

Big Data Analytics Programming

Week-08. Anomaly Detection

Jungwon Seo, 2021-Spring

Anomaly Detection

Overview

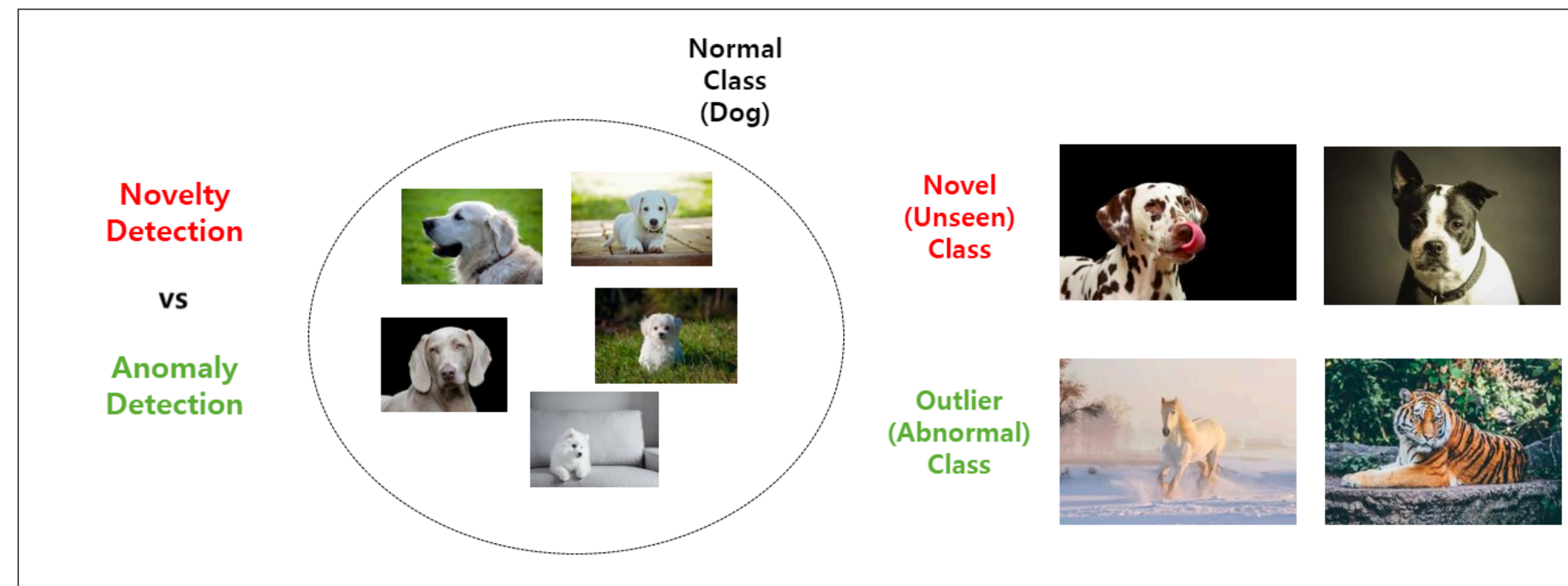
- Anomaly란?
- 전통적인 접근 법
- DNN 기반의 접근법

이상치란?

Anomaly

이상치의 정의

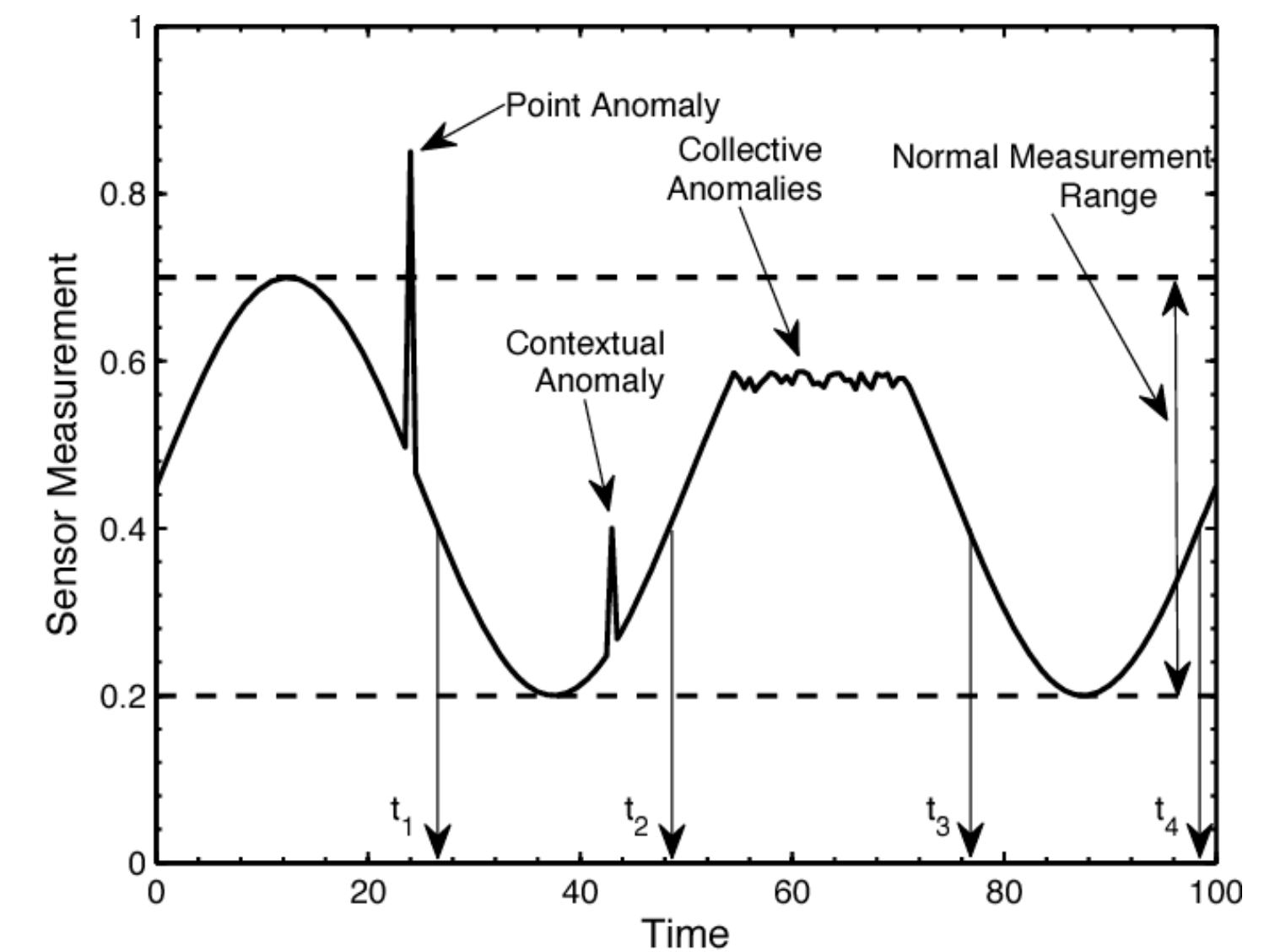
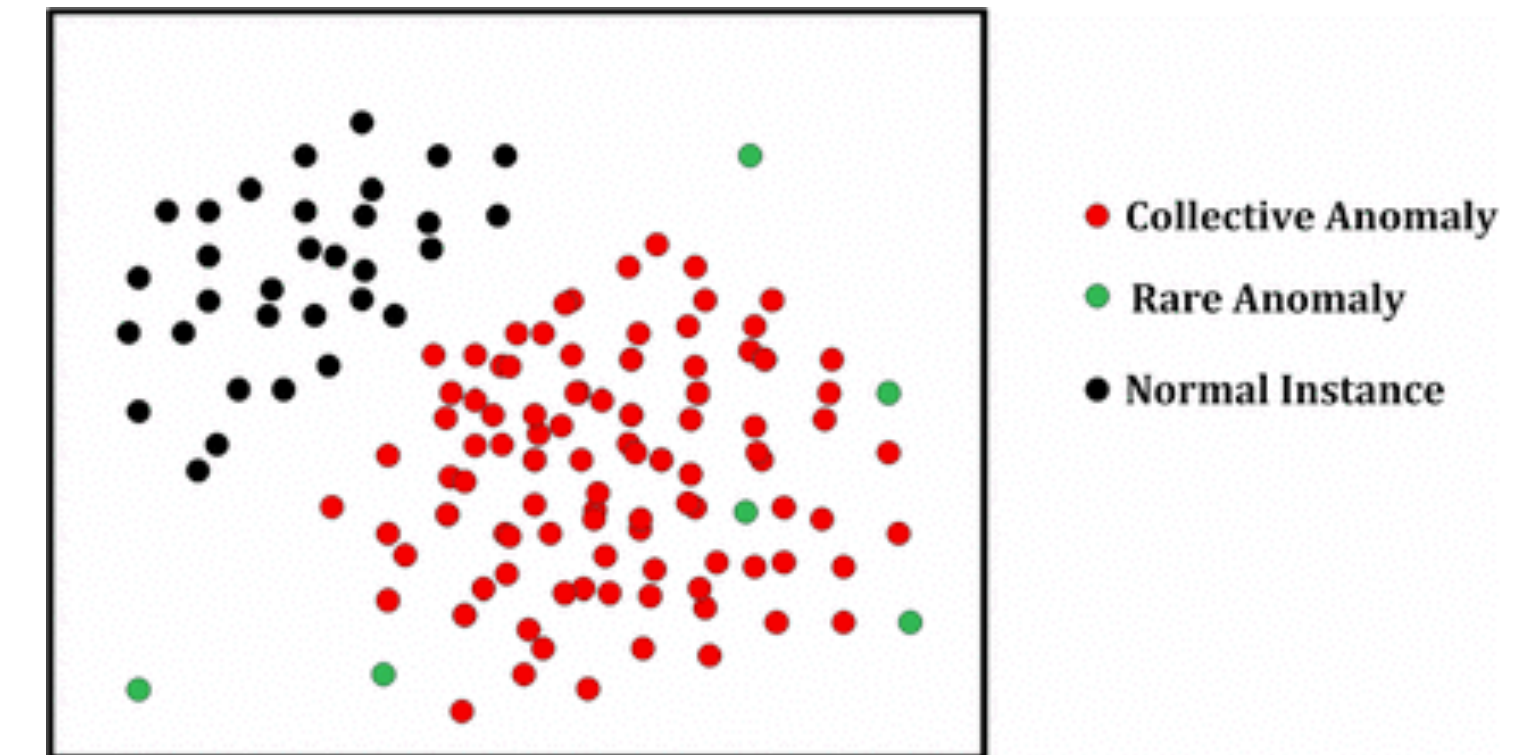
- 대부분의 데이터와 **다른** 희귀 데이터
 - Novelty, Outlier로도 불리울수 있지만, 약간의 뉘앙스 차이가 존재
 - Novelty: 같은 부류지만 이전에 본적이 없는 (Unseen)
 - Novelty Detection: 지금까지는 등장하지 않았지만, 충분히 등장 할 수 있는 sample 찾는 것
 - Outlier: 전혀 관련이 없는
 - Outlier Detection: 등장할 가능성이 데이터에 오염이 발생했을 가능 성이 있는 sample을 찾는 것



Anomaly

Data point of view

- Point anomaly
 - 데이터 셋 하나가 이상
 - 독립적으로 여러개
 - 일반적인 Outlier (정상 데이터 범위 내에 없는)
- Collective anomaly
 - 데이터 셋 내의 여러 **관련된/연결된** 데이터셋이 이상
 - 다른 특성과의 관계가 깨져버린 상황
- Contextual anomaly
 - 전체적인 데이터셋의 맥락을 고려했을 때 이상



(c) Point, Contextual and Collective Anomalies [10]

Anomaly Detection

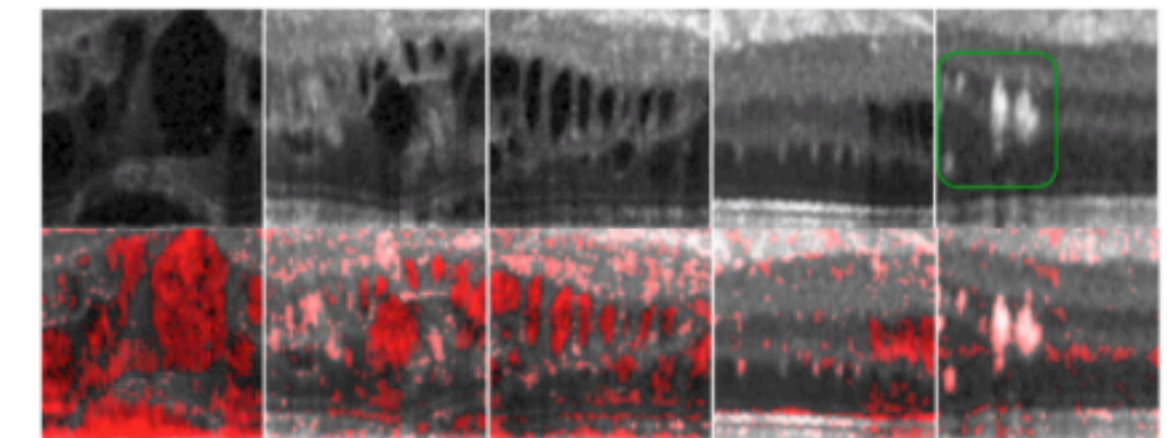
Anomaly Detection은 어디에 쓸 수 있을까?

- 사례

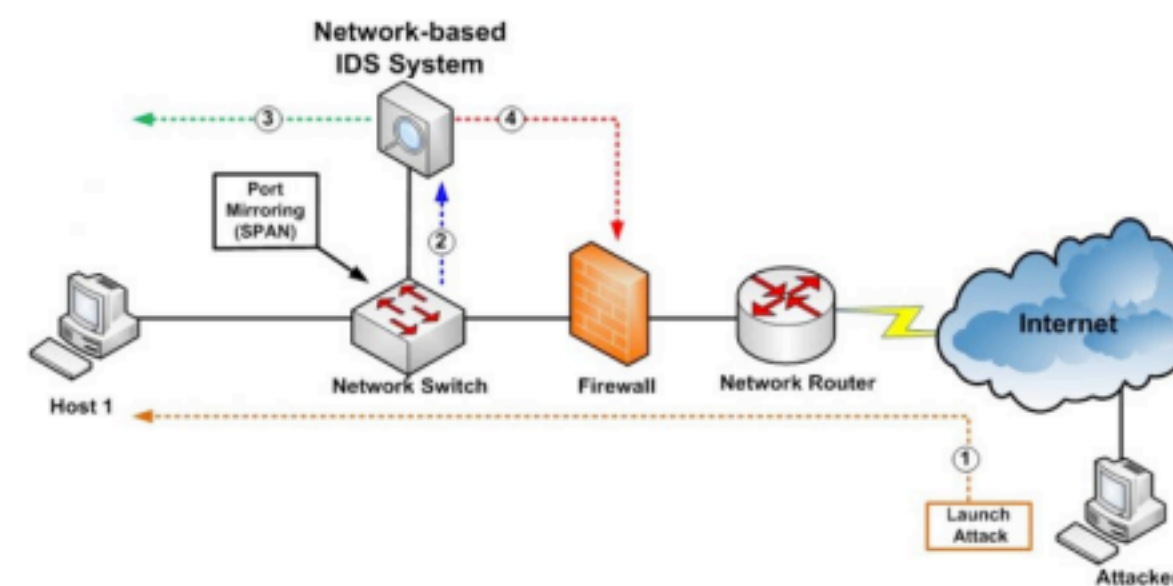
- 신용 카드 사기 탐지
- 통신 사기 탐지
- 네트워크 침입 탐지, 결함 탐지
- Video Surveillance
- 제조업 공정과정에서 이상탐지



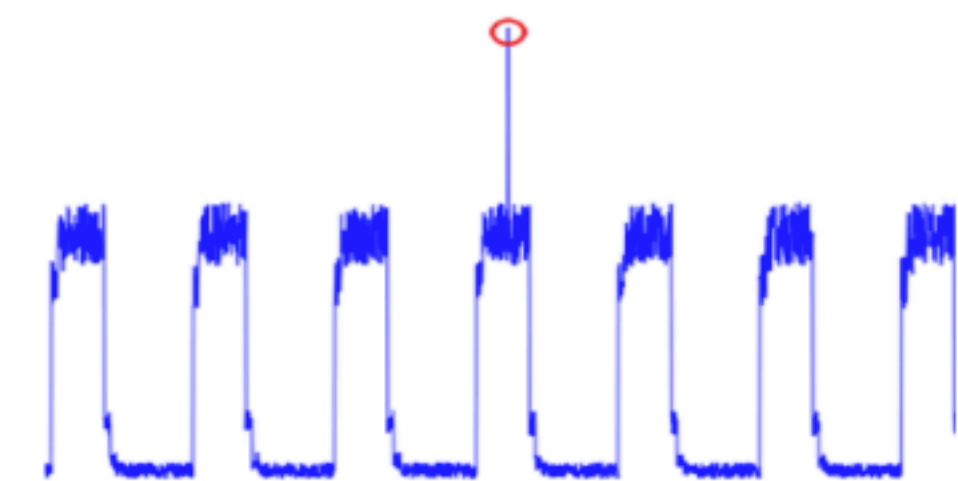
(a) Illegal Traffic Flow detection



(b) Detecting Retinal Damage



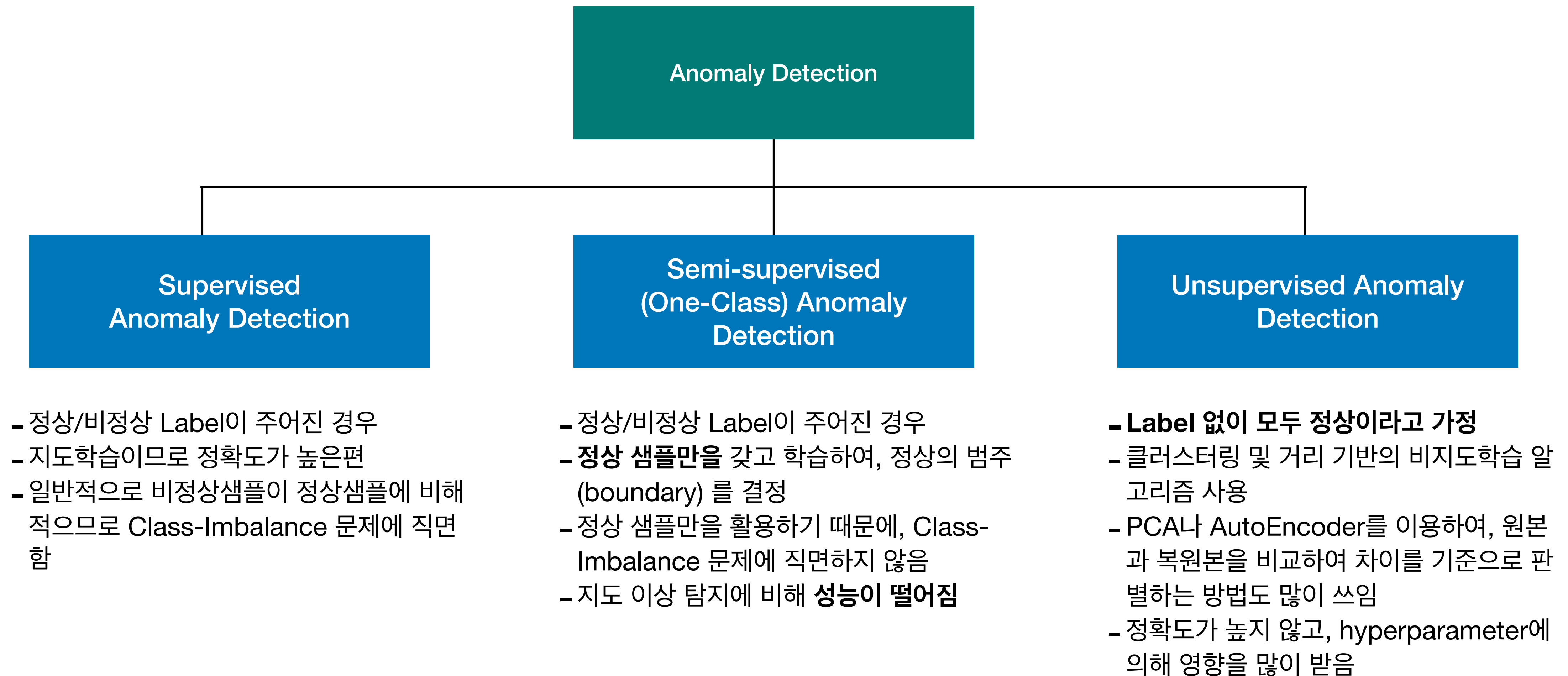
(c) Cyber-Network Intrusion detection



(d) Internet Of Things (IoT) Big-Data Anomaly detection

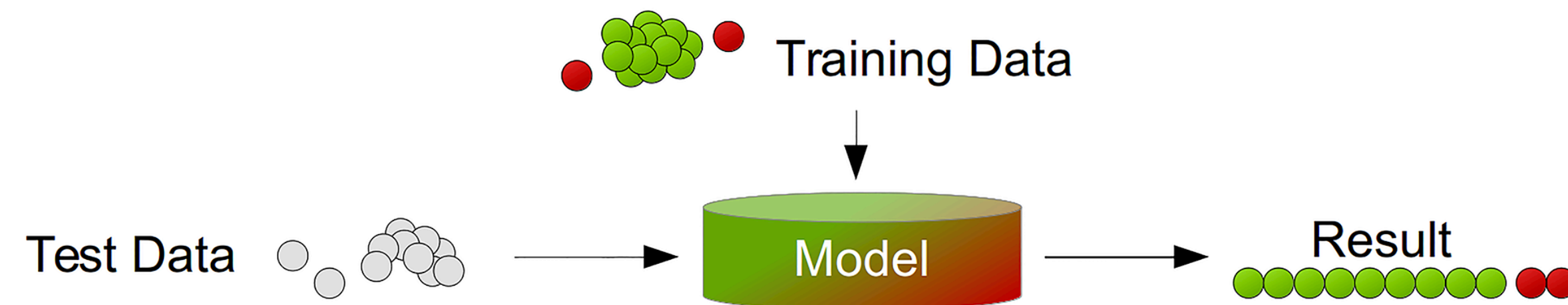
Anomaly Detection

Anomaly Detection의 접근법

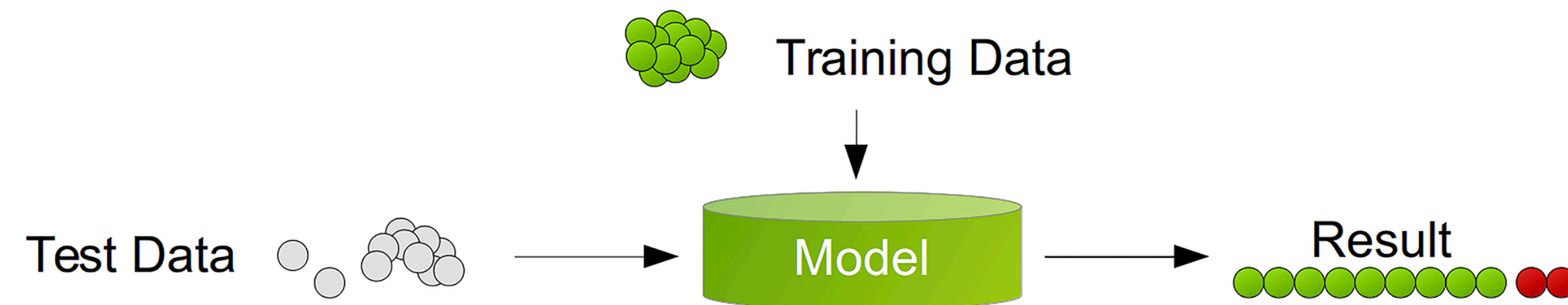


Anomaly Detection

Anomaly Detection의 접근법



(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection



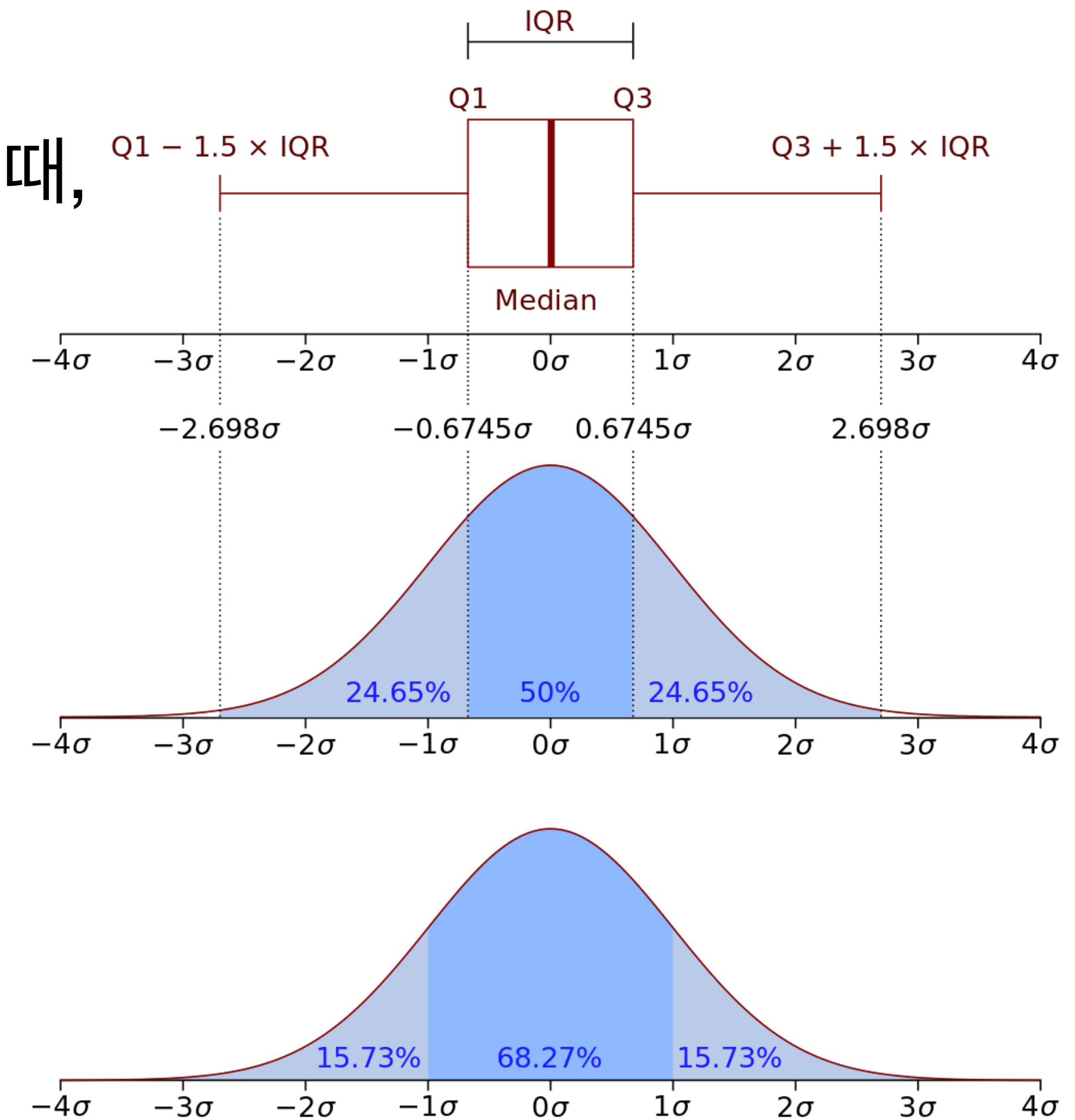
(c) Unsupervised anomaly detection

Conventional Techniques

Box Plot

박스플롯을 활용한 통계적인 접근

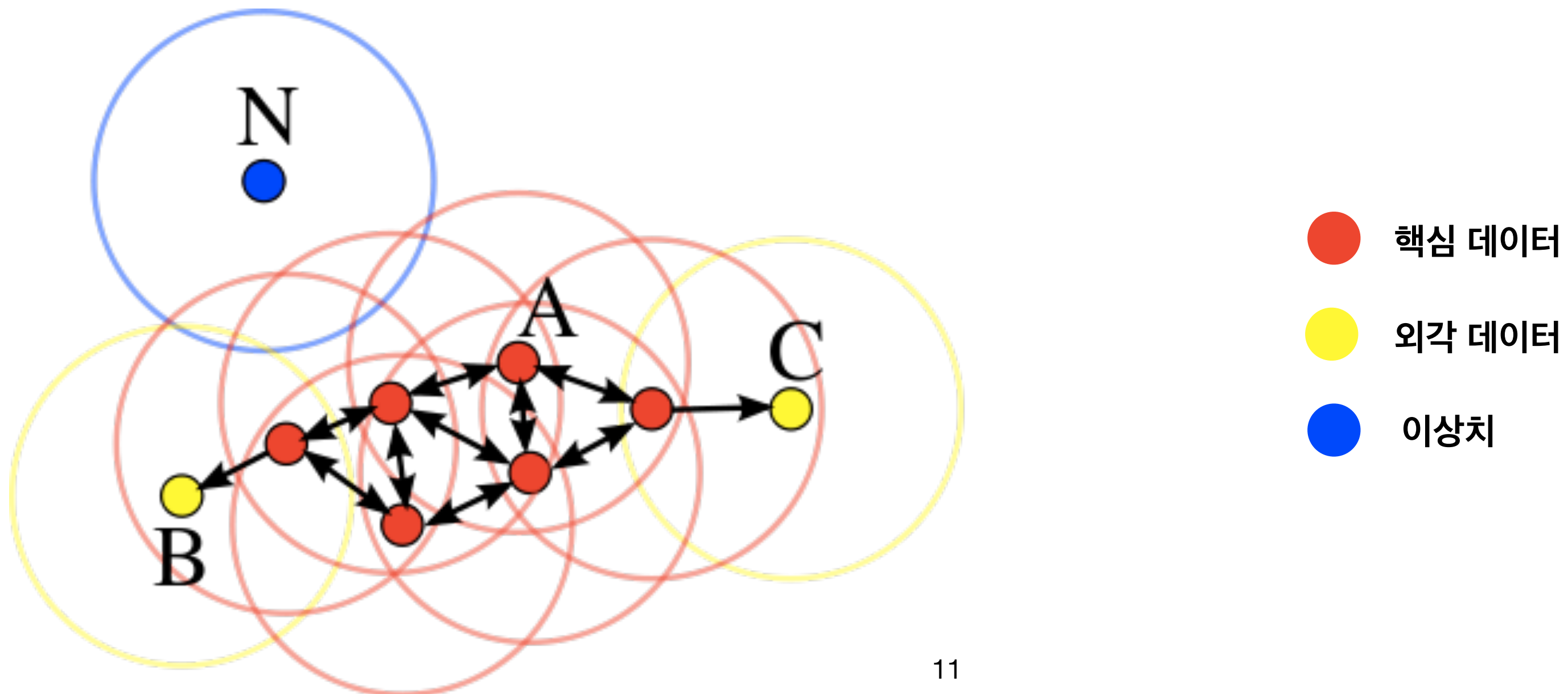
- 만약 어떤 데이터가 특정 확률 분포를 따른다고 가정 했을때, 대부분의 데이터들이 일정 영역 안에 들어온다면, 그 영역 밖에 존재하는 데이터는 이상치라고 판단한다.
- Box plot에서의 이상치 범위는 :
 - $Q1 - 1.5 \times IQR$ 과 $Q3 + 1.5 \times IQR$
- 정규분포에서의 이상치 범위(3시그마 규칙):
 - 평균에서 양쪽 끝 3시그마(표준편차)의 범위에 99.7%의 값이 들어간다



DBSCAN

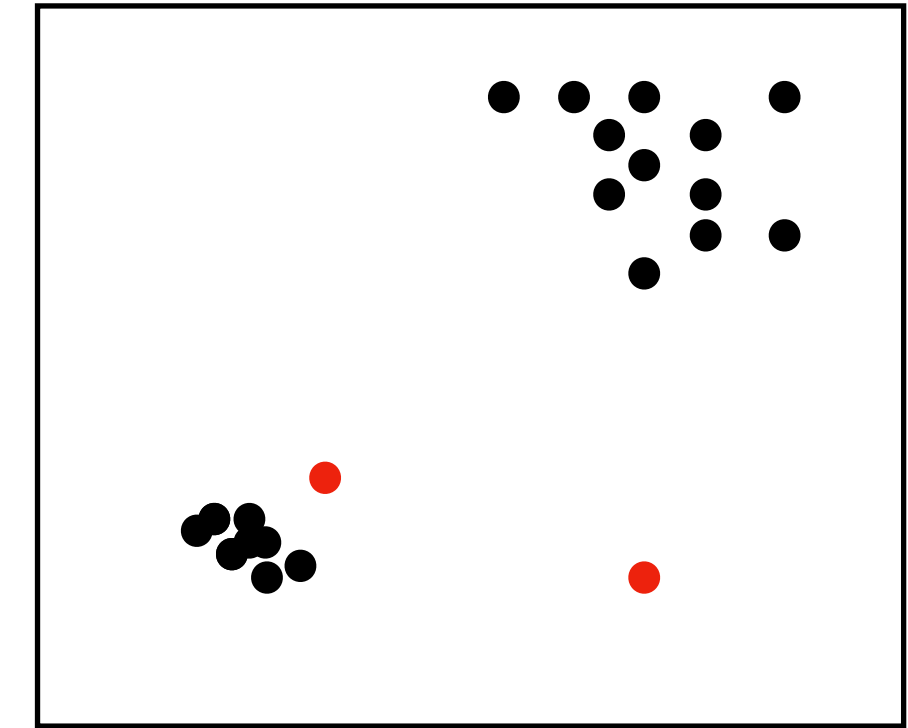
밀도기반 클러스터링 (Unsupervised)

- 데이터의 분포와 밀도를 고려하여 클러스터를 구성
- 클러스터링에 사용되었지만, 과정 중에 이상치를 탐색 가능
- 이상 데이터의 기준: 핵심 및 외각 데이터에 속하지 못한 데이터



Local Outlier Factor

지역의 밀도를 같이 고려하자



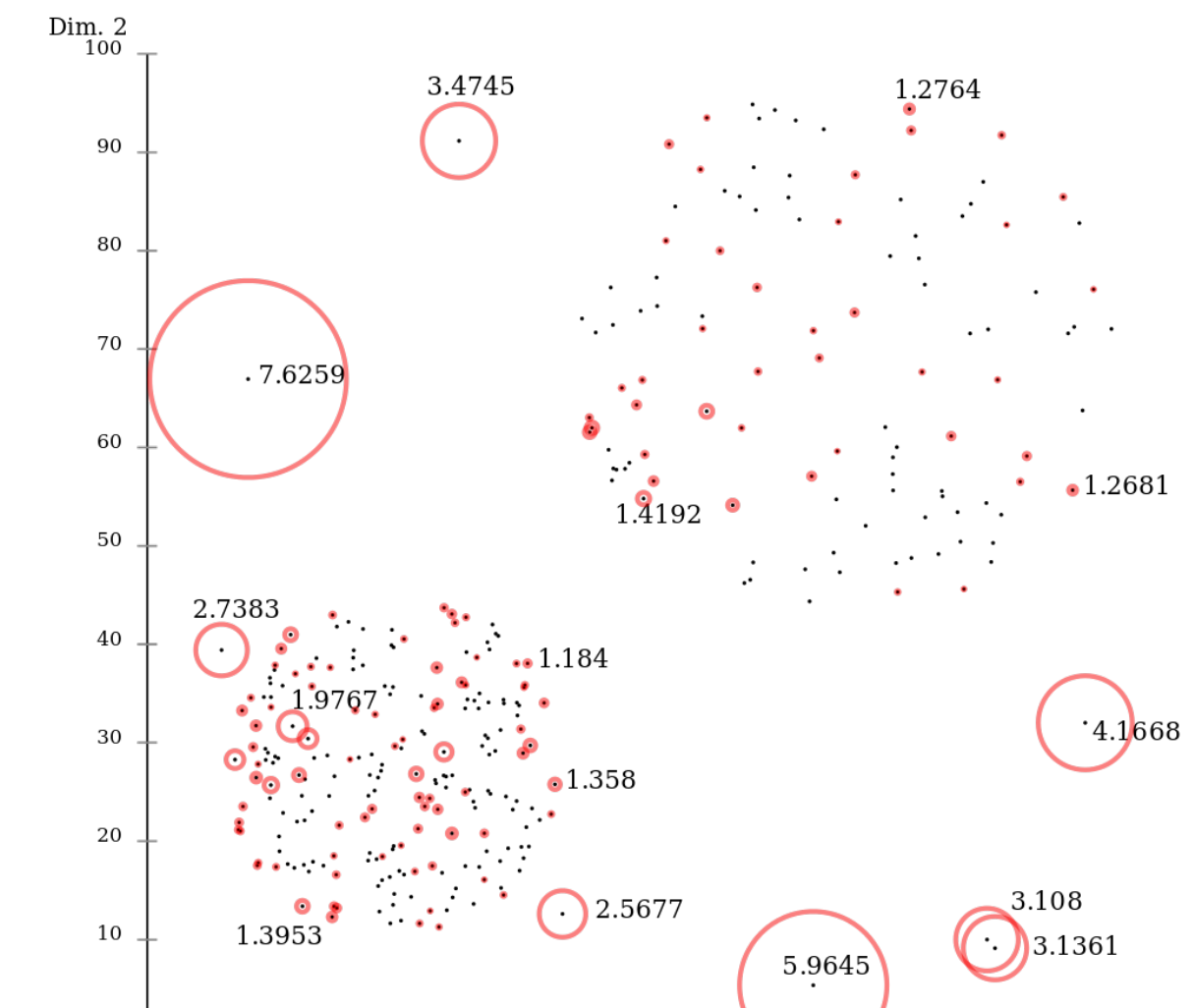
- 기존 밀도기반 알고리즘들의 한계
 - 밀도가 상이한 클러스터가 존재하는 경우 파라미터를 결정하기가 어렵다 (우상단 그림)
 - 반경, 반경 내 속해야하는 데이터 포인트의 수
- 구성요소
 - **k_distance(p)** : k 번째로 가까운 데이터와의 거리
 - **reachability distance(p,o)** : 주변 데이터 o를 고려한 거리
 - $\text{reachability-distance}_k(A,B) = \max\{k\text{-distance}(B), d(A,B)\}$
 - **local reachability density(p)** : p 주변의 k-neighbor들과의 reach dist의 역수

$$\text{lrd}_k(A) := 1 / \left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right)$$

- **Local Outlier Factor(p)** :

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}_k(B)}{\text{lrd}_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}_k(B)}{|N_k(A)| \cdot \text{lrd}_k(A)}$$

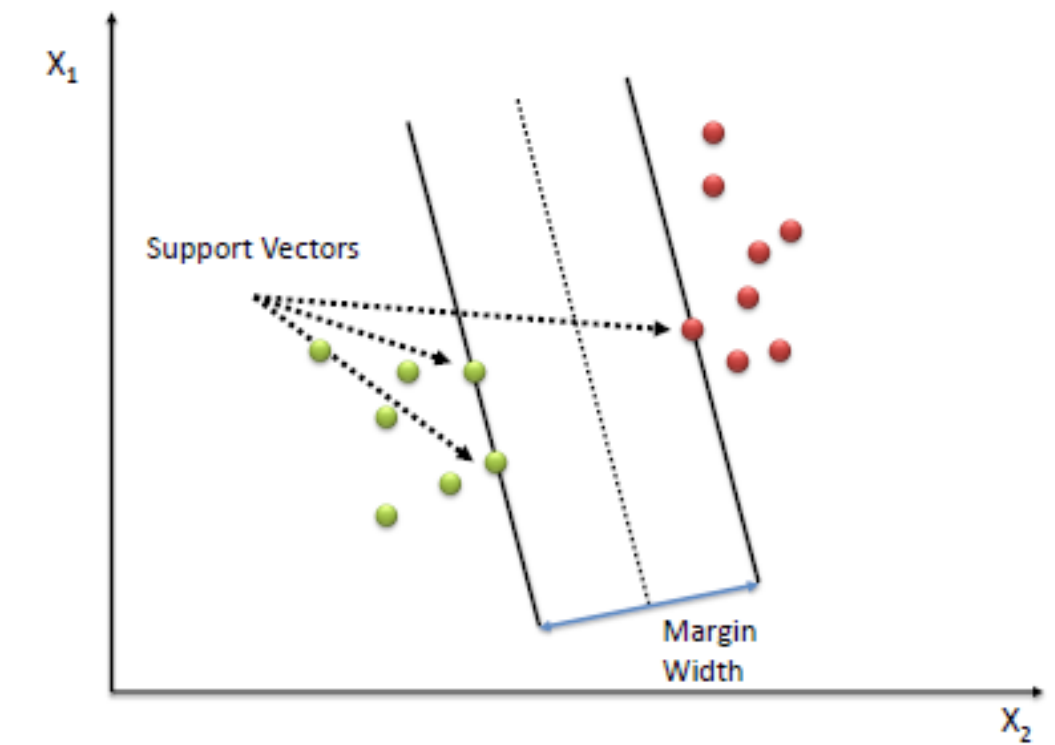
- $\text{LOF}(k) \sim 1$ means Similar density as neighbors,
- $\text{LOF}(k) < 1$ means Higher density than neighbors (Inlier),
- $\text{LOF}(k) > 1$ means Lower density than neighbors (Outlier)



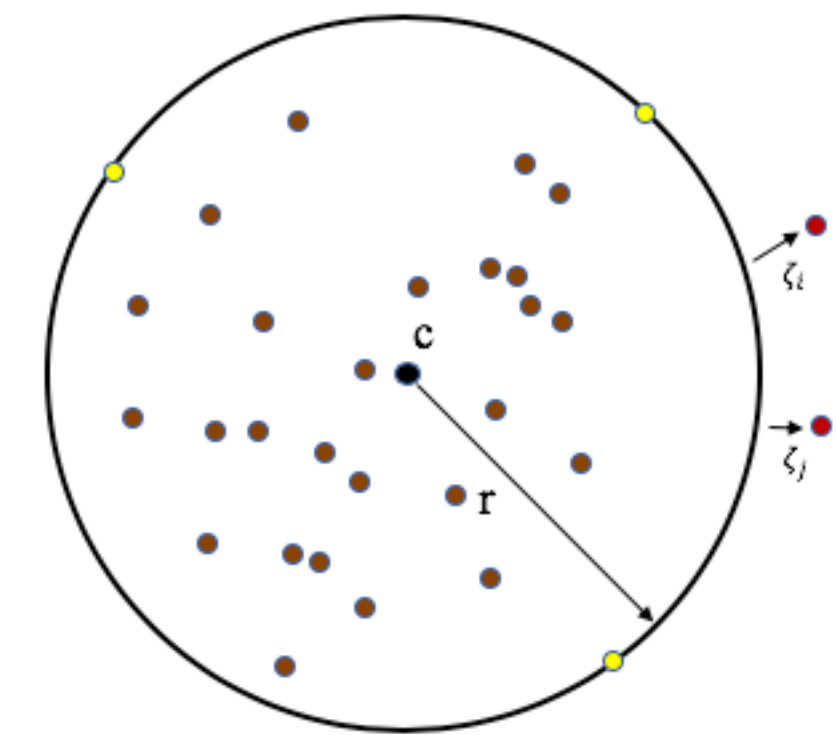
One-Class SVM

정상의 Boundary를 구할 수 있다면?

- Support Vector Machine
 - Margin이 최대가 되는 경계면(hyperplane)을 찾는 방식으로 분류 문제에 활용
 - DNN 이전에 가장 많이 쓰이던 모델
- One-class SVM
 - Unsupervised Learning 또는 Semi-supervised Learning
 - 정상 데이터로만 훈련을 진행하므로써, 정상의 **영역**을 계산
 - Finding the smallest hypersphere
 - 정상 Boundary의 밖에 위치하는 데이터 포인트들은 이상으로 간주



SVM Classifier



One-class SVM

Isolation Forest

정상에 집중하기 보다 비정상에 집중하는..

- Decision Tree

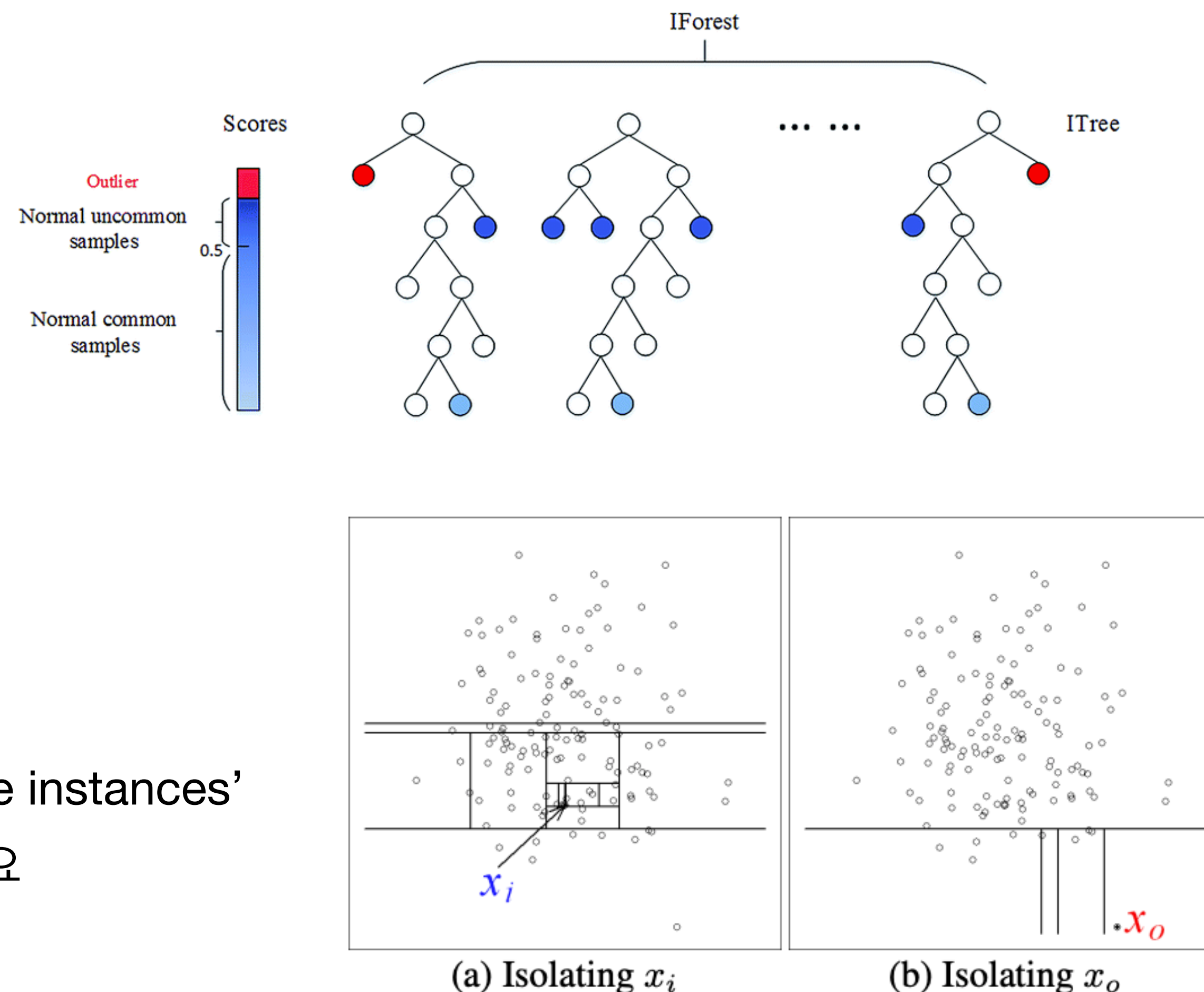
- Feature 단위로 데이터셋을 분류하는 기법
 - If $x_1 > 0$, if $x_2 == \text{"animal"}$,...

- Random Forest

- Decision Tree의 앙상블 버전
- 앙상블 기법: 서로독립적인 모델들의 집단 지성을 이용

- Isolation Forest

- Isolation : ‘separating an instance from the rest of the instances’
- 비정상적 데이터의 경우 Isolate하는데 더 적은 파티션이 필요
 - Tree기준으로 보면 Root에 더 가깝다.



Other techniques

- Distribution Based
 - Gaussian Mixture Model
 - Elliptic Envelop
- Dimensionality Reduction Based
 - Linear Dimensionality Reduction : PCA
 - Non-linear Dimensionality Reduction : Manifold Learning

Deep Learning for Anomaly Detection

Deep Learning for Anomaly Detection

딥러닝을 활용한 이상탐지

- Why Deep Learning?
 - 딥러닝을 활용하는 다른 분야와 같은 이유
 - 대량의 Feature와 Non-linearity가 존재하는 데이터셋에 대해서 이상탐지를 하기 위해
- How?
 - Auto-encoder Approach
 - One-class Neural networks (OC-NN)

Auto Encoder

인코딩한 것과 디코딩한 것의 차이가 크다면?

- Motivation

- DNN을 어떻게 비지도학습에 활용 할 수 있을까?
 - 차원축소
 - Representation

- Encoding

- Input to representation

- Decoding

- Representation to output

- 만약에 X 와 X' 이 차이가 크다면, 이상치라고 볼 수 있지 않을까?

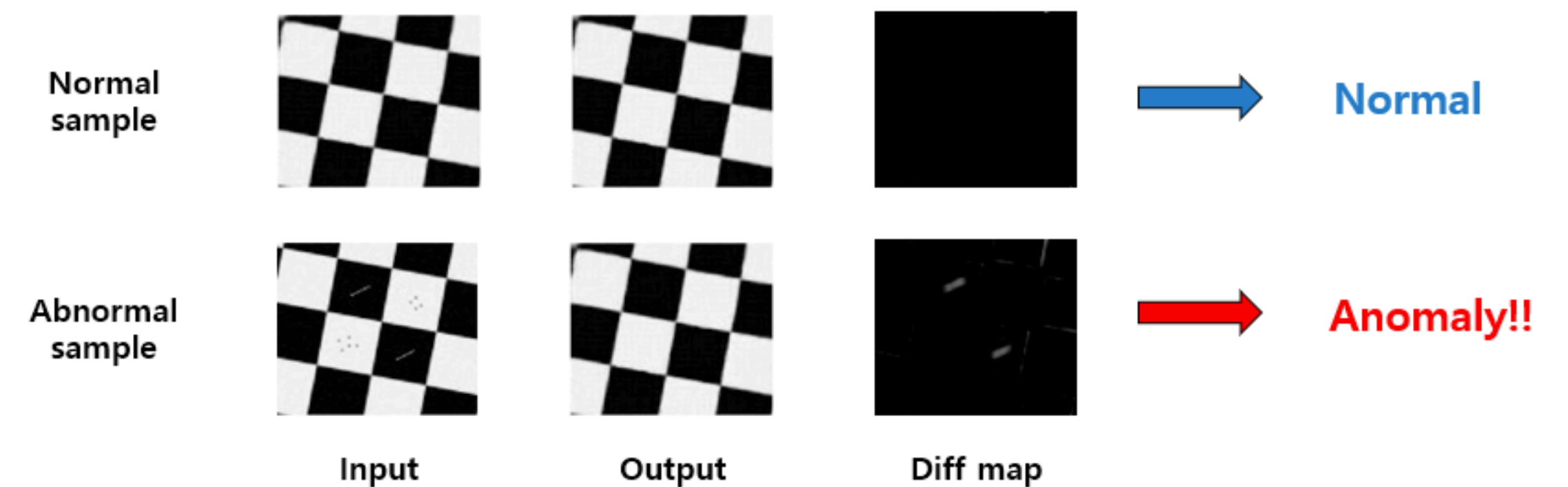
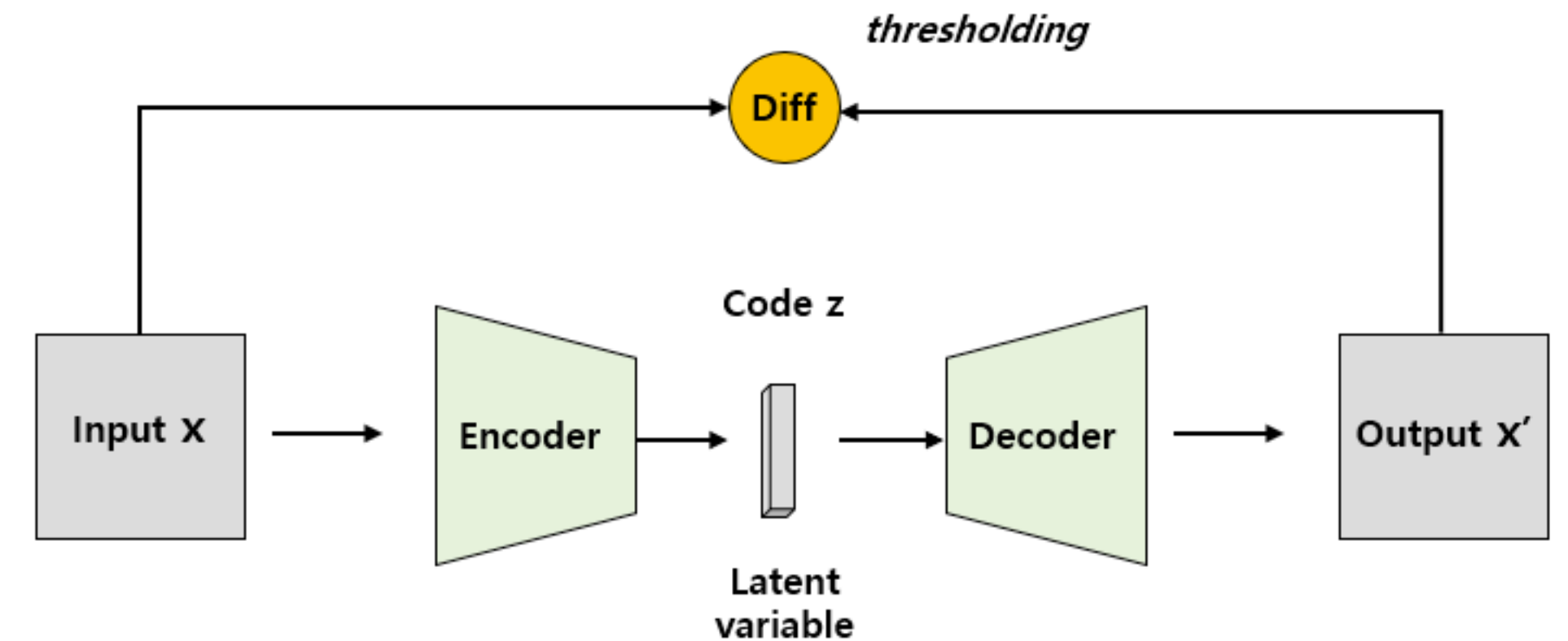


그림 출처: Improving Unsupervised Defect Segmentation by Applying Structural Similarity To Autoencoders, 2019 arXiv

One-class Neural networks (OC-NN)

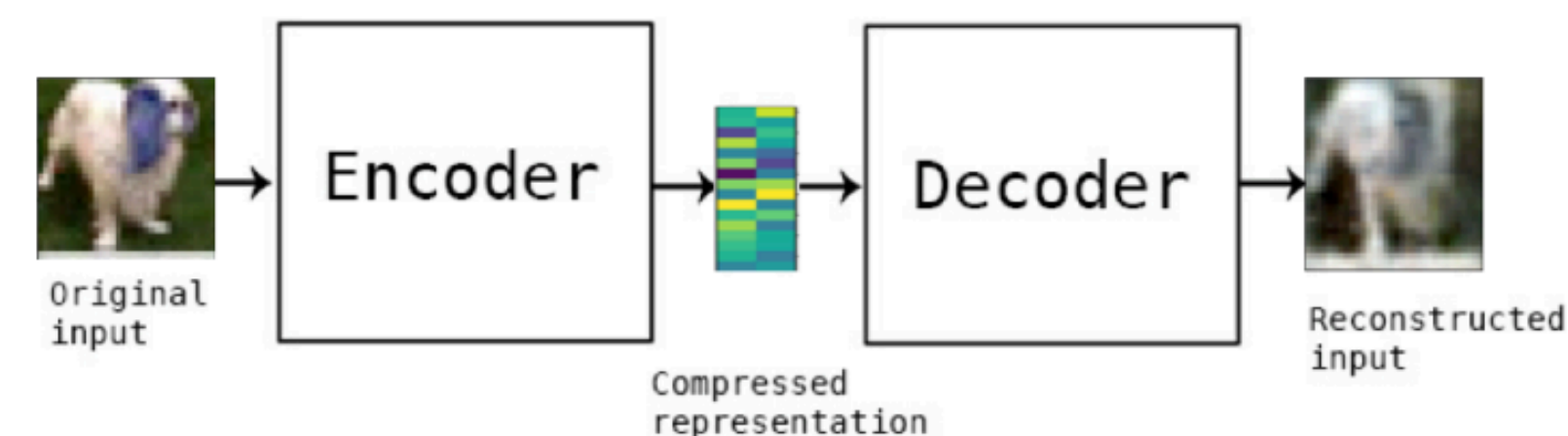
SVM 대신 Neural Networks로

- Motivation

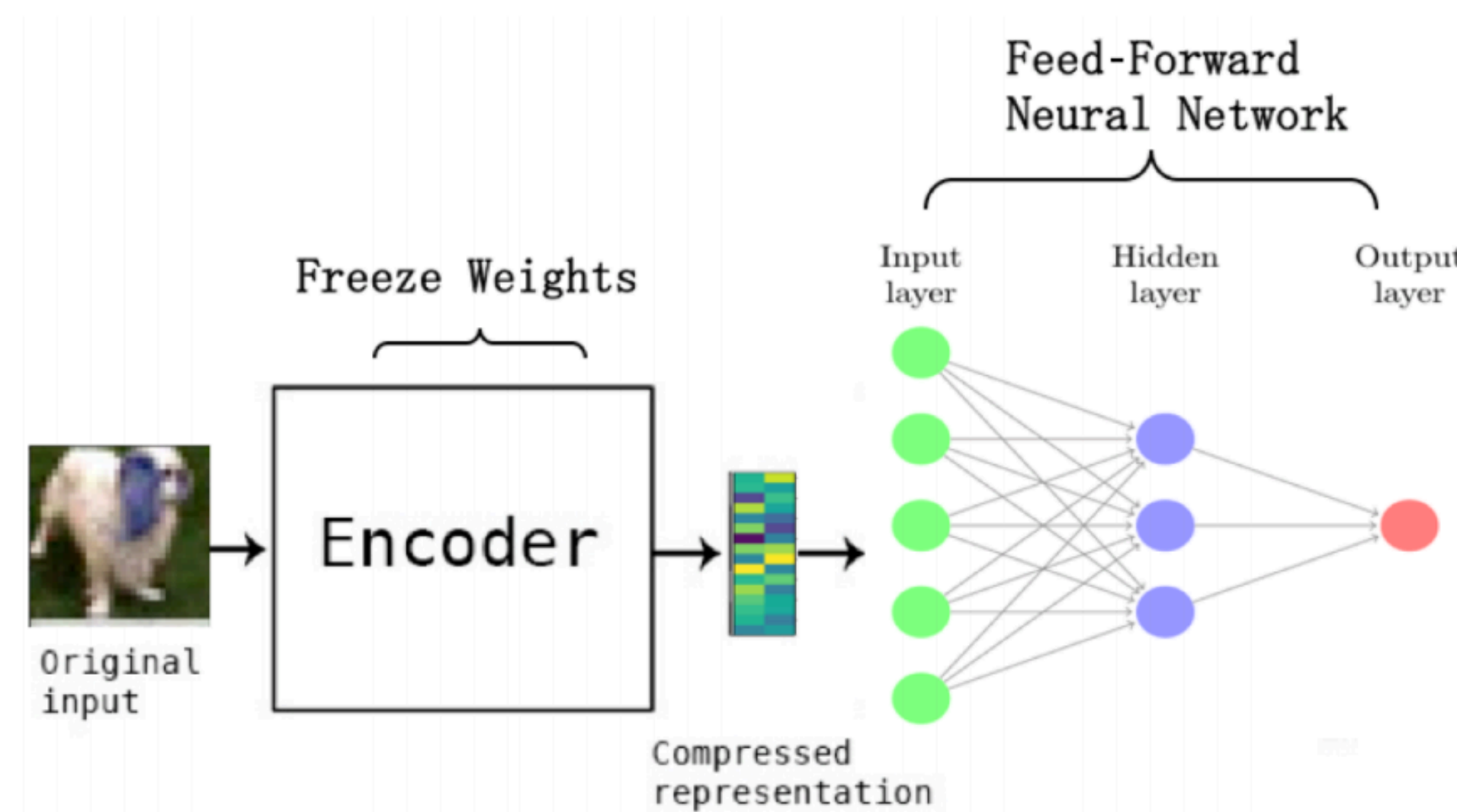
- DNN이 비선형성을 띄는 데이터셋에 잘 동작한다면, SVM 말고, NN을 쓰는게 더 좋지 않을까?

- 동작방식

- 오토인코더 이용한 인코더 확보
- 인코더를 이용해, Input의 차원축소
- 축소된 인풋을 이용해 One-Class 학습



(a) Autoencoder.



(b) One-class neural networks.

E.O.D