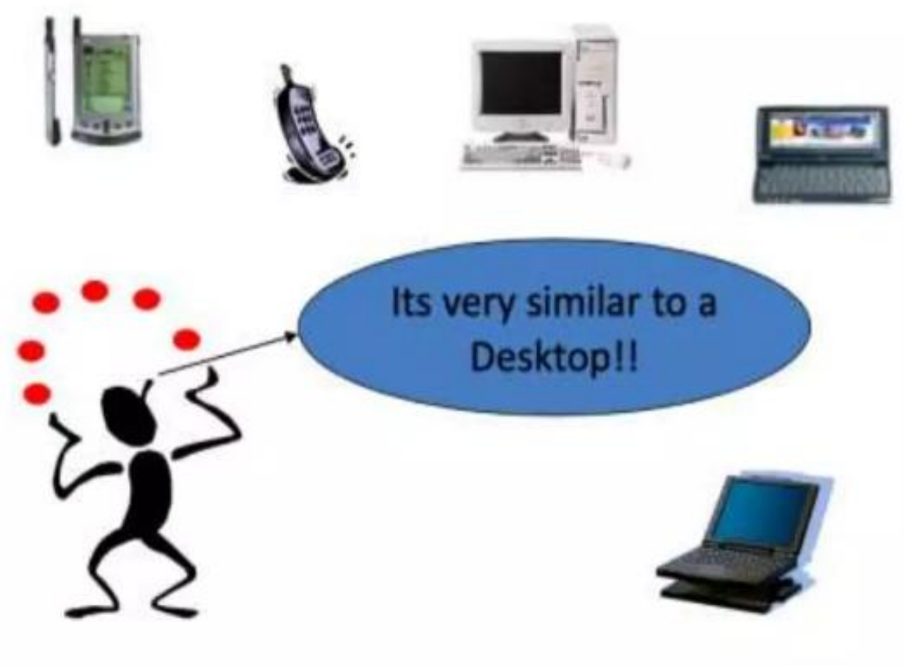
K Nearest Neighbor Classification



Instance-Based Learning

KNN: Alternate Terminologies

- Instance Based Learning
- Lazy Learning
- Case Based Reasoning
- Exemplar Based Learning

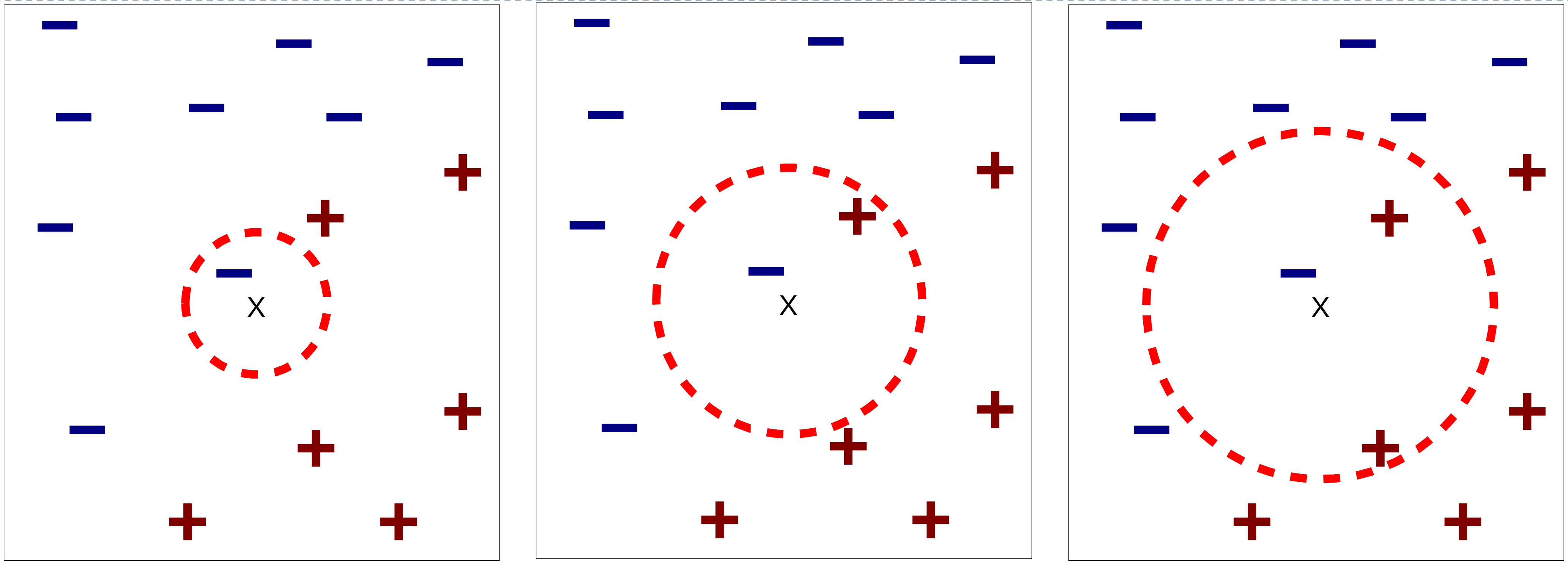
What is k-NN?

- A powerful classification algorithm used in pattern recognition.
- K- nearest neighbors stores all available cases and classify new cases based on similarity measure(eg. Distance function)
- It is one of the top data mining algorithm used today.
- A non-parametric lazy learning algorithm (instance based learning method)

Basic Idea

- *k*-NN classification rule is to assign to a test sample the majority category label of its *k* nearest training samples
- In practice, *k* is usually chosen to be odd, so as to avoid ties
- The $k = 1$ rule is generally called the nearest-neighbor classification rule

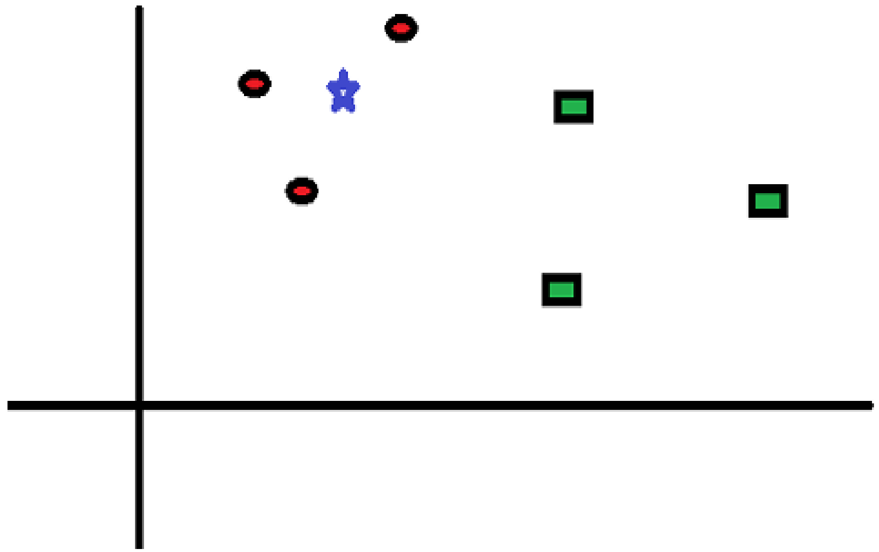
Definition of Nearest Neighbor

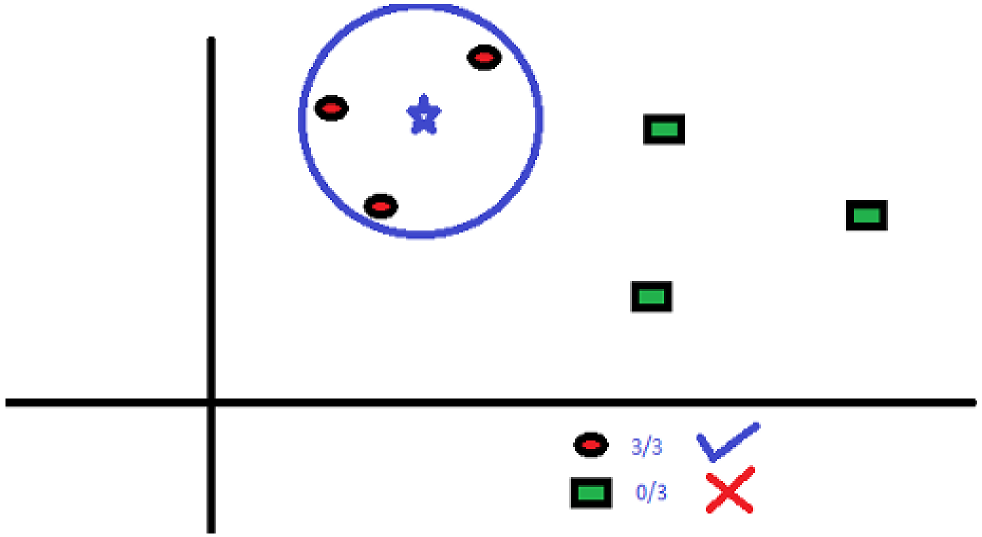


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

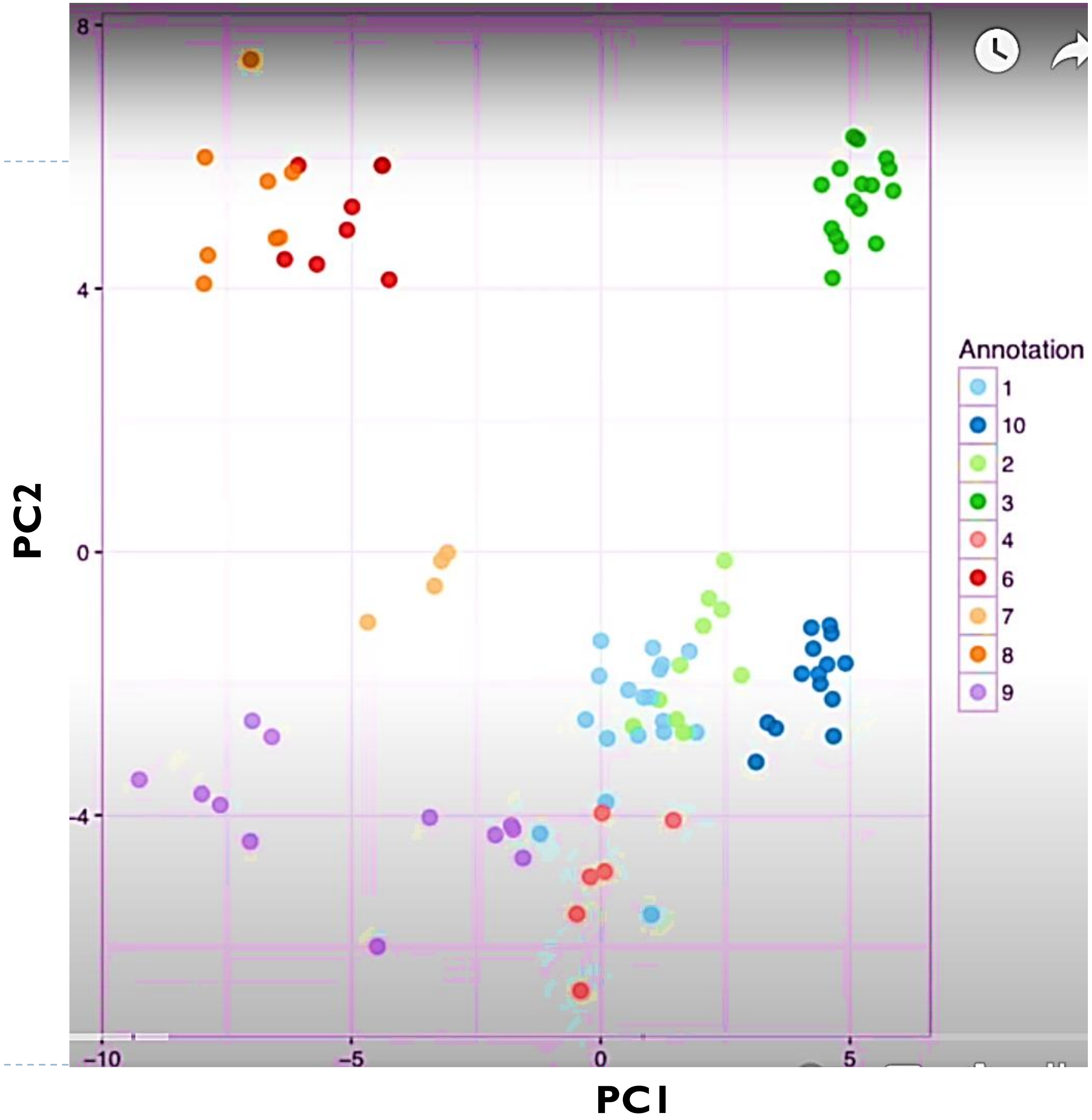
Algorithm



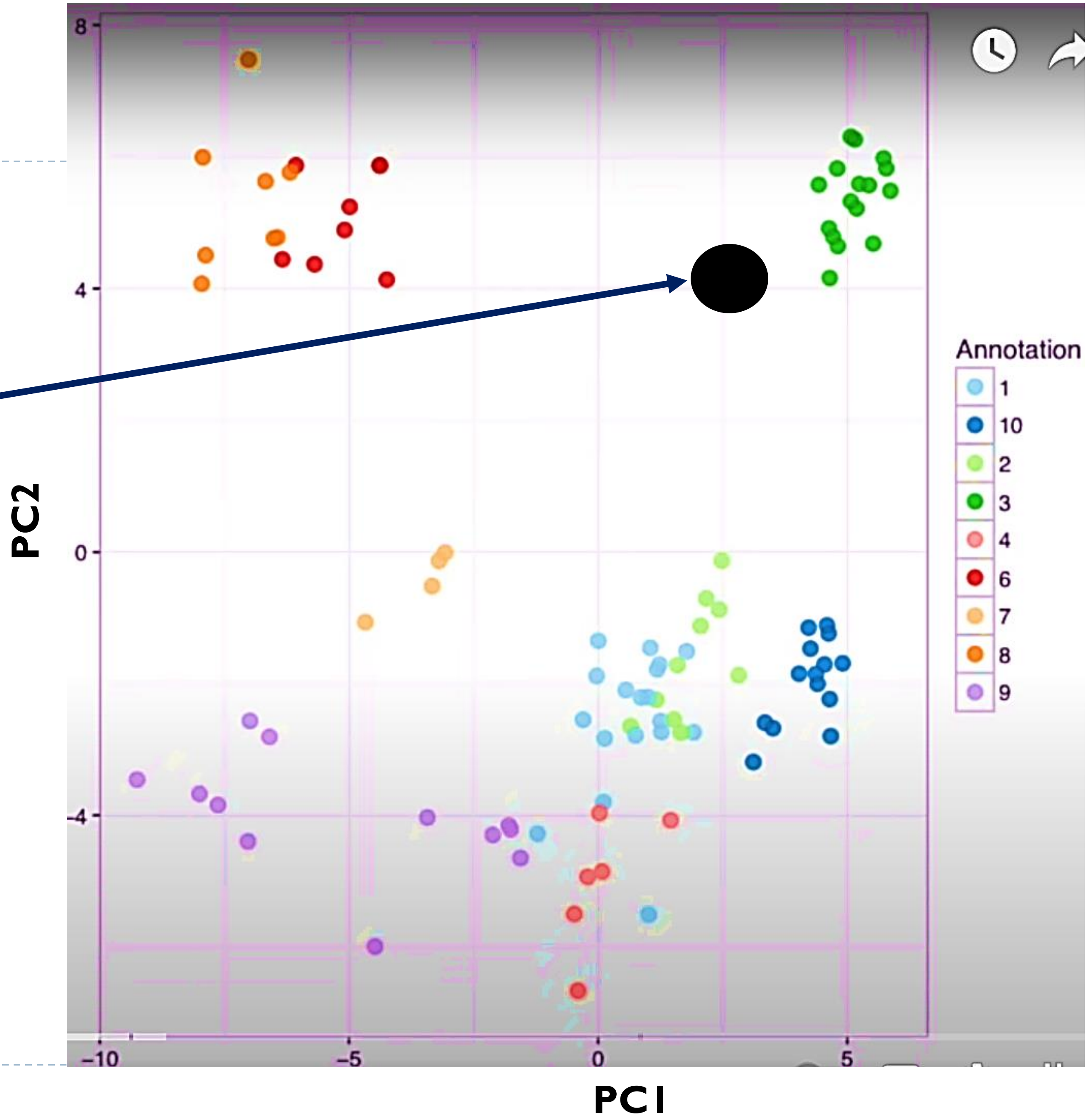


Step1:

- Start with a dataset with known categories.
- Cluster the data

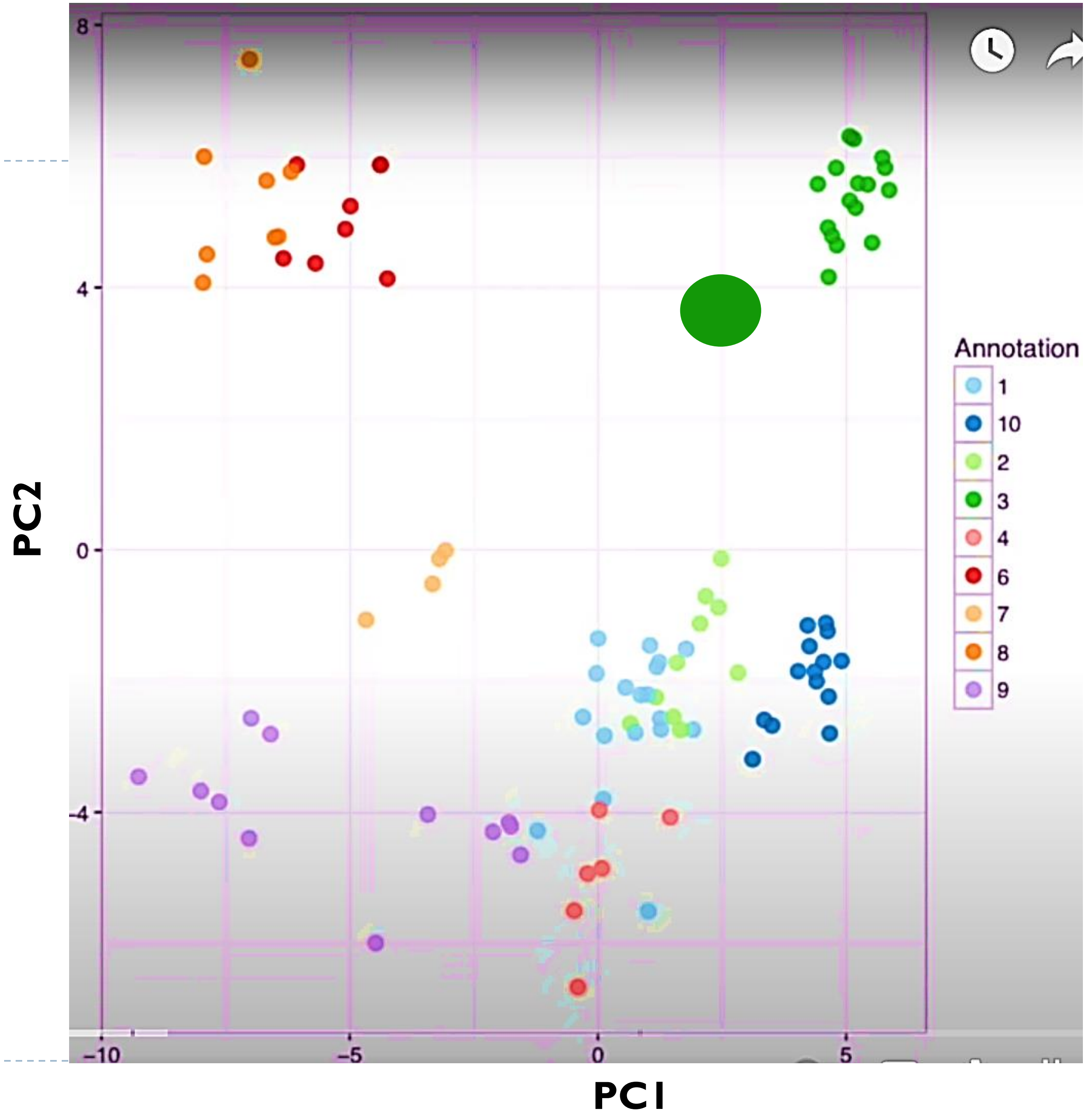


Step2:
Add a new sample with unknown category



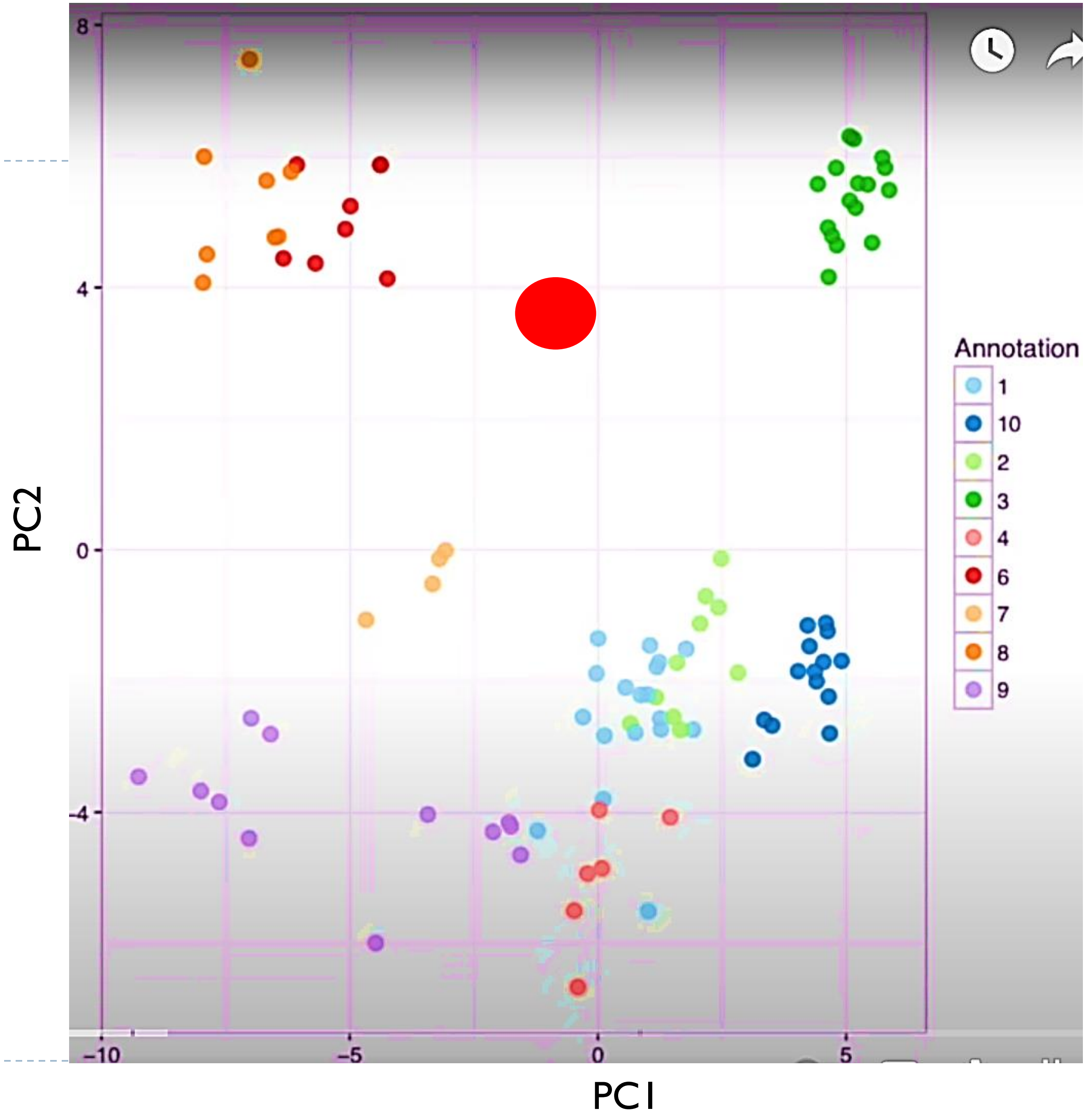
Step3:

- Classify the new sample based on the nearest annotated cell, k-NN
- If k=1, only one nearest neighbour is used
- In this case, the category is **GREEN**
- If k=11, 11 nearest neighbours will be used and the category is still **GREEN**



- k = 11, a new sample is mid-way between red and green
- Pick the category that "gets the most votes"
- In this case:
- 7 nn are **RED**
- 3 nn are **ORANGE**
- 1 nn is **GREEN**
- Most votes are for **RED,**

So, final assignment is **RED**



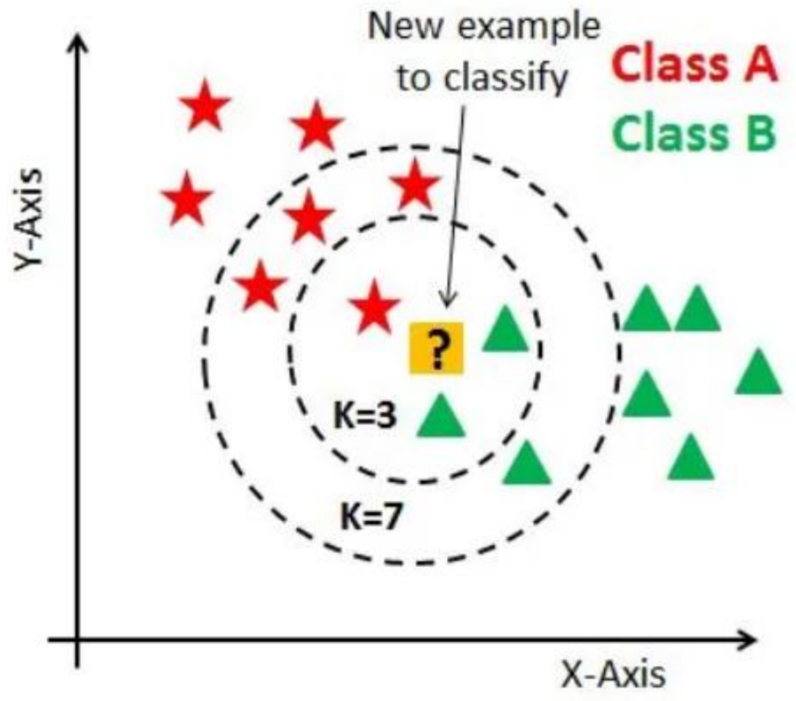
Nearest-Neighbor Classifiers: Issues

- The value of *k*, the number of nearest neighbors to retrieve
- Choice of Distance Metric to compute distance between records
- Computational complexity
  - Size of training set
  - Dimension of data

Value of K

- 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

Rule of thumb:
K = sqrt(N)
N: number of training points



How to choose k?

- There are no pre-defined statistical methods to find the most favorable value of K.
- Initialize a random K value and start computing.
- Choosing a small value of K leads to unstable decision boundaries.
- The substantial K value is better for classification as it leads to smoothening the decision boundaries.
- Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.

Distance Metrics

Minkowsky:
$$D(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r}$$

Euclidean:
$$D(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan / city-block:
$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

Camberra:
$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Chebychev:
$$D(\boldsymbol{x},\boldsymbol{y}) = \max_{i=1}^{m} |x_i - y_i|$$

Quadratic:
$$D(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^T Q (\boldsymbol{x}-\boldsymbol{y}) = \sum_{j=1}^{m} \left(\sum_{i=1}^{m} (x_i - y_i) q_{ji} \right) (x_j - y_j)$$

Q is a problem-specific positive definite $m \times m$ weight matrix

Mahalanobis:
$$D(\boldsymbol{x},\boldsymbol{y}) = [\det V]^{1/m} (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}} V^{-1} (\boldsymbol{x}-\boldsymbol{y})$$

V is the covariance matrix of $A_1..A_m$, and A_j is the vector of values for attribute j occuring in the training set instances $1..n$.

Correlation:
$$D(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{m} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x_i})^2 \sum_{i=1}^{m} (y_i - \overline{y_i})^2}}$$

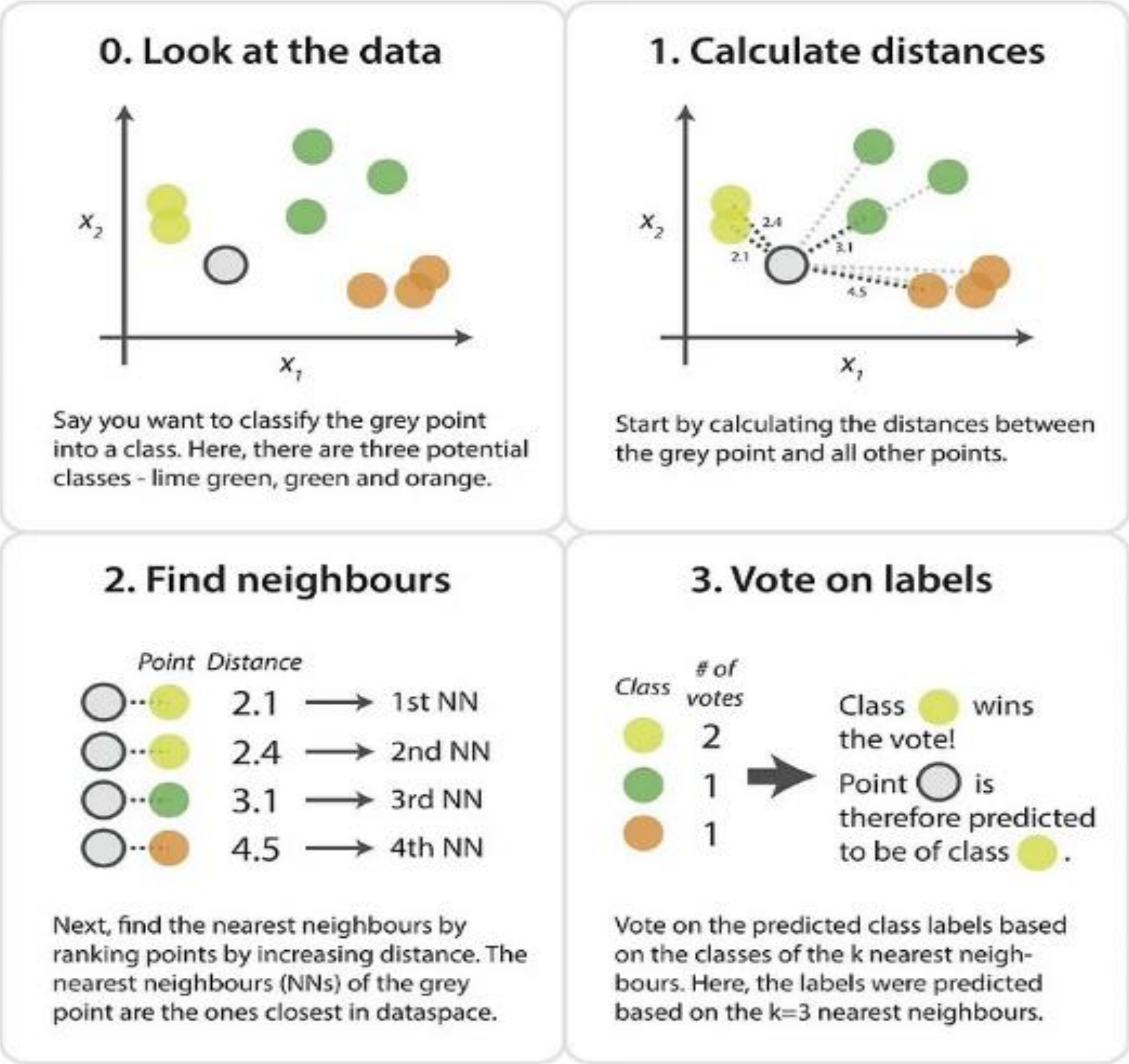
$\overline{x_i} = \overline{y_i}$ and is the average value for attribute i occuring in the training set.

Chi-square:
$$D(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \frac{1}{sum_i} \left(\frac{x_i}{size_{\boldsymbol{x}}} - \frac{y_i}{size_{\boldsymbol{y}}} \right)^2$$

sum_i is the sum of all values for attribute i occuring in the training set, and $size_x$ is the sum of all values in the vector $\boldsymbol{x}$.

Kendall's Rank Correlation:
$$D(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{i-1} \mathrm{sign}(x_i - x_j)\,\mathrm{sign}(y_i - y_j)$$

sign(x)=-1, 0 or 1 if $x < 0$, $x = 0$, or $x > 0$, respectively.


0. Look at the data
x_2
x_1
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.
1. Calculate distances
x_2
2.4
2.1
3.1
4.5
x_1
Start by calculating the distances between the grey point and all other points.
2. Find neighbours
Point Distance
2.1 → 1st NN
2.4 → 2nd NN
3.1 → 3rd NN
4.5 → 4th NN
Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.
3. Vote on labels
Class
# of votes
2
1
1
Class wins the vote!
Point is therefore predicted to be of class .
Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Problem 1

X1 (acid durability, seconds)	X2 (strength, kg/sq.meter)	Y (Classification)
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

$(7-3)^2 +$	$(7-7)^2$			
$(7-3)^2 +$	$(4-7)^2$			
$(3-3)^2 +$	$(4-7)^2$			
$(1-3)^2 +$	$(4-7)^2$			

$(7-3)^2 +$	$(7-7)^2$	16		
$(7-3)^2 +$	$(4-7)^2$	25		
$(3-3)^2 +$	$(4-7)^2$	9		
$(1-3)^2 +$	$(4-7)^2$	13		

$(7-3)^2$ +	$(7-7)^2$	16	3	
$(7-3)^2$ +	$(4-7)^2$	25	4	
$(3-3)^2$ +	$(4-7)^2$	9	1	
$(1-3)^2$ +	$(4-7)^2$	13	2	

k=3

$(7-3)^2$ +	$(7-7)^2$	16	3	Bad
$(7-3)^2$ +	$(4-7)^2$	25	4	X
$(3-3)^2$ +	$(4-7)^2$	9	1	Good
$(1-3)^2$ +	$(4-7)^2$	13	2	Good

k=3

$(7-3)^2$ +	$(7-7)^2$	16	3	Bad
$(7-3)^2$ +	$(4-7)^2$	25	4	X
$(3-3)^2$ +	$(4-7)^2$	9	1	Good
$(1-3)^2$ +	$(4-7)^2$	13	2	Good