AN INITIATIVE OF:

science, technology & innovation
Department: Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

CSIR Touching lives through innovation | 80th anniversary

nicis
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
DIRISA

An Introduction to Unsupervised Machine Learning

By Khanyisa Mokgolobotho & Simanga Mchunu

25 June 2025

01

INTRODUCTION

"The goal is to turn data into information, and information into insight." - Carly Fiorina

What Is Unsupervised Learning?

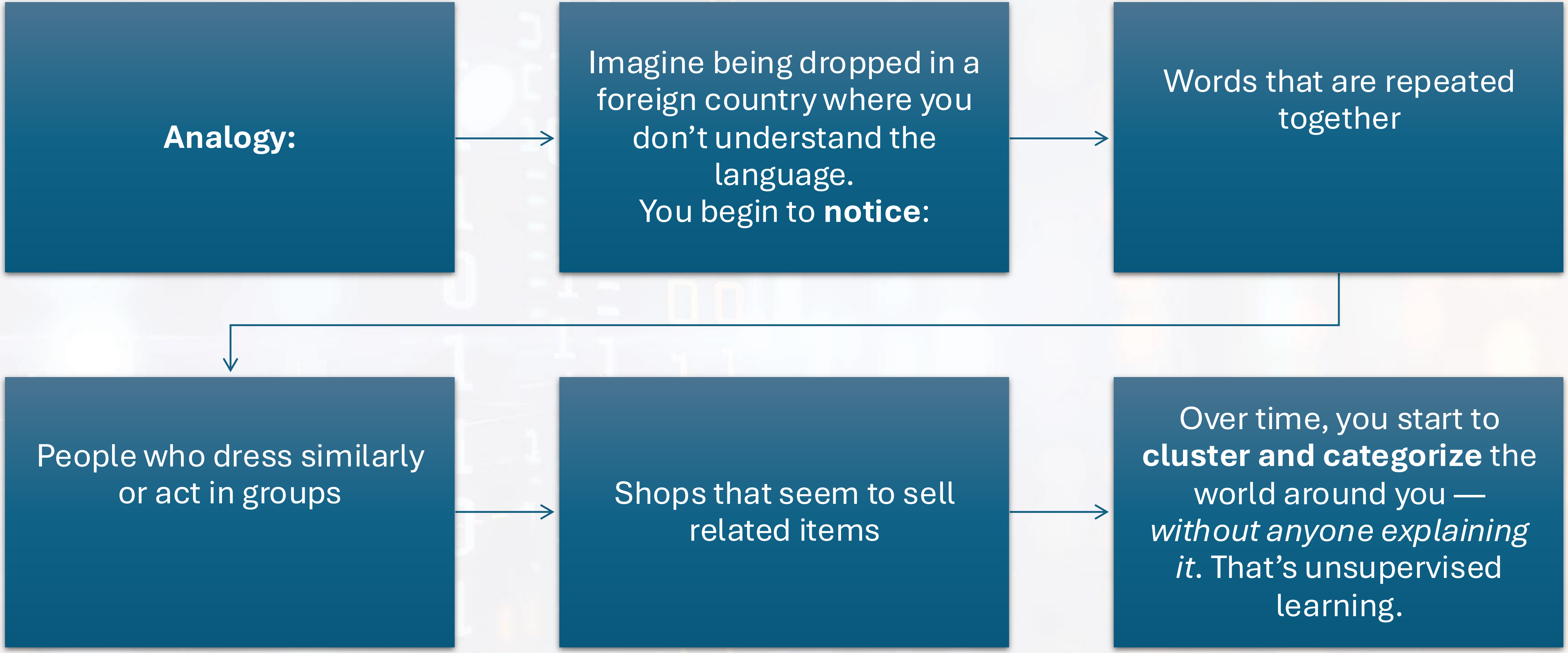
Definition:
Unsupervised learning is a type of machine learning where the algorithm is given **data without labels** and must find structure or patterns on its own.

Key Concept:
The system doesn’t know the "correct answer"—it learns by exploring **relationships**, **similarities**, and **deviations** in the data.

One of the key advantages of unsupervised learning is its ability to uncover previously unknown patterns and relationships. Without the constraints of labeled data, unsupervised algorithms can reveal valuable insights that may not be apparent through other analytical methods. This makes unsupervised learning particularly useful in exploratory data analysis, anomaly detection, and clustering.

“birds of a feather flock together”

What Is Unsupervised Learning?



Analogy:

Imagine being dropped in a foreign country where you don’t understand the language.
You begin to **notice**:

Words that are repeated together

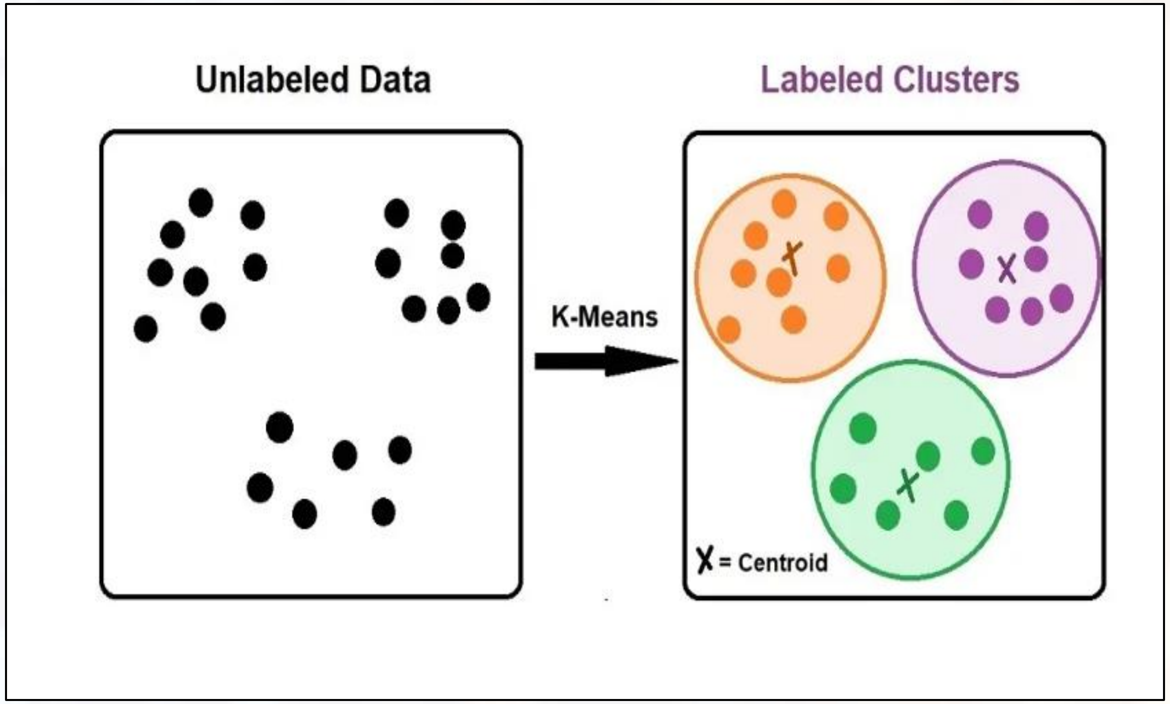
People who dress similarly or act in groups

Shops that seem to sell related items

Over time, you start to **cluster and categorize** the world around you — *without anyone explaining it*. That’s unsupervised learning.

Why should we care?

- **80–90% of real-world data is unlabeled.**
- That means unsupervised learning is your go-to when:
  - You don’t have time/resources to label your dataset
  - You want to explore new patterns before jumping into supervised tasks
  - You need insights that aren’t visible to the naked eye



Popular Unsupervised Learning Techniques

- We'll explore four key techniques:
  - **Clustering**: Grouping data points with similar characteristics
  - **Dimensionality reduction**: Transforming high-dimensional data into lower dimensions for easier analysis
  - **Anomaly detection**: Identifying data points that deviate significantly from the norm
  - **Association rule learning**: Identifying interesting relations between variables in large datasets

Real-World Applications in 2025

01

Healthcare : Cluster patients with similar genetic markers for personalized treatment

02

Finance: Detect rare fraud patterns in credit card transactions

03

Retail: Segment customers for hyper-personalized offers

04

Cybersecurity: Monitor network traffic and flag anomalies in real-time

05

Media Platforms: Use behavior clustering to suggest relevant videos/music

06

Education: Predict dropout risk based on student activity patterns

02

Clustering

Clustering

- **What is Clustering?**
  - Grouping similar data points
  - Uncovers hidden structures within the data
- Let’s say you’re analyzing student performance across multiple faculties. Using clustering, you might discover:
  - A group of students struggling in first-year math, across all departments
  - Another group that excels in theory-based courses but underperforms in practicals

Clustering

Real-World Use Cases:

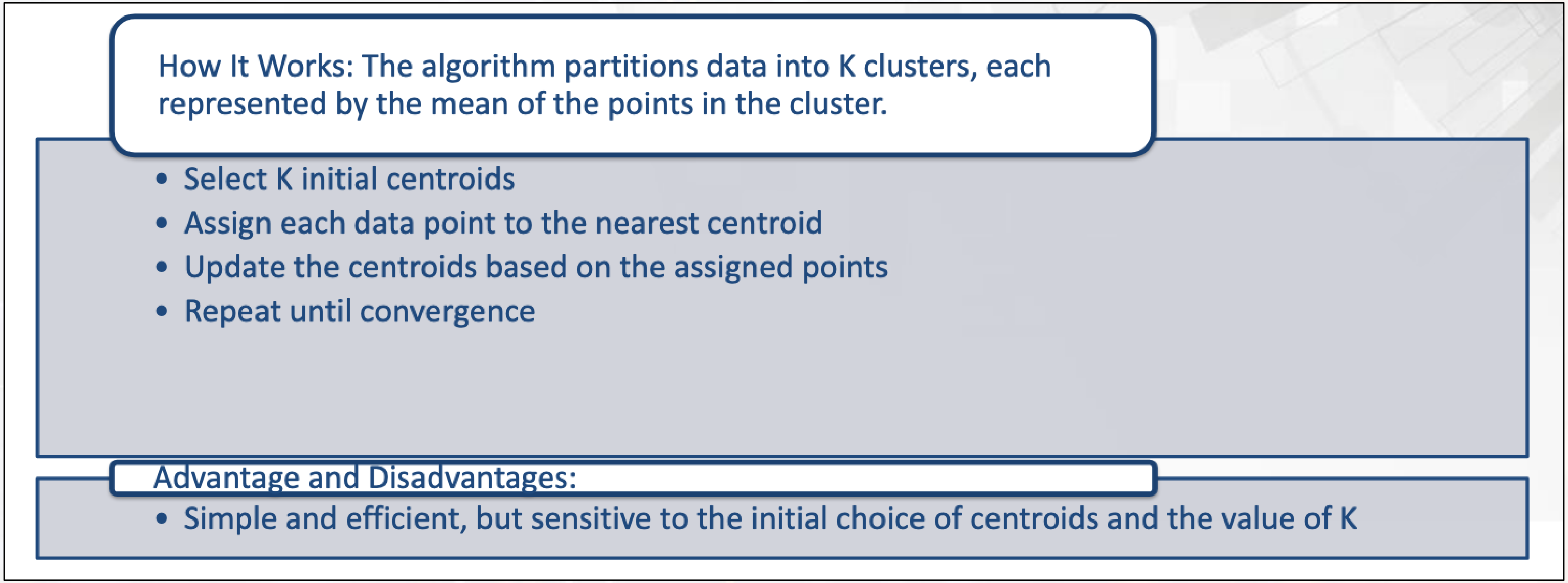
Market Segmentation – Understanding customer types in business

Patient Profiling – Identifying disease subgroups in medical data

Social Network Analysis – Finding communities or influencers

Popular Algorithms: K-Means, DBSCAN, Agglomerative Clustering, Hierarchical clustering

K-Means Clustering



How It Works: The algorithm partitions data into K clusters, each represented by the mean of the points in the cluster.

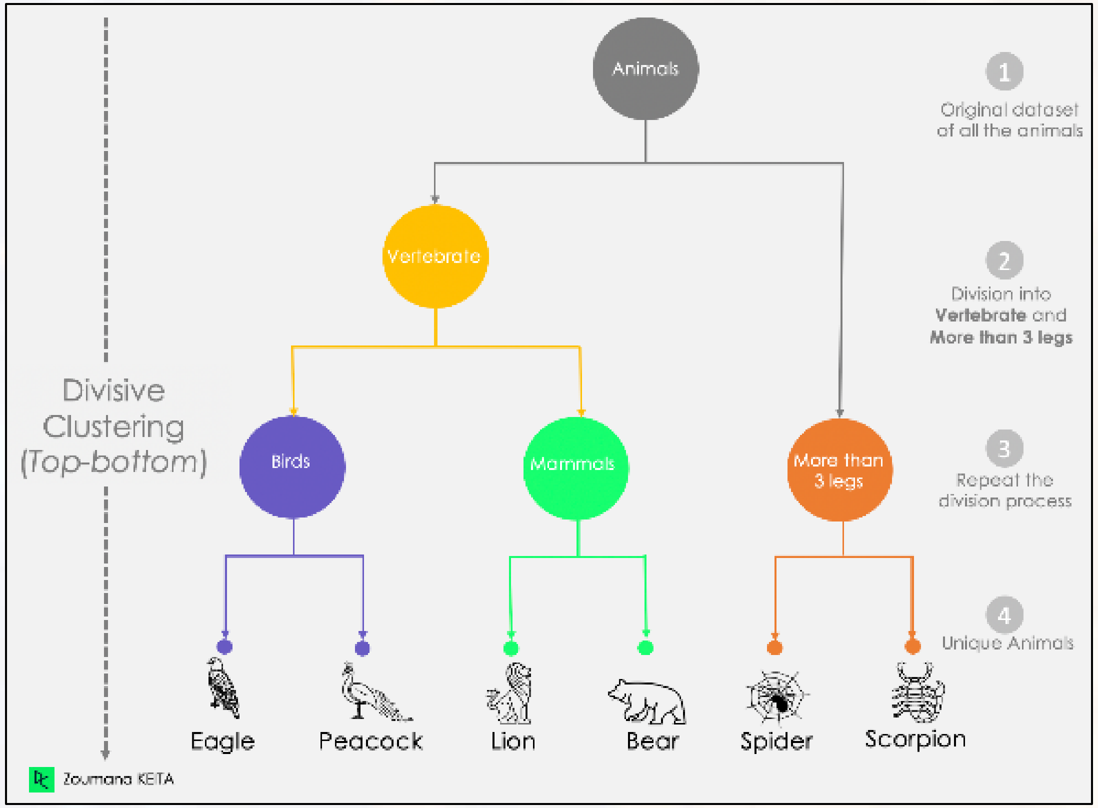
- Select K initial centroids
- Assign each data point to the nearest centroid
- Update the centroids based on the assigned points
- Repeat until convergence

Advantage and Disadvantages:

- Simple and efficient, but sensitive to the initial choice of centroids and the value of K

Hierarchical Clustering

- **Types of Hierarchical clustering:**
  - **Agglomerative**: Bottom-up approach where each data point starts in its own cluster, and clusters are merged iteratively.
  - **Divisive**: Top-down approach where all data points start in one cluster, and splits are performed iteratively.
  - **Dendrograms**: Tree-like diagrams used to visualize the hierarchical structure of clusters.


Animals
1
Original dataset of all the animals
Vertebrate
2
Division into Vertebrate and More than 3 legs
Divisive Clustering (Top-bottom)
Birds
Mammals
More than 3 legs
3
Repeat the division process
4
Unique Animals
Eagle
Peacock
Lion
Bear
Spider
Scorpion
Zoumana KEITA

Source:

03

Dimensionality Reduction

Dimensionality Reduction – Simplifying complex data without losing key information

Why It Matters

Datasets in research or capstone projects often have **dozens or hundreds of variables**. Reducing dimensions helps:

Visualize your data (2D/3D plots)

Improve performance in downstream models

Remove redundant or noisy features

Common Techniques:

PCA (Principal Component Analysis): Projects high-dimensional data into lower dimensions.

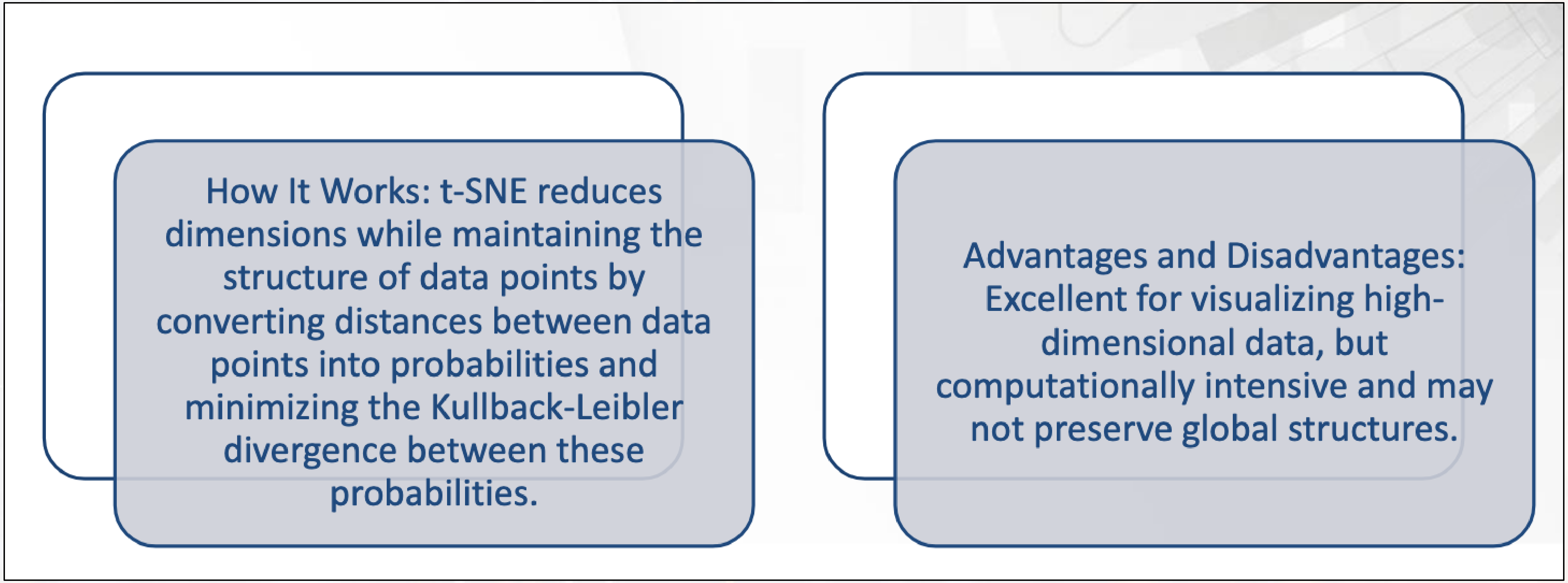
t-SNE: Great for visualizing data clusters in 2D space.

Principal Component Analysis (PCA)

How It Works: PCA transforms data into a new coordinate system such that the greatest variance comes to lie on the first axis, the second greatest variance on the second axis, and so on.

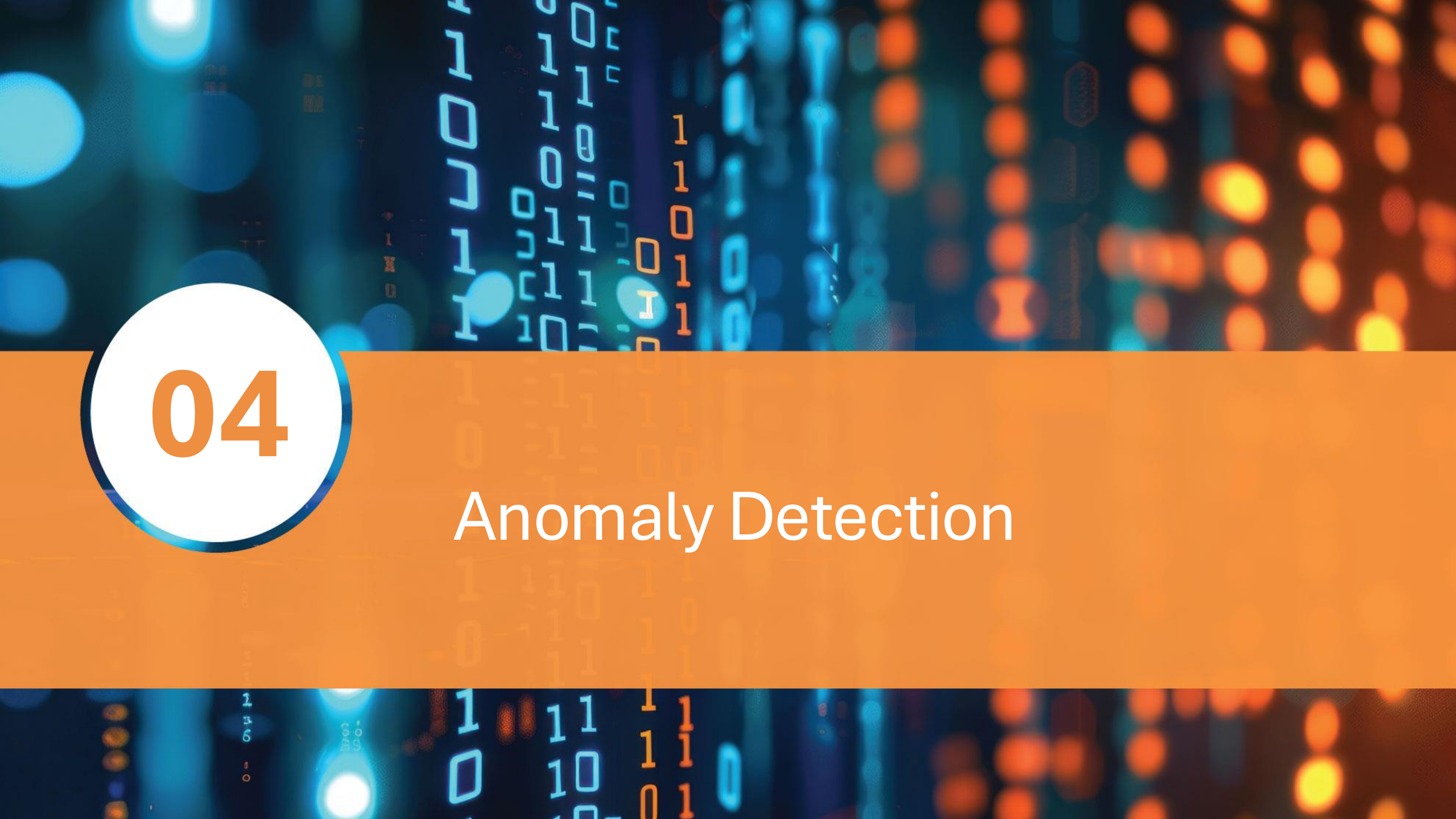
Advantages and Disadvantages: Reduces complexity but can lose interpretability and important information.

t-SNE (t-Distributed Stochastic Neighbor Embedding)



How It Works: t-SNE reduces dimensions while maintaining the structure of data points by converting distances between data points into probabilities and minimizing the Kullback-Leibler divergence between these probabilities.

Advantages and Disadvantages: Excellent for visualizing high-dimensional data, but computationally intensive and may not preserve global structures.

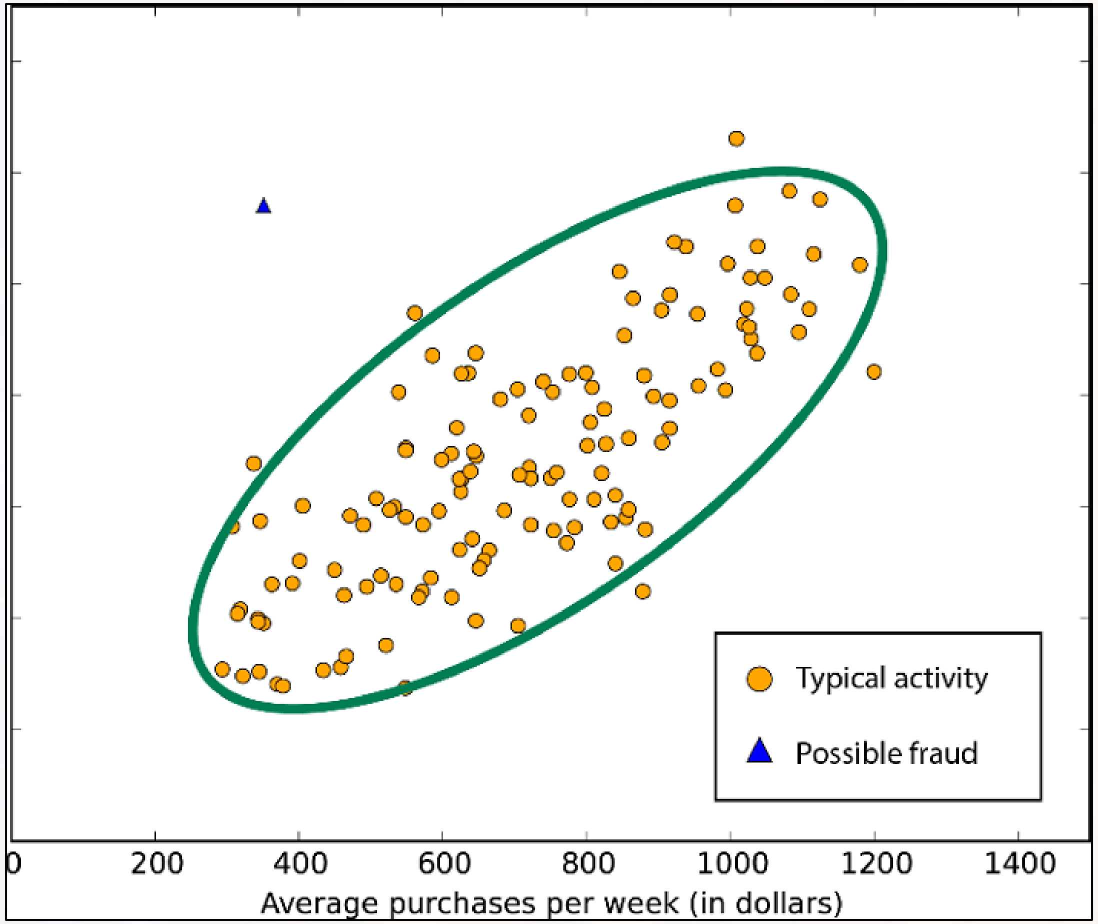


04

Anomaly Detection

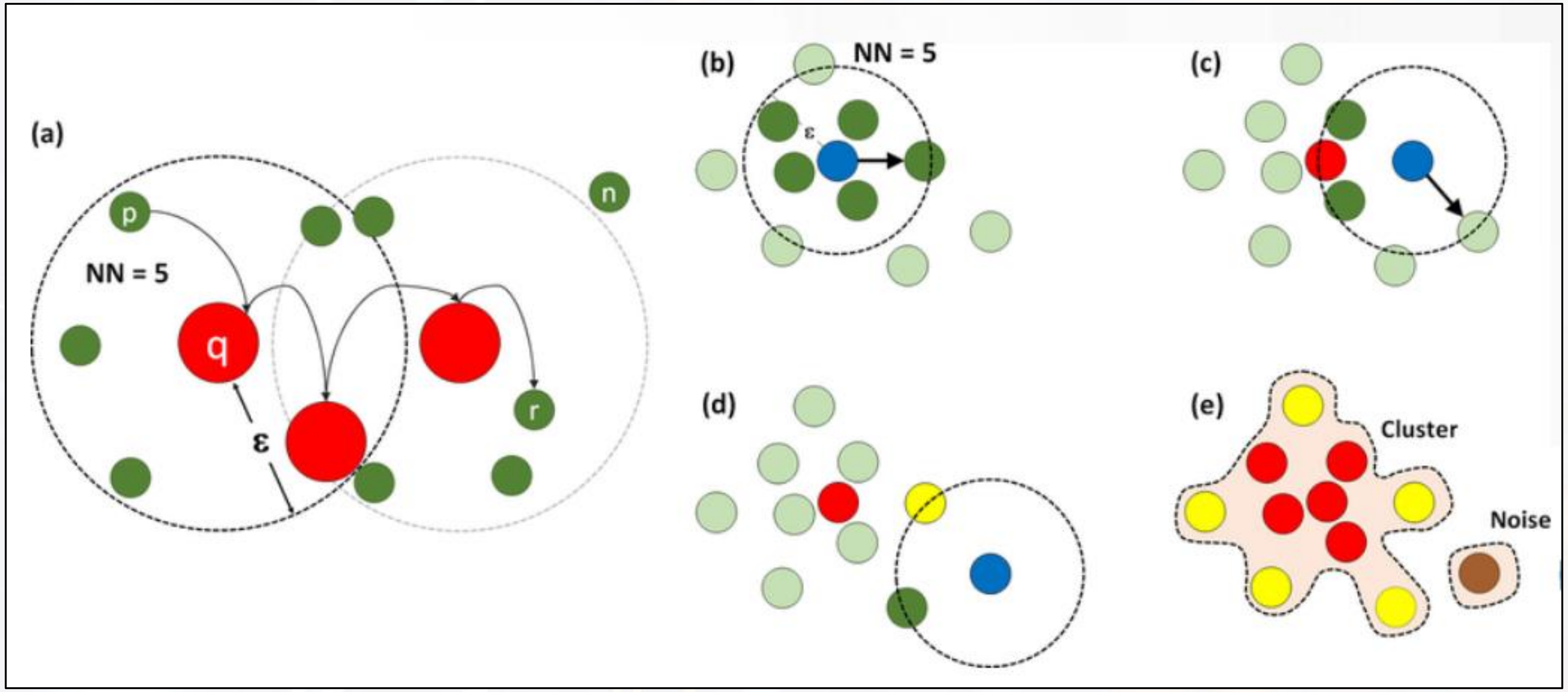
Anomaly Detection

- Identifying outliers in the data
- Data points that deviate significantly from the norm
- Let's say your university is tracking logins to the online learning portal. An anomaly detection model can flag:
  - Suspicious logins at 3AM from another country
  - Students whose activity drops suddenly (could indicate dropout risk)
- **Industry Use Cases:**
  - Fraud Detection in banks
  - Intrusion Detection in cybersecurity
  - Quality control in manufacturing
  - Fraud detection in Finance (unusual transactions)
  - System Failure Detection



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **How It Works**: Clusters data based on the density of points, identifying core points, reachable points, and outliers (noise).
- **Advantages and Disadvantages**: Can find arbitrarily shaped clusters. Robust to noise. Can be sensitive to parameter settings



Association Rule Learning

- **Purpose**: Discover interesting relationships or associations between variables in large datasets.
  - involves identifying rules that describe the co-occurrence of items
  - Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Bread -> Peanut butter
- **Usage**: Commonly used in market basket analysis, recommendation systems, and event correlation.
- **Techniques**: Techniques include the Apriori algorithm, Eclat algorithm, and FP-Growth algorithm.

05

Conclusion

Challenges in Unsupervised Learning

Data Quality: Issues related to noise, missing values, and outliers

Interpretability: Difficulty in interpreting the results.

Scalability: Challenges in scaling algorithms to handle large datasets

Keywords You Should Know

- Clustering – Grouping similar items
- Outlier / Anomaly – A rare or unexpected data point
- Dimensionality – The number of variables/features in a dataset
- PCA – A method to reduce dimensions while keeping important information

What You Should Remember

Unsupervised learning is about letting AI discover structure in data — without labels

It’s used to **group**, **simplify**, and **detect unusual patterns**

In 2025, it powers **AI assistants**, **recommendation engines**, **cybersecurity**, and even **disease diagnosis**

As a student, understanding this will boost your skills in **data science**, **research**, and **AI development**

Remember

The best algorithm is the one that solves your specific problem effectively. Start simple, understand your data, and iterate based on results.

Thank you!

06

Practical

Practical Clustering

- K-Means clustering is a popular unsupervised learning algorithm used to partition data points into distinct groups based on similarity. In this section, we will dive into the theory behind K-Means clustering and explore its implementation in Python using the scikit-learn library.

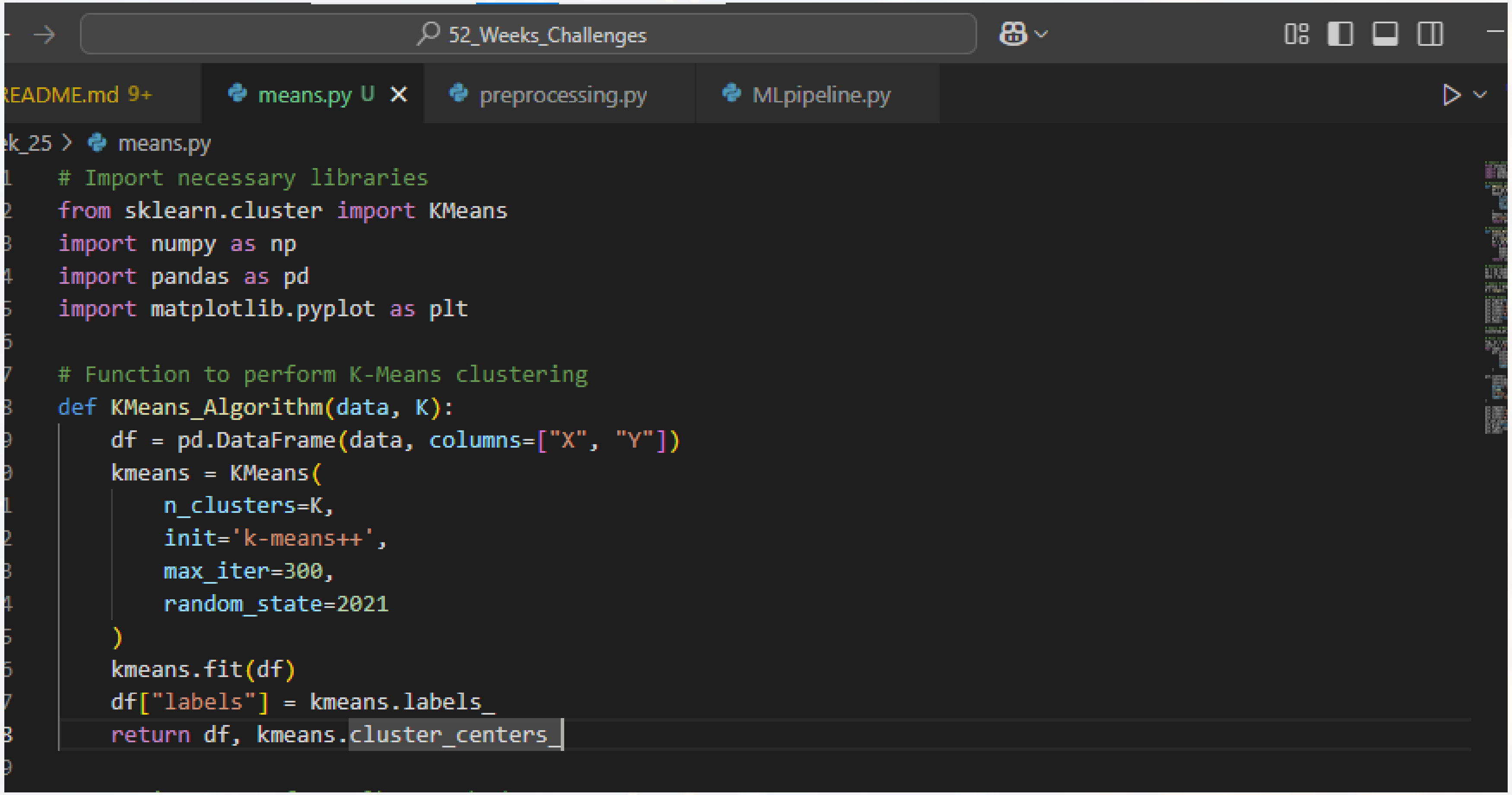
Practical Clustering

- One of the most popular clustering approaches for clustering observations into groups is the unsupervised clustering algorithm **K-Means**.
- Following are conditions for K-Means clustering:
  - number of clusters needs to be specified in advance: K
  - every observation needs to belong to at least one class
  - every observation need to belong to only one class (classes need to be non-overlapping)
  - no one observation should belong to more than 1 class

Practical Clustering

- Mathematically, the within-cluster variation is defined based on the choice of distance measure which you can choose yourself.
  - For instance, as distance measure you can use Euclidean distance, Manhattan distance etc.
- K-means clustering is optimal when the within-cluster variation is the smallest. The within-cluster variation of C_k cluster is a measure W(C_k) of the amount by which the observations in a cluster differs from each other.

Practical Clustering

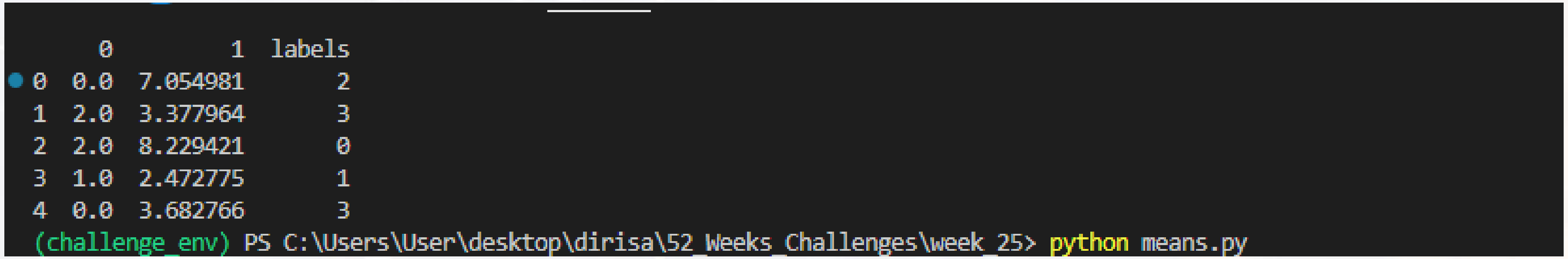


```python
# Import necessary libraries
from sklearn.cluster import KMeans
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Function to perform K-Means clustering
def KMeans_Algorithm(data, K):
    df = pd.DataFrame(data, columns=["X", "Y"])
    kmeans = KMeans(
        n_clusters=K,
        init='k-means++',
        max_iter=300,
        random_state=2021
    )
    kmeans.fit(df)
    df["labels"] = kmeans.labels_
    return df, kmeans.cluster_centers_
```

Practical Clustering

- This script is designed to generate synthetic data, apply K-Means clustering, and assign cluster labels to each data point.
- The K-Means clustering algorithm is an unsupervised machine learning method that groups similar data points into clusters based on their proximity in feature



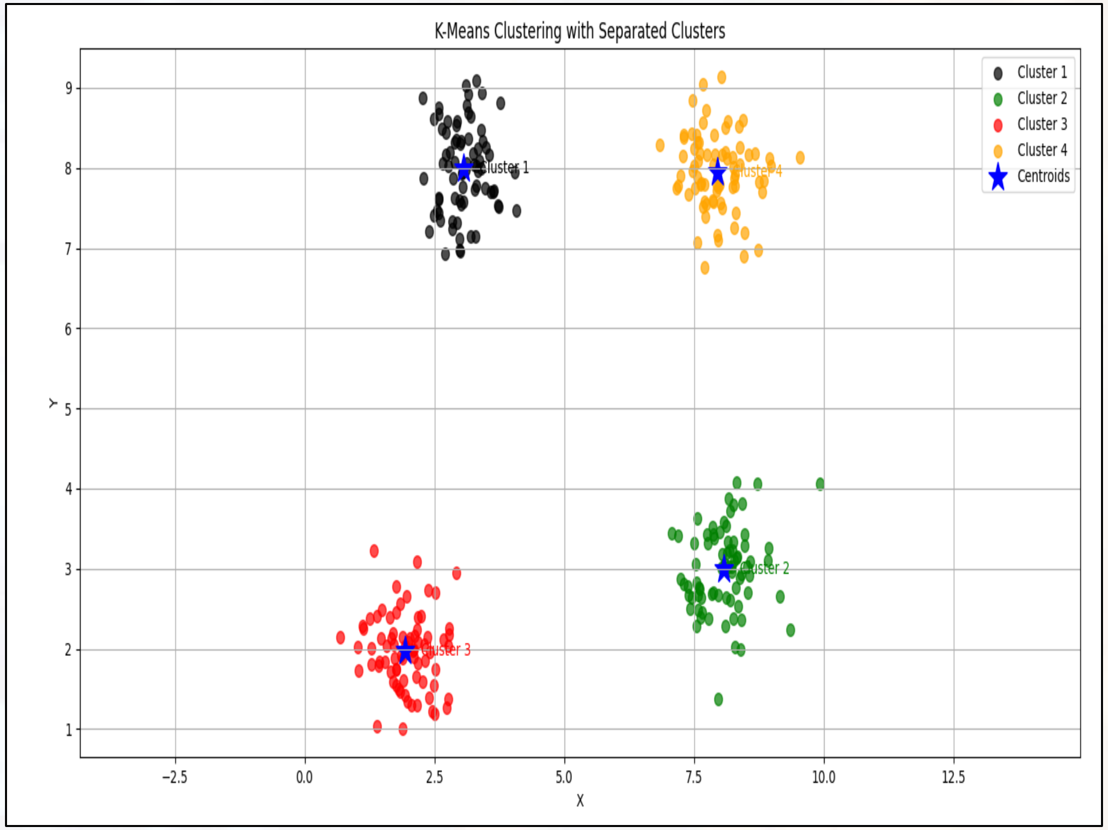
```
     0         1  labels
0  0.0  7.054981       2
1  2.0  3.377964       3
2  2.0  8.229421       0
3  1.0  2.472775       1
4  0.0  3.682766       3
(challenge_env) PS C:\Users\User\desktop\dirisa\52_Weeks_Challenges\week_25> python means.py
```

Practical Clustering

- Let’s briefly understand how the K-Means algorithm works. The algorithm follows these steps:
- **Step 1: Initialization** – Randomly select K centroids, where K represents the desired number of clusters.
- **Step 2: Assignment** – Assign each data point to the nearest centroid based on the Euclidean distance.
- **Step 3: Update** – Recalculate the centroids by taking the mean of all data points assigned to each cluster.
- **Step 4: Repeat** – Repeat steps 2 and 3 until convergence criteria are met (e.g., minimal centroid movement).

Practical Clustering

- In this figure, K-means has clustered these observations into 4 groups. And as you can see from the visualisation, the way observations have been clustered even by the graph seems natural and it makes sense.

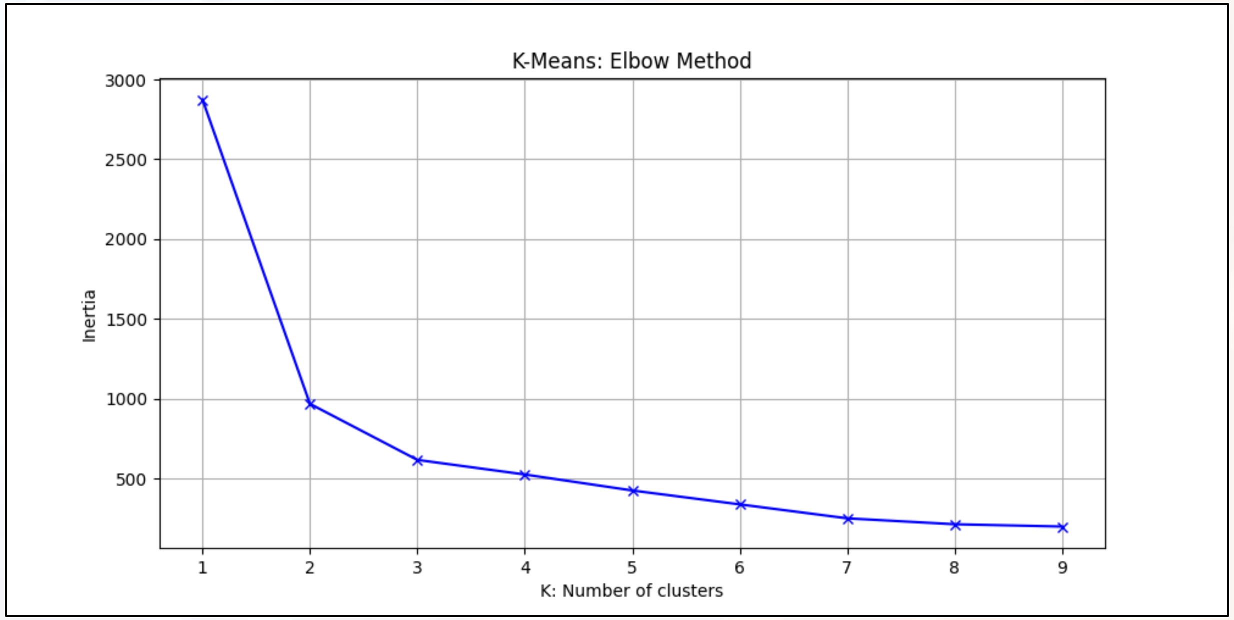


Elbow Method for Optimal Number of Clusters (K)

- One of the most popular methods to determine this optimal value of K, or number of clusters, is the **Elbow Method**.
- To use this approach, you need to know what **Inertia** is.
- Inertia is the sum of squared distances of samples to their closest cluster center.
- So, the Inertia or **within cluster of sum of squares** value gives an indication of how coherent the different clusters are or how pure they are.

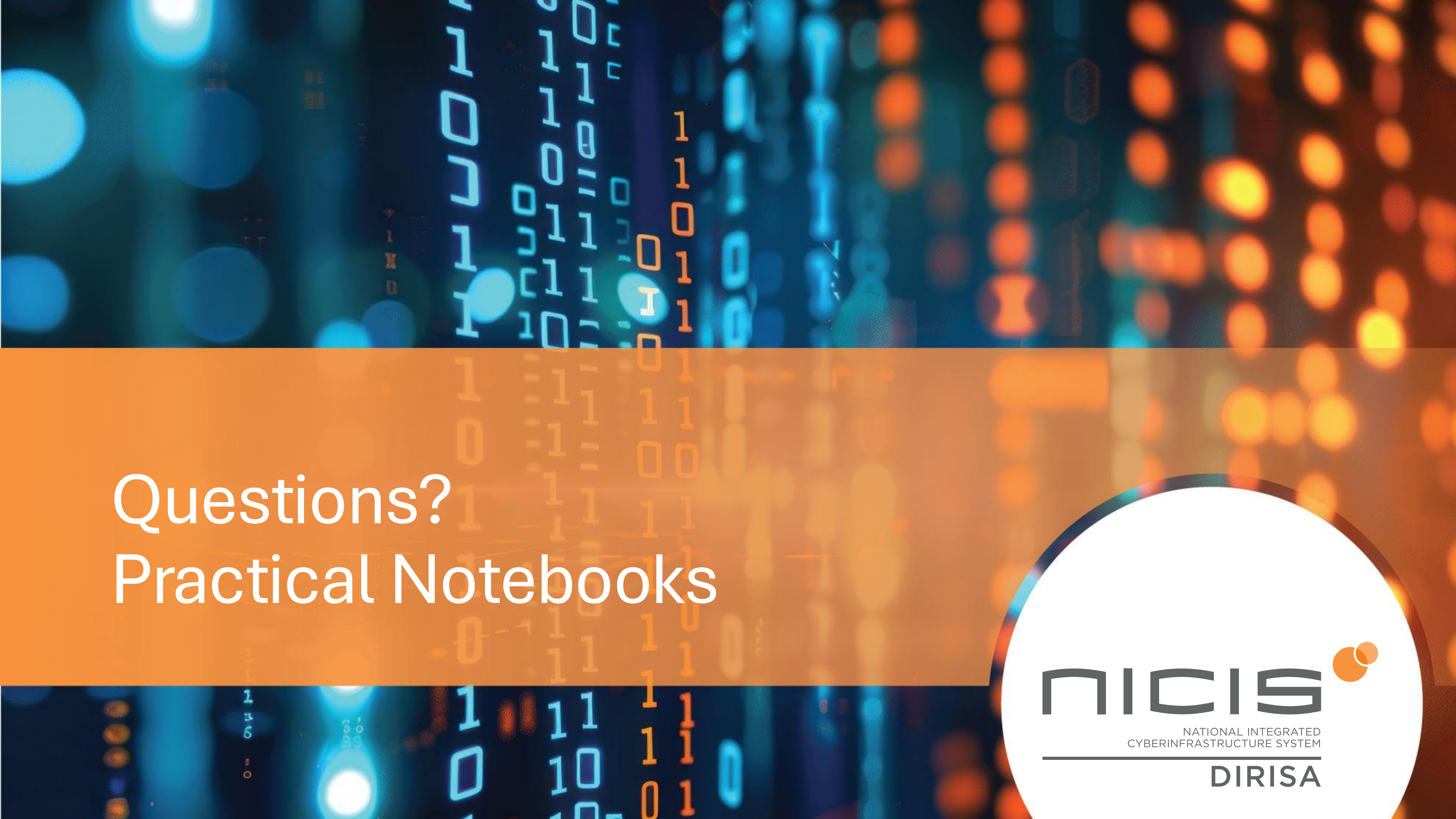
Elbow Method for Optimal Number of Clusters (K)

- Then we can calculate the inertia for different number of clusters K.
- We can plot this as in the following figure where we consider K = 1,2,….,10.
- Then from the graph we can select the K corresponding to the Inertia where the elbow occurs.
- In this case, K = 3 where the Elbow happens.



Elbow Method for Optimal Number of Clusters (K)

- K-Means is a non-deterministic approach and it’s randomness comes in Step 1, where all observations are randomly assigned to 1 of the K classes.
- So as you can see, K-Means clustering offers an efficient and effective approach to grouping data points based on similarity



Questions?
Practical Notebooks

nicis

NATIONAL INTEGRATED
CYBERINFRASTRUCTURE SYSTEM

DIRISA