



인공지능 입문

Part 02. 인공지능과 기술

Chapter 07. 빅데이터

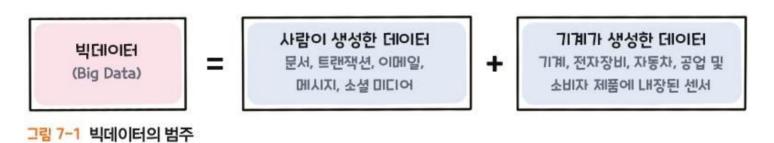
목차

- 1. 빅데이터의 개요
- 2. 빅데이터와 인공지능
- 3. 빅데이터 활용에 대한 규제와 딜레마
- 4. 빅데이터의 개방

01 빅데이터의 개요

Ⅰ. 빅데이터의 개념

- 빅데이터(Big Data)
 - 디지털 환경에서 생성되는 데이터
 - 과거 아날로그 환경에서 생산되던 데이터에 비해 그 규모가 방대하고 생성 주기도 짧으며 수치 데이터뿐만 아니라 문자와 영상 데이터를 포함하는 대규모데이터



I. 빅데이터의 개념

- 빅데이터와 기존 일반 데이터의 차이점
 - 빅데이터는 다양한 방법과 원천, 환경으로부터 수집된 데이터 의미
 - 빅데이터는 병렬 처리를 위한 컴퓨터 시스템이 필요할 정도로 큰 데이터임
 - 빅데이터는 비즈니스 혹은 연구에서 유용한 가치 창출
 - 빅데이터가 창출하는 가치를 보장하기 위해서는 타당성(Validity)과 신뢰성 (Veracity)을 확보해야 함

표 7-1 일반 데이터와 빅데이터의 차이

구분	일반 데이터	빅데이터	
데이터의 원천	내부로부터 수집	외부로부터 수집	
데이터의 형태	정형 데이터가 대부분	비정형 데이터	
분석 방법	모델링	인공지능	
분석 환경	기업 내에 구축된 데이터웨어하우스	클라우드	

I. 빅데이터의 개념

하나 더 알기 데이터 단위

- 비트(Bit) : 가장 작은 데이터 단위
- 8비트 = 1바이트(Byte), 영어나 숫자는 1바이트, 한글은 2바이트
- 1킬로바이트(KB) = 1,024바이트(Byte)
- 1메가바이트(MB) = 1,024킬로바이트(KB)

표 7-2 데이터 단위						
이름	기호	값	이름	기호	값	
킬로바이트	KB	10241 = 210	페타비어트	PB	10245 = 250	
메가바이트	MB	10242 = 220	엑사버이트	EB	10246=260	
기가비어트	GB	1024 ³ = 2 ³⁰	제타바이트	ZB	10247 = 270	
테라바이트	ТВ	10244=240	요타바이트	YB	10248 = 280	



빅데이터의 특징

- 1) 3V:크기 + 속도 + 다양성
- 크기(Volume)
 - 물리적 장치에 저장되는 데이터의 양
- 속도(Velocity)
 - 데이터의 실시간 처리를 보장할 수 있어야 한다는 것
- 다양성(Variety)
 - 다양한 형태의 데이터를 포함하는 것

Ⅱ. 빅데이터의 특징

2) 4V: 3V + 신뢰성

신뢰성(Veracity)

데이터가 얼마나 가치 있고 유용한지를 나타내는 것으로, 빅데이터를 분석하는 데
 있어 기업이나 기관에서 수집한 데이터의 정확성을 살펴보는 것

3) 5V: 4V + 가치

- 빅데이터는 비즈니스에 활용되어 가치를 이끌어 낼 수 있어야 그 의미가 있음
- 빅데이터를 설계하고 수집하기 전에 해당 데이터를 활용하여 무엇을 할 수 있을지
 에 대한 고민이 필요

Ⅱ. 빅데이터의 특징

- 4) 6V:5V + 타당성
- 타당성(Validity)
 - 데이터의 정확성을 의미

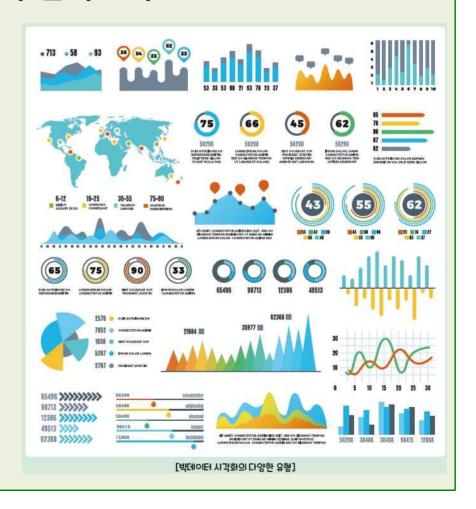
- 5) 7V: 6V + 휘발성
- 휘발성(Volatility)
 - 데이터를 얼마나 오래 저장하고 사용할 수 있는지에 관한 것

Ⅱ. 빅데이터의 특징

하나 더 알기 빅데이터의 특징 좀 더 알아보기

박데이터는 추가적으로 가변성과 시각화의 특징이 있음

- 가변성(Variability) : 데이터의 맥락에 따라 의미가 달라지는 것
- 시각화(Visualization)
 - : 데이터를 시각적으로 표현하는 것

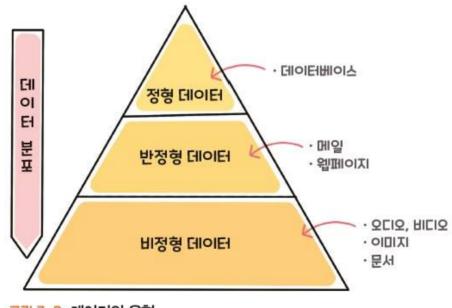


Ⅲ. 빅데이터의 유형



· 빅데이터의 유형

- 정형 데이터
- _ 비정형 데이터
- 반정형 데이터



III. 빅데이터의 유형

1) 정형 데이터

- 정형 데이터(Structured Data)
 - 일정한 형식이나 규칙(Rule)에 맞게 저장된 데이터
 - 가장 쉽게 접할 수 있는 유형은 스프레드시트
 - 정형 데이터의 예 : 스프레드시트, 관계형 데이터베이스, CSV 등

ID	Name	Age	Degree
1	John	18	B.Sc.
2	Jason	32	Ph.D.
3	Robert	52	Ph.D.
4	Ricky	35	M.Sc.
5	Gibb	26	B.Sc.

그림 7-4 정형 데이터의 예

III. 빅데이터의 유형

2) 비정형 데이터

- 비정형 데이터(Unstructured Data)
 - 정형 데이터와 반대되는 개념으로, 정해진 규칙이 없어서 값의 의미를 쉽게 파악하기 힘든 경우가 많음
 - 인공지능 기술의 발전과 함께 비정형 데이터로부터 인사이트(Insight)를 얻는
 사례가 많아지면서 비정형 데이터에 대한 중요성이 부각되고 있음
 - 비정형 데이터의 예 : SNS, 영상, 이미지, 음성, 텍스트 등

Artificial intelligence (AI), is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality. The distinction between the former and the latter categories is often revealed by the acronym chosen, 'Strong' AI is usually labelled as AGI(Artificial General Intelligence) while attempts to emulate 'natural' intelligence have been called ABI(Artificial Biological Intelligence).

III. 빅데이터의 유형

3) 반정형 데이터

- 반정형 데이터(Semi-structured Data)
 - 스키마(Schema)와 메타데이터(Metadata)의 특성을 갖는 데이터
 - 반정형 데이터가 주목을 받는 이유는 인터넷에 존재하는 수많은 데이터를 인공
 지능에 사용할 수 있기 때문
 - 반정형 데이터의 예 : XML, JSON, NoSQL, 로그 등

Ⅲ. 빅데이터의 유형

하나 더 알기 XML, JSON, NoSQL

1. XML(eXtensible Markup Language)

XML은 확장 가능한 마크업 언어라는 뜻으로, HTML처럼 태그(Tag)들을 고정시켜 사용하지 않고 확장이 가능함.

2. JSON(JavaScript Object Notation)

JSON은 데이터를 저장하거나 전송할 때 많이 사용하는 데이터 교환 형식.

```
{
    "employees": [
        {
            "name": "Surim",
            "lastName": "Son"
        },
        {
            "name": "Someone",
            "lastName": "Huh"
        },
        {
            "name": "Someone else",
            "lastName": "Kim"
        }
    ]
}
```

III. 빅데이터의 유형

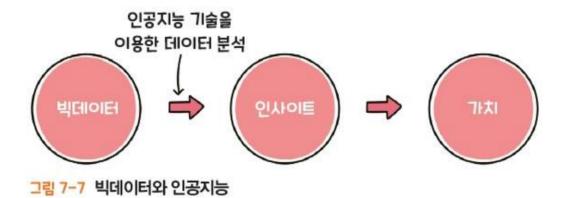
하나 더 알기 XML, JSON, NoSQL

3. NoSQL(Not Only SQL)

NoSQL은 스키마의 특성을 가지며, 유연하고 분산 병렬처리가 쉬워 확장에 유리함. 빅데이터를 다룰 때 많이 사용되는 데이터베이스

비데이터와 인공지능의 관계

- 인공지능은 빅데이터에 대한 분석 방법을 제공
- 인공지능 기술을 이용하면 어떤 사용자라도 쉽고 빠르게 인사이트를 찾아낼수 있음



I. 빅데이터와 인공지능의 관계

1) 인공지능의 주요 원료인 빅데이터

- 인공지능 기술은 빅데이터와 결합되어 대량의 데이터 학습을 진행
- 이를 바탕으로 의사결정을 위한 인사이트 제공 및 미래 발생 상황 예측 가능

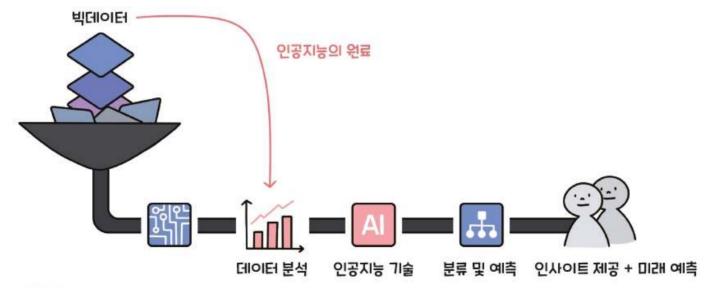


그림 7-8 인공자능의 주요 원료인 빅데이터

비데이터와 인공지능의 관계

2) 인공지능과 융합하는 빅데이터

- 빅데이터와 데이터 분석 기능을 통해 비즈니스 가치를 창출할 때 가장 주목해야 할 기술로 '인공지능과 빅데이터의 융합'

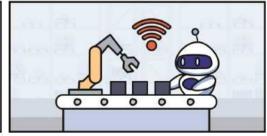


그림 7-9 인공지능과 빅데이터 융합

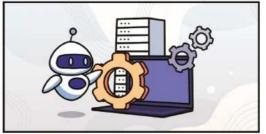
- 비데이터와 인공지능의 관계
 - 2) 인공지능과 융합하는 빅데이터
 - 빅데이터와 인공지능의 융합으로 인한 변화
 - 사물인터넷으로의 연결
 - 초연결 산업 생태계로의 전환
 - 사이버 시스템과의 융합
 - 스마트 머신 등장



사물인터넷으로 연결



초연결 산업 생태계로 전환



사이버 시스템과 융합



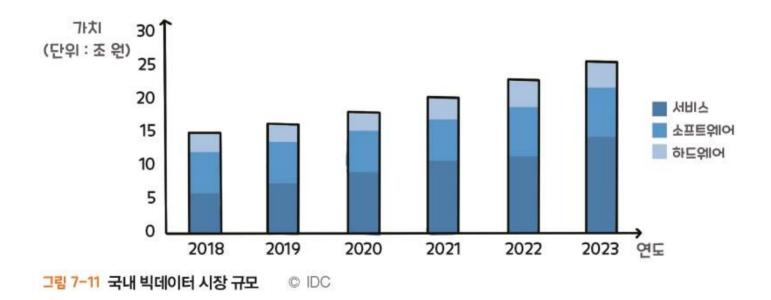
스마트 머신 등장

그림 7-10 빅데이터와 인공자능의 융합으로 변화될 세상

I. 빅데이터와 인공지능의 관계

2) 인공지능과 융합하는 빅데이터

- 빅데이터 플랫폼을 소유하는 국가나 기업이 4차 산업혁명 시대를 지배할 것



Ⅱ. 빅데이터와 인공지능의 활용

- 1) 금융 산업
- 로보어드바이저 (Robo-Advisor)
 - '로봇(Robot)'과 '투자전문가(Advisor)'의 합성어
 - IT기술과 금융이론이 결합되어 컴퓨터가 사람을 대신해 자산을 관리하는 기술



Ⅱ. 빅데이터와 인공지능의 활용

하나 더 알기 핀테크

- 핀테크(FinTech): 금융에 IT기술을 접목하여 복잡하고 어려웠던 금융 업무를 효율적이면서 편리하게 서비스하는 것
- **핀테크 관련 서비스 :** 페이코(PAYCO), 토스(Toss), 카카오페이(KakaoPay) 등



Ⅱ. 빅데이터와 인공지능의 활용

2) 헬스케어 산업

- IBM 왓슨(Watson): 헬스케어 산업에서 빅데이터+인공지능 결합 사례
 - 병원 진료기록, 학계 논문, 생체 데이터 등을 통합 분석하여 환자의 치료 도움
 - 하지만 왓슨의 도입이 업계에 큰 이득을 가져다주진 못해 현재는 해당 서비스를 중단하고 임상시험 서비스로 전환됨



그림 7-13 IBM 왓슨

Ⅱ. 빅데이터와 인공지능의 활용

3) 유통 산업

- 유통 산업에서는 이미지를 이용한 비주얼 검색과 온 · 오프라인 매장에 도입된
 채팅봇이 고객을 직접 응대하는 서비스가 확대되고 있음
- **아마존고(amazon Go) :** 아마존(Amazon)이 운영하는 무인 마트



그림 7-14 아마존고

Ⅱ. 빅데이터와 인공지능의 활용

4) 전자 산업

- 최신 AI 가전제품은 사용자의 생활 습관이나 공간의 특성 등을 스스로 학습하여최적의 작동 방식을 찾아냄
- 삼성전자 무풍 에어컨 : 사용자의 패턴을 학습하는 기능을 갖추고 있음
- LG전자 트롬 세탁기 : 사용자의 패턴을 학습한 후 상황에 맞는 세탁 옵션 추천

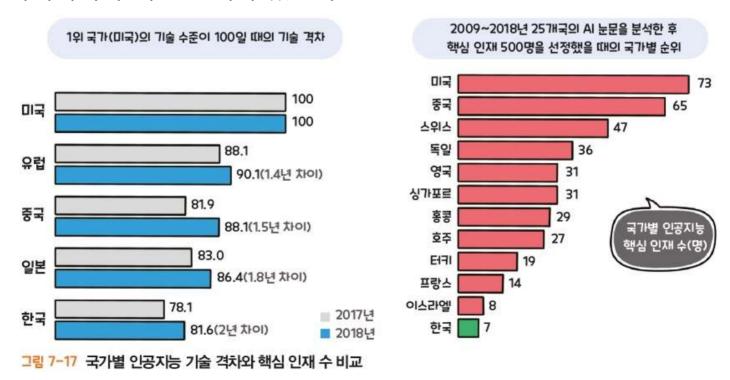


그림 7-15 삼성 전자 무풍 에어컨



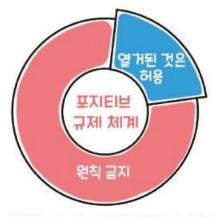
그림 7-16 LG전자 트롬 세탁기

- 우리나라는 4차 산업혁명 핵심 분야 경쟁력이 미국과 유럽 등의 국가에 비해 떨어진다는 평가를 받고 있음
- 이에 대한 원인 중 하나로는 규제 체계를 들 수 있는데, 그렇다면 규제 체계는 빅데이터와 어떤 관계가 있을까?



Ⅰ. 빅데이터에 대한 규제 체계

- 빅데이터에 대한 규제 체계
 - 포지티브(Positive) 규제 : 원칙적으로 금지하면서 예외는 허용하는 방식
 - 네거티브(Negative) 규제 : 원칙적으로 허용 하면서 예외는 금지하는 방식
 - » 네거티브 규제는 법령 등 규정을 제외한 나머지 활동은 모두 허용하기 때문에 기존에 없었던 기술이나 제품을 개발하고 사업화하는 데 유리함

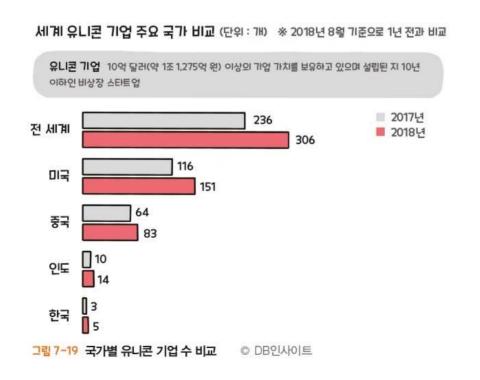






- 현재 우리나라 규제 체계는 포지티브 방식을 따르고 있음
- 포지티브 규제 방식은 사회 ·경제 규모가 작고 단일 기술과 단일 산업이 주종을 이뤘던 시대에는 유효하게 작동했음
- 하지만 새로운 기업들이 빠른 속도로 생겨나고 산업 구조가 복잡해지자 포지 티브 규제가 오히려 발전의 걸림돌이 되고 있음

- 중국의 빅데이터에 대한 규제 체계
 - 네거티브 규제 방식을 채택 중
 - 스타트업이 유니콘 기업으로 성장하는 경우가 많아지고 있음
 - 새로운 아이디어를 다양한 방법으로 시도할 수 있는 환경이 마련되어 있음



- 최근 우리나라도 네거티브 규제로 전환하려는 시도가 이루어지고 있음
- 대표적인 사례가 드론으로, 기존에는 항공촬영이나 관측 등의 분야에만 한정
 됐던 드론의 사용 범위가 사실상 전 분야로 확대되었음





그림 7-20 드론의 활용 범위 확대

II. 빅데이터 활용과 개인정보보호의 딜레마

- 개인정보는 기업과 정부의 경쟁력을 높이는데 매우 유용하게 활용될 수 있음
- 빅데이터에 포함된 개인정보는 보호의 대상이지만, 잘 활용할 수만 있다면 우리
 사회를 더 풍요롭게 만드는 핵심자원
- 하지만 그와 동시에 빅데이터 활용과 개인정보보호라는 딜레마를 마주해야 함

II. 빅데이터 활용과 개인정보보호의 딜레마

1) 개인정보의 개념

• 개인정보

- -「개인 정보 보호법」제2조 제1호 : 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보
- 유럽연합(EU)의 GDPR : 식별되었거나 또는 식별 가능한 자연인(정보주체)과관련된 모든 정보





그림 7-22 EU GDPR 35 / 49

Ⅱ. 빅데이터 활용과 개인정보보호의 딜레마

하나 더 알기

데이터3법

- 데이터3법:데이터 이용을 활성화하는 「개인정보 보호법」, 「정보통신망 이용 촉진 및 정보보호 등에 관한 법률(정보통신망법)」, 「신용정보의 이용 및 보호에 관한 법률(신용정보법)」 등 3가지 법률
 - 개인정보 보호법 : 이름, 주민등록번호, 영상과 같이 개인을 식별할 수 있는 정보의 범위, 법 적용 대상, 수집 및 이용 기준 등을 정함
 - 정보통신망법: 정보통신망을 안전하고 활발하게 이용할 수 있도록 유도하고 이용자의 개인정보를 보호하기 위한 내용을 정의
 - 신용정보법 : 금융 거래 등 상거래에서 개인의 신용을 판단할 때 필요한 신용정보의 이용 범위와 수집, 관리 방법 등을 규정

II. 빅데이터 활용과 개인정보보호의 딜레마

2) 빅데이터 활용과 개인정보보호

• 가명처리

- 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것
- 가명처리된 개인정보인 가명정보를 빅데이터에 활용할 수 있게 되자, 빅데이터 활용과 개인정보 보호 문제는 더욱더 이슈가 되고 있음

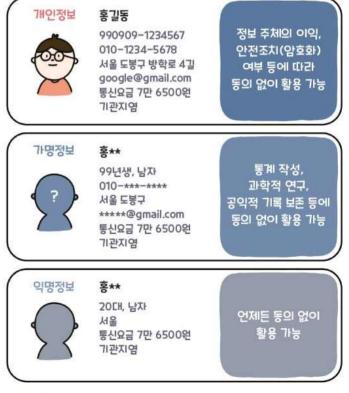


그림 7-23 개인정보, 가명정보, 익명정보의 활용 범위

Ⅱ. 빅데이터 활용과 개인정보보호의 딜레마

- 2) 빅데이터 활용과 개인정보보호
- 가명처리가 이슈가 되는 이유
 - 개인정보는 인공지능 기술을 이용하여 처리되는 과정에서 노출 위험성이 있음
 - 익명 또는 가명처리된 개인정보라 할지라도 관련 데이터의 양이 증가할수록 재 식별 가능성이 높아지기 때문
 - 개인의 입장에서는 공개되어도 무관한 데이터더라도 빅데이터 분석을 통해 개 인정보 주체의 의도와 상관없이 민감한 정보가 추론될 수 있음
 - » 즉, 가명처리를 했다 하더라도 빅데이터가 인공지능 기술과 결합하면 개인을 식별할 수 있는 정보 추출이 가능한 상황이 발생할 수 있는 것

Ⅱ. 빅데이터 활용과 개인정보보호의 딜레마

2) 빅데이터 활용과 개인정보보호

- 개인정보를 보호하면서 데이터를 활용할 수 있는 균형 있는 방안에 대해 모색 할 필요가 있음

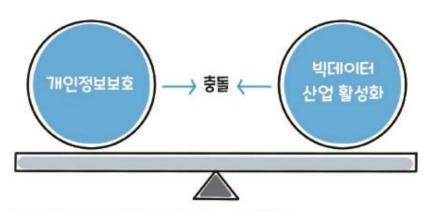


그림 7-24 빅데이터와 개인정보보호의 충돌

II. 빅데이터 활용과 개인정보보호의 딜레마

- 2) 빅데이터 활용과 개인정보보호
- '데이터3법'이 개정된 이후 정보 주체의 동의 없이 개인 정보를 추가적으로 이용 및 제공할 때에는 다음 4가지 요건을 모두 갖춰야 함
 - 당초 목적과의 상당한 관련성이 있을 것
 - 추가 이용에 대한 영향도 예측이 가능할 것
 - 제3자의 이익을 침해하지 않을 것
 - 가명처리를 의무화할
- 개정은 되었으나 여전히 개인정보를 빅데이터에 활용하기에는 제약이
 많은 상황이라 우리나라의 빅데이터 발전 속도는 더디다고 볼 수 있음

Ⅱ. 빅데이터 활용과 개인정보보호의 딜레마

- 3) 개인정보보호와 이용을 위한 균형 있는 제도
 - 정부가 법률과 구체적인 가이드라인을 제시하지 않는다면 빅데이터와 개인정보
 를 이용하는 기업들은 혁신적인 비즈니스모델 개발과 적용에 어려움을 겪을 것임
 - 뿐만 아니라 데이터 분석가의 역할도 중요한데, 데이터 분석 과정에서 개인정보
 가 추출되거나 유출되지 않도록 기술적 방안을 마련해야 함



그림 7-25 데이터 분석가의 중요성

04 빅데이터의 개방

. 공공데이터 개방의 필요성

- 빅데이터의 효과를 단기간 내에 얻을 수 있는 방법은 정부가 소유하고 있는
 데이터를 활용하는 것
- 정부가 빅데이터를 개방한다면 민간에서는 큰 수고 없이 서비스를 개발할 수있을 것
 - » 이러한 이유로 정부가 소유한 공공데이터의 개방이 필요

Ⅱ. 공공데이터 개방의 사례

- 미국은 2009년에 데이터 포털사이트(data.gov)를 공개하여 공공데이터를 개방
- 미국 정부의 데이터 개방 범위는 농업, 환경, 에너지 등의 분야를 포함하여 재정 투명성을 위한 예산, 지출, 계약, 공무원 급여까지 다양함



그림 7-26 미국의 데이터 포털사이트: data.gov

Ⅱ. 공공데이터 개방의 사례

- 영국은 2010년 고든 브라운 총리가 웹과 링크드 데이터의 창시자로 유명한 팀 버너스 리를 초빙해 데이터 포털사이트(data.gov.uk)를 만들면서 공공데이터 개방 시작
- 이후 2018년 '오픈데이터 찾기(Find Open Data)'라는 이름으로 개편됐음



그림 7-27 영국의 데이터 포털사이트: data.gov.uk

Ⅱ. 공공데이터 개방의 사례

우리나라는 OECD가 발표한 '2019년 OECD 정부백서'에서 공공데이터 개방 지수 0.93점(1점 만점)으로 1위를 기록했음

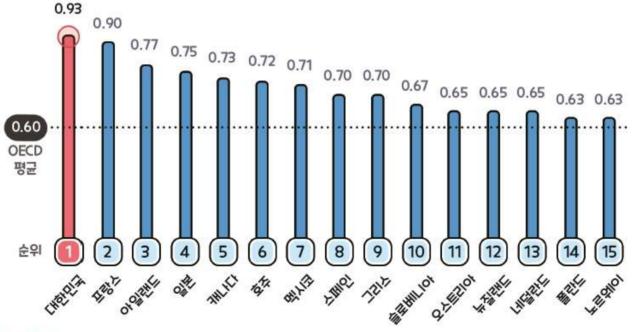
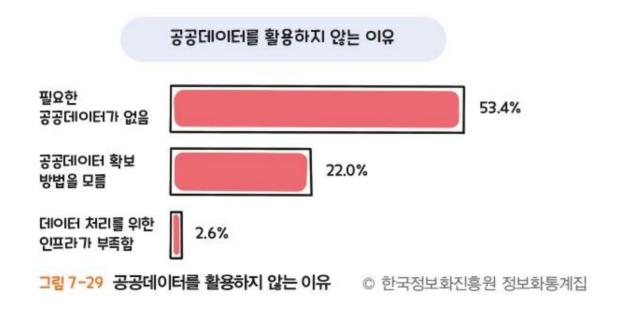


그림 7-28 전 세계 공공데이터 개방 지수

III. 공공데이터 활용을 위한 방안

우리나라의 공공데이터 개방 지수는 높은 반면, 실제 공공데이터 활용률은 그다지 높지 않은 이유는 필요한 데이터를 찾을 수 없기 때문에 활용도가 낮은 것으로 나타났음



III. 공공데이터 활용을 위한 방안

이러한 문제를 개선하기 위해 행정안전부는 향후 3년간 범정부 공공데이터 정책 방향을 제시할 '제3차(2020~22) 공공데이터 제공 및 이용활성화 기본계획' 확정

표 7-3 정부가 공개할 예정인 6개 분야 46개 과제에 대한 공공데이터

분야	개방 예정인 공공데이터
자율주행(11개)	정밀도로지도, 주행환경 인식센서 융합정보, 자율주행 딥러닝 학습 정보 등
스마트시티(6개)	스마트 전력거래, 디지털 트윈 정보, 세종시 스마트에너지 정보 등
헬스케어(8개)	해부학 그림 및 의료행위 그림 정보, 한의약 전주기 정보, 식중독균 유산균 유전체 정보 등
금융정보(5개)	상장사 공시주식 정보, 비상장사 공시 재무제표, 주택저당채권 정보 등
생활환경(7개)	굴뚝 대기오염물질 정보, 산림 미세먼지 정보, 산업부문 온실가스 배출정보 등
재난안전(9개)	구조구급활동 정보, 산사태 정보, 안전·취약시설물관리 정보, 국가화재 정보 등

Thank You!

