Clustering

Sayan Ranu

Nick Mckeown Chair Associate Professor

Department of Computer Science and Engineer

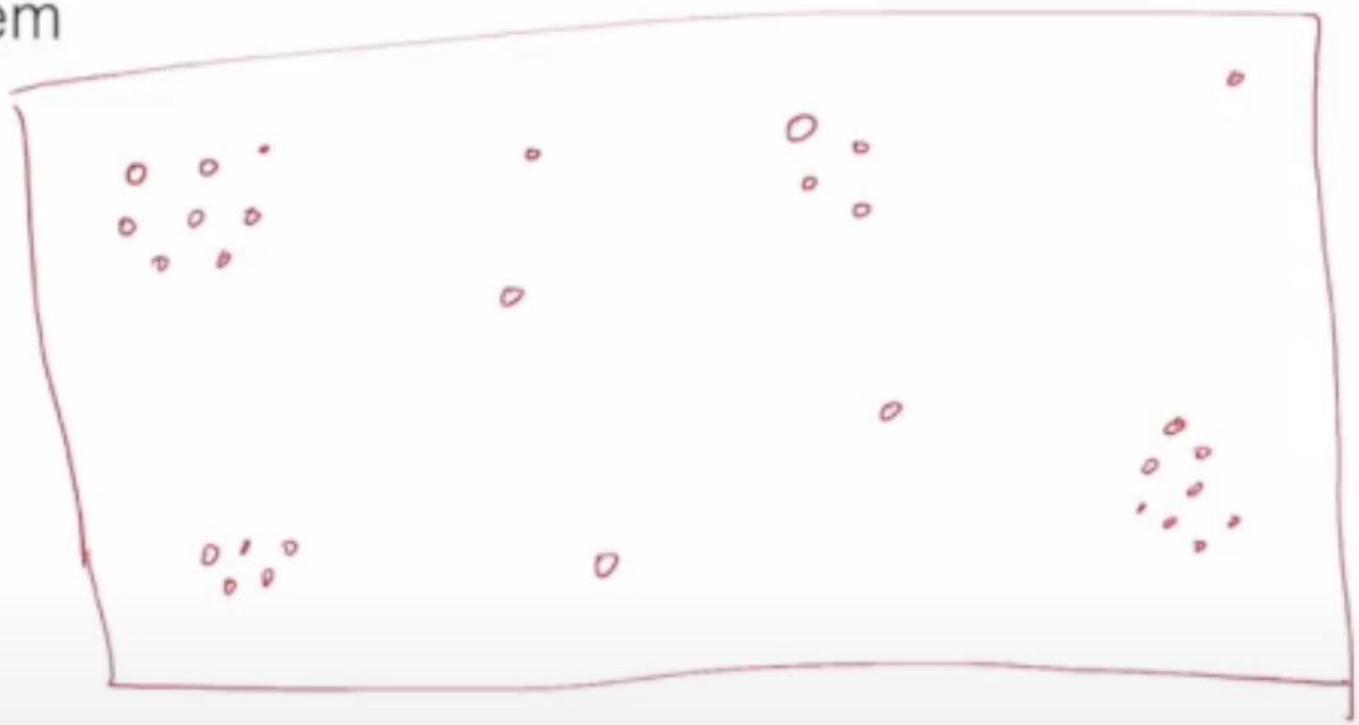
School of AI

IIT Delhi

Please ask questions!

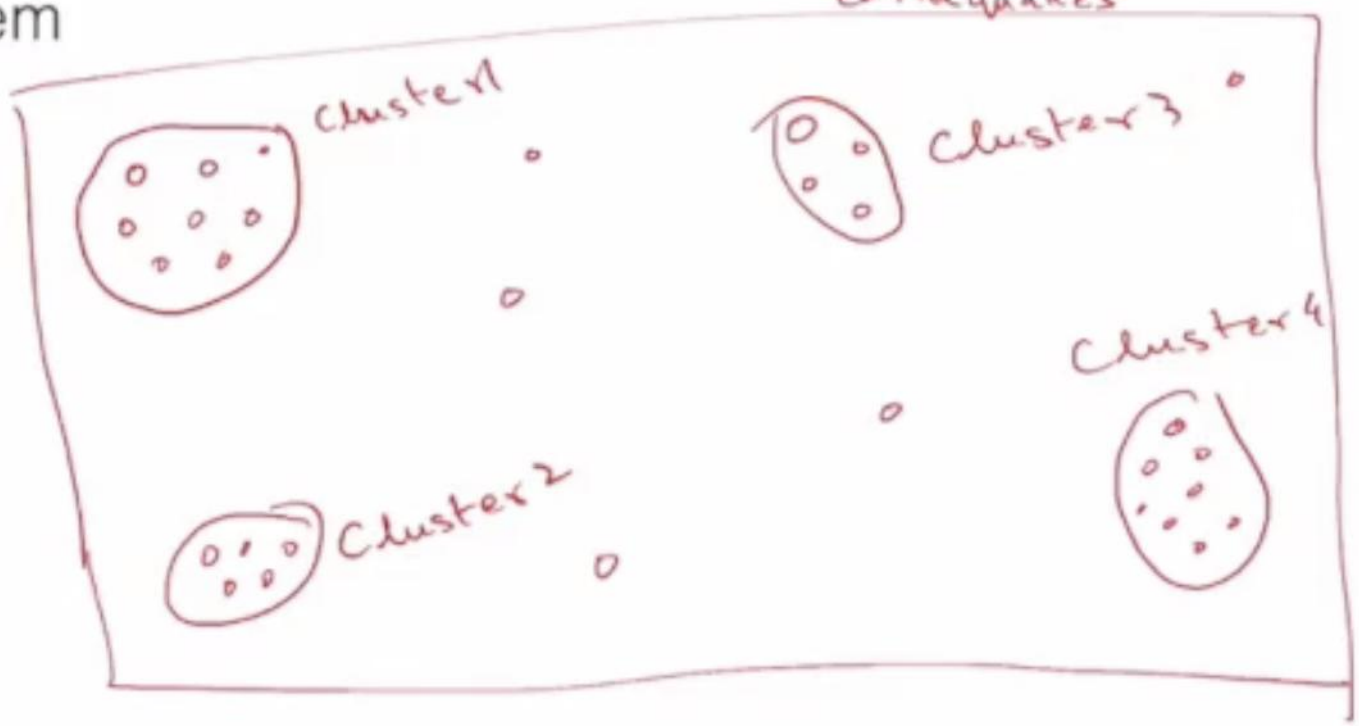
All the airports in that spa

The problem



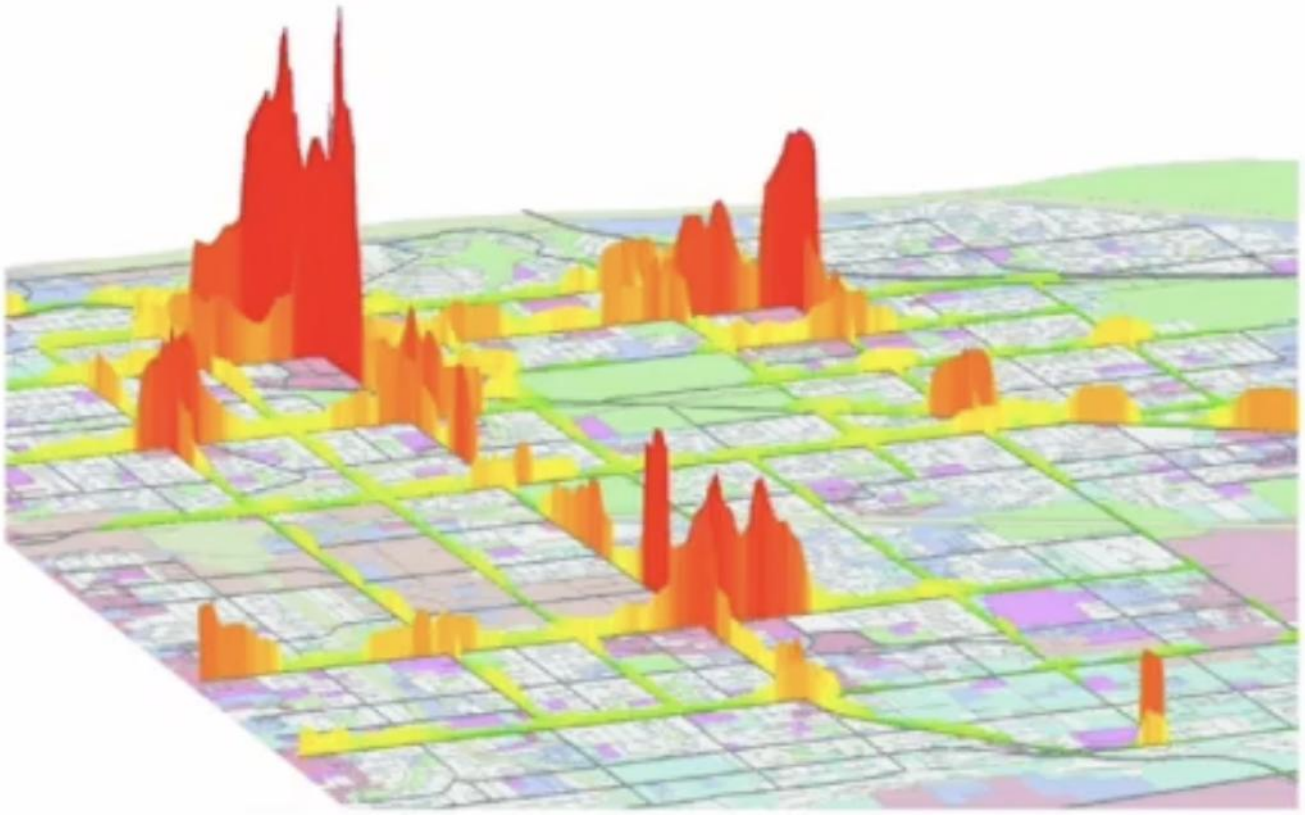
The problem

All the airports in that space
crime event
earthquakes

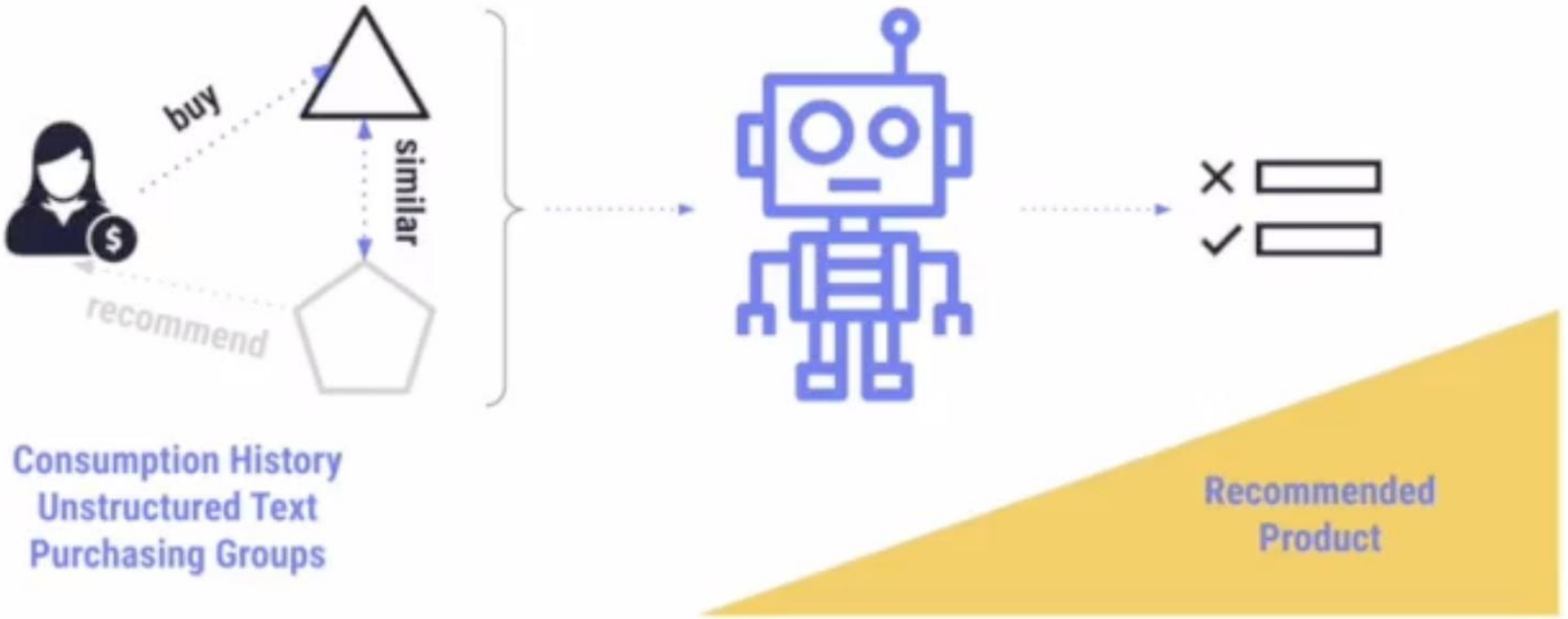


point (x, y)
(lat, long)

Crime Hotspots

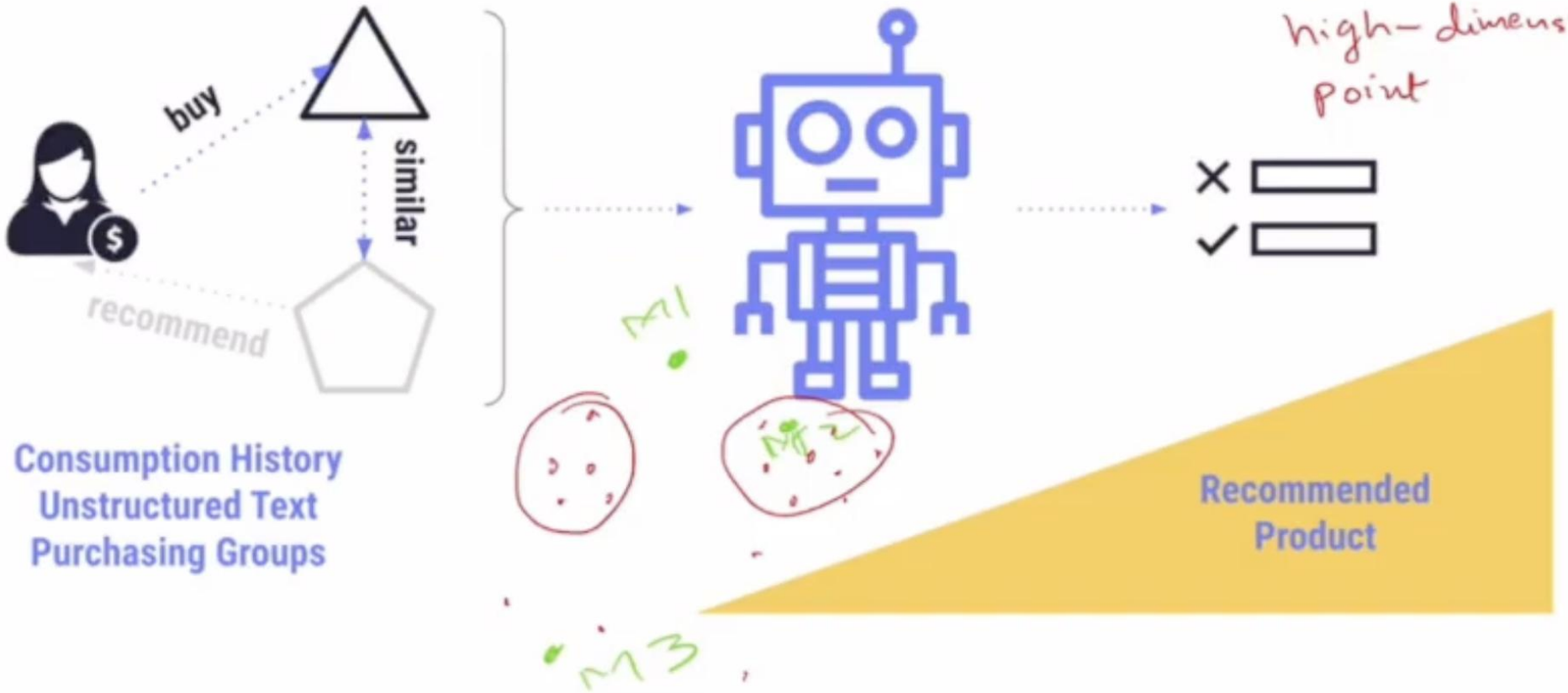


Recommendation Engines



Youtube, Netflix, Amazon Prime

Recommendation Engines



Length
Language
video → [ ]
high-dimensional
point

Market Segmentation

Supervised

Car → BMW

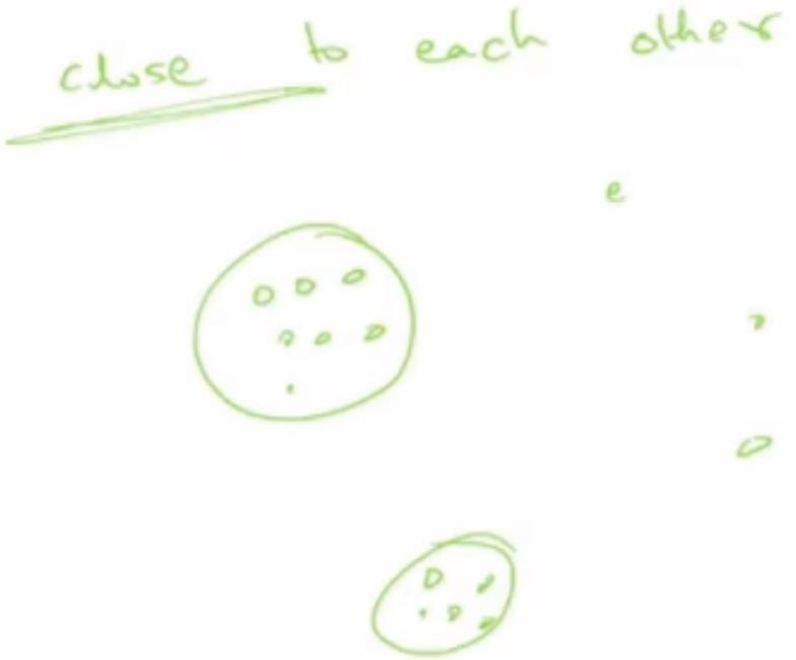
Customer →

[↓ ↓ ↓ ↓]

Age Salary Education

What do we need to cluster data?

1. Dataset of Points
2. Distance Function
   - Typical to assume a *metric* distance function



What do we need to cluster data?

close to each other

distance in km / time

1. Dataset of Points
2. Distance Function
   - Typical to assume a *metric* distance function

Triangular

(1) $d(x,y) = 0$ if $x = y$

> 0 if $x \neq y$

(2) $d(x,y) = d(y,x)$

(3) $d(x,y) \leq d(x,z) + d(z,y)$

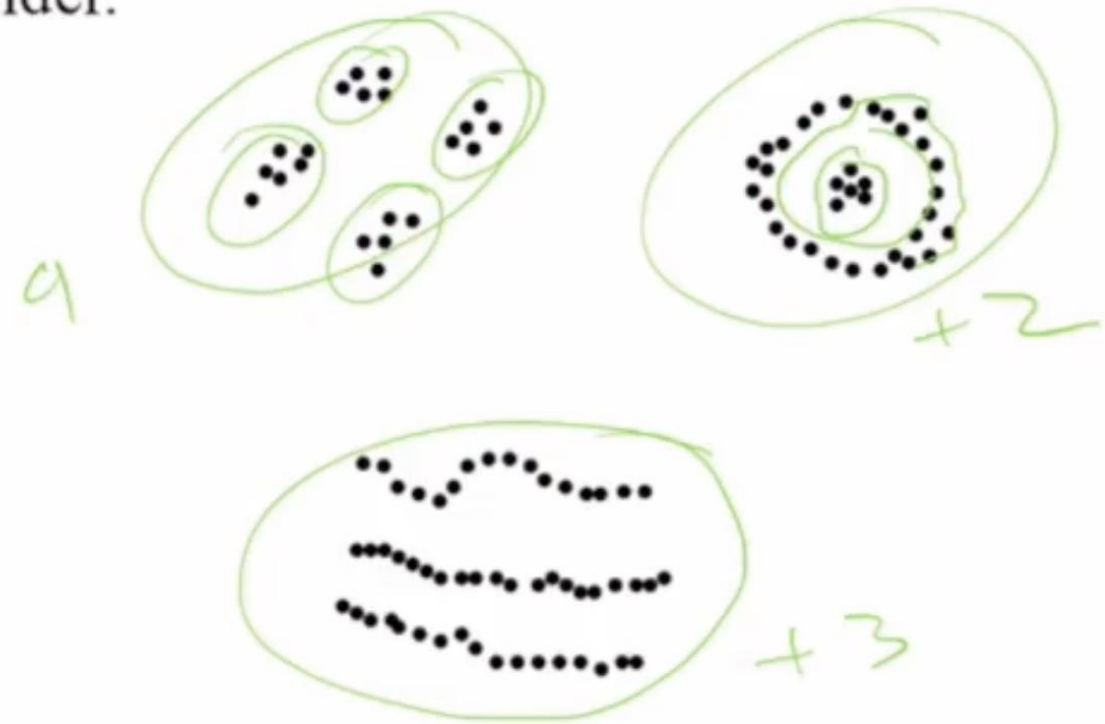
Euclidean distance

$P_1 = (x_1, y_1)$

$P_2 = (x_2, y_2)$

$d(P_1, P_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Clustering is an inherently ill-defined problem since the correct clusters depend upon context and are in the eye of the beholder.



How many clusters are there?

The *K-Means* Clustering Method

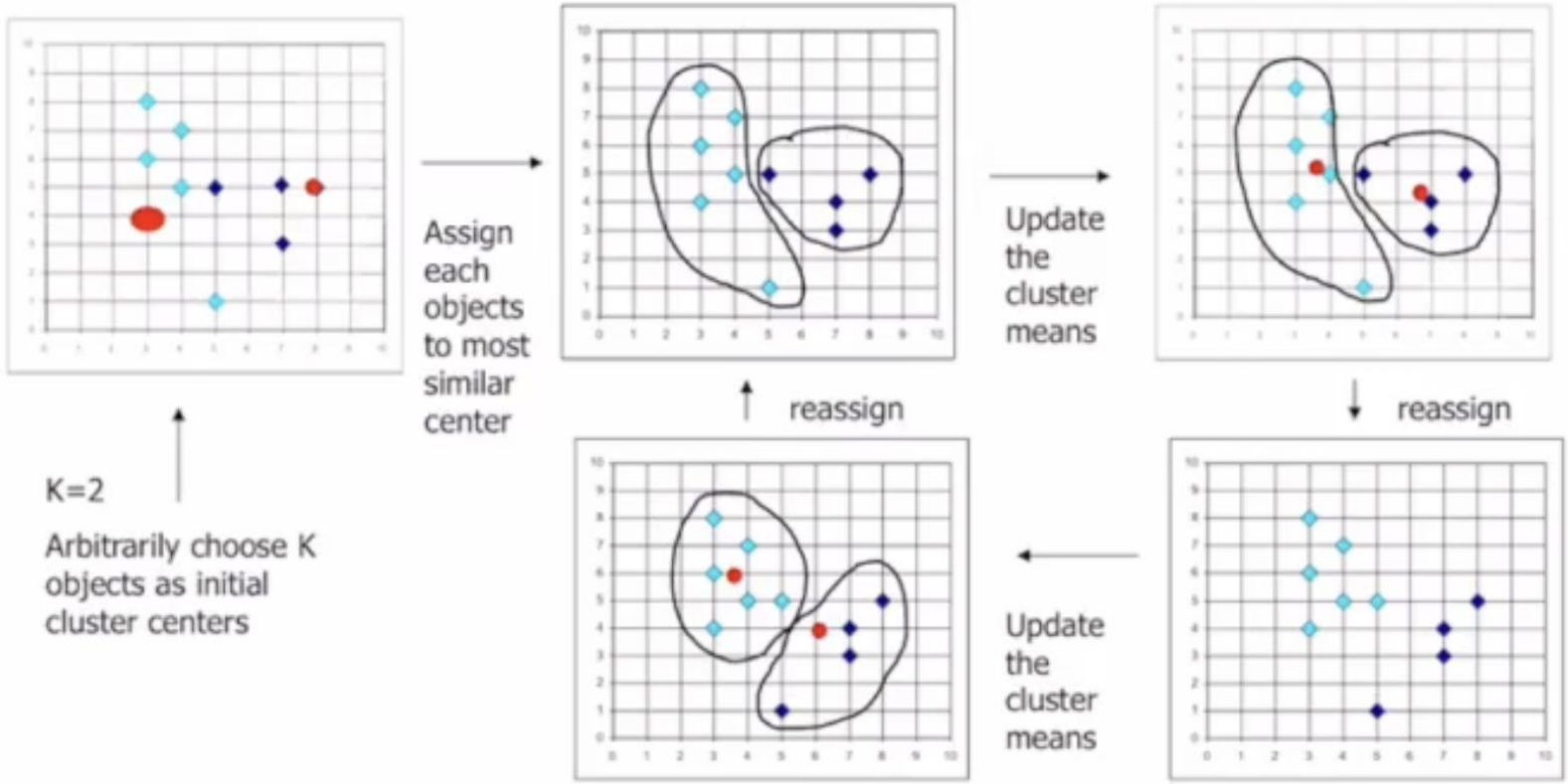
- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment

The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects into *k* nonempty subsets
  - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when no more new assignment

The *K-Means* Clustering Method

- Example

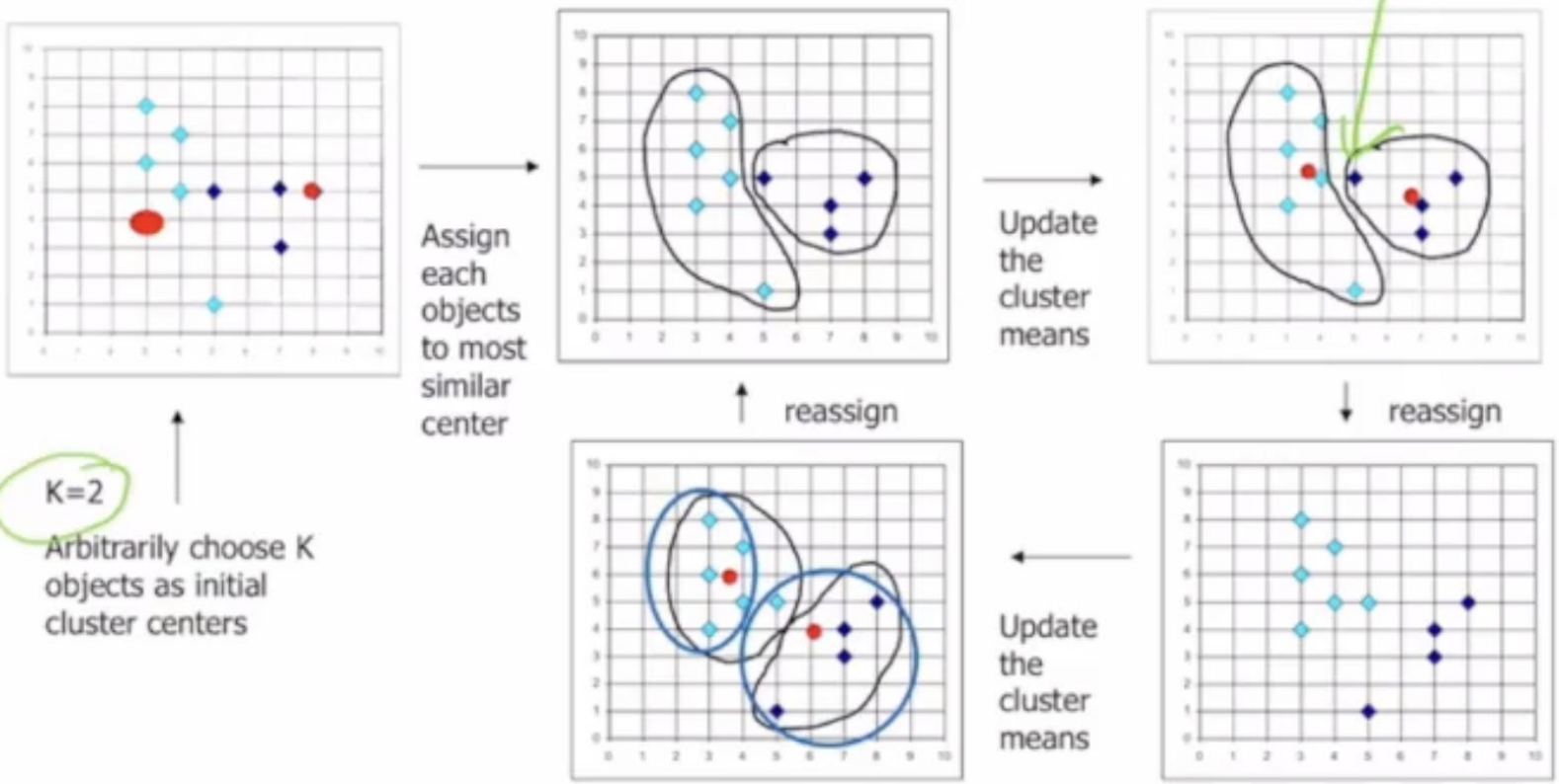


Time: Linear to the no. of points in the dataset

Saturation / Convergence

The *K-Means* Clustering Method

- Example

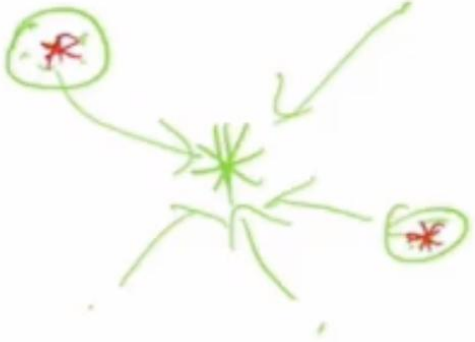


S

Properties of k-means

- How do you select the initial k clusters?
  - This is a black art.
  - Just apply your heuristic and hope it works
    - Choose diverse centers
- How do you find out the value of k?
  - Identify the value of k where cluster quality is best
  - Heuristic: Elbow Plot

Elbow plot

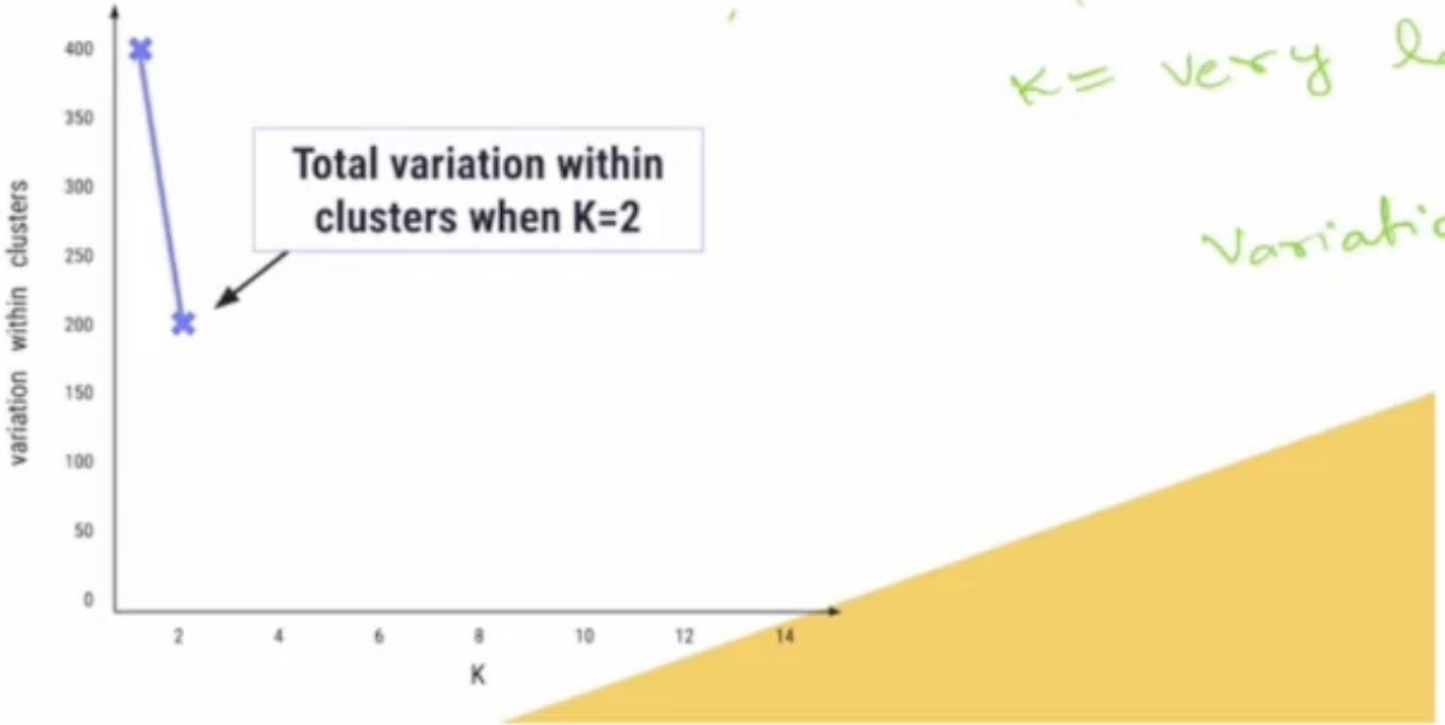


K = 1 ⇒ Variation within a cluster

K = 2

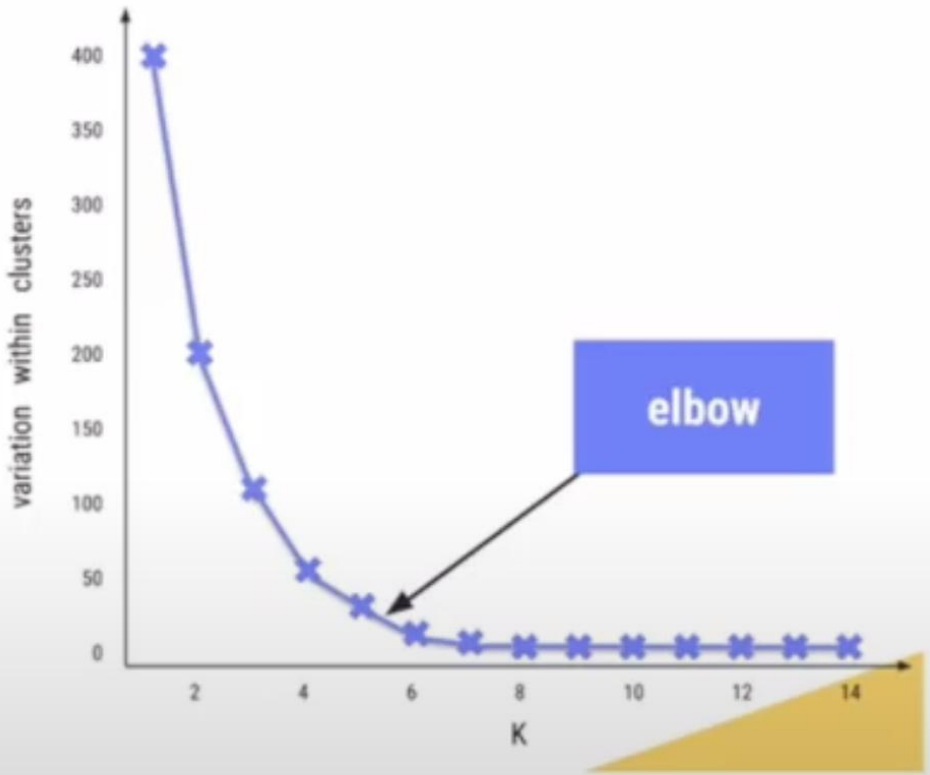
K = very large number

Variation: Avg dist from the centroid



Elbow plot

4, 5, 6



Weakness of k-means

- Can it identify clusters of all shapes?
  - Limited to convex shapes
- Prone to noise
  - Outliers generate spurious centroids
- How do you cluster non-vector data?
  - Text
  - Graphs
  - Time-series