

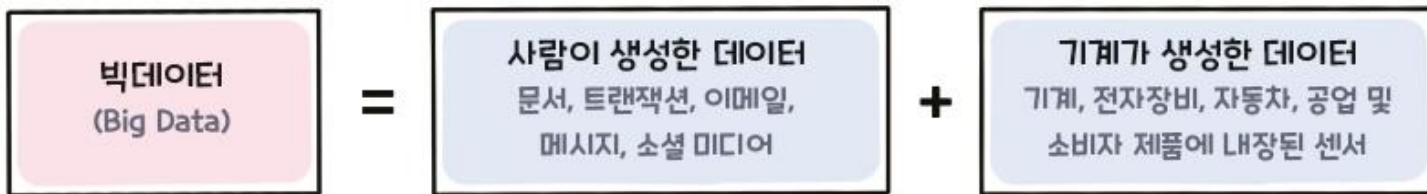
03

Hyperscale Data Analysis

❖ Concepts

■ Big Data

- Data generated in a digital environment
- Big data containing text and video data as well as numerical data, is massive compared to data produced in analog environments in the past



❖ Concepts

- Difference Between Big Data and Traditional General Data
 - Big data means collected from various methods, sources, and environment
 - Big data is big enough to require a computer system for parallel processing
 - Big data creates value for business or research
 - Validity and reliability must be secured to ensure the value created by big data

구분	일반 데이터	빅데이터
데이터의 원천	내부로부터 수집	외부로부터 수집
데이터의 형태	정형 데이터가 대부분	비정형 데이터
분석 방법	모델링	인공지능
분석 환경	기업 내에 구축된 데이터웨어하우스	클라우드

❖ Concepts

하나 더 알기 데이터 단위

- 비트(Bit) : 가장 작은 데이터 단위
- 8비트 = 1바이트(Byte), 영어나 숫자는 1바이트, 한글은 2바이트
- 1킬로바이트(KB) = 1,024바이트(Byte)
- 1메가바이트(MB) = 1,024킬로바이트(KB)

표 7-2 데이터 단위

이름	기호	값	이름	기호	값
킬로바이트	KB	$1024^1 = 2^{10}$	페타바이트	PB	$1024^5 = 2^{50}$
메가바이트	MB	$1024^2 = 2^{20}$	엑사바이트	EB	$1024^6 = 2^{60}$
기가바이트	GB	$1024^3 = 2^{30}$	제타바이트	ZB	$1024^7 = 2^{70}$
테라바이트	TB	$1024^4 = 2^{40}$	요타바이트	YB	$1024^8 = 2^{80}$

❖ Features

- 3V : Volume + Velocity + Variety
 - Volume
 - Amount of data stored on the physical device
 - Velocity
 - Ensure real-time processing of your data
 - Variety
 - Containing various forms of data

❖ Features

■ 4V : 3V + Veracity

- Veracity
 - It's an indication of how valuable and useful your data is, and it's about the accuracy of data collected by companies or institutions in analyzing big datas

■ 5V : 4V + Value

- Value
 - Big data is meaningful only when it can be used for business to drive value
 - Before designing and collecting big data, you need to think about what you can do with it

❖ Features

■ 6V : 5V + Validity

- Validity
 - Accuracy of data

■ 7V : 6V + Volatility

- Volatility
 - About how long data can be stored and used

❖ Features

하나 더 알기 초거대 데이터 특징 좀 더 알아보기

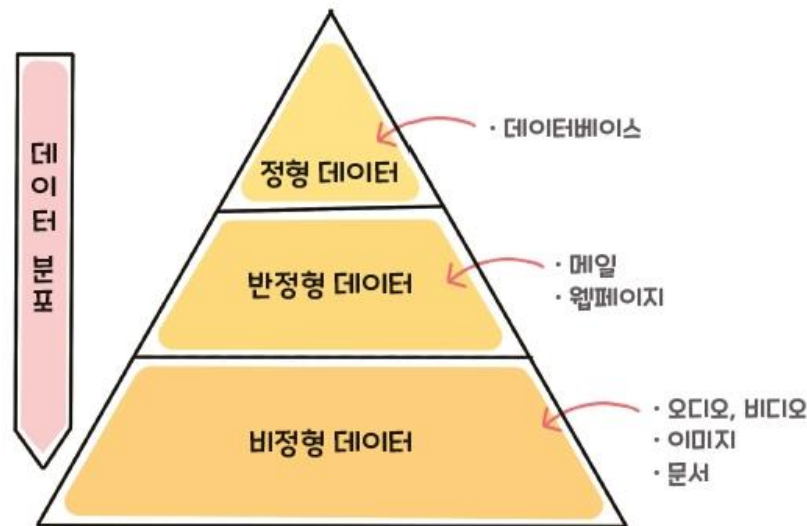
초거대 데이터는 추가적으로 가변성과 시각화의 특징이 있음

- **가변성(Variability)** : 데이터의 맥락에 따라 의미가 달라지는 것
- **시각화(Visualization)**
: 데이터를 시각적으로 표현하는 것



❖ Kinds of Data

- Structured data
- Unstructured data
- Semi-structured data



❖ Kinds of Data

■ Structured data

- Data stored according to a certain format or rule
- (Examples) spreadsheets, relational databases, CSVs, etc

ID	Name	Age	Degree
1	John	18	B.Sc.
2	Jason	32	Ph.D.
3	Robert	52	Ph.D.
4	Ricky	35	M.Sc.
5	Gibb	26	B.Sc.

❖ Kinds of Data

■ Unstructured data

- Contrary to structured data, it is often difficult to grasp the meaning of values because there are no fixed rules
- With the development of artificial intelligence technology, the importance of unstructured data is increasing as more and more cases of acquiring insights from unstructured data
- Examples of unstructured data: SNS, video, image, voice, text, etc.

Artificial intelligence(AI), is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality. The distinction between the former and the latter categories is often revealed by the acronym chosen. 'Strong' AI is usually labelled as AGI(Artificial General Intelligence) while attempts to emulate 'natural' intelligence have been called ABI(Artificial Biological Intelligence).

❖ Kinds of Data

■ Semi-structured data

- Data with characteristics of schema and metadata
- Semi-structured data is attracting attention because it can use a lot of data that exists on the Internet for artificial intelligence
- (Examples) XML, JSON, NoSQL, logs, etc.

```
<University>
  <Student ID="1">
    <Name> John </Name>
    <Age> 18 </Age>
    <Degree> B.Sc. </Degree>
  </Student>
  <Student ID="2">
    <Name> Jason </Name>
    <Age> 32 </Age>
    <Degree> Ph.D. </Degree>
  </Student>
  ...
</University>
```

❖ Kinds of Data

하나 더 알기

XML, JSON, NoSQL

1. XML(eXtensible Markup Language)

XML은 확장 가능한 마크업 언어라는 뜻으로, HTML처럼 태그(Tag)들을 고정시켜 사용하지 않고 확장이 가능함.

2. JSON(JavaScript Object Notation)

JSON은 데이터를 저장하거나 전송할 때 많이 사용하는 데이터 교환 형식.

```
{
  "employees": [
    {
      "name": "Surim",
      "lastName": "Son"
    },
    {
      "name": "Someone",
      "lastName": "Huh"
    },
    {
      "name": "Someone else",
      "lastName": "Kim"
    }
  ]
}
```

[JSON 표현식]

21

❖ Kinds of Data

하나 더 알기

XML, JSON, NoSQL

3. NoSQL(Not Only SQL)

NoSQL은 스키마의 특성을 가지며, 유연하고 분산 병렬처리가 쉬워 확장에 유리함. 빅데이터를 다룰 때 많이 사용되는 데이터베이스

❖ Relationship for AI

- Artificial intelligence provides an analysis method for big data
- AI technology makes it easy and fast for any user to find Insight



❖ Relationship for AI

- Artificial intelligence technology combines with big data to conduct massive data learning
- Based on this, you can provide insights for decision making and predict future events



❖ Relationship for AI

- **'Convergence of artificial intelligence and big data'** is the most notable technology when creating business value through big data and data analysis functions

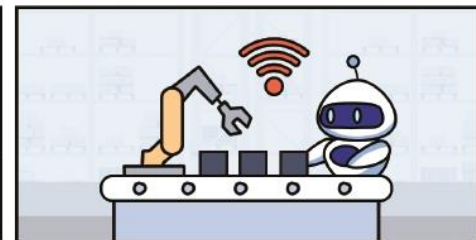


❖ Relationship for AI

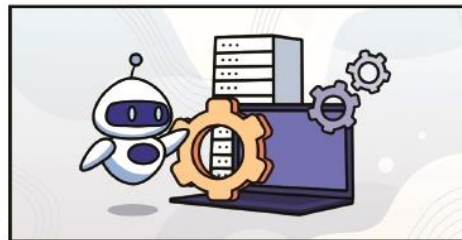
- Changes due to the convergence of hyperscale data and AI
 - Connecting to the Internet of Things(IoT)
 - Transition to hyper-connected industrial ecosystem
 - Convergence with cyber systems
 - Smart machine appears



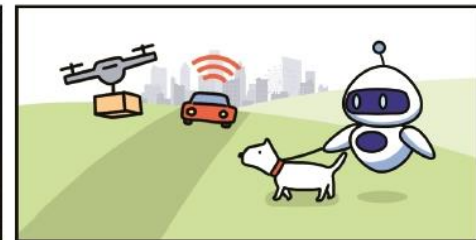
사물인터넷으로 연결



초연결 산업 생태계로 전환



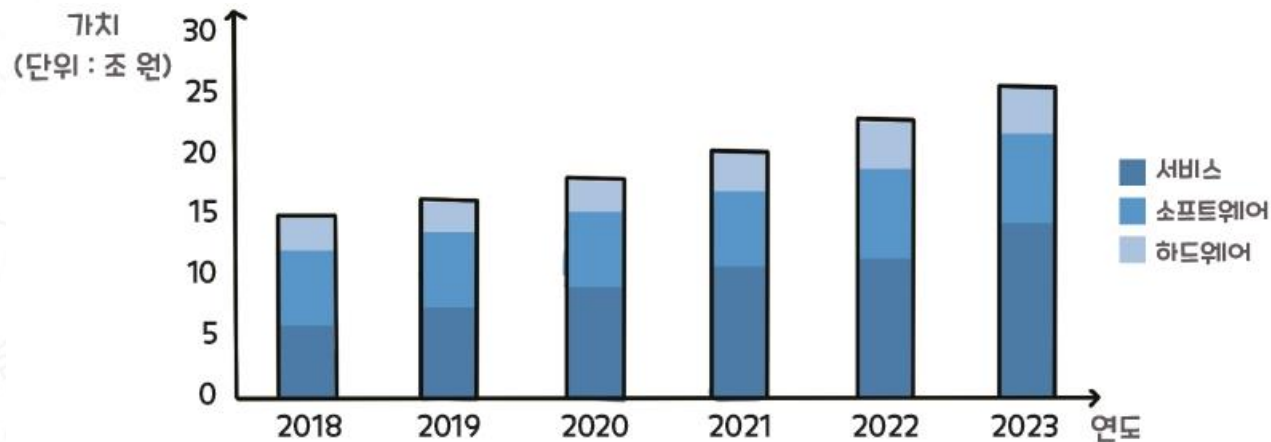
사이버 시스템과 융합



스마트 머신 등장

❖ Relationship for AI

- Countries or companies that own big data platforms will dominate the era of the 4th Industrial Revolution

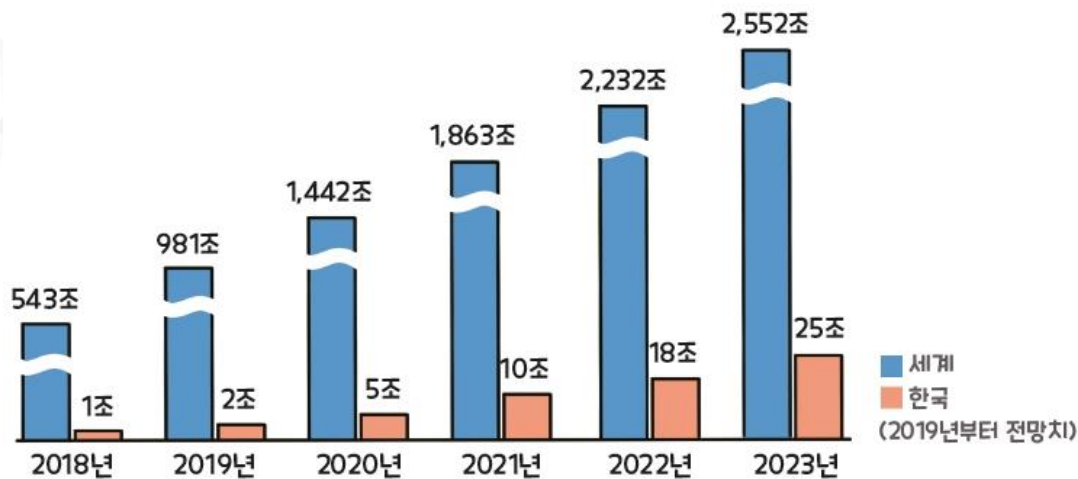


❖ Applications

■ Finance

• Robo-Advisor

- A compound word of 'Robot' and 'Investment Specialist'
- Technology that combines IT technology and financial theory to allow computers to manage assets on behalf of people



❖ Applications

하나 더 알기 핀테크

- **핀테크(FinTech)** : 금융에 IT기술을 접목하여 복잡하고 어려웠던 금융 업무를 효율적이면서 편리하게 서비스하는 것
- **핀테크 관련 서비스** : 페이코(PAYCO), 토스(Toss), 카카오페이(KakaoPay) 등



❖ Applications

■ Healthcare

• IBM Watson

- Integrated analysis of hospital medical records, academic papers, and biometric data to help patients with treatment
- But Watson's introduction didn't benefit the industry much, so it's now discontinued and turned into a clinical trial service



❖ Applications

▪ Distribution

- In the distribution industry, visual search using images and chatbots introduced in online and offline stores are expanding
 - Amazon Go : Unmanned mart run by Amazon



❖ Applications

▪ ICT

- The latest AI appliances self-learn the user's lifestyle and space characteristics to find the best way to operate
 - 삼성전자 무풍 에어컨 : The ability to learn the user's pattern
 - LG전자 트롬 세탁기 : Recommend context-sensitive laundry options after learning your pattern



❖ Necessity of Public Data

- The short-term way to achieve the effects of big data is to utilize government-owned data
- If the government opens up big data, the private sector will be able to develop services without much effort

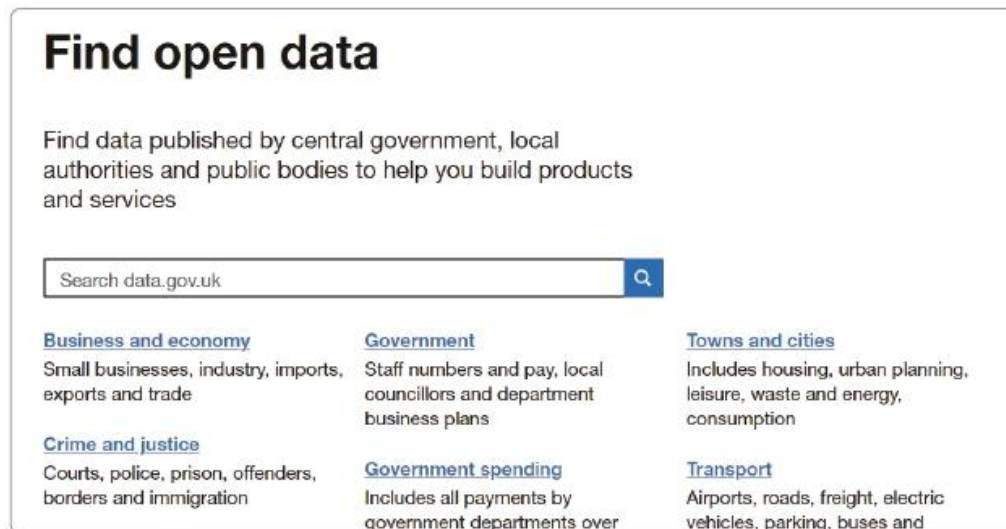
❖ Examples of Public Data

- US launches data portal site (data.gov) in 2009 to open public data
 - The U.S. government's data openness ranges from agriculture, environment and energy sectors to budgets, expenditures, contracts, and civil servants' salaries for financial transparency



❖ Examples of Public Data

- UK began opening public data in 2010 when Prime Minister Gordon Brown invited Tim Berners-Lee, the founder of web and linked data, to create a data portal site (data.gov.uk)
 - Since 2018, it has been reorganized under the name of "Find Open Data"



Find open data

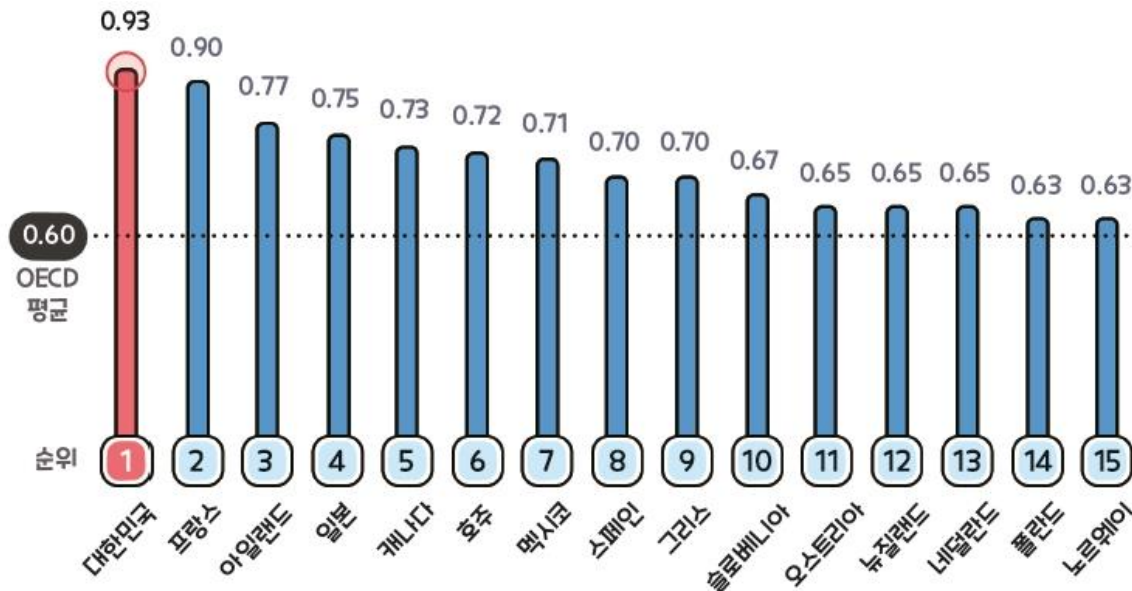
Find data published by central government, local authorities and public bodies to help you build products and services

Search data.gov.uk

Business and economy Small businesses, industry, imports, exports and trade	Government Staff numbers and pay, local councillors and department business plans	Towns and cities Includes housing, urban planning, leisure, waste and energy, consumption
Crime and justice Courts, police, prison, offenders, borders and immigration	Government spending Includes all payments by government departments over	Transport Airports, roads, freight, electric vehicles, parking, buses and

❖ Examples of Public Data

- South Korea ranked first in the "2019 OECD White Paper" released by the OECD with 0.93 points in the public data opening index



❖ Utilization of public data

- While Korea's public data openness index is high, the reason why the actual public data utilization rate is not so high is that the utilization is low because the necessary data cannot be found



❖ Utilization of public data

분야	개방 예정인 공공데이터
자율주행(11개)	정밀도로지도, 주행환경 인식센서 융합정보, 자율주행 딥러닝 학습 정보 등
스마트시티(6개)	스마트 전력거래, 디지털 트윈 정보, 세종시 스마트에너지 정보 등
헬스케어(8개)	해부학 그림 및 의료행위 그림 정보, 한의약 전주기 정보, 식중독균 유산균 유전체 정보 등
금융정보(5개)	상장사 공시주식 정보, 비상장사 공시 재무제표, 주택저당채권 정보 등
생활환경(7개)	굴뚝 대기오염물질 정보, 산림 미세먼지 정보, 산업부문 온실가스 배출정보 등
재난안전(9개)	구조구급활동 정보, 산사태 정보, 안전·취약시설물관리 정보, 국가화재 정보 등

❖ Transparency and Reliability

- Artificial intelligence is likely to cause unexpected errors due to the complexity of biased data or algorithms
- The problem is that current artificial intelligence is a black box system that cannot explain the reason for its decision
- Unless the transparency of the learning process and results of AI is verified, reliability is of course a problem that must be raised

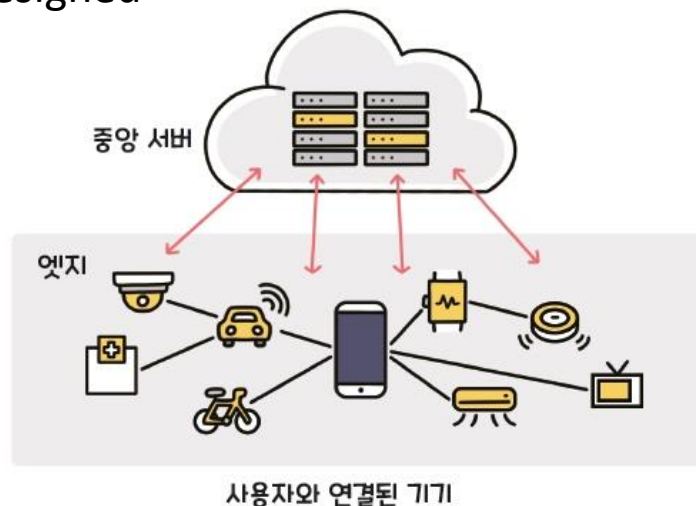


❖ Transparency and Reliability

- Various policies and guidelines for AI reliability verification are being announced, but they are mainly related to follow-up measures, so they cannot be a fundamental prevention
- Verification of data, the beginning of artificial intelligence, can be said to be the beginning of guaranteeing the reliability of artificial intelligence

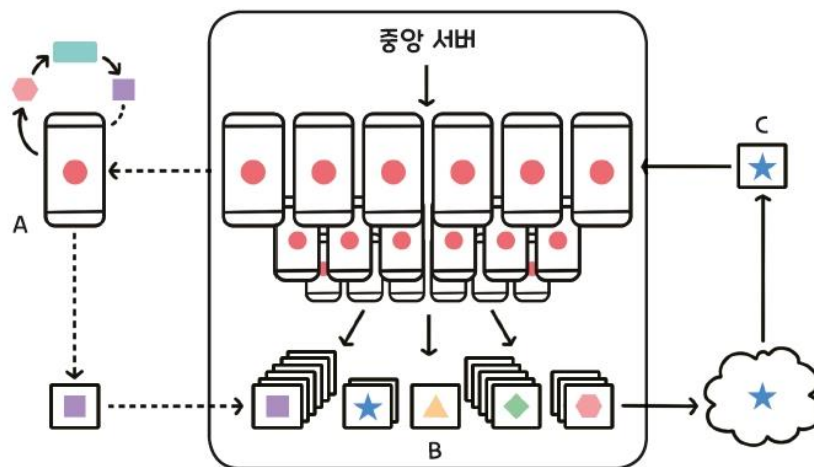
❖ FL, Federated Learning

- The performance of AI is proportional to the amount of data
- Data is typically collected at the edge and sent to the central server, whereby the central server analyzes the data and sends the results back to the edge
 - The problem is that network overload occurs in this process, so federated learning is designed



❖ FL, Federated Learning

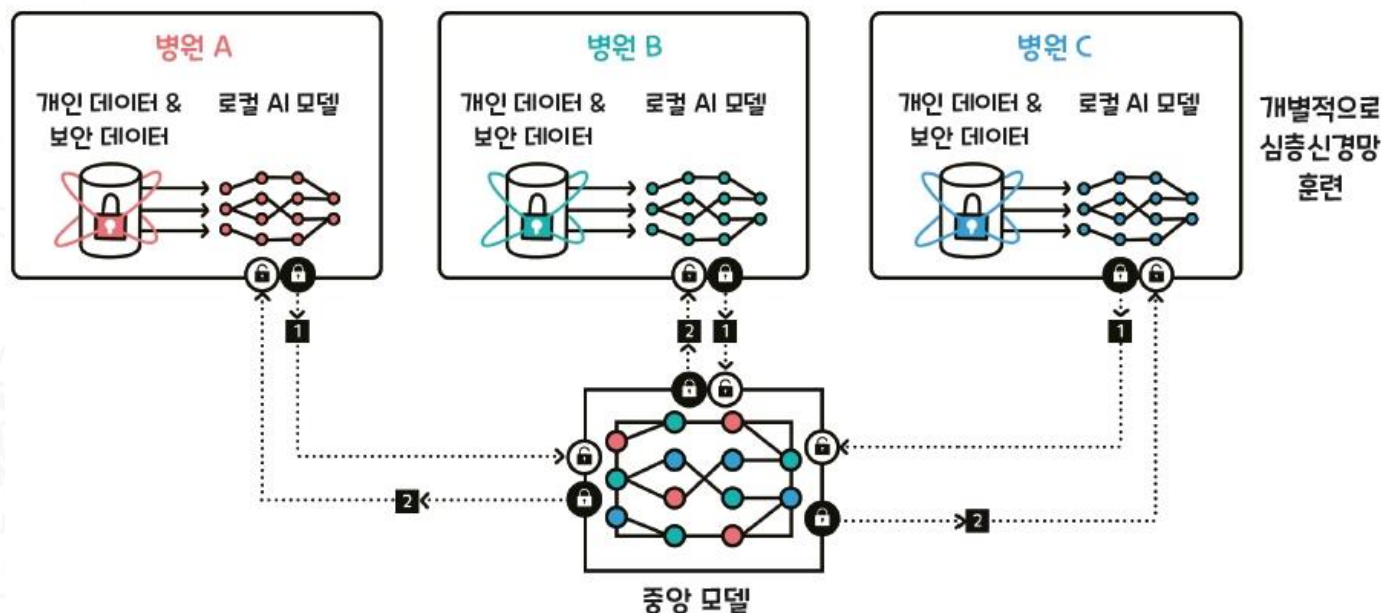
- Technology to process a plurality of individual devices and learning data



- A : 개인의 스마트폰을 이용하여 모델을 학습시킵니다. 이 과정에서 파라미터들이 수정됩니다.
- B : 신경망에서 변경된 수치값을 압축하고 암호화한 후 중앙 서버로 보냅니다.
- C : 중앙 서버에서는 모든 정보들을 합쳐서 중앙 모델에 반영하여 학습하고 그 결과값을 개별 기기에 보냅니다.
- A~C 과정을 반복합니다.

❖ FL, Federated Learning

- Federated learning has the advantage of protecting sensitive data, making it especially useful in medical institutions





❖ Model compression

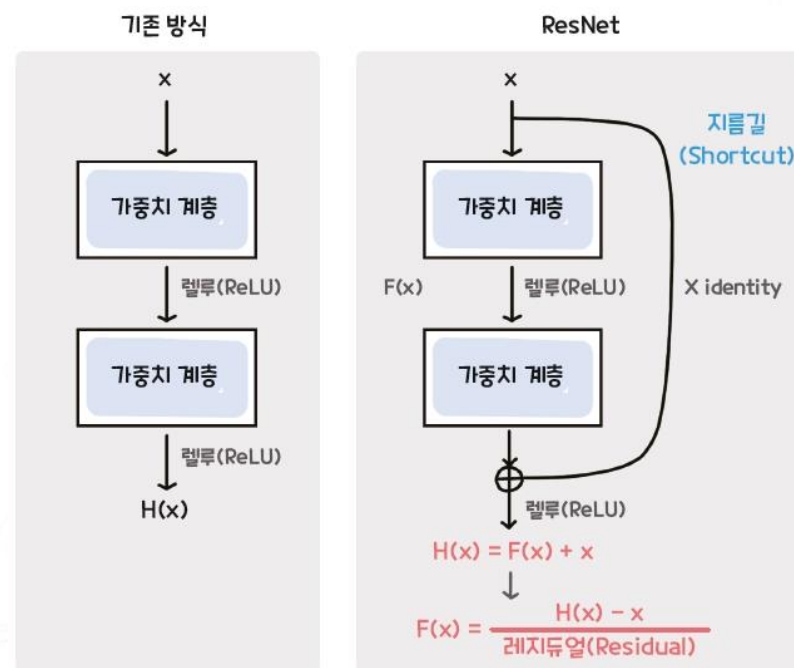
- Techniques to reduce and lighten artificial intelligence
- Specialized in reducing inference time by reducing or lightening the model size
- How to model compression
 - Pruning, quantization, distillation, low-rank approximation, etc.

❖ Model compression

- Shallow deep learning architecture
 - ResNet, DenseNet, MobileNet, ShuffleNet

- ResNet

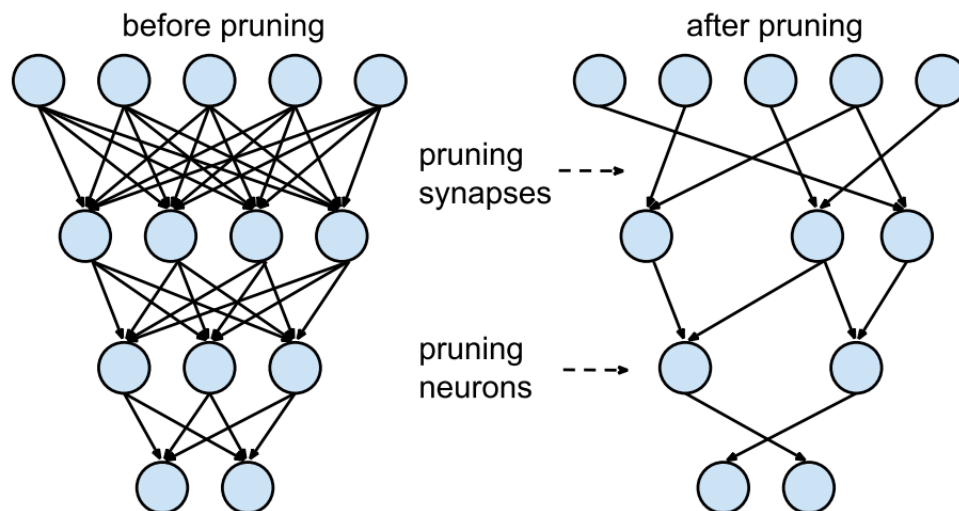
- ResNet is an image classification algorithm developed by Microsoft
- The key to this algorithm is the introduction of the concept of Residual Block to create a short cut that adds input to the output

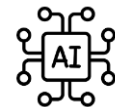


❖ Model compression

■ Pruning

- To find and remove parameters that are less important among the parameters of a neural network
- Even if the parameters are less influential, removing them may result in loss of accuracy, so additional learning is required when pruning is performed





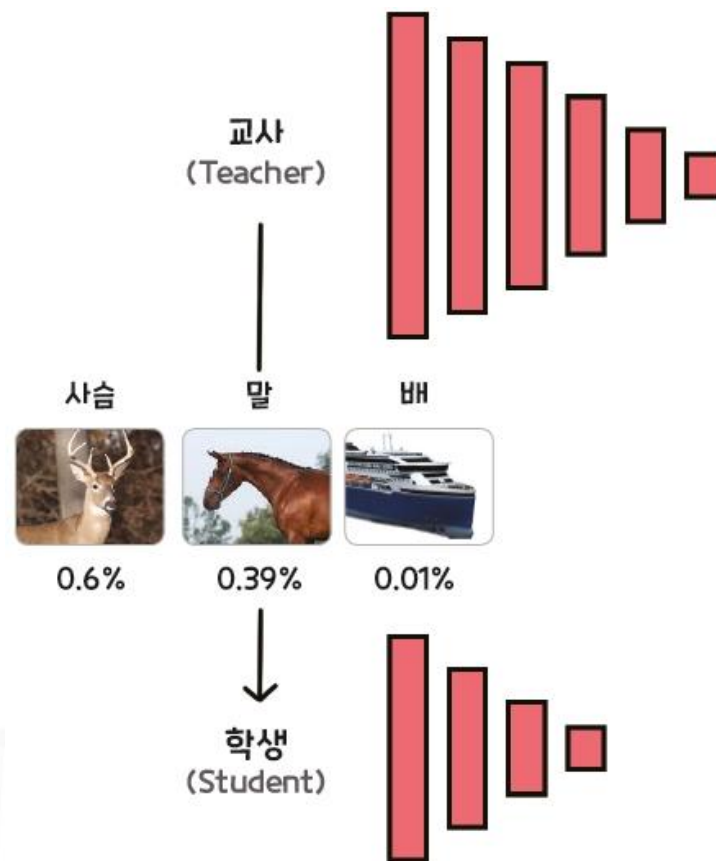
❖ Model compression

▪ Quantization

- To improve computational efficiency by reducing the precision of parameters appropriately
- Learning and reasoning in deep neural networks often eliminates the need for 32-bit or 64-bit floating-point precision
- 16-bit and 8-bit precision allows faster computation with relatively small loss of accuracy
- Recently, algorithms using 4-bit, 2-bit, and 1-bit are also being studied, and in this case, accuracy loss is large, so algorithms that allow learning by considering a small number of bits are required

❖ Model compression

- Knowledge distillation
 - A model for learning a student network from a well-learned teacher network, referred to as a teacher-student network





❖ Model compression

▪ Low-Rank Approximation

- Many convolutional operations in convolutional neural networks mainly use matrix multiplication, which allows for faster computation even if approximate solutions are obtained by reducing the rank during the convolutional operation