
# The Principles of Clustering

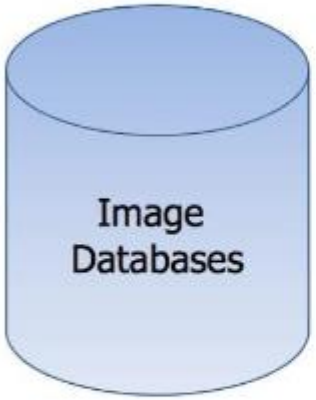
Alexander G. Ororbia II

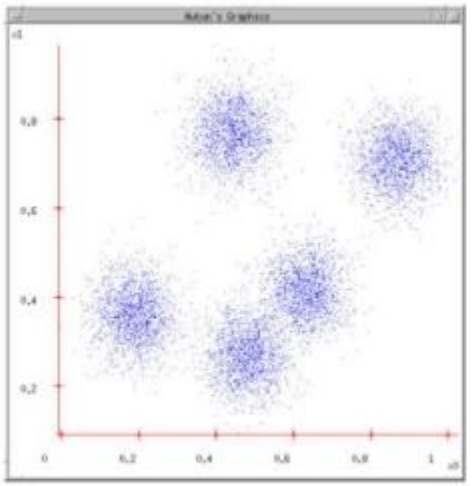
Introduction to Machine Learning

CSCI-335

4/6/2026

# Clustering at a Glance



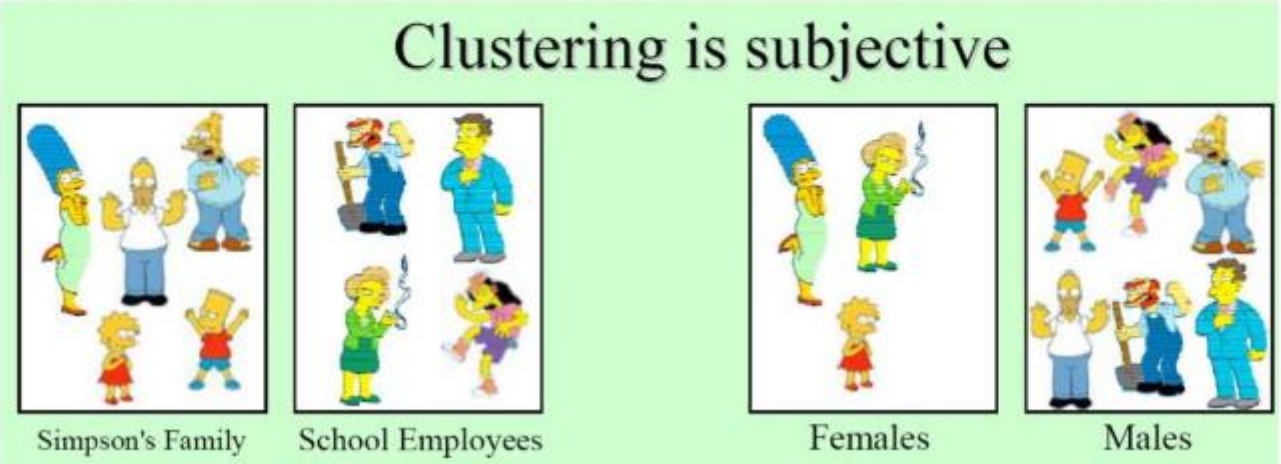

Goal of clustering:
Divide objects into groups and objects within a group are more similar than those outside the group.

# Basic Ingredients of Clustering

- First we need to pick similarity/dissimilarity function?
- The algorithm figures out the grouping of objects based on the chosen dissimilarity/dissimilarity function:
  - Points within a cluster is similar
  - Points across cluster are not similar
- Issues for clustering:
  - How to represent objects? (vector space? Normalization)
  - What is similarity/dissimilarity function?
  - What are the algorithm steps?

# The Subjectivity of Clustering – What is Good Clustering?

What is considered similar/dissimilar?



***White board time!*** Developing the K-means algorithm

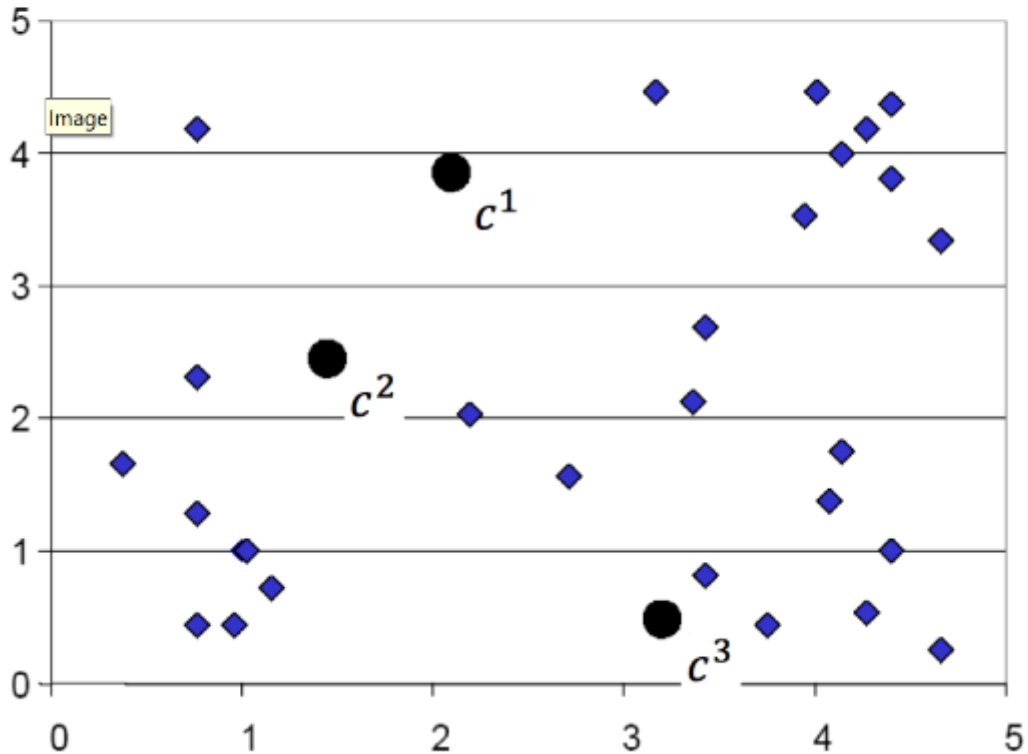
**Algorithm 1** The K-means algorithm for partitioning – each cluster center represented as centroid mean of cluster.

**Input:** Dataset $\mathcal{D}$ containing $N$ objects
**Hyperparameters:** Number of clusters $K$, number of iterations $T$
**function** SIMULATE($\mathcal{D}, K, T$)
    Arbitrarily assign $K$ objects in $\mathcal{D}$ as initial centroids
    **for** $t = 1$ to $T$ **do**
        (Re)assign each $\mathbf{p}$ to cluster it is most similar to, i.e., $\arg\min_j\{\text{dist}(\mathbf{p}, \mathbf{c}_1), ..., \text{dist}(\mathbf{p}, \mathbf{c}_K)\}$
        Update cluster means – re-caclulate centroid mean values for objects in each cluster:
            $$\mathbf{c}_i = \frac{1}{|C_i|}\sum_{j=1}^{|C_i|}\mathbf{p}_j, \text{for } j = 1, 2, ..., K$$
    **Return** The set of $K$ clusters $\{C_1, ..., C_K\}$

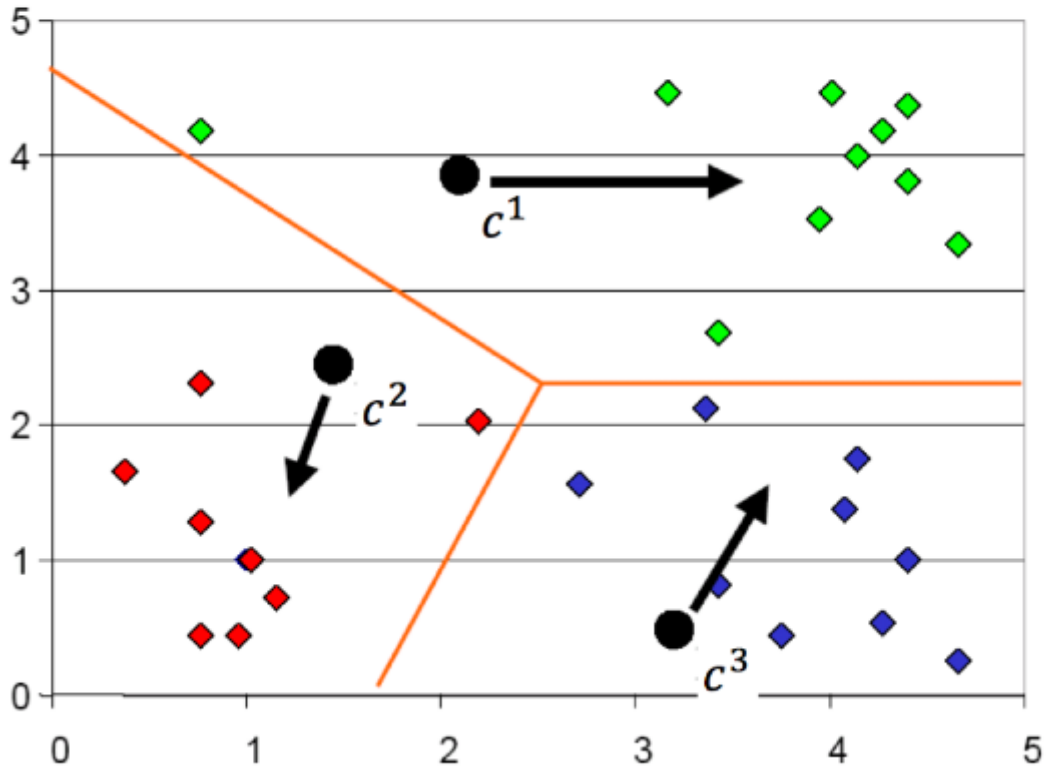
**Algorithm 2** The K-medoids algorithm for partitioning based on representative objects.

**Input:** Dataset $\mathcal{D}$ containing $N$ objects
**Hyperparameters:** Number of clusters $K$, number of iterations $T$
**function** SIMULATE($\mathcal{D}, K, T$)
    Arbitrarily choose $K$ objects in $\mathcal{D}$ as initial representatives
    **for** $t = 1$ to $T$ **do**
        Assign each remaining object to cluster with nearest representative
        Randomly choose nonrepresentative object $\mathbf{o}_{rand}$
        Compute total new cost $\mathbf{F}_{t+1}$ of swapping representative object $\mathbf{o}_j$ with $\mathbf{o}_{rand}$
        If $\mathcal{F}_t - \mathcal{F}_{t+1} > 0$ then swap $\mathbf{o}_j$ with $\mathbf{o}_{rand}$ to form new set of $K$ representatives
    **Return** The set of $K$ clusters $\{C_1, ..., C_K\}$

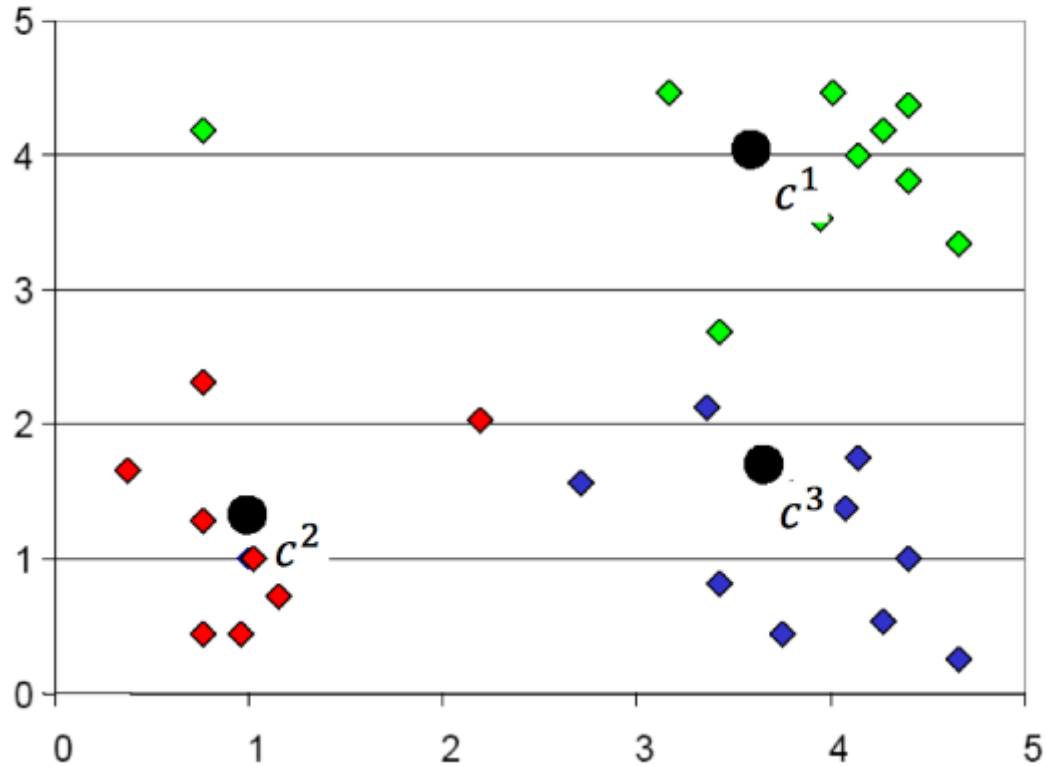
# K-Means step 1



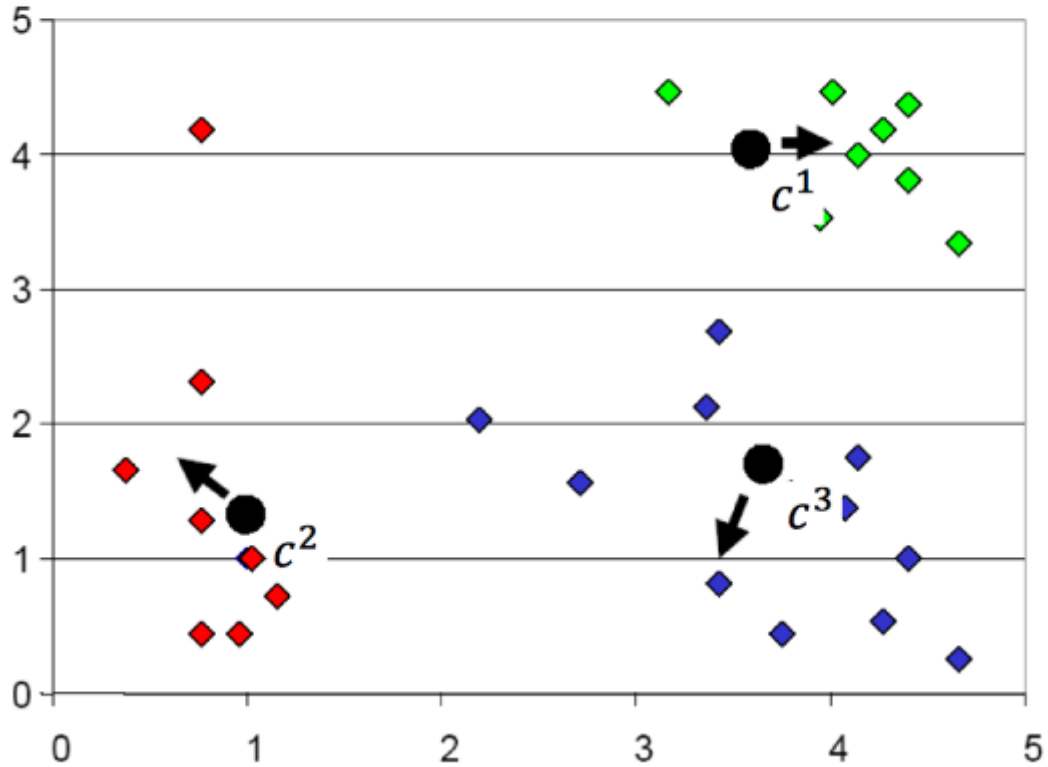
# K-Means step 2



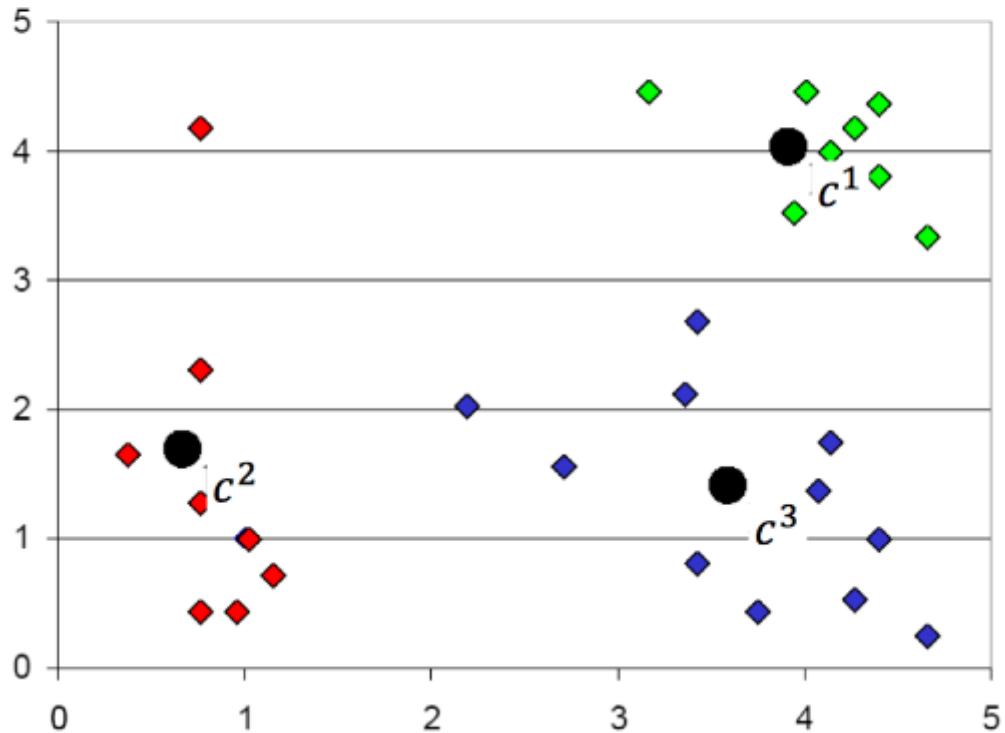
# K-Means step 3



# K-Means step 4



# K-Means step 5



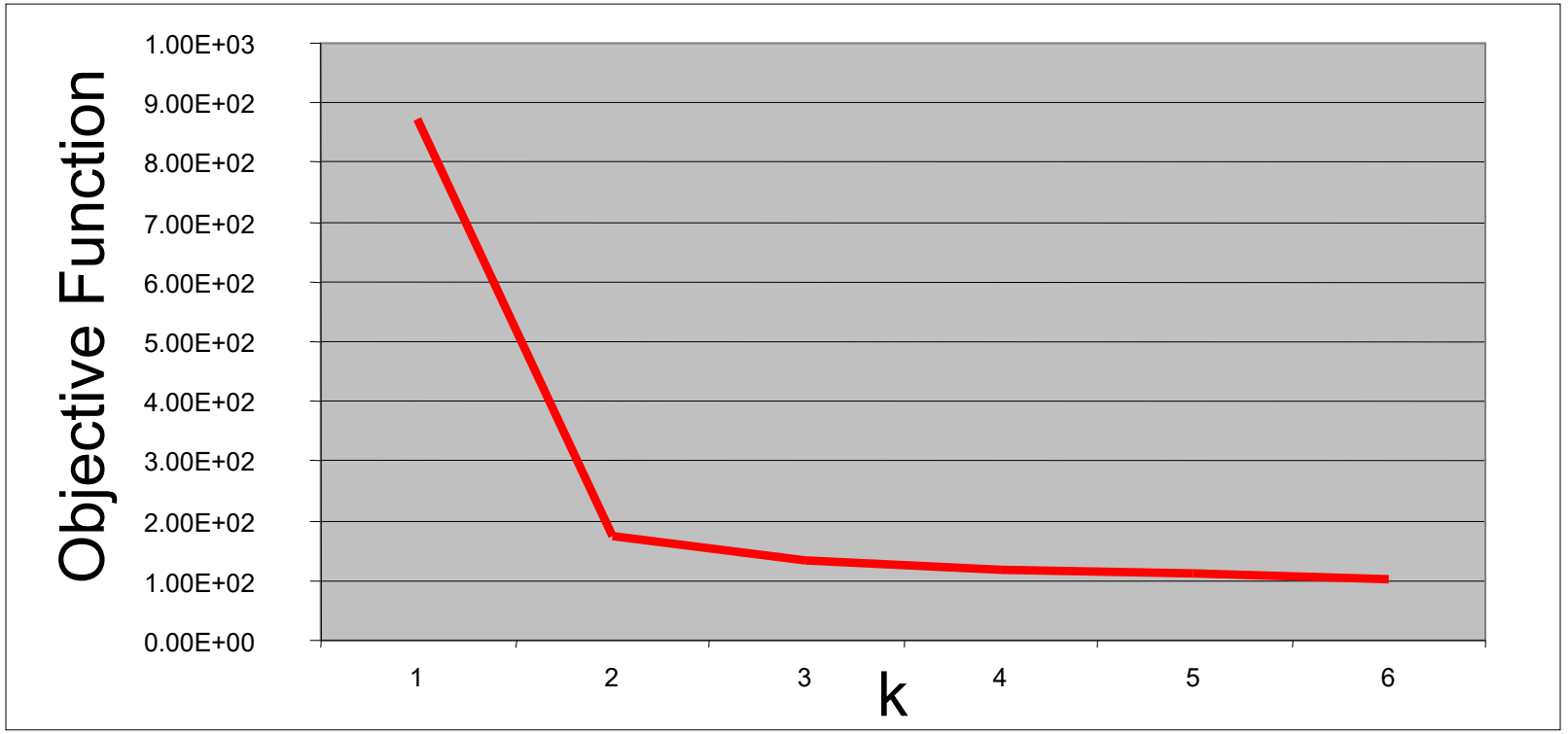
# What Is A Good "Clustering"?

- **Criterion**: A *good* clustering will produce high quality clusters in which:
  - Intra-class/cluster similarity is high
  - Inter-class/cluster similarity is low
  - Measured quality of a clustering depends on both pattern representation and similarity measure used

# The Elbow Method

- To determine *K*, plot *K*-means/medoids objective function values for *K* = 1, 2.,..,$K_{max}$

- Abrupt change at *K* = 2 (plot below) = highly suggestive of two clusters in data; Determining number of clusters also known as "*knee finding*" or "*elbow finding*"




- Other approaches might work better, e.g., Silhouette score, etc.

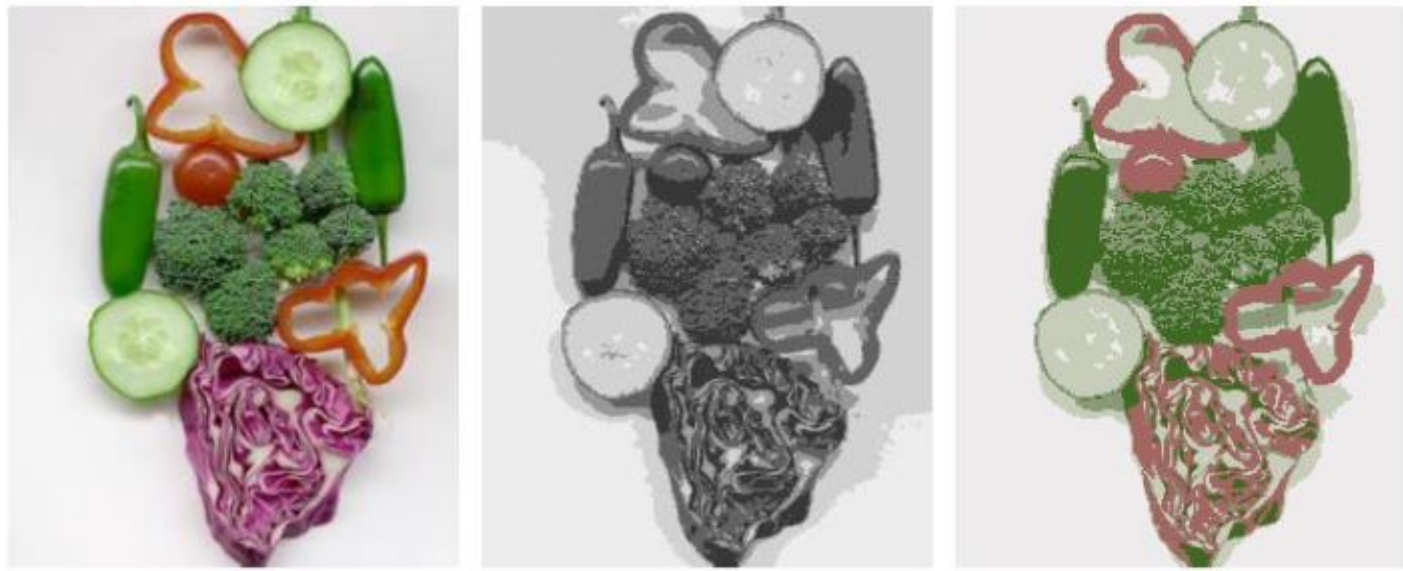


Image | Clusters on intensity | Clusters on color

Clustering using intensity only and color only

* Pictures from Mean Shift: A Robust Approach toward Feature Space Analysis, by D. Comaniciu and P. Meer http://www.caip.rutgers.edu/~comanici/MSPAMI/msPamiResults.html

# QUESTIONS?