Machine Learning

Chapter 1 머신러닝 개요 (intro)

START

Smart Media
스마트미디어인재개발원

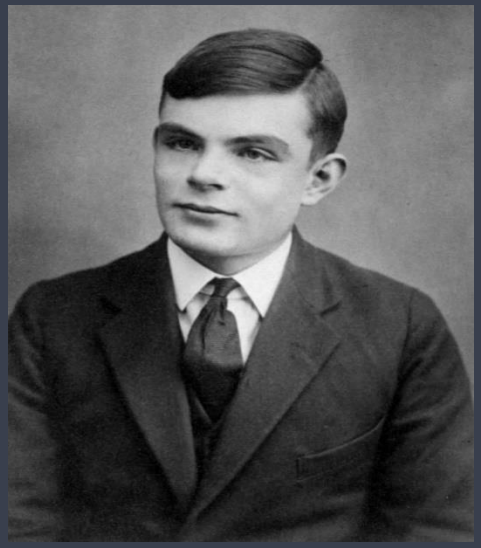
학습목표

- Machine Learning 개념을 이해 할 수 있다.
- Machine Learning의 종류 및 과정을 알 수 있다.
- 기계학습과 관련된 기본 용어를 알 수 있다.

머신러닝이란?

Smart Media
스마트미디어인재개발원

머신러닝(Machine Learning) 출현



앨런 튜링(1912-1954)

기계는 **인간과 같은 사고**를 할 수 있는가?

머신러닝(Machine Learning) 발전

컴퓨팅 속도 + 초고속 인터넷

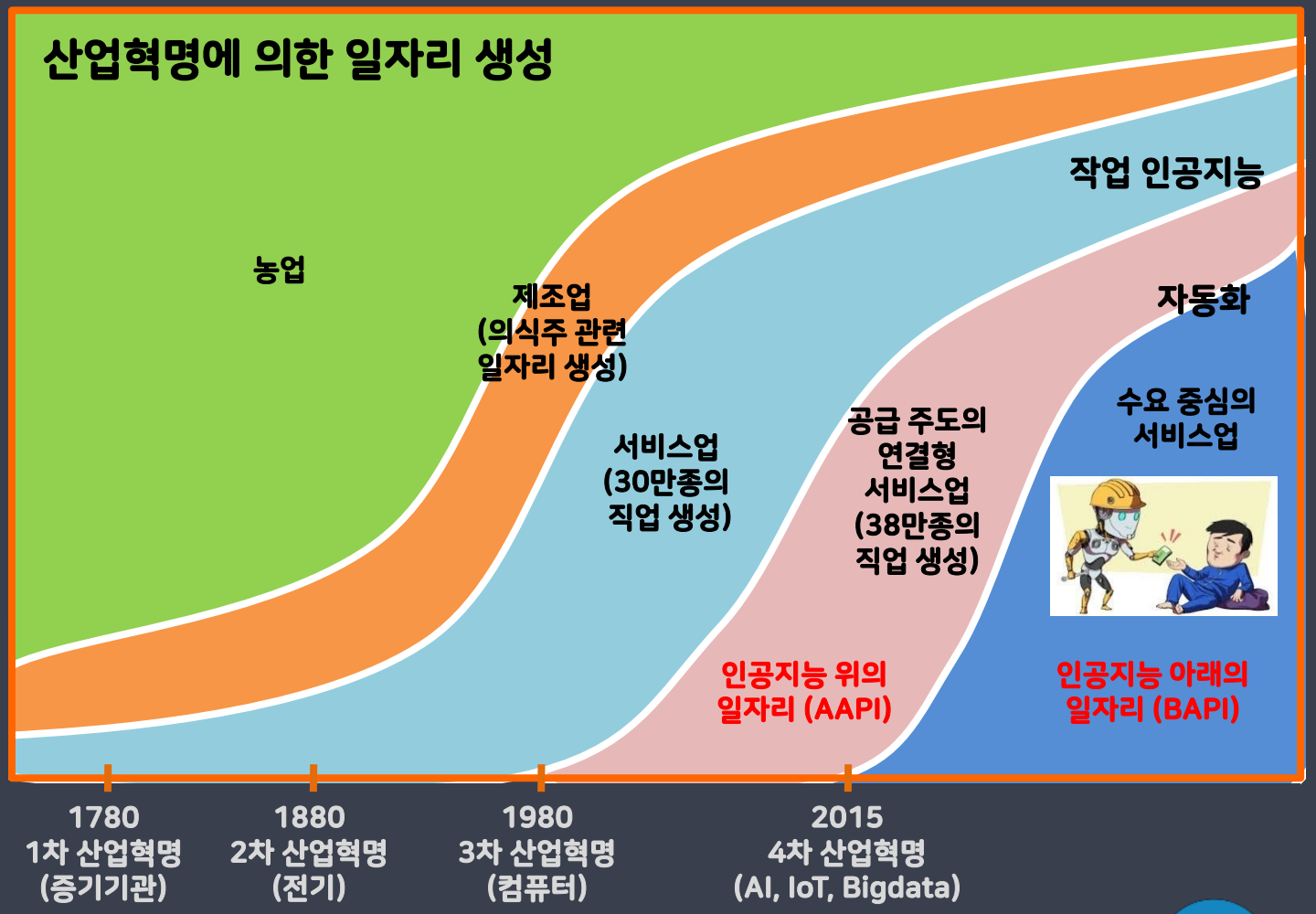
↓

많은 양의 데이터 (학습 재료)

↓

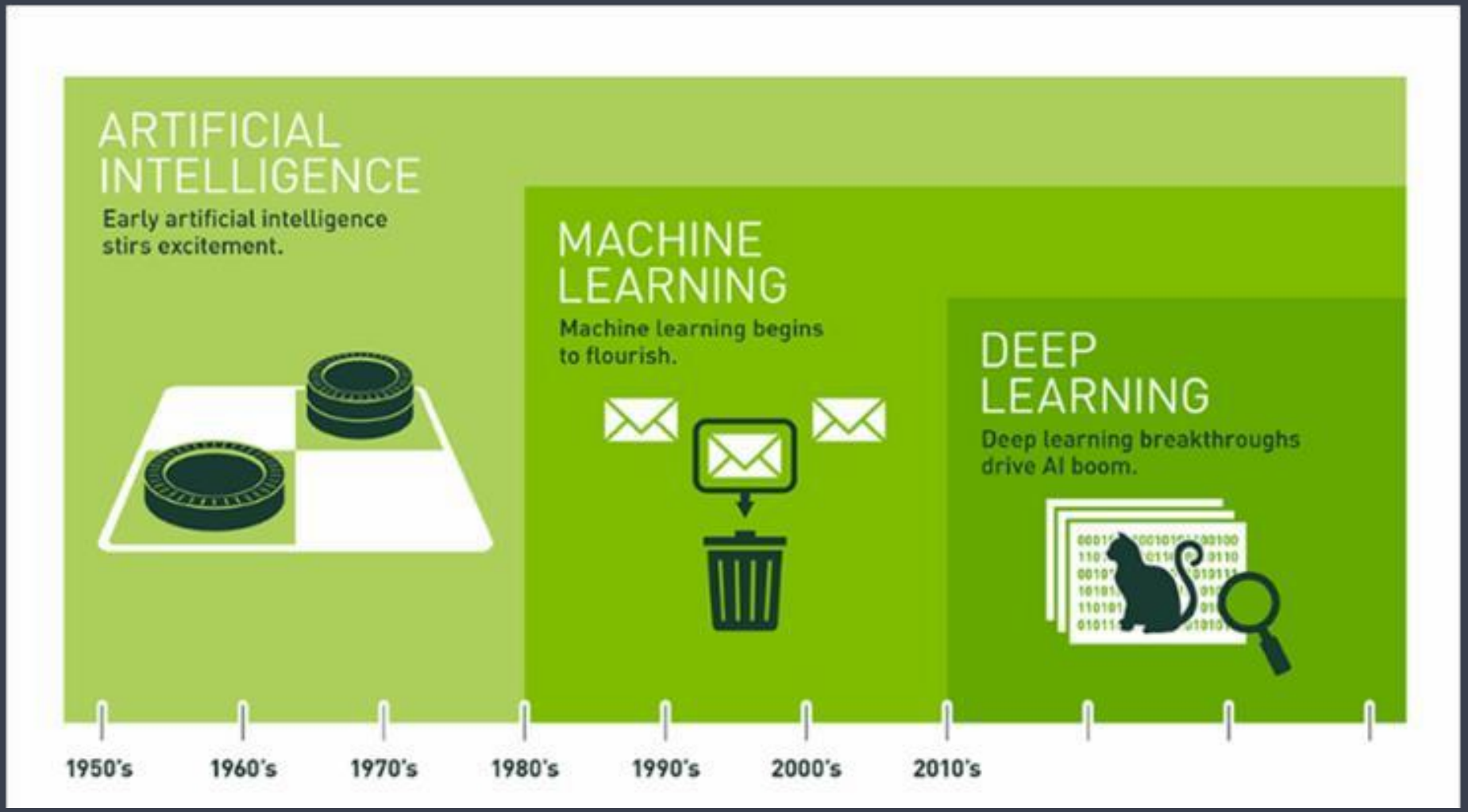
인공지능 발전의 가속화

머신러닝(Machine Learning) 발전



Smart Media
스마트미디어인재개발원

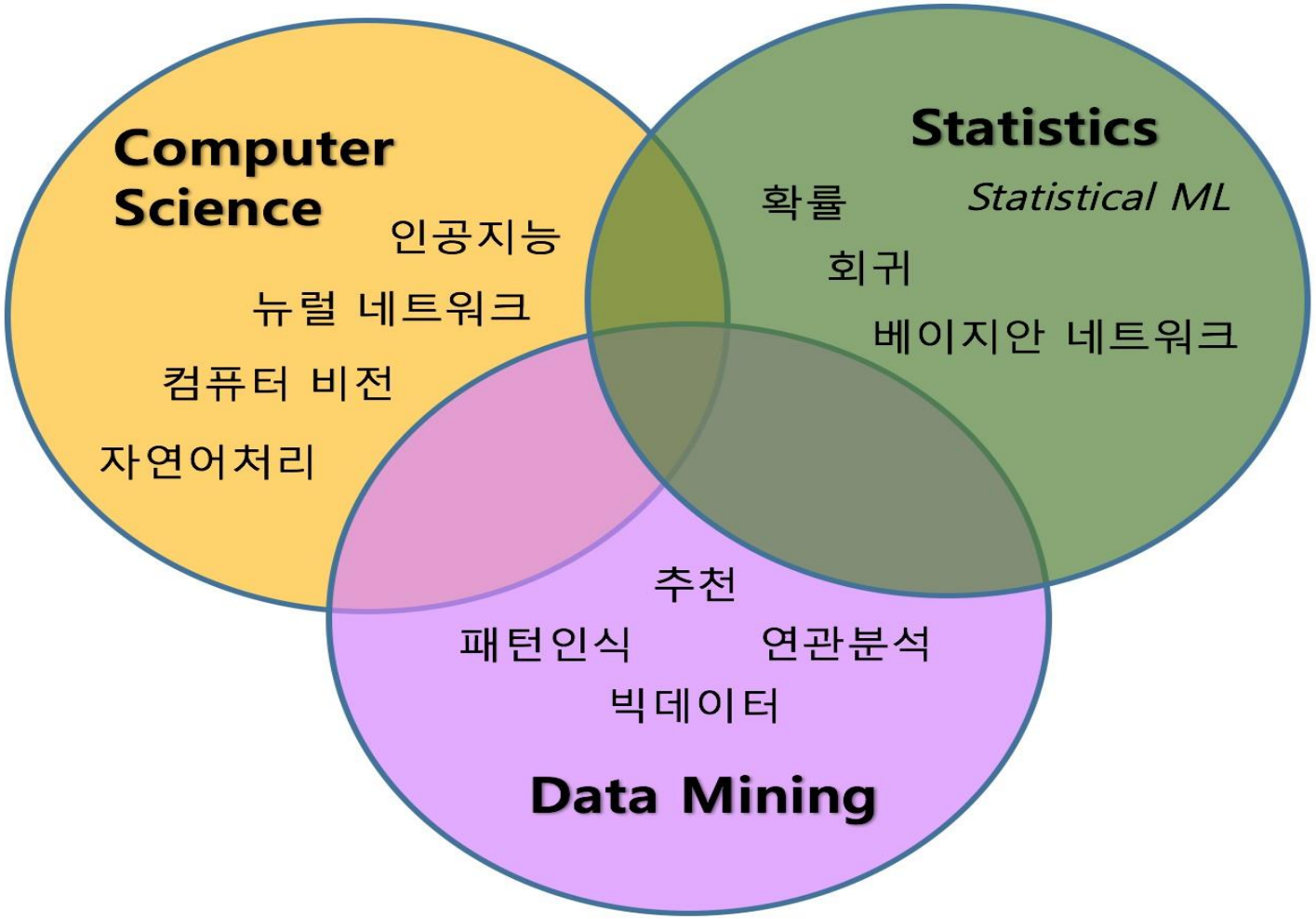
인공지능, 머신러닝, 딥러닝



머신러닝(Machine Learning)이란?

- 데이터를 기반으로 학습을 시켜서 예측하게 만드는 기법
- 인공지능의 한 분야로 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
- 통계학, 데이터 마이닝, 컴퓨터 과학이 어우러진 분야

머신러닝

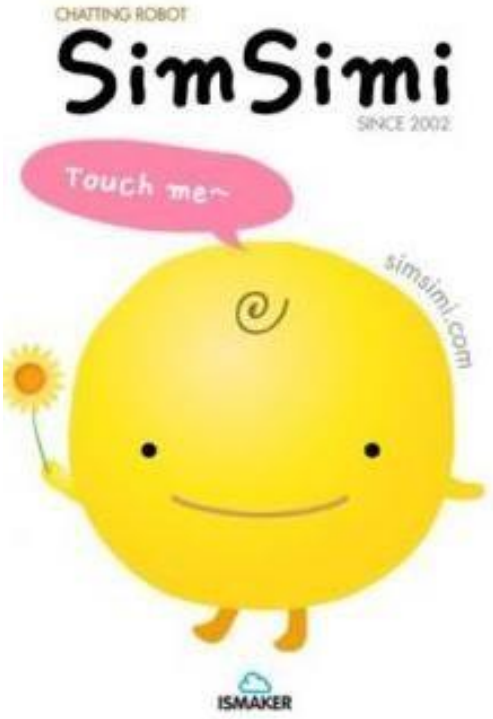

Computer Science
인공지능
뉴럴 네트워크
컴퓨터 비전
자연어처리
Statistics
확률
Statistical ML
회귀
베이지안 네트워크
추천
패턴인식
연관분석
빅데이터
Data Mining

Rule-based expert system

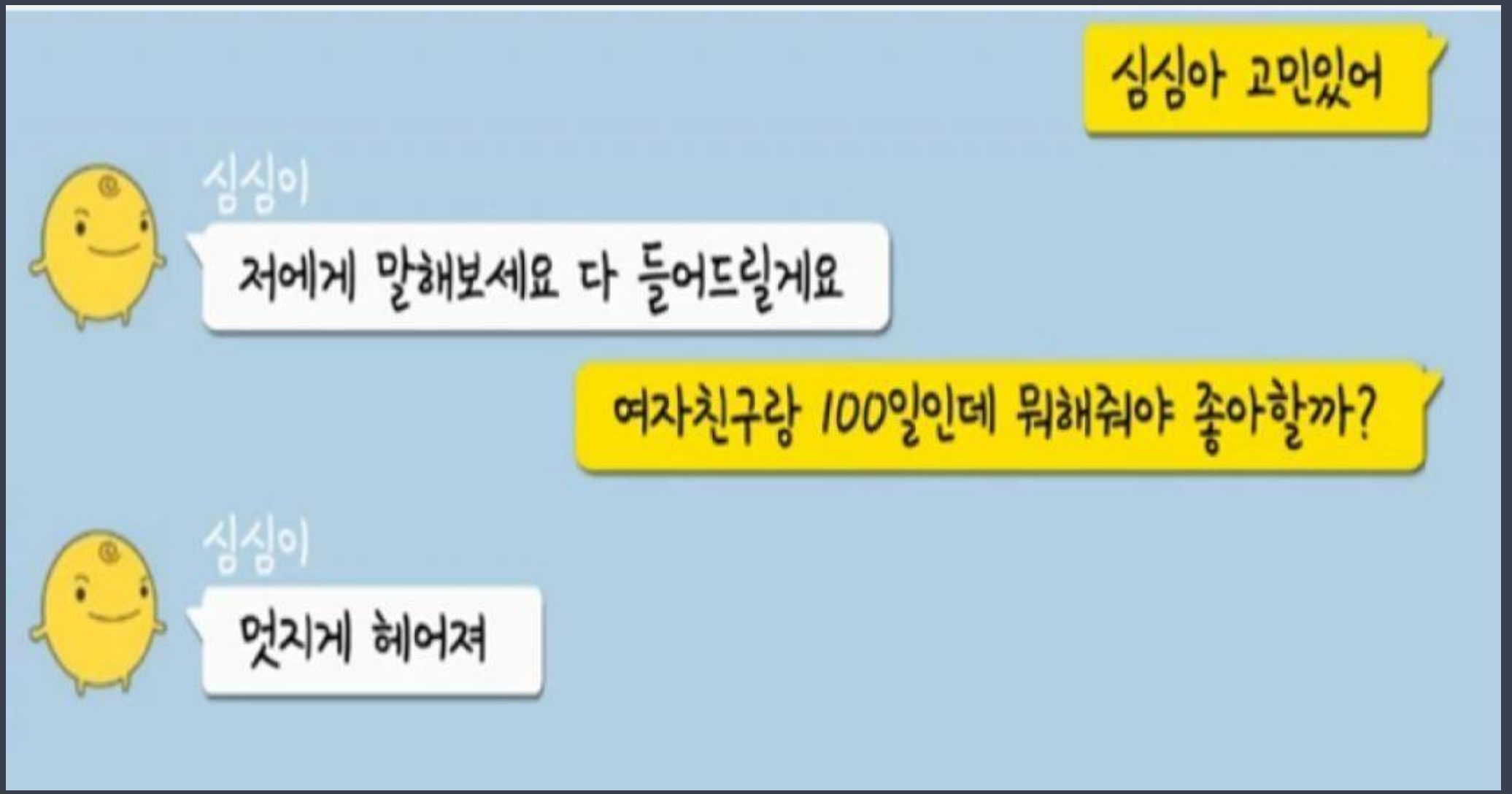
Rule-based expert system
(규칙 기반 전문가 시스템)

"if"와 "else"로 하드 코딩된 명령을 사용하는 시스템

Rule-based expert system



Rule-based expert system



Rule-based expert system

넌 재밋냐 사는게

싱싱이

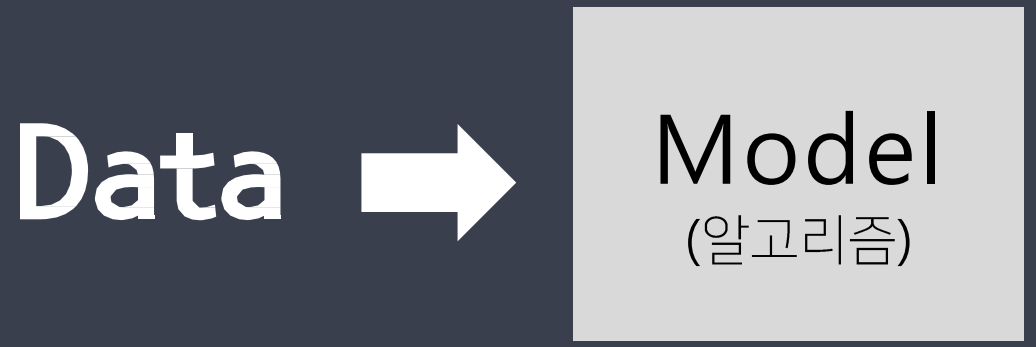
재밋겠냐 하루종일 대답이나 하고 앉아있는데...

Rule-based expert system 문제점

- 스팸 메일 필터
- 얼굴 인식 시스템

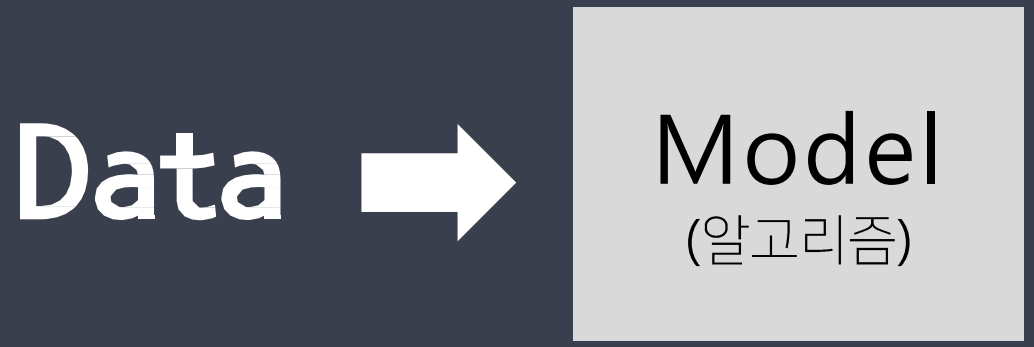
많은 상황에 대한 규칙들을 모두 만들어 낼 수 없다

머신러닝(Machine Learning)



학습을 통해 기계가 **스스로 규칙**을 만들어낸다.

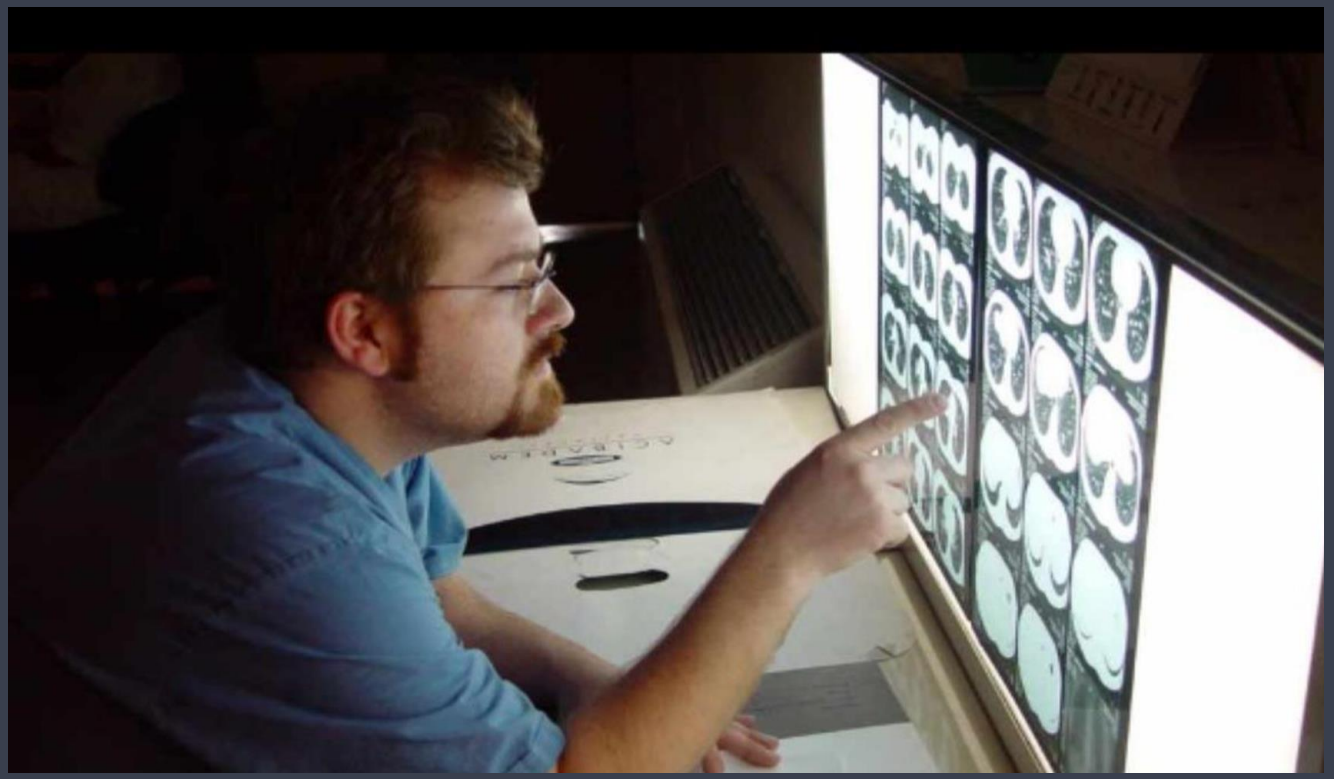
머신러닝(Machine Learning)



데이터를 이용하여 **특성과 패턴을 학습**하고,
그 결과를 바탕으로 미지의 데이터에 대한 **미래결과(값, 분포)를 예측**하는 것

의료 인공지능 분야 사례

영상 의료/병리 데이터의 분석 및 판독 (Deep Learning)
- 영상의학과 전문의



산업 인공지능 분야 사례

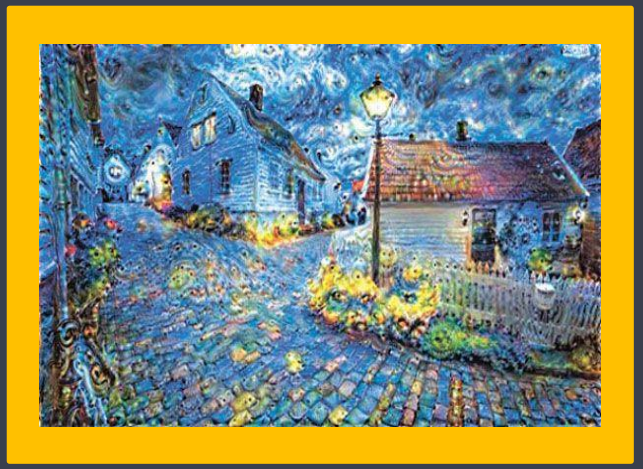
신제품 마케팅
- 인공지능 트렌드 분석 시스템 '엘시아(LCIA)'

'엘시아(LCIA)'의 분석으로
소비자 트렌드 키워드를 도출해
혼술족을 주요 타겟으로 출시한
'꼬깔콘 버팔로윙맛'

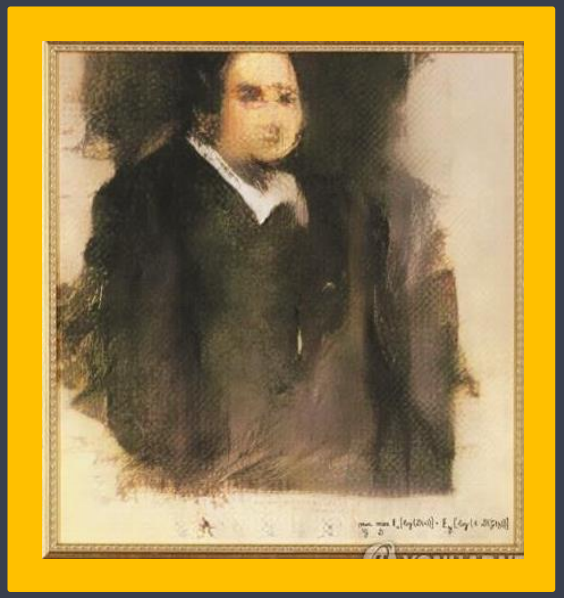
Smart Media
스마트미디어인재개발원

예술 인공지능 분야 사례

무엇이 진짜일까요?



구글 딥드림(Deep Dream)



에드몽 드 벨라미 인공지능 '오비어스'의 작품

사과나무 - 구스타프 클림프

인공지능의 작품은 예술인가? 기술인가?

Smart Media
스마트미디어인재개발원

머신러닝(Machine Learning) 종류

지도학습 (Supervised Learning)

비지도학습 (Unsupervised Learning)

강화학습 (Reinforcement Learning)

머신러닝(Machine Learning) 종류

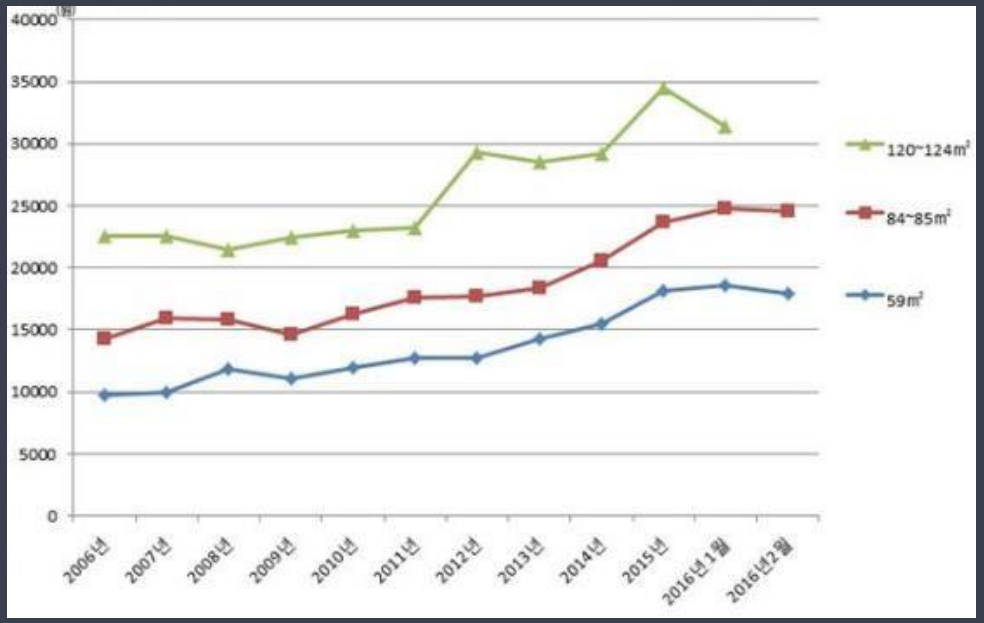
지도 학습 (Supervised Learning)

- 데이터에 대한 Label(명시적인 답)이 주어진 상태에서 컴퓨터를 학습시키는 방법.
- 분류(Classification)와 회귀(Regression)로 나뉘어진다.

머신러닝(Machine Learning) 종류

지도 학습 (Supervised Learning)

스팸 메일 분류



집 가격 예측

머신러닝(Machine Learning) 종류

Kaggle Titanic 데이터

Feature

Class Label

Survived	Pclass	Name	Sex	Age	SibSp
0	3	Braund, Mr	male	22	1
1	1	Cumings, M	female	38	1
1	3	Heikkinen,	female	26	0
1	1	Futrelle, M	female	35	1
0	3	Allen, Mr. W	male	35	0
0	3	Moran, Mr	male		0
0	1	McCarthy,	male	54	0
0	3	Palsson, M	male	2	3
1	3	Johnson, M	female	27	0
1	2	Nasser, Mr	female	14	1
1	3	Sandstrom	female	4	1

머신러닝(Machine Learning) 종류

분류 (Classification)

- 미리 정의된 여러 클래스 레이블 중 하나를 예측하는 것.
- 속성 값을 입력, 클래스 값을 출력으로 하는 모델
- 붓꽃(iris)의 세 품종 중 하나로 분류, 암 분류 등.
- 이진분류, 다중 분류 등이 있다.

머신러닝(Machine Learning) 종류

회귀 (Regression)

- 연속적인 숫자를 예측하는 것.
- 속성 값을 입력, 연속적인 실수 값을 출력으로 하는 모델
- 어떤 사람의 교육수준, 나이, 주거지를 바탕으로 연간 소득 예측.
- 예측 값의 미묘한 차이가 크게 중요하지 않다.

머신러닝(Machine Learning) 종류

비지도 학습 (Unsupervised Learning)

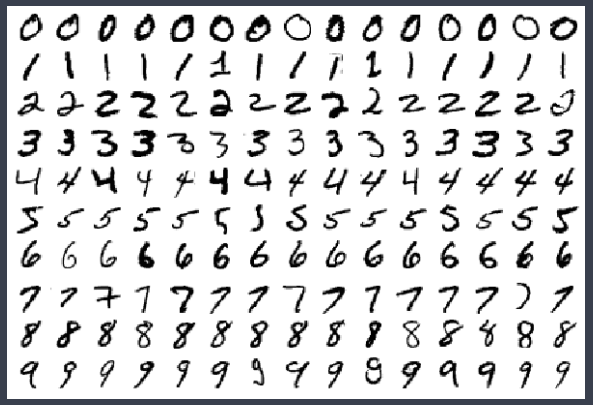
- 데이터에 대한 Label(명시적인 답)이 없는 상태에서 컴퓨터를 학습시키는 방법.
- 데이터의 숨겨진 특징, 구조, 패턴을 파악하는데 사용.
- 데이터를 비슷한 특성끼리 묶는 클러스터링(Clustering)과 차원축소(Dimensionality Reduction)등이 있다.

머신러닝(Machine Learning) 종류

비지도 학습 (Unsupervised Learning)

이미지 감색 처리

소비자 그룹 발견을
통한 마케팅



손 글씨 숫자 인식

머신러닝(Machine Learning) 종류

강화 학습 (Reinforcement Learning)

- 지도학습과 비슷하지만 완전한 답(Label)을 제공하지 않는 특징 이 있다.
- 기계는 더 많은 보상을 얻을 수 있는 방향으로 행동을 학습
- 주로 게임이나 로봇을 학습시키는데 많이 사용

머신러닝(Machine Learning) 종류

강화 학습 (Reinforcement Learning)

머신러닝(Machine Learning)이 유용한 분야

- 기존 솔루션으로는 많은 수동 조정과 규칙이 필요한 문제
- 전통적인 방식으로는 전혀 해결 방법이 없는 복잡한 문제
- 새로운 데이터에 적응해야하는 유동적인 환경
- 대량의 데이터에서 통찰을 얻어야 하는 문제

머신러닝(Machine Learning) 과정

1. Problem Identification(문제정의)
2. Data Collect(데이터 수집)
3. Data Preprocessing(데이터 전처리)
4. EDA(탐색적 데이터분석)
5. Model 선택, Hyper Parameter 조정
6. Training(학습)
7. Evaluation(평가)

머신러닝(Machine Learning) 과정

1. Problem Identification(문제정의)

- 비즈니스 목적 정의
  모델을 어떻게 사용해 이익을 얻을까?
- 현재 솔루션의 구성 파악
- 지도 vs 비지도 vs 강화
- 분류 vs 회귀

머신러닝(Machine Learning) 과정

2. Data Collect(데이터 수집)

- File (CSV, XML, JSON)
- Database
- Web Crawler (뉴스, SNS, 블로그)
- IoT 센서를 통한 수집
- Survey

머신러닝(Machine Learning) 과정

3. Data Preprocessing(데이터 전처리)

- 결측치, 이상치 처리

- Feature Engineering (특성공학)

  Scaling (단위 변환),
  Transform (새로운 속성 추출),
  Encoding (범주형 -> 수치형),
  Binning (수치형 -> 범주형)

머신러닝(Machine Learning) 과정

4. EDA(탐색적 데이터분석)

- 기술통계, 변수간 상관관계
- 시각화
  pandas, matplotlib, seaborn
- Feature Selection (사용할 특성 선택)

머신러닝(Machine Learning) 과정

5. Model 선택, Hyper Parameter 조정

- 목적에 맞는 적절한 모델 선택

- KNN, SVM, Linear Regression, Ridge, Lasso, Decision Tree, Random forest, CNN, RNN …

- Hyper Parameter
  model의 성능을 개선하기위해 사람이 직접 넣는 parameter

머신러닝(Machine Learning) 과정

6. Model Training(학습)

- model.fit(X_train,y_train)
  train 데이터와 test 데이터를 7:3 정도로 나눔

- model.predict (X_test)

머신러닝(Machine Learning) 과정

머신러닝(Machine Learning) 과정

7. Evaluation(평가)

- accuracy(정확도)
- recall(재현율)
- precision(정밀도)
- f1 score

머신러닝(Machine Learning) 과정

100명이 있다고 가정했을 때 그 중에 5명이 암이다.

만약 100명 전부 암 환자가 아니라고 예측한다면
Accuracy(정확도)는 95%

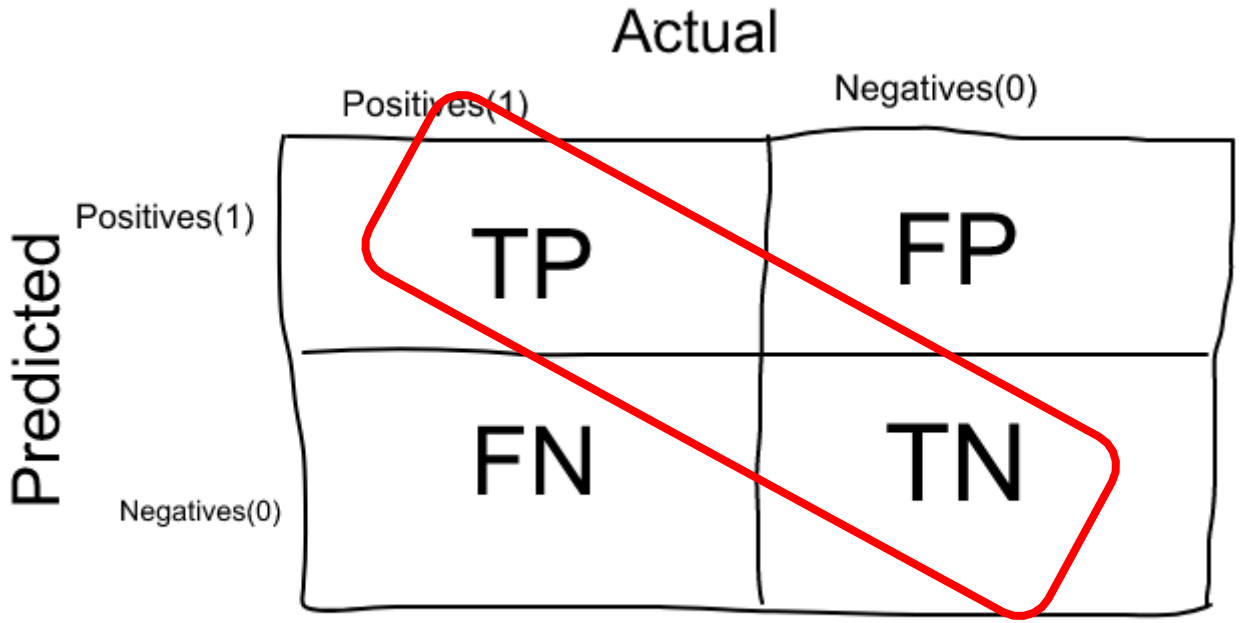
머신러닝(Machine Learning) 과정

환자 100명이 있다고 가정했을 때 그 중에 5명이 암이다.

만약 100명 전부 암 환자가 아니라고 예측한다면
Recall(재현율)은 0%

머신러닝(Machine Learning) 과정

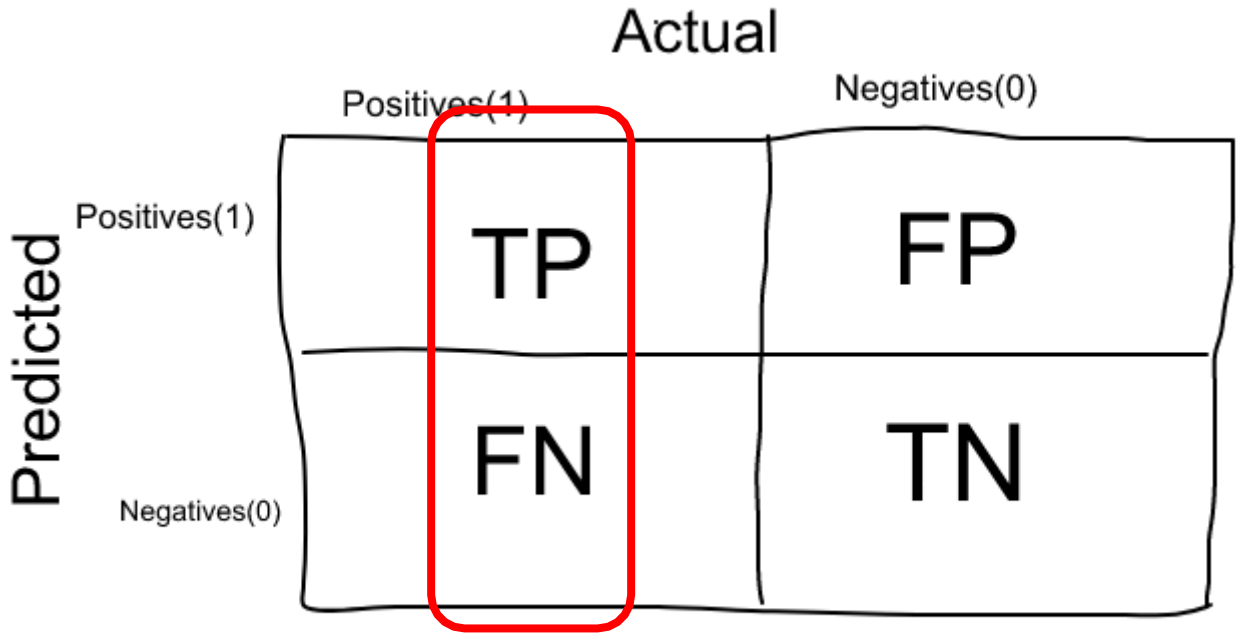
Accuracy(정확도) = (TP+TN) / (TP+FP+FN+TN)



Confusion Matrix

머신러닝(Machine Learning) 과정

Recall(재현율) = TP / (TP+FN)



Confusion Matrix

머신러닝(Machine Learning) 과정

머신러닝(Machine Learning) 과정

번호 [1, 2, 3, 4, 5, 6]
정답 : [음치,음치,음치,음치,가수,가수]

예측 : [음치,음치,가수,가수,가수,가수]

음치 기준 정확도와 재현율은 몇일까?

머신러닝(Machine Learning) 과정

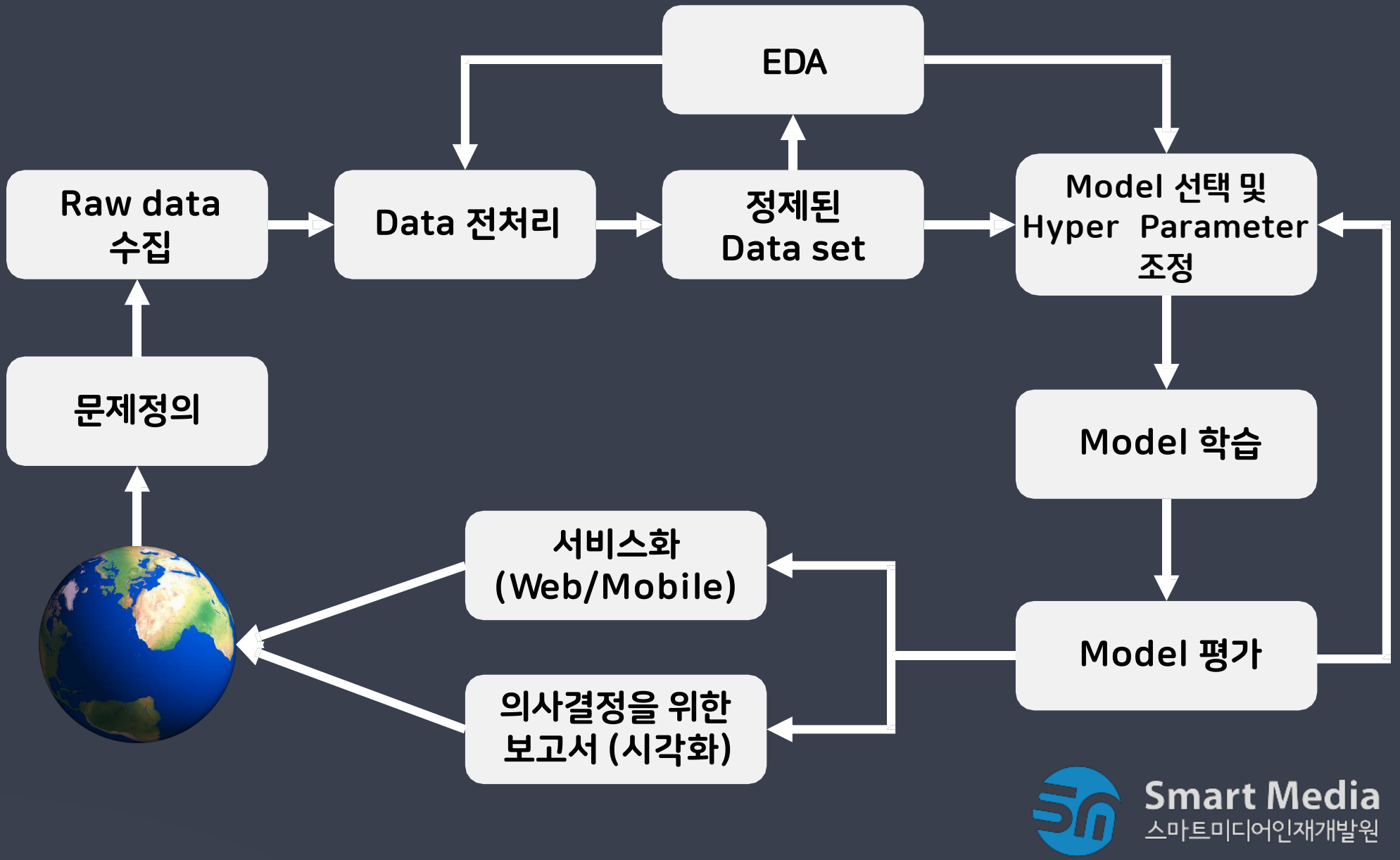
번호 [1, 2, 3, 4, 5, 6]
정답 : [음치,음치,음치,음치,가수,가수]

예측 : [음치,음치,가수,가수,가수,가수]

음치 기준 정확도와 재현율은 몇일까?

정확도 4/6 = 0.66, 재현율 2/4 = 0.50

머신러닝(Machine Learning) 과정



머신러닝(Machine Learning) 실습

scikit-learn

- 파이썬에서 쉽게 사용할 수 있는 머신러닝 프레임워크,라이브러리
- 회귀,분류,군집,차원축소,특성공학,전처리,교차검증,파이프라인 등 머신러닝에 필요한 기능을 갖춤
- 학습을 위한 샘플 데이터도 제공

머신러닝(Machine Learning) 실습

비만도 데이터를 이용해 학습 해 보자.

머신러닝(Machine Learning) 실습

	A	B	C	D
1	Gender	Height	Weight	Label
2	Male	174	96	Obesity
3	Male	189	87	Normal
4	Female	185	110	Obesity
5	Female	195	104	Overweight
6	Male	149	61	Overweight
7	Male	189	104	Overweight
8	Male	147	92	Extreme Obesity
9	Male	154	111	Extreme Obesity
10	Male	174	90	Overweight
11	Female	169	103	Obesity
12	Male	195	81	Normal
13	Female	159	80	Obesity
14	Female	192	101	Overweight
15	Male	155	51	Normal
16	Male	191	79	Normal
17	Female	153	107	Extreme Obesity
18	Female	157	110	Extreme Obesity
19	Male	140	129	Extreme Obesity
20	Male	144	145	Extreme Obesity

회귀 문제?

분류 문제?

머신러닝(Machine Learning) 실습

시각화를 통해 분류가 가능한 문제인지 확인해보자.