
데이터 분석과 머신러닝 기본이론



Contents

- 1. 개요
- 2. 배경지식
- 3. 빅데이터 시대의 데이터사이언스
- 4. 머신러닝 기본이론

Contents

- 1. 개요
- 2. 배경지식
- 3. 빅데이터 시대의 데이터사이언스
- 4. 머신러닝 기본이론

1. 개요

1.1 생략

1. 개요

1.2 사례-1 : 개발자들의 관심분야

✓ 전세계 개발자 대상 설문조사 결과

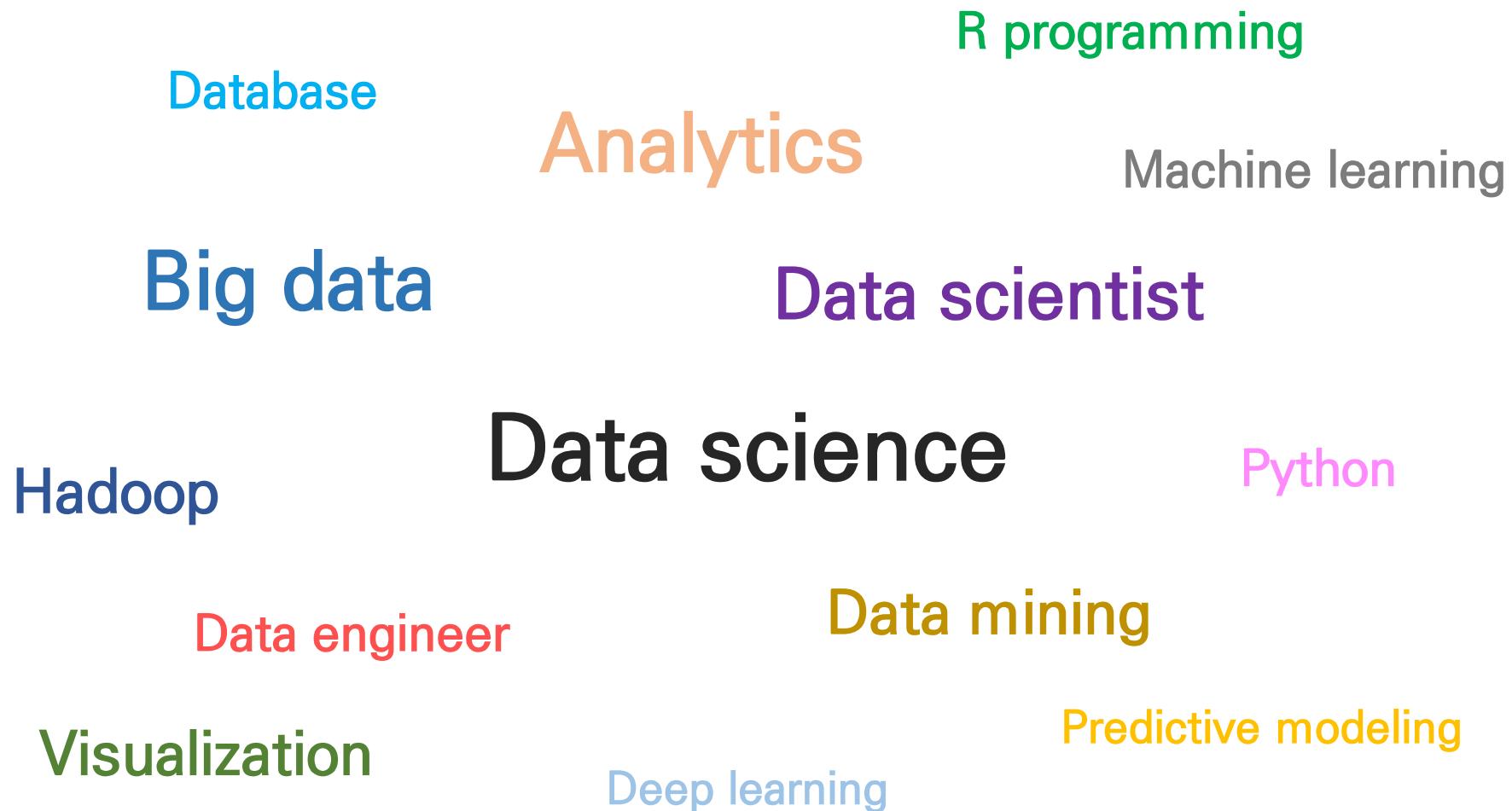


<출처 : <https://www.datasciencecentral.com/forum/topics/most-popular-data-science-keywords-on-dsc>>

1. 개요

1.2 사례-2 : 데이터사이언스 관련 키워드

✓ 구글 키워드 검색 통계



〈출처 : <https://www.datasciencecentral.com/forum/topics/most-popular-data-science-keywords-on-dsc>〉

Contents

- 1. 개요
- 2. 배경지식
- 3. 빅데이터 시대의 데이터사이언스
- 4. 머신러닝 기본이론

2. 배경지식

2.1 데이터 : 정형 vs 비정형

✓ 데이터란?

- 수, 영상, 단어 등의 형태로 된 의미의 단위, 연구나 조사 등의 바탕이 되는 재료

✓ 정형(Structured) 데이터

- 고정된 필드에 저장된 데이터
- 예 : 관계형 데이터베이스, 스프레드시트 등

호선	액번호	여	명	개	일평균	1월	2월	3월	4월	5월	6월	7월
1	150	서울역(1)	2E+07	93562	2823972	2542481	2885932	2918104	2993718	2770656	2864737	
1	151	A창(1)	9478426	44701	1367861	1114504	1380702	1432474	1460584	1334755	1387544	
1	152	종2(1)	1.7E+07	78856	2529090	2098588	2445715	2433679	2496189	2269646	2510207	
1	153	종3(1)	1.2E+07	54481	1729440	1465221	1692299	1638923	1734230	1600924	1692896	
1	154	종5(1)	1E+07	49299	1532069	1303823	1558679	1532793	1571939	1423812	1456749	
1	155	동대문(1)	5418592	25741	743915	682919	820750	799545	837528	774367	759568	
1	156	신수동(1)	5552611	26201	791093	674345	833421	825577	841804	776073	810298	
1	157	제기동	7678291	36429	1191456	915810	1121384	1135625	1148519	1080821	1084676	
1	158	청량리	9466711	45160	1412901	1207810	1441740	1383931	1416339	1311201	1292782	
1	159	동묘앞(1)	4059046	19514	543429	534707	635722	596043	642396	582161	524586	
2	201	사봉(2)	7951336	37210	1144461	931226	1142301	1211281	1199824	1105991	1216244	
2	202	율지로3가	1.6E+07	73925	2295751	1911496	2302489	2323887	2346445	2200448	2366896	
2	203	율지로4가	7152870	33466	1044646	850724	1030217	1052964	1077860	1001067	1095392	
2	204	율지로4가	3584934	16800	529106	429286	520078	517250	538099	507019	544096	
2	205	동대문역	5455416	25774	728454	658851	830546	794505	870377	782308	790375	

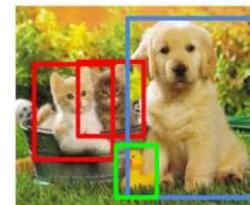
〈지하철 승하차 인원 집계결과 데이터〉

✓ 비정형(Unstructured) 데이터

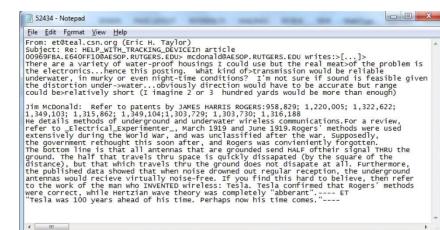
- 고정된 필드에 저장되어 있지 않은 데이터
- 예 : 텍스트, 이미지, 동영상, 음성 데이터 등



CAT



CAT, DOG, DUCK



```
"header": {
    "title": "The JSON example",
    "descriptionText": "This is some title text."
},
"content": {
    "title": "The content example text",
    "elements": [
        {
            "title": "The first element",
            "mainText": "First element main text",
            "additionalText": "First element additional text"
        },
        {
            "title": "The second element",
            "mainText": "Second element main text",
            "additionalText": "Second element additional text"
        }
    ]
}
```

✓ 반정형(Semi-structured) 데이터

- 고정된 필드에 저장되어 있지 않지만, 메타데이터나 스키마 등을 포함하는 데이터
- 예 : XML, HTML, JSON 형태 데이터 등

2. 배경지식

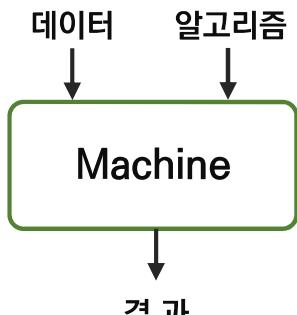
2.2 데이터사이언스와 머신러닝

✓ 데이터사이언스(Data Science)

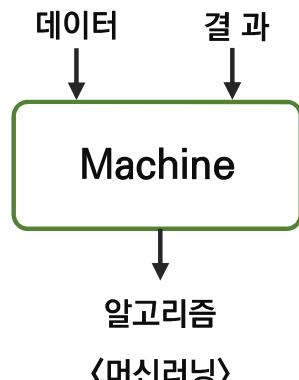
- 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야

✓ 머신러닝(Machine Learning)

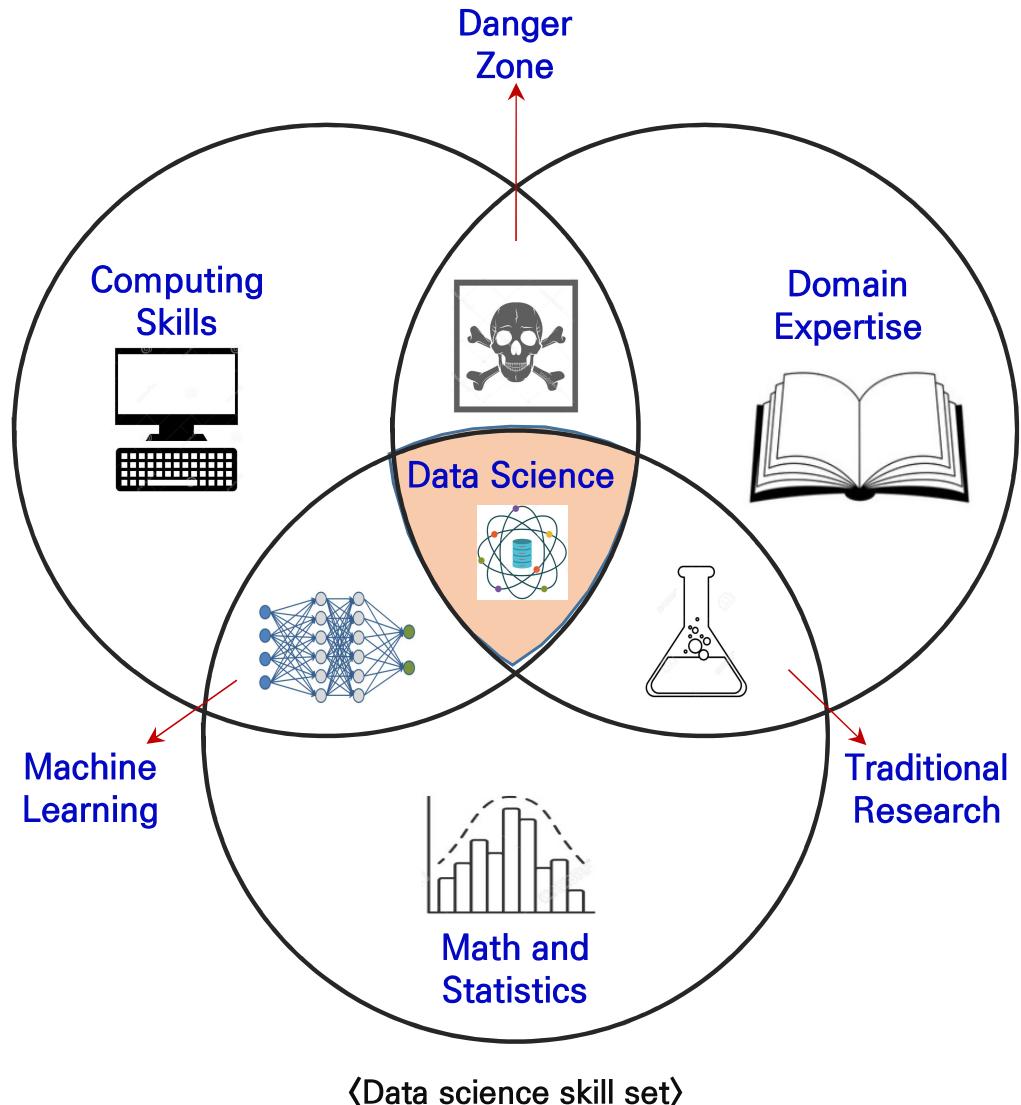
- 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
- 데이터사이언스의 영역에 포함됨



〈전통적인 프로그래밍〉

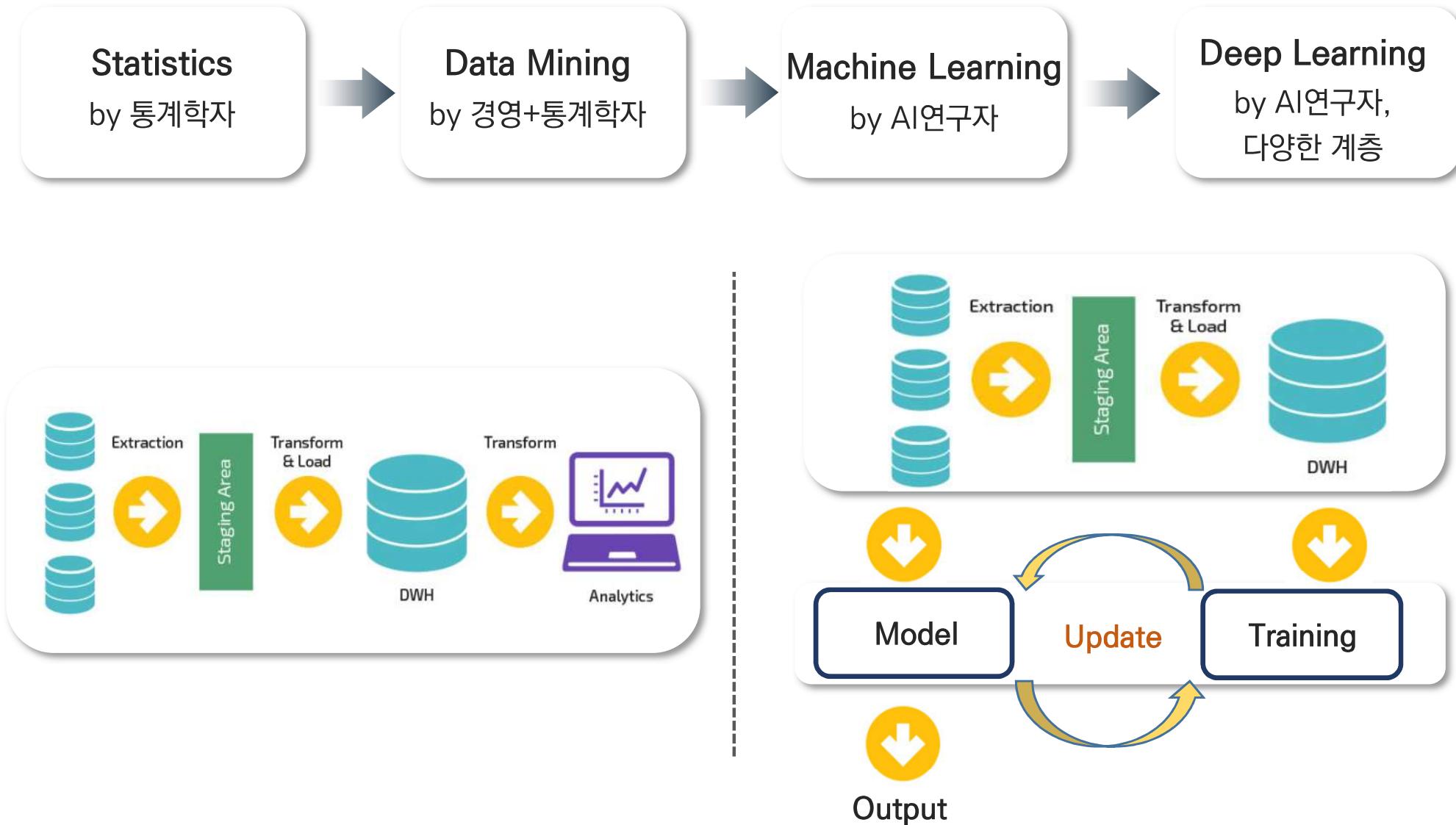


〈머신러닝〉



2. 배경지식

2.3 데이터사이언스 발전과정(알고리즘 분야)



2. 배경지식

2.4 데이터분석 절차



〈출처 : <https://learn.g2.com/data-analysis-process>〉

Contents

- 1. 개요
- 2. 배경지식
- 3. 빅데이터 시대의 데이터사이언스
- 4. 머신러닝 기본이론

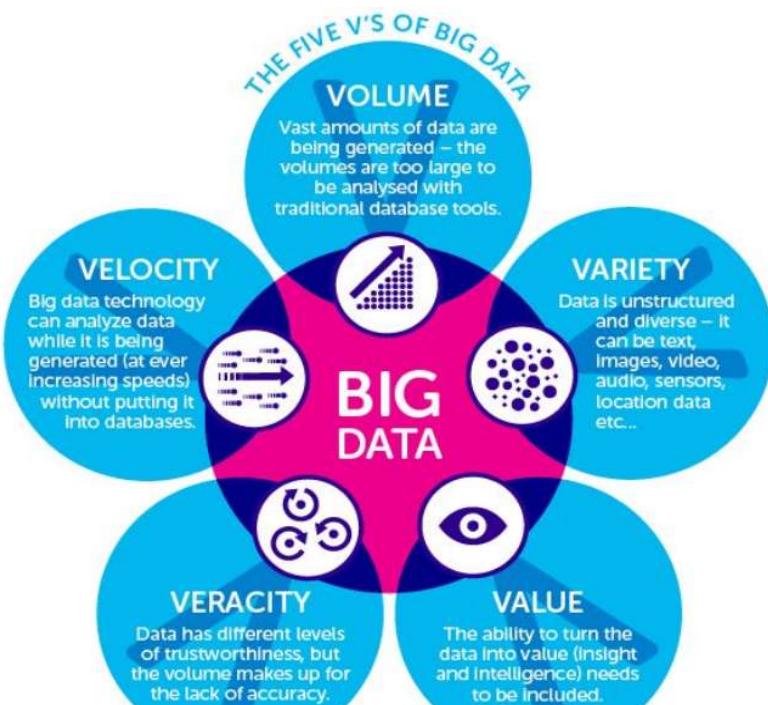
3. 빅데이터 시대의 데이터사이언스

3.1 빅데이터의 특징

✓ 3V + 2V → 5V

- Volume + Variety + Velocity
- Veracity + Value

- Volume(크기) : 수십 테라, 페타 바이트 이상 규모
- Variety (다양성) : 정형, 비정형(계량화 힘든 것들)을 포함
- Velocity(속도) : 빠른 속도로 생산, 짧은 시간만 유의미, 적시에 분석 필요
- Veracity(정확성) : 적절하고, 정확하고, 관련성이 있는 것
- Value(가치) : 가치를 창출할 수 있는 유용한 것



〈출처 :<https://twitter.com/CRUKresearch>

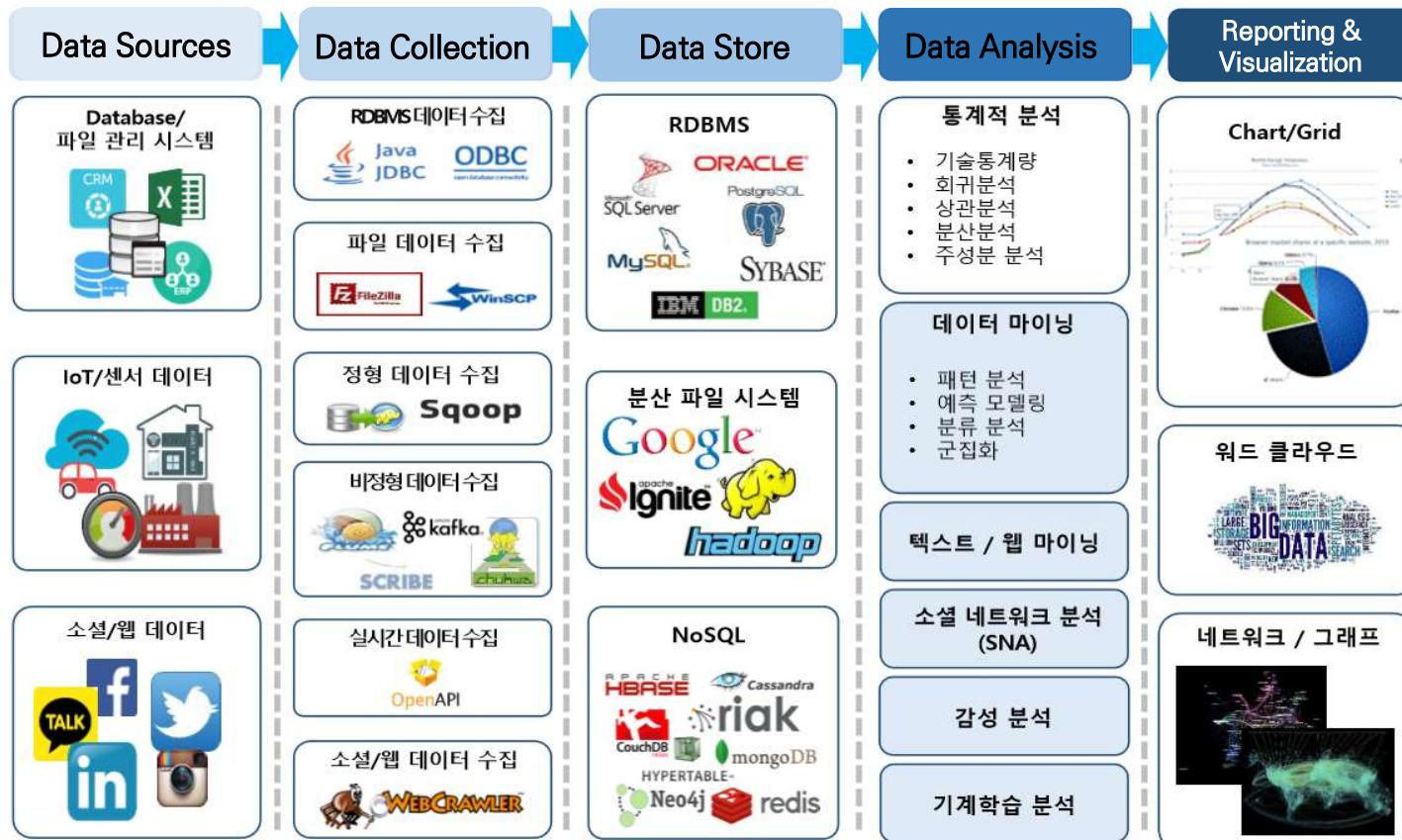
3. 빅데이터 시대의 데이터사이언스

3.2 빅데이터 접근방법

✓ 빅데이터를 효과적으로 활용하기 위한 핵심기술

- 1) 클라우드, 2) 분산저장, 3) 고속처리, 4) 그리고 분석

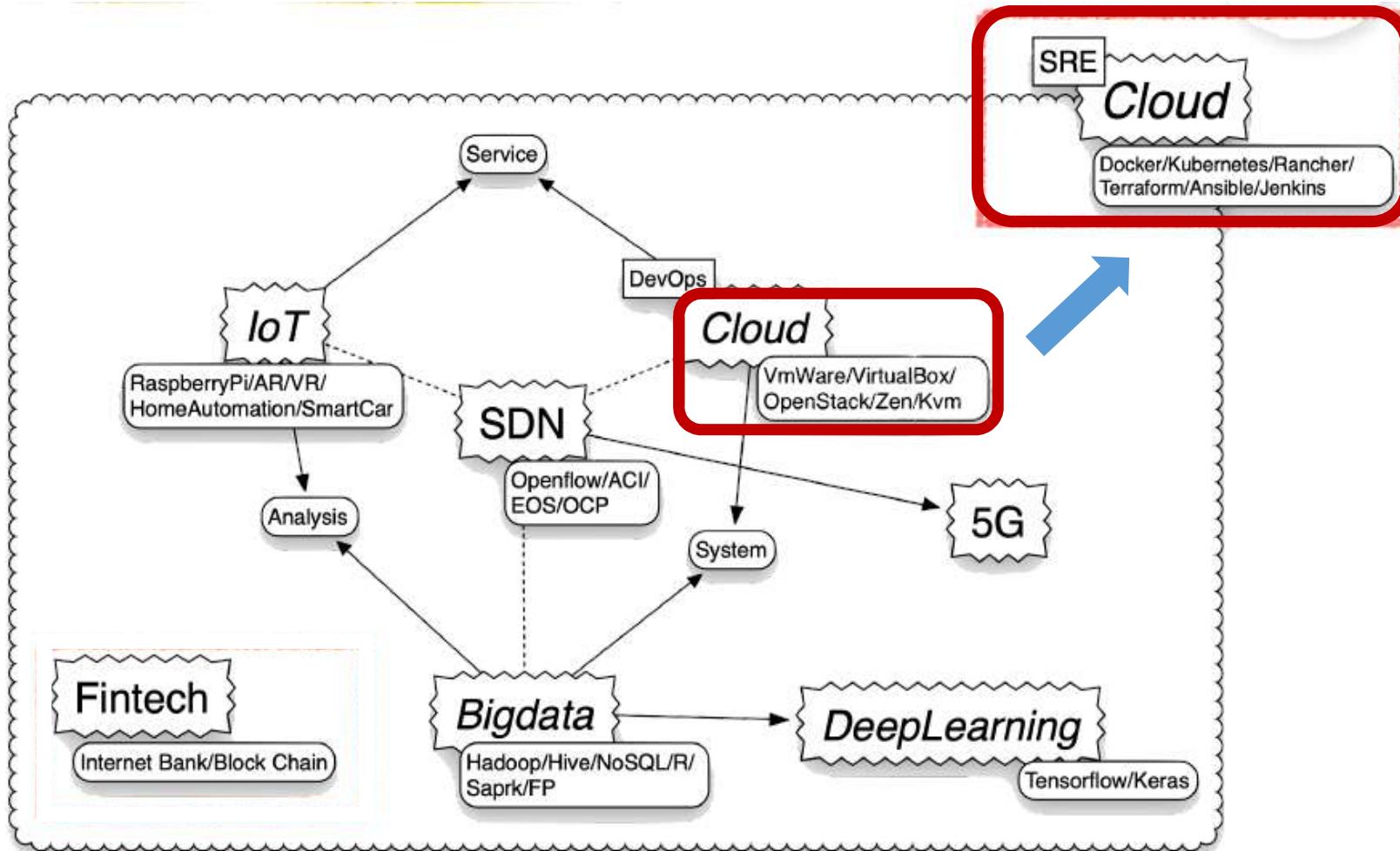
✓ 빅데이터 생태계



3. 빅데이터 시대의 데이터사이언스

3.3 핵심기술-1 : 클라우드

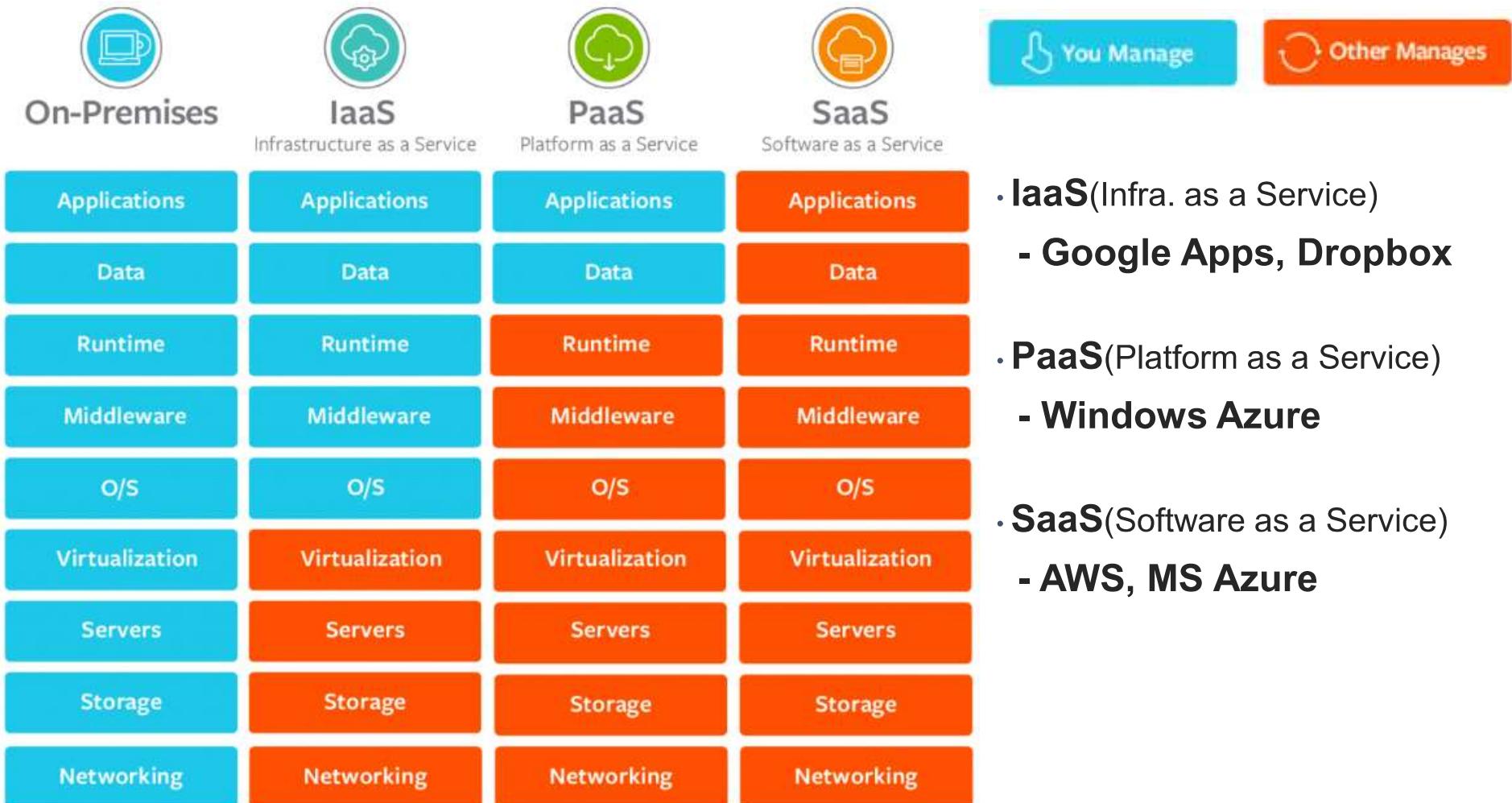
- ✓ Hypervisor → SRE(Site Reliability Engineering)



3. 빅데이터 시대의 데이터사이언스

3.3 핵심기술-1 : 클라우드

✓ IaaS/ PaaS/ SaaS



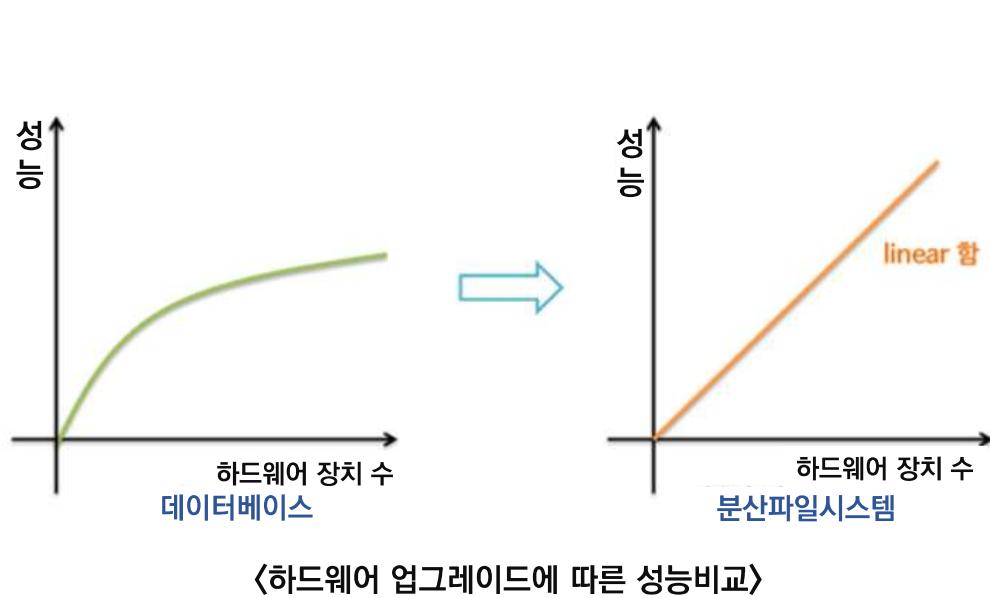
〈출처 : bmc blogs, <https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/>〉

3. 빅데이터 시대의 데이터사이언스

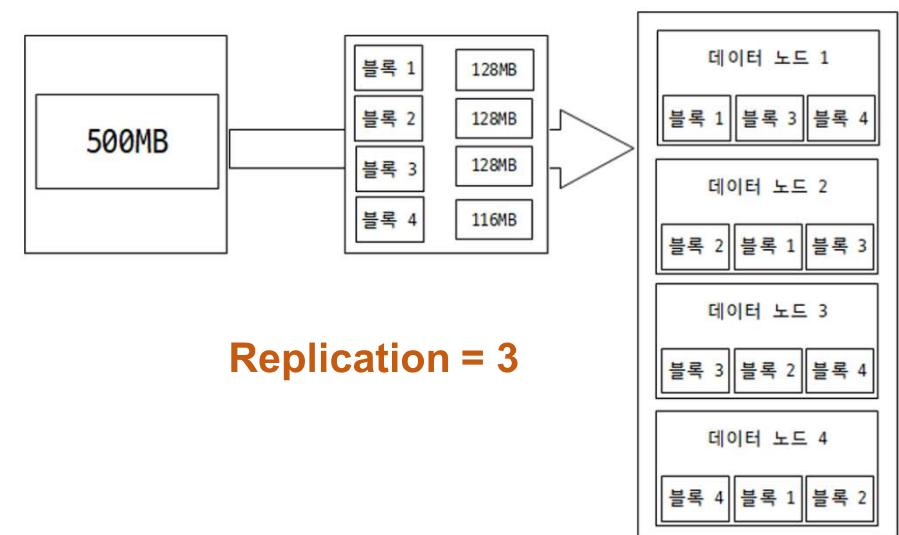
3.3 핵심기술-2 : 분산저장

✓ 하둡(Hadoop)

- 대용량 데이터를 분산저장, 처리 할 수 있는 자바 기반의 오픈소스 프레임워크
- 분산파일 시스템인 HDFS(Hadoop Distributed File System)에 데이터를 저장
- 분산 파일처리 시스템인 맵리듀스(MapReduce)를 이용해 데이터를 처리



〈하드웨어 업그레이드에 따른 성능비교〉



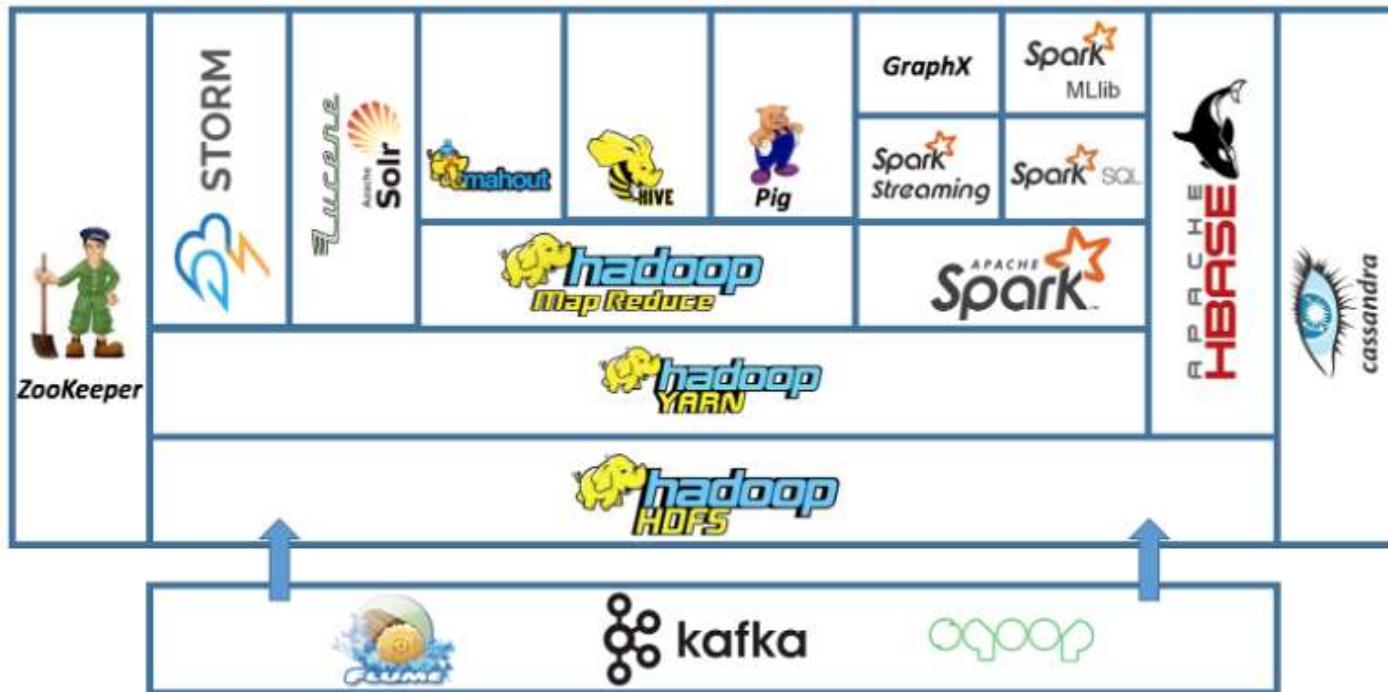
〈하둡 파일시스템 개념〉

3. 빅데이터 시대의 데이터사이언스

3.3 핵심기술-2 : 분산저장

✓ 하둡 에코시스템(Hadoop ecosystem)

- 하둡을 잘 사용하기 위해 제공되는 다양한 서브 프로젝트가 상용화 된 것
- 하둡의 기능을 보완하고 효율적으로 적용하기 위해 사용됨
- 하둡의 제한적인 성능을 대체하기도 함 : Hive, Spark 등



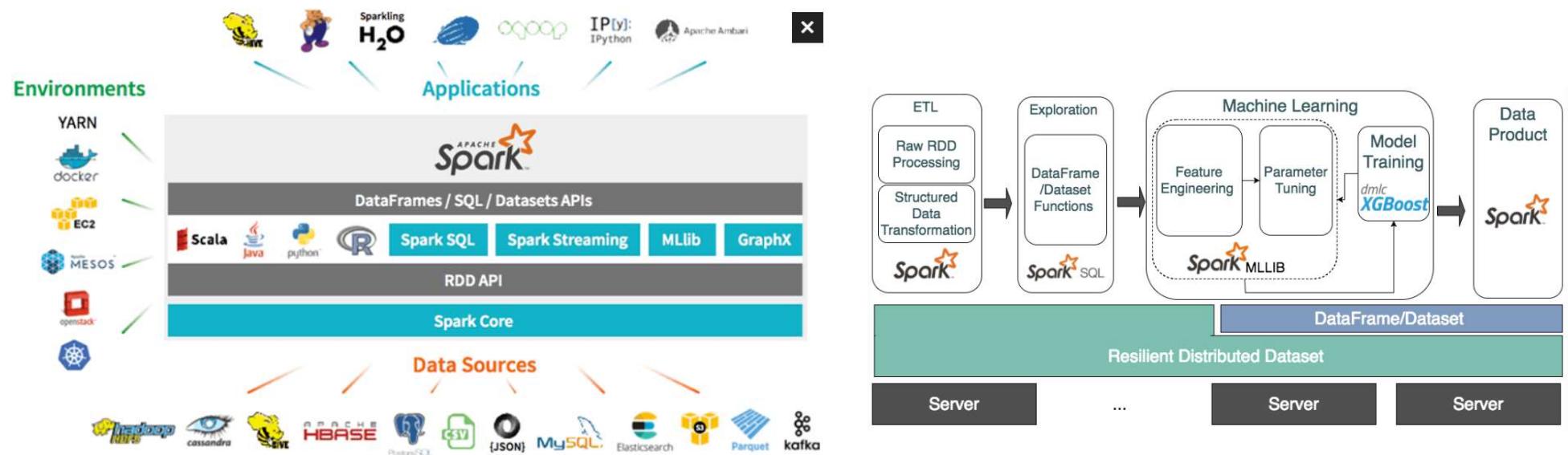
〈출처 : IT개발 전문가 그룹 OpenESD, http://www.openesd.com/?page_id=17

3. 빅데이터 시대의 데이터사이언스

3.3 핵심기술-3 : 고속처리

✓ 스파크(Spark)

- 빅데이터 고속 처리를 위한 병렬분산처리 플랫폼
- 반복되는 Machine learning, Data streaming, 쿼리, 배치 등 넓은 영역의 작업을 간단하게 표현
- In-memory 기반의 고속 데이터 처리: 디스크 기반의 MapReduce 작업의 느린 부분을 개선
- 하둡 기반 시스템 및 기존 DB 기술과 호환 : HDFS, Mysql, ...



〈출처 : SPARK FOR BIG DATA ANALYTICS, <https://www.jenunderwood.com/2016/10/16/spark-big-data-analytics-part-1/>〉

3. 빅데이터 시대의 데이터사이언스

3.3 핵심기술-4 : 분석(머신러닝, 딥러닝)

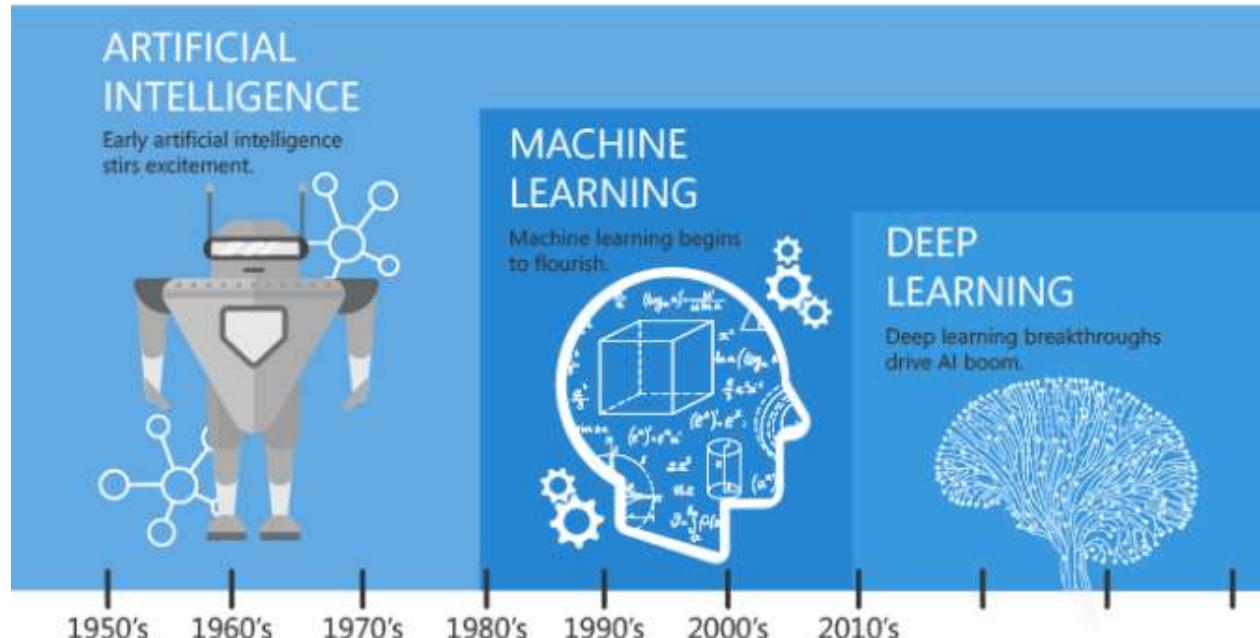
✓ 머신러닝(Machine learning)

- 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야
 - 모델을 학습시키기 위한 데이터가 필요

✓ 딥러닝(Deep learning)

- 여러 비선형 변환 기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습 알고리즘의 한 분야

* 추상화 : 다량의 데이터나 복잡한 자료들 속에서 핵심내용을 요약하는 작업



〈출처 <https://www.linkedin.com/pulse/ai-machine-learning-evolution-differences-connections-kapil-tandon/>〉

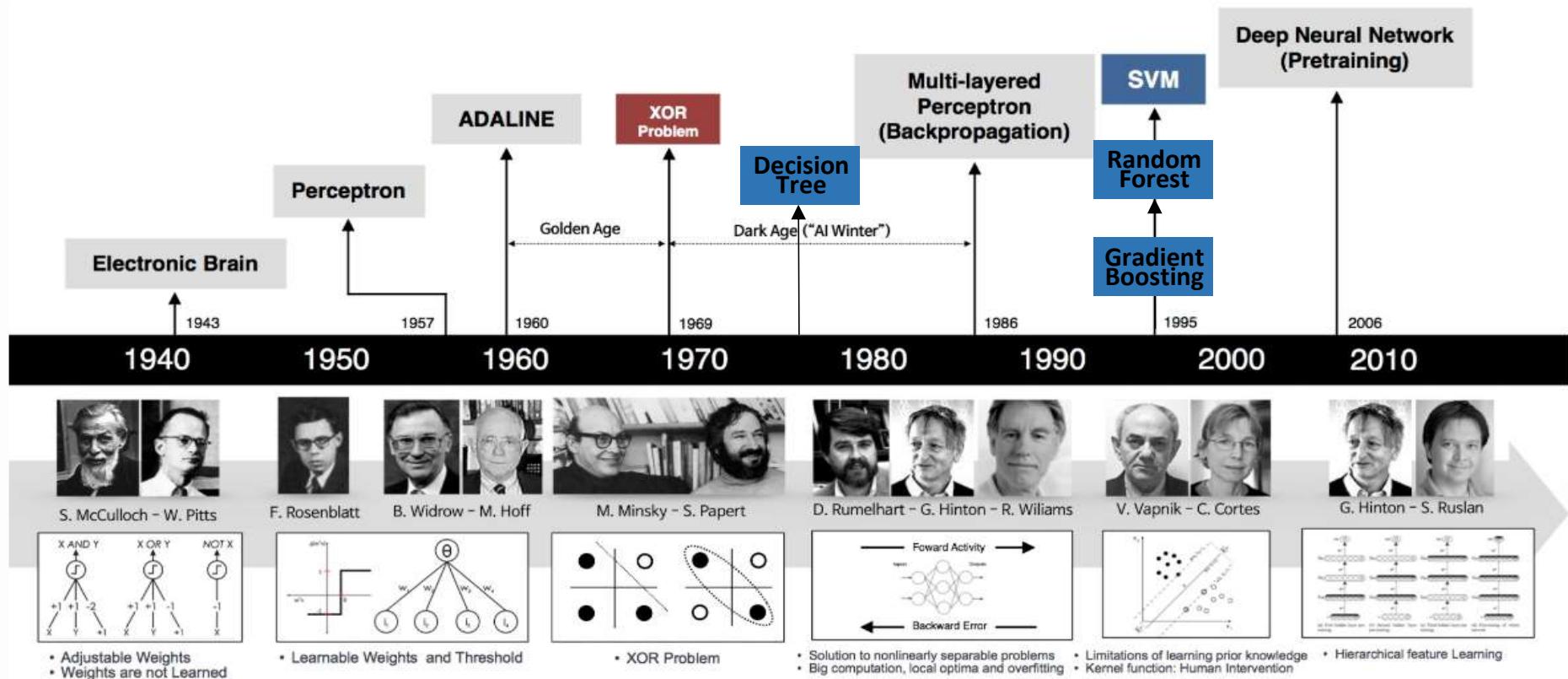
Contents

- 1. 개요
- 2. 배경지식
- 3. 빅데이터 시대의 데이터사이언스
- 4. 머신러닝 기본이론

4. 머신러닝 기본이론

4.1 머신러닝의 역사

✓ 머신러닝 발전과정



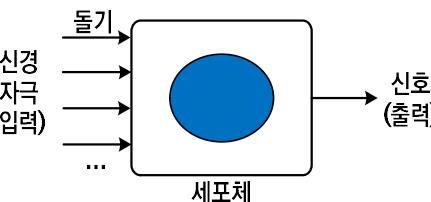
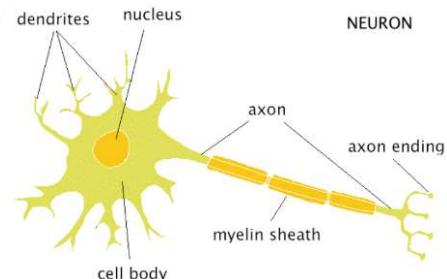
<출처 : Deep Learning 101 – Part 1: History and Background, http://beamlab.org/deeplearning/2017/02/23/deep_learning_101_part1.html>

4. 머신러닝 기본이론

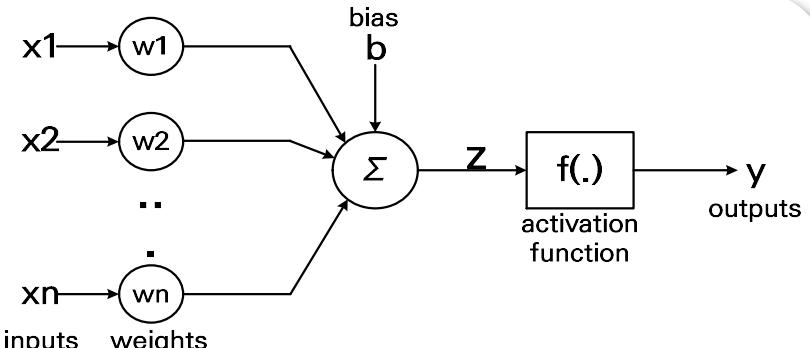
4.1 머신러닝의 역사

✓ MCP 뉴런(McCulloch-Pitts Neurons, 1943)

- 인공 신경망의 시초가 된 최초의 수학적 모델



〈뉴런(Neuron)의 구조(좌)와 단순화된 모델(우)〉



〈MCP 뉴런 모델〉

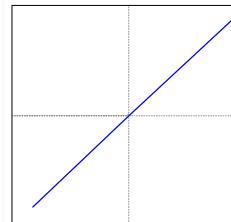
✓ 퍼셉트론(Perceptron, 1957)

- 학습을 하는 인공 신경망

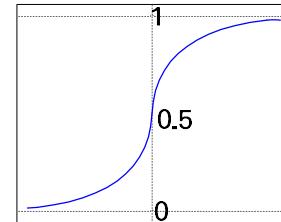


〈Frank Rosenblatt〉

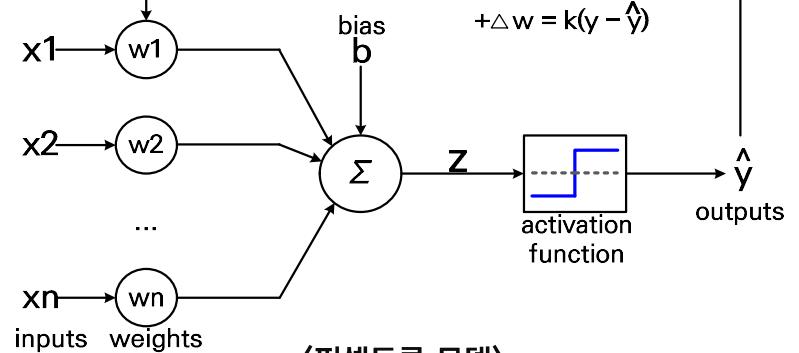
활성화 함수(Accivation function)



〈Adaline〉



〈Sigmoid〉



〈퍼셉트론 모델〉

4. 머신러닝 기본이론

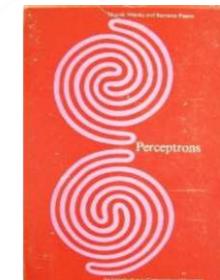
4.1 머신러닝의 역사

✓ AI Winter(1974 ~ 1980)

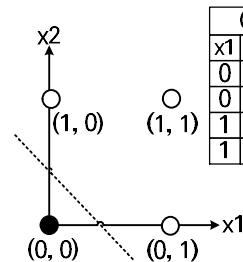
- 연결주의(Connectionism) vs 기호주의(Symbolism)



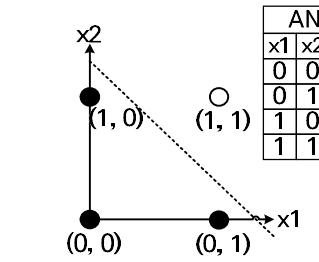
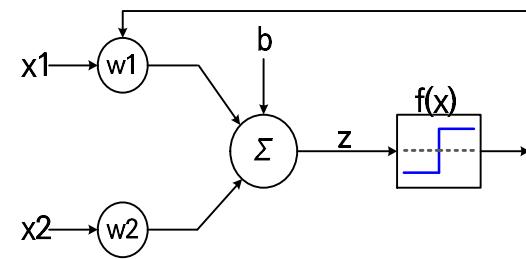
V
S



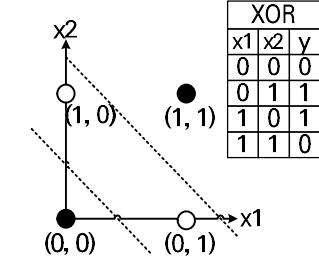
〈Minsky의 퍼셉트론〉



〈OR 연산〉



〈AND 연산〉



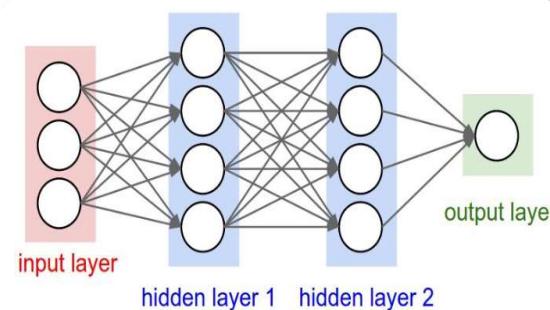
〈XOR 연산〉

✓ Deep Neural Net(1986)

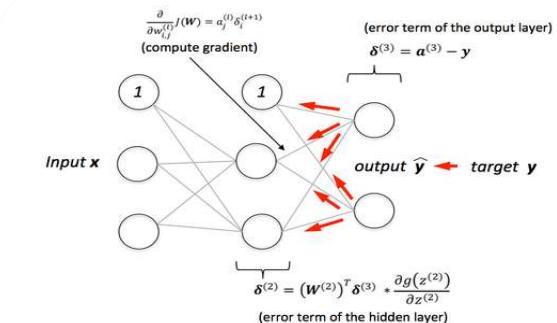
- 다층 퍼셉트론(Multilayer perceptron), 역전파(Backpropagation) 알고리즘을 통한 학습 성능 개선



〈Geoffrey Hinton〉



〈심층 신경망(Deep neural net)〉



〈역전파 알고리즘〉

4. 머신러닝 기본이론

4.1 머신러닝의 역사

✓ Statistical Approaches to Machine Learning(1995)

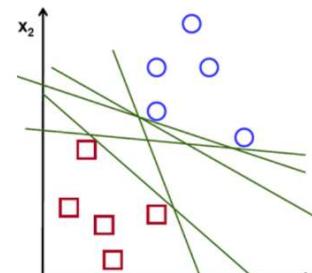
- 2nd AI Winter
- SVM(Support Vector Machines), 랜덤 포레스트(Random Forest), 그래디언트 부스팅(Gradient Boosting)



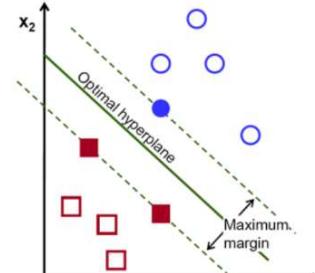
〈V. Vapnik〉



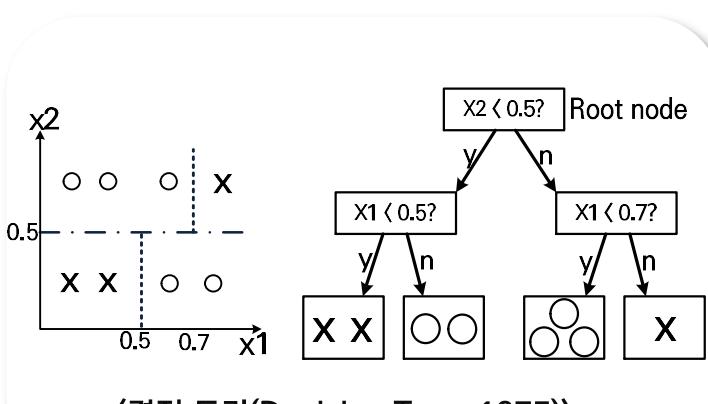
〈C. Cortes〉



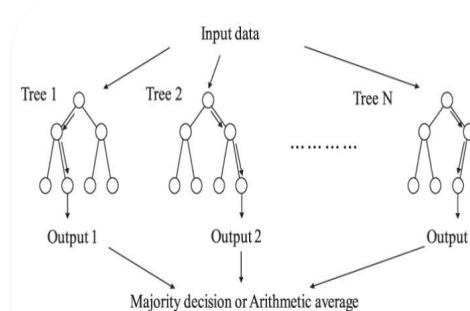
〈Logistic Regression〉



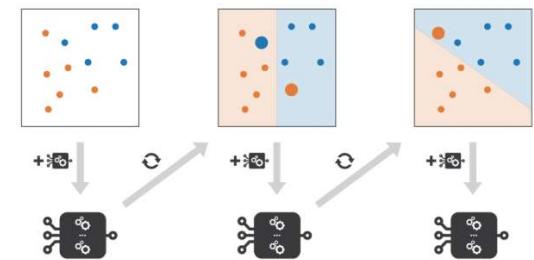
〈SVM with maximum margin〉



〈결정 트리(Decision Tree, 1975)〉



〈Random forest〉



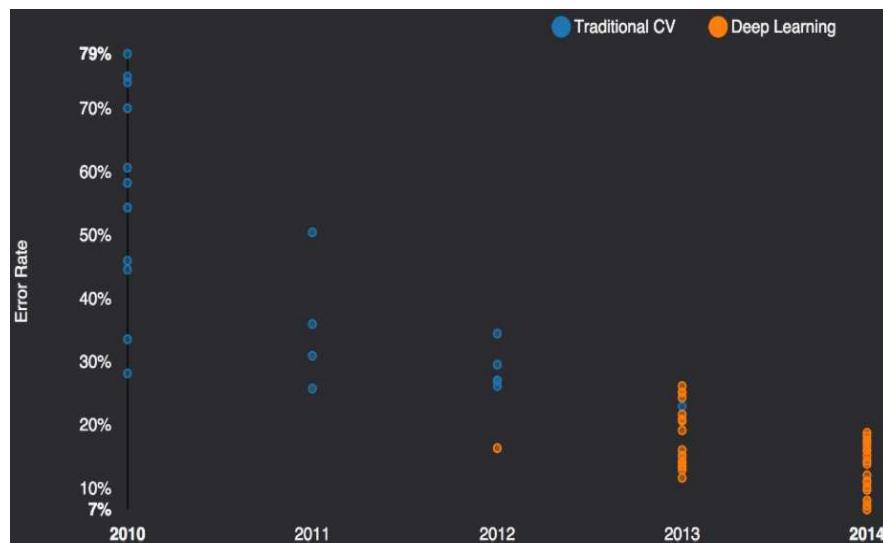
〈Gradient boosting〉

4. 머신러닝 기본이론

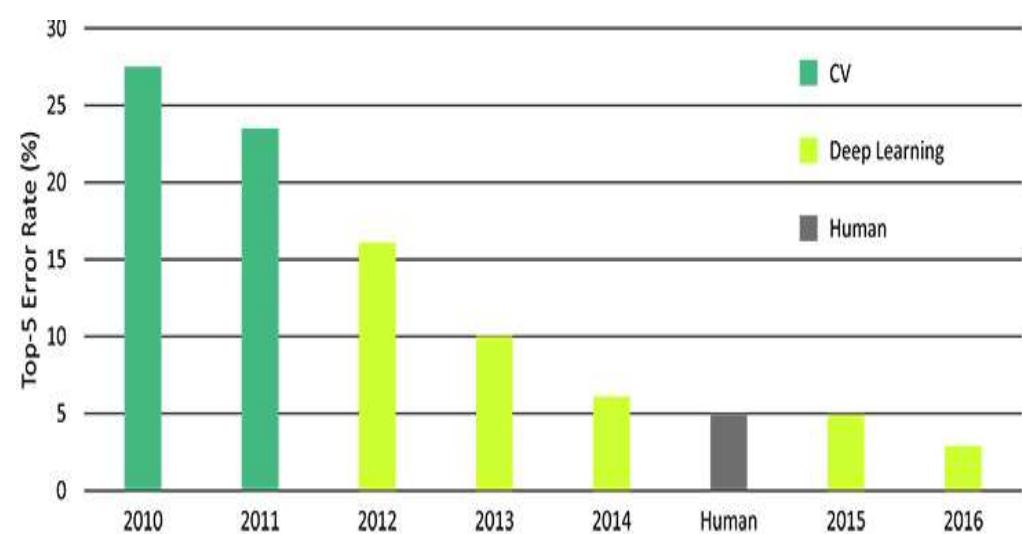
4.1 머신러닝의 역사

✓ 딥러닝의 시대(2006 ~)

- IMAGENET에서의 성공 : Traditional CV(Computer Vision) → Deep learning



〈출처 : <https://medium.com/global-silicon-valley/machine-learning-yesterday-today-tomorrow-3d3023c7b519>〉



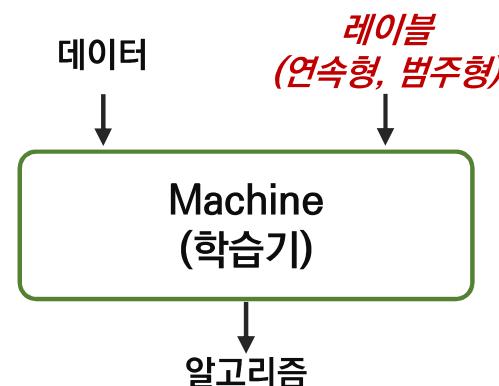
〈출처 : <https://www.dsiac.org/resources/journals/dsiac/winter-2017-volume-4-number-1/real-time-situ-intelligent-video-analytics>〉

4. 머신러닝 기본이론

4.2 머신러닝 개요

✓ 머신러닝의 종류

- 레이블(Label)의 유무에 따른 구분 :
 - 지도학습(Supervised Learning)
 - 비지도학습(Unsupervised Learning)
- 레이블(Label)의 유형에 따른 구분 :
 - 회귀(Regression, 연속형 레이블)
 - 분류(Classification, 범주형 레이블)

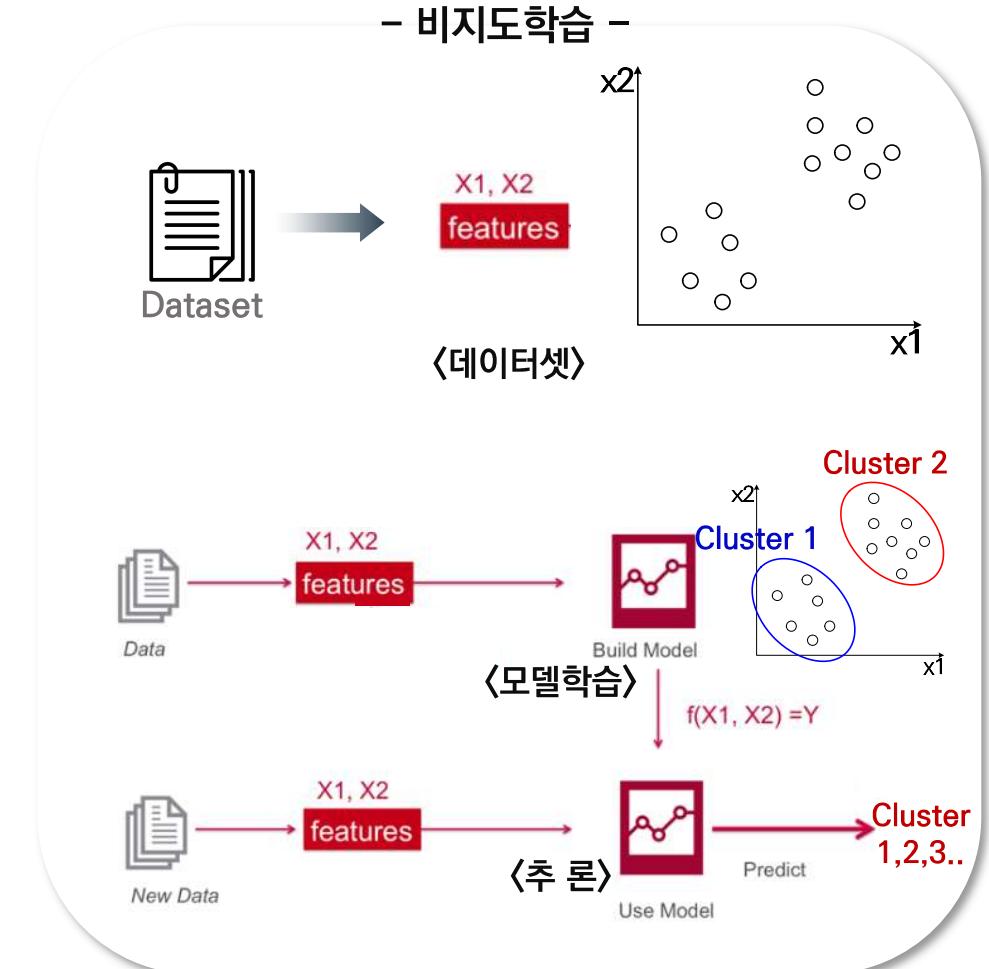
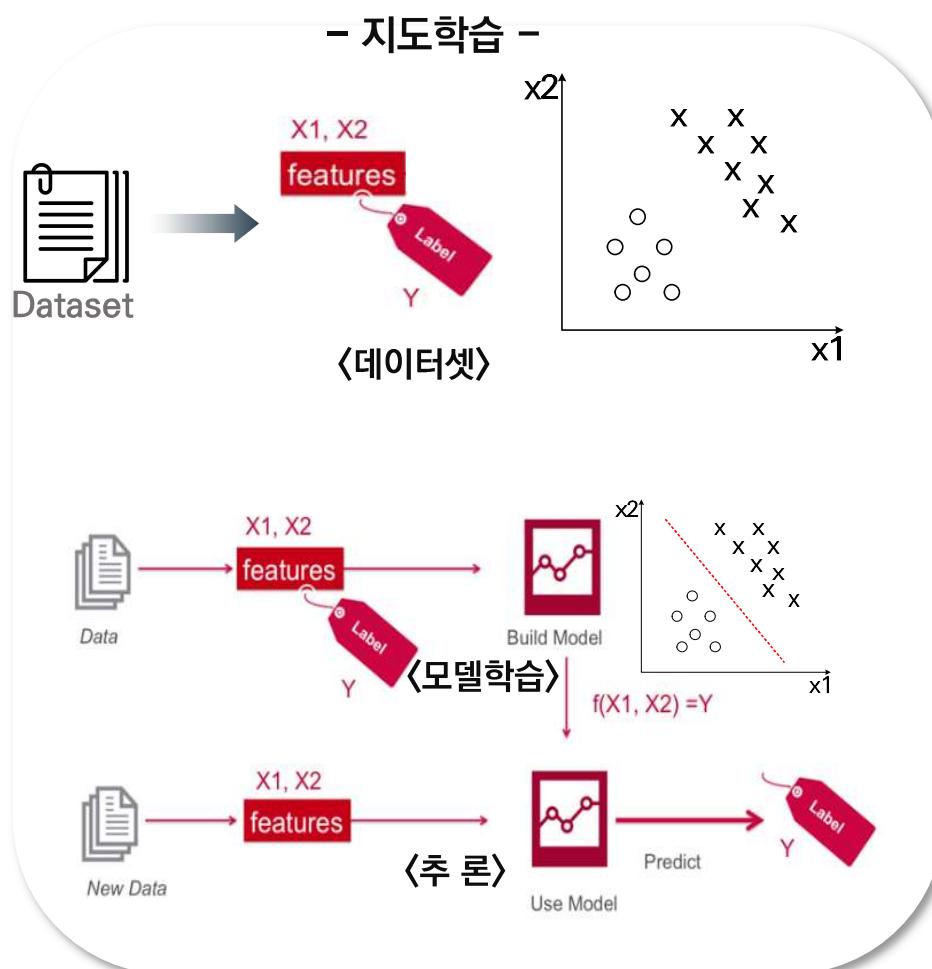


4. 머신러닝 기본이론

4.2 머신러닝 개요

✓ 지도학습 vs 비지도학습

- 지도학습 : 데이터로부터 피처(Input)와 레이블(Output)과의 관계를 함수로 유추하는 방법
- 비지도학습 : 데이터에 레이블이 없을 때 데이터의 관계를 추론하는 방법



4. 머신러닝 기본이론

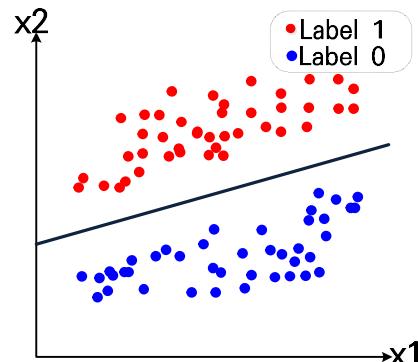
4.2 머신러닝 개요

✓ 분류 vs 회귀

- **분류(Classification)** : 미리 정의된, 가능성 있는 여러 클래스 레이블 중 하나를 예측
 - 이진분류(Binary classification), 다중분류(Multiclass classification)
 - 예 : 유방암 진단, 신용카드 사기 예측, 기계설비 이상진단, 스팸메일 분류
- **회귀(Regression)** : 연속적인 숫자(실수)를 예측
 - 학업 성취도 예측, 부동산 가격 예측, 주식가격 예측

- 분류 -

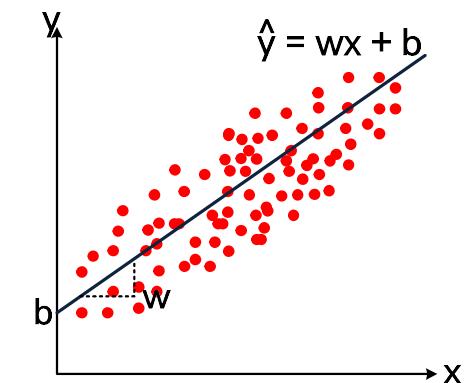
환자나이 (x1)	중앙크기 (x2)	양성 /음성(y)
50	5	1
30	1	0
25	4	1
44	2.3	0
38	1.8	0



〈유방암 진단〉

- 회귀 -

학습시간 (x)	성적 (y)
10	90
8	87
7.5	70
5	40
3	33

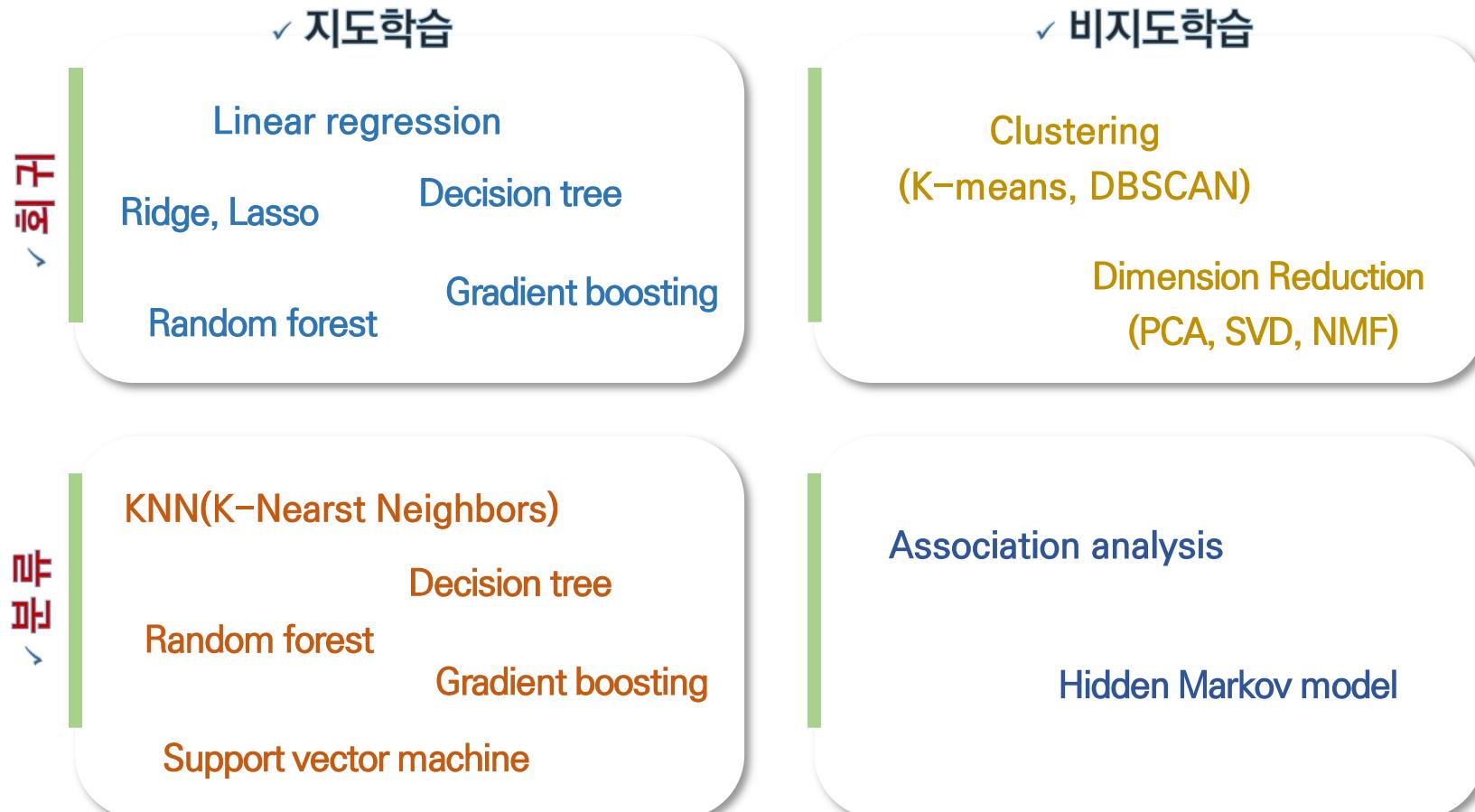


〈학업성취도(성적) 예측〉

4. 머신러닝 기본이론

4.2 머신러닝 개요

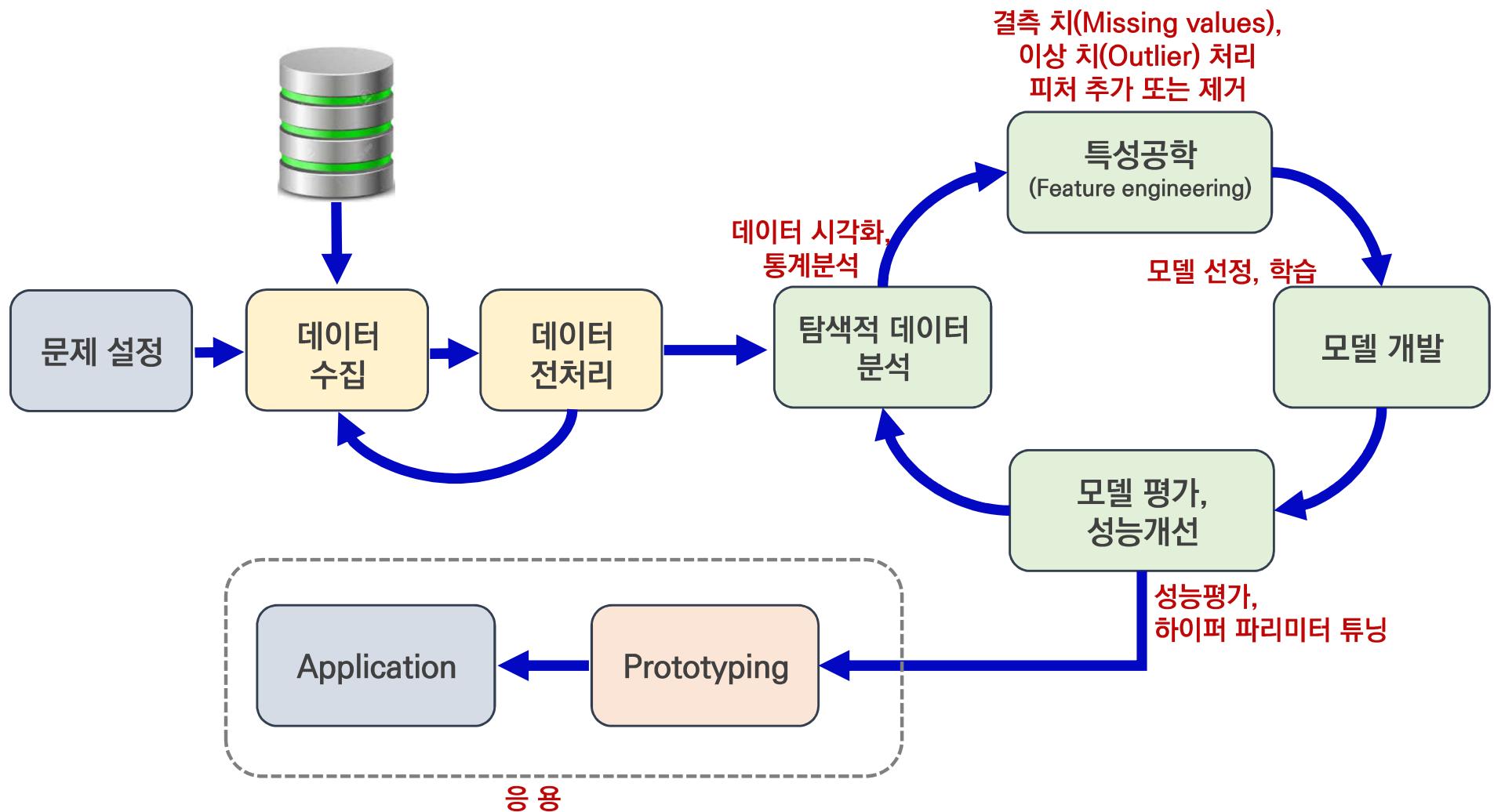
✓ 머신러닝 알고리즘



4. 머신러닝 기본이론

4.2 머신러닝 개요

✓ 머신러닝 모델 개발 절차



4. 머신러닝 기본이론

4.3 모델개발 환경

✓ 파이썬(Python)

- 1991년 귀도 반 로섬(Guido van Rossum)이 발표한 고급 프로그래밍 언어
- 인터프리터식, 객체 지향적, 동적 타이핑 대화형 언어
- 최근버전 : 3.7.4



Python 1: 1994
Python 2: 2000
Python 3: 2008

Rank	Language	Type	Score
1	Python	🌐💻🖱️	100.0
2	Java	🌐📱💻🖱️	96.3
3	C	📱💻🖱️	94.4
4	C++	📱💻🖱️	87.5
5	R	💻	81.5
6	JavaScript	🌐	79.4
7	C#	🌐📱💻🖱️	74.5
8	Matlab	💻	70.6
9	Swift	📱💻	69.1
10	Go	🌐💻🖱️	68.0

⟨IEEE Spectrum language ranking, 2019⟩

4. 머신러닝 기본이론

4.3 모델개발 환경

✓ 파이썬이 데이터분석, 머신러닝에 널리 사용되는 이유

- 오픈소스(Open source)
- NumPy, Pandas, Scikit-learn, Tensorflow 등 다양한 수치해석 및 분석 라이브러리를 보유
- 범용 언어로서 독립 실행 어플리케이션이나 웹 서버 개발이 쉽고 다른 언어와 연동이 쉬움



NumPy



Pandas



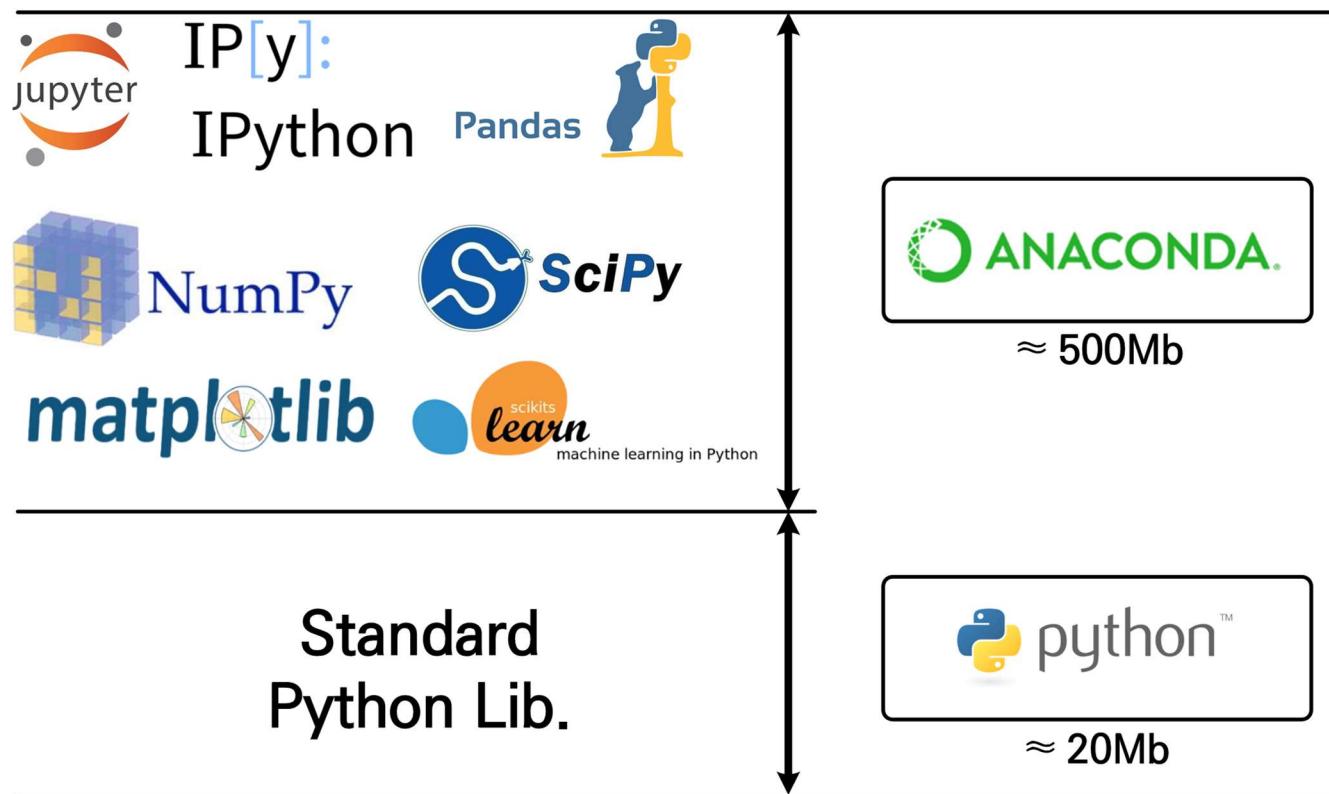
TensorFlow

4. 머신러닝 기본이론

4.3 모델개발 환경

✓ 아나콘다(Anaconda)

- 파이썬과 그와 관련된 라이브러리 관리와 배포를 단순하게 할 목적으로 만들어진 과학 패키지 배포판



감사합니다.

Kwang Myung Yu

www.github.com/sguys99 sguys99@naver.com