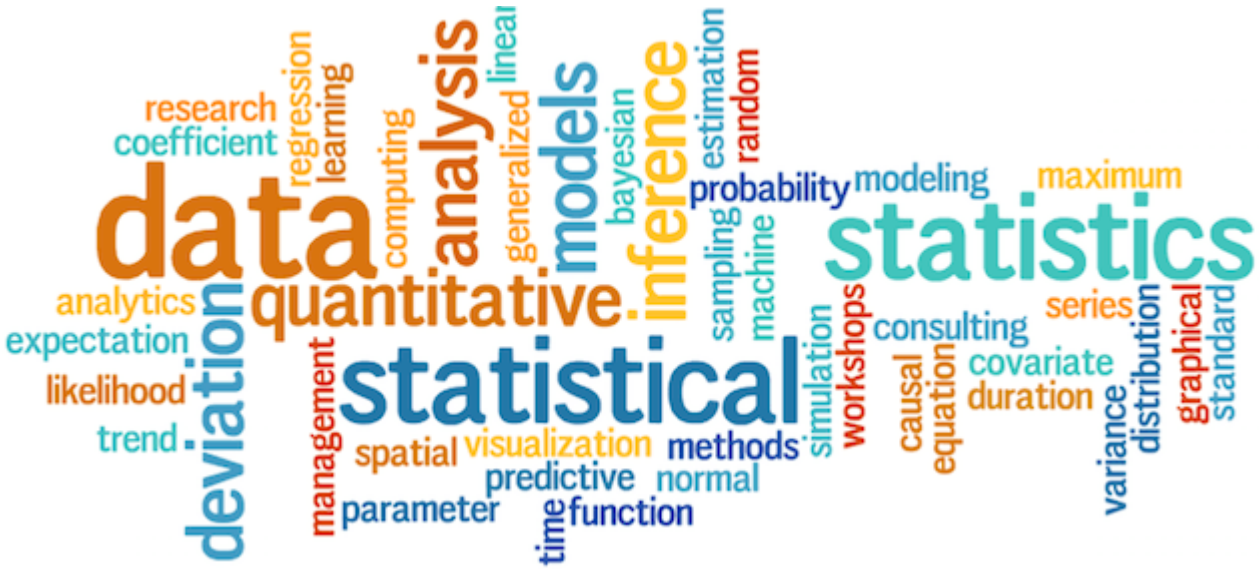



# 통계기반 데이터 분석(Ch 1)

데이터사이언스 & A.I 전공 정화민 교수

# 교수소개

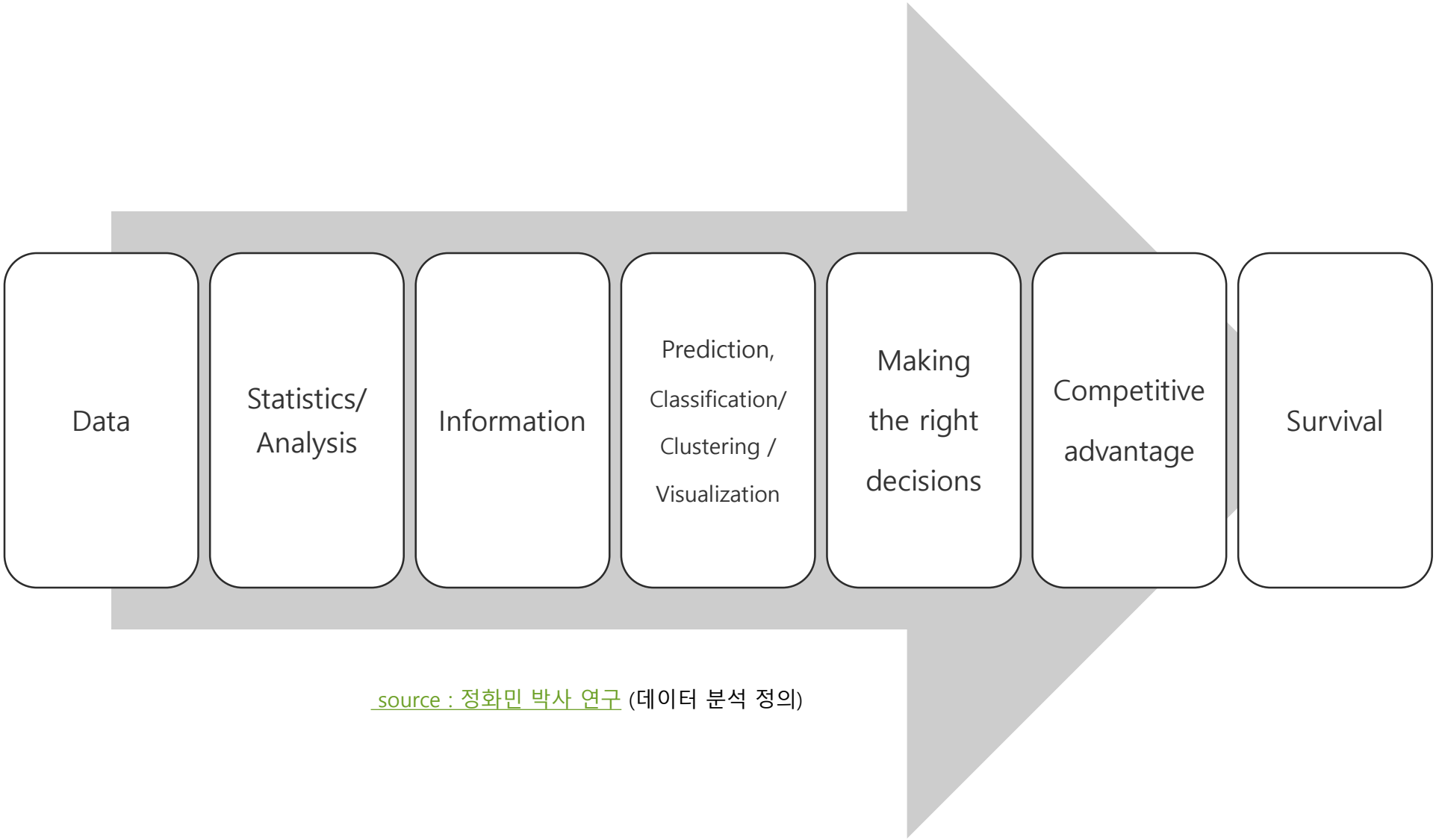
## Jeong Hwa Min(Ph.D)

CEO of TauData. Co.,Ltd.

- Adjunct professor of Sogang University (Graduate school)
- Major : Data Science & AI

- More than 19 years of Data Analysis and AI algorithm expert
- Statistical Analysis expert

# 데이터 분석은 4차산업시대 생존을 위한 필수 기술




source : 정화민 박사 연구 (데이터 분석 정의)

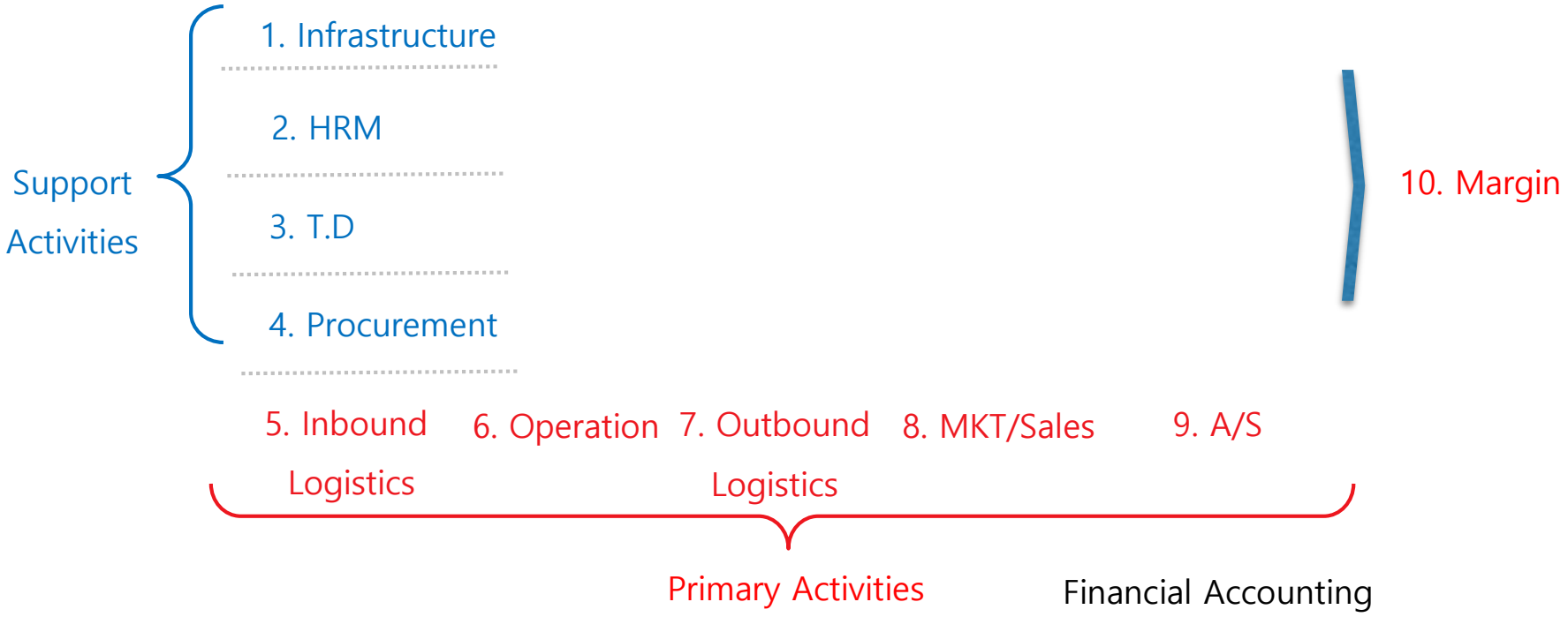
# 우리의 측우기는 과학적 데이터수집 기기

- 고대로부터 강우량 측정, 1442년 전국 350군데 측우관측소 운영.
- 우리의 측우기는 서양의 측우기 보다 200년 앞서 개발되었음.
- 우리민족의 DNA 속에는 데이터 수집, 분석적 기질이 있다.

8만 대장경

조선왕조실록

# 기업경영에서의 데이터 분석 (예: Value chain)

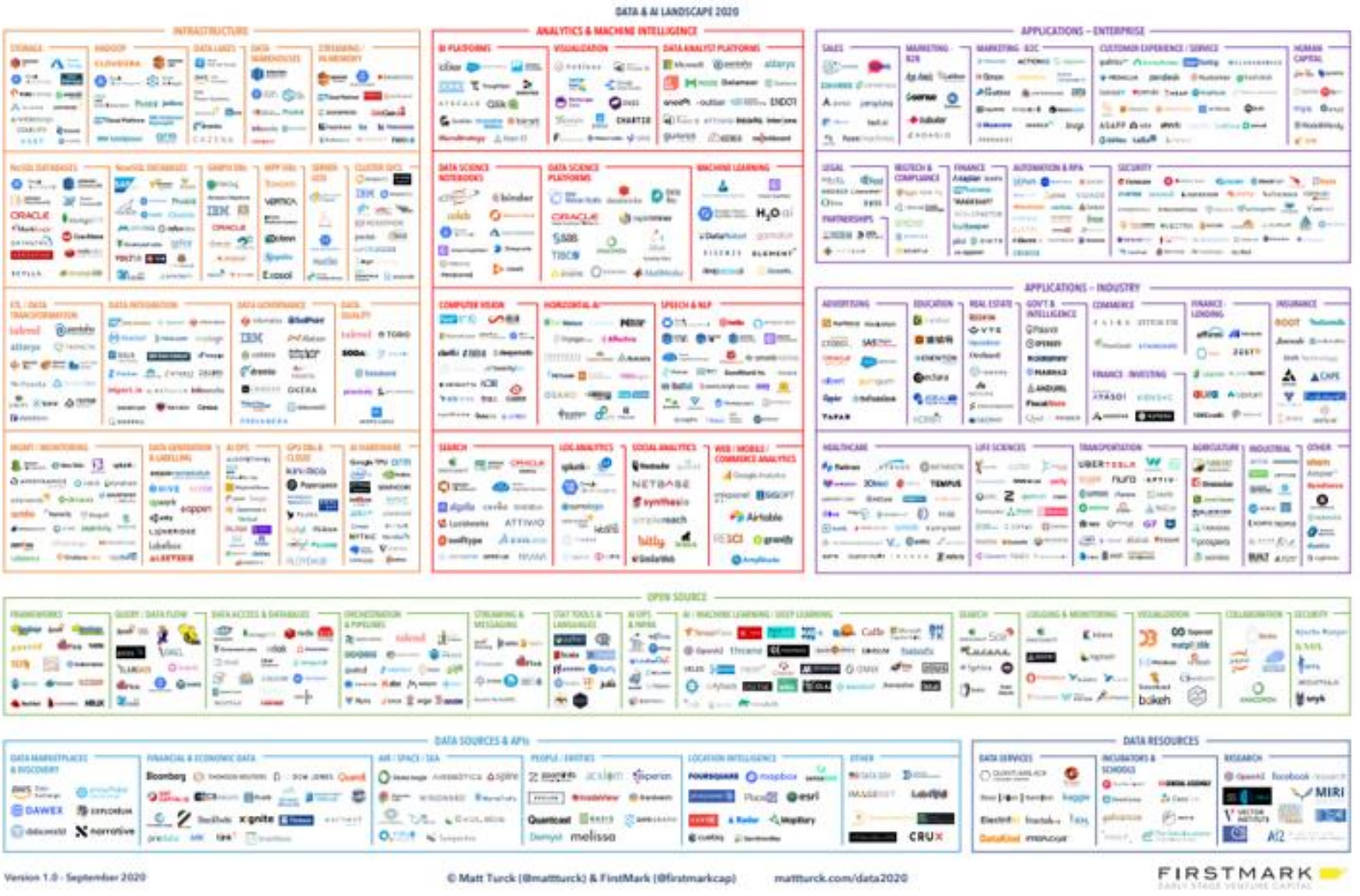



**ERP?**

**CRM?**

# 빅데이터 Landscape

## 빅데이터 Landscape




Source: 2020_Matt_Turck_Big_Data_Landscape

# 50 Best Jobs in America for 2022

## 2017년 미국 최고 최악의 직업


Source :중앙일보 2017.05.01 기사 미국최고 직업은 '통계 학자'

## 2022년 미국 최고 직업

glassdoor

Search for job titles, companies, or keywords | Soul

Jobs | Companies | Salaries | Careers | For Employers | Post Jobs

### 50 Best Jobs in America for 2022

Best Jobs | 2022 | United States | Share

| | Job Title | Median Base Salary | Job Satisfaction | Job Openings | |
|---|---|---|---|---|---|
| #1 | Enterprise Architect | $144,997 | 4.1/5 | 14,021 | View Jobs |
| #2 | Full Stack Engineer | $101,794 | 4.3/5 | 11,252 | View Jobs |
| #3 | Data Scientist | $120,000 | 4.1/5 | 10,071 | View Jobs |
| #4 | Devops Engineer | $120,095 | 4.2/5 | 8,548 | View Jobs |
| #5 | Strategy Manager | $140,000 | 4.2/5 | 6,977 | View Jobs |
| #6 | Machine Learning Engineer | $130,489 | 4.3/5 | 6,801 | View Jobs |
| #7 | Data Engineer | $113,960 | 4.0/5 | 11,821 | View Jobs |
| #8 | Software Engineer | $116,638 | 3.9/5 | 64,155 | View Jobs |
| #9 | Java Developer | $107,099 | 4.1/5 | 10,201 | View Jobs |
| #10 | Product Manager | $125,317 | 4.0/5 | 17,725 | View Jobs |

Source: glassdoor.com

# 빅데이터 전문분야

| | |
|---|---|
| 기술 | 하둡분산처리, 배치처리(Map reduce), 실시간 분산기술, 머신러닝 등 |
| 분석 | 빅데이터 분석 방법론, 분석기법 (회귀분석, 분산분석, 요인분석, 로지스틱 회귀분석, 상관분석, 시계열, 인공신경망, 데이터 시각화, 딥러닝, 의사결정나무, 연관성, 군집분석, 시각화, SNA 등) |
| 기획 | 빅데이터 기획과정, 빅데이터 분석방법론, 빅데이터 과제발굴 및 사업관리, 빅데이터 기획요소발굴 |
| 제조 | 생산자동화, 품질자동화, 자동화와 빅데이터, 제조 현장에서의 빅데이터 검색, 제조 빅데이터 분석, 빅데이터 시각화, 기초 데이터 분석 |
| 의료 | 확률분포, 생존분석, 위험함수 와 생존함수, COX 비례위험모형, 데이터 마이닝, 비모수통계, 상관분석, 회귀분석, 분산분석 등 |
| 금융 | 신용평가모형, 빅데이터 분석, 재무데이터수집 방법, 재무비율 및 재무지표 분석, 시계열 분석, 기업매출 예측, 잔여이익모형, 주식가치평가, 자산포트폴리오 최적화 모델 |
| 유통 | 매출분석, 상권분석, 매출 예측 및 의사결정, 마케팅 효과분석, 수요예측, 위경도 데이터 시각화 등 |
| 공공/선거 등 | 선거 당선자 예측, 통신 빅데이터 이용 공공 정책 수립, 만족도 분석 (T분석, 분산분석 등), 위경도 데이터 시각화, SNA 분석 등 |

데이터 기획, 데이터 수집, 데이터 분석(통계), 시각화 -> Data Scientist 필요함.

# 마케팅분야에서의 빅데이터 활용

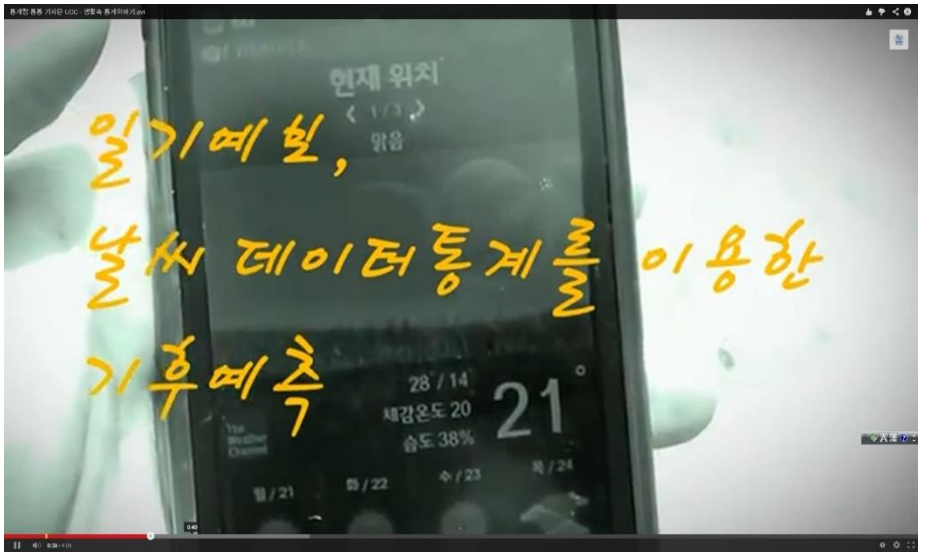
| 마케팅 분야에서의 빅데이터 활용 | | |
|---|---|---|
| CRM | 내용 | 고객분석을 통해 차별적인 경쟁력을 확보하여 높은 성과로 연계 시킬수 있다. |
| | 비고 | 고객충성도 제고, 이탈고객 파악, 잠재고객 파악 등 |
| 맞춤형광고 | 내용 | 효과적인 마케팅을 위한 개별 소비자의 행동파악이 가능하다. |
| | 비고 | 개별소비자 선호제품, 구매촉진, 구매이력, 유사 타 이용자와의 행동을 토대로 광고 제공 |
| 통신 | 내용 | 빅데이터 분석을 통해 수요를 분산시킴으로써 인프라 비용을 절감할 수 있다. |
| | 비고 | 시간대나 이용장소, 이용자 수 등에 이존하는 트래픽을 고려 집중되는 시간에 요금정책 할인 정책 제시 등 |
| 스마트 그리드 (Smart Grid) | 내용 | 방대한 스마트미터의 정보를 집약해 실시간 전력 이용량 측정, 발전량 조절 할 수 있다. |
| | 비고 | 시간데 따른 발전량관리, 가정에서의 효율적 배전 방법 모색 등 |
| 기업의 자산 라이프사이클 관리 | 내용 | 기업이 보유한 다양한 자산들은 적절한 시기에 보수 및 수리가 요구되며, 이 작업의 효율화를 통하여 비용절감을 할 수 있다. |
| | 비고 | 자동차, 기업내 자산관리, 건축산업에서 활용 할 수 있다. |
| 국가적 가치 및 경제적 가치 | 내용 | 공공분야 빅데이터 활용, 정책 및 의사결정에 도움.<br>정부, 기업, 의료, 학술연구 분야에서 그 가치가 입증되고 있음. 정부의 예산절감, 변화에 대한 신속한 대처, 정부신뢰도 향상을 가져올 수 있다. |
| | 비고 | 공공데이터, 소셜데이터 등을 분석하여 대내외의 이슈와 변화를 감지하고 대책을 수립함과 동시에 공공 데이터 공개로 국가 운영을 투명화, 효율화 할 수 있다. |

# 빅데이터 분석기법

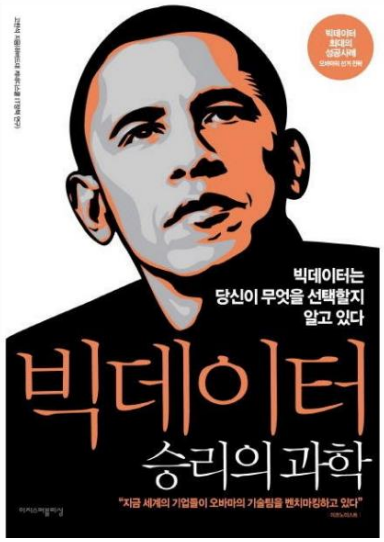
| 빅데이터의 분석 기법 | |
|---|---|
| 데이터 마이닝<br>(Data Mining) | 데이터 마이닝은 데이터 가운데 숨겨져 있는 유용한 상관관계를 발견하여, 미래에 실행 가능한 정보를 추출해 내고 의사 결정에 이용하는 과정을 말한다 (두산백과) |
| 텍스트 마이닝<br>(Text miniing) | 텍스트 마이닝은 대규모의 문서(text)에서 의미 있는 정보를 추출하는 것을 말한다. |
| 오피니언 마이닝<br>(Opinion mining) | 오피니언 마이닝이란 어떤 사안이나 인물, 이슈, 이벤트에 대한 사람들의 의견이나 평가, 태도, 감정 등을 분석하는 것을 말한다. (Liu, 2007) |
| 웹 마이닝<br>(Web mining) | 웹마이닝은 인터넷을 이용하는 과정에서 생성되는 웹 로그(web log) 정보나 검색어로부터 유용한 정보를 추출하는 웹을 대상으로 한 데이터 마이닝을 말한다. (정용찬, 2012) |
| 소셜 분석, 소셜마이닝<br>(Social mining) | 소셜 네트워크의 분석은 수학의 그래프 이론에 뿌리를 두고 있으며, 소셜 네트워크 연결구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾는데 주로 활용된다. 소비자의 흐름이나 패턴 등을 분석하고, 판매나 홍보 마케팅 분야뿐만 아니라 사회의 흐름과 트렌드, 여론변화 추이를 읽어내는 새로운 마이닝 기법이다 (하연, 2012) |
| 현실 마이닝<br>(Reality mining) | 사람들의 행동패턴을 예측하기 휘해 사회적 행동과 관련된 정보를 기기를 통해 얻고 분석하는 기술이다. (정지선, 2012) |
| 군집분석<br>(Cluster Analysis) | 군집분석은 개인이나 여러 개체 중에서 비슷한 속성을 가진 대상을 몇 개의 집단으로 그룹화하고 각 집단의 특성을 파악함으로써 데이터 전체의 구조에 대해 이해하고자 하는 탐색적 분석 기법이다.(김정숙, 2011) |

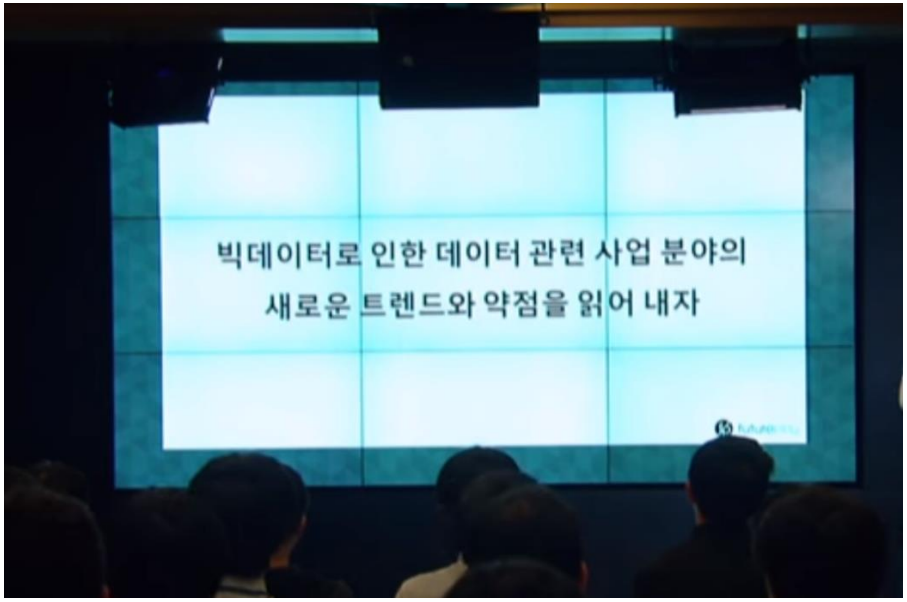
# 빅데이터 활용사례


100년의 영욕
희망 대한민국
빅데이터의 본격적 활용에 나선 정부

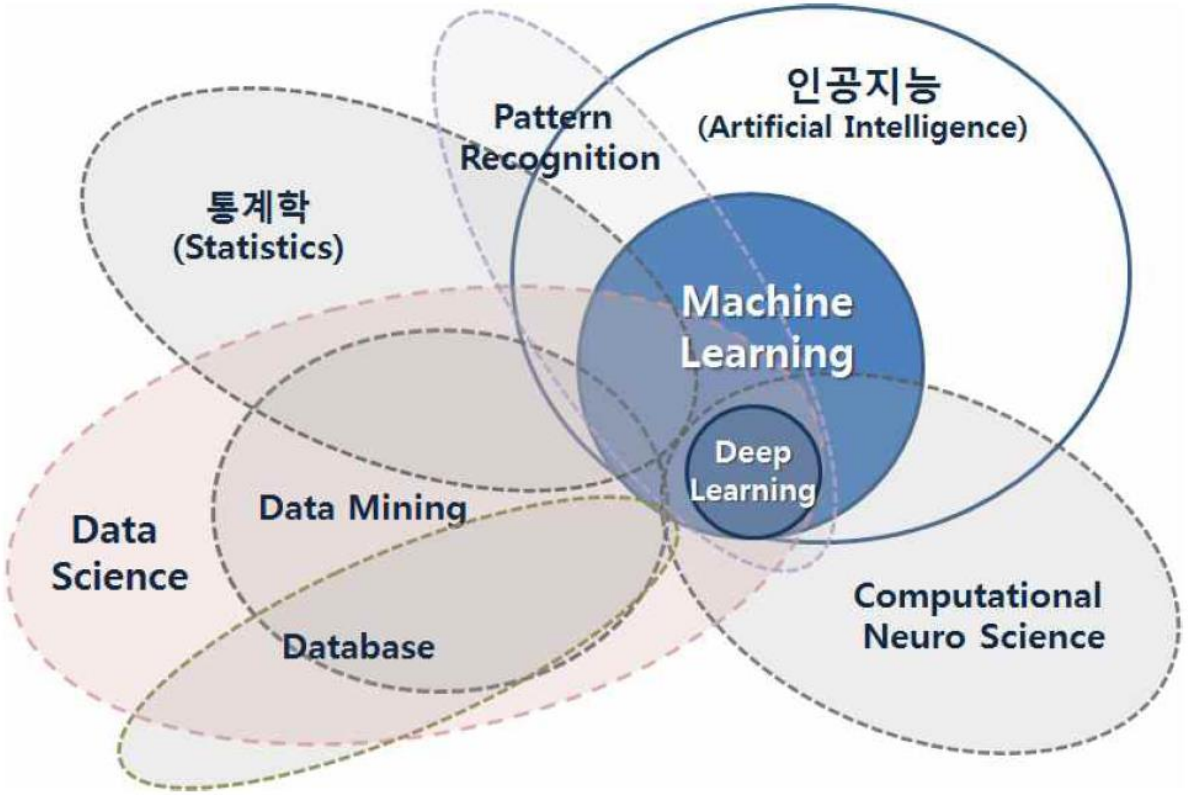

현재 위치
일기예보,
날씨 데이터통계를 이용한
기후예측
28 / 14
체감온도 20
습도 38%
21


VTN HD
SCIENCE
영화 'B형 남자친구'
조사 당시에는
B형 남자에 대한 탐론이 유행했던 때


빅데이터는
당신이 무엇을 선택할지
알고 있다
빅데이터
승리의 과학


빅데이터로 인한 데이터 관련 사업 분야의
새로운 트렌드와 약점을 읽어 내자

# 통계학, 인공지능, 머신러닝, 딥러닝과의 관계




[그림] 머신러닝과 여러 학문과의 관계

Source: NCS 머신러닝 기반 데이터 분석

# 통계학

- 크림전쟁

  - 오스만 제국의 쇠락과 러시아 제국의 남하정책
  - 러시아 제국 견제를 위해 영국, 프랑스 등의 충돌

- 나이팅게일의 활약

- 플로렌스 나이팅게일 (Florence Nightingale, 1820~1910), 백의의 천사
- 전투에 의한 사망자 보다 병원환경이 열악함에 의한 사망자가 더 많음을 밝힘
- 크림전쟁에서 나이팅게일은 정부의 요청으로 야전병원으로 파견, 병원행정의 정상화
- 환자에 대한 정확한 기록과 관리, 병원 내 사망률을 획기적으로 줄임
  -> 위생의 중요성을 데이터 분석으로 사회에 알림


[그림] 흑해 주변 크림반도

Source: https://commons.wikimedia.org


[그림] 크림전쟁

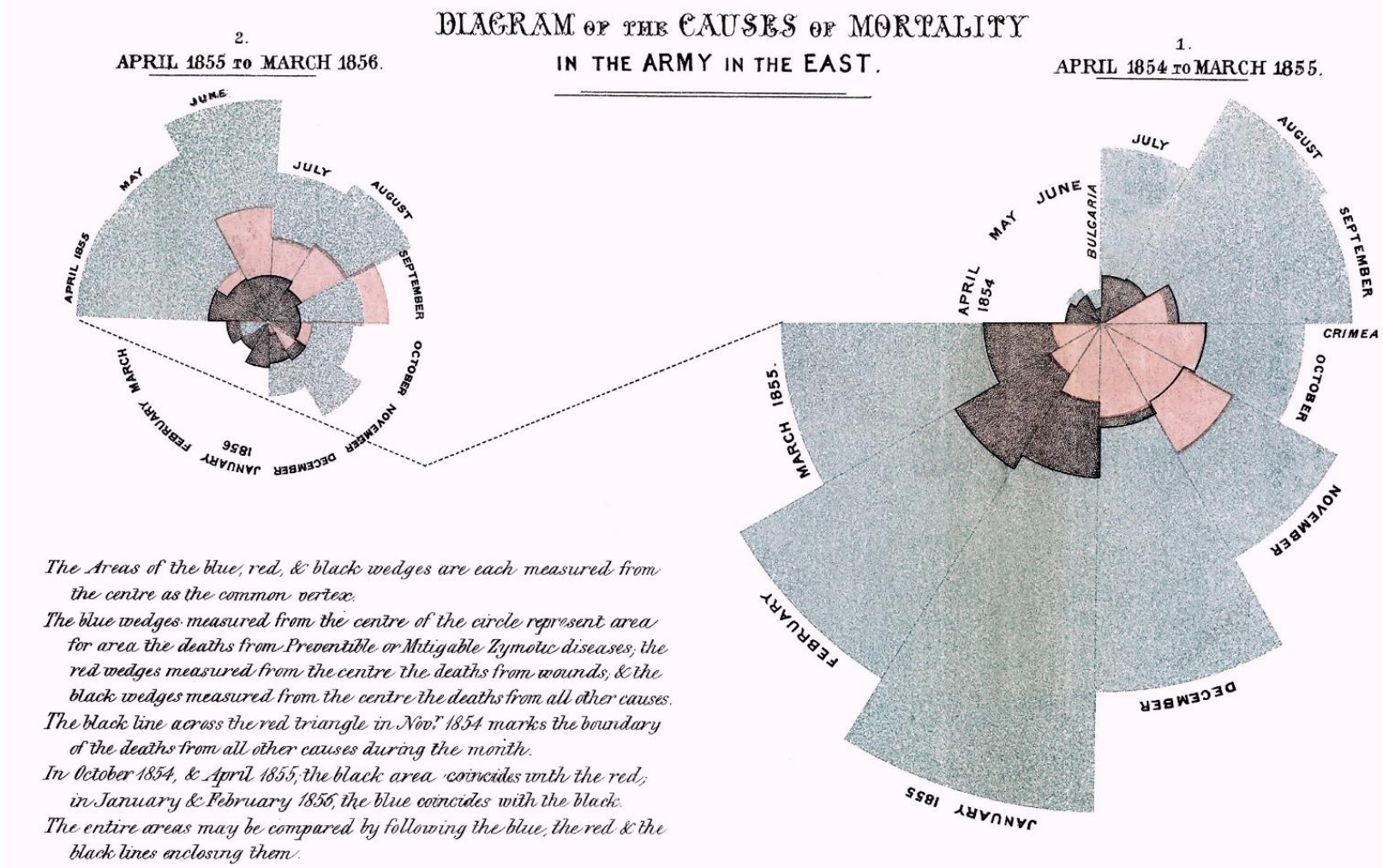
Source: EBS, 통계로 세상을 치료하다

[그림] 나이팅게일

Source: EBS, 통계로 세상을 치료하다

# 통계학

- 나이팅게일의 장미도표(Rose Diagram)




[그림] 장미도표
Source: wikipedia

# 통계학

- **통계학(Statistics, 統計學)**

- 사전적 정의(위키피디아)
  - 수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문

- R. A. Fisher의 통계학
  - 관찰자료에 수학적 원리를 적용하는 응용수학의 한 분야

- 다양한 사회 현상에 대해 자료를 바탕으로 신뢰할 만한 정보를 제공하여 사회 현상을 파악하게 하는 학문
  - 보다 효율적인 정보 전달 방법을 연구하는 분야 또한 중요한 통계학의 분야

- 연구대상 : R. A. Fisher
  - 모집단, 변동량, 자료축약방법

- **모집단과 표본**

  - 모집단

    - '우리가 알고자 하는 대상 전체'
    - 조사 대상의 범위
    - 전수조사 : 모집단 전체를 조사하는 방법

  - 표본

    - '모집단으로부터 조사하기 위해 선택된 조사대상'
    - 모집단 전체를 조사하는 것이 불가능하거나, 수류탄과 같이 조사하면 사라지는 특성을 가진 조사대상, 시간적/공간적 제약이 있을 시 모집단을 잘 대표할 수 있는 조사대상으로 실제 조사대상이 됨
    - 표본조사 : 표본을 조사대상으로 조사하는 방법

## ▪ 통계에서의 데이터 예시

| 변수유형 | 자료유형 | 인구주택 총조사 자료 | 예 |
|---|---|---|---|
| 질적 변수 | 명목형 자료 | 성별, 배우자와의 관계 | 거주지역, 혈액형 등 |
| | 순서형 자료 | 학력 | 학점, 설문문항 등 |
| 양적 변수 | 이산형 자료 | 출생아 수 | 형제 수, 수강과목 수 등 |
| | 연속형 자료 | 연령 | 키, 몸무게 등 |

- **- 척도 : 자료들을 측정하기 위한 측정 도구**

▫ 명목척도와 서열척도

- 각각 명목형 자료, 순서형 자료에 사용하는 척도입니다.

▫ 등간척도와 비율척도 : 양적변수에 사용되는 척도로 사칙연산이 가능합니다.

- 자료의 각 값이 동일한 간격으로 이뤄진 자료를 등간척도라 하고 비율척도와의 차이는 절대 영(0)점이 존재하면 비율척도라 합니다. 절대 영점은 값의 크기가 0으로 결측과 구별해 주시기 바랍니다.
- 기온에서 0도는 절대영점일까 생각해 봅시다.

# 통계 분석 프로그램 R , R studio 설치

- **R**

  - 통계 계산과 그래프 작성을 위한 프로그래밍 언어이자 소프트웨어 환경

  - 1993년 뉴질랜드 오클랜드 대학의 로스 이하카(Ross Ihaka)와 로버트 젠틀맨 (Robert Gentleman)에 의해 공개

  - R 코어팀에 의해 지속적으로 개발

  - GPL하에 배포되어 비용 부담 없이 자유롭게 사용

  - 통계소프트웨어 개발과 자료 분석에 널리 사용되고 있으며, 통계학자들뿐만 아니라 계량 연구를 하는 분야에서 폭넓게 사용

- **홈페이지**
  - http://www.r-project.org

# 통계 분석 프로그램 R , R studio 설치

- **R Studio**

  - R Studio는 R을 위한 통합개발환경(IDE)입니다.

    - 통합개발환경(IDE, Integrated Development Environment)
      - 코딩, 디버그, 컴파일, 배포 등 프로그램 개발에 관련된 모든 작업을 하나의 프로그램 안에서 처리하는 환경을 제공하는 소프트웨어이다(위키피디아, 통합개발환경)
    - R 통합개발환경을 위해 R Studio는 콘솔, 직접 코드를 실행시킬 수 있는 구문강조(Syntax Highlighting) 기능이 있는 편집기와 그림을 그리고 코드 이력을 기록하고 코드 내 버그 찾기 기능과 작업공간 관리 기능을 수행합니다.
    - 오픈소스로 개발되어 있어 무료로 사용할 수 있으며, 일정 비용을 지불하고 각종 지원을 받을 수 있는 유료버전(기업용)이 있습니다.
    - 홈페이지 : https://www.R Studio.com