주성분 분석
K-Means 클러스터링

김수인

KAIST
http://suin.kim

강의 목표

- 30분 강의
- elice.io 문제를 알고리즘을 이해하고 풀 수 있게 돕는 것
  - 이 데이터에 왜 이 알고리즘을 사용할까?
  - 파라미터를 어떻게 바꾸면 어떻게 동작할까?
  - 알고리즘이 내보내는 결과를 어떻게 해석할까?
  - 앞으로 내가 회사나 연구실에서 이 알고리즘을 사용할 때 주의해야 할 점은?

주성분 분석

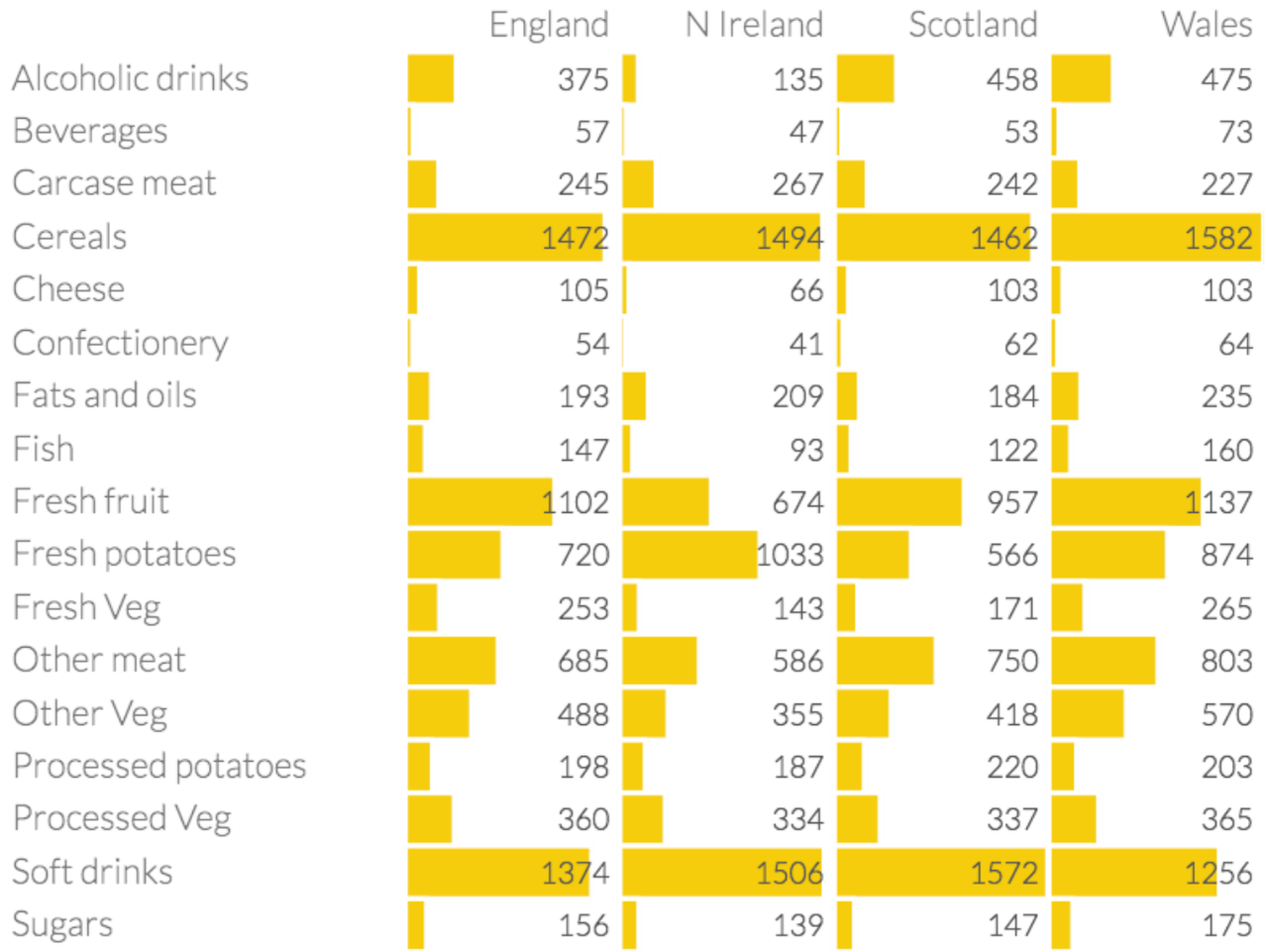
고차원의 데이터를
저차원의 데이터로 환원시키는 기법

주성분 분석

고차원의 데이터를 정보를 최대한 유지하면서
저차원의 데이터로 환원시키는 기법

주성분 분석: 왜 사용하는가 (1)

- 고차원의 데이터를 사람이 이해가능하게 시각화

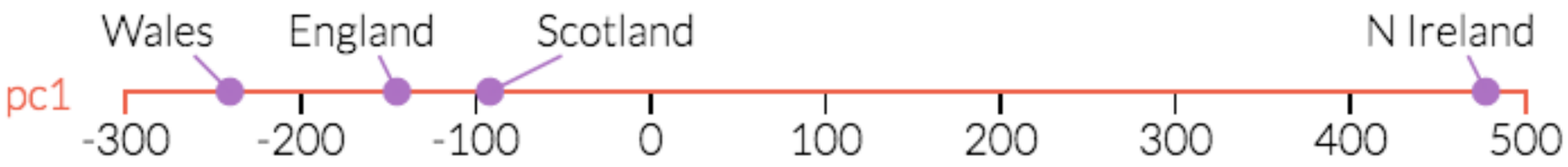


source: http://setosa.io/ev/principal-component-analysis/

주성분 분석: 왜 사용하는가 (1)

- 고차원의 데이터를 사람이 이해가능하게 시각화

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



source: http://setosa.io/ev/principal-component-analysis/

주성분 분석: 왜 사용하는가 (2)

- 변수간의 숨겨진 연관성을 분석

주거지역 정보 데이터

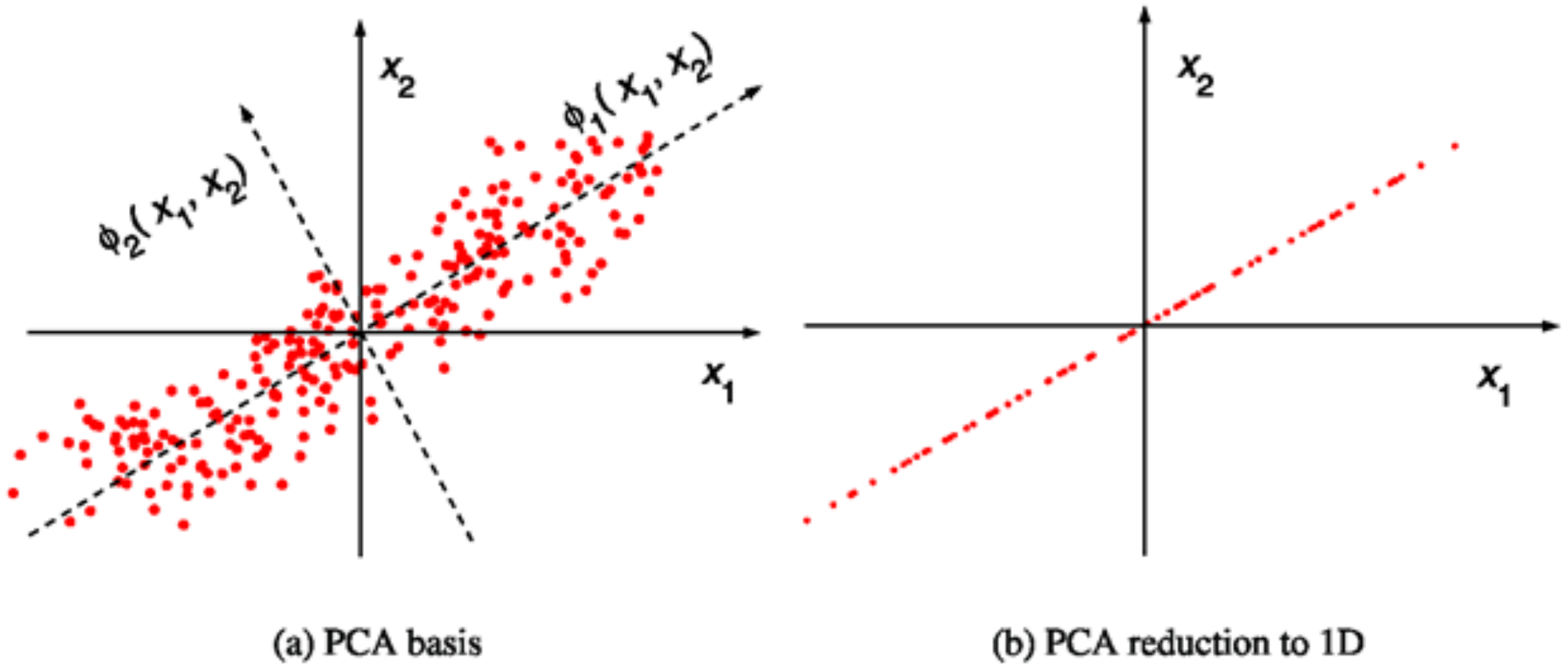
	Principal Component		
Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144
Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

1. 주거시설, 의료시설, 교통, 문화, 오락시설은 서로 관련되어 있음
2. 두 번째 주성분은 의료시설에만 연관된 것임
3. 범죄율 ↑ 오락시설 ↑

source: https://onlinecourses.science.psu.edu/stat505/node/54

주성분 분석: Motivation

- 2차원의 데이터를 1차원으로 축소하기



(a) PCA basis

(b) PCA reduction to 1D

image source: http://what-when-how.com/face-recognition/face-recognition-in-subspaces-face-image-modeling-and-representation-part-1/

주성분 분석: Motivation

- 2차원의 데이터를 1차원으로 축소하기

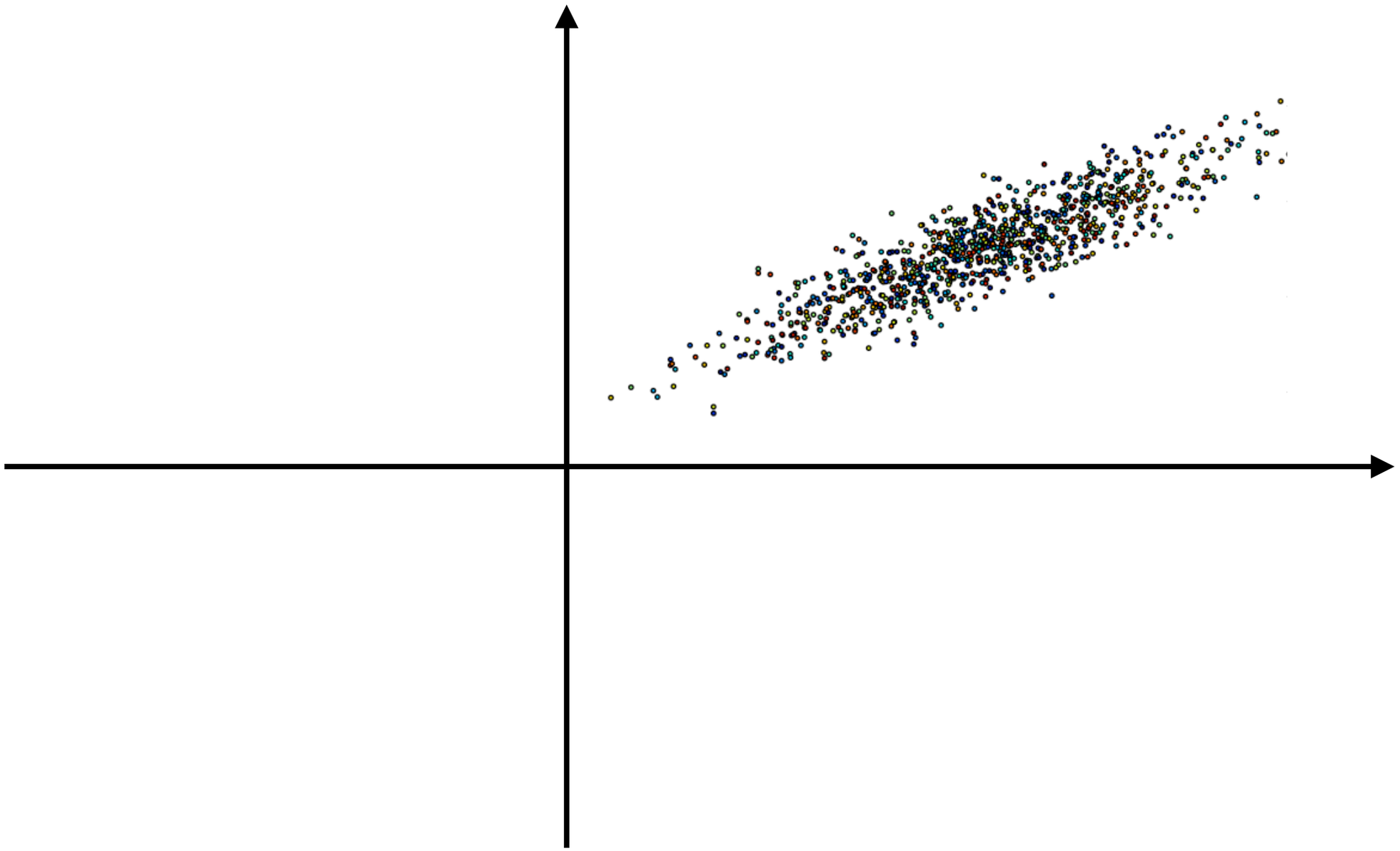


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

- 2차원의 데이터를 1차원으로 축소하기

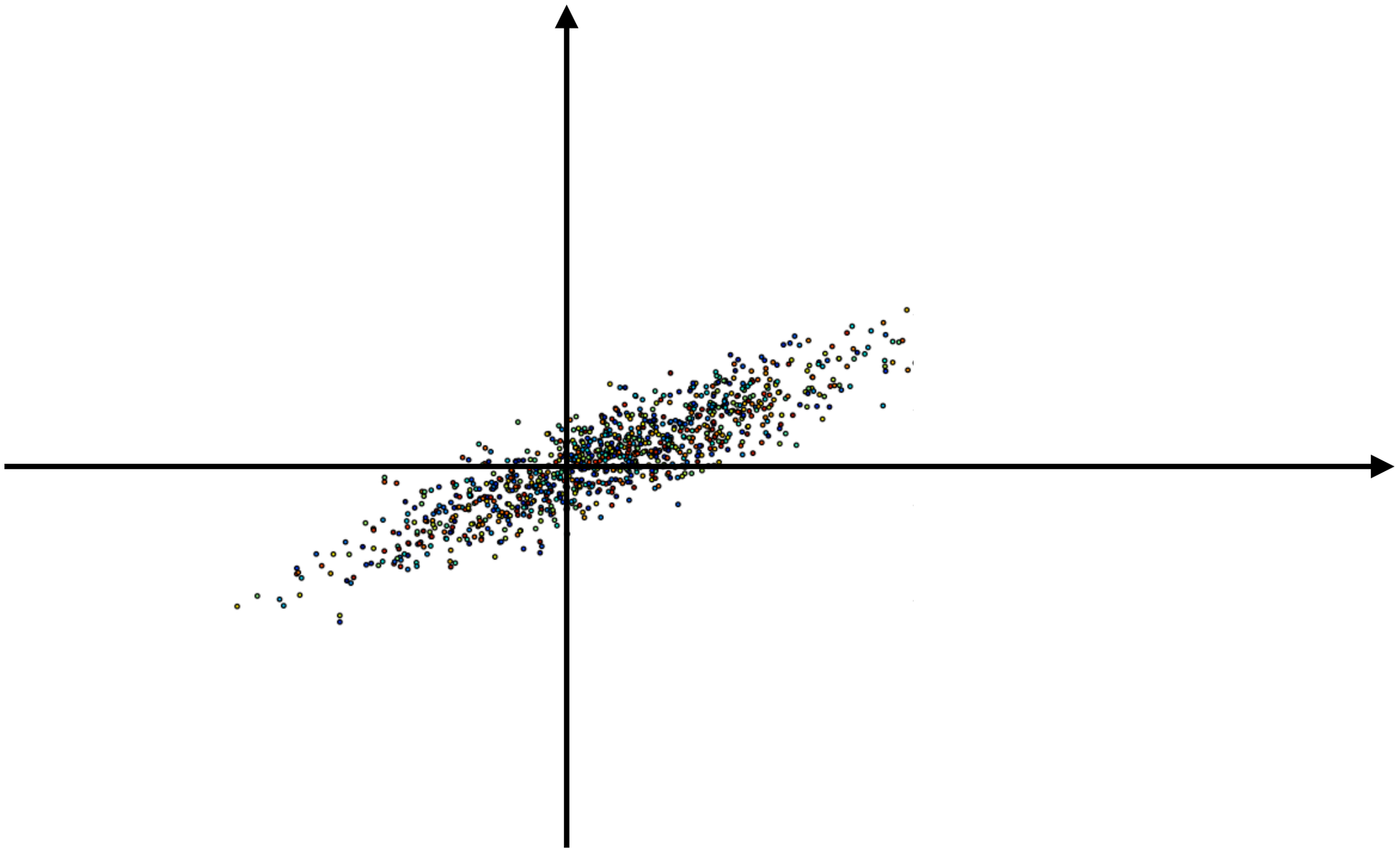


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

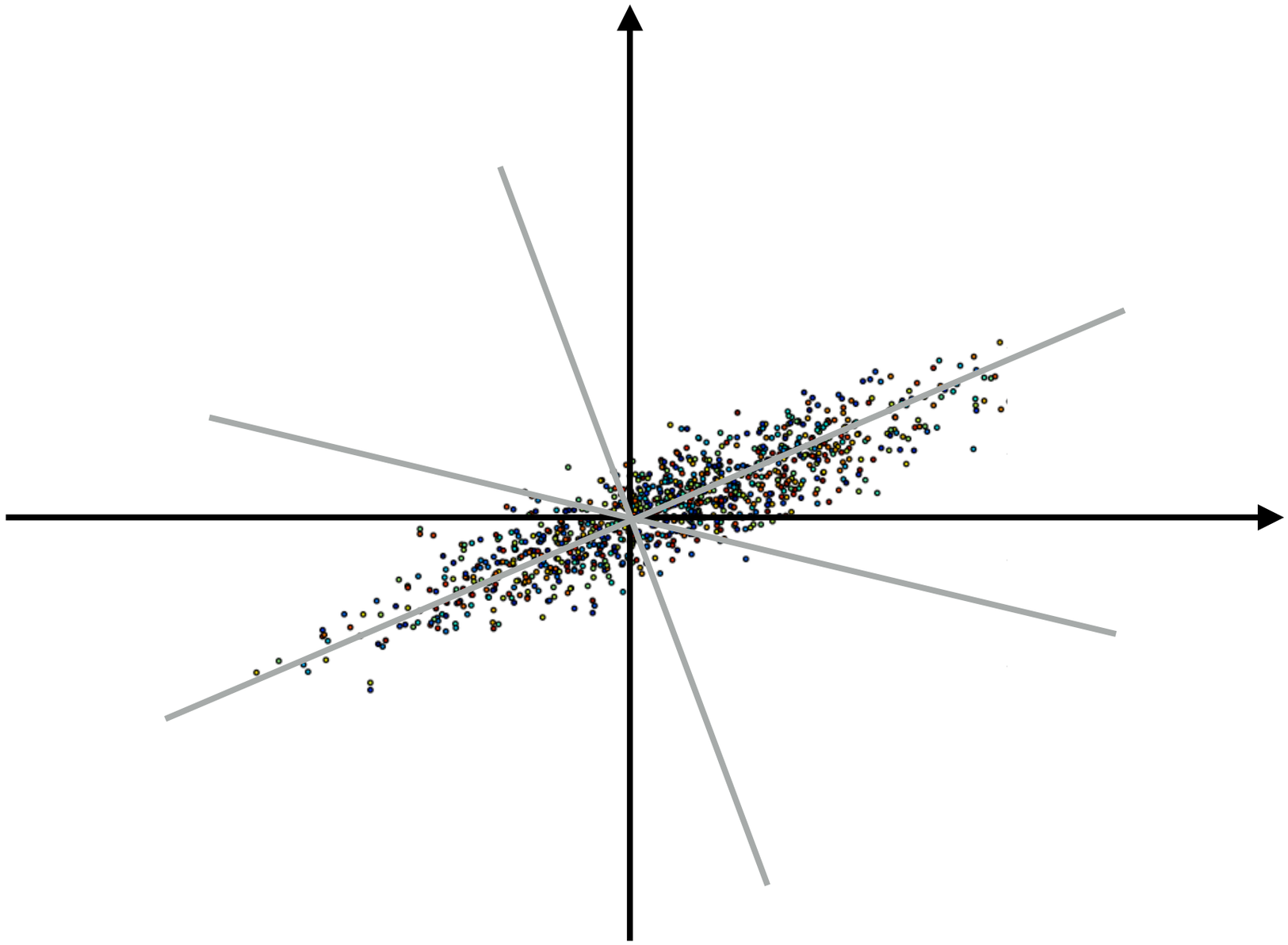


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

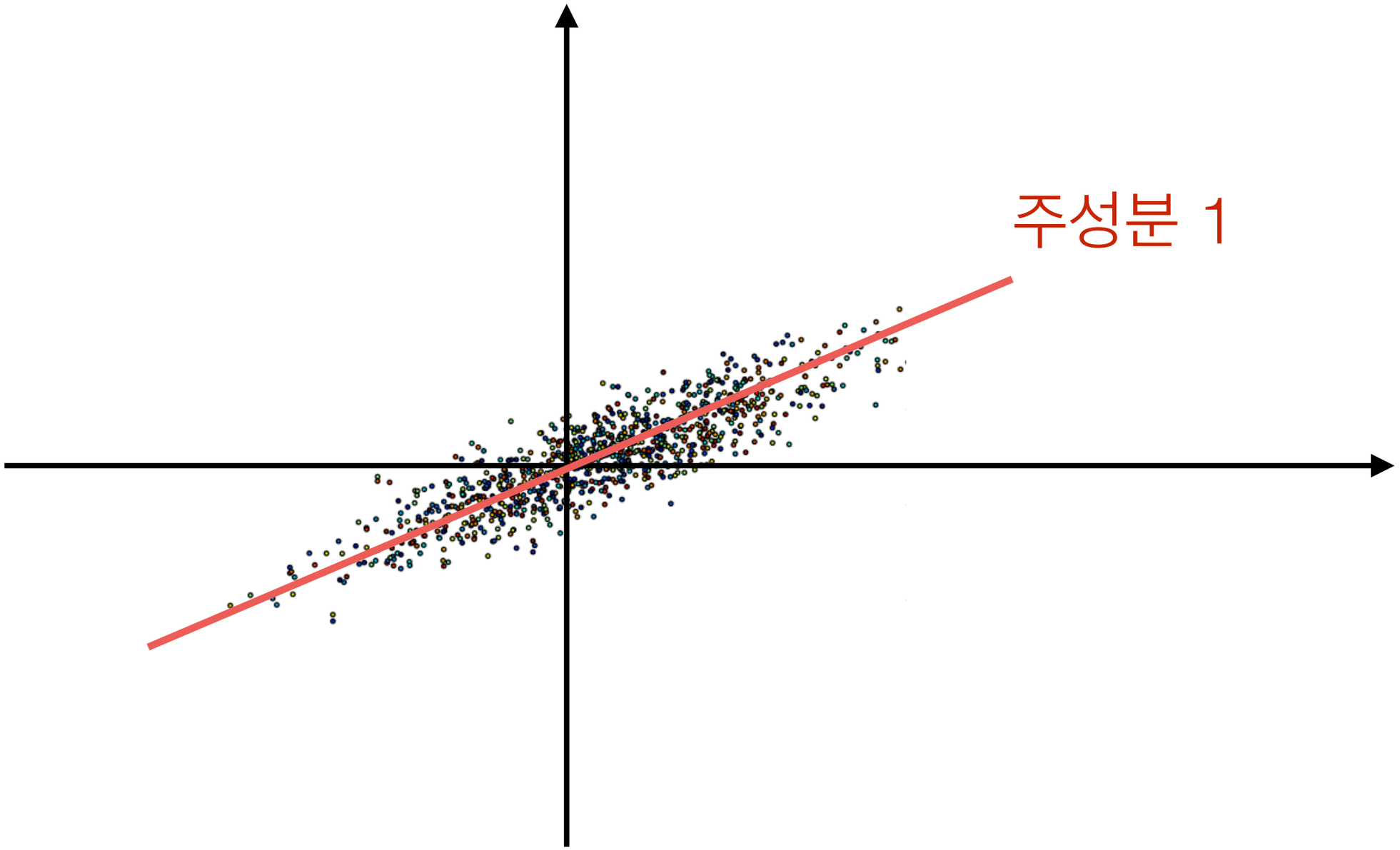
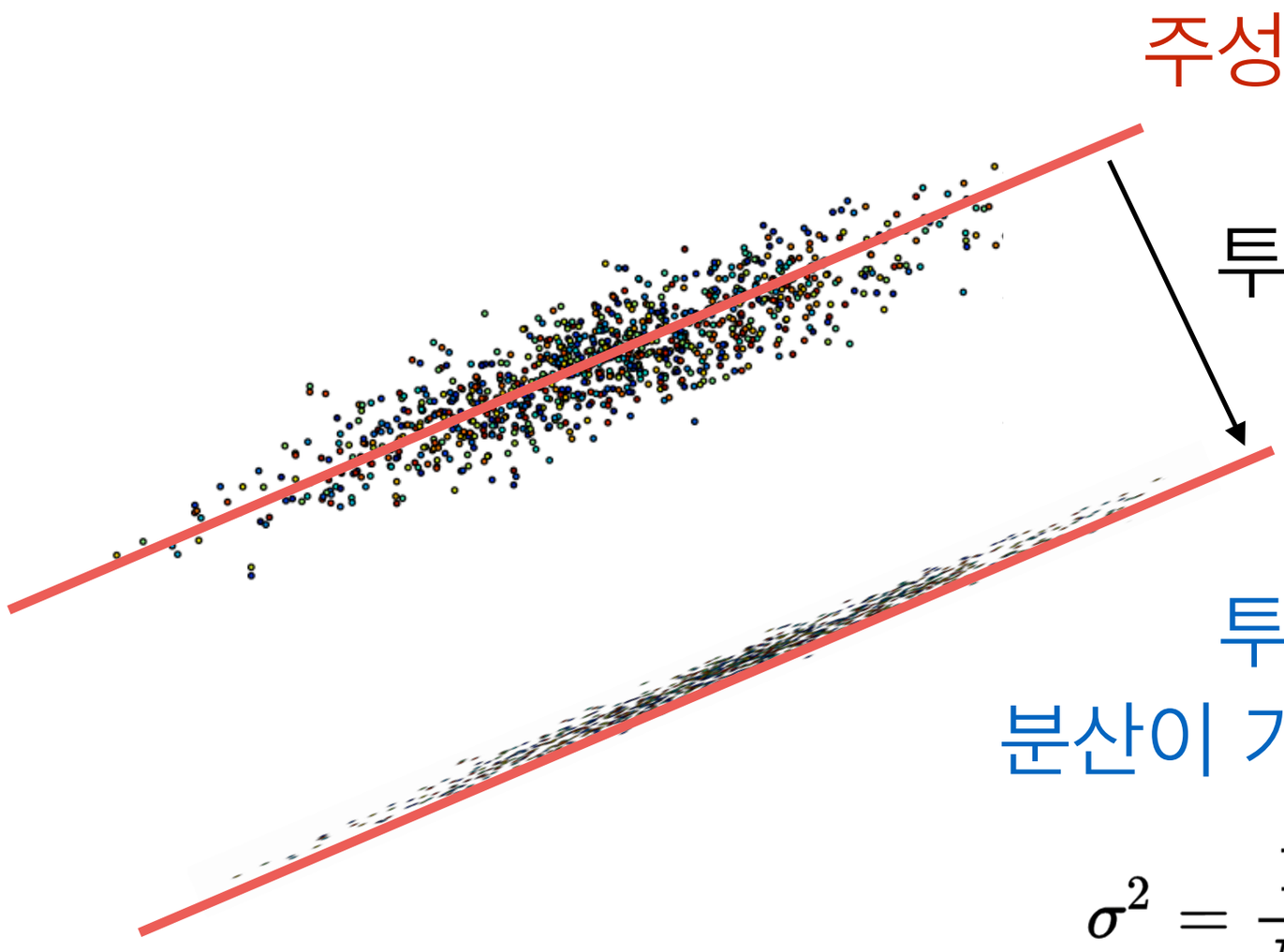


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation



투사했을때
분산이 가장 커지는 방향

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

- 주성분을 만드는 것은 데이터를 축에 맞추어 이동하는것과 같음

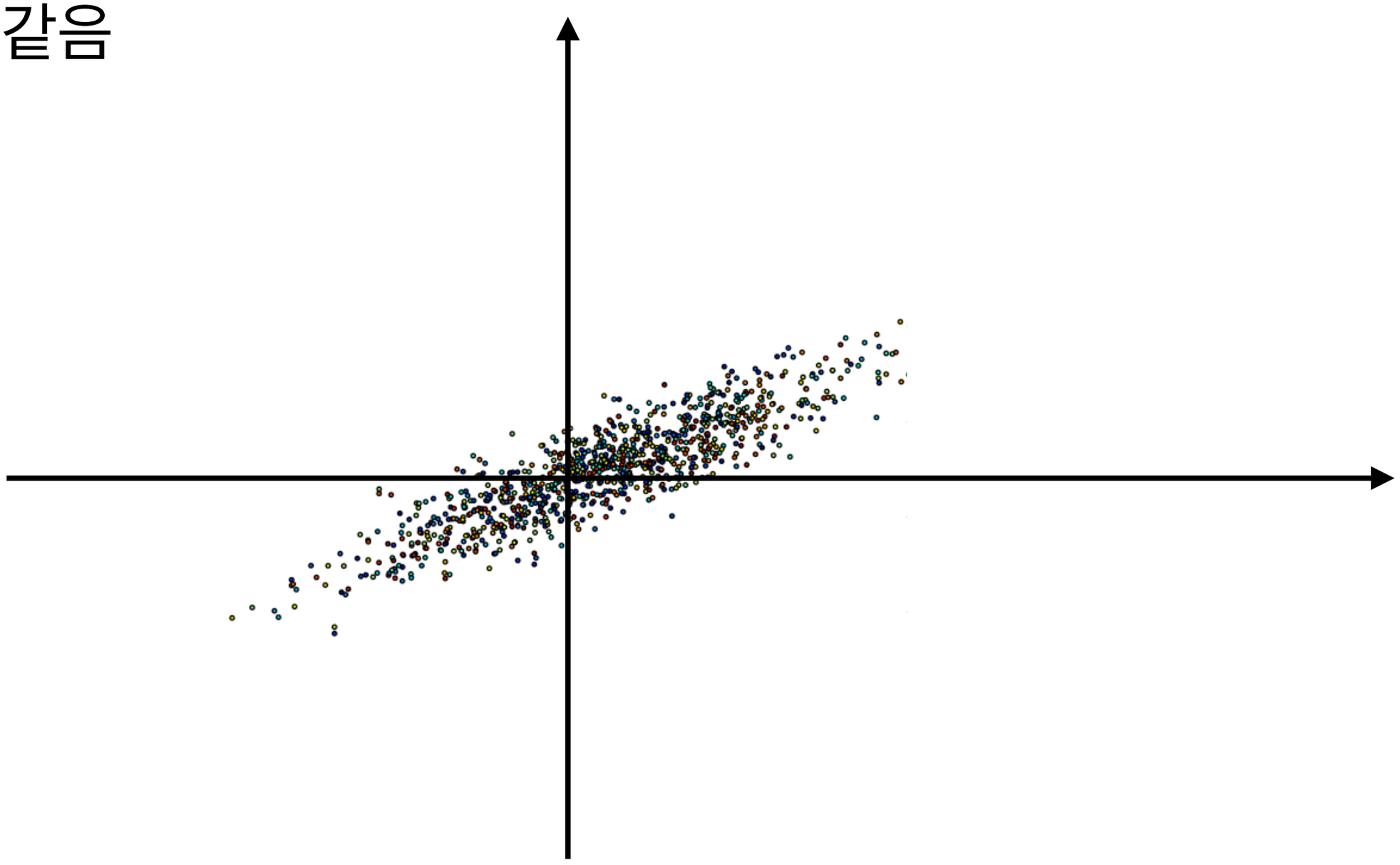


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

- 주성분을 만드는 것은 데이터를 축에 맞추어 이동하는것과 같음

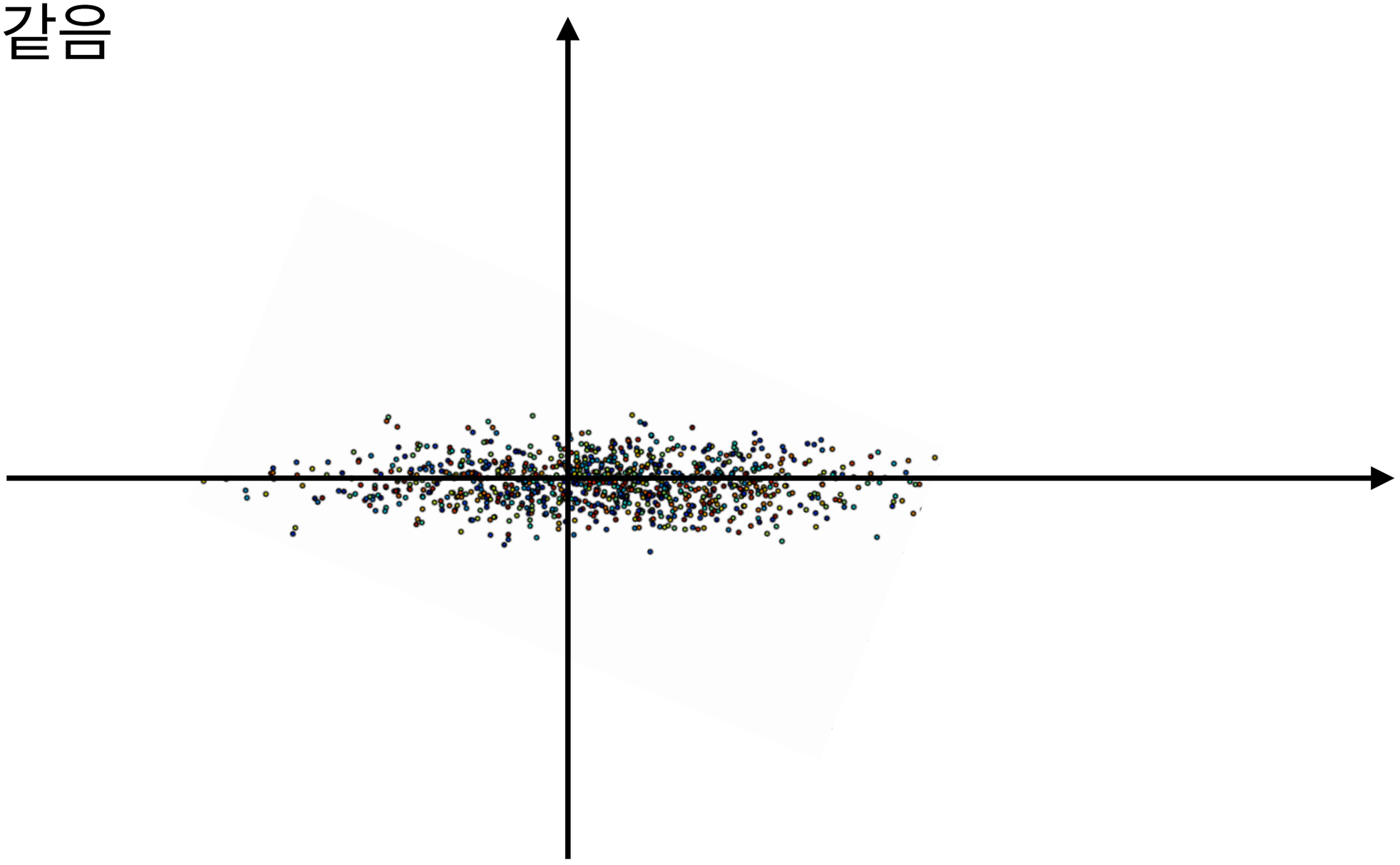
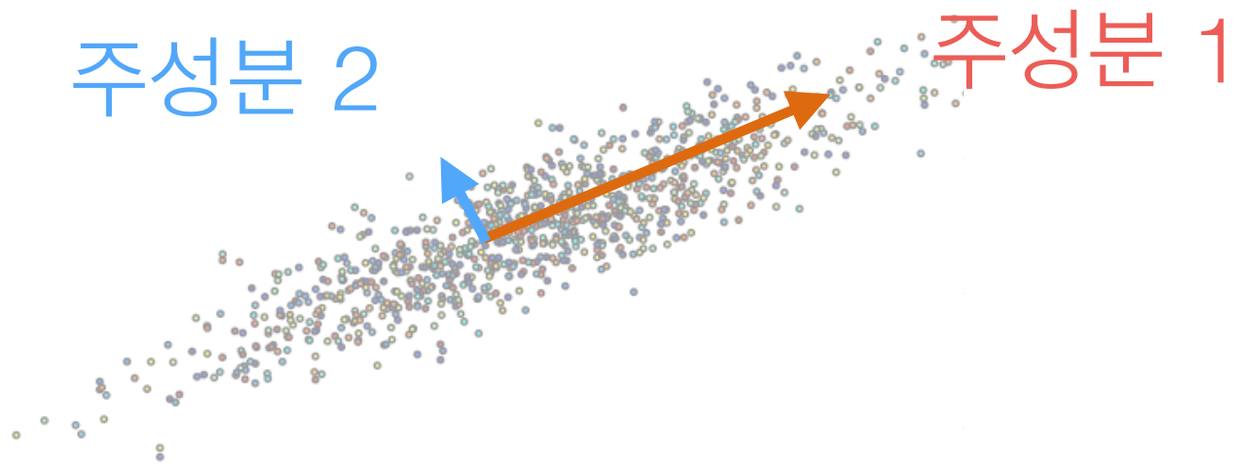


image source: http://austingwalters.com/pca-principal-component-analysis/

주성분 분석: Motivation

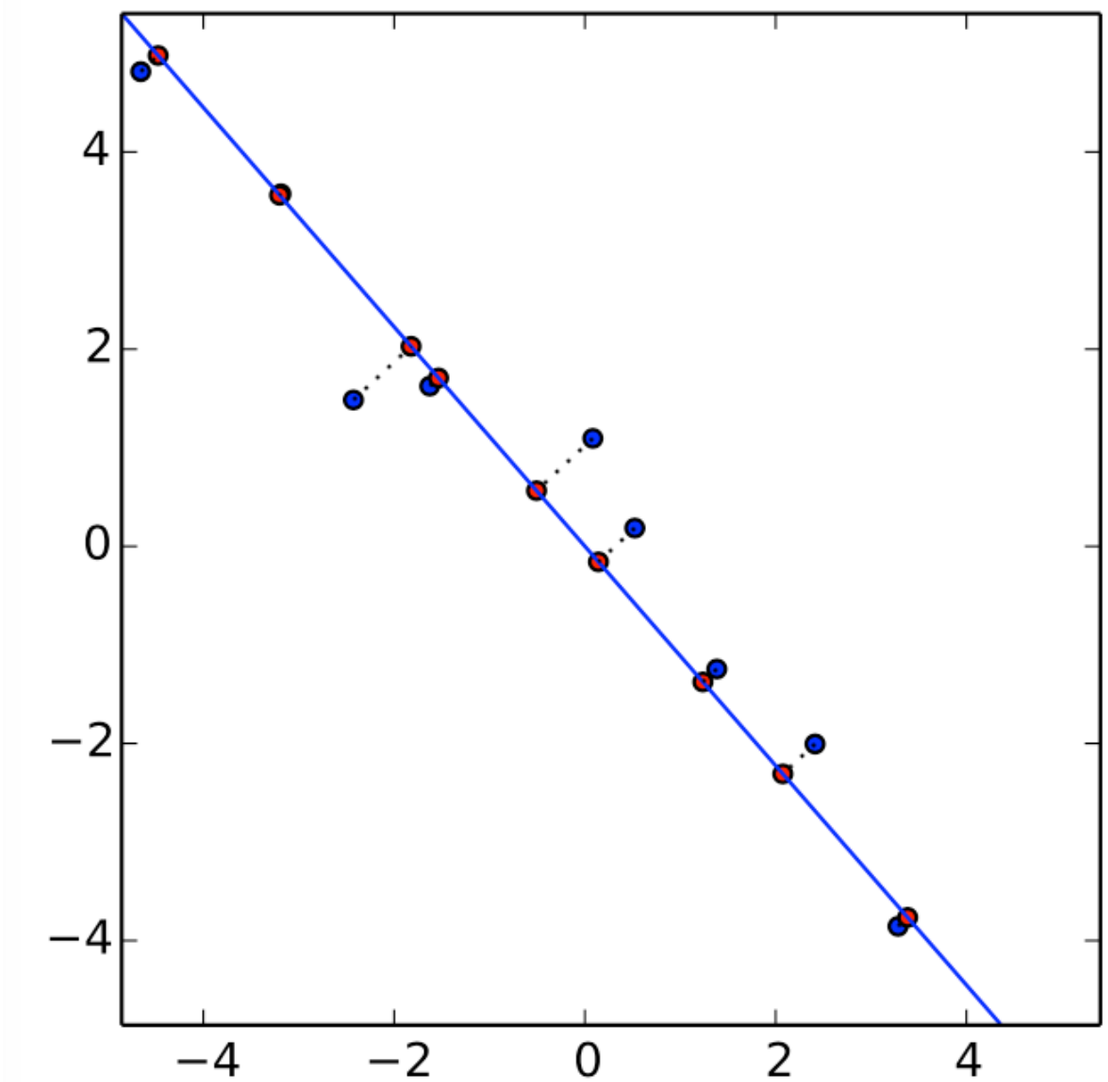
- N 차원 데이터 집합에 대해 최대 N 개의 서로 수직인 주성분 벡터를 발견할 수 있음



- 첫 번째 주성분은 그 방향으로 데이터들의 분산이 가장 큰 축
- 두 번째 주성분은 첫 번째 주성분으로 표현할 수 없는 축 중에서 가장 분산이 큰 축
- …

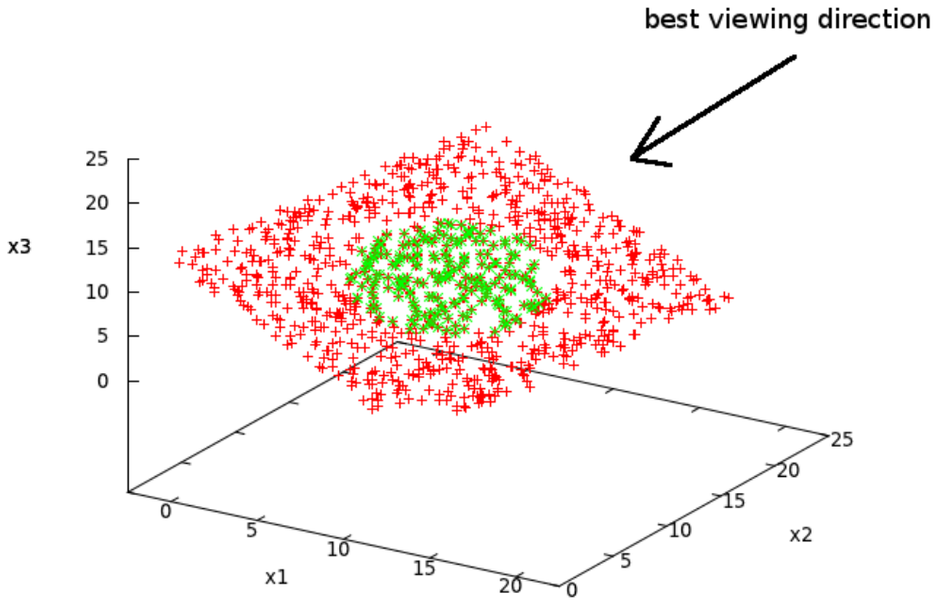
image source: http://austingwalters.com/pca-principal-component-analysis/

elice.io 에서 실행해보기

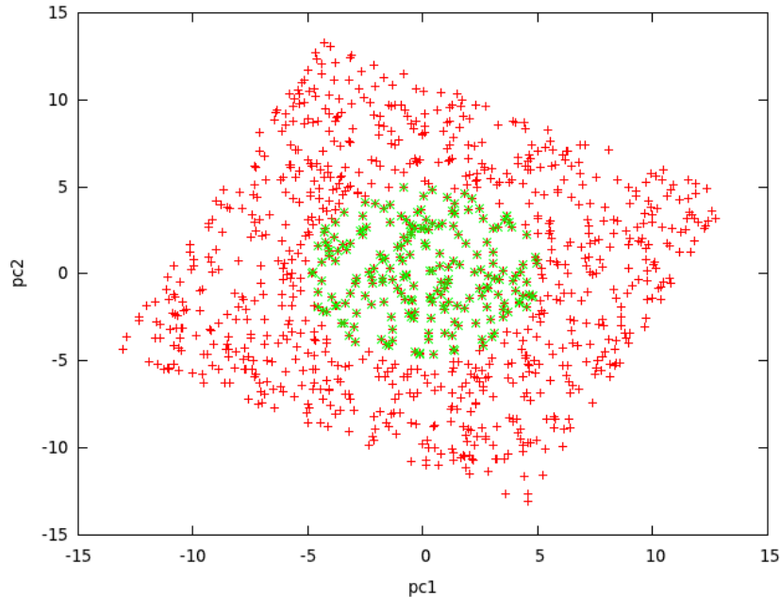


고차원 데이터 분석

- 3차원 데이터
  - 첫 번째 주성분을 찾고 투사: 1차원
  - 두 번째 주성분을 찾고 투사: 2차원
- http://setosa.io/ev/principal-component-analysis/



3D

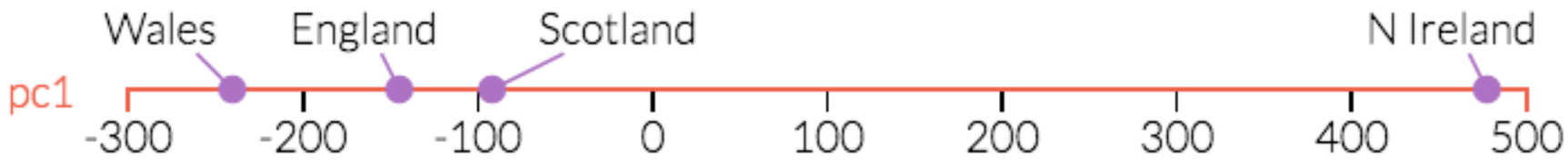


2D

주성분 분석: 왜 사용하는가 (1)

- 고차원의 데이터를 사람이 이해가능하게 시각화
- 첫 번째 주성분에 17차원 데이터를 투사하여 1차원으로 축소한 결과

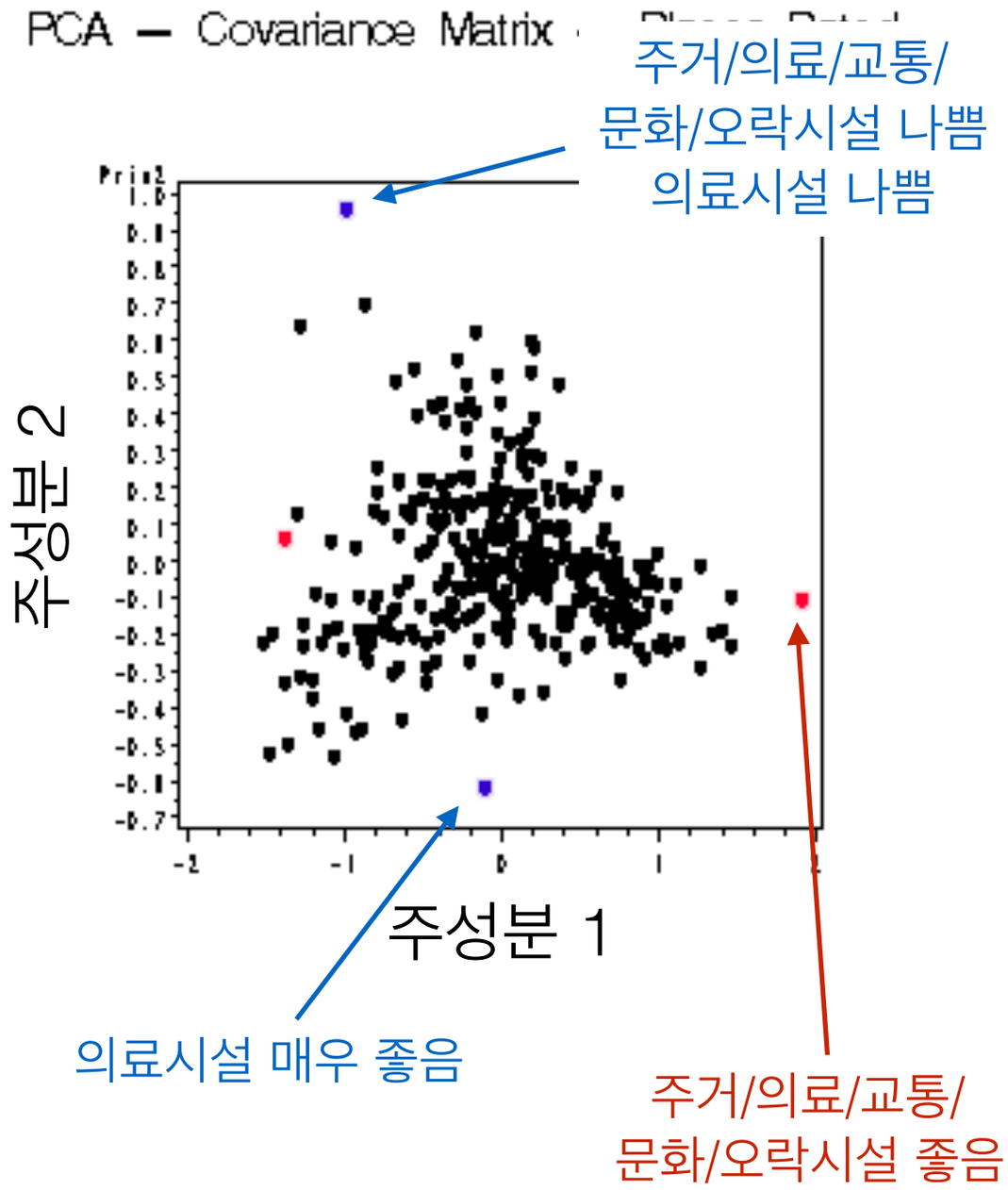
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



source: http://setosa.io/ev/principal-component-analysis/

주성분 분석: 왜 사용하는가 (2)

	Principal Component		
Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144
Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

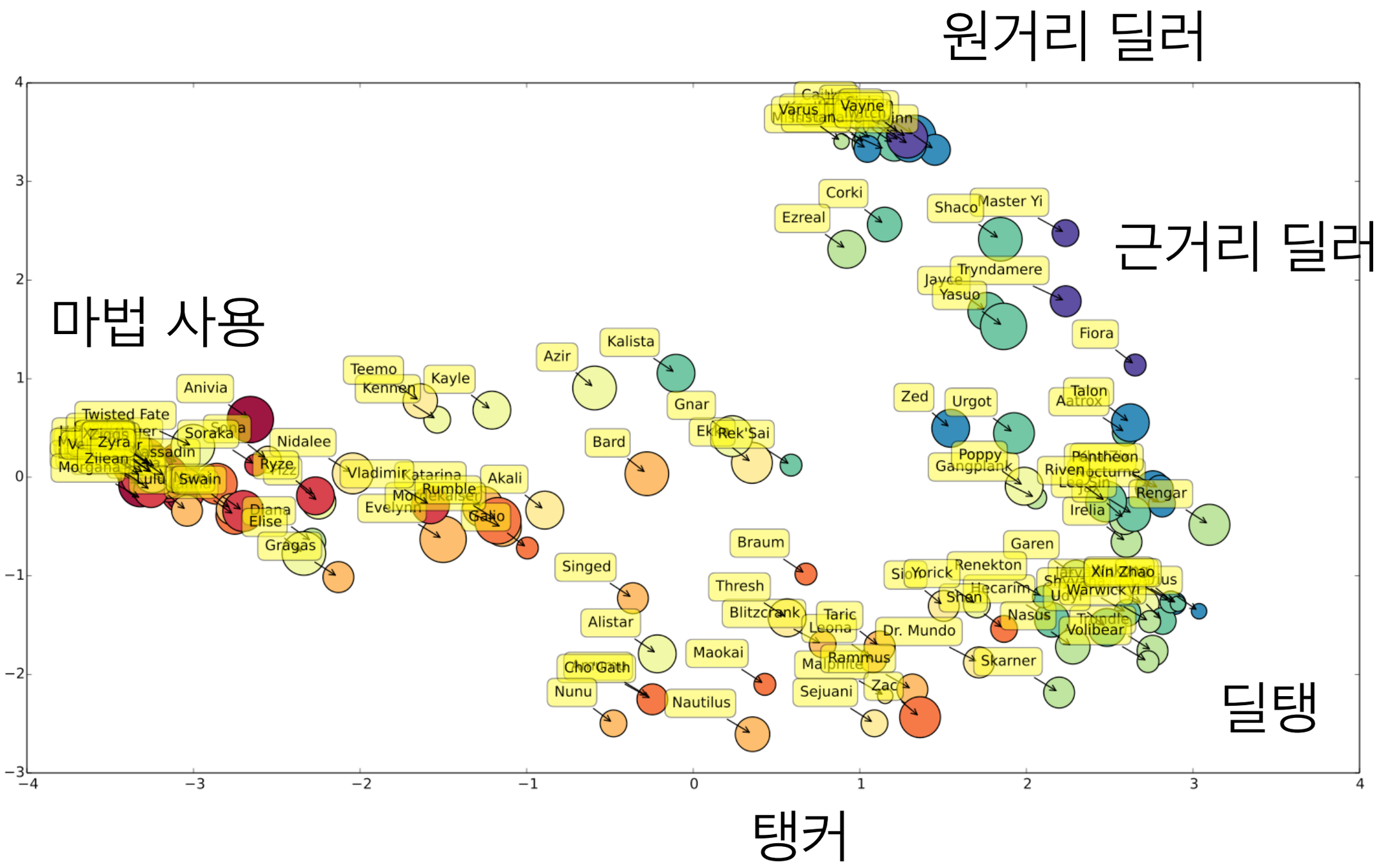


1. 주거시설, 의료시설, 교통, 문화, 오락시설은 서로 관련되어 있음
   0.985: 이 주성분은 Arts 를 측정하는 성분임
2. 두 번째 주성분은 의료시설에 음으로 연관됨

source: https://onlinecourses.science.psu.edu/stat505/node/54

LoL 데이터 분석

174차원 → 2차원



LoL 주성분 살펴보기

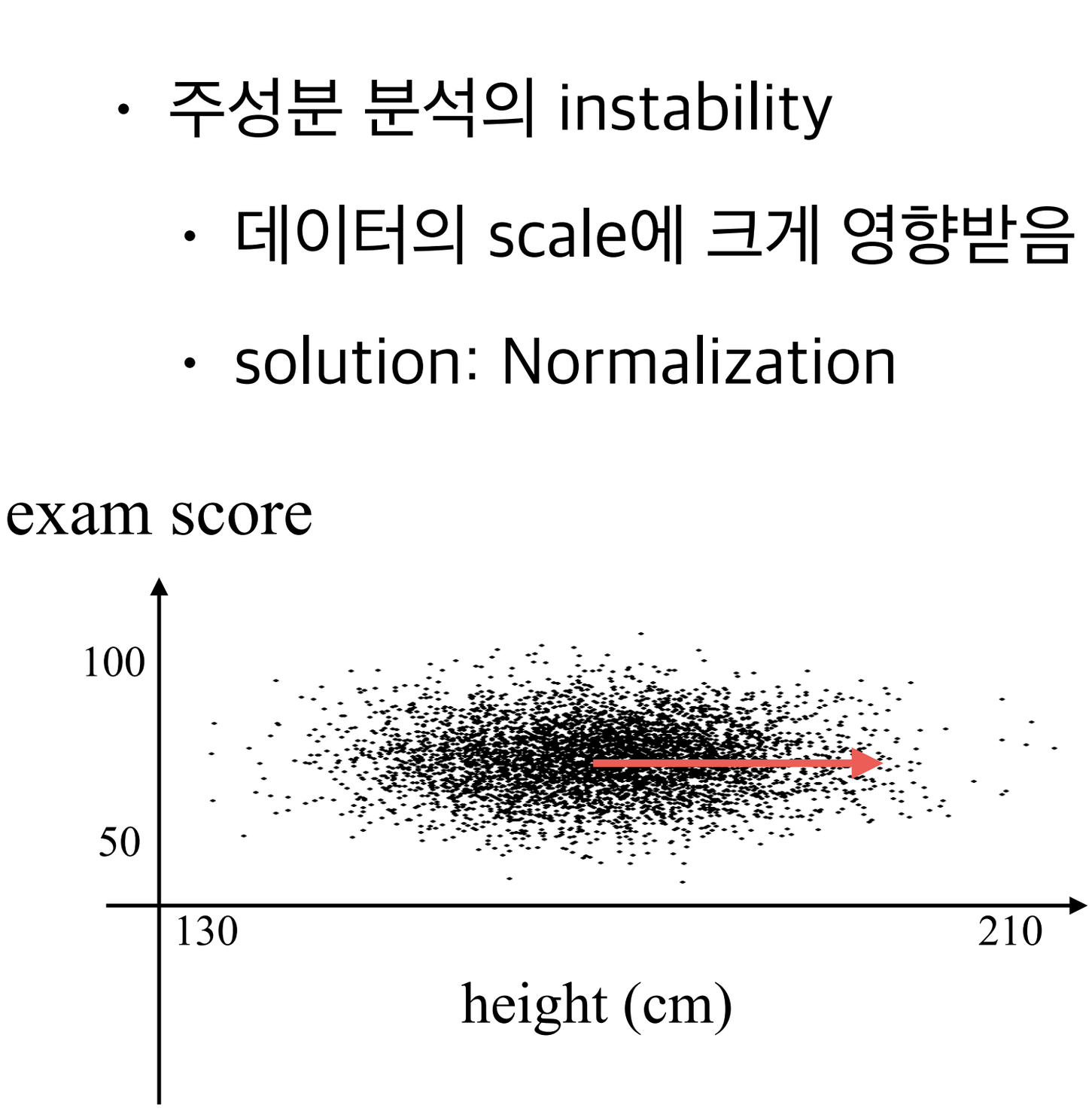
- Analysis: “A subjective decision”
- 마법 — 마법 기술 범위
- 공격력 — 마법 저항

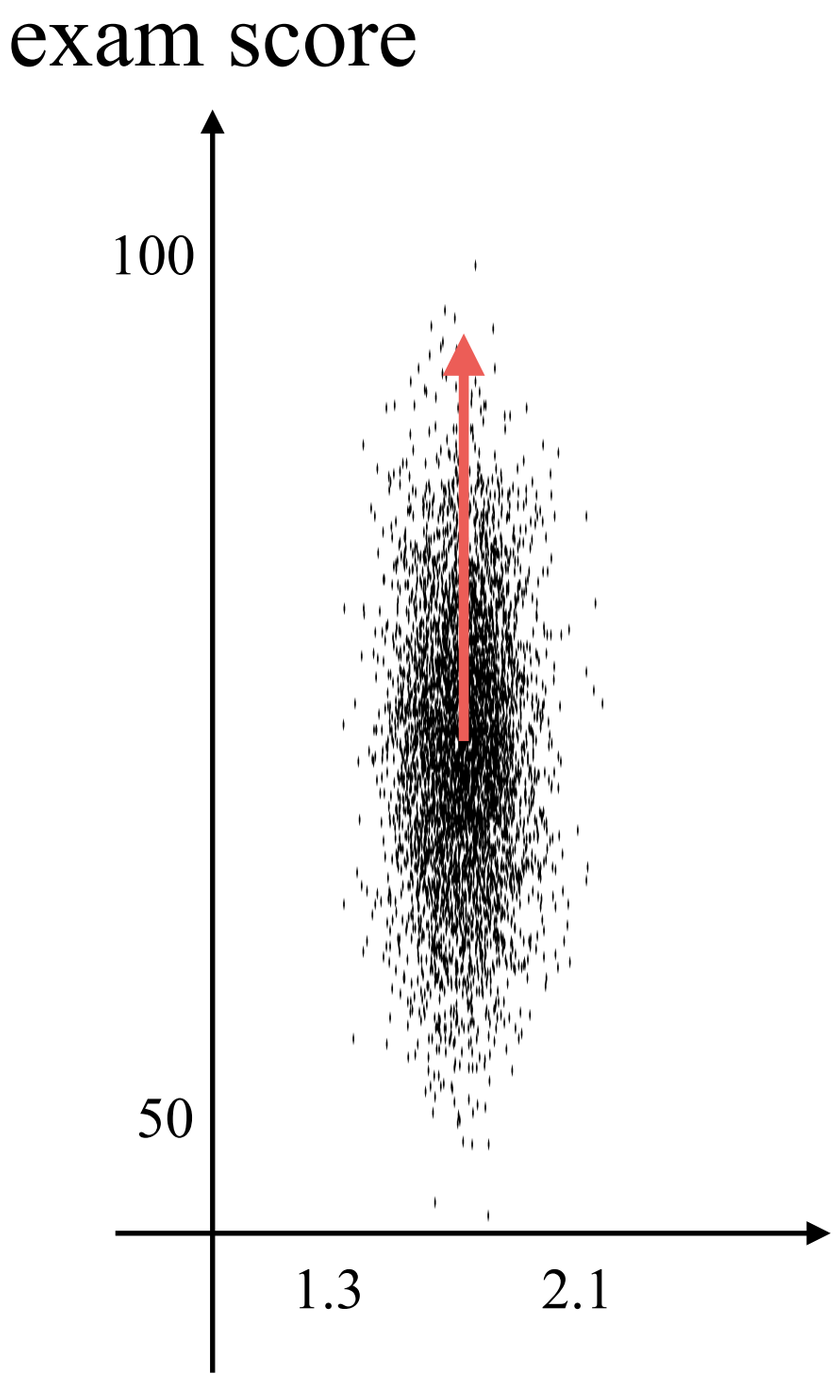
주성분 1

magic',	-1.13E-01
'difficulty',	-3.64E-02
'attack',	1.04E-01
'defense',	2.54E-02
'armorperlevel',	-7.62E-03
'mpperlevel',	-4.80E-02
'attackspeedoffset',	-3.39E-02
'hp',	6.99E-02
'attackspeedperlevel',	2.64E-02
'attackrange',	-1.05E-01
'attackdamageperlevel',	1.82E-02
'critperlevel',	1.01E-28
'spellblockperlevel',	1.32E-01
'crit',	0.00E+00
'spellblock',	1.31E-01
'attackdamage',	4.20E-02
'armor',	4.84E-02
'hpregenperlevel',	1.20E-02
'movespeed',	5.67E-02
'mpregenperlevel',	-5.23E-02

Achtung!

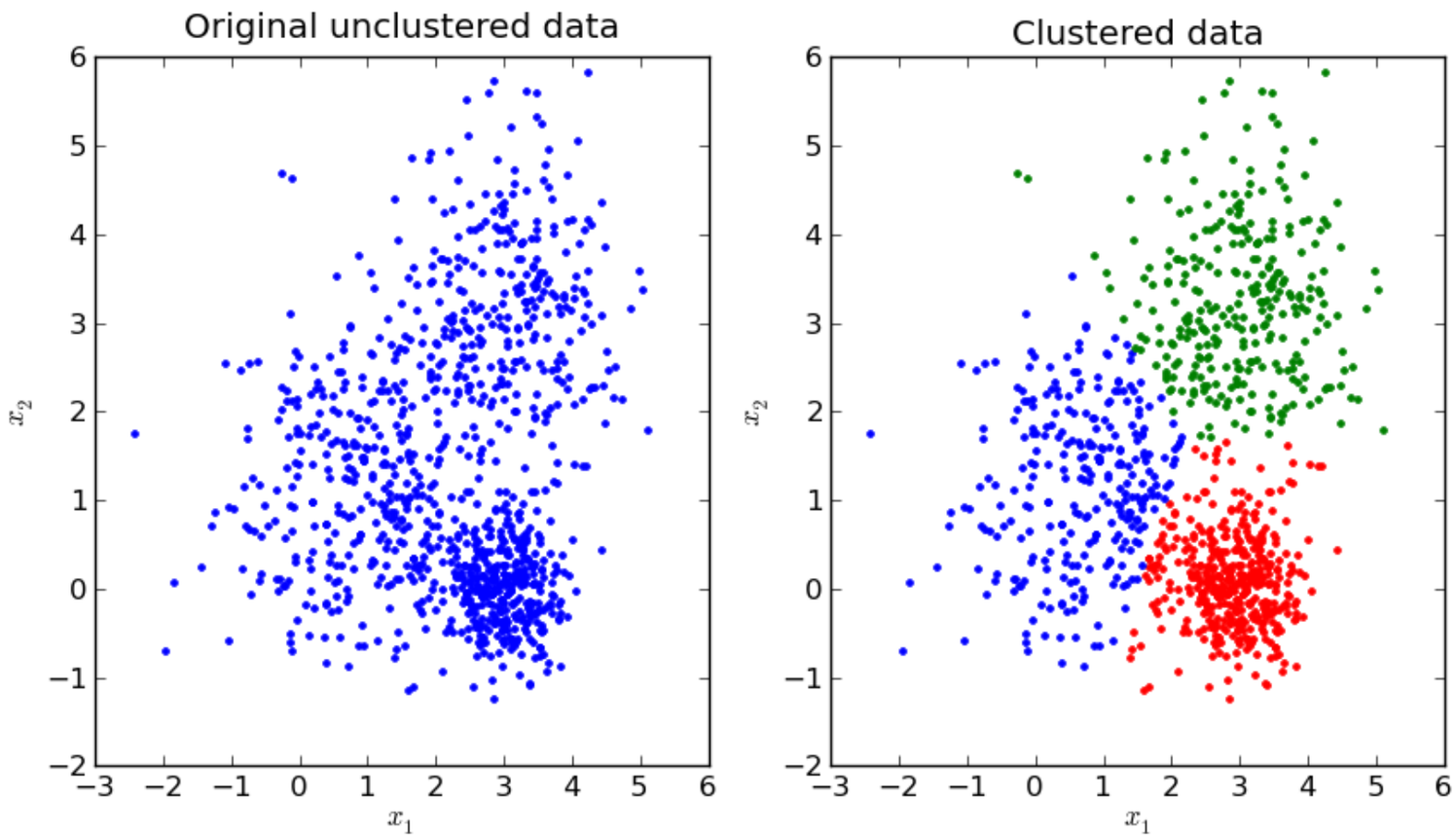
- 주성분 분석의 instability
  - 데이터의 scale에 크게 영향받음
  - solution: Normalization





클러스터링

주어진 데이터를 클러스터로 묶는 알고리즘



source: http://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue

클러스터링: 왜 사용하는가

- Unsupervised Learning (비지도학습)
- **레이블 없이** 데이터들을 묶어 (grouping) 데이터를 설명하는 클러스터를 찾고자 할 때

클러스터링: 왜 사용하는가

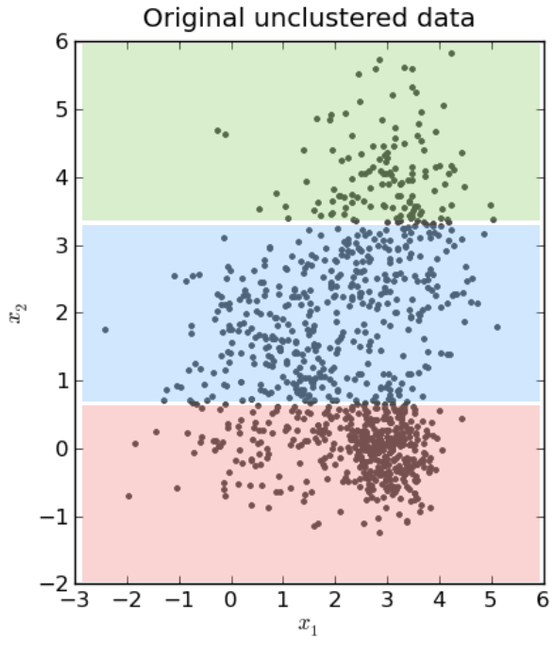
- **마케팅**: 비슷한 성향을 가진 고객들을 발견하고 분류해서 타깃 광고를 수행하고자 할 때
- **보험**: 고객들을 risk factor 에 따라 분류하여 사고 위험이 높은 고객들을 선별
- **도시 계획**: 집들을 모양, 가치, 위치에 따라 분류
- **뉴스 분석**: 뉴스 기사들을 토픽에 따라 분류 (bag-of-words)

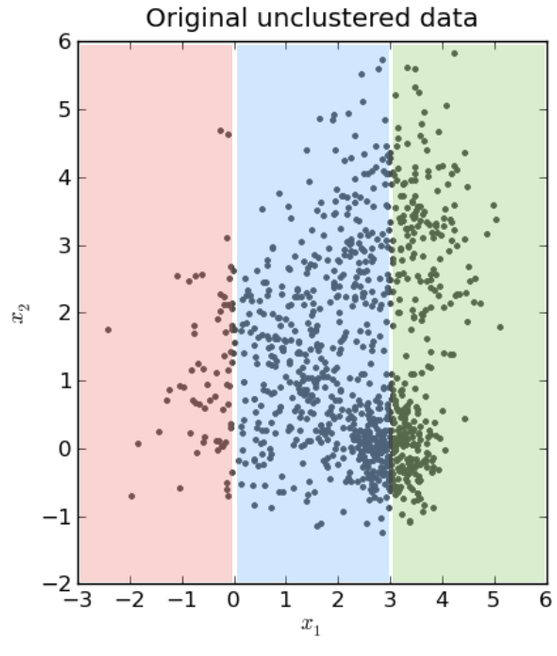
source: sabanciuniv 2002 CS525

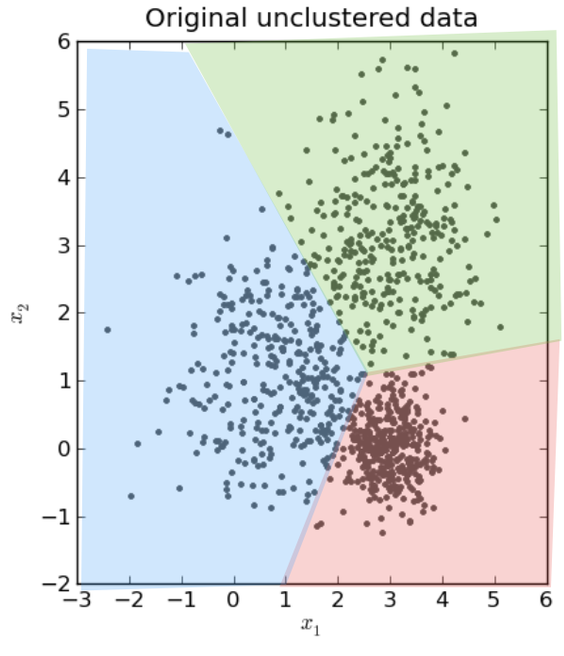
K-Means 클러스터링

주어진 데이터를 K개의 클러스터로 묶는 알고리즘

목표: 각 클러스터 내의 분산의 합을 최소화

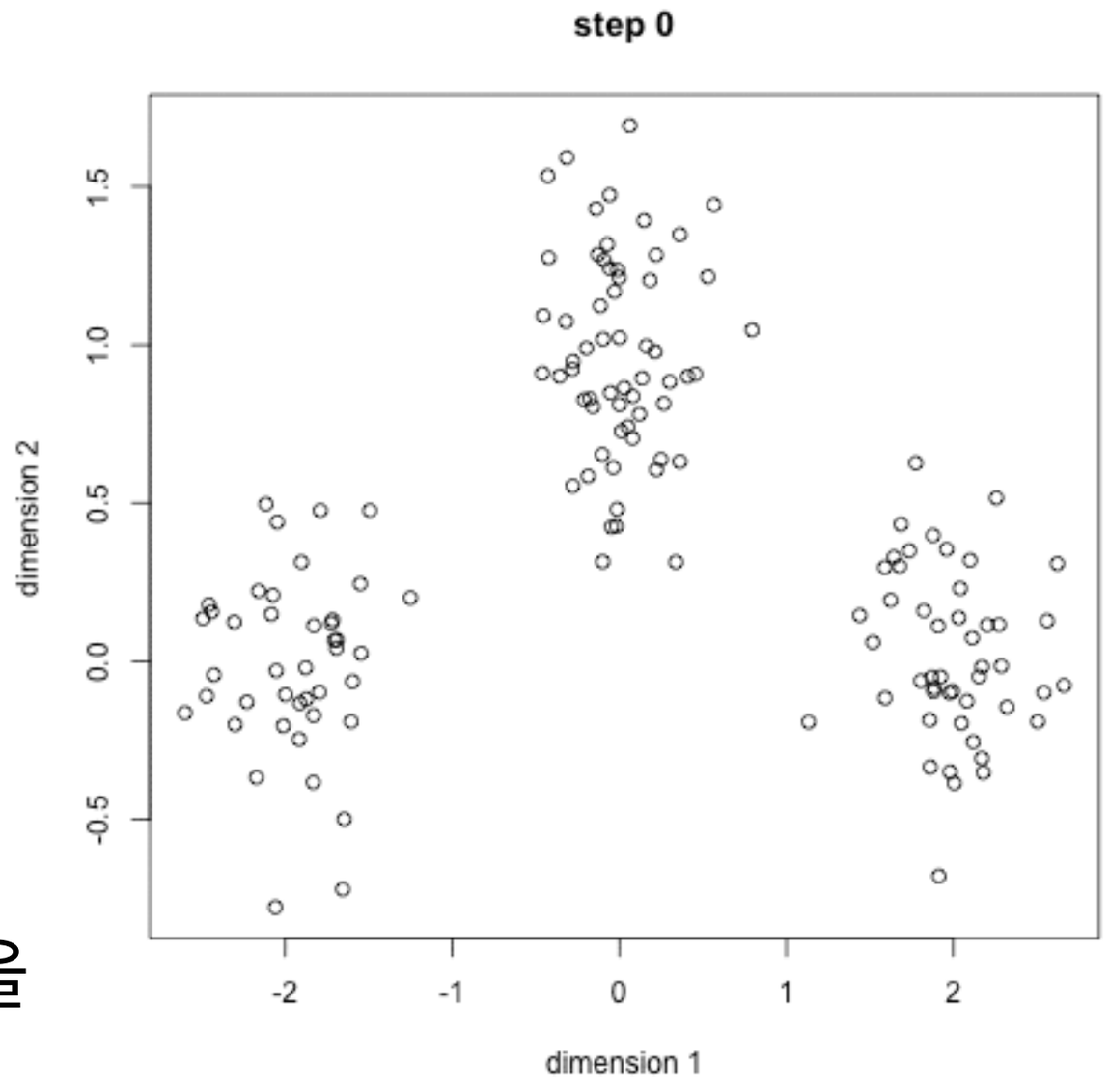






K-Means 클러스터링: 알고리즘

- **시작**: K개의 시작점을 각 클러스터의 중심으로 설정
- **반복**
  - 각각의 데이터 포인트를 그 포인트에서 가장 **가까운** (Euclidean distance) 클러스터 중심에 할당
  - 할당된 데이터 포인터의 중심을 클러스터 중심으로 재설정

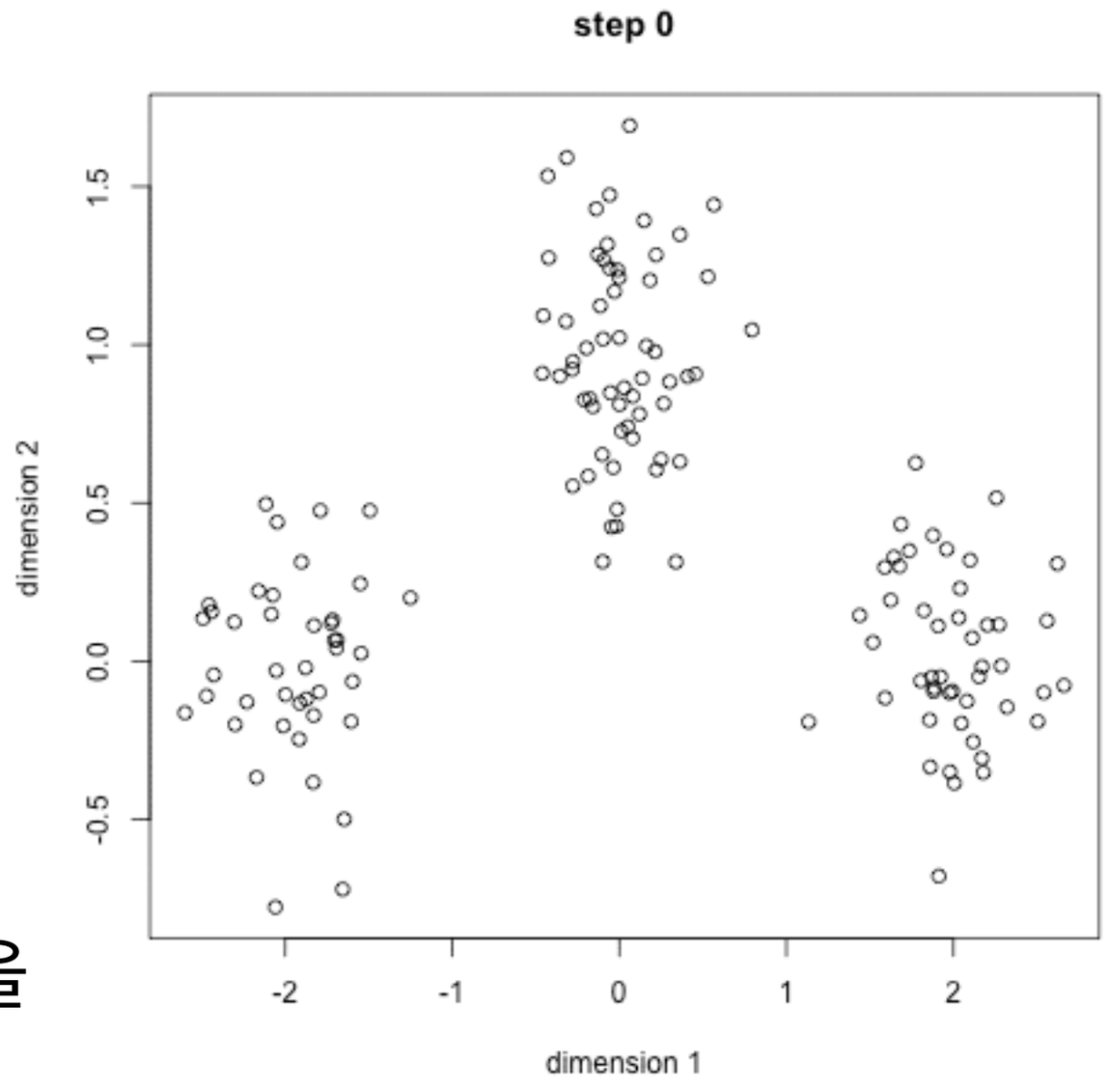


- **멈춤**: 데이터 포인트의 소속 클러스터가 바뀌지 않을 때

source: http://simplystatistics.org/2014/02/18/k-means-clustering-in-a-gif/

K-Means 클러스터링: 알고리즘

- **시작**: K개의 시작점을 각 클러스터의 중심으로 설정
- **반복**
  - 각각의 데이터 포인트를 그 포인트에서 가장 **가까운** (Euclidean distance) 클러스터 중심에 할당
  - 할당된 데이터 포인터의 중심을 클러스터 중심으로 재설정

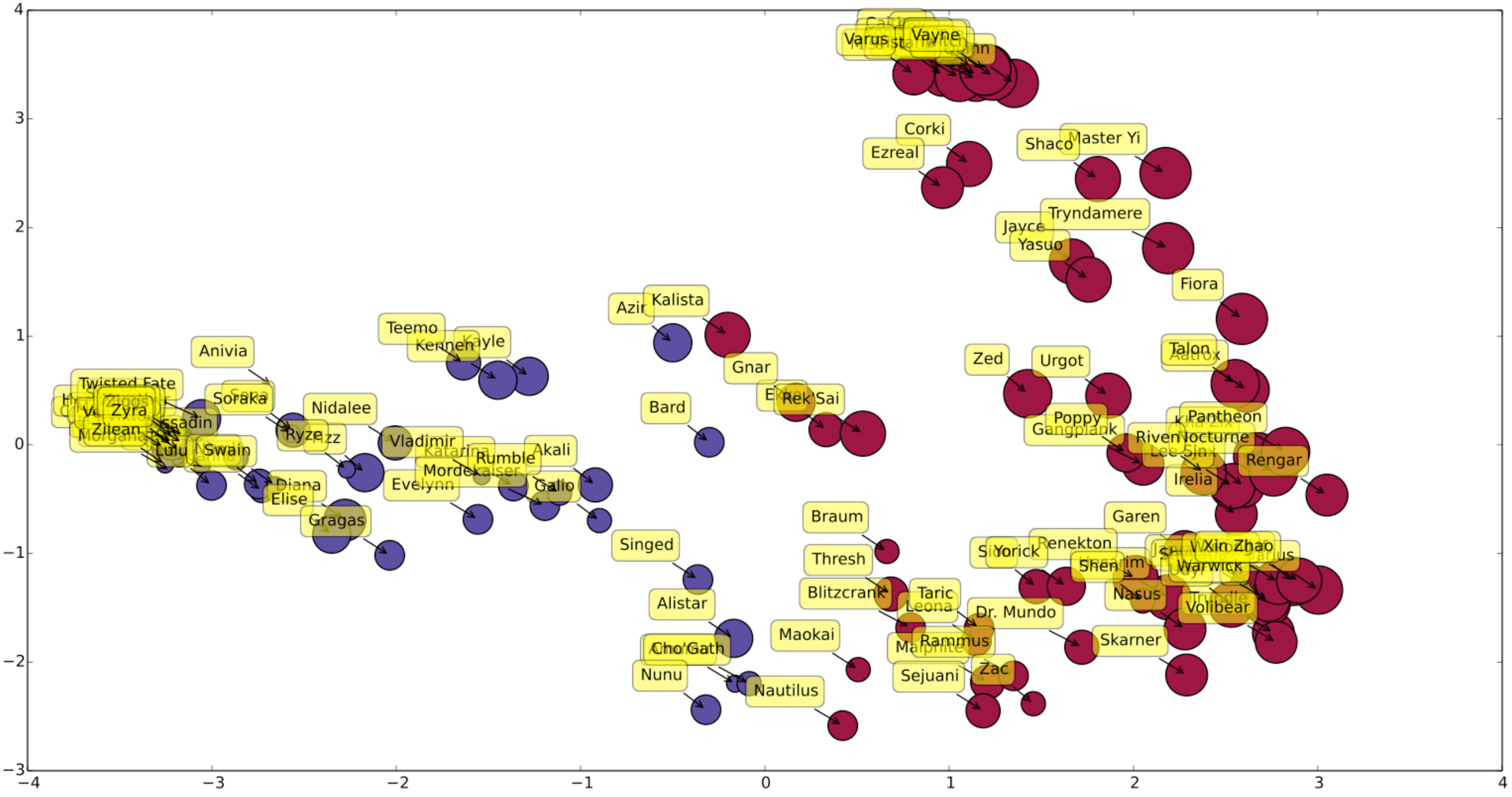


- **멈춤**: 데이터 포인트의 소속 클러스터가 바뀌지 않을 때

source: http://simplystatistics.org/2014/02/18/k-means-clustering-in-a-gif/

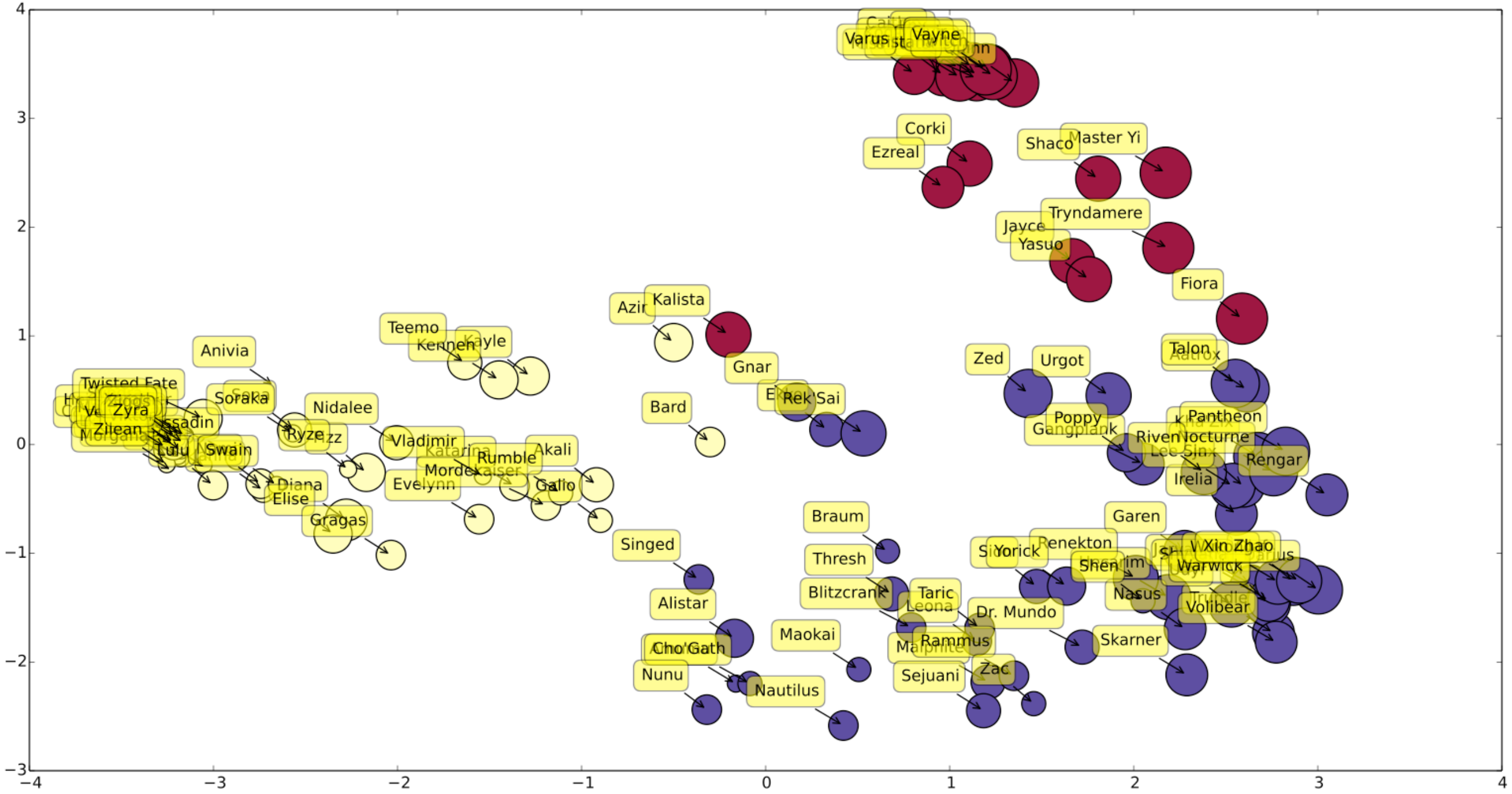
elice.io 에서 실행해보기

2차원 데이터 클러스터링 (PCA 적용: 174차원 → 2차원)



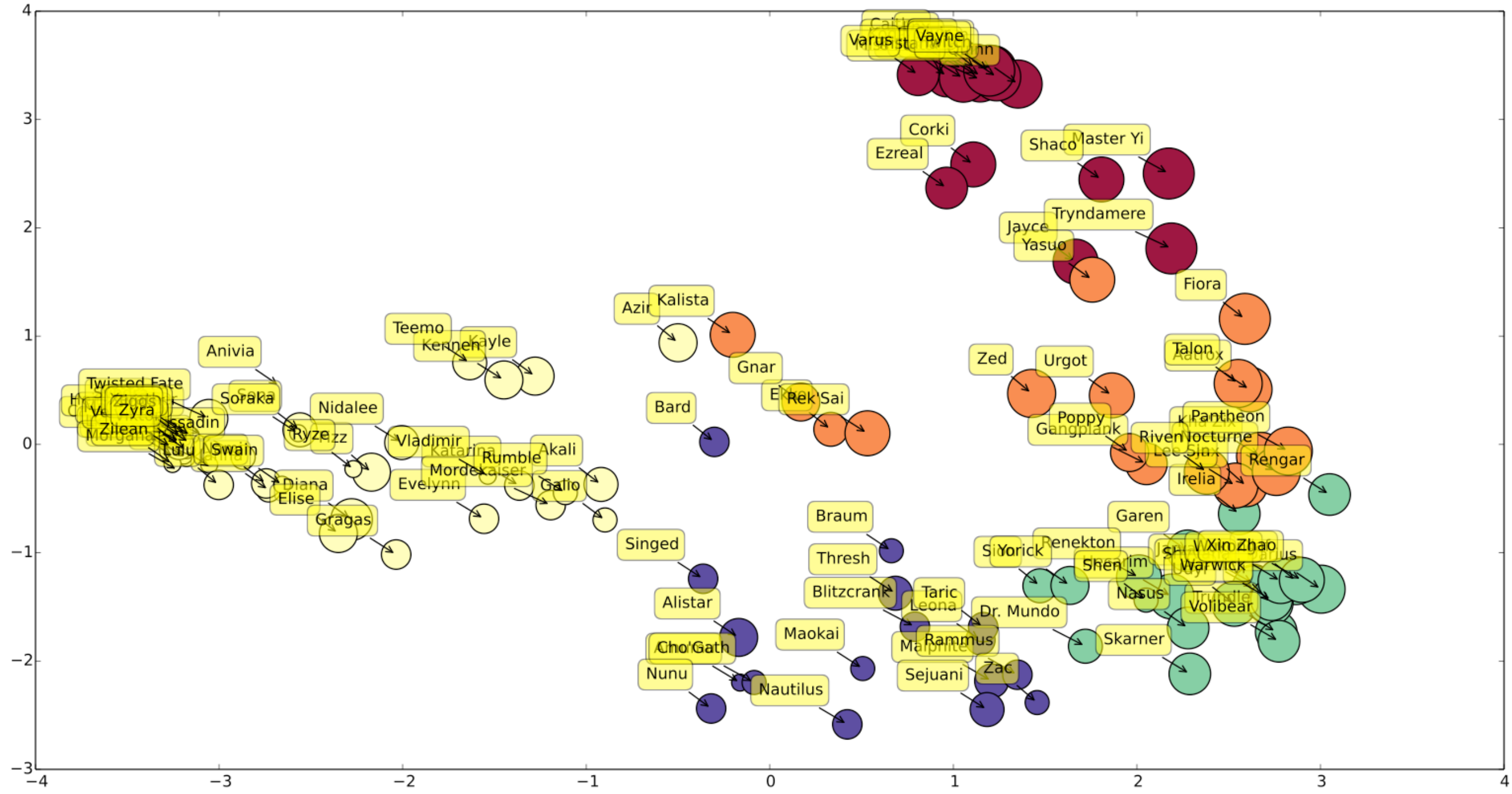
K = 2

elice.io 에서 실행해보기



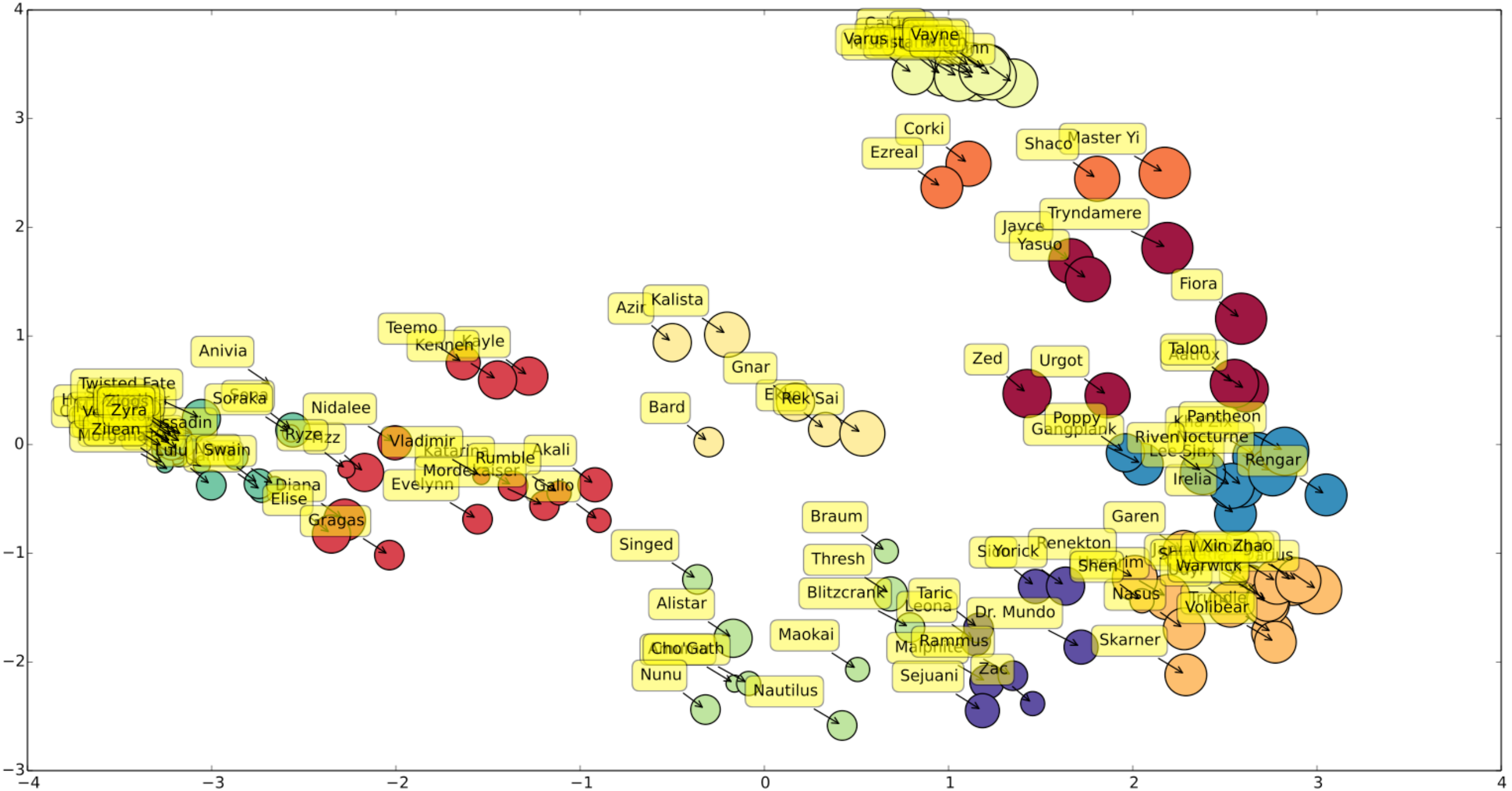
K = 3

elice.io 에서 실행해보기



K = 5

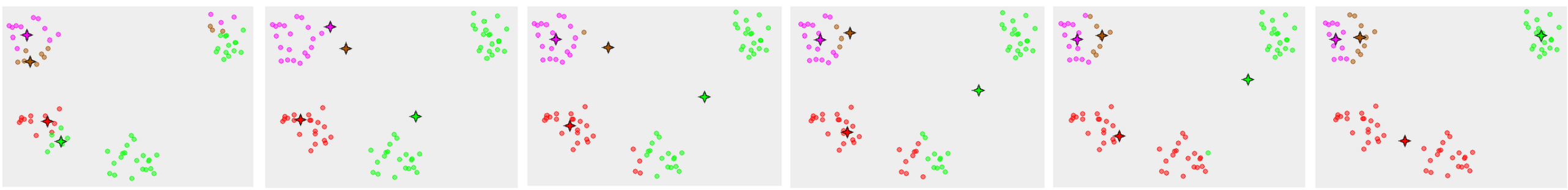
elice.io 에서 실행해보기



K = 10

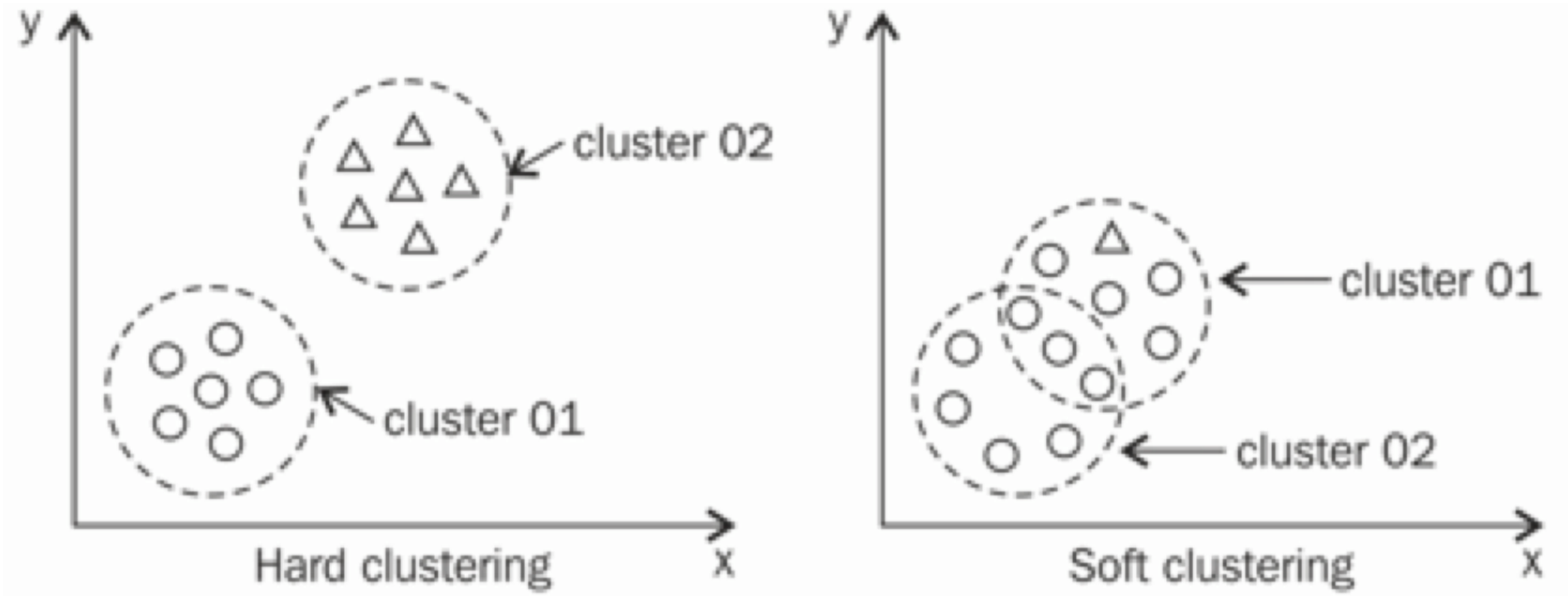
Discussion

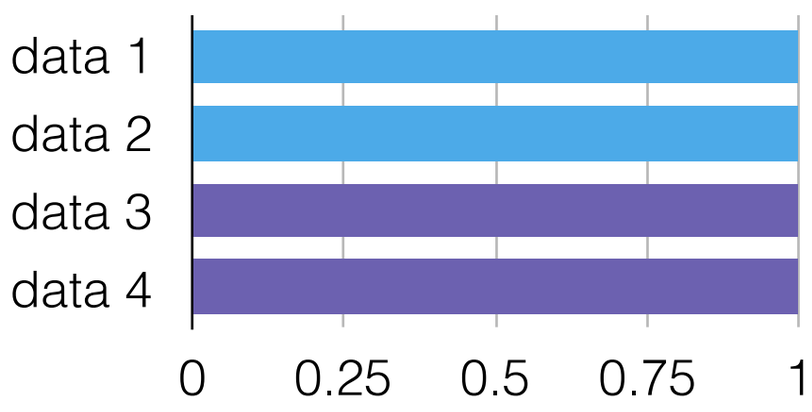
- K 의 수를 잘못 설정하면 성능이 나빠질 수 있음
  - Soft clustering
  - Non-parametric density estimation
- Local minima에 수렴할 경우 빠져나오지 못할 수 있음
- 시작점을 설정하는 다양한 방법

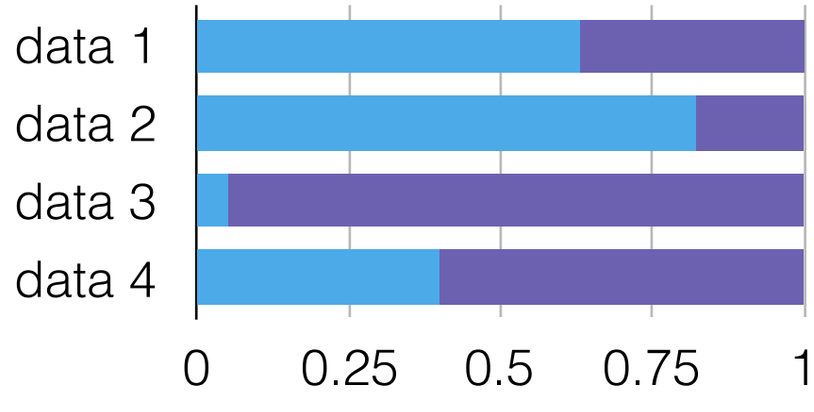


source: https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence_to_a_local_minimum.png

Hard vs. Soft Clustering







Gaussian Mixture Model

source: http://blog.guillaumeagis.eu/k-means-clustering-apache-mahout/