CS500-Data Science Tools and Technique

K-Nearest Neighbor (KNN)

A Supervised Machine Learning Algorithm

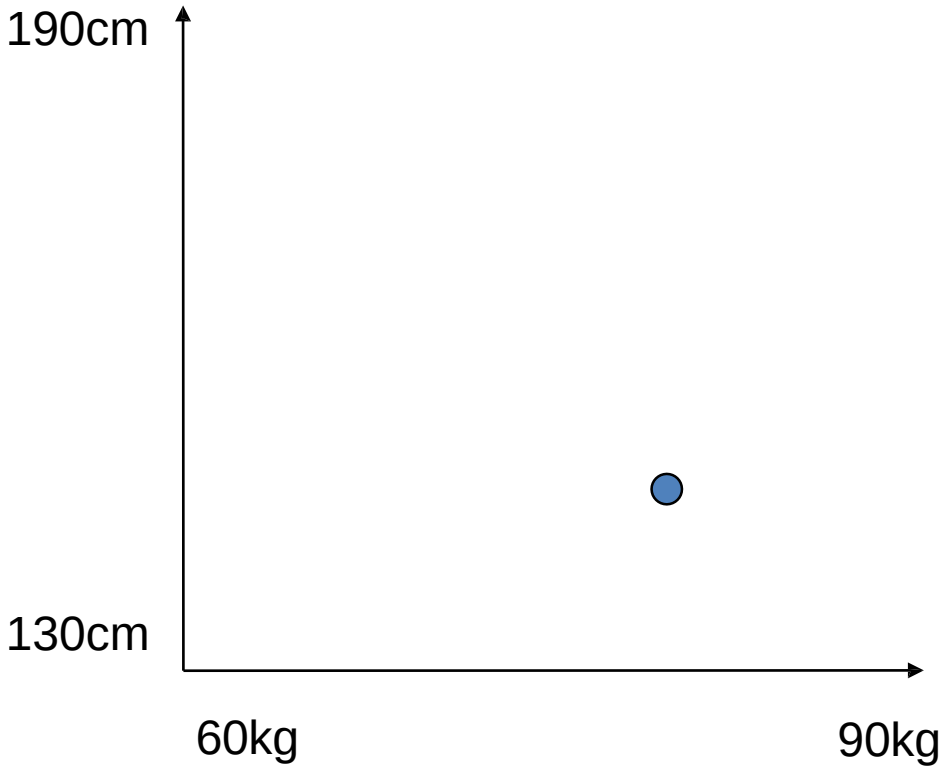
Bahar Ali
PhD Scholar,
National University Of Computer and Emerging Sciences,
Peshawar.

K Nearest Neighbor & Neural Networks

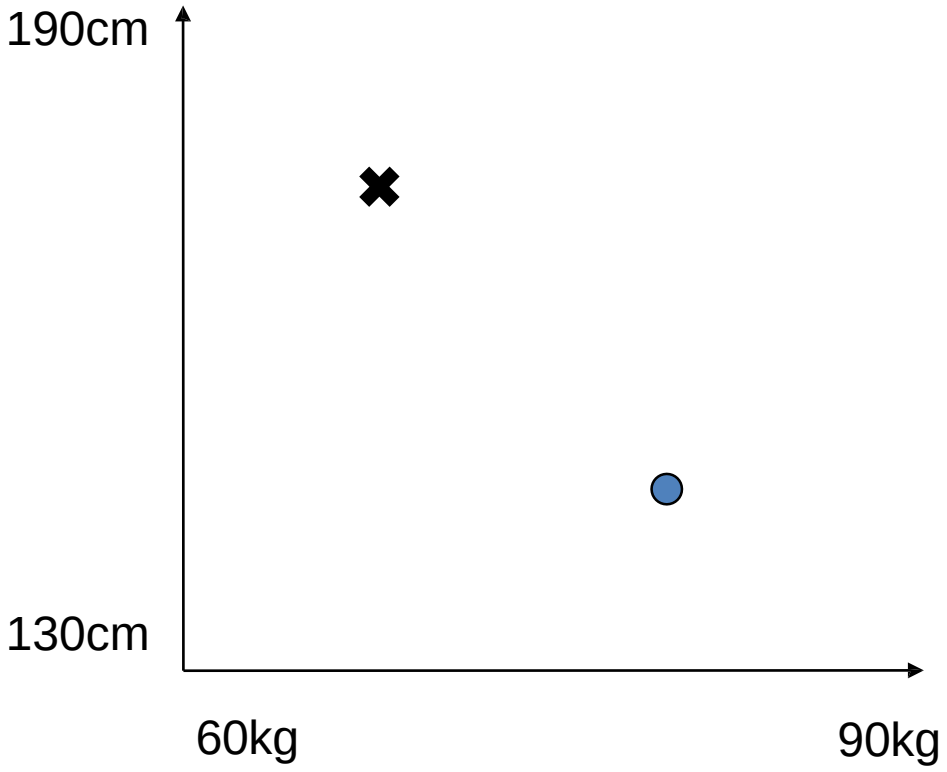
Can we LEARN to recognise a rugby player?

What are the “features” of a rugby player?

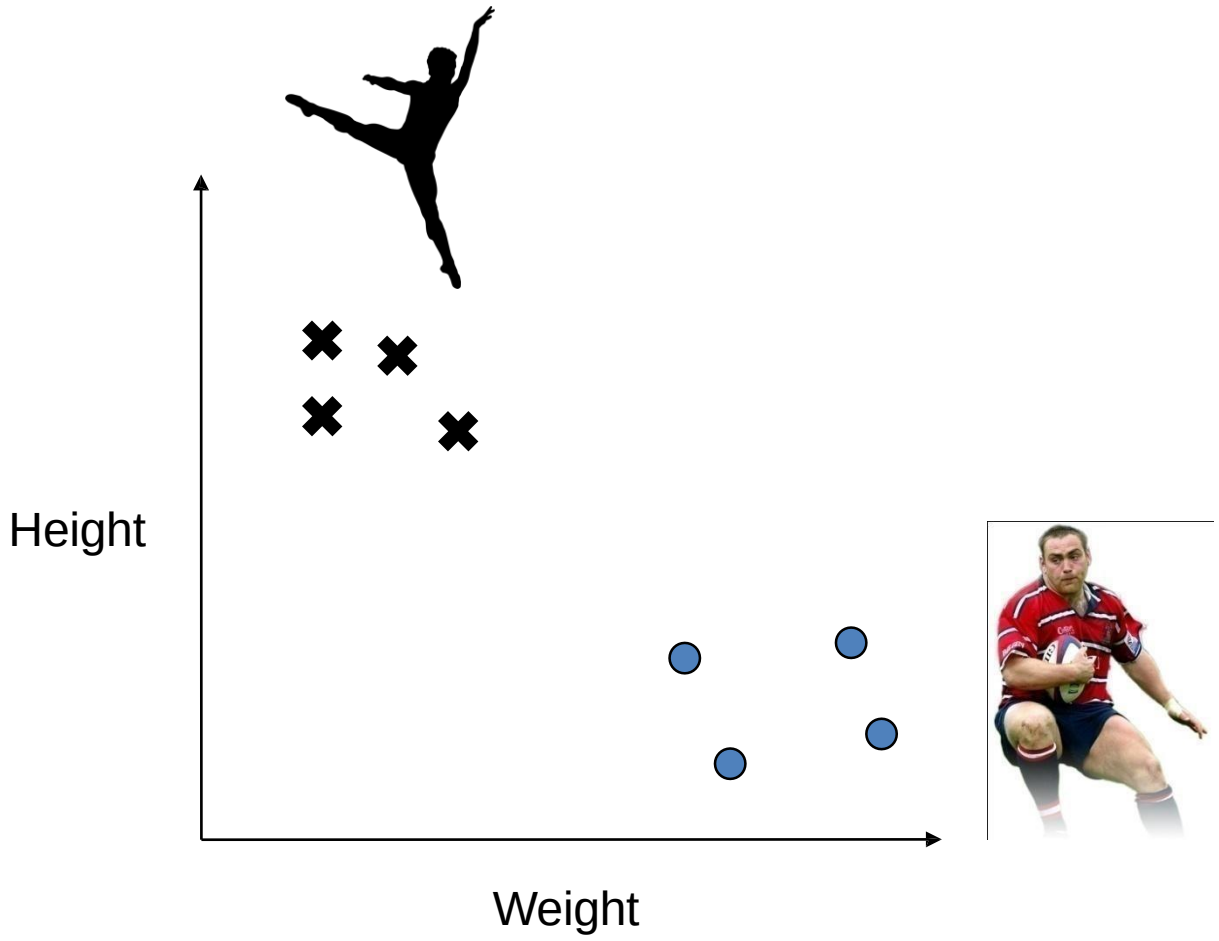
Rugby players = short + heavy?



Ballet dancers = tall + skinny?

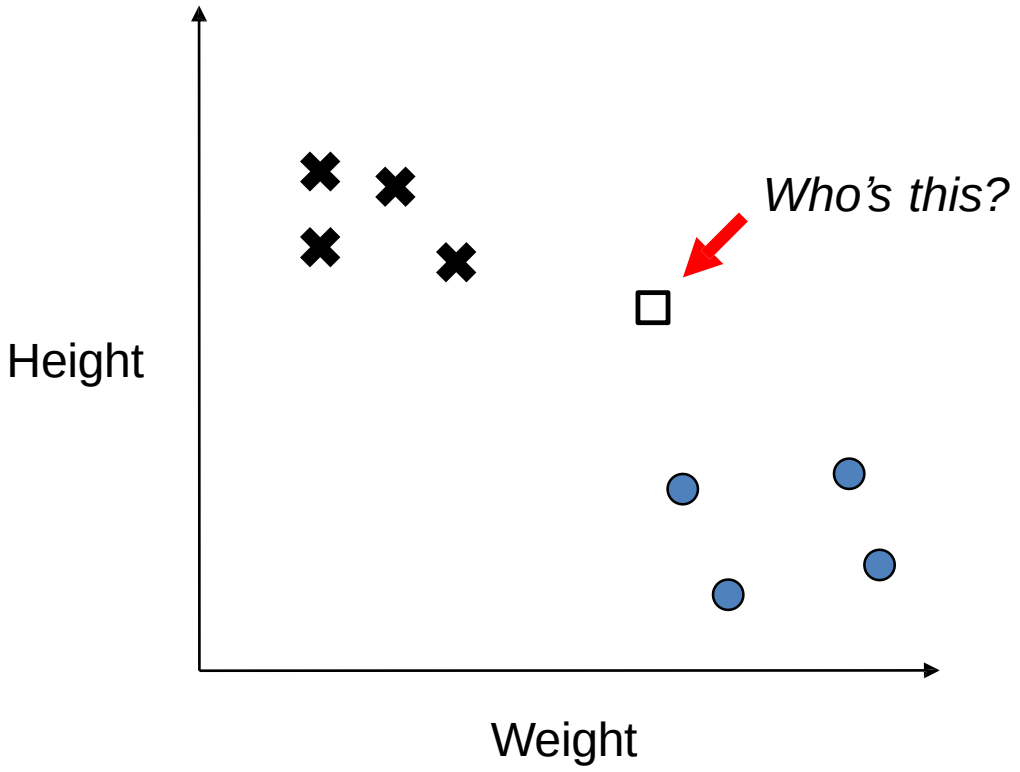


Rugby players “cluster” separately in the space.



K Nearest Neighbors

The K-Nearest Neighbour Algorithm



The K-Nearest Neighbour Algorithm

Step 1: Determine the value for K

Step 2: Calculate the distances between the new input (test data) and all the training data. The most commonly used metrics for calculating distance are ***Euclidean, Manhattan and Minkowski***

Step 3: Sort the distance and determine k nearest neighbors based on minimum distance values

Step 4: Analyze the category of those neighbors and assign the category for the test data based on majority vote

Step 5: Return the predicted class

Distance functions

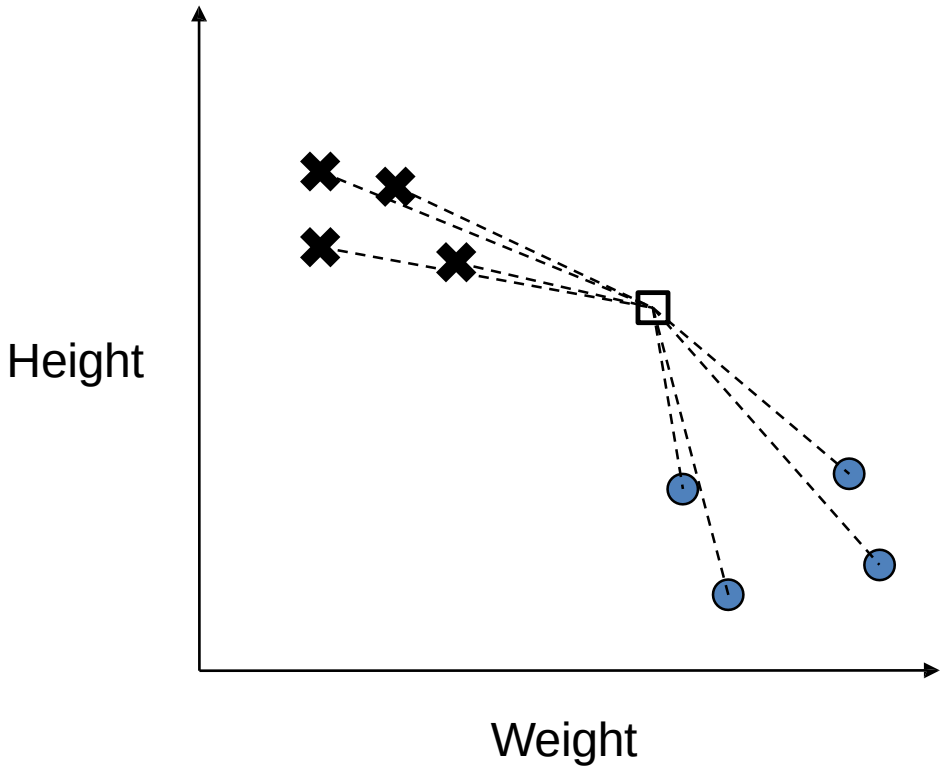
Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

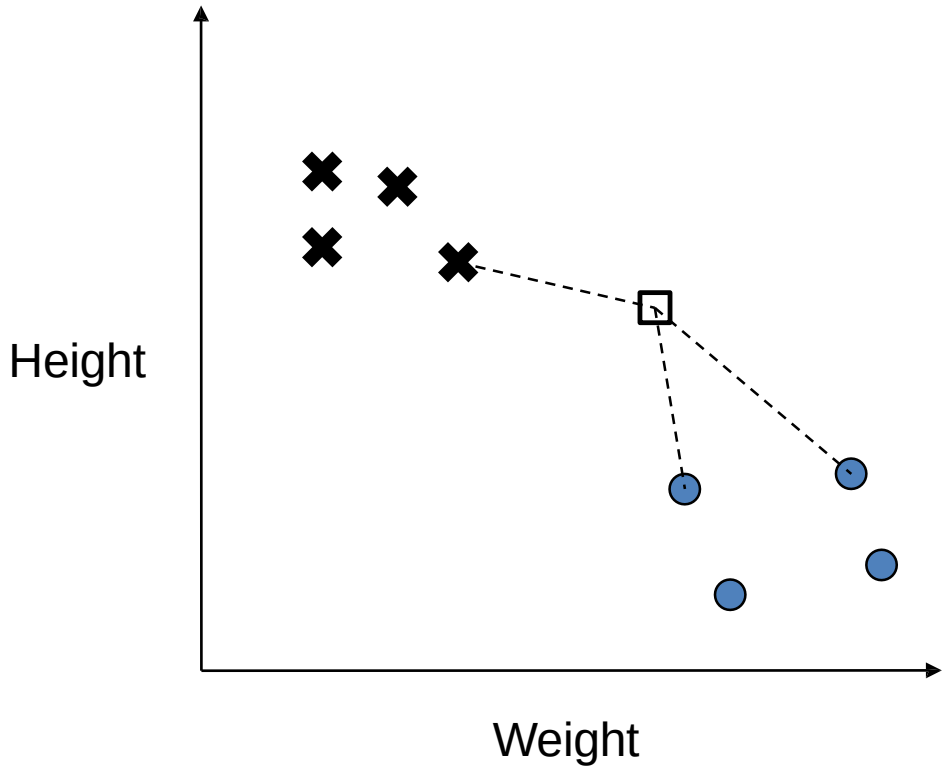
The K-Nearest Neighbour Algorithm

1. *Measure distance to all points*



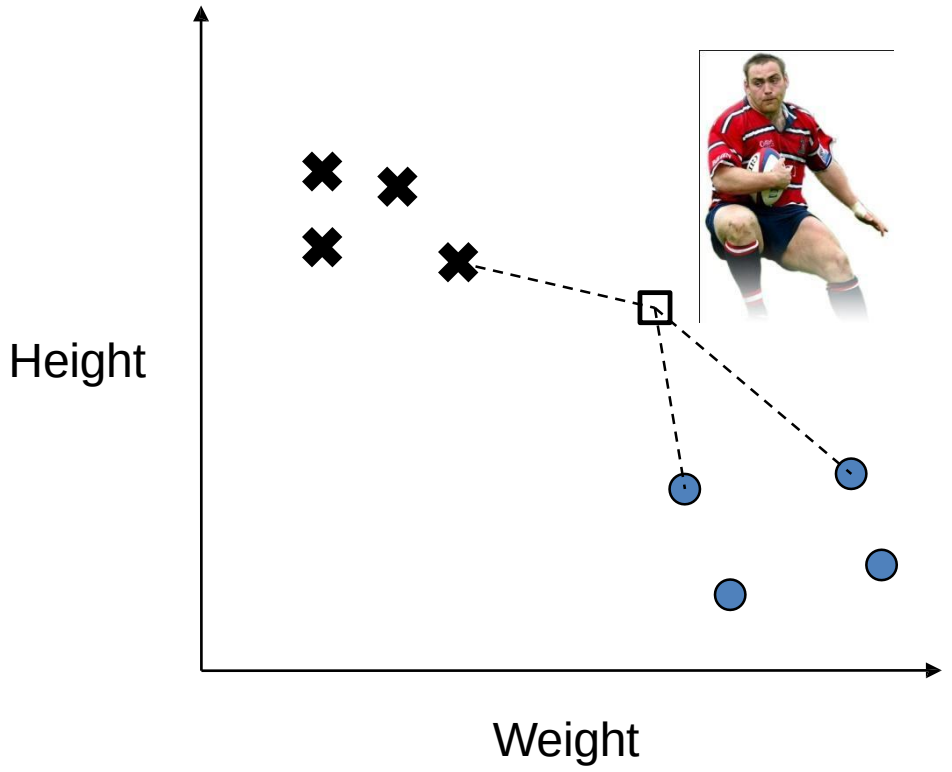
The K-Nearest Neighbour Algorithm

1. *Measure distance to all points*
2. *Find closest "k" points* ← **(here k=3, but it could be more)**



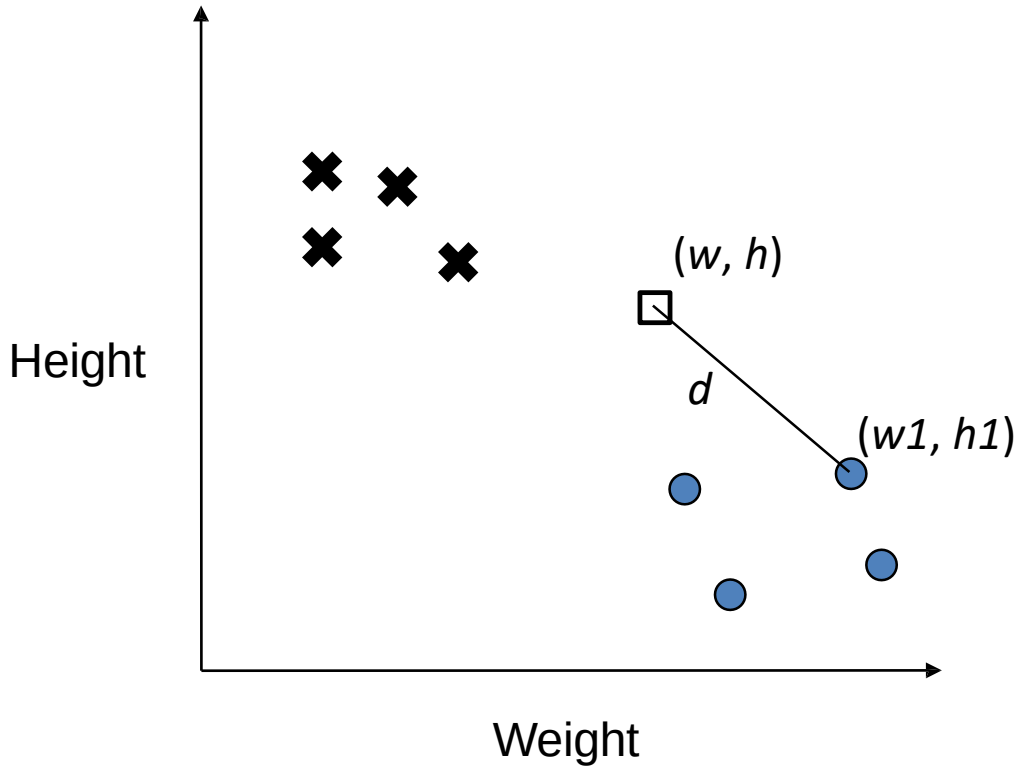
The K-Nearest Neighbour Algorithm

1. *Measure distance to all points*
2. *Find closest "k" points* ← **(here k=3, but it could be more)**
3. *Assign majority class*



“Euclidean distance”

$$d = \sqrt{(w - w_1)^2 + (h - h_1)^2}$$



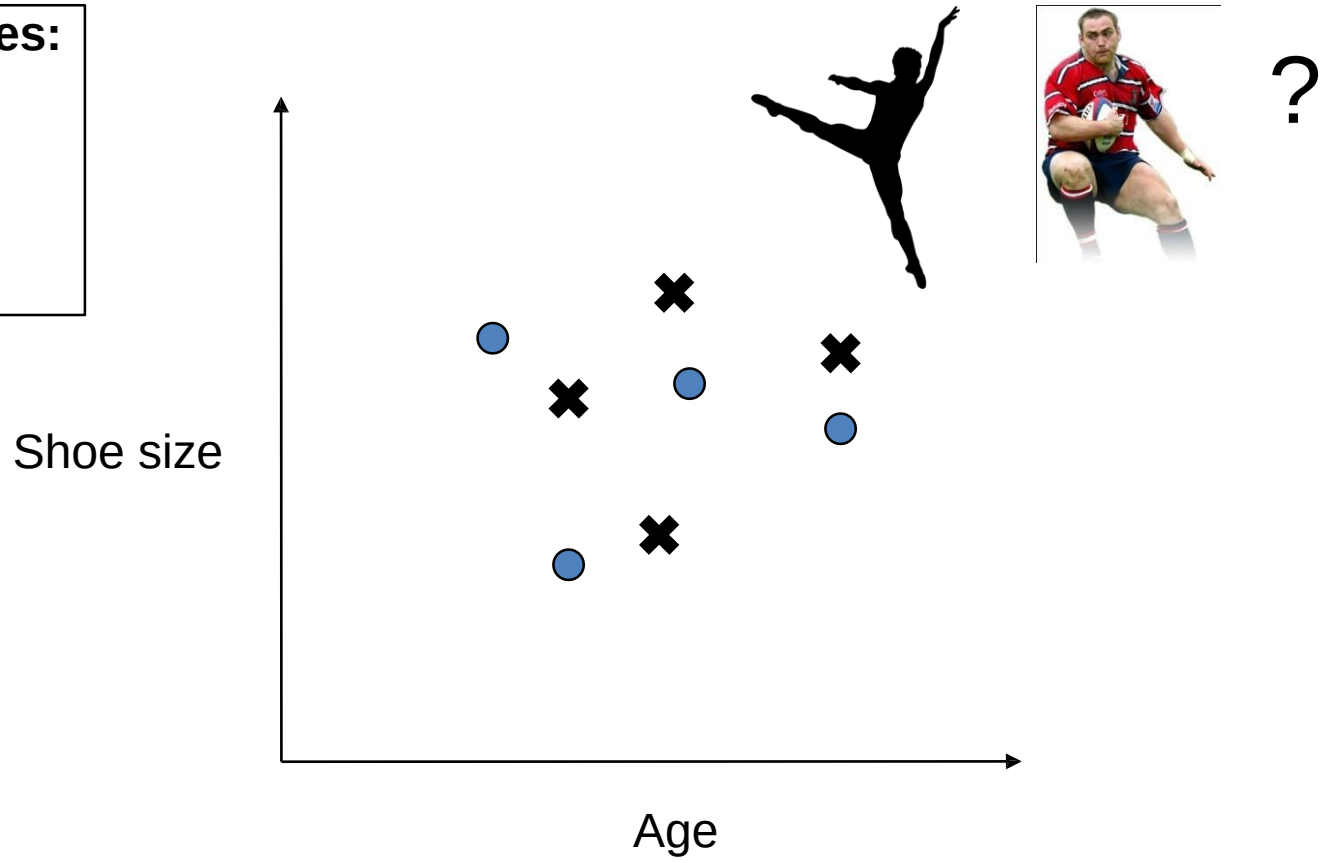
Euclidean distance still works in 3-d, 4-d, 5-d, etc….

$$d = \sqrt{(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2}$$

x = Height
y = Weight
z = Shoe size

Choosing the wrong features makes it difficult, too many and it's computationally intensive.

Possible features:
- Shoe size ✓
- Height
- Age ✓
- Weight



?

K-Nearest Neighbour Model

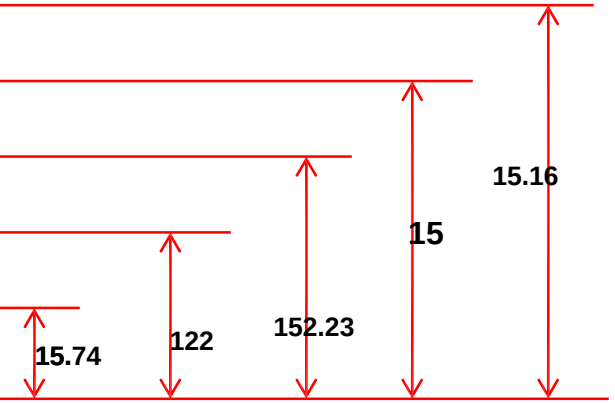
- Example : Classify whether a customer will respond to a survey question using a 3-Nearest Neighbor classifier

Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?

K-Nearest Neighbour Model

- Example : 3-Nearest Neighbors

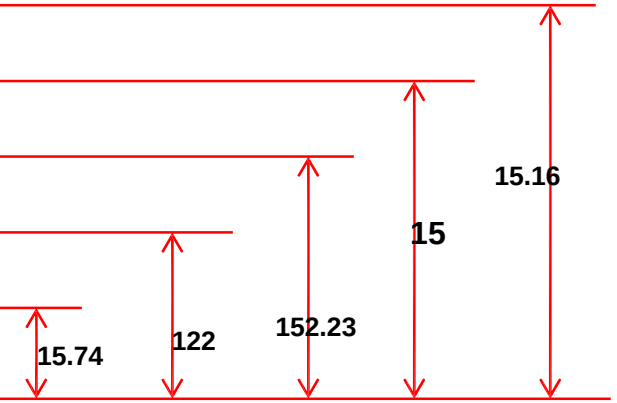
Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?



K-Nearest Neighbour Model

- Example : 3-Nearest Neighbors

Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?

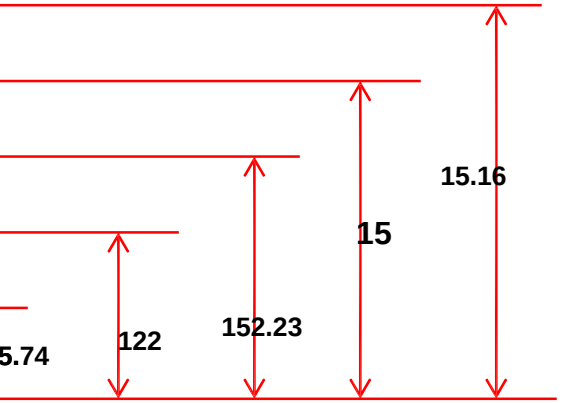


Three nearest ones to David are: No, Yes, Yes

K-Nearest Neighbour Model

- Example : 3-Nearest Neighbors

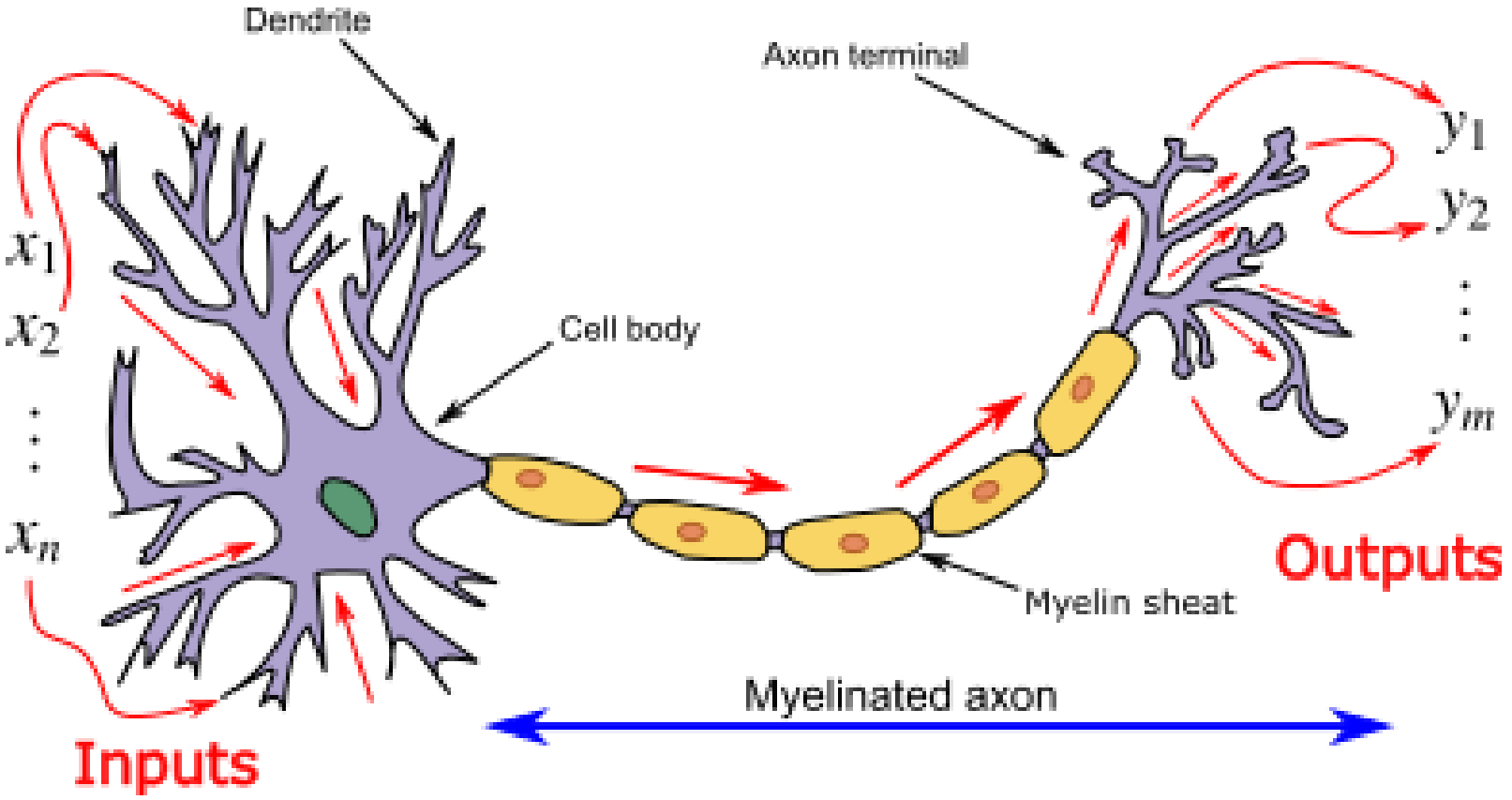
Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22	50K	2	Yes
Hannah	63	200K	1	No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	Yes?



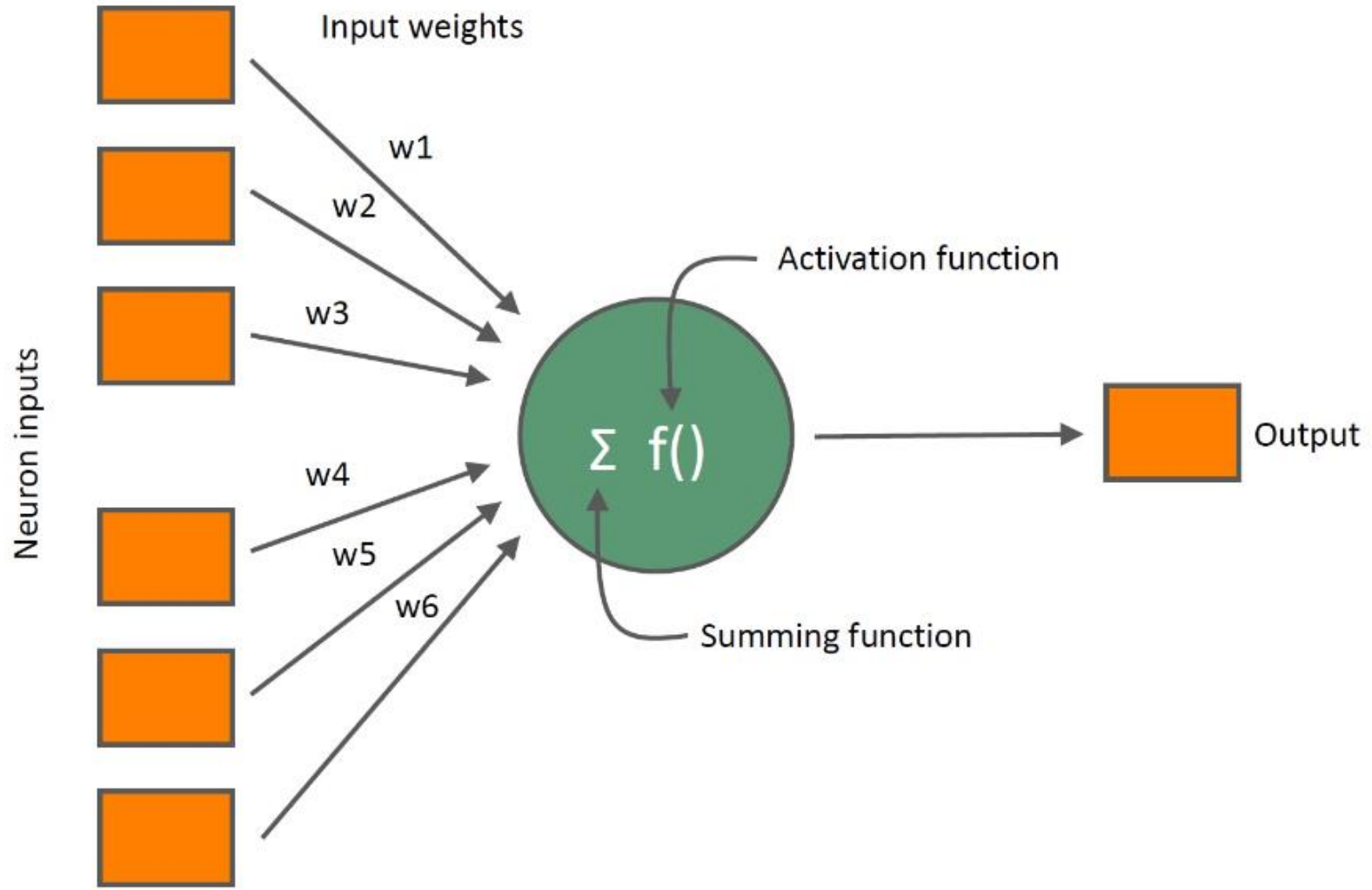
Three nearest ones to David are: No, Yes, Yes

Artificial Neural Network

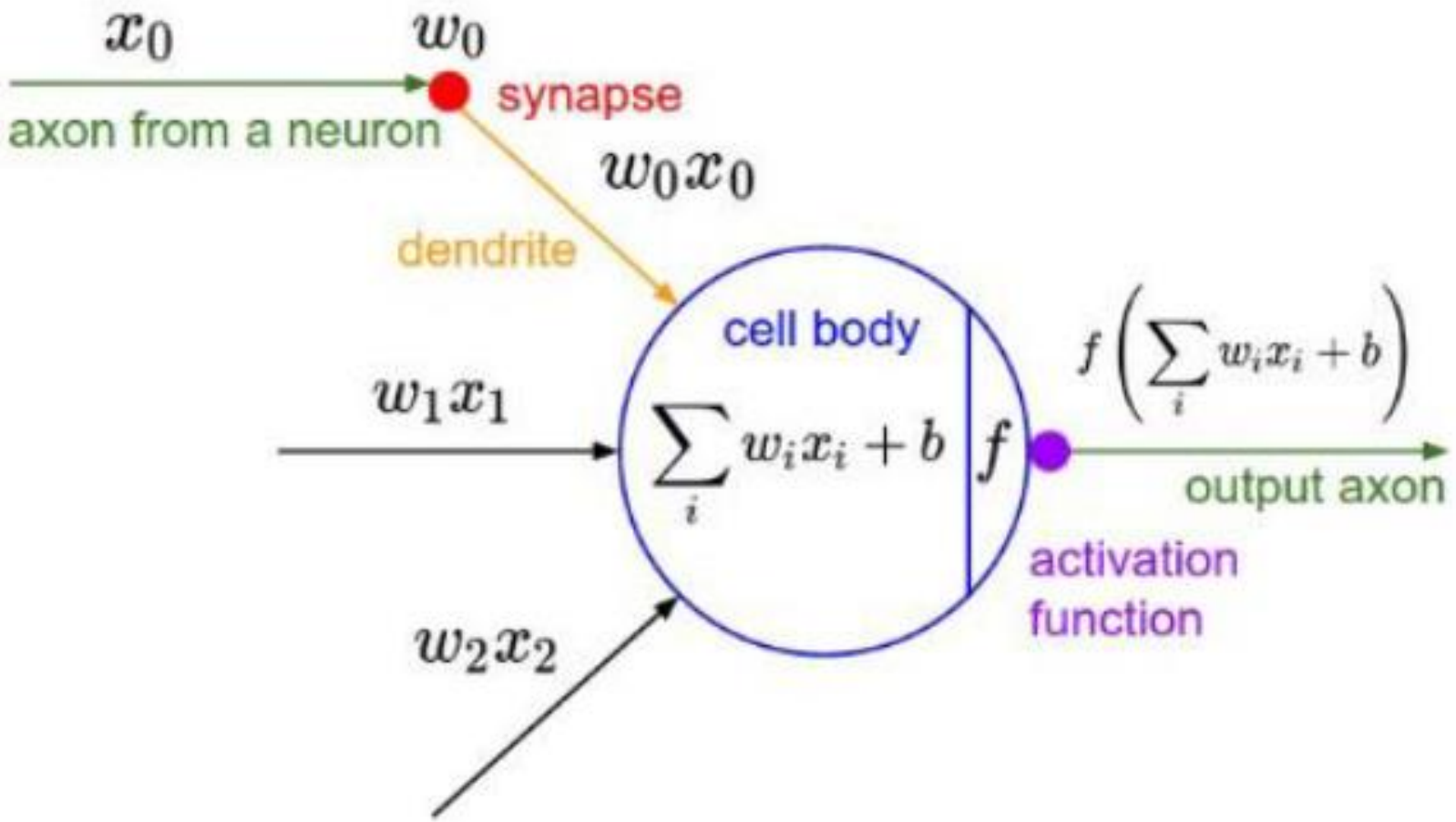
Biological Neuron/ Biological Neural Network



Artificial Neuron/ Artificial Neural Network



Artificial Neuron/ Artificial Neural Network



Artificial Neural Network Layers