빅데이터 예측분석

2018년 9월 / 서강대학교 정보통신 대학원 정 화민 교수 (MIS Ph.D)

Contents I

데이터 분석의 필요성

우리의 측우기는 과학적 데이터수집 기기

고대로부터 강우량 측정, 1442년 전국 350군데 측우관측소 운영.

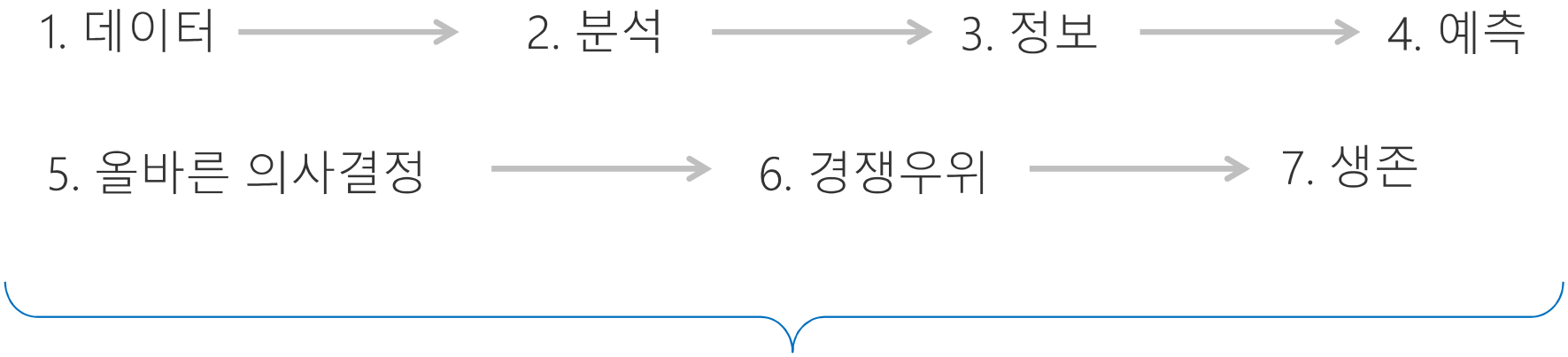
우리의 측우기는 서양의 측우기 보다 200년 앞서 개발되었음.

우리민족의 DNA 속에는 데이터 수집, 분석적 기질이 있다.

8만 대장경

조선왕조실록

데이터 분석과 활용은 생존이다.



데이터 분석을 통한 미래예측과 생존의 과정

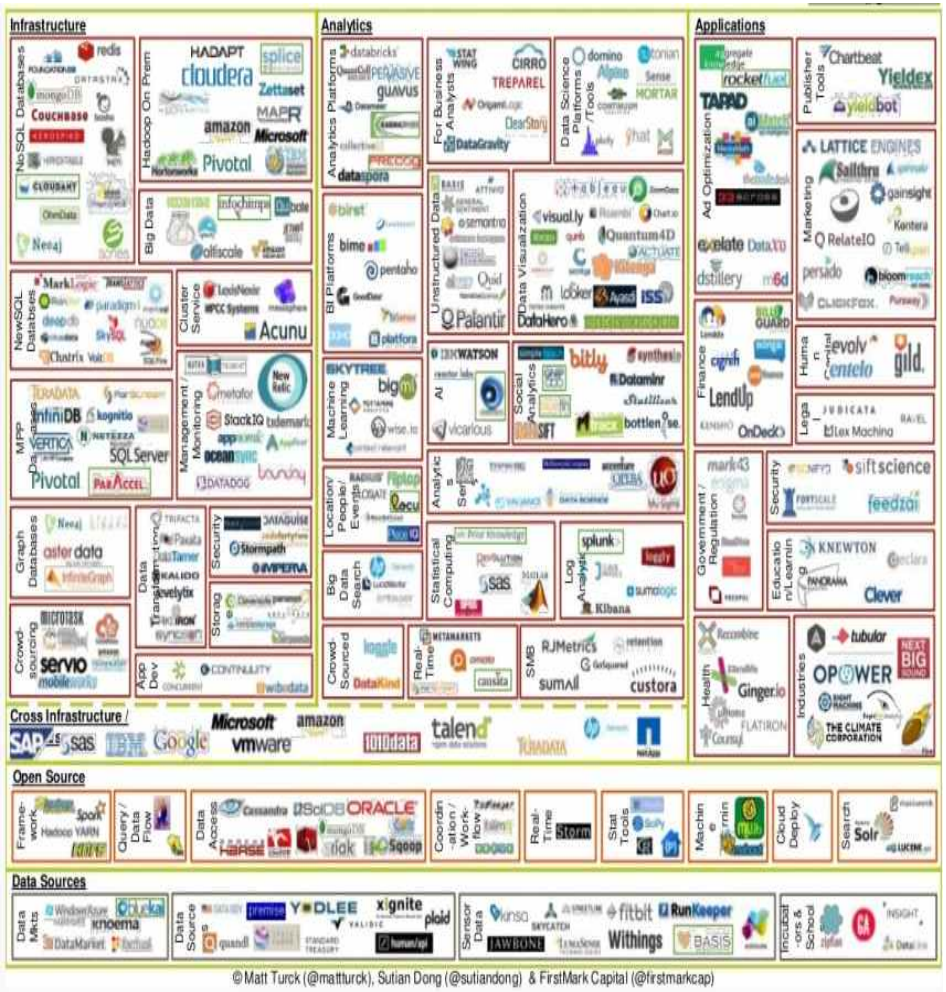
기업경영에서의 데이터 분석 (예: Value chain)

ERP(전사적 자원계획) ?

CRM(고객관계관리) ?

빅데이터 분석에는 적합한 프로그램과 통계적 지식필요

2017년 빅데이터 Landscape



2017년 미국 최고/최악의 직업

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.

국내 시장

국내 데이터 산업 사업체 (2016년 기준 6,726)

(단위 : 개, 억 원)

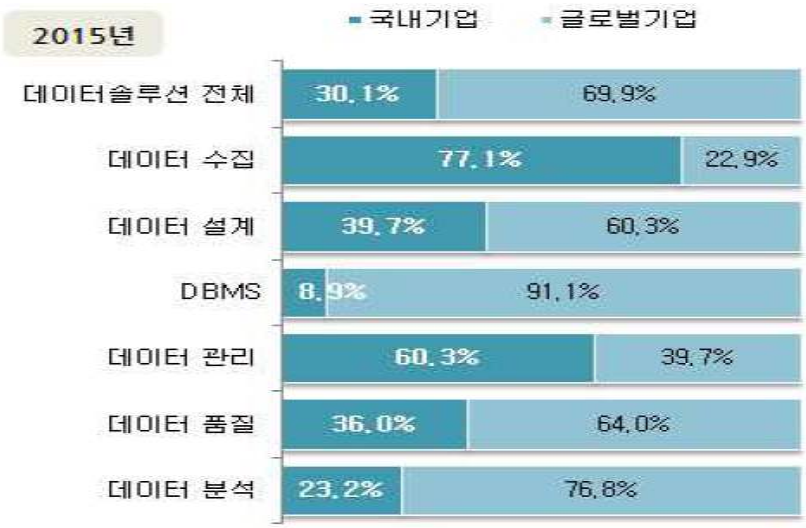
종사자 기준	사업체수	비중
1~10인 미만	5,338	79.4%
10~50인 미만	892	13.3%
50~300인 미만	416	6.2%
300인 이상	80	1.2%
전체	6,726	100.0%

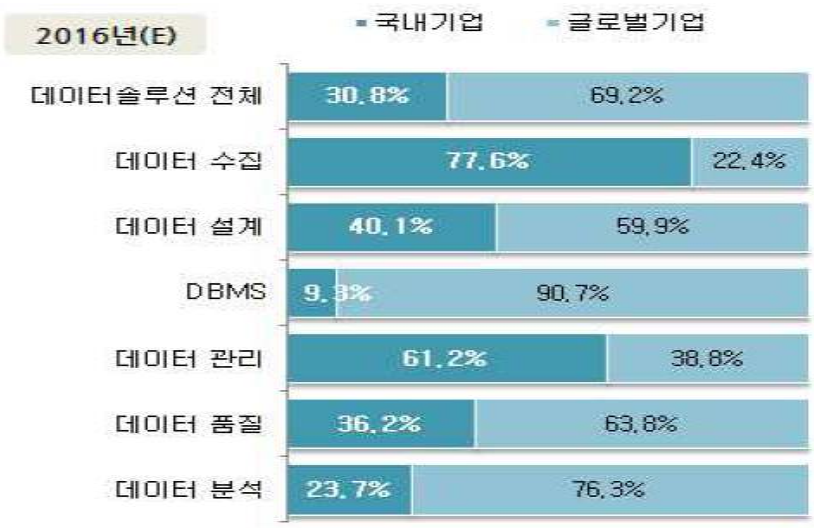
데이터 산업 시장규모 (2016년 기준 국내 시장규모 13조 6천억)

(단위 : 억 원)

구 분	2010년	2011년	2012년	2013년	2014년	2015년	2016년(E)	증감률 '15~'16	CAGR '10~'16
데이터솔루션	6,725	8,717	10,487	10,789	13,619	14,124	14,876	5.3%	14.1%
데이터 구축/컨설팅	37,407	43,180	47,715	49,985	53,730	55,280	55,651	0.7%	6.8%
데이터서비스	42,242	43,218	47,317	52,258	57,329	64,151	66,305	3.4%	7.8%
전체	86,374	95,115	105,519	113,032	124,678	133,555	136,832	2.5%	8.0%

국내 데이터솔루션 기업시장 점유(국내 점유율 약 30%)





Source: 데이터 산업백서(데이터산업진흥원, 2016.)

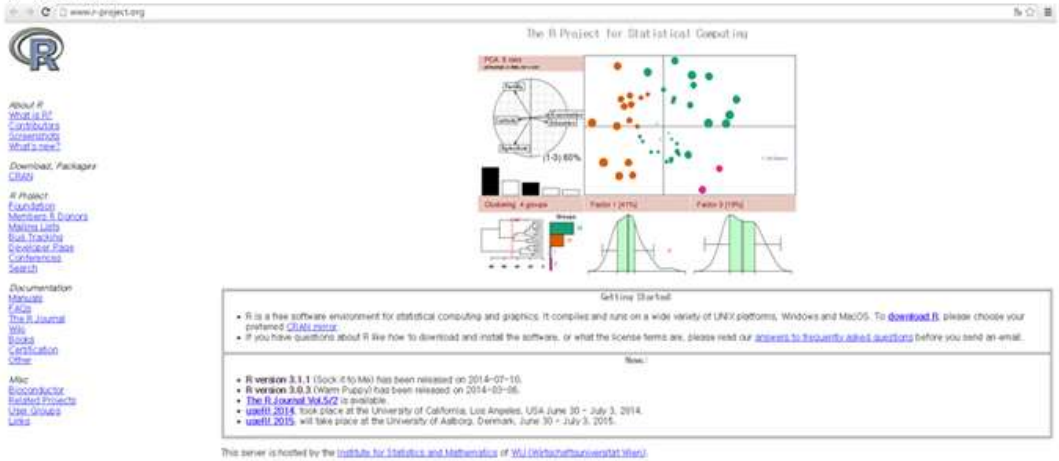
Contents II

빅데이터란?

빅데이터


big data
BIG DATA
빅데이터 경영을바꾸다
BIG DATA
빅 데이터, 세상을 이해하는 새로운 방법
Budget expectations
$32.4 billion in 2017
BIG DATA
빅 데이터의 충격
"DATA IS THE NEW OIL."
빅데이터 경영을바꾸다
DATA BIG DATA
ANALYSIS
COLLECTION
SIZE
MANAGEMENT
STORAGE

분석도구 : R


The R Project for Statistical Computing
Getting Started
News

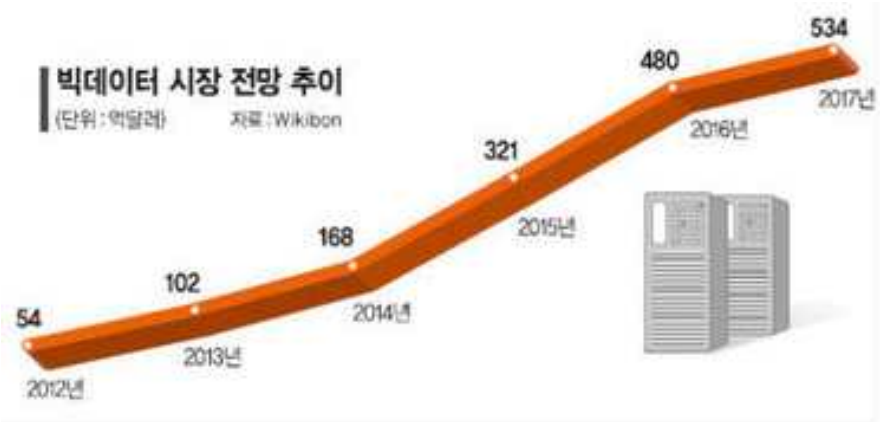
빅데이터는 21세기 원유

빅데이터

기존 데이터의 수집, 관리 및 처리 소프트웨어의 수용한계를 넘어서는 크기의 데이터를 말한다.

빅데이터의 처리는 분석기술과 표현기술로 구분

(출처: 위키백과)



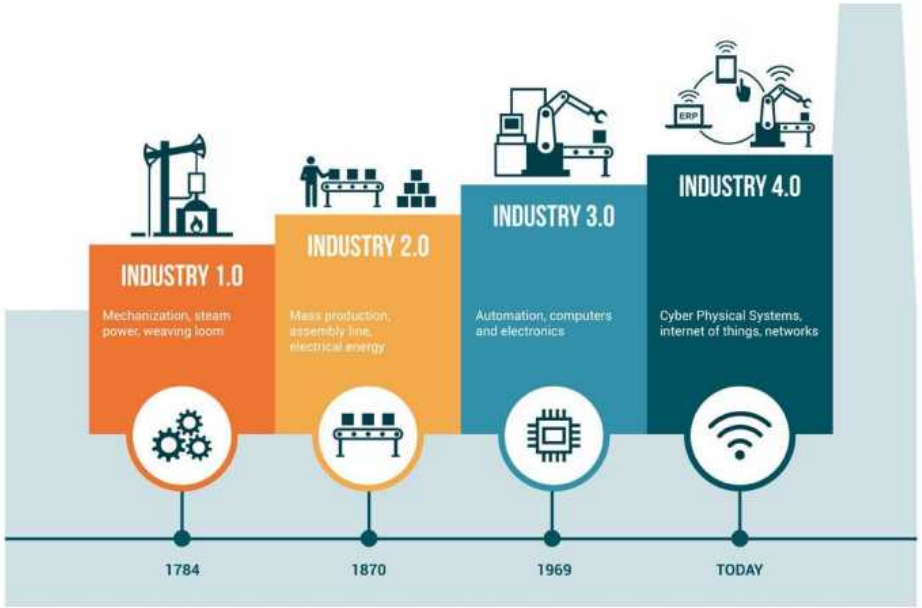
빅데이터 정의

기관	정의	참고
맥킨지	기존방식의 저장, 관리 분석할 수 있는 범위를 초과하는 규모의 데이터	데이터 규모
IDC	다양한 종류의 데이터로부터 낮은 비용으로 가치를 추출하고 데이터 초고속 수집, 발굴 분석을 지원	업무 수행
가트너	빅데이터는 21세기의 원유	데이터 활용

빅데이터 출현

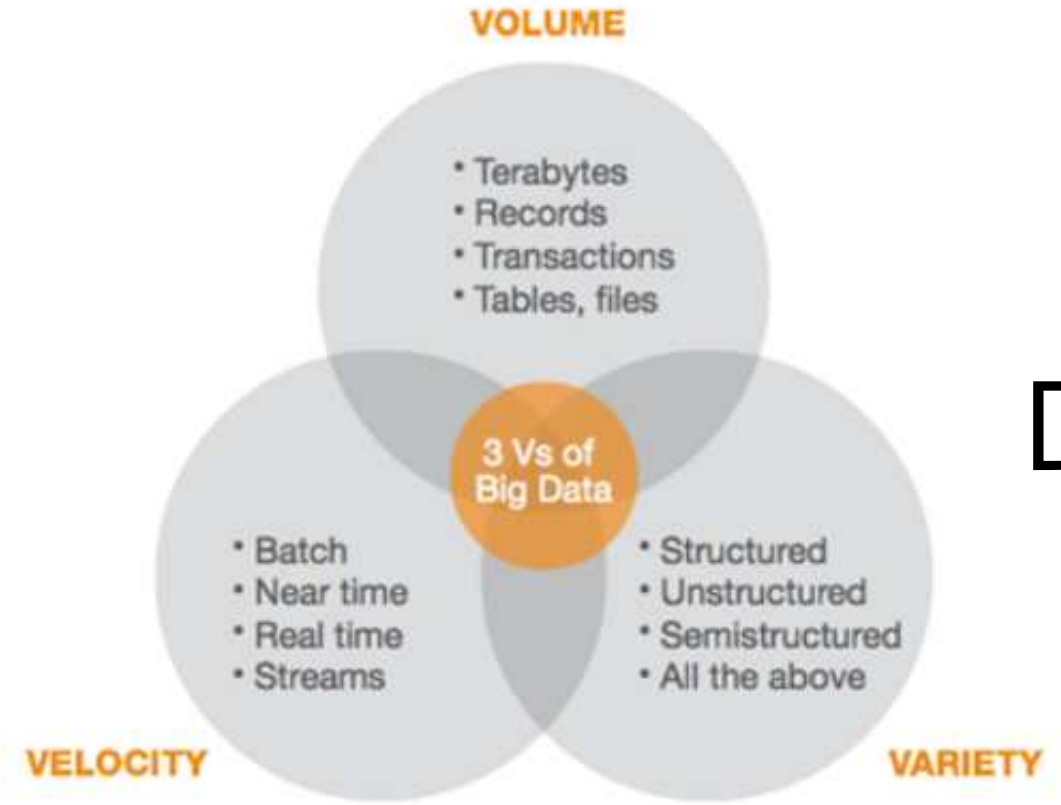

ICT패러다임 변화
데이터의 양적팽창
기공/분석
신가치 창출
기술 발달
(고성능, 고집적화,
모바일화
데이터 생성, 수집,
관리
데이터 활용:
지원+역량+기술
경제적 가치 창출
비용 저하
(저장매체, 대역폭,
속도)
새로운 ICT
패러다임 견인
빅데이터 출현
소셜 네트워크
서비스 발달
사회적 문제 해결
그림자 정보
사물 정보 증가

빅데이터 등장배경


INDUSTRY 1.0
Mechanization, steam power, weaving loom
INDUSTRY 2.0
Mass production, assembly line, electrical energy
INDUSTRY 3.0
Automation, computers and electronics
INDUSTRY 4.0
Cyber Physical Systems, internet of things, networks
1784
1870
1969
TODAY

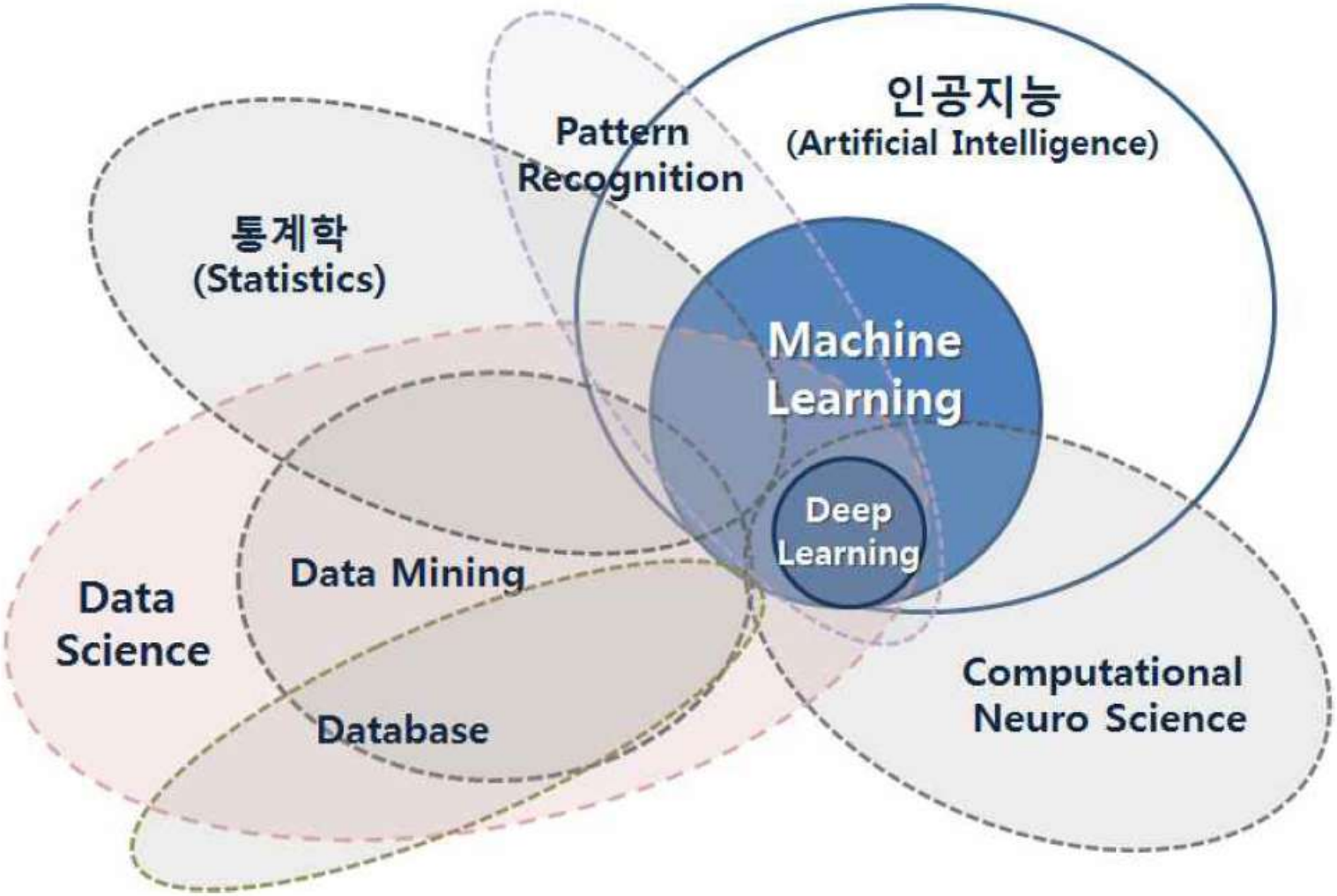

2018 This Is What Happens In An Internet Minute
60 SECONDS
facebook
973,000 Logins
18 Million Text Messages
YouTube
4.3 Million Videos Viewed
375,000 Apps Downloaded
174,000 Scrolling Instagram
481,000 Tweets Sent
1.1 Million Swipes
187 Million Emails Sent
936,073 Views
67 Voice-First Devices Shipped
38 Million Messages
25,000 GIFs Sent via Messenger
2.4 Million Snaps Created
$862,823 Spent Online
266,000 Hours Watched
NETFLIX
3.7 Million Search Queries
Google
Created By:
@LoriLewis
@OfficiallyChadd

빅데이터의 정의 3V



데이터 크기
데이터 다양성
데이터 속도

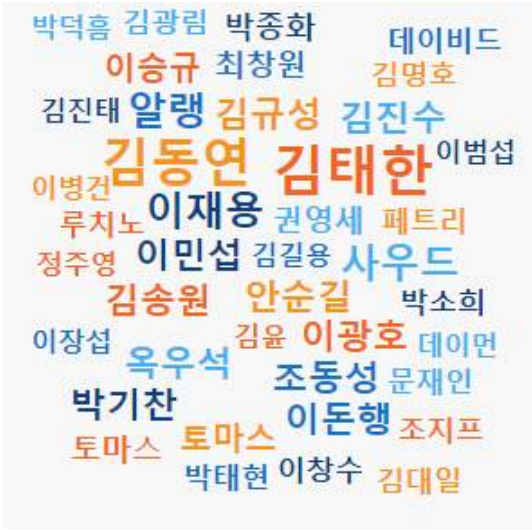
빅데이터 관련 학문


인공지능
(Artificial Intelligence)
Pattern
Recognition
통계학
(Statistics)
Machine
Learning
Deep
Learning
Data Mining
Data
Science
Database
Computational
Neuro Science

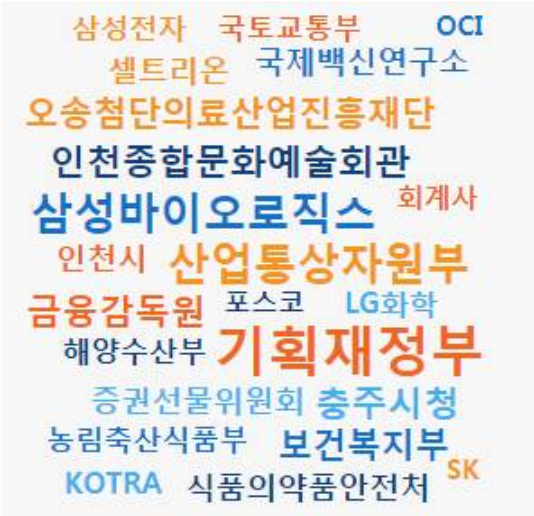
바이오 산업관련 Data Mining

키워드

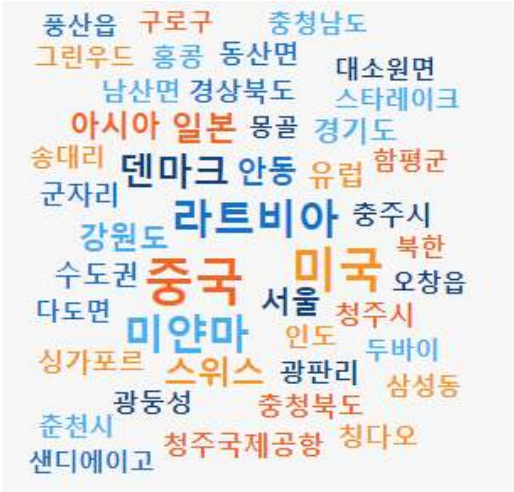
인물



기관



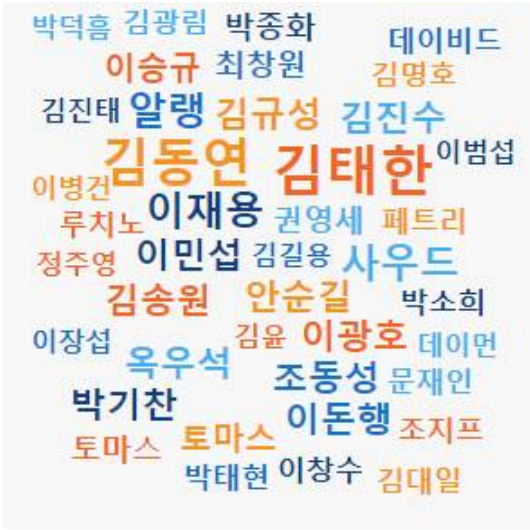
장소



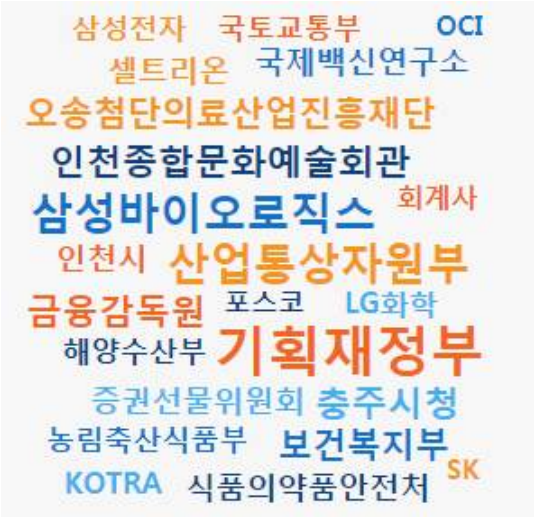
바이오 산업관련 Data Mining

키워드

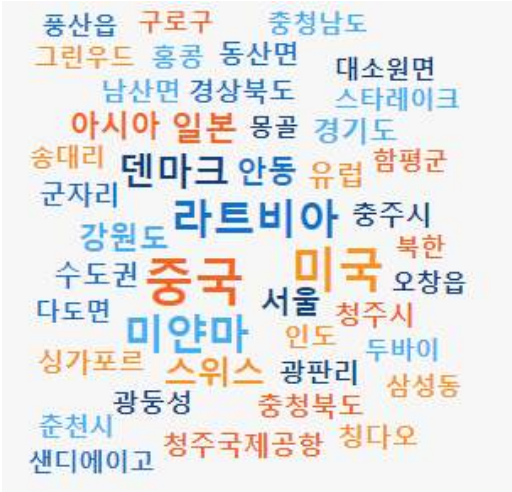
인물



기관



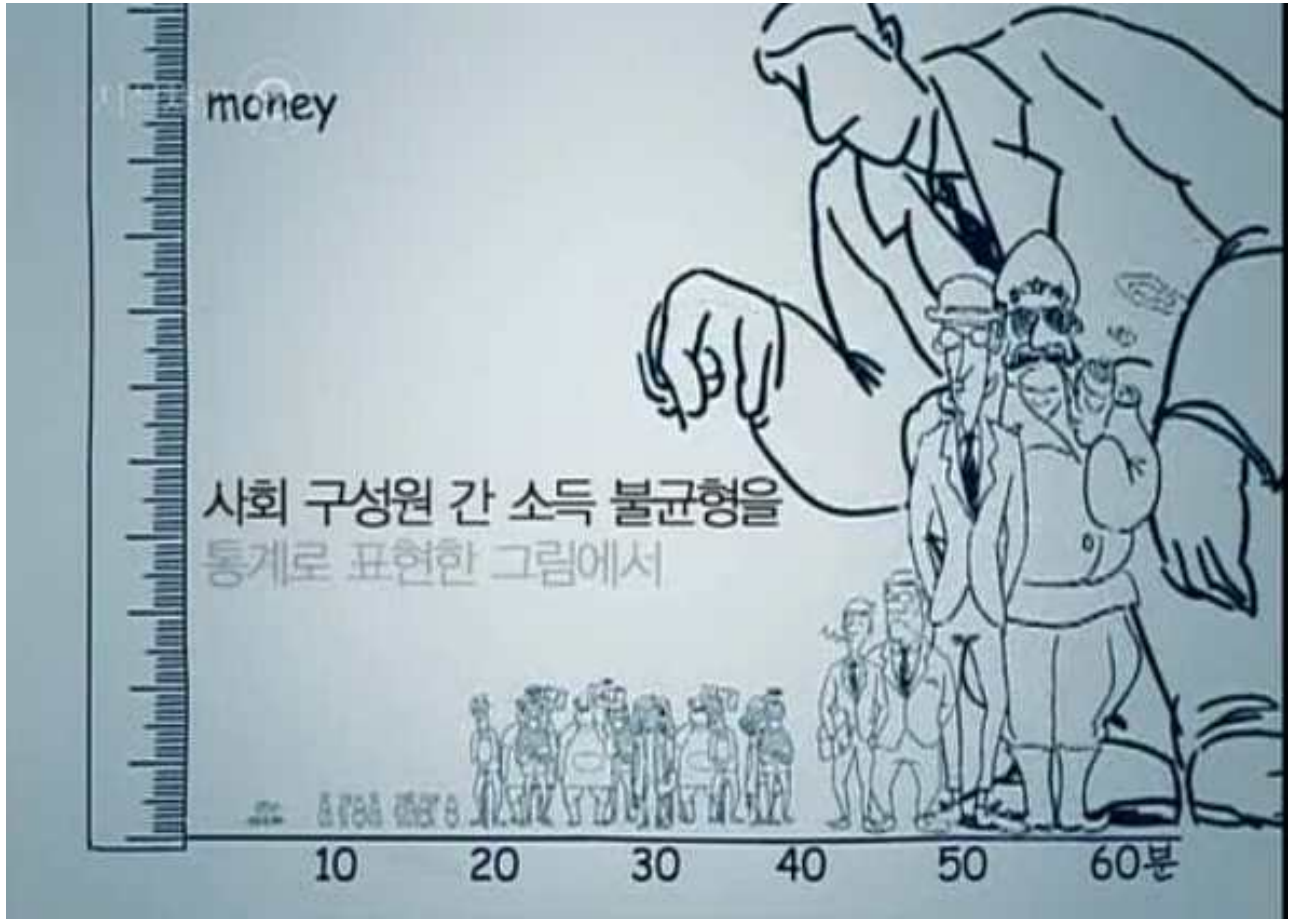
장소



Contents III

데이터 분석 기초/ 통계

평균의 함정



마케팅분야에서의 빅데이터 활용

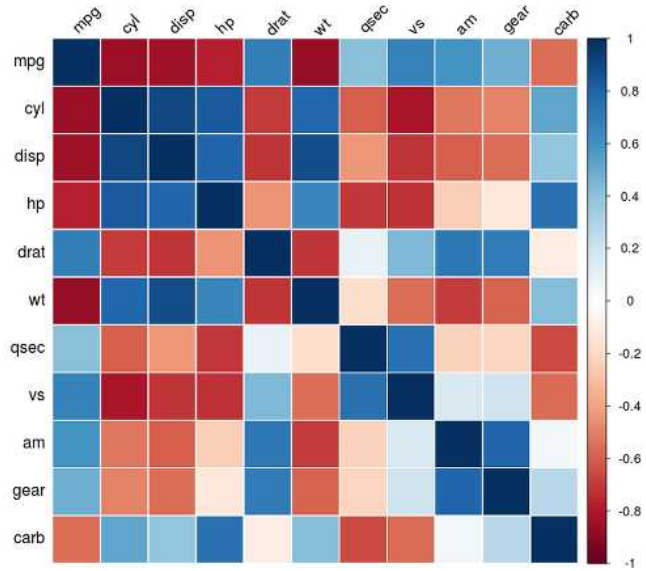
마케팅 분야에서의 빅데이터 활용		
CRM	내용	고객분석을 통해 차별적인 경쟁력을 확보하여 높은 성과로 연계 시킬수 있다.
	비고	고객충성도 제고, 이탈고객 파악, 잠재고객 파악 등
맞춤형광고	내용	효과적인 마케팅을 위한 개별 소비자의 행동파악이 가능하다.
	비고	개별소비자 선호제품, 구매촉진, 구매이력, 유사 타 이용자와의 행동을 토대로 광고 제공
통신	내용	빅데이터 분석을 통해 수요를 분산시킴으로써 인프라 비용을 절감할 수 있다.
	비고	시간대나 이용장소, 이용자 수 등에 이존하는 트래픽을 고려 집중되는 시간에 요금정책 할인 정책 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. 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 (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 (Text miniing)	텍스트 마이닝은 대규모의 문서(text)에서 의미 있는 정보를 추출하는 것을 말한다.
오피니언 마이닝 (Opinion mining)	오피니언 마이닝이란 어떤 사안이나 인물, 이슈, 이벤트에 대한 사람들의 의견이나 평가, 태도, 감정 등을 분석하는 것을 말한다. (Liu, 2007)
웹 마이닝 (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 (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 (Reality mining)	사람들의 행동패턴을 예측하기 휘해 사회적 행동과 관련된 정보를 기기를 통해 얻고 분석하는 기술이다. (정지선, 2012)
군집분석 (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)

통계기법을 이용한 상관관계 분석

누가 누구를 더 좋아할까?



설문데이터를 활용한 상관분석

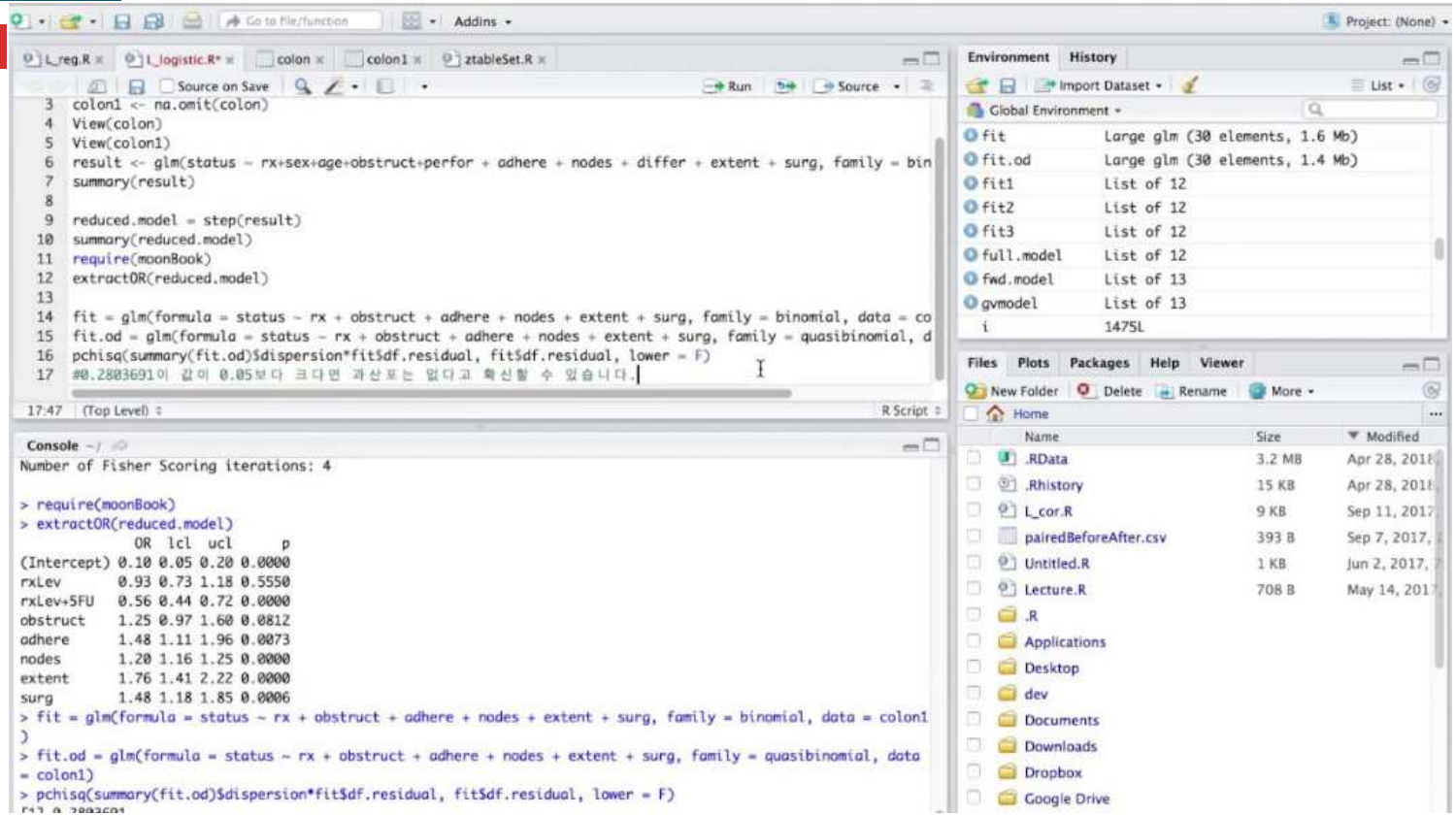
설문지번호	덕선	택이	선우	동룡	정환
1	7	7	6	4	7
2	7	7	6	5	5
3	6	5	7	5	5
4	5	4	4	7	4
5	5	5	6	6	7

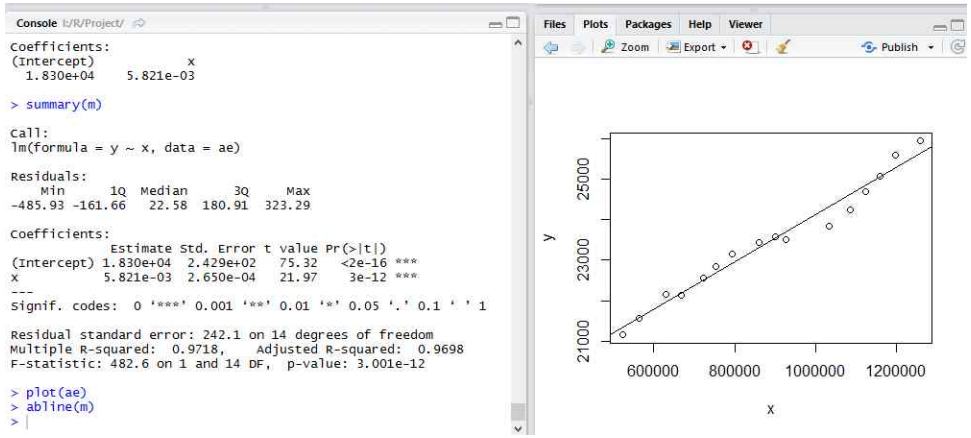
데이터 코딩

	덕선	택이	선우	동룡	정환
덕선	1				
택이	0.931695	1			
선우	0.456435	0.442269	1		
동룡	-0.87706	-0.84984	-0.72058	1	
정환	0.186339	0.444444	0.442269	-0.52298	1

분석결과 : 덕선과 택이는 아주 높은 상관관계(+)가 있음

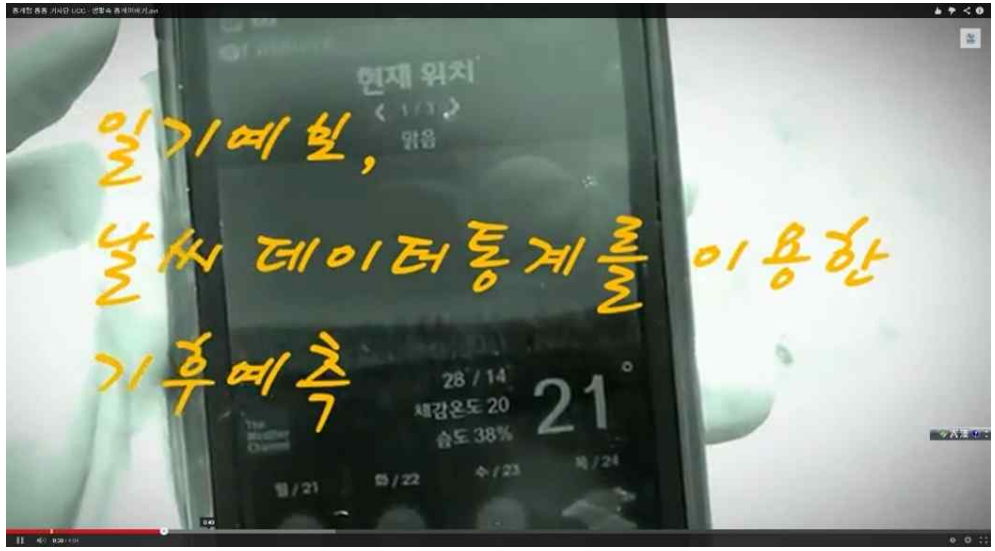
인과관계를 밝히는 회귀분석 (R)



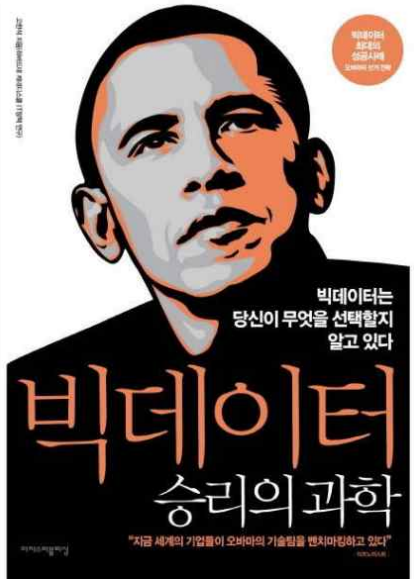


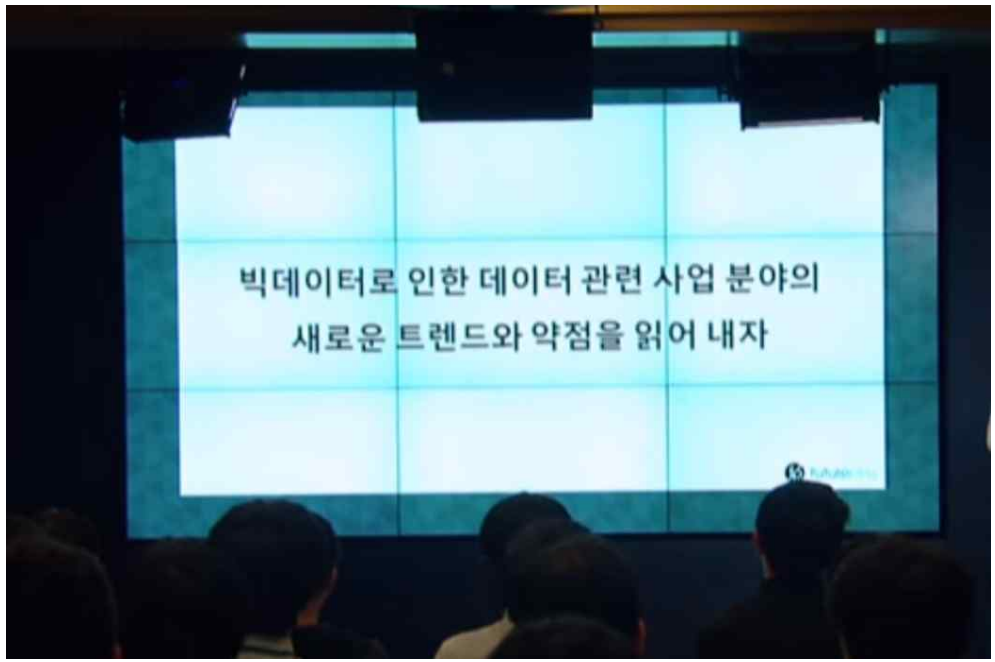
빅데이터 활용사례


100분의 행복
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 담론이 유행했던 때


빅데이터는
당신이 무엇을 선택할지
알고 있다
빅데이터
승리의 과학

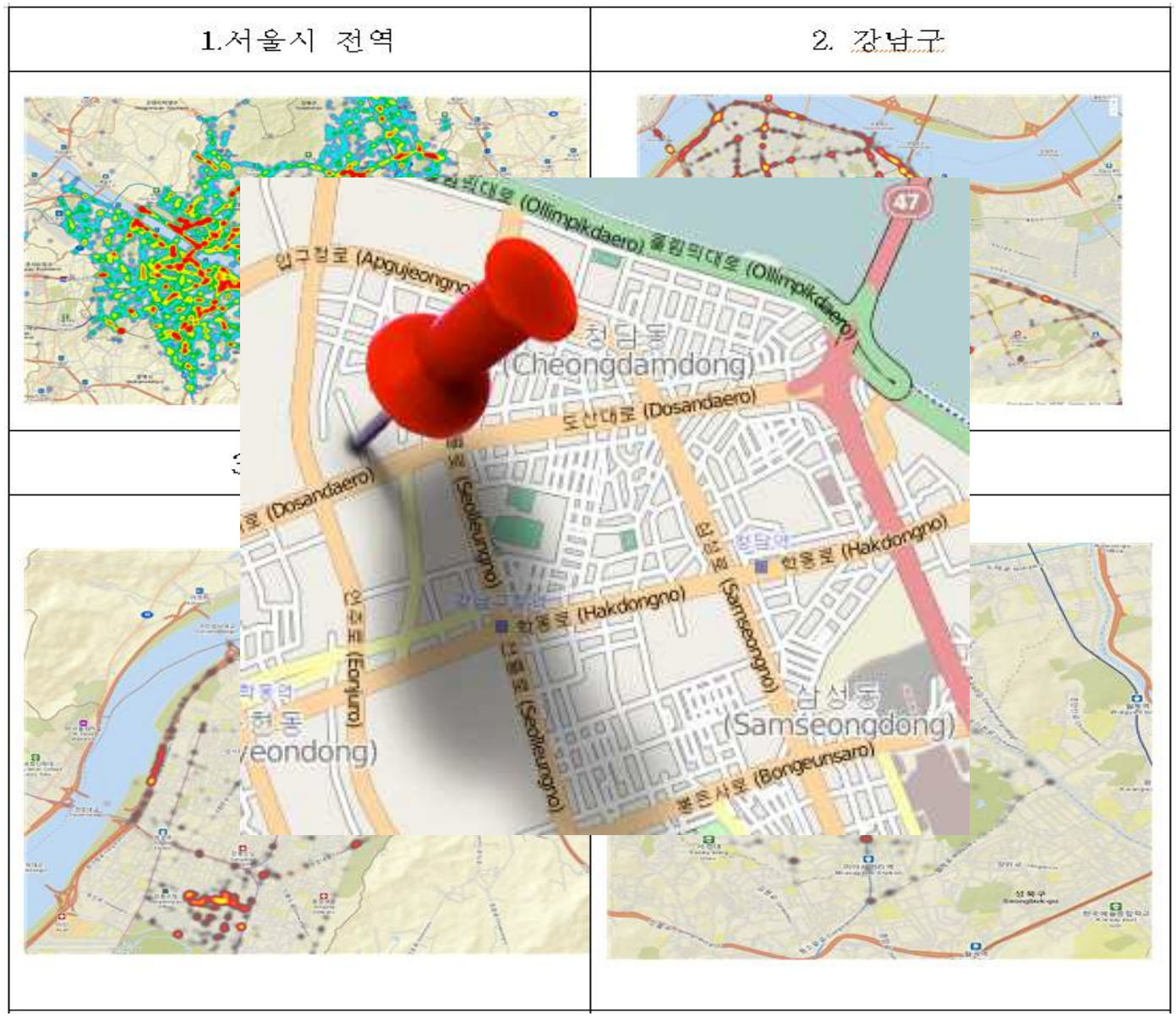

빅데이터로 인한 데이터 관련 사업 분야의
새로운 트렌드와 약점을 읽어 내자

통신 빅데이터 활용

통신 데이터를 활용한 유동인구 시각화

교통빅데이터 활용 시각화

위경도 데이터를 활용한 교통



데이터를 분석하여 정보로 만들고
이 정보를 통하여 미래를 예측하면
올바른 의사결정을 할 수 있다. 이는
개인 , 기업 , 그리고 정책을 수행하는
정부기관의 경쟁우위를 가져오고
생존을 약속한다.

감사합니다.

정화민 교수
e-mail : vivahyatt@gmail.com