

# 데이터 과학

## L12: Principal Component Analysis

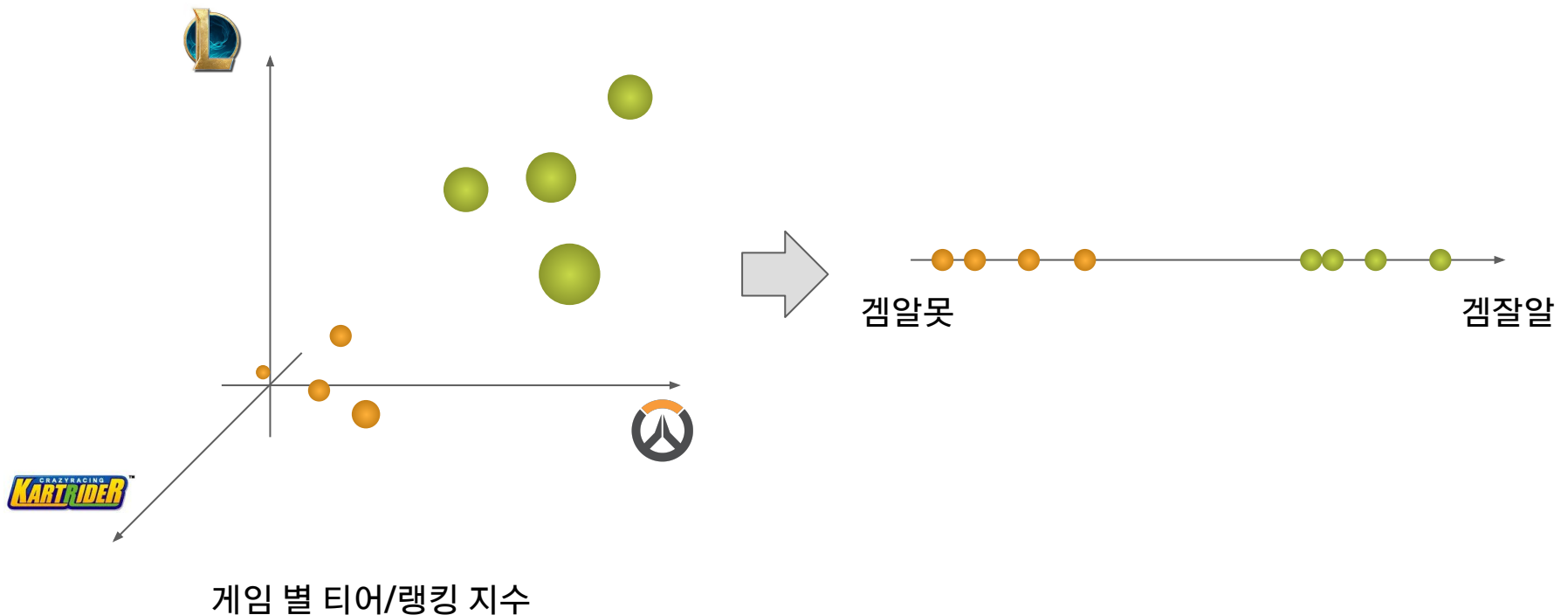
Kookmin University

# Principal Component Analysis

- 주 성분 분석

- 데이터의 분포를 결정하는 핵심 성분 찾기

- 예) 원래 데이터: 게임별 티어 → 주 성분: 게임DNA

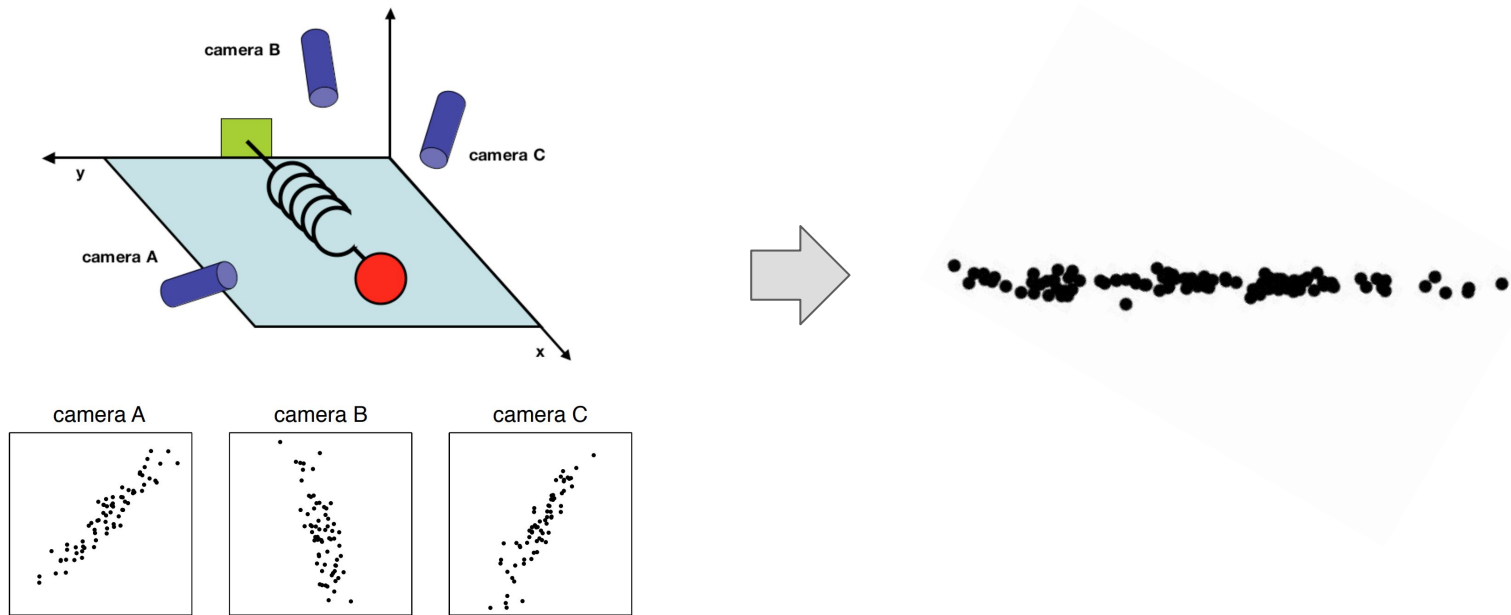


# Principal Component Analysis

- 주 성분 분석

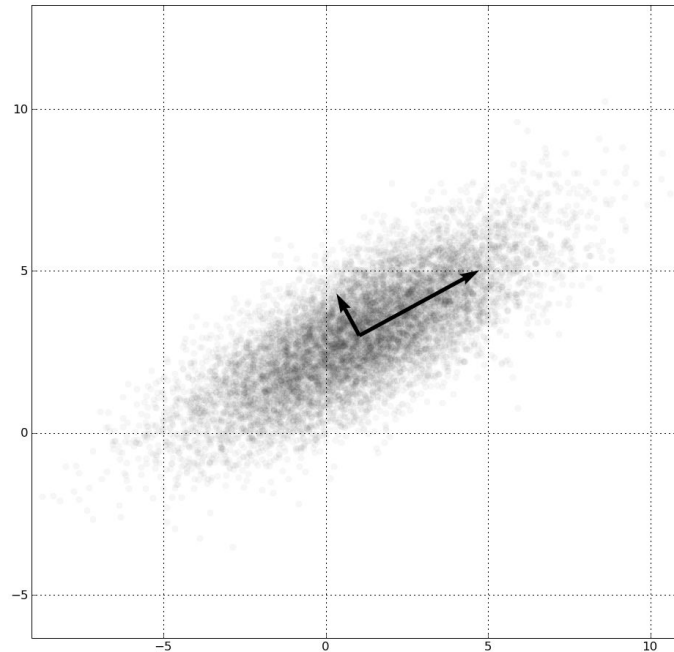
- 데이터의 분포를 결정하는 핵심 성분 찾기

- 예) 원래 데이터: 게임별 티어 → 주 성분: 게임DNA
    - 예) 원래 데이터: 카메라별 공의 위치 → 주 성분: 스프링의 힘



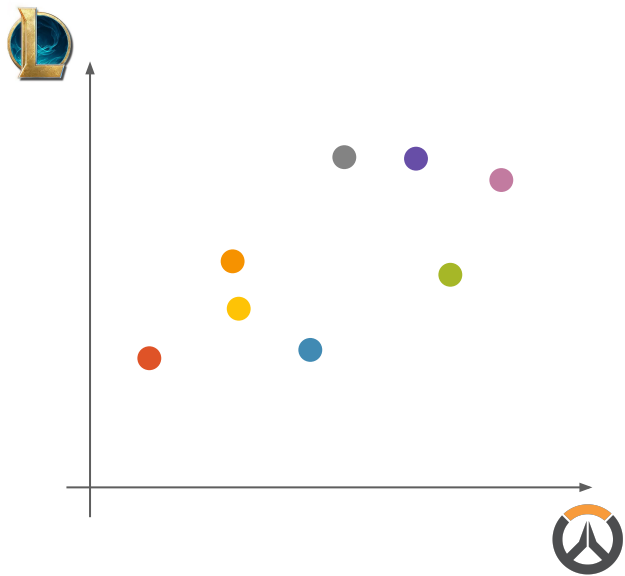
# Principal Component Analysis











- 주 성분 분석
  - 분산을 최대화 하면서 서로 직교하는 새로운 축을 찾음



# 차원 축소

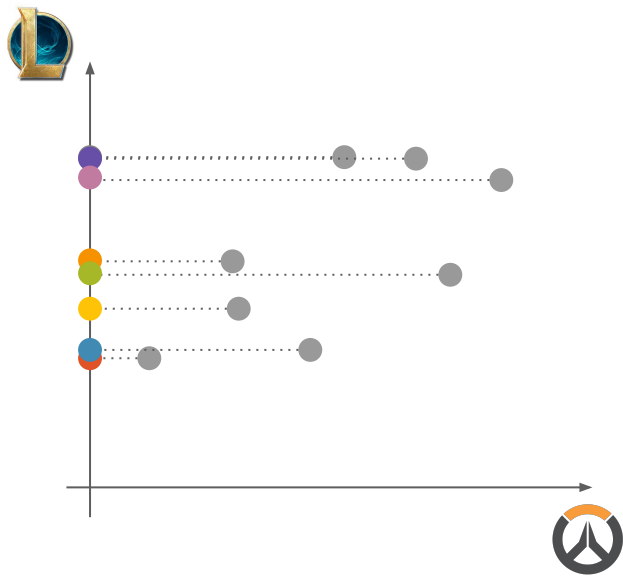
- 차원 축소 방법





								
	A	B	C	D	E	F	G	H
	3.1	3.4	4.6	3.2	7.9	7.8	4.4	7.5
	1.0	4.2	4.0	5.7	6.2	8.1	9.3	9.9

# 차원 축소

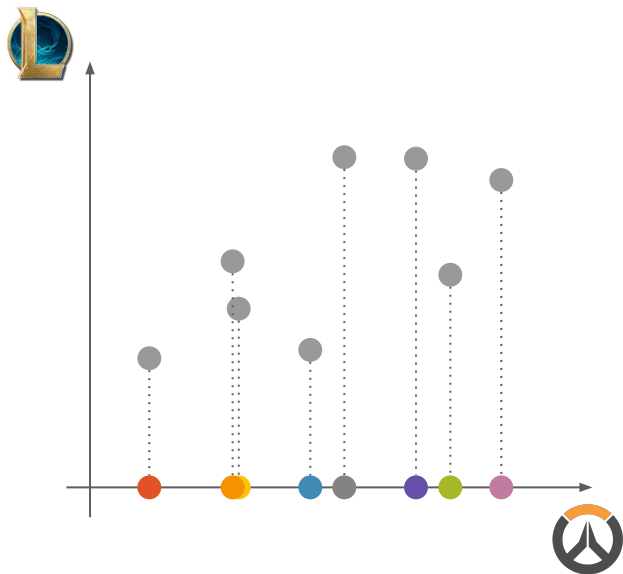
- 차원 축소 방법
  - 방법1. 아무 차원이나 지운다.





	A	B	C	D	E	F	G	H
	3.1	3.4	4.6	3.2	7.9	7.8	4.4	7.5
	1.0	4.2	4.0	5.7	6.2	8.1	9.3	9.9

# 차원 축소

- 차원 축소 방법
  - 방법1. 아무 차원이나 지운다.
    - 어떤 차원을 지우는 것이 더 좋은가?



	A	B	C	D	E	F	G	H
	3.1	3.4	4.6	3.2	7.9	7.8	4.4	7.5
	1.0	4.2	4.0	5.7	6.2	8.1	9.3	9.9

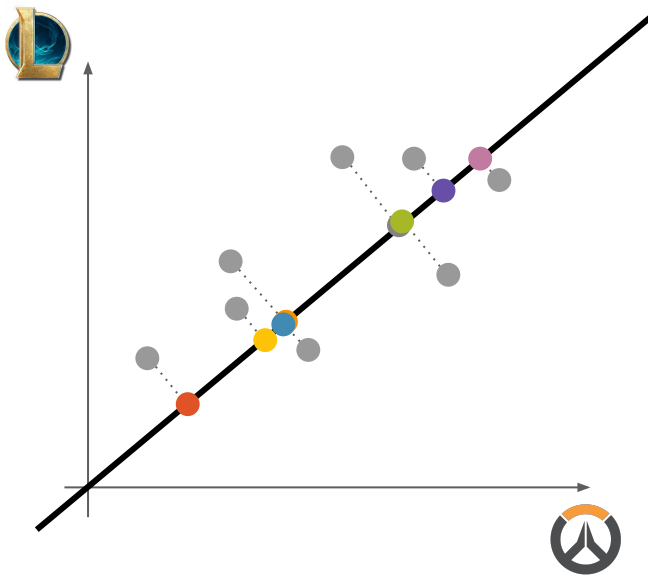
# 차원 축소



- 차원 축소 방법

- 방법2. 새로운 축(선분)을 찾는다. = 주 성분 찾기

- 분산을 최대로..!

- 어떻게 찾아..? → Gradient Descent 등 활용



	A	B	C	D	E	F	G	H
	3.1	3.4	4.6	3.2	7.9	7.8	4.4	7.5
	1.0	4.2	4.0	5.7	6.2	8.1	9.3	9.9
?	3.3	5.4	6.1	6.5	10.0	11.2	10.3	12.4

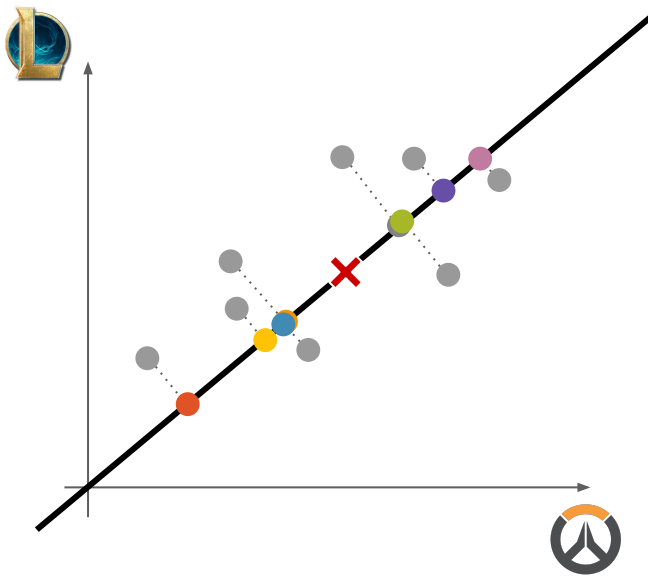
<http://i.imgur.com/Uv2dlsH.gif>



# 주성분 찾기

- 분산 구하기

$$\text{var}(X) = E((X - \mu)^2)$$



?

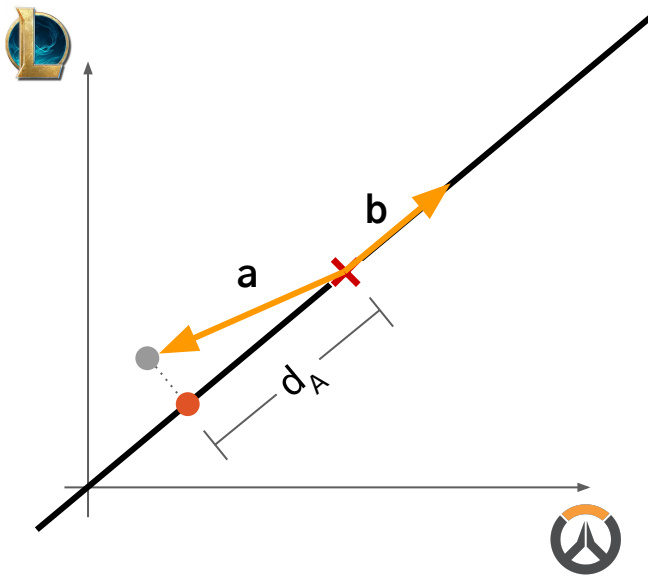
3.3	5.4	6.1	6.5	10.0	11.2	10.3	12.4
-----	-----	-----	-----	------	------	------	------

✗ 평균 = 8.2

# 주성분 찾기

- 분산 구하기

$$\text{var}(X) = E((X - \mu)^2)$$



?

A	B	C	D	E	F	G	H
3.3	5.4	6.1	6.5	10.0	11.2	10.3	12.4

✗ 중심점, 평균 = 8.2

$$d_A = |a \cdot b| = |\text{값} - \text{평균}|$$

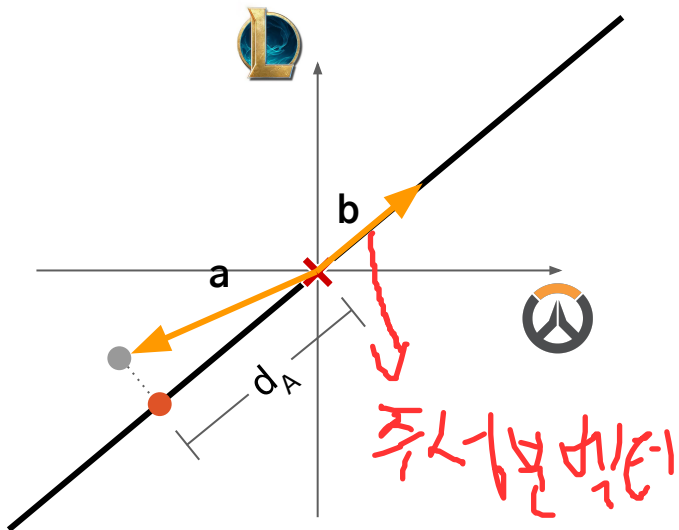
단, b는 단위벡터

# 주성분 찾기

- 분산 구하기

- 축을 옮기면 분산을 구하기 더 구하기 쉬워진다.

$$\text{var}(X) = E((X - \cancel{\mu})^2)$$



Eigen vector of PC1

Eigen value of PC1 (분산)

?

A	B	C	D	E	F	G	H
-4.9	-2.8	-2.1	-1.6	1.9	3.1	2.1	4.3

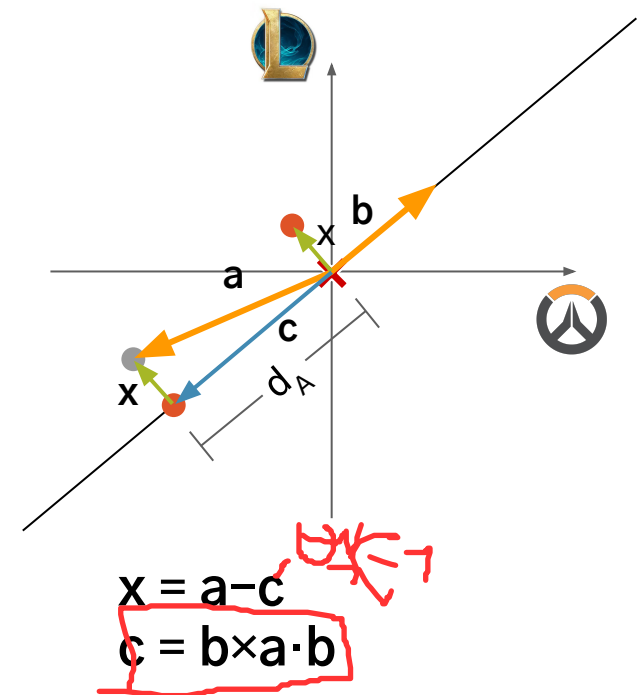
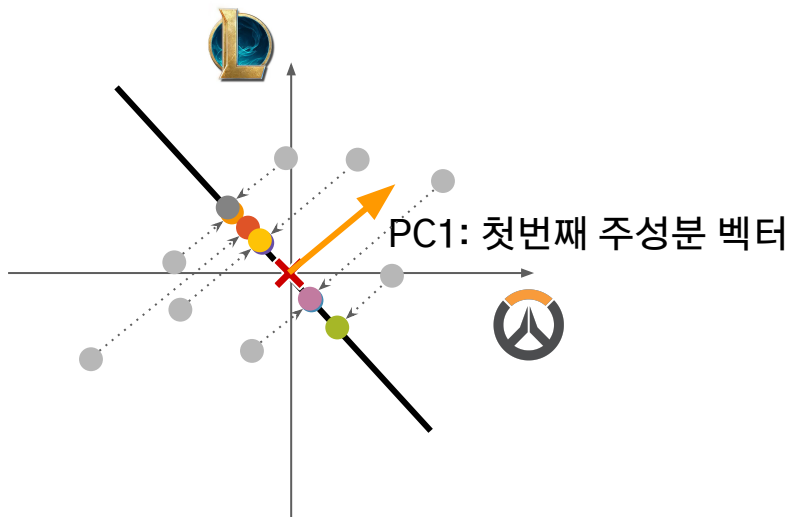
× 중심점, 평균 = 0

$$d_A = |a \cdot b| = |\text{값} - \cancel{\text{평균}}|$$

단, b는 단위벡터

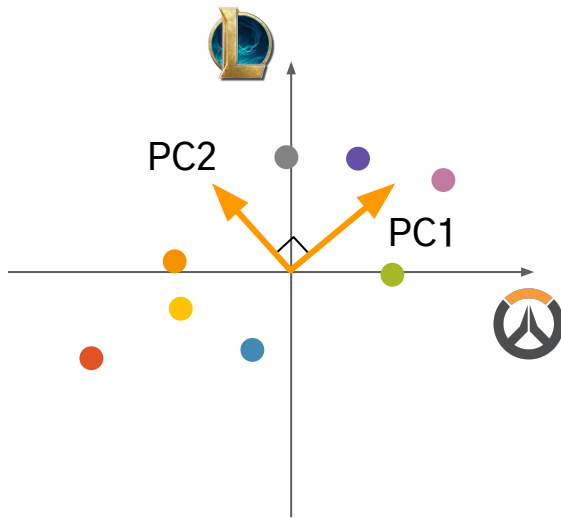
# 주성분 찾기

- 두 번째 주성분 찾기
  - 데이터에서 첫 번째 주성분을 제거 → 다시 주성분 탐색



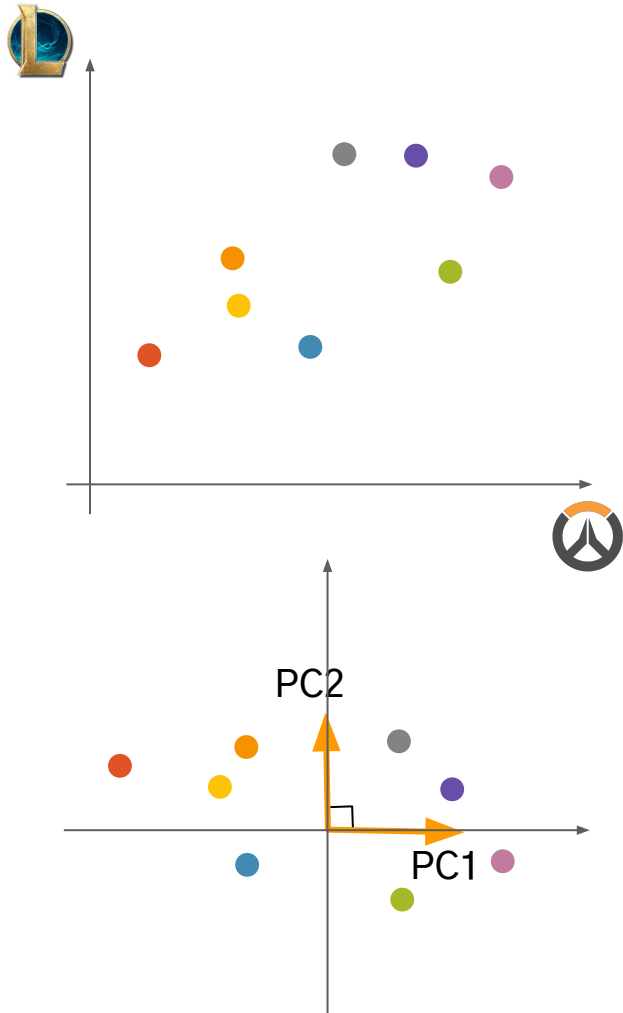
# 주성분 찾기











- 두 번째 주성분 찾기
  - 데이터에서 첫 번째 주성분을 제거 → 다시 주성분 탐색



Q.  $n$ 차원 데이터에서 주성분의 수는?

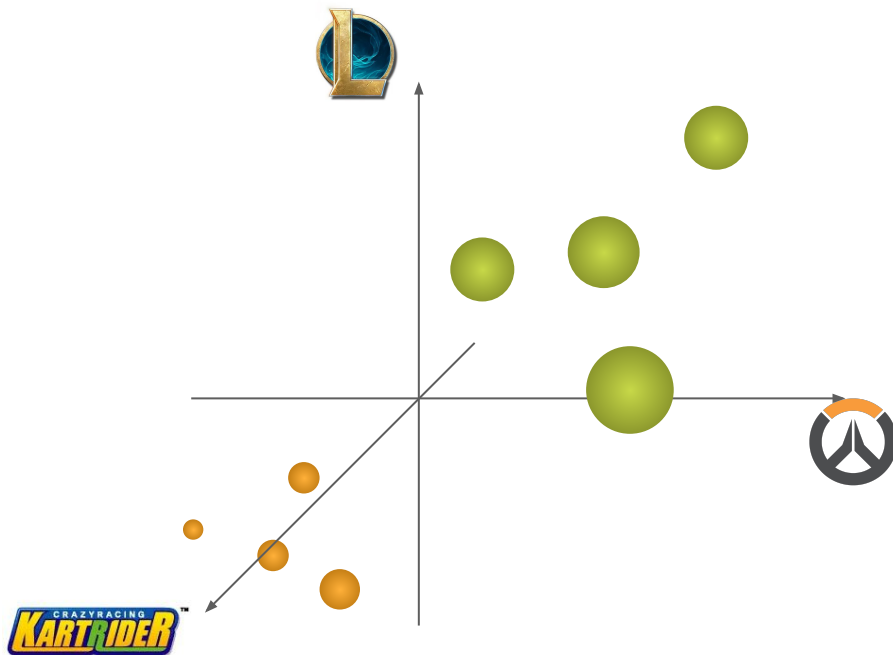
# 주성분으로 데이터 표현



	 A	 B	 C	 D	 E	 F	 G	 H
	3.1	3.4	4.6	3.2	7.9	7.8	4.4	7.5
	1.0	4.2	4.0	5.7	6.2	8.1	9.3	9.9

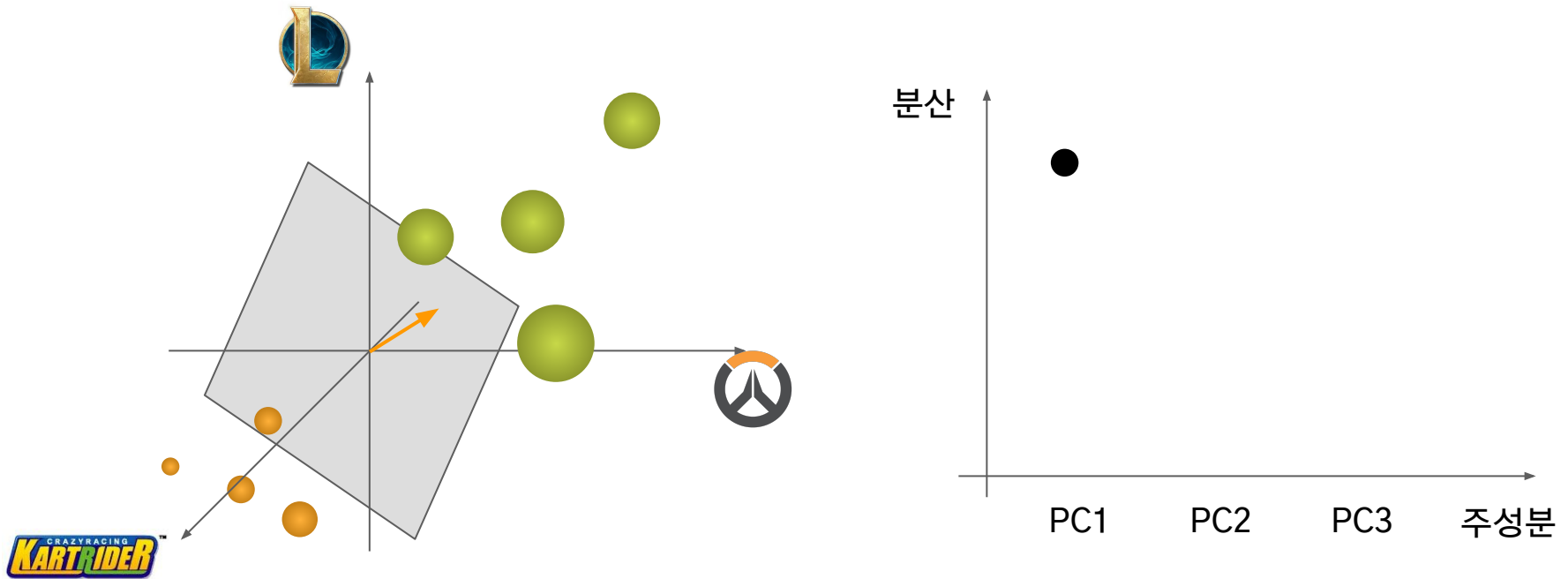
	A	B	C	D	E	F	G	H
PC1	-4.9	-2.8	-2.1	-1.6	1.9	3.1	2.1	4.3
PC2	1.1	0.7	1.5	-0.3	1.7	0.3	-0.9	-0.3

# 3차원 예시



# 3차원 예시

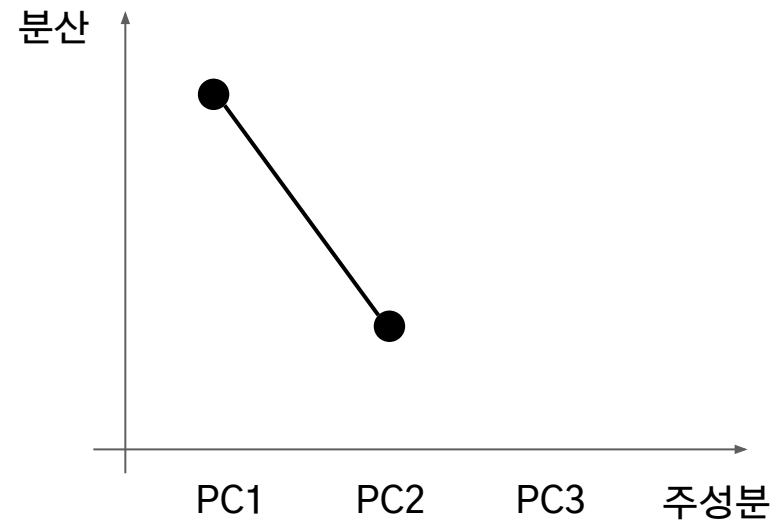
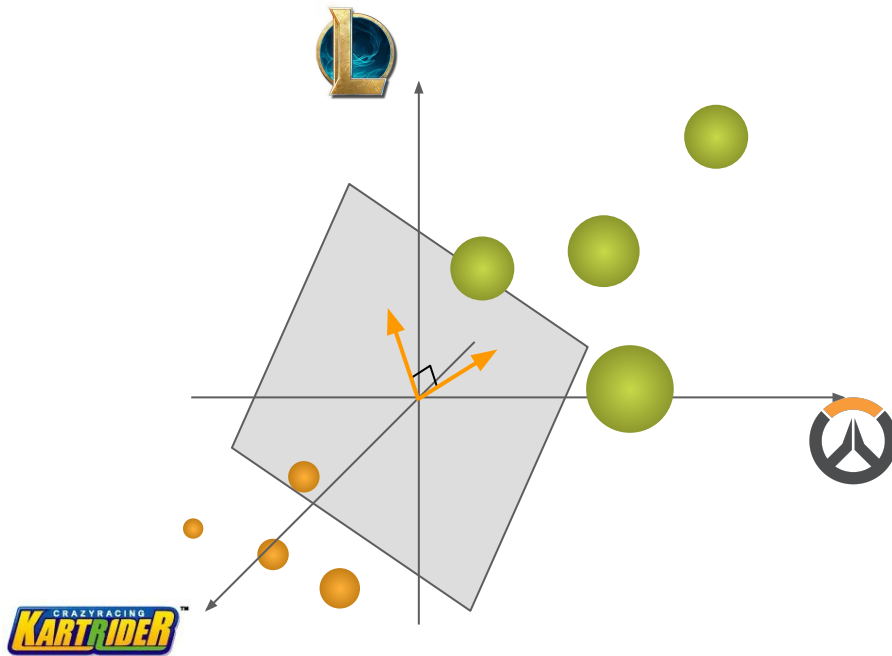
- PC1 찾기: 사영했을 때 분산이 가장 커지는 벡터





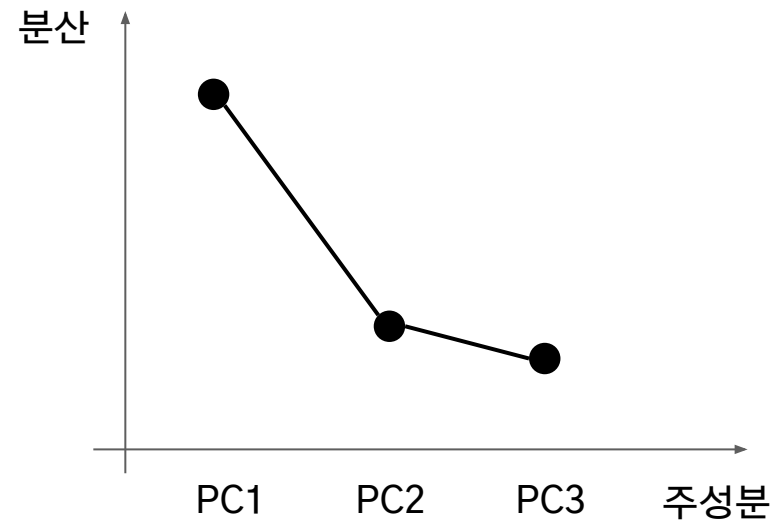
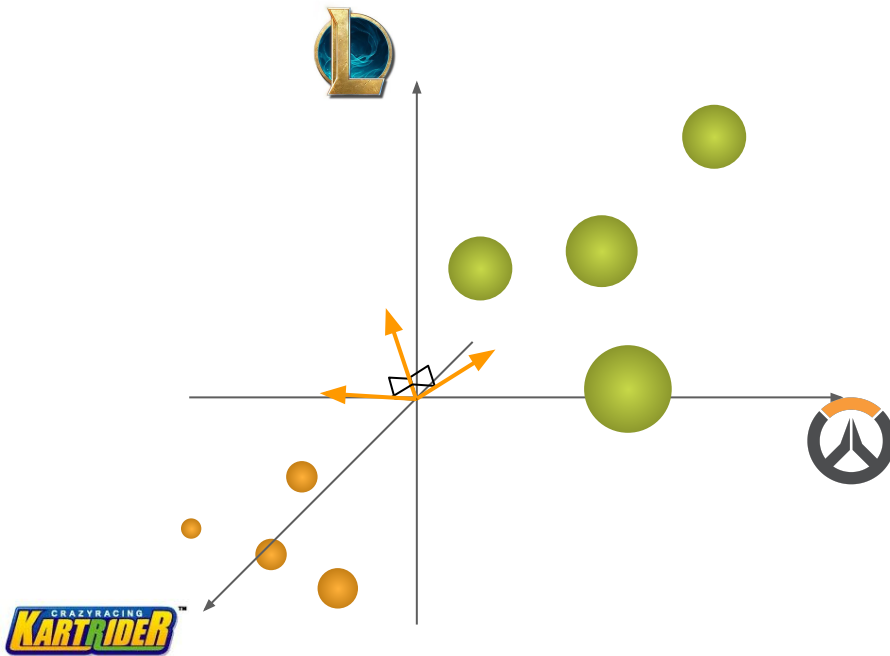
# 3차원 예시

- PC1의 직교평면에서 PC2 찾기



# 3차원 예시

- PC1과 PC2에 모두 직교하는 벡터 = PC3

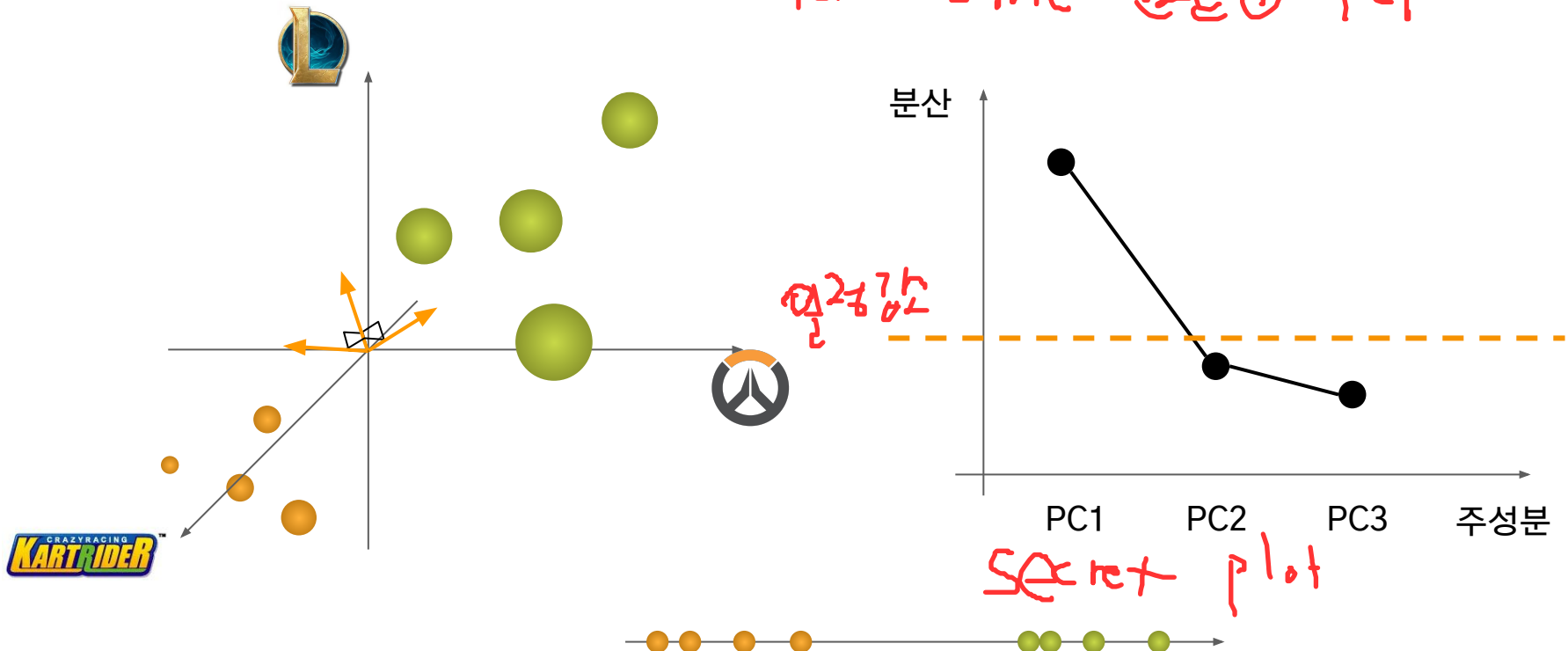


# 3차원 예시

PC가 각분수를 분할이 쉽다!

- PC1과 PC2에 모두 직교하는 벡터 = PC3

low rank = 분산 ↓ 구조는 PC



$$\overset{\substack{\uparrow \\ \text{matrix}}}{A} v = \lambda v \leftarrow \text{eigen vector}$$

$\uparrow$   
eigen value

## Questions?