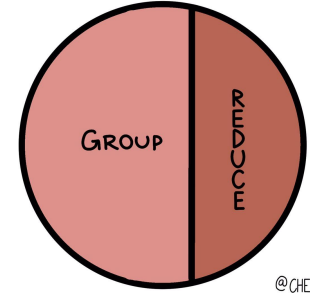


SIMPLIFY



@CHELSEAPARLETT

# K-Means

Dr. Chelsea Parlett-Pelleriti

# Unsupervised Machine Learning

# Clustering

# K-Means

$$C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K W(C_k) \right\}$$

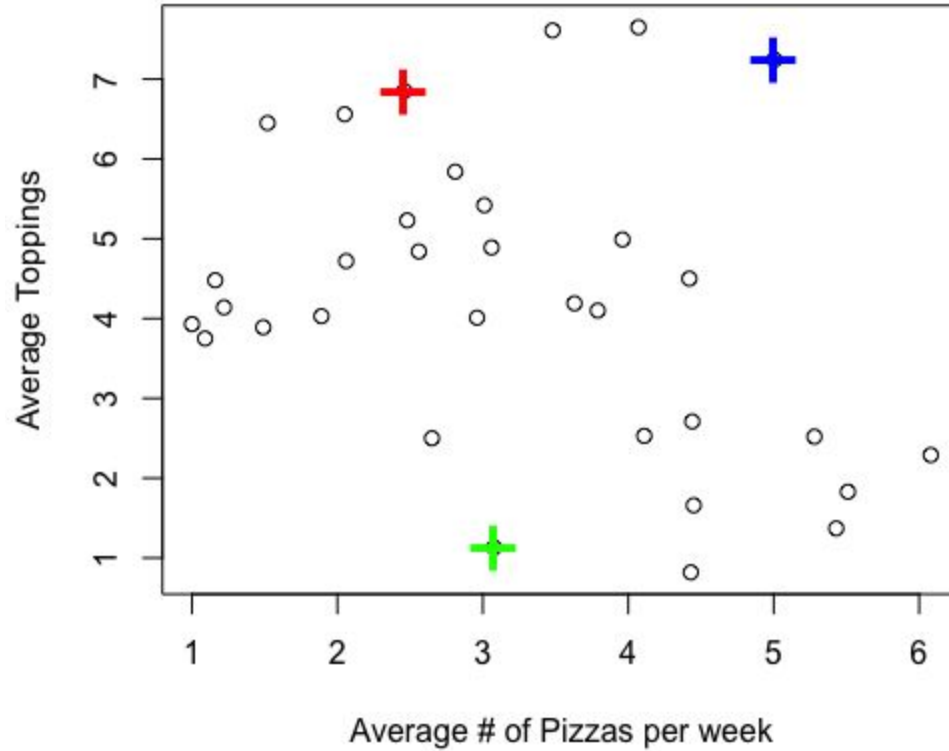
$$W(C_k) = \frac{1}{C_k} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

# K-Means Algorithm

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers
4. Repeat 2 and 3 until either:
  - a. Cluster membership does not change
  - b. Centers change only a tiny amount

# K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers

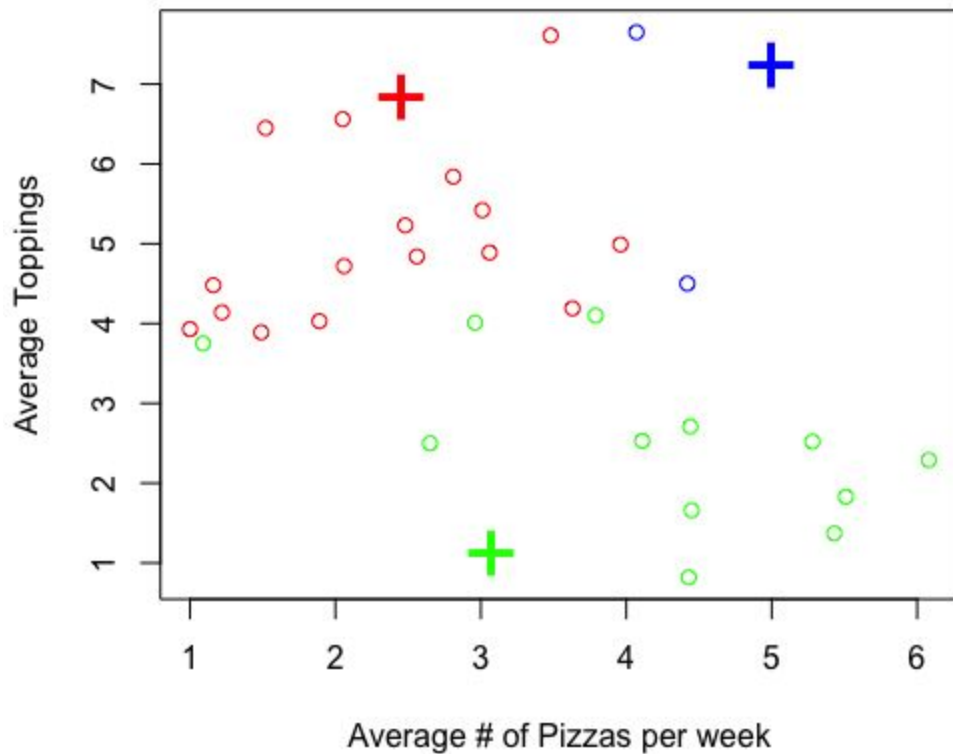


1

# 2

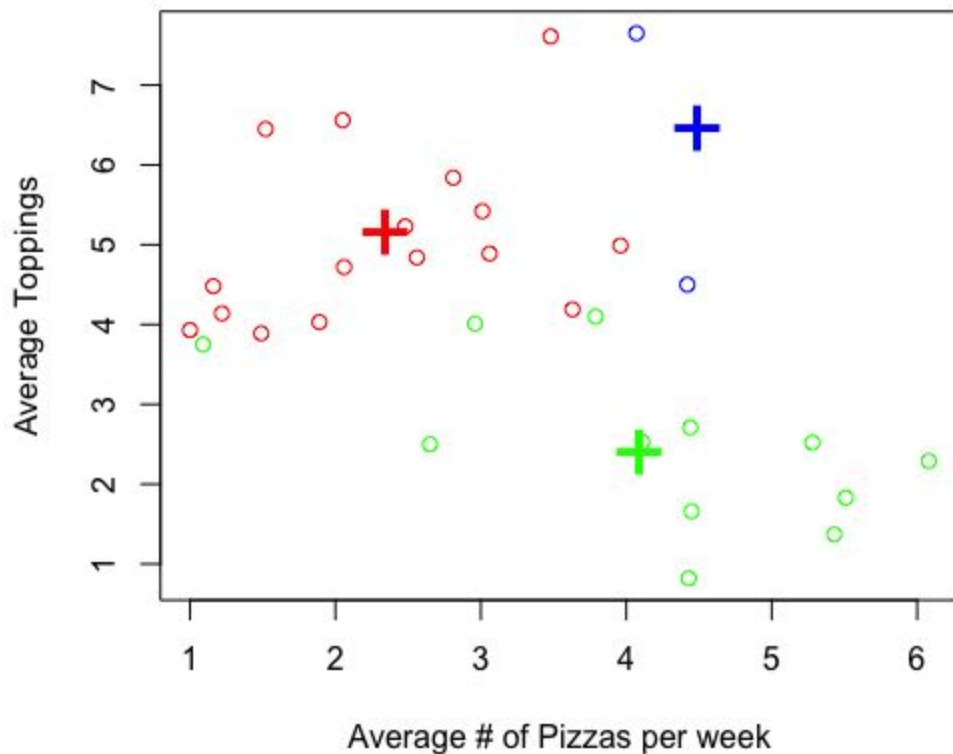
## K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



# K-Means

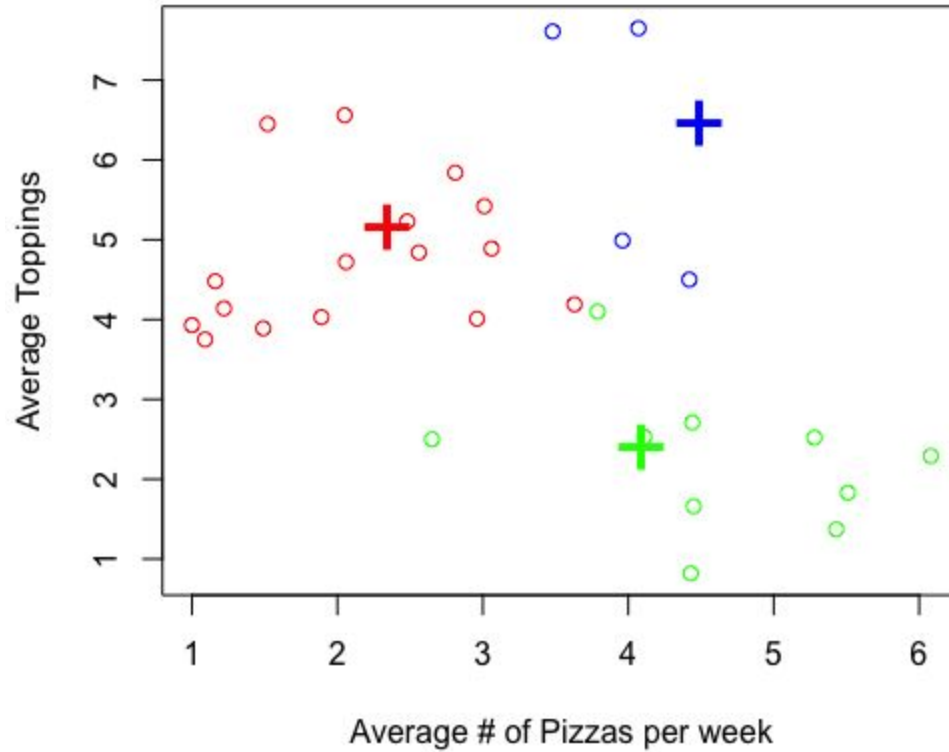
1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers





# K-Means

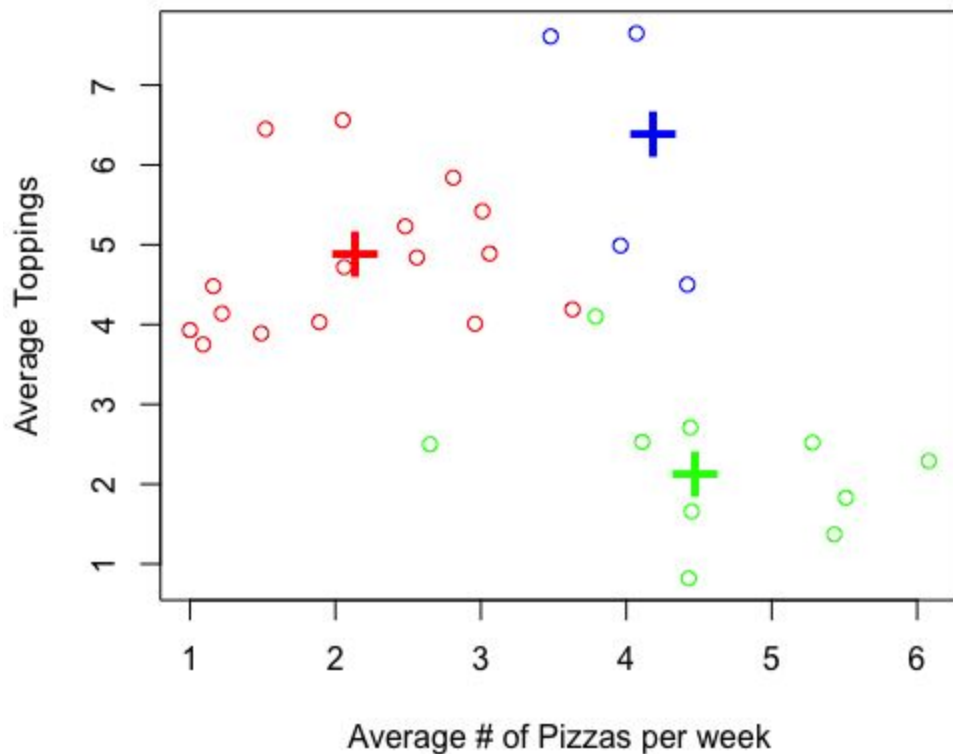
1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



2

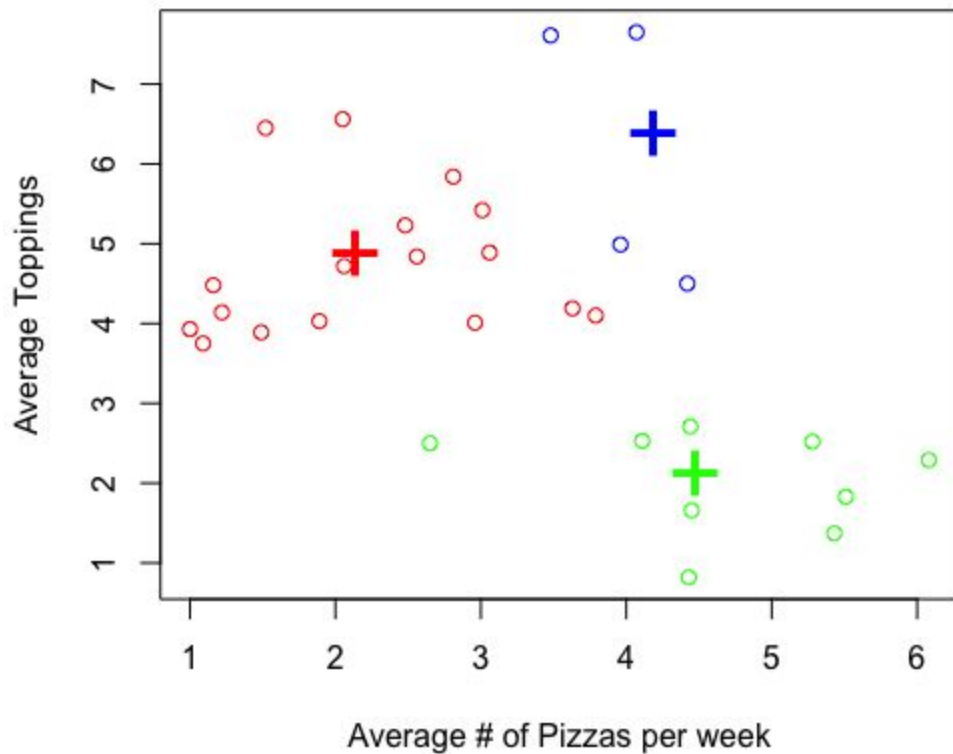
# K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



# K-Means

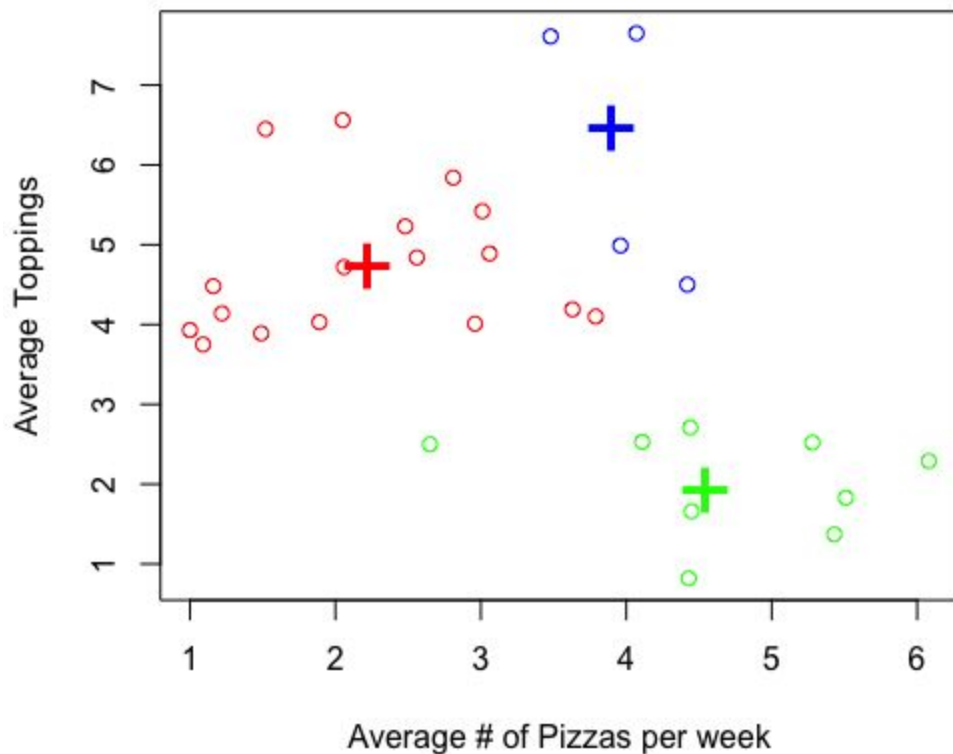
1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



2

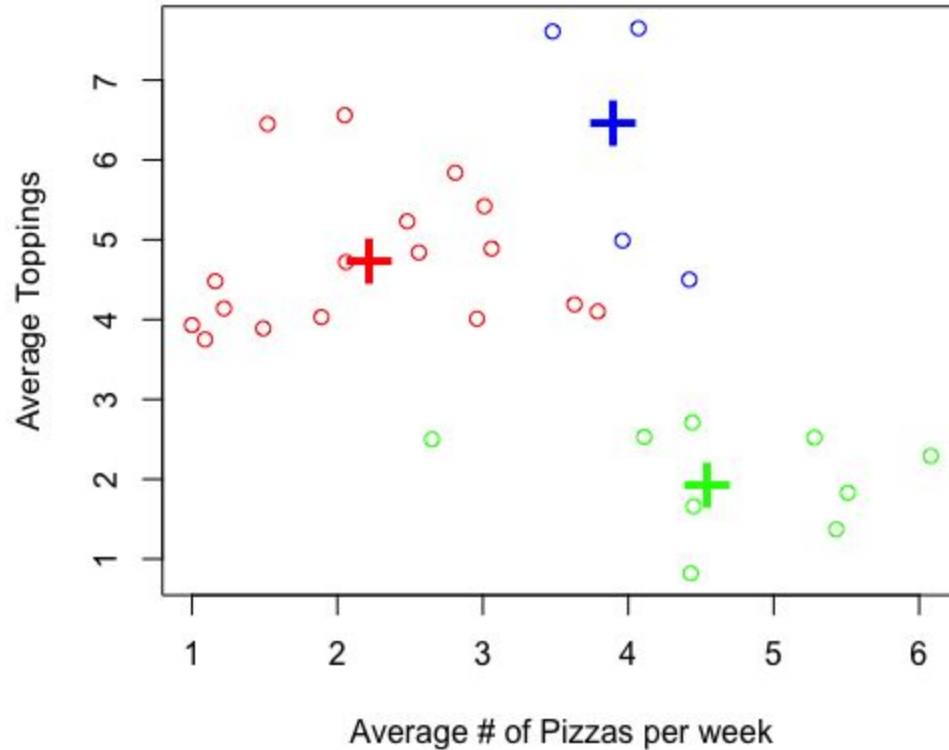
# K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



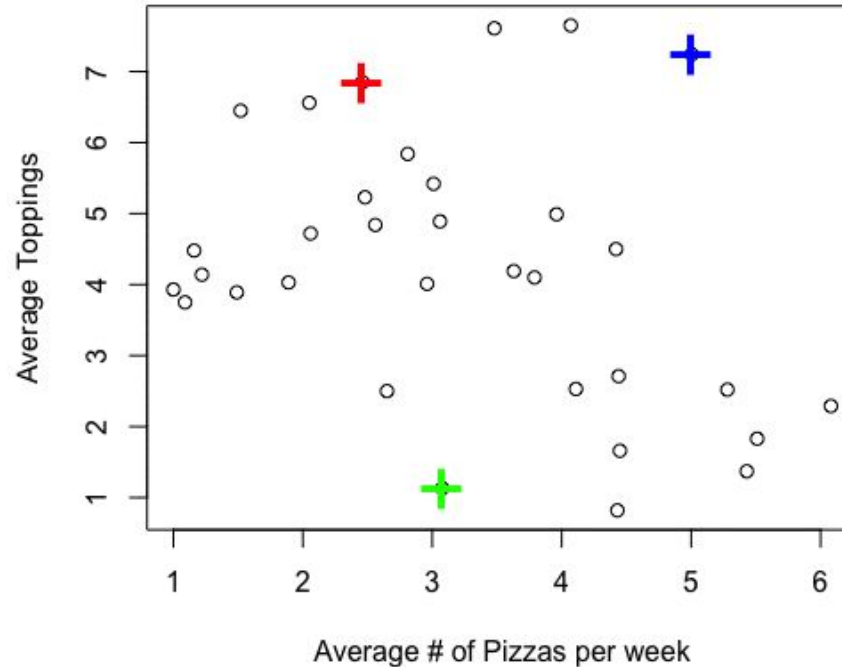
# K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



# K-Means

1. Choose  $k$  random points to be cluster centers
2. For each data point, assign it to the cluster whose center is closest
3. Using these assignments, recalculate the centers



# K-Means

## Assumptions

- Spherical Clusters
- Roughly the same # in each cluster

# Evaluating Unsupervised Models

Cohesion:

Separation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$



# Applications

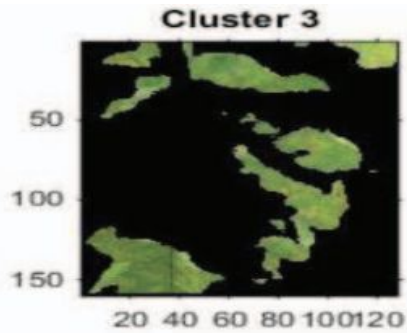


Fig. 8: Only Leaf

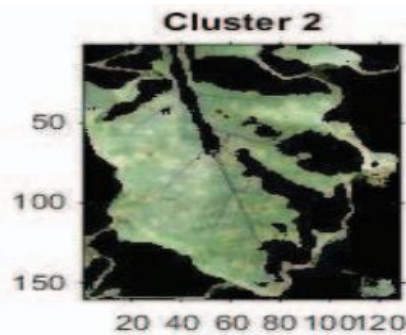


Fig. 7: Both Brinjal and Leaf

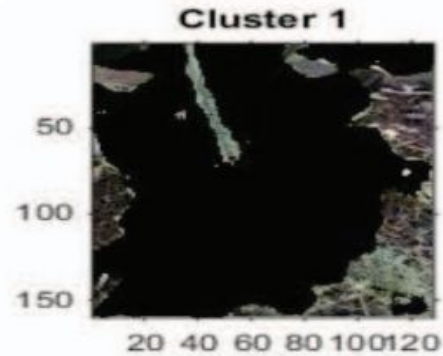


Fig. 6: Only Brinjal

# Applications

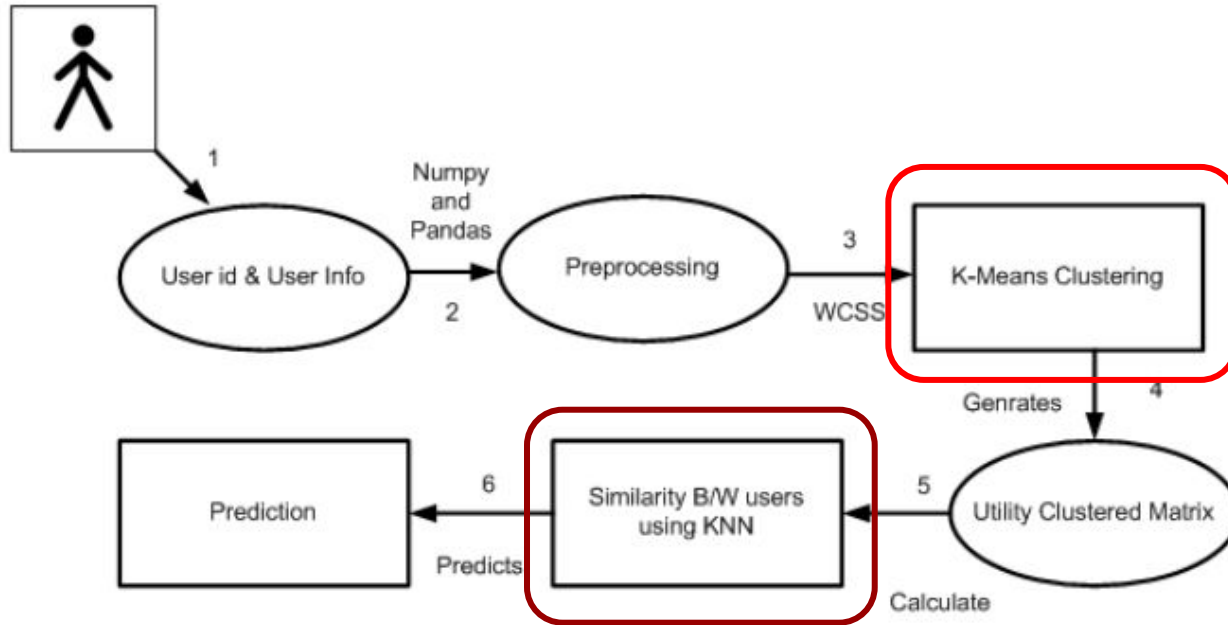


Fig. 2: Process Flow Diagram