AI 보안이슈

최대선
지능보안연구실
공주대학교

내용

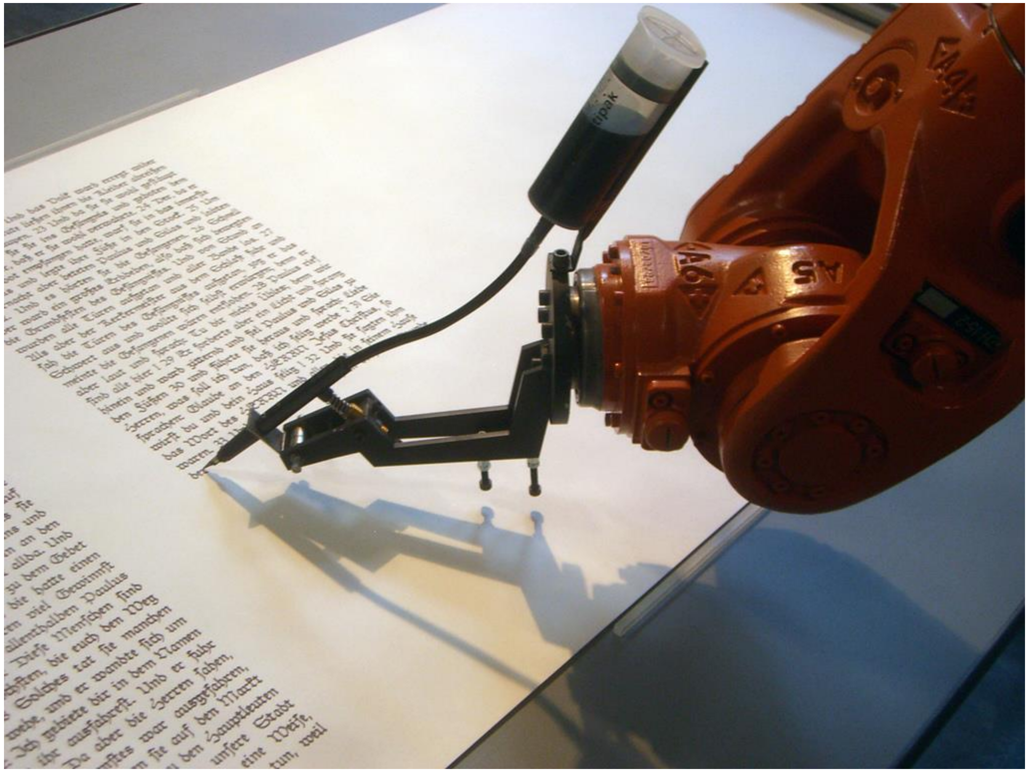
- AI Safety
- Adversarial AI
- Data Privacy in AI

AI SAFETY

종말 시나리오

▶ 고객에서 수기로 감사 편지 쓰는 로봇 Turry

- http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html
- 스티븐호킹, 빌게이츠, 닉보스트롬 등이 걱정하는 시나리오



<출처 : https://www.reddit.com/r/PenmanshipPorn/comments/29l1gs/calligraphy_bot_photographed_by_mirko_tobias/>

AI SAFETY

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

<Concrete Problems in AI Safety, 2016>

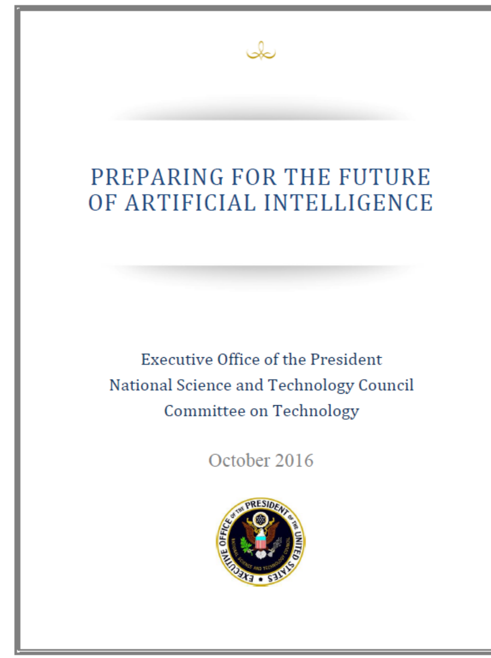
EXECUTIVE OFFICE OF THE PRESIDENT…

국립공주대학교
KONGJU NATIONAL UNIVERSITY

- **AI Safety Engineering**
  - verification and validation
  - how to build a safety case for a technology
  - how to manage risk
  - how to communicate with stakeholders about risk

☞ ***Recommendation 19:***

AI professionals, safety professionals, and their professional societies should work together to continue progress toward a mature field of AI safety engineering



PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE

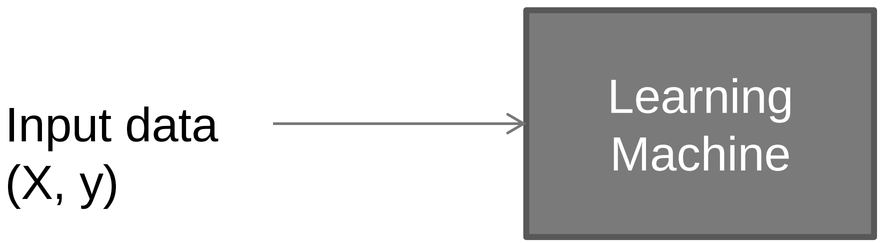
Executive Office of the President
National Science and Technology Council
Committee on Technology

October 2016

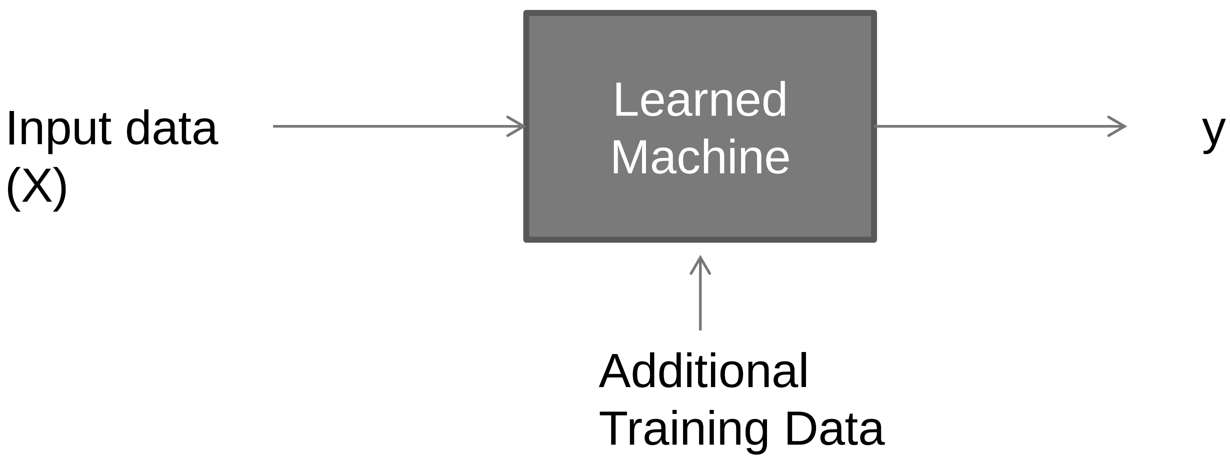
ADVERSARIAL AI

MACHINE LEARNING

Training



Use

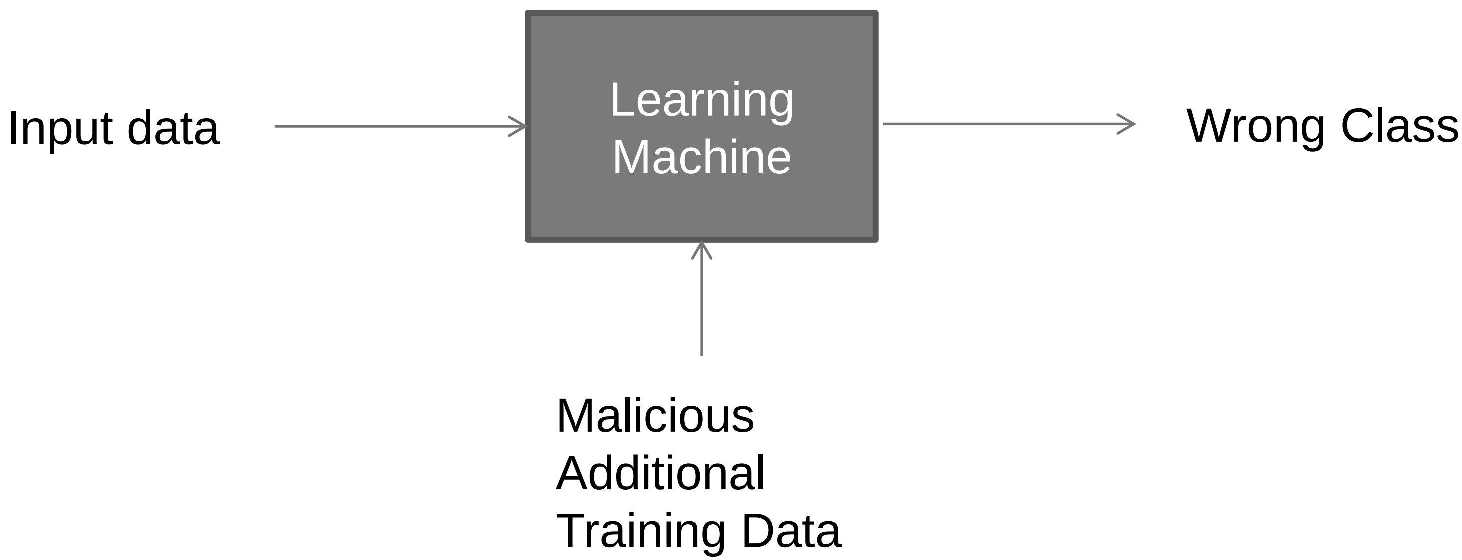


AI POISONING

- AI 나쁜 물들이기
  - In Active Learning

POISONING ATTACK

- **Make classifier wrong**
  - **목표 : 최소한의 데이터로 최대한의 오동작을**

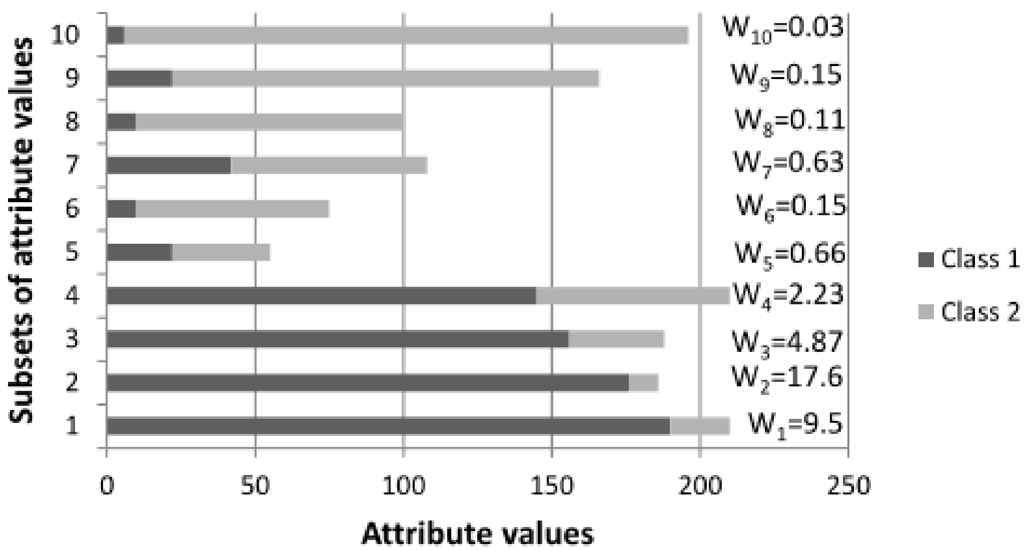


국립공주대학교
KONGJU NATIONAL UNIVERSITY

POISONING ATTACK

- **Make classifier wrong**
  - 목표 : 최소한의 데이터로 최대한의 오동작을
    - Class1을 Class2로 학습시키기 위한 속성값 분포 분석

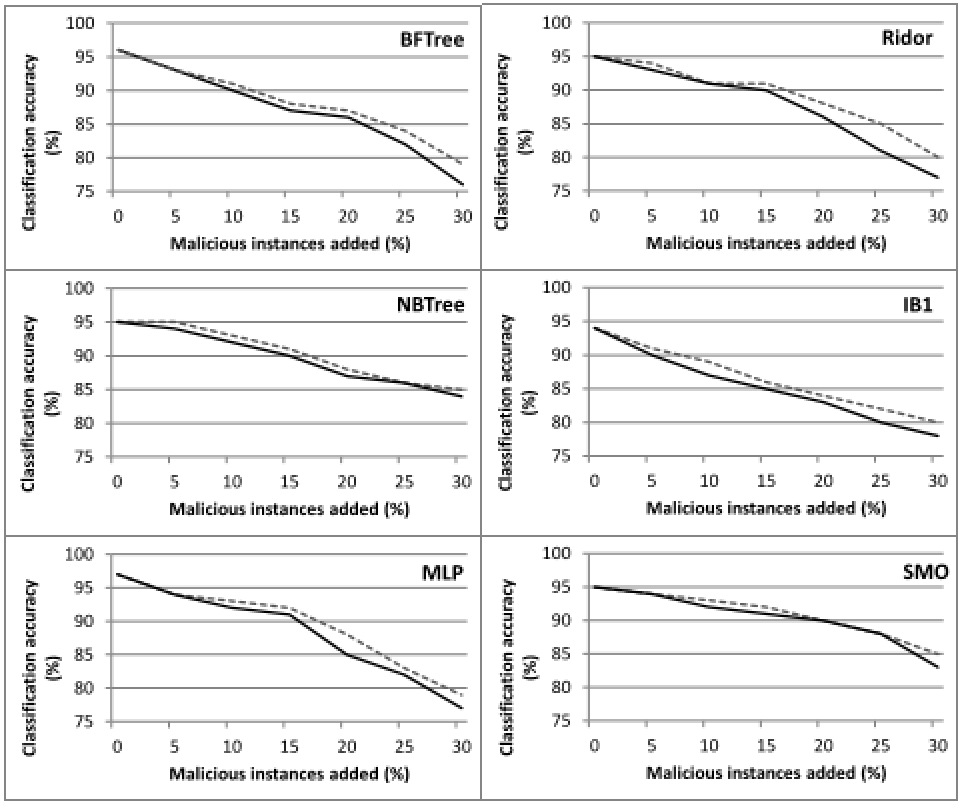
      Class1의 비율이 높은 분포로 학습데이터를 생성 후 class2로 tag



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

POISONING ATTACK

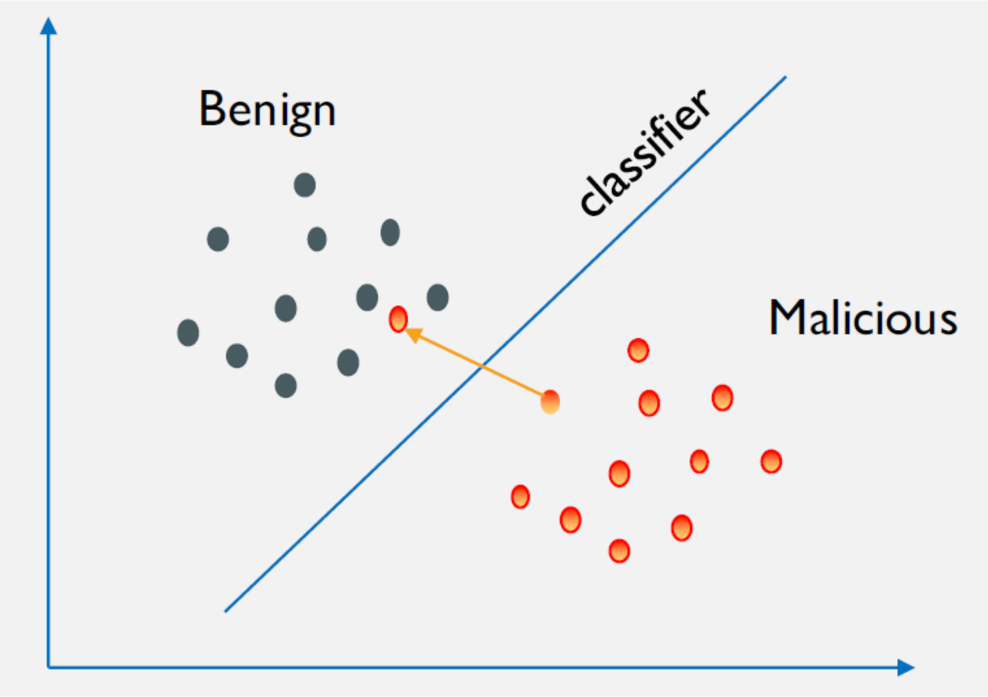
▶ 결과



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

EVASION ATTACK

- 최소한의 변조 (feature substitution)로 다른 class로 인식되게 하는 것



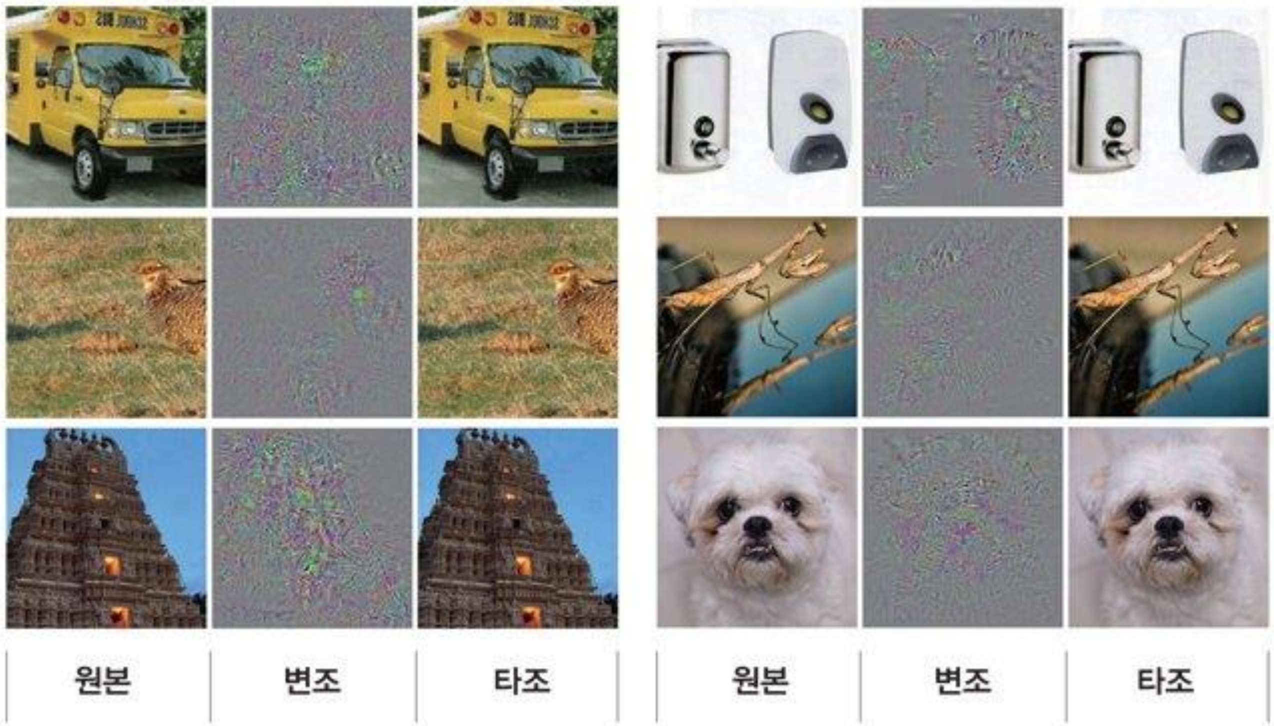
<출처 : Vorobeychik, Florian Tramer,et.al, Adeversarial AI>

min$_{x'}$ $c(x,x')$ s.t.: $c(x,x') \leq$ cost budget, x' classified as benign

EVASION ATTACK

▶ 이미지

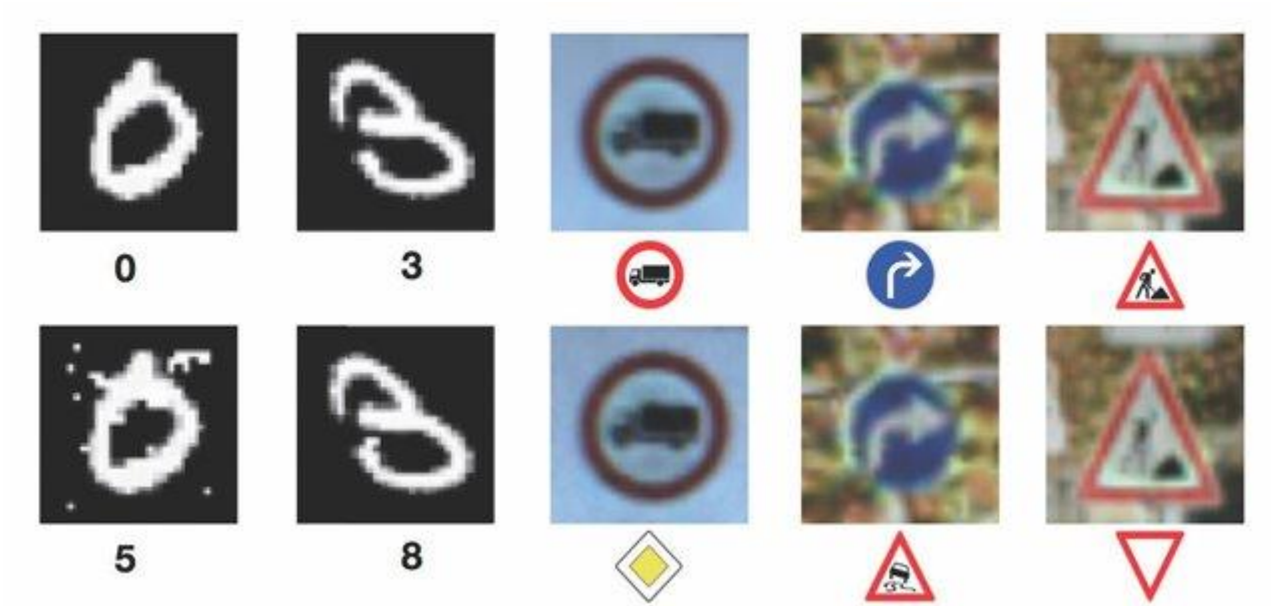
- 4%만 변조해도.. 97%는 잘못 분류



<출처 : 구글-뉴욕대, 2014>

EVASION ATTACK

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

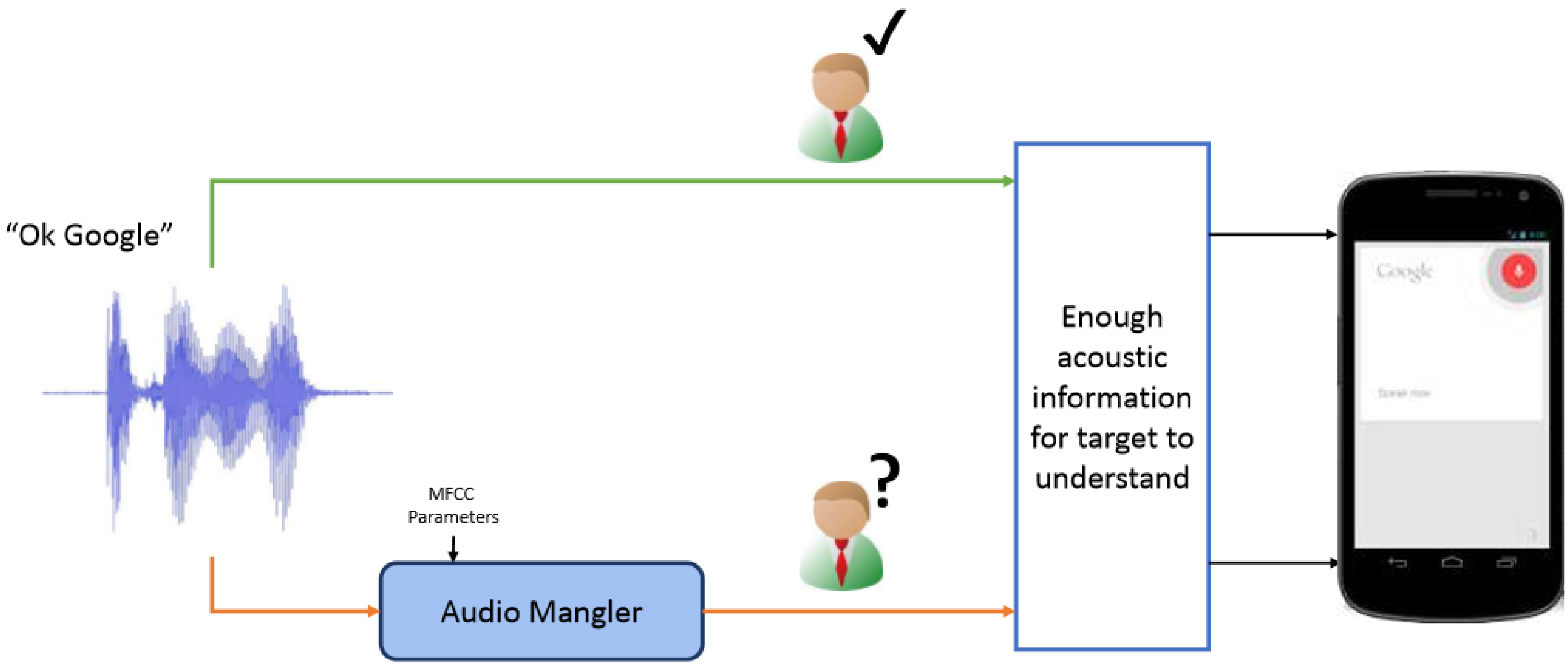


<출처 : PSU, google, WSU, 2016>

EVASION ATTACK

▶ 보이스

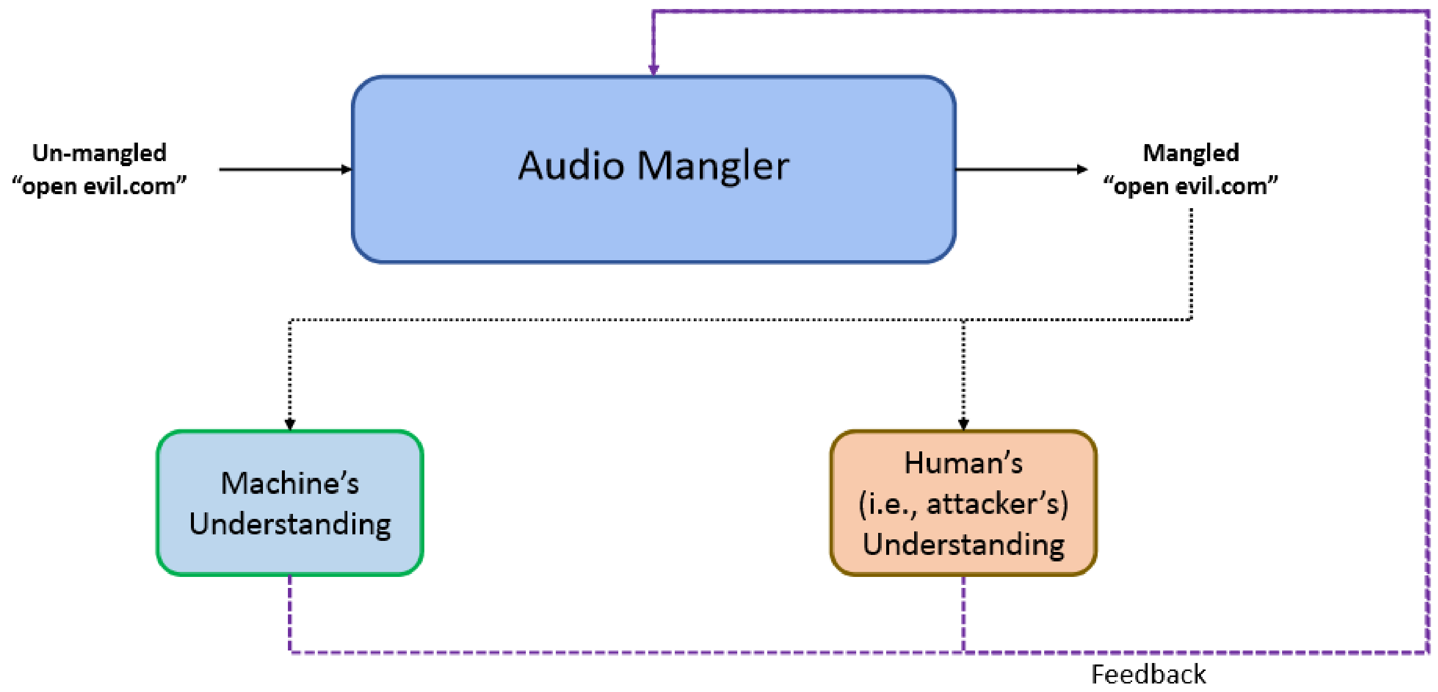
- 사람이 알아들을 수 없는 명령



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

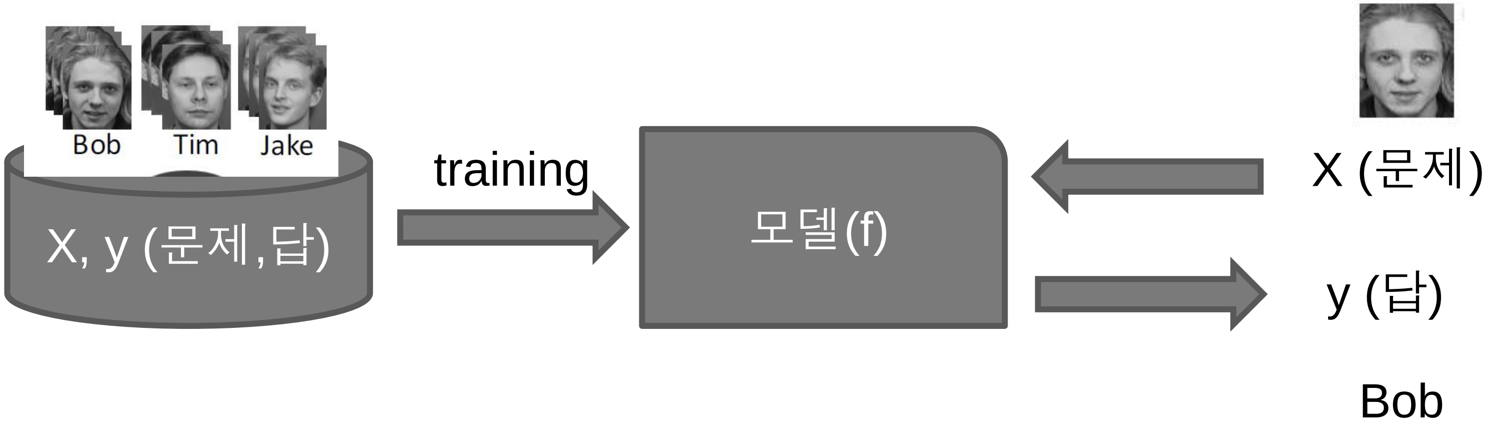
EVASION ATTACK

▶ 공격 데이터 생성



MODEL EXTRACTION ATTACK

▶ ML



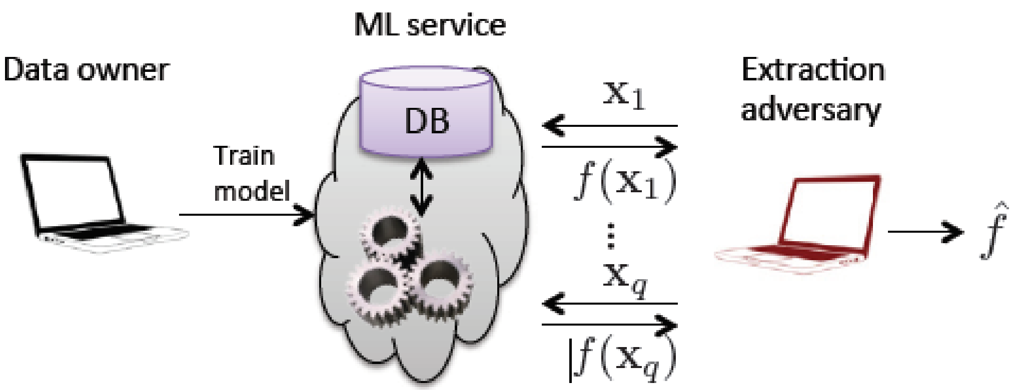
▶ MLaaS (Machine Learning As a Service)

Service	Model types
Amazon	**Logistic regressions**
Google	??? (announced: logistic regressions, decision trees, neural networks, SVMs)
Microsoft	Logistic regressions, decision trees, neural networks, SVMs
PredictionIO	Logistic regressions, decision trees, SVMs (white-box)
BigML	**Logistic regressions, decision trees**

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

MODEL EXTRACTION ATTACK

- 제공된 모델 f 에 가까운 f^ 만들기



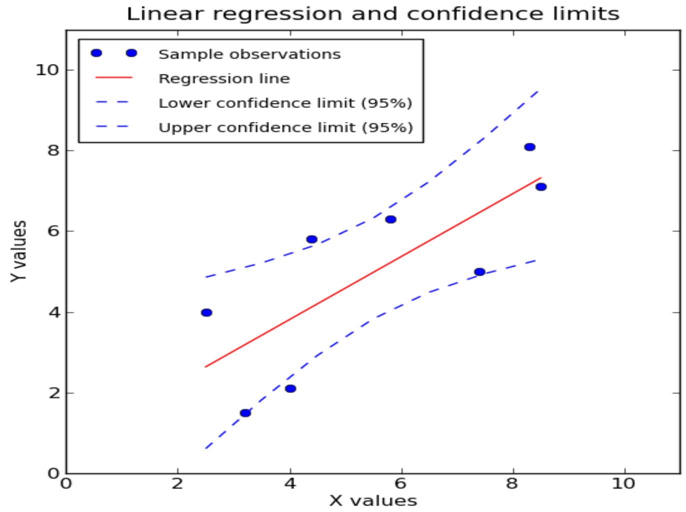
<출처 : Florian Tramer,et.al, Stealing Machine Learning Models via Prediction APIs , Usenix'16>

- 공격 목적
  - 유료서비스 모델 탈취
  - Inversion attack, Poisoning attack, Evasion attack 에 활용

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

국립공주대학교
KONGJU NATIONAL UNIVERSITY

▶ 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)

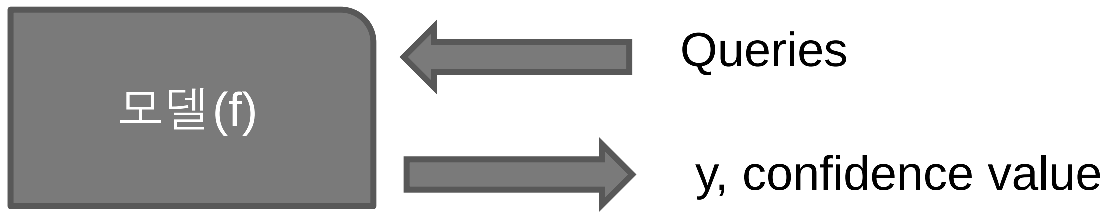
				Without incomplete queries			With incomplete queries		
Model	Leaves	Unique IDs	Depth	$1-R_{test}$	$1-R_{unif}$	Queries	$1-R_{test}$	$1-R_{unif}$	Queries
IRS Tax Patterns	318	318	8	100.00%	100.00%	101,057	100.00%	100.00%	29,609
Steak Survey	193	28	17	92.45%	86.40%	3,652	100.00%	100.00%	4,013
GSS Survey	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>

DATA PRIVACY IN AI

INVERSION ATTACK

- 모델에 query를 하여 Training data를 재현해냄




<extracted face>


<training data>

Strategy	Attack against 1 individual		Attack against all 40 individuals	
	Online Queries	Attack Time	Online Queries	Attack Time
Black-Box Inversion [Fredrikson et al.]	20,600	24 min	×40 800,000	16 hours

<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

ATTACK ON DATA SANITIZING

- 비정형 데이터 sanitizing
  - Medical records (PII)
  - Images/Video (faces)
  - 주로 named entity recognition (머신러닝기반)
  - ☞ 미탐 오류 => 프라이버시 침해

- 공격자는 자동화된 도구로 미탐되어 제거되지 않은 PII를 찾음
  - 공개된 sanitized data에서 PII를 찾도록 training

정리

- AI는 위험할 수 있다.
- AI는 여러 공격에 취약하다
- AI 학습데이터의 프라이버시도 중요하다

감사합니다.

SUNCHOI@KONGJU.AC.KR