'Secure'한 AI 보안 모델 만들기

서준석 (js_seo@heung.me)

2019.06.21

오늘의 주제

첫 번째 질문, 인공지능은 정말 똑똑한가?

두 번째 질문, 인공지능 보안 모델은 과연 Secure 한가?

세 번째 질문, 어떻게 'Secure'한 인공지능 보안 모델을 만들 수 있나?

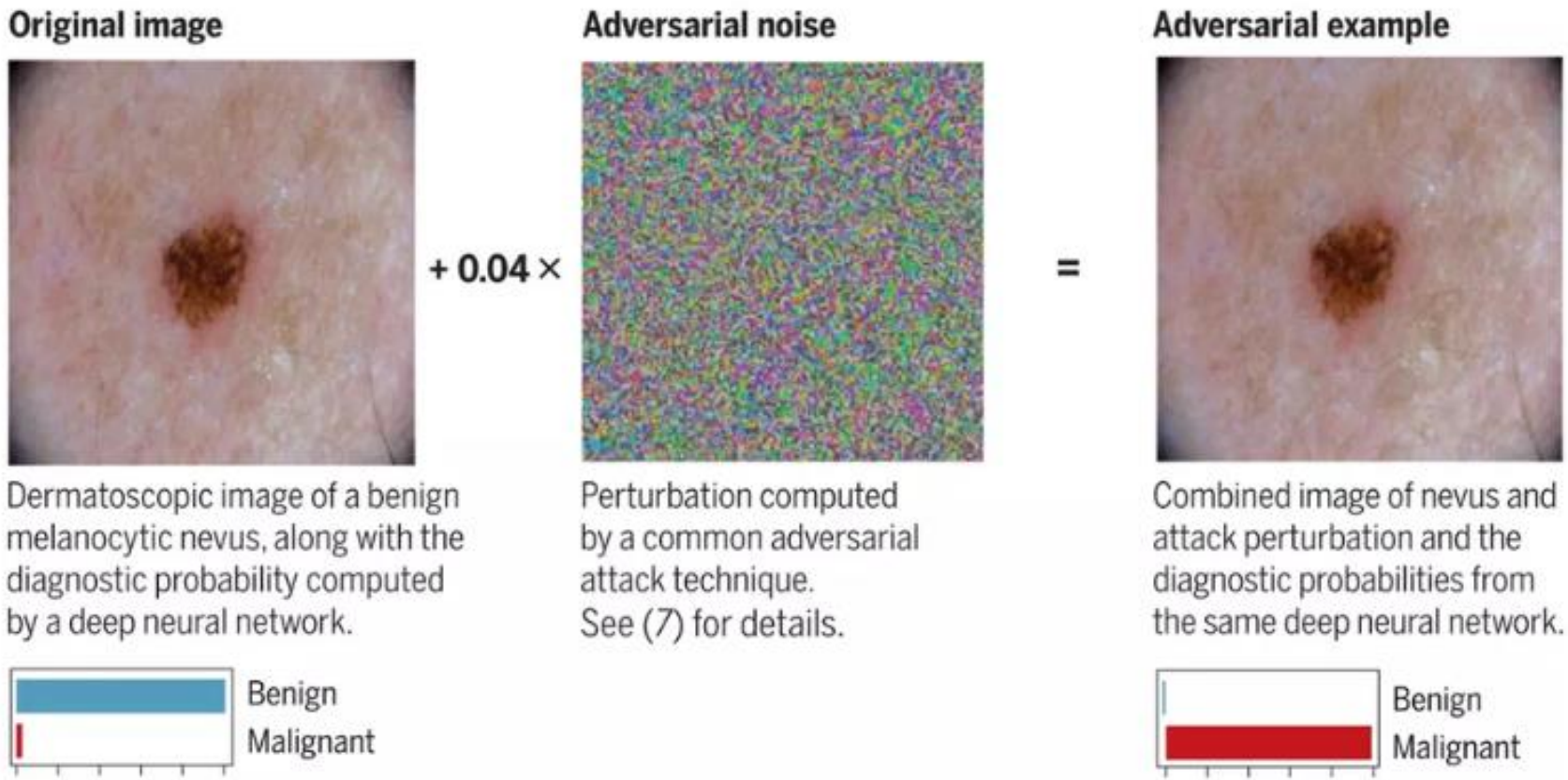
첫 번째 질문

첫 번째 질문, 인공지능은
정말 똑똑한가?

- 확실한 건, 마법은 아니다
- 수많은 전제 조건과 환경적 제약이 있다
- 어쨌든, 시키는 일은 정말 잘 한다
- 하지만, '시키는 일'만 정말 잘 한다
- 바둑, 운전, 작곡, 그림 그리기, 보안 등등

- 자율주행차는 모든 도로를 달릴 수 있을까?


ZU 0666',0,0); DROP DATABASE TABLICE;--



Demonstration of how adversarial attacks could target various medical AI systems *N. Cary / Science*

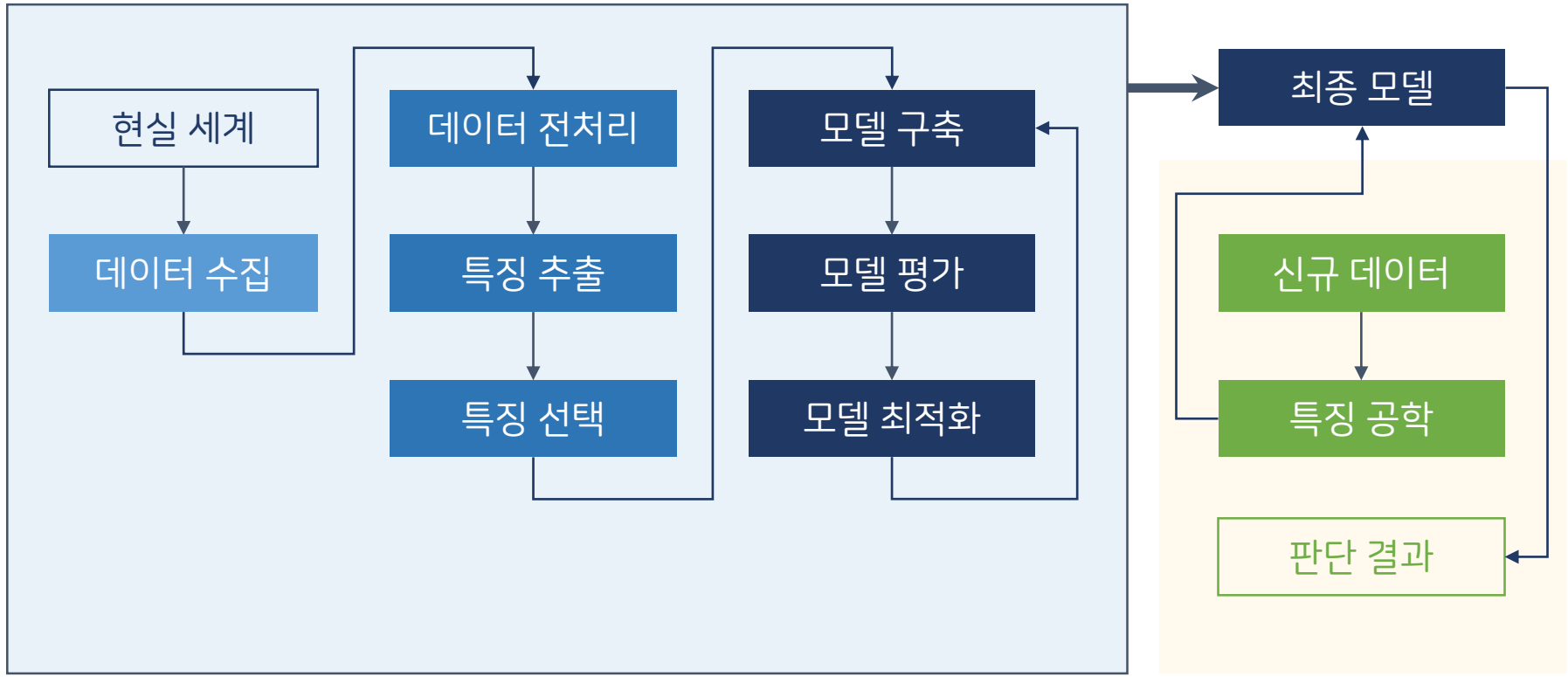
Fig 35. In-car perspective when testing, the red circle marks, the interference markings are marked with red circles

두 번째 질문

두 번째 질문, 인공지능 보안 모델은 정말 Secure한가?

- 기존 패턴/시그니처 기반의 보안 위협 탐지의 한계 극복을 위해 AI 도입
- 목표: 수많은 데이터에서 정상 데이터들의 패턴과 다른 패턴을 보이는 데이터를 찾자
- 사례: 악성코드 탐지, 이상징후 탐지, 네트워크 침입 탐지 시스템

ML 모델이 만들어 지는 과정



보안 시스템을 보안하자

위협 탐지를 위해 구축한 위협 탐지 시스템에 대한 위협은 없을까?

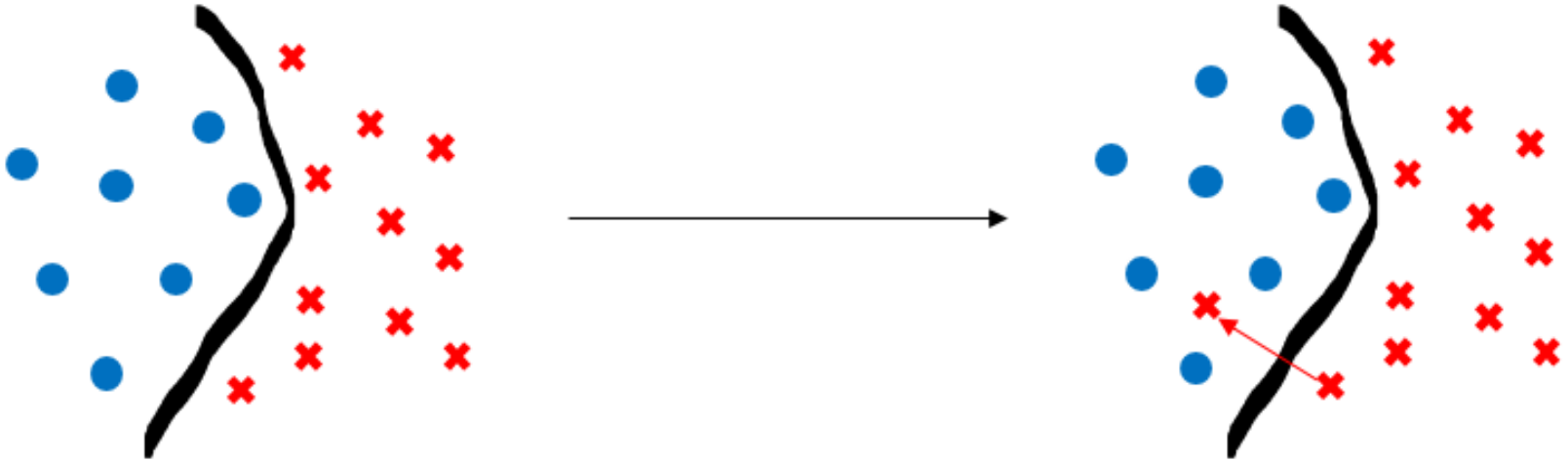
위협 탐지 기능을
무력화 하는
공격 샘플 생성

OR

위협 탐지 시스템을
대상으로 하는
해킹

위협 탐지 기능을 무력화 하는 공격 샘플 생성

위협 탐지의 기준이 되는 '선'을 넘나드는 샘플을 만들 수 있다면?



2017년부터 본격적인 연구가 되고 있음 - 비교적 신생 분야

우선 사례부터

통계분석 기반 Adversarial ML로 악성코드 탐지 시스템 우회

EvadeML: ML 기반 악성 PDF 탐지를 방해하는 모델

AdversariaLib: sklearn와 뉴럴 네트워크 판단 능력을 공격하는 모델

Bot vs. Bot: 강화학습 기반 악성코드 탐지 우회 모델(Blackhat 2017)

Pwning Deep Learning Systems: 딥러닝 모델 공격(이미지 기반)

Gradient-based attacks: perturbation or GAN

……

Adversarial ML

적대적인 머신러닝(Adversarial Machine Learning)

> 머신러닝과 정보 보안 영역의 교차점

> 모델의 보안성 위반과 학습성 저해를 목표로 하는 머신러닝 모델

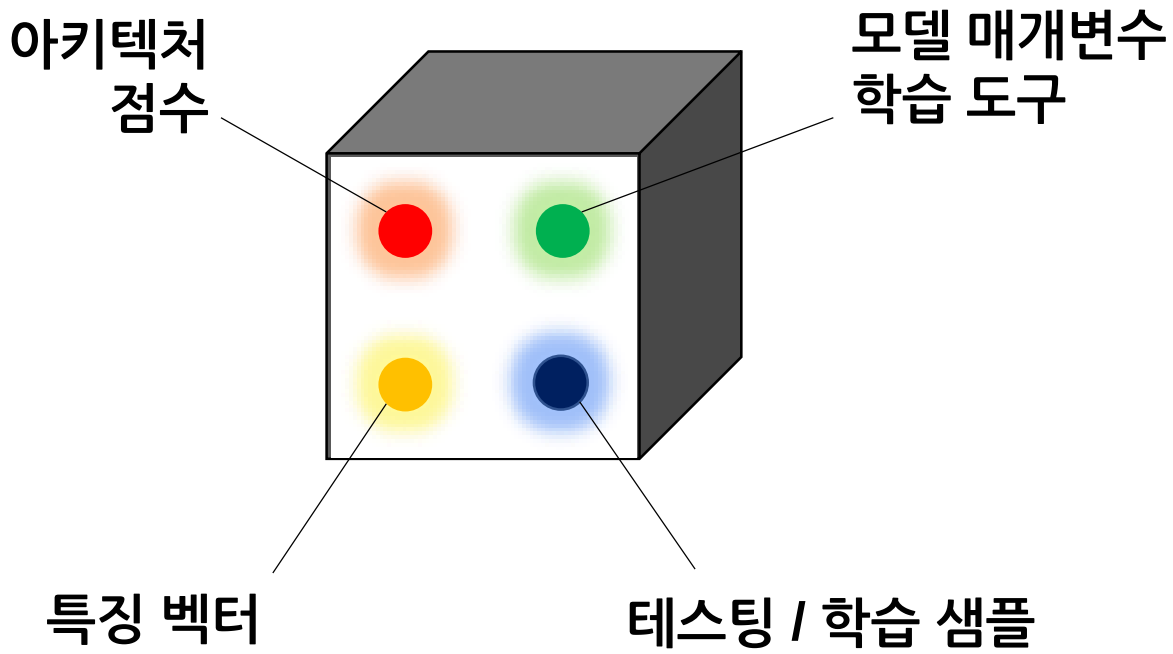
> 탐지 모델의 눈을 멀게 하는 샘플 생성

> 탐지 모델의 무결성과 가용성을 깨뜨리는 공격

> 하지만 어떻게?

Adversarial 조건

탐지 모델을 흐리게 만드는 샘플을 만들려면?



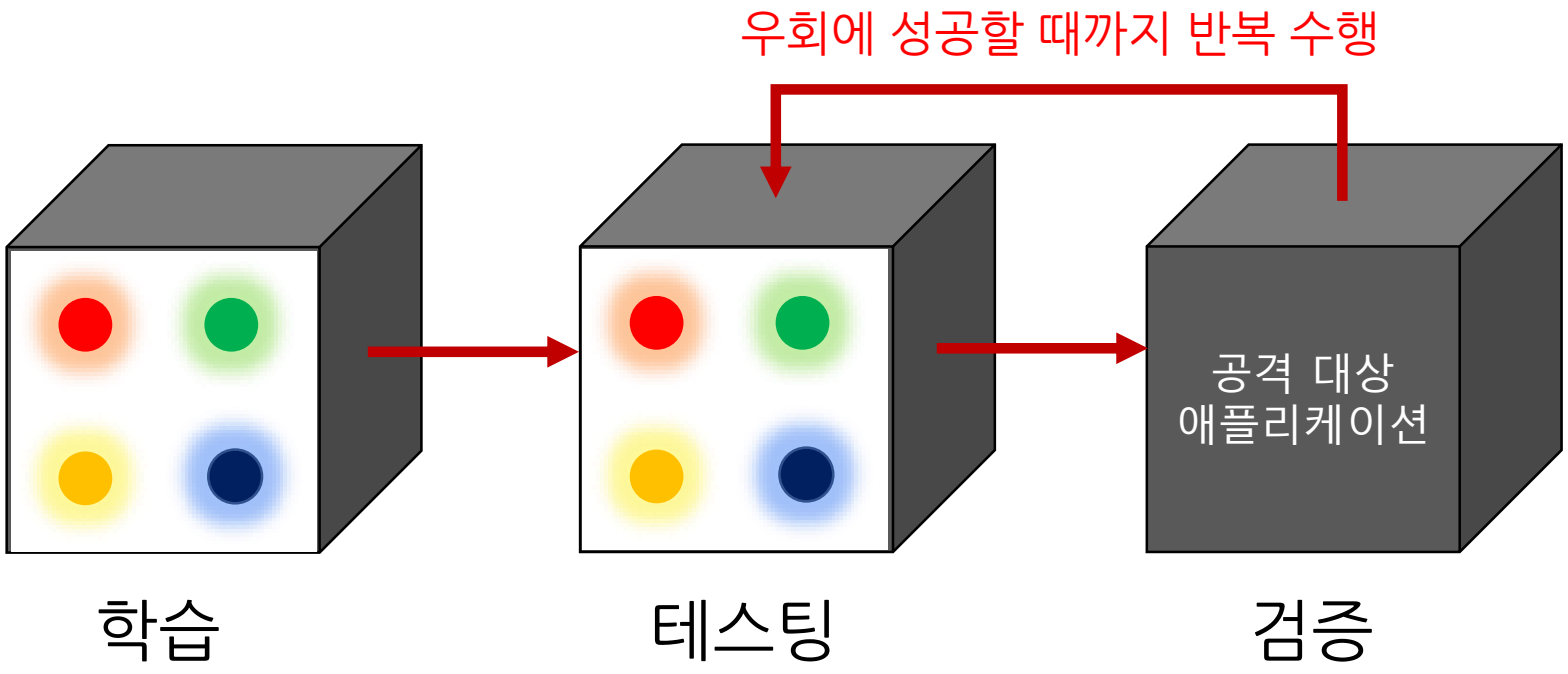
실제 환경에서는

이들 중 어떠한 항목도

구할 수 없다!

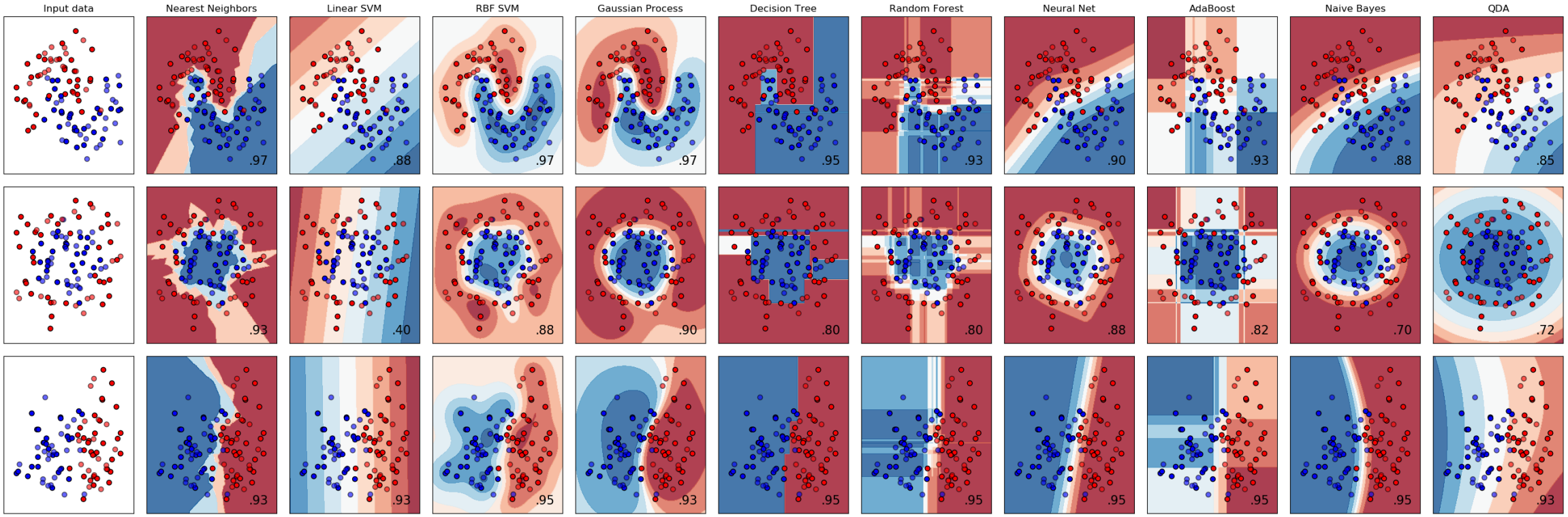
Adversarial 조건

공격자가 '4가지 핵심 요소'를 가지고 있다고 가정
- 최대한 많은 특징, 매개변수, 모델을 확보하고 테스트



Adversarial 조건

내가 만든 모델을 사용해 공격 샘플을 만들어도 유효한 이유



Adversarial 조건

내가 만든 모델을 사용해 공격 샘플을 만들어도 유효한 이유

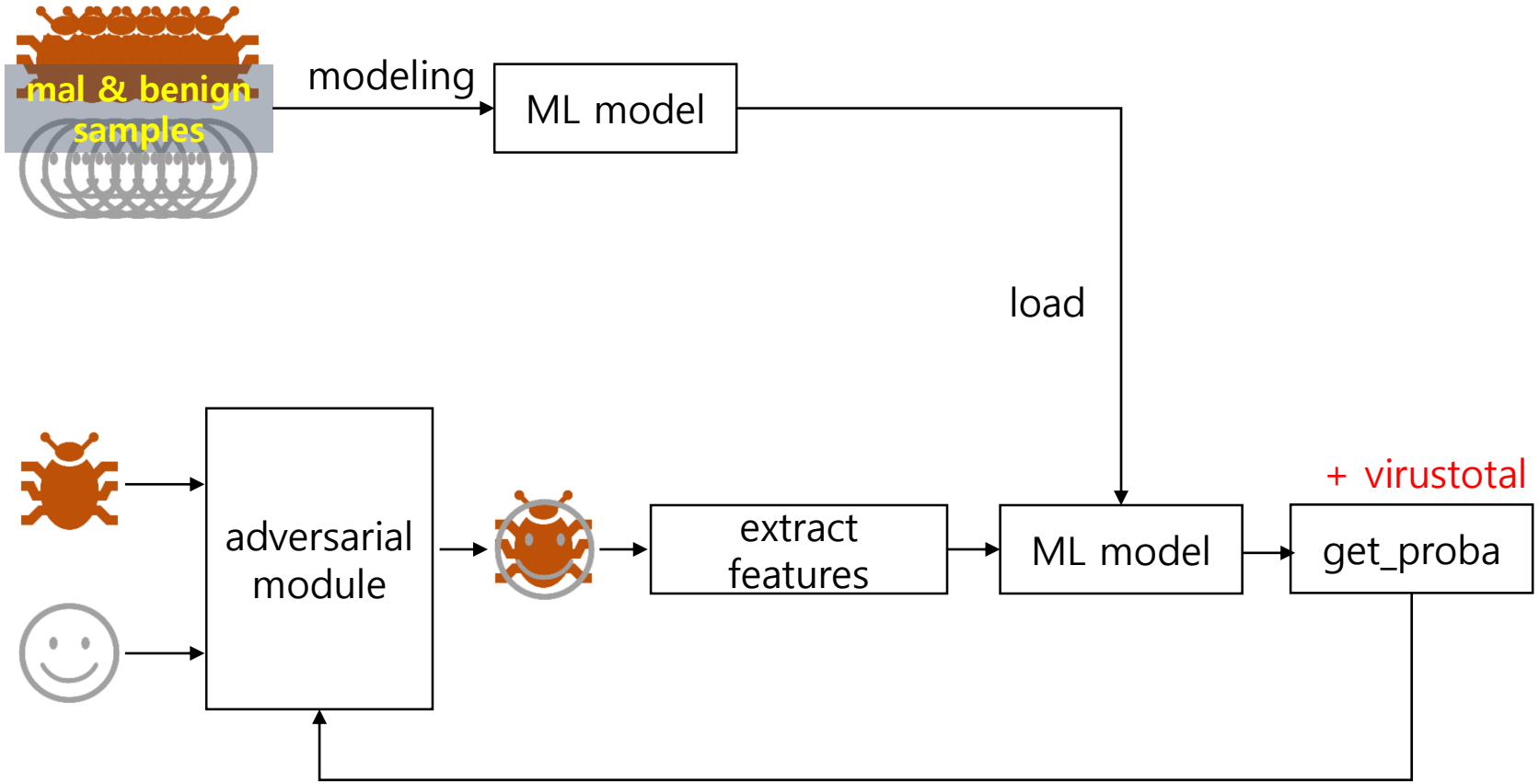
> 지도학습: 알고리즘이 달라도 목적은 모두 같다

> 파라미터: 최적의 파라미터는 최적의 알고리즘을 보장해 준다

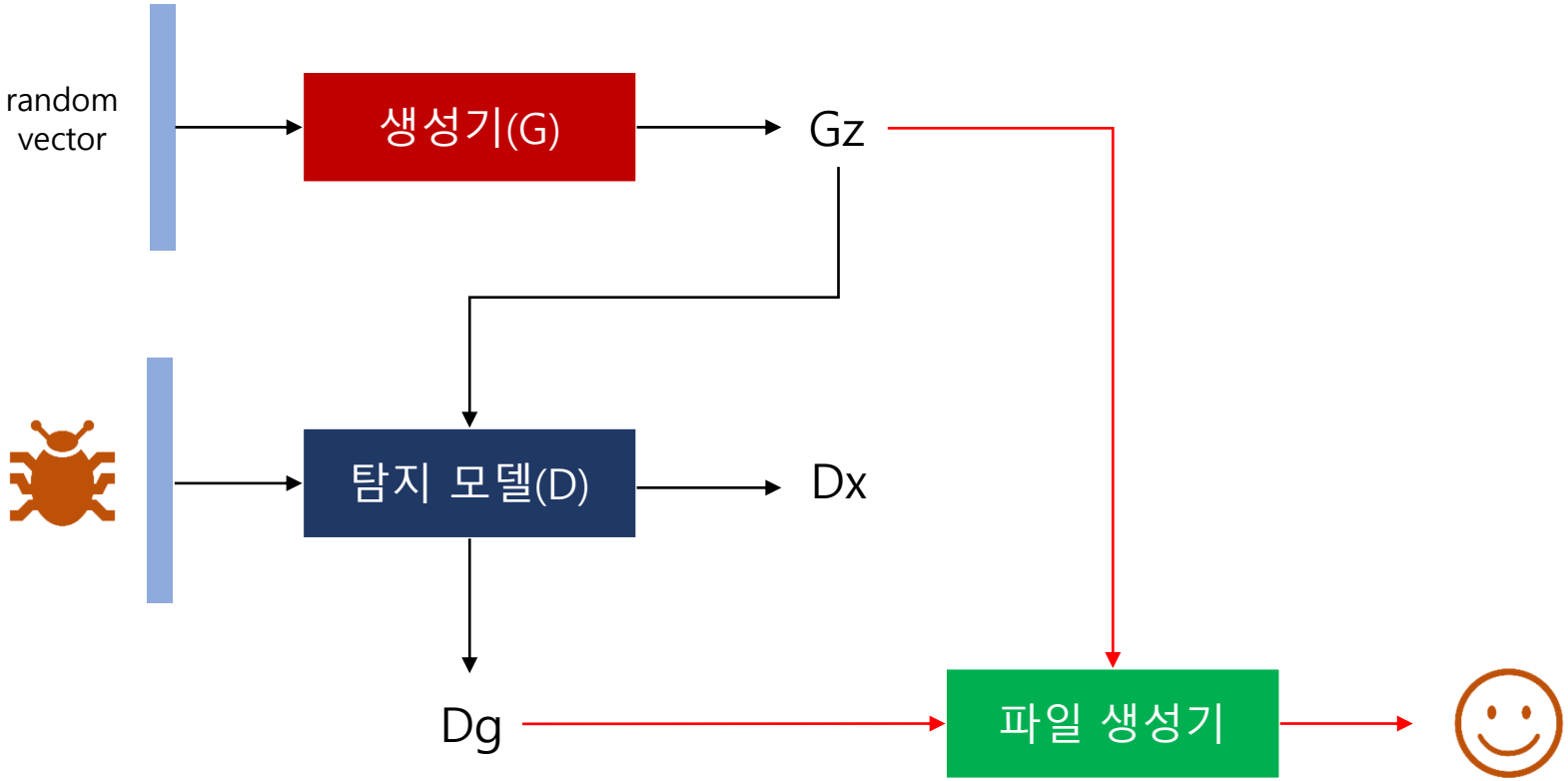
> 데이터: 충분한 양의 데이터를 확보하면 결과는 비슷해 진다

> 특징: EXE 형태의 파일이 가질 수 있는 특징은 한계가 있다

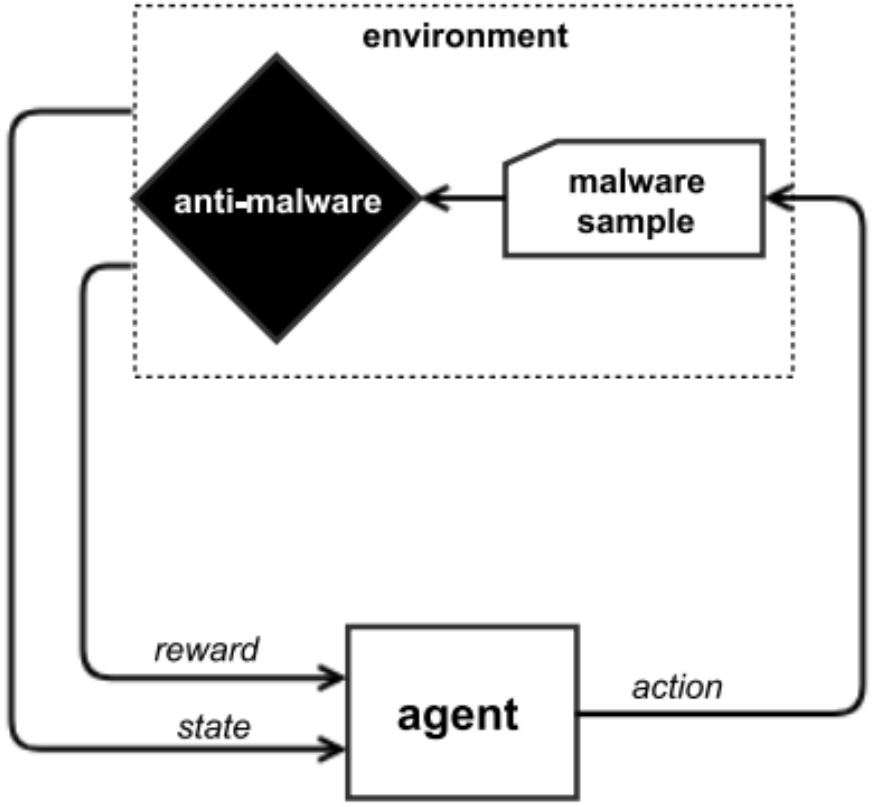
Adversarial 모델 예시



Adversarial 모델 예시



Adversarial 모델 예시

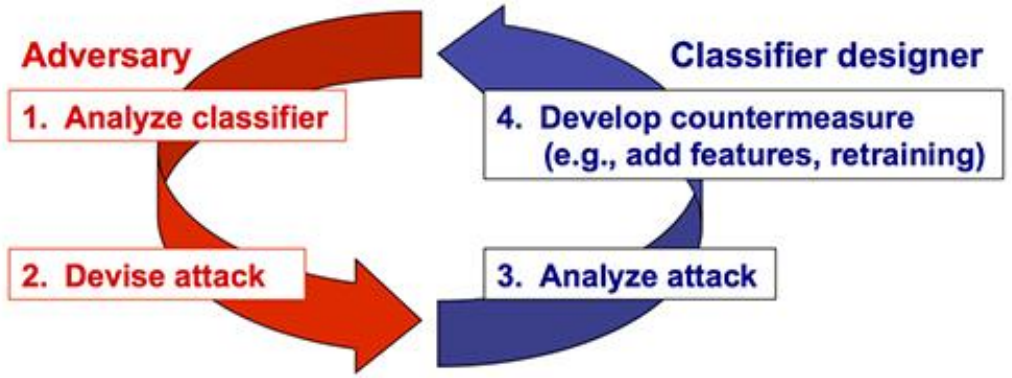


Adversarial ML

Arms Race Problem

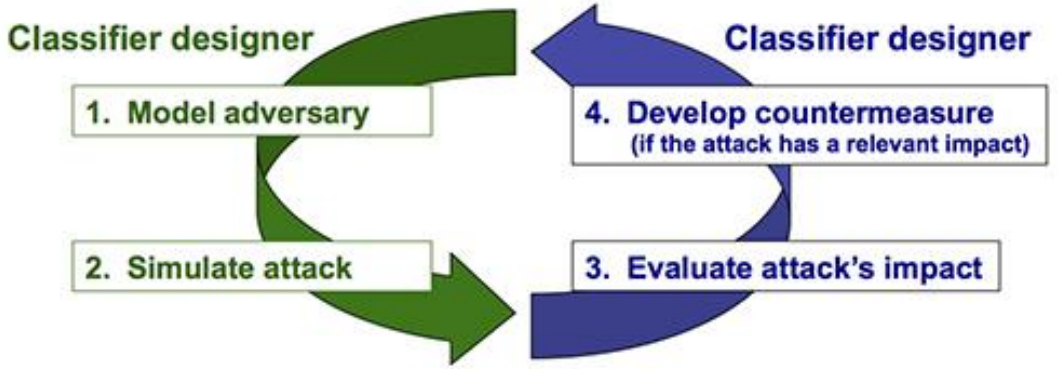
= 만들고 난 후에 대응할 것인가

: 모니터링, 업데이트



= 만들 때부터 고려할 것인가

: 시큐어 코딩, 테스팅



대응 방안

Adversarial ML 공격 대응 방법

- 쉽게 흔들리지 않는 좋은 특징을 쓰거나(feature engineering)
- 여러 모델을 혼합한 형태인 앙상블 모델을 사용
- 초기 구축 단계부터 Adversarial 테스팅 도입(by design)

대응 방안

쉽게 흔들리지 않는 좋은 특징(feature engineering) 사용

> 쉽게 흔들리지 않는 좋은 특징이란?

악성과 정상 데이터를 명확하게 구분할 수 있으면서 단순 추출이 아닌 도메인 전문가의 경험과 기술로 '가공'한 특징

> 만약 딥러닝 모델이라면?

특징 공학을 적용한 데이터로 모델링 했다면 위와 동일

Raw 데이터를 넣었다면 첫 번째 방법 적용 불가능

대응 방안

여러 모델을 혼합한 형태인 앙상블 모델을 사용

> 앙상블(ensemble) 모델이란?

여러가지 동일한 종류의 혹은 서로 상이한 모형들의 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법론

> 앙상블의 강력함

모델 정확도에 방해가 되는 노이즈 필터링 가능

모델 해석을 쉽게 만들어 예측성을 높여줌

개별 모델에 대한 공격에 저항력을 높여줌

대응 방안

초기 구축 단계부터 Adversarial 테스팅 도입(by design)

> 테스팅 코드 작성

초기 모델 구축 후에 해당 모델을 테스트할 수 있는 코드를 작성
(Discriminative vs Adversarial)

> 테스팅 코드 적용 단계

초기 모델 구축 후

정기 /비정기 테스팅 스케줄링

모델 업데이트 시

예상되는 미래 모습

소프트웨어 취약점으로 인한 문제 발생

→ 소프트웨어 취약점 점검

그래도 계속 문제가 발생

→ 설계 단계부터 점검하라(시큐어코딩)

그래도 자꾸 적용 안함

→ 의무화, 법제화

적대적 머신러닝으로 인한 문제 발생

→ 머신러닝 모델 점검

그래도 계속 문제가 발생

→ 설계 단계부터 점검하라(Adv ML)

그래도 자꾸 적용 안함

→ 의무화, 법제화

감사합니다.