교육 분야 빅데이터 활용 방안

2018.12.20.

책임연구원 이종현

KERIS 한국교육학술정보원
KOREA EDUCATION AND RESEARCH INFORMATION SERVICE

CONTENTS

Ⅰ. 한국교육학술정보원(KERIS) 소개

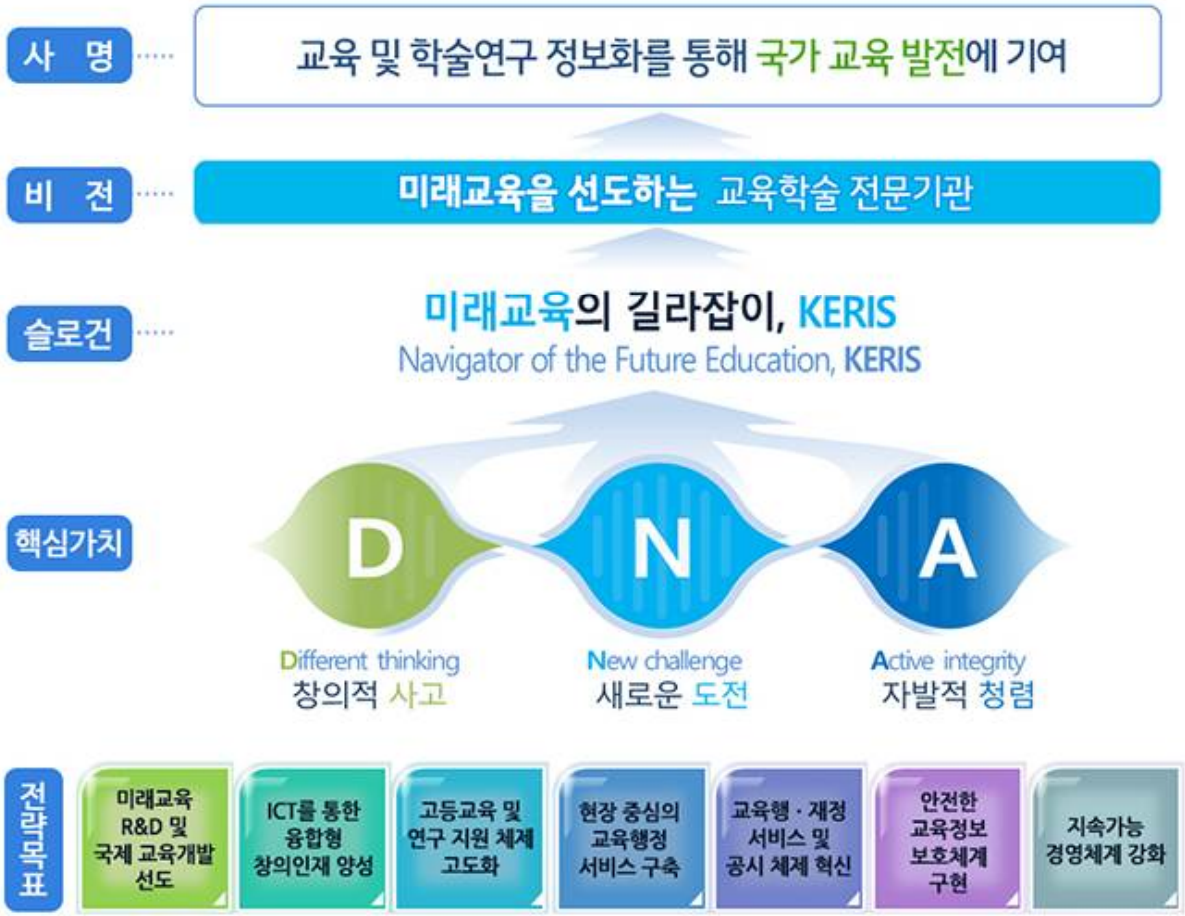
Ⅱ. 빅데이터 활용 가치 및 국내외 추진 동향

Ⅲ. 국내외 빅데이터 활용 사례

Ⅳ. 2018년 KERIS 빅데이터 활용 추진 현황

Ⅰ 한국교육학술정보원 소개

Ⅰ 한국교육학술정보원 소개



≫ **설립근거: 한국교육학술정보원법**
≫ **설립연도: 1999년**
≫ **소재지: 대구 동구 혁신도시**
≫ **직원수: 300여명**
≫ **대표서비스: RISS, NEIS, 에듀넷**

II 국내외 빅데이터 추진 현황

II 빅데이터 활용 가치 및 국내외 추진 동향

1. 빅데이터의 활용 가치

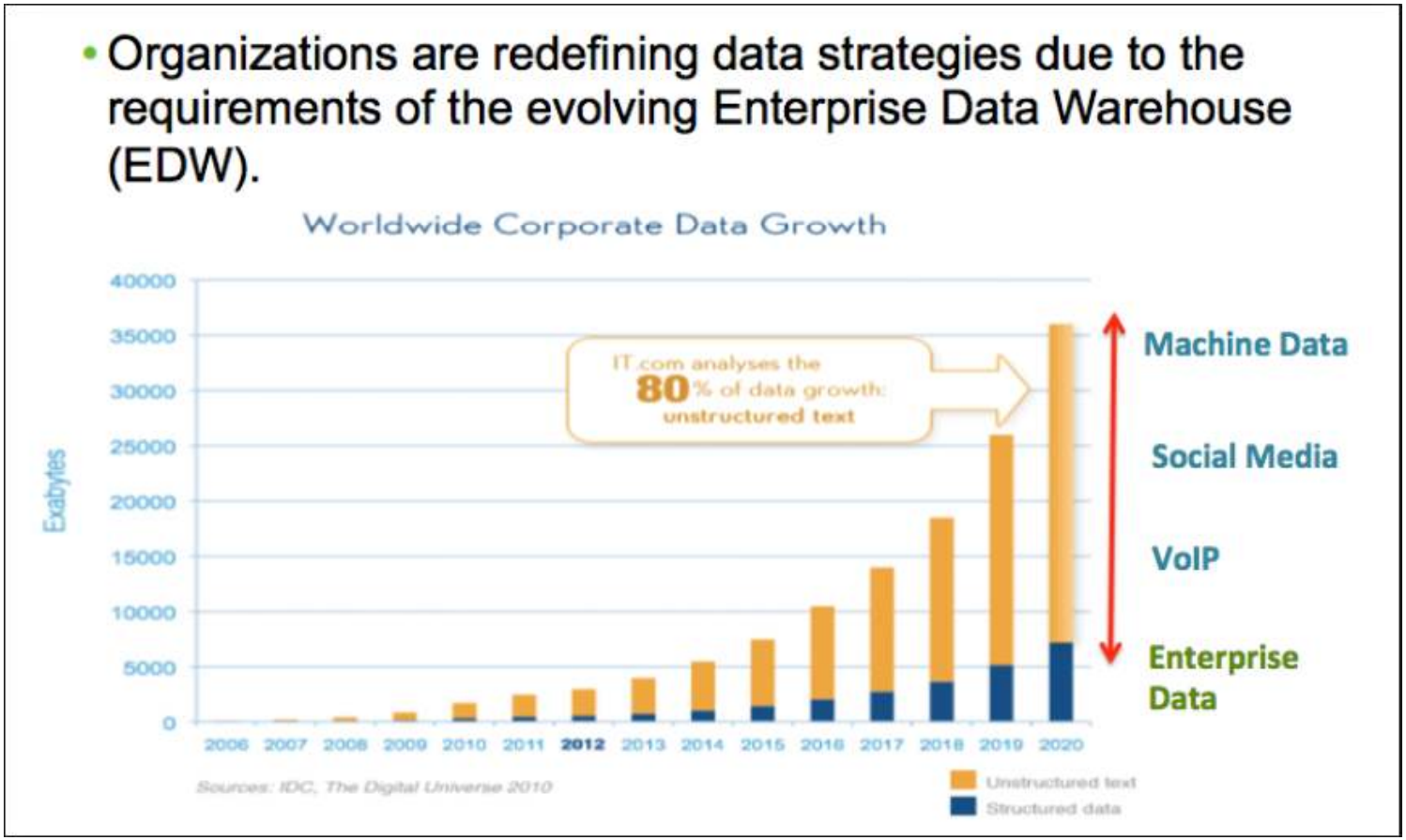
2. 국내외 정책 추진 동향

3. 빅데이터 분석 기술 동향

4. 통계처리, BI(Business Intelligence), 빅데이터 기술 비교

1 | 빅데이터의 활용 가치(#1)

≫**Social Media의 보편적 활용과 IoT의 확산은 데이터의 폭발적인 증가로 이어질 전망(IDC, 2010)**
특히 IT.com은 비정형의 텍스트가 데이터 증가의 80%를 차지할 것으로 분석함



1 | 빅데이터의 활용 가치(#2)

≫ **빅데이터의 정의는 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 데이터 셋 규모로, 그 정의는 주관적으로 앞으로도 계속 변화될 것(맥킨지 글로벌 인스티튜트)**

가트너

- 3V로 정의: Volume(크기), Variety(다양성), Velocity(속도)
- 5V로 정의: 3V + Veracity(정확성), Value(가치)

Volume(크기)

일반적으로 Terabyte~Petabyte 의미

[참고]
1024Gigabyte = 1Terabyte,
1024Terabyte = 1Petabyte,
1024Petabyte = 1Exabyte,
1024Exabyte = 1Zettabyte,
1024Zettabyte = 1Yottabyte

Variety(다양성)

정형, 반정형, 비정형 데이터 망라

- (정형데이터) 관계형 DBMS, 구조화, 연산가능
- (반정형데이터) XML, 구조화, 연산불가능
- (비정형)TEXT/이미지 비구조화, 연산불가능

Velocity(속도)

실시간 처리가 불가능한 데이터 생성 속도

[참고] 1분당 생성되는 데이터양
구글서치: 2억개
이메일: 2억 4백만개
페이스북: 350GB
유튜브: 72시간

Veracity(정확성)

편향되거나 노이즈가 없는 데이터 속성을 의미

Value(가치)

데이터로부터 가치를 추출할 수 있는 근원적인 목표를 달성할 수 있는가?

1 | 빅데이터의 활용 가치(#3)

마윈(알리바바그룹 회장, 2015년)

"1930년대엔 사람들이 '보이지 않는 손'이 시장에 있다고 믿었기 때문에, 그래서 시장경제가 이긴 것"이라며 "하지만 손에 데이터를 쥐고 있는 지금의 우리는 예전엔 보이지 않던 그 손을 볼 수 있게 됐다"

쇤베르거(옥스포드 교수, 2015년)

"빅데이터는 새로운 시각으로 세상을 보게 해 주는 안경"

4차 산업혁명 핵심요소

손정희(소프트뱅크 회장, 2017년)

"데이터는 산업혁명 시대의 석유 같은 자원이다"

제프 베조스(아마존 CEO, 2017년)

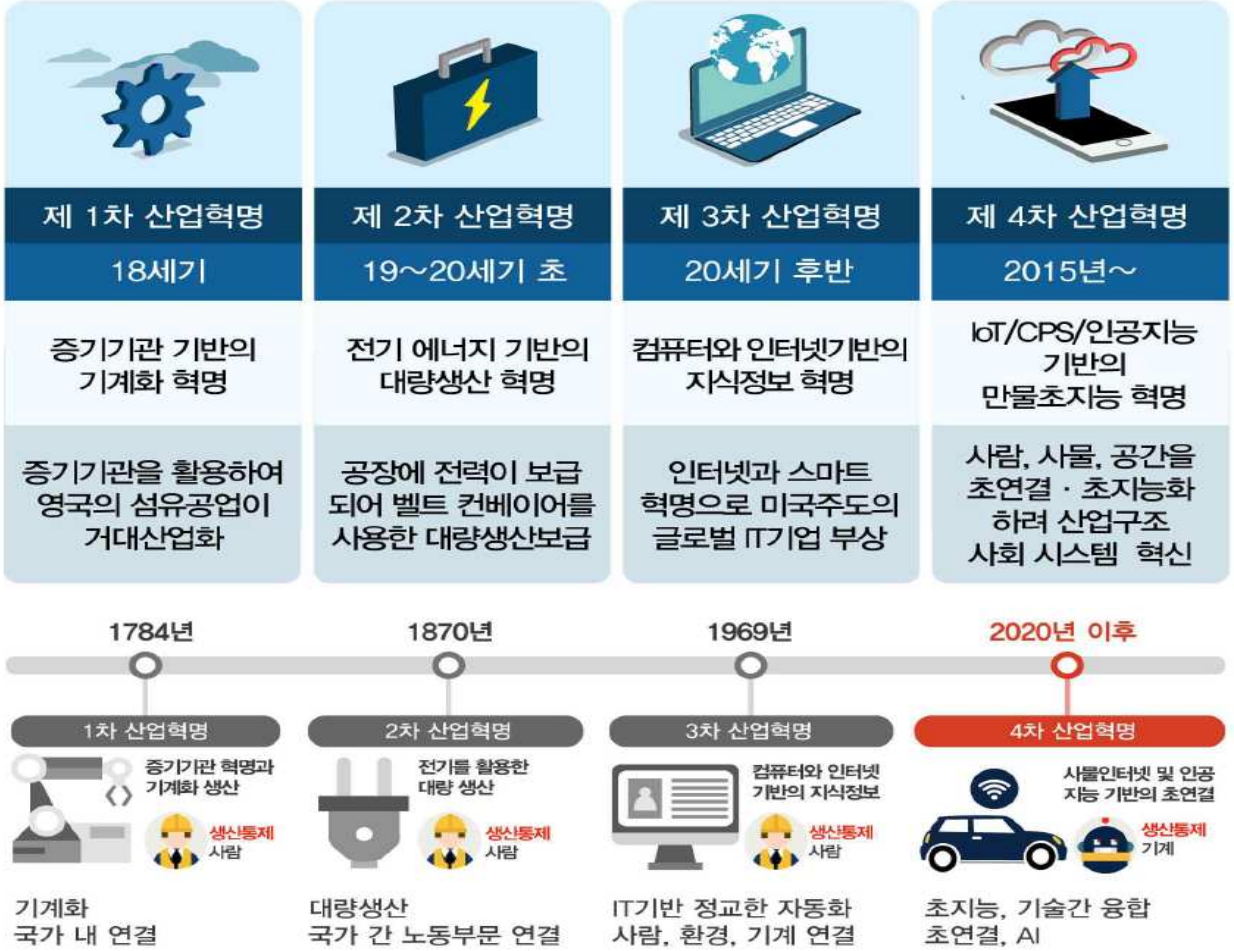
"데이터가 상품이나 서비스로 언제 중요해질지 알 수 없기 때문에 데이터를 절대 버리지 않는다"

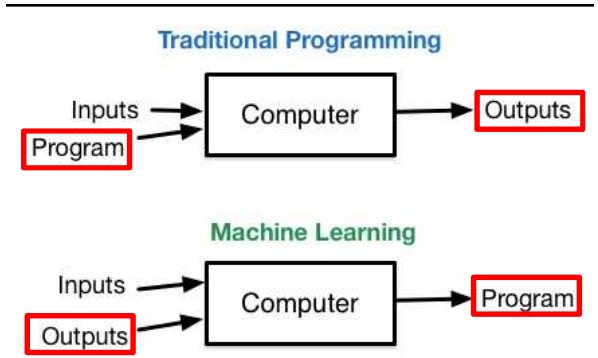
1 | 빅데이터의 활용 가치(#4)

미래 사회 특징	빅데이터의 역할과 가치	
불확실성	통찰력	· 현실세계 데이터 기반의 패턴분석, 미래 전망 · 다양한 가능성 시나리오, 시뮬레이션 제공 · 다각적인 상황이 고려 된 통찰력과 유연성 확보
리스크	대응력	· 환경, 소셜 데이터 분석을 통한 이상 징후 감지 · 이슈의 빠른 분석을 통한 실시간 의사결정 지원 · 국가, 기업 경영 투명성 제고 및 비용 절감
스마트	경쟁력	· 상황인지, 인공지능 기반의 신규 서비스 창출 · 개인화, 지능화 기반 차세대 사업 모델 발굴 · 평판, 트렌드 분석을 통한 기업 경쟁력 확보
융 합	창조력	· 타분야간의 결합을 통한 새로운 지식의 발견 · 상관관계 이해를 통한 시행착오 최소화 · 방대한 데이터 활용을 통한 新융합 시장 창출

(출처) 전북대학교(2017년)

1 | 빅데이터의 활용 가치(#5)





(출처) 전북대학교(2017년)

(Amazon Go) https://www.youtube.com/watch?v=NrmMk1Myrxc

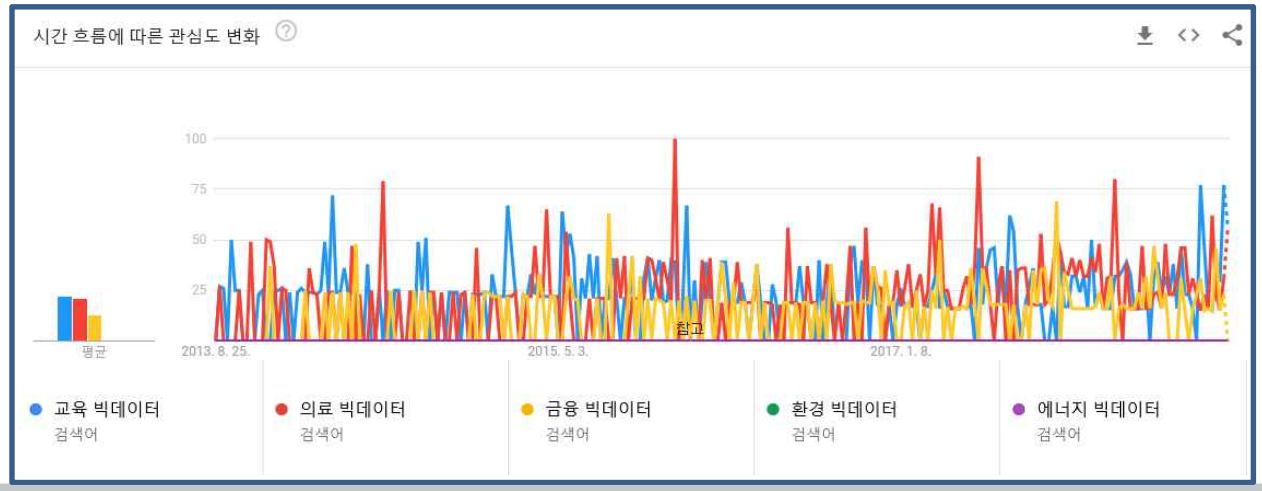
1 | 빅데이터의 활용 가치(#6)

≫**영국 경제 비즈니스 연구센터는 빅데이터가 2012년에서 2017년까지 영국 산업에 2,160억 파운드 (한화 310조 4천억원)의 경제효과와 5만 8천개의 일자리 창출효과도 있을 것으로 전망**

Economic Benefits	2011	2012-17
Business efficiency	17,379	149,471
Business innovation	2,865	24,062
Business creation	4,843	42,430
Total	**25,087**	**215,964**

(출처) http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf

≫**국내 주요 빅데이터 분야별 구글 검색 관심도**



- **5년 평균 구글 검색 빅데이터 분석 결과: ①교육(22),②의료(21),③금융(13) 분야 순으로 관심도 높음**
- **조사기간: 2013.8.23.~2018.8.23. (최근 5년간)**

2 | 국내외 정책 추진 동향(#1)

≫**2017년에 발표된 스위스 국제경영개발대학원(IMD) 국가경쟁력 평가 결과에 따르면 디지털경쟁력 순위에서 우리나라의 빅데이터 사용 및 활용능력 수준은 평가대상 63개 국가 중 56위로 평가됨**

DIGITAL TRENDS - OVERALL

KOREA REP.

OVERALL PERFORMANCE (63 countries)

Overall | Factors

1 19

14 Knowledge

17 Technology

24 Future readiness

10 Adaptive attitudes

3 48 Business agility

23 IT integration

The direction of the triangle indicates the performance change from the last year:
△ improved or stable
▽ declined

OVERALL & FACTORS - 5 years	2013	2014	2015	2016	2017
OVERALL	14	21	18	17	19
Knowledge	14	12	13	15	14
Technology	13	18	16	13	17
2 Future readiness	19	25	24	25	24

FUTURE READINESS

Subfactors	2013	2014	2015	2016	2017
Adaptive attitudes	10	11	8	8	10
Business agility	31	43	38	43	48
IT integration	19	21	17	21	23

Adaptive attitudes	Rank
E-Participation	4
Internet retailing	4
Tablet possession	34
Smartphone possession	8
Attitudes toward globalization	31

Business agility	Rank
Opportunities and threats	46
Innovative firms	32
4 Agility of companies	34
▷ Use of big data and analytics	56
Knowledge transfer	32

IT integration	Rank
▶ E-Government	3
Public-private partnerships	29
Cyber security	49
Sofware piracy	24

- 공공데이터 개방은 영국의 56.6% 수준
- 선진국 대비 빅데이터 기술격차는 3.1년

(출처) IMD WORLD DIGITAL COMPETITIVENESS RANKING 2017

2 | 국내외 정책 추진 동향(#2)

≫**미국 및 유럽 등 주요 선진국은 4차 산업혁명의 패러다임 속에서 2010년대부터 빅데이터의 중요성과 가치를 인식하고, 국가적 차원에서 빅데이터산업의 활성화를 위해 다각적인 노력을 기울이고 있음**

미국

빅 데이터 전략 수립 및
범정부 협력거버넌스 구축

- THE FEDERAL BIG DATA RESEARCH AND DEVELOPMENT STRATEGIC PLAN을 발표 (2016.05)
  - –빅데이터 R&D의 7개 전략과 18개 세부과제

영국

민간 활용 촉진을 위한
공공데이터 개방 중심의 정책

- 데이터 전략 위원회(DBS)에서 'A strategy for UK data capability' 를 발표 (2013.02)
  - –빅데이터 전문인력양성소(Open Data institute)를 설립 : 5년간 140억원을 투자

유럽연합(EU)

국가 위기관리를 위한
민관협력형 지원체계 확립

- 유럽진행위원회(EC)는 빅데이터 가치연합(BDVA)과의 민관협력 사업을 추진키로 합의
  - –Horizon 2020 연구혁신(Research and Innovation) 프로그램 : 25억 유로 투자

일본

빅 데이터 활용으로
Active Japan 구현

- 총무성 산하 빅데이터 활용 특별 부회에서 「빅데이터 활용 기본전략 발표」(2012.05)
  - –2013년 89.3억엔 투자 예정

(출처) 전북대학교(2017년)

2 | 국내외 정책 추진 동향(#3)

≫**국내 정책동향**

2012년 스마트 국가 구현을 위한 빅데이터 마스터플랜

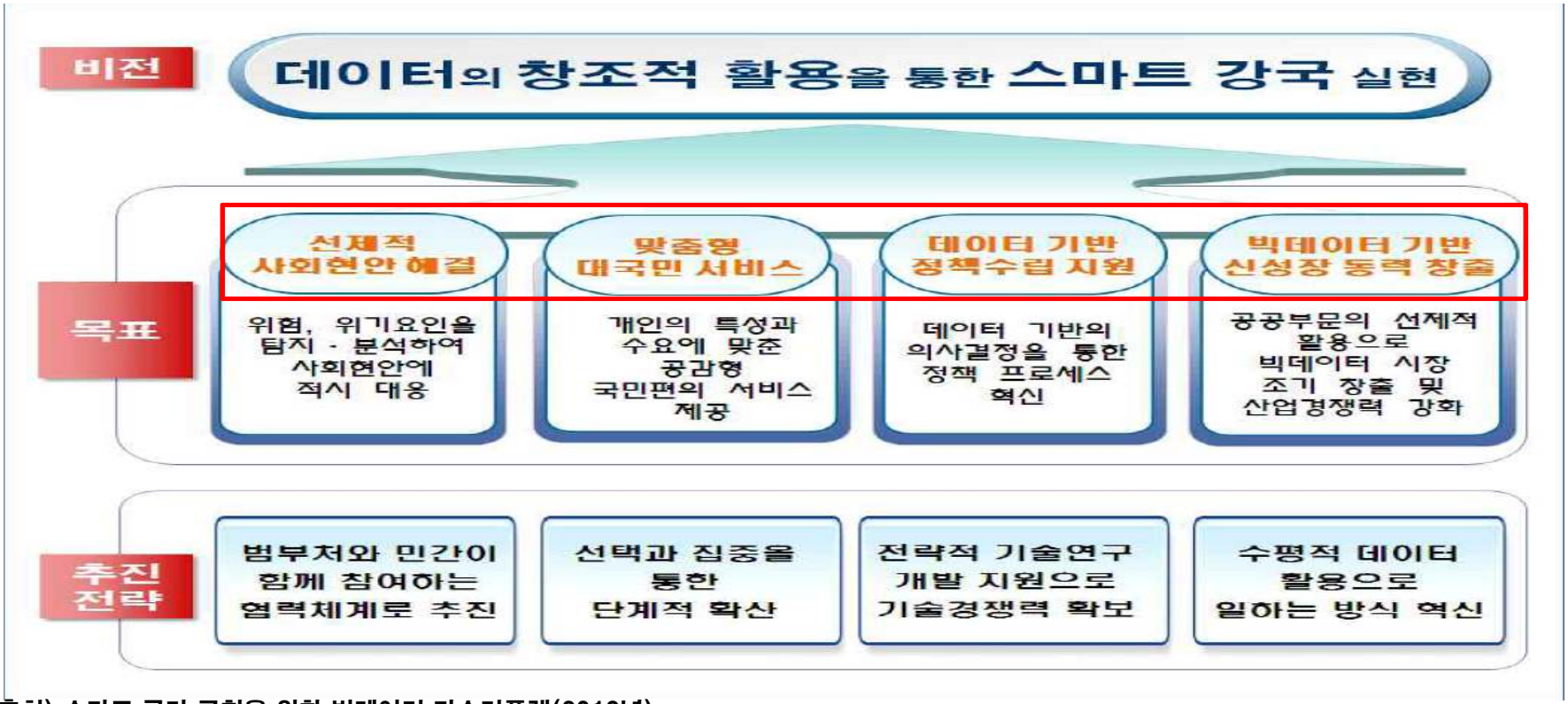
2016년 민관 합동 빅데이터 T/F

2017년 4차산업혁명위원회

2 | 국내외 정책 추진 동향(#4)

≫**2012년 교육부, 행안부 등 5개 기관 합동으로 데이터 기반 사회현안 해결을 위한 마스터플랜 수립**

2012년 스마트 국가 구현을 위한 빅데이터 마스터플랜

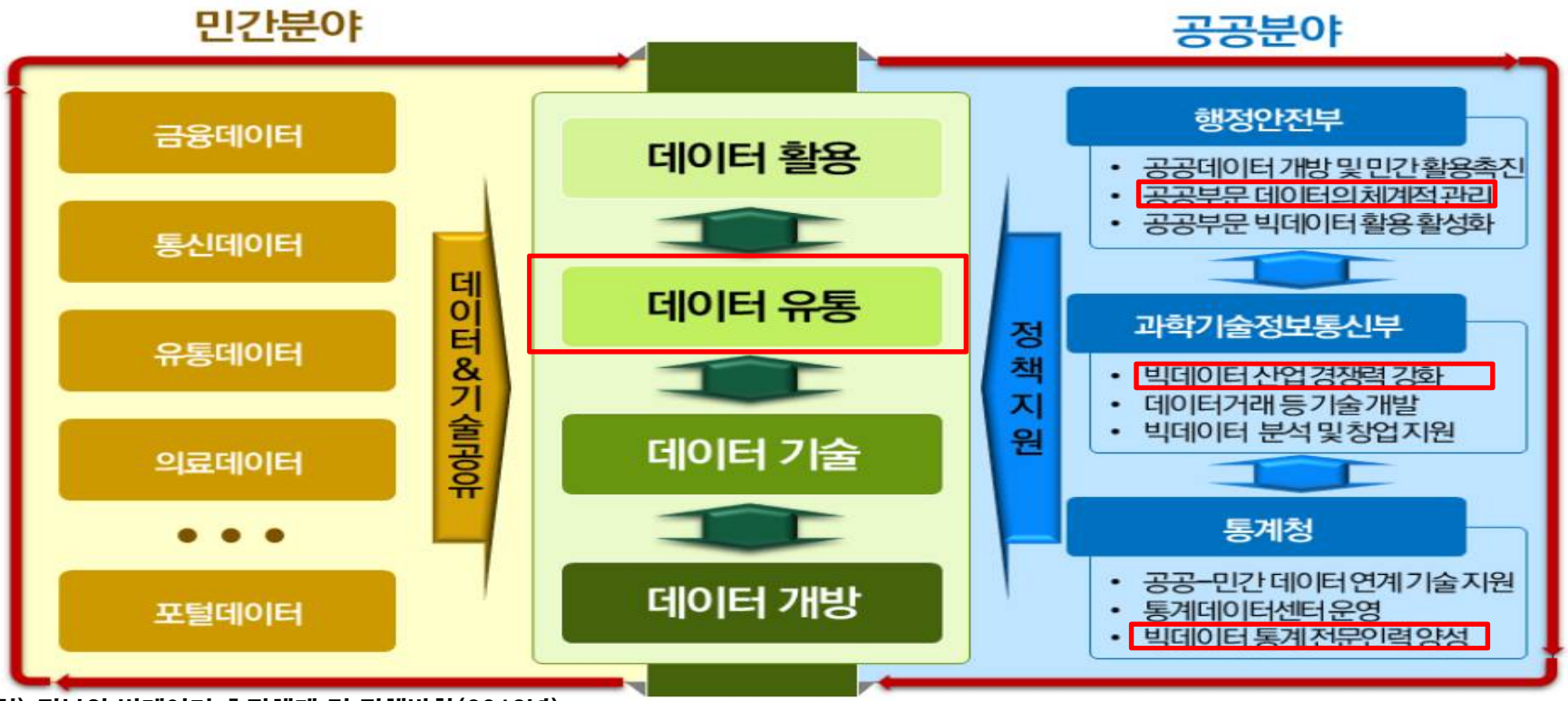


(출처) 스마트 국가 구현을 위한 빅데이터 마스터플랜(2012년)

2 | 국내외 정책 추진 동향(#5)

≫**행안부, 과기정통부, 통계청이 공공과 민간 빅데이터 간 상호 연계 및 활성화를 지원하기 위한 TF 출범**

2016년 민관 합동 빅데이터 T/F



(출처) 정부의 빅데이터 추진체계 및 정책방향(2018년)

2 | 국내외 정책 추진 동향(#6)

≫**문재인 정부 출범 이후 국가차원의 4차 산업혁명 시대 국가전략 수립을 위한 대통령 직속기구로 설립**

2017년 4차산업혁명위원회(#1)

비전 「모두가 참여하고 모두가 누리는 사람 중심의 4차 산업혁명」 구현

지능화 혁신 프로젝트 추진

[기술] 성장동력 기술력 확보	[산업] 산업 인프라 · 생태계 조성	[사회] 미래사회 변화 대응
지능화 기술경쟁력 확보 혁신성장동력 육성 R&D체계 혁신	초연결 지능형 네트워크 구축 데이터 생산·공유 기반 강화 신산업 규제개선 중소벤처/지역거점 성장동력화	핵심인재 성장지원 미래사회 교육 혁신 일자리 안전망 확충 사이버역기능·윤리 대응 강화

(출처) 4차산업혁명위원회 보도자료(2017년)

2 | 국내외 정책 추진 동향(#7)

≫**4차산업혁명위원회 규제혁신 해커톤(2018.2.~2018.4.)을 통해 개인정보보호 규제혁신 방안 도출**

2018년 4차산업혁명위원회(#2)

(출처) 한국인터넷진흥원(2018년)

2 | 국내외 정책 추진 동향(#8)

≫문재인 대통령 데이터경제 활성화 규제혁신 현장방문 축사(2018.8.31.) 발췌

- "이제 우리는 빅데이터, 인공지능, 자율주행차 등 새로운 도전에 직면했습니다. 산업화 시대는 석유가 성장의 기반이었습니다. 4차 산업혁명 시대, 미래 산업의 원유가 바로 데이터입니다. 데이터와 인공지능의 결합이 다양한 새로운 산업을 만들어 낼 것입니다."

- "데이터의 적극적인 개방과 공유로 새로운 산업을 도약시켜야 합니다. 데이터 규제혁신의 목표는 분명합니다. 데이터의 개방과 공유를 확대해 활용도를 높이는 것입니다. 이를 통해 신기술과 신산업, 새로운 제품과 서비스를 창출하는 것입니다."

- "우리 혁신성장의 미래가 데이터에 있습니다. 데이터를 기반으로 하는 신산업, 신기술을 위해 데이터 규제혁신이 빠르게 이뤄져야 합니다."

2 | 국내외 정책 추진 동향(#9)

≫(과기정통부) 빅데이터 전문센터 지정 및 네트워크 협의체 운영(2018년말 사업종료)

- (개요) 분야별 빅데이터센터 간 협력과 시너지 창출을 위해 빅데이터 협력 네트워크 조성
  - 빅데이터 네트워크 협의체 참여를 통해 국내 빅데이터 생태계 확산을 선도할 민간 대표 기업 및 연구기관을 빅데이터 전문센터로 지정
  - 2018년 30여개 기관·기업으로 구성된 빅데이터 협의체 운영
- (주요 협력 아젠다) 신사업 수요창출, 데이터 개방/유통 표준화, 우수사례 발굴, 일자리 창출
- (향후계획) 2019년 연간 800억원(100여개 기관 협업) 규모의 사업 기획 중

2 국내외 정책 추진 동향(#10)

≫(행정안전부) 공공빅데이터 센터 구축 사업(예정)

- (추진배경) 공공빅데이터센터의 근거 법률인「데이터기반행정 활성화에 관한 법률」이 '17년 12월 정부안으로 국회에 제출됨에 따라 법 제정과 함께 공공빅데이터센터가 신속히 구성되어 가동될 수 있도록 필요한 준비를 사전에 진행
- (개요) 정부통합데이터 분석기능과 공공/민간 빅데이터센터의 허브기능 수행

<공공빅데이터센터의 기본기능>

2 | 국내외 정책 추진 동향(#11)

- 해외 정책 시사점: 빅데이터의 가치에 대한 높은 인식으로 활용 활성화

1) 산재된 빅데이터의 수집과 활용을 총괄하는 기관의 중요성을 인식하고 있음

2) 빅데이터 활성화를 위해 개인별 맞춤형 서비스가 중요함을 인식하고, 개인정보보호와 빅데이터 효율적 활용 사이의 접점을 찾는 노력 중

3) 빅데이터 분야 우월적 기술과 인력 확보가 국가 경쟁력의 핵심이라 판단하고, 전문 인력 양성 분야를 투자정책의 핵심으로 강조

4) 빅데이터가 향후 경제성장의 핵심산업임을 인식하고 빅데이터 기반의 국가전략과 정책개발에 역량을 집중하고 있음

2 | 국내외 정책 추진 동향(#12)

- 국내 정책 시사점: 빅데이터 관련 정책 추진 중이나 가시적 성과는 미약

1) 정부 차원의 빅데이터 컨트롤 타워의 역할이 다소 모호함

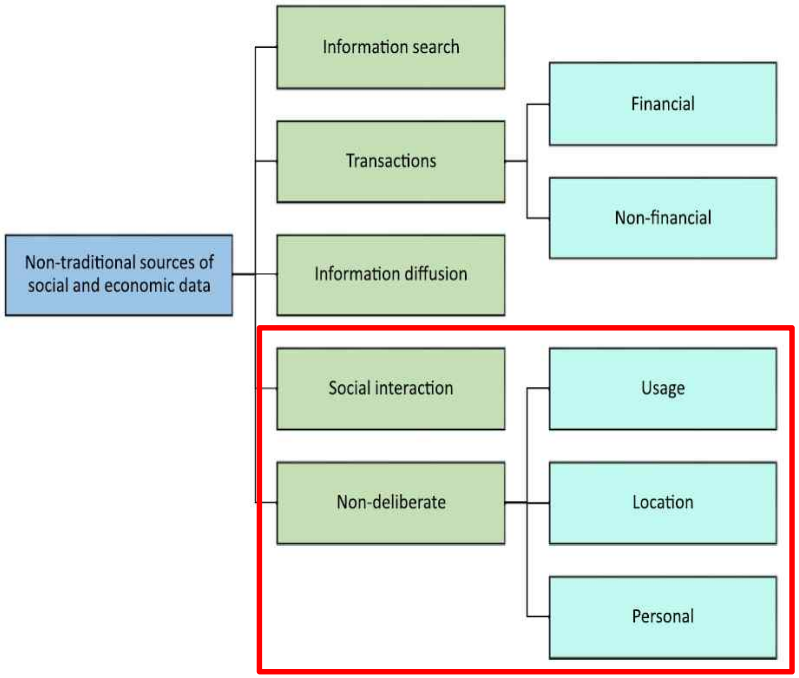
2) 빅데이터 센터 간 협업을 위해 과기정통부는 민간분야, 행정안전부 공공분야에 중점을 두고 국가차원의 빅데이터 공유 플랫폼 구축을 위해 노력하고 있음

3) 엄격한 개인정보보호 규제에 따라 데이터 활용도는 선진국 대비 낮은 상황

4) 공공데이터 개방 및 활용 수준도 선진국 대비 저조한 수준

3 | 빅데이터 분석 기술 동향(#1)

≫빅데이터 생산관점(수집대상) 분류체계

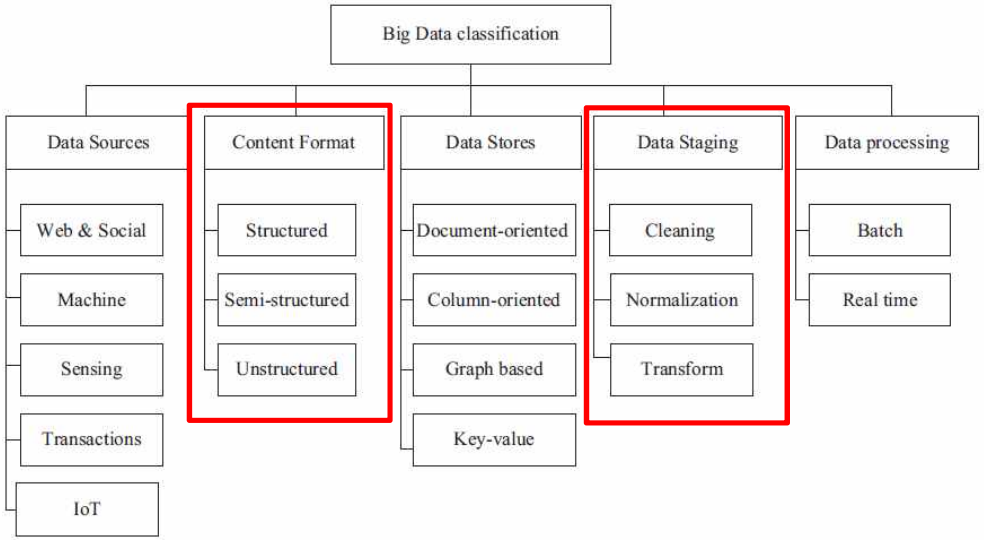


분류체계		설명
대분류	소분류	
Information search		– 이용자 관심사에 대한 주제 검색 시 능동적으로 생성되는 데이터 – 예: 검색엔진, 구글 트렌드
Transactions	Financial	– 이용자가 제품을 구매하거나 서비스를 이용하기 위해 비용을 지불하는 이벤트 – 예: E-bannking, e-commerce, urban sensors(tolls, 신용카드 리더기, 소매점 스캐너, 대중교통 카드 리더기 등)
	Non-financial	– 제품을 구매하거나 서비스를 이용하기 위해 정보를 상대방에게 제공할 때 발생하는 이벤트 – 예: E-govemment, e-recruiting
Information diffusion		– 이용자는 본인 또는 본인이 대표하는 기관의 마케팅 목적을 위해 또는 공적인 이미지를 만들기 위해 정보 또는 지식을 전파함 – 데이터는 능동적으로 생성됨 – 예: Corporate websites, apps, Wiki pages
Social interaction		– 이용자는 정보, 의견, 아이디어를 다른 이용자들과 공유함. 데이터는 능동적으로 생성됨 – 예: Social Networking Sites, opinion platforms, blogs
Non-deliberate	Usage	– 디바이스를 이용하여 언제, 어디서, 어떻게 활동했는지에 대한 연관 데이터를 단순한 사실 형태로 생산 – 데이터는 이용자의 행동의 결과로부터 수동적으로 생산됨 – 예: Web cookies, Internet Protocol, Sensors for self-tracking
	Location	– 모바일폰의 사용을 통해 이용자의 위치와 연관된 데이터가 생산됨 – 예: GPS, GSM, Bluetooth, WIFI Points 등
	Personal	– 이용자가 의식하거나(구매를 위해 입력하는 경우) 의식하지 못하는(개인수입을 추론하는데 활용되는 데이터) 형태로 개인정보(나이, 성별, 등)를 생산함 – 디바이스나 도구를 활용하여 특정 목적을 달성하는 과정에서 생산됨 – 예: Forms, profiles, type of searches or purchases

(출처) Desamparados Blazquez et al. / Technological Forecasting & Social Change 130 (2018) 99-113

3 | 빅데이터 분석 기술 동향(#2)

≫빅데이터 기술관점 분류체계

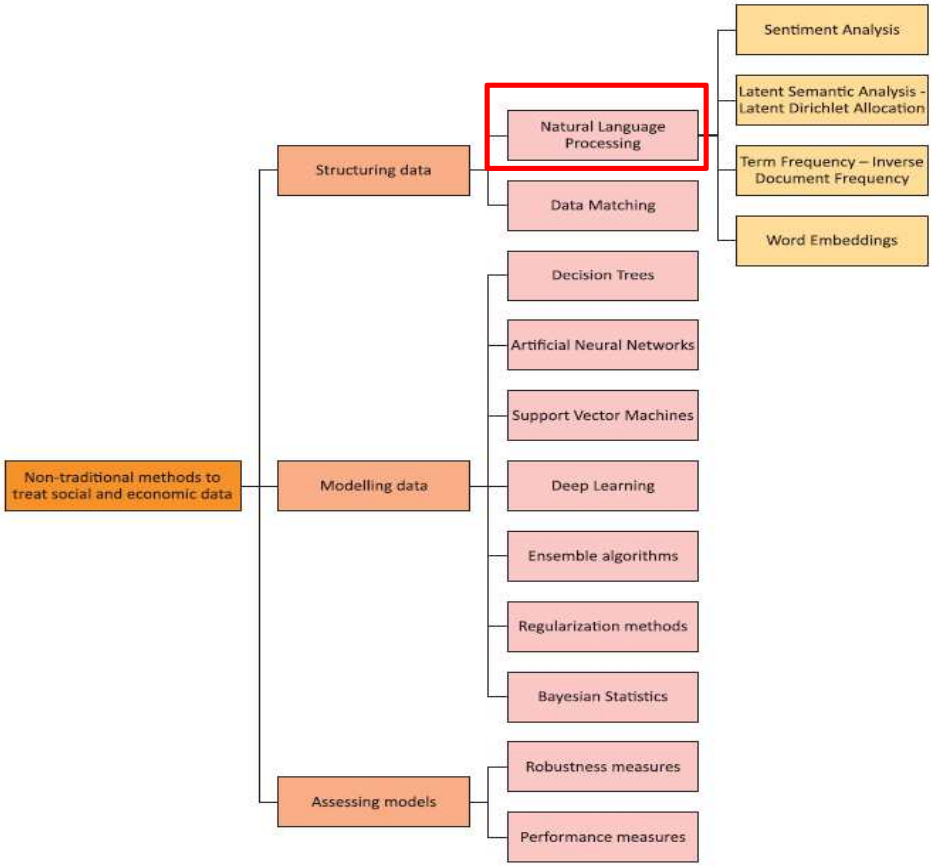


분류체계	소분류	설명
Data Sources	Social media	- Facebook, Twitter 등 SNS 및 블로그
	Machine-generated data	- 인간의 개입없이 하드웨어 또는 소프트웨어가 생산하는 데이터
	Sensing	- 물리적인 양을 측적하고 신호로 전송하는 장치
	Transactions	- 시간을 축으로 이벤트에 따라 데이터 생성
	IoT	- 인터넷의 구성원으로서 유일하게 구분되는 객체
Content Format	Structured (정형)	- RDBMS에 구축되어 SQL 또는 프로그램 언어로 관리되는 데이터 - 입력, 저장, 검색, 분석이 용이함. 숫자, 단어, 날짜 등의 형태로 구조화된 형태로 존재
	Semi-Structured (반정형)	- 전통적인 방식으로 데이터베이스에 테이블 형태로 구축된 데이터가 아닌 특정 파일 형식의 데이터 - 데이터를 수집하고 분석하기 위해 특정 파일 형식에 적합한 프로세스를 구비해야함
	UnStructured (비정형)	- 텍스트 메시지, 위치정보, 비디오, 소셜미디어 데이터 등 형식이 특정화되지 않은 데이터
Data Stores	Document-oriented	- 표준 형식(JSON, XML, PDF, MS Word 등)으로 저장 - RDBMS의 레코드와 유사하나 문서단위로 검색됨
	Column-oriented	- BigTable 구조의 데이터베이스처럼 컬럼 패밀리라는 개념을 활용함 - 성과 이름 컬럼이 이름이라는 컬럼 패밀리는 상위 컬럼 개념으로 통합관리됨
	Graph database	- Neo4j와 같은 graph 모델(노드, 엣지, 속성, 관계)을 저장하고 표현할 수 있는 데이터베이스
	Key-value	- 대규모 RDBMS를 저장하고 접근할 수 있는 방식 - 데이터의 각 조각(row, column, version)이 키 값을 기준으로 저장되고 검색됨
Data Staging	Cleaning	- 불완전하고 타당성이 부족한 데이터를 식별해내는 과정
	Transform	- 분석에 적합한 형태로 데이터를 변환하는 과정
	Normalization	- 중복성을 최소화하기 위해 데이터베이스 스키마를 구축하는 방법
Data Processing	Batch	- 처리속도가 긴 일괄처리 업무를 위해 최근 몇 년간 MapReduce 기반 시스템을 많은 기관이 도입함
	Real time	- 실시간 빅데이터 처리를 위해 가장 유명한 도구는 S4임(simple scalable streaming system)

(출처) I.A.T. Hashem et al. / Information Systems 47 (2015) 98-115

3 | 빅데이터 분석 기술 동향(#3)

≫빅데이터 분석 방법론(#1)



‣ 빅데이터 대부분의 데이터 원천은 사람이 직접 생성한 텍스트 형태여서 이를 완벽하게 이해하고 활용하기 위해서는 자연어 처리를 위한 특정 기술이 필요함. Natural Language Processing(NLP)은 컴퓨터가 자연어 텍스트를 이해하고 구조를 형성하여 다양한 응용 프로그램에서 활용할 수 있는 방법을 탐색하는데 초점을 맞춘 방법론임 (Chowdhury, 2005)

‣ NLP는 그 자체로 정보검색 시스템을 위한 쉬운 인터페이스를 제공하고 동시에 다양한 방식으로 텍스트를 구성하여 기본 정보를 보다 쉽게 추출 할 수 있는 일련의 기술을 포괄하는 분석 방법임

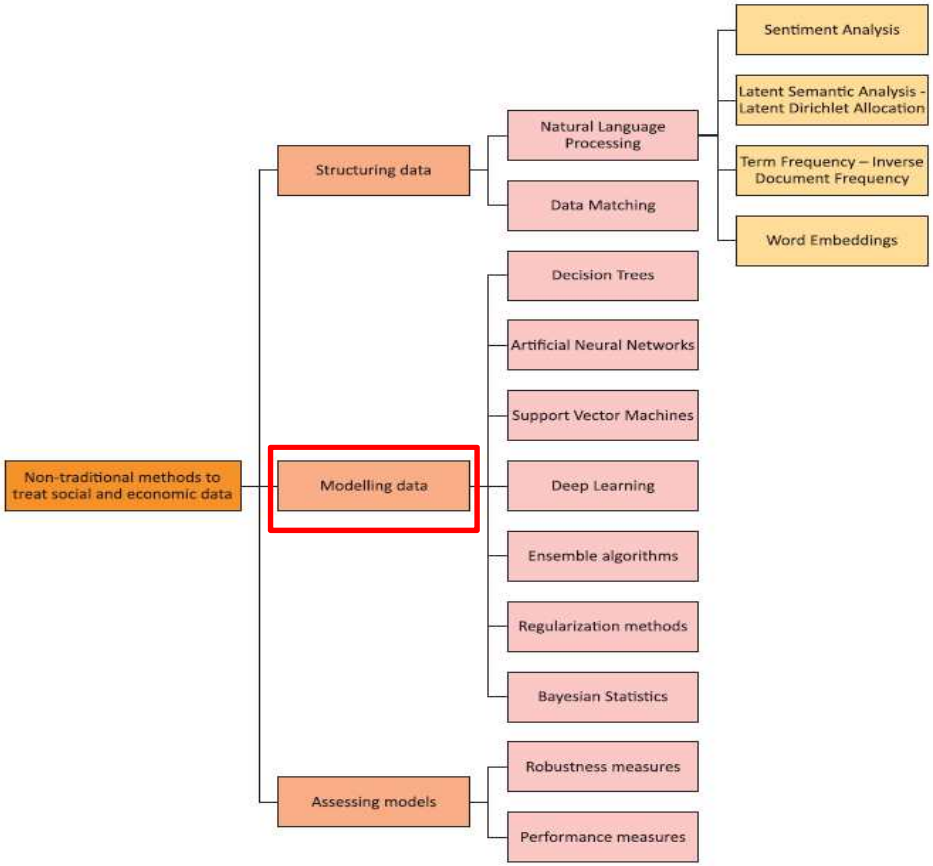
‣ 사회적 분석을 위한 몇 가지 흥미로운 NLP 기술로는 감정 분석 (Opinion Mining이라고도 함), 잠정 의미론적 분석(LSA), 잠재 Dirichlet 할당(LDA), TF-IDF(용어 빈도-문서 빈도 반전) 및 Word Embedding이 있음

(출처) Desamparados Blazquez et al. / Technological Forecasting & Social Change 130 (2018) 99-113

(Google Duplex) https://www.youtube.com/watch?v=JvbHu_bVa_g

3 | 빅데이터 분석 기술 동향(#4)

≫빅데이터 분석 방법론(#2)



‣ 데이터 모델링(및 그 관계)은 빅데이터 분석의 주요 프로세스로서 데이터 세트의 차원을 줄이는 것과 모델링 기술을 데이터에 적용하여 결과를 얻는 것을 포함함. 이용가능한 데이터의 유형과 분석의 목적에 따라 데이터 모델링을 위한 두 가지 패러다임, 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)이 적용될 수 있음(Hastie et al., 2013)

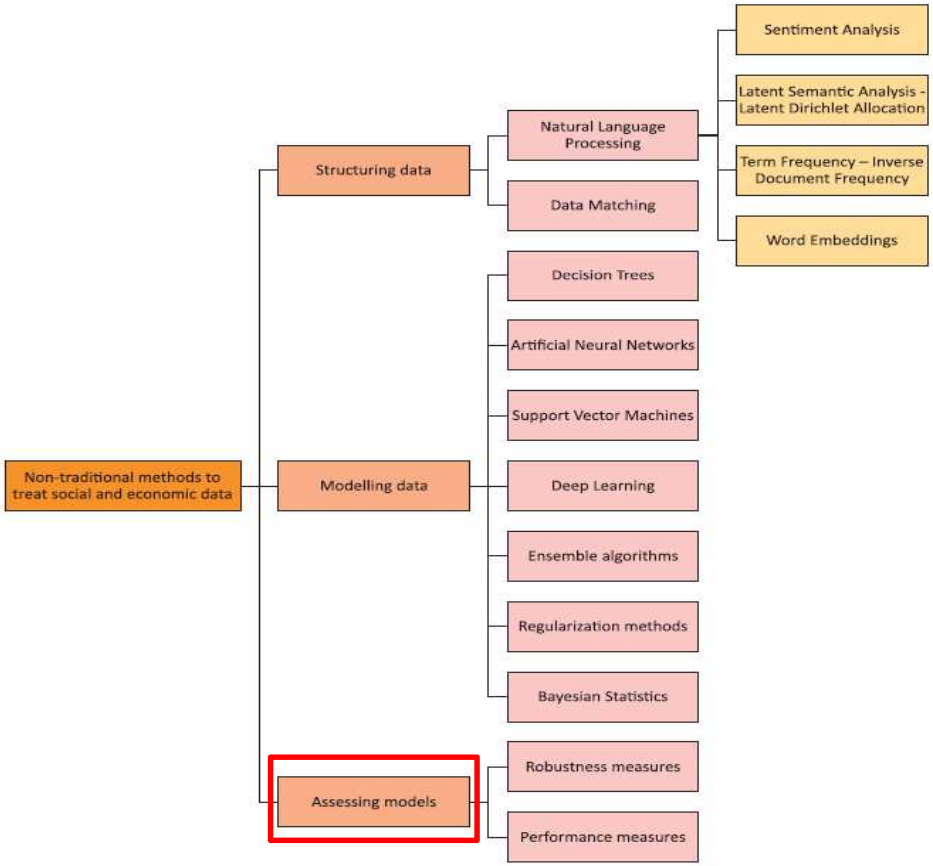
‣ 지도학습과 비지도학습에 다양한 기계학습 기술을 적용할 수 있음. 현재는 지도학습이 다양한 분야, 특히 예측이 필요한 경우 범용적으로 활용되고 있음

‣ 의사결정 트리, ANN, SVN, 딥러닝을 개선하여 성능 향상에 중점을 두고 개발된 Ensemble(앙상블) 알고리즘은 처리과정에 무작위성을 추가하여 과적합 문제를 해결하는데 적합한 방법론임

(출처) Desamparados Blazquez et al. / Technological Forecasting & Social Change 130 (2018) 99-113

3 | 빅데이터 분석 기술 동향(#4)

≫빅데이터 분석 방법론(#2)

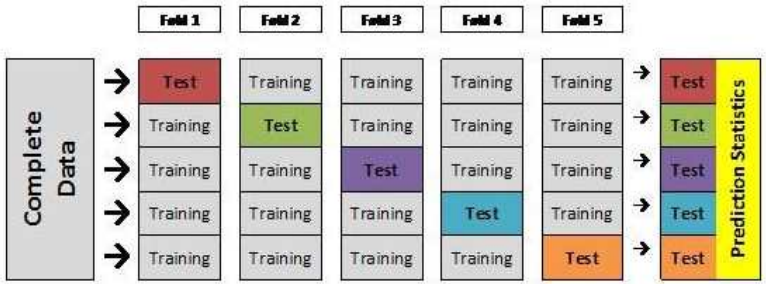


‣ 예측모델의 타당성(validity)과 적용가능성(applicability), 예측품질(quality of the predictions)을 판단하기 위해 정확성(Performance)과 유효성(robustness) 측면의 평가가 필요함. 정확성 평가는 예측모델이 학습에 사용된 데이터에 얼마나 적합한지, 정확한 결과를 산출하는지 평가하는 것이고, 유효성 평가는 학습하지 않은 다른 데이터셋에 대해 얼마나 잘 동작하는지 여부를 판단하는 것임

‣ 정확성과 유효성을 검증하는 방법은 홀드아웃(holdout) 방식과 K-Fold 교차 유효성 검증(cross-validation)의 크게 두 가지 방법이 활용됨

- 홀드아웃 검증은 초기 관측치의 75%~90%를 학습에 활용하고 나머지 관측치를 예측 성능 평가에 활용하는 방법

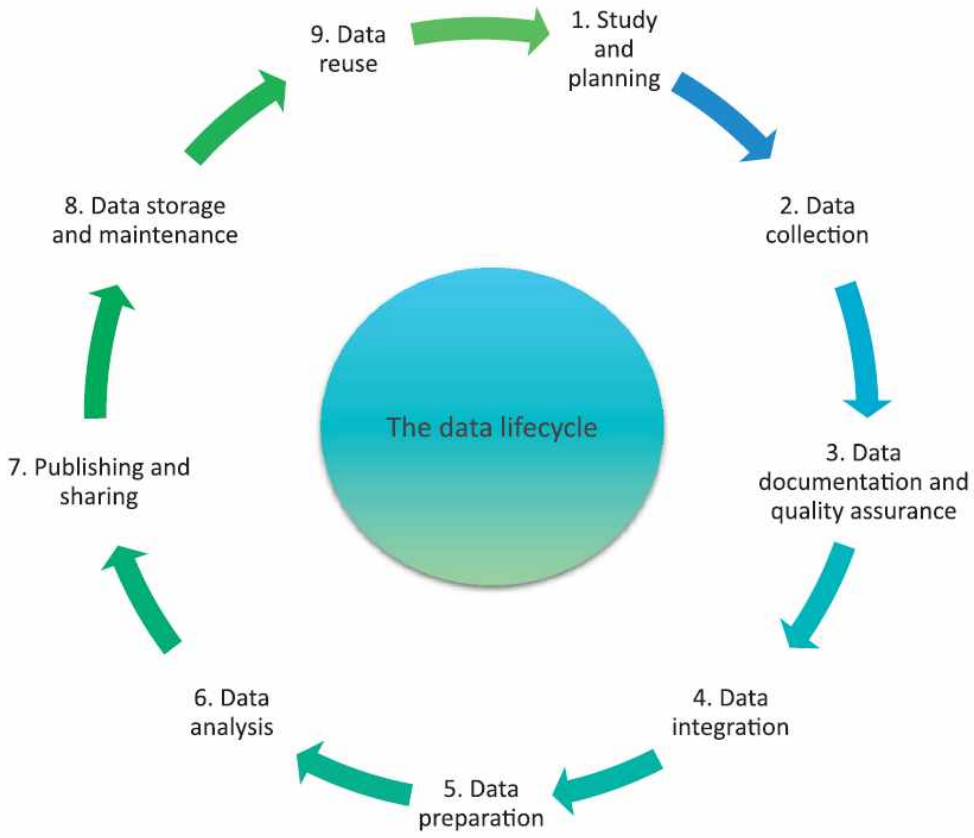
- K-Fold 교차 유효성 검증은 관측 데이터를 동일한 크기의 K개로 분할하여, 분할된 1개 영역과 나머지 K-1 영역을 학습과 평가에 반복적으로 활용하는 방법



(출처) Desamparados Blazquez et al. / Technological Forecasting & Social Change 130 (2018) 99-113

3 | 빅데이터 분석 기술 동향(#5)

≫빅데이터 생명주기



단계구분	설명
(1단계) Study and planning	- 연구(또는 프로젝트) 설계 과정 - 연구단계별 수집 대상 데이터 및 분석방법 선정 - 개인정보보호를 포함한 법제도적 문제 파악
(2단계) Data collection	- 내외부 데이터를 수집하는 과정
(3단계) Data documentation and quality assurance	- 수집된 데이터의 품질확인 및 메타정보 추출 과정 - 메타정보는 출처, 추출날짜, 데이터 형식, 추출방식 등을 포함 - 데이터 품질 및 유효성을 보장하기 위해 결측치 및 오입력 데이터에 대한 검증이 필요하고, 개인정보보호를 위한 익명화 등이 필요함 - 데이터 출처의 신뢰성 확인 역시 필요함
(4단계) Data integration	- 서로 다른 데이터 출처로부터 수집된 데이터를 일관성 있게 융합하여 연속적인 데이터화 하는 과정 - 데이터 속성 식별, 동일한 단위적용 등이 포함됨 - 통합된 데이터로부터 추출된 자료가 익명성을 보장해야함
(5단계) Data preparation	- 분석도구에 적합한 형태로 통합된 데이터를 정제하는 과정
(6단계) Data analysis	- 데이터 분석, 결과 수집 및 해석, 결론 도출의 과정 - 빅데이터 처리를 위한 통계적 기법과 연산 도구가 적용되는 단계
(7단계) Publishing and sharing	- 분석된 결과 및 결론을 보고서화 하거나 예측모델로 구축하는 과정 - 출판물은 저작권을 설정하고, 예측모델은 서비스에 배포하는 과정을 포함
(8단계) Data storage and maintenance	- 통합(학습 및 검증대상) 데이터에 대한 보존 과정 - 메타데이터 백업과 불필요한 개인정보를 삭제하는 활동 포함 - 법·제도적 제약사항을 고려하여 보존처리
(9단계) Data reuse	- 이전에 수집, 처리, 분석, 저장한 데이터를 재사용하는 과정 - 새로운 가설의 테스트 또는 새로운 연구를 위한 준비단계에 활용

(출처) Desamparados Blazquez et al. / Technological Forecasting & Social Change 130 (2018) 99-113

3 | 빅데이터 분석 기술 동향(#6)

- 분석 기술 시사점: 빅데이터 대부분을 차지하는 비정형 소셜미디어 정보와 IoT 정보 처리가 중요한 이슈로 부상함

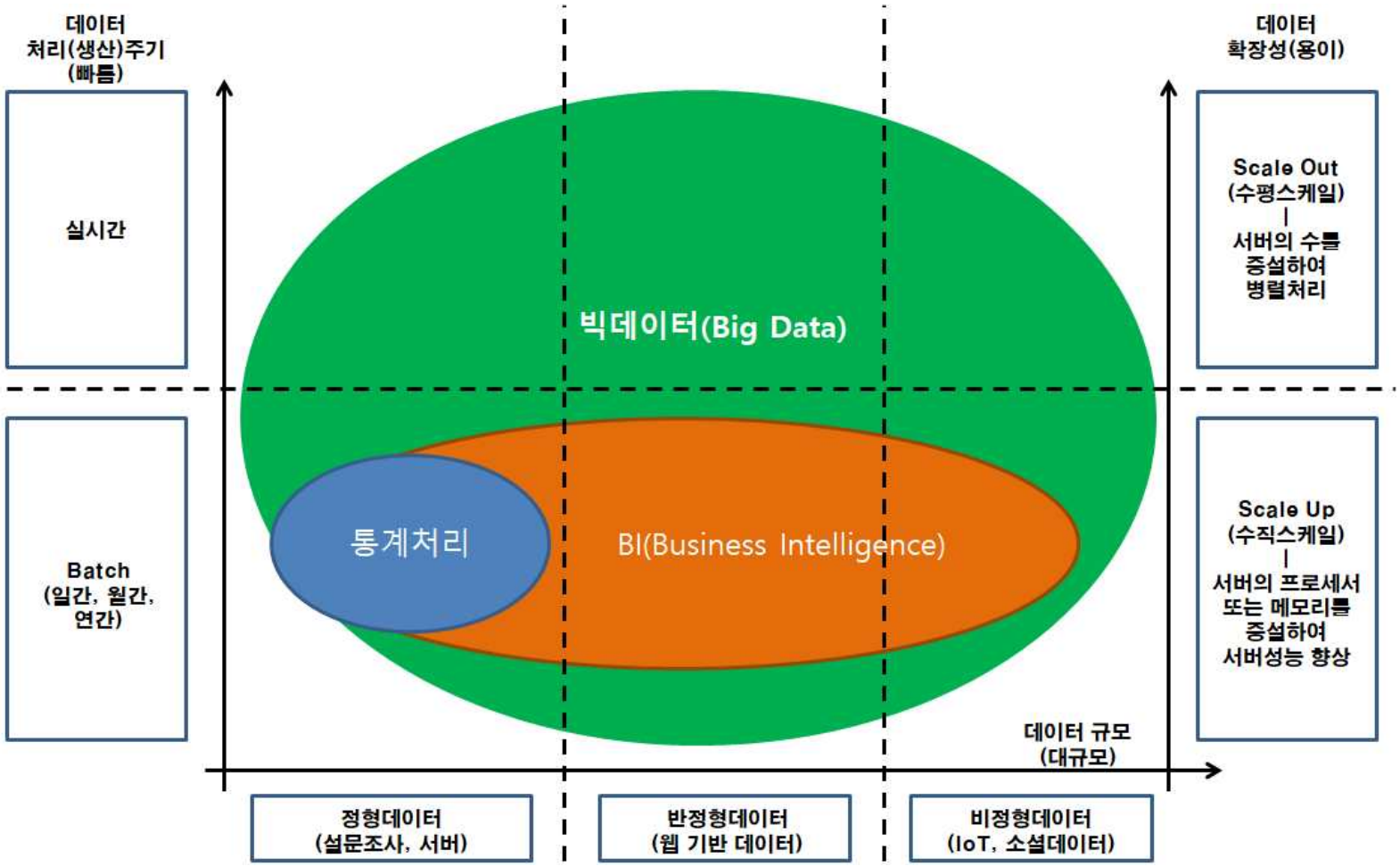
1) 빅데이터 예측모델링 방법은 다양한 알고리즘이 활용되고 있으며, 최근에는 과적합(overfitting) 이슈를 해결하기 위한 기법들의 연구가 활발함
2) 다양한 기계학습 기술 중 베이지안 방법론이 간편하고 효율적인 구현과 높은 예측 정확도로 인해 인기를 끌고 있음
3) 예측모델의 유효성 향상을 위해 트레이닝 데이터를 효과적으로 활용하는 방안에 대한 방법론 연구도 활발함
4) 빅데이터를 넘어 스마트데이터를 지향하기 위해 빅데이터 라이프사이클에 관한 연구도 활발함

4 통계처리, BI(Business Intelligence), 빅데이터 기술 비교(#1)

[표] 통계처리, BI, 빅데이터 분석 비교표

구분	통계처리	BI	빅데이터 분석
주요 목적	- 표본데이터 분석 기반 일반화(예측)	- 내부 경영데이터를 중심으로 분석·가시화하여 기업 내 프로세스 개선	- 전체 모수를 대상으로 군집·관계·분류·예측 분석 수행 - 내부 경영데이터와 외부 소셜 및 IoT 데이터를 종합적으로 분석하여 예측모델 개발
처리 데이터	- 정형자료	- 정형자료 위주	- 정형, 반정형, 비정형
데이터 처리속도	- Batch (일간, 월간, 연간)	- Batch 처리 중심 (일간, 월간, 연간) - 실시간 데이터 처리 가능	- 실시간 데이터 처리 중심 - Batch (일간, 월간, 연간)
시스템 규모 확장성	- Scale Up (서버의 프로세서 또는 메모리 증설)	- Scale Up (서버의 프로세서 또는 메모리 증설)	- Scale Out (서버의 수를 증설하여 병렬처리)
주요 활용층	- 개별 연구자	- 기업경영층 (전사차원 활용) - 부서별 의사결정 시 주로 활용	- 개별 연구자 - 기업경영층 (전사차원 활용) - 데이터 과학자 - 일반 직원 (상시 활용)

4 | 통계처리, BI(Business Intelligence), 빅데이터 기술 비교(#2)



III 국내외 빅데이터 활용 사례

II 국내외 빅데이터 활용 사례

1. IR(Institutional Research) 활용 사례

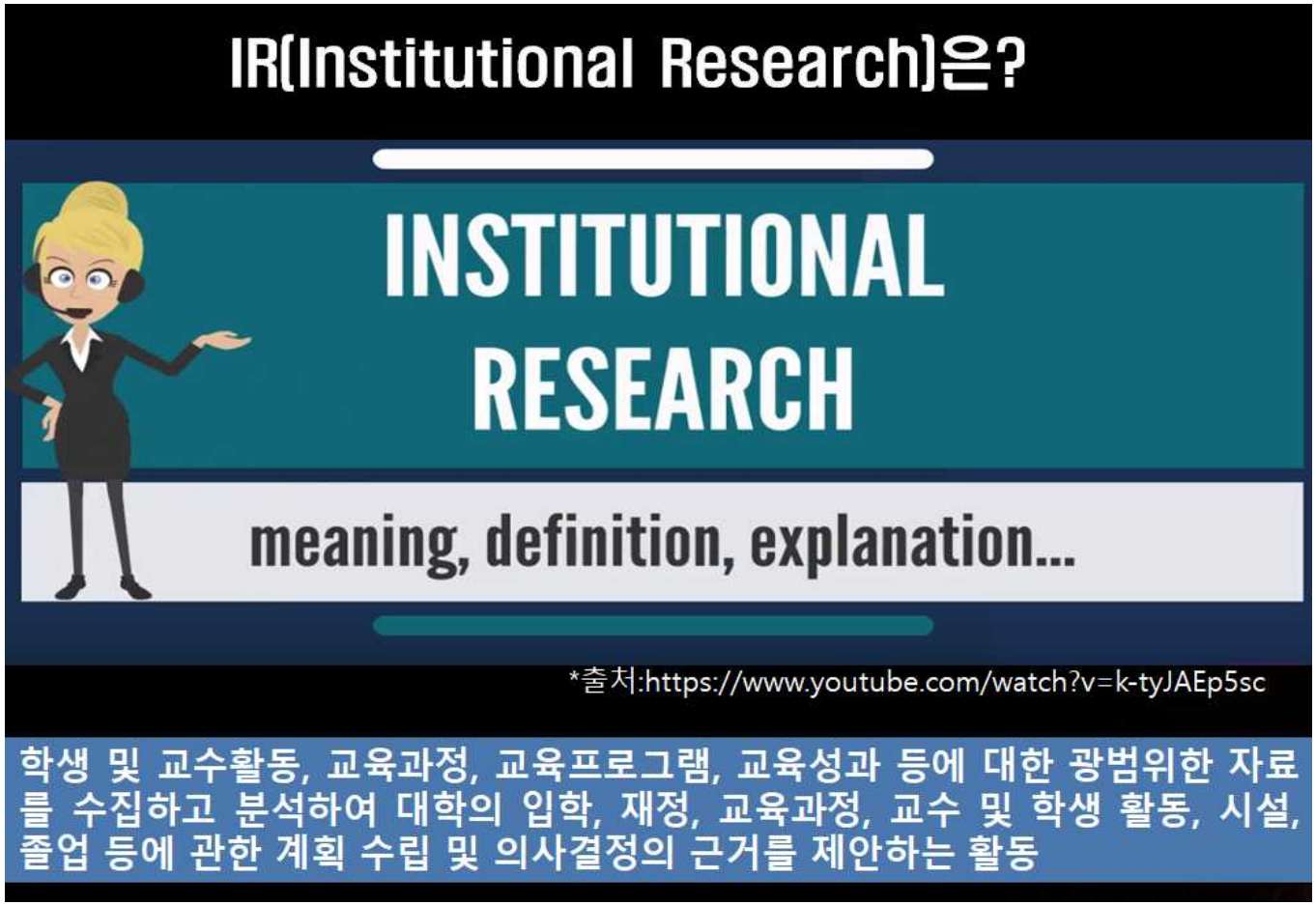
2. 지능형(적응형) 학습서비스

3. 소셜(Social) 빅데이터 분석

4. 공공 분야 빅데이터 활용 사례

1 IR(Institutional Research) 활용 사례(#1)

≫**IR이란?**



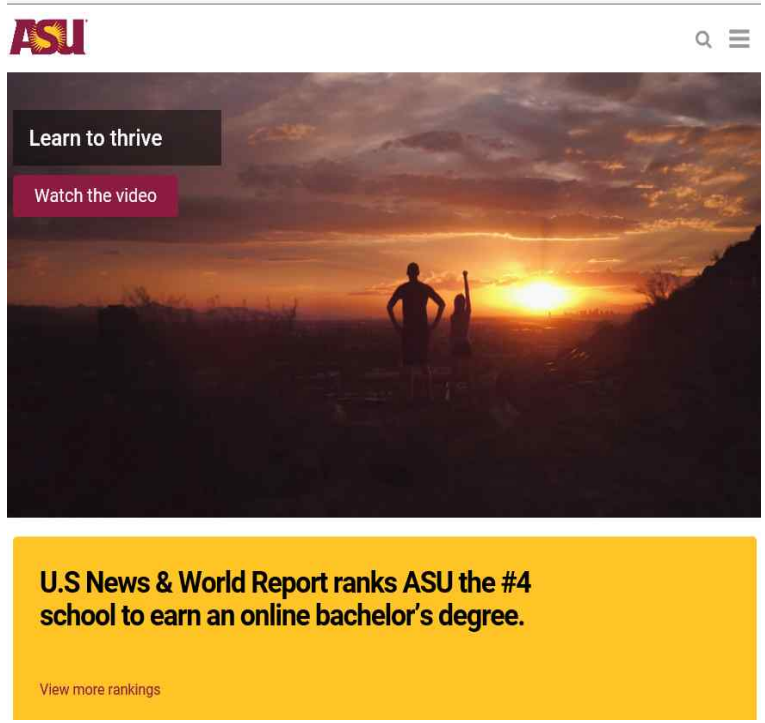
1 IR(Institutional Research) 활용 사례(#2)

애리조나 주립 대학

- **(배경) 입학생들 중 다수가 대학 수준에 미치지 못하는 수학 (교과) 준비도를 보이고 있으므로 기초 수학 과목부터 중도 포기 및 낮은 수준의 학업 성취를 보임**
- **(실행) 데이터 분석(learning analytics) 및 문항반응모델(item response model) 기술을 활용해서 학생의 수준별 학습자료 및 학습과정을 제공하는 적응형 학습 플랫폼(adaptive learning platform) 제공**

  - 학습 시간, 학습 참여, 문제풀이 등 데이터를 수집하여 개별적으로 학생 분석

  *adaptive Learning Analytics 모델 개발을 꾸준히 추진하고 있으며, 최근 수학 외 다른 과목으로도 확대 적용 중



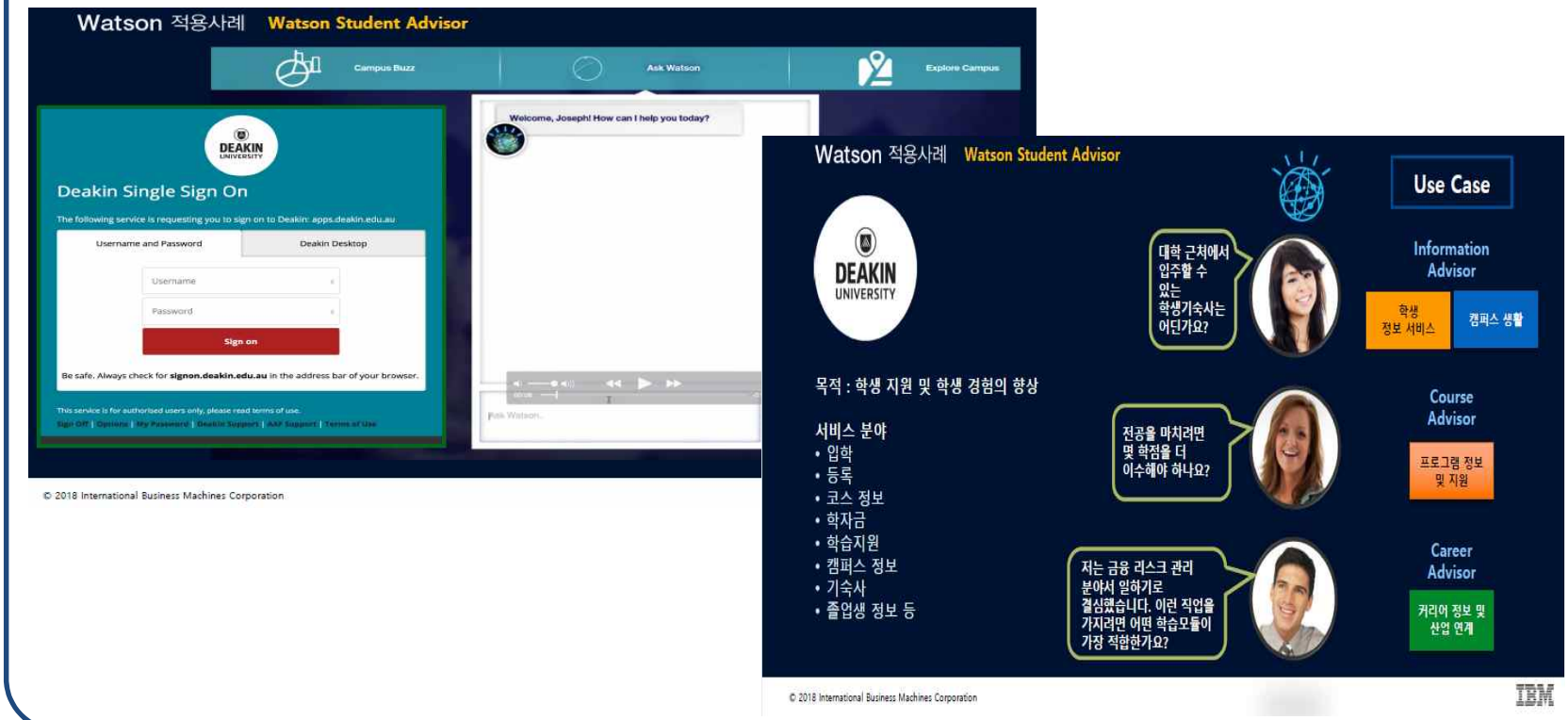
- **(성과) 대학 기초수학 과정 이수율이 65%에서 85%로 향상되었으며, 학생들의 수업 만족도도 향상되어 고급 과정에서의 이수율에도 긍정적 효과가 반영됨**

1 IR(Institutional Research) 활용 사례(#3)

빅데이터와 AI의 활용 : DEAKIN University (호주)

- **Student Advisor는 AI왓슨(IBM)에 의해 24시간 365일 제공**

IBM 왓슨이 제공하는 가치 : 전문가서비스확대, 신속한답변, 사용자 경험, 온디맨드, 개인화, 지속적 개선 등



* 출처 : 한국IBM/CTL리더 김경전 상무 발표자료 재구성

1 IR(Institutional Research) 활용 사례(#4)

빅데이터의 활용 : 입학생 모집 전략 (IBM Big Data & Analytics Hub)

"3 ways universities are leveraging big data analytics for recruitment and retention"

- **(*Going straight to the source*) 입학생 모집을 위한 핵심성과지표(Key Performance Indicators)를 분석하여 광범위한 홍보 대신 선택과 집중 전략 수립**

  * (과거) 표준 시험 점수, 고등학교 내신성적, 교과외 활동 등과 같은 계량지표를 참조하여 학생 선발

  * (현재) 학업 이력, 인구통계학적 정보, 지리적 정보 등을 고려하여 선발 시 등록율을 높일 수 있을 것인지에 대한 예측

  * 전략적 등록관리(strategic enrollment management)에 관해서는 입학허가 결정에 영향을 미치는 변수 분석(한석수, 2006, 한국교육)

- **(*Looking at social media signals*) 사회 여러 분야와 마찬가지로 소셜 미디어를 활용해서 잠재적 입학생들의 관심과 학업에 대한 행동모델 등을 분석**

  * (미국) 대학 입시 담당자들 중 40%는 지원자의 소셜 미디어를 추가적으로 분석했다고 응답 (2015년 Kaplan 조사)

- **(*Keeping students on track*) 빅데이터를 활용하면 학업 중 발생하는 문제에 학교가 개입하여 해결 가능**

  * 선수 학습에서 문제가 발생하면 처방없이 고급과정으로 넘어갔을 때 성공적인 학업성과를 기대할 수 없으므로 사전에 문제 해결을 위한 방안 마련과 상담 등의 조치를 통해 학습 과정을 학교가 지원

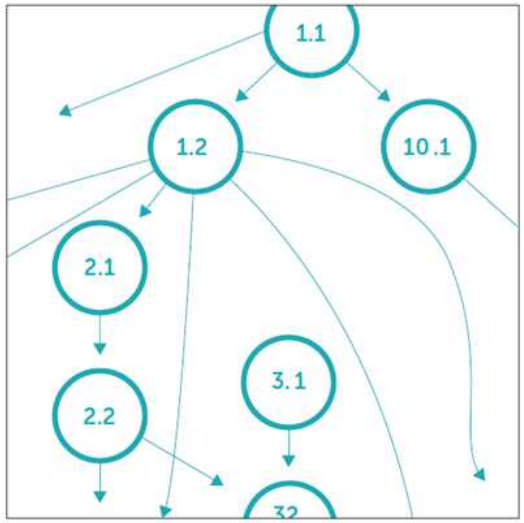
2 | 지능형 학습서비스(#1)

- The Knewton Platform 소개
  - 기업개요: 호세 페레이라(Jose Ferreira)가 08년 창립한 적응형학습 엔진개발 전문 기업으로써 학생의 문제풀이 시간 등의 데이터 측정을 바탕으로 학습분석 및 코칭
  - 적용 사례: 아리조나주립대(ASU) Math Readiness 과정

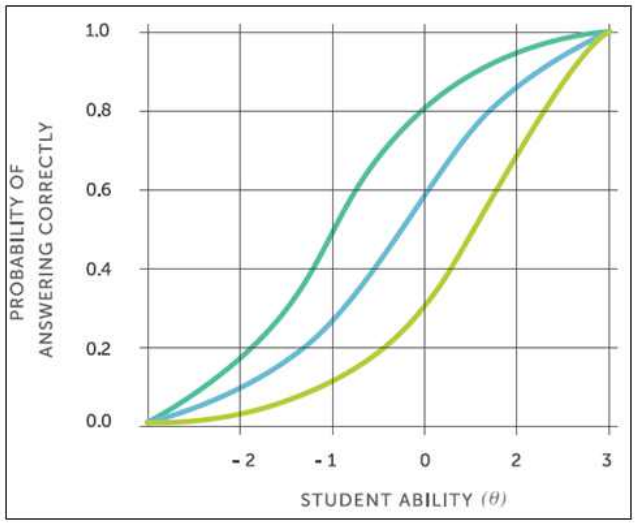
    * 수업 포기비율 감소(13%→6%), 수업 완료비율 증가(66%→75%)

2 | 지능형 학습서비스(#2)

- The Knewton Platform 구성 개념
  - (엔진 설계개념) 문항반응이론(Item Response Theory) 기반 예측 알고리즘
  - (학습지식 구조) 개념 간 관계(위계)를 설정하여 저난이도~고난이도 그래프(맵)을 구현함



A representation of the Knewton Knowledge graph



Typical item response functions

(출처) Kevin Wilson et al. / A KNEWTON WHITEPAPER (2015)

2 | 지능형 학습서비스(#3)

- The Knewton Platform 주요 기능
  - (확장된 IRT 모델 적용) 학생의 성취도가 시간이 지나며 변화한다는 점을 허용하고 학습지식 그래프의 개념간 상관관계를 식별함
  - (문항추천 엔진) 학습지식 그래프 기반으로 개념 간 관계를 고려하여 학습자에게 적합한 난이도의 문항을 추천함
  - (학생 성취도 분석엔진) 학생이 교육자료를 학습한 시간과 학습과정의 이벤트를 수집하고 패턴을 분석하여 학생의 성취도를 추론함
  - (학생 성취도 예측엔진) 학생의 성취도 분석결과를 토대로 개념을 이해하는데 걸리는 시간을 예측함

2 | 지능형 학습서비스(#4)

- 노리(Knowre) 스마트 스쿨수학 소개
  - 개인교사가 학생들의 문제점과 수준을 진단하고 솔루션을 제시하듯 개인화된 교육을 제공하는 적응적 수학 교육학습 플랫폼
  - 국내 및 미국 대상 솔루션 공급(BtoB 형태 비즈니스) 현황
    - (대교:국내) 러닝센터에서 중학교 커리큘럼 주교재로 활용
    - (EBS:국내) '15년 12월 EBS 수목달 중등수학 오픈
    - (미국 공교육) 70개 학교, 2만명 학생에게 서비스 제공
    - (미국 사교육) Sylvan Learning*과 시범서비스 진행 중

      * 미국 최대의 러닝센터 프랜차이즈

(출처) 노리 스마트 스쿨수학 소개자료 (2018)

2 | 지능형 학습서비스(#5)

- 노리(Knowre) 스마트 스쿨수학
  - (엔진 설계개념) 문항반응이론(Item Response Theory) 기반 예측 알고리즘
  - (학습지식 구조) 개념 간 관계(위계)를 설정하여 저난이도~고난이도 그래프(맵)을 구현함
  - (문항추천 엔진) 학습지식 그래프 기반으로 개념 간 관계를 고려하여 학습자에게 적합한 난이도의 문항을 추천함
  - (학생 성취도 분석엔진) 문항풀이 결과와 정답률을 근거로 성취도 평가

(출처) 노리 스마트 스쿨수학 소개자료 (2018)

2 | 지능형 학습서비스(#6)

▪ 노리(Knowre) 스마트 스쿨수학

- 개인교사가 학생들의 문제점과 수준을 진단하고 솔루션을 제시하듯 개인화된 교육을 제공하는 적응적 수학 교육학습 플랫폼

1:1 학습코칭 방식을 디지털화한 학습법

노리의 디지털화된 단계별 문제풀이 가이드

"이차방정식 $x^2+5x-35=3x$ 의 해를 구하시오."

단위 지식	개념	수학적 표현
등식의 성질 – 뺄셈	등식의 양변에 같은 식을 빼어도 등식은 성립한다.	$x^2+2x-35=0$
이차식의 인수분해	이차 다항식을 두 개 이상의 인수의 곱으로 나타낸다.	$(x-5)(x+7)=0$
$A \times B = 0$ 식 풀이	$A \times B = 0$ 이면 $A=0$ 또는 $B=0$ 이다.	$x-5=0$ or $x+7=0$
$x+A=B$ 꼴 정수계수 방정식	$Ax-B=0$ 을 계산하면 $x=B/A$ 와 같다.	$x=5,-7$

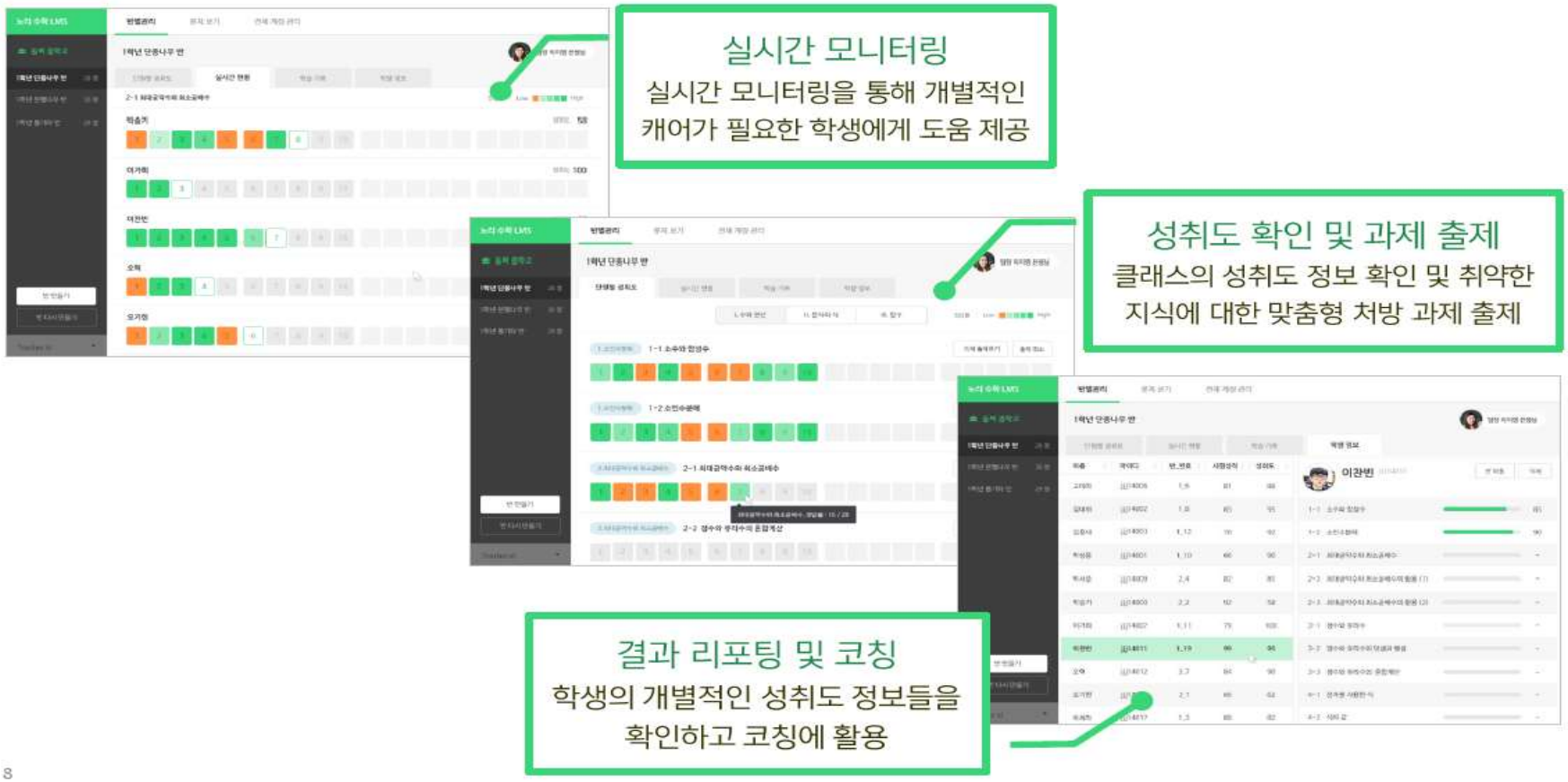
$x=5,\ -7$

학생의 문제풀이 과정에서 오답 발생 시,
하위단계로 문제를 쪼개어 과외선생님처럼 단계별 풀이 유도

(출처) 노리 스마트 스쿨수학 소개자료 (2018)

2 | 지능형 학습서비스(#7)

- 노리(Knowre) 스마트 스쿨수학
  - 교사 및 관리자가 학생의 학습정보를 실시간으로 확인하여 효과적인 코칭 가능



8

(출처) 노리 스마트 스쿨수학 소개자료 (2018)

2 | 지능형 학습서비스(#8)

- Kidaptive ALP(Adaptive Learning Platform) 소개
  - '11년 실리콘밸리에서 창업하여 빅데이터 플랫폼 기반으로 학생들의 학습행동과 문제풀이 데이터를 분석하여 학습자 맞춤형 교육학습 서비스 제공
  - 국내 대상 솔루션 공급(BtoB 형태 비즈니스) 현황
    - (웅진) '18년 2월 웅진 북클럽 AI학습코칭 서비스 오픈
    - (청담러닝) '16년 호두잉글리시 서비스 오픈

(출처) Kidaptive ALP 소개자료 (2018)

2 | 지능형 학습서비스(#9)

- Kidaptive ALP(Adaptive Learning Platform)
  - (엔진 설계개념) 문항반응이론(Item Response Theory) 기반 예측 알고리즘
  - (학습지식 구조) 평가문항에 대한 정/오답 결과를 근거로 베이지안 확률론을 적용하여 저난이도~고난이도 항목을 구분함
  - (문항추천 엔진) 학습주제간 상관관계와 문항지식 난이도를 분석하여 적합한 난이도의 문항을 추천함
  - (학생 성취도 분석엔진) 학생이 교육자료를 학습한 시간과 학습과정의 이벤트를 수집하고 패턴을 분석하여 학생의 성취도를 추론함

(출처) Kidaptive ALP 소개자료 (2018)

2 | 지능형 학습서비스(#10)

▪ Kidaptive ALP(Adaptive Learning Platform)

- **학습자의 학습 수준을 고려하여 문항의 체감난이도, 응답 시간, 정답을 맞출 확률 제공. 이를 통해 아는 문제와 모르는 문제 등을 구별하여 최적의 경로와 최단 시간에 학습 목표에 도달할 수 있게 함**

- 문제풀이에 최소 40초 이상 소요되는 문제를 'A학생'은 8초 만에 풀어 틀림
- 'A학생'이 정답을 맞출 확률이 낮은 문제를 열심히 풀었으나 틀림. 연계 개념 강의를 수강하여 지도 필요
- 문제풀이에 최소 50초 이상 소요되는 문제를 'A학생'은 1초 만에 풀어, 이해를 하지 않고 푼 것으로 분석됨

(출처) Kidaptive ALP 소개자료 (2018)

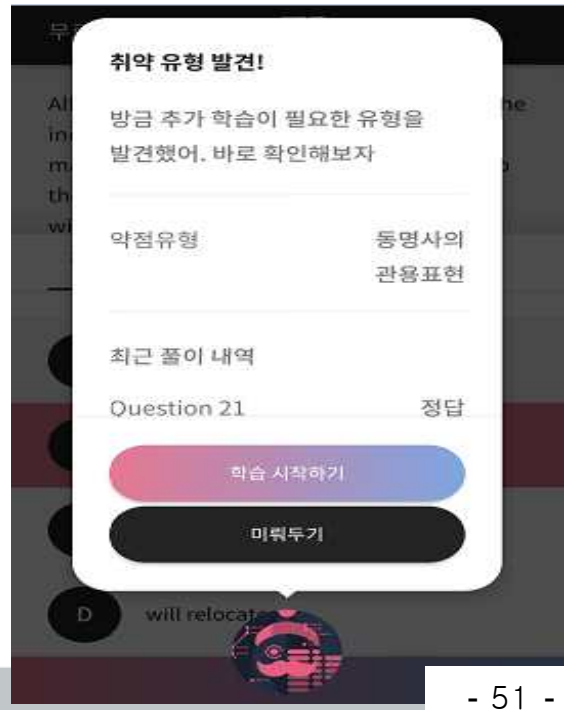
2 | 지능형 학습서비스(#11)

- Riiid 소개.
  - '13년 설립된 에듀테크 스타트업으로써 '17년 6월 AI기반 산타토익 런칭
  - 산타토익 다운로드 수는 20만건, 앱스토어 교육 매출 부분 순위 2위('17년 기준)
  - (엔진 설계개념) 데이터 중심 협업 필터링(fully data-driven collaborative filtering) 기반 예측 알고리즘
  - (학습지식 구조) 다수의 이용자가 사전에 문제를 풀어낸 결과를 기반으로 문항 간 상관관계와 문항의 저난이도~고난이도를 식별함
  - (문항추천 엔진) 문항 간 상관관계와 문항 난이도를 분석하여 적합한 난이도의 문항을 추천함

(출처) Kangwook Lee et al. / NIPS Workshop (2016)

2 | 지능형 학습서비스(#12)

- Riiid 산타토익 서비스 화면
  - 문제 풀이 결과에 대한 양적 분석 자료 제공, 유사 난이도 문제 식별로 효율적인 학습환경 제공
  - 유형별 학습자 성취도 분석 및 취약점 발견 시 추가학습 콘텐츠 제공



3 소셜(Social) 빅데이터 분석 (#1)

▪ 배경 및 목적

- 사회적 여론은 현상과 이슈를 발생하고, 이와 같은 사회적 목소리는 국가의 현안과 정책에 영향을 미침
- 최근 정보기술의 발전으로 과거 인쇄 미디어와 TV방송을 중심으로 전달되던 언론정보가 소셜미디어를 통해 빠르게 확산되고 재생되는 흐름을 보임
- 교육현장의 의견을 과학적으로 쉽고 빠르게 수집하여 교육정책에 반영할 수 있는 빅데이터 분석체제 마련

3 | 소셜(Social) 빅데이터 분석 (#2)

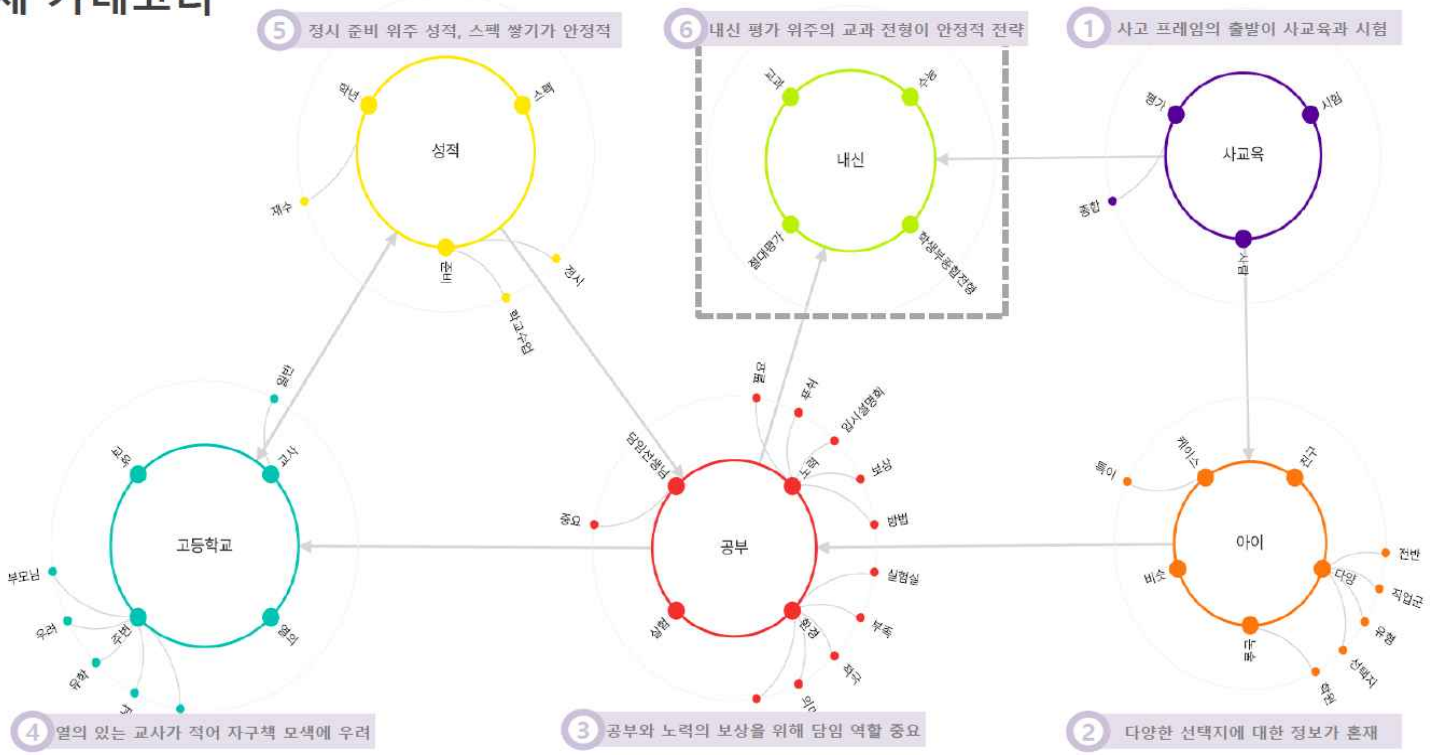
▪ 선행연구 분석

- ✓ 트위터를 활용한 소셜 네트워크 연구결과, 국제회의 네트워크 분석을 통해 참석자 중 유력자(영향력자)를 파악 및 주요 정보원에 대한 분석 가능(이미경 외, 2015)
- ✓ 그동안 언론학계에서 의제설정 연구를 위해 수작업에 의존하던 방법 대신에 정책의제의 변화추이를 쉽고, 편하게 분석할 수 있는 체제의 확보 필요성 확인 (박대민, 2016)
- ✓ 2015개정 교육과정 정책에 대한 언론보도 내용을 빅데이터 분석기법을 활용하여 연구결과, 언론 매체의 관심도 및 정책의 주제어, 정책의 주요 쟁점은 정책의 형성·결정·집행 단계에 따라 변화하는 것으로 나타남 (유예림, 2017)

3 | 소셜(Social) 빅데이터 분석 (#3)

- [소셜 빅데이터 분석기술을 FGI(Focus Group Interview)에 활용한 사례

학생, 학부모 그룹 의미망:
주제 카테고리



3 | 소셜(Social) 빅데이터 분석 (#4)

≫교육 분야 민간 빅데이터 분석 · 활용 사례(#1): 시대에듀

구분		내용
업체명		- 시대에듀
업체개요		- 국가기술자격 분야 출판종수 최대 기업
분석목적		- 효과적인 마케팅 전략수립을 통해 경쟁우위 선점
데이터	내부	- 회원데이터, 구매데이터, 수강데이터, 로그데이터
	외부	- SNS(뉴스, 트위터, 커뮤니티, 블로그, 카페) 데이터
분석방법		- 회원, 구매, 로그데이터 기반 군집분석 · (회원데이터) 나이, SMS/문자 신청여부 · (구매데이터) 구매금액, 품목, · (로그데이터) 홈페이지 유입경로(링크/검색어), 방문주기, 평균 페이지 이동 수, 체류시간, 상품/동영상 열람정보

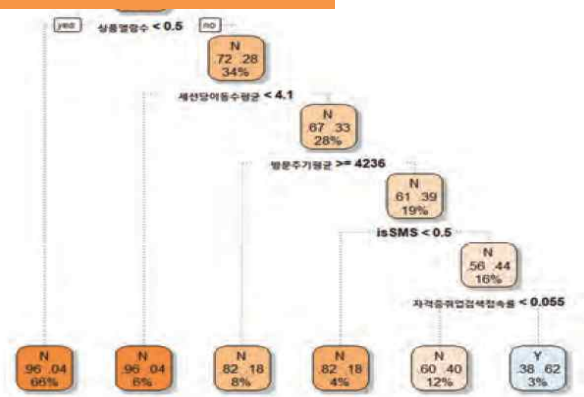
1. 고객분류

3. 시대에듀 검색 유입군 I (19%)	5.시대에듀 검색 유입군 II (시대에듀 정보 적극활용자)(5%)
• 대부분 포털을 통한 검색을 통해 접속하며, **접속시 대부분 시대에듀 관련하여 검색하여 유입**되며 '에듀'보다는 '고시' 키워드로 접속(포털에서 가장 많이 시대에듀 관련하여 검색) • 방문은 있는 편이며, 방문 시 비교적 여러 페이지를 보지만, 그룹 5에 비해서는 적은 편 • 낮은 연령대, 문자 및 메일링 신청도 적은 편	• 대부분 포털을 통한 검색을 통해 접속하며, **접속시 대부분 시대에듀로 관련하여 검색하여 유입**되며 '에듀 ' 보다는 '고시' 키워드로 접속 • **방문수가 가장 많으며, 방문시 가장 오랜시간 머물며 여러 페이지를 보고 상품열람 및 검색도 가장 활발** • 연령대가 가장 높고, 문자신청, 메일링신청도 꽤 있는 편

2. 자격 /취업 관심군 (32%)	4. 마케팅페이지 주유입군 (6%)	6. 학원 /강의 관심고객 (4%)
• **주로 자격증이나 교재 검색을 통해 접속하며,** 시대에듀 관련 검색은 없는 편 • 방문수가 적은 편이며, 상품열람 검색도 적은편 • 메일링신청, 문자신청이 가장 많으며 연령대도 두번째로 높은 편	• 주로 포털검색이 아닌 직접접속 또는 마케팅페이지 링크를 타고 접속 • 메일링 신청과 문자신청이 적은 편 • 방문수가 비교적 적고, 홈페이지 내 이용률도 낮은 편	• 직접접속보다는 포털 검색을 통해 주로 접속하며, 주로 학원 / 강의 검색을 통해 접속 • 가장 긴 주기로 사이트에 접속, 접속시 비교적 많은 상품열람을 하며, 여러 키워드로 검색

연령대도 낮으며
• 방문수가 가장 없고, 시대에듀 방문시 가장 적은 페이지를 보고
• 상품열람, 검색도 가장 없음
• 포털을 통해 검색하여 접속하는 경우가 가장 적으며, 검색시에도 시대에듀 관련 검색은 가장 적음

2. 예측모델링



3. 강의분석

3 | 소셜(Social) 빅데이터 분석 (#5)

≫교육 분야 민간 빅데이터 분석 · 활용 사례(#2): 와신교육

구분		내용
업체명		- 와신교육
업체개요		- 독학재수학원으로 시작한 업체로, 사업다각화의 관점에서 원어민 화상영어 서비스 '스테디톡' 추진
분석목적		- Target층 분석을 통한 제한된 홍보 자원의 효과성 극대화
데이터	외부	- 뉴스, 트위터, 커뮤니티, 블로그, 카페
분석방법		- (핵심고객군 분석) SNS데이터를 활용하여 화상·전화영어 언급 직업군 분석 - (서비스 선택 결정요소 분석) 화상·전화영어와 함께 언급되는 서비스 선택기준에 대한 키워드 분석 - (홍보채널 분석) 화상·전화영어 언급량을 채널별로 분석

1. 고객층 식별

[화상 / 전화영어 초중고등생 세부 학년 키워드 분석]

2. 주요 마케팅 채널 식별

[화상 / 전화영어 주요 홍보채널 분석]

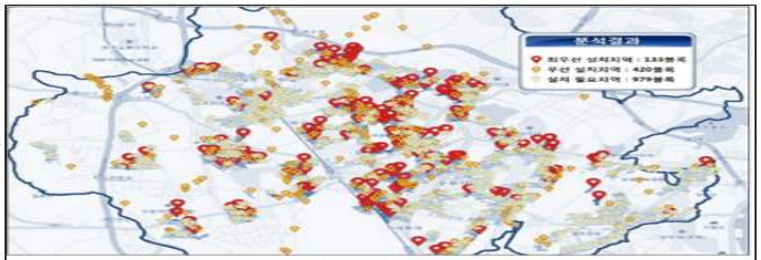
3. 교육과정 개편

4 공공 분야 빅데이터 적용 사례(#1)

≫경기도 CCTV 설치지역 분석 및 모델 고도화

구분	내용
기관명	- 경기도
배경	- 경기도 31개 시군에서는 매년 지속적으로 CCTV를 설치하고 있으나, 설치지역에 대한 객관적 선정방안 부재 - CCTV 설치 담당자가 최적화된 설치 위치를 알고 있어도 이를 뒷받침할 만한 객관적 데이터가 부재 - 대부분 민원에 근거하여 특정 지역에만 집중적으로 CCTV가 설치됨에 따라 방범 효율성이 저하 ※ 서울특별시, 광주광역시, 울산광역시, 수원시, 천안시, 고양시 등의 지자체에서 표준분석모델을 활용 중
활용데이터	
분석내용	- CCTV 우선 설치지역을 분석하기 위해 분석 공간 확정 후 범죄 예측지수를 모델링하여 감시 취약지수를 개발하고, CCTV우선설치 지역분석 최종지수를 도출 - CCTV 속성변수 분석 및 유동인구 분포지수를 활용하여 CCTV 운영관리를 위한 분석 진행

구분	데이터명	출처	사이즈(개수)
공공	CCTV 설치현황	지자체	6,578
	CCTV 설치예정현황		620
	CCTV 설치요청 민원데이터		914
	가로등·초·보안등(방범등)설치현황	국립재난안전연구원	121,571
	개발정보	지자체	41
	어린이집현황	경기도 교육청	12,591
	학교기본현황(유치원, 초·중·고)		4,455
	용도지역지구 도시지역	국토교통부	106,262
	연속지적도	경기도	4,812,860
	하천용도지역	하천지리정보시스템	3,893
	공원	경기도청	6,288
	어린이놀이시설 현황	놀이시설안전관리시스템	14,199
	상가업소데이터	소상공인시장진흥공단	135,361
	가중치 항목의 중요도	지자체	77
민간	경기도 내국인 유동인구	민간업체	4,129,517
	경기도 외국인 유동인구		4,129,517
	도로링크		1,946,621
	아파트 단지 정보		24,713
	도로명주소 건물데이터		1,579,235
	인구정보		71,740
	블록 영역		83,392
	가구정보		71,110
	거처유형별 주택정보		785,038
	표준지 공시지가		59,843
	버스정류장		146,452
	건물유형별 가구 비율		533,666

분석결과	- (CCTV 우선설치지역 분석 결과 및 시각화) 최종지수를 평가 결과에 따라 6등급으로 구분하고 범죄예측지수와 CCTV감시 취약지수가 높은 곳을 CCTV 우선설치지역으로 선정 - (CCTV 운영관리 분석 결과 및 시각화) CCTV 운영관리 목적에 따라 크게 'CCTV 성능 개선', 'CCTV 가까움'으로 구분, 'CCTV 성능 개선'은 '설치일자 오래됨', '화소 수 낮음', '개수 부족'의 지표 값으로 구분하여 시각화
활용 및 기대효과	- 도민의 CCTV 설치요청 민원데이터 및 분석 결과와 민원 데이터의 타당성을 검토하여 필요 지역에 CCTV를 우선설치 - CCTV 운영관리의 우선순위 제공을 통해 운영관리 효율성 증대 - 도민 안전관련 추가 정책 고려 시, 참고 자료로 활용
시각화	 CCTV 설치 후보지역 분포도  CCTV 우선 설치지역 분석결과

4 공공 분야 빅데이터 적용 사례(#2)

≫한국도로공사 인공지능 기반 도로포장 파손 실시간 탐지

구분	내용
기관명	- 한국도로공사 ICT센터
배경	- 도로망의 지속적인 증가로 공용기간이 상당히 경과한 노후 노선이 늘어나면서 도로시설물에 대한 유지관리 중요성 강조 - 도로 포장은 주행의 안전성과 쾌적성에 주요한 역할을 차지하므로 도로관리기관의 효율적인 유지관리가 필수 - 교통하중, 기후환경 등 다양한 요인에 의해 발생되는 포트홀 등의 도로포장 파손은 차량 파손과 교통사고의 원인이 되어 신속한 보수가 요구
활용데이터	
분석내용	- 1단계: 파손위험후보 이미지 추출을 위한 실시간 영상 이미지 분석 ∘ 촬영영상 이미지에 대한 정사영 변환, 노이즈 제거, 전처리를 통해 관심영역 추출 - 2단계: 딥러닝 모델을 통한 파손부 예측 분류 ∘ 전국 고속도로 4984km 동영상 촬영후 이미지 추출 및 분류
분석결과	- 고속도로 포장 일상유지보수 업무에 시스템 활용 추진 ∘ 파손정보 수집 및 보수물량 산출로 과학적 유지보수 수행 - 포장관리 통합시스템에 정보연계를 통한 포장상태 종합관리 가능 ∘ 정밀조사장비를 통한 포장상태조사(매 2년) 결과와의 융합 분석으로 노면상태와 내부의 구조적 상태의 체계적 분석을 통한 최적유지보수 시행 - 안전한 도로주행을 위한 실시간 도로안전정보 제공 가능 ∘ C-ITS* 및 자율주행 도로환경 도로상태에 대한 실시간 안전정보제공 * C-ITS(Cooperative Intelligence Transportation Systems): 차세대 지능형교통시스템으로 차량주행 중 운전자에게 급정거, 낙하물 등의 사고위험정보를 실시간으로 제공하는 시스템 ∘ 공공데이터 공유를 통한 융합 교통정보(경로안내시 안전도로 안내 등) 서비스 제공
활용 및 기대효과	- 도로포장 일상점검의 혁신을 통한 스마트 유지관리체계 실현 - 도로파손부 신속한 유지보수로 고객 피해 최소화 및 교통사고 예방 - 도로안전정보 제공을 통한 도로 주행안전도 향상 - 인공지능 기반의 실시간 도로파손탐지 기술 실용화(특허등록 진행 중)로 도로유지관리 기술 선도 - 기술 실용화에 따른 경제적 효과로 5년 기준 총 415억 원 절감
시각화	

IV 2018년 KERIS 빅데이터 활용 업무추진 현황

IV 2018년 KERIS 빅데이터 활용 추진 현황

1. 나이스 연계 교육문제 현안 해결 가능성 탐색 연구

2. 빅데이터를 활용한 학업중단 학생 대응 모델 연구

3. 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석

1 나이스 연계 교육문제 현안 해결 가능성 탐색 연구(#1)

▪ 연구개요

- 연구명 : 나이스연계 교육문제 현안 해결 가능성 탐색 연구
- 연구기간 : 2018. 6 – 2018. 10(5개월)
- 연구수행기관 : 건양대학교

1 나이스 연계 교육문제 현안 해결 가능성 탐색 연구(#2)

- 배경
  - 교육은 사회 현안에 민감한 분야라는 특성이 있어 데이터를 활용한 사회 현안 예측 및 대응을 위한 거시적 차원의 연구 수행이 필요함
  - 특히 나이스는 지난 10여 년간 데이터를 축적하고 있어 과거 개인정보 이슈로 민감했던 인식의 변화 확인 필요

- 목적
  - 최근 빅데이터 활용의 가치 확산으로 나이스 데이터 활용에 대한 인식도 파악
  - 교육 데이터 분석을 통해 지능정보사회에서 새로운 가치 창출과 교육문제 현안 해결에 기여할 수 있는 방안을 모색

1 나이스연계 교육문제 현안 해결 가능성 탐색 연구(#3)

▪ 나이스 데이터 활용 인식 조사 : 2,437명

※ 조사 기간 : 2018. 8.23 ~30(교육부, 교육청, 각급 학교 대상 공문협조에 의한 온라인 설문)

구분		참가자(명)	비율(%)
나이	20대	210	8.6
	30대	669	27.5
	40대	742	30.4
	50대	753	30.9
	60대 이상	63	2.6
조직 구분	교육부	15	0.6
	시도교육청	148	6.1
	각급 학교	2,237	91.8
	기타 기관	37	1.5
담당 업무	정책입안자	9	0.4
	학교관리자(보직자)	249	10.2
	교사- 직원	1,974	81.0
	기타업무	205	8.4

1 나이스연계 교육문제 현안 해결 가능성 탐색 연구(#4)

- 데이터를 활용한 교육현안 해결을 위한 주요 주제 도출(20개)

도출 주제	도출 주제
공기질 개선	학교 운영 예산
식중독 예방	학생 생활 건강
위기 학생 지원	진로 교육
다문화 지원	농산어촌 학교 지원 시설
저소득층 아동	교육복지
한 부모 가정 지원	교육재정
학교 폭력	저출산 고령화
교육시설 안전	학교 운동장 유지 관리
학교 통폐합/신설 예측	학교 정보화 사업 추진
보조교사 수요 예측	보안 및 개인정보

1 | 나이스 연계 교육문제 현안 해결 가능성 탐색 연구(#5)

≫**예측모델(개발중)과 데이터셋 예시**

1. 사회적 약자의 기본 학습권 보장을 위한 2중 언어 수요 예측 모델

Data set
- 다문화 학생
- 언어
- 연령

예측 결과
- 언어별 교사 수
- 언어별 예산
- 부족 시기 예측

2. 안심하고 안전하게 다니는 학교(교실 공기질 예측 모델)

Data set
- 학생수
- 미세먼지 농도
- (학교주소)
- 교실내 공기질 측정값

예측 결과
- 학교 및 교실 별 미세먼지 위험도 예측결과
- 지원 우선순위 설정
- 교실 환기 시점 및 시간

3. 안심하고 안전하게 다니는 학교(급식/영양관리 강화)

Data set
- 메뉴
- 조리법
- 식중독 지수

예측 결과
- 식중독 위험도 예측
- 레시피 추천

4. 보건실 활용 현황을 통한 질병 예방 시점 예측(학생 생활 건강)

Data set
- 보건일지
- 응급처치

예측 결과
- 질병 위험도 예측
- 질병 예방 교육 시점 예측

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#1)

▪ 연구개요

- 연구명 : 빅데이터를 활용한 학업중단 학생 대응 모델 연구
- 연구기간 : 2018. 7. 12 ~ 12. 14 (약 5개월)
- 연구수행기관 : 이화여자대학교

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#2)

- 배경
  - 학업중단은 개인적으로 청소년의 사회적 자립 및 성장을 저해하고 사회적으로는 인적자원 손실 등 문제를 야기하므로 학업중단을 예방하는 체계적인 관리 체계를 구축하기 위해서 학업중단 위기에 놓인 학생들을 조기에 발견하는 것이 무엇보다 중요

- 목적
  - 빅데이터를 활용한 학업중단 학생 대응 모델 연구를 통해 잠재적 학업중단 위기학생을 사전에 파악하고 학교·가정·사회에서 학업중단 예방을 위해 선제적으로 대응

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#3)

≫**연구의 내용 및 방법**

- 국내 학업중단 예방 제도 및 프로그램 분석
- 해외 학업중단 예측·관리를 위한 제도 및 프로그램 분석
- 학업중단 학생 대응을 위한 빅데이터 활용 주요 요소 도출
- 학업중단 위기학생의 효율적인 대응을 위한 모델 제안

- 문헌 및 선행 연구 분석
- 국내외 학업중단 예측 모형 사례조사
- 빅데이터 분석 : HGLM 및 랜덤포레스트 분석
- 전문가 자문단 구성 및 협의회 개최

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#4)

≫**해외 학업중단 예측 · 관리를 위한 프로그램(미국 EWS)**

- Early Warning System에서는 학업중단 위기학생을 예측하기 위해 5가지 위기 지표를 활용함

지표	기준 (flag)
출석 (Attendance)	수업시간의 10% 또는 그 이상 결석
과정 이수 (Curse Faiures)	한 학기 이상 교육과정 이수 실패
학점 (Grade point average)	2.0점 이하 (4.0점 기준)
CCSR on-track indicator	두 학기 이상 핵심 교육과정 이수 실패, 또는 진급기준에 미달 성적
행동 (Behavior)	지역별 정의 및 기준

[EWS 지표 및 기준]

*CCSR on-track indicator는 시카고 교육청에서 '학점'과 '성적'을 조합하여 개발한 조기경보지표

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#5)

≫해외 학업중단 예측 · 관리를 위한 프로그램(호주 Student Mapping Tool)

- 호주 빅토리아 교육부에서는 위기학생(disengaged students)을 진단할 수 있는 진단 시스템을 개발하여 운영
- Student Mapping Tool(SMT) 컴퓨터 소프트웨어를 통해 다양한 자료를 종합하여 위기학생 및 잠재 위기학생 판별

학생의 위기 여부를 판단할 수 있는 근거자료	
가정배경 자료	학생 입학시 수집되는 가정에 대한 기본적인 배경정보
출석 자료	학생의 학교 결석 유형 및 기간
정부 자료	학생과 학부모의 동의하에 정부로부터 받을 수 있는 교육, 건강, 복지 관련 종합 자료
교사 관찰자료	학생의 학급생활이나 수업시간의 학습활동, 행동의 특이성이나 변화 등을 기록할 수 있는 자료

2 | 빅데이터를 활용한 학업중단 학생 대응 모델 연구(#6)

≫해외 학업중단 예측 · 관리를 위한 프로그램(호주 Student Mapping Tool)

- 호주 빅토리아 주정부는 2004년 SMT를 개발하여, 2006년 시범적용을 통해 2007년부터는 모든 중등학교에 적용할 수 있도록 보급
- SMT는 프로그램화된 엑셀파일(pre-programed execl)로 호주 학업성취도 정보공시홈페이지(NAPLAN)에 **학교, 학년 등의 기본 정보를 입력하면 지역과 학년별로 특화되어 프로그램 되어 있는 엑셀 파일이 재구성되어 다운로드** 받을 수 있도록 개발
- 다운로드 받은 엑셀파일에 다양한 항목을 수치 및 알파벳으로 입력하면, 위기 가능성이 있는 분야의 개별 학생들이 **노란색으로 표시되어 교사나 학교의 관리 및 지도의 대상으로 진단하는 도구**

3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#1)

- 배경 및 목적
  - 대기오염으로 촉발된 공기질에 관한 공공인식이 제고되면서 실내공기질 (Indoor Air Quality)에 대한 관심도 높아지고 있음
  - 국내외적으로 공기질이 학생의 학업성취에 영향을 끼친다는 인식 확산
  - 향후 빅데이터 기반 학교 내 공기질 관리 체제 마련 필요
- 추진내용
  - 학교 내 공기질 관리에 관한 선행연구 분석
  - 공기질 관리에 관한 정책 추진 현황 분석
  - 케이웨더 및 관련 기업 빅데이터 분석 기술동향 분석

3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#2)

▪ [선행연구 분석] 교실 실내공기질의 학생 영향도 확인 및 개선 방안 도출

- ✓ 학생과 교직원의 건강증진 및 쾌적한 학교환경위생을 유지하기 위해 교실내의 공기오염의 종류, 발생원, 인체에의 영향, 관리대책 제안(김윤신, 1991)
- ✓ 학생 설문조사 결과 건강에 가장 큰 해를 끼치는 환경오염으로 실내공기오염이 대기오염보다 심각하다고 인식(김인환, 2005)
- ✓ 초등학교 교실 실내공기질의 현장측정을 통하여 실내공기질을 개선할 수 있는 방안 제안 및 실내공기질의 기초자료를 제시(류지응, 2008)
- ✓ 교실 실내 공기중 유해유기물질의 농도와 학생들의 신경행동기능과 관련성을 평가한 결과, 적으나마 통계적으로 유의한 인지기능 저하가 관찰됨(곽홍탁, 2011)
- ✓ 실내 공기의 높은 이산화탄소 농도는 학생의 주의집중을 방해하며, 문제풀이에서 틀릴 가능성을 높이고, 문제풀이 속도를 느리게 만듦(임완철, 2015)

3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#3)

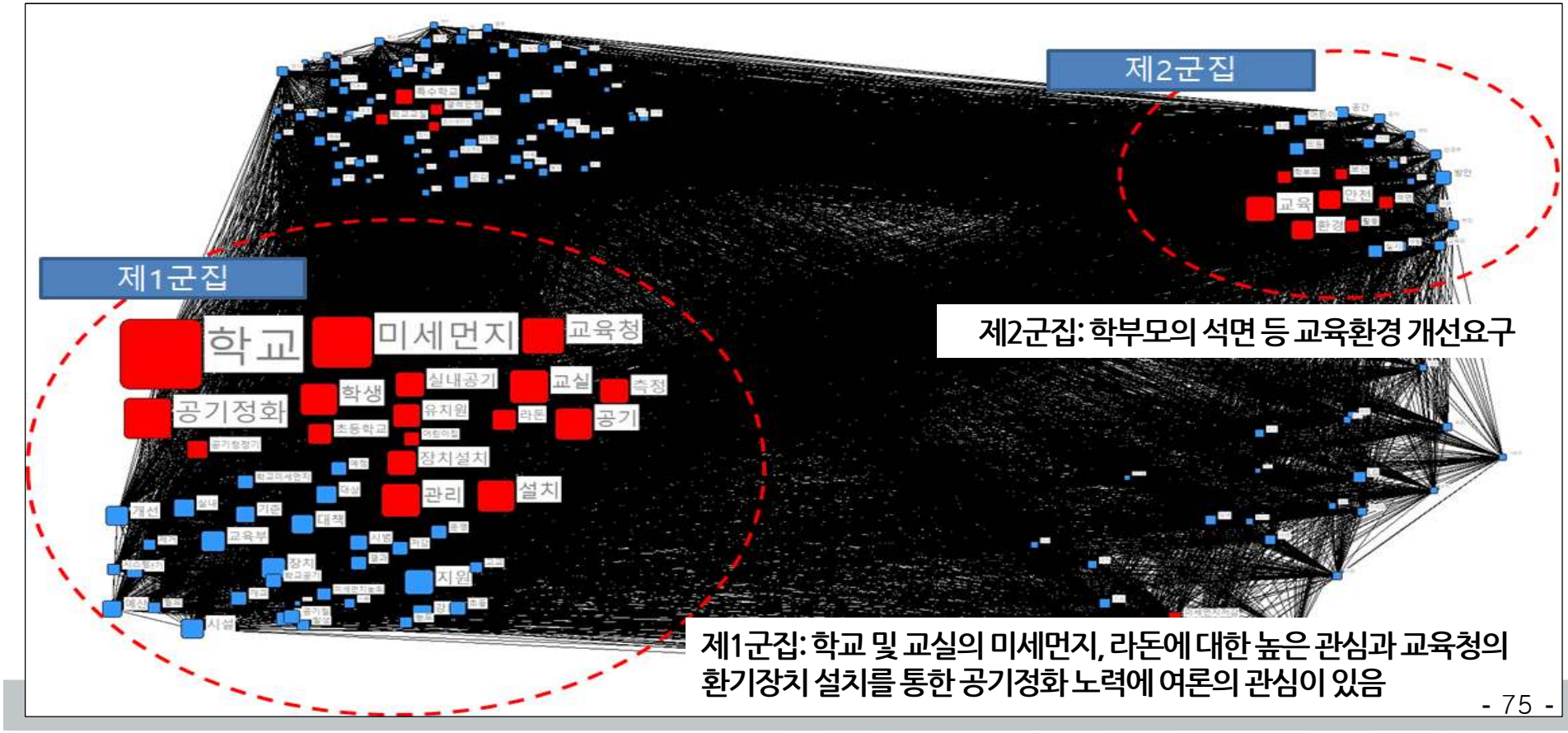
▪ [정책동향 분석] 실내 공기질 관리 관련 정책 추진현황 (3가지)

- (과기부, 환경부, 보건복지부 : 미세먼지 국가전략프로젝트 사업단) 미세먼지의 발생유입, 측정예보, 집진저감, 보호대응 등 4대 분야의 연구개발을 추진하는 미세먼지 국가전략프로젝트 사업단이 2017년 8월에 발족
- (국토부, 과기부 : 스마트시티 국가전략프로젝트 연구개발 실증사업) 실증도시로 선정된 대구에서 미세먼지 등 생활 문제 해결을 위한 리빙랩 *(Living Lab) 구축 운영 예정
- (행안부 : 스마트시티 국가시범도시) 시범도시(부산 에코델타시티, 세종 5-1 생활권)의 기본구상안에 도시문제 해결을 위한 리빙랩 조성, 미세먼지 관련 시민 안전서비스 등 공기질 관련 내용이 포함

***일반적인 실험실 연구와 달리 생활현장에서 시민들의 참여로 진행하는 연구 방식**

3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#4)

- [소셜네트워크 분석]“ 학교공기질” 키워드를 활용하여 최근 1년간 (‘17.8.30~‘18.8.30) 뉴스 데이터 분석



3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#5)

▪ [기술동향 분석]

- (측정 장비) 학교 단위 측정시 기본적으로 실외측정기 1대, 실내측정기 3~5대, DID(모니터링 장비), S/W가 필요
- (분석) 실외와 실내의 공기질 상태를 비교하여, 모니터링 장비 및 모바일 등을 통해 적절한 조치 안내
- (조치) 실내 공기질 개선을 위해서는 반드시 기기가 필요한 것이 아니라 환기만으로도 충분한 해결책이 될 수 있음

3 | 학교 내 공기질 관리를 위한 빅데이터 기술 동향 분석(#6)

≫일선학교 실제 구축사례(케이웨더)

공기질 관제화면

실외 측정기

실내 측정기

Q & A