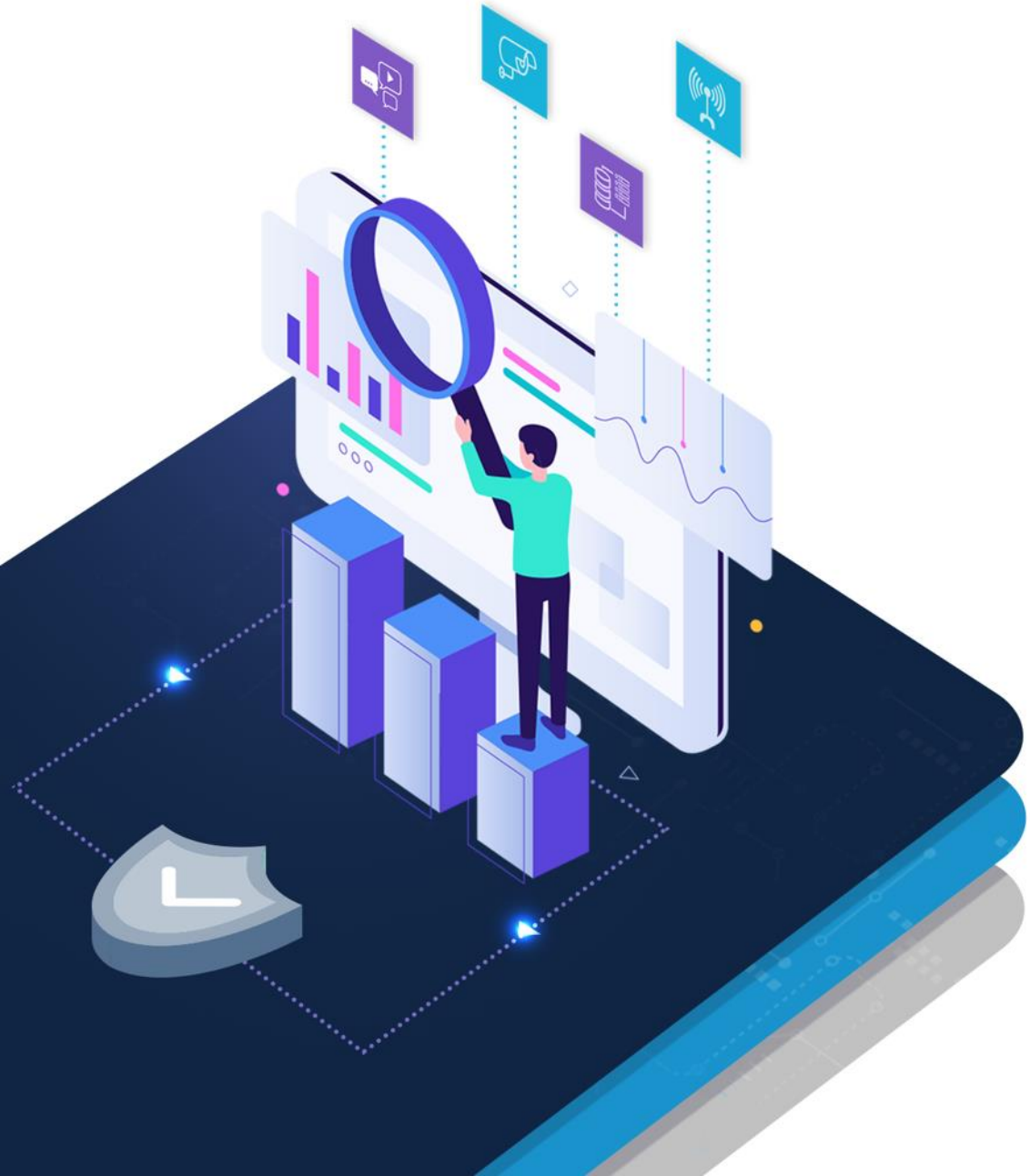
과학기술정보통신부

NIA

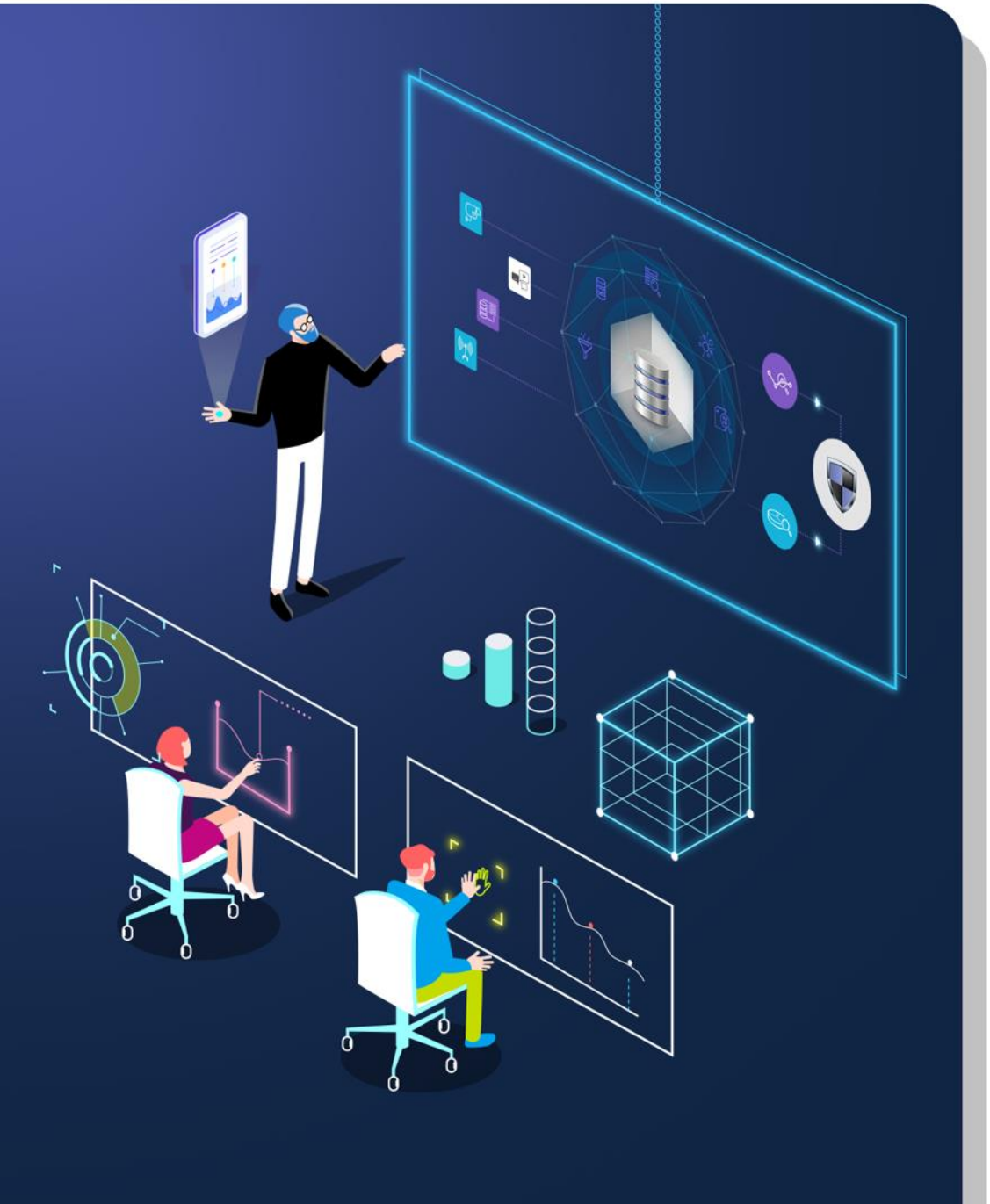
통합데이터지도

데이터 분석과정
〈전처리의 중요성〉

WISEiTECH 위세아이텍 연구소 / 오현우 연구원



목차



Chapter

I 디지털 시대와 데이터 활용

1. 디지털 시대의 D.N.A.와 기업들
2. AI 머신러닝 모델 도입 어려움 요인
3. AI 도입,운영의 저해요인 중요도 변화 전망
4. AI 인력 수급 전망

01 디지털 시대 및 데이터 활용 진입장벽

그림 1-1 ▌디지털 시대의 D.N.A.와 기업들

표 1-1 AI 머신러닝 모델 도입 어려움 요인[9)]

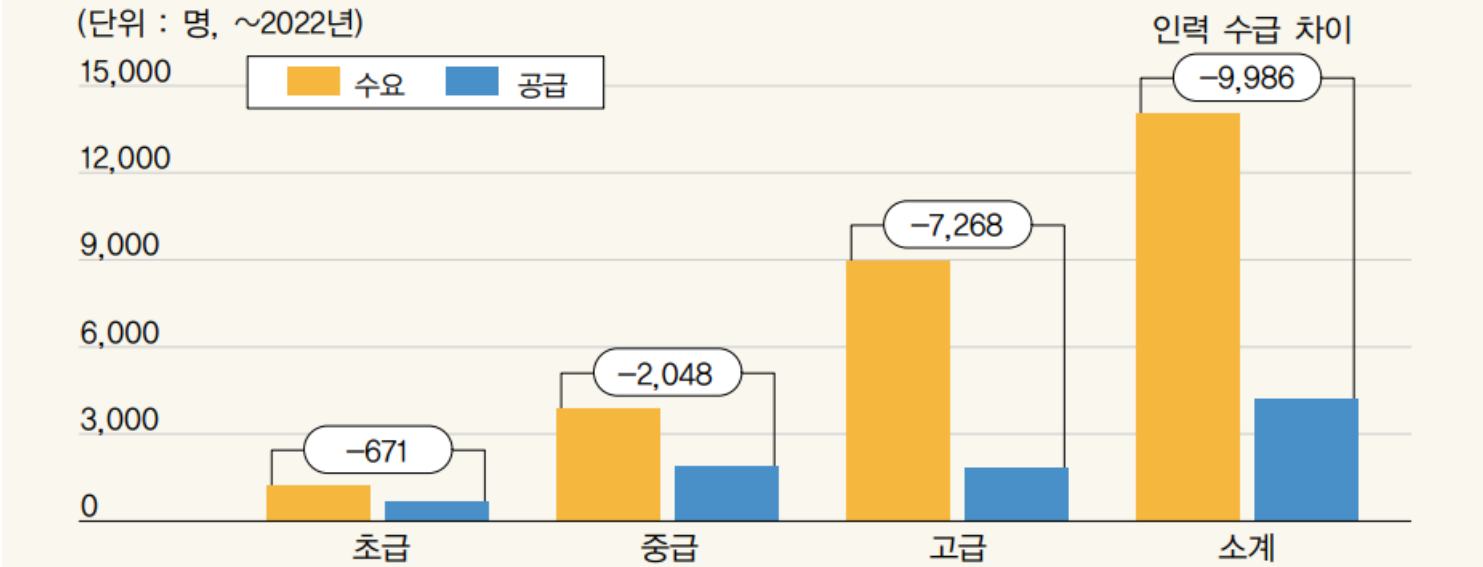
	도입 전	도입 후		
		사전 준비	모델 개발	서비스 운영
1	내부 인재 부족	양질의 데이터 부족		내부 운영 인력 부족
2	초기 투자 비용	과업 구체화의 어려움	모델 개발의 여려움	결과물 검증의 어려움
3	AI 기술 미성숙	모델 선택의 어려움	도메일+AI 인재 부족	모델 고도화의 어려움
4	데이터 부족	문제 정의의 어려움	데이터 전처리 어려움	높은 공급기업 의존성
5	AI 인프라 비용 부족	AI에 대한 이해 부족	모델 검증의 어려움	데이터 관리의 어려움

- **내부 역량** : 도메인, AI 기술자, 모델 운영 및 개선 내부 인력의 부족
- **AI 기술 미성숙** : 데이터 전처리, 모델 선택, 모델 개발의 어려움
- **데이터 의존성** : 양질의 데이터 부족, 데이터 관리의 어려움
- **불확실성** : 결과물 검증의 어려움, 모델 검증의 어려움

표 1-2 AI 도입·운영의 저해요인 중요도 변화 전망[12)]

저해요인	향후 중요도 변화 전망 및 이유	
데이터 의존성	↑	AI 모델의 차별화 요인으로써 데이터의 가치 증가
내부 역량 부족	↑	도메인 지식+데이터 리터러시를 갖춘 내부 인재 확보가 AI 도입의 성패를 좌우
불확실성	↑	기술 이해도가 낮은 다양한 기업의 AI 도입으로 편향, 품질관련 책임분쟁 확산 가능
공급기업 의존성	↑	AI 플랫폼을 통해 AI 기술을 도입하게 됨으로써 플랫폼에 락인될 가능성 증가
AI기술의 복잡성/미성숙	↓	쉬운 AI 개발툴의 확산으로 중요도 감소 예상
비용 부담	↓	다양한 AI 서비스의 등장으로 비용 부담 완화 예상

그림 1-4 ▌AI 인력 수급 전망[21]



Chapter

II 데이터

1. 데이터
2. 빅데이터
3. 데이터 V.S. 빅데이터

02 데이터

데이터(Data)

- **인간 또는 컴퓨터를 비롯한 자동 기기에 의해 행해지는 통신과 해석, 처리로 형식화된 사실과 개념, 명령을 표현한 것. 정보는 특정한 목적에 따라 특유의 형식을 갖고 있다. 정보 처리 분야에서 데이터는 다음 3가지의 뜻을 갖는다.**

A. 컴퓨터 프로그램과 구별했을 때의 데이터이다. 예를 들면, 변수, 정수의 값, 프로그램 파일과 데이터 파일과 같이, 처리하는 것과 처리되는 것의 관점에서 본 경우이다. 그러나 이 관점도 앞뒤의 문맥에 따라 좌우된다. 원시 프로그램은 연결기의 입력 데이터가 되기 때문이다.
B. 처리 프로그램의 관점에서 볼 때 결과, 즉 출력에 대한 입력과 같이 더욱 엄밀한 뜻으로 사용하는 경우이다.
C. 문서, 음성, 화상과 구별하는 경우이다. 이 결과, 데이터 처리와는 별도로 문서 처리, 음성 처리, 화상 처리의 분야가 존재하고 있다는 뜻이다.

- 출처: 데이터[data] (IT용어사전, 한국정보통신기술협회)

02 빅데이터

빅데이터(Big Data)

- 빅 데이터는 통상적으로 사용되는 데이터 수집, 관리 및 처리 소프트웨어의 수용 한계를 넘어서는 크기의 데이터를 말한다. 빅 데이터의 사이즈는 단일 데이터 집합의 크기가 수십 테라바이트에서 수 페타바이트에 이르며, 그 크기가 끊임없이 변화하는 것이 특징이다. 빅데이터라는 용어는 1990년대부터 사용되어 왔으며, 존 매쉬가 이 용어를 대중화하였다.
- 출처: 빅데이터,『 위키백과 』

- 빅데이터는 전통적인 데이터 프로세싱 방법으로 처리할 수 없을 정도로 대규모이거나 복잡한 데이터입니다. 빅데이터는 흔히 'Three V'로 불리는 볼륨(Volume), 다양성(Variety), 속도(Velocity)라는 특성을 가지고 있습니다. 볼륨은 대규모 크기를 의미하며, 다양성은 비표준 형식의 광범위한 범위를 그리고 속도는 신속하고 효율적으로 처리되어야 하는 특성을 의미합니다.
- 출처: 빅데이터,『 레드햇 』

- 빅데이터는 우리가 매일 사용하는 컴퓨터와 모바일 기기, 기계 센서에서 흘러나오는 방대한 제타바이트급 데이터로 구성된 정보의 바다를 가리킵니다.
- 출처: 빅데이터,『 SAP 』

- 빅 데이터란 양(Volume)이 매우 많고, 증가 속도(Velocity)가 빠르며, 종류(Variety)가 매우 다양한 데이터를 말합니다. 이것을 3V라고도 합니다.
- 출처: 빅데이터,『 오라클 』

- 빅데이터는 일반적으로 사용되는 데이터의 한계를 넘어서는, 복잡하고 용량이 크고 끊임없이 변화하는 데이터를 말합니다.
- 출처: 빅데이터,『 통합 데이터 지도 』

빅데이터의 특징

Volume	규모의 증가
Variety	다양성
Velocity	처리속도

3V

Value	가치
Veracity	진실성

+2V

Validity	유효성
Volatility	휘발성

+2V

5V

7V

02 데이터 V.S. 빅데이터

데이터와 빅데이터 비교

구분	데이터	빅데이터
수집	데이터 관리기관	다양한 경로
분석가능 표본	모집단 내의 표본집단	모집단 전체
데이터 표본	기본 정보 파악 가능	기본 정보 파악 불가능
데이터 형식	구조화/정형화 고정적인 데이터 객관화	비구조화/비정형화 실시간 생성 및 확장 가능 데이터 역동성, 상호의존성, 관련성 텍스트, 숫자, 동영상 등 다양한 형태
데이터 분류	관계형 데이터베이스로 분류 가능	관계형 데이터베이스로 분류 불가능
데이터 분석	계량적 모형기반 전통적인 분석기법 통계적 분석	기계학습 등 다양한 인공지능 분석 통계적 분석 데이터 시각화 자연어 처리 텍스트 분석
장점	통제된 정확성	실시간 공개로 인한 투명성 확보
한계점	제한된 표본 자료수집 비용	불순물 신뢰성 결여 개인정보 및 사생활 침해 국가 및 조직의 보안 문제 정보의 불균형 심화 기능 일반화 가능성에 대한 지속적 논의

Chapter

III

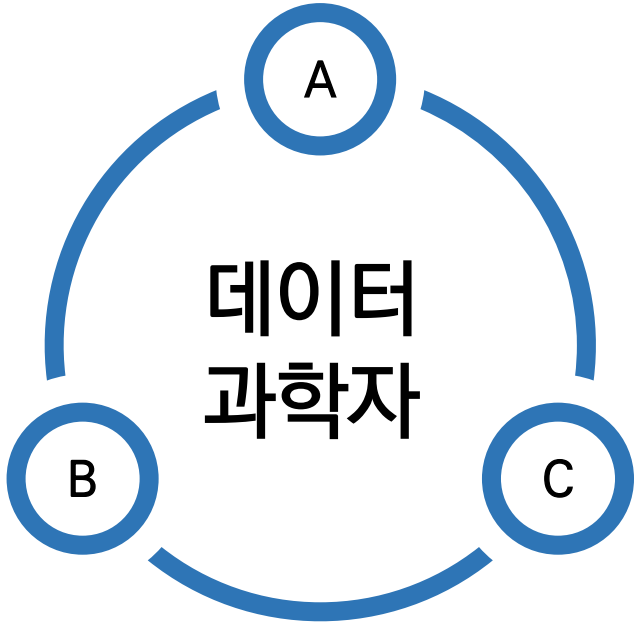
데이터 과학자

1. 데이터 과학자
2. 데이터 과학자의 필요조건
3. 데이터 과학자의 평가기준
4. 데이터 과학자에 대한 고찰

03 데이터 과학자

데이터 과학자(Data Scientist)

- 데이터 엔지니어
- 데이터 분석가
- 머신러닝/딥러닝 엔지니어



- **데이터 과학자의 3가지 타입**

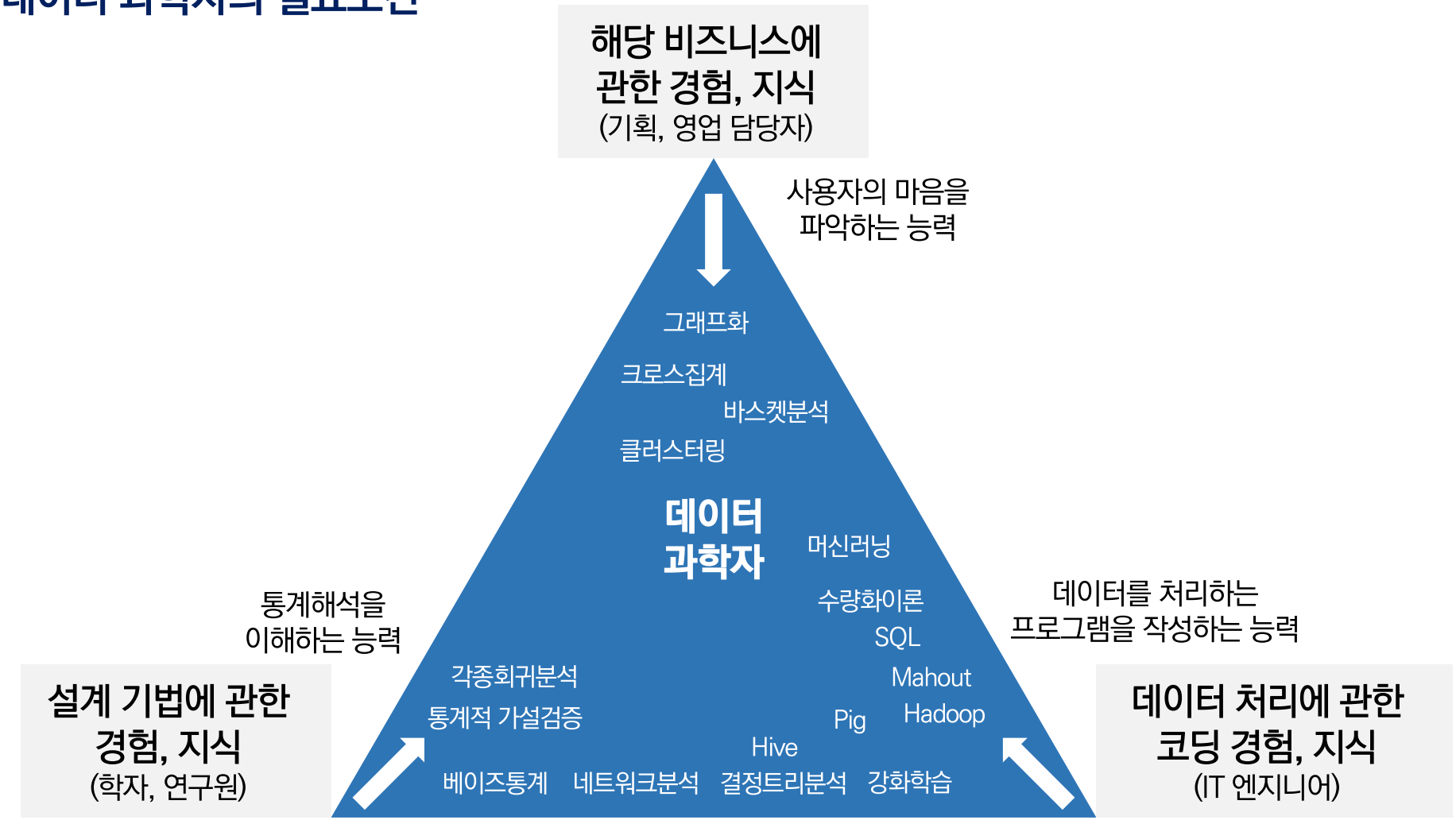
A. 비즈니스 영역 기반의 데이터 과학자

B. 통계학 영역 기반의 데이터 과학자

C. 엔지니어링 영역 기반의 데이터 과학자

03 데이터 과학자

데이터 과학자의 필요조건



03 데이터 과학자

데이터 과학자의 평가기준

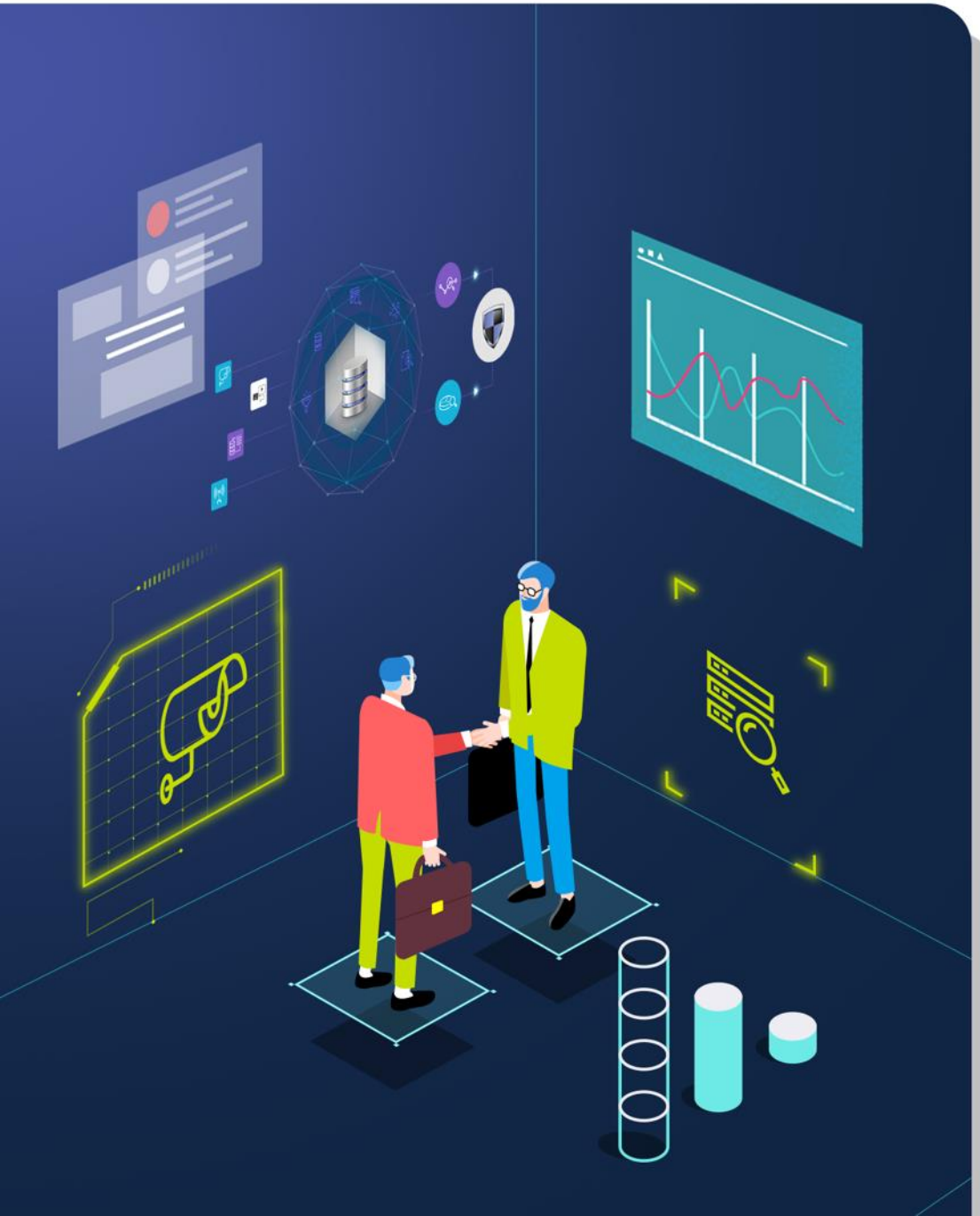
03 데이터 과학자

데이터 과학자에 대한 고찰

- 균형 잡힌 시각
  “데이터 과학자는 만능이 아니지만 만능이다.”

- 기본에 충실하고,
  각 분야의 전문가가 되자!

- ‘새로운 타입의 전문가’를
  어떻게 지속적으로 양성할지 고민



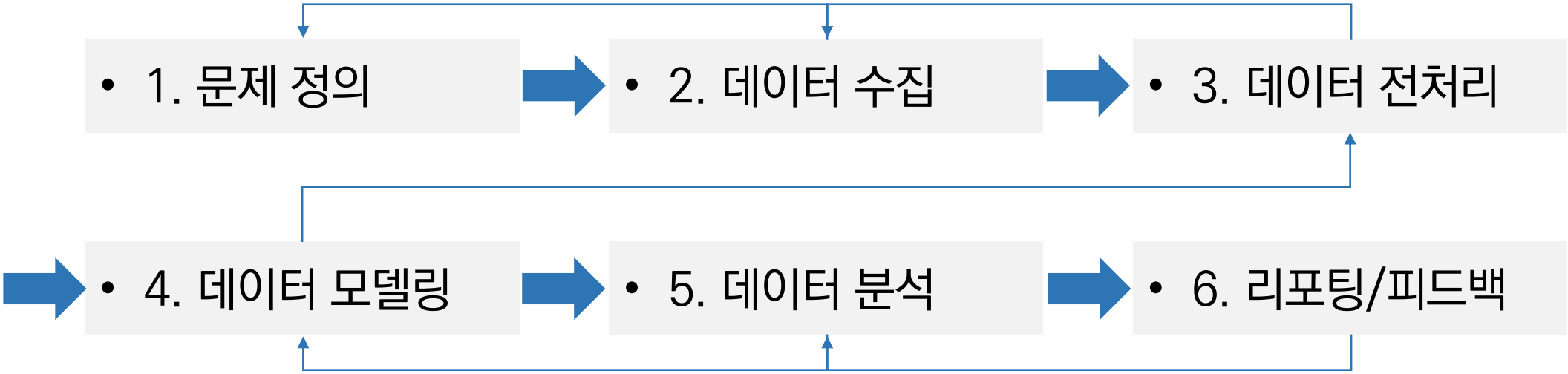
Chapter

Ⅳ 데이터 분석

1. 데이터 분석 프로세스
    1. 문제 정의
    2. 데이터 수집
    3. 데이터 전처리
    4. 데이터 모델링
    5. 데이터 분석
    6. 리포팅/피드백

04 데이터 분석

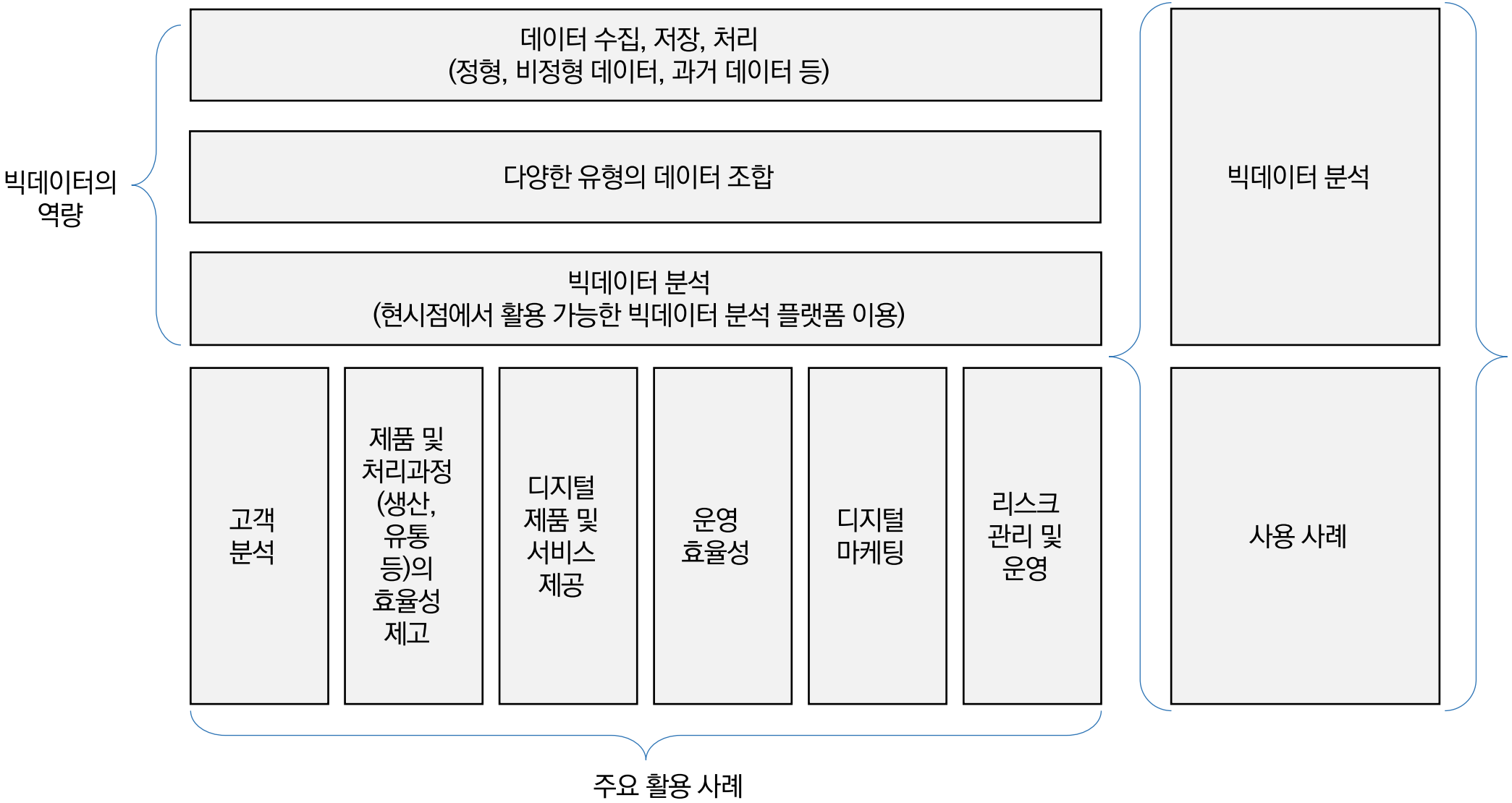
데이터 분석 프로세스



각 단계가 유기적으로 상호작용

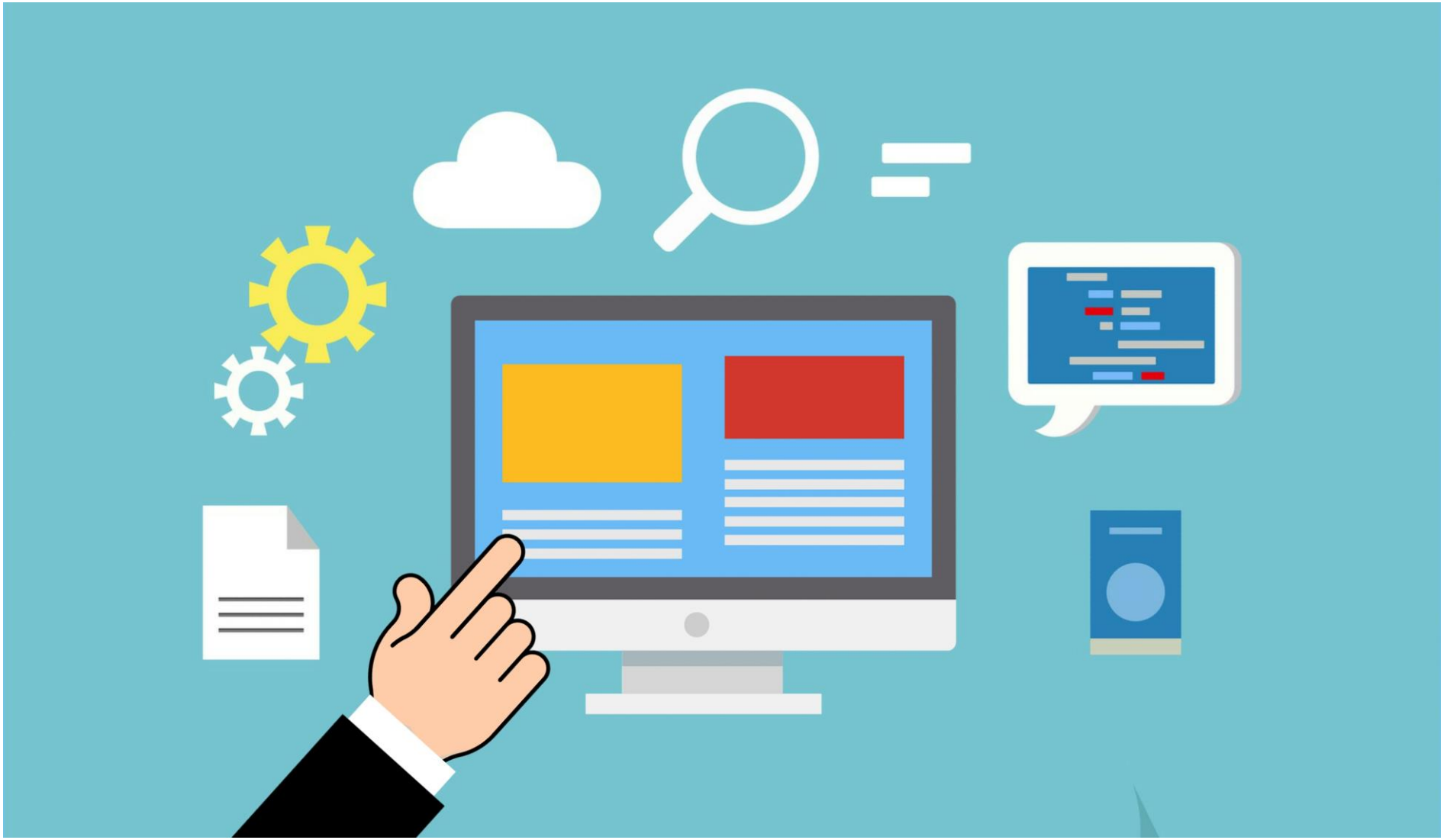
04 데이터 분석 – (1) 문제 정의

빅데이터 분석의 목적(Gartner)



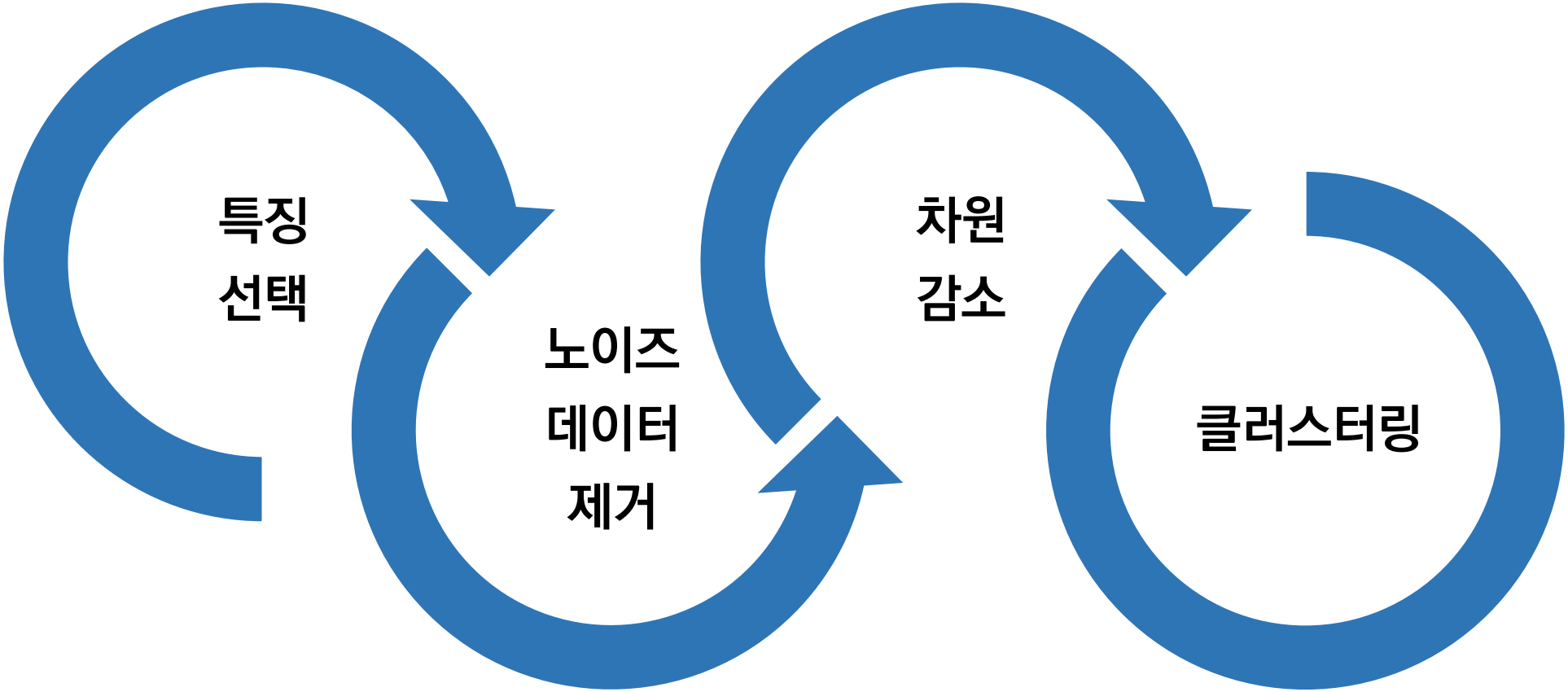
04 데이터 분석 – (2) 데이터 수집

데이터 수집



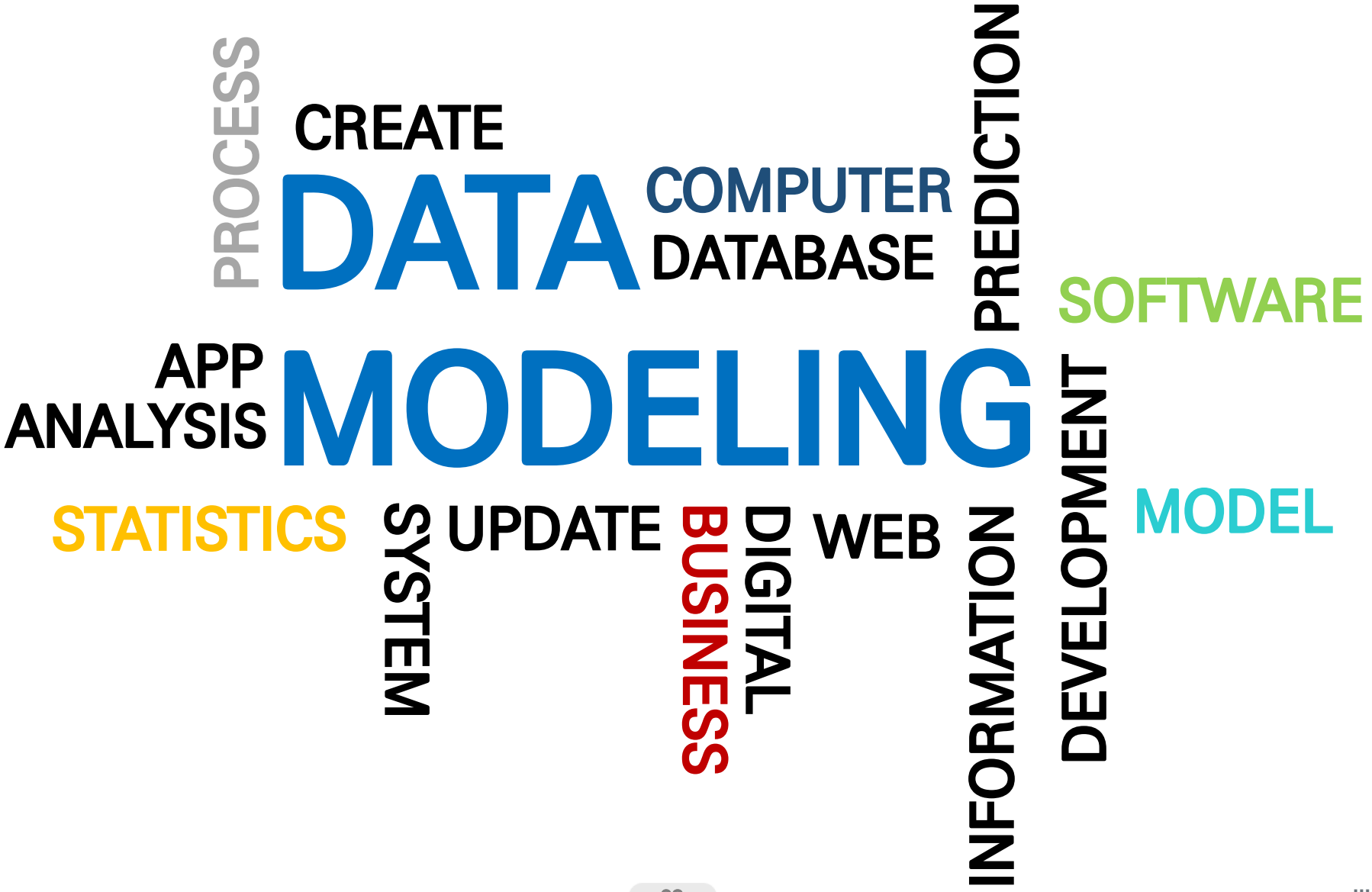
04 데이터 분석 – (3) 데이터 전처리

데이터 전처리



04 데이터 분석 – (4) 데이터 모델링

데이터 모델링

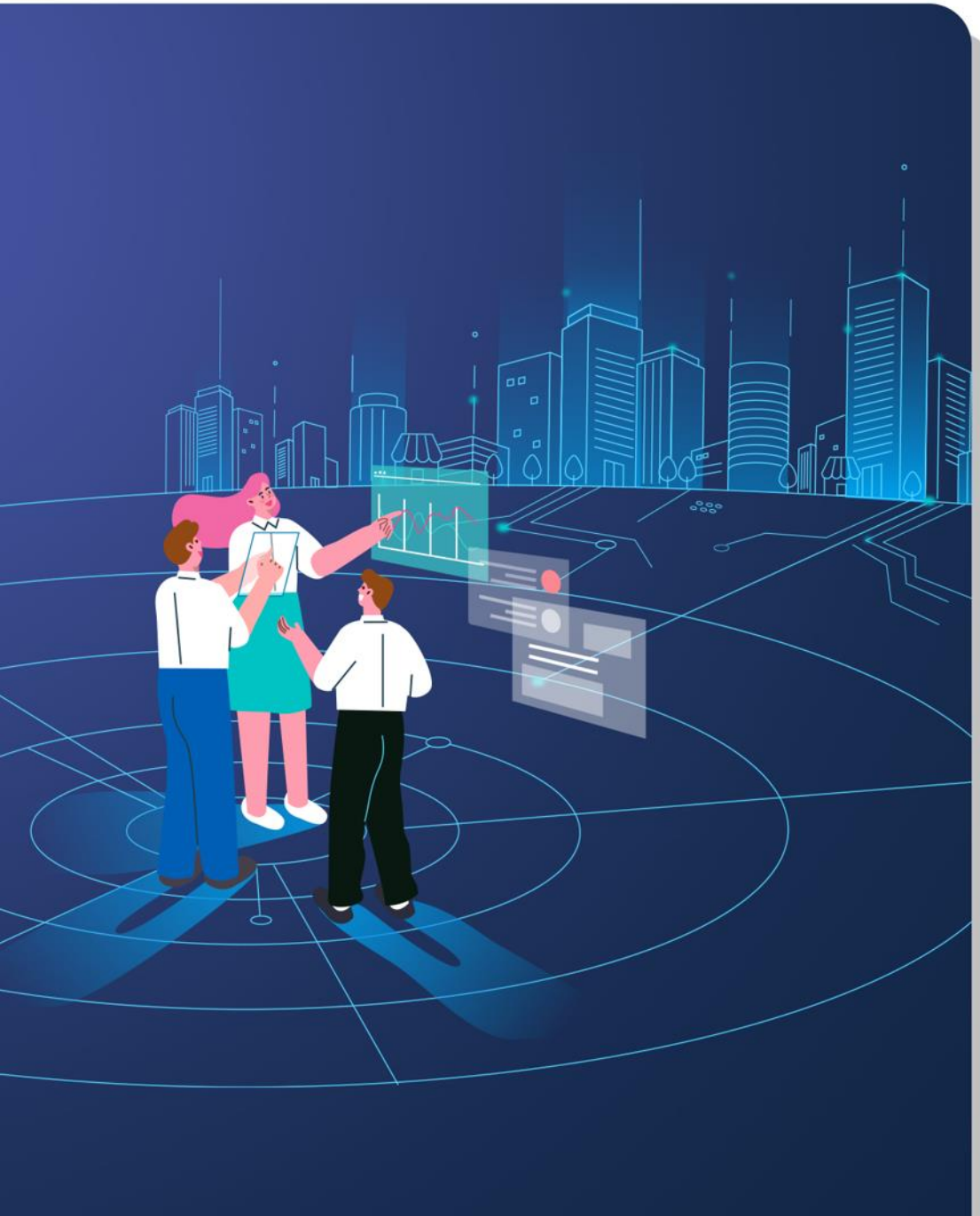


04 데이터 분석 – (5) 데이터 분석

데이터 분석

04 데이터 분석 – (6) 리포팅/피드백

리포팅/피드백

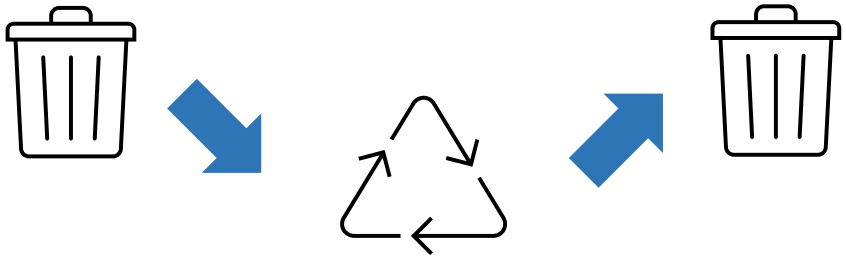


Chapter

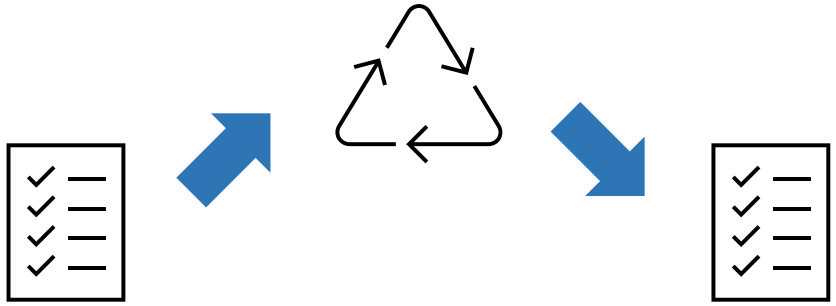
V 데이터 전처리

1. 사례 탐구
    1. 사례#1
    2. 사례#2
    3. 사례#3

사례 탐구



“Garbage In, Garbage Out”



05 데이터 전처리

사례#1

한경 국제

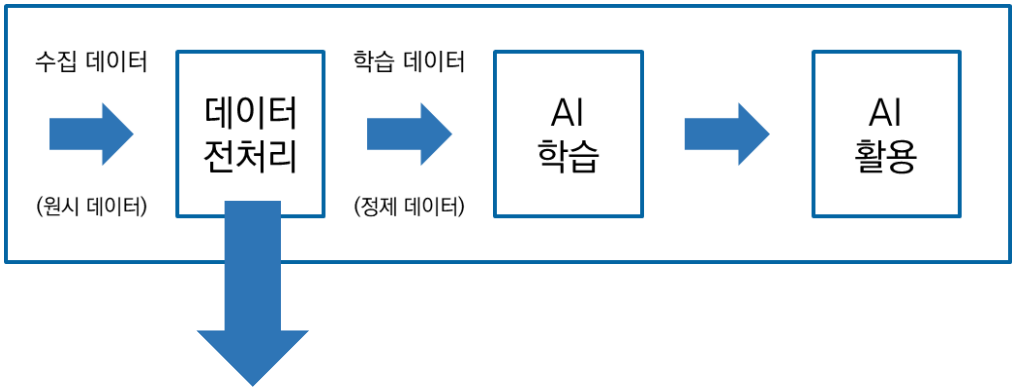
데이터 전처리도 AI의 중요한 요소다

오준호 선임기자

입력 2021.03.18 05:50 수정 2021.03.18 10:36 가가

최두환 포스코ICT 고문

사람들은 AI를 생각할 때 데이터가 많고 알고리즘이 좋으면 AI가 좋은 성능을 발휘할 것이라 여긴다. 그런데 이런 생각에는 허점이 있다. AI에 입력되는 데이터가 올바르지 않으면, AI의 출력 또한 엉망이 된다. 소위 영어로 "Garbage in, garbage out(쓰레기를 넣으면 쓰레기가 나온다)" 현상이 발생하는 것이다. AI가 성능을 발휘하려면 좋은 알고리즘만큼이나 중요한 것이 입력되는 데이터 수준이다.



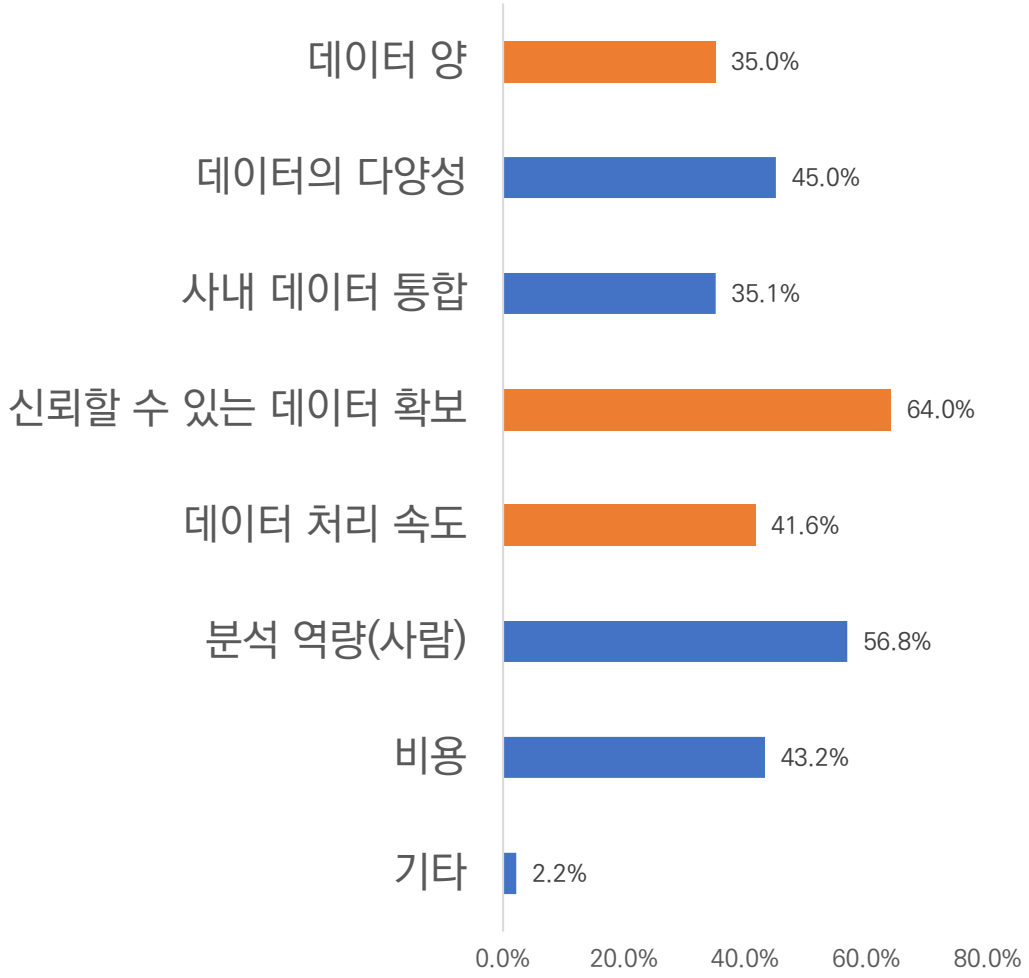
기계적 데이터 전처리
(Mechanical Data Preprocessing)

- 데이터 청소(Data Cleaning)
- 데이터 편집(Data Editing)
- 데이터 라벨링(Data Labeling)

의미적 데이터 전처리
(Semantic Data Preprocessing)

- 데이터 정리(Data Reduction)
- 데이터 재구성(Data Wrangling)

사례#2



빅데이터 프로젝트 추진 시 고민
(출처: 마이크로스트레티지)

뉴스 인물·기업 기획특집 테크가이드 제품가이드 컨퍼런스 테크라이브러리 웨비나

HOME > 뉴스 > 소프트웨어

[기고] 스마트하지 않은 데이터 활용법

데이터넷 | 승인 2022.06.28 09:48 | 댓글 0

데이터 한계·특성 명확히 알고 활용할 수 있는 상태로 지속 관리 필요

[데이터넷] 데이터는 일종의 의사결정 도구로, 데이터를 스마트하게 잘 활용하는 것이 중요하다. 그러나 실제로 많은 기업들이 활용 가능한 데이터 확보에 어려움을 겪고 있으며, 또 원하는 모든 데이터를 수집하는 것도 불가능하다. 이에 확보된 데이터를 이용해 최대한의 결과를 낼 수 있어야 하며, 스마트하지 않은 데이터도 쓸모 있는 데이터로 변환해 활용이 가능하다.

연재순서

스마트하지 않은 데이터 허브를 구별하는 방법
스마트하지 않은 데이터를 활용하는 방법(이번호)
스마트하지 않는 생활을 바꾸는 융복합 서비스

어쩌면 여러분은 스마트하지 않는 데이터까지 활용해야 할 필요성을 전혀 느끼지 못할 수도 있다. 스마트시티 사업은 최신 기술을 이용해 눈길을 끄는 것이 중요하다고 생각되기 때문이다.

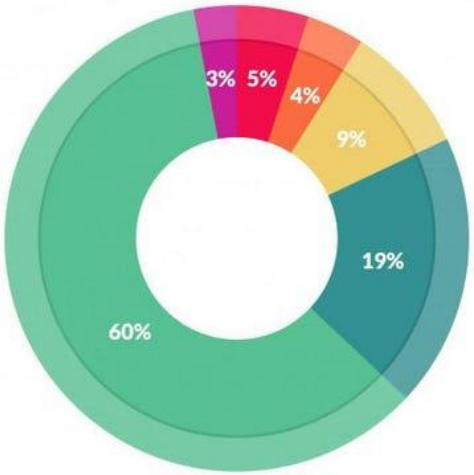
스마트시티 업계 종사자들은 수차례 전시행사를 진행하면서 참관객들이 '디지털 트윈'과 같은 3D 시각화를 선보인 부스에 환호했다는 것을 이미 알고 있다. 시각화 기술이 중요하다는 의견에 필자 역시 동의하지만, 그 이상으로 데이터를 스마트하게 하는 과정이 중요하다.

여전히 데이터를 스마트하게 활용한다는 점이 와닿지 않을 수 있다. 예를 들어 겉보기에는 화려하지만 입으로는 거짓말을 하는 사람이 있다고 가정해보자. 잠깐 바라보는 것은 좋으나 시간이 지날수록 그 매력은 떨어지기

서동재 비투엔 AI연구사업팀 이사
(b2en@b2en.com)

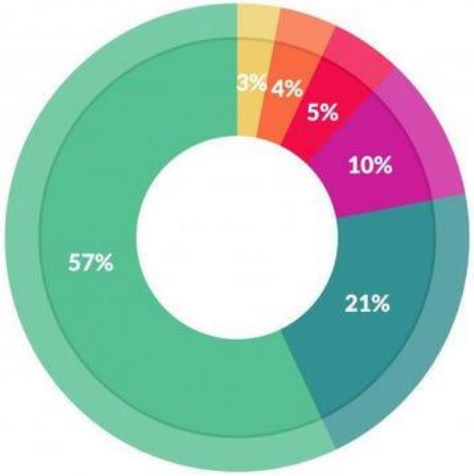
05 데이터 전처리

사례#3



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

데이터 과학자가 많은 시간을 할애하는 작업

데이터 사이언스에서 가장 재미없는 파트

- 데이터 정리 및 구성
- 데이터 세트 수집

각각 79%, 78%의 비중을 차지

06 참고자료

- 사카마키 류지, 사토 요헤이.“비즈니스 활용 사례로 배우는 데이터 분석:R”.한빛미디어.손정도 옮김
- 폴 크리커드.“실무 예제로 배우는 데이터 공학“. 제이펍.류광 옮김
- 유혁(Stephen H. Yu).데이터를 잘 써먹을 수 있는 구체적인 방법들.NIA, IT DAILY
- 김선영.증거기반 정책에서의 빅데이터에 관한 연구.한국정책학회보.제29권.1호(2020.3): 69-90
- 임은택, 김종현, 김광용.“Auto ML 플랫폼 WiseProphet으로 AI 모델 쉽게 개발하기”.청람.2-37
- Kim, D. H., Yoo, S. E., Lee, B. J., Kim, K. T., & Youn, H. Y. (2019). Data preprocessing for efficient machine learning. In Proceedings of the Korean Society of Computer Information Conference (pp. 49-50). Korean Society of Computer Information.
- 데이터 분석절차.Seongkeun.https://velog.io/@osk3856/data-analysis
- 한국경제기사.https://www.hankyung.com/international/article/202103175258i
- 데이터넷 기고문.http://www.datanet.co.kr/news/articleView.html?idxno=173988
- 포브스 기사.https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=2db937fc6f63

Q&A