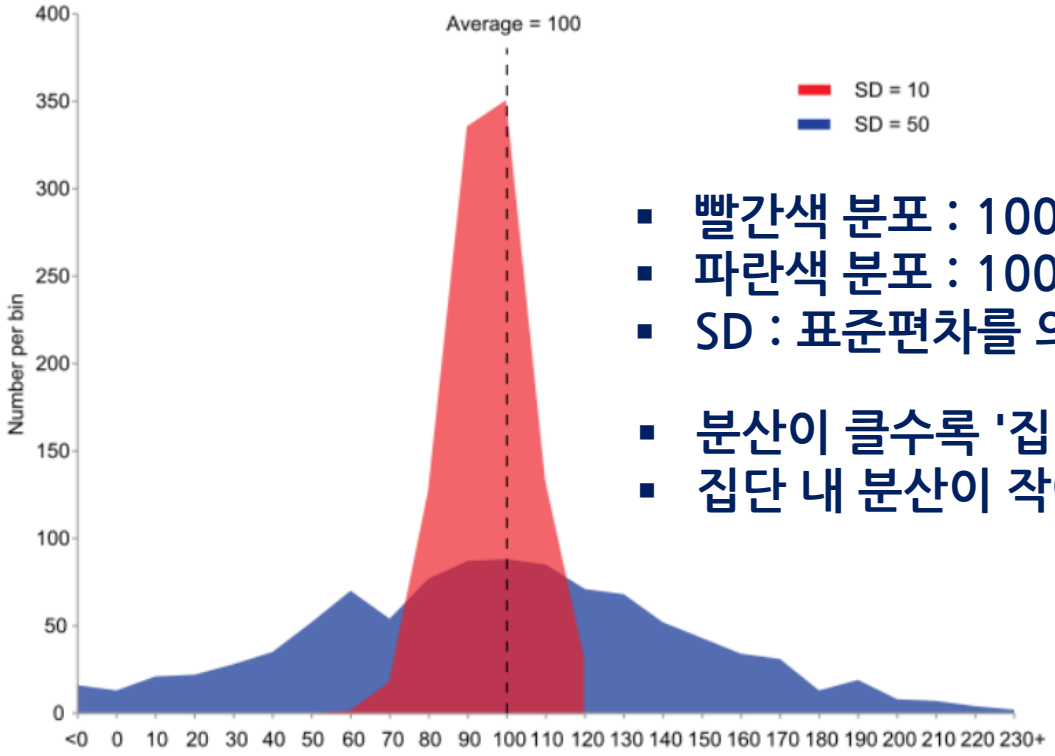

주성분 분석의 분산

분산

모집단(population)
모수(parameter)

추출(sampling) →
← 추론(inference)

표본(sample)
통계량(statistic)

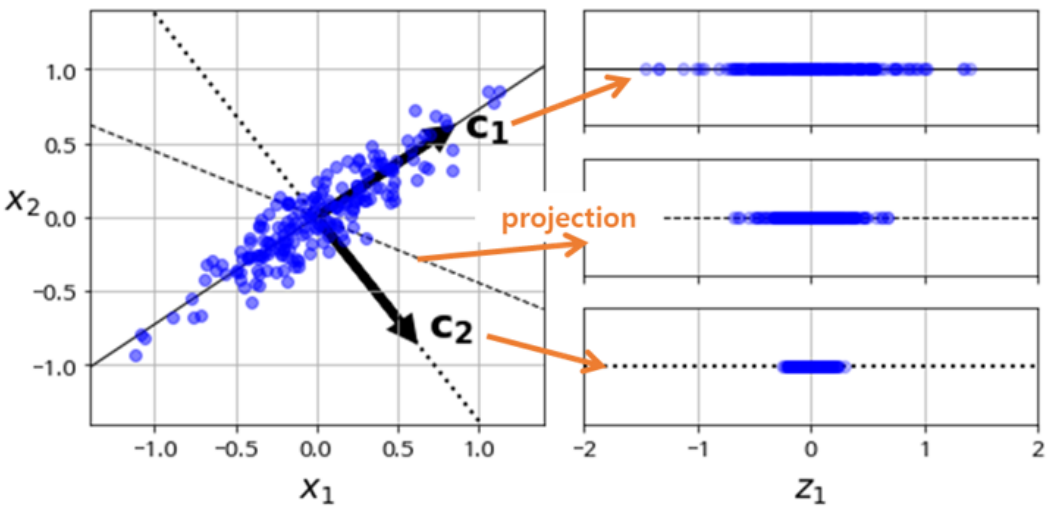


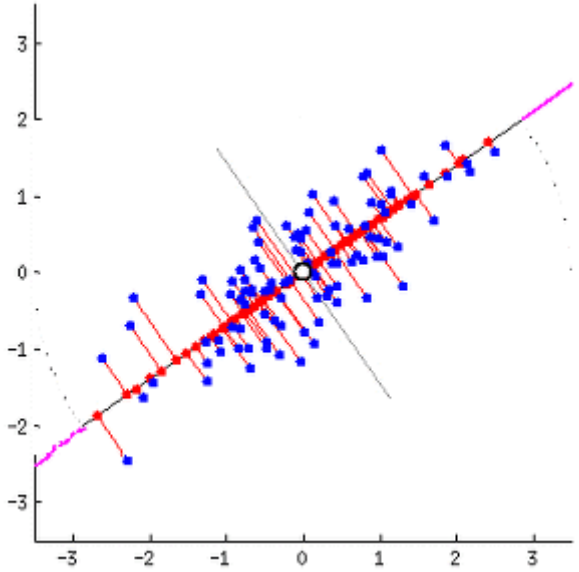
- 빨간색 분포 : 100의 평균값과 100의 분산 값
- 파란색 분포 : 100의 평균값과 2500의 분산 값
- SD : 표준편차를 의미

- 분산이 클수록 '집단의 평균값의 차이'가 무의미해짐
- 집단 내 분산이 작아질수록 평균의 차이가 분명해짐

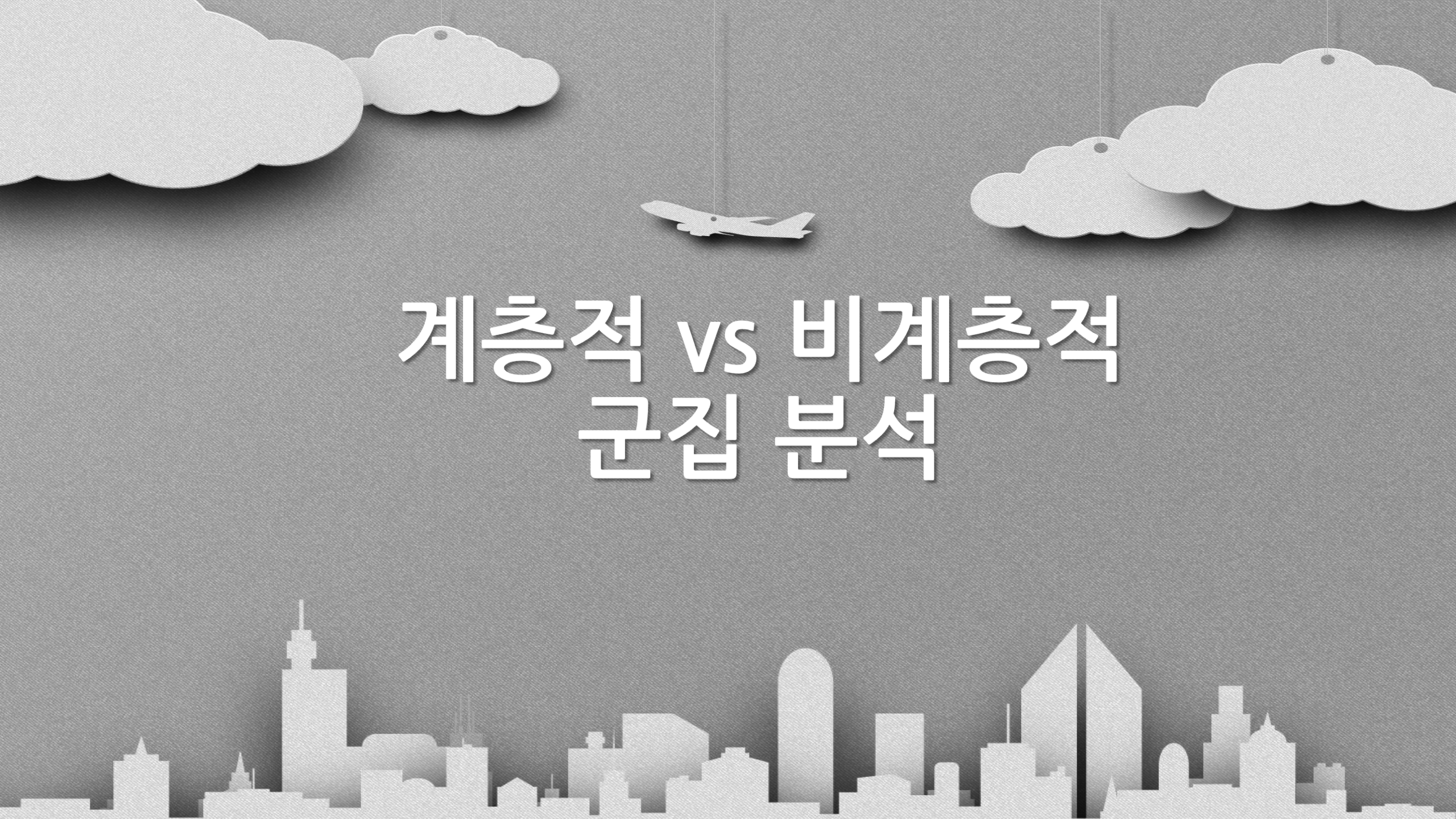
PCA(주성분 분석)

- "PCA는 데이터의 분산이 최대가 되는 축을 찾는다" = "정보의 손실을 최소화 한다 "





- 원본 데이터 셋과 투영(projection)된 데이터셋 간의 분산이 최대가 되는 축
  = 평균제곱거리(재구성 오차)를 최소화 하는 축을 찾음
- PCA 좋은 글 : https://laptrinhx.com/dimensionality-reduction-principal-component-analysis-359354885/



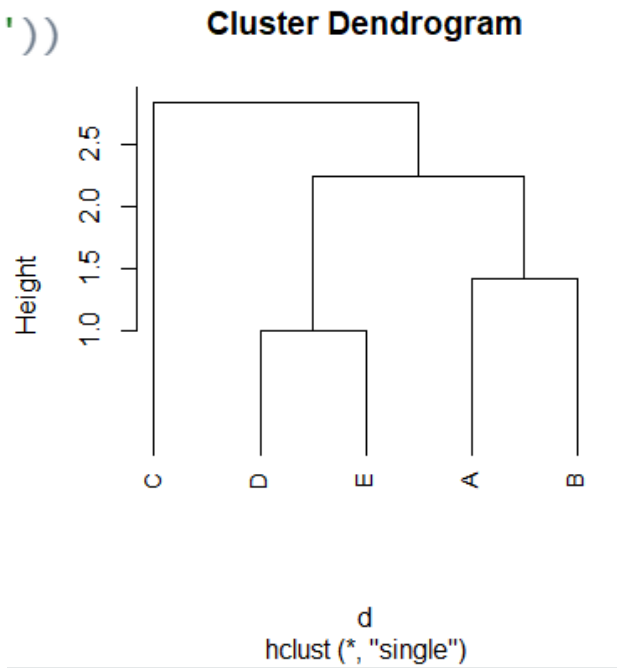
계층적 vs 비계층적 군집 분석

계층적 군집의 예

- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다. 최단연결법을 통해 학생들을 3개의 군집으로 나누면 어떻게 나누어 지는가? (Euclidean 거리 사용)

학생	(키, 몸무게)
A	(1, 5)
B	(2, 4)
C	(4, 6)
D	(4, 3)
E	(5, 3)

```
height <- c(1,2,4,4,5)
weight <- c(5,4,6,3,3)
student <- data.frame(height, weight,
                      row.names=c('A', 'B', 'C', 'D', 'E'))
d <- dist(student)
# single : 최단거리법, complete : 최장거리법
# average : 평균기준법
m <- hclust(d, method='single')
plot(m, hang=-1, cex=0.9)
```



계층적 군집의 예

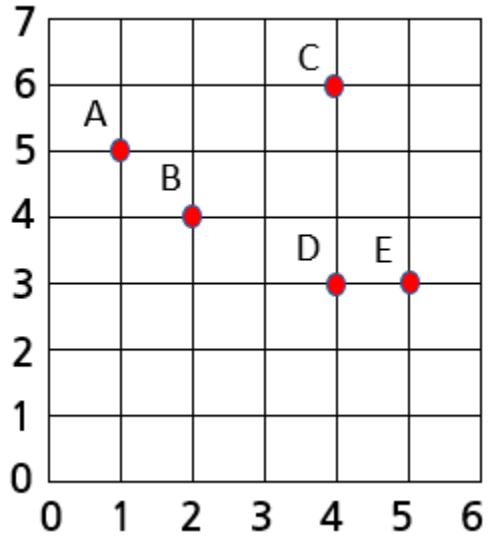
- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다. **최단연결법**을 통해 학생들을 3개의 군집으로 나누면 어떻게 나누어 지는가? (Euclidean 거리 사용)

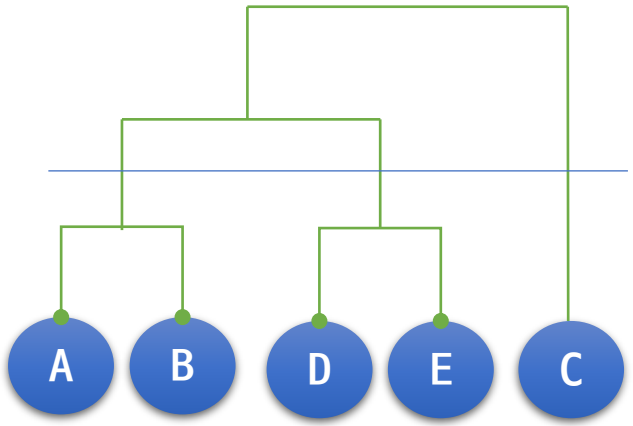
	A	B	C	D
B	2			
C	10	8		
D	13	5	9	
E	18	10	10	1

	A	B	C
B	2		
C	10	8	
DE	13	5	9

	AB	C
C	8	
DE	5	9

	C
ABDE	8





1. 각 학생 사이의 거리를 Euclidean 거리의 제곱으로 표시한 거리표를 작성한다 (표기, 연산의 간략화)
2. 가장 작은 숫자를 찾아 가장 먼저 군집을 형성하는 것을 찾고, 최단거리표를 작성한다 (최단연결법)
3. 그 다음 작은 값들을 찾아가며 계속 군집을 만들고 최단거리표를 다시 작성한다.

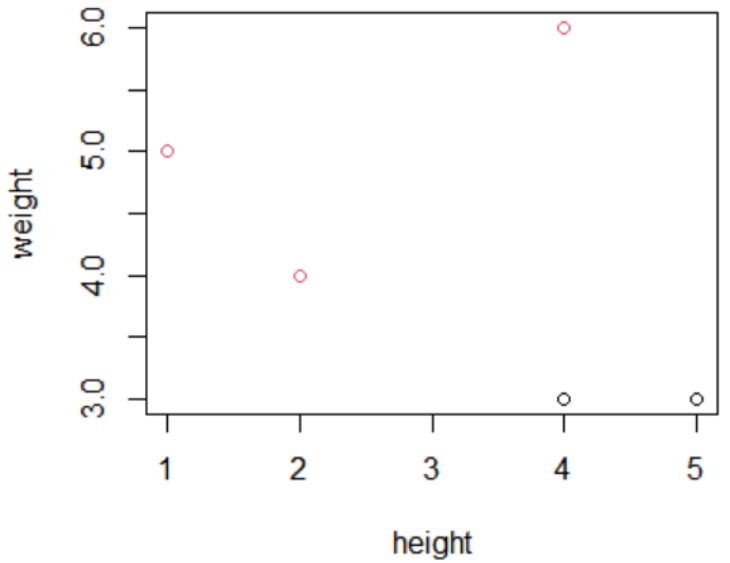
비계층적 군집의 예

- 아래는 학생들의 키와 몸무게를 정규화 한 데이터이다.
  K-means를 사용하여 비계층적 군집을 실행하라

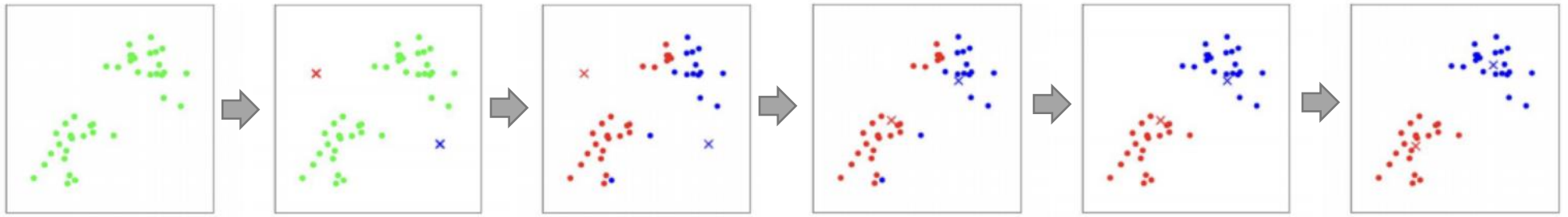
학생	(키, 몸무게)
A	(1, 5)
B	(2, 4)
C	(4, 6)
D	(4, 3)
E	(5, 3)

```
K-means clustering with 3 clusters of sizes 2, 2, 1

Cluster means:
  height weight
1    4.5    3.0
2    1.5    4.5
3    4.0    6.0

Clustering vector:
A B C D E
2 2 3 1 1

Within cluster sum of squares by cluster:
[1] 0.5 1.0 0.0
 (between_SS / total_SS =  91.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```
# 비계층적 군집
km = kmeans(student, 3)
km
plot(student, col=km$cluster)
```

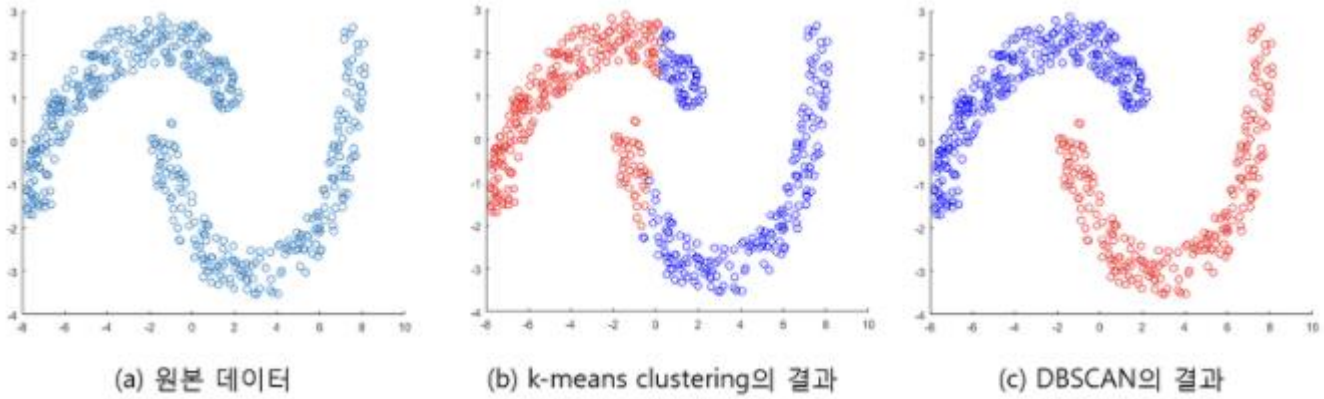


계층적 vs 비계층적 군집

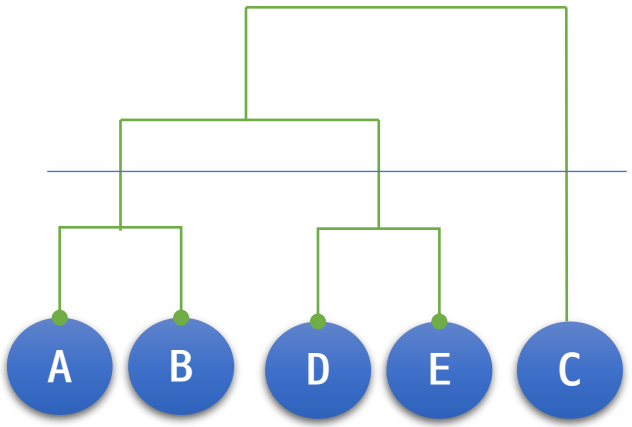


출처 : http://stanford.edu/~cpiech/cs221/img/kmeansViz.png

비계층적 군집



이미지출처 :
https://untitledtblog.tistory.com/146



계층적 군집