An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Unsupervised Learning

iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Let’s learn about Unsupervised Learning!

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

fig ref: pierian data

Supervised Learning

Using historical **labeled** data, predict a label on new data (regression or classification).

Unsupervised Learning

Using **unlabeled** data, discover patterns, clusters, or significant components.

fig ref: pierian data

Unsupervised Learning:

- Clustering:
  - Using features, group together data rows into distinct clusters.
- Dimensionality Reduction:
  - Using features, discover how to combine and reduce into fewer components.

Supervised performance metrics will not apply for unsupervised learning!

Then How can we compare to a correct label answer, if there was no label?

fig ref: pierian data

We will need to figure out other ways of assessing unsupervised model performance or reasonableness.

Infact, our understanding of what “performance” actually means will need to change with unsupervised learning!

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Machine Learning Pathway for Unsupervised Learning

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

fig ref: pierian data

Real World

Raw Data

Process & Store Data

SQL Database, CSV files, Excel, Cloud Storage

fig ref: pierian data

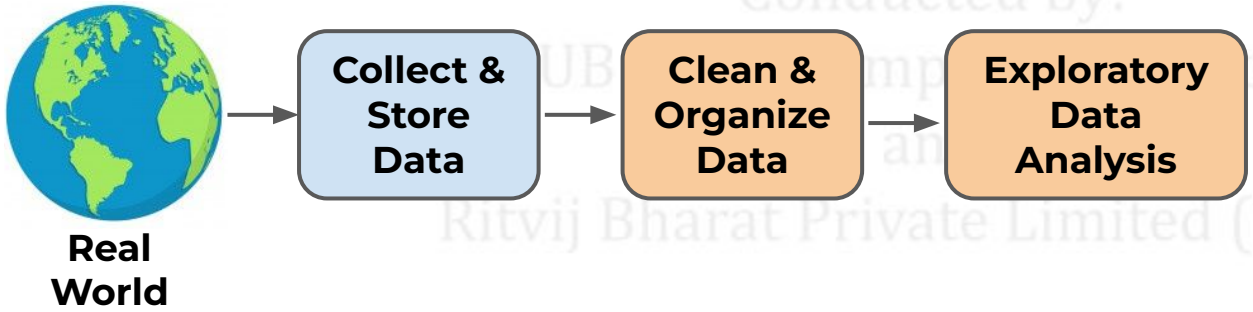
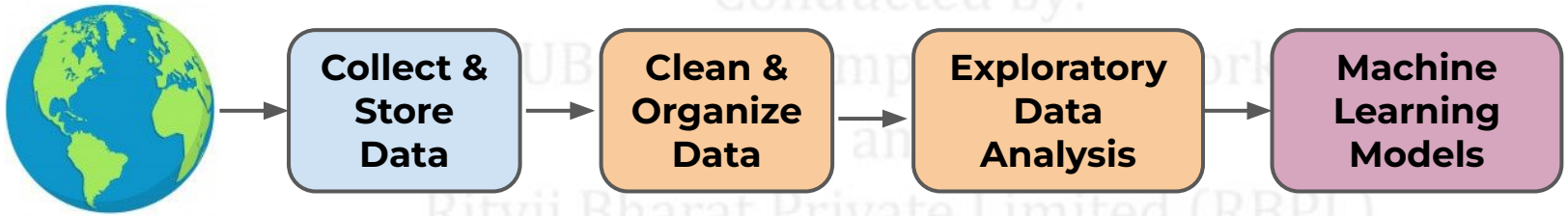


fig ref: pierian data



Supervised Learning:
Predict an Outcome
Unsupervised Learning:
Discover Patterns in Data

fig ref: pierian data

Real World

Collect & Store Data

Clean & Organize Data

Exploratory Data Analysis

Clustering

Dimensionality Reduction

Unsupervised Learning:
Discover Patterns in Data

fig ref: pierian data

Clustering: If we have unlabeled data, can we attempt to cluster or group similar data points together to “discover” possible labels for clusters?

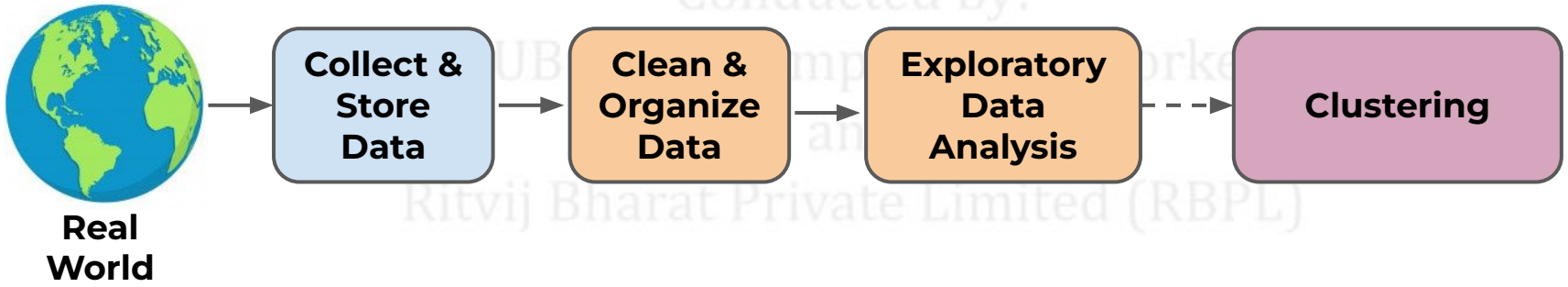


fig ref: pierian data

Dimensionality Reduction: If we have unlabeled data, can we attempt to reduce the number of features by combining them into new components? Do these new components give us further insight for the data?

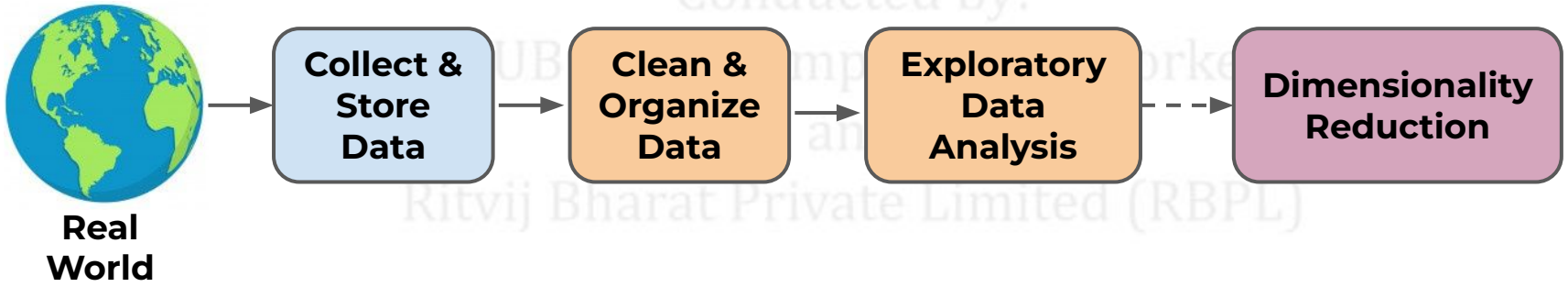


fig ref: pierian data

1. K-Means
2. Hierarchical clustering
3. Dimensionality reduction.

Methods for interpreting the model results

fig ref: pierian data

Things keep in mind:

- *What does it really mean to “discover” labels through clustering?*
- *Without known labels how do we measure performance?*
- *Do combinations of features hold important insights?*

fig ref: pierian data

Let’s get started!

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

K-Means Clustering

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

- Understanding Clustering
- Intuition of K-Means
- Mathematical Theory of K-Means
- Example of K-Means

Presented by:
Shreyas Shukla

fig ref: pierian data

Do not confuse K-Means with KNN!

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

General Concepts

fig ref: pierian data

Clustering uses **unlabeled data**
It looks for similarities between groups (clusters) in order to attempt to segment the data into separate clusters.

Keep in mind that we don't actually know the true correct label for this data!

Imagine an example data set:

X1	X2
2	4
6	3
...	...
1	2

How could we cluster this data together?

X1	X2
2	4
6	3
...	...
1	2

fig ref: pierian data

Intuitively, we see 2 groupings:

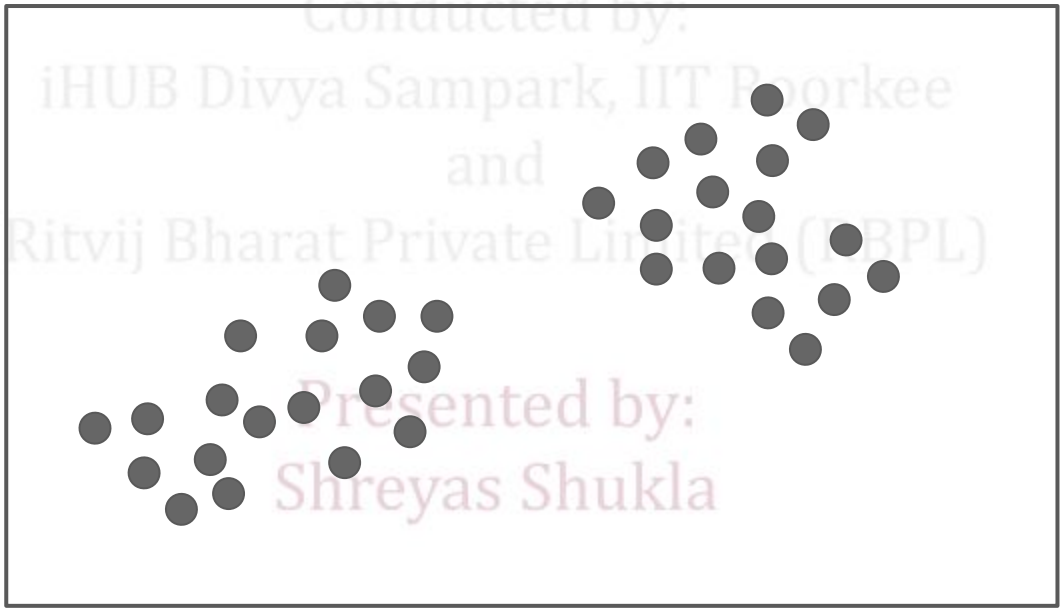


fig ref: pierian data

Note how distance is the intuitive metric:

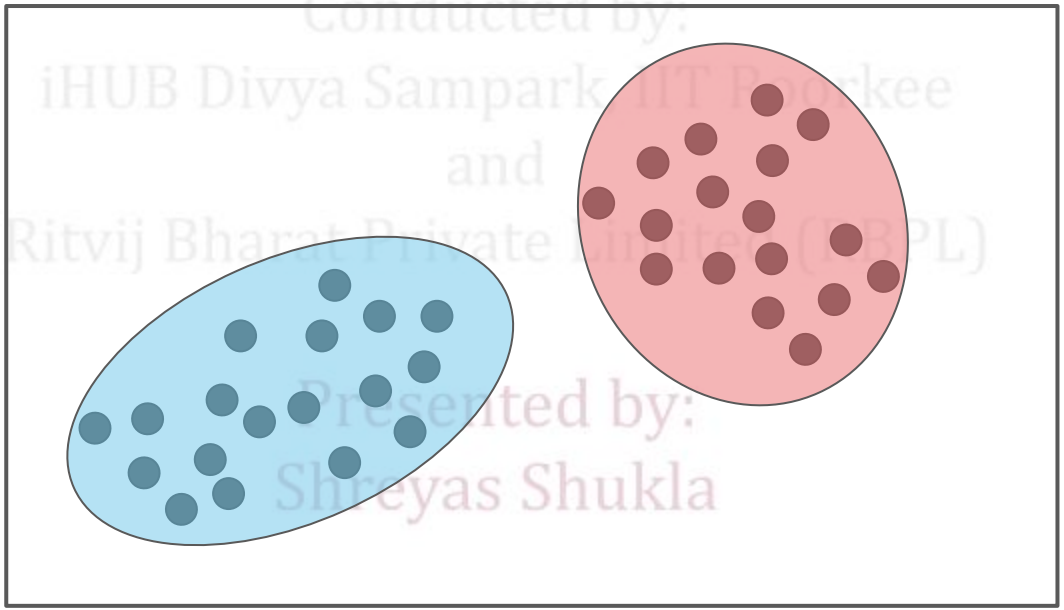


fig ref: pierian data

Assign clusters:

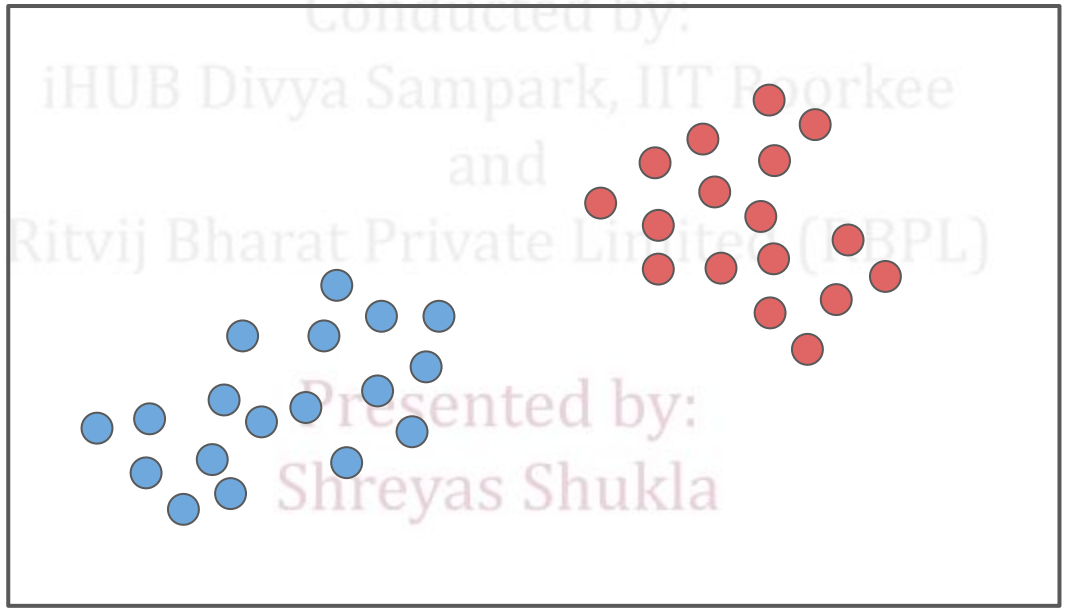


fig ref: pierian data

We don’t actually know for sure if this is a correct way of grouping together these data points, there was no correct label to begin with!

Also what about situations that are not so obvious or multi-dimensional?

2 or 3 clusters could both be reasonable:

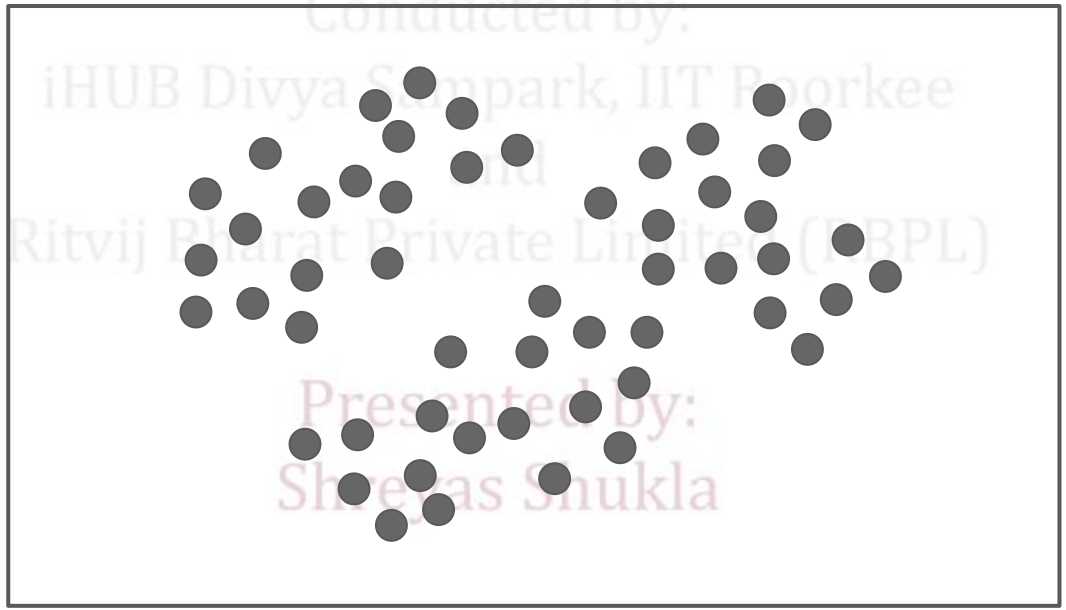


fig ref: pierian data

2 or 3 clusters could both be reasonable:

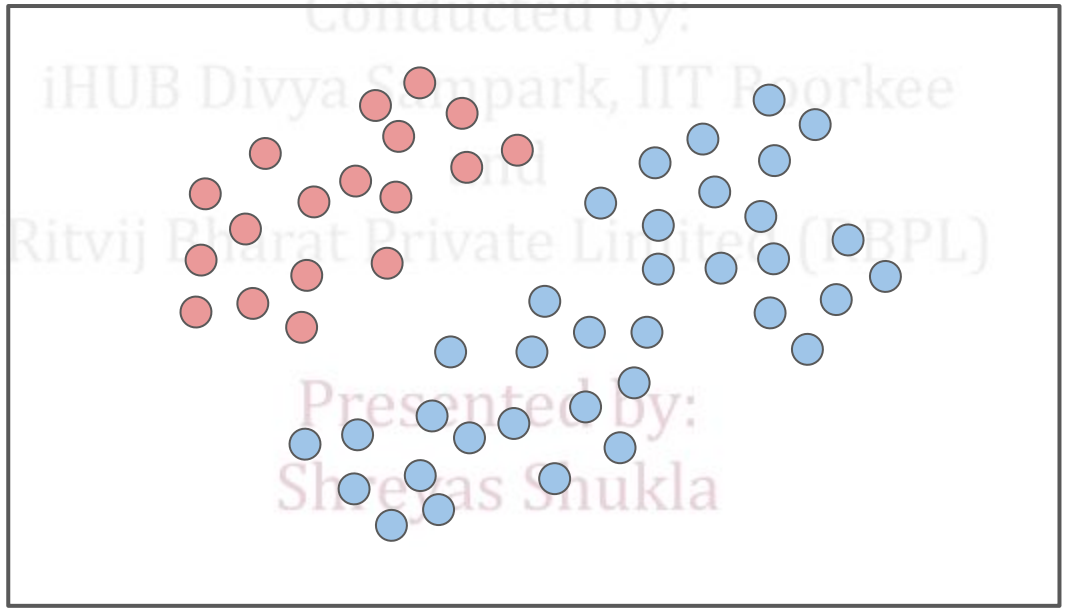


fig ref: pierian data

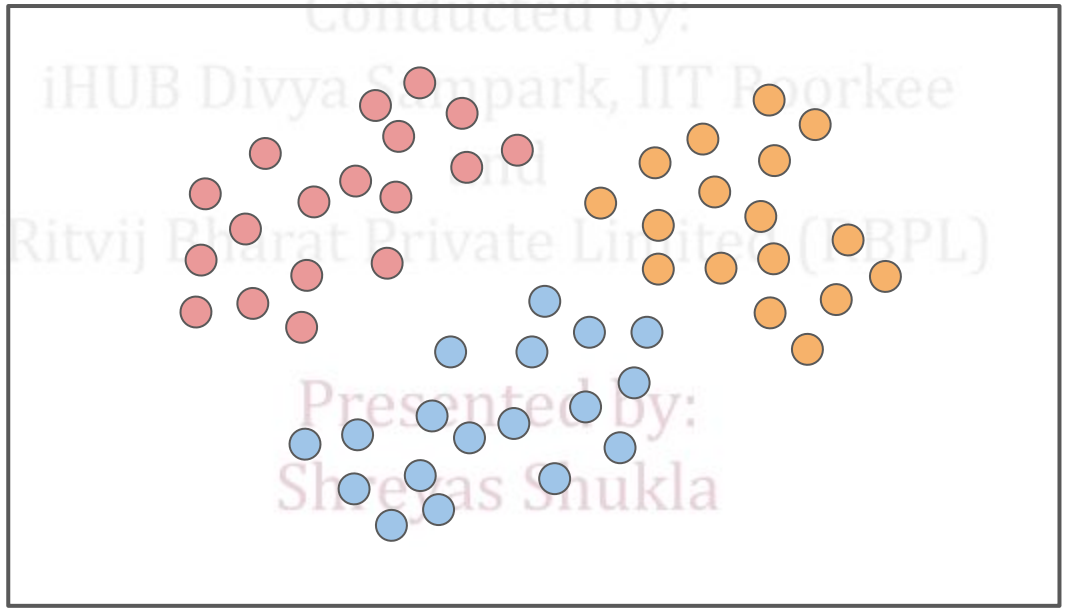


fig ref: pierian data

Different methods can be used to decide!

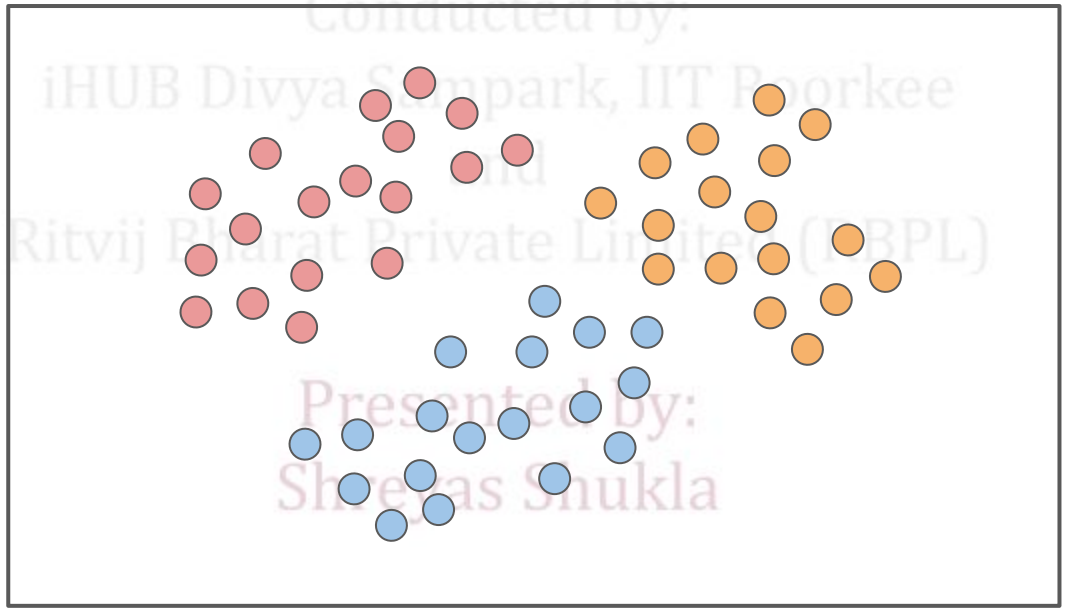


fig ref: pierian data

Clustering doesn't "label" these for you!

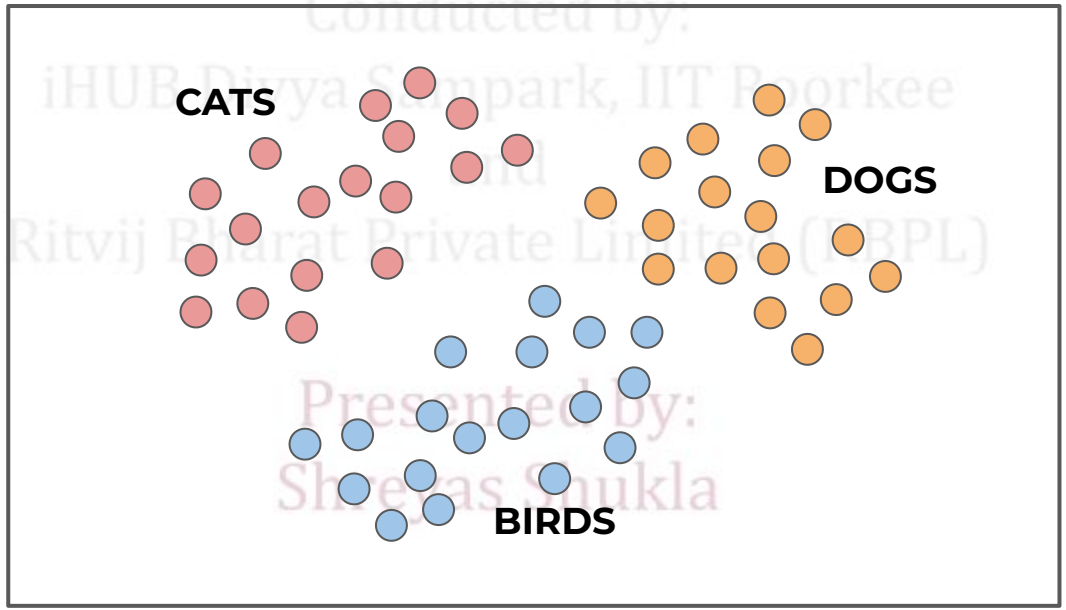


fig ref: pierian data

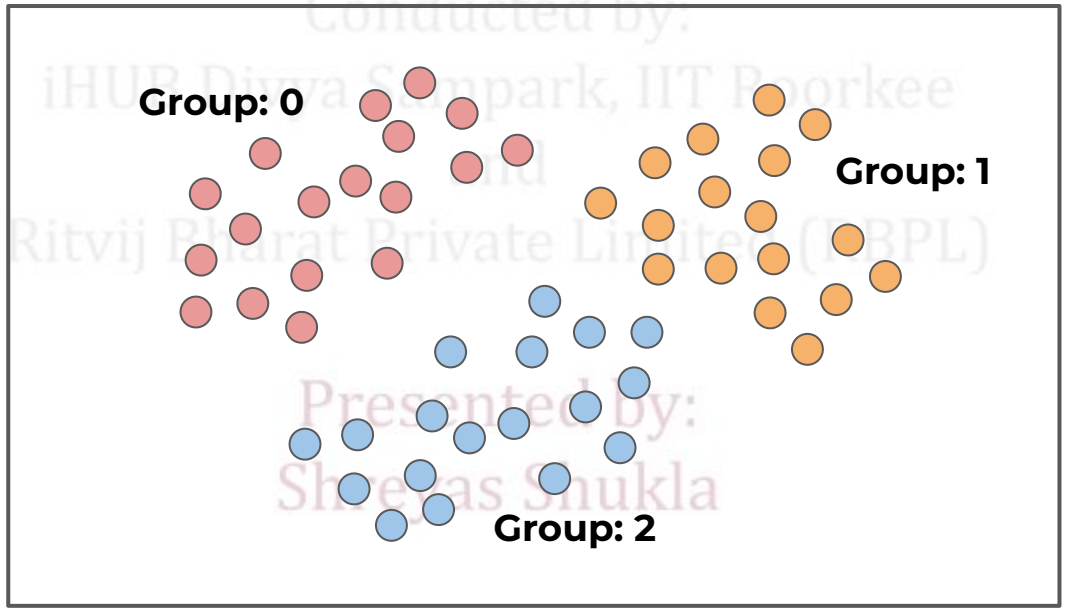


fig ref: pierian data

Main Clustering Ideas:

1. Use features to decide which points are most similar to other points.
2. There is no final correct **y** label to compare cluster results to.
3. Clustering is an unsupervised learning process that "discovers" potential labels.

- *What about a new unlabeled data point?*
- *How do we assign it to a cluster?*
- *Was it the correct cluster for assignment?*

- *How do we assign a new data point to a cluster?*
  - Different approaches depending on the unsupervised learning algorithm used.
  - Use features to assign most appropriate cluster.

- *If we've discovered these new cluster labels, could we use that as a **y** for supervised training?*
  - Yes! We can use unsupervised learning to discover possible labels, then apply supervised learning on new data points.

- *If we've discovered these new cluster labels, could we use that as a **y** for supervised training?*
  - What's the trade-off?

If we've discovered these new cluster labels, could we use that as a ***y*** *for supervised training?*

- Clustering doesn't tell you what these new cluster labels represent, no real way of knowing if these clusters are truly significant.

- How do we decide which number of clusters is best?
- Do we decide or let the algorithm decide?
- How can we measure “goodness of fit” for clustering without a **y** label for comparison?

Machine Learning as an art

What is ground truth?

What trade-offs are we making by using unsupervised learning as a substitute for ground truth of the y label that was not given?

It is much harder to compare unsupervised algorithms against each other due to the lack of ground truth based performance metrics like accuracy or RMSE.

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
and
Ritvij Bharat Private Limited (RBPL)

K-Means Clustering

Intuition and Theory

Presented by:
Shreyas Shukla

fig ref: pierian data

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Working of K-Means clustering

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Presented by:
Shreyas Shukla

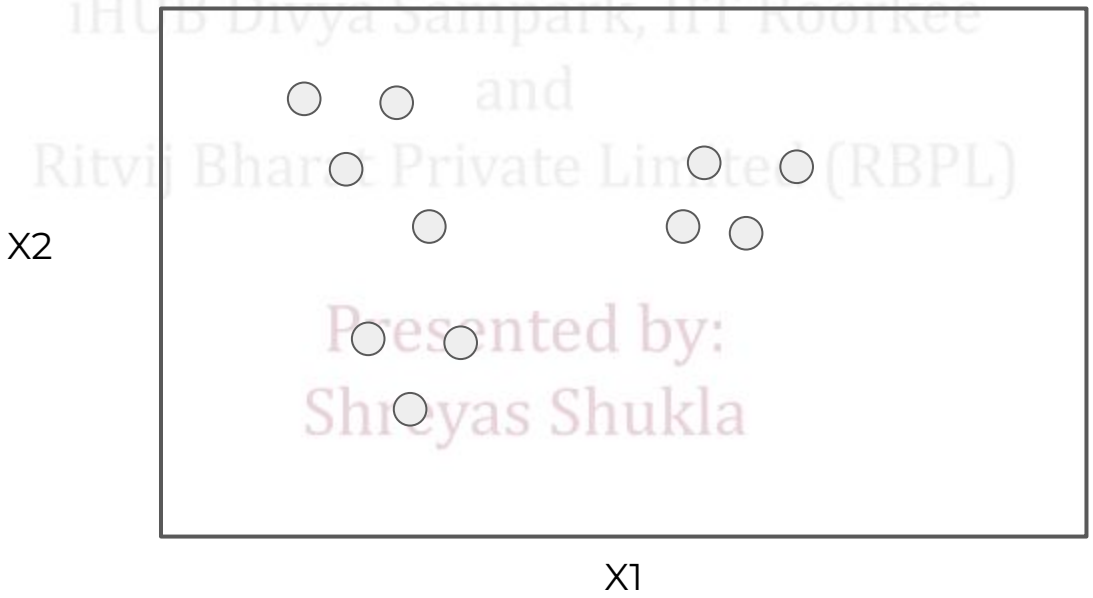
fig ref: pierian data

First, a set of properties each point we must satisfy:

- Each point must belong to a cluster.
- Each point can only belong to one cluster (no single point can belong to multiple clusters).

We'll work with a simple dataset with only 2 features. The process shown here easily extends to **N** feature dimensions.

- Step 0: Start with unlabeled data (only features).



Note: *If we had the group labels, it wouldn't make sense to cluster!*

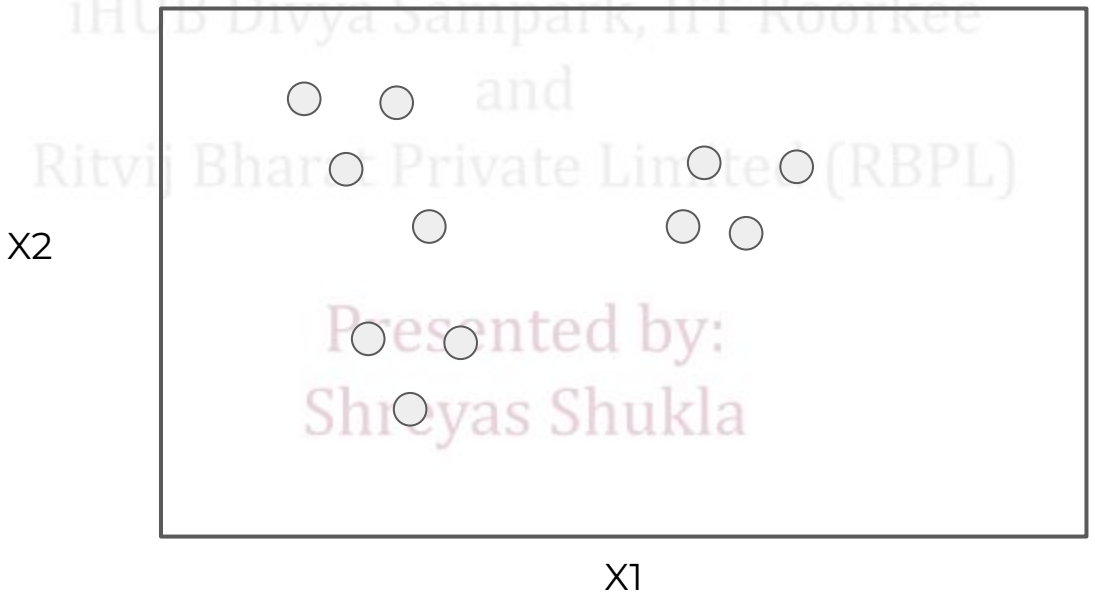


fig ref: pierian data

Step 1: Choose the number of clusters to create (this is the K value).

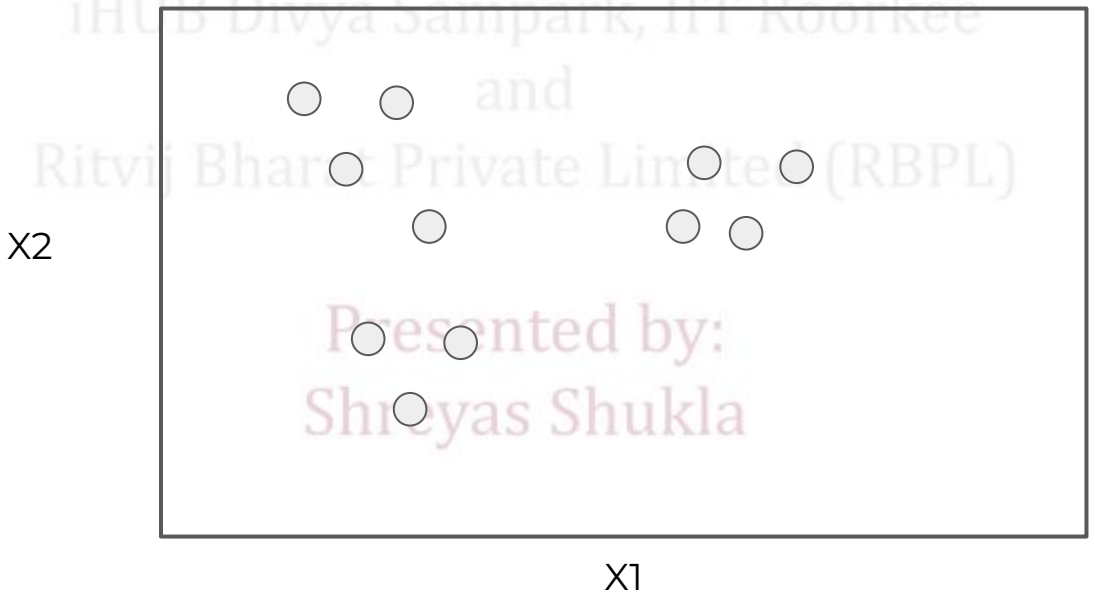


fig ref: pierian data

Step 1: We'll choose K=3. Note in most situations you won't visualize the data!

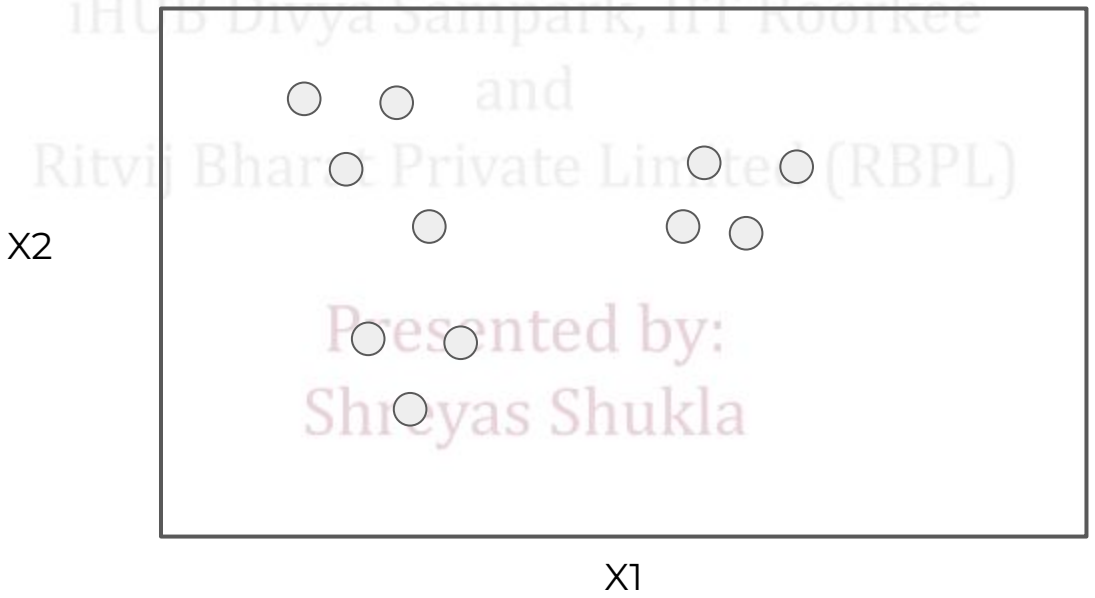


fig ref: pierian data

Step 2: Randomly select K distinct data points.

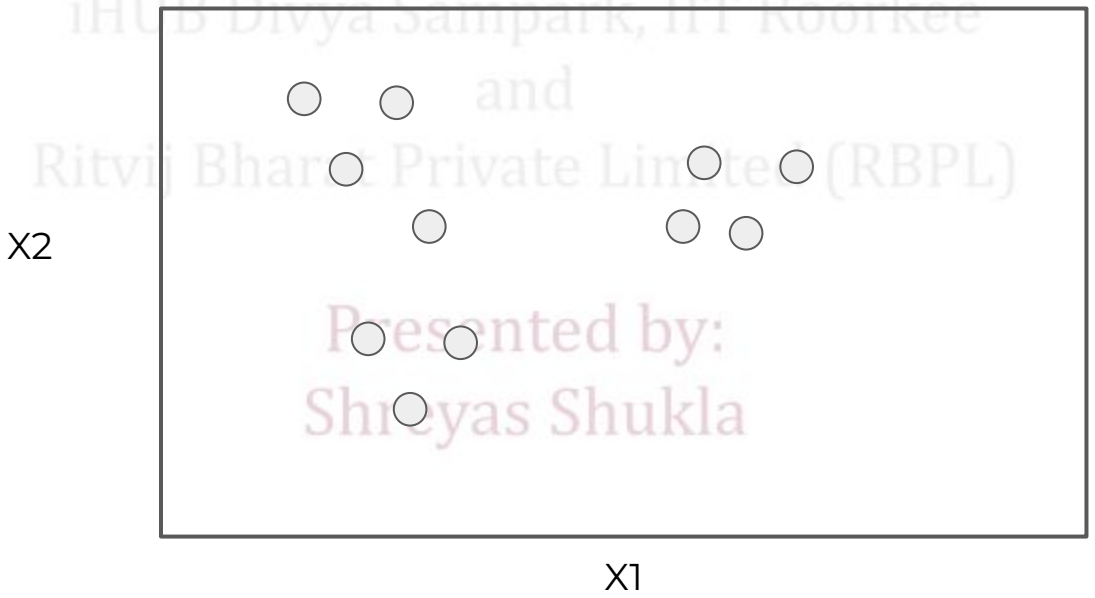


fig ref: pierian data

Step 2: Randomly select K distinct data points. Our K=3. We'll treat these new K points as our cluster points.

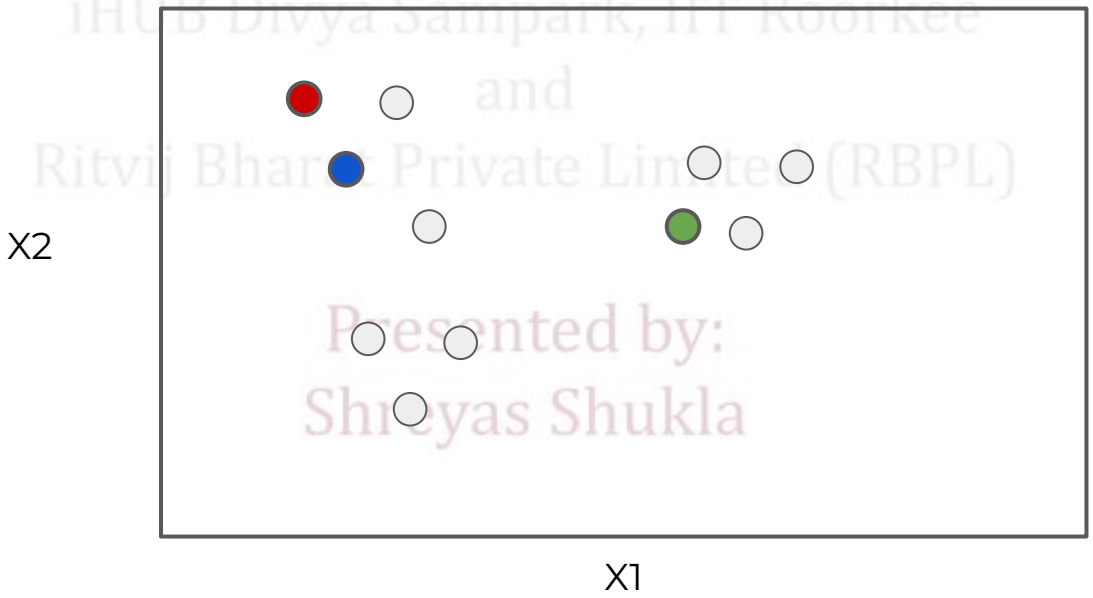


fig ref: pierian data

Step 3: Assign each remaining point to the nearest “cluster” point.

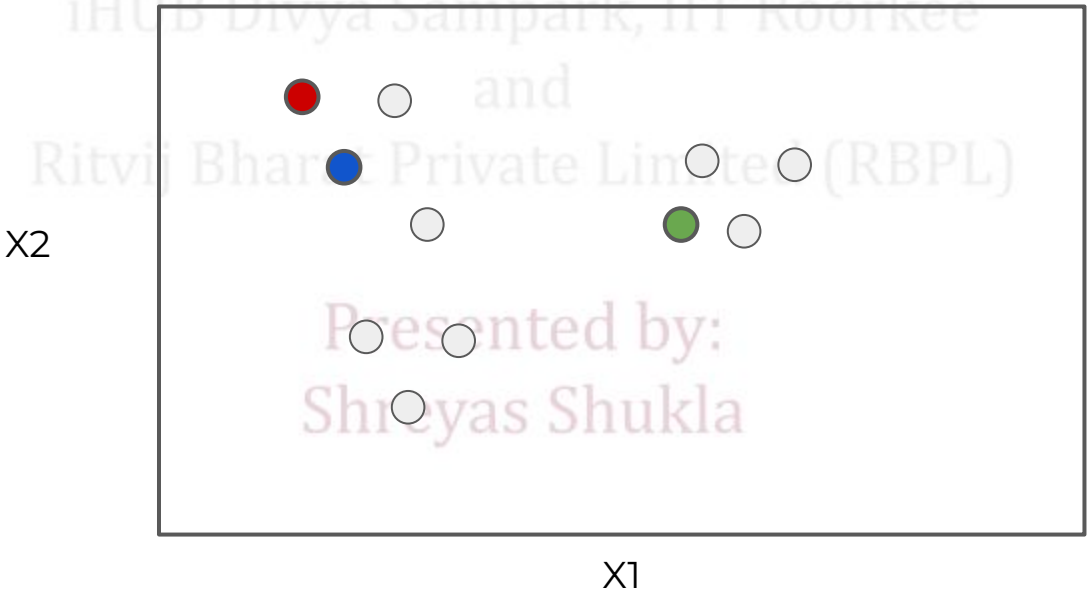


fig ref: pierian data

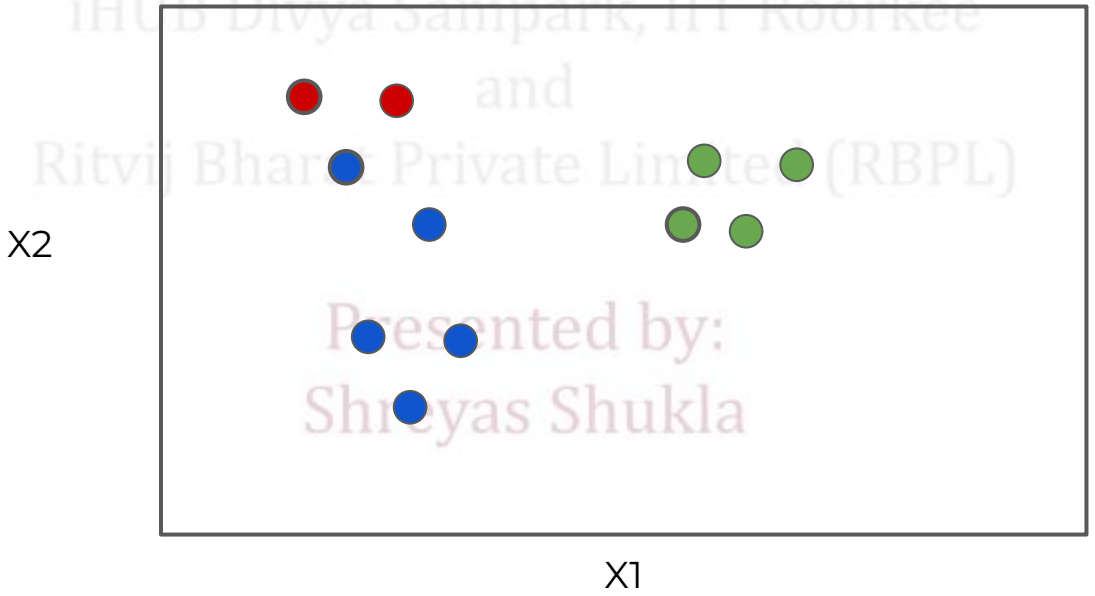


fig ref: pierian data

Step 3: Note how this is using a distance metric to judge the nearest point.

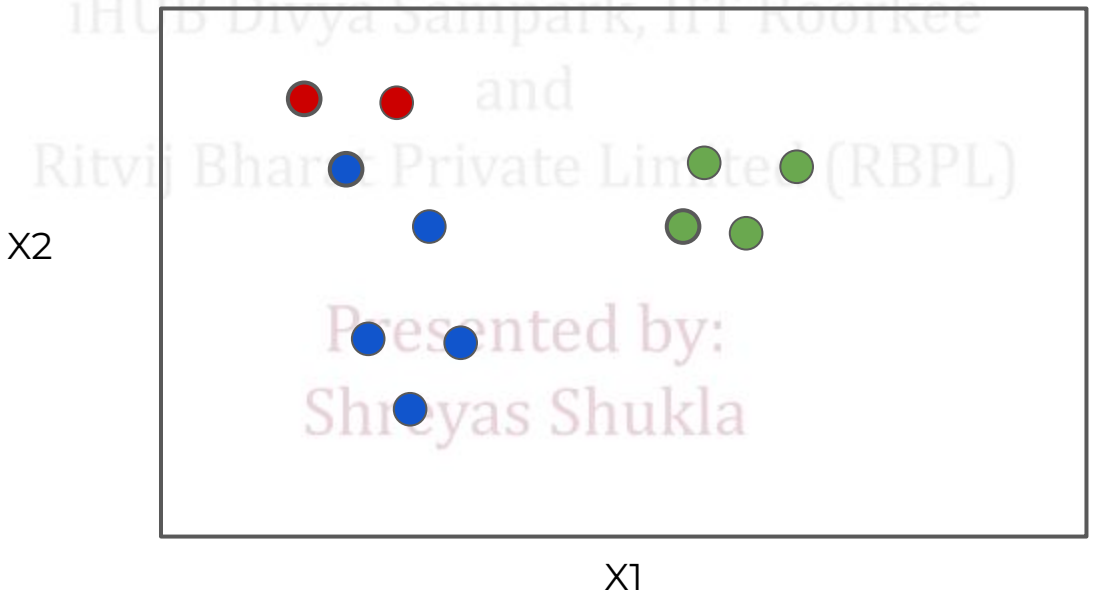


fig ref: pierian data

Step 4: Calculate the center of the cluster points (mean value of each point vector).

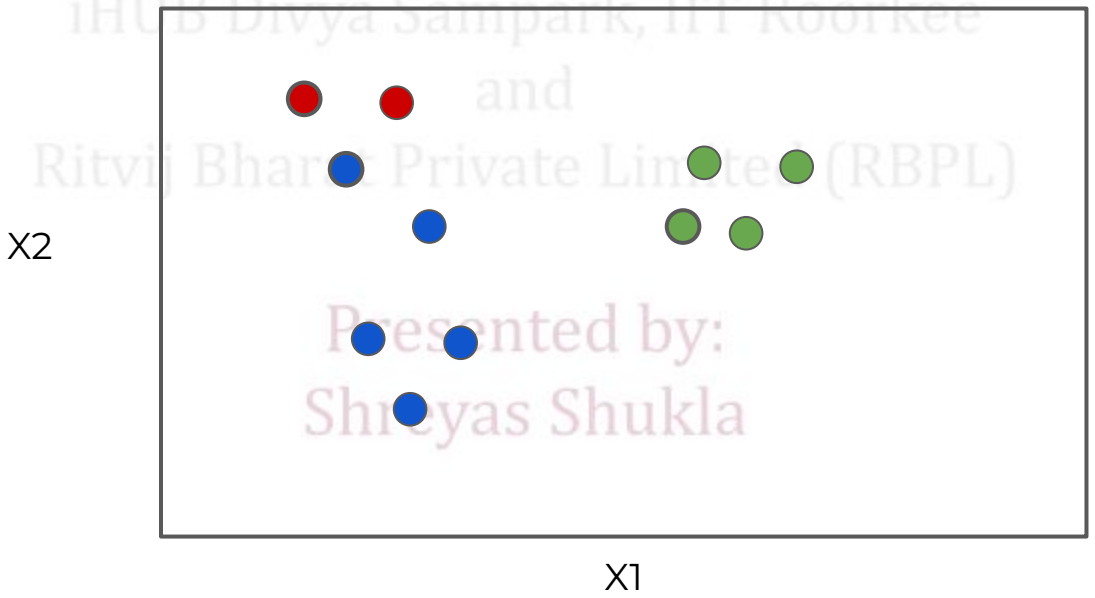


fig ref: pierian data

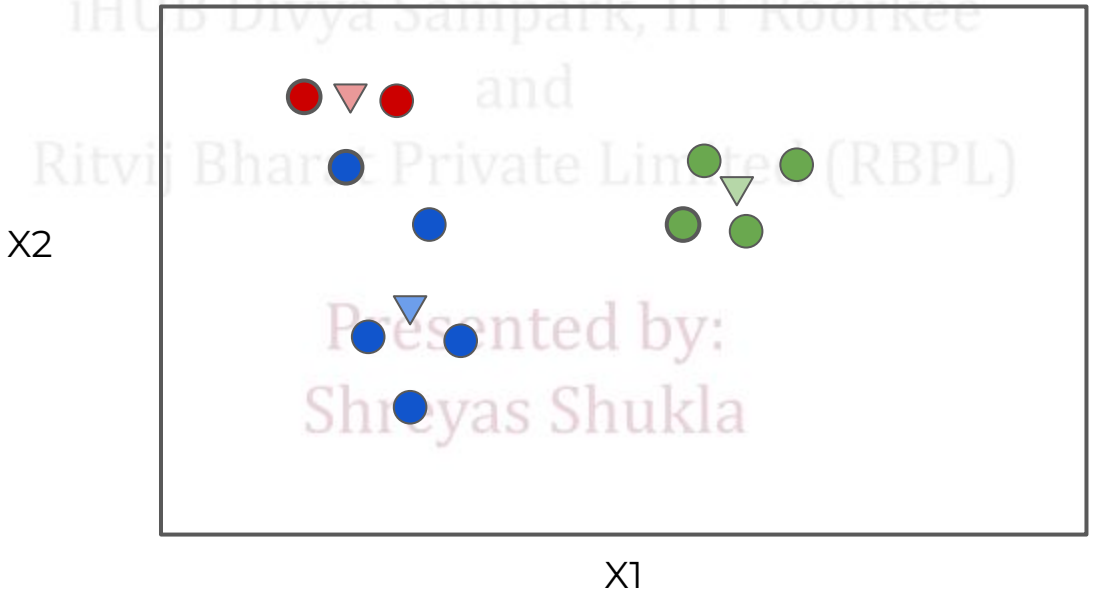


fig ref: pierian data

Step 5: Now assign each point to the nearest cluster center.

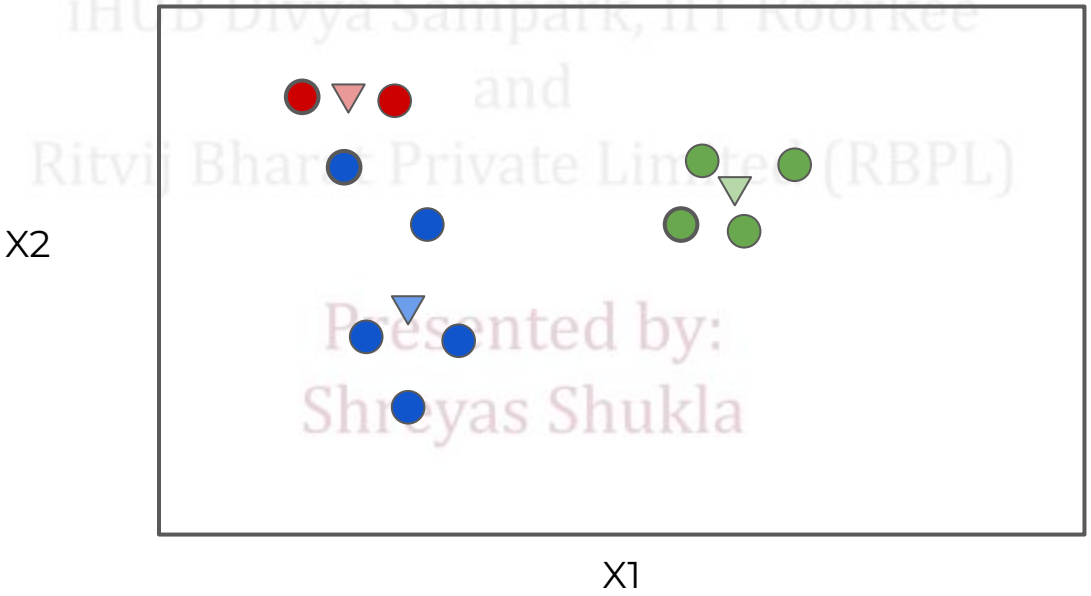


fig ref: pierian data

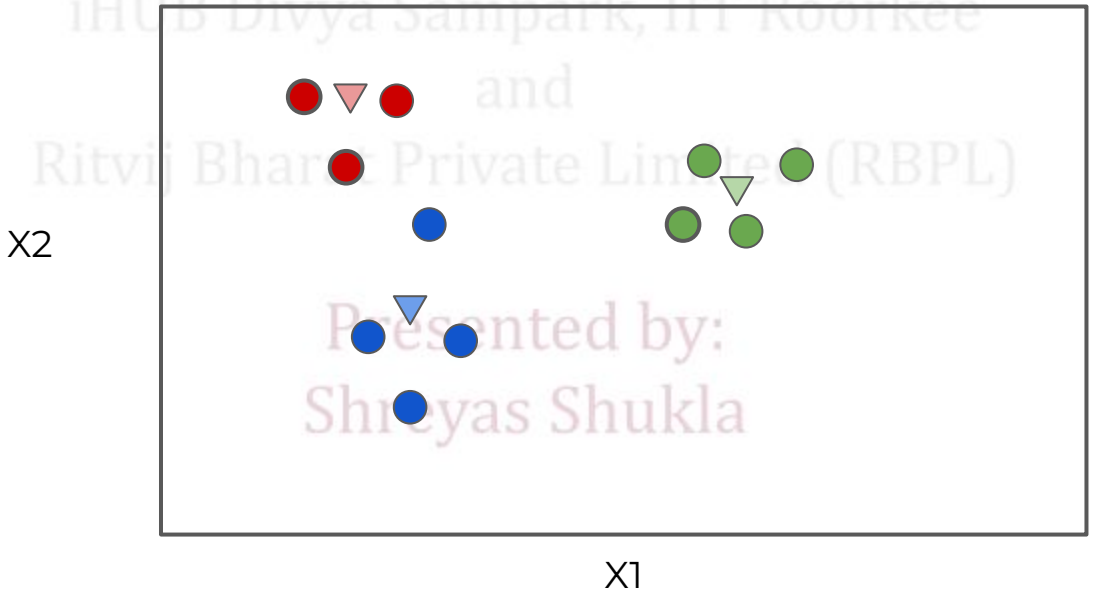


fig ref: pierian data

We repeat steps 4 and 5 until there are no more cluster reassignments.

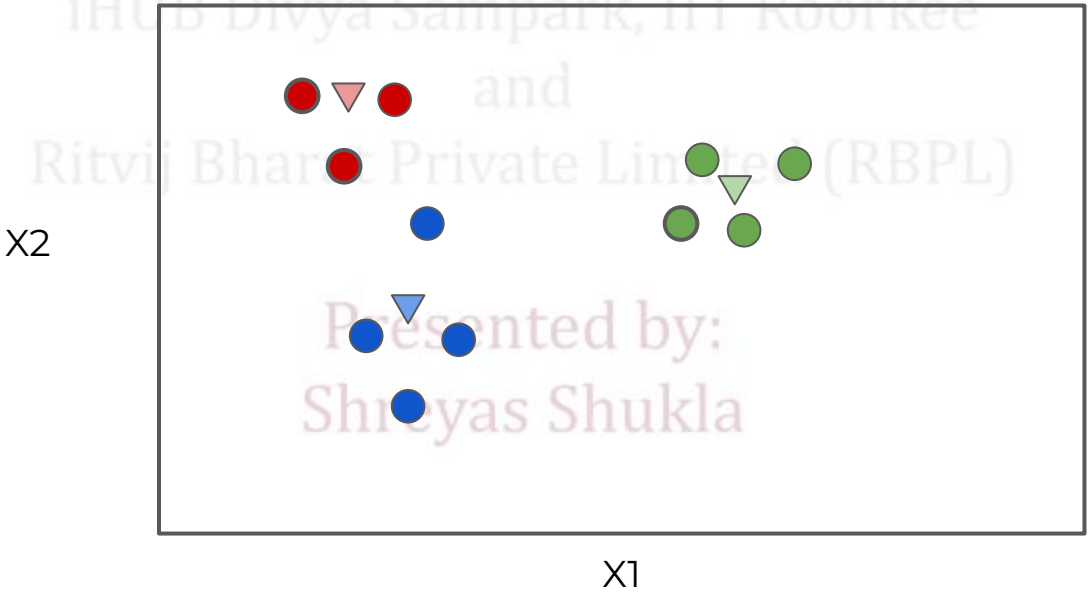


fig ref: pierian data

Step 4b: Recalculate new cluster centers:

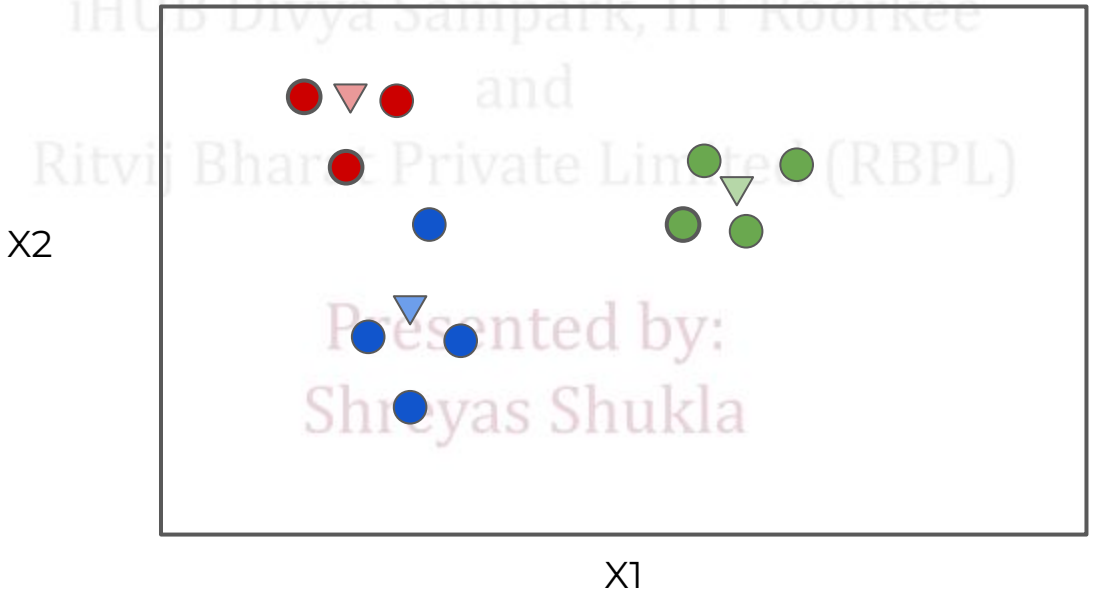


fig ref: pierian data

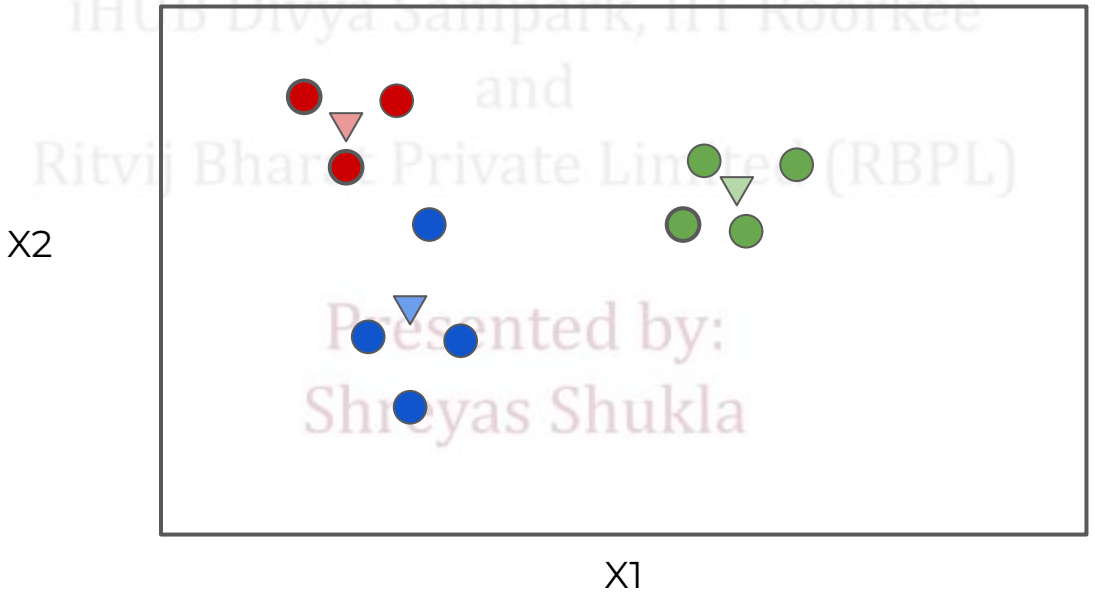


fig ref: pierian data

Step 5b: Assign points to nearest cluster center.

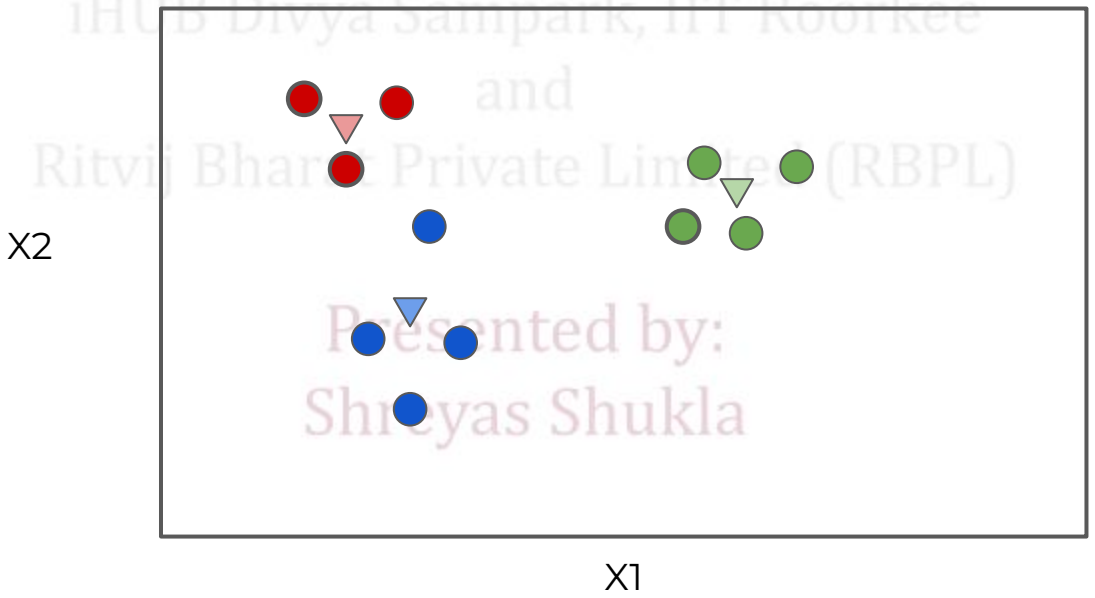


fig ref: pierian data

If there are no more reassignments, we're done! The clusters have been found.

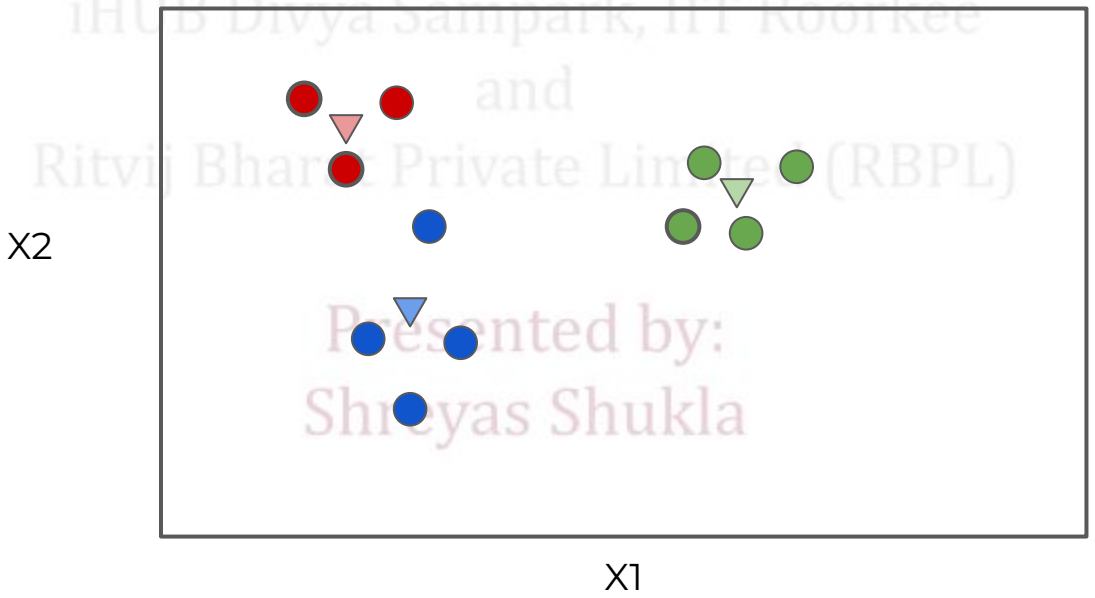


fig ref: pierian data

Upcoming considerations:

- How do we choose a reasonable value for K number of clusters?
- Is there any way we can evaluate how good our current K value is at determining clusters?

Let’s code out an example of K-Means clustering,

then we’ll revisit these considerations when they naturally appear after we find the first set of clusters for a given K.

An Introduction to Machine Learning with Python Programming
11 Sep 2023 - 20 Oct 2023

Conducted by:
iHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited (RBPL)

Let’s Code !!

Presented by:
Shreyas Shukla

fig ref: pierian data