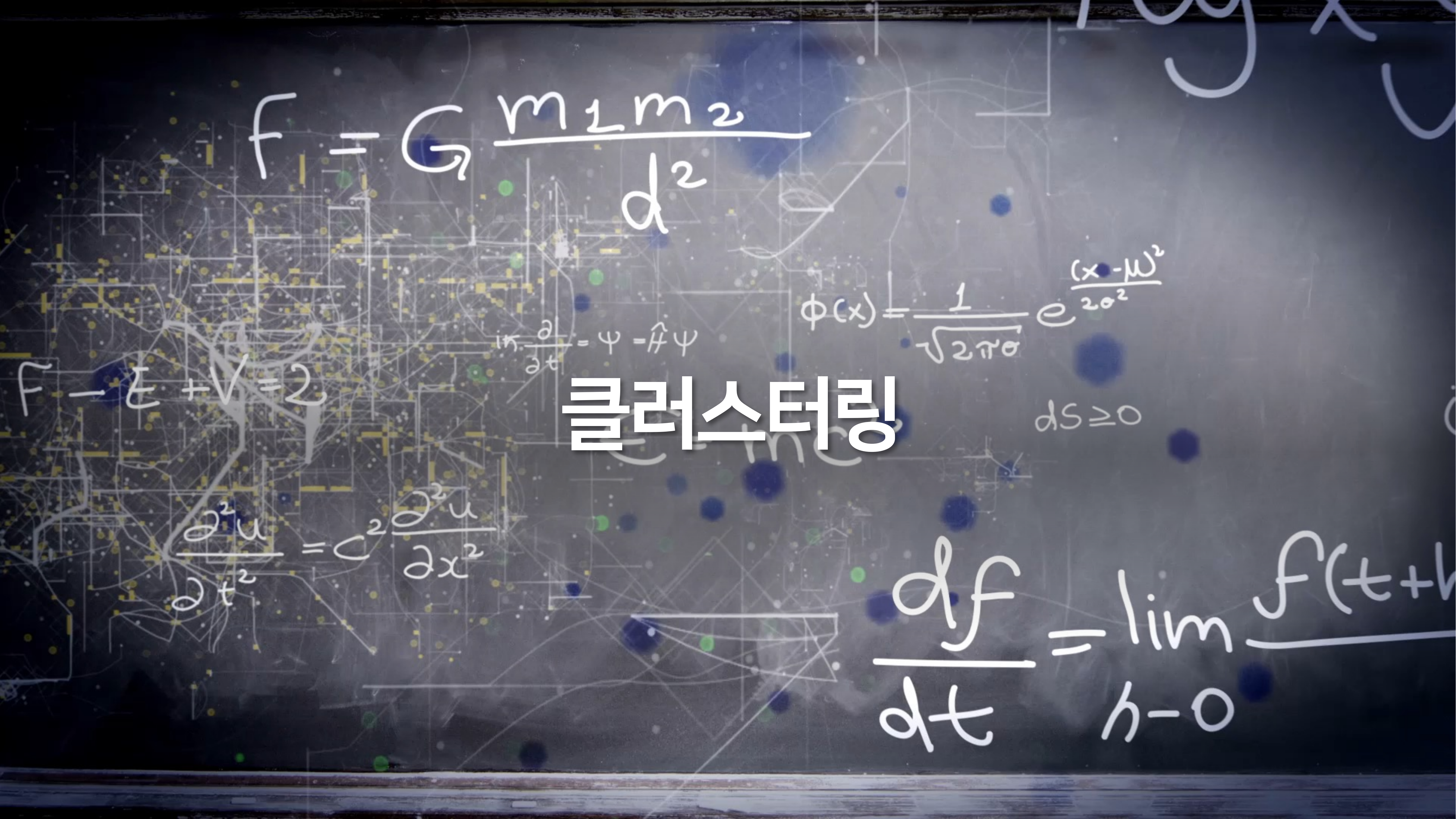
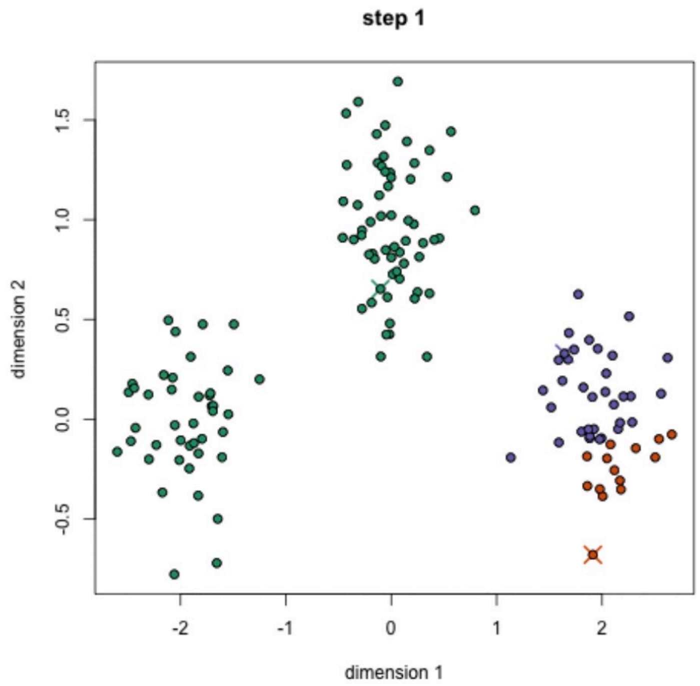


클러스터링

클러스터링 이란

- 데이터 포인트의 그룹화와 관련된 머신러닝 기술
- 데이터 포인트 집합이 주어지면 클러스터링 알고리즘을 사용하여 각 데이터 포인트를 특정 그룹으로 분류 가능
- 이론적으로 같은 그룹에 속한 데이터 요소는 비슷한 속성 및 / 또는 피처를 가져야 하지만 다른 그룹의 데이터 요소는 매우 다른 속성 및 / 또는 피처를 가지기도 함
- 클러스터링은 비지도 학습의 한 방법.

1. K-Means Clustering



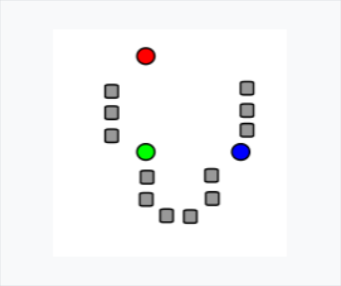
K-Means는 우리가 실제로 수행하는 모든 작업이 포인트와 그룹 중앙 사이의 거리를 계산하므로 매우 빠름 →
매우 적은 계산량으로 선형 복잡도 O (n)을 갖음

Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm

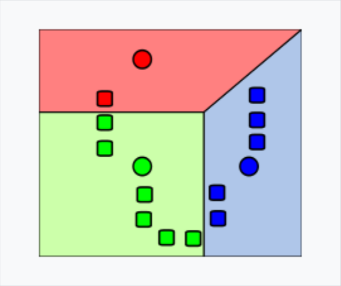
https://ieeexplore.ieee.org/document/5453745

1. Steps in K-means clustering

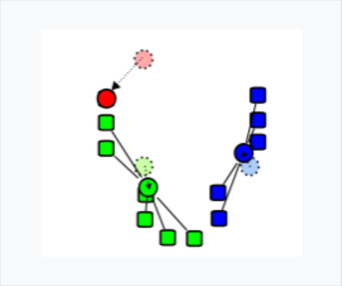
Demonstration of the standard algorithm



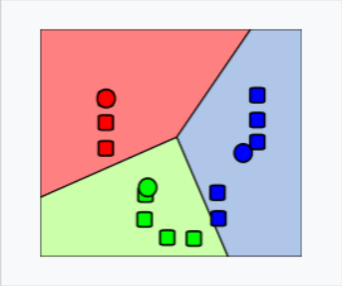
1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).



2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the k clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance.[8] (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.)

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \le \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \le j \le k \right\},$$

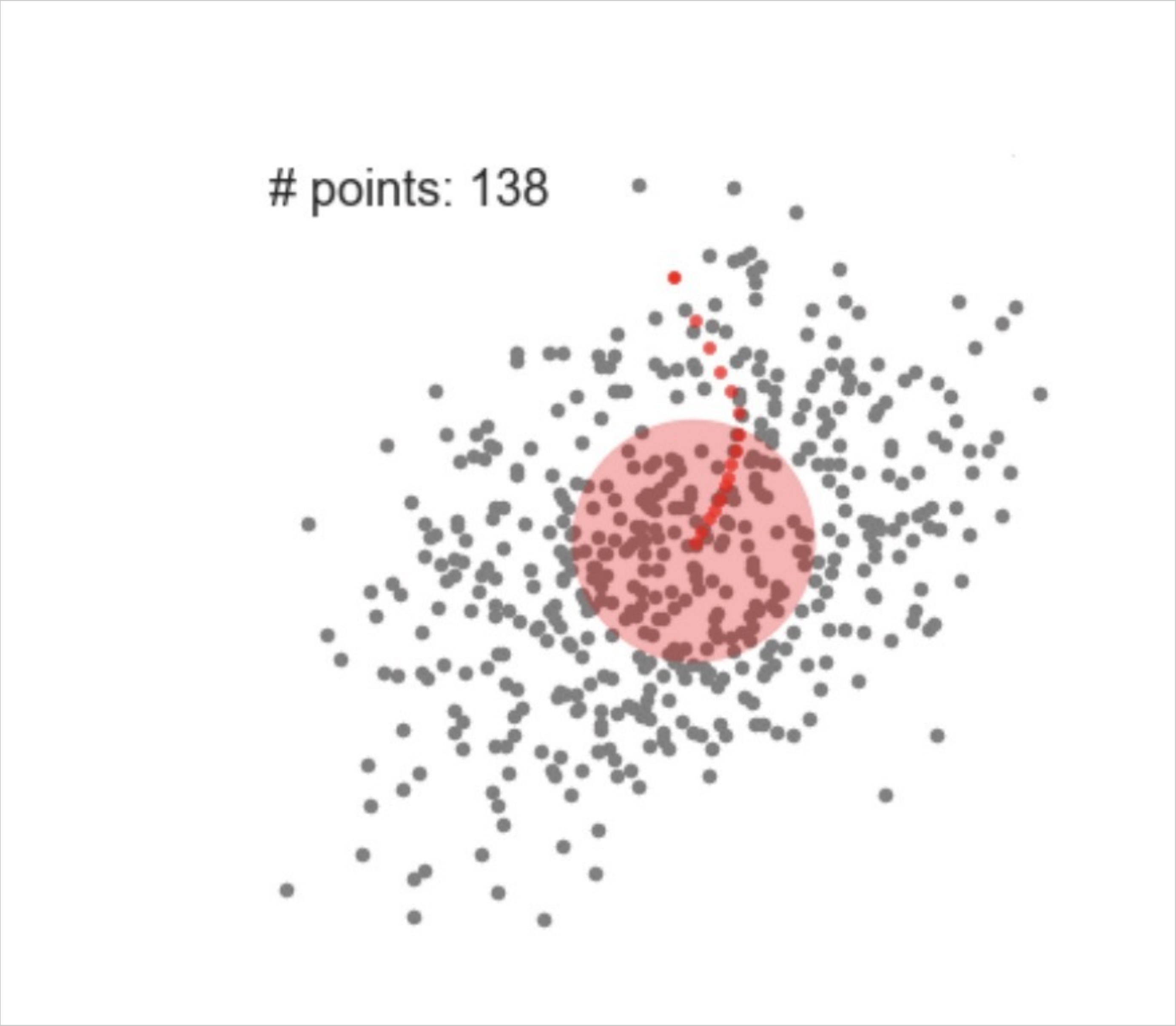
where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

Update step: Recalculate means (centroids) for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

2. Mean-Shift Clustering

- **데이터 포인트의 밀집된 영역을 찾기 위해 시도하는 슬라이딩 윈도우 기반 알고리즘.**
- 중심점에 대한 후보를 슬라이딩 윈도우 내의 포인트의 평균으로 업데이트하여 작동하는 각 그룹 / 클래스의 중심점을 찾는 것이 목표= centroid 기반 알고리즘
- **후보 윈도우는 후 처리 단계에서 필터링되어 거의 중복을 제거하여 최종 세트의 중심점과 해당 그룹을 형성.**



2. Steps in Mean-Shift Clustering

- 평균 이동이 자동으로 이를 감지하여, 클러스터 수를 선택할 필요가 없음.
- 윈도우 사이즈 설정 어려움 (not trivial)

Mean shift is a procedure for locating the maxima—the modes—of a density function given discrete data sampled from that function.[1] This is an iterative method, and we start with an initial estimate x. Let a kernel function $K(x_i - x)$ be given. This function determines the weight of nearby points for re-estimation of the mean. Typically a Gaussian kernel on the distance to the current estimate is used, $K(x_i - x) = e^{-c||x_i - x||^2}$. The weighted mean of the density in the window determined by K is

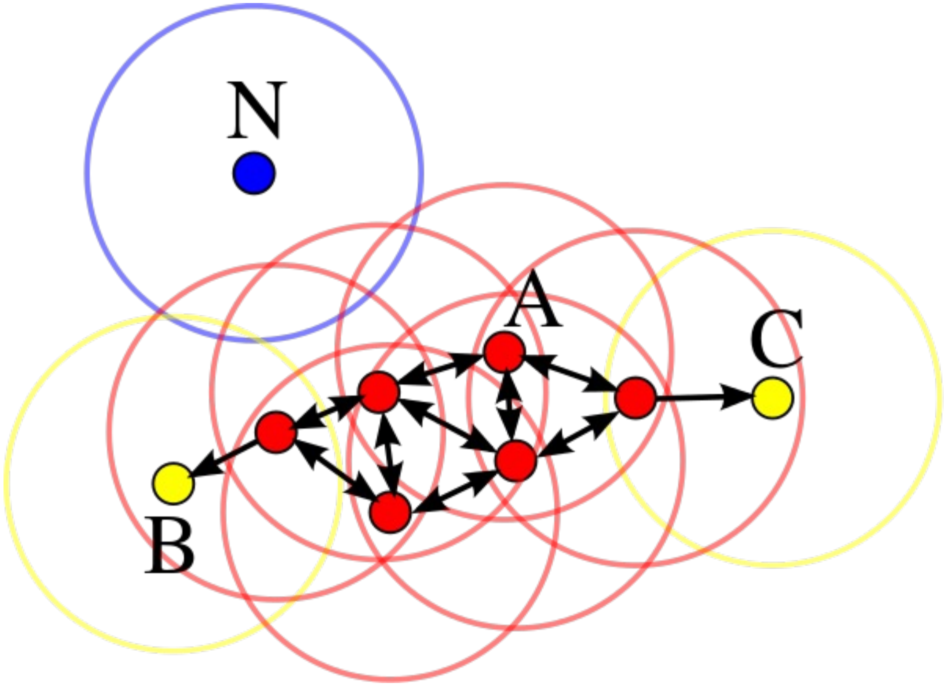
$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

where $N(x)$ is the neighborhood of x, a set of points for which $K(x_i - x) \neq 0$.

The difference $m(x) - x$ is called *mean shift* in Fukunaga and Hostetler.[3] The *mean-shift algorithm* now sets $x \leftarrow m(x)$, and repeats the estimation until $m(x)$ converges.

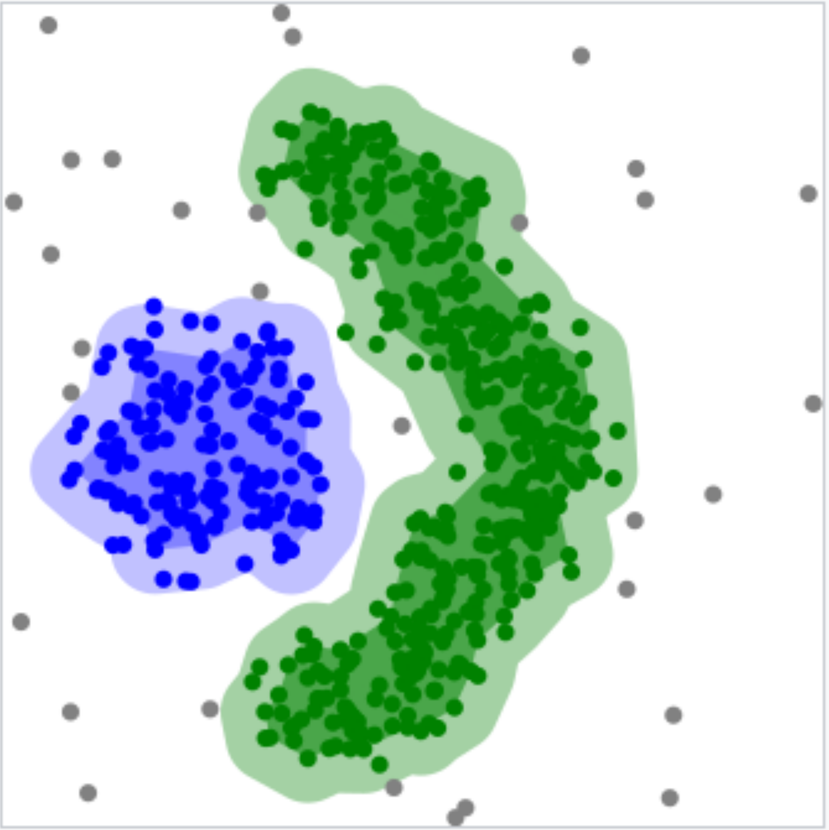
3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

- A point p is a *core point* if at least minPts points are within distance ε of it (including p).
- A point q is *directly reachable* from p if point q is within distance ε from core point p. Points are only said to be directly reachable from core points.
- A point q is *reachable* from p if there is a path $p_1, \ldots, p_n$ with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i. Note that this implies that the initial point and all points on the path must be core points, with the possible exception of q.
- All points not reachable from any other point are *outliers* or *noise points*.
- Now if p is a core point, then it forms a *cluster* together with all points (core or non-core) that are reachable from it. Each cluster contains at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.



3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

- 클러스터 집합을 전혀 필요로 하지 않음.
- 데이터 포인트가 매우 다르더라도 이상 치를 노이즈로 설정가능
- 임의로 크기가 정해지고 임의로 모양이 지정된 클러스터를 매우 잘 찾을 수 있음
- 클러스터의 밀도가 다양 할 때 잘 수행되지 않음



DBSCAN can find non-linearly separable clusters. This dataset cannot be adequately clustered with k-means or Gaussian Mixture EM clustering.

4. Gaussian Mixture Models (GMM)을 사용한 Expectation-Maximization (EM) 클러스터링

- In statistics, EM (expectation maximization) algorithm handles latent variables, while GMM is the Gaussian mixture model.

1. (E-step) For each i, j, set

$$w_j^{(i)} := p\left(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma\right)$$

2. (M-step) Update the parameters

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$

$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

5. (Agglomerative) Hierarchical Clustering

- seeks to build a hierarchy of clusters, usually presented in a dendrogram.

- Agglomerative: This is a "bottom-up" approach: Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

- Divisive: This is a "top-down" approach: All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.