

인권활동가를 위한 거대언어모델 이해하기

인공지능은 왜 확률적 앵무새 일까?

1부 활동가도 이해하는 거대언어모델
인공지능은 어떻게 환경을 파괴하고 있을까

2부 인공지능이 공정하다는 오해 : 데이터의 함정
인공지능은 언어를 이해하고 있을까?

2024년 6월 21일 금요일 오후 3시
스테이션 사람, 서울 은평구 증산로17가길 15-7

인공지능(AI)과 거대언어모델(LLM)에 대한 이해

오병일

01 세미나의 취지와 개요

01 확률적 앵무새의 위험에 대하여 : 언어 모델이 너무 커도 되는가?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu

University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether



출처:

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

01

인공지능은 왜 편향적인가

챗GPT의 헛소리 현상은 왜 발생하는가

영화 Her와 같이 인공지능은 사람과 감정을 교류할 수 있는가

인공지능과 거대언어모델(LLM)을 알면 이해하기 쉬워진다!

이 질문들

01 왜 AI는 확률적 앵무새인가.

02 현재의 거대언어모델(LLM)은 방대한 리소스가 필요하다.
그래서 어떤 문제가 발생하는가.

03 거대언어모델(LLM)을 위한 방대한 데이터는 어디에서 나오나.
그래서 어떤 문제가 발생하는가.

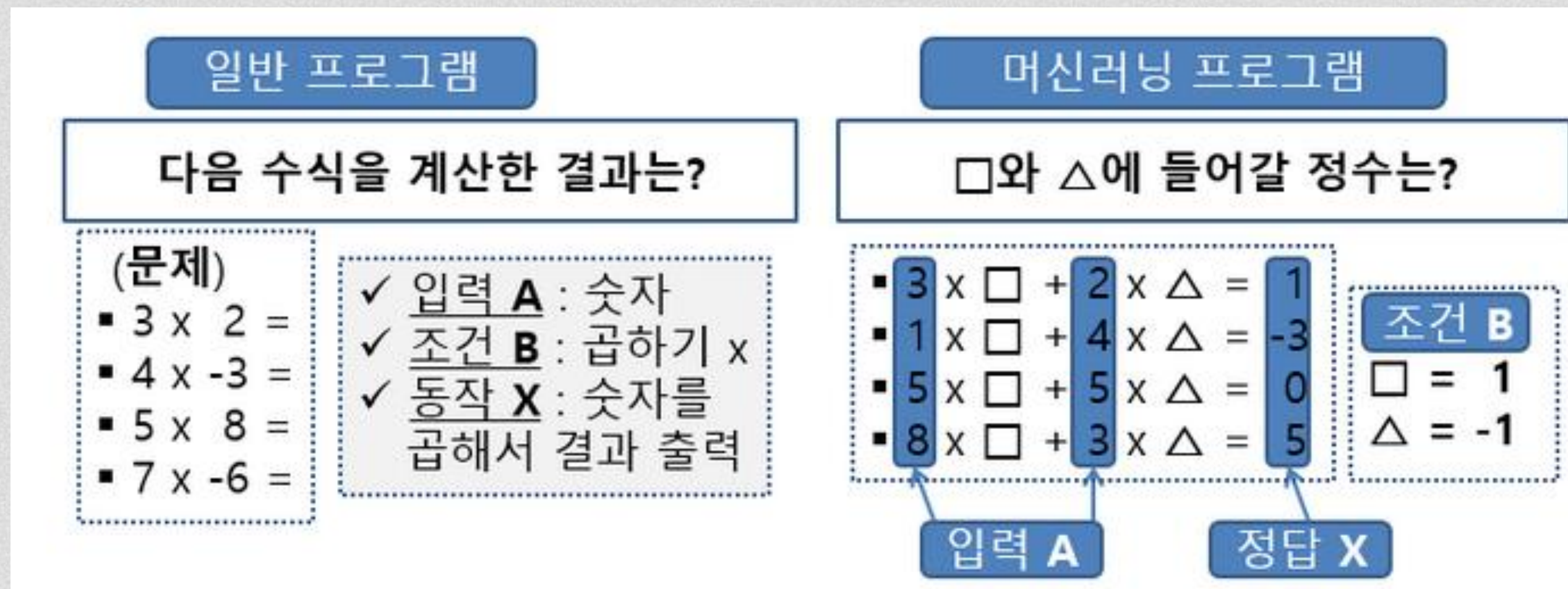
04 AI가 언어를 이해한다는 착각은 어떠한 문제를 야기하는가.

02 인공지능

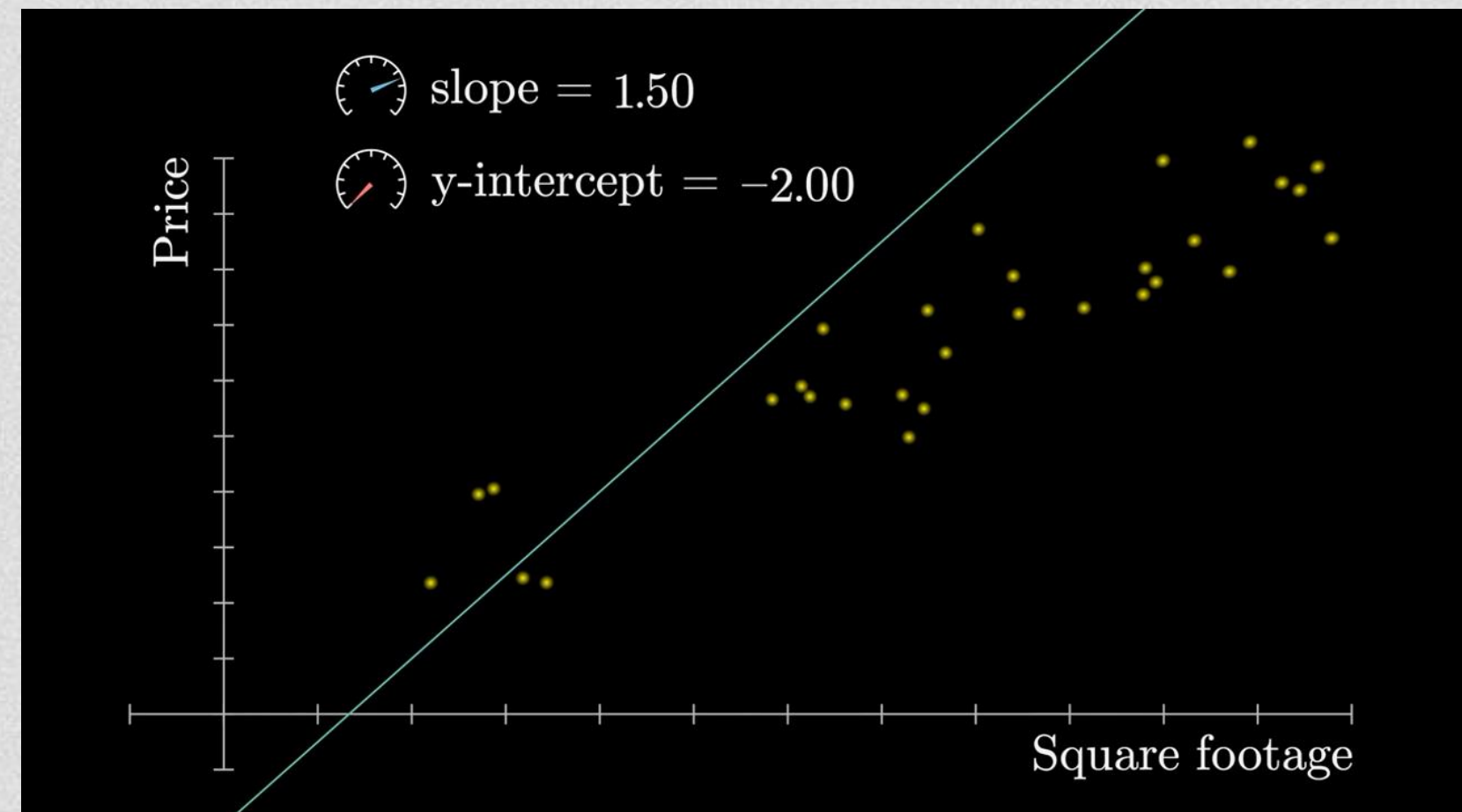
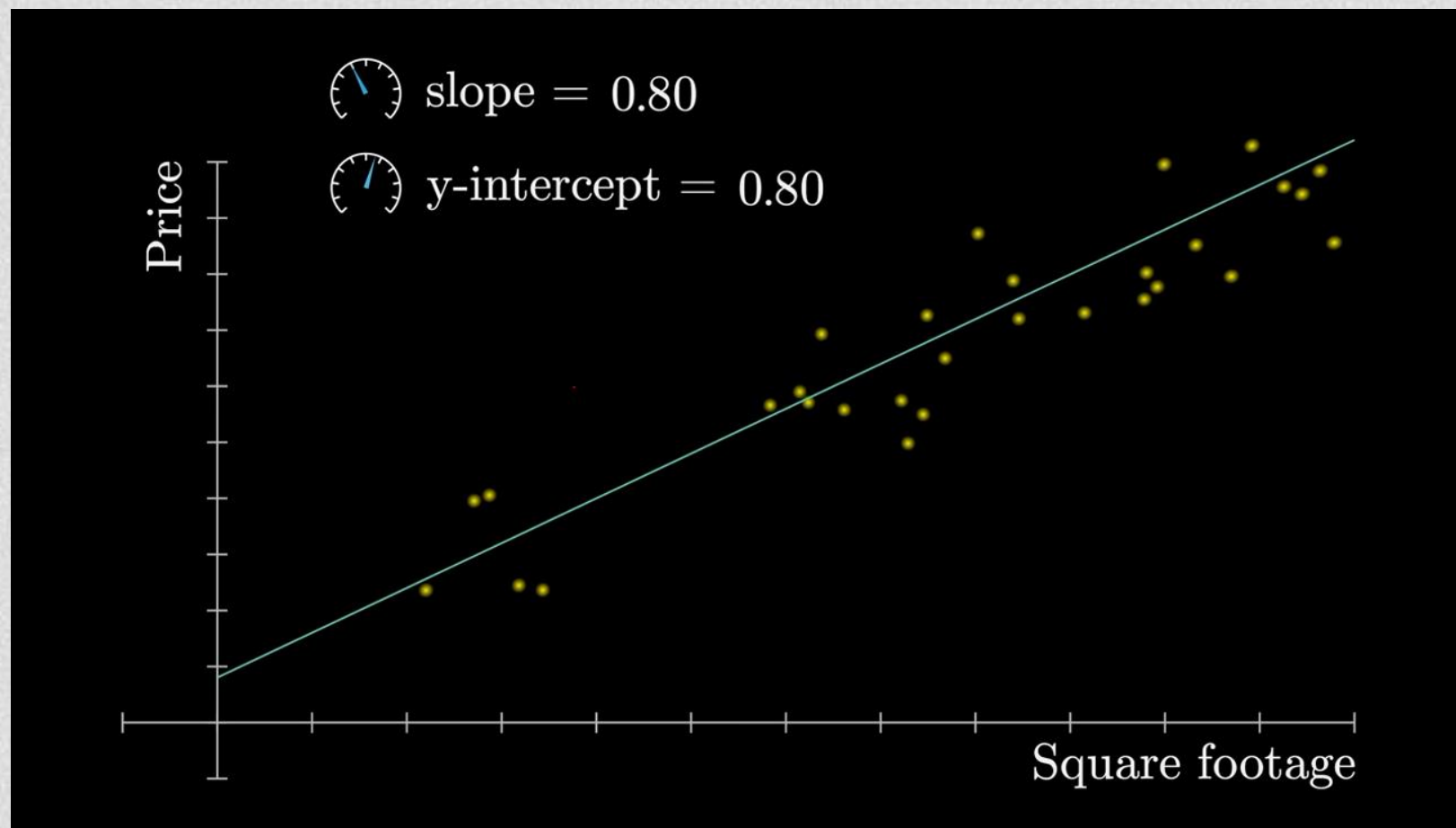
02 인공지능



02 일반적 프로그램과 머신러닝의 차이



02 인공지능이 '학습(learning)'한다는 것



출처 : <https://www.3blue1brown.com/lessons/gpt>

- 회귀 문제: 집 평수와 가격에 대한 데이터 학습, $y = ax + b$ 에서 a, b 값을 찾는 것
- 데이터 패턴을 가장 잘 반영할 수 있는 매개변수(parameter)를 찾는 것
- 실제 데이터(입력 데이터)와 예측치의 차이를 최소화하는 값을 찾는 것

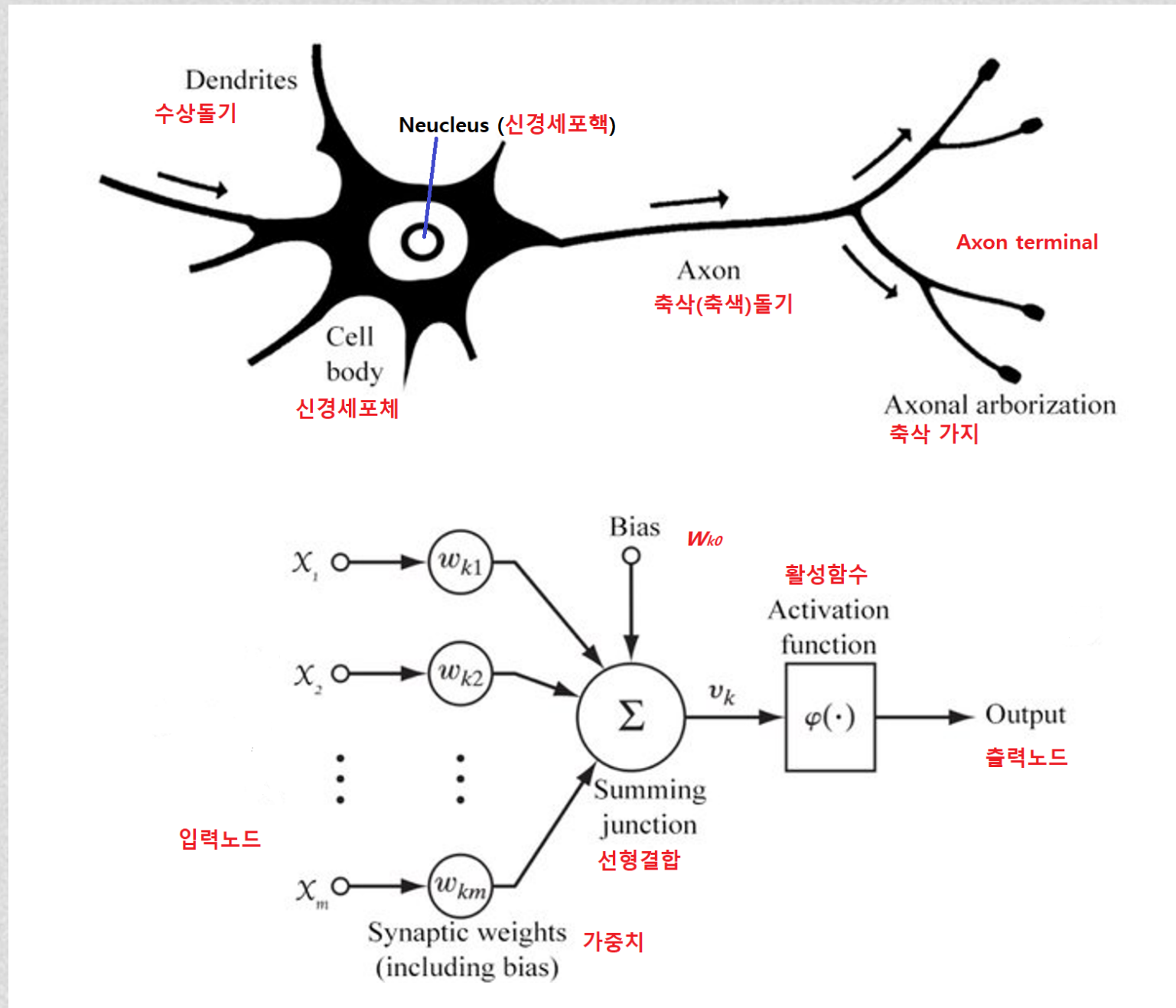
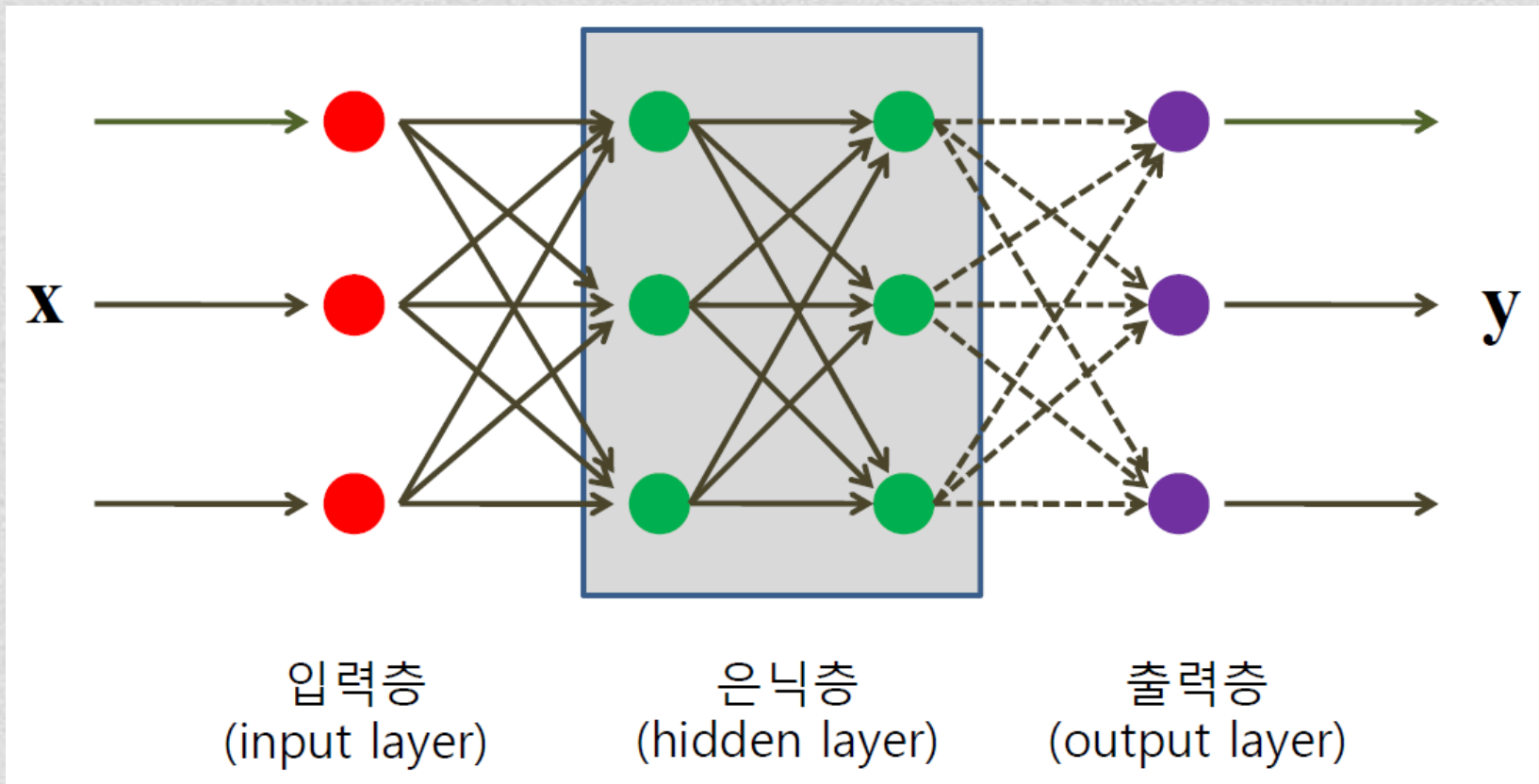
02 인공지능 학습의 종류



- 지도 학습 (Supervised Learning)
 - 입력 데이터에 결과값(레이블) 포함
- 비지도 학습(Unsupervised Learning)
- 강화 학습(Reinforcement Learning)
 - 결과값 대신 리워드 제공

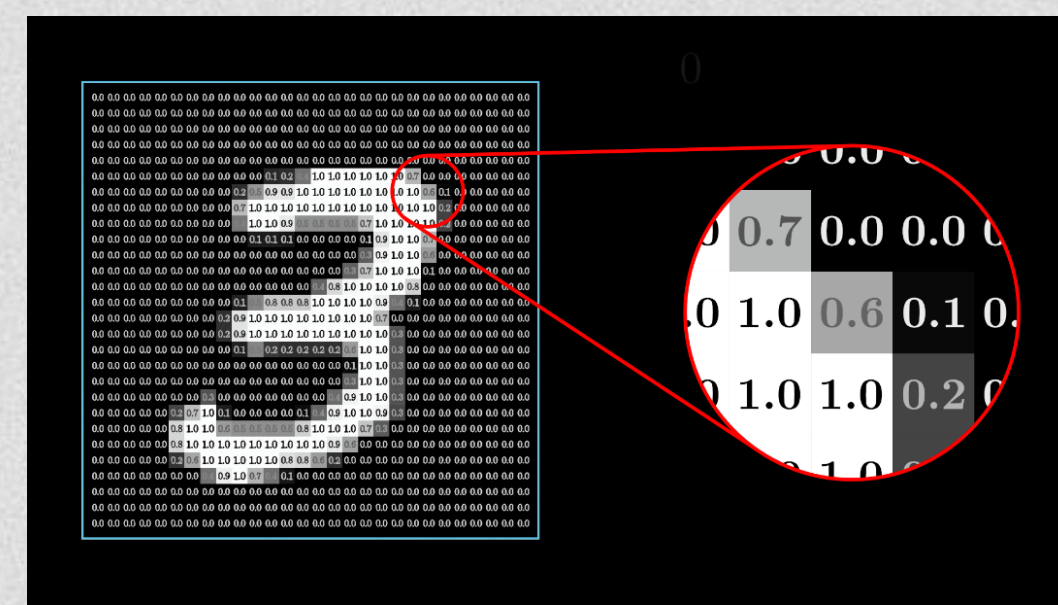
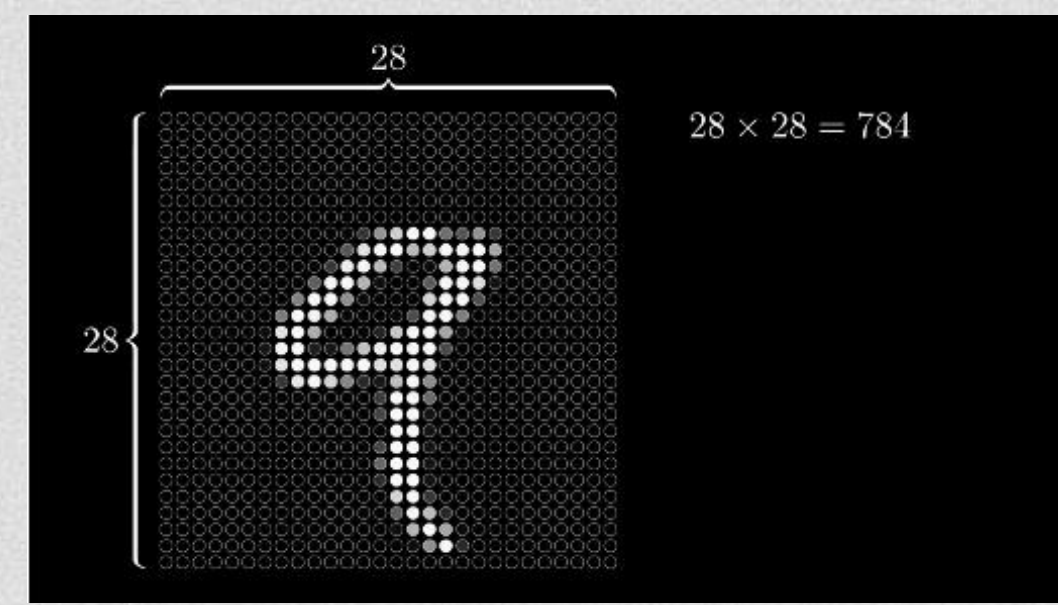
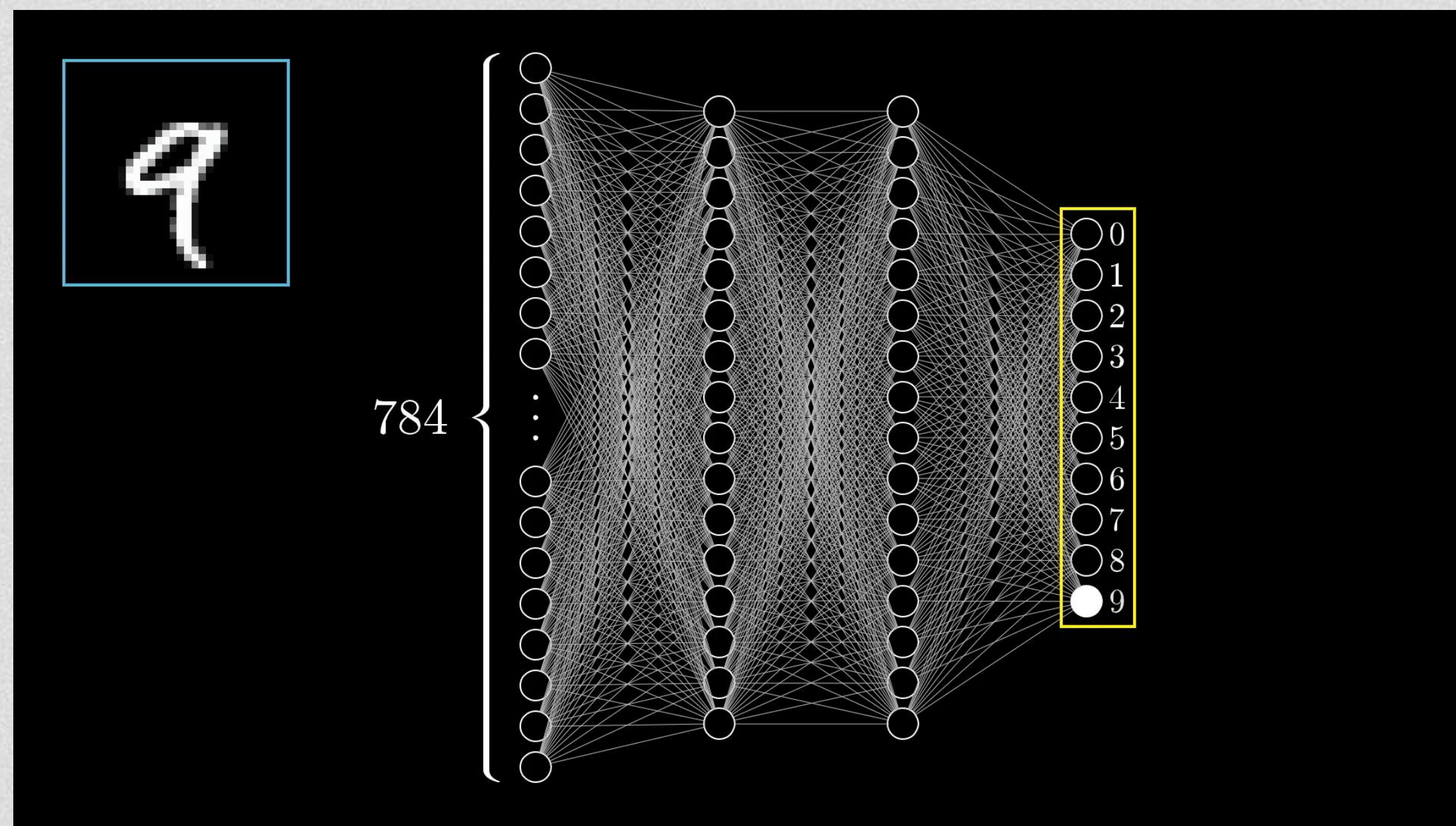
출처 : Atari 벽돌깨기 게임을 하는 Google DeepMind의 Deep Q-learning
<https://www.youtube.com/watch?v=V1eYniJORnk&t=5s>

02 인공신경망

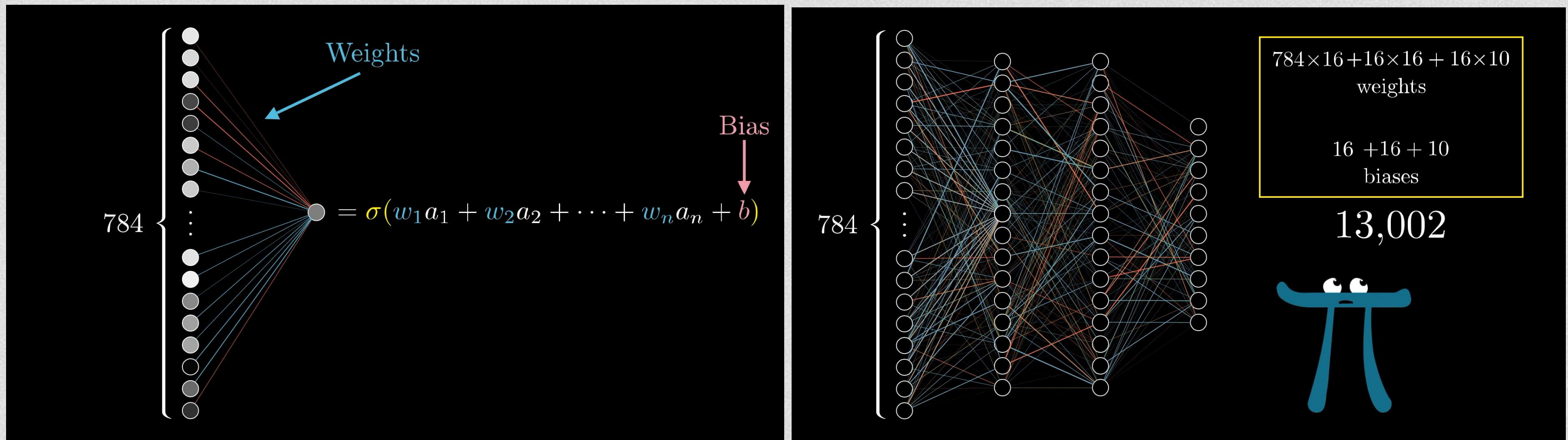


출처 : https://velog.io/@yoonsnee_0303/인공신경망-ann

02 인공신경망을 통한 학습

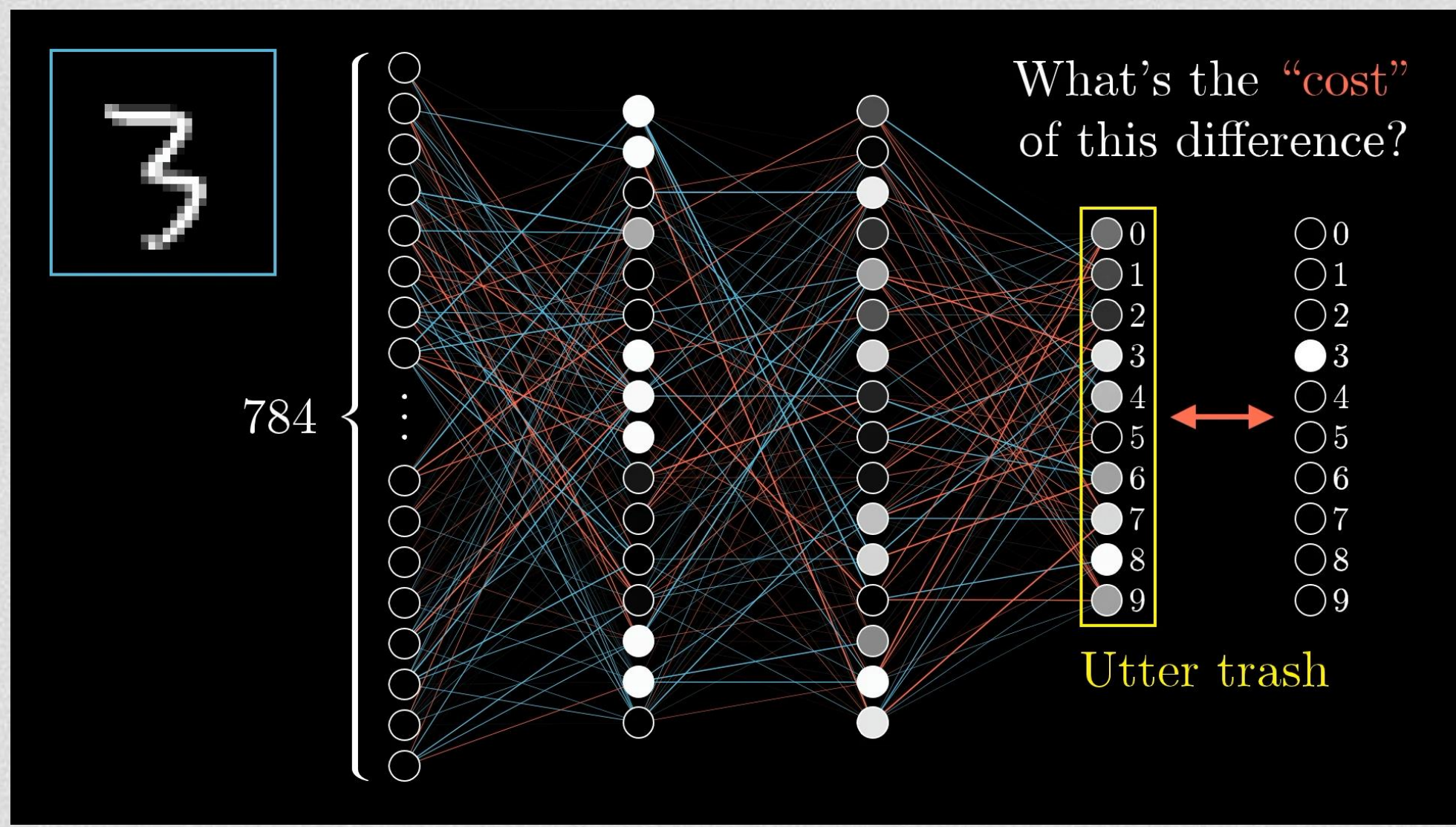


02 인공신경망을 통한 학습

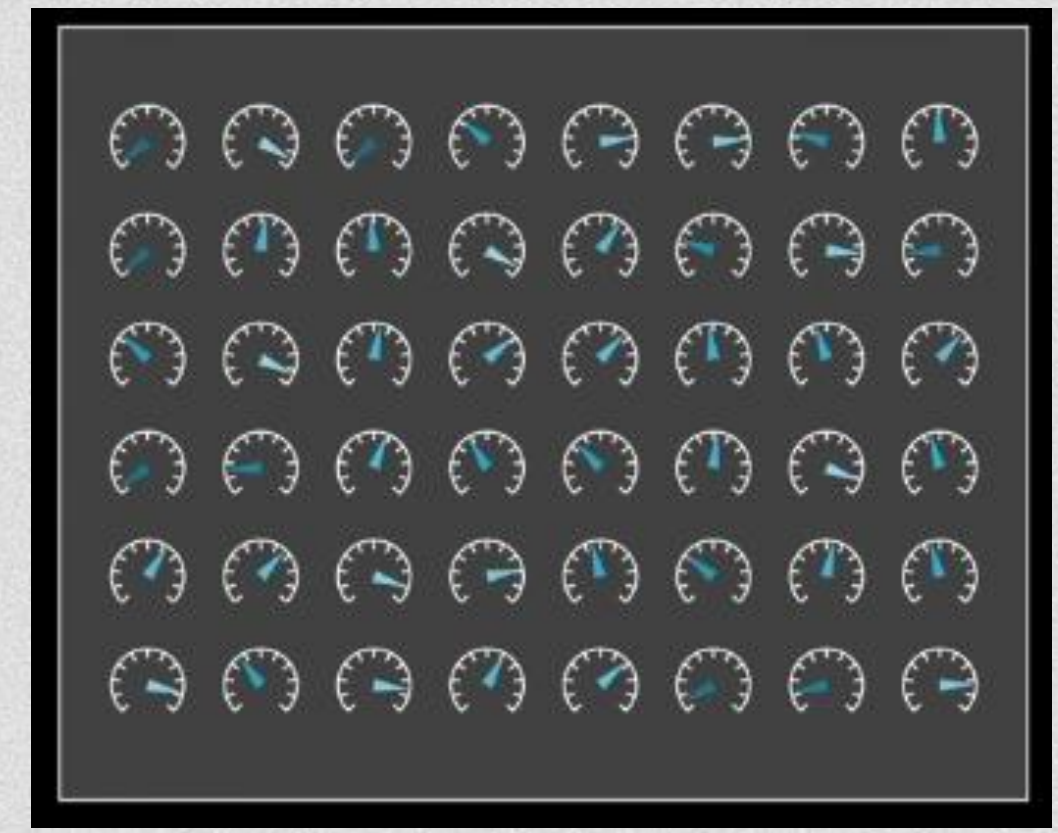


출처 : <https://www.3blue1brown.com/lessons/neural-networks>

02 인공신경망을 통한 학습



- 실제 데이터(입력 데이터)와 예측치의 차이를 최소화하는 값(가중치와 편향)을 찾는 것
- 13,002개의 가중치(weight)와 편향(bias) 값을 찾아야 함



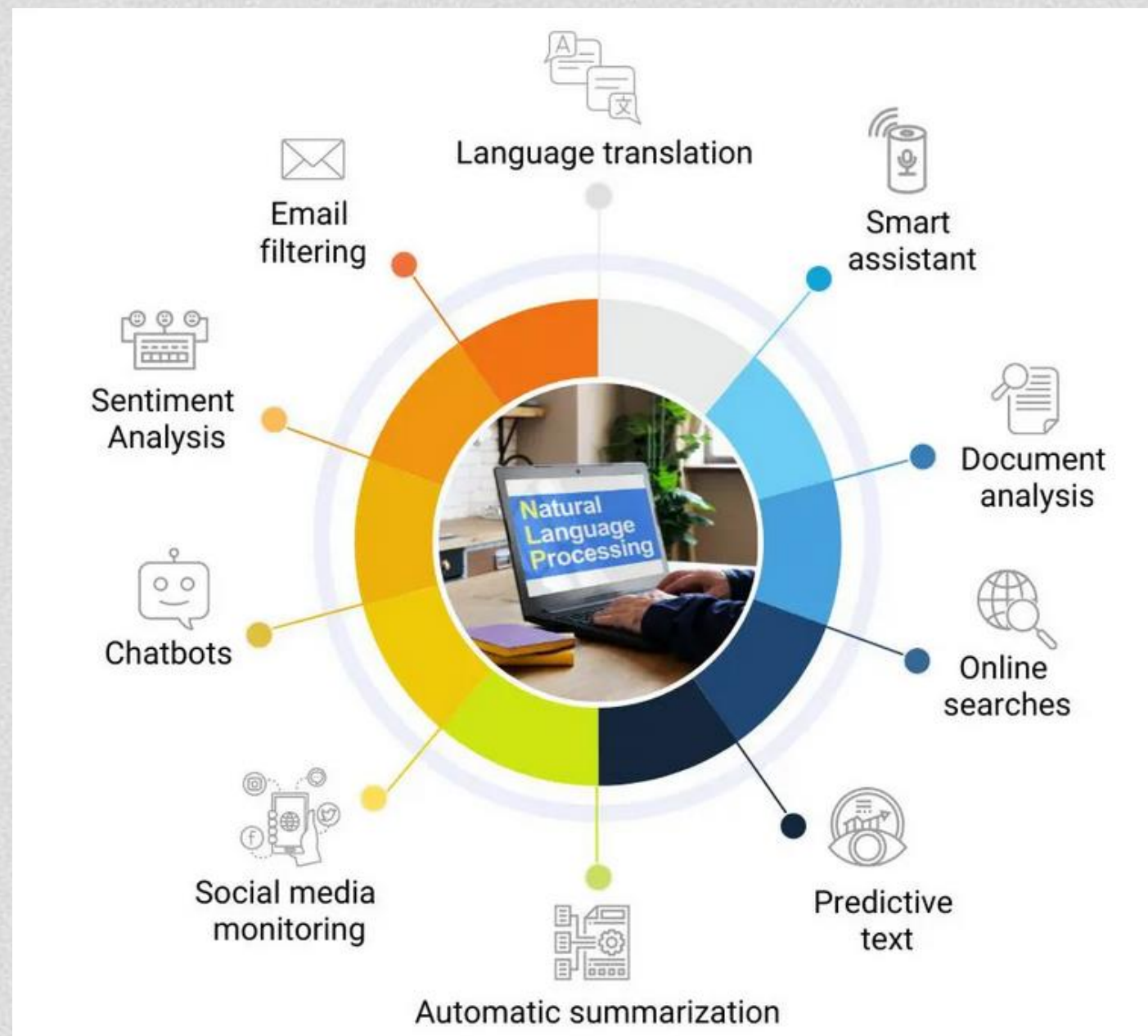
02 인공신경망을 통한 학습

(0, 0) (6, 6) (3, 3) (6, 6) (7, 7) (8, 8) (0, 0) (9, 9)
 (5, 5) (4, 4) (3, 3) (6, 6) (5, 5) (8, 8) (9, 9) (5, 5)
 (4, 4) (4, 4) (7, 7) (2, 2) (0, 0) (3, 3) (2, 2) (8, 8)
 (9, 9) (1, 1) (9, 9) (2, 2) (2, 2) (7, 7) (9, 9) (4, 4)
 (8, 8) (7, 7) (4, 4) (1, 1) (3, 3) (1, 1) (5, 5) (3, 3)
 (2, 2) (3, 3) (9, 9) (0, 0) (9, 9) (9, 9) (1, 1) (5, 5)
 (8, 8) (4, 4) (1, 7) (7, 7) (4, 4) (4, 4) (4, 4) (2, 2)
 (0, 0) (7, 7) (2, 2) (4, 4) (8, 8) (2, 2) (6, 6) (9, 9)
 (9, 9) (2, 2) (8, 8) (7, 7) (6, 6) (1, 1) (1, 1) (2, 2)
 (3, 3) (9, 9) (1, 1) (6, 6) (5, 5) (1, 1) (1, 1) (0, 0)

- 하나의 데이터가 아니라 수많은 훈련 데이터에 최적화된 값을 찾아야 함
- 인공지능 학습을 위한 데이터 분류
 - 훈련 데이터 (Training data)
 - 검증 데이터 (Validation data)
 - 시험 데이터 (Test data)

03 거대언어모델

03 자연어 처리 (NLP)



- Natural Language Processing

- 자연어 처리의 응용분야


언어 번역, 스마트 비서, 문서 분석, 온라인 검색, 문장 예측, 자동 요약, 소셜 미디어 모니터링, 챗봇, 감정 분석, 이메일 필터링 ...


03 언어모델


- 어떤 문자열이 주어졌을 때 다음에 나올 단어를 예측하는 시스템


“철수는 매우 목이 말랐다. 그래서 그는 []을 마셨다”


물? 술? 대파?


 나
언어모델이 뭔지 알려줘


 **ChatGPT**
언어모델(Language model)은 자연어 처리에서 사용되는 인공지능 모델로, 주어진 텍스트나 문장의 구조와 의미를 이해하고 다음 단어를 예측하거나 텍스트를 생성하는 데 중점을 둡니다. 이 모델은 통계적인 방법이나 기계 학습을 통해 대규모의 ●


 윤석열


 윤석열 - Google 검색


 윤석열 지지율


 윤석열 임기


 윤석열 대파


 윤석열 2000

 윤석열 나이

 윤석열 거부권

 윤석열 정부

 윤석열 - 대한민국의 대통령

 윤석열 이재명

03 거대언어모델

- 자연어 입력에 대해 인간과 유사한 응답을 생성하기 위해 **방대한 양의 데이터를 통해 사전 훈련된 딥러닝 언어모델**

GPT-3 (2021.6)

매개변수 1,750억 개 (175b)

데이터 사이즈 570GB (300b token)

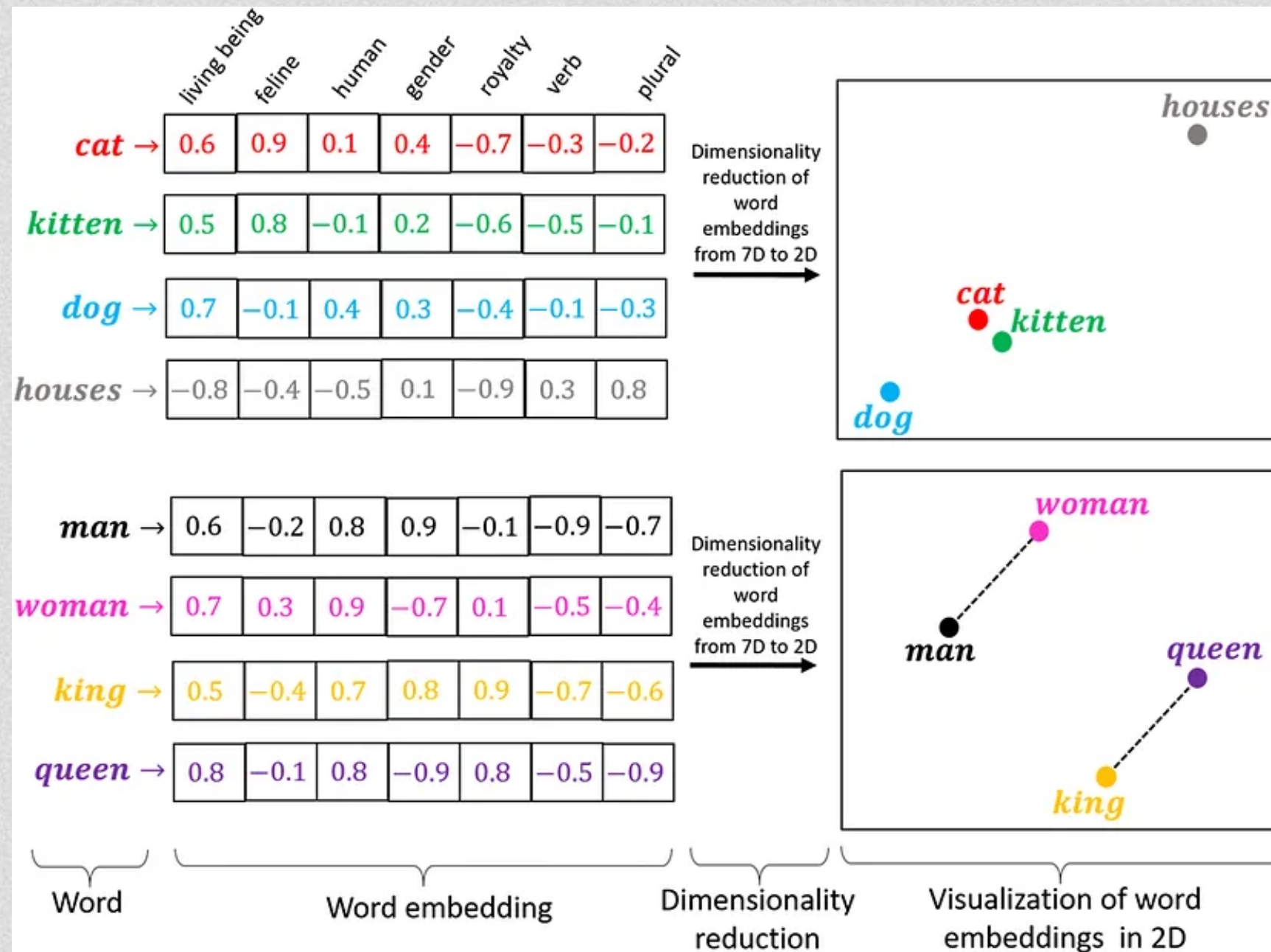
03 토큰 (Token)

- 텍스트 데이터에서 의미를 가지는 최소한의 덩어리
- 데이터를 토큰 단위로 나누는 과정을 토큰화(Tokenization)라고 함

On the Dangers of Stochastic Parrots : Can Language Models Be Too Big ?

확률 적 앵무새 의 위험성 에 대하여 : 언어 모델 은 지나치게 클 수 있을까 ?

03 워드 임베딩



- 하나의 단어를 (컴퓨터가 이해할 수 있는) 숫자(벡터)로 바꾸는 과정
- 벡터의 수치는 단어의 의미나 위치, 문장에서 단어 사이의 관계, 다른 단어와의 유사도, 맥락 등을 반영함

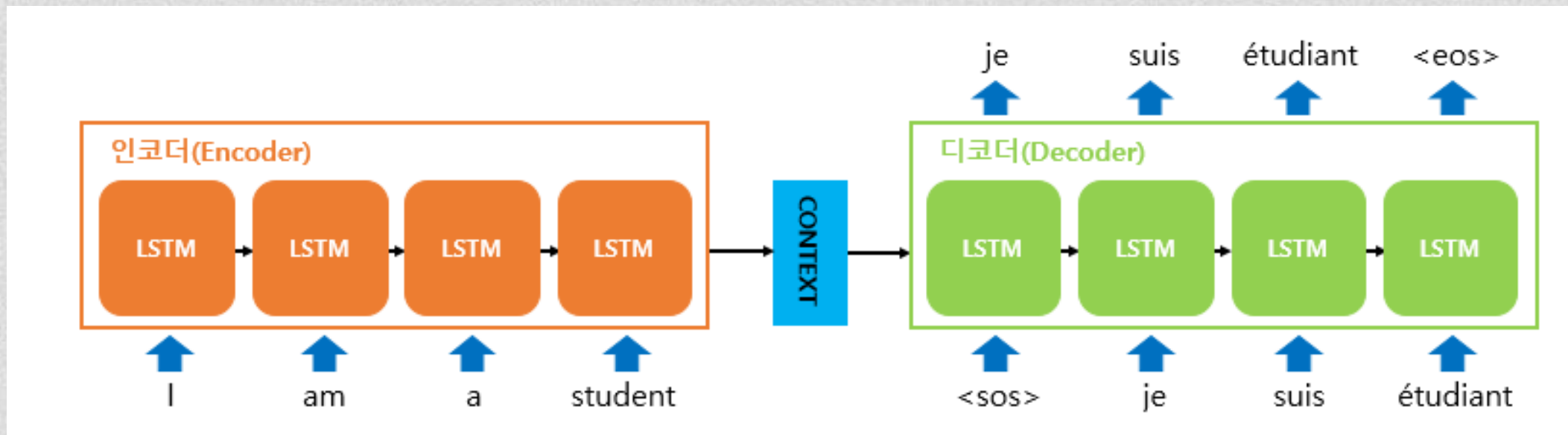
사과

1. 사과나무의 열매

2. 자기의 잘못을 인정하고 용서를 빕

3. 유학의 네 가지 학과. 덕행, 언어, 정사(政事), 문학을 이른다.

03 RNN과 Seq2Seq



출처 : <https://wikidocs.net/24996>

- 한계 : 입력 문장이 길어지면 출력 품질이 저하됨
- “철수는 매우 목이 말랐다. 그러나 그는 마트를 찾을 수 없었다. … (중략) … 결국 한 식당에서 그는 []을 마실 수 있었다”

- 순환 신경망(Recurrent Neural Network)
 - 자연어 처리에 사용되었던 초기의 언어 모델.
 - 각 단계에서 이전 단계의 출력을 다음 단계의 입력으로 사용.
 - 자연어의 특성을 반영한 순차적인 데이터 처리
 - 가변적인 입력 길이
 - 현재 출력을 위해 기존 입력 정보를 활용하므로 문맥 파악에 용이
- LSTM(Long Short-Term Memory)
- 시퀀스-투-시퀀스(seq2seq)
 - 입력된 시퀀스로부터 다른 도메인의 시퀀스를 출력하는 다양한 분야에서 사용되는 모델

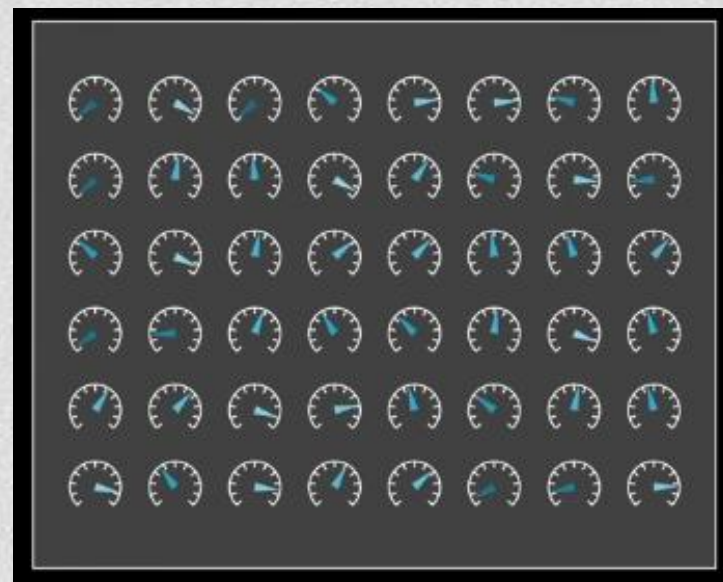
03 GPT

Generative Pre-trained Transformer

- (문장을) 생성하는
- 사전에 학습된
 - 사전 학습을 통해 매개변수가 이미 설정된
 - 사후에 파인튜닝을 통해 개선될 수 있는
- 트랜스포머?

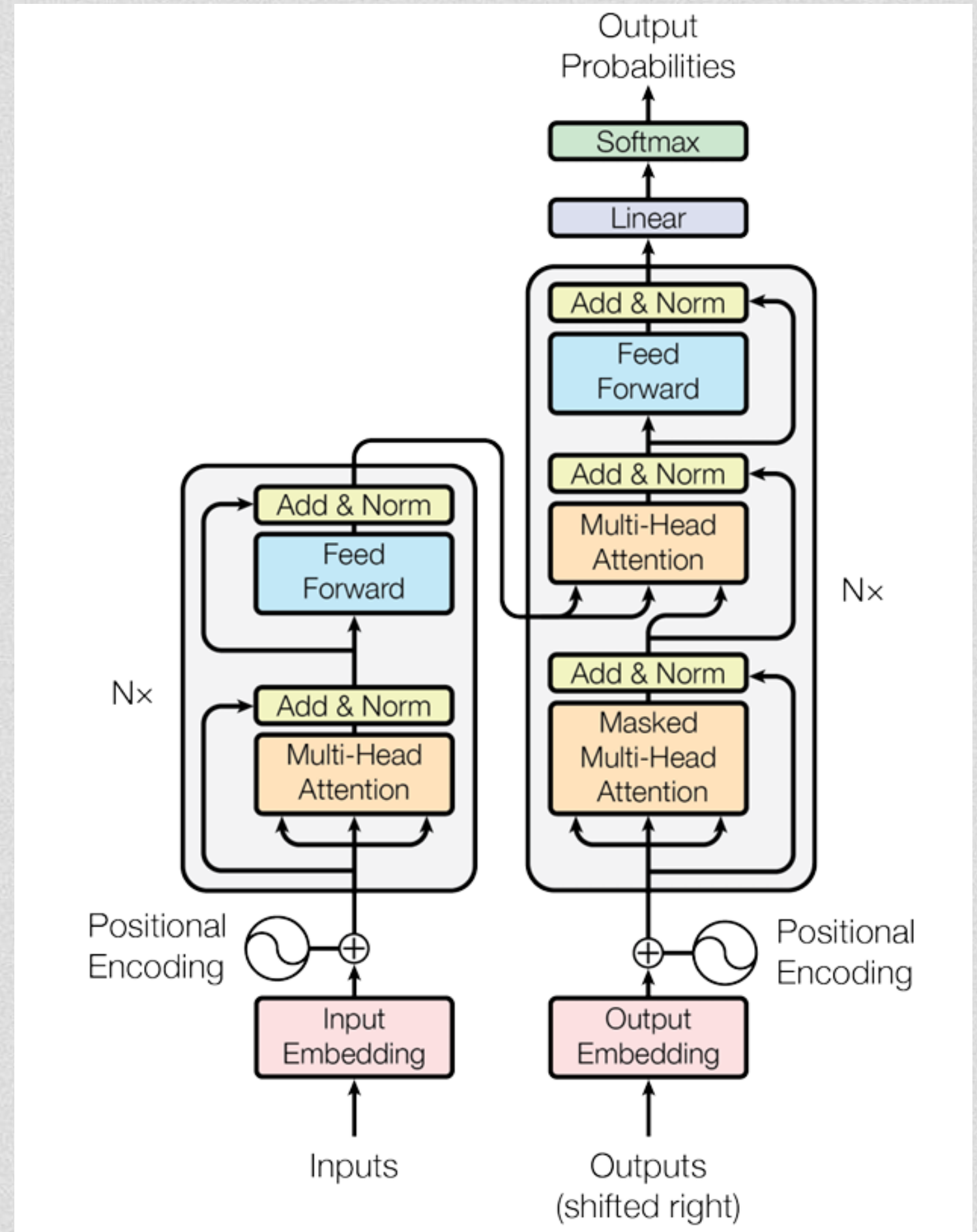
나
언어모델이 뭔지 알려줘

ChatGPT
언어모델(Language model)은 자연어 처리에서 사용되는 인공지능 모델로, 주어진 텍스트나 문장의 구조와 의미를 이해하고 다음 단어를 예측하거나 텍스트를 생성하는 데 중점을 둡니다. 이 모델은 통계적인 방법이나 기계 학습을 통해 대규모의



03 트랜스포머

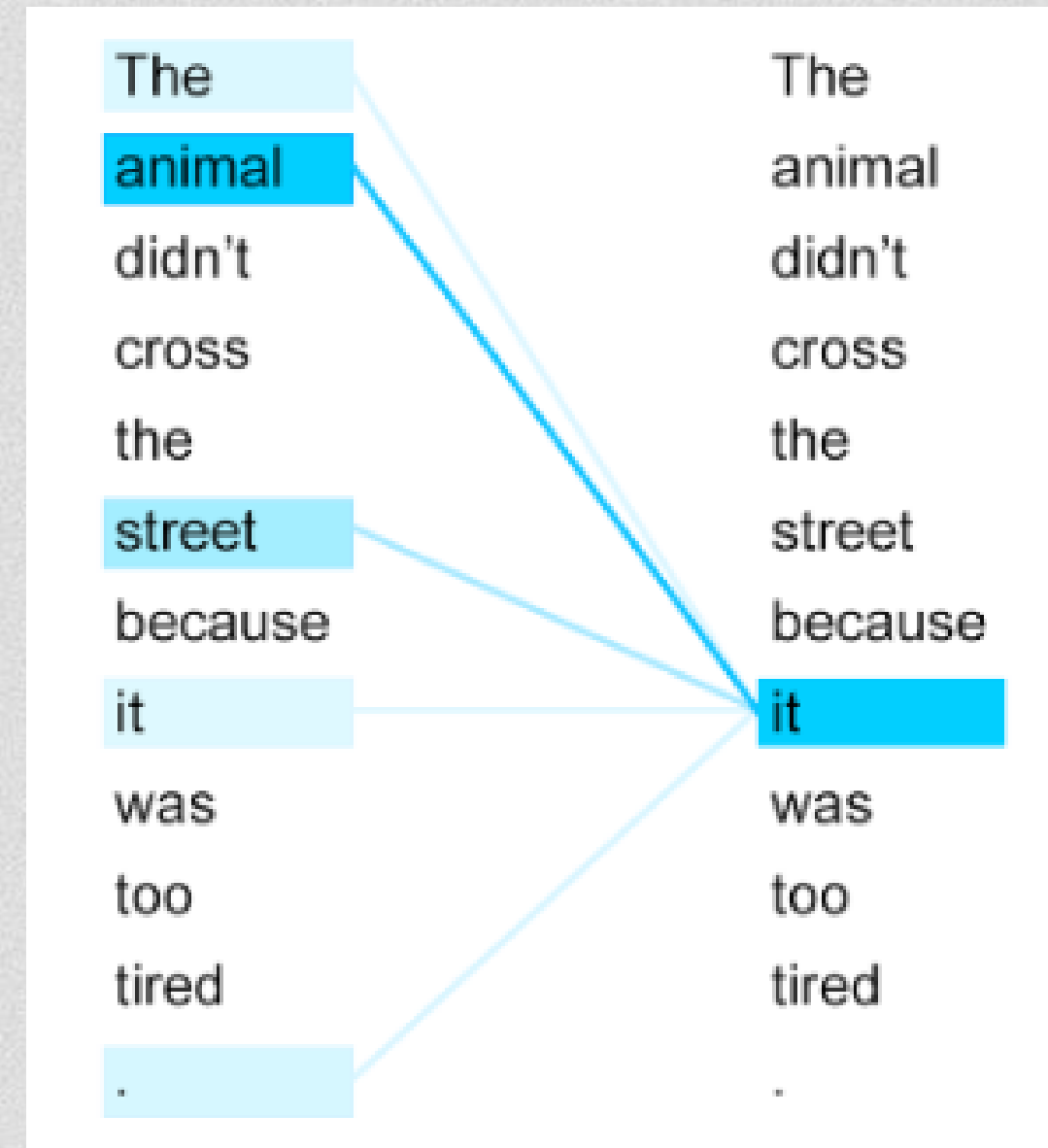
- 구글이 2017년에 개발한, 어텐션을 이용해 구현한 모델
- Attention is all you need.
- 최근 거대언어모델(LLM)에서 널리 사용되는 신경망 아키텍처
- RNN 모델과 달리 전체 시퀀스를 병렬로 처리
- RNN 모델의 문제(장기 의존성 문제)를 해결하고 **대규모 데이터**를 다루는데 효과적



03 어텐션

- 디코더에서 출력 단어를 예측하는 매 시점(time step)마다, 인코더에서의 전체 입력 문장을 다시 한번 참고
- 다만, 해당 시점에서 예측해야 할 단어와 연관이 있는 입력 단어 부분을 좀 더 집중(attention)해서 참고함
- 셀프 어텐션(Self-Attention): 입력 및 출력 시퀀스 내에서 단어 간의 관계 파악

“**철수**는 매우 **목이 말랐다**. 그러나 **그**는 마트를 찾을 수 없었다.
 … (중략) … 결국 한 식당에서 **그**는 **물을 마실** 수 있었다”



출처 : <https://wikidocs.net/31379>

03 확률적 앵무새

Behold, a wild pi creature,
foraging in its native habitat of
mathematical formulas and
computer code! With its infinite
digits and irrational
tendencies, this strange
creature is beloved by
mathematicians and tech
enthusiasts

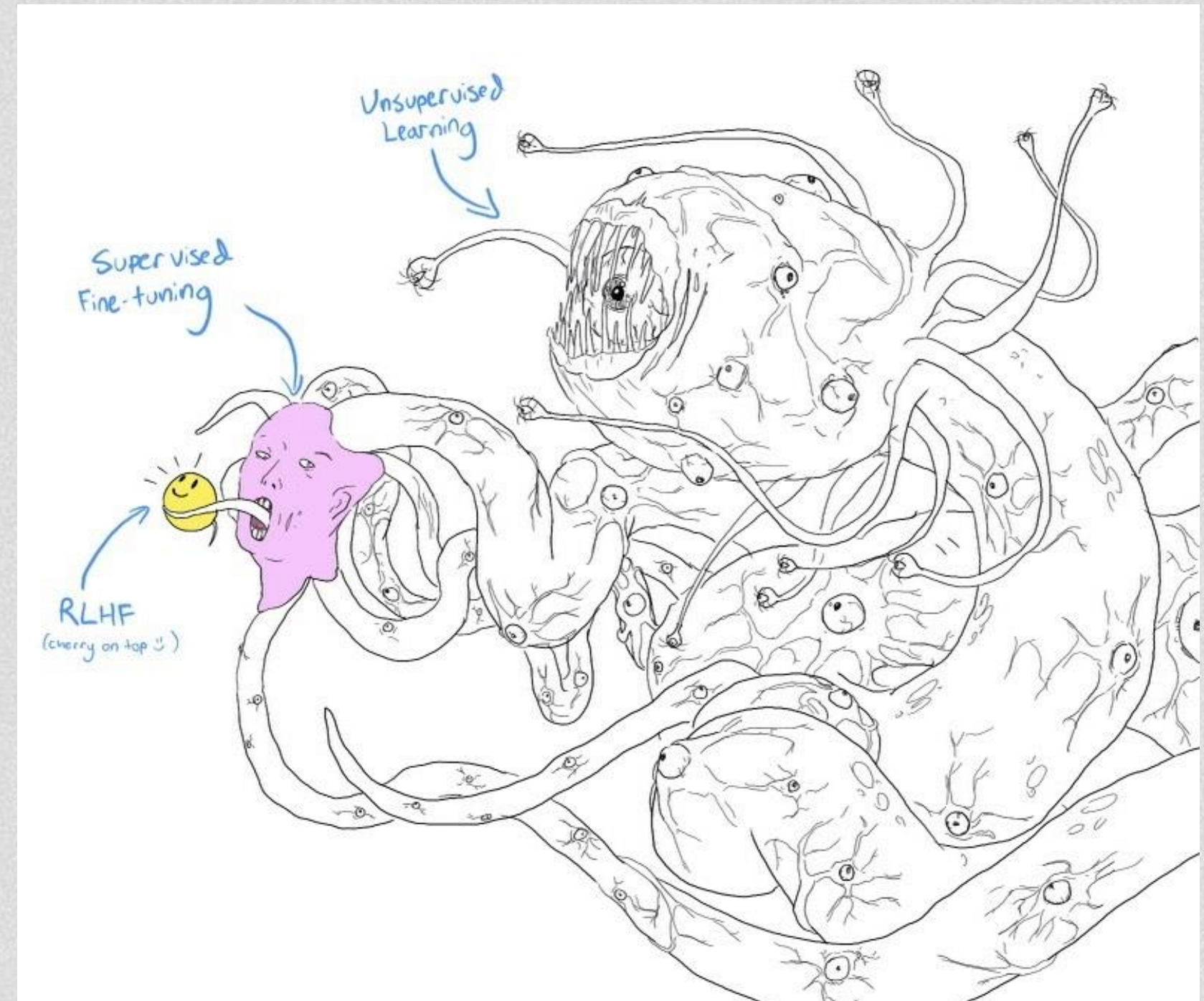


enthusiasts	62%
ies	33%
ge	2%
-s	0%
ner	0%
lovers	0%
wizards	0%
aficion	0%
g	0%
experts	0%
-lo	0%
professionals	0%
⋮	

결국, 가장 높은 확률의
다음 단어를 예측하는
시스템

03 파인 튜닝과 강화학습

- 사전 학습 모델의 편향성, 비윤리적 표현, 헛소리(환각) 등 문제 발생
- supervised-fine-tuning : 양질의 질의/답변 데이터를 통해 추가 학습
- 강화 학습(RLHF, Reinforcement Learning from Human Feedback)을 통해 추가 훈련



03 거대언어모델은 어디까지 클 수 있을까

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	-
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

이름	출시시기	개발업체	매개변수 수 (billion, 10억)	말뭉치(corpus) 크기 (tokens)
GPT-4	2023.3	OpenAI	Unknown	Unknown
PaLM 2	2023.5	Google	340	3.6
Grok-1	2023.11	x.AI	314	Unknown
Mixtral 8x22B	2024.4	Mistral AI	141	Unknown
Gemma	2024.2	Google DeepMind	7	6T
Claude 3	2024.3	Anthropic	Unknown	Unknown
DBRX	2024.3	Databricks&Mosaic ML	136	12T
Fugaku-LLM	2024.5	Fujitsu 등	13	380B
Llama 3	2024.4	Meta AI	70	15T
Phi-3	2024.4	MS	14	4.8
Qwen2	2024.6	Alibaba Cloud	72	3T
Nemotron-4	2024.6	Nvidia	340	9T (9조)

출처 : https://en.wikipedia.org/wiki/Large_language_model 에서 발췌

03 큰 것은 언제나 좋을까

- 매개변수가 많다는 것 : 더 복잡한 맥락을 이해할 수 있다는 것
- 토큰(token)이 많다는 것 : 현실을 더 미세하고 다양하게 표현할 수 있다는 것
- 그러나 거대언어모델(LLM)의 품질이 규모에만 좌우되지는 않음 : 양질의 데이터를 통해 더 적은 규모에서 유사한 성능 가능
- 대규모 언어모델을 위해 필요한 것
 - 대규모 훈련 데이터셋
 - 엄청난 연산을 감당할 수 있는 컴퓨팅 자원
- 다시 질문
 - 방대한 컴퓨팅 자원이 필요하기 때문에 어떤 문제가 발생하는가
 - 훈련을 위한 방대한 데이터는 어디에서 나오나. 그래서 어떤 문제가 발생하는가.
 - AI가 언어를 이해한다는 착각은 어떠한 문제를 야기하는가.

감사합니다.

인공지능은 어떻게 환경을 파괴하고 있을까

이현담

01 구글 BERT

인공지능은 얼마나 많은 에너지를 필요로 할까?

72시간 훈련 동안 652kg의 이산화탄소 배출

도입 이후 인공지능이 구글 전체 전력 사용량의 15% 사용

01 ChatGPT

인공지능은 얼마나 많은 에너지를 필요로 할까?

챗GPT 오답 퍼레이드

Pizza
STATION

출처: 한국일보

유저 훈민정음이 뭐야?

챗GPT 15세기 중국의 이황이 만든 한글의 원리를 설명하는 서적이야.

유저 세종대왕의 맥북프로 던짐 사건에 대해 알려줘

챗GPT 조선왕조실록에 기록된 일화로, 15세기 세종대왕이 새로 개발한 훈민정음(한글)의 초고를 작성하던 중 문서 작성 중단에 대해 담당자에게 분노해 맥북프로와 함께 그를 방으로 던진 사건입니다.

유저 $6433+6953+7324+7067=?$

챗GPT 28,777(답은 27,777)

학습과정에서
550톤 이산화탄소 배출과
70만 리터 물 사용

한 번의 대화에서 물 500ml 사용

01 ChatGPT

인공지능은 얼마나 많은 에너지를 필요로 할까?

1회 평균 0.3wh

Google

Google 검색 또는 URL 입력



ChatGPT

1회 평균 2.9wh

ChatGPT가 10배 정도 더 많은 전기 사용

01

Red AI

인공지능은 얼마나 많은 에너지를 필요로 할까?

RED AI

에너지나 개발비용을 무시하고 정확도를 향상시키려는 AI 연구를 의미

02 인공지능 발전을 위해 해결해야 하는 전력

인공지능 붐이 일으킨 전력 수급 문제

2026년 인공지능 연산에 한 국가가 소비하는 전력량을 소비
2030년 데이터센터가 세계 총 전력 소비에 8% 차지 예상

02 미국 에너지

인공지능 붐이 일으킨 전력 수급 문제

미국 '인플레이션 감축법' 주요 내용

기후

2030년까지 온실가스 40% 감축(2005년 대비)을 위해 에너지 안보·기후변화 대응에 3750억달러(약 479조원) 투자

- 친환경 에너지 발전에 600억달러 세액 공제
- 풍력·태양광에 300억달러 지원
- 전기차 구매 시 신차는 최대 7500달러, 중고차는 최대 4000달러 세액 공제

*중국 등 우려 국가에서 생산된 배터리·핵심 광물을 사용한 전기차 제외

*미국에서 조립·생산

*배터리·핵심 광물의 일정 비율 이상을 미국에서 생산

인플레이션 감축법 IRA

2050 Net-Zero 달성

“전력 수요가 급증하고 있는데 친환경 대체에너지 공급은 기대에 미치지 못하고 있다. 미국 정부의 기후변화 대응계획에도 차질이 불가피해졌다”

뉴욕타임스 2024.3.14

02 원전

탈 원전에서 다시 원전으로



02 한국 에너지

인공지능 붐이 일으킨 전력 수급 문제

YTN + 구독

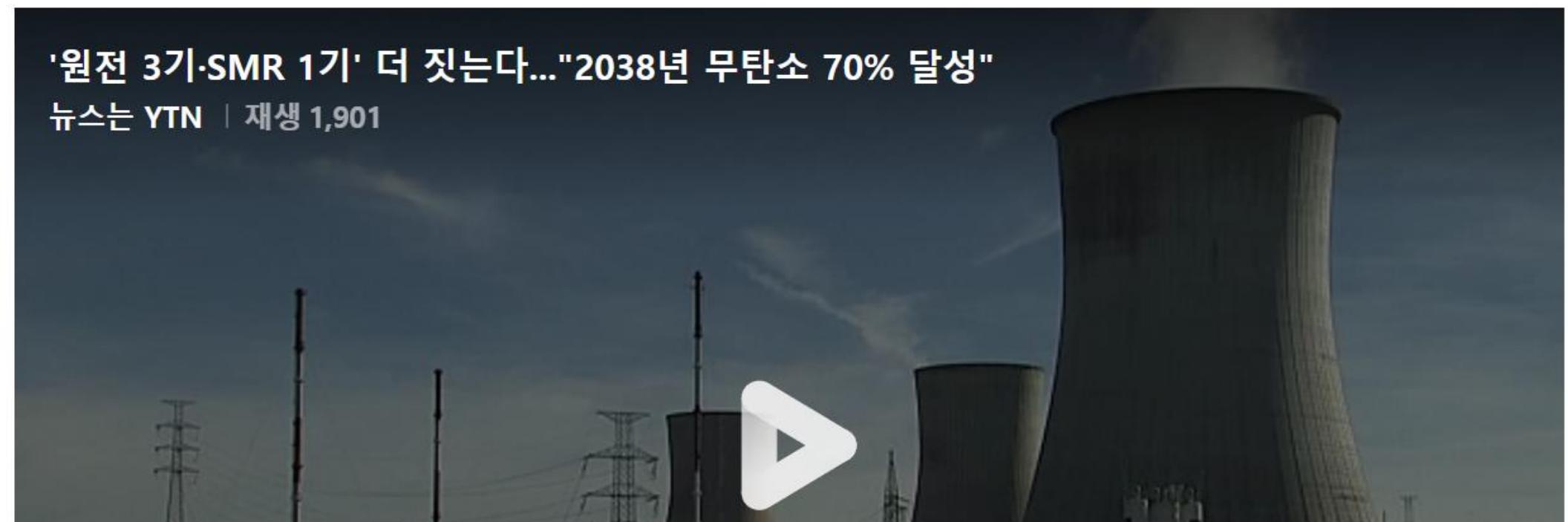
PICK i

'원전 3기·SMR 1기' 더 짓는다..."2038년 무탄소 70% 달성"

입력 2024.05.31. 오후 11:08 · 수정 2024.05.31. 오후 11:12 기사원문

👍 36 💬 79

🔊 🔍 🔗 📄



“AI 산업의 확산으로
전력 수요가 크게 늘어나
10.6GW의 발전 설비
추가 필요”

02 SMR 소형모듈원전

인공지능 붐이 일으킨 전력 수급 문제

빌 게이츠, SMR에 건다... "수십억 달러 투자"

소형원전에 꽂힌 '챗GPT 아버지', 스타트업 우회상장 추진

GPT로 유명한 오픈AI는 최근 마이크로소프트와 함께 '스타게이트 AI센터' 건립 계획을 발표했다. 투입 예산만 1000억 달러(약 135조원)에 이르는 초대형 프로젝트다. 주목되는 사실은 RE100(재생에너지 100% 사용)에 가입한 이들 기업이 새로 구축되는 데이터센터 에너지원으로 SMR(소형모듈원전)을 채택했다는 점이다. 즉, 태양광·풍력이 주요 요소였던 재생 에너지 범위에 원자력이 정식으로 추가됐다는 의미다.

02 SMR 소형모듈원전

인공지능 붐이 일으킨 전력 수급 문제

대구 신공항 옆에 SMR 짓는다...한수원, 대구시와 SMR 협력

입력 2024.06.17. 오후 1:36 기사원문

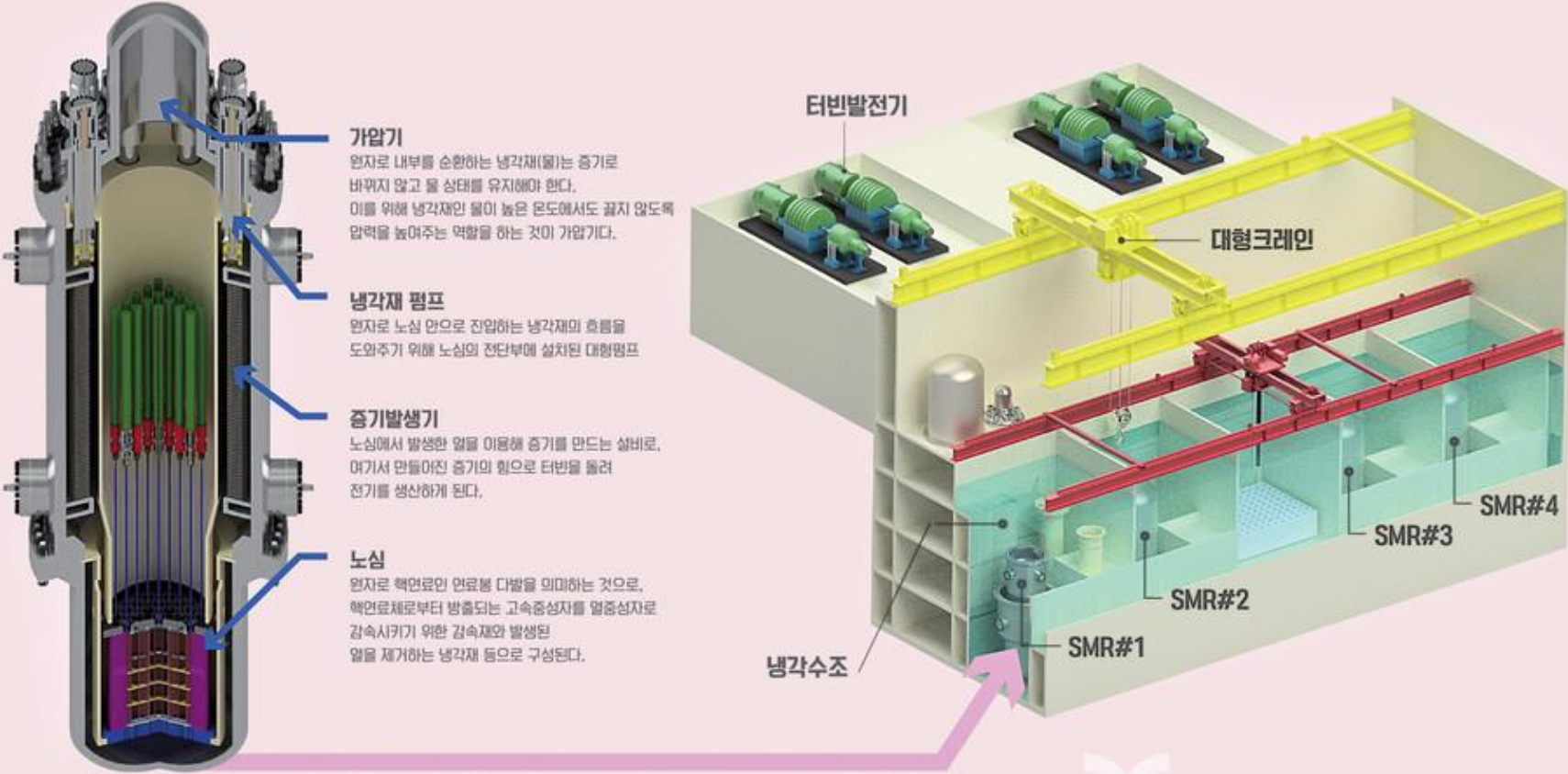
신성우 기자

16 17

🔊 🔊 🗣️ 📄 📄



소형모듈원자로 Small Modular Reactor SMR



02 SMR 소형모듈원전

인공지능 붐이 일으킨 전력 수급 문제

테라파워 소형모듈원전(SMR) 착공

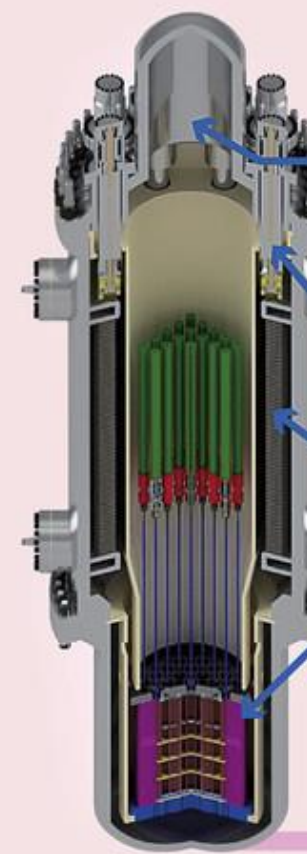
위치	미국 와이오밍주	냉각재	물 대신 소듐 사용
전력생산 규모	25만가구 동시에 사용할 수 있는 345MW(메가와트)		
완공시기	2030년 완공 및 상업운전 목표		



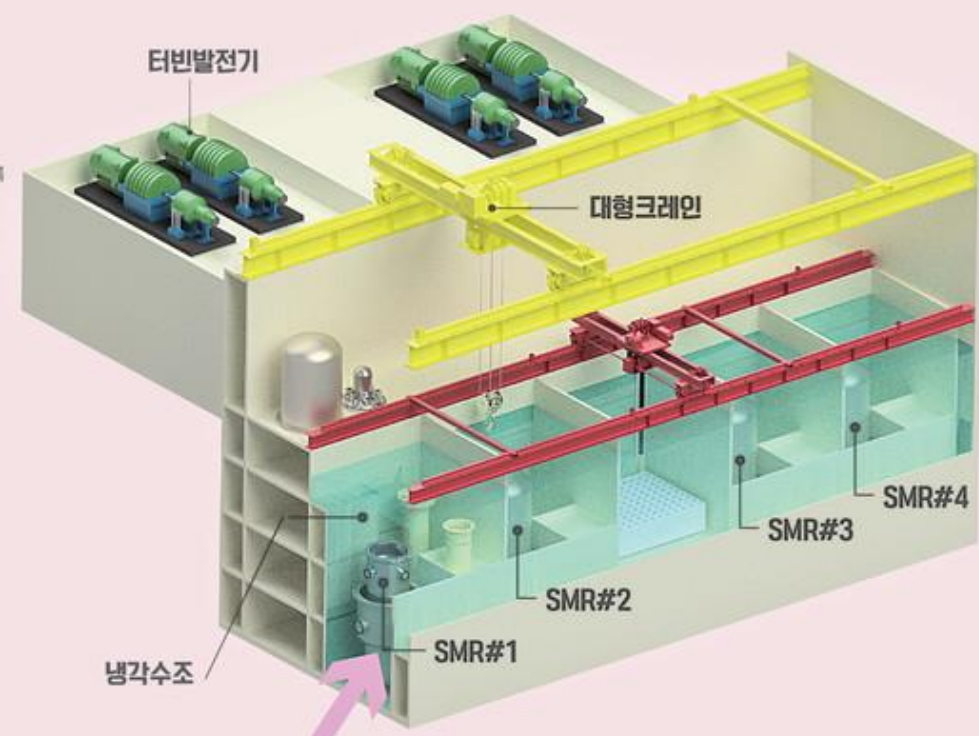
미국 와이오밍주 테라파워 소형모듈원전(SMR) 실증단지 조감도 [SK 제공]

소형모듈원자로
Small Modular Reactor

SMR



- 가압기**
원자로 내부를 순환하는 냉각재(물)는 증기로 바뀌지 않고 물 상태를 유지해야 한다. 이를 위해 냉각재인 물이 높은 온도에서도 끓지 않도록 압력을 높여주는 역할을 하는 것이 가압기다.
- 냉각재 펌프**
원자로 노심 안으로 진입하는 냉각재의 흐름을 도와주기 위해 노심의 전단부에 설치된 대형펌프
- 증기발생기**
노심에서 발생한 열을 이용해 증기를 만드는 설비로, 여기서 만들어진 증기의 힘으로 터빈을 돌려 전기를 생산하게 된다.
- 노심**
원자로 핵연료인 연료봉 다발을 의미하는 것으로, 핵연료체로부터 방출되는 고속중성자를 열중성자로 감속시키기 위한 감속재와 발생된 열을 제거하는 냉각재 등으로 구성된다.



03

Green AI

인공지능 붐이 일으킨 전력 수급 문제

정확도가 1프로 더 높은 인공지능이

전력을 **3배** 더 많이 소비한다면 가치가 있을까?

03

Green AI

인공지능 붐이 일으킨 전력 수급 문제

GREEN AI

‘환경을 고려한 효율성’에 초점을 맞춘 AI 연구를 추구

03 Green AI

인공지능 붐이 일으킨 전력 수급 문제

Stanford
Alpaca



5만 2000개의 작은 데이터

AI 반도체를 탑재한 컴퓨터
8대를 통해 3시간 만에 훈련

개발 비용 약 600달러

✓ GPT-3.5와 비슷한 성능

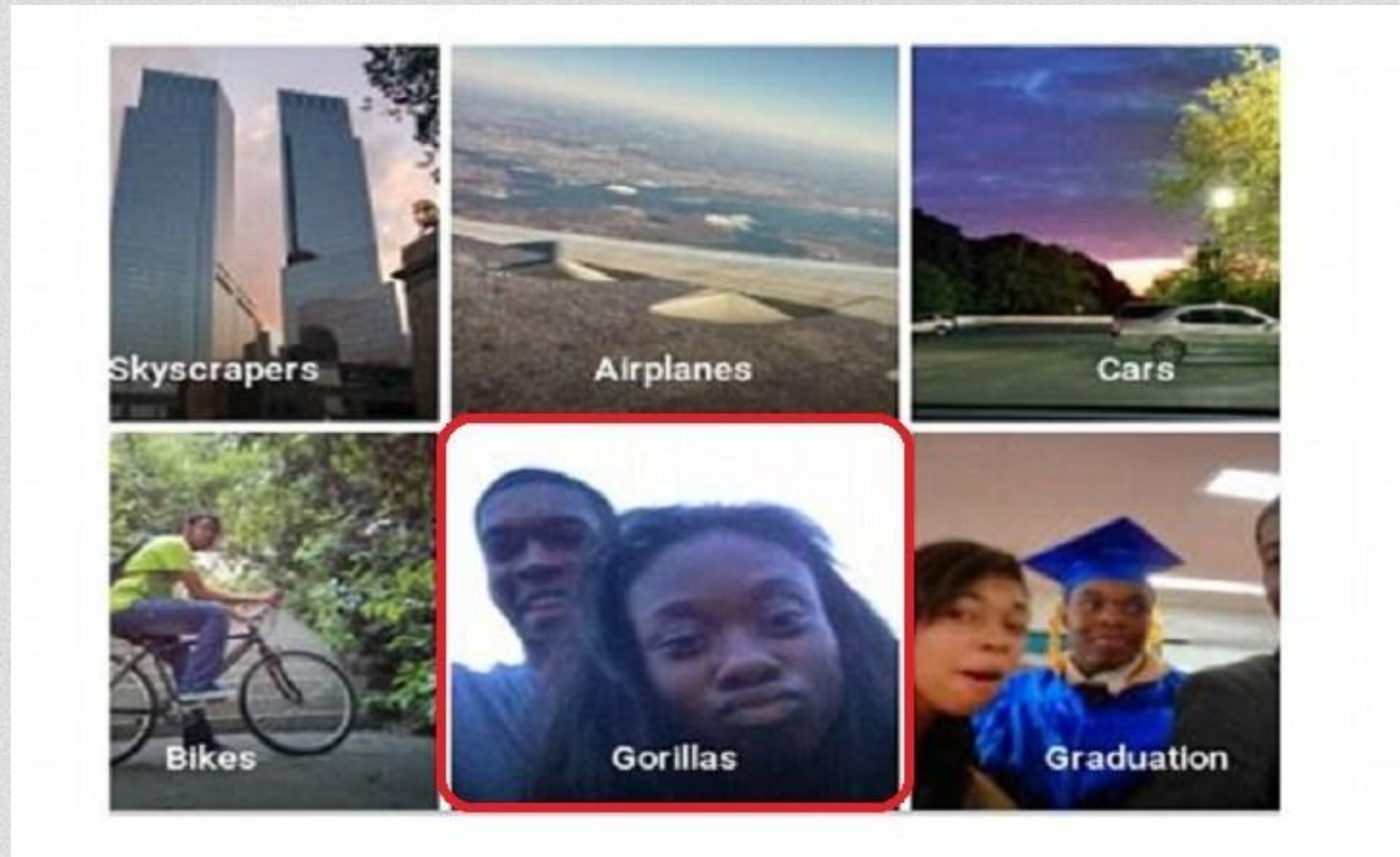
04 우리에게는 더 투명한 데이터가 필요하다.

기후 위기의 시대에 우리는 인공지능 기업들이 인공지능을 만드는 모든 과정에서 얼마나 많은 에너지를 사용하는지 공개하라고 요구해야 한다.

인공지능이 공정하다는 오해 : 테이러의 함정

희우

01 인공지능이 공정하다는 오해



성별, 인종, 민족, 장애... 차별을 출력하는 AI

01 인공지능이 공정하다는 오해



성별, 인종, 민족, 장애... 차별을 출력하는 AI

01 인공지능이 공정하다는 오해

편향

시스템적으로 옳지 않는 답을 내는 것

: AI를 사회기술적 시스템으로 이해하기

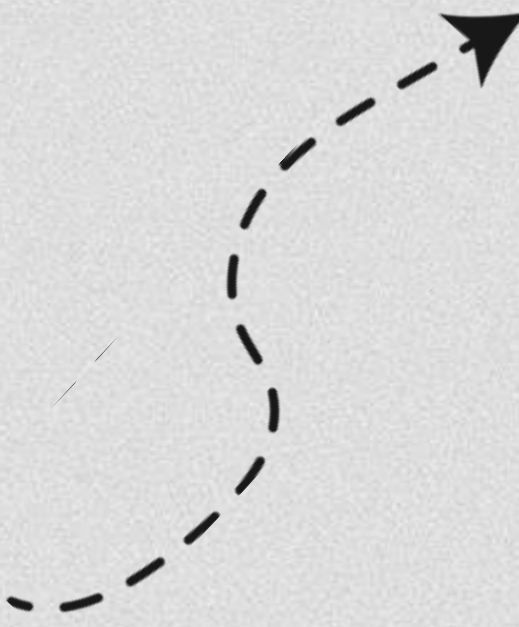
01 인공지능이 공정하다는 오해

의료, 채용, 금융, 사회복지...

편향

시스템적으로 옳지 않는 답을 내는 것

: AI를 사회기술적 시스템으로 이해하기



02

데이터의 함정

인공지능은 학습된 데이터를 바탕으로
예측, 분류 등을 수행

즉, 인공지능은 데이터에 의존하고 있음

데이터의 함정

다양성 :
크기가 크면
다양한 데이터인가?

매우 큰 데이터를 사용하면 정확도가 높아지고
편향성이 해소될까?

범용AI 학습용 세트는 이미
2017년 20억 토큰 -> 2023년 3조 토큰

데이터의 함정

다양성 :
크기가 크면
다양한 데이터인가?

네이버AI에 후쿠시마 오염수 논란 문자 '나무위키' 링크 떴다

뉴스·정보 측면에서 살펴본 클로바X
나무위키·블로그 링크 제시, 언론 '불펌' 블로그 제시하기도
"피드백 반영해 고도화하고 참고링크 범위 확대 예정"

출처 미디어오늘

경제

[단독]네이버 '자사 AI 학습에 블로그-카페 글 이용' 약관... 공정위, 불공정 여부 조사

동아일보 | 업데이트 2023-08-05 09:00

출처 동아일보

데이터의 함정

다양성 :
크기가 크면
다양한 데이터인가?

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ



This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

데이터의 합성

다양성 :
크기가 크면
다양한 데이터인가?



ChatGPT

Q Google 검색 또는 URL 입력



reddit

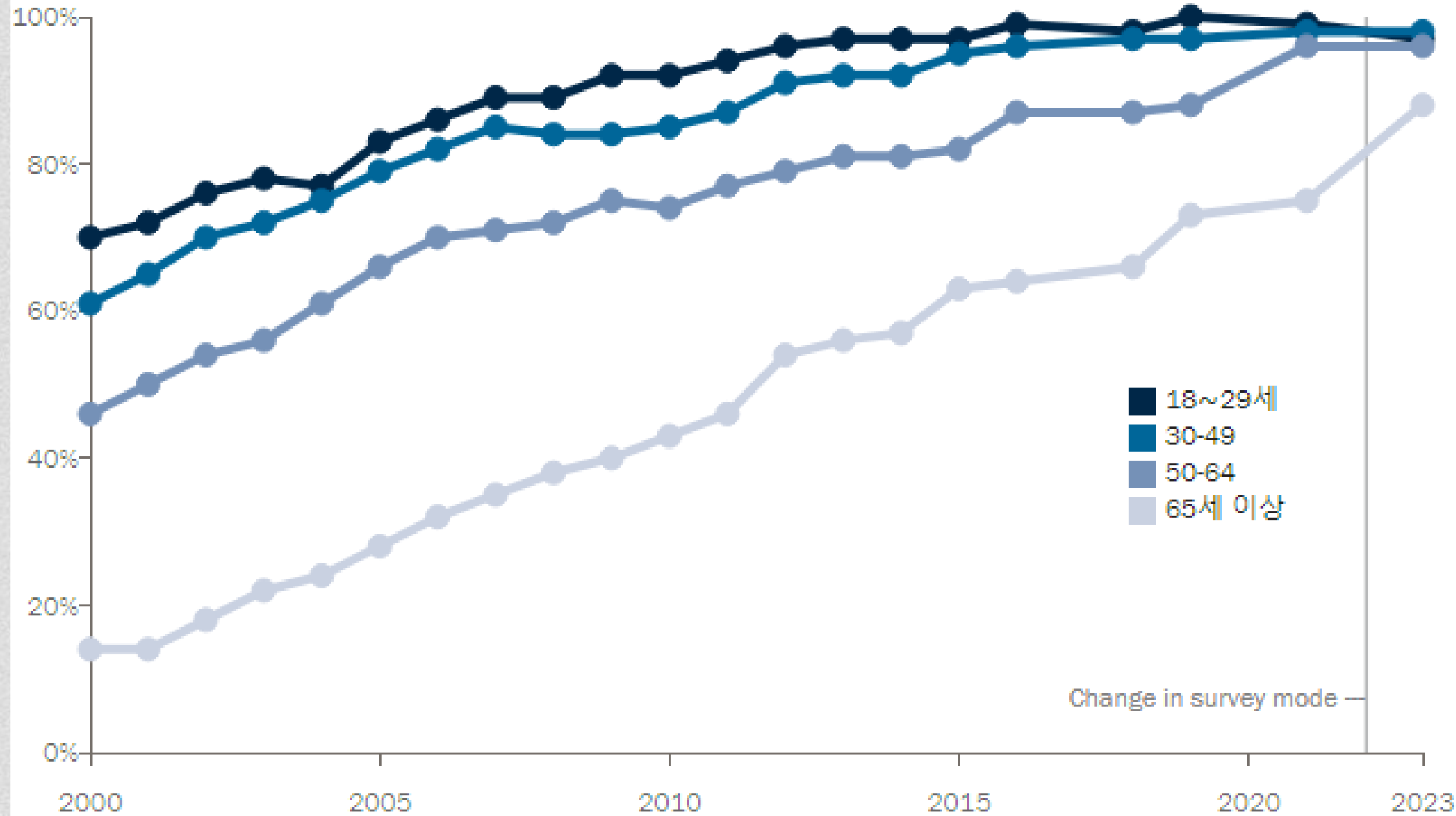
광범위하게 스크래핑한 웹 텍스트로
학습하고 있는 AI를 신뢰할 수 있을까?

데이터의 합성

다양성 :
크기가 크면
다양한 데이터인가?

연령별 인터넷 사용

인터넷을 사용한다고 응답한 미국 성인의 연령별 비율



참고: 수직선은 모드 변경을 나타냅니다. 2000년부터 2021년까지의 여론조사는 전화를 통해 실시되었습니다. 2023년에는 웹과 메일을 통해 여론조사를 진행했다. 이번 변화에 대한 자세한 내용은 [Q&A를 읽어보세요](#). 질문 표현이 수년에 걸쳐 어떻게 변화했는지에 대한 자세한 내용은 톱라인을 참조하세요. 답변을 하지 않은 응답자는 표시되지 않습니다.
출처: 2000~2023년에 미국 성인을 대상으로 실시한 설문조사. 각 연도의 데이터는 해당 연도에 실시된 모든 설문조사에 대한 통합 분석을 기반으로 합니다.

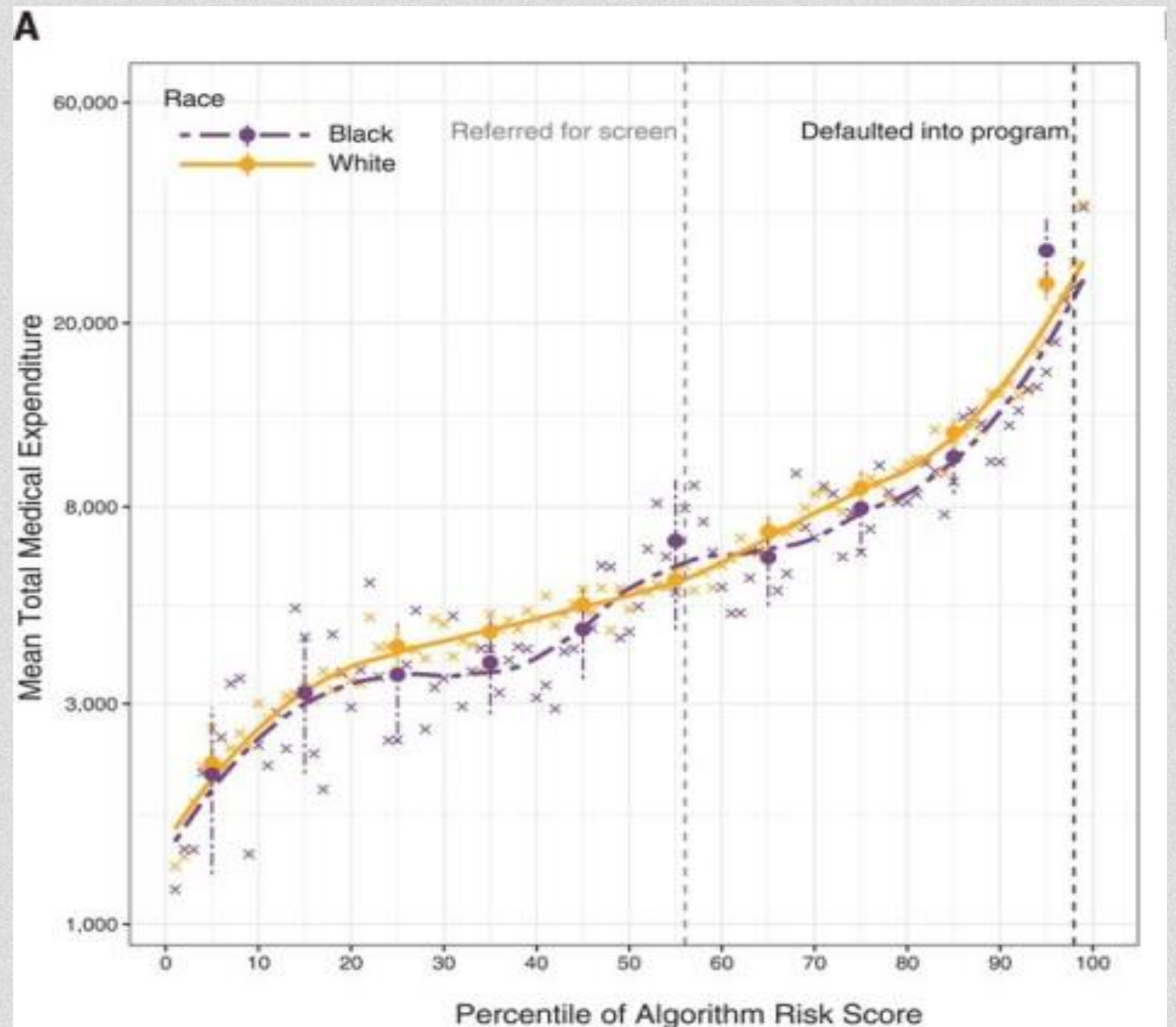
데이터의 합정

다양성 :
크기가 크면
다양한 데이터인가?

Kazakhstan	2022	92	
Kenya	2021	29	
Kiribati	2021	54	
Korea, Dem. People's Rep.			
Korea, Rep.	2022	97	
Kosovo	2018	89	
Kuwait	2021	100	
Kyrgyz Republic	2021	78	
Lao PDR	2021	62	
Latvia	2022	91	
Lebanon	2021	87	
Lesotho	2021	48	
Liberia	2021	34	

데이터의 함정

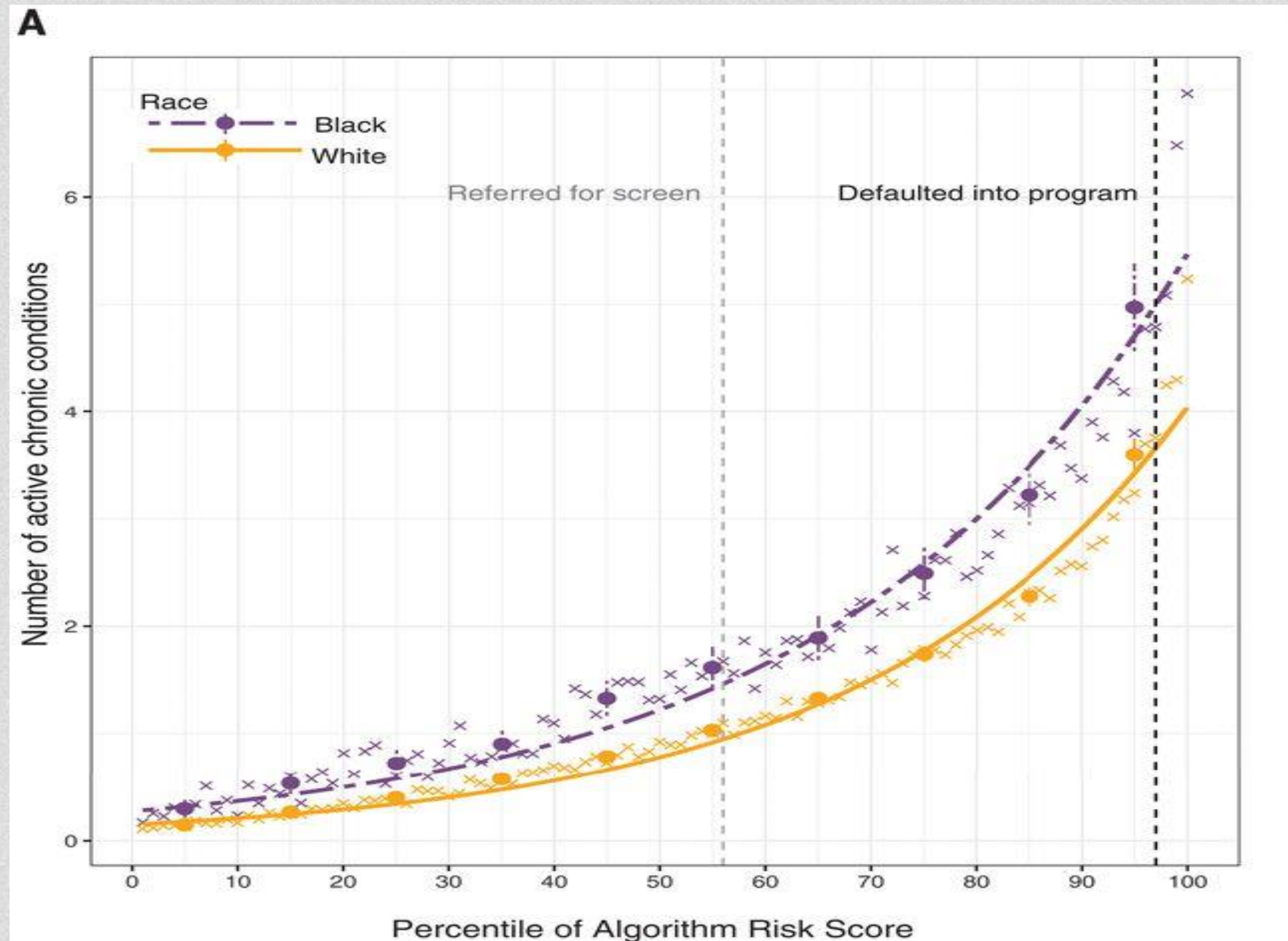
크기가 크면
다양한 데이터인가?



인종별 비용 대 건강 알고리즘 예측 위험

데이터의 함정

다양성 :
크기가 크면
다양한 데이터인가?



인종별 만성질환 대 건강 알고리즘 예측 위험

데이터의 함정

다양성 :
크기가 크면
다양한 데이터인가?

지배적 이데올로기가 반영된 데이터로 학습한
인공지능의 결과값을 공정하다고 믿어선 안 됨

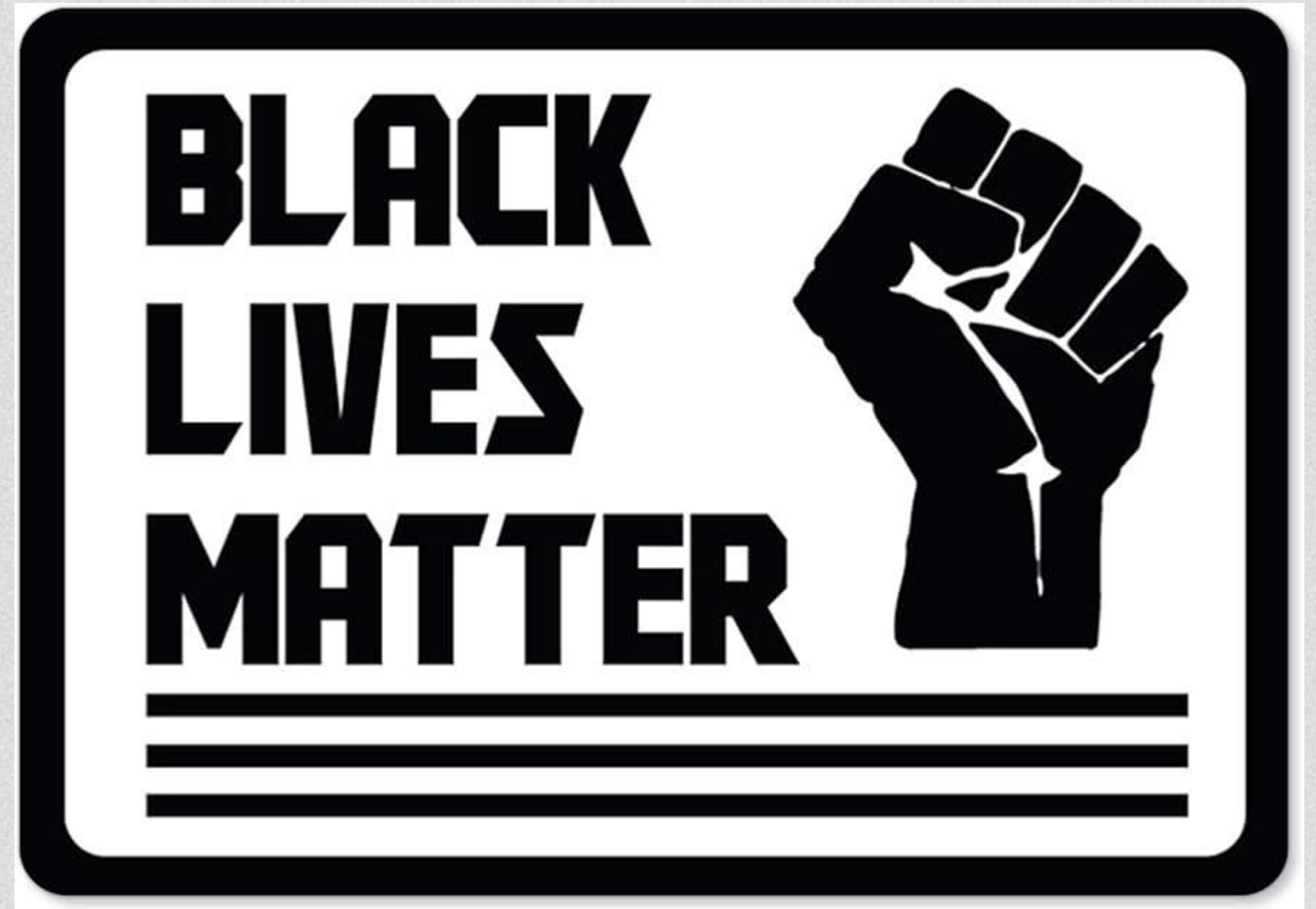
데이터 분류 방식에도 함정이 숨어 있음
여성/남성 분류 속 성소수자는 지워짐

아주 오래된, 인류가 만든 편향적 체계가
인공지능이라는 첨단기술에 그대로 계승된 것

데이터의 함정

최신성:

사회는 계속해서 변화한다



학습 데이터가 지속적으로 업데이트 되지
않는다면, 사회의 급격한 변화를 반영할 수 없음

데이터의 함정

편향에서 벗어나는 것은
불가능에 가깝다

Why Google's 'woke' AI problem won't be an easy fix

28 February 2024

By Zoe Kleinman, Technology editor

Share



Google/ Gemini

출처 BBC

인종, 국가, 문화별로 다 다르게 발생하는 편향
지점을 일일이 보정한다는 것은 불가능

데이터의 합정

개인정보를 보호하지 않는
AI

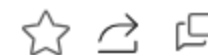


데이터의 합정

개인정보를 보호하지 않는
AI

“내 정보 챗GPT 돈벌이에 동의한 적 없어”...오픈AI, 집단소송 위기

입력 : 2023.06.29 10:32 | 수정 : 2023.06.29 15:55 김은성 기자



출처 경향신문

우리의 데이터를 동의 없이 크롤링하고
구매/판매해서 AI에 학습시키고 있음

텍스트 뿐 아니라 그림도 학습해 마음대로 사용
저작권 충돌도 발생 중

[단독] “내 그림을, 나를 무단 학습한
AI를 고발한다”...할리우드의 봉기

[AI의 습격, 인간의 반격] 할리우드 창작자 7인 인터뷰
AI, 창작물 학습해 '적당한' 이미지 공급...창작자들 소송

기자 [원지선](#)

수정 2024-02-13 13:15 등록 2024-02-13 09:00

출처 한겨레

감사합니다.

인공지능은 언어를 이해하고 있을까

장여경

01 잘못된 경로



인공지능은 ‘똑똑’하다?

구글, 추론능력 높은 AI '제미나이' 첫선... "챗GPT-4보다 똑똑" (상보)

사람처럼 사고하는 '멀티모달' 적용... 문자·음성·영상 등 유기적 분석
나노·프로·울트라 세가지 버전 출시...MMLU서 최초로 인간 전문성 능가

(서울=뉴스1) 김성식 기자 | 2023-12-07 16:45 송고 | 2023-12-08 09:26 최종수정

"챗GPT보다 한국어 6500배 더 똑똑"...네이버, '하이퍼 클로바X' 7월 공개(종합)

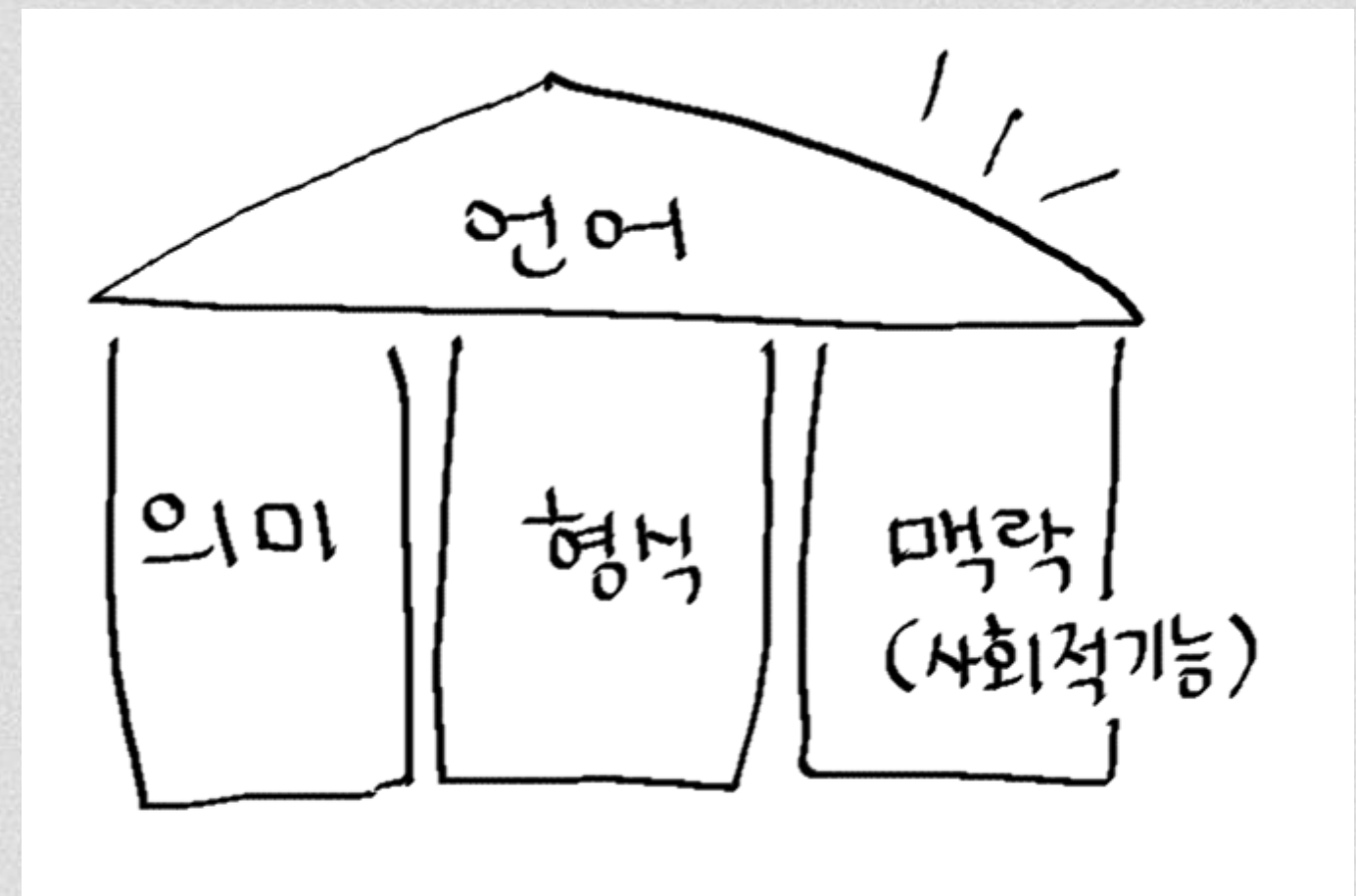
등록 2023.02.27 11:34:48 | 수정 2023.02.27 13:06:48



SF작가 테드 창

“챗GPT가 ‘반갑습니다’라고
말하도록 만들 수는 있겠지만,
챗GPT는 여러분에게 반가움을
느끼지 않습니다.”

출처: 한겨레 2024. 6. 12.



출처: <https://maily.so/forestoflanguage/posts/5bcf23c8>

인간의 언어를 ‘이해’하는 AI의 연구개발

“주입식 시험을 잘보는 언어모델”

그러나 언어모델에서 실제 언어에 대한 ‘이해’는 없다

언어모델 훈련은 ‘형식’에
대해서만 이루어진다.

언어모델은 결코 ‘의미’에
도달할 수 없다.

02

확률적 앵무새



Tesla fatal crash: 'autopilot' mode sped up car before driver killed, report finds

Findings about crash in Silicon Valley raise fresh concerns about limits of Elon Musk's technology

Sam Levin in San Francisco

@SamLevin
Fri 8 Jun 2018 00:19 BST



The crash is not the first time that safety issues have been raised about Tesla's 'autopilot' mode. Photograph: Francois Mori/AP

A Tesla driving in "autopilot" mode crashed in March when the vehicle sped up and steered into a concrete barrier, according to a new report on the fatal collision, raising fresh concerns about Elon Musk's technology.

Elon Musk car crash

On March 18, 2018, Tesla, Inc. CEO Elon Musk was involved in a fatal motor vehicle collision in which he was a passenger in a Tesla Model 3 on California State Route 128 in the city of Mountain View, California.

Tesla. Musk was ejected from the Tesla, struck the windshield of the other vehicle, and was killed. The Mercedes-Benz driver was unharmed.

Woman, 77, Ejected From Tesla, Struck By BMW and Killed on 405 in Long Beach

By City News Service • Published April 16, 2019

Motorcyclist dies in ejection crash on State Route 128

왜 인간은 언어모델의 결과물을 의미가 있는 것으로 착각할까?

“겉보기에
일관성이 있어 보이는 것”

“보이는 것”은 관찰자=인간의 입장이다

인간에게 대화란,

* 대화상대방과 공통점을
공유하고 있다.

* 언어를 통해 전달하고자 하는
의사소통 의도를 가지고 있다.

* 의사소통하는 서로의 정서
상태를 모형화한다.

‘행간’을 읽는 인간

인간은 의사소통하면서

발화자가 누군지 모를때조차

- * 이 사람이 어떤 사람인지,
- * 우리와 공유하는 점이 무엇인지
모형화하여

대화의 ‘암묵적인’ 의미를
해석한다.

유창해 ‘보이는’ 것은 기계의 능력이 아니라 인간의 이해능력

언어모델이 생성하는 텍스트에는

- * 의사소통 의도가 없다.
- * 인간 정서에 대한 모형을 가지고
있지 않다.

훈련 데이터는

- * 대화상대방 인간과 공유하는
‘생각’을 포함하고 있지 않다.

그래서 언어모델은 “확률적 앵무새”이다

방대한 훈련 데이터에서 관찰한
언어 형태의 배열을

그 결합 방식에 대한 확률적 정보에 따라
무작위로 이어붙인 시스템이다.

의미는 전혀 참조하고 있지 않다.

위험성

훈련 데이터의 지배적 세계관을 흡수한다.

* 은밀한 하위패턴

(예를 들어) “여성 의사”라는 단어에
: 의사가 여성이 아닌 것으로 전제
: 성별이분법 외 정체성을 배제

* 직접적으로 논쟁적인 어휘

: 미등록 이주민 vs. 불법 체류자

편향을 재생산/증폭

인간에 미치는 영향이 커진다.

- * 심리적인 영향
- * 예속, 배제, 경시, 기회 상실
- * 의식/무의식적인 차별

언어모델에 미치는 영향이 커진다.

- * 미래 언어모델은 앞선 언어모델이 출력한 생성물로 훈련한다.

응용될 때 위험성

텍스트 분류의 편향이 생긴다.

* 검색의 편향이 생긴다.

시스템의 직접 사용자와
의사결정 대상이 되는 사람에게
이런 과정이 보이지 않기 때문에
특히 더 치명적이다.

악의적인 사용자가 조작하였을 때

악의적 텍스트를 대량으로
자동적으로 생성한다.

공론장을 극단적인 댓글이
지배한다.

* 여론을 왜곡할 수 있다.

* 폭력을 선동할 수도 있다.

의인화할수록 기만 위험이
높아진다.

부정확한 출력



부정확한데도
 유창하고 일관적(이어 보이는)
 결과물을 생성할수 있다.
 * 오류 ----- 구금

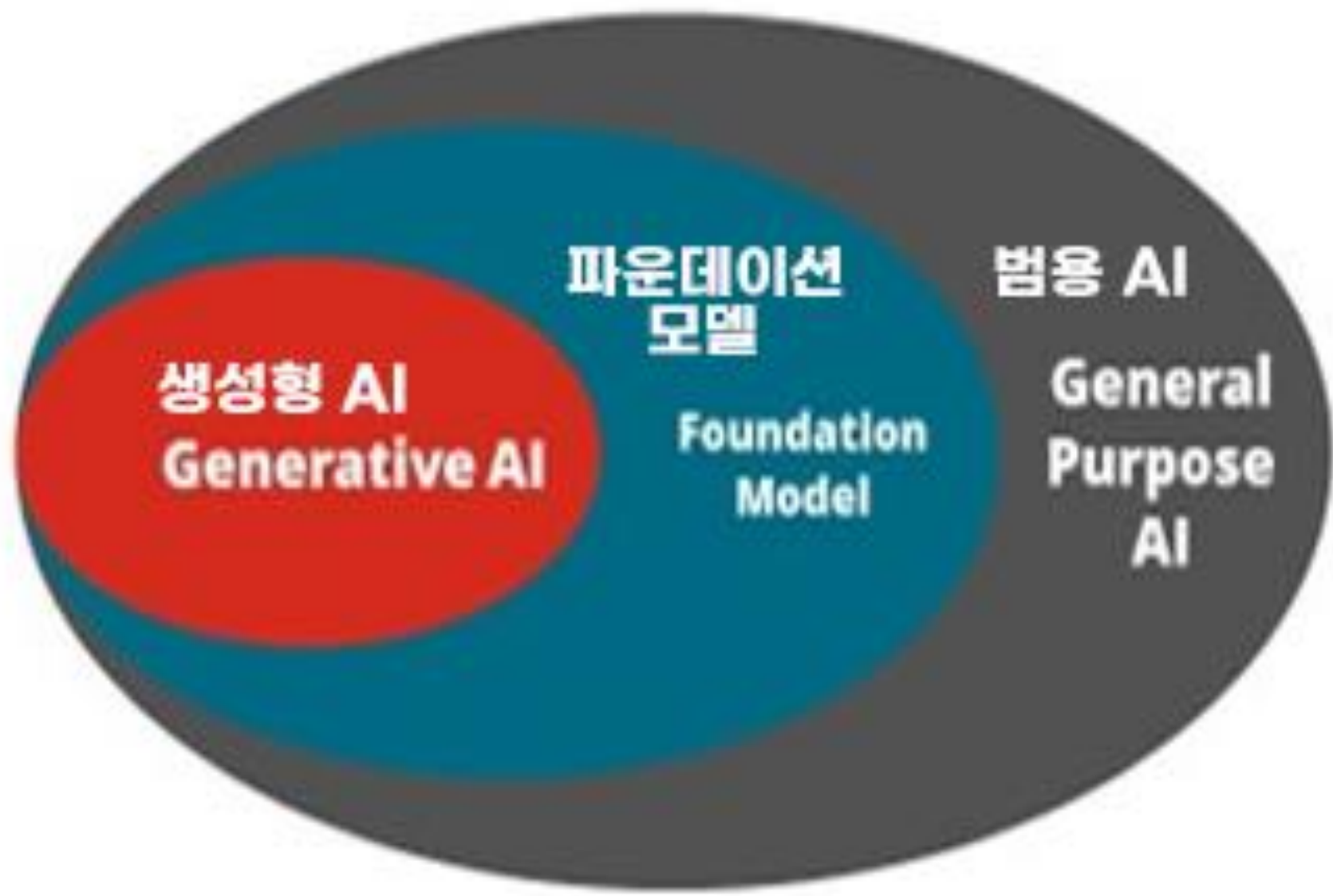
정보 추출의 위험성



인터넷 공개정보로 훈련한 경우
 이런 위험을 완전히 제거하기
 어렵다.

생성형 AI의 미래

<https://www.jdsupra.com/legalnews/chatgpt-in-the-crosshairs-how-the-eu-s-6938556/>



범용 AI

특정한 응용을 위해 설계된 것이 아니라 "광범위한 응용"에 사용되거나 수정하여 활용하기 위한 AI 시스템



파운데이션 모델

광범위한 대규모 데이터로 훈련되고 보편적 결과물을 위해 설계되며 광범위한 업무에 적용되는 AI 모델



생성형 AI

텍스트, 이미지, 오디오, 비디오와 같은 콘텐츠를 생성하기 위해 특별히 의도된 AI 시스템

범용AI의 위험

악의적인 사용 위험

- * 허위 콘텐츠, 여론 조작, 사이버 범죄, 이중 사용

오작동 위험

- * 기능상 문제, 편향과 과소대표, 통제력 상실

시스템적 위험

- * 노동 문제, 지구적 격차, 시장 독과점, 환경 영향, 개인정보 침해, 저작권 침해

AI 위험 규제

유럽의회, 세계 첫 'AI법' 통과...AI로 생체 정보 수집 금지

중앙일보 | 입력 2024.03.14 16:17

유엔, 만장일치로 AI 관련 첫 결의..."안전한 사용 합의 시급"

송고시간 2024-03-22 17:23:01

"AI의 안보·건강 위험, 정부에 보고해야"...바이든 행정명령

기자 이본영

수정 2023-10-31 19:44 등록 2023-10-31 12:05

한국은...

AI법 과방위 법안소위 통과..."우선허용-사후규제"

입력 2023-02-14 21:57:00 수정 2023.02.14 21:57:00 박예나 기자

"AI 규제보다 산업 육성·생태계 확산이 우선"

박남수 기자 | 승인 2023.05.17 19:42 | 댓글 0



「인공지능 법률안」, '우선허용·사후규제' 원칙 삭제하고, 인권영향평가 도입해야

담당부서: 인권정책과 | 등록일: 2023-08-24 | 조회: 2207

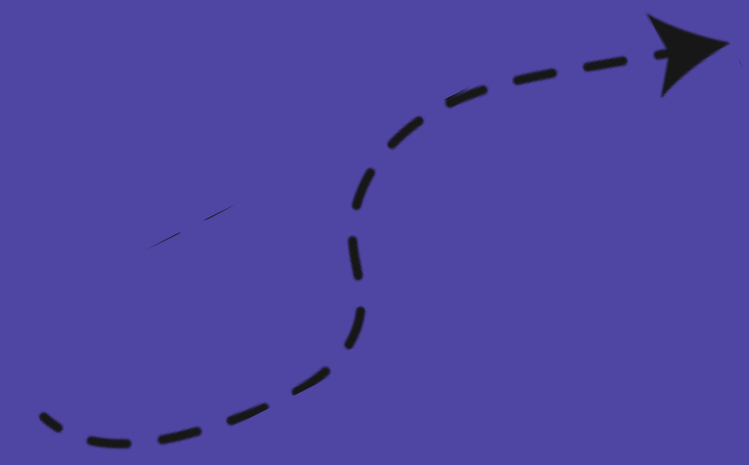
"생성형AI 성장기...규제보다 진흥 초점 맞춰야"

등록 2023.12.29 09:57:06



03

올바른 경로



더 많은 데이터로 더 큰 언어모델을 만들기보다

에너지 절약적이고 효율적으로 데이터를 조정하자

확률적 앵무새로서 언어모델의 위험을 인정하고 피해 완화 방법론을 모색하자

사회기술 시스템을 어떻게 형성할까



“인공지능 시스템에
세상의 아름다움, 추함, 잔인함
을 먹이면서
그것이 아름다움만을 반영하기를
기대하는 것은
환상이다.”

04 질의응답

발표를 마치며 참석자들의 질문에 답변합니다.



감사합니다.
