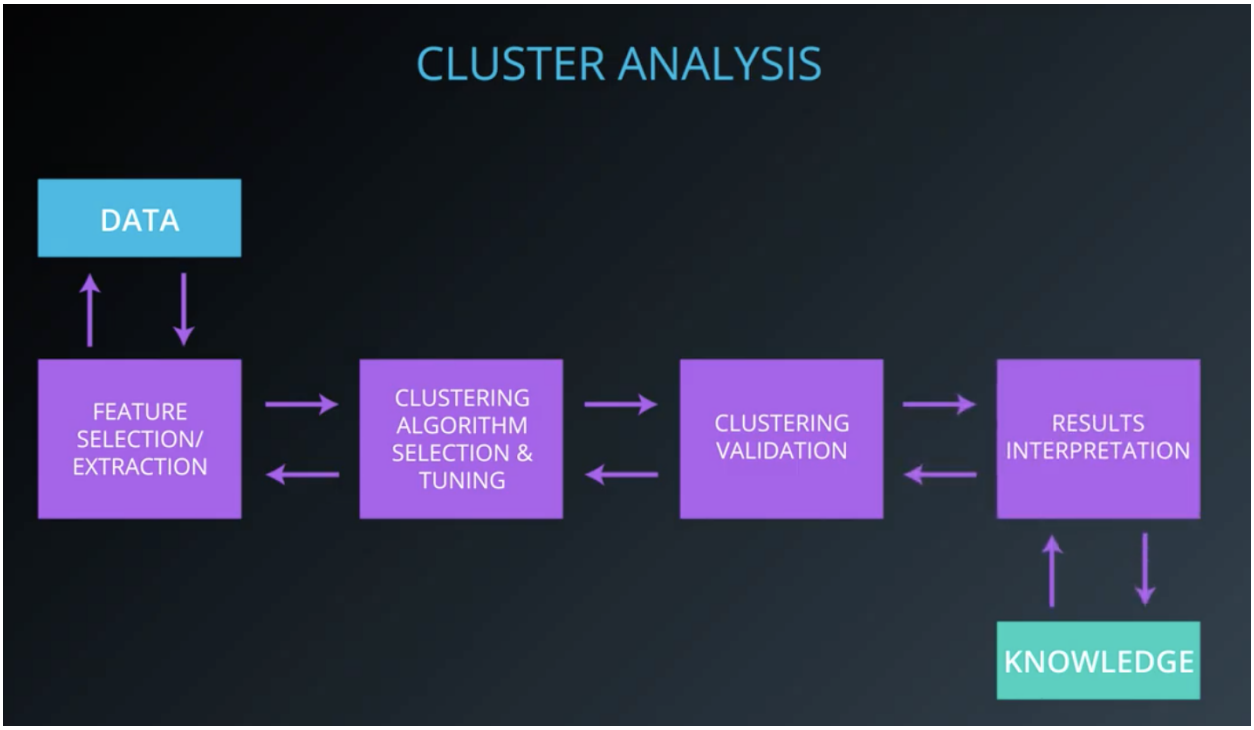
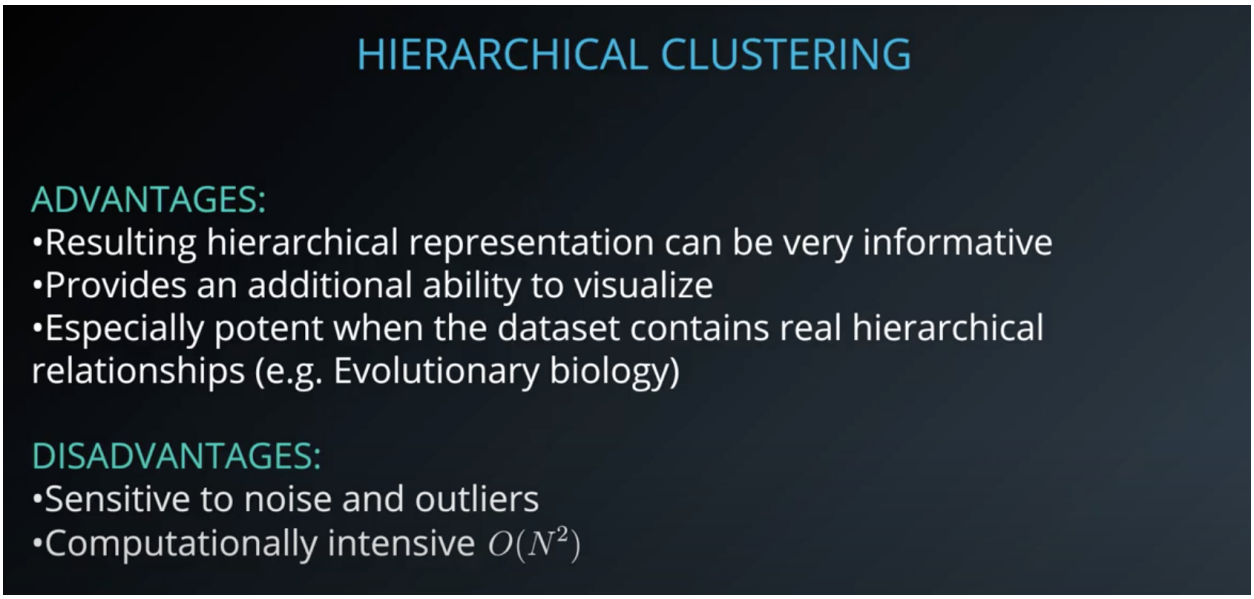
Udacity Nanodegree — Clustering

Process for clustering



K-means Clustering

Hierarchical Clustering



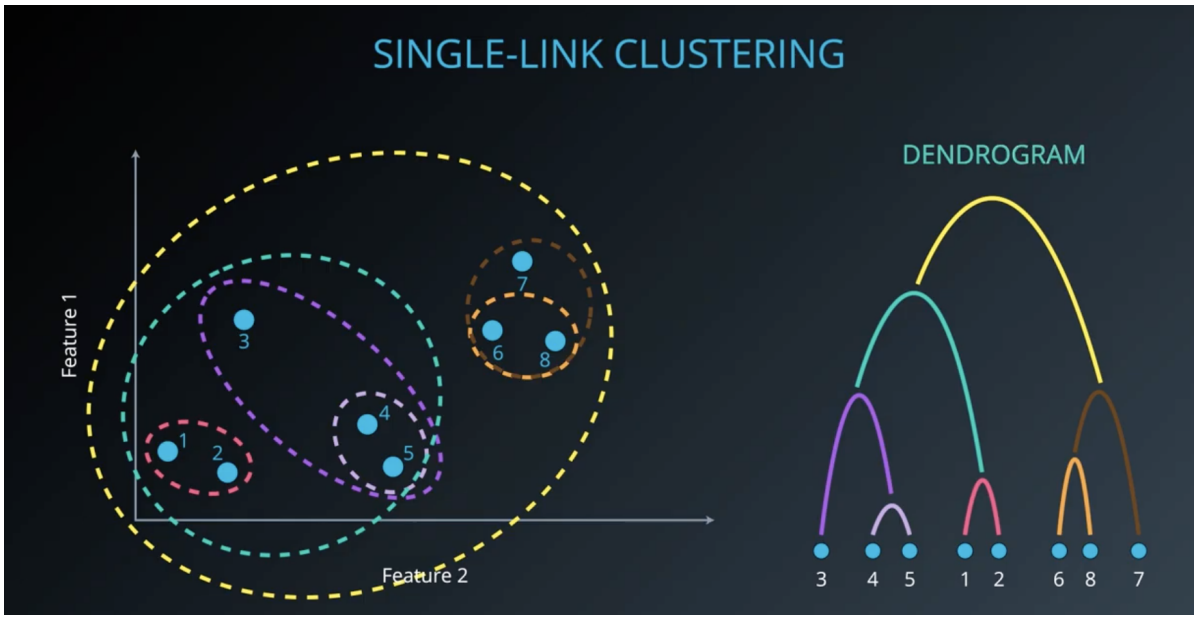
Application

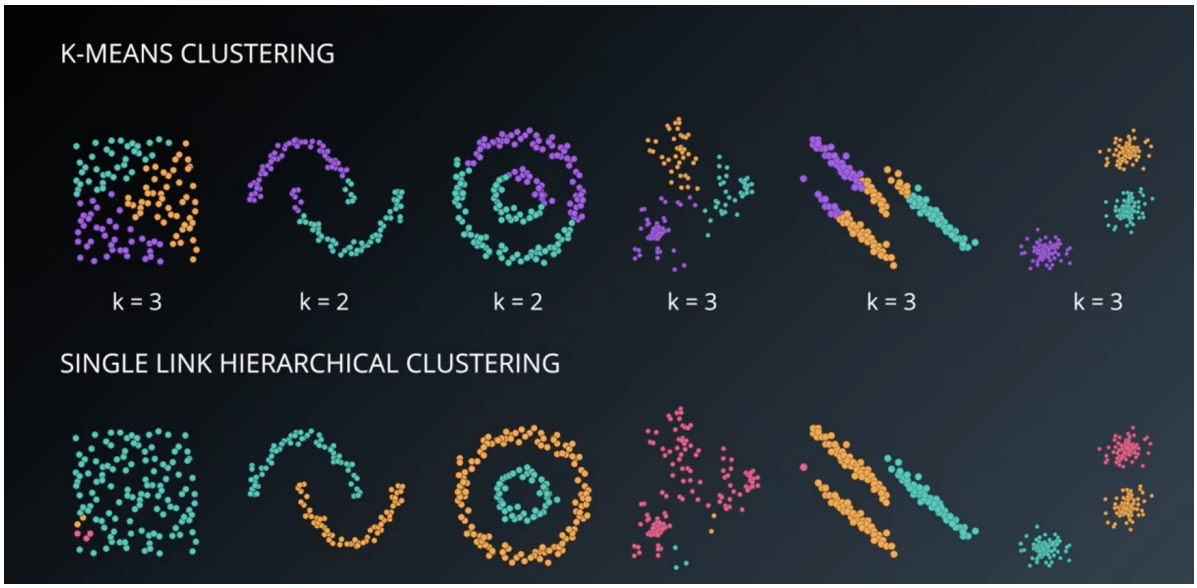
Paper: **Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi**

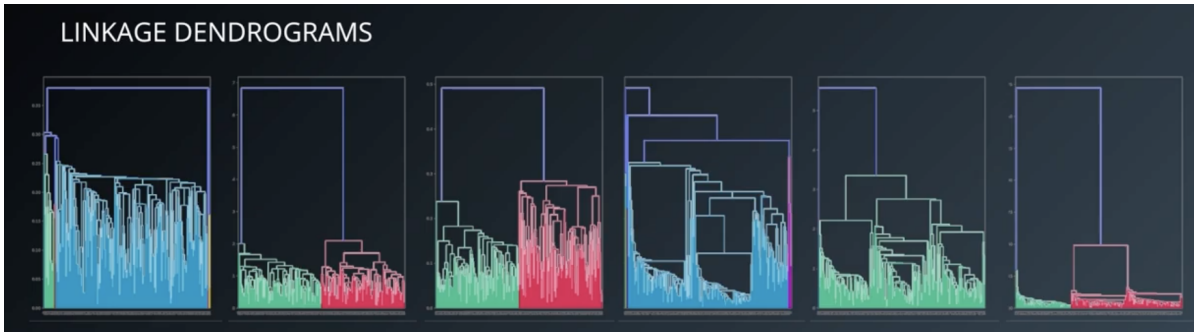
Paper: **Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency**

▼ Hierarchical Clustering - Single Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the closest point between two clusters to determine which cluster to join. e.g. 7 & 6+8, closest to 6, belong to 6+8

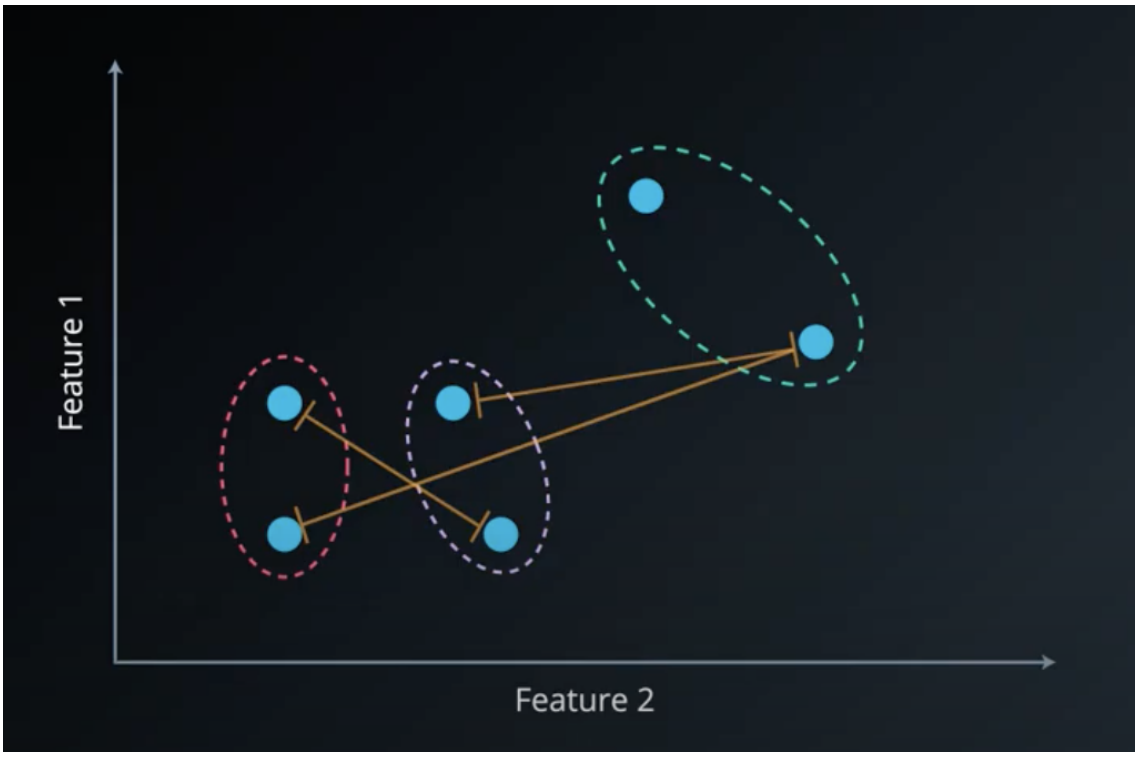






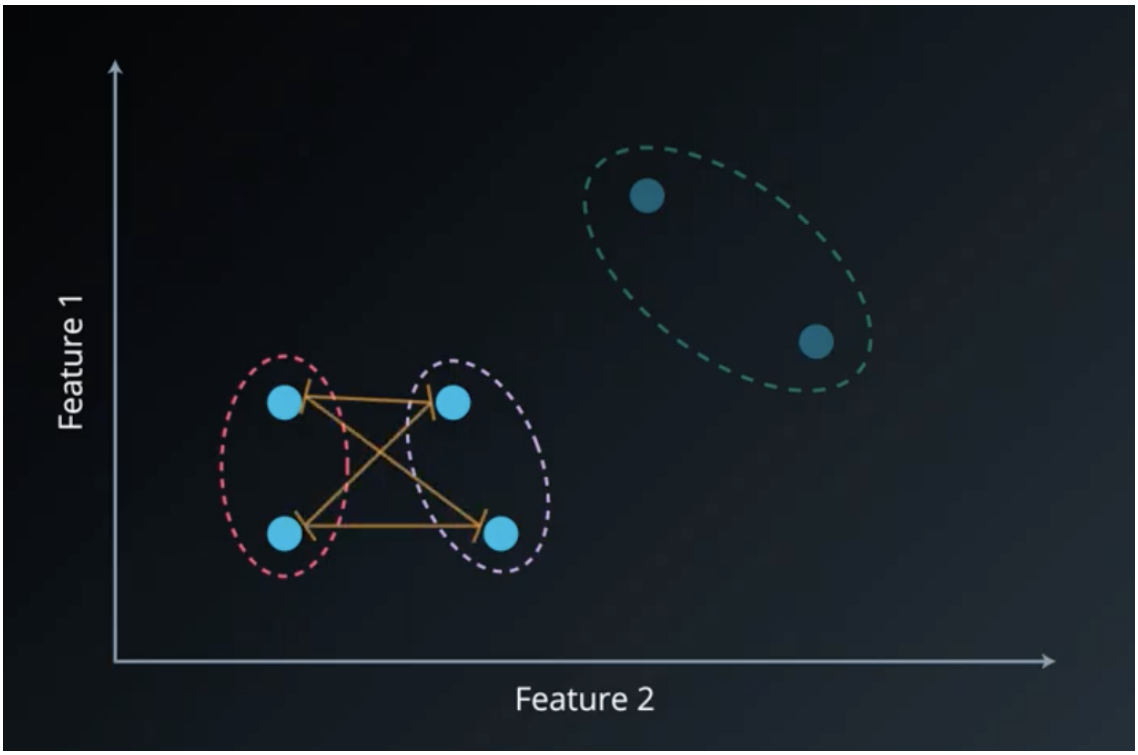
▼ Hierarchical Clustering - Complete Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the farthest point between two clusters to determine which cluster to join. e.g. 7 & 6+8, the farthest is 8, belong to 6+8

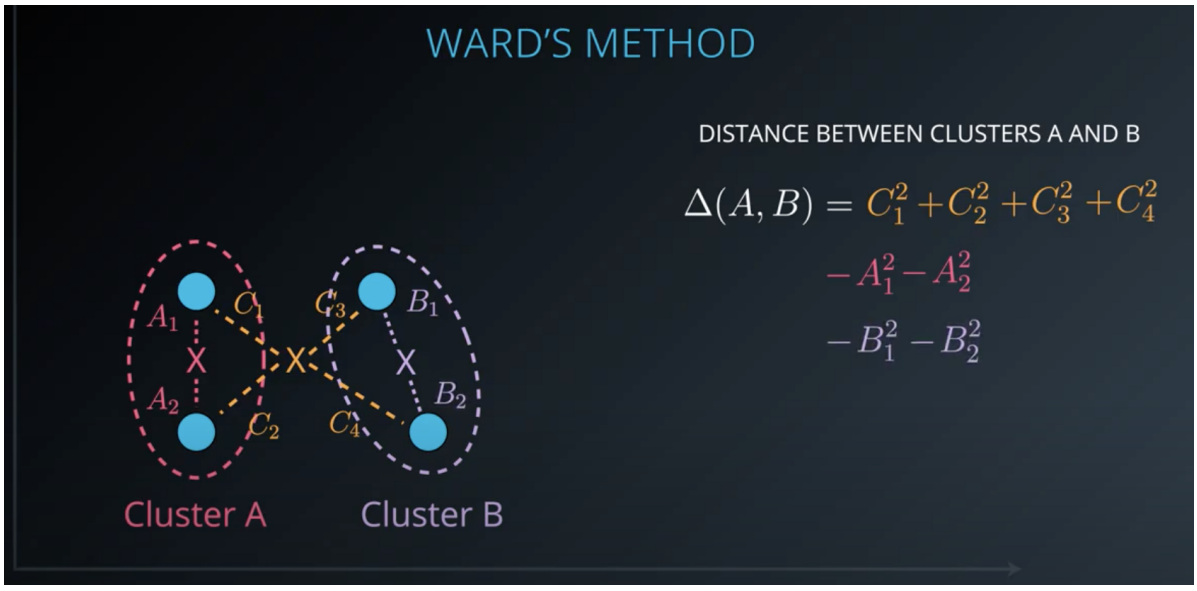


▼ Hierarchical Clustering - Average Link Clustering

1. Look at the closest point for a point(1-2, 4-5)
2. If there is already clusters, find the average distance for every point between two clusters to determine which cluster to join. e.g. 7 & 6+8, the farthest is 8, belong to 6+8

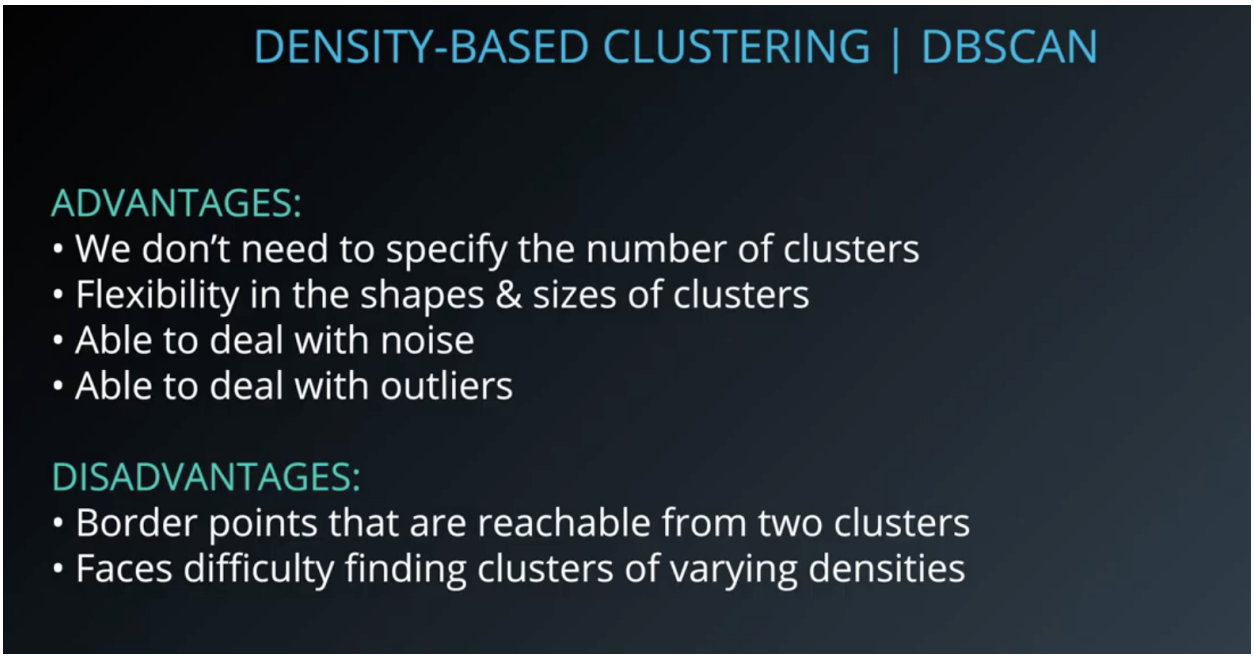


▼ Hierarchical Clustering - Ward's method



Density-based clustering

DBSCAN clusters points based on density(stamps) and in that process it identifies noise.



- For border problem, if a point os reachable from two clusters, the clustering become arbitrary
- For varying densities, we can use HDBSCAN as a alternative

Application

Paper: **Traffic Classification Using Clustering Algorithms**

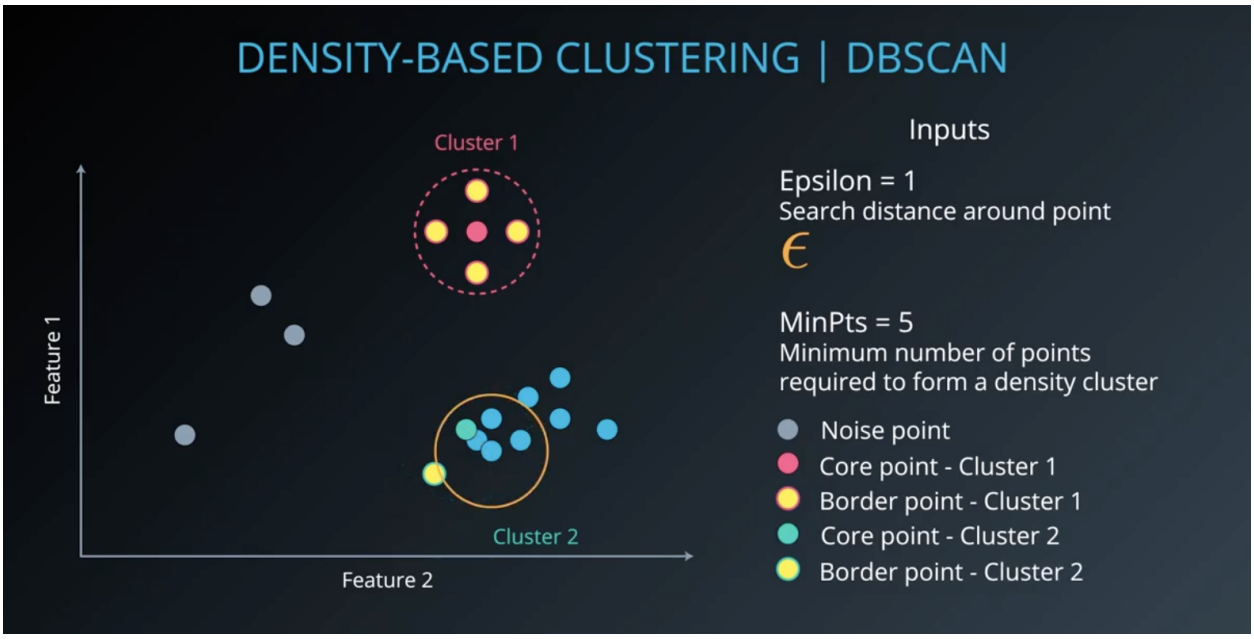
Paper: **Anomaly detection in temperature data using dbscan algorithm**

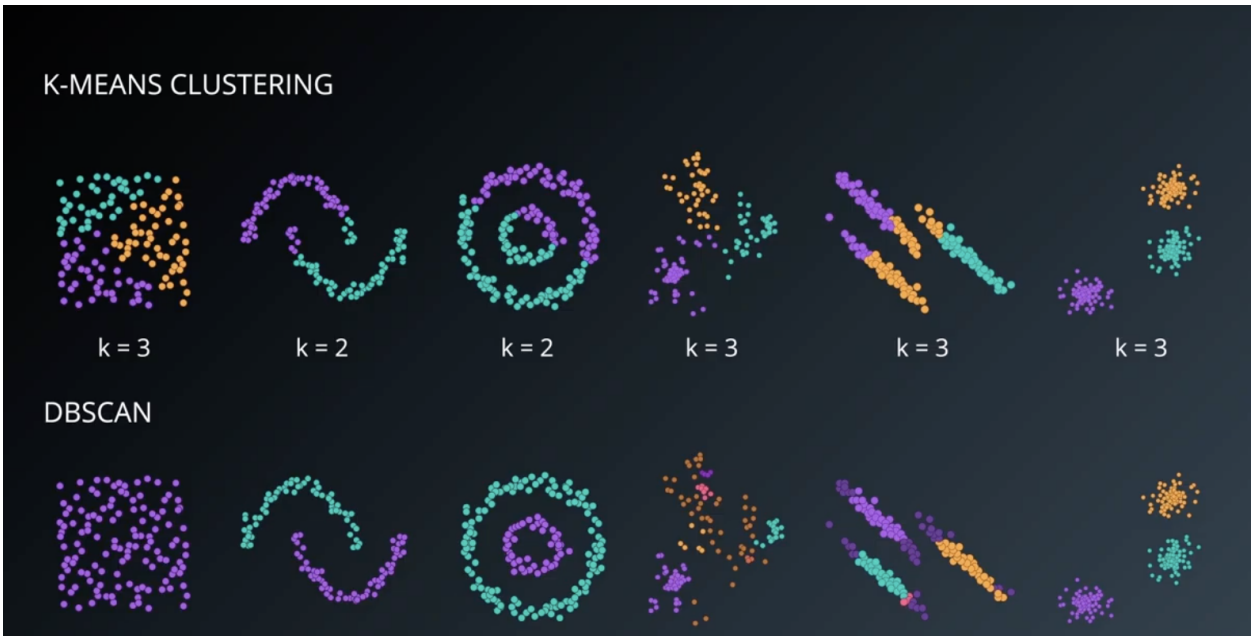
Paper: **Hierarchical density based clustering**

Variables

eps: distance for a cluster(radius from the center)

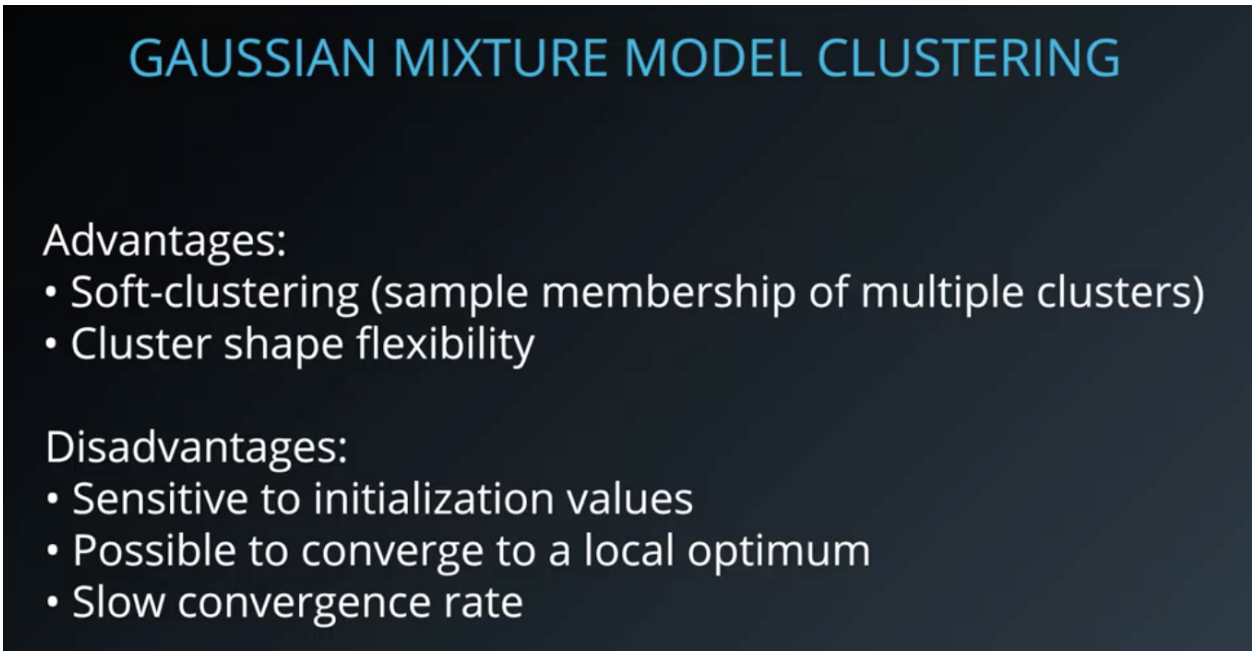
min_samples: min samples need to include to form a cluster





Gaussian Mixture Model Clustering (GMM)

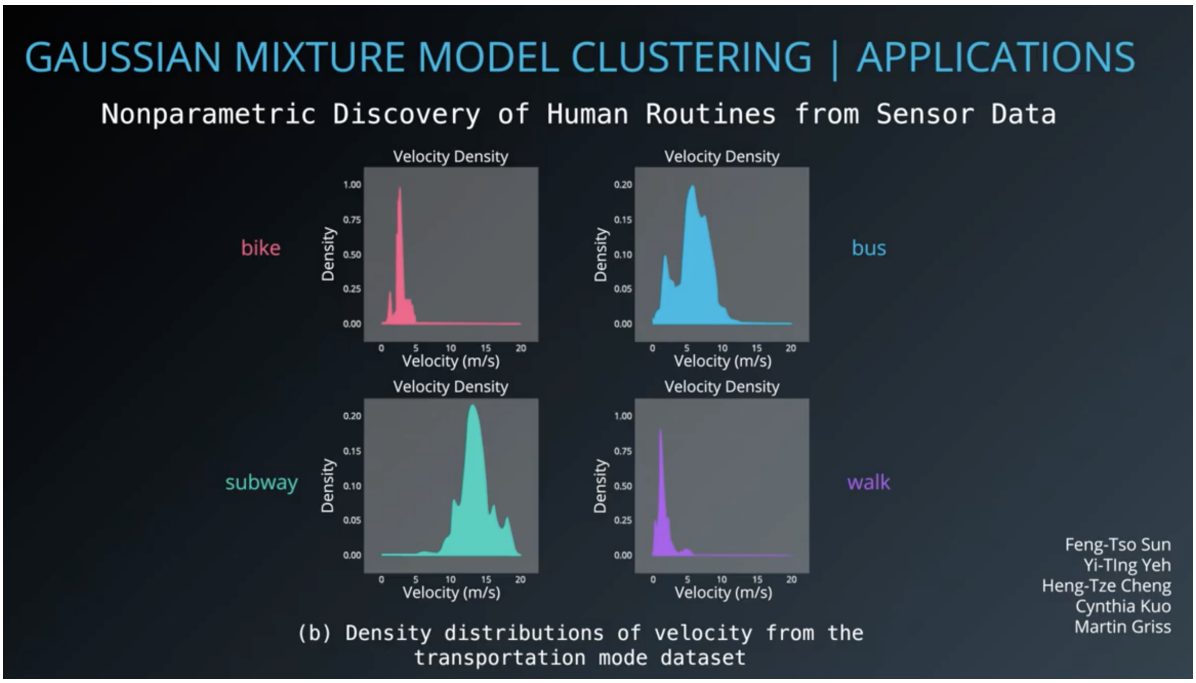
Assume each cluster following a normal distribution



soft clustering: Fuzzy clustering (also referred to as soft clustering or soft k-means) is a form of clustering in which each data point can belong to more than one cluster.

Application

▼ Paper: **Nonparametric discovery of human routines from sensor data** [PDF]



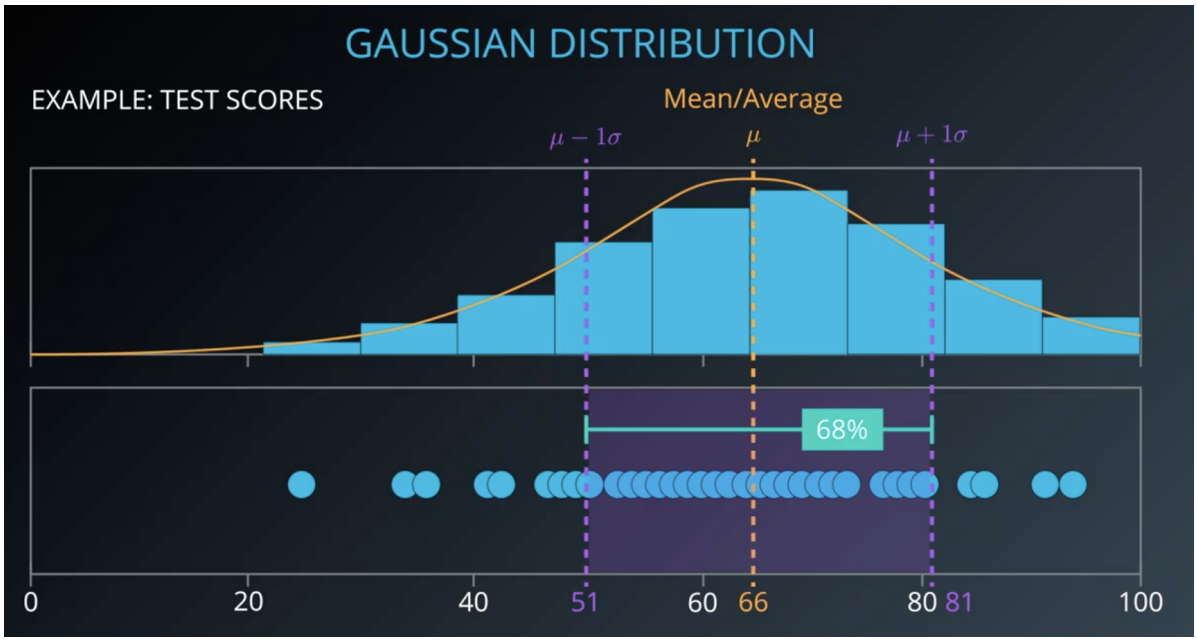
Paper: **Application of the Gaussian mixture model in pulsar astronomy** [PDF]

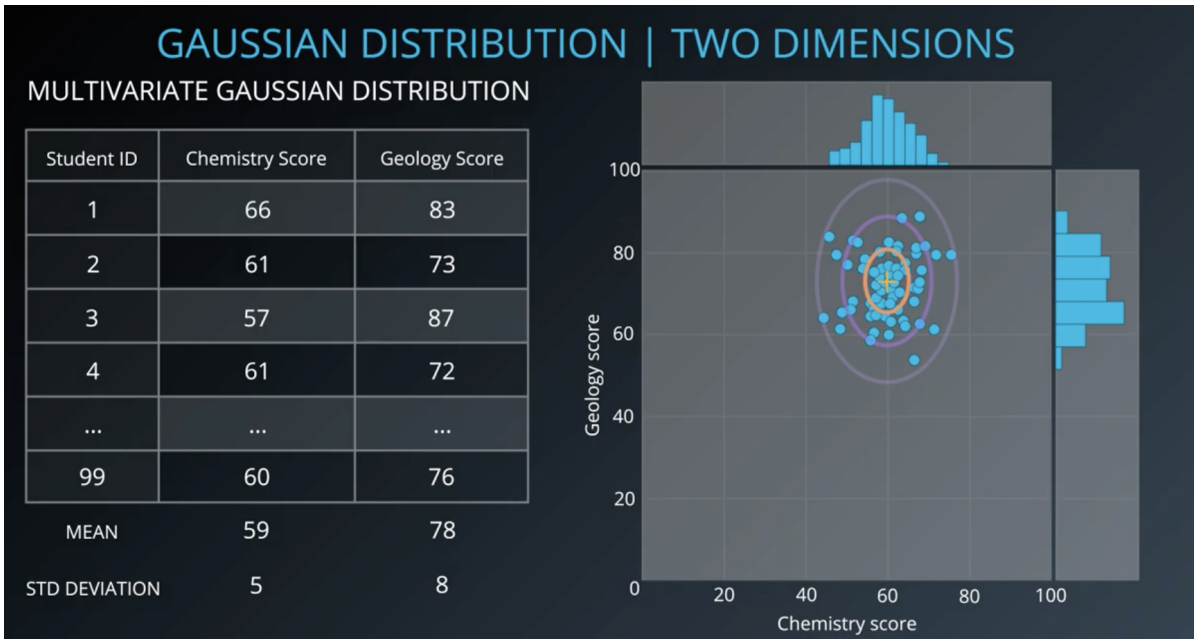
Paper: **Speaker Verification Using Adapted Gaussian Mixture Models** [PDF]

Paper: **Adaptive background mixture models for real-time tracking** [PDF]

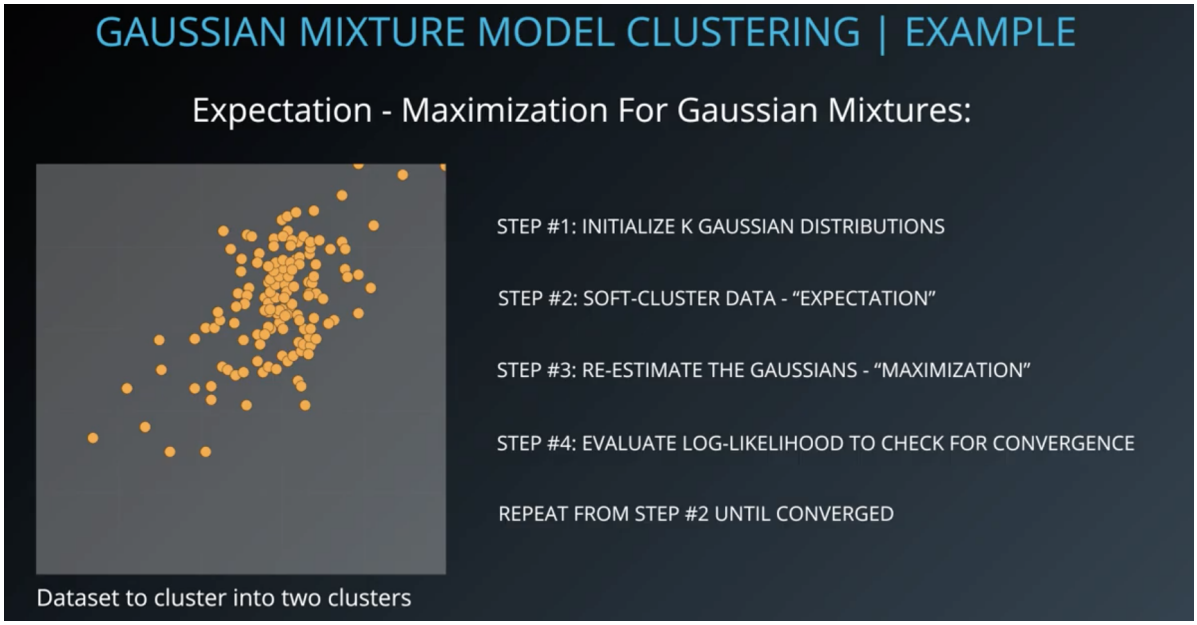
Background subtraction Video: **https://www.youtube.com/watch?v=lLt9H6RFO6A**

▼ Gaussian distribution(Normal distribution)

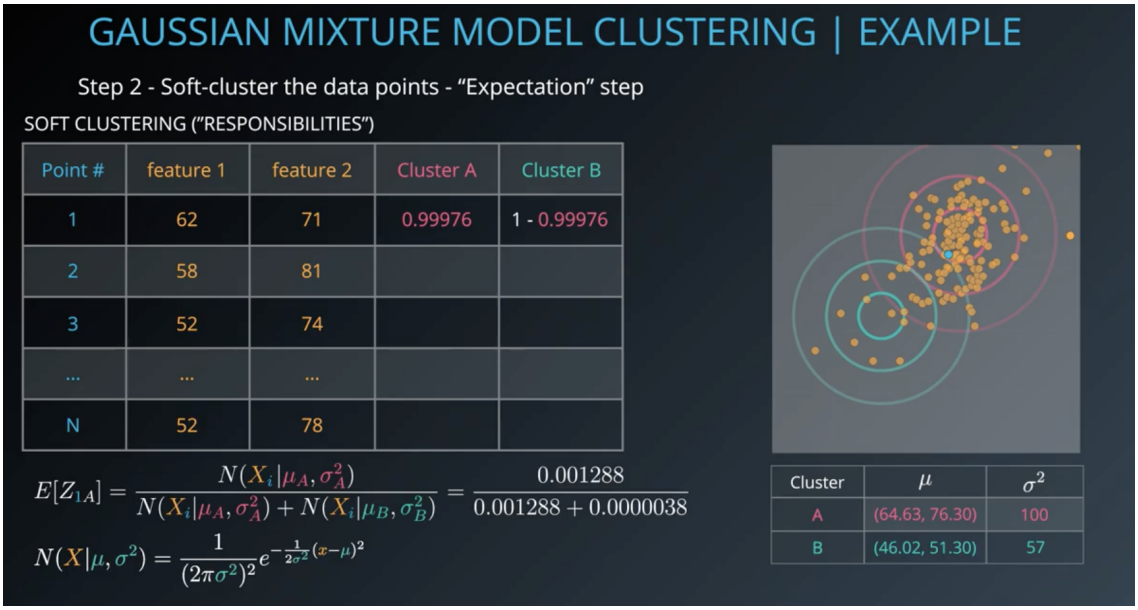




▼ Steps for GMM



1. Step one: two random points for gaussian distribution
2. Using expectation formula to calculate the prob of being in cluster A and B



3. Calculate Cluster A's μ by calculate the weighted average of Cluster A

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - "Maximization" step

NEW GAUSSIAN PARAMETERS

Cluster	new μ	new σ^2
A		
B		

$$\text{new } \mu_A = \frac{\sum_{i=1}^{N} E[Z_{ij}]X_i}{\sum_{i=1}^{N} E[Z_{ij}]} = \frac{0.99937\text{x}(62\ 71) + 0.9998\text{x}(58\ 81) + \ldots}{0.99937 + 0.9998 + 0.55818 + \ldots}$$

Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...	...	...	...	...
N	52	78	0.99133	0.00867

Calculate Cluster A's σ^2

GAUSSIAN MIXTURE MODEL CLUSTERING | EXAMPLE

Step 3 - Re-estimate parameters of Gaussians - "Maximization" step

NEW GAUSSIAN PARAMETERS

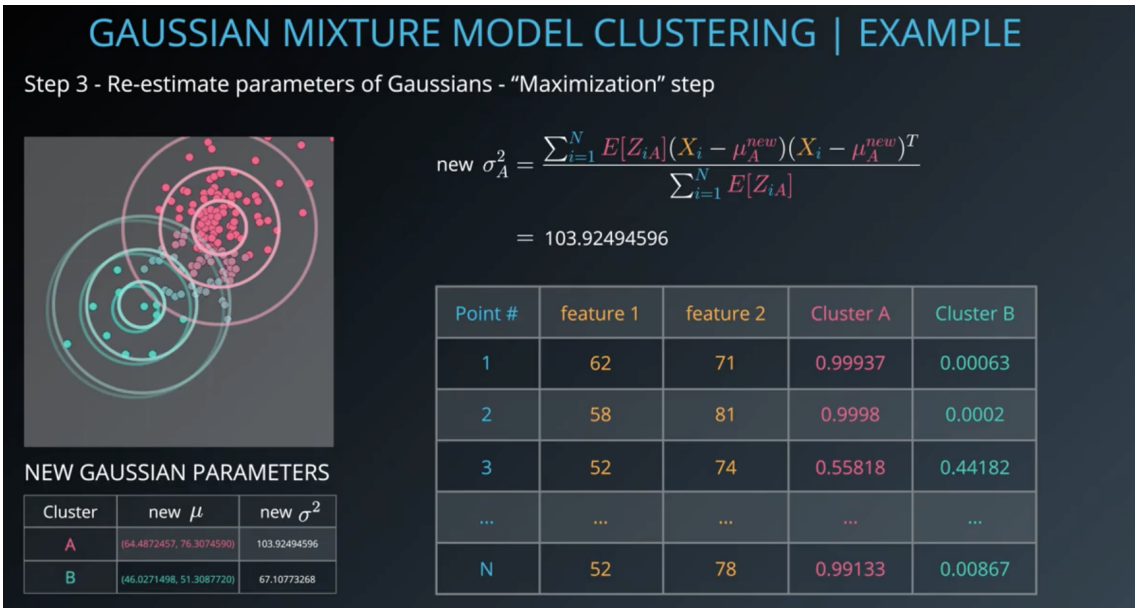
Cluster	new μ	new σ^2
A	(64.4872457, 76.3074590)	103.92494596
B	(46.0271498, 51.3087720)	67.10773268

$$\text{new } \sigma_A^2 = \frac{\sum_{i=1}^{N} E[Z_{iA}](X_i - \mu_A^{new})(X_i - \mu_A^{new})^T}{\sum_{i=1}^{N} E[Z_{iA}]}$$

$$= 103.92494596$$

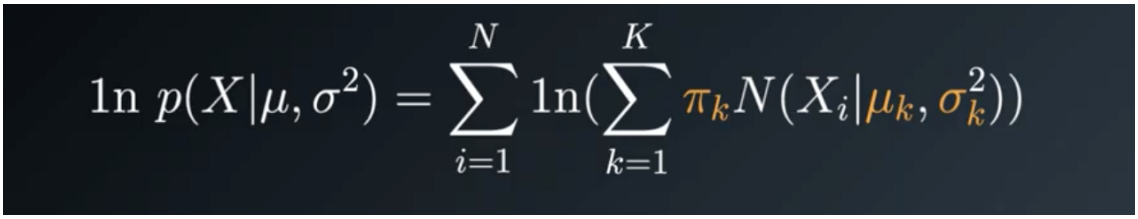
Point #	feature 1	feature 2	Cluster A	Cluster B
1	62	71	0.99937	0.00063
2	58	81	0.9998	0.0002
3	52	74	0.55818	0.44182
...	...	...	...	...
N	52	78	0.99133	0.00867

Move the center and variance



4. Step 4: Evaluate log-likelihood: max this likelihood formula

   Choose good mixing coefficient(π_k), μ_k, σ_k^2



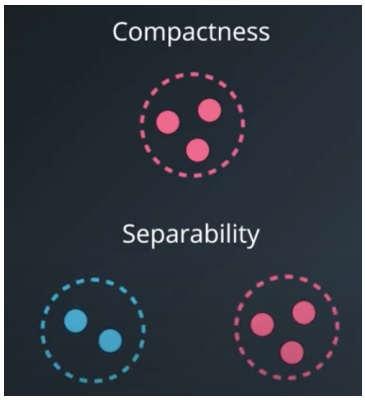
Clustering Validation

- external indices: we have labels
- internal indices: no labels
- relative indices:

Validation indices include

- compactness: how close the points in each cluster are to each other

- separability: how far distinct clusters are from each other



silhouette: for circular data

DBSCAN validation:

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=83C3BD5E078B1444CB26E243975507E1?doi=10.1.1.707.9034&rep=rep1&type=pdf

▼ Questions

what is the difference between standardized and normalized?