

Welcome to

*DS3010:*  
*DS-III: Computational Data Intelligence*  
***Unsupervised Learning***

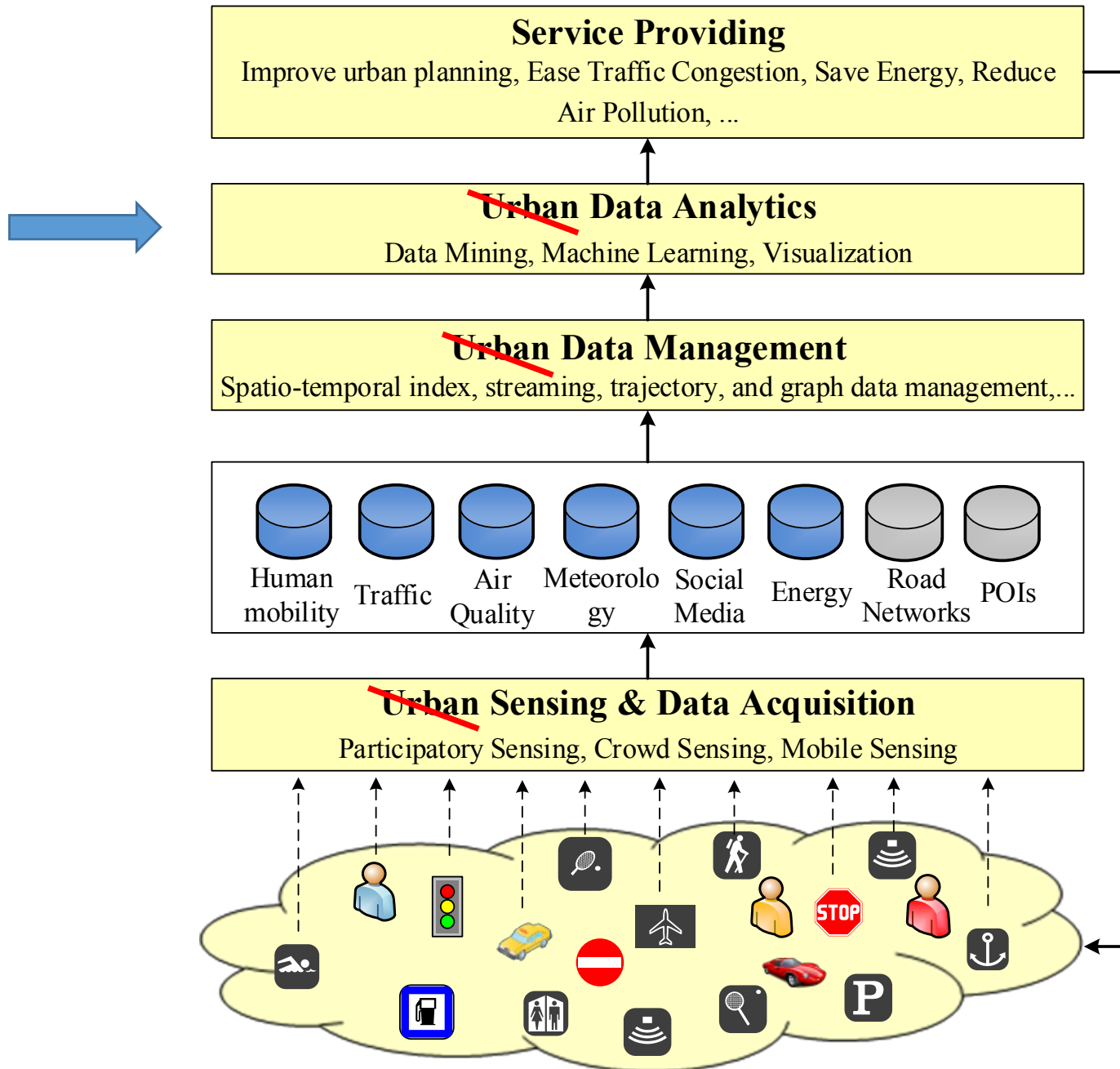
Prof. Yanhua Li

Time: 11:00am – 12:50pm M & R

Location: HL 114

D-term 2022

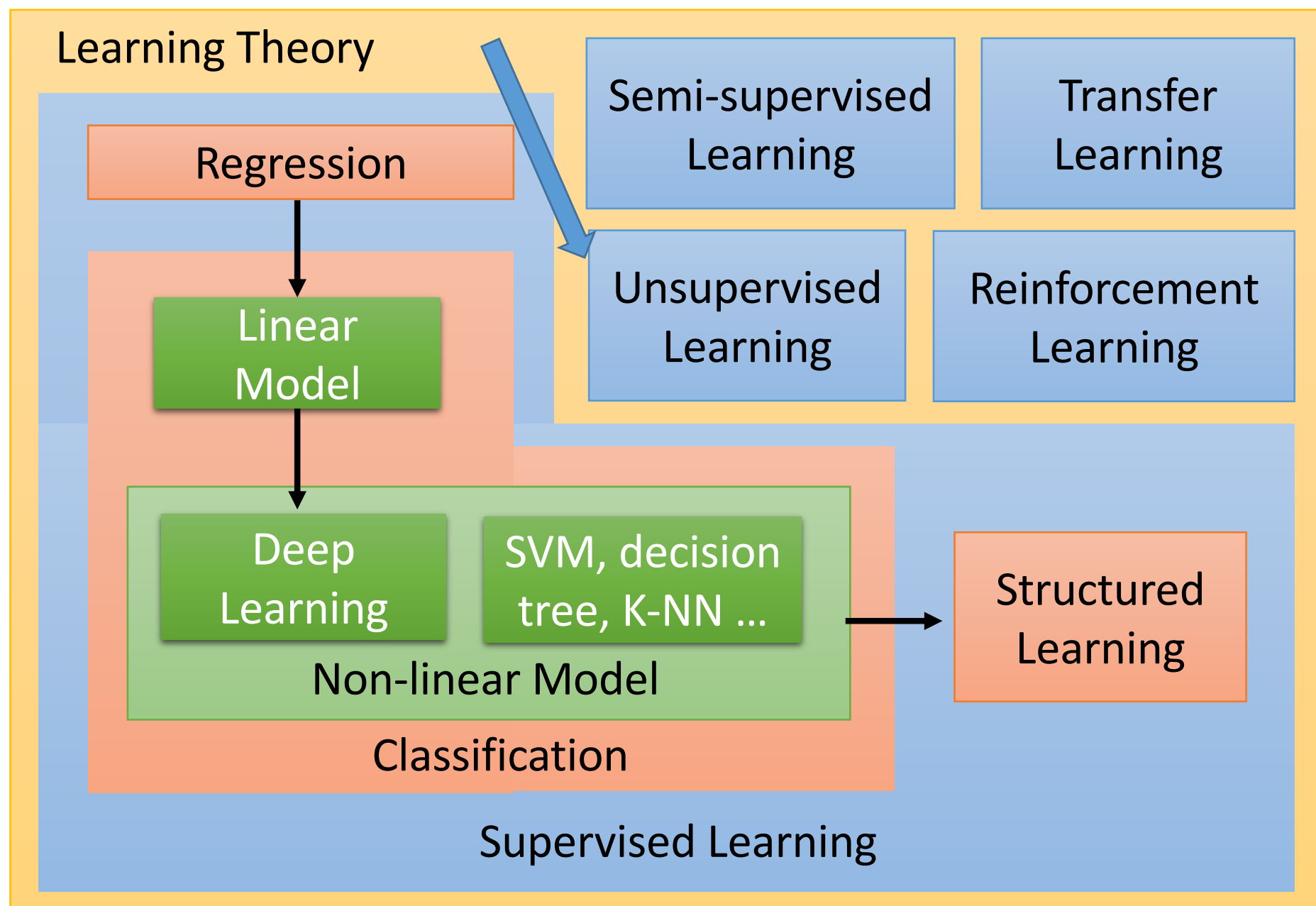
# Data pipeline



**Urban Computing: concepts, methodologies, and applications.**  
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.

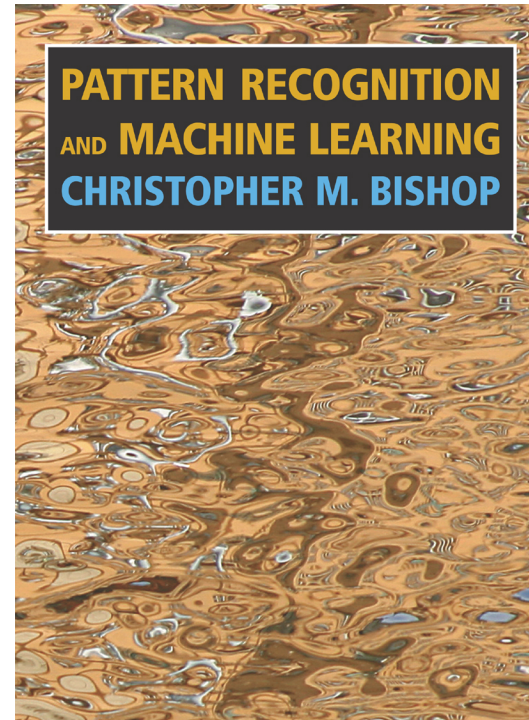
# Learning Map

scenario task method



# References

Unsupervised  
Learning:  
Clustering

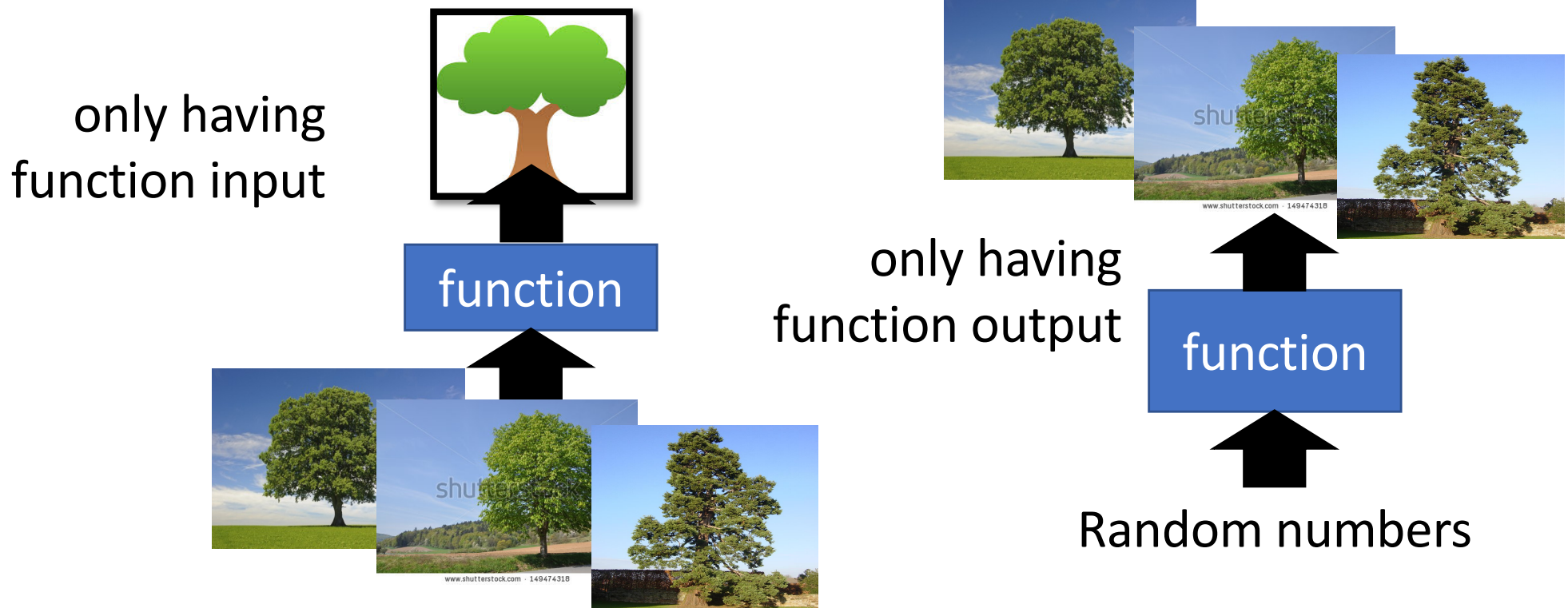


Bishop: Chapter 9.1

# Unsupervised Learning

# Unsupervised Learning

- Clustering & dimension reduction (**our focus**)
- Generation



# Clustering

$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Cluster 1



Open question: how many clusters do we need?

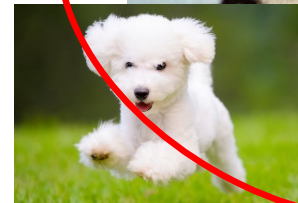


Cluster 3

$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

Cluster 2



- K-means

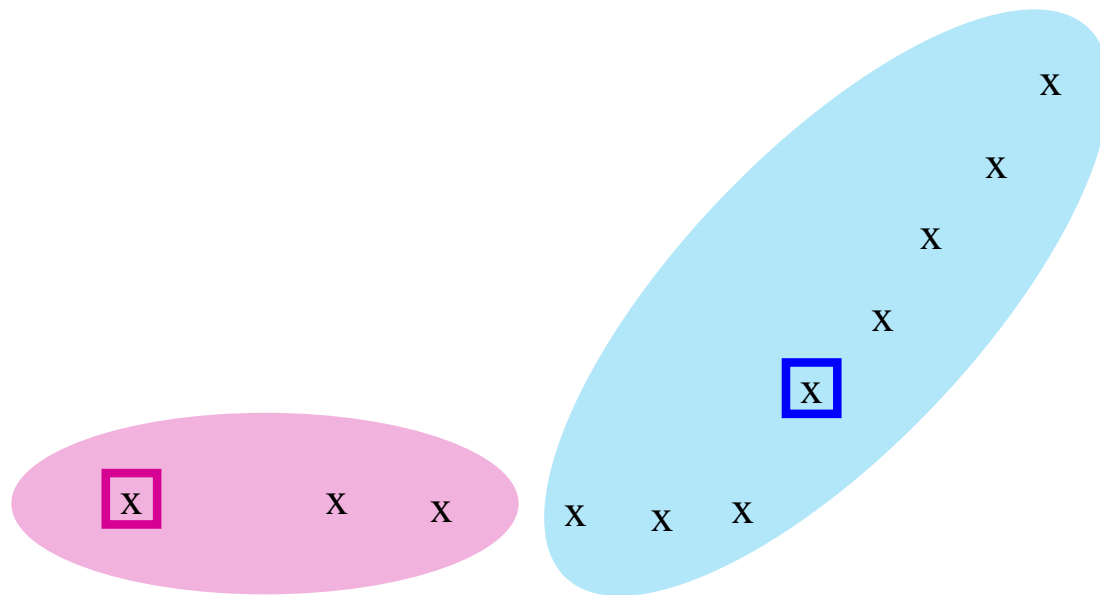
- Clustering  $X = \{x^1, \dots, x^n, \dots, x^N\}$  into K clusters
- Initialize cluster center  $c^i$ ,  $i=1,2, \dots, K$  (K random  $x^n$  from  $X$ )

- Repeat

- For all  $x^n$  in  $X$ :  $b_i^n = \begin{cases} 1 & x^n \text{ is most "close" to } c^i \\ 0 & \text{Otherwise} \end{cases}$

- Updating all  $c^i$ : 
$$c^i = \sum_{x^n} b_i^n x^n / \sum_{x^n} b_i^n$$

# Example: Assigning Clusters

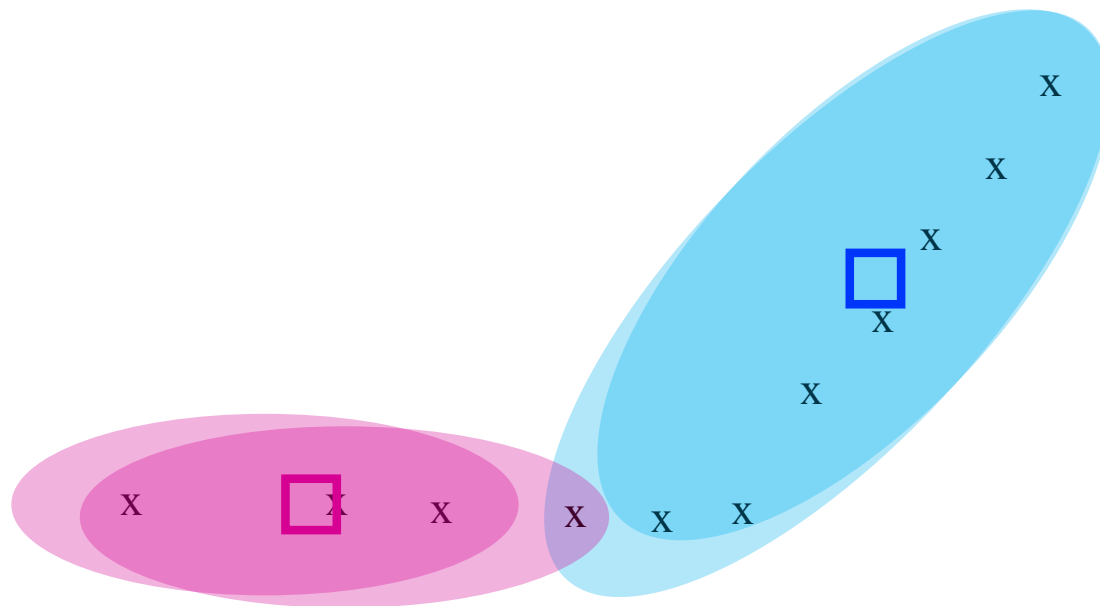


x ... data point  
□ ... centroid

**Clusters after round 1**



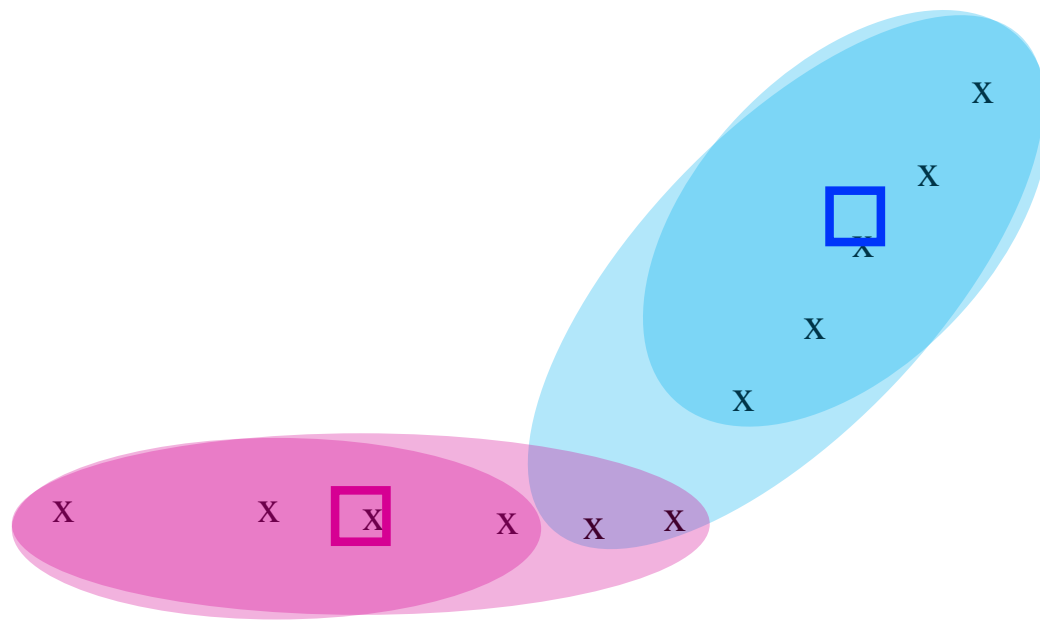
# Example: Assigning Clusters



x ... data point  
□ ... centroid

**Clusters after round 2**

# Example: Assigning Clusters



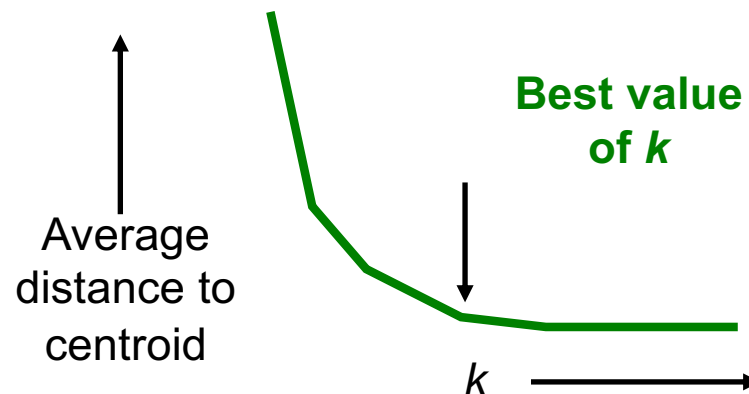
x ... data point  
□ ... centroid

**Clusters at the end**

# Getting the $k$ right

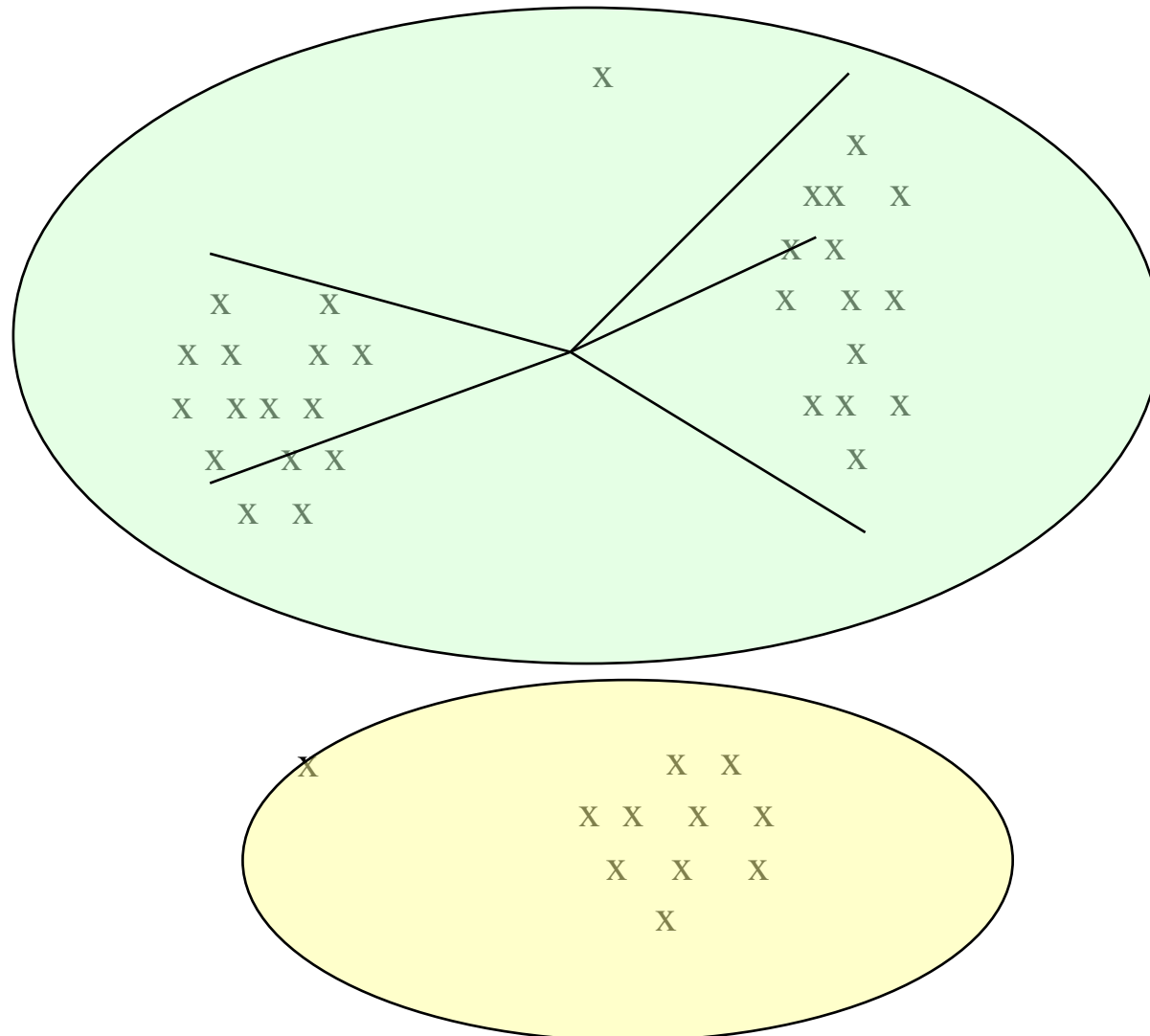
## How to select $k$ ?

- ❖ Try different  $k$ , looking at the change in the average distance to centroid as  $k$  increases
- ❖ Average falls rapidly until right  $k$ , then changes little



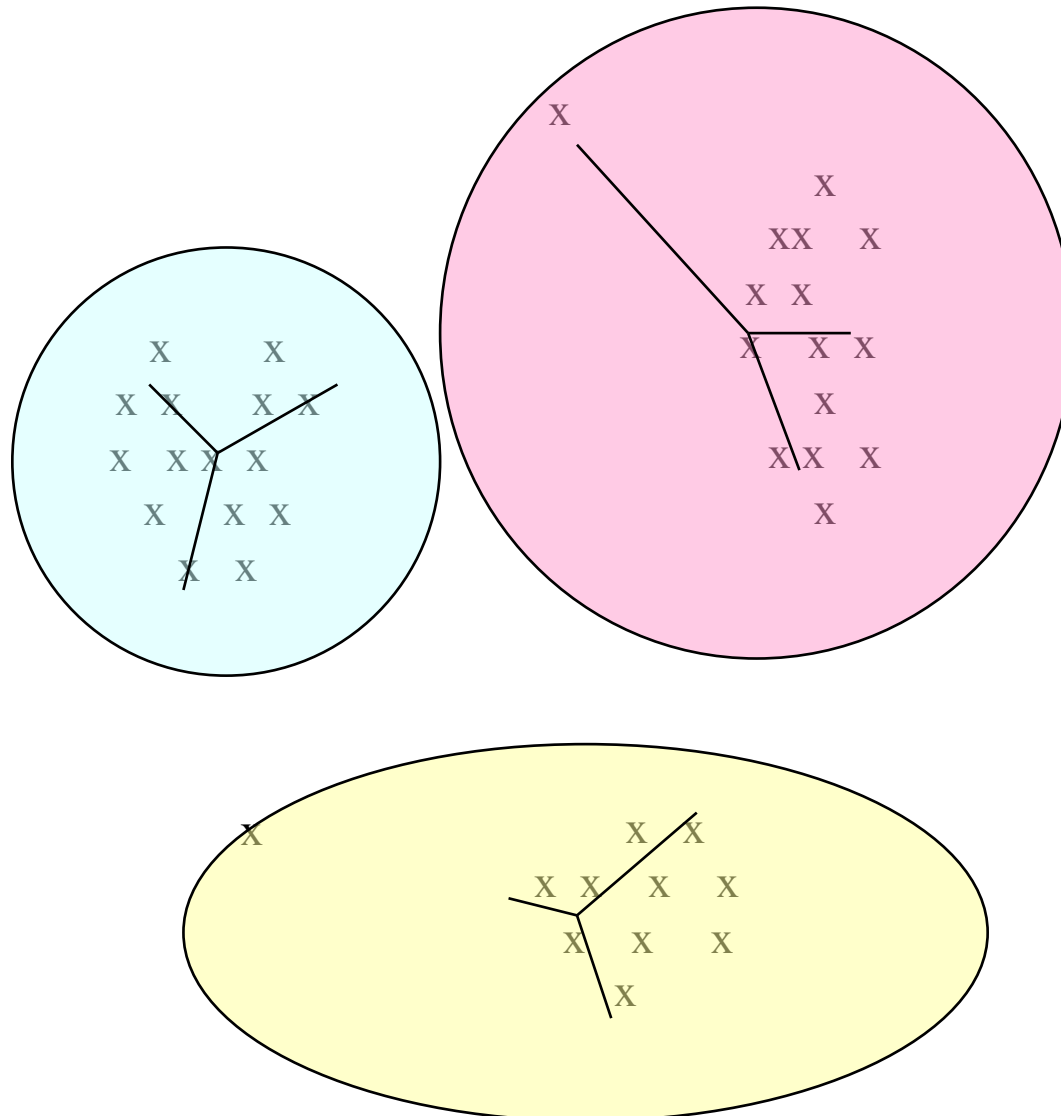
# Example: Picking $k=2$

**Too few;**  
many long  
distances  
to centroid.



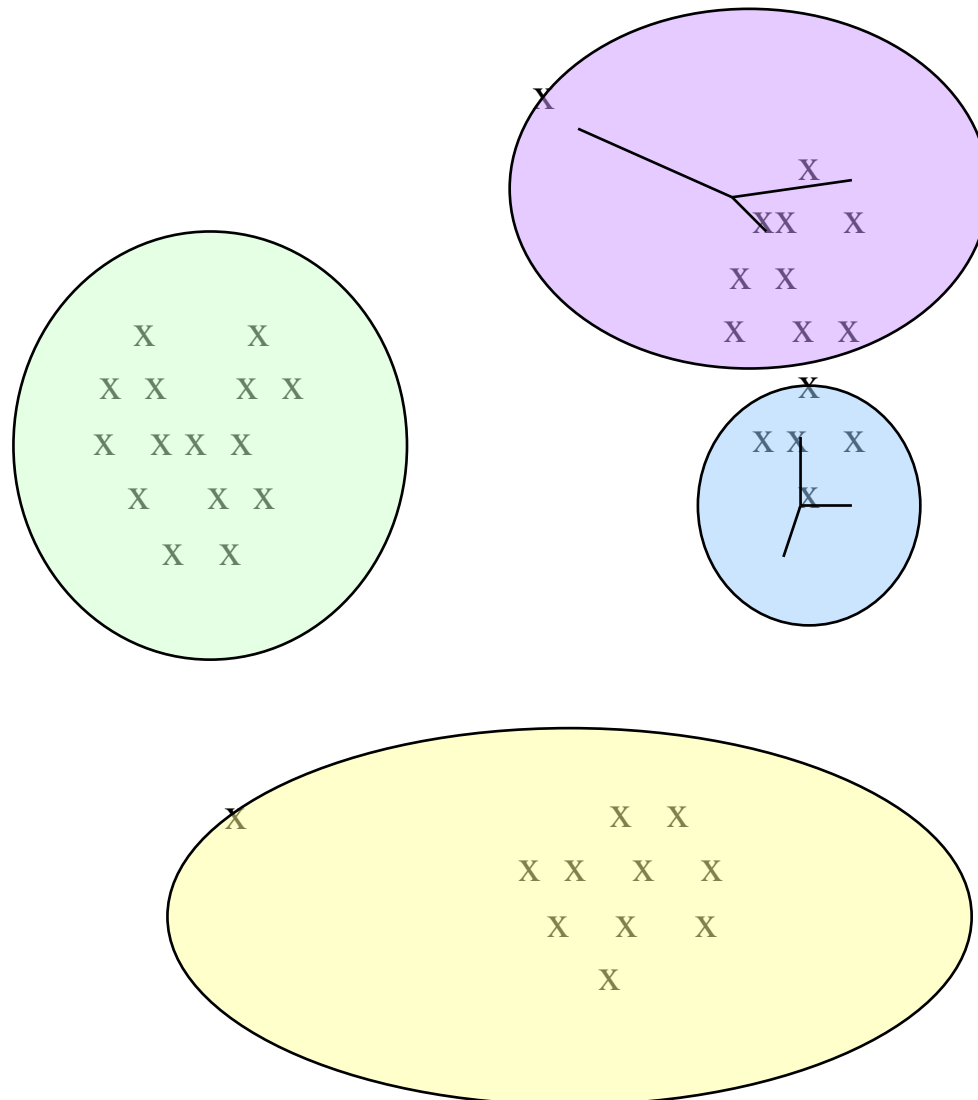
# Example: Picking $k=3$

**Just right;**  
distances  
rather short.



# Example: Picking $k$

**Too many;**  
little improvement  
in average  
distance.

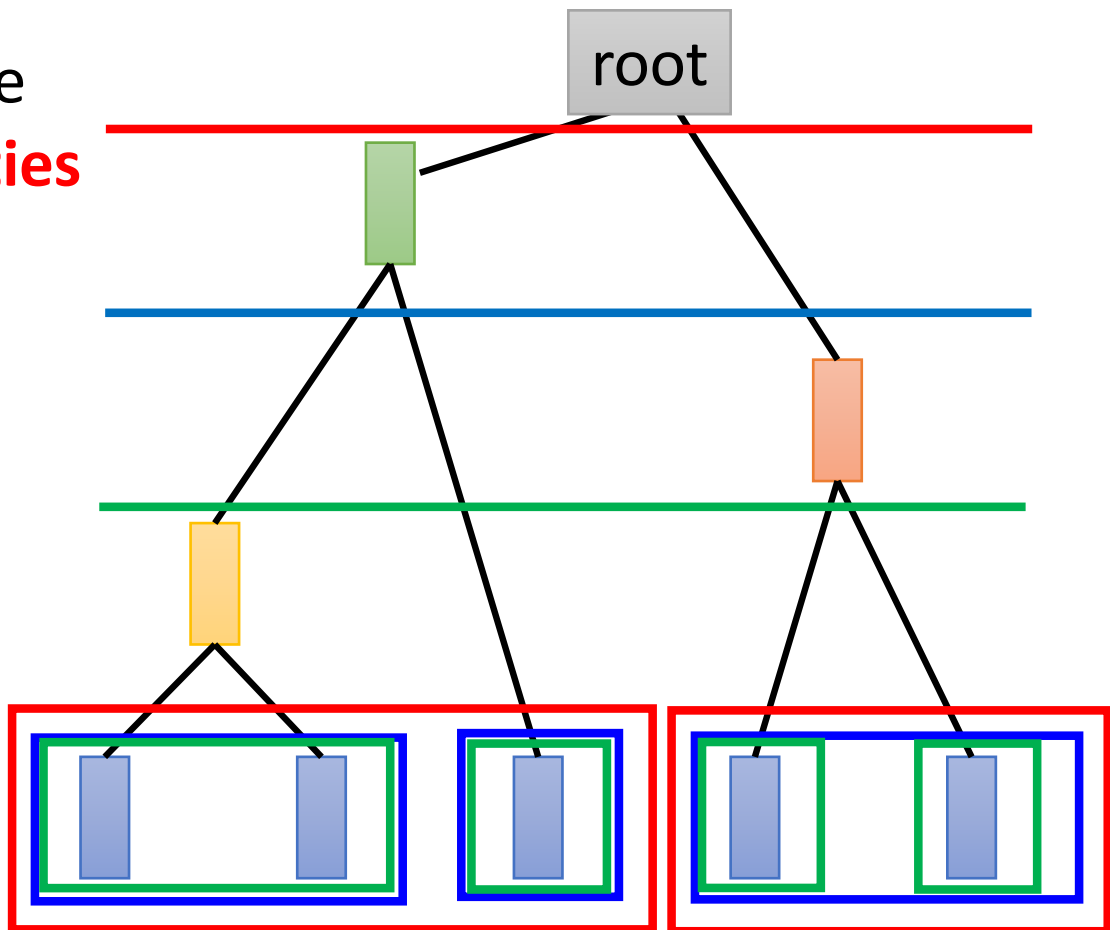


# Clustering

- Hierarchical Agglomerative Clustering (HAC)

Step 1: build a tree  
based on **similarities**  
between clusters

Step 2: pick a  
threshold

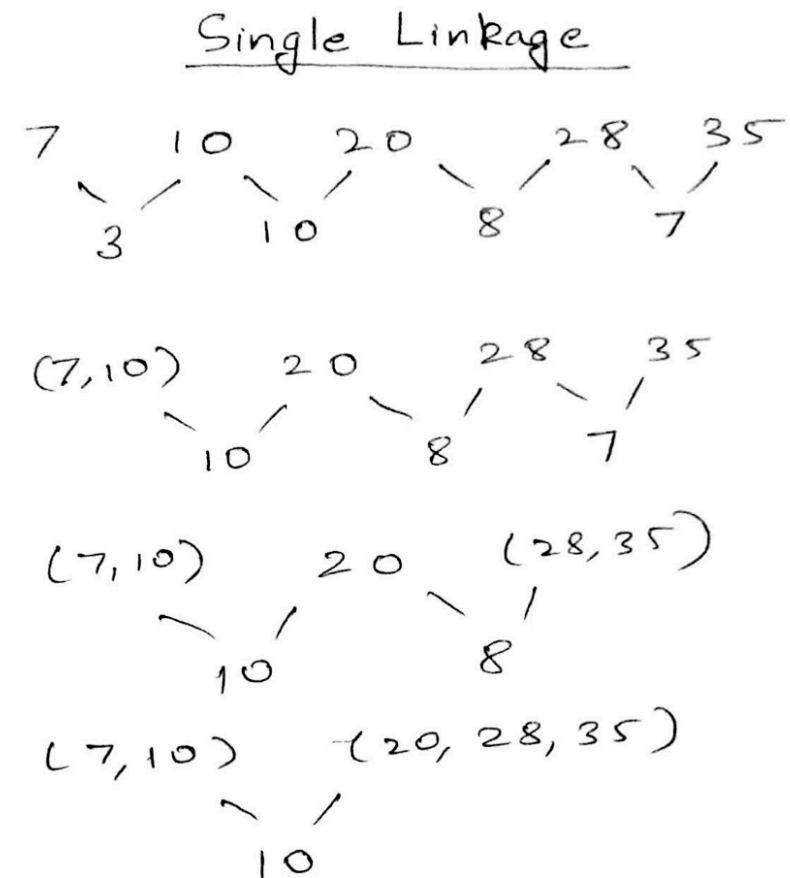
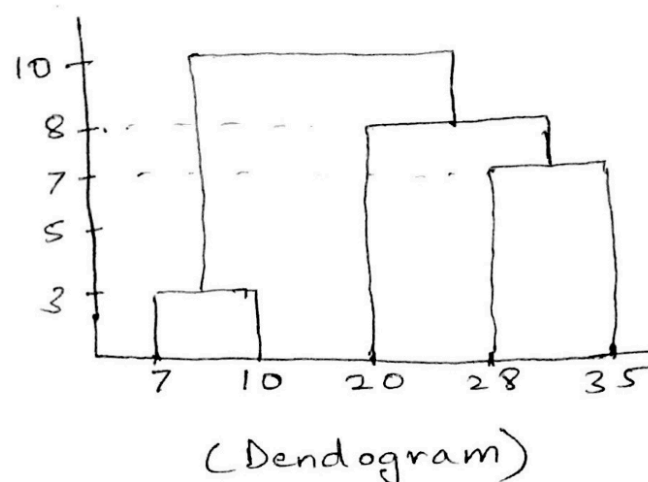


# Using single linkage similarity

- Hierarchical Agglomerative Clustering (HAC)

Step 1: build a tree  
based on **similarities**  
between clusters

Step 2: pick a  
threshold



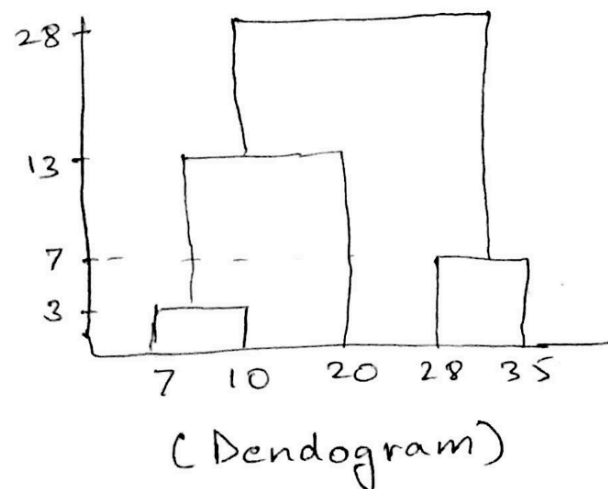


# Using complete linkage similarity

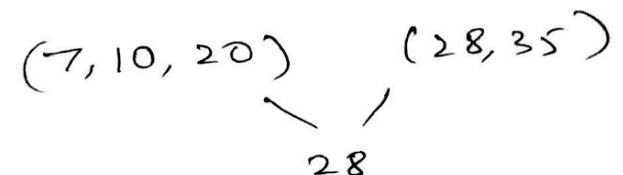
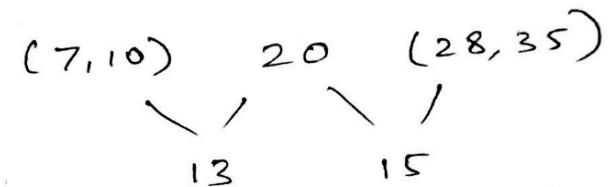
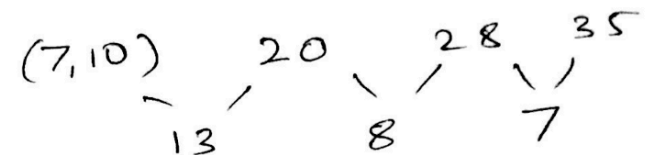
- Hierarchical Agglomerative Clustering (HAC)

Step 1: build a tree  
based on **similarities**  
between clusters

Step 2: pick a  
threshold



Complete Linkage



# Questions