Data Science with Python

Dimensionality Reduction

Agenda

01 What is Dimensionality Reduction?

02 Benefits of Dimensionality Reduction

03 Principal Component Analysis (PCA)

04 How does PCA work?

05 Factor Analysis

06 How does Factor Analysis work?

07 Factor Analysis vs PCA

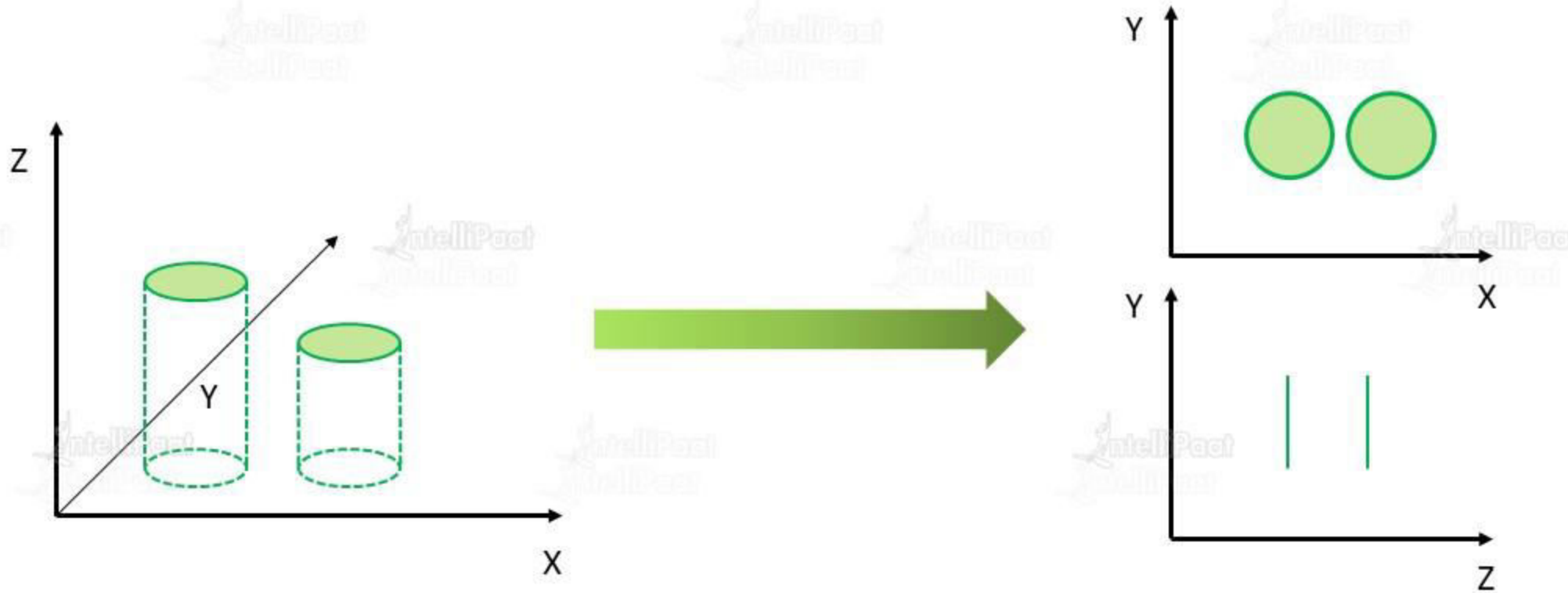
08 What is LDA?

09 How does LDA work?

What is Dimensionality Reduction?

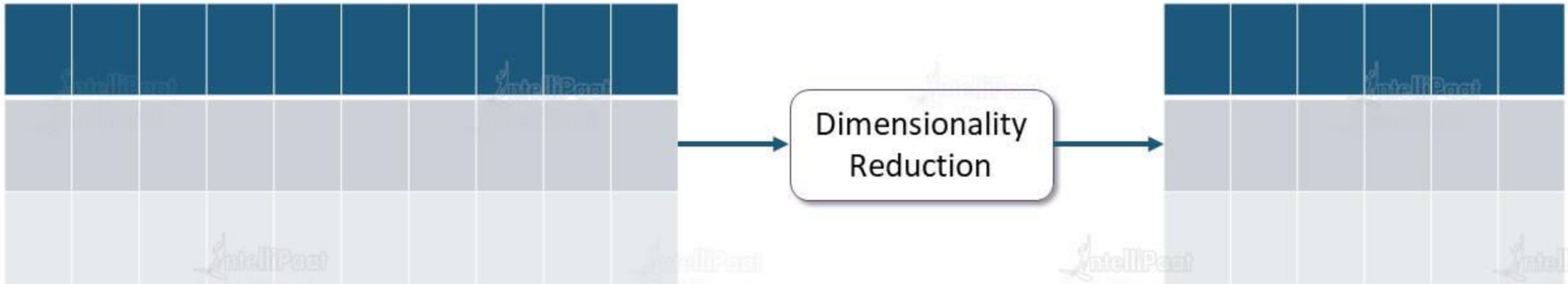
What is Dimensionality Reduction?

The process of converting a dataset of vast dimensions into data with lesser dimensions to reduce the complexity of data by keeping the relevant structure



What is Dimensionality Reduction?

Suppose, we have a table with 10 columns, but around 6 columns might be useless for our analysis. Thus, we can remove the columns using dimensionality reduction



What is Dimensionality Reduction?

Components of Dimensionality Reduction

Feature Elimination

Feature Extraction

What is Dimensionality Reduction?

Feature Elimination

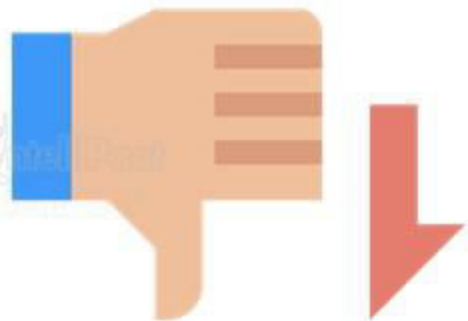
Feature Extraction

Removing some variables completely if they are:

- Redundant with other variables
- Not providing any new information

Sets a smaller dataset

Might loose some data



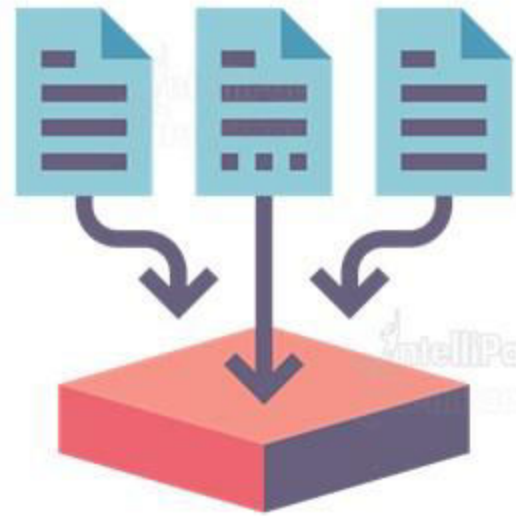
What is Dimensionality Reduction?

Feature Elimination

Feature Extraction

Extracting new variables from old variables

PCA works based on feature extraction. This is the first step we do on our preprocessed dataset

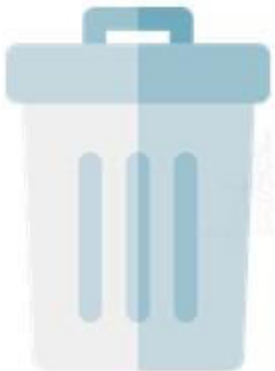


Benefits of Dimensionality Reduction

Benefits of Dimensionality Reduction

When we keep on adding features without increasing the training samples, the dimension keeps growing. This will give us an overfitted ML model that works great with the training data but might fail on new data. This makes the model unpredictable

This is where dimensionality reduction helps!
We remove irrelevant and redundant features, and we also create new features.



Benefits of Dimensionality Reduction

It decreases the unwanted dimensions in Machine Learning. Each data will be saved with little incremental information. The data has to be treated to reduce the number of dimensions

When you checking out a bikes dataset, you don't require the below tools. So, they are unwanted.

- GPS sensors
- Gyro meters
- Flexible
- Video feeds
- Smart devices

Benefits of Dimensionality Reduction

Image processing is an application of dimensionality reduction

Benefits of Dimensionality Reduction

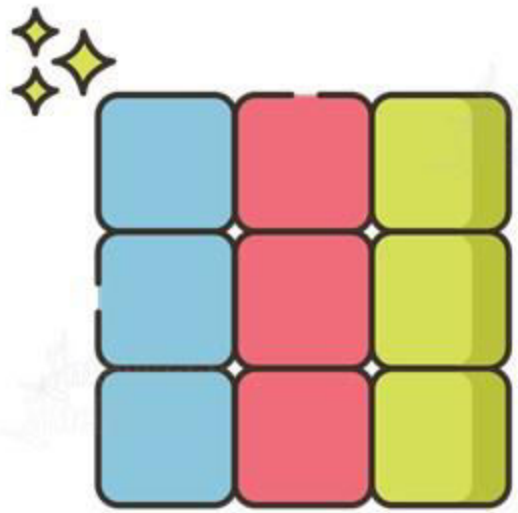
In short dimensionality reduction helps us reduce the size of data to be processed in such a way that it does not affect the performance of our model to a large degree

Principal Component Analysis

Principal Component Analysis

PCA is a dimensionality reduction technique used to create a dataset of a lower dimension without loosing any valuable information

PCA is used to figure out patterns and correlations in a dataset



Principal Component Analysis

PCA is a dimensionality reduction technique used to create a dataset of a lower dimension without loosing any valuable information

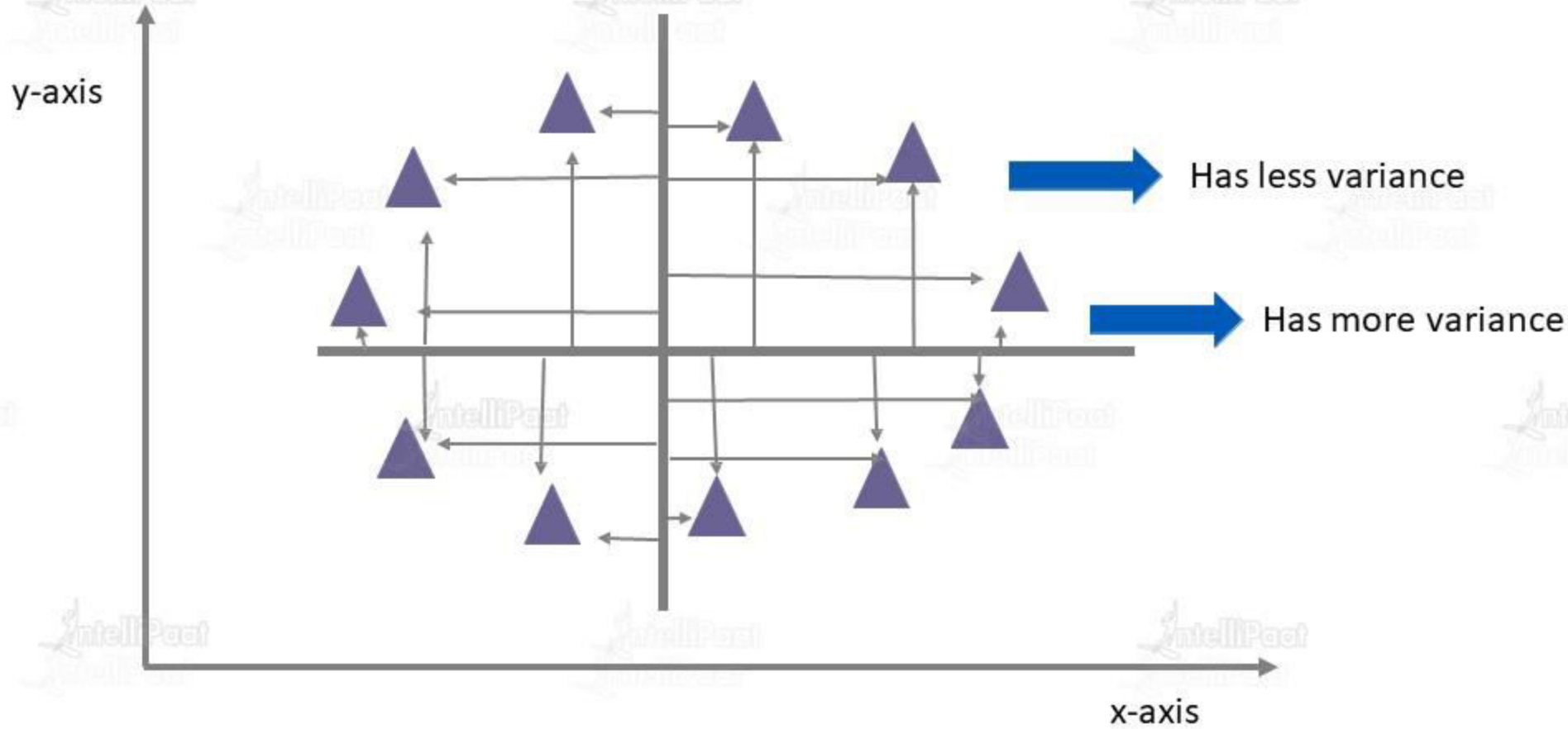
Name	Number of Wheels	Color	Height	Number of Seats
Mercedes	4	Red	4 feet	5
BMW	4	Blue	3 feet	5
Marco Polo	6	Blue	8 feet	10
Volkswagen	4	White	5 feet	5

Has low variance

Has high variance

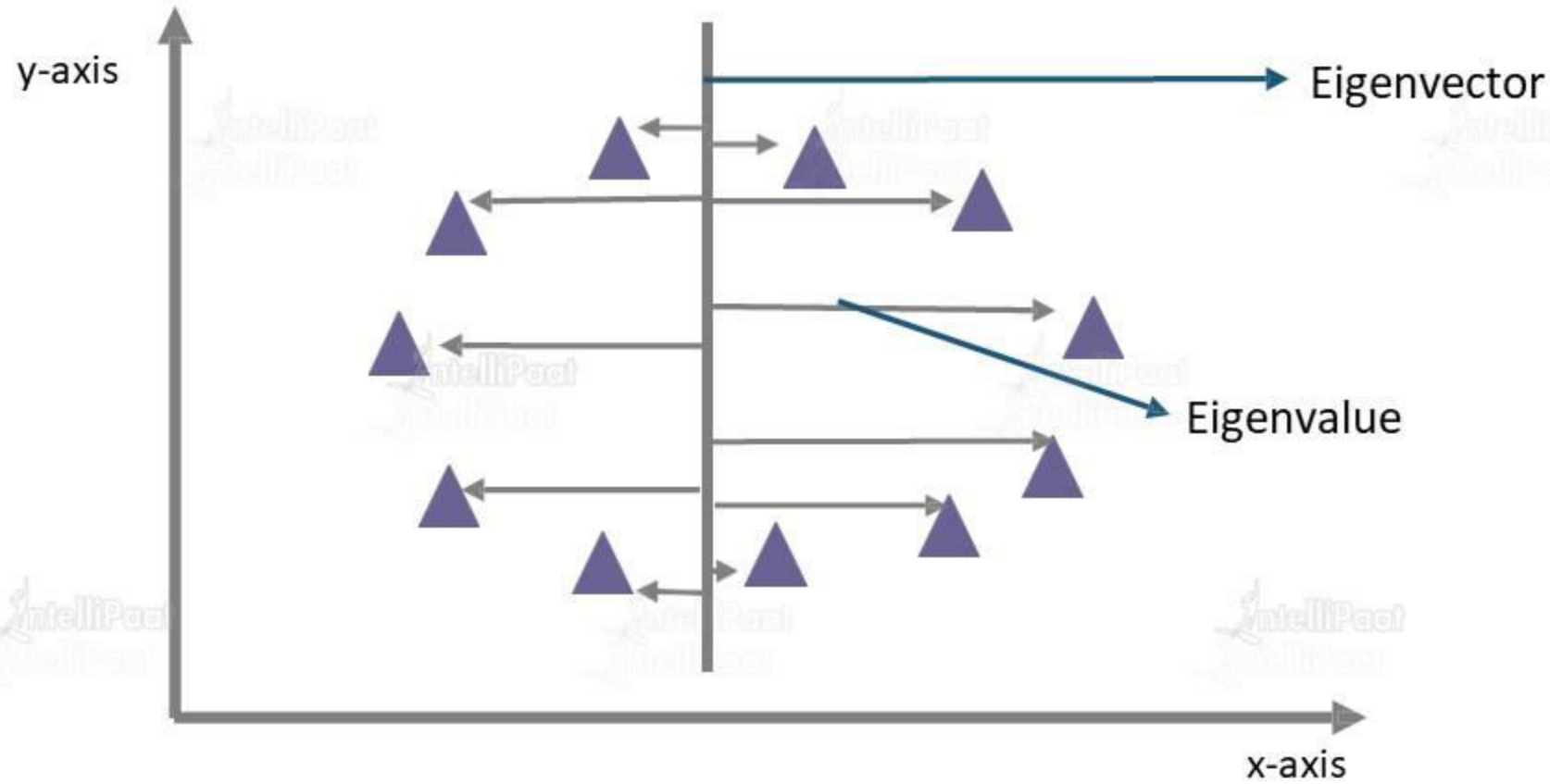
How does PCA work?

How does PCA work?

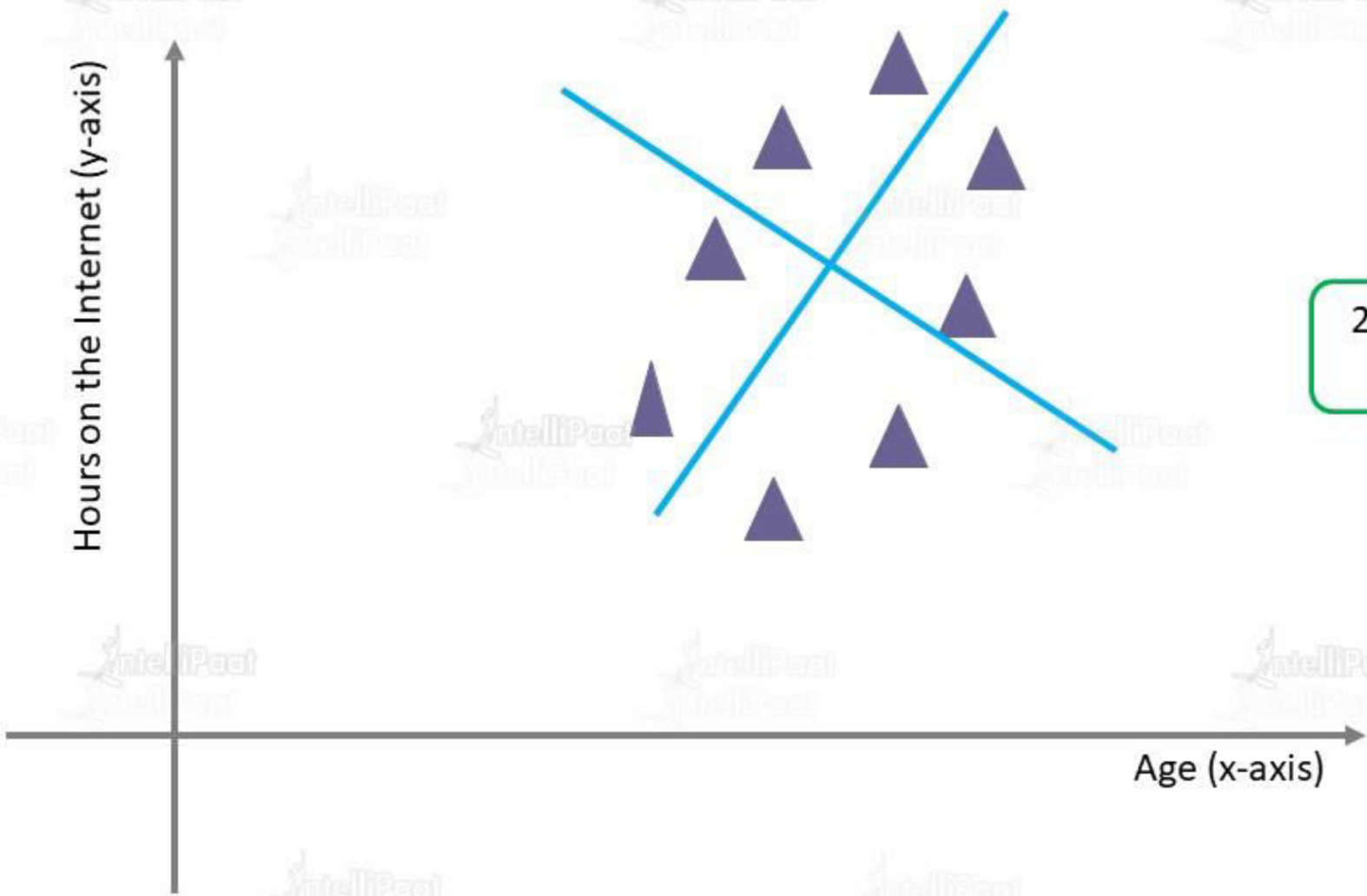


How does PCA work?

How to find out the principal component?
Eigenvectors and Eigenvalues

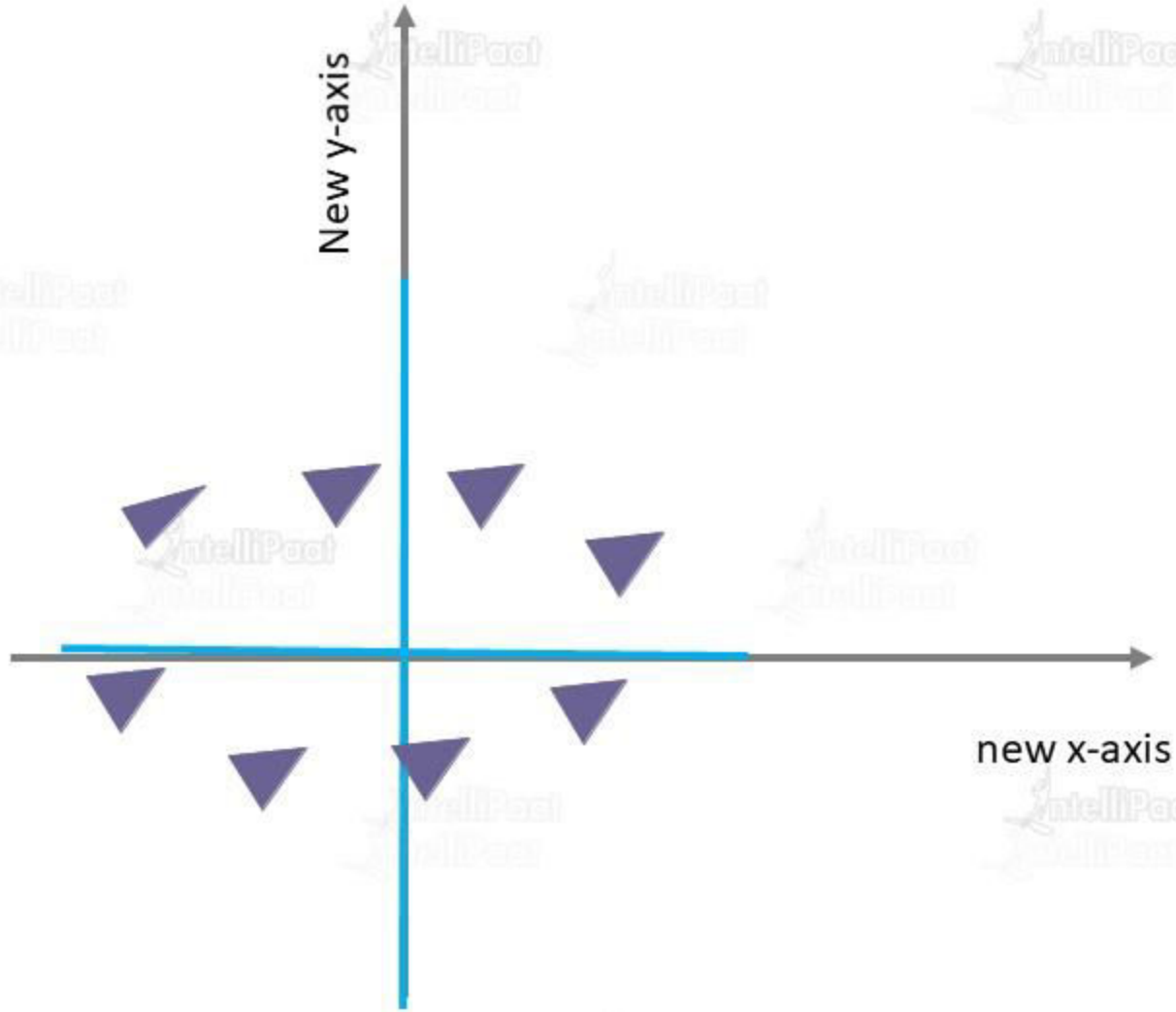


How does PCA work?



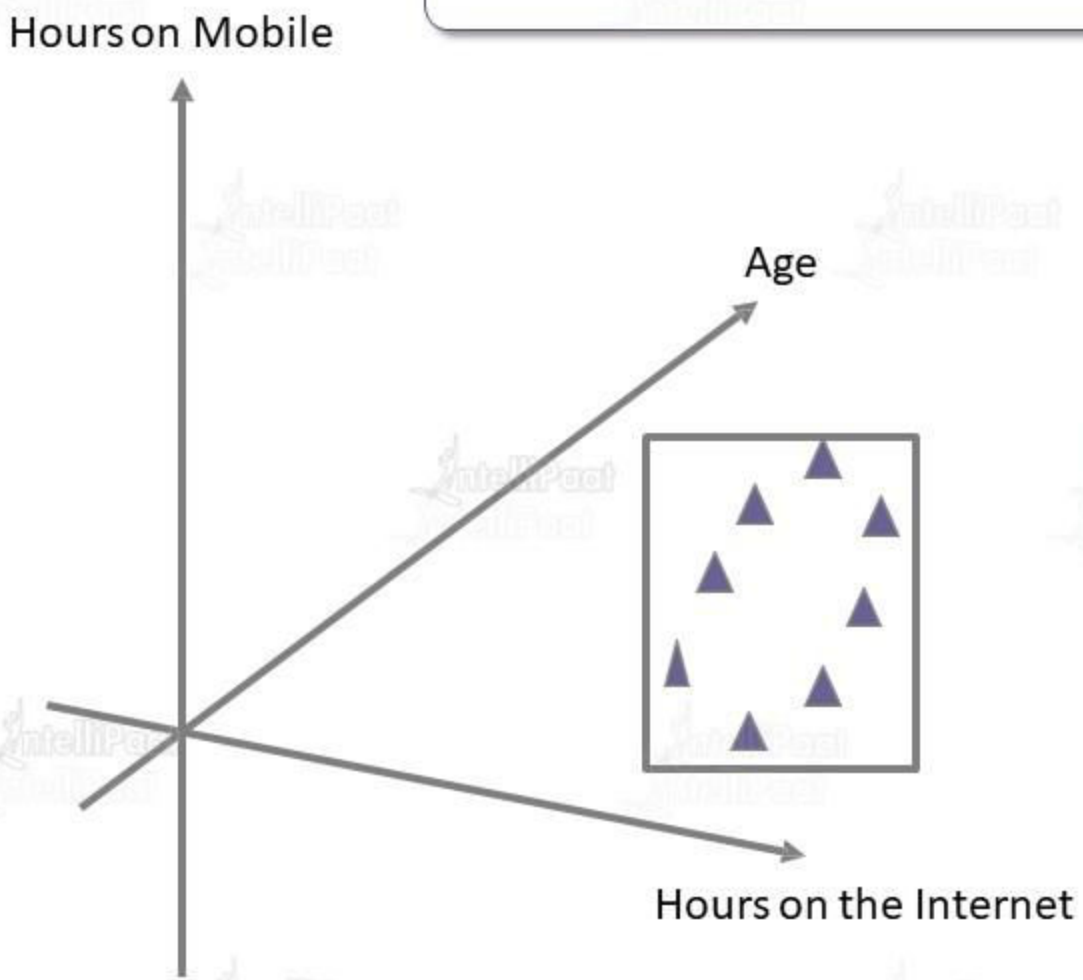
2 dimension = 2 eigenvectors and 2 eigenvalues

How does PCA work?



How does PCA work?

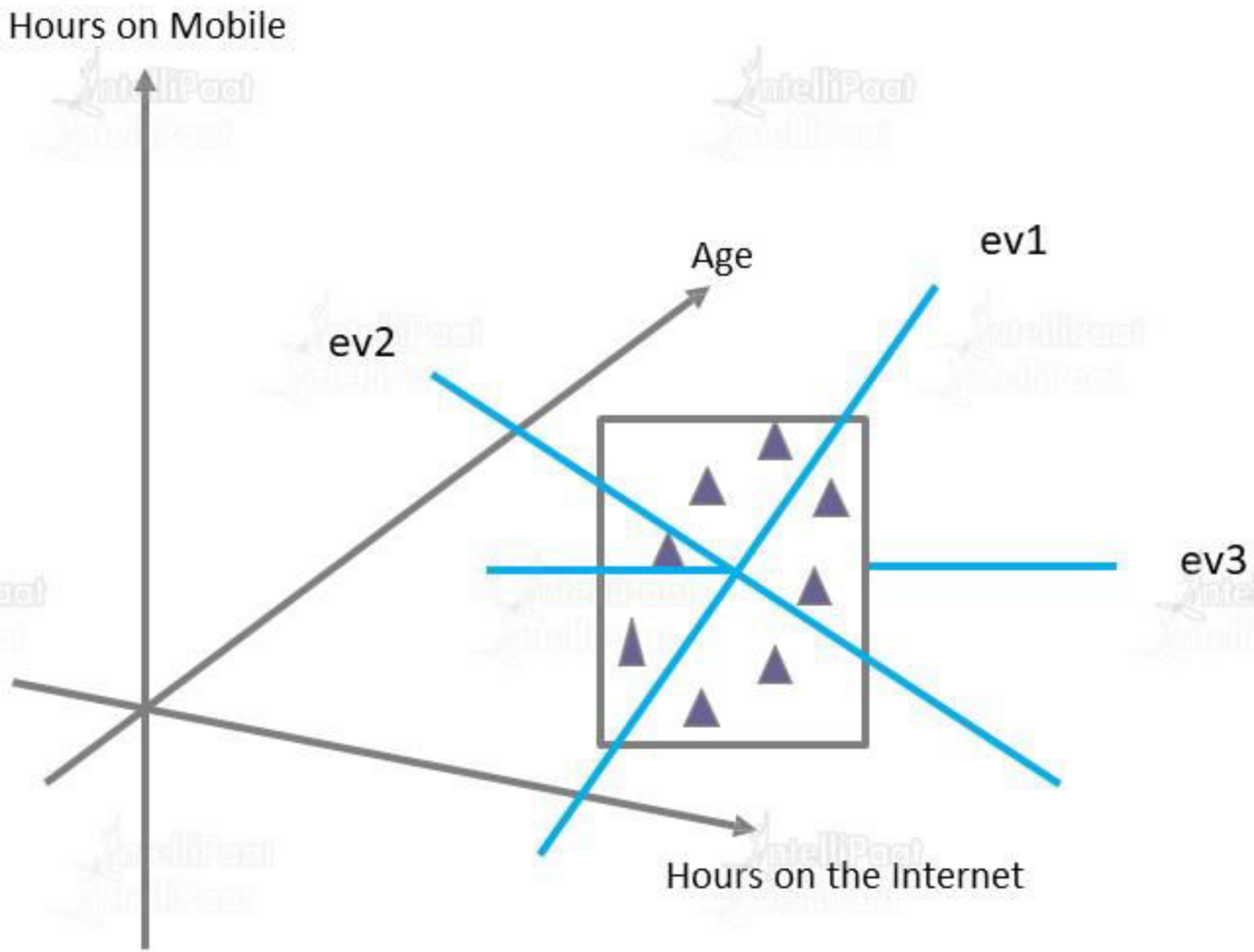
How does PCA reduce the dimensions?



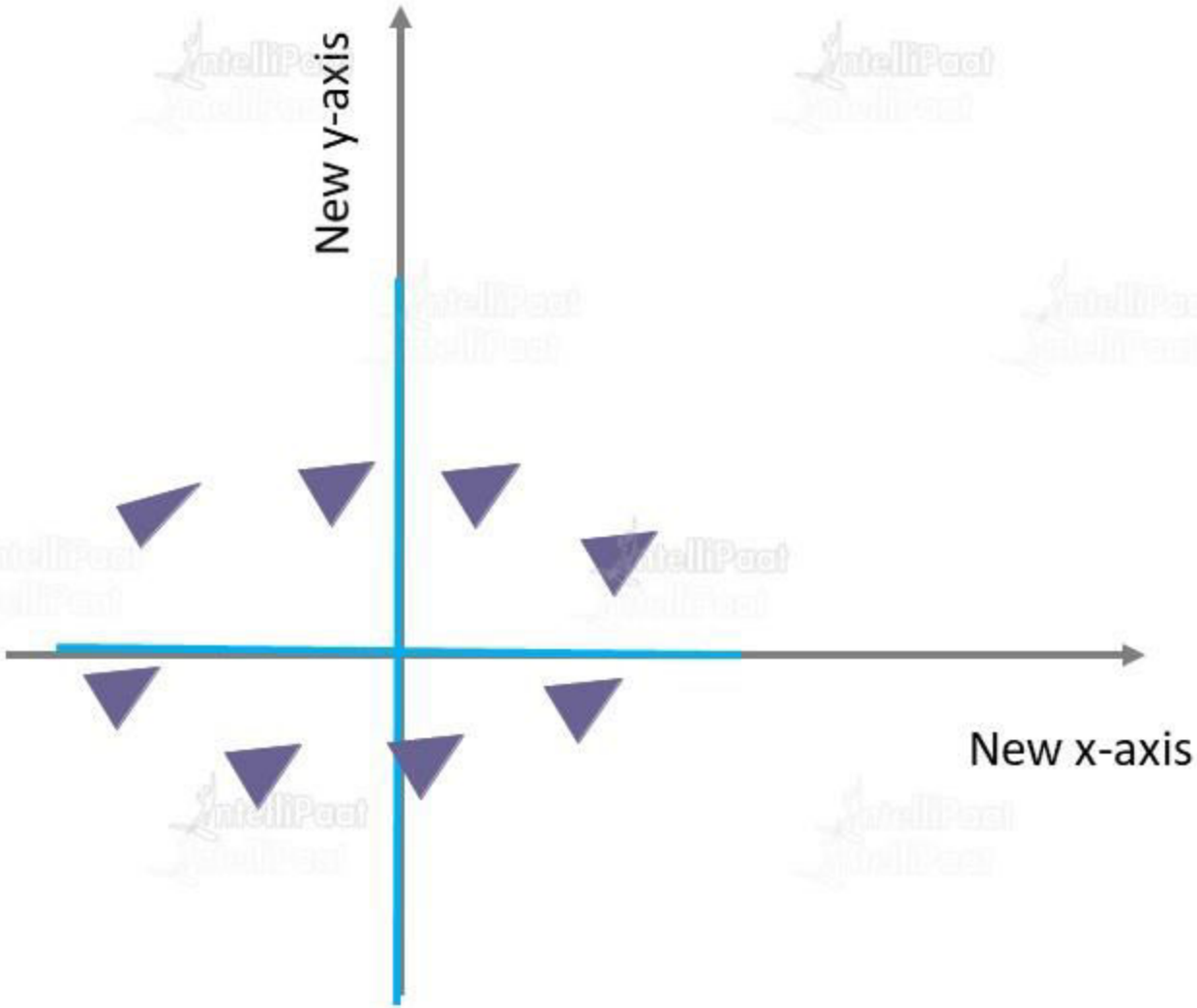
PCA reduces data by removing unnecessary dimensions

How does PCA work?

3 dimensions = 3 eigenvectors and 3 eigenvalues



How does PCA work?



3D is reduced to a 2D problem

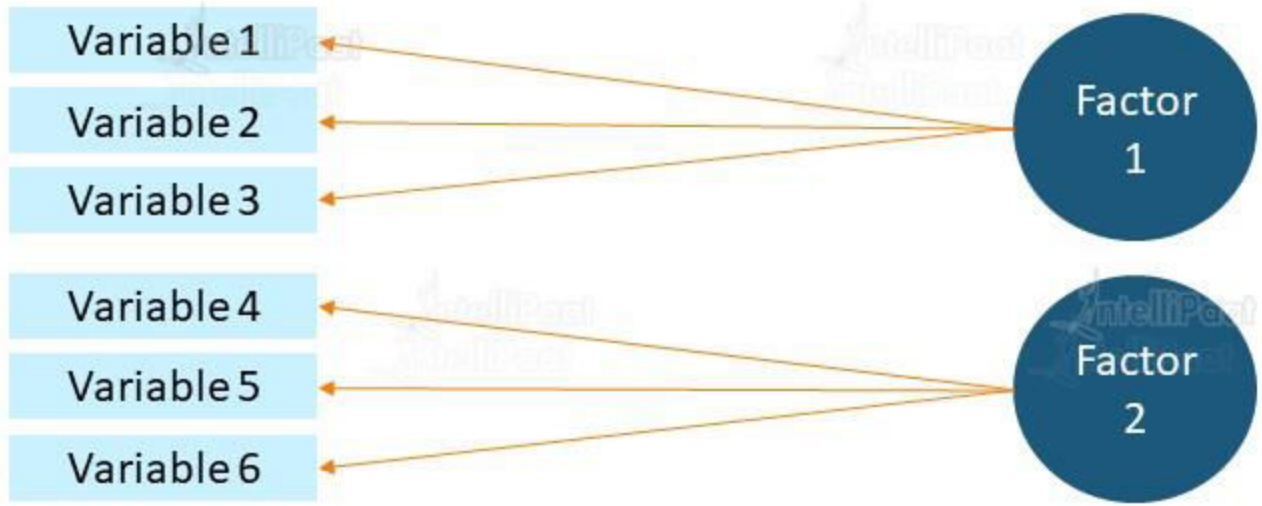
PCA: Hands On

Factor Analysis

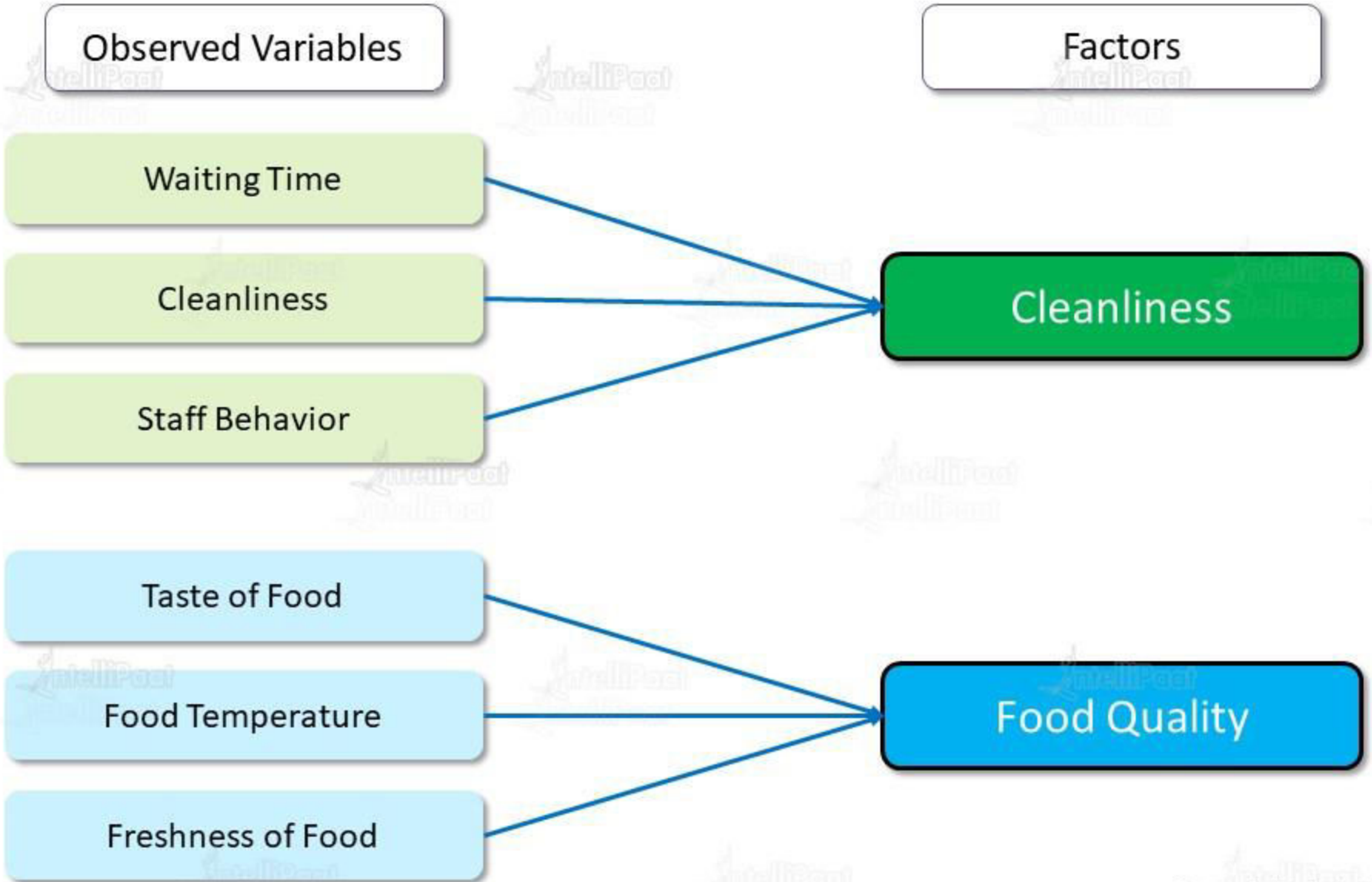
Factor Analysis

It is a data analysis method we can use to search for significant underlying trends or factors from a set of observed variables

It is widely used in market research, finance, and advertising. PCA and CFA are types of Factor Analysis



Factor Analysis



How does Factor Analysis work?

How does Factor Analysis work?

The purpose of Factor Analysis is to reduce the number of observed variables and find unobservable variables

Step 1

Feature Extraction

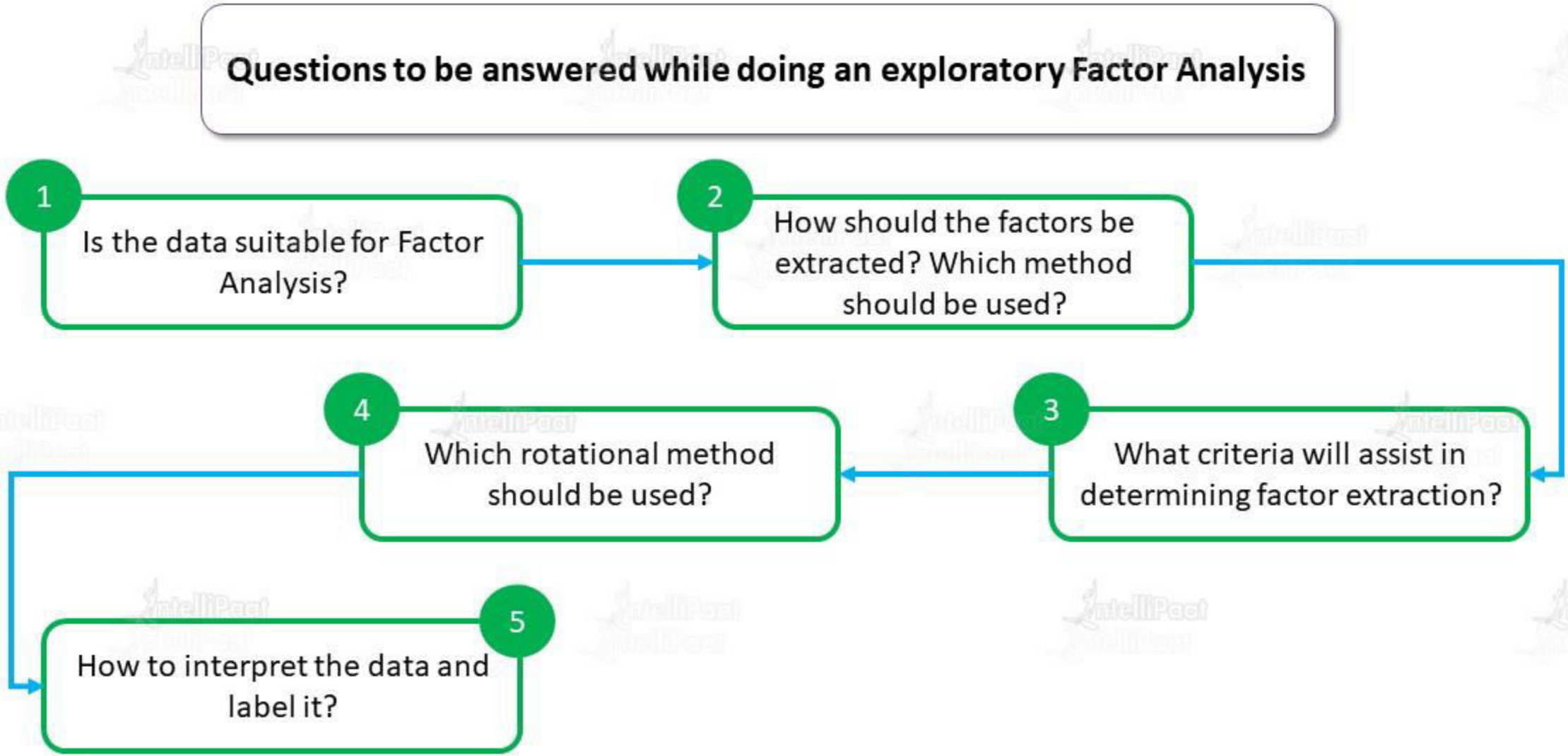
Approach for extraction selected using variance partitioning methods such as PCA to calculate no of factors

Step 2

Feature Rotation

Here, we convert factors into uncorrelated factors to improve the overall interpretability

How does Factor Analysis work?



Factor Analysis vs PCA

Factor Analysis vs PCA

1. PCA components explain the maximum amount of variance, while FA explains the covariance in data
2. PCA components are fully orthogonal to each other, whereas FA does not require factors to be orthogonal
3. PCA components is a linear combination of the observed variables, while, in FA, the observed variables are linear combinations of the unobserved variable or factor
4. PCA components are uninterpretable. In FA, underlying factors are labelable and interpretable
5. PCA is a kind of dimensionality reduction method, whereas FA is the latent variable method
6. Although PCA is a type of Factor Analysis, PCA is observational, whereas FA is a modeling technique

Factor Analysis: Hands On

What is LDA?

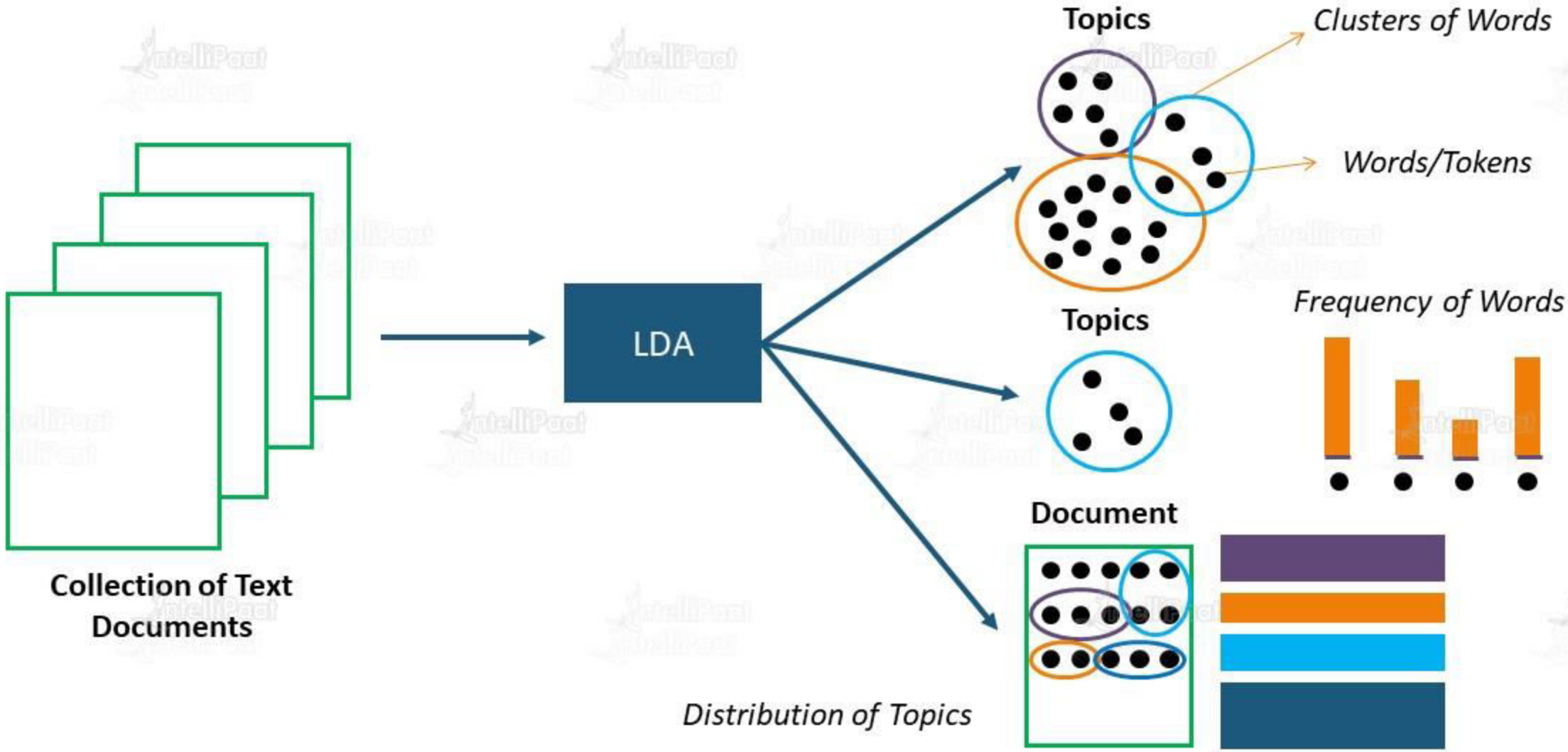
What is LDA?

In NLP, the Latent Dirichlet Allocation (LDA) is a generative model that helps in explaining the sets of observations by unobserved groups showing why some parts of the data are similar. It also helps heavily in topic modeling

What it topic modeling?

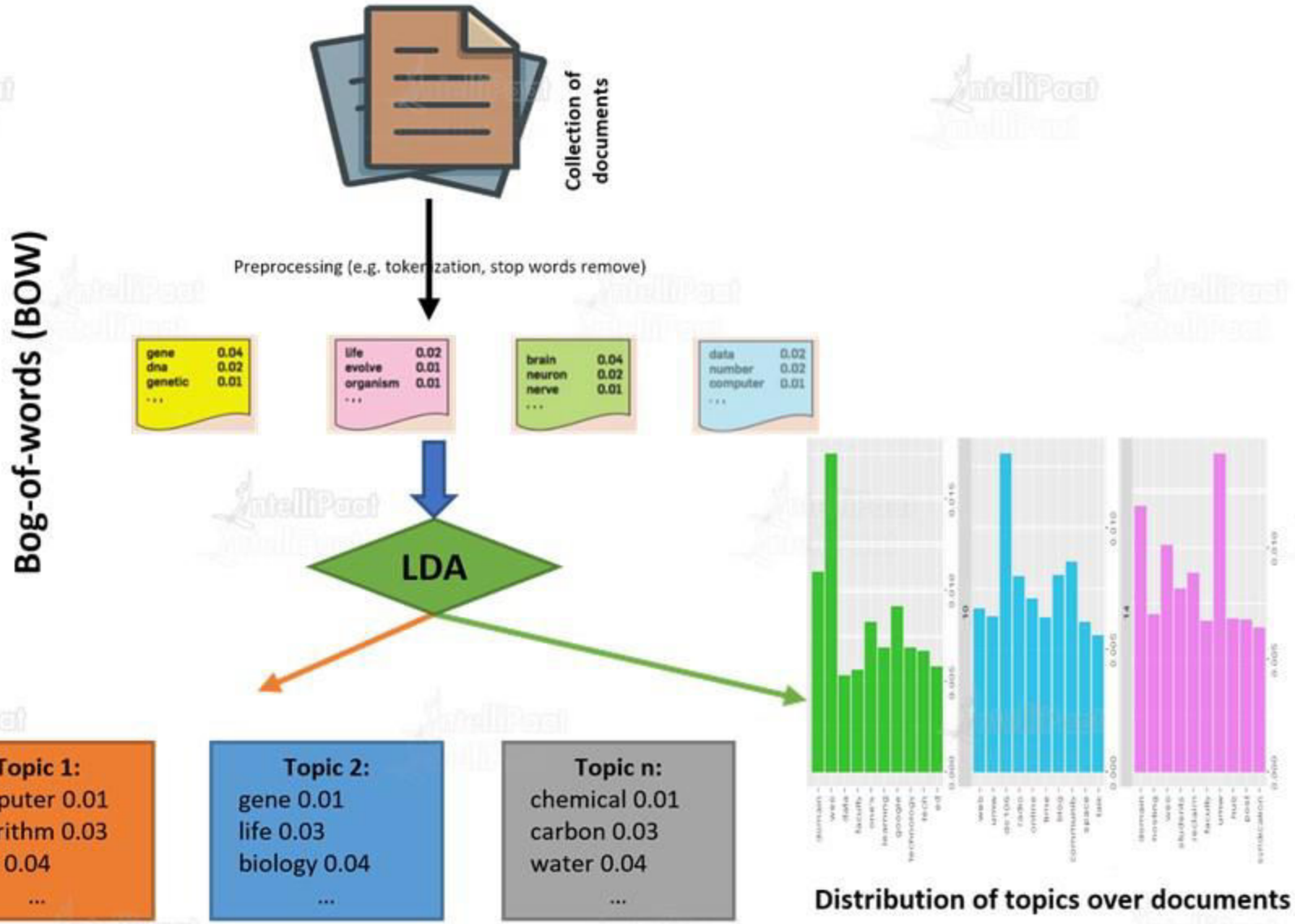
This is a part of unsupervised NLP, where we can represent a text file using several topics that can explain the underlying information of that document

What is LDA?



How does LDA work?

How does LDA work?



LDA: Hands On

India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

support@intellipaat.com

24/7 Chat with Our Course Advisor