION

- reduces the volume of the original data and represents it in a much smaller form.

Methods : Parametric(Regression and Log-Linear), Non-Parametric(Histogram, Clustering, Sampling, Data Cube Aggregation)

DATA COMPRESSION

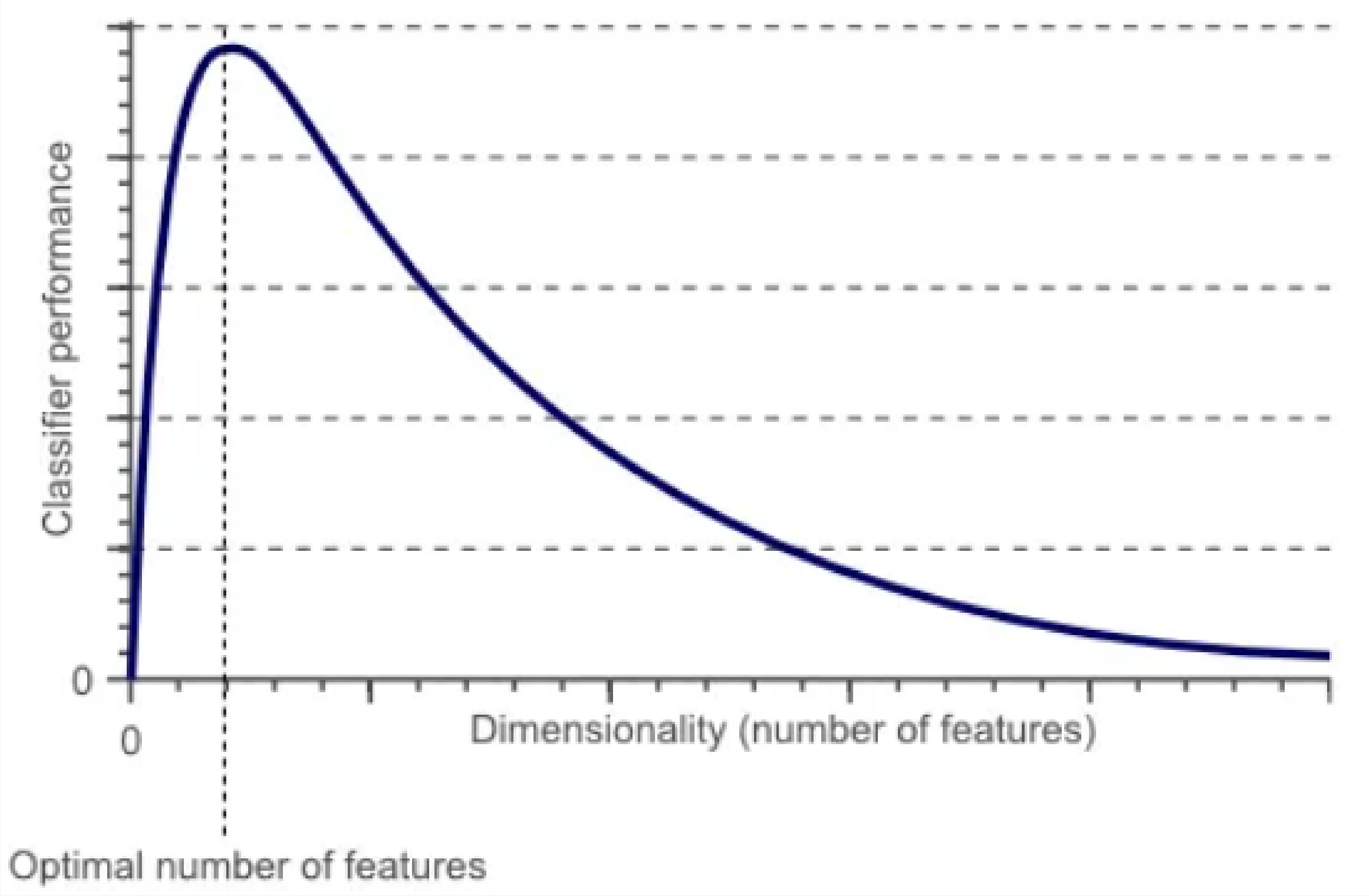
- is a technique where the data transformation technique is applied to the original data in order to obtain compressed data.

Nature - lossless, lossy

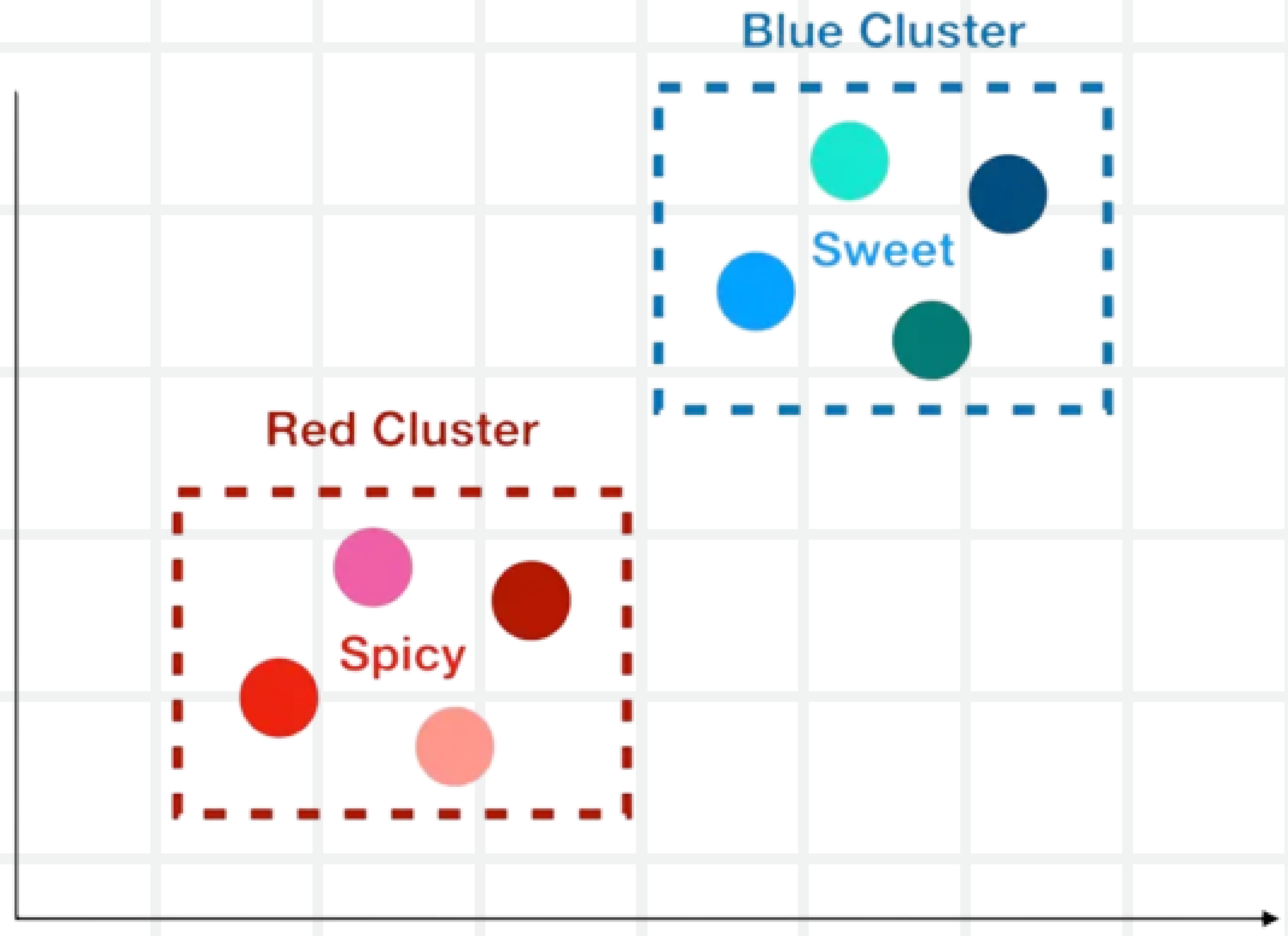
- Dimensionality reduction and numerosity reduction techniques can also be considered forms of data compression.

WHAT IS THE CURSE OF DIMENSIONALITY?

Problem - "Excessively large datasets lead to overfitting, where the model is influenced by outliers and noise."



EXAMPLE FOR CURSE OF DIMENSIONALITY



Our 2 Color Based Clusters of Candy Flavor

EXAMPLE FOR CURSE OF DIMENSIONALITY

	Reddish	Bluish
	1	0
	1	0
	1	0
	1	0
	0	1
	0	1
	0	1
	0	1

Perfect Clusters

EXAMPLE FOR CURSE OF DIMENSIONALITY

	Red	Maroon	Pink	Flamingo	Blue	Turquoise	Seaweed	Ocean
	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	0	0	0	1	0	0	0	0
	0	0	0	0	1	0	0	0
	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	1	0
	0	0	0	0	0	0	0	1

High Dimensional Data Makes Trouble For Clustering

WHAT IS DIMENSIONALITY REDUCTION?

- a statistical/ML-based technique which reduces the number of features in the dataset and obtain a dataset with an optimal number of dimensions.
- One of the most common ways to accomplish Dimensionality Reduction is Feature Extraction, which reduces the number of dimensions by mapping a higher dimensional feature space to a lower-dimensional feature space.
- The most popular technique of Feature Extraction is Principal Component Analysis (PCA)

WHAT IS PRINCIPAL COMPONENT ANALYSIS (PCA)?

- introduced by the mathematician Karl Pearson in 1901.
- maps a higher dimensional feature space to a lower-dimensional feature space.
- identify correlations and patterns in a dataset so that it can be transformed into a dataset of significantly fewer dimensions without loss of any important information.

WHEN TO USE PCA?

UNSUPERVISED ALGORITHM:

It does not require there to be a specific outcome variable you are trying to predict in your dataset.

MANY CORRELATED FEATURES:

The transformed features are guaranteed to be independent of one another no matter how highly correlated the input features were.

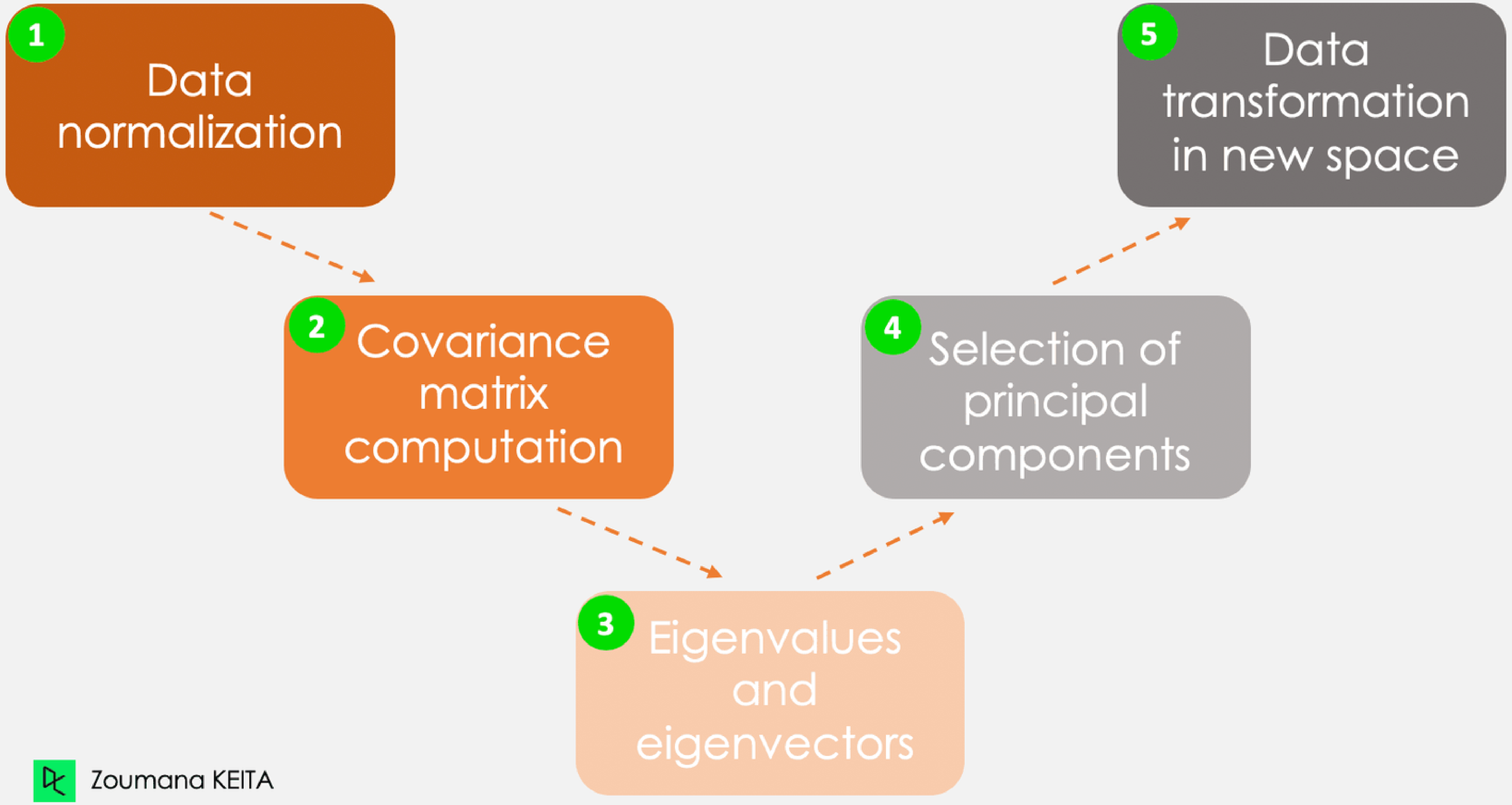
CONTINUOUS DATA(NUMERIC VALUE):

PCA is intended to be used on a set of numeric features.

WHY PCA IS USEFUL?

- **Guaranteed to produce uncorrelated features**: as correlated features tend to cause problems for a lot of machine learning algorithms, the transformed features that come out of the model are guaranteed to be uncorrelated.
- **Relatively fast** : compared to other dimensionality reduction techniques.
- **Not sensitive to choice of seed(initialization conditions)**: a deterministic algorithm, which means that it will always produce the same result when applied to the same dataset.
- **No hyperparameters** : no additional step of hyperparameter tuning.
- **Popular and well studied**: the most common dimensionality reduction techniques, plenty of resources.

HOW DOES PCA WORKS?



Zoumana KEITA

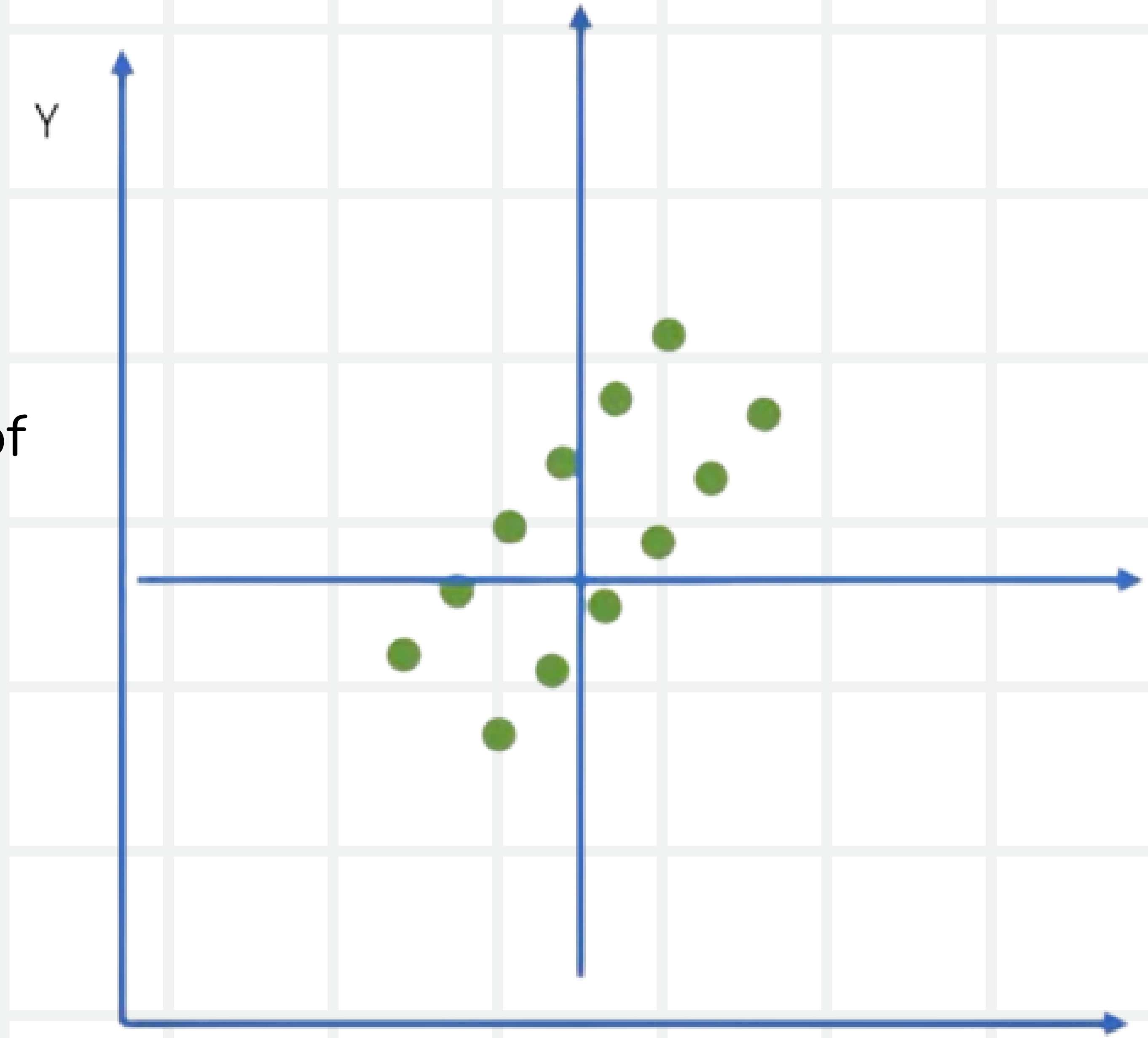
STEP 1 - DATA NORMALIZATION

The following information has different scales and performing PCA using such data will lead to a biased result.

Monthly expenses: $300
Age: 27
Rating: 4.5

Data Normalization ensures that each attribute has the same level of contribution, preventing one variable from dominating others.

	Student 1	Student 2	Student 3	Student 4	...
Math	95	88	93	75	...
Reading	9	8	10	7	...

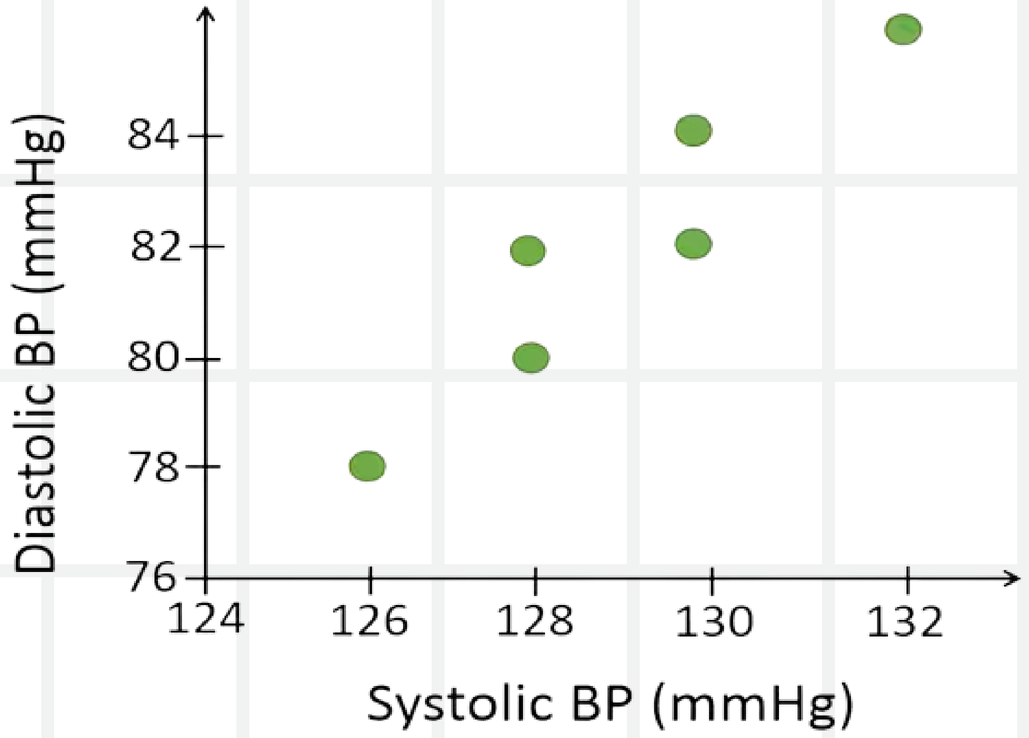


To transform the variables of the same standard, you can follow the following formula.

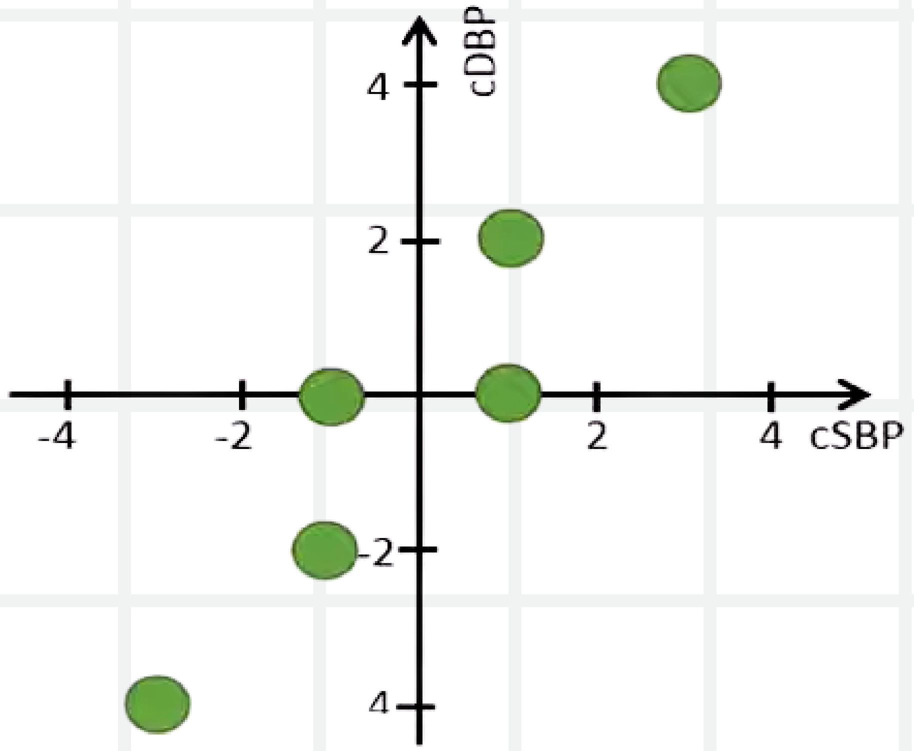
$$Z = \frac{VALUE - MEAN}{STANDARD\ DEVIATION}$$

$$\text{MEAN} = \frac{Sum\ of\ the\ terms}{Total\ number\ of\ terms}$$

$$\text{STANDARD DEVIATION} = \sqrt{\frac{\in(x - mean)^2}{n}}$$



Before



After

STEP 2 - COVARIANCE MATRIX

A covariance matrix is a N x N symmetrical matrix that contains the covariances of all possible data sets.

Features = X1, X2, X3, ..., Xn

$$Cov\ Matrix = \begin{bmatrix} Var(x1,x1) & Cov(x1,x2) & \dots\dots\dots & Cov(x1,xn) \\ Cov(x2,x1) & Var(x2,x2) & \dots\dots\dots & Cov(x2,xn) \\ Cov(xn,x1) & Cov(xn,x2) & \dots\dots\dots & Var(xn,xn) \end{bmatrix}_{n \times n}$$

Properties:

1. Square Matrix
2. Symmetric

$$\text{Covariance} = \frac{Sum\ (\ X-(Mean\ of\ X)(\ Y-(Mean\ of\ Y))}{Number\ of\ data\ points}$$

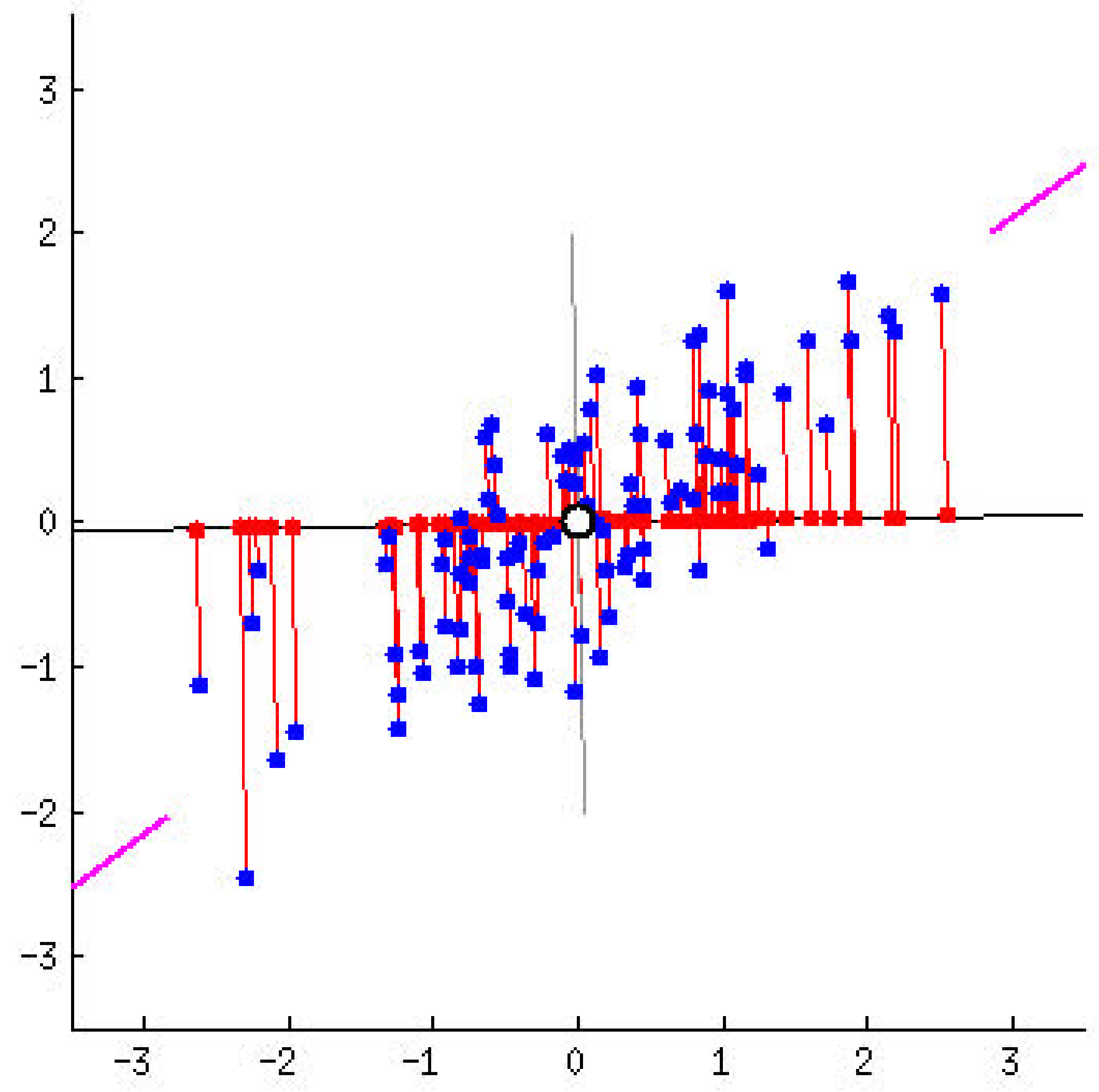
+ve Trend   -ve Trend   No Trend

STEP 3 - EIGENVECTORS AND EIGENVALUES

$$\mathbf{Av} = \lambda\mathbf{v}$$

λ, called the corresponding eigenvalue.

- Once the eigenvector components have been computed, define eigenvalues in descending order (for all variables) and now we will get a list of principal components.
- The eigenvalues represent the principal components and these components represent the direction of data.
- If the line contains large variables of large variances, then there are many data points on the line. Thus, there is more information on the line too.

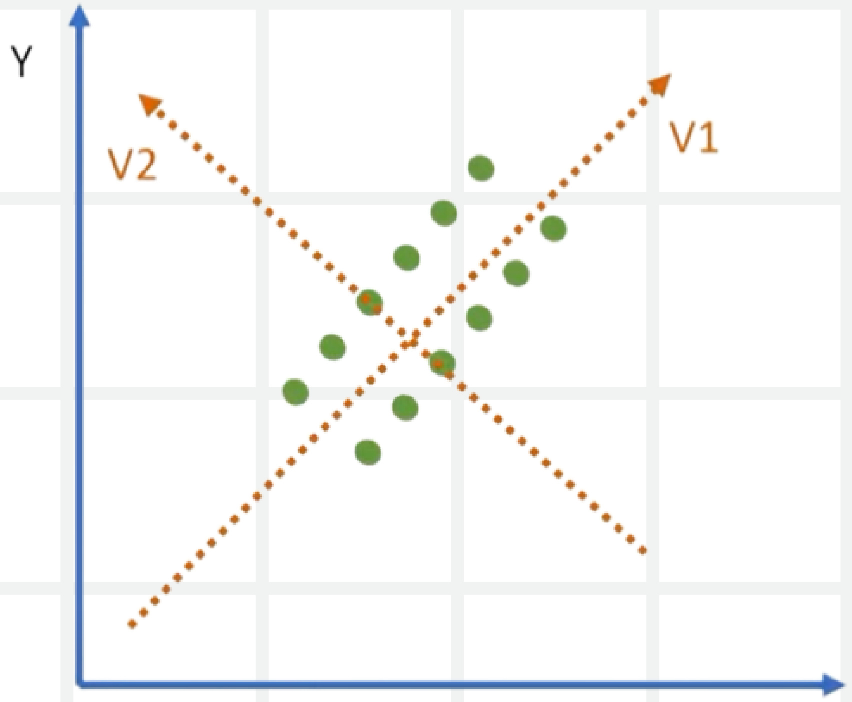


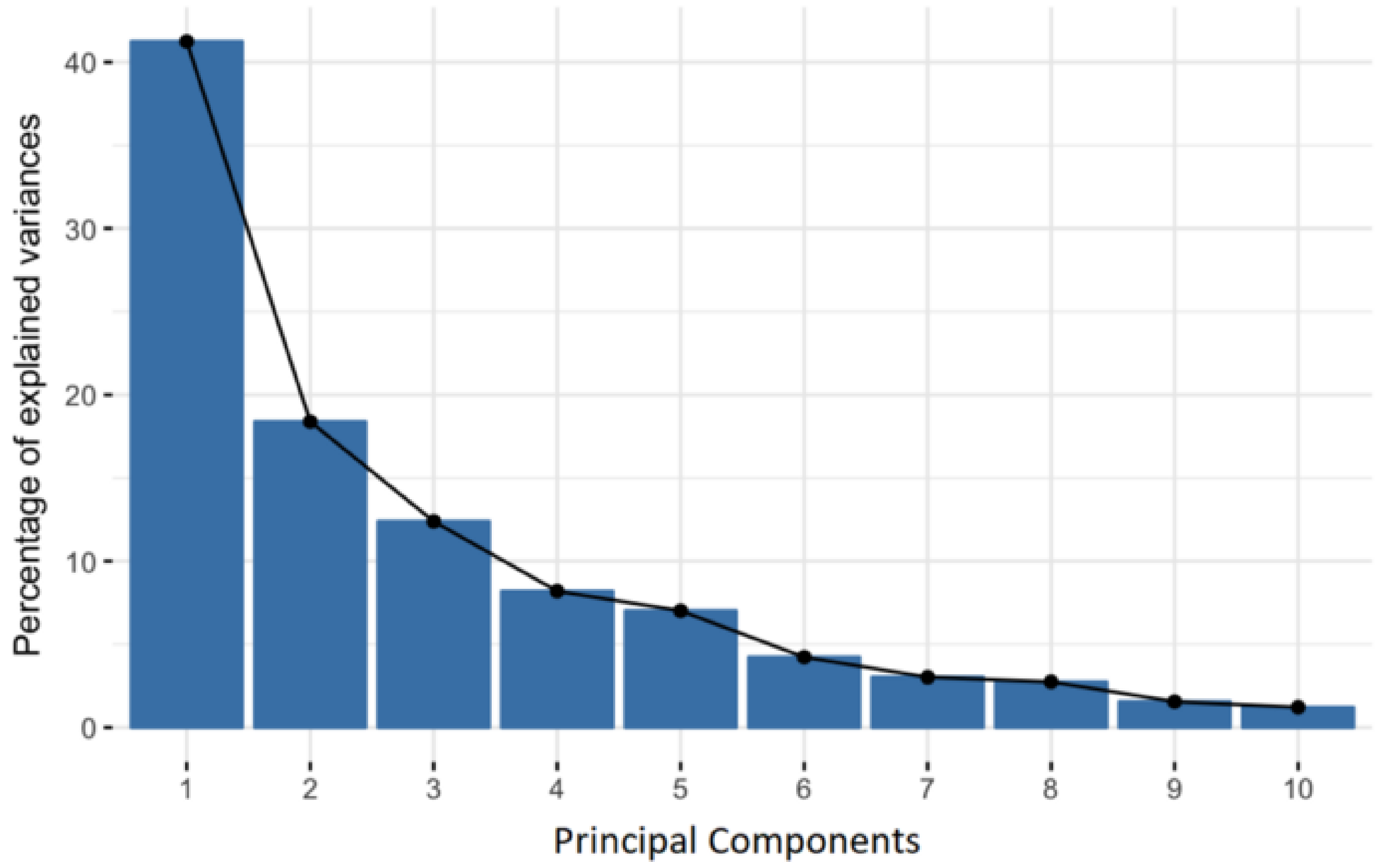
STEP 4 - SELECTION OF PRINCIPAL COMPONENTS

- There are as many pairs of eigenvectors and eigenvalues as the number of variables in the data.
- The eigenvector with the highest eigenvalue corresponds to the first principal component.
- The second principal component is the eigenvector with the second highest eigenvalue, and so on.

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$





STEP 5 - DATA TRANSFORMATION IN NEW DIMENSIONAL SPACE

- Still now, apart from standardization, we haven't made any changes to the original data.
- We have just selected the Principal components and formed a feature vector.
- Yet, the initial data remains the same on their original axes.
- This step aims at the reorientation of data from their original axes to the ones we have calculated from the Principal components.
- This can be done by the following formula.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Note: It is important to remember that this transformation does not modify the original data itself but instead provides a new perspective to better represent the data.

PCA'S MATHEMATICAL EXAMPLE

PCA Problem

Given the Following data, use PCA to reduce the dimension from 2 to 1.

Feature	Example 1	Example 2	Example 3	Example 4
x	4	8	13	7
y	11	4	5	14

Step(1) : Data-Set

No of features, $n = 2$

No of Samples, $N = 4$

Step(2): Computation of mean of variables

$\bar{x} = \frac{4+8+13+7}{4} = 8$, $\bar{y} = \frac{11+4+5+14}{4} = 8.5$

Step(3) Computation of Covariance matrix

Ordered pairs are - (x,x), (x,y), (y,x), (y,y)

$\text{cov}(x,x) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i-\bar{x})^2$

$= \frac{1}{4-1}[(4-8)^2 + (8+8)^2 + (13-8)^2 + (7-8)^2]$

$= 14$

$\text{cov}(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$

$= \frac{1}{4-1}[(4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)]$

$= -11$

$\text{cov}(y,x) = -11$, $\text{cov}(y,y) = 23$

Covariance Matrix $\Rightarrow$ $S = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix}$

$= \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$

Step(4) Eigen Value, Eigen Vector, Normalized Eigen vector

(i) Eigen values; $\lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$

$\det(S - \lambda I) = 0$

$\det\left(\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix}\right) = 0$

$(14-\lambda)(23-\lambda) - (-11\times-11) = 0$

$\lambda^2 - 37\lambda + 201 = 0$

$\lambda = 30.3849, 6.6151$

$\lambda_1 > \lambda_2$,

$\lambda_1 = 30.3849 \Rightarrow$ First Principal Component.

$\lambda_2 = 6.6151$

(ii) Eigen Vector of λ_1, $U_1 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$

$(S - \lambda_1 I)U_1 = 0$

$\begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$

$\begin{bmatrix} (14-\lambda_1)u_1 - 11u_2 \\ -11u_1 + (23-\lambda_1)u_2 \end{bmatrix} = 0$

$(14-\lambda_1)u_1 - 11u_2 = 0$

$-11u_1 + (23-\lambda_1)u_2 = 0$

$\frac{u_1}{11} = \frac{u_2}{14-\lambda_1} = t$

where t=1,

$U_1 = 11 \quad , U_2 = 14 - \lambda_1$

Eigen Vector U_1 of $\lambda_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix}$

$= \begin{bmatrix} 11 \\ 14-30.3849 \end{bmatrix} = \begin{bmatrix} 11 \\ -16.3849 \end{bmatrix}$

(iii) Normalize the eigen vector V_1 ;

$$e_1 = \begin{bmatrix} 11 / \sqrt{11^2 + (16.385)^2} \\ -16.3849 / \sqrt{11^2 + (-16.385)^2} \end{bmatrix}$$

$$= \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

For λ_2 ,

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step (5) Derive new Dataset

	Ex 1	Ex 2	Ex 3	Ex 4
First principal component (PC1)	P_{11}	P_{12}	P_{13}	P_{14}

$P_{11} = e_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \end{bmatrix}$

$= [0.5574 \; -0.8303] \begin{bmatrix} -4 \\ 2.5 \end{bmatrix}$

$= -4.3052$

$P_{12} = e_1^T \begin{bmatrix} 8-8 \\ 4-8.5 \end{bmatrix}$

$= [0.5574 \; -0.8303] \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$

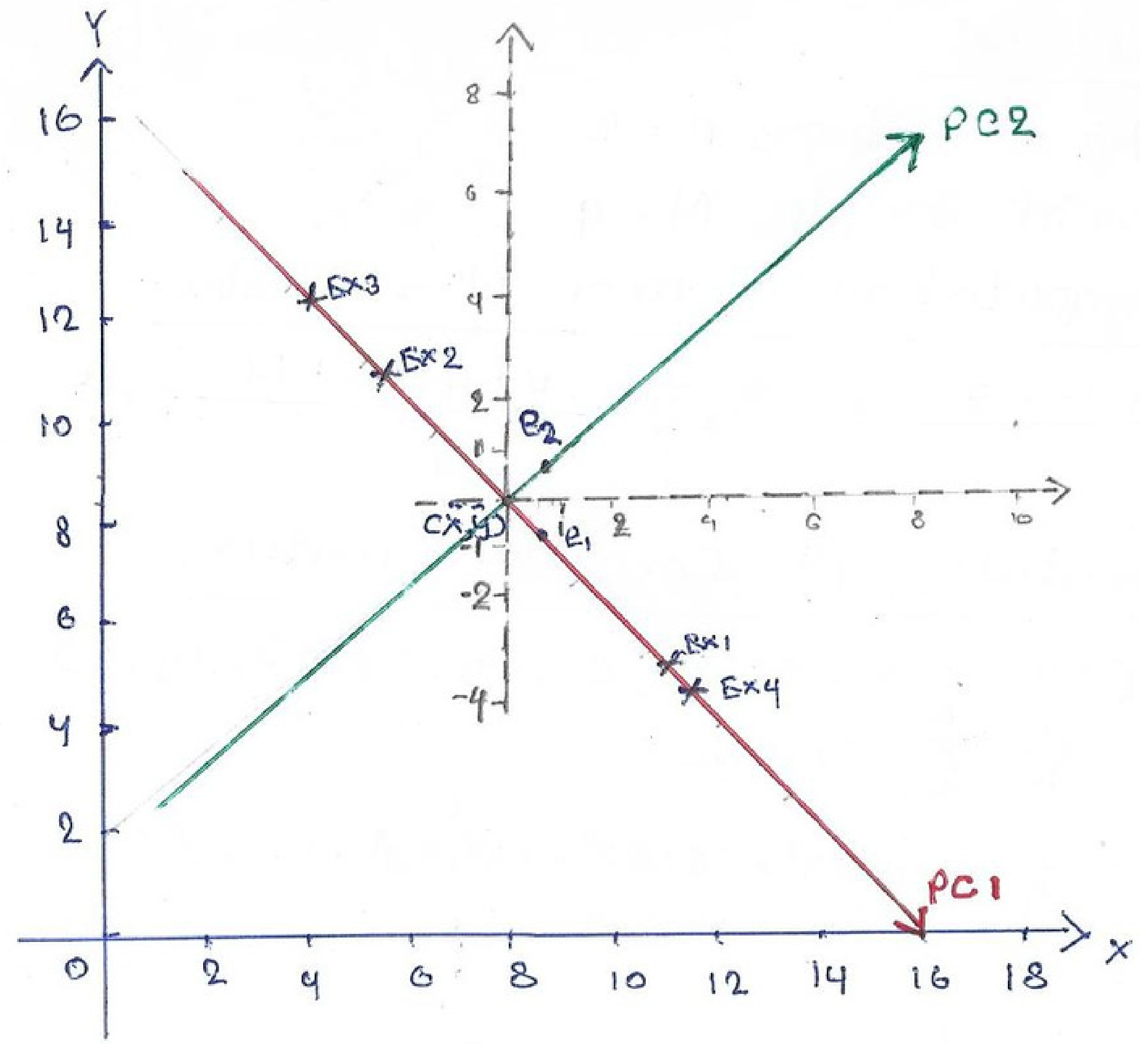
$= 3.7361$

$P_{13} = 5.6928$

$P_{14} = -5.1238$

	Ex 1	Ex 2	Ex 3	Ex 4
PC 1	-4.3052	3.7361	5.6928	-5.1238

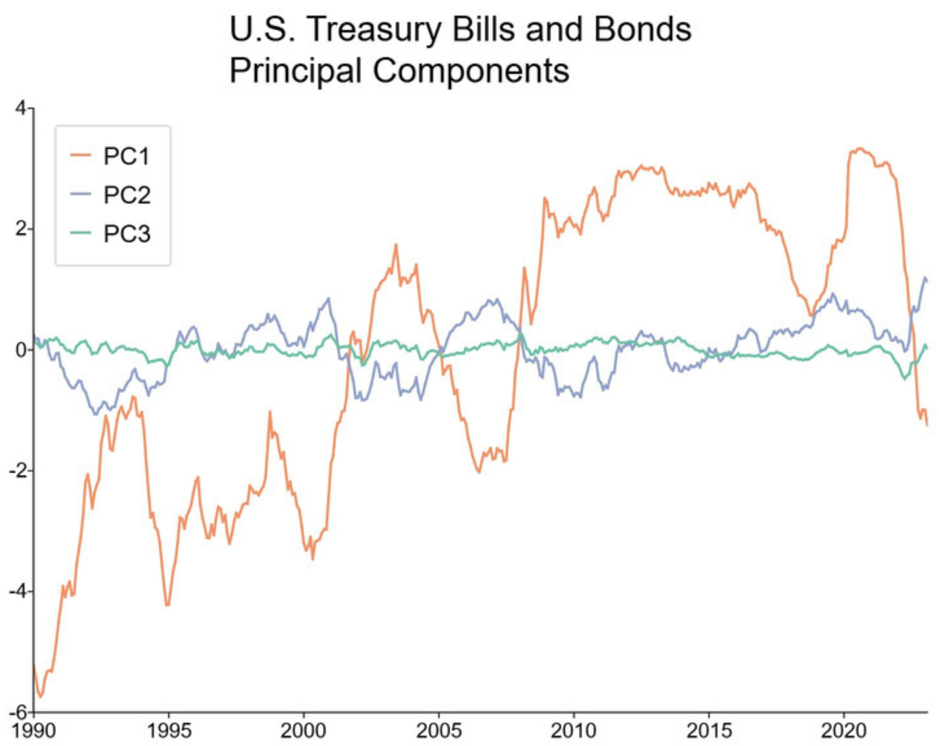
Cordinate System For Principal Components



APPLICATIONS OF PRINCIPAL COMPONENT ANALYSIS

FINANCE

- PCA can assist in forecasting stock prices by reducing dimensionality and identifying significant components that capture most of the data's variability, benefiting experts in their analysis.



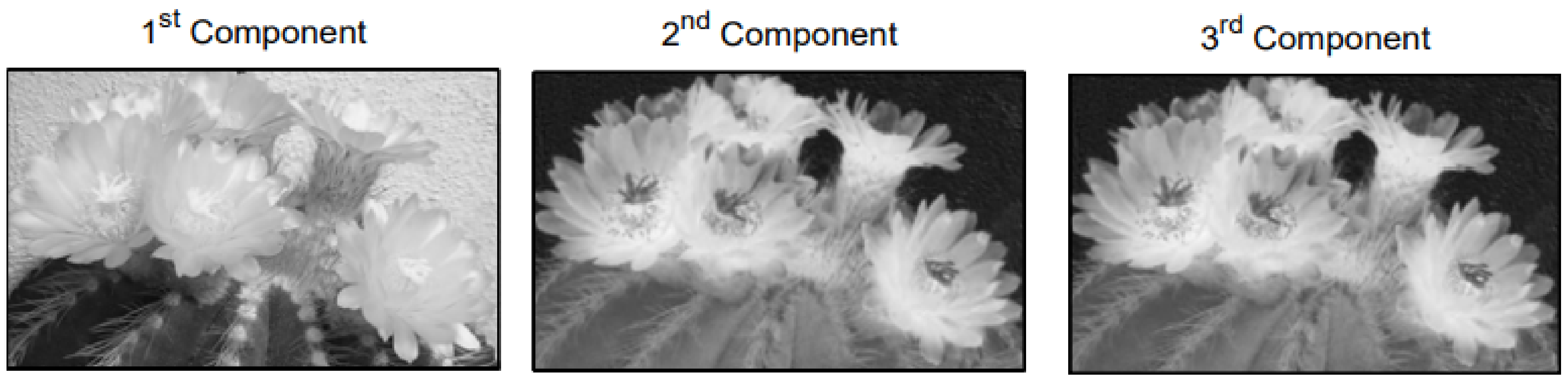


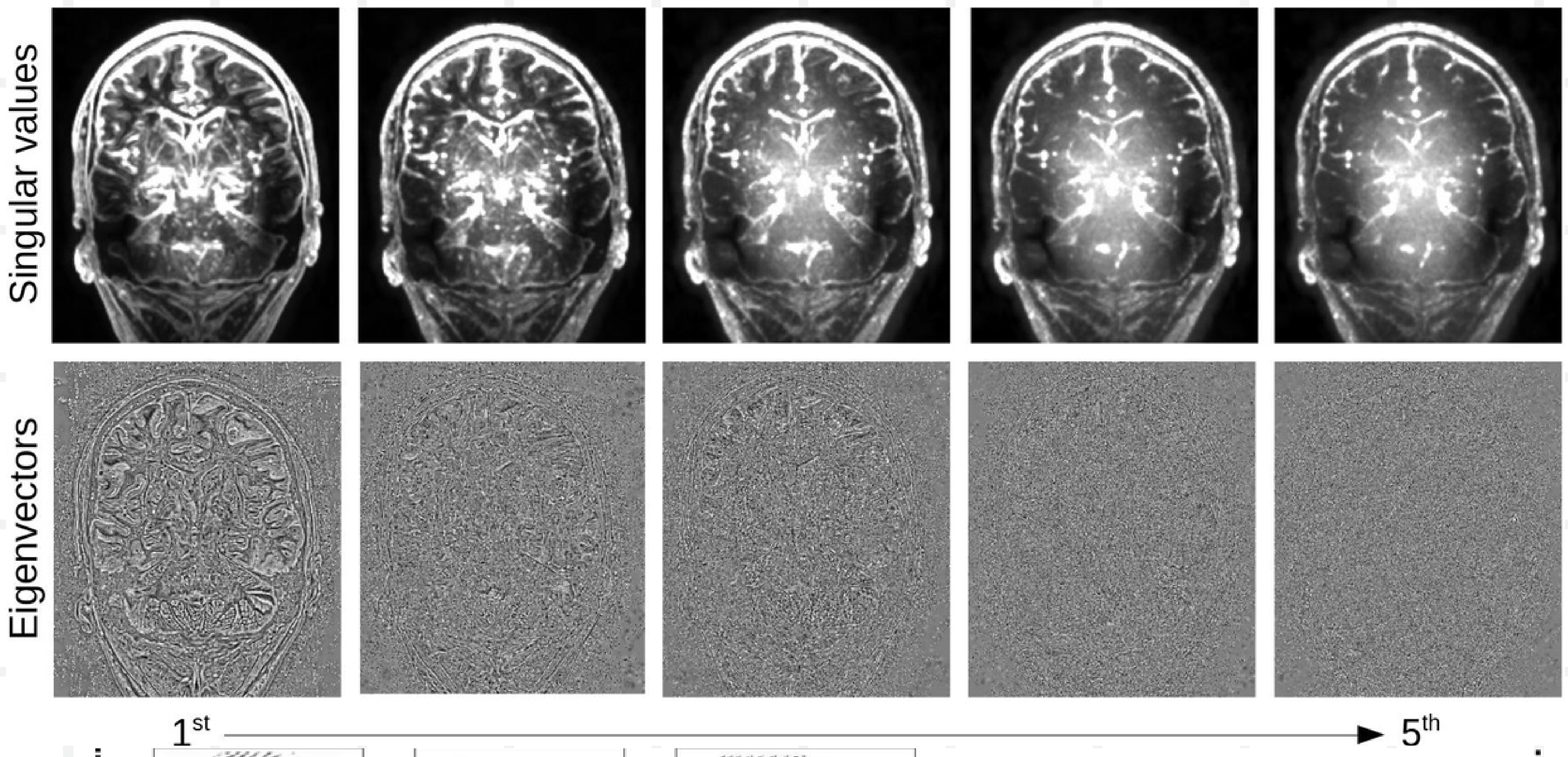
Figure 2: PCA of a selected image

IMAGE PROCESSING

- An image is made of multiple features.
- PCA is mainly applied in image compression to retain the essential details of a given image while reducing the number of dimensions.
- PCA can be used for more complicated tasks such as image recognition.

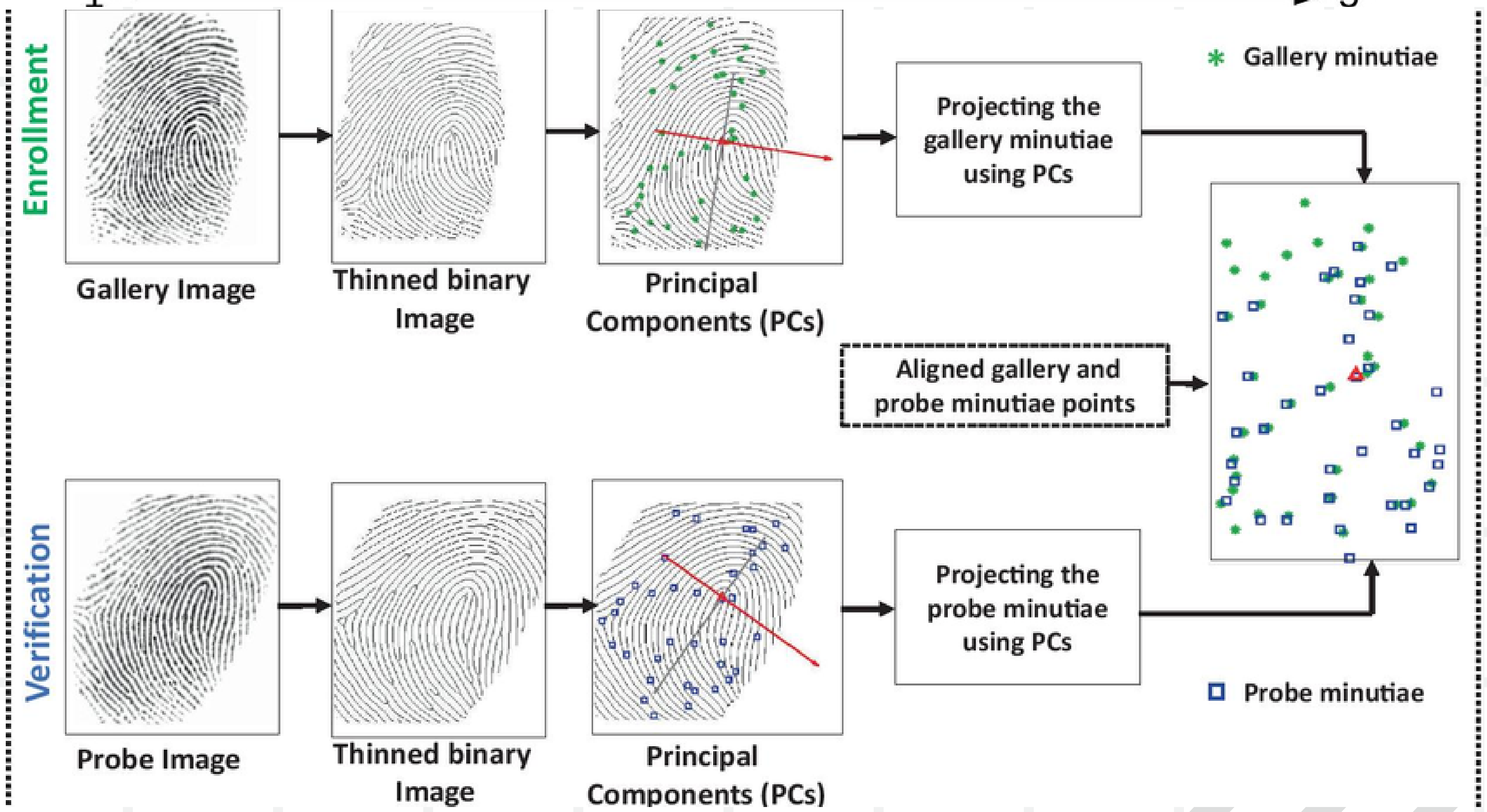
HEALTHCARE

- PCA is used in magnetic resonance imaging (MRI) scans to reduce the dimensionality of the images for better visualization and medical analysis.
- It can also be integrated into medical technologies to recognize a given disease from image scans.



SECURITY

- Biometric systems used for fingerprint recognition can integrate technologies leveraging principal component analysis to extract the most relevant features, such as the texture of the fingerprint and additional information.



LIMITATIONS OF THE PCA

Principal Component Analysis (PCA) is a great dimensionality reduction technique, but it does have some limitations:

1. **Linearity:** PCA assumes a linear relationship between variables, which means it might not work well for datasets with nonlinear relationships.
2. **Data loss:** PCA reduces the dimensionality by projecting data onto a new subspace, leading to some loss of information. The explained variance of the retained components may not capture all the important patterns in the original data.
3. **Interpretability:** The transformed components in PCA are combinations of the original features, making it harder to interpret their physical meaning compared to the original variables.

LIMITATIONS OF THE PCA

4. **Outliers:** PCA is sensitive to outliers as they can influence the principal components, potentially leading to distorted results.
5. **Scaling:** PCA is sensitive to the scale of features, so it's essential to standardize or normalize the data before applying PCA to ensure meaningful results.
6. **Large datasets:** For very high-dimensional datasets, PCA can become computationally expensive and may not be feasible to apply. Find another technique or feature selection.
7. **Selecting the number of components:** Determining the appropriate number of principal components to retain can be subjective and requires careful consideration.

REAL-WORLD EXAMPLE OF PCA IN PYTHON

Breast Cancer accuracy evaluation using Logistic Regression Algorithm before and after applying PCA

THANK YOU