Robert Collins
CSE586

K-Means Intro

Reading: Chapter 13.4.4, Prince book

K-means Introduction

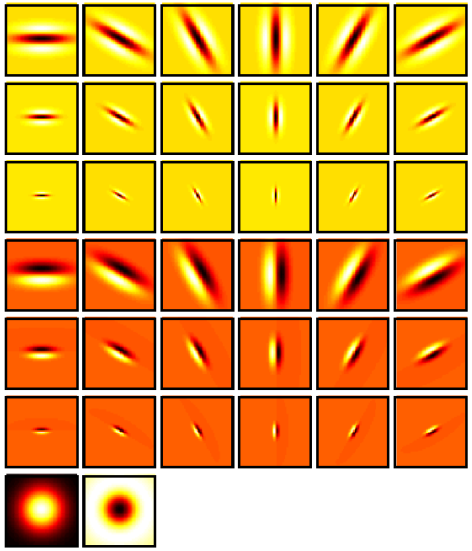
- K-means is a well-known method of clustering data.
- Determines location of clusters (cluster centers), as well as which data points are "owned" by which cluster.
- Motivation: Commonly used to build a vocabulary of visual words prior to bag of words processing

Assumptions

- You know how many clusters you want.
- Clusters are roughly spherical
- Radius of each cluster is roughly equal
- Number of points in each cluster is roughly the same

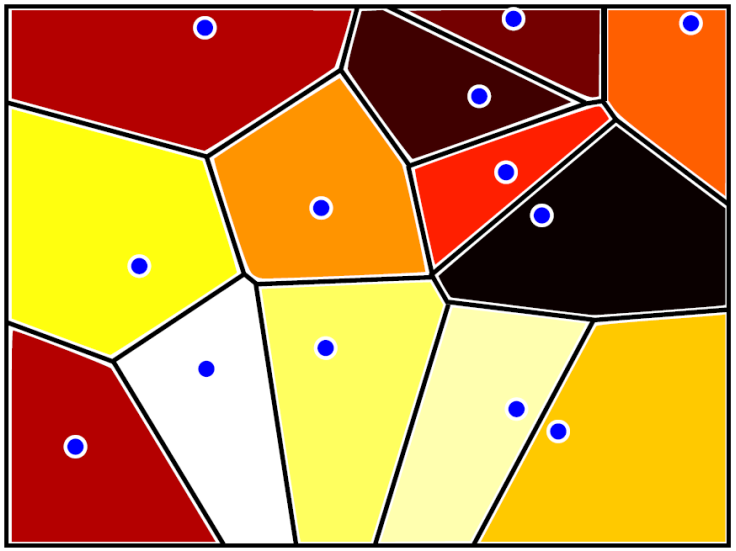
Use for Computing Visual Words

Filter bank responses



K-means

Clusters in filter space



- Each cluster is a visual word.
- Discrete set of words.

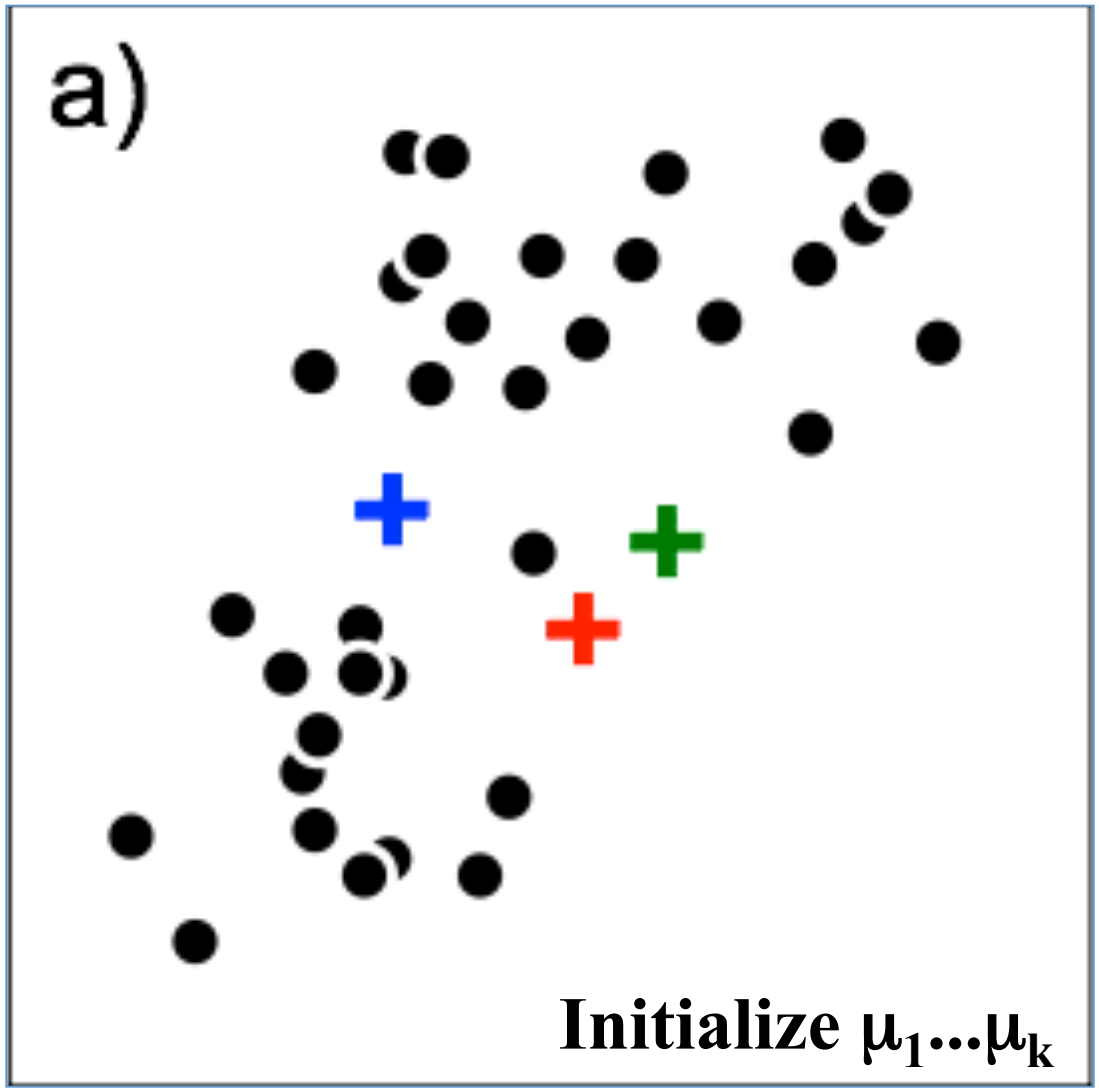
Outline

- Will derive on the board
- Overview:
  - Spherical Gaussians
  - Likelihood ratio classifier
  - K-means objective function
  - Interleaved solution for centers and point labels

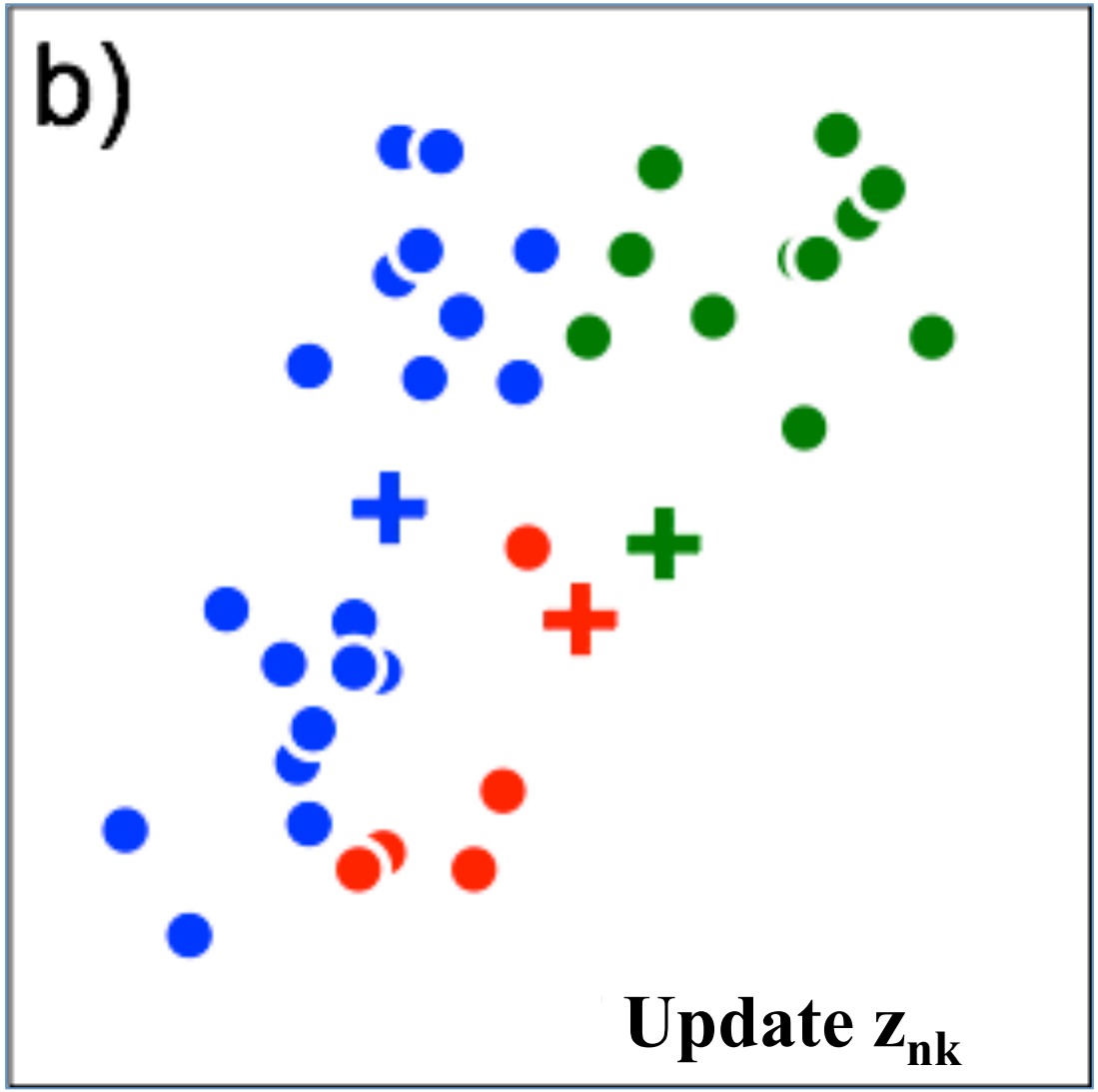
K-Means Algorithm

- Given N data points $x_1, x_2, \ldots, x_N$
- Find K cluster centers $\mu_1, \mu_2, \ldots, \mu_K$ to minimize $\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|x_n - \mu_j\|^2$
  (z_{nk} is 1 if point *n* belongs to cluster *k*; 0 otherwise)
- Algorithm:
  - initialize K cluster centers $\mu_1, \mu_2, \ldots, \mu_K$
  - repeat
    - set z_{nk} labels to assign each point to closest cluster center
    - revise each cluster center μ_j to be center of mass of points in that cluster $\mu_j = \frac{\sum_{n=1}^{N} z_{nj} x_n}{\sum_{n=1}^{N} z_{nj}}$
  - until convergence (e.g. z_{nk} labels don't change)

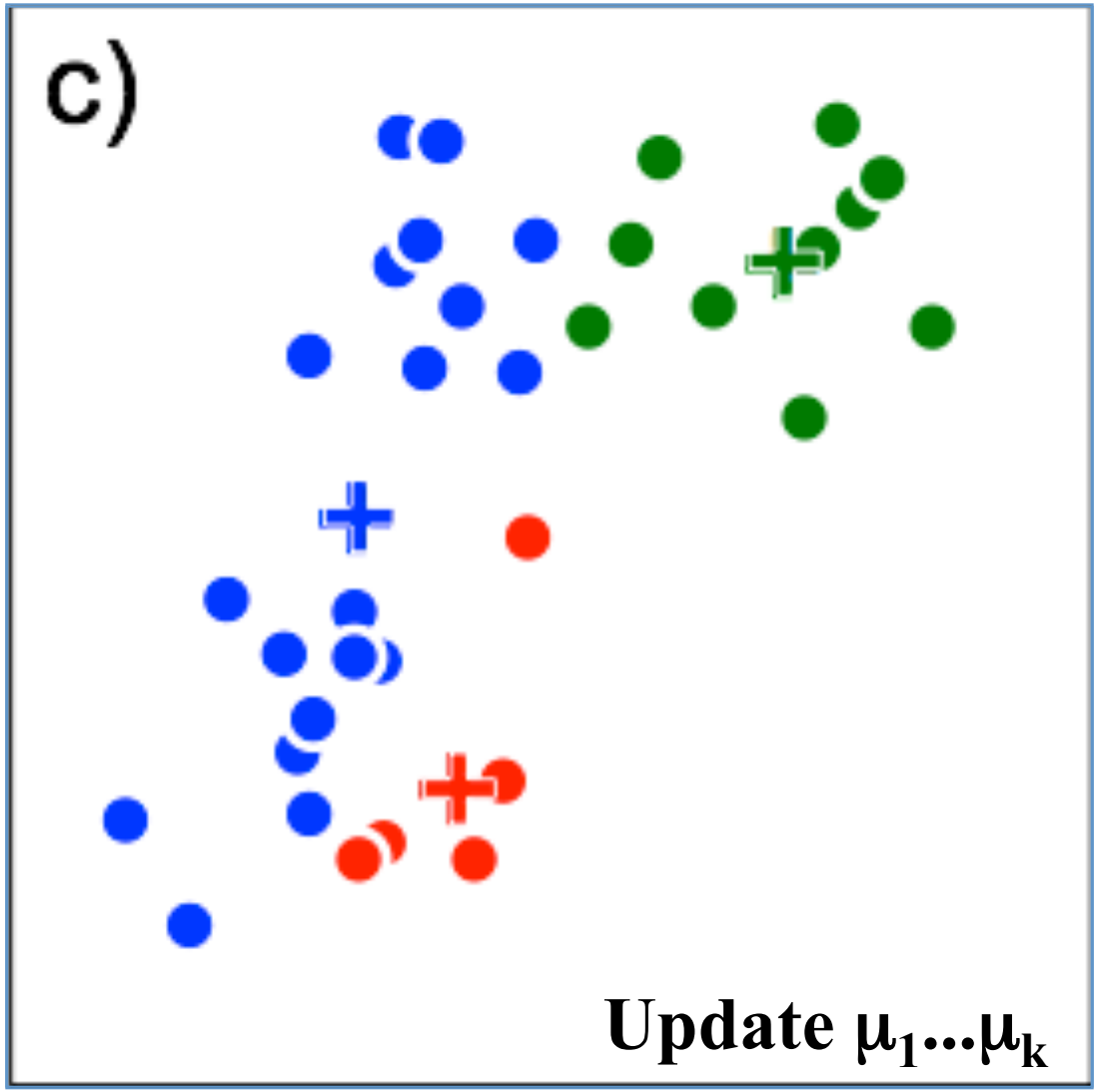
Example



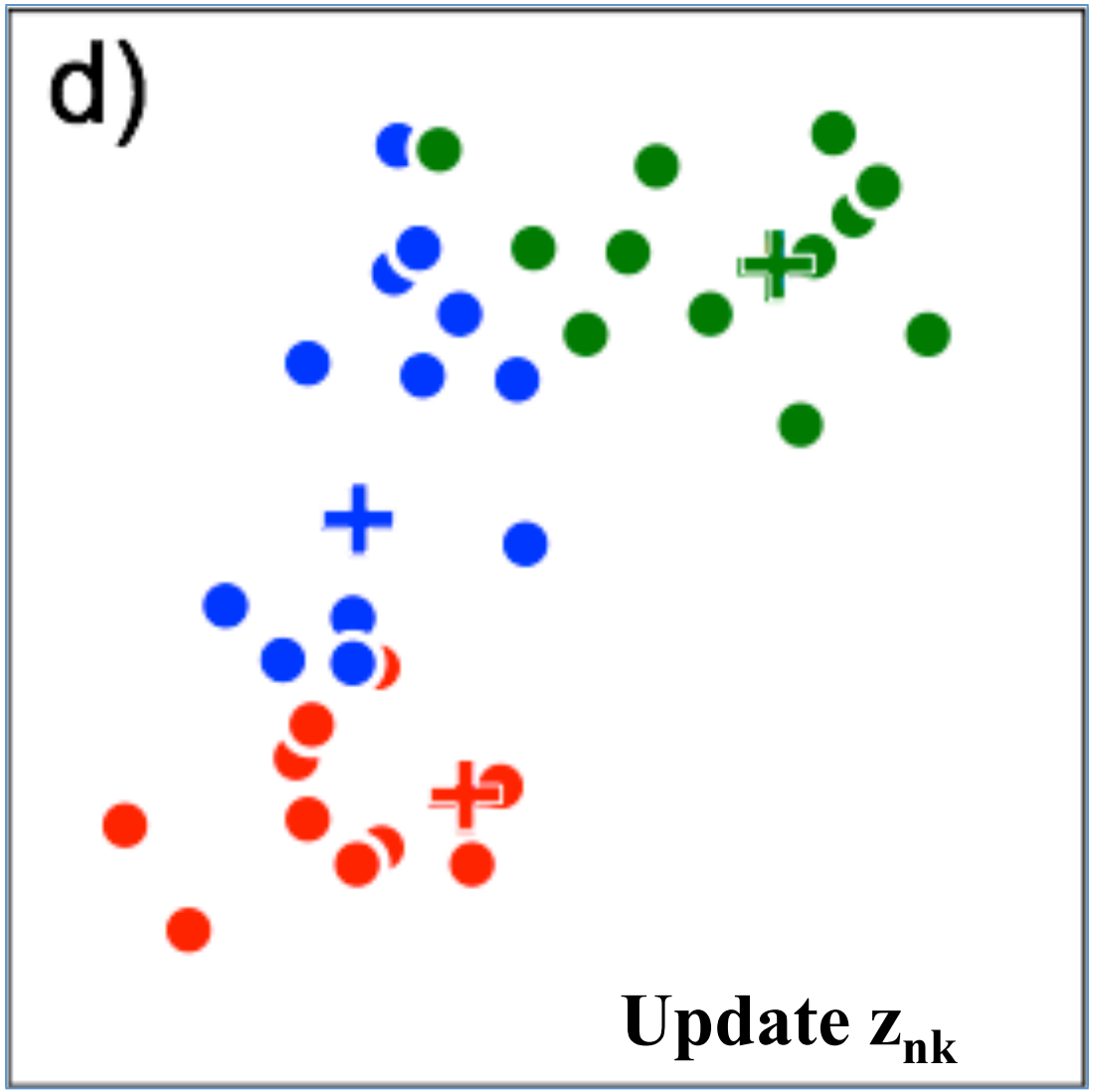
Example



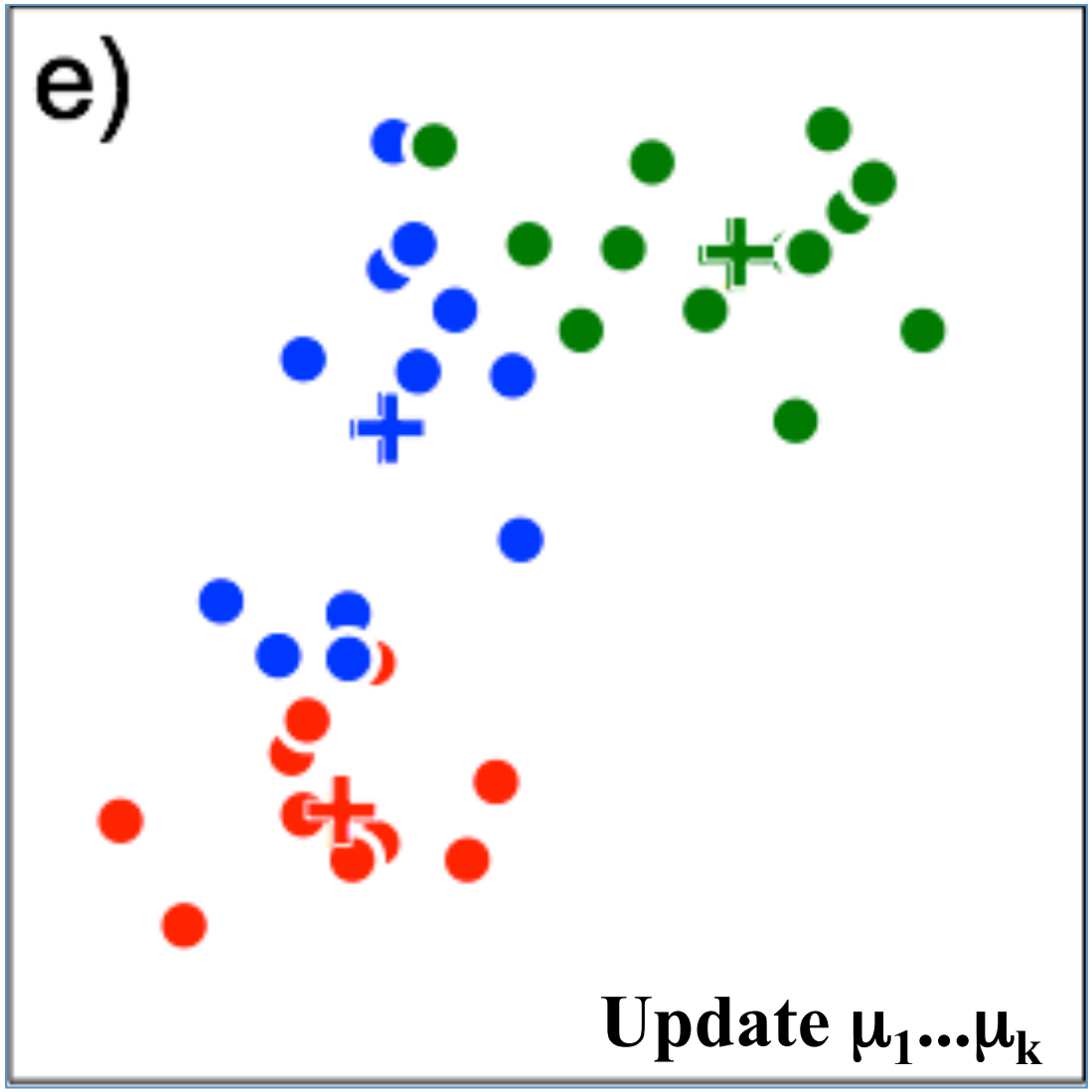
Example



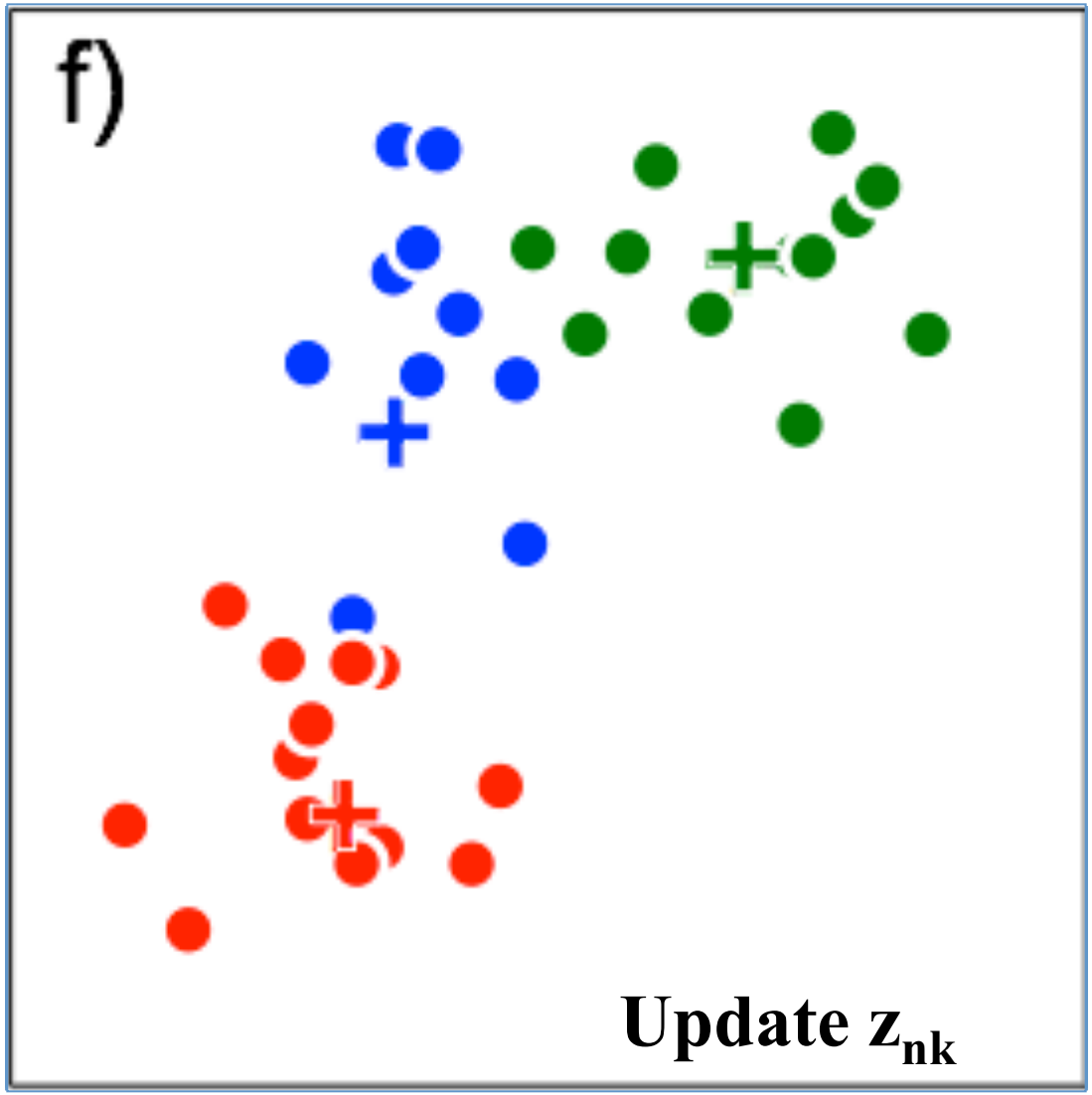
Example



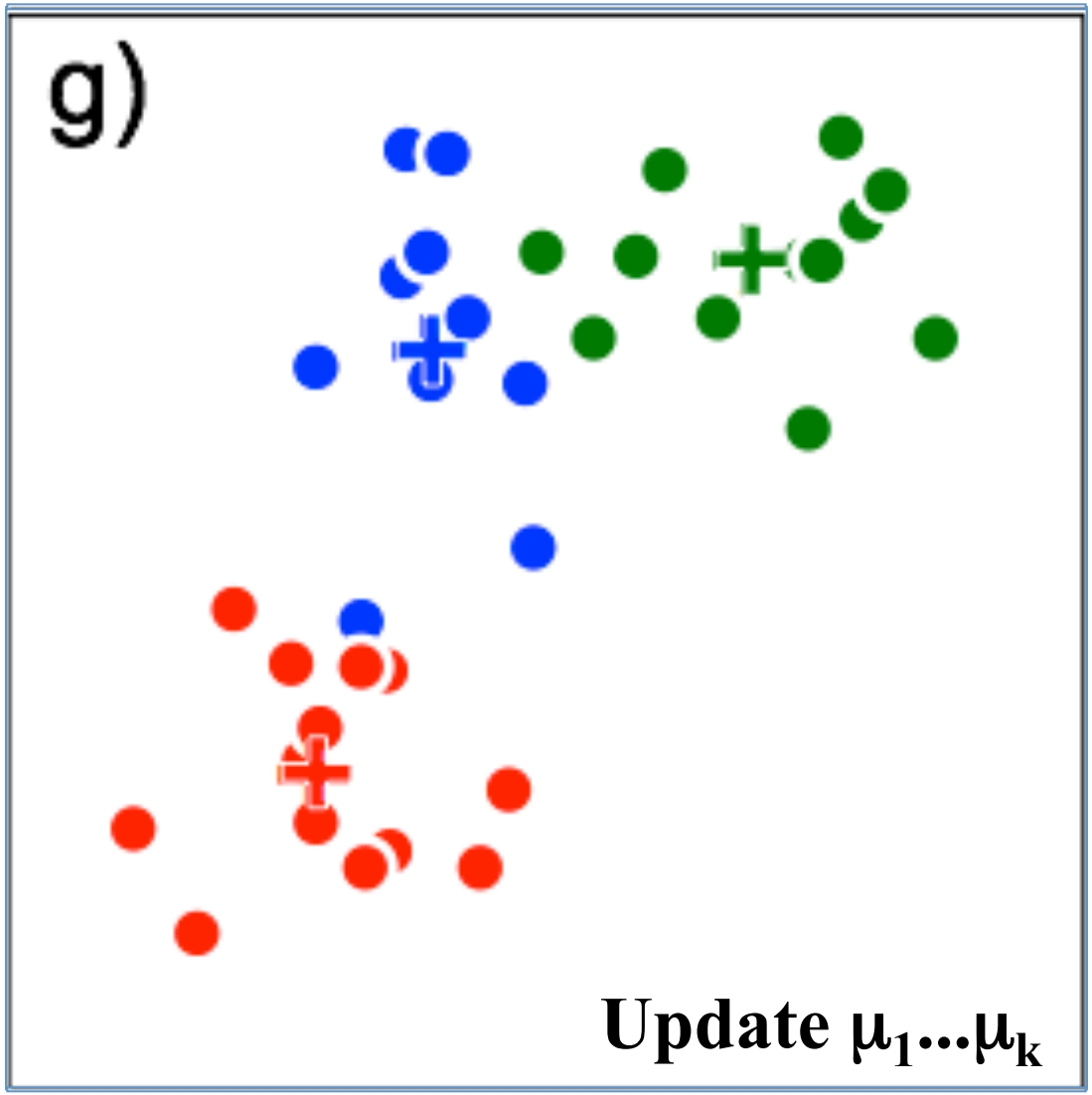
Example



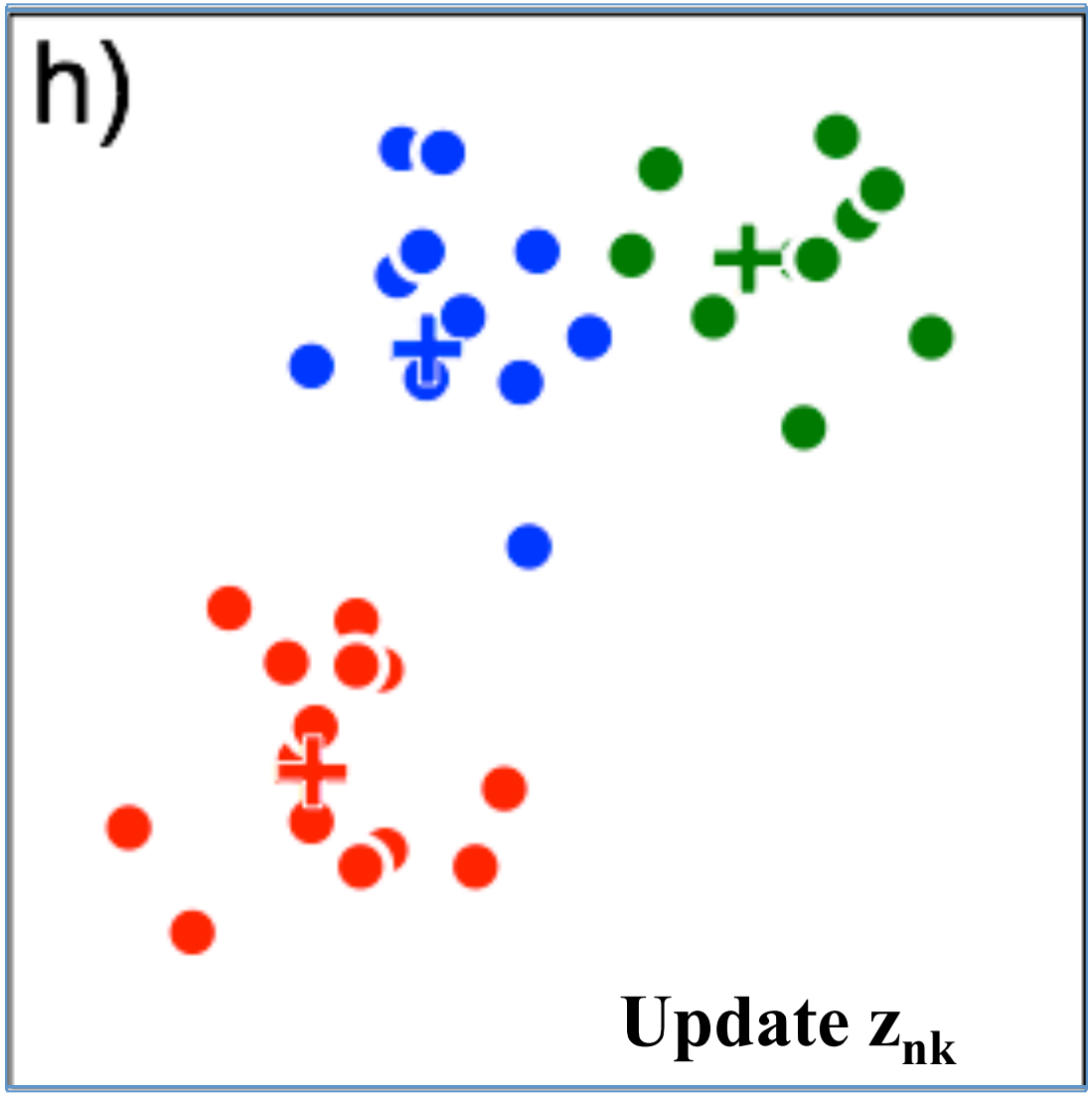
Example



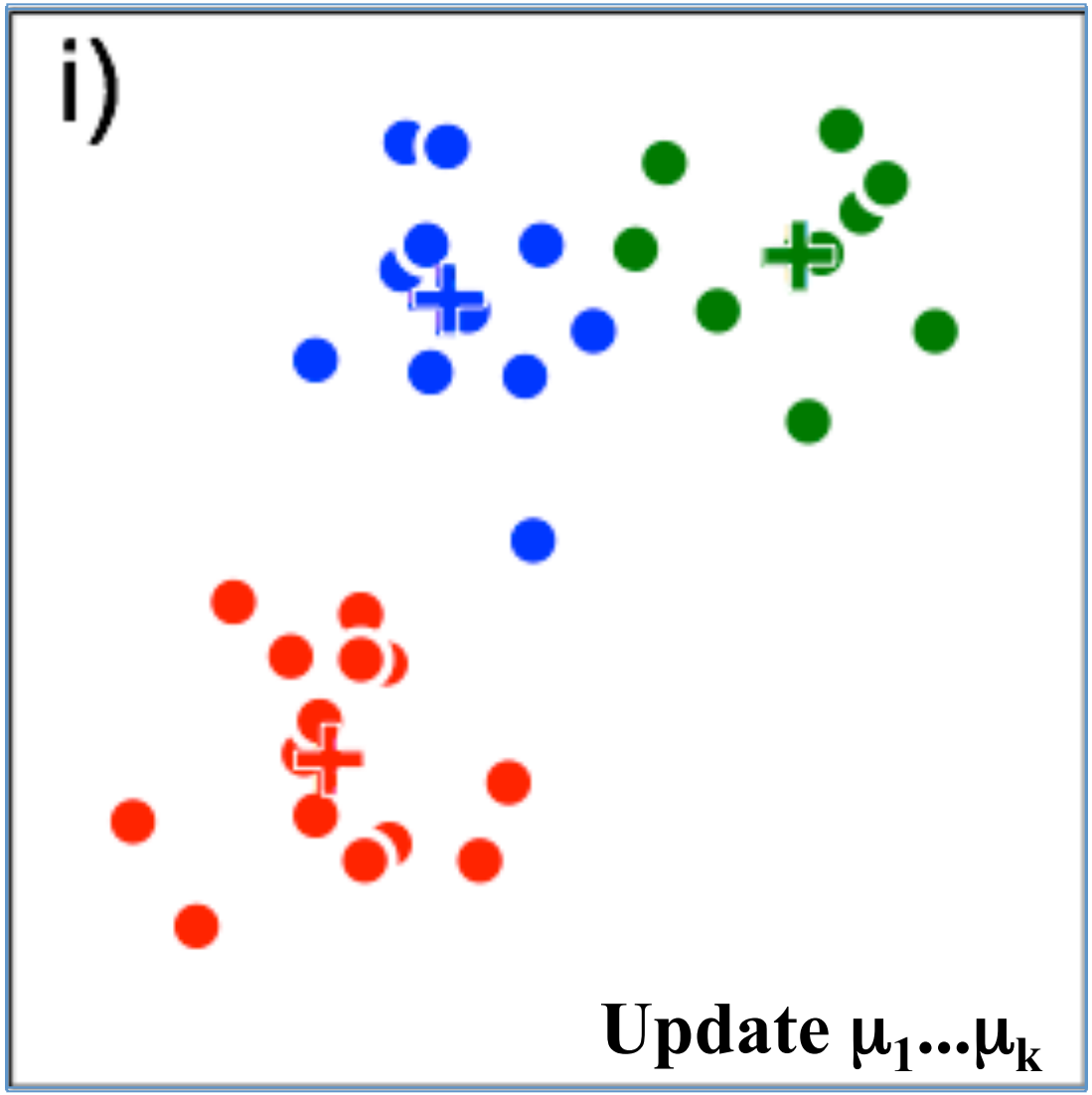
Example



Example



Example



K-means Pros and Cons

Pros:

Fast and easy to code

Guaranteed to converge

Cons:

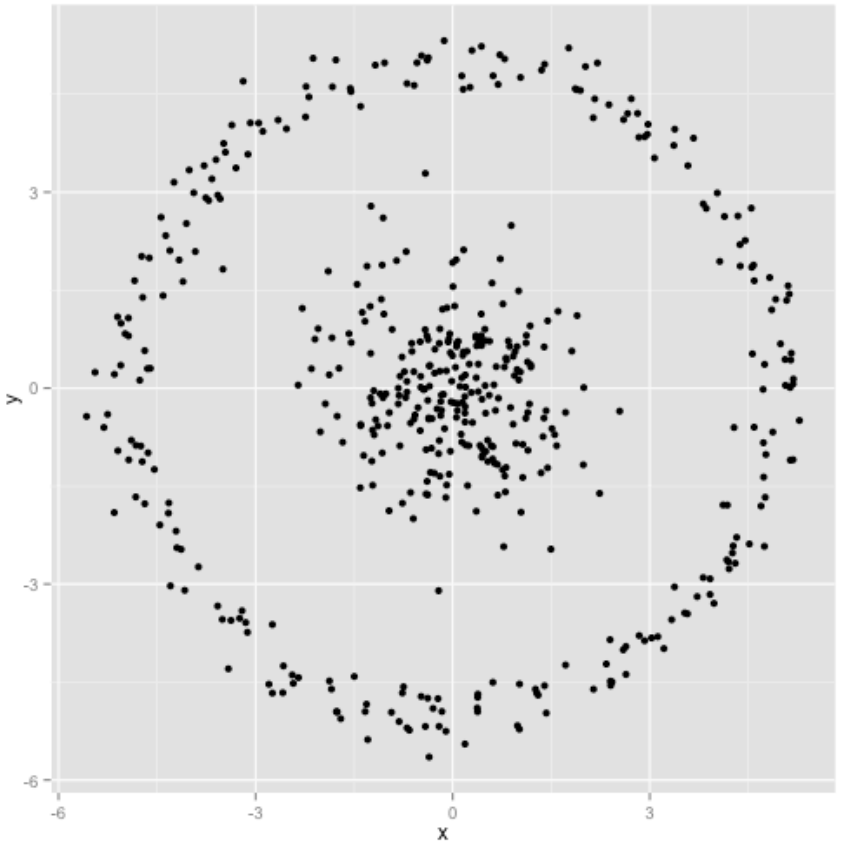
Converges to local minimum

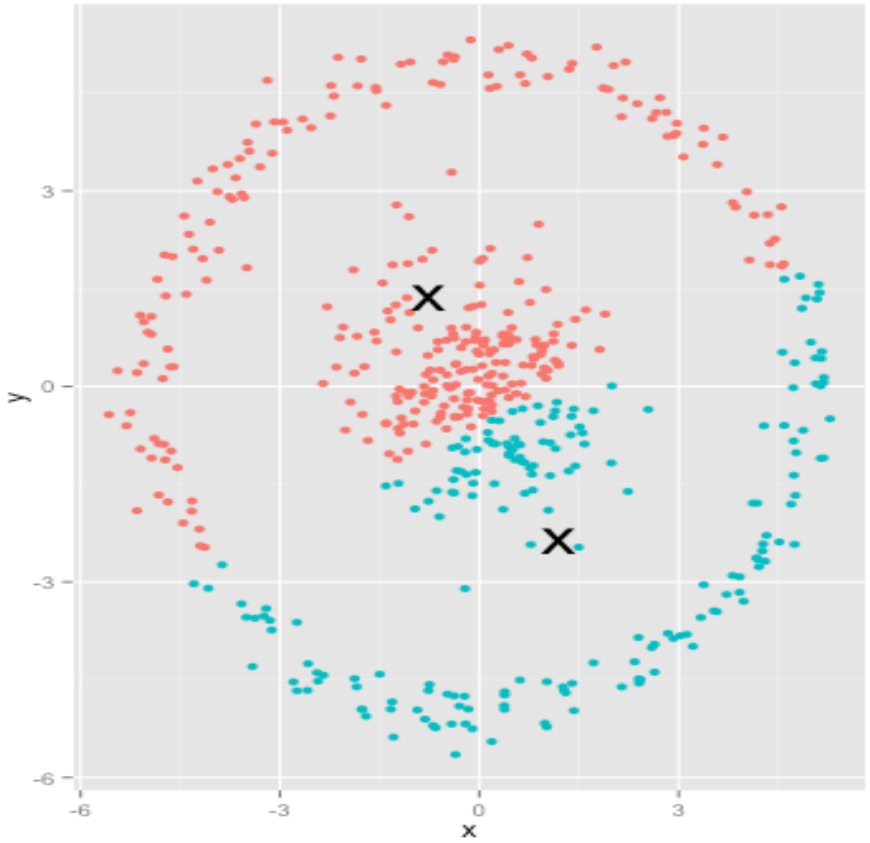
Need to know number of clusters K

Assumptions introduce limitations

spherical clusters, same variance, same # of points
see http://varianceexplained.org/r/kmeans-free-lunch/

Example of Limitations

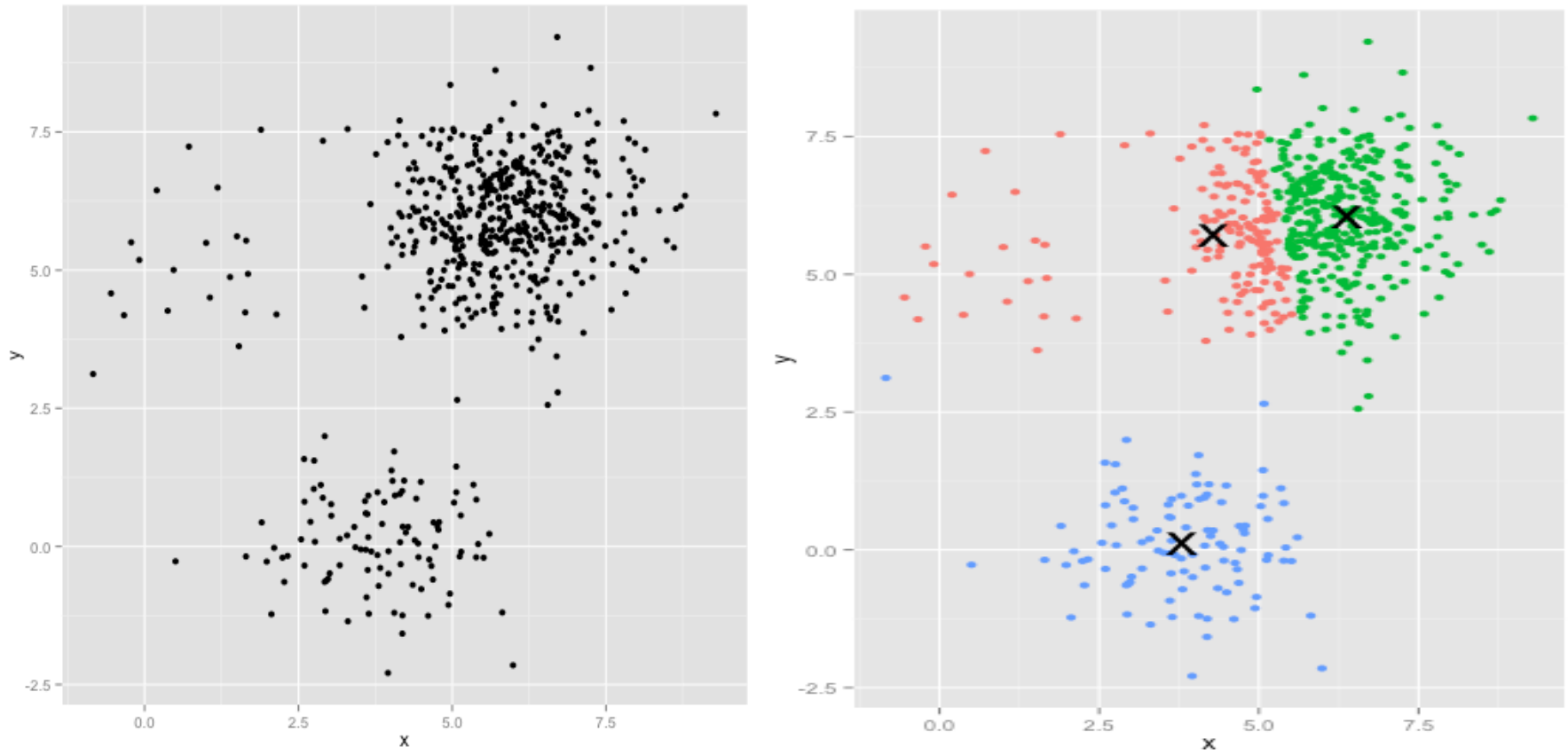




http://varianceexplained.org/r/kmeans-free-lunch/

Robert Collins
CSE586

Example of Limitations



http://varianceexplained.org/r/kmeans-free-lunch/

On the other hand

For this application we are not trying to faithfully describe exact number and shape of clusters in the data, we are discretizing the data into words that will become histogram buckets. Therefore we can live with K-means limitations.