Introduce
Big Data & AI & Data Science

윤요섭

A Table of Contents.

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

Q.

빅데이터란?

1 사람들이 생각하는 빅데이터

2 빅데이터 정의

3. 빅데이터가 각광 받는 이유

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

빅데이터 정의

빅데이터란 무엇인가요?

Big + Data 니까
큰 데이터 아닌가요?

틀린 말은 아니지만 빅데이터의 정의를
단순히 큰 데이터라고 정의할 수는 없다.

빅데이터란?

일반적 정의에서 빅데이터는 큰 용량과 복잡성으로
기존 애플리케이션이나 툴로는 다루기 어려운
데이터 셋의 집합을 의미한다.

- 2021빅데이터 분석기사 필기

기존 데이터베이스 관리도구의 능력을 넘어서는
대량(수십 테라바이트)의 정형 또는 심지어
데이터베이스 형태가 아닌 비정형의 데이터 집합
조차 포함한 데이터로부터 가치를 추출하고
결과를 분석하는 기술

- 위키백과 “빅데이터”

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

빅데이터 정의

기관	정의
가트너	빅데이터란 향상된 시사점과 더 나은 의사결정을 위해 사용되는 비용 효율이 높고 혁신적이며 대용량, 고속 및 다양성의 특징을 가진 정보 자산을 말한다. **(2012)**
매킨지	빅데이터란 일반적으로 데이터베이스 소프트웨어가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터를 말한다.
한국데이터산업진흥원	빅데이터란 데이터에 대한 기존의 접근 방식으로는 얻을 수 없었던 통찰과 가치를 창출하는 모든 것을 말한다.

빅데이터, 인공지능, 데이터 과학 개요

빅데이터 정의

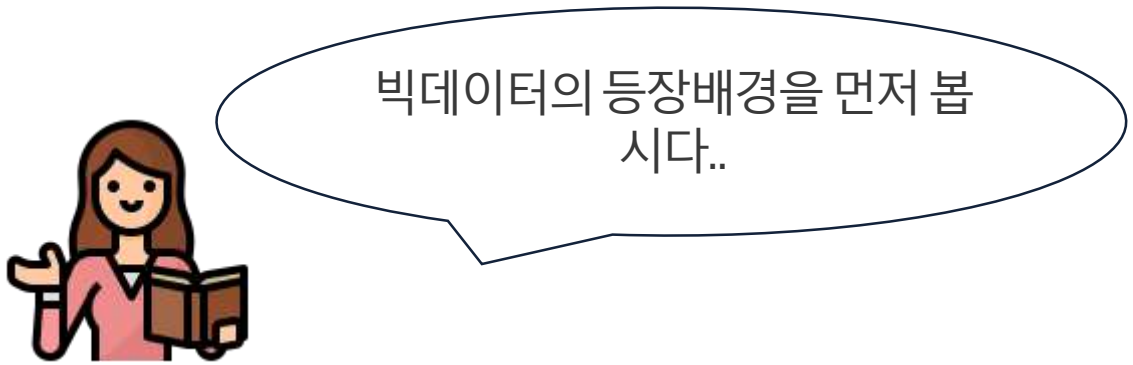
기관	정의
가트너	의사결정을 위해 사용되는 비용 효율이 높고 혁신 정보 자산을 말한다. **(2012)**
	장, 관리, 분석할 수 있는 범위를
한국데	는 얻을 수 없었던 통찰과 가치를 창출

빅데이터의 정의를 기존에 처리 할 수 없
던 양으로 정의하는 곳도 있고
가치와 통찰로 정의하는 곳도 있는데
왜 정의가 다 달라요?

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

빅데이터의 배경



데이터 저장 용량

과거에는 대량의 용량의 데이터를 저장할 수 있는 공간이 없었었다.

데이터 처리 기술

이미지, 오디오, 비디오 등과 같은 비정형 데이터를 처리하고 저장하고 활용할 수 있을 만한 기술이 없었다.

데이터의 양

데이터를 많이 수집하려고 해도 데이터를 많이 모으기가 어려운 환경

데이터의 가치

데이터 전체를 파악하고 활용할 만한 가치가 있다고 판단하지 않았으며, 데이터의 가치를 극대화 시킬 만한 분석 기술
부재

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

빅데이터의 배경

빅데이터의 등장배경을 보면 각 기관 어떤 기준에 따라 정의했는지 알 수 있어요

빅데이터 등장

데이터 저장 용량

과거에는 대량의 용량의 데이터를 저장할 수 있는 공간이 없었었다.

↓

대량의 용량의 데이터를 저장할 수 있게 되었다.

데이터 처리 기술

이미지, 오디오, 비디오 등과 같은 비정형 데이터를 처리하고 저장하고 활용할 수 있을 만한 기술이 없었다.

↓

비정형 데이터 처리 기술 증가

데이터의 양

데이터를 많이 수집하려고 해도 데이터를 많이 모으기가 어려운 환경

↓

데이터를 수집하기가 보다 쉬워지고 다양한 곳에서 데이터를 얻을 수 있게 됨

데이터의 가치

데이터 전체를 파악하고 활용할 만한 가치가 있다고 판단하지 않았으며, 데이터의 가치를 극대화 시킬 만한 분석 기술 부재

↓

데이터 가치 활용 및 분석 기술의 발달

빅데이터, 인공지능, 데이터 과학 개요

빅데이터의 특징

빅데이터의 특징:3V+2V

크기

Volume

대량의 데이터 증가 발생으로 기존 데이터 수집, 관리 한계

가치

Value

데이터 전체를 파악하고 패턴을 발견하기가 어렵게 되면서 가치의 중요성 강조

정확성

Veracity

빅데이터 기반의 예측 분석 결과에 대한 신뢰성이 중요하게 됨

다양성

Variety

비정형 데이터의 발생으로 다양한 데이터 형식 증가

속도

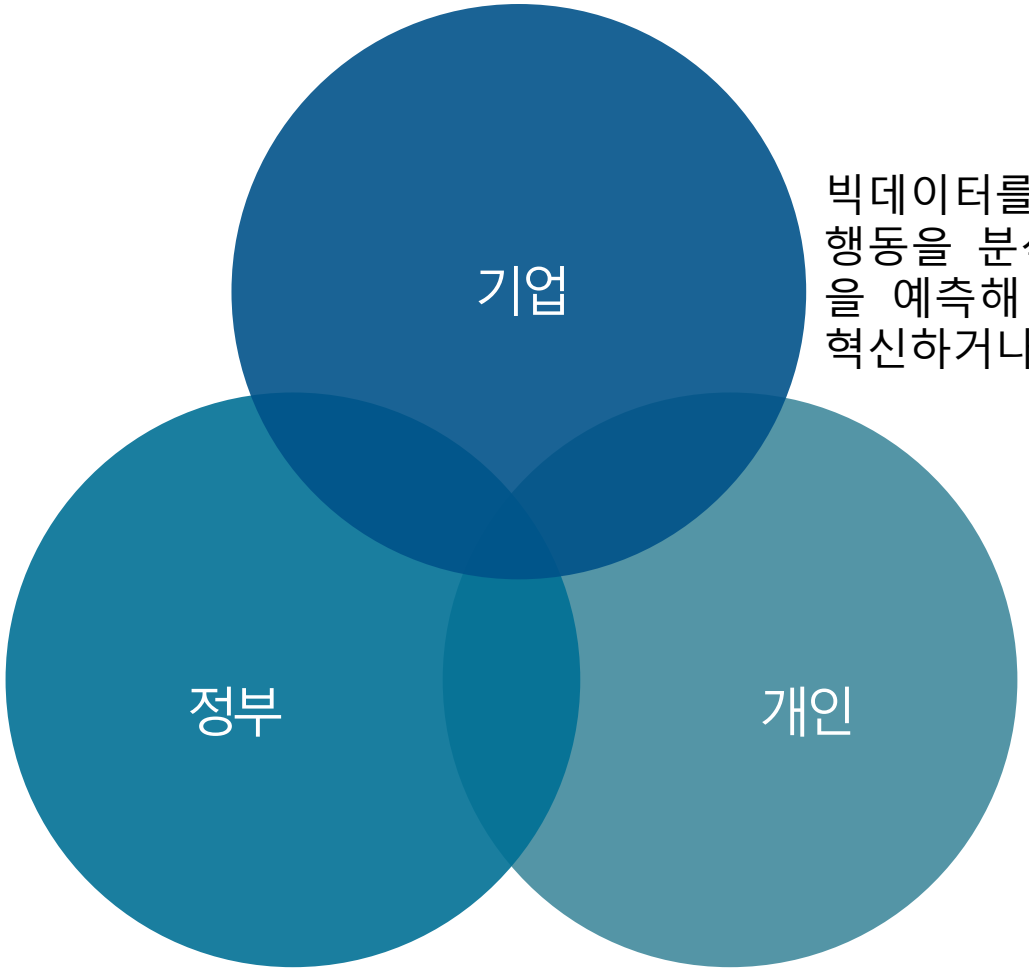
Velocity

실시간 정보 발생으로 데이터의 유입, 처리 속도 요구

빅데이터, 인공지능, 데이터 과학 개요

빅데이터가 각광받는 이유

빅데이터의 영향



빅데이터를 활용해 소비자의 행동을 분석하고, 시장 변동을 예측해 비즈니스 모델을 혁신하거나 신사업 발굴

빅데이터 활용 부문은 크게 환경 탐색, 상황 분석, 비래 대응으로 나누어서 미래성장 전략 등으로 활용

개인의 목적에 따라 빅데이터 활용이 확산되면서 스마트 라이프로의 변화

빅데이터, 인공지능, 데이터 과학 개요

빅데이터 영향

산업	빅데이터 활용
금융 서비스	신용점수 산정, 사기 탐지, 가격 책정, 프로그램 트레이딩, 클레임 분석, 고객 수익성 분석
에너지	트레이딩, 공급/수요 예측
병원	가격 책정, 고객 로열티, 수익 관리
정부	사기탐지, 사례관리, 범죄 방지, 수익 최적화
소매업	판촉, 매대 관리. 수요 예측, 재고 보충, 가격 및 제조 최적화
제조업	공급사슬 최적화, 수요 예측, 재고 보충, 보증서 분석, 맞춤형 상품 개발, 신상품 개발
운송업	일정 관리, 노선 배정, 수익 관리
헬스케어	약품 거래, 예비 진단, 질병 관리
커뮤니케이션	가격 계획 최적화, 고객 보유, 수요 예측, 생산 능력 계획, 네트워크 최적화, 고객 수익성 관리
서비스	콜센터 직원 관리, 서비스-수익 사슬 관리
온라인	웹 매트릭스, 사이트 설계, 고객 추천
모든 사업	성과 관리

Q.

인공지능이란?

1 인공지능 정의

2 인공지능이 각광 받는 이유

3. 인공지능의 종류 소개

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

인공지능 정의

인공지능이란?

인간의 학습능력, 추론능력, 지각능력을 인공적으로 구현하려는 컴퓨터 과학의 세부분야 중 하나이다.

- 위키백과“인공지능 ”

강인공지능

어떤 문제를 실제로 사고하고 해결할 수 있는 컴퓨터 기반의 인공적인 지능을 만들어 내는 것

약인공지능

인간의 오감을 컴퓨터로는 처리하기 어려웠던 문제를 컴퓨터로 처리하고 수행하게 만드는 기능

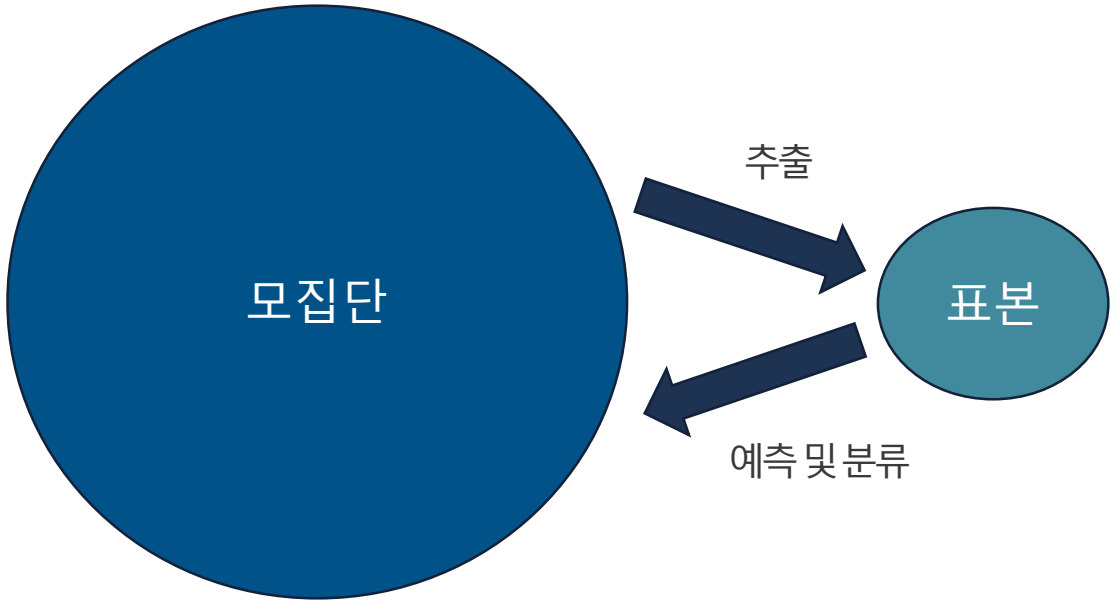
인공지능은

인간이 가지고 있는 지능을 컴퓨터로 구현하는 것을 최종 목표로 하고 있지만 아직까지 그런 인공지능은 존재하지 않아요.

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

인공지능이 각광 받는 이유 - 모집단과 표본



표본으로 모집단을 예측하거나 분류하기 위해서는 표본의 모집단의 특성과 유사해야 함

Train Data
- 인공지능이 학습할 데이터

Target Data

Train Data로 인공지능이 학습을 한 뒤, Target Data와 인공지능이 예측한 답을 서로 비교하여 모델의 성능을 높인다.

Test Data

?

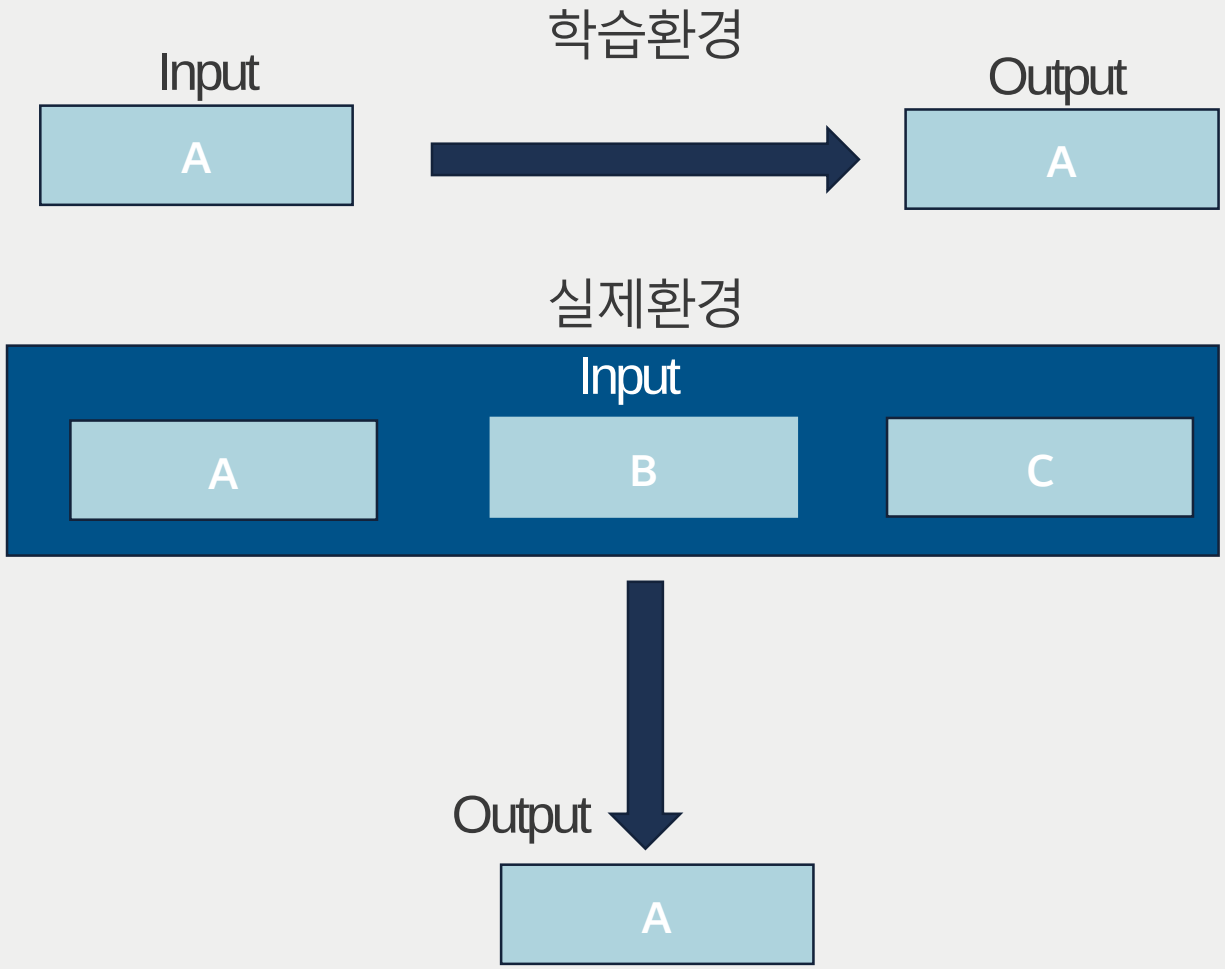
학습한 데이터를 바탕으로 이제 실제 Test Data를 입력하여 인공지능이 Target 값을 예측하는 원리

따라서, 인공지능이 잘 학습하려면 표본의 데이터가 모집단의 특성과 유사해야만 학습이 잘 이루어질 수가 있다.
(Train Data., Target Data가 엉망이면 성공적인 예측이 이루어질 수가 없음)

빅데이터, 인공지능, 데이터 과학 개요

인공지능이 각광 받는 이유 - 모집단과 표본

표본과 모집단의 특성이 다를 때의 문제



표본은 모집단의 특성을 최대한 반영할 수 있어야함

그렇지 않을 경우, 실제 데이터에 대한 예측과 분류가 의미가 사라짐

↓

모집단의 특성을 최대한 잘 반영하려면

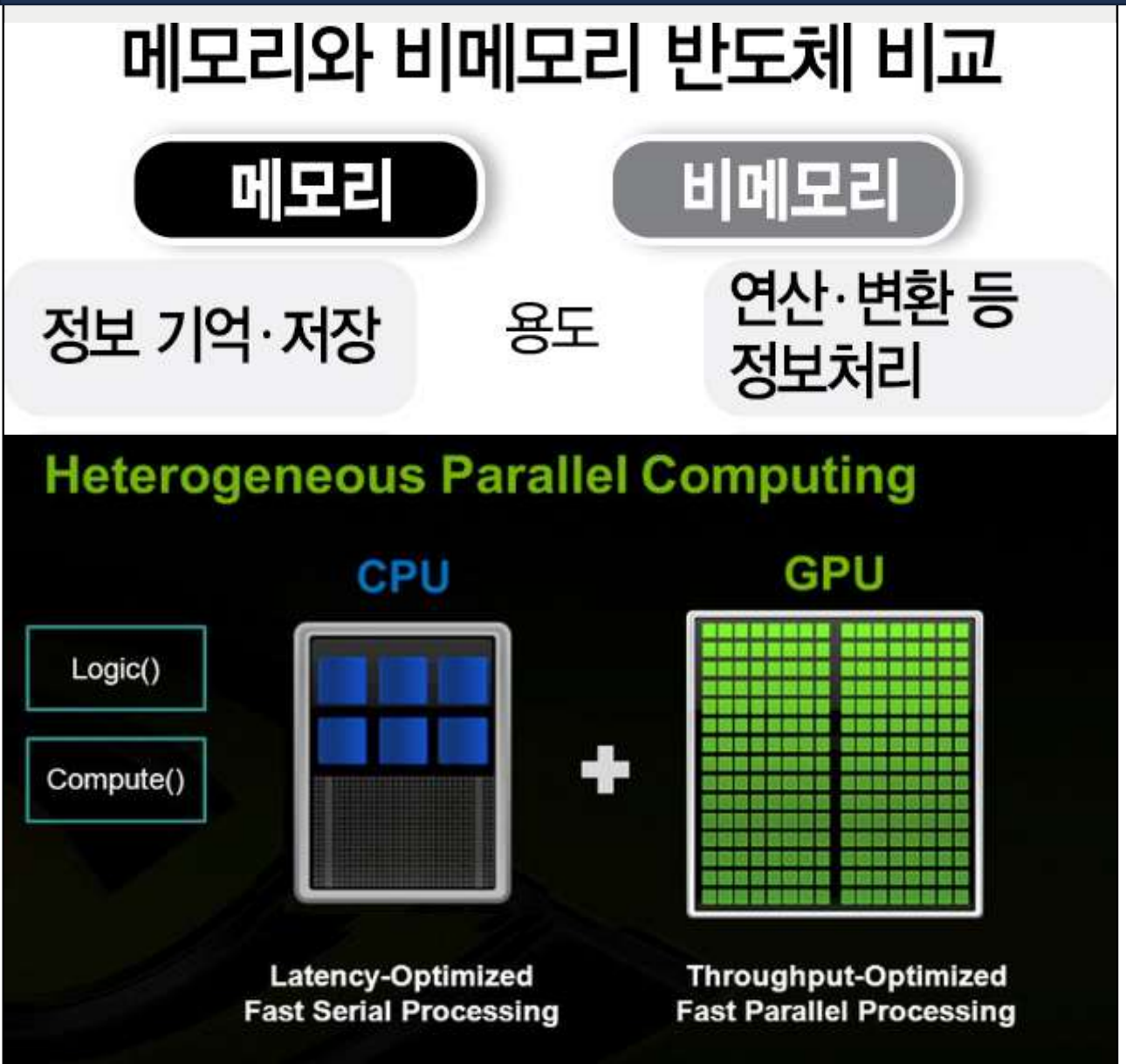
데이터의 양과 질이 좋아야 한다.

↓

빅데이터의 발전으로 수많은 데이터를 수집 및 처리할 수 있는 기술이 발전하면서 인공지능이 발전할 수 있는 계기가 됨.

빅데이터, 인공지능, 데이터 과학 개요

인공지능이 각광 받는 이유 - GPU의 발전



인공지능은 엄청난 연산 능력을 필요로 한다.

아무리 다양한 많은 데이터를 확보한다고 해도 인공지능이 연산할 수 있는 컴퓨팅 파워를 가지지 못하면, 인공지능을 학습 시킬 수가 없다.

⬇

반도체의 발전으로 많고 다양한 데이터를 빠르게 학습시킬 수 있는 환경이 갖춰졌다.

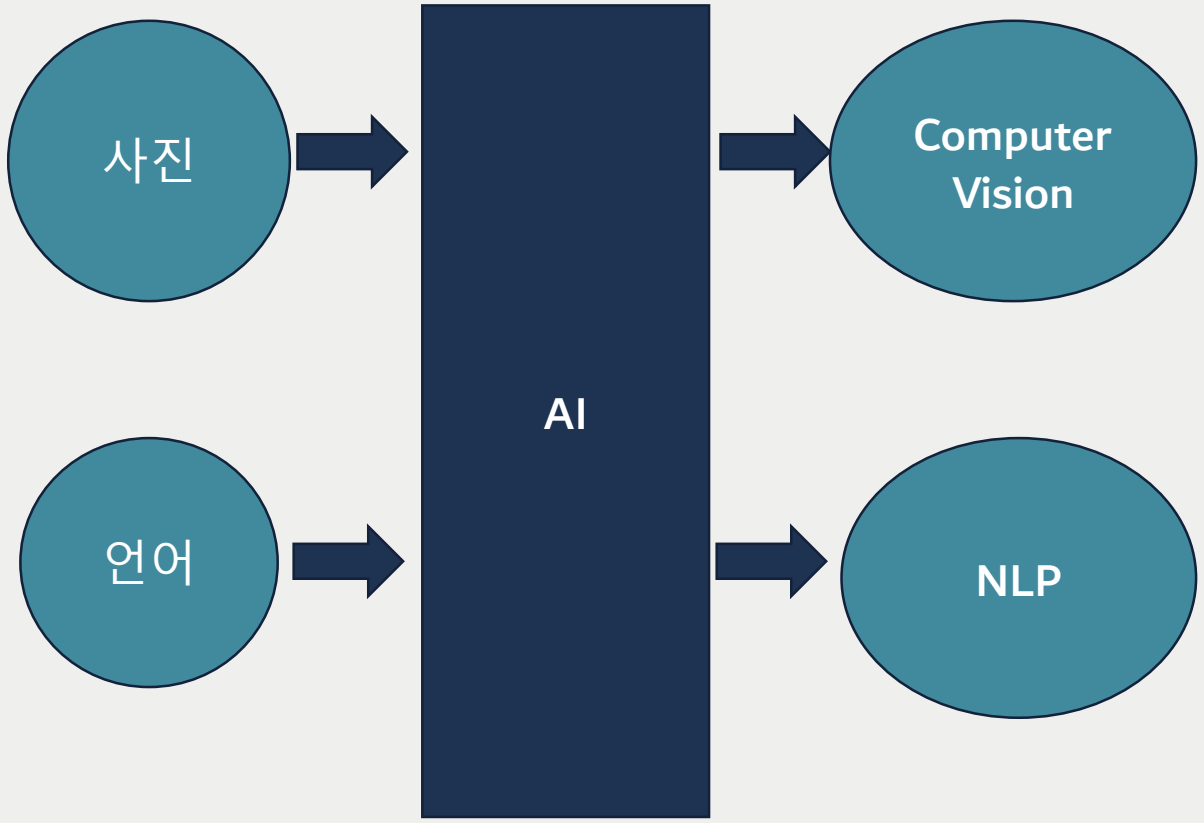
비메모리 반도체 중 GPU 발전 – 처리 연산 속도

메모리 반도체 발전 – 고용량 메모리로 학습 데이터 저장

빅데이터, 인공지능, 데이터 과학 개요

인공지능이 각광 받는 이유 - 비정형 데이터 처리

비정형 데이터 처리



비정형 데이터란 미리 정해진 구조가 없고,
고정된 필드에 저장 되지 않은 데이터를 의미

- 정형 데이터 : RDB에 저장할 수 있는 구조가 있는 데이터
- 반정형 데이터 : JSON, XML 등등
- 비정형 데이터 : SNS댓글, 오디오, 이미지, 동영상 등등

빅데이터에서 비정형 데이터가 차지하는 비율이 : 약 80%

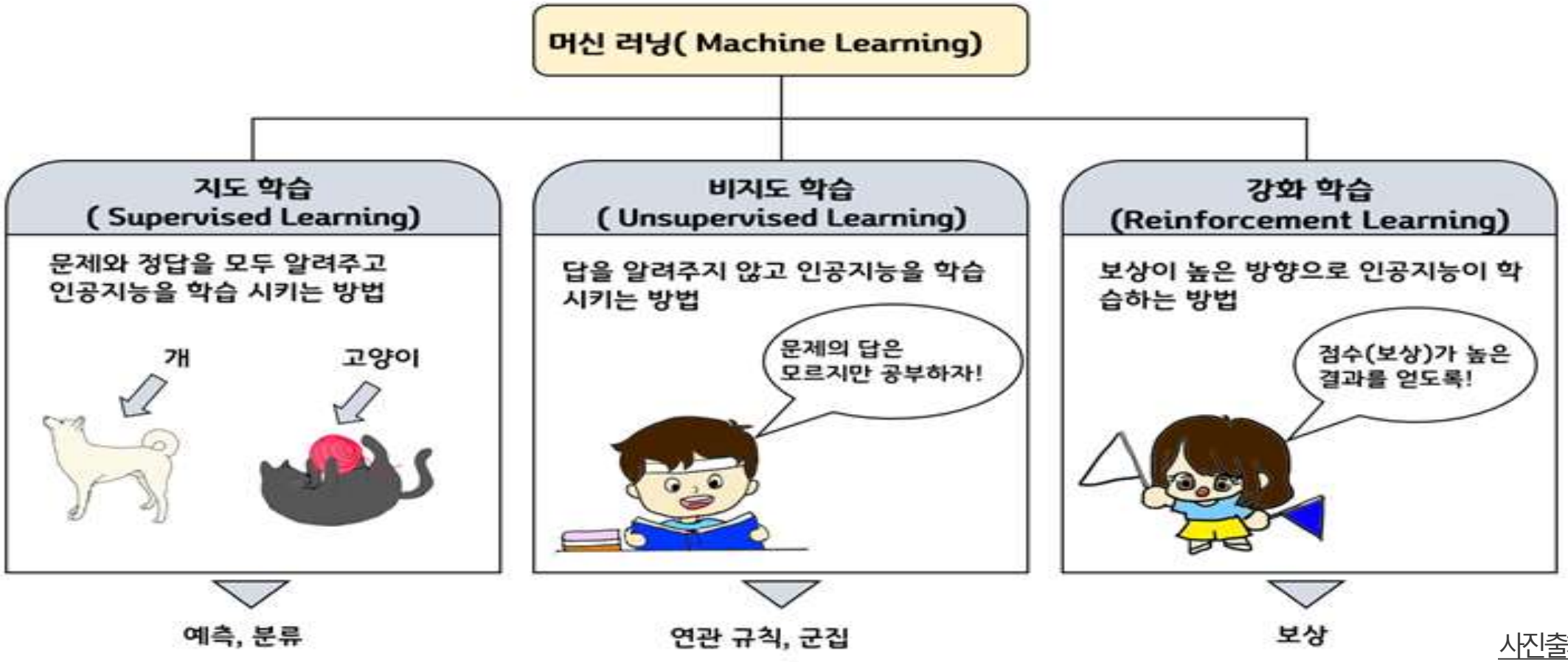
비정형 데이터를 처리하는 기술이 증가

비정형 데이터를 통해 인공지능이 학습하고,

예측, 분류 할 수 있는 기술 또한 발전

인공지능 종류

-데이터의 정답을 주고 학습하는 것인지에 대한 유무



사진출처:

인공지능 종류

-학습방법에 대한 분류

머신러닝

학습데이터를 독립변수(원인변수)로 놓고
맞추고자 하는 데이터(칼럼)를
종속변수(결과변수)로 설정한다.

↓

학습데이터로 학습한 뒤, 실제데이터를 주고
예측 및 분류 작업을 한다.

딥러닝

머신러닝의 한 종류로 컴퓨터가 사람처럼 사
고할 수 없을 까라는
아이디어에서 시작

↓

입력층, 은닉층, 출력층으로
이루어져 있음.
사람의 뉴런처럼 컴퓨터에서는
퍼셉트론이 이 역할을 함.

↓

신경망을 구축하여 예측 및 분류
작업을 진행

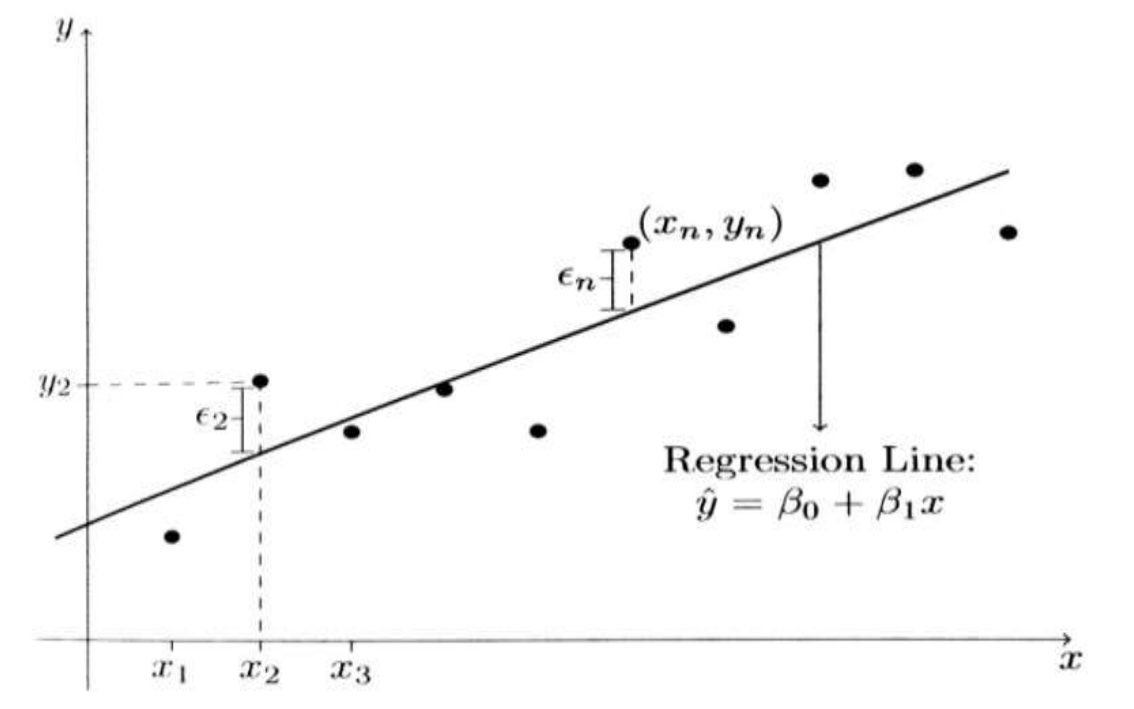
강화학습

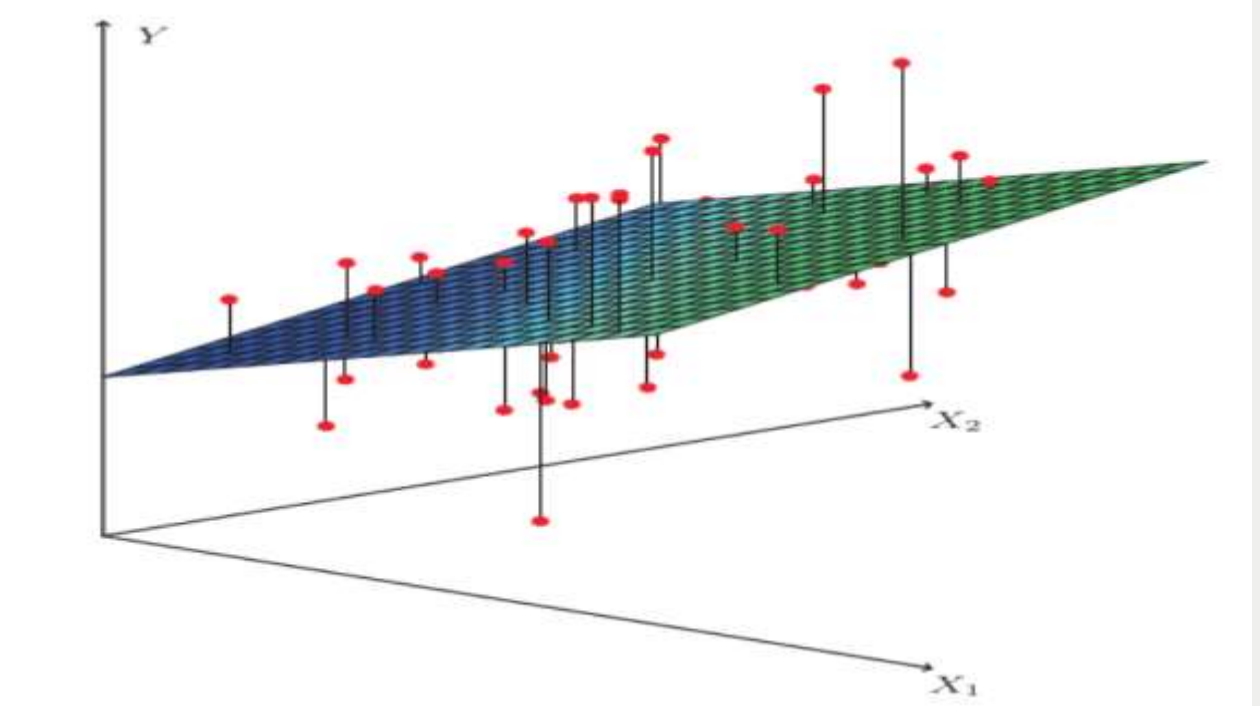
아이를 훈육할 때와 마찬가지로 어떤
사건이 발생했을 때, 보상 혹은 페널티
를 부여하면서 학습을 진행

↓

보상의 기댓값을 극대화 시키는 것을
목표로 학습을 진행함

머신러닝-회귀분석





- 회귀식: $Y = \beta_0 + \beta_1 X + \epsilon$
- 설명: 독립변수와 종속변수가 1개씩일 때, 이 둘 사이의 인과관계를 분석하는 것으로, 두 변수의 관계가 선형이다.

1. 다중회귀분석

- 회귀식: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$
- 설명: 독립변수가 2개 이상이고 종속변수가 하나일 때 사용 가능한 회귀분석으로, 독립변수와 종속변수의 관계가 선형으로 표현된다. 단순회귀분석이 확장된 형태이다.

머신러닝–회귀분석 예시

-보스턴 집값 예측 문제

범죄율	재산세율	주택당 방수	…	인구중 하위 계층 비율	집값

범죄율부터 인구 중하위 계층 비율까지

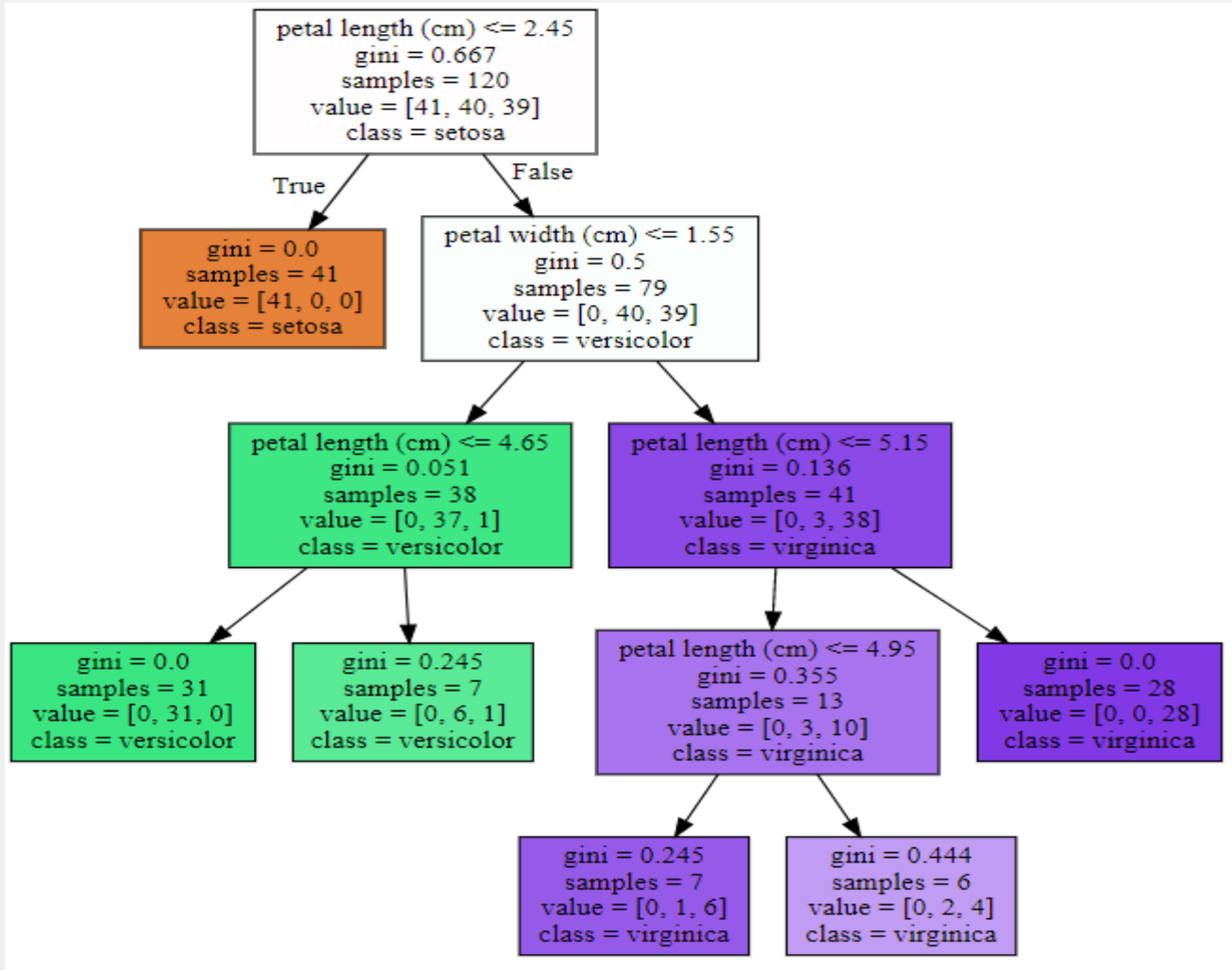
전부 독립 변수로 입력한 뒤

$$Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta kXk+\epsilon$$

다중 선형 회귀분석 식을 통해

집값(Y)을 예측한다.

머신러닝-분류분석



좌측 사진은 붓꽃 데이터를 보고 붓꽃의 종류를 맞추는 과정에서 머신러닝이 분류하는 과정을 시각화 한 그림이다.

분류분석은 일종의 스무고개 게임으로 생각하면 편하다.

카테고리컬한 독립변수들을 대상으로 일종의 기준을 두고 맞는지 아닌 지에 따라 분류하여 최종 종속변수를 맞춘다.

머신러닝–분류분석 예시

-타이타닉생존여부 분류문제

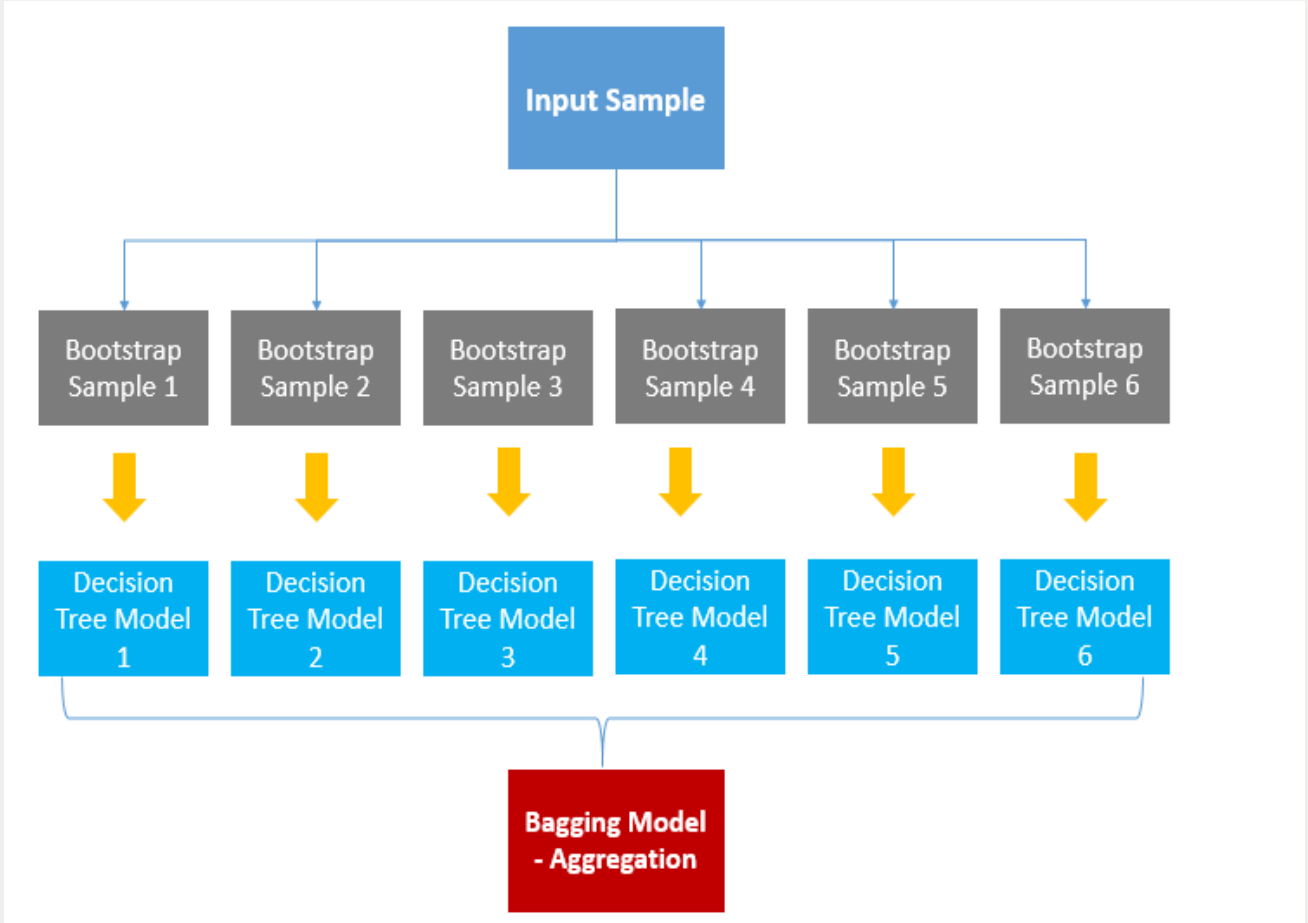
티켓 등급	성별	객실 번호	…	탑승장소	생존유무

타겟 데이터가 0과 1일 경우 이진분류,

0~9 처럼 여러 개일 경우 다진분류 문제로 바라 볼 수 있다.

회귀분석과 달리 독립변수들이 Categorical 한 특성을 가지고 있다.

머신러닝- 앙상블 학습



여러 개의 결정 트리를 결합하여 하나의 결정 트리보다 더 좋은 성능을 내는 머신러닝 기법

대표적으로 독립적으로 각각의 모델을 학습시켜 집계하는 배깅과 각 모델에 가중치를 부여하여 모델 성능을 개선하는부스팅이 있다.

빅데이터, 인공지능, 데이터 과학 개요

머신러닝 - Overfitting

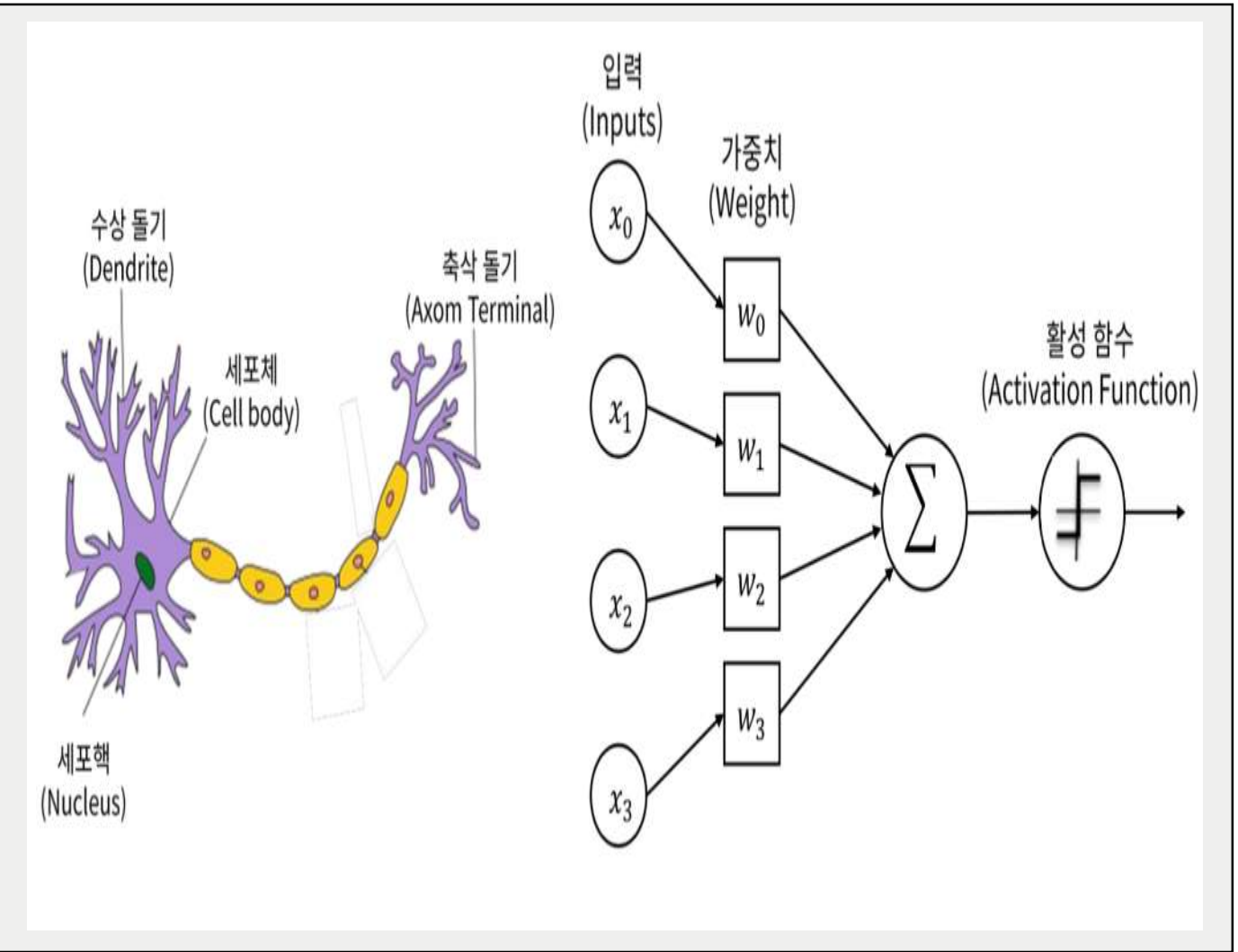
모의고사 성적 실제 시험 성적

과대적합(Overfitting)은 학습 데이터의 일반적인 특성을 학습 하는 것이 아니라 학습 데이터 자체를 학습해버리면서, 실제 환경에서 오차가 증가해버리는 현상

쉬운 예시로 설명

문제집을 보고 내용을 공부해야 실제 시험에서 점수가 잘 나오는데, 문제집의 문제와 답 번호만 외우고 시험 보러 가는 것이라고 보면 된다.

딥러닝-퍼셉트론

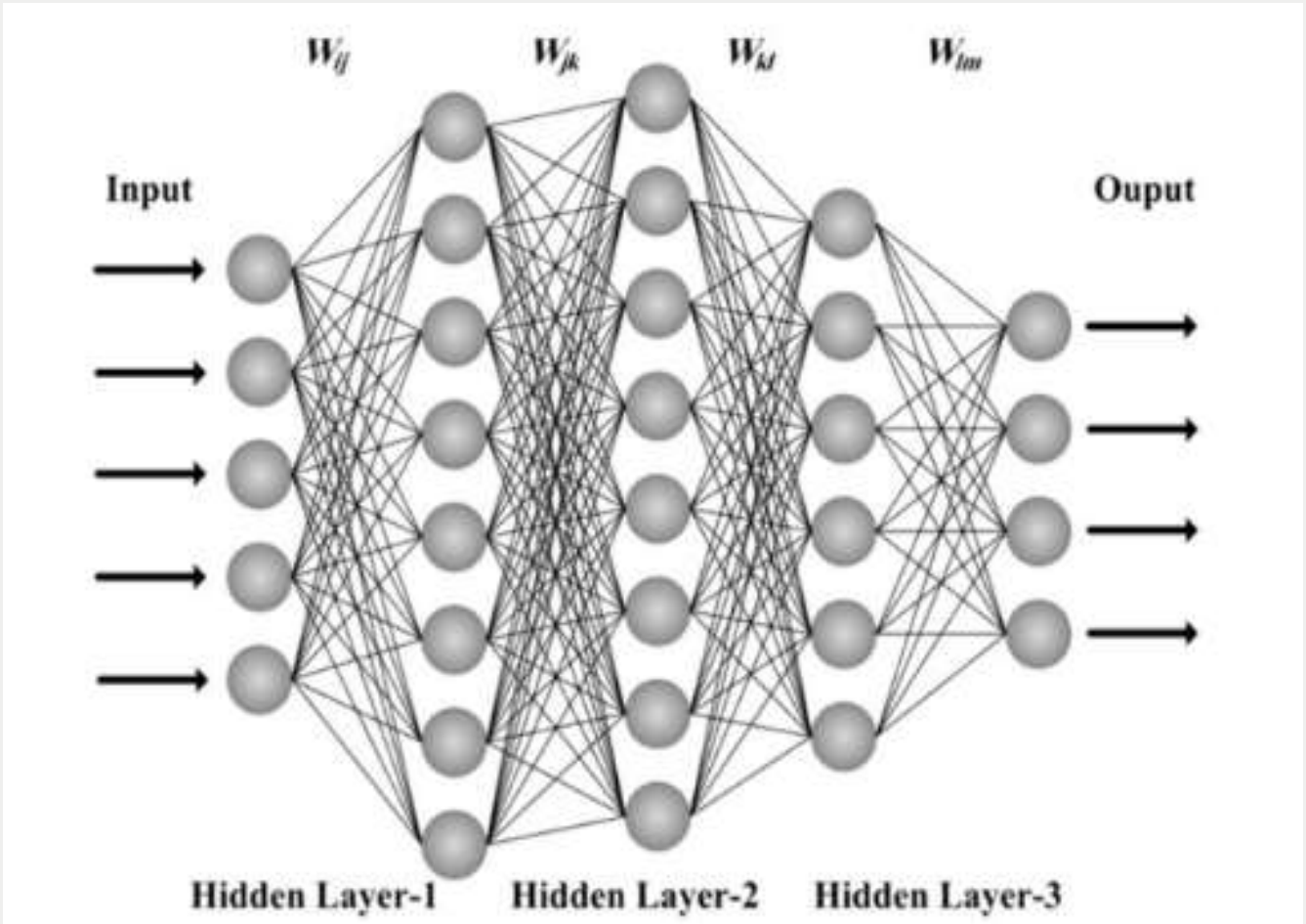


사람은 약 천 억개의 뉴런을 가지고 있으며, 천 억개의 뉴런들이 서로 신호를 주고 받는다.

딥러닝은 사람의 뉴런처럼 퍼셉트론이라는 개념으로 컴퓨터에 신경망을 구축하여 예측 및 분류 작업을 하는 학습 방법이다.

각 퍼셉트론에 어떤 입력값이 들어오면 가중치를 곱하고 활성함수를 통해 0 혹은 1로 output을 내보낸다.

딥러닝

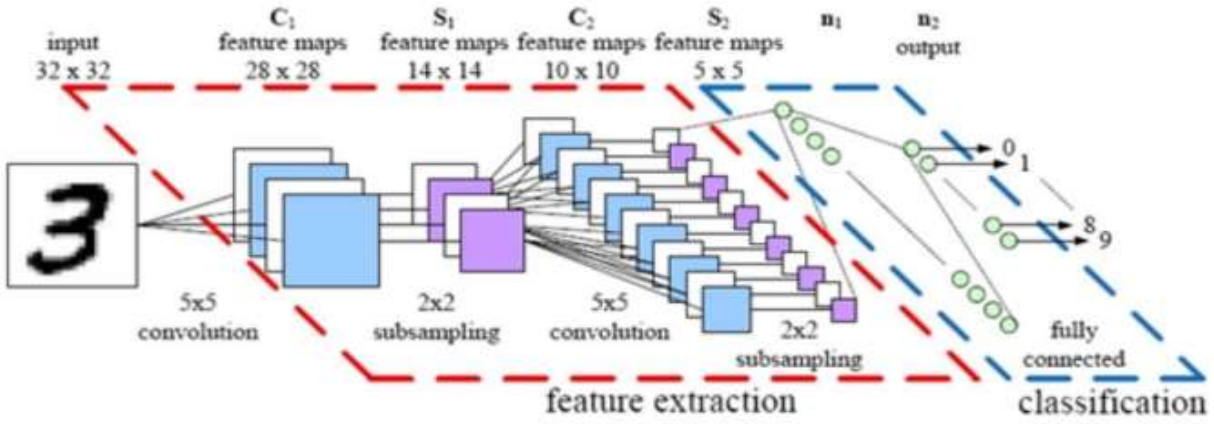


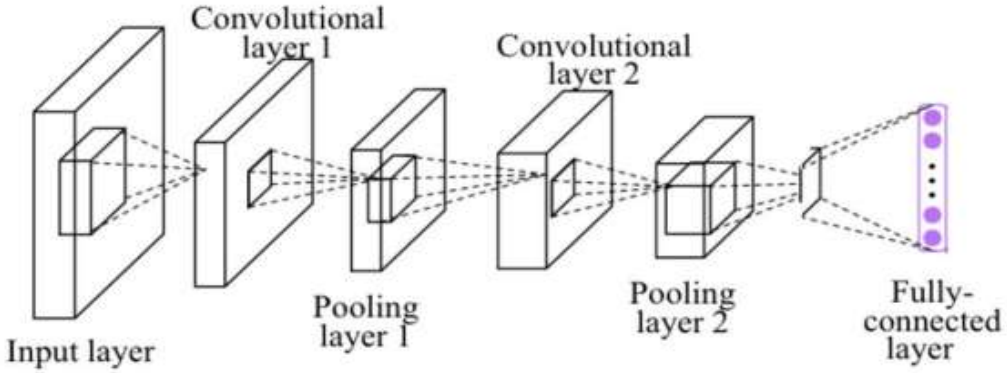
이러한 퍼셉트론이 다층으로 구성되어 있을 때, 이를 딥러닝이라고 한다.

입력층, 은닉층, 출력층으로 구성되어 있으며 각 은닉층을 통해 Input 값에 가중치를 곱하여 output을 출력한다.

딥러닝 – CNN(합성곱 신경망)

Simple CNN





airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

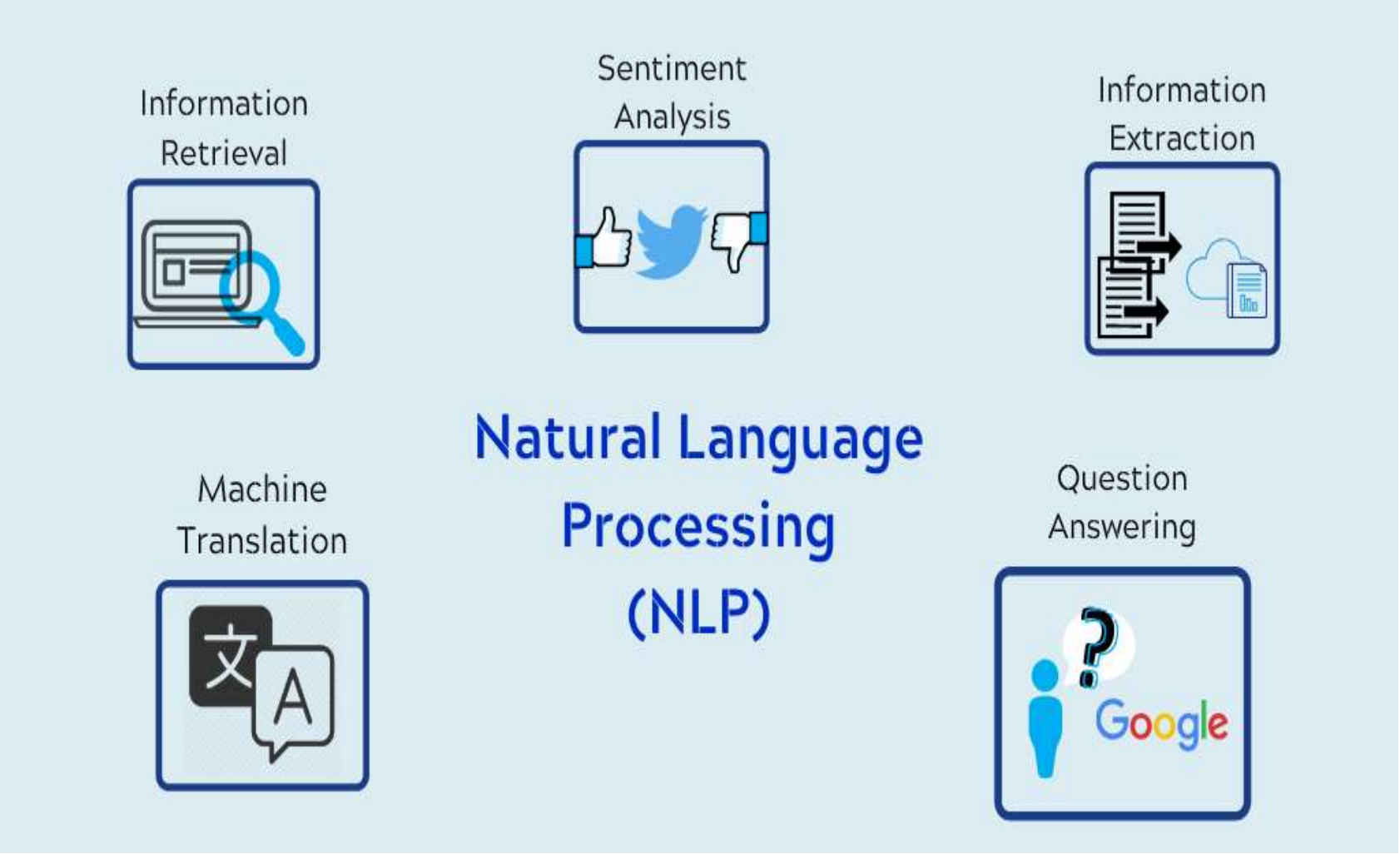
딥러닝 – Computer Vision

CNN이 발달하면서, 한 이미지에서
객체와 그 경계 상자를 탐지하는
알고리즘 분야를

컴퓨터 비전 분야라고 한다.

주로 이미지나 영상을 분석하는
인공지능의 한 분야이다.

딥러닝–NLP



인간의 언어를 컴퓨터에게 학습시켜서

인간의 언어를 토대로 번역, 텍스트분석, 감성분석 등을 수행하는

인공지능의 분야를 NLP라고 한다.

빅데이터, 인공지능, 데이터 과학 개요

인공지능 종류 - 딥러닝

1. 다음은 한국 사람이 쓴 리뷰다. 어떤 내용인가?

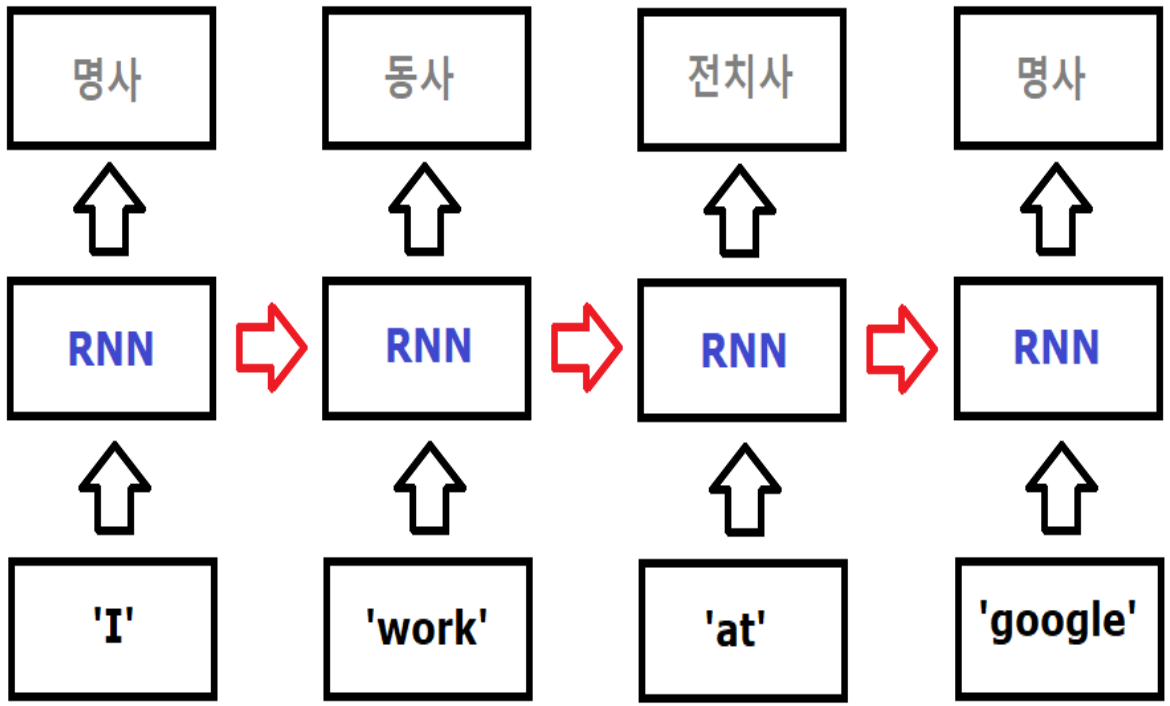
이 꼈쑫뚲하욼쑦 절떆 깨찌맊요 띠어렵꼮 뿔찐쬴해니꺆 깨찌뫄쎘욬 깨삐싺꼬 쮠짱이 끼에쏚 뤼븁깡요혜썼욦

우리는 어떻게 이 리뷰를 해석 할 수 있었을까?

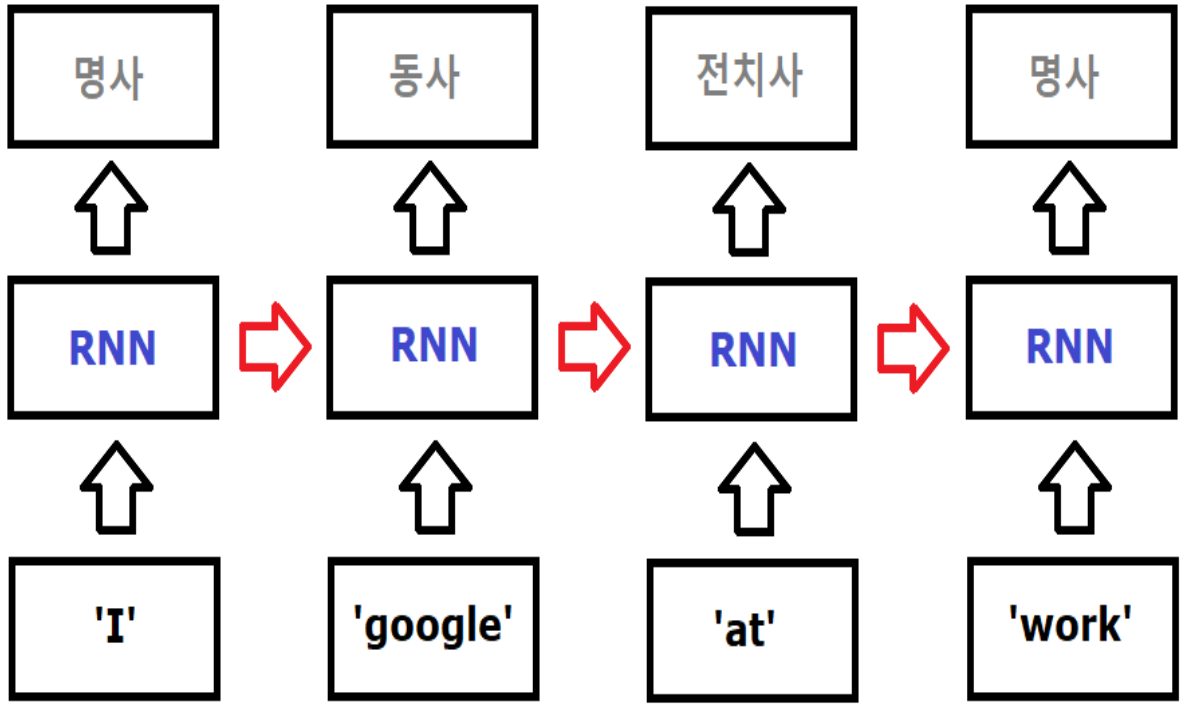
이 게스트하우스 절대 가지마요 더럽고 불친절하니까 가지마세요 개비싸고 주인장이 계속 리뷰강요했어요

딥러닝 – RNN(순환 신경망)

언어는 단어의 순서에 따라 의미가 달라지며, 뒷 단어가 앞 단어에 영향을 받는다.

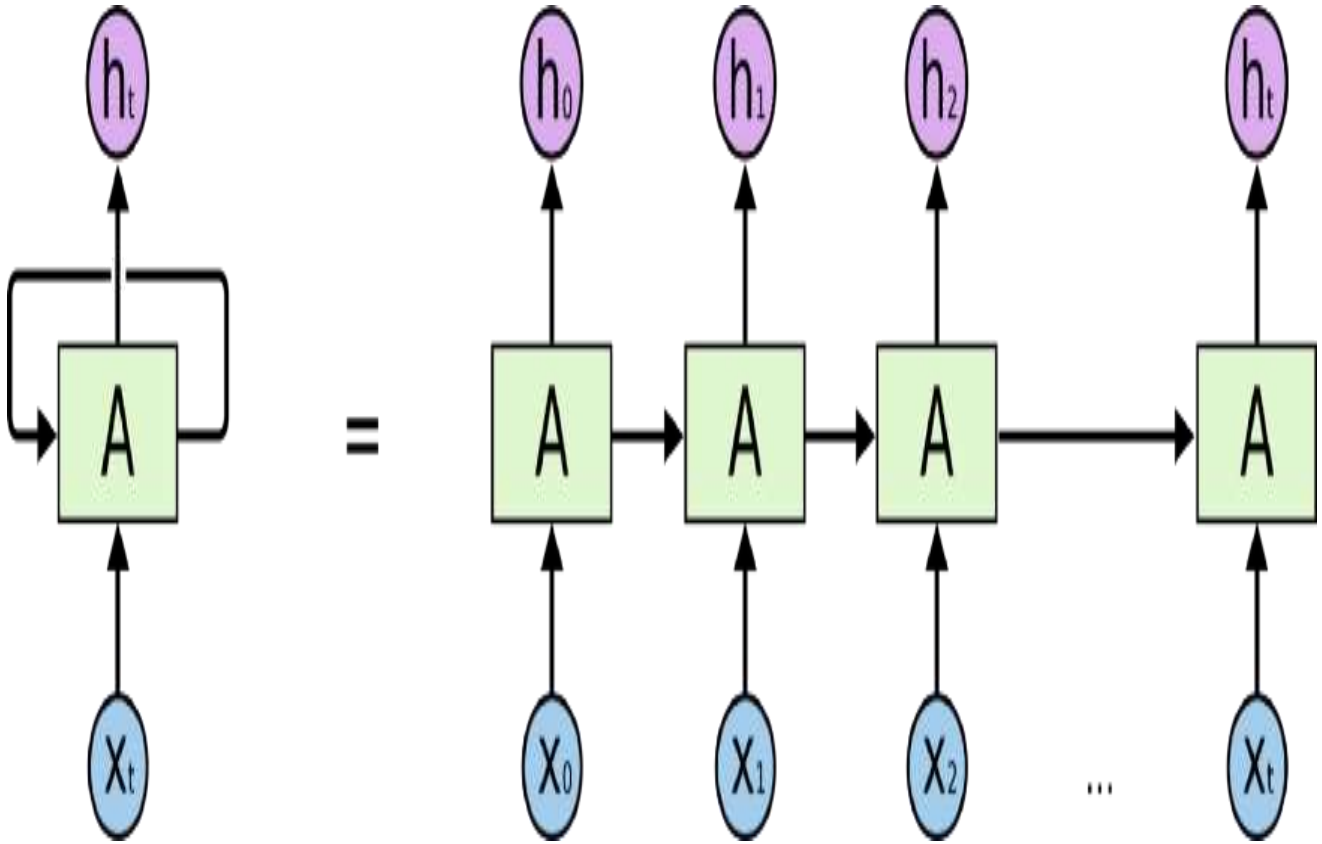


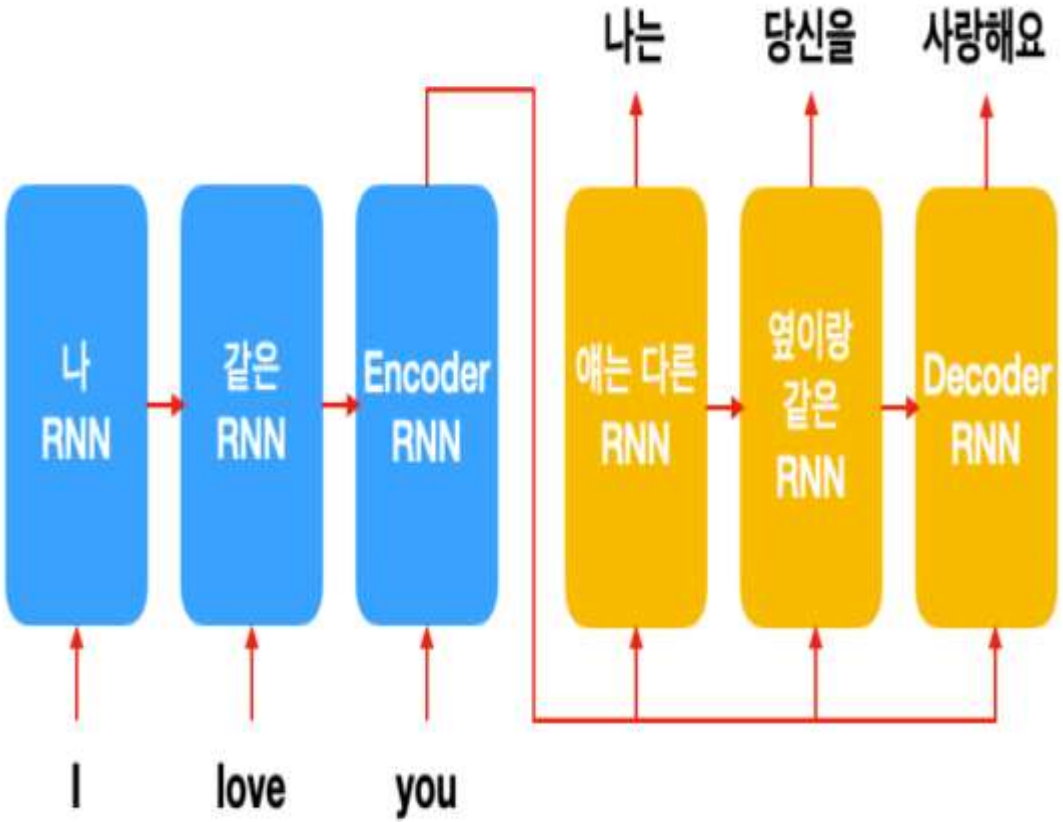
나는 구글에서 일한다



나는 직장에서 구글링 한다

딥러닝 – RNN(순환 신경망)



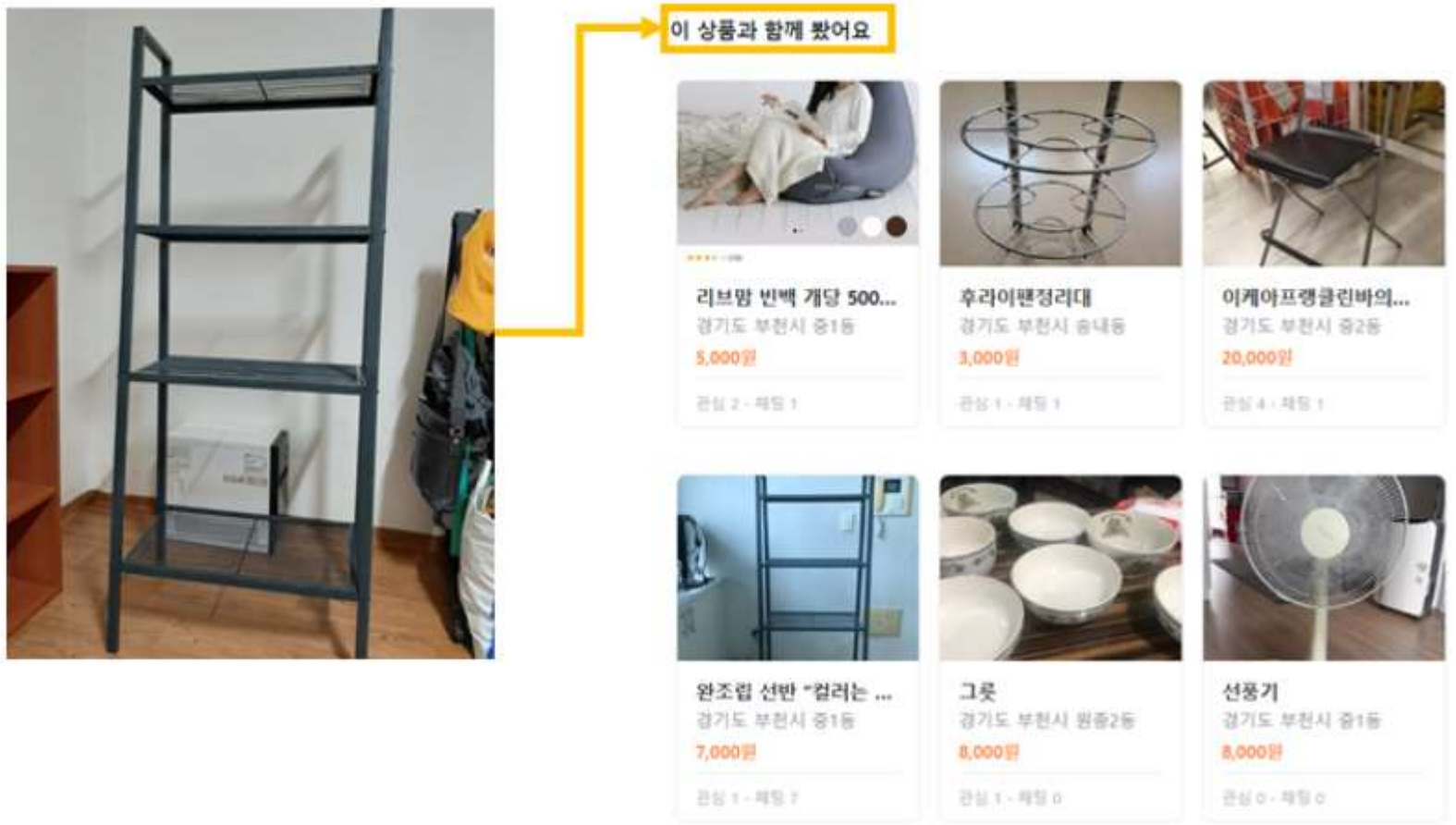



we're
doomed

빅데이터, 인공지능, 데이터 과학 개요

인공지능 종류 - 추천시스템

딥러닝-추천시스템



추천시스템은

사용자들의 행동 정보를 바탕으로

데이터들을 행렬로 변환한 뒤,

거리를 계산하여 유사도를 측정한 뒤,

이를 바탕으로 관련 상품을 추천하는

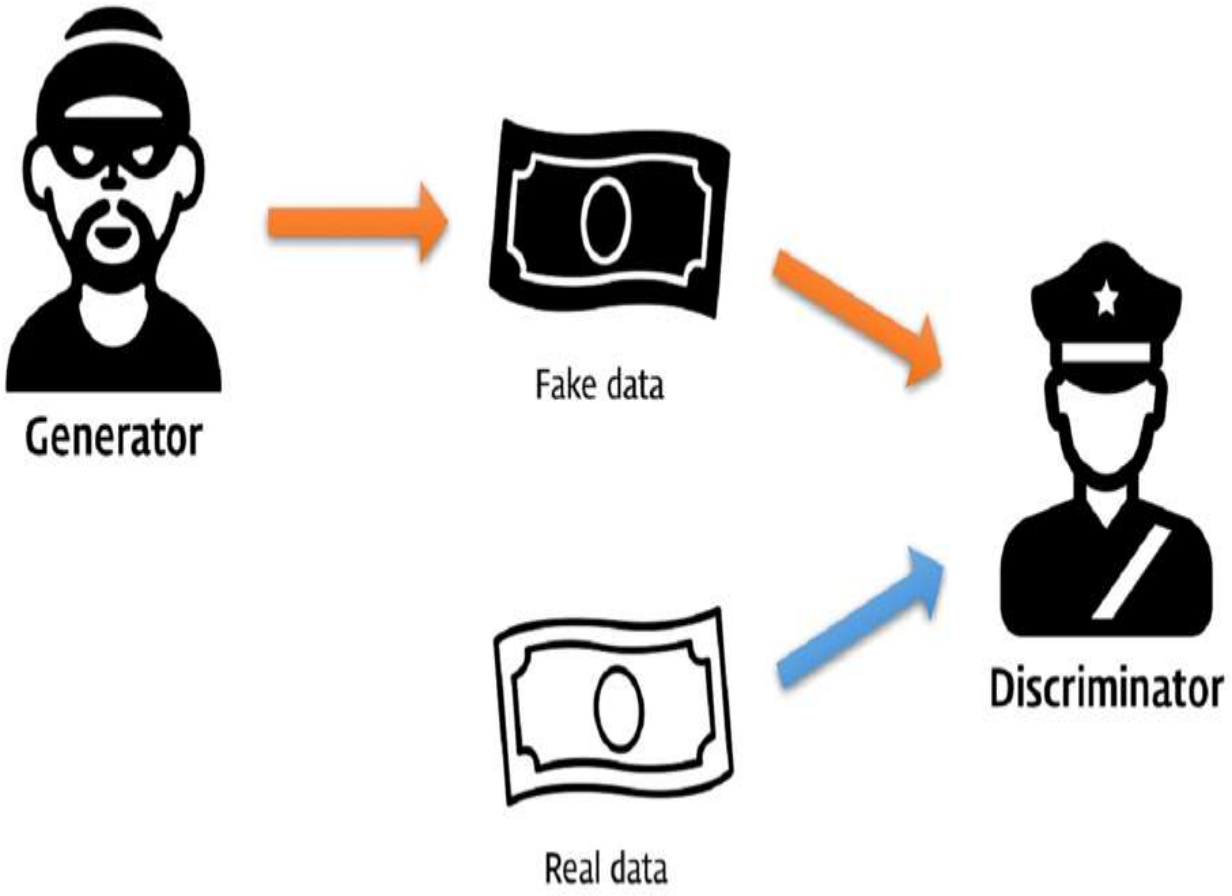
인공지능의 한 분야이다.

강화학습

강화학습은사전지식이없어도시행착오를통해학습하는방법으로

다른머신러닝분야와는다르게순차적으로행동을결정해야하는문제를다루는인공지능의한분야이다.

적대적 생성 신경망(GAN)



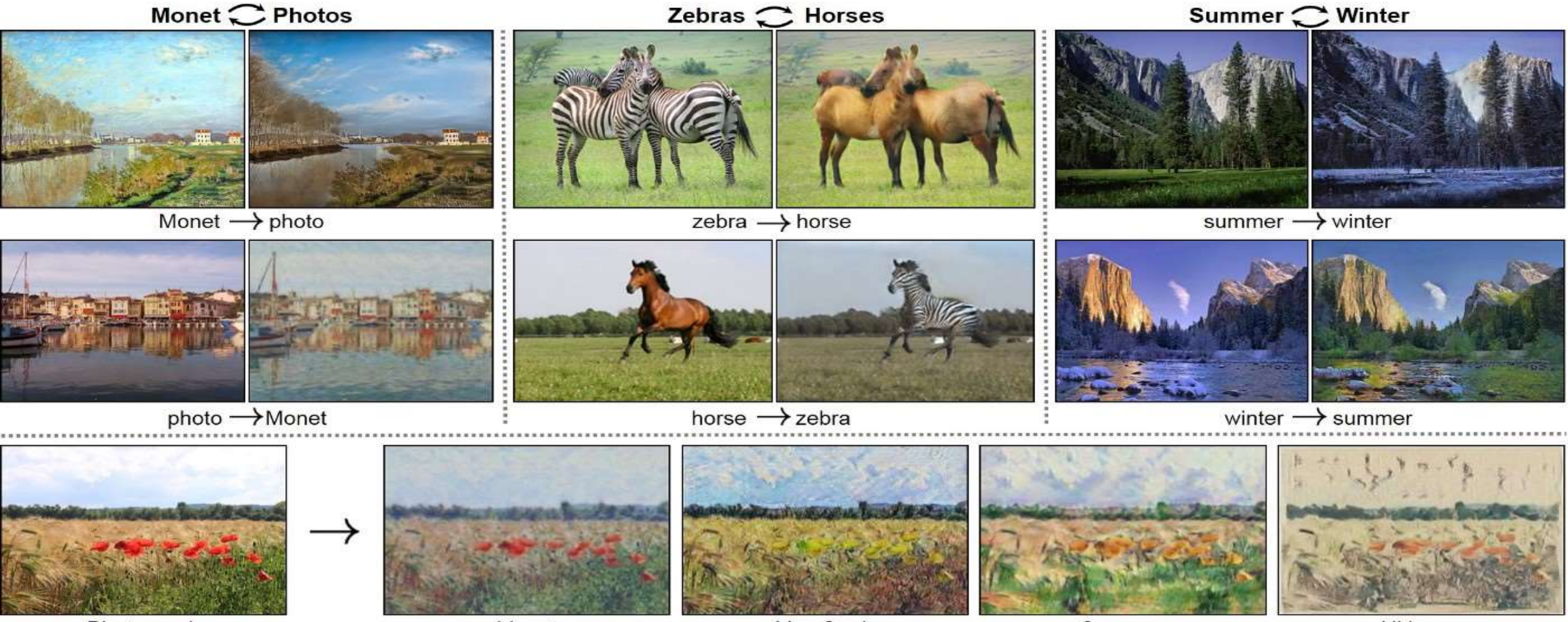
적대적 생성 신경망은 생성기 신경망과 판별기 신경망으로 구성되어 있다.

생성기 신경망: 가짜 데이터를 진짜로 속여보자!

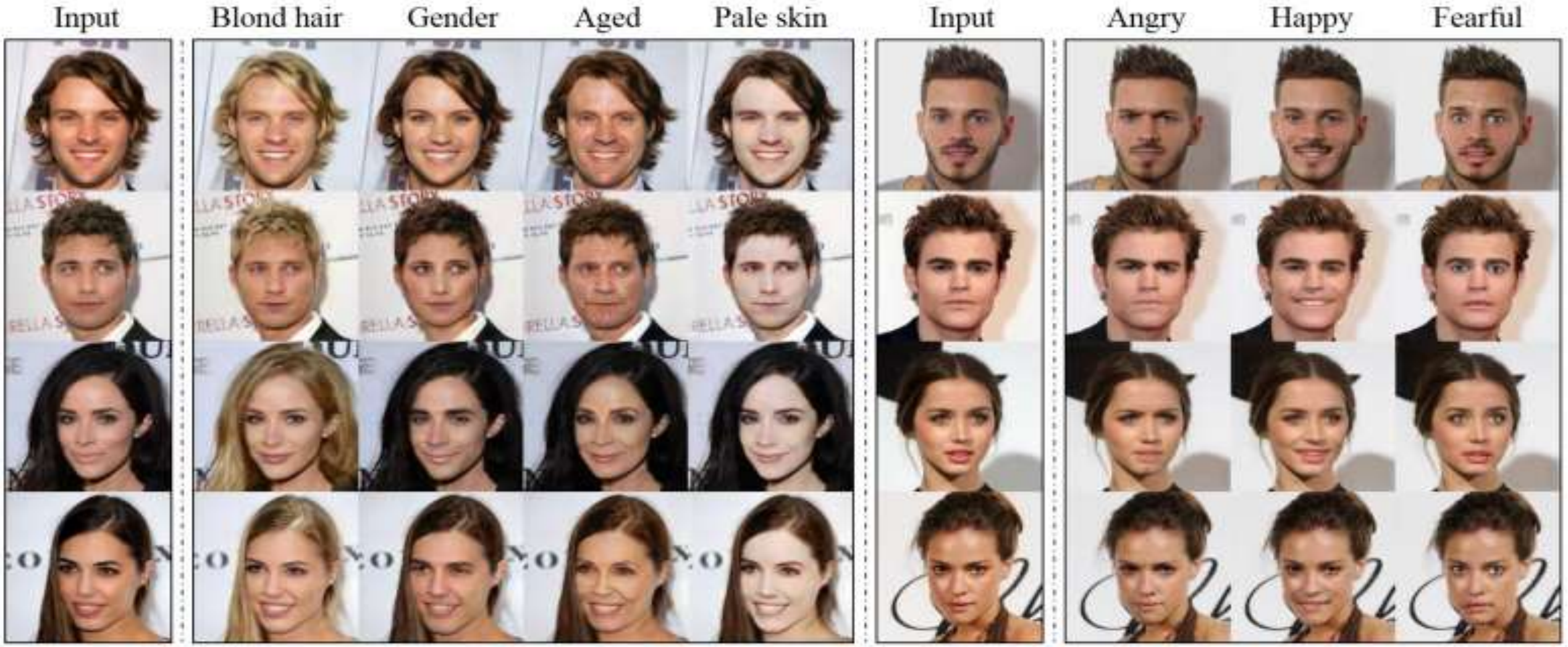
판별기 신경망: 가짜 데이터를 찾자!

생성기와 판별기를 서로 경쟁하며 발전시킨다.

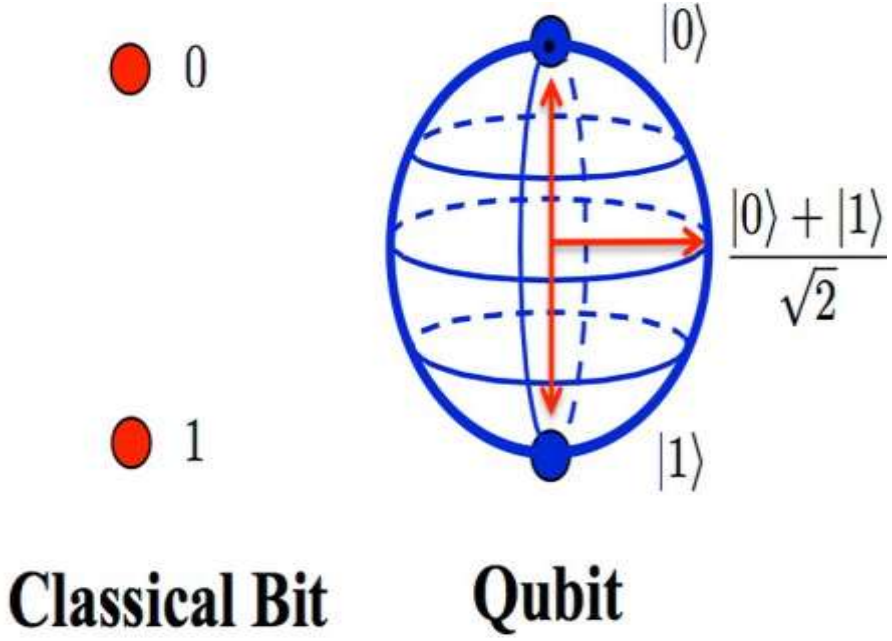
적대적 생성 신경망(GAN) 활용 사례-1



적대적 생성 신경망(GAN) 활용 사례-2



양자컴퓨터



기존 컴퓨터 vs 양자 컴퓨터

구분	기존 컴퓨터	양자컴퓨터
연산 개념도	0 1 정보를 0이나 1로 표현 0 1 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 1 1 1 0 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1개씩 순차적으로 계산 많은 시간 소요	0 1 0과 1을 중첩 0 1 0 0 1 0 0 1 1 1 0 1 0 1 0 0 1 1 0 1 1 0 0 1 0 0 합쳐서 한번에 계산 순식간에 계산
기본 단위	Bit(0 또는 1)	Qubit
연산 방법	논리 표에 의한 계산	행렬 함수에 의한 계산

양자 컴퓨터 상에서는 0과 1의 비트 체계가 아닌 Qubit라는 비트를 사용

구글은 2019년 초전도 방식의 54큐비트 양자컴퓨터 '시커모어'를 통해 슈퍼컴퓨터가 1만년 동안 수행해야하는 연산을 200초 만에 해결했다는 내용의 논문이 발표됐다.

빅데이터, 인공지능, 데이터 과학 개요

추가자료 : 머신러닝 관련 추가 자료

Machine Learning 정리 자료 클릭

Machine_Learning 정리

Supervised Learning (지도 학습)

Regression Analysis(회귀분석)

회귀분석은 통계학에서 관찰된 연속형 변수들에 대해 독립변수와 종속변수 사이의 인과관계에 따른 수학적 모델인 선형적 관계식을 구하여 어떤 독립변수가 주어졌을 때 이에 따른 종속 변수를 예측한다. 또한 이 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석방법이다.

회귀분석의 가정

> **선형성**

- 독립변수가 변화할 때 종속변수가 일정한 크기로 변화한다면 선형성을 만족한다고 볼 수 있다.
- 산점도를 통해 확인 할 수 있다.

빅데이터, 인공지능, 데이터 과학 개요

추가 자료 : 강화학습 관련 추가 자료

Reinforcement Learning 정리 자료 클릭

강화학습이란?

Introduce Reinforce Learning

강화학습 기초 1편부터 FINAL 편까지 총 4편으로 구성되어 있음.

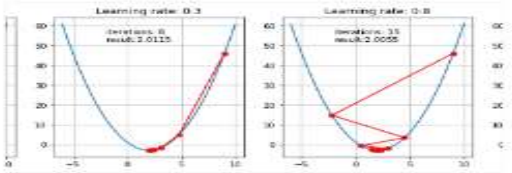
-파이썬과 케라스로 배우는 강화학습 책 기반으로 만들어진 자료

빅데이터, 인공지능, 데이터 과학 개요

추가 자료 : 키워드로 보는 데이터 분석

데이터분석 핵심 키워드 포스팅 자료 (velog.io/@yunyoseob)

1. Gradient Descent : 경사하강법



키워드로 공부하는 데이터 분석 첫 번째 포스팅은 경사하강법입니다.경사하강법이란 함수의 기울기(경사)를 구하여 기울기가 낮은 쪽으로 계속 이동시켜서...

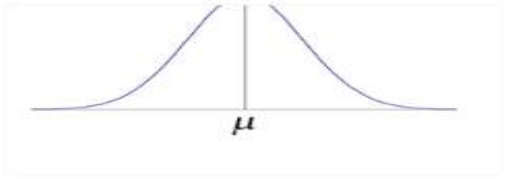
2022년 2월 1일

2. Gradient Vanishing : 기울기 소실

오늘 포스팅할 2번째 키워드로 공부하는 데이터 분석의 키워드는 Gradient Vanishing입니다.퍼셉트론이란 입력층과 출력층으로만 구성된 최초의 인공신...

2022년 2월 2일

3. Gaussian Distribution : 정규분포



오늘 포스팅할 3번째 키워드로 공부하는 데이터 분석의 키워드는 Gaussian Distribution(정규분포)입니다. 사진 출처 : 엣지있는 인공지능 : 정규분포정규...

2022년 2월 7일

Q.

데이터 사이언스이란?

1 데이터 사이언스 정의

2 데이터 사이언스 역할

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

데이터 사이언스 정의

데이터 사이언스이란?

데이터로부터 의미 있는 정보를 추출해내는 학문이다

- 2021 빅데이터 분석기사 필기

데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.

- 위키백과“데이터 사이언스 "

데이터 사이언스는 “데이터 공학, 수학, 통계학 , 컴퓨터 공학, 시각화, 해커의 사고방식, 해당분야의 전문 지식을 종합한 학문“ 이다.

- 위키피디아

데이터 과학과 빅데이터, 인공지능은

엄밀히 말하자면 서로 다르지만 보통

혼용해서 사용하기도 합니다.

Part 1,

빅데이터, 인공지능, 데이터 과학 개요

데이터 사이언스 정의

데이터 마이닝 vs 데이터 사이언스

데이터 마이닝은 주로 분석에 포커스를 두지만,

데이터 사이언스는 분석 뿐만 아니라, 이를

효과적으로 구현하고 전달하는 과정, 궁극적으로

전략적 인사이트 도출을 위한 일련의 행위까지

모두 포괄하는 광의의 개념입니다.

데이터 사이언스에 대한 역할

데이터 사이언스는 전략적 통찰을 추구하고

비즈니스 핵심 이슈에 답하고, 사업의 성과를

견인해 나갈 수 있어야 한다.

Part 2,

빅데이터 프로젝트와 빅데이터 관련 직업 소개

Part 2,

빅데이터 프로젝트

빅데이터 프로젝트 단계

수집 및 저장

로그수집기, 크롤링 ,센싱, RSS 리더 , OpenAPI 등을 이용하여 데이터 수집

분산파일 시스템, NoSQL, 병렬 DBMS 등을 이용해 대용량 데이터를 저장

>>

전처리

대용량 데이터를 처리하기 위한 분산 처리 기술
(하둡, H베이스, 맵리듀스 등)

>>

분석

파이썬, R 등을 이용하여 분석

데이터 분석에 있어서 탐색적 데이터 분석 (EDA), Feature Engineering, Model Selection 등의 작업을 진행함

>>

모델링

데이터 사전분석을 마친 뒤, 본격적으로 튜닝을 시작하는 시기, Hyper Parameter 튜닝 등의 작업을 하며, 모델을 고도화, 분석 모형을 평가 및 개선하는 작업을 수행

>>

시각화 및 서비스

분석 결과를 시각화 하고 서비스를 개시하는 단계

Part 2,

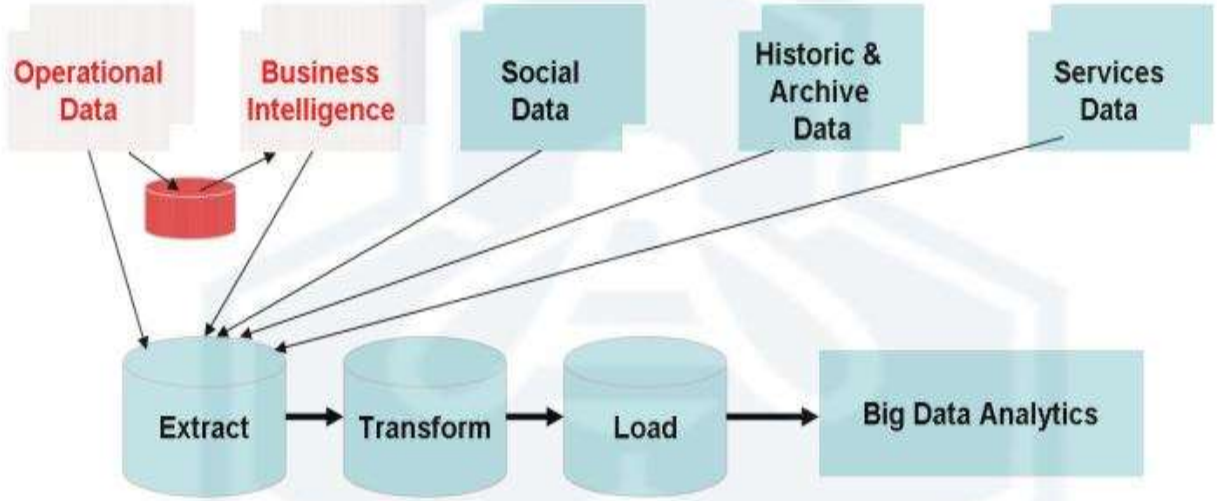
빅데이터 프로젝트

수집 및 저장, 전처리

실시간 처리 vs 배치 처리

빅데이터 실시간 적용 사례

Big Data Batch Processing



Traditional Systems use Proprietary Databases (Oracle, etc)

Big Data Systems use Open-source highly parallel systems (Hadoop, etc)

- Initial Indexing only by time
- Both techniques highly batch orientated
- Real-time or near real-time virtually impossible

Part 2,

빅데이터 프로젝트

수집 및 저장, 전처리

OLTP vs OLAP

구분	OLTP	OLAP
목적	비즈니스 활동 지원	비즈니스 활동에 대한 평가,분석
주 트랜잭션 형태	SELECT, INSERT, UPDATE, DELETE	SELECT
속도	수초 이내	수초 이상 수분 이내
데이터 표현 시간	실시간	과거
관리단위	테이블	분석된 정보
최적화 방법	트랜잭션 효율화, 무결성의 극대화	조회 속도, 정보의 가치, 편의성
데이터의 특성	트랜잭션 중심	정보 중심
예시	회원정보 수정	1년간의 주요 인기 트랜드
	상품주문	한달간의 항목별 수입, 지출
	댓글 남기기 및 수정	10년간 A회사의 직급별 임금 상승률

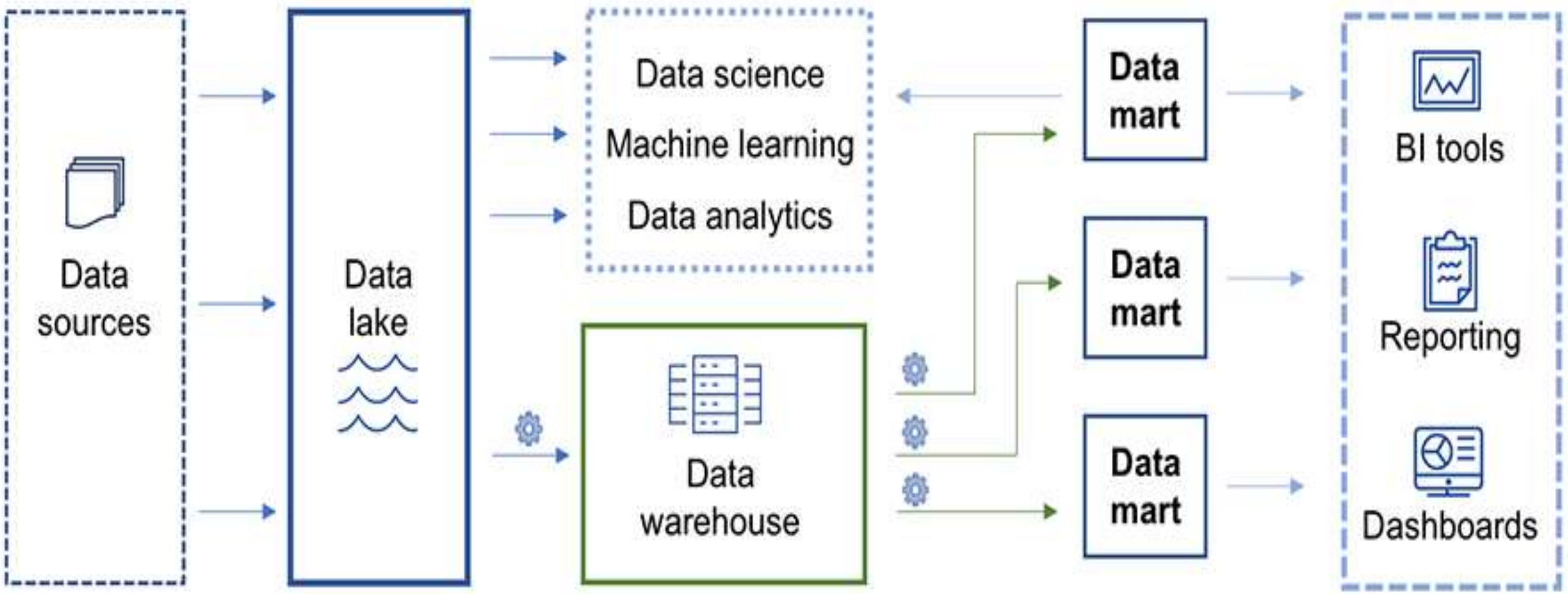
Data vs Information

Part 2,

빅데이터 프로젝트

수집 및 저장, 전처리

OLAP



Part 2,

빅데이터 프로젝트

수집 및 저장, 전처리

빅데이터 아키텍처에서 하둡 플랫폼과 카프카의 역할

Part 2,

빅데이터 프로젝트

수집 및 저장, 전처리

Hadoop

하둡(Hadoop)이란 대용량 데이터를 분산처리 할 수 있는 자바 기반의 오픈 소스 프레임 워크

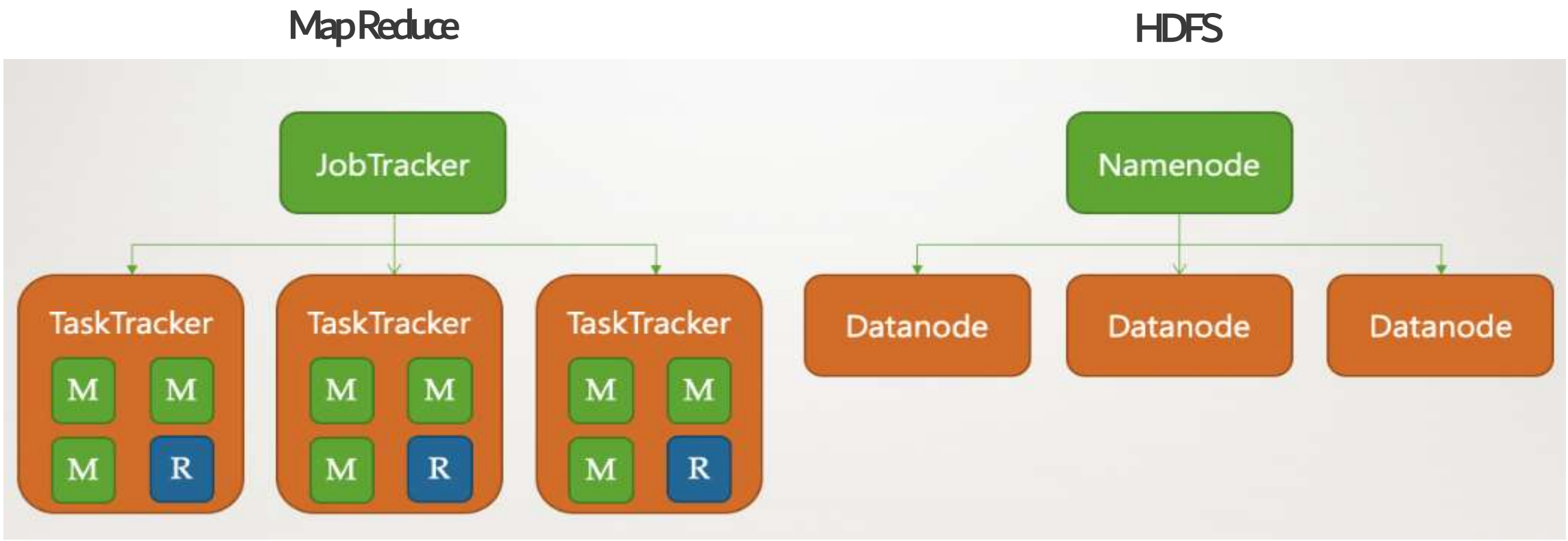
Hadoop 탄생 역사

년도	역사
2002	아파치 루씬(텍스트 검색 라이브러리)의 일환으로 너치 프로젝트 시작(오픈 소스 웹 검색 엔진)
2003	구글에서 실제로 운영되는 GFS라는 구글 분산 파일시스템 논문 출시
2004	NDFS라는 너치 분산 파일시스템을 오픈 소스로 구현
2004	구글에서 맵리듀스를 소개하는 논문 발표
2005	너치 개발자들이 너치 내부에 맵리듀스를 구현
2006	NDFS와 맵리듀스는 하둡이라는 이름으로 너치에서 독립하여 분리

빅데이터 프로젝트

수집 및 저장, 전처리

Hadoop v1의 SPOF(단일 고장점 문제)

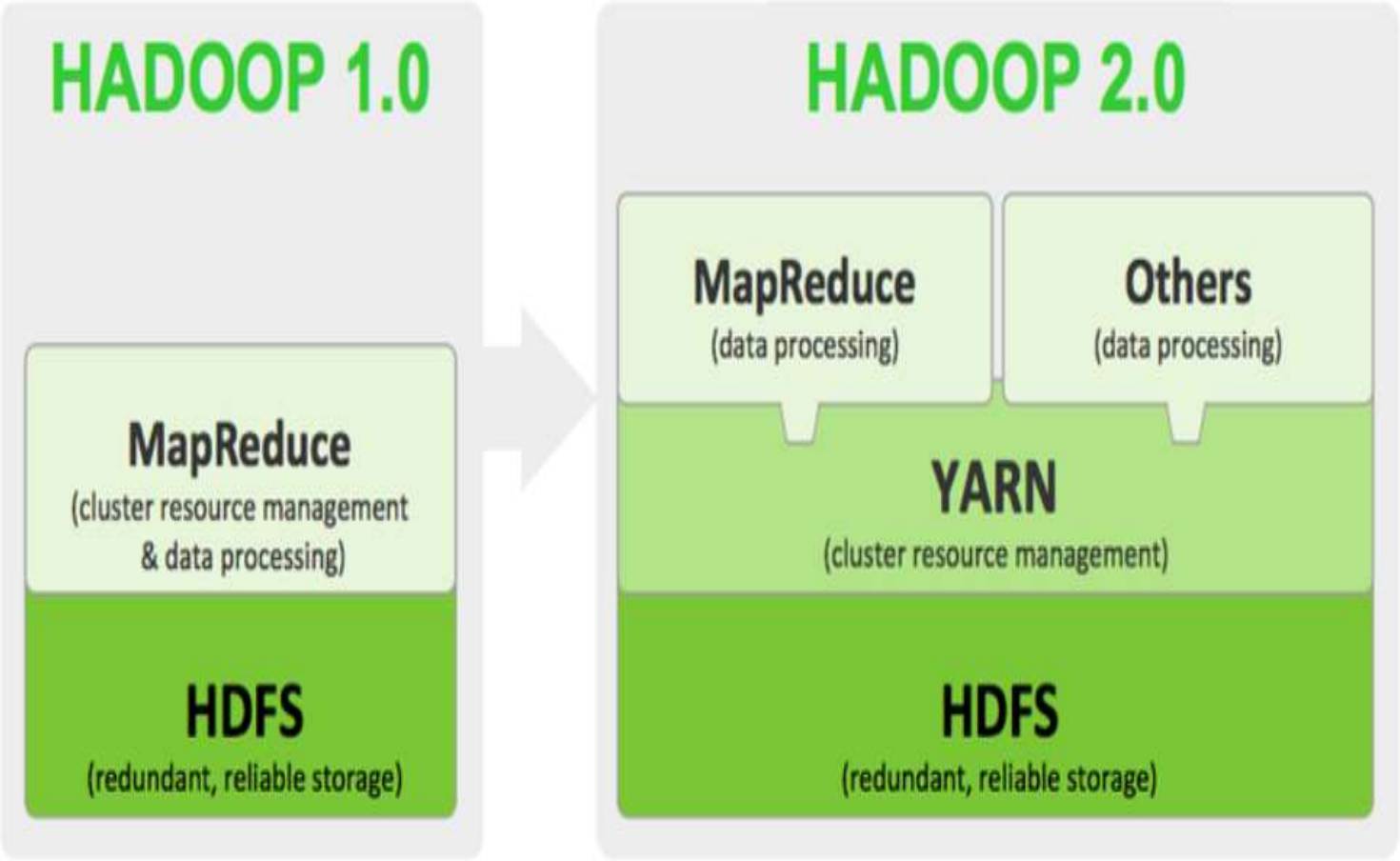


HDFS의 Namenode와 JobTracker가 문제가 생기는 순간 하둡 전체가 멈춰버리는 현상

빅데이터 프로젝트

수집 및 저장, 전처리

Hadoop v2의 YARN



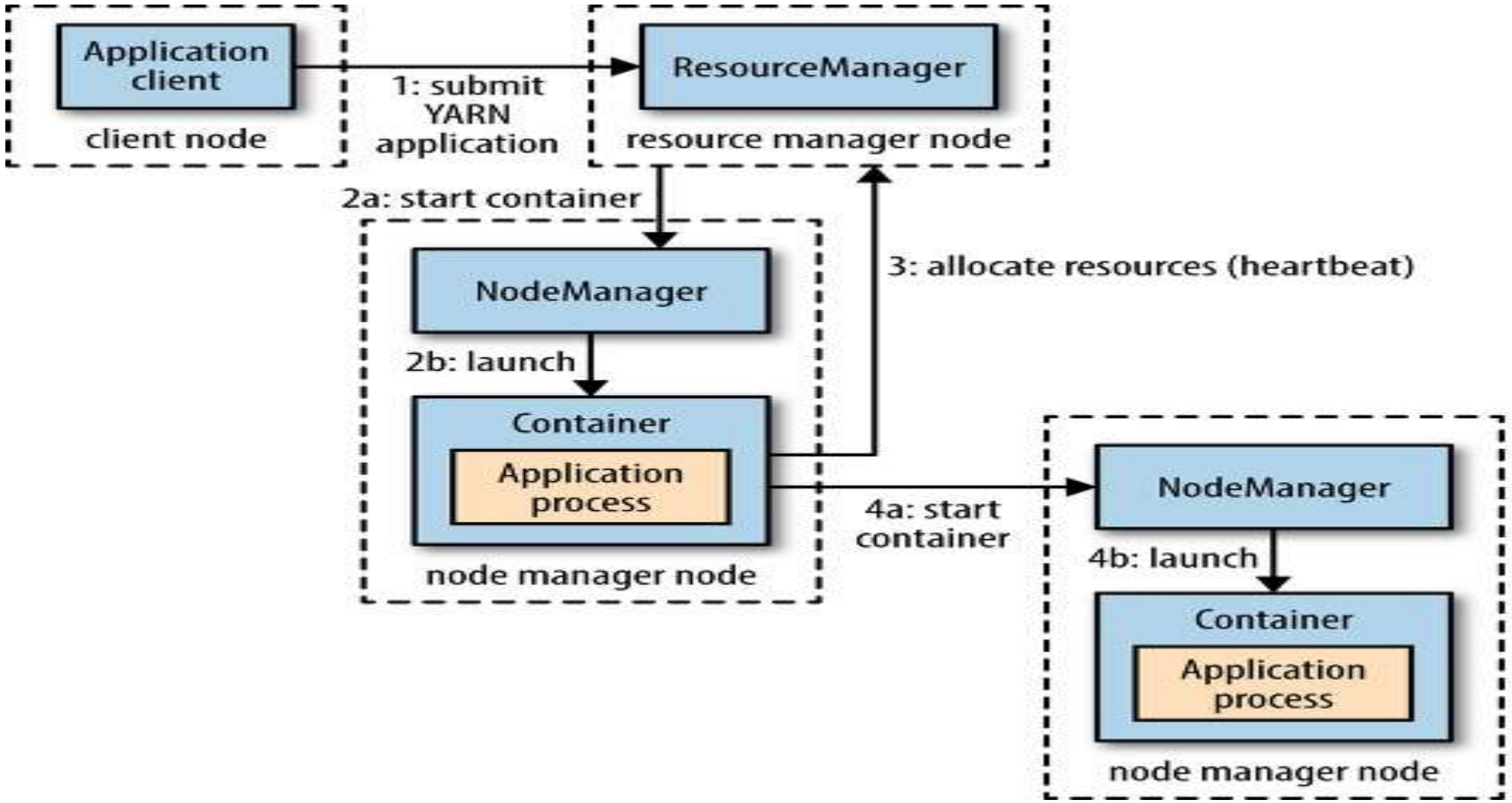
MapReduce의 단일 고장점 문제와 확장성 범위 문제를 해결하는 YARN이 탄생

YARN은 리소스 매니저와 노드 매니저로 나뉘어 다른 어플리케이션을 실행시킬 수 있다.

빅데이터 프로젝트

수집 및 저장, 전처리

YARN 구조



YARN 탄생 이후,

HDFS와 YARN을 기준으로 수많은 컴포넌트 들이 결합되면서 하나의 생태계 가구축되는데,

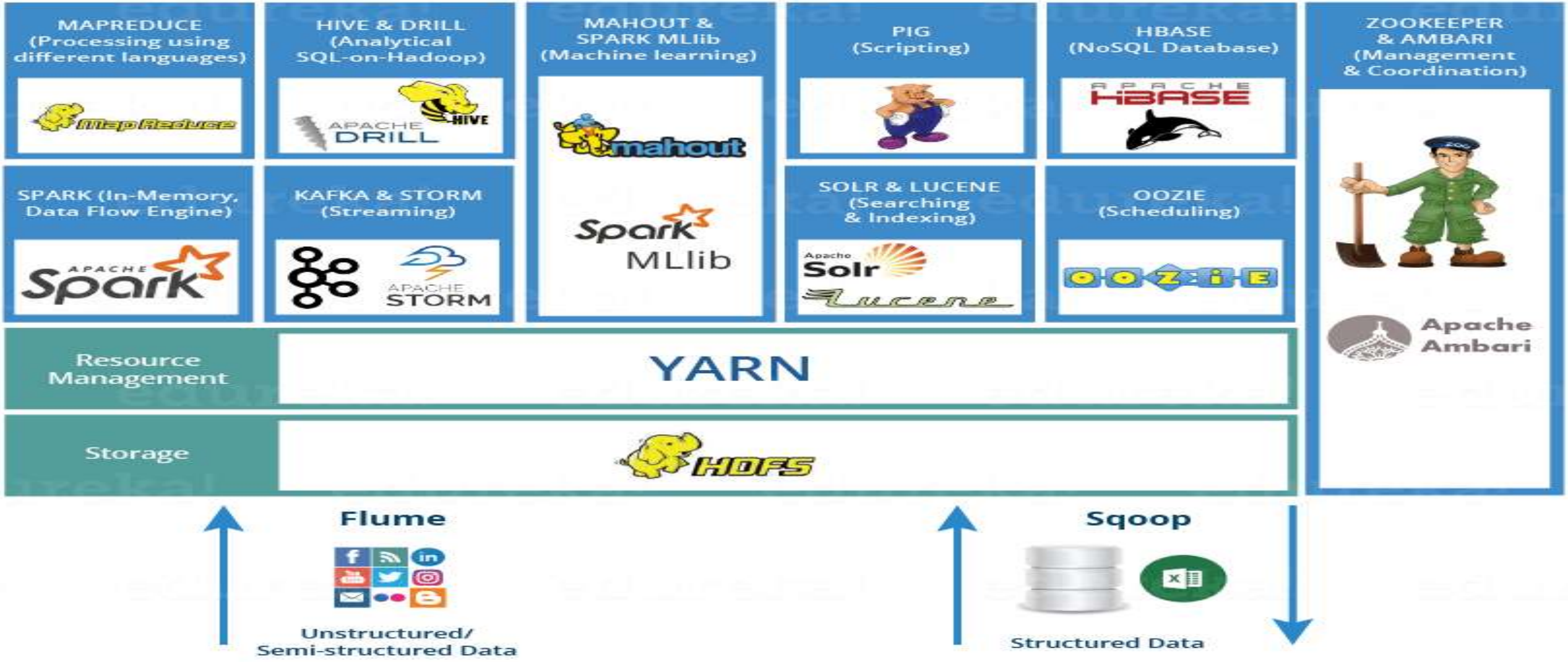
이 생태계를 하둡 에코 시스템이라고함

Part 2,

빅데이터 프로젝트

수집 및 저장, 전처리

Hadoop Ecosystem

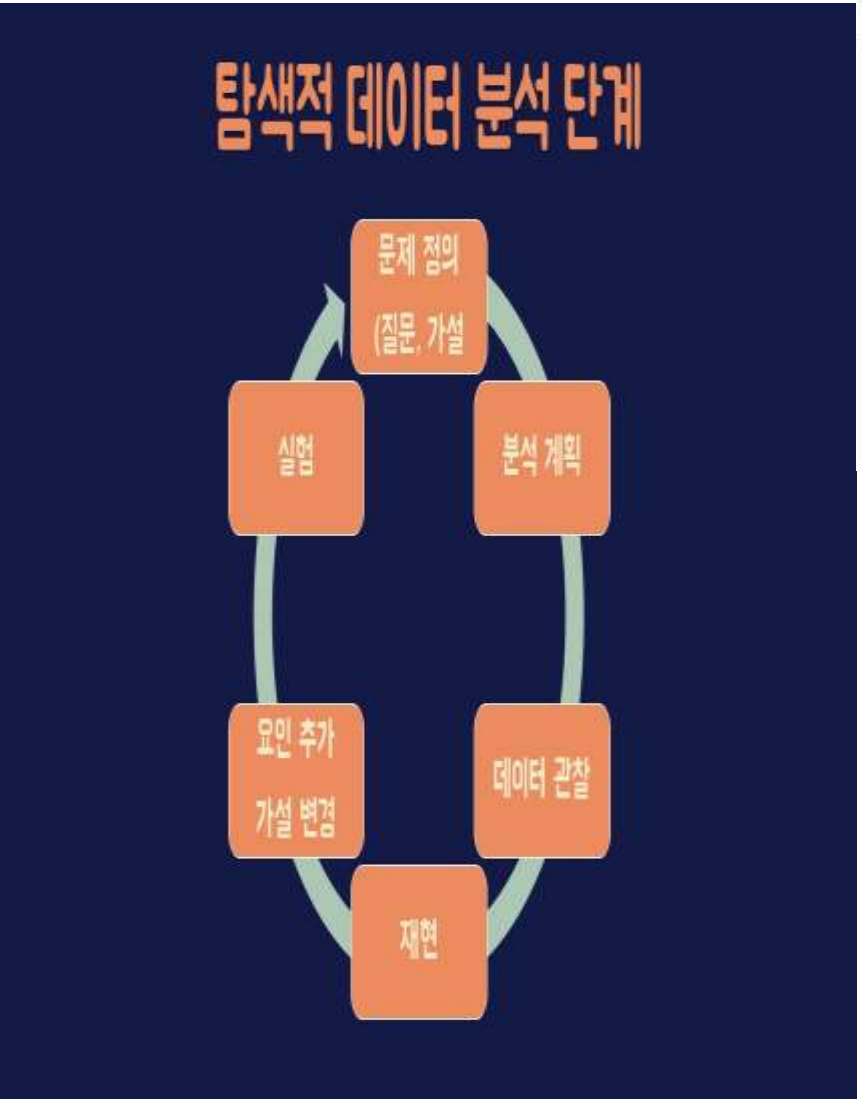


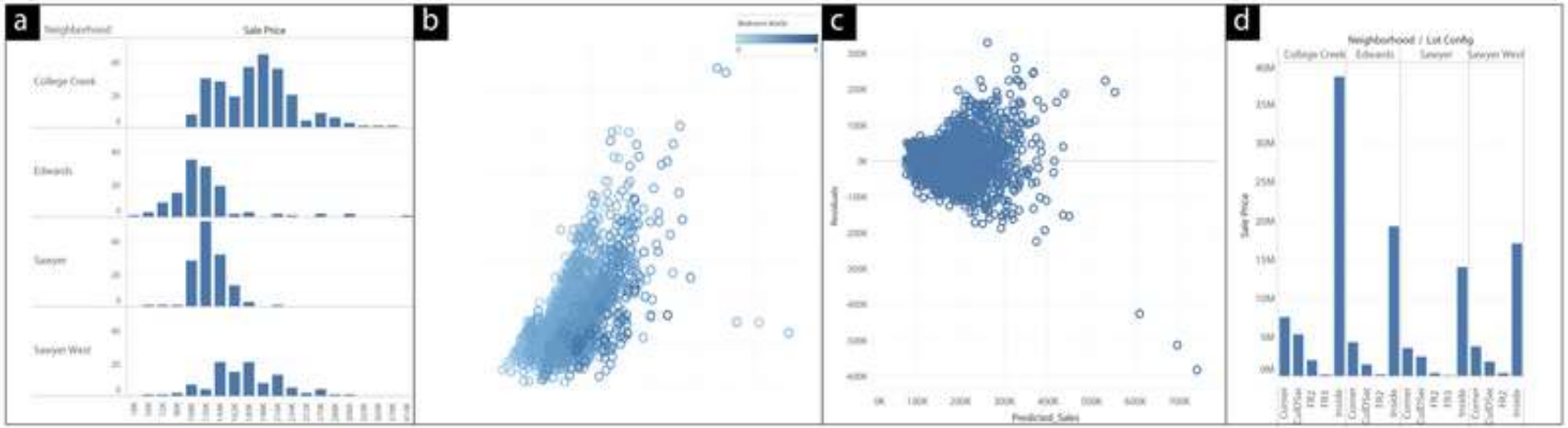
Part 2,

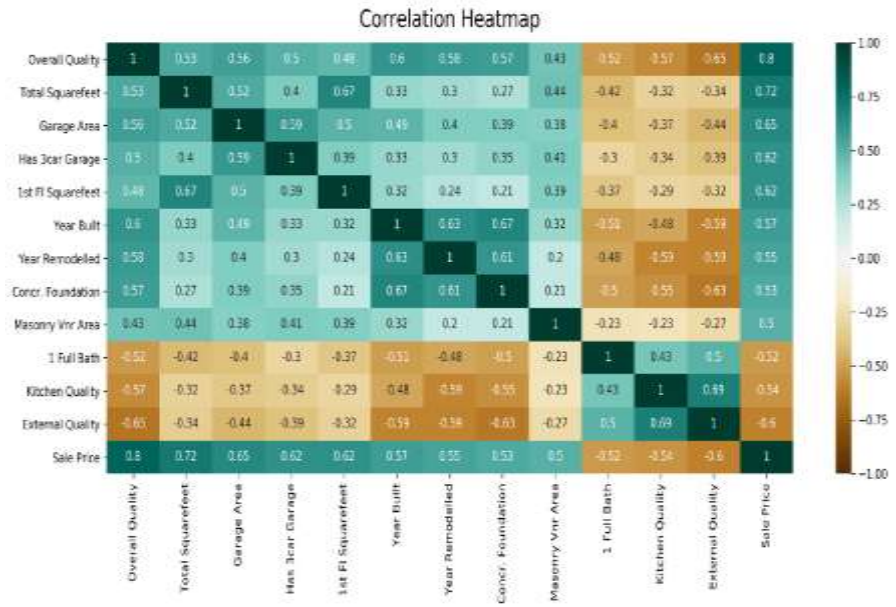
빅데이터 프로젝트

분석

EDA(Exploratory Data Analysis)








```
print(lr.summary())
```

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.910
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	461.2
Date:	Wed, 29 Sep 2021	Prob (F-statistic):	4.73e-71
Time:	19:09:53	Log-Likelihood:	-270.60
No. Observations:	140	AIC:	549.2
Df Residuals:	136	BIC:	561.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.3346	0.357	12.139	0.000	3.628	5.041
TV	0.0538	0.002	34.539	0.000	0.051	0.057
Newspaper	0.0063	0.007	0.902	0.369	-0.008	0.020
Radio	0.1100	0.010	10.609	0.000	0.090	0.131

Omnibus:	18.669	Durbin-Watson:	2.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.404
Skew:	-0.643	Prob(JB):	1.52e-07
Kurtosis:	4.932	Cond. No.	443.

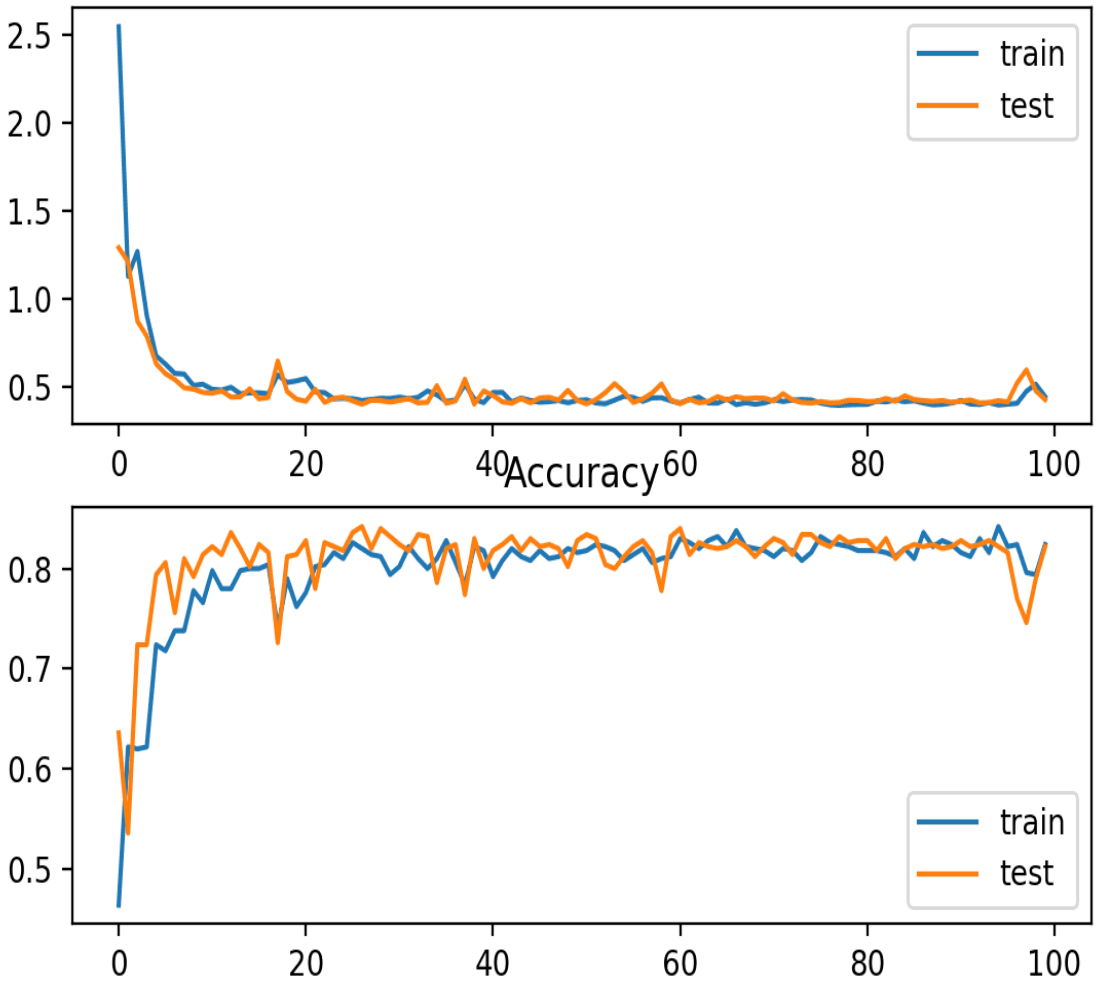
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Part 2,

빅데이터 프로젝트

모델링

Loss & Accuracy



Hyper Parameter Tuning

```
Fitting 2 folds for each of 12 candidates, totalling 24 fits
[CV] max_features=2, n_estimators=10 .....................................
[CV] max_features=2, n_estimators=10 .....................................
[CV] max_features=2, n_estimators=20 .....................................
[CV] max_features=2, n_estimators=20 .....................................
[CV] ................... max_features=2, n_estimators=10, total=   2.0s
[CV] max_features=2, n_estimators=30 .....................................
[CV] ................... max_features=2, n_estimators=10, total=   2.5s
[CV] max_features=2, n_estimators=30 .....................................
[CV] ................... max_features=2, n_estimators=20, total=   3.9s
[CV] max_features=4, n_estimators=10 .....................................
[CV] ................... max_features=2, n_estimators=20, total=   4.7s
[CV] max_features=4, n_estimators=10 .....................................
[CV] ................... max_features=4, n_estimators=10, total=   3.1s
[CV] max_features=4, n_estimators=20 .....................................
[CV] ................... max_features=2, n_estimators=30, total=   5.9s
[CV] max_features=4, n_estimators=20 .....................................
[CV] ................... max_features=4, n_estimators=10, total=   4.2s
[CV] max_features=4, n_estimators=30 .....................................
[CV] ................... max_features=2, n_estimators=30, total=   7.3s
[CV] max_features=4, n_estimators=30 .....................................
[CV] ................... max_features=4, n_estimators=20, total=   6.3s
[CV] bootstrap=False, max_features=2, n_estimators=10 ....................
[CV] ................... max_features=4, n_estimators=20, total=   7.7s
[CV] bootstrap=False, max_features=2, n_estimators=10 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=10, total=   3.0s
[CV] bootstrap=False, max_features=2, n_estimators=20 ....................
[CV] ................... max_features=4, n_estimators=30, total=   9.1s
[CV] bootstrap=False, max_features=2, n_estimators=20 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=10, total=   4.0s
[CV] bootstrap=False, max_features=2, n_estimators=30 ....................
[CV] ................... max_features=4, n_estimators=30, total=  11.4s
[CV] bootstrap=False, max_features=2, n_estimators=30 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=20, total=   6.1s
[CV] bootstrap=False, max_features=4, n_estimators=10 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=20, total=   8.4s
[CV] bootstrap=False, max_features=4, n_estimators=10 ....................
[CV] . bootstrap=False, max_features=4, n_estimators=10, total=   5.0s
[CV] bootstrap=False, max_features=4, n_estimators=20 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=30, total=   9.1s
[CV] bootstrap=False, max_features=4, n_estimators=20 ....................
[CV] . bootstrap=False, max_features=2, n_estimators=30, total=  12.8s
[CV] bootstrap=False, max_features=4, n_estimators=30 ....................
[CV] . bootstrap=False, max_features=4, n_estimators=10, total=   7.7s
[CV] bootstrap=False, max_features=4, n_estimators=30 ....................
[CV] . bootstrap=False, max_features=4, n_estimators=20, total=  10.0s
[CV] . bootstrap=False, max_features=4, n_estimators=20, total=  14.8s
[CV] . bootstrap=False, max_features=4, n_estimators=30, total=  12.2s
[CV] . bootstrap=False, max_features=4, n_estimators=30, total=  17.5s
[Parallel(n_jobs=-1)]: Done  24 out of  24 | elapsed:   55.2s finished
```

Part 2,

빅데이터 프로젝트

시각화 및 서비스

분석 결과 시각화 및 보고서 작성

서비스 개시

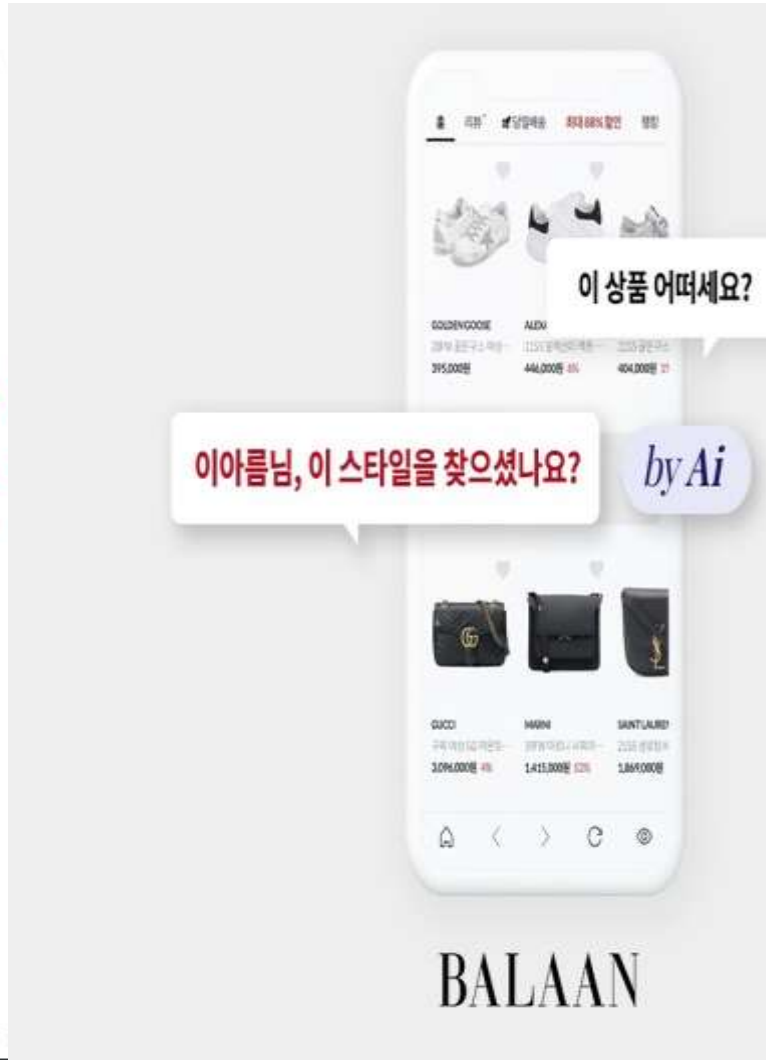
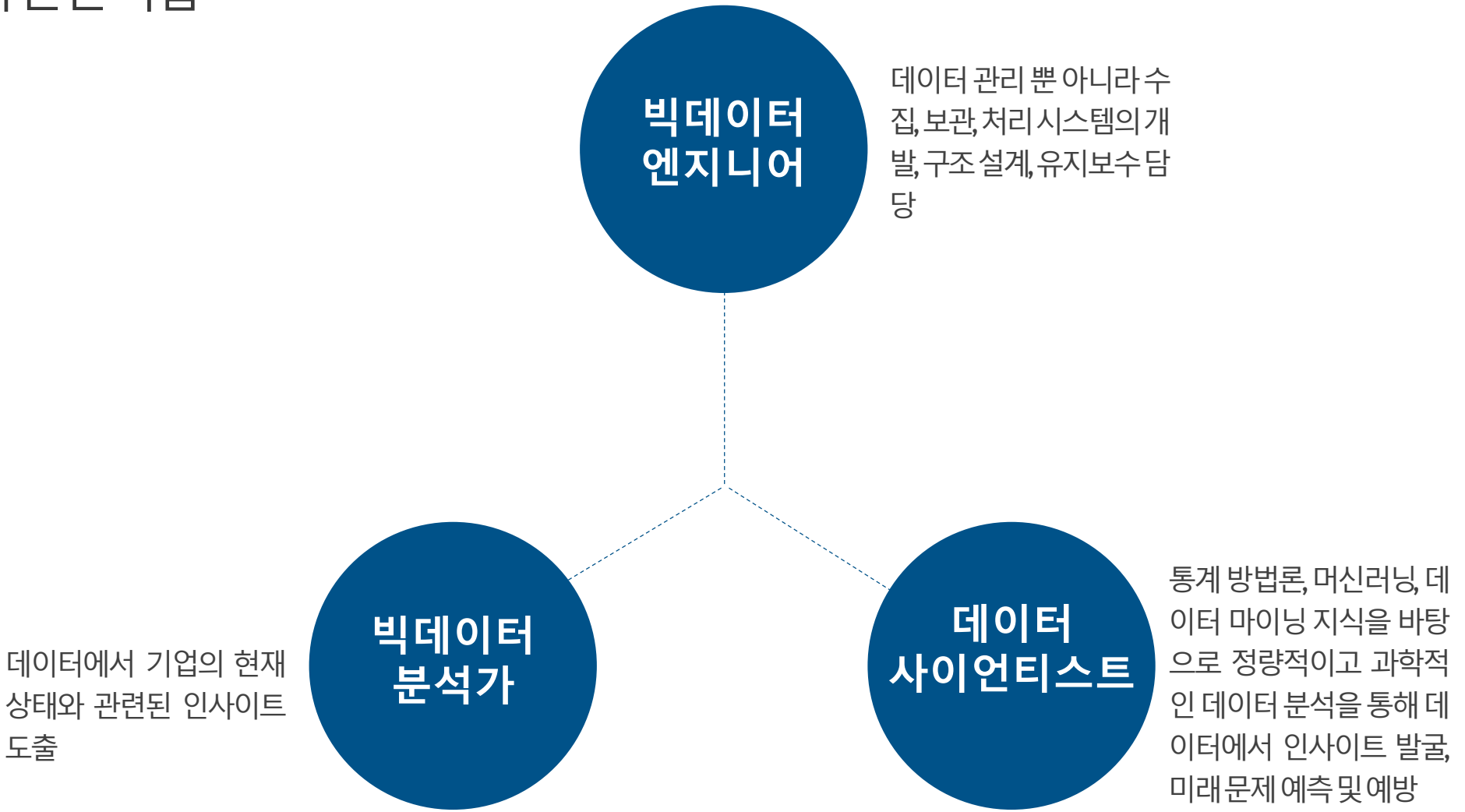


그림4 24시간 상담이 가능한 'KT AI 통화비서'

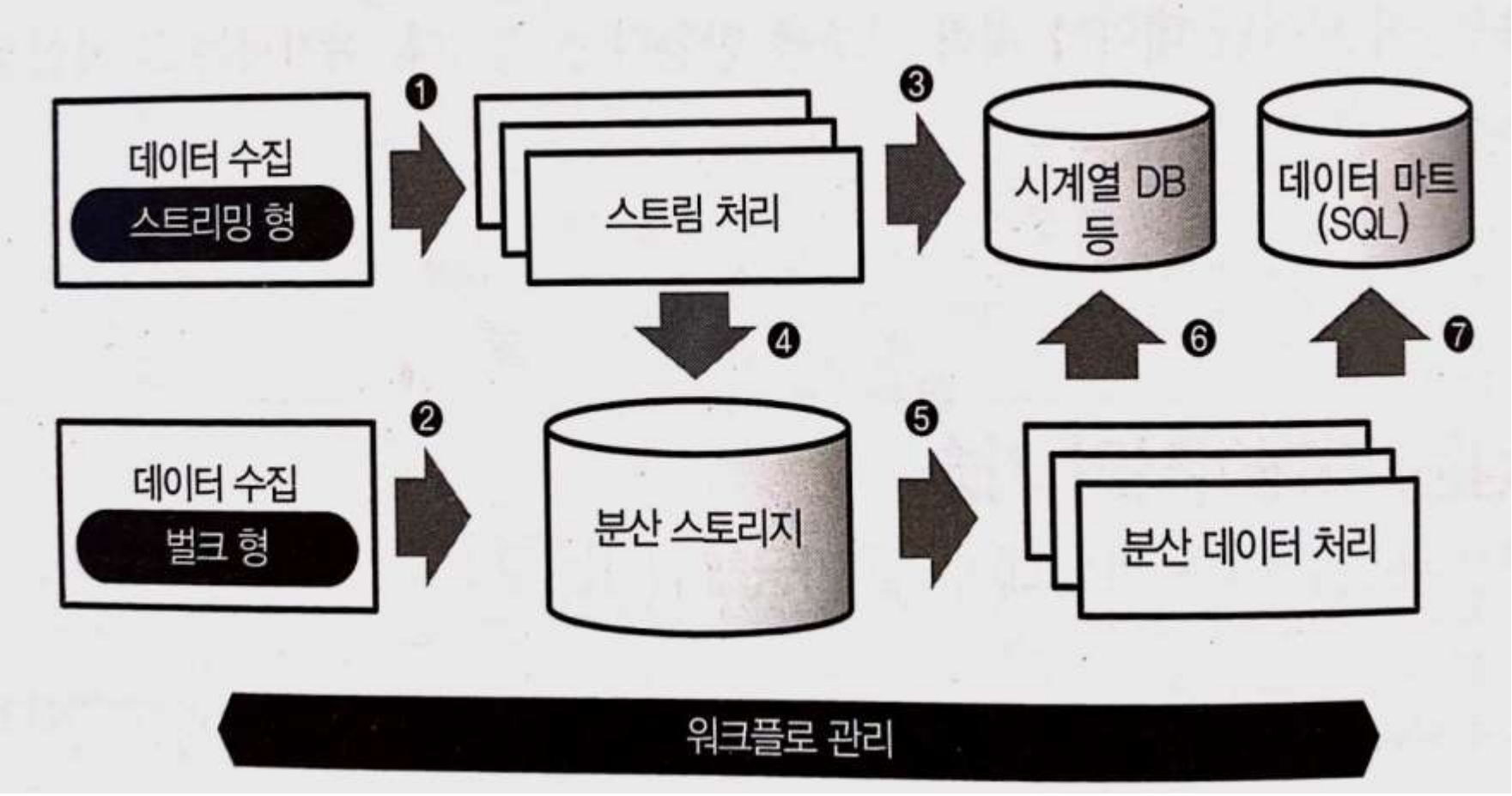
※ 출처: KT홈페이지

Part 2, 빅데이터 관련 직업 소개

빅데이터 관련 직업



빅데이터 엔지니어의 역할



데이터를 수집해서 안정적으로 데이터를 딜리버리 하는 역할이 데이터 엔지니어의 역할이다.

Part 2,

빅데이터 관련 직업 소개

빅데이터 엔지니어

빅데이터 엔지니어 채용공고

모집부문 / 상세내용

사용 기술

Python Kafka Elasticsearch Spark Hadoop Redis
MongoDB NoSql Spring Boot Java

주요업무

[이런 업무를 담당합니다]

• 줌닷컴/줌투자/검색 줌에서 발생하는 다양한 로그성 데이터를 처리/가공하는 업무를 수행합니다.
• 대용량 데이터의 처리를 위한 인프라를 운영하는 업무를 수행합니다.
• 검색 데이터 수집 및 가공처리를 위한 ETL파이프라인을 구축/운영합니다.
• 검색 데이터 수집을 위한 크롤링 업무를 수행합니다.

자격요건

[이런 역량을 갖춘 분을 찾습니다]

• 신입 또는 경력 2년 이상이신 분
• java 및 spring/spring boot 개발 역량이 있으시며 해당 개발 언어에 대한 기본 이해가 있으신 분
• database에 대한 이해 및 설계 가능하신 분: DB 정규화,DB INDEX,트랜잭션에 대한 개념의 이해 및 프로시저 구축 가능 하신 분
• 쿼리 작성 가능하신 분:테이블 조인 및 서브 쿼리 작성 가능하신 분
• nosql에 대한 이해도가 있으시며 활용 경험이 있으신 분 :mongo db,redis
• ETL/ELT 파이프라인 구축 경험이 있으신 분
• 파이썬으로 개발 경험이 있으신 분

우대사항

[이런 역량이 있으면 더욱 좋습니다]

• kafka에 대한 이해가 있으신 분
• NiFi에 대한 활용 경험이 있으신 분
• Elastic Search에 대한 활용 경험 및 구조에 대한 이해도가 있으신 분
• 대용량 데이터 분산 처리 시스템(Hadoop)에 대한 이해도와 활용 경험이 있으신 분
• Hadoop EcoSystem에 대한 이해도가 있으신 분
• 경력이 2년 이상이신 분

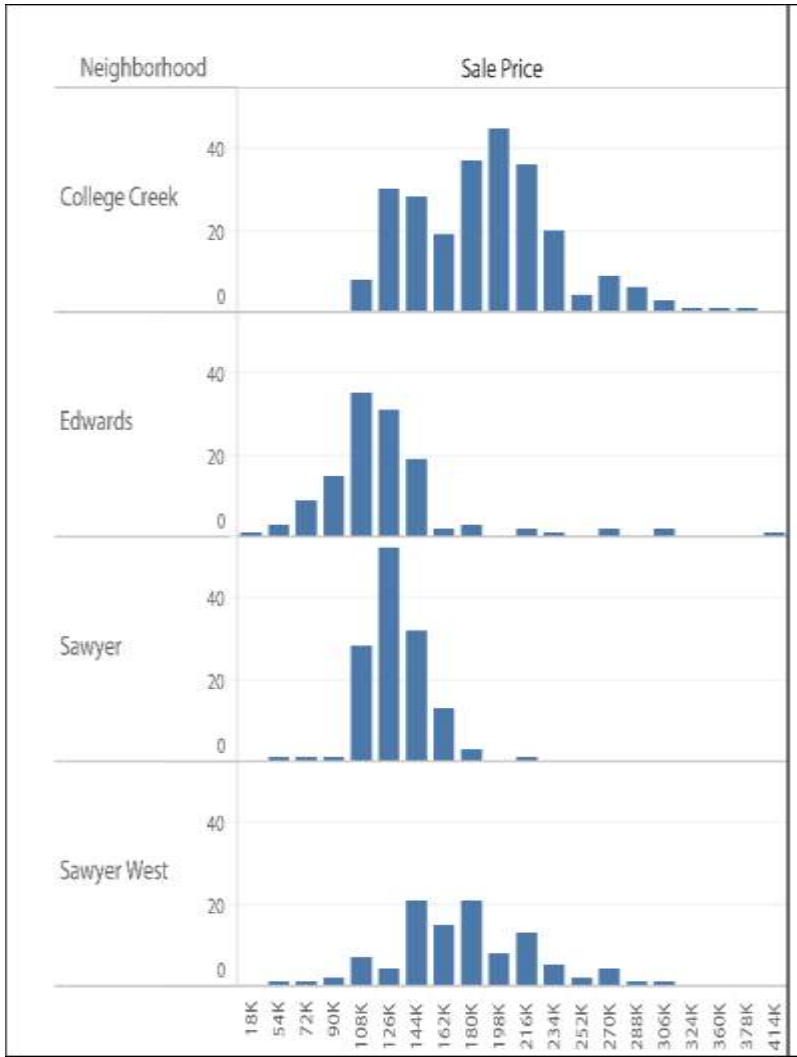
빅데이터와 관련하여 실시간 처리할 때 사용하는 **Kafka**, 대용량 데이터를 배치 처리 할 때 사용하는 **Hadoop**, 이 외에도 **Key-Value** 형태로 비정형 데이터를 저장할 수 있는 **Elastic Search, MongoDB** 등등의 데이터 베이스 지식을 필요로 한다.

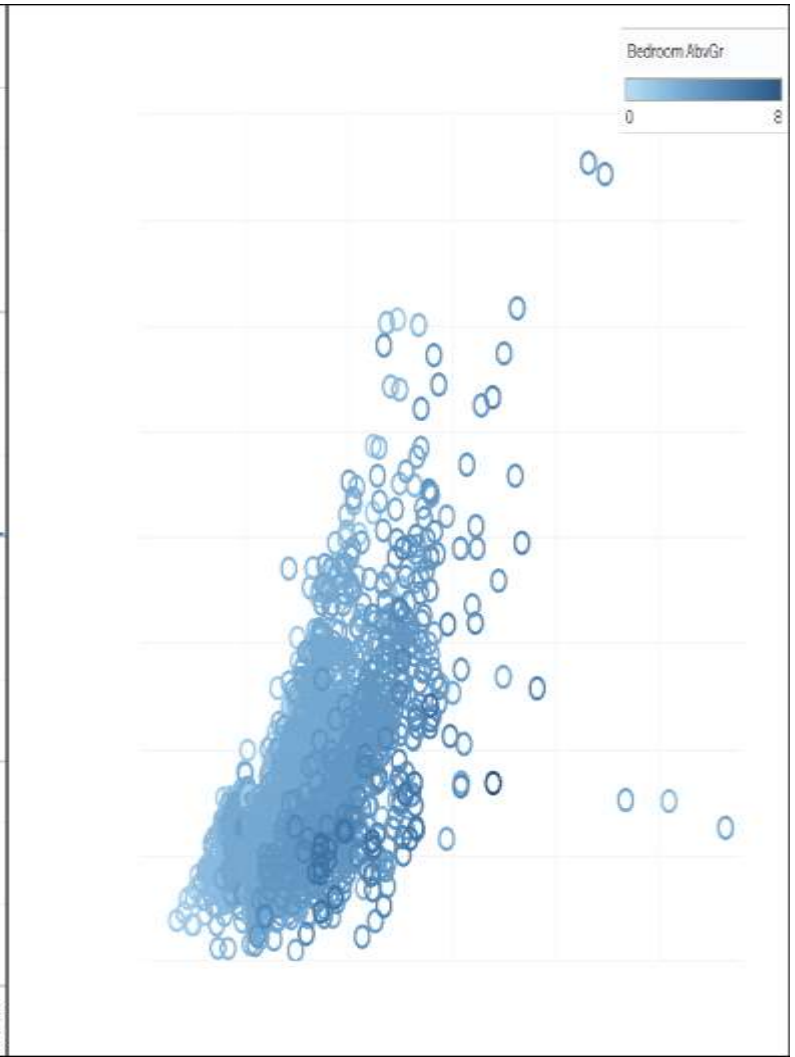
또한, 엔지니어인 만큼, **Java, Python, Spring boot**의 지식 또한 필요로 하는 것을 볼 수 있다.

빅데이터 관련 직업 소개

빅데이터 분석가

빅데이터 분석가의 역할





빅데이터 엔지니어가 수집과 저장, 처리를 맡았으면 빅데이터 분석가는 수집된 데이터를 기반으로 탐색적 데이터 분석 등의 작업을 수행한다.

통계, 수학적 지식을 기반으로 각 데이터 간의 상관관계, 인과관계 등을 분석하여 시각화 하고

비즈니스 인사이트를 도출하는 작업을 수행한다.

Part 2,

빅데이터 관련 직업 소개

빅데이터 분석가

빅데이터 분석가의 채용 공고

공통 자격요건

- 학력 : 대졸 이상 (4년)
- 나이/성별 : 무관

데이터분석가 | Process Intelligence팀 0명 (팀원 0명)

담당업무

- 데이터 분석 기반 프로세스 개선 컨설팅
- 프로세스마이닝 기반 데이터분석
- ProDiscovery에 필요한 분석기능 연구

지원자격

- R, Python, SQL 등 데이터 분석 도구의 능숙한 사용 능력 (하나 이상 필수)
- 데이터 추출, 전처리, 분석, 인사이트 도출까지 데이터 분석 전과정 경험

우대사항

- 프로세스 마이닝 지식보유
- 관련분야 대학원 전공 (산업공학, 통계학, 빅데이터 등)
- 프로젝트 매니지먼트 경험
- 문제 정의와 가설 수립을 잘하는 자

빅데이터 분석가의 경우 데이터 추출, 전처리 분석, 인사이트 도출에 대한 이해를 요구한다.

데이터를 추출하는 데 사용되는 **SQL**에 대한 지식을 필요로하고, 탐색적 데이터 분석**(EDA)**과정과

데이터를 분석하기 위해 필요한 언어인 **Python** 혹은 **R**에 대한 이해를 요구하는 것을 볼 수 있다.

통계학, 수학에 대한 심층적인 이해와 비즈니스 인사이트를 발굴해야 하는 만큼 산업공학, 통계학, 빅데이터 대학원 이상을 우대하는 것을 확인 할 수 있다.

Part 2,

빅데이터 관련 직업 소개

빅데이터 과학자

2012년 10월 하버드 비즈니스 리뷰

Harvard Business Review

Analytics And Data Science | Data Scientist: The Sexiest Job of the 21st Century

Subscribe Sign In

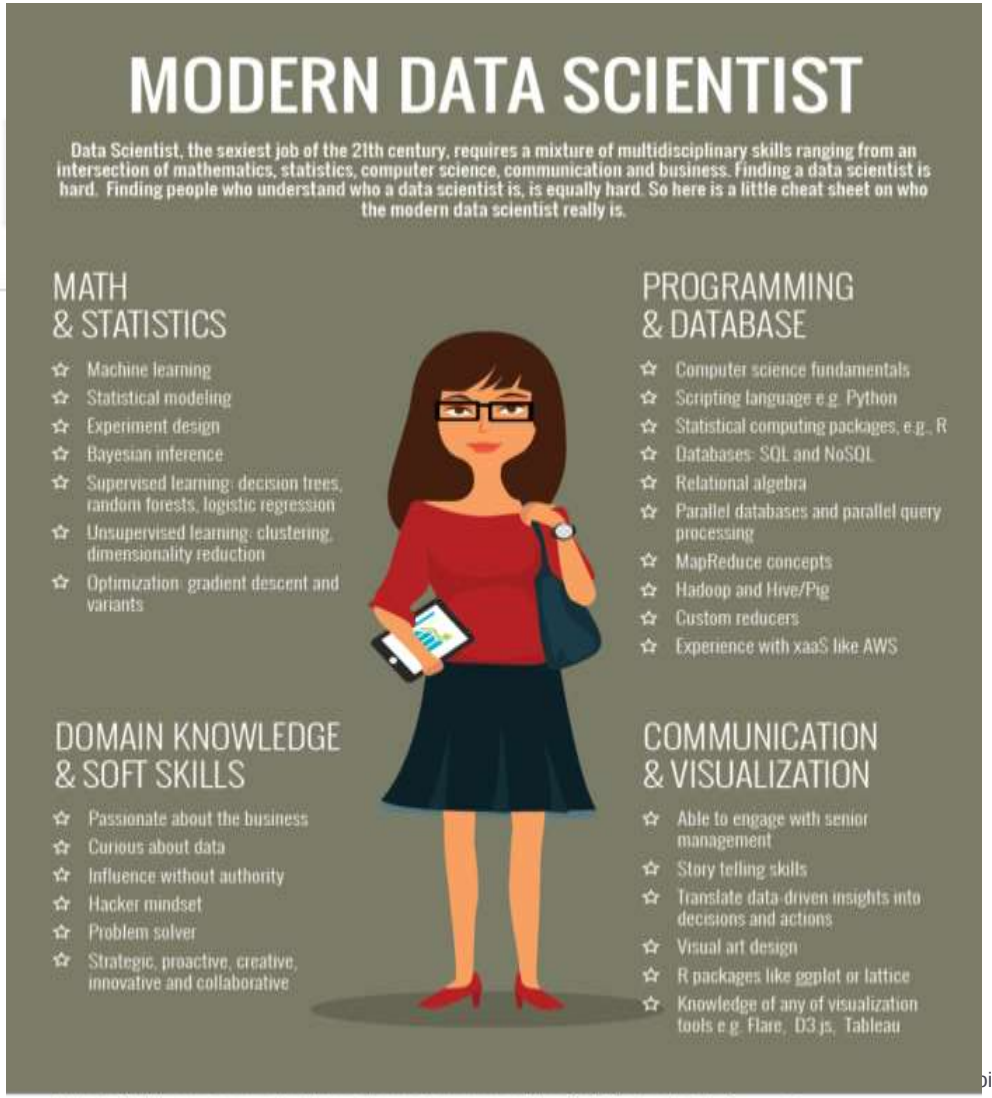
Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

데이터과학자

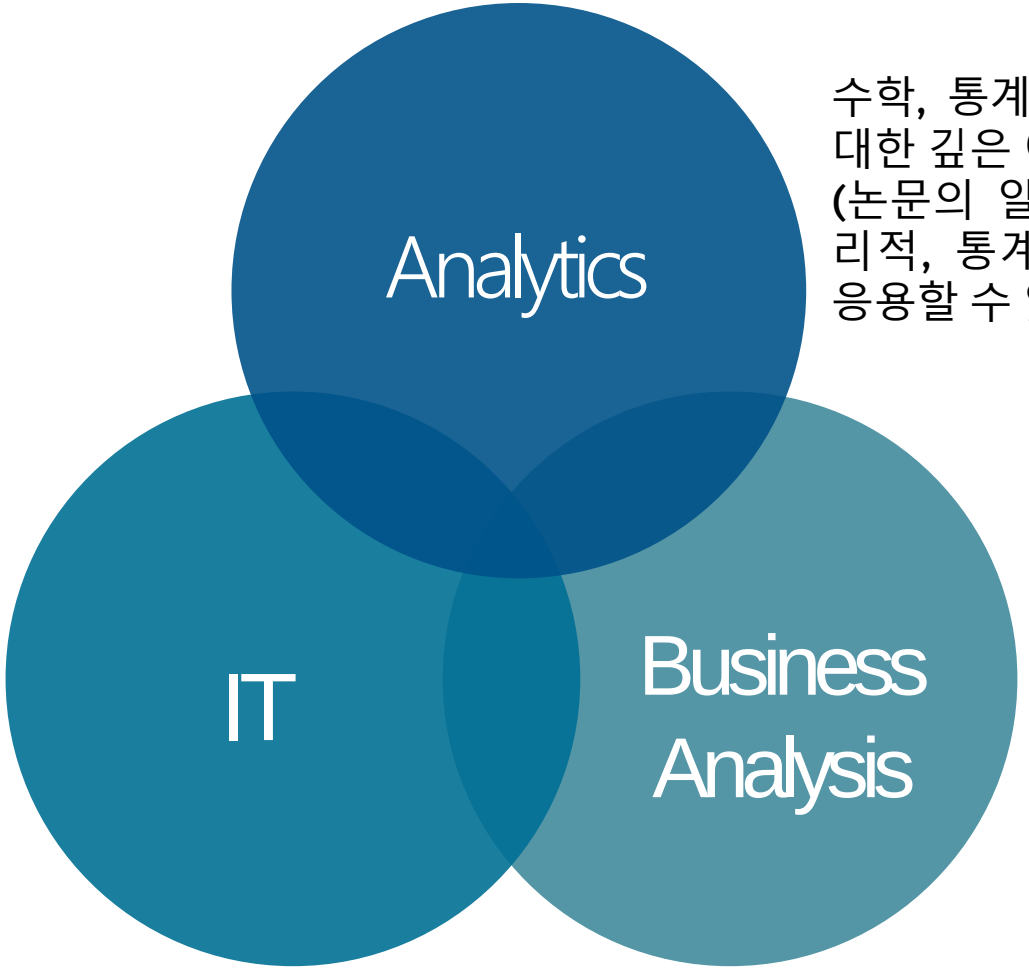


Part 2,

빅데이터 관련 직업 소개

데이터 과학자

데이터 사이언스의 구성요소



수학, 통계학, 데이터 분석에 대한 깊은 이해 (논문의 알고리즘을 보고 수리적, 통계적으로 해석하고 응용할 수 있는 역량)

프로그래밍, 데이터 엔지니어링 (논문의 수식과 통계적인 부분을 보고 코딩과 엔지니어링으로 구현해낼 수 있는 역량)

비즈니스 인사이트를 도출하고 비즈니스 문제를 해결해낼 수 있는 역량, 창의성, 커뮤니케이션, 스토리텔링, 시각화 등

Part 2,

빅데이터 관련 직업 소개

데이터 과학자

데이터 사이언스의 역량

빅데이터에 대한 이론적 지식
(수학, 통계, 경영, IT)

통찰력 있는 분석(창의적 사고, 호기심, 논리적 비판)

설들력 있는 전달(스토리텔링, 시각화)

다분야 간 협력(커뮤니케이션)

Part 2,

빅데이터 관련 직업 소개

데이터 과학자

데이터 과학자의 채용 공고

모집부문 / 상세내용

사용 기술

Spark Pandas SQL PyTorch TensorFlow MachineLearning BigData R Python

주요업무

- 데이터 수집
- 통계분석
- 머신러닝 모델 개발

자격요건

- 통계 및 데이터마이닝 관련 실무 경험이 있는 분
- 데이터 기반 실험 및 모델을 실제 서비스에 적용하고 개선하여 임팩트를 창출해 본 경험이 있는 분
- 논문을 이해하고 활용하는데 어려움이 없는 분
- 데이터 분석/모델링을 위한 언어 및프레임워크 (Python, R, SQL, Pandas, TensorFlow, Pytorch, Spark 등)를 사용해본적이 있으신 분
- 회사 내부 구성원들과 협업 및 다양한부문과의 원활한 커뮤니케이션 할 수 있는 역량이 있는 분

우대사항

- 통계적 추론, Supervised learning, 자연어 처리, 데이터 인프라에 대한 이해도 가 높으신 분
- 감성분석 실무경험이 있으신분
- 부동산 도메인에 대한 이해도가 높고, 부동산, 금융 데이터 분석 경험이 있으신 분
- 대용량 데이터 분석 및 빅데이터 플랫폼 활용 경험을 가지신 분(Hadoop, Spark 등)
- 데이터를 기반으로 의사결정을 해나가는 조직에서의 데이터 활용 경험이 있으신 분

논문 이해와 활용도, 데이터 분석 능력, 커뮤니케이션, 도메인 지식, 빅데이터 플랫폼 활용 능력, 데이터 기반 의사결정 등의 역량을 필요로 하는 것을 볼 수 있다.

빅데이터 관련 직업 소개

부록

빅데이터 관련 참고하면 좋은 자료

빅데이터 관련한 직업 및 역할에 대한 소개를 담은 책

7편으로 이루어져 있고, 데이터 인프라에 대한 소개를 담은 영상(영상 총 길이: 약 100분)

데이터 분석 가이드 책으로 데이터 수집부터 시각화까지 폭넓게 다루고 있는 책

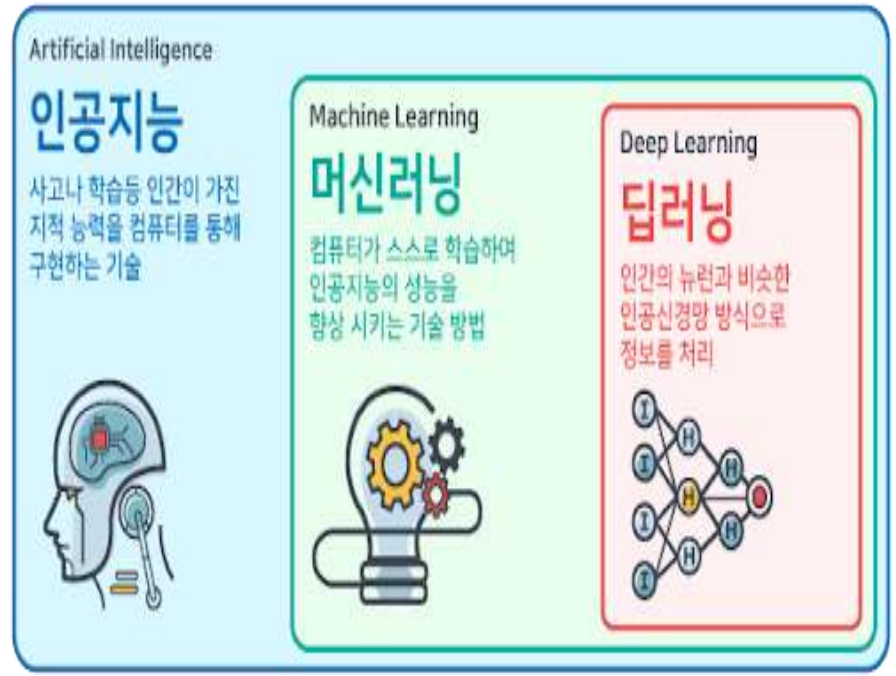
빅데이터 관련 직업 소개

부록 2

인공지능 관련 참고하면 좋은 자료

인공지능의학습법과 응용사례(안될과학)

TheAgeofAI 시리즈. 인공지능이어떻게활용되는지에대한시리즈영상으로총9개의영상이있다.



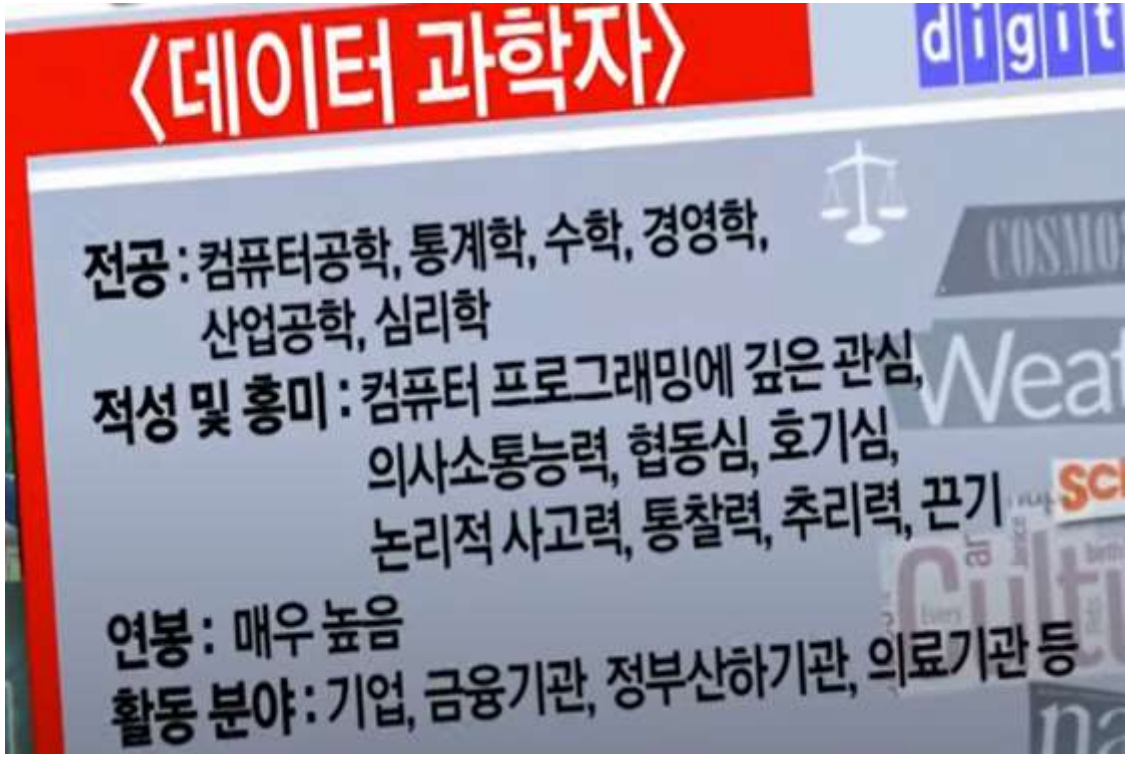
딥러닝의종류,한방에정리하기

Part 2,

빅데이터 관련 직업 소개

부록 2

데이터 과학 관련 참고하면 좋은 자료



[중등과학1]과학과나의미래-미래직업데이터과학자

[삼성SDS브이로그]데이터사이언티스트는대체무슨일을하는사람들이지?

끝