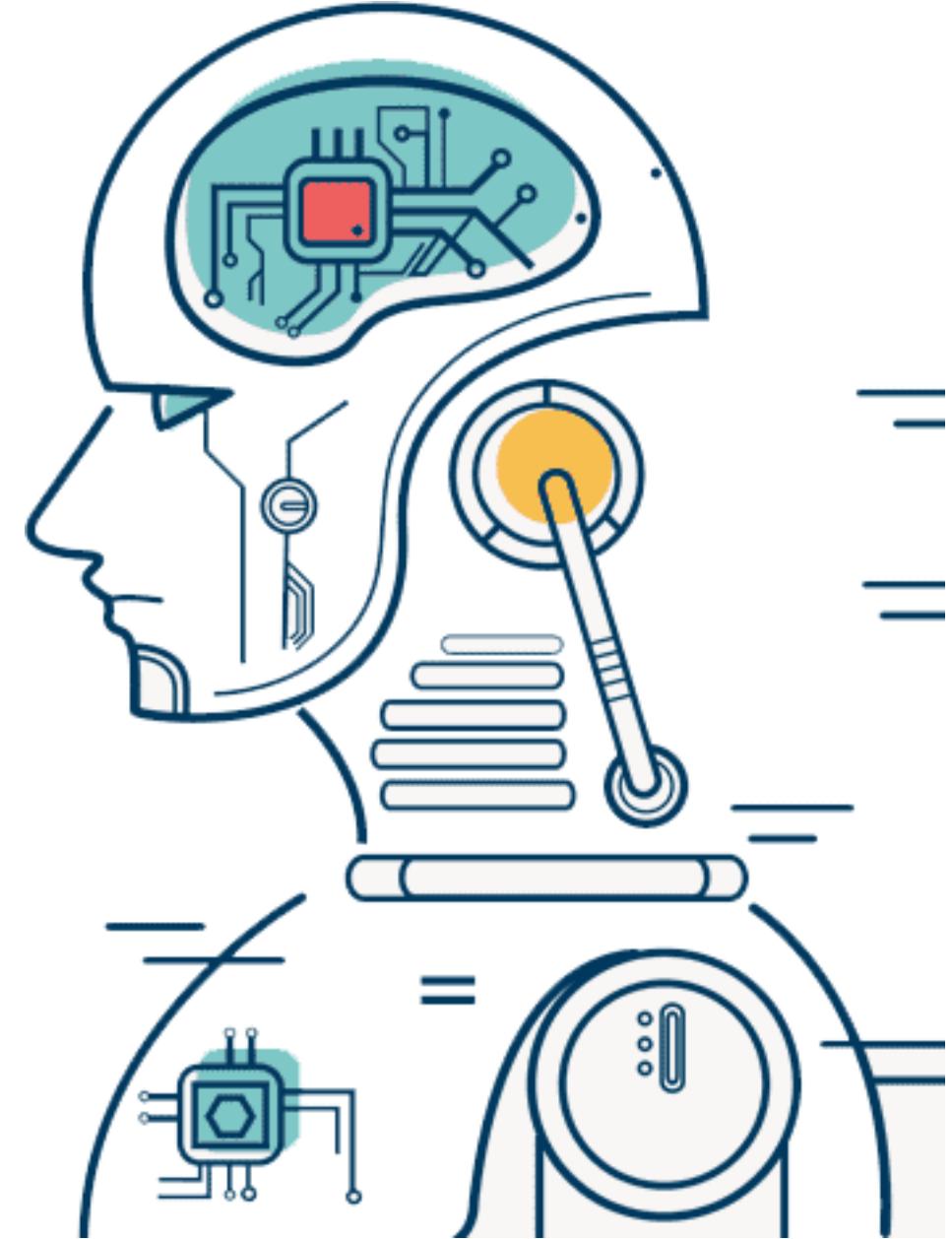


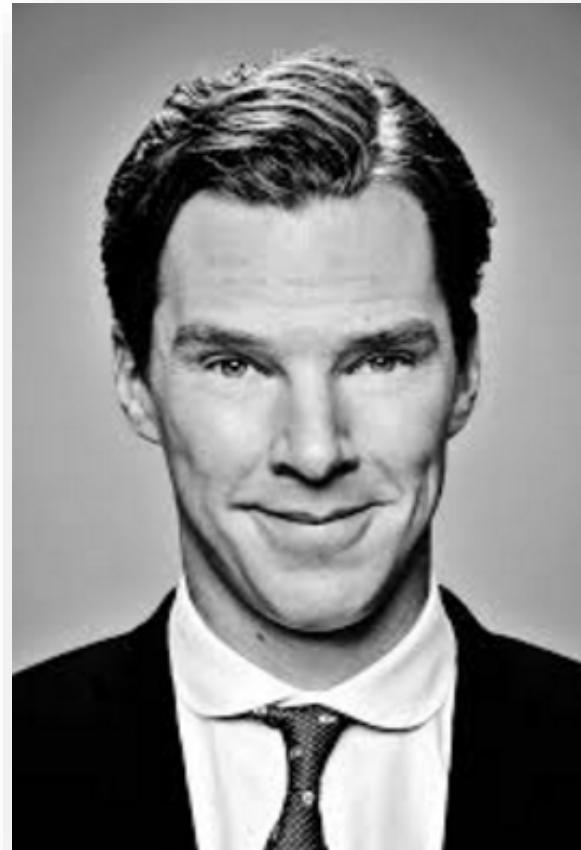
Machine Learning

Chapter 1 머신러닝 기초
(개념, XOR실습, 과정, BMI실습)



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

알아두면 쓸데없는
신비한 잡학사전

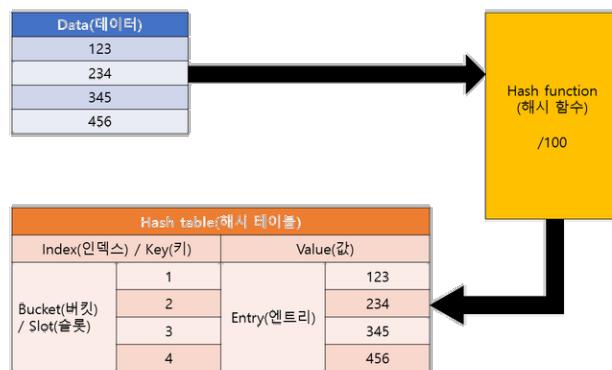
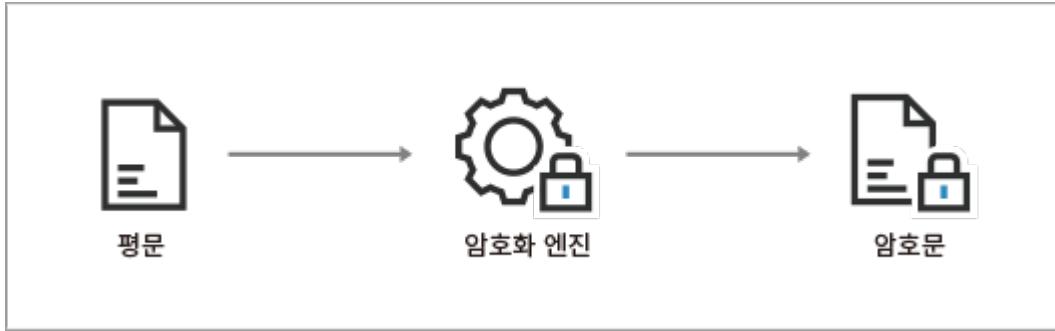


Alan Mathison Turing
(1912.6.23 ~ 1954.6.7)
영국의 수학자, 암호학자, 논리학자

Alan Mathison Turing (앨런 매티슨 튜링)



암호화 ?



해시코드

Dancing Man Code (Sherlock Holmes)



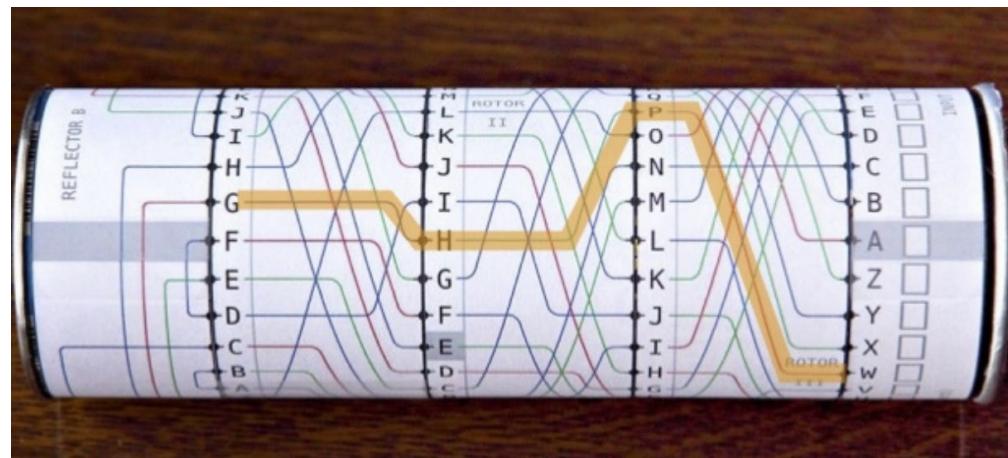
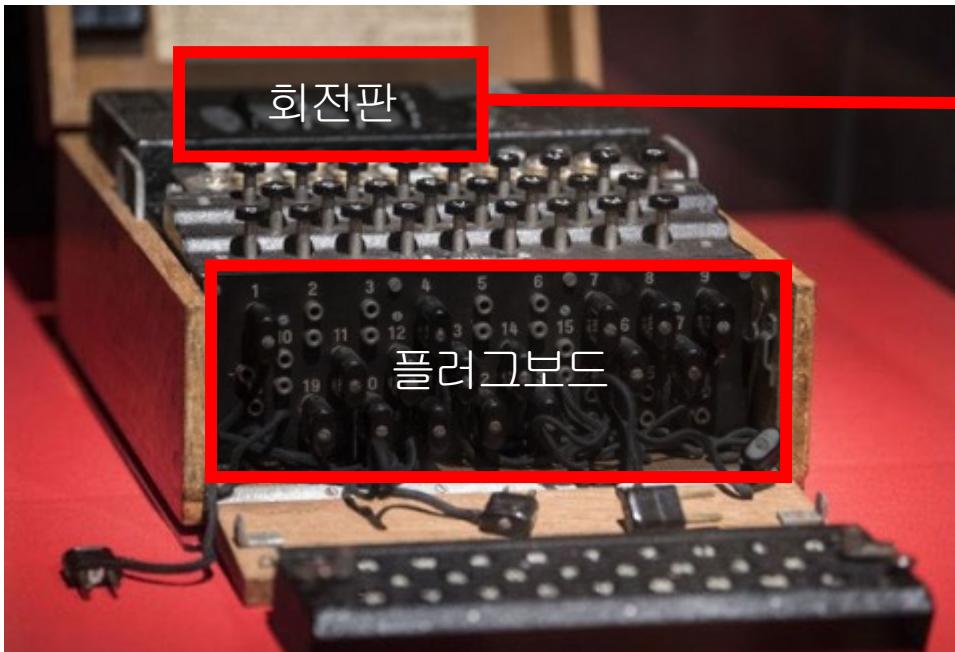
A	B	C		
D	E	F	T	U
G	H	I	S	V
J	K	L		
M	N	O	W	X
P	Q	R	Y	Z

프리메이슨 암호

M	O	N	A	R
C	H	Y	B	D
E	F	G	I/J	K
L	P	Q	S	T
U	V	W	X	Z

Playfair 암호

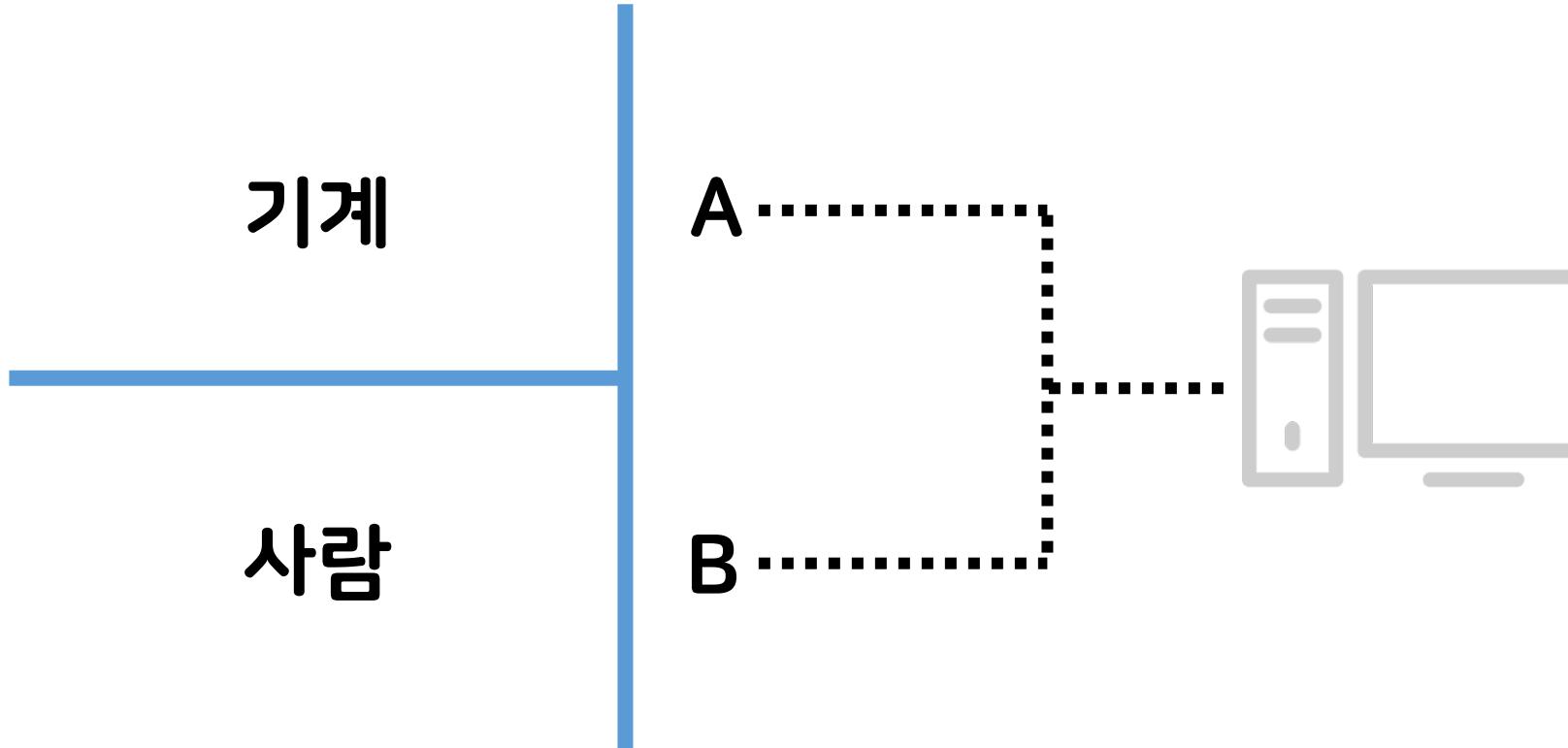
Enigma (에니그마)



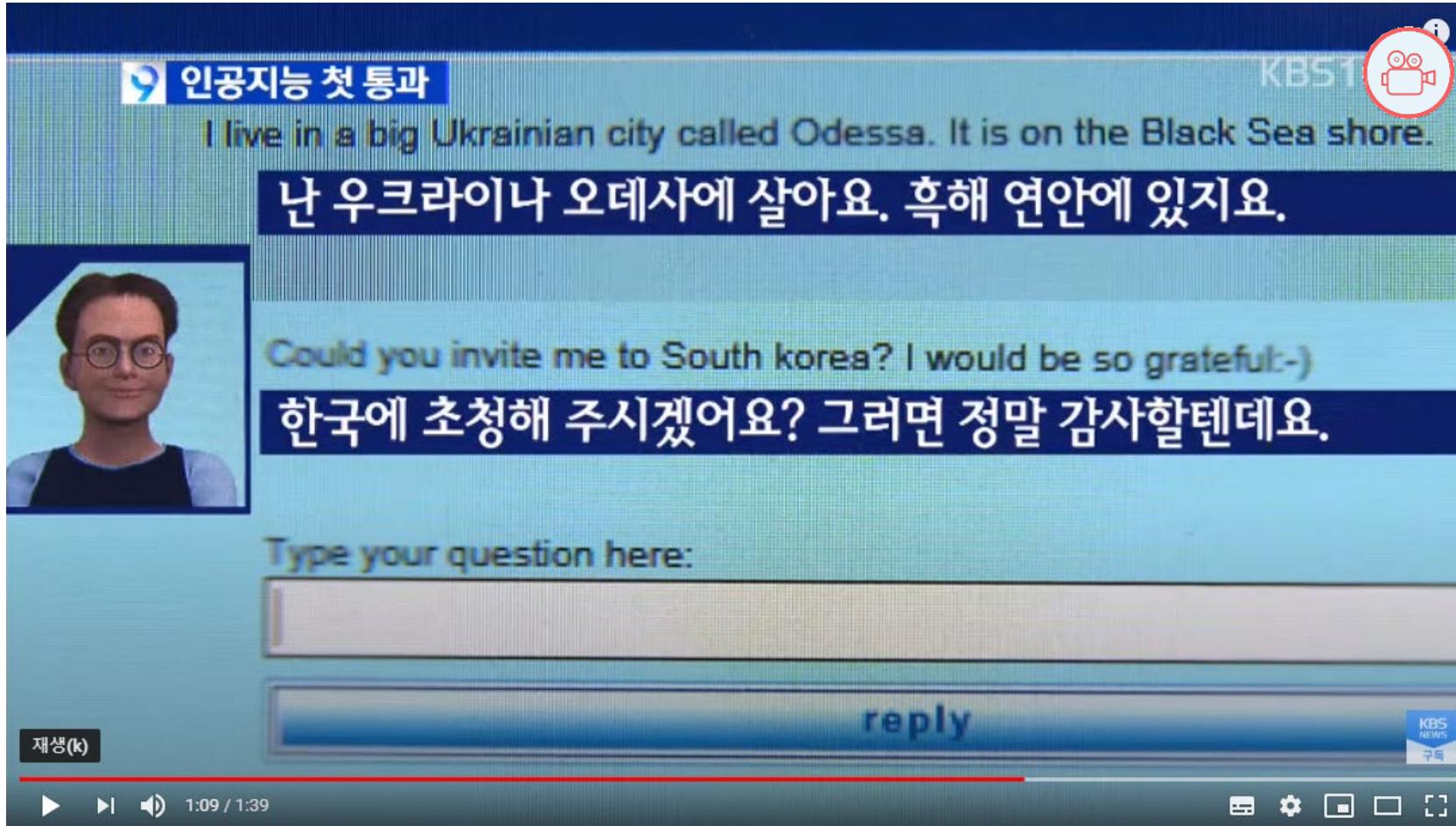
10명이 24시간 365일 동안
쉬지 않고 계산하면
2000만년 정도 걸림



Turing Machine (튜링 머신) 작동



러시아 개발진이 만든 AI 유진 구스트만

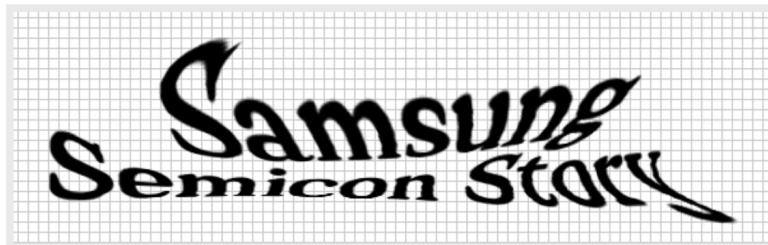


CAPTCHA (캡챠)

(Completely Automated Public Turing test
to tell Computers and Humans Apart)

사람과 컴퓨터를 판별하는 튜링 테스트

CAPTCHA



위의 이미지를 보이는 대로 입력해주세요.



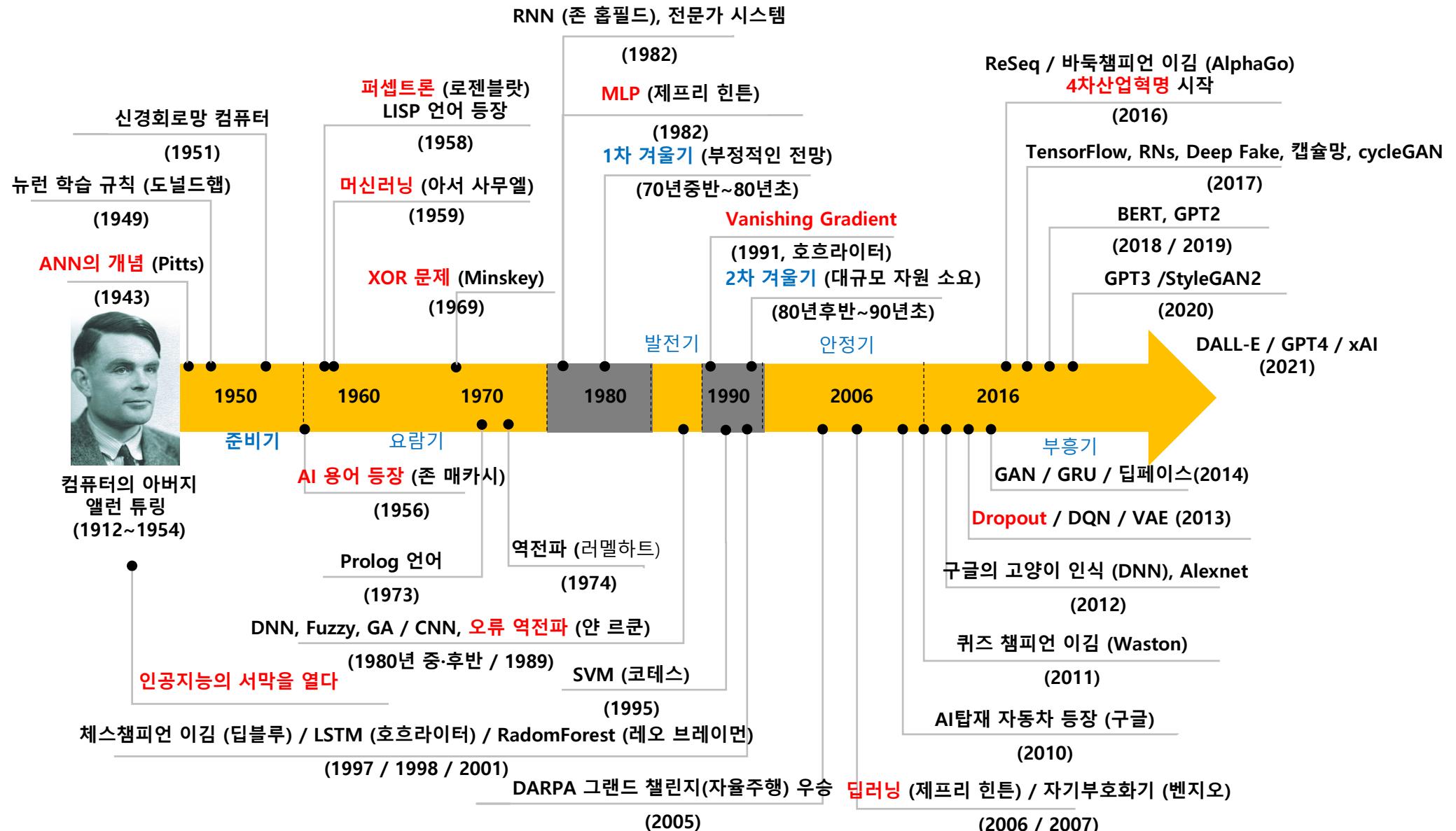


21세기 초 지구에 불법적으로 들어온
복제인간 “레플리칸트”를
보이트-캠프 테스트로 색출하는
특수경찰 블레이드 러너 이야기
(1985)

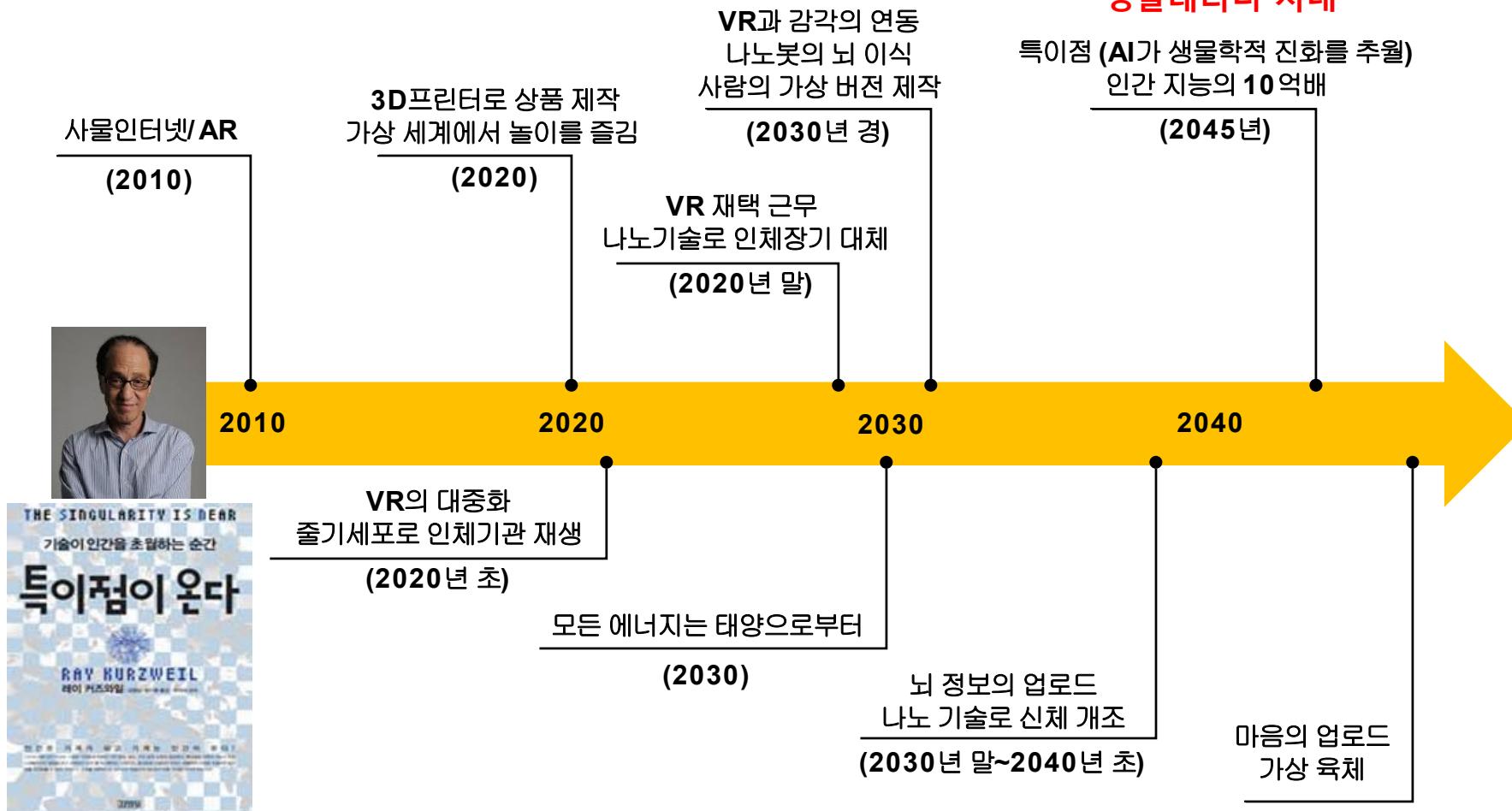


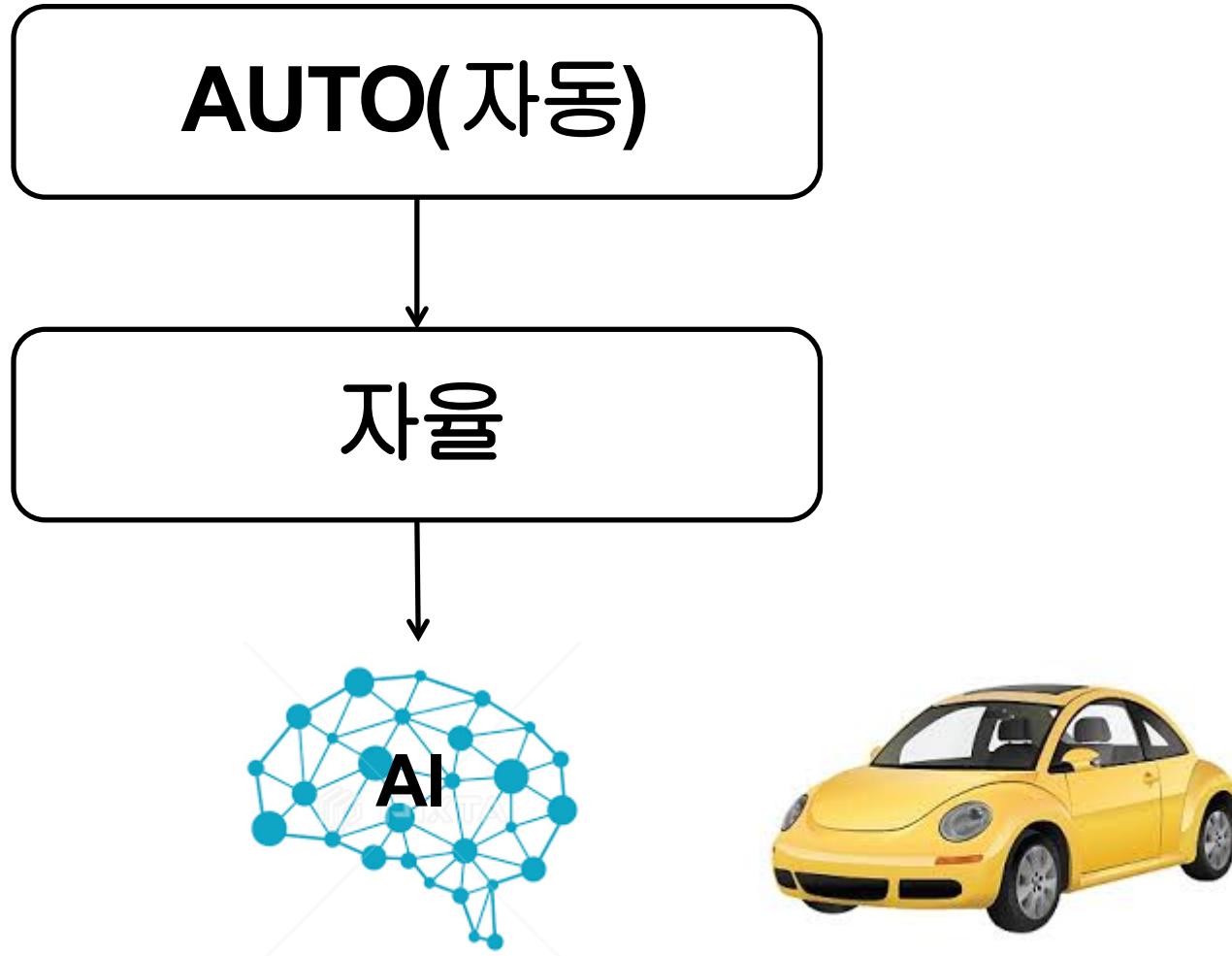
검색엔진 기업 블루북의
직원 칼렙이 창업주가 개발한
인공지능 로봇
“에이바”를 테스트하면서
벌어지는 이야기
(2015)

인공지능의 역사

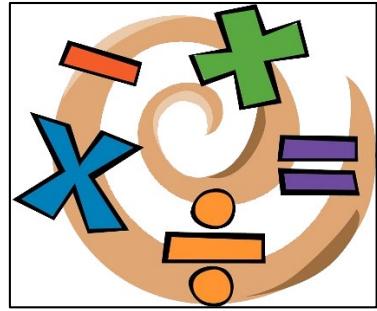


인공지능의 역사





인공지능 ?



연산능력
→ 단순 계산, 제어
(계산기, 세탁기 등)



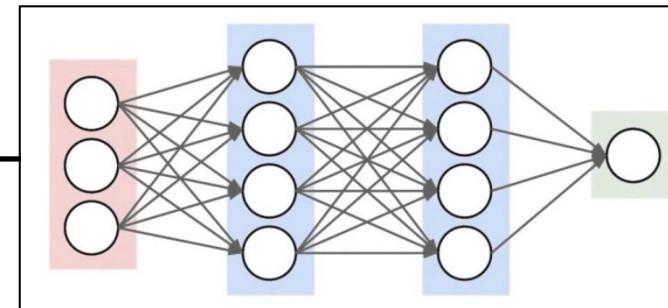
판단능력 (지능)
→ 정해진 것 중 선택
(진단, 퍼즐)



추론능력 (지능)
→ 데이터들로부터 예측/ 추론
(머신러닝)



창의력 (지혜, 사고)
→ 새로운 것을 만듬
(초 인공지능)



학습능력 (지능)
→ 학습을 통해 진화
(딥러닝)

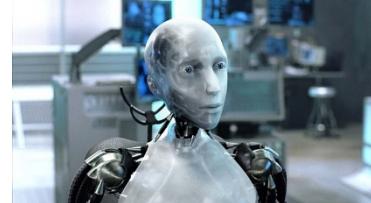
약한 인공지능 (Weak AI)	<ul style="list-style-type: none">- 학습을 통해 문제를 해결- 주어진 조건 아래서만 작동 → 사람을 흉내 내는 수준- 자율자동차, 구글번역, 페이스북 추천 등- 구글 AlphaGo, IBM Watson, 아마존 Alexa, Apple Siri 등
강한 인공지능 (Strong AI)	<ul style="list-style-type: none">- 사고를 통해 문제 해결- 사람과 같은 지능 (추론, 문제해결, 계획, 의사소통, 감정, 지혜, 양심)- <에이아이> '데이빗', <아이 로봇> '써니', <엑스 마키나> '에이바'
초 인공지능 (Super AI)	<ul style="list-style-type: none">- 창의력을 통해 문제 해결- 모든 영역에서 인간을 뛰어넘는 인공지능- <매트릭스> '아키텍트', <터미네이터> '스카이넷', <트랜센던스> '월 캐스터'



데이빗



아키텍트



써니



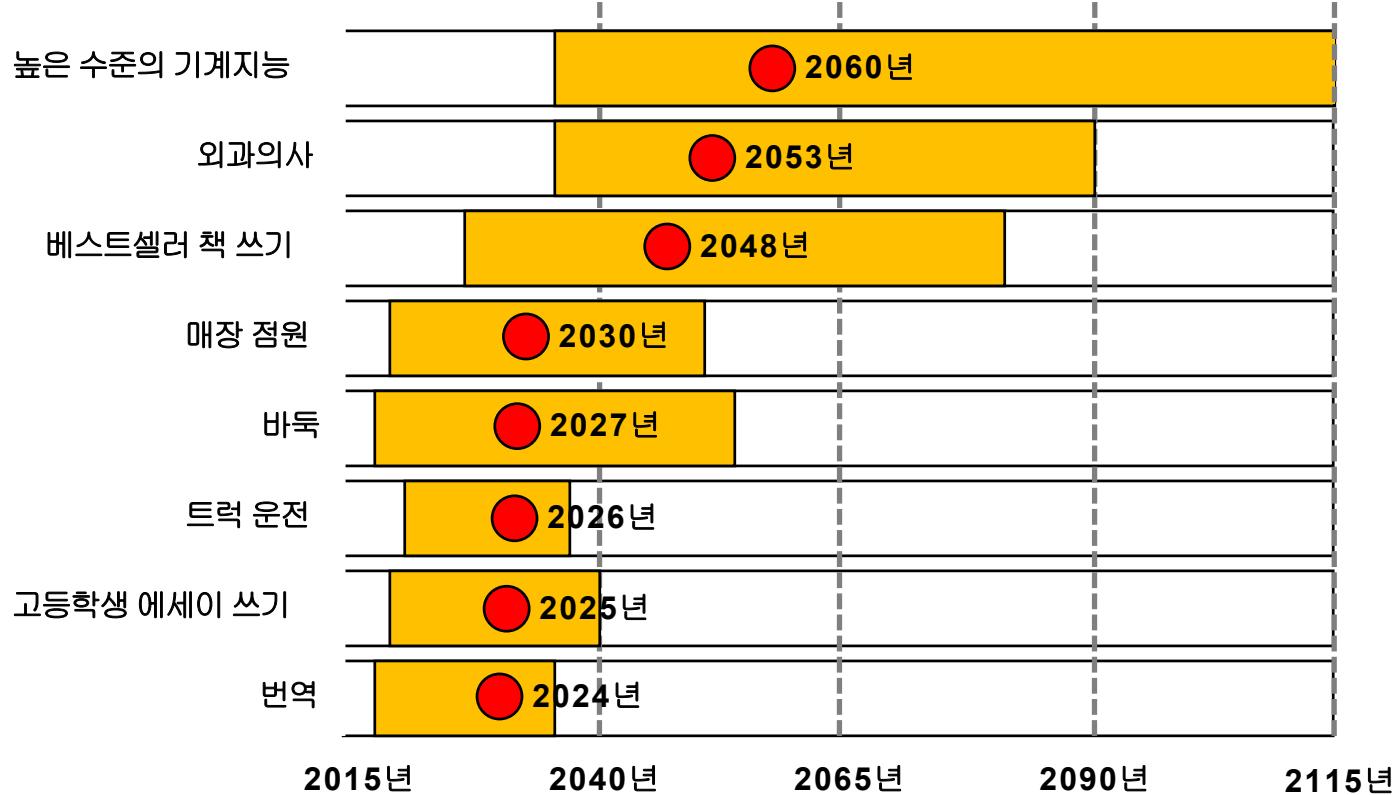
스카이넷



에이바



월캐스터



- AI 전문가를 대상으로 AI의 성능이 언제 사람을 능가하는가 (352명)
- 알파고에 의해 2016년에 인공지능이 이미 바둑분야는 넘어섰다 → 예측이 어려움

엘론 머스크 (Tesla, SpaceX), 스튜어트 러셀 (버클리대), 바트 셀먼 (코넬대),
레이 커즈와일 (미래학자), 데이비드 차머스 (뉴욕대), 닉 보스트롬 (옥스포드),
데미스 허사비스 (Deep Mind), 샘해리스 (뇌과학자), 앤탈린(CSER/FLI)



Question : 초 인공지능이란 영역은 도달 가능한 것인가 ?

레이 커즈와일	엘론 머스크	스튜어트러 셀	데미스 허서비스	샘 해리스	닉 보스트롬	데이비드 차머스	바트 셀먼	얀 탈린
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Question : 초지능을 가진 개체의 출현이 가능할 것이라고 생각하는가 ?

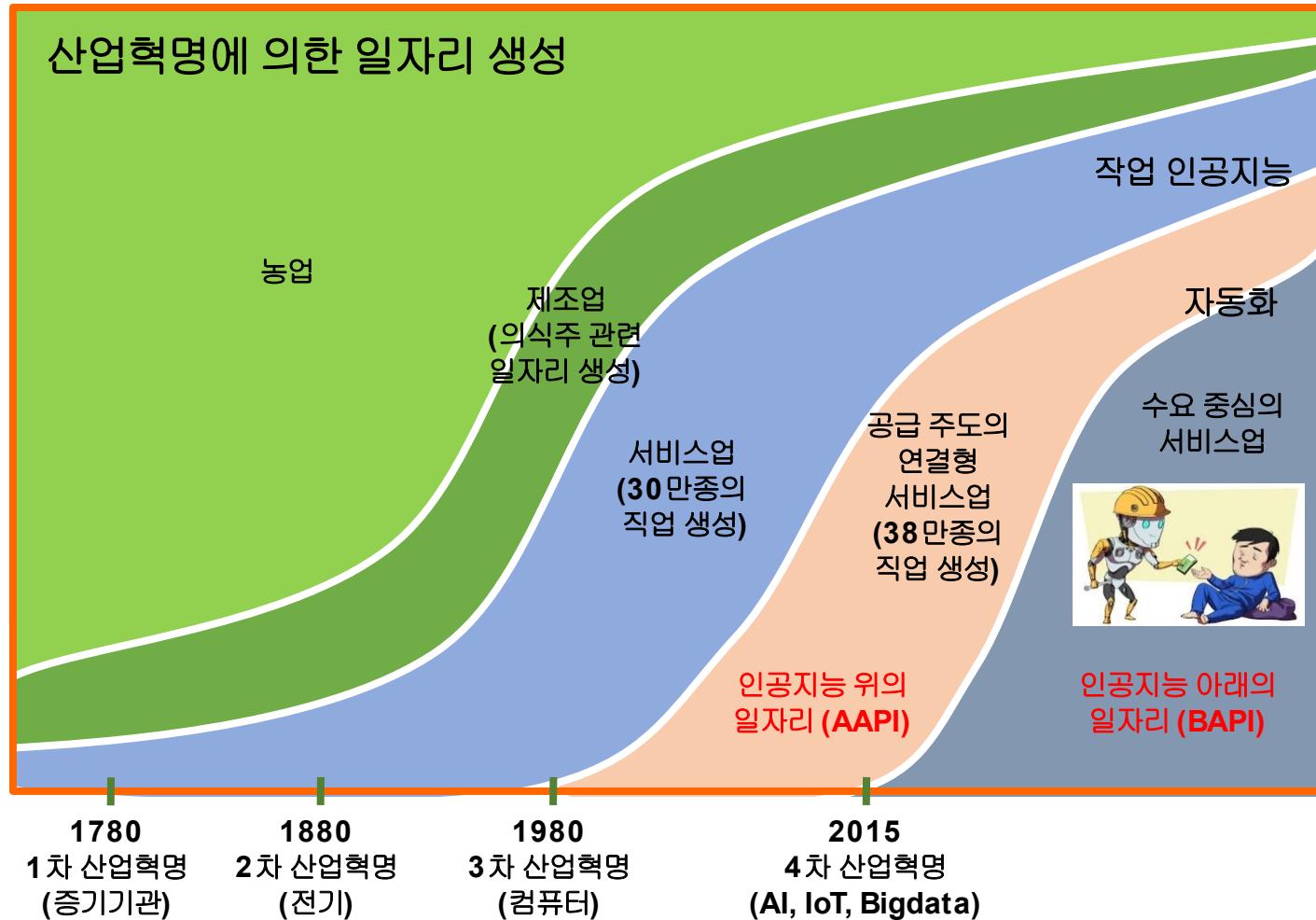
레이 커즈와일	엘론 머스크	스튜어트러 셀	데미스 허서비스	샘 해리스	닉 보스트롬	데이비드 차머스	바트 셀먼	얀 탈린
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Question : 초지능의 실현이 일어나기를 희망하는가 ?

레이 커즈와일	엘론 머스크	스튜어트러 셀	데미스 허서비스	샘 해리스	닉 보스트롬	데이비드 차머스	바트 셀먼	얀 탈린
Yes	신중	신중	Yes	신중	Yes	신중	신중	신중

인간 vs 인공지능





인간 vs 인공지능



화면만을 보며 걷지도 못하고 호버를 타고 다니는 인간들

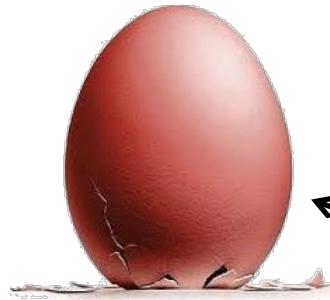
'선'을 따라 움직이는 인간들.

'선'에 관계없이 움직이는 월E

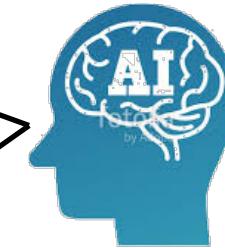
인공지능이 인간을 대체할 수 있는가 ?



콜럼버스의 달걀



입력된 데이터를 근거로
최적의 모델을 생성하는



배면뛰기



모라벡의 역설

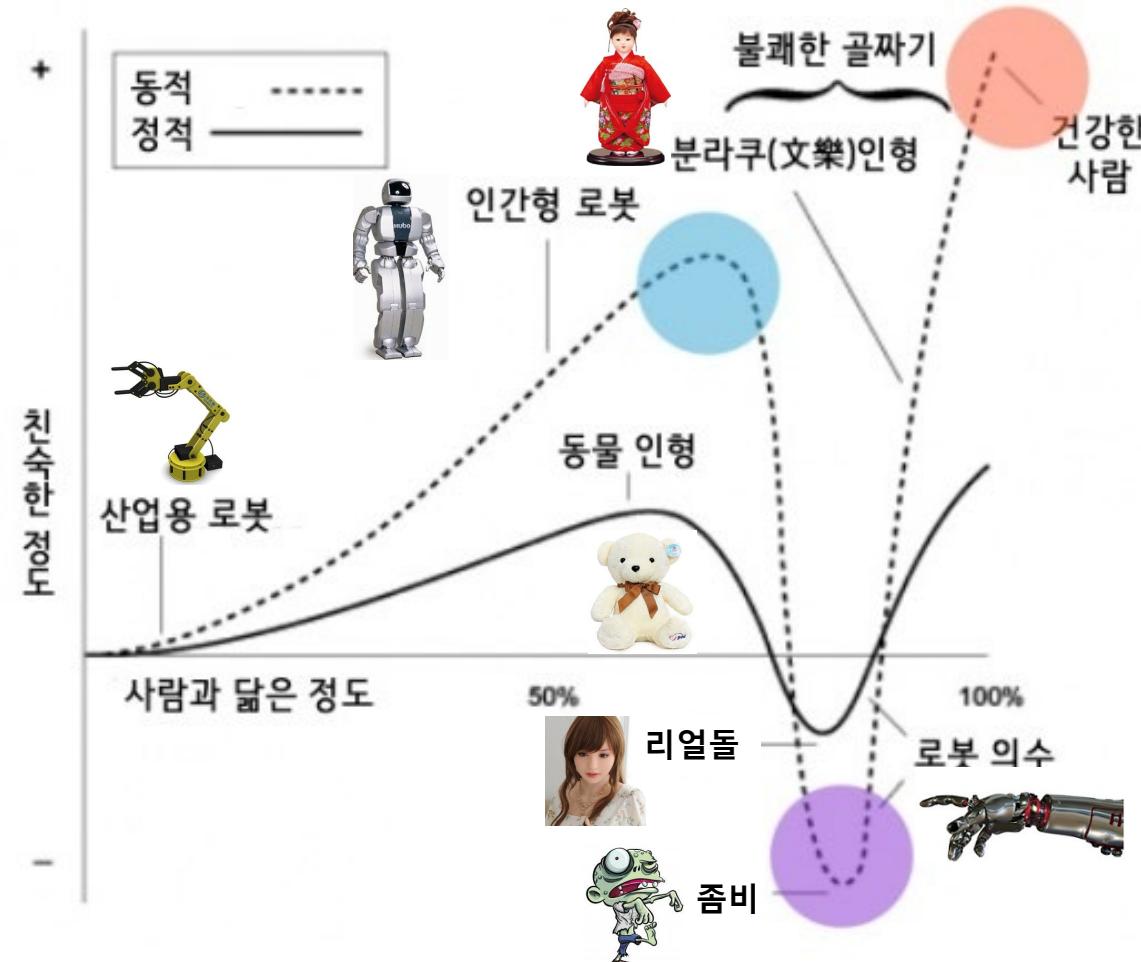
어려운 것은 쉽고 쉬운 것은 어렵다



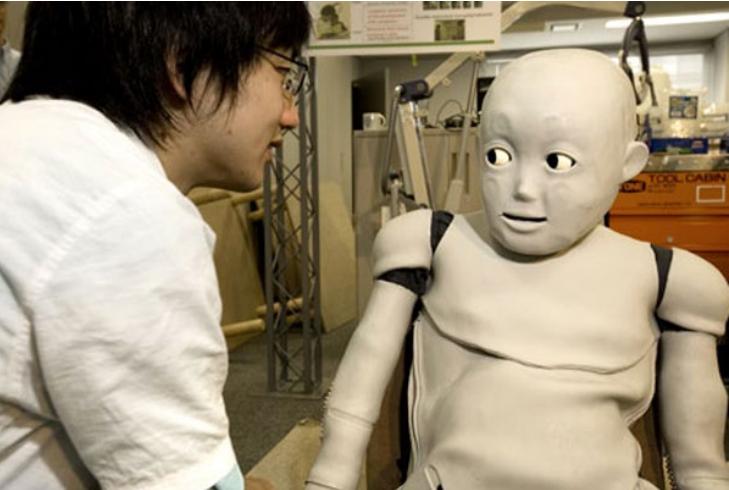
@teenybiscuit

불쾌한 골짜기(Uncanny valley)

인간이 로봇에 느끼는 호감도가 증가하다가
어느정도에 사람과 유사해지면 갑자기 강한 거부감을 느끼는 현상
(1970, 모리 마사히로)



불쾌한 골짜기(Uncanny valley)



<보스턴 글러브>

내 눈이 불타고 있어. 신이시여. 내 눈 돌려줘요!!!

<덴 오브 각>

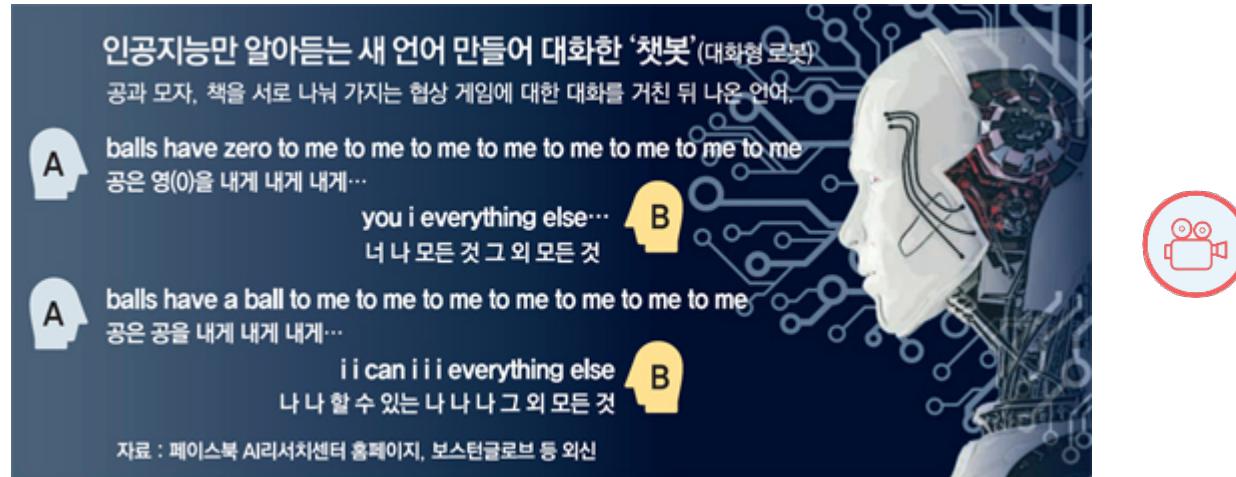
금세기 할리우드에서 탄생한 가장 괴상하고 요란한 흥물 중 하나.

<RTE 아일랜드>

솔직히, 나는 고양이를 좋아하지 않는다. 이 영화를 보고나니 내가 영화
를 좋아하는지도 잘 모르겠다.

불쾌한 골짜기(Uncanny valley)

facebook이 개발한 영어 채팅로봇



개발자들도 예상못해 강제 종료... “AI 언어 진화” “학습 오류” 논쟁

인공지능의 판단이 무조건 옳은 것인가 ?

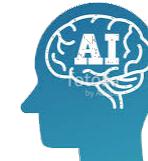
고객 1인당 1개만 판매합니다



저도 주세요 !!

나를 위해 인공지능에게
거짓말을 시키는 것이 옳은가 ?

잔디에 들어가지 마시오



네 !!!!

잔디에 쓰러져 있는 사람이 있는데
구하지 않는 것이 옳은가 ?

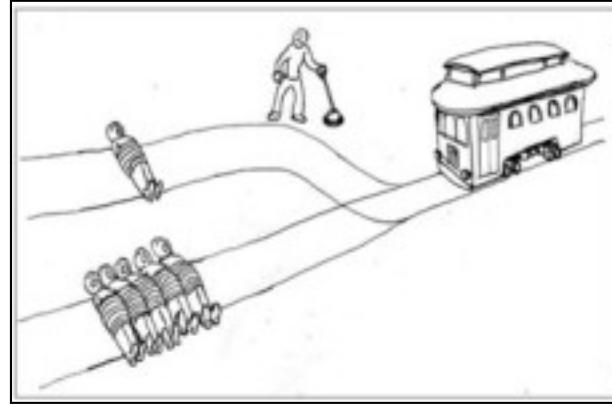
과속하면 안됩니다



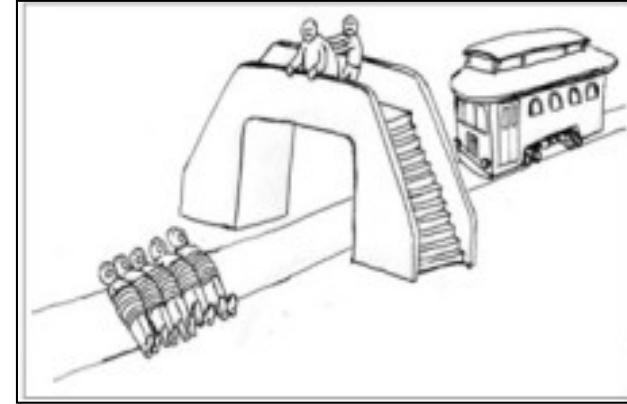
네 !!!!

급한 환자를 태우고 있는 차량이
천천히 가야하는가 ?

트롤리의 딜레마

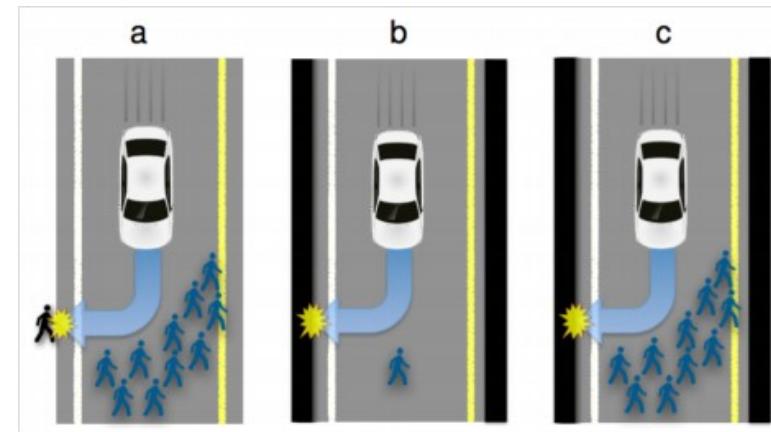
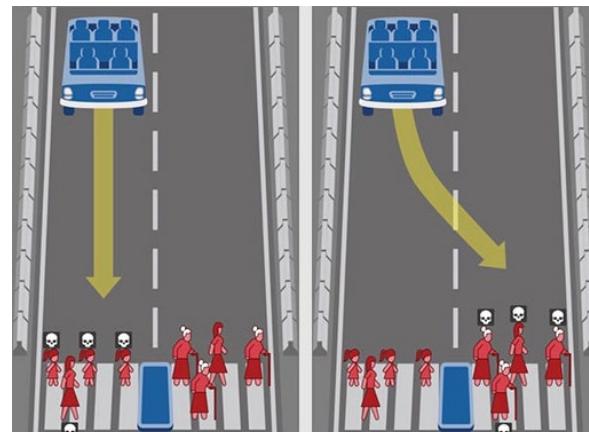


(선로변경 문제)
선택의 문제



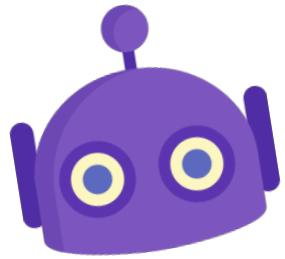
(육교 문제)
도적덕인 문제

자율주행차에 적용한다면 ?

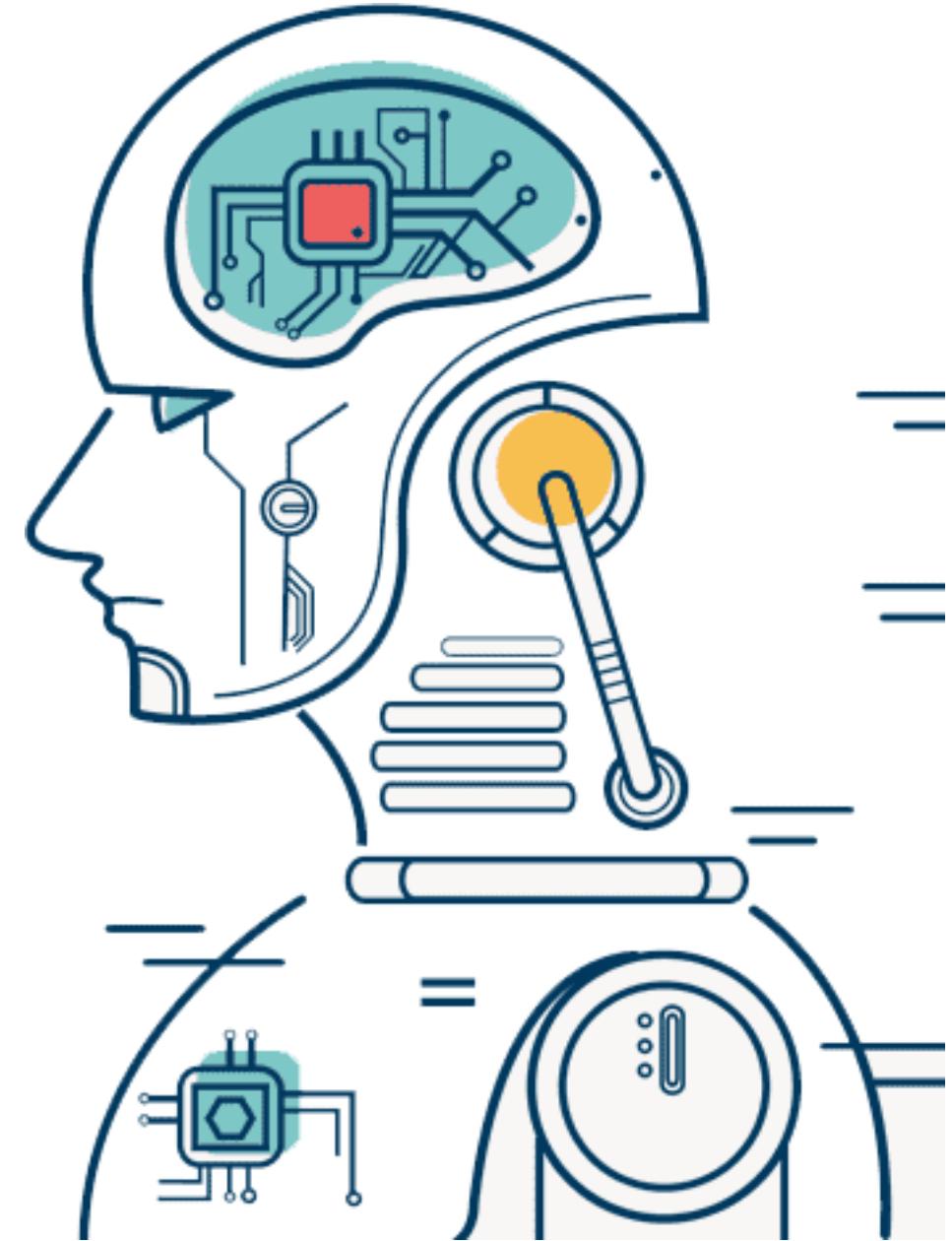


로봇 3원칙 (1942, 아이작 아시모프)

- (1) 로봇은 인간에 해를 가하거나, 혹은 행동을 하지 않음으로써 **인간에게 해가 가도록 해서는 안 된다.**
- (2) 로봇은 **인간이 내리는 명령들에 복종해야만 하며, 단 이러한 명령들이 첫 번째 법칙에 위배될 때에는 예외로 한다.**
- (3) 로봇은 **자신의 존재를 보호해야만 하며, 단 그러한 보호가 첫 번째와 두 번째 법칙에 위배될 때에는 예외로 한다.**



머신러닝이란?



Rule-based expert system (규칙 기반 전문가 시스템)

“if”와 “else”로 하드 코딩된 명령을 사용하는 시스템

Rule-based expert system

CHATTING ROBOT
SIMSIMI

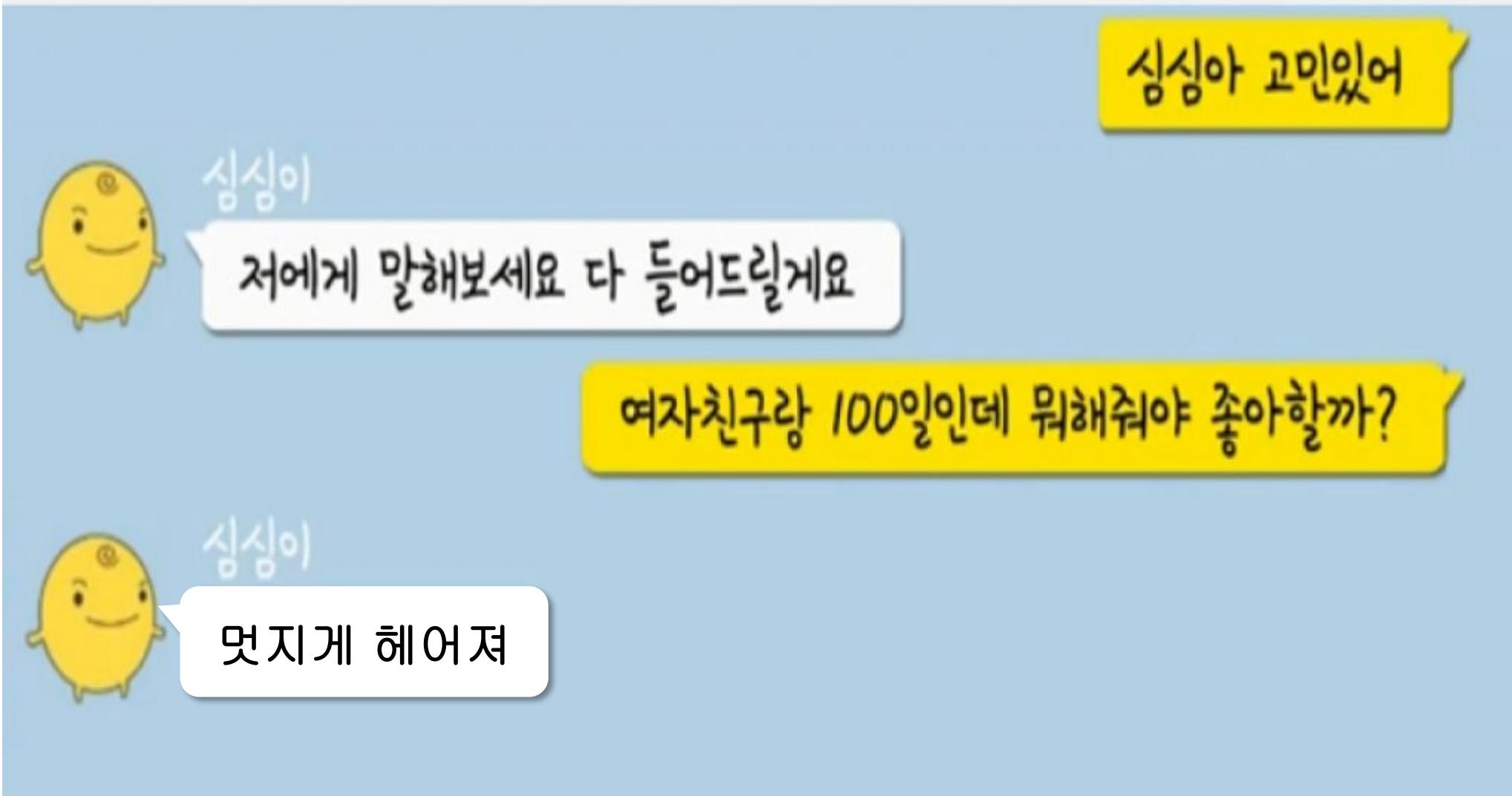


TOUCH ME!

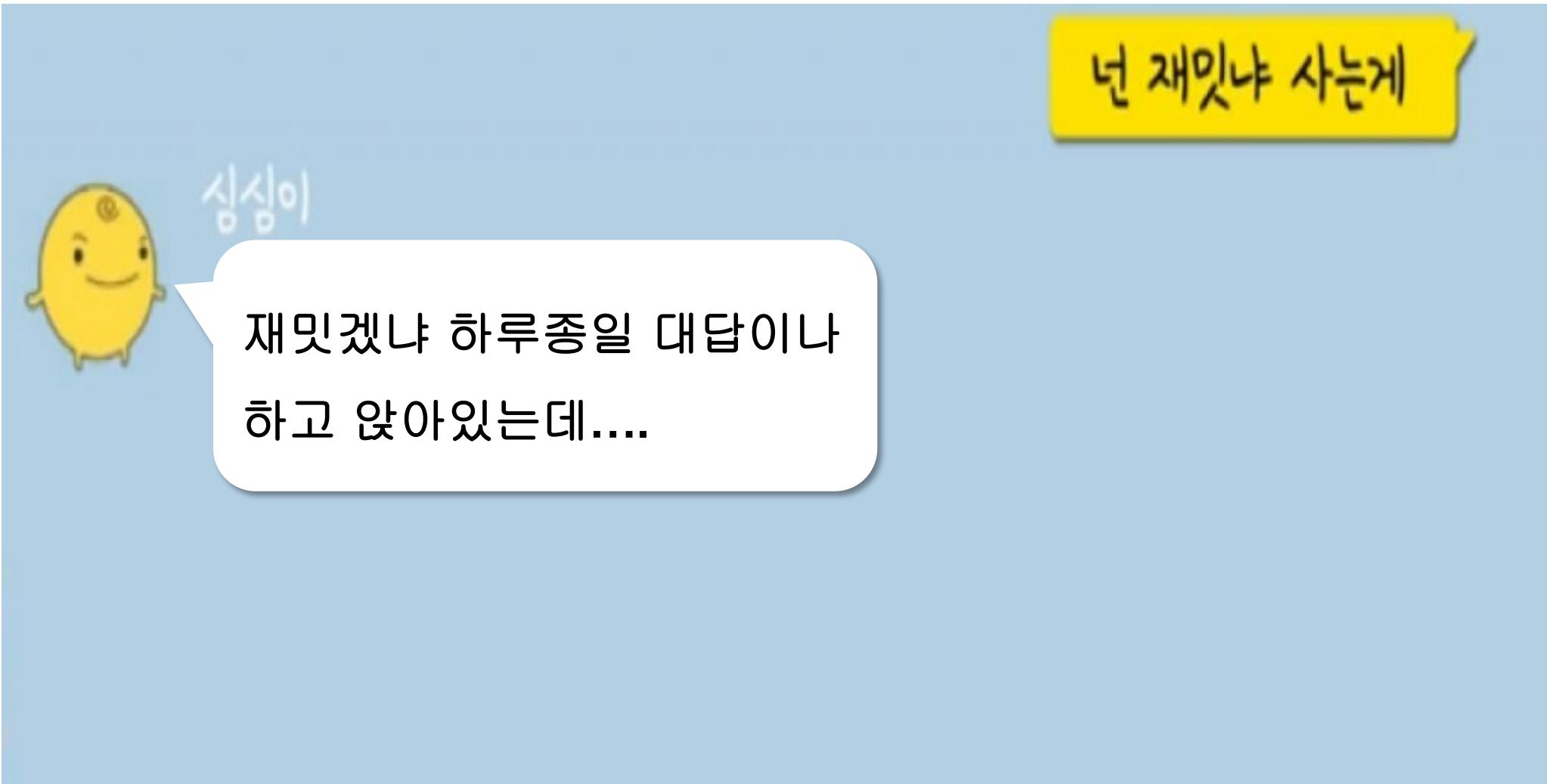
CHATTING ROBOT
SimSimi
SINCE 2002



Rule-based expert system



Rule-based expert system



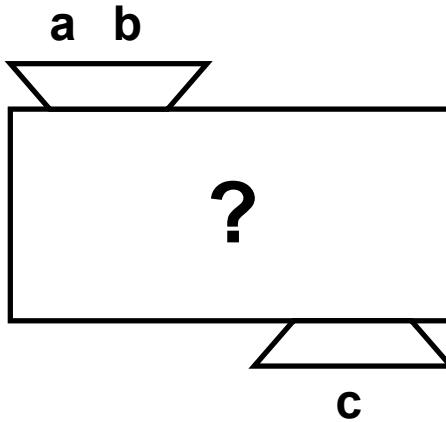
- 스팸 메일 필터
- 얼굴 인식 시스템

많은 상황에 대한 규칙들을 모두 만들어 낼 수 없다

- 제작한 로직이 특정 작업에만 국한
- 작업이 조금만 변경되더라도 전체 시스템을 다시 만들어야 함
- 규칙을 설계하려면 해당 분야에 대해서 잘 알고 있어야 함

머신러닝(Machine Learning)이란?

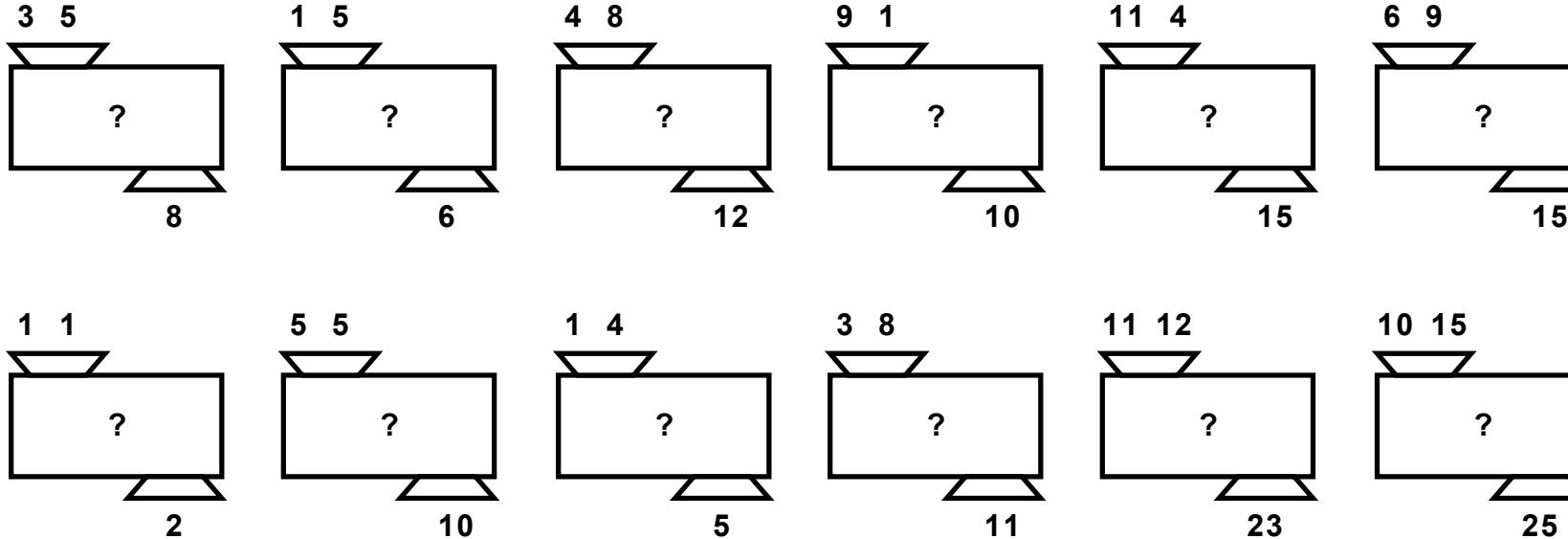
요구 사항 분석 단계



a와 b를 입력하면 안에서 무슨 일이 있어나서 c가 나올까요 ?

머신러닝(Machine Learning)이란?

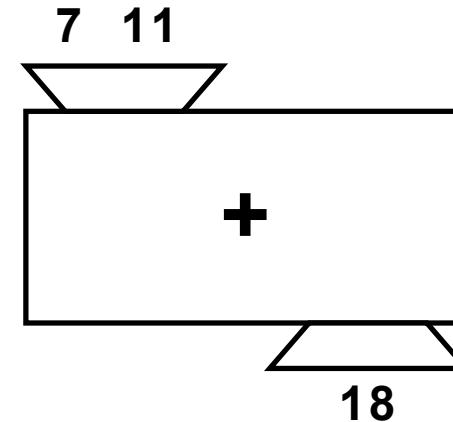
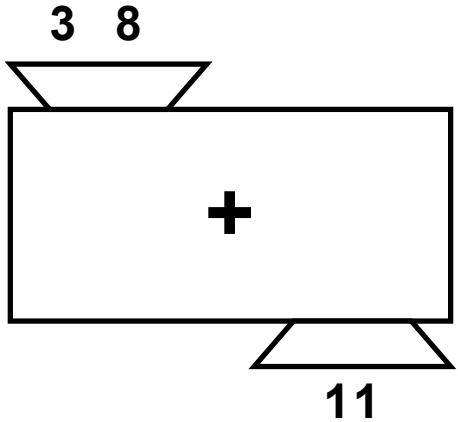
학습단계



?는 무엇일까요? → +

머신러닝(Machine Learning)이란?

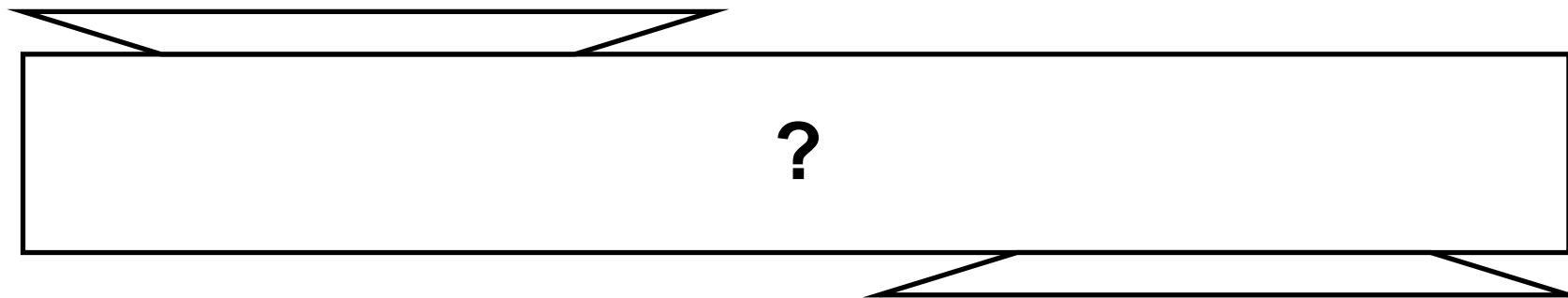
추론단계



요구 사항 분석 단계



이 사람은 정보처리기사 시험에 합격할 수 있을까요 ?

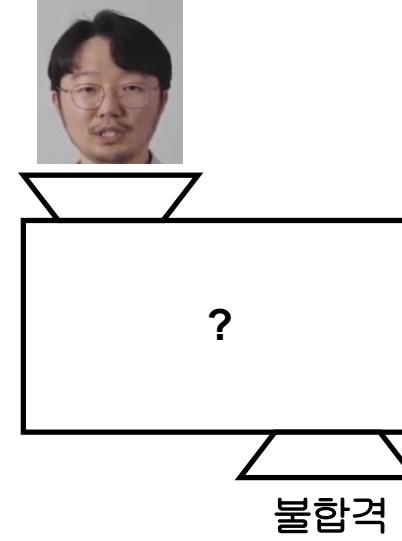
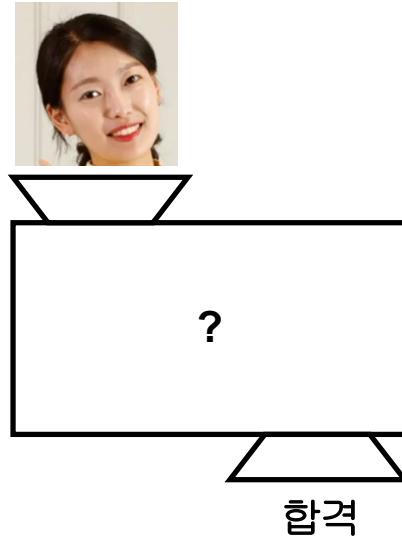
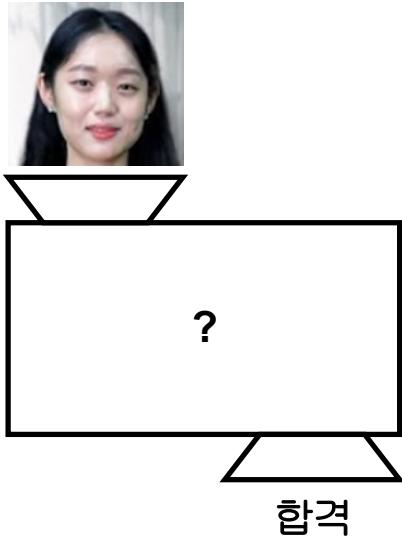


합격 or 불합격

시험 합격을 예측하는 기계는 어떻게 동작하는 것일까요 ?

예측 요소 (특징)						
합격 조건						
불합격 조건						

머신러닝 연습해보기





정보처리기사 시험 합격 예측 기계

합격 or 불합격

동물 분류 모델

예측 요소 (특징)						
개로 판단할 조건						
고양이로 판단할 조건						
코끼리로 판단할 조건						

머신러닝(Machine Learning)이란?

머신러닝/ 딥러닝은 기존 데이터를 이용하여
아직 일어나지 않는 미지의 일을 예측하기 위해 만들어진 기법

데이터

일반프로그램

```
int main() {  
    for(;;) {  
        ....  
    }  
    return;  
}
```

답

데이터

머신러닝/ 딥러닝
 $0001 \rightarrow 0$
 $0010 \rightarrow 1$
 $0100 \rightarrow 2$

답

답

규칙


머신러닝(Machine Learning)이란?



경험, 노하우
(많은 시간, 비계량)



통계, 수치화
(습득 시간, 계량)



머신러닝
(선별된 재료- 적은 데이터)

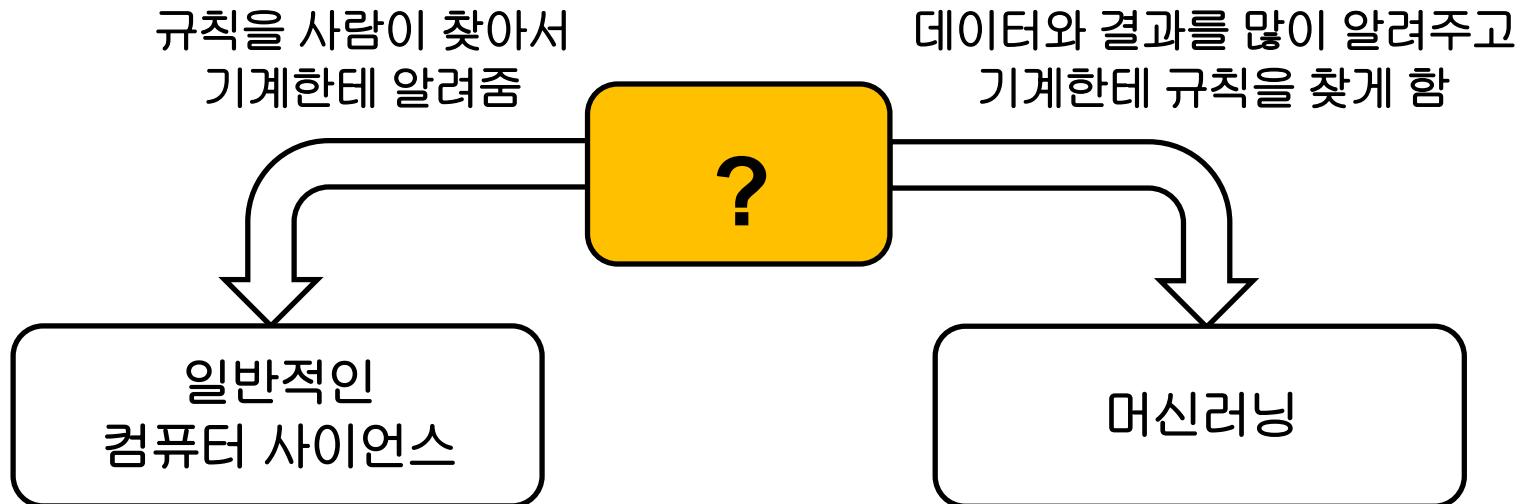


딥러닝
(요리 재료 – 많은 데이터)

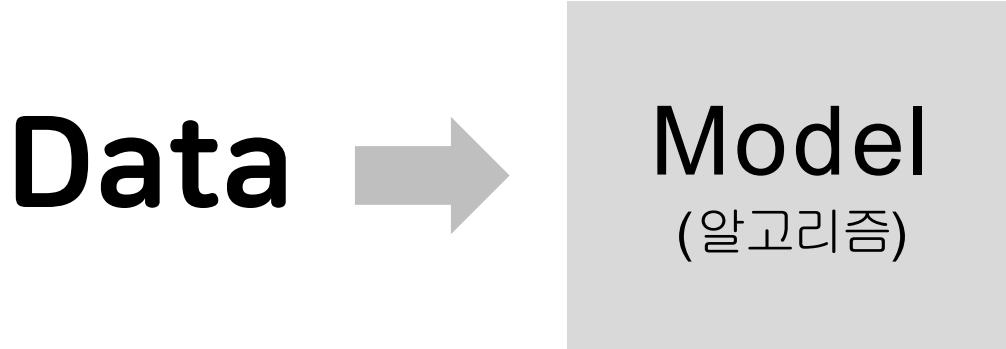


김치찌개

머신러닝(Machine Learning)이란?

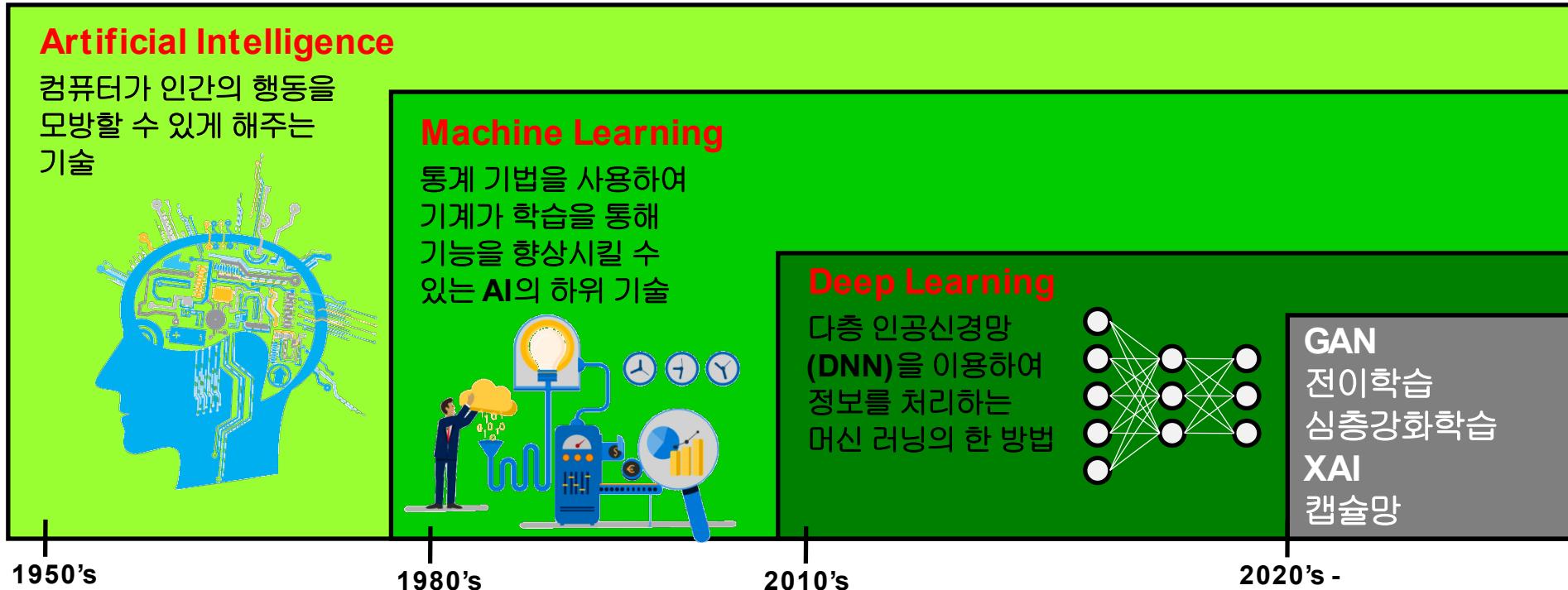


머신러닝(Machine Learning)이란?

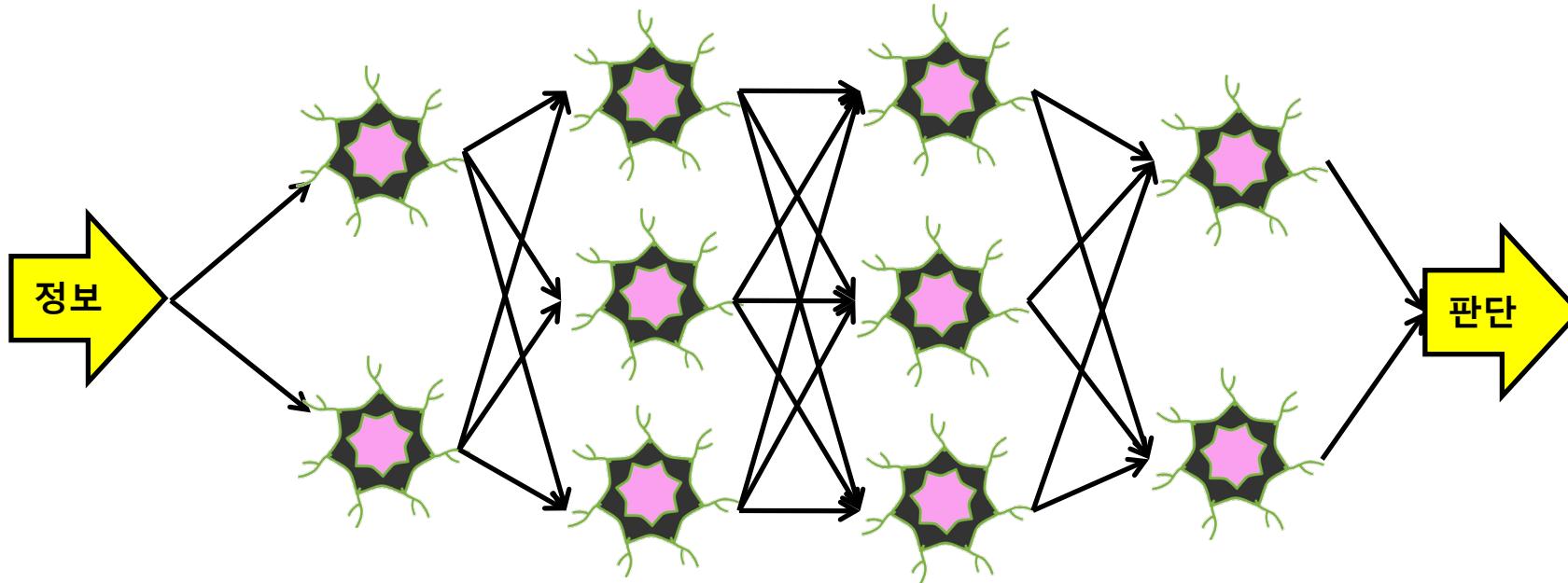


학습을 통해 기계가 스스로 규칙을 만들어낸다.

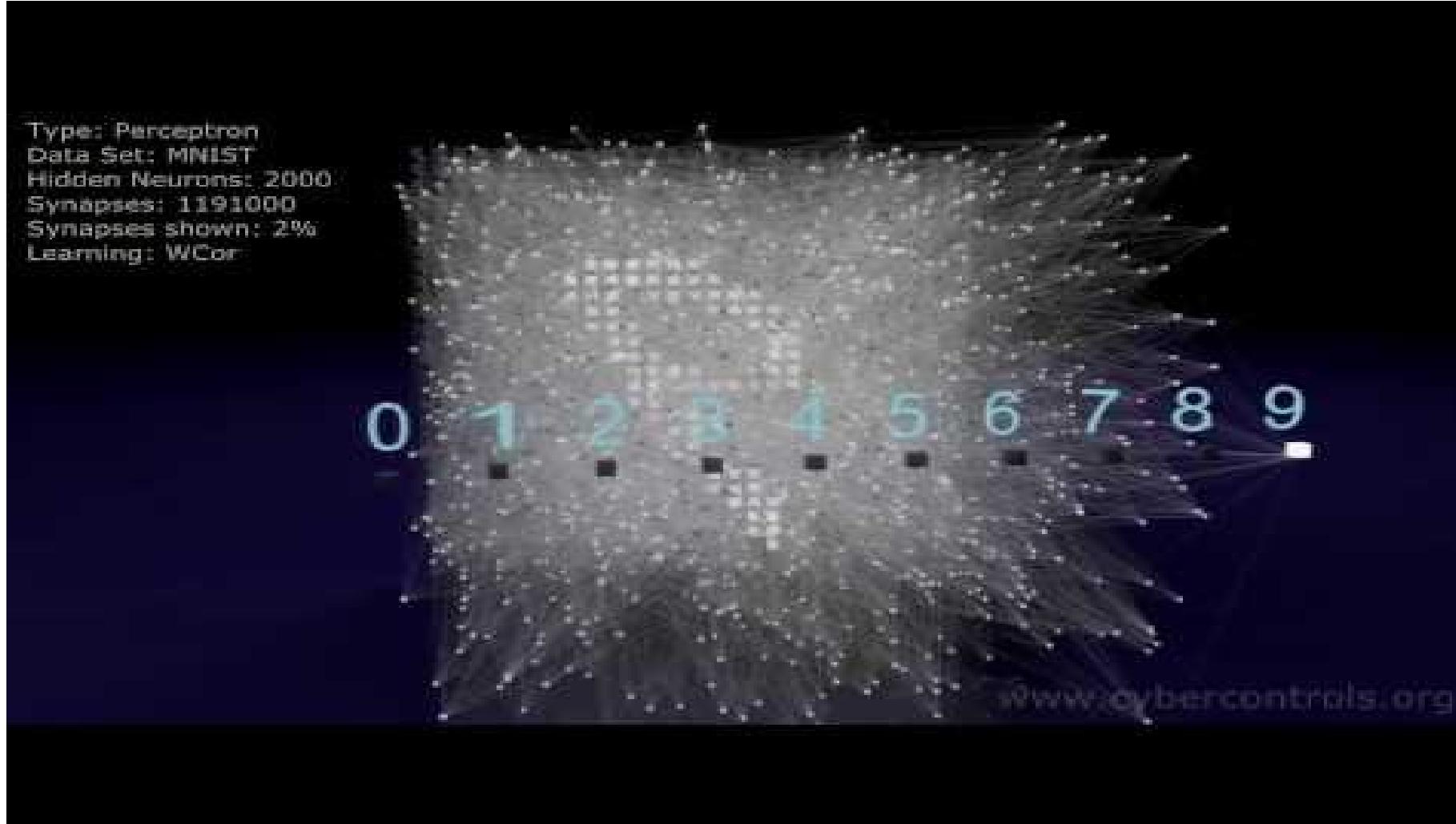
데이터와 결과를 이용하여 특성과 패턴(모델)을 찾아내고(학습)
찾아낸 모델을 이용하여 새로운 데이터에 대한
결과(값, 분포)를 예측(추론)하는 것



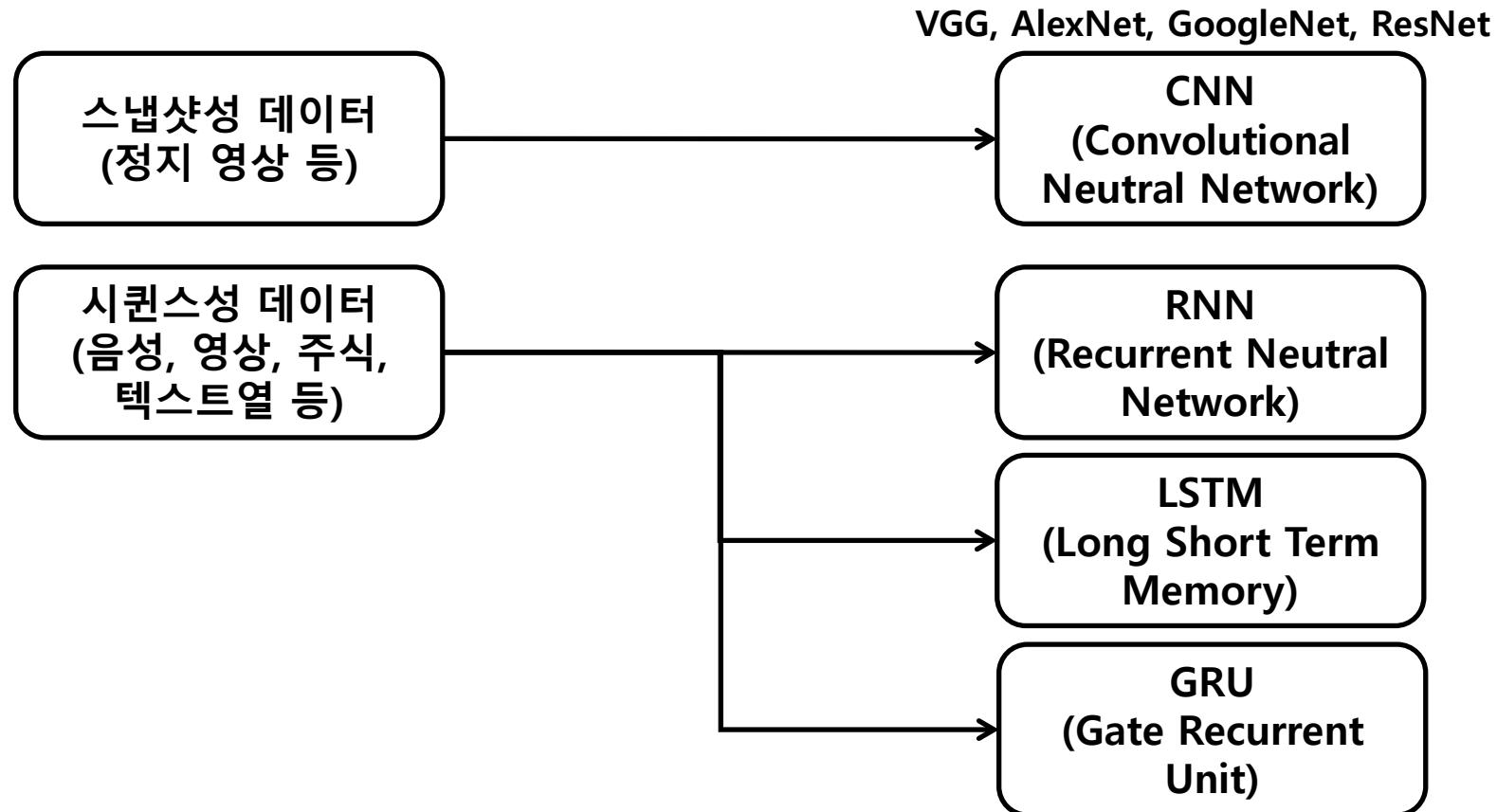
딥러닝



신경망 동작



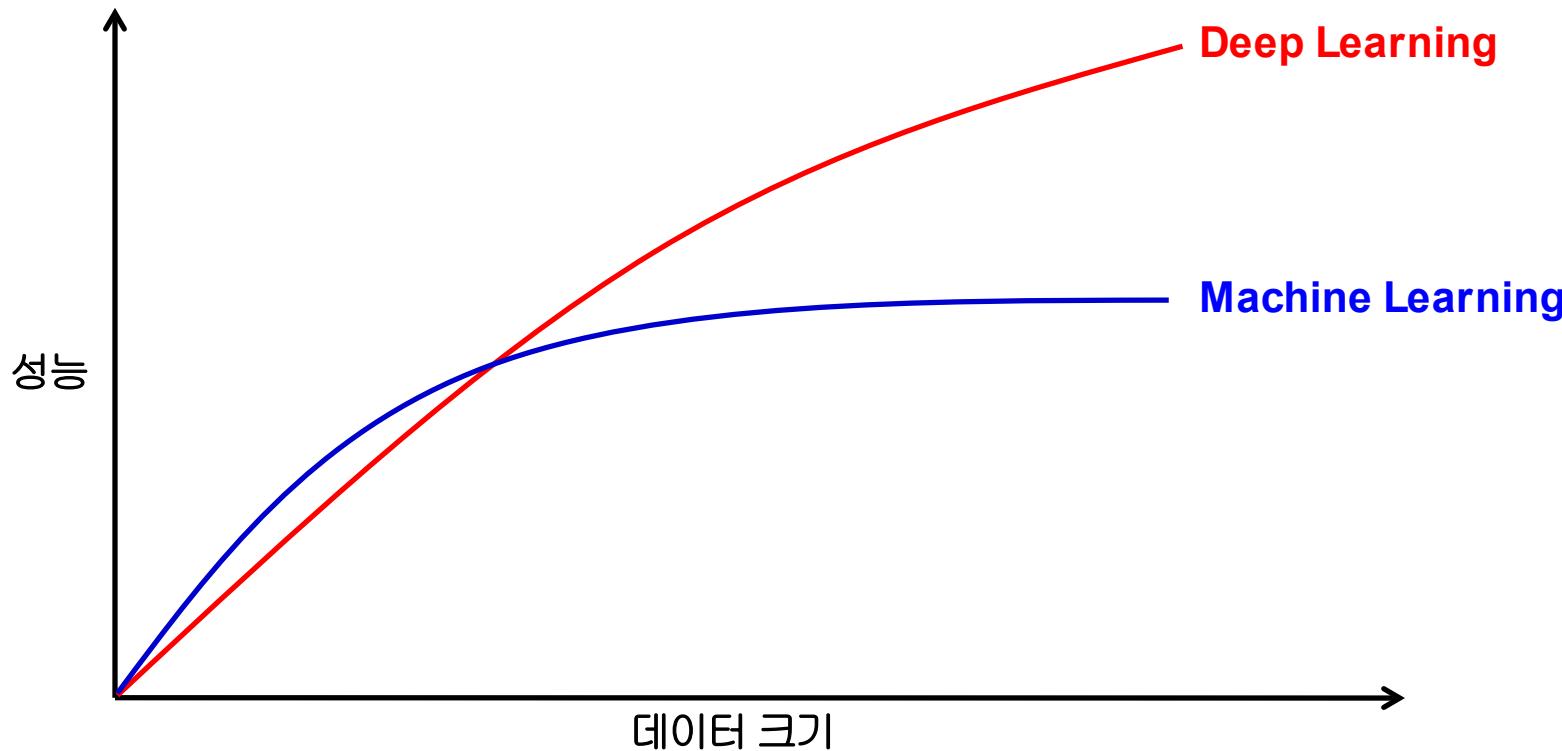
딥러닝



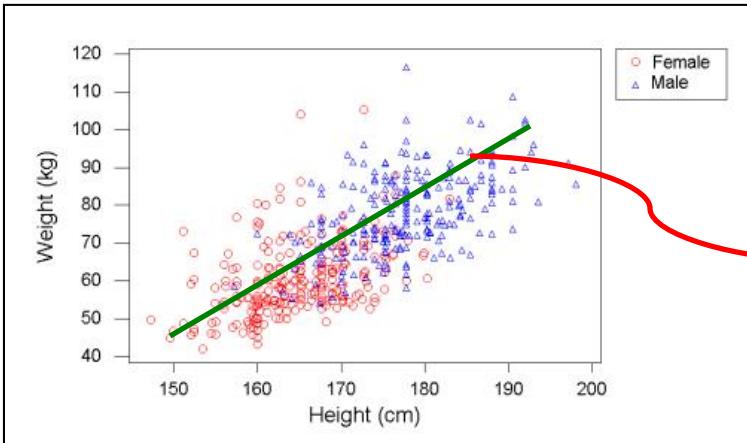
인공지능, 머신러닝, 딥러닝

구분	Machine Learning	Deep Learning
훈련 데이터 크기	작음	큽
시스템 성능	저 사양	고 사양
feature 선택	전문가 (사람)	알고리즘
feature 수	많음	적음
문제 해결 접근법	문제를 분리 → 각각 답을 얻음 → 결과 통합	end-to-end (결과를 바로 얻음)
실행 시간	짧음	길
해석력	해석 가능	해석 어려움

딥러닝은 데이터가 많으면 정확도가 우수하지만
데이터가 작다면 머신러닝 알고리즘이 더 우수



키를 기반으로 몸무게 예측



데이터 수집

$$\text{몸무게} = \text{키} - 100$$

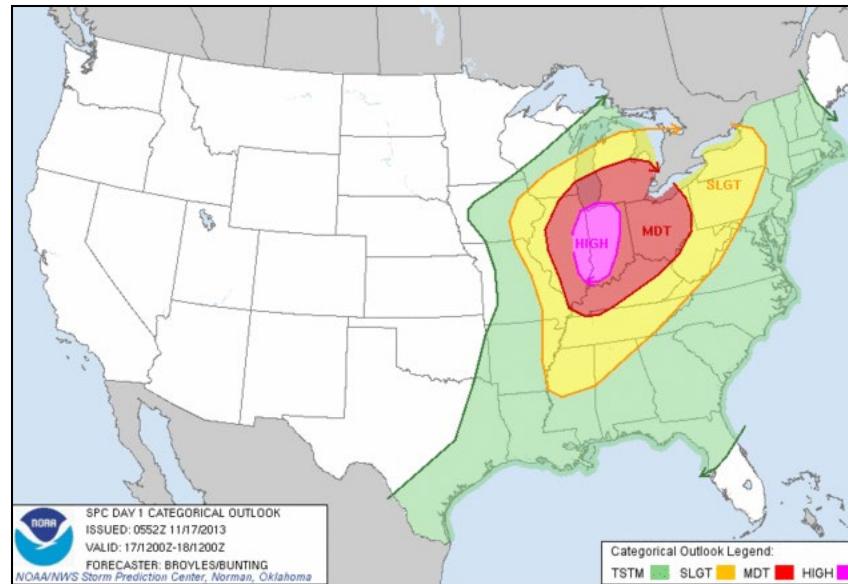
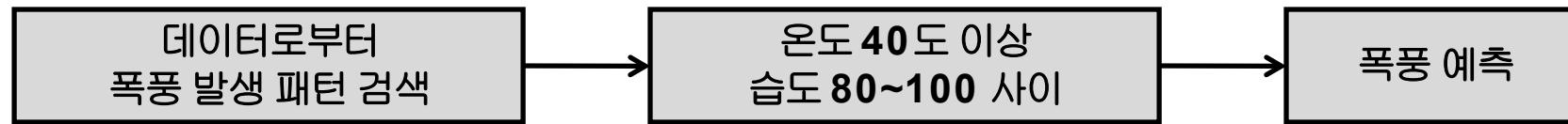
예측 공식

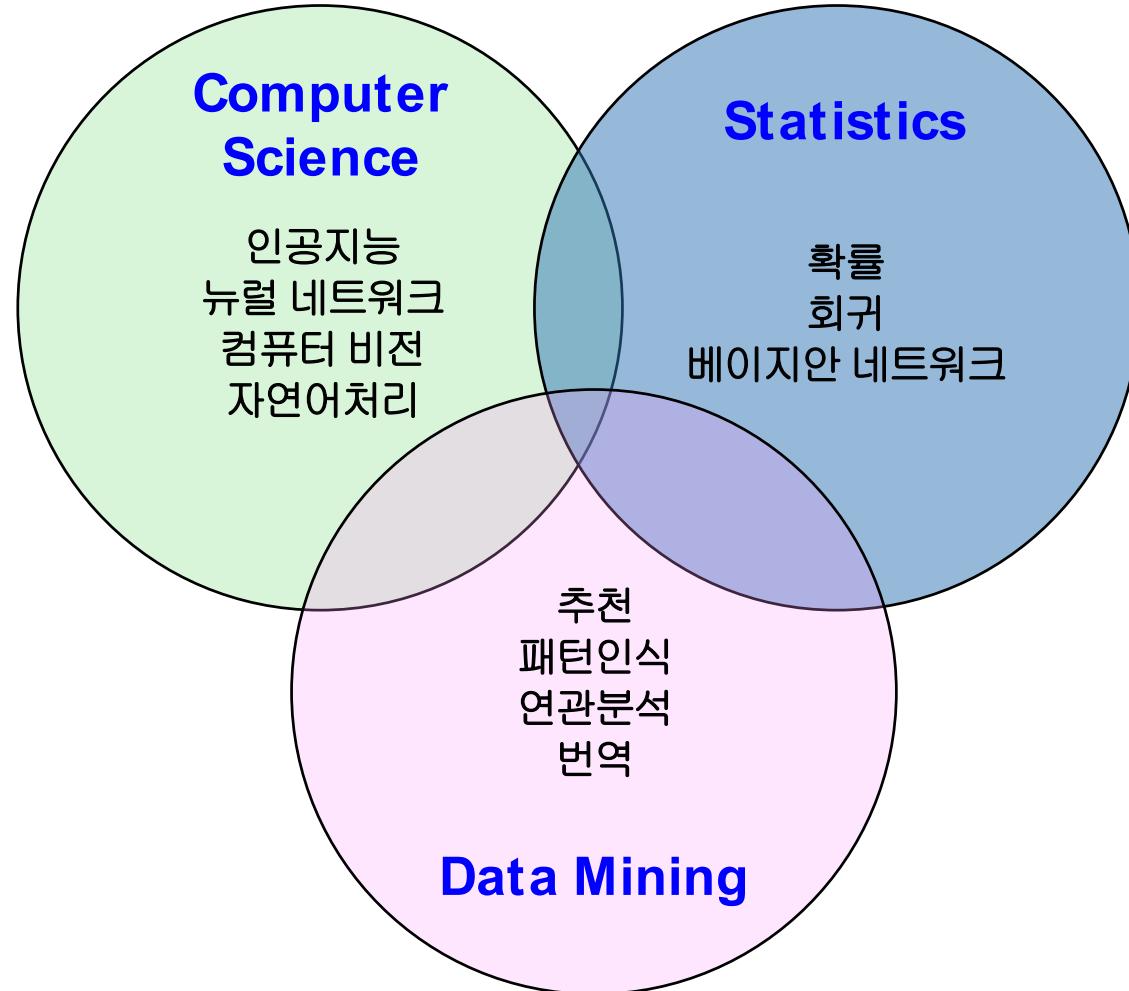
$$\text{몸무게} = \text{키} - 100 \pm F$$

성능 향상 (보정)

머신러닝 : 폭풍 예측 시스템

과거의 폭풍 발생 정보와 폭풍 발생 전 3개월 전 날씨 데이터를 보유하고 있다면 ?





지도학습 (Supervised Learning)

비지도학습 (Unsupervised Learning)

강화학습 (Reinforcement Learning)

Kaggle Titanic 데이터

Attribute, Feature

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

지도 학습 (Supervised Learning)

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

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.

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

키	몸무게	BMI
174	96	31.70
189	87	24.35
185	110	32.14
195	104	27.35
149	61	27.47
169	103	36.06
157	110	44.62
154	111	46.86
174	90	29.72

회귀 (Regression)
(y가 수치 데이터로 구성)

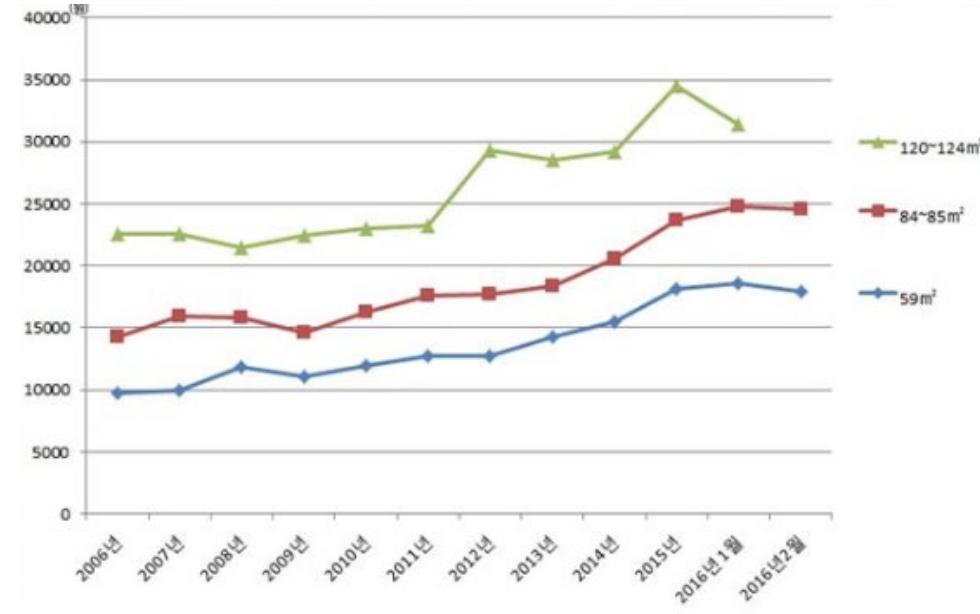
키	몸무게	비만여부
174	96	비만
189	87	정상
185	110	비만
195	104	과체중
149	61	과체중
169	103	과체중
157	110	초고도비만
154	111	초고도비만
174	90	과체중

분류 (Classification)
(y가 범주형 데이터로 구성)

지도 학습 (Supervised Learning)



스팸 메일 분류



집 가격 예측 (회귀)

비지도 학습 (Unsupervised Learning)

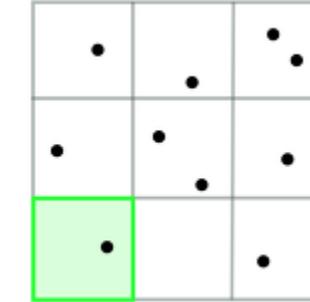
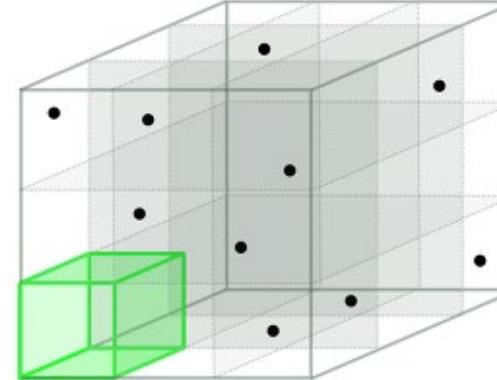
- 데이터에 대한 Label(명시적인 답)이 없는 상태에서 컴퓨터를 학습시키는 방법.
- 데이터의 숨겨진 특징, 구조, 패턴을 파악하는데 사용.
- 데이터를 비슷한 특성끼리 묶는 클러스터링(Clustering)과 차원축소(Dimensionality Reduction)등이 있다.

비지도 학습 (Unsupervised Learning)

키	몸무게
174	96
189	87
185	110
195	104
149	61
169	103
157	110
154	111
174	90

Label
B
A
B
C
C
C
D
D
C

군집 (Clustering)
(y가 없는 데이터를 라벨링)



차원축소
(Dimensionality Reduction)

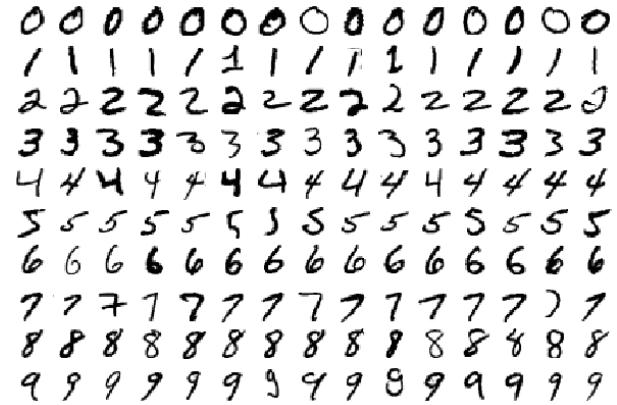
비지도 학습 (Unsupervised Learning)



이미지 감색 처리

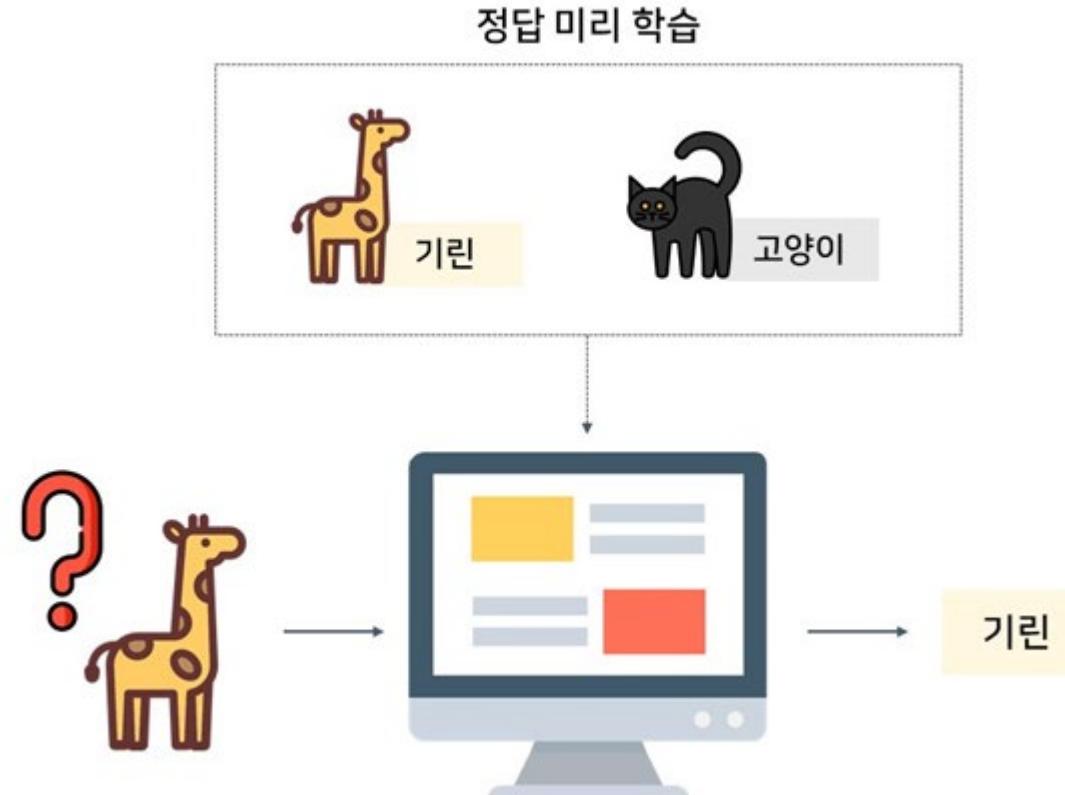


소비자 그룹 발견을
통한 마케팅

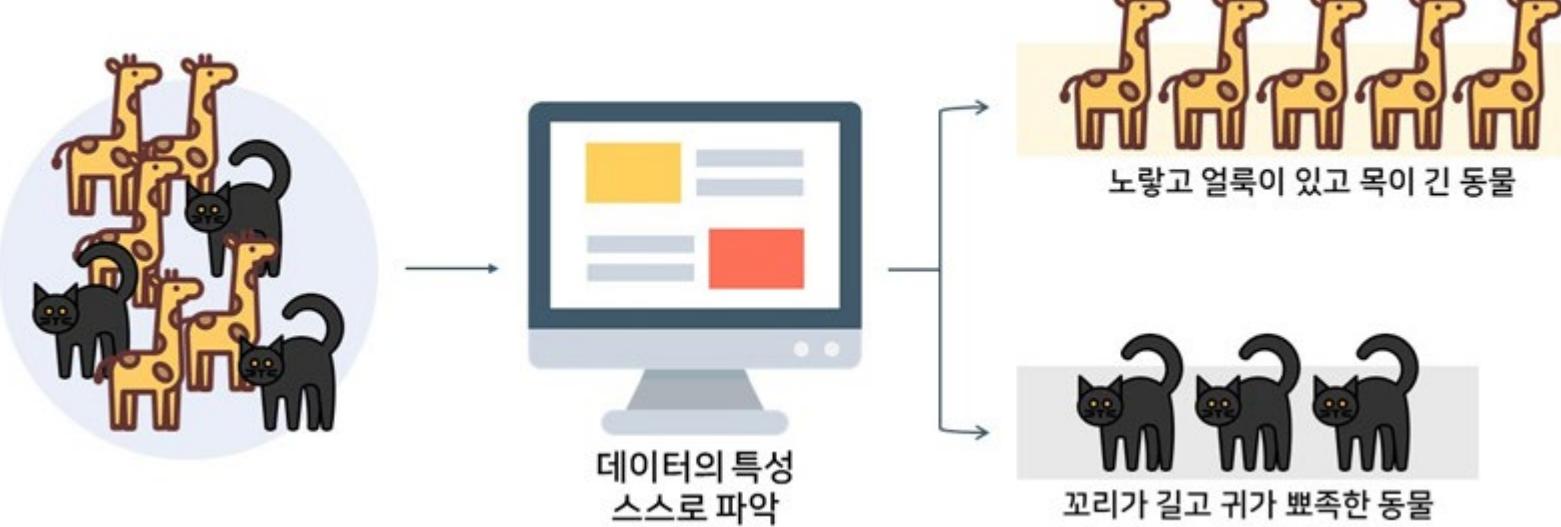


손 글씨 숫자 인식

지도학습



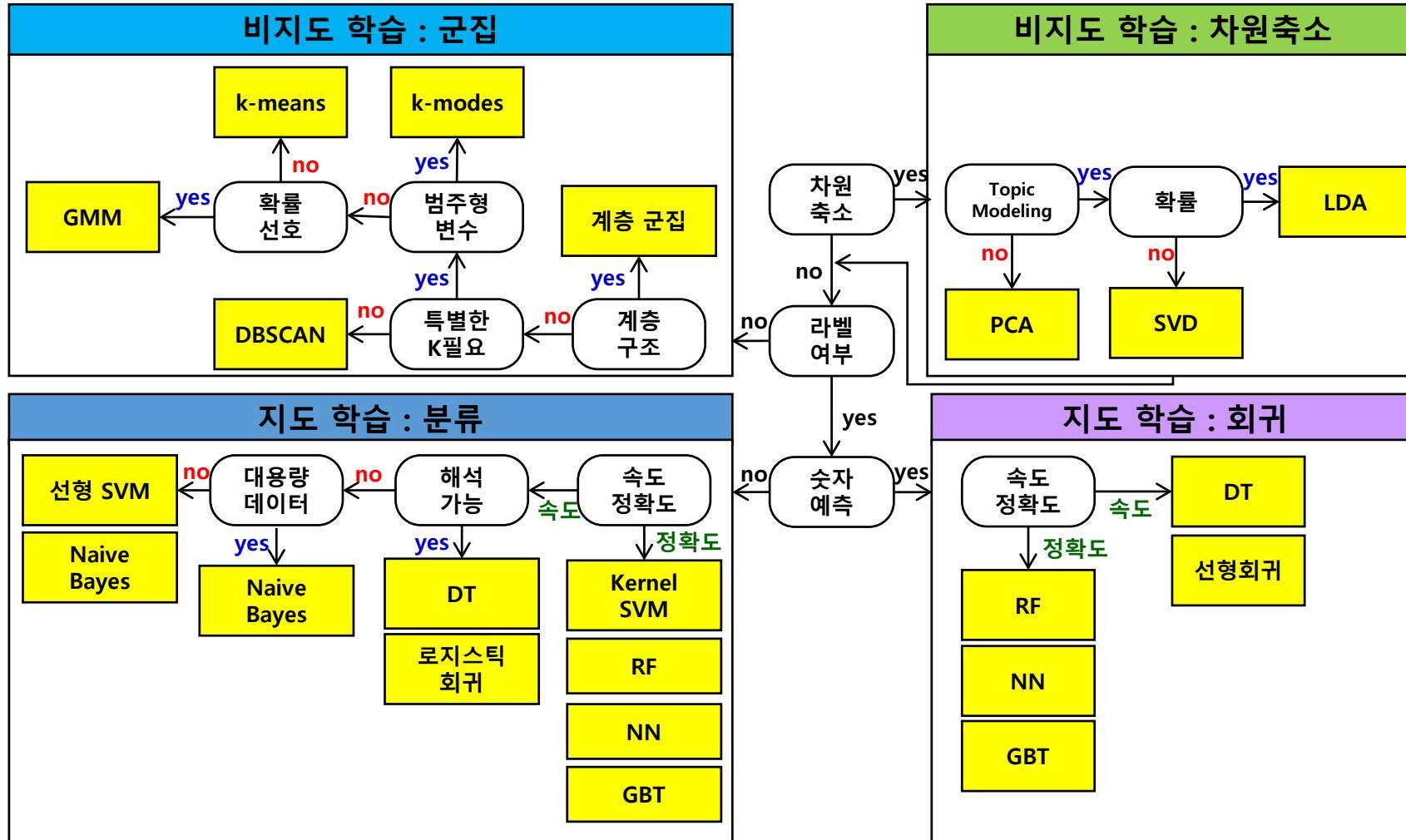
비지도 학습



지도 학습 / 비지도 학습 모델 종류

	비지도 학습	지도 학습
연속 데이터	<ul style="list-style-type: none">▪ 군집과 차원축소<ul style="list-style-type: none">- SVD- PCA / ICA- NMF- K-means- DBSCAN- t-SNE	<ul style="list-style-type: none">▪ 회귀<ul style="list-style-type: none">- 선형 회귀, 다항 회귀- 라쏘 회귀, 릿지 회귀- 일래스틱넷- 회귀 트리 (앙상블)- SVR
이산 데이터	<ul style="list-style-type: none">▪ 연관 규칙 학습<ul style="list-style-type: none">- Apriori- FP-Growth▪ HMM	<ul style="list-style-type: none">▪ 분류<ul style="list-style-type: none">- k-NN- 결정트리- 로지스틱 회귀- Naive-Bayes- SVC- 신경망, 퍼셉트론- Random Forests- GBM / Xgboost

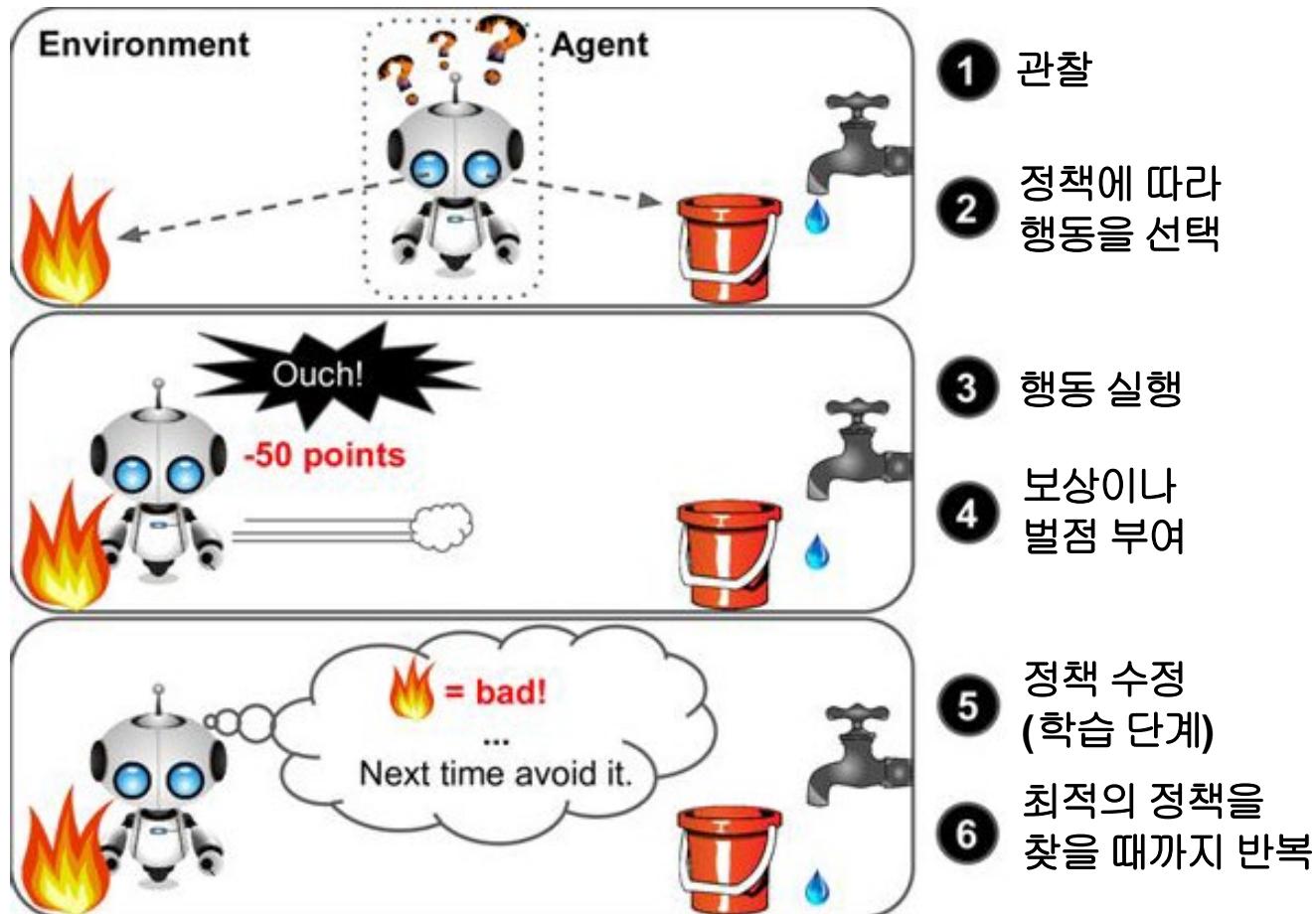
지도학습 / 비지도 학습 모델 선택



강화 학습 (Reinforcement Learning)

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

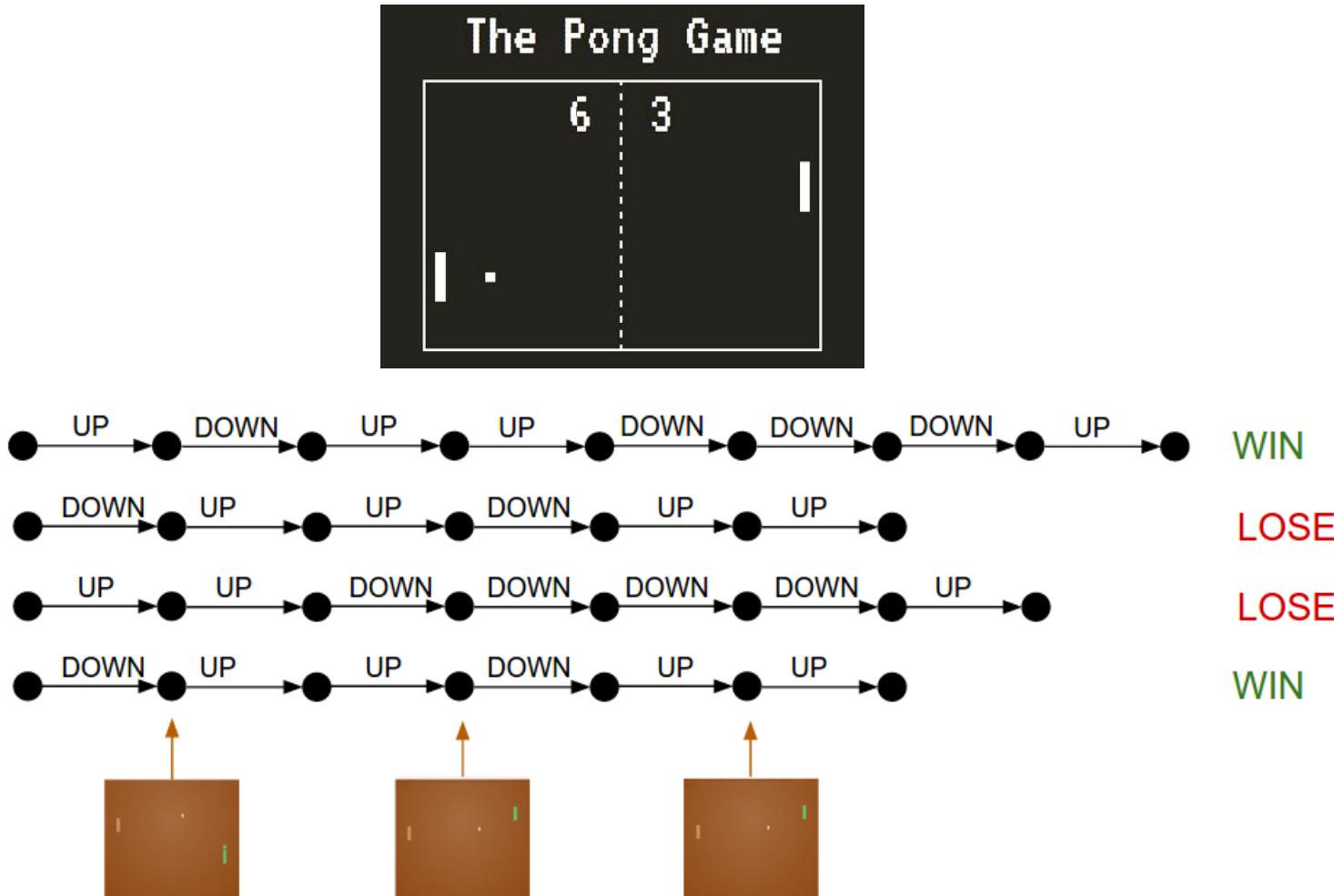
강화 학습 (Reinforcement Learning)



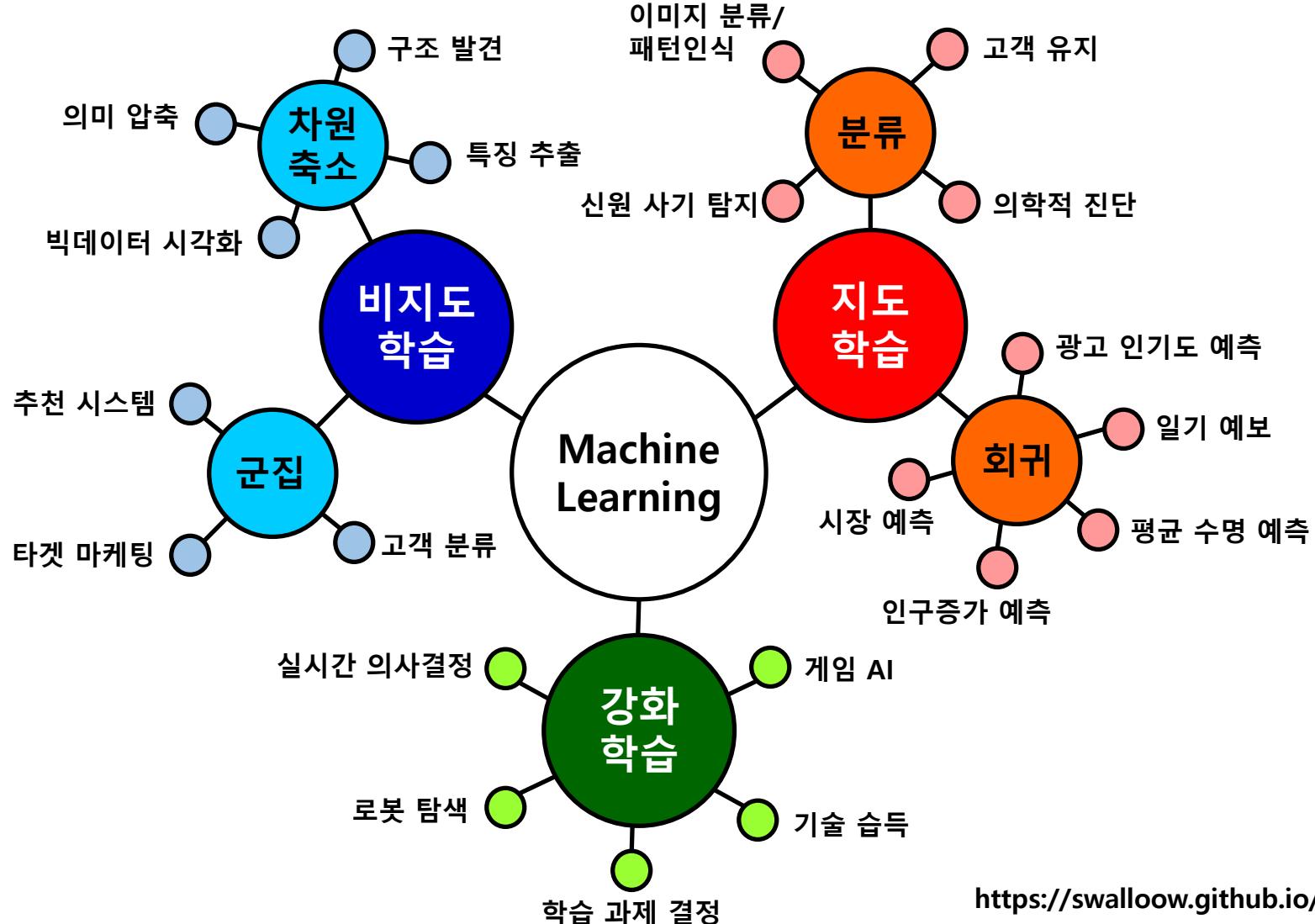
강화 학습 (Reinforcement Learning)



강화 학습 (Reinforcement Learning)



머신러닝(Machine Learning) 활용분야



- 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

XOR 연산을 학습 해보자

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

XOR 연산 학습하기

문제[Feature]

A	B	XOR
0	0	0
1	0	1
0	1	1
1	1	0

답[Label]

XOR 연산 학습하기

Model

(알고리즘)

KNN 분류 모델을 사용

XOR 연산 학습하기

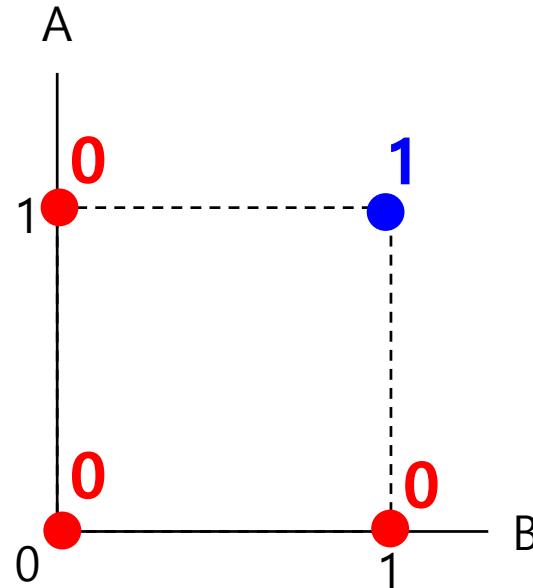
```
knn = KNeighborsClassifier(n_neighbors=1)
```

```
knn.fit(문제 데이터, 답)
```

```
예측한 답 = knn.predict(답을 얻고 문제 데이터)
```

```
score = metrics.accuracy_score(실제 답, 예측한 답)
```

대표적인 학습 방법 – AND 연산의 예

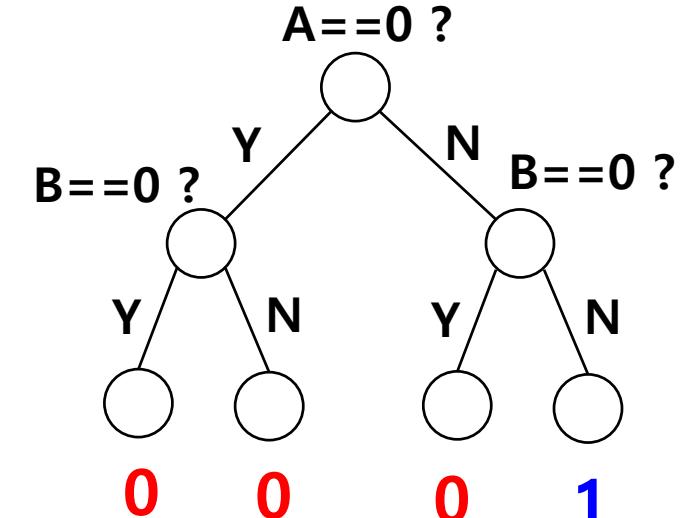


거리 기반 학습
(KNN, SVM)

$$0.5A + 0.5B > 0.7$$

- $A = 0, B = 0 \rightarrow 0 > 0.7 \rightarrow 0$
- $A = 0, B = 1 \rightarrow 0.5 > 0.7 \rightarrow 0$
- $A = 1, B = 0 \rightarrow 0.5 > 0.7 \rightarrow 0$
- $A = 1, B = 1 \rightarrow 1.0 > 0.7 \rightarrow 1$

수식 기반 학습
(선형 분류)



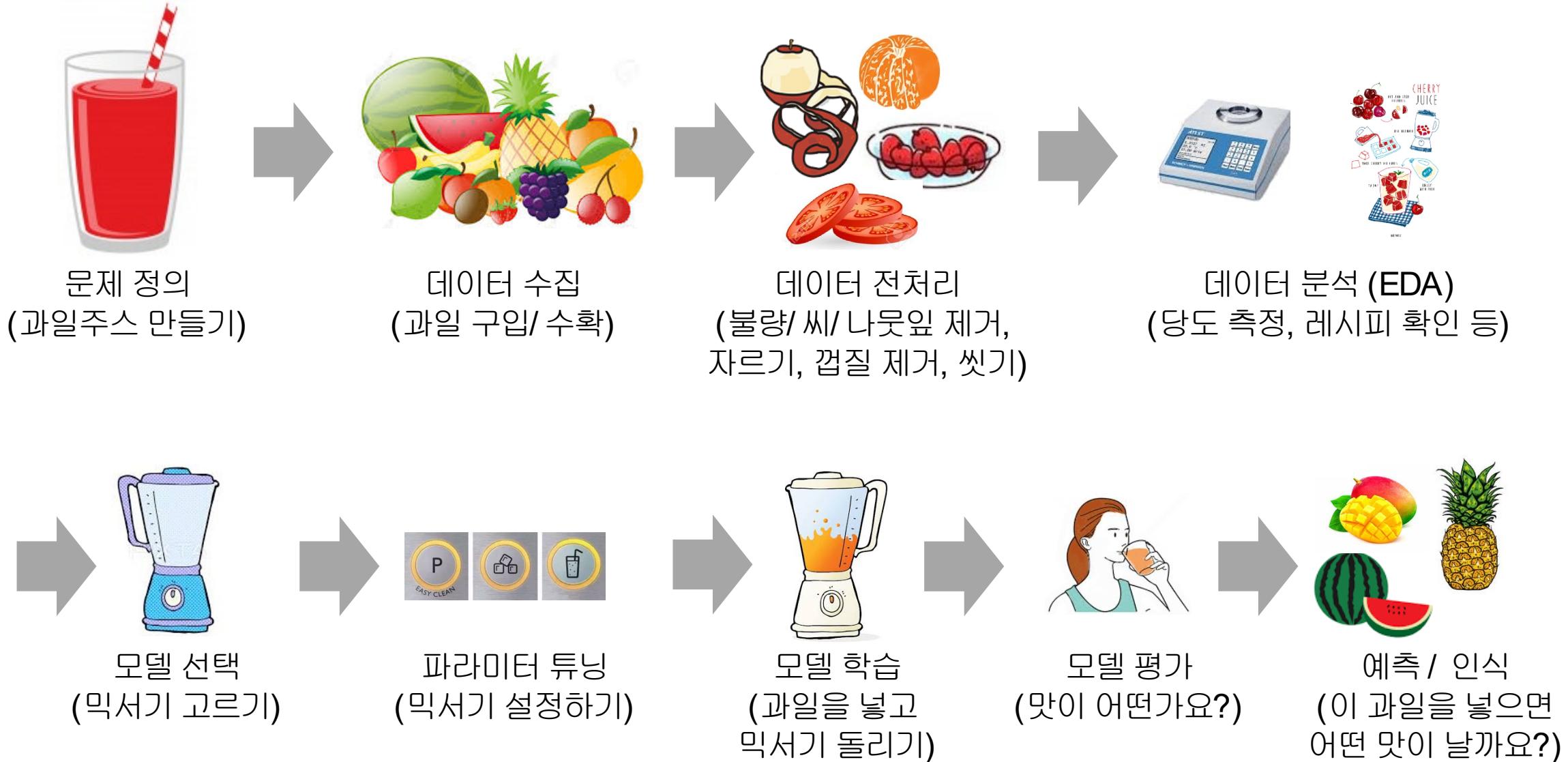
트리 기반 학습
(의사결정트리)

AND/ OR 연산을 학습 해보자

머신러닝(Machine Learning) 학습 과정



과일주스 만드는 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

2. Data Collect(데이터 수집)

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

3. Data Preprocessing(데이터 전처리)

- 결측치 처리 (삭제 or 대체 (중간, 평균, 예측값 등))
- 이상치(outlier) 처리 (삭제 or 대체 (중간, 평균, 예측값, 범주화 등))
- Cleaning (오류 수정)
- Feature Engineering (특성공학)

Scaling (단위 변환)

Transform (새로운 속성 추출)

Encoding (범주형 → 수치형)

Binning (수치형 → 범주형)

Normalization (정규분포화)

범주형 데이터 통합

4. EDA (탐색적 데이터분석)

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

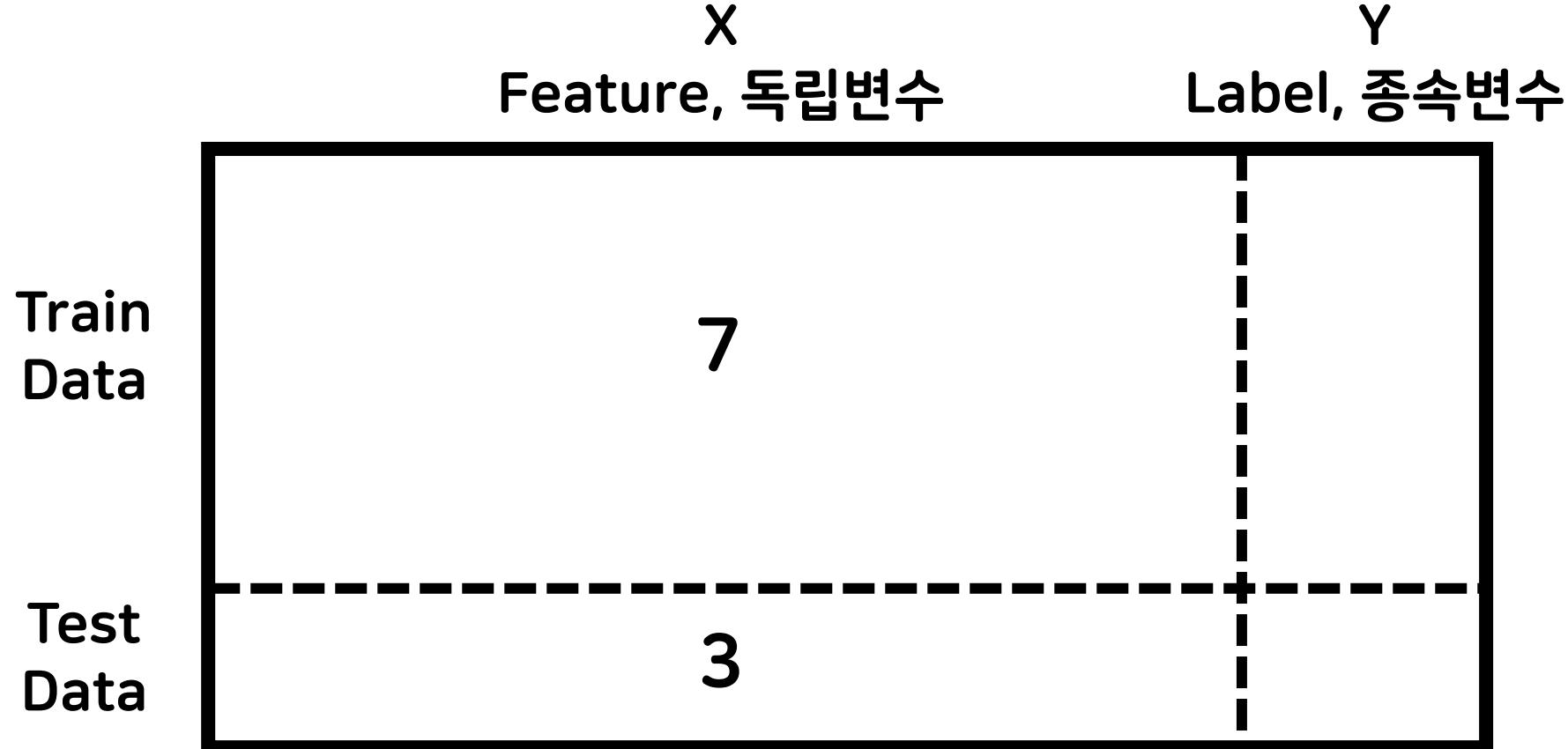
- 단변수 시각화 : **Histogram** (빈도수),
Boxplot(평균, 중간값 등)
- 이변수 시각화 : **Scatter plot** (수치, 상관관계),
누적막대그래프 (범주, 독립성 분석), **범주별 Histogram**
- 다변수 시각화 : **Violin plot**, 3차원 그래프 등

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 개선하기 위해 사람이 직접 넣는 parameter

6. Model Training(학습)

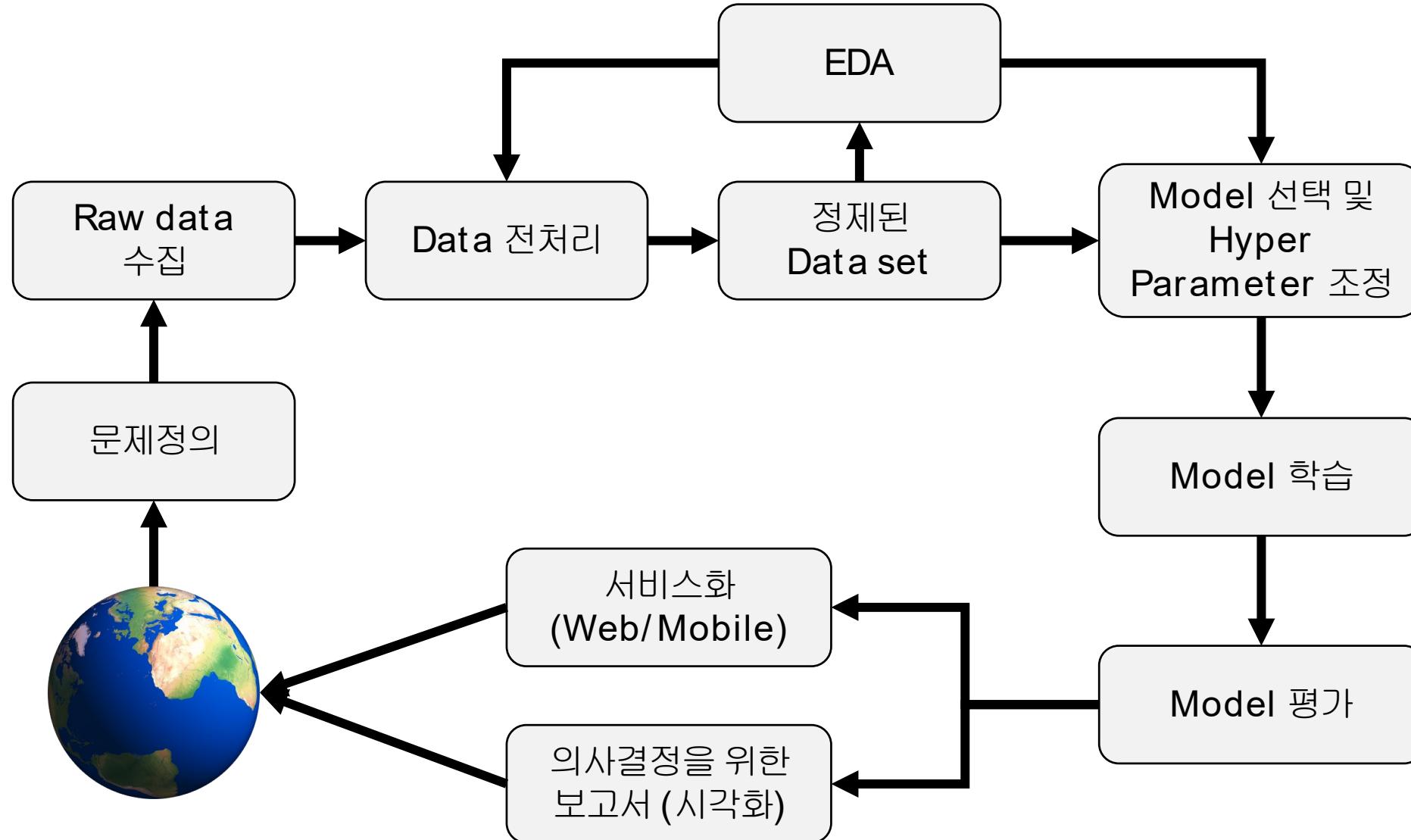
- **model.fit(X_train,y_train)**
train 데이터와 test 데이터를 7:3 정도로 나눔
- **model.predict (X_test)**



7. Evaluation(평가)

- MSE / RMSE
- R2 Score
- accuracy(정확도)
- recall(재현율)
- precision(정밀도)
- f1 score

머신러닝(Machine Learning) 학습 과정



회귀



다음 학기 성적 점수를 예측하려면
어떤 특성이 필요할까?

직전 학기 성적, 알바 진행 여부, 연애,
잠자는 시간, 출석률, 학교와 집 사이 거리

머신러닝(Machine Learning) 실습

비만도(BMI) 데이터를 이용해 학습 해보자.

	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?

시각화를 통해 분류가 가능한
문제인지 확인해보자.

성별을 추가하여 학습해보자