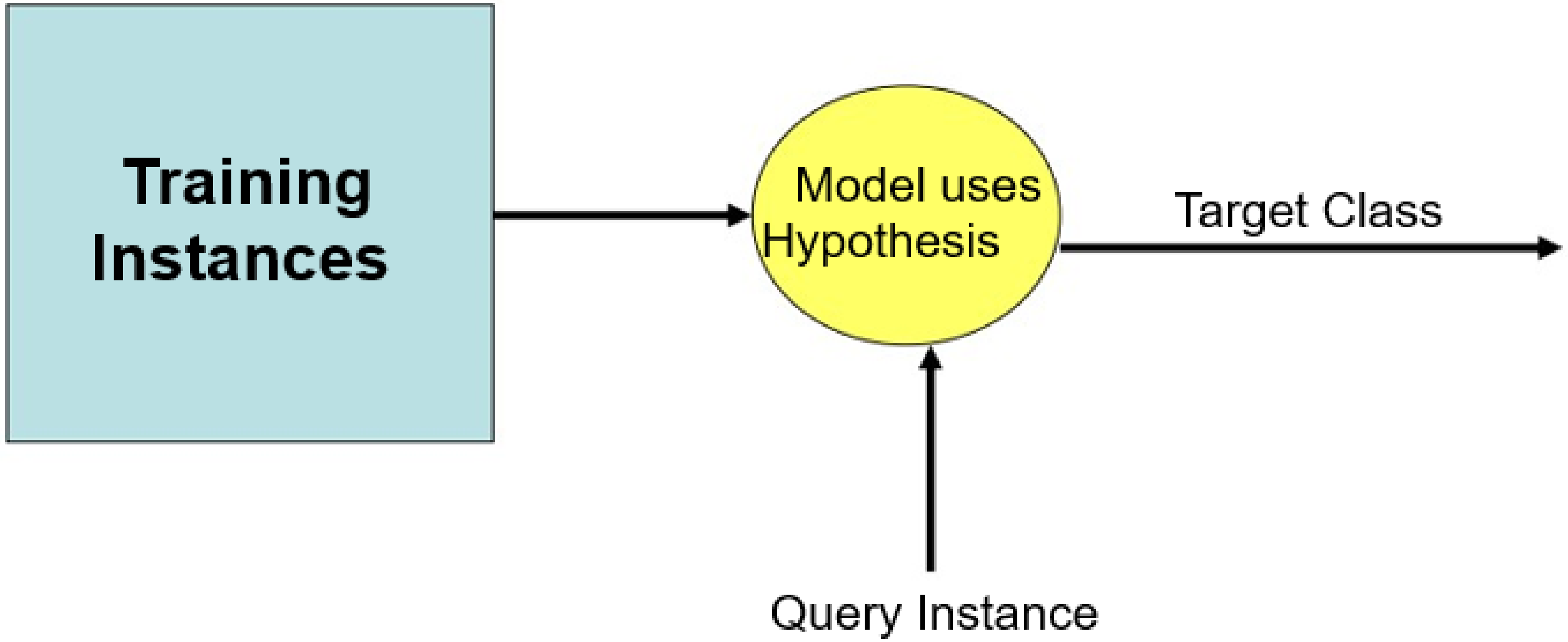
Instance Based Learning

AIML/ BDA

Topics covered

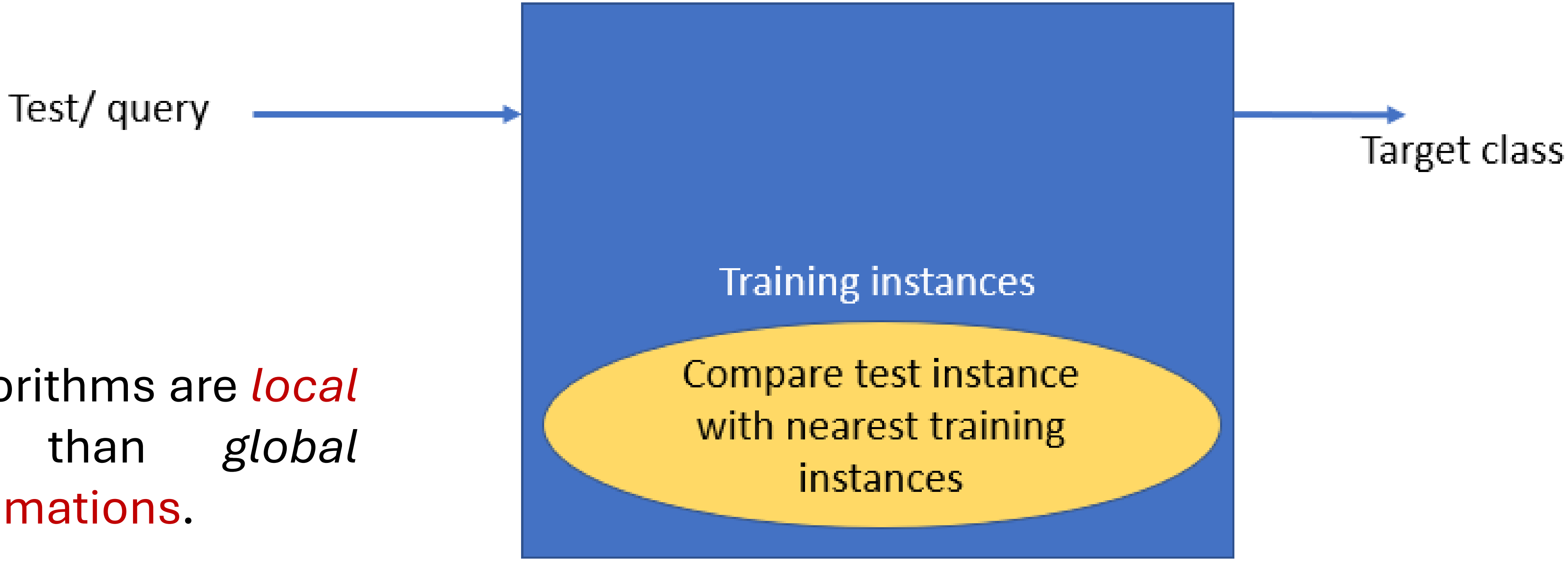
- K-Nearest Neighbors (K-NN) concept
- Distance metrics
- K-NN for classification

Machine Learning Classification



Instances Based Learning

- Training instances are stored in memory
- For a test (unseen) instances
- Compare test instances with instances seen in training and gives result
- Also known as Memory based learning



Test/ query

Target class

Training instances

Compare test instance with nearest training instances

IBL algorithms are *local* rather than *global* approximations.

Instances Based Learning

- IBL methods learn by storing the training data.

- When a new query instance is encountered, a set of similar related instances is retrieved from memory and used to classify the new query instance.

- For each distinct query, IBL construct different approximation to the target function.
- These are *local* rather than *global* approximations.

Advantages and Disadvantages of IBL

Advantage:

- Suitable for problems with very complex target functions.

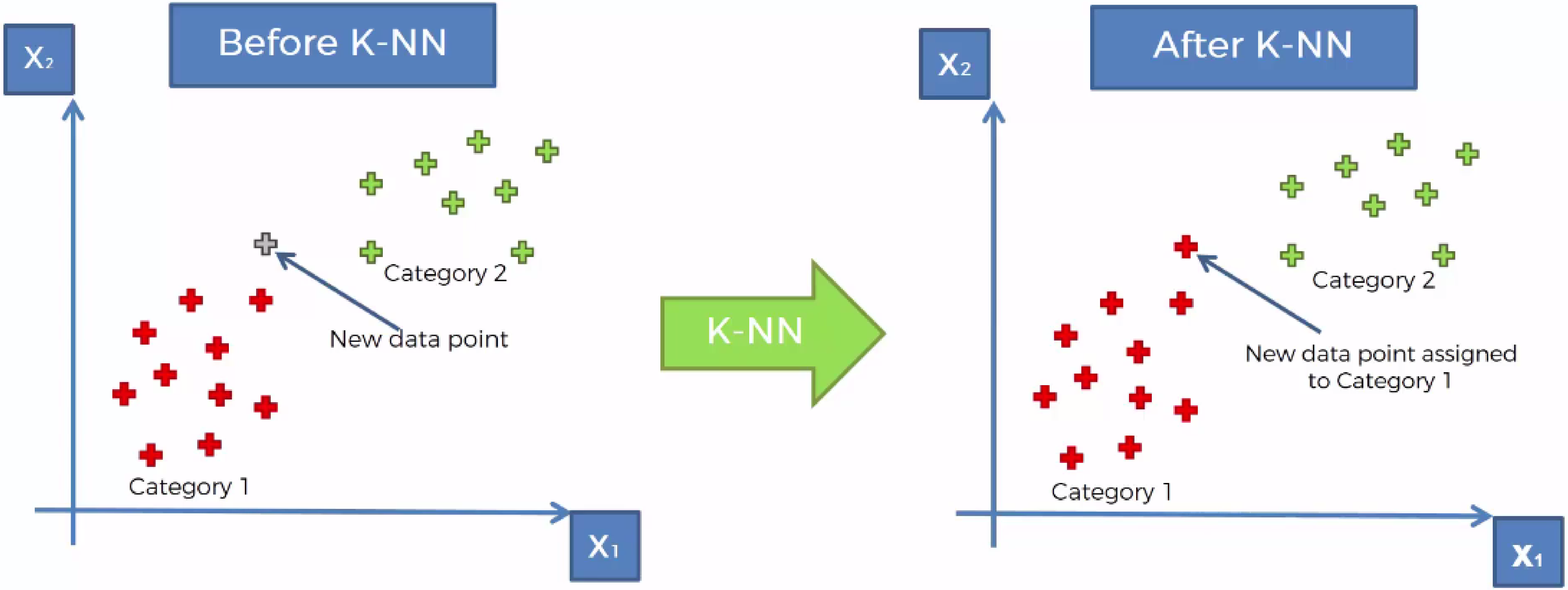
Disadvantage:

- The cost of classifying new instances - high.
- Considered all attributes of the instances – dimension increase

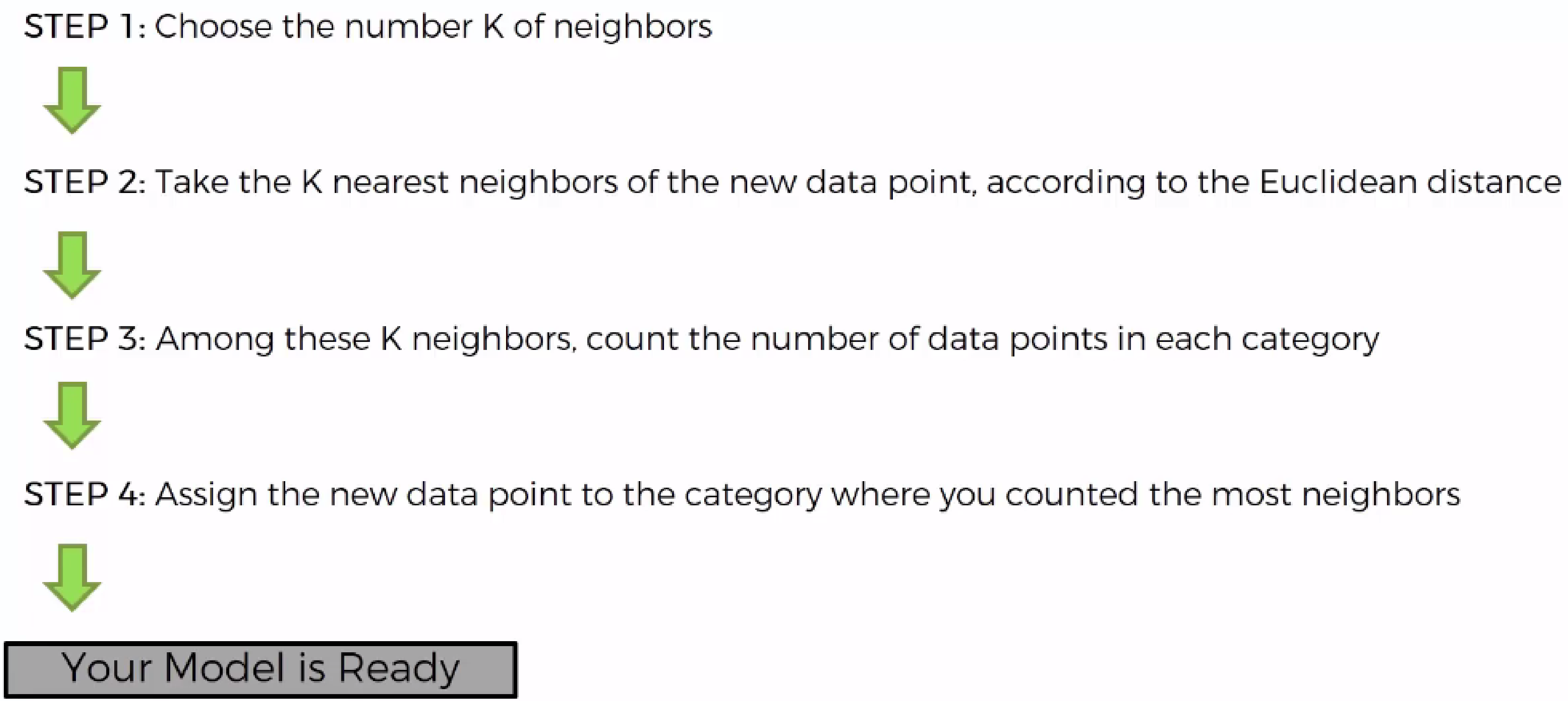
Comparison

	Instance-based learning	Other learnings
In Memory	**Training Instances**	Model / Hypothesis
Hypothesis	Every time a new hypothesis is generated	Hypothesis is same for all future examples

What K-NN does for you



How did it do that ?



STEP 1: Choose the number K of neighbors

STEP 2: Take the K nearest neighbors of the new data point, according to the Euclidean distance

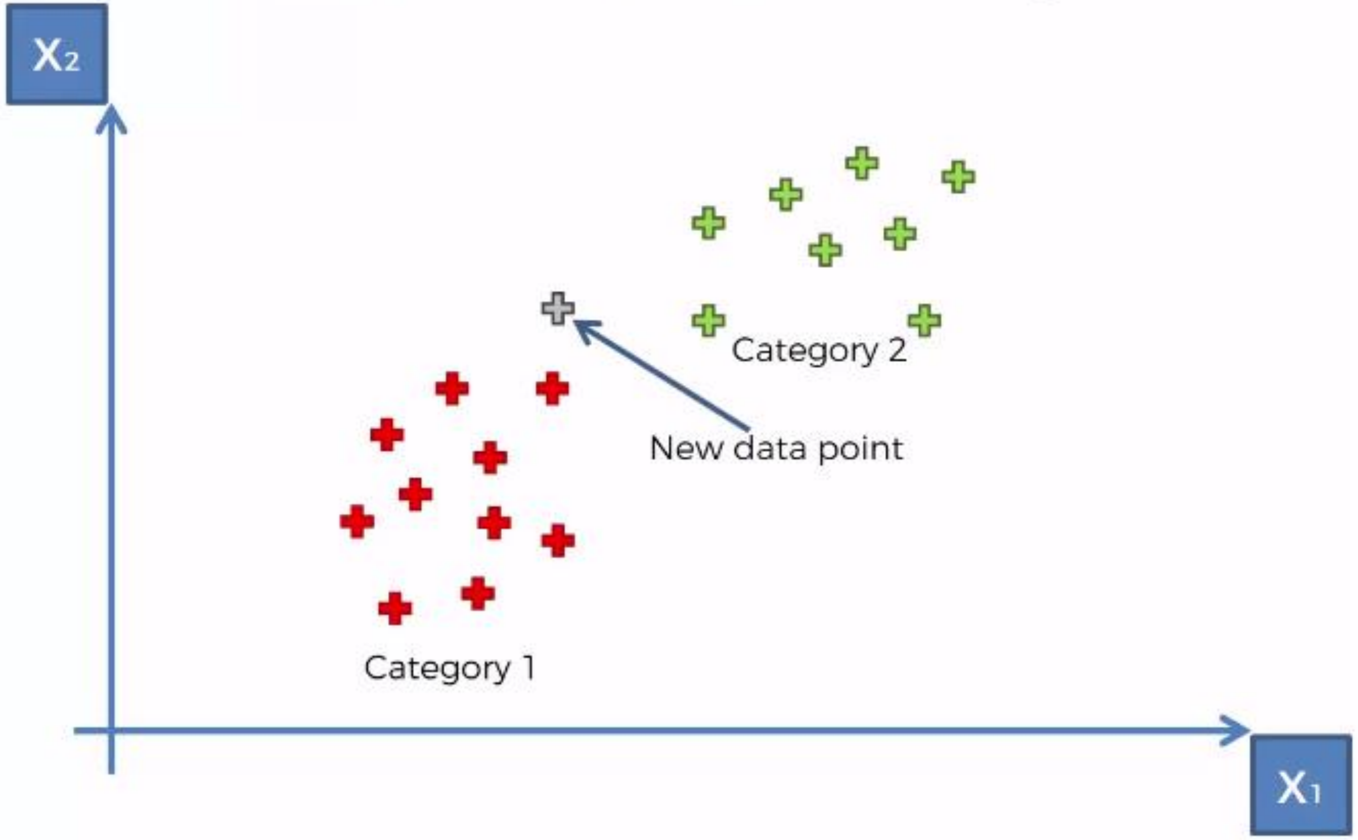
STEP 3: Among these K neighbors, count the number of data points in each category

STEP 4: Assign the new data point to the category where you counted the most neighbors

Your Model is Ready

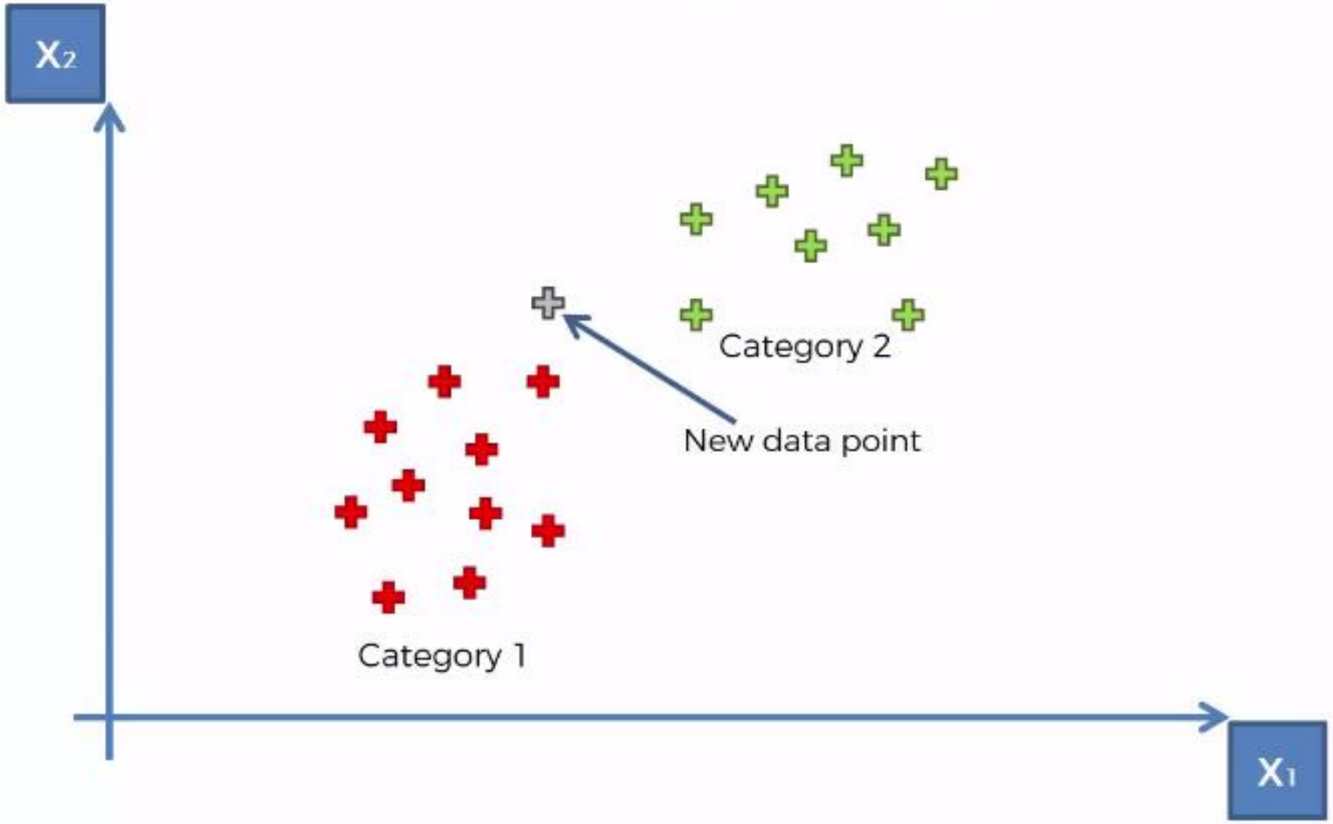
K-NN algorithm

STEP 1: Choose the number K of neighbors: K = 5



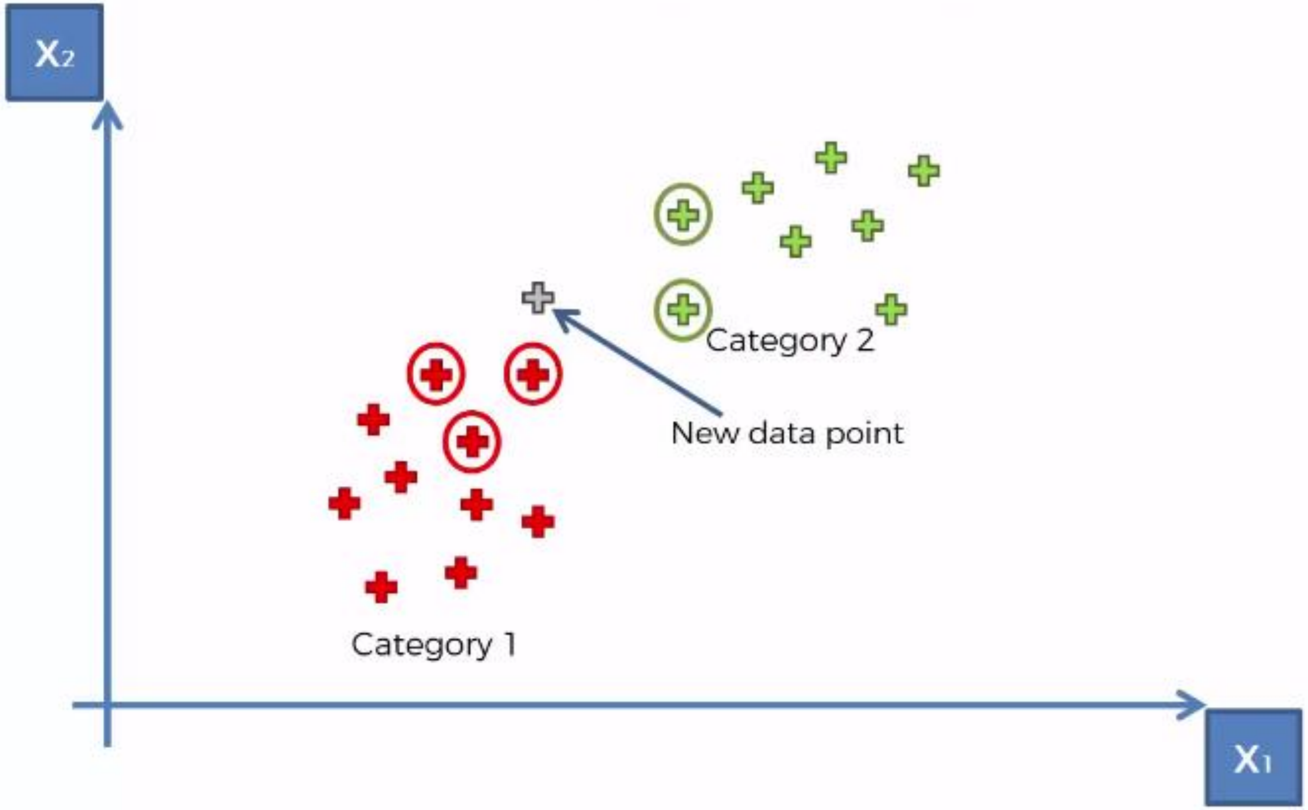
K-NN algorithm

STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance



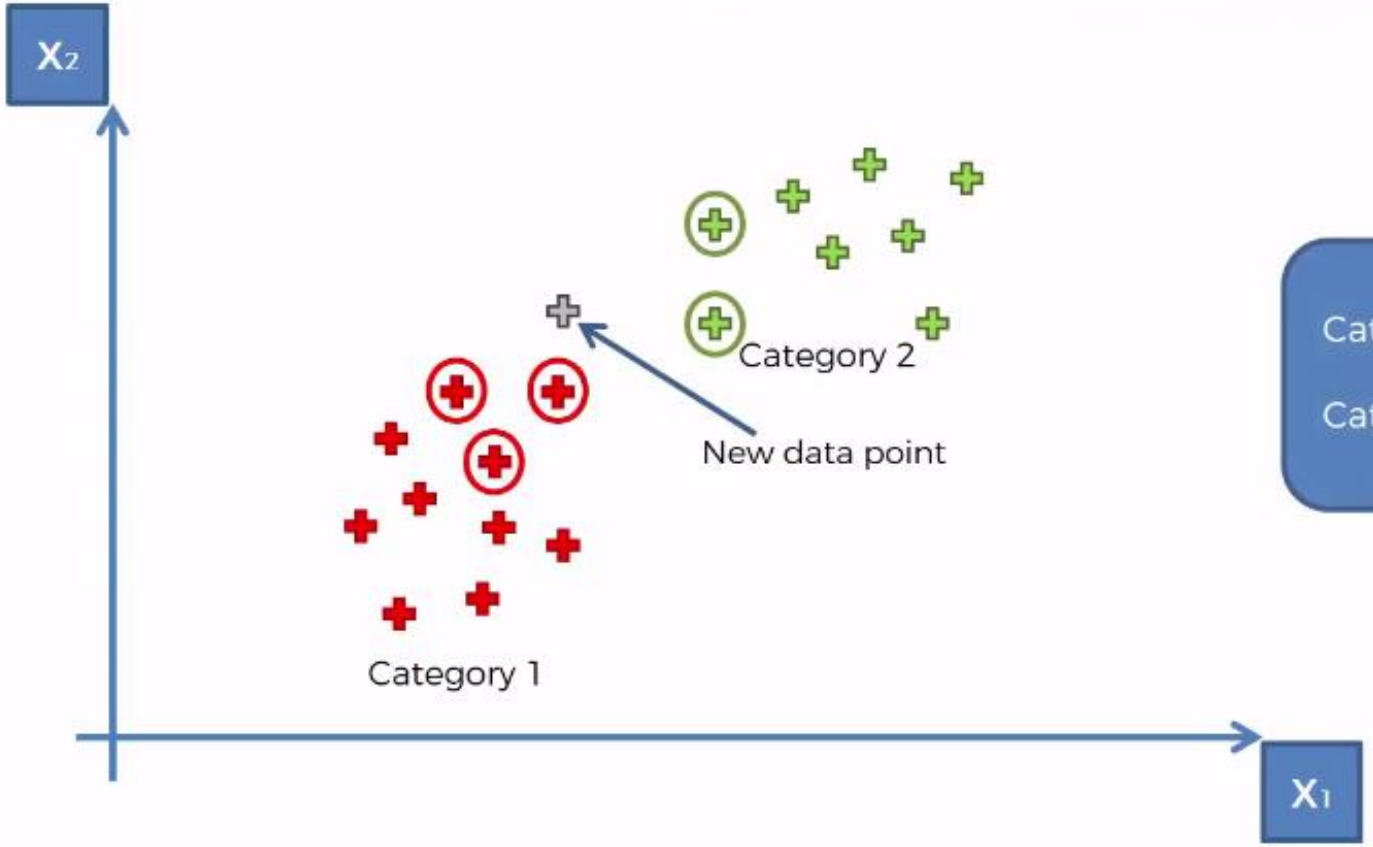
K-NN algorithm

STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance



K-NN algorithm

STEP 3: Among these K neighbors, count the number of data points in each category

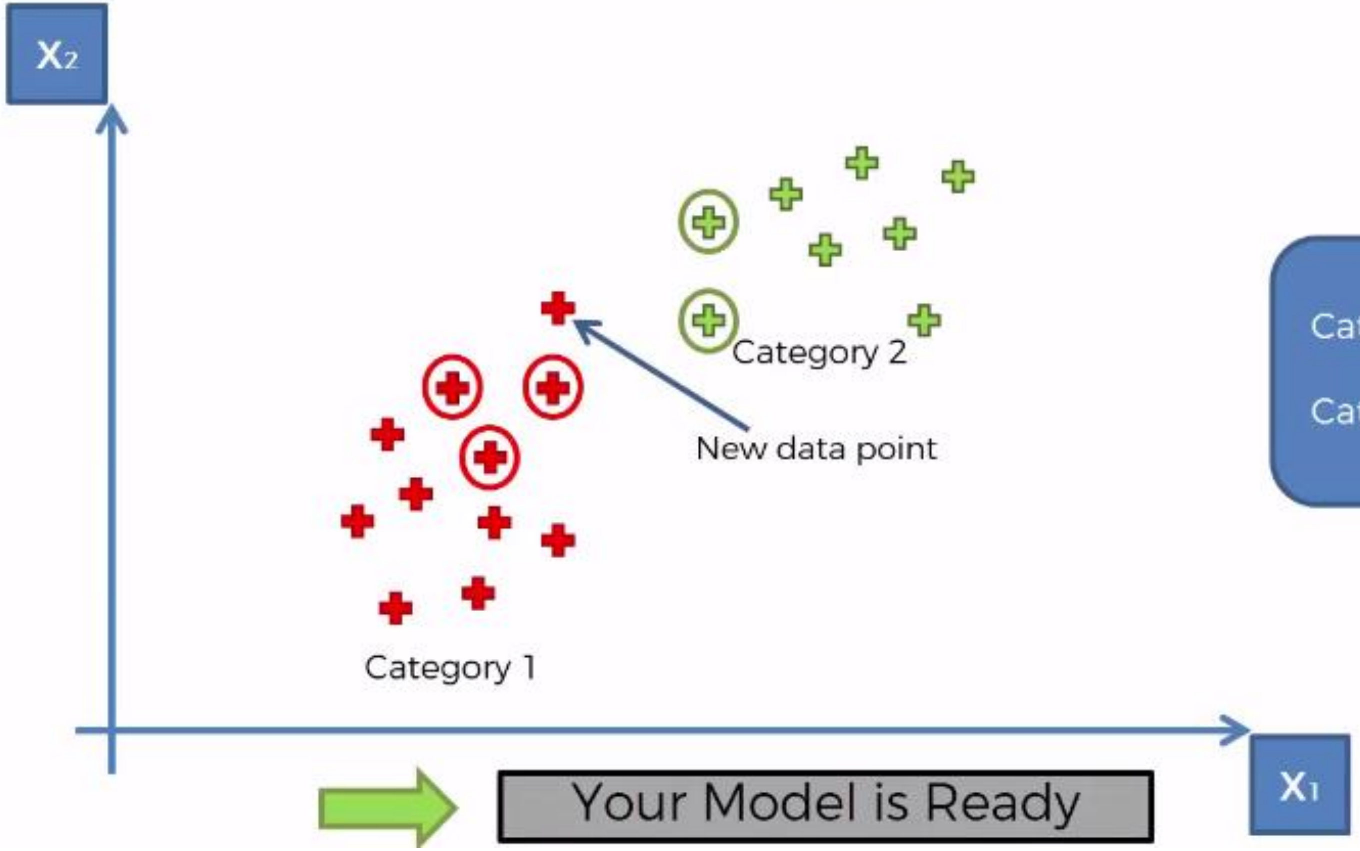


Category 1: 3 neighbors

Category 2: 2 neighbors

K-NN algorithm

STEP 4: Assign the new data point to the category where you counted the most neighbors



Category 1: 3 neighbors

Category 2: 2 neighbors

Your Model is Ready

Example

- A company produce tissues (used by biological labs).

- The company’s objective is to predict how well their products are accepted by their clients.

- They conducted a survey with their clients to find the acceptance of the product. Quality is based on acid durability and strength parameter.

Example

- The data set pertains to a company that produces tissues for use in biological labs.

Name	Acid Durability	Acid Strength	Acceptability
Type-1	7	7	Low
Type-2	7	4	Low
Type-3	3	4	High
Type-4	1	4	High

Test data: **Type-5** **Acid Durability = 3** **Strength = 7**

- **Built a classifier to predict a new type of tissue**

- Apply the Euclidian distance measure for the data to find the distances from the new data **Type-5.**

Name	Acid Durability	Strength	Distance	NeighorRank
Type-1	7	7	Sqrt((7-3)^2+(7-7)^2) =4	3
Type-2	7	4	Sqrt((7-3)^2+(4-7)^2)=5	4
Type-3	3	4	Sqrt((3-3)^2+(4-7)^2)=3	1
Type-4	1	4	Sqrt((1-3)^2+(4-7)^2)=3.6	2

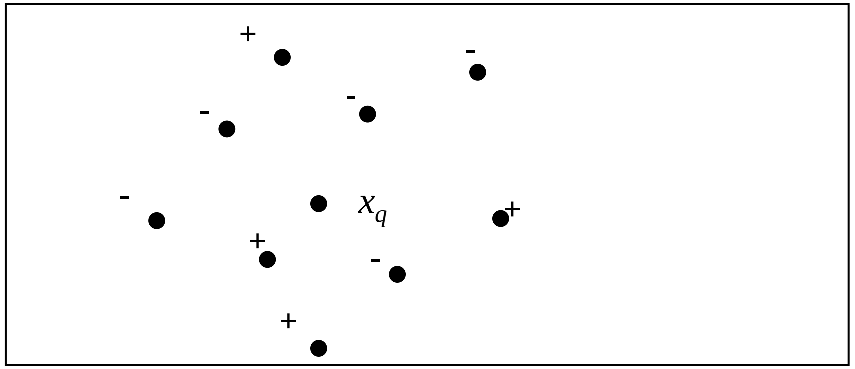
- **If k =1**, ONE immediate neighbor Type 3 Good (new type is = High)
- **If k =2**, TWO immediate neighbor Type 3, Type 4 = High; (new type is = High)
- **If k =3**, THREE immediate neighbor Type 3, Type 4 = High & Type 1 = Low (but the probability of High is high so, consider new type is classified as High)

Example 2

- Assume a Boolean target function and a 2-dimensional instance space (shown in figure).
- Determine how the *k*-Nearest Neighbour Learning algorithm would classify the new instance x_q for k = 1,3,5 and 7.
- The + and – signs in the instance space refer to positive and negative examples respectively.

Distance from query instance	Classification
1.00	+
1.35	-
1.40	-
1.60	-
1.90	+
2.00	+
2.20	-
2.40	+
2.80	-

Distance from query instance	Classification
1.00	+
1.35	-
1.40	-
1.60	-
1.90	+
2.00	+
2.20	-
2.40	+
2.80	-



1-NN	+
3-NN	-
5-NN	-
7-NN	-

Selection of K value ?

- Try many different values for K and see what works best for your problem.
- K value should be an odd number (3, 5, 7, 9, etc.).

How does the efficiency and accuracy of k-NN search change as k increases?

- If we have sufficiently large number of training experiences the accuracy should increase
- The computational complexity of KNN increases with the size of the training dataset.
    - The time to calculate the prediction will also increase.
    - In that sense less efficient

- **KNN is a Lazy Learning algorithm – why?**

- No learning of the model/ algorithm

- It “memorizes” the training dataset

- DT **algorithm** learns its model during training time

- **KNN is a Non-Parametric algorithm – why?**

- It makes no assumptions about the functional form of the problem being solved.

- Is KNN supervised or unsupervised learning algorithm?

- **KNN** is a **supervised** learning algorithm, uses labeled data for classification problem.

- Note: K-means is an **unsupervised** learning algorithm used for clustering problem

Thank you