Machine Learning

K-Means

Machine Learning

비지도학습(unsupervised)

1.데이터에서 숨겨진 패턴이나 고유 구조체 탐색

2.클러스터링(clustering)

-탐색적 데이터 분석을 통해 데이터에서 숨겨진 패턴과 그룹을 찾는다.

-데이터 집합을 클러스터cluster라는 그룹으로 분할

-어느 클러스터에 속할지 예측

-k-means, hierarchical 클러스터링, Gaussian 혼합 모델, Hidden Markov 모델, 퍼지C-means 클러스터링

Machine Learning

비지도학습으로 가능한 문제

1. 블로그 글의 주제 구분

-사전에 어떤 주제인지 알지 못함

-얼마나 많은 주제가 있는지 모름

2.고객들의 취향이 비슷한 그룹으로 묶기

-쇼핑사이트 : 부모 ,독서 광,게이머

-어떤 그룹이 있는지 알 수 없음

-얼마나 많은 그룹이 있는지 알 수 없음

3.비정상적 웹사이트 접근 탐지

-웹 트래픽만 관찰 가능

-어떤 것이 정상이고 비정상인지 알지 못함

K-Means 단계

1. Data준비
2. 얼마나 많은 클러스터를 만들지 결정

   ex)100명고객티셔츠(s,m,l사이즈)

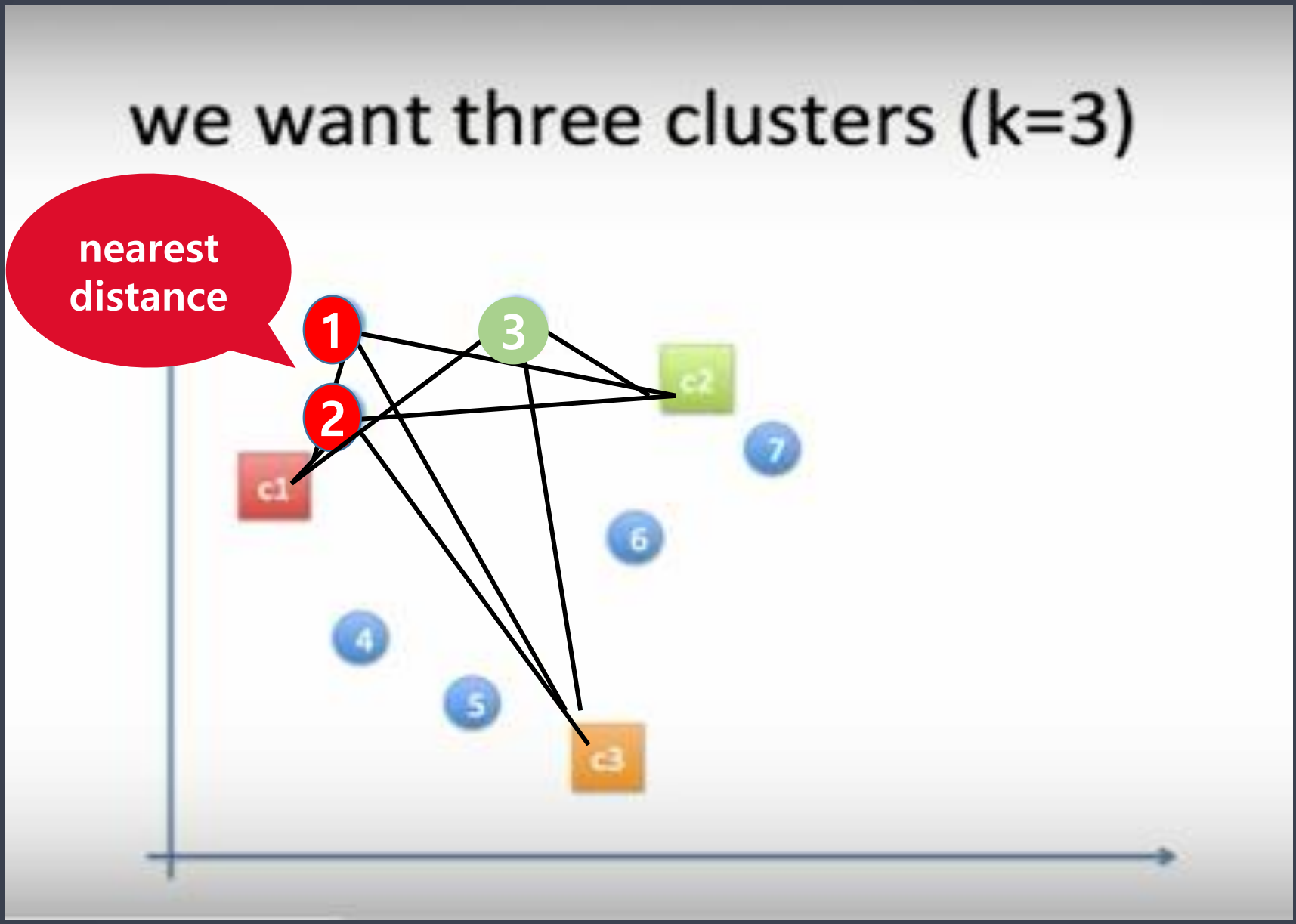
3. 클러스터의 초기중심설정

   - ✓ 랜덤으로 중심점 설정
   - ✓ 수동으로 중심점 설정
   - ✓ Kmean++

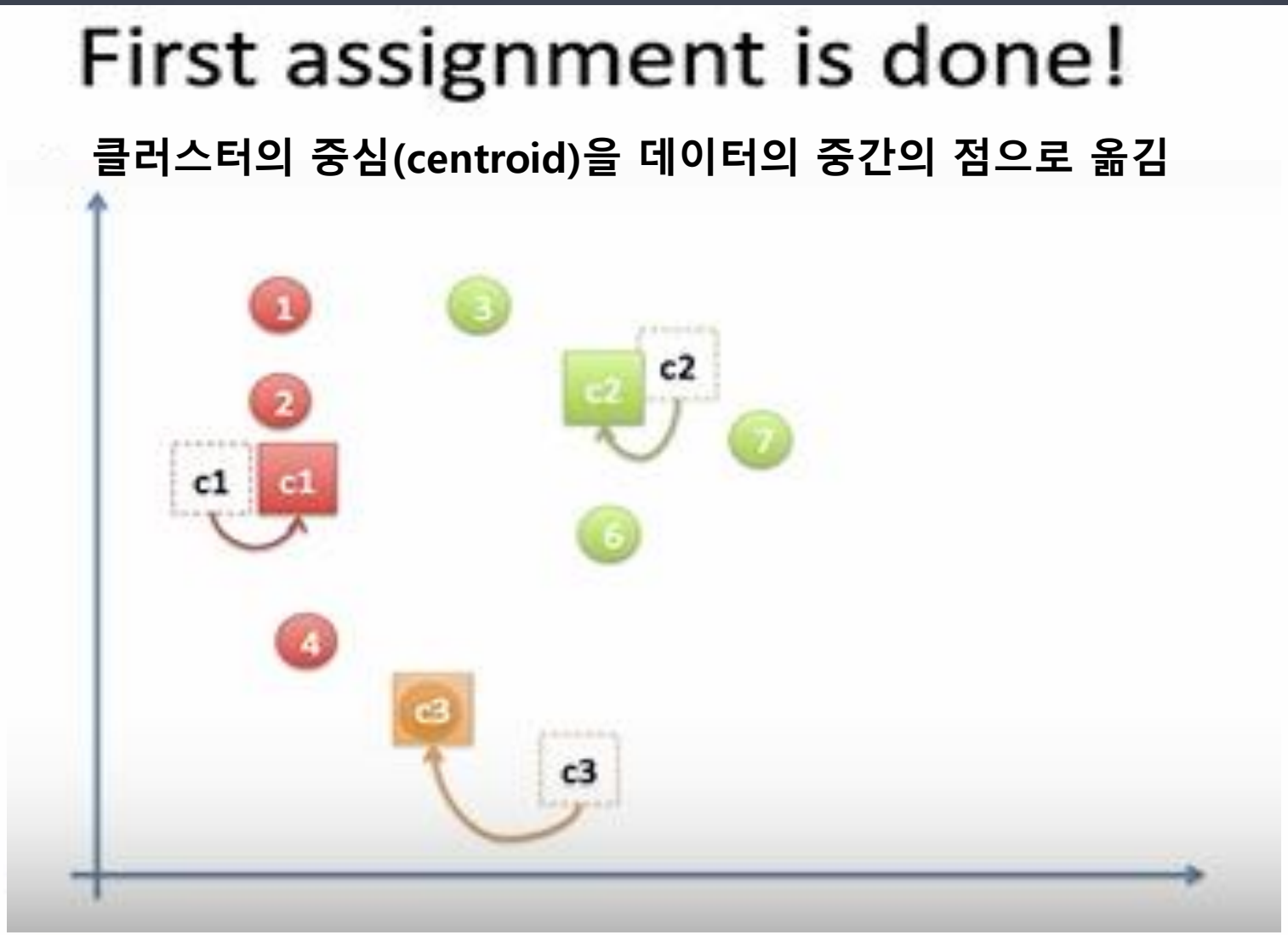
K-Means 단계

1. Data준비
2. 얼마나 많은 클러스터를 만들지 결정
3. 클러스터의 초기중심(centroid)설정
4. 가장가까운 클러스터로 데이터 포인트 할당
5. 클러스터의 중심(centroid)을 데이터의 중간의 점으로 옮김
6. 4번과 5번 단계 반복:중심이 바뀌지 않을 때 까지 반복

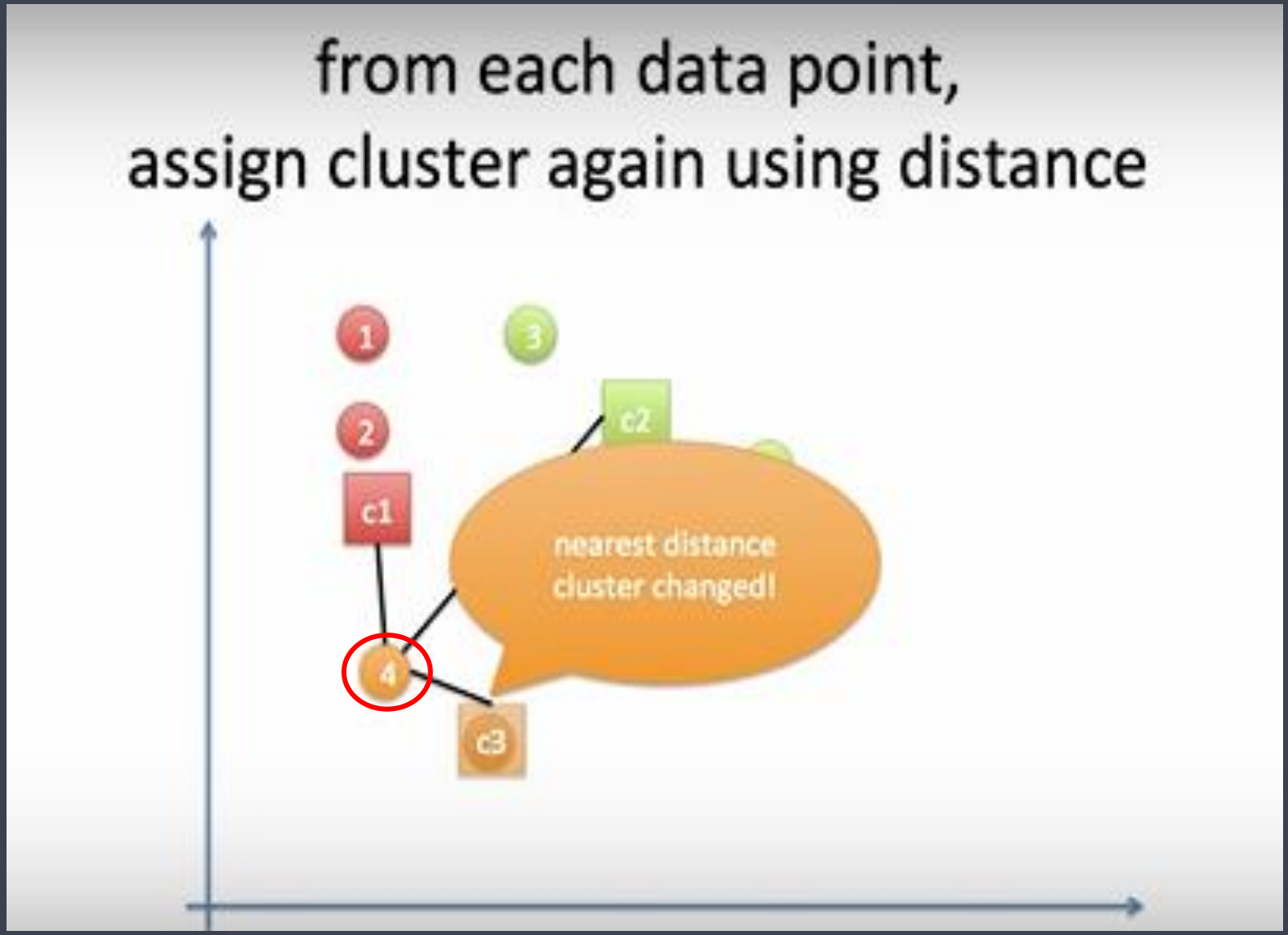
K-Means



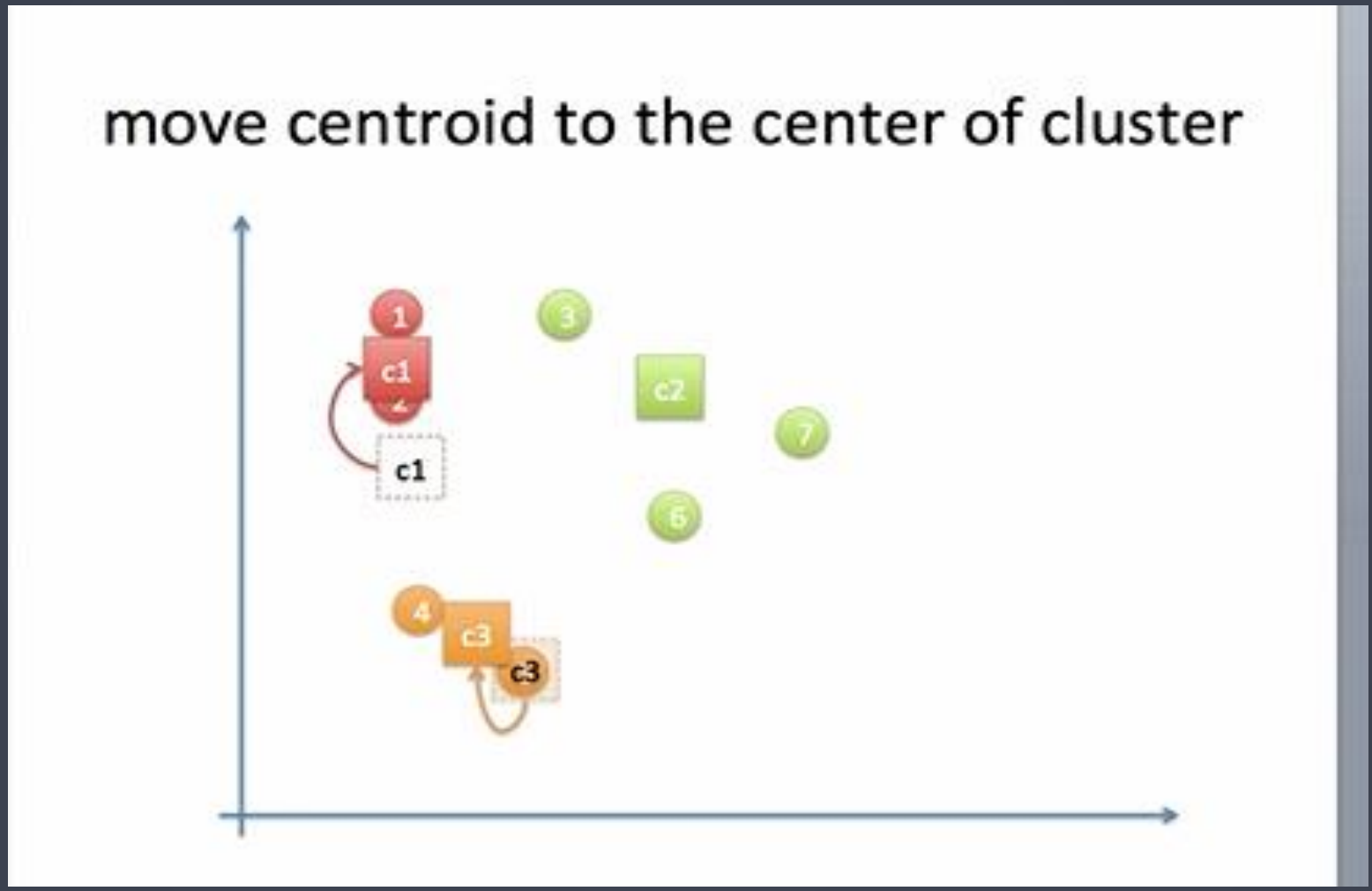
K-Means


First assignment is done!
클러스터의 중심(centroid)을 데이터의 중간의 점으로 옮김
1
2
3
4
6
7
c1
c1
c2
c2
c3
c3

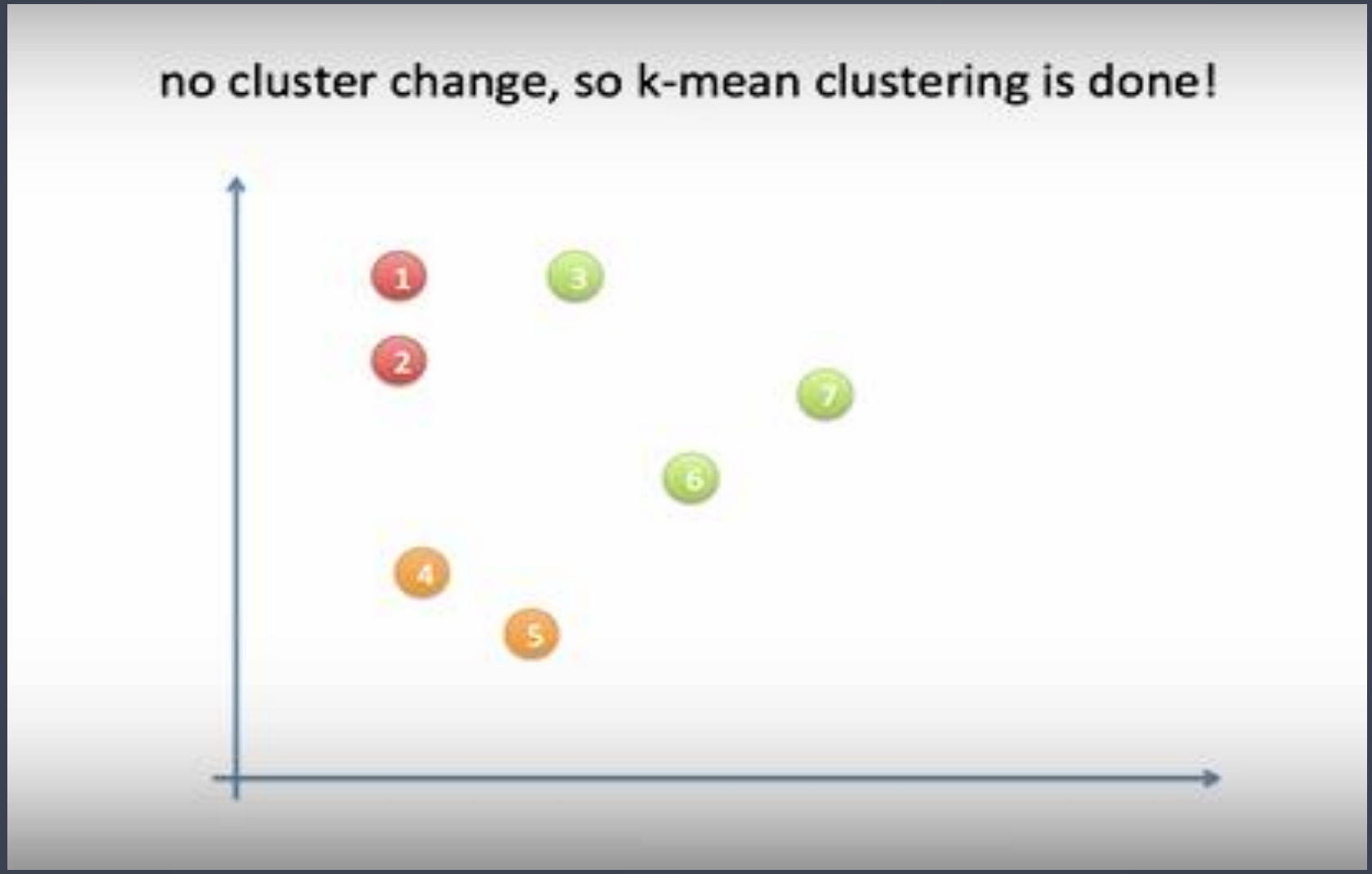
K-Means


from each data point,
assign cluster again using distance
1
3
2
c2
c1
nearest distance
cluster changed!
4
c3

K-Means



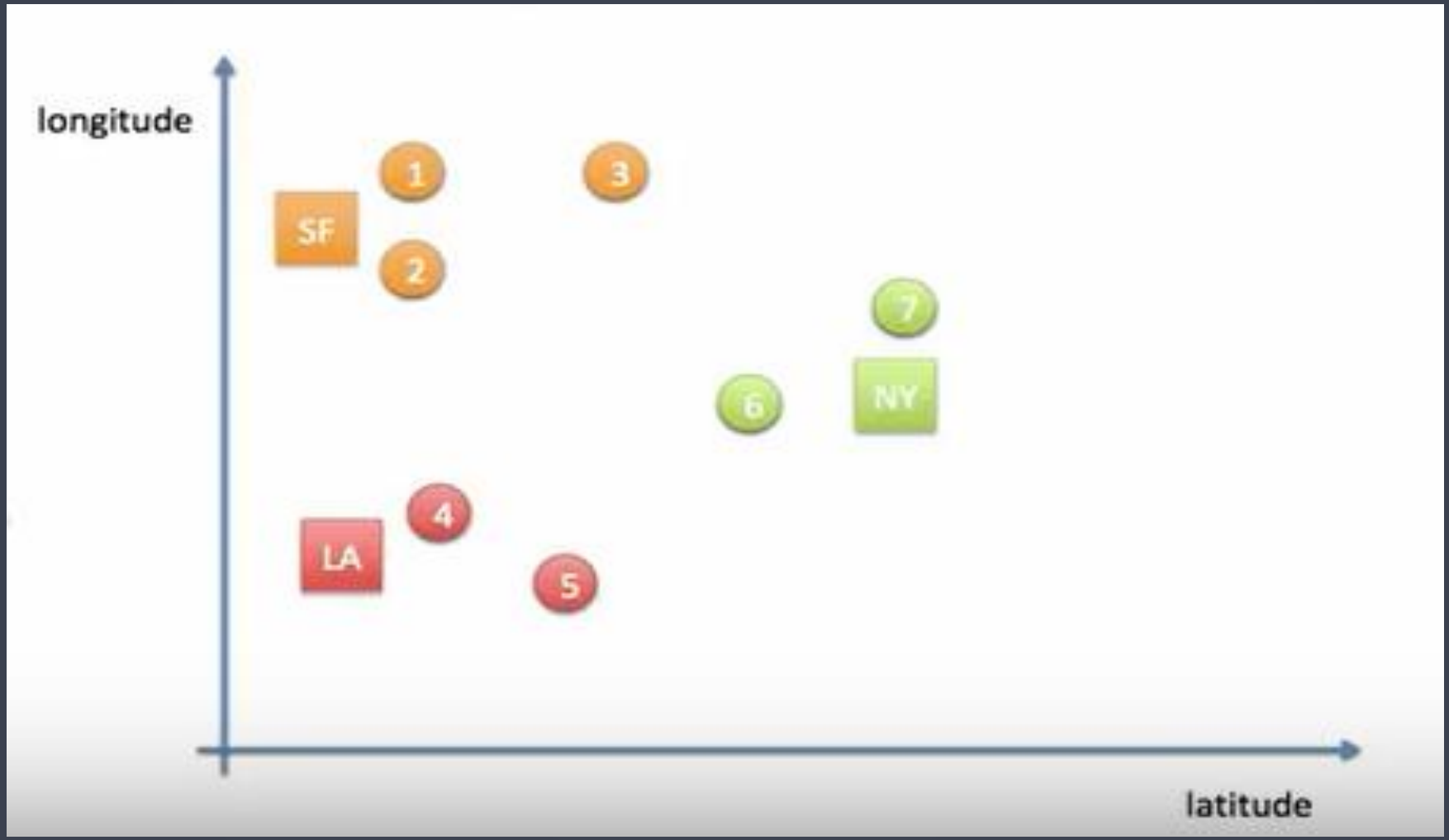
K-Means



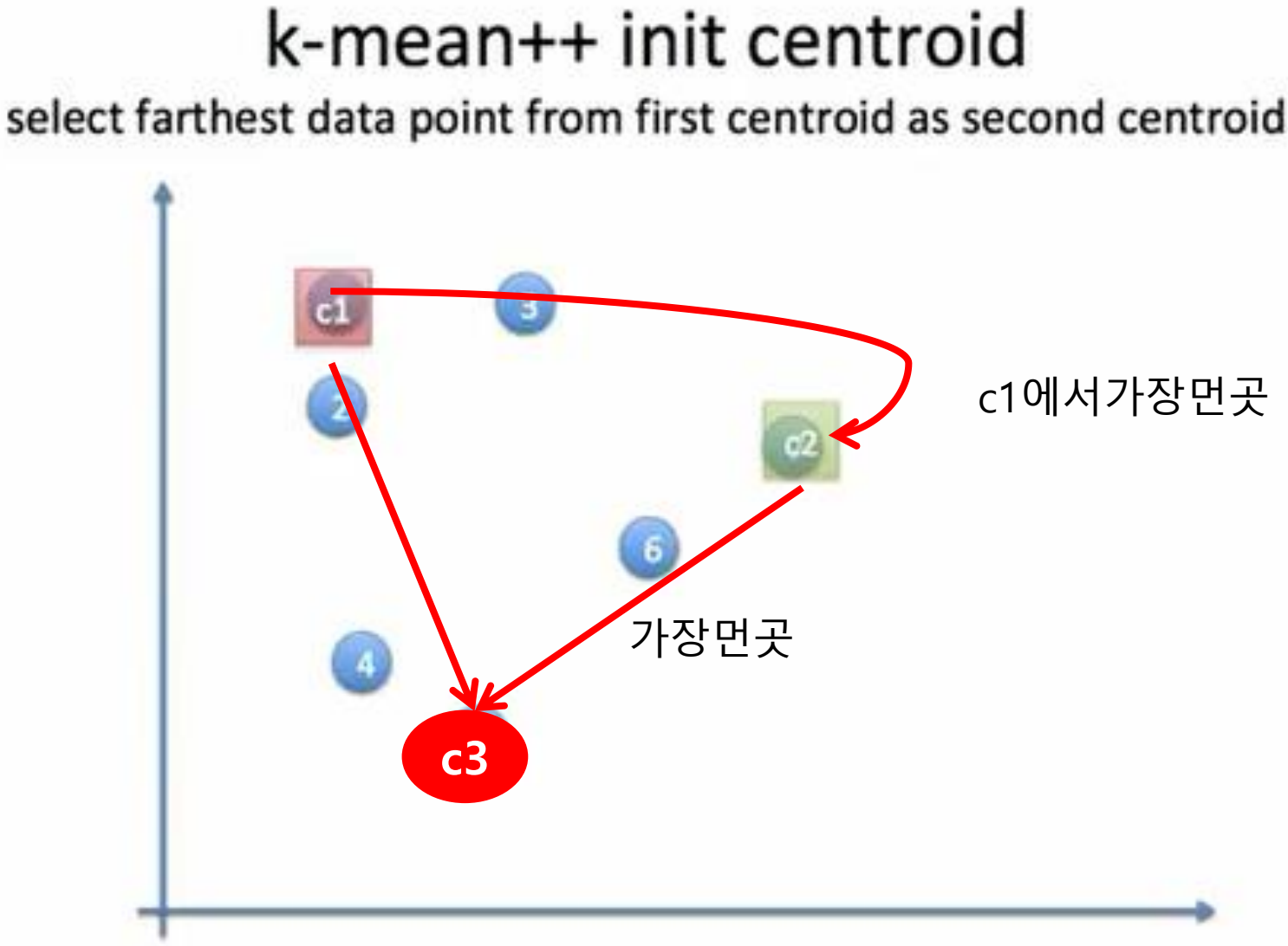
How to init centroid

1. Randomly choose
2. Manually assign init cenroid
3. k-mean++

manually assign init cenroid



k-mean++


k-mean++ init centroid
select farthest data point from first centroid as second centroid
c1
3
2
c2
6
4
c3
c1에서가장먼곳
가장먼곳