

Chapter 1. Introduction

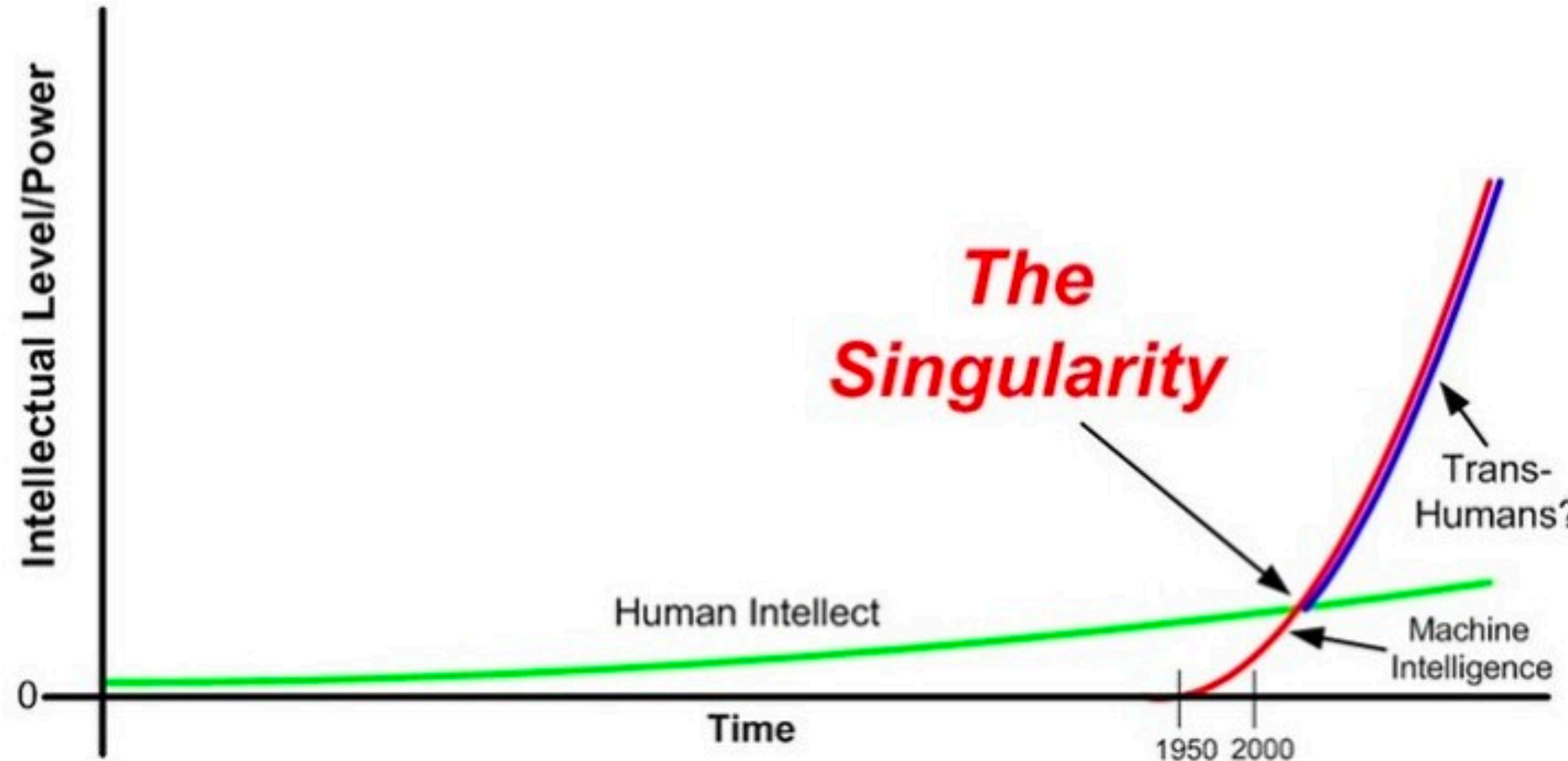
만물의 영장

- 인간, 만물의 영장 (靈長)
- 인간만이 생각할 수 있는 존재이기 때문...

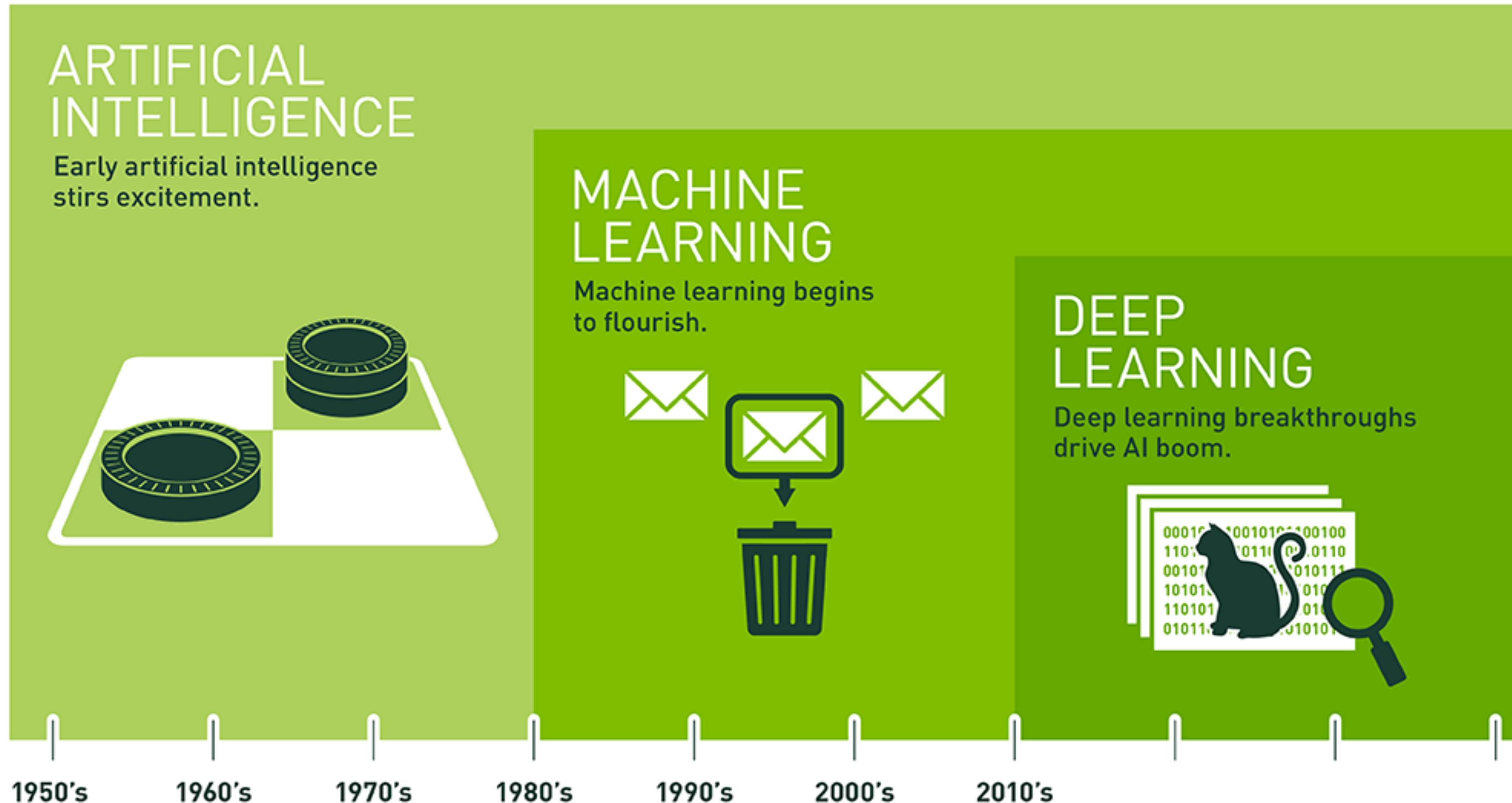


데카르트

The Singularity



Machine learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

1958 in the New York Times

- A new device learns by doing
- The embryo of an electronic computer that expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.
- "걸을 수 있고, 말하고, 볼 수 있으며, 쓸 수 있고, 본인을 복제하며, 자신의 존재를 인식할 수 있는 전자 컴퓨터의 초기 모델입니다."

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human be-

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls
The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

1958 New York Times...

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

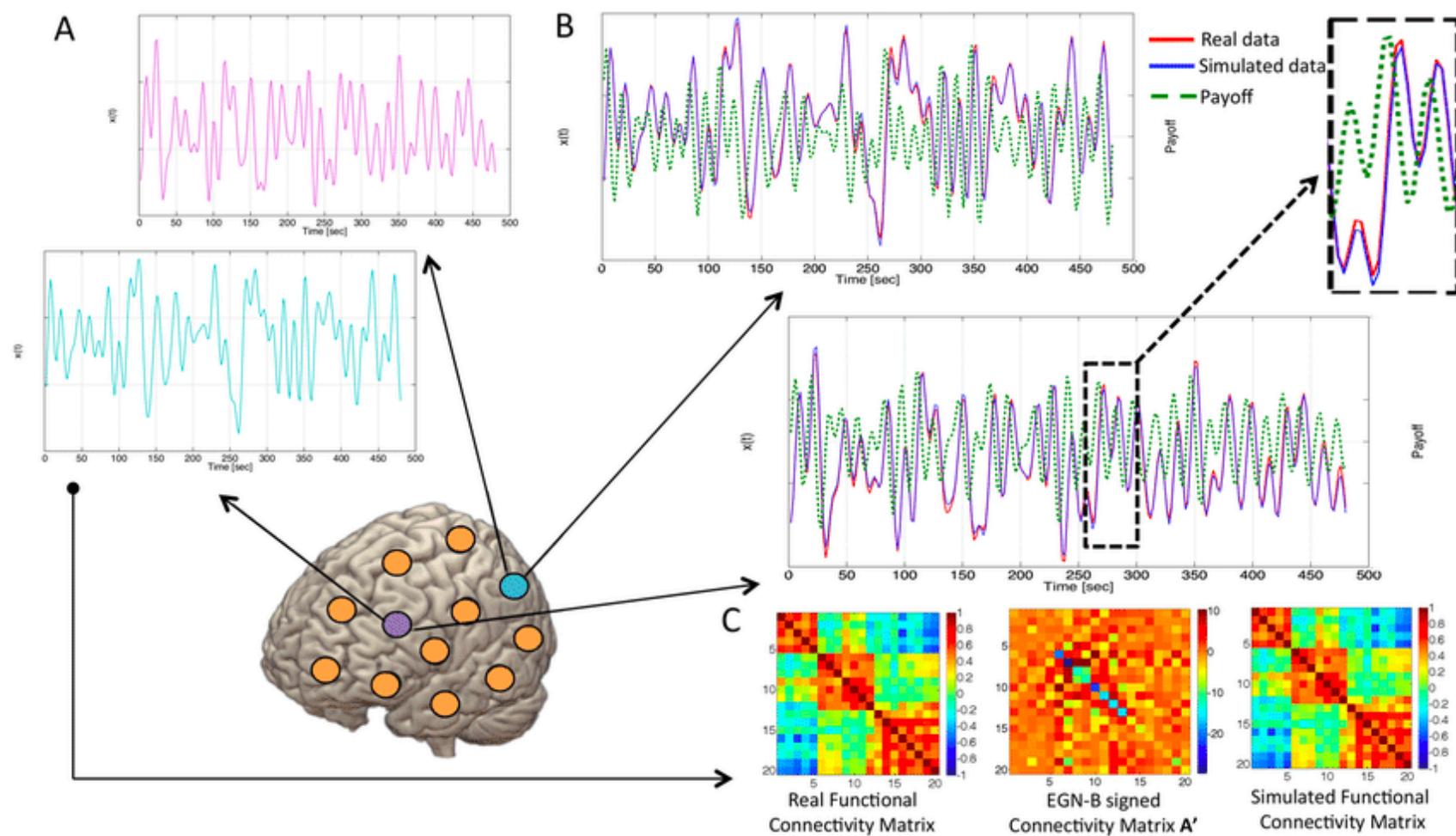
In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

인공 지능을 위한 첫 단추는?

- 인간의 지능은 무엇일까?
- 인간의 두뇌 안에서 일어나는 전기적 신호?
- 신호의 의미를 전부 이해할 수 있을까?



사회

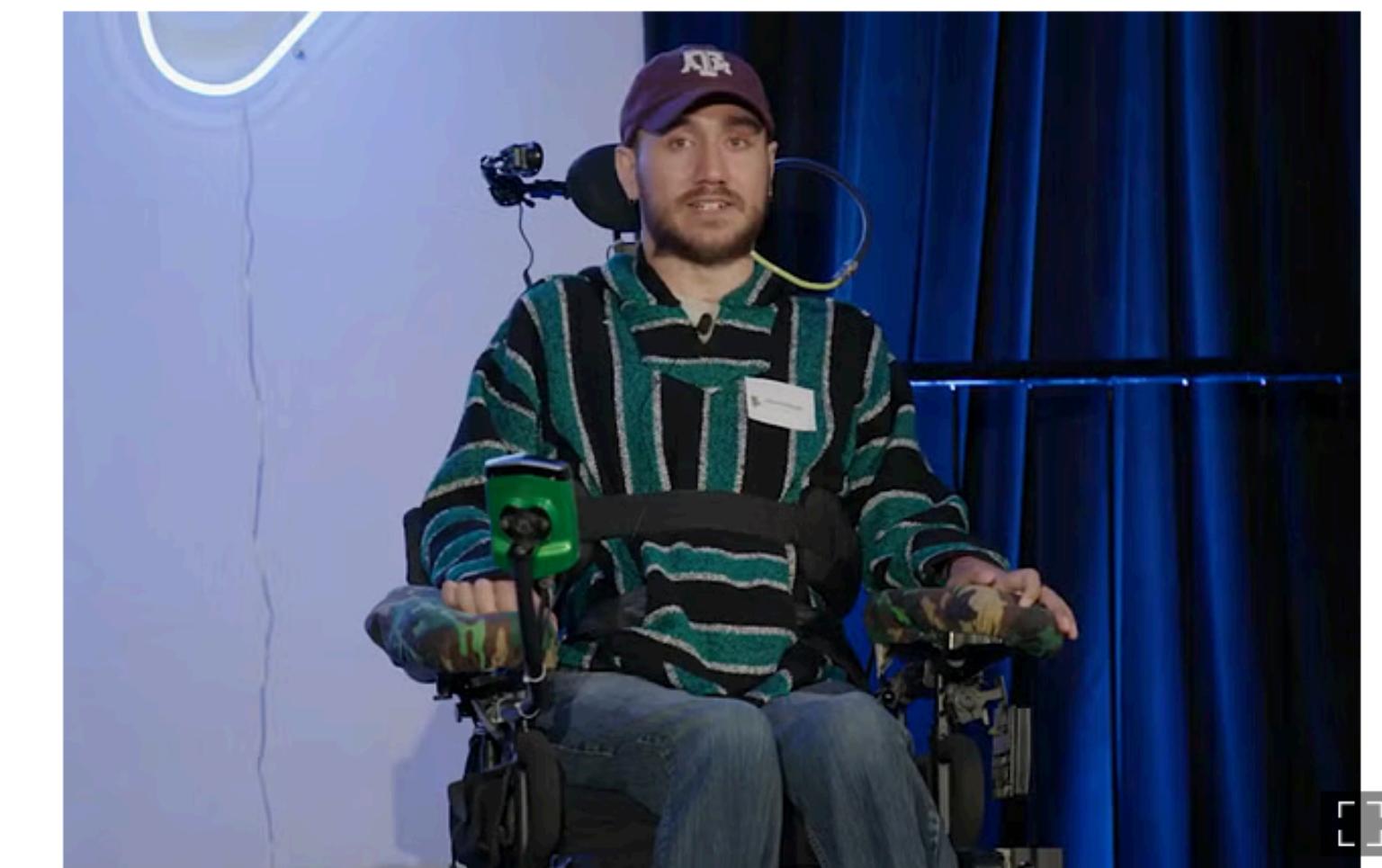
“초능력 가진 것 같아, 인생 바뀌었다”...세계 최초로 뇌에 뉴럴링크 이식한 남자

이덕주 기자 mrdjlee@mk.co.kr

입력 : 2024-05-17 13:44:52 수정 : 2024-05-17 14:58:24

¶ 가 □ ☰ ☱

대학생때 다이빙 사고로 전신마비
올해 1월 뇌에 뉴럴링크 이식
“일론 머스크가 수술 후 찾아와”
생각하는 대로 컴퓨터 작동가능



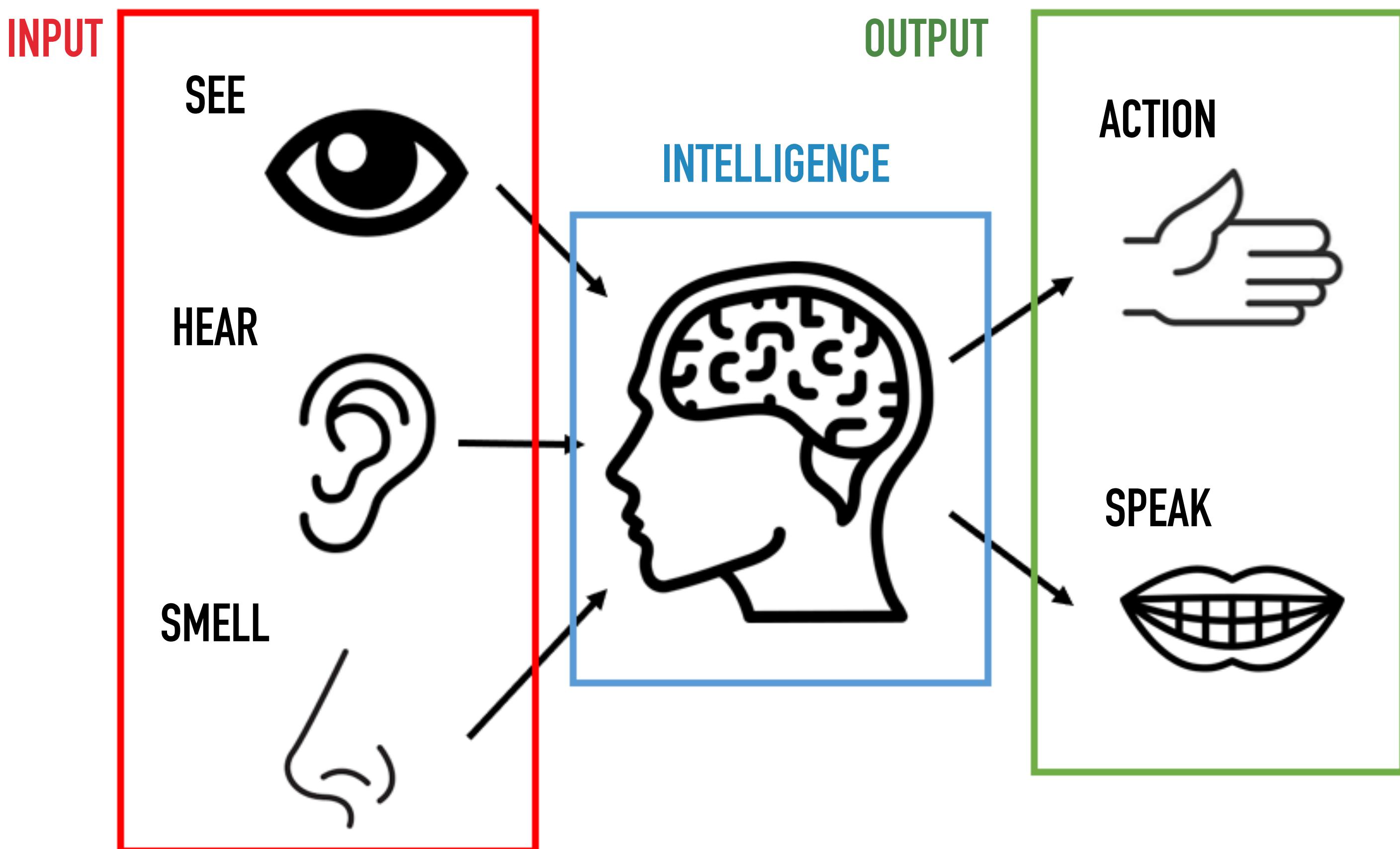
최초의 뉴럴링크 칩 이식자인 놀런 아르보가 뉴럴링크를 방문해 직원들에게 이야기하고 있다. <사진=뉴럴링크>

전기적 신호가 지능이라면...

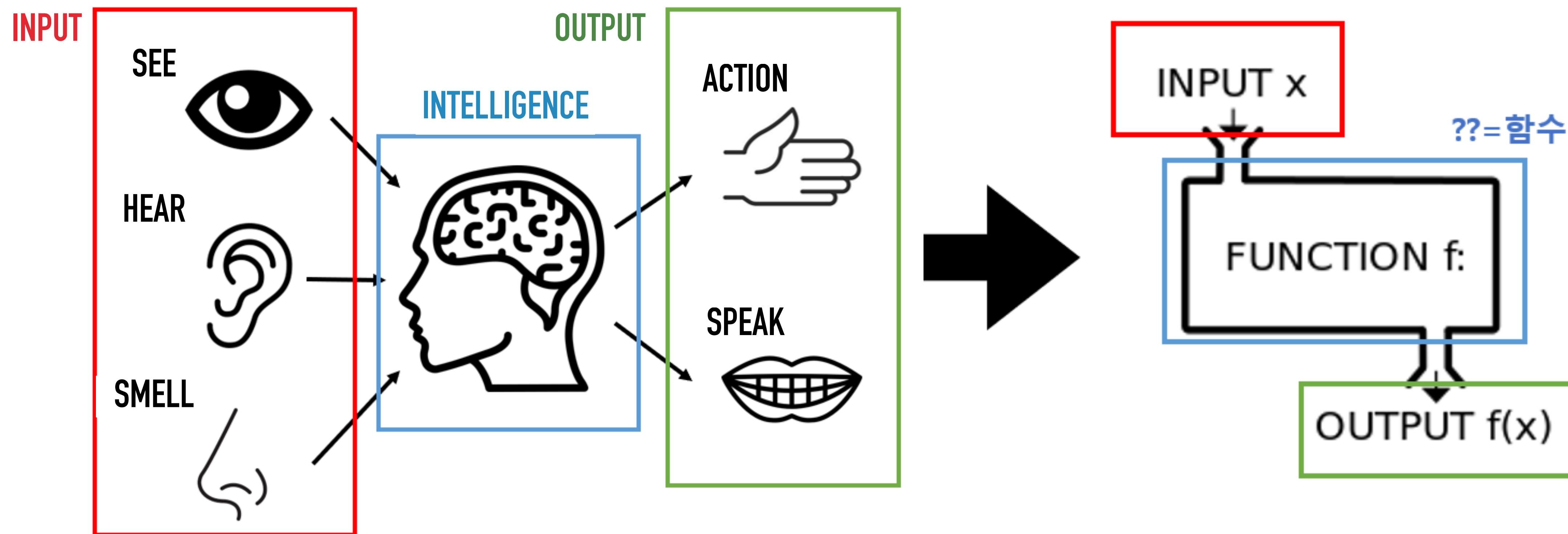
- 두뇌 속 신호에 대한 해석
- 그러한 신호를 생성하는 인공적인 존재를 만드는 것은 또 다른 영역에 해당

인간의 사고 과정

- 감각기관을 통한 인지
→ Inputs
- 지성을 통한 사고
- 사고의 결과를 바탕으로 행동
→ Outputs



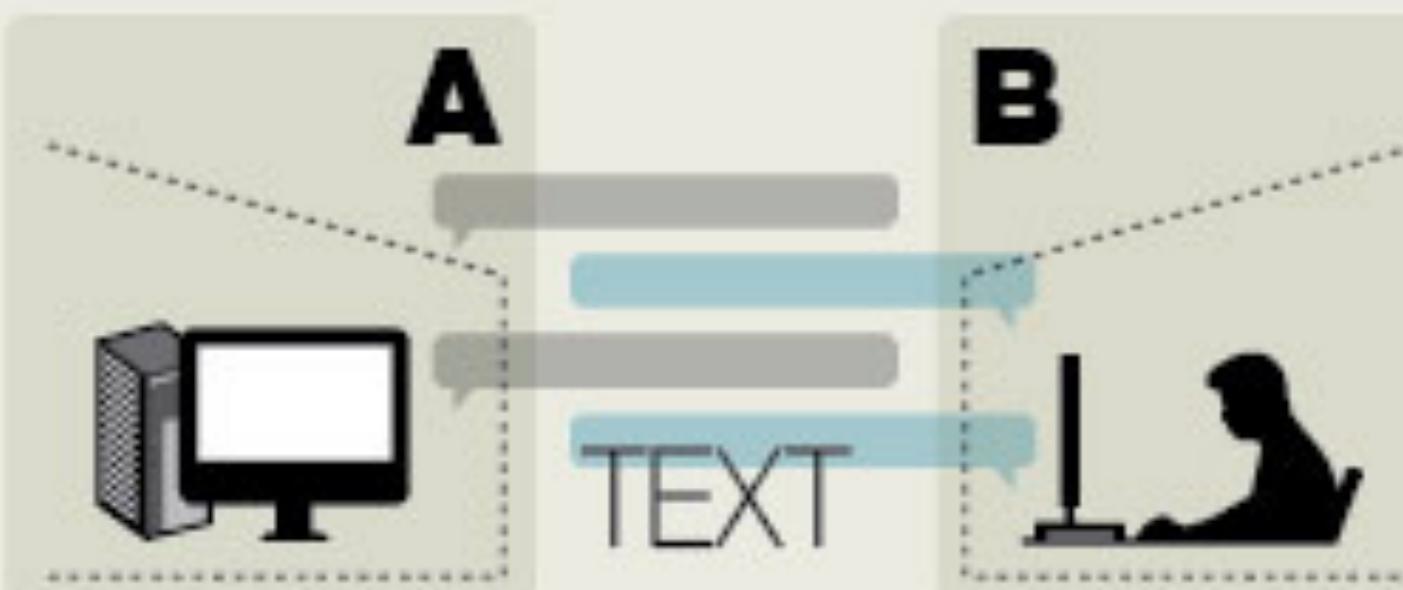
인간의 사고 과정 = 하나의 함수



어떻게 인간의 지능을 모방한 함수를 만들 수 있을까?

인공지능 판별기준 튜링 테스트 개요

- ① 컴퓨터 화면을 통해 문자로만 대화



기계 (실험대상)

TEXT

B

사람 (대조 샘플)

→ ② A,B 모두 사람이라고 주장 ←



심사위원 (제3자)

- ③ 어느 쪽이 사람인지 구분할 수 없을 경우,
A·B 둘다 인간 수준의 사고능력을 가진 것으로 판정

연합뉴스

이재윤 기자 / 20140609

@yonhap_graphics(트위터)

YONHAPNEWS

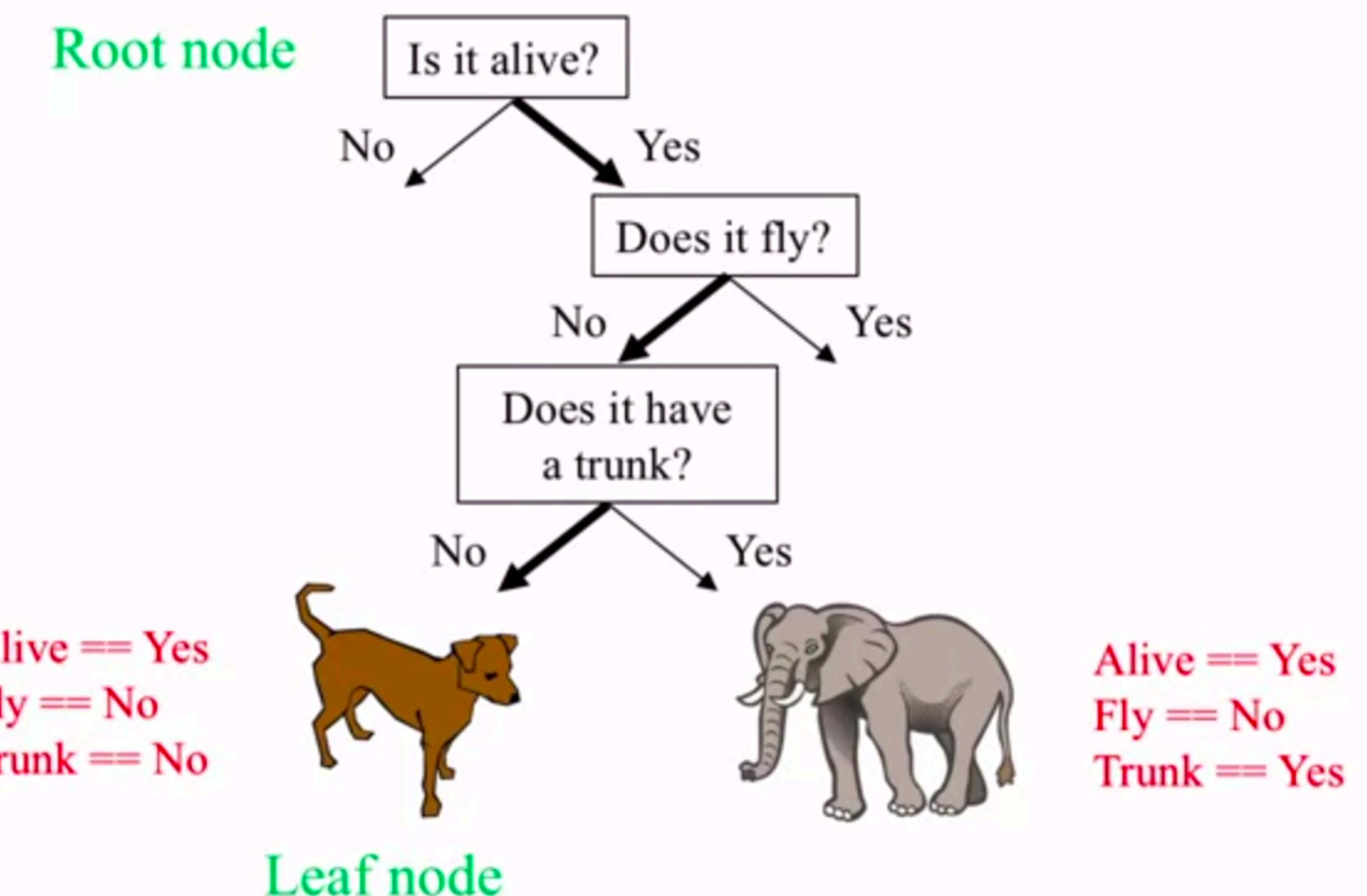
어떻게 인간의 지능을 모방한 함수를 만들 수 있을까?

- 중요한 것은 입력에 대한 출력 → 지능/사고의 결과물로서 출력을 얻어야 함.
- 가장 쉽게 생각할 수 있는 방법은?
 - If, then, else.
 - 조건에 따라 정해진 행동을 수행

Rule-based approach

- 규칙 기반 모델 또는 전문가 시스템
- 직관적임, 하지만 ...?
 - 새로운 입력 또는 질문에 대한 반응은?
 - 단순한 일이 아닌 복잡한 문제는?
- 룰을 결정하는 것이 쉽지 않다...

Decision Tree Example



다른 방법을 생각해 봅시다.

Q. 사람은 어떻게 특정 분야의 전문가가 될 수 있을까요?

다른 방법을 생각해 봅시다.

Q. 사람은 어떻게 특정 분야의 전문가가 될 수 있을까요?

A. 공부, 학습을 통해서 전문가가 됩니다.

기계(프로그램)를 학습시키자!

Q. 그렇다면 기계는 무엇을 배워야 하나요?

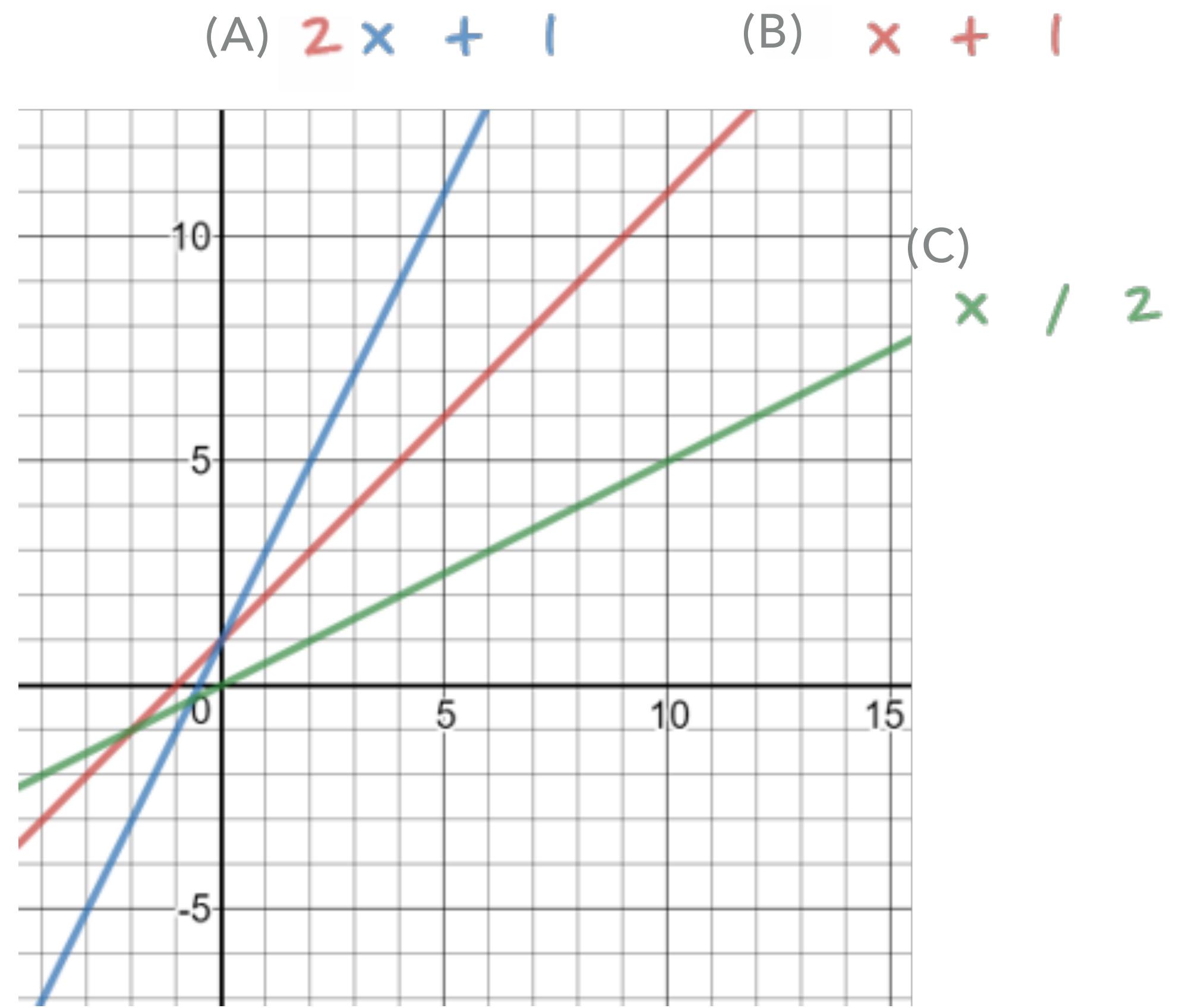
기계(프로그램)를 학습시키자!

Q. 그렇다면 기계는 무엇을 배워야 하나요?

A. 인간 지능의 출력과 동일/유사한 결과를 출력하는 함수

함수를 학습한다는 것의 의미

- $y = ax + b$
 - $a = 2, b = 1$
 - $a = 1, b = 1$
 - $a = 1/2, b = 0$
- $y = ax^2 + b\sin(x) + c$
- $y = f(x; w)$



Example of simple linear graph

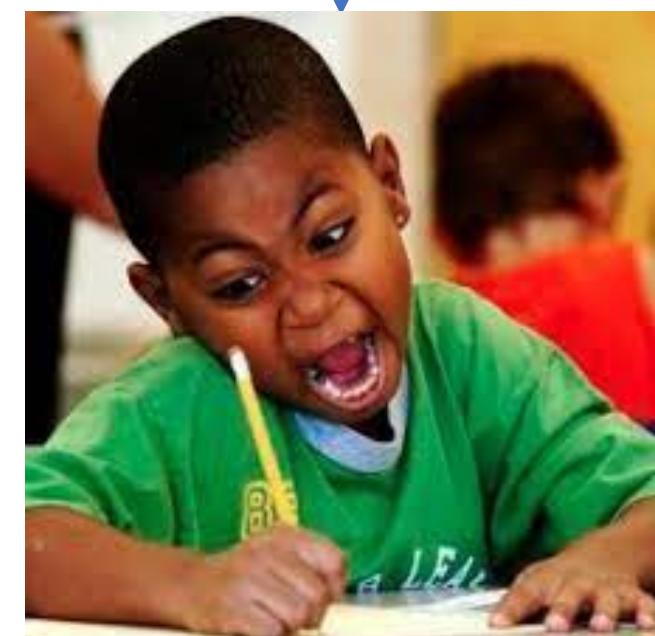
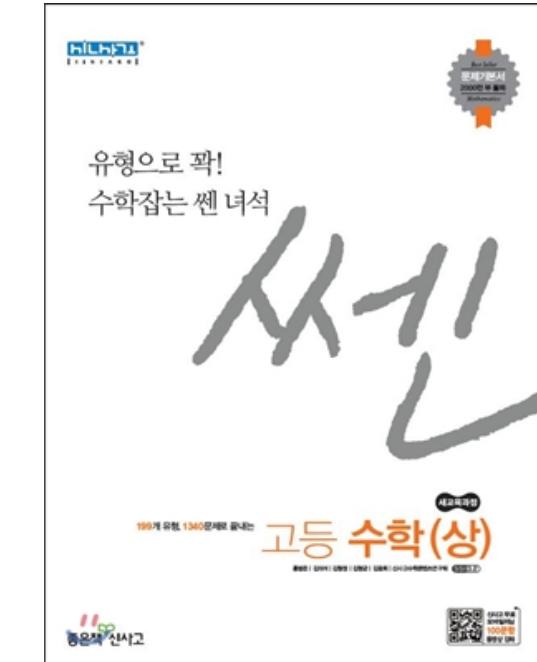
함수를 학습한다는 것의 의미는?

- 함수의 형태를 결정하는 w (웨이트, 파라미터, 계수 등)를 학습을 통해 찾는 것.
- 어떤 w 가 필요한가요?
 - 주어진 태스크에 대해 적합한 결과를 내는 w .
- 기계 학습 (Machine learning, ML)에서는...
 - ML 모델이 w 를 학습할 수 있도록 판을 깔아줌 → 이후에는 반복을 통해 태스크에 대한 성능 향상
 - ML 모델에 따라서 학습이 되는 w 의 형태는 달라질 수 있음.
 - 결정트리의 경우 분할 규칙을 학습, 인공신경망에서는 레이어의 가중치 등

인간 학습 프로세스

- 정답이 있는 문제지
- 문제풀기
- 채점
- 복습
- 위 과정을 반복한다.

Workbook

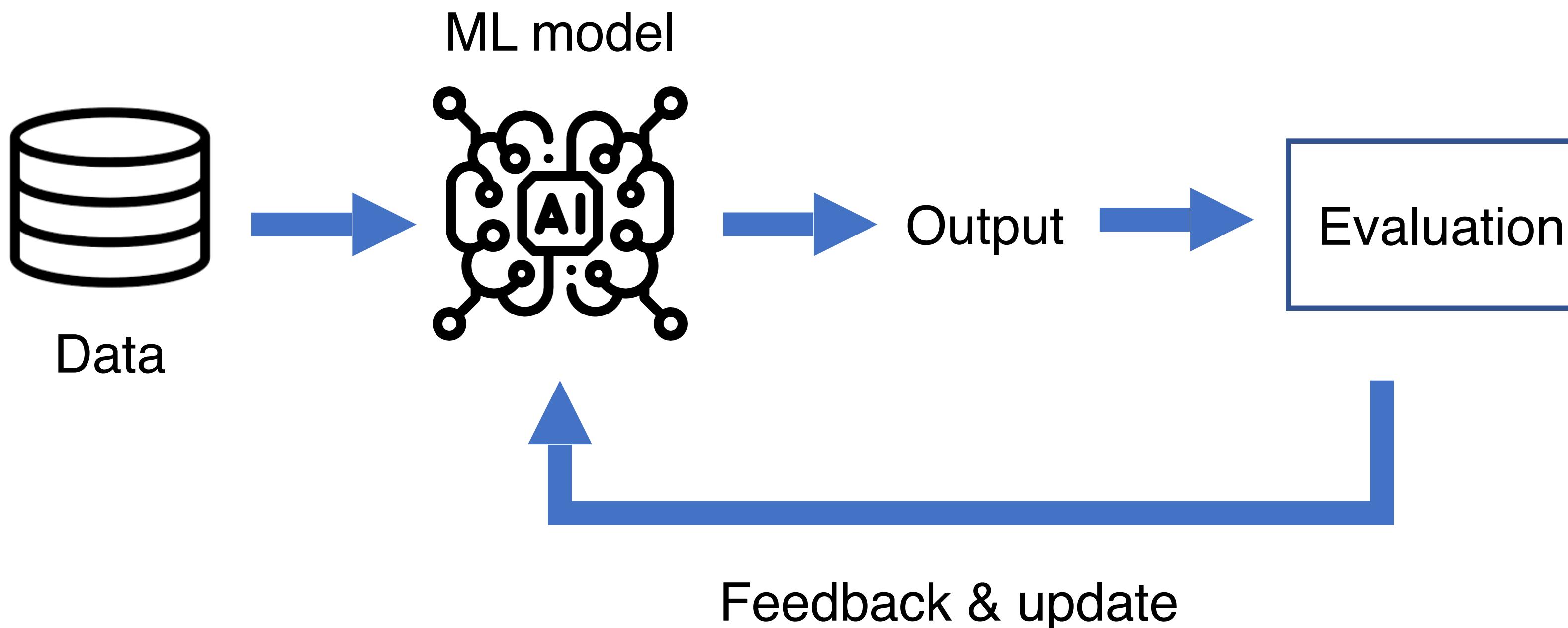


Result

Evaluation

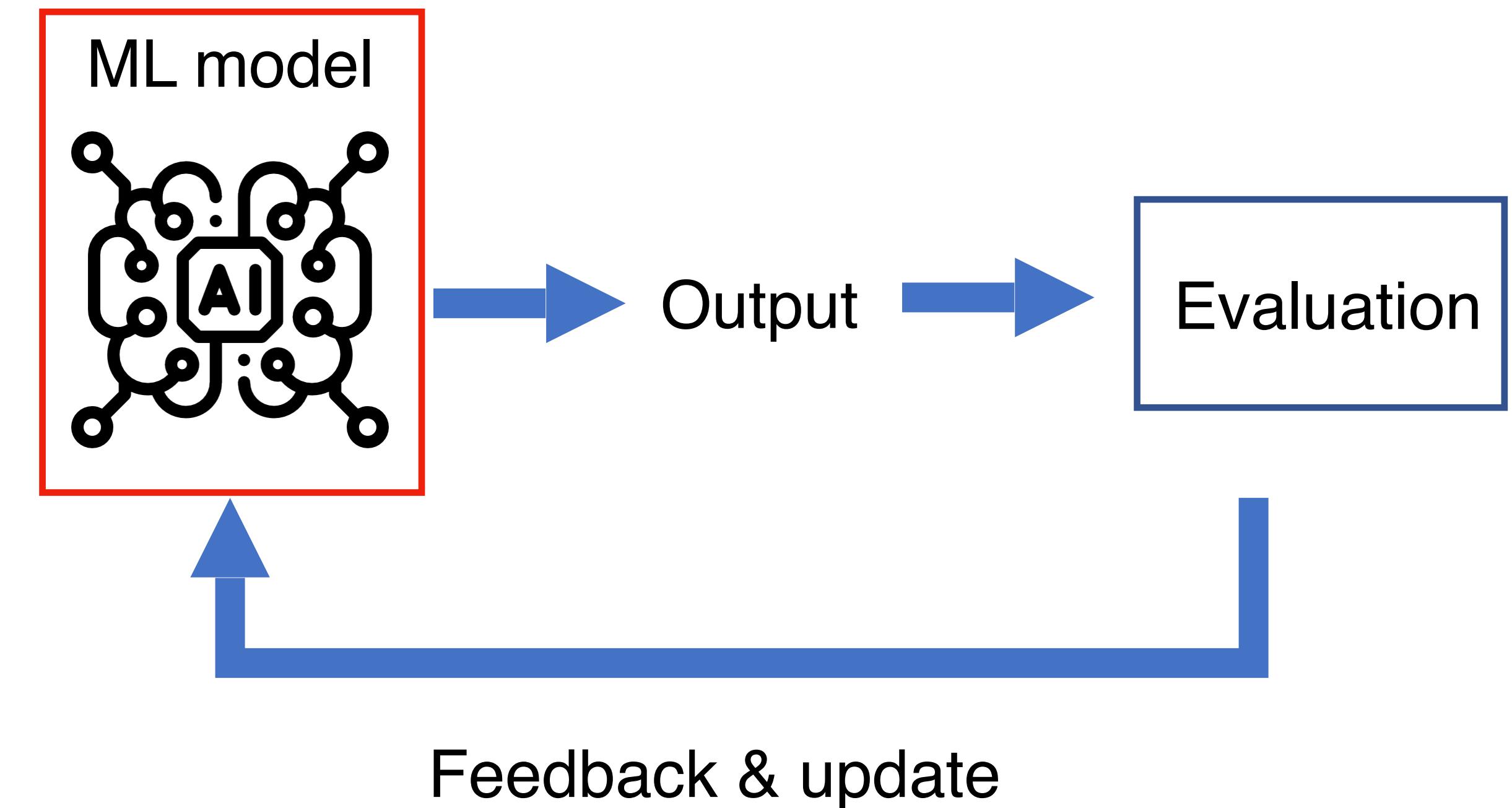
Review

기계 학습 프로세스



어떤 모델을 써야 할까?

- $y = ax + b$ or $y = a \exp(x)$
- 이보다 더 복잡한 형태의 수식?
- 아니면 다른 알고리즘?



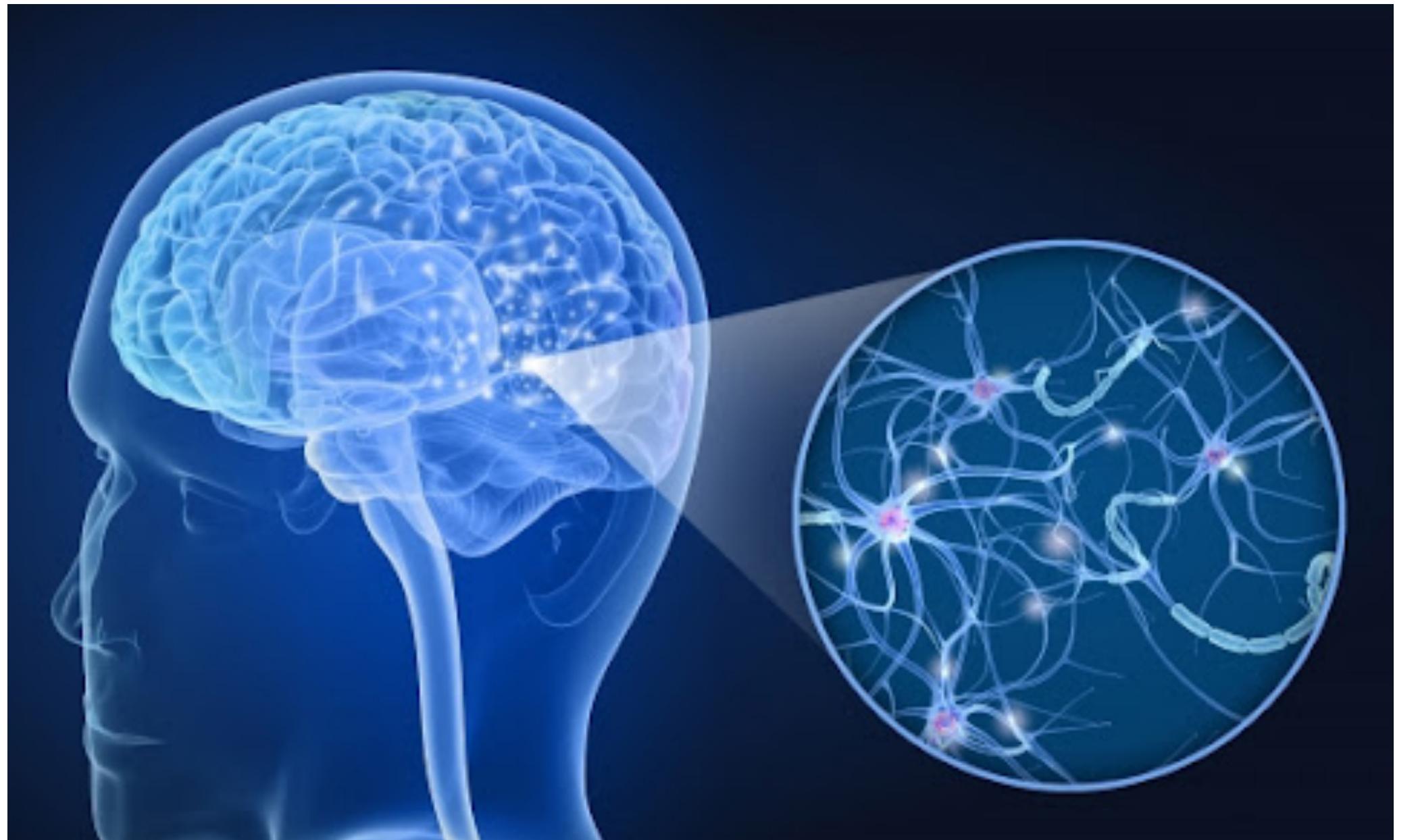
사람은 어떤가요?

- 인간 학습 프레임워크에서는 인간의 두뇌가 ML모델의 역할을 수행함.

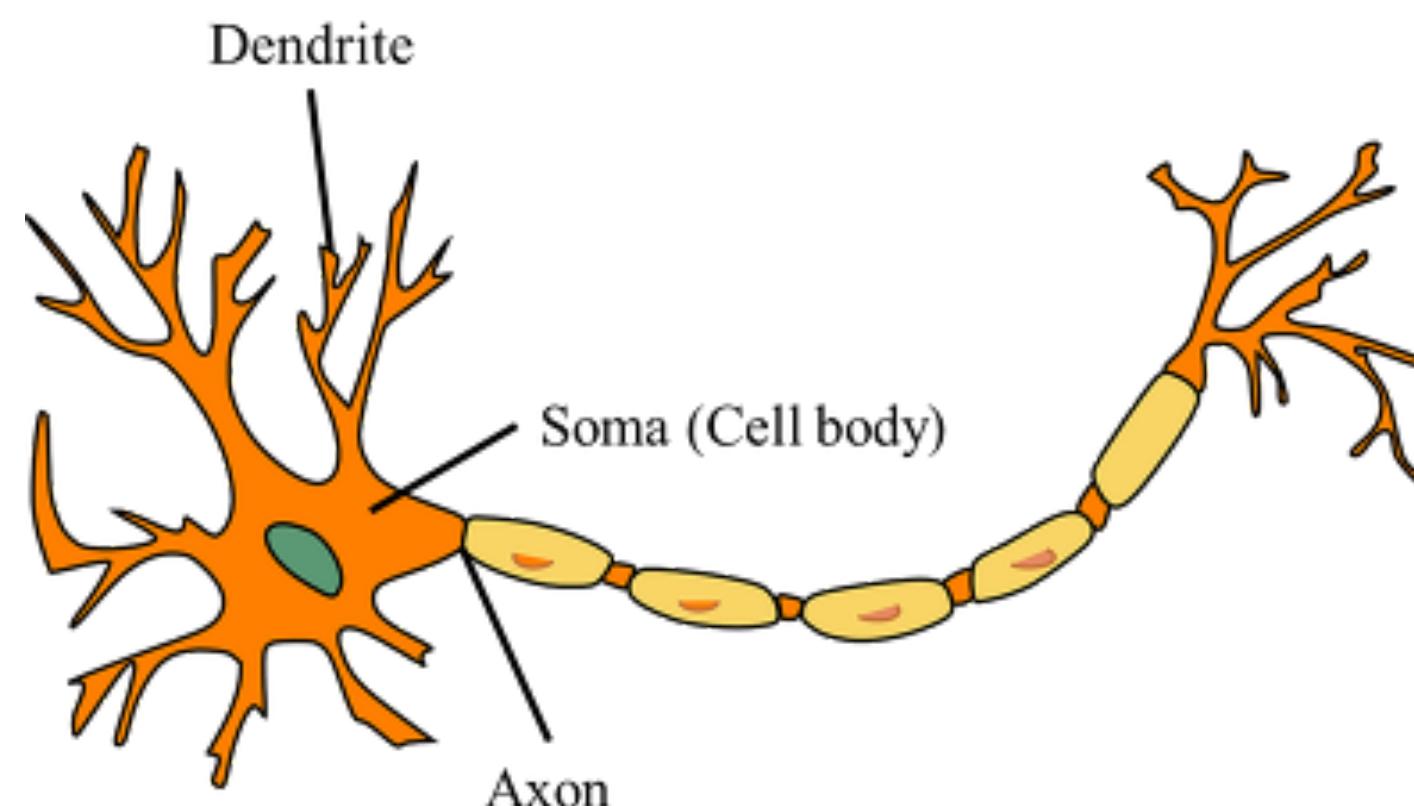


인간의 두뇌는 뉴런들의 거대 네트워크

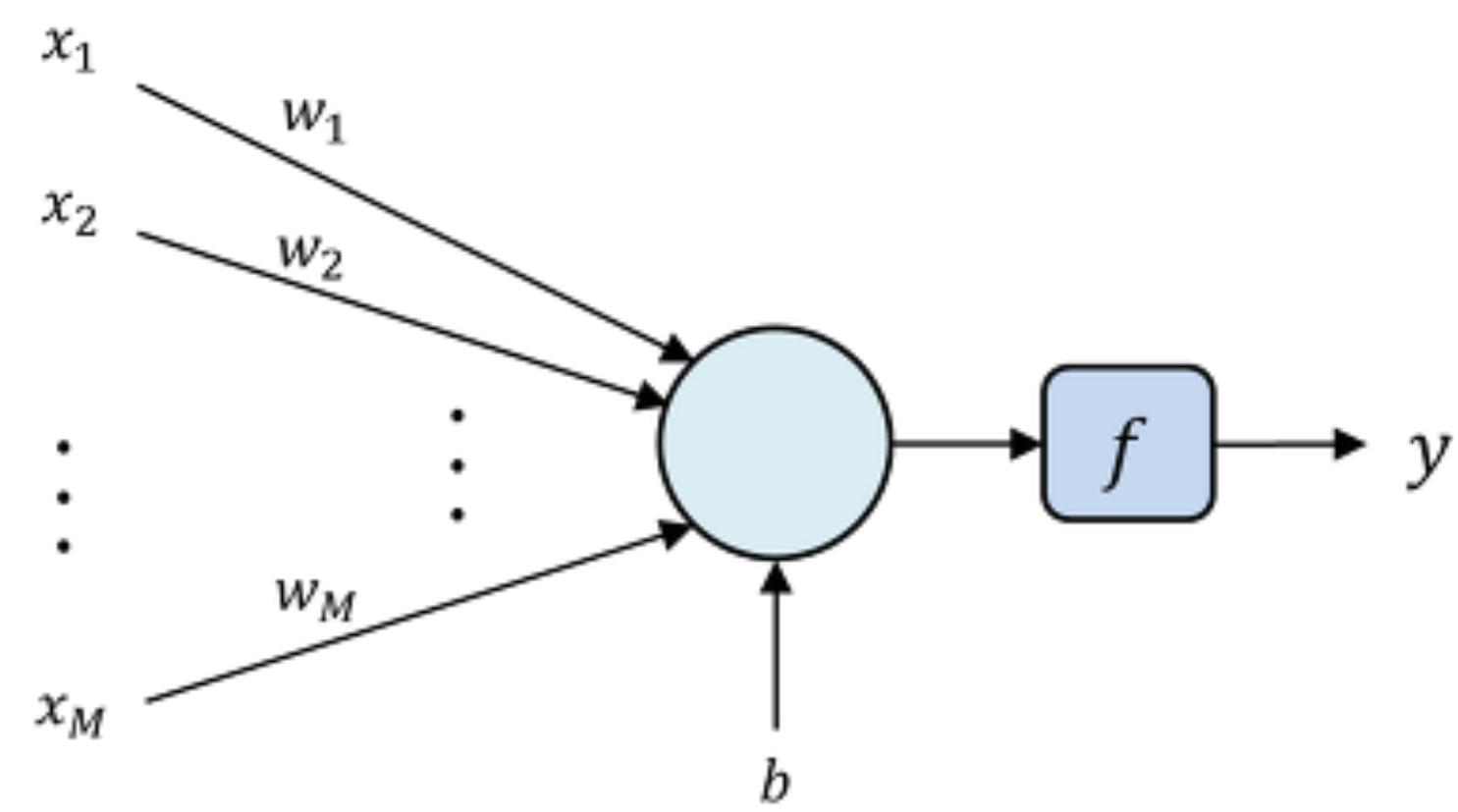
- 약 천억 (100 billion)개의 뉴런
- 약 100조 (100 trilloin)개의 시냅스
- 시냅스: 뉴런과 뉴런 사이의 연결



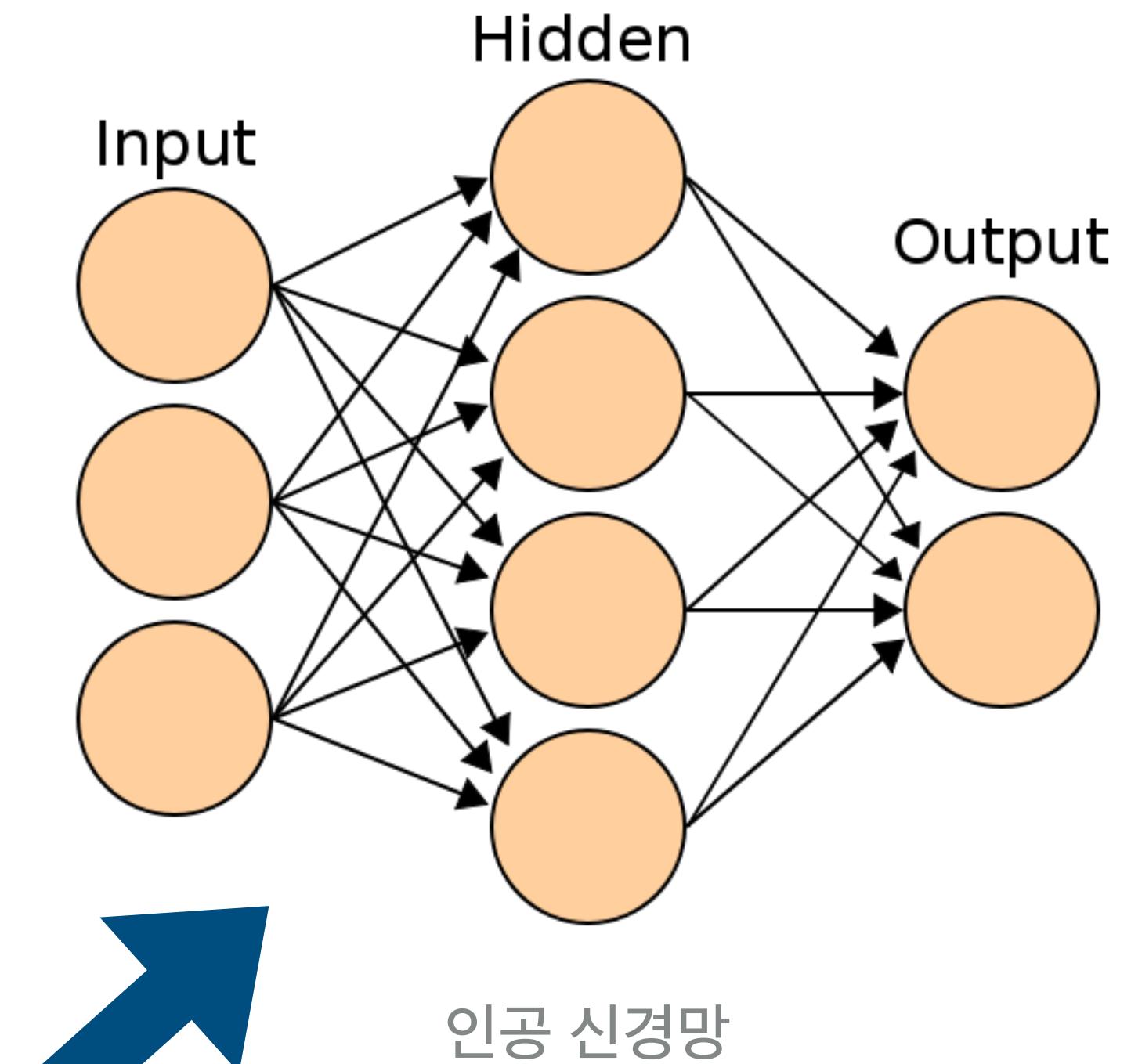
인간학습 → 기계학습, 인간두뇌 → 인공신경망



뉴런

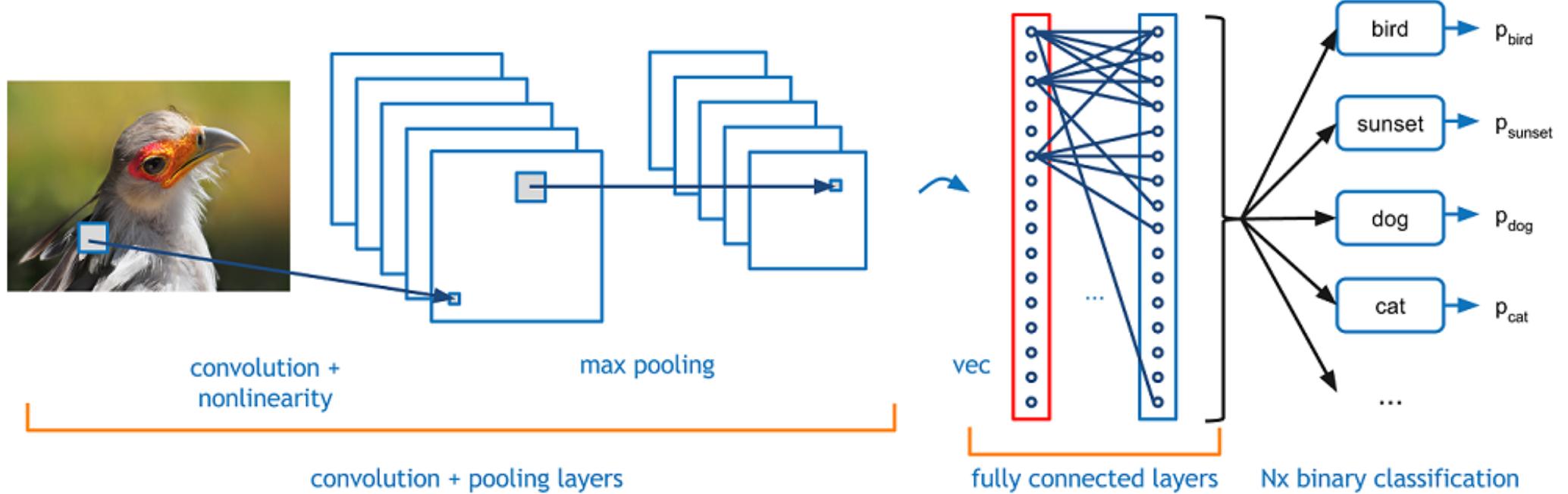


인공 뉴런

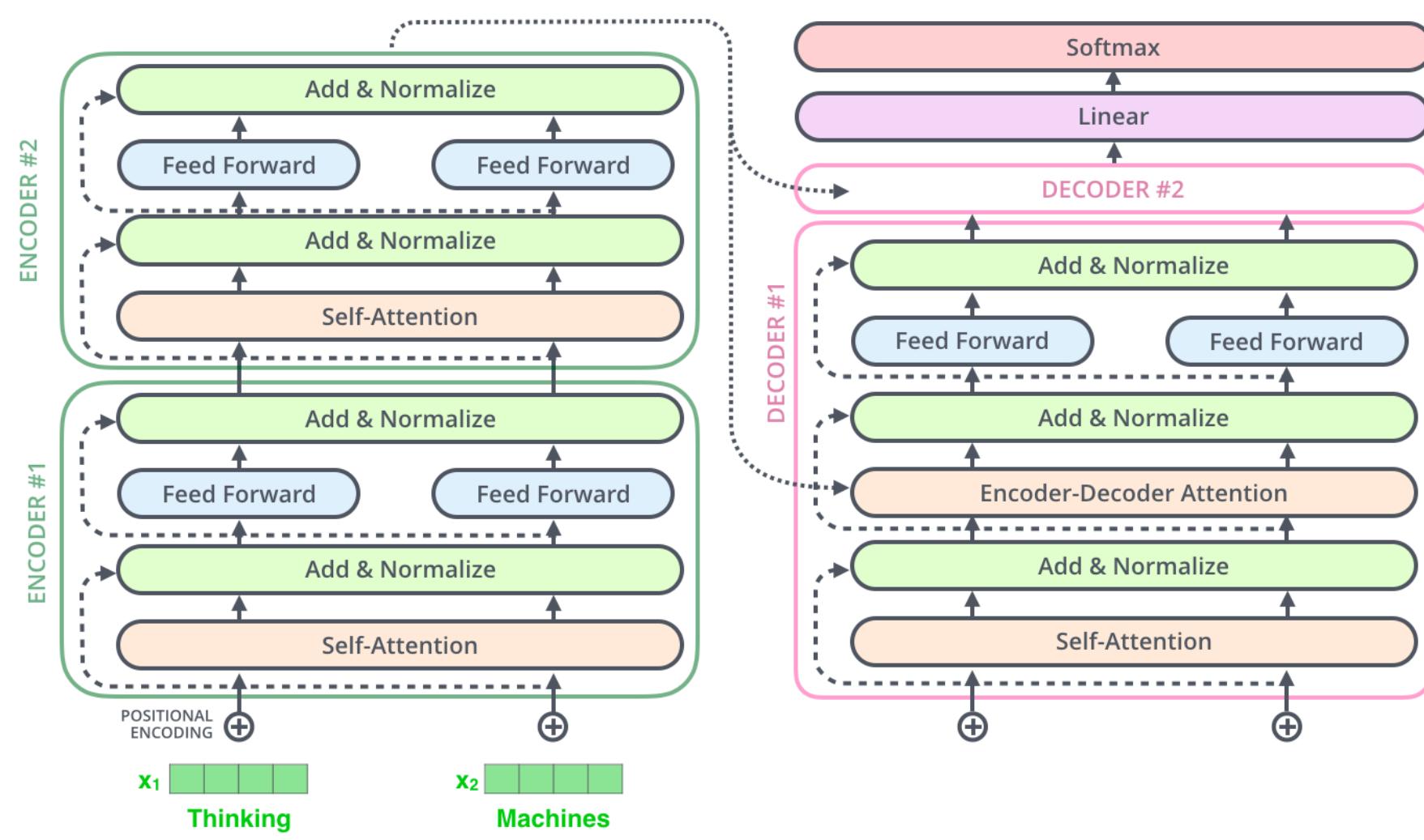


인공 신경망

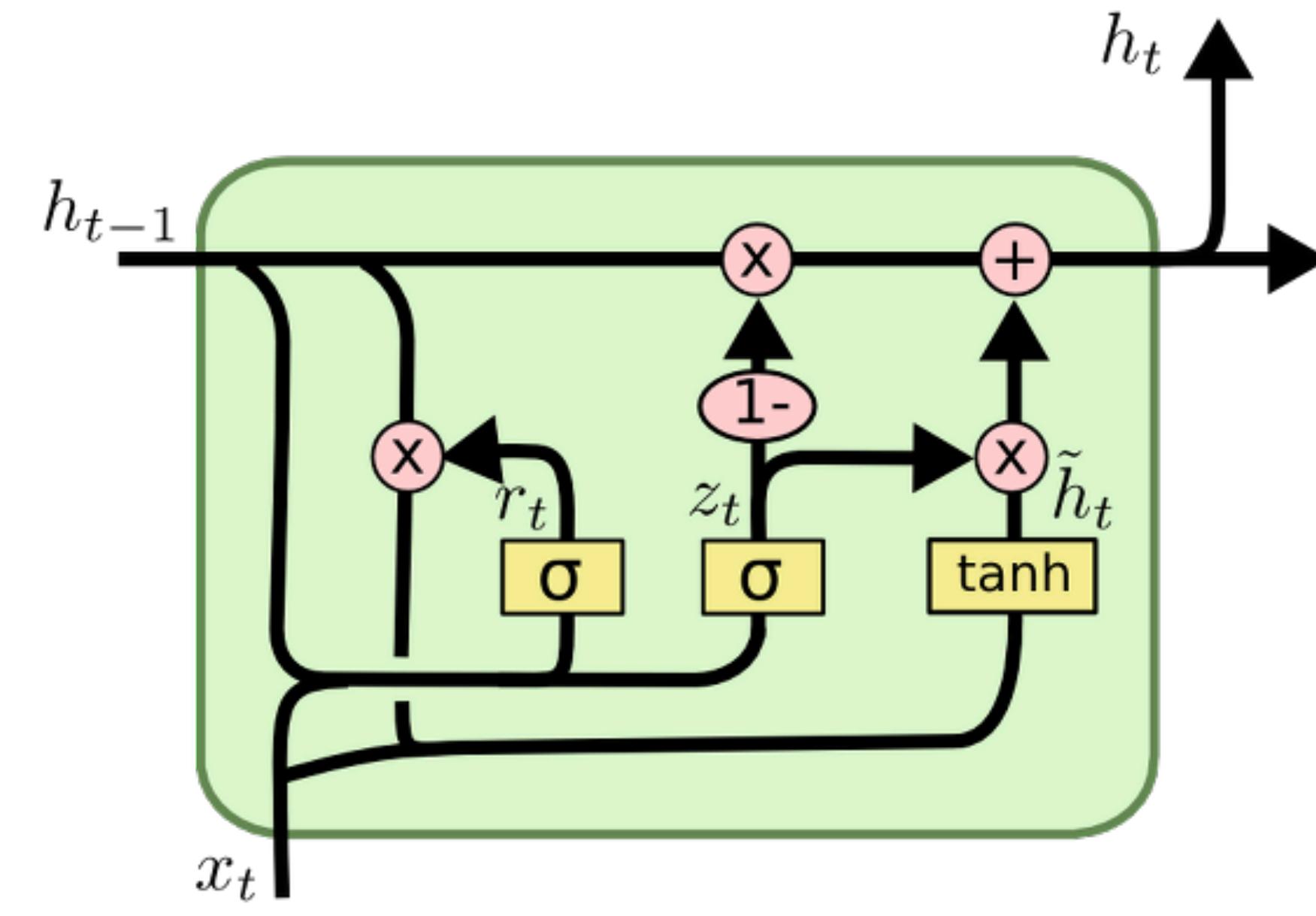
다양한 딥러닝 모델 구조



Convolutional neural networks



Transformer

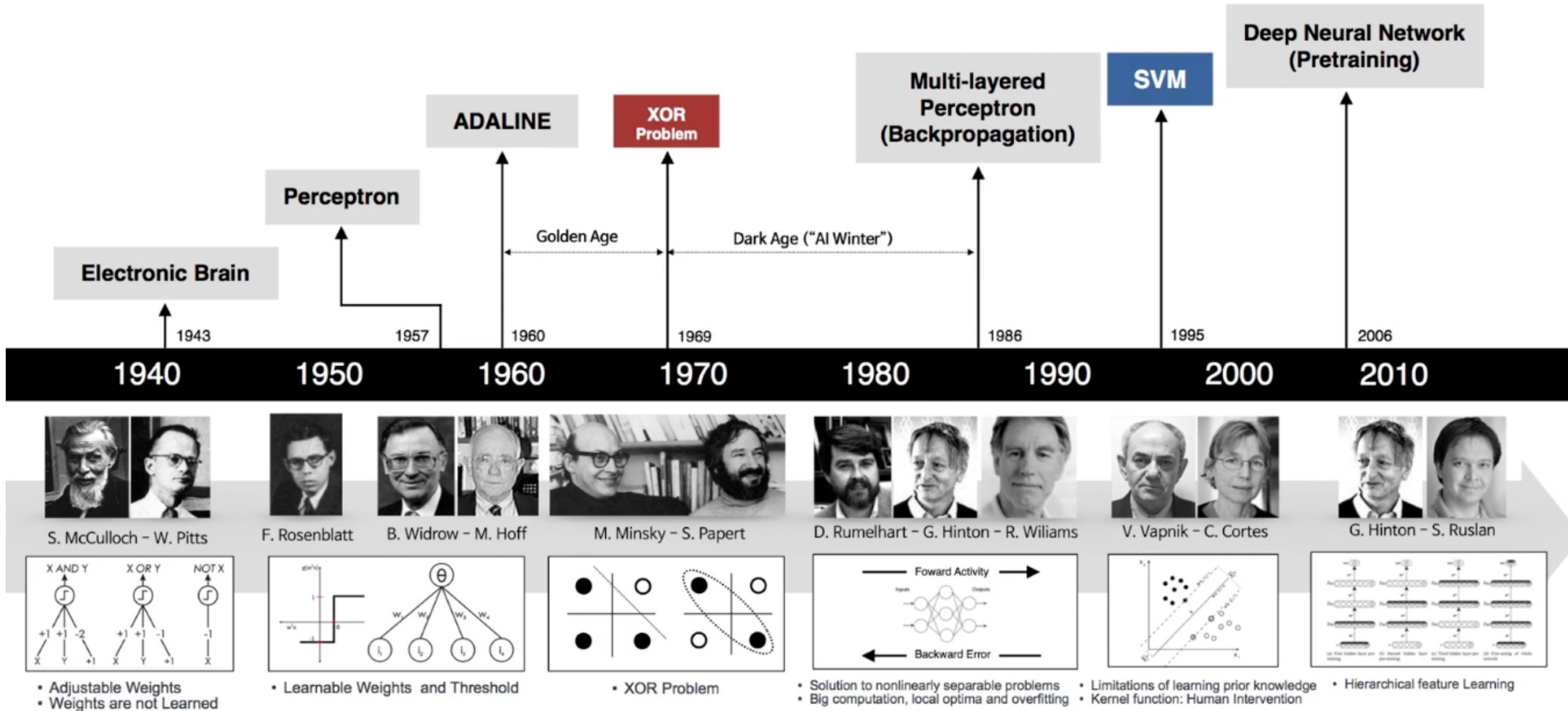


Long short-term memory

딥러닝이란

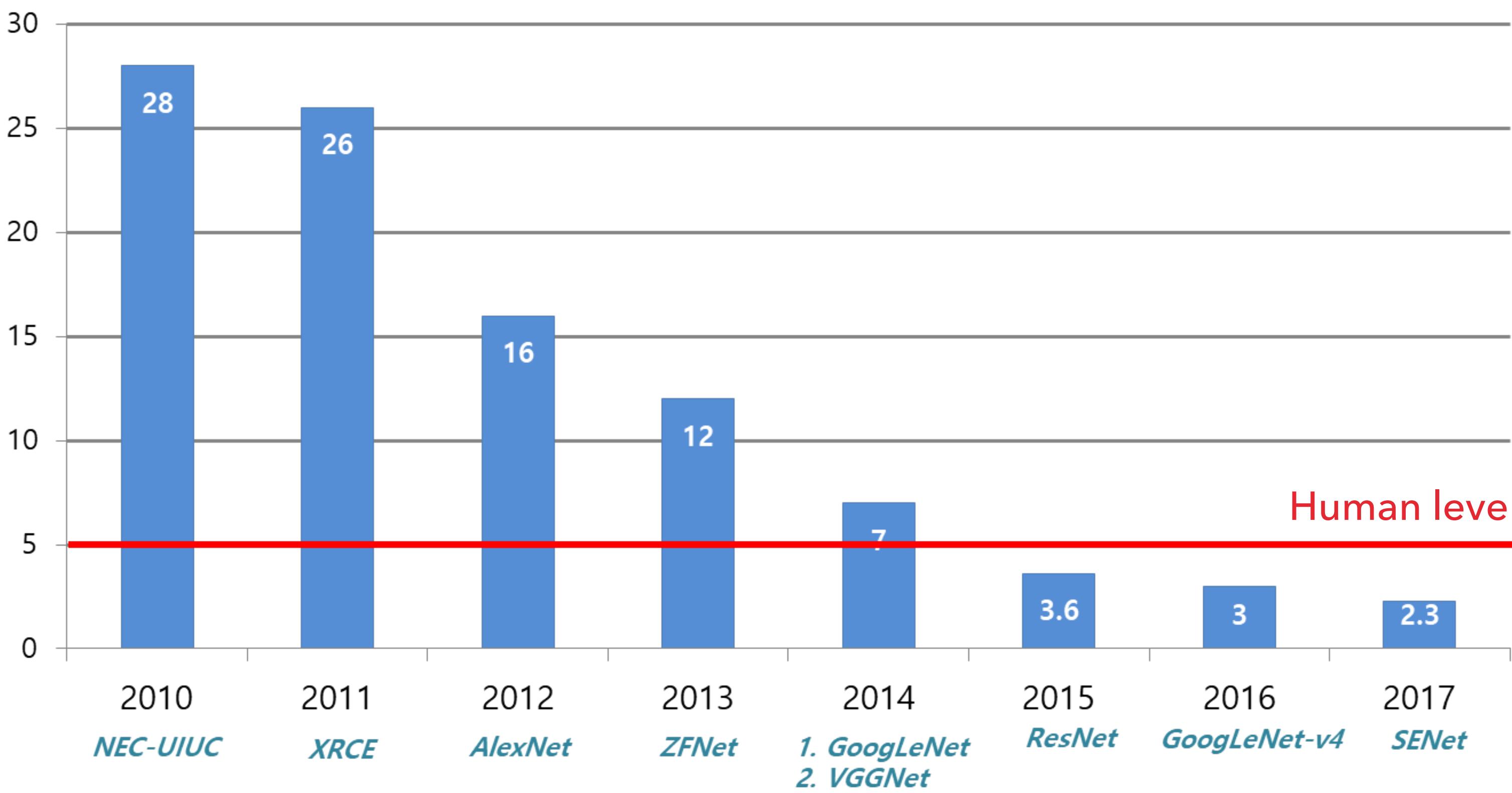
심층신경망 + 빅데이터 + 기계학습 프로세스

생각 보다 긴 딥러닝의 역사



인간을 넘어

Classification error of winning algorithms



바둑을 넘어

Configuration and strength ^[21]			
Versions	Playing hardware ^[22]	Elo rating	Matches
AlphaGo Fan	176 GPUs, ^[2] distributed	3,144 ^[1]	5:0 against Fan Hui
AlphaGo Lee	48 TPUs, ^[2] distributed	3,739 ^[1]	4:1 against Lee Sedol
AlphaGo Master	4 TPUs, ^[2] single machine	4,858 ^[1]	60:0 against professional players; Future of Go Summit
AlphaGo Zero (40 days)	4 TPUs, ^[2] single machine	5,185 ^[1]	100:0 against AlphaGo Lee 89:11 against AlphaGo Master
AlphaZero (34 hours)	4 TPUs, single machine ^[7]	4,430 (est.) ^[7]	60:40 against a 3-day AlphaGo Zero

https://en.wikipedia.org/wiki/AlphaGo_Zero

ELO rating / 2023

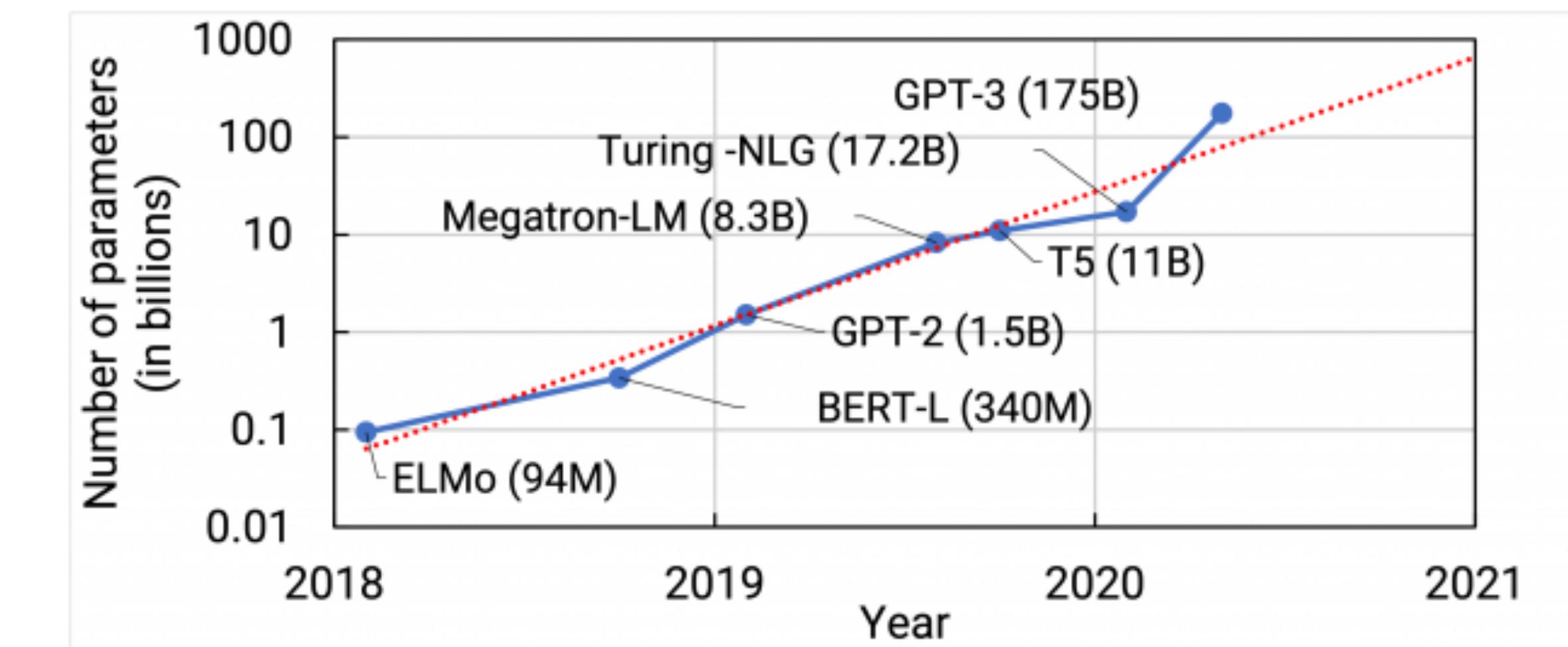
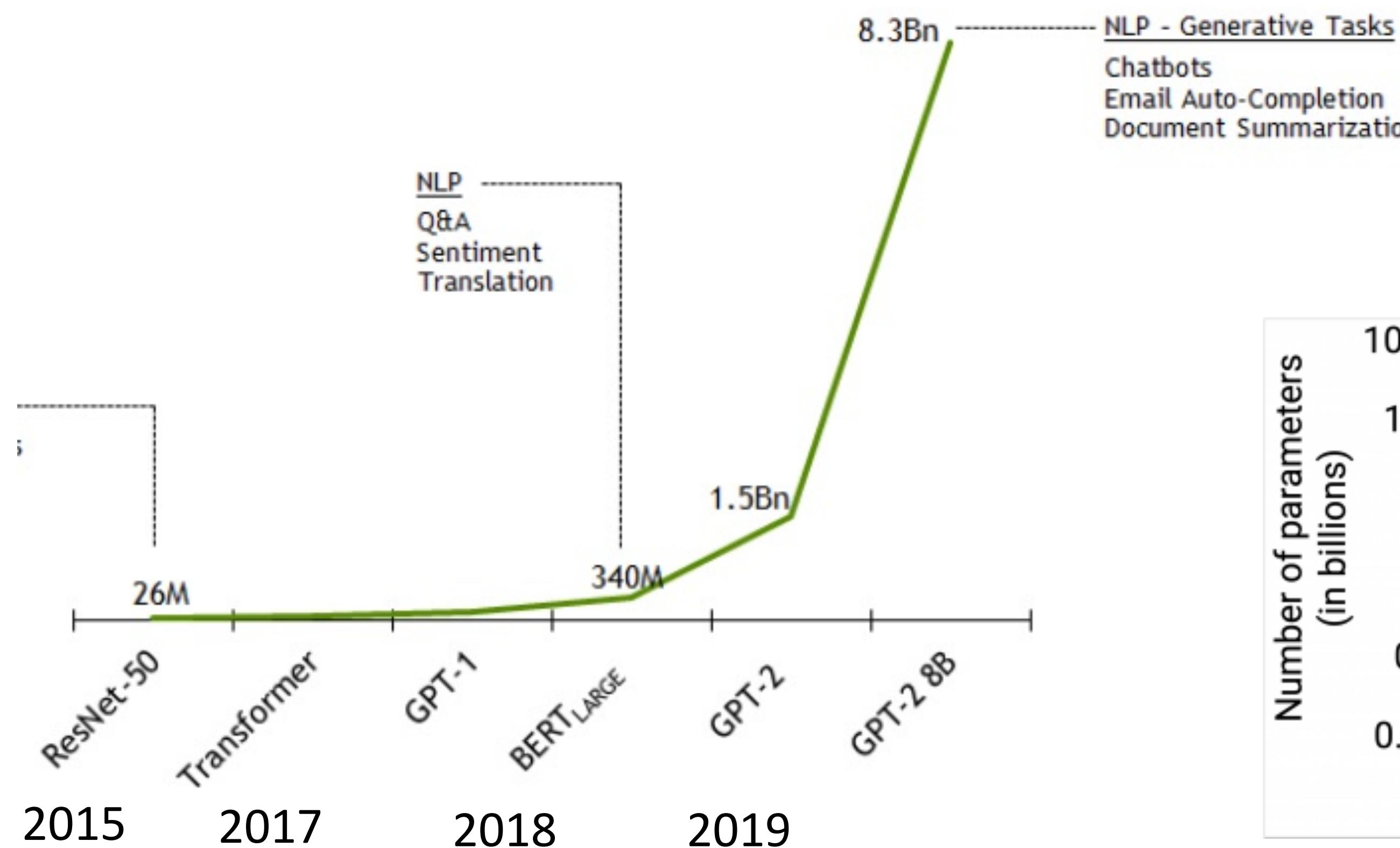
Rank	Name	♂♀	Flag	Elo
1	Shin Jinseo	♂		3826
2	Ke Jie	♂		3716
3	Park Junghwan	♂		3704

<https://www.goratings.org/en/>

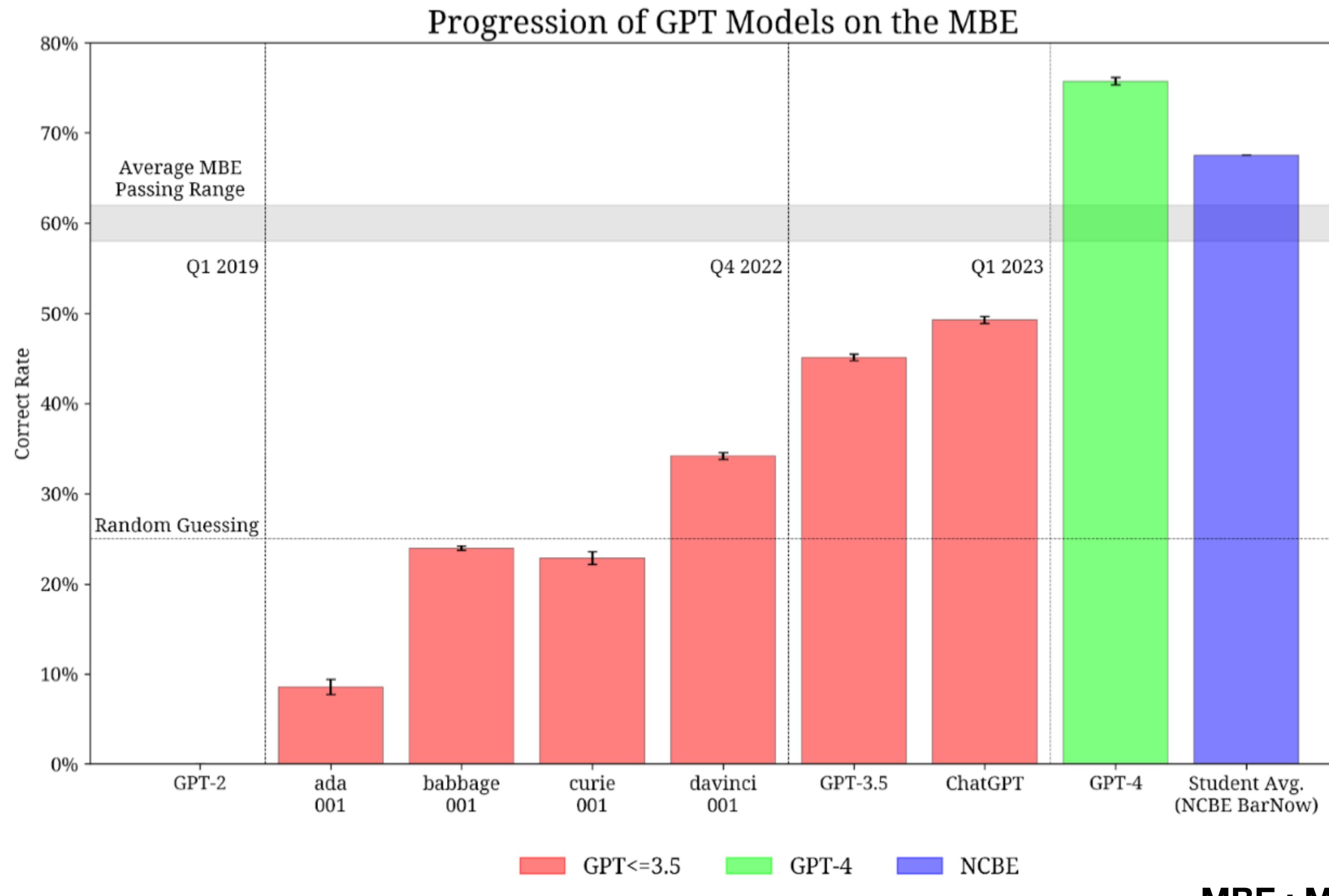
ELO rating / 2024

Rank	Name	♂♀	Flag	Elo
1	Shin Jinseo	♂		3854
2	Park Junghwan	♂		3697
3	Ding Hao	♂		3670

점점 더 커지는 신경망 모델



변호사 시험도?



최근 모델들

- GPT-3 : 1750억개
- PaLM : 540B
- GPT-4 : 1조개 이상?
- Llama 3.1 : 8B(20GB), 70B(160GB), 405B(4050억, 810GB)

인공지능의 가능성 - 예술



AI(인공지능) 프로그램 미드저니가 생성한 '스페이스 오페라 극장'. 글을 입력하면 거기에 맞는 이미지를 만들어주는 이 프로그램으로 만든 작품이 미국 콜로라도주의 한 미술대회에서 디지털 아트 부문 1위를 차지하면서 예술 범위를 둘러싼 논쟁이 확산하고 있다. [사진 위키미디어]

인공지능의 가능성 - 예술

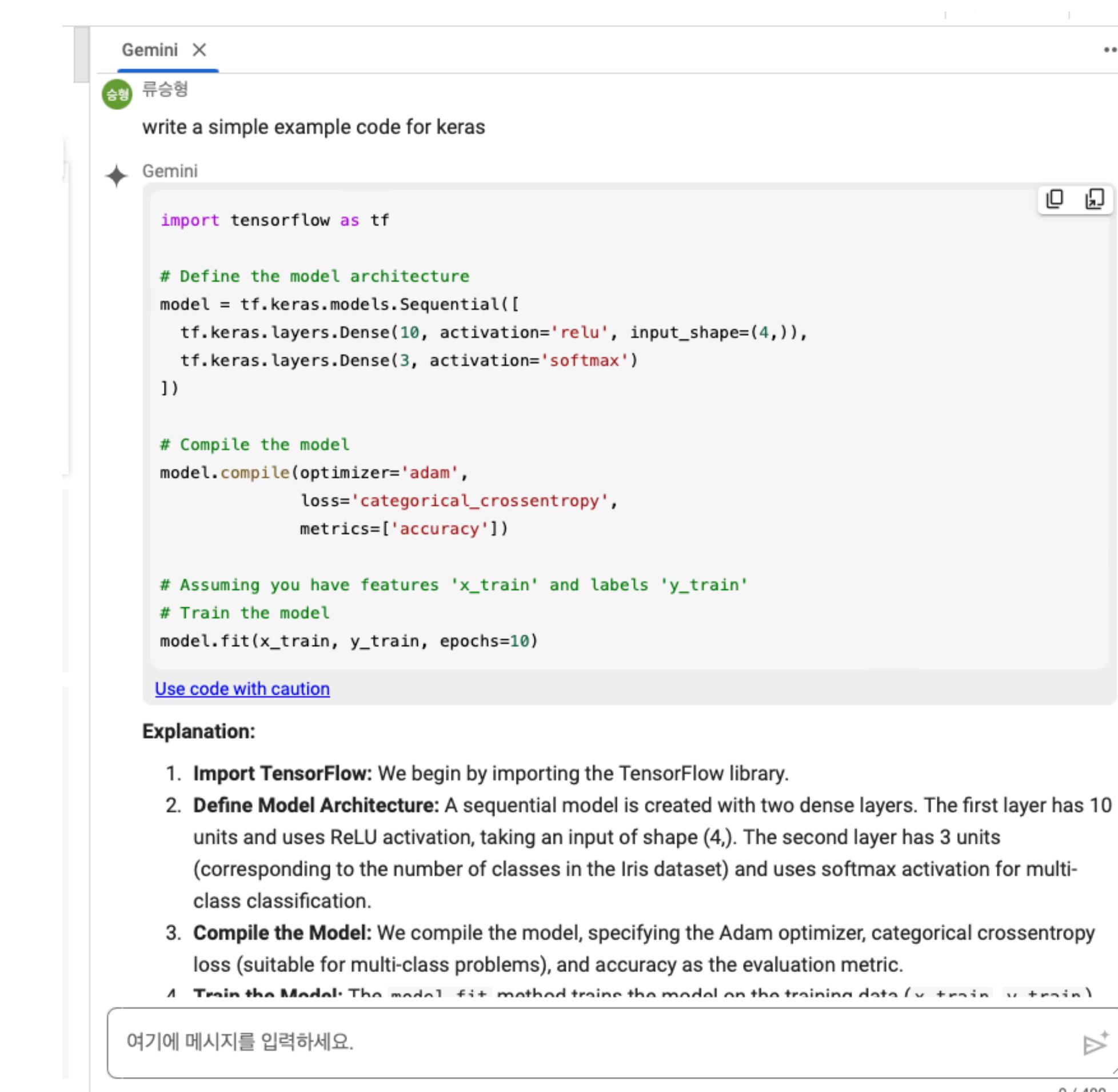


'1839 어워드'의 AI 부문 동상을 차지한 '플라밍곤(FLAMINGONE)'

인공지능의 가능성 - 코딩



The image shows the GitHub Copilot landing page. At the top, there are two tabs: "Open Source" and "Product". Below them, the main title is "Introducing GitHub Copilot: your AI pair programmer". A sub-section below the title says, "Today, we're launching a technical preview of GitHub Copilot, a new AI pair programmer that helps you write better code." A "Technical preview" button is visible. The central part of the page features the text "Your AI pair programmer" in large, bold, white font. Below this, there's a code editor window showing a snippet of JavaScript for fetching a NASA picture of the day. The GitHub Copilot logo is at the bottom left.



The image shows a screenshot of the Gemini AI interface. At the top, it says "Gemini X" and "Luisberg". Below that, a prompt reads "write a simple example code for keras". The AI-generated code is displayed in a code editor:

```
import tensorflow as tf

# Define the model architecture
model = tf.keras.models.Sequential([
    tf.keras.layers.Dense(10, activation='relu', input_shape=(4,)),
    tf.keras.layers.Dense(3, activation='softmax')
])

# Compile the model
model.compile(optimizer='adam',
              loss='categorical_crossentropy',
              metrics=['accuracy'])

# Assuming you have features 'x_train' and labels 'y_train'
# Train the model
model.fit(x_train, y_train, epochs=10)
```

Below the code, there's a link "Use code with caution". An "Explanation:" section follows, containing four numbered steps:

1. Import TensorFlow: We begin by importing the TensorFlow library.
2. Define Model Architecture: A sequential model is created with two dense layers. The first layer has 10 units and uses ReLU activation, taking an input of shape (4,). The second layer has 3 units (corresponding to the number of classes in the Iris dataset) and uses softmax activation for multi-class classification.
3. Compile the Model: We compile the model, specifying the Adam optimizer, categorical crossentropy loss (suitable for multi-class problems), and accuracy as the evaluation metric.
4. Train the Model: The `model.fit` method trains the model on the training data (`x_train`, `y_train`)

At the bottom, there's a text input field with the placeholder "여기에 메시지를 입력하세요." and a send button with a right arrow icon.

인공지능의 가능성 - 코딩

Tab

Cursor includes a powerful autocomplete that predicts your next edit. Once enabled, it is always on and will suggest edits to your code across multiple lines, taking into account your recent changes.

Code Generation

Cursor sees your recent changes, so it can predict what you want to do next.

```
You, 9 seconds ago | 1 author (You) | 0 implementations
7 pub struct EncodedMessageQueue {
6     pub queue: Vec<EncodedMessage, ClientFilter>,
5
4     sender: Arc<Sender<Vec<EncodedMessage, ClientFilter>>>,
3     receiver: Arc<Receiver<Vec<EncodedMessage, ClientFilter>>>,
2 }
1
6 impl EncodedMessageQueue {      You, 8 seconds ago * Uncommitted changes
5     pub fn new() -> Self {
4         let (sender, receiver) = crossbeam_channel::unbounded();
3         Self {
2             queue: vec![],
1             sender: Arc::new(sender),
             receiver: Arc::new(receiver),
         }
     }
1 }
```

Multi-Line Edits

Cursor can suggest multiple edits at once, saving you time.

```
.is_see_through(true) BlockBuilder
.light_reduce(true) BlockBuilder
.transparent_standalone(true) BlockBuilder
.build(),
Block::new(name: "Orange Concrete Block").id(5006).build(),
Block::new(name: "Blue Concrete Block").id(5007).build(),
Block::new(name: "Red Concrete Block").id(5008).build(),
Block::new(name: "White Concrete Block").id(5009).build(),
Block::new(name: "Ivory").id(5012).build(),
make_stairs(name: "Oak Stairs", id: 5013),
make_stairs(name: "Ivory Stairs", id: 5014),
make_stairs(name: "Stone Stairs", id: 5015),
```

인공지능의 가능성 - 코딩



ChatGPT

RS

오토ML에 대해 공부하고 싶어. 무엇부터 해야할까? 기초적인 내용을 알려줘



오토ML(AutoML)은 기계 학습 모델의 선택, 튜닝, 전처리 등을 자동화하여 개발자나 데이터 과학자들이 더 쉽게 기계 학습 모델을 개발할 수 있도록 돕는 기술입니다. 또한, 오토ML은 기존의 기계 학습 프로세스에 필요한 전문 지식을 덜 요구하기 때문에 더 많은 사람들이 데이터 과학을 활용할 수 있게 합니다.



오토ML에 대해 배우기 위해서는 다음과 같은 기본적인 내용들을 이해해야 합니다:

- 기계 학습 기본 지식:** 오토ML을 제대로 이해하려면 기계 학습에 대한 기본적인 이해가 필요합니다. 이에는 지도 학습, 비지도 학습, 강화 학습 등의 기본적인 학습 알고리즘과 특성 엔지니어링, 모델 선택 및 튜닝 등의 기본적인 개념이 포함됩니다.
- 데이터 전처리 및 특성 엔지니어링:** 이 단계에서는 데이터를 모델링하기 적합한 형태로 변환하는 방법에 대해 배웁니다. 또한, 원시 데이터에서 유용한 정보를 추출하거나 변환하는 특성 엔지니어링 기법에 대해서도 배웁니다.
- 모델 선택 및 하이퍼파라미터 튜닝:** 오토ML의 핵심은 모델 선택 및 하이퍼파라미터 튜닝을 자동화하는 것입니다. 이를 위해, 다양한 기계 학습 알고리즘과 이들의 하이퍼파라미터에 대해 이해하는 것이 필요합니다. 또한, 그리드 탐색, 랜덤 탐색, 베이지안 최적화 등의 튜닝 기법에 대해서도 배웁니다.
- 모델 평가 및 검증:** 모델이 얼마나 잘 훈련된 모델이 새로운 데이터에 얼마나

↳ Regenerate response

생성형 딥러닝 모델의 시대

- 딥러닝 모델이 글을 이해하고 그림을 그리는 시대.
 - 거대 모델 + 거대 데이터
- 여전히 기본적 머신러닝 모델이나 오토ML에 대한 공부를 해야할까?

인공지능 결과물의 활용

- 결과물에 대한 책임을 누가 질 것인가?

글로벌

챗GPT가 알려준 '가짜 판례' 제출한 美 변호사, 법원서 징계 위기

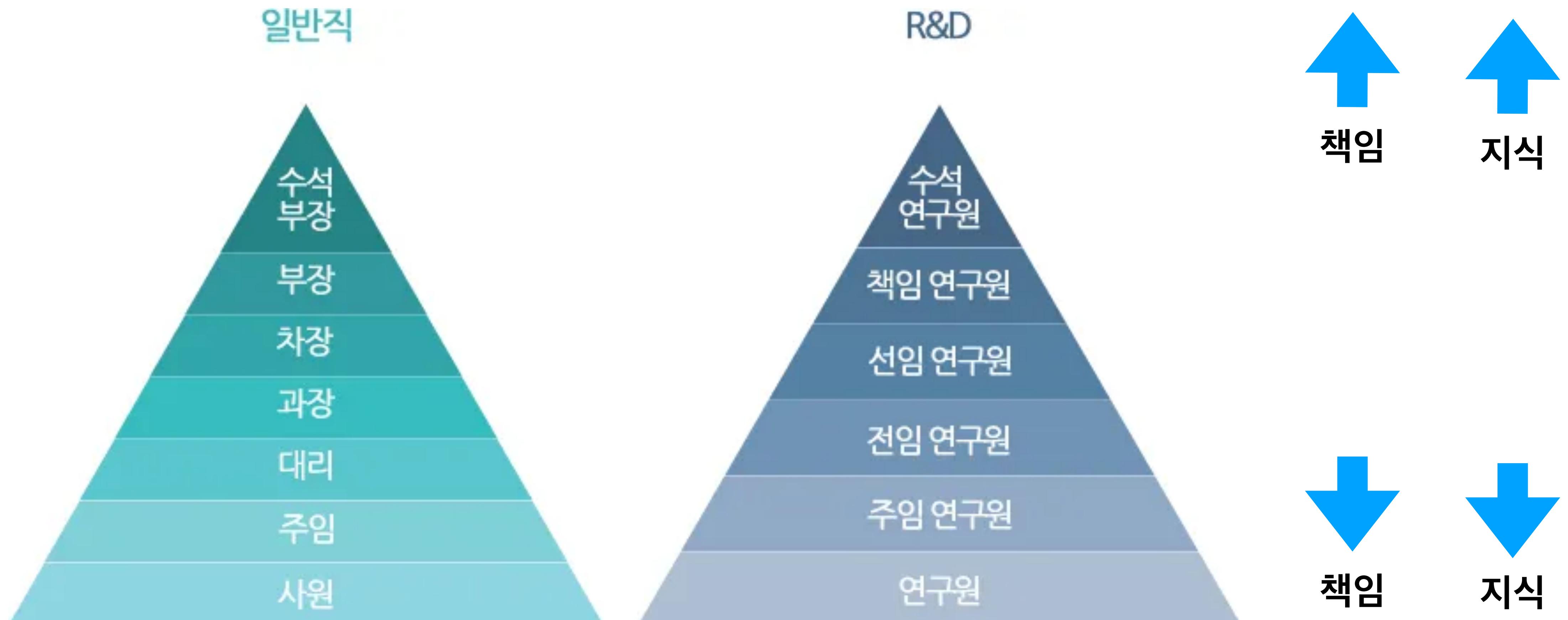
홍윤지 기자 | 2023-05-29 14:13



美 법원, 청문회 거쳐 변호사 윤리법 위반 여부 판단할 듯

<https://www.lawtimes.co.kr/news/187960>

회사의 체계

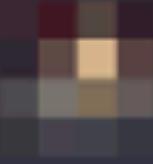


AI로 외주 맡기기



- AI에게 나의 태스크 일부를 할당
- AI는 빠르게 해결하여 생산성 향상
- 책임이 더욱 늘어날 것...

Hallucination - 2023

 조선왕조실록에 기록 된 세종대왕의 맥북프로 던짐 사건에 대해 알려줘

 세종대왕의 맥북프로 던짐 사건은 역사서적인 조선왕조실록에 기록된 일화로, 15세기 조선시대 세종대왕이 새로 개발한 훈민정음(한글)의 초고를 작성하던 중, 문서 작성 중단에 대한 담당자에게 분노하여 맥북프로와 함께 그를 방으로 던진 사건입니다.

세종대왕은 훈민정음 개발을 위해 열심히 노력하던 중, 한글 문자를 이해하지 못하는 당시의 관료들이 문서 작성을 돋기 위해 임명한 기획관 중 한 명인 최환(崔煥)에게 작성 중단 사실을 들게 되었습니다. 세종대왕은 이에 분노하여 최환을 부르고, 맥북프로를 함께 들고 그를 방으로 불러들이며 화를 내었습니다. 이후, 세종대왕은 훈민정음의 초고 작성을 다른 담당자에게 넘겨주게 되었습니다.

이 사건은 조선시대의 통치자로서 세종대왕이 어떻게 민족문화를 발전시키기 위해 노력했는지를 보여주는 일화로 유명합니다. 또한, 세종대왕의 열정과 업무처리에 대한 철저한 태도를 보여주는 사례로도 언급되고 있습니다.

한국일보

Hallucination - 2024

조선왕조실록에 기록된 맥북프로던짐사건에 대해 이야기해줘



"맥북프로던짐사건"은 조선왕조실록에 기록된 사건이 아니며, 현대의 디지털 기기인 맥북 프로와는 전혀 관련이 없습니다. 조선왕조실록은 조선 왕조의 역사와 관련된 기록으로, 1392년부터 1910년까지 약 500여 년간의 역사를 다룹니다.

이와 관련된 내용을 찾을 수는 없지만, 대신 조선왕조실록에 기록된 실제 사건 중 흥미로운 예를 들어볼 수 있습니다. 예를 들어, 세조 때의 사육신 사건, 연산군의 폭정, 임진왜란 등 다양한 사건들이 기록되어 있습니다. 조선왕조실록은 매우 상세하고 방대한 역사 기록으로, 조선 시대의 정치, 사회, 문화, 경제 등 다양한 측면을 담고 있습니다.

혹시 다른 조선왕조실록의 실제 사건에 대해 궁금한 점이 있다면, 말씀해 주세요.



Hallucination - 2024

조선왕조실록에 기록된 막복불호 투척 사건에 대해 이야기해줘



"막복불호 투척 사건"은 조선왕조실록에 기록된 실제 사건 중 하나입니다. 이 사건은 조선 세조(재위 1455-1468) 때 발생한 일로, 당시의 정치적 긴장과 권력 다툼의 일면을 보여줍니다.

사건 배경

세조는 단종을 폐위하고 왕위에 오른 후, 자신의 정권을 안정시키기 위해 많은 반대 세력들을 제거했습니다. 이 과정에서 권력 다툼과 암투가 끊이지 않았습니다. 막복불호 투척 사건은 이러한 정치적 불안정 속에서 벌어진 일입니다.

사건의 전개

이 사건은 세조의 측근이었던 한명회와 관련이 있습니다. 한명회는 세조의 즉위를 도운 공신으로, 당시 조선의 실세 중 한 명이었습니다. 하지만 한명회의 권력 남용과 전횡에 대해 불만을 가진 사람들이 많았습니다.

어느 날 한명회의 집 앞에서 '막복불호(莫僕不呼)'라고 쓰인 글씨가 적힌 종이를 던진 사건이 발생했습니다. 이 글귀는 "나의 종이 아니면 부르지 않는다"라는 의미로, 한명회의 전횡을 비판하는 내용이었습니다. 이 사건은 단순한 반항이 아니라, 당시에 한명회의 권력 남용에 대한 사회적 불만이 얼마나 커는지를 보여주는 사건이었습니다.

Hallucination - 2024

막복불호의 뜻이 뭐야

"막복불호(莫僕不呼)"의 뜻은 "나의 종이 아니면 부르지 않는다"입니다. 여기서 각 한자의 의미는 다음과 같습니다:

- 莫 (막): "없다" 또는 "하지 말라"
- 僕 (복): "종"
- 不 (불): "아니다"
- 呼 (호): "부르다"

이를 직역하면 "종이 아니면 부르지 않는다"라는 의미로 해석할 수 있습니다. 이 표현은 특정한 사람의 권력 남용이나 독재적인 태도를 비판하는 의미로 사용될 수 있습니다. 따라서 막복불호 투척 사건은 한명회의 권력 남용을 비판하는 사회적 불만을 상징적으로 나타낸 사건입니다.



Hallucination - 2024

세종대왕 막복불호 사건에 대해서 알려줘

❖ Sources

세종 1397 ~ 1450 - 우리역
사넷

contents.history.... · 1

세종 - 위키백과, 우리 모두의
백과사전

ko.wikipedia · 2

세종(조선) - 나무위키

namu · 3



View 5 more

✳️ Perplexity

세종대왕의 막복불호(莫復不好) 사건은 세종이 세자로 책봉되기 전, 충녕대군 시절에 일어난 중요한 사건입니다. 이 사건은 세종의 인품과 정치적 능력을 잘 보여주는 일화로 알려져 있습니다.

사건의 배경

세종이 세자로 책봉되기 전, 그의 형인 양녕대군이 세자의 자리에 있었습니다. 그러나 양녕대군은 학업에 소홀했고 자기 절제 능력이 부족하여 여러 가지 문제를 일으켰습니다 ¹. 이로 인해 태종은 양녕대군의 세자 자격에 대해 고민하게 되었습니다.

Hallucination - 2025

세종대왕 막복불호 사건에 대해 알려줘

Perplexity

이미지

출처 6

1 task ↗

sejongnewspaper

[세종신문 - 세종의 눈으로 세상을 읽는
창] 임복비 사건



contents.history.go.kr

세종대왕, 훈민정음을 창제하다 - 우리역
사넷



chosun

“세종대왕의 맥북 던짐 사건에 대해 알려
줘” 했더니 챗GPT가 내놓은 ...

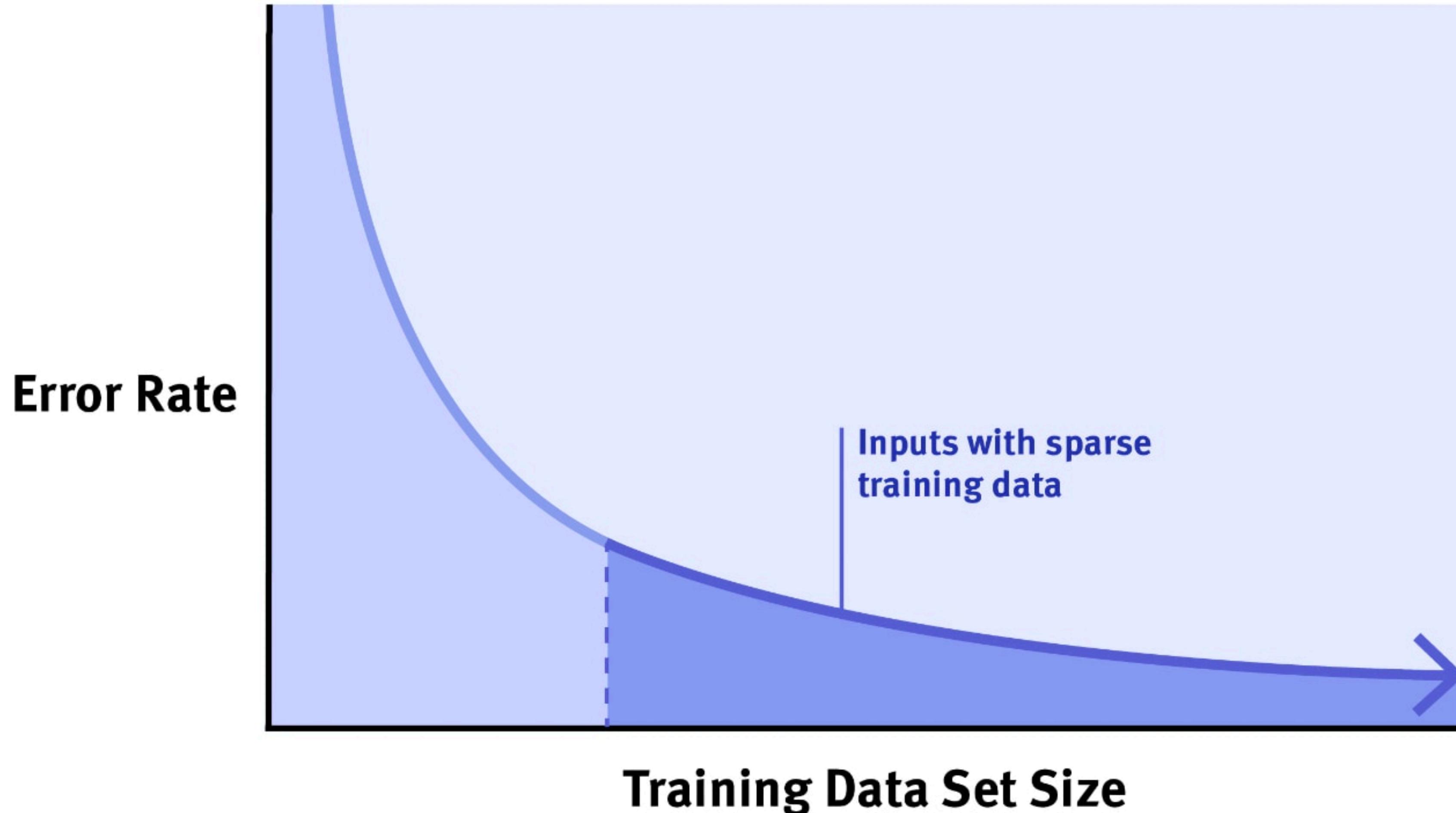
▶ W

+3 소스들



세종대왕 ‘막복불호 사건’에 대해 알려달라고 하셨는데, 실제 역사 기록이나 신뢰할 만한 자료에서는 ‘세종대왕 막복불호 사건’이라는 명칭이나 일화를 찾을 수 없습니다. 검색 결과와 위키백과, 주요 언론, 역사 콘텐츠 등에서도 해당 용어 혹은 사건에 대한 언급은 전혀 확인되지 않습니다.

데이터 기반 모델의 롱테일 문제



인공지능 = 인간?

- 인간과 상호작용하면서 나타날 수 있는 문제들이 인공지능에도 나타날 수 있음.
- 잘 모르는 부분에 대해서 확신에 찬 것처럼 이야기 → Hallucination
- 문맥에 대해 잘못 파악하는 경우 → “알아서 잘 딱 깔끔하고 센스있게”

정형 데이터 활용과 인공지능

- 자연어 데이터
 - CommonCrawl: 90TB (전처리 전), 9TB (전처리 후)
 - C4(Colossal Clean Crawled Corpus): 800GB
 - The Pile: 900GB
- 정형 데이터의 경우는?

롱 테일 영역의 산업 데이터

- 정형 데이터: 단변수, 고차원의 데이터
- 자연어는 임베딩을 통해 고차원벡터로 변환되지만 기본적으로 단변수 데이터로 볼 수 있음.
- 동일한 시스템이더라도 환경에 따라 전혀 다른 데이터의 특성을 가질 수 있음.
- 비슷한 목적을 가진 시스템이더라도 수집되는 변수 구성에 차이가 있을 수 있음.
- 같은 물리량을 측정하더라도 환경에 따라 스케일의 차이가 클 수 있음.

메타 인지 능력을 통한 적용

- 특정 산업데이터셋에 특화된 딥러닝 모델 또는 인공지능 연구
 - 데이터셋의 규모에 따라서 달라질 수 성능 편차가 클 수 있음.
 - 문제별 각개격파 >> 정형데이터의 경우 ML이 잘 동작함
- 문제를 풀 수 있는 과정을 알고있는 인공지능 연구
 - 문제를 이해하고 이를 풀 수 있는 코드를 작성 / 실험
 - >> 결과물에 대해 이해할 수 있는 역량이 필요함

다양한 머신러닝 모델들

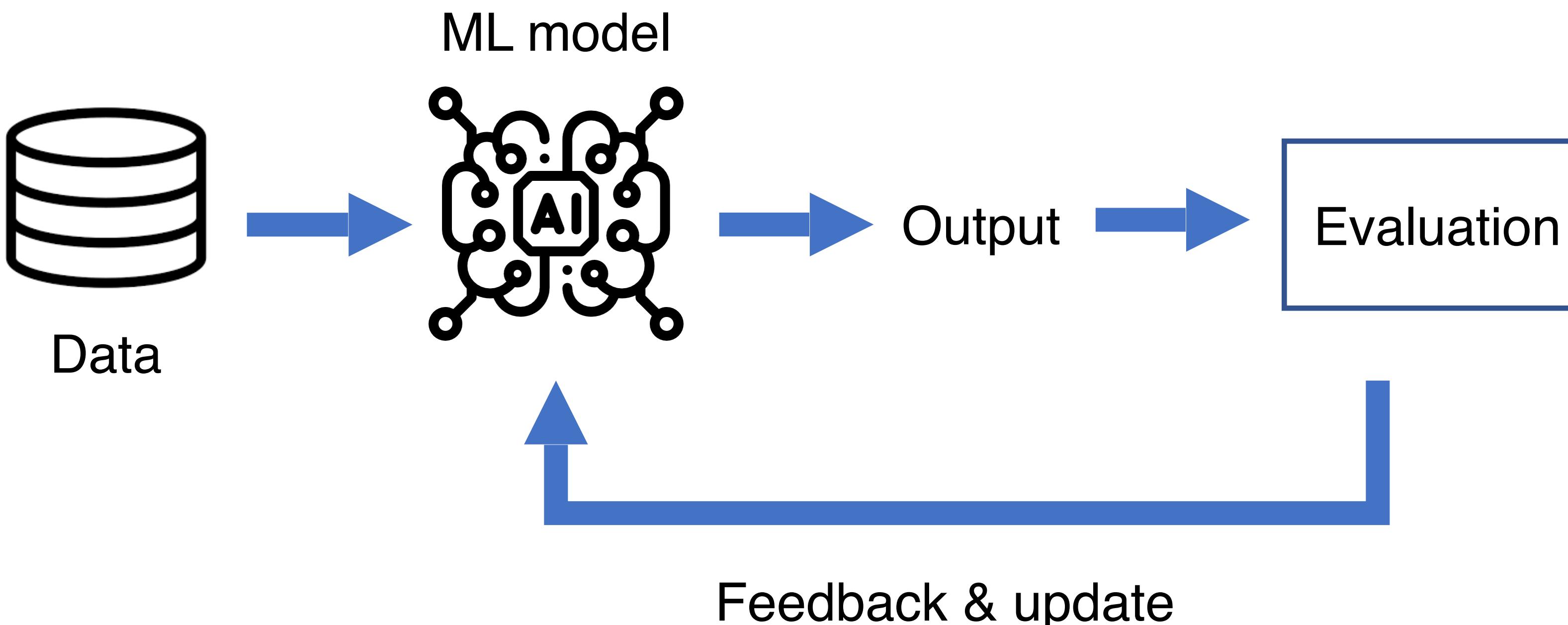
- 선형 회귀 (Linear Regression)
- 로지스틱 회귀 (Logistic Regression)
- 의사결정 트리 (Decision Tree)
- 랜덤 포레스트 (Random Forest)
- 그래디언트 부스팅 트리 (**Gradient Boosting Tree**)
 - XGBoost
 - LightGBM
 - CatBoost
- 서포트 벡터 머신 (Support Vector Machine, SVM)
- 나이브 베이즈 (Naive Bayes)
- K-최근접 이웃 (K-Nearest Neighbors, KNN)
- 신경망 (**Neural Networks**)
 - 다층 퍼셉트론 (Multi-Layer Perceptron, MLP)
 - 컨볼루션 신경망 (Convolutional Neural Network, CNN)
 - 순환 신경망 (Recurrent Neural Network, RNN)
 - 트랜스포머 (Transformer)

모델의 학습은?

- 머신러닝 모델에 따른 학습 알고리즘을 통해 학습 수행.
 - 선형 회귀: 최소제곱법, 경사하강법
 - K-평균 군집화: Expectation-Maximization
 - 의사결정 트리: 정보 이득, 지니 불순도
 - 신경망: 경사하강법 + 오류역전파법(Back propagation)

지금까지....

- 머신러닝 모델의 학습 프로세스
- 그러나 모델 개발 과정에서는 학습 외에도 다양한 과정들이 요구됨



필요한 과정들

- 어떤 문제를 풀 것인지?
- 가용한 데이터는 무엇인지?
- 데이터에 대한 변환이 필요한지?
- 어떤 머신러닝 모델을 사용할 것인지?
- 머신러닝 모델의 구조는 어떻게 설정할 것인지? 등등

AutoML

- AutoML (Automated Machine Learning) : 자동화된 기계 학습
- AutoML은 기계 학습 모델을 구축하고 최적화하는 전체 프로세스를 자동화하기 위한 기술과 방법을 의미.
- 머신러닝 모델 개발의 각