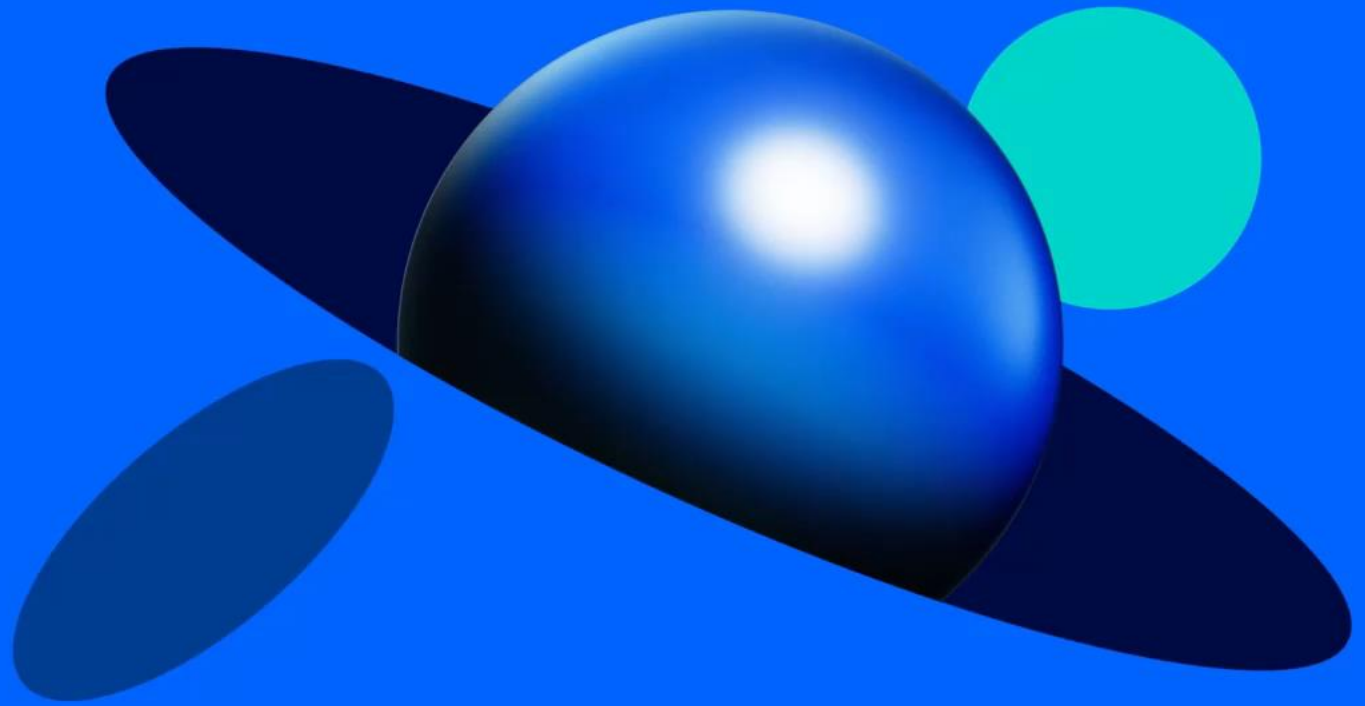
NHN Cloud make IT 2023

시계열 데이터 속에 숨어 있는 이상 징후를 찾는 딥 러닝 기술



NHN Cloud / 클라우드AI팀
박현목

목차

- 이상 탐지란?
- 이상 탐지, 어떻게 활용할 수 있을까?
- 이상 탐지를 위한 고민들
- 이상 탐지 기술 개발
- 이상 탐지 서비스

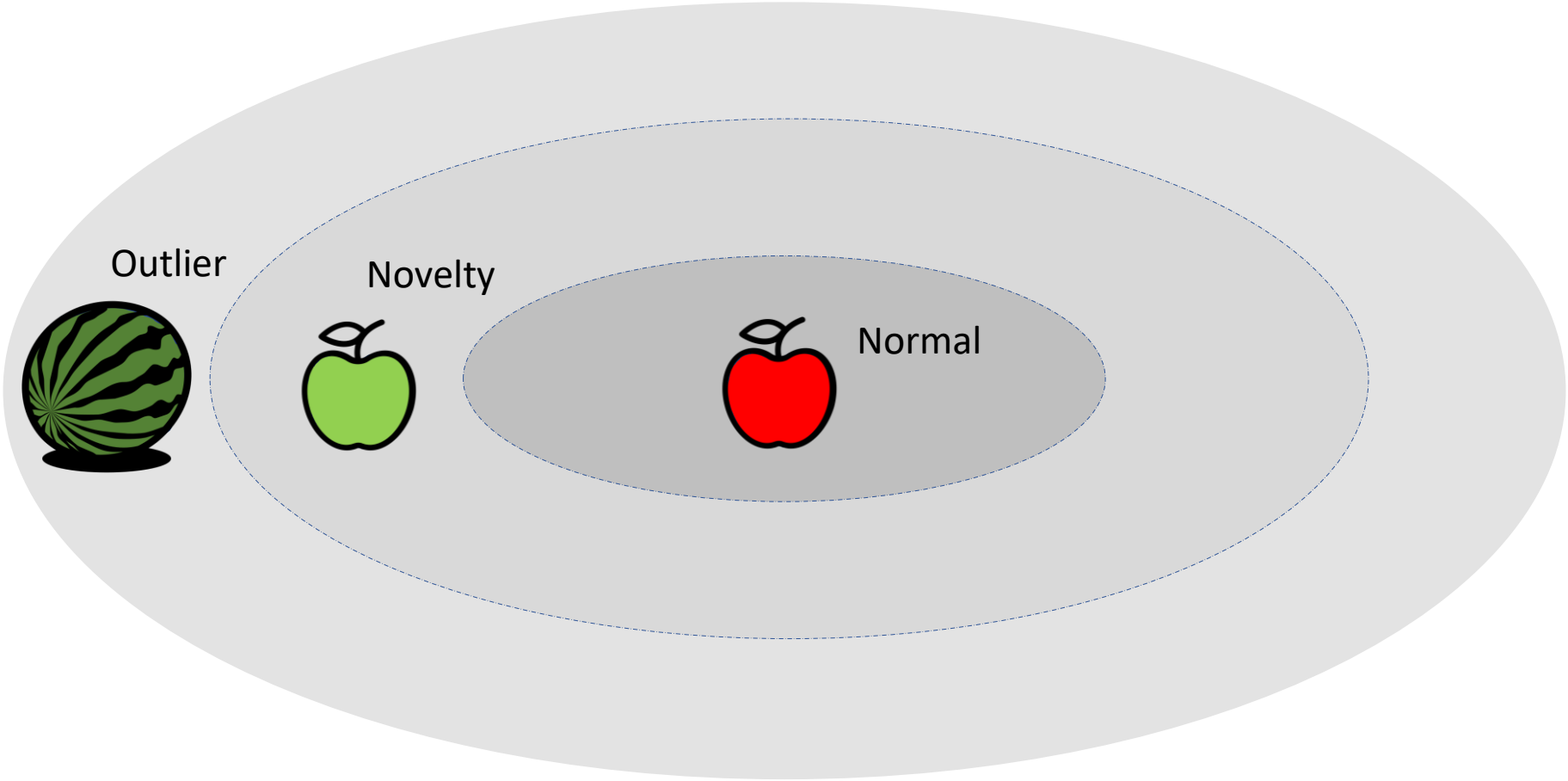
NHN Cloud make IT 2023

이상 탐지란?

이상 탐지란?

이상 탐지(anomaly detection)

- 정상적인 동작에서 벗어난 데이터 포인트, 이벤트를 식별하는 데이터 분석 과정



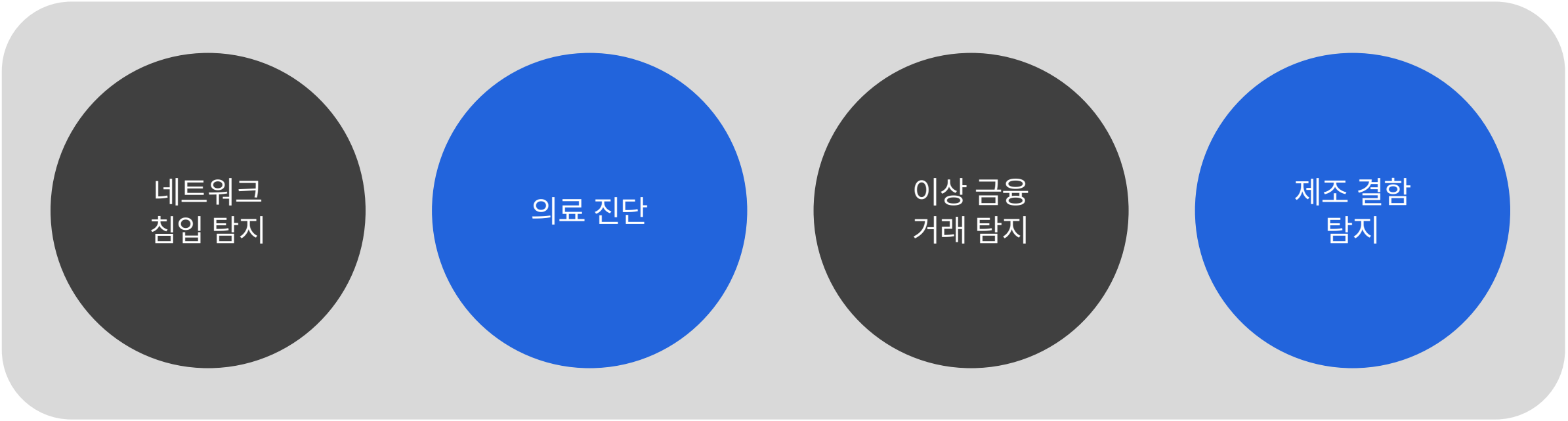
NHN Cloud make IT 2023

이상 탐지, 어떻게 활용할 수 있을까?

이상 탐지, 어떻게 활용할 수 있을까?

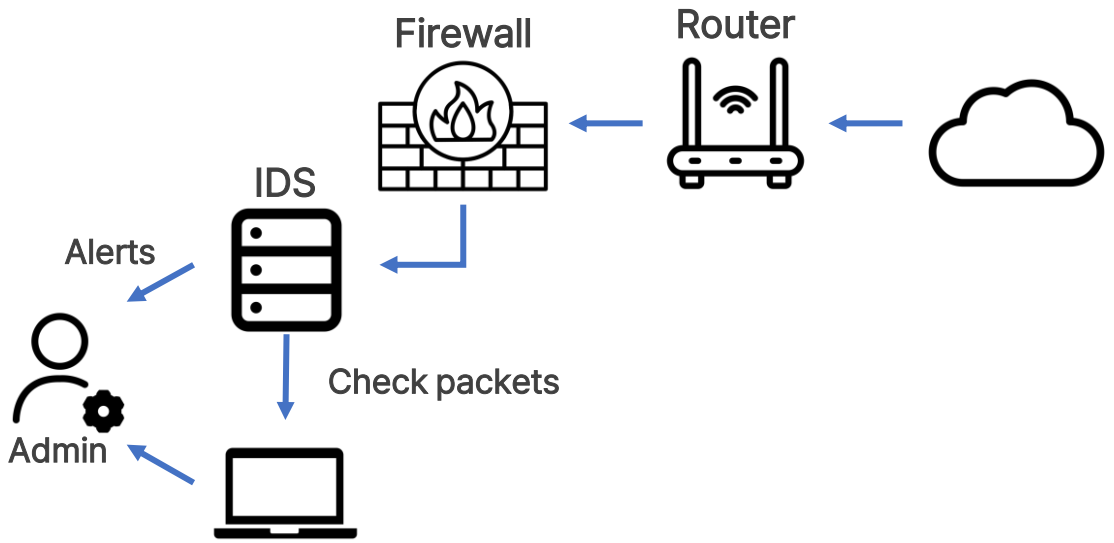
활용 사례

- 실시간 모니터링이 필요한 분야
- 안정적인 시스템 유지, 관리와 결함 탐지가 중요한 분야

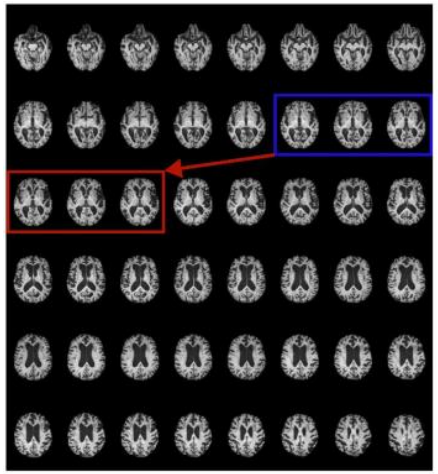


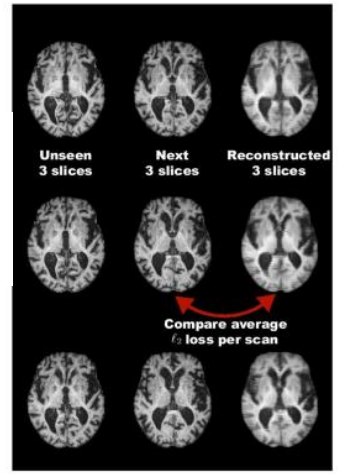
이상 탐지, 어떻게 활용할 수 있을까?

네트워크
침입 탐지



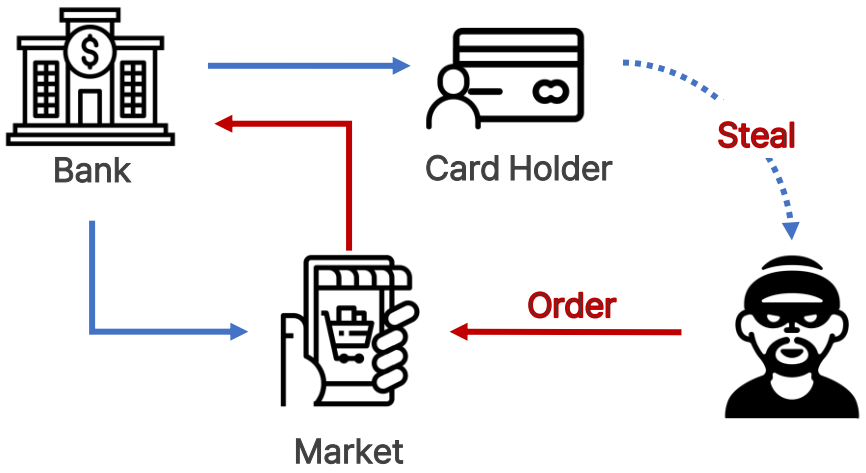
의료 진단



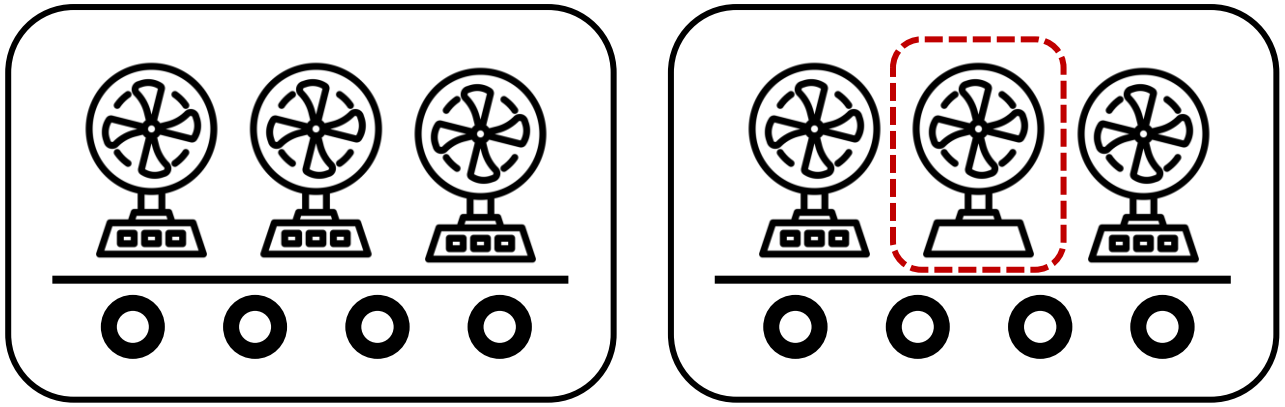


이상 탐지, 어떻게 활용할 수 있을까?

이상 금융
거래 탐지



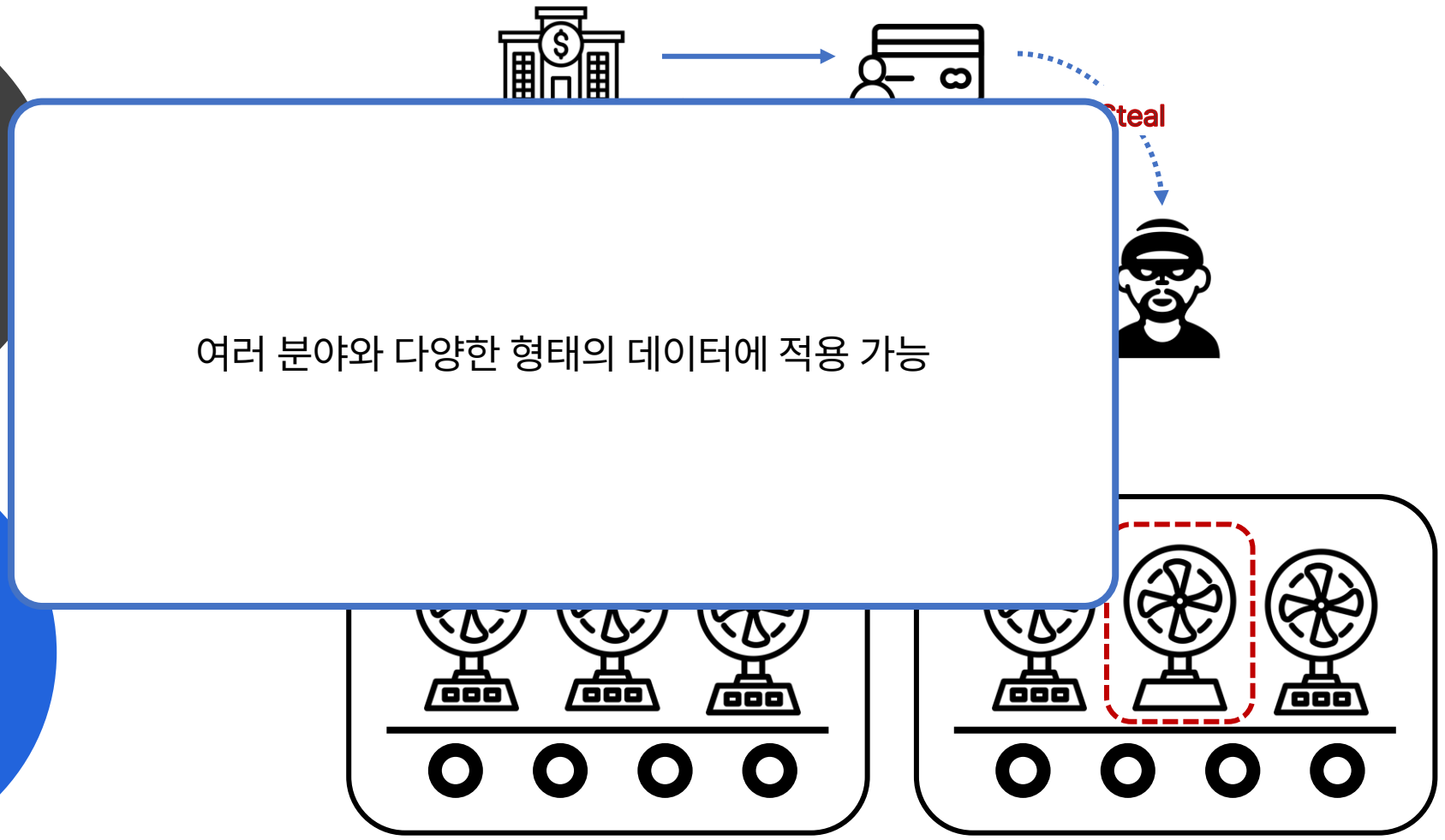
제조 결함
탐지



이상 탐지, 어떻게 활용할 수 있을까?

이상 금융
거래 탐지

제조 결함
탐지



여러 분야와 다양한 형태의 데이터에 적용 가능

시계열 이상탐지

왜 우리는 시계열에 주목했을까?

- 단순하지만 우리의 생활 속에 깊게 자리 잡은 데이터

숫자 123 + 시간  = 시계열 데이터

시계열 이상탐지

왜 우리는 시계열에 주목했을까?

- 단순하지만 우리의 생활 속에 깊게 자리 잡은 데이터

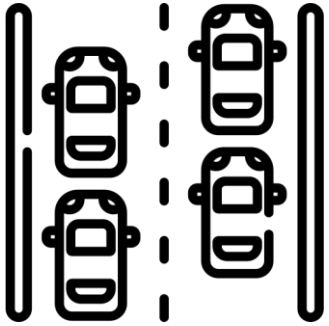
모두 시계열 데이터

주식

인터넷 속도

심장 박동

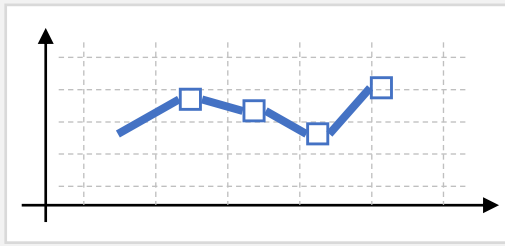
시간당 교통량

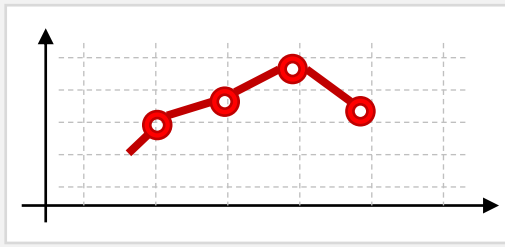


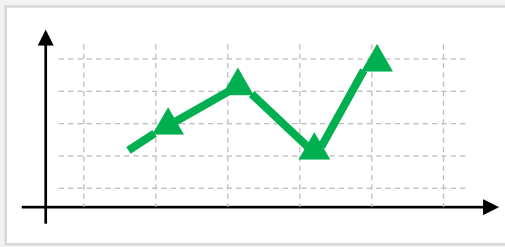
배경 지식

시계열 데이터란?

- 시간 순서대로 나열된 테이블 데이터







	00:00:00	00:01:00	00:02:00	00:03:00	...
A	102	100	90	130	...

	00:00:00	00:01:00	00:02:00	00:03:00	...
B	0.1	1.5	2.3	0.3	...

	00:00:00	00:01:00	00:02:00	00:03:00	...
C	1002	1200	960	1300	...

배경 지식

시계열 데이터란?

- 시간 순서대로 나열된 테이블 데이터

[데이터 수(N)] X [데이터 측정 시간(T)]

다변량

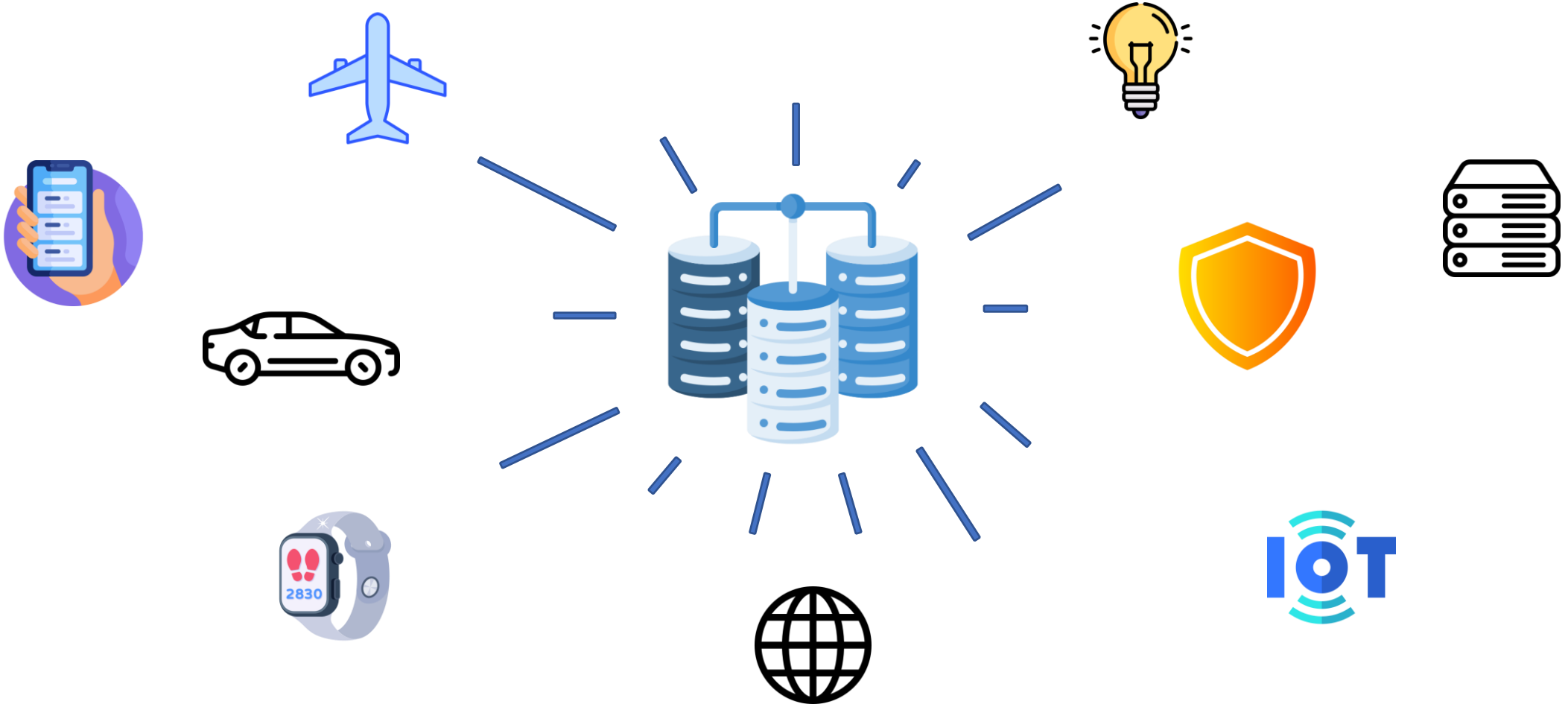
단변량

	00:00:00	00:01:00	00:02:00	00:03:00	...
A	102	100	90	130	
B	0.1	1.5	2.3	0.3	
C	1002	1200	960	1300	
D					
...					...

왜 이상 탐지 시스템이 필요할까?

지금은 "빅 데이터" 시대

넘쳐나는 데이터에 대한 관리, 검증 및 안정화는 필수



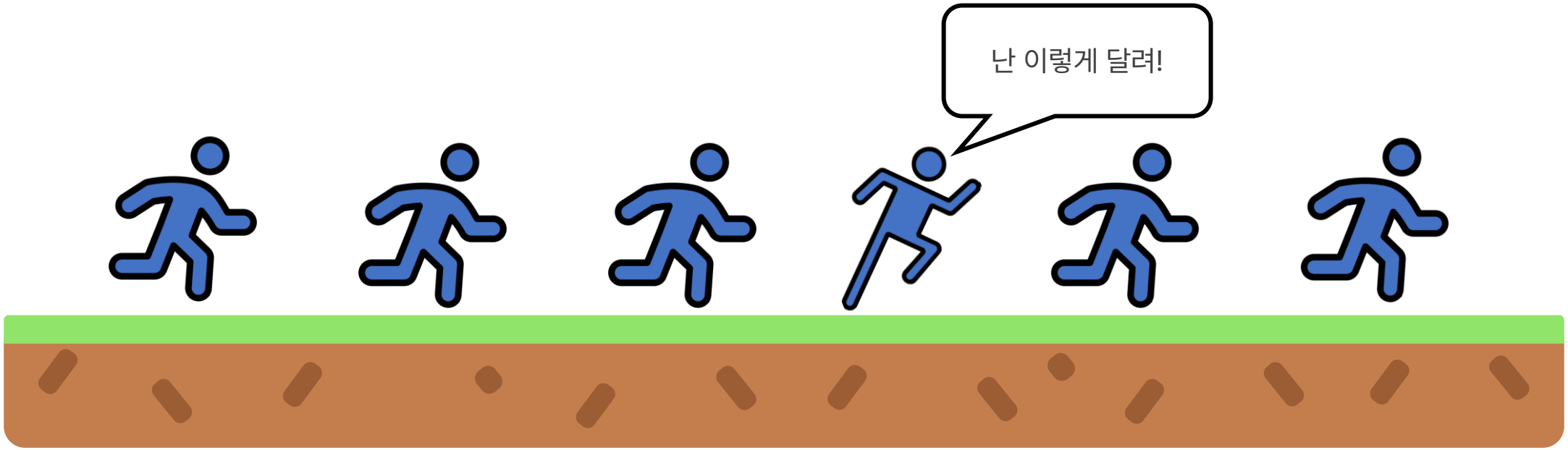
NHN Cloud make IT 2023

이상 탐지를 위한 고민

이상 탐지를 위한 고민들

이상 탐지의 목적은?

- 정상적인 동작에서 벗어난 데이터를 찾는 것
- Q) **정상적**이라는 것은?
  - 모든 비정상 동작을 하나하나 정의하기는 현실적으로 불가능
  - 다른 데이터들과 비교해서 패턴에서 벗어난 데이터를 **비정상 데이터**로 식별



이상 탐지를 위한 고민들

이상 탐지의 목적은?

- 정상적인 동작에서 벗어난 데이터를 찾는 것
- Q) **정상적**이라는 것은?
  - 모든 비정상 동작을 하나하나 정의하기는 현실적으로 불가능
  - 다른 데이터들과 비교해서 패턴에서 벗어난 데이터를 **비정상 데이터**로 식별

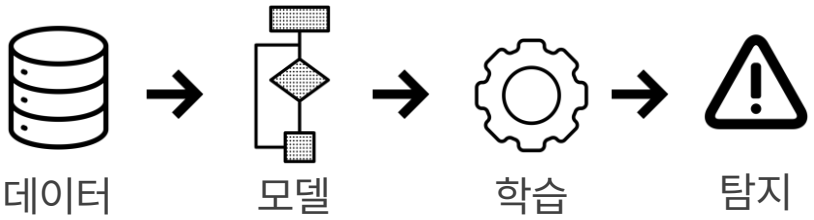
이상 탐지의 역사

Rule Based

- 이상 동작의 규칙을 정하여 탐지
- 경험적인 요소가 중요
- 매우 간단하게 적용 가능
- 규칙을 벗어난 동작을 탐지할 수 없음

Machine Learning

Traditional



- 확률, 통계적 접근법
- 준수한 성능
- 복잡한 모델링이 어려움

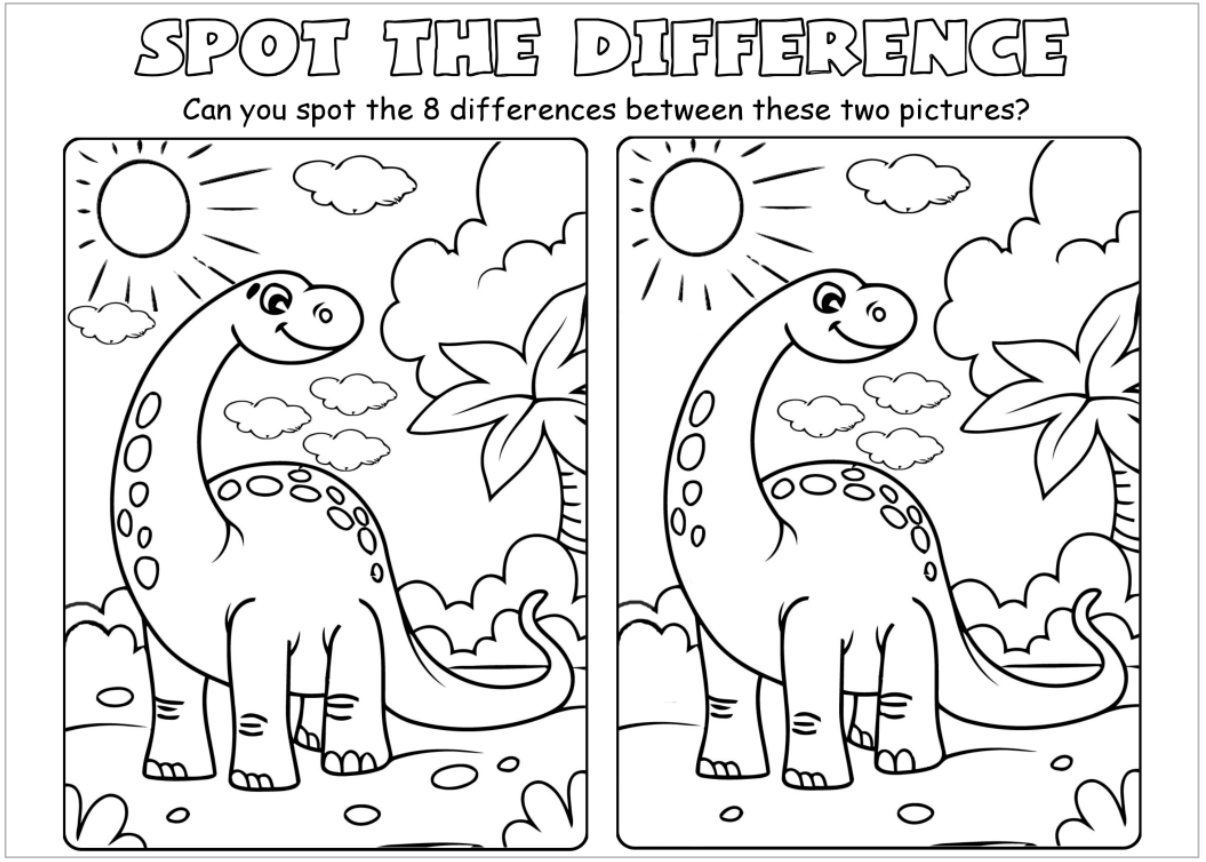
Deep Learning

- Representation Learning
- 활발한 연구가 진행 중
- 적절한 모델과 데이터 필요

이상 탐지가 어려운 이유

이해하기 어려운 데이터

- 데이터를 이해하기 위한 **도메인 전문 지식**이 필요
- 전문가도 정상, 이상을 확실하게 구분하기 어려운 경우가 다수



출처: https://www.printitfree.net/dinosaur-spot-the-difference-picture-puzzl

이상 탐지가 어려운 이유

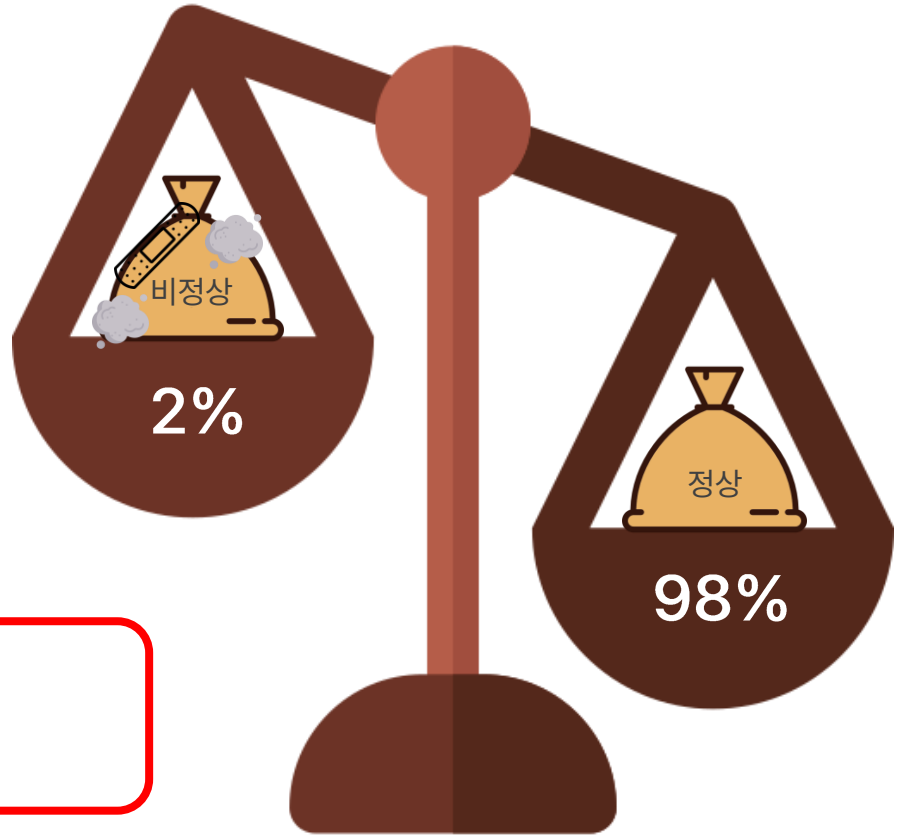
데이터 불균형

- **정상 데이터**와 비교해 너무 적은 **비정상 데이터**의 수

이상 탐지가 어려운 이유

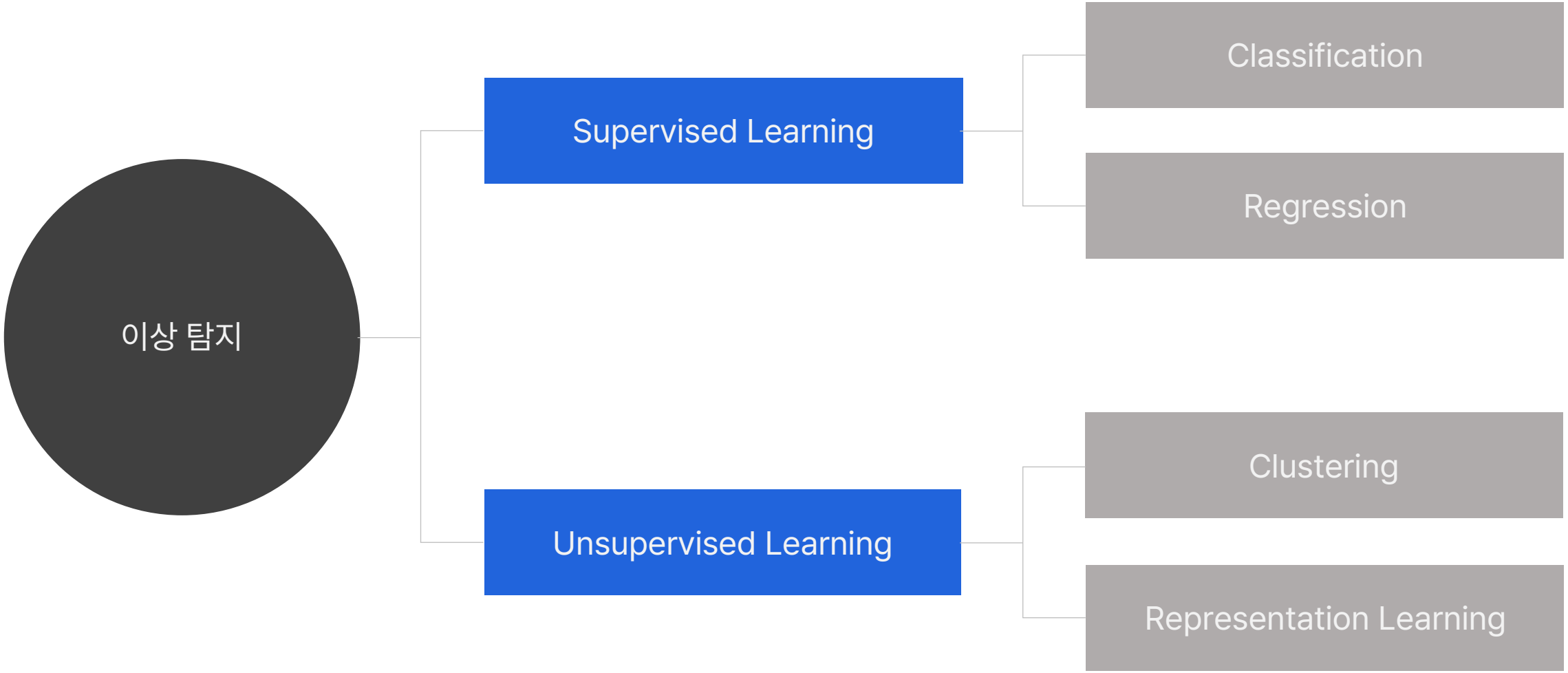
데이터 불균형

- **정상 데이터**와 비교해 너무 적은 **비정상 데이터**의 수



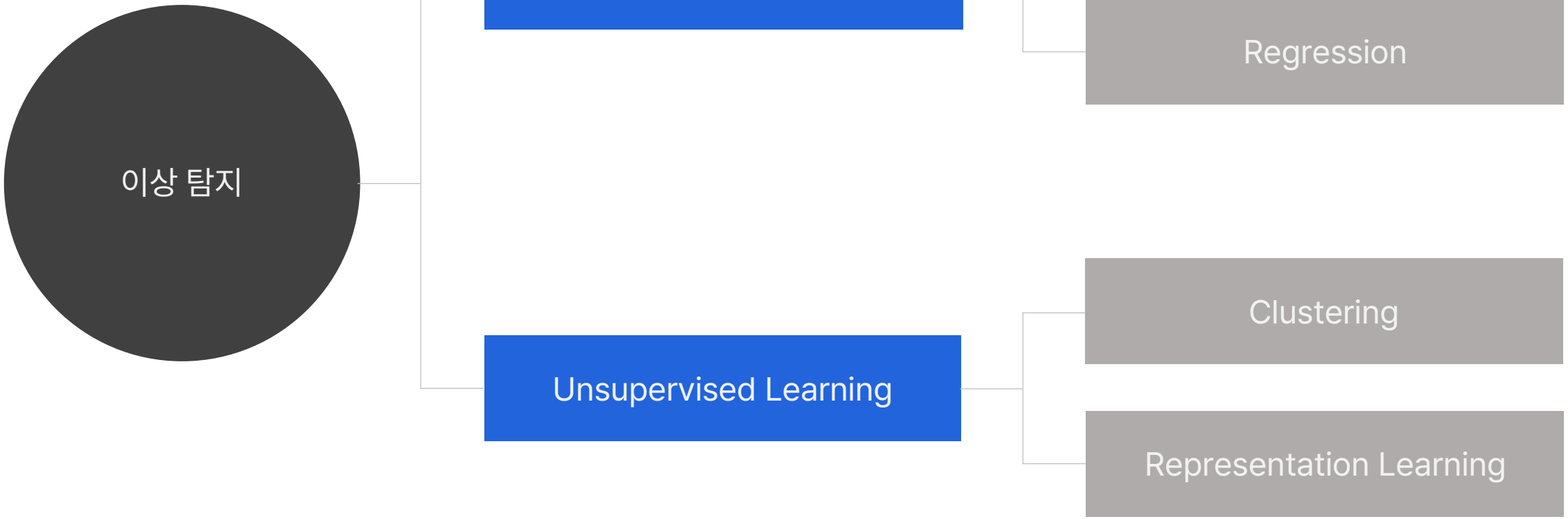
정상, 비정상 데이터의 차이를
충분히 분석하기 어려움

까다로운 데이터... 어떻게 학습할까?

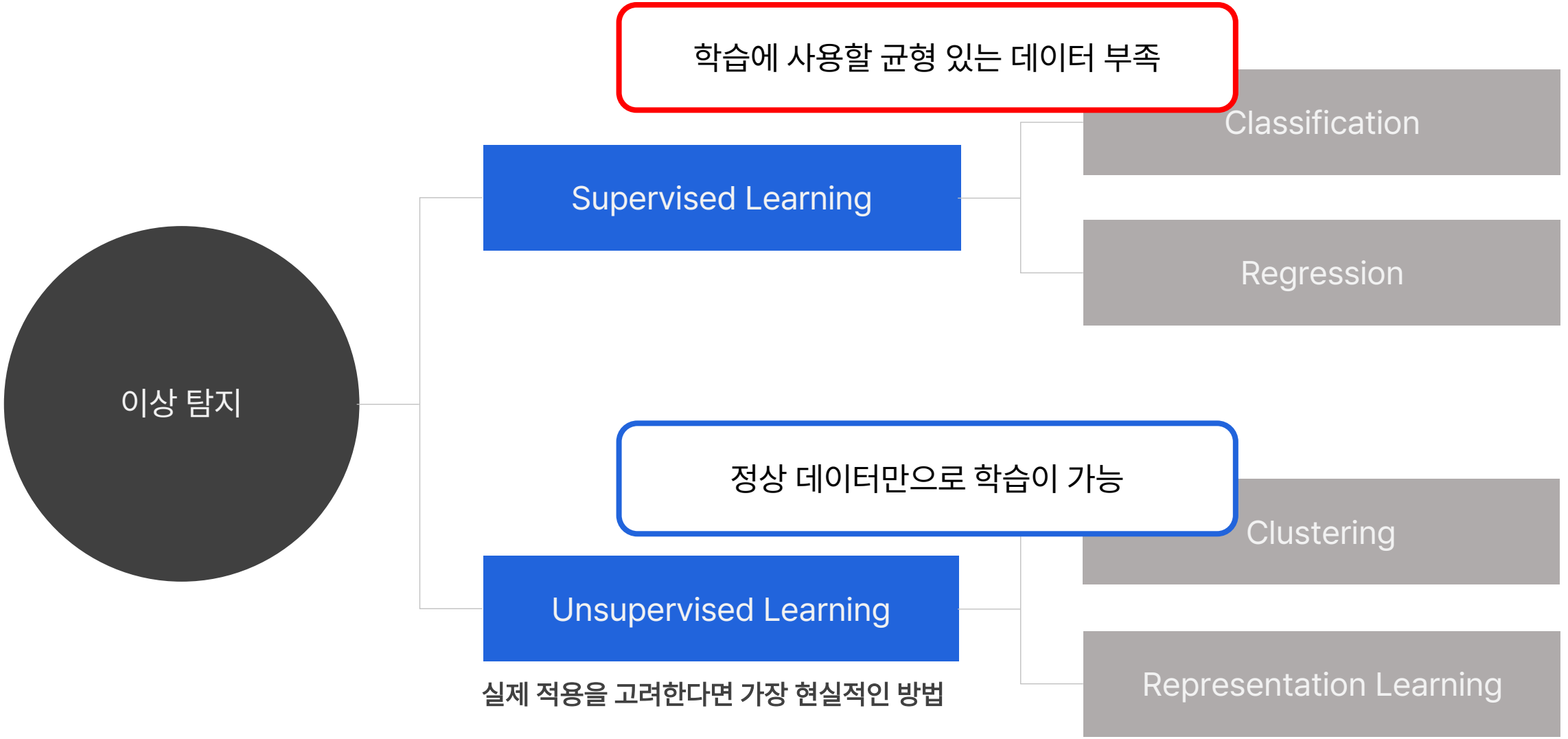


까다로운 데이터... 어떻게 학습할까?

학습에 사용할 균형 있는 데이터 부족



까다로운 데이터... 어떻게 학습할까?

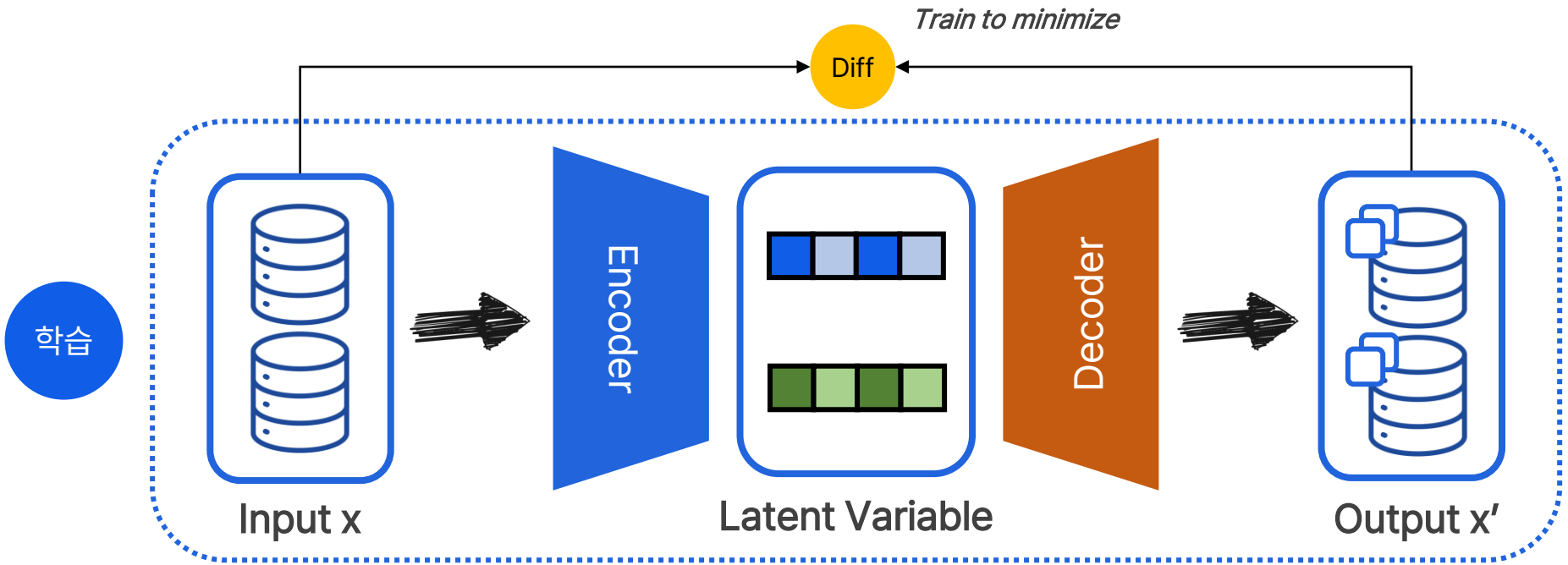


실제 적용을 고려한다면 가장 현실적인 방법

비지도 학습

무엇을 학습할까?

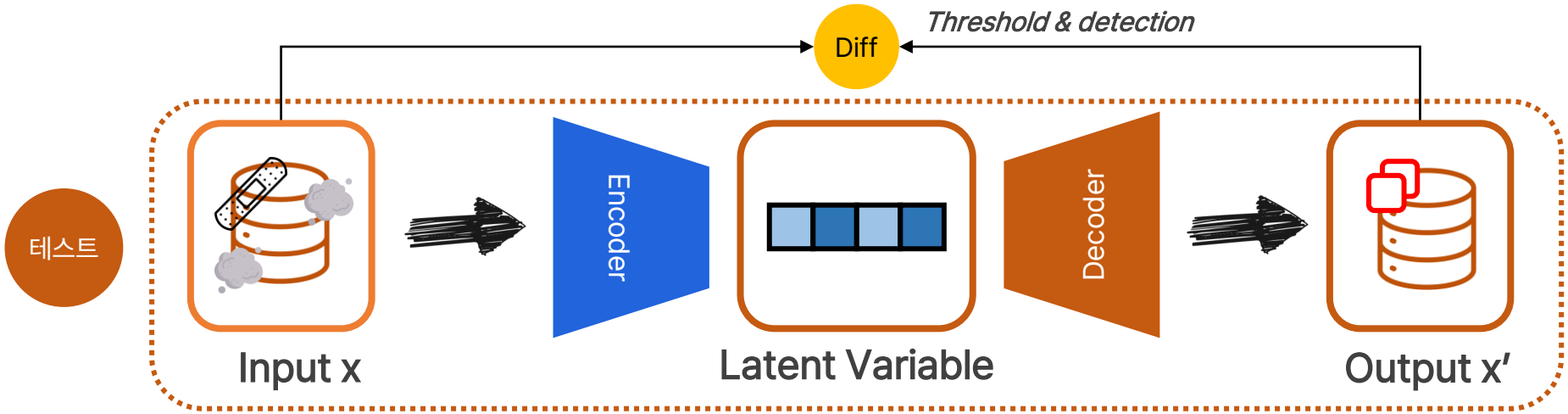
- 입력 데이터(정상 데이터)를 **"원래대로"** 복원(reconstruction)하도록 학습



비지도 학습

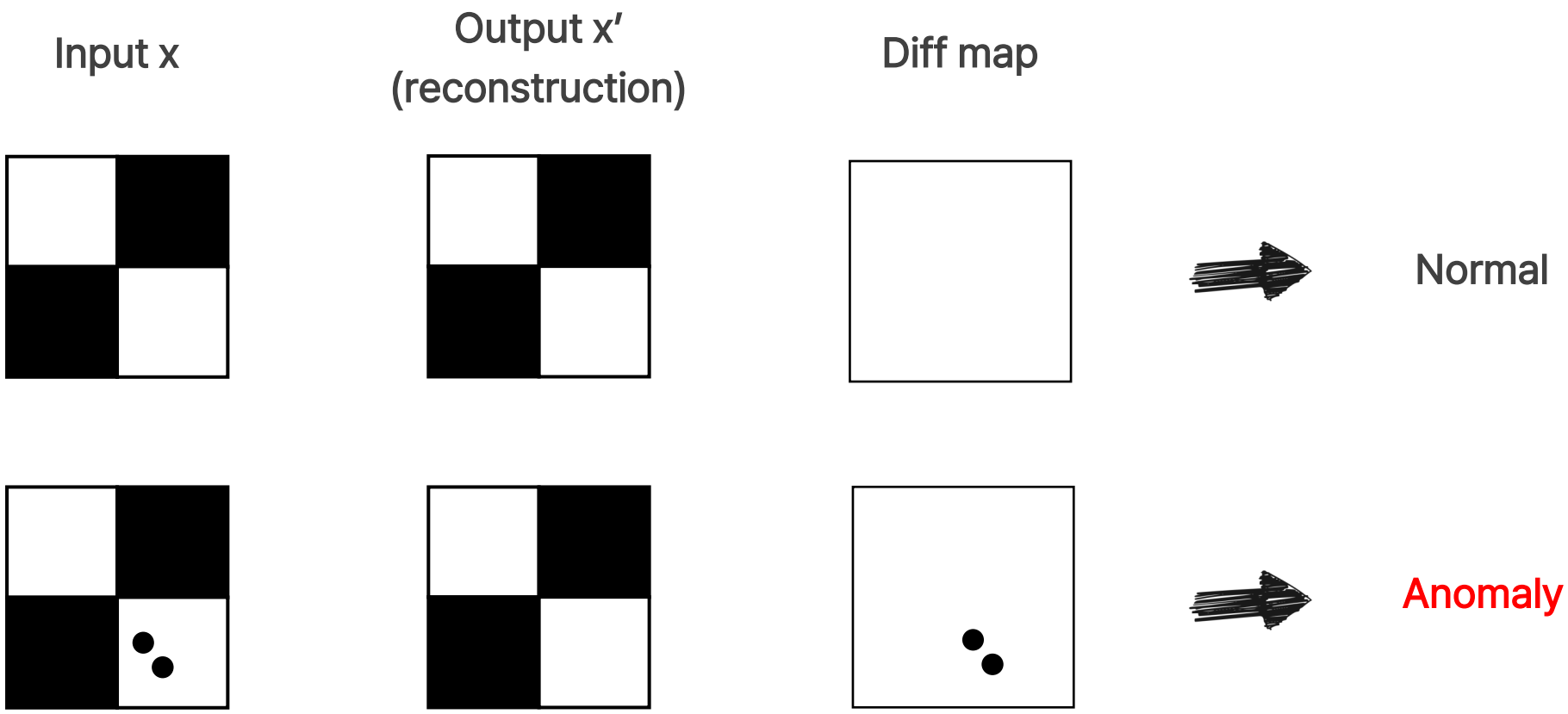
이상 데이터 탐지

- 학습한 정상 패턴을 바탕으로 **비정상 데이터**를 복원



비지도 학습

이상 데이터 탐지



NHN Cloud make IT 2023

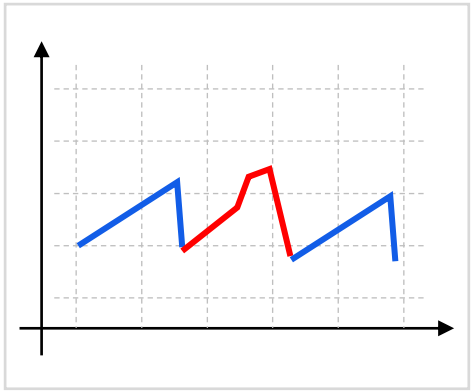
이상 탐지 기술 개발

시계열 이상 탐지 과정

시계열 입력 → 복원 → 탐지

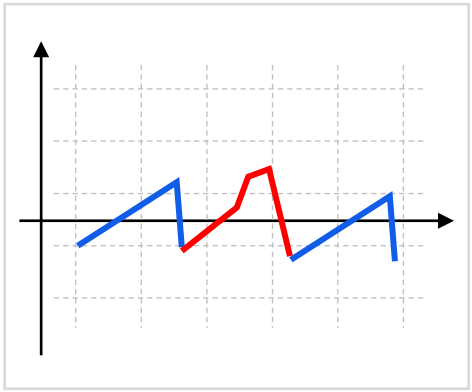
시계열 이상 탐지 과정

시계열 입력 — 복원 — 탐지



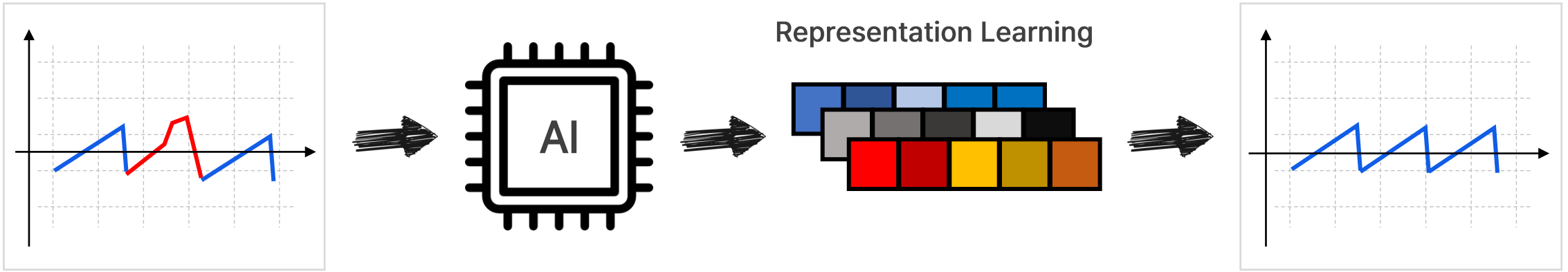
데이터 전처리
(Preprocessing)

정규화



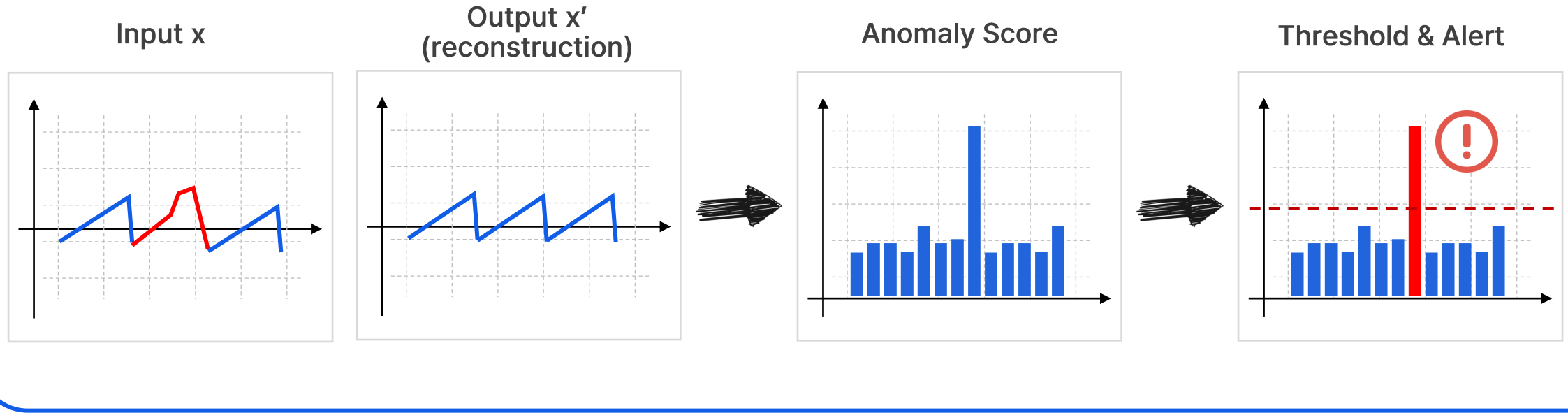
시계열 이상 탐지 과정

시계열 입력 → 복원 → 탐지



시계열 이상 탐지 과정

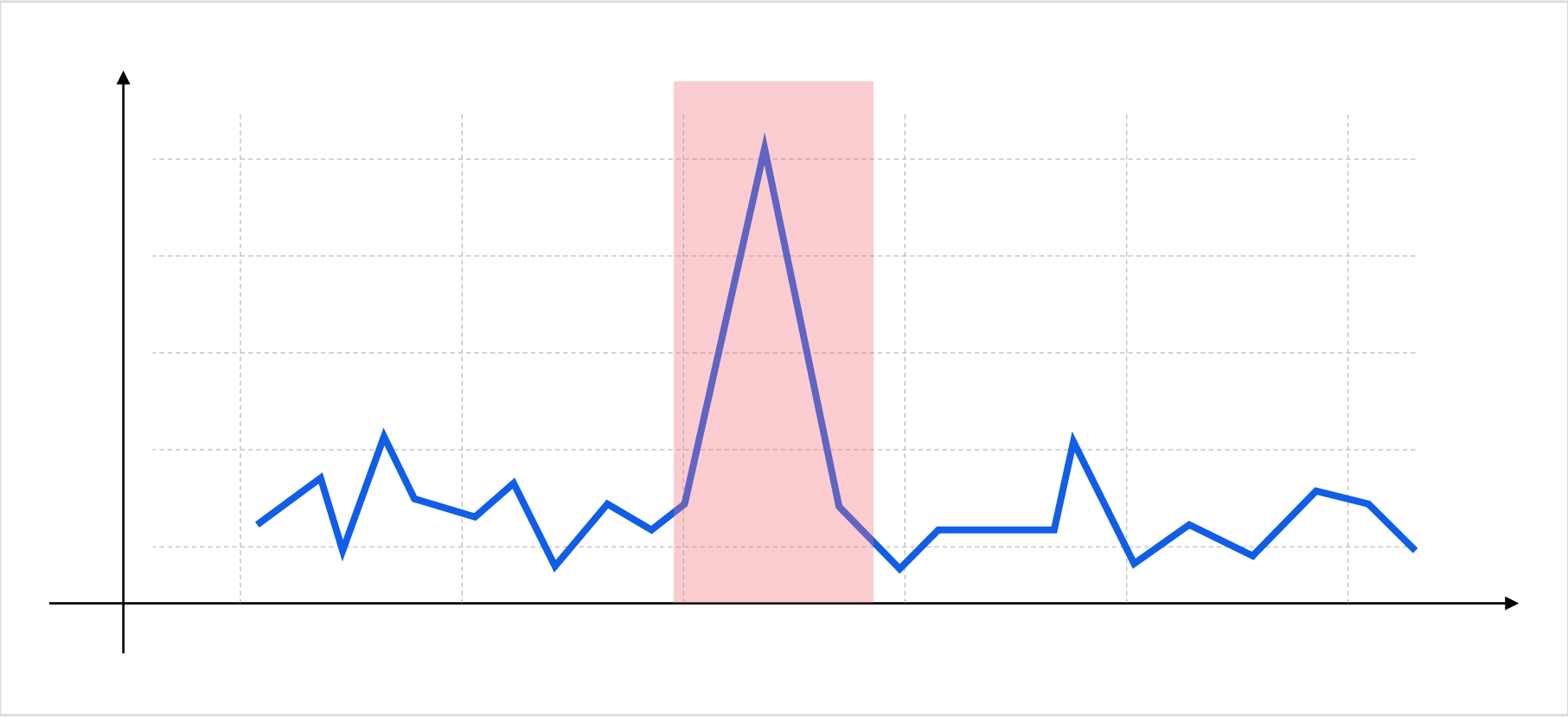
시계열 입력 → 복원 → 탐지



시계열에서의 이상 데이터

전역 이상(global anomaly)

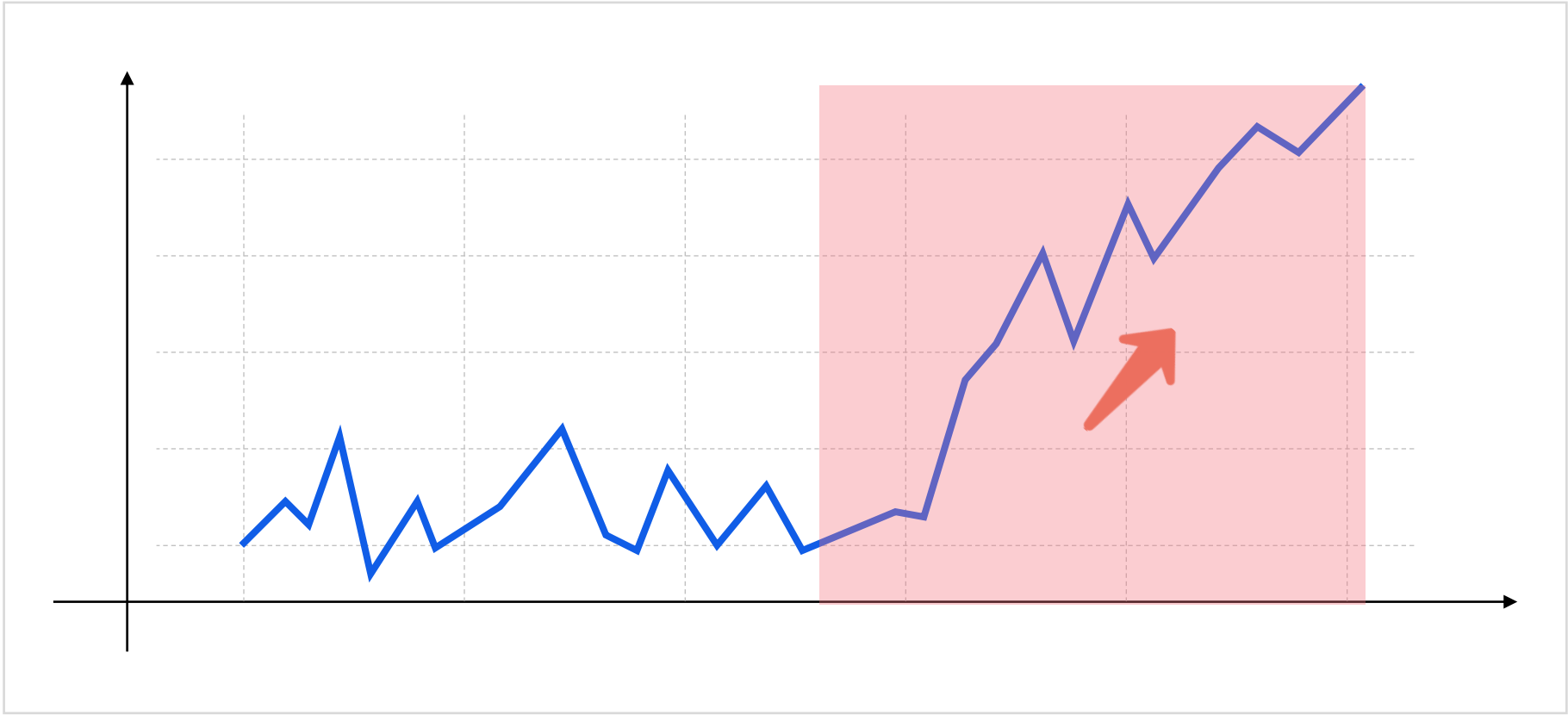
- 전역적인 범위에서 정상 범주로부터 큰 편차가 있는 데이터



시계열에서의 이상 데이터

추세 이상(trend anomaly)

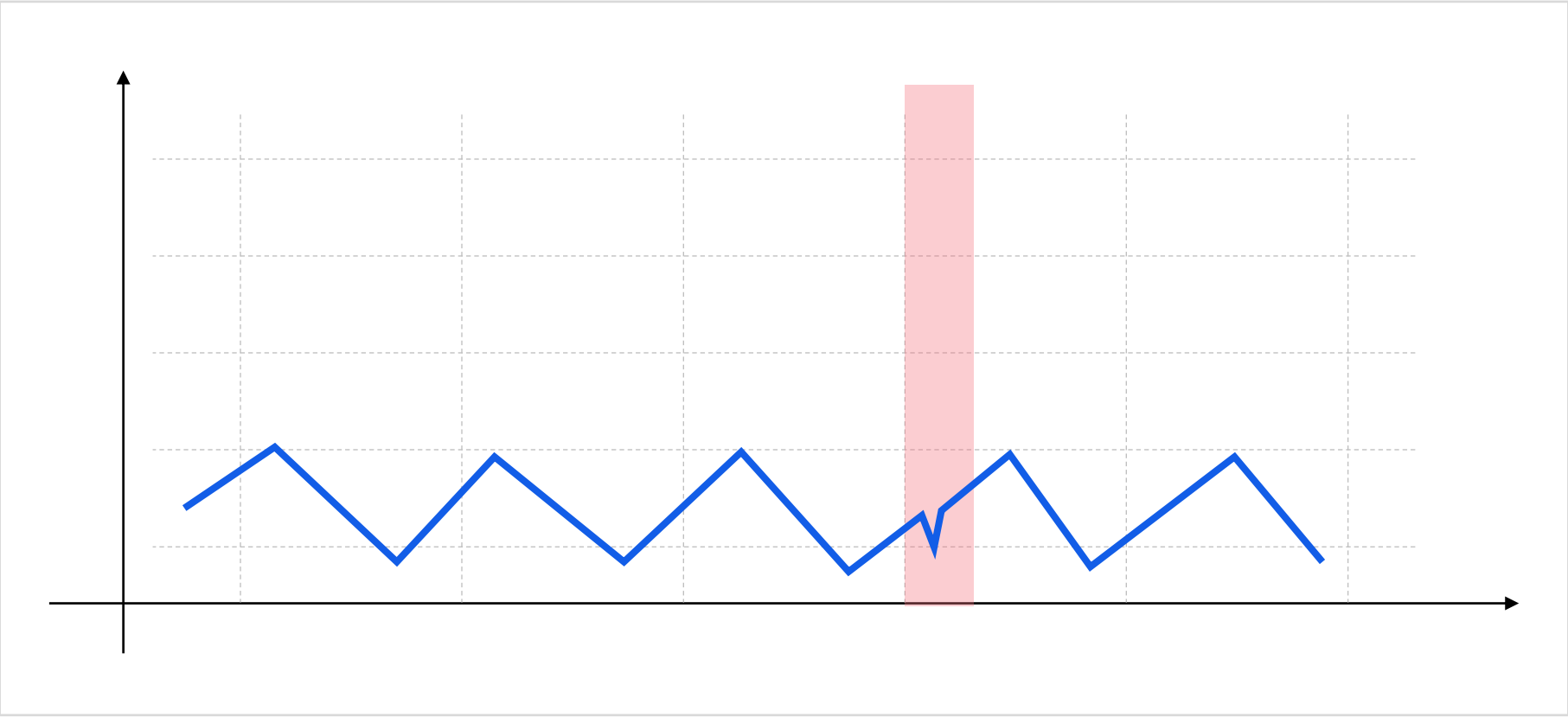
- 장기적인 패턴이나 방향에 이상이 있는 데이터
- 평균이 크게 변하는 데이터



시계열에서의 이상 데이터

맥락 이상(contextual anomaly)

- 특정 상황이나 문맥에서 정상 범주로부터 편차가 있는 데이터



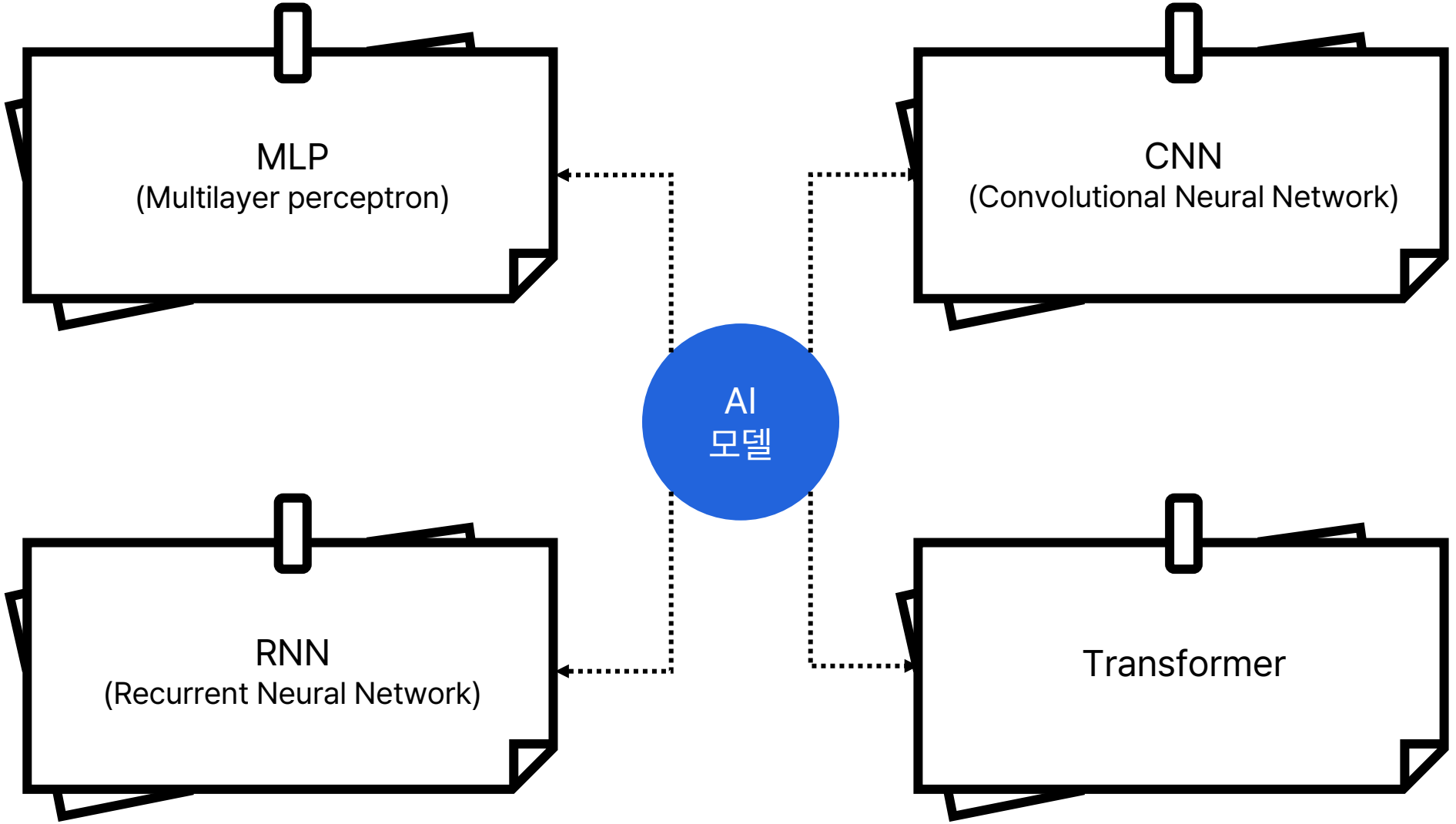
시계열 이상 탐지 딥 러닝 모델

어떤 모델들을 사용해야 할까?

- 데이터를 바라보는 관점에 따라 다양한 모델 적용 가능
- 시간의 흐름에 따라 변화하는 데이터의 특징을 학습하는 것이 중요

시계열 이상 탐지 딥 러닝 모델

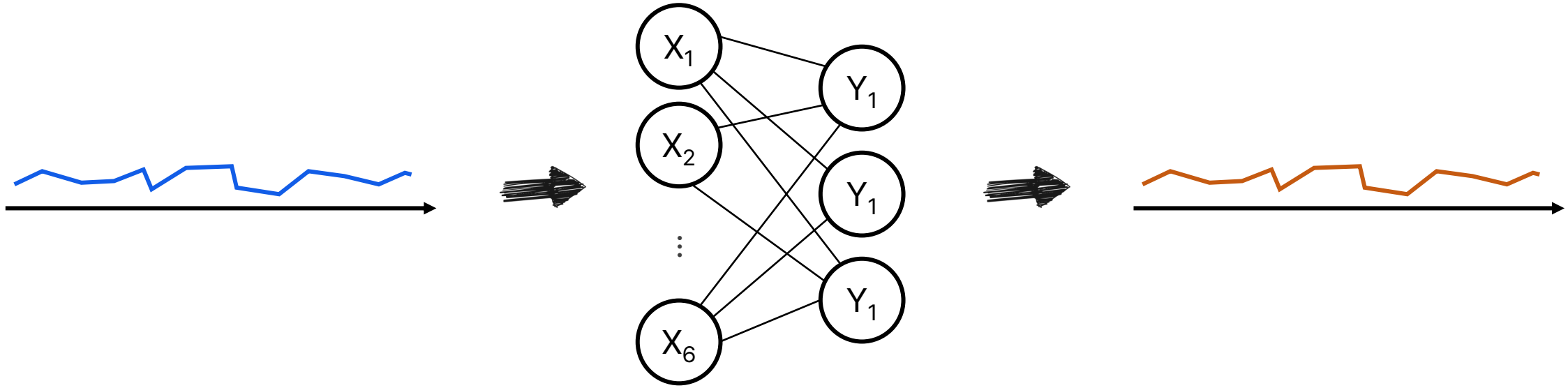
다양한 모델들이 존재



시계열 이상 탐지 딥 러닝 모델

Multilayer Perceptron

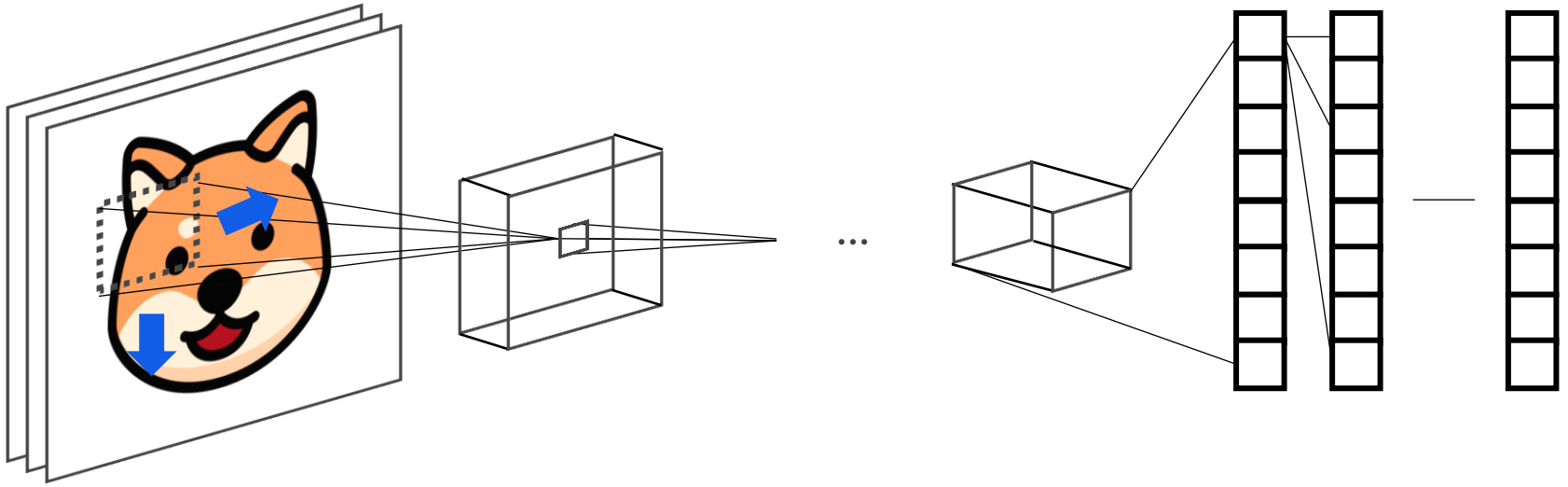
- 복잡한 딥 러닝 모델의 구성요소
- 다층 퍼셉트론을 사용하여 간단하게 모델 구성



시계열 이상 탐지 딥 러닝 모델

CNN(Convolutional Neural Network)

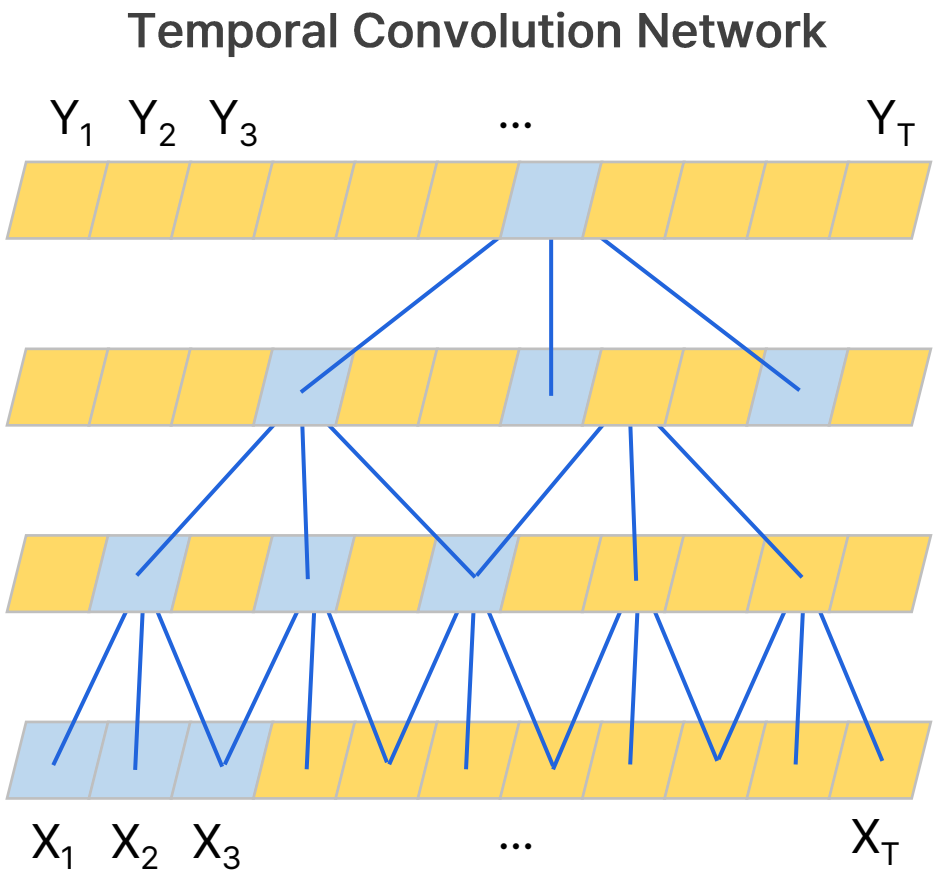
- 이미지, 그리드 데이터를 효과적으로 학습하는 모델
- 시계열 데이터를 다루기 위해 변형



시계열 이상 탐지 딥 러닝 모델

CNN(Convolutional Neural Network)

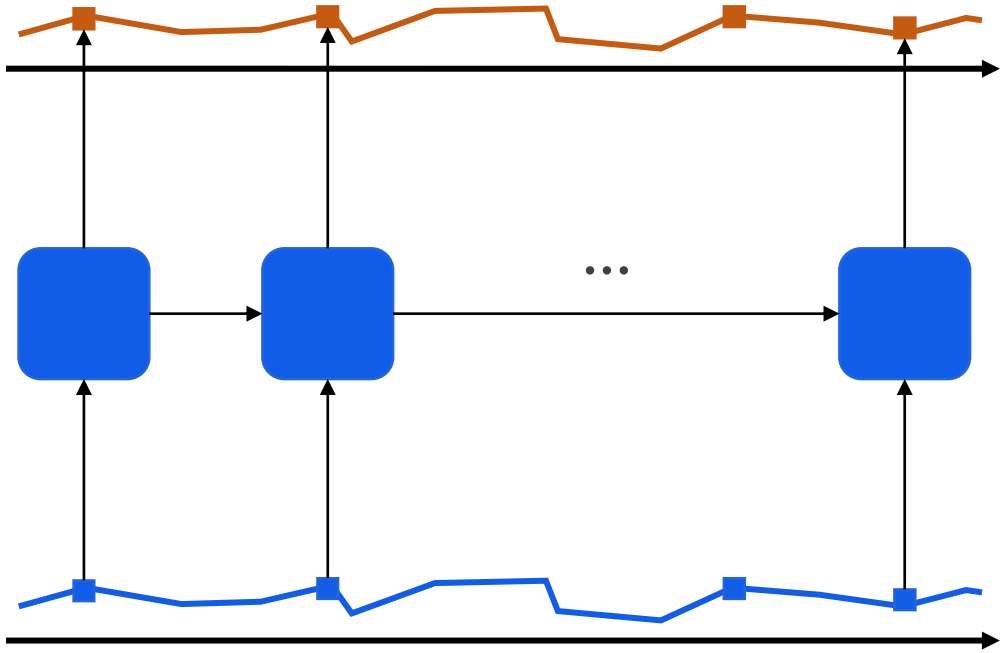
- 이미지, 그리드 데이터를 효과적으로 학습하는 모델
- 시계열 데이터를 다루기 위해 변형



시계열 이상 탐지 딥 러닝 모델

RNN(Recurrent Neural Network)

- 순차 데이터(sequence data)를 효과적으로 학습하는 모델
- 기존의 출력, 시계열 데이터를 순서대로 모델에 입력



시계열 이상 탐지 딥 러닝 모델

트랜스포머

- 순환 구조 없이도 순차 데이터를 효과적으로 학습
- GPT(Generative Pre-trained Transformer)에 사용된 모델
- 트랜스포머의 강점
  - 장기간 의존성
  - 병렬 연산
  - 복잡한 패턴 학습

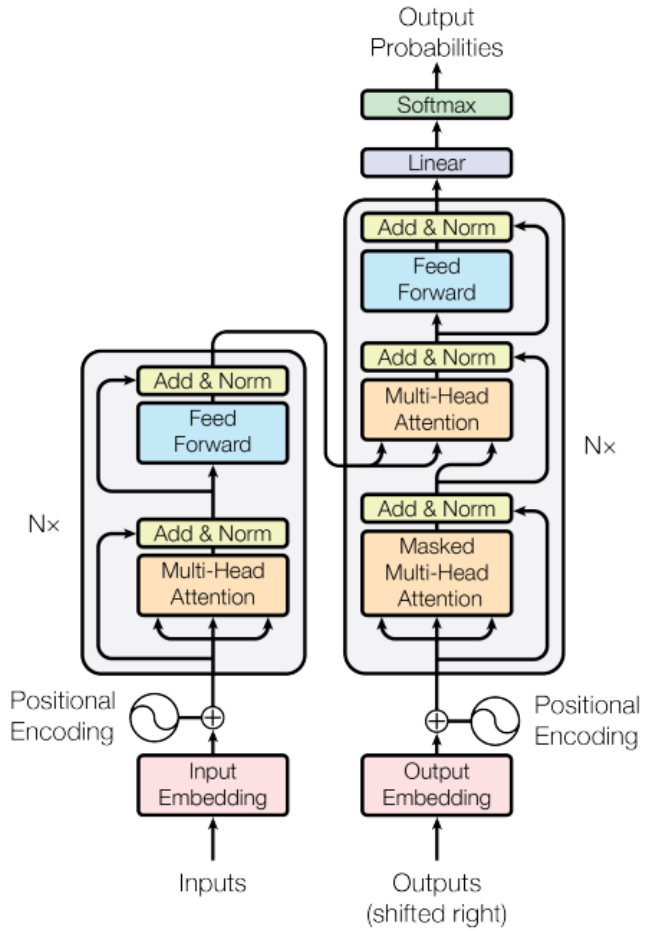
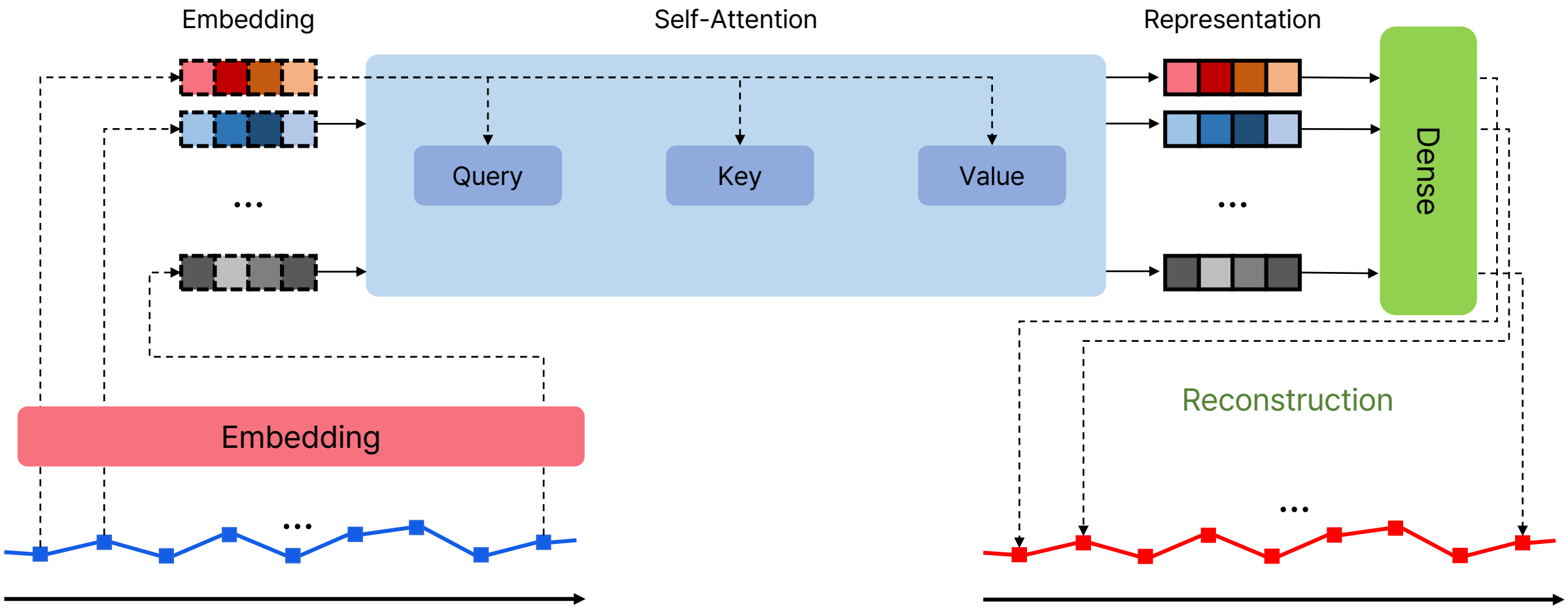


Figure 1: The Transformer - model architecture.

출처: https://arxiv.org/abs/1706.03762 "Attention Is All You Need"

트랜스포머, 자세히 알아보자

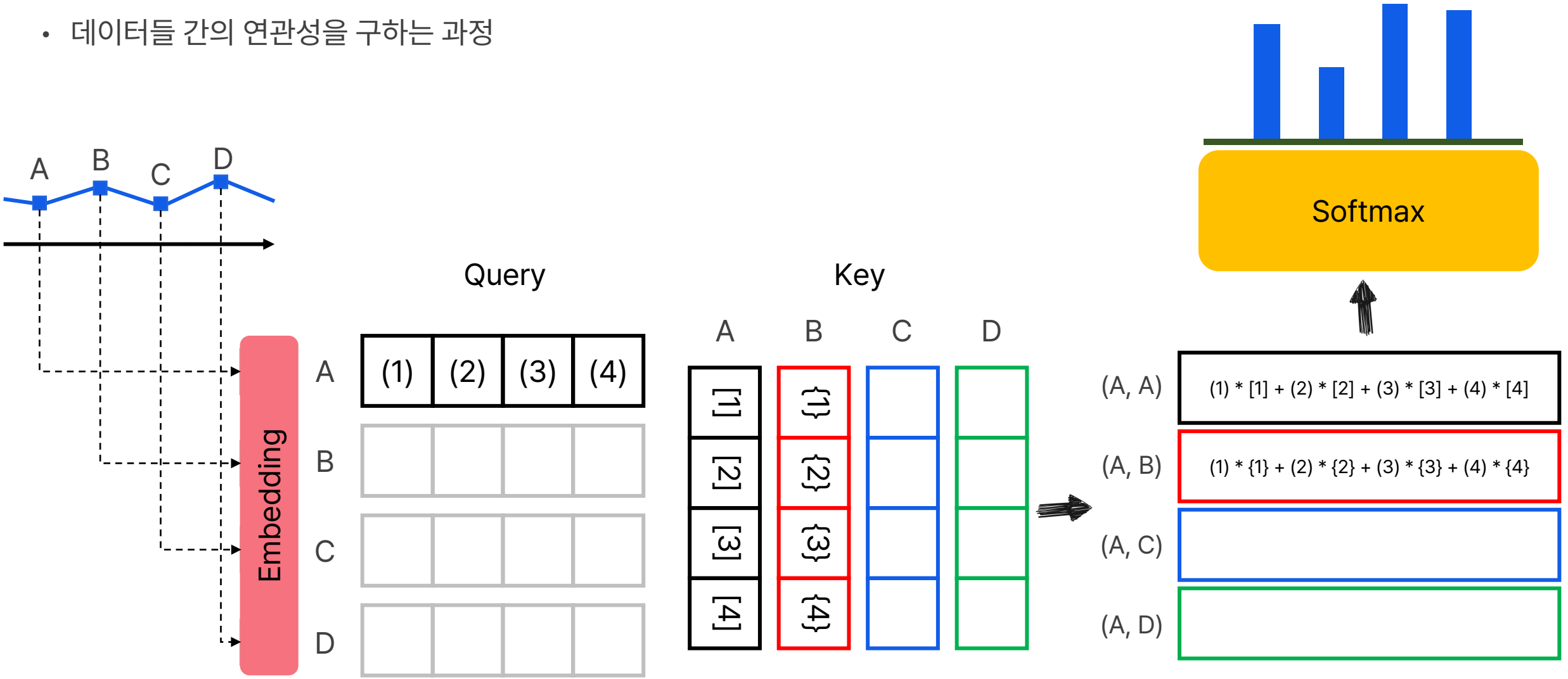
모델 구조



트랜스포머, 자세히 알아보자

Self-Attention

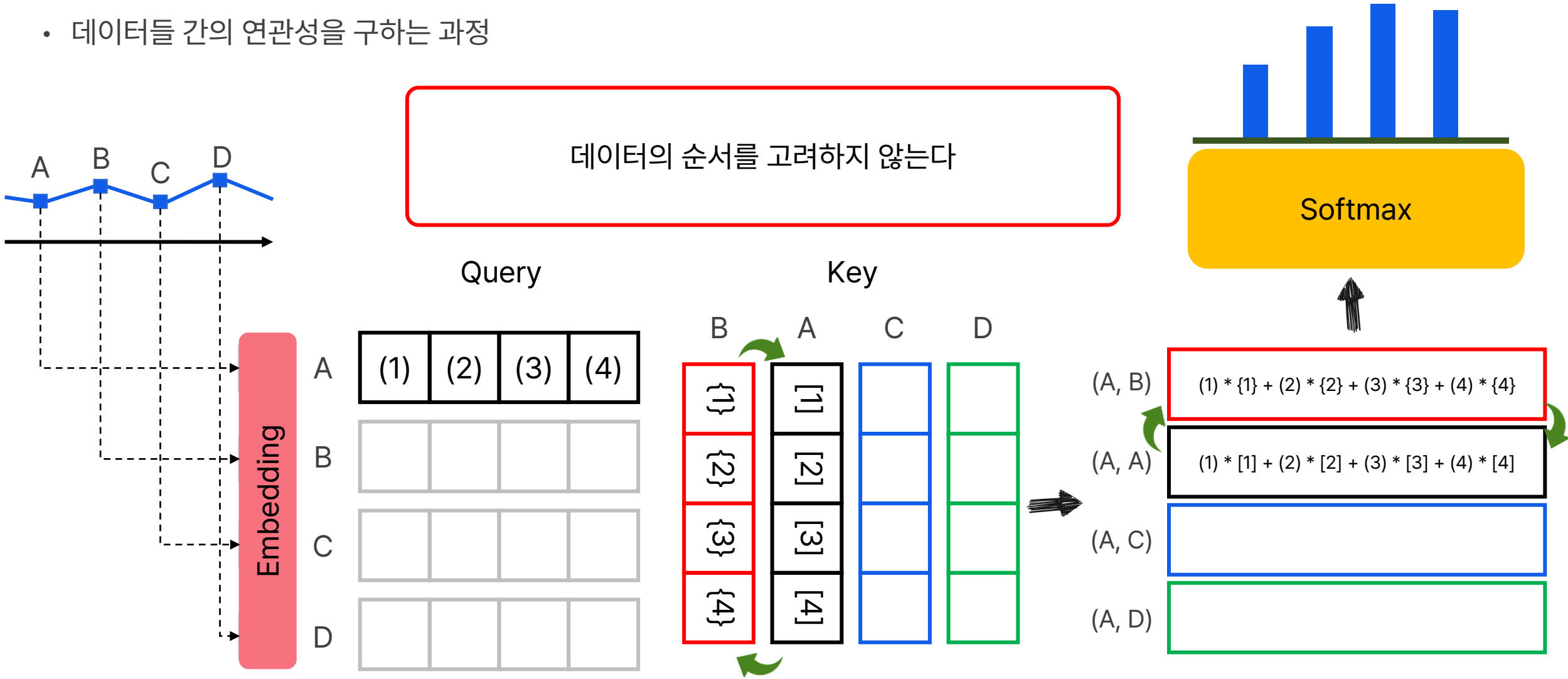
- 데이터들 간의 연관성을 구하는 과정



트랜스포머, 자세히 알아보자

Self-Attention

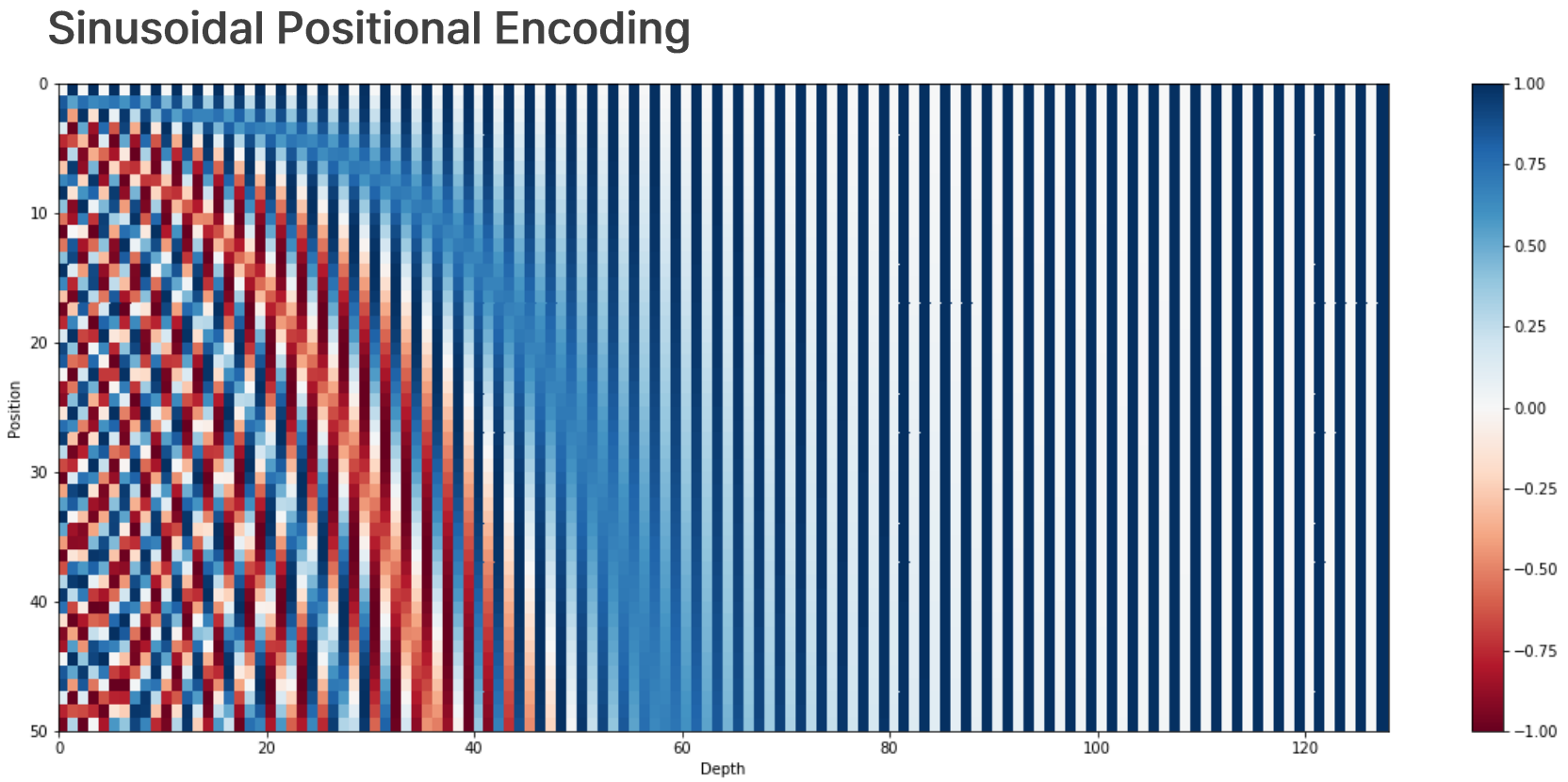
- 데이터들 간의 연관성을 구하는 과정



트랜스포머, 자세히 알아보자

Positional Encoding

- 위치 정보가 필요한 Self-attention의 문제점을 보완



시행 착오

실전에서 직면한 다양한 문제들

다양한 데이터 분포	2% 부족한 패턴 학습 능력	임계값 설정
시시각각 달라지는 데이터 분포	이벤트, 시간에 의한 패턴을 학습하지 못함	사용자마다 다른 이상 데이터 분포
데이터 일부로 전체 분포를 알 수 없음		사용자마다 다른 최적의 임계값
↓	↓	↓
데이터 정규화도 학습!	시간 정보를 함께 학습!	자동, 동적으로 임계값 탐색!

다양한 데이터의 분포

데이터 정규화

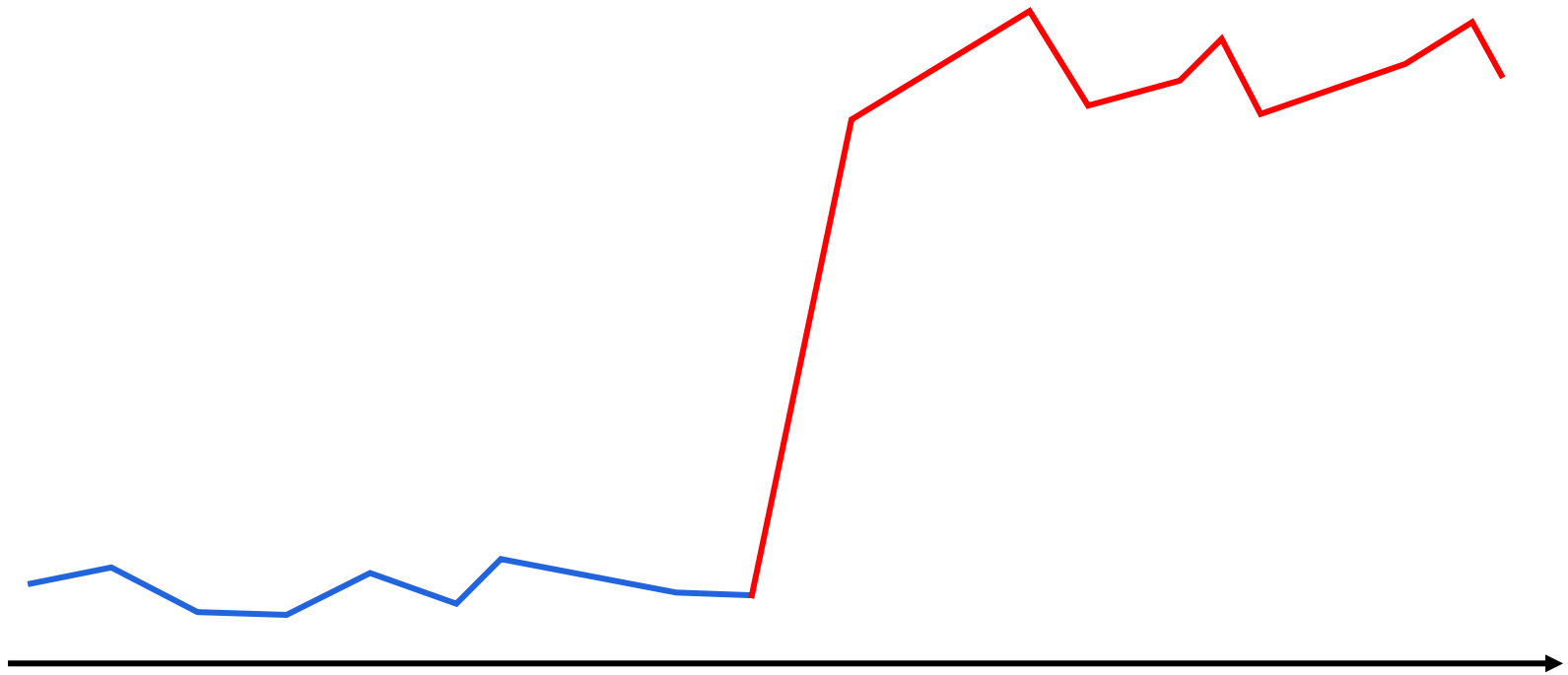
- 데이터의 분포를 일치시켜 주는 과정

안정적인 학습을 위해!

다양한 데이터의 분포

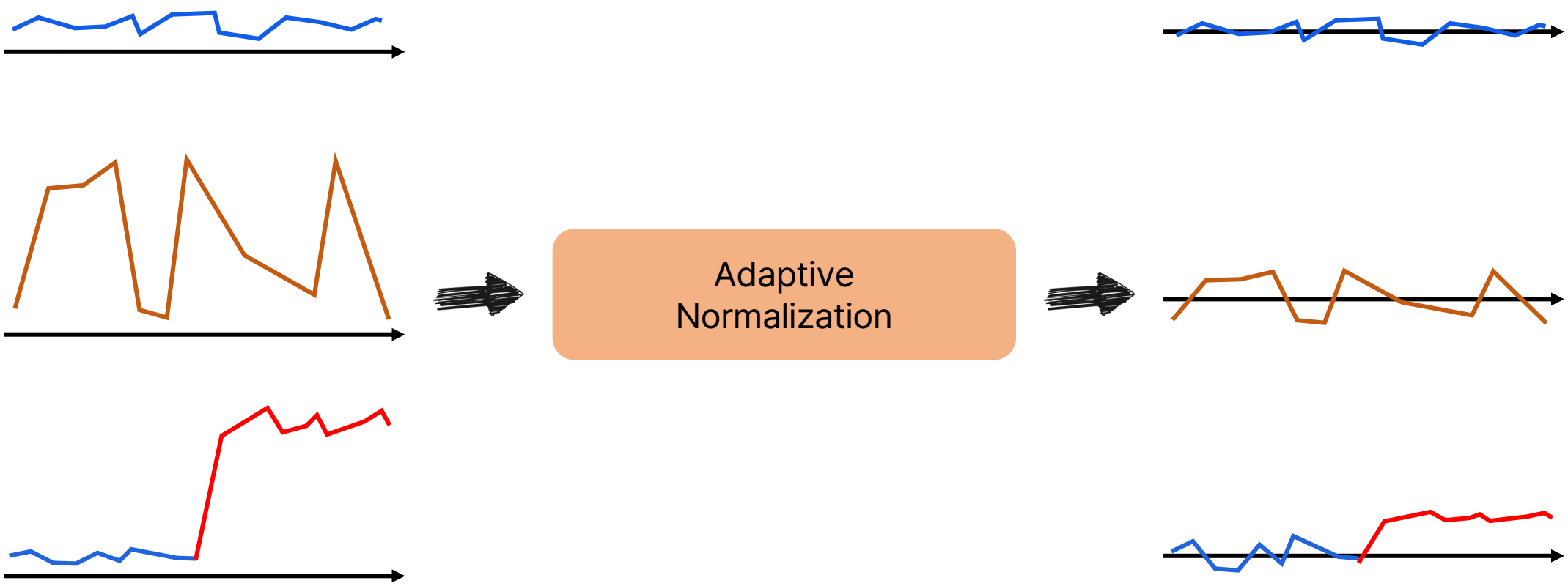
데이터 정규화, 생각보다 복잡했다

- 같은 데이터 안에서도 달라지는 데이터 분포



다양한 데이터의 분포

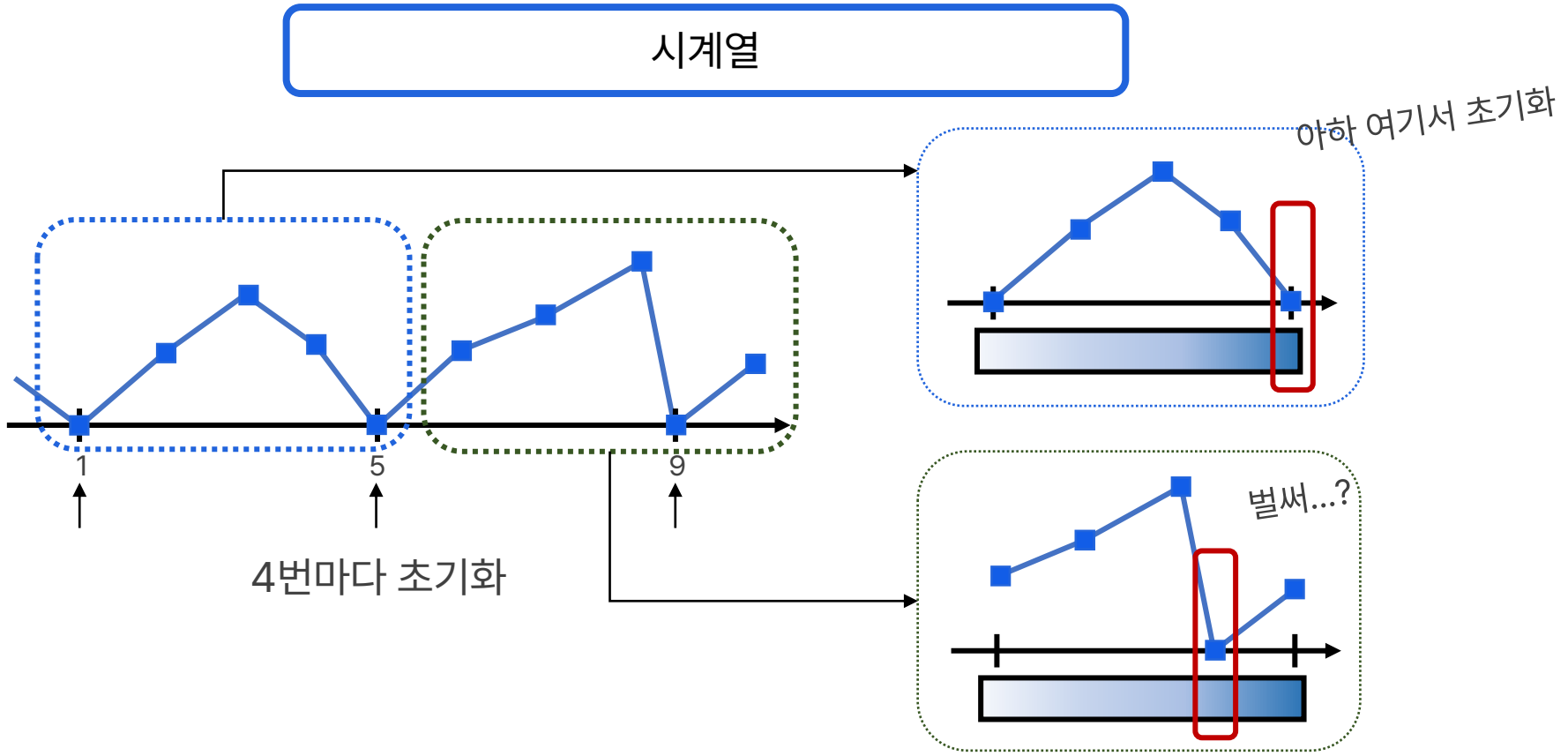
데이터 정규화, 학습하자!



2% 부족한 패턴 학습 능력

기존 Positional Encoding의 한계점

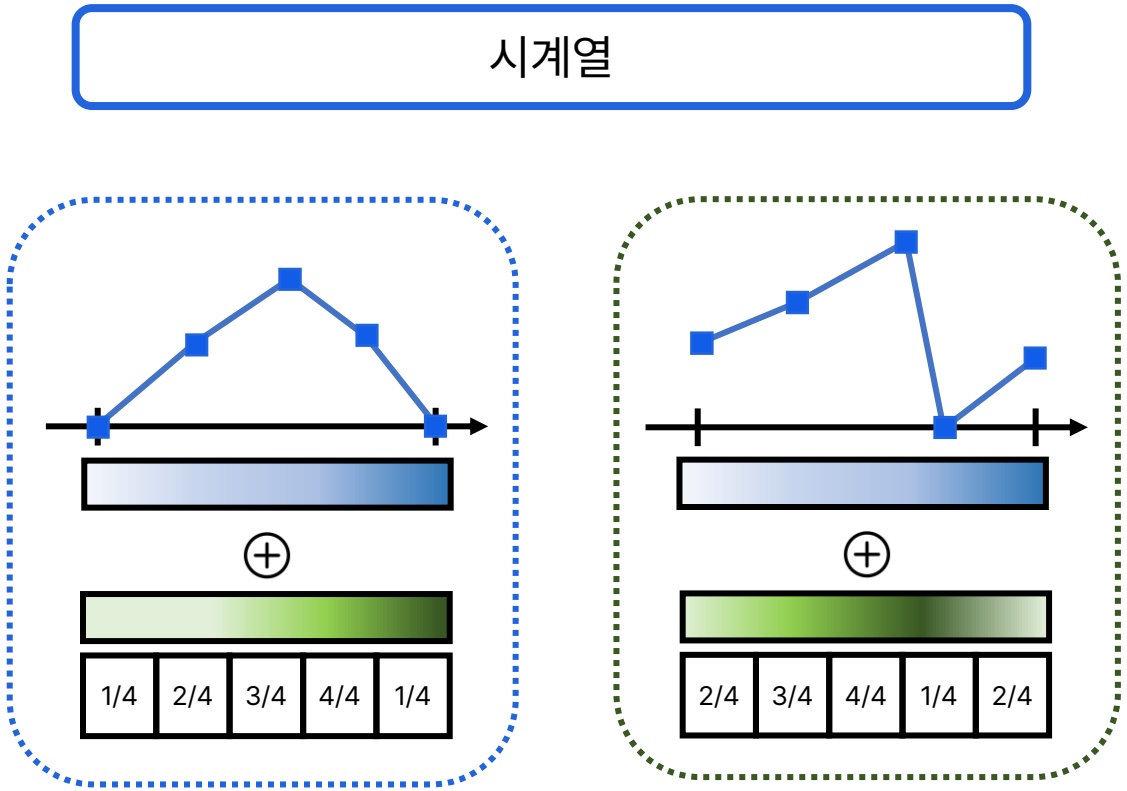
- 절대적인 위치 정보를 알 수 없음



2% 부족한 패턴 학습 능력

기존 Positional Encoding의 한계점

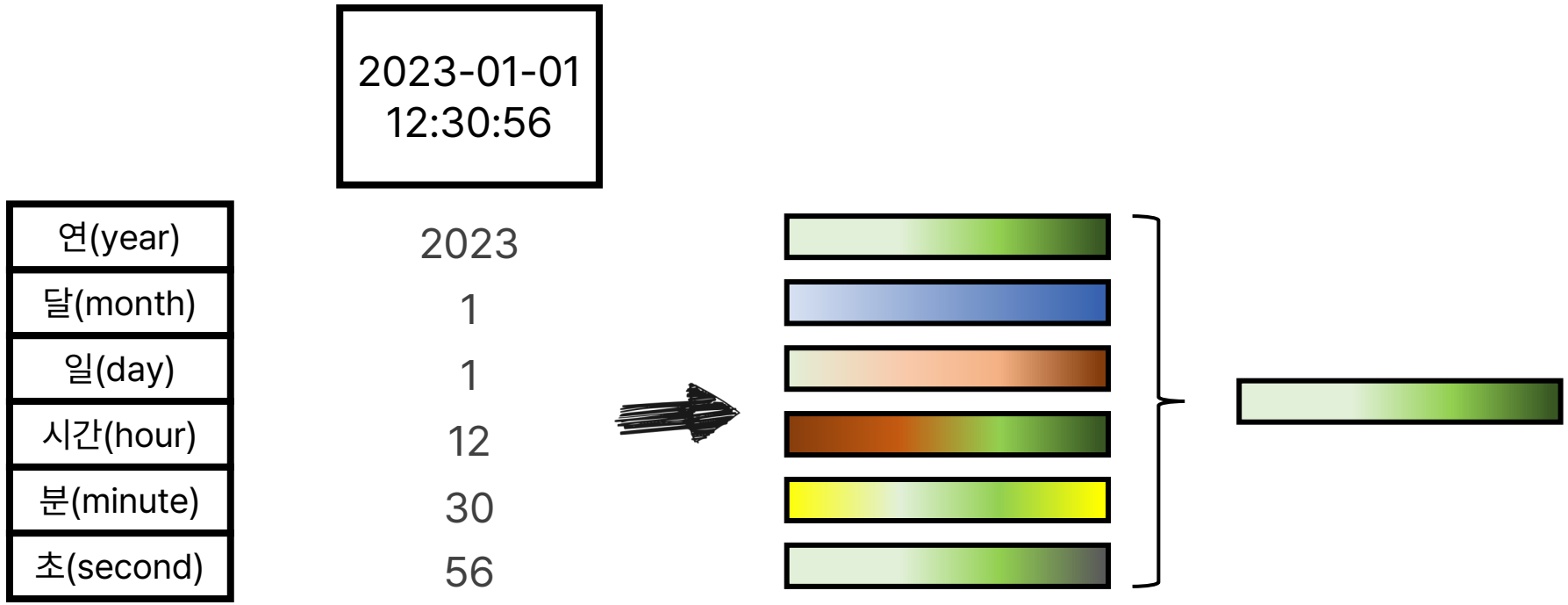
- 절대적인 위치 정보를 알 수 없음



2% 부족한 패턴 학습 능력

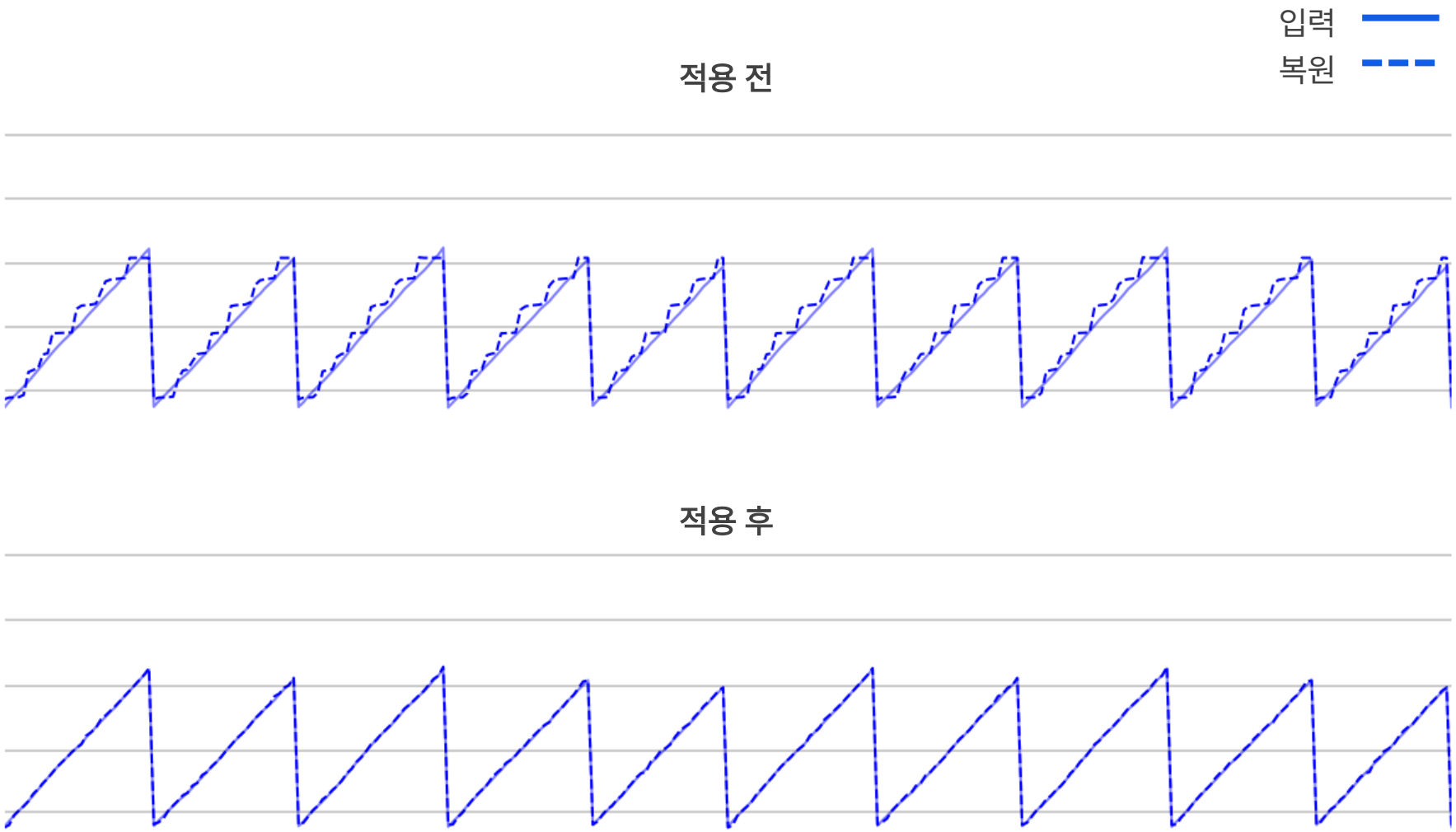
Timestamp Positional Encoding

- 시계열 데이터에서 쉽게 얻을 수 있는 시간 정보
- 시간 정보를 사용하여 인코딩 형성
- 데이터의 주기성, 특별한 이벤트 정보(명절, 휴일 등)를 반영할 수 있음



2% 부족한 패턴 학습 능력

Timestamp Positional Encoding



탐지의 마무리는 임계값 설정

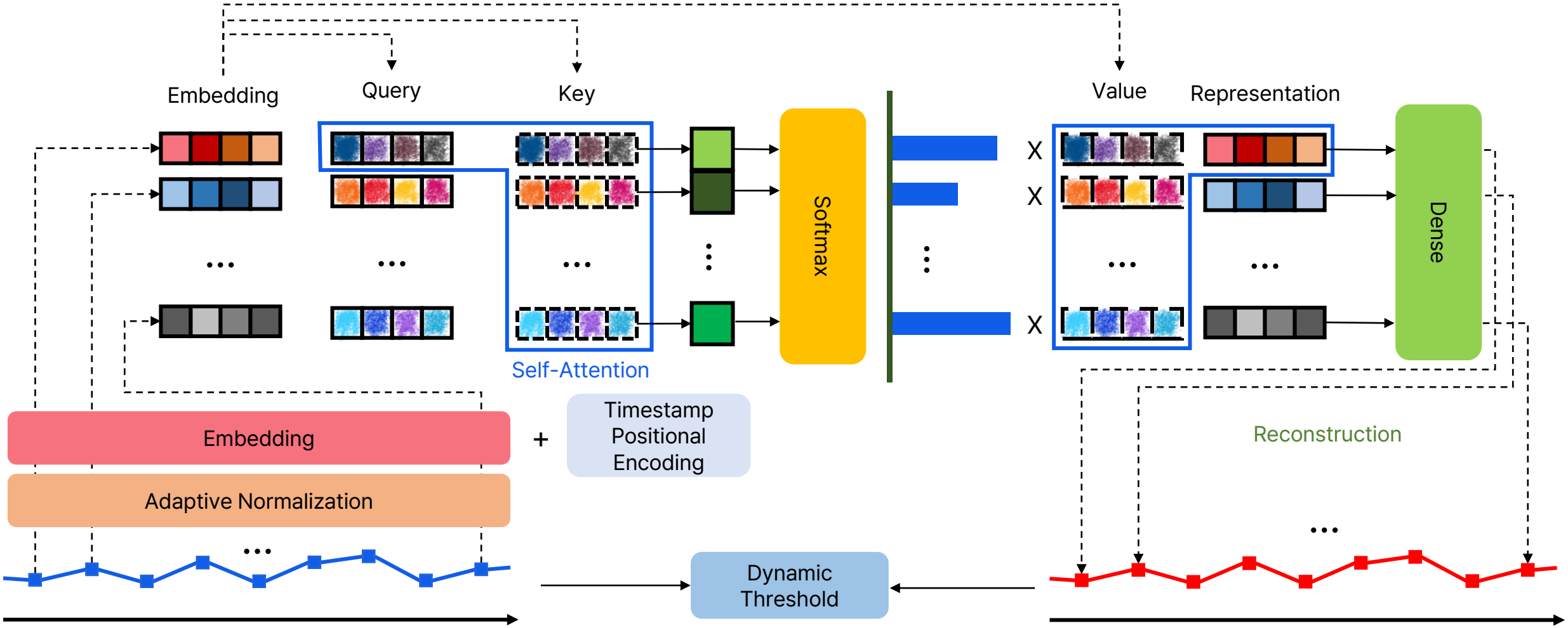
정답이 있긴 한 걸까..?

- 어노말리 스코어에 임계치를 설정하여 탐지
- 적절한 임계값을 설정하는 것은 매우 중요한 과정
- 아직도 많은 연구가 필요한 어려운 문제

정적(Static)	동적(Dynamic)
미리 정한 규칙에 따라 임계값 고정	조건에 따라 동적으로 임계값 조절
사전에 충분한 데이터 확보가 필요	유연하게 대처가 가능
사용자, 환경에 따라 유연하게 대처할 수 없음	아직은 많은 연구가 필요

트랜스포머, 자세히 알아보자

모델 구조



NHN Cloud make IT 2023

이상 탐지 서비스

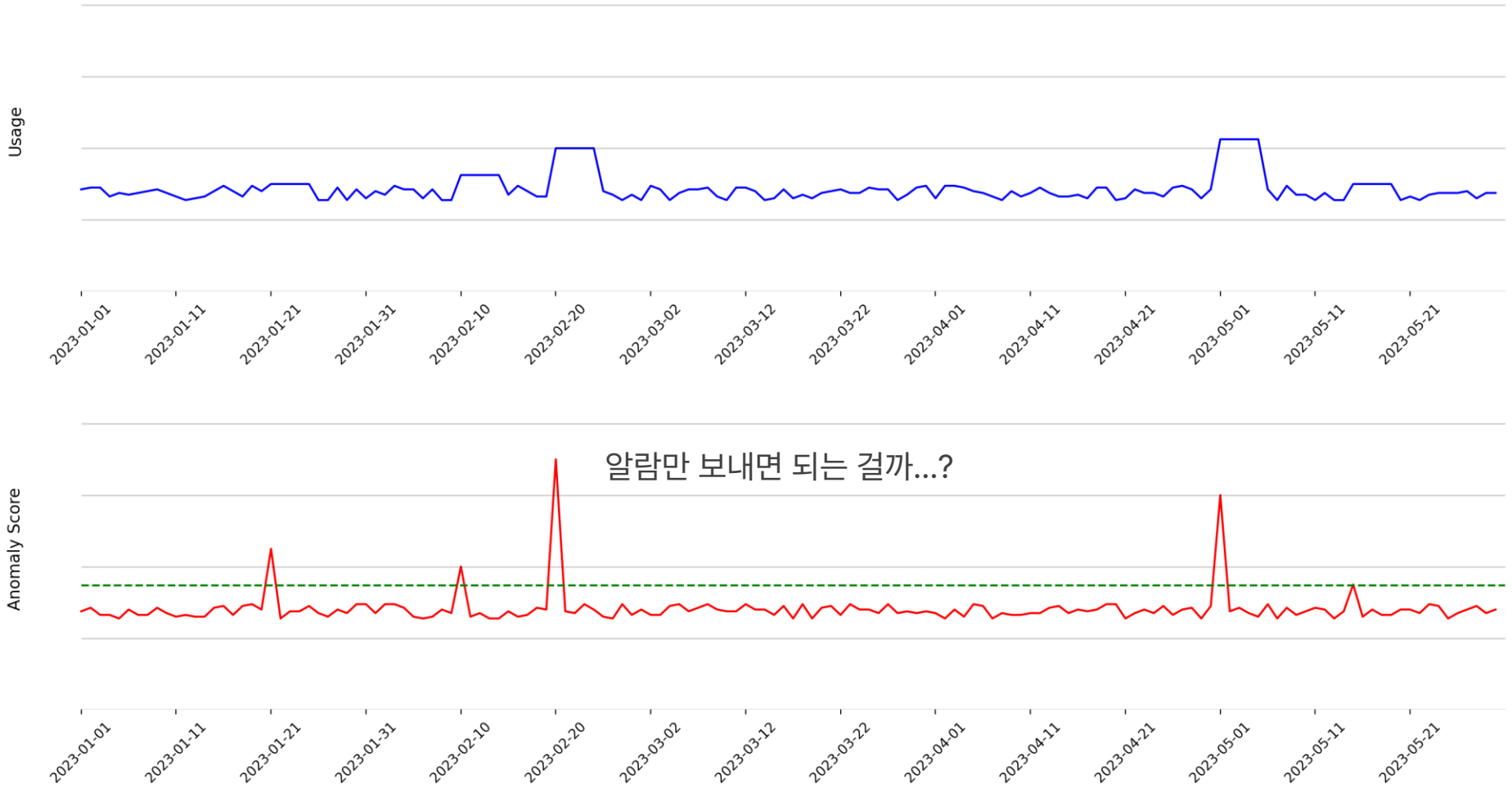
이상 탐지 서비스

어떤 서비스를 제공할 수 있을까?

- 핵심 기능
  - 신뢰 구간(confidence interval)
  - 다변량 & 단변량 이상 탐지
  - 원인 분석(root cause analysis)
  - 임계값 설정(threshold setting)

다소 아쉬웠던 정보량

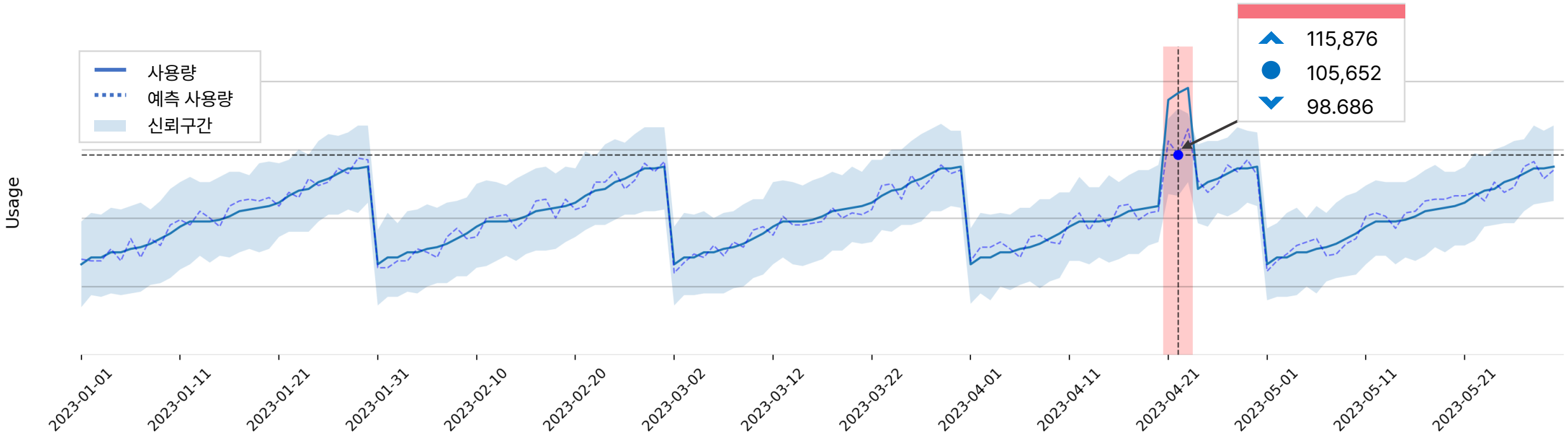
사용자가 보게 될 화면



사용자에게 도움이 되는 정보

신뢰 구간(confidence interval)

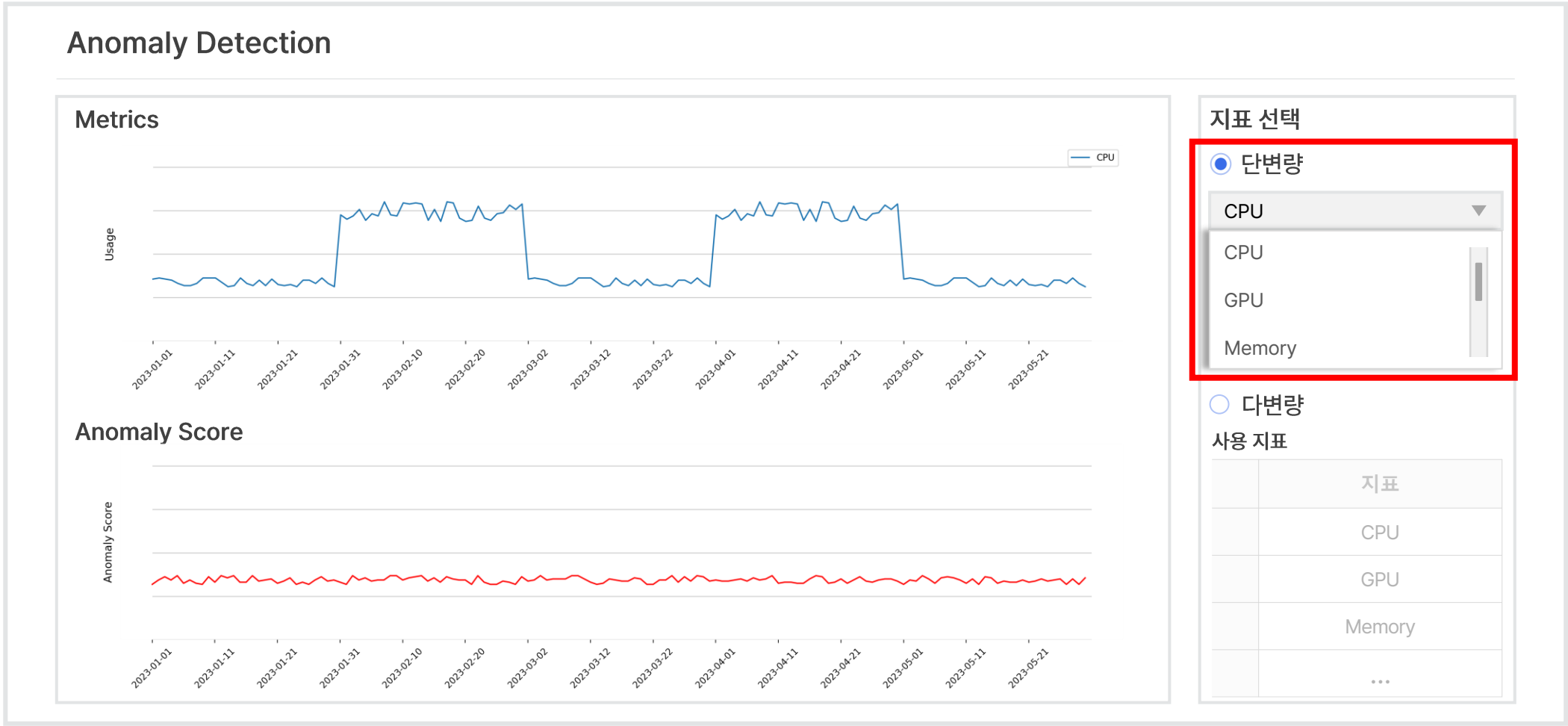
- 탐지된 이상 데이터의 직관적인 이해
- 이상 탐지에서 더 나아가 데이터를 이해



이상 탐지 서비스

다변량 & 단변량 이상 탐지

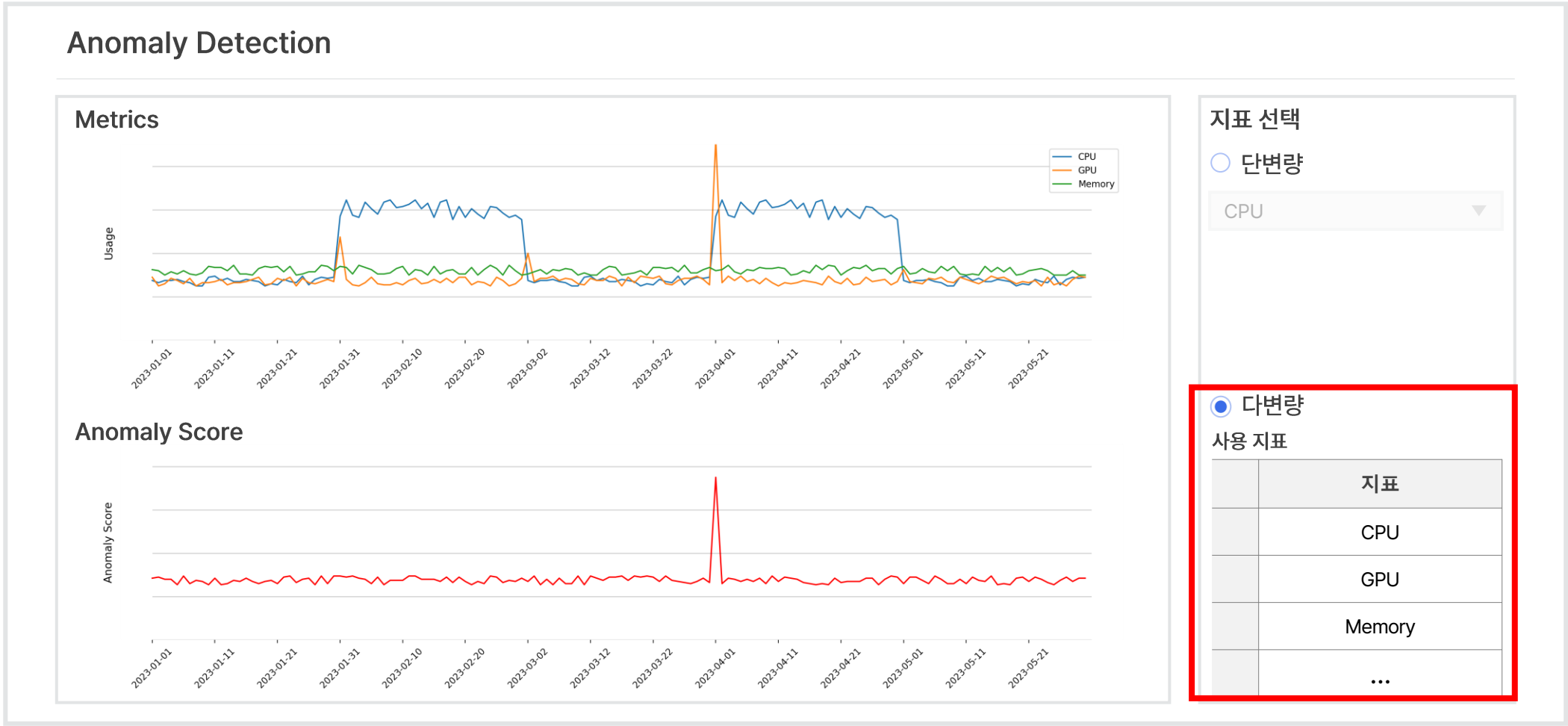
- 단일 지표, 전체 지표(다변량) 중 원하는 지표 선택



이상 탐지 서비스

다변량 & 단변량 이상 탐지

- 단일 지표, 전체 지표(다변량) 중 원하는 지표 선택



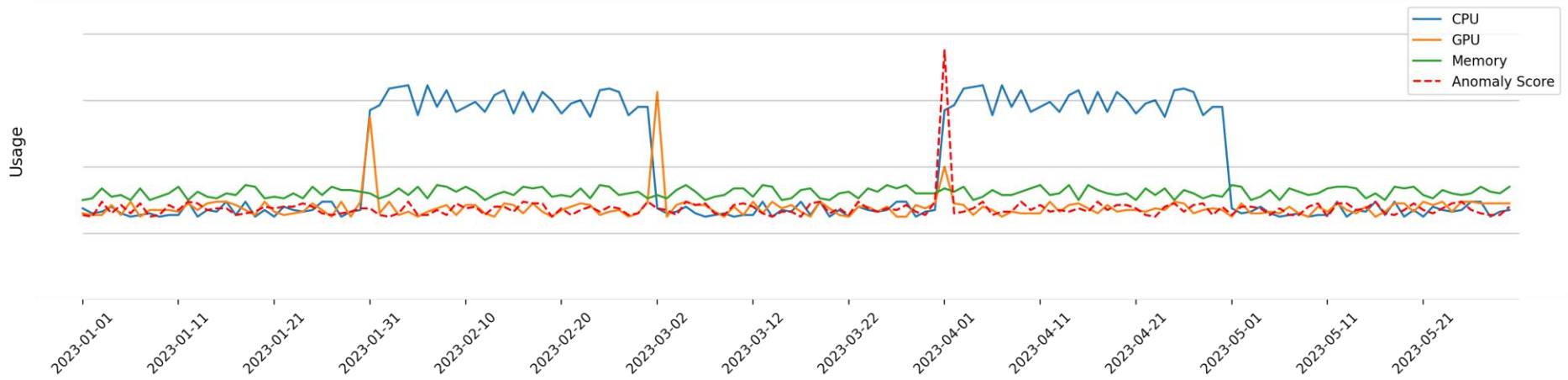
이상 탐지 서비스

원인 분석(root cause analysis)

- 탐지된 이상 데이터의 원인 분석

Root Cause Analysis

Metrics



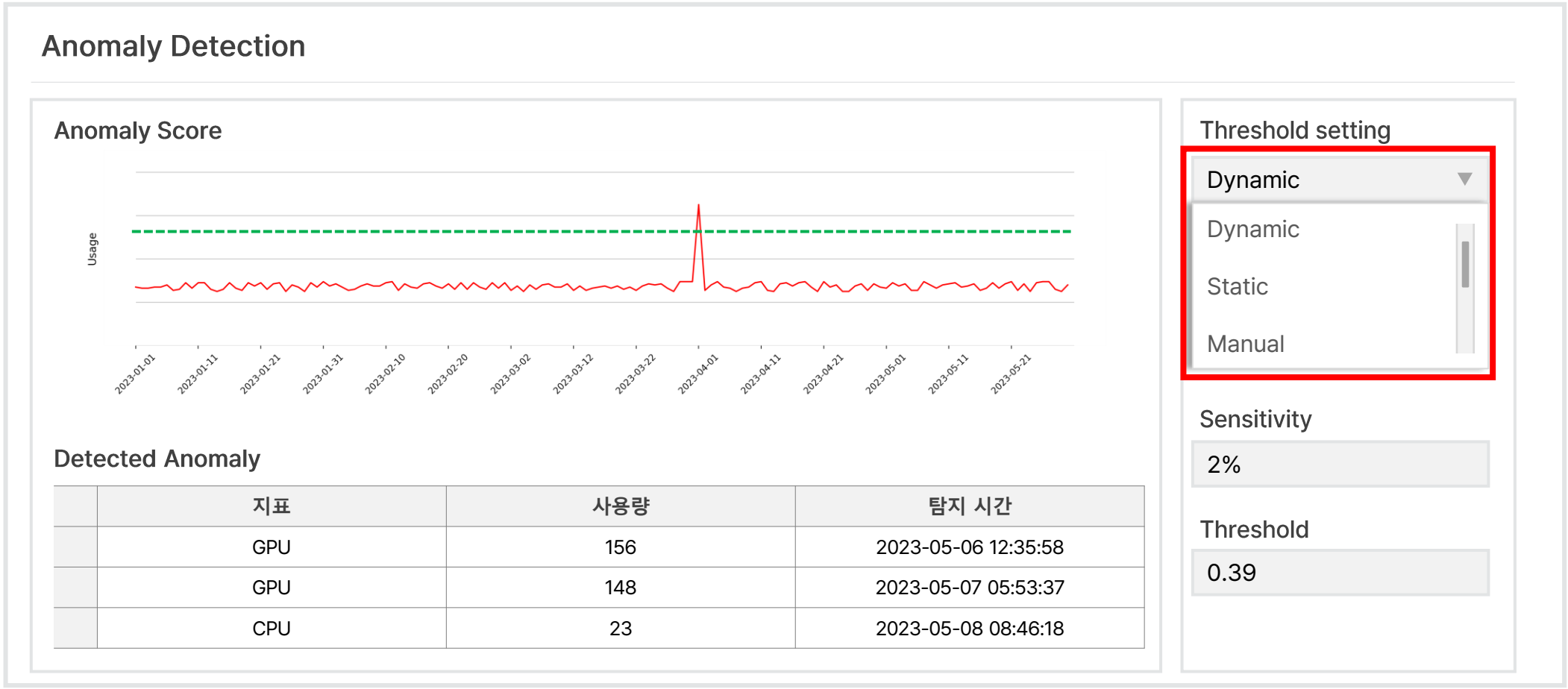
Cause Analysis

	지표	영향도(%)
	CPU	2.46
	GPU	95.96
	Memory	1.58

이상 탐지 서비스

임계값 설정(threshold setting)

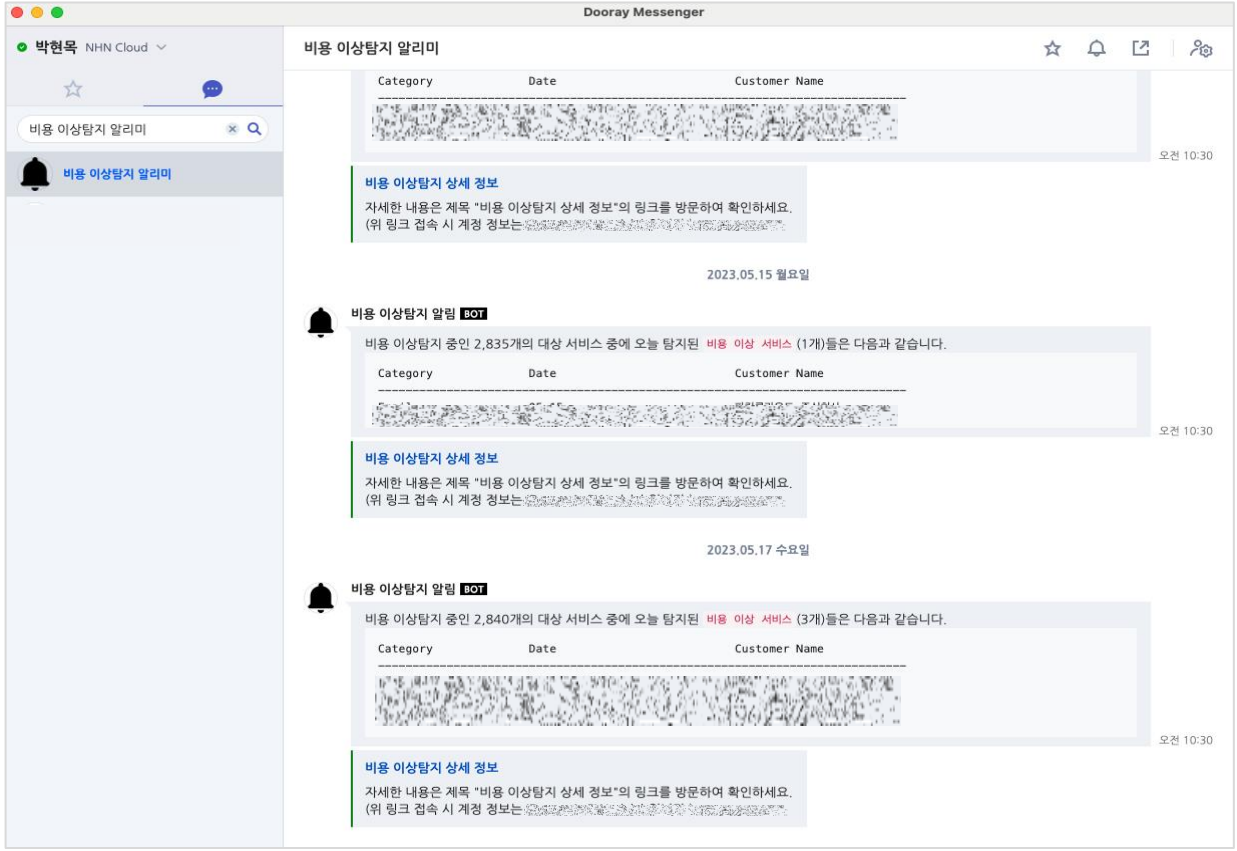
- 동적, 정적, 매뉴얼 등 자유롭게 설정 가능



실전을 통한 경험 축적

비용 이상 탐지

- 클라우드 서비스 비용 대상으로 시범 서비스 운영
- 이메일, 메신저, SMS 알림



마치며

열심히 연구 중입니다...

- 정확하고 안정적인 이상 탐지를 위한 꾸준한 연구
- 비용, 모니터링, 서비스, 보안 등 다양한 분야로의 확장

Q&A

NHN Cloud make IT 2023

고맙습니다.