

Vector Data 이해하기

신한별

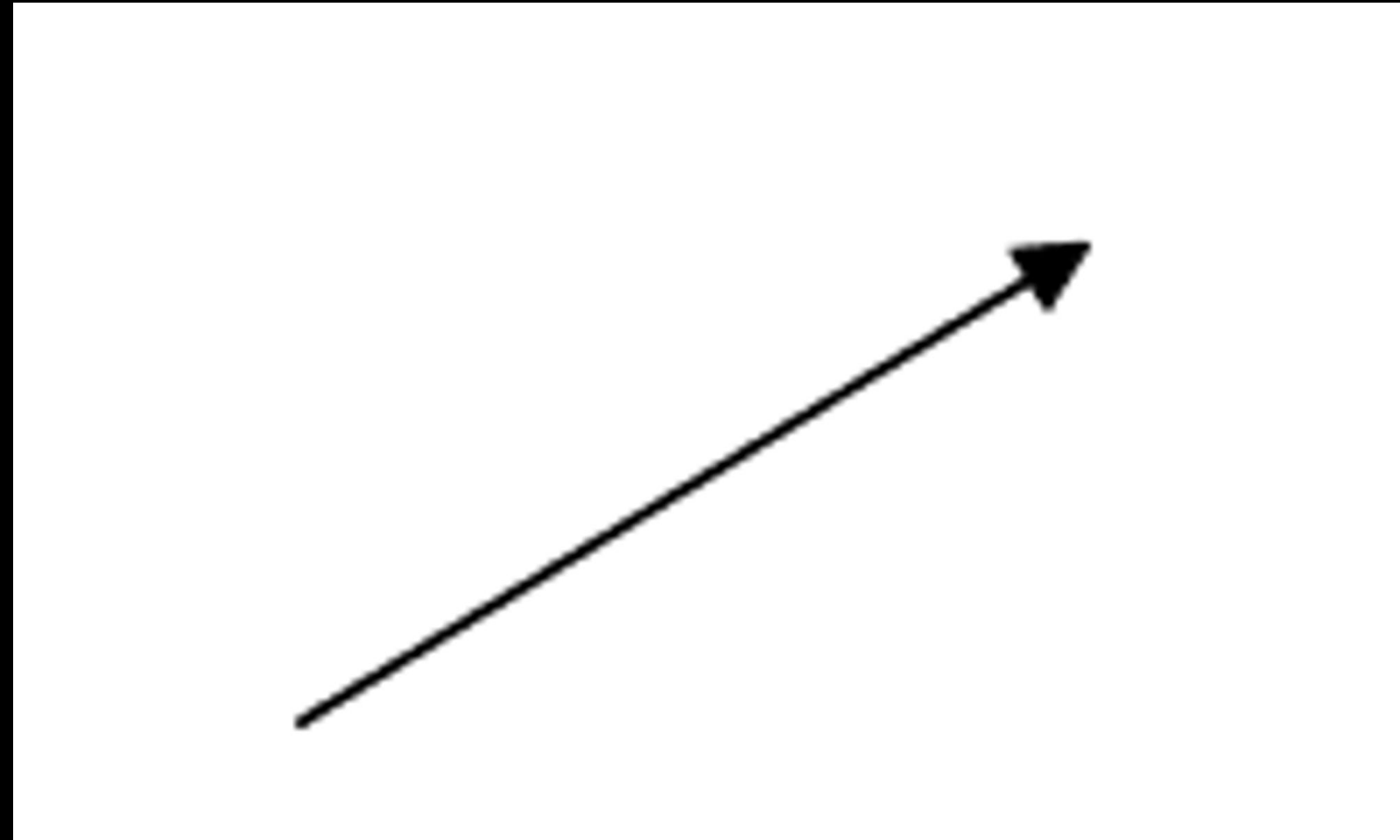
<https://github.com/shinhanbyeol>

Table of Contents

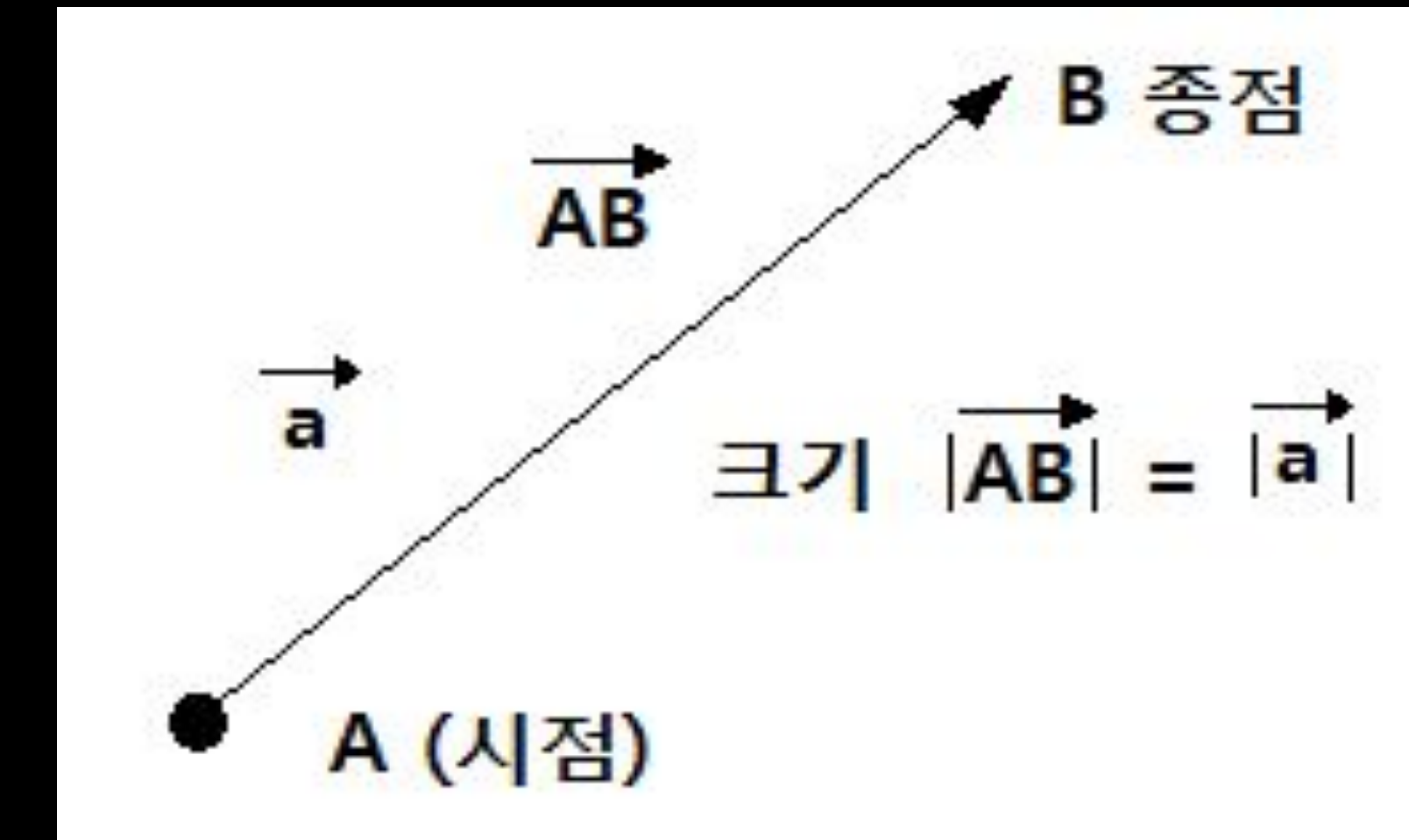
- 01 **Vector** 란 무엇인가?
- 02 **Vector** 데이터 활용사례
- 03 **AI** 가 **Vector data** 를 이용하는 법
- 04 **Vector** 데이터 이용해보기

01 **Vector** 란 무엇인가?

Vector 의 어원 및 의미



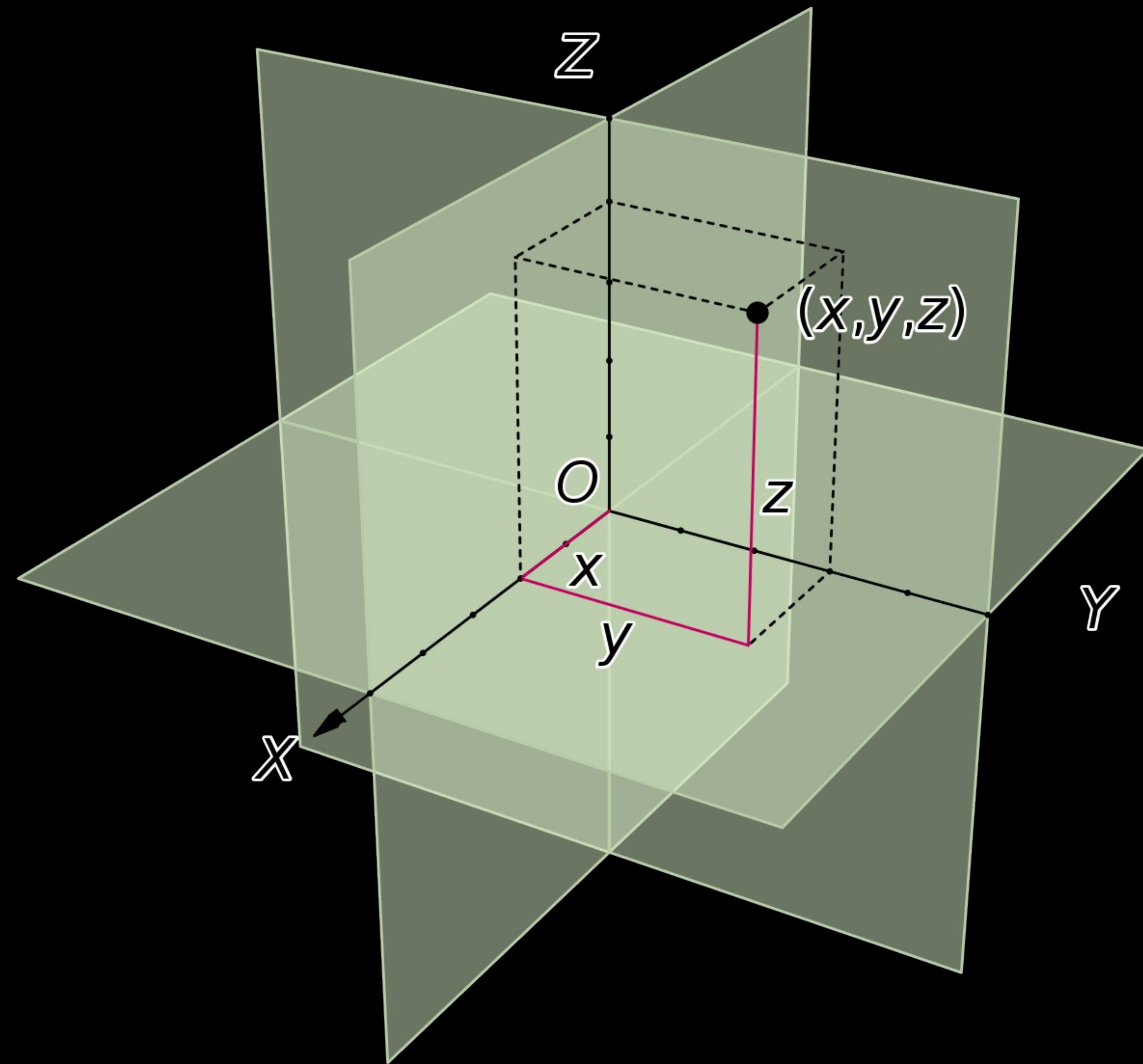
벡터의 어원은 ‘운반하다’라는 뜻의 라틴어에서 비롯되었으며, 크기와 방향을 갖는 물리량을 의미합니다.



우리가 통상적으로 이야기하는 벡터는 유클리드 벡터 라고 하며, 유클리드 공간에서의 기본 단위를 벡터라고 합니다. 벡터는 보통 3차원 벡터일 경우 (x,y,z) 2차원 벡터일 경우 (x,y) 등의 원소 데이터를 가지게 됩니다.

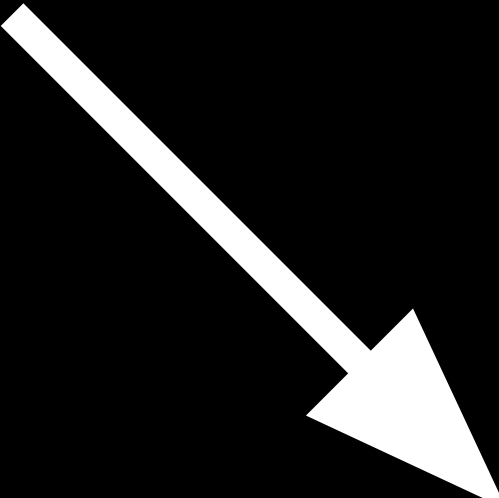
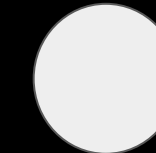
벡터는 점데이터 뿐만 아니라 원점 또는 시점으로부터 유클리드 거리를 통해 힘의 크기 또는 방향 등을 표현 할 수도 있습니다.

유클리드 공간이란?



유클리드 공간은 기하학 원론의 저자인 유클리드의 이름을 따온 공간으로, 우리가 살고 있는 3차원 공간을 포함해서 평면, 수직선은 물론 그 이상의 다차원 공간까지 표현하는 공간입니다.

Vector data 관

Vector data =  =  = $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$

02 **Vector** 데이터 활용 사례

Vector 데이터 활용사례: GIS



GIS(지리 정보 시스템)은 AI 나 Vector database 가 유행 하기도 전부터 전통적으로 Vector 데이터를 활용하던 시스템 입니다.

지리 정보 시스템은 대체로 위도, 경도 데이터를 다루고 있으며 이는 Vector 의 형태로 데이터를 처리하는 사례중 가장 보편적인 사례 일 것입니다.

GIS 는 현실세계의 지리 정보를 데이터화 구조화 해서 저장하는 역할 뿐 아니라 활용가치가 있는 데이터들과 연계하여 vector 에 다양한 속성을 부여 할 수도 있습니다.

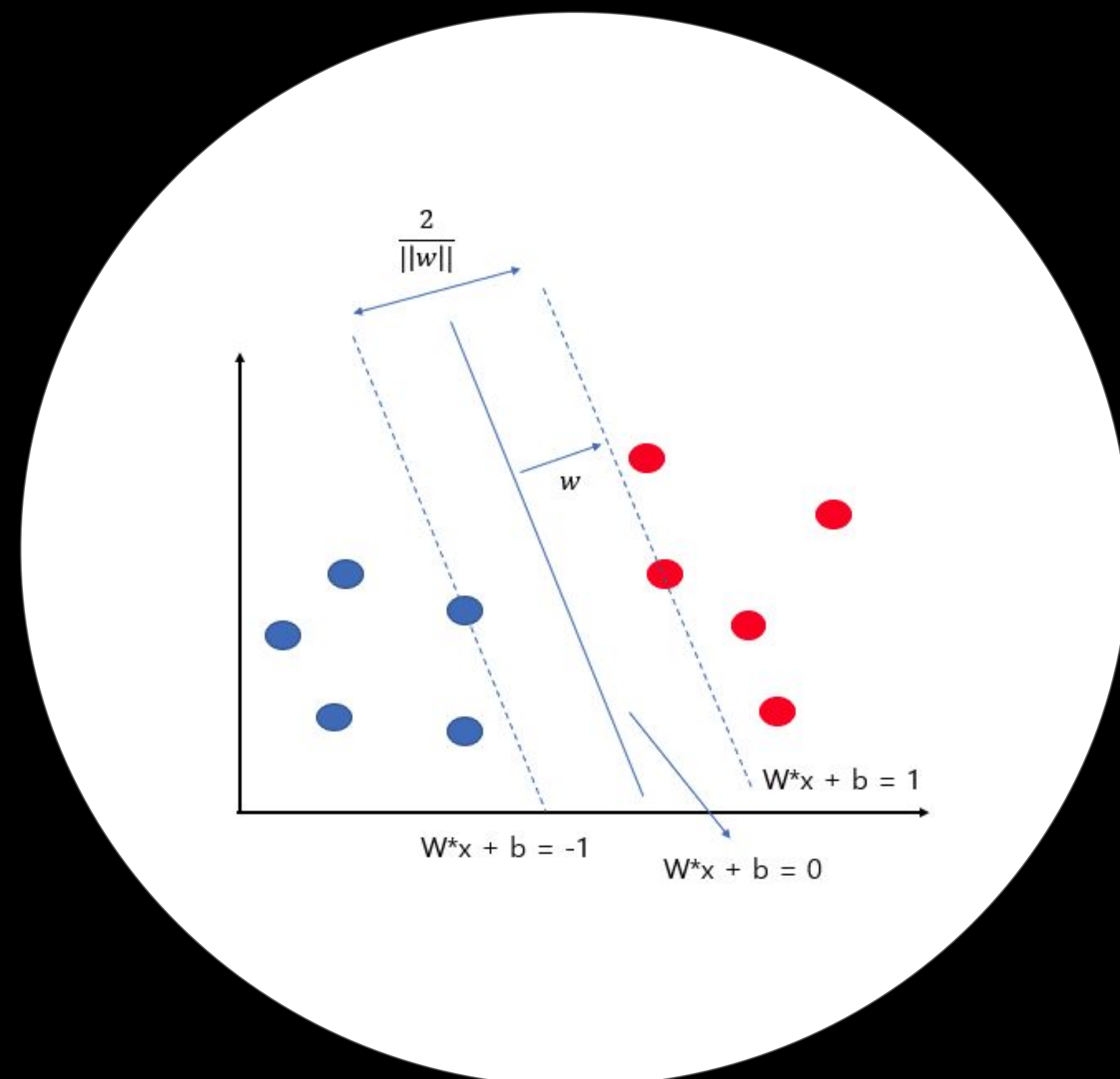
Vector 데이터 활용사례: 서포트 벡터 머신 (SVM)

서포트 벡터 머신(support vector machine, SVM)은 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 **분류**와 **회귀 분석**을 위해 사용한다.

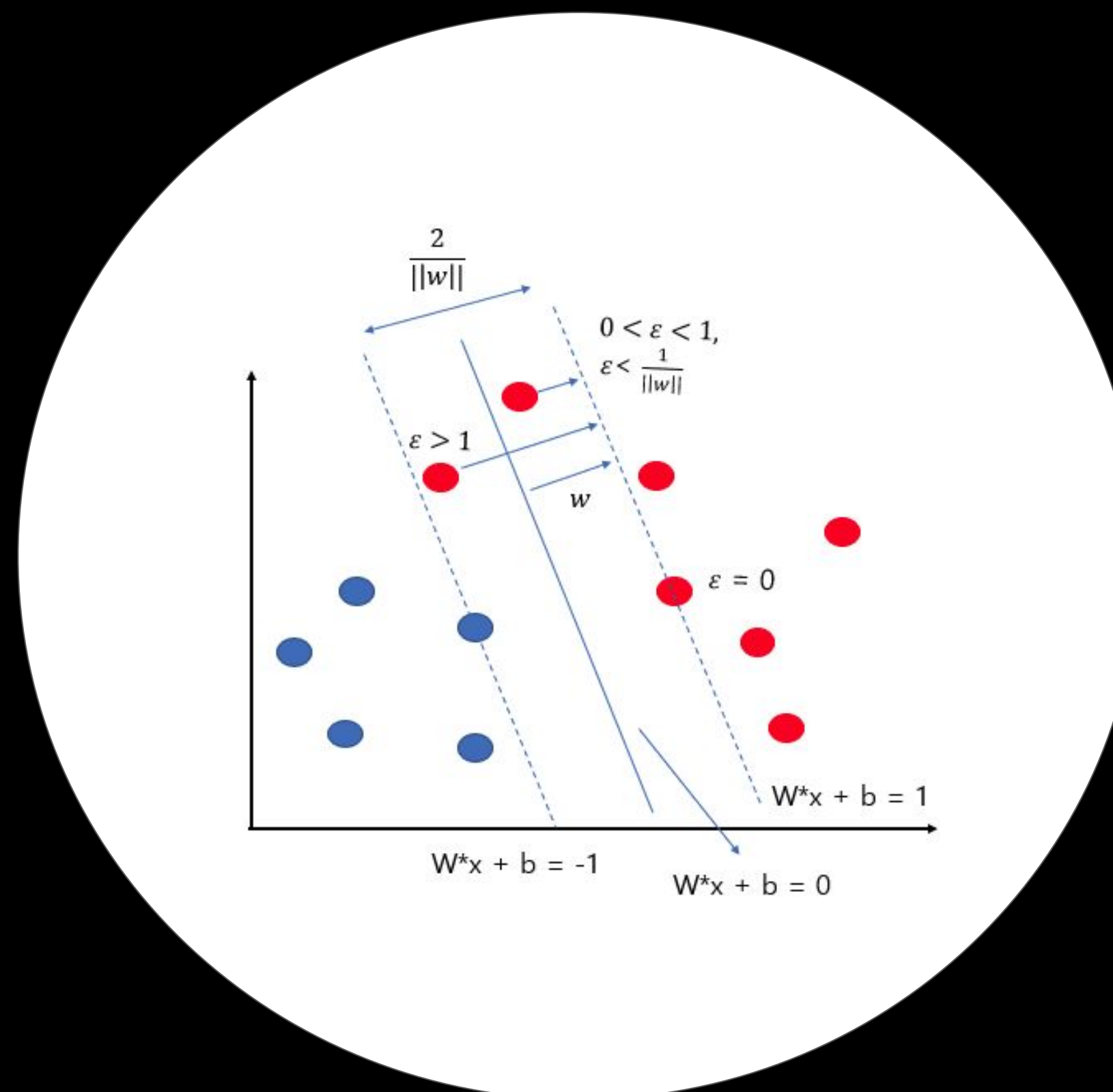
데이터를 분류하는 것은 **기계 학습**에 있어서 일반적인 작업이다. 주어진 데이터 점들이 두 개의 클래스 안에 각각 속해 있다고 가정했을 때, 새로운 데이터 점이 두 클래스 중 어느 것에 속하는지 결정하는 것이 목표이다.

즉 N차원을 공간을 (N-1)차원으로 나눌 수 있는 **초평면**을 찾는 분류 기법입니다.

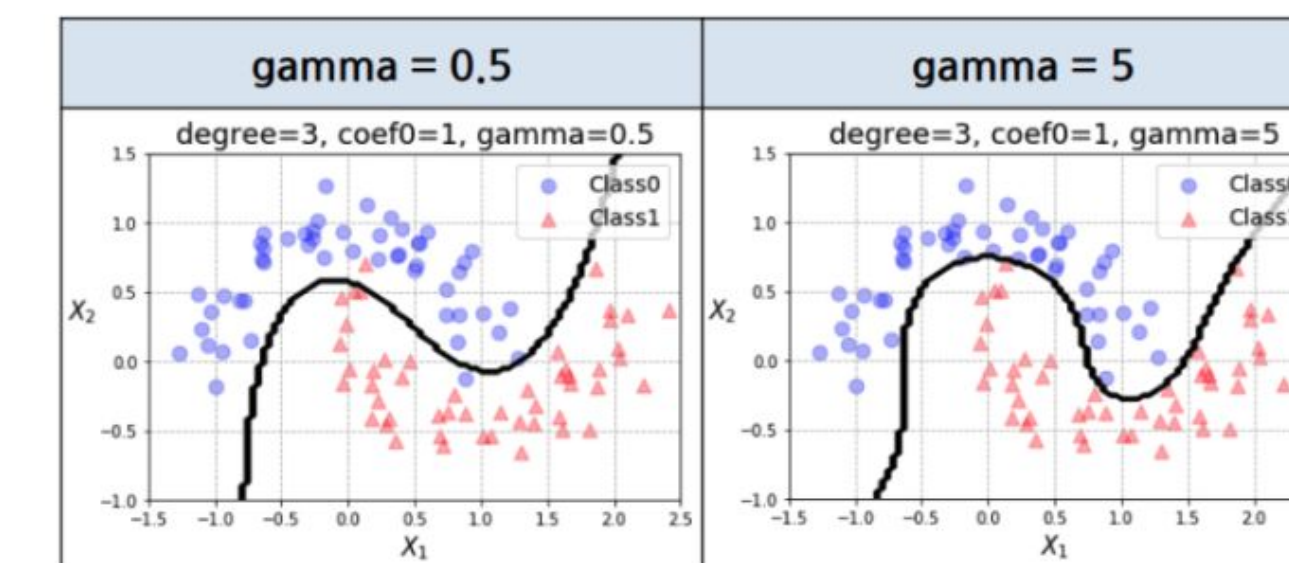
선형 SVM 하드마진



선형 SVM 소프트마진



비선형 SVM



Vector 를 이용한 데이터 사례: 서포트 벡터 머신 (SVM)

여러 프로그래밍 언어에 지원되는 서포트 벡터 머신 라이브러리 목록

C/C++

- SVM light: 다양한 운영체제에서 사용 가능.
- mySVM: SVM light 기반 최적화, Windows와 Linux 지원.
- LIBSVM: 다양한 언어 지원, C/C++ 포함.
- GPDT: C++로 구현, Linux 지원.

Python

- scikit-learn: 범용 머신 러닝 라이브러리, 다양한 운영체제 지원.
- PyML: 서포트 벡터 머신에 초점, Linux와 Mac OS X 지원.
- LIBSVM: 다양한 언어 지원, Python 포함.

Java

- LIBSVM: 다양한 언어 지원, Java 포함.

MATLAB

- LIBSVM: 다양한 언어 지원, MATLAB 포함.
- Matlab SVM Toolbox: 그래픽 사용자 인터페이스 제공.
- Spider: 매트랩용 머신 러닝 패키지.
- LS-SVMlab: 최소제곱법 서포트 벡터 머신 구현.
- SVM Tools: 다양한 서포트 벡터 머신 버전 지원.

R

- e1071: LIBSVM 기반.

Vector 를 이용한 데이터 사례: 벡터 데이터 장점, 목적

1. 비정형 데이터, 반정형 데이터의 분류
2. 방대한 양의 고차원 데이터를 효율적으로 저장
3. 분류되어 Vector화 된 데이터들의 색인 및 검색
3. 유사도 측정
4. 머신러닝, 기계검색에 활용

03 **AI** 가 **Vector data** 를 이용하는 법

AI 가 **Vector data** 를 이용하는 법

"이데아는 현상 세계 밖의 세상이며 이데아는 모든 사물의 원인이자 본질"
- 플라톤

Generally, for Plato, things were classified in relation with the distance that separates them from their archetypal forms, which yields some order (or pre-order) on them.

플라톤의 경우 사물은 원형과 분리되는 거리에 따라 분류되었으며, 이는 사물에 어떤 질서(또는 사전 질서)를 부여합니다.

AI 가 **Vector data** 를 이용하는 법

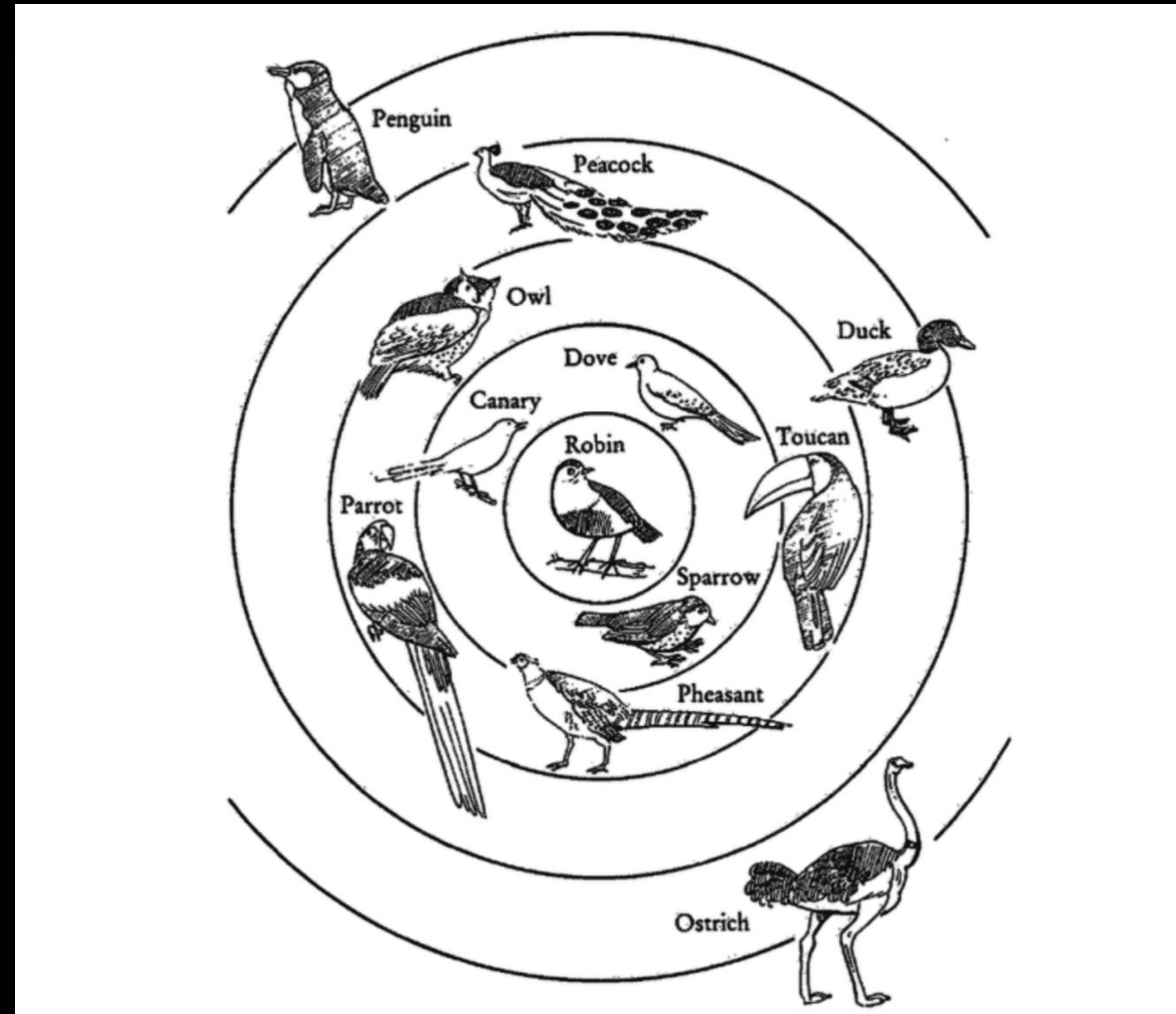
분류 (Classification)

분류 문제는 과학 연구의 기본 주제 중 하나입니다. 예를 들어, **수학**, **물리학**, **자연과학**, **사회과학**, 그리고 물론 도서관과 **정보과학**은 모두 분류법을 활용합니다. 분류는 주문 및 구성에 매우 유용한 도구입니다. 이는 지식을 늘리고 정보 검색을 촉진하는 데 도움이 되었습니다.

대략적으로 말하면 '분류'는 일반적으로 개체보다 적은 클래스나 그룹에서 개체를 공유, 배포 또는 할당하는 작업으로 구성됩니다.

AI 가 Vector data 를 이용하는 법

추운 남극에 사는 새
a cold Antarctic bird

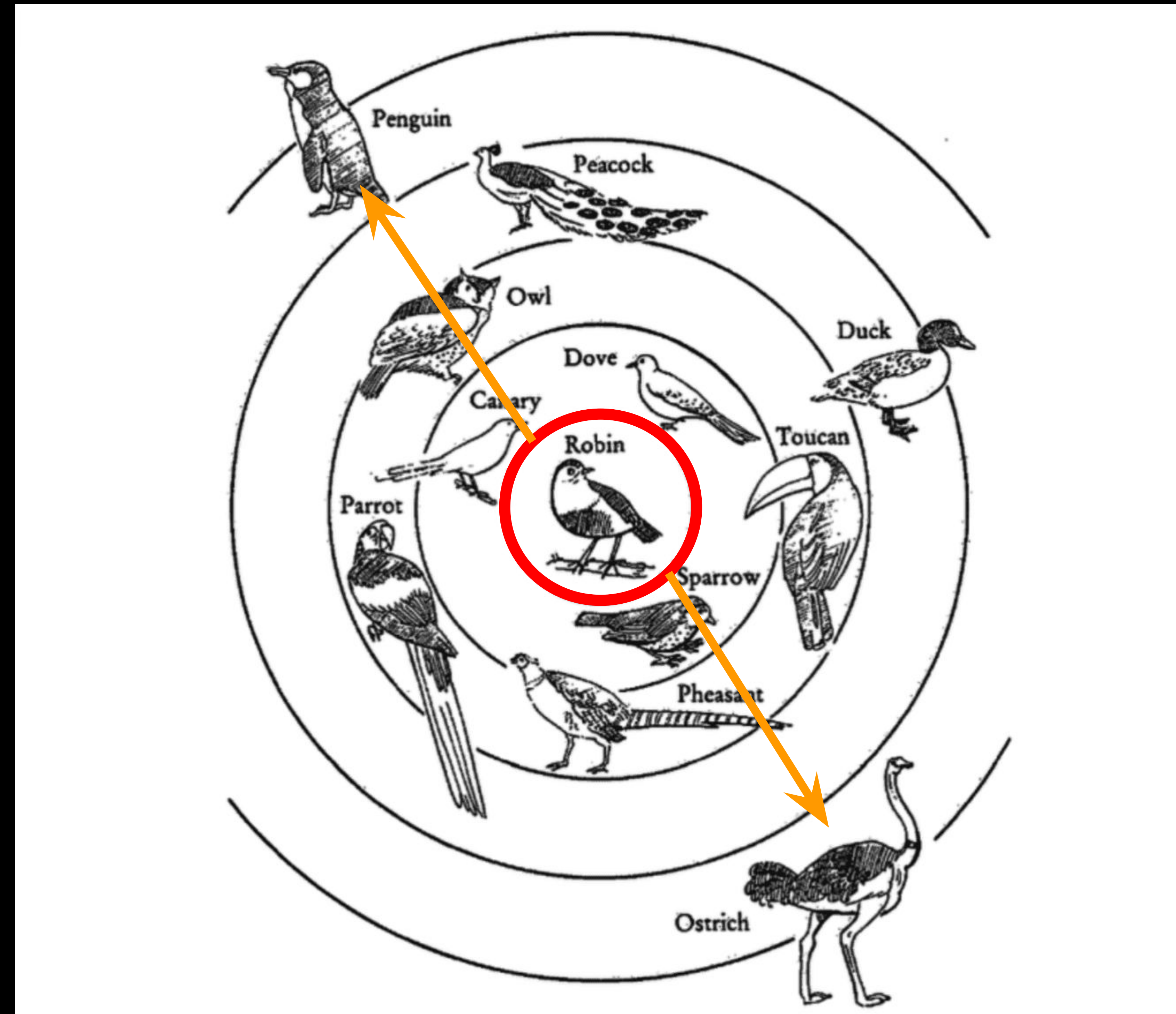


날 지는 못하지만 빨리
달리는 새

A bird that can't fly but
runs fast

AI 가 Vector data 를 이용하는 법

추운 남극에 사는 새
a cold Antarctic bird



날 지는 못하지만 빨리
달리는 새

A bird that can't fly but
runs fast

AI가 Vector data를 이용하는 법

추운곳에 사는 새?

답안 표시



❖ 추운 곳에 사는 새는 다양한 종류가 있습니다. 그 중 대표적인 몇 가지를 소개해 드리겠습니다.

- **황제펭귄:** 황제펭귄은 지구상에서 가장 큰 펭귄입니다. 남극 대륙과 주변 섬에 서식하며, 추운 기후를 견딜 수 있도록 두꺼운 깃털과 지방층을 가지고 있습니다. 황제펭귄은 탁월한 수영 실력을 가지고 있으며, 물 속에서 먹이를 찾습니다.



황제펭귄

일반적인 새?

답안 표시



❖ 일반적인 새는 다양한 종류가 있지만, 그 중에서도 대표적인 몇 가지를 소개해 드리겠습니다.

- **참새:** 참새는 전 세계적으로 가장 흔하게 볼 수 있는 새 중 하나입니다. 도시와 시골 어디에서나 볼 수 있으며, 곡식, 씨앗, 과일 등을 먹이로 합니다. 참새는 울음소리가 매우 시끄럽고, 집단으로 모여 생활하는 특징이 있습니다.



참새

날 지는 못하지만 빨리 달리는 새?

답안 표시



- ❖ 날지 못하지만 빨리 달리는 새는 바로 타조입니다. 타조는 아프리카에 서식하는 조류로, 지구상에서 가장 큰 새입니다. 키는 최대 2.7m, 무게는 최대 150kg까지 자랄 수 있습니다. 타조는 날개가 매우 작아 날 수 없지만, 다리가 매우 튼튼하고 발가락에는 날카로운 발톱이 있습니다. 타조는 시속 최대 70km의 속도로 달릴 수 있으며, 한 번에 최대 5m를 도약할 수 있습니다.



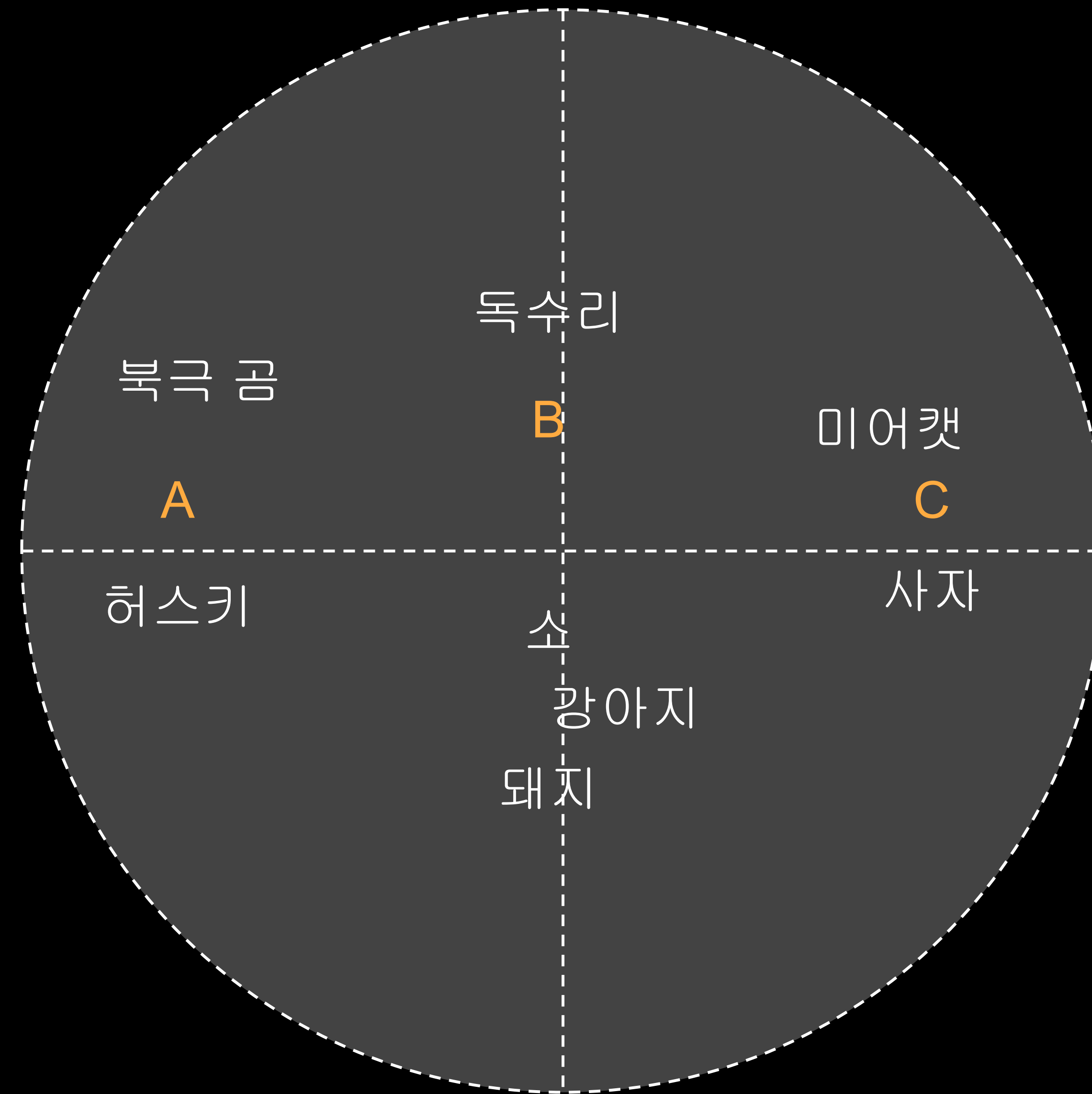
타조

AI가 Vector data를 이용하는 법

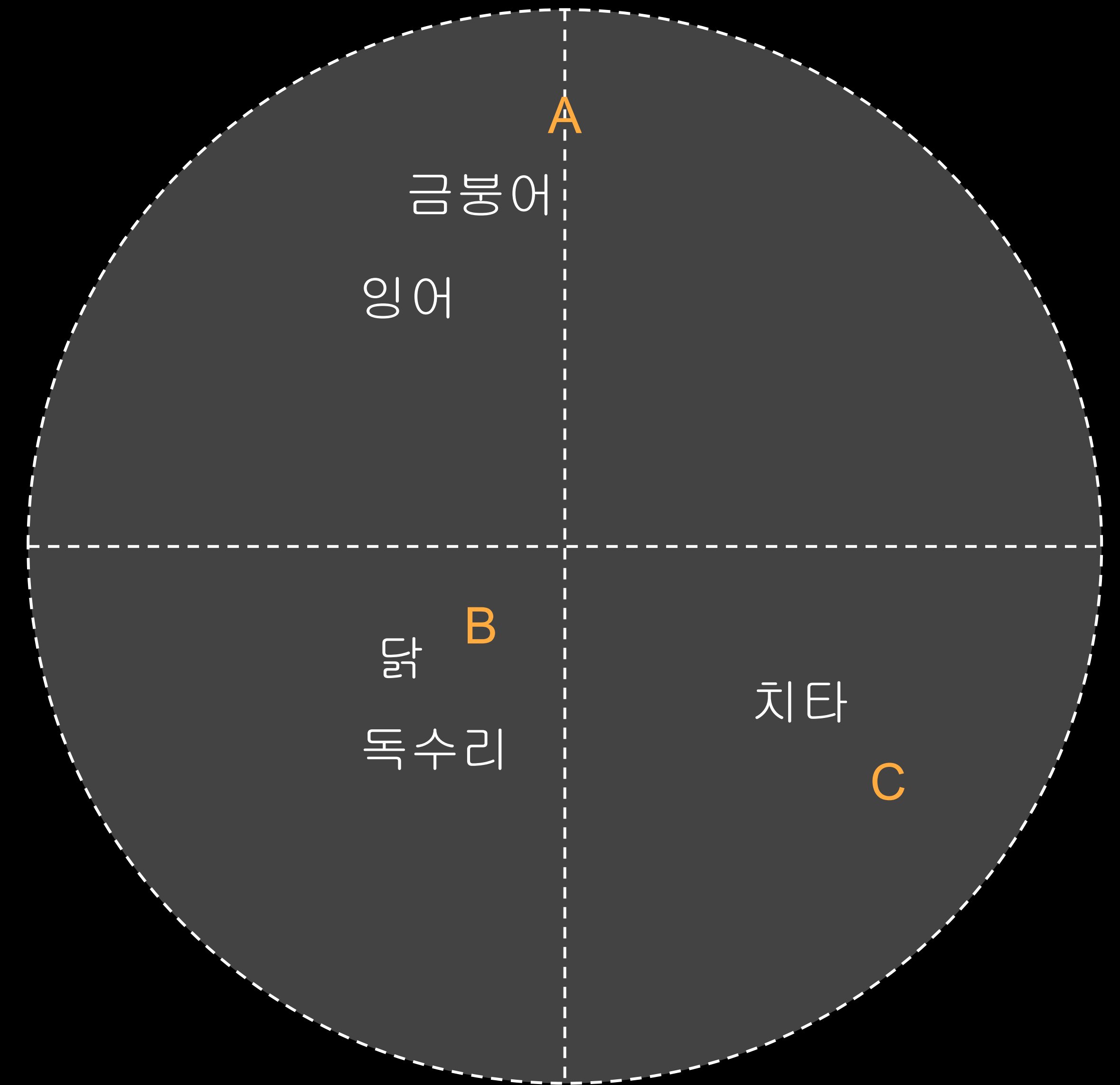
맞추어 보세요 **A, B, C**는 어떤 동물 일까요?



종별 분류



기후별 분류



특징별 분류

AI 가 **Vector data** 를 이용하는 법

A : 펭귄

B : 참새

C : 타조

AI 가 **Vector data** 를 이용하는 법

임베딩 (**Embedding**)

임베딩은 머신 러닝 모델과 시맨틱 검색 알고리즘에서 사용하도록 설계된 텍스트, 이미지, 오디오와 같은 값 또는 개체의 표현입니다. 임베딩은 이러한 개체를 각 개체가 가지고 있거나 가지고 있지 않은 **요소** 또는 **특성**, 개체가 속한 **범주**에 따라 **수학적 형태**로 변환합니다.

AI 가 **Vector data** 를 이용하는 법

임베딩 (Embedding)



[0,2]



[1,2]



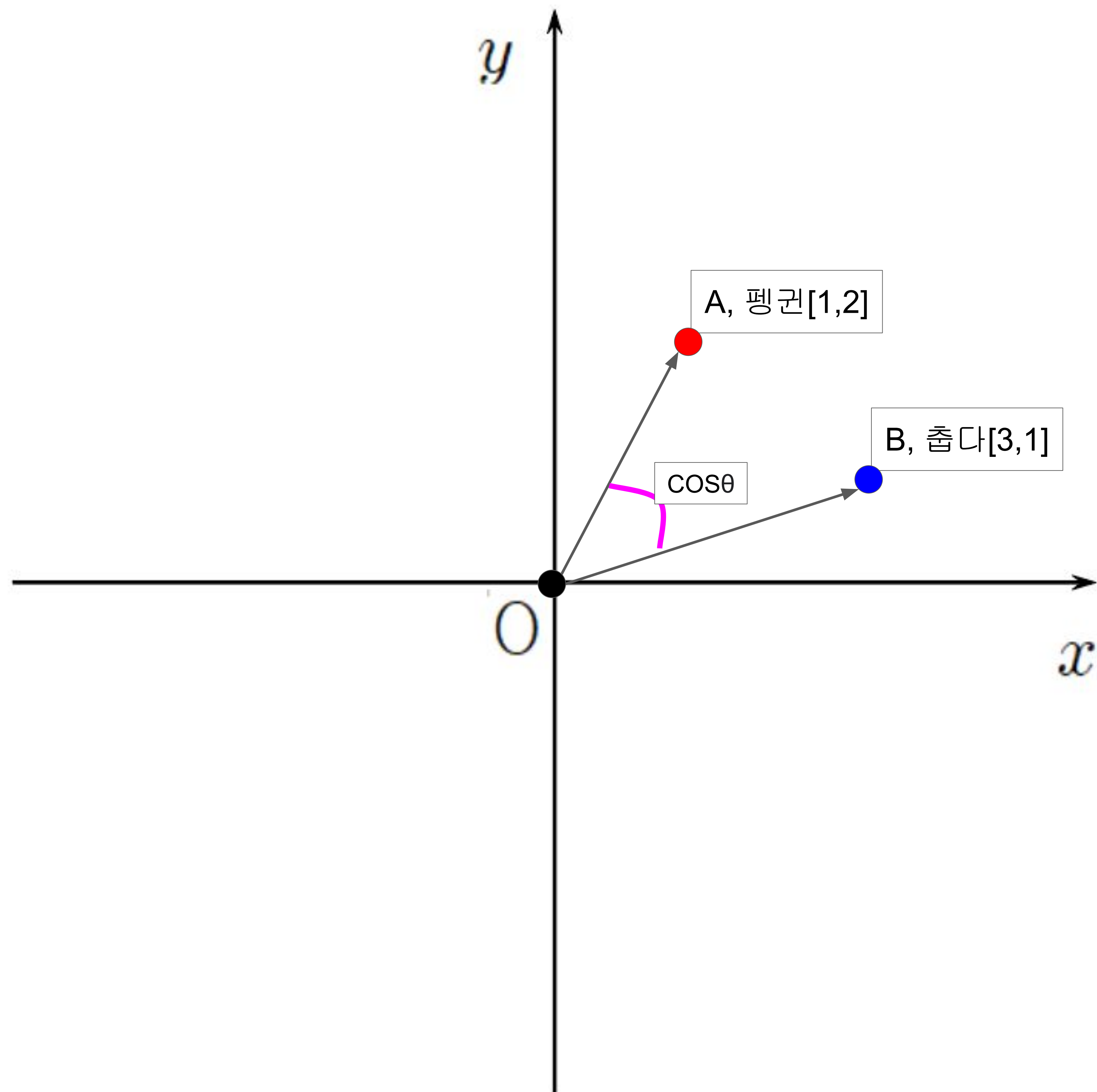
[3,0]

AI 가 **Vector data** 를 이용하는 법

유사도(**Similarity**) 측정

유사도(similarity)란 두 데이터가 얼마나 같은지 나타내주는 척도

AI 가 Vector data 를 이용하는 법



내적 유사도 측정법 cosine 유사도 측정법

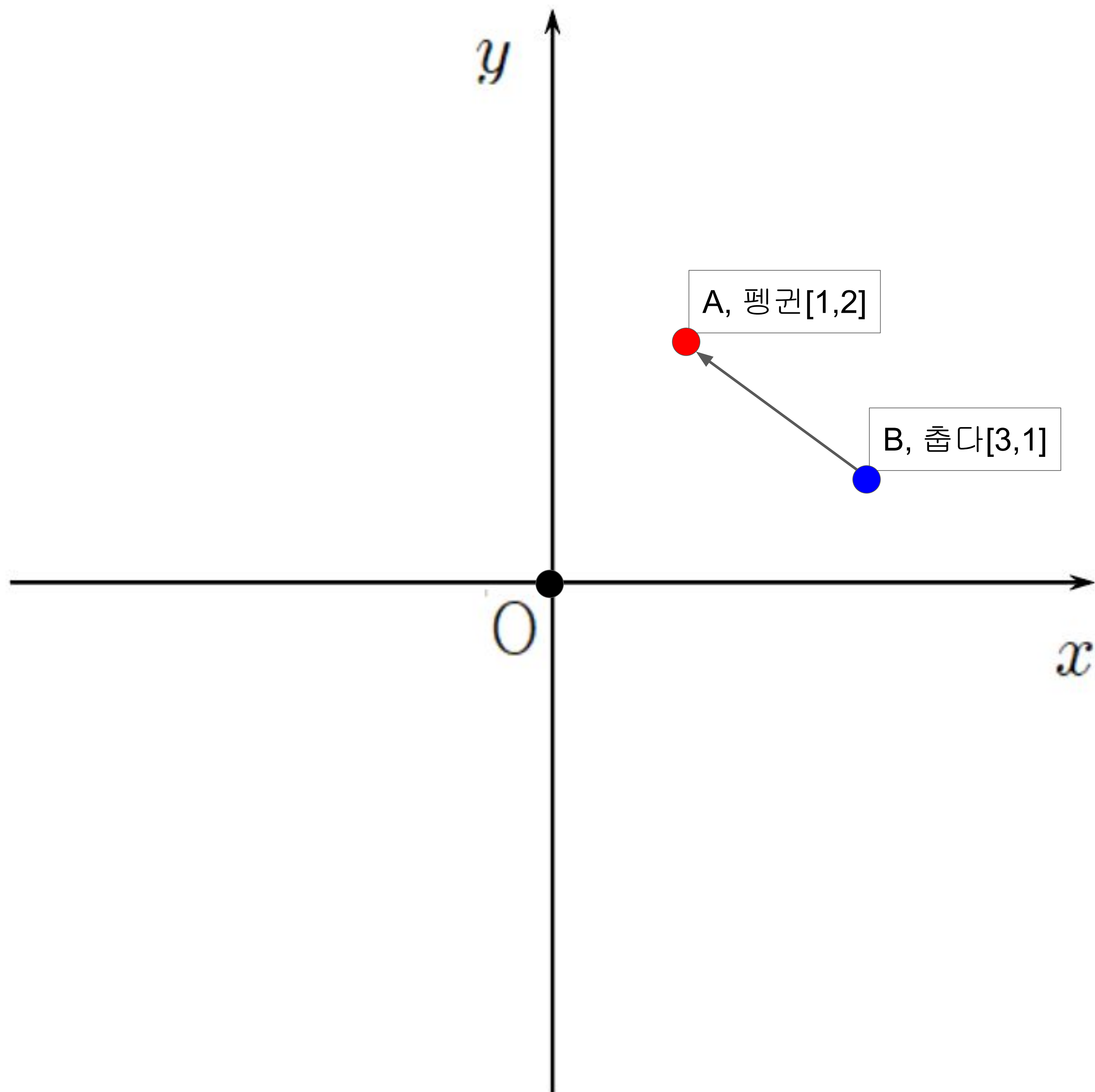
내적공간의 두 벡터간 각도의 코사인값을 이용하여 측정된 벡터간의 유사한 정도를 구하는 측정법입니다.
측정방식 및 수식은 다음 아래와 같이 설명 할수 있습니다

$$A \cdot B = |A| |B| \cos\theta$$

$A \cdot B$ 는 벡터간의 성분을 곱한 결과 이므로 행렬 곱셈 공식에 의하여 다음 아래 와 같이 계산합니다.

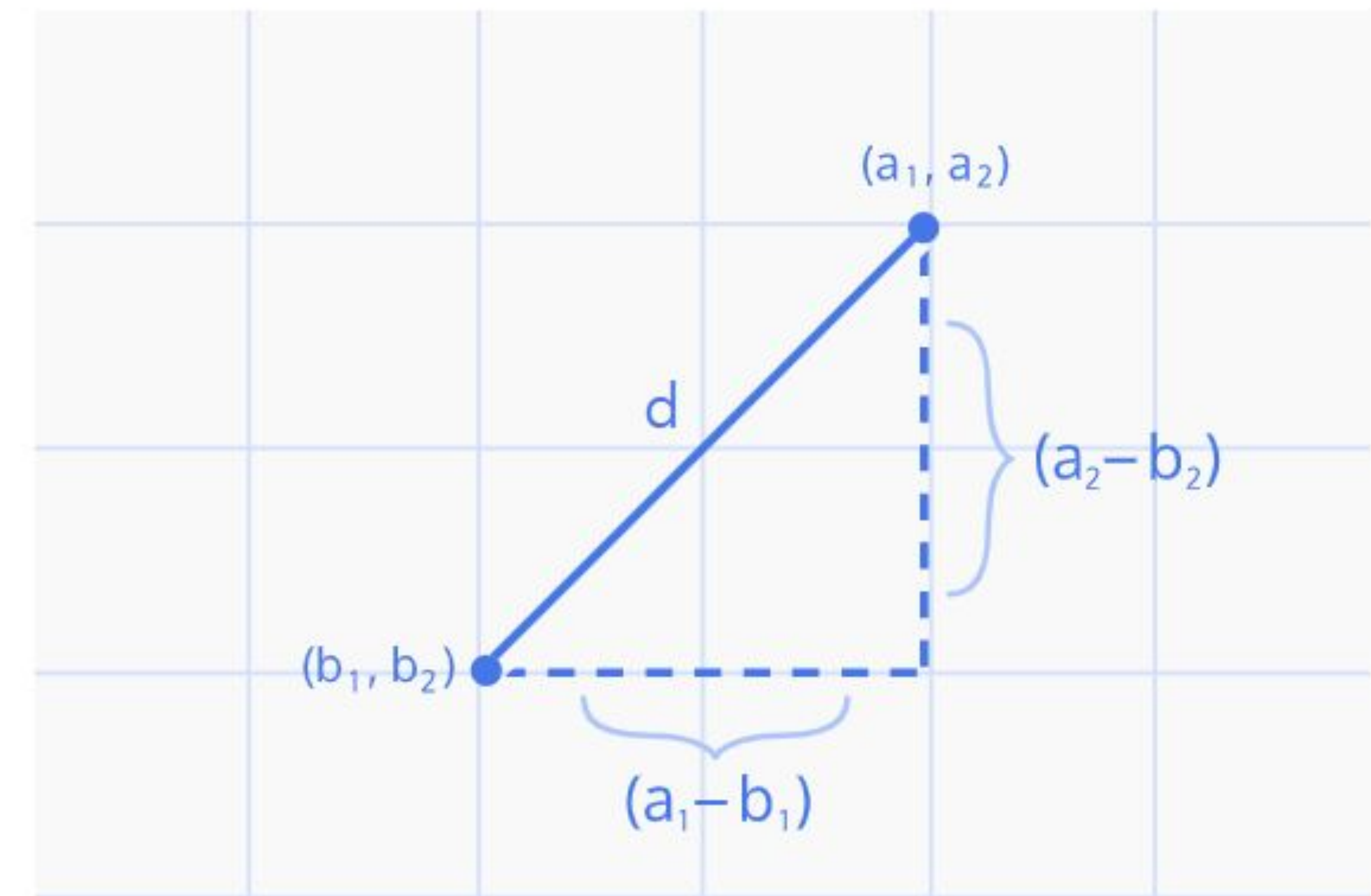
$$A(x) \cdot B(x) + A(y) \cdot B(y) = 5$$

AI 가 Vector data 를 이용하는 법



유클리디안 거리 유사도 측정법

유클리드 거리(Euclidean Distance)는 e두 점 사이의 거리를 계산하는 기법입니다 두 점 사이의 거리를 구하는 공식은 다음 아래와 같습니다.



$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

AI 가 **Vector data** 를 이용하는 법

Python 을 이용한 **vector data** 유사도 측정 실습

https://github.com/shinhanbyeol/study-vector-data/blob/main/vector_sample.py

Q & A
감사합니다.