

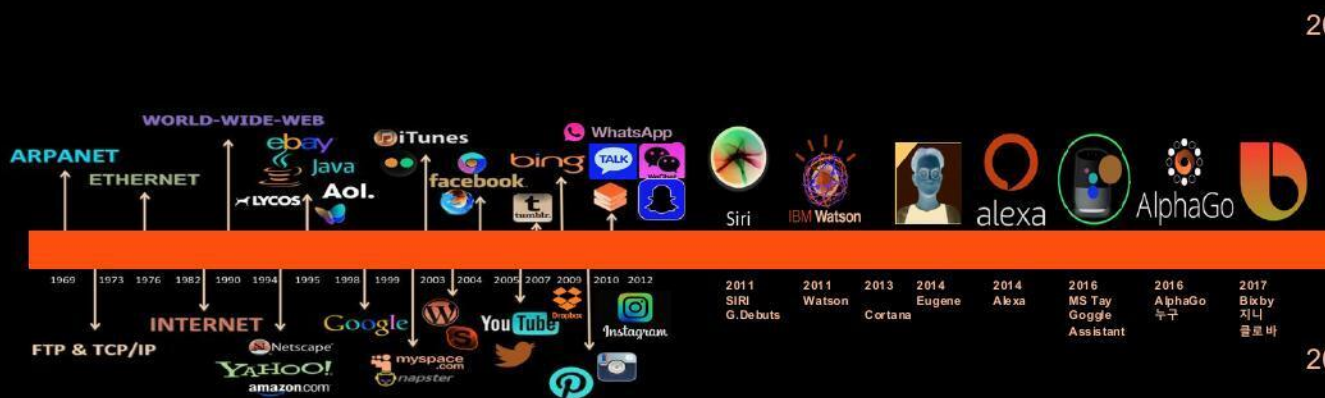
BigData

이해와 활용, 처리과정

빅데이터 분석 개론
이 강윤 (keylee@gachon.ac.kr)



Internet Timeline



Daum(1995)
Naver(1998)
Yahoo Korea(1997)
CyWorld(1998)



ICBM
ICBMS
AICBM
CAMSS
SMAC
ABC
ABCD
AI/BI/CI
AIoT
DNA

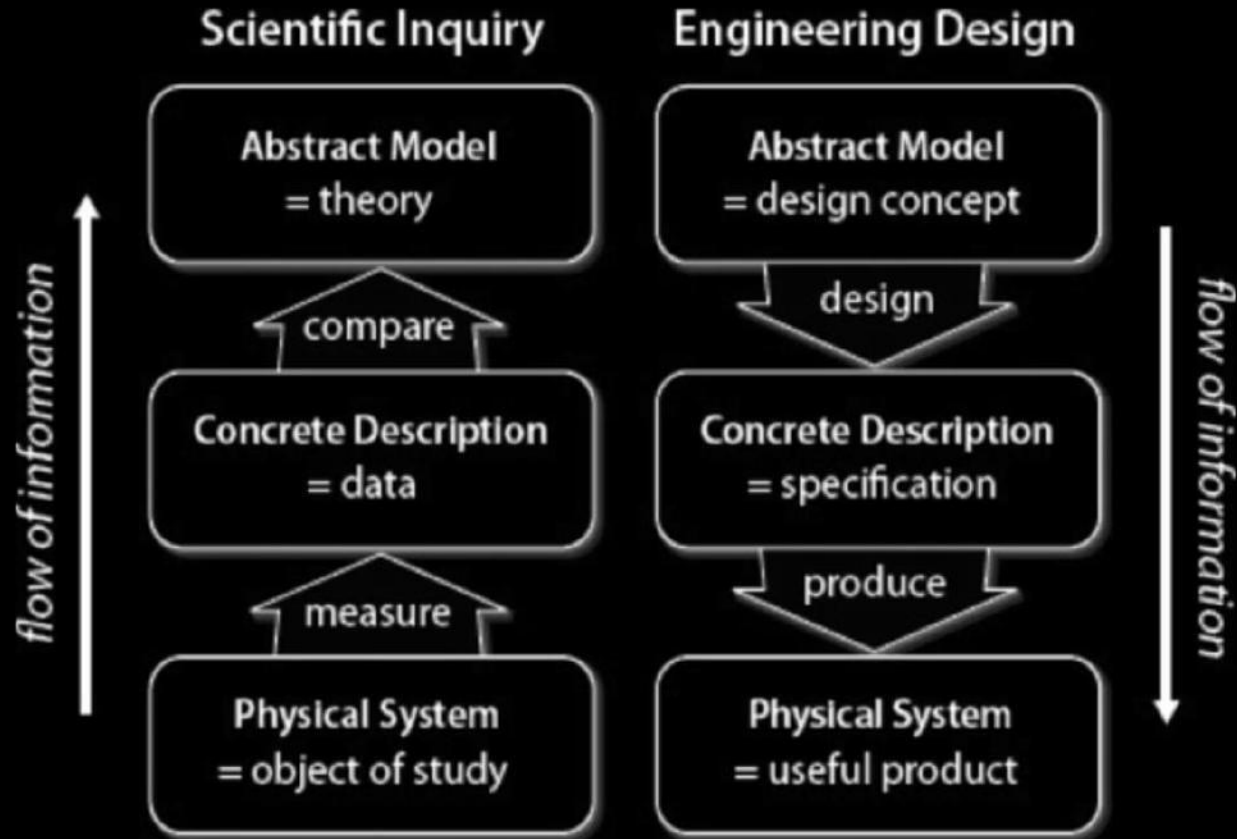
2020.6 OpenAI
ChatGPT



2024

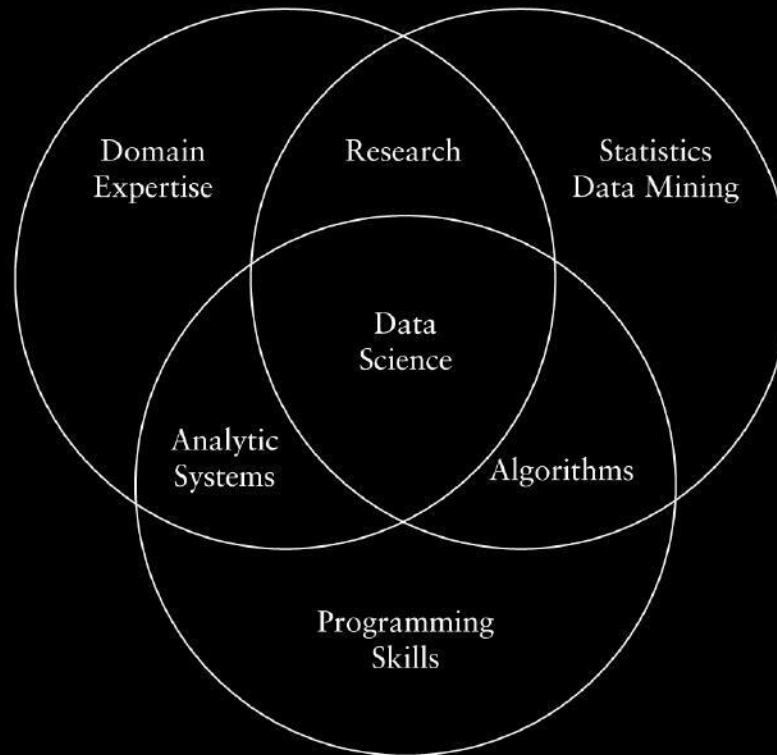


Science vs Engineering



The Antiparallel Structures of Scientific Inquiry and Engineering Design

Data Science – 3 interdisciplinary areas



01. 빅데이터의 이해

– 빅데이터의 정의

- 디지털 환경에서 발생하는 대량의 모든 데이터
- 대규모의 데이터를 저장·관리·분석할 수 있는 하드웨어 및 소프트웨어 기술, 데이터를 유통·활용하는 모든 프로세스를 포함
- 빅데이터 플랫폼을 구성하는 하드웨어, 소프트웨어, 애플리케이션 간의 유기적 순환에 의해 가치를 창출

표 2-1 빅데이터의 정의

기관	정의
맥킨지	일반적인 데이터베이스 소프트웨어가 수집, 저장, 관리, 분석할 수 있는 범위를 초과하는 대규모의 데이터다.
가트너	향상된 시사점과 더 나은 의사결정을 위해 사용되는 것으로 비용 효율이 높고 혁신적이며 대용량 고속 및 다양성을 가지는 정보 자산이다.
위키피디아	기존 데이터베이스 관리 도구의 수집, 저장, 관리, 분석 역량을 넘어서는 대량의 정형 또는 비정형 데이터셋 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술이다.
국가전략위원회	대용량 데이터를 활용 및 분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술이다.
삼성경제연구소	기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터 집합으로 대규모 데이터와 관계된 기술 및 도구(수집, 저장, 검색, 공유, 분석, 시각화 등)를 모두 포함한다.
한국정보화진흥원	저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터와 이것을 저장, 관리, 분석할 수 있는 하드웨어 및 소프트웨어 기술, 데이터를 유통 및 활용하는 과정을 통틀어 나타낸다. 즉, 빅데이터를 구성하는 하드웨어, 소프트웨어 그리고 이를 포괄하는 모든 프로세스를 의미하는 거대 플랫폼이다.



그림 2-1 빅데이터 플랫폼

01. 빅데이터의 이해

■ 빅데이터의 개념

– 빅데이터의 출현

- 기술의 발달과 비용 저하, 소셜 네트워크 서비스 발달, 그림자 정보와 사물 정보 증가 등의 ICT 패러다임의 변화
- 빅데이터에 전문 역량과 기술을 더하여 전략적으로 활용할 방법이 주목됨
- 경제적 가치 창출, 사회 문제 해결, 새로운 ICT 패러다임 견인이라는 신가치 창출

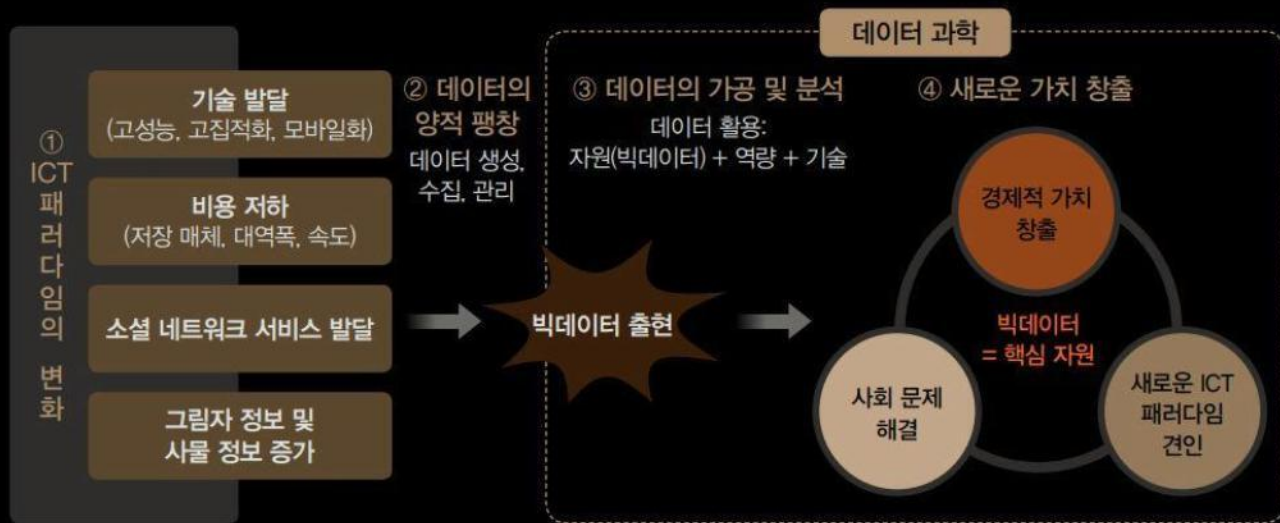


그림 2-2 빅데이터의 출현과 신가치 창출의 흐름

01. 빅데이터의 이해

■ 빅데이터의 분류

– 정형 데이터

일정한 규칙으로 체계적으로 정리된 것으로 그 자체로 해석이 가능하여 바로 활용할 수 있음

– 반정형 데이터

고정된 필드에 저장되어 있지는 않지만 XML, HTML 등의 메타데이터와 스키마를 포함하는 것으로 파일 형태로 저장

– 비정형 데이터

고정된 필드나 스키마가 없는 것

스마트 기기에서 페이스북, 트위터, 유튜브 등으로 생성되는 소셜 데이터

IoT 환경에서 생성되는 위치 정보나 센서 데이터와 같은 사물 데이터 등

표 2-2 정형화 정도에 따른 빅데이터의 분류

구분	설명	수집 및 처리 난이도
정형 데이터	• 고정된 필드에 저장 • 관계형 데이터베이스처럼 스키마 형식에 맞게 저장 • 예: RDB, 스프레드시트	• 내부 시스템에 의한 데이터라 수집하기 쉬움 • 파일 형태의 스프레드시트는 형식을 가지고 있어 처리하기 쉬움 • 처리 난이도: 하
반정형 데이터	• 고정된 필드에 저장되어 있지는 않지만 메타 데이터나 스키마 등을 포함 • 예: XML, HTML, JSON, 웹 문서, 웹 로그	• API 형태로 제공되므로 데이터 처리 기술이 필요함 • 처리 난이도: 중
비정형 데이터	• 데이터 구조가 일정하지 않음 • 규격화된 데이터 필드에 저장되지 않음 • 예: 소셜 데이터, 텍스트 문서, 이미지/동영상/음성 데이터, 문서 파일(PDF)	• 파일을 데이터 형태로 파싱해야 하므로 처리하기 어려움 • 처리 난이도: 상

01. 빅데이터의 이해

■ 빅데이터의 특징

– 데이터 측면

초기에는 빅데이터의 특징을 3V로 일컬어지는 규모, 다양성, 속도로 나타냄

빅데이터를 통한 가치 창출이 중요해지면서 정확성과 가치를 추가한 5V로 나타냄

표 2-3 빅데이터의 특징 - 데이터 측면

구분	특징	설명
1차 특징	규모	• ICT 기술의 발전으로 디지털 정보량이 기하급수적으로 폭증하여 제타바이트 시대로 진입
	다양성	• 데이터의 종류 증가: 로그 기록, 소셜/위치/소비/현실 데이터 등 • 데이터의 유형 다양화: 텍스트 외에 멀티미디어 등의 비정형 데이터 증가
	속도	• 센서, 모니터링 등의 사물 정보와 스트리밍 등의 실시간 정보가 증가하면서 데이터의 생성 및 이동(유통) 속도 증가 • 대규모 데이터를 처리하고 가치 있는 정보를 활용하기 위한 데이터 처리 및 분석 속도 증가
추가 특징	정확성	• 방대한 데이터를 기반으로 분석을 수행하므로 정확성 향상
	가치	• 빅데이터 분석으로 도출된 최종 결과물이 문제 해결을 위한 통찰력을 제공하므로 새로운 가치 창출 가능

01 빅데이터의 개념과 처리 과정

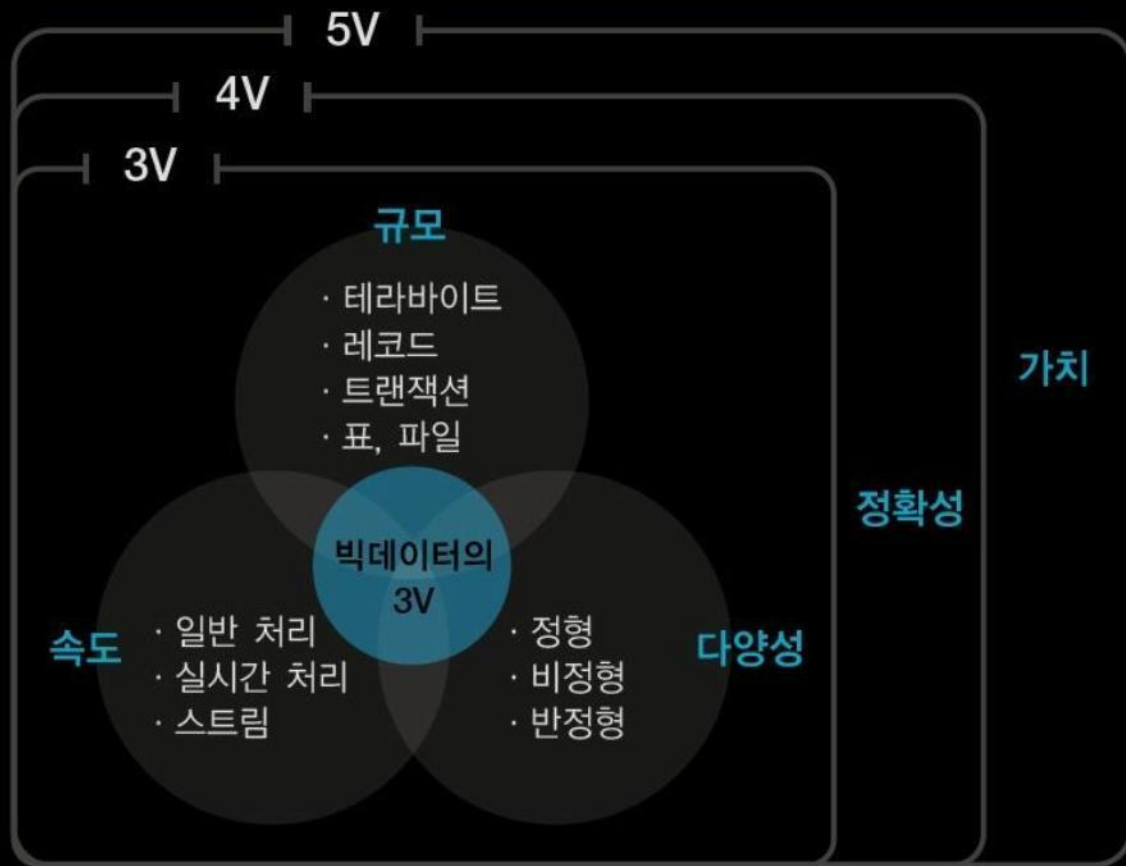


그림 1-2 빅데이터의 속성 [02]

01. 빅데이터의 이해

■ 빅데이터의 특징

- 분석 환경 측면

데이터 분석 시스템의 구성 요소인 데이터, 하드웨어, 소프트웨어 분석 방법은 분석 환경에 따라 다른 특징을 나타냄

표 2-4 빅데이터의 특징 - 분석 환경 측면

요소	과거의 데이터 분석 환경	현재의 빅데이터 분석 환경
데이터	<ul style="list-style-type: none">• 정형화된 수치 중심의 자료	<ul style="list-style-type: none">• 비정형의 다양한 데이터• 예: 문자 데이터(SMS, 검색어), 영상 데이터(CCTV, 동영상), 위치 데이터 등
하드웨어	<ul style="list-style-type: none">• 고가의 저장 장치• 데이터베이스• 대규모 데이터웨어하우스	<ul style="list-style-type: none">• 클라우드 컴퓨팅: 비용 대비 효율성 증대
소프트웨어 분석 방법	<ul style="list-style-type: none">• 관계형 데이터베이스: RDBMS• 통계 패키지: SAS, SPSS• 데이터 마이닝• 머신러닝• 지식 발견	<ul style="list-style-type: none">• 오픈 소스 형태의 무료 소프트웨어• 오픈 소스 통계 솔루션: R• 텍스트 마이닝• 오피니언 마이닝• 감성 분석

01. 빅데이터의 이해

■ 빅데이터의 특징

– 처리 방식 측면

빅데이터는 기존 데이터베이스 관리 시스템(DBMS)으로 처리하던 것에 비해 100배 이상 많은 정형, 비정형 데이터를 처리

표 2-5 빅데이터의 특징 - 처리 방식 측면

구분	이전의 데이터 처리 방식	빅데이터 처리 방식
데이터 트래픽	• 테라바이트 수준	• 페타바이트 수준: 최소 100테라바이트 이상 • 정보의 장기간 수집 및 분석 • 방대한 처리량
데이터 유형	• 정형 데이터 중심	• 비정형 데이터 비중이 높음: SNS 데이터, 로그 파일, 클릭스트림 데이터, 콜센터 로그 통신, CDR 로그 등 • 처리 복잡성 증대
프로세스 및 기술	• 단순한 프로세스 및 기술 • 정형화된 처리 및 분석 결과 • 원인 및 결과 규명 중심	• 다양한 데이터 소스와 복잡한 로직 처리 • 처리 복잡도가 높아 분산 처리 기술 필요 • 새롭고 다양한 처리 방법 필요: 정의된 데이터 모델/상관관계/절차 등이 없음 • 상관관계 규명 중심 • 하둡, NoSQL 등 개방형 소프트웨어 사용

NOTE_ 디지털 데이터 단위

- 1테라바이트(TeraByte: TB)=1024GB
- 1엑사바이트(ExaByte: EB)=1024PB
- 1요타바이트(YottaByte: YB)=1024ZB

- 1페타바이트(PetaByte: PB)=1024TB
- 1제타바이트(ZetaByte: ZB)=1024EB

01. 빅데이터의 이해

■ 빅데이터의 가치

– 혁신과 창조의 도구

- 빅데이터 분석이 제공하는 스마트 서비스는 기존 비즈니스에 효율화, 개인화, 그리고 미래 예측력을 통한 혁신을 제공
- 단순히 새로운 기술이나 비즈니스 모델이 아니라 새로운 패러다임으로의 변화를 의미
- 빅데이터 자체부터 이를 활용한 사용자 애플리케이션까지 광범위하여 빅데이터 플랫폼과 에코시스템으로 확장

표 2-6 빅데이터 분석을 통한 비즈니스 혁신 방향

방향	설명
효율화	<ul style="list-style-type: none">• 빅데이터를 이용해 과거 및 현재의 현상을 파악할 수 있다.• 물류, 재무, 기획, 마케팅 등 경영 전반의 데이터를 실시간으로 분석한 후 최선의 의사결정을 할 수 있다.
개인화	<ul style="list-style-type: none">• 온라인 이용자의 활동 정보와 SNS 등으로 축적된 개인 정보를 결합하여 사용자에게 특화된 서비스를 제공할 수 있다.• 현재 개인 정보는 광고 분야에 활용 중인데 이를 넘어 의료, 교육 등 모든 분야로 확대가 가능하다.
미래 예측력	<ul style="list-style-type: none">• 과거 및 실시간 데이터를 분석하여 축적한 개인 정보로 개인 또는 조직 전체의 행동 및 의사결정 패턴을 도출할 수 있다.• 미래에 적용 가능한 시나리오를 제시하고 예측 가능한 행동 및 발생 가능한 문제점을 사전에 방지하는 서비스가 가능하다.

01. 빅데이터의 이해

■ 빅데이터의 가치

– 사회·경제적 가치

- 빅데이터는 정치, 사회, 경제, 문화, 과학 기술 등 사회 전반에 걸쳐 가치 있는 정보를 제공
- 데이터의 도입과 활용은 산업 경쟁력 제고, 생산성 향상, 혁신을 위한 새로운 가치 창출을 할 것으로 기대

표 2-7 맥킨지가 제시한 빅데이터를 이용한 사회·경제적 가치 창출 방법

방법	설명
정보의 투명성	<ul style="list-style-type: none"> • 이해 관계자가 적시에, 보다 쉽게 빅데이터에 접근할 수 있게 하는 것만으로도 가치 창출이 가능하다. • 예: 공공 부문에서 분리된 부서가 관련 데이터에 보다 쉽게 접근할 수 있으면 데이터 검색과 처리 시간이 절감된다.
실험을 통한 소비자의 요구 발견, 트렌드 예측, 성과 관리	<ul style="list-style-type: none"> • 더 많은 거래 데이터를 디지털 형태로 축적함에 따라 더욱 정확하고 상세하게 소비자 요구를 발견하거나 트렌드 예측을 할 수 있다. • 예: 관리자가 빅데이터를 사용하여 자연스럽게 발생하거나 통제된 실험으로 일어나는 성과의 변동성을 분석하고 나아가서 근본적인 원인과 결과를 분석하면 더 높은 수준으로 성과를 관리할 수 있다.
소비자 맞춤 비즈니스를 위한 고객 세분화	<ul style="list-style-type: none"> • 빅데이터를 통해 더 구체적으로 고객을 세분화하여 고객의 요구에 맞는 더 정확한 맞춤형 서비스를 제공할 수 있다. • 예: 공공 부문에서 시민을 세분화하여 필요한 서비스를 제공할 수 있다.
자동화된 알고리즘을 통한 의사결정	<ul style="list-style-type: none"> • 빅데이터 기술을 사용하여 전체 데이터셋을 정교하게 분석함으로써 의사결정을 개선하고 위험을 최소화할 수 있으며 가치 있는 인사이트를 발굴할 수 있다. • 예: 판매 정보에 실시간 대응하여 재고 및 가격을 자동으로 조정하는 자동화 알고리즘은 소매업체의 의사결정을 최적화할 수 있다.
새로운 비즈니스 모델, 상품, 서비스의 혁신	<ul style="list-style-type: none"> • 빅데이터를 통해 새로운 상품 및 서비스를 개발하거나 기존 상품 및 서비스를 강화하여 완전히 새로운 비즈니스 모델을 개발할 수 있다. • 예: 실시간 위치 데이터를 이용하여 자동차를 운전하는 장소와 방법에 따라 내비게이션을 제공하고 상해보험 가격도 책정하는 완전히 새로운 위치 기반 서비스가 가능하다.

02. 빅데이터의 활용

■ 빅데이터의 역할

- 미래 사회의 특성은 불확실성, 리스크, 스마트, 융합으로 대변됨
- 빅데이터를 활용해 여러 가지 가능성에 대한 시나리오 시뮬레이션을 하면 불확실한 상황 변화에 유연하게 대처 가능
- 빅데이터에 기반한 정보 패턴 분석으로 리스크에 대응할 수 있음
- 개인화 및 지능화된 서비스를 제공하여 삶의 질을 향상시킴

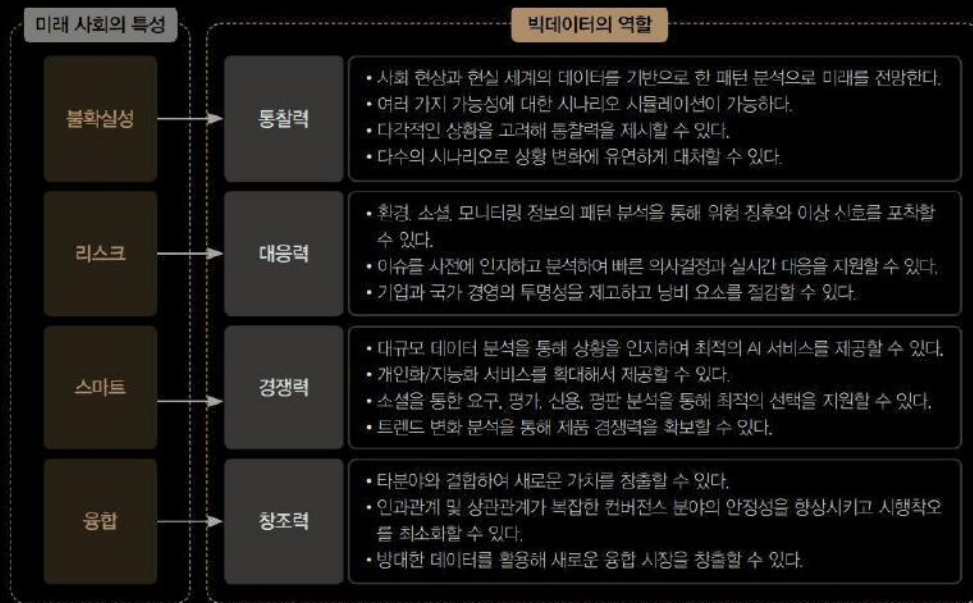


그림 2-3 미래 사회의 특성과 대응되는 빅데이터의 역할

02. 빅데이터의 활용

■ 빅데이터 활용 전략

– 기업의 성공적인 빅데이터 활용

- 리더십, 역량 관리, 기술 도입, 의사결정, 기업 문화가 필요
(미국 메사추세츠공대(MIT)의 에릭 브린올 프슨교수와 앤드류 맥아피 교수)

표 2-8 성공적인 빅데이터 활용 조건

조건	내용
리더십	목표 설정을 위해 빅데이터를 활용한 성공이 무엇인지를 명확히 정의하고 이를 강력하게 추진할 수 있는 리더십이 필요하다.
역량 관리	데이터 과학자, 시스템 개발자 등과 같은 전문 인력의 역량을 관리해야 한다.
기술 도입	빅데이터 관련 시스템에 최적화된 기술을 도입하고 조직 내·외부의 데이터를 통합 및 가시화하는 기술을 도입해야 한다.
의사결정	빅데이터 분석에 기반한 의사결정으로 조직의 유연성을 보장해야 한다.
기업 문화	빅데이터를 활용할 수 있는 조직 문화가 필요하다.

- 자원, 기술, 인력의 3가지 요소에 대한 전략을 수립

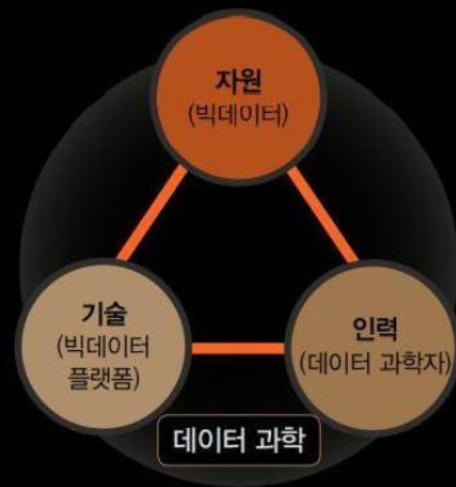


그림 2-4 성공적인 빅데이터 활용을 위한 3요소

02. 빅데이터의 활용

■ 빅데이터 활용 전략

– 활용 가능한 빅데이터 발견하기

- 가트너: 미래 사회에는 '데이터 경제 시대'가 도래할 것으로 전망
- 상호 연결과 협력으로 데이터 활용 영역이 확장되면 데이터 자원이 단계적으로 무한해질 것
- 그에 따라 자원을 확보하는 방안도 단계적으로 확장

표 2-9 데이터 자원 확보를 위한 단계적 방법

단계	내용과 과제	방법
저장	<ul style="list-style-type: none">• 조직의 독자적인 데이터를 생성 및 저장하는 단계• 인터넷을 통해 외부 데이터 수집(검색) 가능• 데이터의 신뢰성과 품질 제고를 위한 노력이 필요	생성, 저장, 수집(검색)
공유	<ul style="list-style-type: none">• 기업 데이터를 외부 기관과 상호 교환하는 단계• 1:1 또는 1:n의 공유 및 연계 가능	연계, 공유
통합	<ul style="list-style-type: none">• 특정 활동이나 목적을 위하여 연합, 그룹, 클럽이 상호 협력하는 공동의 장(집단)을 형성하는 단계• 표준 데이터 풀과 연계하여 국경을 초월한 정보 교환과 상호 이용이 가능	참여, 협력
공동 창출	<ul style="list-style-type: none">• 오픈 플랫폼으로 데이터를 공유하는 단계• 상호 협력과 참여로 공동의 자원을 창조	오픈, 창조

02. 빅데이터의 활용

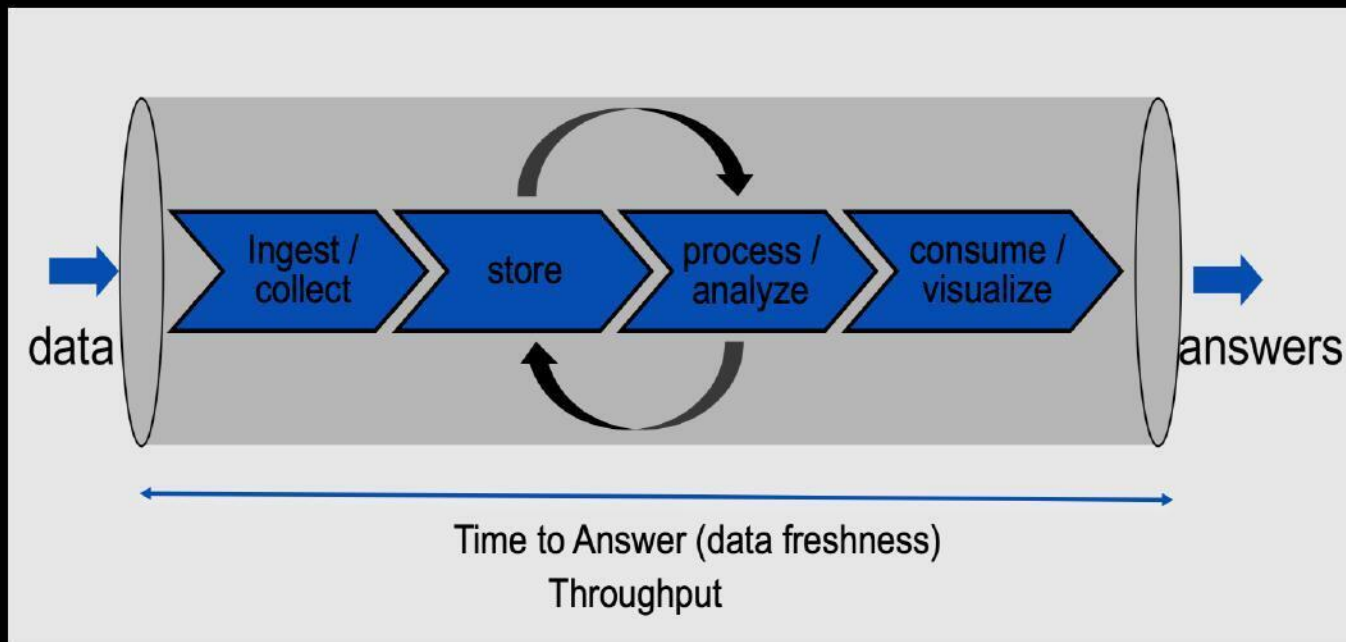
■ 빅데이터 활용 전략

– 빅데이터 처리 단계와 신기술 이해하기

- 빅데이터는 데이터의 수집 → 저장 → 처리 → 분석 → 표현의 단계를 거치며 세부 영역과 관련 기술이 개발
- 조직과 기업의 혁신 전략으로 적용할 수 있게 빅데이터 플랫폼, 빅데이터 분석 기법 및 기술에 대한 이해가 필요
- 분석 기술
 - 통계, 데이터 마이닝, 머신러닝, 딥러닝, 자연어 처리, 패턴 인식, 소셜 네트워크 분석, 비디오·오디오·이미지 프로세싱 등
- 빅데이터의 활용·분석·처리를 포함하는 인프라
 - BI, DW, 클라우드 컴퓨팅, 분산 데이터베이스 (NoSQL), 분산 병렬 처리, 분산 파일 시스템 등
- 빅데이터 관련 신기술
 - 대용량 데이터 처리를 위한 분산 처리 기술인 하둡과 인메모리, 의미 분석 기술인 데이터 마이닝, 자연어 처리, 머신러닝, 딥 러닝, 그리고 비정형 데이터 처리를 위한 NoSQL 기술

02. 빅데이터의 활용

▪ Simplified BigData Processing



수집 → 저장 → 처리 → 분석 → 표현

02. 빅데이터의 활용

빅데이터 처리 과정과 기술 (5단계)



그림 1-7 빅데이터 처리 과정 [09]

02. 빅데이터의 활용

■ 빅데이터 활용 전략

– 빅데이터 처리 단계와 신기술 이해하기

표 2-10 빅데이터의 처리 단계별 기술 영역

단계	기술 영역	내용
데이터 소스	내부 데이터	데이터베이스, 파일 관리 시스템
	외부 데이터	파일, 멀티미디어, 스트리밍
수집	크롤링 <small>crawling</small>	검색 엔진 로봇을 이용한 데이터 수집
	ETL: 추출 <small>Extraction</small> , 변환 <small>Transformation</small> , 적재 <small>Loading</small>	소스 데이터의 추출, 전송, 변환, 적재
저장	데이터 관리: NoSQL	비정형 데이터 관리
	저장소	빅데이터 저장
	서버	초경량 서버
처리	맵리듀스 <small>mapReduce</small>	데이터 추출
	작업 처리	다중 작업 처리
분석	신경 언어 프로그래밍 <small>NLP, Neuro Linguistic Programming</small>	자연어 처리
	머신러닝	데이터 패턴 발견
	직렬화 <small>serialization</small>	데이터 간 순서화
표현	시각화 <small>visualization</small>	데이터를 도표나 그래픽으로 표현
	획득 <small>acquisition</small>	데이터의 획득 및 재해석

02. 빅데이터의 활용

5개 레이어



02. 빅데이터의 활용

■ 빅데이터 활용 전략

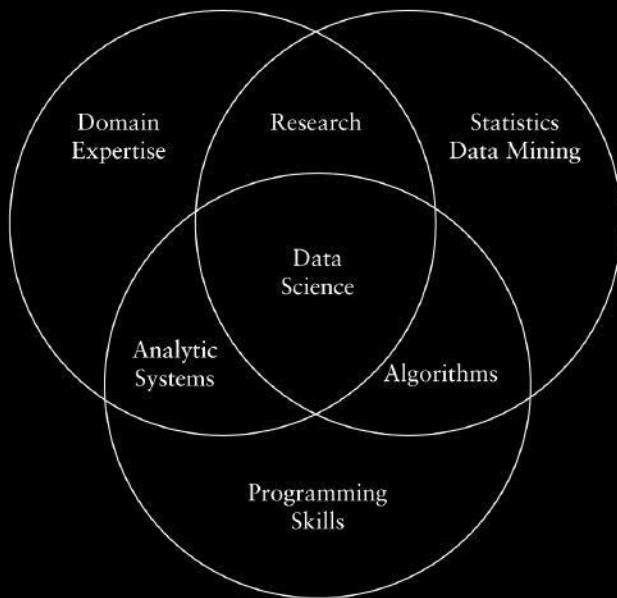
– 데이터 과학자 역량 강화하기

- 빅데이터 시대에는 데이터를 분석하고 관리할 수 있는 인력에 대한 중요성이 커짐
- 대규모 데이터 속에서 숨겨진 정보를 찾아내는 데이터 과학자는 ‘빅데이터 시대의 연금술사’

• 존 라우치: 데이터 과학자에게 필요한 6가지 기본 자질

- ① 수학 역량
- ② 공학 역량
- ③ 데이터를 분석할 때 필수적인 가설을 세우거나 검증할 때 필요한 비판적 시각
- ④ 이를 잘 작성할 수 있는 글쓰기 역량
- ⑤ 다른 사람에게 잘 전달할 수 있는 대화 능력
- ⑥ 호기심과 개인의 행복

- 데이터 과학자는 외부보다는 내부 인력으로 내재화하여 활용



Gachon
Cognitive Computing Lab

