Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Unsupervised Learning

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Let’s learn about Unsupervised Learning!

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

***Supervised performance metrics will not apply for unsupervised learning!***

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Machine Learning Pathway for Unsupervised Learning

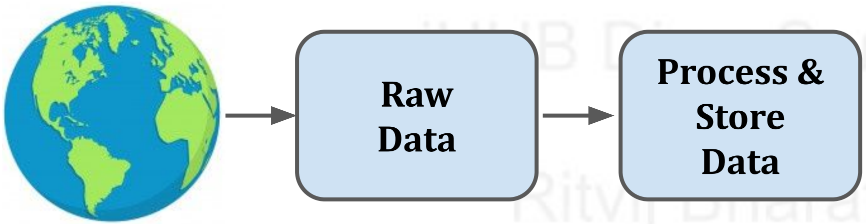
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

IHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla



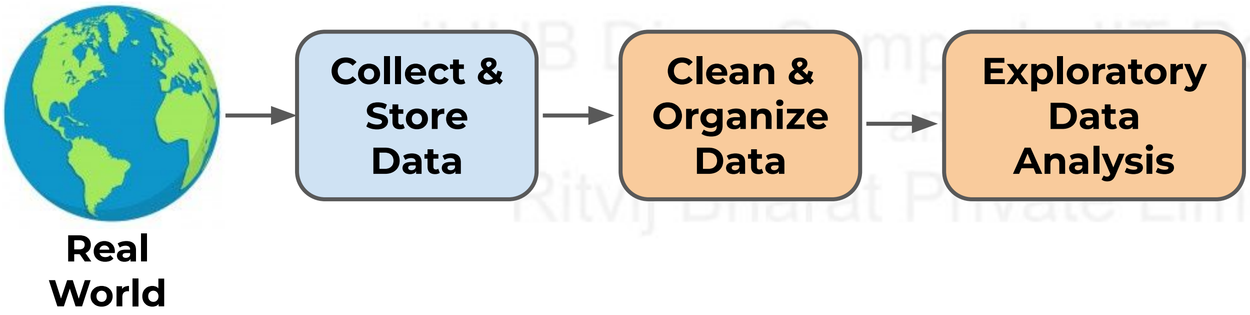
Real World

Raw Data

Process & Store Data

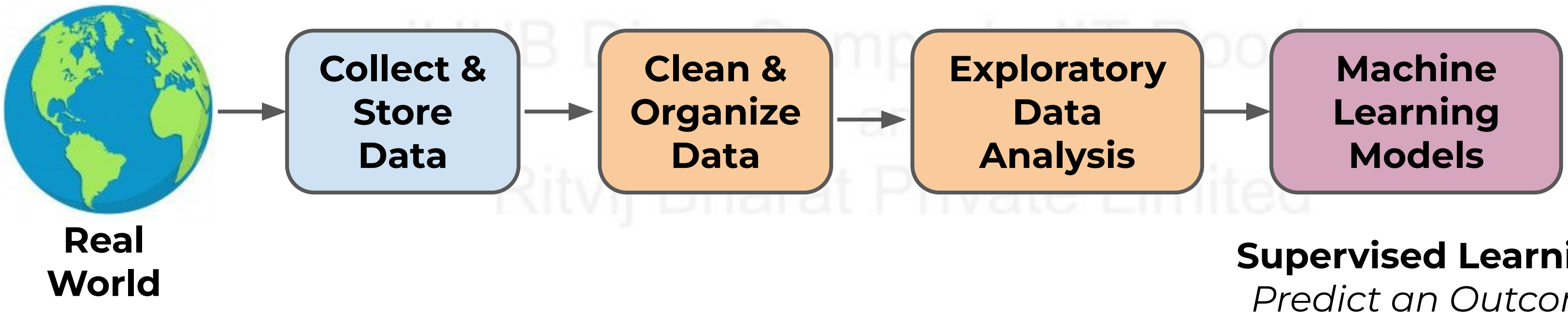
SQL Database, CSV files, Excel, Cloud Storage

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB Divya Sampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



Real World → Collect & Store Data → Clean & Organize Data → Exploratory Data Analysis

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

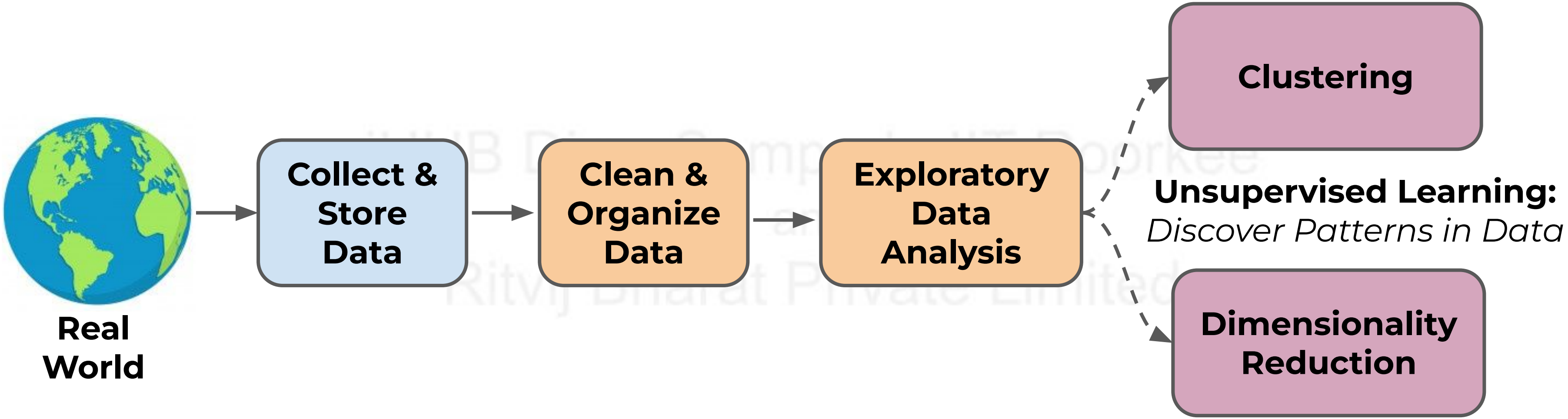


IHUB Divya Sampark, IIT Roorkee and Ritvij Bharat Private Limited

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

Led by : Shreyas Shukla

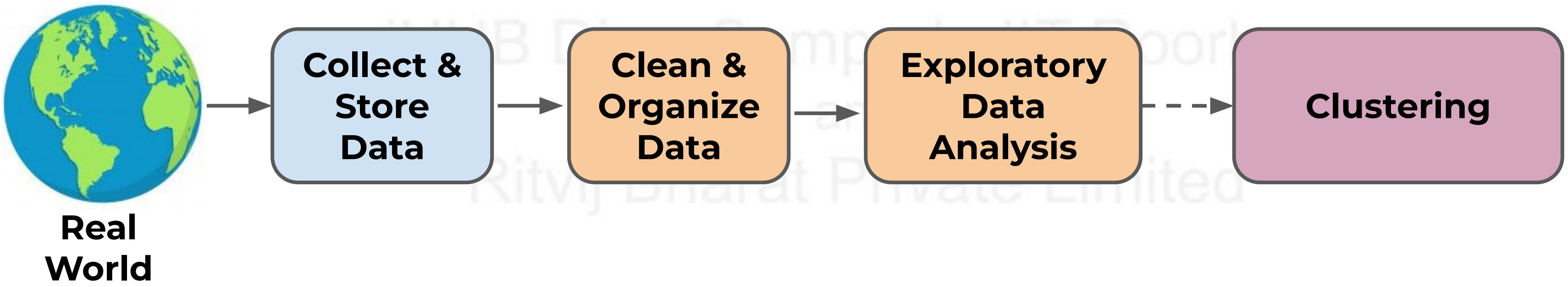
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB DivyaSampark IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



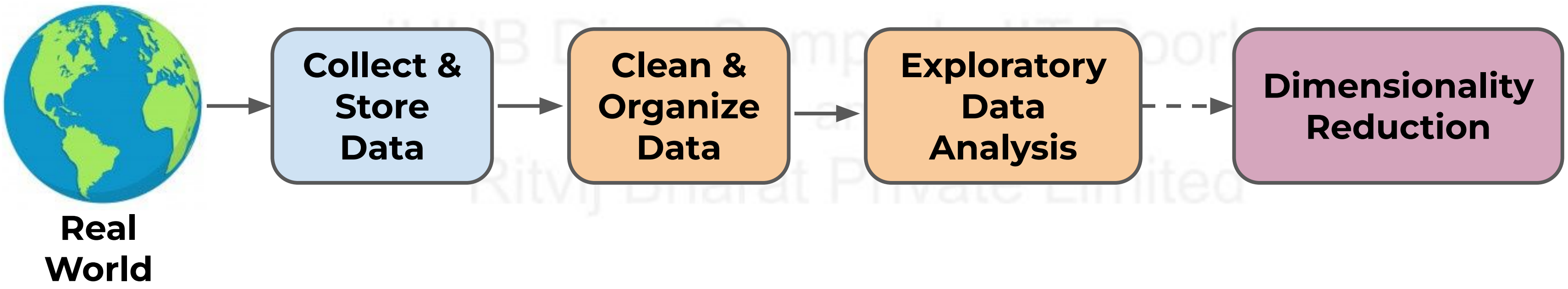
**Unsupervised Learning:**
*Discover Patterns in Data*

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB Divya Sampark IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



**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?



Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)
IHUB Divya Sampark IIT Roorkee and Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Methods for interpreting the model results

Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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?*

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

# Let’s get started!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

# K-Means Clustering

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

***Do not confuse K-Means with KNN!***

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

# General Concepts

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Imagine an example data set:

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

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

Led by : Shreyas Shukla

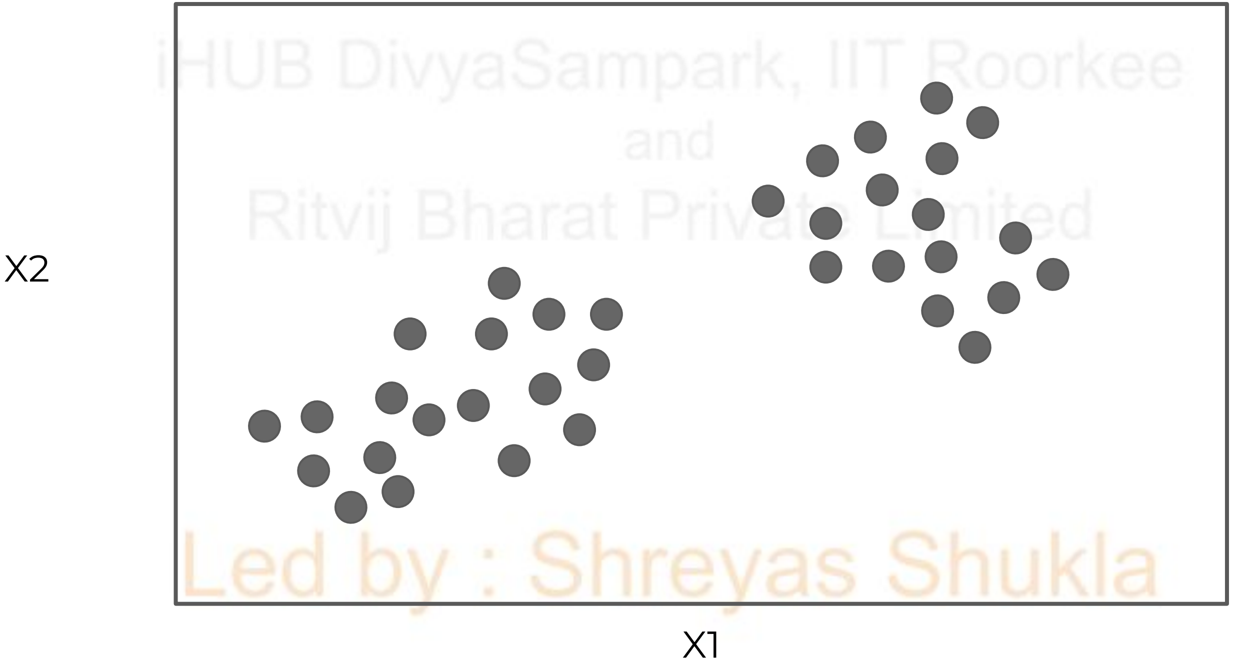
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

# How could we cluster this data together?

| X1 | X2 |
|---|---|
| 2 | 4 |
| 6 | 3 |
| … | … |
| 1 | 2 |

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

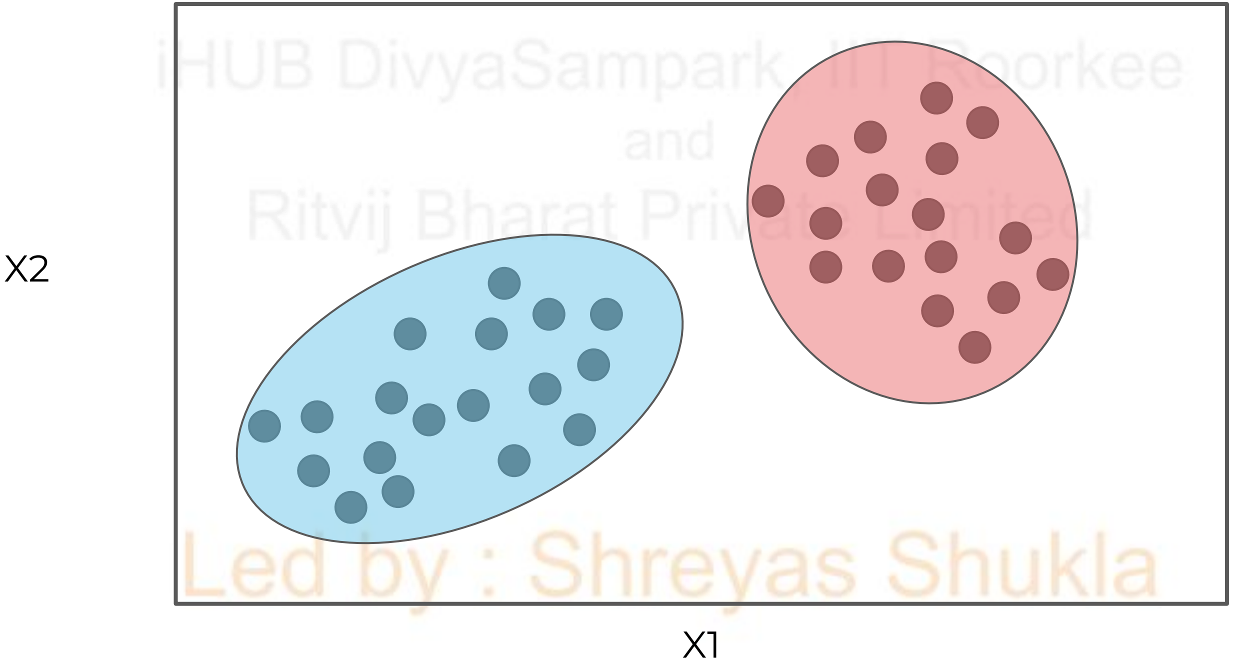
# Intuitively, we see 2 groupings:



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Note how distance is the intuitive metric:

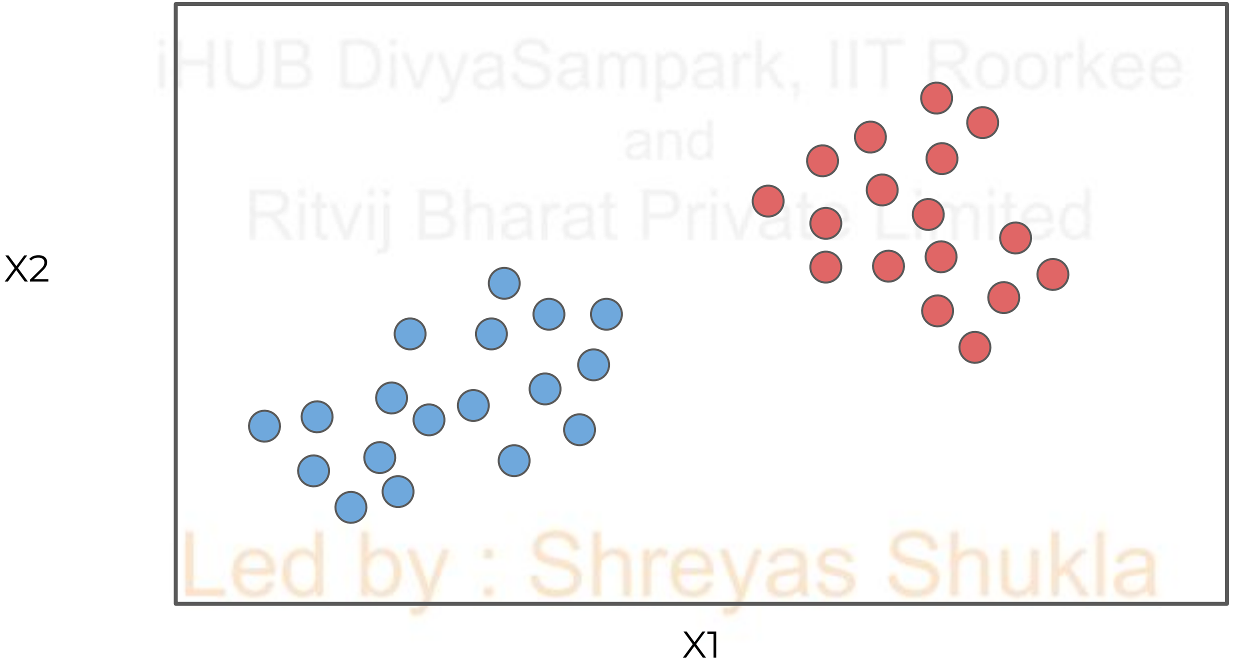


iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Assign clusters:

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited



Led by : Shreyas Shukla

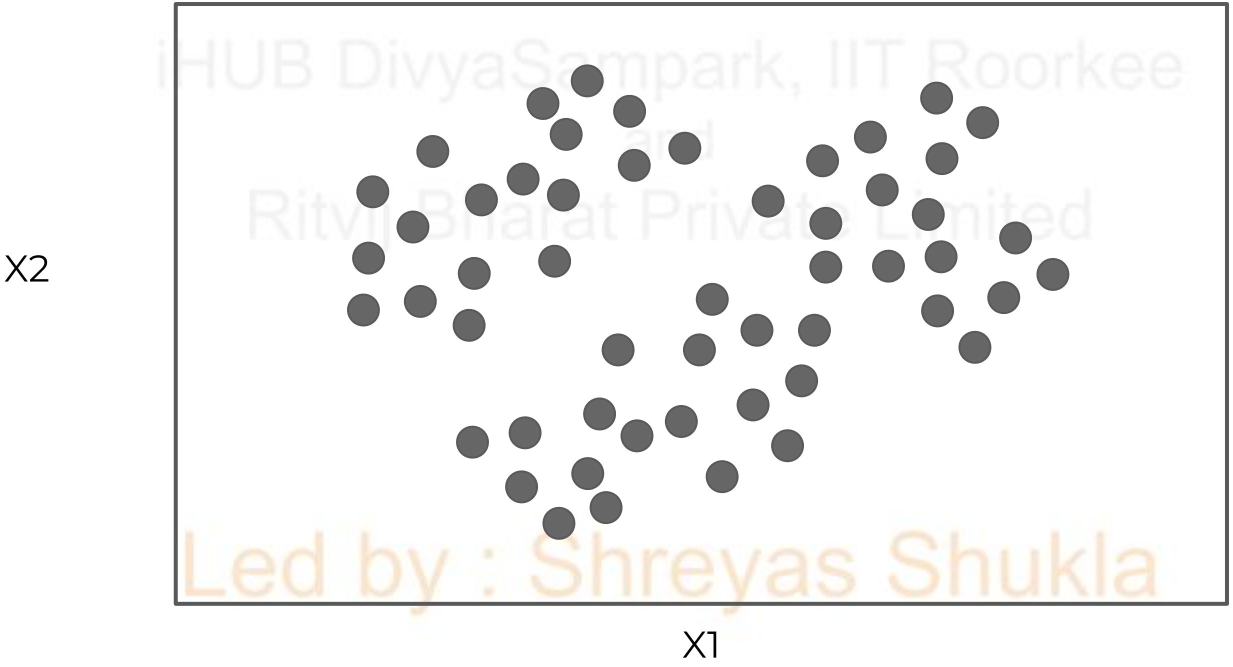
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
IHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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?

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

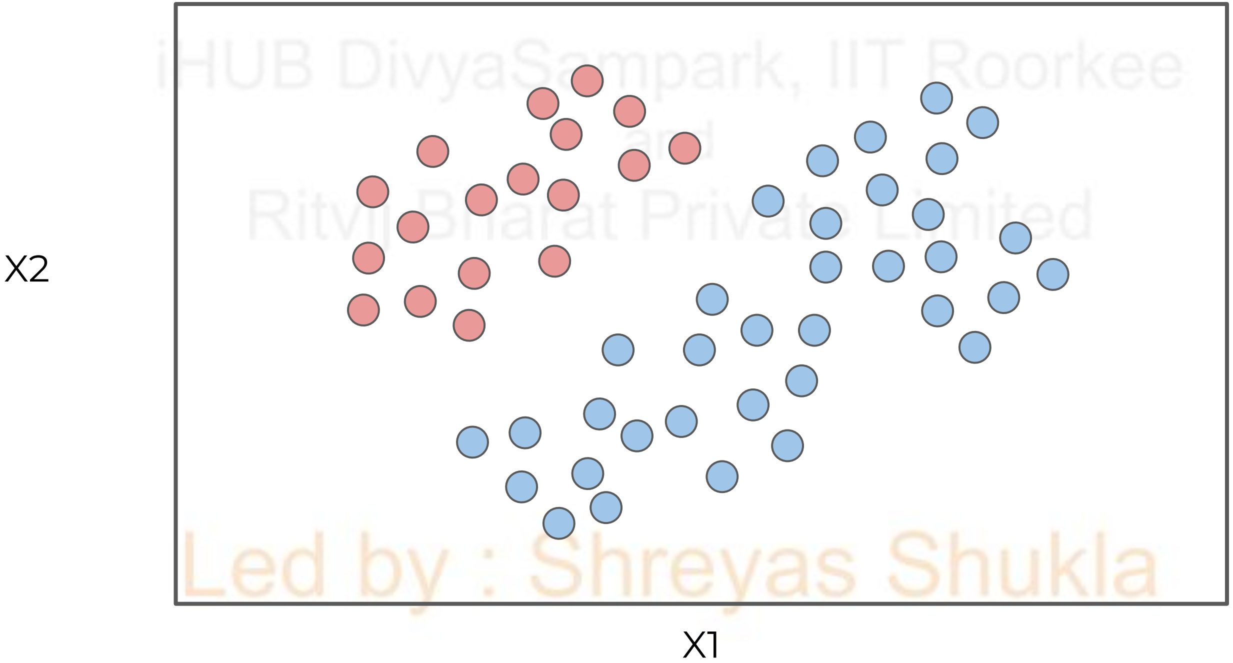
# 2 or 3 clusters could both be reasonable:



iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

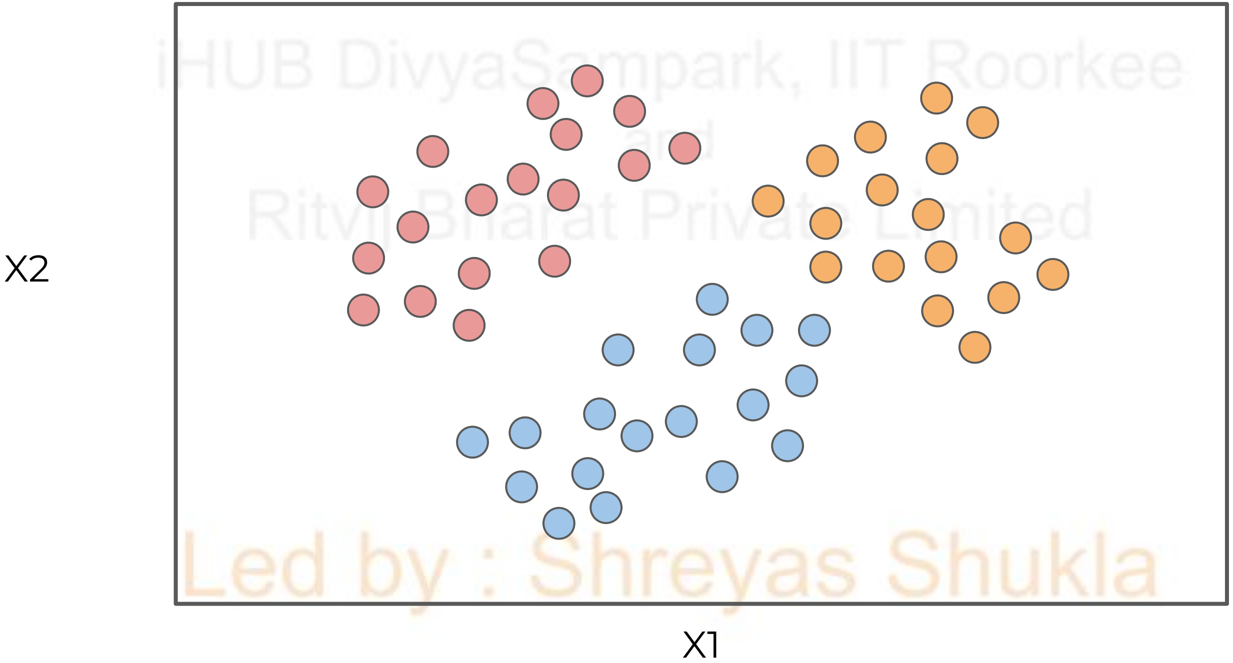
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# 2 or 3 clusters could both be reasonable:



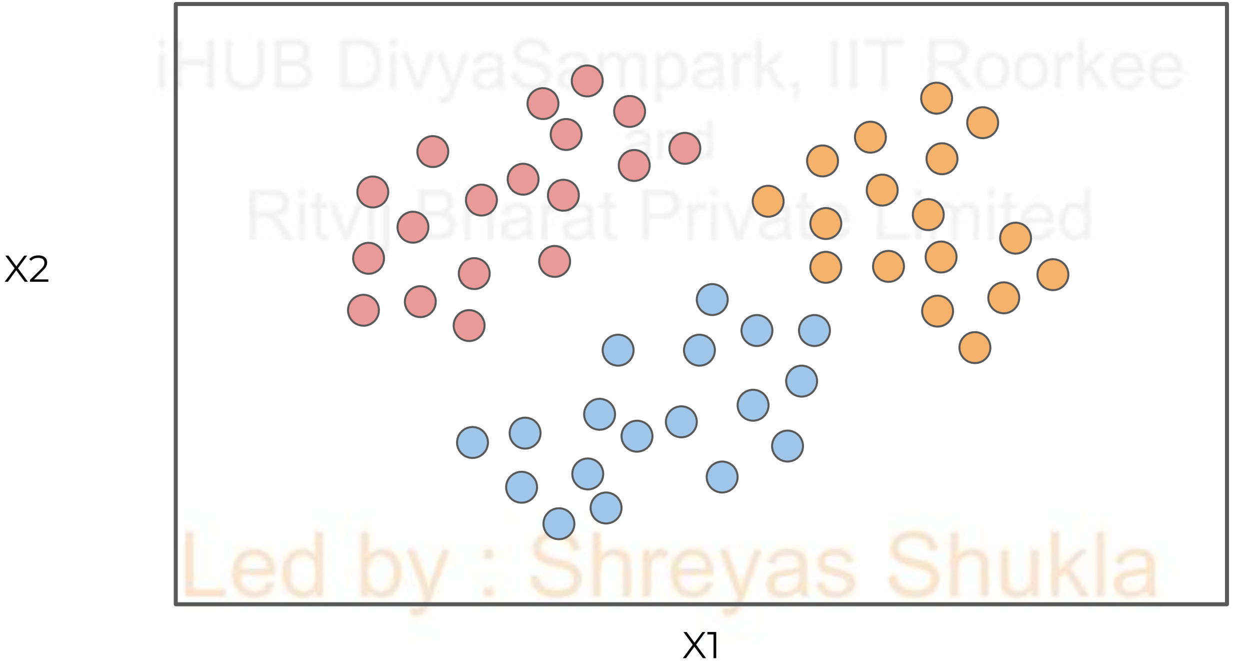
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



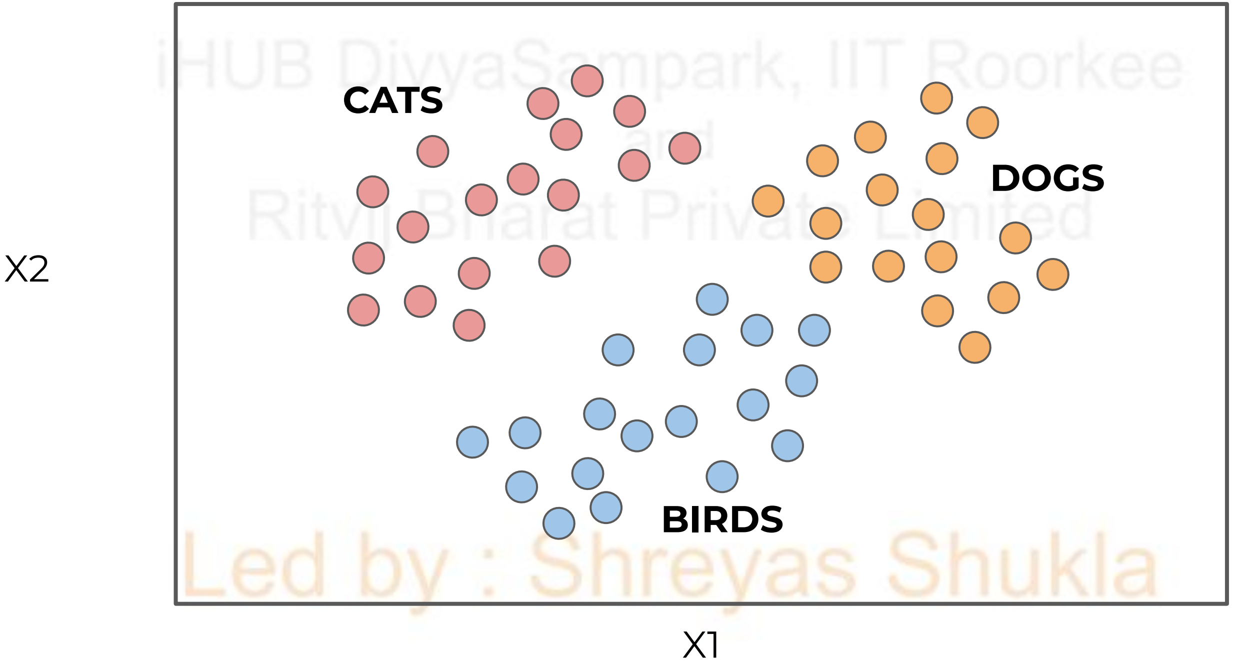
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

# Different methods can be used to decide!



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Clustering doesn’t “label” these for you!

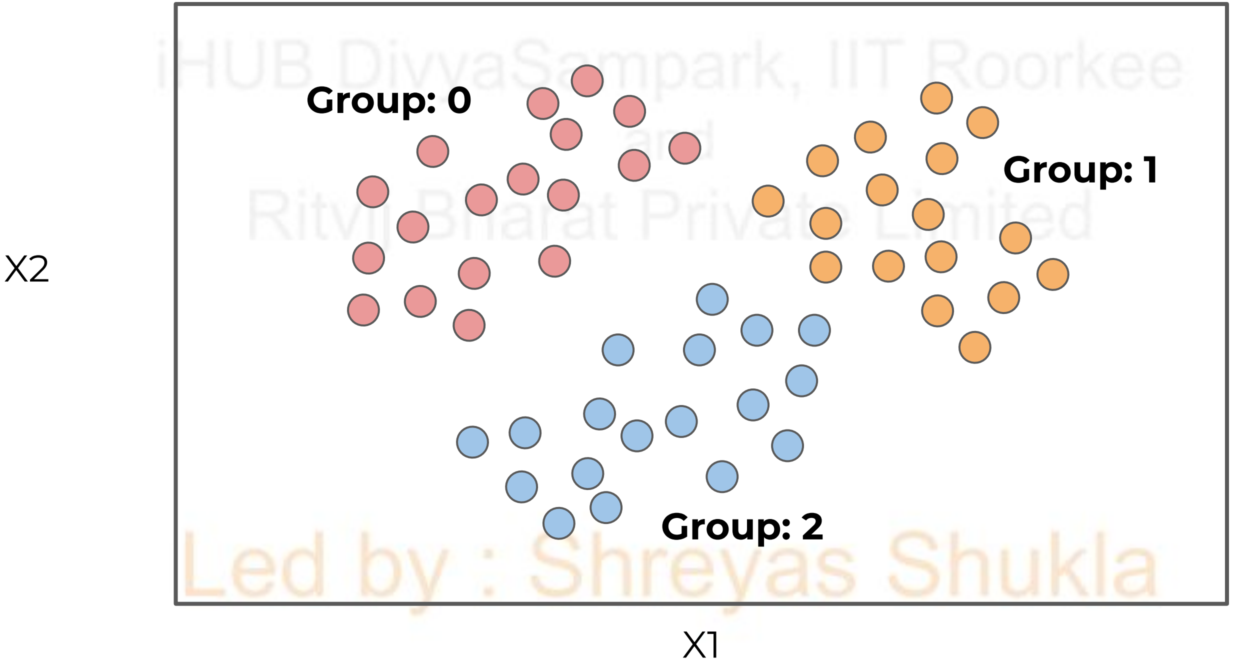


iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

○ *What about a new unlabeled data point?*

○ *How do we assign it to a cluster?*

○ *Was it the correct cluster for assignment?*

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

- *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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
vij Bharat Private Limited
Led by : Shreyas Shukla

- ○ *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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee,
and
vij Bharat Private Limited
Led by : Shreyas Shukla

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

*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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

- 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?

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

# Machine Learning as an art

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

# K-Means Clustering

Intuition and Theory

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

# Working of K-Means clustering

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

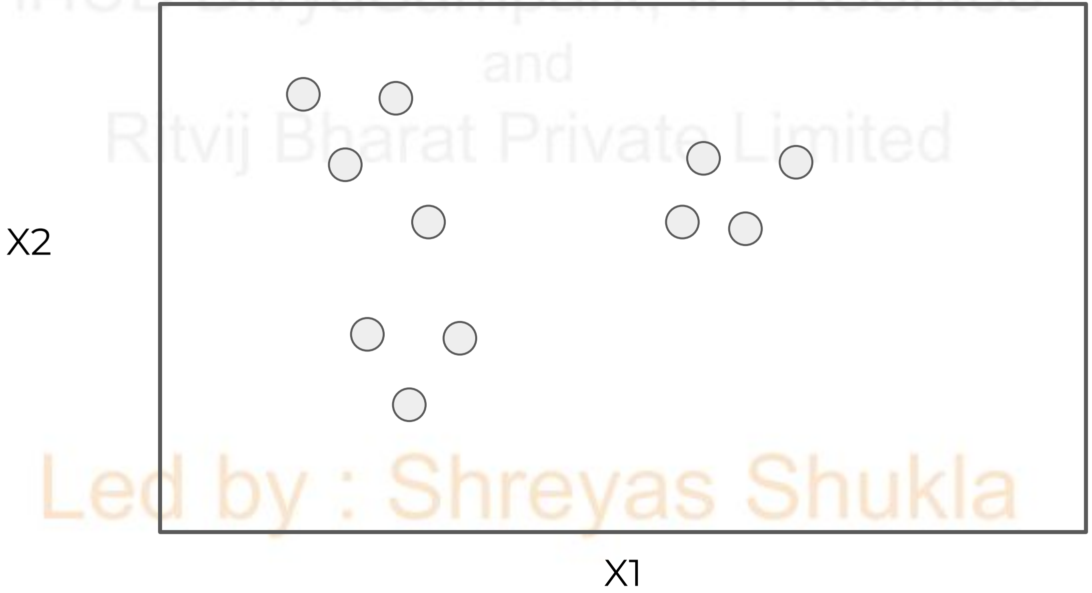
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

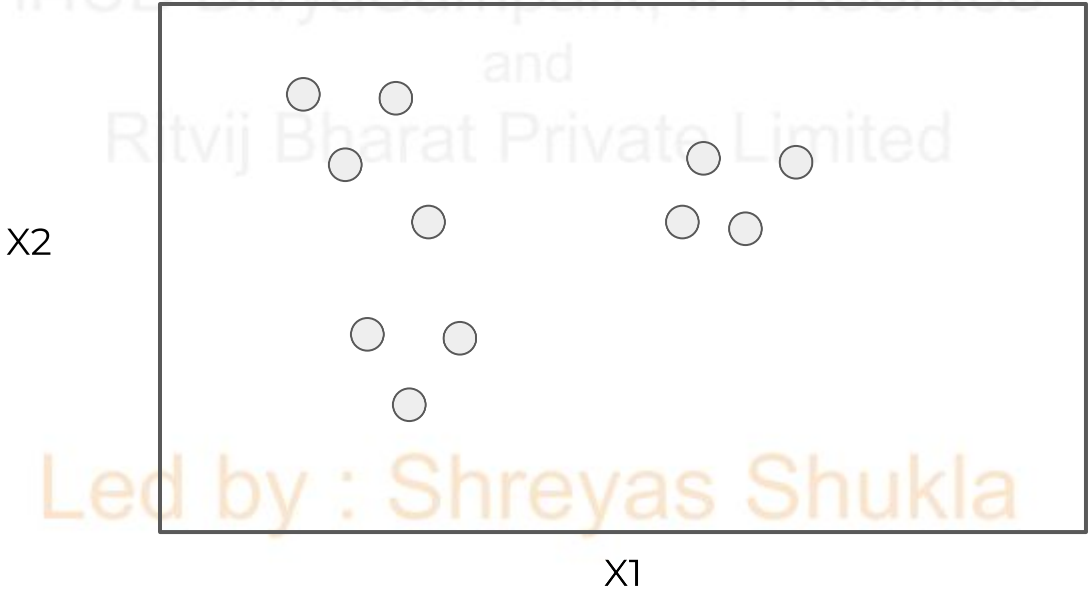
- Step 0: Start with unlabeled data (only features).



Mastering Machine Learning with Python (27th Aug 2024 - 16th Oct 2024) iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited Led by : Shreyas Shukla

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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

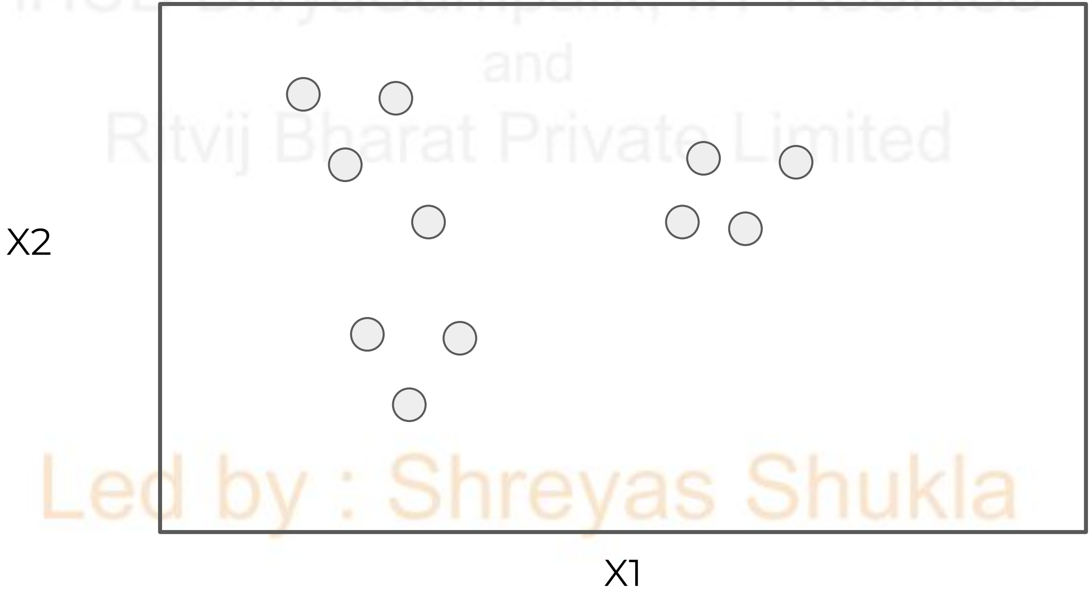


# Step 1: Choose the number of clusters to create (this is the K value).

Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)

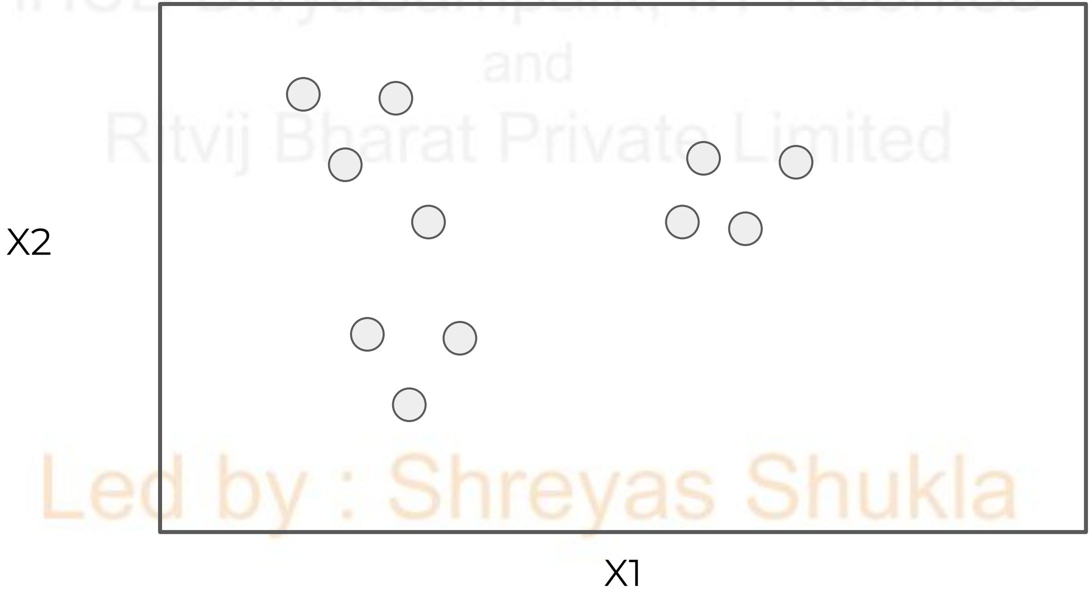
iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited

Led by : Shreyas Shukla



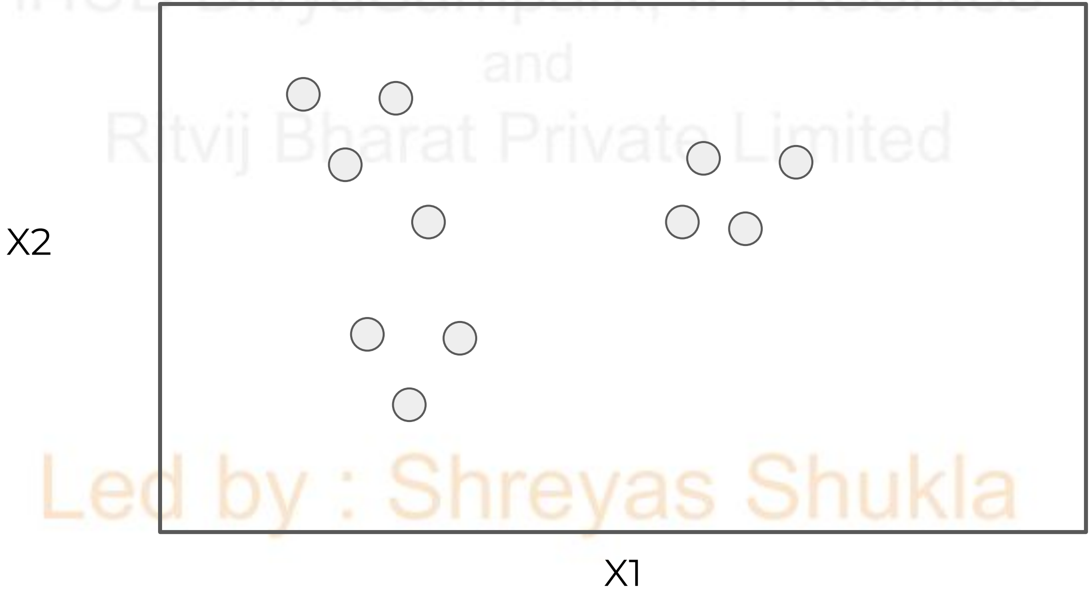
# Step 1: We'll choose K=3. Note in most situations you won't visualize the data!

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



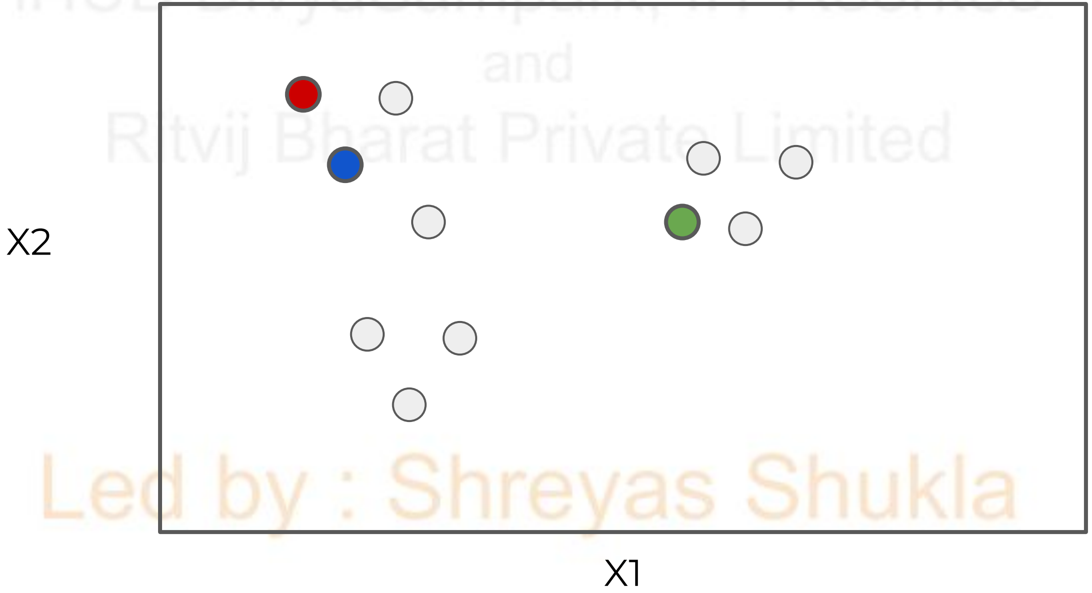
# Step 2: Randomly select K distinct data points.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



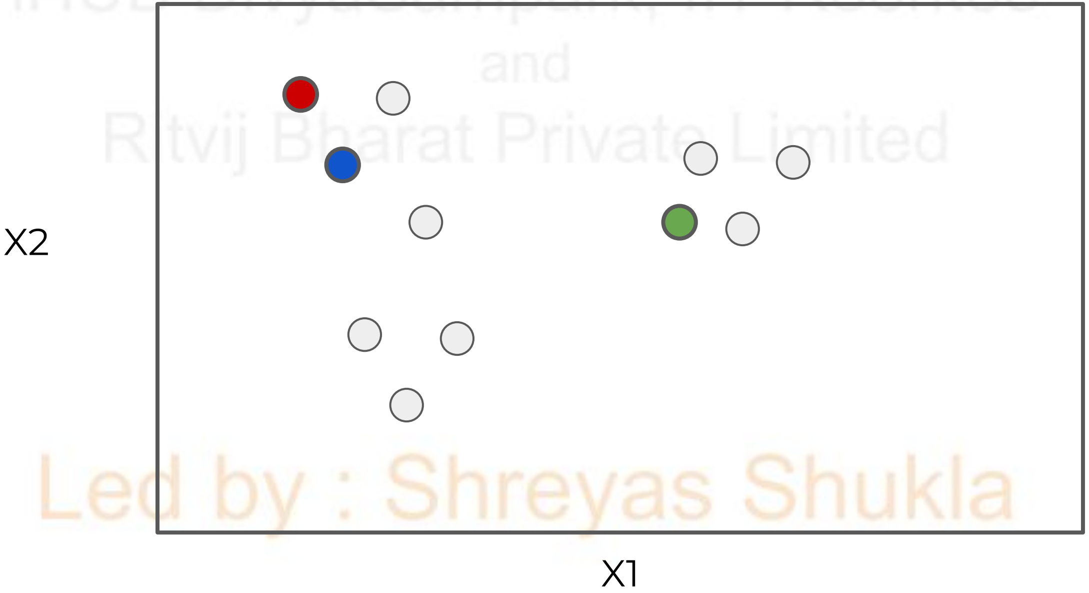
Step 2: Randomly select K distinct data points. Our K=3. We'll treat these new K points as our cluster points.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



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

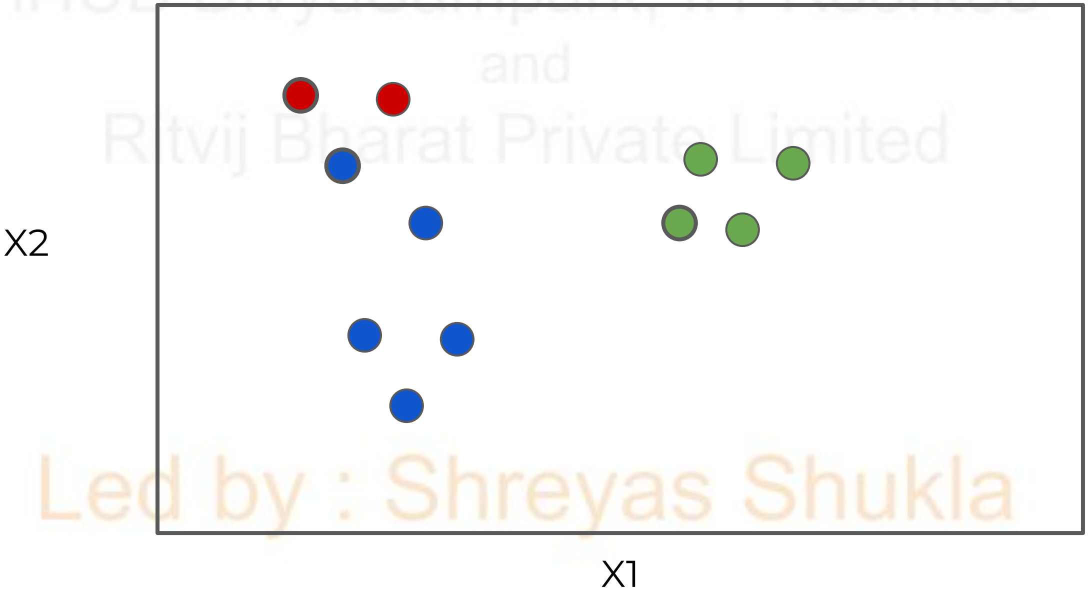
Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

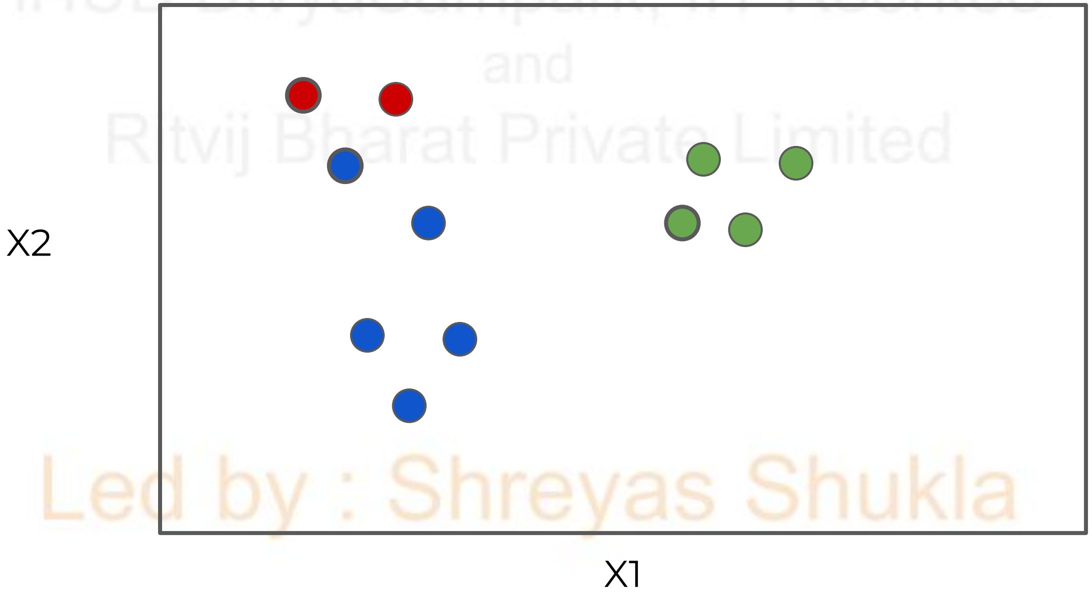
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla



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

Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

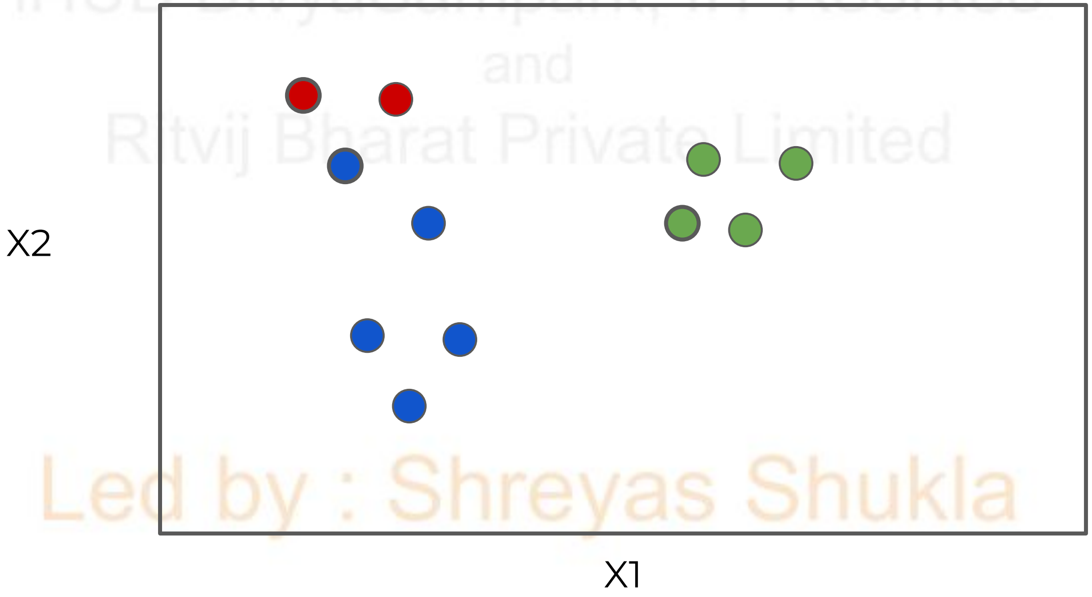


# Step 4: Calculate the center of the cluster points (mean value of each point vector).

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

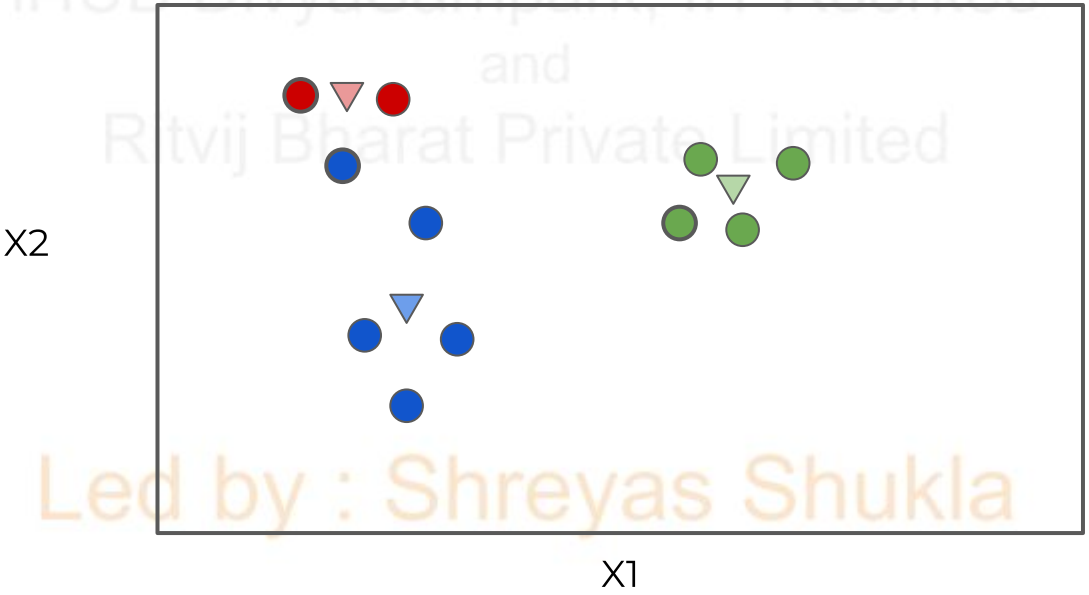
Led by : Shreyas Shukla



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

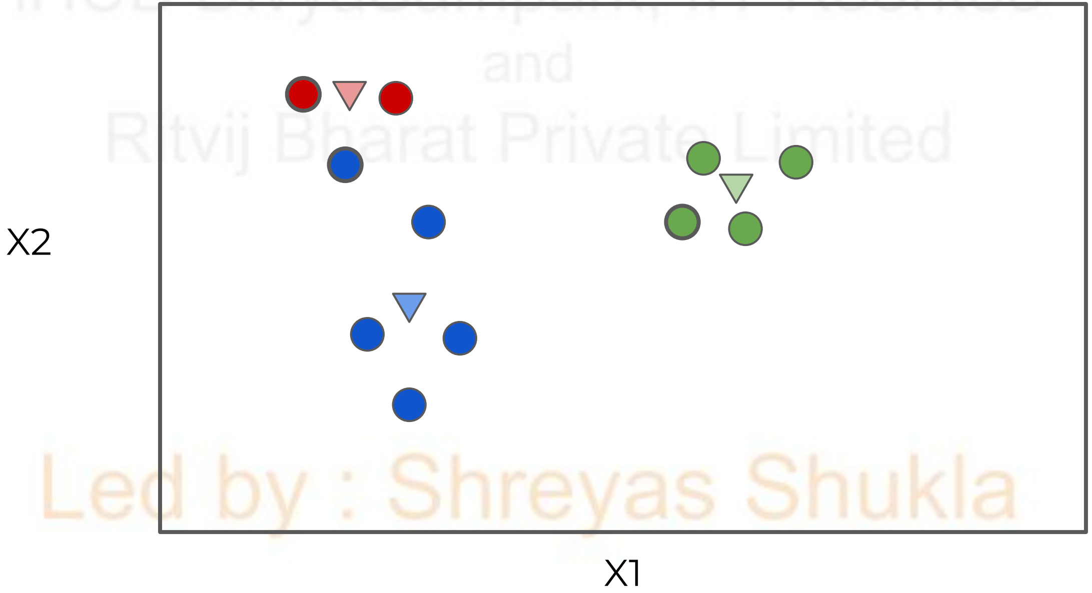
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

Led by : Shreyas Shukla

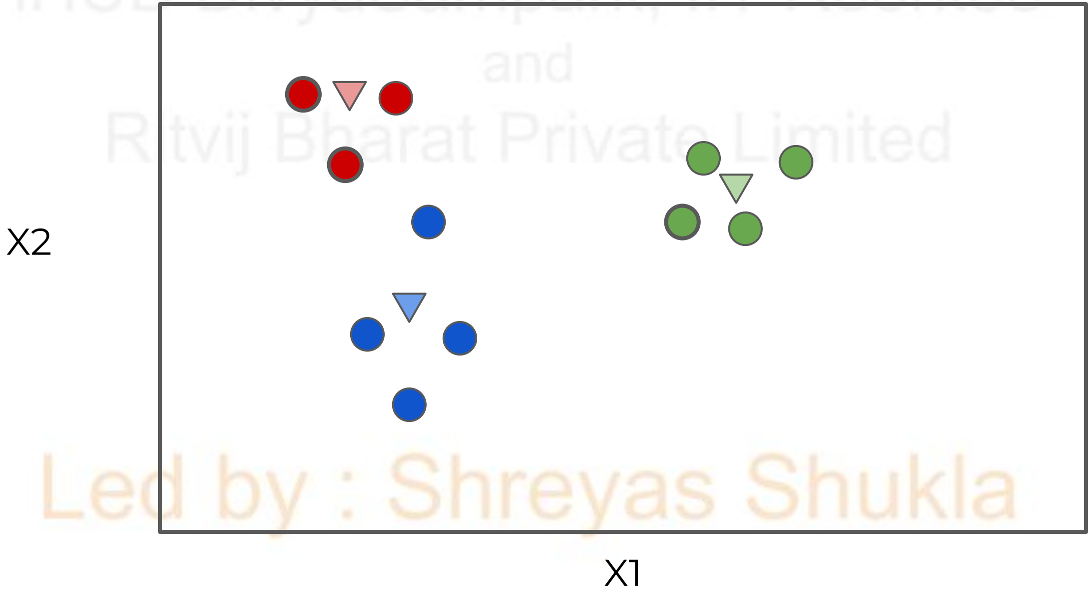


# Step 5: Now assign each point to the nearest cluster center.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

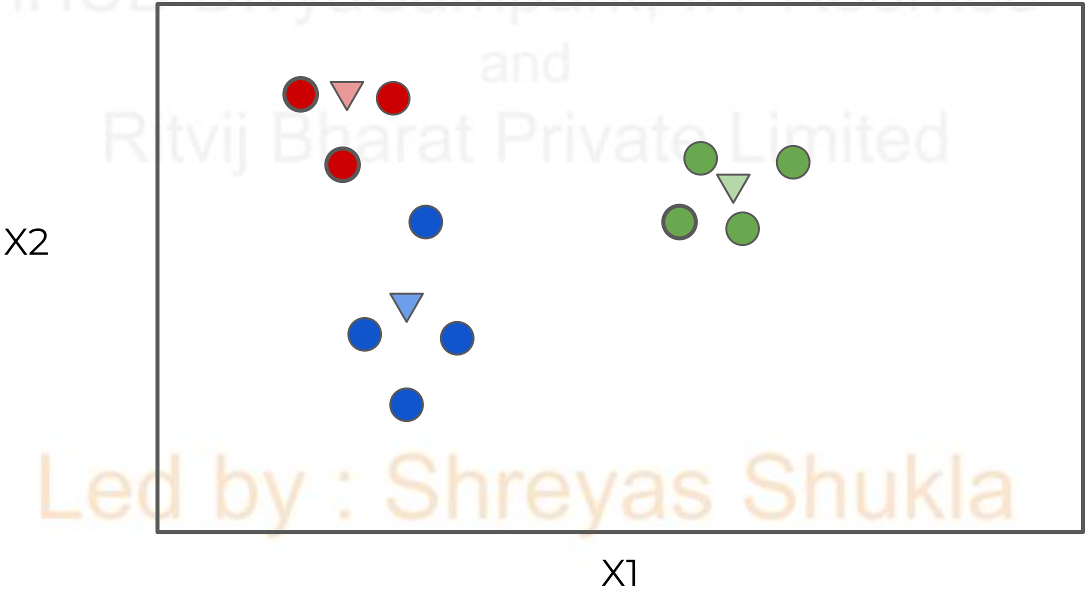


Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

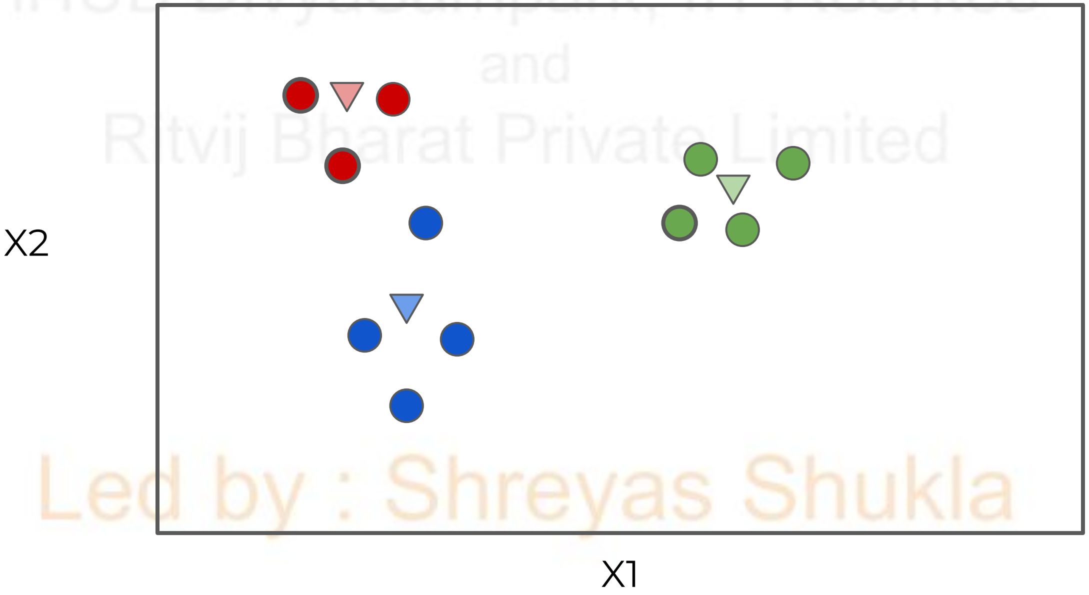


# Step 4b: Recalculate new cluster centers:

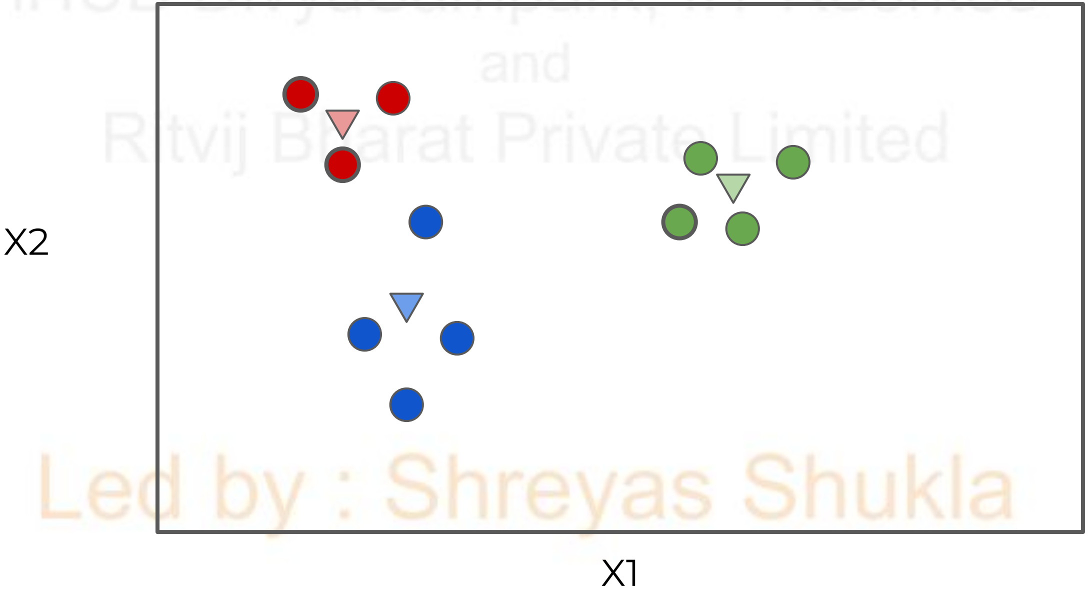
Mastering Machine Learning with Python (27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee and Ritvij Bharat Private Limited

Led by : Shreyas Shukla

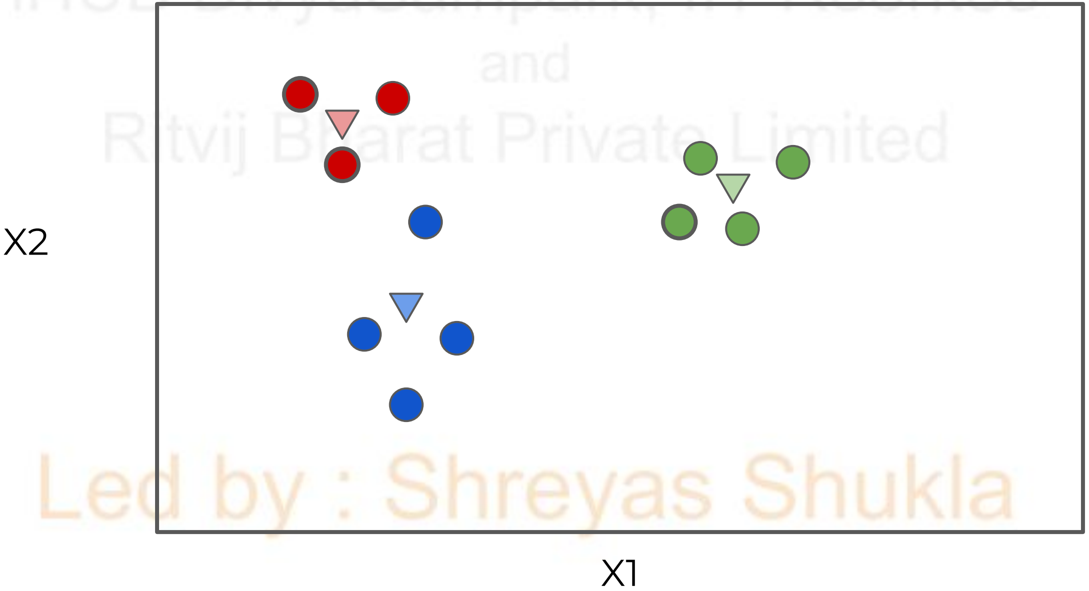


Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



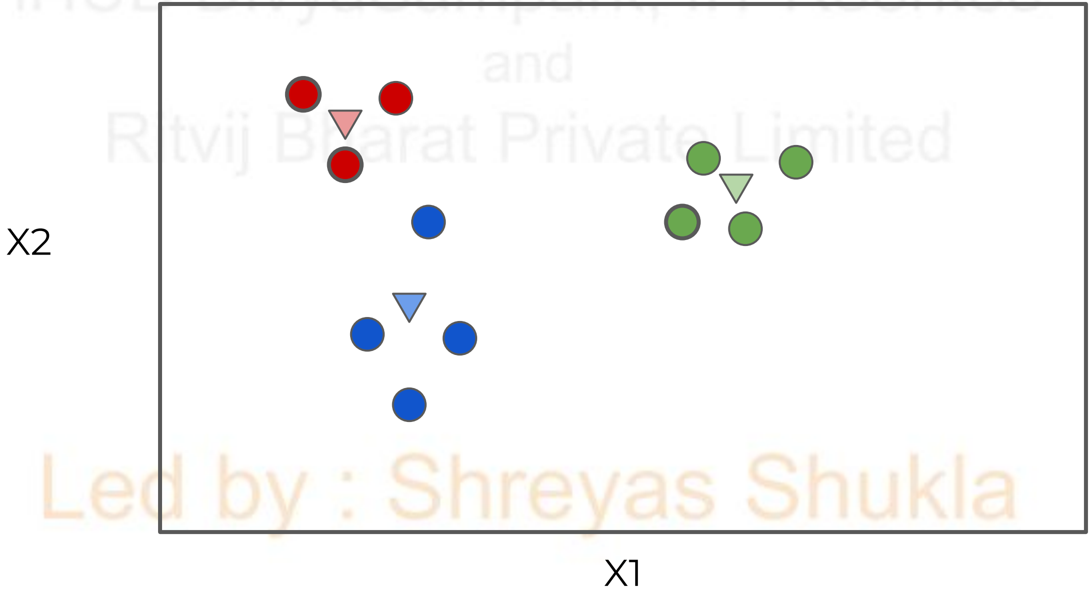
# Step 5b: Assign points to nearest cluster center.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



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

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla



Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

**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?

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited
Led by : Shreyas Shukla

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.

Mastering Machine Learning with Python
(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

# Let’s Code !!

Led by : Shreyas Shukla