ECSQL

Group 8

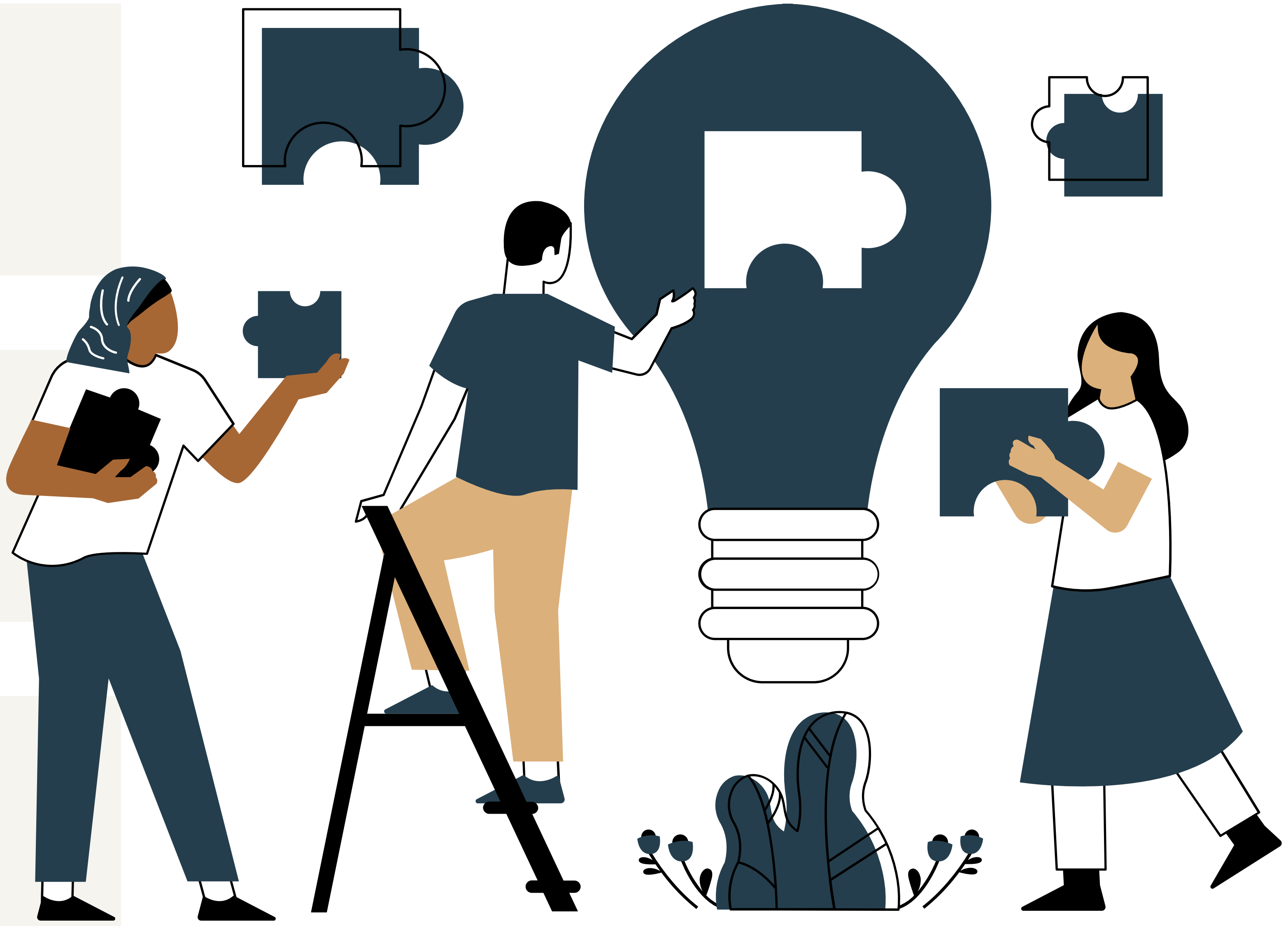
R10945002 林柏詠
R11945005 郭庭沂
R11945044 張瑜倢
R11945018 曾于瑄

Outline

1 Motivation

2 Workflow

3 Design Architecture

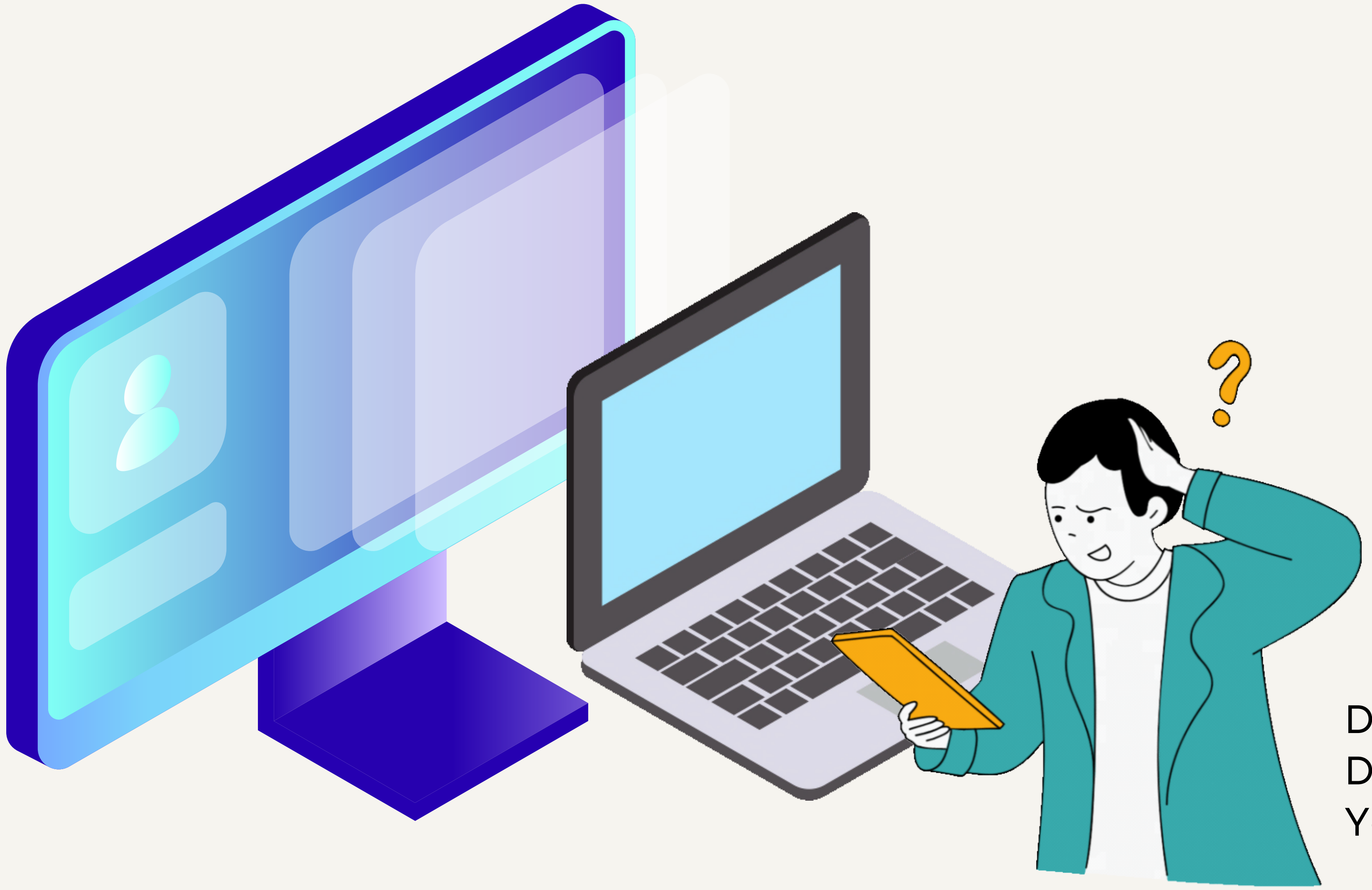


Motivation

Slow Excel Opening
due to Large Data Volume

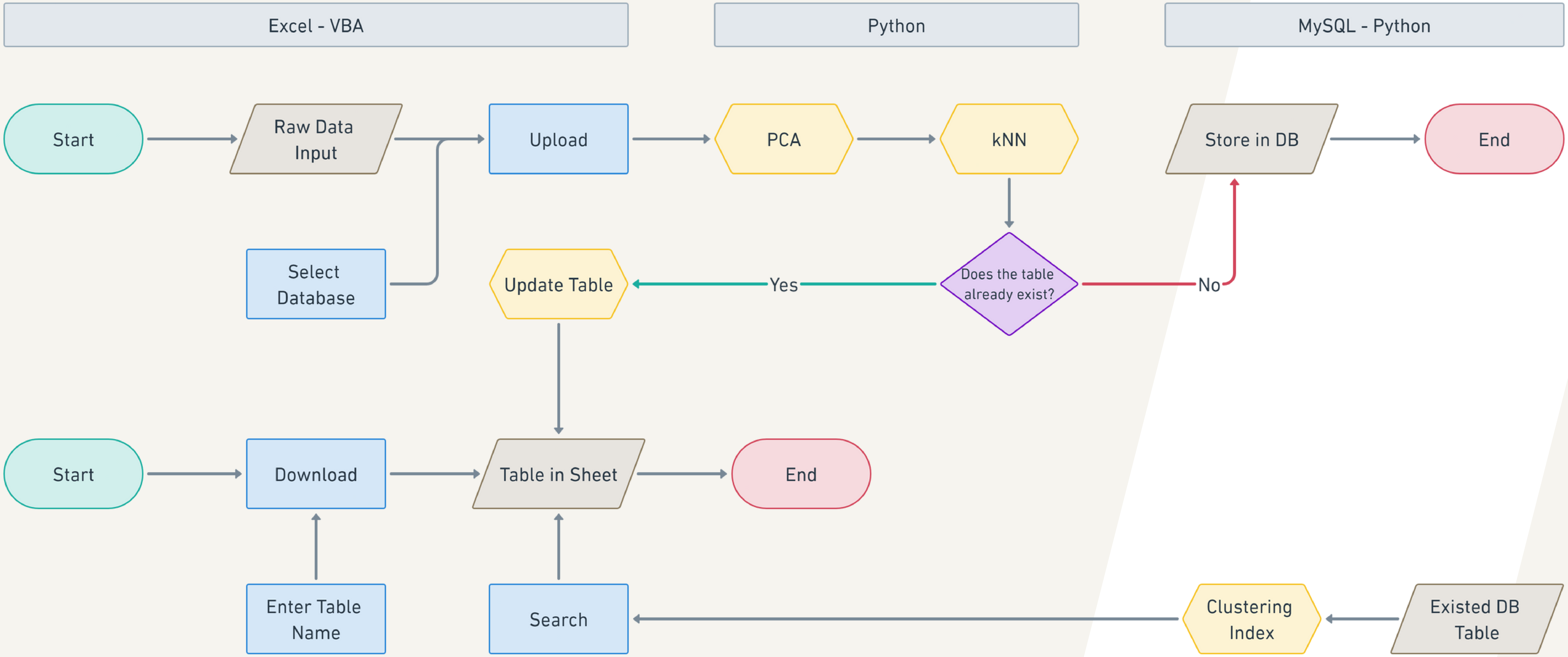
Database Access
Difficulty for Non-programmers

Search Time
Impact from Massive Data Volume

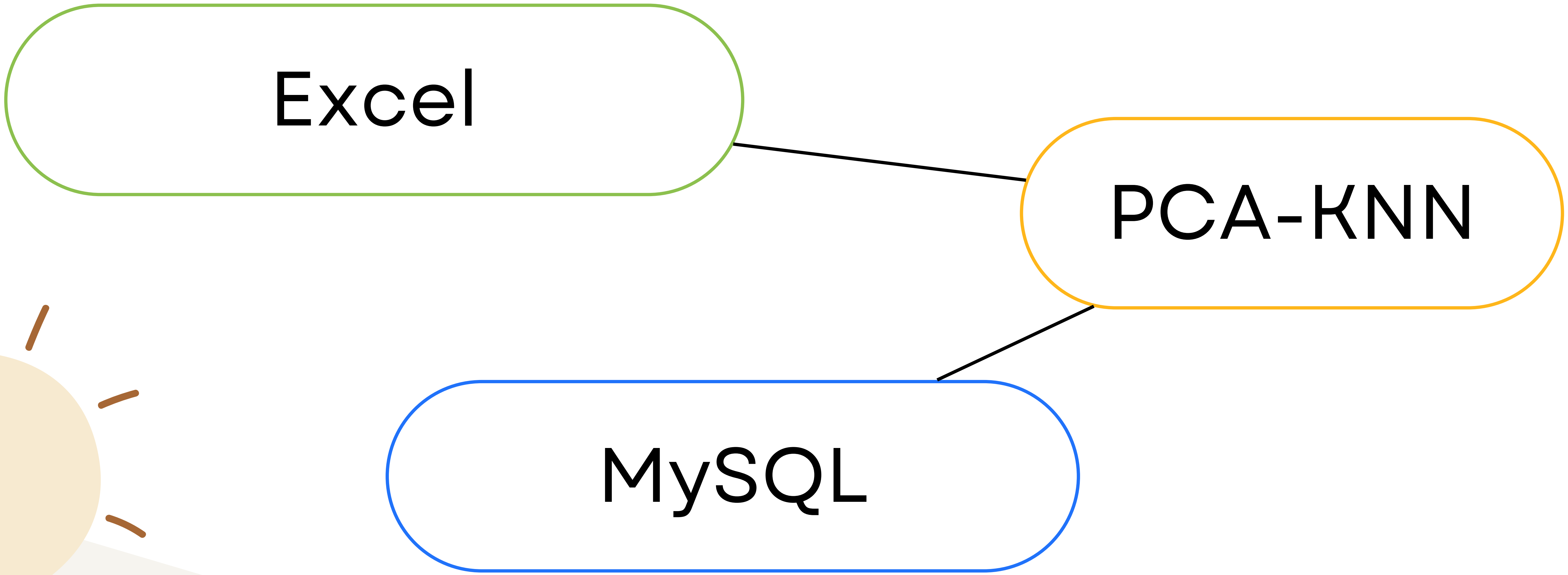


Doctors,
DB Beginners,
You and me?

Workflow


Excel - VBA
Python
MySQL - Python
Start
Raw Data Input
Upload
PCA
kNN
Store in DB
End
Select Database
Update Table
Yes
Does the table already exist?
No
Start
Download
Table in Sheet
End
Enter Table Name
Search
Clustering Index
Existed DB Table

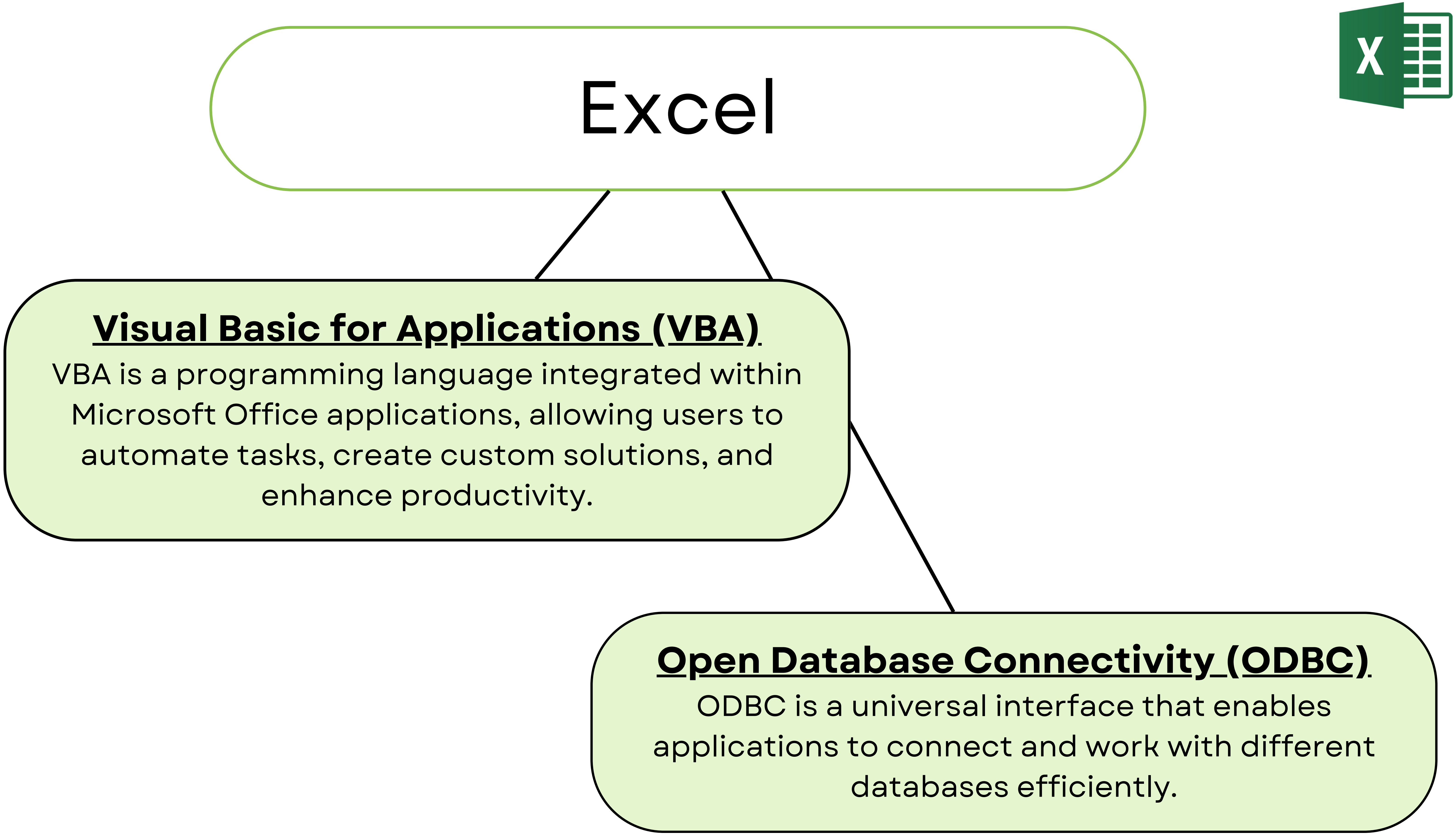
Design Architecture



Excel

PCA-KNN

MySQL



Excel

Visual Basic for Applications (VBA)

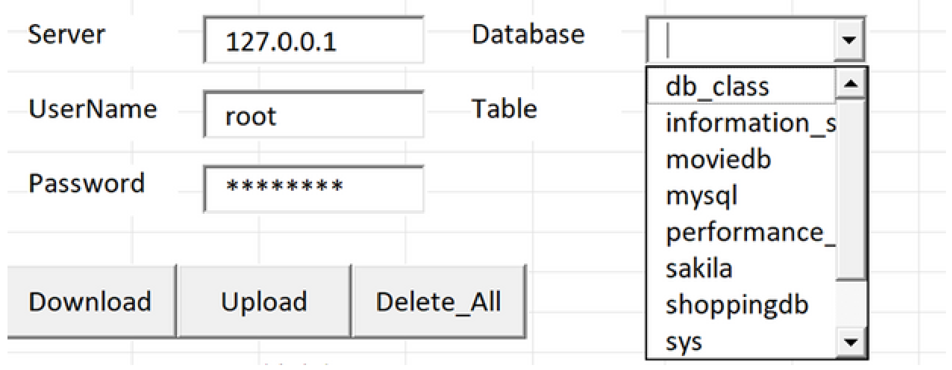
VBA is a programming language integrated within Microsoft Office applications, allowing users to automate tasks, create custom solutions, and enhance productivity.

Open Database Connectivity (ODBC)

ODBC is a universal interface that enables applications to connect and work with different databases efficiently.

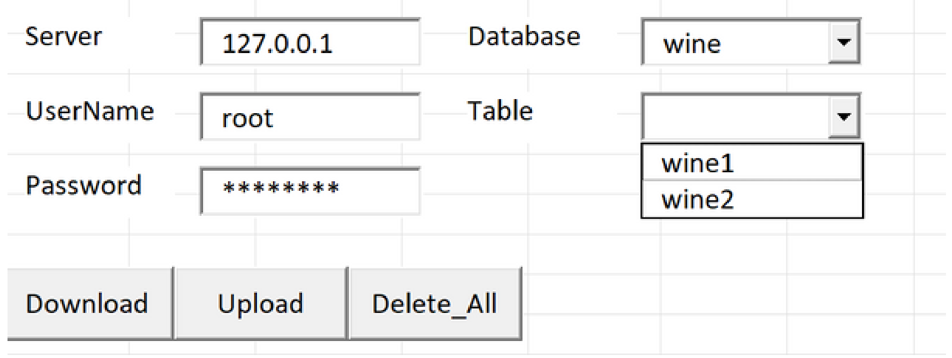
Excel

Connect to Database



Server	127.0.0.1	Database	
UserName	root	Table	
Password	********		

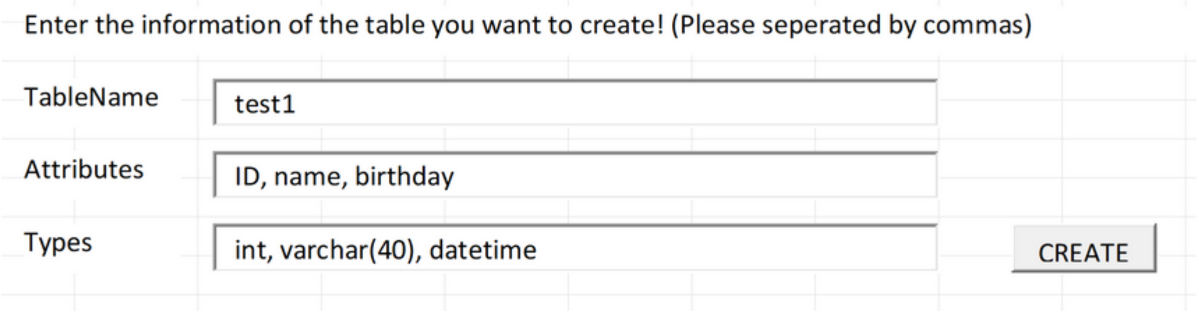
db_class
information_s
moviedb
mysql
performance_
sakila
shoppingdb
sys



Server	127.0.0.1	Database	wine
UserName	root	Table	
Password	********		

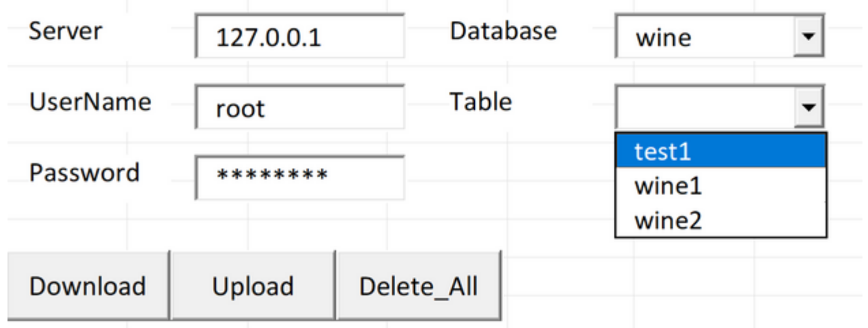
wine1
wine2

Create Table /
Upload and Download Data



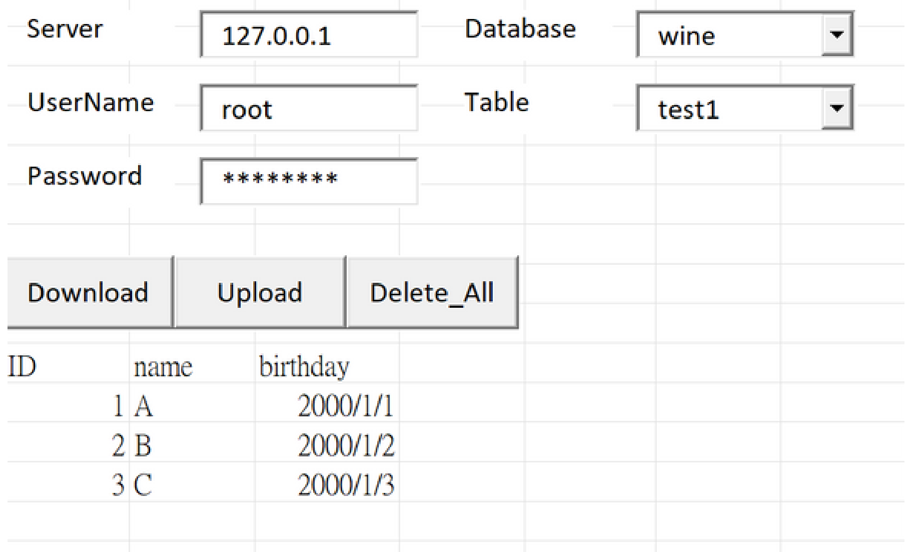
Enter the information of the table you want to create! (Please seperated by commas)

TableName	test1
Attributes	ID, name, birthday
Types	int, varchar(40), datetime



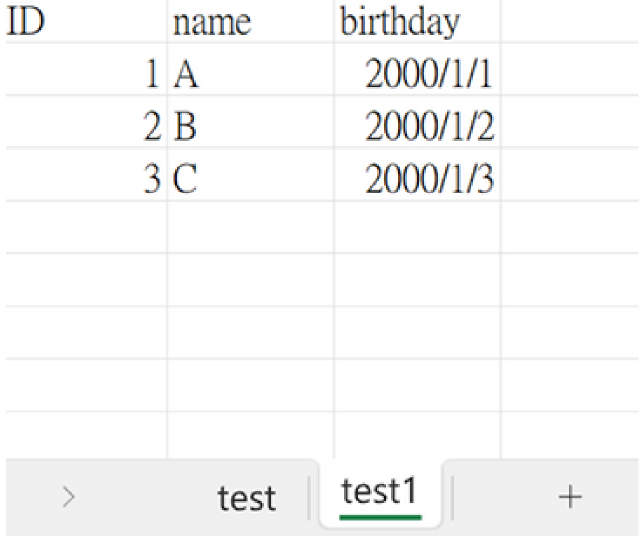
Server	127.0.0.1	Database	wine
UserName	root	Table	
Password	********		

test1
wine1
wine2

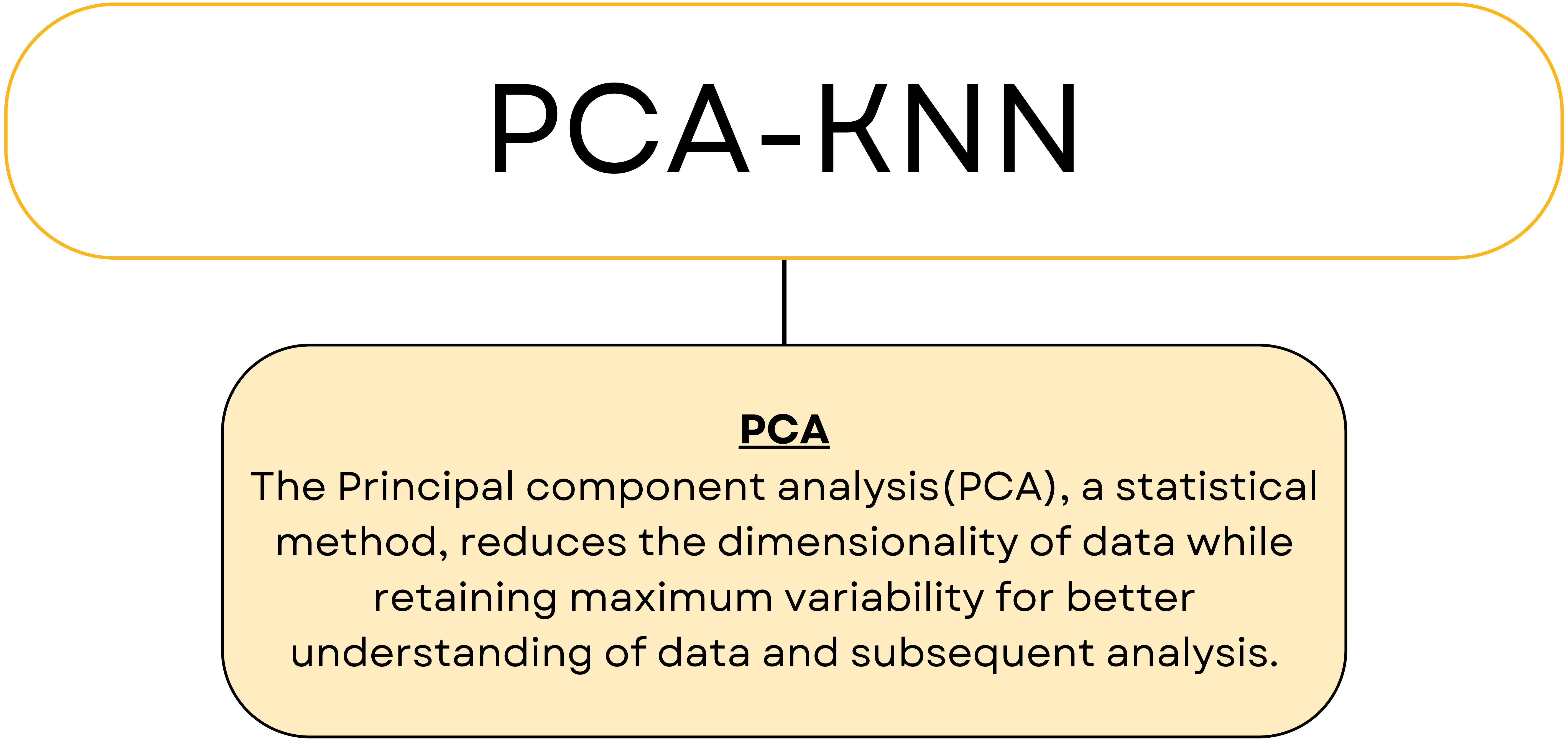


Server	127.0.0.1	Database	wine
UserName	root	Table	test1
Password	********		

ID	name	birthday
1	A	2000/1/1
2	B	2000/1/2
3	C	2000/1/3



ID	name	birthday
1	A	2000/1/1
2	B	2000/1/2
3	C	2000/1/3



PCA-KNN

PCA

The Principal component analysis(PCA), a statistical method, reduces the dimensionality of data while retaining maximum variability for better understanding of data and subsequent analysis.

PCA

High Dimensional Data Challenges	
01	Feature correlation
02	Computing costs
03	Overfitting

01 Multiple solution problems or even redundancy

02 Increased memory requirements and reduced operational efficiency

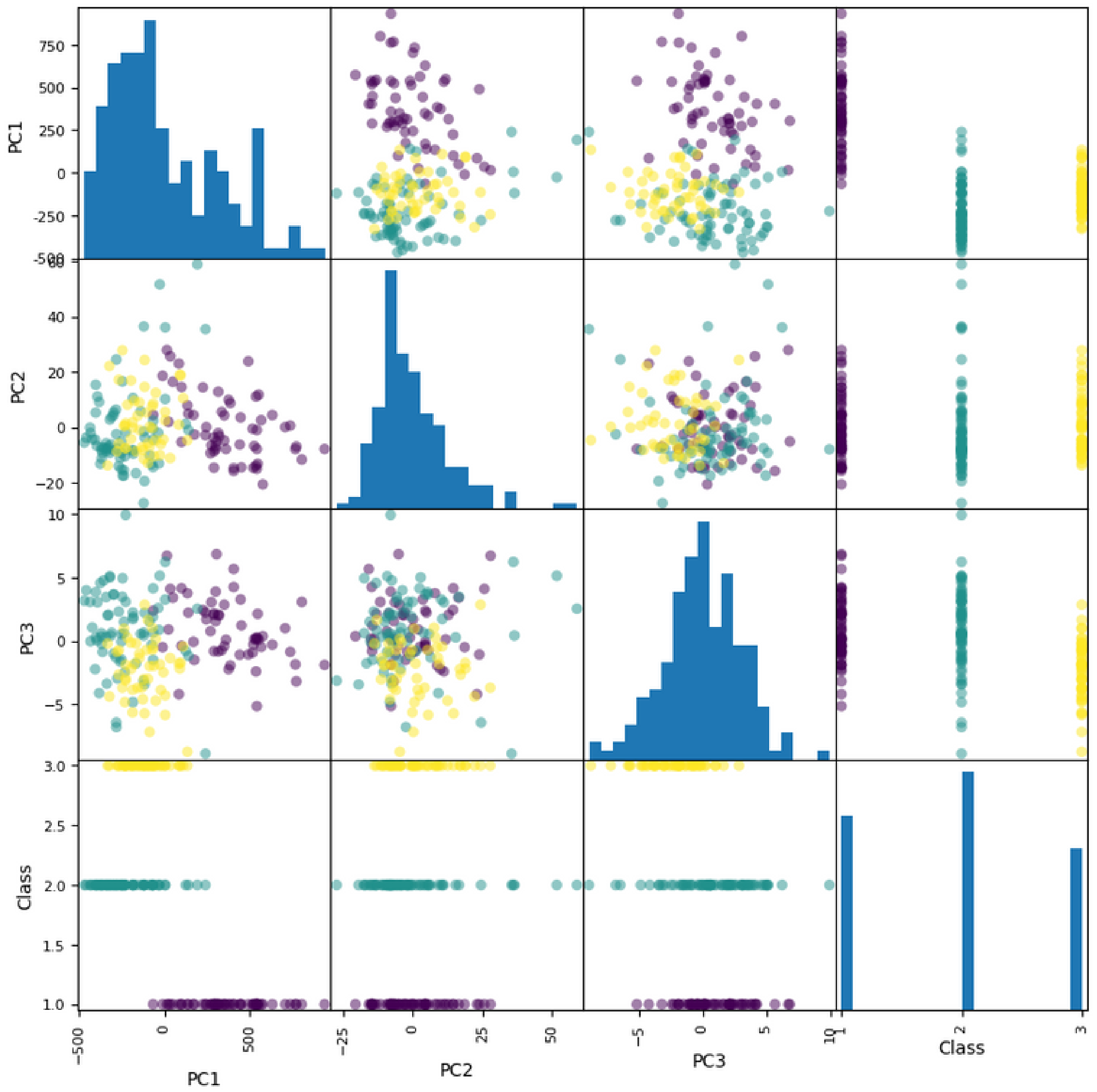
03 Captures too much noise and detail to generalise to new data.

PCA

From the original 13 dimensions (excluding Labels), we retained **3 principal components** and measured the contribution of each principal component to the variability of the original data by the **explained variance ratio**.

In addition, the **Scatter Matrix** allows us to observe correlations between variables, scattering patterns and possible trends.

Principal Components	**PC1**	**PC2**	**PC3**
Explained Variance Ratio	99.81	0.17	0.01



The scatter matrix of PCA with labels

PCA-KNN

KNN

The k-nearest neighbour (KNN), a supervised algorithm, predicts the classification of unlabeled data by taking into account the features and labels of the training data.

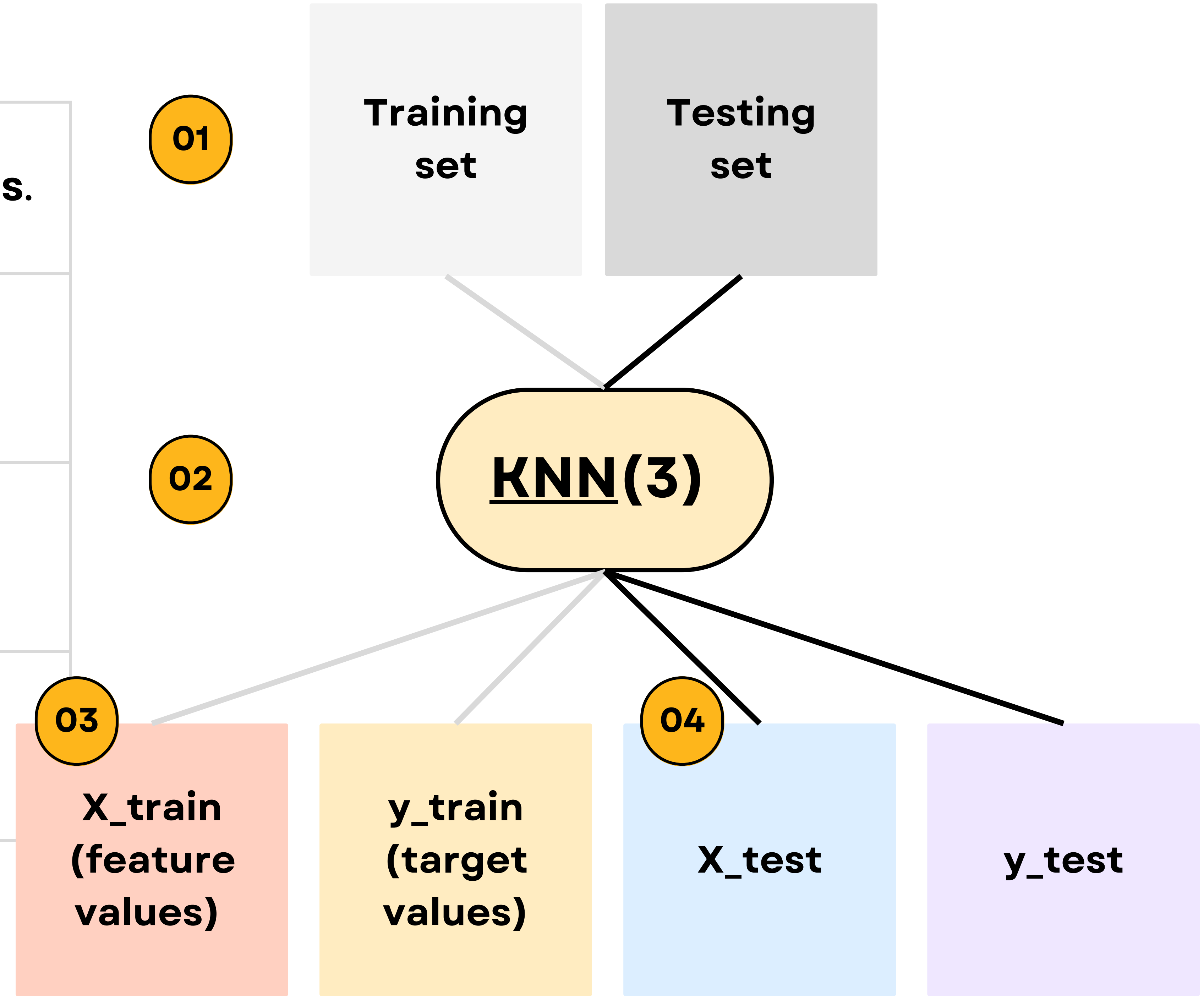
KNN

01 Split the dataset into **training** and **testing sets.**

02 Creat a **k-nearest neighbors (KNN) model.** (N_neighbors parameter is set to 3)

03 Use the training data by passing **X_train** and **y_train.**

04 Predict the classes for the testing data by passing **X_test.**



01

Training set

Testing set

02

KNN(3)

03

X_train (feature values)

y_train (target values)

04

X_test

y_test

KNN

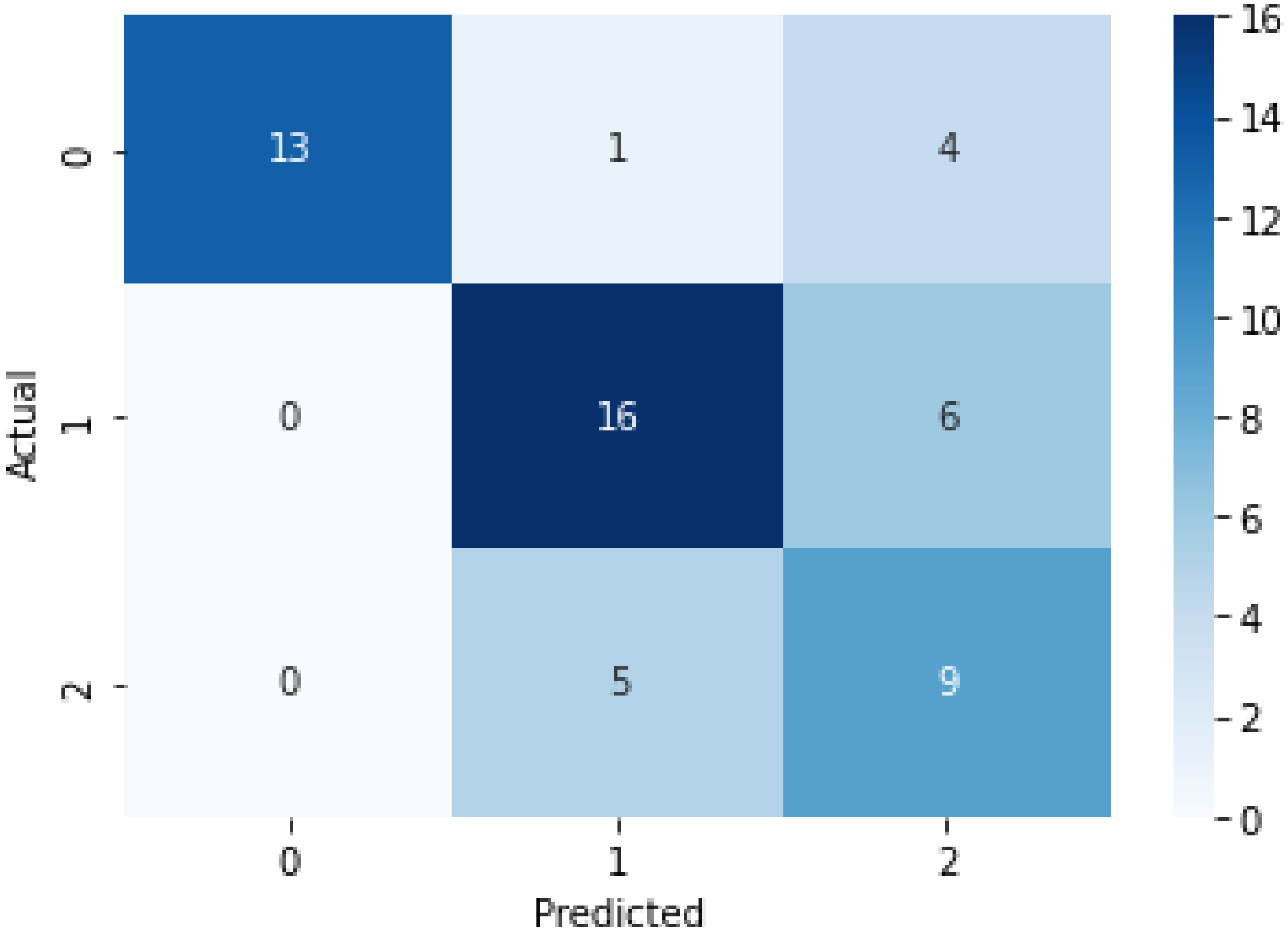
05

Evaluate the accuracy of the KNN model by **comparing** its **predictions with the true labels** of the **testing data**.

Print the accuracy score on the **training data**.

06

The matrix represents the performance of the KNN model in terms of **classifying the samples into their respective classes.**



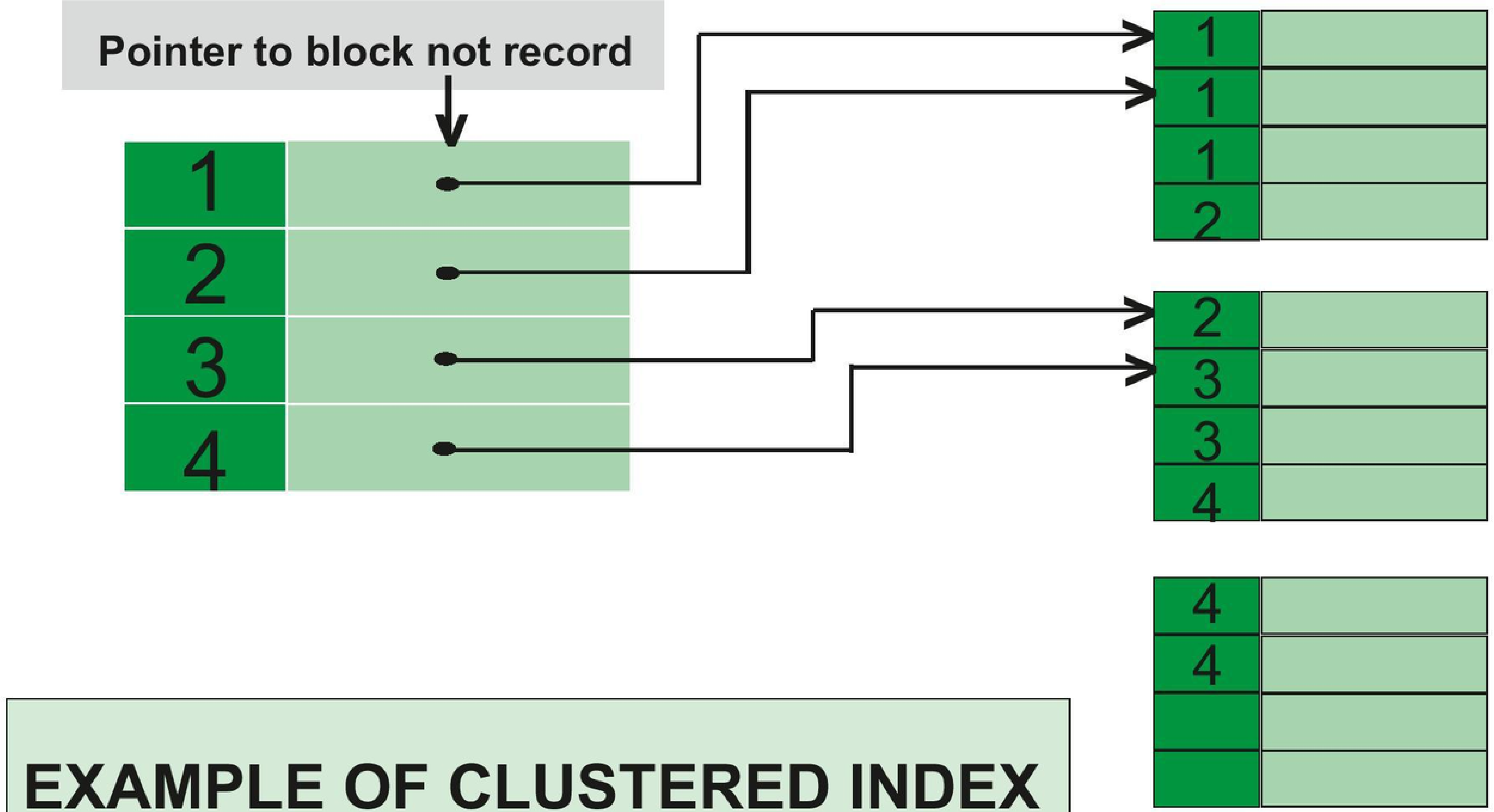
The accuracy of the KNN model's prediction

Clustering index

Create table

abcdefg

Index



Pointer to block not record

EXAMPLE OF CLUSTERED INDEX

Future work

01 **More focus on efficacy of searching step**

02 **Clustering implementation**

03 **Overall code integration**

Video Download Link