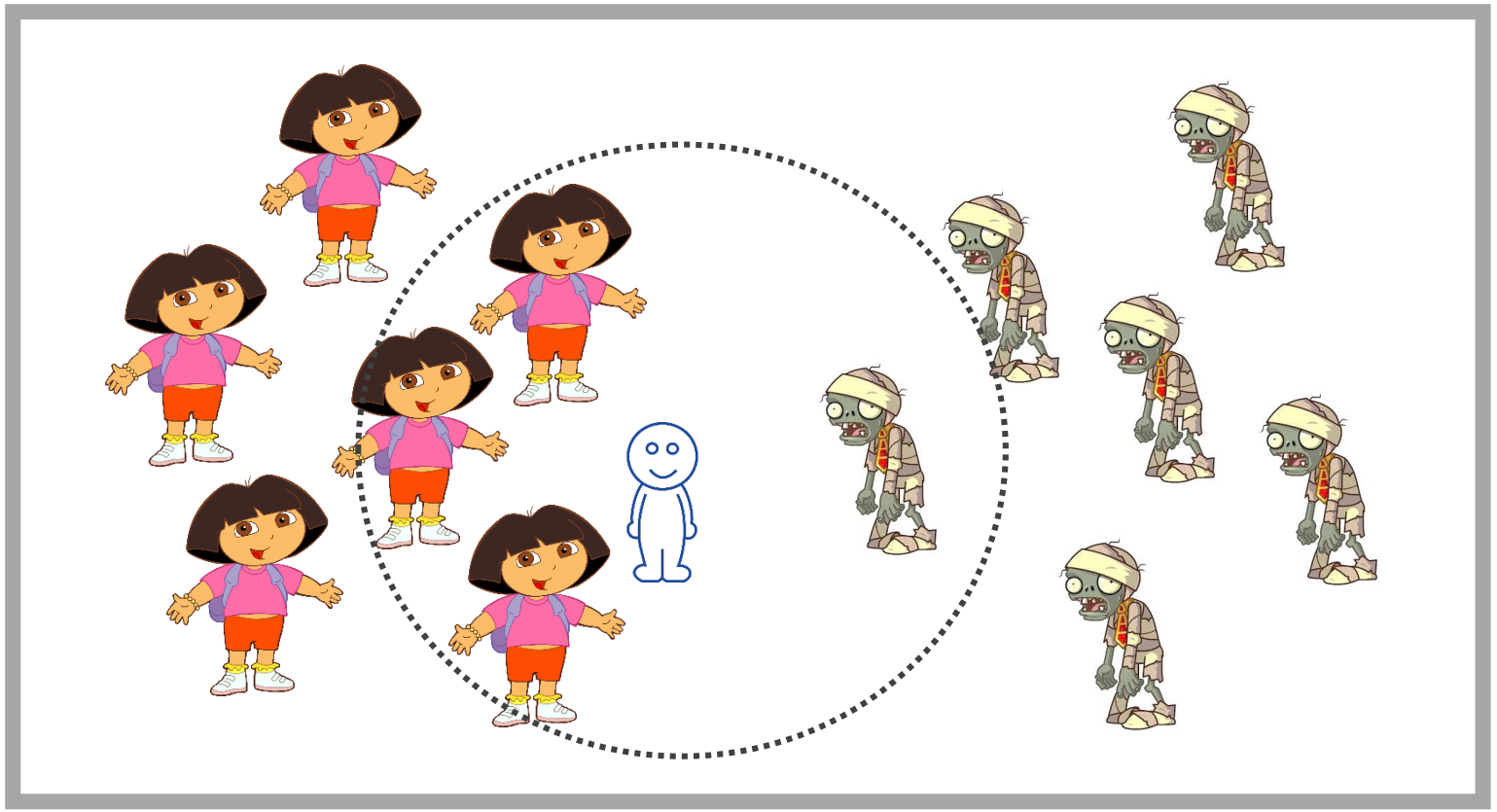
K-Nearest Neighbors

KNN Intuition

Have you ever wondered about the difference between Zombie and Human

KNN Intuition



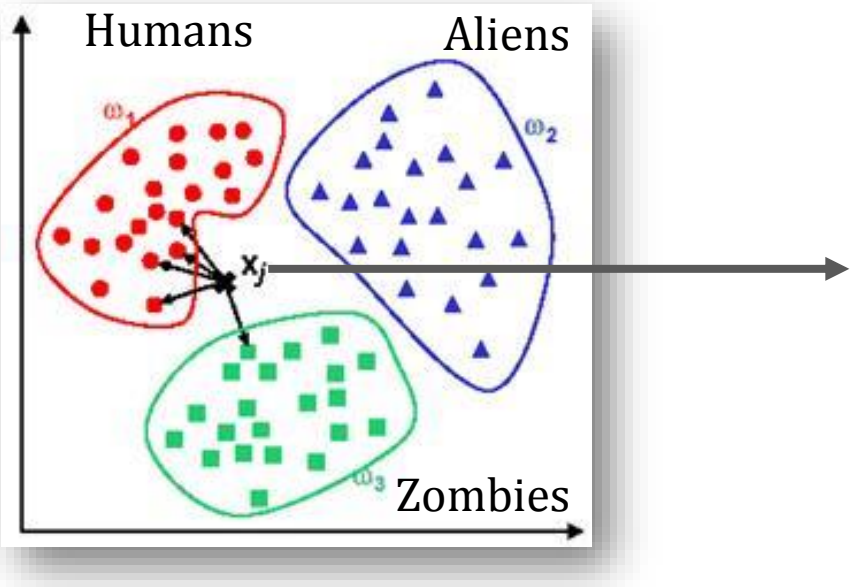
Hence The New Character Is Human

Predictions on new data point is done by searching through the entire training set for the K most similar instances

KNN Making Predictions

Training Will Be On
Entire Dataset

Predictions made by KNN are directly
with training dataset



What if I have a new data point

The search will be through the entire training
set for the K most similar instances

How To Determine Most Similar K Instances

Distance measure to Calculate The Similarity

Choose the distance metric based on our data properties

Distance Metrics	Data
Euclidean distance	Real-valued input variables (Widths and Heights)
Manhattan Distance	Real-valued input variables which not similar in type (Age, Gender, Height)
Hamming Distance	Binary vectors

Experiment with different distance metrics

How To Decide The K Value

Try many different values for K

1 to 21

K can be found by algorithm tuning

Thumb Rule

Square root of number of variables

Try K = ODD

Try K = Even

KNN Algorithm

Predictions are based on K-most similar instances

Used

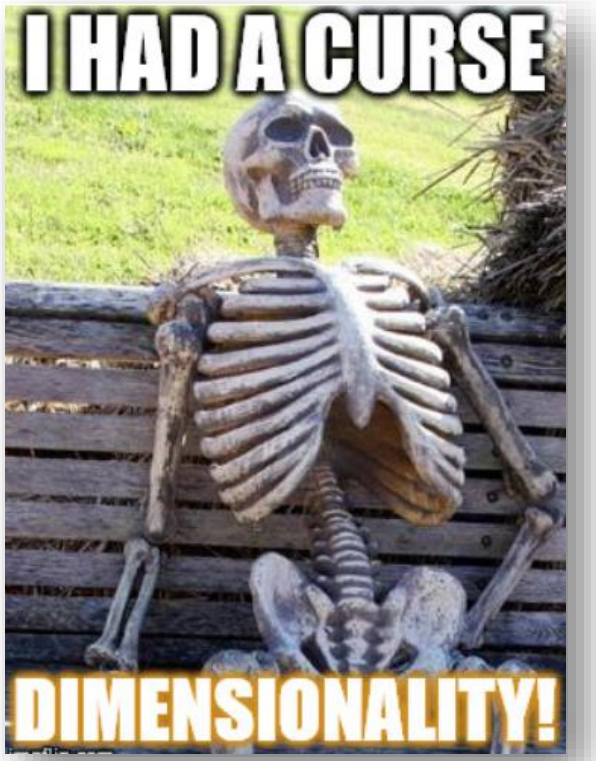
Regression
Mean or Median

Classification
Mode

Curse of Dimensionality

Increasing the dimensions increases the distance between two points

KNN is best for lower dimensional data



Prepare Data For KNN

Rescale Data

Impute Missing Data

Lower Dimensional Data

How KNN Works

Pick A Number Of Neighbors

↓

Choose a Distance Method

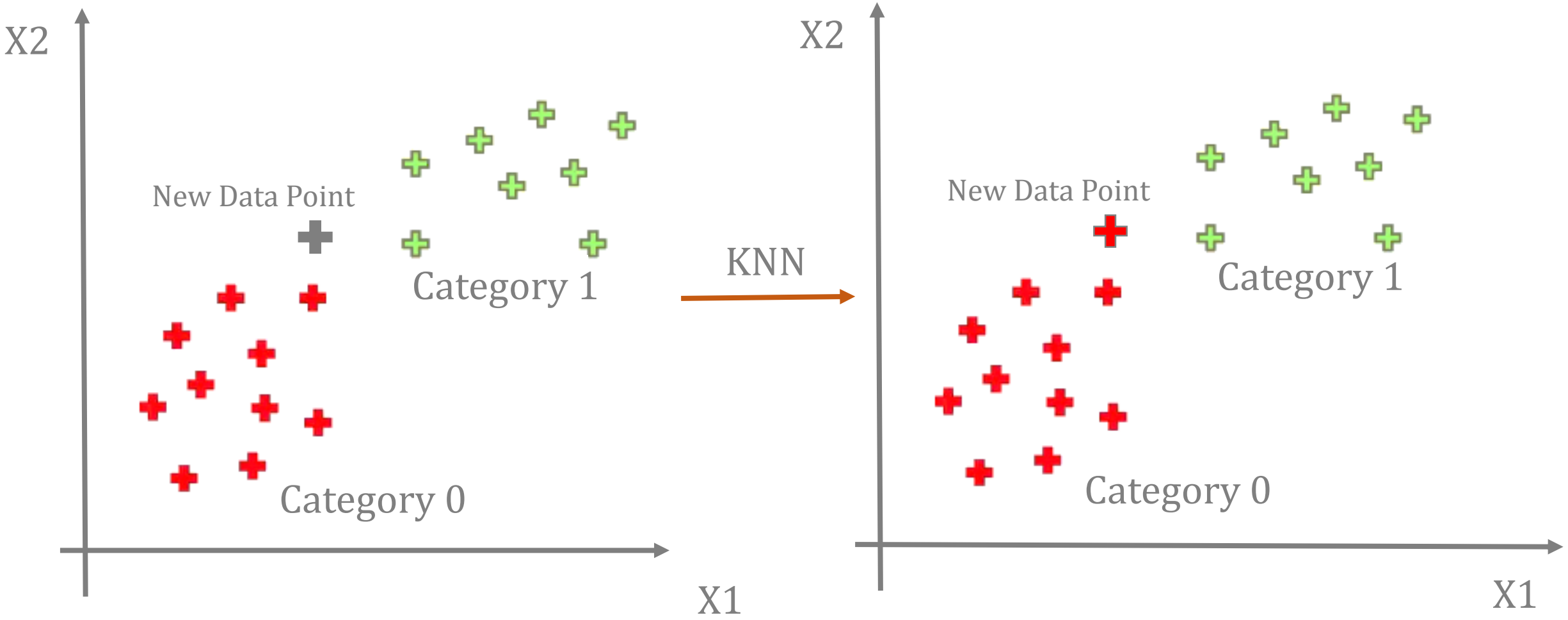
↓

For a New Point, Identify The Number Of Nearest Neighbors

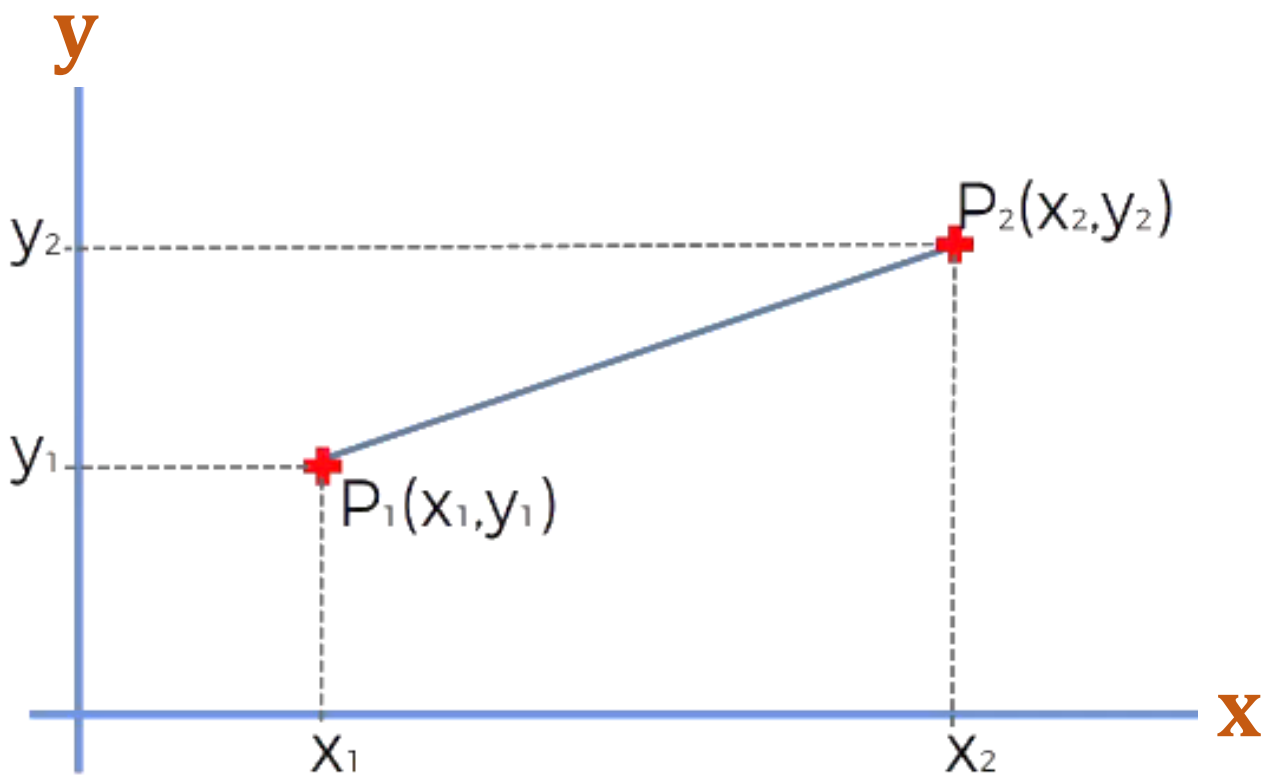
↓

Vote or Mode If it is a classification
Or
Consider A Mean/Median For Regression

KNN With Example



KNN With Example



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

KNN With Example

K = 5

Category 1 has 2 Points near to it

Category 0 has 3 points near to it

KNN With Example

Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	
158	58	M	4.2	
158	59	M	3.6	
158	63	M	3.6	
160	59	**M**	2.2	**3**
160	60	**M**	1.4	**1**
163	60	**M**	2.2	**3**
163	61	**M**	2.0	**2**
160	64	**L**	3.2	**5**
163	64	L	3.6	
165	61	L	4.0	
165	62	L	4.1	
165	65	L	5.7	
168	62	L	7.1	
168	63	L	7.3	
168	66	L	8.6	
170	63	L	9.2	
170	64	L	9.5	
170	68	L	11.4	
161	**61**			

→ Predicted Value

A small value of k = Large variance in predictions

A large value of k = Large model bias

END.