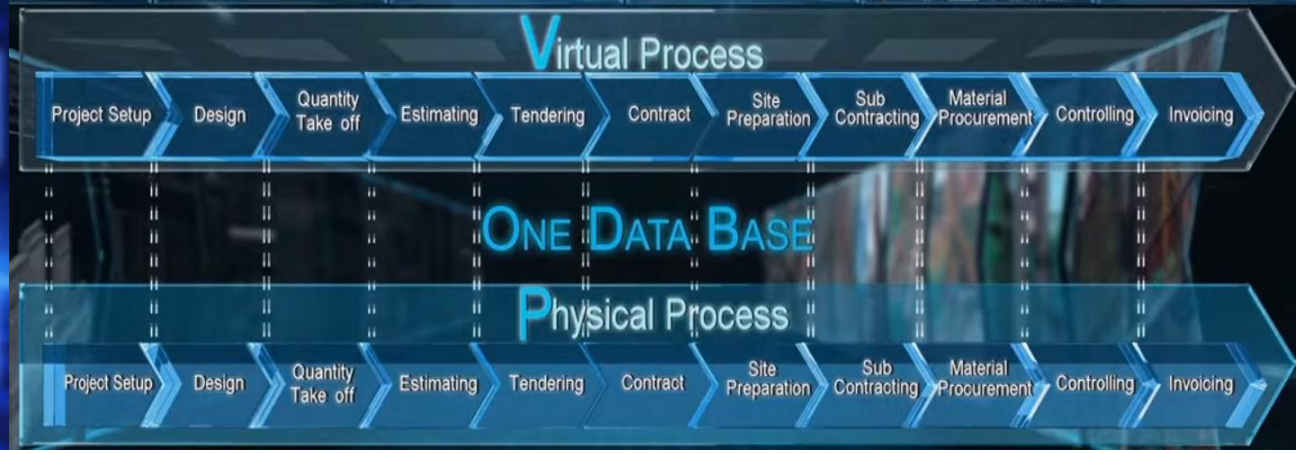


Big Data concept & 활용 Idea

- 빅데이터 이해
- AI, Machine Learning 이해
- 적용을 위한 Idea 개발 프로세스

목차

- **데이터 분석 개념 및 절차**
- **활용사례 및 시사점**
- **활용이 어려운 이유**
- **AI, Machine learning 기본 지식**
- **Open source 활용 및 Demo**
- **데이터분석 아이디어 개발 절차**

빅데이터 개념

통계, DM, AI/ML 데이터 분석
(Open source 활용)

어떻게 활용할 것인가?

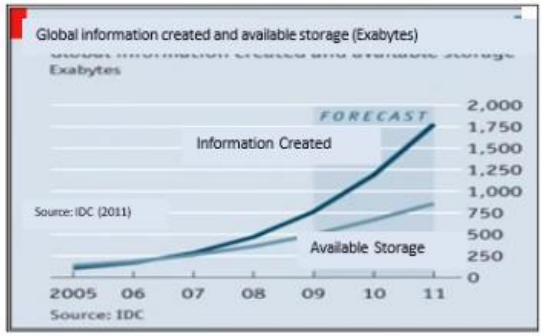
Source : 510. BDA_concept_Idea_AJOU

빅데이터 이해

video clip

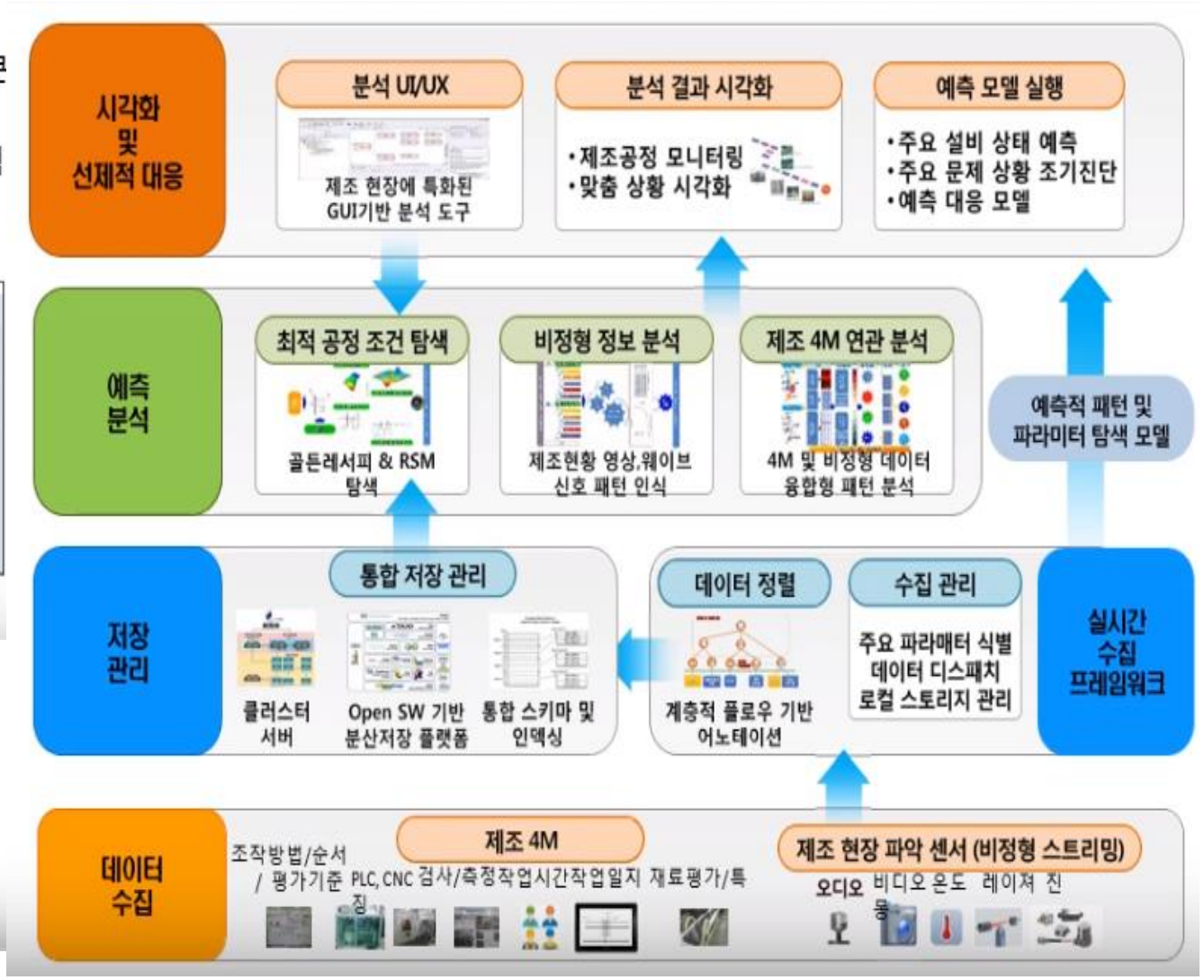
빅데이터란 ?

- 기존 컴퓨팅 기술로는 저장, 관리, 분석이 불가능할 정도로 큰 데이터 집합과 관련 기술, 인력 등을 포괄하는 의미
- IT기술에서 출발했으나 정치, 사회, 문화, 등 삶 전체의 이슈, 혁러다임으로 부각 (Economist, Gartner, McKinsey, NYT)



전세계 데이터는 매년 40% 증가

- 대학병원 미숙아실의 첨단 의료장비들에서 생성되는 스트림 데이터
  - 초당 10,000건의 데이터 발생
- 빅데이터 기술로 실시간 통합, 분석하면 응급상황 조기 예측
  - 기존방식보다 24시간 전에 예측
- 미숙아 사망률 감소는 물론 의사와 간호사 노동 감소



Source : https://youtu.be/oz4Ir3TCcxk

빅데이터(Big Data) 분석으로 할 수 있는 일이 계속 증가하고 있다?

- 빅 데이터 시대
  - ✓ 데이터가 급속한 속도로 증가 : 2020 (40 ZB)
  - ✓ exabyte = 10^{18} 바이트 = 미의회도서관 데이터의 4천 배 크기
  - ✓ 생성된 데이터 중 0.5%만 분석

기업에 적용할 수 있는 것이 무엇이 있을까 ?

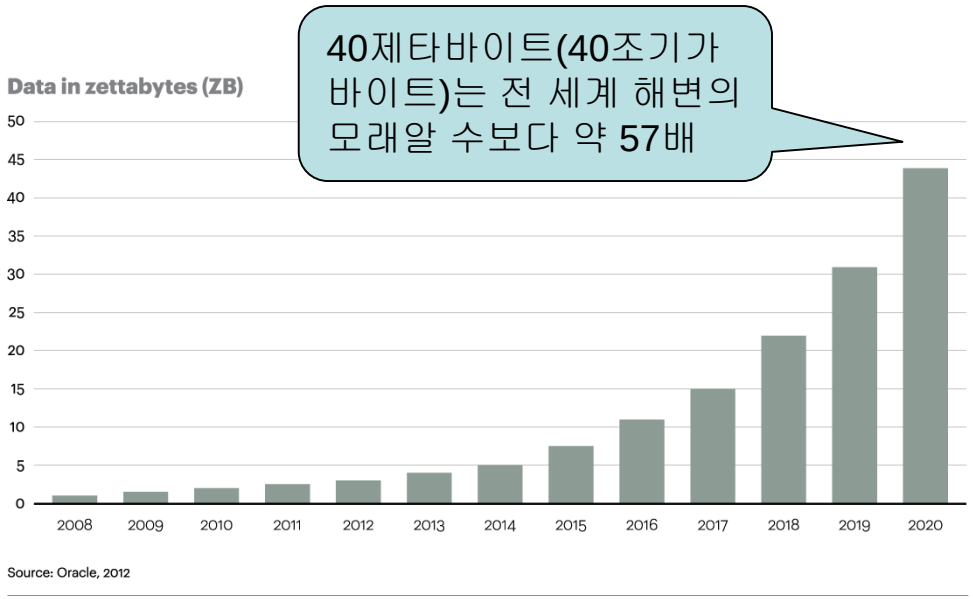
적용을 위한 준비 사항은?

Source : 김종우, 김선태, *경영을 위한 데이터마이닝*, 한경사, 2009

빅데이터 외형적 의미

- **빅 데이터 시대 : 데이터가 급속한 속도로 늘어나고 있다!**
  - ✓ 전 세계의 데이터량은 18개월마다 2배
  - ✓ 지난 5년간 데이터 양은 800% 증가
  - ✓ 전체 데이터의 90%는 2년 이내의 생성
  - ✓ 스마트폰, 소셜미디어, 멀티미디어 콘텐츠 활용 증대
  - ✓ 30억 기가바이트의 데이터가 매일 생산되지만, 이 중 0.5%만이 분석
  - ✓ 2009년 0.9 제타 바이트 였던 데이터량이 2020년 35 제타 바이트 로 44배 규모로 증가할 것을 예측

- **빅 데이터 의미 : 일반적인 DB SW로 관리하기 어려운 정도의 큰 규모의 데이터**
  - ✓ 현재로는 수십 테라에서 향후 페타, 엑사 바이트 정도 크기의 대용량 데이터를 의미
  - ✓ 페타바이트(petabyte) = 10^{15} 바이트
  - ✓ 엑사바이트(exabyte) = 10^{18} 바이트
  - ✓ 미의회도서관 데이터의 4천 배 크기



Source : 김종우, 김선태, *경영을 위한 데이터마이닝*, 한경사, 2009

AI & Deep Learning

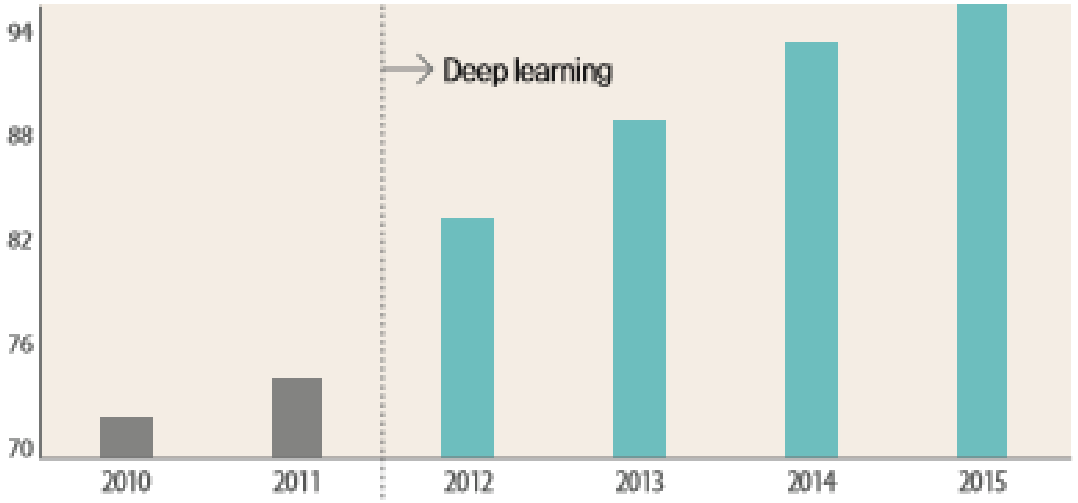
- **Deep Neural Network**
- **2010년부터 시작된 Imagenet Large-Scale Visual Recognition Challenge**

  ✓ **1000가지 물체 종류 중 이미지를 맞추는 문제**

그림7 ILSVRC에서는 컴퓨터로 이미지 안의 물체를 정확히 맞추는 문제를 풀기 위해 전 세계의 연구자들이 치열한 경쟁을 벌인다.

출처: Microsoft

그림8 ILSVRC 역대 우승팀의 정확도. 2015년에는 인간의 인지 능력을 따라잡는 데 성공한다

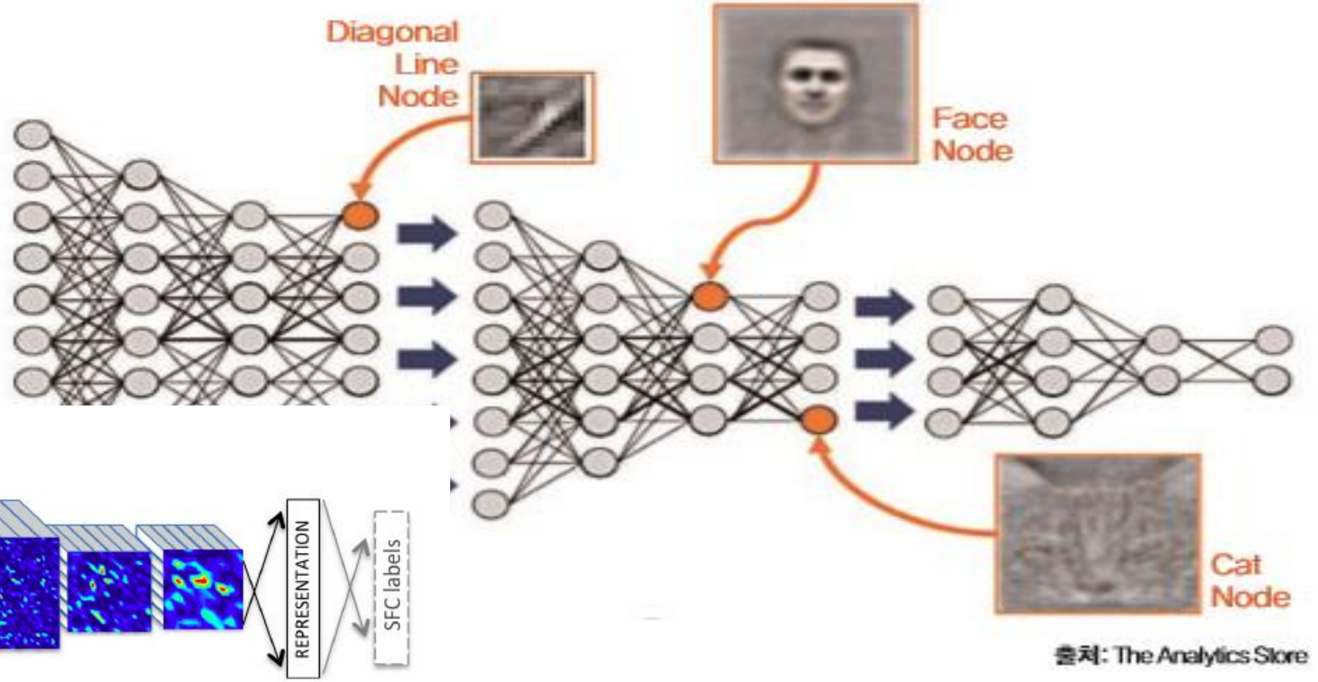


Source : 김종우, 김선태, *경영을 위한 데이터마이닝*, 한경사, 2009

Deep Learning

- ILSVRC 2012
  - ✓ 캐나다 토론토대학의 Geoffrey Hinton 교수팀의 SuperVision
  - ✓ 다음해 DNN Research 설립 : 구글에 인수
- 2015년 구글 95.2% : 사람 95%
- 페이스북의 DeepFace
  - ✓ 2014.6 : 97.35% vs 인간은 97.53%
- 구글 FaceNet
  - ✓ 2015.6 : 99.63%

그림6 구글이 2012년에 발표한 딥러닝 실험 결과. 대량의 이미지 데이터로부터 "대각선" "사람 얼굴" "고양이 얼굴" 등의 개념을 스스로 도출해 내었다.



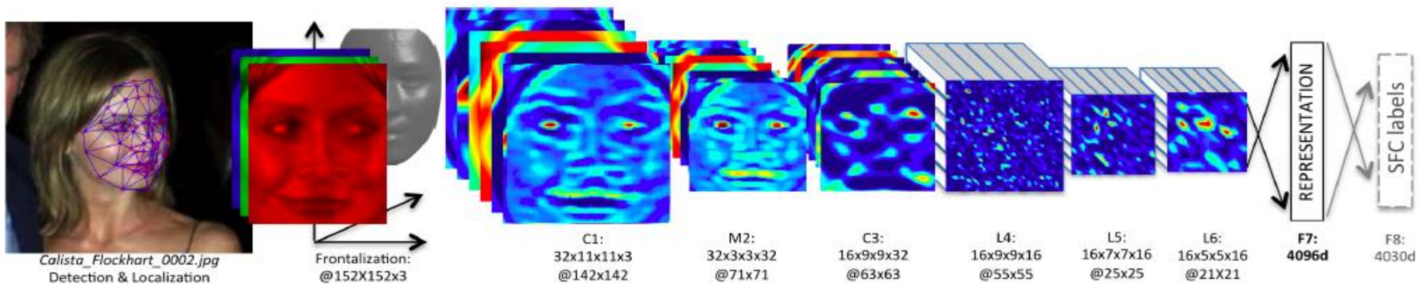
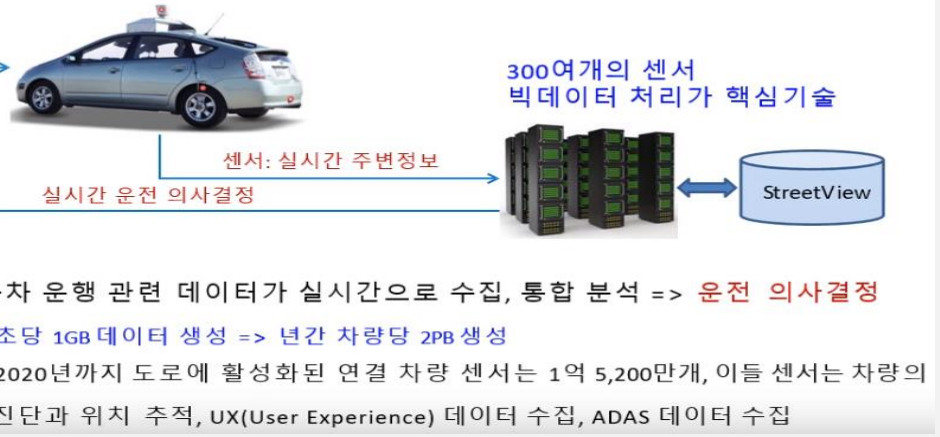


Figure 2. **Outline of the *DeepFace* architecture.** A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate outputs for each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

Source : 김종우, 김선태, 경영을 위한 데이터마이닝, 한경사, 2009

빅데이터 & 융합

초기 검증



- 자동차 운행 관련 데이터가 실시간으로 수집, 통합 분석 => 운전 의사결정
  - 초당 1GB 데이터 생성 => 년간 차량당 2PB 생성
  - 2020년까지 도로에 활성화된 연결 차량 센서는 1억 5,200만개, 이들 센서는 차량의 진단과 위치 추적, UX(User Experience) 데이터 수집, ADAS 데이터 수집

The ScienceTimes
June 06,2019

Source :https://youtu.be/oz4Ir3TCcxk https://www.samsungsds.com/global/ko/support/insights/self-driving-car.html,

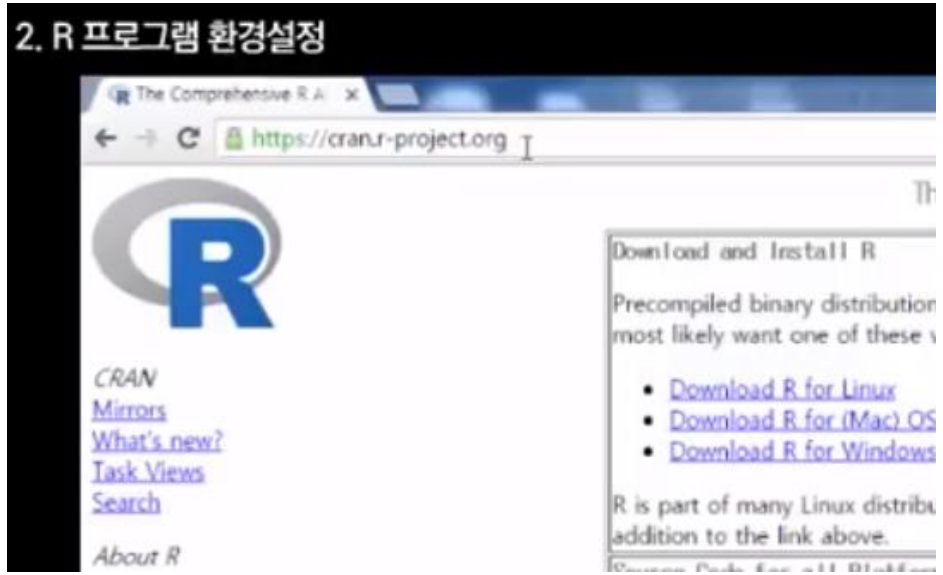
빅데이터 학습을 위한 R

데이터를 구성하고 있는
물리적 하드웨어,
애플리케이션,
소프트웨어를 포괄하는
거대 플랫폼

- 단순한 데이터의 크기가 아님
- 데이터의 형식과 처리 속도 등을 함께 아우르는 개념
- 기존 방법으로는 데이터의 수집, 저장, 검색, 분석 등이 어려운 데이터의 총칭

[자료 출처] https://www.itworld.co.kr

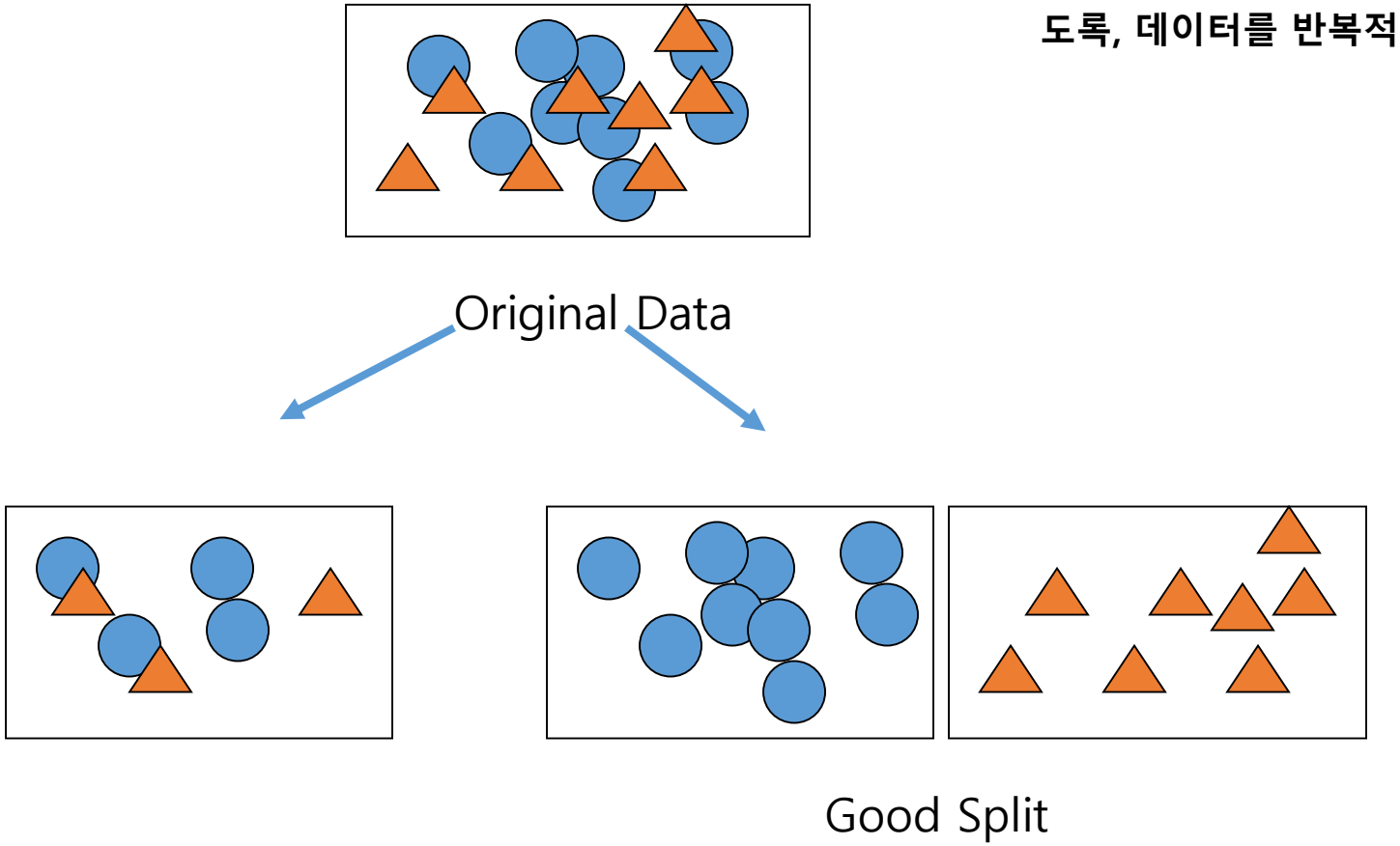
- 탐색어 여론 - 검색어 관련 단어들의 순위별 연관어를 분석하여 상황별 여론을 분석



Source : https://youtu.be/3ihf6VQBTts

개념은 의외로 단순 : Decision Tree 방법

수치형, 범주형 입력변수를 기반으로 분할하기



- 모든 의사결정나무 알고리즘들은 기본 절차에 공통점을 가지고 있음
  - 목표변수 측면에서 부모노드보다 더 순수도(purity)가 높은 자식들이 되도록, 데이터를 반복적으로 더 작은 집단으로 나눈다(repeatedly split)

Data Mining 분류

- 의사결정나무
- 최근접 이웃 기술(Nearest neighbor techniques)
- 인공신경망
- 연결분석 / 회귀분석 모형
- 생존 분석 / 장바구니 분석

- 자동 군집 탐지
- 자기 조직화 지도(self organization map, SOM)

- 연관성 규칙
- 군집화

Source : 김종우, 김선태, *경영을 위한 데이터마이닝*, 한경사, 2009

Big Data tools

1) Data Storage and Management

2) Data Cleaning


3) Data Mining

discovering insight in data

시사점

?

4) Data Visualization

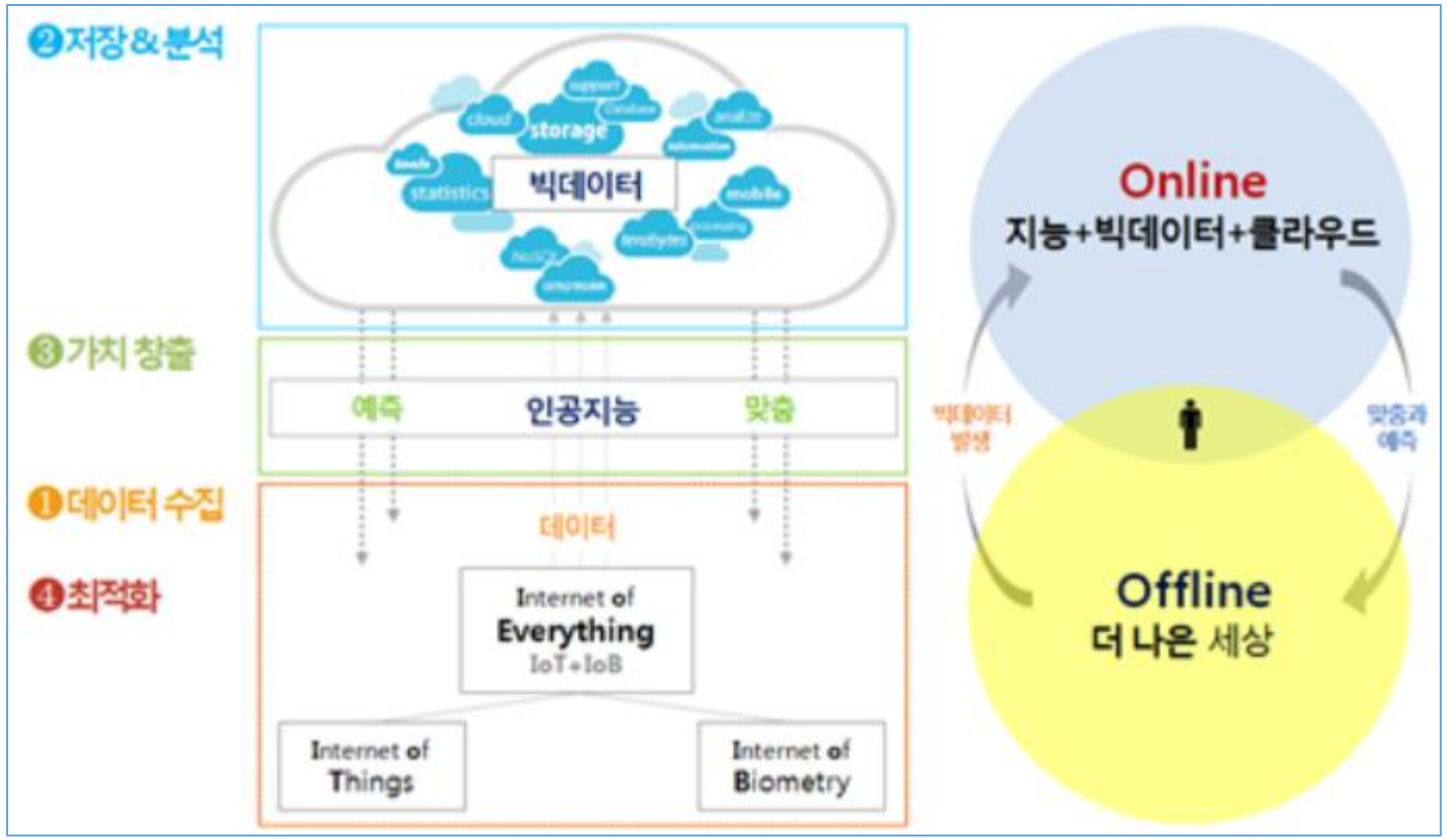
5) Data reporting

Source : Big Data Tools and Technologies | Big Data Tools Tutorial | Big Data Training | Simplilearn, https://youtu.be/Pyo4RWtxsQM

데이터의 중요성

4차 산업혁명은 데이터화 – 정보화 – 지능화 –스마트화 과정을 통하여 가상의 현실화(데이터의 아날로그화)로 더 나은 세상으로 데이터 활용 (데이터의 O2O 모델)

데이터 수집부터 스마트와 프로세스 (KCERN O2O 평행모델)



데이터의 가치 실현

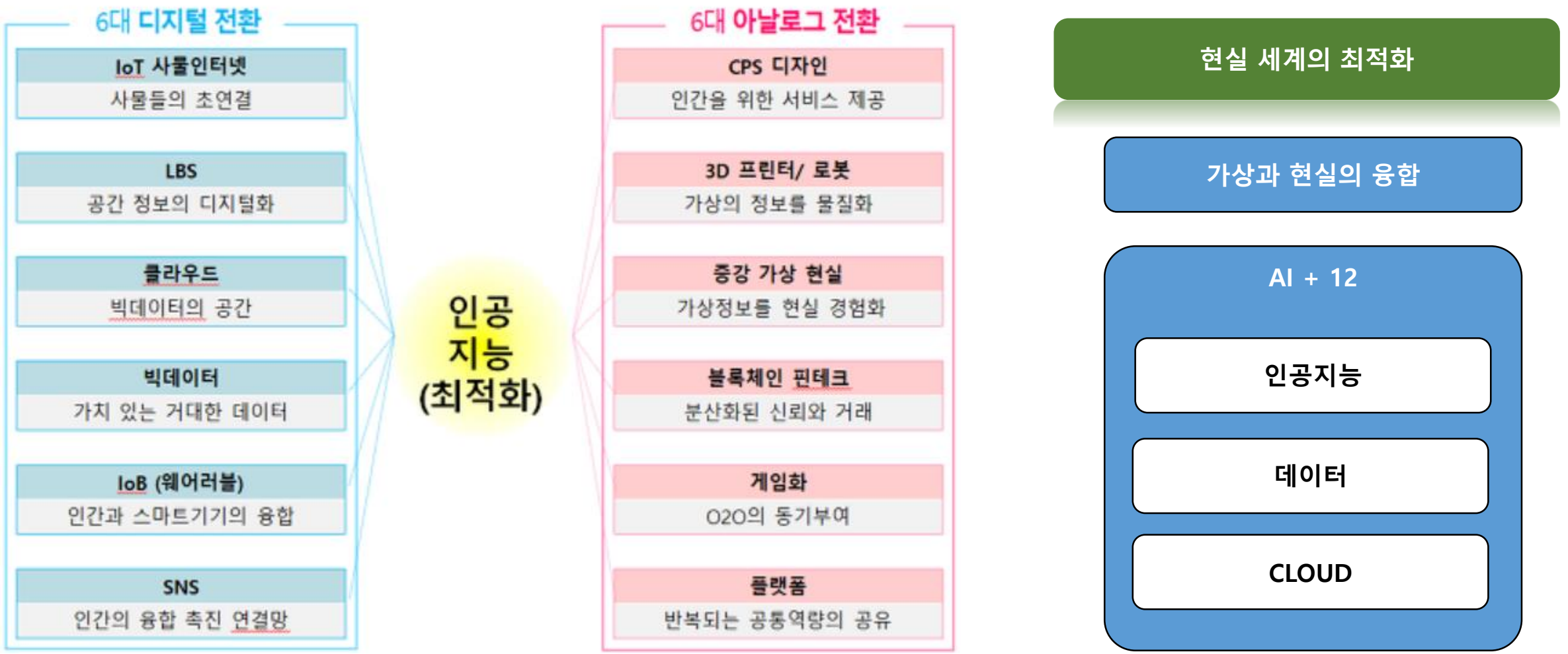
데이터 활용 역량

데이터 이해

Source : http://kcern.org/bbs/board.php?bo_table=kcern_monthly_late&wr_id=122&ipwm=1&pim=1&pim=1

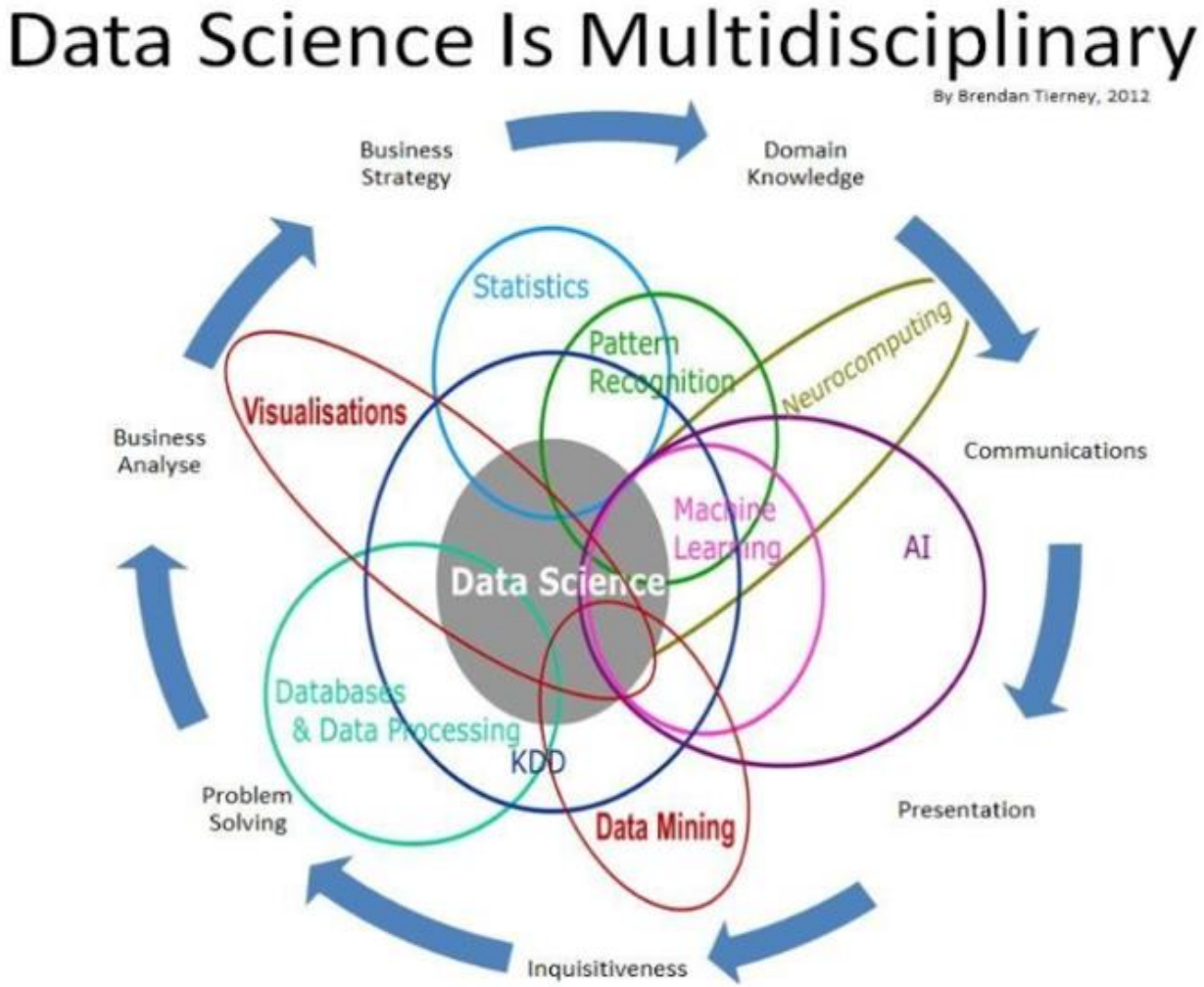
AI + 12 TECH

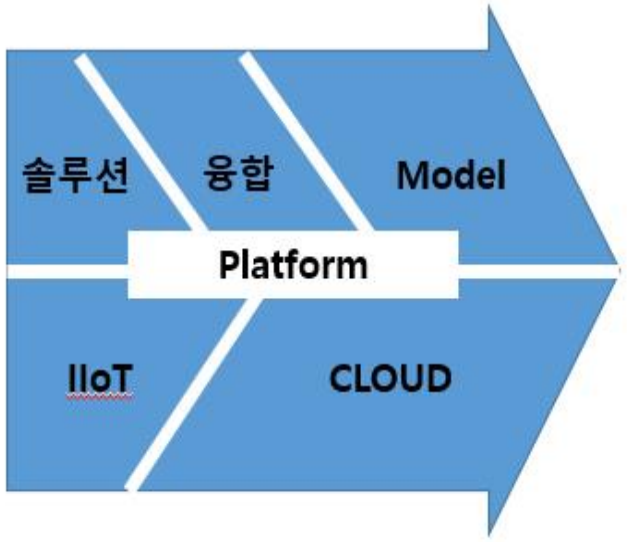
4차 산업혁명은 인간을 위한 현실과 가상의 융합으로 사회문제를 해결하는 가치와 그 것을 가능케 하는 데이터와 방법으로 구성



Source :

데이터 분석 방법은 다양





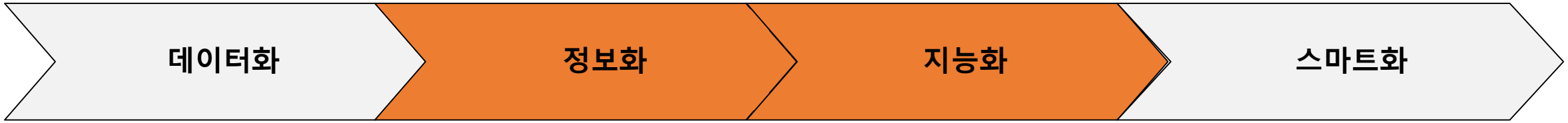
스마트화

데이터의 가치 실현

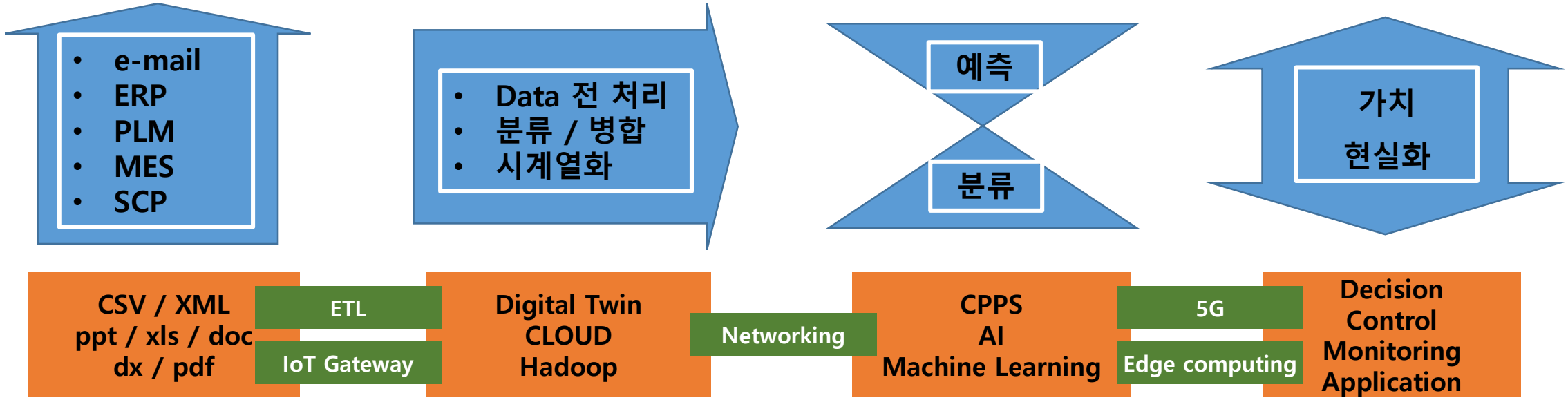
Source : https://www.getfilecloud.com/blog/2018/08/will-machine-learning-replace-data-scientists/#.XCwHh1wzZPY

데이터의 스마트화 프로세스

현실세계의 데이터를 가상세계에서 Digital Twin으로 구조화하고 AI를 활용 여러가지 상황으로 Simulation 하여 최적의 의사결정이 될 수 있도록 가치를 현실화



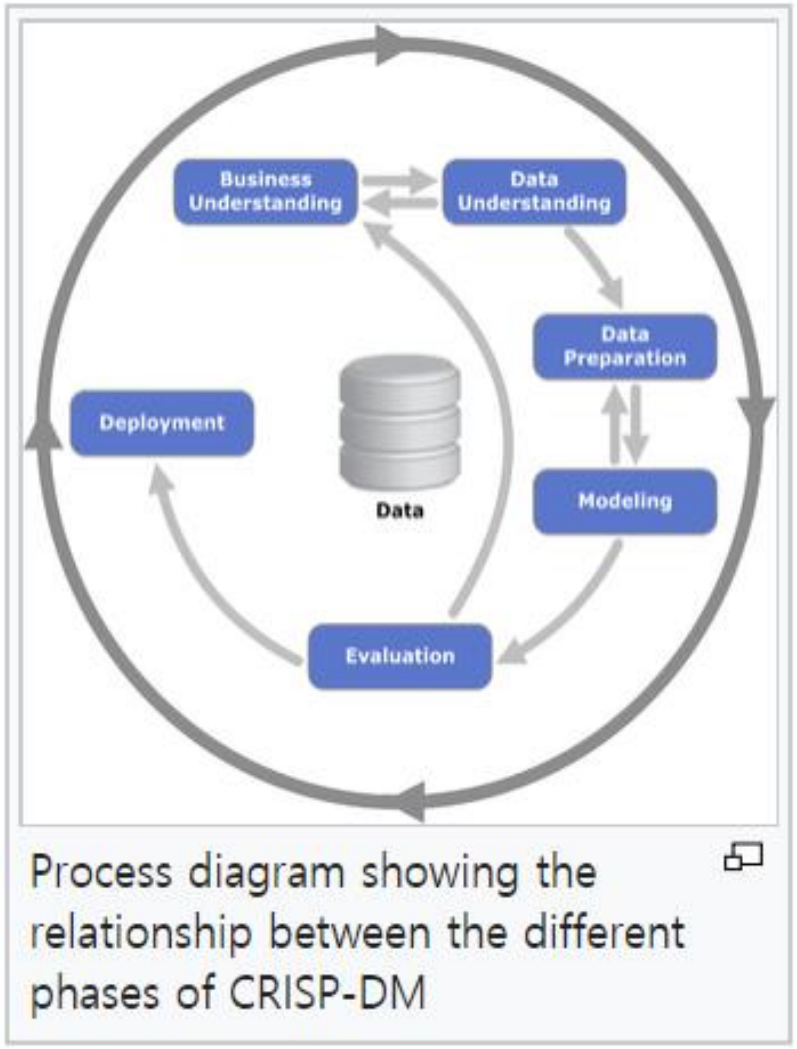
- 현실의 시간·공간·인간을 각종 센서로 데이터화
- 클라우드에 빅데이터를 만드는 정보화 단계
- 분석, 구조화로 미래에 대한 예측
- 개별 사물과 개인에 대한 시공간의 맞춤 정보화
- 가상 세계에서 최적화한 예측과 맞춤의 가치를 현실화



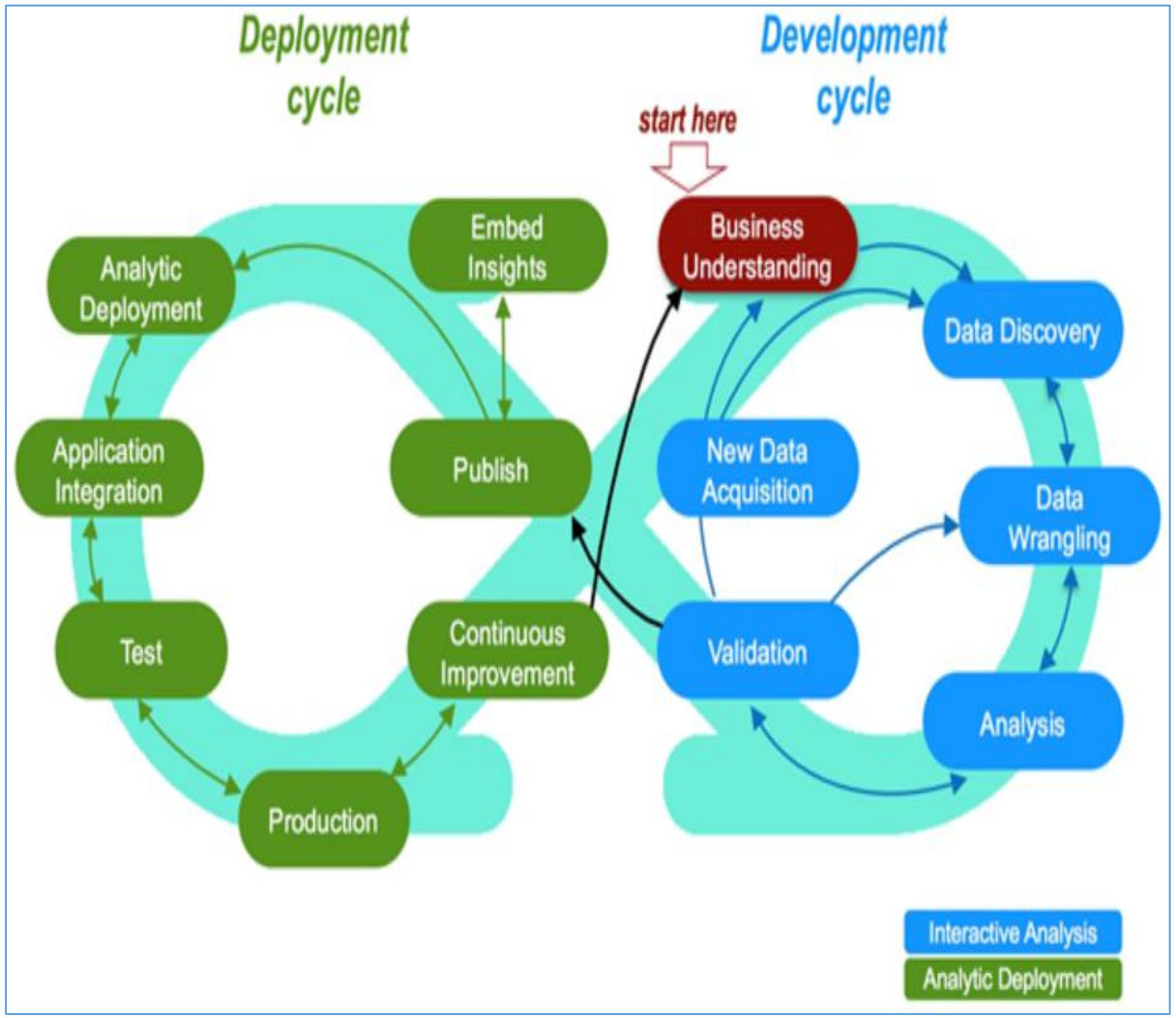
Source :

데이터 분석 절차

Cross-industry standard process for data mining, known as CRISP-DM



Process diagram showing the relationship between the different phases of CRISP-DM



장애 요인

?

Source : https://amsterdam.luminis.eu/2018/08/17/the-forgotten-step-in-crisp-dm-and-asum-dm-methodologies/

데이터화

사물인터넷(IoT)과 생체인터넷(IoB)으로 현실의 시간·공간·인간의 현실 세계를 데이터화 위해, 데이터의 활용 기준에 따라 데이터의 저장, 추출, 변화 및 클라우드로 이동

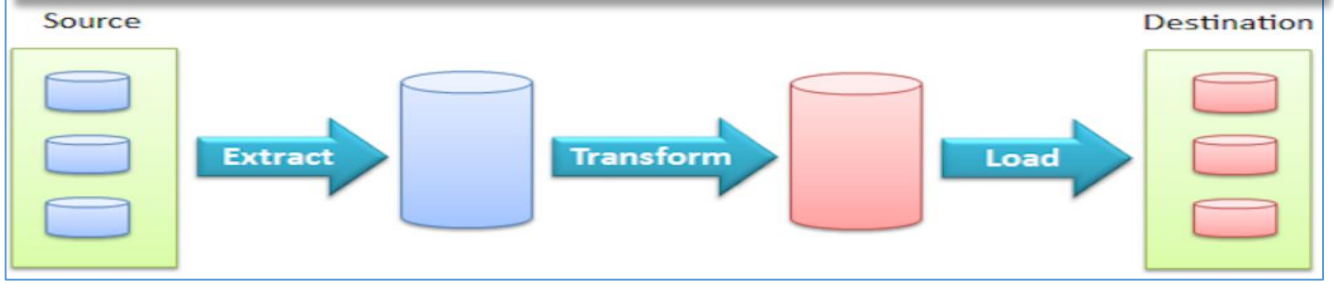
- ERP/PLM
- 2D, 3D
- e-mail, 보고서
- 설비 Data

Data 검색 및 분석을 위한 선택

Process Map Data Flow Diagram	Categories
Source	Dataset widgets

- 원하는 데이터를 찾아서 색인화
- Dataset 간의 상관 관계 형상화
- 사용자들이 의미 있는 widget 활용

ETL (Mapping Table, API)



Data Lake 이동위한 Auto XML, 설계 data 전송을 위한 Automation ML

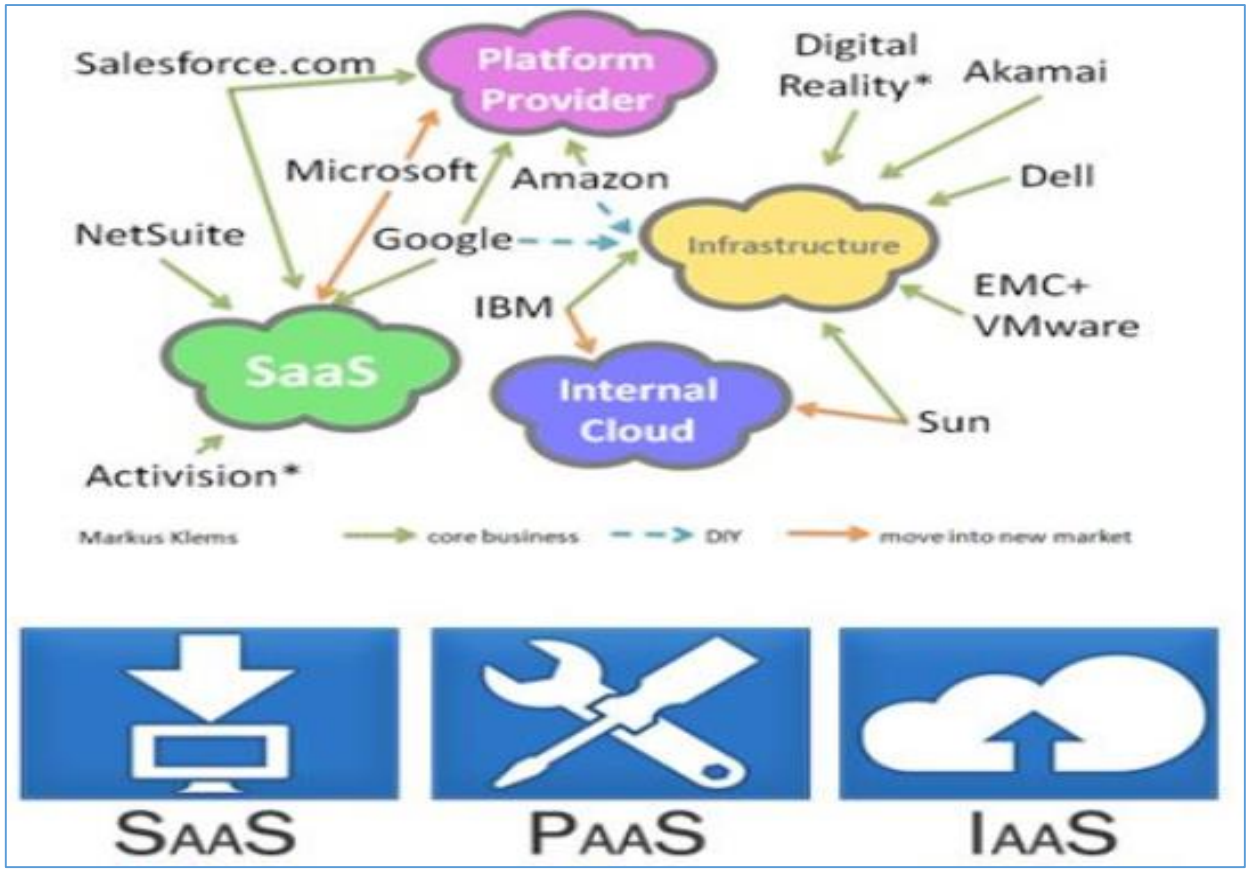
CLOUD
(Data Lake)

- Green Plum
- Hadoop

Source :

정보화 – 클라우드에 정보 Loading

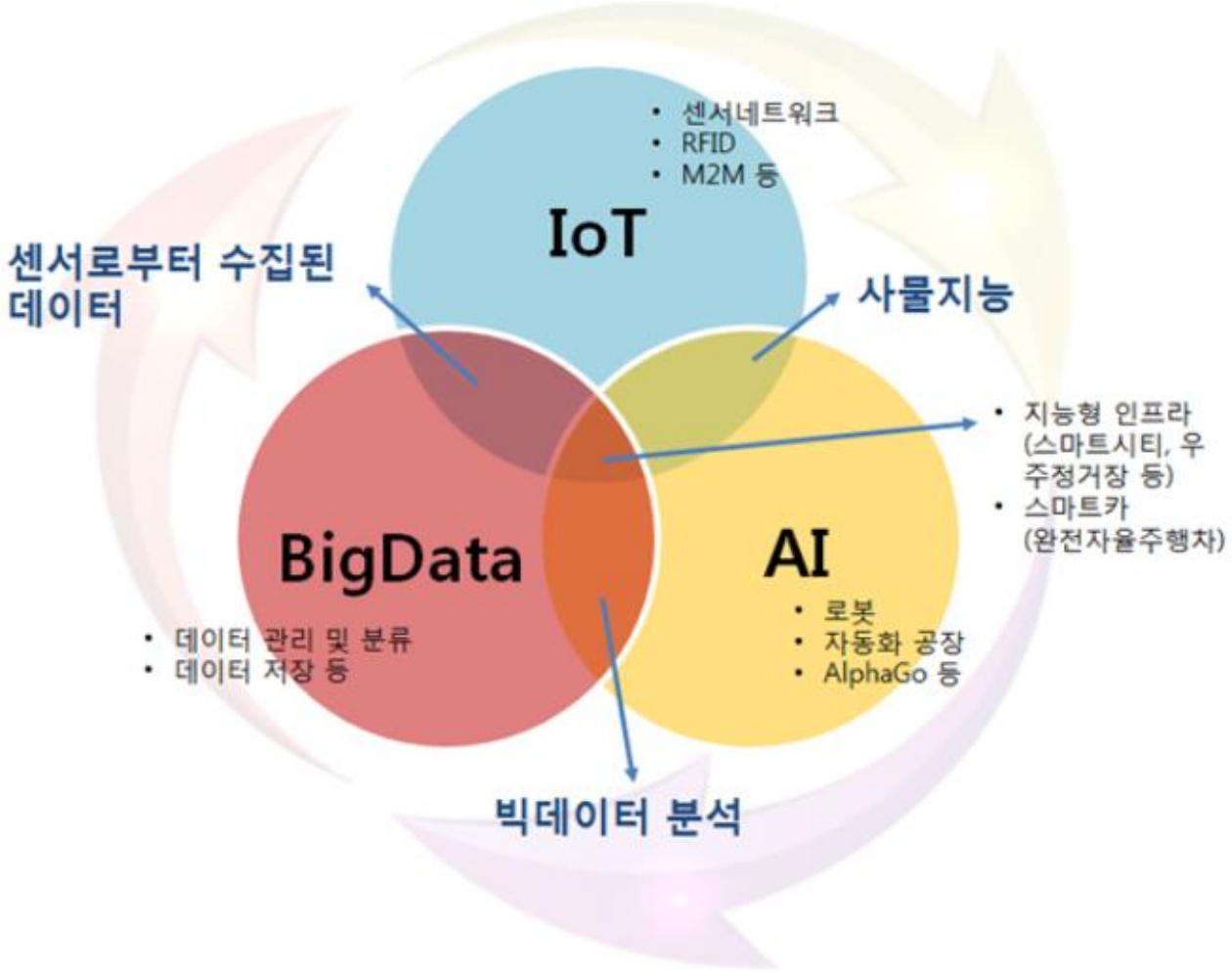
클라우드에 빅데이터를 만드는 것으로, 현실 세계에 흩어진 데이터들이 통합되면서 융합의 가치를 만들 수 있도록 관련된 데이터 간의 네트워크 정보화가 이뤄져 요소 데이터들이 모인 빅데이터가 부분과 전체를 통합하도록 하여 생태계의 입체적 구조가 드러나게 한다.



Source : http://www.software.kr/um/um03/um0305/um030504/um030504View.do?postId=35302&cpage=3
https://slidesplayer.org/slide/11093295/

지능화 - 개념

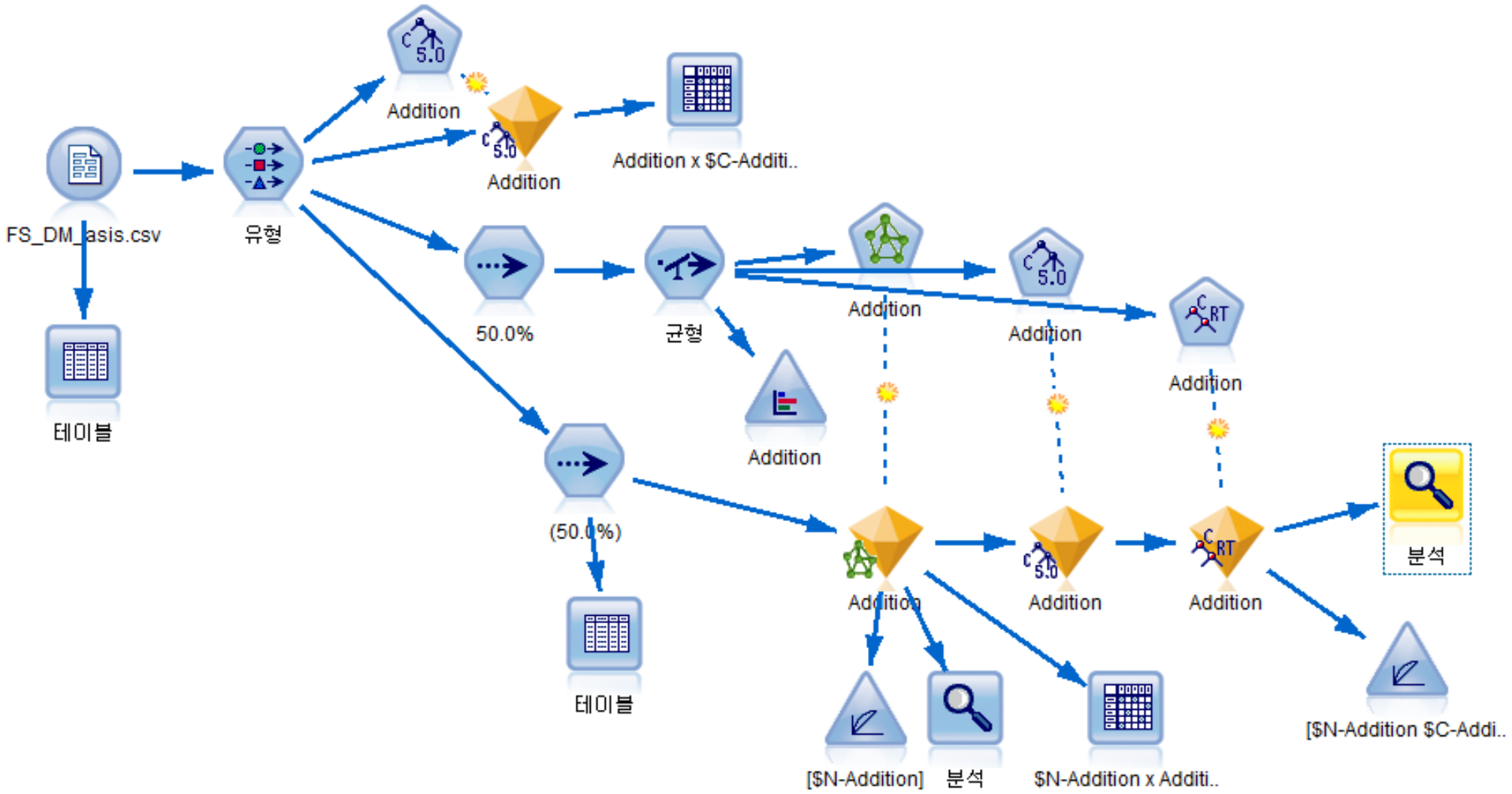
인공지능을 통하여 클라우드에 모인 빅데이터를 분석하고 구조화해 미래에 대한 예측과 개별 사물과 개인에 대한 시공간의 맞춤을 제공하는 것



Source : http://www.itdaily.kr/news/articleView.html?idxno=82425

지능화 – 분석 Modeling case

데이터가 존재하면 Open source 혹은 상용 프로그램을 이용 여러가지 분석을 통해 의미 있는 시사점 도출



Source :

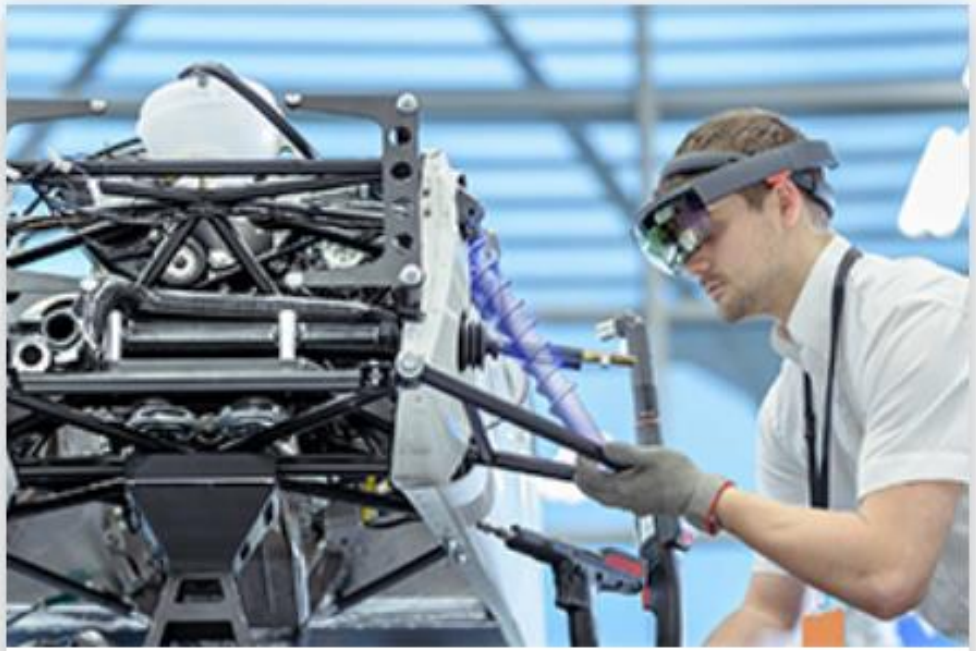
스마트화 - 가상의 현실화

가상 세계에서 최적화한 예측과 맞춤의 가치를 현실화하는 것으로 물리적 로봇, 소프트웨어 로봇으로 행동화와 인간의 의사결정을 지원하여 최적의 결과가 도출될 수 있도록 한다.

혼합현실을 이용한 물리적 디자인의 시각화

Source : https://dynamics.microsoft.com/ko-kr/mixed-reality/layout/ https://blog.skhynix.com/1888

스마트화 – AR/VR



AR Solutions for Manufacturing

Increase technician proficiency, maximize manufacturing output, and reduce assembly errors.

AR Solutions for Service

Reduce service costs and resolution times, while boosting first-time fix rates and customer satisfaction.

Source : https://www.ptc.com/en/products/augmented-reality

목차

Source : 510. BDA_concept_Idea_AJOU

use case summary (1)

- DuPont : R&D 분석 Baseline 자동 추천 : 분산된 R&D 경험과 지식을 자산화, 개발 건별 Formula 및 Recipe의 Base Case 도출 자동화, 개발소요시간 25%, 양산 정확도 10% up
- KONE : 제품을 IoT로 연결 실 사용 정보를 신 제품 개발에 활용, 각 엘리베이터 실시간 작동 상태 , Machine to Machine 작용
- Daimler : 실린더 및 엔진 생산공정에 예방정비 설비운영 적용
- Woodside : Engineering 지식 자산화한 Engineering Assist 시스템(Willow), 생산기지의 운영역량 향상 (IBM Watson, Cognitive Advisor : https://www.ibm.com/watson/stories/woodside/)
- T-mobile : 통신망 정비, 정비효율과 안전 (Cognitive : 각 통신탑 비행경로 학습, visual inspection, 안테나 각도 계측, Drone)
- Mitsubishi : 대형 플랜트 수익성, Project 계약 Risk 평가 관리 / 자동화된 분석/예측 기술 (요건 추출, 관련 정보 제공, 규제/법령 현지사업 난이도, Risk 평가)
- 아우디 잉골슈타트 공장 : (인체공학 : 작업자 부상, 피로 방지)

- The North Face : 인간형 구매 상담 서비스, COMPASS, 고객경험 혁신
- Verizon : 선제적 서비스 제공 고객 경험, 고객별 문의 패턴, 성향 분석 예측기반 Self service, 웹에서 고객별 FAQ Page 개인화, 고객 문의시 Watson Virtual agent 이용 사전 예측하여 담당자 지원 , AI Smart cell center / 고객 경험

제조업

- 설비 안전 검사, 자재 이동 검사
- 석유/화학 : Plant 안전, 유출 감지, 오염원 감지
- 철강사 : 수학적 모형으로 상태 예측하여 최적 운전 조건 도출, 공정 데이터 활용 Predictive Model (Deep Learning) + Control Model
- 자동주차 Robot / 증강현실 활용한 테스트

Source :

use case summary (2)

amazon Deep Learning : https://youtu.be/RonzxMpdTDk

- Eyes & Editors : 추천 (1995)
- 평가 기반 추천 : 행동 이벤트, Rating (매출 35 %)
- 주문 전에 배송 계획 예측 / 물류 KIVA 로봇 /드론 배송
- amazon alexa (음성 서비스 : 음성+머신러닝+CLOUD) 타사에서 해당 플랫폼에서 alexa api 이용 , amazon Go

구글

- 구글의 독감예보 서비스
- 자동 캡션 / 자율 주행차
- 웰드렌즈 기술을 이용 번역 앱
- Google Deep Dream : 그림, 소설, 시, 작곡, 영화 대본

- 뉴욕 시의 빅데이터를 활용한 범죄 감소
- 서울시의 심야버스 노선 결정
- IBM 왓슨(Watson) 의료 지원
- 트위터를 통한 주가 예측 사례
- 디지털 헬스케어

- 제조분야 빅데이터 분석 활용 : https://youtu.be/LM0BEb_cH2w
  - ✓반도체 가상 계측기 : 데이터를 이용 품질 변수 측정기
  - ✓반도체 품질관리 시스템 (이미지 빅데이터)
  - ✓유리기판 생산과정 (비 정형 데이터)

- 공정 Big Data를 활용한 생산성 혁신 (분석 문화, 시스템, 조직)

https://youtu.be/rtuCX5vDnLU

  - ✓반도체 가상 계측기 : 데이터

Source :

use case summary (3)

원료 Lot 관리 : https://youtu.be/rtuCX5vDnLU

- 원료 Lot 관리에 따른 품질 불량율의 Trend : 특정 원료가 동일 기간, 동일 Line에서 불량률이 높음, 원인 파악으로 조치
- 자동 처리 원료 투입 데이터의 오류 (Lot no 잘못)
- 데이터의 관리 오류 인식으로 데이터 처리 프로세스 바꿈
- 원료 입고부터 원료 성적서(COA) 처리 등 변경 (1.2년 소요)

AI기반 분석플랫폼 : https://youtu.be/VSheDrrilmw

- 설비 유지보수, 품질이상 원인 분석, 제품 불량 사전 예측
- 데이터 관리를 위한 프로세스 설명

설비 유지보수 시점예측 : https://youtu.be/VSheDrrilmw

- Hybrid 모델 (일반적 모델+딥러닝 모델)로 설비의 종합 건강도
- 설비의 펌프, 모터의 물리적 특성과 구간 (설비 시작, 가동 중, 완료 시점)별 모델이 다양
- 품질이상 원인분석 : 품질 이상 원인을 찾고 과거 데이터에서 재현성 검증 (사후 불량 원인분석에서 사전불량 예측 체계)

설비 예지보전 : https://youtu.be/rtuCX5vDnLU

- 모터의 진동 센서 대신 전류, 전압 데이터 관리
- 마지막 공정에서 불량이 난 경우 공정관 데이터 관리를 통해 어느 공정에서 불량이 발생한 것인지 파악 가능

Source :

use case summary (4)

해외사례 : https://youtu.be/bY6ZzQmtOzk

- BDO bank : 사기, 부정행위 감지
- Rolls-Royce : 엔진 설계에 적용하여 성능 개선
- Starbuck : 적절한 매장 위치 결정
- 데이터 분석의 활용 범주 : Descriptive, Diagnostic, Predictive , Prescriptive (항공 운임 결정)

Big data application domains : https://youtu.be/bY6ZzQmtOzk

- Healthcare / Education / Marketing / Telecommunications
- Ecommerce / Media & entertainment
- Government

BMW : https://youtu.be/SUIcf2U6pu4

- Virtual world
- 금형, 조립, 설비 예지정비
- 조립, 적절한 체결

- SIEMENS : MindSphere (솔루션 공급사)
- GE : Digital thread (Product lifecycle 상에서 데이터 활용)
- Adidas speed factory : 3D Printer, Robot 이용 고 자동화 및 맞춤생산 / Digital Clone Factory, 맞춤생산 cell (여러 해 동안 시행 착오를 거침)
- PTC : ThingWorx Platform

Source :

시사점 1 : 융합(사용된 업무와 적용 기술) & 활용

적용 업무 / 산업

Marketing

- 고객 만족도
- 제품 추천

Operation Excellence

- 생산성, 품질, 안전 향상
- 유연 생산, 품질 예측관리
- 설비관리
- Risk 관리

Strategy

- 적절한 의사 결정

의료

Infrastructure

제조

통신

안전/공공

석유/화학

보험

건설

광산

적용 기술 : 디지털기술, Robot, Block chain + 인공지능

고도 자동 생산/조립/물류

- Robot
- 3D Printing

데이터 취득/통신/제어

- IIoT, SCP (저전력 통신, 저가)
- CLOUD, LoRa, Big Data 플랫폼

Business Process/Solution

- Digitization IT
- ERP, PLM, MOM / 기준정보

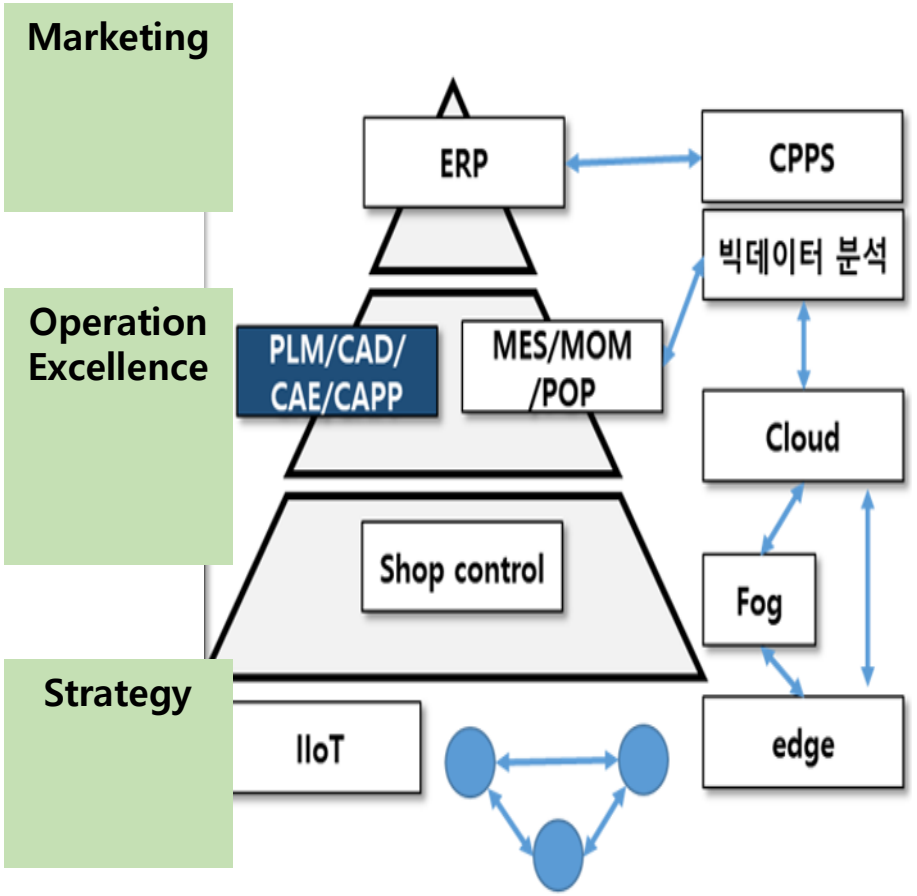
인공지능 (AI, Cognitive Computing)

- Big Data Analytics
- Machine Learning
- 자연어 처리
- ML Vision

Source :

시사점 2 : 왜, 이러한 사례가 중소 제조 기업 내부에서 활용되지 않을까?

기업 내부



기업 외부와 협업

제작

설계

발주처

중소기업 입장

자금, 인력, 인프라 측면에서 기존 사례에서 찾으려면 어려움 ,

- **사례 : 원료 Lot 관리 사례**
- 발주처에 부품 적기 공급

그래서, 기업 환경에 맞는 성공 사례를 직접 만든다고 생각하면 다른 차원의 기회가 될 수 있음 (아직 실제 사례가 없는 use case Idea 개발)

✓ 공정관리, 품질관리, 제작 진도
✓ 수요 예측
✓ 설비 Layout / 최적 계획

Source :

시사점 : 어디에 집중하고 어떻게 접근할 것인가?

빅데이터 프로젝트의 실패 이유 (Gartner 참조)

가치 & 적용 영역의 한계

- 데이터 활용 / AI 새로운 가치
- 융합 : 실물 + 데이터 연계

- 기업 모든 프로세스에 빅데이터 적용 할 수 있는가?
- 어렵다면 무엇이 장애 요인?
- 어느 수준으로 할 것인가? (Global company의 low cost countries 자회사

혁신에 대한 접근

- Platform Holder or participant (System of sytems)
- 고객 가치, 제품 혁신, 일하는 방식 변화 vs 단일문제 해결
- 타사 사례 추종자 vs 사례에서 insight 로 작지만 강한 기업
- 목적 중심 vs 수단중심

추진 기업의 역량

- Process
- 솔루션 활용
- 데이터의 축적 및 관리 수준
- 조직, 인력, 문화, 제도

무엇이 중요한가

Smart Factory 기술

Cognitive computing

- Robot, AR, IIoT, Platform, CLOUD, AI, CPPS, 통신, …

Solutions

?

Source :

빅데이터 Gartner 시사점

전체 빅데이터 프로젝트의 75% 실패 (Gartner : 목적이 아닌 수단을 우선시)

- ✓ 데이터 분석의 목적이 아니라 분석의 수단을 우선시 한 경우
- ✓ 분석할만한 데이터 부족
- ✓ 산업별 Best Practice (use case) 부재
- ✓ 전문 인력 부재

가치 & 적용 영역 선택

혁신에 대한 접근

어떤 목적 / 어떻게 활용

추진 기업의 역량

데이터

중소기업 대안은?

?

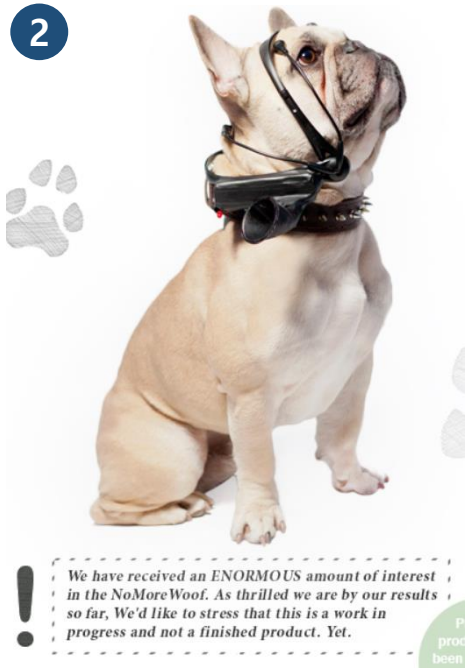
Source :

use case : 어떤 의미가 있는가?

1

https://youtu.be/rVlhMGQgDkY

2



http://www.nomorewoof.com/

3

https://www.zdnet.com/video/big-data-in-action-in-a-mercedes-f1-simulator/

Source :

AI가 융합되면 할 수 있는 것은? 농사를 위한 농기구를 떠올려 보라!

[Boston Dynamics]

고객가치 & 기업가치

프로세스와 연계 & 준비
일하는 방식의 변화

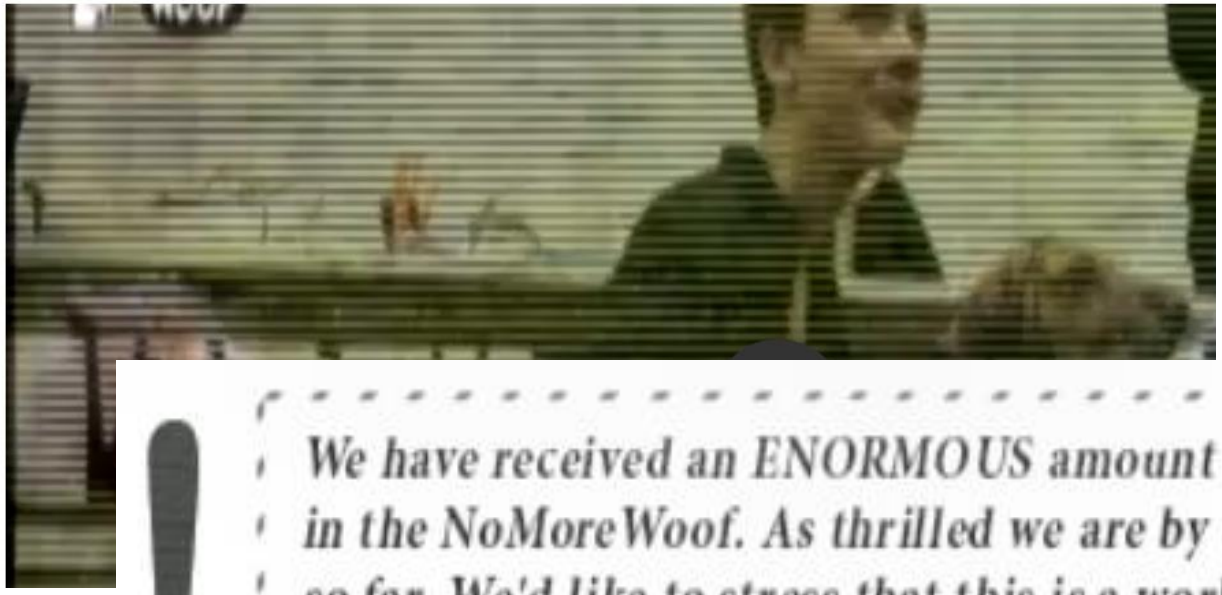
활용 가치

Source : https://youtu.be/rVlhMGQgDkY / https://interestingengineering.com/boston-dynamic-videos-foreshadow-equally-fascinating-and-terrifying-future

빅데이터 분석으로 할 수 있는 것들이 무엇인가 ? 여기서 시사점은?

Talking Dog Device Ready to Hit Market Soon

By LIZ FIELDS Feb. 26, 2014



We have received an ENORMOUS amount of interest in the NoMoreWoof. As thrilled we are by our results so far, We'd like to stress that this is a work in progress and not a finished product. Yet.

We have received an ENORMOUS amount of interest in the NoMoreWoof. As thrilled we are by our results so far, We'd like to stress that this is a work in progress and not a finished product. Yet.

Source : abc NEWS / http://www.nomorewoof.com/

데이터분석의 가치를 어떻게 이야기 하고 있는가 ?

- 승리의 여백을 1000분의 1초 단위로 재는 상황에서 **데이터의 신속한 처리와 분석이** 승리의 관건이다.

- "데이터는 매우 중요하다. 데이터가 없으면 우리는 거의 결정을 내릴 수 없다."라고 해리스가 말한다. "그 데이터는 구조화 될 수도 있고 구조화되지 않을 수도 있다. 운전자가 우리에게 무언가를 말한다면, 우리는 그것을 데이터로 증명한다. 우리는 차량에서 구성 변경을 지원하는 데이터에서 이상 징후를 찾는다."

수많은 적용 사례들이 있음에도

우리 기업에는 왜 쉽게 적용하지 못할까 ?

HPE HPC & AI 글로벌 사례

Source : https://www.zdnet.com/video/big-data-in-action-in-a-mercedes-f1-simulator/

목차

Source : 510. BDA_concept_Idea_AJOU

이렇게 많은 시도가 있는 가운데, 우리 제조기업에 적용은 왜 어려운가 ?

Algorithm 과 AI를 이용한 다양한 적용 사례들이 계속 소개되지만, 업무에 활용은 잘 되지 못하고 있음

❶ **Top 10 Machine Learning Algorithms**

1. Naïve Bayes Classifier Algorithm
2. K Means Cluste
3. Support Vector
4. Apriori Algorith
5. Linear Regressi
6. Logistic Regress
7. Artificial Neura
8. Random Forest
9. Decision Trees
10. Nearest Neighb

❷ **ML 적용 분야**

- Finance
  - Asset allocation
  - Algo trading
- Fraud detection
- Cybersecurity
- eCommerce
- Search
- Manufacturing
- Medicine
- Law
- Business Analytics
- Ad serving
- Recommendation engines
- Smart homes
- Military
- UAV (cars, drones etc.)
- Scientific discovery
- Mathematical theorem
- Route Planning
- Virtual Assistants
- Personalisation
- Compose music
- Write stories

❸ **Top Machine Learning Solutions**

- Alteryx
- AWS SageMaker
- Google Machine L
- Microsoft Azure M
- RapidMiner
- SAP Leonardo
- SAS Visual Data M

❹ **106 STARTUPS TRANSFORMING HEALTHCARE WITH AI**

?

활용이 쉽지 않은 이유

Source : https://www.dezyre.com/article/top-10-machine-learning-algorithms/202 https://yalantis.com/blog/trends-in-machine-learning-in-past-years/

Smart Factory use case

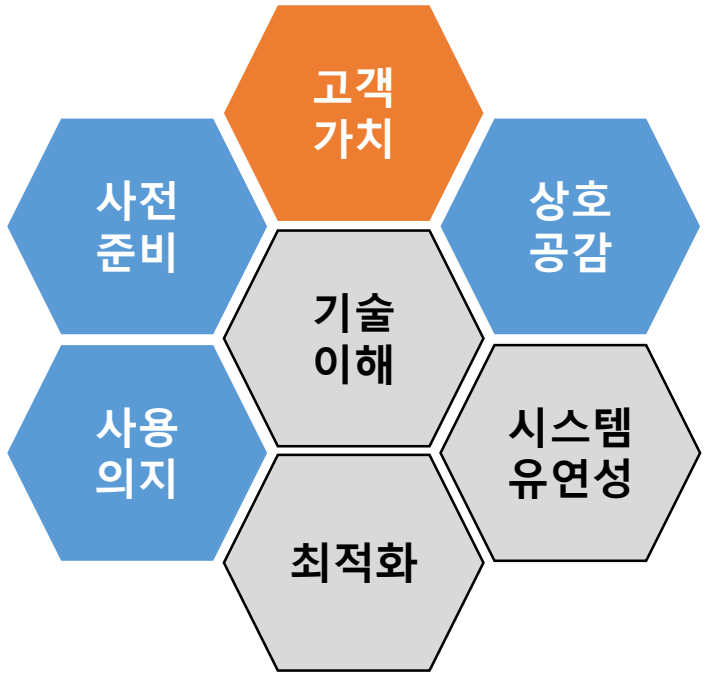
사례

현장 자동화 및 로봇 도입

생산 최적화 등 디지털 과제

데이터 분석 과제 및 Platform 구축

시사점

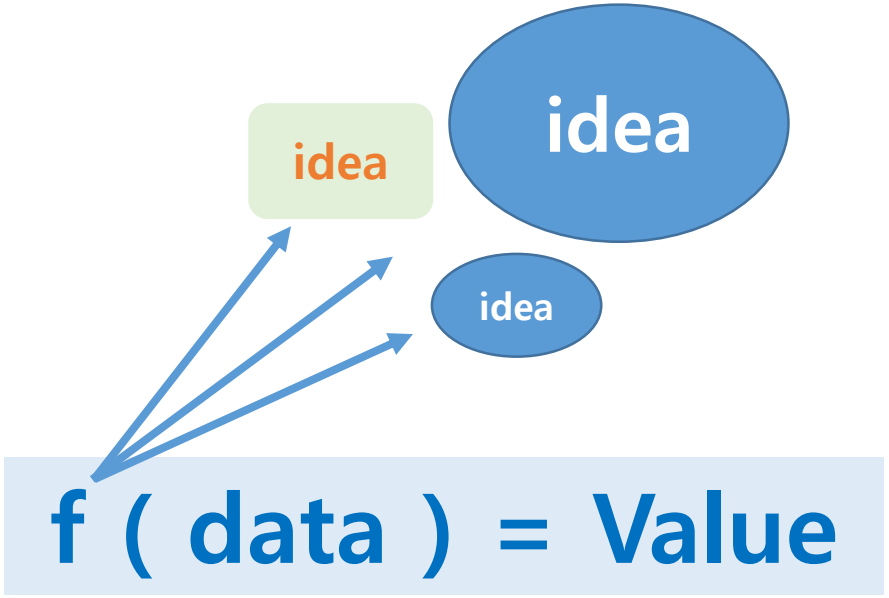


Source : 221. Smart_factory_case_study.pptx

데이터 그리고, 가치에 대한 이해 요약

데이터

데이터 자체가 가치를 보장하지 않는다!



운영기준, 역량, ...

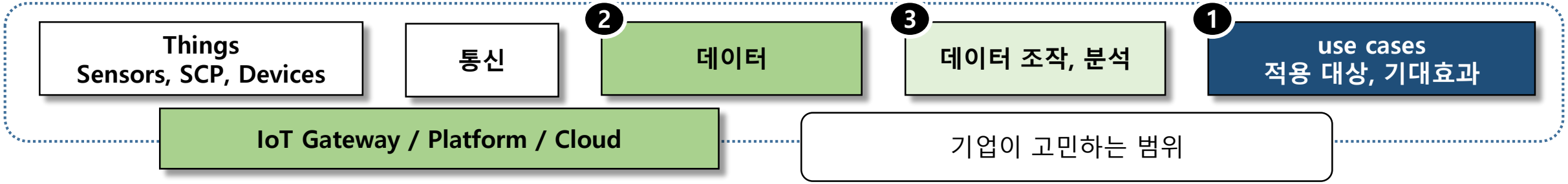
가치

Source : ghSeo

왜 가치 실현이 쉽지 않는가 ?

사용해야 할 이유와 필요성을 모르기 때문이다.

고객(기업)이 기대사항에 대한 솔루션 공급사의 제시하는 방식에서 차이가 있음



중요도

- 사용자 입장 중요도 순서 : 1번이 없으면 나머진 불 필요
- 솔루션 공급사 (3번 강조) : 3 ➔ 2 ➔ 1
- 기업에 3번 전문조직이 있음에도 활성화 안되는 이유

해결 방안

- 데이터 사용자가 use case 발굴을 손쉽게 할 수 있어야 하며, 거기서 Idea 도출
- Pilot으로 검증하여 확신을 갖고
- 본 과제로 구체화 필요 (솔루션 공급사, 전문가 협업)

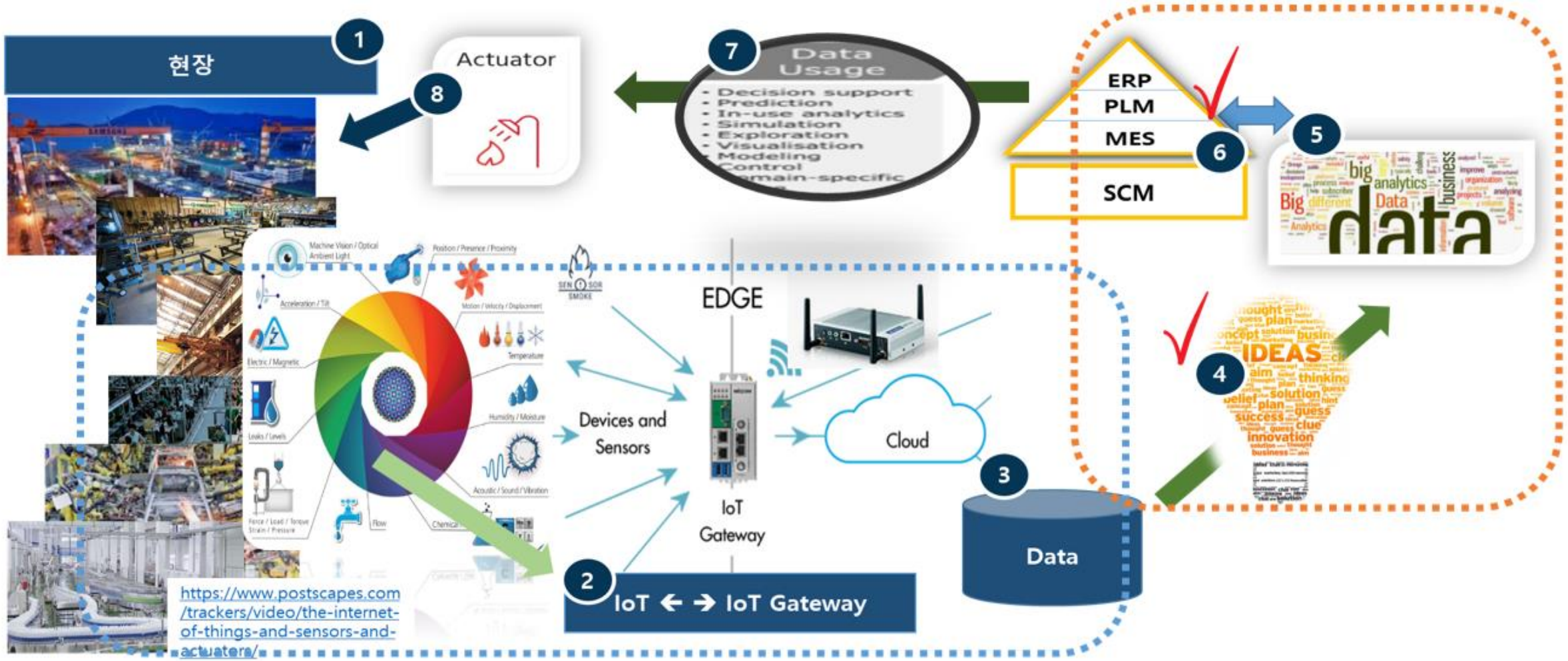
비용 부담 없이 use case 도출

사용자 편의성

Source :

중요한 것은 AI/ML, Cloud 및 솔루션이 아니라 그 것을 활용한 새로운 가치 이다.

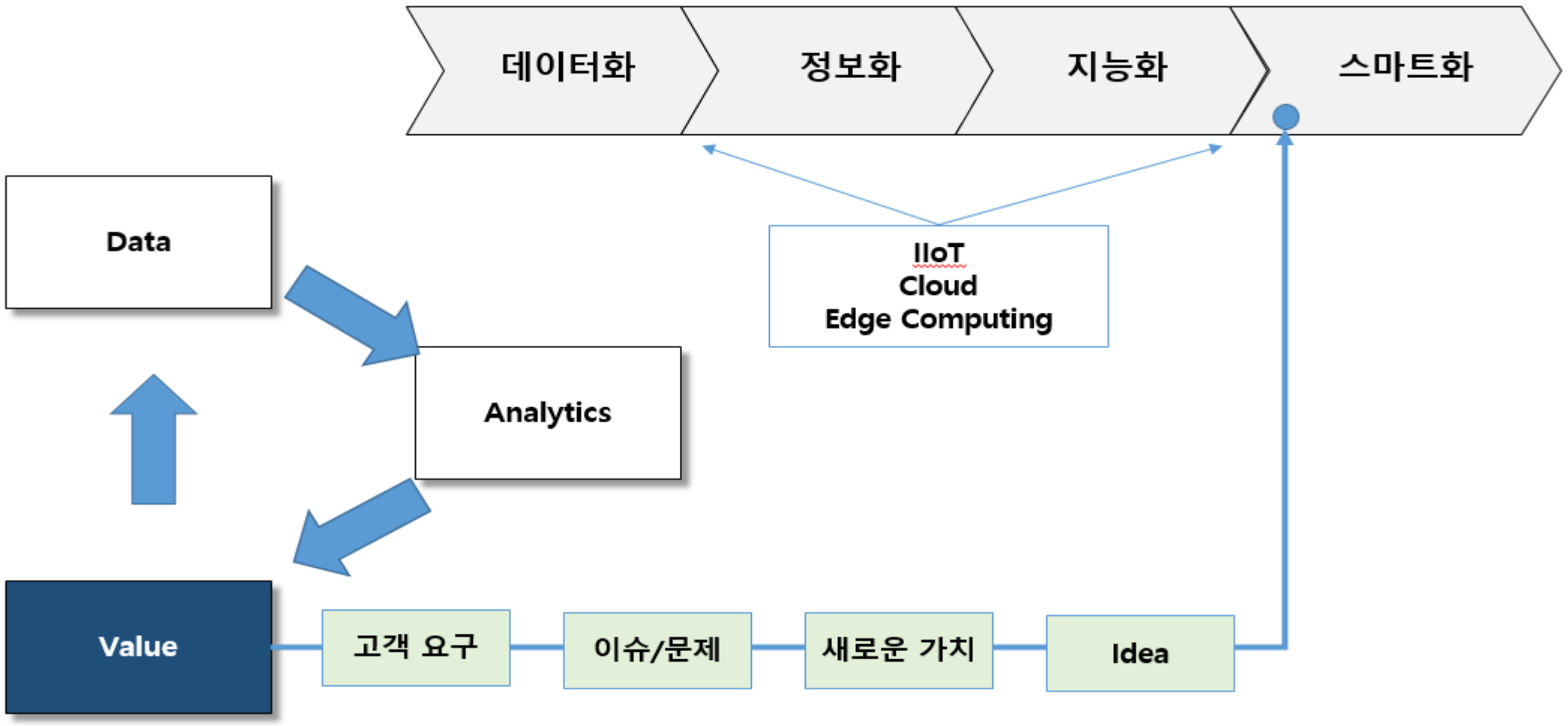
데이터 분석 flow



https://marketingland.com/whats-big-idea-3-fundamentals-successful-digital-creative-153747

Source :

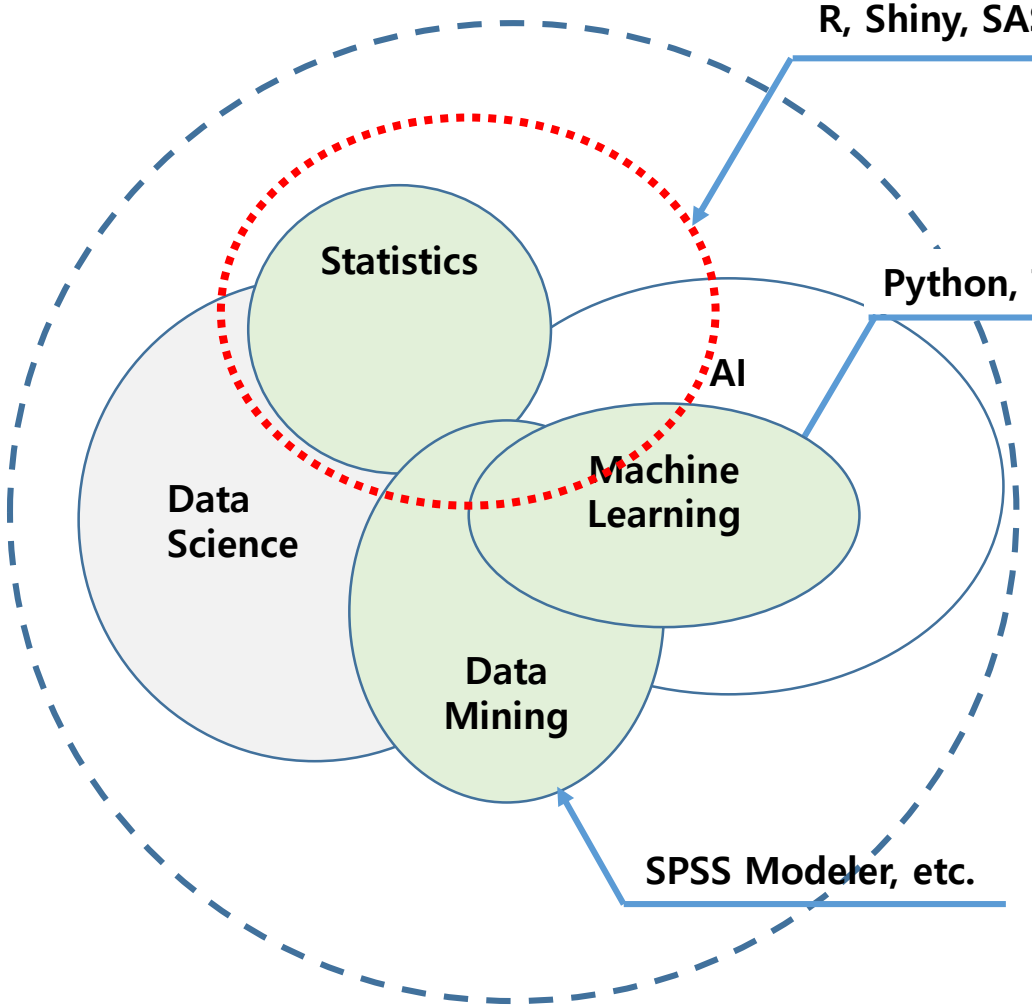
가치가 만들어지는 절차와 데이터의 흐름/작동하는 순서는 다르다!



Source :

목차

데이터의 가치화를 위해 분석과 솔루션에 대한 개념 이해가 중요하다.



통계 | Data Mining | AI/ML

분석에 사용하는 Tool

R, python, SAS, SPSS Modeler, alteryx, Apache Spark, QlikView, splunk>, rapidminer

Source :

통계 기본을 정확하게 이해하고 활용 범위를 확대하면서 기대효과 검증 방식으로 접근

개념에 대한 정확한 이해 부족으로, 적용 가능한 대상이 많음에도 적용하지 못하고 있음.

- 공분산
- 독립 t-Test (일표본, 대응표본, 독립표본)
- ANOVA (one-way, two-way, MANOVA)
- 요인분석 (PCA/FA)
- 상관분석
- 신뢰도 분석
- 회귀분석 / 다중 회귀분석
- 로지스틱
- 판별분석
- 군집분석
- 경로분석 / 구조분석

개념 이해 → **적용 대상의 확대** → 분석 결과 해석

- 기존에 사용하는 생산, 품질 외 거의 모든 프로세스에 사용 가능
- 적용을 위한 새로운 시도 필요

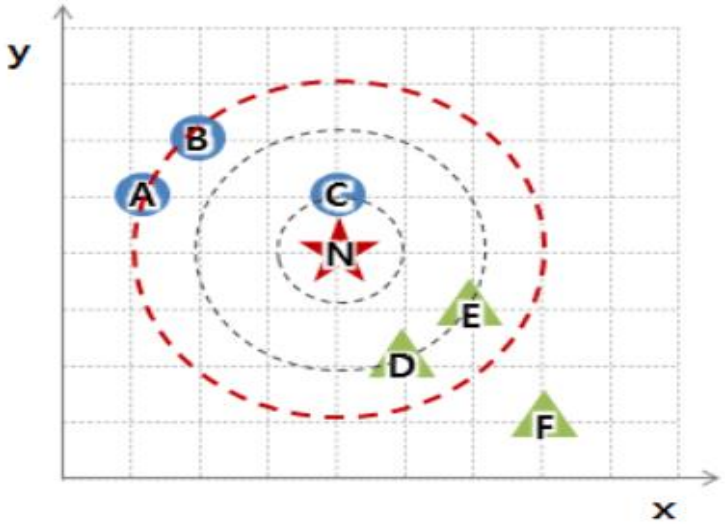
- 전제 조건의 이해
- 분석 결과의 올바른 해석

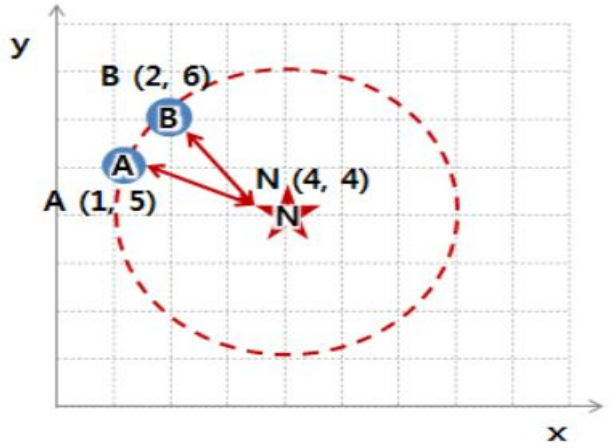
- 솔루션 도입이 우선이 아니라,
- 적용 대상을 넓혀서 기대효과 여부를 검증하는 것이 우선
- 그러기 위해 사용이 용이한 Open source 활용
- 기대효과 가시화되면, 업무에 지속적인 적용을 위해 전문 솔루션 검토 시작

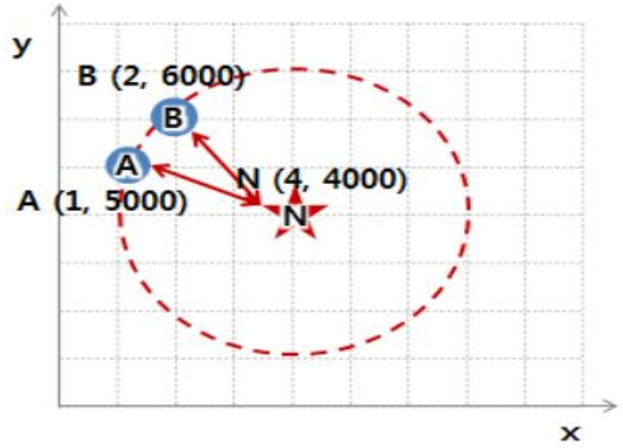
Source :

왜 통계인가? 직관적으로 인식하는 것과 분석을 통한 결과는 다르다.

Data Mining : k-NN (k-최근접 이웃 알고리즘)







Unit	Ax	Ay	Nx	Ny	유클리드 거리
$	1	5	4	4	3.162
	Bx	By			
	2	6	4	4	2.828

Unit	Ax	Ay	Nx	Ny	유클리드 거리
₩	1	5000	4	4000	1,000
	Bx	By			
	2	6000	4	4000	2,000

- 변수의 단위에 따라 분석 결과 상이
- 데이터의 표준화, normalization 필요
- 최적의 k 찾기

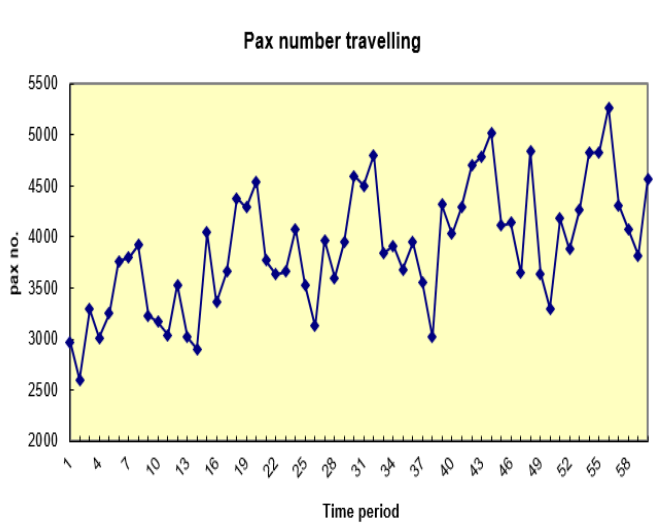
Source : https://kkokkilkon.tistory.com/14

계절지수 고려

일반 회귀분석

Number of Passengers

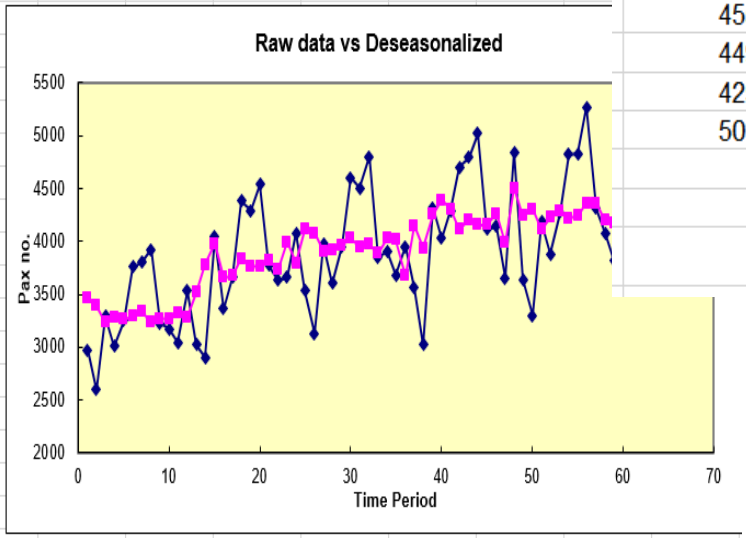
Month	1st	2nd	3rd	4th	5th
Jan	2970	3020	3530	3560	3640
Feb	2600	2900	3130	3020	3300
Mar	3290	4040	3970	4320	4180
Apr	3010	3360	3600	4030	3880
May	3250	3660	3950	4290	4270
June	3760	4380	4600	4700	4820
July	3800	4290	4500	4790	4830
Aug	3920	4540	4800	5020	5270
Sept	3230	3770	3840	4110	4310
Oct	3170	3630	3910	4140	4080
Nov	3040	3660	3680	3650	3820
Dec	3530	4080	3950	4840	4560



계절지수 + 일반 회귀분석

Number of Passengers

Month	1st	2nd	3rd	4th	5th	Mo. Avg	S.I.
Jan	2970	3020	3530	3560	3640	3344	0.86
Feb	2600	2900	3130	3020	3300	2990	
Mar	3290	4040	3970	4320	4180	3960	
Apr	3010	3360	3600	4030	3880	3576	
May	3250	3660	3950	4290	4270	3884	
June	3760	4380	4600	4700	4820	4452	
July	3800	4290	4500	4790	4830	4442	
Aug	3920	4540	4800	5020	5270	4710	
Sept	3230	3770	3840	4110	4310	3852	
Oct	3170	3630	3910	4140	4080	3786	
Nov	3040	3660	3680	3650	3820	3570	
Dec	3530	4080	3950	4840	4560	4192	
						3896.5	



deseason 이용

Forecast	Actual	Deviation
3827.7	3860	-32.3
3436.7	3700	-263.3
4570.4	5100	-529.6
4144.2	4620	-475.8
4519.6	4840	-320.4
5201.7	5230	-28.3
5211.1	5380	-168.9
5547.9	5860	-312.1
4555.6		
4495.5		(2,131)
4256.0		
5017.4		
	Bias	-266.33
	MAD	266.33
	MSE	100,823

ML

단순 회귀

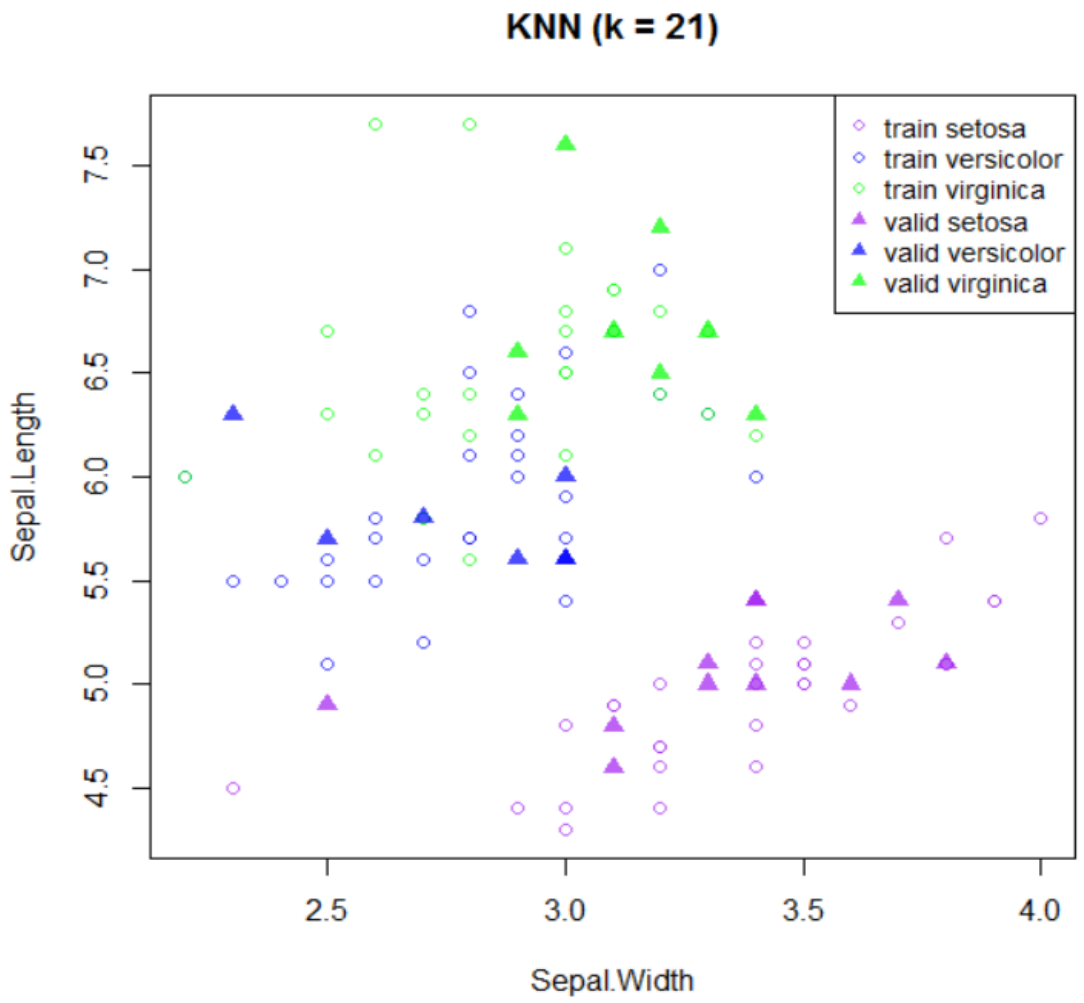
	Forecast	Actual	Deviation
61	4558.2	3860	698.2
62	4579.2	3700	879.2
63	4600.2	5100	-499.8
64	4621.2	4620	1.2
65	4642.2	4840	-197.8
66	4663.2	5230	-566.8
67	4684.2	5380	-695.8
68	3255.8	5860	-2604.2
			(2,986)

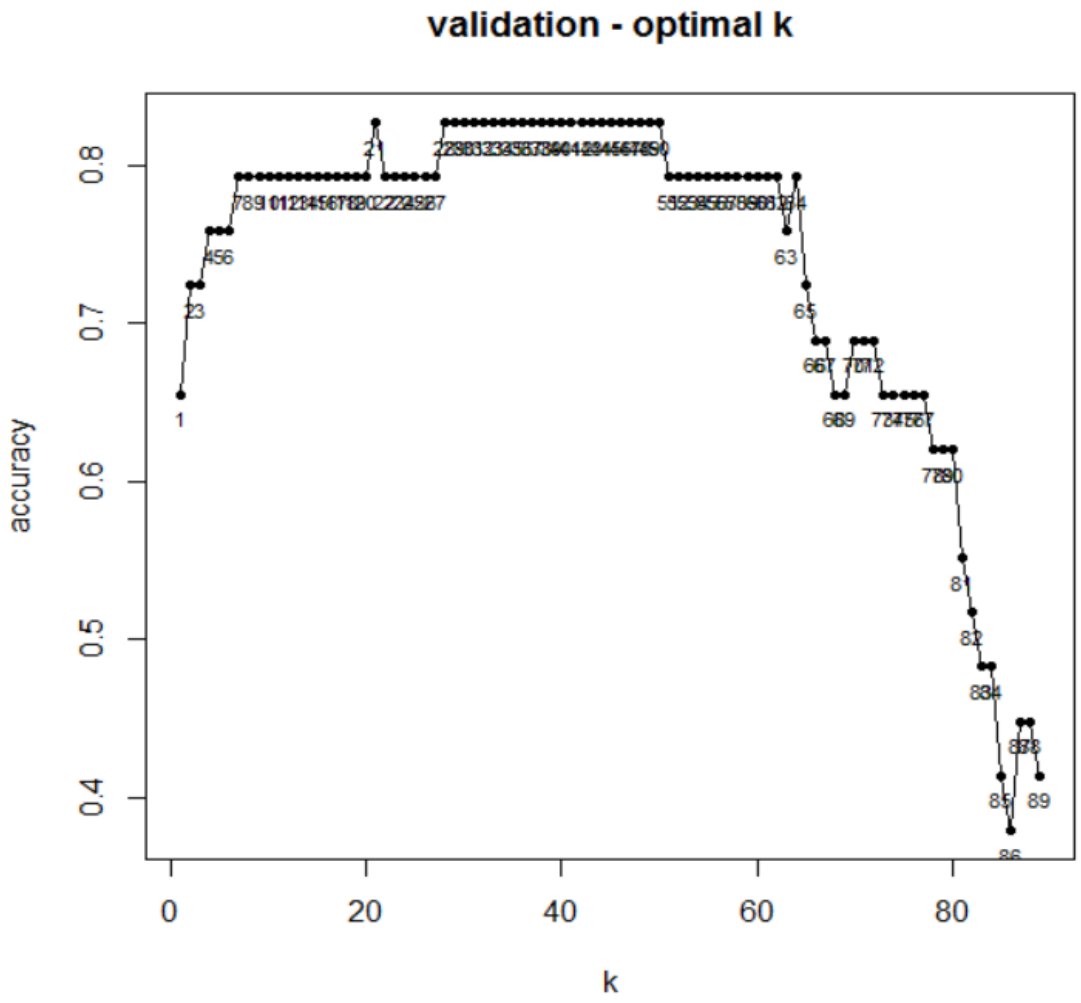
	계수		
Y 절편	3255.847458		
X 1	21.00500139	Bias	-373.24
		MAD	767.87
		MSE	1,142,046

Source :

의사결정에 명확한 기준을 제공한다.

k-NN (k-최근접 이웃 알고리즘) : 21-NN 모델의 분류 정확도는 78.1% 어떤 시사점이 있는가?





Source : 꼬낄콘의 분석일지, https://kkokkilkon.tistory.com/14

통계 개념의 이해로 활용에 대한 시각을 넓히자.

(311) Statistics Basic

- Statistics basic
- 통계분석 활용

통계를 활용하는 능력

학문이나 이론으로 접근하는 것이 아니라, 활용을 위해

개념을 정립한 후에, 필요 시에 찾아서 활용

(312) 통계 분석기법 및 결과 해석

- Basic
- R을 이용한 통계 분석

AI/ML 을 이용한 활용에 대한 아이디어 개발

필요한 것을 사용하는 시대, 사용자 중심으로 변화에 대응

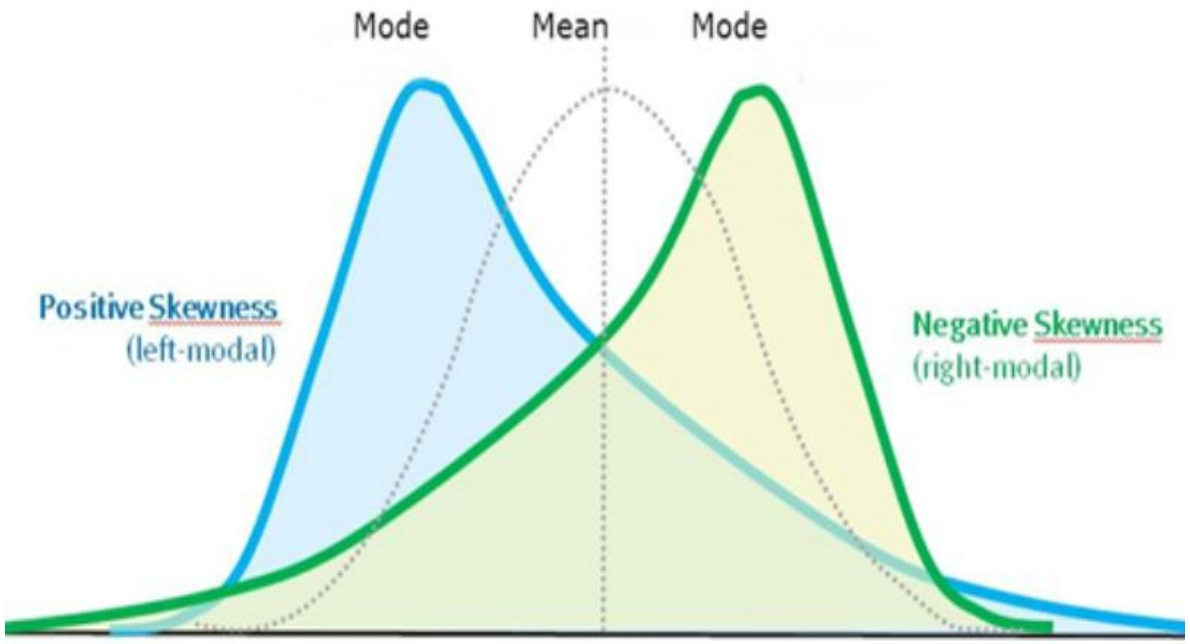
(Open source, Library, API Economy + Cloud)

Source : 311. DA_Statistics._theory

통계 활용 - 모 집단 특성 파악 (기술통계)

모 집단 특성 파악 : 기술통계

- 기술통계 숫자의 의미도 중요하지만,
- why 시각으로 더 깊이 들어가 보면 다른 차원을 볼 수 있다



평균

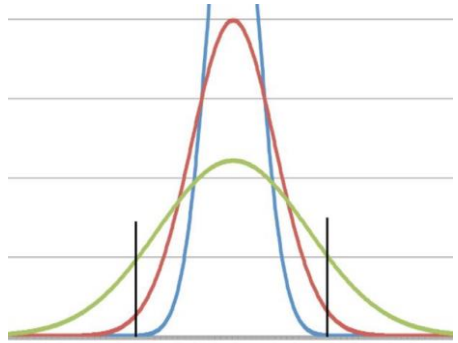
분산, 표준 편차

왜도

$$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$

첨도

$$Kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4}$$



Source :

통계 활용 – 추론과 비용 절감

대통령 선거 결과 추론
몇 명을 표본으로 조사를 해야 하는가?

표준정규분포와 **모집단 비율 추정할 때 표본의 크기**

- **신뢰수준 95%, 허용 오차 +- 3.1 %**

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

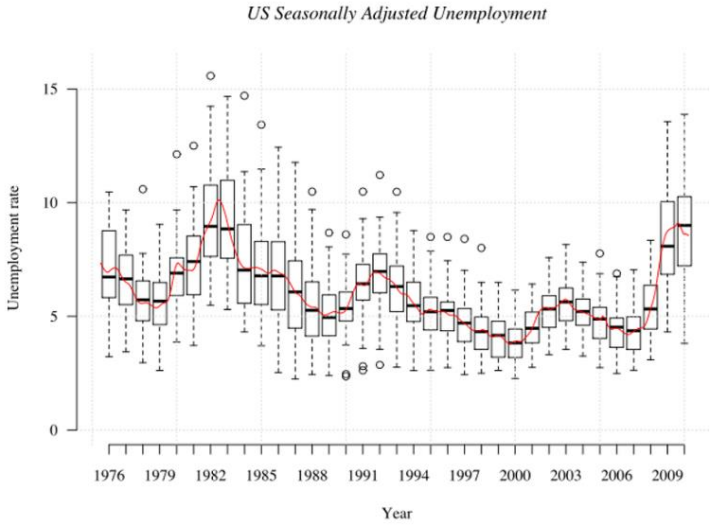
$$n = \frac{Z^2}{4e^2}$$

표본크기	최대허용오차
1,000명	3.10%
1,500명	2.53%
2,000명	2.19%
2,500명	1.96%
3,000명	1.79%

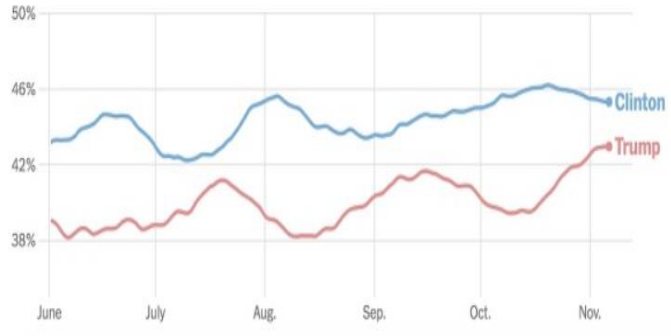
Source :

통계분석에 왜 R인가?

기술 통계



추리 통계



명령어
length()
summary()
mean()
var()
sd()
quantile()
fivenum()
IQR()
boxplot()
pairs()
hist()
stem()
qqnorm()

쉬운 명령어

다양한 시각화 방법

Source :

R : 아주 간단한 명령어로 다양한 분석 값을 손쉽게 얻을 수 있다.

- **평균 :** mean()
- **분산 - 관찰지의 퍼진 정도 :** var()
- **표본분산 :** var(data)*(length(data)-1)/length(data)
- **표준편차 - 관찰치의 퍼진 정도 :** sd()
- sqrt(var(data))

- **표준오차 - 추정치의 표준편차 :** **s**d(data)/sqrt(length(data))
- **변동계수 :** sd(data)/mean(data)

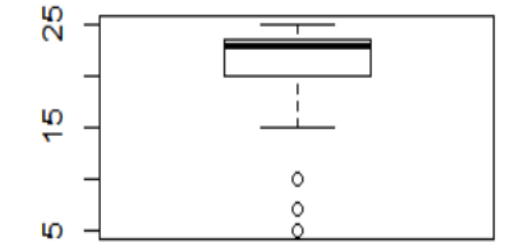
- **Boxplot**
- boxplot(data, col="blue")

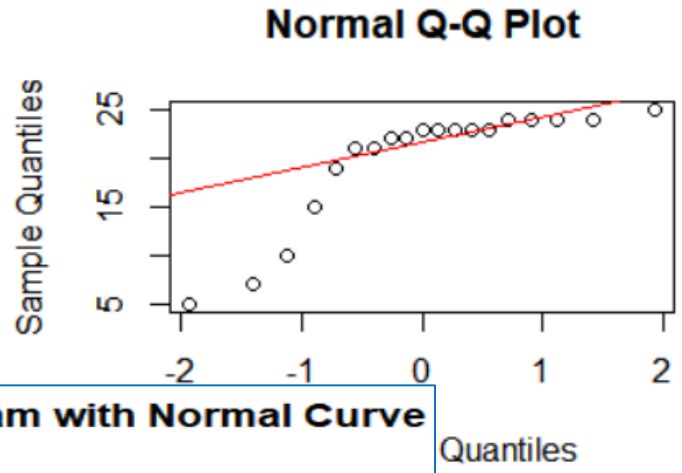
- **Q-Q Normality plot - 데이터가 정규분포에 얼마나 근접**
- qqnorm(data)
- qqline(data)

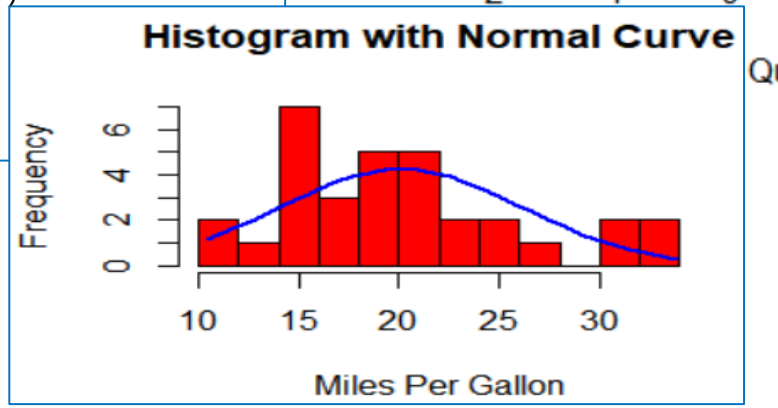
- **표본추출 :** sample()

- **히스토그램**
- hist(data, probability=TRUE)
- lines(density(data), col="red")

```
x <- c(5, 7, 10, 15, 19, 21, 21, 22,
22, 23, 23, 23, 23, 23, 24, 24, 24,
24, 25)

length(x)
max(x)
min(x)
range(x)
mean(x)
median(x)

summary(x)  # different method

quantile(x, type = 1)


boxplot(x)
```

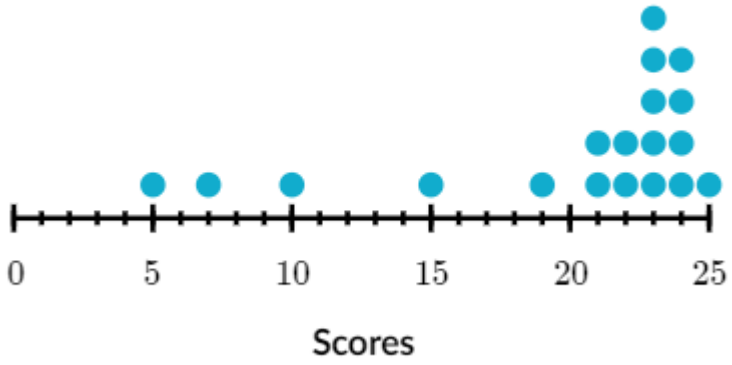






Source : sta_descriptive.R

얼마나, 효과적인 지는 엑셀과 비교하면 알 수 있다. 서로 사용 용도가 다르다.

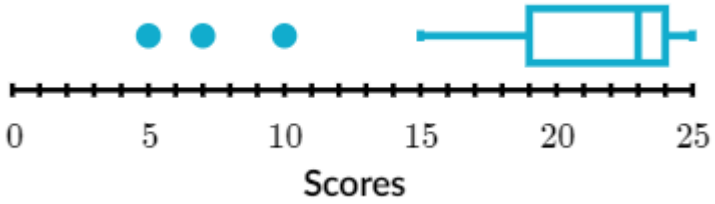


5, 7, 10, 15, 19, 21, 21, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25

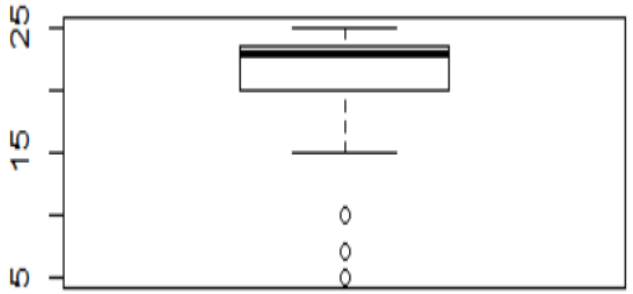
$$Q_1 - 1.5 \cdot \text{IQR} = 19 - 1.5(5)$$

$$= 19 - 7.5$$

$$= 11.5$$



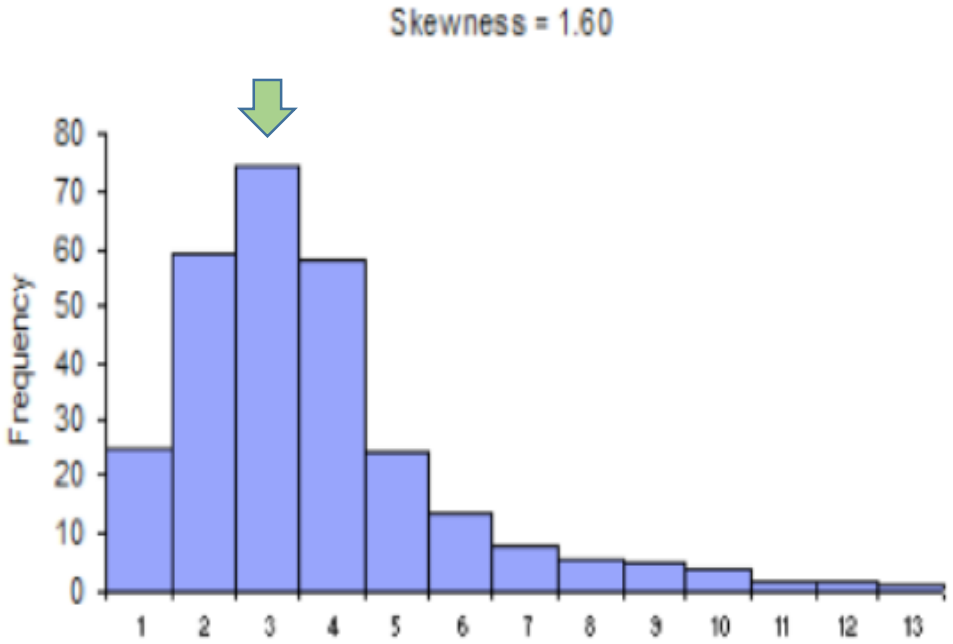

```
> quantile(x, type = 1)
  0%  25%  50%  75% 100%
   5   19   23   24   25
```



Source : https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule

기술통계

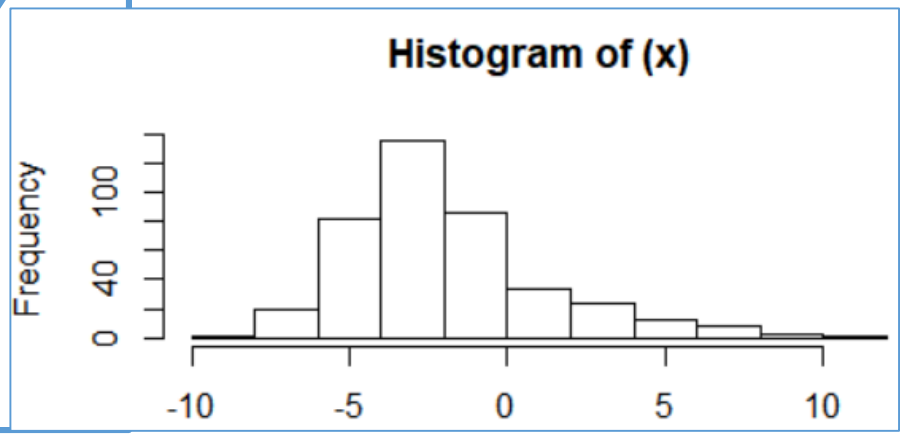
기술 통계	결과의 해석 및 plot	집단의 재 분류



R Code

```
a <- rnorm( 50, 0, 2 )
b <- rnorm( 300, -3, 2)
c <- rnorm( 50,  4, 4 )

x <- c(a, b, c)

hist((x))
```

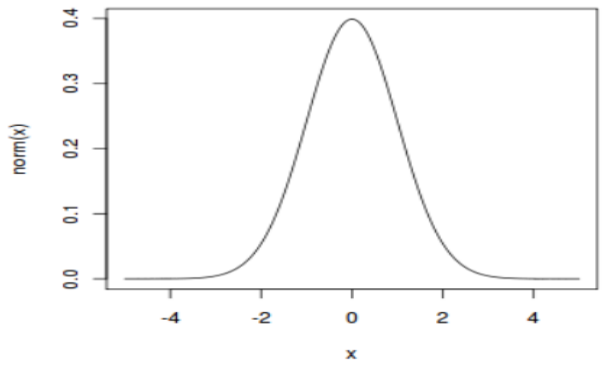
- Mode를 발생시키는 요인 확인
- 요인에 따른 재 분류
- ....



Source :

가설 검정

모집단의, 평균, 표준편차를 알고 있음

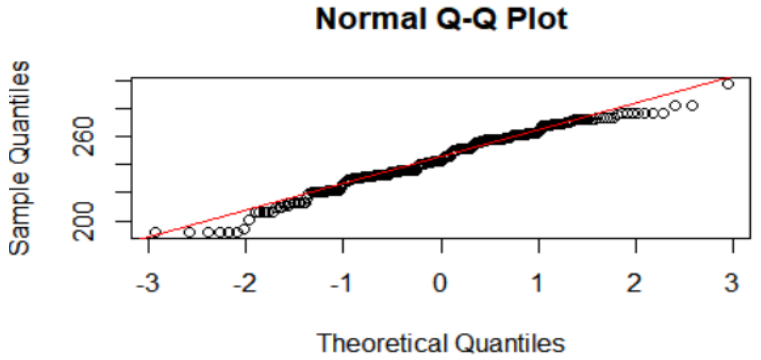


t.test 일표본 가설 검정

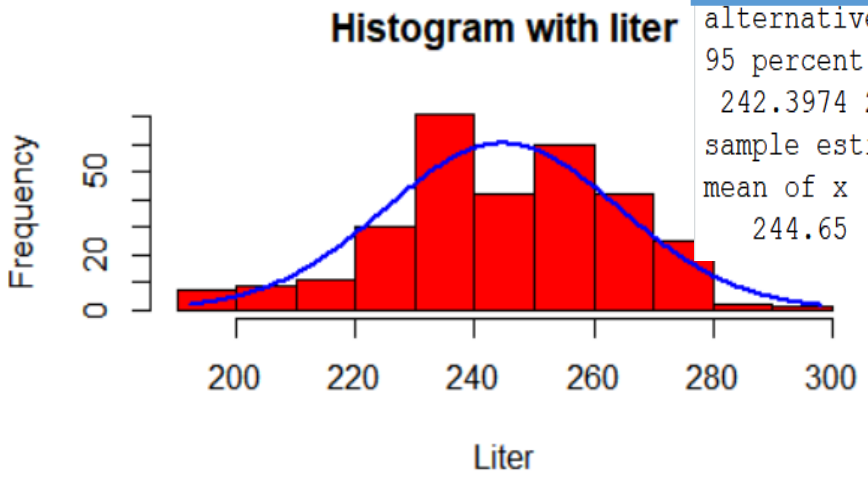
t.test

```
> t.test(x$liter, mu=250)

        One Sample t-test

data:  x$liter
t = -4.6739, df = 299, p-value = 4.477e-06
alternative hypothesis: true mean is not equal to 250
95 percent confidence interval:
 242.3974 246.9026
sample estimates:
mean of x
   244.65
```

정규성 가정
(normality assumption)



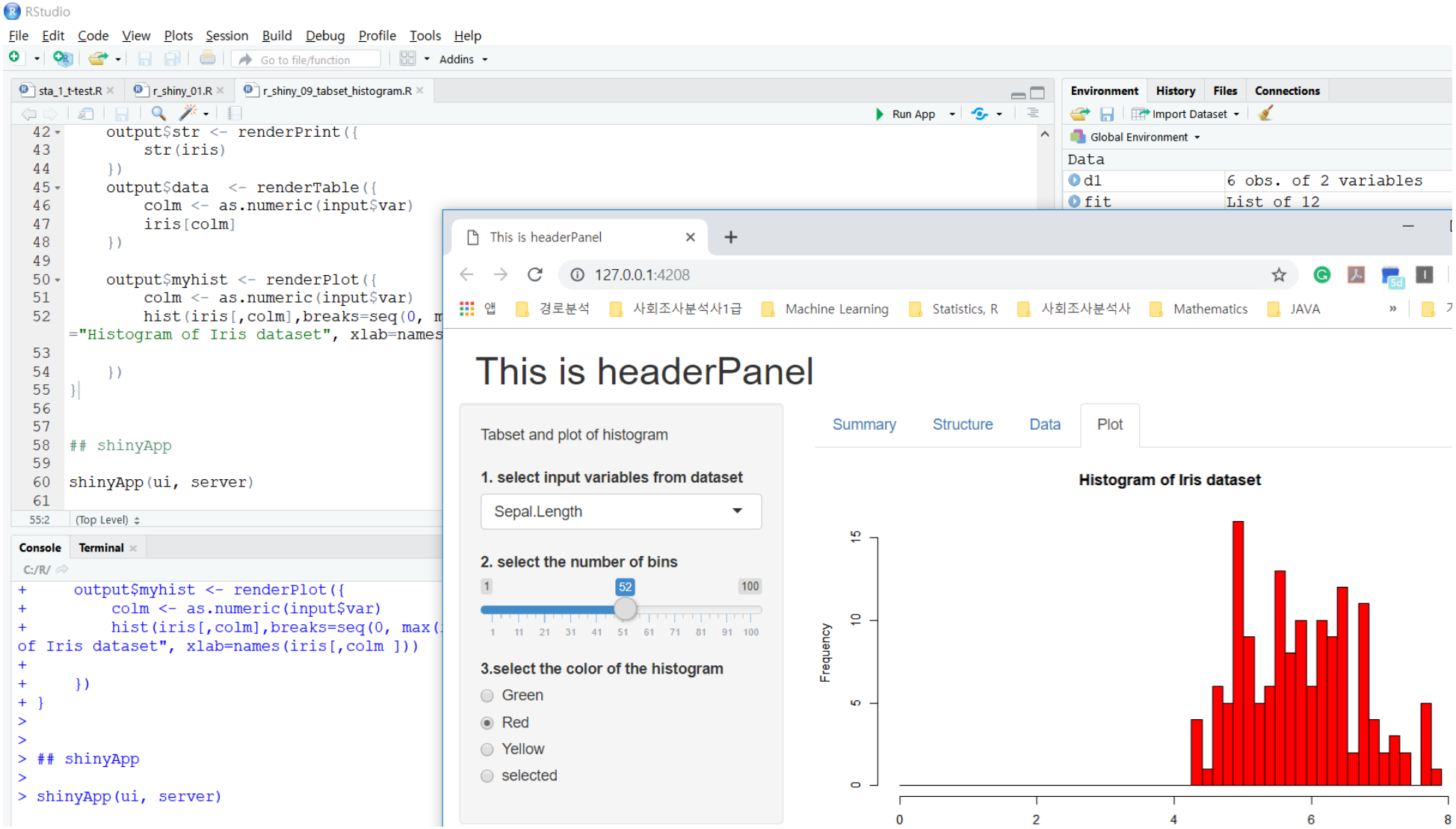

```
shapiro.test(x$liter)

        Shapiro-Wilk normality test

data:  x$liter
W = 0.97886, p-value = 0.0002061
```



등분산성 가정
(homogeneity of variance)
p-value > 0.05

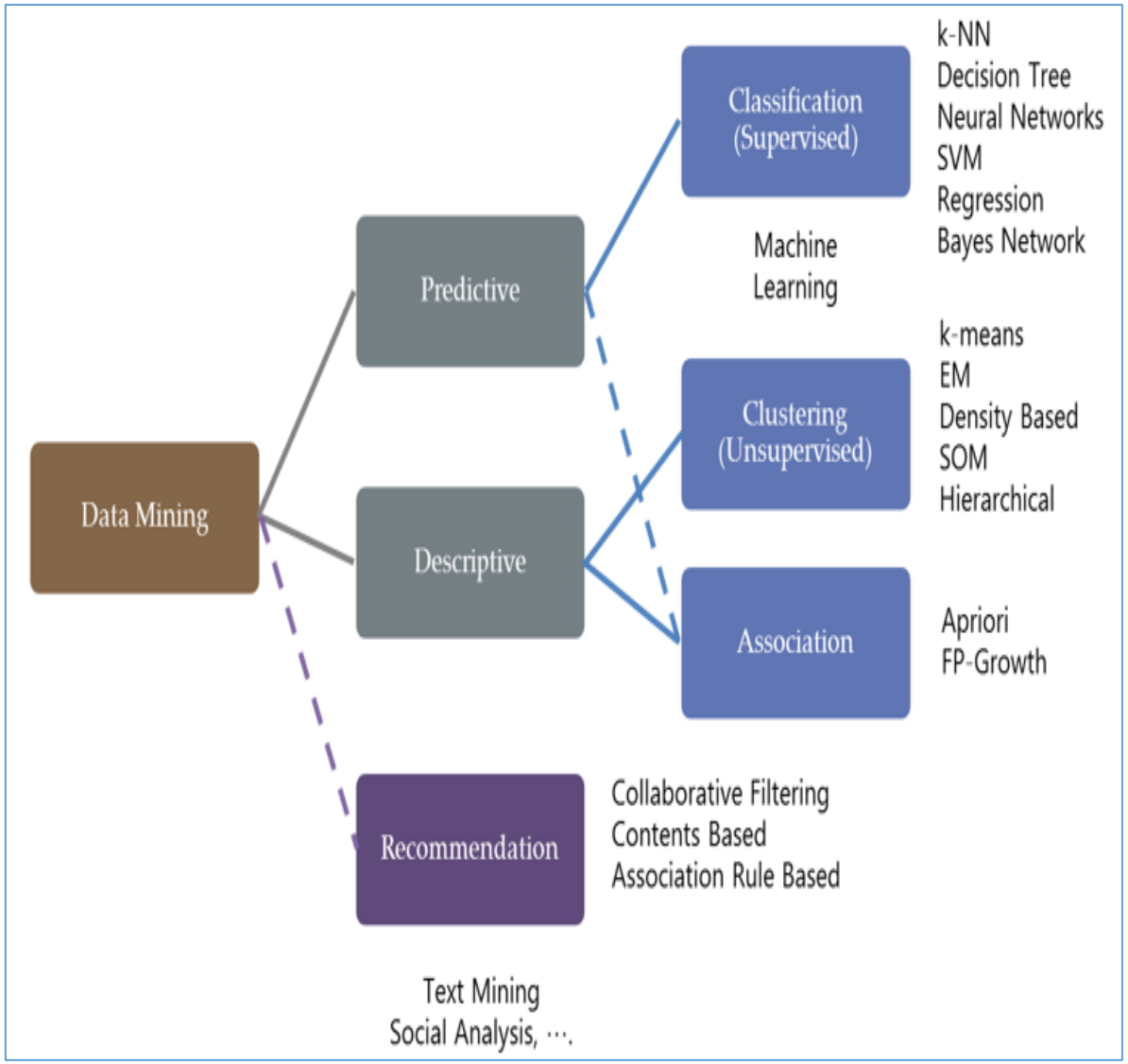
Source : sta_1_t-test.R

R demo

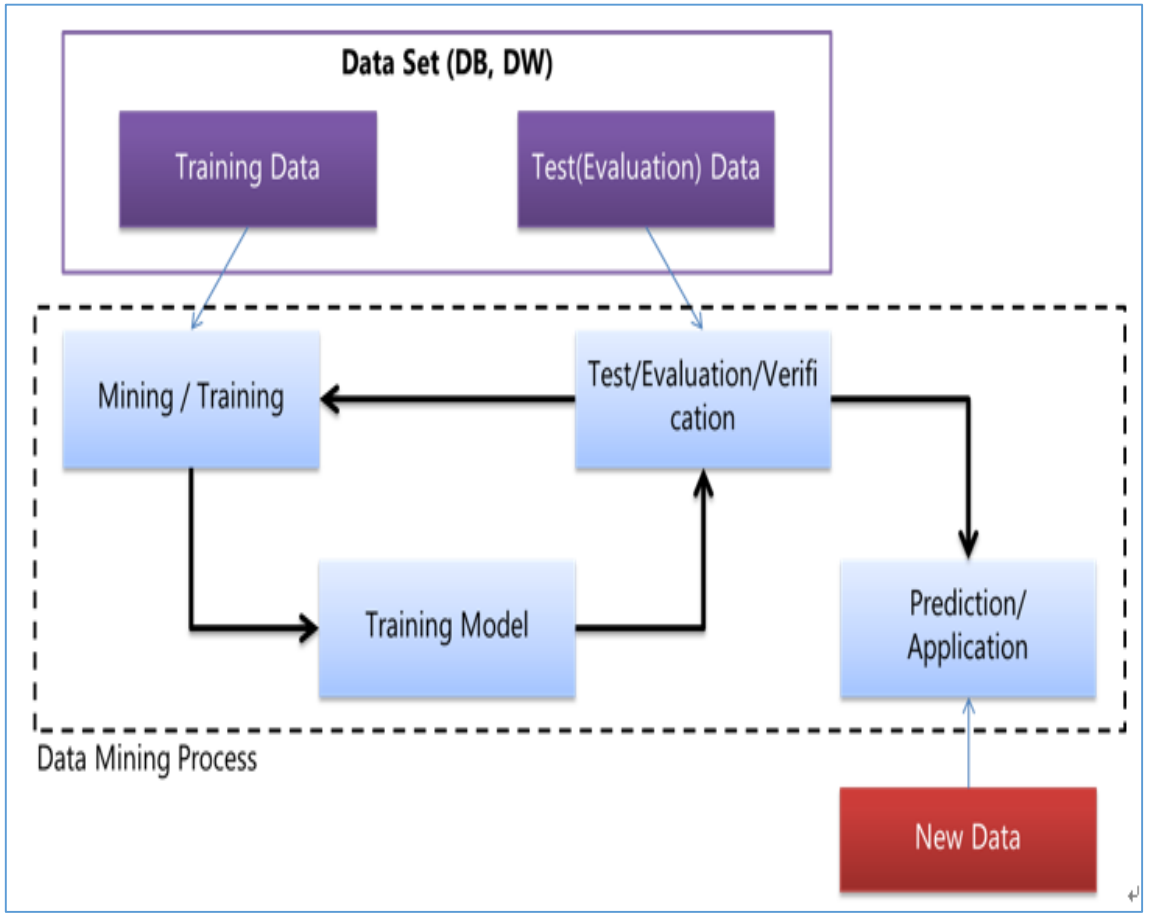


Source :

Data Mining – 분석 Algorithm 및 ML을 활용하여 데이터를 분석한다.



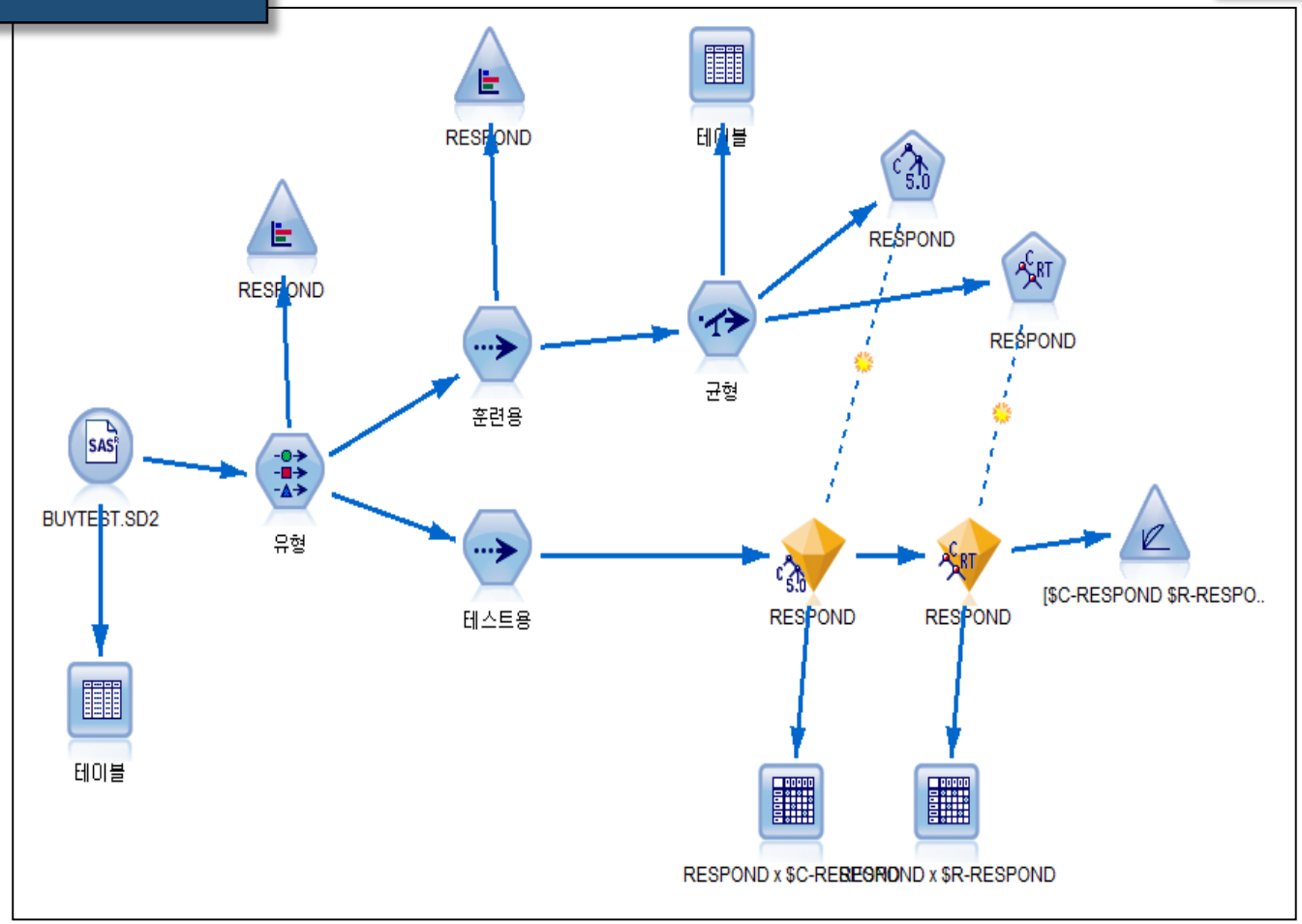
실무 활용 솔루션



Source :

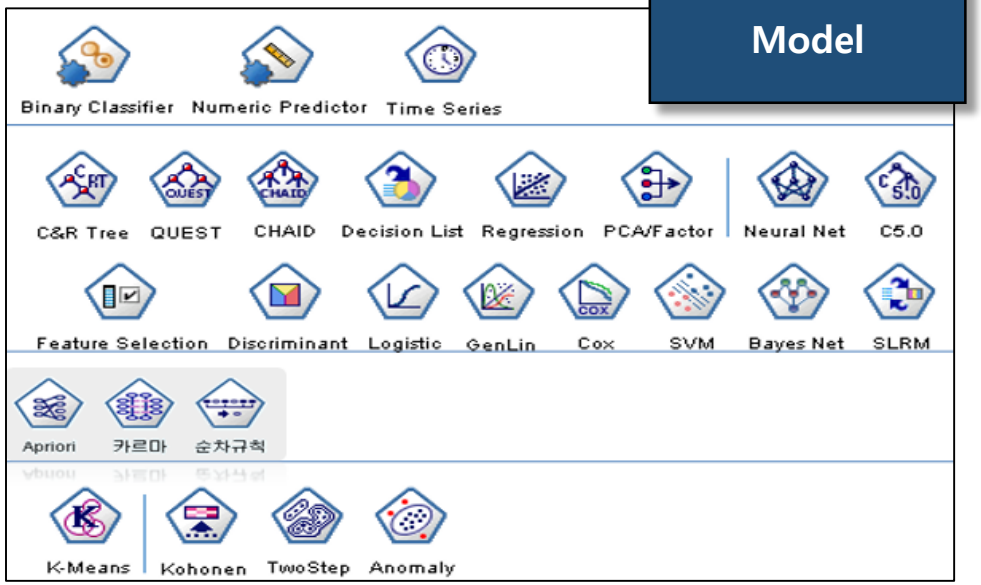
Data Mining 솔루션

Work Flow



Input

Model



Output

Source :

ML 개념 소개

(321) ML Basic Concept

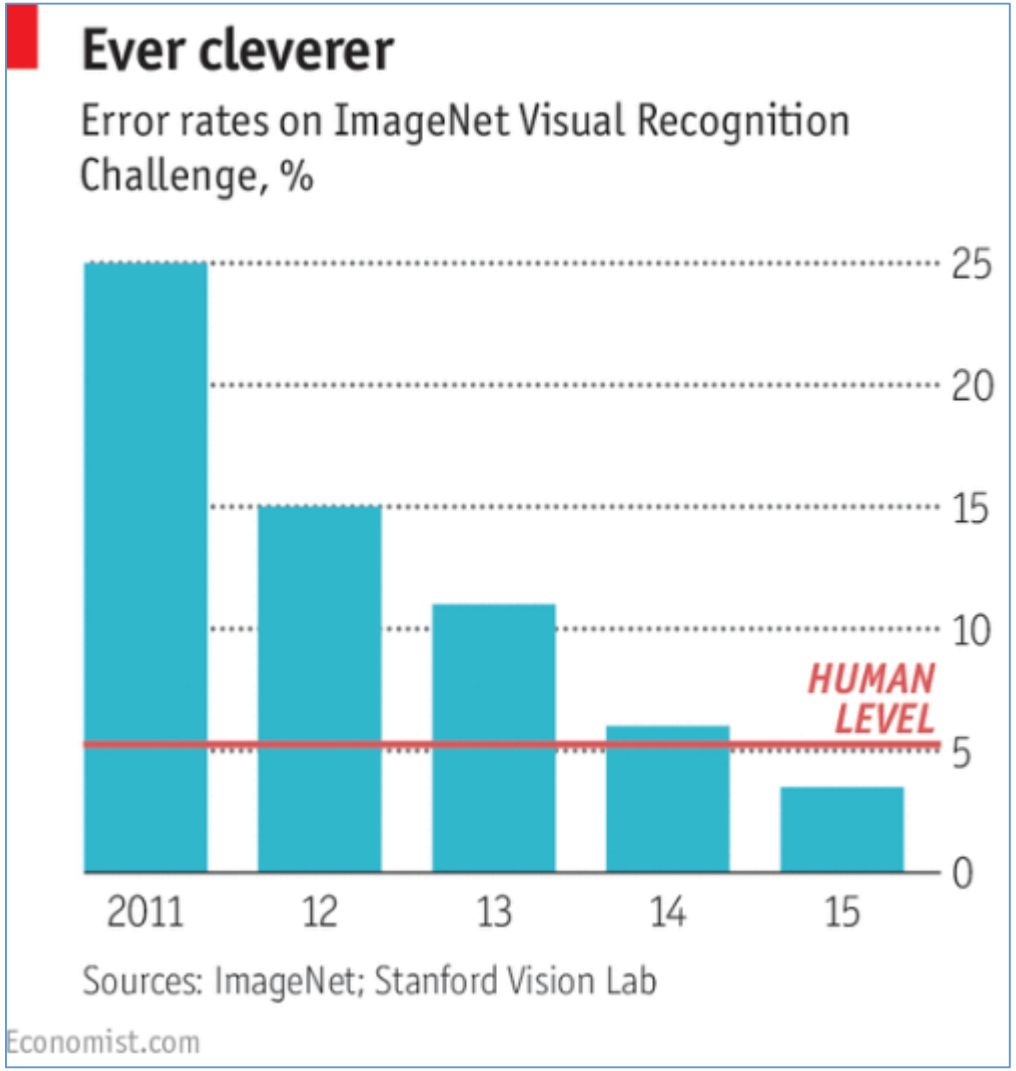
- Python, TensorFlow, TensorBoard, Matplotlib
- AI, xor, ML, Deep Learning
- CNN
- RNN
- 활용

(322) ML Algorithm (p1~p3)

- Linear Regression
- Logistic Classification
- Softmax
- AI, ML & Deep Learning
- CNN, RNN

Source : 321. ML_concept_basic.pptx, 322. ML_Algorithm_p1_lm_softmax , 322. ML_Algorithm_p2_cnn_xor , 322. ML_Algorithm_p3_rnn

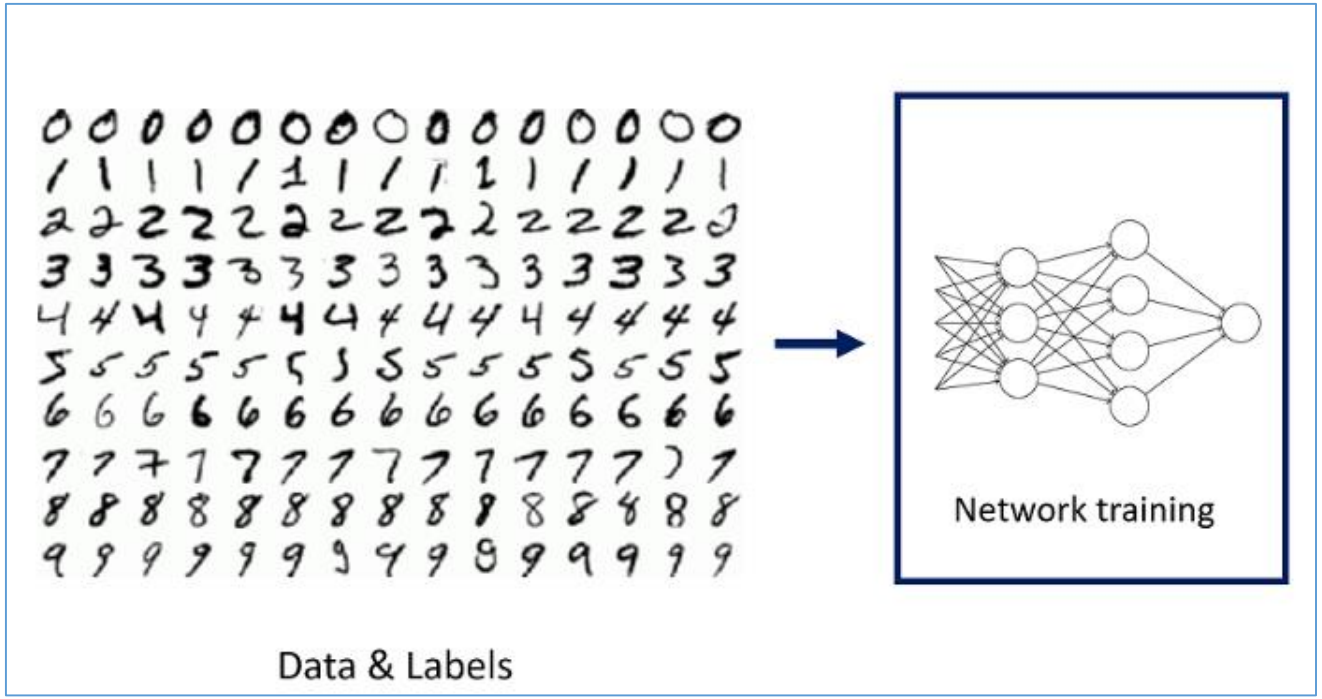
모든 데이터는 가치가 있는가 ?



Q1 : 데이터 자체가 가지는 가치에 대한 인식 (3가지)

Q2 : 가치화 아이디어

Q3 : 실현하기 위한 도구

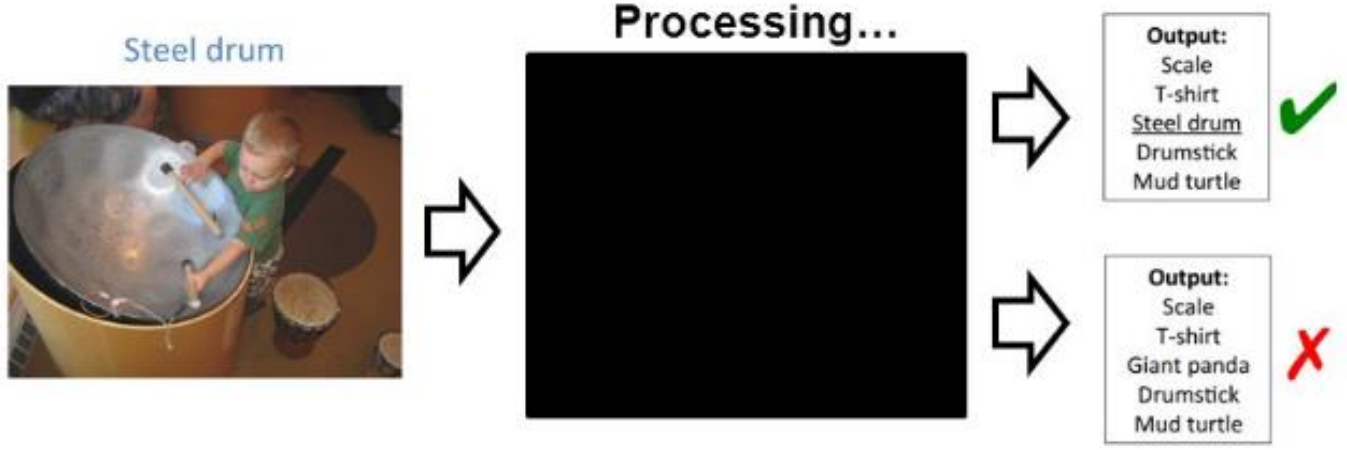


Source :

데이터 분석방식에 따라 새로운 가치가 생성될 수 있는가?

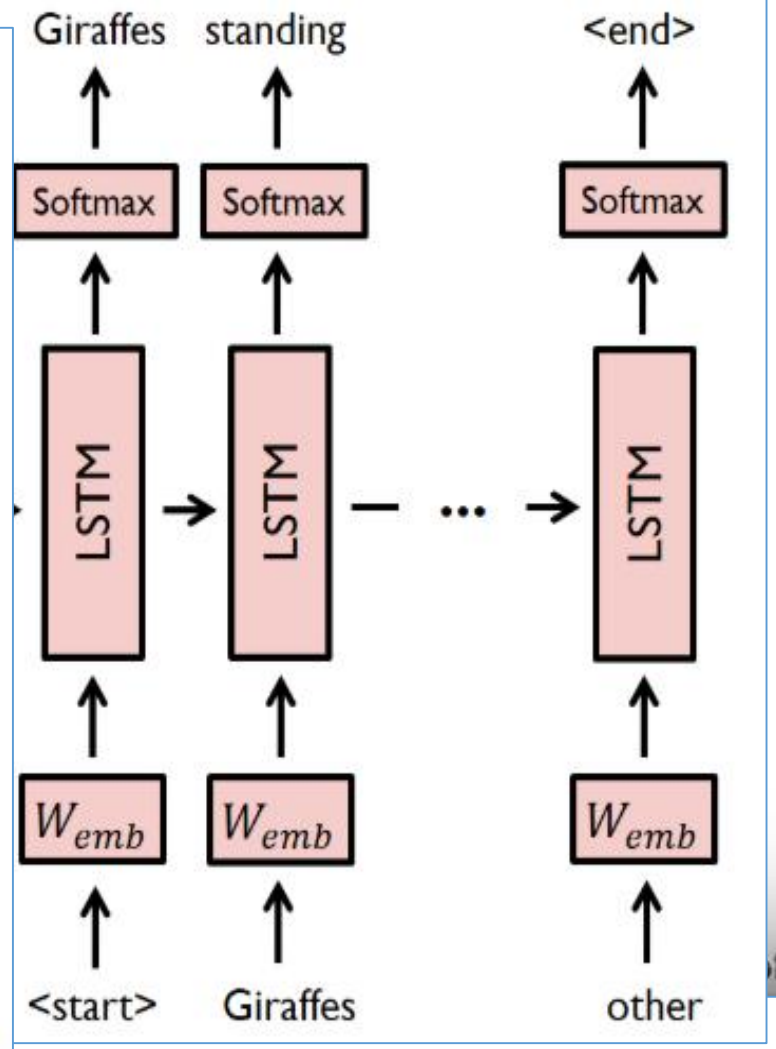
ILSVRC

- **ImageNet Large Scale Visual Recognition Challenge**
- **An image classification challenge with 1,000 categories (1.2 million images)**



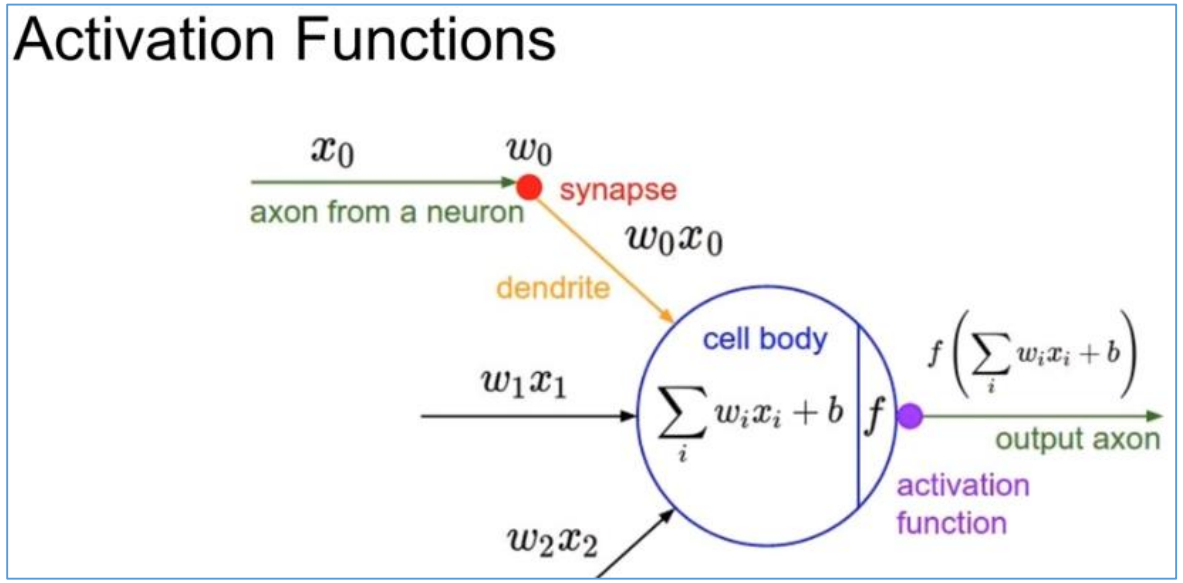
reference : http://www.image-net.org/challenges/LSVRC/2013/slides/ILSVRC2013_12_7_13_clsloc.pdf

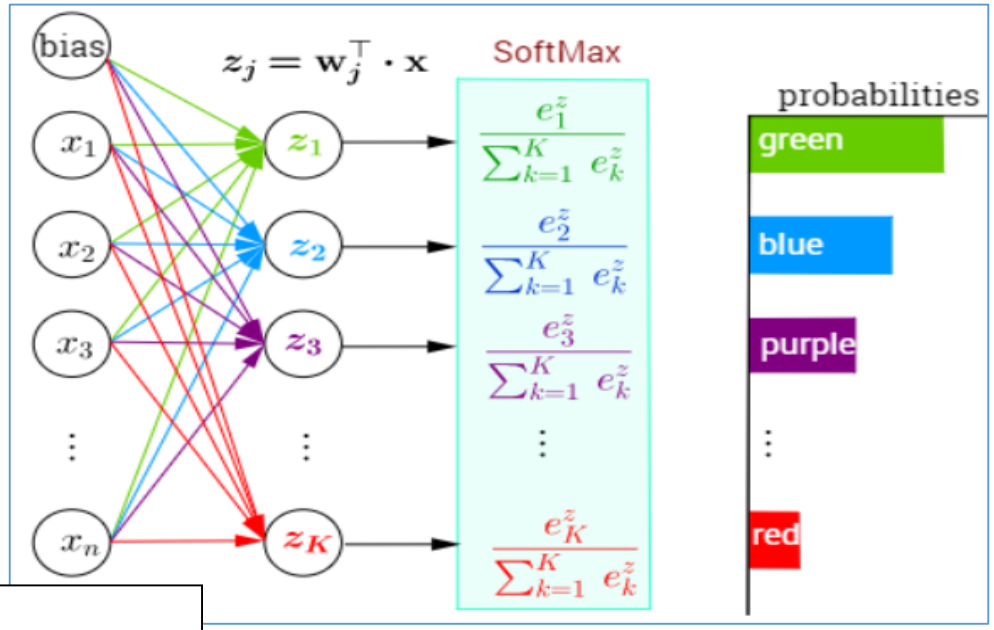
KAIST

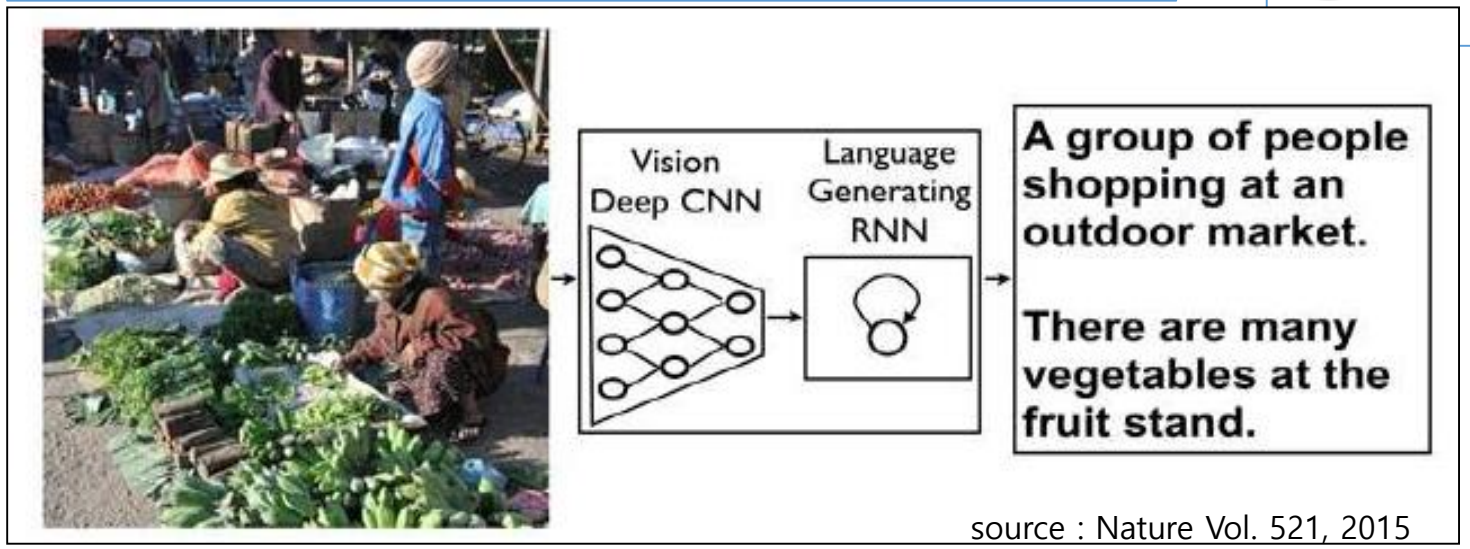


Source : Analytics Vidhya , https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/

AI / Machine Learning



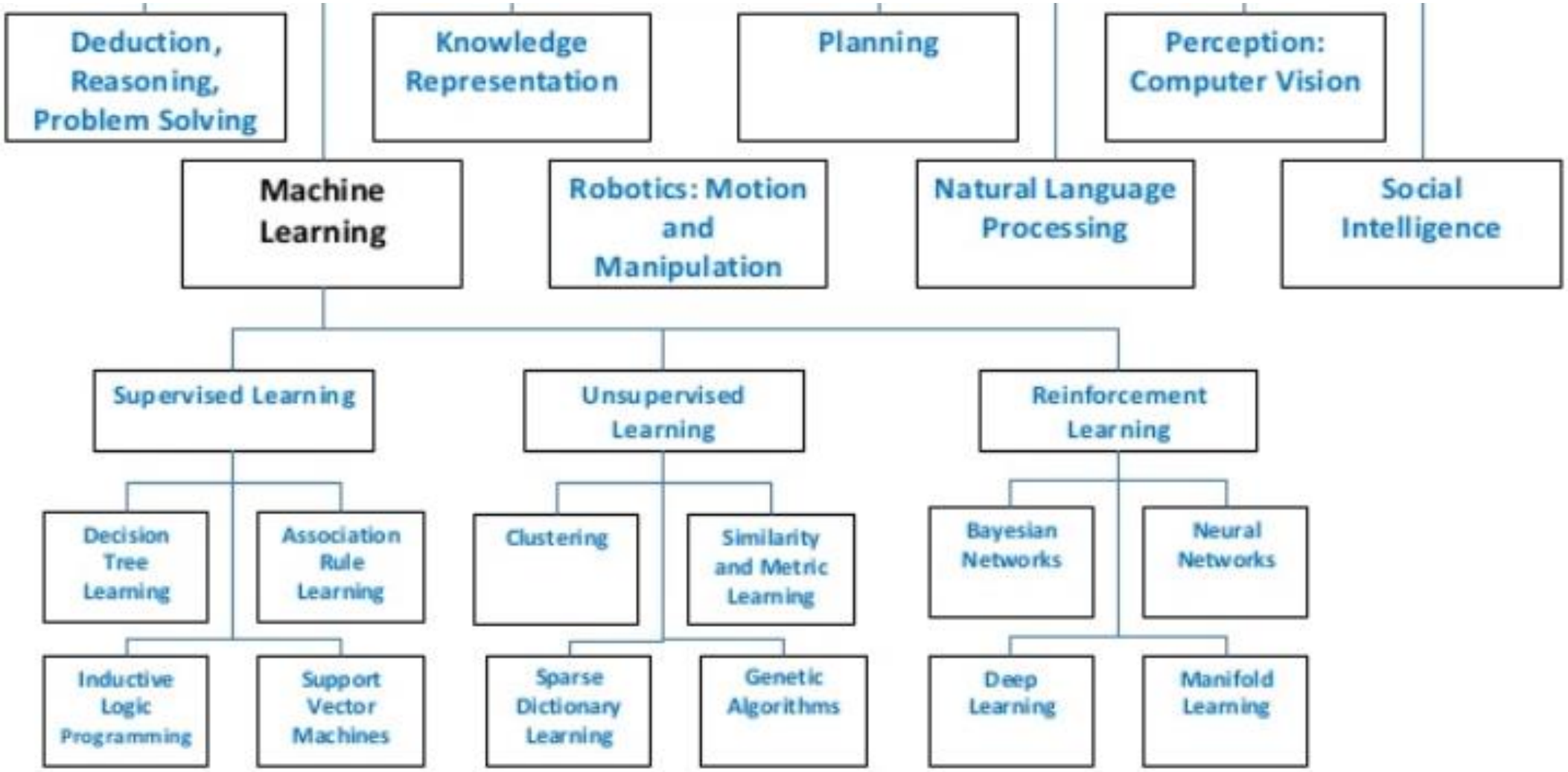




source : Nature Vol. 521, 2015

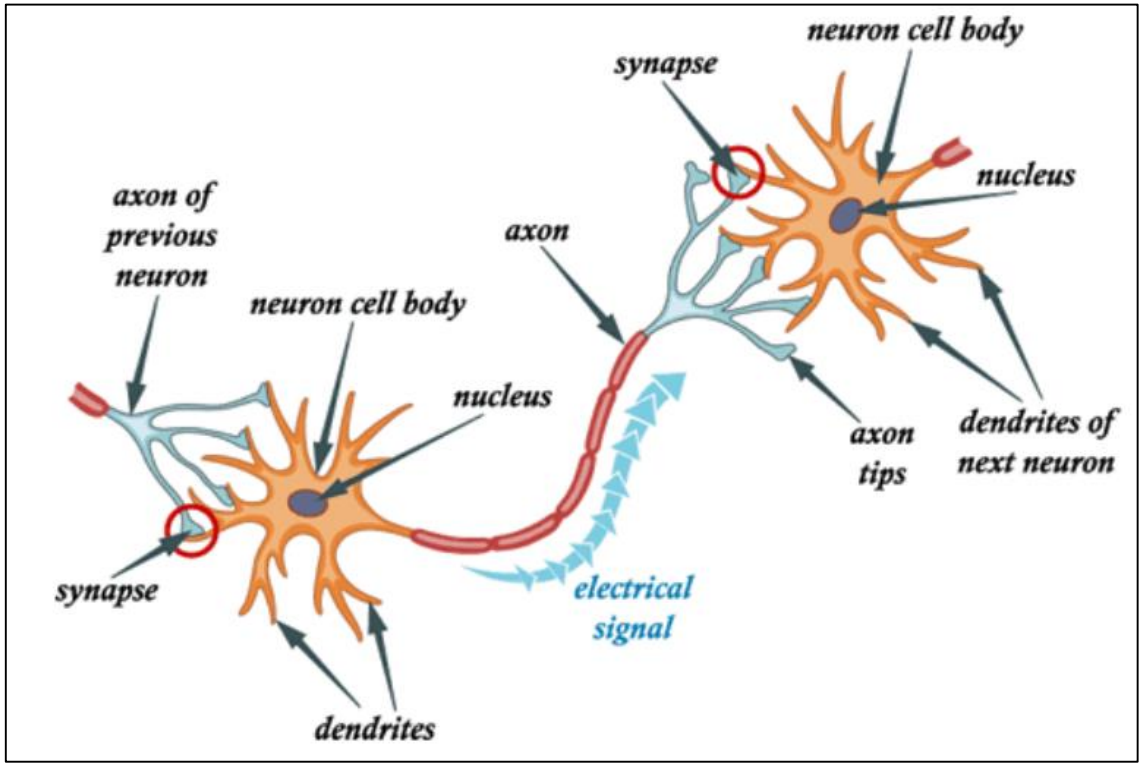
Source : https://youtu.be/n7DNueHGkqE https://sebastianraschka.com/faq/docs/softmax_regression.html

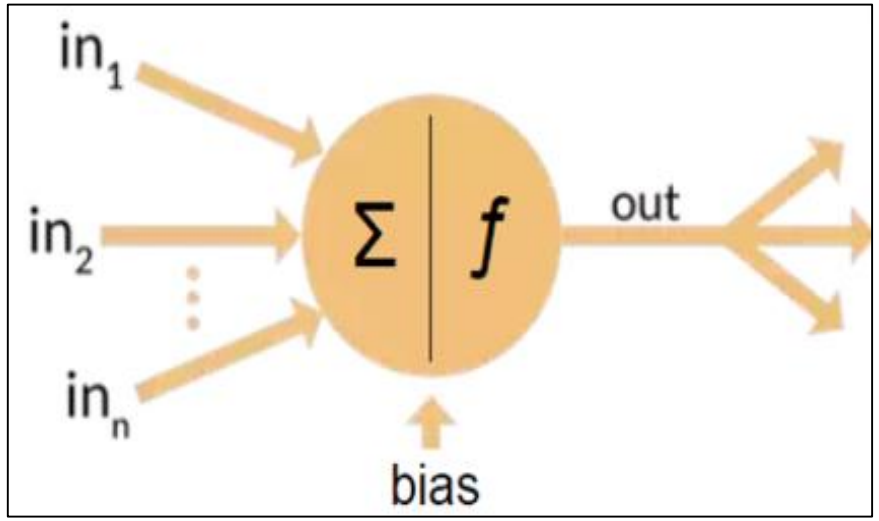
분석에 대한 분류가 아니라 적용하여 효과를 가시화 하는 것이다.



Source : Deepdive in AIML venture landscape By Ajit Nazre Rahul Garg https://www.slideshare.net/ajitnazre/deepdive-in-aiml-venture-landscape-by-ajit-nazre-rahul-garg

AI의 시작과 진화를 보면 그러한 사실을 증명한다.



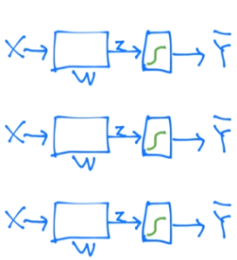


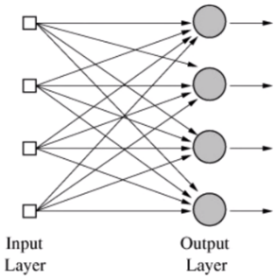
Source : https://towardsdatascience.com/the-differences-between-artificial-and-biological-neural-networks-a8b46db828b7

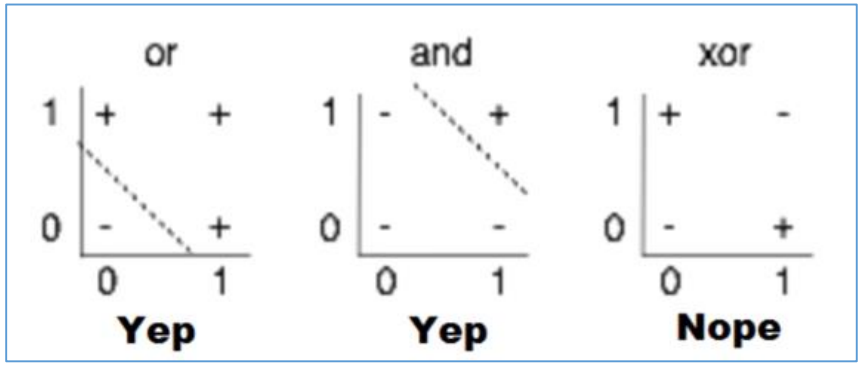
딥러닝 시작과 XOR 문제

xor linearly separable

Logistic regression units

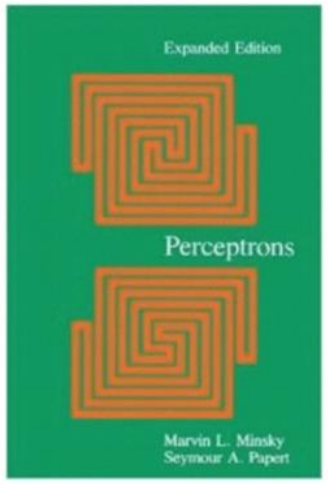






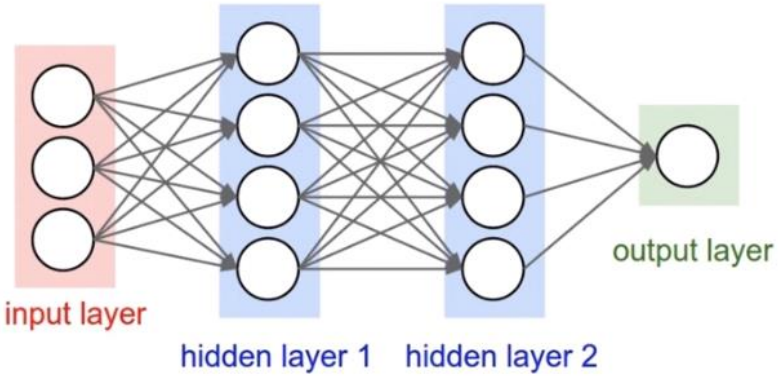
Perceptrons (1969)

by Marvin Minsky, founder of the MIT AI Lab



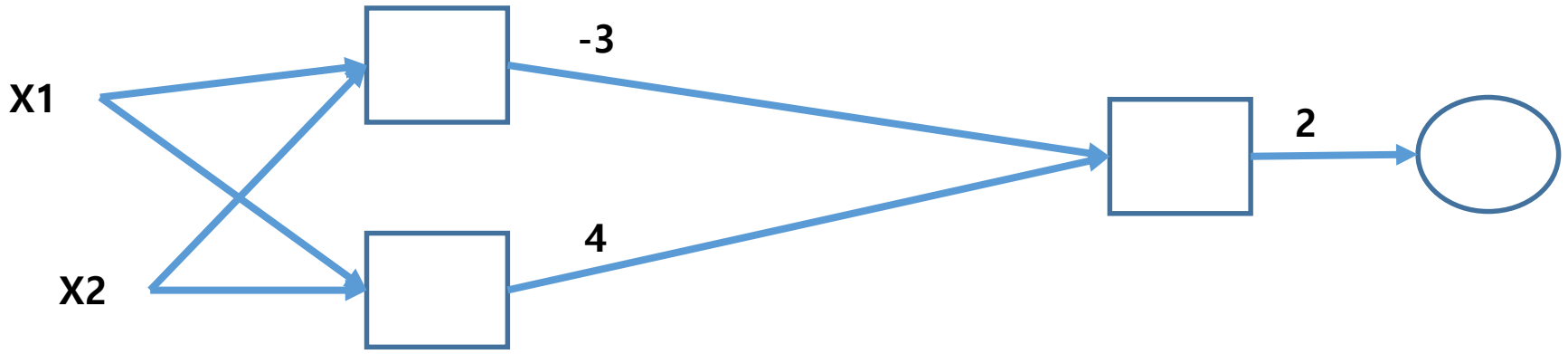
- We need to use MLP, multilayer perceptrons (multilayer neural nets)
- No one on earth had found a viable way to train MLPs good enough to learn such simple functions.

xor 은 multi layer 로 가능하나, weight, bias 를 학습 시킬 수 없다.



Source : https://youtu.be/n7DNueHGkqE

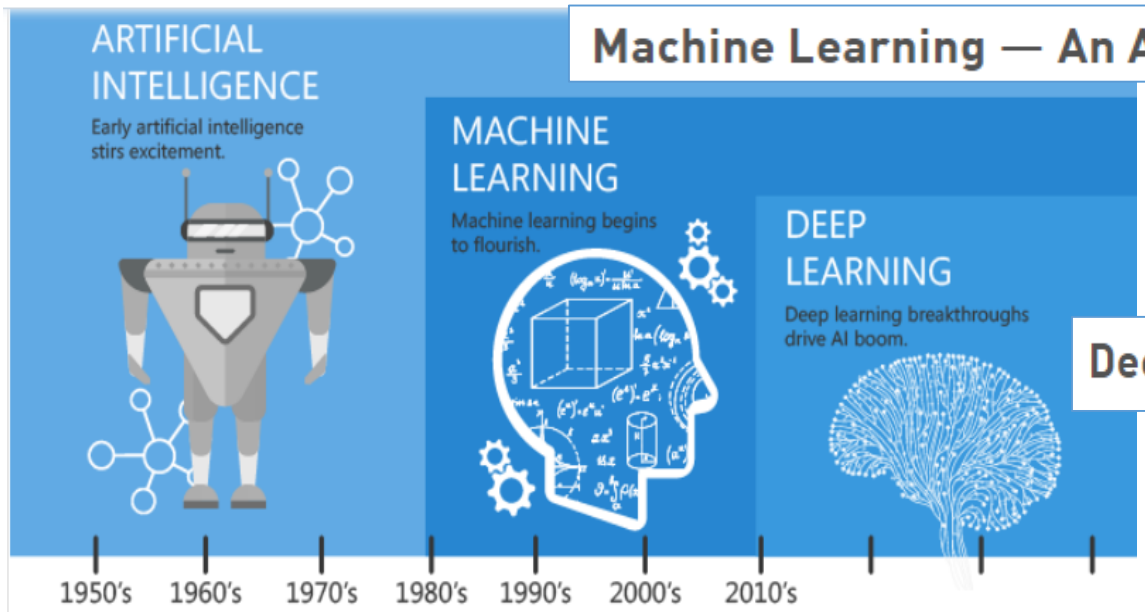
Neural net for XOR



$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} * \underset{\text{weight}}{\begin{pmatrix} 2 & -5 \\ 2 & -5 \end{pmatrix}} \underset{\text{bias}}{\begin{pmatrix} -3 & 4 \end{pmatrix}} \Rightarrow \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{pmatrix} * \underset{\text{weight}}{\begin{pmatrix} -5 \\ -5 \end{pmatrix}} \underset{\text{bias}}{\begin{pmatrix} 2 \end{pmatrix}} \Rightarrow \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

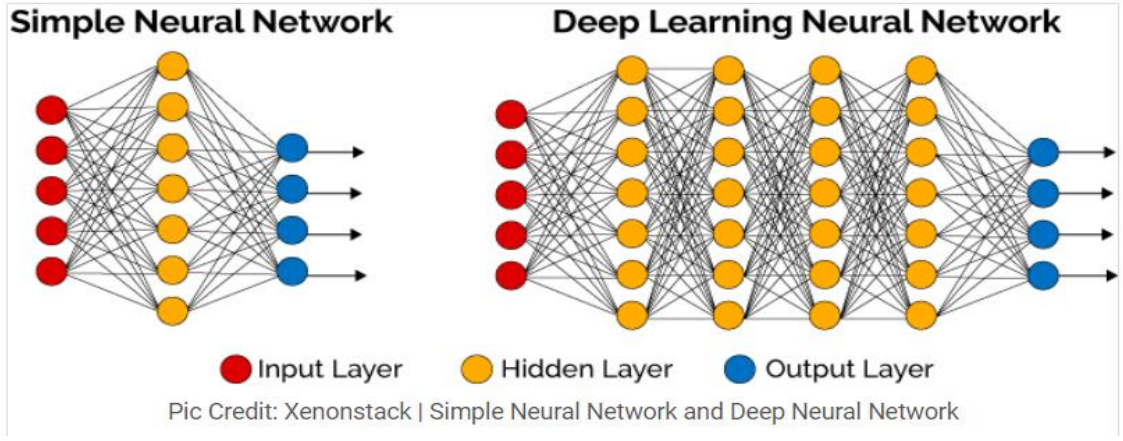
Source :

ML & DL

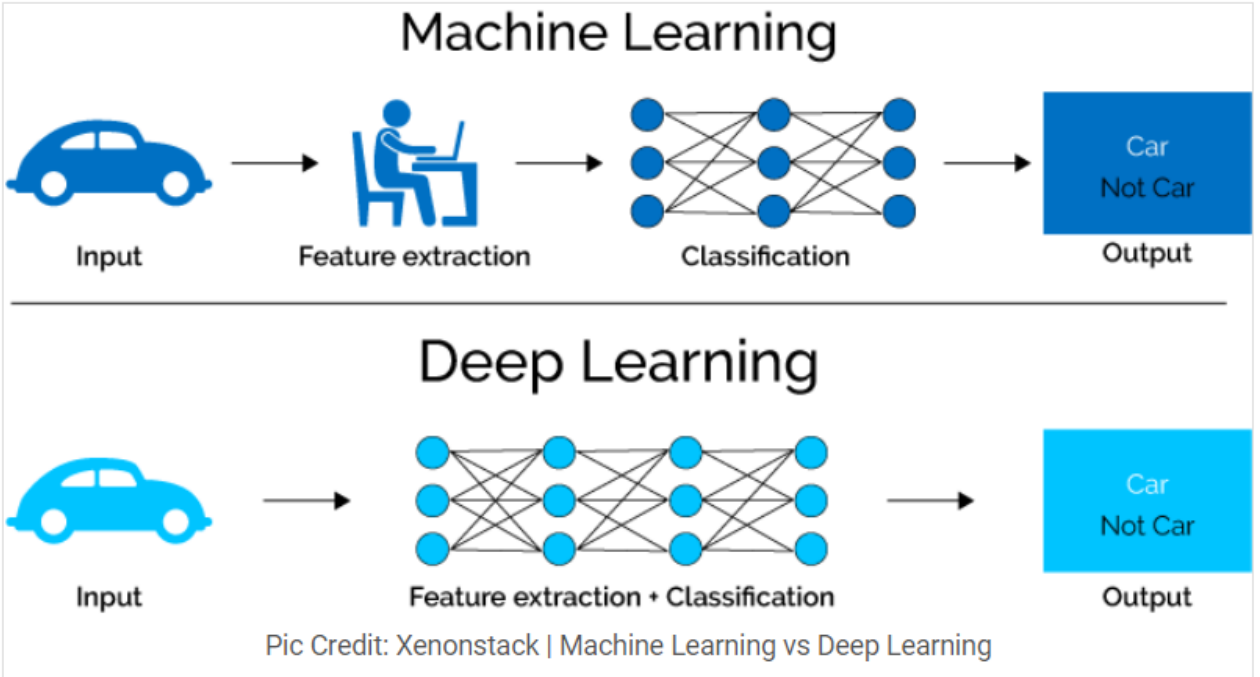


Machine Learning — An Approach to Achieve Artificial Intelligence

Deep Learning — A Technique for Implementing Machine Learning



Pic Credit: Xenonstack | Simple Neural Network and Deep Neural Network

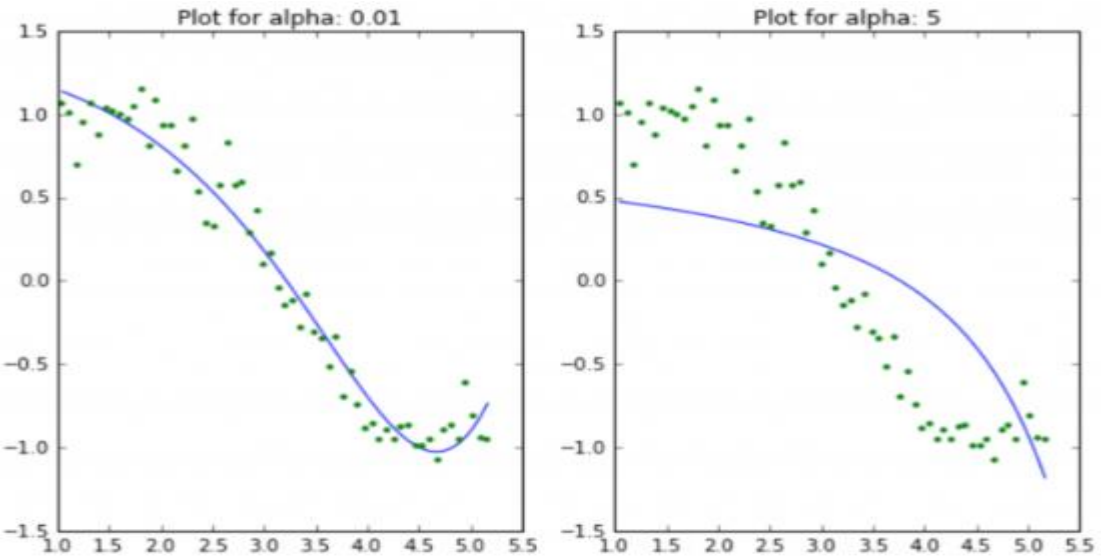


Pic Credit: Xenonstack | Machine Learning vs Deep Learning

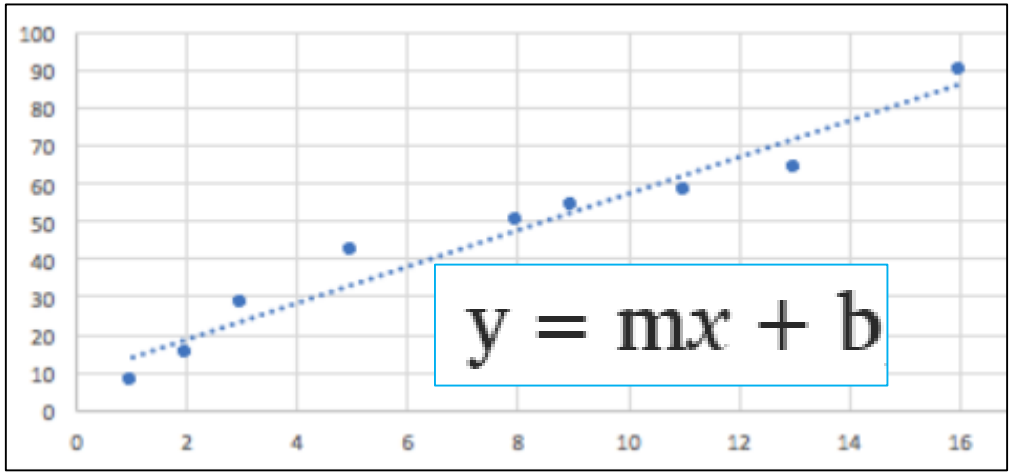
Source : http://aimagnifi.com/blog/index.php/2017/10/13/what-is-the-difference-between-machine-learning-and-deep-learning/

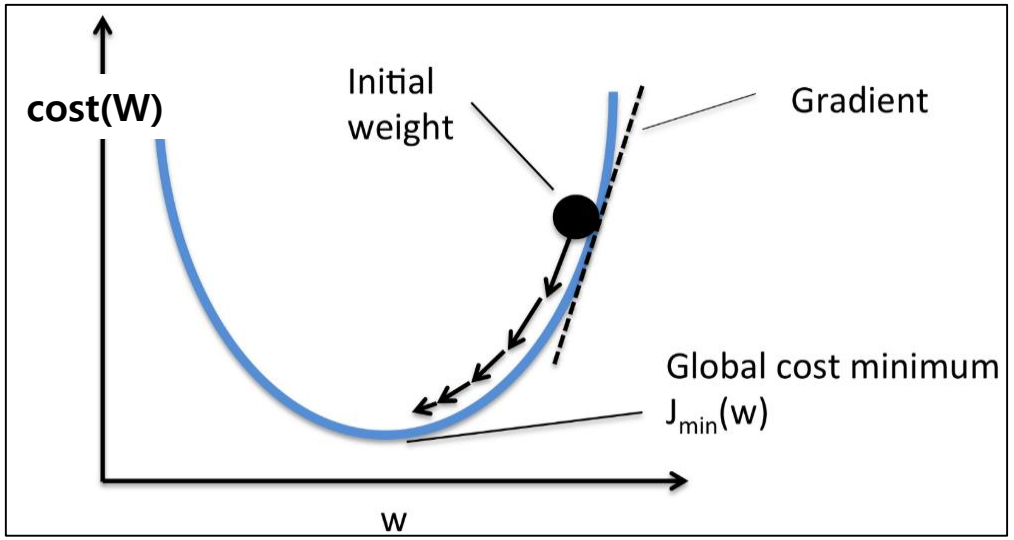
Linear regression – Hypothesis & Cost curve

- **Hypothesis :**
- **Cost function :**
- **Optimizer & Training**



https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/

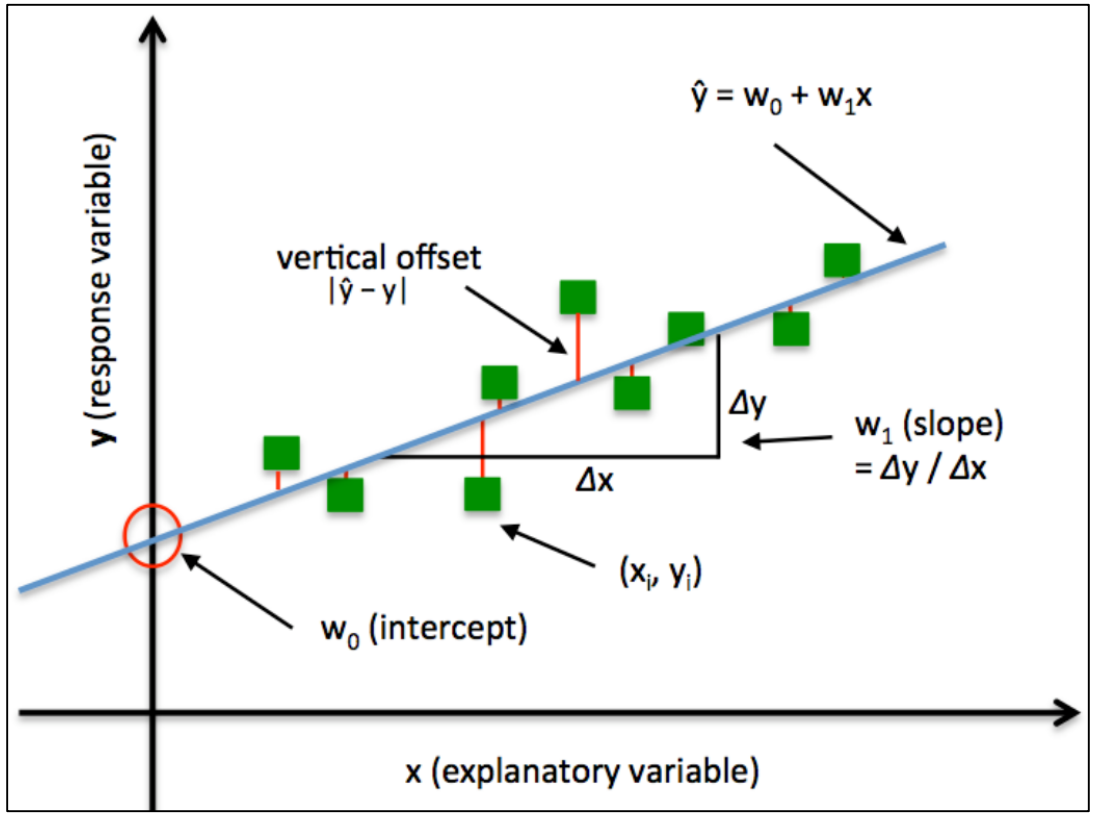




Source : https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/

Linear regression – Cost minimize

최적의 선형 회귀선을 찾기 위한 적절한 Cost function 을 이용 최소의 비용에 해당하는 W, b 값을 도출



$$Cost(W) = \sum_{1}^{m} (Wx^i - y^i)^2$$

Source : https://byungjun0689.github.io/1.-Basic_Deep_Learning/

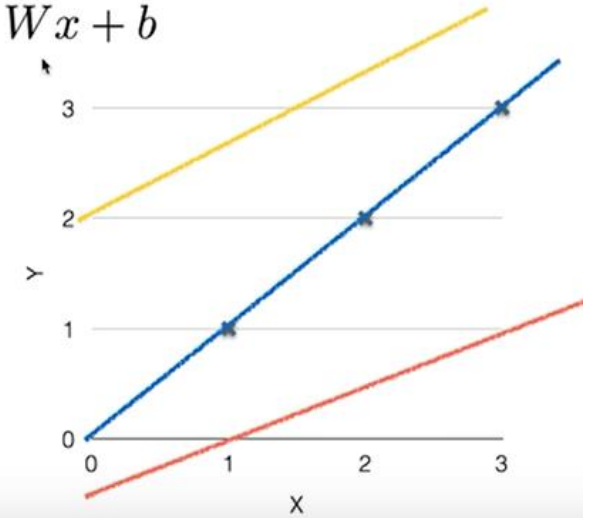
Linear Regression의 Hypothesis, cost

집단을 가장 잘 설명하는 선형회귀선(가설)을 찾기 위해 비용함수의 최소값을 만족하는 W, b 값을 찾음

Hypothesis	Cost Function	W, b 값

(Linear) Hypothesis

$$H(x) = Wx + b$$

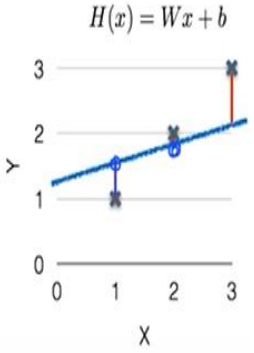


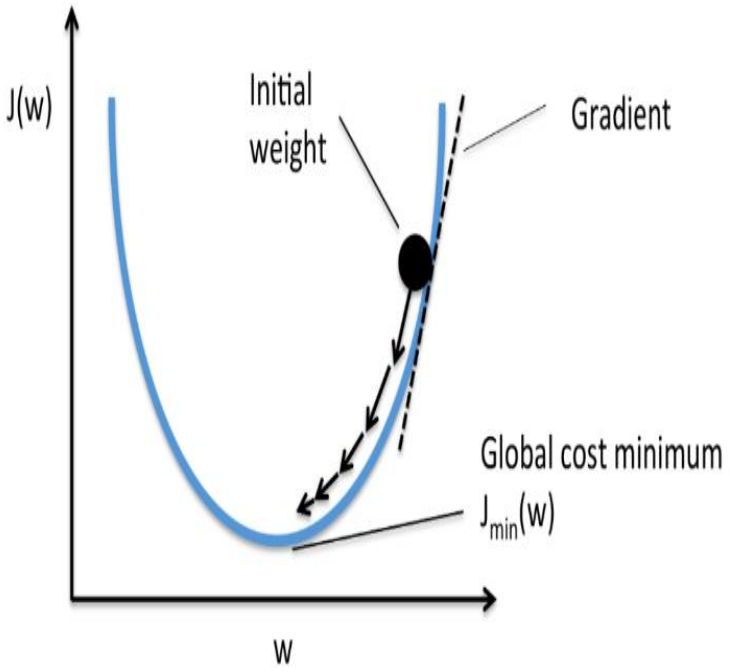
Cost function

- How fit the line to our (training) data

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$





출처 : https://youtu.be/Hax03rCn3UI?list=PLlMkM4tgfjnLSOjrEJN31gZATbcj_MpUm

Cost function 확인

W 변화에 따른 Cost 확인

What cost(W) looks like?

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

x	Y
1	1
2	2
3	3

- W=1, cost(W)=?

- W=1, cost(W)=0
- W=0, cost(W)=4.67
- W=2, cost(W)=4.67

```
import tensorflow as tf
import matplotlib.pyplot as plt
X = [1, 2, 3]
Y = [1, 2, 3]

W = tf.placeholder(tf.float32)
# Our hypothesis for linear model X * W
hypothesis = X * W

# cost/loss function
cost = tf.reduce_mean(tf.square(hypothesis - Y))
# Launch the graph in a session.
sess = tf.Session()
# Initializes global variables in the graph.
sess.run(tf.global_variables_initializer())
# Variables for plotting cost function
W_val = []
cost_val = []
for i in range(-30, 50):
    feed_W = i * 0.1
    curr_cost, curr_W = sess.run([cost, W], feed_dict={W: feed_W})
    W_val.append(curr_W)
    cost_val.append(curr_cost)

# Show the cost function
plt.plot(W_val, cost_val)
plt.show()
```

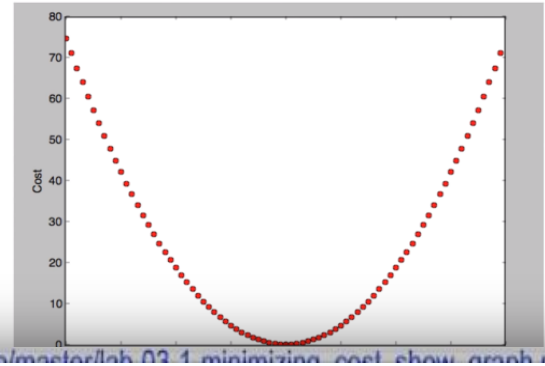
matplotlib http://matplotlib.org/users/installing.html

$$H(x) = Wx$$

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

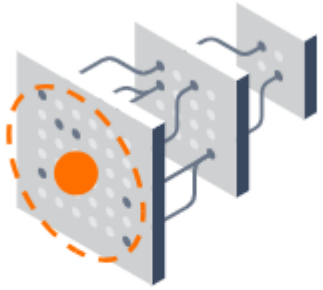
How to minimize cost?

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$



Source : https://youtu.be/Y0EF9VqRuEA

Python / TensorFlow



TensorFlow Mechanics

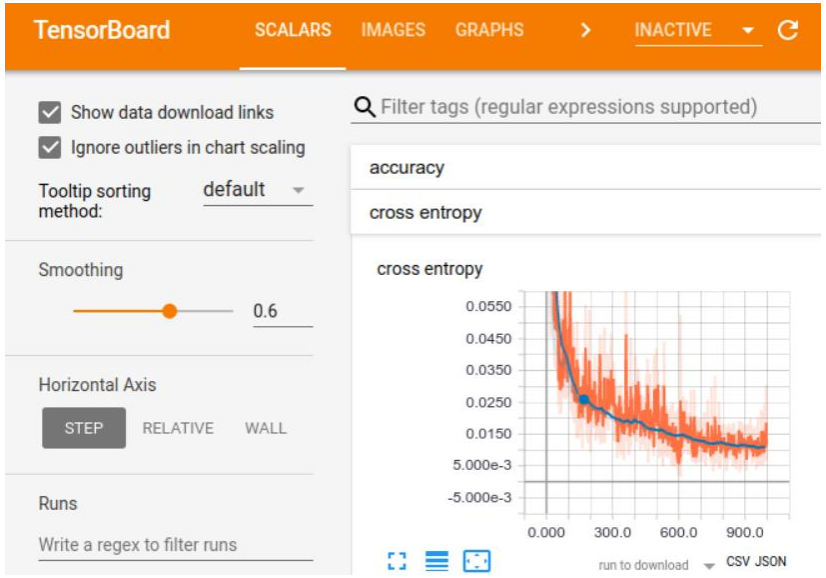
2 feed data and run graph (operation)
sess.run (op, feed_dict={x: x_data})

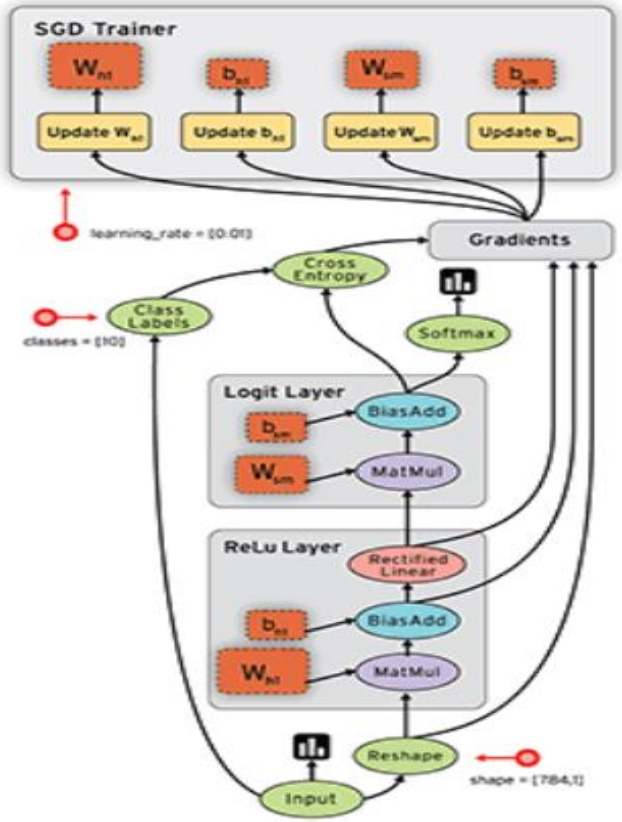
1 Build graph using TensorFlow operations

3 update variables in the graph (and return values)

WWW.MATHWAREHOUSE.COM

Easy model building





Source : https://www.tensorflow.org/

Gradient Descent algorithm (경사하강)

가설이 1차 선형이고 cost 함수가 2차 포물선인 경우 적용하여 W, b 값을 도출

What cost(W) looks like?

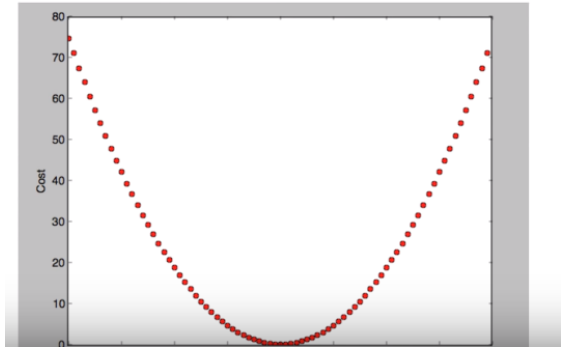
$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

x	Y
1	1
2	2
3	3

- W=1, cost(W)=?
- W=1, cost(W)=0
- W=0, cost(W)=4.67
- W=2, cost(W)=4.67

How to minimize cost?

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$



- Minimize cost function
- Gradient descent is used many minimization problems
- For a given cost function, *cost (W, b)*, it will find *W, b* to minimize cost
- It can be applied to more general function: *cost (w1, w2, …)*

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

Source :

Gradient Descent algorithm (경사하강) 을 수식으로 구현 (cost 최적화 algorithm)

가설이 1차 선형이고 cost 함수가 2차 포물선인 경우 적용하여 W, b 값을 도출

```
W = tf.Variable(tf.random_normal([1]), name='weight')
X = tf.placeholder(tf.float32)
Y = tf.placeholder(tf.float32)

# Our hypothesis for linear model X * W
hypothesis = X * W

# cost/loss function
cost = tf.reduce_sum(tf.square(hypothesis - Y))

# Minimize: Gradient Descent using derivative: W -= learning_rate * derivative
learning_rate = 0.1
gradient = tf.reduce_mean((W * X - Y) * X)
descent = W - learning_rate * gradient
update = W.assign(descent)

# Launch the graph in a session.
sess = tf.Session()
# Initializes global variables in the graph.
sess.run(tf.global_variables_initializer())
for step in range(21):
  sess.run(update, feed_dict={X: x_data, Y: y_data})
  print(step, sess.run(cost, feed_dict={X: x_data, Y: y_data}), sess.run(W))
```

```
import tensorflow as tf
x_data = [1, 2, 3]
y_data = [1, 2, 3]
```

- Minimize cost function
- Gradient descent is used many minimization problems
- For a given cost function, *cost (W, b)*, it will find *W, b* to minimize cost
- It can be applied to more general function: *cost (w1, w2, ...)*

$$cost(W) = \frac{1}{2m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2)x^{(i)}$$

$$W := W - \alpha\frac{\partial}{\partial W}cost(W)$$

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

Source : https://youtu.be/Y0EF9VqRuEA

Summary – 6 steps

Data

- x_train = [1, 2, 3]
- y_train = [2, 4, 6]

Weight bias

- W = tf.Variable(tf.random_normal([1]), name='weight')
- b = tf.Variable(tf.random_normal([1]), name='bias')

hypothesis

- hypothesis = x_train * W + b

cost

$$cost(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})^2$$

optimizer

manual step :
(W- 기울기 * lr) ➔ update

$$W := W - \alpha\frac{1}{m}\sum_{i=1}^{m}(Wx^{(i)} - y^{(i)})x^{(i)}$$

```
learning_rate = 0.1
gradient = tf.reduce_mean((W * X - Y) * X)
descent = W - learning_rate * gradient
update = W.assign(descent)
```

함수 사용
train= tf.train.GradientDescentOptimizer.minimize(cost)

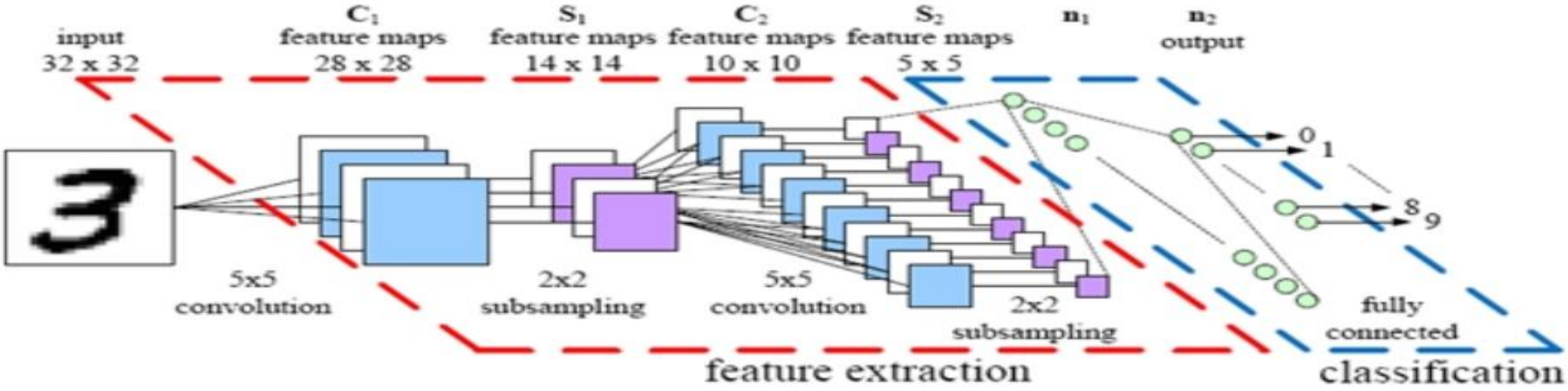
build graph in a session

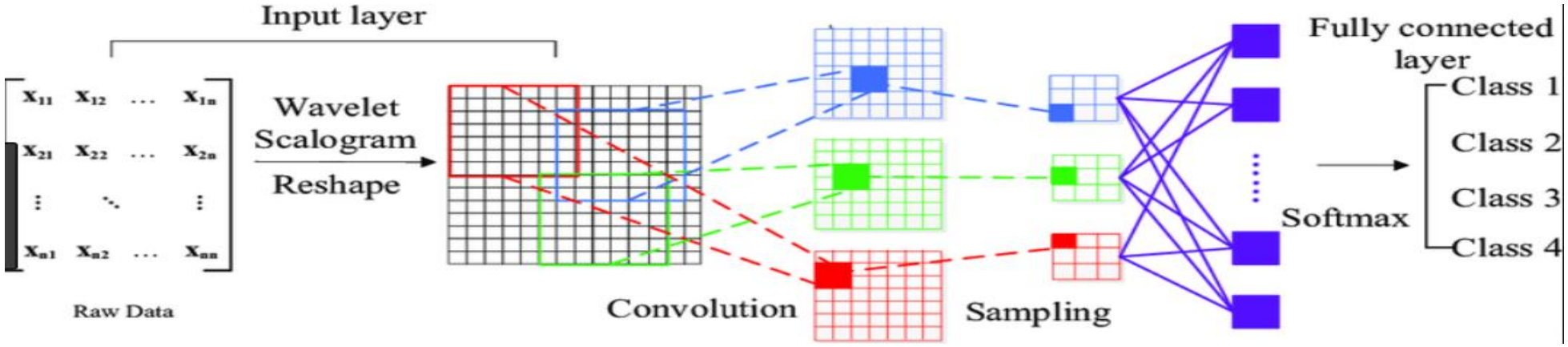
sess.run(update)

sess.run(train)

Source :

CNN





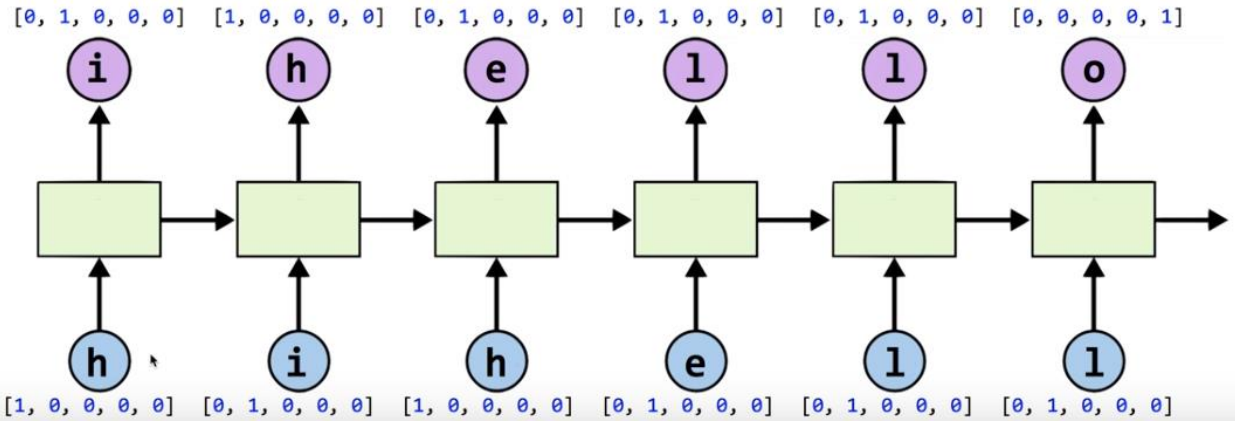
Source : https://deepnotes.io/ https://www.researchgate.net/figure/Architecture-of-convolutional-neural-network-model_fig4_322325843

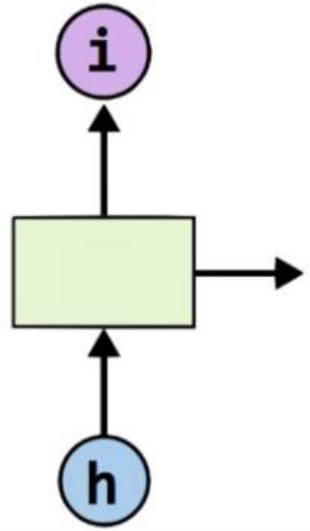
RNN : hi hello teaching

- text: 'hihello'
- unique chars (vocabulary, voc):
  h, i, e, l, o
- voc index:
  h:0, i:1, e:2, l:3, o:4

One-hot encoding

```
[1, 0, 0, 0, 0],   # h 0
[0, 1, 0, 0, 0],   # i 1
[0, 0, 1, 0, 0],   # e 2
[0, 0, 1, 0, 0],   # l 3
[0, 0, 0, 1, 0],   # l 3
[0, 0, 0, 0, 1],   # o 4
```





Source : https://youtu.be/39_P23TqUnw

목차

- 데이터 분석 개념 및 절차
- 활용사례 및 시사점
- 활용이 어려운 이유
- AI, Machine learning 기본 지식
- **Open source 활용 및 Demo**
- 데이터분석 아이디어 개발 절차

 # Open source 활용 및 시스템 Demo

- ✓ 통계, ML Quick review
- ✓ R
- ✓ Shiny
- ✓ Python (Linear Regression, NN for XOR)

- 기술 통계를 프로세스에 활용

표본

$$\text{평균 } \overline{X} = \frac{\sum X_i}{n}$$

$$\text{분산 } S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

$$\text{표준편차 } S = \sqrt{S^2} = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

$$Sk = \frac{3(\overline{X} - Md)}{S}$$

Sk : 피어슨의 비대칭도

S : 표준편차

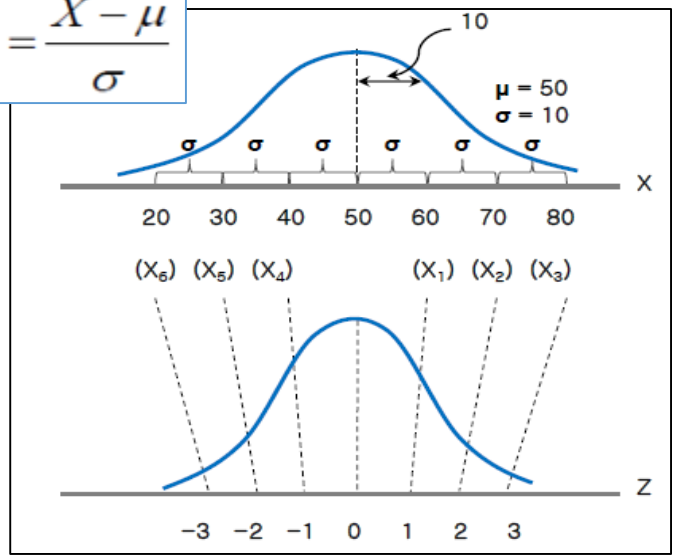
Md : 중앙값

$$S = \sqrt{\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2}$$

$$Z = \frac{X - \mu}{\sigma}$$

- 정규분포(평균 μ, 분산σ²)

확률변수 X는 X ~ N(μ, σ²)



어디에 어떻게 활용하는가 ?

?

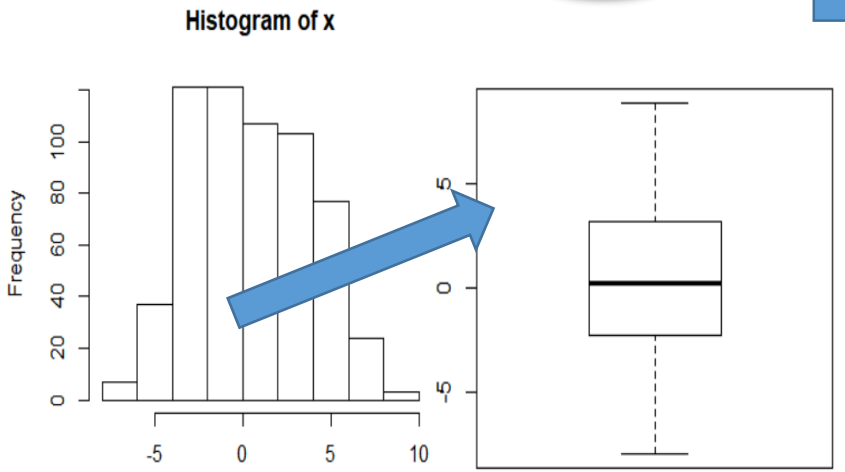
- 같은 평균이면 모집단이 같은가?
- 생산성이 관리되고 있는가?
- 개선의 효과가 있는가?
- 차이가 난 부분의 원인?
- 여러 변수 중에서 같은 부류는?

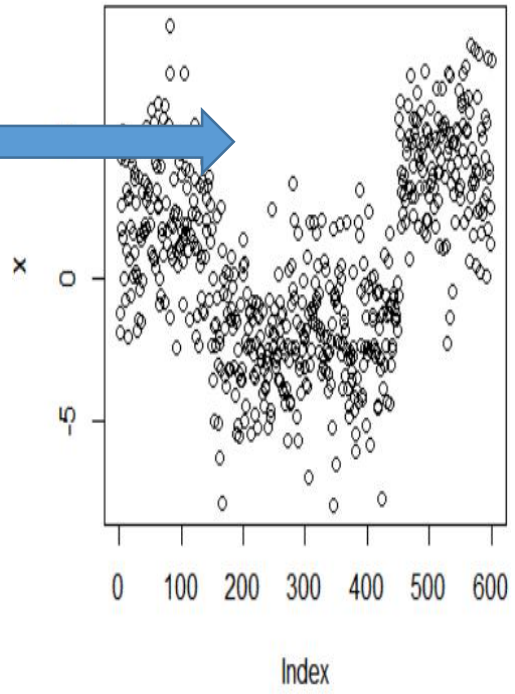
Source :

- 탐색만으로도 많은 시사점을 찾을 수 있다.

탐색	분류	해석	의미 조치

표본






```
a <- rnorm( 150, 2, 2 )
b <- rnorm( 300, -2, 2)
c <- rnorm( 150,  4, 2 )

x <- c(a, b, c)

summary(x)

hist(x)

boxplot(x)

plot(x)
```

모 집단

- 모집단 구성하는 소분류 a, b, c 분류를 찾아냄
- 차이 발생 이유를 해석
- 원하는 소분류의 강화 조치

Source :

R & Rstudio 설치

The R Project for Statistical Computing

Getting Started

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

RStudio

- **R의 통합개발환경(IDE, Integrated Development Environment)**
- **RSudio, StatEt …**

Installers for Supported Platforms

Installers	Size	Date
RStudio 1.2.1335 - Windows 7+ (64-bit)	126.9 MB	2019-04-08
RStudio 1.2.1335 - Mac OS X 10.12+ (64-bit)	121.1 MB	2019-04-08

RStudio Desktop
Open Source License

FREE

DOWNLOAD

Learn More

RStudio Desktop
Commercial License

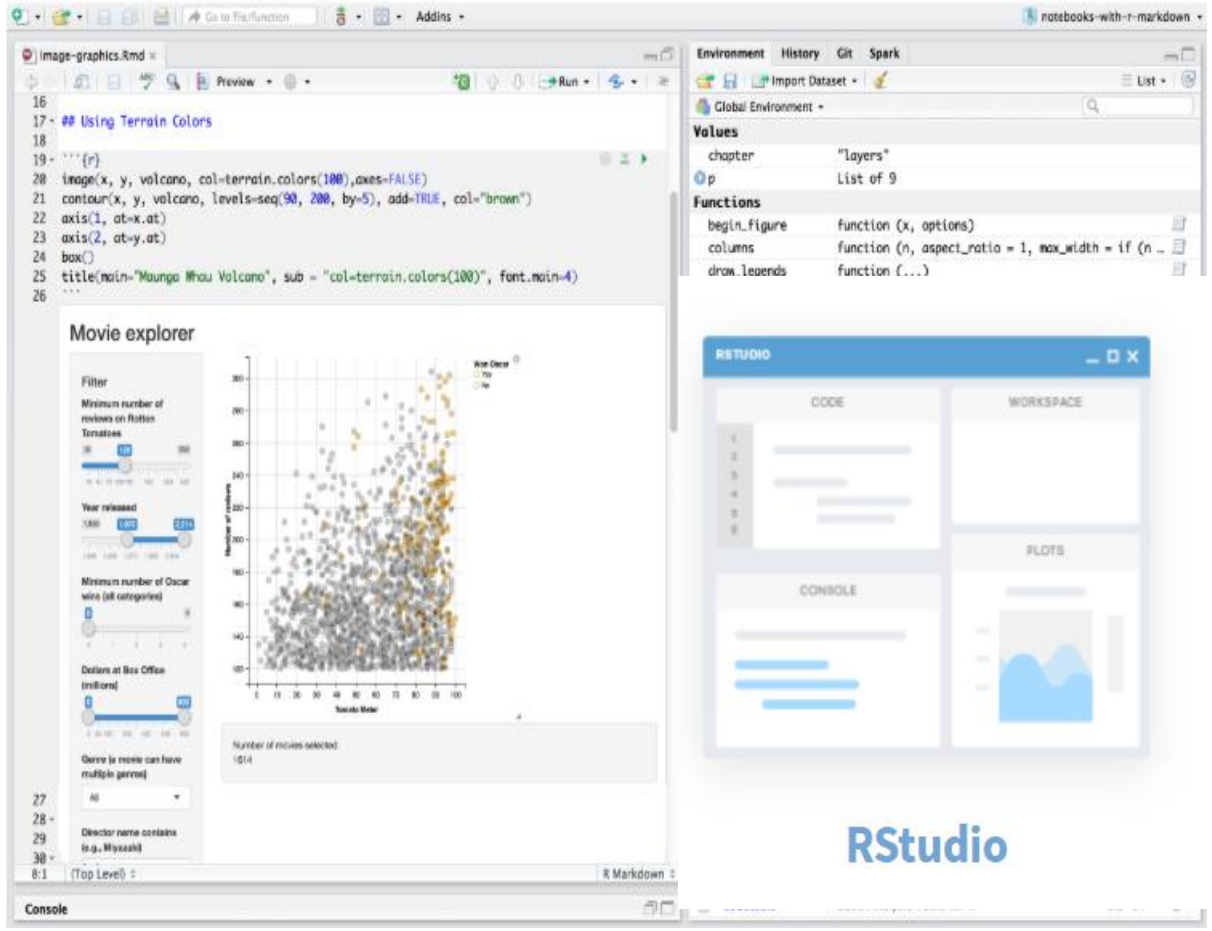
$995 per year

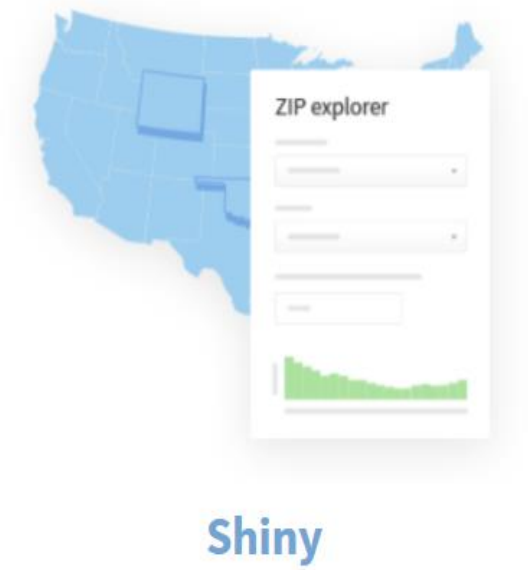
BUY

Learn More

Source : https://www.r-project.org/ https://www.rstudio.com/

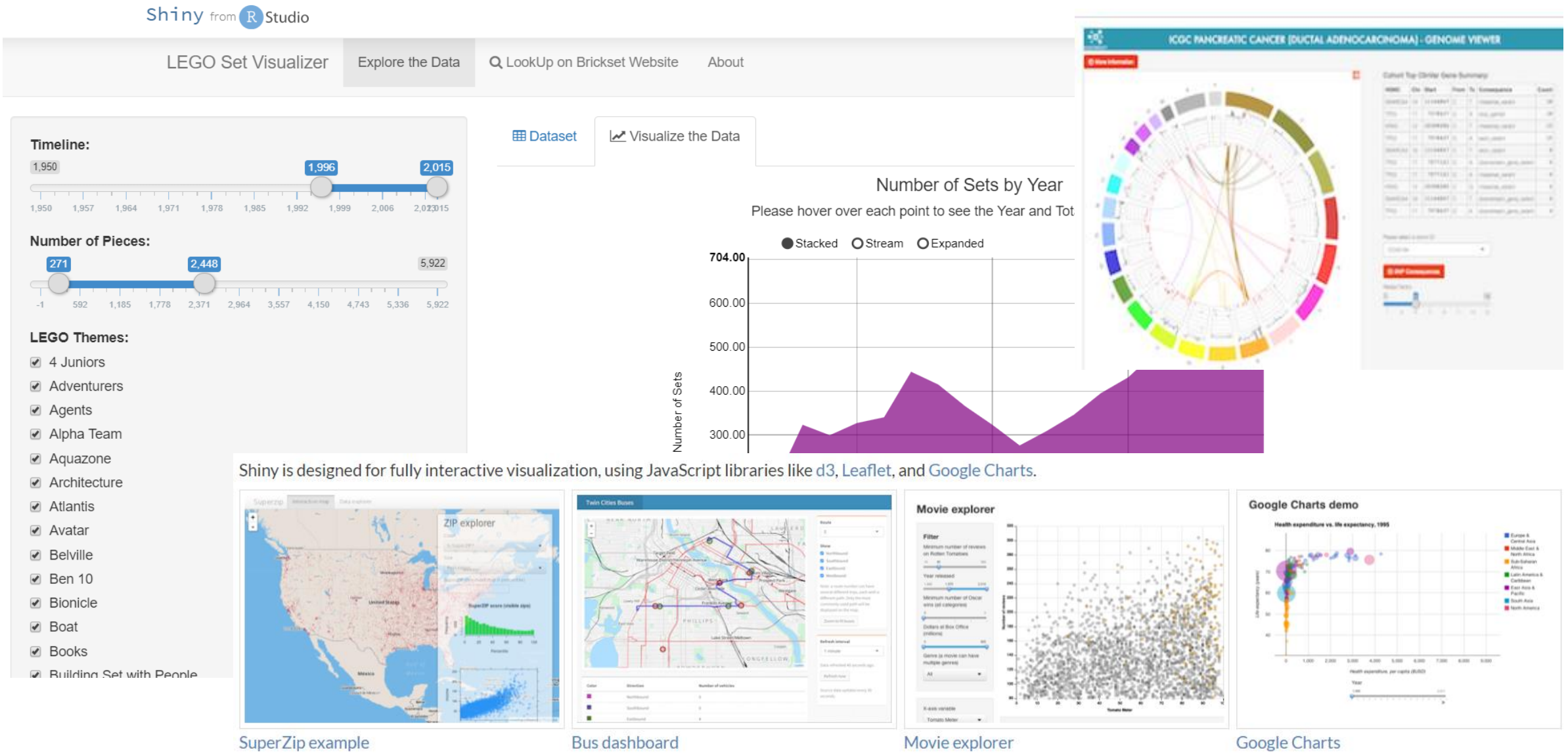
R의 Library는 계속 발전하고 있으며, 사용하기에 유용하다.





Source : https://www.r-project.org/ https://www.rstudio.com/

Shiny



Source：https://shiny.rstudio.com/

Machine Learning concept

- **Machine Learning 개념 이해를 위한 시스템 Demo**

  - ✓ **Linear Regression / Logistic / softmax**
  - ✓ **AI, XOR & Deep Learning (NN for XOR)**
  - ✓ **TensorFlow / Tensorboard**
  - ✓ **NN, ReLu, Xavier, Dropout, and Adam**
  - ✓ **RNN Basics**

시스템 Demo 과정에서 이해할 부분

ML vs DL

R vs Python

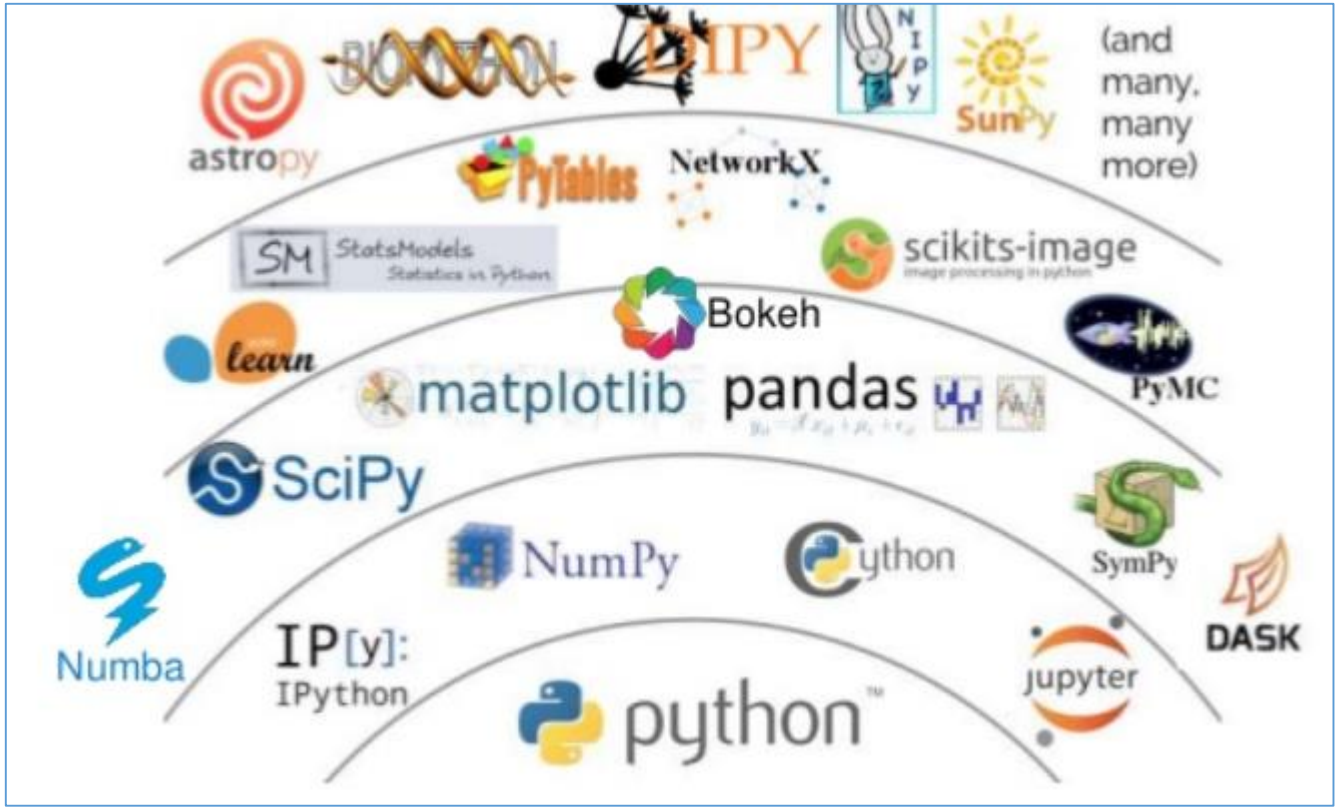
Linear Regression
Excel vs ML

CPPS – AI – BDA ?

Source : 321. ML_concept_basic.pptx

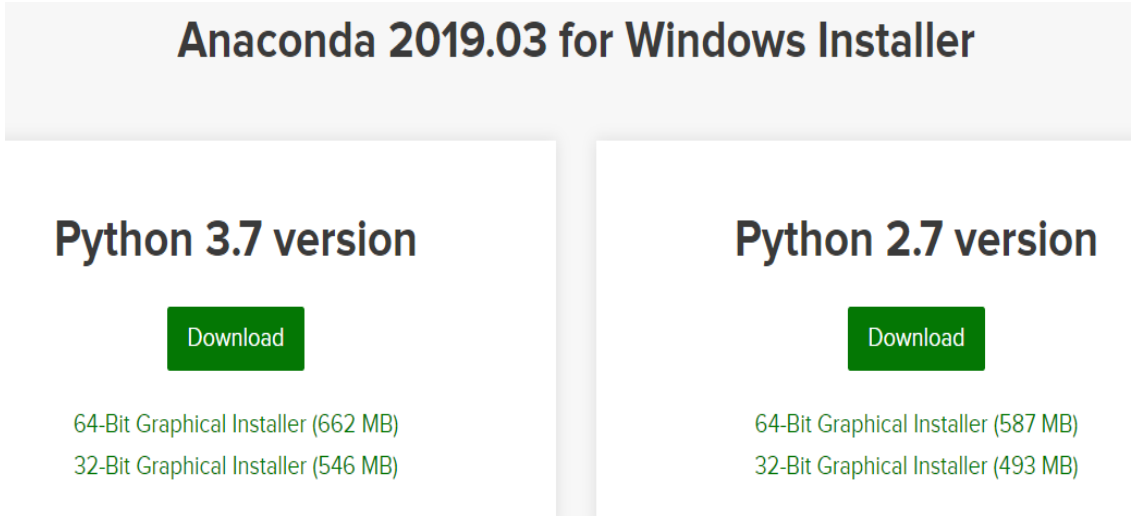
Python / tensorflow / numpy – Anaconda : spyder / Jupyter

Python은 함께 사용할 여러 Library 와 버전 문제가 되지 않도록 환경을 구분하여 운영하는 것에 필요한 배포판 Anacoda 를 이용 설치



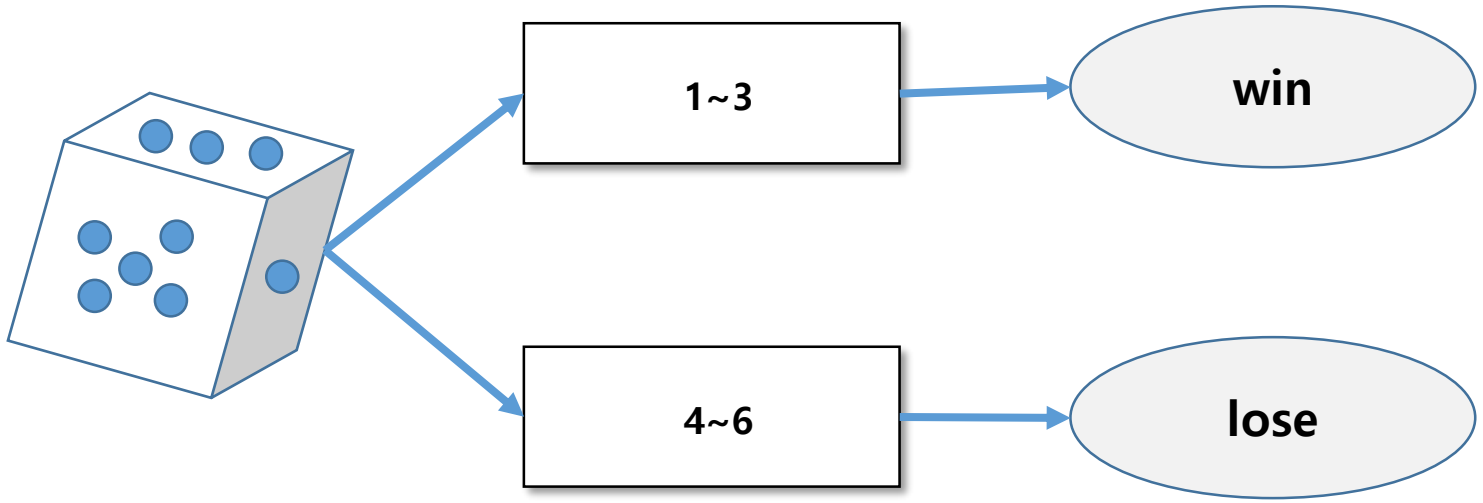
Source : https://www.slideshare.net/teoliphant/python-for-data-science-with-anaconda

AI / ML 을 위한 Anaconda 설치



Source : https://www.anaconda.com/distribution/

python 이용 AI 개념 이해 : 주사위 1~3 vs 4~6



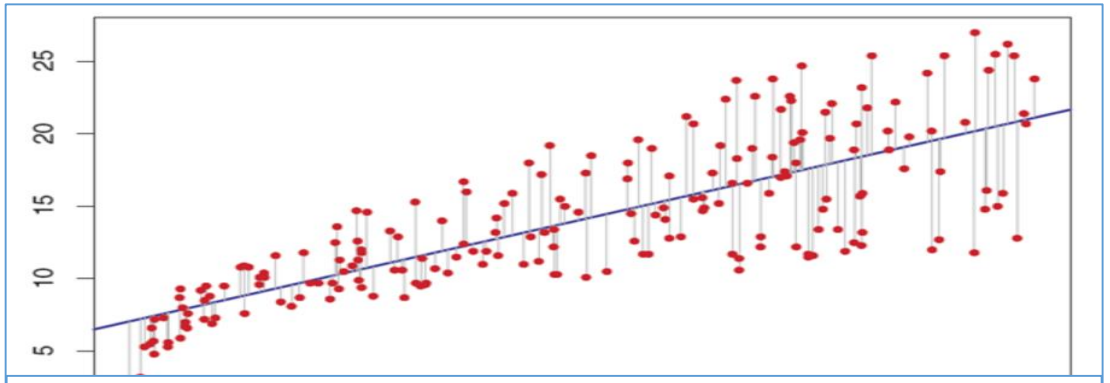
AI

RNN

Source : ML_python_demo

Linear regression

```
import tensorflow as tf
tf.set_random_seed(777)  # for reproducibility

1 x_train = [1, 2, 3]
y_train = [2, 4, 6]

2 W = tf.Variable(tf.random_normal([1]), name='weight')
b = tf.Variable(tf.random_normal([1]), name='bias')

3 hypothesis = x_train * W + b

4 cost = tf.reduce_mean(tf.square(hypothesis - y_train))

5 optimizer =
  tf.train.GradientDescentOptimizer(learning_rate=0.01)
train = optimizer.minimize(cost)
```



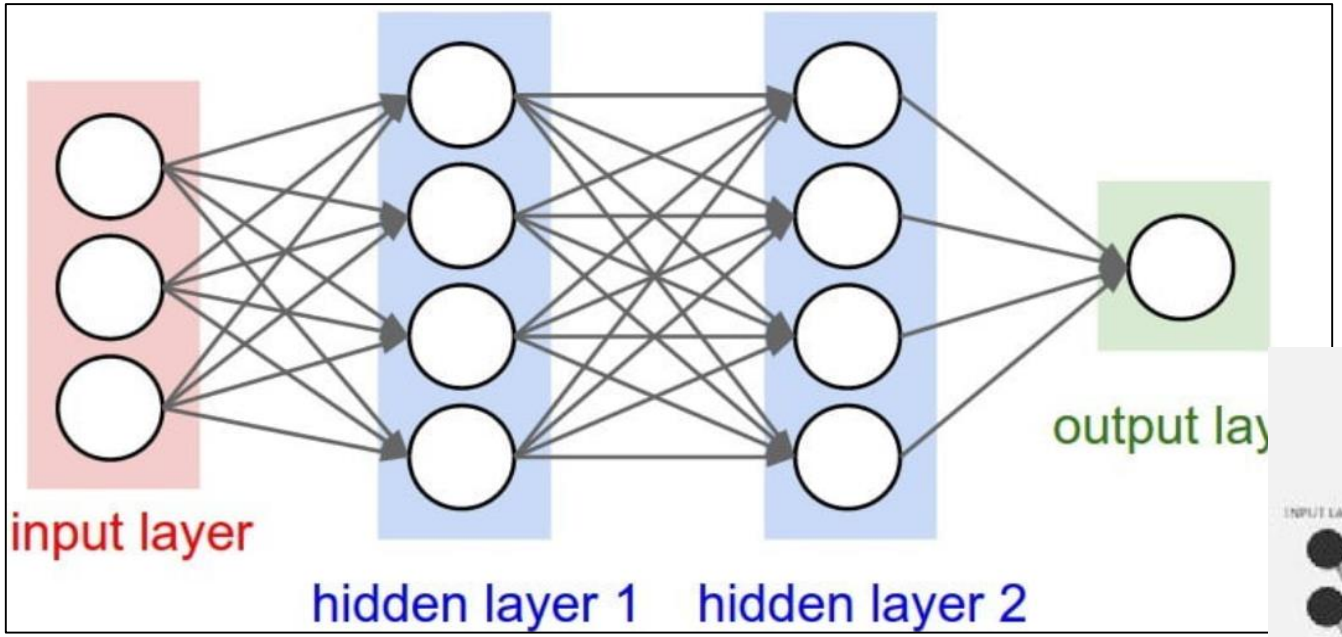
Source : ** https://www.matlabsolutions.com/blog/tensorflow-linear-regression-understanding-the-concept.php

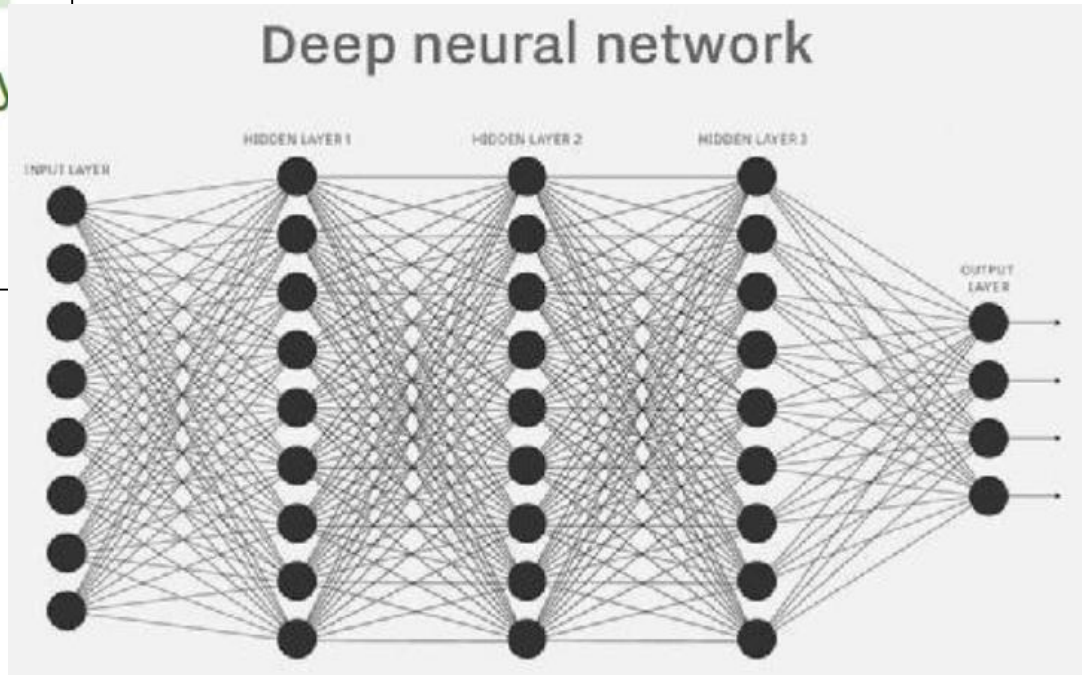
```
6 sess = tf.Session()

sess.run(tf.global_variables_initializer())

# Fit the line
# range 1001 : W 2.03 -> 2001 W  1.993

for step in range(2001):
    sess.run(train)
    if step % 20 == 0:
        print(step, sess.run(cost), sess.run(W), sess.run(b))
```

Source :

Concept – Deep learning

다층구조, backpropagation에서 적절한 weight 값을 찾기 위한 초기값과 함수





Source : https://searchenterpriseai.techtarget.com/definition/neural-network

Neural net for XOR

```
import tensorflow as tf
import numpy as np

tf.set_random_seed(777)  # for reproducibility
learning_rate = 0.01

x_data = [[0, 0],
          [0, 1],
          [1, 0],
          [1, 1]]
y_data = [[0],
          [1],
          [1],
          [0]]
x_data = np.array(x_data, dtype=np.float32)
y_data = np.array(y_data, dtype=np.float32)

X = tf.placeholder(tf.float32, [None, 2], name='x-input')
Y = tf.placeholder(tf.float32, [None, 1], name='y-input')
```

```
with tf.name_scope("layer1"):
    W1 = tf.Variable(tf.random_normal([2, 2]), name='weight1')
    b1 = tf.Variable(tf.random_normal([2]), name='bias1')
    layer1 = tf.sigmoid(tf.matmul(X, W1) + b1)

    w1_hist = tf.summary.histogram("weights1", W1)
    b1_hist = tf.summary.histogram("biases1", b1)
    layer1_hist = tf.summary.histogram("layer1", layer1)


with tf.name_scope("layer2"):
    W2 = tf.Variable(tf.random_normal([2, 1]), name='weight2')
    b2 = tf.Variable(tf.random_normal([1]), name='bias2')
    hypothesis = tf.sigmoid(tf.matmul(layer1, W2) + b2)

    w2_hist = tf.summary.histogram("weights2", W2)
    b2_hist = tf.summary.histogram("biases2", b2)
    hypothesis_hist = tf.summary.histogram("hypothesis", hypothesis)

# cost/loss function
with tf.name_scope("cost"):
    cost = -tf.reduce_mean(Y * tf.log(hypothesis) + (1 - Y) *
                           tf.log(1 - hypothesis))
    cost_summ = tf.summary.scalar("cost", cost)

with tf.name_scope("train"):
    train = tf.train.AdamOptimizer(learning_rate=learning_rate).minimize(cost)
```

Source :

Neural net for XOR

```
predicted = tf.cast(hypothesis > 0.5, dtype=tf.float32)
accuracy = tf.reduce_mean(tf.cast(tf.equal(predicted, Y), dtype=tf.float32))
accuracy_summ = tf.summary.scalar("accuracy", accuracy)

# Launch graph
with tf.Session() as sess:
    # tensorboard --logdir=./logs/xor_logs
    merged_summary = tf.summary.merge_all()
    writer = tf.summary.FileWriter("./logs/xor_logs_r0_01")
    writer.add_graph(sess.graph)  # Show the graph

    # Initialize TensorFlow variables
    sess.run(tf.global_variables_initializer())

    for step in range(10001):
        summary, _ = sess.run([merged_summary, train], feed_dict={X: x_data, Y: y_data})
        writer.add_summary(summary, global_step=step)

        if step % 100 == 0:
            print(step, sess.run(cost, feed_dict={
                  X: x_data, Y: y_data}), sess.run([W1, W2]))
```

```
Hypothesis:  [[1.02593222e-04]
 [9.99915719e-01]
 [9.99913216e-01]
 [1.01901845e-04]]
Correct:  [[0.]
 [1.]
 [1.]
 [0.]]
Accuracy:  1.0
```

```
# Accuracy report
h, c, a = sess.run([hypothesis, predicted, accuracy],
                   feed_dict={X: x_data, Y: y_data})
print("₩nHypothesis: ", h, "₩nCorrect: ", c, "₩nAccuracy: ", a)
```

Source :

통계(Excel 회귀분석) vs ML linear regression

Excel 회귀분석

x	y
1	2
2	3
3	7
4	9
5	11

요약 출력

회귀분석 통계량	
다중 상관계수	0.986
결정계수	0.973
조정된 결정계수	0.964
표준 오차	0.730
관측수	5

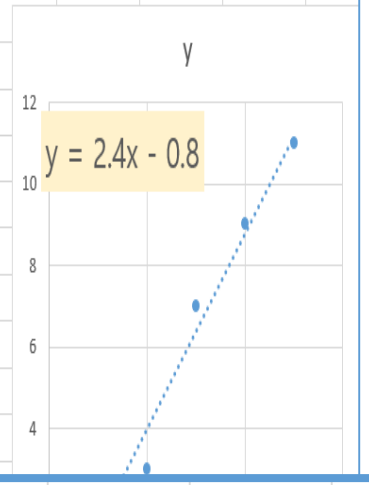
분산 분석

	자유도	제곱합	제곱 평균
회귀	1	57.6	57.6
잔차	3	1.6	0.533
계	4	59.2	

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	-0.8	0.766	-1.044	0.373	-3.238	1.638	-3.238	1.638
x	2.4	0.231	10.392	0.002	1.665	3.135	1.665	3.135

잔차 출력

관측수	예측치 y	잔차
1	1.6	0.4
2	4	-1
3	6.4	0.6
4	8.8	0.2
5	11.2	-0.2



		y=18+x1*3 + x2*4
13	2	65
15	7	91
1.7	9	59.1

ML 결과

A	B	C
1	1	25
2		
3		
4		
5		
6		
7		
8	5	62
9	6	69
10	7	76
11	8	83
12	9	90

```
# Ask my score
print("Your score will be ", sess.run(
    hypothesis, feed_dict={X: [[13., 2.]]}))

print("Other scores will be ", sess.run(hypothesis,
                                        feed_dict={X: [[15., 7.], [1.7, 9.]]}))
```

```
 [95.978294]]
Your score will be  [[68.44238]]
Other scores will be  [[100.62118 ]
 [ 48.523952]]
```

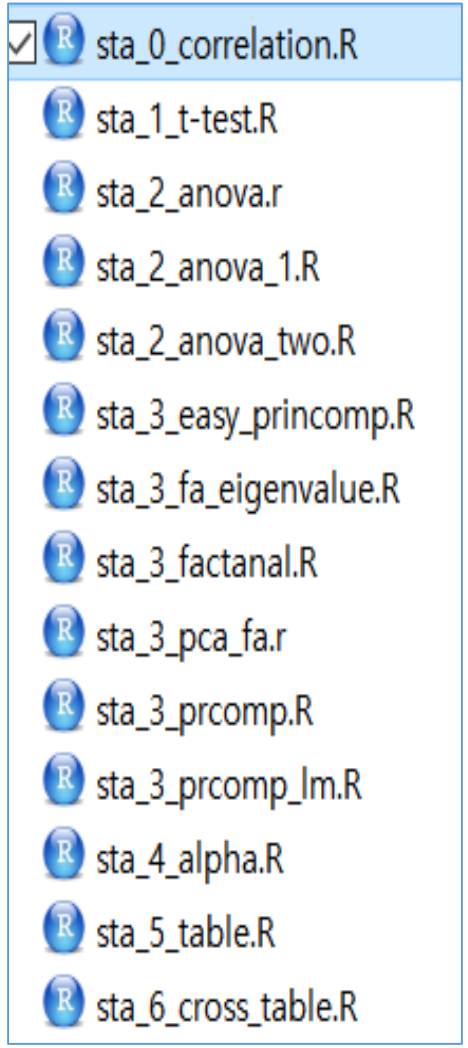
C:₩anaconda₩pytfworks/ mul-reg.csv

Source : 520. Statistics_R_ML.pptx / 312. DA_R1_concept.pptx, 312. DA_R2_code / 420. PL_R.pptx / 통계 R 학습_roc.xlsx / ML_01_from_01

시스템 DEMO : 통계분석 R code

주요 통계분석 R 활용

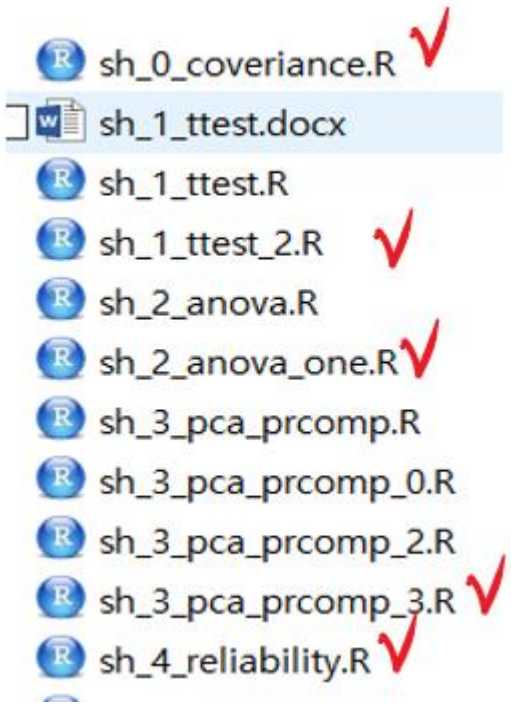
- 공분산
- 독립 t-Test (일표본, 대응표본, 독립표본)
- ANOVA (one-way, two-way, MANOVA)
- 요인분석 (PCA/FA)
- 신뢰도 분석
- 상관분석
- 교차분석
- 회귀 / 다중 회귀분석
- 로지스틱
- 군집분석
- …



계절지수를 통한 적정 개념에 대한 이해

Source : 430. System_demo_R_Python.pptx

시스템 DEMO : Shiny web



Shiny 확장성

Source : 430. System_demo_R_Python.pptx

시스템 DEMO : (Python, Tensorflow)

- Linear Regression
- Logistic / Softmax
- CNN
- RNN
- TensorBoard (anaconda/py2.7)
- 통계/수학적 기법(excel) vs ML

Source : 430. System_demo_R_Python.pptx

목차

데이터 분석 Idea 개발 프로세스 활용

개념을 이해하고 즉시 시도하는 단계

1. 빅데이터, 디지털 기술 이해
2. 고객가치 아디디어, 적용을 위한 use case 개발
3. Open source 활용 Pilot으로 개념 검증

데이터의 스마트화 프로세스

4. 데이터화
5. 정보/지능/융합 (지능화)
6. 가치실현 (스마트화)

IIoT
CLOUD
Edge computing
IoT Platform

데이터 분석 솔루션
가시화
분석 모델
융합

Source :

데이터 분석(고객가치) Idea 개발 프로세스

1. 데이터, 디지털 기술 이해
2. 고객 가치 Idea
3. 방향 설정 및 목표/기대효과
4. 데이터화
5. 정보/지능/융합
6. 가치 실현

- 디지털 기술 이해
- 데이터 이해
- 데이터 분석 이해
  - ✓ 통계 / Data Mining
  - ✓ AI / Machine Learning

- 고객 가치 Idea 도출/개발

- 즉시 적용할 수 있는 것부터 시작

디지털 분임조 활동

- Use case Idea / Piot
- 사전 준비사항 도출
- 적용을 위한 준비

- 데이터화
- IIoT
- IoT Gate way
- CLOUD

- 외부 전문가와 협업 (데이터 준비, 적용 모델 선택, 평가)

Data Lake

- 고객 가치 (데이터 분석 + H/W)
- 제품의 디지털화
- Telemetric / SCP
- Robot / Machine
- AR / VR / MR
- Connected IIoT

CPPS

Shop Control (real time)

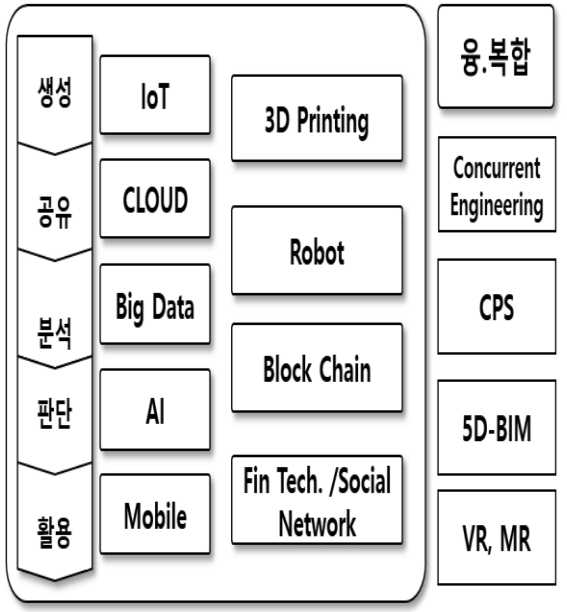
Edge Computing

Source :

자율 학습과 Open source를 이용하는 디지털 분임조 같은 스스로 학습 조직의 운영 검토

자율 학습 (기술 이해 및 융/복합 Case study)

디지털 기술



Program Language

- ✓ Python
- ✓ JSP

데이터 분석

- ✓ R (통계 이해)
- ✓ Machine Learning

웹 시스템 개발

- ✓ 정보 조회 및 제어 (통신)
- ✓ Data 분석 시스템

시스템 개발과정의 결과물

- 고객 경험 이해
- Digital 개념 이해
- 필요역량 및 준비사항

Next Steps

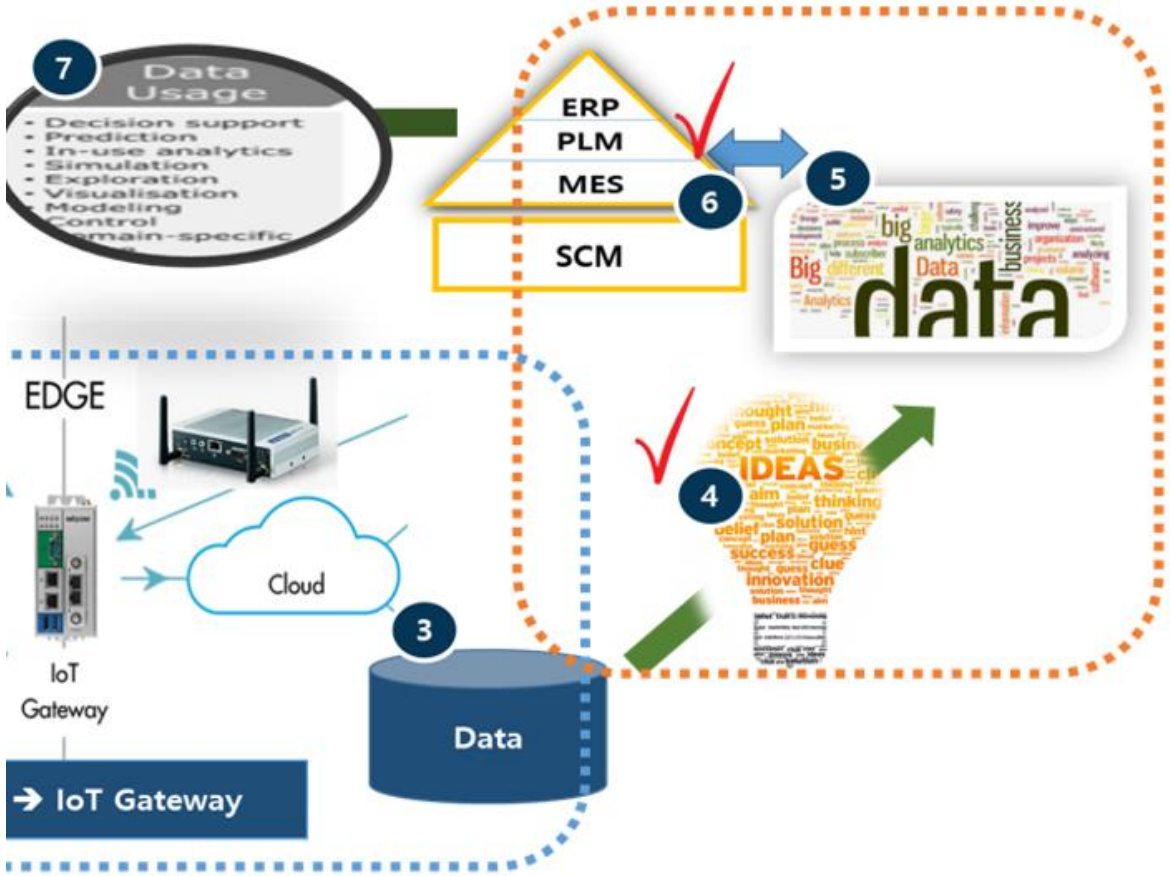
- Business Model
- Roadmap 준비
- Approach 방식

- ✓ Disruptive Innovation
- ✓ Open Innovation

Source :

❷ 고객 가치 Idea : 데이터 활용을 위한 흥미로운 질문을 해보라.

문제 해결과 고객을 위한 새로운 가치에 대하여 막연히 AI를 적용할 수 있지 않을까 하고 생각에서 출발하여, AI 를 어떤 방식으로 사용하고 그로 인한 기대효과 및 고객측면에서 가치에 대한 아이디어를 도출



새로운 가치에 대한 고민으로부터 시작된다.

흥미로운 질문을 해보라!

Source :

데이터 분석절차의 시작은 새로운 가치에 대한 고민으로부터 시작된다.

적용 Idea
얻고자 하는 것, 누가, 왜 원하는가 ?

어디를 대상으로

어떻게

- Decision Making (w/Report)
- Monitoring
- Real time control
- Optimization

- Cost
- Speed
- Quality
- Risk

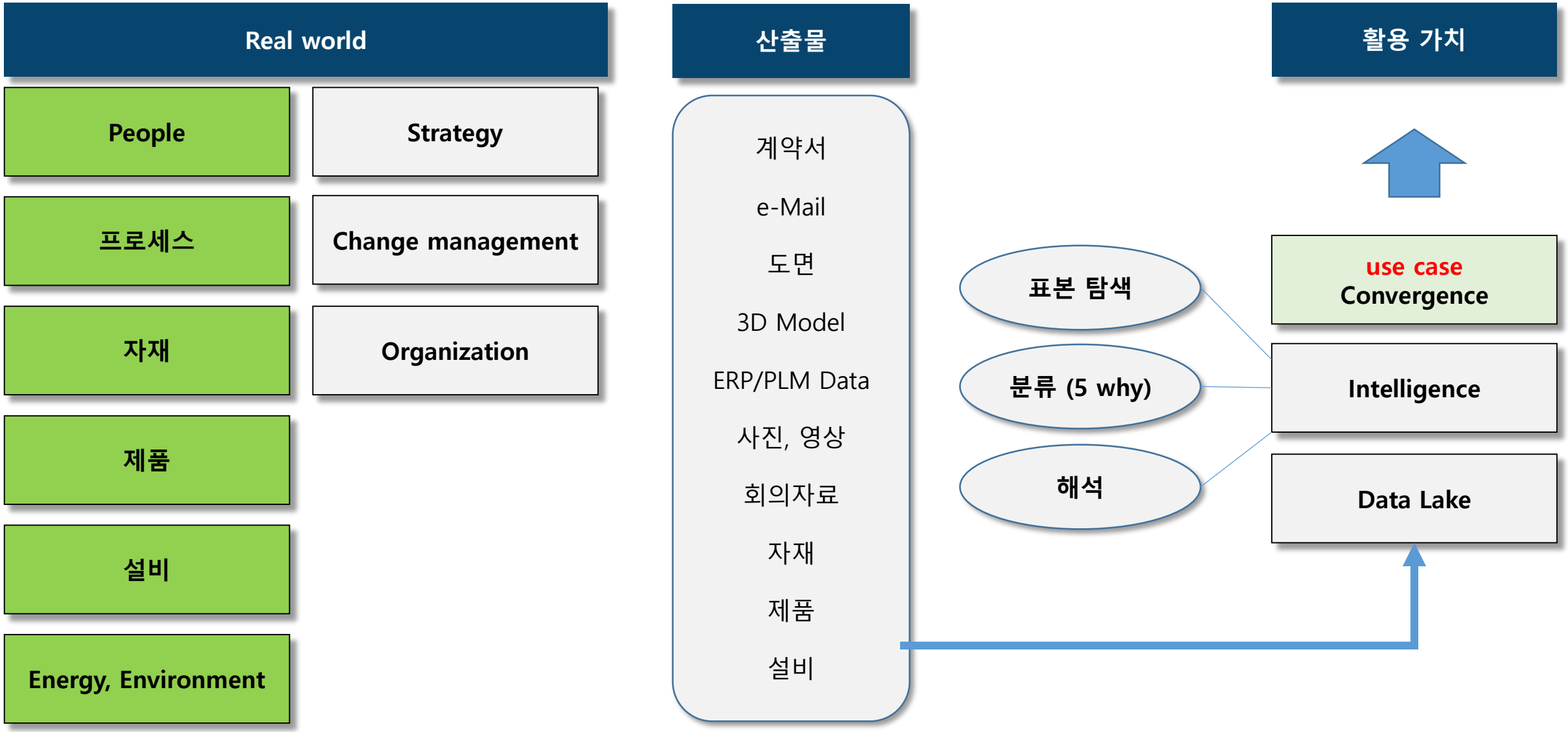
Q1 : 어디에, 왜 사용하여 어떤 가치를 낼 수 있는가 ?
- ✓ 남들은 어떻게 하는가 ?
- ✓ 어떤 솔루션들이 있는가 ?

Q2 : 우리도 가능할까 ? 가능하다면 어떤 준비가 필요할까 ?

Q3 : 남들이 하지 않는 새로운 가치를 내기 위해선 ?

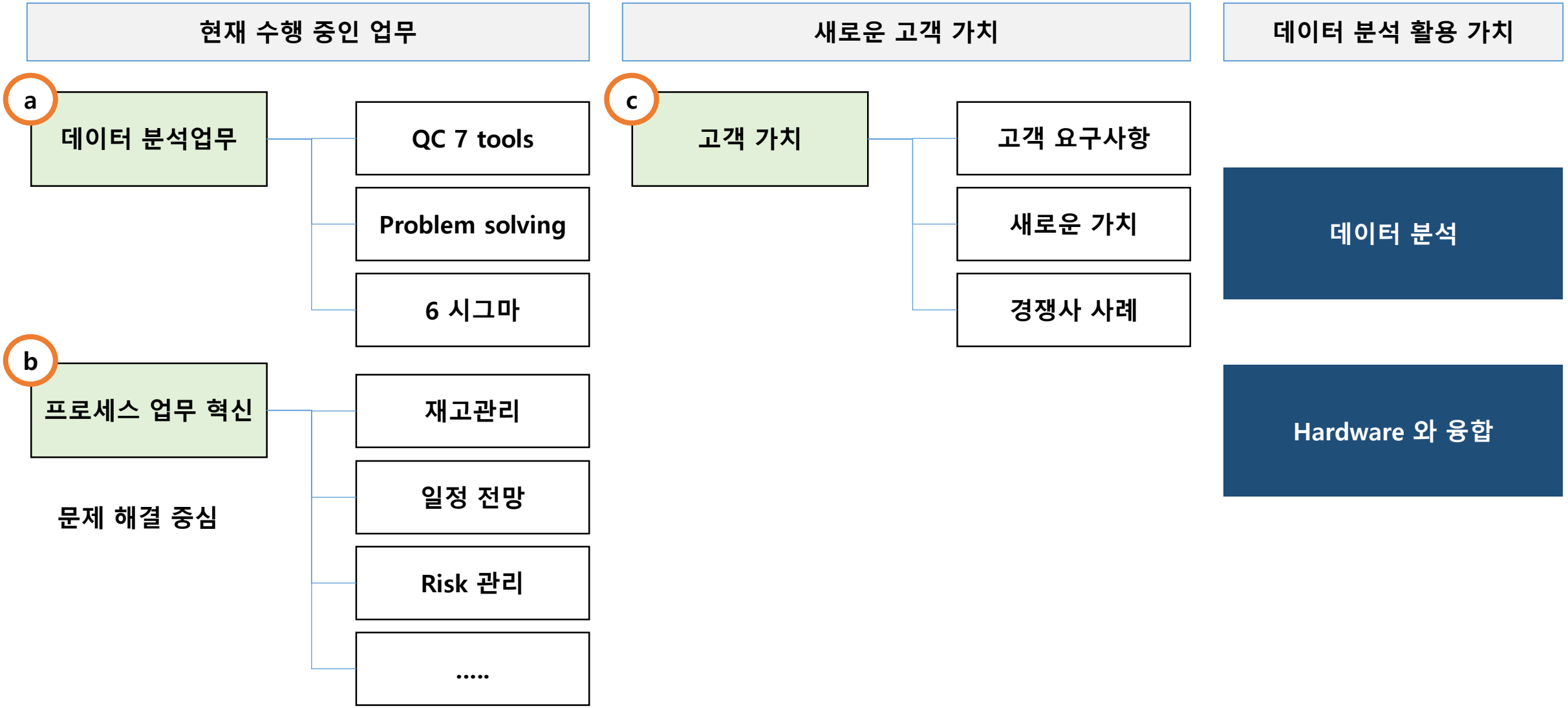
Source :

즉시 적용할 수 있는 것부터 시작



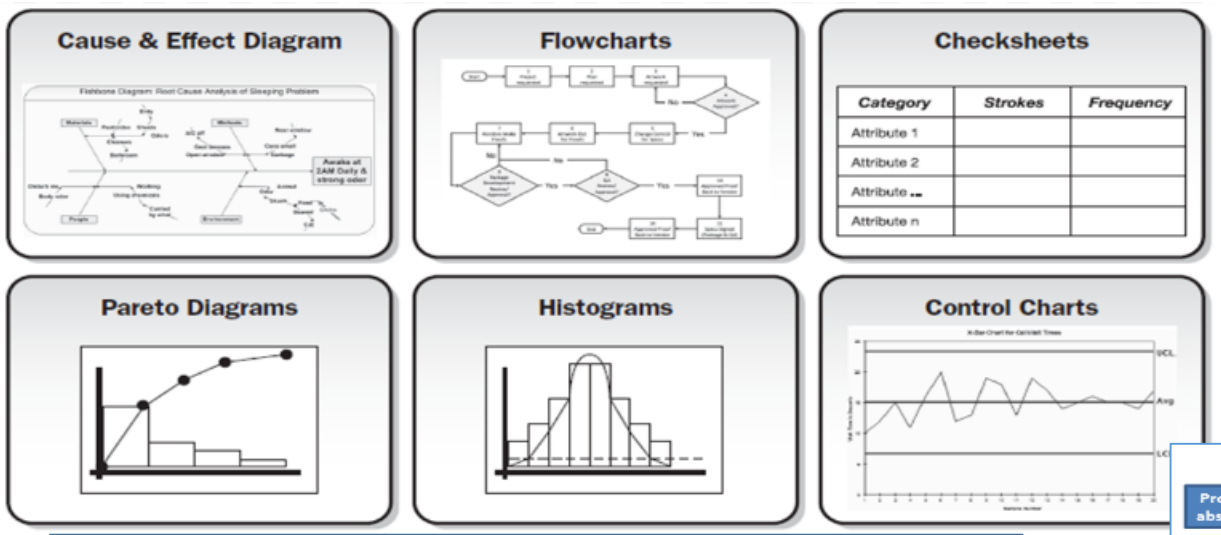
Source :

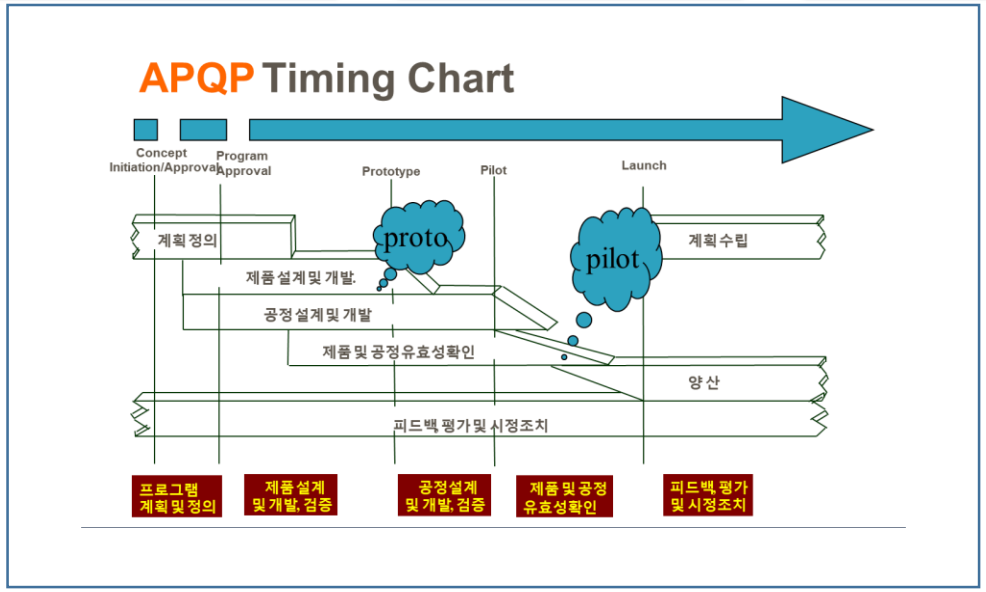
적용 대상

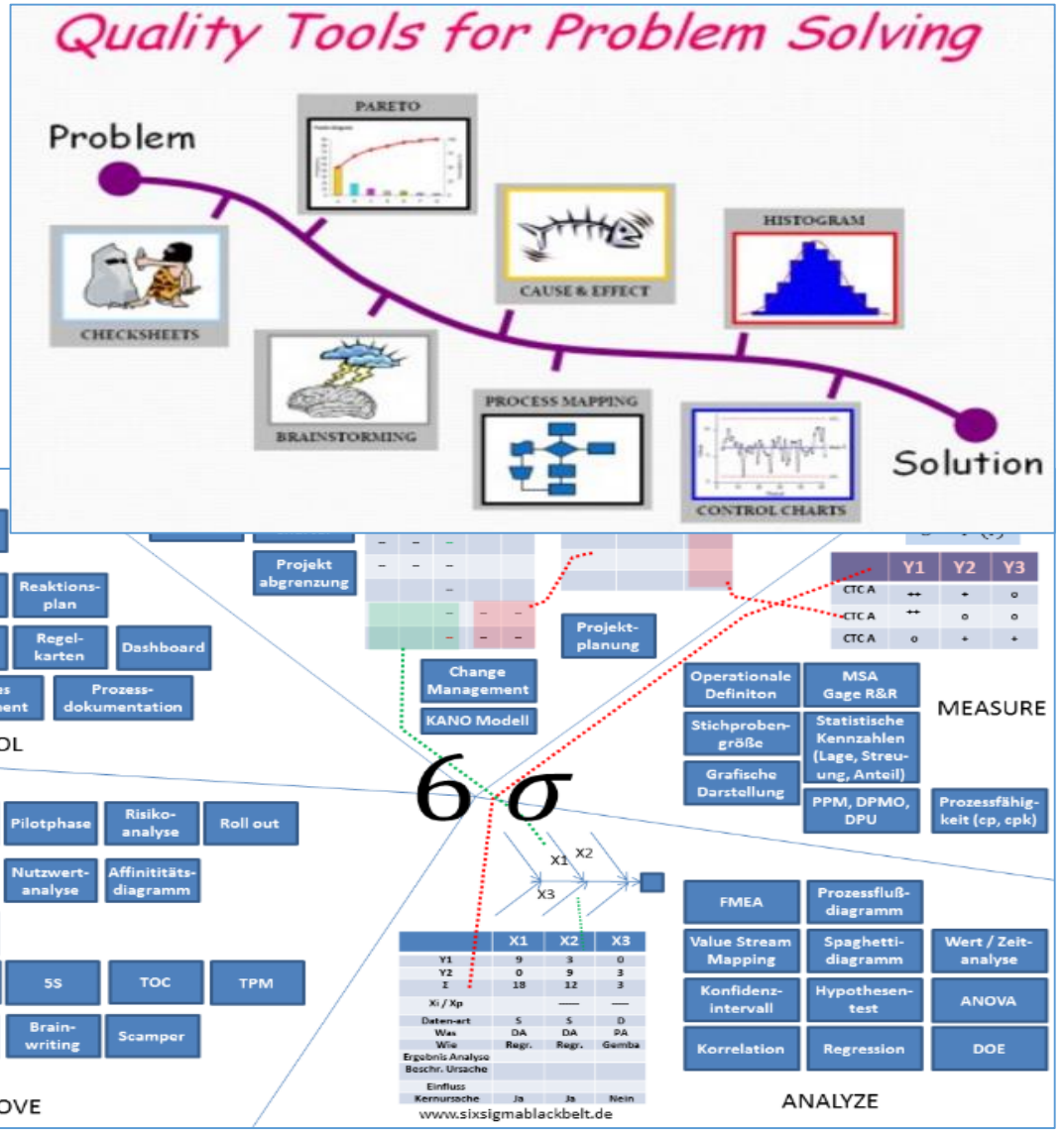

현재 수행 중인 업무
새로운 고객 가치
데이터 분석 활용 가치
a
데이터 분석업무
QC 7 tools
Problem solving
6 시그마
b
프로세스 업무 혁신
재고관리
일정 전망
Risk 관리
.....
문제 해결 중심
c
고객 가치
고객 요구사항
새로운 가치
경쟁사 사례
데이터 분석
Hardware 와 융합

Source :

a QC 7 tools, 6 시그마 등 : 특정 목적 기반 데이터 분석



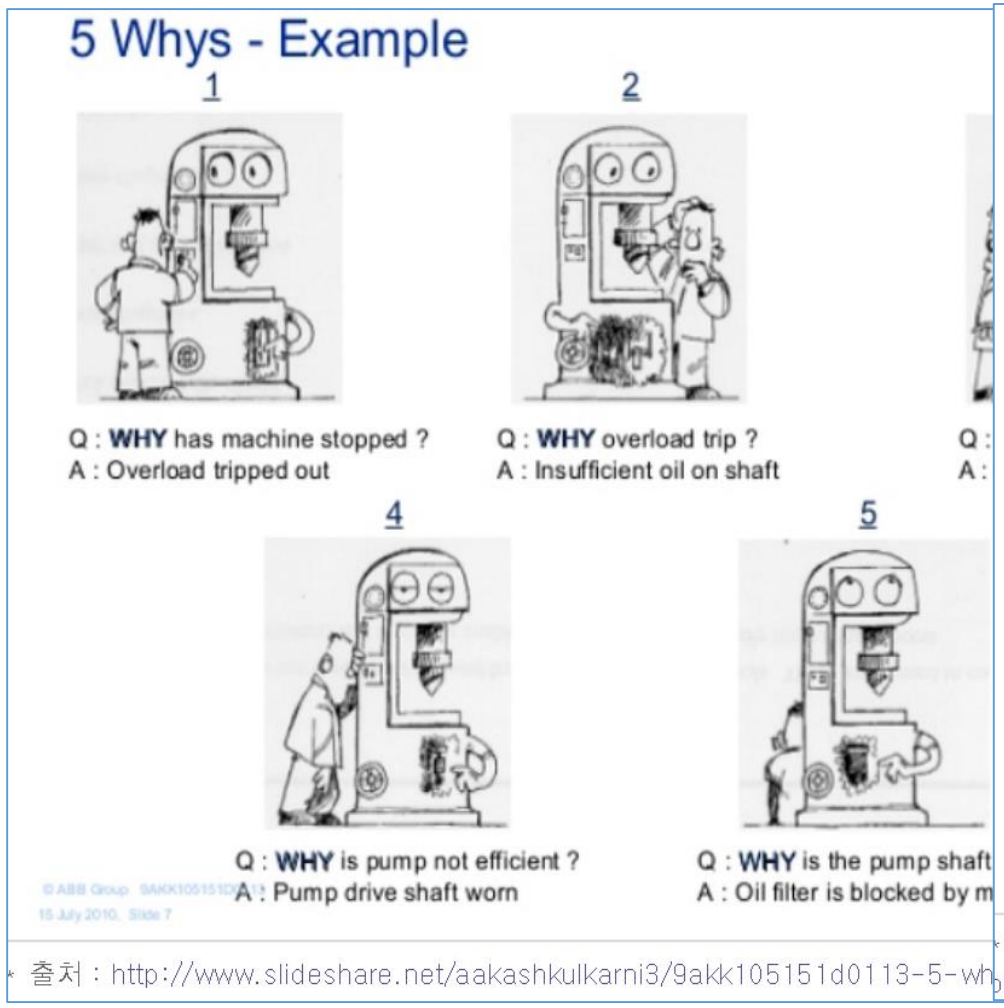




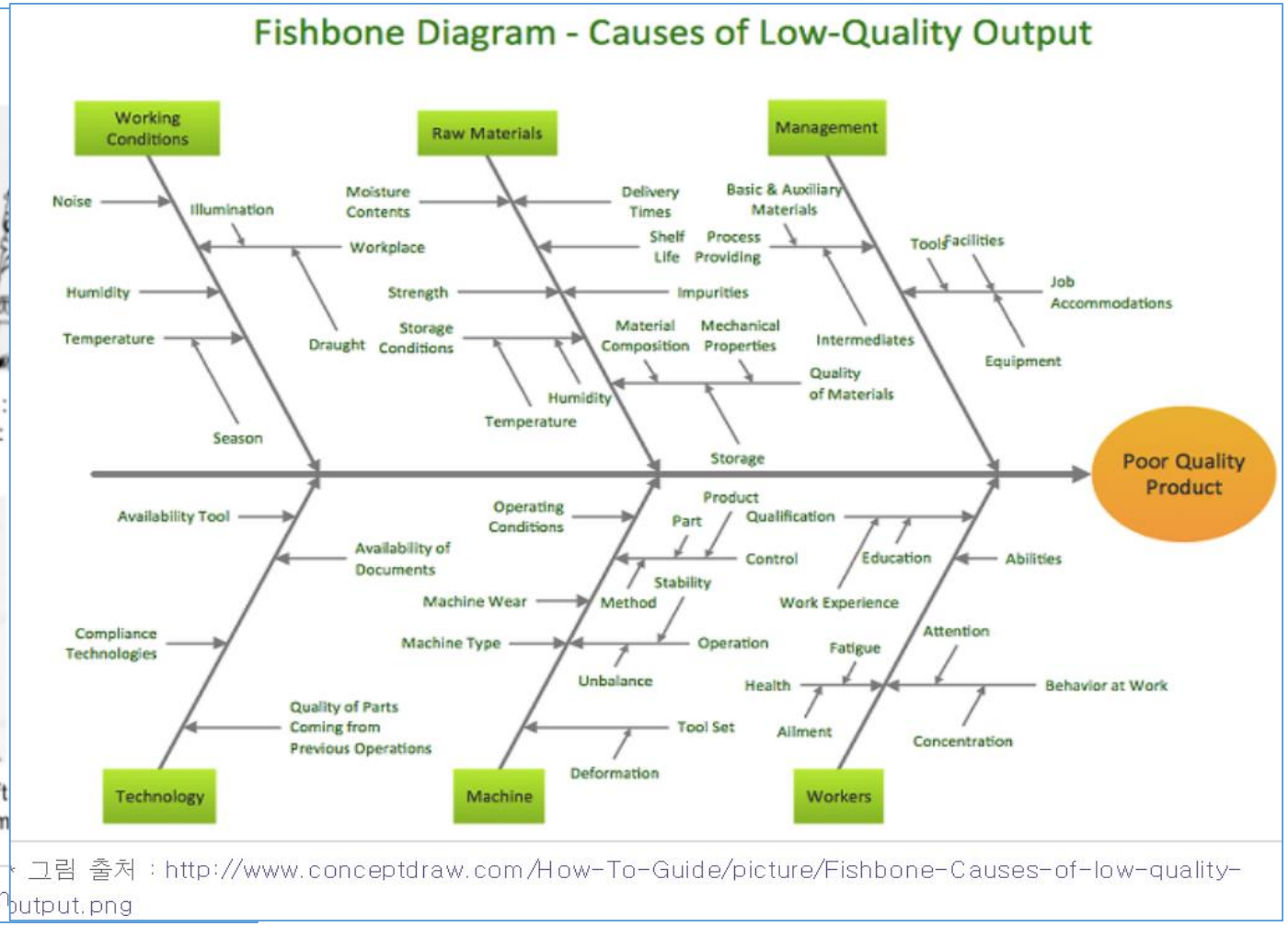
Source :

QC 7 tools – 원인 분석과 결과에 대한 해석을 위해 통계 활용을 재 점검

Cause & Effect | **Pareto Diagrams** | **Histograms** | **Control Charts** | **Scatter Diagrams**



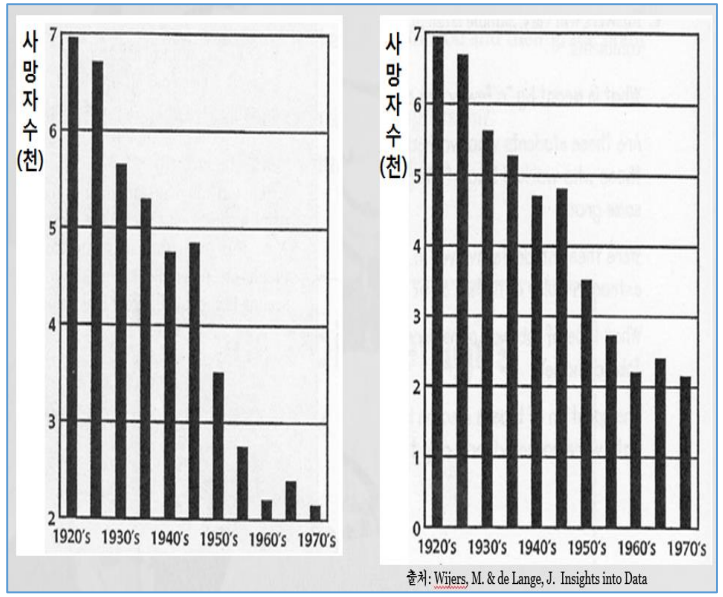
* 출처 : http://www.slideshare.net/aakashkulkarni3/9akk105151d0113-5-wh

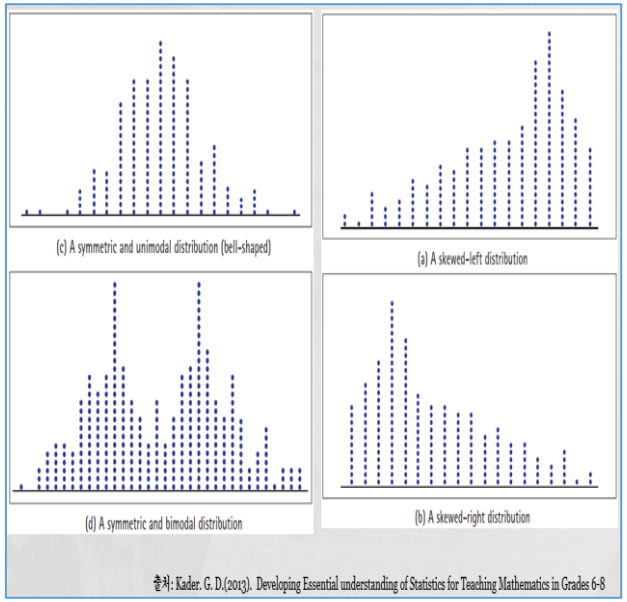


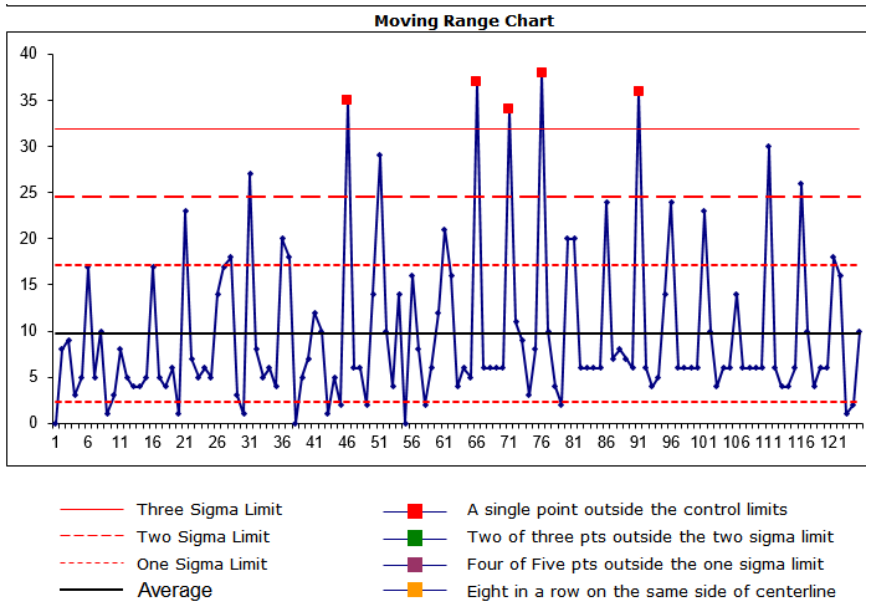
* 그림 출처 : http://www.conceptdraw.com/How-To-Guide/picture/Fishbone-Causes-of-low-quality-output.png

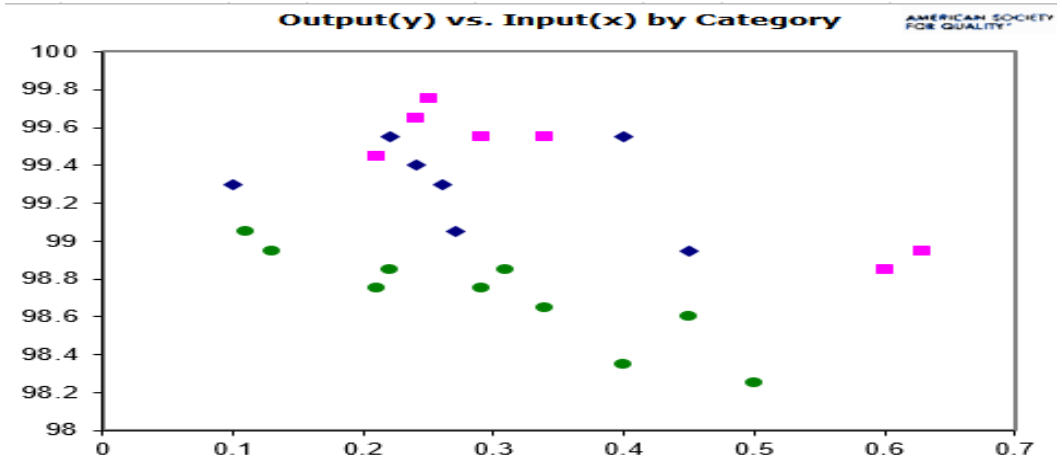
Source :

Problem solving & 6 시그마 : 올바른 의사결정을 위해 가능한 수단 활용









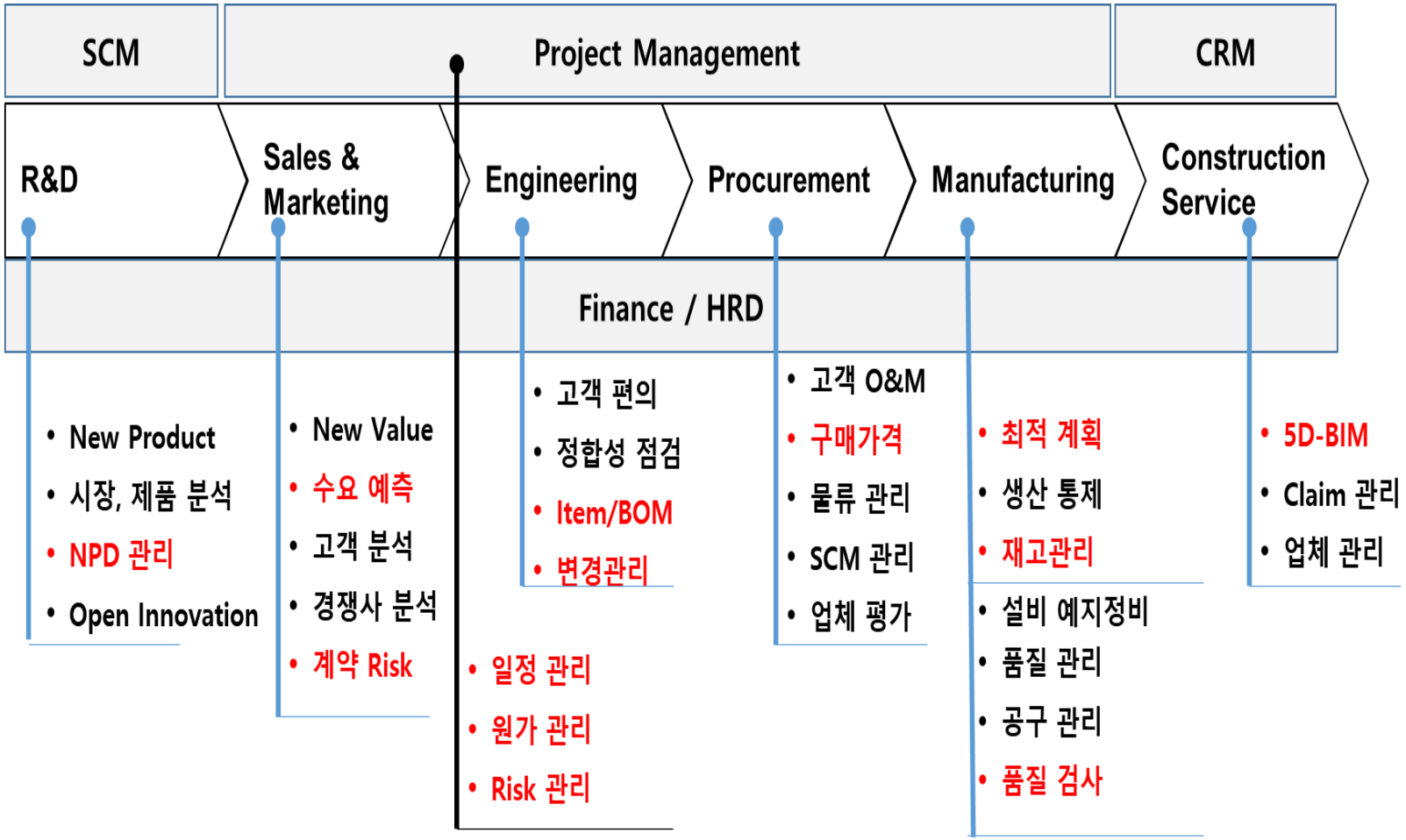
7 BASIC QUALITY TOOL TEMPLATES

These templates will help you get started using the seven basic quality tools. Just download th

- Cause and Effect / Ishikawa / Fishbone Diagram Template (Excel)
- Check Sheet Template (Excel)
- Control Chart (Excel)
- Histogram (Excel)
- Pareto Chart (Excel)
- Scatter Diagram Template (Excel)
- Stratification Diagram (Excel)

Source : https://asq.org/quality-resources/seven-basic-quality-tools

b 프로세스별 AI, ML 적용 가능 use case Idea 대상



?

- 혁신과 효율 관점으로 어떻게...

Source : 332. DA_use_case

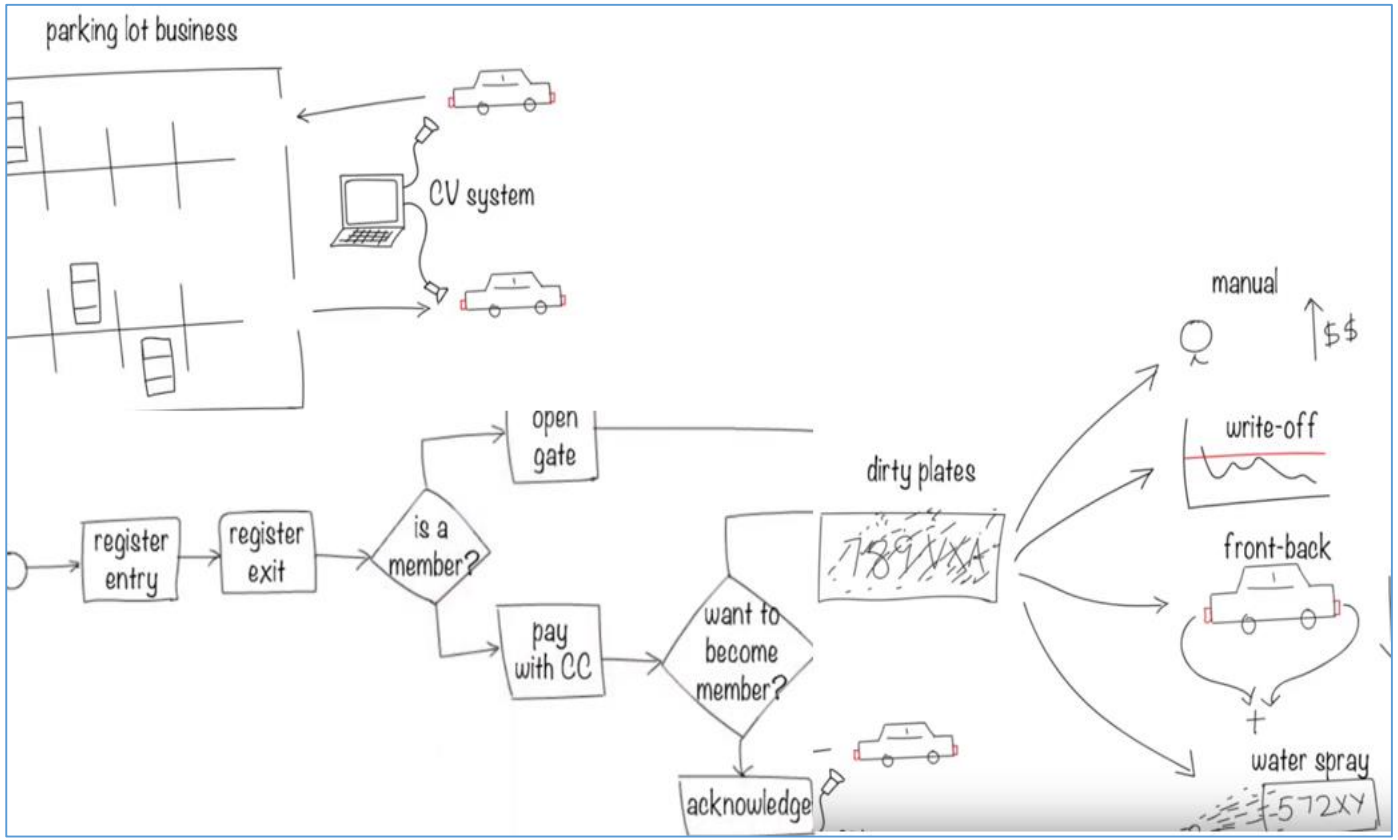
문제 해결을 위한 시도

- 문제 정의

- 문제해결을 위한 각종 방법론 (7 Steps, TRIZ) +
- 다양한 디지털 기술

- Pilot으로 신속한 검증

사례 참조



Source : how to apply ai in business, https://youtu.be/N_eHmaRf9T4

- 재고관리에도 AI를 적용할 수 있다.

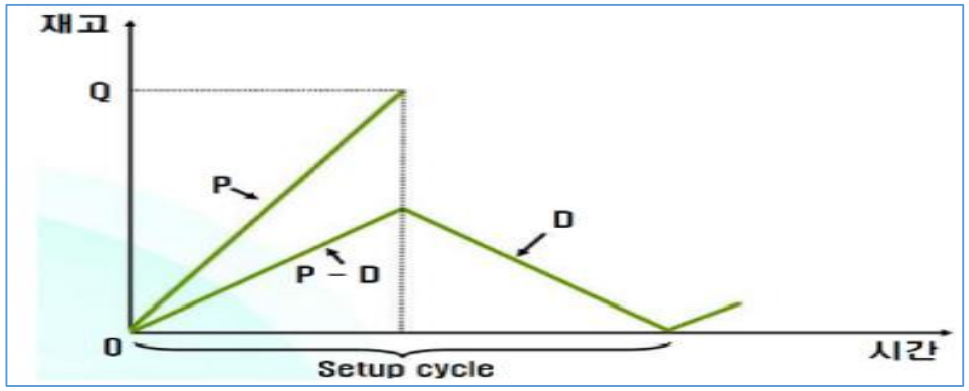
EOQ | **EPQ** | 연간 수요 | 평균 수요 | 표준 편차

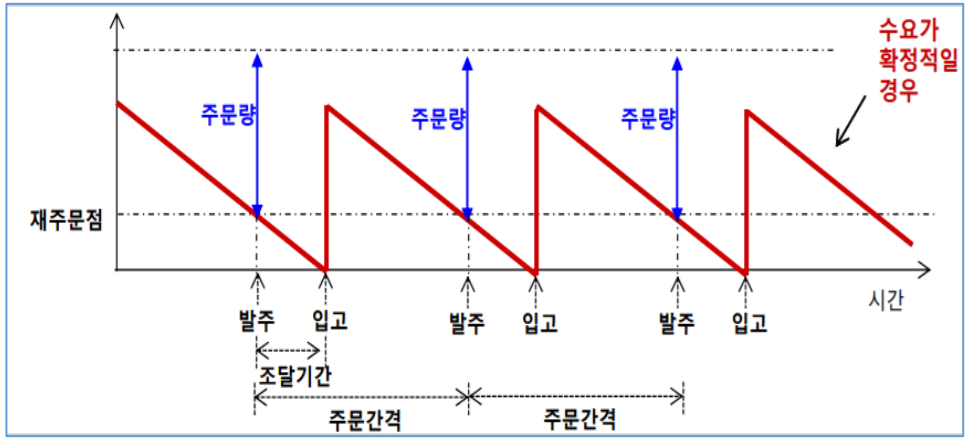
- 경제적주문량 모형 - EOQ (Economic Order Quantity).
- 경제적생산량 모형 - EPQ (Economic Production Quantity).
- 수량할인모형 (Quantity Discount).
- 확률적모형 (Stochastic models).
- 단일기간모형 (Single Period model).

$$Q = \sqrt{\frac{2\overline{D}S}{H}}$$

$$R = \mu + z\sigma$$

$$s = z\sigma$$

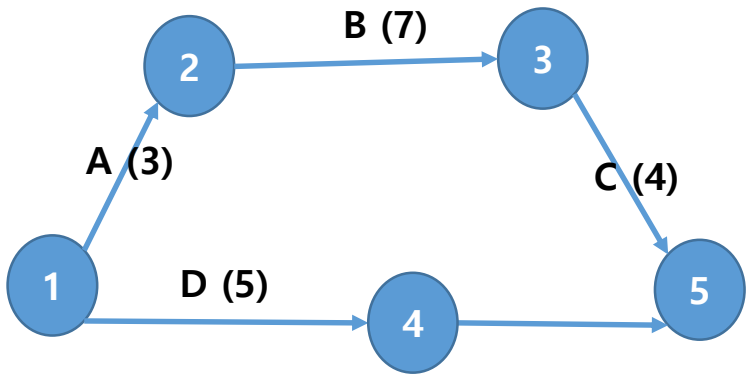




Source :

- AI이용 일정관리를 기존 방식에서 벗어나 효과적으로 할 수 있다.

pert/cpm 이용한 기존 방식



기대소요시간

$$t_e = \frac{(t_o + 4t_m + t_p)}{6}$$

분산

$$\sigma_t^2 = \left(\frac{t_p - t_o}{6}\right)^2$$

작업	시간추정치			기대소요시간	분산
	to	tm	tp		
A	2	3	4	3.00	0.11
B	4	7	10	7.00	1.00
C	2	4	6	4.00	0.44
D	4	5	6	5.00	0.11
E	2	3	4	3.00	0.11

$$Z = \frac{X - \mu}{\sigma}$$

Path	기대소요시간	분산	z (15)	확률 (%)
A-B-C	14.00	1.56	0.80	0.79
D-E	8.00	0.22	14.85	100

Source :

c 새로운 가치를 위한 시도

고객 공감

- **비즈니스 설계 모델 기법 활용**
  - ✓ Customer insights (고객관점, segment, The Empathy Map)
    새로운 혁신적인 비즈니스 모델 Ideation (자원, 제안(offer=value proposition),고객, 재무) 주도
  - ✓ Visual thinking / Prototyping / Storytelling / Scenarios
- **Design Thinking**
- **창의적 Idea**

사례 참조

easyJet

zipcar
wheels when you want them

amazon web services™

23andMe

CEMEX

비즈니스 모델이란 모방이나 벤치마킹이 아니라,
새로운 가치를 창조함으로써,
수익을 창출하는 새로운 메커니즘

Source :

3 방향 설정 및 목표/기대효과

Use case

- Use case / Piot
- 사전 준비사항 도출
- 적용을 위한 준비

적용 범위 확대

- 기존 통계 분석업무
- 프로세스 단위 분석업무 확대

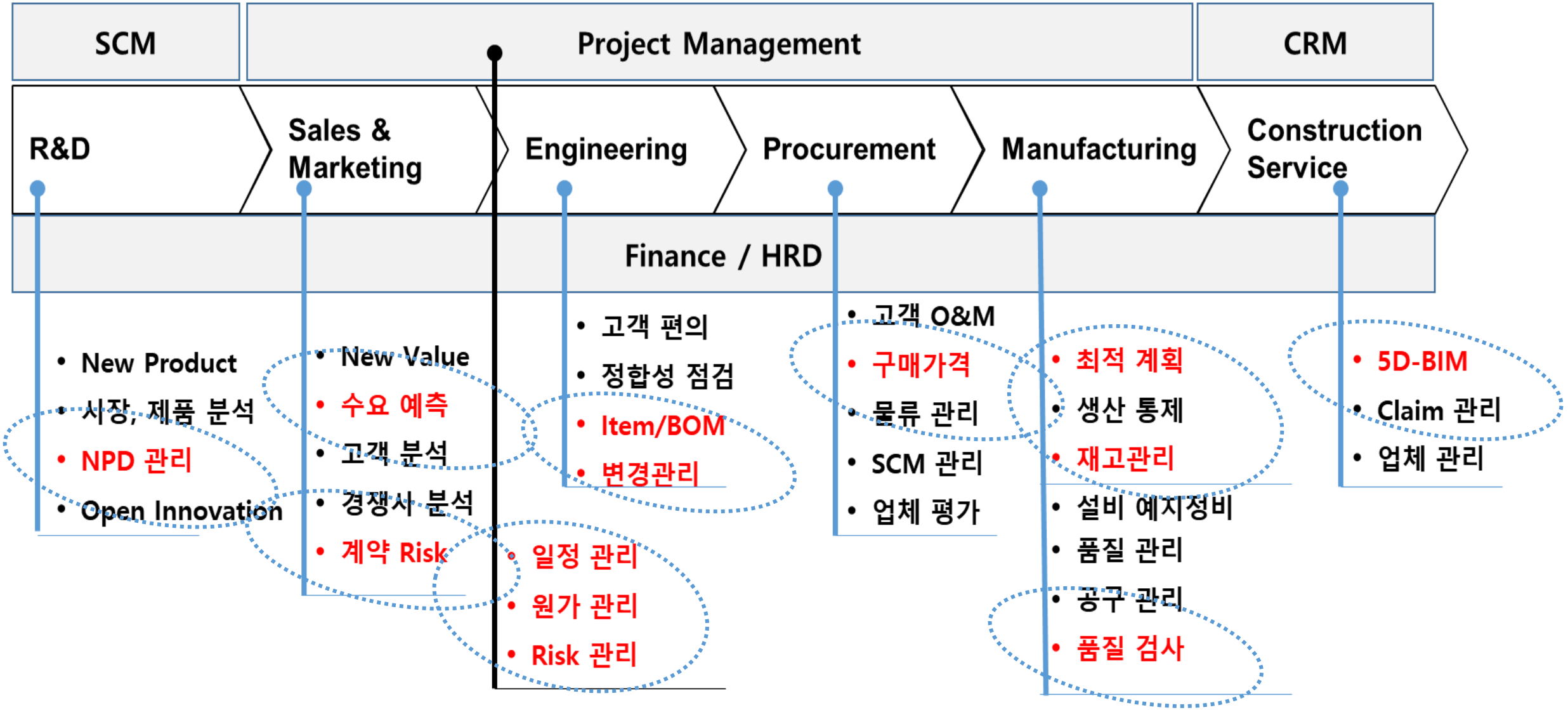
- 새로운 고객 가치
- 데이터 분석을 통한 경쟁 차별화
  - ✓ 원격 지원
  - ✓ 새로운 데이터 분석 가치

적용 대상업무
(use case Idea 대상)

기존 통계 분석업무	프로세스 단위 분석업무 확대
고객 요구사항	새로운 고객가치
핵심 업무	경쟁 차별화

Source :

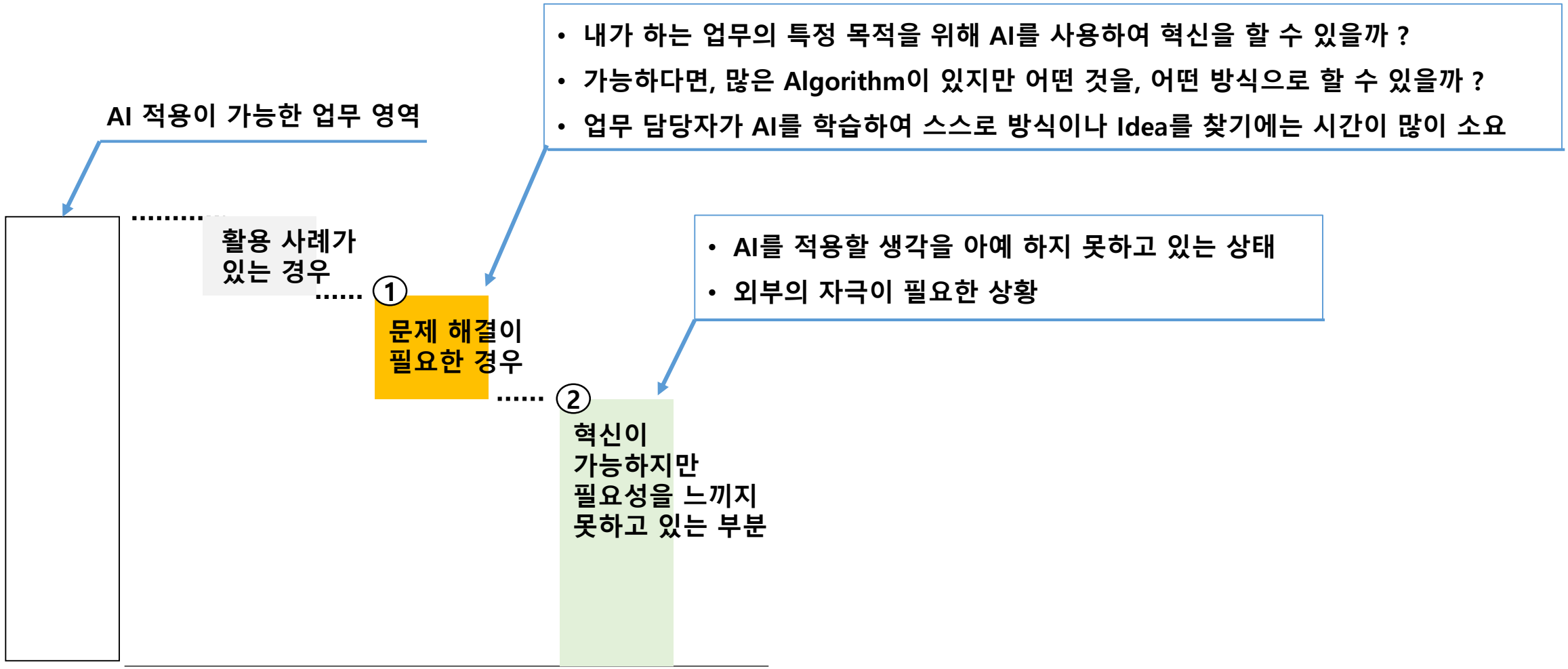
use case 개발 : AI/ML으로 새로운 가치 도출을 위한 적용대상은 상상하기 나름 ~



Source :

use case를 고민하고 고객 가치를 위한 Idea 도출에 투자하라

이용 가능한 업무 영역이 있음에도, AI 적용 여부 결정에 많은 시간이 소요되거나, 아예 적용할 생각조차 하지 못하고 있음



Source :

- use case의 시작은 자신의 업무에서 찾는 것 부터 시작하여 새로운 고객가치로 확대

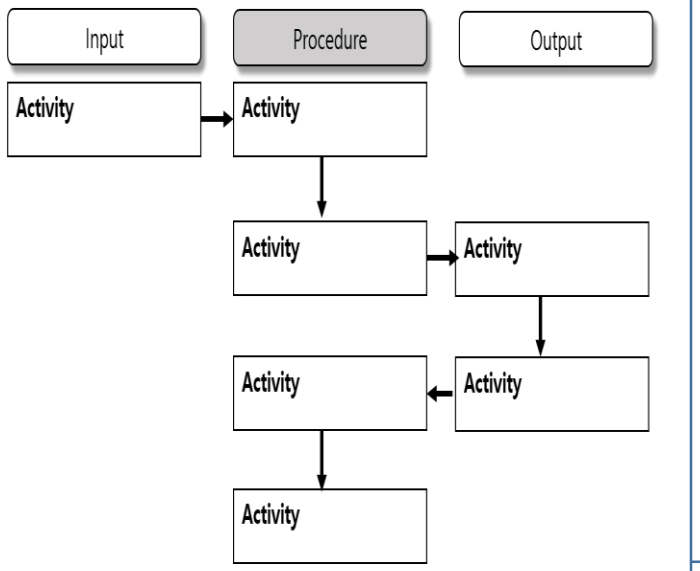
제조 현장의 실제 업무 상황의 문제해결을 위해 사용한 방법 (Open source ML Algorithm, 사용 절차)과 그로 인한 기대효과를 제시한 것

업무 상황

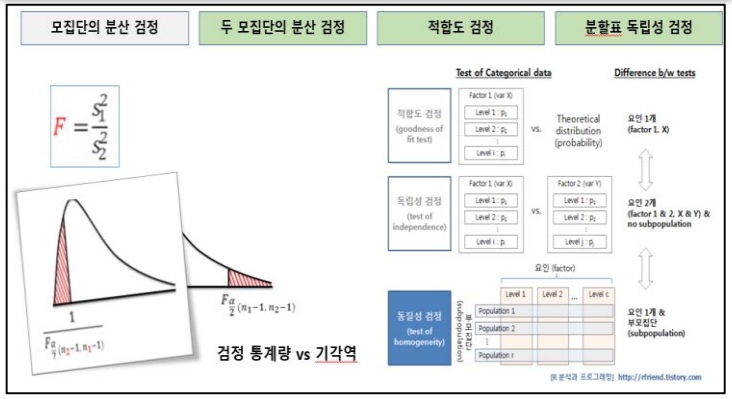
Domain : ...

Domain : ...

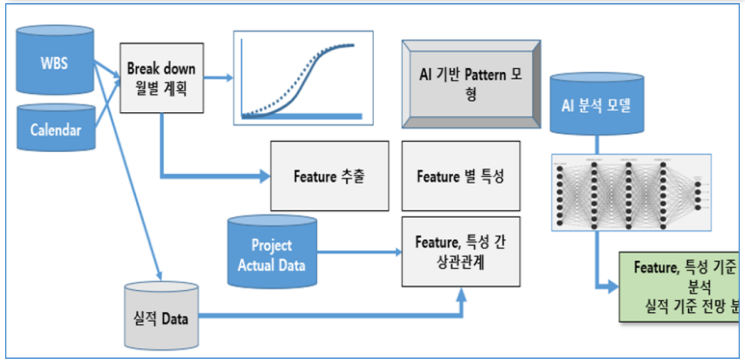
Domain : 수주 제조업 use case



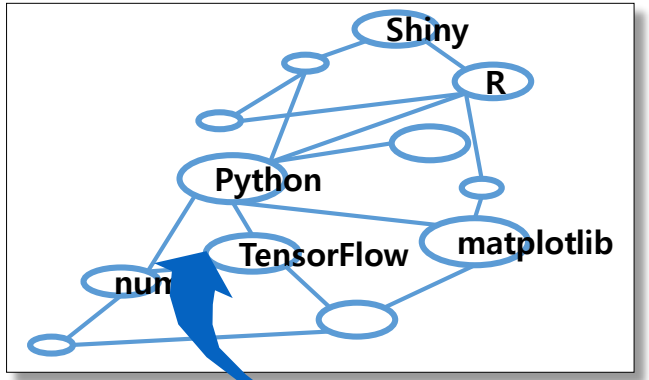
통계 분석 방안

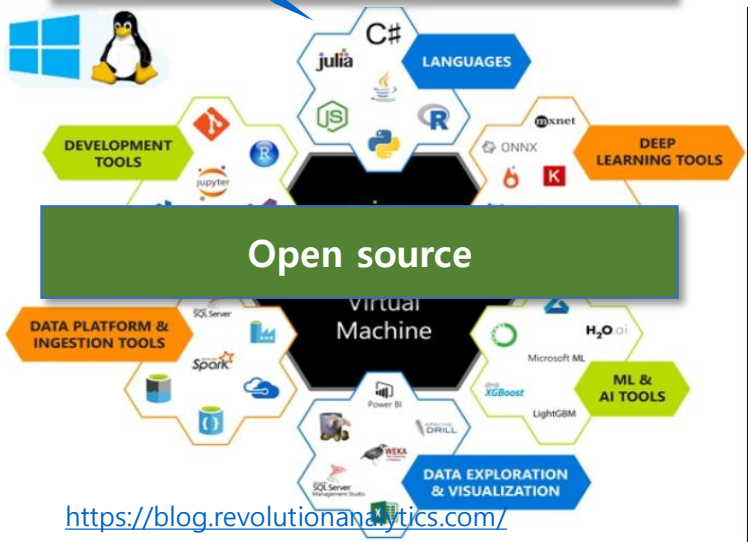


Machine Learning 분석 방안



Open source 활용





https://blog.revolutionanalytics.com/

Source : 331. DA_use_case_procedure

- 문제 해결을 위한 use case 방법 예시

다양한
강종 및 사이즈에
따른 수요 예측

설비 Layout
최적화

디지털 기술 및

데이터 분석 이해

문제 정의 및 데이터 준비

- 정확한 수요 예측이 진정 문제인가?
- 주문 후 공급 기간의 단축인가?
- 강종, 사이즈 종류
- 수요 기업과 주문 실적
- 강종, 사이즈에 영향을 주는 요소는?
- 제품 공급 프로세스는 최적인가?

문제 정의 및 데이터 준비

- 문제 정의
- 공급 제품별 공정/기계 Flow
- Flow별 작업기간
- Cost function 정의 (Lead time, 원가, 설비 가동률, TCO 기준)

입력 데이터

↓

가설, Cost Function 정의

↓

모델 선택

↓

모델 학습, 평가

↓

모델 학습, 평가

Source :

- 활용할 수 있는 알고리즘, 필요한 Device를 search 하여 활용할 수 있는 역량이 경쟁력

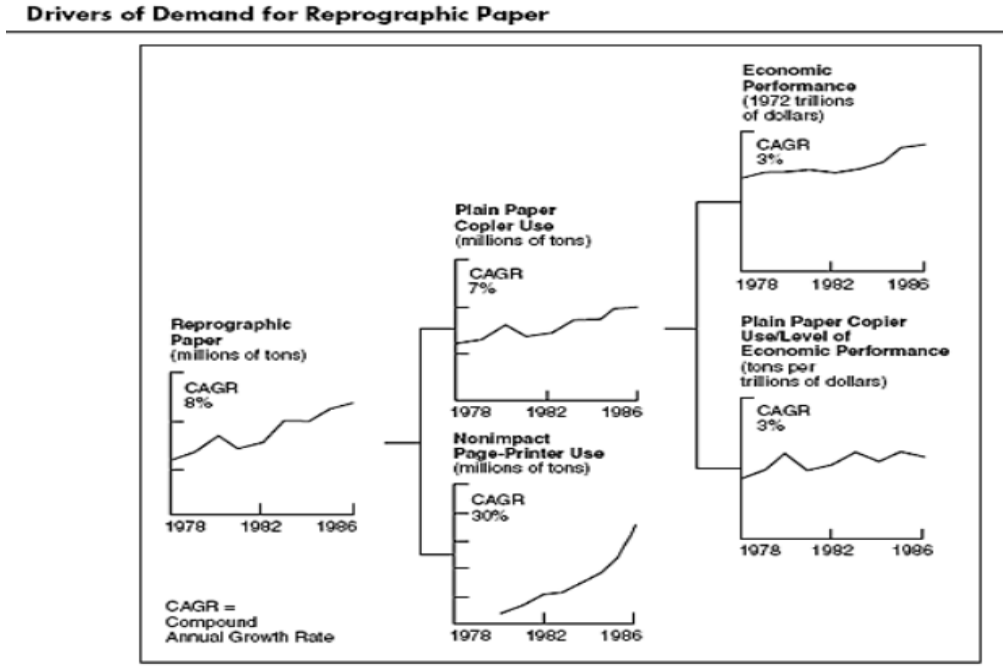
1. Define the market.
2. Divide total industry demand into its main components.
3. Forecast the drivers of demand in each segment and project how they are likely to change.
4. Conduct sensitivity analyses to understand the most critical assumptions and to gauge risks to the baseline forecast.

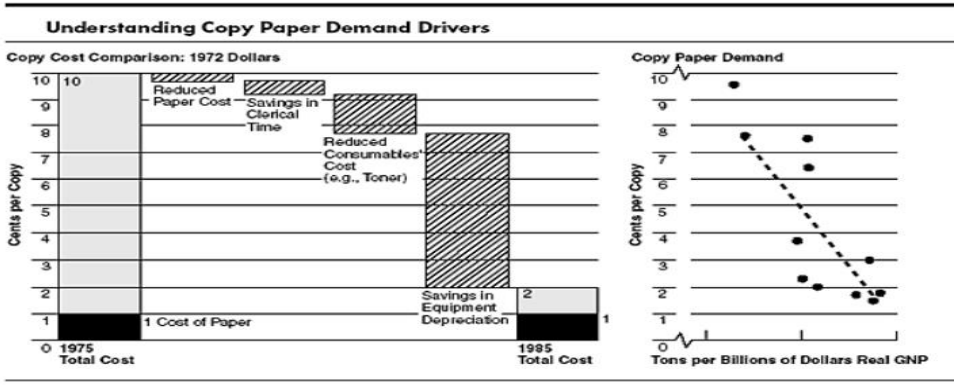
Four Steps to Forecast Total Market Demand

by William Barnett

FROM THE JULY 1988 ISSUE

Harvard Business Review





Understanding Copy Paper Demand Drivers

Source : https://hbr.org/1988/07/four-steps-to-forecast-total-market-demand

4 데이터화

데이터화

- 데이터화
- IIoT / IoT Gate way
- CLOUD
- Data Lake

데이터 활용한 가치 | Data architecture | CLOUD

Use case | Edge Data 처리 | IIoT – Data - 분석

전문가 협업

- 외부 전문가와 협업 (데이터 준비, 적용 모델 선택, 평가)

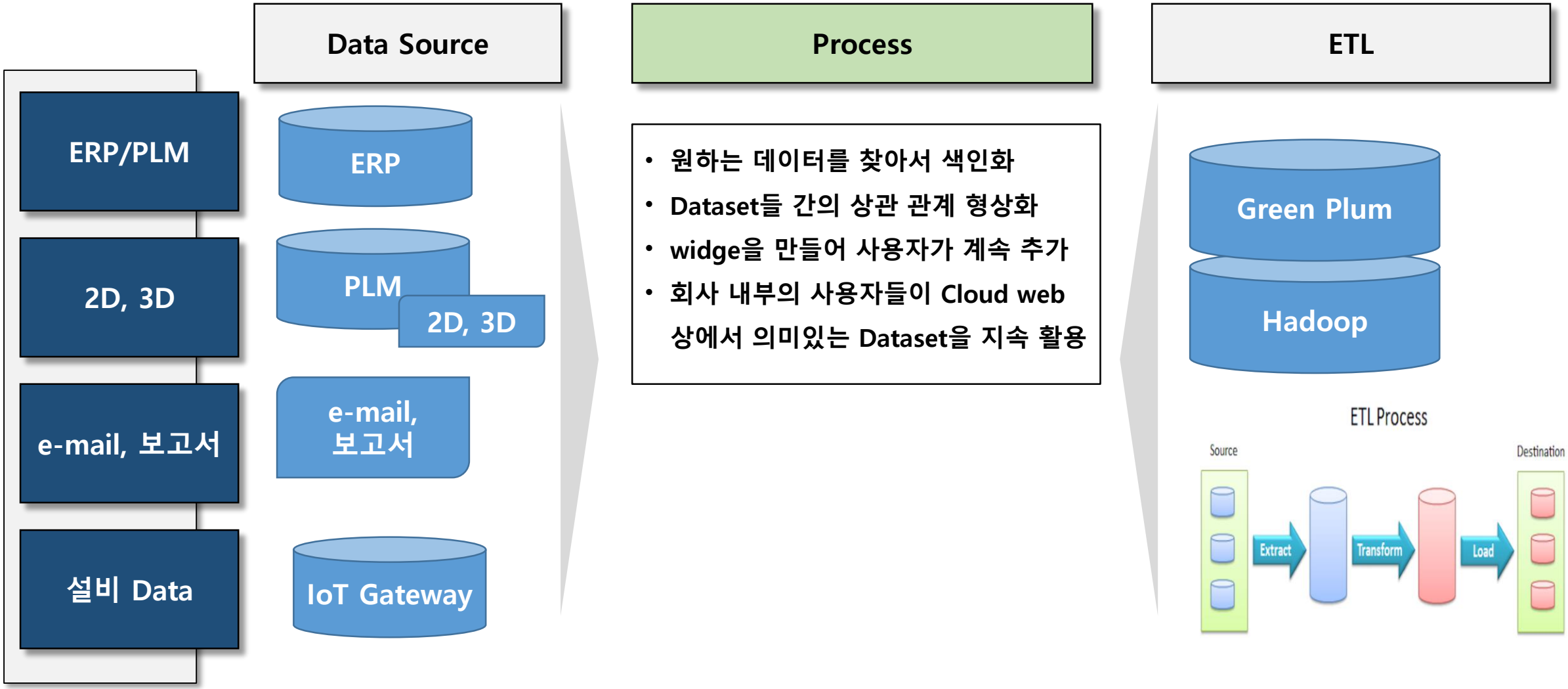
데이터 저장, 활용을 위한 Infra 구성

분석 Tool 활용

Source :

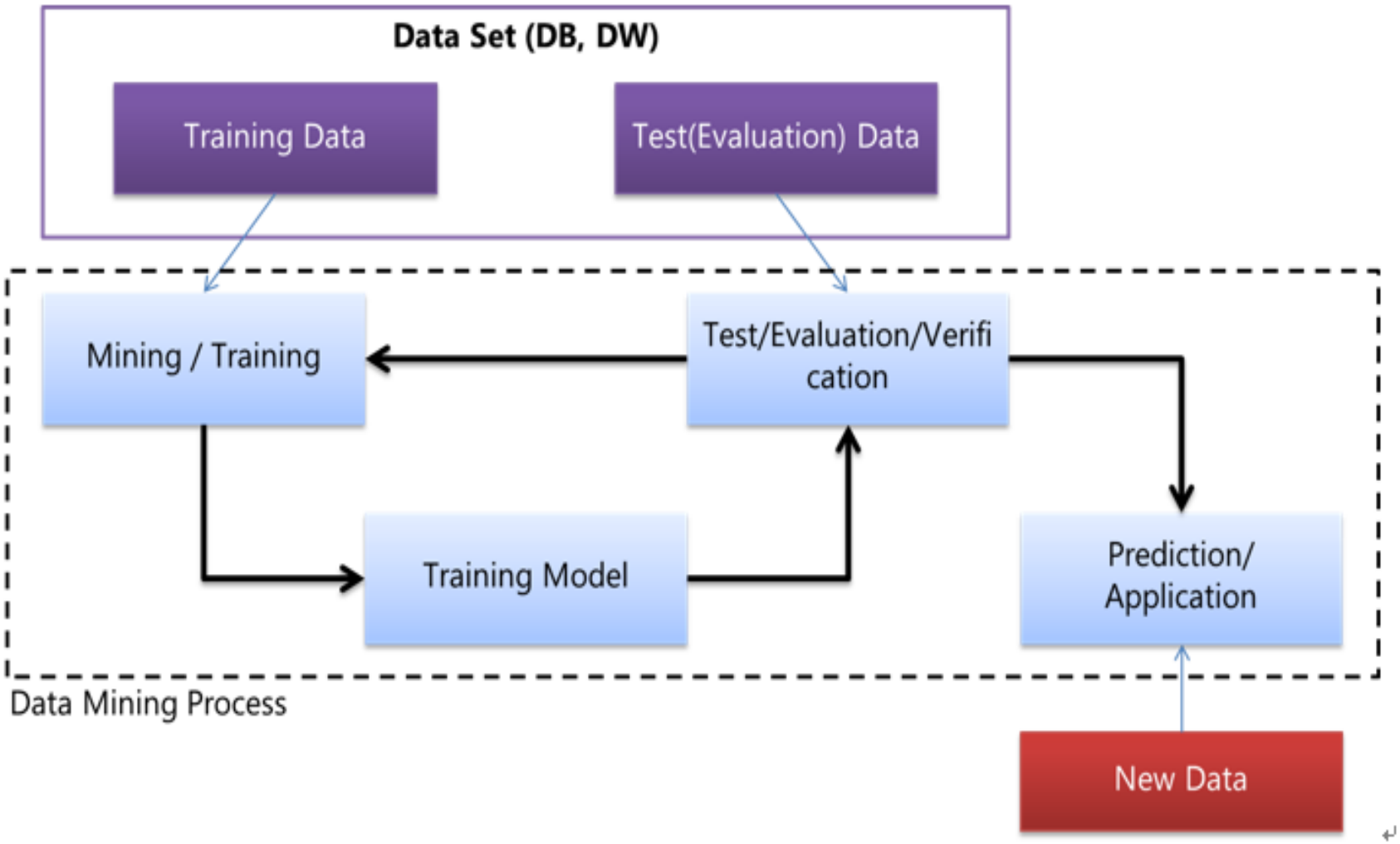
- Data Lake – Concept : automation xml , 설계 data

원하는 Category, Data source를 선택하여 자신의 작업 공간에 import하여 원하는 작업을 실행



Source :

데이터 지능화 (모델분석) Flow



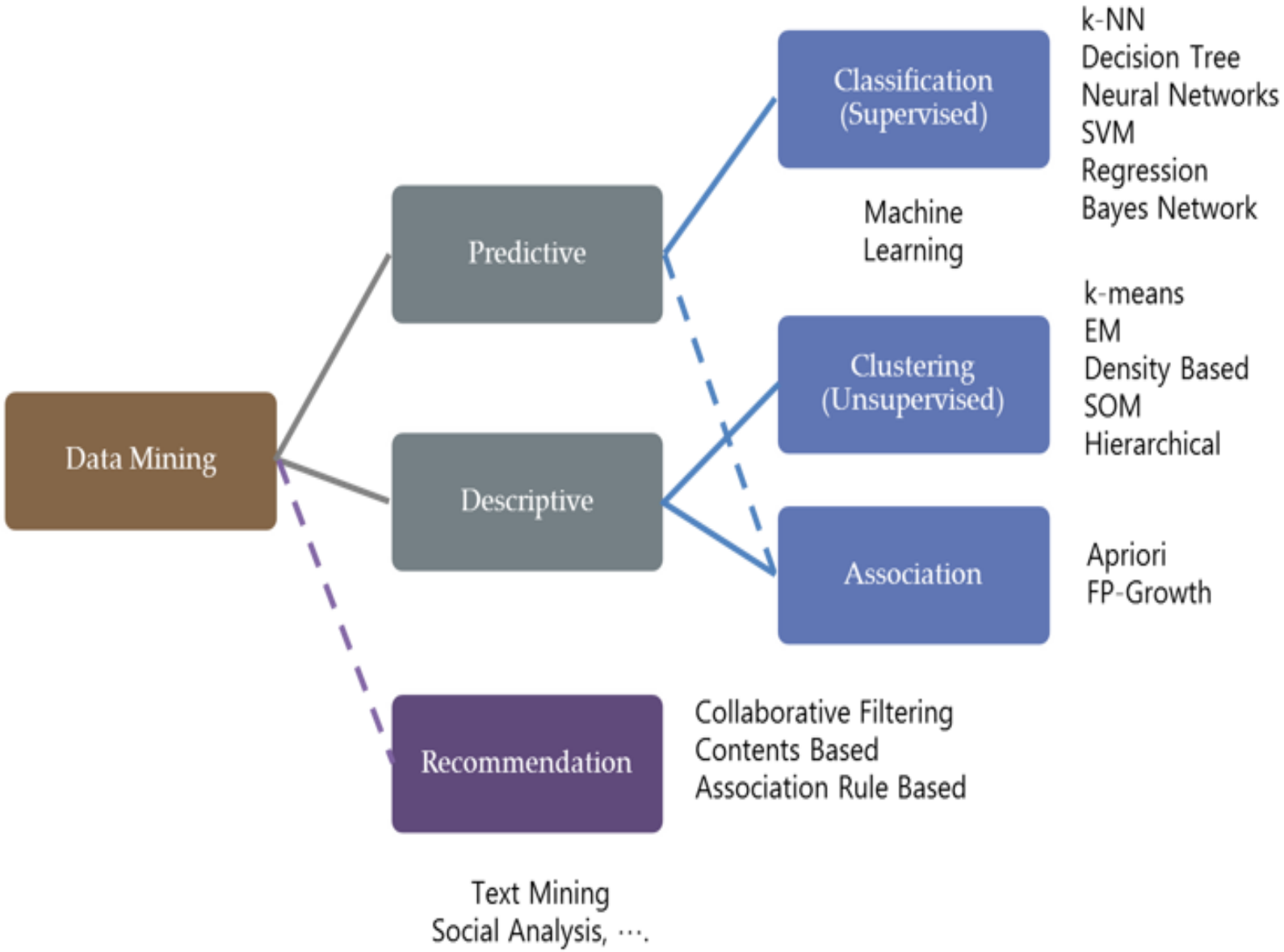
Training, Test 구분

normalization
, 균형화

초기값, learning rate size

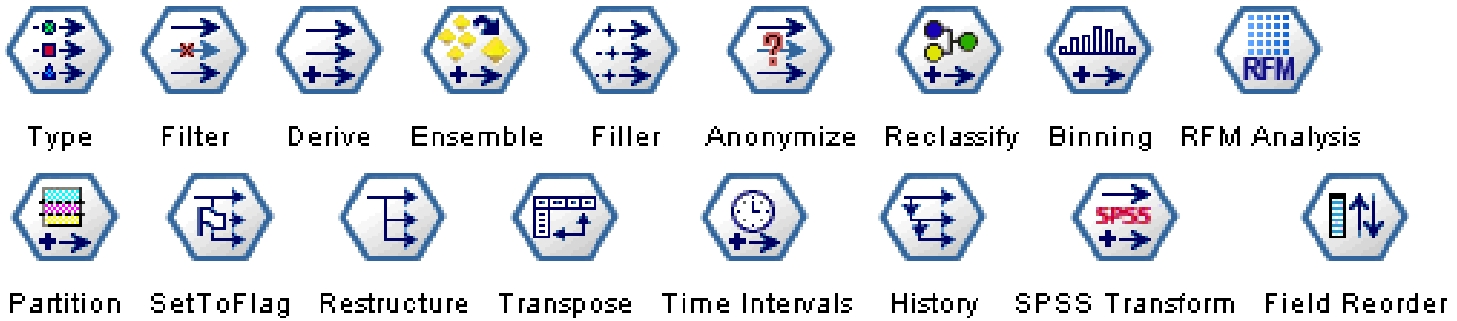
Source :

데이터 마이닝 기법



Source :

전문 Tool 사용

구분	기능	노드
Sources	데이터 연결 노드	
Record Ops	레코드 단위 데이터 변환작업 노드	
Field Ops	필드 단위 데이터 변환작업 노드	
Graphs	데이터 도식화 노드	

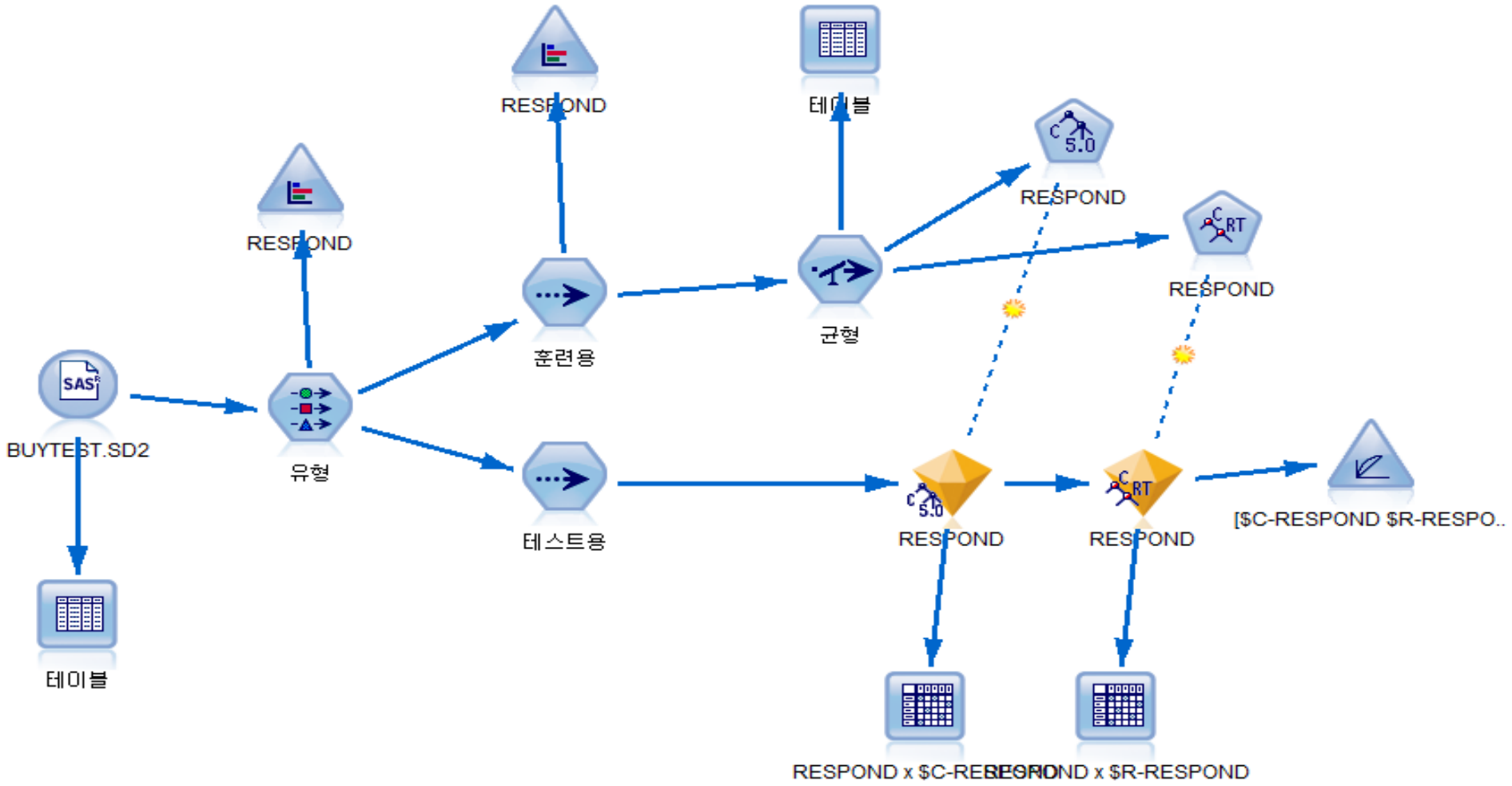
Source :

전문 Tool 사용

구분	기능	노드	
Modeling	모델링 노드	Automated	 Binary Classifier Numeric Predictor Time Series
		Classification	 C&R Tree QUEST CHAID Decision List Regression PCA/Factor Neural Net C5.0 Feature Selection Discriminant Logistic GenLin Cox SVM Bayes Net SLRM
		Association	 Apriori 카르마 순차규칙
		Segmentation	 K-Means Kohonen TwoStep Anomaly
Output	결과 출력 노드		 Table Custom Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals SPSS Output
Export	데이터 내보내기 노드		 Database Flat File SPSS Export Dimensions SAS Export Excel Publisher

Source :

의사결정나무 추론을 위한 클레멘타인 흐름도



Source :

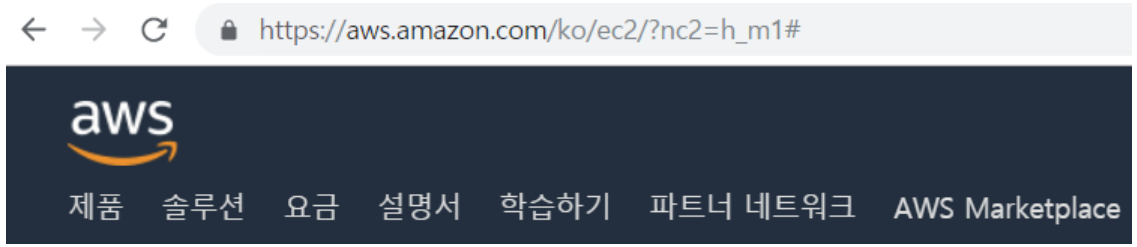
- CLOUD

CLOUD 장점

- 탄력성
- 확장성
- 민첩성

CLOUD 서비스

- Google (Google Cloud Platform), MS Azure , amazon (AWS), IBM Cloud
- 기업이 선택하여 사용하기 용이하지 않음
- 클라우드 서비스 브로커리지(CSB) 활용



접근 방법

- 활용 관점
- 변경 가능성 염두 (복수 사용)
- 기업의 Business Model, Platform 활용 전략에 대한 방향이 결정되고,
- 이를 기반 OE 측면에서 프로세스 혁신에 따라 표준과 기업 고유 업무를 구분
- CLOUD에서 처리할 부분과 아닌 것으로 구분 / 단, 기업의 독특하고 고유한 부분이 정말 필요한 것인지 판단
- Hybrid 고려

Source :

- Big Data, Big Questions. Does Edge Computing Have the Answers?

- ***How much data is too much? Edge Computing states a case for smaller data analysis***

- **Too much poor data**
- **Improved data collection**
- **Filtering data on the edge**
- **Real-time data processing**
- **Edge is the future of big data**

Not all data generated by industrial IoT sensors is useful, so why not filter it at the edge?

Source : VIA, https://www.viatech.com/en/2018/10/big-data-big-questions-does-edge-computing-have-the-answers/

5 융합

가치 구분

- 새로운 고객 가치
- Operation Excellence
- Product Innovation

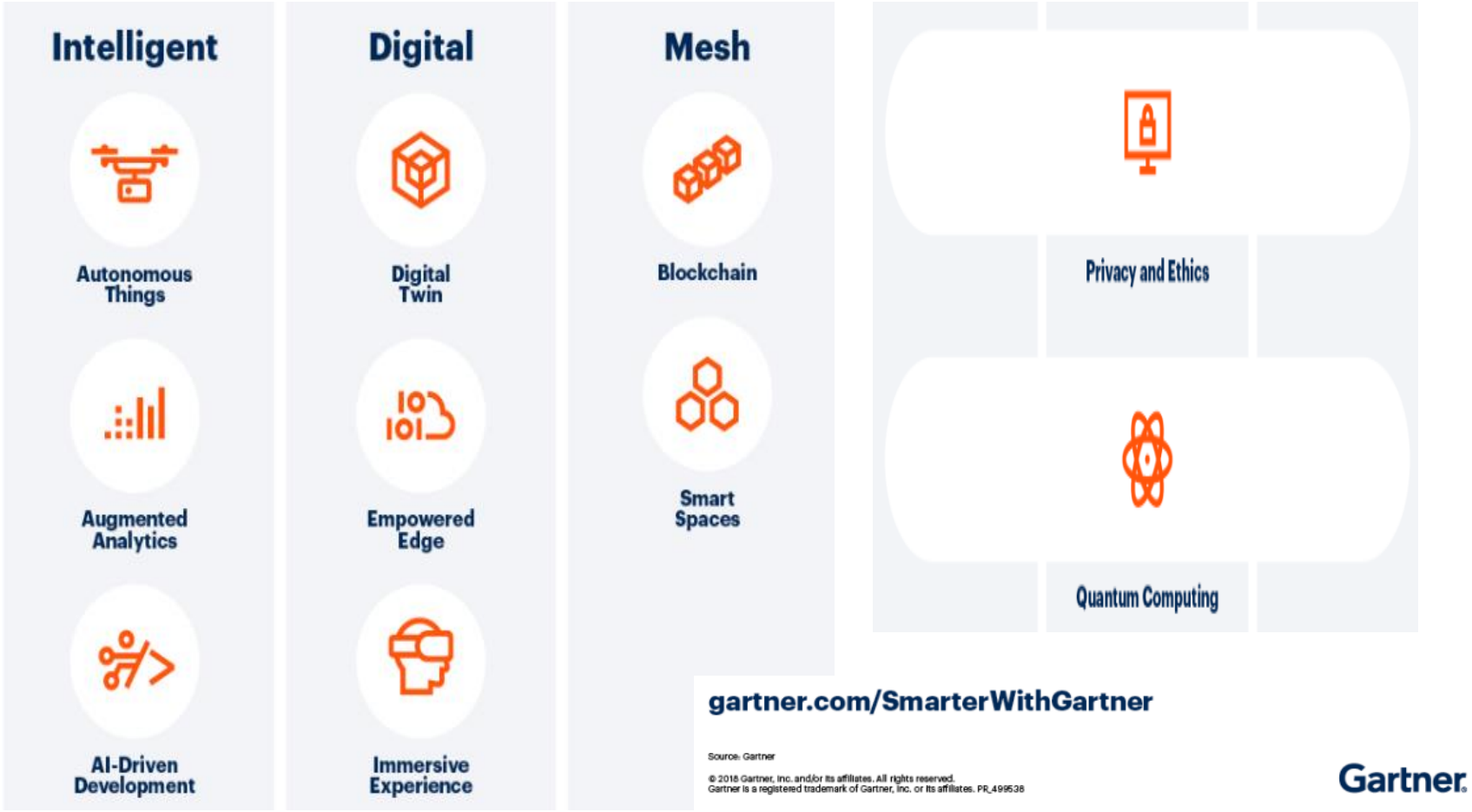
융합

- Data 자체
- 데이터 분석 및 융합을 통한 가치
- 연결 (H/W, S/W, 가치)
- 적절한 Hardware

Source :

디지털 시대 핵심은 데이터와 기술의 융합에 의한 Intelligence이다.

Gartner Top 10 Strategic Technology Trends for 2019



데이터

기술의 융합 & Intelligence

Source : October 15, 2018Contributor: Kasey Panetta , https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019/

6 가치 실현

고객 가치

- 고객 가치 (데이터 분석 + H/W)
- Telemetric M / SCP
- AR / VR / MR

Operation Excellence

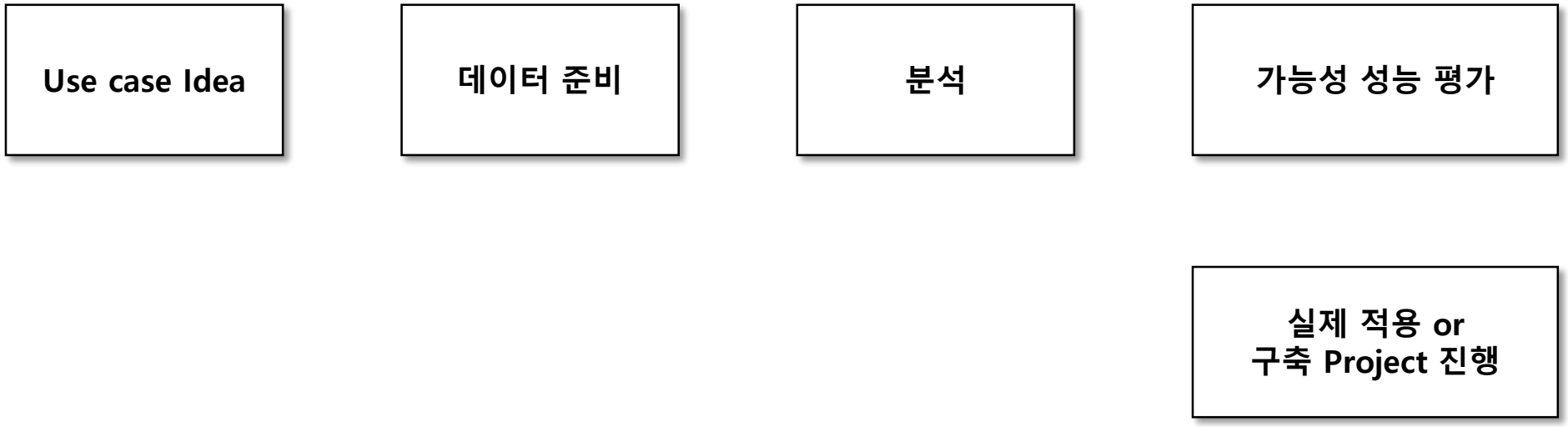
- 최적 공정계획 / 수요 예측 : CPPS
- 공정 진척에 따른 기계, 설비 자동 제어 (Shop control, Edge computing)
- Robot / Machine
- Connected IIoT

Product Innovation

- 제품의 디지털화
- 제품과 디지털 기술의 융합

Source :

Use case Idea 기반 데이터 분석 Pilot (예시)



Source : use case Idea 기반 사례

Wrap-up

Source :

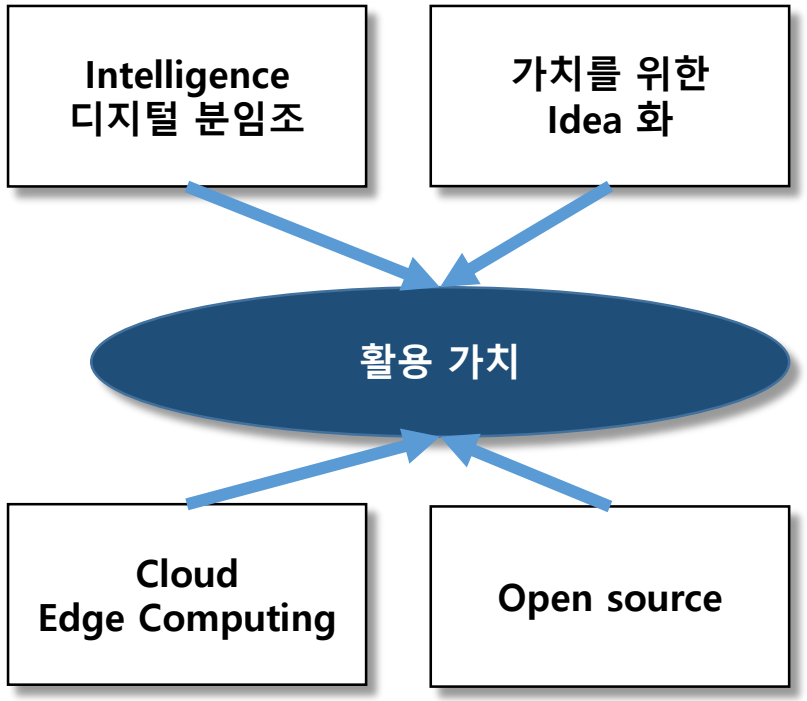
디지털 시대 미래에 어떻게 대응할 것인가?

고객가치를 위한 Idea

- 새로운 변화의 이해를 위해 Open source & API Economy
- 지속적인 혁신을 위한 방안 필요 (고객가치 아이디어, 신속한 시도 & 축적을 통한 혁신)
- 여러 조직이 함께 협업하는 방식 고민
- 디지털 분임조

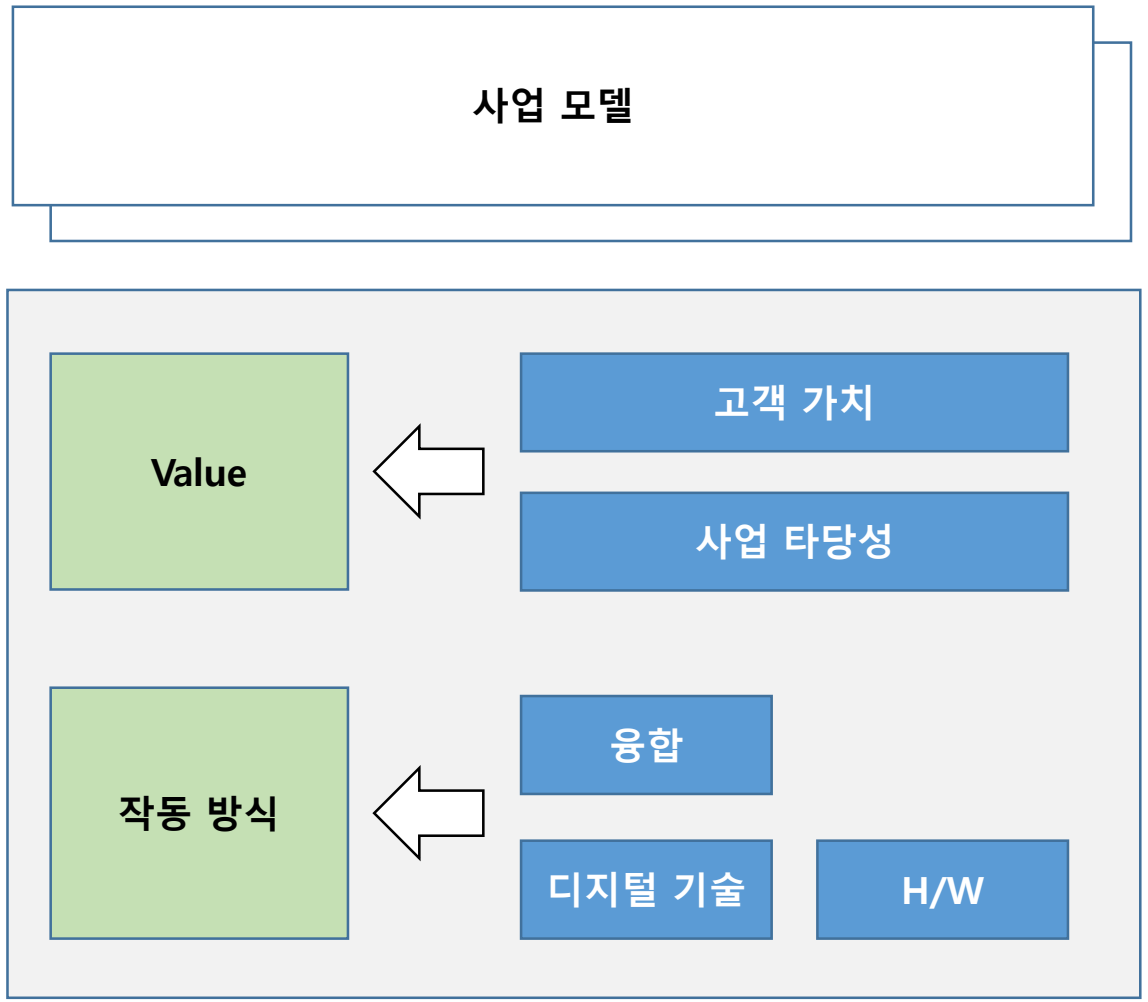
Connectivity 가치화 융합

- 가치와 융합 측면에서 Cloud, Edge computing 이해
- 디지털 기술, H/W, 데이터의 융합으로 새로운 가치 구현
- 신속한 검증으로 기대효과 여부 판단
- 신속한 적용 및 지속적인 고도화



Source :

향후 10년, 어떻게 변화에 대응할 것인가 ?



미래 사회에 대한 Insight

향후 5~10년 후 기회를 위하여 지금 무엇을 하여야 할까 ?

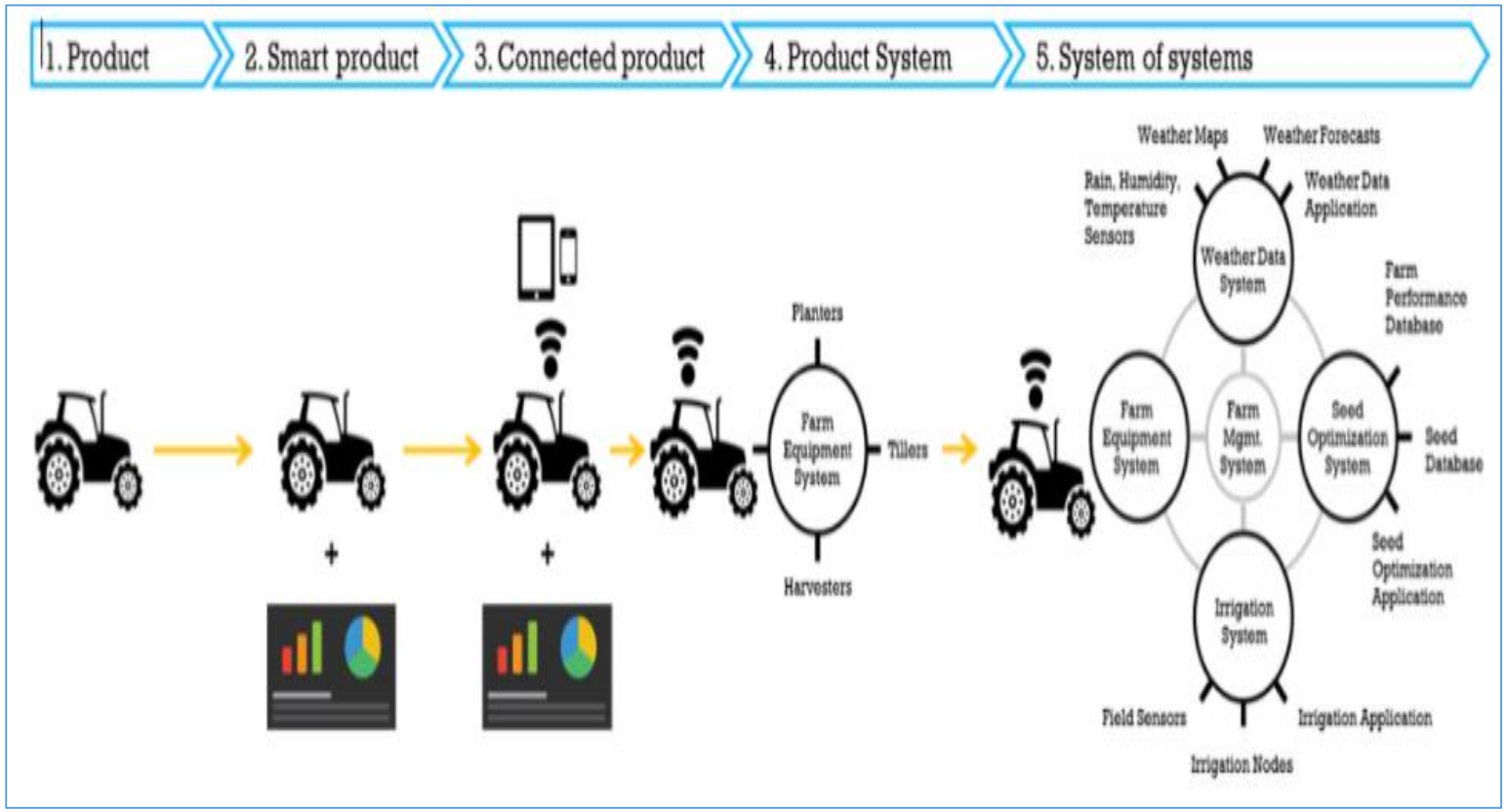
- 방향에 대한 인지
- 변화에 대응할 수 있는 역량
- 유연함으로 가치에 대한 시도

Source :

고객가치의 실현모델을 위한 제품/서비스, CLOUD 기반 SCP와 API 활용 역량 준비

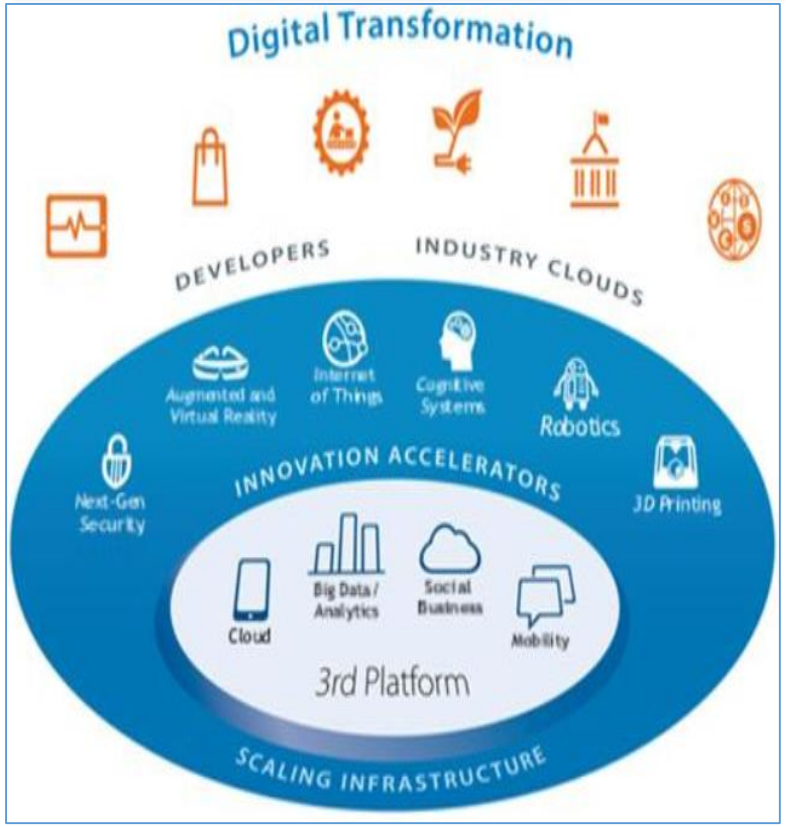
기업이 필요로 하는 신속한 제품 애플리케이션 개발 및 운영을 지원할 수 있는 체계와 제품 내부 및 외부에서 생성되는 엄청난 양의 데이터를 수집, 분석 및 활용을 위한 IT Infra 체계에 대한 준비

SCP로 인한 산업 경쟁구도의 변화 방식 (2014 년)



* SCP (Smart Connected Product)
출처 : "HOW SMART, CONNECTED PRODUCTS ARE TRANSFORMING COMPETITION," HBR, NOVEMBER 2014

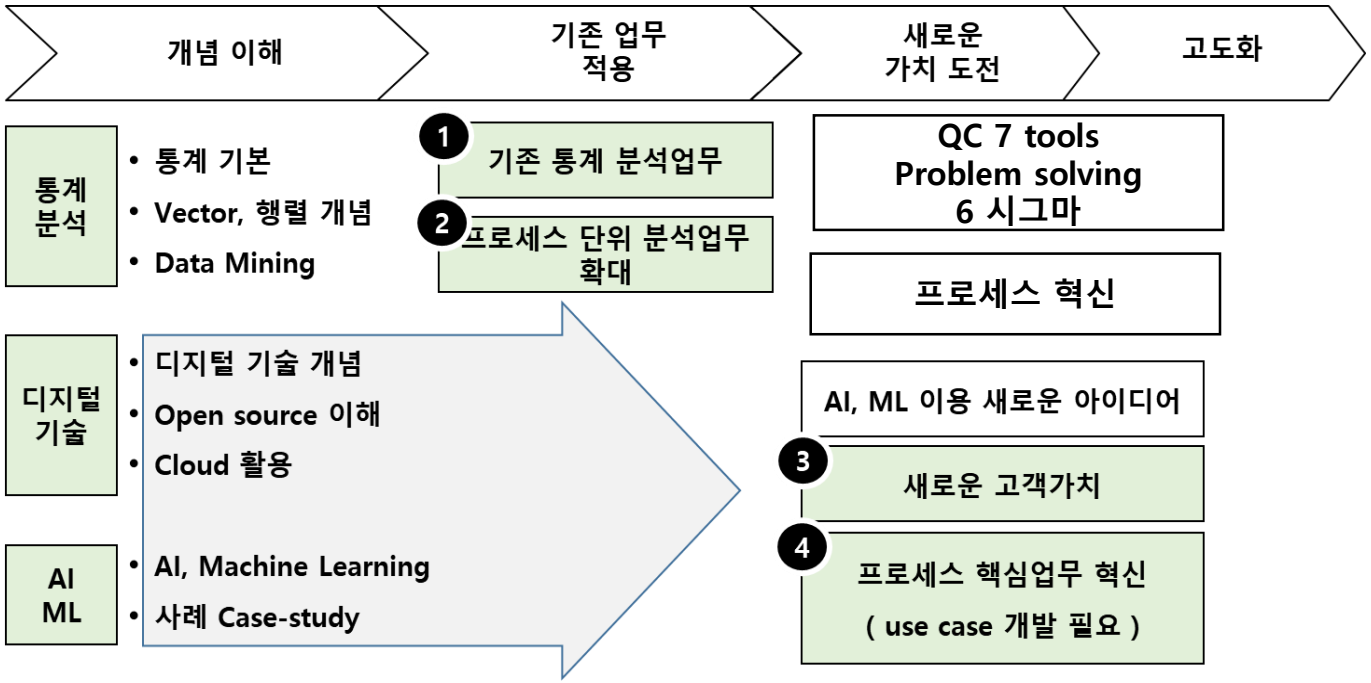
역량과 시스템의 준비
DX Group, Scale up

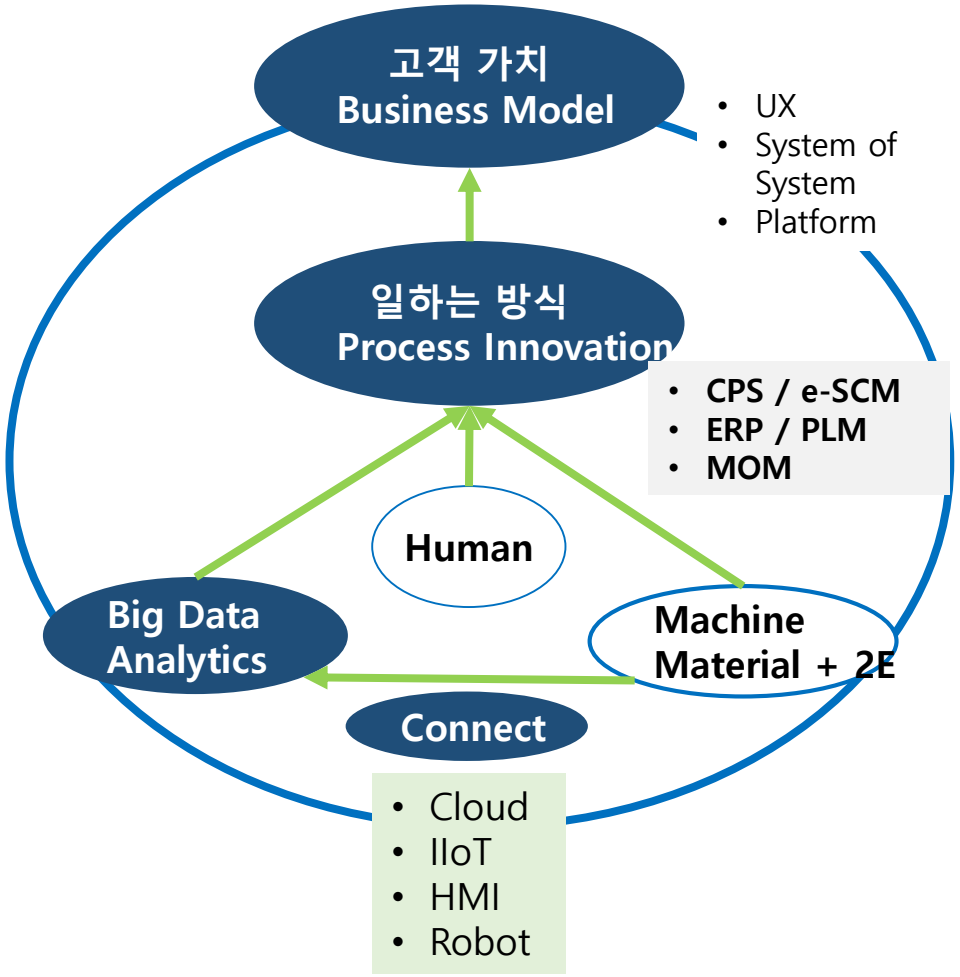


출처: Digital Transformation and 3rd Platform Leadership: What You Need to Know, IDC, https://youtu.be/LW1AmvsLa5c

빅데이터는 고객가치와 기업 가치 증대를 위한 활용 측면에서 바라보아야 한다.

- 빅데이터 개념과 흐름을 이해하고자 하는 것은 그 자체가 아니라
- 활용하여 고객가치의 실현이며, 그 것을 위한 방법을 알고자 하는 것이다.
- 개념에 대한 이해가 되었다면 데이터 분석 Flow를 따라 가는 것이 아니라 가치를 위한 고민이 가장 우선이다.
- 고객가치 /융합 /BDA /Idea /이해 위한 적용 절차

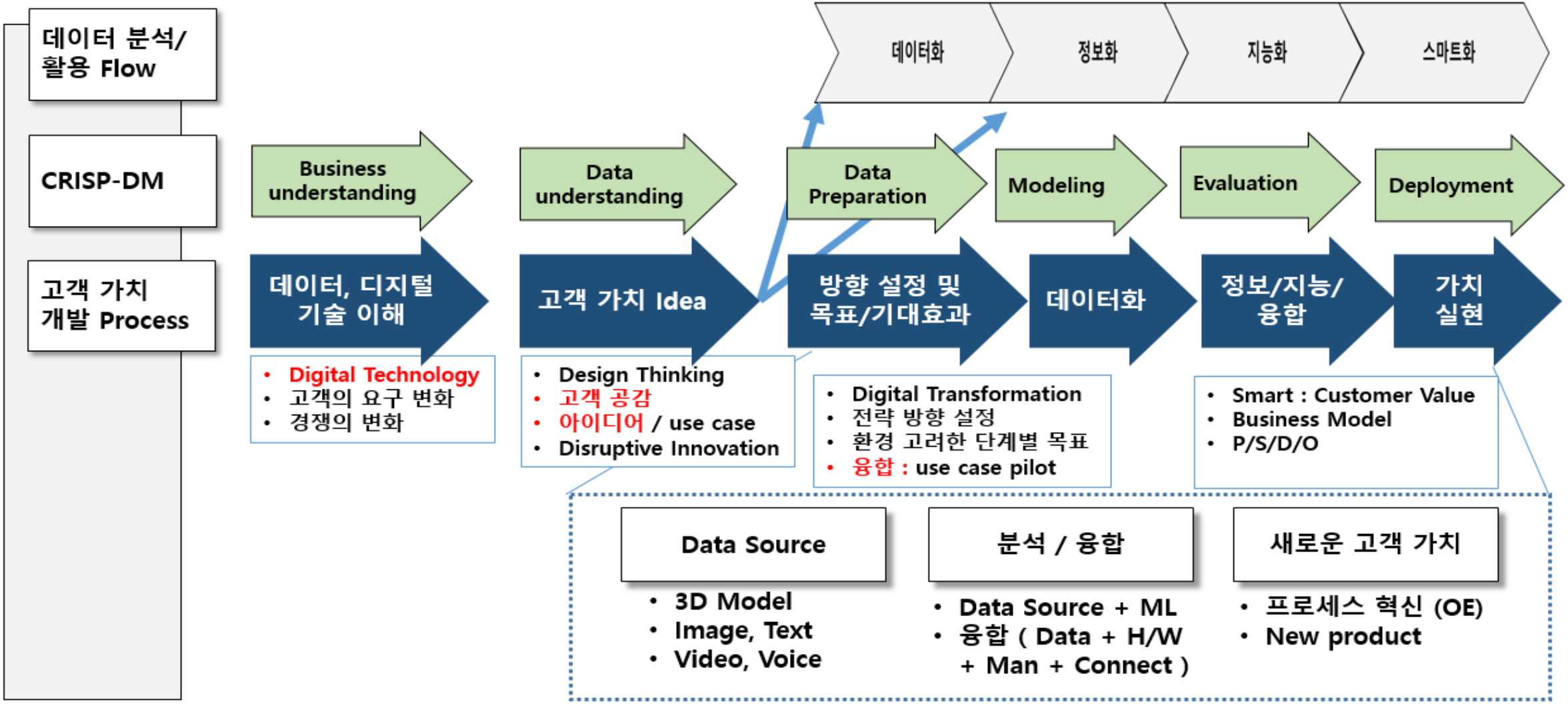




Source :

데이터 분석(고객가치) Idea 개발 Process (이해 → Idea → 융합 → BDA 활용)

새로운 가치혁신을 위해 AI Startup 기술 참조하고 hardware를 융합하는 창의력 발휘

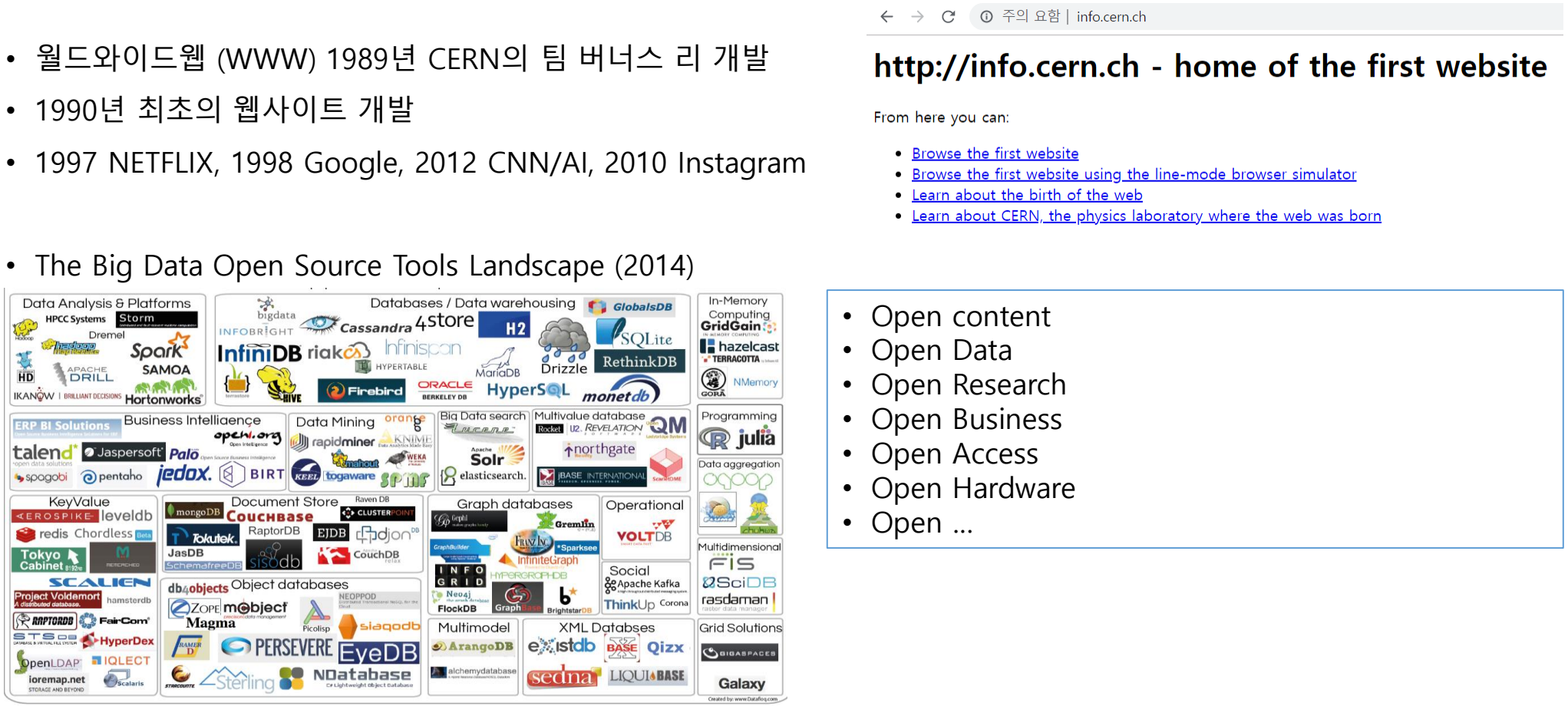


Source :

Open source의 가치 발견과 활용 (완전한 Free는 아님)

web의 시작

- 월드와이드웹 (WWW) 1989년 CERN의 팀 버너스 리 개발
- 1990년 최초의 웹사이트 개발
- 1997 NETFLIX, 1998 Google, 2012 CNN/AI, 2010 Instagram



Open source

- The Big Data Open Source Tools Landscape (2014)

- **github**

- Open content
- Open Data
- Open Research
- Open Business
- Open Access
- Open Hardware
- Open …

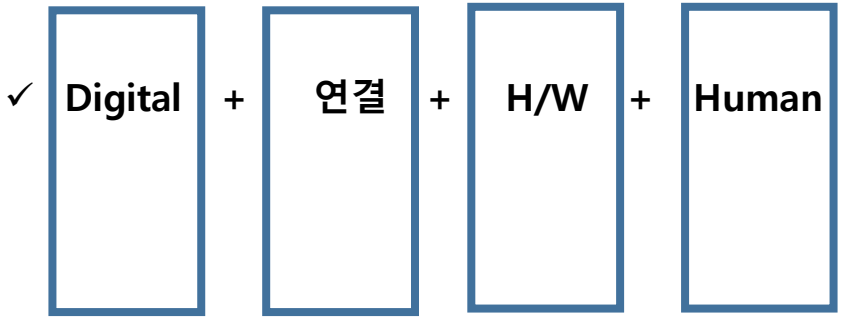
Source : https://datafloq.com/big-data-open-source-tools/os-home/

그래서, 기업에 어떻게 활용할 것인가 ?

- 빅데이터(BDA) 가치에 집중하되, 고객에 대한 공감부터 시작하여 관련 아이디어 도출...

- (기업) 스스로 시도할 수 있는 준비 (디지털 분임조)
  - ✓ 즉시 할 수 있는 것부터 시작 : QC 7 Tools, Problem solving, 6 σ
  - ✓ 프로세스 혁신

- (개인) 가치의 실현을 목적으로 Open source의 활용을 고려해 보자.



Source :

End of Documents