Techtonic 2018

-

Thu . Nov 15

-

SAMSUNG SDS Tower
West Campus B1F
Magellan Hall /Pascal Hall

SAMSUNG SDS
Realize your vision

데이터 사이언티스트를 위한

리얼 제조 현장에서의 데이터 분석 노하우

삼성SDS 조성호

Techtonic 2018

Agenda

- 제조 현장 환경
- 데이터 확보
- 데이터 정제
- Lessons Learned

리얼 제조 현장에서의 데이터 분석 노하우

제조 현장 환경

제조 현장의 고민

대용량 영상 데이터를 처리하는 제조 환경에서, 딥러닝 기반 검사 시스템 구축 시 경험한 어려움들

“대량 데이터, 고된 정제 작업”

Ex. 제조 결함 유형 분류

- 10만 건 이미지 Labeling 작업
  전문 인력 활용하여 25일 소요
  → 정제 일관성 80%

“경험에 의존하는 개발”

최적 모델 생성 탐색에 많은 시간 소요

- 네트워크 3개 이상 탐색
  (ResNet→DenseNet → VGG → AlexNet → ResNet)
- Layer 수 및 각종 학습 파라미터
  조합 케이스 별 성능 확인 및 수동 트래킹

“신규 데이터, 다양한 검사 환경”

현장 데이터 적용 시 모델 성능 예측/유지의 어려움

- Ex) 제조 결함 이미지 10만 건 활용하여 학습한 모델
  실제 현장 데이터 적용 시, 20~25% 하락

생각과는 다른 현장 현황

Supervised Learning을 위해 정제가 가능한 Data를 충분히 확보하고 있는 제조 현장은 아주 적음

Smart Factory

- Smart Factory!!! 제조 현장은 이렇겠지?
  왠지 데이터가 잘 관리되어 있을 듯한 환경

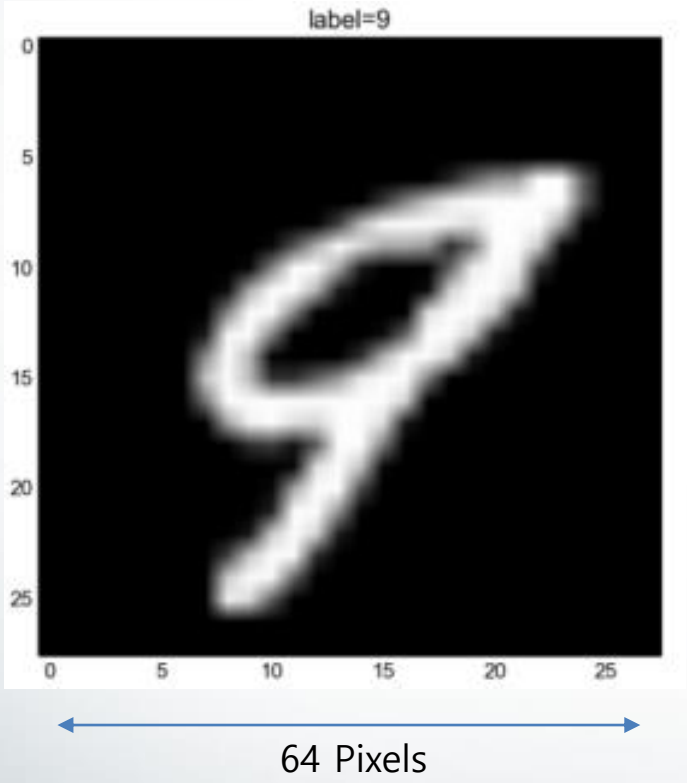
Real World

- 현실은?
  실제 분석 요청을 받은 이미지들 … ….

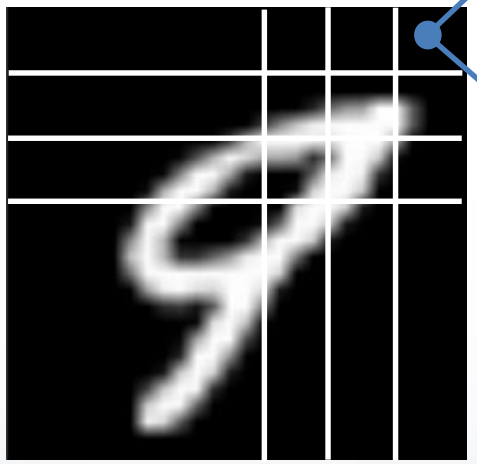
주변에서 쉽게 구할 수 있을까?

MNIST와 같은 학습용 데이터는 제조 현장의 실 데이터와는 많은 차이가 있어 해당 데이터들을 이용한 모델은 현장 적용이 어려움

- **학습용 데이터의 특성**



- 전체 크기는 작음
- 대상의 크기는 큼
- 중앙에 위치
- 비교적 뚜렷한 명암과 특징



깔끔한 배경

0	0	0
0	0	0
0	0	0

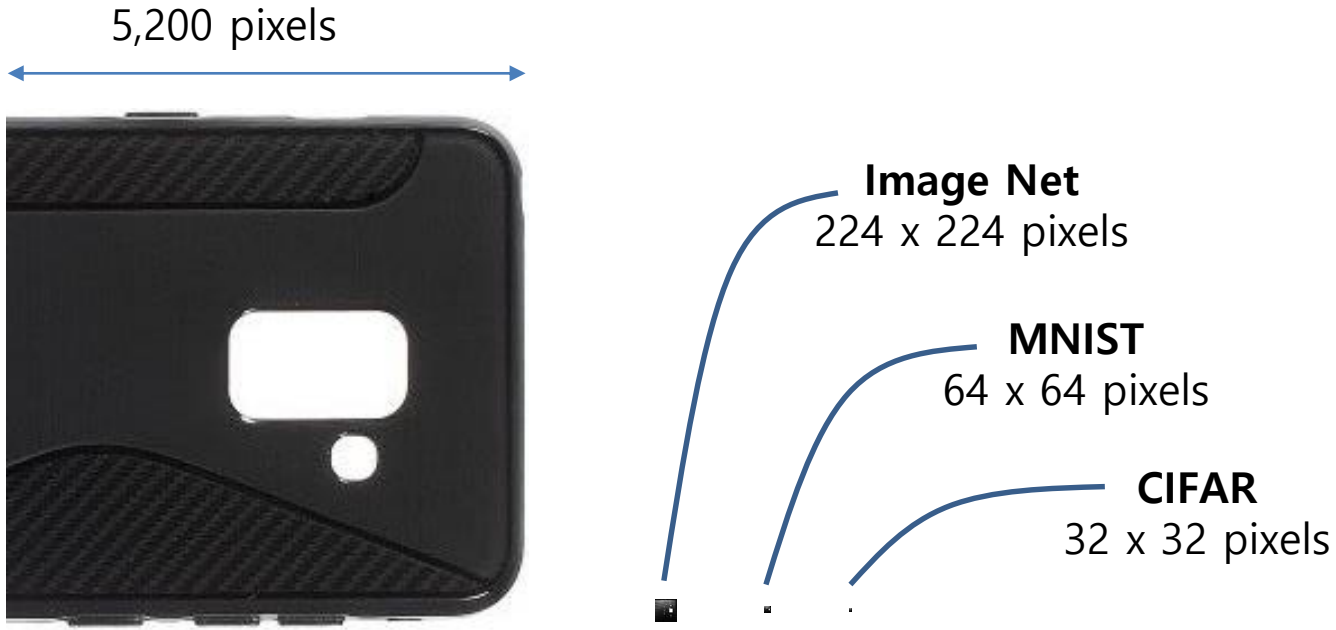
배경에 노이즈가 있다면?

0	0	0
0	15	0
25	0	10

주변에서 쉽게 구할 수 있을까?

접사경, 라인 스캐너, 고해상도 산업용 카메라, SEM(전자 현미경) 등을 통한 이미지 수집
→ 아주 작은 이미지부터 4K ~ 12K 고해상도 이미지 다양

학습 데이터와 실 데이터 해상도 비교



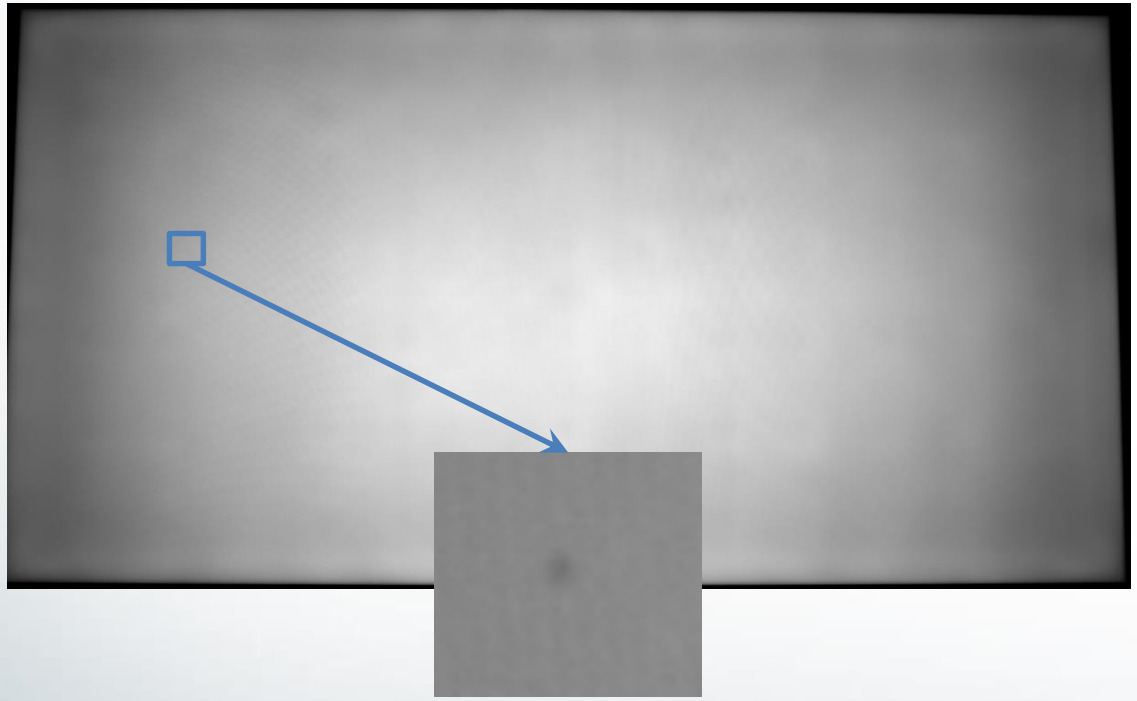
- **다양한 사이즈, 다양한 해상도**
- **보통은 고해상도, Big Size**
- **다양한 위치**
- **식별 어려움**
  - Viewpoint Variation
  - Illumination
    Deformation
  - Occlusion
    Background Clutter
  - Intra-class Variation

주변에서 쉽게 구할 수 있을까?

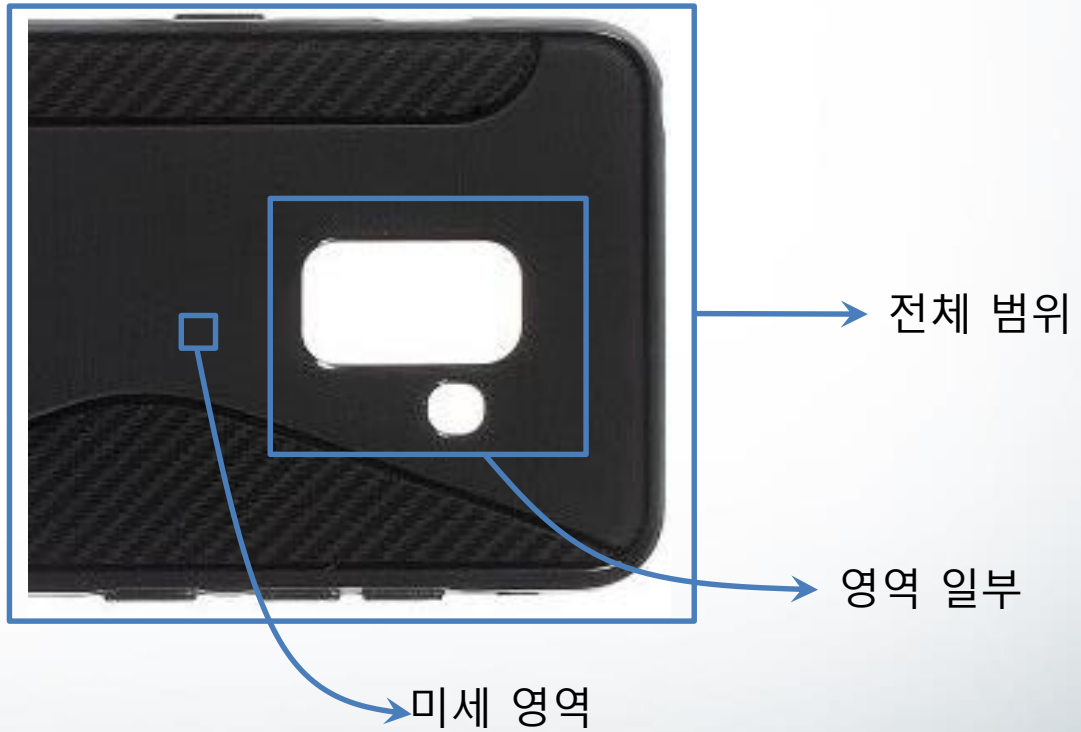
학습 데이터들이 대상이 고정되거나 하나의 특성만을 반영하는 데 비해
분석 대상은 몇 Pixels 수준의 몹시 작거나 복합적임

- **학습 데이터와 실 데이터 특성 비교**

아주 작거나



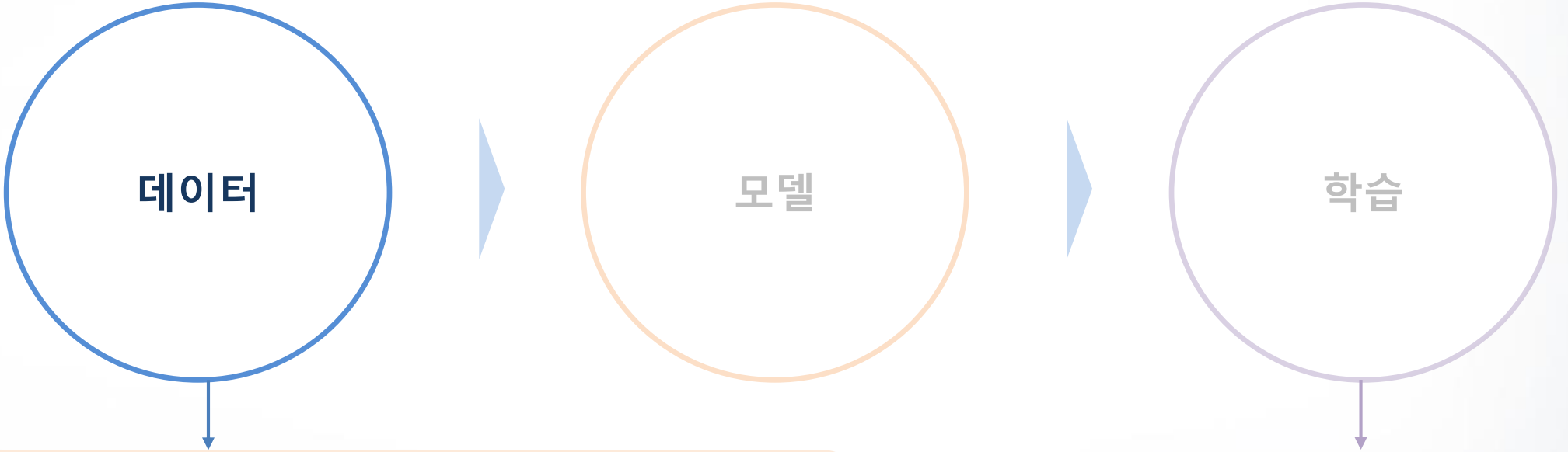
복합적



문제 영역

제조 현장에서 AI의 수준은,
학습에 활용되는 지식의 원천인 데이터의 Size와 Quality에 의해 큰 영향을 받음

AI 확보 과정



✓ 제조 현장에서 종종 부딪히는 상황

→ **데이터 확보** : 데이터가 부족하거나 불균형 하거나

→ **데이터 정제** : 사람마다 다른 기준, 방대한 데이터 …

???

높은 성능을 기대하기 어려움

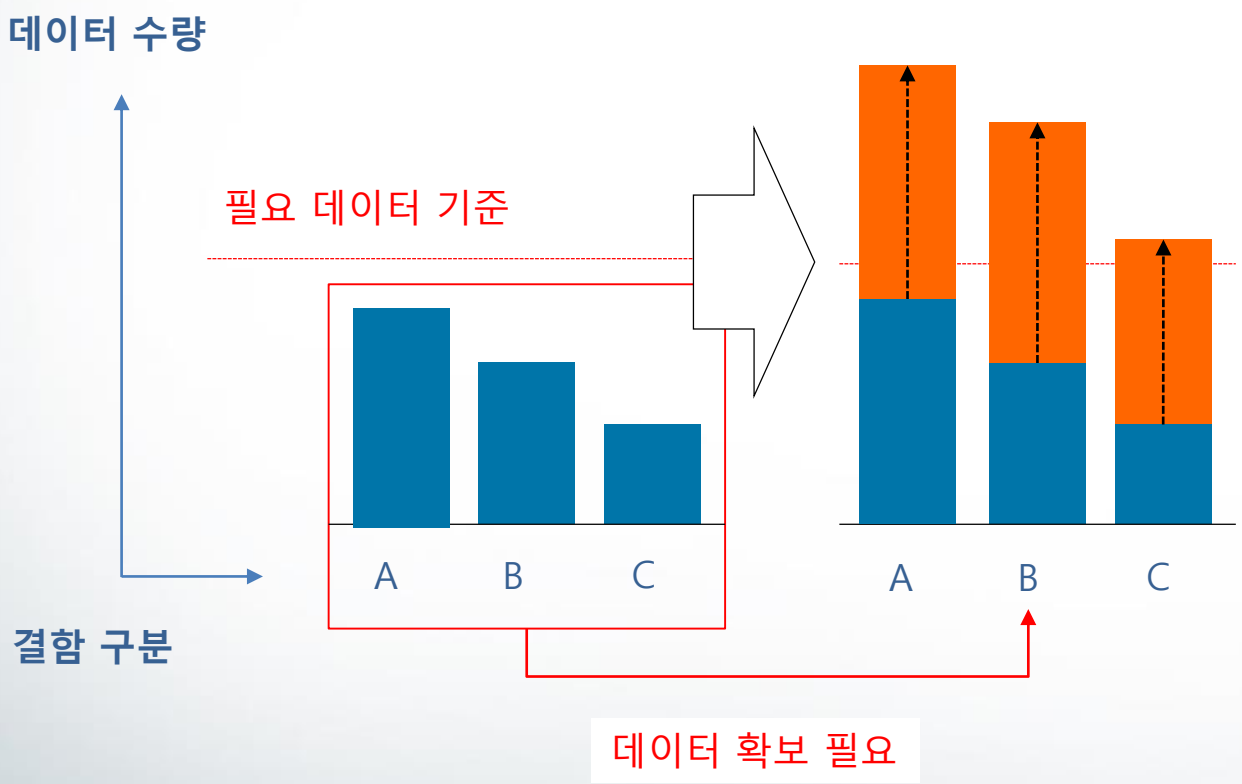
리얼 제조 현장에서의 데이터 분석 노하우

데이터 확보

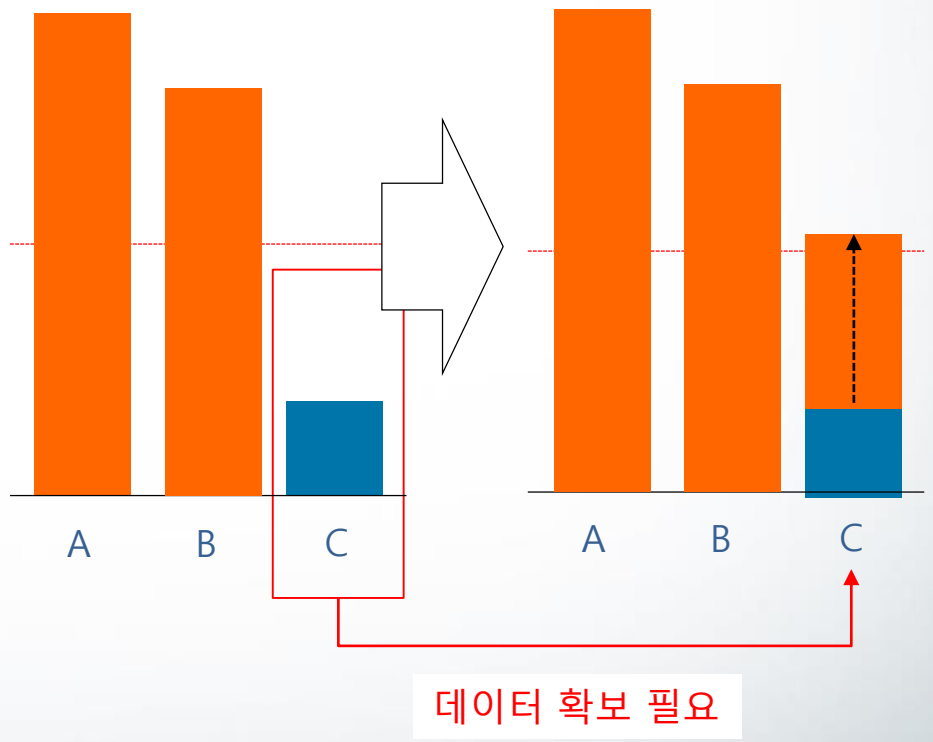
Problem

제조 데이터는 대량의 정상건, PPM 수준의 결함 등
데이터의 부족과 불균형 데이터의 분포 → 성능 향상을 위해 해결책 필요

[학습 데이터 부족]



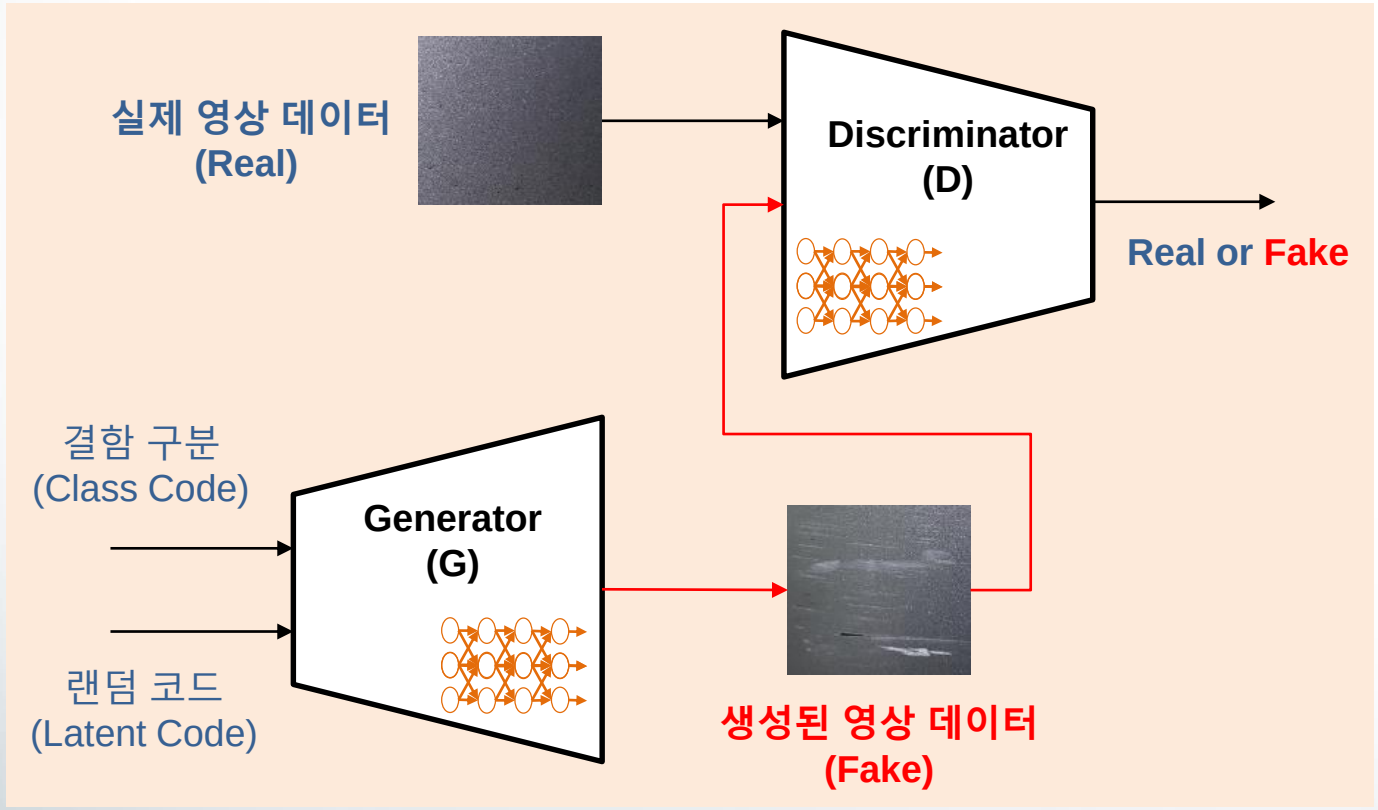
[불균형 학습 데이터]



Approach

데이터의 분포/형태를 학습하고, 이를 활용하여 유사 데이터를 생성 (데이터 생성 AI)

데이터 생성 AI 모델 : GAN(Generative Adversarial Networks)



[학습/활용 과정]

[D] Real/Fake 데이터 구분

[G] Fake 데이터 생성 (D를 속임)

학습

[D] Real/Fake 데이터 구분 불가

[G] Real과 유사한 Fake 데이터 생성

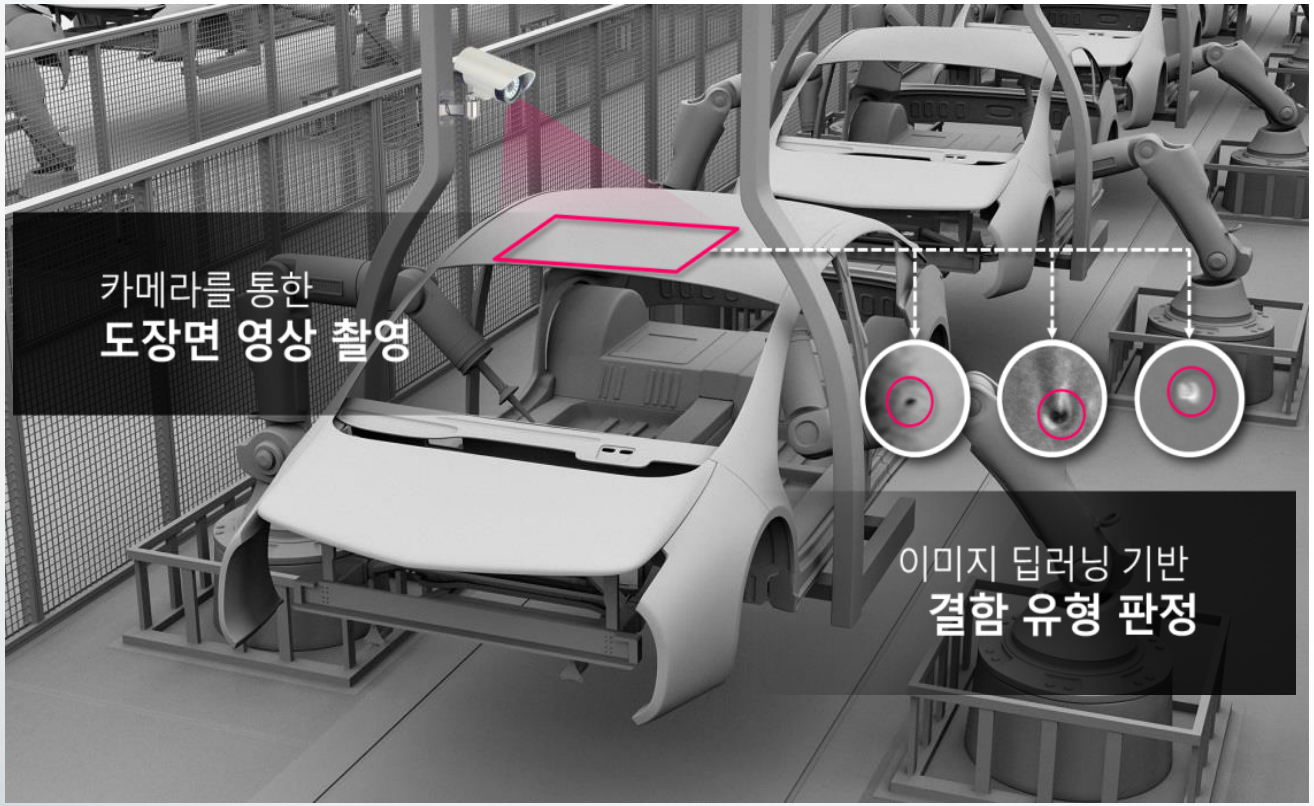
활용

[G] 원하는 Class에 해당되는 Fake데이터를 Random 생성 (Real과 거의 유사)

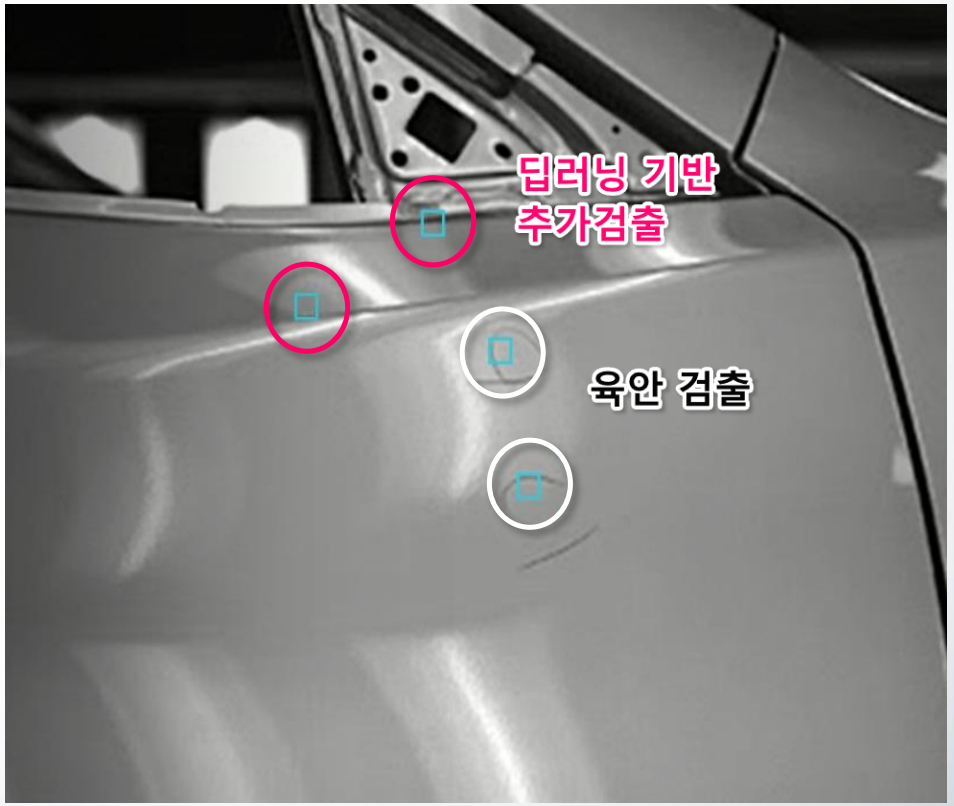
Case Study

딥러닝 학습을 통해 생성된 Classifier를 이용하여 외관 결함에 대한 판정

도장면 이미지 확보 후 딥러닝 기반 결함 유형 판정



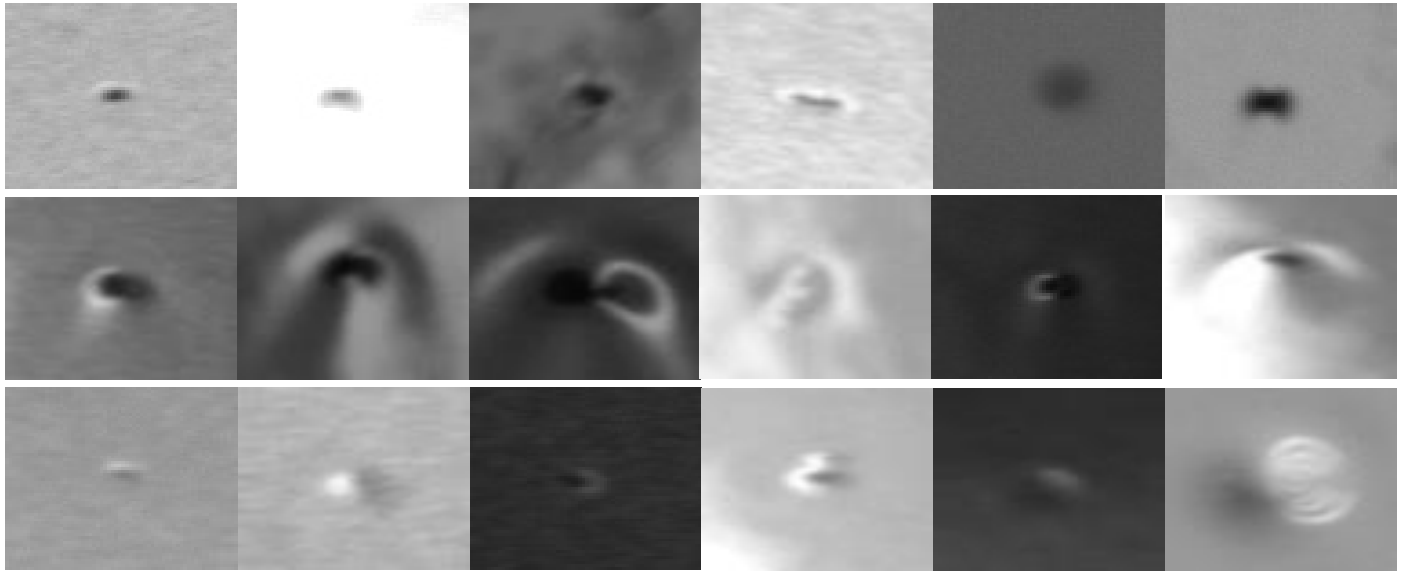
추가 불량 검출 확인



Case Study : 데이터 현황

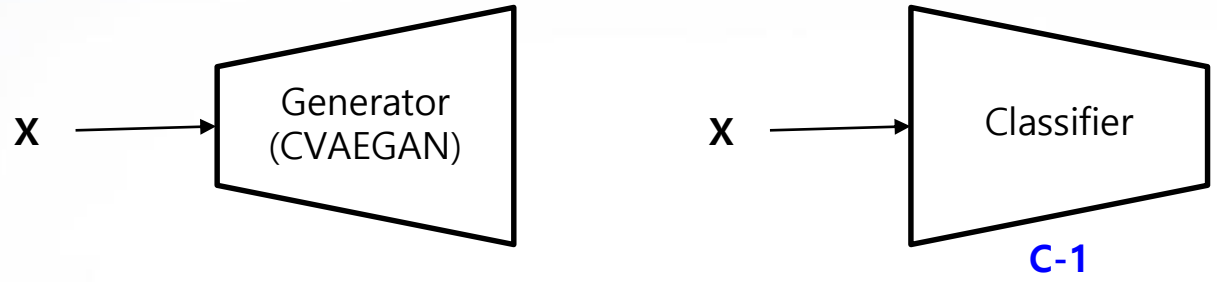
전체적으로 부족한 학습 데이터를 보완하기 위해 데이터 추가 생성 필요

- **Pinhole** : 검은 점 형상의 구멍 혹은, 분화구 형상이 매우 완만
- **Crater** : 분화구 형상이 명확 혹은 여러 겹이며 중심에 구멍 혹은 핵이 위치
- **Pollution** : 볼록하게 솟아 오른 형상이며, 반사 영역과 그림자 영역이 명확한 경우

Class	Size	Original image (예시)
Pinhole	808	
Crater	1,136	
Pollution	1,314	

Case Study : 데이터 생성(GAN)

Step-1. Generator & Classifier training



X : Real Data for Training

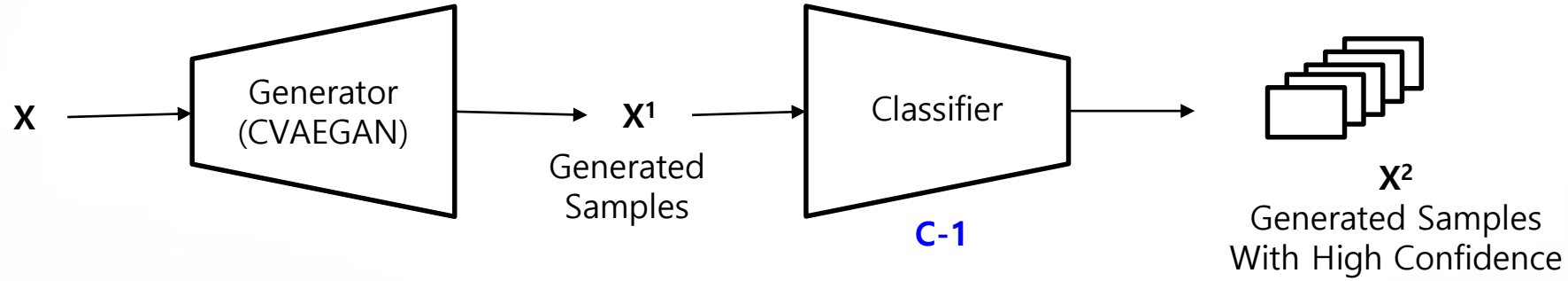
X^1 : Generated Samples

X^2 : High Confidence X^1

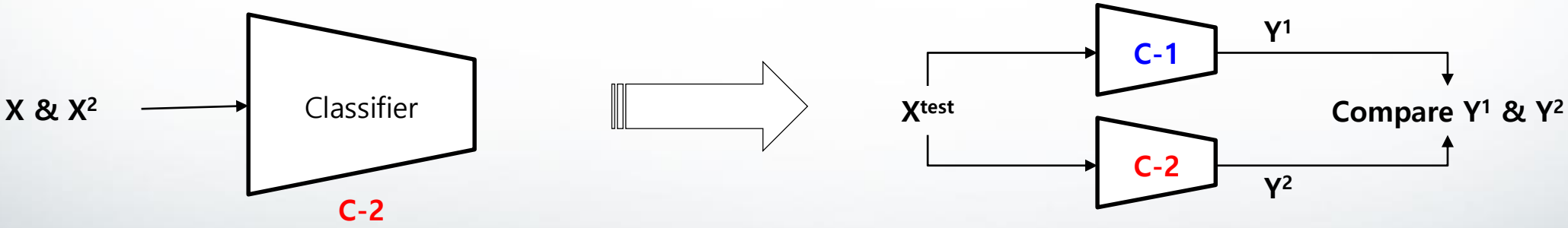
X^{test} : Real Data for Test

Y^1 & Y^2 : Classification Result

Step-2. Training sample generation



Step-3. Second classifier training & Evaluation test

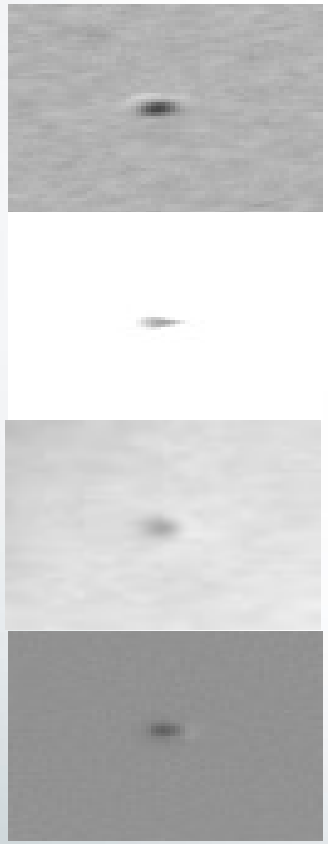


Case Study : 생성 결과

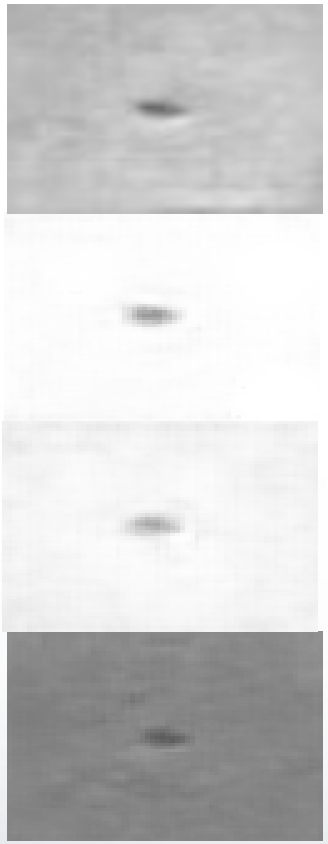
도장 결함 원본 이미지와 GAN에 의해 생성된 이미지 비교

[Pinhole]

Original

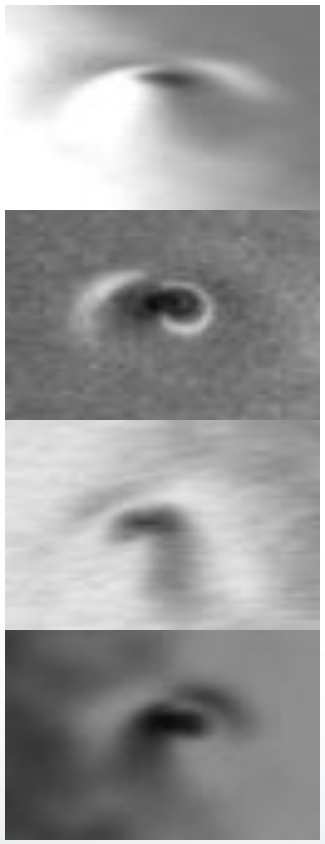


Generated



[Crater]

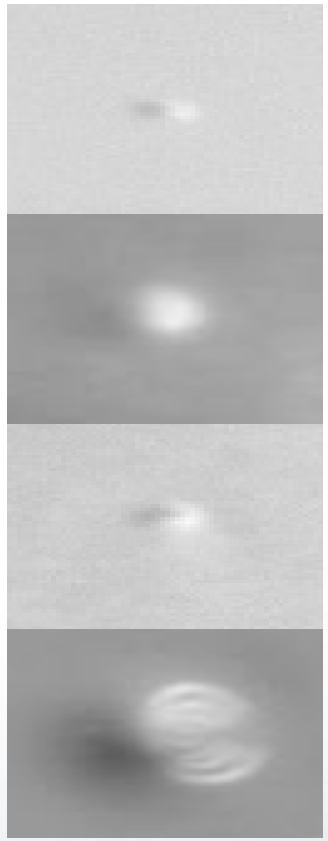
Original



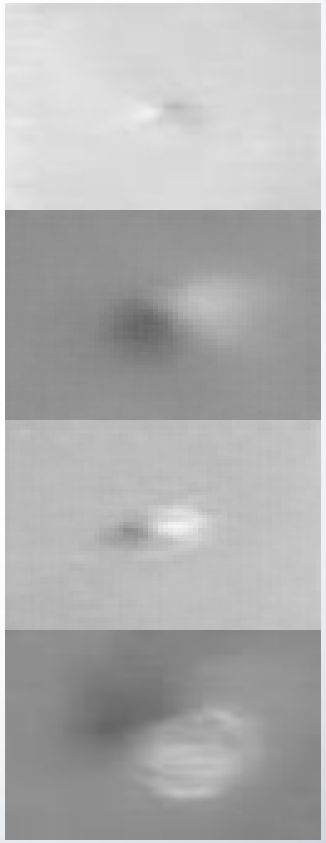
Generated

[Pollution]

Original

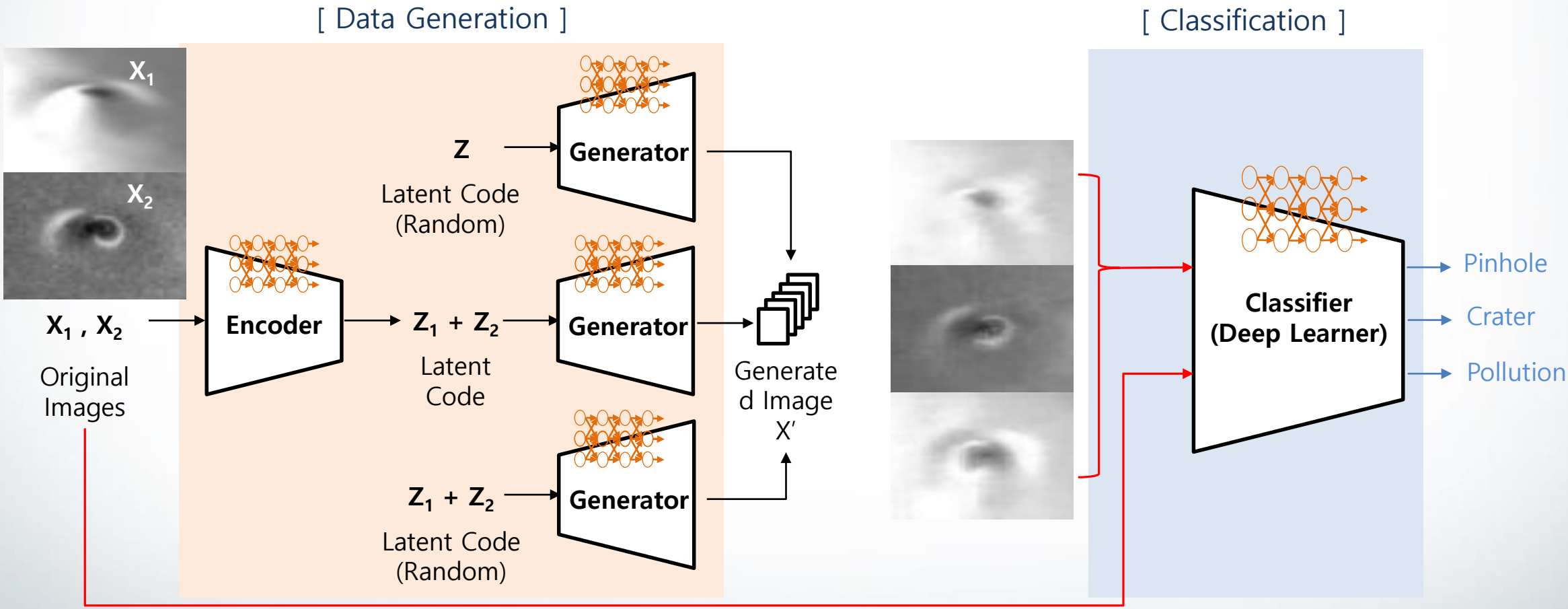


Generated



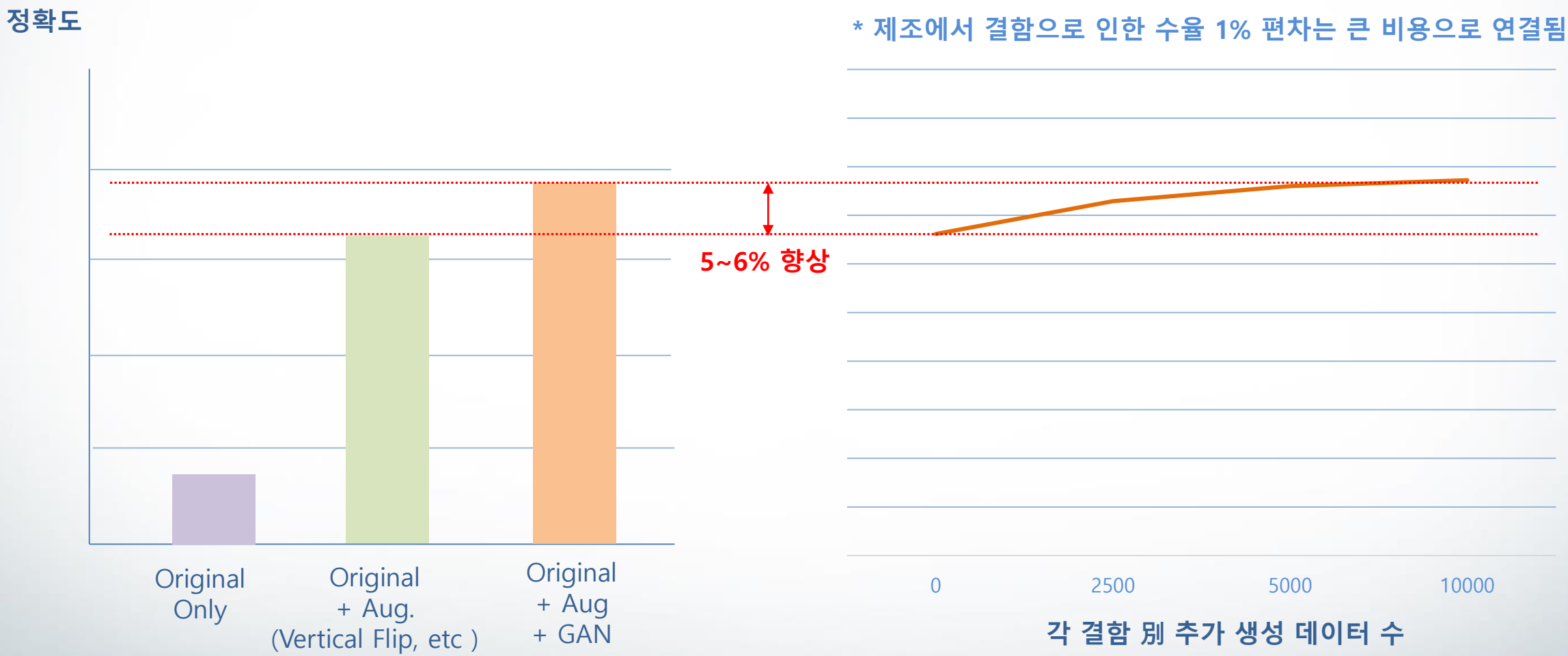
Case Study : 학습 적용

1. 원본 데이터의 분포/형태를 학습하고, 이를 활용하여 유사 데이터를 생성
2. 생성된 데이터를 이용하여 Classifier 학습



Case Study : 적용 효과

GAN을 이용한 데이터 생성을 통한 학습 효과 : 5~6%*의 분류 정확도 향상을 보임



※ 데모 영상은 비공개 처리 되었습니다.

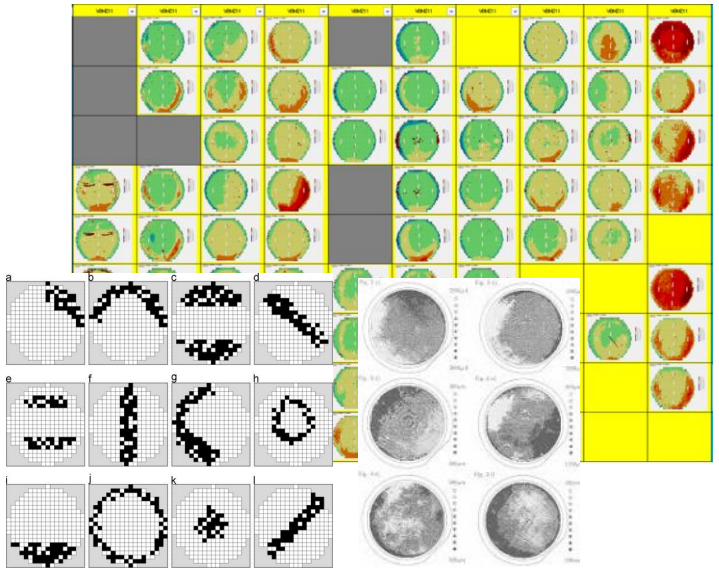
오픈소스 Ai분석 플랫폼 Brightics Studio

데이터 정제

Problem

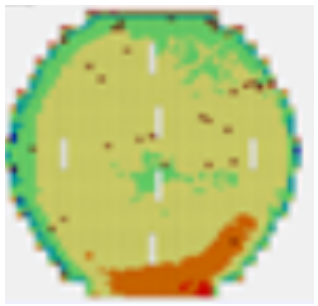
너무 많은 이미지, 통일이 불가능한 사람마다 다른 분류 기준, 지루한 데이터 정제 작업과 낭비

너무 많은 데이터와 조건들



Occlusion
Viewpoint Variation
Intra class Variance
Deformation
Clutter
Illumination

사람마다 상이한 기준 해석



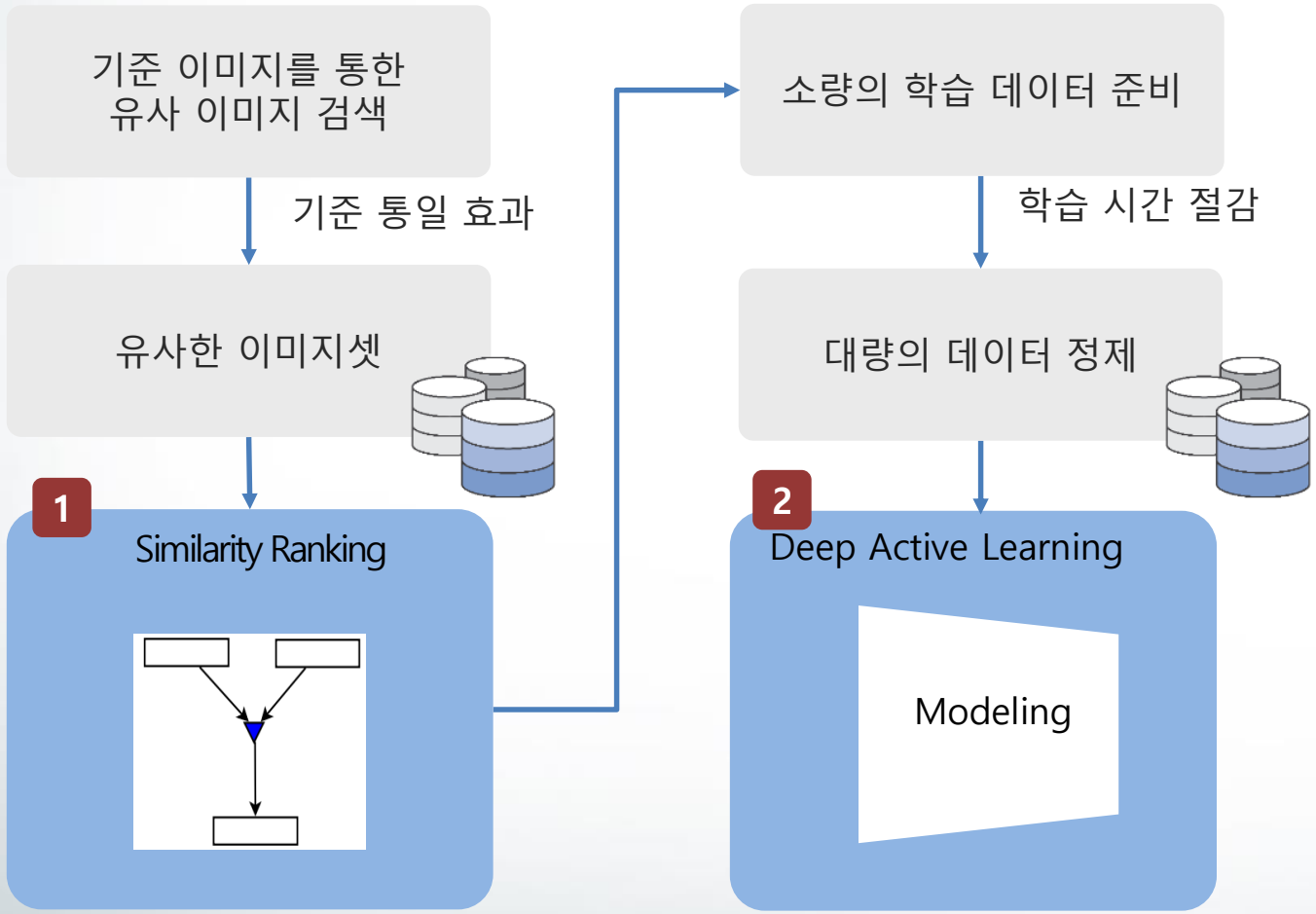
기준 이미지

사람마다 다른 해석

데이터 정제를 효율적, 효과적으로 지원하는 도구가 필요

Approach

데이터 정제, 표준화를 위한 프로세스



[데이터 정제 도구]

1

대량 데이터에서 이미지 질의를 통해
유사 이미지 수집

전달

2

수집된 유사 이미지들의 Ranking Score 이용,
일부 분류하고 대량 이미지 정제에 사용

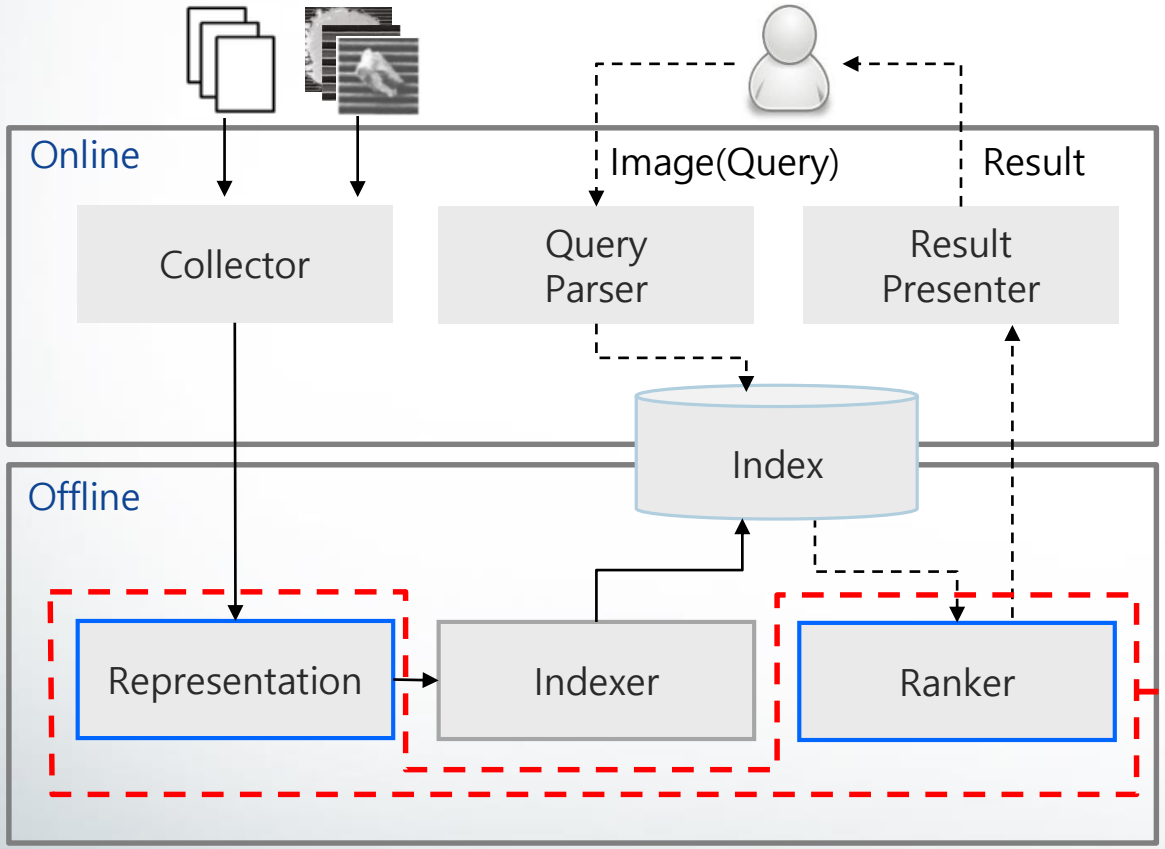
활용

학습 데이터 분류 기준 통일
학습 데이터 정제 시간 절감

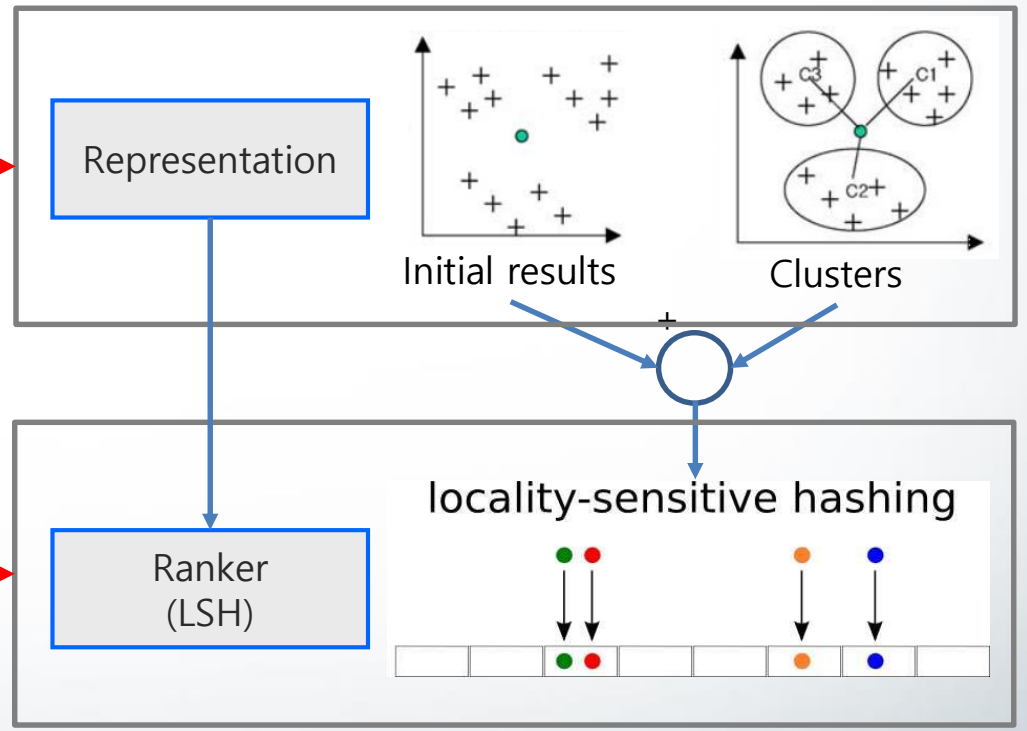
Approach

SDS의 이미지 검색 기술인 Visual Search 기술 중 Similarity Ranking 이용
→ 대량 데이터에서 유사 이미지 검색을 통해 후보 이미지 추출

SDS Visual Search Architecture



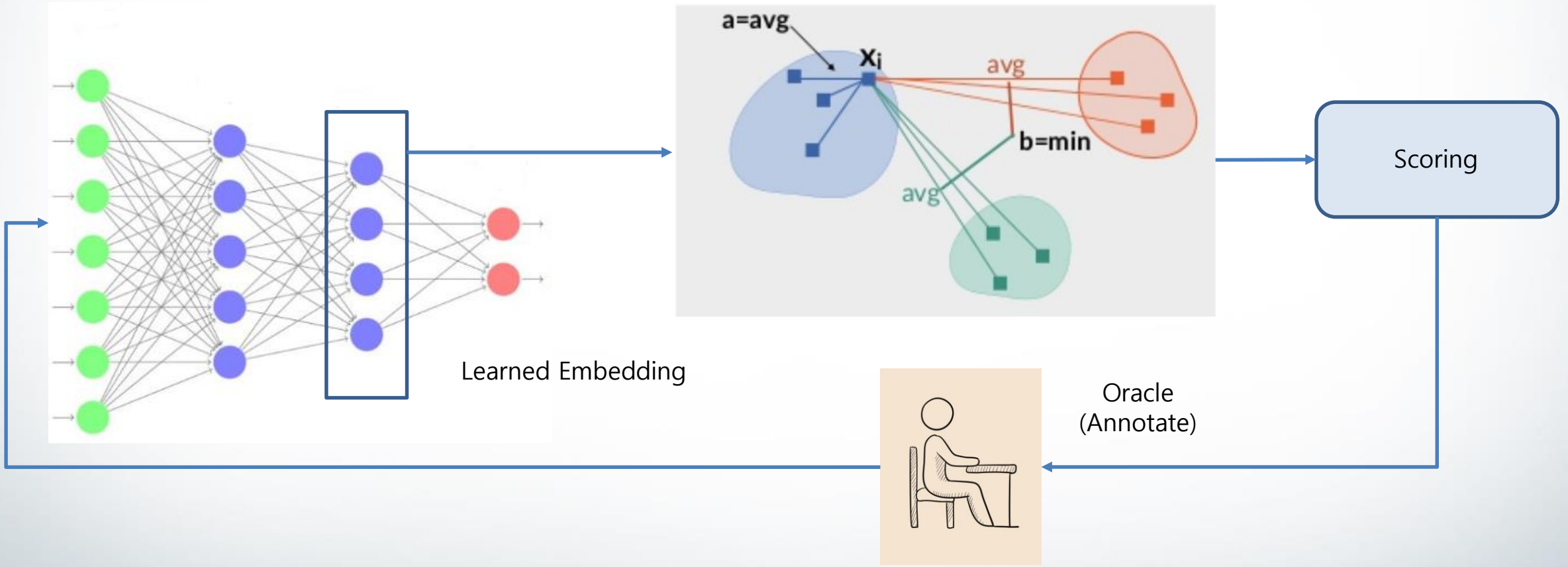
Similarity Ranking



Approach

Active Learning 기반 소량 학습 데이터 셋을 반복 적용하여 대량 데이터에 대한 효과적 분류 수행

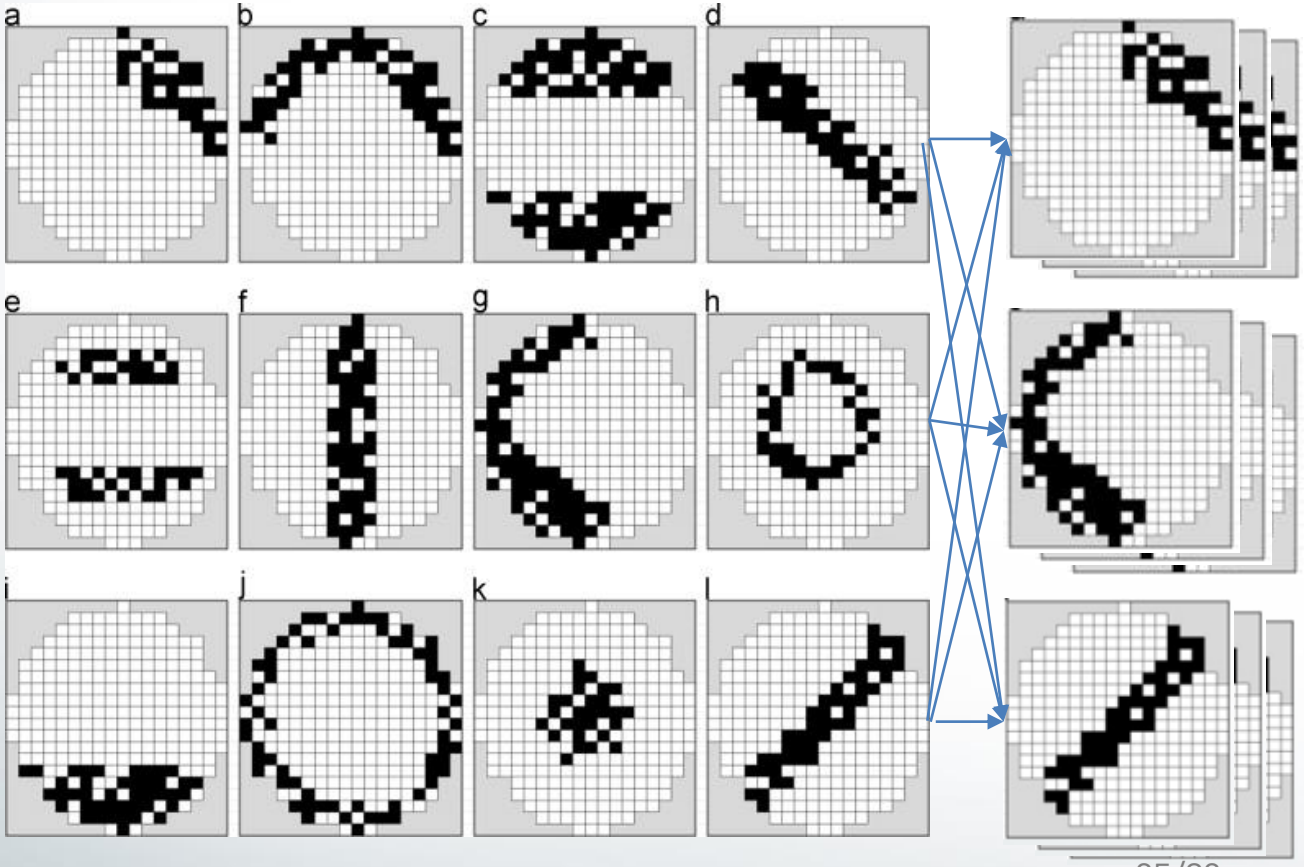
Deep Active Learning



Case Study

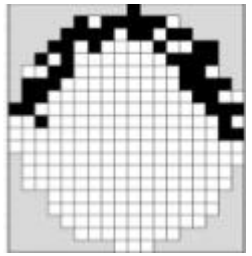
제조 결함 패턴이나 이미지를 분류하고 종합적 판단하여 수율 개선에 활용

딥러닝 기반 결함 패턴 분류



수율 개선 분석

분류 결과 결과



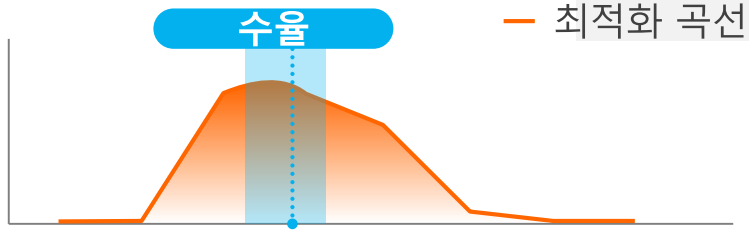
- 결함 패턴 분류
- 유형 분석 및 추적

Legacy

기 견적 DB

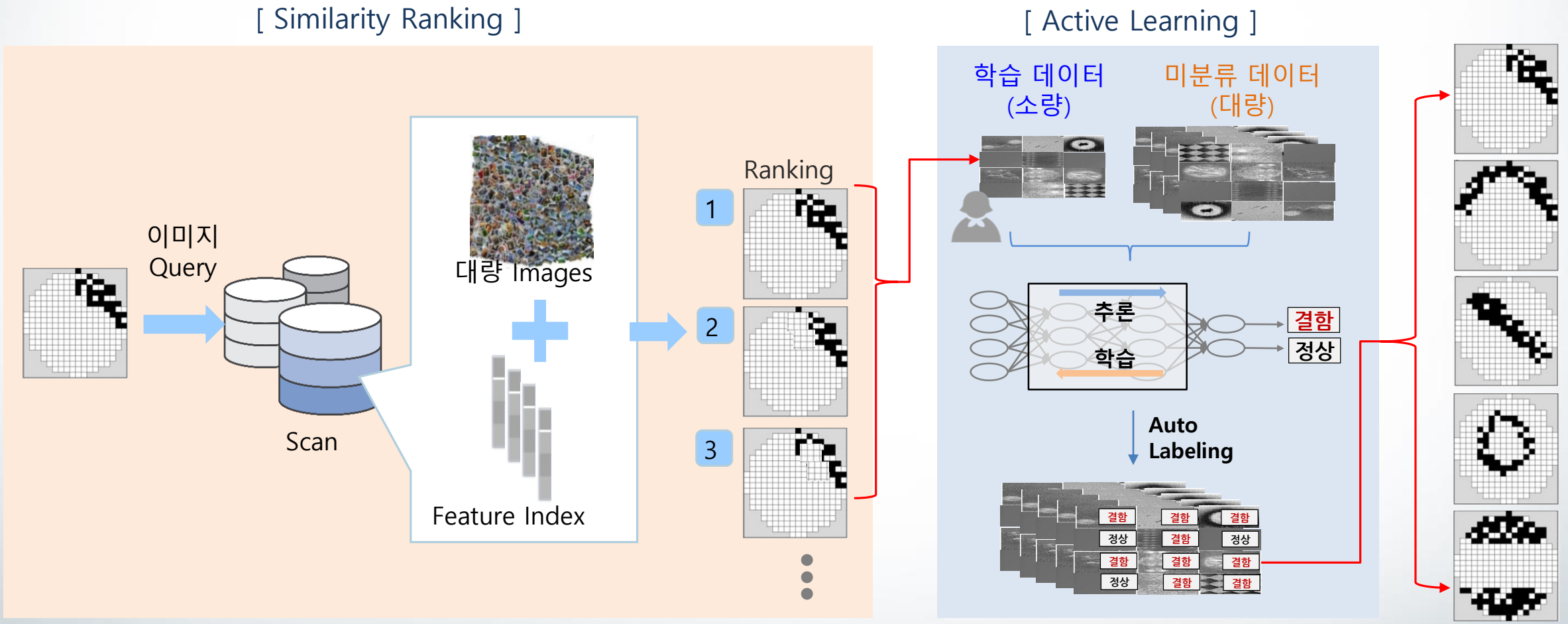
관련 설비 이력

설비 최적화 or 설계 변경



Case Study : 유사도 분석을 통한 데이터 정제

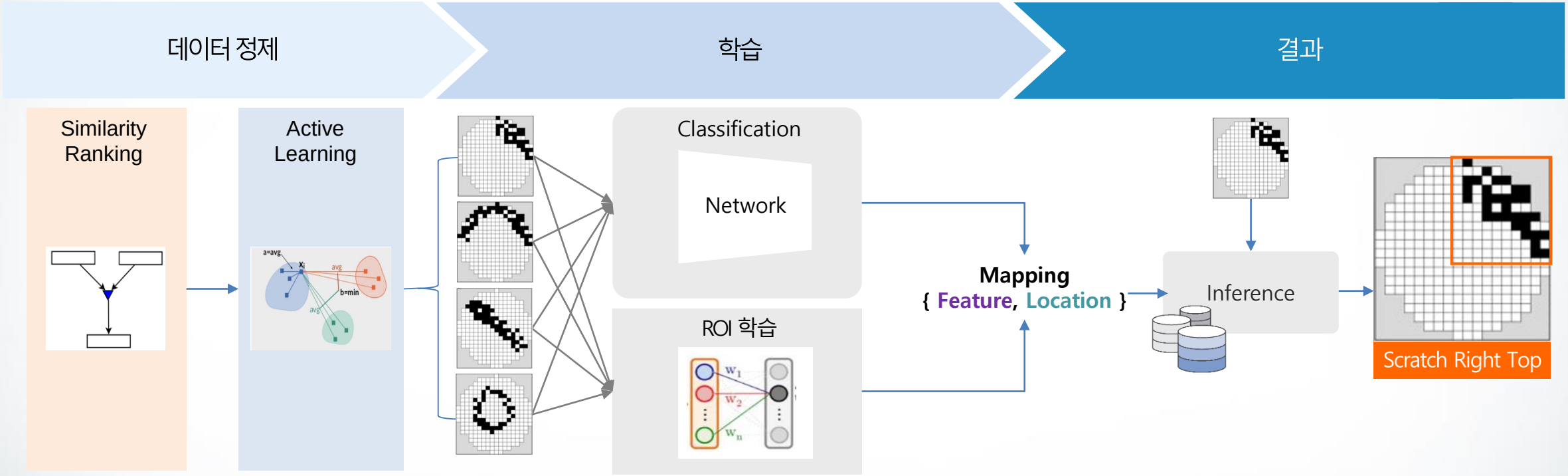
유사도 분석(Similarity Ranking)을 통해 대량 데이터에서 후보 Dataset을 추출하고 해당 Dataset을 학습하여 생성한 모델을 다시 대량 데이터에 적용(Active Learning)해 필요한 학습 데이터 정제



※ 데모 영상은 비공개 처리 되었습니다.

Case Study : 모델 적용

대량 정제된 데이터를 적용하여 모델 학습 수행

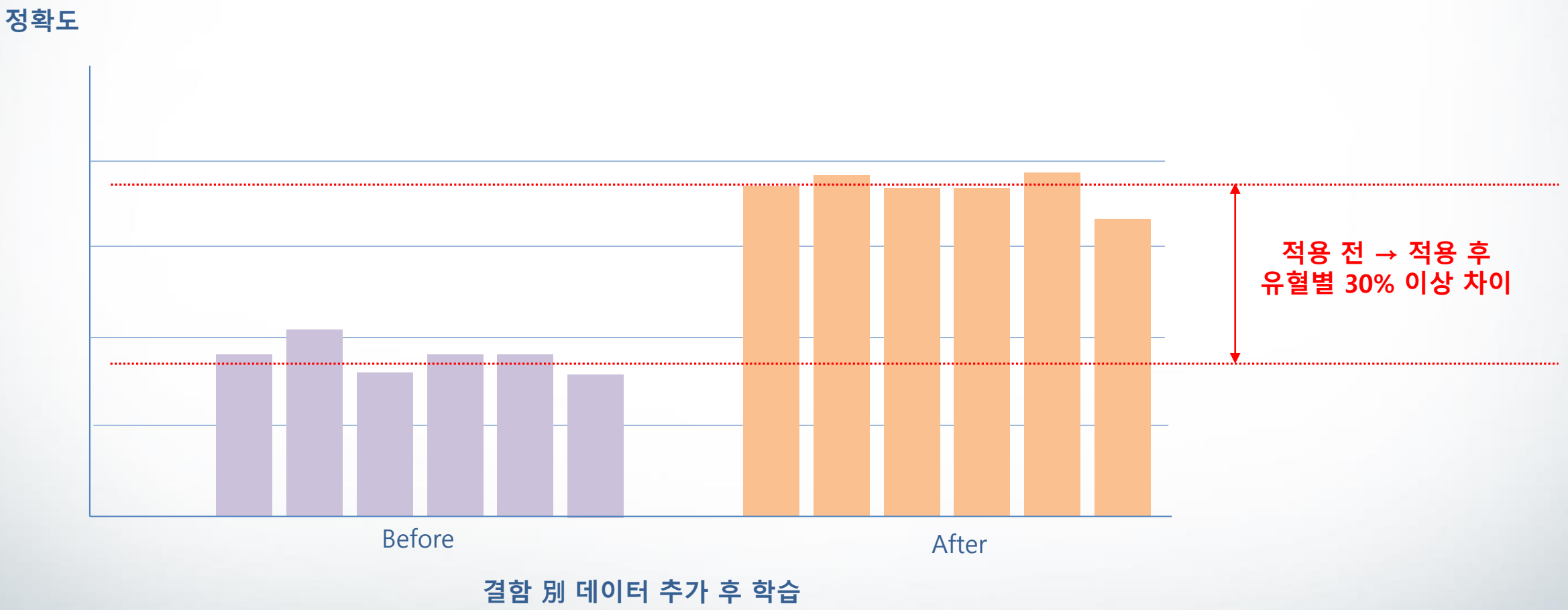


▸▸▸대량 데이터 정제, 학습 → Supervised Learning 단점 극복

대량 데이터 정제 시간을 효과적으로 줄이고 정제된 대량 데이터를 투입하여 모델 성능을 개선

Case Study : 적용 효과

데이터 정제를 이용한 정확도 향상과 작업 시간 단축
→ 검수 결과 부적합, 재작업이 반복되어 30일 간 20인 기준 4만 장 수집 → 수시 수집/정제



※ 데모 영상은 비공개 처리 되었습니다.

리얼 제조 현장에서의 데이터 분석 노하우

Lessons Learned

Lessons Learned

데이터 수집

GAN

- 출현 빈도가 낮은 결함 패턴에 대한 분류 성능 강화
- 생성할 데이터 개수에 대한 적절한 기준 파악 필요

데이터 정제

Similarity Ranking
Active Learning

- 충분하고 적절한 Labeling 데이터 확보에 대한 어려움을 보완
- 배경이나 특징이 너무 복잡할 경우에 대해 보완 필요

Q & A

Thank you