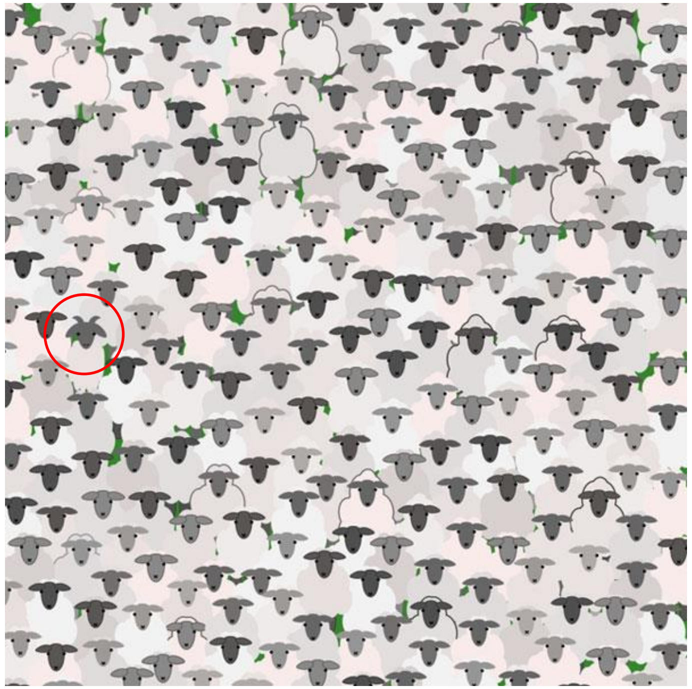


Anomaly Detection: Overview

Pilsung Kang

School of Industrial Management Engineering

Korea University

Machine Learning

- Definition
  - ✓ A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)

Supervised Learning

- Goal: predict a single "target" or "outcome" variable
- Finds relations between X and Y
- Train (learn) data where target value is known
- Score data where target value is not known

Unsupervised Learning

- Explores intrinsic characteristics
- Estimates underlying distribution
- Segment data into meaningful groups or detect patterns
- There is no target (outcome) variable to predict or classify

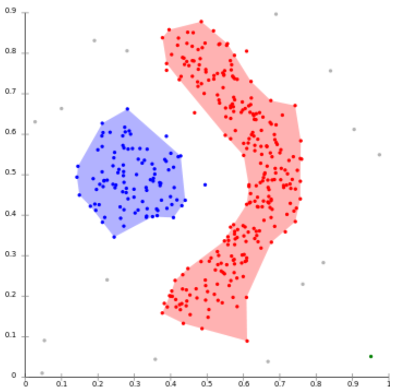
Unsupervised Learning

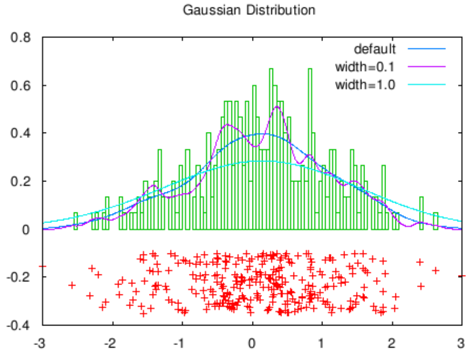
A given dataset **X**

	Var. 1	Var. 2	…	Var. d
Ins. 1	..	..	…	..
Ins. 2	..	..	…	..
…	…	…	…	…
Ins. N	..	..	..	..

Unsupervised learning

- Explores intrinsic characteristics
- Estimates underlying distribution
- Density estimation, clustering, association rule mining, network (graph) analysis, etc.

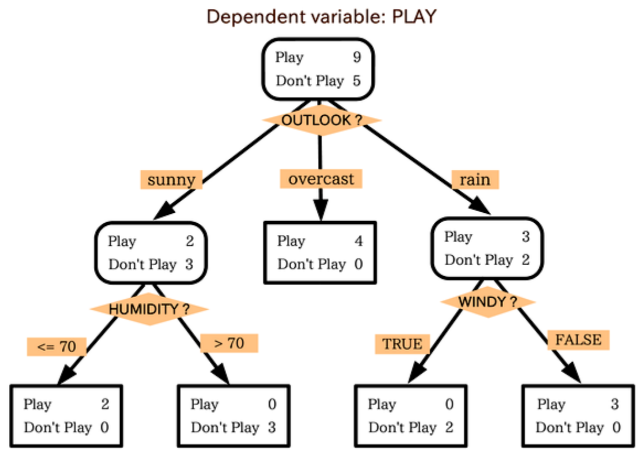




Supervised Learning

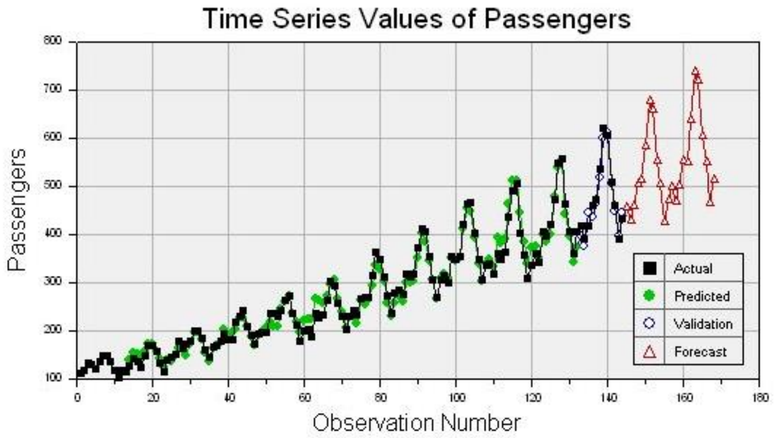
A given dataset **X** & **Y**

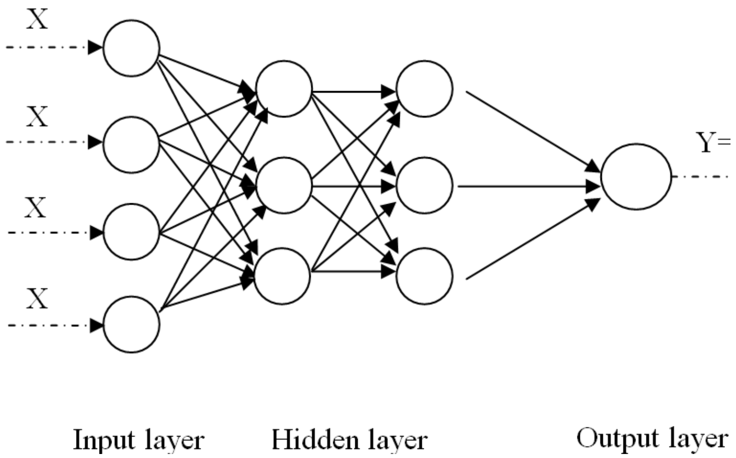
	Var. 1	Var. 2	…	Var. d		Y
Ins. 1	..	..	…	..		..
Ins. 2	..	..	…	..	$y = f(x)$	..
…	…	…	…	…		…
Ins. N	..	..	..	..		..



Supervised learning

- Finds relations between X and Y: estimate the underlying function $y = f(x)$
- Classification, regression, novelty detection





고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Anomaly Detection

- What is abnormal/novel data (outliers)?

> *"Observations that deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism (Hawkins, 1980)"*
>
> *"Instances that their true probability density is very low (Harmeling et al., 2006)"*

- Outliers are different from noise data
  - ✓ Noise is random error or variance in a measured variable
  - ✓ Noise should be removed before outlier detection
- Outliers are interesting
  - ✓ It violates the mechanism that generates the normal data

Anomaly Detection

- Applications: Industrial Monitoring

MakinaRocks   Products   About   Careers   Resources   Blog   Contact

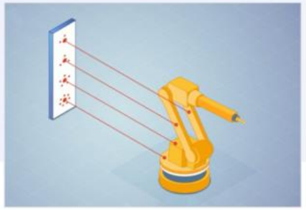
Main / Products / Anomaly Detection Suite

Anomaly Detection Suite

Detect Early, Act Smart

During the manufacturing process, a variety of anomalies such as malfunctions of equipment or defects in parts of an assembled product may occur. It is imperative to detect such anomalies early on, and act quickly in order to minimize the unexpected downtime in manufacturing or power generation, or to assure product quality. Our ML-based anomaly detection (suite) has the capacity to identify various anomalies and aid in scoping out the root causes.

How it works



Build

a ML-based novelty detection model specific to an application (e.g., robot arm)

Operationalize

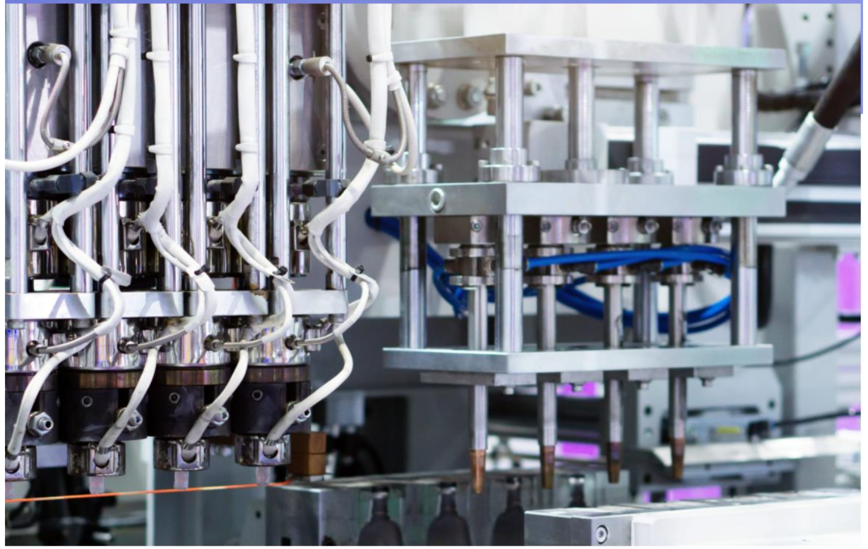
the model that evolves with data

Scale

to expand the number of assets and systems via cloud

ANOMALY DETECTION SUITE 01

Predictive Maintenance for Semiconductor Processing Equipment



Goal

To detect abnormal patterns and predict remaining time to failure of semiconductor processing equipment ahead of time to minimize downtime losses and excessive maintenance costs

Our Approach

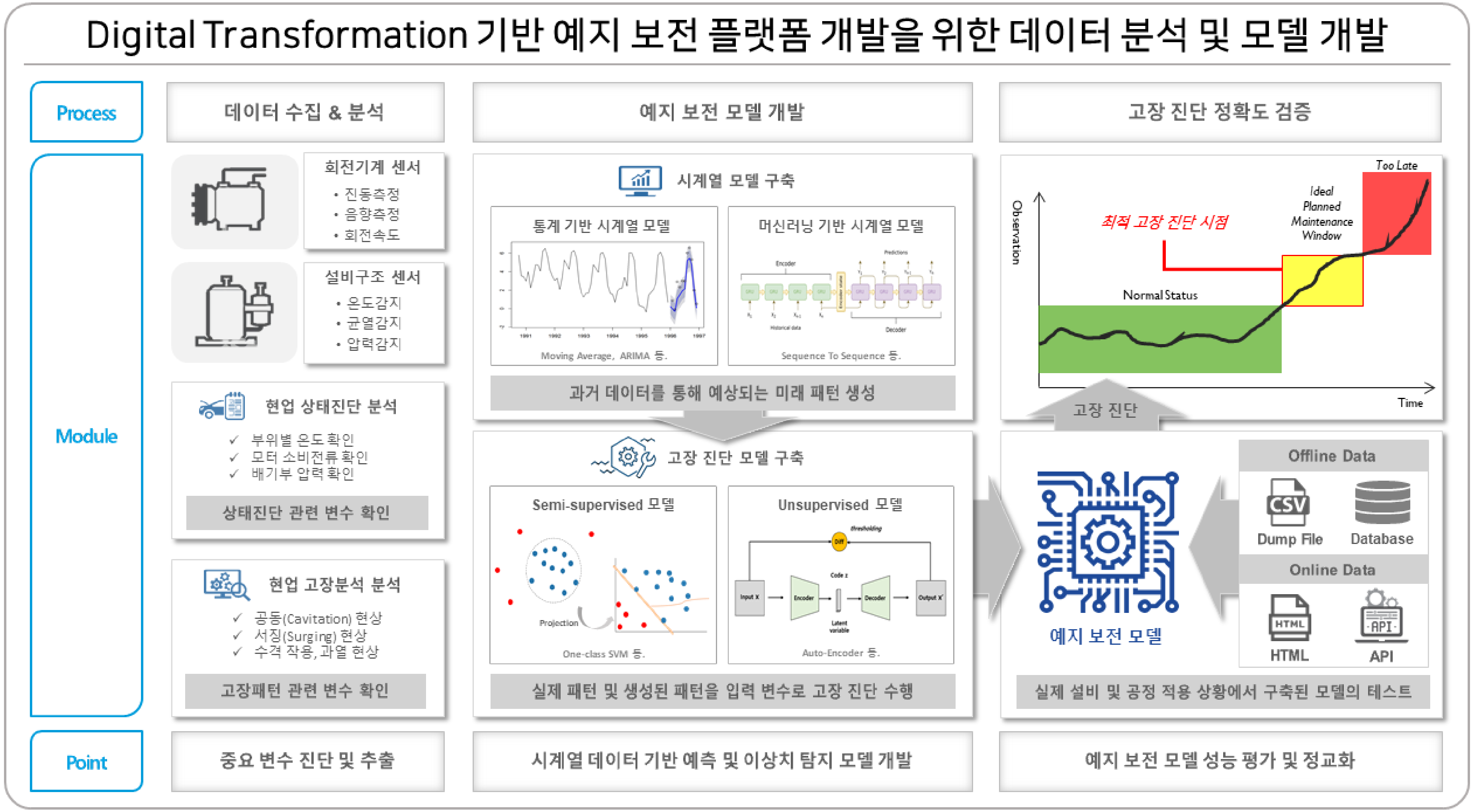
We formulated the problem as semi-supervised novelty detection to overcome lack of failure samples in production. On top of novelty detection results, we developed a method to estimate time-to-failures (TTF) of semiconductor processing equipment. To adapt to production environment change, a continual learning scheme was developed as well, and is now ready to apply.

Results

Improved Time-to-Failure prediction with 90% + accuracy and less than 1% false alarm rate

Anomaly Detection

- Applications: Industrial Monitoring

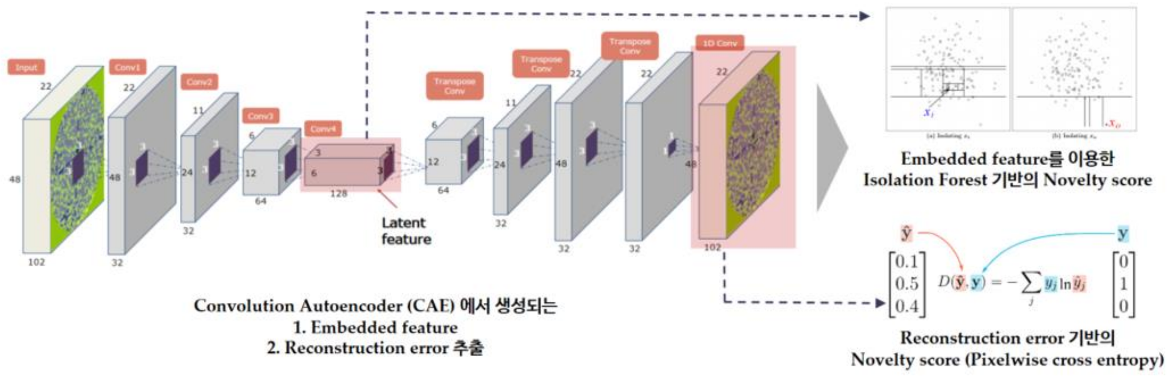


고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Anomaly Detection

- Applications: Industrial Monitoring



오늘 생성된 WBM중에서

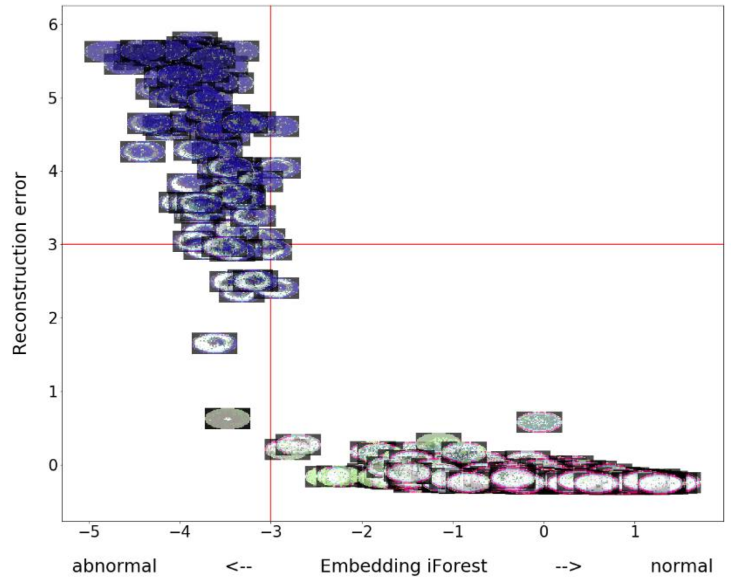
이상치는 어떤 것일까?

오늘 생성된 WBM중에서

과거 WBM 패턴으로 보았을 때,
이상치는 어떤 것일까?

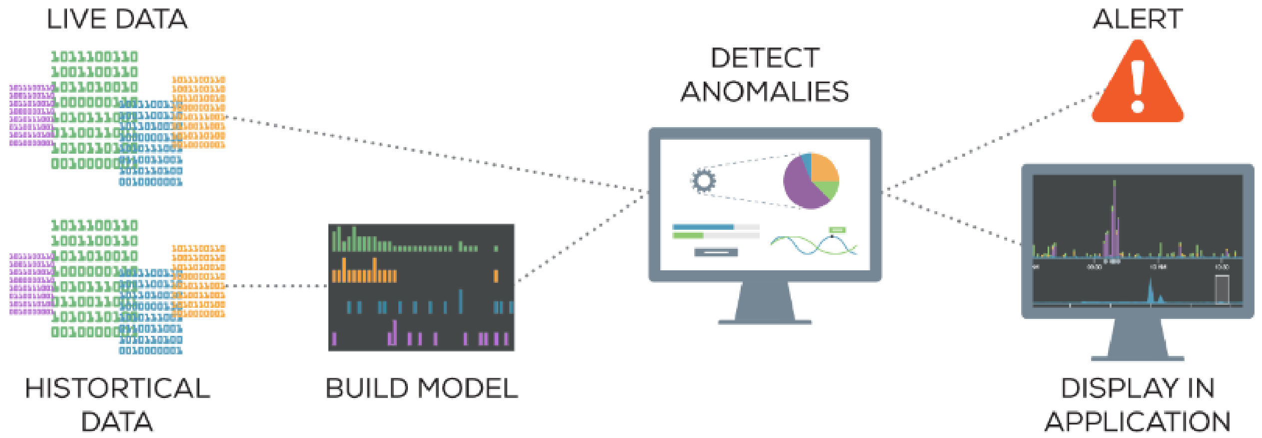
오늘 생성된 WBM중에서

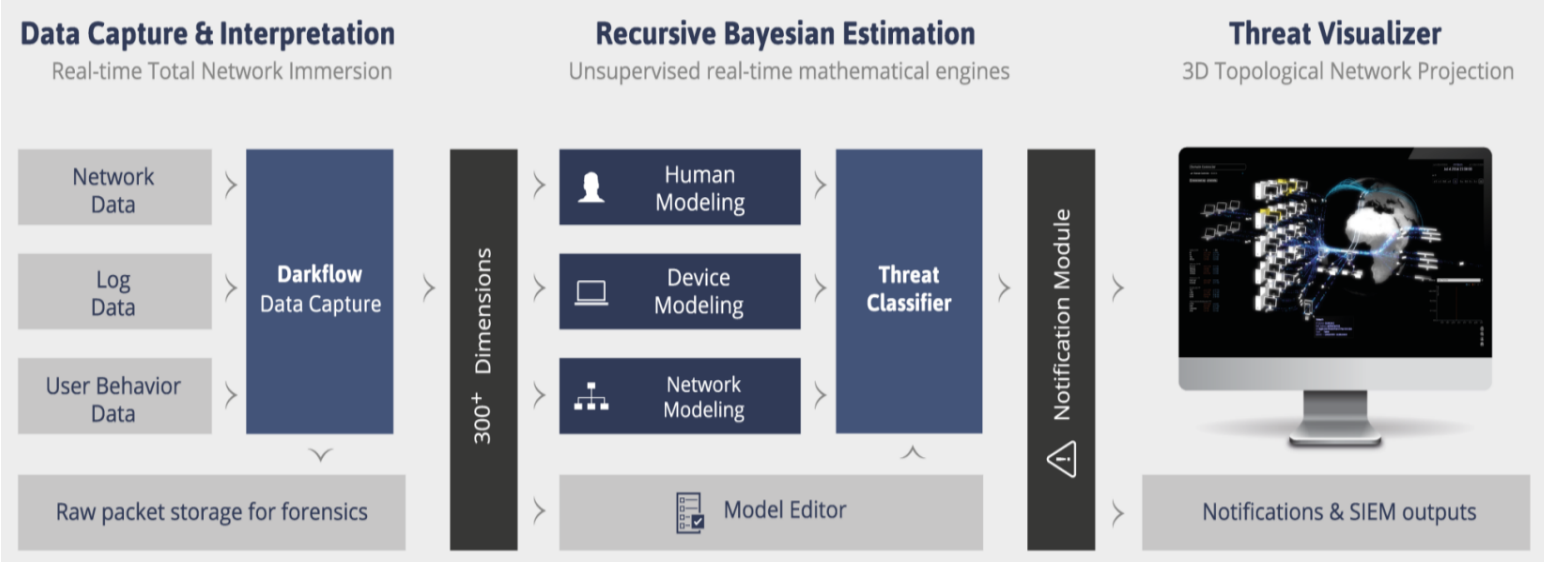
과거 WBM 패턴으로 보았을 때,
WBM중 어느 칩에서 이상치가
크게 발생 하였을까?



Anomaly Detection

- Applications: System Security





Anomaly Detection

Data Preparation

Vectorization

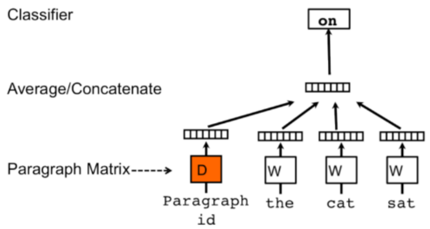
Host-based

ADFA-LD Dataset

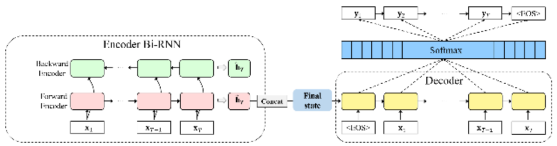
Syscall Trace

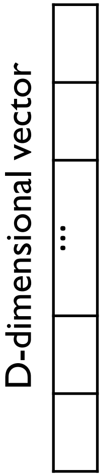
```
265 104 265 104 3 175
104 142 3 3 3 104 146
265 104 142 142 175
146 142 146 142 265 3
175 175 142 142 175
119 265 142 146 265
146 119 142 146 142
142 142 142 146 104
265 3 119 3 265 119
146 146 146 265 146
142 142 146 142 119
```

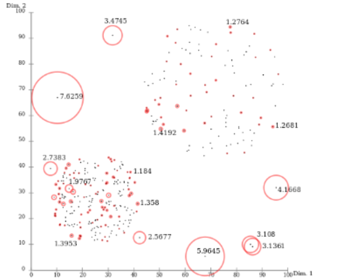
Doc2vec

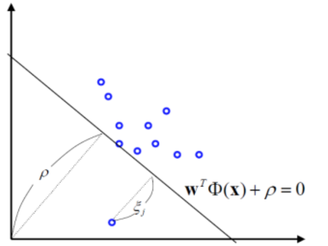


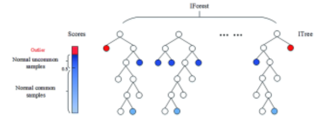
RNN-AutoEncoder

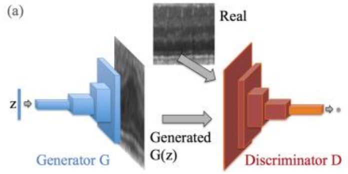








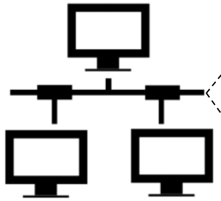


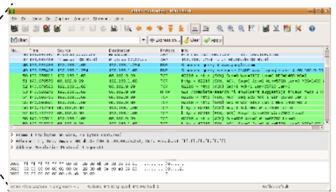


Network-based

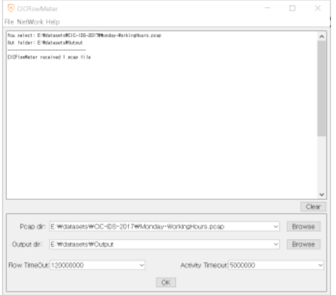
CICIDS2017 Dataset

Packet Capture

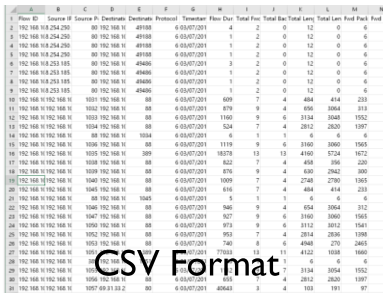




CIC-FlowMeter



Features



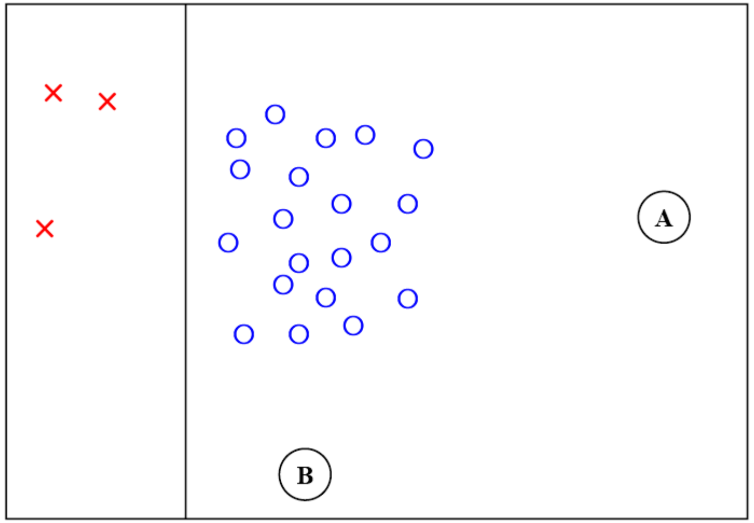
CSV Format

고려대학교 KOREA UNIVERSITY

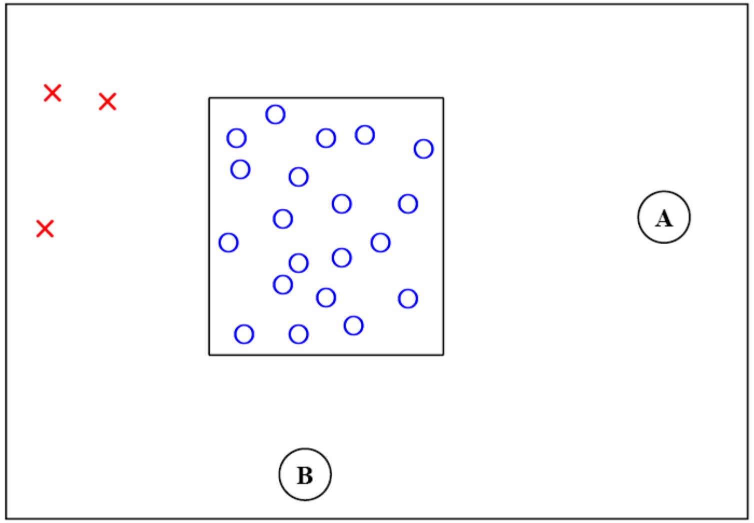
DSBA Data Science & Business Analytics

Anomaly Detection

- Classification vs. Anomaly Detection



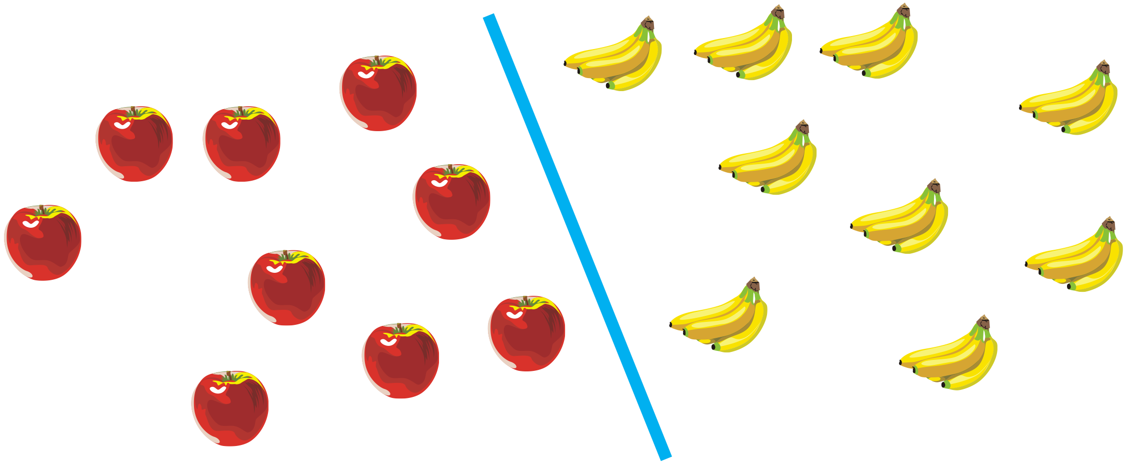
Binary classification



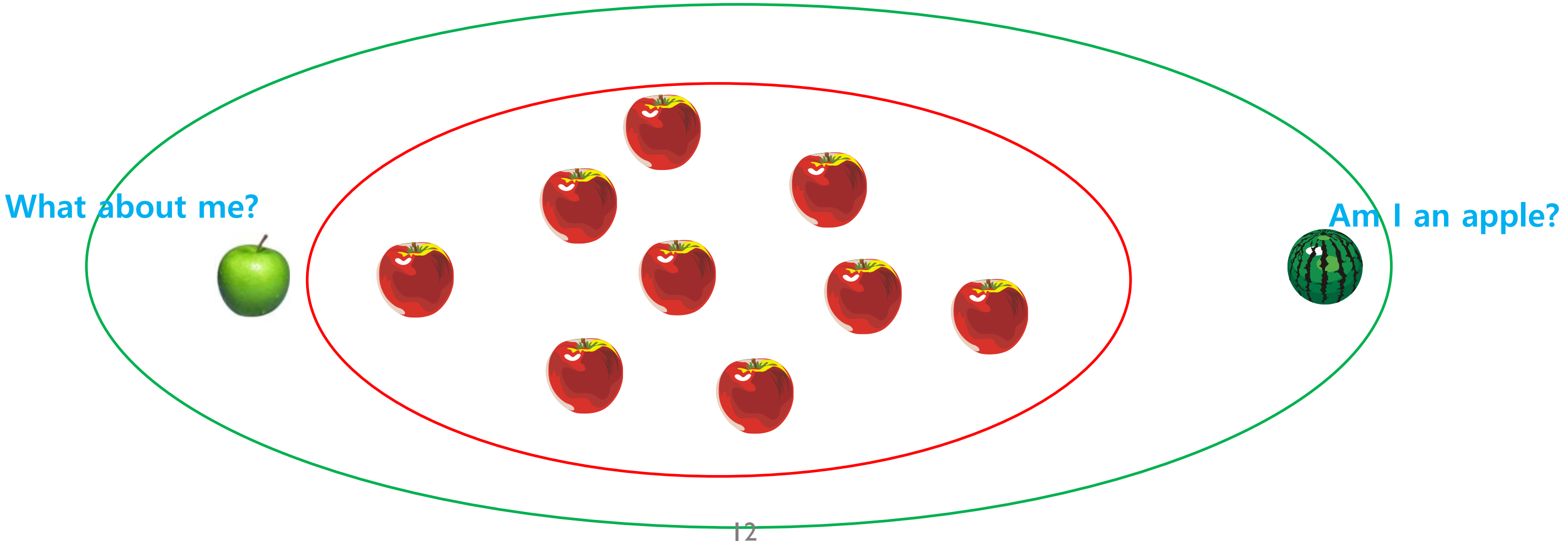
Anomaly detection

Anomaly Detection

- The way by which the classification and novelty detection learns from data
  - ✓ Classification

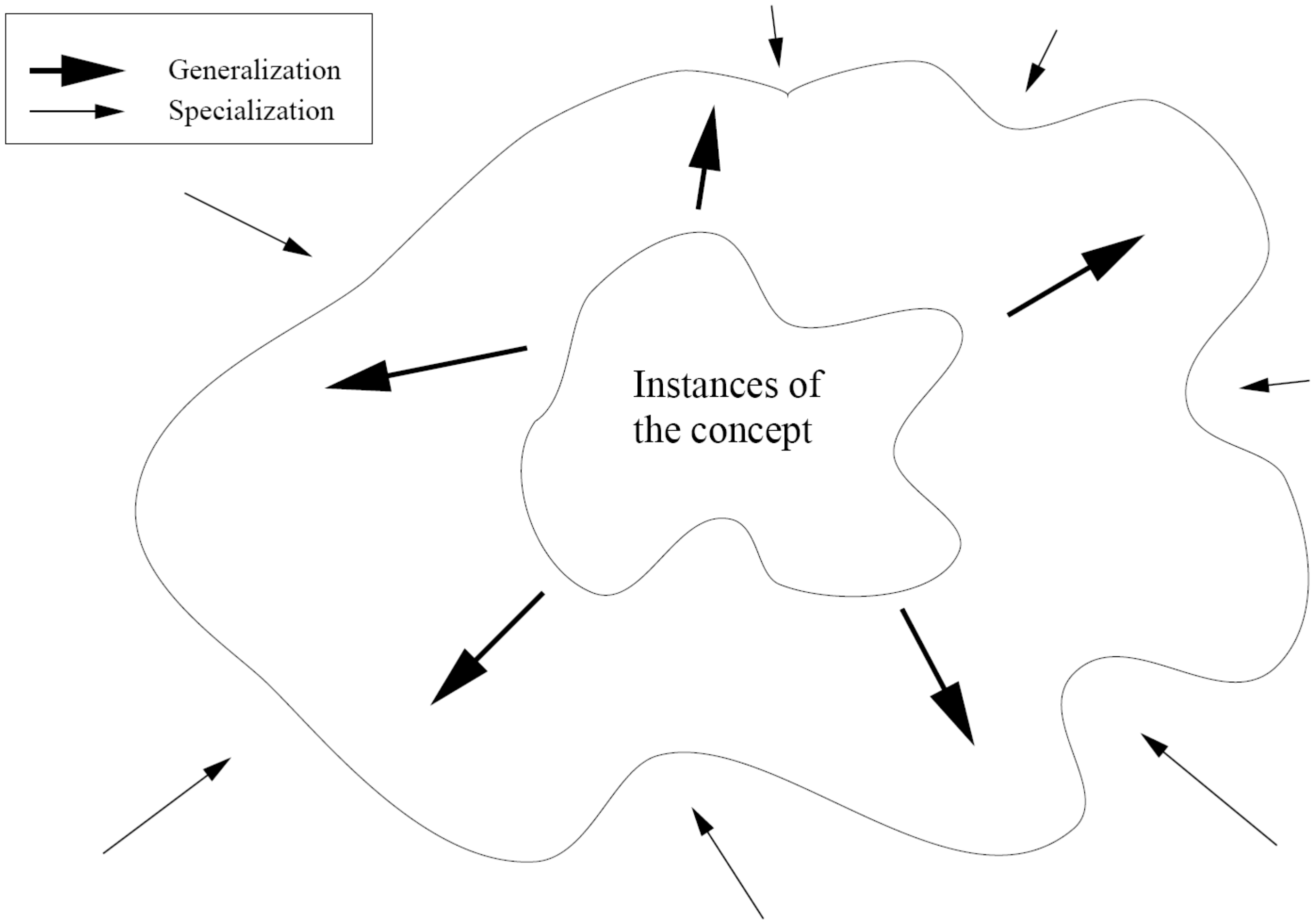


  - ✓ Anomaly detection



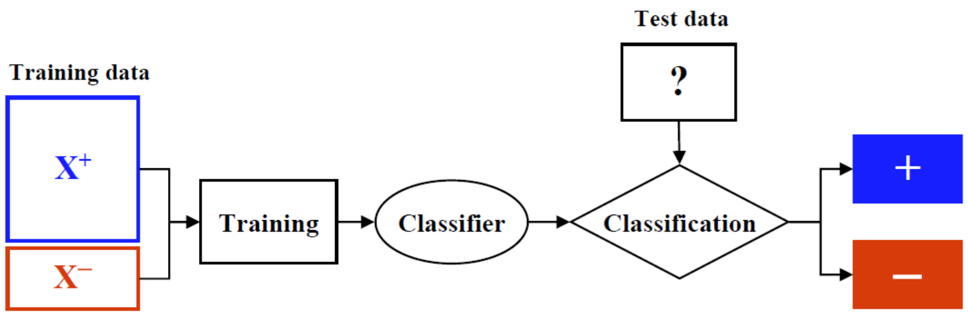
Anomaly Detection

- Generalization vs. Specialization

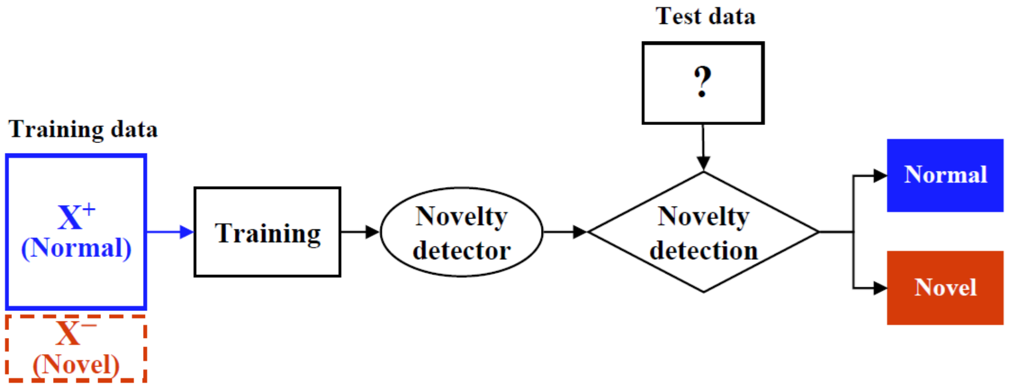


Anomaly Detection Approach

- Assumption
  - ✓ There are considerably more "normal" observations than "abnormal" observations in the data
  - ✓ Classification

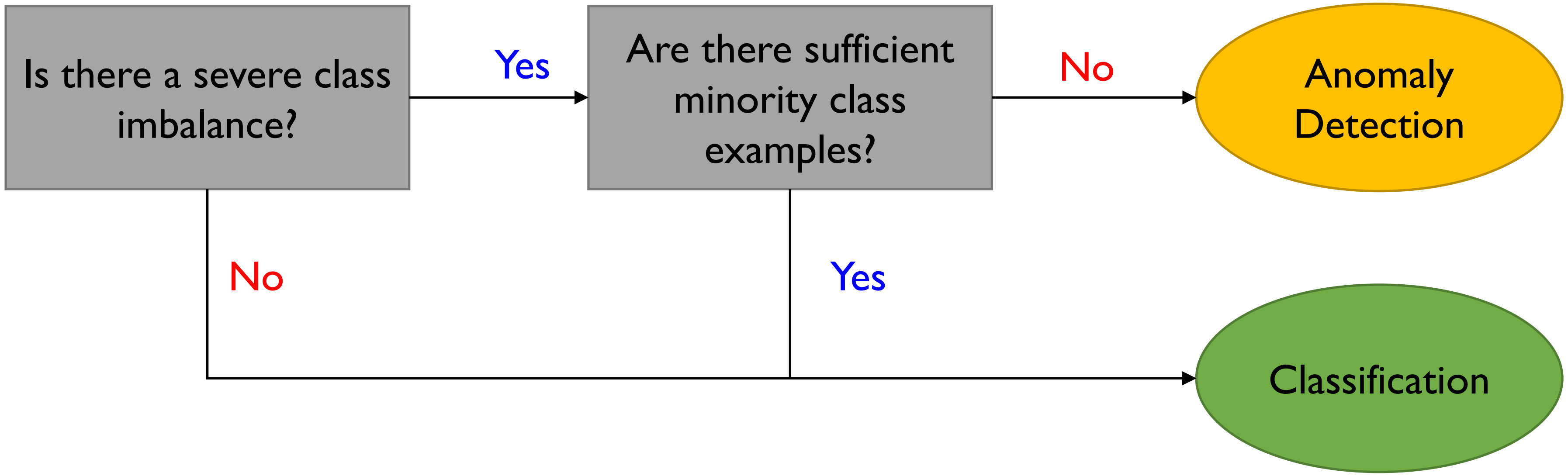


  - ✓ Anomaly detection



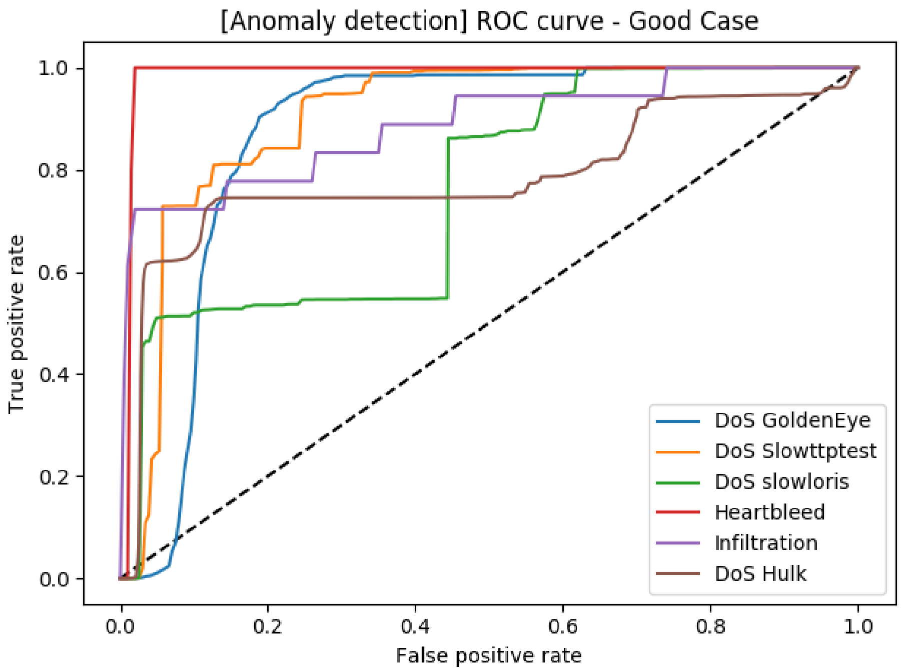
Anomaly Detection

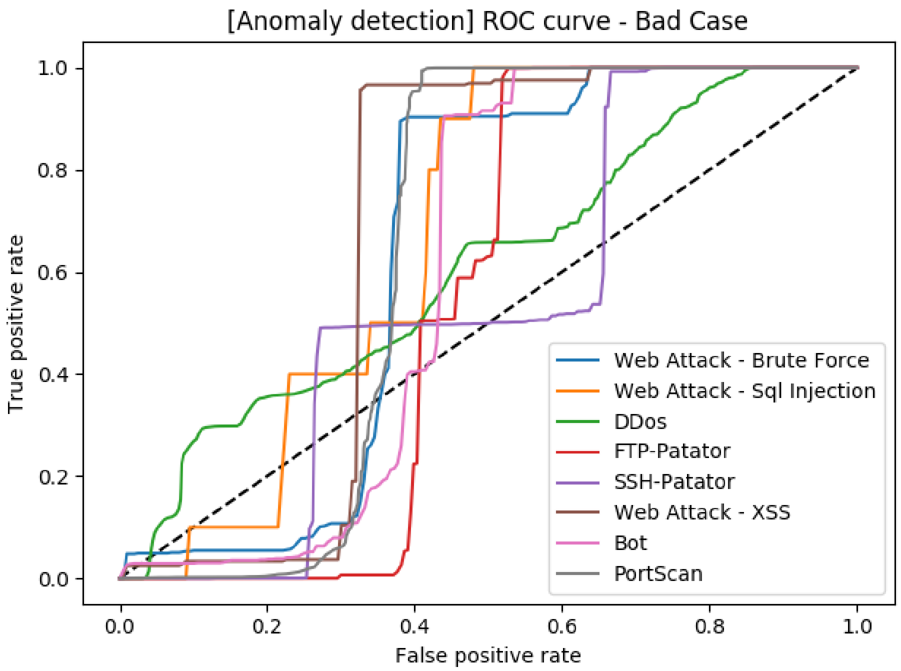
- Classification vs. Anomaly Detection
  - ✓ Which one to use?



Anomaly Detection

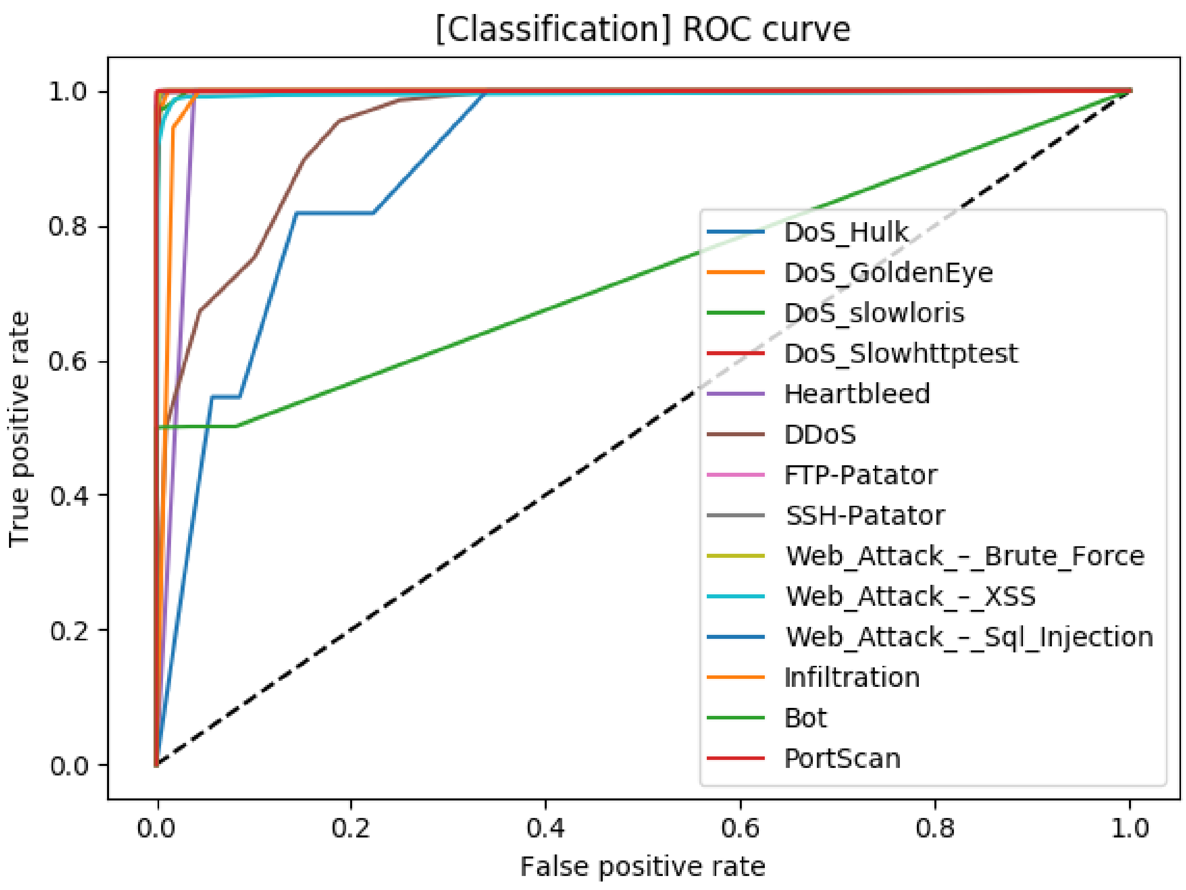
- Classification vs. Anomaly Detection
  - ✓ Performance comparison for network traffic anomaly detection





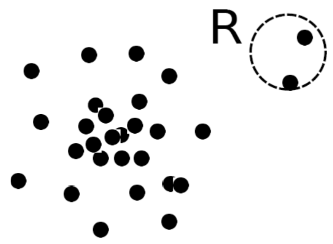
Anomaly Detection

- Classification vs. Anomaly Detection
  - ✓ Performance comparison for network traffic anomaly detection

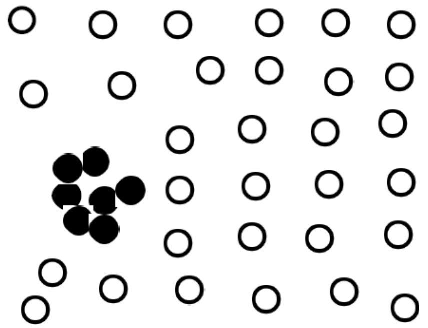


Type of Abnormal Data (Outliers)

- Global outlier
  - ✓ Object that significantly deviates from the rest of the data set
  - ✓ Ex) Credit card fraud detection
  - ✓ Issue: find an appropriate measurement of deviation



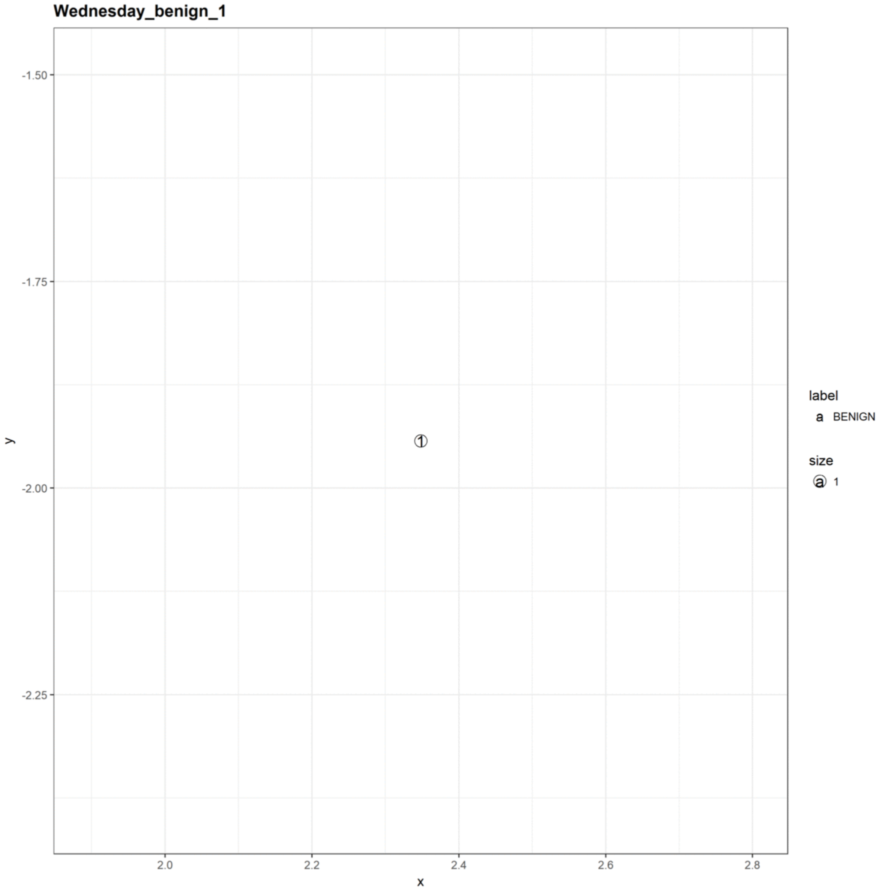
- Contextual outlier (local outlier)
  - ✓ Object that deviates significantly based on a selected context
  - ✓ Ex) 30℃ in Alaska vs. 30℃ in Sahara
  - ✓ Issue: How to define or formulate meaningful context?



- Collective outlier
  - ✓ A subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers
  - ✓ Ex) Denial-of-Service (DoS) attack

Type of Abnormal Data (Outliers)

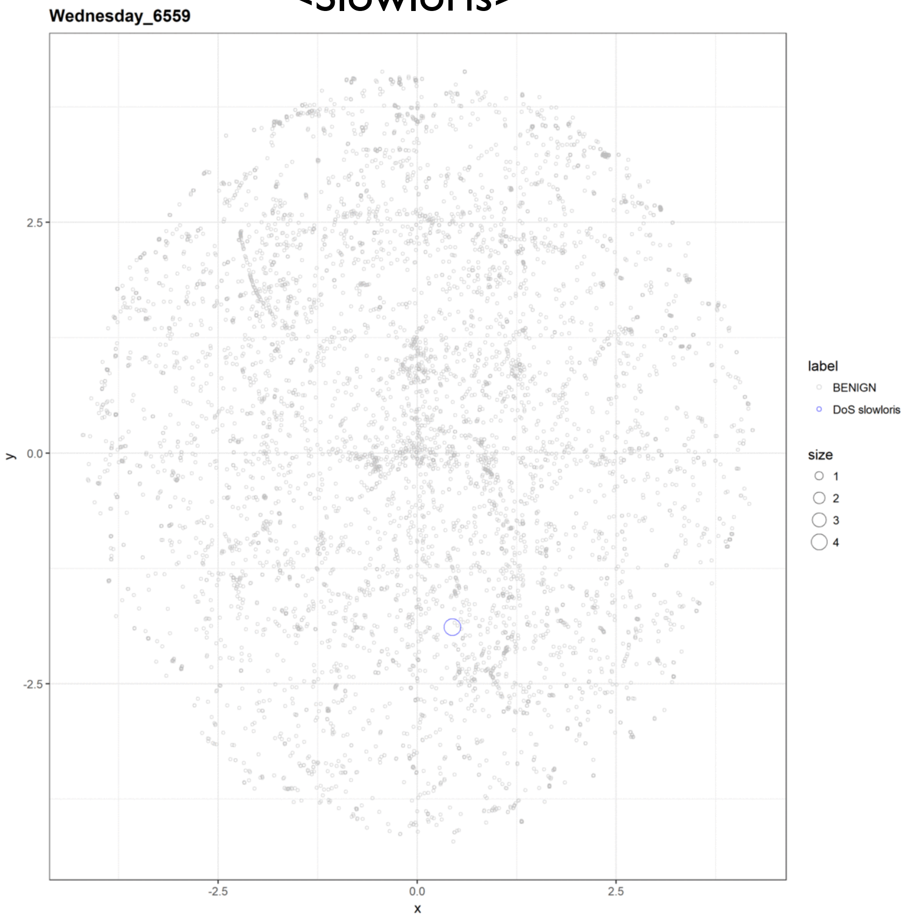
- Collective outlier: An example
  - ✓ Normal traffic (animation by Minsik Park)



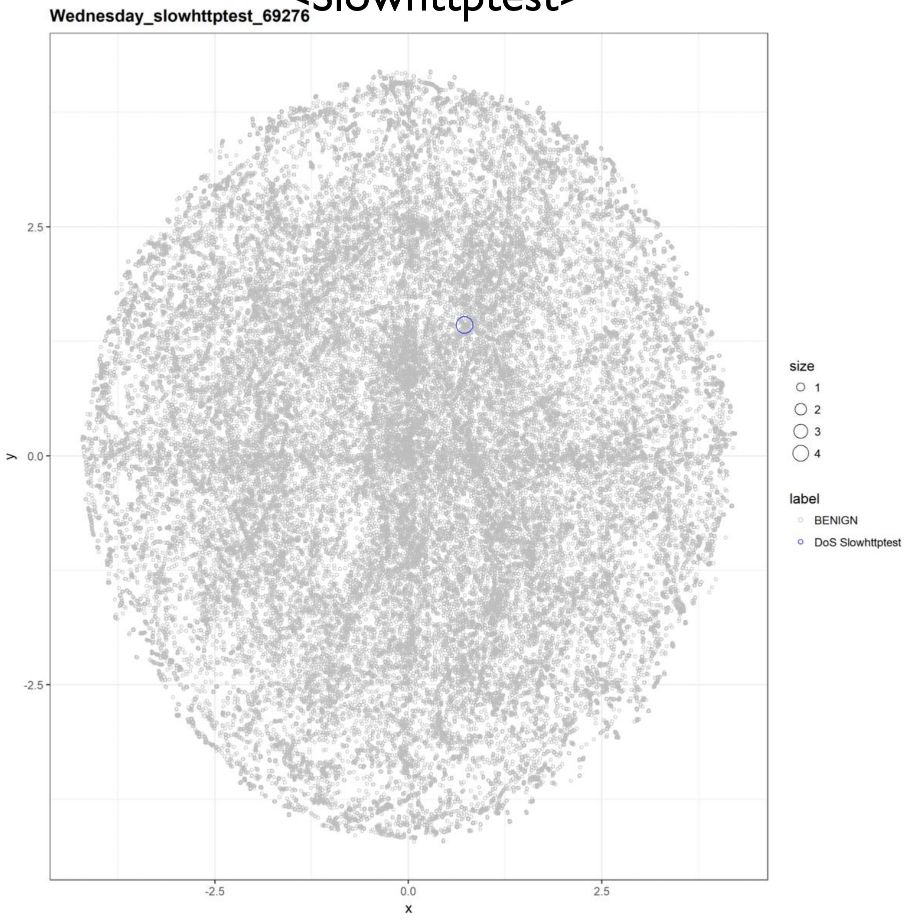
Type of Abnormal Data (Outliers)

- Collective outlier: An example
  - ✓ DoS traffic (animation by Minsik Park)

<Slowloris>



<Slowhttptest>



Challenges

- Modeling normal objects and outliers properly
  - ✓ The border between normal and outlier objects is often a gray area

- Application-specific outlier detection
  - ✓ Choice of distance measure among objects and the model of relationship among objects are often application-dependent
  - ✓ E.g., clinic data: a small deviation could be an outlier; while in marketing analysis, larger fluctuations

- Understandability
  - ✓ Understand why these are outliers: Justification of the detection
  - ✓ Specify the degree of an outlier: the unlikelihood of the object being generated by a normal mechanism

Challenges

- Novelty detection actually matters

○ 분석목표 : 제품 원료품질 검사 데이터(X)로 제품불량(Y)을 예측

○ 데이터 분포

- X 변수 : 재료 품질검사 데이터 560 개 항목
- Y 범주 : 불량, 정상
- 비중 : 정상 93%, 불량 7%

○ 분석기법

- 지도학습 : KNN, Random Forest
- 비지도학습 : Gausian Mixture 모델

○ 분석과정 및 결과

1) 전체 데이터로 학습
   - Accuracy 93%
   - 문제점 : 불량데이터에 대한 예측성능이 매우 낮음 (Precision 0.38, Recall 0.01, f1-score 0.02)
2) 정상, 불량 학습데이터 비중을 맞춤 (1 : 1)
   - Accuracy 66%
   - 문제점 : 불량데이터 예측성능은 향상되었으나 (Precision 0.38, Recall 0.71), 전체 Accuracy 낮아 짐 (66%)
3) 이상치 탐지 분석 (가우시안 혼합모델)
   - 비정상 데이터에 대한 예측성능이 지도학습과 유사한 수준으로 낮음
4) PolyNomial 방법
   - X 변수를 증가시키는 방법 : 분류분석이고, 이미 X변수가 많아서 시도하지 않음

○ 문의사항

- 판단기준 : 편중된 데이터에 대해서 어느 정도의 수치가 나왔을 때 연관성이 있다고 판단해야 할지 판단기준은?
- 분석방향 : 편중된 데이터에 대해서 불량(이상치)을 예측할 수 있도록 모델학습 시, 시도해 볼만한 방법은?

Performance Measures

- Performance Measures
  - ✓ Confusion matrix for novelty detection

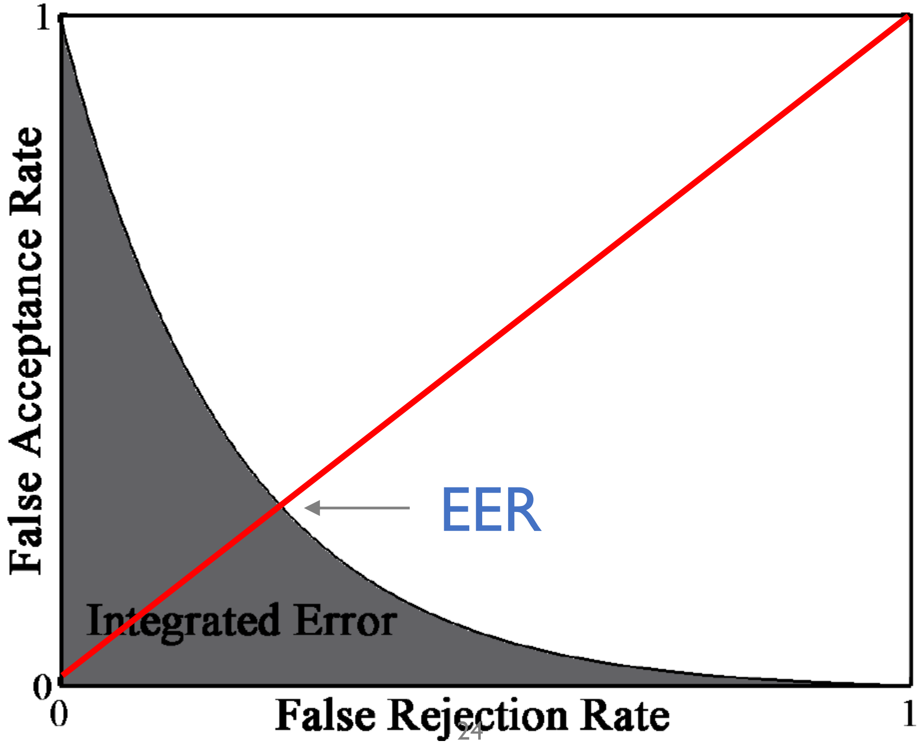
		Predicted class	
		Abnormal	Normal
Actual class	Abnormal	A	B
	Normal	C	D

  - ✓ Performance measures when the cut-off (threshold) is set

Metric	Description
Detection Rate	(Identified as abnormal)/(Actually abnormal) = A/(A+B)
False Rejection Rate (FRR)	(Rejected as abnormal)/(Actually normal) = C/(C+D)
False Acceptance Rate (FAR)	(Accepted as normal)/(Actually abnormal) = B/(A+B)

Performance Measures

- To evaluate an intrinsic performance of novelty detection algorithms
  - ✓ Equal error rate (EER): Error rate where the FAR and FRR are the same
  - ✓ Integrated Error (IE): the area under the FRR-FAR curve
    - AUROC for classification: the higher the better
    - IE for anomaly detection: the lower the better




Q & A