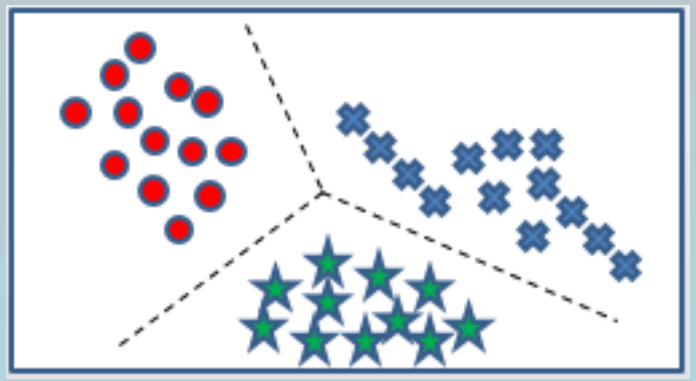
Artificial Intelligence

Chapter 8 Machine Learning (Unsupervised Learning)

Recap (Chap 7)

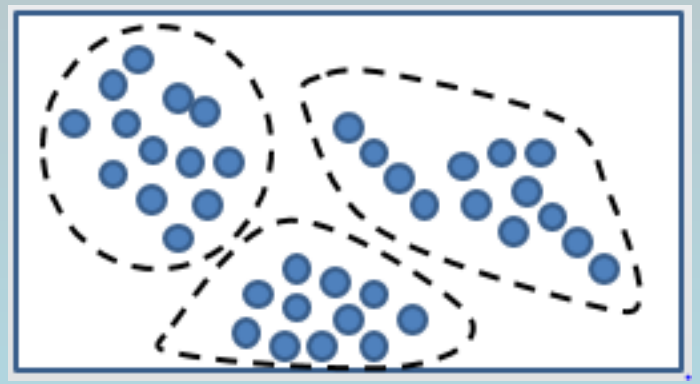
Machine Learning

Supervised



Classification

Unsupervised

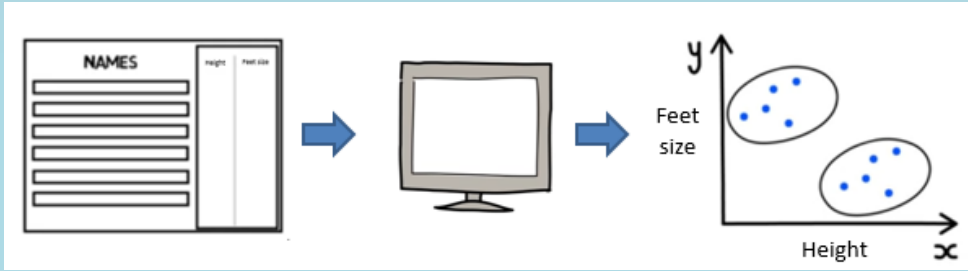


Clustering

X (Features) | y (Label)

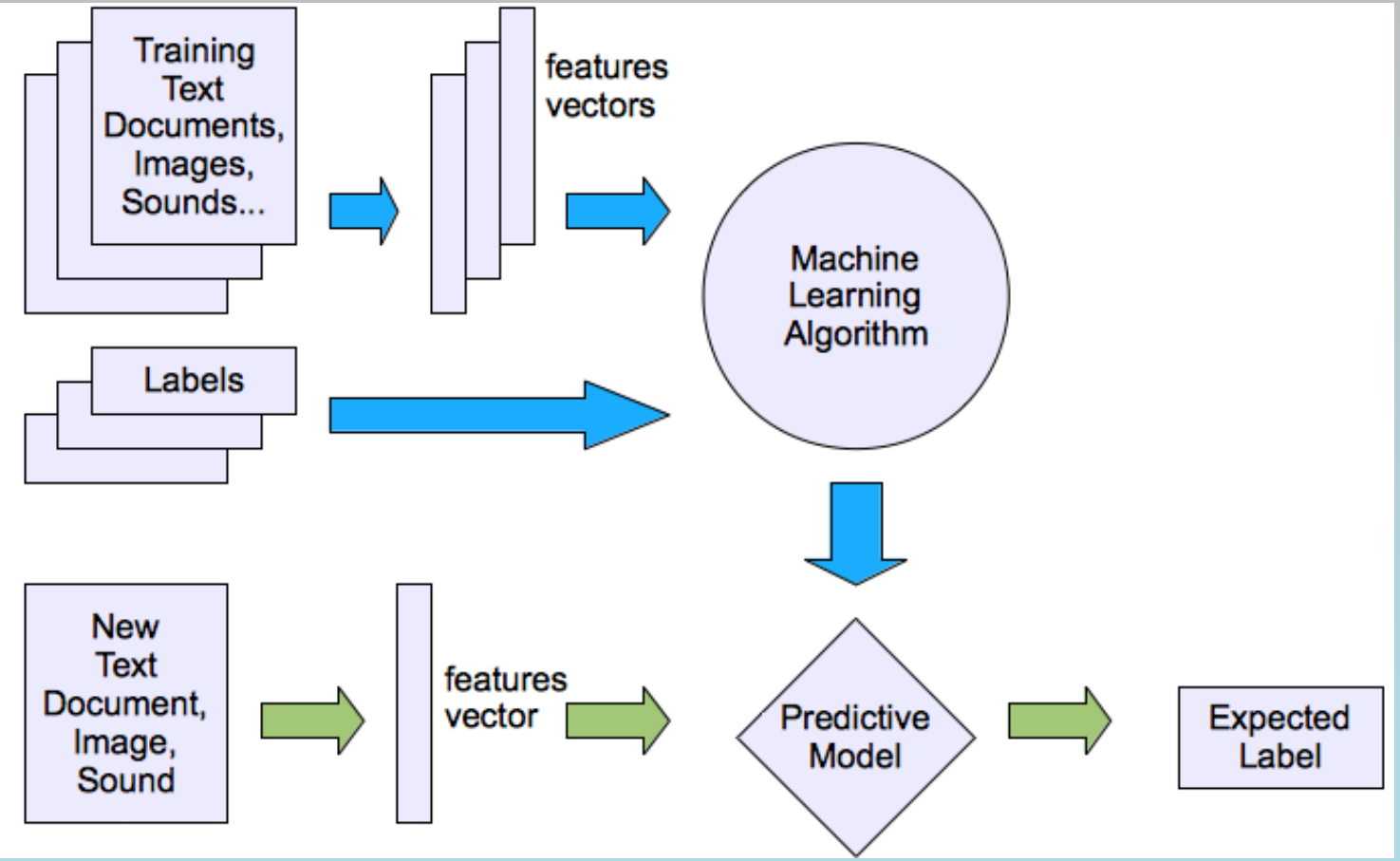
Color	Size	Fruit
Red	Big	Apple
Orange	Big	Orange
Red	Small	Grapes
Red	Big	Apple
Orange	Big	Orange

16



Recap (Chap 7)

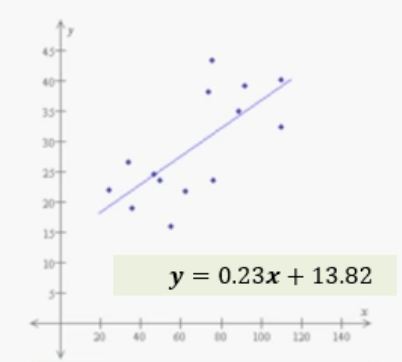
Steps of Supervised Learning



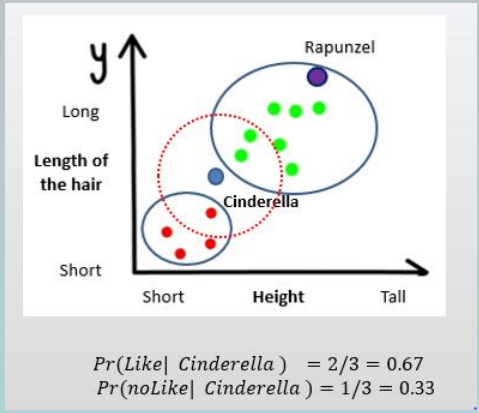
Recap (Chap 7)

Supervised Learning Algorithms/ Models

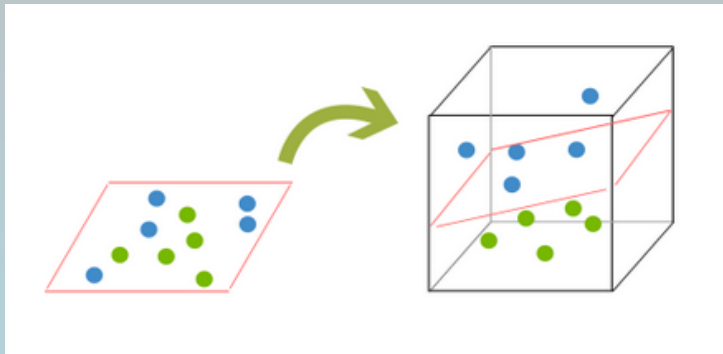
Regression



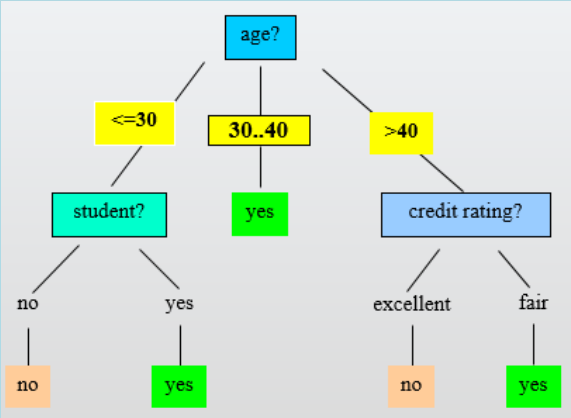
KNN



SVM



Decision Tree



Naïve Bayes

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 2/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

Recap (Chap 7)

Assessing Classifier

		Predicted		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- Accuracy: Overall, how often is the classifier correct?
- (TP+TN)/total = (100+50)/165 = 0.91

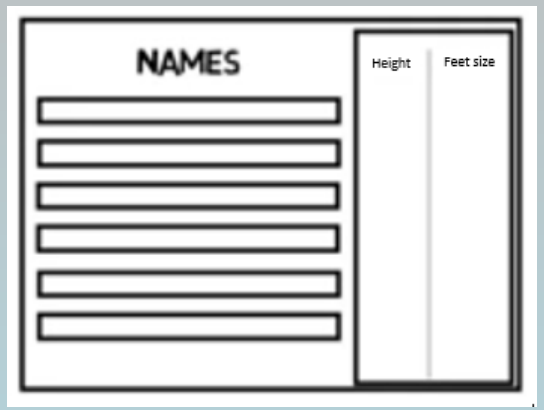
- Precision: When it predicts yes, how often is it correct?
- TP/predicted yes = 100/110 = 0.91

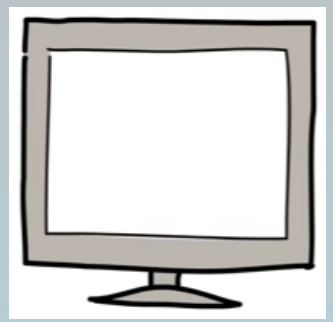
- Recall: When it's actually yes, how often does it predict yes?
- TP/actual yes = 100/105 = 0.95

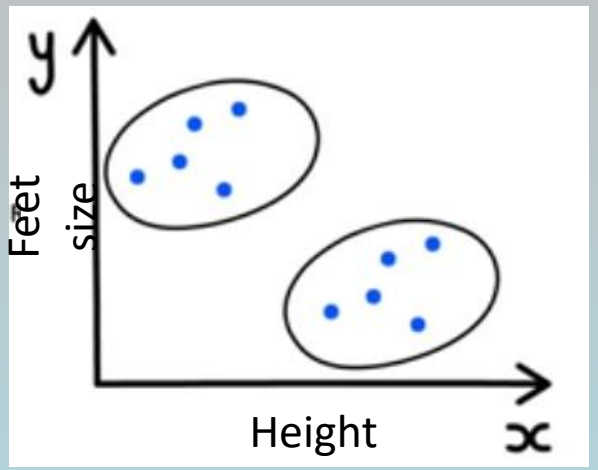
Machine Learning

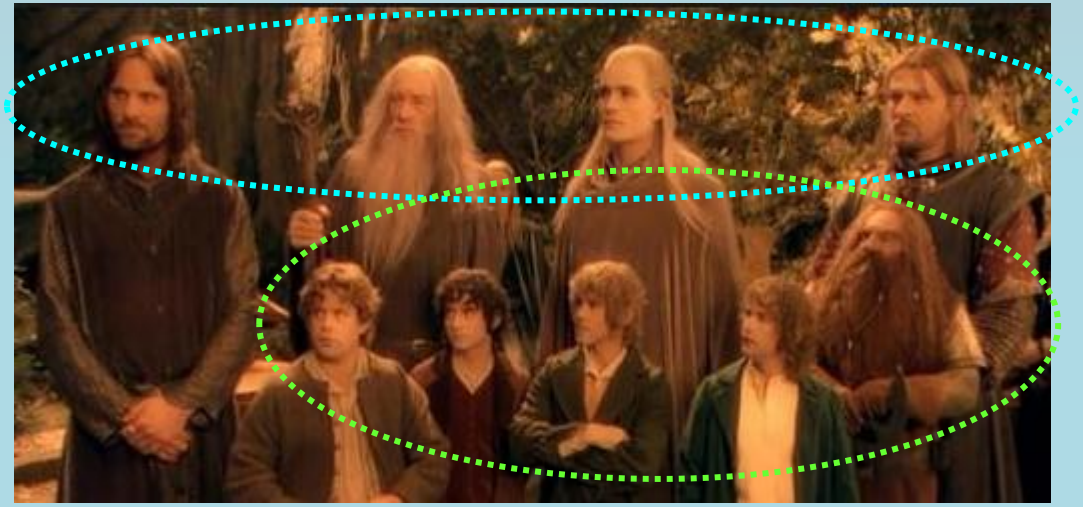
UNSUPERVISED LEARNING (CLUSTERING)

General of Unsupervised Learning









NO Labelled Data

Categorized into two groups.

Real-life Application of Clustering

Search Engine

Google Scholar

Clustering techniques

Articles Case law

[PDF] A comparison of document **clustering techniques**

M Steinbach, G Karypis, V Kumar - KDD workshop on text mining, 2000 - matlabi.ir

This paper presents the results of an experimental study of some common document **clustering techniques**: agglomerative hierarchical **clustering** and K-means.(We used both a "standard" K-means algorithm and a "bisecting" K-means algorithm.) Our results indicate that ...

☆ 99 Cited by 3114 Related articles All 18 versions »

A comparison of neural network and fuzzy **clustering techniques** in segmenting magnetic resonance images of the brain

LO Hall, AM Bensaid, LP Clarke... - IEEE transactions on ..., 1992 - ieeexplore.ieee.org

Magnetic resonance (MR) brain section images are segmented and then synthetically colored to give visual representations of the original data with three approaches: the literal and approximate fuzzy c-means unsupervised **clustering** algorithms, and a supervised ...

☆ 99 Cited by 723 Related articles All 10 versions

MRI segmentation using fuzzy **clustering techniques**

MC Clark, LO Hall, DB Goldgof... - IEEE Engineering in ..., 1994 - ieeexplore.ieee.org

The authors' main contribution is to build upon their earlier efforts by expanding the tissue model concept to cover a brain volume. Furthermore, processing time is reduced and accuracy is enhanced by the use of knowledge propagation, where information derived from ...

☆ 99 Cited by 271 Related articles All 4 versions

Spectral–spatial classification of hyperspectral imagery based on partitional **clustering techniques**

re.ieee.org
posed. The
cation and the

Google Spiderman

All Images Videos News Maps More Settings Tools SafeSear

tom holland wallpaper cartoon homecoming deadpool civil war drawing bla

Real-life Application of Clustering

Business and Marketing

Grouping of customers.

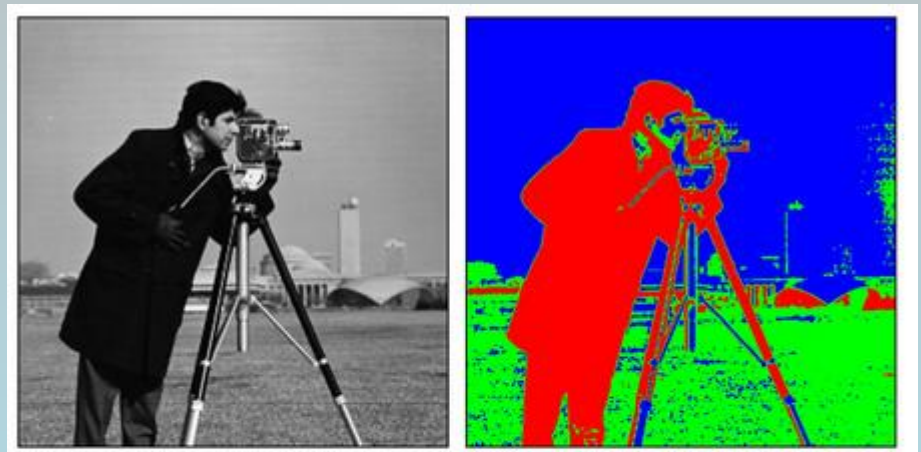
76, 753 prepaid subscribers

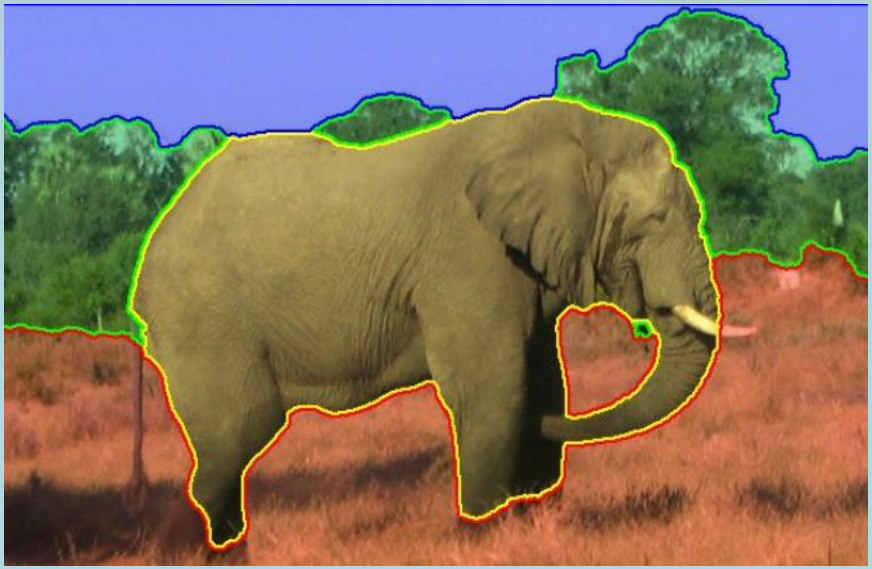
Prepaid Telekom Customers segmentation using K-means

Total= 76,753	Calls (%)	Data (%)	SMS (%)
Cluster 1	38	29	7
Cluster 2	79	20	1
Cluster 3	46	24	30
Cluster 4	31	42	27
Cluster 5	7	76	17
Cluster 6	44	41	15
Cluster 7	3	94	3

Real-life Application of Clustering

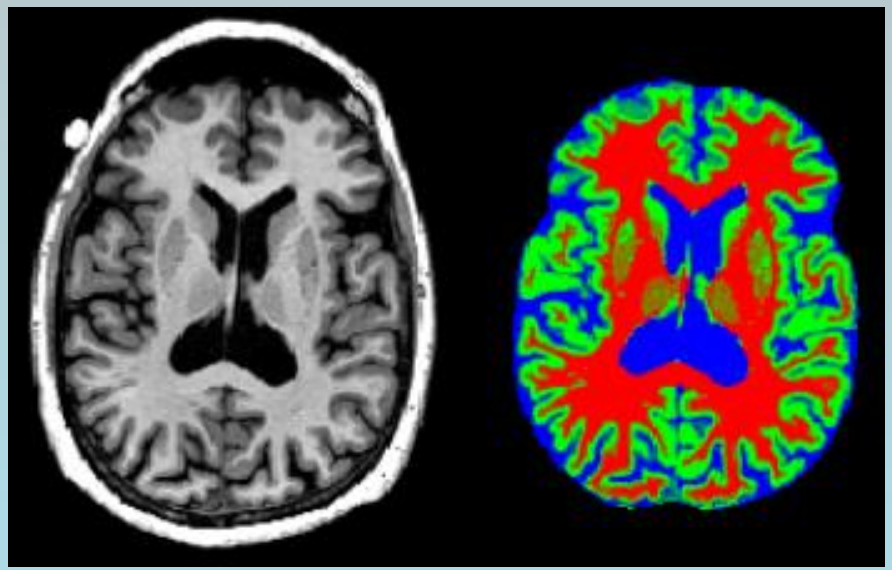
Image segmentation

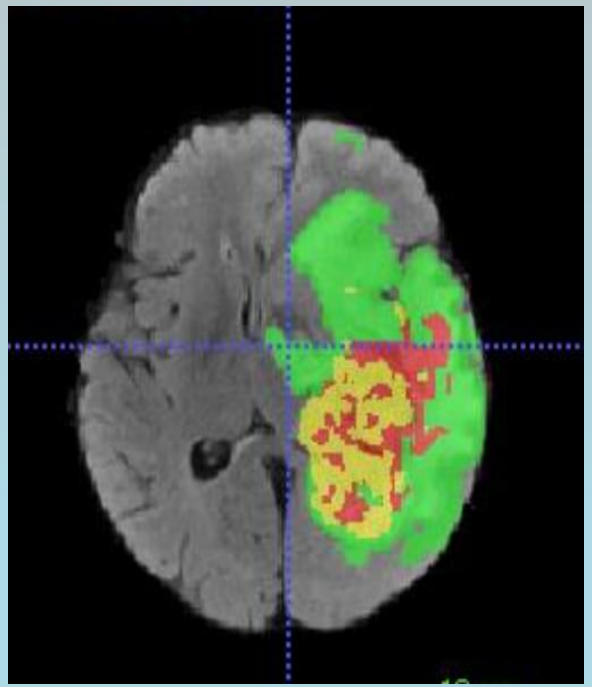




Real-life Application of Clustering

Medical Imaging





Clustering Methods

- Clustering Methods
  - K- Means
  - Gaussian Mixture Model
  - Mean-Shift
  - Hierarchical Clustering

1. K-means

- K-means partitioning (or clustering) N data points into K disjoint cluster centroid, c_j containing data points so as to minimize the sum-of-squares criterion

$$J = \sum_{j=1}^{K} \sum_{x \in c_j} \|x - c_j\|^2$$

- Let the set of data points (or instances) $\mathcal{D}$ be $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$,
  where $\boldsymbol{x}_n = (A_1, A_2, \ldots, A_r)$ is a vector in a real-valued space $A \subseteq R^r$, and r is the number of attributes (dimensions) in the data.

- K is specified by the user

1. K-means

Algorithm:

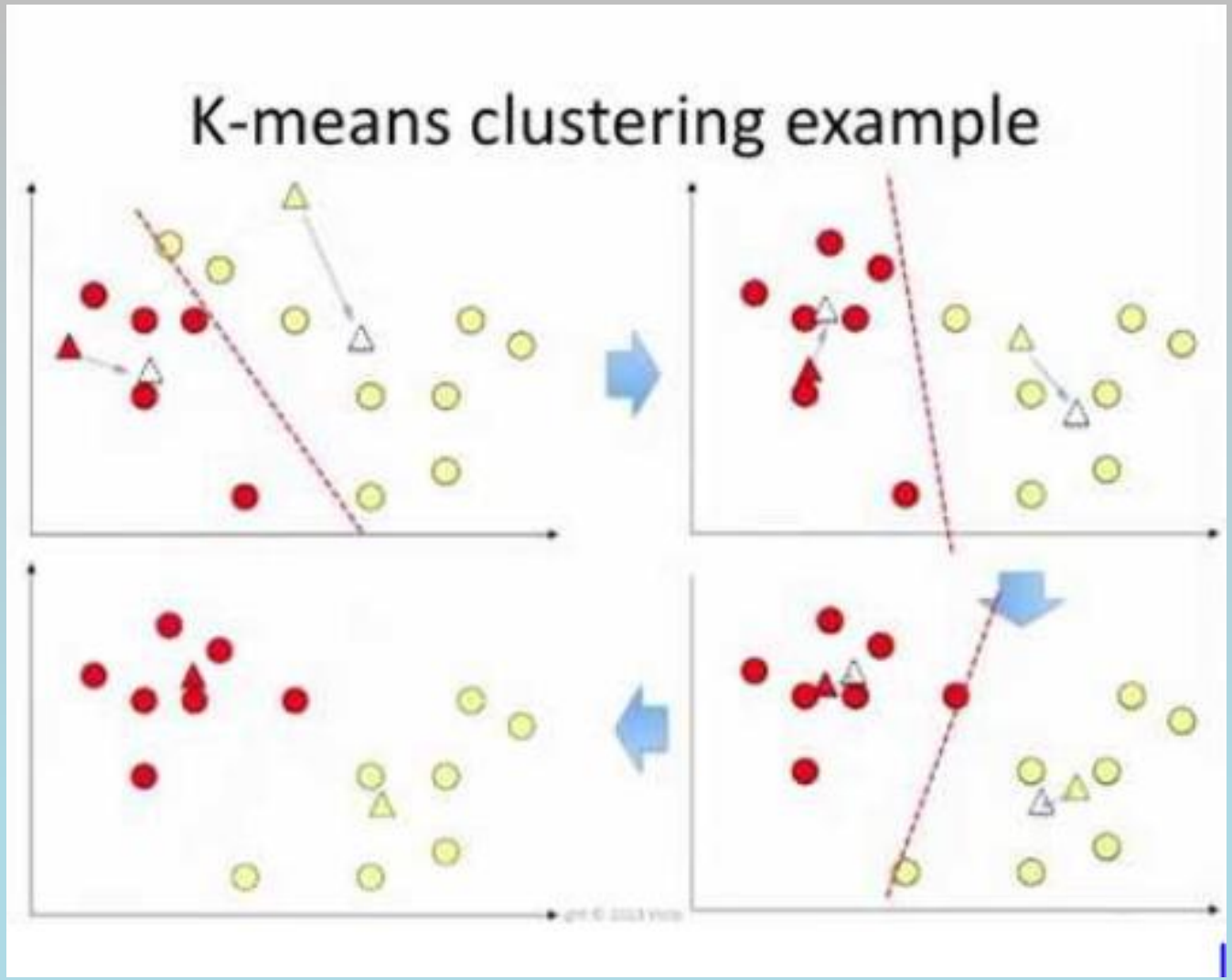
- Given *k*, the *k-means* algorithm works as follows:

  1) Randomly choose *k* data points (seeds) to be the initial centroids, cluster centers

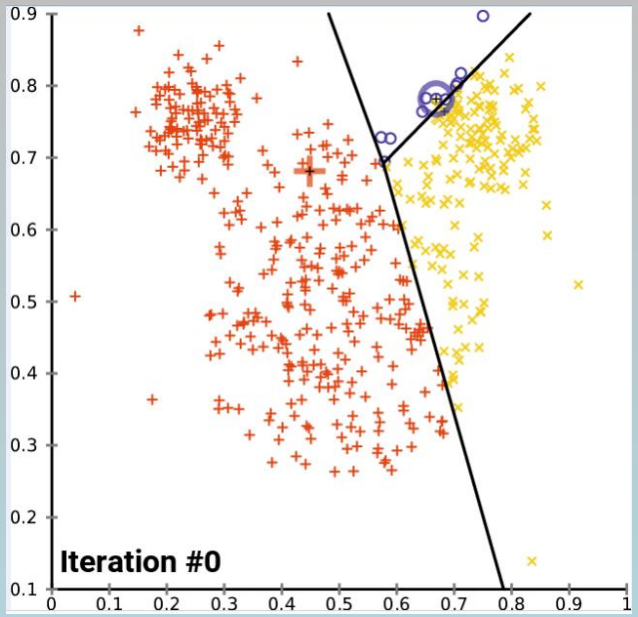
  2) Assign each data point to the closest centroid

  3) Re-compute the centroids using the current cluster memberships.

  4) If a convergence criterion is not met, go to 2.


K-means clustering example

1. K-means

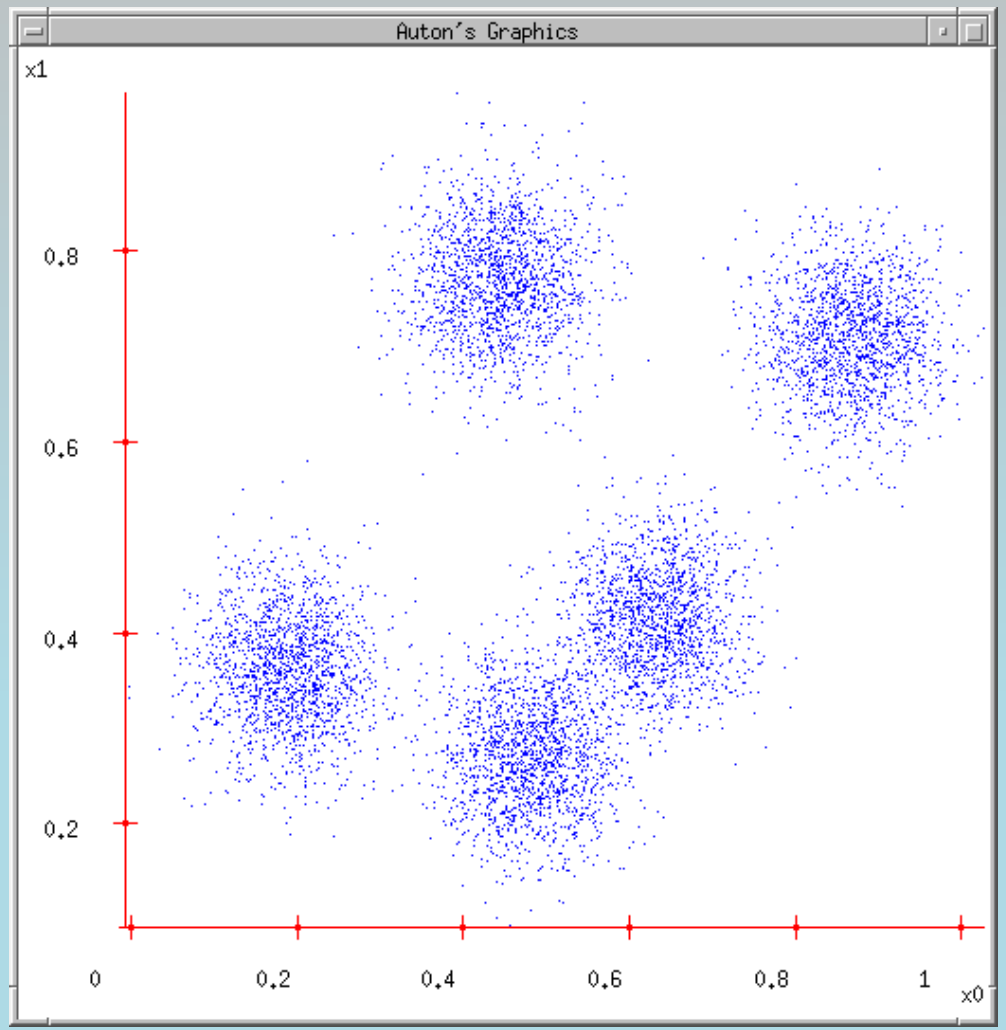


Randomly select 3 centroids.

Iteration #0

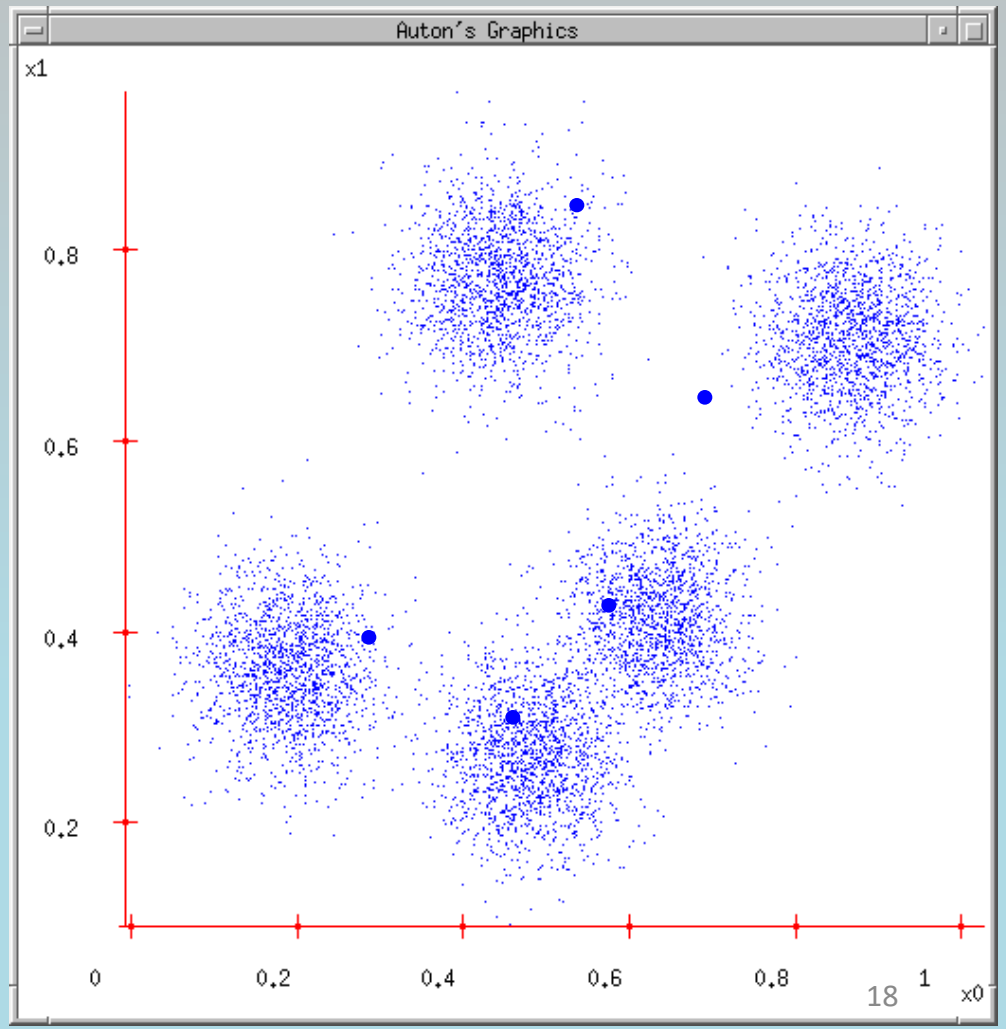
1. K-means

1. Number of clusters (e.g. k=5)



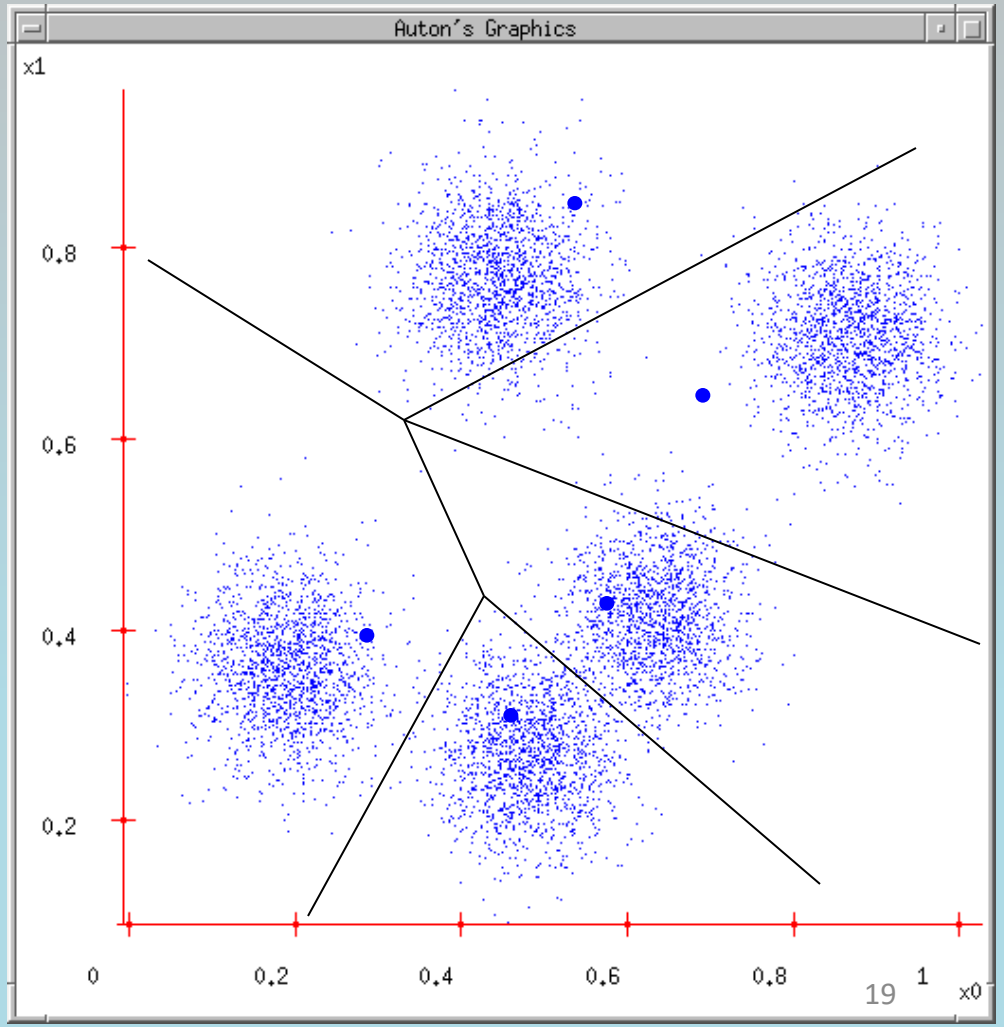
1. K-means

1. Number of clusters (e.g. k=5)
2. Randomly guess k cluster Center locations



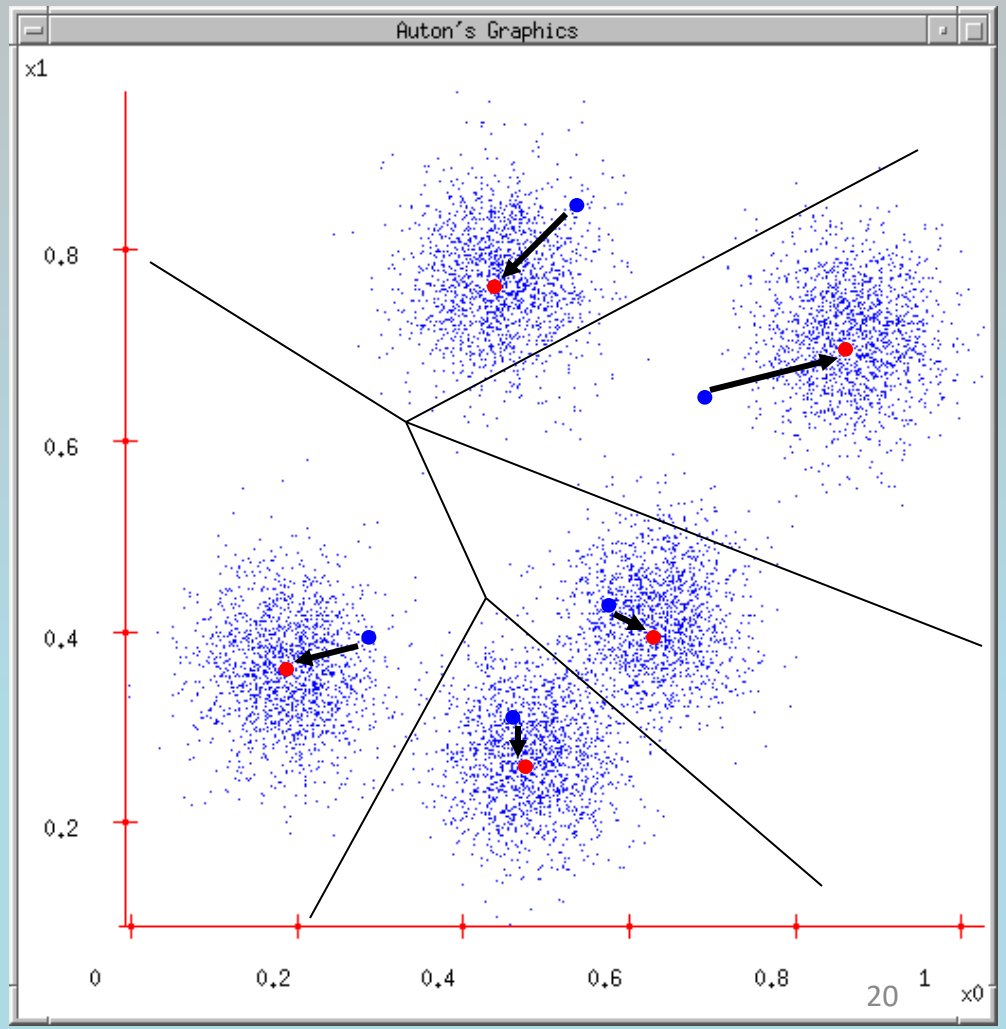
1. K-means

1. Number of clusters (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Repeat:
   I. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



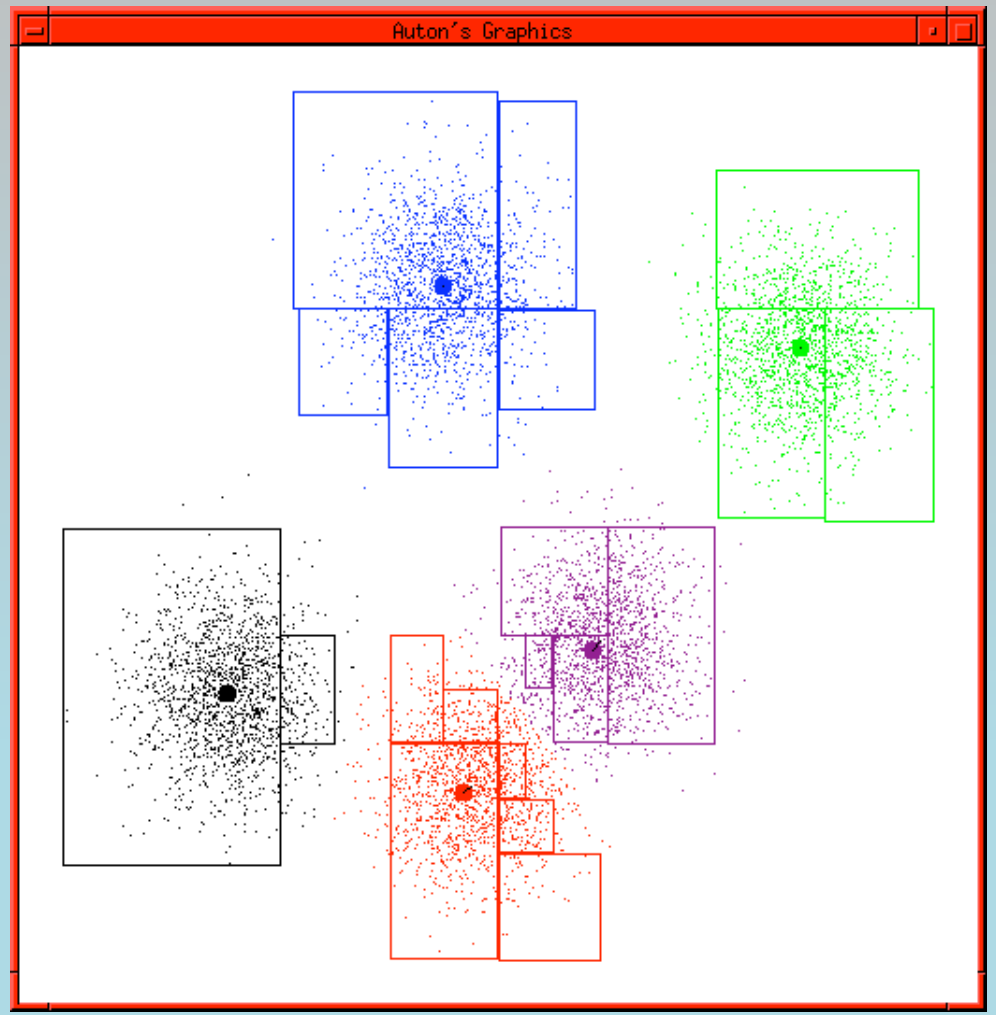
1. K-means

1. Number of clusters (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Repeat:
   I. Each datapoint finds out which Center it's closest to.
   II. Each Center finds the centroid of the points it owns
4. Stop when the centroids don't change.



1. K-means

- K-Means Animation



Example generated by Andrew Moore using Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999.

- Example

data	Height	Weight
	185	72
	167	56
	168	60
	179	68
	182	72
	188	77
	180	71
	180	70
	183	84
	180	88
	169	47
	177	76

$$d(x_1, c_1) = \sqrt{(185 - 186)^2 + (72 - 7} = 1.41$$

$$d(x_1, c_2) = \sqrt{(185 - 165)^2 + (72 - 4} = 32.06$$

1.Assign k=2

2.Randomly guess k cluster Center locations

Centroid,	Height	Weight
c1	**186**	**71**
c2	**165**	**47**

3.Each datapoint finds out which Center it's closest

	Height	Weight	to *c1*	to *c2*	*c1*	*c2*
1	185	72	1.4142	32.0156	☺	
2	167	56	24.2074	9.2195		☺
3	168	60	21.0950	13.3417		☺
4	179	68	7.6158	25.2389	☺	
5	182	72	4.1231	30.2324	☺	
6	188	77	6.3246	37.8021	☺	
7	180	71	6.0000	28.3019	☺	
8	180	70	6.0828	27.4591	☺	
9	183	84	13.3417	41.1461	☺	
10	180	88	18.0278	43.6578	☺	
11	169	47	29.4109	4.0000		☺
12	177	76	10.2956	31.3847	☺	

- Example

data	Height	Weight
	185	72
	167	56
	168	60
	179	68
	182	72
	188	77
	180	71
	180	70
	183	84
	180	88
	169	47
	177	76

3. Each datapoint finds out which Center it's closest

Height	Weight	to *c1*	to *c2*	*c1*	*c2*
185	72	1.4142	32.0156	☺	
167	56	24.2074	9.2195		☺
168	60	21.0950	13.3417		☺
179	68	7.6158	25.2389	☺	
182	72	4.1231	30.2324	☺	
188	77	6.3246	37.8021	☺	
180	71	6.0000	28.3019	☺	
180	70	6.0828	27.4591	☺	
183	84	13.3417	41.1461	☺	
180	88	18.0278	43.6578	☺	
169	47	29.4109	4.0000		☺
177	76	10.2956	31.3847	☺	

4.Each center finds the new centroid points

$$c_1 = \left(\frac{1}{9}(185 + 179 + \cdots), \quad \frac{1}{9}(72 + 68 + \cdots)\right)$$

$$c_2 = \left(\frac{1}{3}(167 + 168 + \cdots), \quad \frac{1}{3}(56 + 60 + \cdots)\right)$$

New Centroid,	Height	Weight
c1	**181.56**	**75.33**
c2	**168.00**	**54.33**

5.Repeat step 3 & 4 until convergence.

- Second iteration

Height	Weight	to *c1*	to *c2*	*c1*	*c2*
185	72				
167	56				
168	60				
179	68				
182	72				
188	77				
180	71				
180	70				
183	84				
180	88				
169	47				
177	76				

New Centroid,	Height	Weight
c1	**181.56**	**75.33**
c2	**168.00**	**54.33**

1. K-means

Stopping/convergence criterion

1. no (or minimum) re-assignments of data points to different clusters,
2. no (or minimum) change of centroids, or
3. minimum decrease in the **sum of squared error** J,

$$J = \sum_{j=1}^{K} \sum_{x \in c_j} \|x - c_j\|^2$$

1. K-means

Problems with K-Means

- ***Very*** sensitive to the initial points.
  - Do many runs of k-Means, each with different initial centroids.

- Must manually choose k.
  Learn the optimal k for the clustering.
  (Note that this requires a performance measure.)

1. K-means

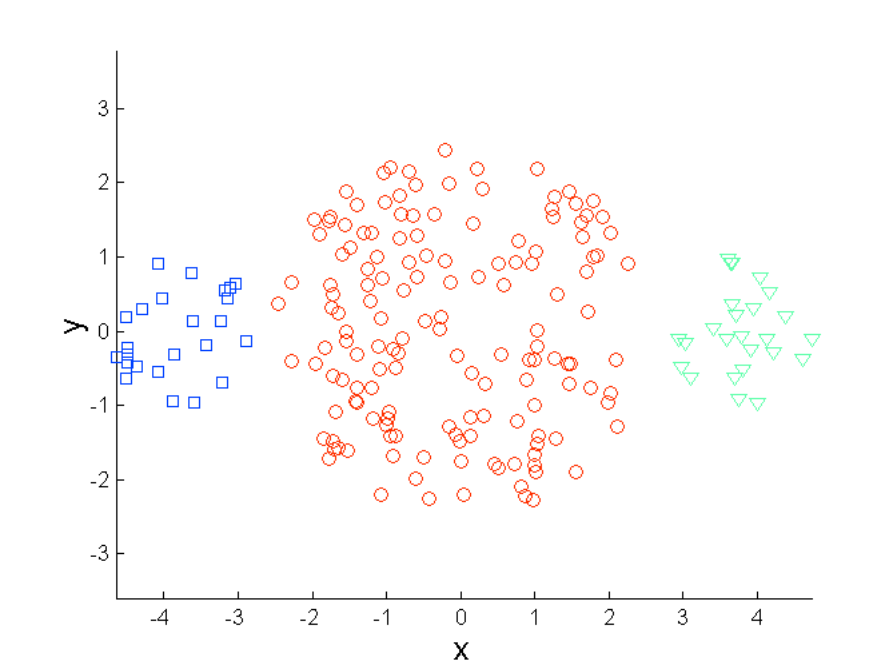
Problems with K-Means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

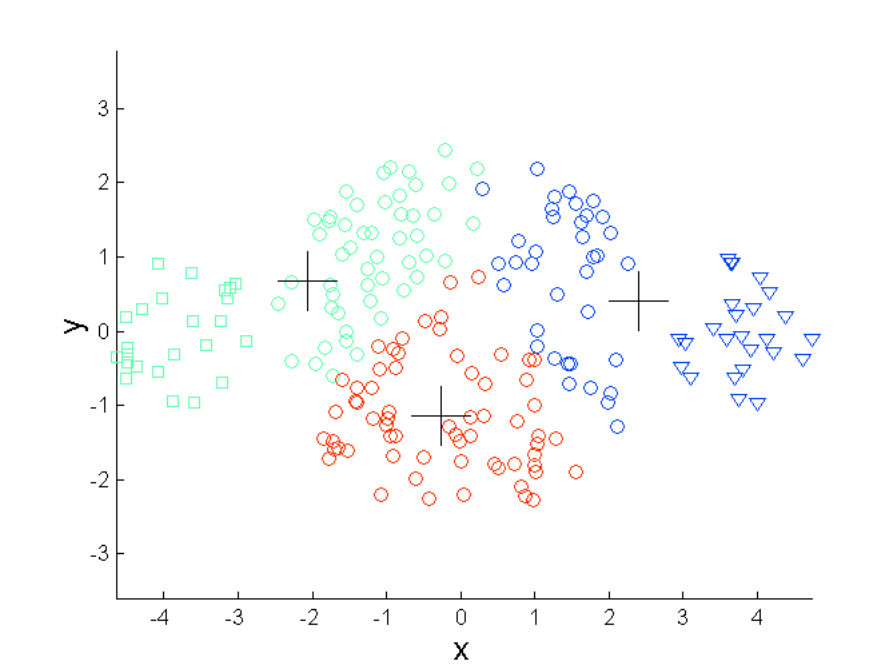
- K-means has problems when the data contains outliers.

1. K-means

Problem with K-means: Differing Sizes



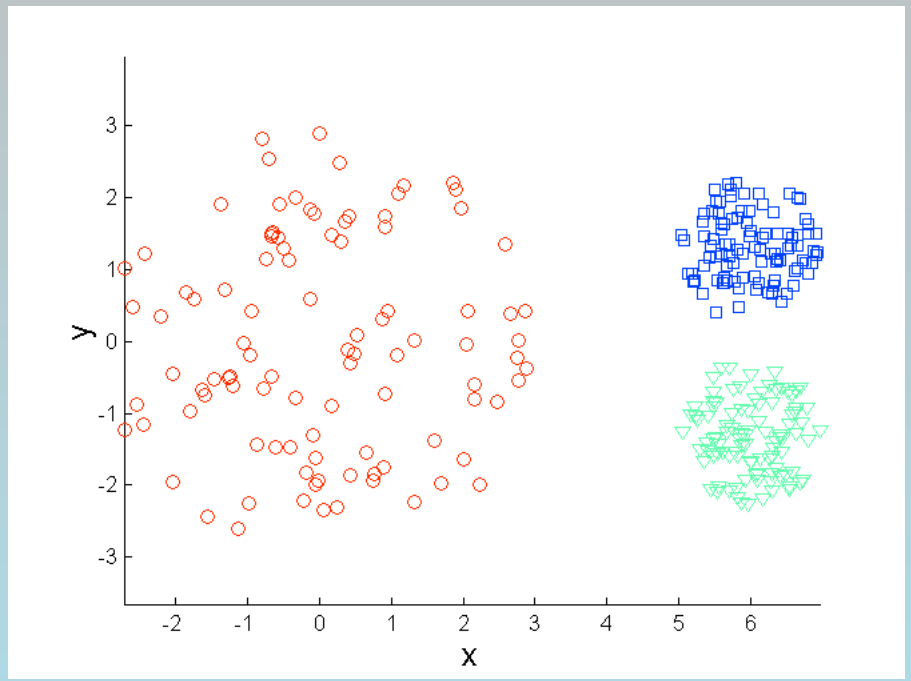
Original Points



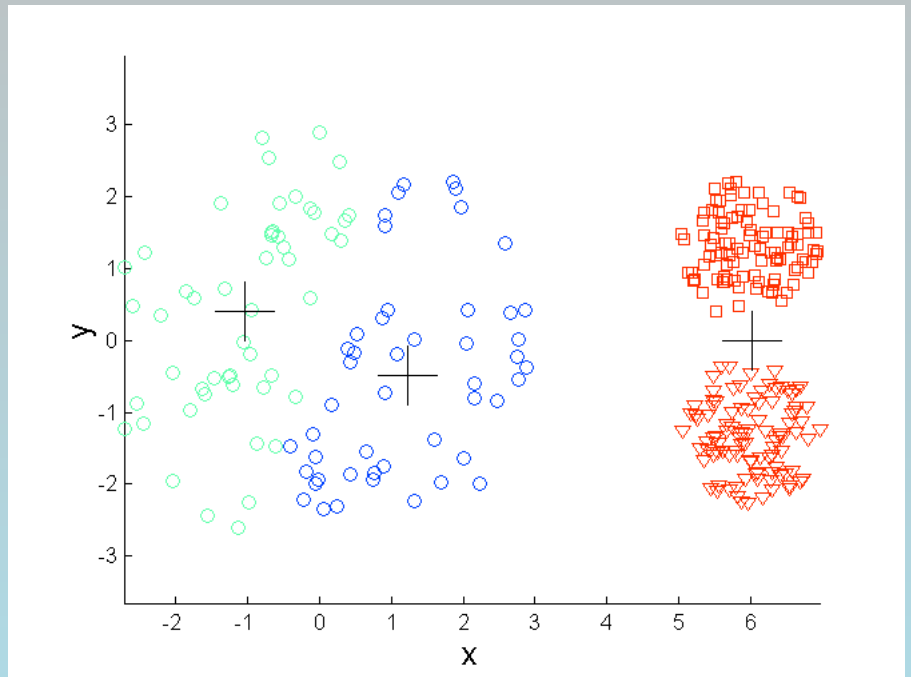
K-means (3 Clusters)

1. K-means

Problem with K-means: Differing Density



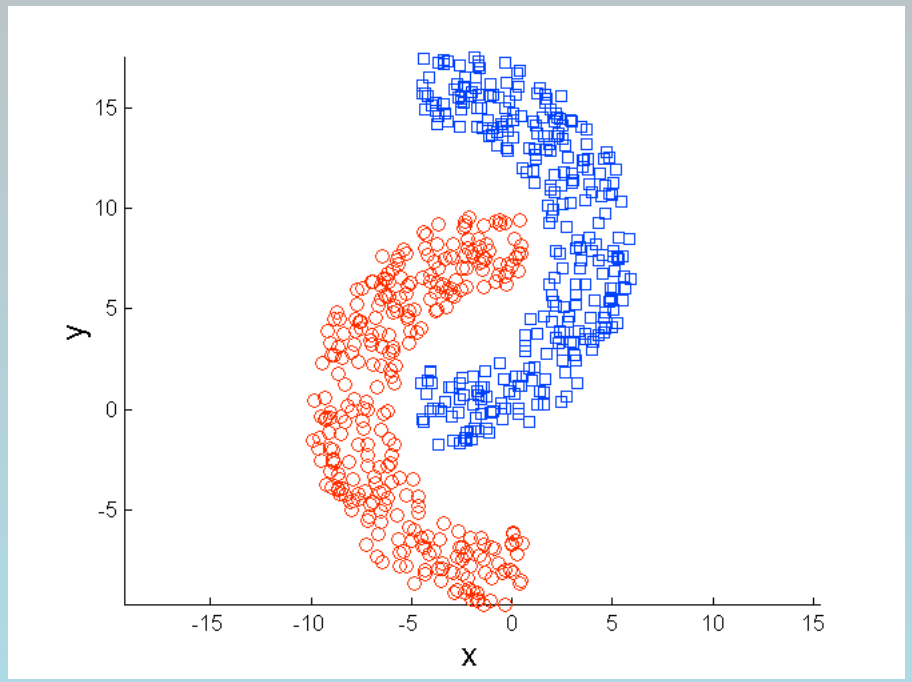
Original Points



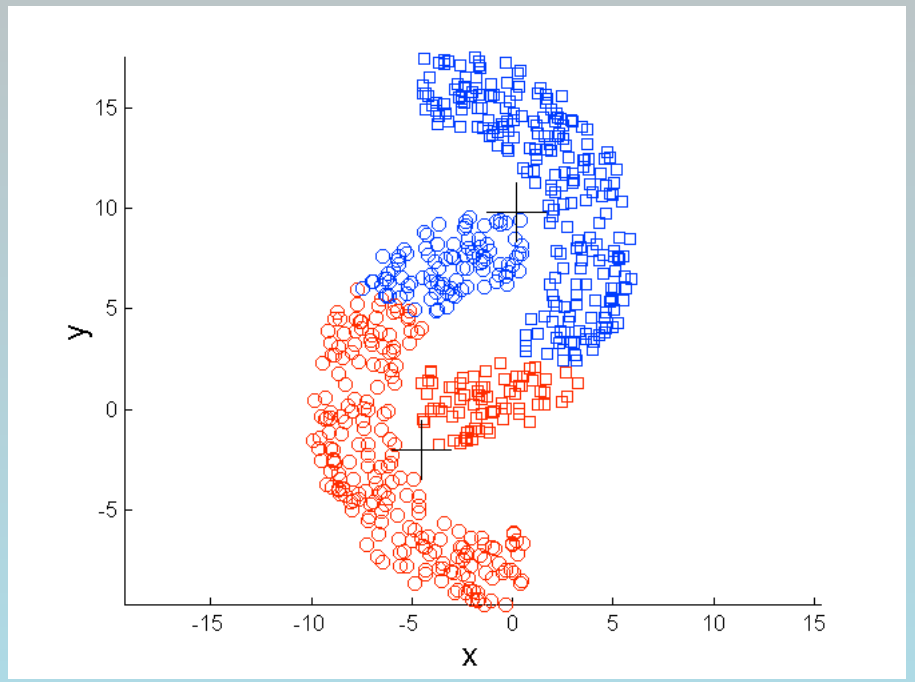
K-means (3 Clusters)

1. K-means

Problem with K-means: Non-globular Shapes



Original Points



K-means (2 Clusters)

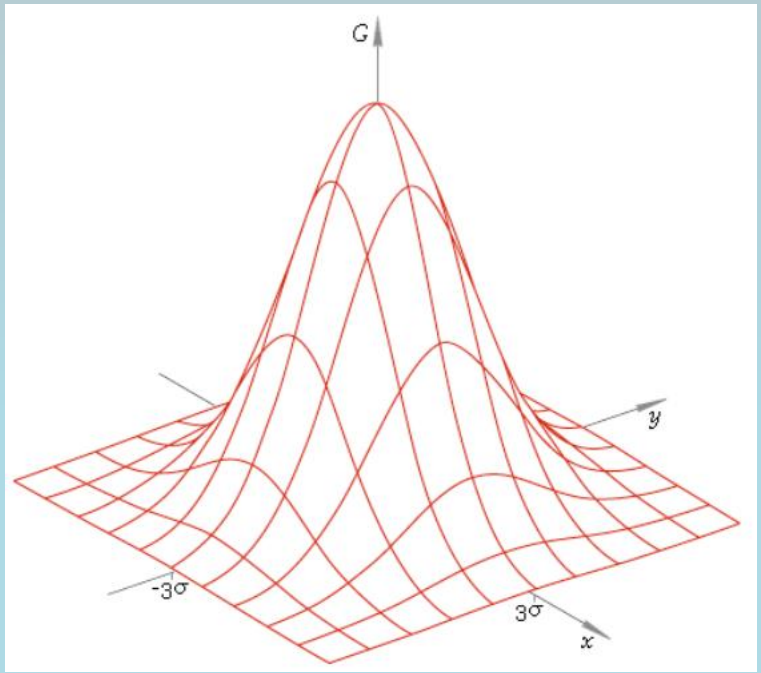
K-means summary

- Despite weaknesses, *k*-means is still the most popular algorithm due to its simplicity, efficiency and
  - other clustering algorithms have their own lists of weaknesses.
- No clear evidence that any other clustering algorithm performs better in general
  - although they may be more suitable for some specific types of data or applications.
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!

2. Gaussian Mixture Model

- What is a Gaussian?

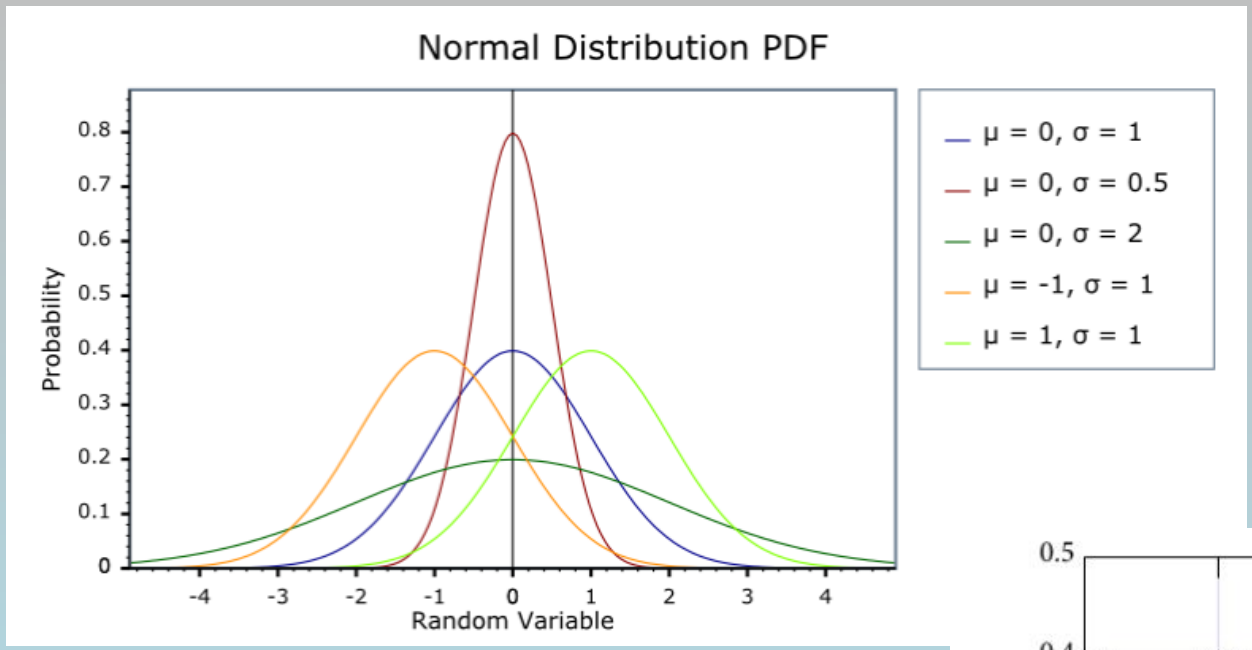
$$p(X) = \mathcal{N}(X|\mu, \Sigma)$$

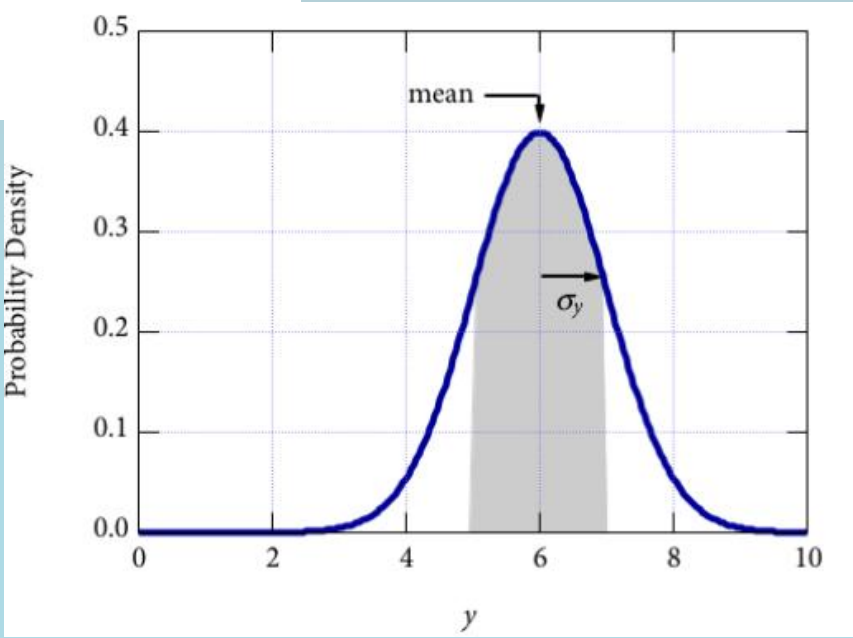


$\mathcal{N}$ = Gaussian ==Normal distribution

μ = Mean

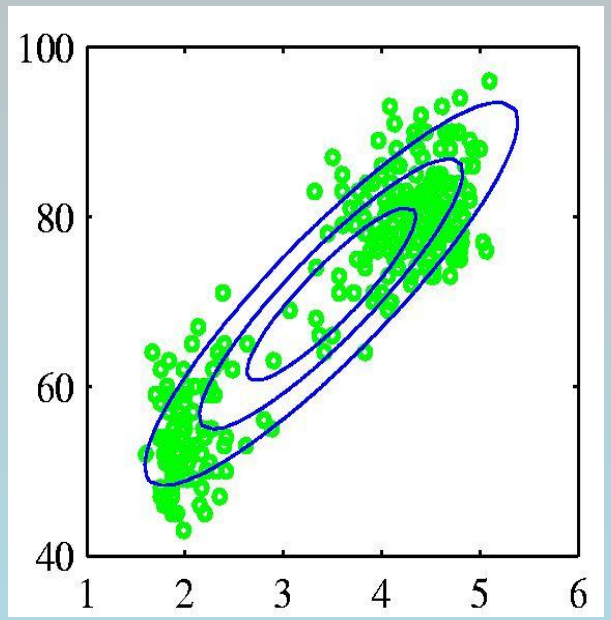
Σ = Variance

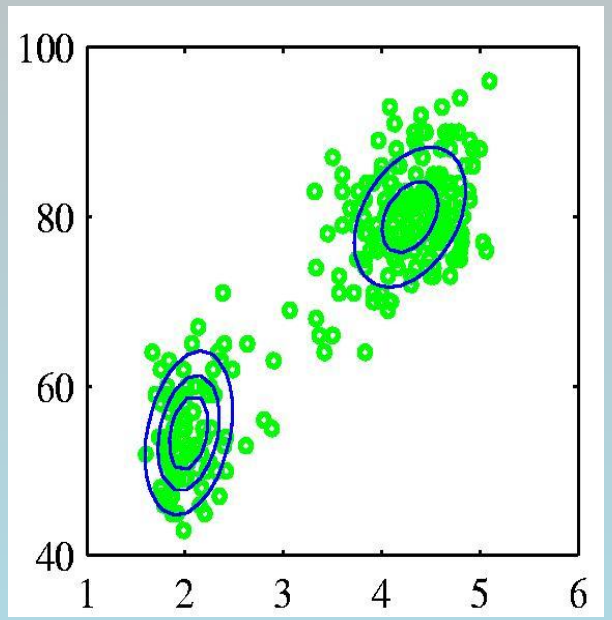

Normal Distribution PDF
Probability
Random Variable
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
-4
-3
-2
-1
0
1
2
3
4
μ = 0, σ = 1
μ = 0, σ = 0.5
μ = 0, σ = 2
μ = -1, σ = 1
μ = 1, σ = 1


Probability Density
mean
σ_y
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
y

2. Gaussian Mixture Model

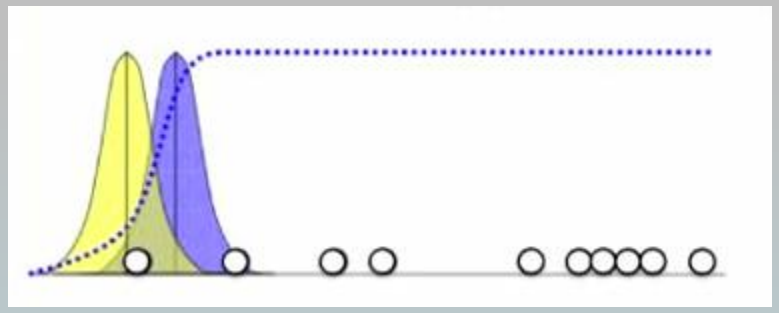
GMM: When one Gaussian is not enough

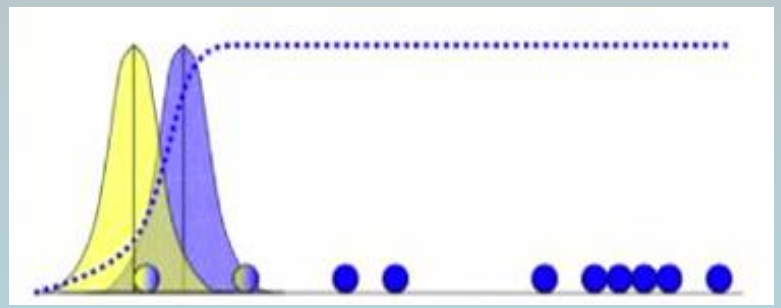


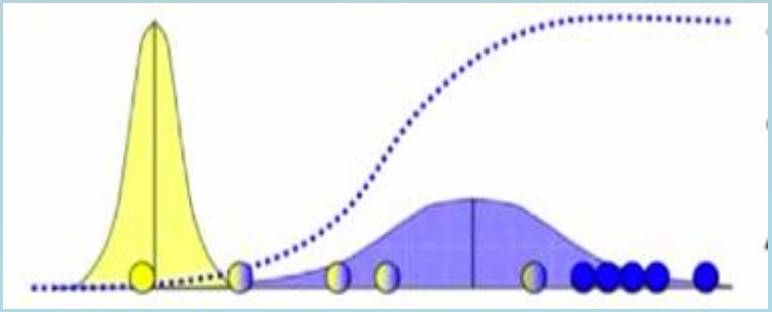


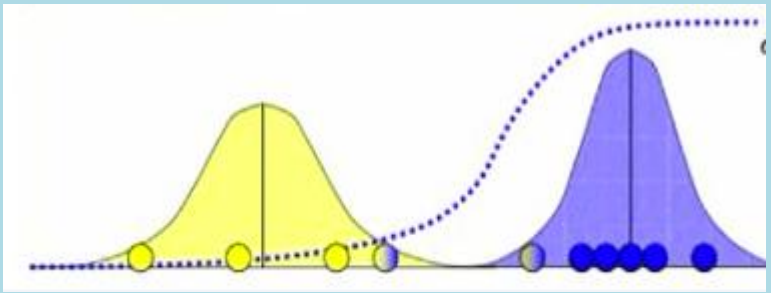
Gaussian Mixture Model (Multiple Gaussians)

- Real world datasets are rarely unimodal!

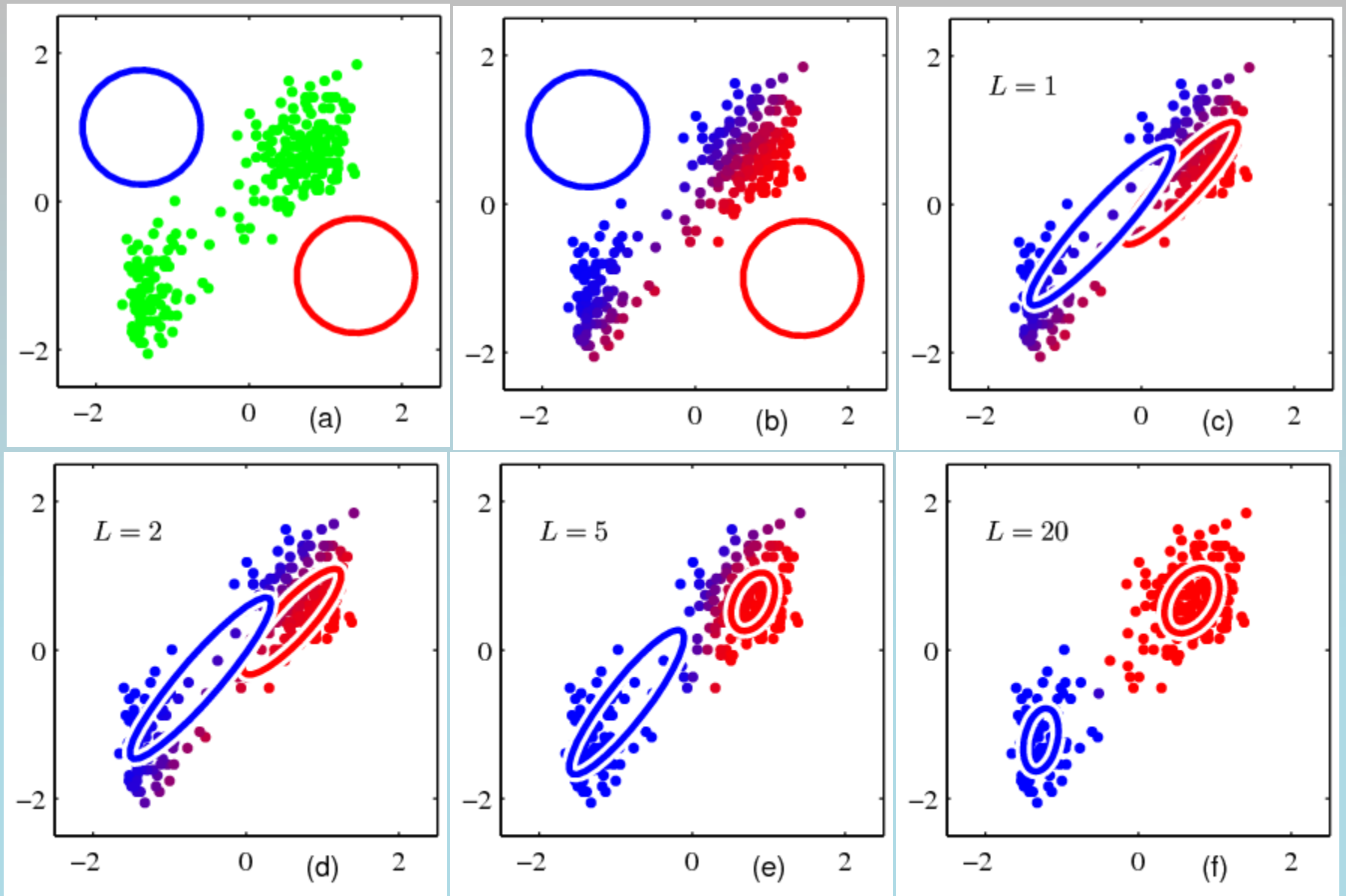






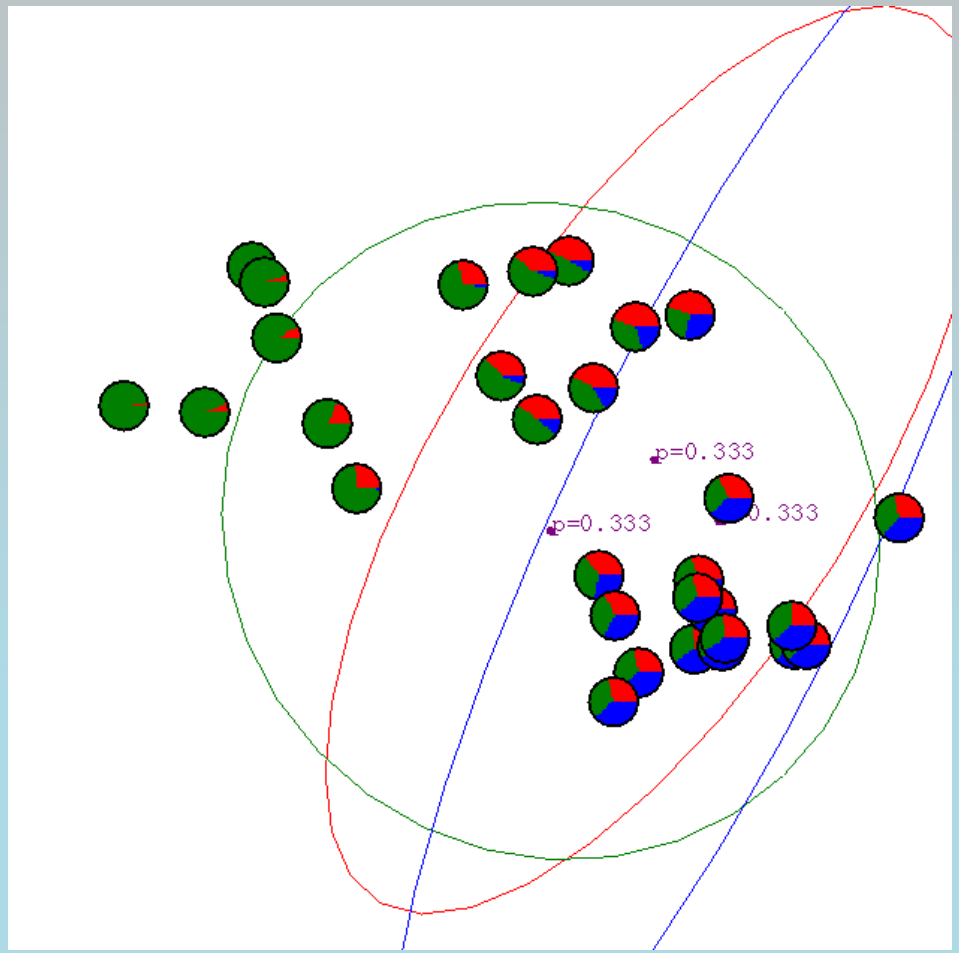


2. Gaussian Mixture Model



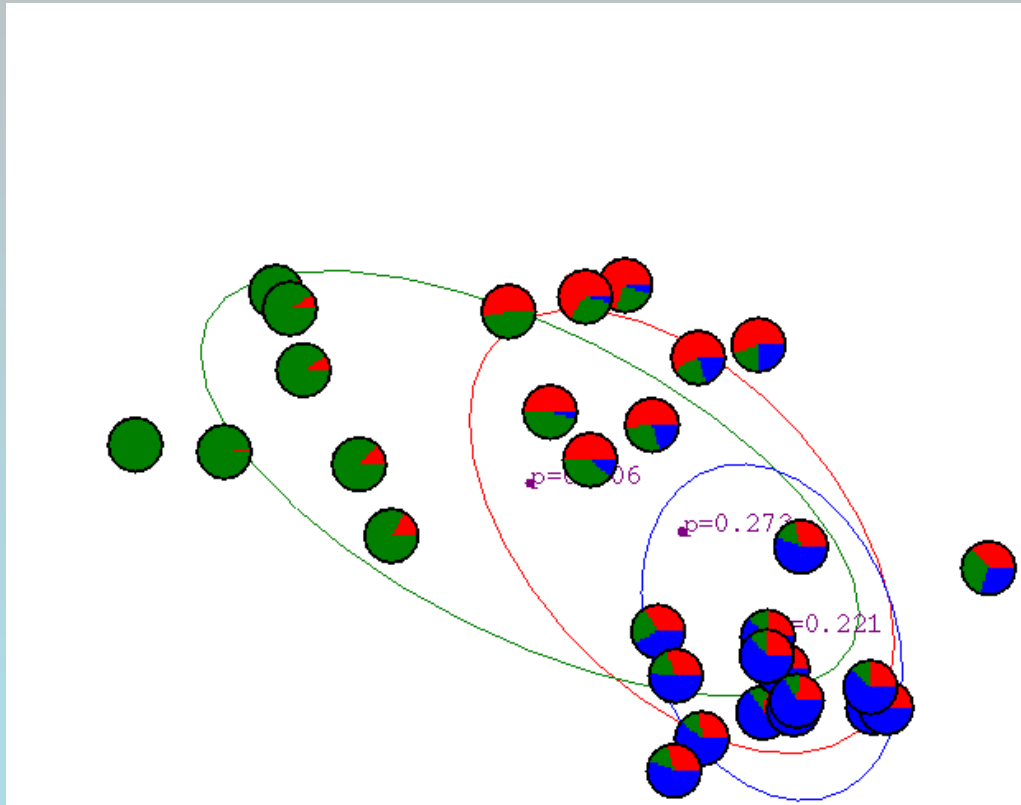
2. Gaussian Mixture Model

Another GMM Example: Start



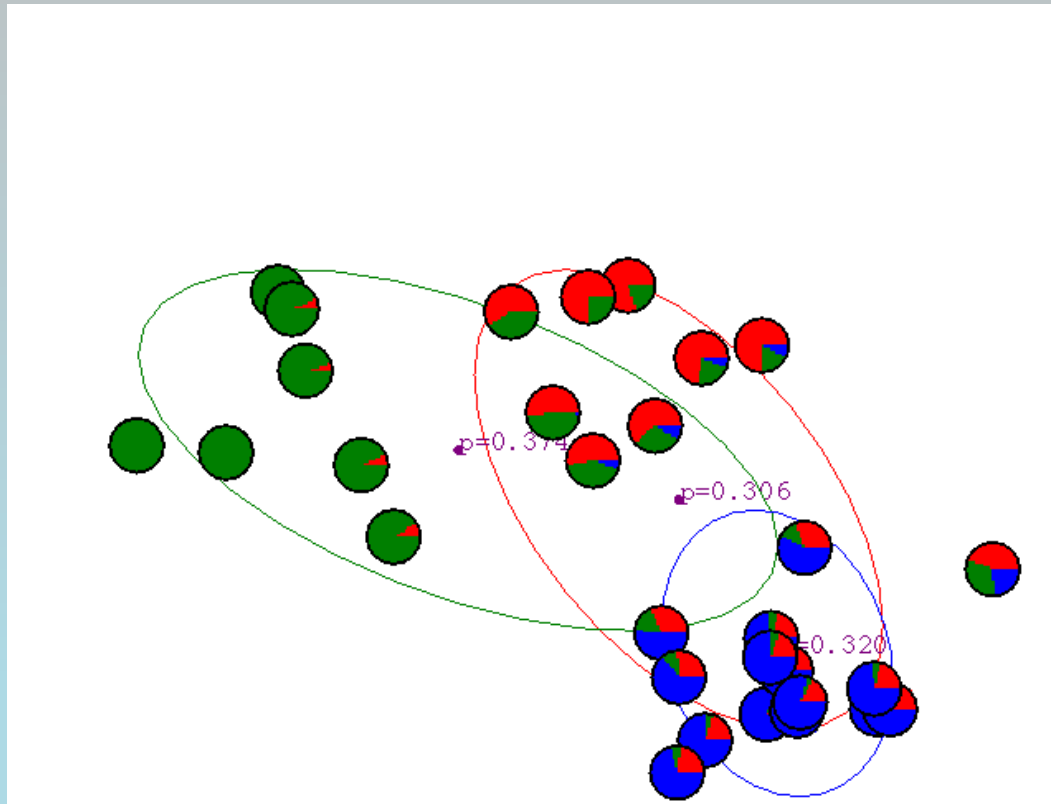
2. Gaussian Mixture Model

After First Iteration



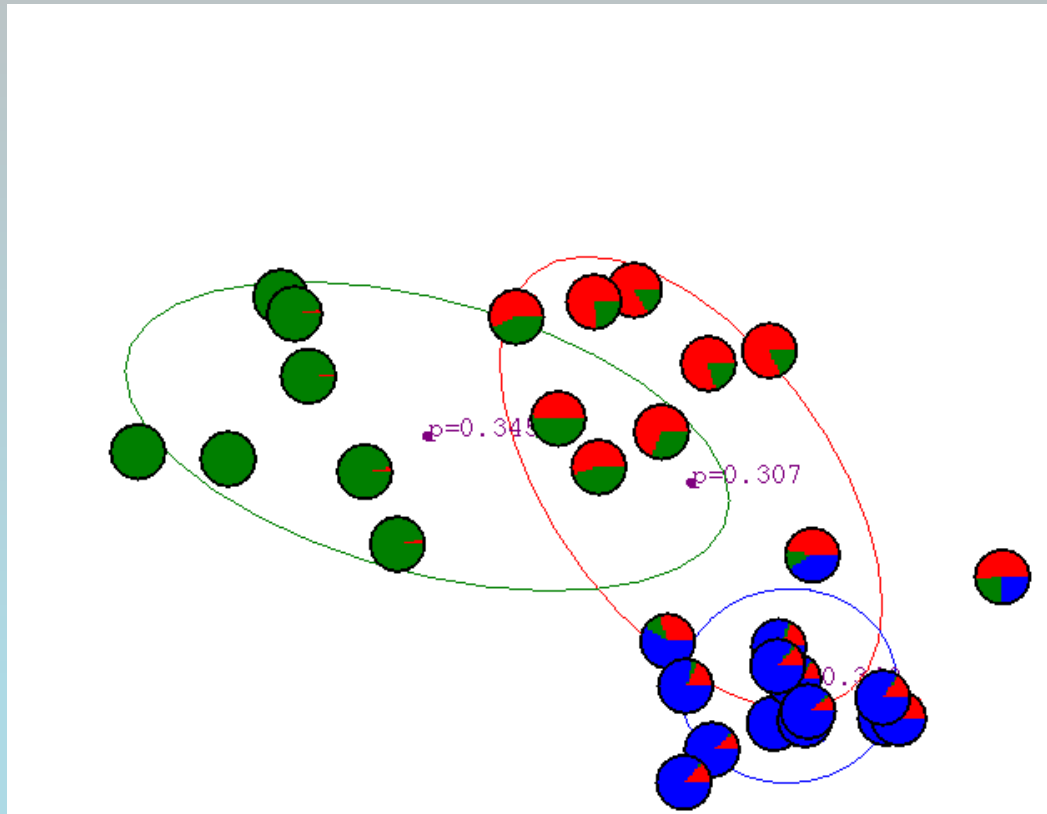
2. Gaussian Mixture Model

After 2nd Iteration



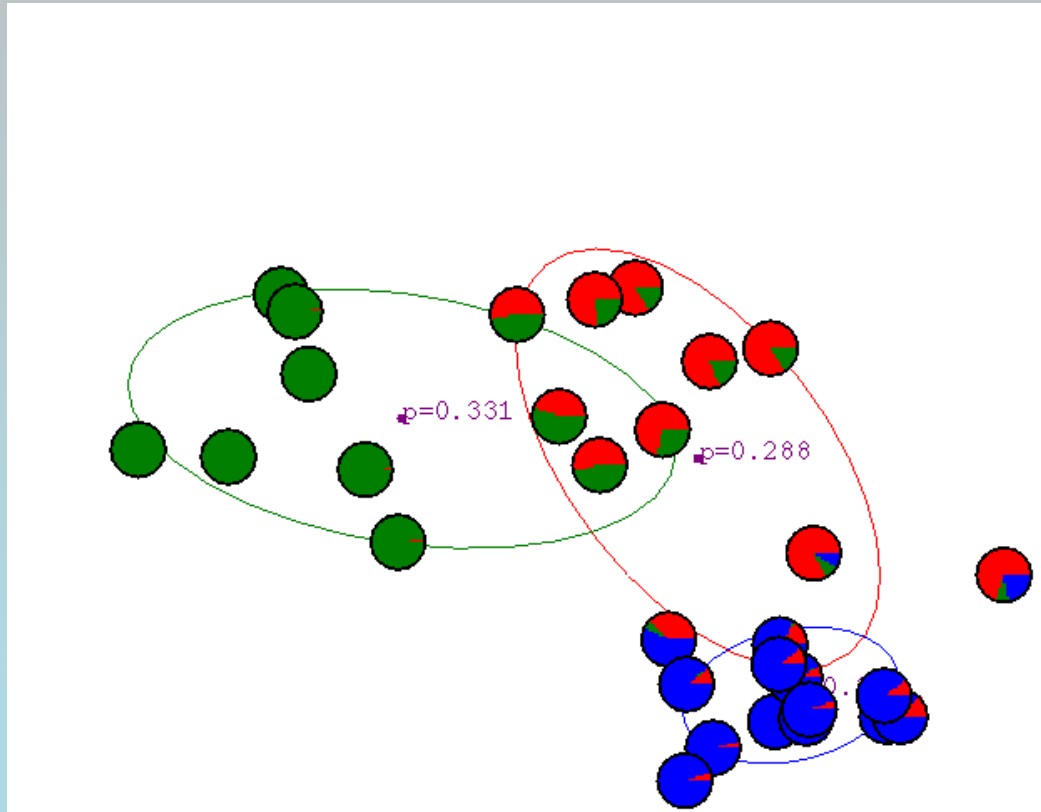
2. Gaussian Mixture Model

After 3rd Iteration



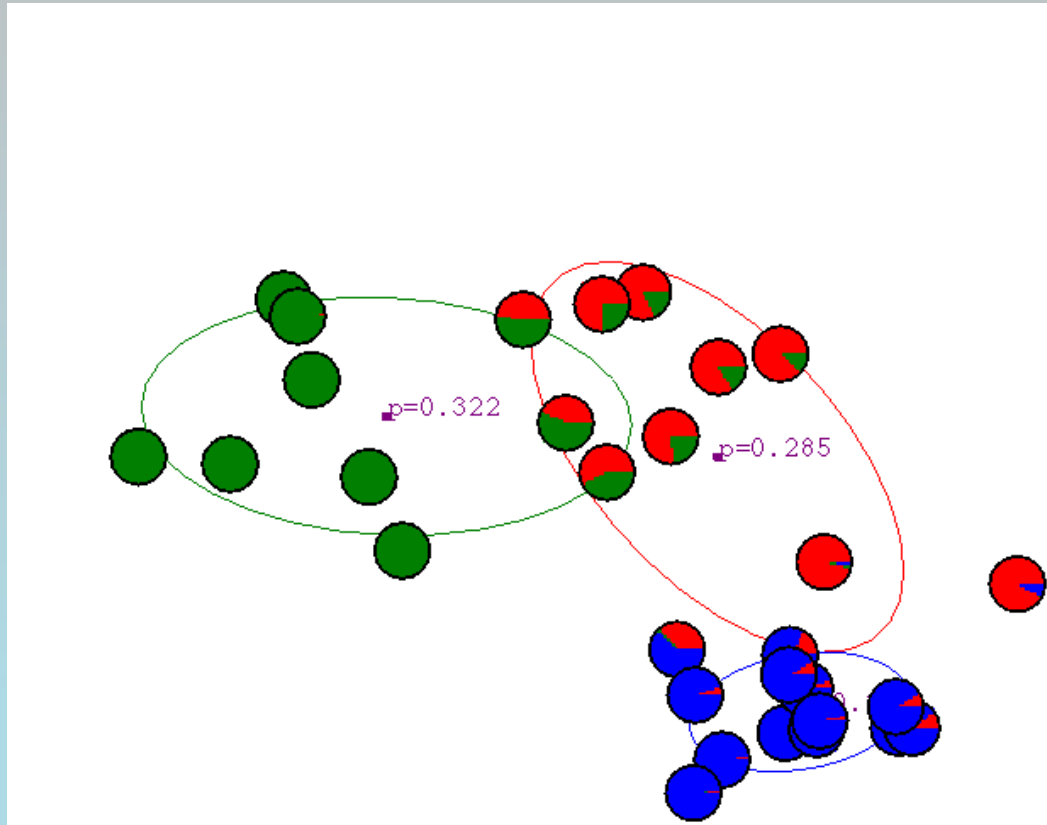
2. Gaussian Mixture Model

After 4th Iteration



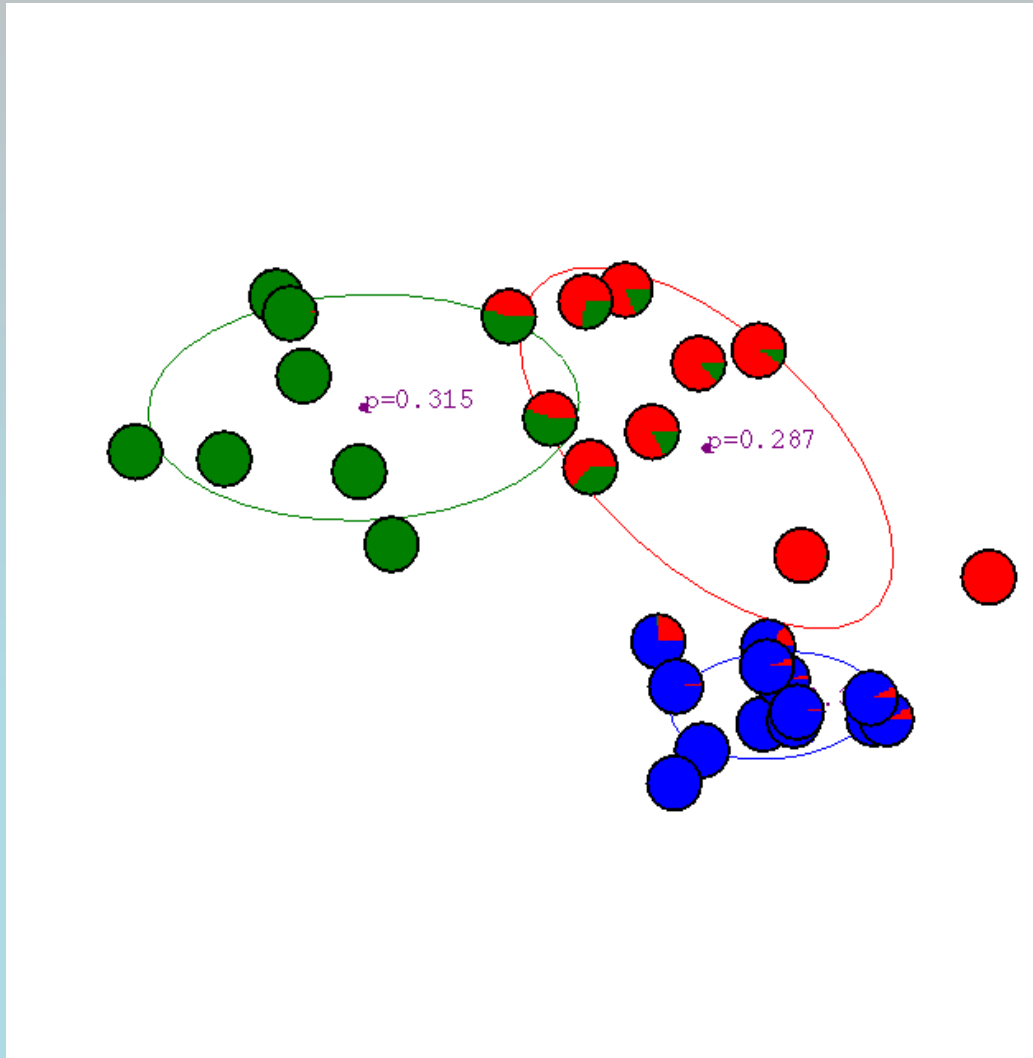
2. Gaussian Mixture Model

After 5th Iteration



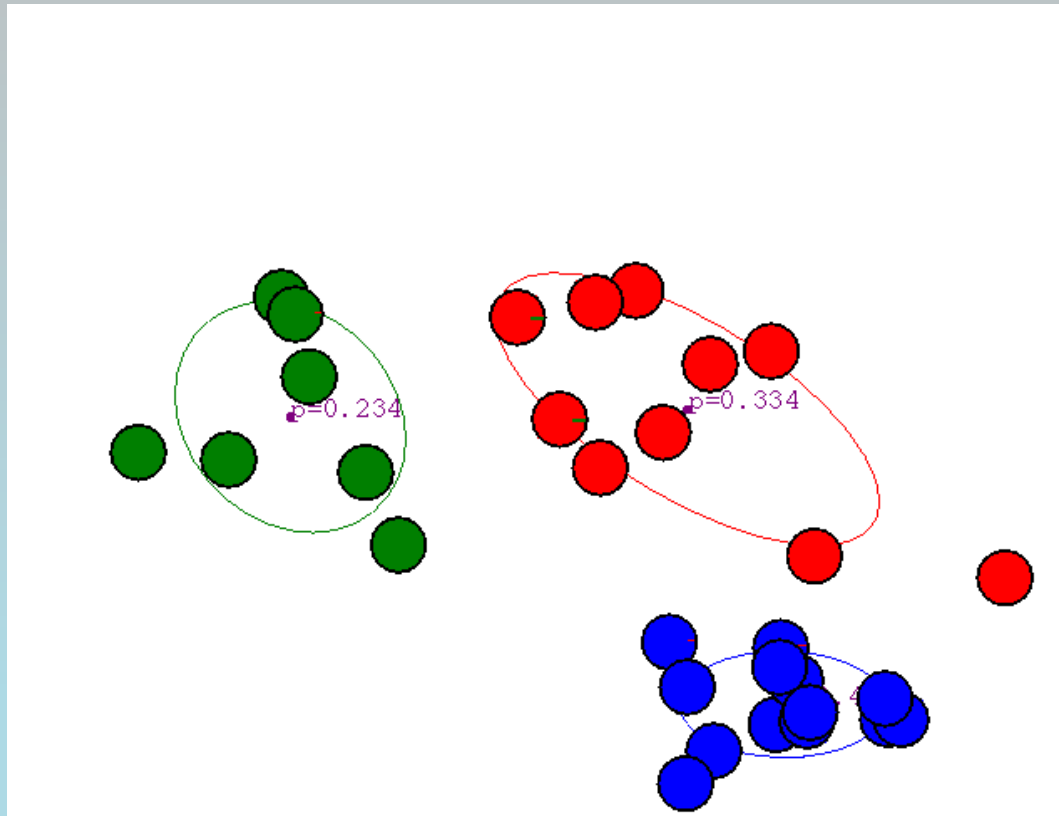
2. Gaussian Mixture Model

After 6th Iteration



2. Gaussian Mixture Model

After 20th Iteration



3. Mean Shift

- The mean shift algorithm seeks a *mode* or local maximum of density of a given distribution.

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

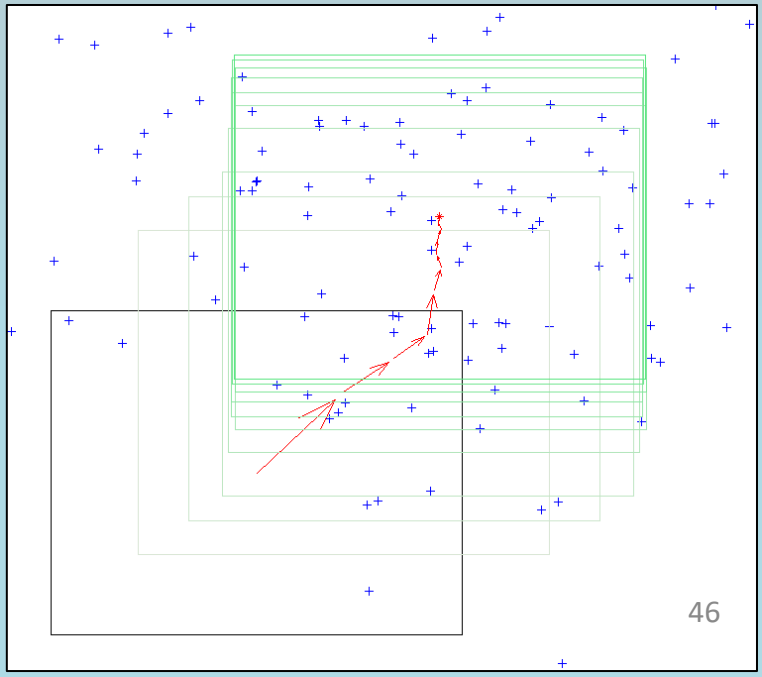
$K(x)$ is a kernel function that determines the weight of nearby points for re-estimation of the mean. The popular one:

1) Flat Kernel
2) Gaussian Kernel

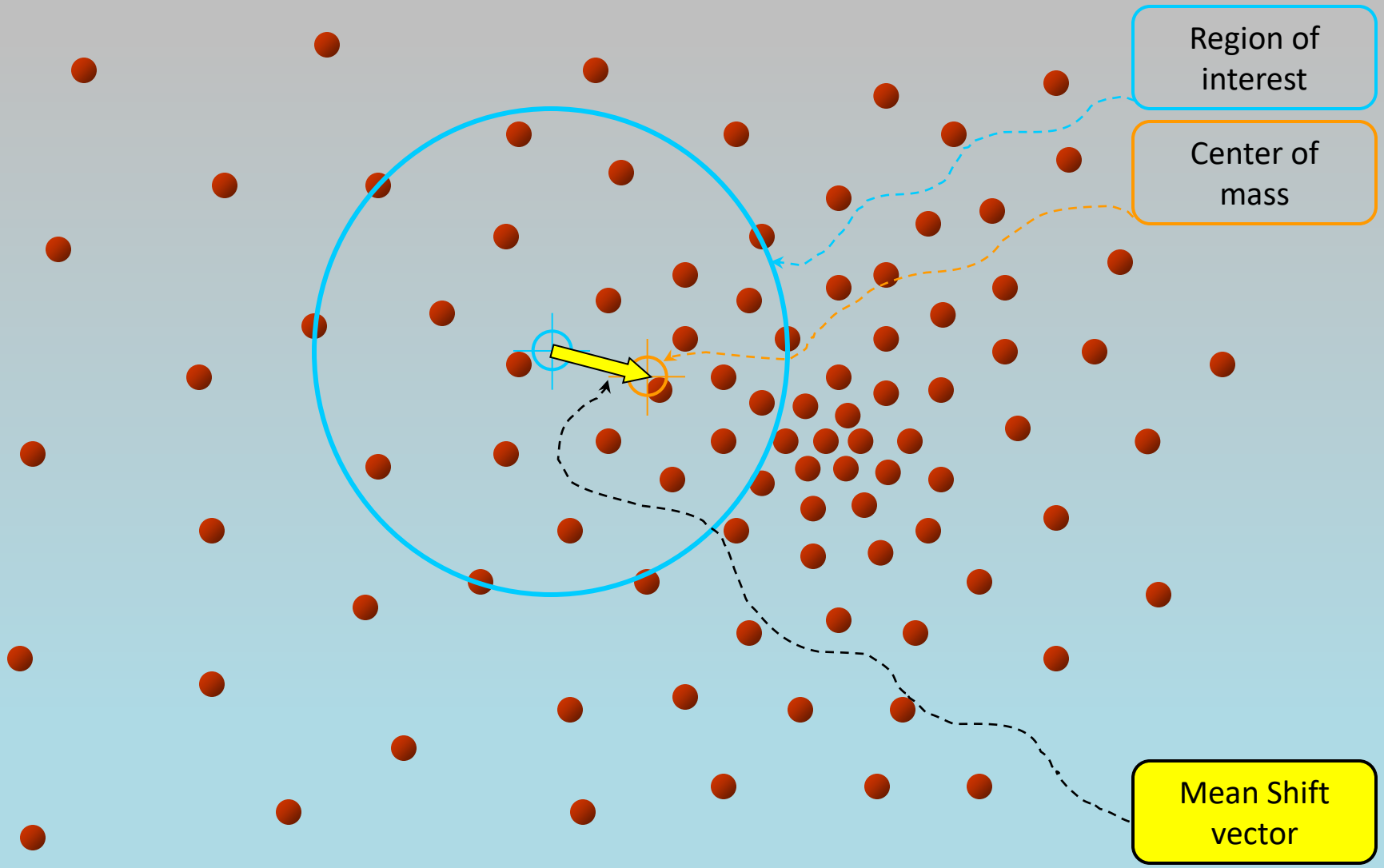
- where $N(x)$ is the neighborhood of x, a set of points for which $K(x) \neq 0$.
- The difference $m(x) - x$ is called mean shift.

3. Mean Shift

- **Algorithm**:
  1. Choose a search window (width and location)
  2. Compute the mean of the data in the search window
  3. Center the search window at the new mean location
  4. Repeat until convergence

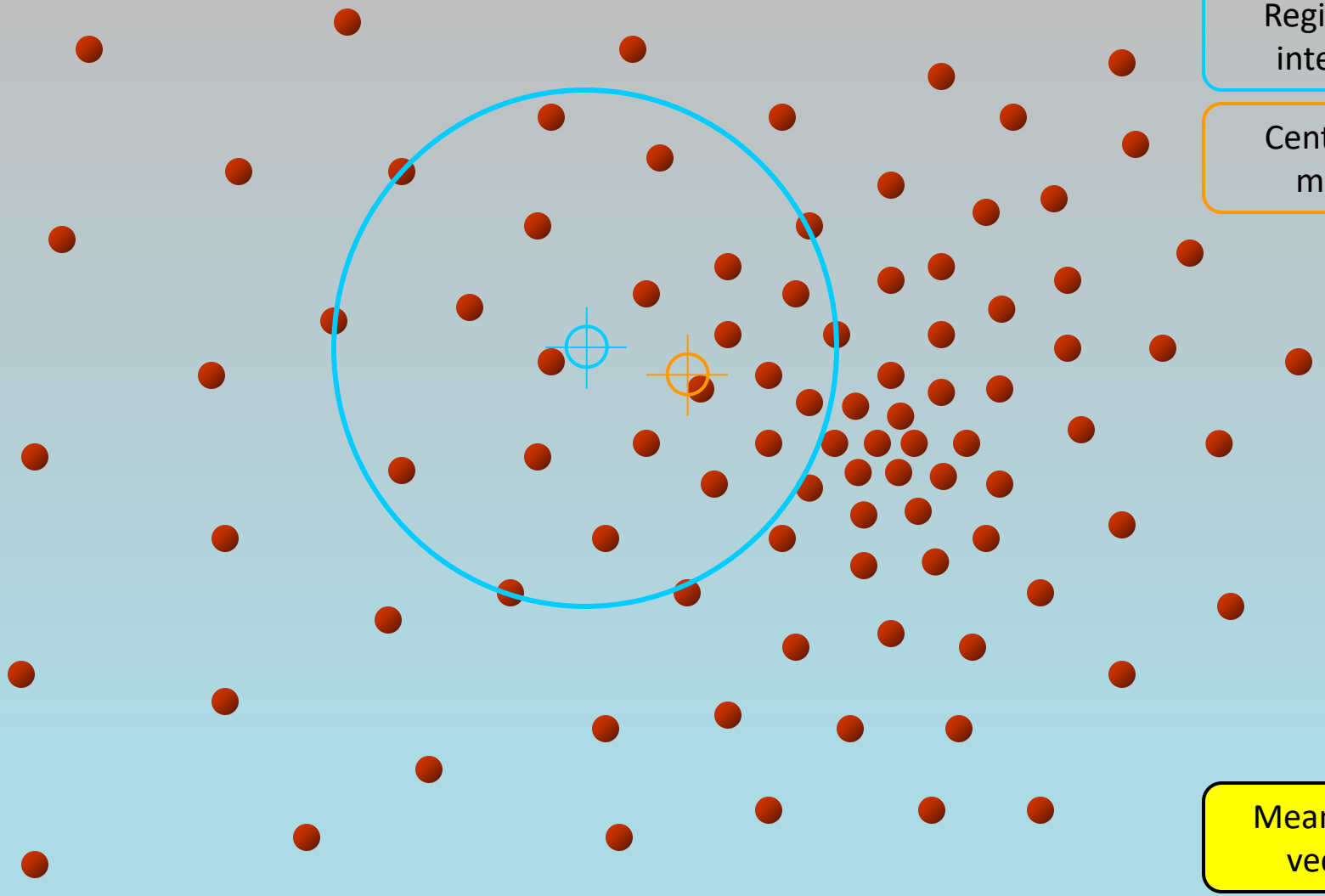


3. Mean Shift



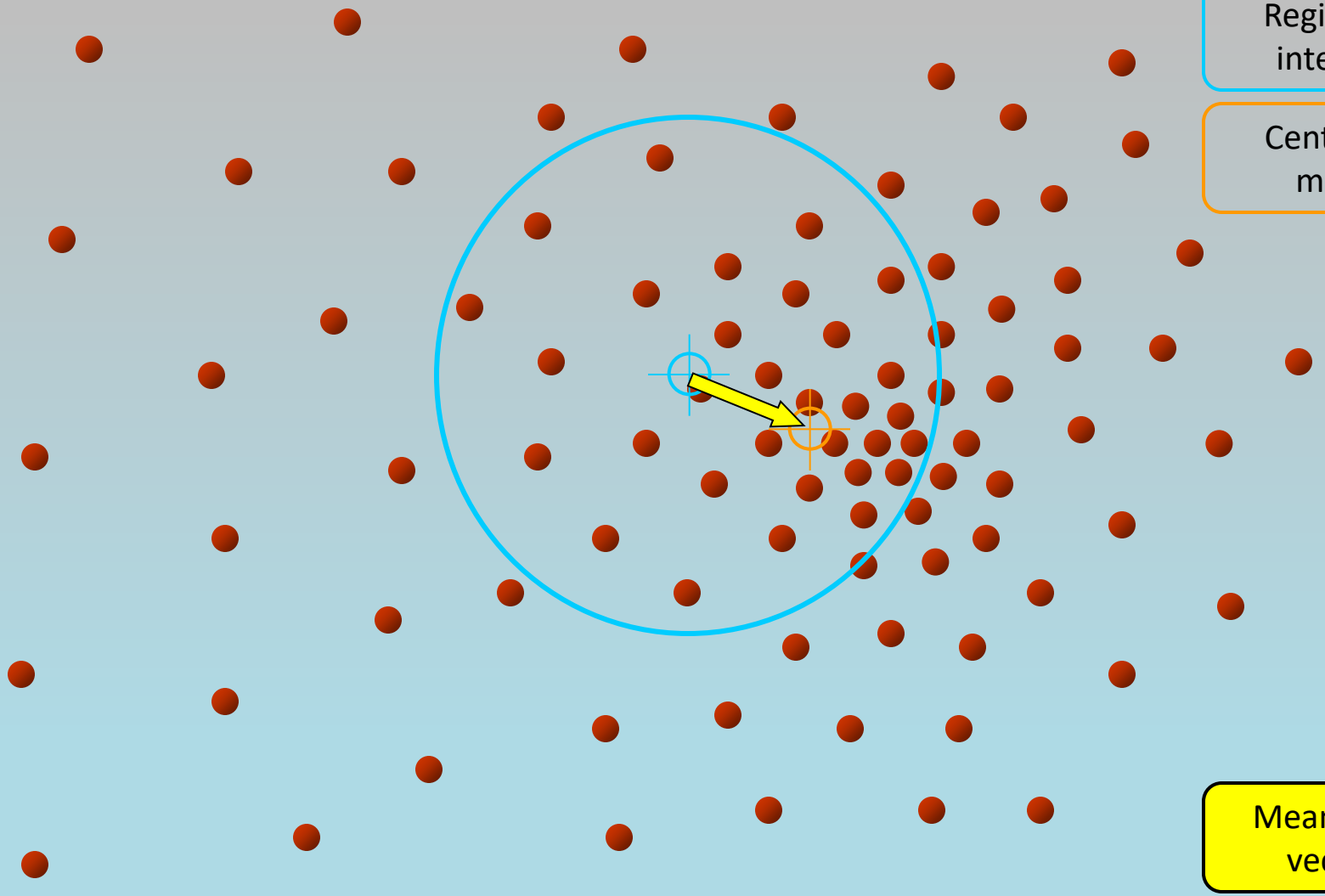
Adopted from Slide by Y. Ukrainitz & B. Sarel

3. Mean Shift



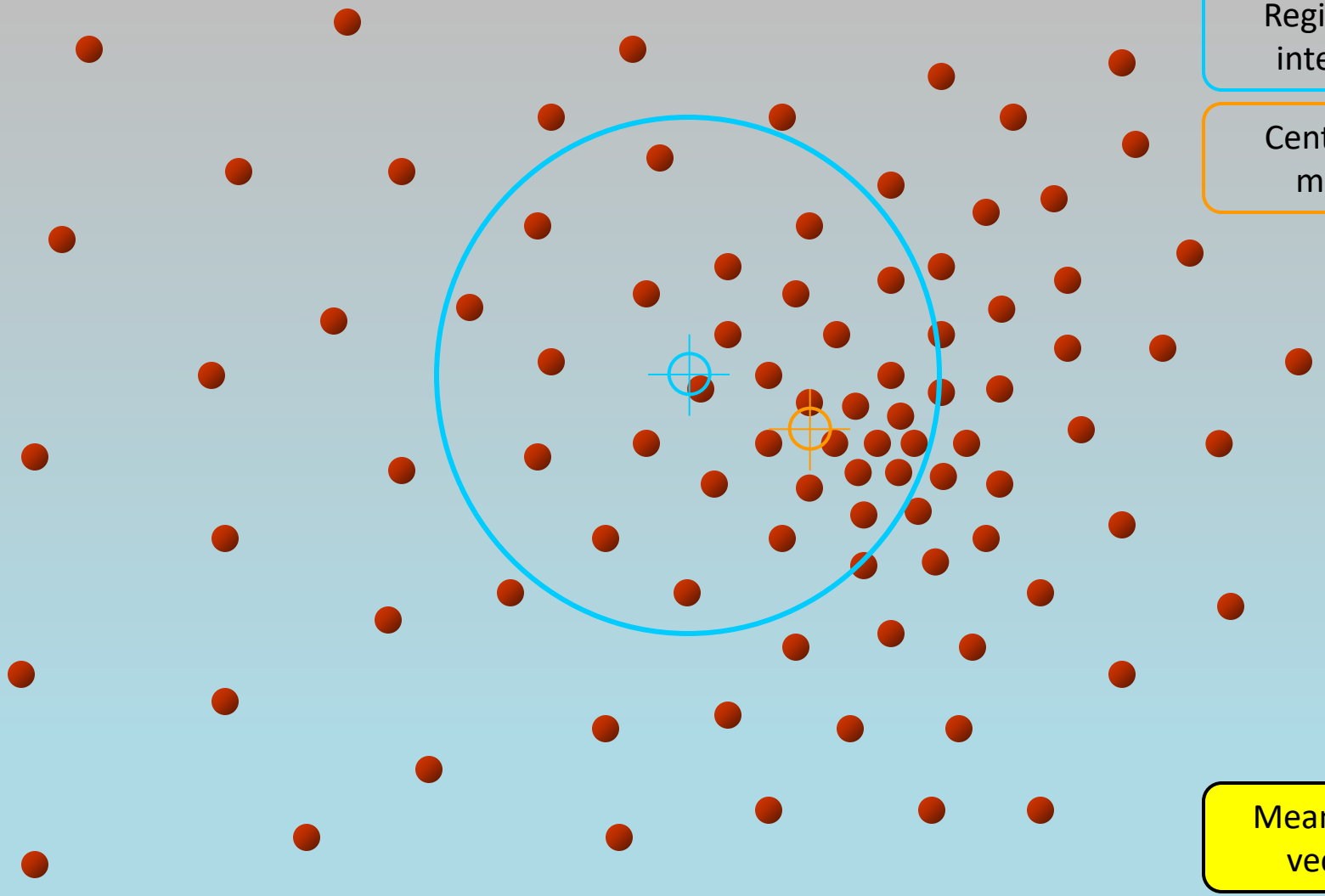
Adopted from Slide by Y. Ukrainitz & B. Sarel

3. Mean Shift



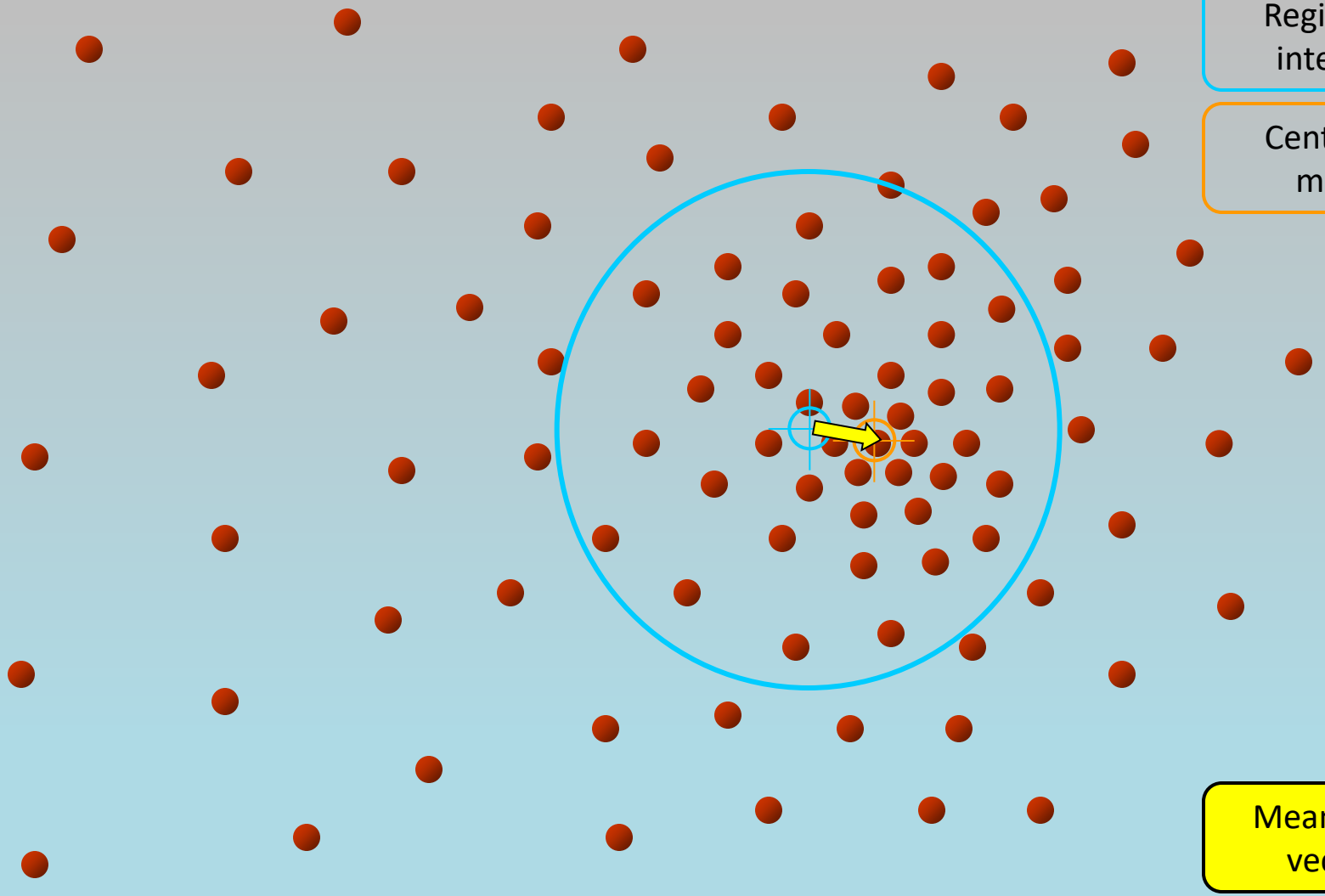
Adopted from Slide by Y. Ukrainitz & B. Sarel

3. Mean Shift



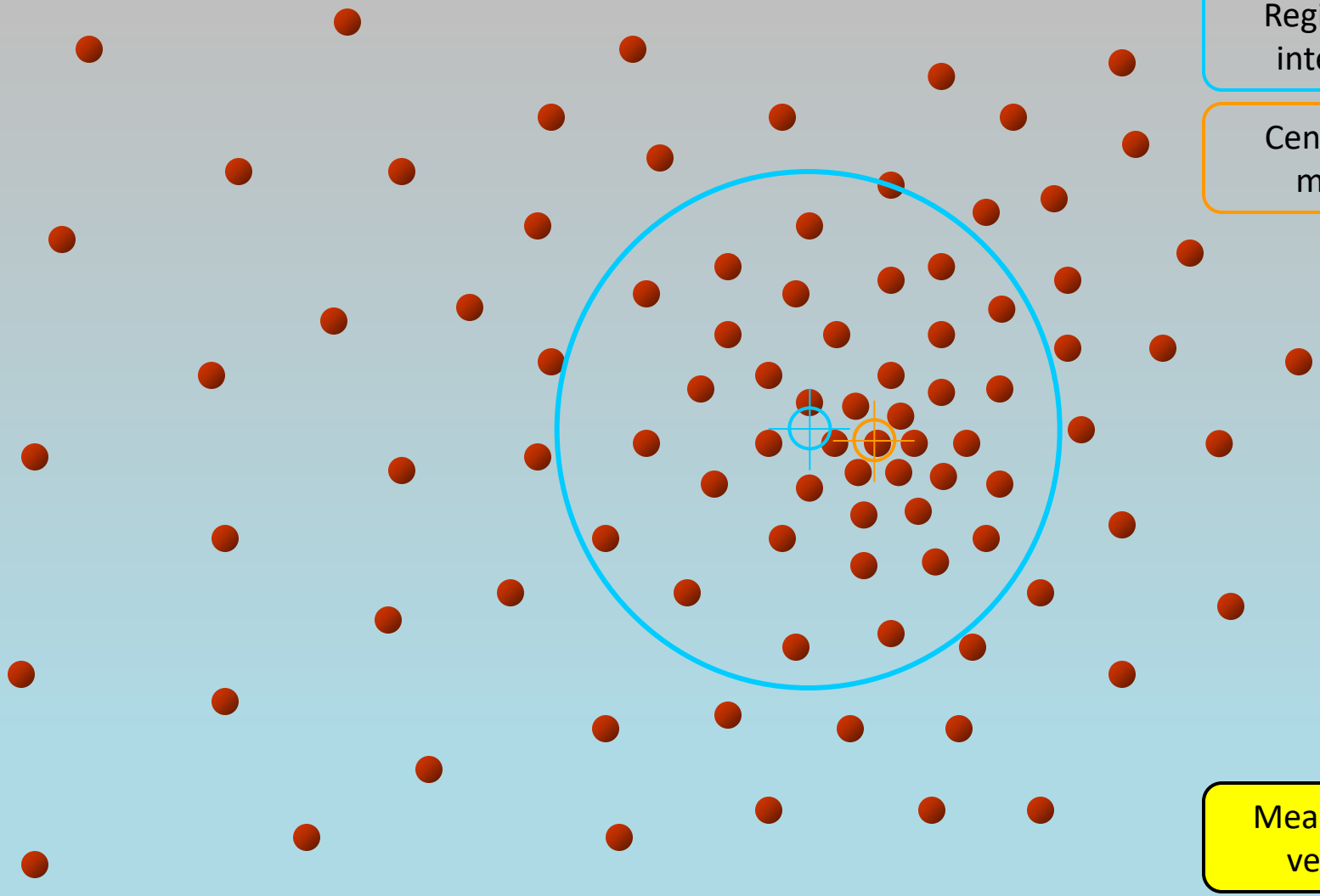
Adopted from Slide by Y. Ukrainitz & B. Sarel

3. Mean Shift



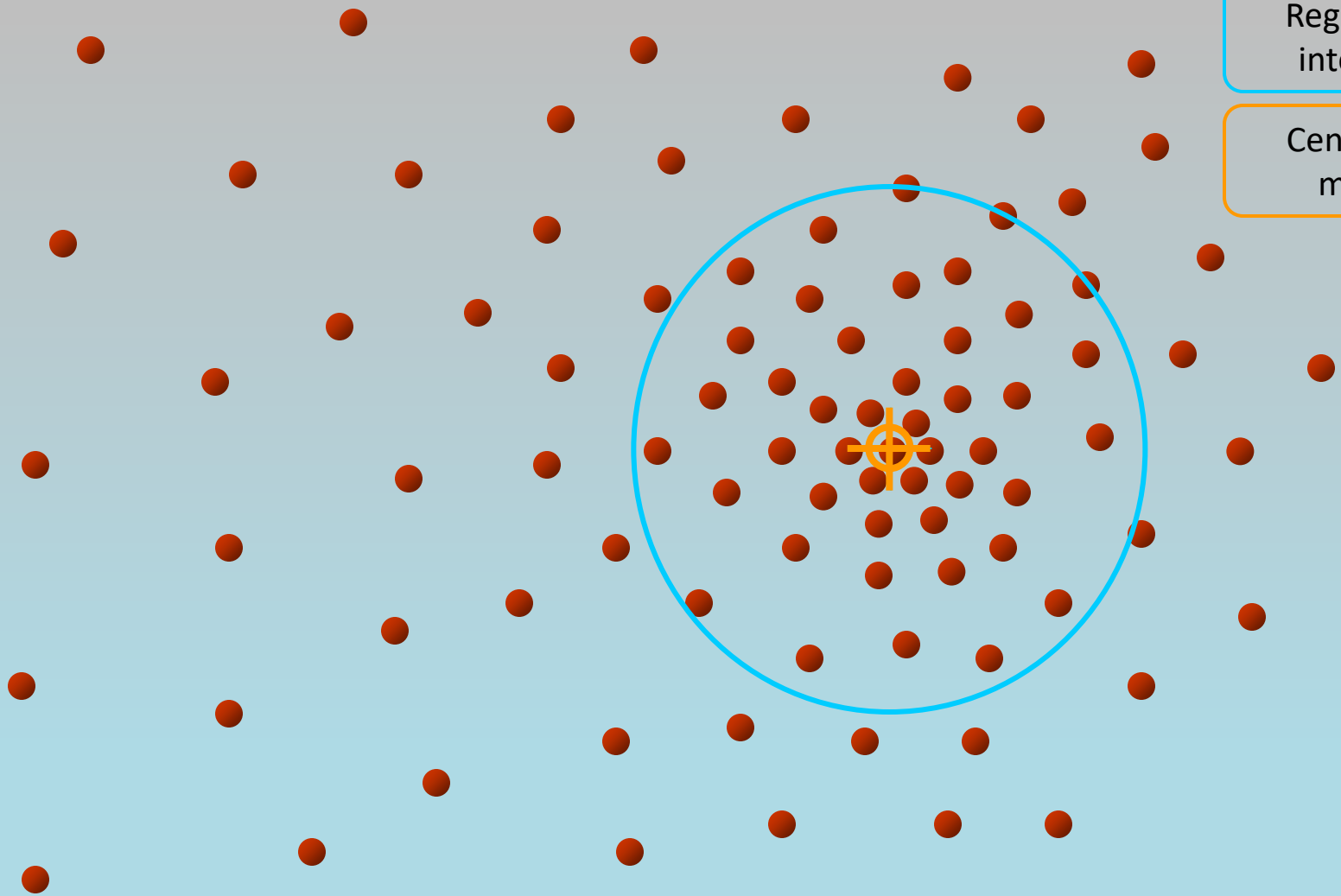
Adopted from Slide by Y. Ukrainitz & B. Sarel

3. Mean Shift



Adopted from Slide by Y. Ukrainitz & B. Sarel

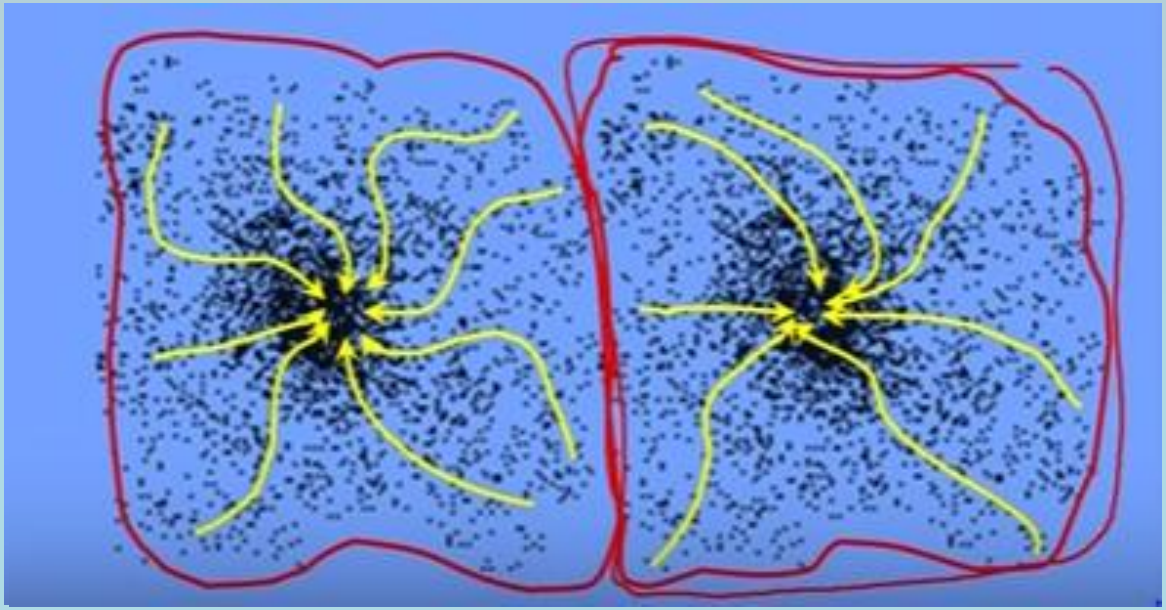
3. Mean Shift



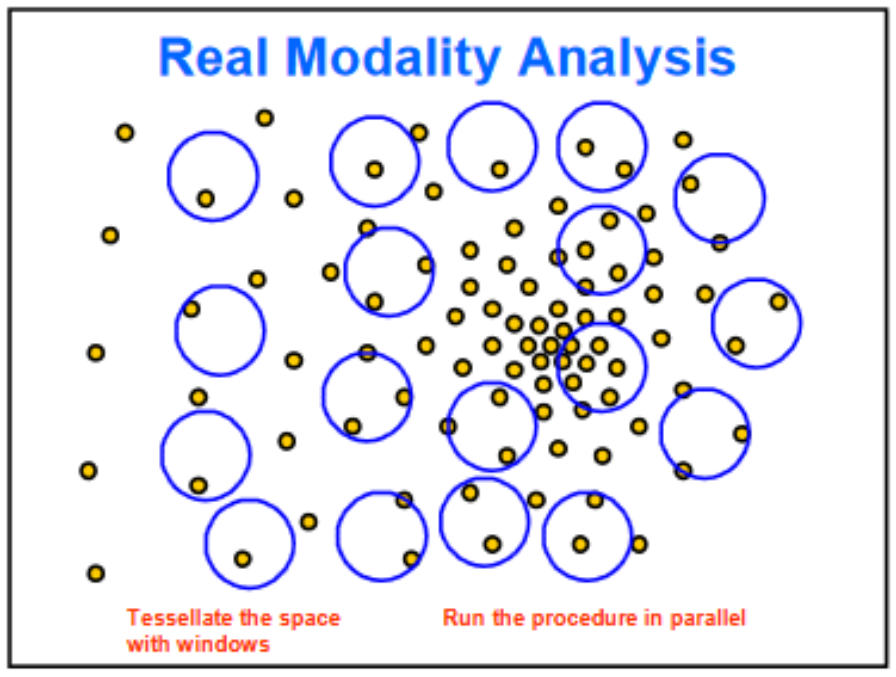
Adopted from Slide by Y. Ukrainitz & B. Sarel

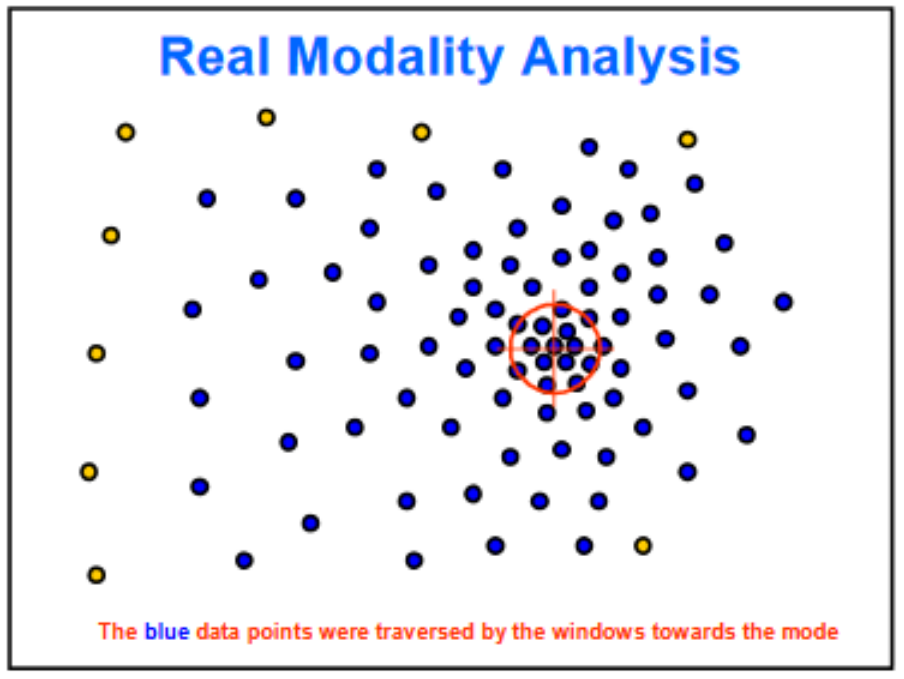
3. Mean Shift

- Cluster: all data points in the attraction basin of a mode
- Attraction basin: the region for which all trajectories lead to the same mode



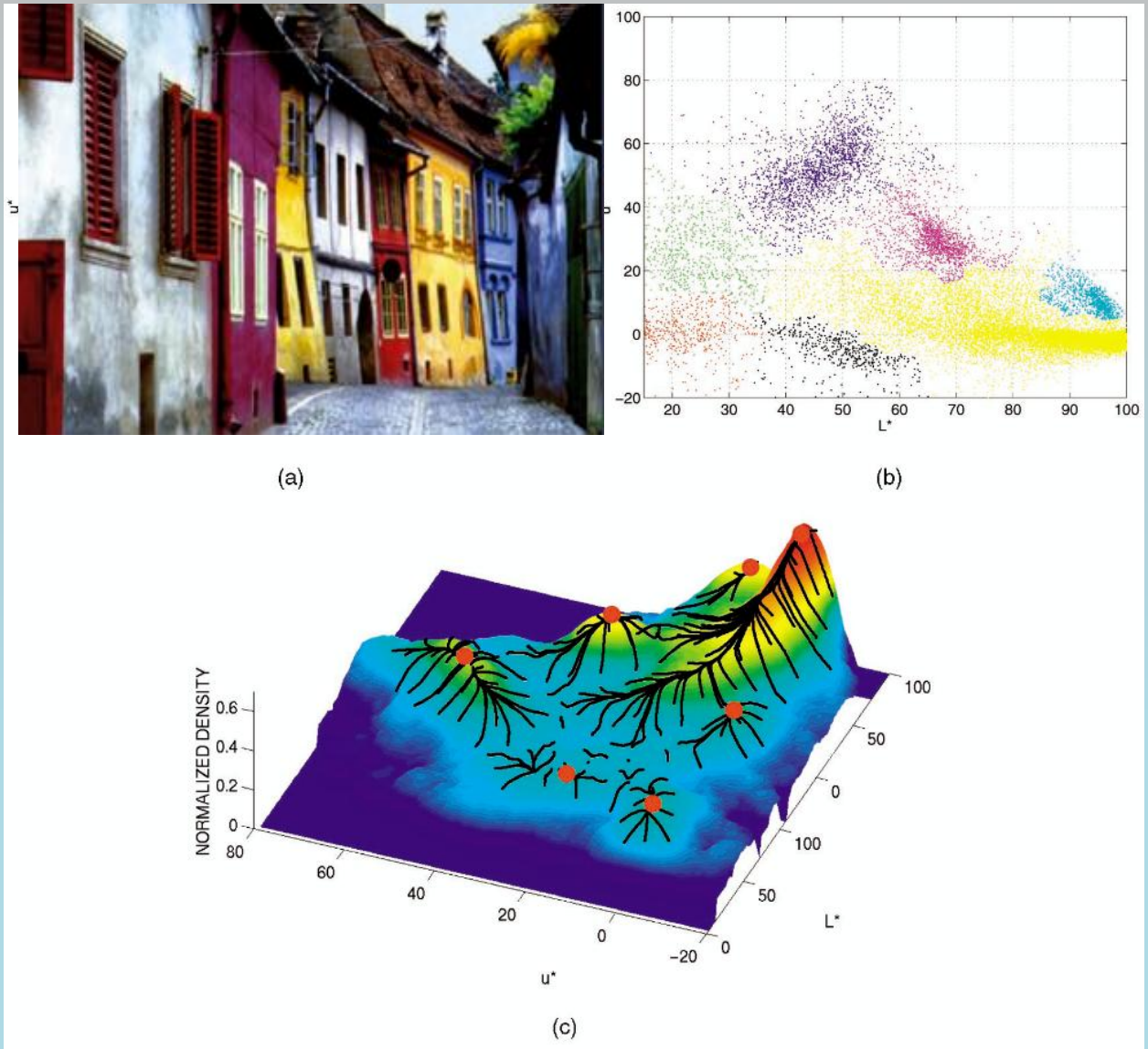
Slide by Y. Ukrainitz & B. Sarel


Real Modality Analysis
Tessellate the space with windows
Run the procedure in parallel


Real Modality Analysis
The blue data points were traversed by the windows towards the mode

3. Mean Shift

- Another example:



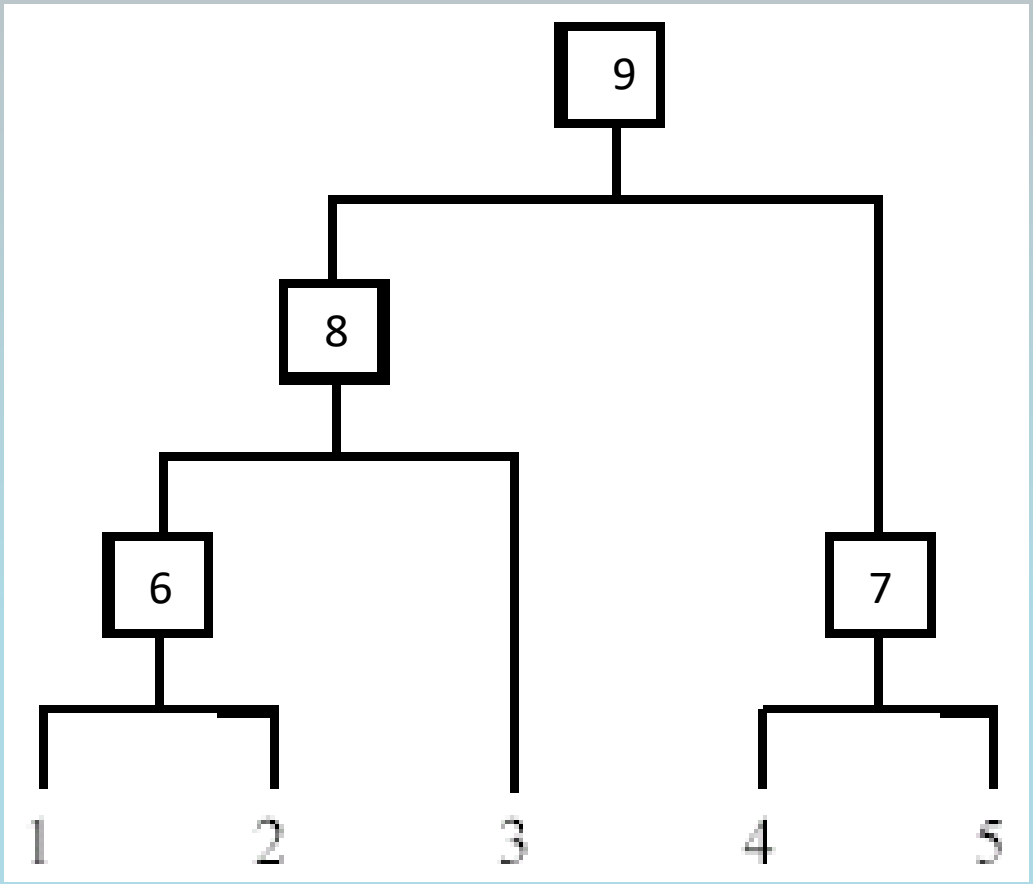
3. Mean Shift

Mean shift advantages and disadvantages

- Advantages
  - ❖Does not assume number of clusters
  - ❖Just a single parameter (window size)
  - ❖Finds variable number of modes
  - ❖Robust to outliers
- Disadvantages
  - ❖Output depends on window size
  - ❖Computationally expensive

4. Hierarchical clustering

- Produce a nested sequence of clusters, a tree, also called Dendrogram.



4. Hierarchical clustering

Types of hierarchical clustering

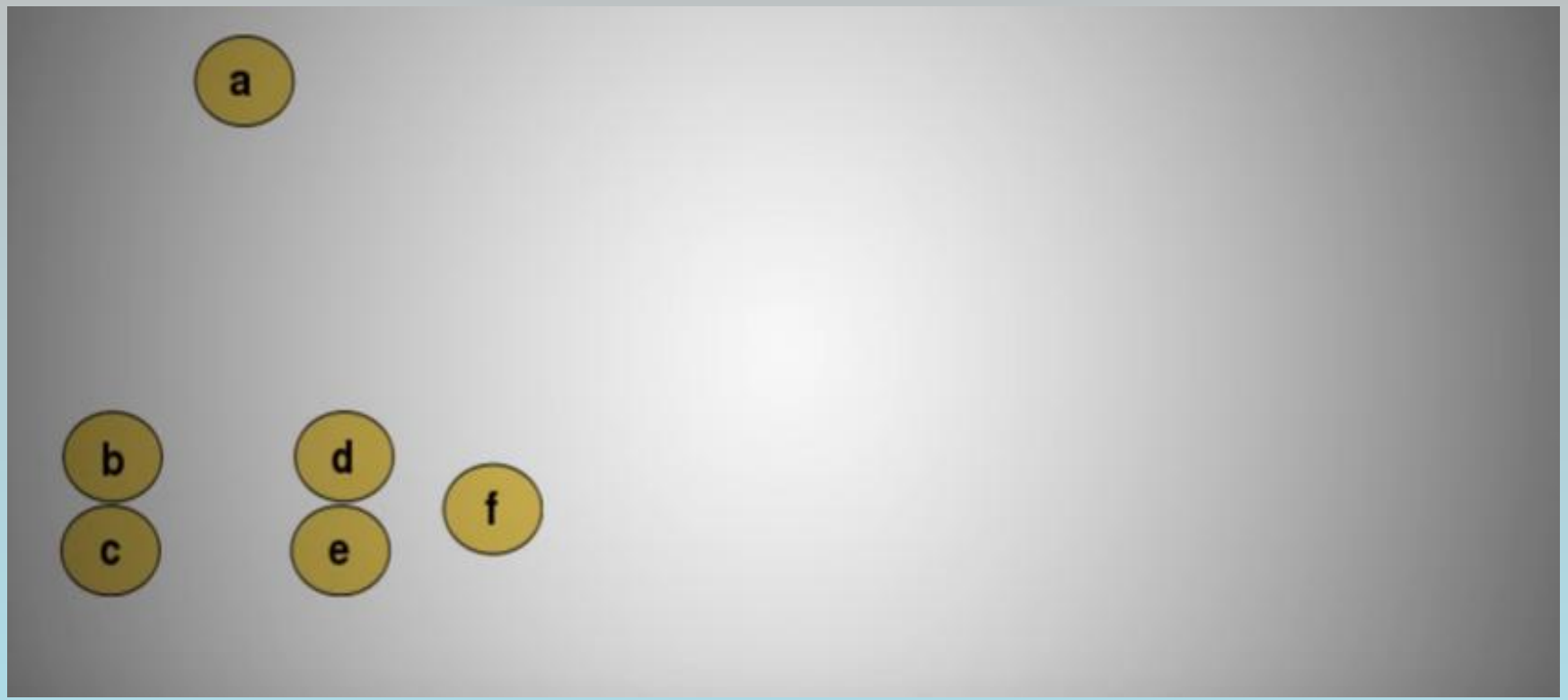
- Agglomerative (bottom up) clustering: It builds the dendrogram (tree) from the bottom level, and
  - merges the most similar (or nearest) pair of clusters
  - stops when all the data points are merged into a single cluster (i.e., the root cluster).
- Divisive (top down) clustering: It starts with all data points in one cluster, the root.
  - Splits the root into a set of child clusters. Each child cluster is recursively divided further
  - stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point

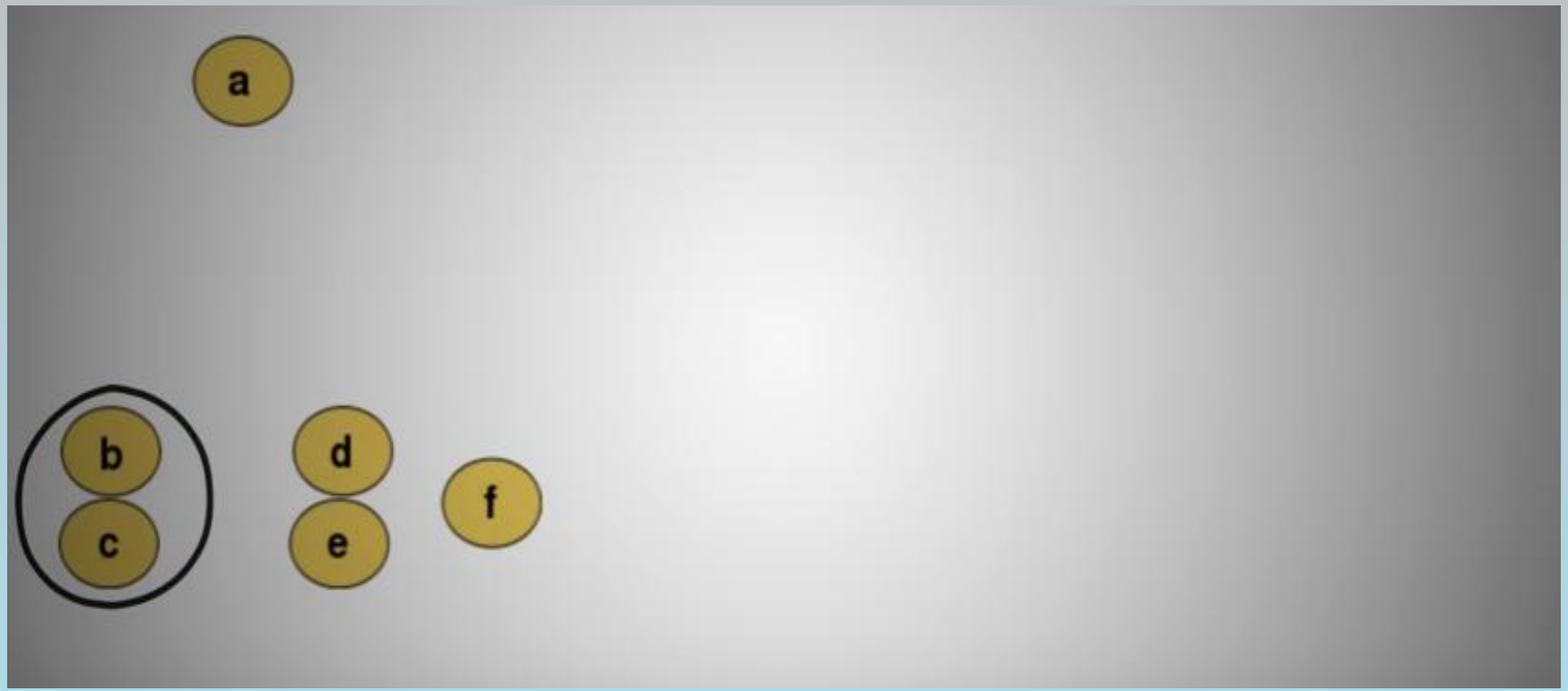
4. Hierarchical clustering

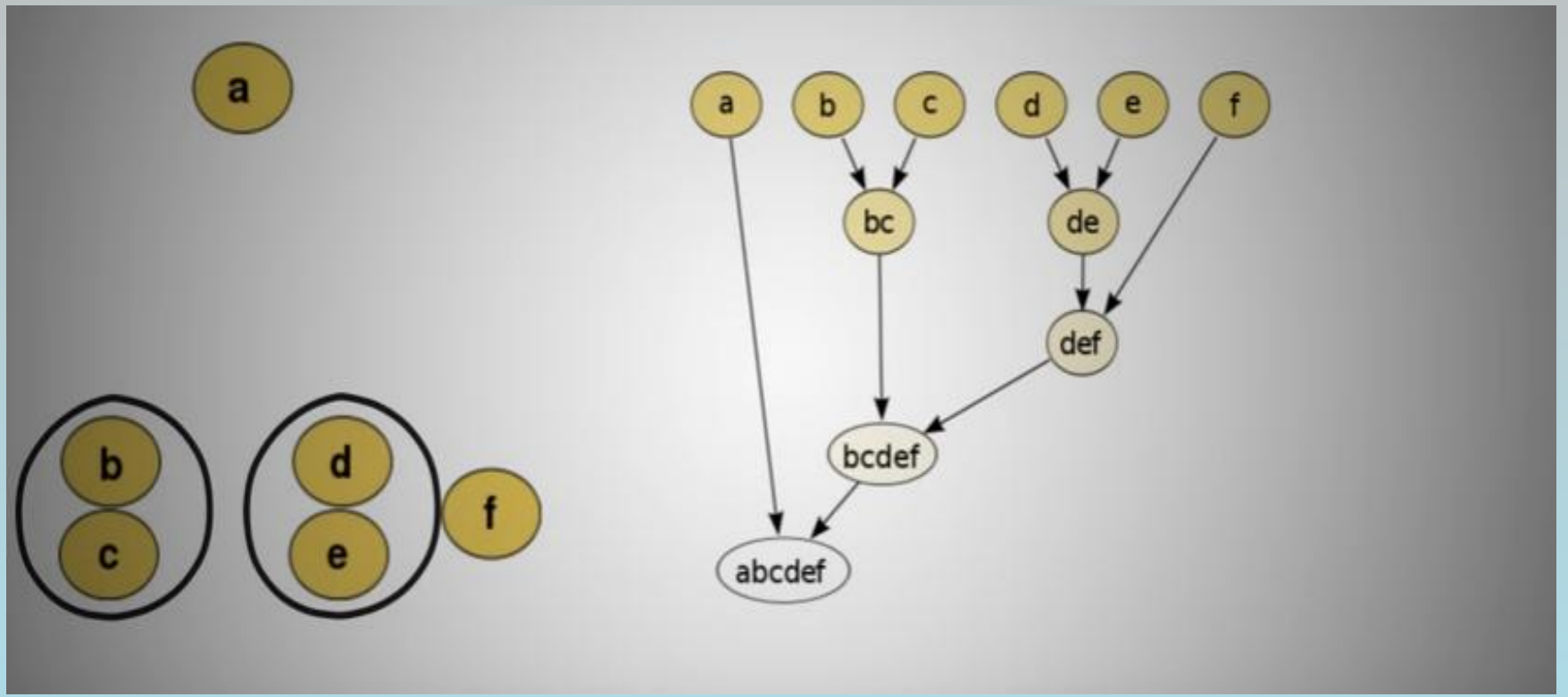
Agglomerative clustering (more popular then divisive methods).

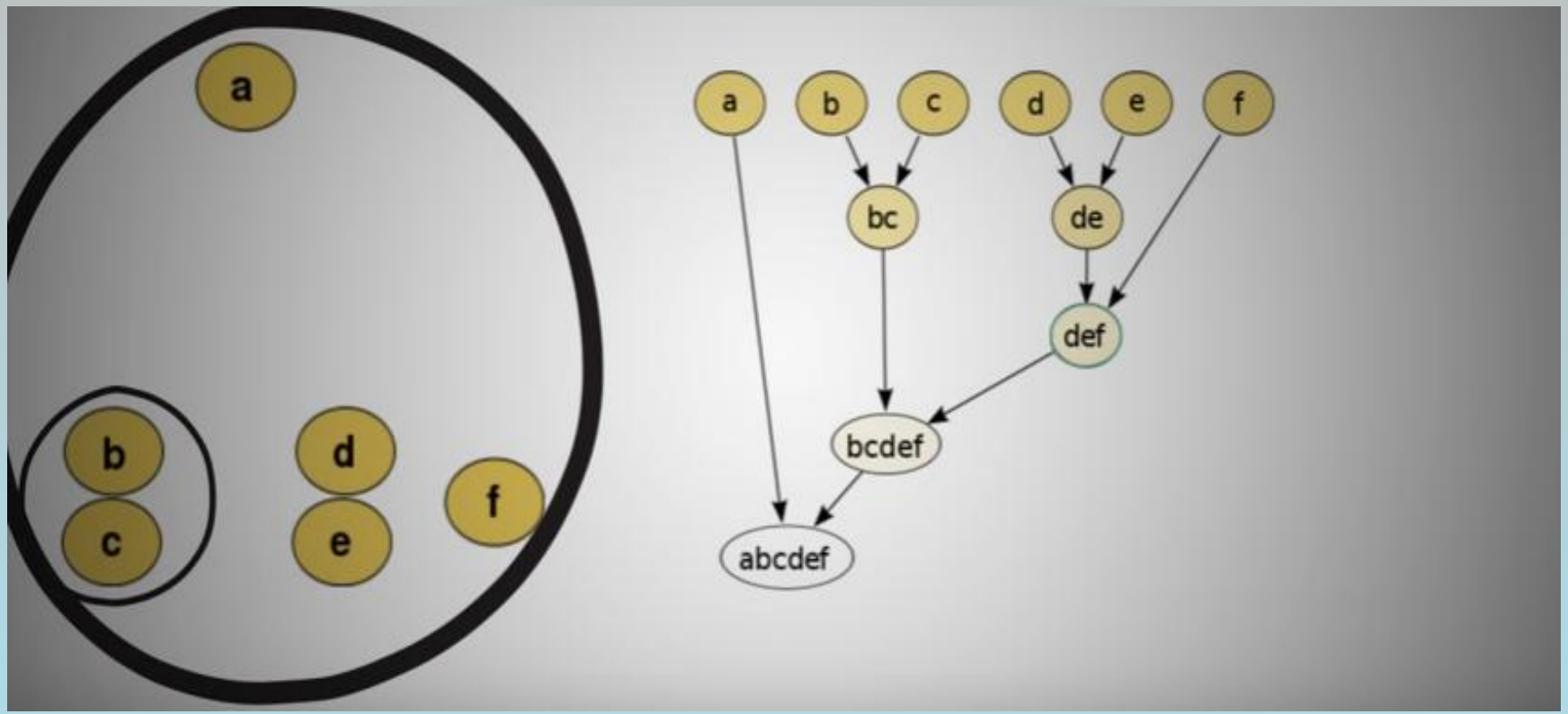
Algorithm:

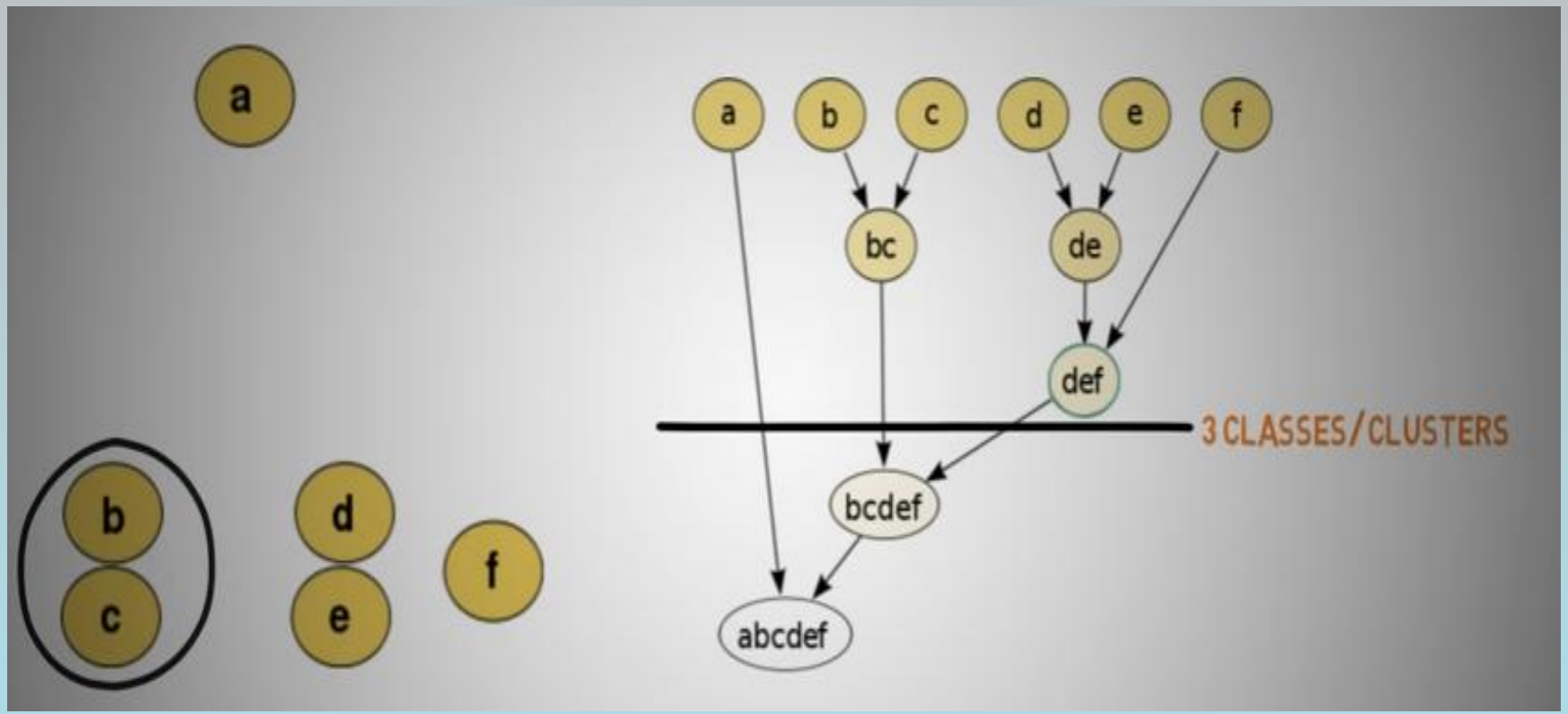
1. At the beginning, each data point forms a cluster (also called a node).
2. Merge nodes/clusters that have the least distance.
3. Go on merging until there is only one cluster left.

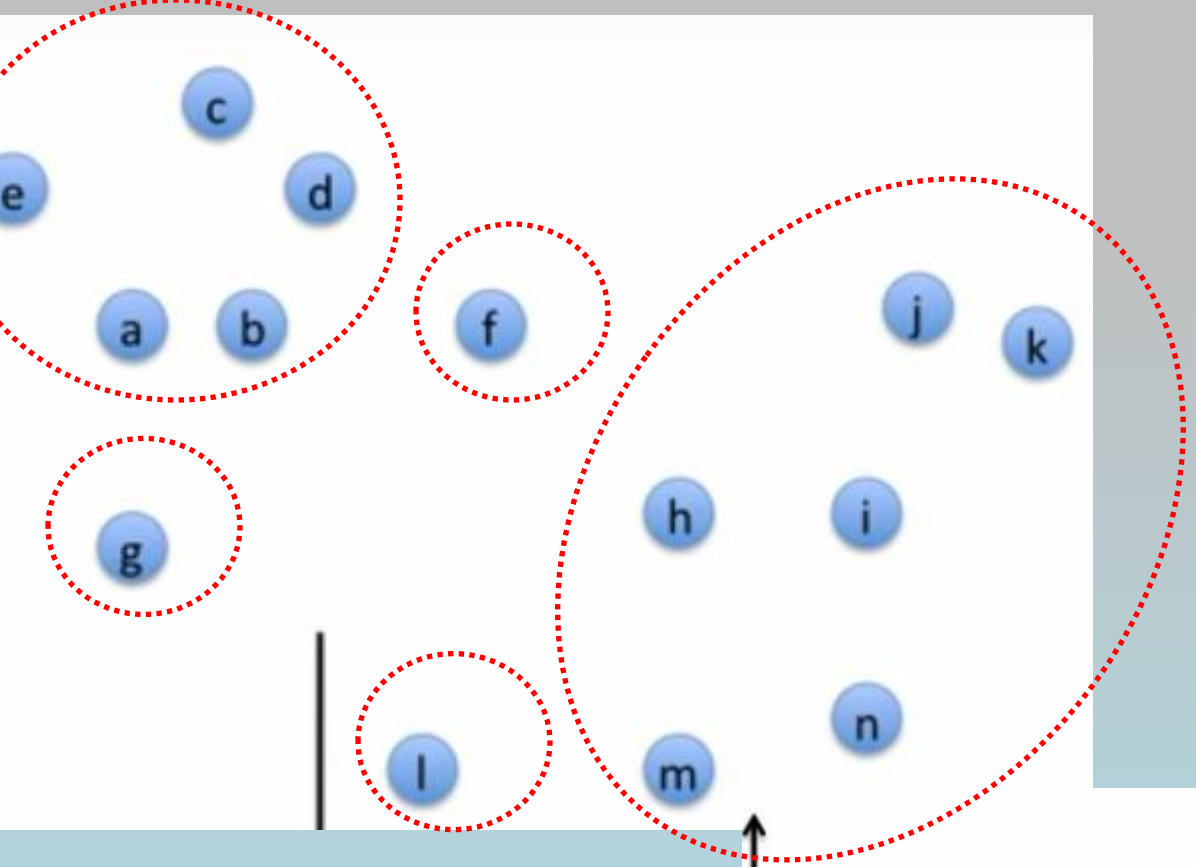

a
b
c
d
e
f

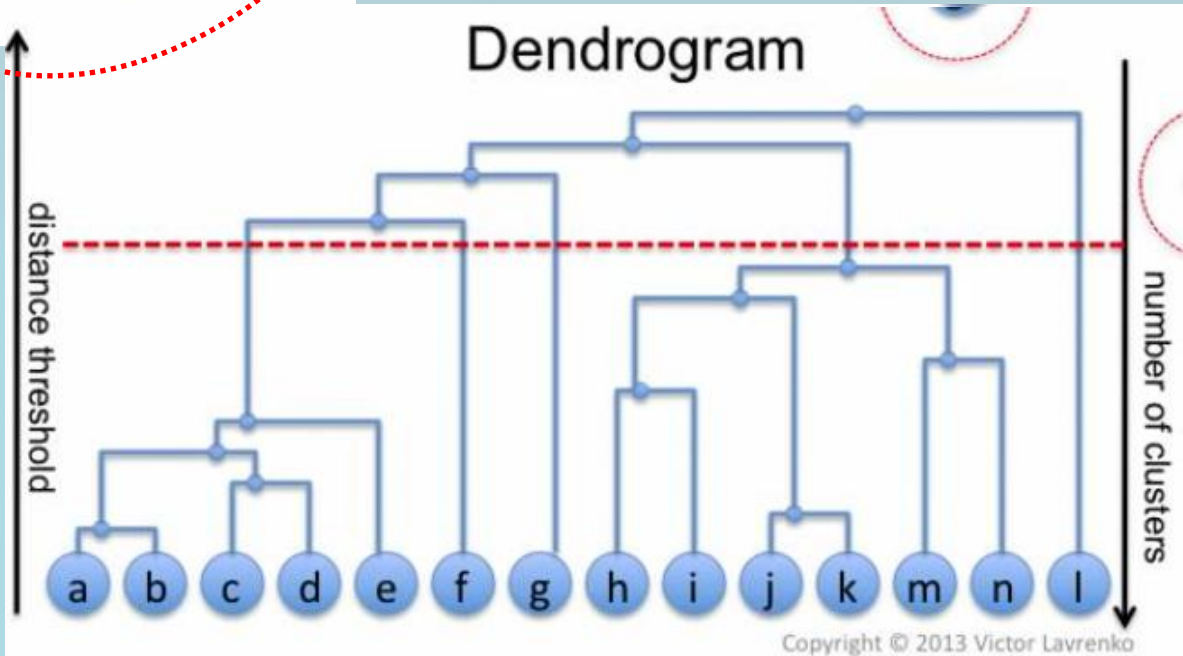

a
b
c
d
e
f

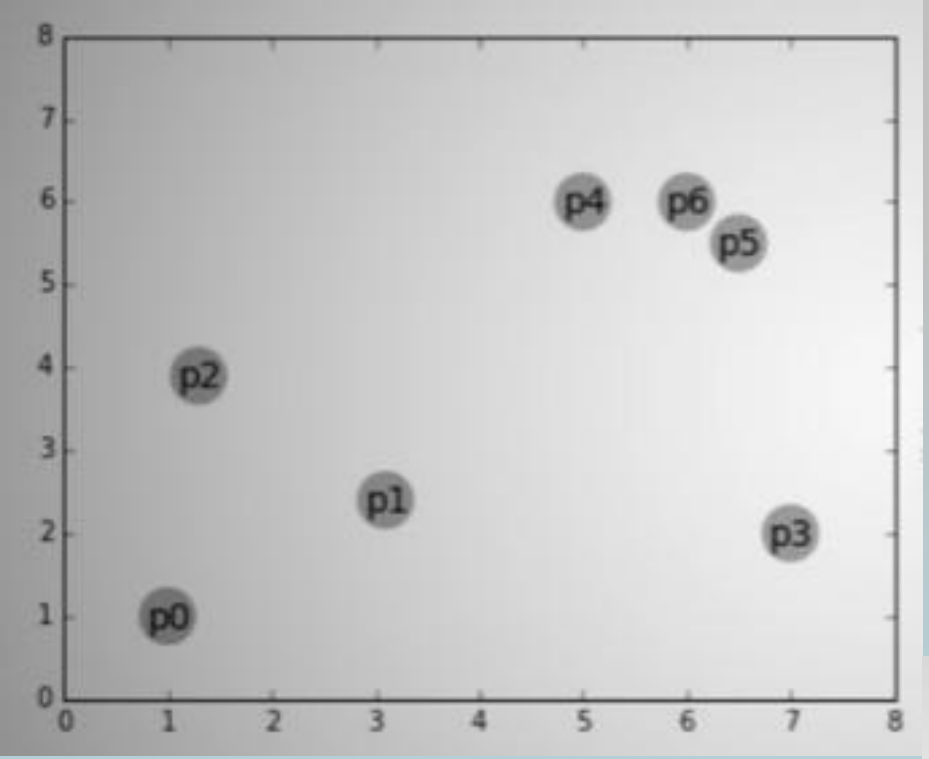

a
b
c
d
e
f
a
b
c
d
e
f
bc
de
def
bcdef
abcdef


a
b
c
d
e
f
a
b
c
d
e
f
bc
de
def
bcdef
abcdef

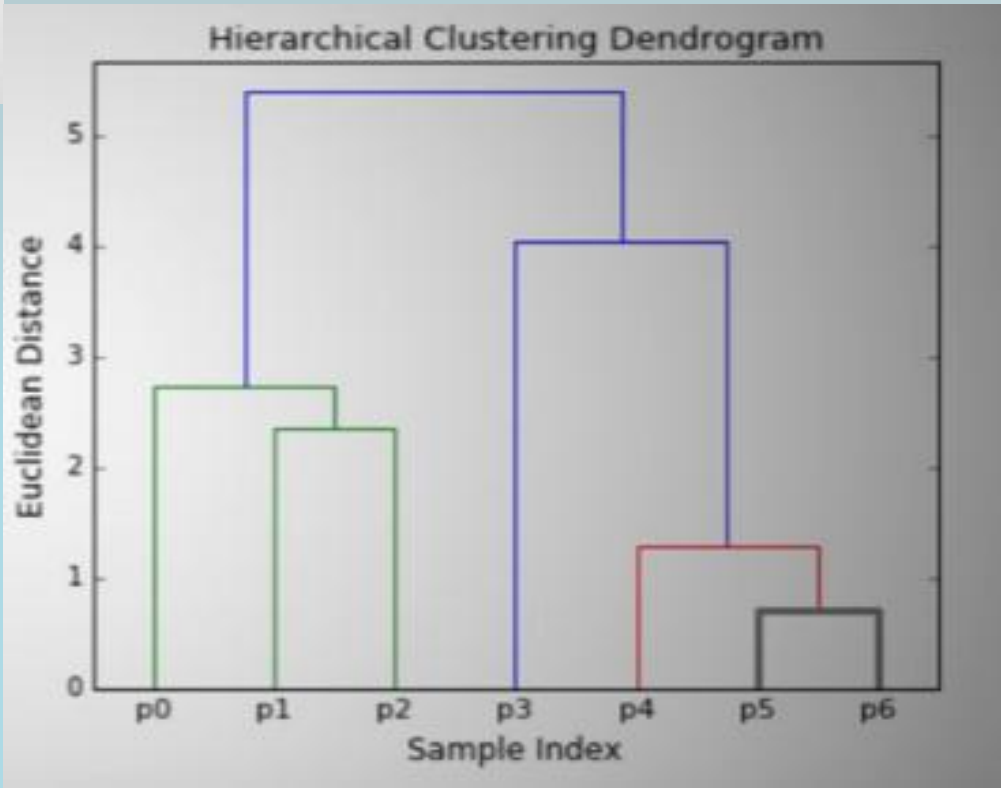

a
b
c
d
e
f
bc
de
def
3 CLASSES/CLUSTERS
bcdef
abcdef
a
b
c
d
e
f


c
e
d
a
b
f
j
k
g
h
i
n
m
l


Dendrogram
distance threshold
number of clusters
a
b
c
d
e
f
g
h
i
j
k
m
n
l
Copyright © 2013 Victor Lavrenko



Hierarchical Dendrogram



Assessing Clustering

Major issue - labeling

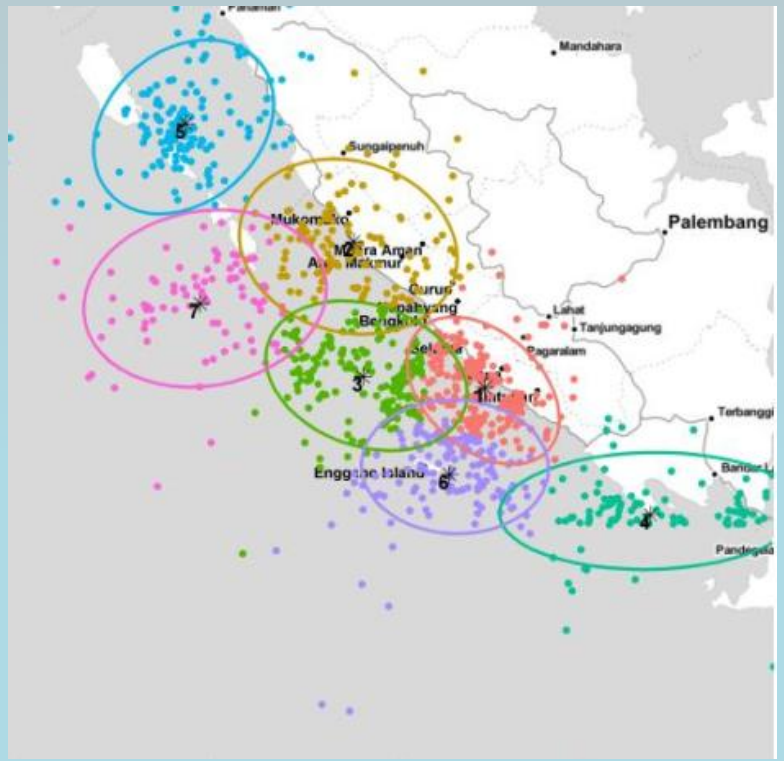
- After clustering algorithm finds clusters - how can they be useful to the end user?

- Need a concise label for each cluster
  - In search results, say “Animal” or “Car” in the ***jaguar*** example.
  - In topic trees (Yahoo), need navigational cues.
    - Often done by hand, a posteriori.

Cluster Evaluation: hard problem

- The quality of a clustering is very hard to evaluate because
  - We do not know the correct clusters

Total= 76,753	Calls (%)	Data (%)	SMS (%)
Cluster 1	38	29	7
Cluster 2	79	20	1
Cluster 3	46	24	30
Cluster 4	31	42	27
Cluster 5	7	76	17
Cluster 6	44	41	15
Cluster 7	3	94	3

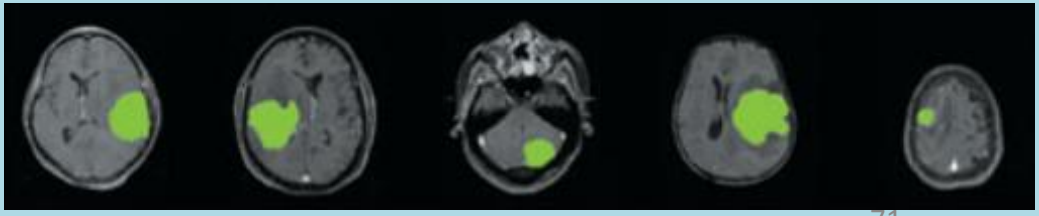


Cluster evaluation: ground truth

- Use of ground truth
- After clustering, a confusion matrix is constructed. From the matrix, we compute various measurements such as precision, recall and F-score.

Input images (original images)

Clustered tumor region



Ground truth for the tumor

Evaluation based on internal information

- **Intra-cluster cohesion** (compactness):
  - Cohesion measures how near the data points in a cluster are to the cluster centroid.
  - Sum of squared error (SSE) is a commonly used measure.
- I**nter-cluster separation** (isolation):
  - Separation means that different cluster centroids should be far away from one another.
- In most applications, expert judgments are still the key.