TEXAS A&M UNIVERSITY
Engineering

ECEN 758 Data Mining and Analysis: Lecture 7, Representative Clustering II

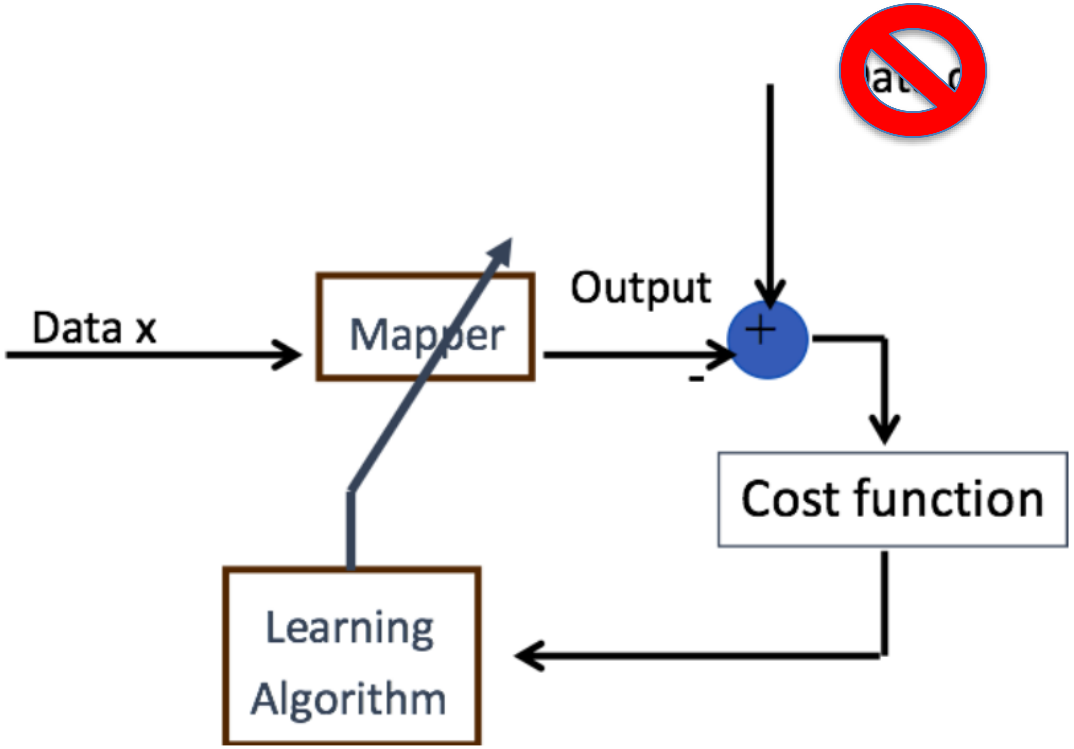
Joshua Peeples, Ph.D.
Assistant Professor
Department of Electrical and Computer Engineering

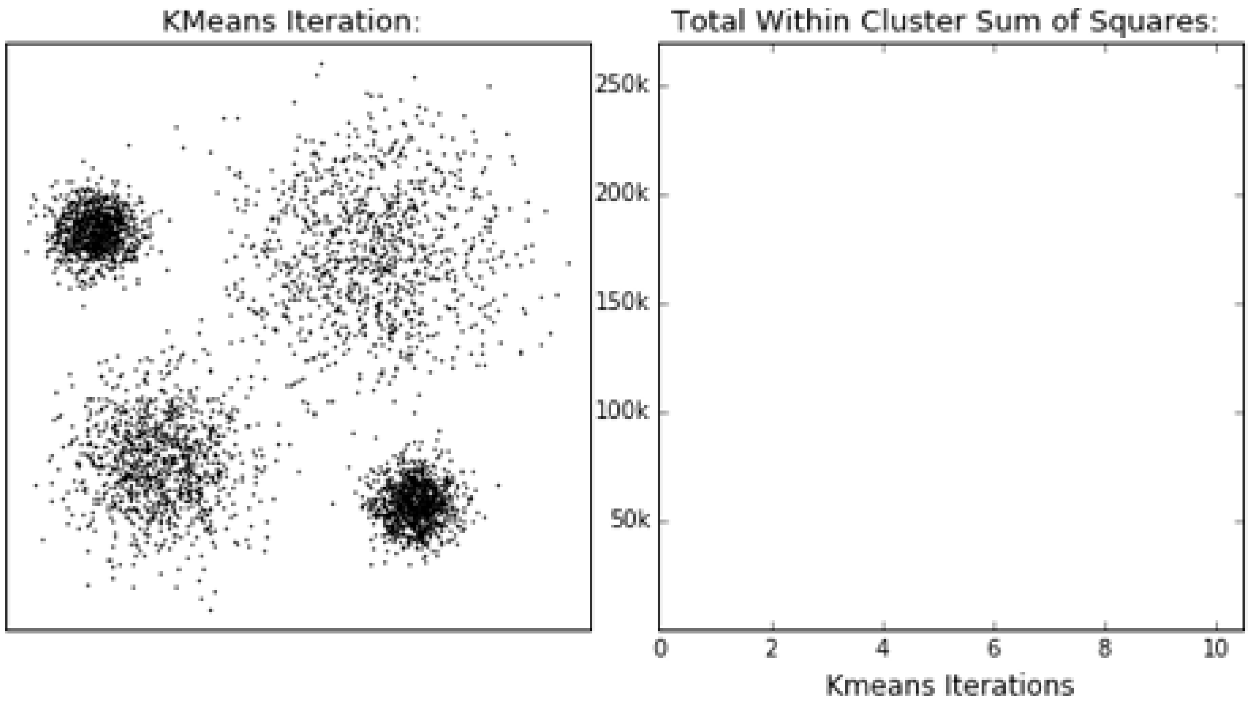
Announcements

- Assignment #1 solutions available on Canvas
- Assignment #2 will be released next Wednesday (09/18)
  - Please upload submission as single PDF
  - Please upload Python code (.py, ipynb)
  - Do not include screenshots of code in submission

Last Lecture

- Representative Clustering I





Gif from: D. Sheehan, Clustering with Scikit with GIFs

Today

- Representative Clustering II
- Reading: MMDS Chapter 7
- Supplemental reading: ZM Chapter 13 and 17

Unsupervised Learning: Clustering

Clustering Overview

- Clustering:
  - Unsupervised learning – just data, no labels
  - Similarity/Dissimilarity in the data
  - Can provide insights when we have no preconception of data

Clustering Overview

- We will discuss several variants of clustering
  - **Representative-based Clustering**
  - Hierarchical Clustering
  - Density-Based Clustering

Representative-based Clustering

Representative-based Clustering

- Goal: partition data into *k* groups or clusters
- Clusters:
  - Representative of data points in group (also called centroid)
  - Common choice is mean
- Brute force solution not ideal
  - Generate all possible partitions

$$\boldsymbol{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$$

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{x_j \in C_i} \boldsymbol{x}_j$$

k-Means Clustering Review

k-Means Algorithm: Objective

- Sum of squared errors (SSE) objective function
- Goal: find clustering to minimize SSE
- Greedy iterative approach
  - Can converge to a local optima
- Two steps to achieve minima

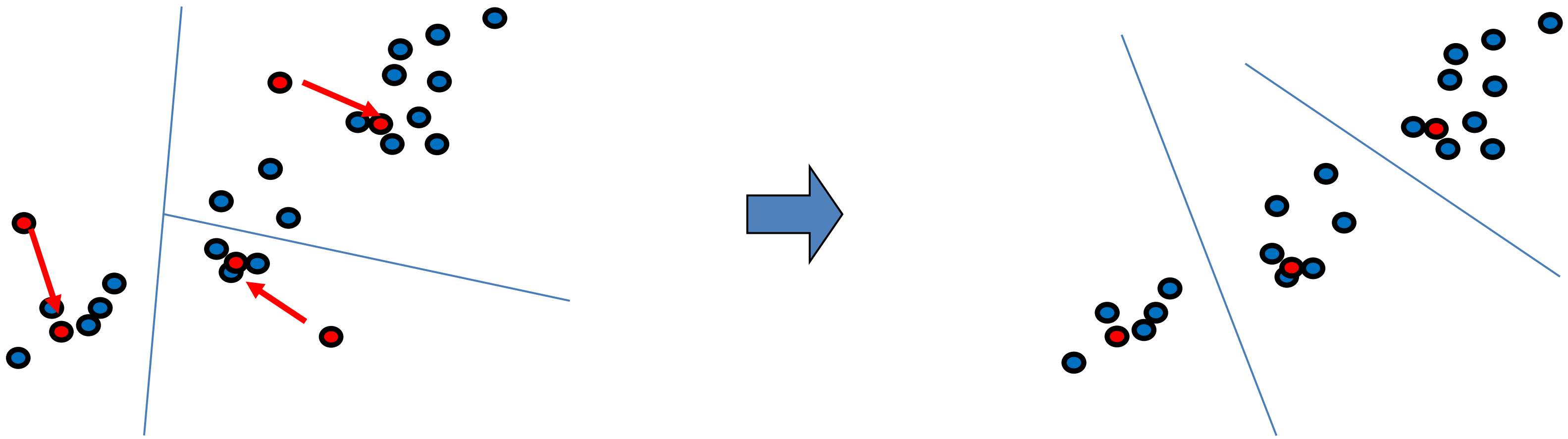
$$SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \{SSE(\mathcal{C})\}$$

Phase I: Update Assignments

- For each point, re-assign to closest mean: $a_{ij} = \underset{k}{argmin}\, dist(x_i, c_k)$

- Choose among $[c_1, .. \, c_k]$ the mean which minimizes the distance between x_i and c_k, and assign that value of [1..k] to a_{ij}

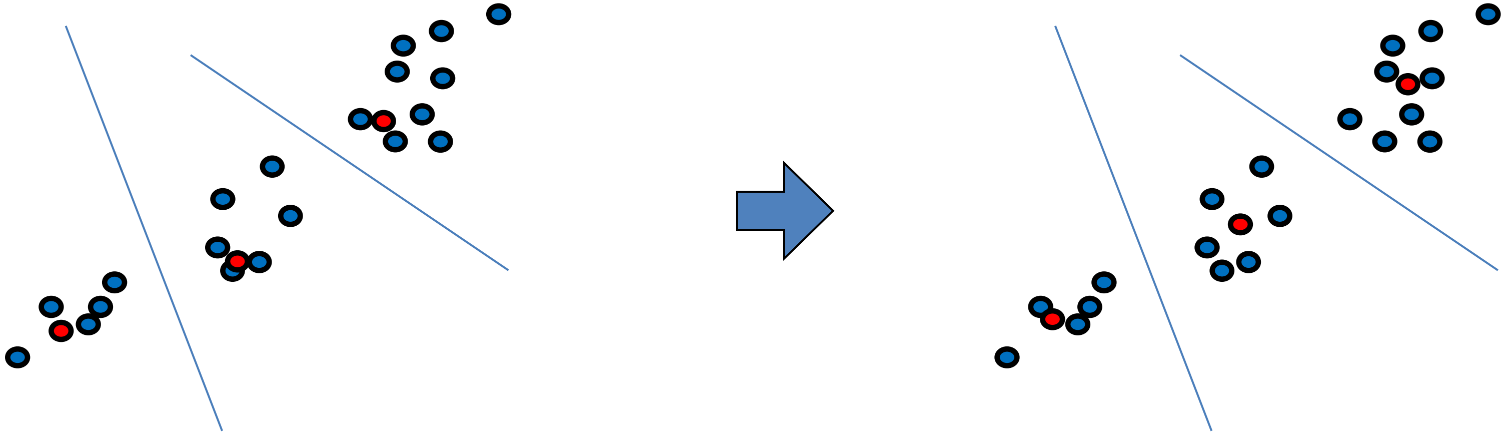


Phase II: Update Means

TEXAS A&M UNIVERSITY Engineering

- Move each mean to the average of its assigned points:
- Select the points which are assigned to the mean point c_k (i.e. those with $a_{ij} = k$.) Average these points and assign that new value to c_k

$$c_k = \frac{1}{|\{i: a_{ij} = k\}|} \sum_{i:a_{ij}=k} \boldsymbol{x}_i$$



What are disadvantages of k-means?

k-Means Disadvantages

TEXAS A&M UNIVERSITY Engineering

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing (not ideal for large datasets)
- Selecting number of clusters (*k*)
- “Crisp”/Hard clustering

k-Means Disadvantages

- **Linear boundaries between clusters**
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing
- Selecting number of clusters (*k*)
- “Crisp”/Hard clustering

k-Means Disadvantage: Linear Boundaries

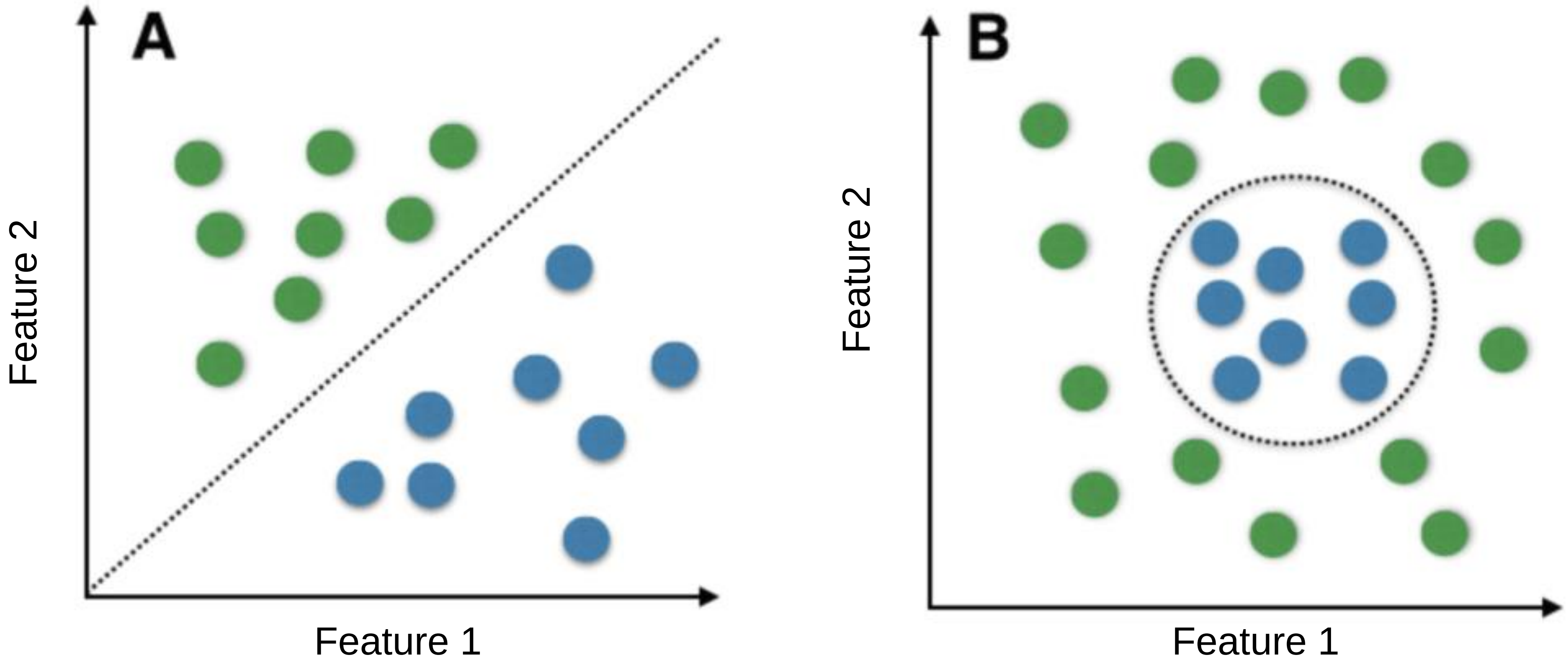


Image from: Alokesh985, Introduction to Support Vector Machines (SVM).

k-Means Disadvantages

- Linear boundaries between clusters
- **Only uses Euclidean distance**
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing
- Selecting number of clusters (*k*)
- "Crisp"/Hard clustering

k-Means Disadvantage: Euclidean Distance

- Spherical cluster assumption:
  - Radius equal to the distance between the centroid and the furthest data point
- Sensitive to points far away (i.e., outliers)

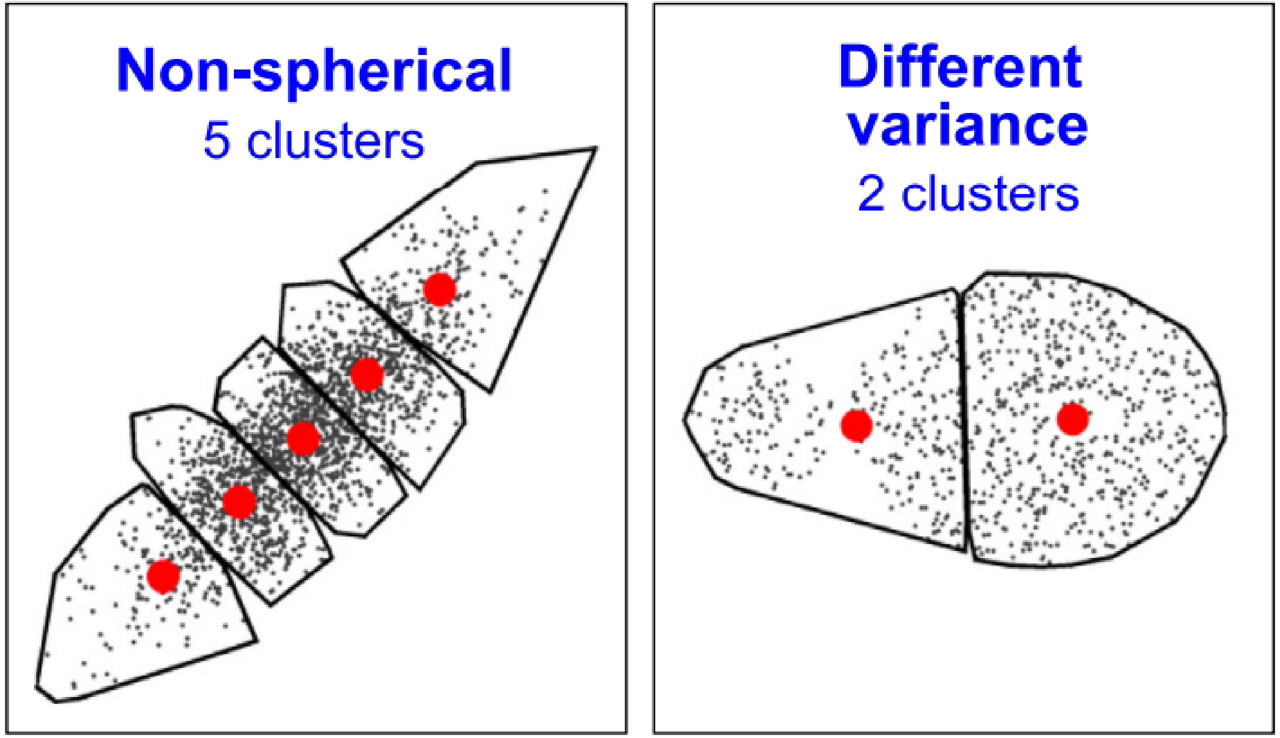


Image from: Fränti, Pasi, and Sami Sieranoja. "How much can k-means be improved by using better initialization and repeats?." Pattern Recognition 93 (2019): 95-112.

k-Means Disadvantages

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- **Non-symmetrical clusters**
- Initialization
- Batch processing
- Selecting number of clusters (*k*)
- “Crisp”/Hard clustering

k-Means Disadvantage: Non-symmetrical clusters

- Difficulty with clustering data of different sizes and density

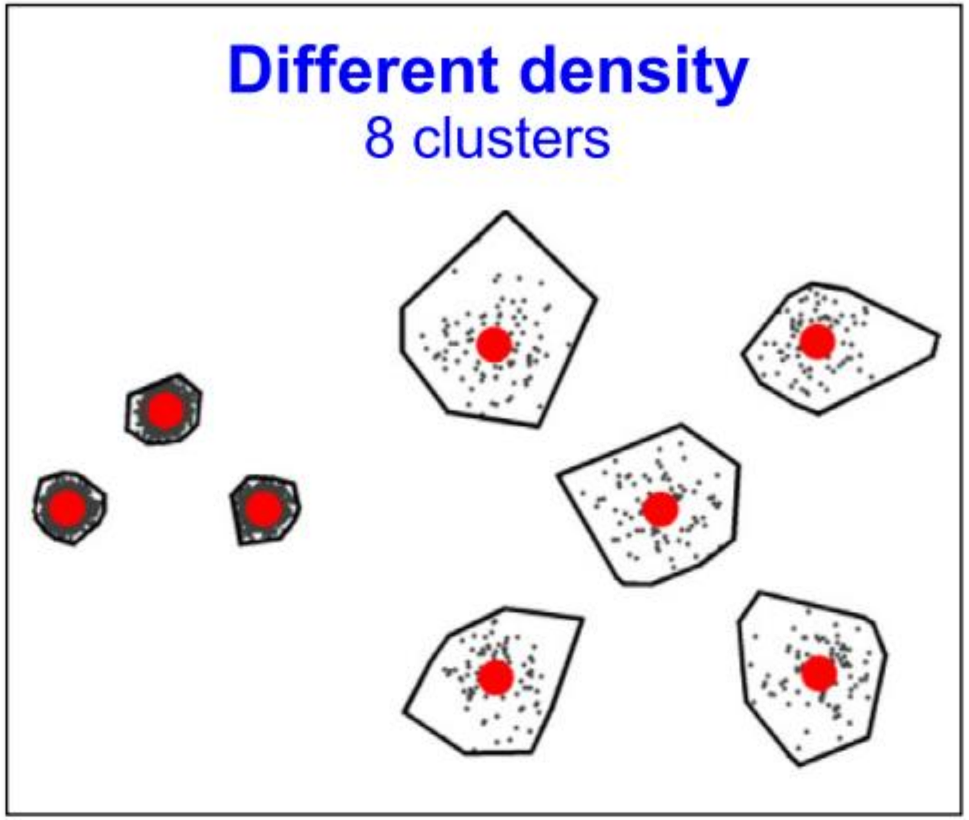


Image from: Fränti, Pasi, and Sami Sieranoja. "How much can k-means be improved by using better initialization and repeats?." Pattern Recognition 93 (2019): 95-112.

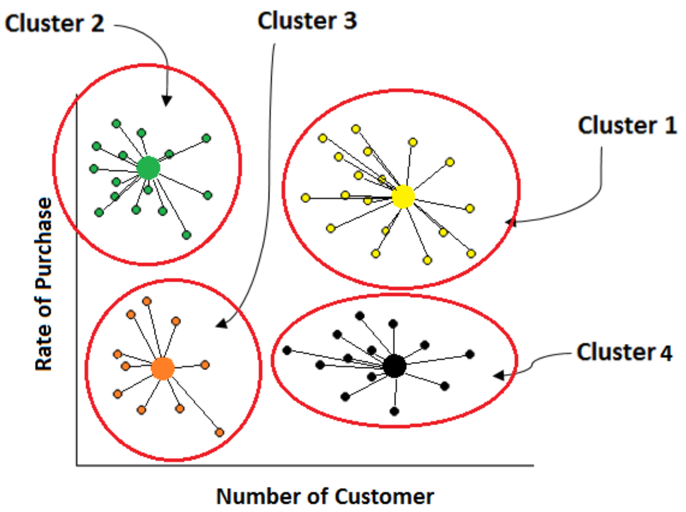
k-Means Disadvantages

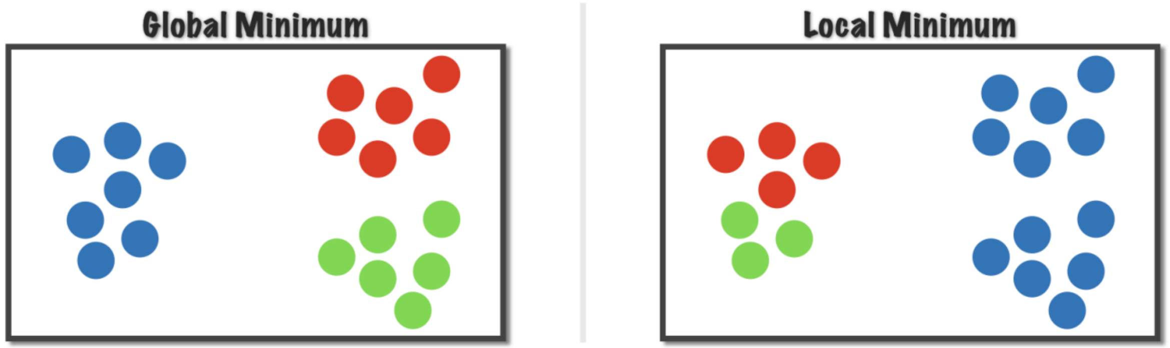
TEXAS A&M UNIVERSITY Engineering

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- **Initialization**
- Batch processing
- Selecting number of clusters (*k*)
- “Crisp”/Hard clustering

k-Means Disadvantage: Initialization

- Non-deterministic approach
- May get stuck in local optima





Images from: Amit Chauchan, Fully Explained K-means Clustering with Python and Alan Jeffares, K-means: A Complete Introduction

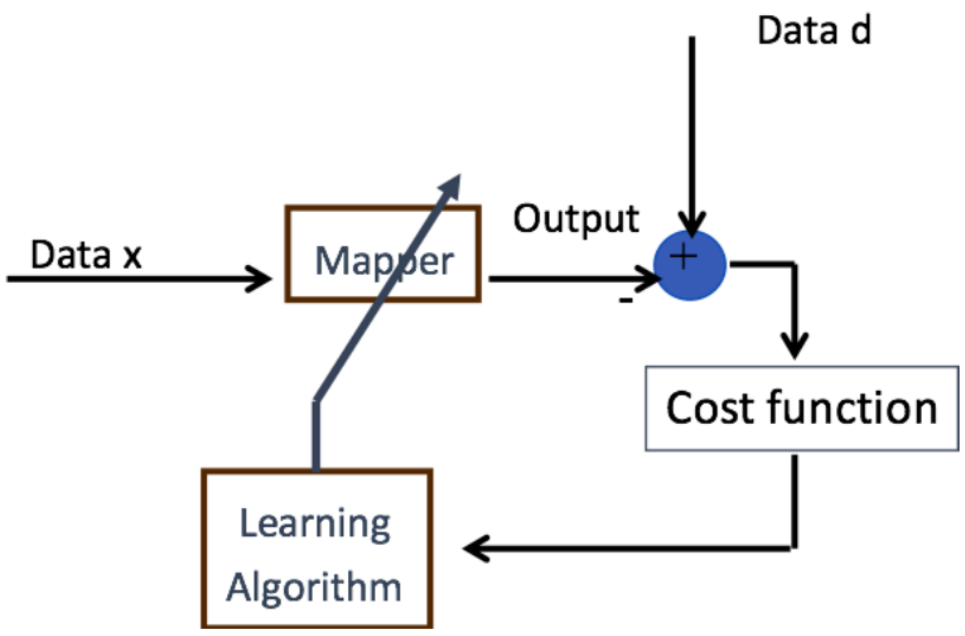
k-Means Disadvantages

TEXAS A&M UNIVERSITY Engineering

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- **Batch processing**
- Selecting number of clusters (*k*)
- “Crisp”/Hard clustering

Batch Training

- Batching relies on accumulating errors over multiple training observations (“batch”) prior to updating model parameters
- Batching is controlled by an additional hyperparameter (e.g., batch size)
- Three batch modes:
  - Online (one training sample)
  - **Batch** (all training samples)
  - Mini-batch (subset of training samples)



k-Means Disadvantage: Batch Processing

TEXAS A&M UNIVERSITY Engineering

- Batch size = All training data
- Advantage: “Smoother” training
- Disadvantage: Usually converges to local optima, computationally expensive (memory)

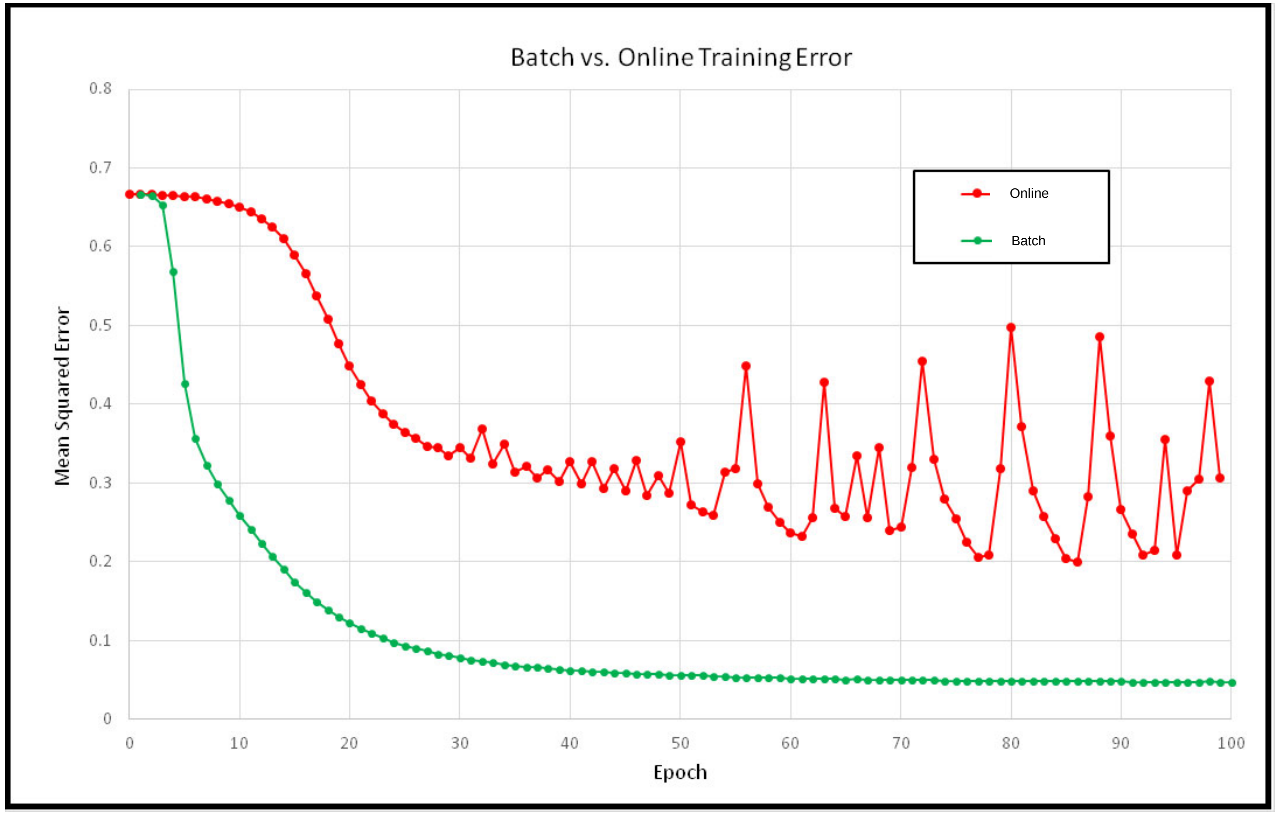


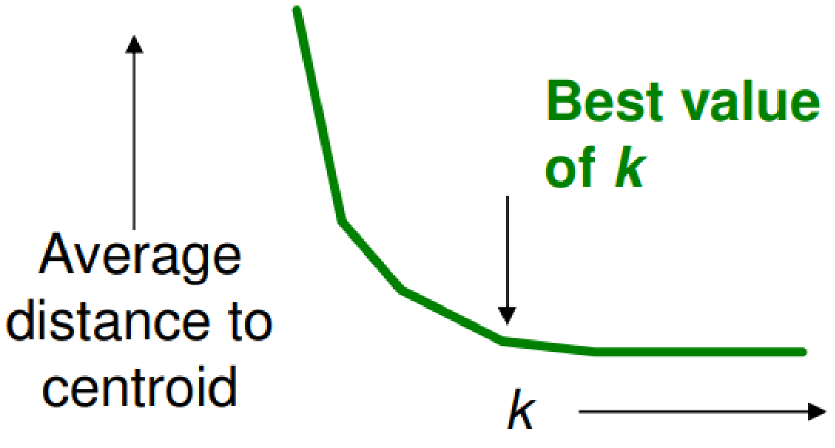
Image from: James McCaffery, Understanding Neural Network Batch Training: A Tutorial.

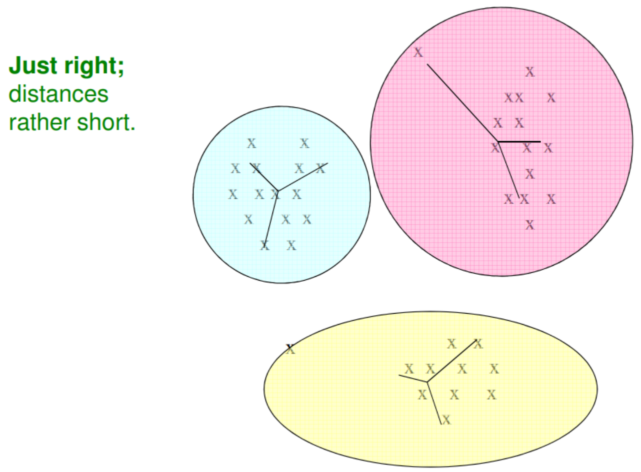
k-Means Disadvantages

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing
- **Selecting number of clusters (*k*)**
- “Crisp”/Hard clustering

k-Means Disadvantage: Choosing *k*

- *k* is hyperparameter to determine number of clusters
- Results heavily dependent on *k*
- Selecting *k*
  - Try different values and look at change in average distance to centroid
    - Average falls rapidly until right *k*, then changes little ("elbow method")





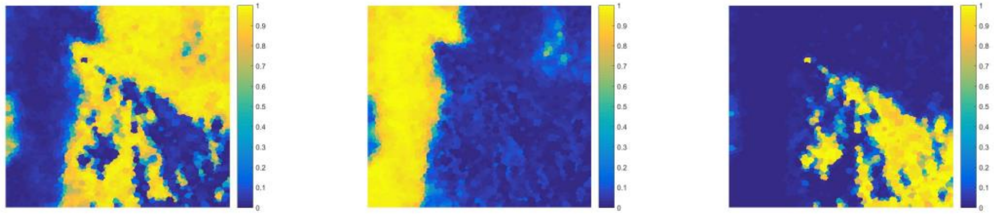
k-Means Disadvantages

- Linear boundaries between clusters
- Only uses Euclidean distance
  - Assumes spherical clusters
  - Sensitive to outliers
- Non-symmetrical clusters
- Initialization
- Batch processing
- Selecting number of clusters (*k*)
- **“Crisp”/Hard clustering**

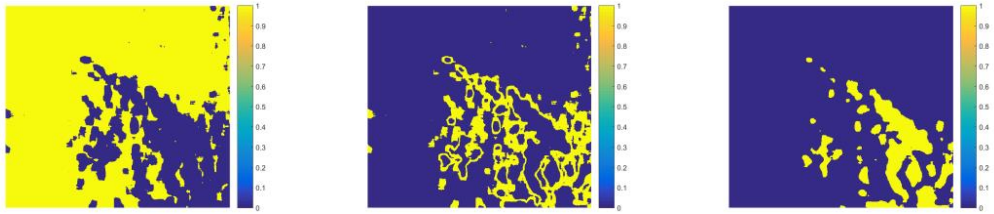
k-Means Disadvantage: “Crisp”/Hard Clustering

- Points can only “belong” to one cluster
- Different applications may require “soft” clustering
    - Points may belong to more than one group

Input Image



(h) FLICM Cluster 1 (i) FLICM Cluster 2 (j) FLICM Cluster 3



(k) K-Means Cluster 1 (l) K-Means Cluster 2 (m) K-Means Cluster 3

J. Peeples, et al. Possibilistic fuzzy local information C-means with automated feature selection for seafloor segmentation. SPIE, 2018.

Extensions of k-Means

Extensions of k-Means

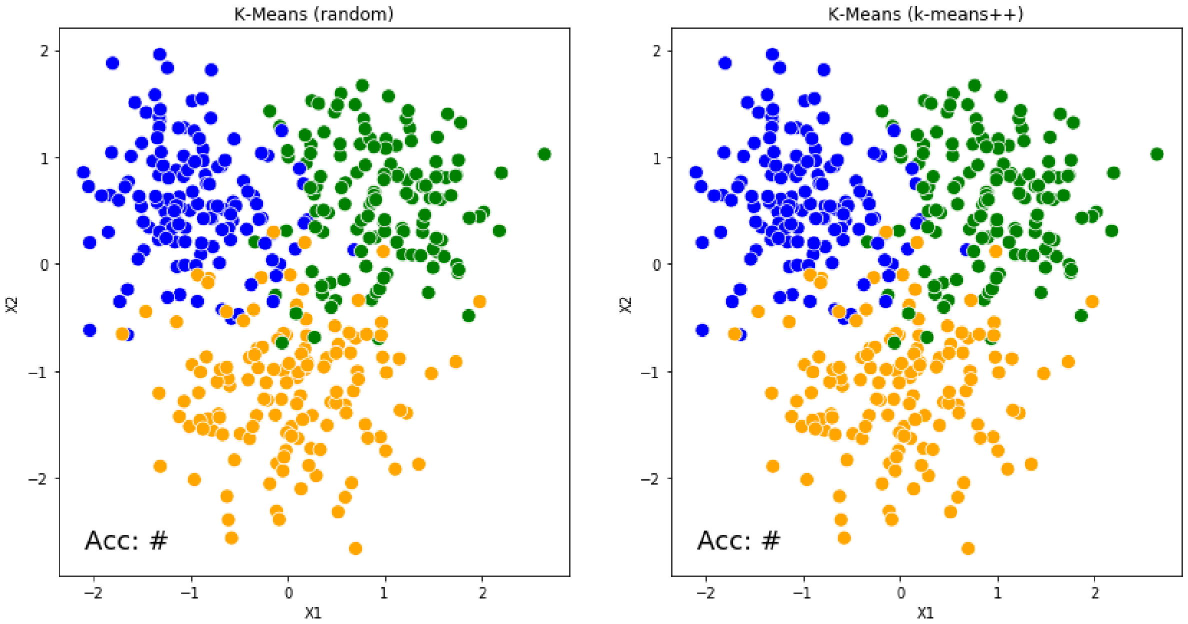
- k-Means++
- Kernel k-Means
- Mini-batch k-Means
- Bradley-Fayyad-Reina (BFR) Algorithm
- Clustering Using Representatives
- Alternative representative clustering approaches
    - k-Mediods, Affinity Propagation, Gaussian Mixture Models

Extensions of k-Means

- **k-Means++**
- Kernel k-Means
- Mini-batch k-Means
- Bradley-Fayyad-Reina (BFR) Algorithm
- Clustering Using Representatives (CURE)
- Alternative representative clustering approaches
    - k-Mediods, Affinity Propagation, Gaussian Mixture Models

k-Means++

- Used to improve initialization
- Pick centroids far from one another
- Steps:
  - Select initial center from data point at random
  - Select next center based on proximity
  - Repeat until *k* centers are chosen
  - Apply standard *k*-Means algorithm



$$\frac{D(x')^2}{\sum_{x \in \mathcal{X}} D(x)^2}$$

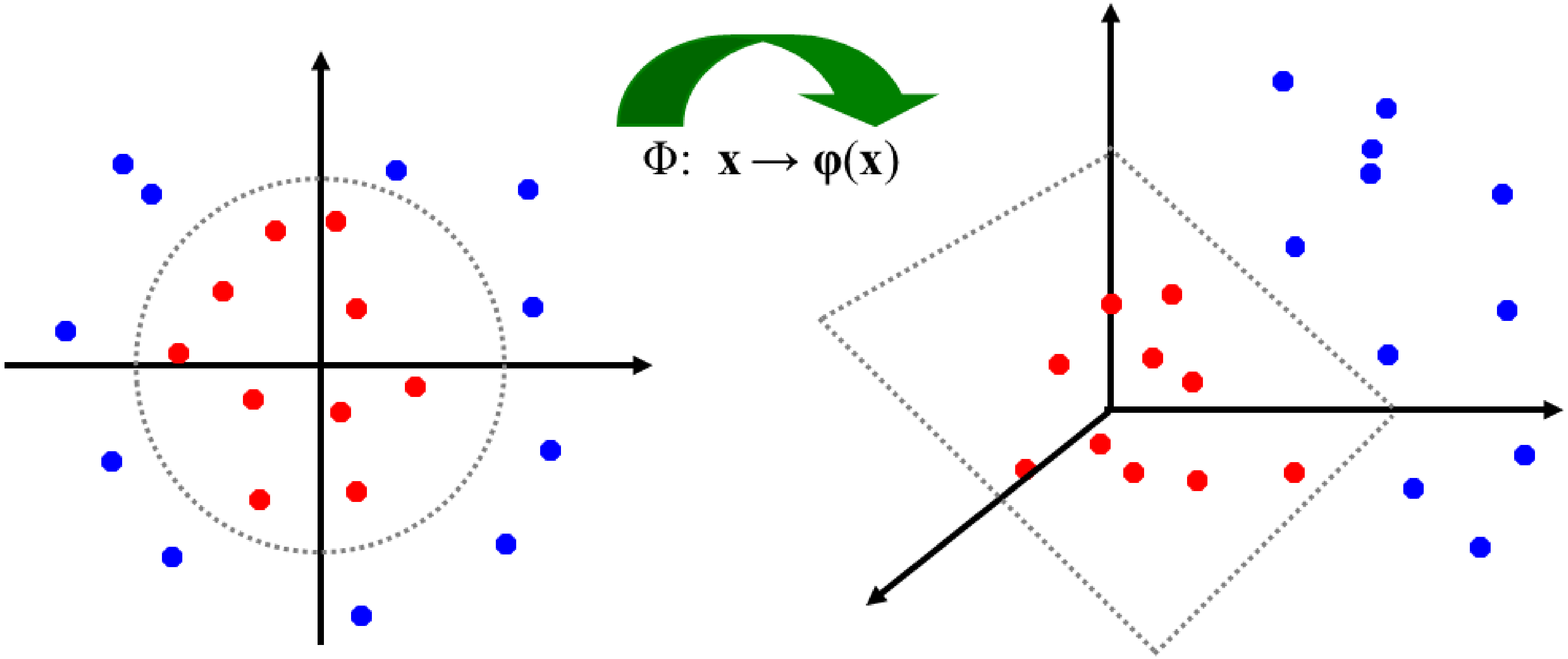
Gif from: F. Busko, Animation of k-Means Clustering

Extensions of k-Means

- k-Means++
- **Kernel k-Means**
- Mini-batch k-Means
- Bradley-Fayyad-Reina (BFR) Algorithm
- Clustering Using Representatives (CURE)
- Alternative representative clustering approaches
  - k-Mediods, Affinity Propagation, Gaussian Mixture Models

Kernel Trick

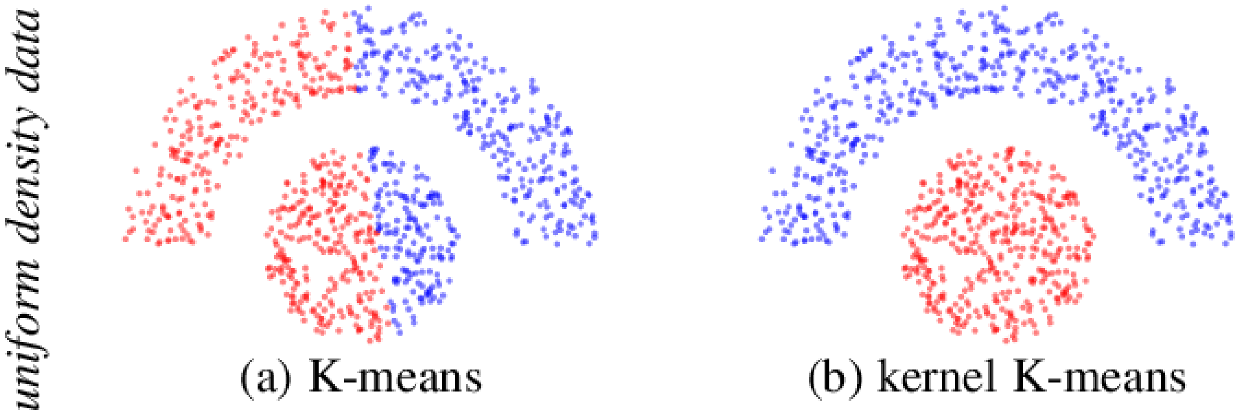
- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



[Source: Ray Mooney, UT]

Kernel k-Means

- Use “kernel trick” on data to extract nonlinear boundaries
- Can rewrite objective in terms of kernel function



$$\min_{\mathcal{C}} SSE(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \phi(\mathbf{x}_j) - \boldsymbol{\mu}_i^{\phi} \right\|^2 = \sum_{j=1}^{n} K(\mathbf{x}_j, \mathbf{x}_j) - \sum_{i=1}^{k} \frac{1}{n_i} \sum_{\mathbf{x}_a \in C_i} \sum_{\mathbf{x}_b \in C_i} K(\mathbf{x}_a, \mathbf{x}_b)$$

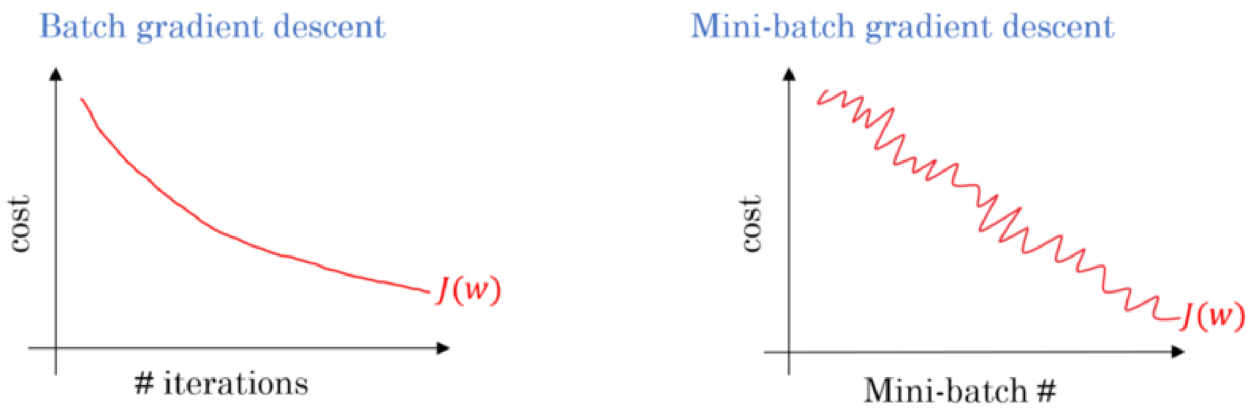
Image from: D. Marin, et al, Kernel clustering: Density biases and solutions.

Extensions of k-Means

- k-Means++
- Kernel k-Means
- **Mini-batch k-Means**
- Bradley-Fayyad-Reina (BFR) Algorithm
- Clustering Using Representatives (CURE)
- Alternative representative clustering approaches
  - k-Mediods, Affinity Propagation, Gaussian Mixture Models

Mini-batch k-Means

- Batch size = selected by user
- Trade off between online and batch learning
- Smaller batches = more randomness
- Large batches = "smoother" training



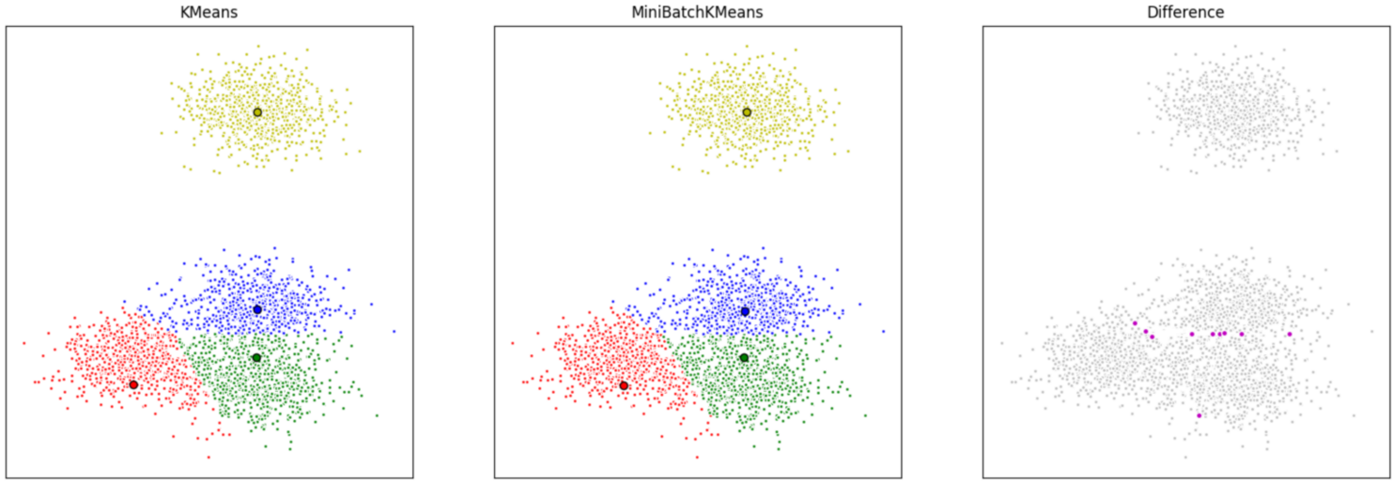


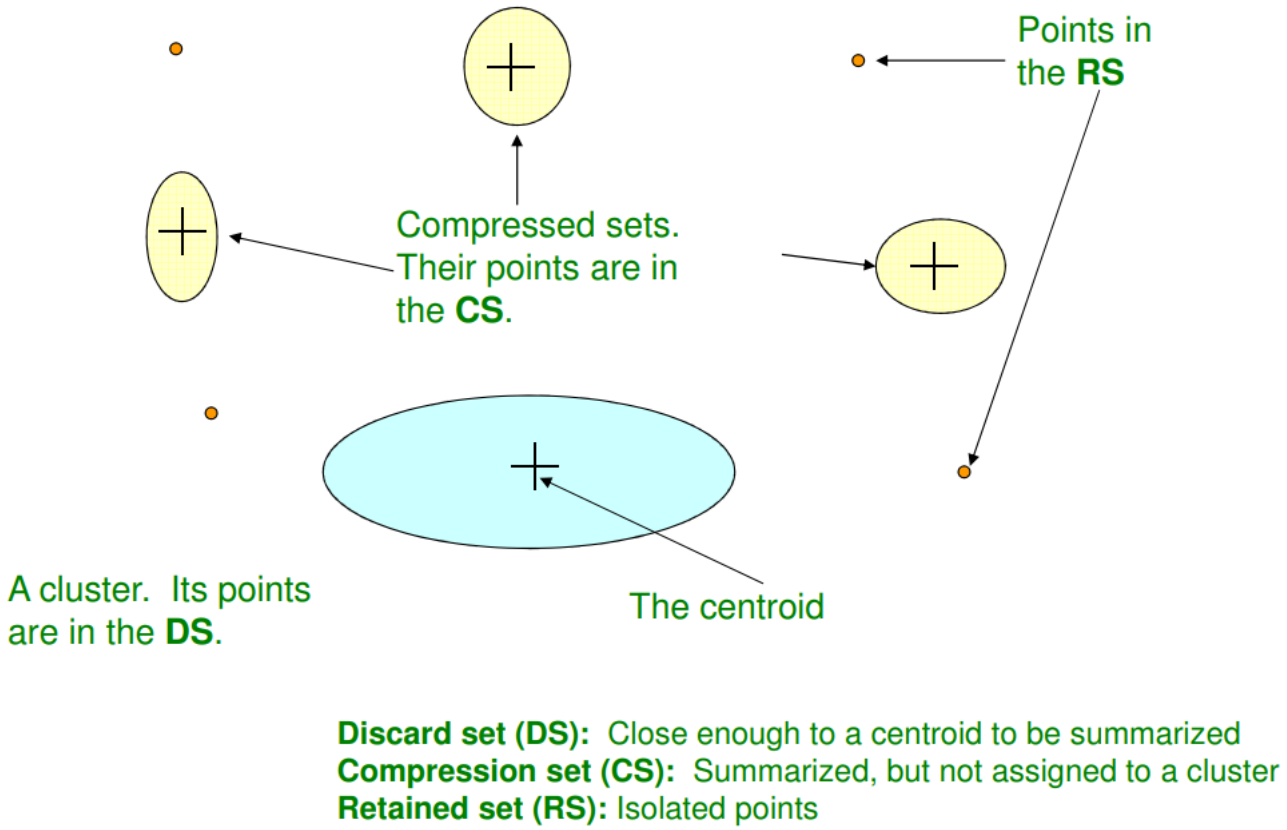
Image from: Cross Validated, Understanding mini-batch gradient descent.

Extensions of k-Means

- k-Means++
- Kernel k-Means
- Mini-batch k-Means
- **Bradley-Fayyad-Reina (BFR) Algorithm**
- Clustering Using Representatives (CURE)
- Alternative representative clustering approaches
  - k-Mediods, Affinity Propagation, Gaussian Mixture Models

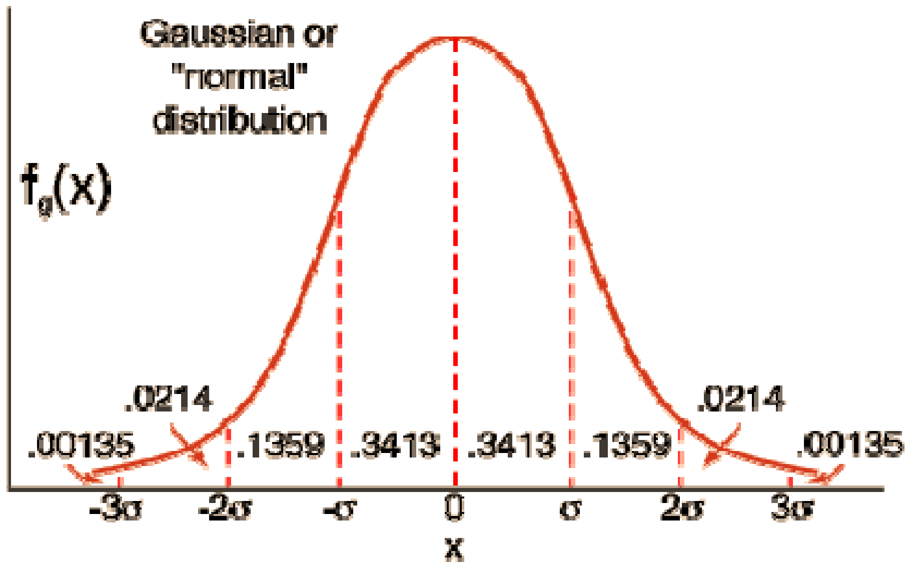
Bradley-Fayyad-Reina (BFR) Algorithm

- Extension to *k*-means to large data
- Clusters assumed to be normally distributed
- Three data points:
  - Discard set
  - Compression set
  - Retained set



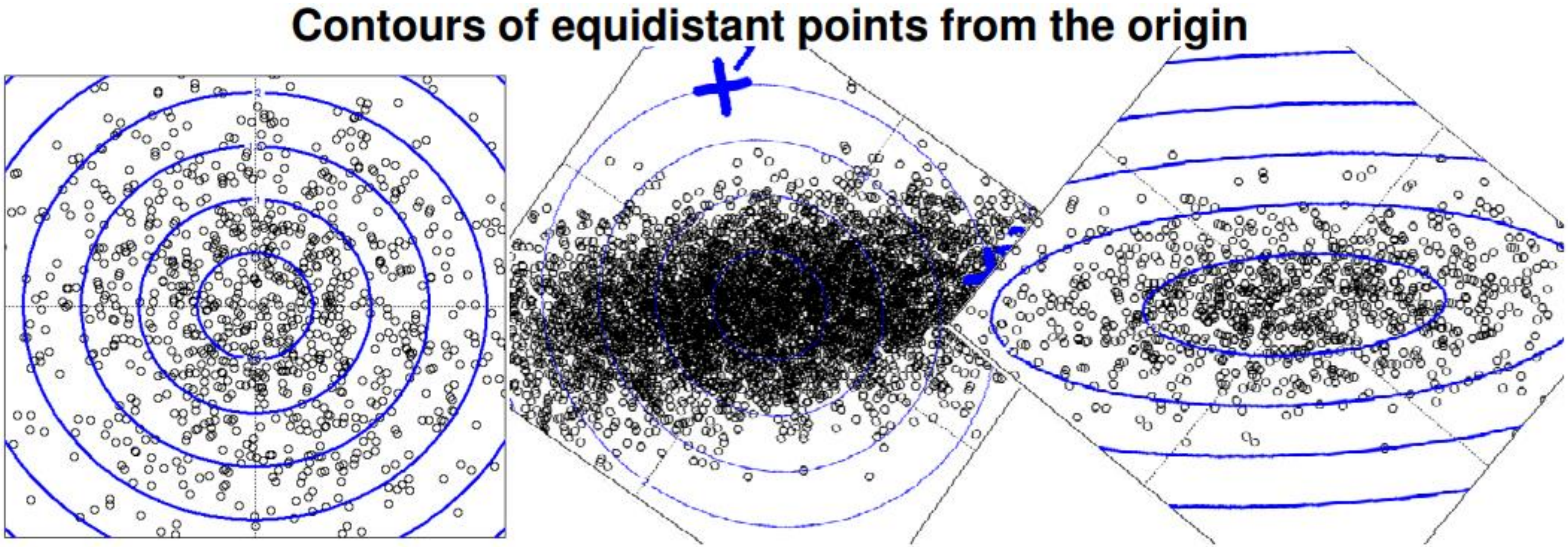
Bradley-Fayyad-Reina (BFR) Algorithm

- Use Mahalanobis distance to decide "closeness"
- High likelihood of the point belonging to current nearest centroid



$$(\mathbf{x}-\mu)^T\,\Sigma^{-1}\,(\mathbf{x}-\mu)$$

Bradley-Fayyad-Reina (BFR) Algorithm



Uniformly distributed points, Euclidean distance

Normally distributed points, Euclidean distance

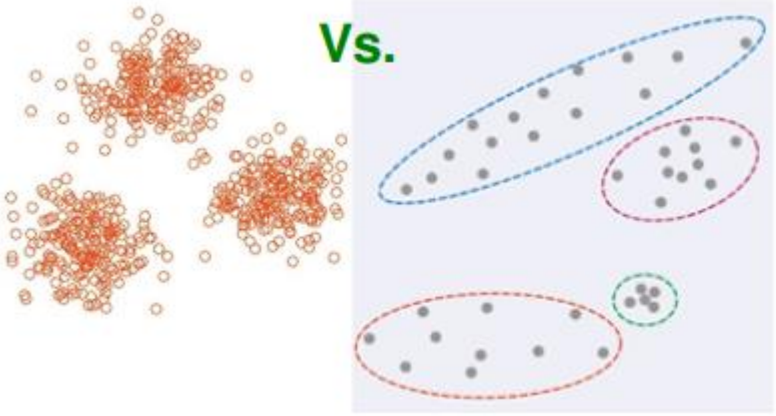
Normally distributed points, Mahalanobis distance

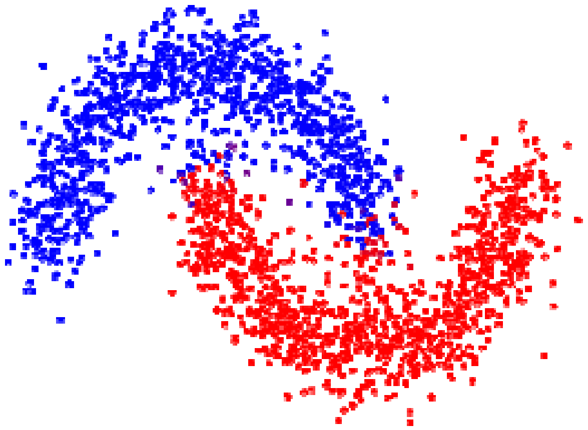
Extensions of k-Means

- k-Means++
- Kernel k-Means
- Mini-batch k-Means
- Bradley-Fayyad-Reina (BFR) Algorithm
- **Clustering Using Representatives (CURE)**
- Alternative representative clustering approaches
  - k-Mediods, Affinity Propagation, Gaussian Mixture Models

Clustering Using Representatives (CURE)

- BFR/k-Means assume normally distributed clusters in each dimension
- CURE algorithm
  - Assumes Euclidean distance
  - Allows for cluster of any shape
  - Use collection of representative points to represent cluster





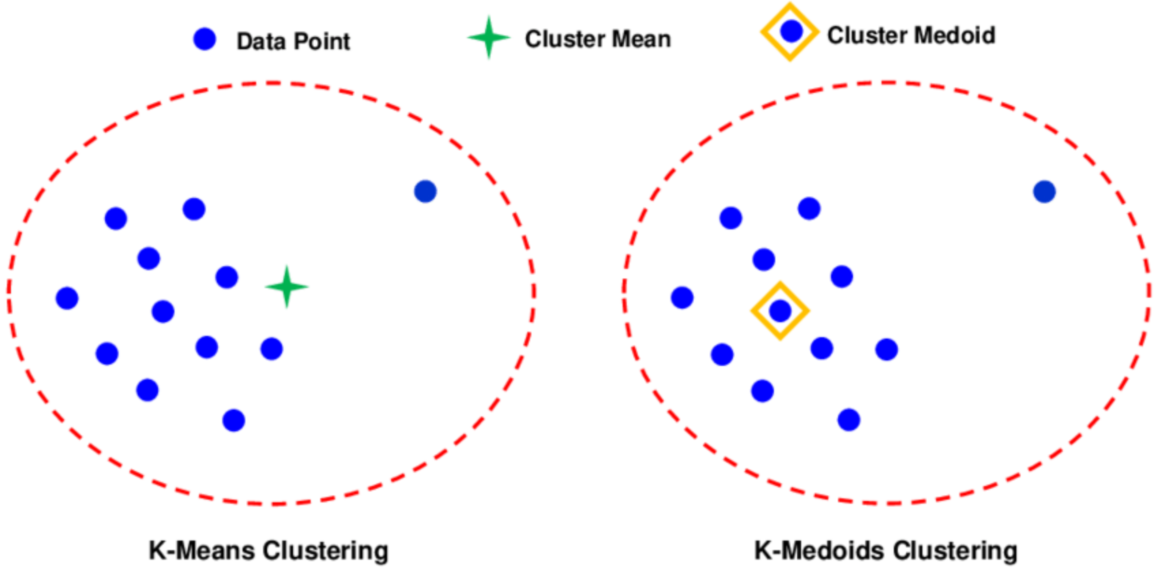
Extensions of k-Means

TEXAS A&M UNIVERSITY Engineering

- k-Means++
- Kernel k-Means
- Mini-batch k-Means
- Bradley-Fayyad-Reina (BFR) Algorithm
- Clustering Using Representatives (CURE)
- **Alternative representative clustering approaches**
    - k-Mediods, Affinity Propagation, Gaussian Mixture Models

k-Mediods

- "Partitioning Around Mediod" (PAM)
- Mediod is point with minimal dissimilarity with all other points in cluster (exemplars)
- Can use any distance measure
- More robust to outliers



$$Cost(C^1, \ldots, C^k, z^{(1)}, \ldots, z^{(k)}) = \sum_{j=1}^{k} \sum_{i \in C^j} d(x^{(i)}, z^{(j)})$$

$$\text{exemplars:} \quad \{z^{(1)}, \ldots, z^{(k)}\} \subseteq \{x^{(1)}, \ldots, x^{(n)}\}$$

Image from: A. Entezami, An innovative hybrid strategy for structural health monitoring by modal flexibility and clustering methods.

k-Mediods

- "Partitioning Around Mediod" (PAM)
- Mediod is point with minimal dissimilarity with all other points in cluster (exemplars)
- Can use any distance measure
- More robust to outliers

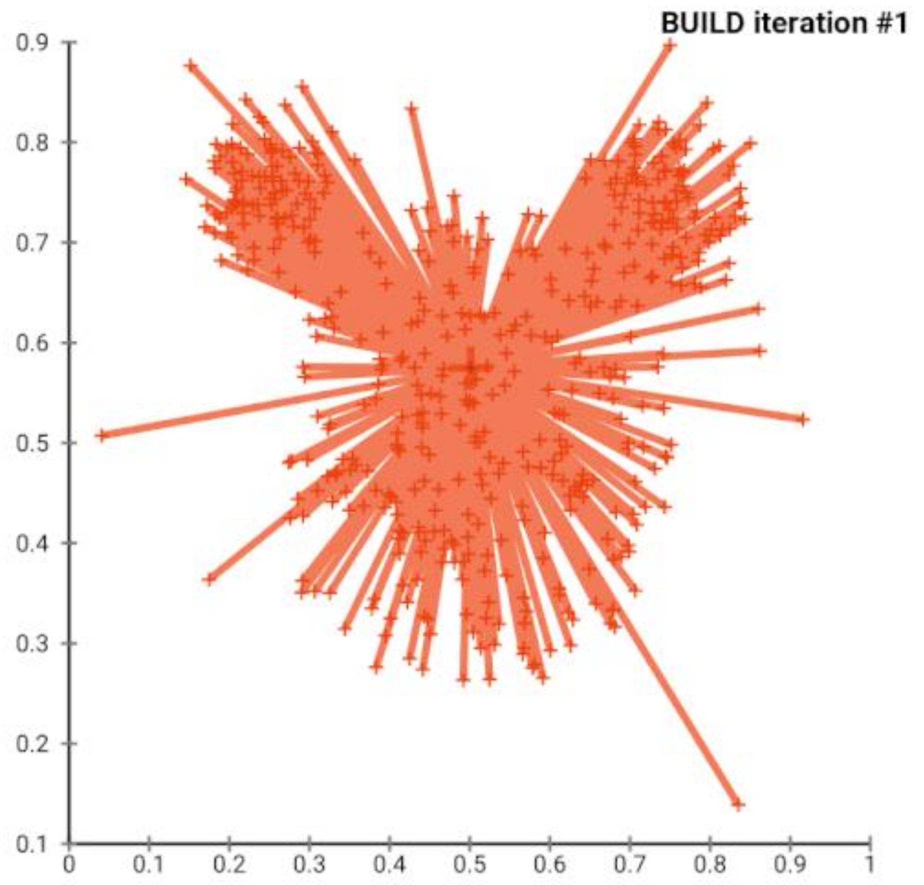
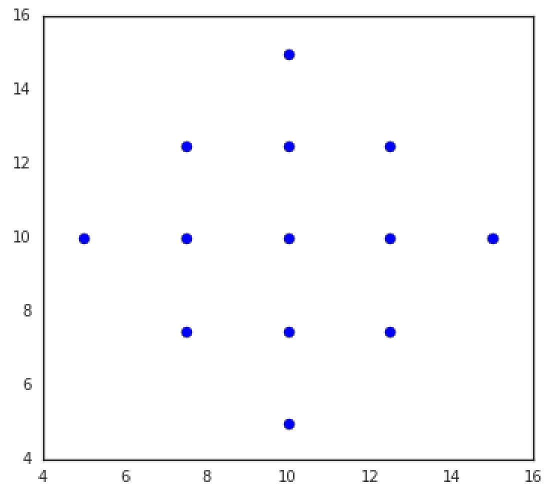
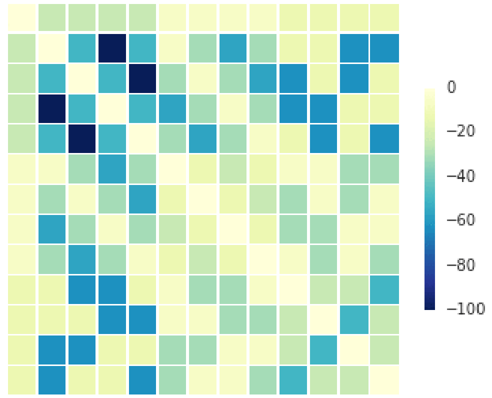


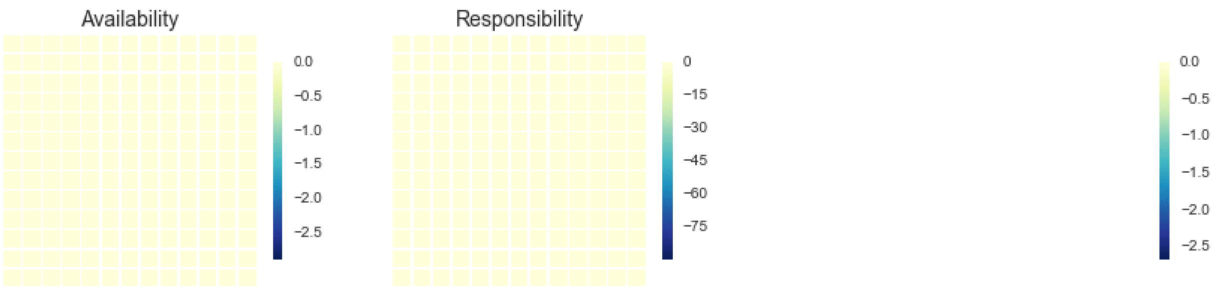
Image from: Z. Ahmed, K Medoids Clustering — An approach to Unsupervised Learning Algorithm.

Affinity Propagation

- Cluster centers are data points (exemplars)
- Do not need to specify number of clusters!
  - Still must set two hyperparameters: preference and damping
- Uses three matrices:
  - Similarity
  - Availability
  - Responsibility

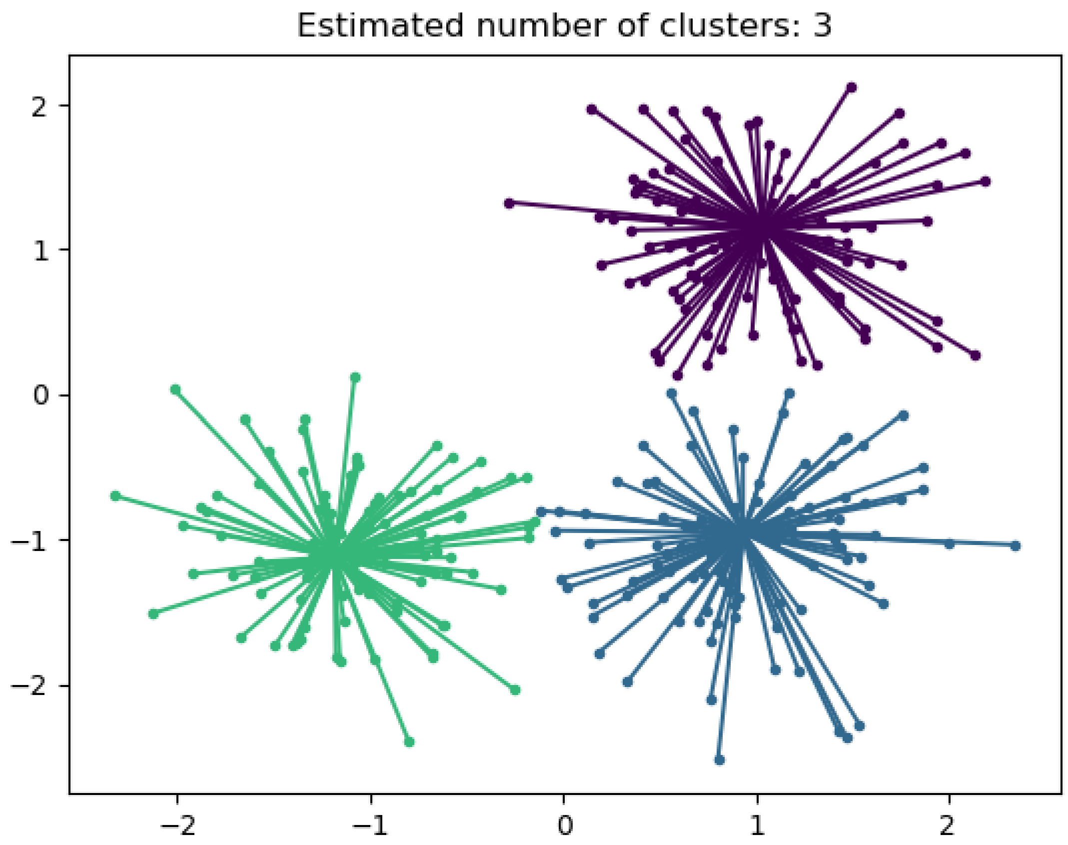






Affinity Propagation

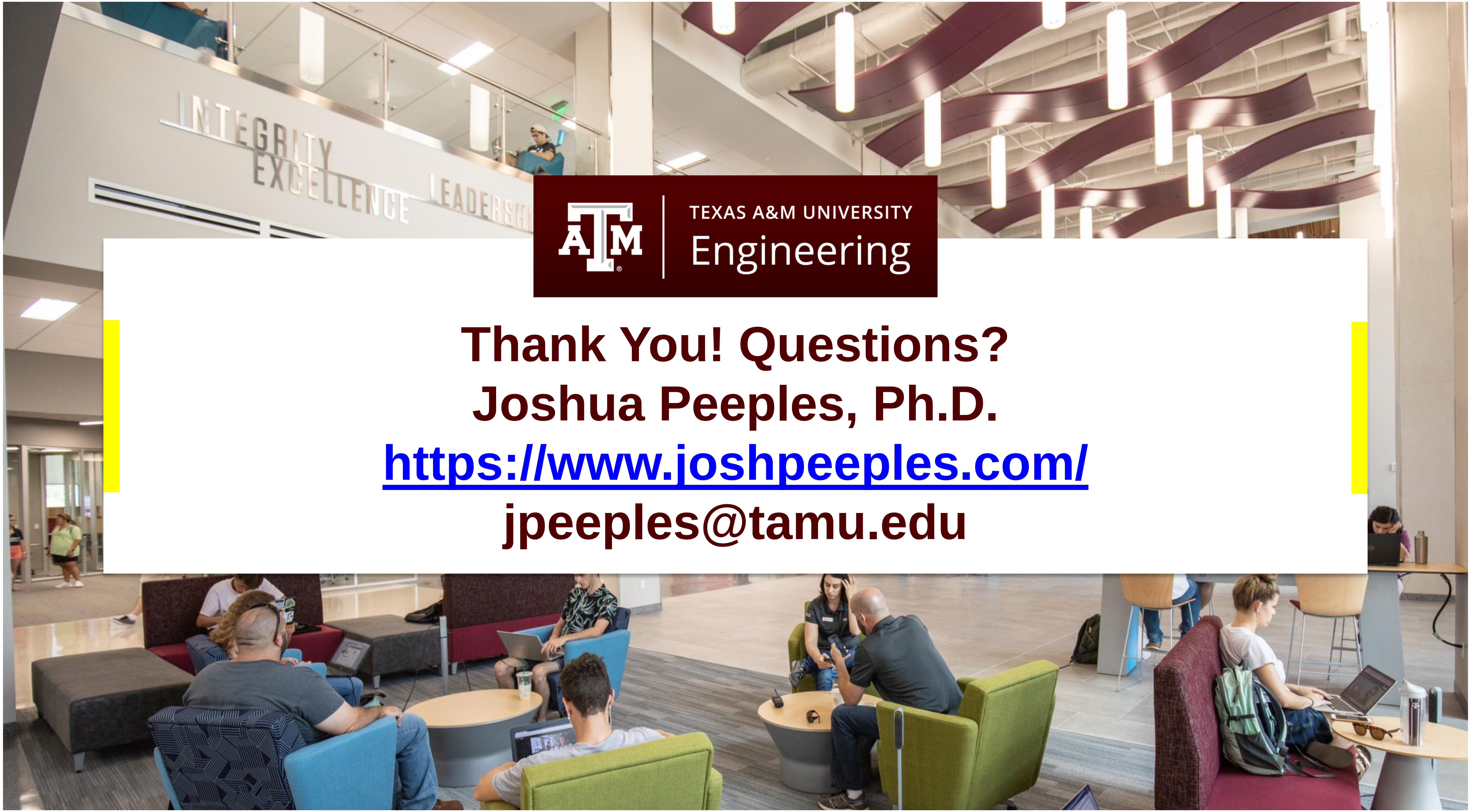
- Cluster centers are data points (exemplars)
- Do not need to specify number of clusters!
  - Still must set two hyperparameters: preference and damping
- Uses three matrices:
  - Similarity
  - Availability
  - Responsibility
- Deterministic



Next class

TEXAS A&M UNIVERSITY Engineering

- Gaussian Mixture Models



TEXAS A&M UNIVERSITY
Engineering

Thank You! Questions?

Joshua Peeples, Ph.D.
https://www.joshpeeples.com/
jpeeples@tamu.edu

TEXAS A&M UNIVERSITY Engineering

Supplemental Slides

Useful Links

TEXAS A&M UNIVERSITY Engineering

- Clustering Algorithms Overview
- Sklearn Clustering
- Kernel k-Means implementation