빅데이터 시대의 조망과 인재 양성

KISTEP 수요 포럼

2016년 10월 5일

박성현
서울대 통계학과 명예교수
전 한국과학기술한림원 원장

목 차

1. 인류사회의 변천과정

인류는 3번의 물결(미래학자 Alvin Toffler)을 거침

□ 1차 물결: 농경사회 (8000 B.C. 경): 독립시대

□ 2차 물결: 산업사회 (18-19 세기): 산업혁명, 경쟁시대

□ 3차 물결: 정보화 사회 (20 세기 중반 이후):

- 원천정보의 생성자가 리더십 발휘
- 보유하고 있는 정보의 양과 활용 정도에 따라 국력을 평가
- 통계학, 정보과학, 전산과학 등은 정보화 사회를 이끄는 학문

제4의 물결은? –21세기의 지식사회

지식사회(knowledge society)?

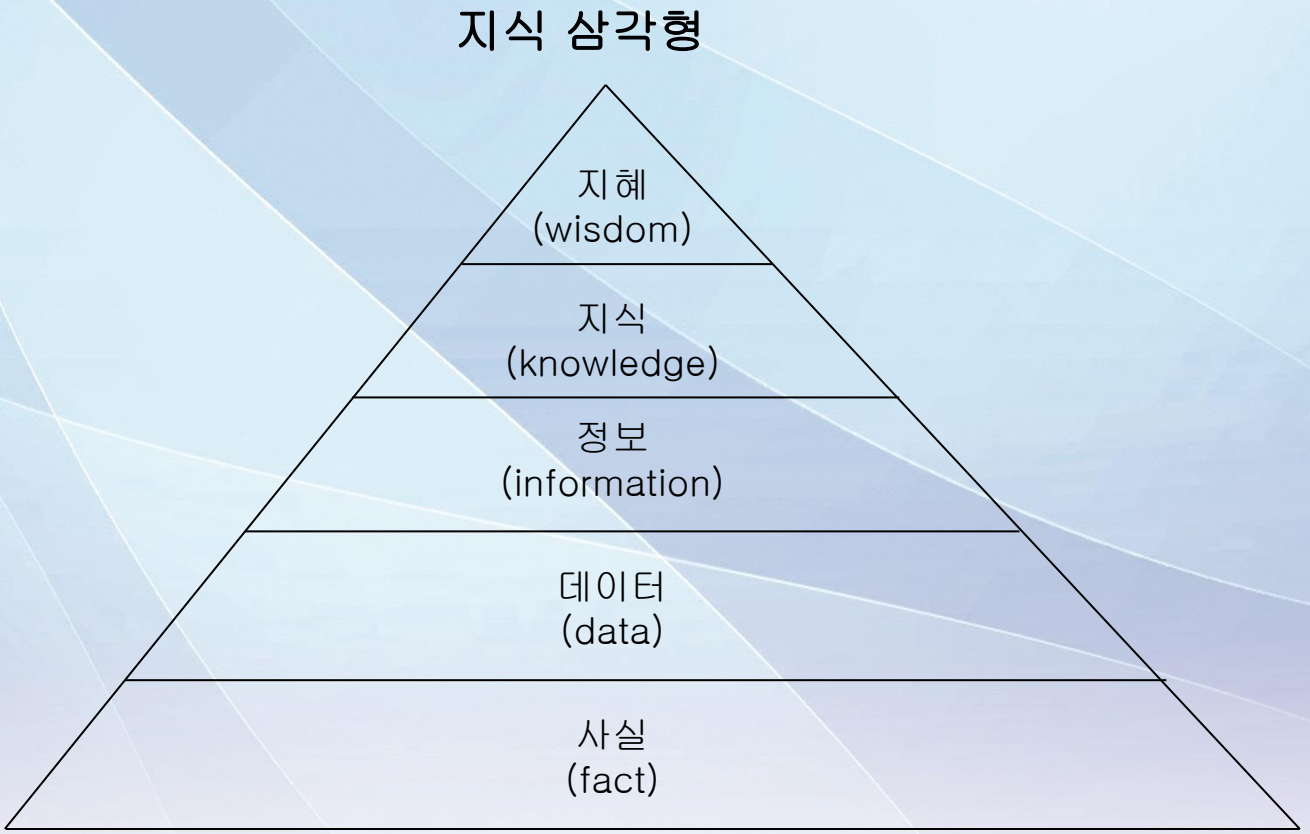
- 지식이란 정보를 집적하고 체계화하여 어떤 유용성을 갖도록 한 것으로, 의사결정의 기준을 제공하는 집약된 인식
- 21세기는 지식경쟁력을 가진 나라가 선진국이며, 지식창출 기반을 조성하는 빅데이터의 역할이 매우 중요해짐
- 지식사회에서는 과학기술과 지식 창출이 국가경쟁력을 결정하는 가장 중요한 요소

인류 과학기술 발전사

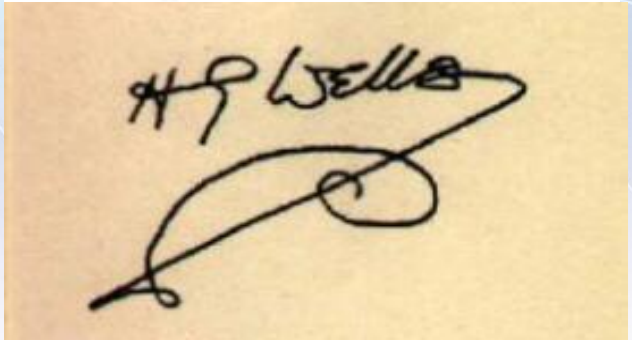
- 인류의 뛰어난 창의력은 놀라운 국부창출을 가져옴
- 산업혁명 이전까지 인류는 가난한 삶

: 과학기술 진보가 인간의 풍요 · 건강한 삶 창출

현재 우리는 정보화사회와 지식사회의 중간에 있으며,
향후 완전한 지식사회(제4차 산업혁명)로 갈 것으로
예측됨. 빅데이터가 중요한 역할을 할 것임.



“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write”.(1866-1946)



2. 통계학의 발전과 데이터 기술

통계학(統計學, Statistics)의 어원

- Latin어의 Status(State의 의미)가 root word이며,
- Political Arithmetic(정치 산술)의 의미로 시작된 학문
- Statistics란 단어는 Encyclopedia Britannica (1797)에 처음 등장
  초기 통계학에서는 인구조사가 중요

CENSUS(인구조사)

Latin어의 Censere가 root word이며, taxation의 의미.

B.C.3000경, Babylonia, Egypt, China에서 이미 시행.

B.C. 1440경, 구약 민수기 1장, 모세가 Sinai광야에서 이스라엘 백성의 Census 실시. 20세 이상으로 싸움에 나갈 수 있는 남자 (50세 이하), 12지파 603,550명을 계수

현대 통계학(Modern Statistics)이란?

- 통계학은 사회, 자연 및 인간생활의 온갖 현상을 연구하기 위하여, 불확실성(uncertainty)이 내포된 데이터(data)의 선택, 관찰, 분석, 추정 및 검정을 통하여, 의사결정(decision-making)에 필요한 정보(information)의 획득과 처리 방법을 연구하는 학문. (20세기 초에 영국을 중심으로 발전됨)

- 20세기 중반에 미국으로 이동. 최근에는 독일, 프랑스, 일본, 인도, 한국 등에서도 많은 연구가 이루어지고 있음.

21세기 지식기반 정보화 시대에 사용 가능한 통계학의 다른 이름

- Data Information Science
- Data Science
- Data Technology
- Decision-making Science
- Statistical Information Science
- Informative Statistical Science
- 기타

DT(데이터 기술) 이란

- 데이터 기술이란 데이터의 측정, 수집, 저장, 검색 기술에서부터 시작하여, 데이터의 분석 및 해석 능력, 데이터로부터의 모형화 기술, 데이터로부터의 진단, 관리 및 예측 기술을 다루는 과학적 방법론.

- - 박성현: “데이터기술의 경제학”, 한국경제신문 다산 칼럼, 2001.12.3.

- Park, S. H. and Suh, M. W. (2008): Data technology as a new discipline for broader application of statistics, Journal of Data Science, 6(3), p. 357-368.
- Erto, P., Pallota, G. and Park, S. H. (2008): An example of data technology product: a control chart for Weibull processes, International Statistical Review, 76(2), p. 157-166.

DT를 6T에 추가해야

과학기술의 발전과 경제사회 변혁을 주도할 미래 유망 신기술

IT: 정보기술
NT: 나노기술
ET: 환경기술
BT: 생명공학기술
ST: 우주항공기술
CT: 문화기술

여기에 **DT: 데이터 기술** 이 추가되어야

IT와 DT의 차이점

	IT	DT
관련된 주 학문	컴퓨터 공학, 전자공학, 통신공학, 제어계측공학, 정보공학 등	수학, 통계학, 산업공학, 정보과학, 전산과학, 경영과학 등
주요 제품	통신장비, 전자장비 등 하드웨어	DBMS, CRM, SPC, 암호시스템, Data-mining, 통계패키지, 시뮬레이터 등의 각종 소프트웨어
주요 특징	정보(글, 그림, 소리 등)를 전달하는 공학적인 기술	다량의 데이터로부터 현상을 파악 하고, 효율성을 극대화하며, 미래를 예측하는 과학
우리나라의 수준	상	중 하

DT 는 국가경쟁력의 요소로 중요성이 날로 증대되고 있음

DT를 배경으로 한 통계학의 미래

통계학은 DT를 중심으로 발전할 것이며, DT에서는 DMAMP (Define, Measure, Analyze, Model, Predict)이라는 문제해결 사이클을 사용할 것으로 예측됨.

DT에서는 빅데이터, 6T (IT, BT, CT, ST, NT, ET)등의 분야에 통계적 분석 방법론을 Fusion 시켜 현재를 분석관리하고, 미래를 예측하고 대비하는 기술을 발전시킬 것임.

결론: DT를 중심으로 하는 통계학의 발전은 국가운영, 기업운영, 개인의 생활에 중요한 Infra 역할을 할 것이며 통계학은 미래학문으로 중요한 위치를 점할 것임.

3. 데이터 폭발과 빅데이터 시대

- McKinsey Global Institute의 보고:
  매달 300억 개의 콘텐츠가 페이스북에서 유통
  1분마다 24시간 분량의 동영상이 유튜브에 올라옴
  매월 트위터에는 1억1천만 개의 정보가 트윗됨
  월마트에는 시간당 100만개의 거래정보가 축적됨

- IDC(Internet Data Center):
  2012년 생성된 데이터 양은2.8ZB(zettabyte)
  * 1 ZB = 1조 기가바이트(GB)
  = 400만개의 미국의회도서관(장서보유 1억 4,200만권)에 해당
  생성되는 데이터 양은 2년에 2배씩 증가
  데이터는 CCTV 정보, 실시간 센서 정보, 지리 정보, 각종 SNS 정보 등도 포함됨
  * 1 GB = 1,000 MB, 1MB = 1,000,000 Bytes

빅데이터란?

- 빅데이터란 “기존의 DB가 저장, 관리, 분석할 수 있는 범위를 벗어나고 비정형 데이터를 포함한 큰 규모의 데이터 집합(또는 그 분석 기술 포함)”

- 데이터의 분류
  - 정형(구조적, structured) 데이터:
    정해진 서식에 따라 특정 형식에 맞춰 잘 구조화되어 관리되는 데이터
  - 비정형(비구조적, unstructured) 데이터:
    일정한 형식이 없는 SNS 데이터, 뉴스 게시물, 블로그, 태그 정보, 온라인 커뮤니티 게시판의 게시물, 유투브 동영상, 음악, 사진 등.

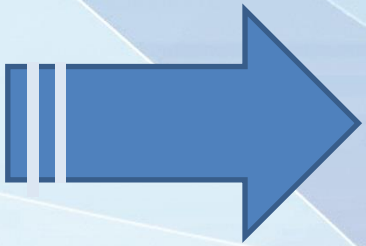
빅데이터 개념의 발전과정

- 1970년대와 그 이전: Data Analysis (DA)
- 1980년대: Data Base (DB) Management
- 1990년대: Data Mining (DM), Knowledge Discovery in Database (KDD)
- 2000 년대: Business analytics, Bioinformatics
- 2010년대: Big Data (Analytics)

빅데이터의 핵심 역량은 무엇일까?

빅데이터의 필요 핵심 역량은 최소한 IT, 분석, 업무 전문 역량 등 3가지가 있다

빅데이터
구현/분석/활용에
필요한 역량



IT 역량 — 구현

통계 분석 역량 — 분석

업무에 대한
전문 역량 — 활용

빅데이터에 대한 평가

- 유명한 과학저널 <네이쳐(Nature)>는 2008년 9월호에서 “인터넷 이후 기업에 가장 큰 영향을 미칠 것으로 기대되는 것이 빅데이터”
- 다보스 세계경제포럼은 2012년 “올해 가장 주목해야 할 과학기술 1위는 빅데이터”를 꼽음
- 유래미래보고서는 2030년 10대 메가트렌드의 하나로 “진정한 빅데이터 시대의 도래”를 선정(2012년 발행)
- 최근 우리 정부도 계속하여 국가 성장동력 중의 하나로 빅데이터를 지목

빅데이터의 적용 사례 (예)

1. 오바마를 당선시킨 빅데이터 활용
   – 두 번의 선거에서 1억 명이 넘는 유권자의 성향 분석

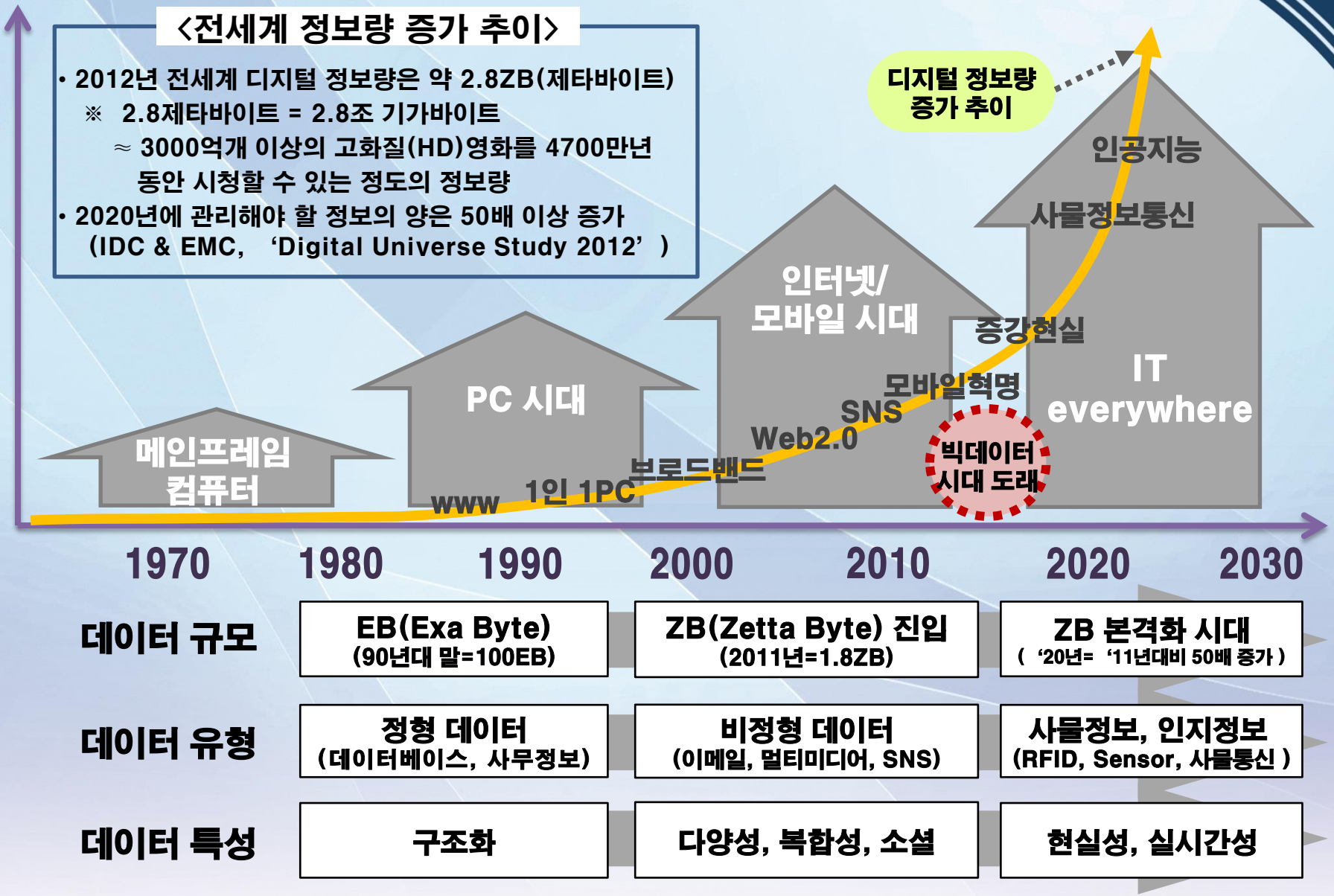
2. 서울 시장 선거에서의 예측
   – SNS 분석을 통하여 인기도 조사 (SAS 회사 사례)

3. 신용카드사의 고객맞춤형 홍보 전력
   – <한국경제신문 기사>
   최근 아이를 출산한 김씨는 L카드사로부터 아기와 유아용품으로 만 채워진 전단지를 받음(빅데이터 분석에 의하여 임부복이나 튼살 크림을 구매한 고객 가운데 산부인과나 산후조리원 결제 실적이 있는 회원을 별도 구분)

빅데이터의 영향

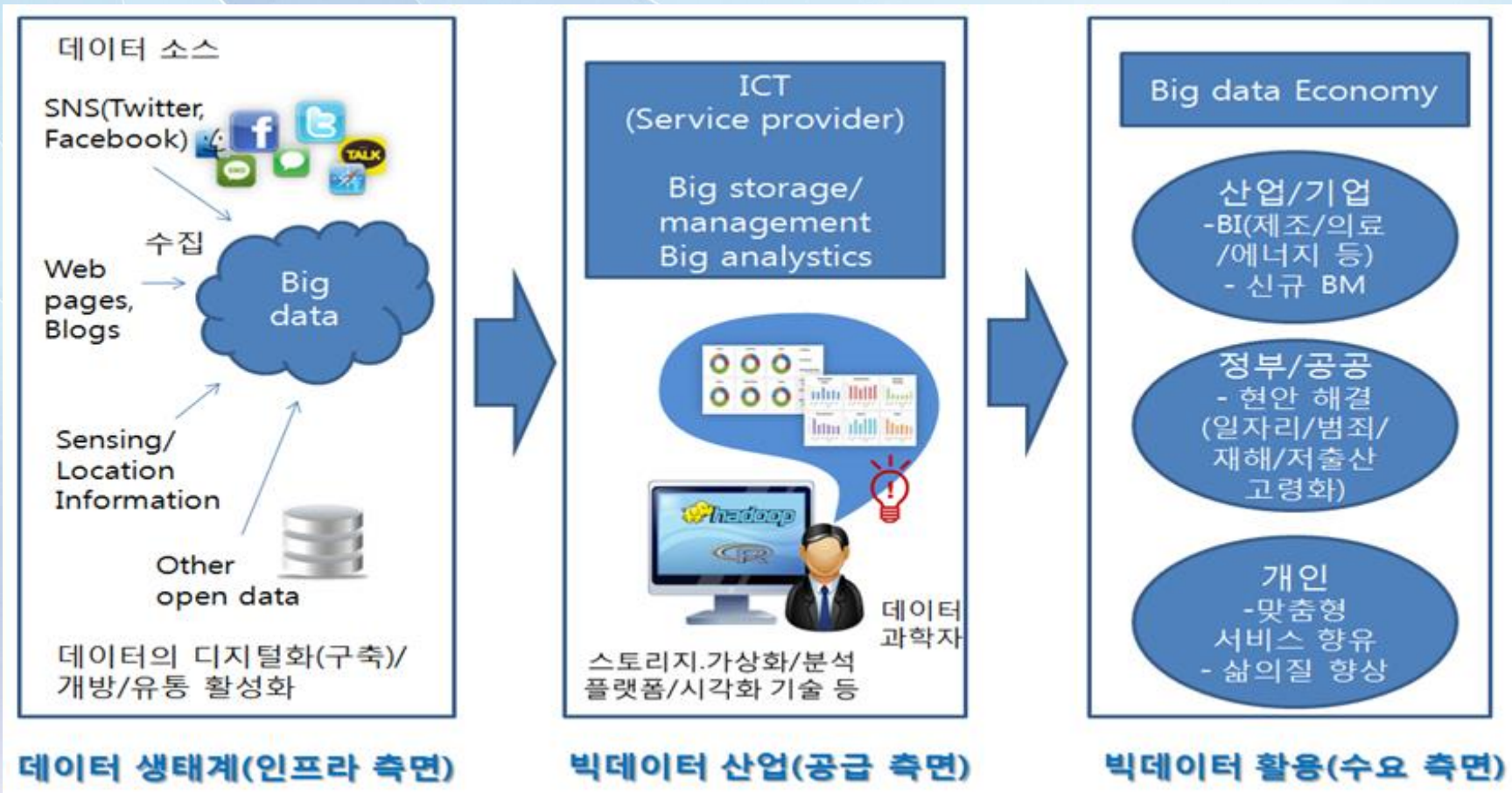
1. 제4차 산업혁명의 기반으로 차세대 지식사회의 총아 역할
   (IoT, IoE, 자율주행 자동차, 인공지능, 스마트공장 등의 기초 기술)
2. 고객세분화를 통한 고객만족경영 (CRM(고객관계경영)포함)과 판매 촉진
3. 제품/서비스 개발, 품질경영의 고도화
4. 데이터 개방을 촉진하는 정부3.0이 성공적으로 시행되면 공공행정의 선진화, 투명성 제고에 기여
5. 선거운동, 사회복지 활동에 활용
6. SNS 등을 통한 국민 여론을 실시간으로 탐색하여 국정에 반영

4. 빅데이터 시대에 필요한 데이터 과학자



빅데이터의 의의

"빅데이터는 단순한 산업이 아니라 인터넷처럼 경제사회 전반에서 혁신을 주도하는 일종의 '플랫폼'(GPT : General Purpose Tech)"



데이터 사이언티스트(Data scientist)

– 데이터 사이언티스트란 "데이터 분석을 통하여 유용한 정보를 추출하는 전문가이며, 고객의 행동이나 시장 주기 같은 구조화되지 않은 데이터의 숨겨진 패턴을 찾아냄으로써 새로운 기회를 창출하는 업무를 수행하는 사람" 이다. (정보통신용어사전)

– 데이터 사이언티스트가 갖추어야 할 지식:
  • IT 분야: DB 관리, 프로그래밍, 빅데이터 처리 소프트웨어 이해 (예: Hadoop)
  • 통계 분야: 데이터 분석기법, 선형 모형, 다변량 통계, R, 통계 패키지 등
  • 경영, 산업공학 분야: 기업경영, 인지공학, 품질경영, 소비자와의 커뮤니케이션 스킬 등

– 기업에서는 전산, 통계분석, 품질전문가 등을 데이터 사이언티스트로 키우는 것이 현명. 미국에서는 가장 인기 있는 직종으로 부상

빅데이터 분석 소프트웨어의 출현

Google은 데이터파일의 효율적인 관리를 위하여 Google File System을 개발하고, 이를 바탕으로 MapReduce라는 모델을 개발(특정용어 등의 출현빈도를 세거나 다양한 비구조적 데이터를 처리)

아파치 소프트웨어재단(ASF)은 Yahoo의 지원을 받아 MapReduce을 기반으로 하는 오픈 소스 프로그램 Hadoop을 개발.

현재 이를 이용한 패키지화에 여러 기업이 참여하고 있음(IBM, SAS, Oracle, EMC, Cloudera, SAP 등)

주요 빅데이터 분석 기술

- 텍스트 마이닝 (Text mining)
- 오피니언 마이닝 (Opinion mining)
- 소셜네트워크 분석 (Social network analysis)
- 군집분석 (Cluster analysis)
- 다변량분석 (Multivariate analysis)
- 동적 그래픽스 (Dynamic graphics)
- 데이터 시각화 (Data visualization)
- 기타 통계분석 (회귀분석, 범주형 자료분석, 예측 모델링 등)

제4차 산업혁명과 빅데이터

혁명	시기	주요 특징
제1차 산업혁명	18세기 말	영국: 증기기관, 도시 공장, 공작기계
제2차 산업혁명	20세기 초	미국: 테일러 시스템, 대량생산, 자동화 표준관리, 경영학, 통계학 발전
제3차 산업혁명	20세기 중반	미국: 컴퓨터, ICT 발전, 데이터 과학
제4차 산업혁명	21세기	(?): 인터넷, 빅데이터, 사물인터넷, Smart Factory, Smart Farming, 지능형 로봇, 융합과학기술, 소프트웨어 중심 지식 사회

** 제4차 산업혁명은 소프트웨어와 데이터 기반의 지능디지털기술변환 (intelligent digital technology transformation)에 의한 혁명으로 우리 사회를 크게 변화시킬 것임.

제4차 산업혁명이 일자리와 교육에 주는 영향

1. 일자리에 주는 충격
 - 다보스포럼(2016.1.18)에서 발표한 '미래고용보고서'에 의하면 세계 고용의 65%를 차지하는 주요 15개국에서 2020년까지 710만개의 일자리가 사라지고, 200만개의 새 일자리가 생겨, 510만개의 일자리가 순 감소.
 - 사무관리직에서 큰 감소. 반면에 STEM 분야와 경영, 금융 전문서비스 분야에서 증가.

2. 교육에 주는 충격
 - 일자리에 맞추어 인재 양성은 불가피.
 - 다보스포럼의 '미래고용보고서'는 교육목표로 1-5위를 '복잡한 문제를 푸는 능력', '비판적 사고', '창의력', '사람관리', '협업능력'을 지적함.

5. 미국의 데이터 과학자 양성 프로그램

Data source: Il-Yeol Song and Yongjun Zhu (2015); "Big data and data science; what should we teach?", Expert Systems, Version of Record Online; 9. Oct. 2015, Wiley Publishing Ltd.

Four categories of data science program in USA

1. bachelor's program – not very common yet
2. master's program – most popular
3. certificate program – graduate level, less than one year, mostly offered on-line such as Columbia University.
4. specification/concentration in doctoral program – rarest (only one in Univ. of Washington)

Bachelor and master programs (total 42)

Bachelor's programs

College/school/department	Number of programs
University/joint departments	3
Computer science	3
Data science	2
Business	1

Total 9 universities

Bachelor and master programs (as of 2014, total 42 universities)

Master's programs

College/school/department	Number of programs
University/joint departments	17
Information science	7
Computer science	3
Statistics	3
Information technology	1
Operations research	1
Professional studies	1

Core courses in master programs of 15 universities (that show curricula on their websites)

Course	Number of universities offering the course
Exploratory data analysis	10
Database	10
Data mining	9
Data visualization	8
Statistical modelling	8
Machine learning	6
Information retrieval	5

Course	Number of universities offering the course
Information and social network analysis	4
Data warehouse	4
Introduction to data science	3
Research methods	3
Social aspects of data science	3
Algorithms	2
Data cleaning	2
Text mining	2
Healthcare analytics	2

6. 한국에서의 데이터 과학자 양성 방안

현재의 상황:

국내는 2014년부터 데이터사이언스 학과를 일반대학원에 석(박)사 과정으로 설립하기 시작.

- 국민대학교 일반대학원(경영대) 데이터사이언스 학과 2014년 설립 (국내 최초 석박사 과정)
- 단국대학교 일반대학원(공대) 데이터사이언스학과(석사과정) 2015년
- 성균관대학교(문과대학) 문헌정보학과 데이터사이언스 전공(석사과정)으로 시작(2015년), 데이터사이언스학과로 독립할 예정.
- 건국대학교 일반대학원(상경대) 데이터사이언스학과(석사과정) 2016년
- 서울과학기술대학교 일반대학원 데이터사이언스학과(석사과정) 2016년
- 세종대학교 SW 융합대학(신설) 데이터사이언스학과 설립예정
- 기타

사례1: 국민대학교 일반대학원 데이터 사이언스 학과 커리큘럼

이수영역		과목명
공통		데이터 사이언스 개론, 경영통계, 데이터베이스관리론, R프로그래밍
전공	경영학:	경영정보시스템, 오퍼레이션스애널리틱스, 마케팅애널리틱스, 파이낸스애널리틱스, 경영최적화와 시뮬레이션
	통계학:	통계자료처리론, 선형통계분석론, 다변량통계분석론, 데이터마이닝
	데이터 과학:	빅데이터분산처리론, 빅데이터통합과 모델링, 소셜네트워크분석, 텍스트마이닝과 소셜애널리틱스, EDA와 빅데이터시각화, 빅데이터프로젝트

* 경영, 통계, 전산과학 균형적인 커리큘럼

사례2: 단국대학교 공과대학 대학원 데이터사이언스 학과 커리큘럼

이수영역	과목명
공통	경영의사결정론, 선형통계분석 , 경영정보시스템
전공필수	빅데이터 처리
전공선택	기계학습, 데이터 마이닝, SAP 프로젝트 1,2 데이터분석 사례연구, 빅데이터 마케팅, R 프로그래밍, 데이터분석 및 기획, 데이터분석 프로젝트, 데이터베이스 관리, 경영 최적화 이론, 이미지 처리와 인식, 인공지능 응용
연구	세미나 1,2,3

* 전산과학과 프로젝트 중심의 커리큘럼

대학에서의 데이터 사이언티스트 양성 방안

(1) 데이터사이언스 학과를 일반대학원에 설치하여 석사(박사)과정으로 운영하고, 관련 있는 모든 학과들(경영대, 전산과학과, 통계학과, 산업공학과 등)이 협력하여 데이터사이언스 학과를 공동 운영하는 것이 좋다.
운영 예: 학과장은 관련 학과들이 돌아가면서 담당하고, 대학원장에게 직접 보고하는 형태가 바람직함. 커리큘럼의 균형 잡힌 운영을 위하여 단과대학에 예속되는 것은 피하는 것이 좋음.

(2) 학과의 성격상 실제의 프로젝트를 수행하도록 권유할 필요가 있으며, 산학협력을 강조할 필요가 있음. 산학협력 보고서로 학위논문을 대체.

(3) 이 학과는 융합학문이므로, 기초적인 과목들(통계, 전산, 경영, 산업공학)을 공부할 필요가 있으며, 졸업필수 학점을 최소 10개 과목(30학점)으로 하는 것도 바람직함.

(4) 초기에는 장학금 등을 지급하여 우수한 학생들이 참여하도록 유도.
참여하는 교수들에게 한국연구재단에서 '융합과학기술' 장려의 측면에서 연구비 지급이 필요함.

기업에서의 빅데이터 인재 활용 방안

(1) 현재에는 데이터 사이언티스트를 구할 수 없으므로, 기존의 전산요원, 통계분석 요원(또는 Six Sigma 전문가), 품질관리기사 등을 활용할 필요가 있음.

(2) Six Sigma 경영을 하면서 양성된 BB (black belt), MBB (master black belt) 등이 있으면 이들에게 부족한 부문(전산, 경영 등)을 교육하여 임시 데이터 사이언티스트로 활용할 수 있음.

(3) 빅데이터를 도입하면서 팀활동을 통한 우수한 사례를 만들 필요가 있으며, 이는 기업에 확산하는데 도움이 될 것임. 팀활동이 중요하며, 사장은 이런 팀활동에 무게를 실어주어야 함.

(4) 기업에서 필요한 양질의 데이터를 확보하기 위하여, 데이터 확보 시스템을 만들고, 이를 다루는 전문요원을 확보할 필요가 있음.

7. 미래를 위한 교육개혁 방향

빅데이터를 포함하여 제4차 산업혁명에 대비한 우리의 교육개혁 방향을 다음과 같이 하는 것이 바람직함.

(1) 프로젝트 학습과 심층학습 중심의 교사(교수) 학습 방식 장려

(2) 선다형 평가방식의 혁신과 대학의 입학생 선발방식 혁신

(3) 초.중.고 교육에서 공교육 역량의 강화

(4) 각 대학에 데이터 사이언티스트를 양성하는 프로그램의 도입

(5) 소프트웨어 교육의 강화와, 컴퓨팅적 사고와 통계적 사고를 함양하는 교육 프로그램 장려.

- The best way to predict the future is to create it.

  (Peter Drucker)

- 한 사람이 꾸는 꿈은 꿈으로 그치지만 많은 사람이 같이 꾸는 꿈은 현실이 된다.

  (징기스칸)

감사합니다.