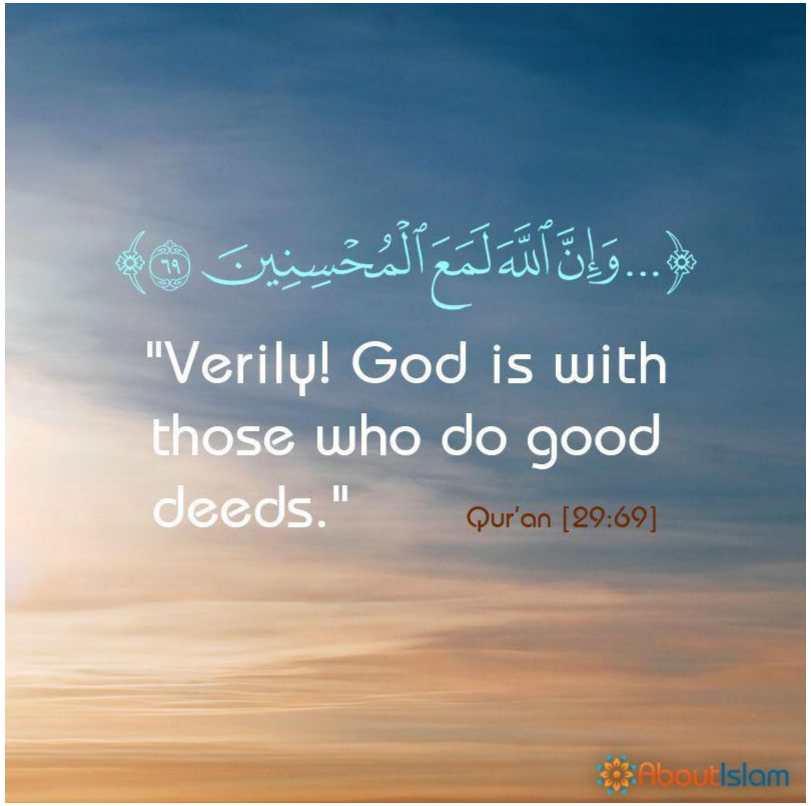
Data Mining

K-NEAREST NEIGHBOR

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology
UNIVERSITY OF SARGODHA, SARGODHA

Lesson from Al-Quran

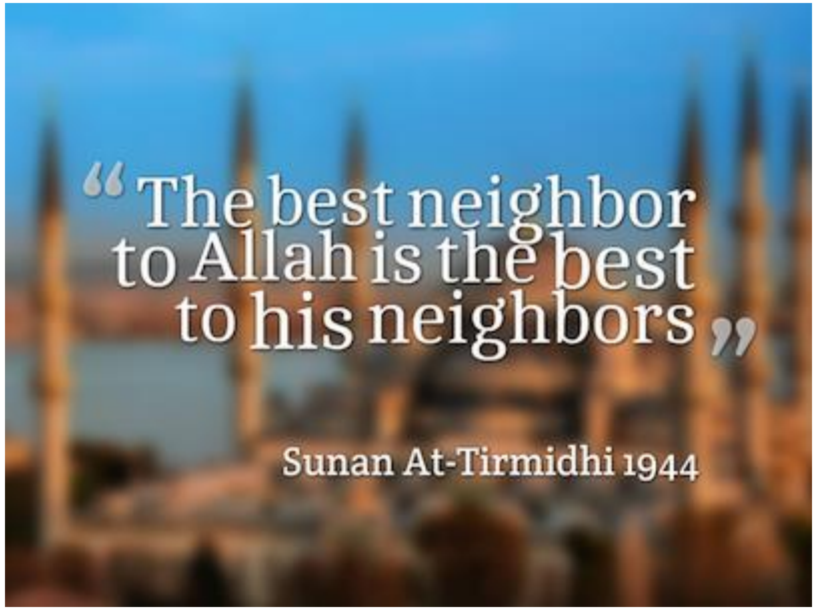


﴿...وَإِنَّ ٱللَّهَ لَمَعَ ٱلْمُحْسِنِينَ ﴿٦٩﴾﴾

"Verily! God is with those who do good deeds."

Qur'an [29:69]

AboutIslam



"The best neighbor to Allah is the best to his neighbors"

Sunan At-Tirmidhi 1944

Agenda

- Lazy algorithms
- KNN
  - Basic Idea
  - Nearest neighbor concept
  - Classification using KNN
  - Value of K
  - Eamples
  - Exercise
  - To do Task

Basic Idea

Idea:

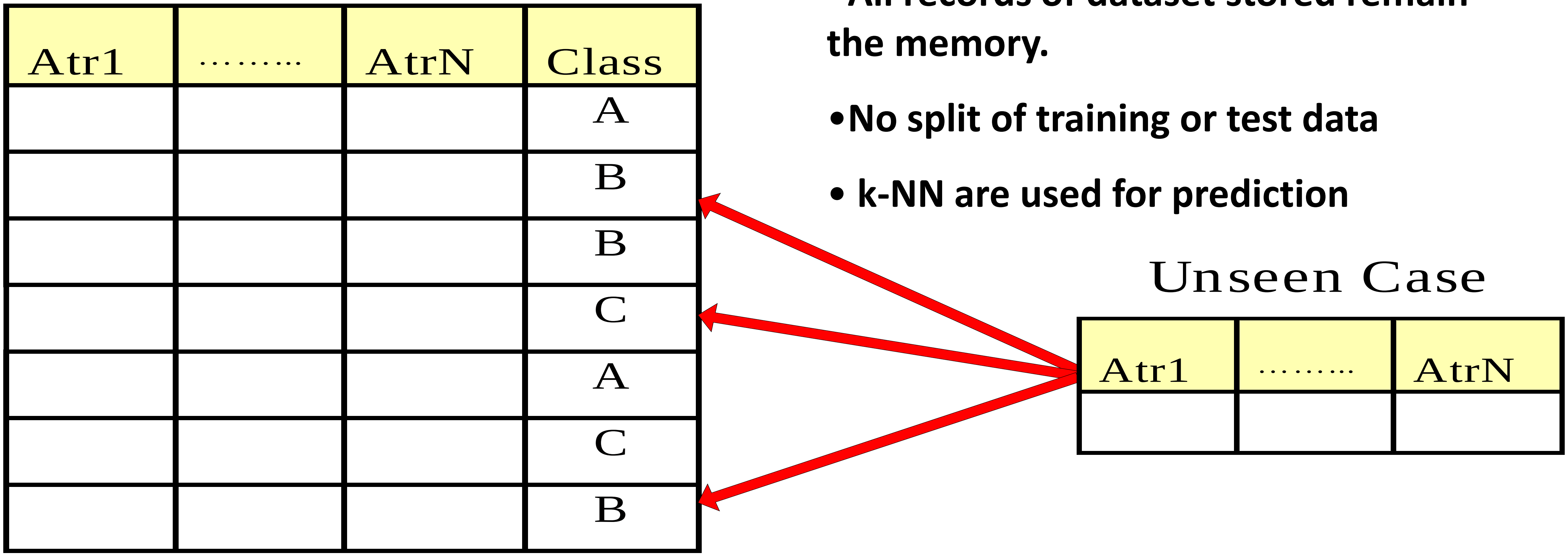
- k-NN stands for k-Nearest neighbour
- k-NN is a simple algorithm
- Predicts the class of a new case (object or instance) based on a similarity with the already instances having class labels

One Algo: Different Names

- K-Nearest Neighbors
  - Considers k neighbors only
- Memory-Based Reasoning
  - Required data for computation
- Instance-Based Learning
- Example-Based Reasoning
- Case-Based Reasoning
  - Takes existing examples into account, considers test instance and computes result
- Lazy Learning
  - All computation deferred until classification decision

Instance-Based Classifiers

- All records of dataset stored remain the memory.
- No split of training or test data
- k-NN are used for prediction



Set of Stored Cases

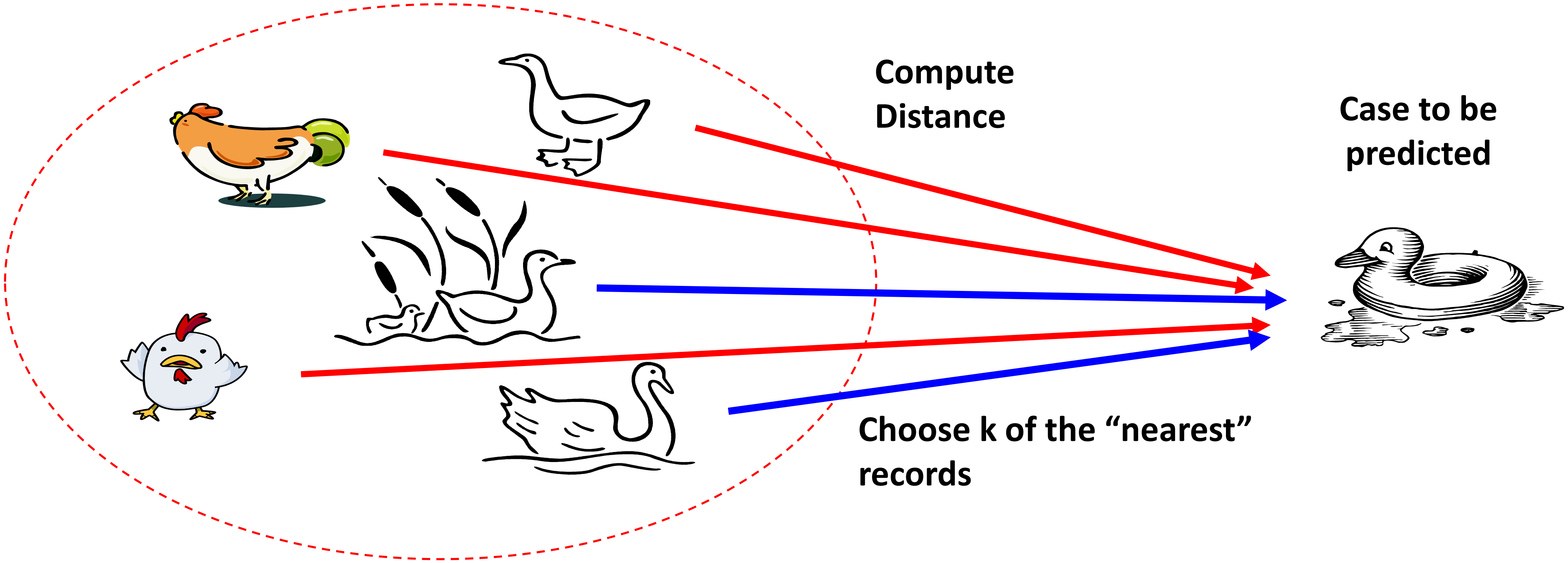
Atr1	………	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

Unseen Case

Atr1	………	AtrN

Nearest Neighbor Classifiers

- Basic idea:
  - If it swims like a duck, quacks like a duck, then it's probably a duck



Compute Distance

Case to be predicted

Data

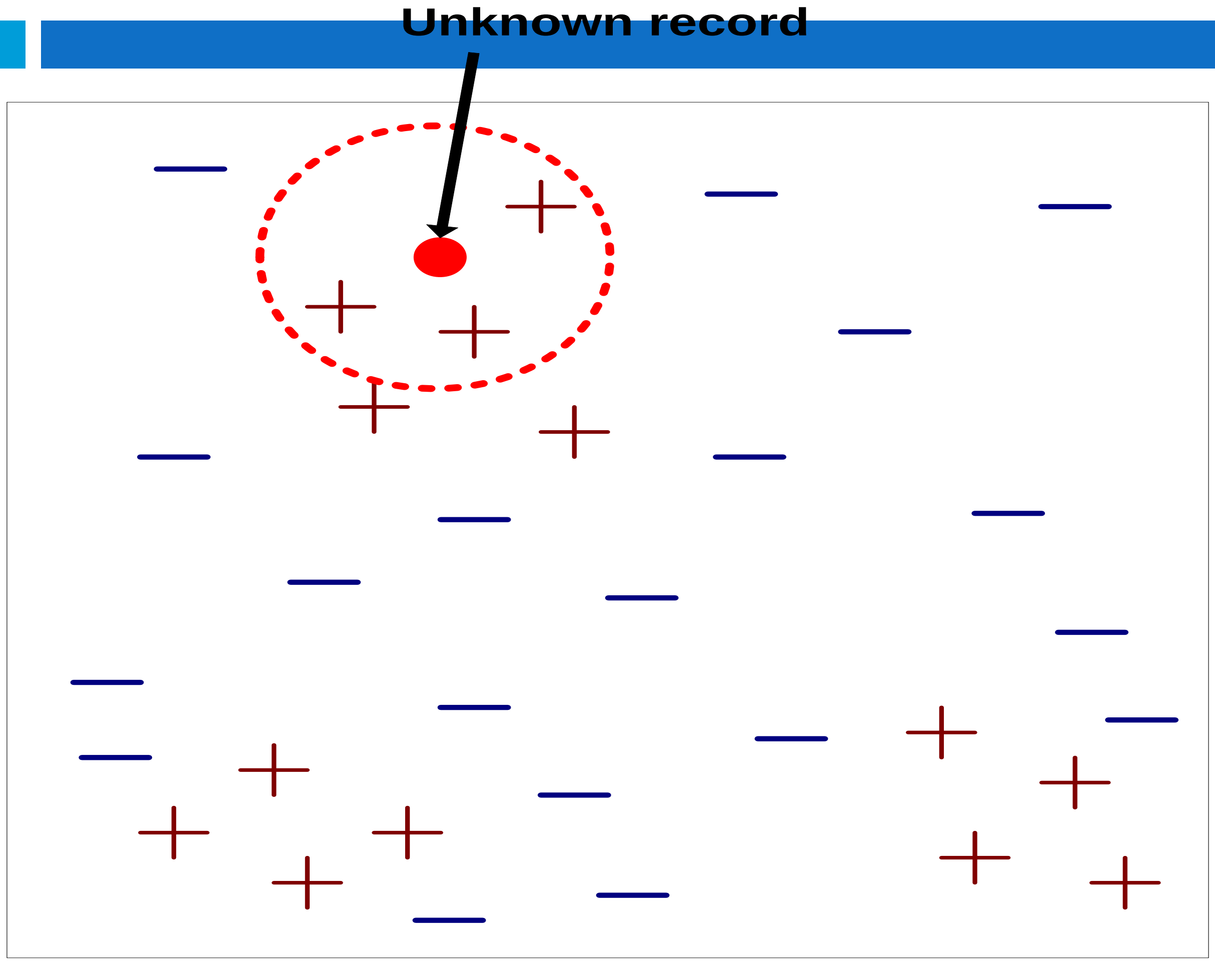
Records

Choose k of the "nearest" records

Number of Neighbors

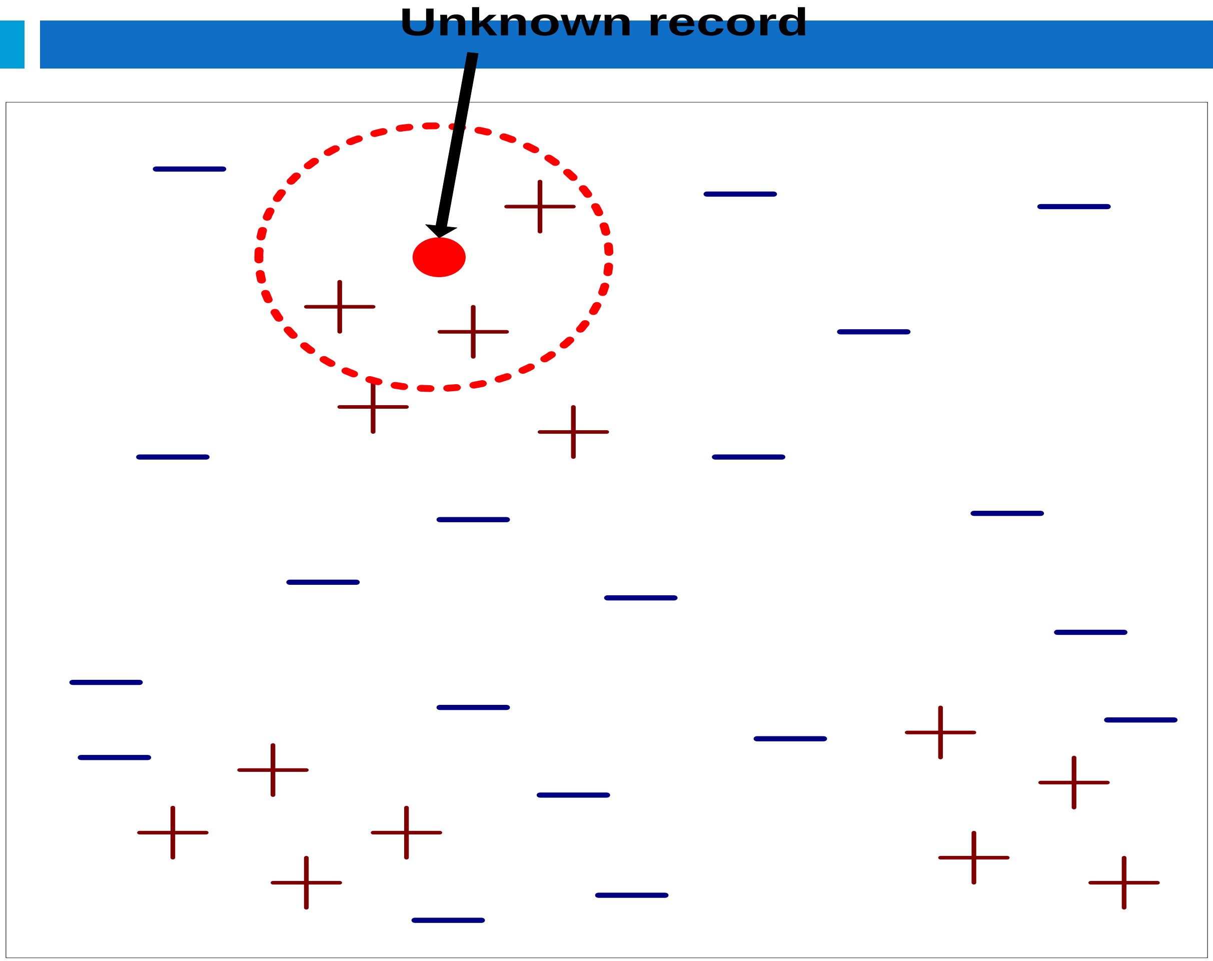
- If K=1,
  - select the nearest neighbor
- If K>1,
  - For classification select based on k neighbors.

Nearest-Neighbor Classifiers



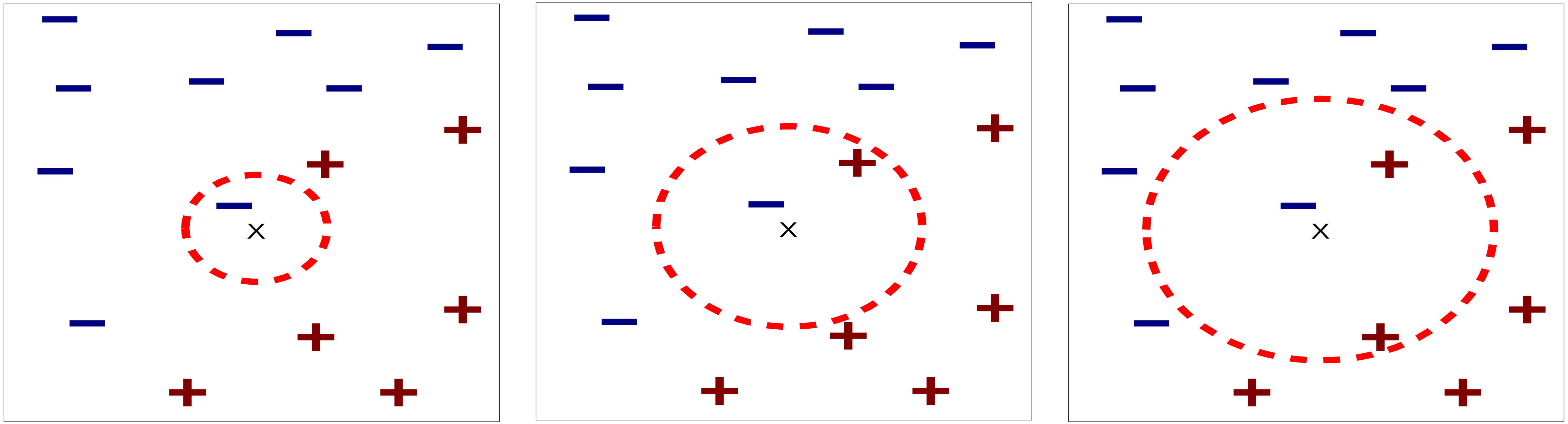
- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

Nearest-Neighbor Classifiers



- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

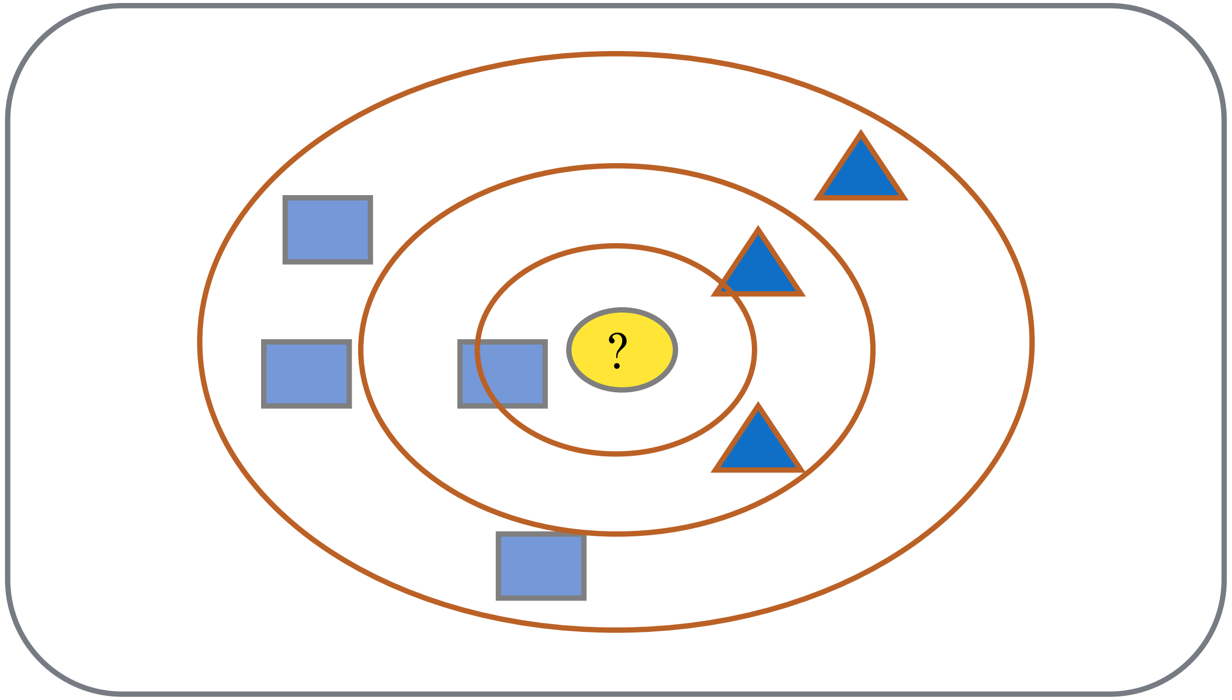
Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

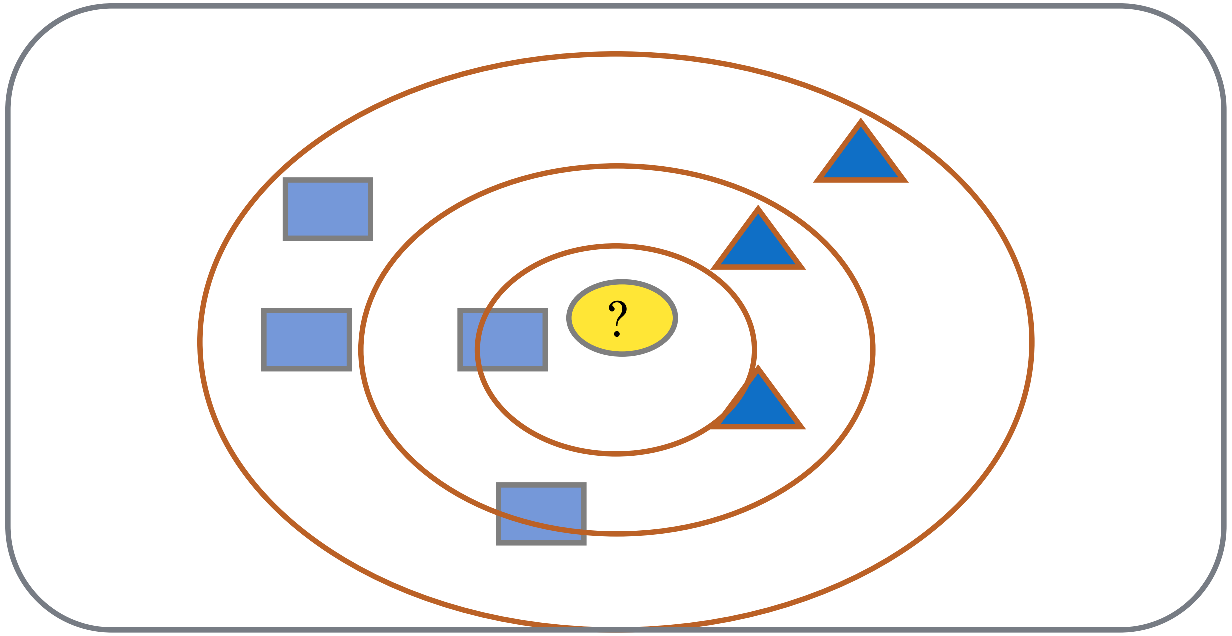
K-nearest neighbors of a record x are data points that have the k smallest distance to x

k NEAREST NEIGHBOR



- $k = 1$:
  - Belongs to square class
- $k = 3$:
  - Belongs to triangle class
- $k = 7$:
  - Belongs to square class

k NEAREST NEIGHBOR

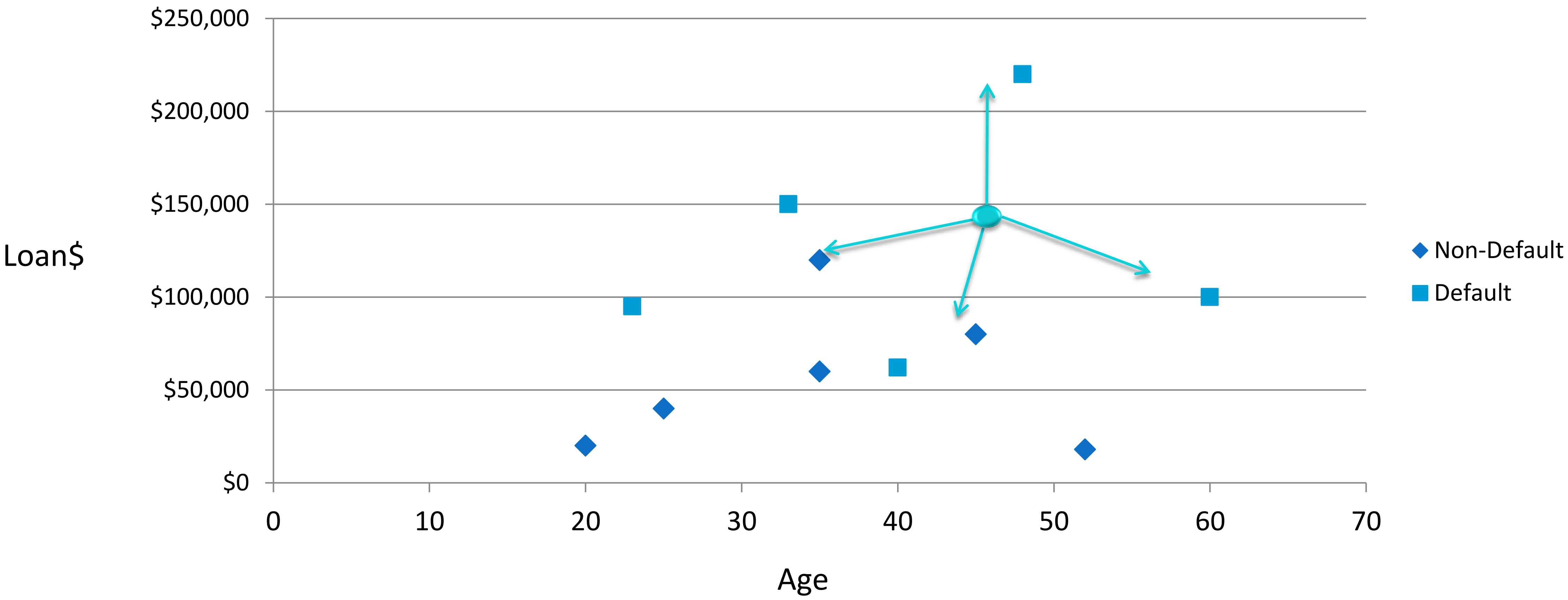


- $k = 1$:
  - Belongs to square class
- $k = 3$:
  - Belongs to triangle class
- $k = 7$:
  - Belongs to square class

□ Choosing the value of *k*:

- **If *k* is too small, sensitive to noise points**
- **If *k* is too large, neighborhood may include points from other classes**
- **Choose an odd value for *k*, to eliminate ties**

KNN Classification



KNN Classification – Distance

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
48	**$142,000**	**?**	

Euclidean Distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

k NEAREST NEIGHBOR

- □ Accuracy of **all** NN based algorithms depends on a data model.
- □ Scaling issues/ WHY NORMALIZATION
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes.
  - Examples
    - Height of a person may vary from 4’ to 6’
    - Weight of a person may vary from 100lbs t 300lbs
    - Income of a person may vary from $10k to $500k

KNN Classification – Distance

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
48	**$142,000**	**?**	

Euclidean Distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

KNN Classification – Standardized Distance

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

Nearest Neighbor Classification

- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors

Example
(Test: Durability:3, Strength:7, Class;?)

Type No	Item Durability	Item Strength	Class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good

Example

Type No	Item Durability	Item Strength	Class	Distance
Type-1	7	7	Bad	Sqrt$((7-3)^2 + (7-7)^2) = 4$
Type-2	7	4	Bad	
Type-3	3	4	Good	
Type-4	1	4	Good	

Type No	Item Durability	Item Strength	Class	Distance
Type-1	7	7	Bad	Sqrt($(7-3)^2$ + $(7-7)^2$) = 4
Type-2	7	4	Bad	5
Type-3	3	4	Good	3
Type-4	1	4	Good	3.6

Type No	Item Durability	Item Strength	Class	Distance	Rank
Type-1	7	7	Bad	Sqrt($(7-3)^2 + (7-7)^2$) = 4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2

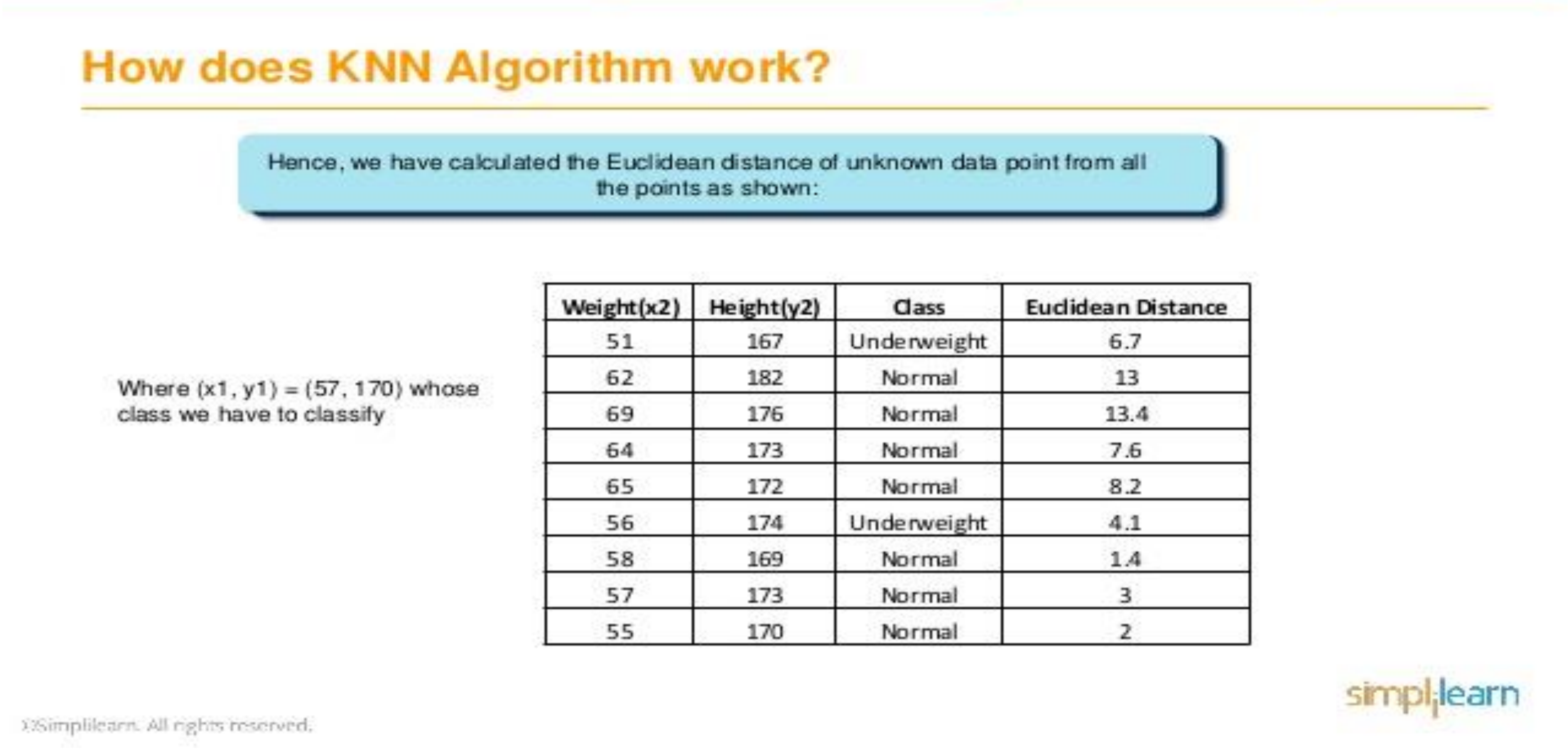
Merits and Demerits

- **Advantages**
  - **Simple technique that is easily implemented**
  - **Can work with relatively little information**
  - **Well suited for multi classes problems**
  - **Learning is simple (does not involve preprocessing)**
  - **Performs best in some cases (gene and protein identification)**
- **Dis-advantages**
  - **Memory issues and expensive computation for large datasets**
  - **Low Accuracy if presence of noisy or irrelevant features**
  - **Feature selection problem**

Exercise: KNN using R

- KNN tutorial using R language
  - https://www.youtube.com/watch?v=lDCWX6vCLFA
- KNN using Python
  - From Scratch
  - https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/
  - Using SkLearn Library
  - https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/
- Dataset
  - UCI dataset
  - Kaggle data
  - Use any data you like
- Use tool such as Weka or RapidMiner

Exercise: do it yourself



How does KNN Algorithm work?

Hence, we have calculated the Euclidean distance of unknown data point from all the points as shown:

Where (x1, y1) = (57, 170) whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

I am my nearest neighbour.

Tacitus