

Machine Learning in Security

Introduction

- **Artificial Intelligence: A Modern Approach**

1. Thinking like Human
2. Acting like Human
3. Thinking Rationally
4. Acting Rationally

Definition: Artificial intelligence (AI) is [intelligence](#) demonstrated by [machines](#), unlike the **natural intelligence** [displayed by humans](#) and [animals](#), which involves consciousness and emotionality. (from Wikipedia)

Human: Cognition → Processing(Analyzing) → Decision Making → Action

- **How to determine the optimal action?**

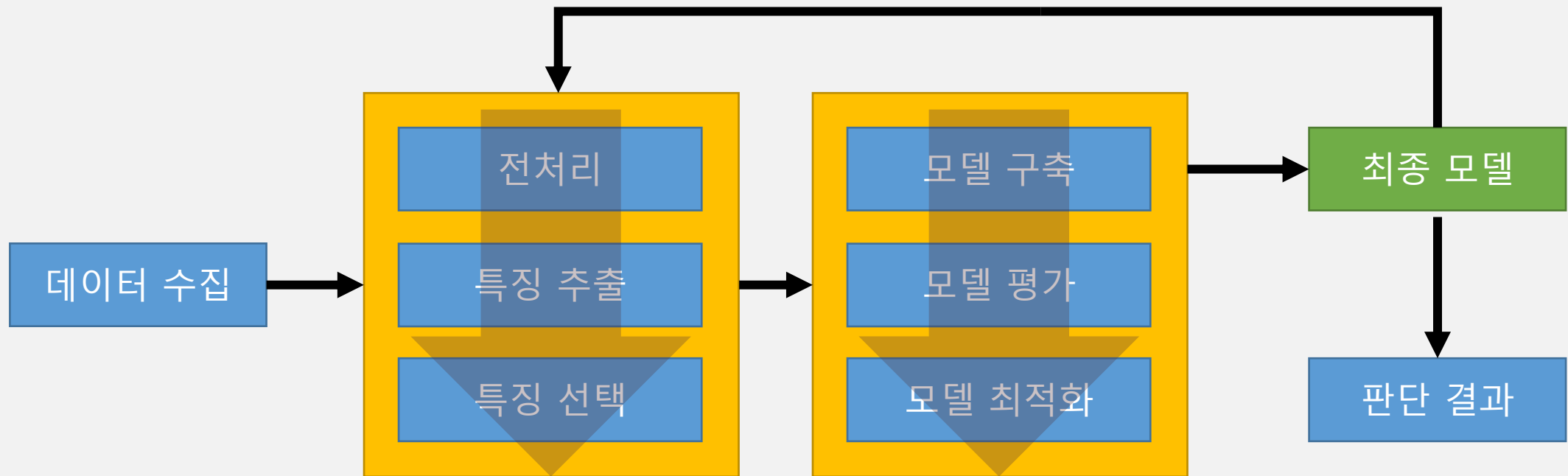
1. Data-driven (데이터 중심) : 관찰과 가정을 바탕으로 경험을 통해 판단
 - 수많은 데이터에서 공통적인 '패턴'을 찾는 문제
2. Algorithm-driven (알고리즘 중심) : 수학과 공학적인 방식으로 판단
 - 수많은 해(답안)나 경우의 수 중에서 '최적의 해'를 찾는 문제

Machine Learning (머신러닝; 기계학습)

- 패턴 인식: 데이터 안에 숨겨진 '진짜 의미' 를 찾는 기술
- 지도 학습 (Supervised Learning), 비지도 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning)

보안에서의 빅데이터와 머신러닝(Machine Learning)

- 머신러닝 프로세스



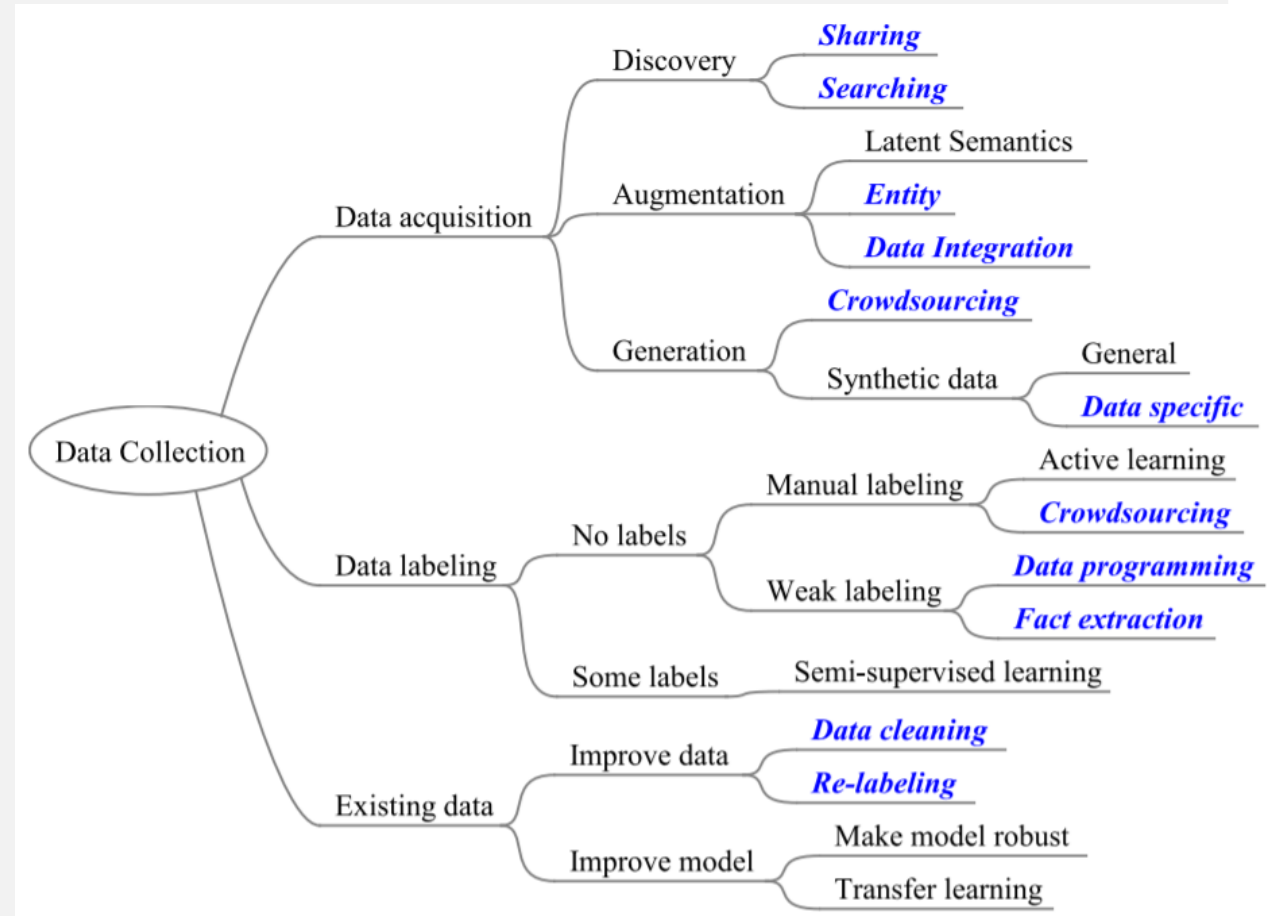
1. 데이터 수집

- 양질의 데이터
- 컴퓨터가 이해할 수 있는 형식
- 다양한 채널을 통한 수집 (하드웨어를 통한 수집, 웹 크롤러 등)
- 데이터의 양은? (가능한 많은 데이터 확보가 중요)
- 해결하고자 하는 문제 영역에 도움이 되는 많은 양의 데이터가 없다면?

보안에서의 빅데이터와 머신러닝(Machine Learning)

1. 데이터 수집 (Data Collection)

- 양질의 데이터
- 컴퓨터가 이해할 수 있는 형식
- 다양한 채널을 통한 수집
 - 하드웨어(센서, 카메라)를 통한 수집
 - 웹 크롤러를 통한 웹 문서 수집
 - 시스템, 서버 로그 등
- 데이터의 양은?
 - 가능한 많은 데이터 확보가 중요
- 해결하고자 하는 문제 영역에 도움이 되는 많은 양의 데이터가 없다면?



[그림 출처]

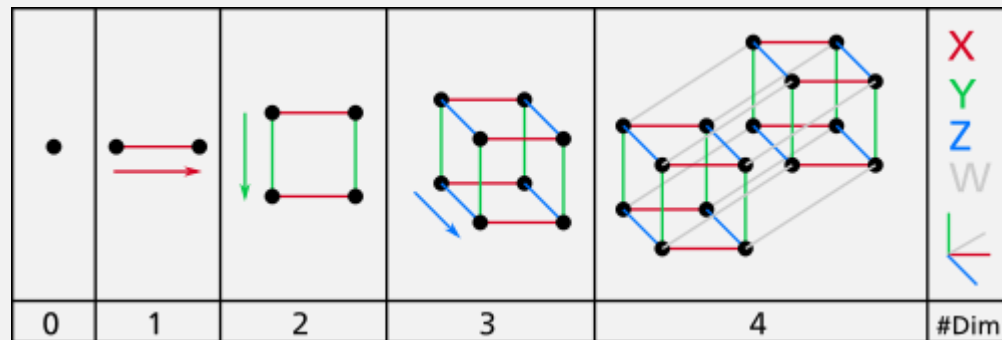
A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective (<https://arxiv.org/abs/1811.03402>)

2. 특징 공학 (Feature Engineering)

- 모델의 성능에 영향을 끼치는 핵심 요소 (가장 중요함)
- 특징(Feature)
 - [Feature](#) is an individual measurable property or characteristic of a phenomenon being observed (from Wikipedia)
- 모든 도메인에 적용가능한 공통적인 기술 보다는 특정 도메인에 특화된 기술들이 더 많음
- 데이터 관찰 (데이터를 이해하는 중요한 과정)
 - 어떤 형식으로 저장되어 있는지
 - 누락된/잘못된 값은 없는지
 - 데이터의 분포는 어떠한지
- 시각화 (보통 데이터 관찰과 전처리과정과 함께 진행됨)
 - 시각화된 결과물을 토대로 데이터 기저에 숨겨진 특성 파악 및 의사결정 수행


2. 특징 공학 (Feature Engineering)

- 특징 추출 (데이터 전처리 후)
 - 데이터를 그대로 사용
 - 통계 수치를 활용
 - 도메인 특성을 토대로 추출
- 추출 시 가능한 모든 특징을 뽑아 두는 것이 좋음
- 특징 선택 (특징 추출 후)
 - 이전 단계에서 추출한 특징 중 모델링에 사용할 특징 조합을 선택하는 단계
 - 특징이 많으면 무조건 좋을까? → **Curse of Dimensionality**



[그림 출처] 차원 - 위키피디아
(<https://ko.wikipedia.org/wiki/%EC%B0%A8%EC%9B%90>)

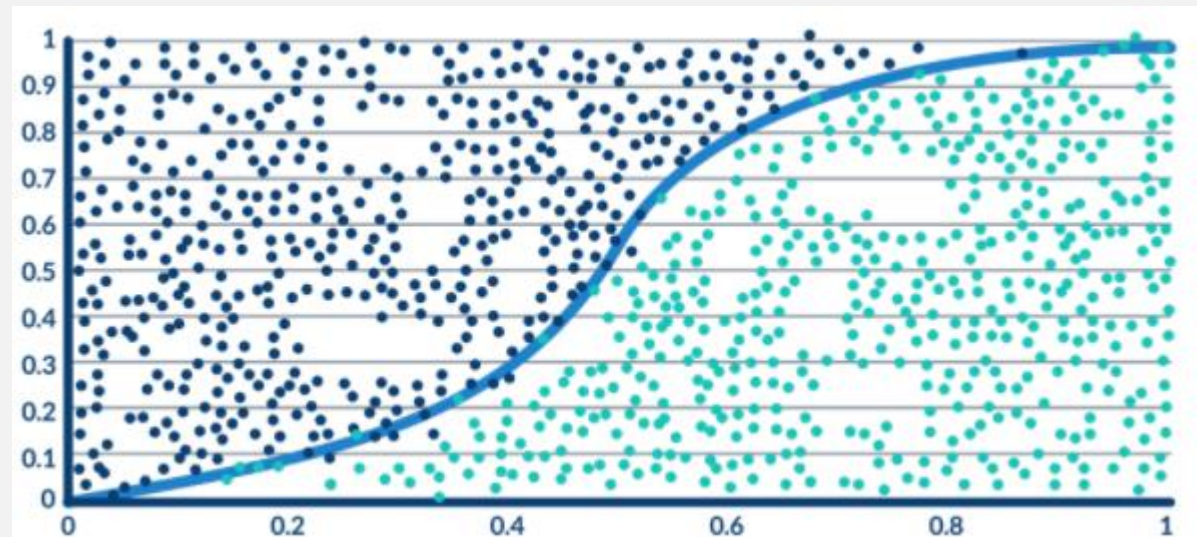
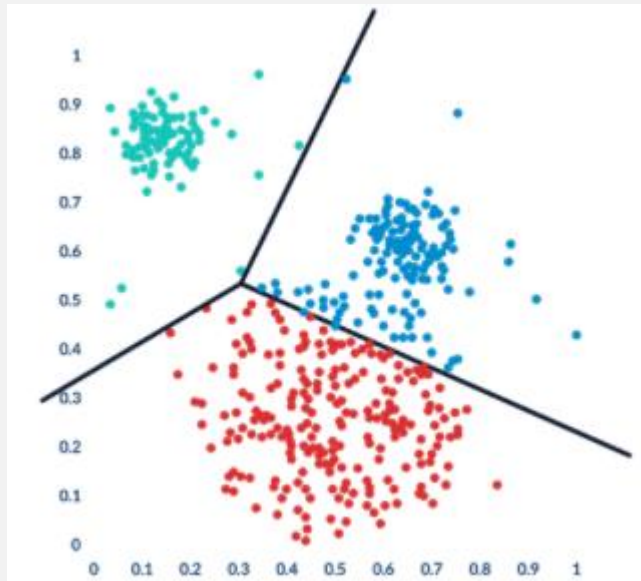
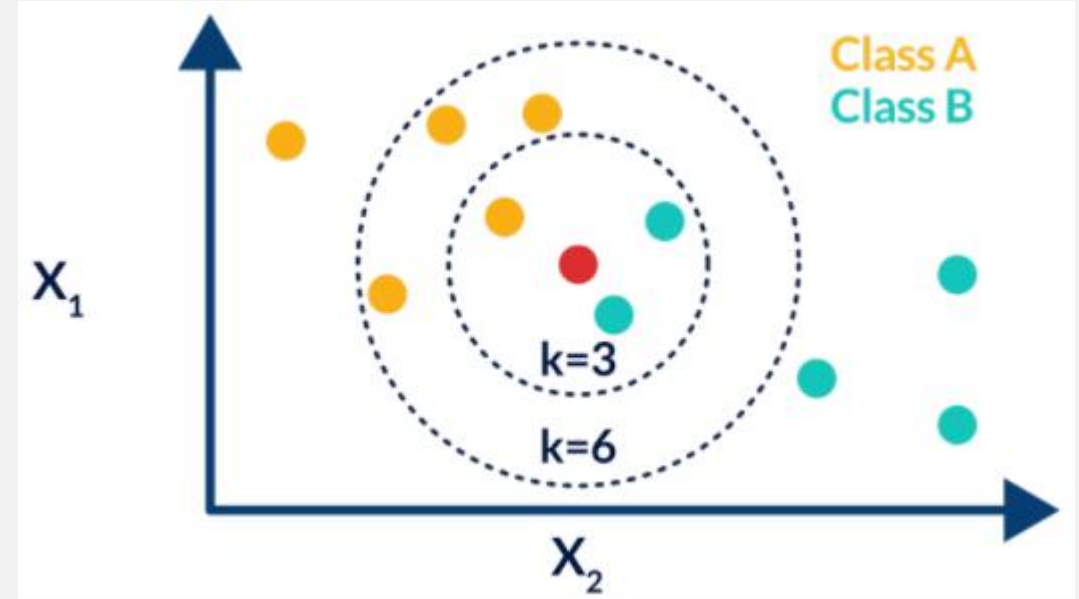
2. 특징 공학 (Feature Engineering)

- 특징 선택 (최적의 조합 선택)
 - 최적의 조합을 찾을 때 까지 특징 개수를 하나씩 늘리거나 줄임
 - 문제에 적용 가능한 모델을 3~4가지 정한 뒤에 선택한 특징 조합을 대입 
(분류 문제인 경우 선택한 특징 조합을 'SVM', 'Random Forest', 'Deep Neural Network' 등의 모델에 각각 넣고 정확도를 계산함)
 - 도메인 지식을 바탕으로 휴리스틱에 의존해 특징을 선택하는 방법도 있음
- 최적의 특징 조합 선택 후
 - 모델링에 적합한 형태로 변환해야 함
 - 정규화: 모든 특징 값의 범위를 동일한 범위로 맞추는 과정
 - 스케일링: 정규 분포 형태로 변환해주는 과정
 - 범주형 데이터 처리: 범주형 데이터 → 수치형 데이터로 변환

보안에서의 빅데이터와 머신러닝(Machine Learning)

3. 모델링 (Modeling)

- 모델 구축 → 모델 평가 → 최적화 의 과정
 - 결국, 최종 목표는 데이터를 가장 잘 표현해주는 선 (직선, 곡선 등)



3. 모델링 (Modeling)

- 예측: 학습을 위해 정답지가 필요함 (지도학습)
- 분류: 정답지가 필요 없고, 데이터의 분포와 상대적 위치를 기반으로 학습함 (비지도학습)
- 모델 구축:
 - 배깅: 전체 데이터를 여러 개의 샘플로 나누어 (여러 번) 학습한 뒤 전체 결과 집계
 - 부스팅: 랜덤 샘플을 이용하되(배깅과 유사), 가중치를 부여한 학습 (순차학습)
 - 스택킹: 서로 다른 모델을 조합해 최고의 성능을 내는 모델을 생성
- 모델 평가(검증):
 - 전체 데이터 중에 80%를 학습에 사용
 - 나머지 20%를 테스트에 사용 (비율은 달라질 수 있음)
 - K-Fold 교차 검증 기법: 데이터를 K개로 나눈 뒤에 테스트를 K번 수행