02강

빅데이터의 이해와 활용

빅데이터의 개요 2

방송대 통계·데이터과학과
이긍희 교수

한국방송통신대학교
Korea National Open University

학습목차

빅데이터의
이해와 활용

1 데이터 분석

1. 데이터 분석

- 20세기 데이터 분석
  - 표본조사와 실험계획법 기반 양질 데이터(스몰 데이터) 기반 성장
    - 표본조사 : 여론조사, 국가통계
    - 실험계획법 : 신약개발
  - 양질 데이터 + 공정한 통계모형 → 의미 있는 결과
    - 변수간 인과구조를 밝혀서 의미를 설명
    - 공정한 데이터 수집과 공정한 분석과정 통계학
    - 여론조사, 품질혁신, 국가통계 등 20C 발전 소프트인프라

1. 데이터 분석

< 통계학 >

- 모집단을 여러 가정과 공정한 절차 체계로 구축,
  이를 기반한 미지의 모집단을 소수의 데이터(표본)로 추정
- 통계학은 모집단의 가정, 좋은 데이터, 공정한 방법론을 통해
  가장 그럴듯한 결과를 도출

1. 데이터 분석

- 빅데이터 기반 데이터 분석 : 머신러닝
  - 머신러닝(기계학습) : 빅데이터 일부로 학습시켜 만든 모형
    - 인과구조 설명할 수 없음
    - 예측 결과가 통계모형보다 좋음

1. 데이터 분석

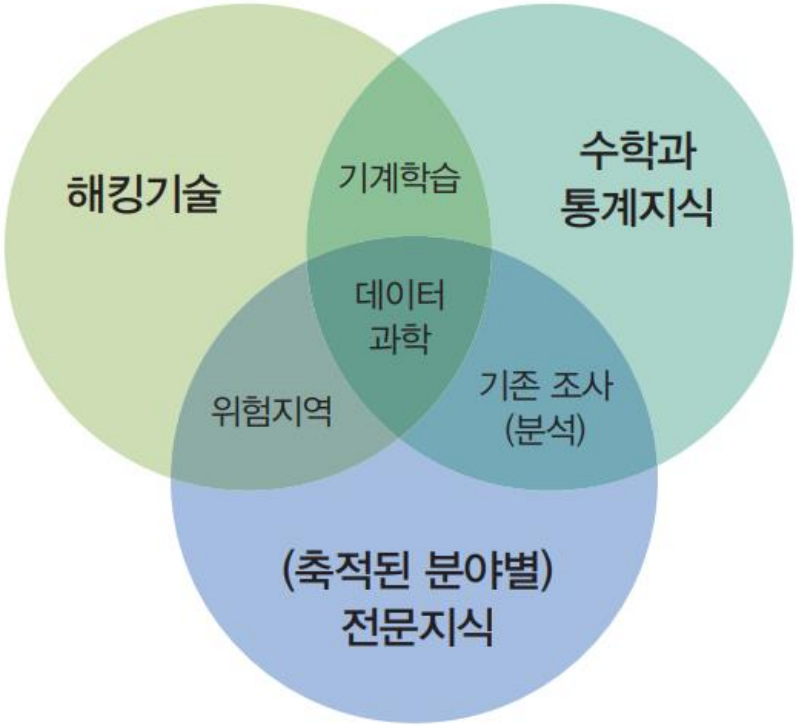
- 빅데이터 기반 데이터 분석 : 머신러닝
  - 빅데이터 시대에는 모형의 설명보다 예측력을 중시
    - 기계번역 발전의 길 : 규칙기반 → 통계기반
      → 빅데이터 딥러닝 기반
  - 빅데이터 분석 : 분산은 작지만 편의가 존재
    - 인과구조를 알 수 없고, 상관관계 아는 한계
    - 데이터과학자의 통찰이 필요

빅데이터의
이해와 활용

2 데이터 과학자

2. 데이터 과학자

- **데이터과학** : 수학·통계학, 해킹 기술(코딩 기술)과 해당 분야 전문지식이 종합된 분야로 정의



출처: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

2. 데이터 과학자

- 데이터과학자(data scientist)
  - 해당 분야 전문지식을 바탕으로 데이터를 수집·저장·가공하고, 다양한 원천의 데이터를 결합·분석하며, 이로부터 새로운 가치를 만드는 일을 하는 사람
  - <하버드 비즈니스 리뷰> : 21세기 가장 매력적인 직업
  - 가트너의 정의 : 빅데이터로부터 인사이트를 추출
    - 다양한 분야의 기술을 겸비해 팀으로 높은 성과를 내는 사람

2. 데이터 과학자

- 데이터과학자의 기술
  - 하드 스킬(hard skill) : 빅데이터 이론적·기술적 지식
    - 데이터베이스, 프로그래밍, 통계학, 딥러닝, 머신러닝, 텍스트마이닝 등
  - 소프트 스킬(soft skill)
    - 통찰력(창의적 사고, 호기심, 논리적 비판), 스토리텔링, 전달 능력, 다른 분야와 소통·협력능력

2. 데이터 과학자

데이터 관련 직무

데이터 공학자

- 데이터 수집, 보관, 저장, 관리, 정제, 컴퓨팅 환경 제공
- 프로그램 언어, 클라우드 환경, 하둡, 스파크 등

데이터 분석가

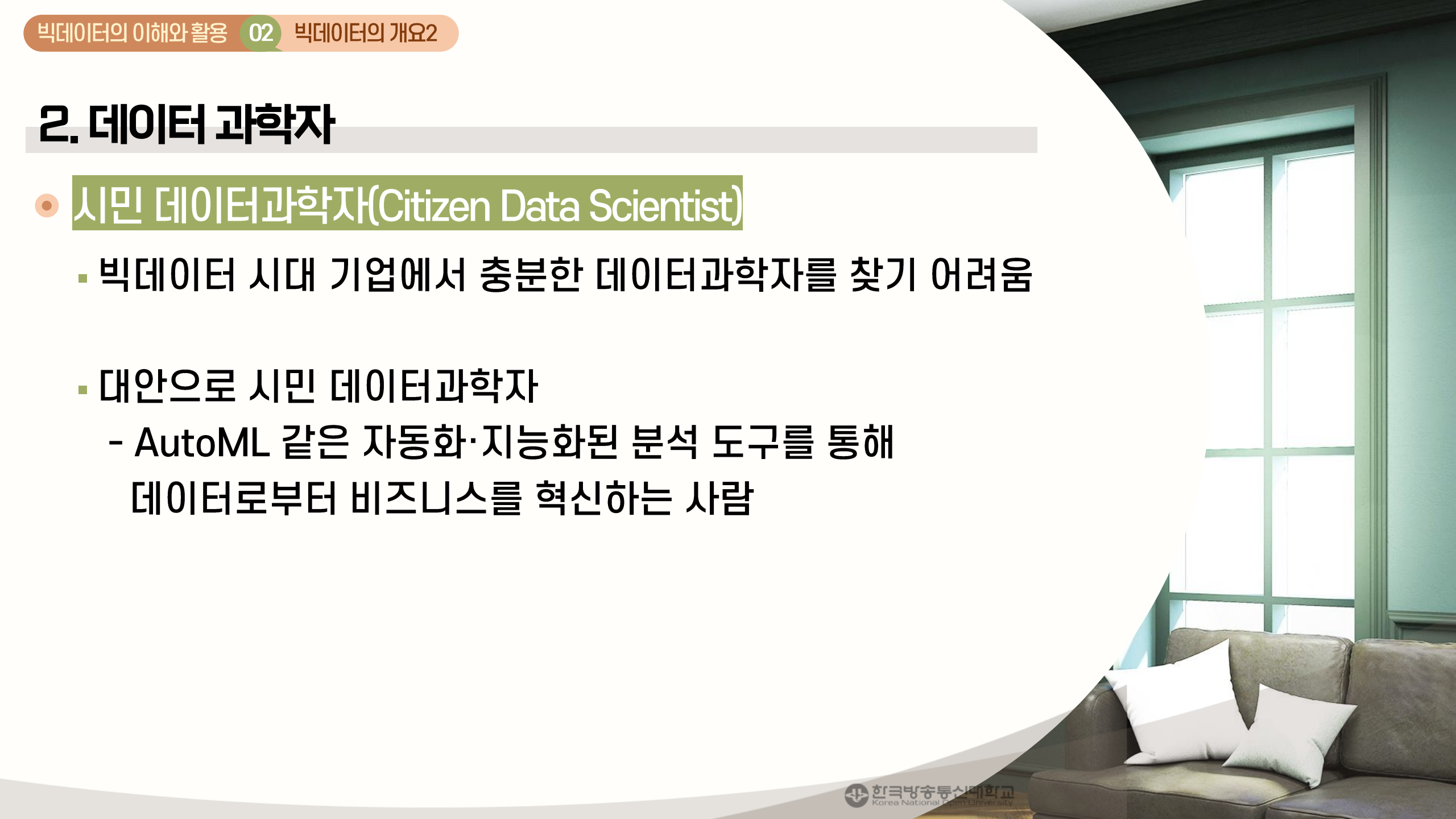
- 데이터 분석 및 시각화, A/B테스트
- 통계학, R, Python, 시각화도구, 클라우드 활용

데이터과학자

- 새로운 가치(서비스) 만들거나, 예측알고리즘 개발 등
- 머신러닝, 통계학, 코딩능력

2. 데이터 과학자

- 시민 데이터과학자(Citizen Data Scientist)
  - 빅데이터 시대 기업에서 충분한 데이터과학자를 찾기 어려움
  - 대안으로 시민 데이터과학자
    - AutoML 같은 자동화·지능화된 분석 도구를 통해 데이터로부터 비즈니스를 혁신하는 사람



▪ 스캐너데이터를 이용한 소비자물가지수 작성

발간등록번호
11-1240000-000957-01

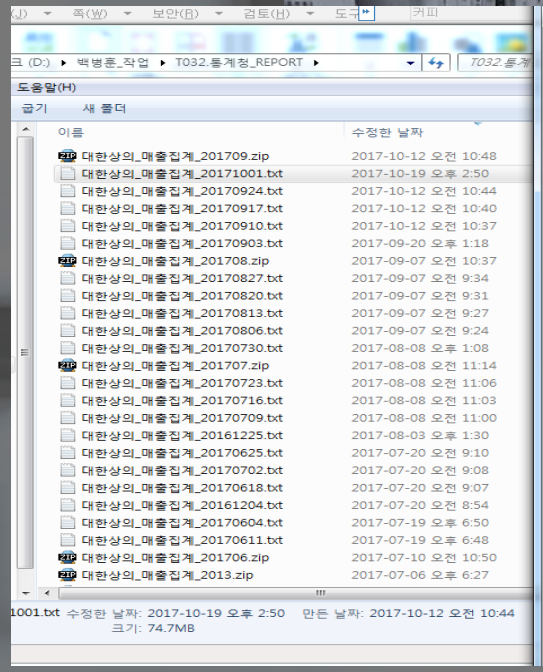
소비자물가지수 생산을 위한
스캐너 데이터 활용방안 연구

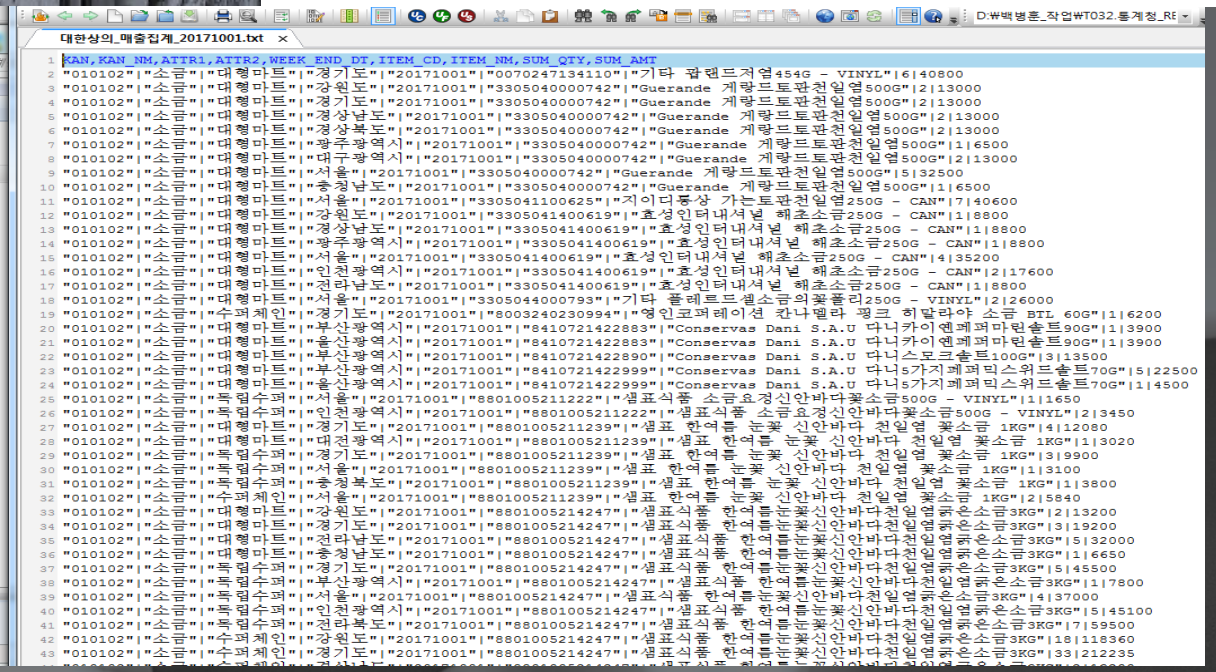
—

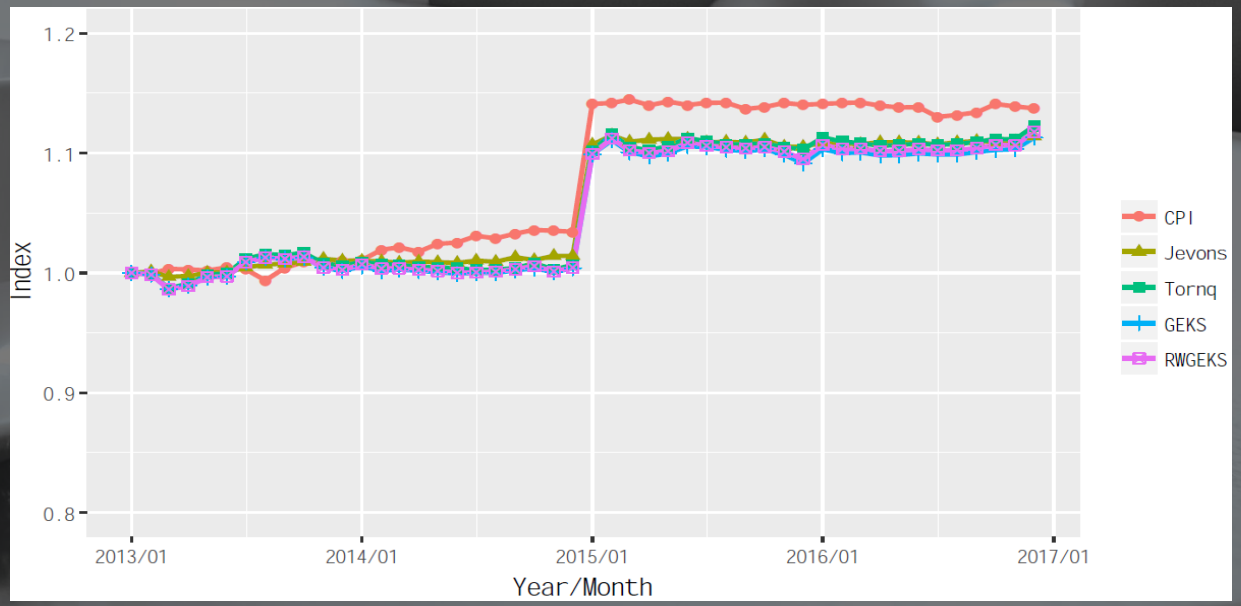
2017. 11

통계청

스캐너데이터를 이용한 소비자물가지수 작성







빅데이터의
이해와 활용

3 데이터 경제

3. 데이터 경제

- 데이터 생태계
  - 데이터 공급자 : 데이터 수집·저장 데이터 생산자와 데이터 유통 서비스 제공자 등
  - 데이터 수요자 : 생산된 데이터를 활용하여 업무를 하는 기업·정부, 데이터 서비스 이용 개인

3. 데이터 경제

- 생산요소로서 데이터의 속성
  - 무한히 소모되지 않고 복제될 수 있는 비경합성
    → 한계비용 제로의 속성
  - 데이터가 최신일수록, 모일수록, 사용할수록, 정확할수록, 다른 데이터와 결합할수록 그 가치는 높아짐

3. 데이터 경제

- 국가간, 기업간 격차 : 정부 과도한 지배력 완화 위한 제도 마련
  - 글로벌 플랫폼 기업은 미국 소재 : 국가간 격차
  - 빅테크 기업 다양한 데이터 기업 인수, 합병 : 기업 간 격차
  - 개별 기업들도 데이터 수집 위해 노력 : 독자 몰, 구독 서비스
    - 나이키 : 아마존에서 팔지 않고 독자 몰 구축, 데이터 수집
    - 미쉐린(타이어), 롤스로이스, GE(항공기 엔진), 테슬라(보험)

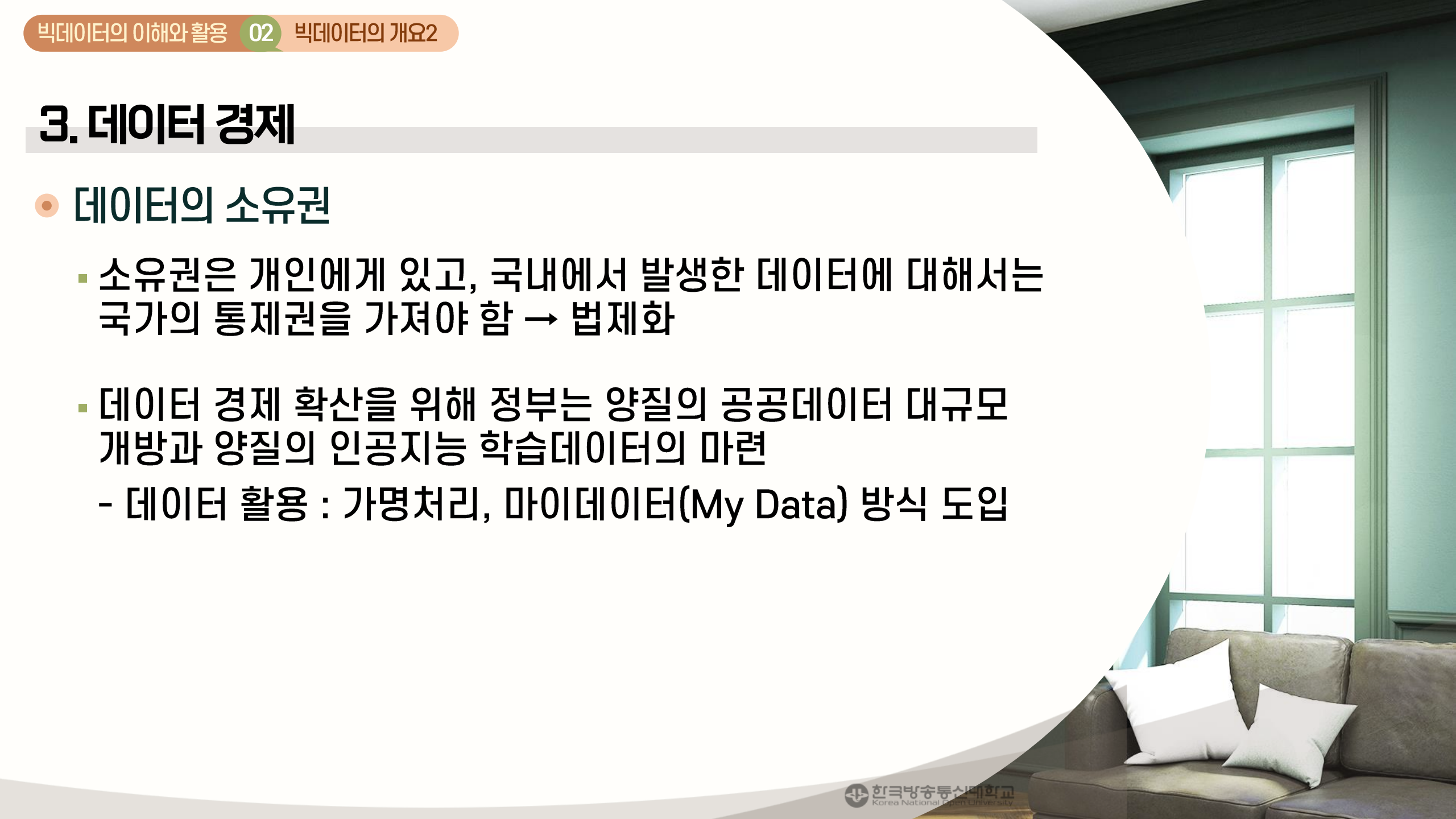
3. 데이터 경제

데이터의 소유권

- 소유권은 개인에게 있고, 국내에서 발생한 데이터에 대해서는 국가의 통제권을 가져야 함 → 법제화

- 데이터 경제 확산을 위해 정부는 양질의 공공데이터 대규모 개방과 양질의 인공지능 학습데이터의 마련

  - 데이터 활용 : 가명처리, 마이데이터(My Data) 방식 도입



3. 데이터 경제

- 데이터의 소유권
  - 가명처리 : 개인데이터에서 개인정보 일부를 삭제 또는 대체하여 추정정보 없이는 개인을 식별하지 못하게 하는 처리
  - 마이데이터는 정보주체인 개인이 본인의 신용, 자산, 건강 등의 정보를 스스로 관리할 수 있도록 하는 것

교재 읽기 : 18p

21세기 들어서 나타난 빅데이터와 머신러닝은 그동안 옳다고 생각해왔던 "좋은 데이터에 공정한 규칙 기반 통계모형을 적용하여 얻은 결과가 최상"이라는 우리의 판단 기준을 무너트리기 시작했다. 예를 들어 2016년, 2020년 미국 대선과 영국의 브렉시트(Brexit) 관련 여론조사가 실제 예상과 매우 달랐다. 또한, 모토로라, GE 등 절차 중심의 품질혁신을 추구했던 기업들은 사라지거나 기존 방식의 사업모형을 바꾸고 있다. 이제는"무엇인지는 설명할 수 없지만, 데이터로부터 좋은 결과(예측력이 좋은)를 내는 모형이 새로운 데이터에서 좋은 성과를 내는 모형"이라는 결과 중심으로 과정을 생각하게 되었다.

정리하기

- 데이터과학자는 해당 분야의 전문지식을 바탕으로 데이터를 수집·저장·가공하고, 다양한 원천의 데이터를 결합·분석하고, 이로부터 새로운 가치를 만드는 일을 하는 사람이다.

- 빅데이터 기반 데이터분석은 머신러닝(또는 딥러닝)의 방식으로 진행되는데 이 분석은 인과구조를 설명할 수 없지만 예측결과가 통계모형보다 우수하다.

정리하기

- 빅데이터는 인공지능 모형을 개발할 때 이용되어 새로운 서비스를 만드는 원천으로 이용되고 있고, 이 서비스는 새로운 빅데이터를 만들고 있다.

- 데이터 경제는 데이터가 경제의 기본 생산요소인 자본, 노동에 더해지는 새로운 생산요소가 되어 경제성장과 고용 창출에 주요한 역할을 하는 경제이다.

03 강

다음시간 안내

빅데이터의 수집과 활용 1

수고하셨습니다!