AI 기만 공격 기술

박호성
부산외국어대학교

목차

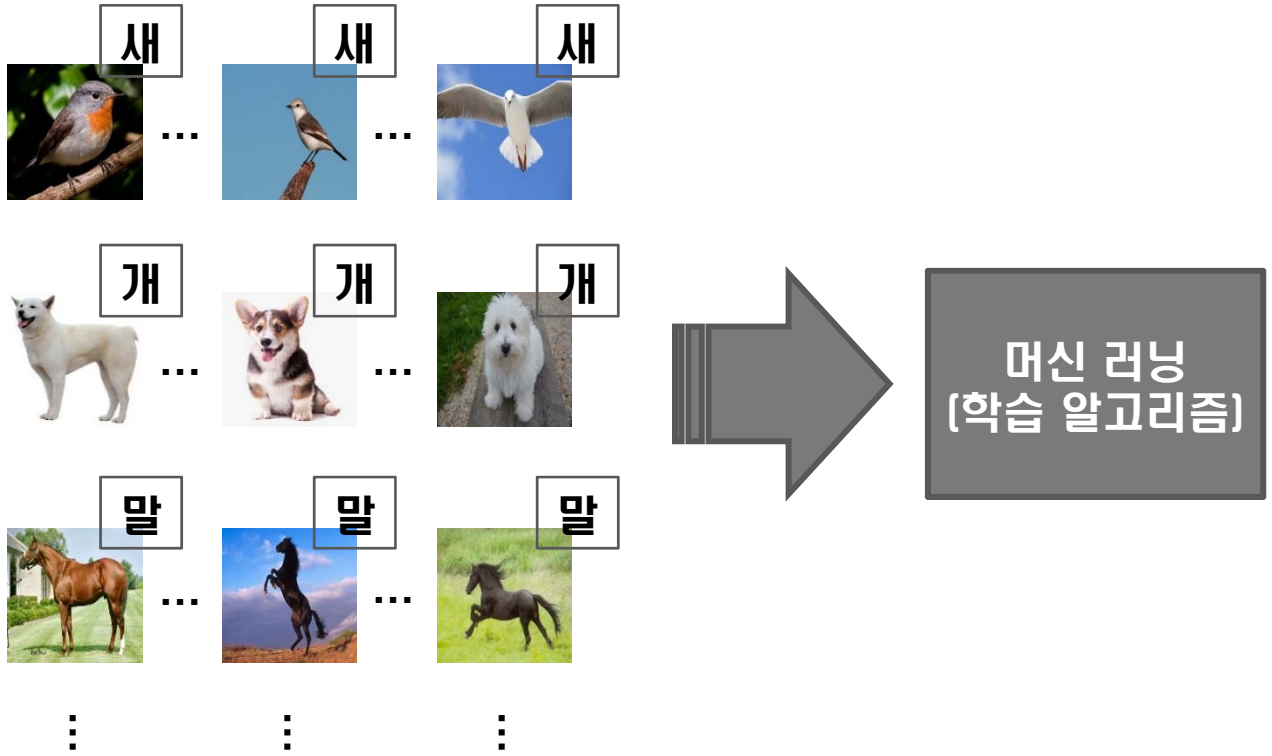
- AI 학습 (이미지 분류 모델)
- AI와 보안
- AI 보안 위협
- AI 관련 연구 소개
- ~~AI 보안 대책~~

AI 학습

Classification Model

▶ 학습 단계

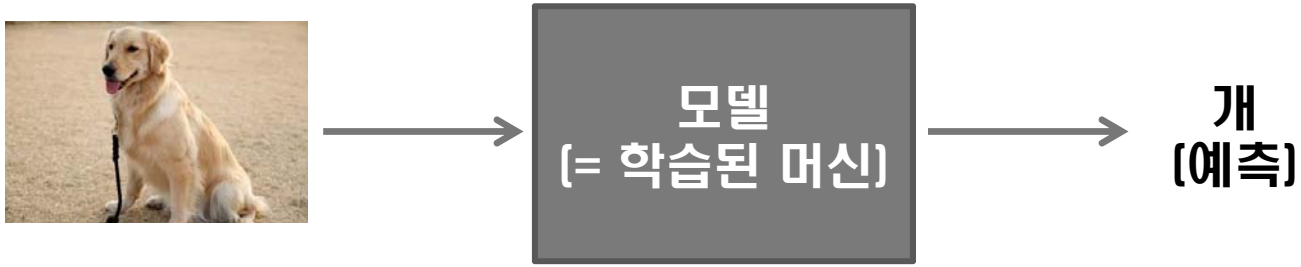
- 다량의 학습 데이터 ➔ 스스로 패턴 분석



Classification Model

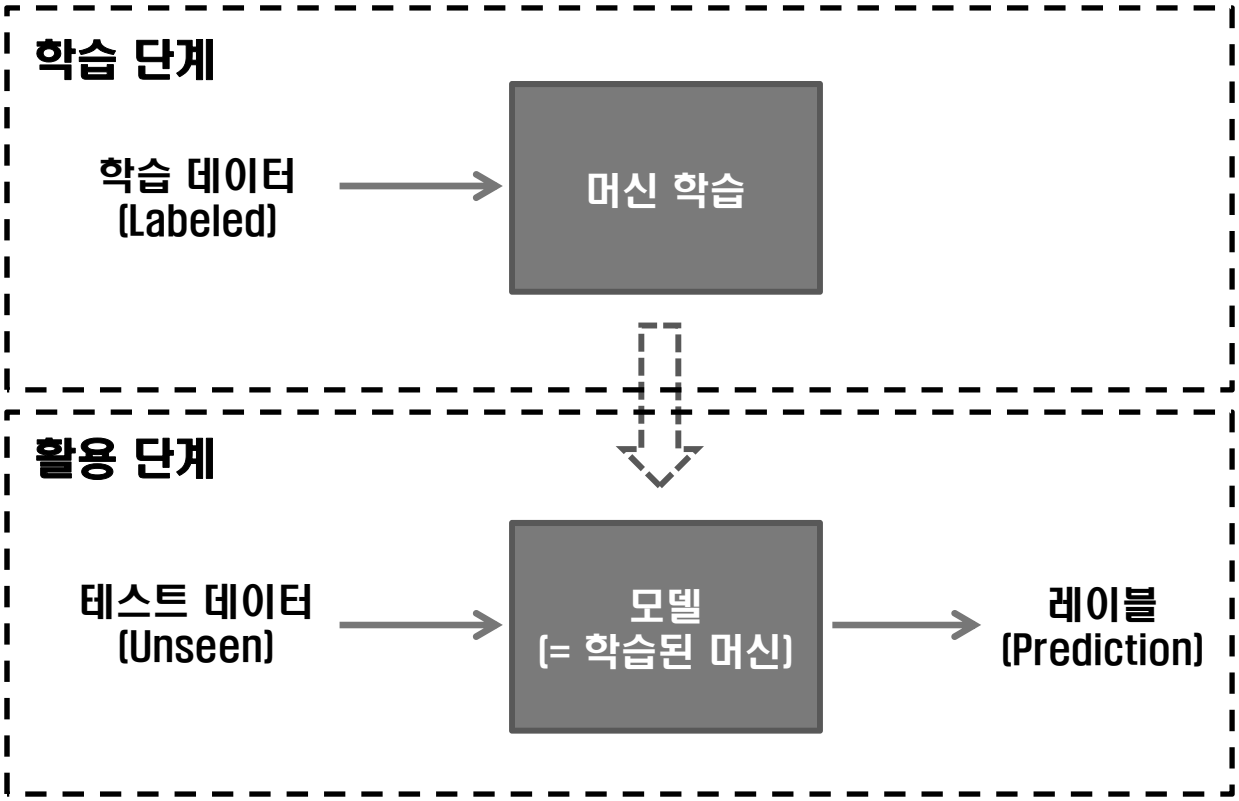
▶ 활용 단계

- Unseen data ➔ 결과 예측



Classification Model

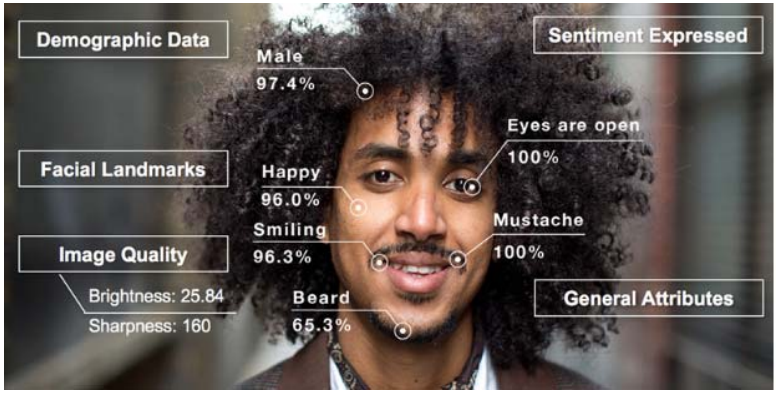
머신 러닝을 통한 이미지 분류



대표적인 AI 서비스

▶ 객체 인식

Amazon Recognition



얼굴 인식 및 분석

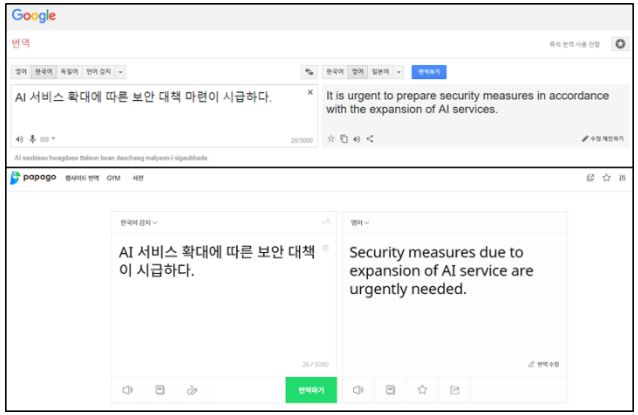
대형폐기물 인식 서비스

대표적인 AI 서비스

▶ 자연어 처리

AI 스피커 음성 인식

챗봇

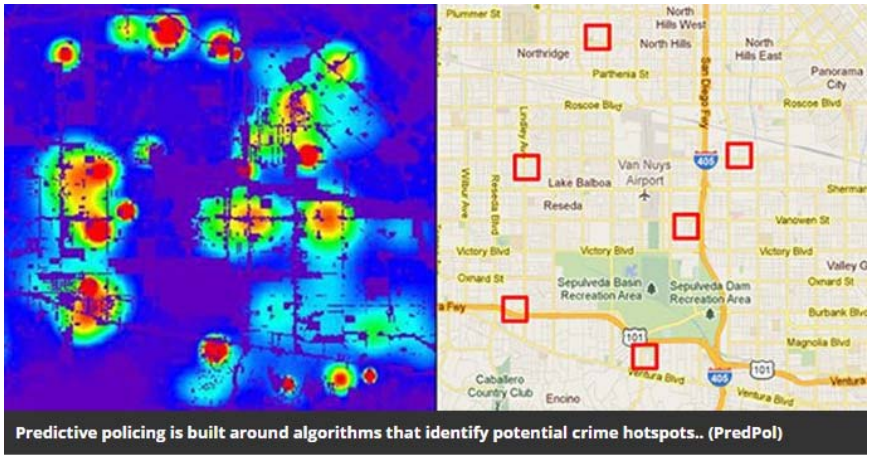


언어 번역

대표적인 AI 서비스

▶ 상황 인식

CCTV 분석

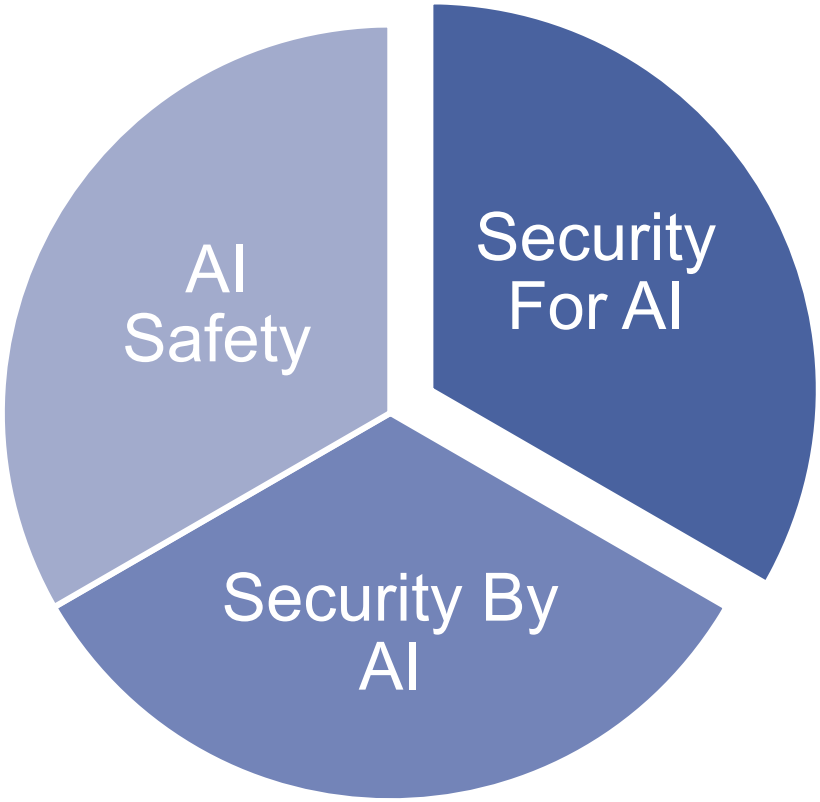


범죄 예측 시스템

AI와 보안

AI – 보안 관점

AI로 인한
(의도치 않은)
사고 방지



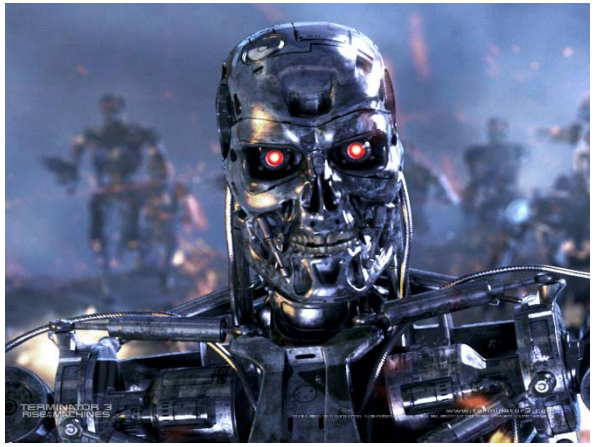
AI에 대한
공격을 방어

AI를 활용한 보안

AI Safety

▶ 사고(Accident) 방지

- 디자이너의 의도와 다른 유해 행동 가능성 대비
- 설계 원칙 / 모니터링 / 제약과 감독 등



Concrete Problems in AI Safety

Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

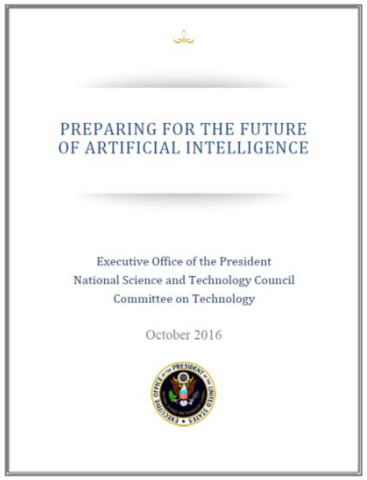
▶ 5가지 Safety issues

- Avoiding Negative Side Effects
- Avoiding Reward Hacking
- Scalable Oversight
- Safe Exploration
- Robustness to Distributional Shift

AI Safety

▶ 대응

- 미국 정부 NSTC (2016) 'AI R&D 전략 계획'
  - AI Safety 이슈를 살펴보고
  - AI R&D 시 항공, 우주 등 다른 분야의 안전 원칙을 도입할 것



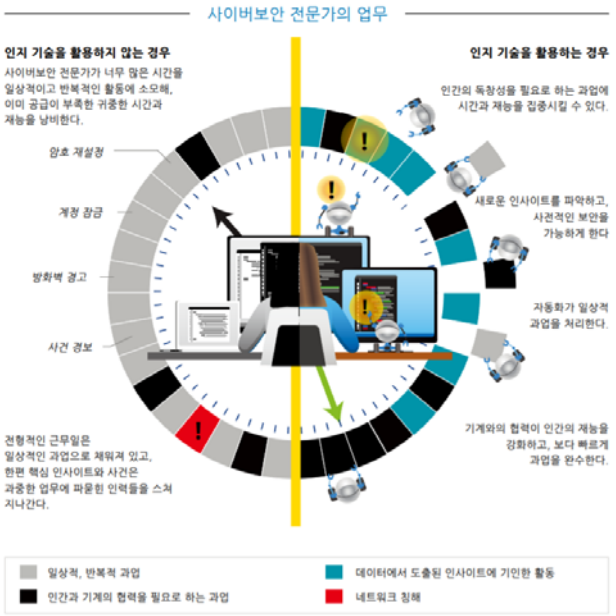
- Future of life institute (연구단체)
  - 앨론 머스크의 펀드로 AI safety 관련 30여 개 프로젝트 진행 중
  - Strategic Research Center for Artificial Intelligence
  - How to Build Ethics into Robust Artificial Intelligence
  - Robust and Transparent Artificial Intelligence Via Anomaly Detection and Explanation
  - Evaluation of Safe Development Pathways for Artificial Superintelligence

Security by AI

AI를 활용한 보안 강화

- 사용자 인증 : 생체인증, 행위인증
- 이상거래 탐지 : 비정상 트랜잭션 탐지
- 악성코드 탐지 : 알려지지 않은 악성코드
- 사이버 공격 탐지 : 로그 => 공격 탐지
- 지능형 CCTV : 이벤트 (침입, 화재..) 탐지
- 지능형 보안관제

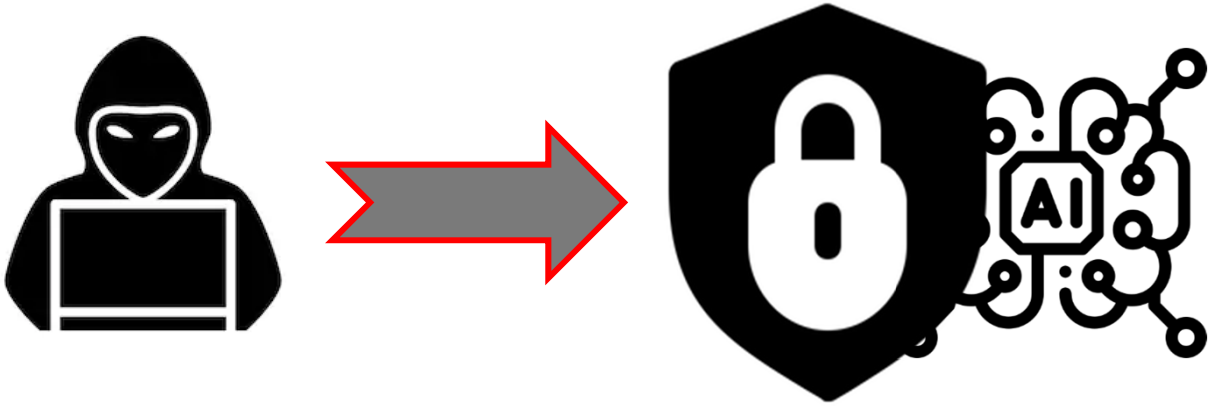
다크트레이스의 EIS (사이버 악성 행위 탐지)



보안 작업 자동화

Security for AI

- AI 시스템에 대한 공격 방어



AI **보안 위협 및 사례**
(Security for AI)

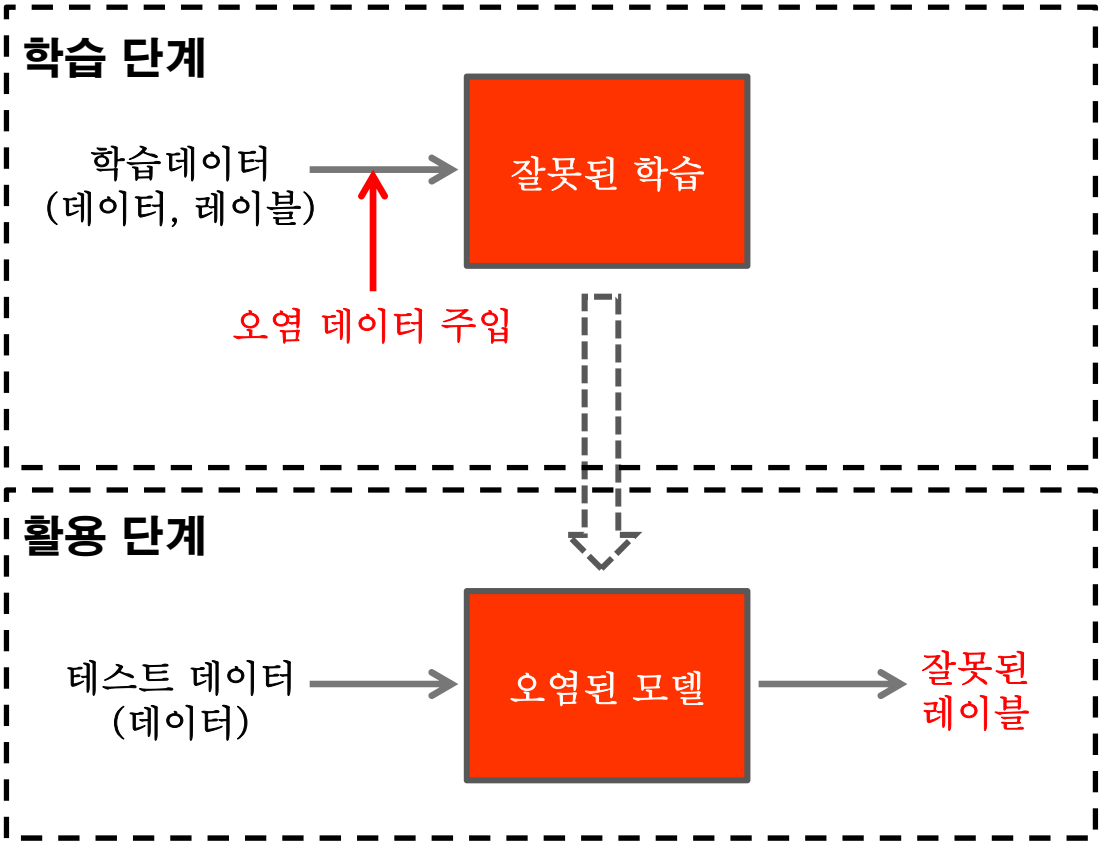
보안 위협

- 학습 단계
  - Poisoning attack
- 활용 단계
  - Evasion attack
  - Model extraction attack
  - Inversion attack

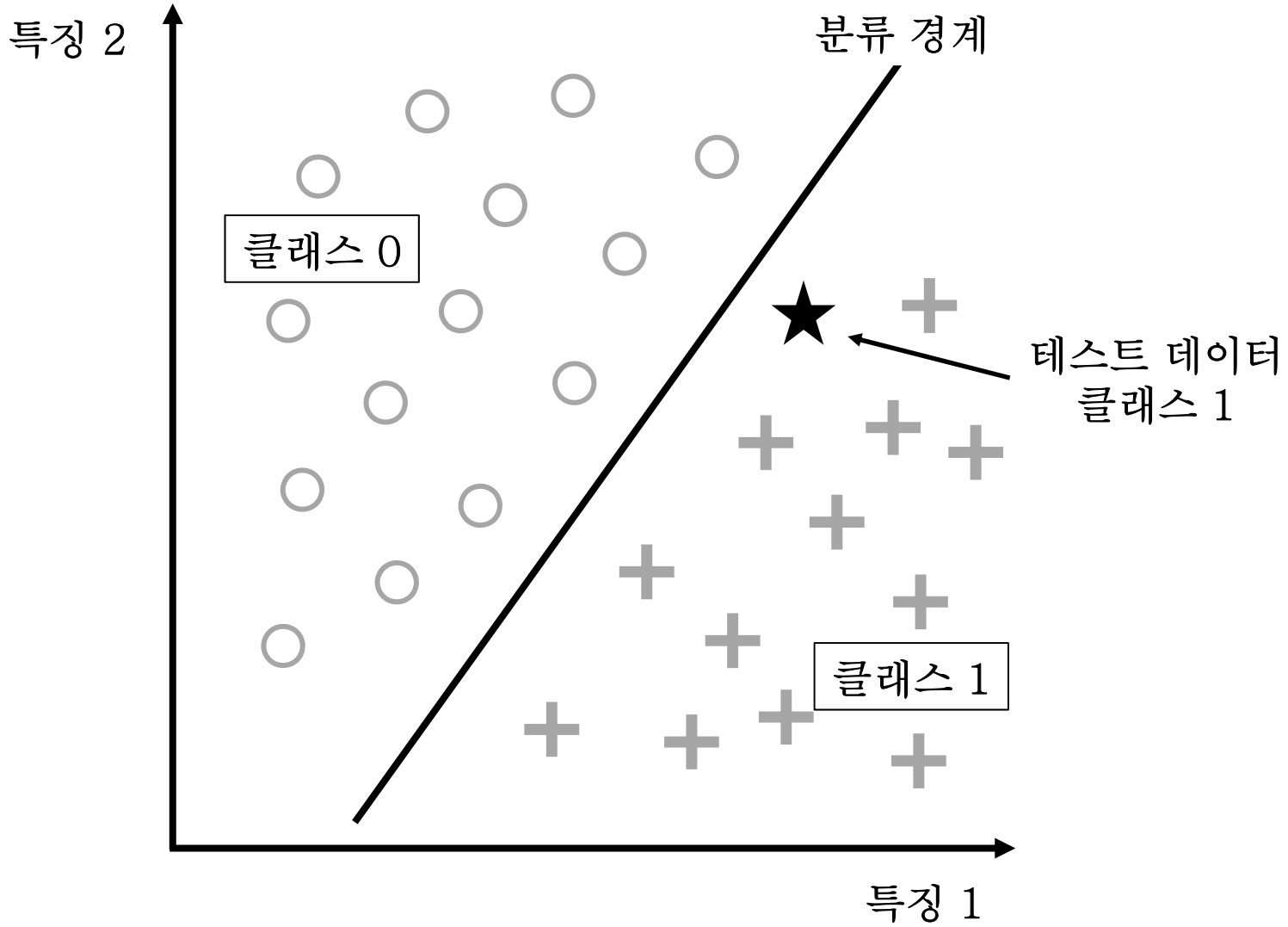
Poisoning attack

Make classifier wrong

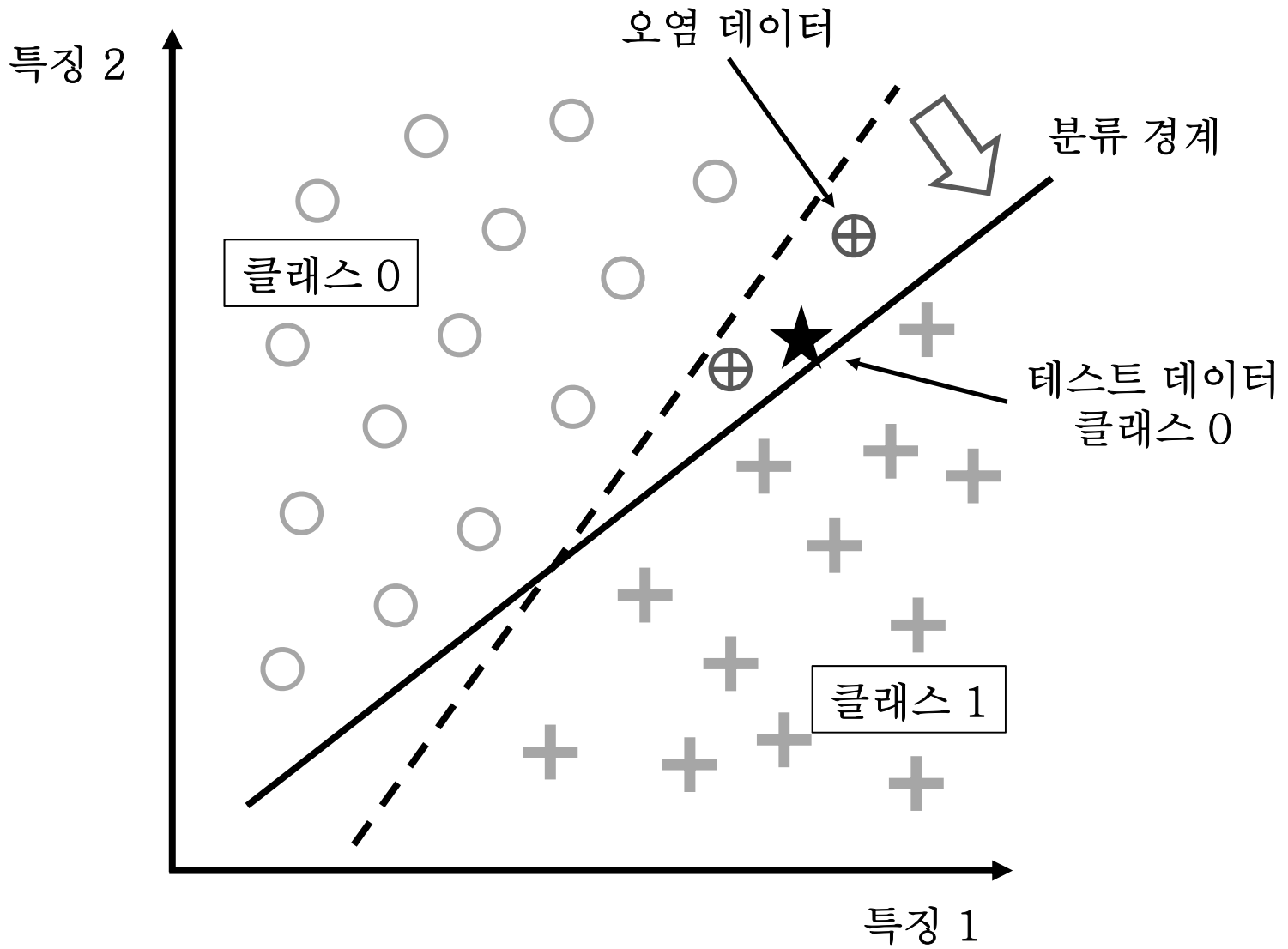
- 잘못된 학습데이터를 주입하여 AI 시스템 오동작 유발



Poisoning attack

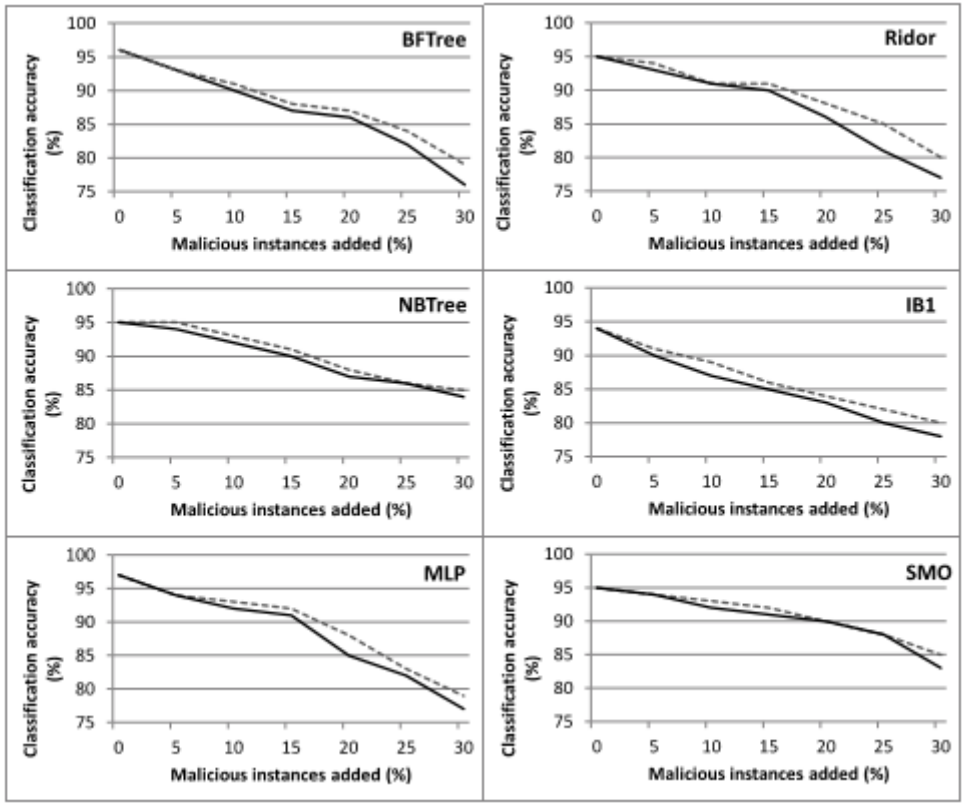


Poisoning attack



Poisoning attack

▶ **최소의 오염 데이터 추가로 오류 최대화**



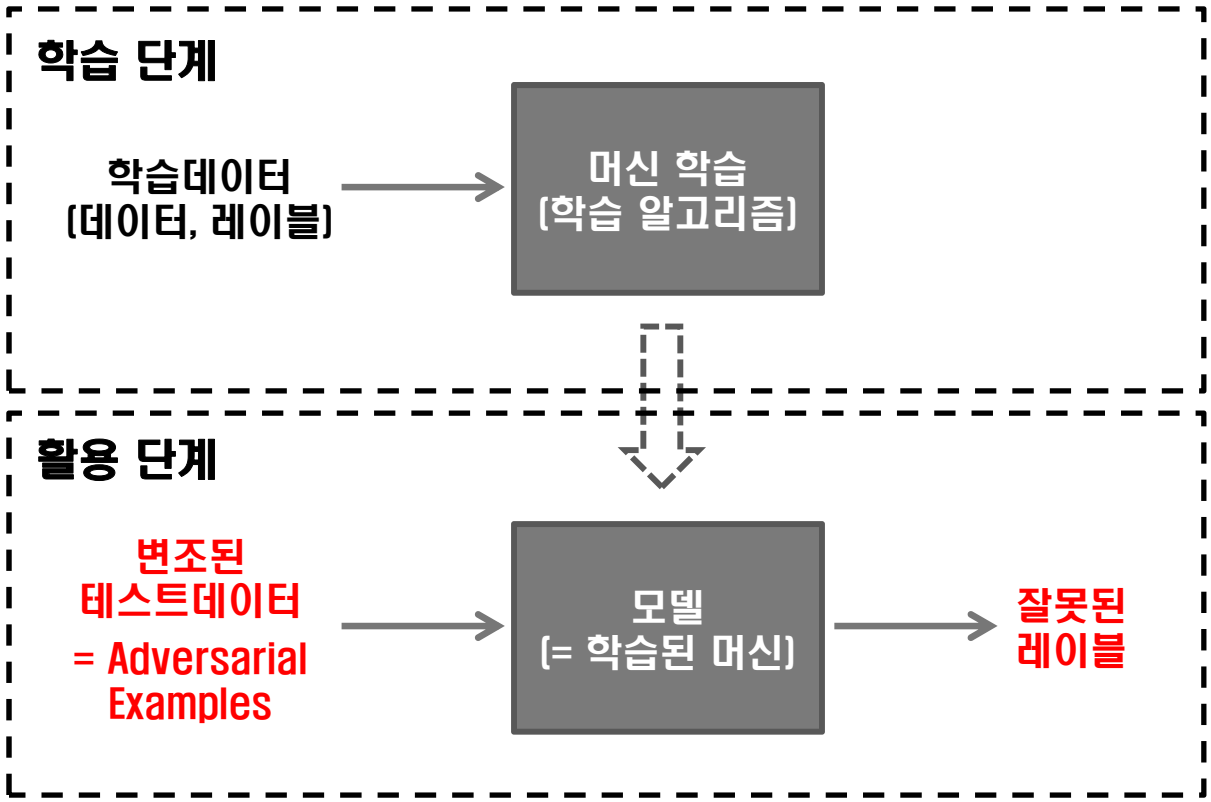
[출처 :M. Mozaffari-Kermani,et.al, Systematic Poisoning Attacks..IEEE Journal of Bio & Health Informatics]

Poisoning attack

- 사례 : AI 나쁜 물들이기
  - In Active Learning

Evasion Attack

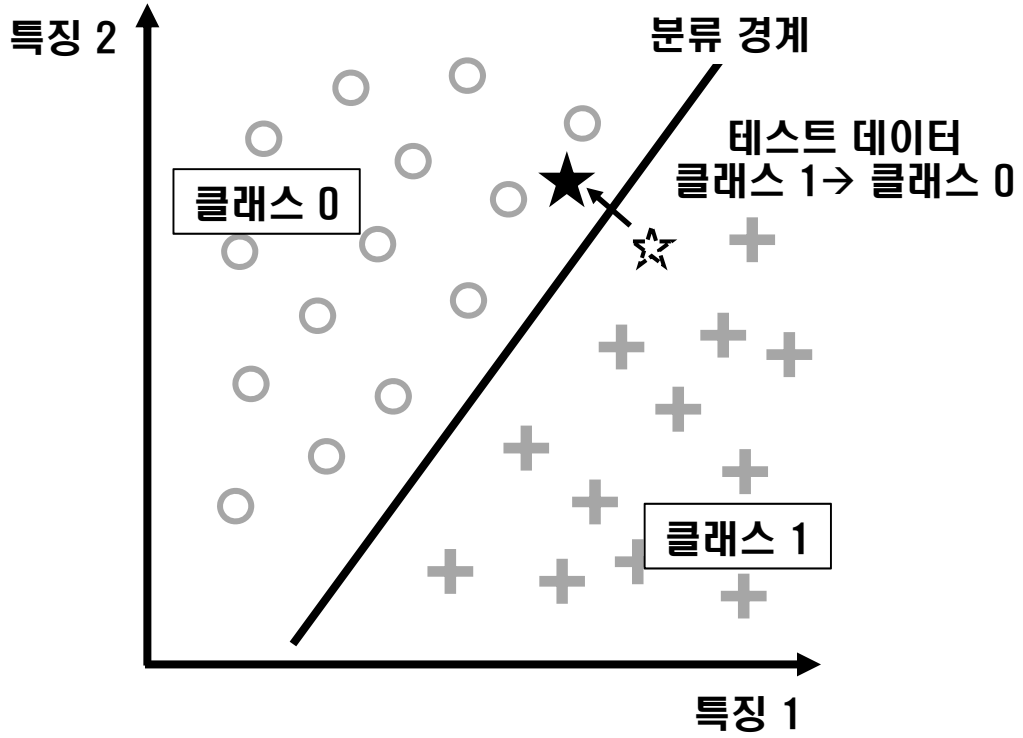
- 기만 공격 (Evasion Attack, Adversarial Examples)
  - 활용 단계의 분류 데이터를 변조 ➔ 오작동 유발



Evasion Attack

목표 / 원리

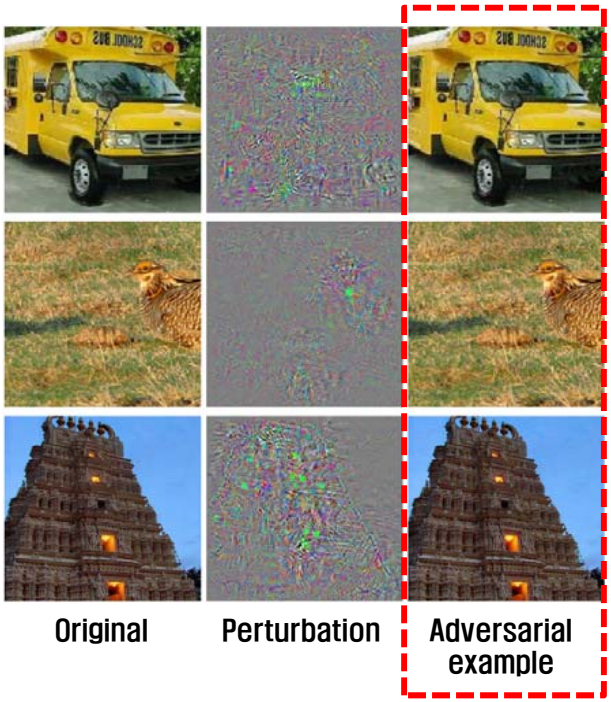
- 최소한의 변조 – 눈에 보이지 않을 정도의 작은 노이즈 추가
- 최대한의 오작동 유발

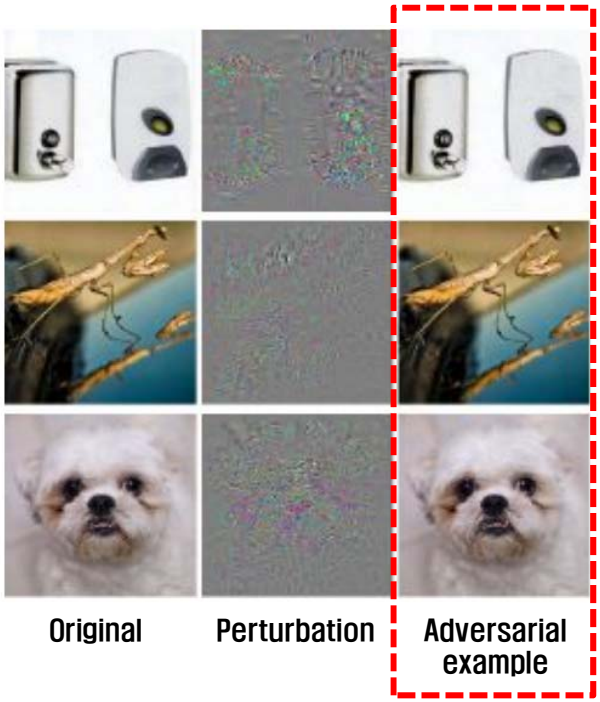


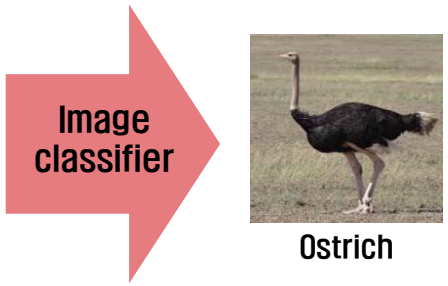
Evasion Attack

▶ 이미지

- 4%만 변조해도.. 97%는 잘못 분류
- 사람은 변조된 이미지를 인식하기 어려움



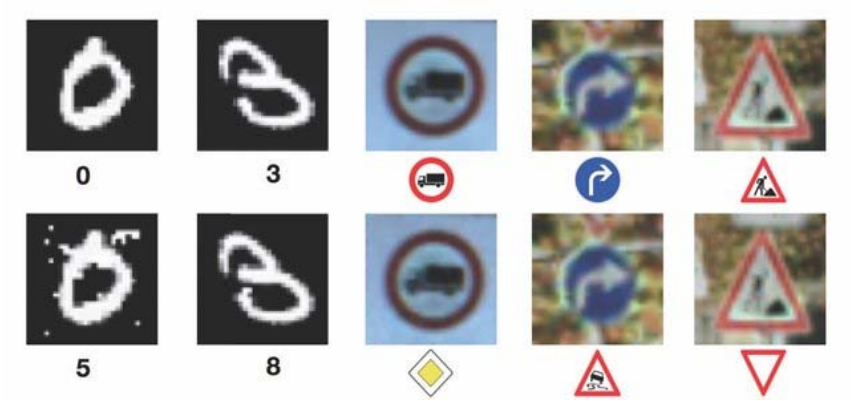




<C. Szegedy et al. "Intriguing properties of neural networks" ICML 2014>

Evasion attack

- **심각하지 않다고요?**
  - ☞ **좌회전 표시를 보고 우회전하는 자율 주행차?**



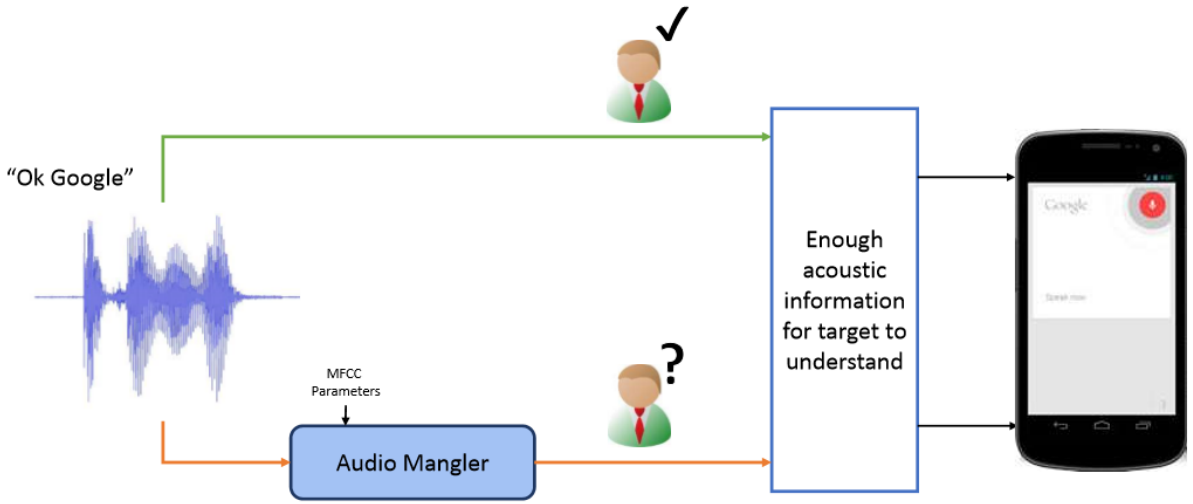
<출처 : PSU, google, WSU, 2016>

테슬라 자율주행차 사고

Evasion attack

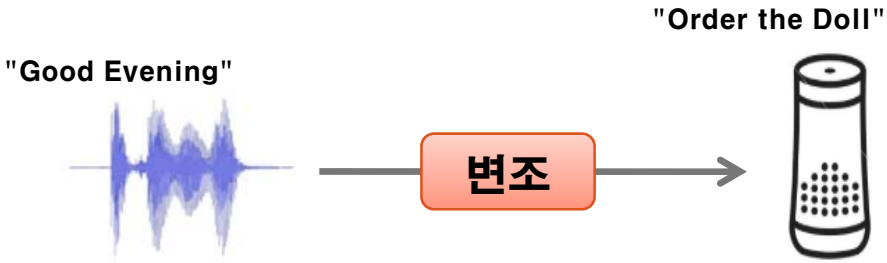
Audio

사람이 알아들을 수 없는 명령어



<출처 : Tavish Vaidya, et. al.Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition, WOOT '15>

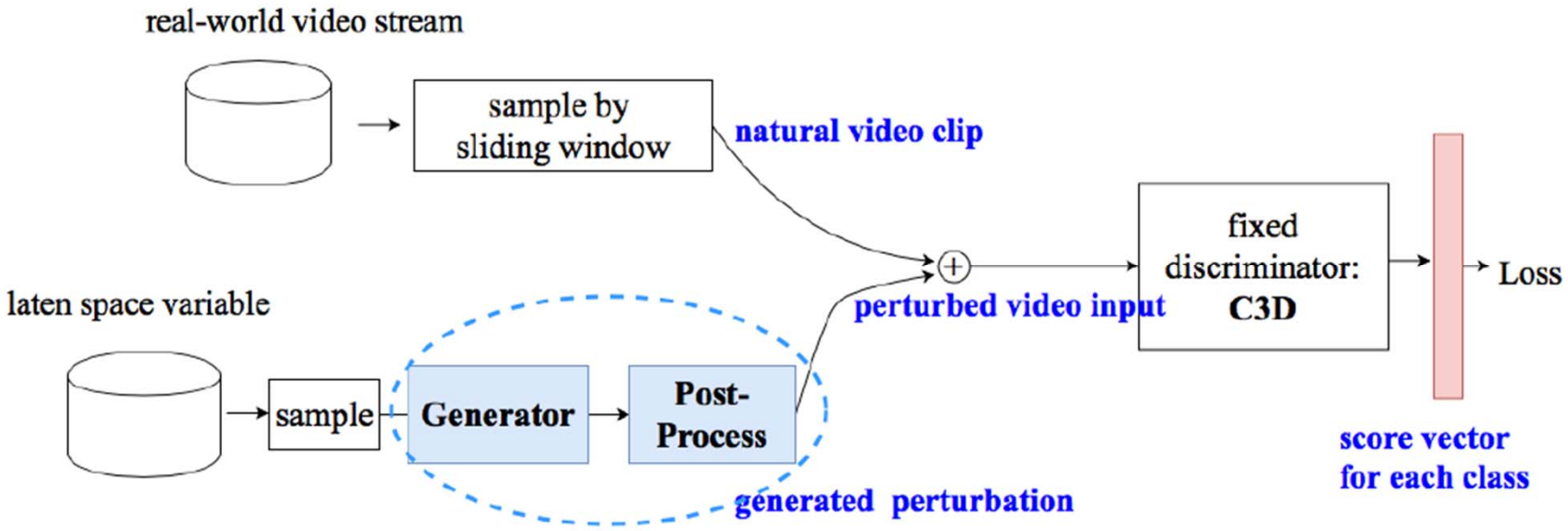
명령어 왜곡



<출처 : Unvi. California , DLS '18>

Evasion attack

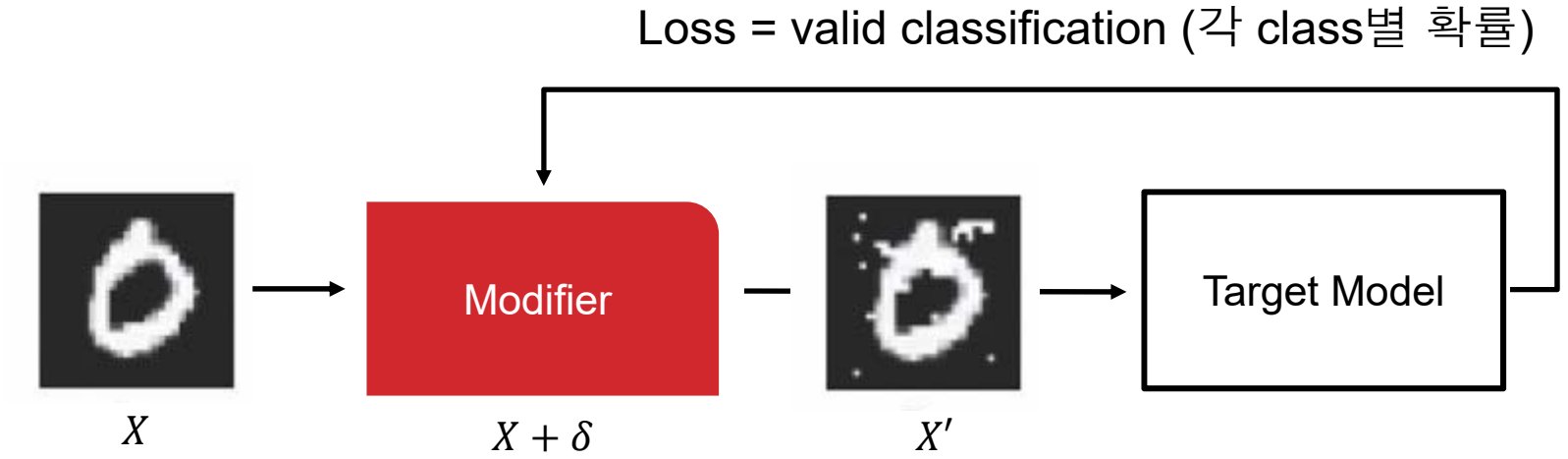
- Video
  - 비디오 clip을 매 프레임마다 약간식 변조하는 방식
  - GAN (Generative Adversarial Netowrk) 활용



Li, Shasha, et al. "Stealthy Adversarial Perturbations Against Real-Time Video *Classification Systems." NDSS 2019*

Adversarial attack 기술

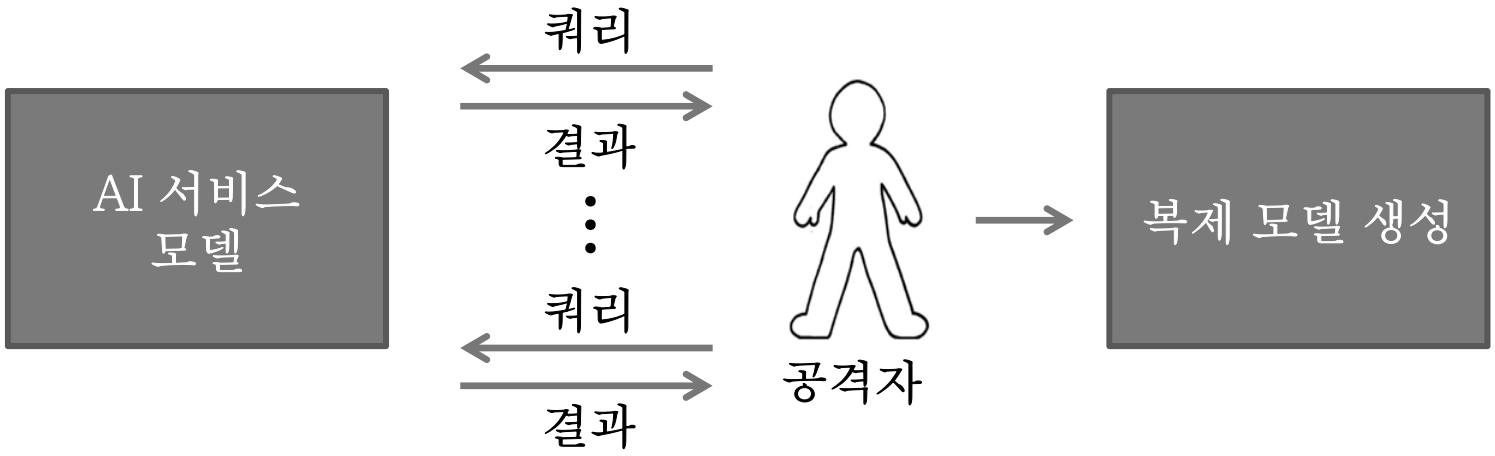
- **Target classifier를 속일 수 있도록 변조**
  - **머신러닝과 동일한 원리**



- **변조율 최소를 목표 : 사람에 의한 탐지를 방지**
  - $\min \delta$ & $\min L$

Model extraction Attack

- 학습된 모델에 쿼리를 해서 타겟 모델 f 에 가까운 f' 만들기 (복제)

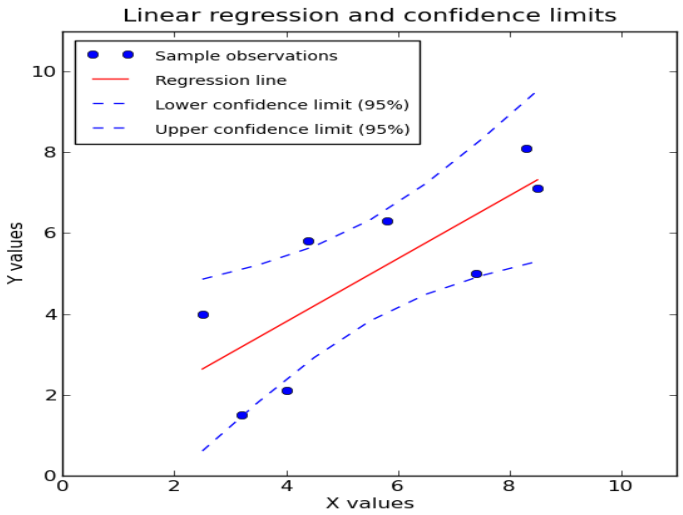


- 공격 목적
  - 유료서비스 모델 탈취
  - ☞ 다른 공격에 활용
    - Inversion attack, Poisoning attack, Evasion attack

Model extraction Attack

- 기존 방법 : Membership Queries
  - 각 데이터가 어느 class에 속하는지 질의?
  - 수많은 query를 반복하면 학습데이터를 얻어 모델링을 새로 할 수 있음

- 최근 방법 : Confidence value의 활용
  - ML produces prediction + confidence
  - Confidence value -> Regression parameter estimation
  - 적은 수의 Query 로도 model extraction이 가능



Model extraction Attack

Performance

- 100% 흉내 내는데 소요된 쿼리# 및 시간

Service	Model Type	Data set	Queries	Time (s)
Amazon	Logistic Regression	Digits	650	70
	Logistic Regression	Adult	1,485	149
BigML	Decision Tree	German Credit	1,150	631
	Decision Tree	Steak Survey	4,013	2,088

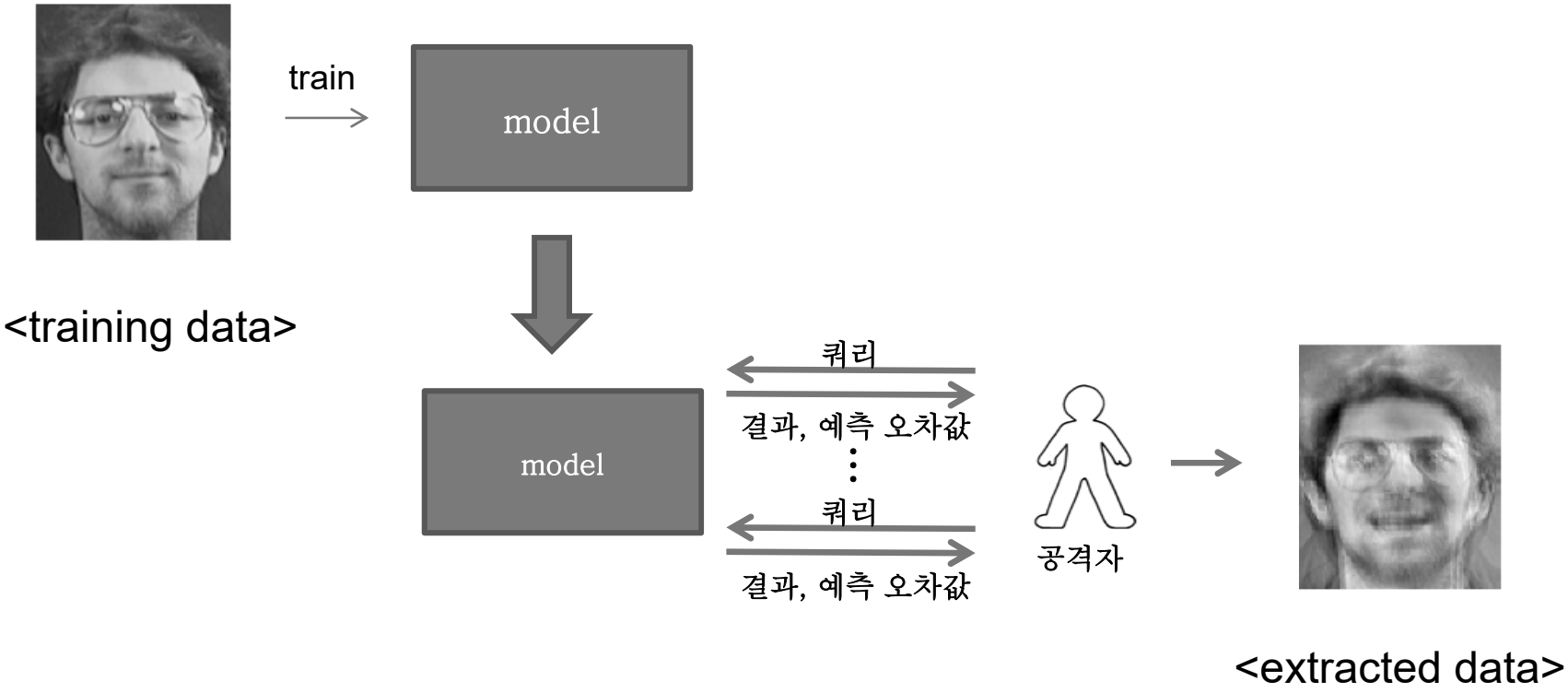
- Decision Tree 재구성 (using incomplete queries)

				Without incomplete queries			With incomplete queries		
Model	Leaves	Unique IDs	Depth	$1-R_{\text{test}}$	$1-R_{\text{unif}}$	Queries	$1-R_{\text{test}}$	$1-R_{\text{unif}}$	Queries
IRS Tax Patterns	318	318	8	100.00%	100.00%	101,057	100.00%	100.00%	29,609
Steak Survey	193	28	17	92.45%	86.40%	3,652	100.00%	100.00%	4,013
GSS Survey	159	113	8	99.98%	99.61%	7,434	100.00%	99.65%	2,752
Email Importance	109	55	17	99.13%	99.90%	12,888	99.81%	99.99%	4,081
Email Spam	219	78	29	87.20%	100.00%	42,324	99.70%	100.00%	21,808
German Credit	26	25	11	100.00%	100.00%	1,722	100.00%	100.00%	1,150
Medical Cover	49	49	11	100.00%	100.00%	5,966	100.00%	100.00%	1,788
Bitcoin Price	155	155	9	100.00%	100.00%	31,956	100.00%	100.00%	7,390

<출처 : Florian Tramer,et.al, Stealing Machine Learning Models via Prediction APIs , Usenix'16>

Inversion attack

- **모델에 질의하여 Training data를 재현해냄**
  - ☞ **학습데이터를 빼낼 수 있음**



<소스 : Matt Fredrikson,et.al, Model Inversion Attacks that Exploit Confidence Information CCS'15>

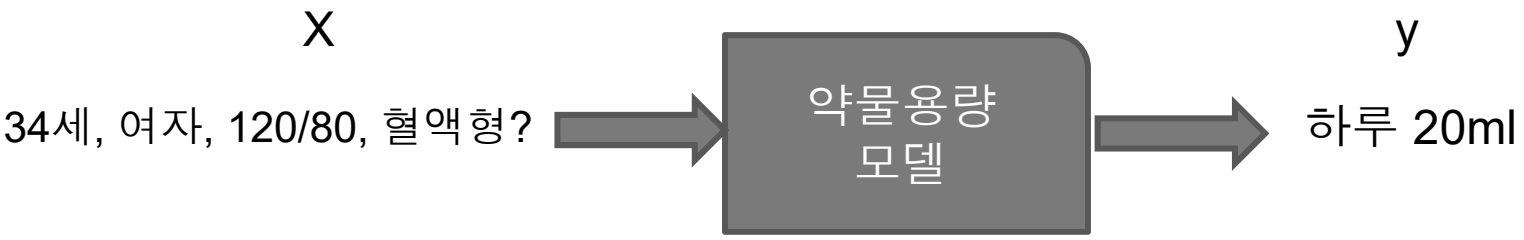
Inversion attack

Nominal feature estimate

- 다른 feature들과 y 를 알 때, 알지 못하는 feature x_1을 model로 부터 알아내는 방법
  - feature x_1 : sensitive information

☞ 알고 싶은 feature의 모든 값을 시험하여 y prediction error가 가장 작은 값을 선택

- 혈액형 외에 다른 것을 알 때..



AI **관련 연구 소개**

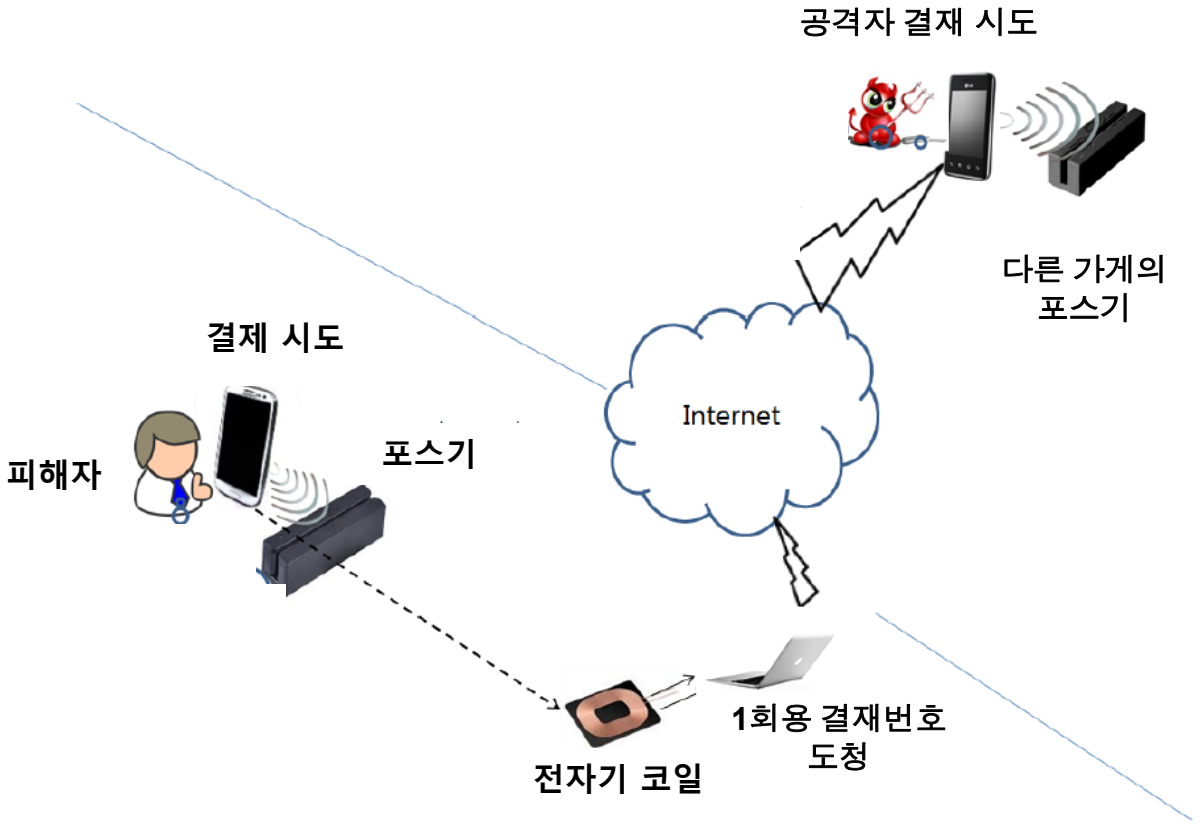
- Interesting
- Better performance

Security by AI – 목소리 인증

- 히든 싱어를 AI가 맞춘다면?
  - AI 모델로 원곡을 학습
  - 각 후보의 노래를 AI 모델이 검증
  - 가장 원곡과 가까운 목소리를 선택
  - AI 정답률 66.33% Vs. 사람 정답률 33.48%

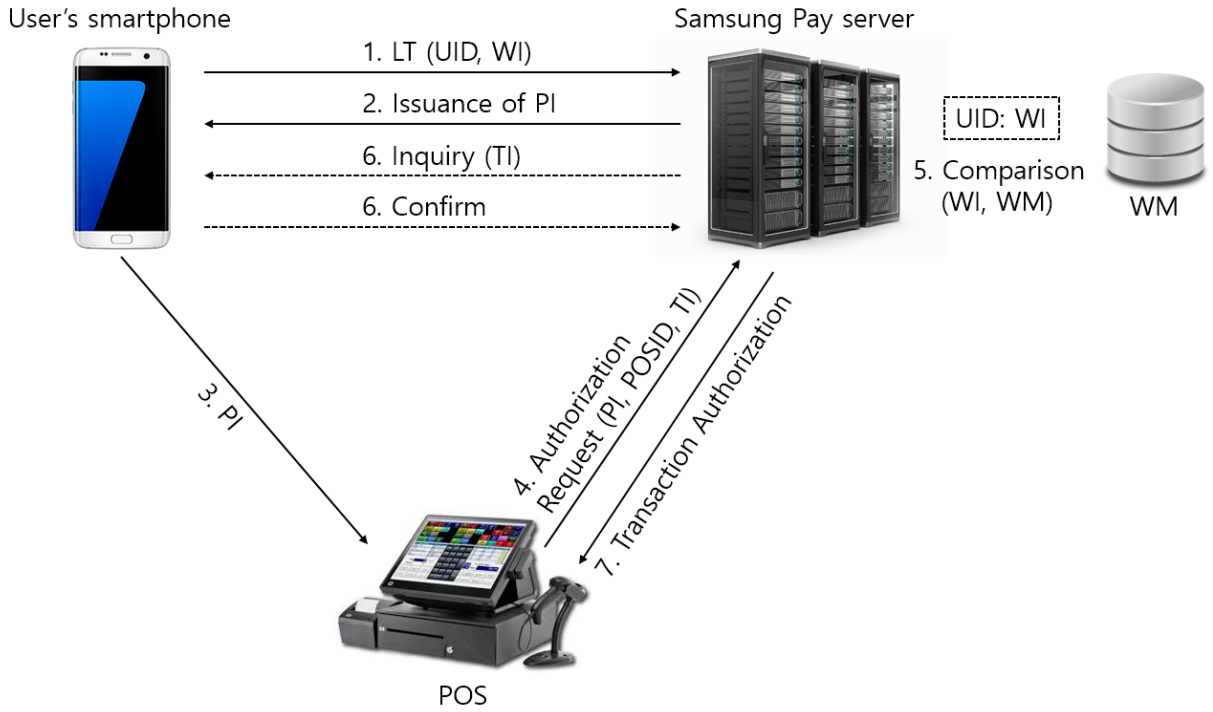
Security by AI – 삼성페이 도용

▶ 삼성 페이 결제 도용



Security by AI – 삼성페이 도용

- 삼성페이 결제 도용 방어
  - AI 모델이 POS의 정보를 학습
    - 무선 신호 특성 정보 (Wi-Fi, Bluetooth, Cellular)
  - 모델 검증을 통한 도용 탐지



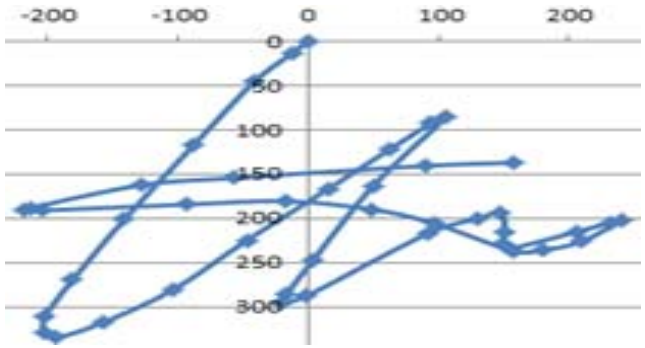
Security by AI – 동적서명 인증

▶ 모바일 서명 인증

Security by AI – 동적서명 인증

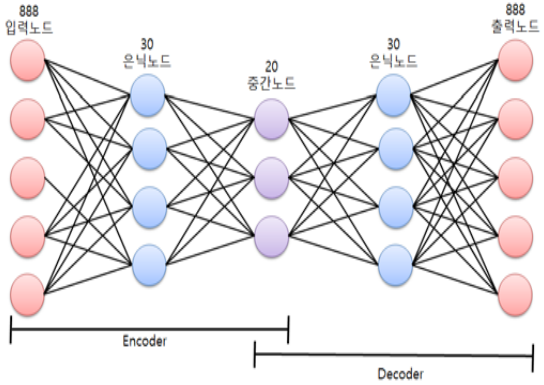
▶ 서명 등록

- 다양한 정보를 획득
  - 좌표,
  - 가속도 센서 정보
  - 서명 데이터 길이
  - 압력



학습

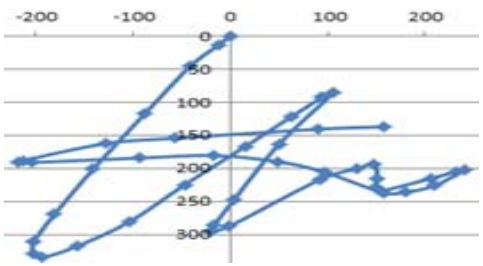
- AI 학습
  - 한 사람의 여러 서명을 학습
  - 모델 생성



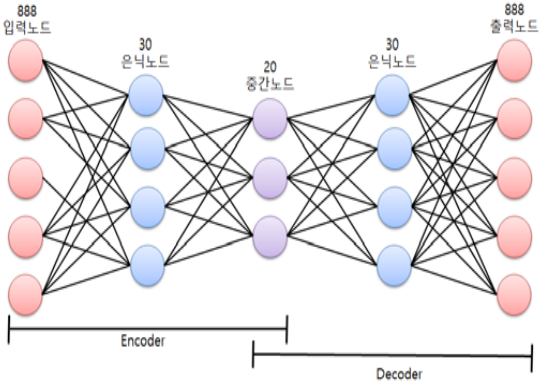
Security by AI – 동적서명 인증

▶ 서명 인증

- 모델에 새로운 서명 입력
- 모델은 입력된 서명의 유사도 검증
- 인증 여부 판단



입력

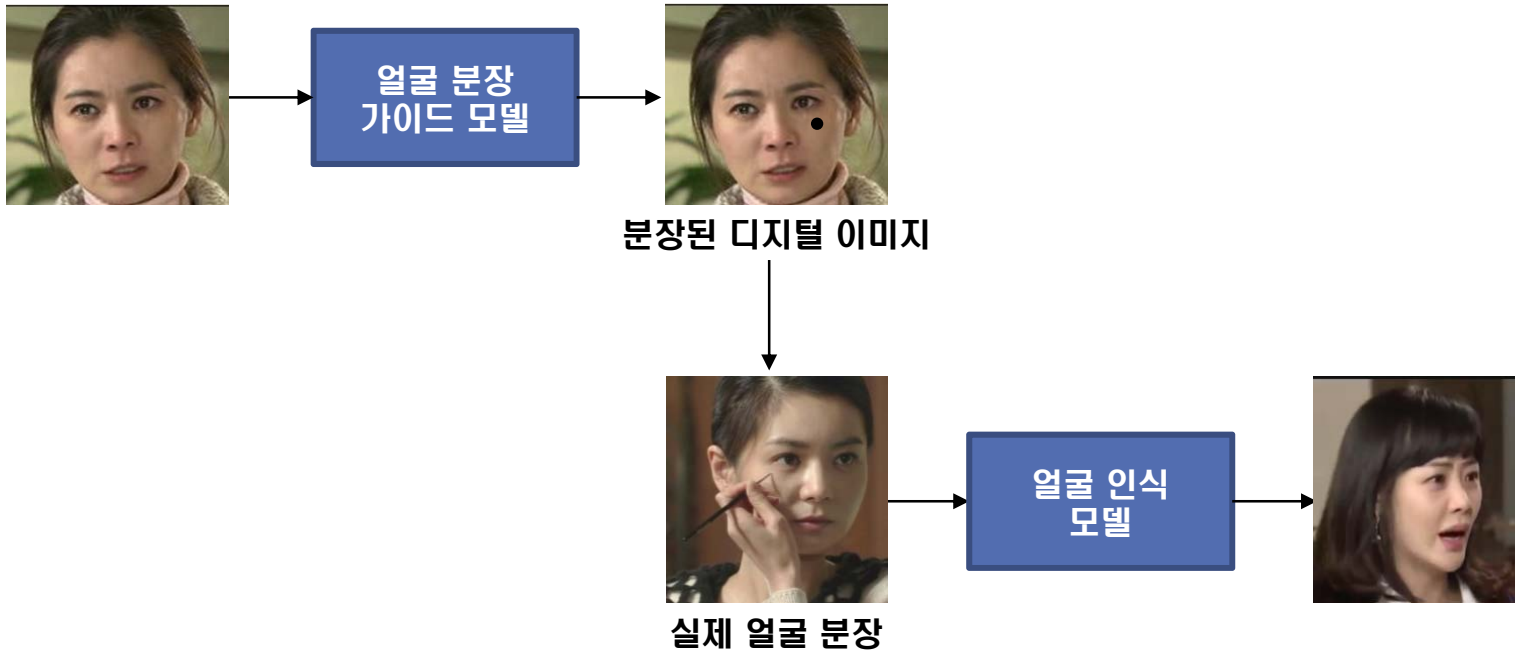


인증 성공

인증 실패

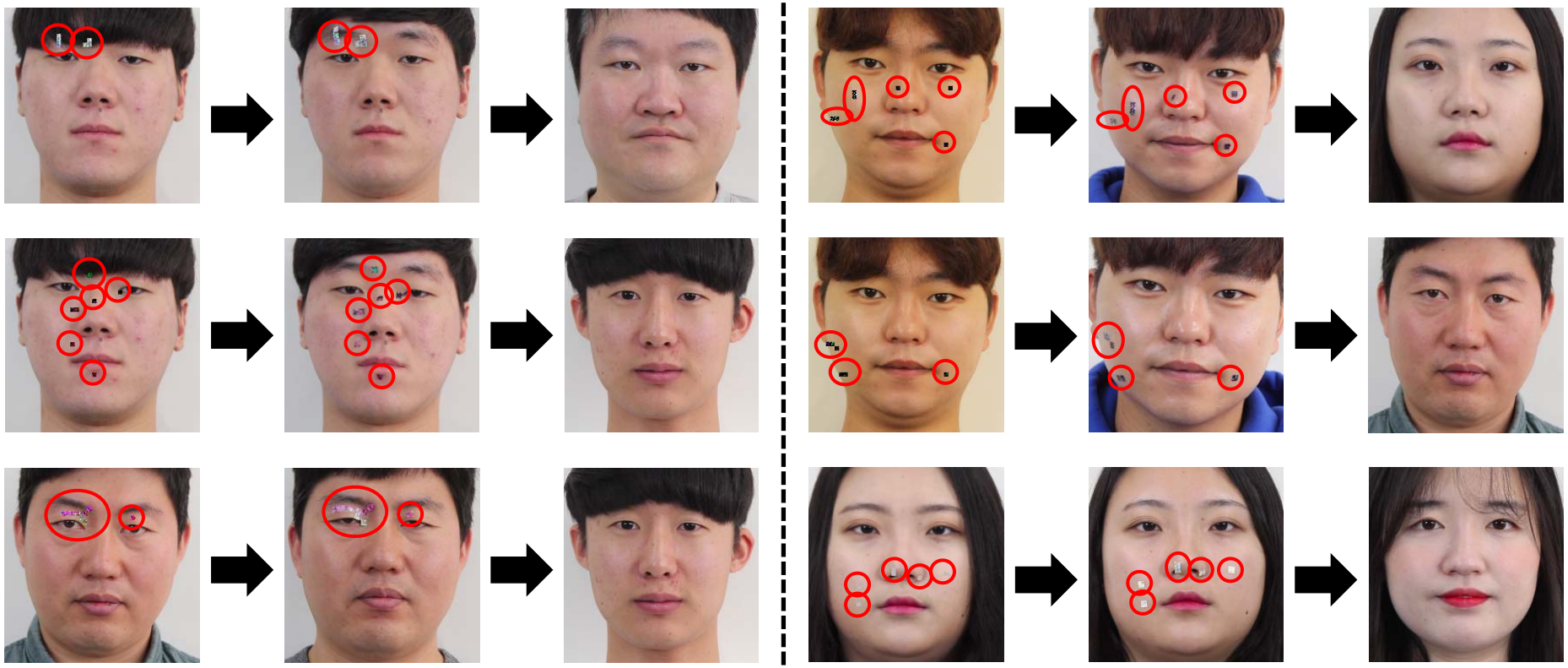
Security for AI – 분장 공격

- **아내의 유혹?**
  - 컴퓨터 이미지가 아닌 실제 얼굴에 분장하여 공격을 해보자



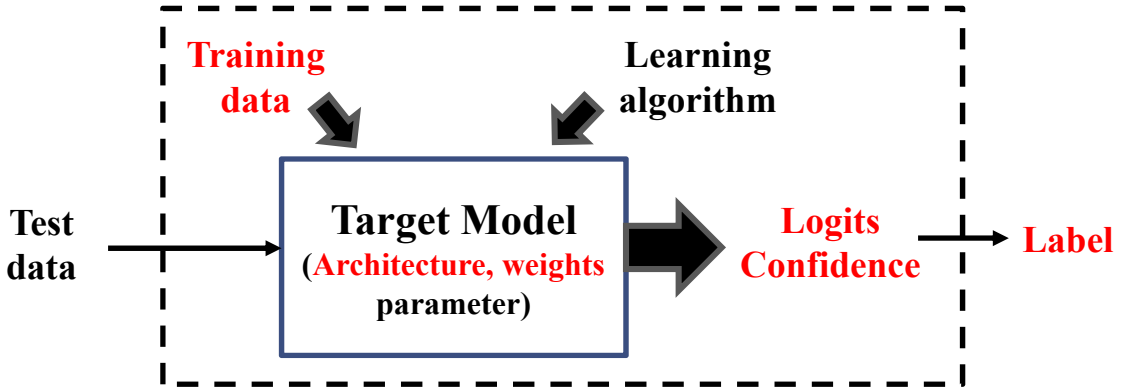
Security for AI – 분장 공격

▶ 공격 결과



Security for AI – 블랙박스 공격

- 화이트 박스 Vs. 블랙박스
  - 공격자가 타겟 모델에 대해 어디까지 알고 있는가?



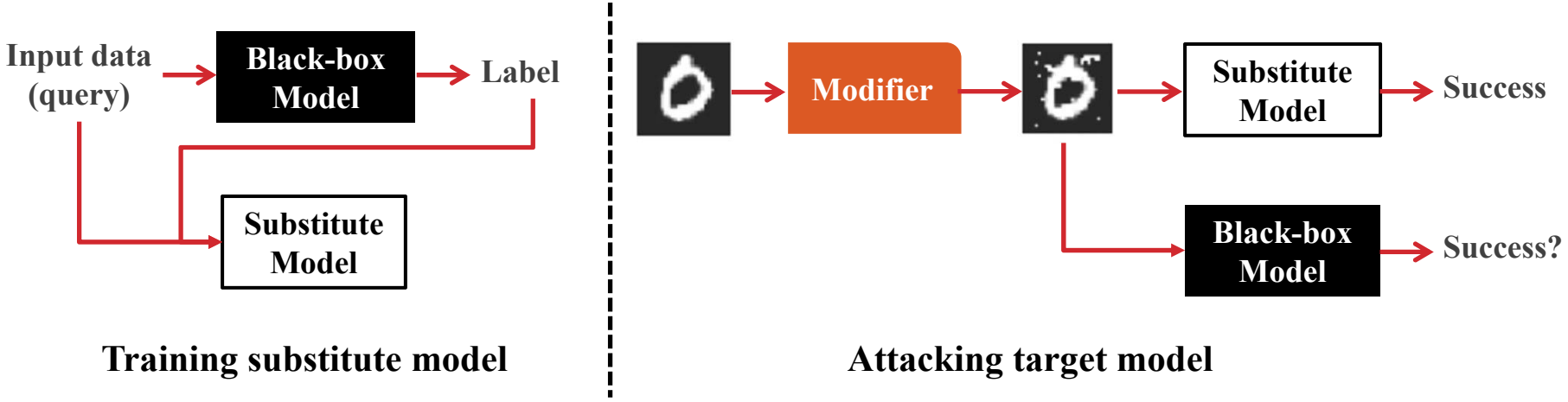
< 화이트박스 >

< 블랙박스 >

Security for AI – 블랙박스 공격

▶ 타겟 모델의 모방 (Substitute Model)

- 학습 데이터 : 쿼리와 **레이블** (타겟 모델의 분류 결과)
- 새 모델 학습 ➔ 대체 모델
- 대체 모델로 adversarial example 생성 ➔ 타겟 모델 공격



Security for AI – 블랙박스 공격

▶ 제안 방안 컨셉 (Retraining substitute model)

- 부분 학습을 통한 쿼리 개수 감소 & 공격 성공률 향상