Clustering

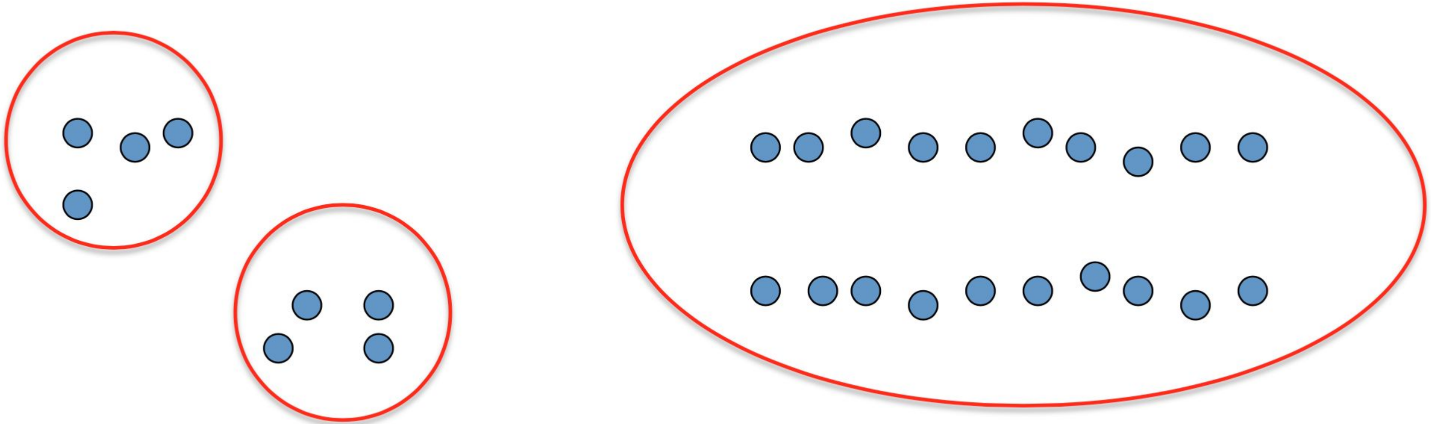
By Saurav

Clustering

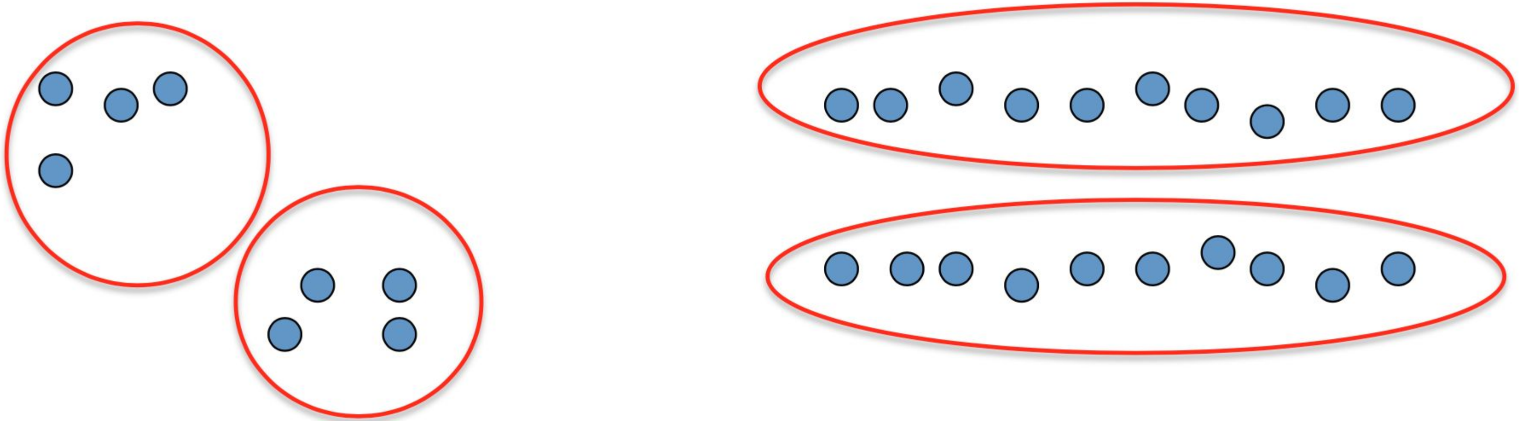
- Unsupervised Learning
- Only input no output data
- To find patterns in data
- To group similar things together

Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



So what do you mean by similar?



- What could “similar” mean?
  - One option: small Euclidean distance (squared)

$$\mathrm{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

  - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

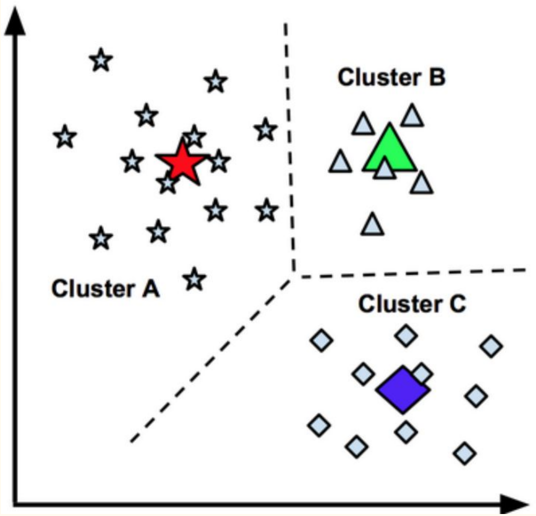
So what is a good cluster?

- Points within same cluster should be similar
- Points between different clusters should be dissimilar.

Two Types of Clustering

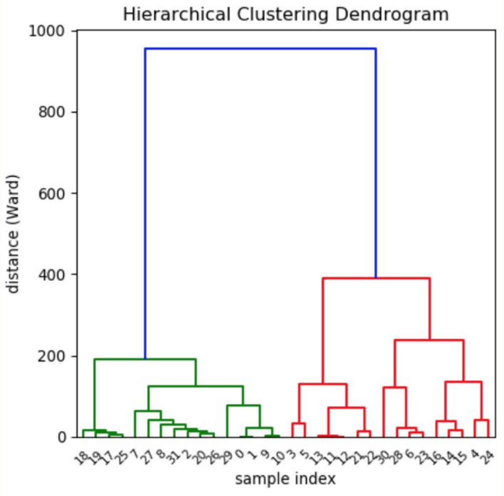
Partition Algorithms

- K Means Algorithm.



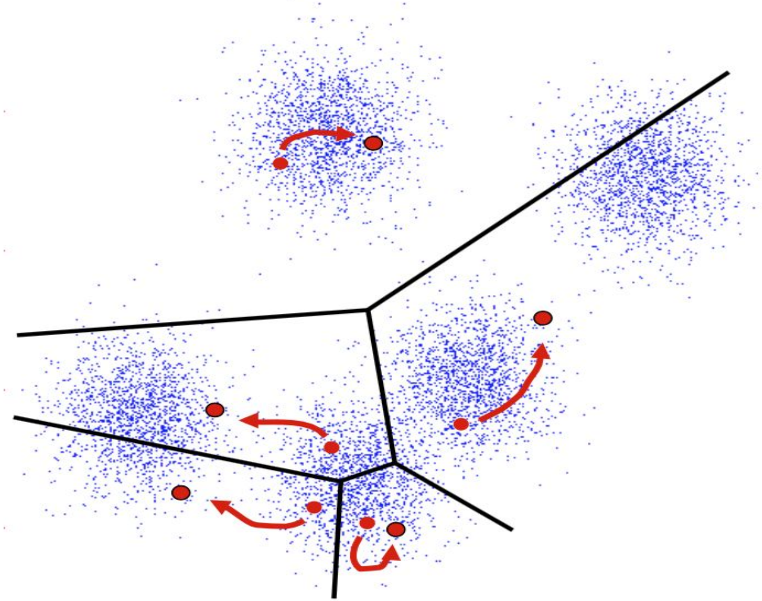
Hierarchical Algorithms

- Bottom -up Algorithm (Agglomerative)



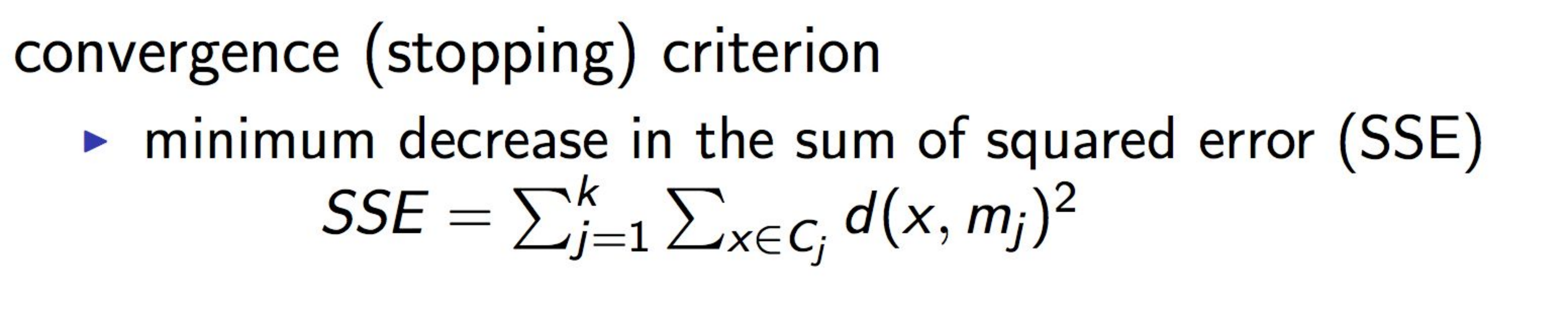
K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



Try K means on these four points (1,1), (2,1), (4,3), (5,4)

So when to stop the clustering?



convergence (stopping) criterion

- minimum decrease in the sum of squared error (SSE)

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} d(x, m_j)^2$$

Within cluster sum of squares (sos)

Total Error = sos(C1) + sos(C2) + sos(C3)

Does the unit of our variable matter in K means? (Like 1 km or 1000 meters)

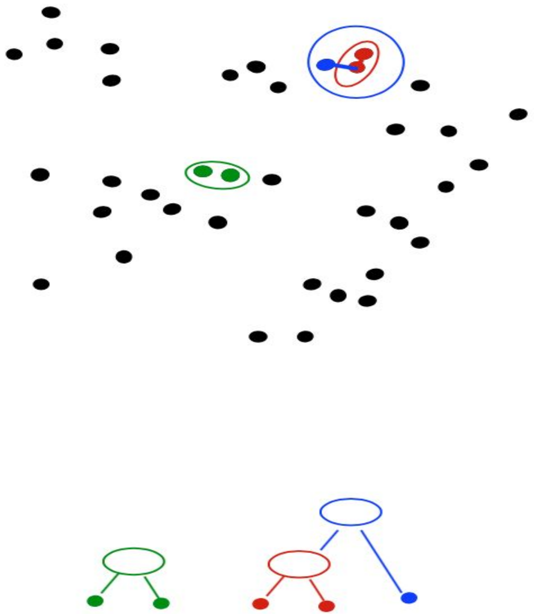
(2,3, 1000) 1000 in meters or (2,3,1) 1 in km.

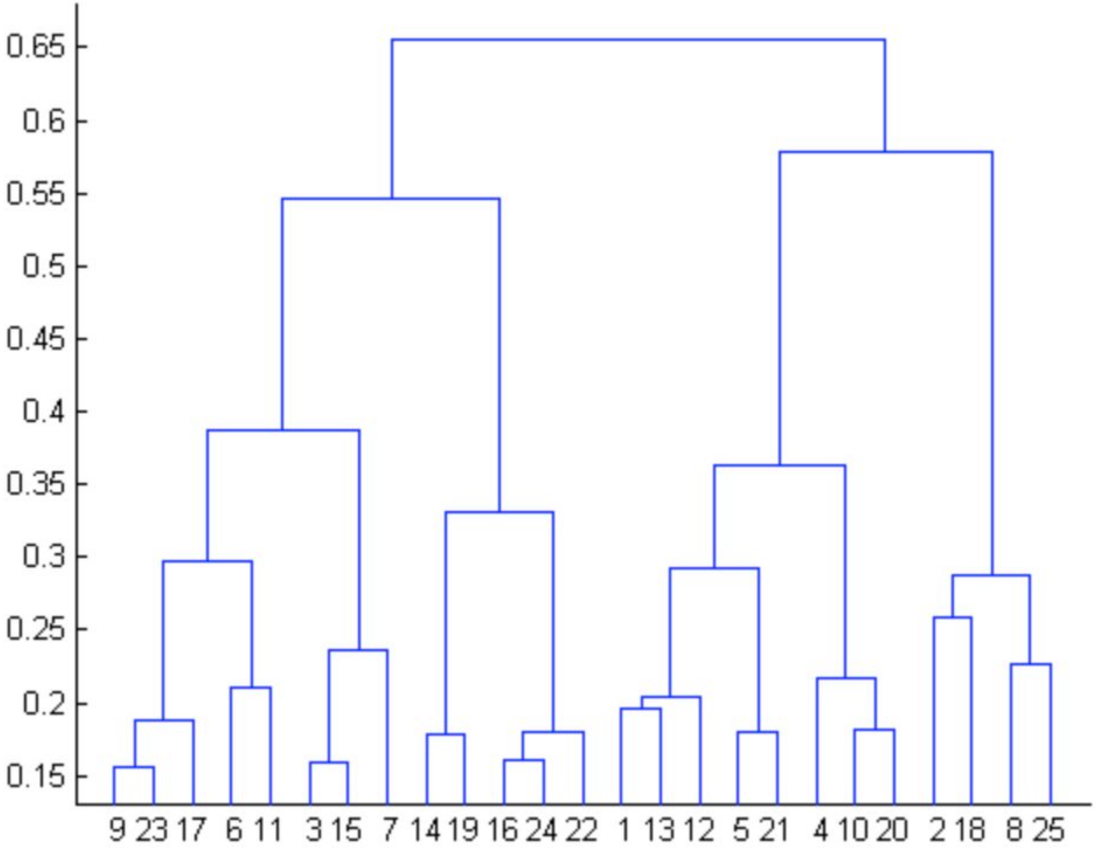
Properties of K Means

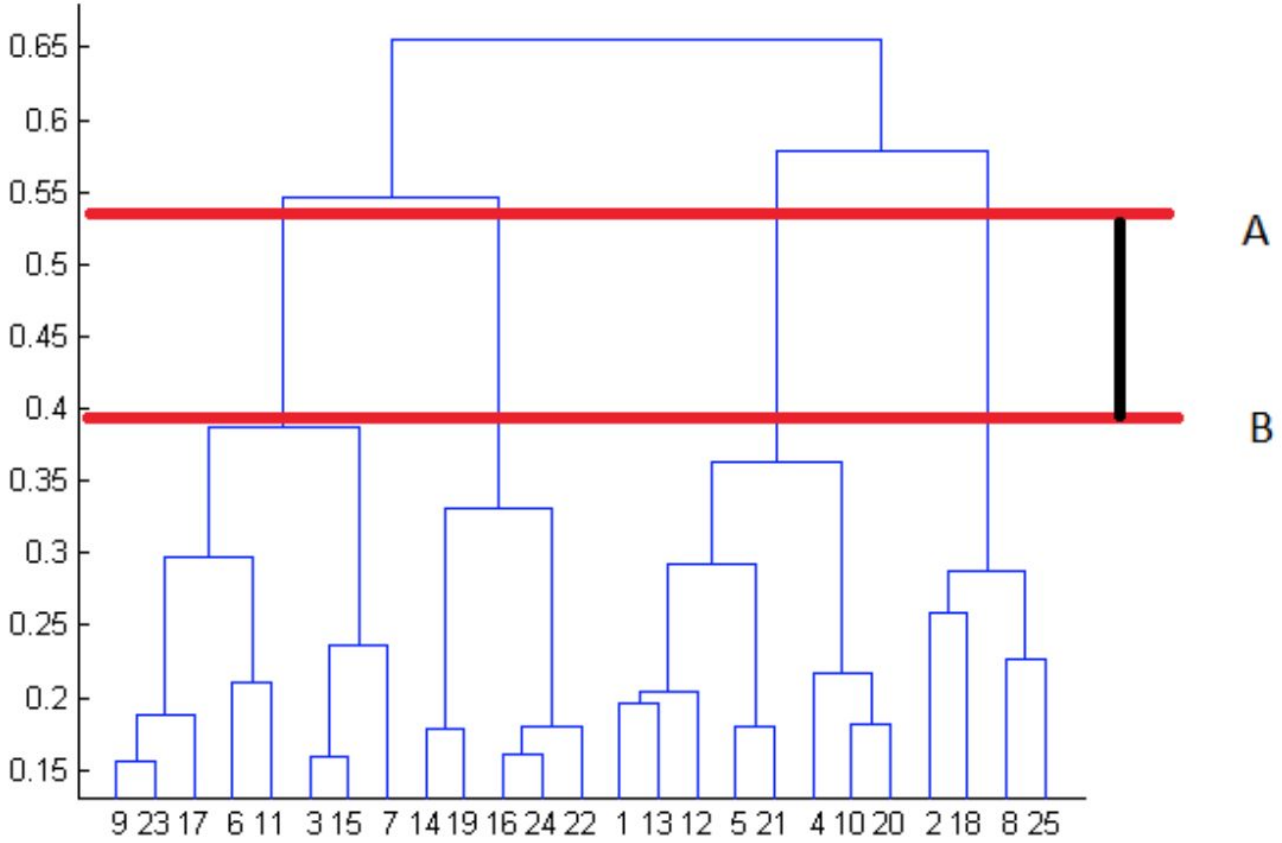
- Need to provide ‘K’ to the algorithm.
- Useful when we already know the number of clusters.
- Useful for big data set.

Agglomerative Clustering

- Agglomerative clustering:
  - First merge very similar instances
  - Incrementally build larger clusters out of smaller clusters
- Algorithm:
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there’s only one cluster left
- Produces not one clustering, but a family of clusterings represented by a dendrogram



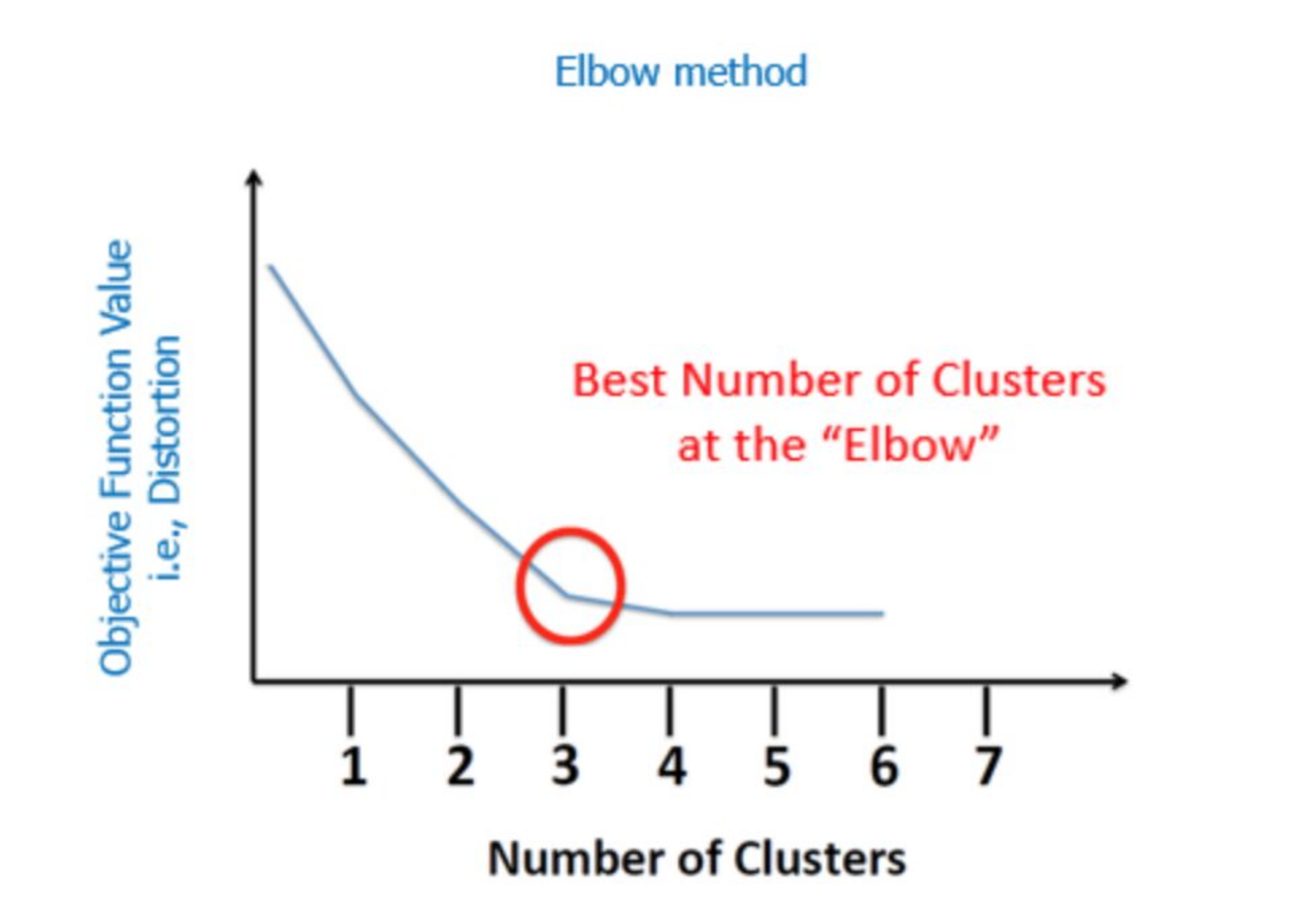

0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
9 23 17 6 11 3 15 7 14 19 16 24 22 1 13 12 5 21 4 10 20 2 18 8 25

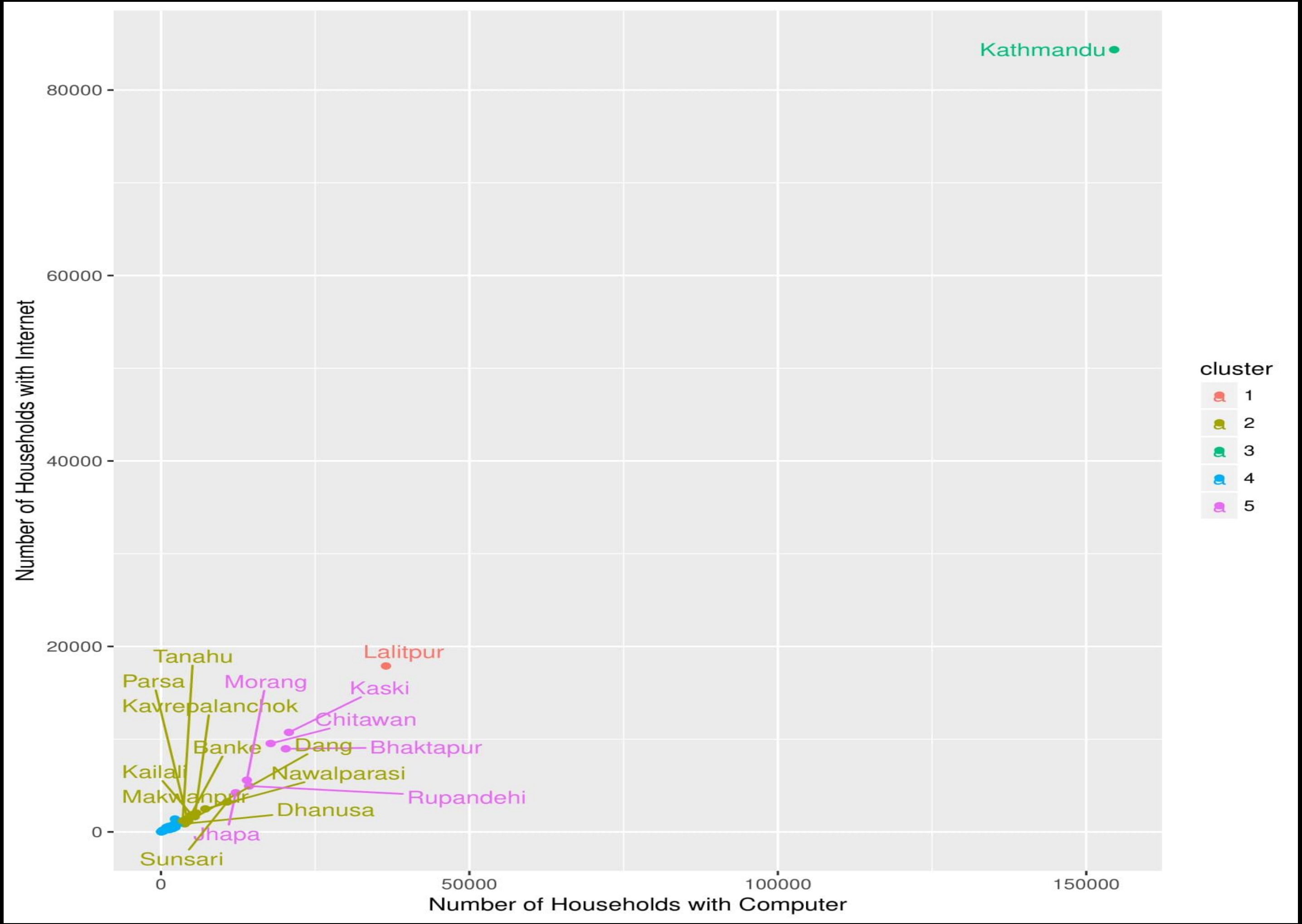

0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
9 23 17 6 11 3 15 7 14 19 16 24 22 1 13 12 5 21 4 10 20 2 18 8 25
A
B

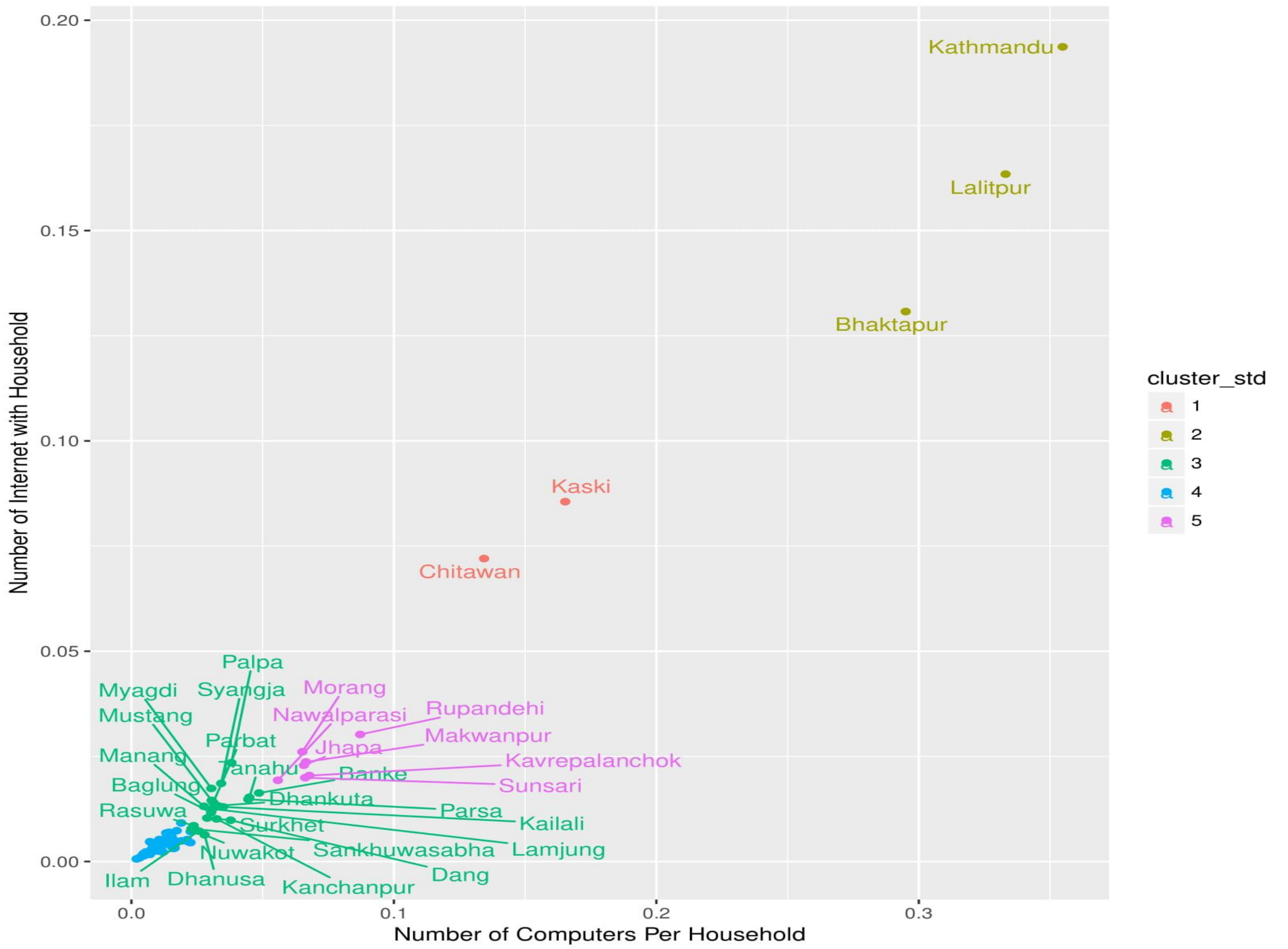
Properties of Hierarchical Clustering

- No need to provide ‘K’ to the algorithm.
- Useful when we don’t know the number of clusters before hand.
- Useful when we want tree structure of our data.
- Not suitable for big data set.

So what is the best cluster size?

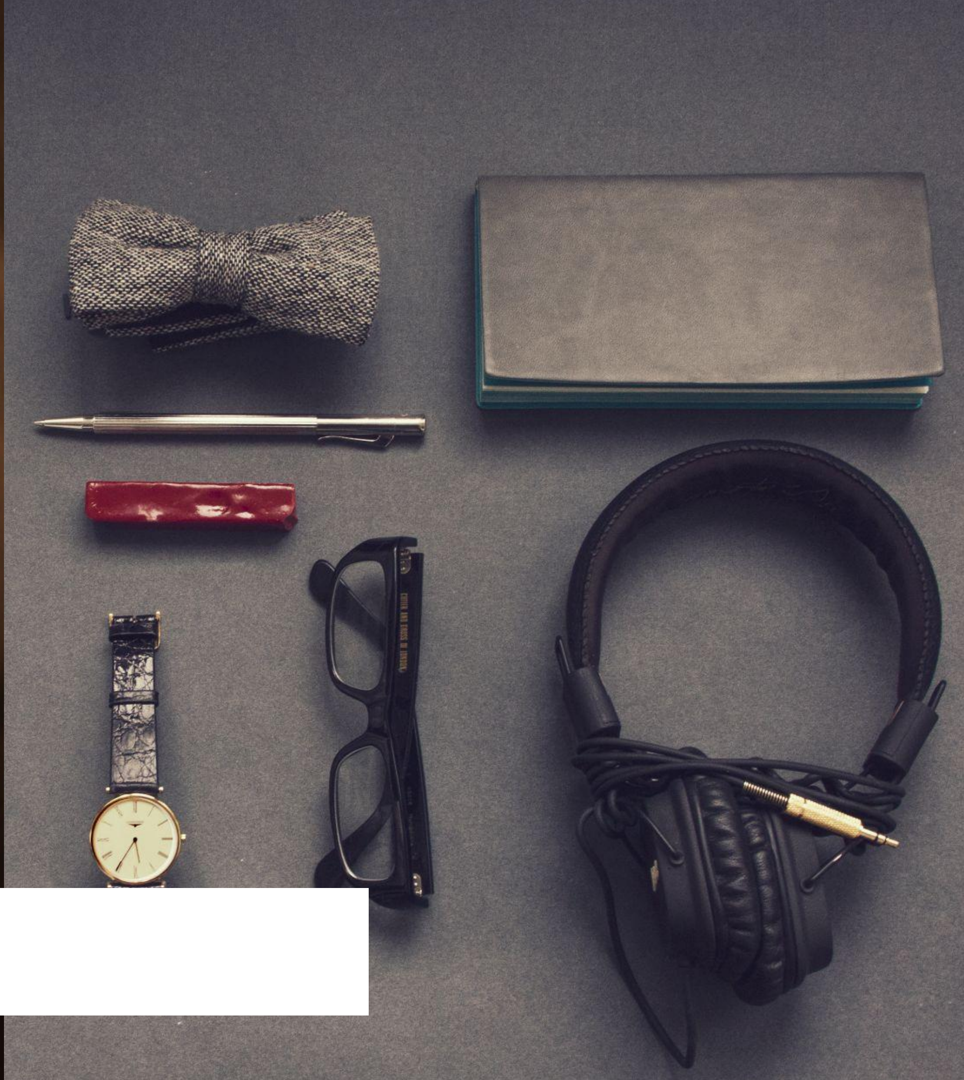

Elbow method
Objective Function Value
i.e., Distortion
Best Number of Clusters
at the "Elbow"
1
2
3
4
5
6
7
Number of Clusters


Kathmandu
Lalitpur
Tanahu
Parsa
Kavrepalanchok
Morang
Kaski
Chitawan
Banke
Dang
Bhaktapur
Kailali
Nawalparasi
Makwanpur
Rupandehi
Dhanusa
Jhapa
Sunsari
80000
60000
40000
20000
0
0
50000
100000
150000
Number of Households with Internet
Number of Households with Computer
cluster
1
2
3
4
5


Number of Internet with Household
0.20
0.15
0.10
0.05
0.00
Number of Computers Per Household
0.0
0.1
0.2
0.3
Kathmandu
Lalitpur
Bhaktapur
Kaski
Chitawan
Palpa
Myagdi
Syangja
Morang
Mustang
Nawalparasi
Rupandehi
Makwanpur
Parbat
Jhapa
Manang
Tanahu
Kavrepalanchok
Banke
Sunsari
Baglung
Dhankuta
Rasuwa
Parsa
Kailali
Surkhet
Sankhuwasabha
Lamjung
Nuwakot
Ilam
Dhanusa
Dang
Kanchanpur
cluster_std
1
2
3
4
5

Possible use cases related to your projects?

End Note!



Let us do it in coding now!