소프트웨어 세미나1

-생활 속 빅데이터-

목차

- 빅데이터란?
- 빅데이터 활용사례
- 빅데이터 분석 툴
- 주요 이슈 및 문제점

01. 빅데이터란?

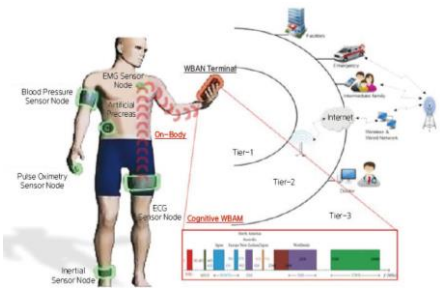
big data

빅데이터는 기존 데이터베이스 관리도구의 능력을 넘어서는 대량(수십 테라바이트)의 정형 또는 데이터베이스 형태가 아닌 비정형 데이터 집합을 포함한 데이터로부터 **가치를 추출**하고 **결과를 분석**하는 기술

비즈니스 데이터

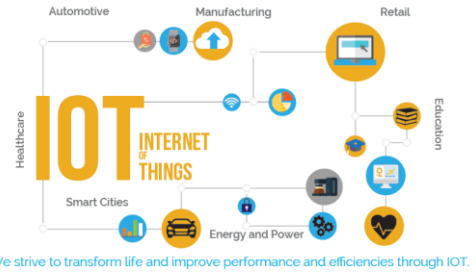
소셜 데이터

의료정보

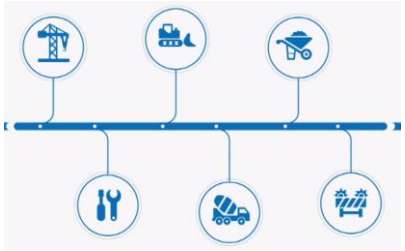


동영상

사물인터넷



기반시설

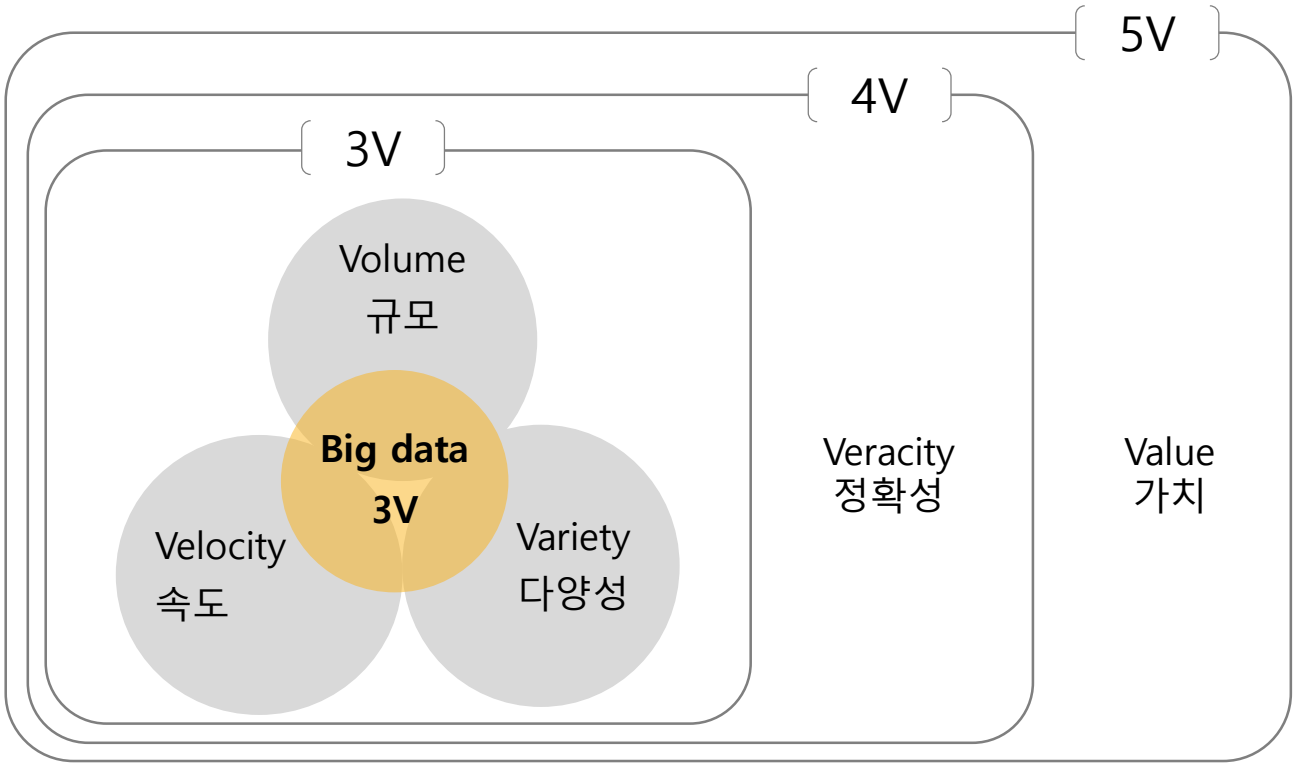


출처 : https://ko.wikipedia.org/wiki/

그림출처 : https://hypertext30.wordpress.com/2014/06/16/펠레의-저주와-빅데이터/

big data의 속성

- 더그 레이니Doug Laney(가트너의 애널리스트)는 ***빅데이터의 속성***을 **3V**(Volume, Velocity, Variety)로 정의했다. 큰 용량, 다양성이 높은 자산, 빠른 속도를 말한다.
- IBM은 여기에 Veracity 정확성을 추가하여 **4V**로 정의했으며,
- 최근 Value 가치를 추가하여 **5V**로 정의하였다.



*가트너Gartner: 미국 시장조사 및 컨설팅 회사로, IT분야의 연구 및 자문 담당.

출처 : 한빛아카데미 '빅데이터 컴퓨팅 기술' 박두순 저

big data는 얼마나 커야 할까

A

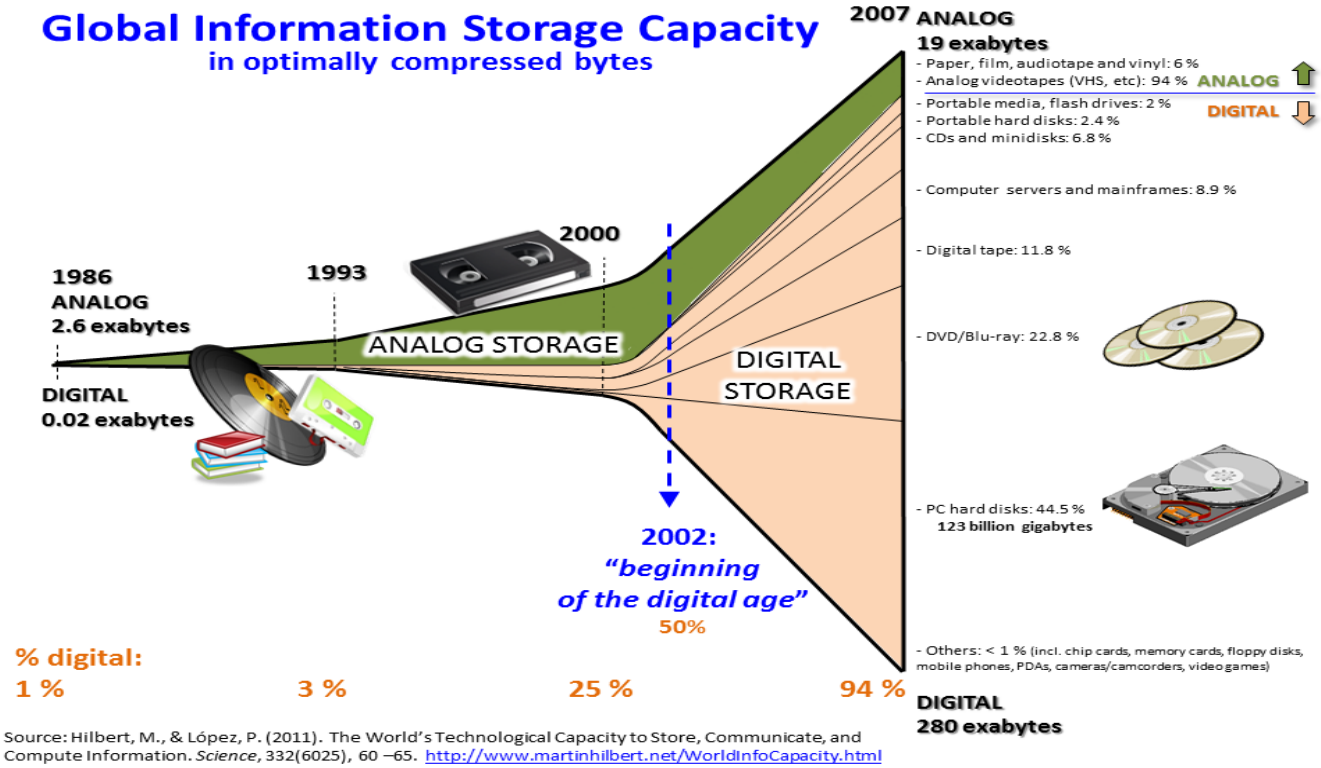
1 byte

1024 byte	=	1 K byte	=	2^{10} byte	(Kilo)
1024 K byte	=	1 M byte	=	2^{20} byte	(Mega)
1024 M byte	=	1 G byte	=	2^{30} byte	(Giga)
1024 G byte	**=**	**1 T byte**	**=**	**2^{40} byte**	**(Tera)**
1024 T byte	=	1 P byte	=	2^{50} byte	(Peta)
1024 P byte	=	1 E byte	=	2^{60} byte	(Exa)
1024 E byte	=	1 Z byte	=	2^{70} byte	(Zetta)

빅데이터는 기존의 IT 기술로는 처리하기 어려운 양을 의미한다. IT기술로 처리하는 것보다 빅데이터 기술을 사용하는 것이 가격, 처리 속도 면에서 훨씬 유리한 것이 빅데이터다. 최소 50TB 정도는 되어야 한다.

big data의 등장 배경

- 1990년대 이후 인터넷이 확산→ 2000년대 이후 개인화 서비스와 SNS의 확산→ 인터넷 서비스가 통신, 음악, 검색, 쇼핑, 동영상 등의 서비스를 제공하는 환경으로 바뀜.
- 스마트폰의 보급은 2020년 40제타 바이트의 데이터 생산을 예측한다.
- 데이터를 저장, 관리, 분석하는데 한계→정보기술의 패러다임이 바뀌어 **빅데이터 개념 등장**

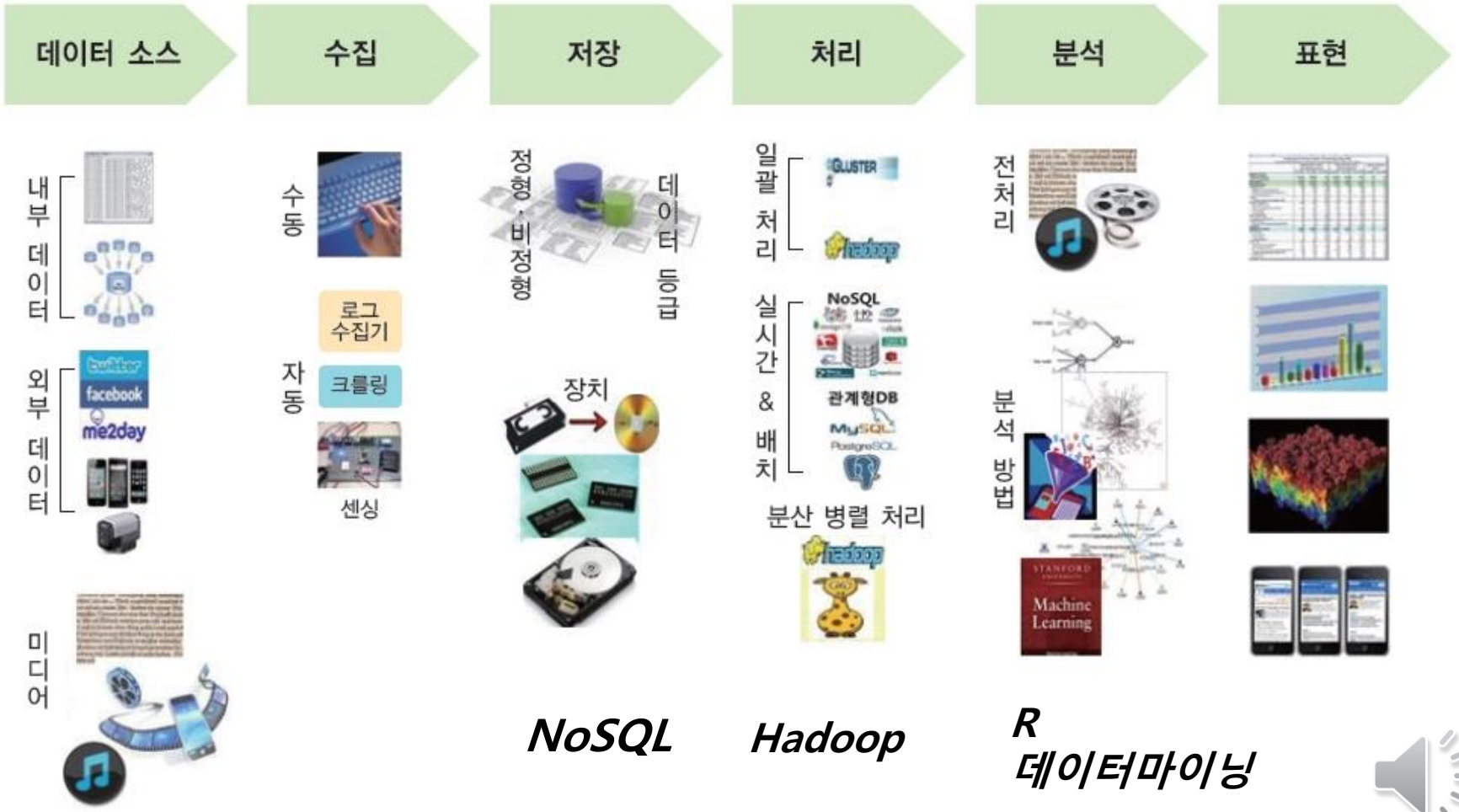


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

출처 : 한빛아카데미 '빅데이터 컴퓨팅 기술' 박두순 저
그림출처 : 워싱턴포스트

big data 처리과정

- 기존의 데이터와 속성이 달라 수집, 저장, 처리, 분석, 시각화하는 새로운 방법이 등장하였고, 이 기술을 위한 새로운 직업이 등장



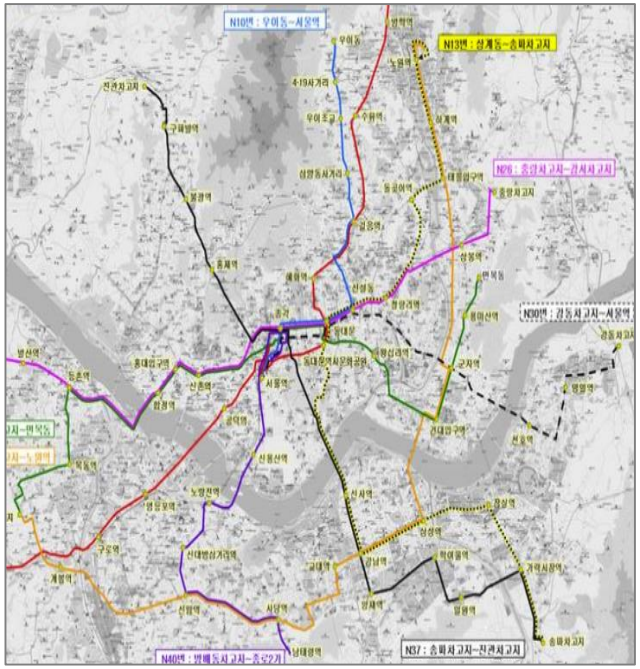
출처 : 한빛아카데미 '빅데이터 컴퓨팅 기술' 박두순 저

Big data

https://www.youtube.com/watch?v=vyRC1bUAwFE 5분

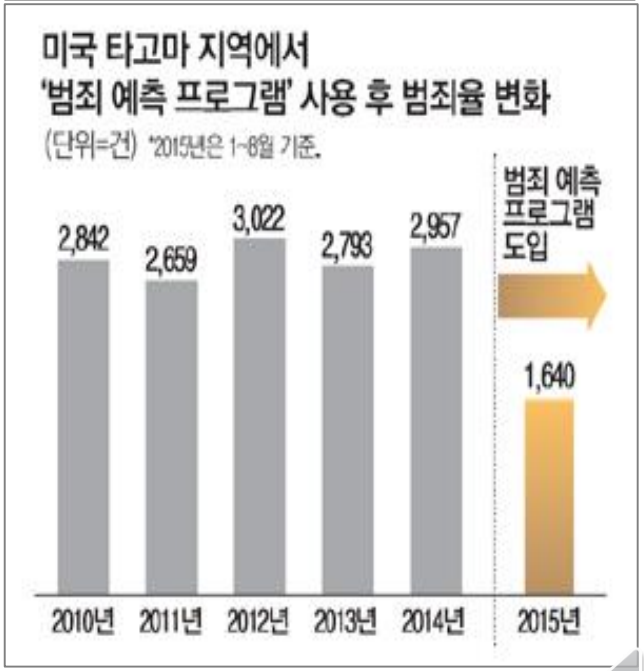
02. 빅데이터의 활용사례

빅데이터 활용 사례


N10번 : 우이동~서울역
N13번 : 상계동~송파차고지
N16 : 중앙차고지~강서차고지
N30번 : 강동차고지~서울역
N37 : 송파차고지~진관차고지
N40번 : 방배동차고지~중구2가


보건의료, 치료에서 예방·건강 관리로…
불치병까지 미리 막는 빅데이터
보건의료 빅데이터 기술분야별 성장률
보건의료 빅데이터 시장 성장률


VOLUME
SCALE OF DATA
VARIETY
FORMS OF DATA
BIG DATA
ANALYSIS OF DATA-FLOW
VELOCITY
UNCERTAINTY OF DATA
VERACITY

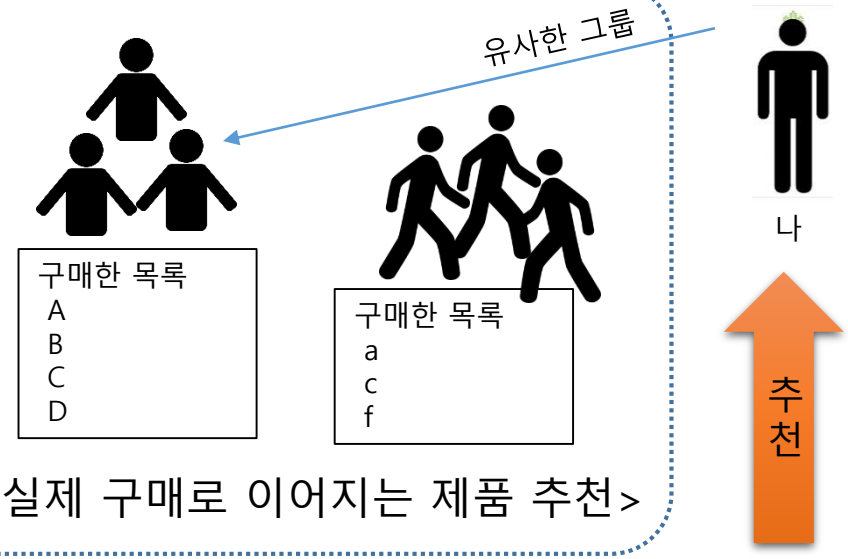

미국 타고마 지역에서
'범죄 예측 프로그램' 사용 후 범죄율 변화
(단위=건) *2015년은 1~8월 기준.
2,842
2,659
3,022
2,793
2,957
1,640
범죄 예측 프로그램 도입
2010년
2011년
2012년
2013년
2014년
2015년


Hot Trend
매뉴팩처링 트랜스포메이션
Manufacturing Transformation
빅데이터,
제조업을 재발견하다

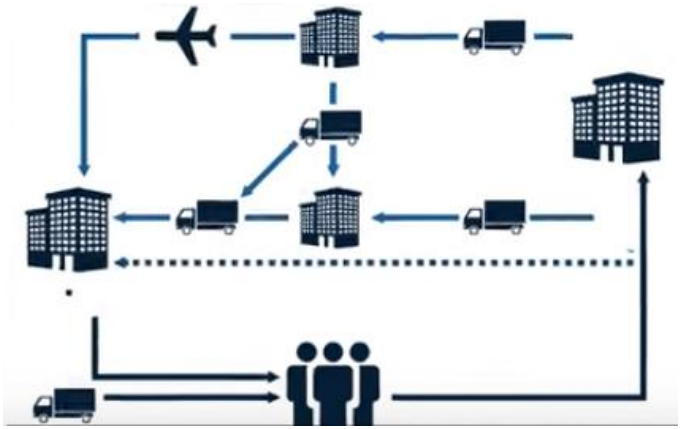
빅데이터 활용 사례1

아마존닷컴 추천 서비스

모든 고객들의 구매 내역을 기록하고 분석하여 고객의 취향과 관심사를 파악하여 추천상품으로 제공함



<실제 구매로 이어지는 제품 추천>

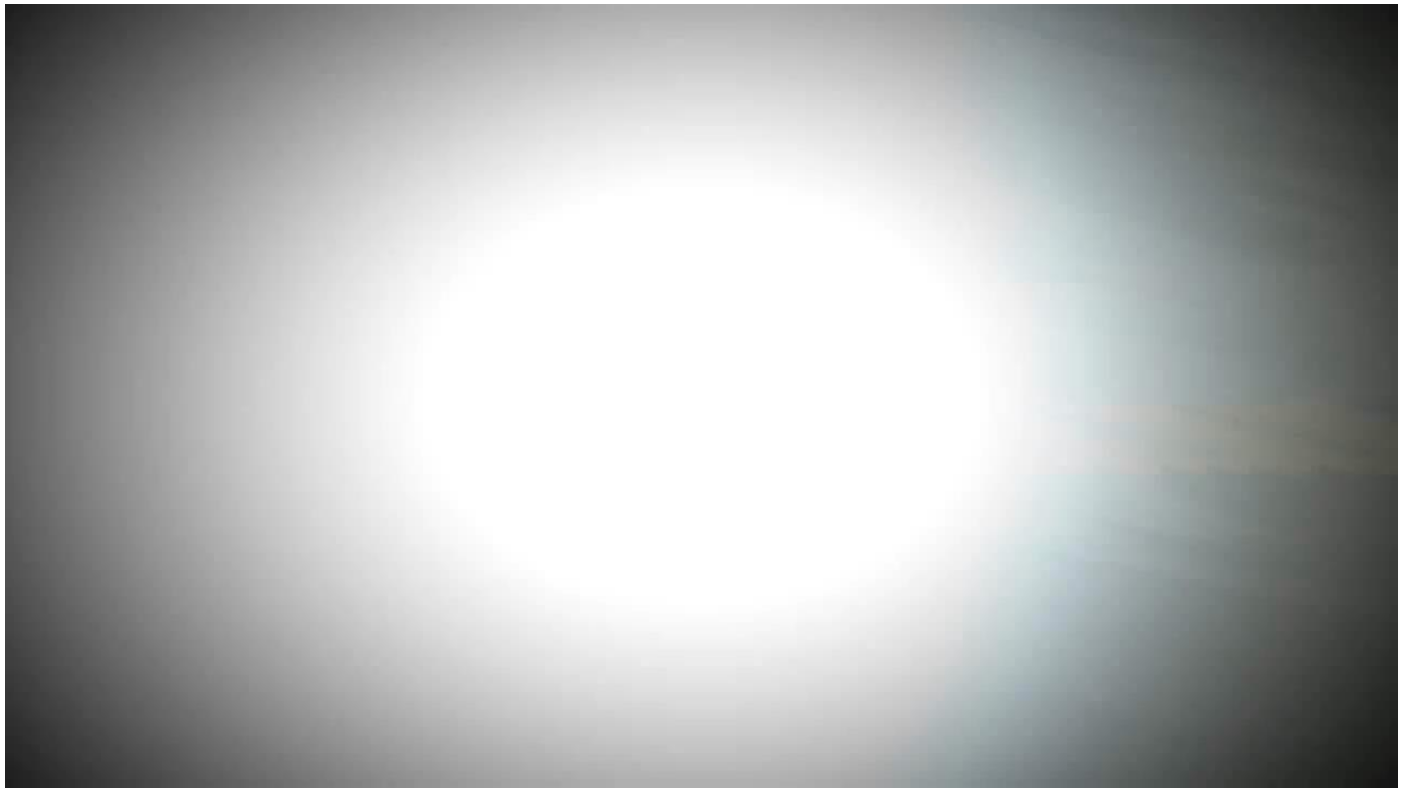


<구매 예측 상품 물류센터에 준비>

빅데이터 활용 사례2

google 트랜드와 감기

구글은 미국 "질병통제예방센터 " 보다 빠르게 독감 바이러스 확산을 예측하였다. 이것은 실시간으로 일어나는 구글 검색 빈도를 파악하여 빅데이터 분석을 통해 **독감지도를 생성**하였기 때문이다.



https://www.youtube.com/watch?v=X4hMFym0-uo 1분36초

빅데이터 활용 사례3

서울시 심야버스 노선

서울시는 KT와 MOU를 맺고 서울시를 1km의 육각형 셀로 자른 후 1,250개의 각 셀에서 심야시간에 전화한 위치와 전화 받은 위치를 분석하여 심야 버스 노선을 완성하였다. 2013년

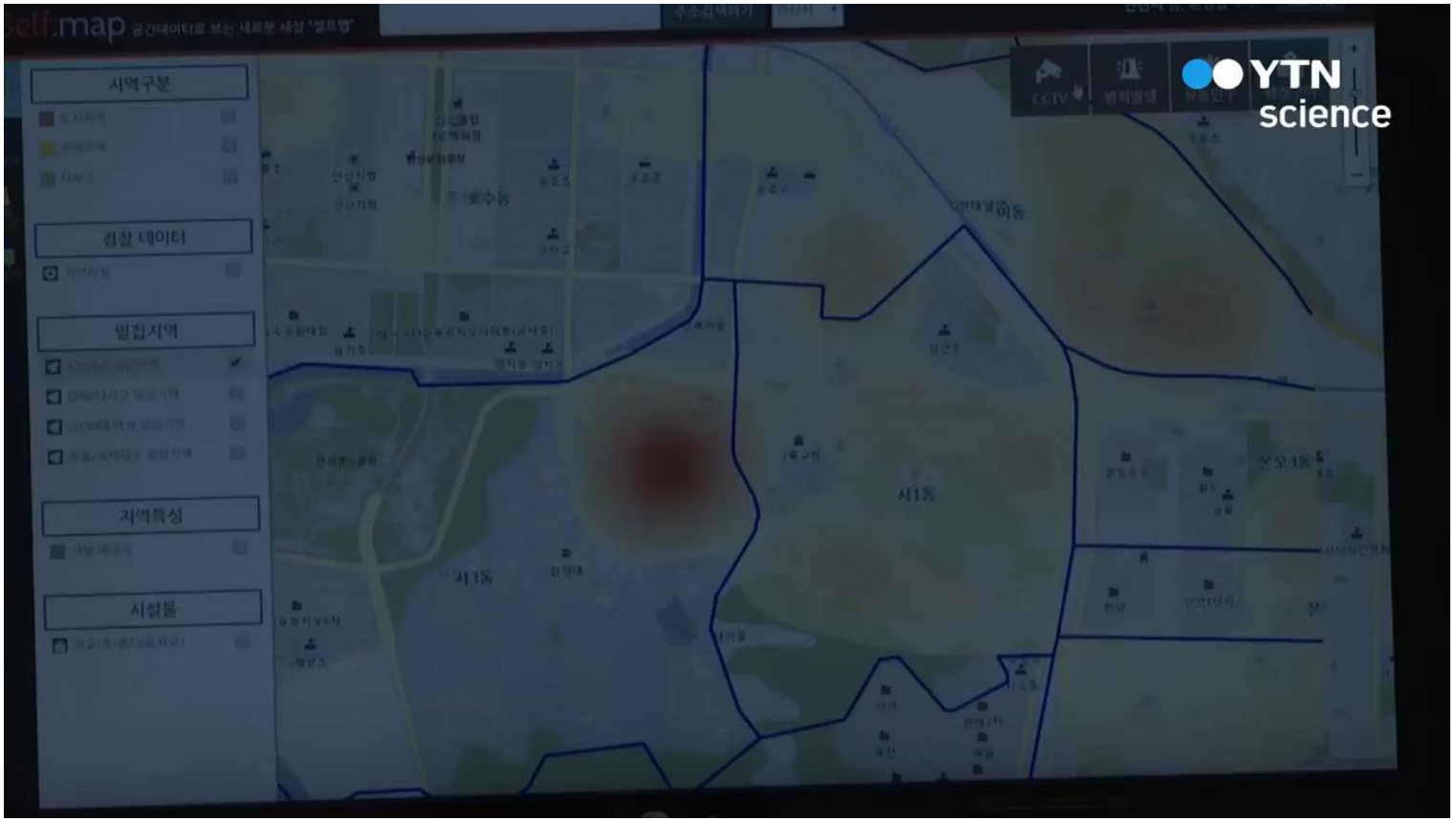
올빼미버스

https://www.youtube.com/watch?v=kJNKIU5fgUU 2분25초

빅데이터 활용 사례4

CCTV와 범죄

한정적인 자원(CCTV)을 어디에 설치하는 것이 최대 효과를 낼 수 있을까.



https://www.youtube.com/watch?v=JZUPnI-mkV8 1분22초

빅데이터 활용 사례5

미국 프로그레시브 보험 'Pay as you drive' 보험 구조

프로그레시브의 'Pay as you drive'

STEP1 — Plug

Day 1

보험 계약자는 보험사가 제공하는 운행정보 측정기기인 스냅샷(Snapshot)을 차량에 설치

STEP2 — Drive

Day 1-30

측정기기는 운전 빈도, 속도, 주행거리, 운전시간대, 급브레이크 횟수 등의 데이터를 수집해 보험사에 전송

STEP3 — Save

Day 31

보험사는 측정기기 설치 후 첫 한 달분의 데이터를 분석해 임시 할인율 결정

최종 요금의 할인 금액은 계약 6개월 이후 결정되며, 최대 30%까지 할인

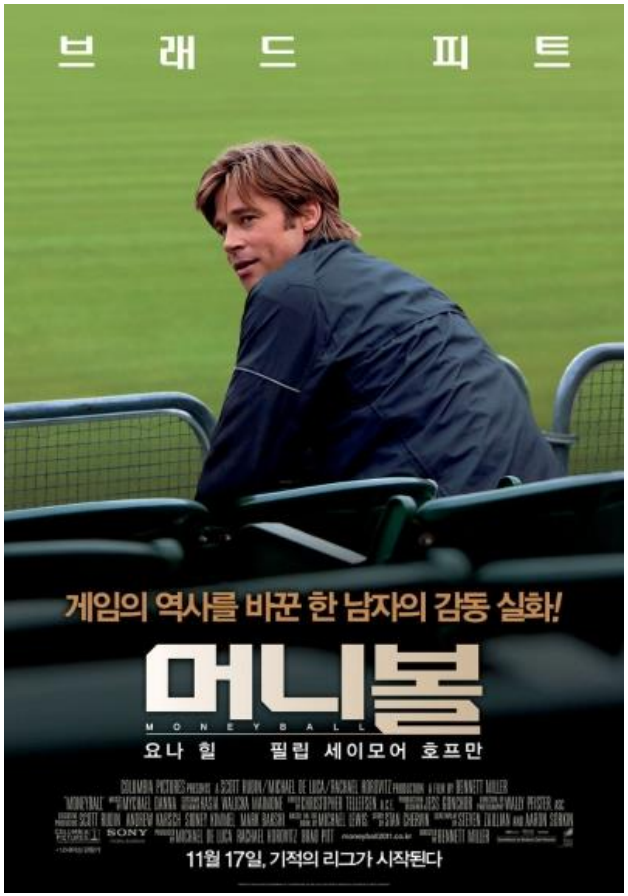
계약자는 자신의 운전 이력과 운전습관 등의 데이터를 인터넷으로 확인하고 점검

〈출처: 프로그레시브 홈페이지 발췌 및 재구성〉

빅데이터 활용 사례6

MLB(Major League Baseball)의 머니볼 이론 및 데이터 야구

- 경기 데이터를 철저하게 분석해 오직 데이터를 기반으로 적재적소에 선수들을 배치해 승률을 높인다는 게임 이론
- 미국 메이저 리그 베이스볼 ”오클랜드 어슬레틱스”의 구단장 ”빌리 빈”이 리그 전체 25위에 해당하는 낮은 구단 지원금 속에서도 최소비용으로 최대효과를 거둔 상황에서 유래
- 최하위 팀을 4년 연속 포스트시즌에 진출시키고 메이저 리그 최초로 20연승의 신기록을 세움.
- 타율, 타점, 홈런 등 흥행 요소만을 중시하던 야구계에서 출루율, 장타율, 사사구 비율이 승부와 관련되어 있음을 간파하고 데이터를 수집, 분석, 활용



https://www.youtube.com/watch?v=ElVkO2m92Gk&t=24s

빅데이터 활용 사례7

스마트 길, 빅데이터 공유

한국도로공사는 최첨단 ICT센터를 구축하고 교통데이터를 생산 수집. 활성화를 위해 "오픈 오아시스" 구축. 교통 빅데이터를 국민과 공유하게 됨. (교통상황 시각화, 교통사고 정보 등)

한국도로공사

https://www.youtube.com/watch?v=O11pKesEXWE 2분20초

빅데이터 활용 사례8

빅데이터가 만든 대박상품

빅데이터가 만든 대박상품, 대왕 요구르트의 등장
- 소비자를 관찰하고 데이터의 패턴을 찾아내라
- 소비자가 원하는 상품 출시로 기업 매출 확대

https://www.youtube.com/watch?v=Spp39BgpsNk 1분30초

빅데이터 활용 사례

	수행기관	프로젝트 명	주요 내용
해외	미국 국립보건원	유전자 데이터 공유를 통한 질병치료체계 마련	75개기업과 제휴하여, 200TB의 유전자정보 수집,일반에 공개. 유전자 비교분석 서비스 제공, DNA 이상에 따른 질병 사전예측 및 대응
	미국 국립보건원	Pillbox 프로젝트를 통한 의료 개혁	약검색 서비스를 통해 지역별 질병통계 분석 이 정보를 기반으로 보건정책 수립 대응
	미국 퇴역군인국	미국 퇴역군인 전자의료기록 분석을 통한 맞춤형 의료서비스 지원	2년간 25개 DW 구축, 전자의료기록(EHR) DB구축 의료서비스에 제공
	캐나다 온타리오 공과대 병원	미숙아 모니터링을 통한 감염 예방 및 예측	미숙아 1명이 일9,000만건 데이터 생성 (바이털 사인을 초당 1,000번 수집) 이상징후를 통해 6~24시간 먼저 감염 확인
	건강보험회사 웰포인트	슈퍼컴퓨터를 활용한 효율적 환자치료	환자 증상, 면담기록등 모든 내역을 저장, 환자치료 가이드라인 제시 기능. 2억 페이지 검색을 3초 내 실시
	구글	검색어 분석을 통한 독감예보 서비스 제공	검색어 쿼리를 조사하여, 지역별 독감동향을 신속히 감지함.
국내	한국인체자원은행 네트워크	정보공유로 생명공학 분야 경쟁력 제고	전국 16개 병원에서 36만명 인체 정보 획득. 특정질환별로 연구자들에게 무료 제공
	DNA Link	유전자 분석시스템으로 맞춤형 건강검진 서비스 제공	4만명 이상 질병관련 분석을 하여 국내 최대 한국인 유전체 DB 구축 1 TB 이상 DB 구축, 기하급수적으로 늘어나는 DNA 데이터 저장 분석 가능
	연세대학교 의료원	후(HooH) 헬스케어시스템	전자진료기록부, 의료영상 전송 등에 대한 전산시스템을 클라우드 기반으로 이전 빅데이터 인프라를 적용하여 방대한 데이터 축적

03. 빅데이터 분석 툴

소셜미디어 분석 툴1

소셜 매트릭스

다음소프트의 무료 제공 툴 http://insight.some.co.kr/

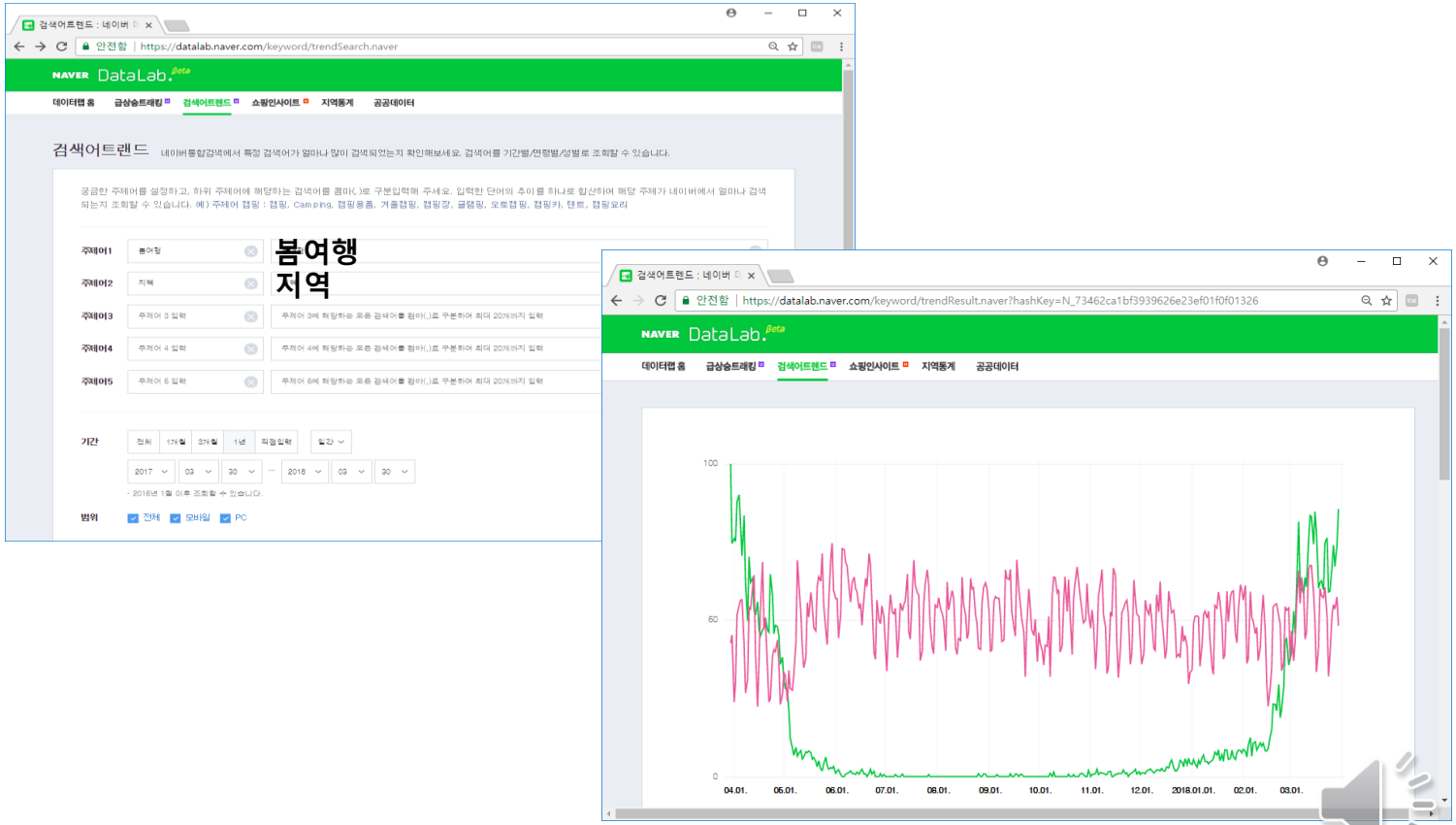
→ 주어진 탐색어를 가지고 트위터와 네이버 블로그를 검색한 후 탐색어와 연관되어 쓰는 단어들을 조사해 **탐색어 맵**을 만든다.

소셜미디어 분석 툴2

네이버 트렌드

네이버의 무료 제공 툴 https://datalab.naver.com/

"**검색어 트랜드**"는 주어진 탐색어를 가지고 단어들을 조사해 **빈도**를 계산한다. "주제어", "기간", "범위", "성별", "연령"을 선택할 수 있으며, 연관검색어 통계는 보여주지 않는다.

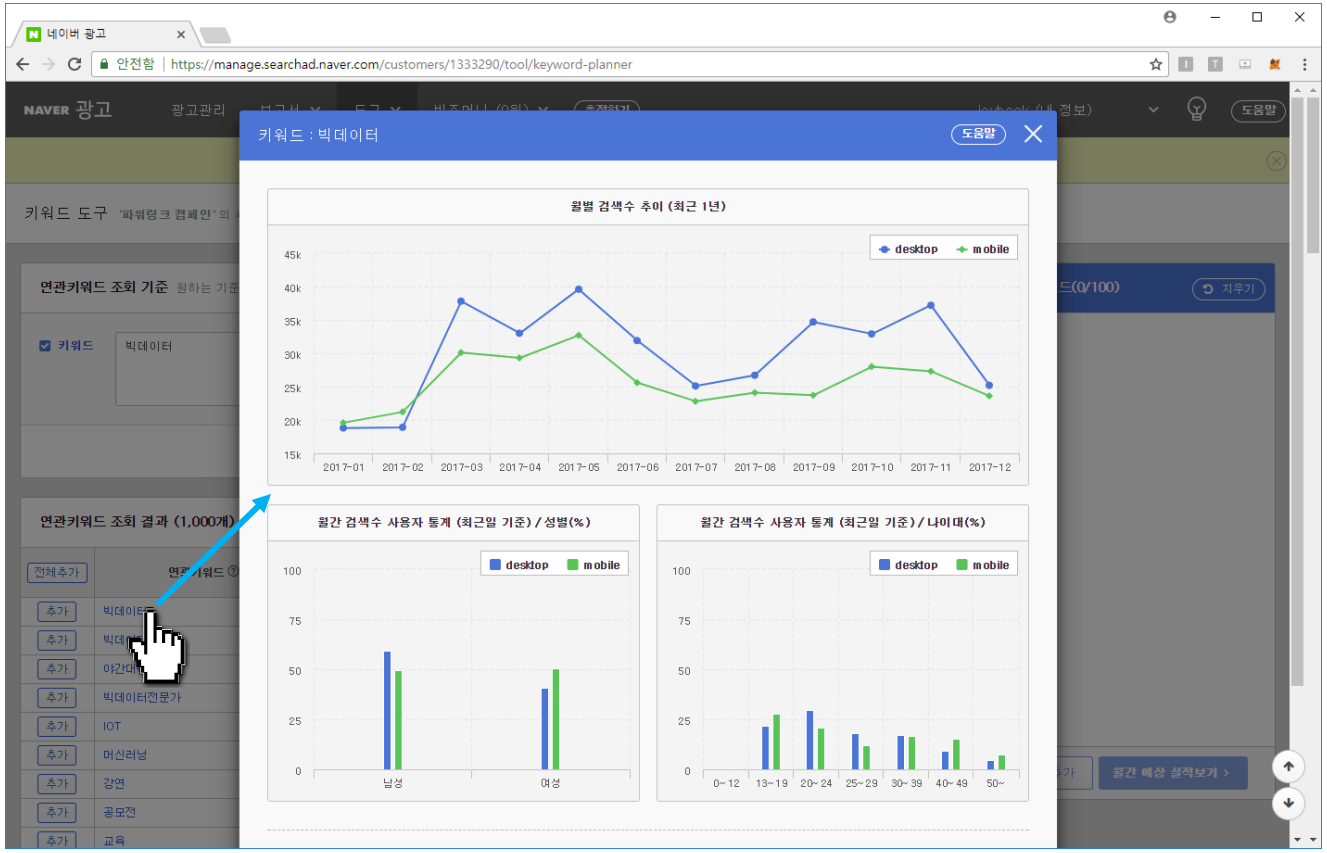


소셜미디어 분석 툴3

네이버 광고

네이버의 광고주를 위한 무료 제공 툴 https://searchad.naver.com/

"광고시스템"에서 [도구]-[키워드 도구] 에서 '검색어'를 입력하여 최근 4주간의 월별추이와 성별추이, 검색연령의 연관 키워드 **빈도**를 보여준다.



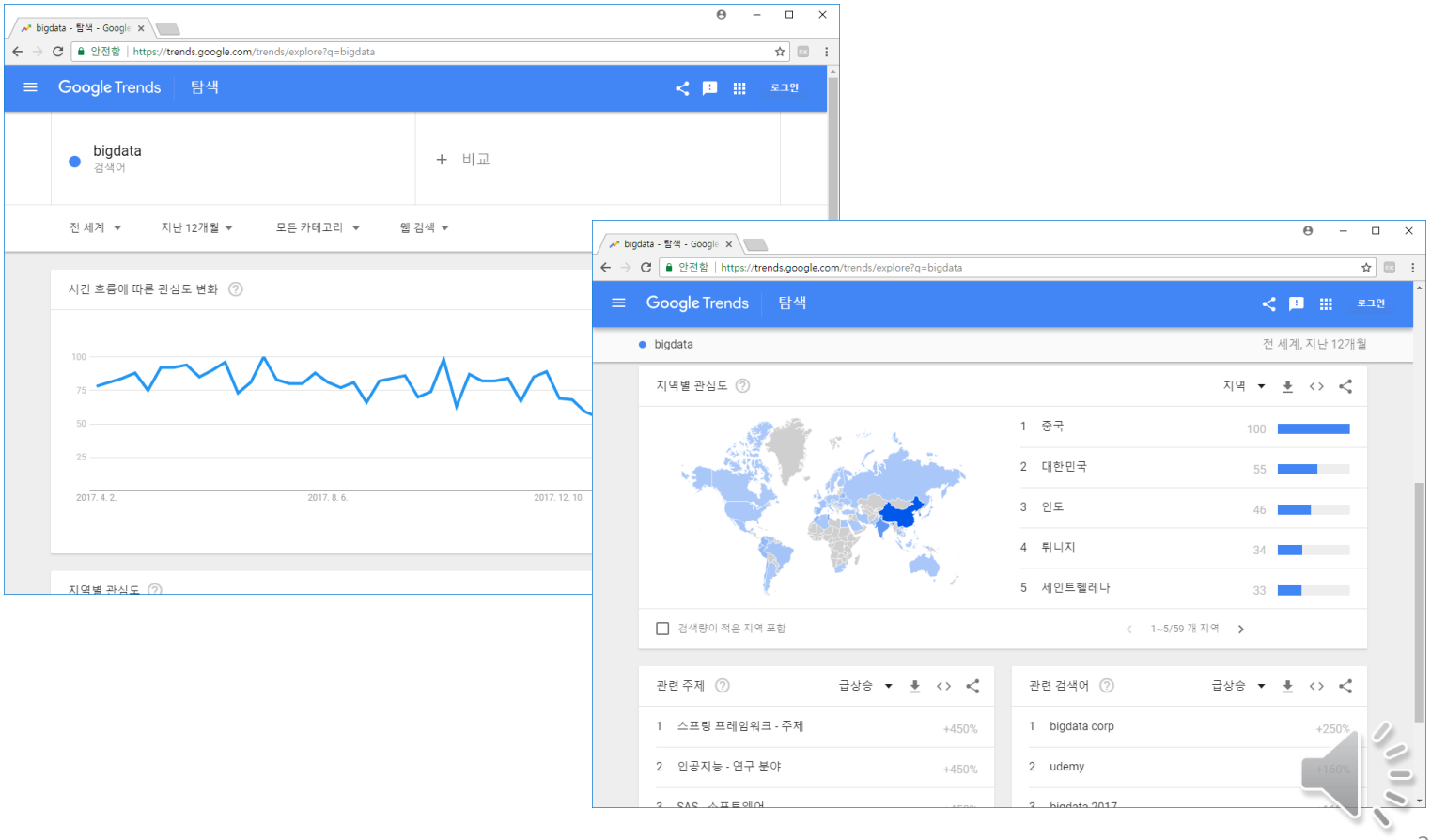
로그인 필요

소셜미디어 분석 툴4

구글 트렌드

구글의 무료 제공 툴 https://trends.google.com/trends/

주어진 탐색어를 가지고 단어들을 조사해 **빈도, 연관검색어** 통계, 지역별 관심도와 관련주제도 보여준다.



04. 주요 이슈 및 문제점

주요 이슈

사람들에게 무엇이 필요한가?

대량의 다양한 데이터로부터 감춰진 정보를 찾아냄

우수한 데이터 과학자(Data Scientist)가 필요함

주요 이슈

- 데이터 과학자 ➔ 미래 핵심 인재

https://www.youtube.com/watch?v=dZZfDj_ieEU 5분

문제점

1. 사생활침해와 개인정보 보호 → 악용사례 급증
2. 빅데이터가 만든 '빅브라더' → 구글 페북은 모든걸 알고 있다.

   빅브라더 : 정보의 독점으로 사회를 통제하는 관리권력, 혹은 그러한 사회체계를 일컫는 말.

3. 데이터 소유권과 저작권 분쟁

   개인의 리포트, 사진, 동영상과 같은 저작물이 공유, 배포되어 수익구조가 형성된 경우에 이 수익을 누가 가져야 할 것인가. 사진의 경우, 필름을 넘겨준 경우 소유권은 의뢰를 한 사람에게 있지만, 작품 사진에 대한 저작권은 직접 사진을 찍은 사진사에게 있다.

4. 빅데이터 분석의 오류

   데이터 분석에 사용되는 기법중 통계학에 기초한 분석은 평균치의 함정에 빠져 심각한 오류를 가져올 수 있다. → 전문성 요구되며 세심한 주의가 필요하다.

Thank you