K-Nearest Neighbours `KNN`

Seminar in Advanced Topics in Data Analytics

Dr. Shay Horovitz

Dor Romano

Time to Play!

21 Questions

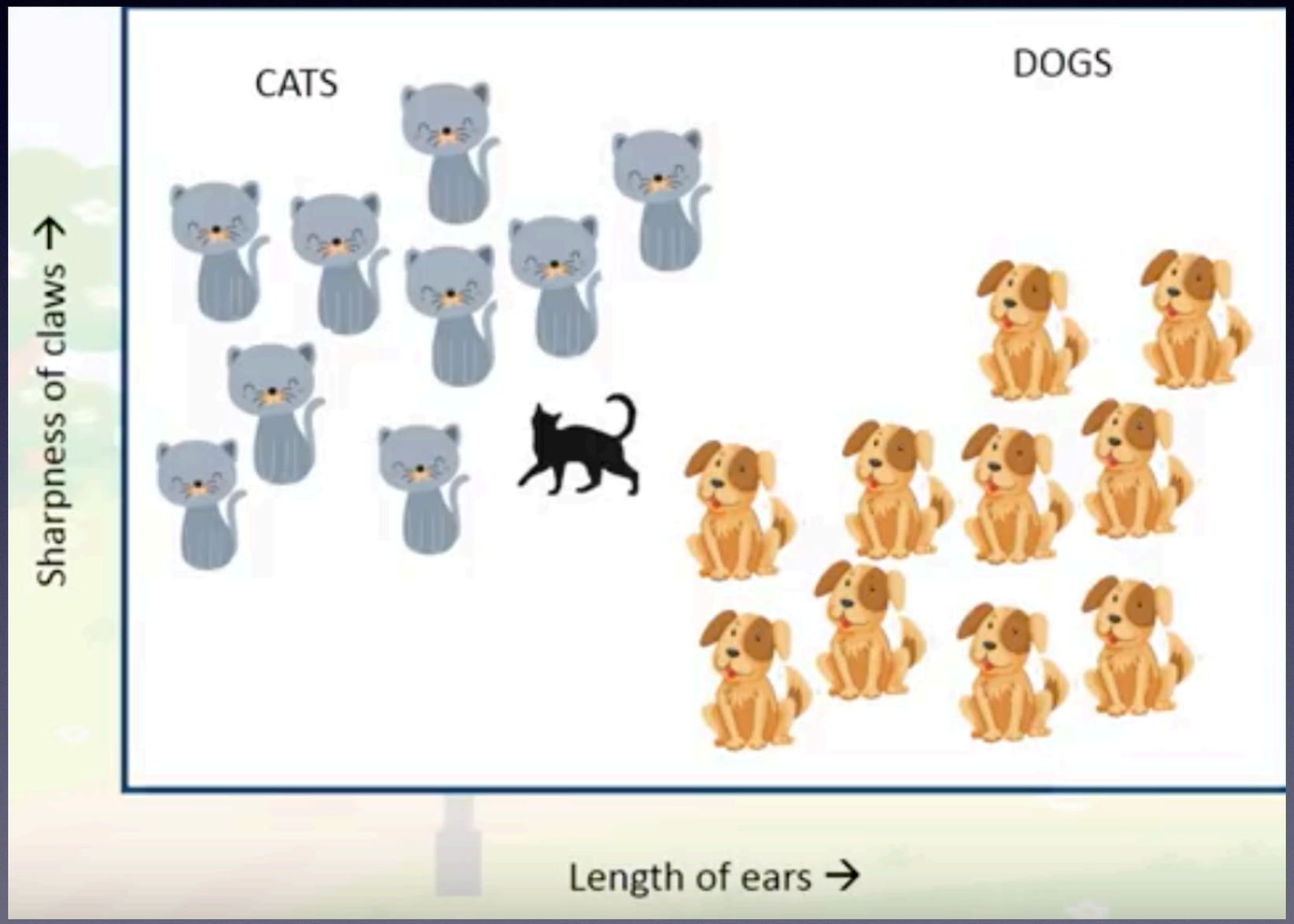
K-Nearest Neighbours
Topics for discussion

- What is it used for?
- Key principles
- How to Choose the factor ‘K’
- Common Uses
- Pros and Cons
- How it works
- Case Study

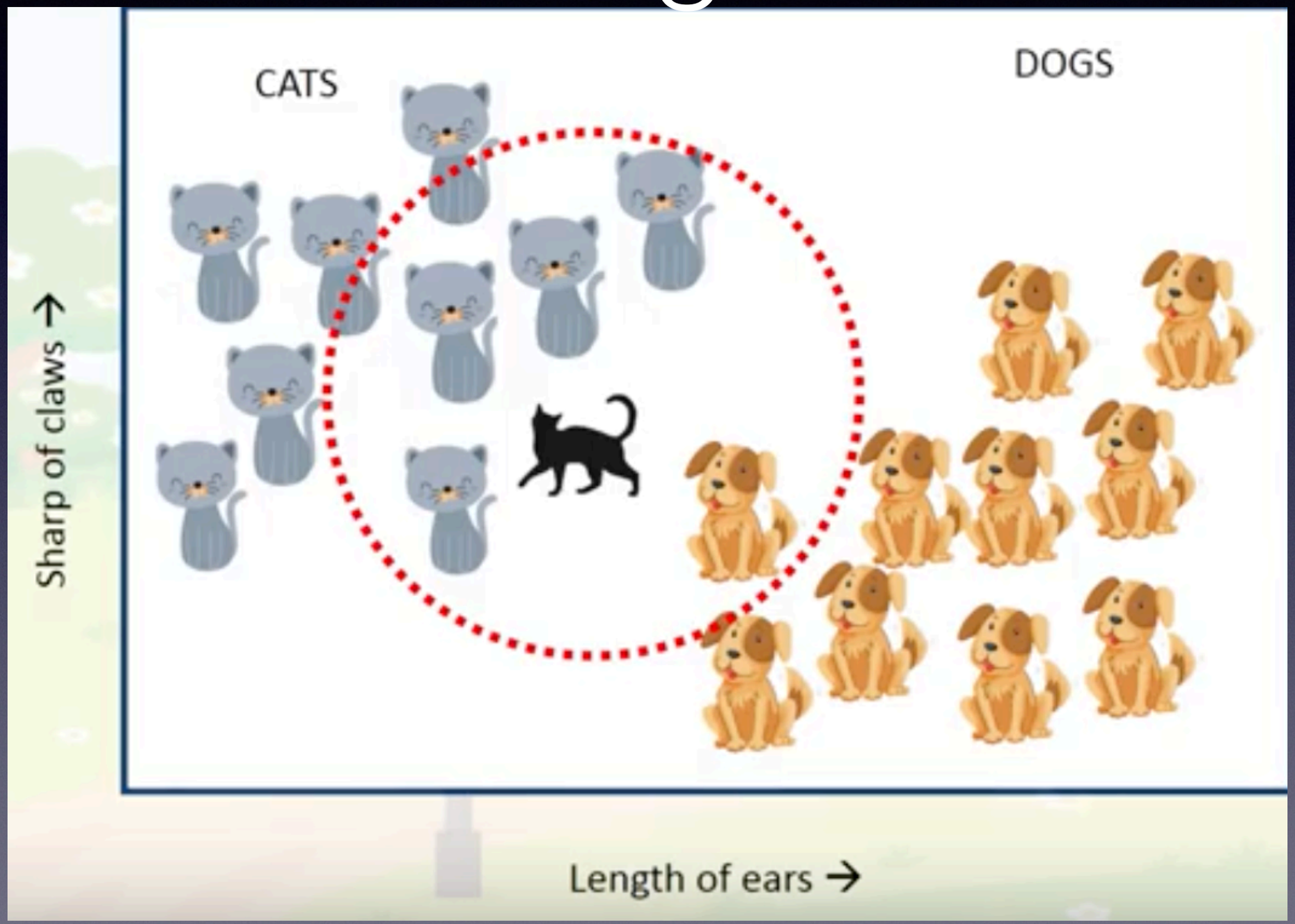
K-Nearest Neighbours
What is it good for?

CATS	DOGS
	
Sharp Claws, uses to climb	Dull Claws
Smaller length of ears	Bigger length of ears
Meows and purrs	Barks
Doesn't love to play around	Loves to run around

K-Nearest Neighbours
What is it good for?



K-Nearest Neighbours
What is it good for?



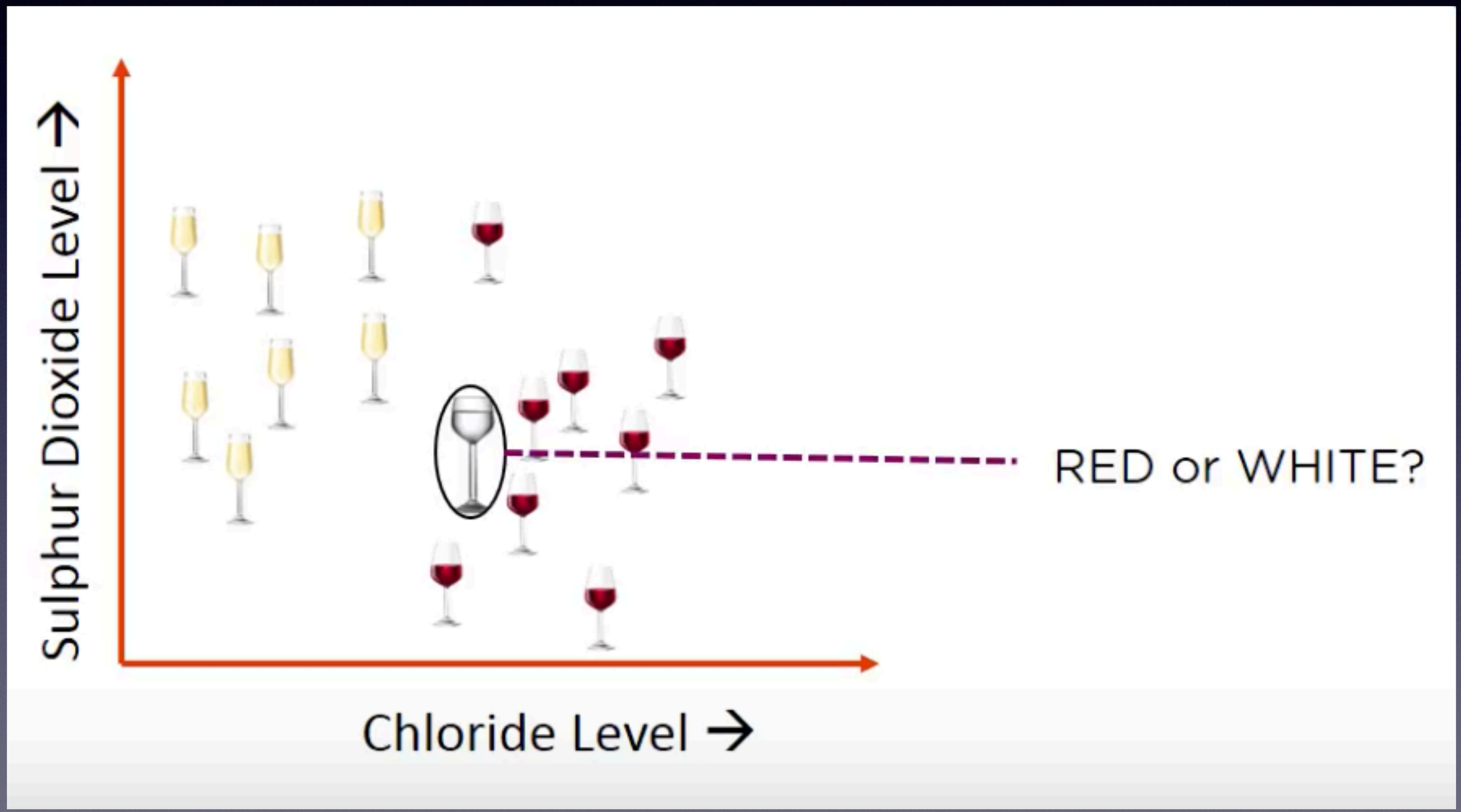
K-Nearest Neighbours
Key Principle

- Classifies data points based on how its neighbours are classified.

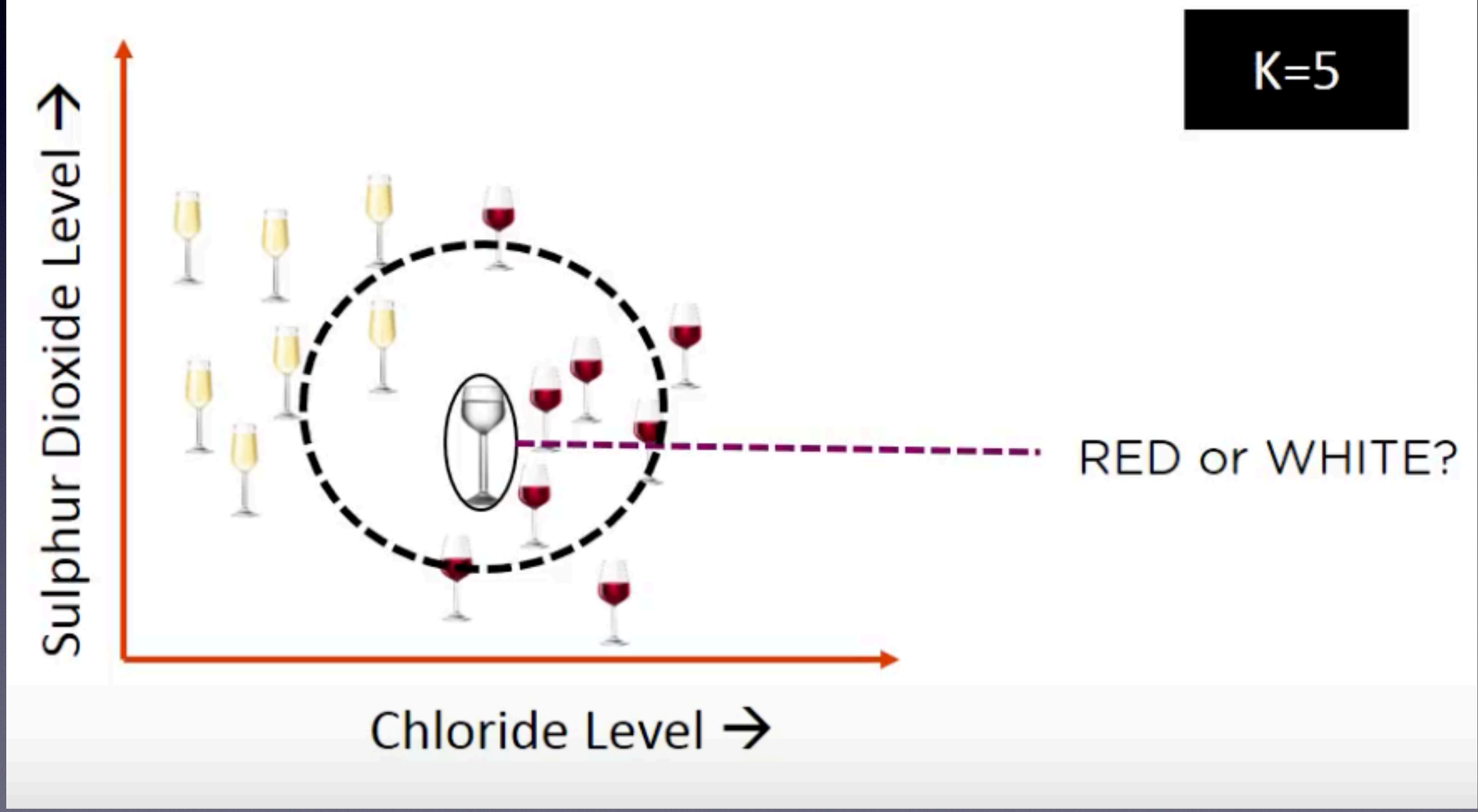
- Stores all available cases and classifies new cases based on a similarity measure.

- *K* is a parameter that refers to the number of nearest neighbours to include in the majority voting process.

K-Nearest Neighbours
What is it good for?

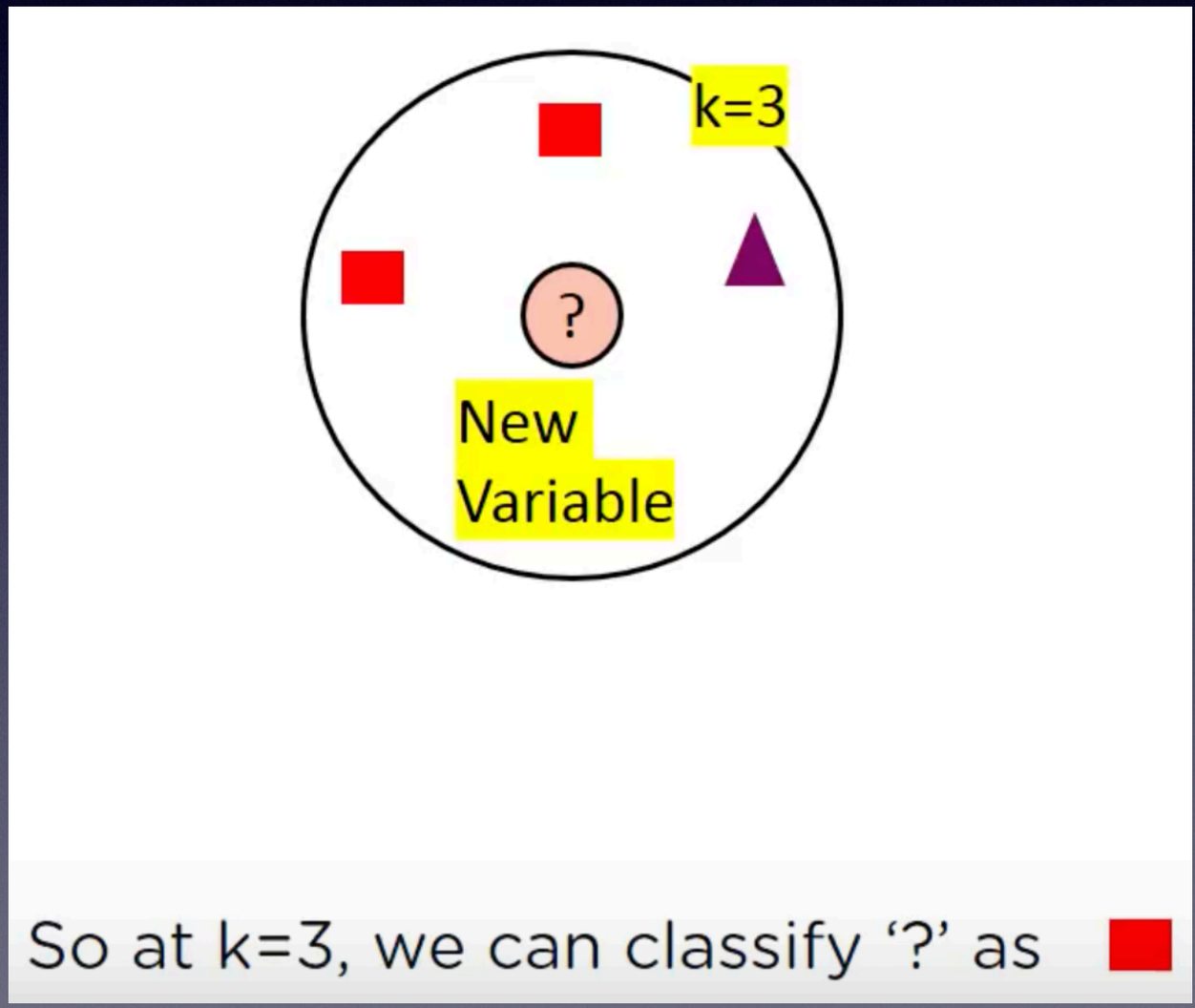


K-Nearest Neighbours
What is it good for?



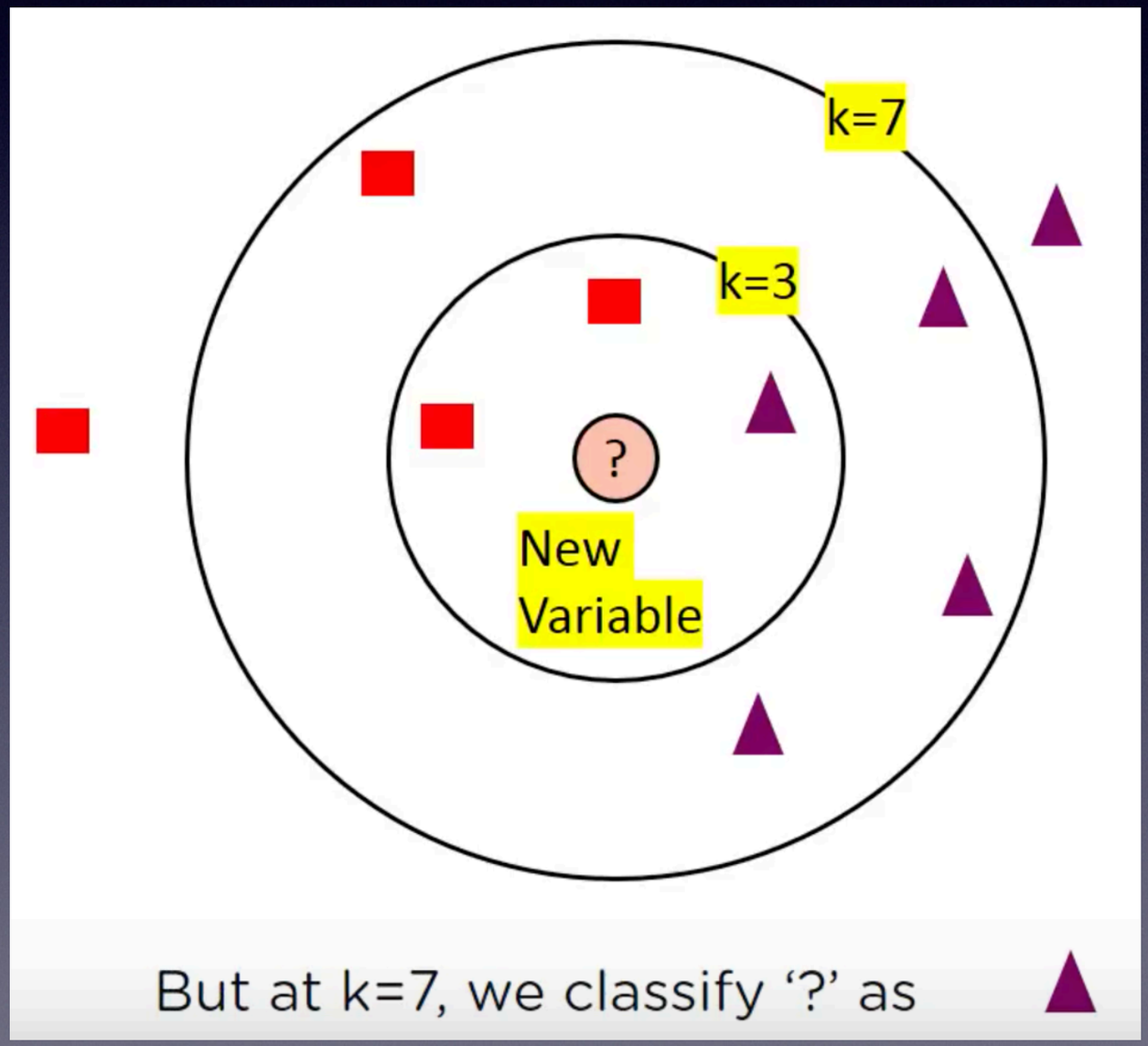
K-Nearest Neighbours

Choosing ‘K’



K-Nearest Neighbours

Choosing ‘K’



K-Nearest Neighbours

Choosing ‘K’

Sqrt(n), Where `n` is the total number of data points

Odd value of *k* is selected to avoid confusion between two classes of data

K-Nearest Neighbours
Common Uses

- Labeled Data
- Noise free Data
- Small Dataset

K-Nearest Neighbours
Pros and Cons

Pros:

- Very simple.
- Training is trivial.
- Works with any number of classes.
- In case you would like to add more data, the algorithm is easy to use.
- Contains a few parameters:
  - k
  - Distance Metric

Cons:

- No data points border line.
- High prediction cost (worse for large data sets).
- Not good with high dimensional data.
- Categorical features don’t work well.

K-Nearest Neighbours

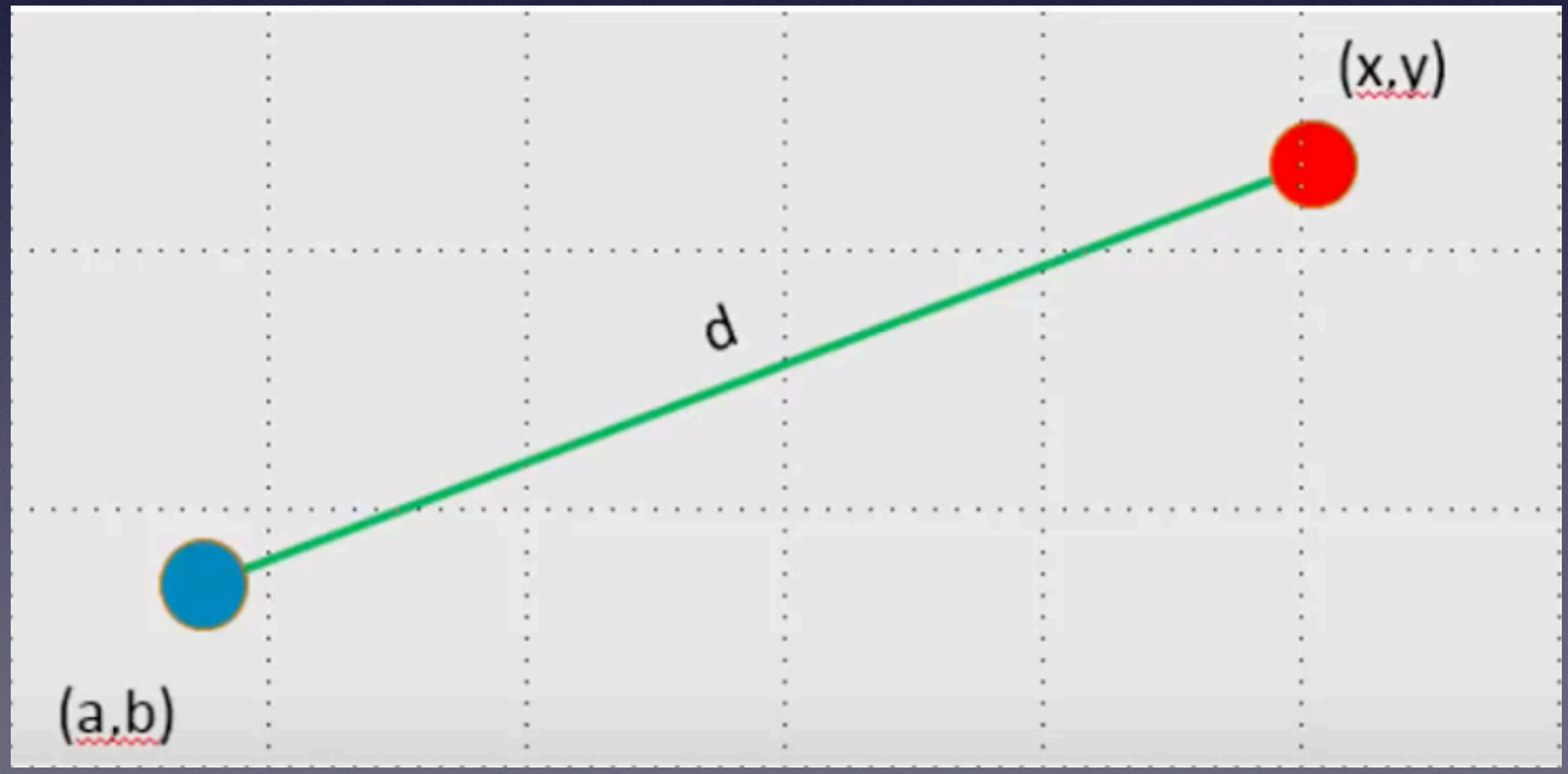
How it works

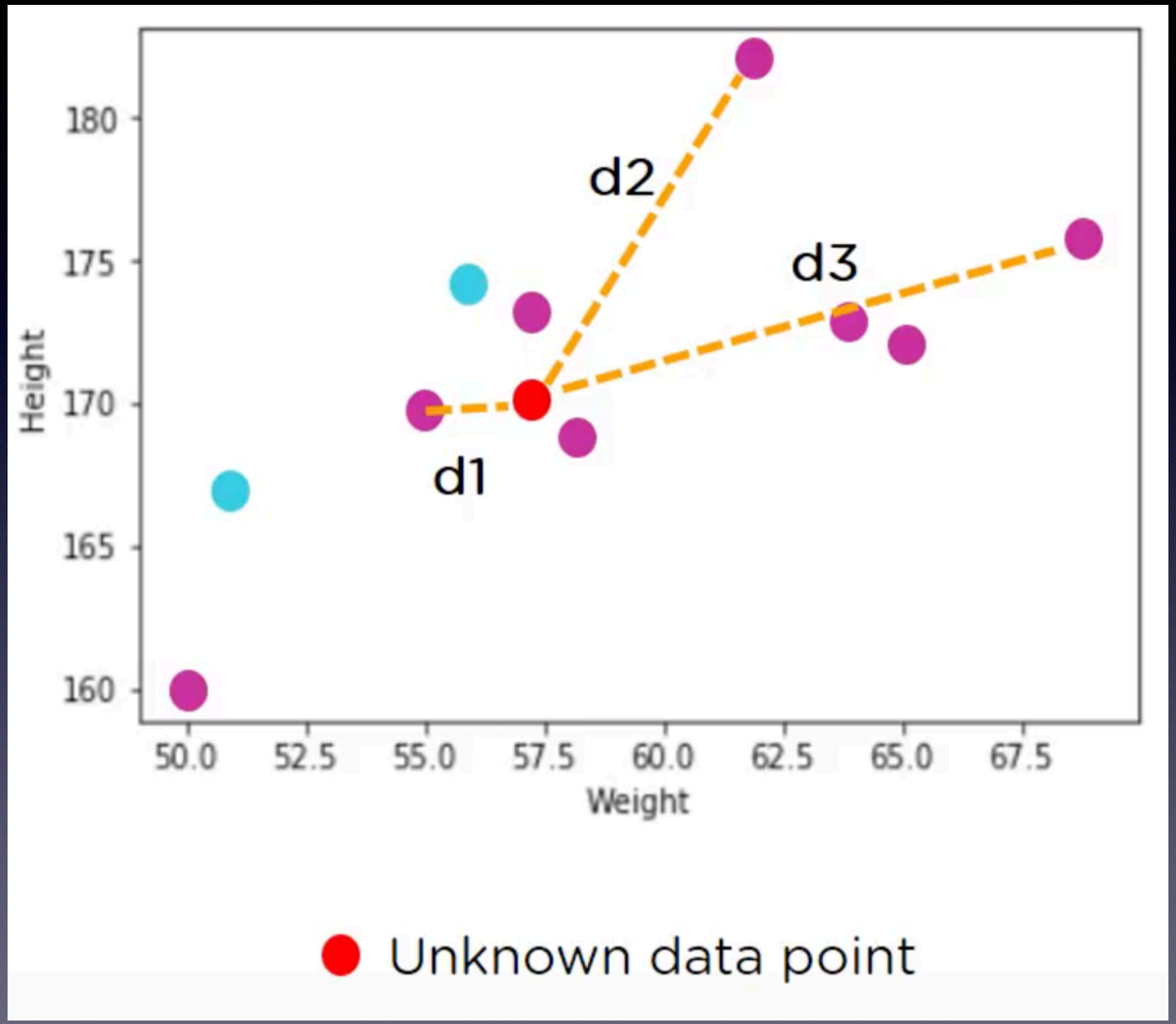
Weight(x2)	Height(y2)	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

K-Nearest Neighbours
How it works

- Calculate the Euclidean Distance

$$dist(d) = \sqrt{(x - a)^2 + (y - b)^2}$$

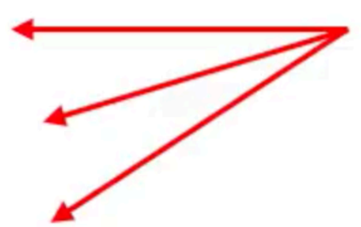



180
175
170
165
160
Height
d1
d2
d3
50.0
52.5
55.0
57.5
60.0
62.5
65.0
67.5
Weight
Unknown data point

K-Nearest Neighbours

How it works

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



k = 3

K-Nearest Neighbours
Case Study

- Example for Notebook

Questions?

Thank You!