R & Python을 활용한 빅데이터 분석 기초

빅데이터는 무엇인가?

빅데이터분석TF팀
민종열 대리

빅데이터서비스개발TF팀
김성훈 대리

한국고용정보원
Korea Employment Information Service

PROFILE

민종열 대리

통계학과 졸업

한국고용정보원 빅데이터분석TF팀 재직 중

빅데이터 기획 및 분석 업무 담당

원내 빅데이터 분석 프로젝트 진행

김성훈 대리

LG CNS 빅데이터개발팀(2013.01 ~ 2015.07)

빅데이터 플랫폼 구성 및 운영

한국고용정보원 빅데이터서비스개발TF팀(2018.12~)

고용정보 추천서비스(THE WORK) 운영 및 개발

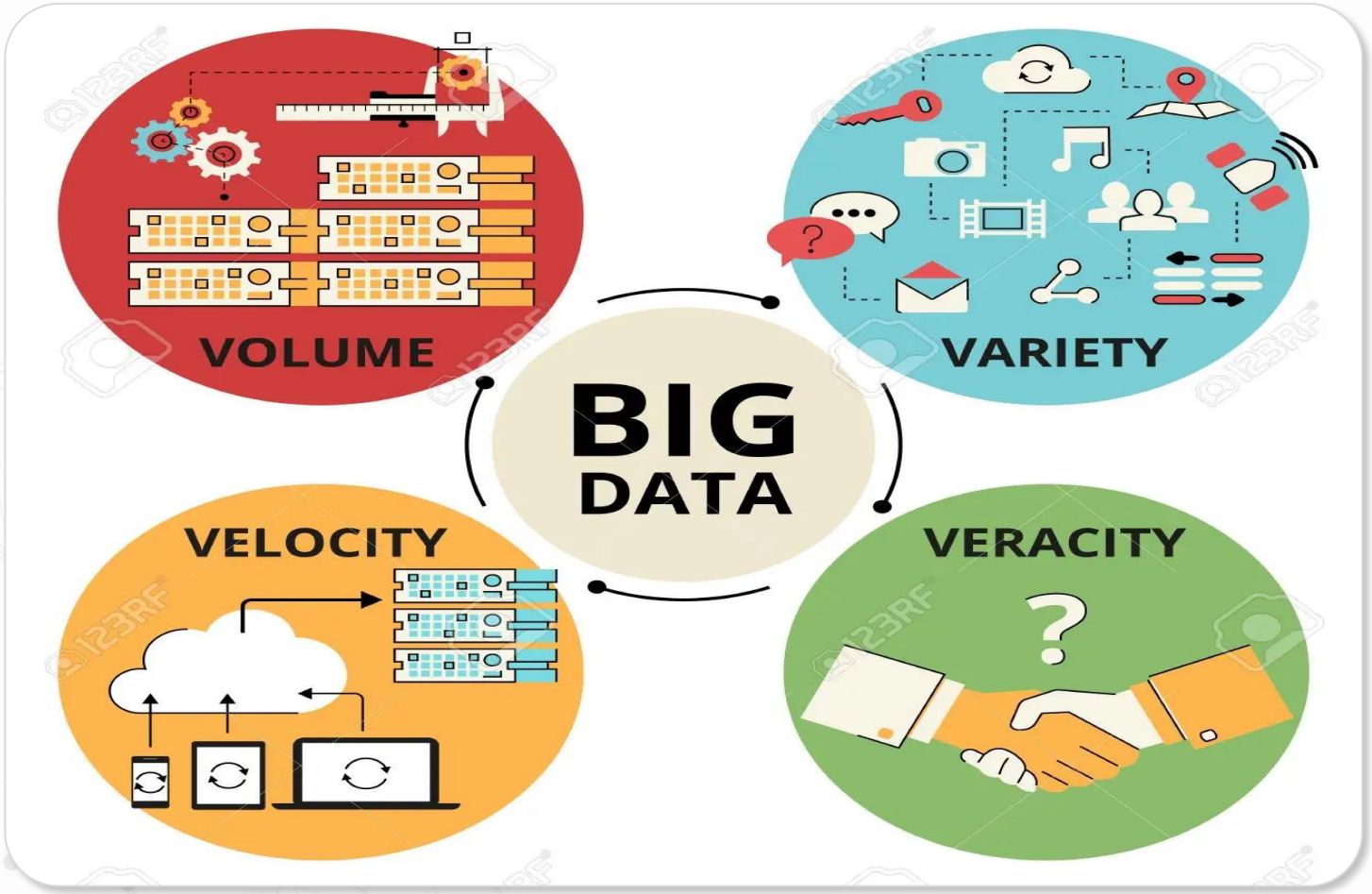
빅데이터 및 AI플랫폼 운영

CONTENTS

Chapter

I 빅데이터란 무엇인가?

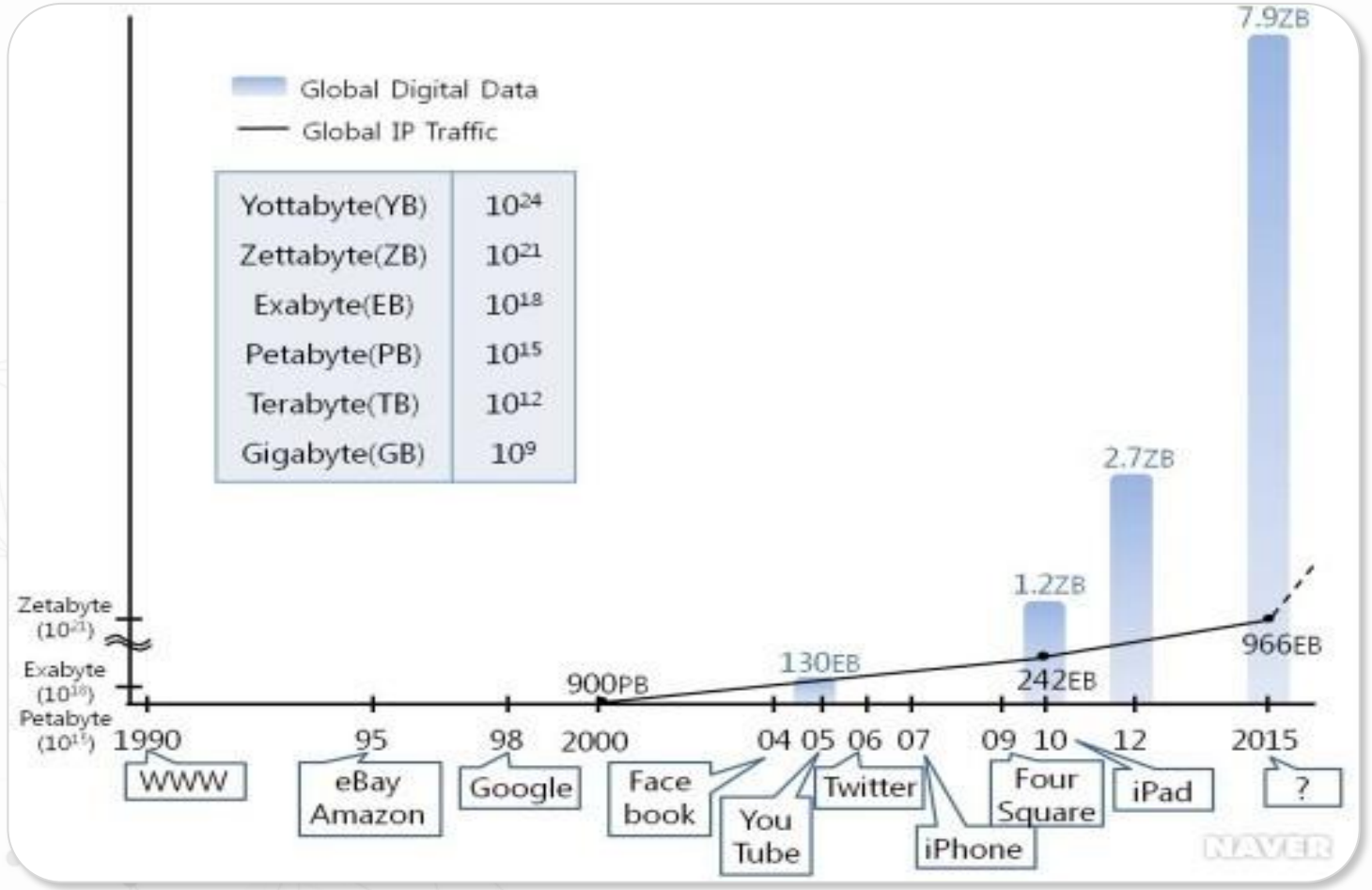
01 빅데이터란?



빅데이터의 3V + V + V

- VOLUME : 양
- VARIETY : 다양성
- VELOCITY : 속도
- VALUE : 가치
- VERACITY : 신뢰성

01 빅데이터 VOLUME

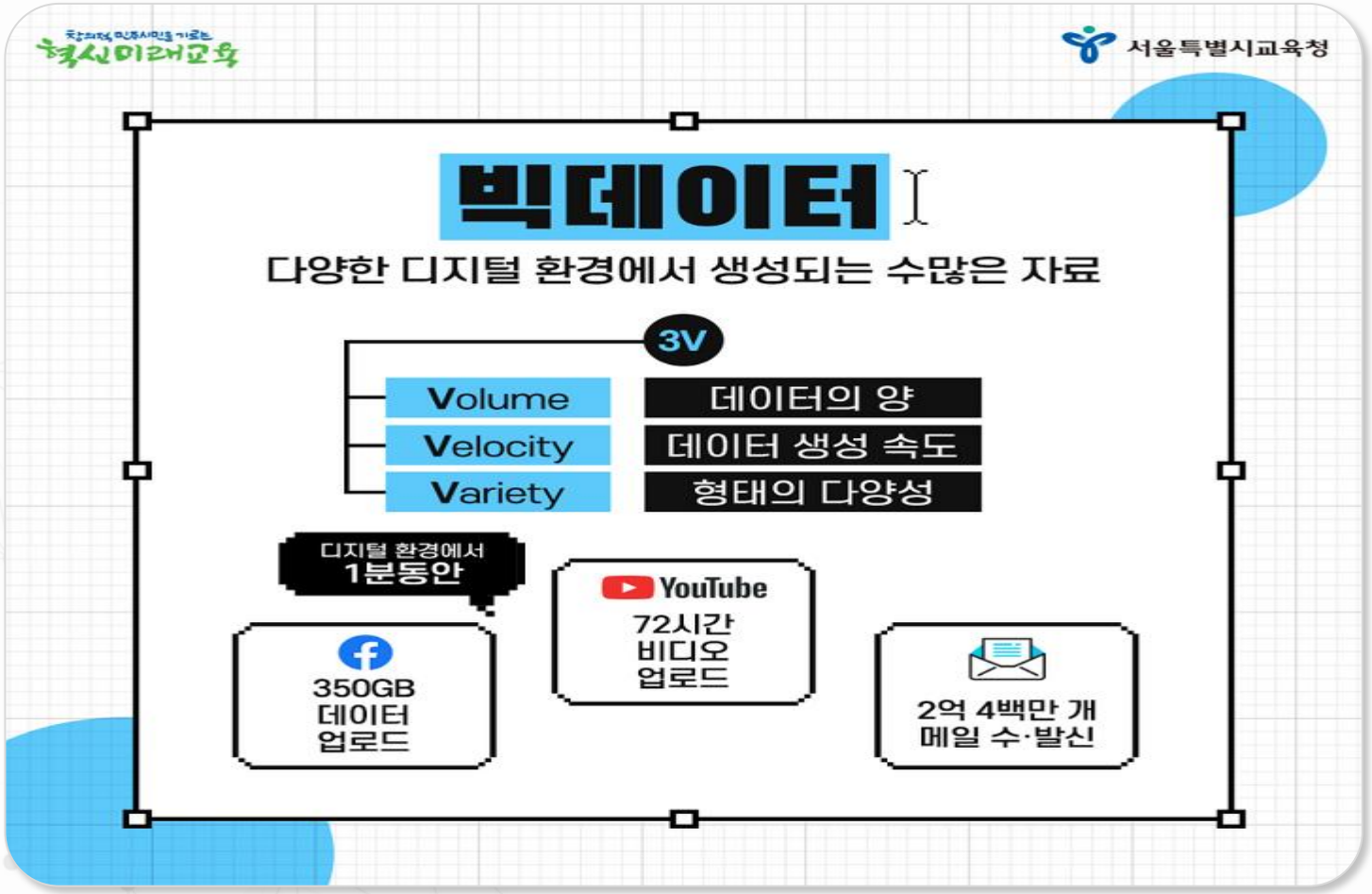


VOLUME(양)

- 2005년 : YOUTUBE : 130EB
- 2010년 : IPAD : 1.2ZB
- 2012년 : 2.7ZB
- 2015년 : 7.9ZB

과연 현재 데이터양은?

01 빅데이터 VARIETY



VARIETY(다양성)

- § YOUTUBE : 영상
- § FACEBOOK : 텍스트, 이미지
- § 메일 : 텍스트
- § 센서데이터 : 로그

미래에 새로 생길 데이터는?

01 빅데이터 VELOCITY

VELOCITY(속도)

- 1980년 : 통화
- 1990년 : 문자
- 2003년 : 문자, 음성, 동영상, 화상통화
- 2009년 : 인터넷
- 2020년 : 인터넷이 3초 안에 실행

2022년 현재는?

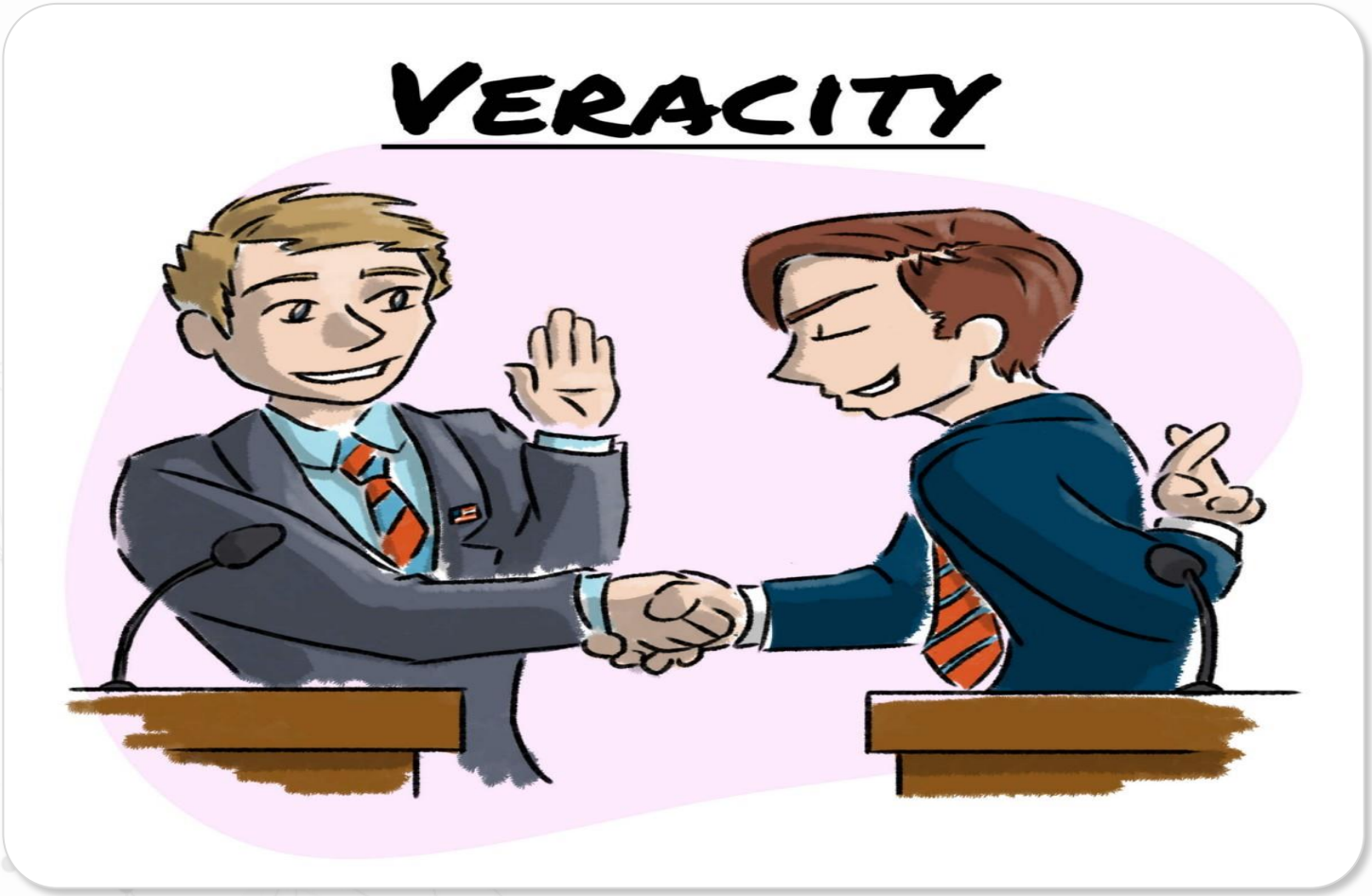
01 빅데이터 VALUE

VALUE(가치)

- § 넷플릭스 추천 알고리즘
- § 유튜브 추천 알고리즘
- § 유튜브 노란딱지 알고리즘
- § 페이스북 광고 알고리즘

빅데이터 활용 가능 여부

01 빅데이터 VERACITY

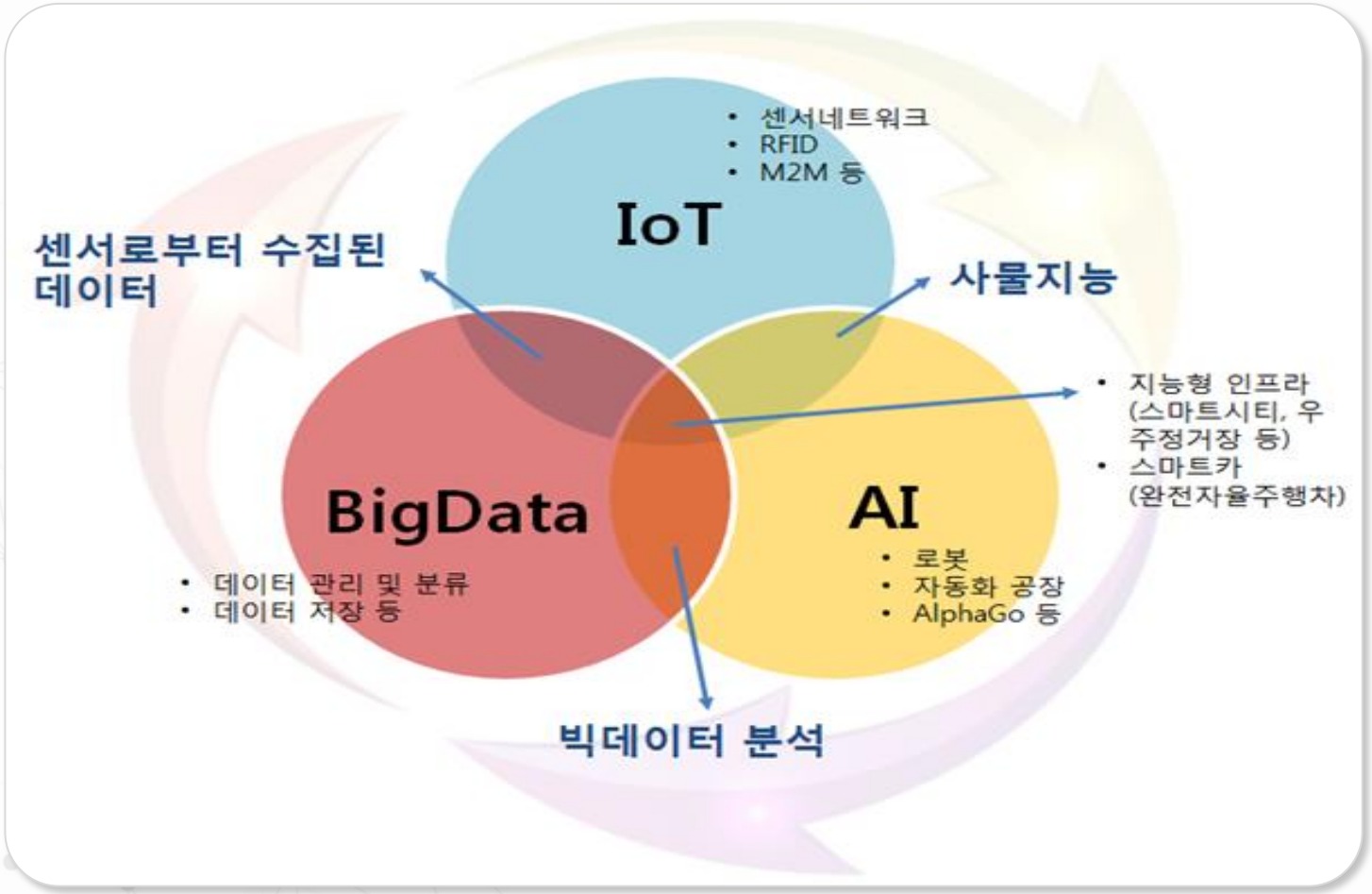


VERACITY(신뢰성)

- 데이터의 신뢰성이 중요해지고 있음
- 데이터가 중요도가 UP!
- 좋은 데이터를 가지고 있는 것이 회사의 자산이 되었음.

통신데이터, 카드데이터 등

01 빅데이터와 연관있는 분야



빅데이터 & AI & IoT

§ 빅데이터 : 관리 및 분류, 데이터 저장

→ 관련 분야 : hive, cloud DB

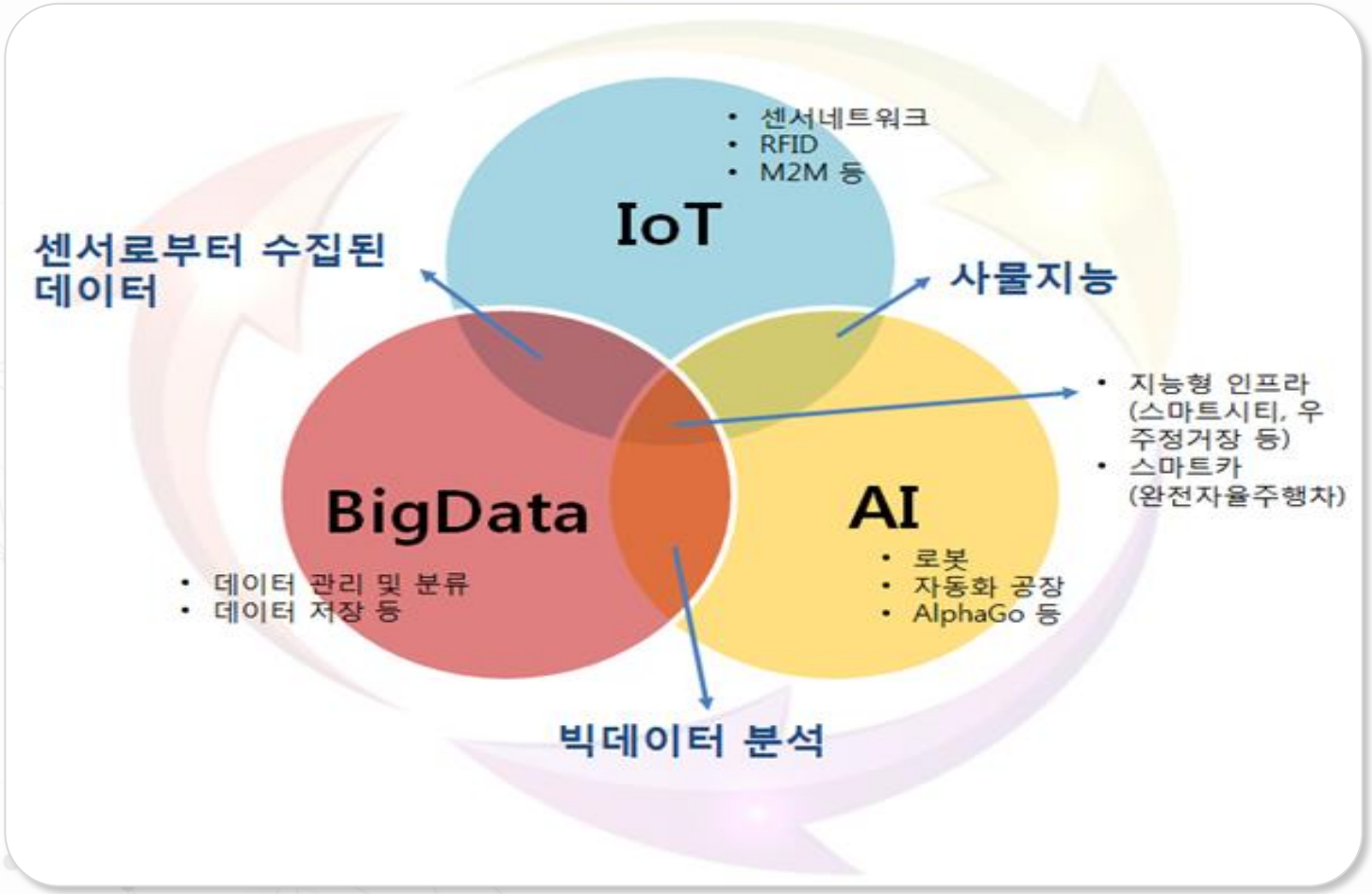
§ AI : AI & 머신러닝 & 딥러닝

→ 관련 분야 : R & Python

§ IoT : Internet Of things(사물인터넷)

→ 관련 분야 : 냉장도 센서

01 현업에서의 활용



빅데이터 & AI & IoT

§ 빅데이터 + AI : 넷플릭스 추천 알고리즘

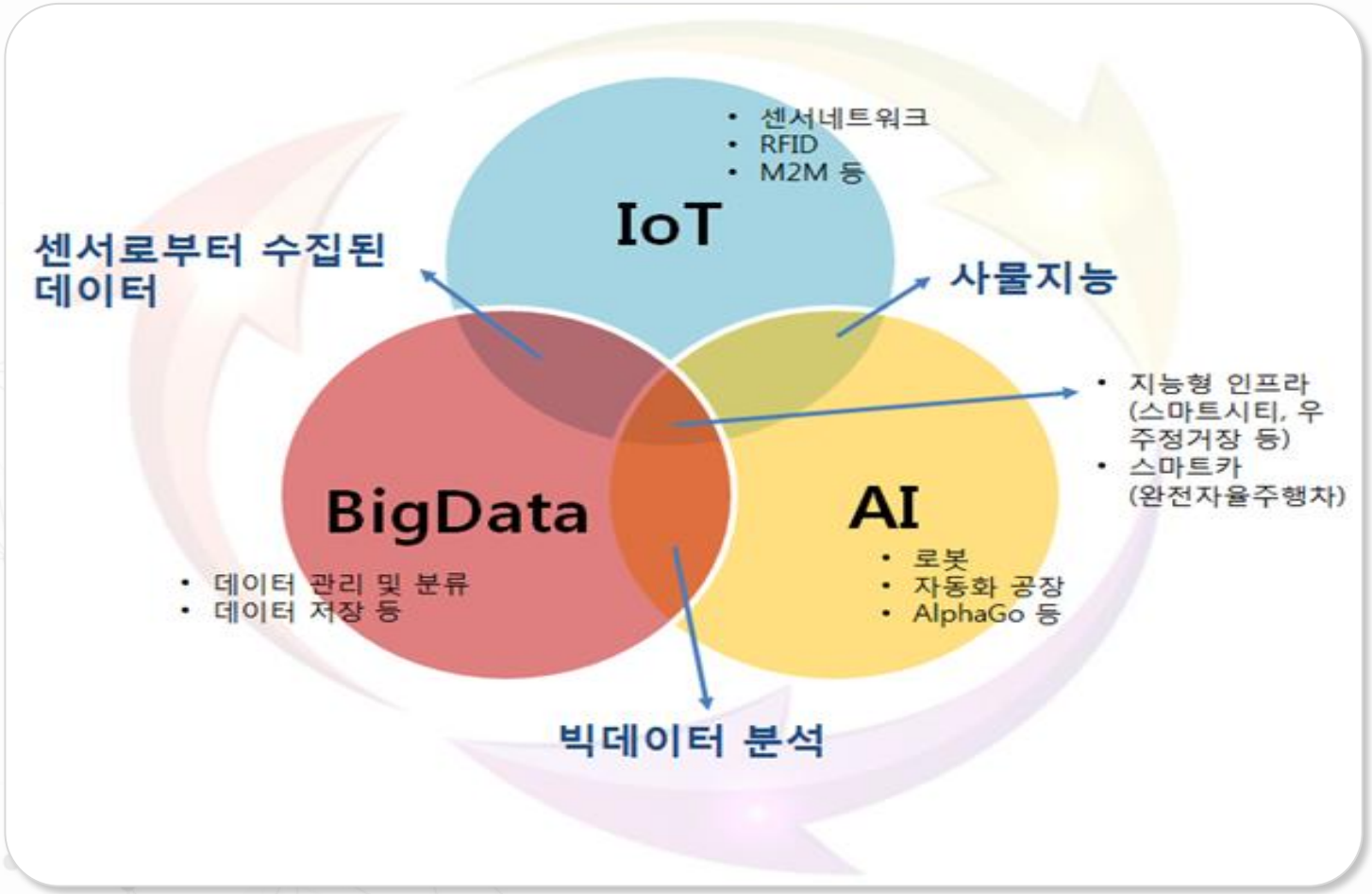
§ IoT + 빅데이터 : 다양한 공간의 데이터 수집

→ 자동차 주행 데이터

§ IoT + AI : 사물지능 (빅스비)

§ 세가지 모두 : 자율주행

01 무서운 예시



만약에 대학교가 이렇다면?

§ 빅데이터 + AI : 시험을 통한 장학금 선발자 선정

§ IoT + 빅데이터 : 모든 좌석에 센서를 달아서 수강 데이터 수집

§ IoT + AI : 센서가 있는 자리에서 조는 경우 교수님께 알람

§ 세가지 모두 : 센서가 있는 자리에 앉아서 졸았던 데이터를 장학금 선발기준에 넣음

01 무서운 예시

빅브라더의 시작

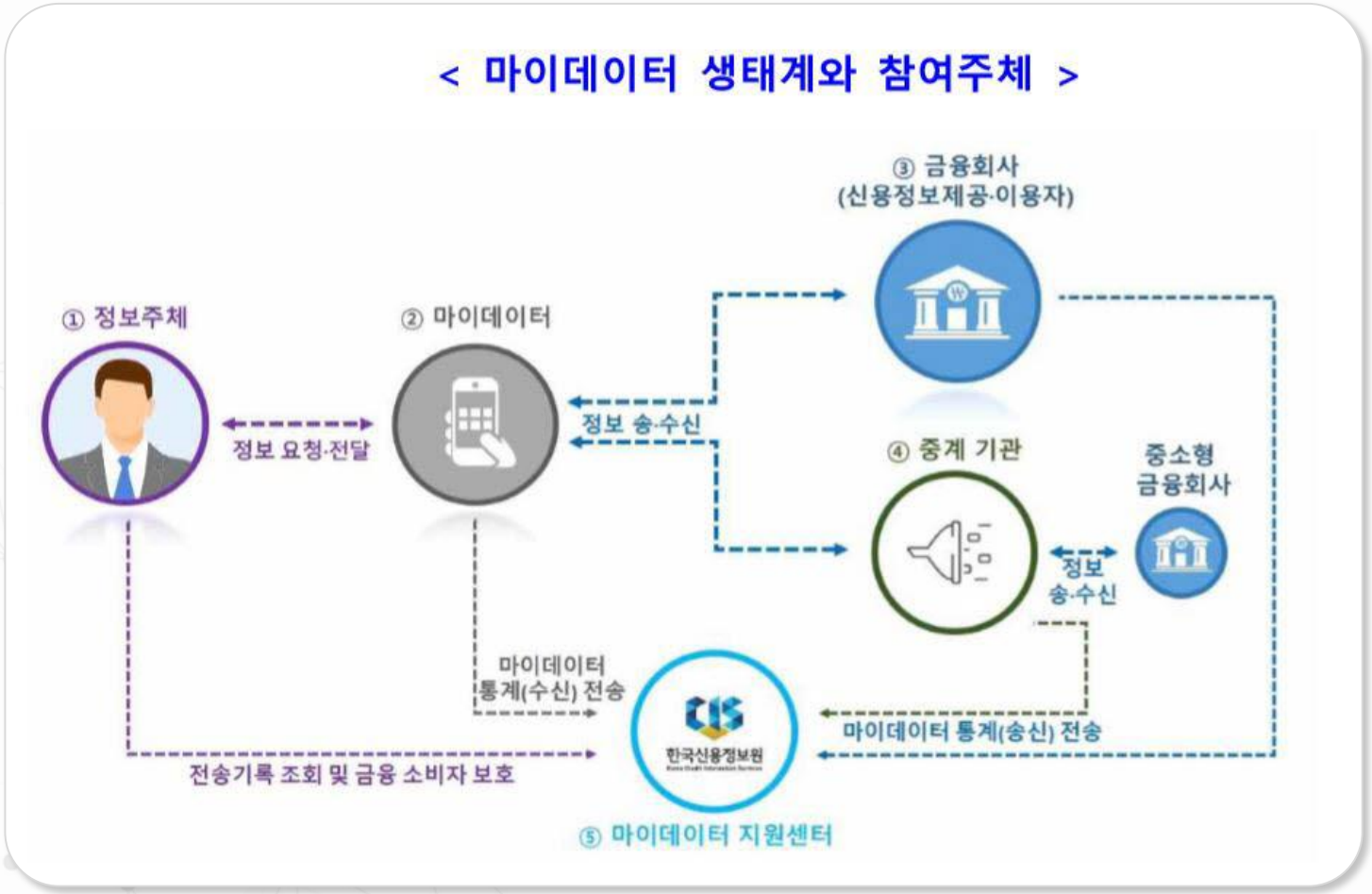
§ 모든 사람들의 데이터는 실시간으로 수집되며
§ 모든 사람들의 데이터는 평생에 걸쳐 저장되며
§ 아무것도 하지 않아도 범죄자로 낙인찍힐 수 있다.

물론! 해당 문제는 철저하게 데이터 보안 및 동의를 통해 관리하고 있음

Chapter

II 현업에서의 빅데이터란?

02 현업에서의 빅데이터 (공공 : 공유, 개방)



공공 빅데이터 (공유, 개방)

§ 데이터는 공유 개방 되어야 한다

§ 데이터는 많은 사람이 활용할 수 있는 형태여야 한다

공유, 개방
-> 마이데이터 사업

02 현업에서의 빅데이터 (공공 : 비식별)

비식별 조치방법

처리기법	조치전	비식별조치후
가명처리	홍길동, 35세, 서울 거주, 한국대 재학	임꺽정, 30대, 서울 거주, 국제대 재학
총계처리	임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm	물리학과 학생키 합 660cm, 평균키 165cm
데이터삭제	주민등록번호 901206-1234567	90년대생, 남자
데이터범주화	홍길동, 35세	홍씨, 30~40세
데이터 마스킹	홍길동, 35세, 서울 거주, 한국대 재학	홍OO, 35세, 서울 거주, OO대학 재학

*자료: 개인정보 비식별 조치 가이드라인

그래픽: 이승현 디자이너

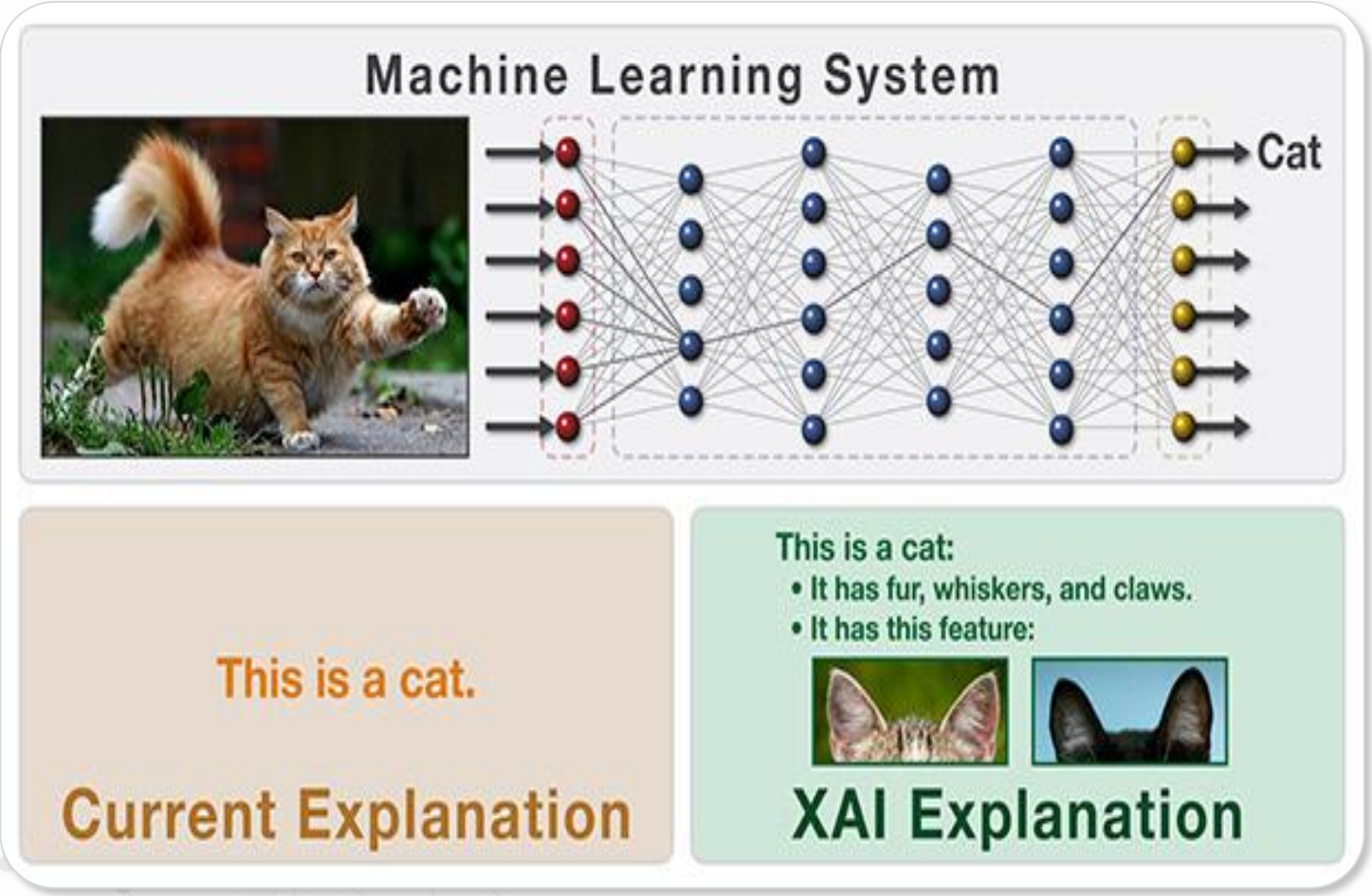
공공 빅데이터 (비식별)

§ 데이터를 통해 누구도 유추되어서는 안 됨

§ 데이터가 공유됨으로 개인정보 누출이 되어서는 안 됨

데이터 3법 (익명 처리)

02 현업에서의 빅데이터 (공공 : 설명 가능 알고리즘)



공공 빅데이터 (XAI)

- XAI : eXplanable AI (설명가능한 AI)
- 왜 그러한 결과가 나왔는지 설명이 가능해야 한다.
- Example : 장학금 선발자를 알고리즘으로 구현한다면 왜(!)에 중심을 둬야 한다.

모든 결과에는 설명이 필요함!

02 현업에서의 빅데이터 (기업 : 결과론적 알고리즘)

기업 빅데이터 (결과론적 알고리즘)

- YOUTUBE 알고리즘 : 나한테 이런걸 왜 추천 해 주는 거지?
- 근데 나는 왜 이걸 보고 있지?
- 근데 왜 이게 재미있지?

결과를 잘 맞추면 GOOD!

02 현업에서의 빅데이터 (기업 : 새로운 경험)

주요 카드사별 빅데이터 분석 활용 서비스

카드사	서비스 명	주요 특징
삼성카드	LINK 비즈파트너	중소가맹점주가 가맹점 전용 홈페이지에 고객에게 제공할 혜택을 직접 등록하면, 삼성카드가 빅데이터 분석 시스템인 스마트 알고리즘을 기반으로 이용 가능성이 높은 고객에게 맞춤형 정보 제공
신한카드	마이샵 파트너	빅데이터 분석 통해 고객이 필요한 쿠폰 혜택 등을 제안, 신한FAN 앱을 통해 마이샵이 추천한 혜택 확인 가능, 선택한 제안은 가맹점에서 카드 결제 시 자동 적용
BC카드	빅데이터 분석 보고서 즉시 발급 서비스	고객이 분석 요청하는 분야를 선택하고 지역, 기간, 주제어 등 세부 정보 입력하면 1시간 이내에 자동으로 보고서 작성
현대카드	피코	고객들이 자주 검색하거나 결제한 품목의 정보를 바탕으로 개인의 취향에 최적화한 해외 패션 사이트 추천
롯데카드	롯데카드 라이프 앱	고객을 200여개 선호 지수로 분류. 위치, 상황, 경험 등 다면적 빅데이터 분석을 통한 '초개인화 서비스'를 적용
KB국민카드	스마트오퍼링 시스템	고객의 향후 소비를 미리 예측해 최적의 시간에 맞춤형 혜택을 제공

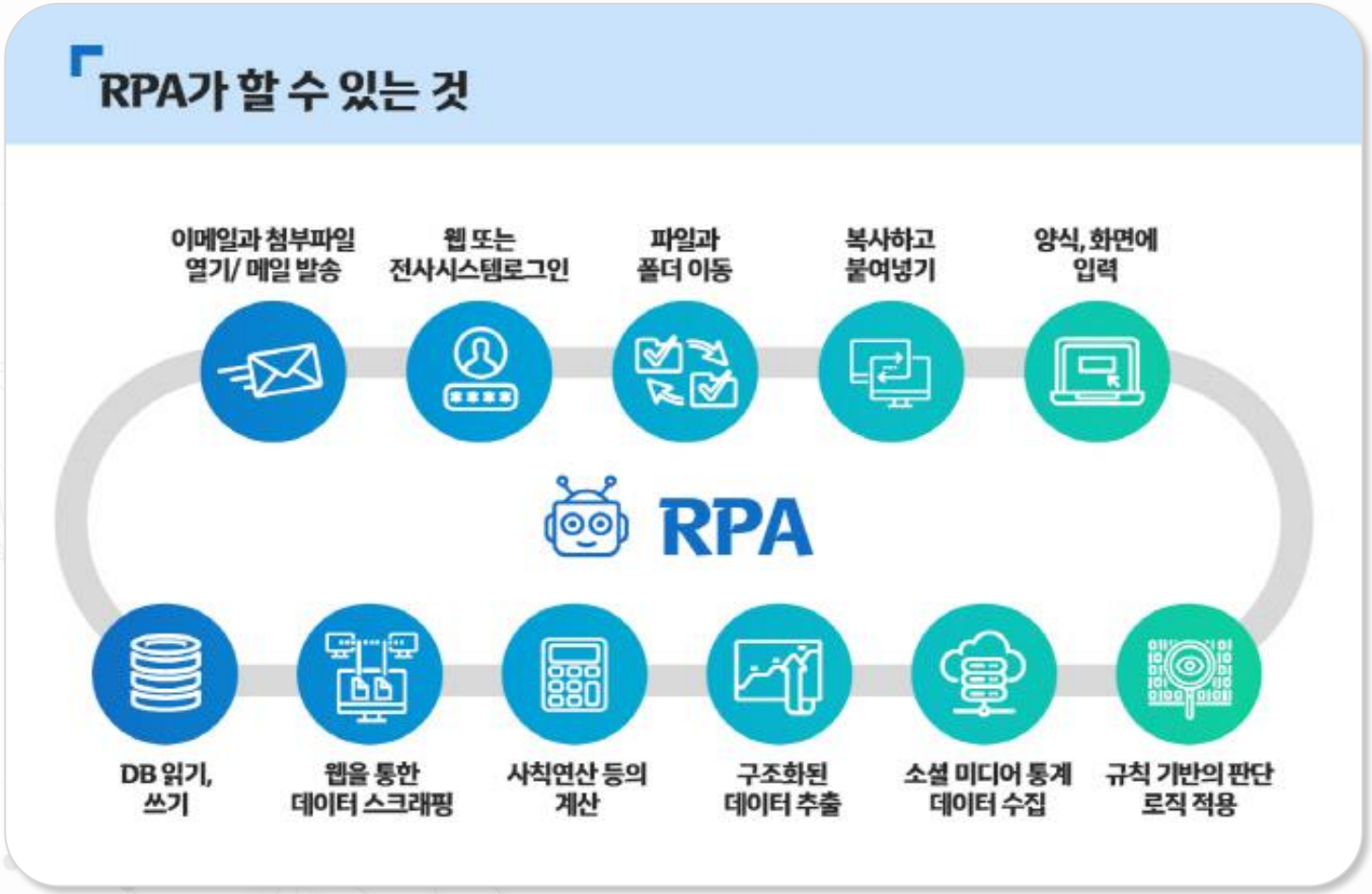
자료: 각 사

기업 빅데이터 (새로운 경험)

- 적은 예산을 효율적으로 활용하여 고객에게 더 적절한 혜택을 주고 (카드사 혜택 변경)
- 적은 예산을 활용해 새로운 경험을 준다. (해외 패션 사이트 추천)

새로운 경험을 통해 새로운 고객을 유치할 수 있음!

02 현업에서의 빅데이터 (회사 내부)



사무업무자동화 (RPA)

- 반복적인 업무를 자동화 함으로써 인력을 고부가 가치 사업에 활용할 수 있음
- 사람이기에 할 수 밖에 없는 실수를 줄일 수 있음
- 짧은 구축 시간

회사 내부적으로도 활용 가능!

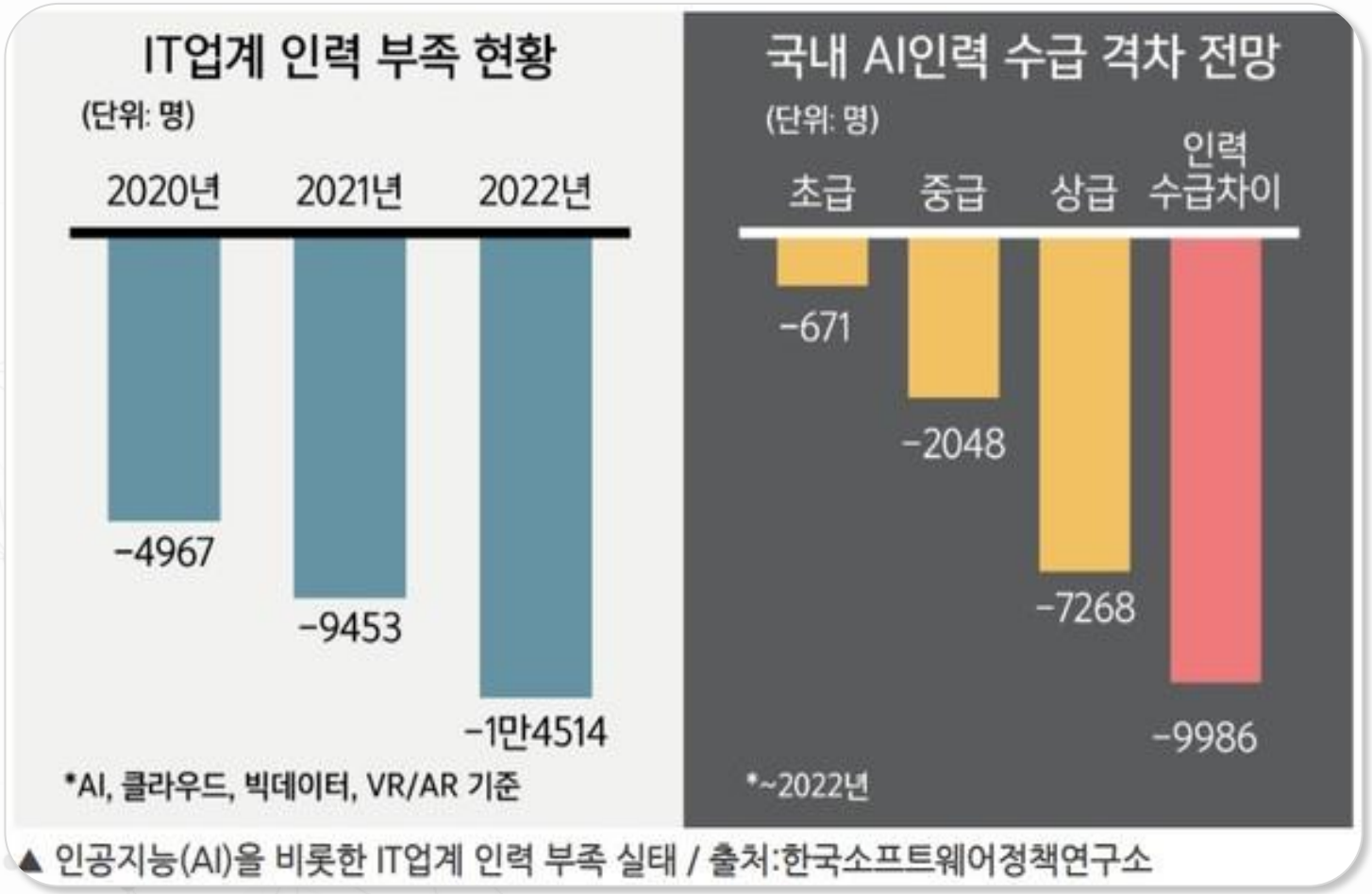
02 빅데이터의 중요성

기업 CEO의 외침

§ 빅데이터는 공공, 기업 모든 영역에서 가장 중요한 가치로 떠오르며 데이터를 많이 가지고 있는 기업이 미래의 선두주자가 될 수 있다고 수많은 기업 CEO가 얘기를 함

미래에는 데이터를 지배하는 기업이 모든 것을 얻는다!

02 빅데이터 인재 부족



▲ 인공지능(AI)을 비롯한 IT업계 인력 부족 실태 / 출처:한국소프트웨어정책연구소

빅데이터 인재 부족

§ 빅데이터의 중요성에 반해 빅데이터 인력은 시간이 지날수록 부족하다.

§ 이유 : 잘하는게 어려움

§ → 다양한 분야에서의 지식이 필요

§ → 데이터 저장, 관리, 분석, 해석 등

빅데이터를 잘하기 위해서는 너무나도 할 것이 많다.

Chapter

III 빅데이터 역량을 위한 준비

빅데이터 직군 및 관련 전공

데이터 아키텍처

- 데이터 요건 분석
- 데이터 표준화
- 데이터 모델링
- 테이터베이스 설계

산업공학과 컴퓨터공학

빅데이터 관리

- 데이터 수집
- 데이터 저장
- 데이터 관리
- 데이터 추출

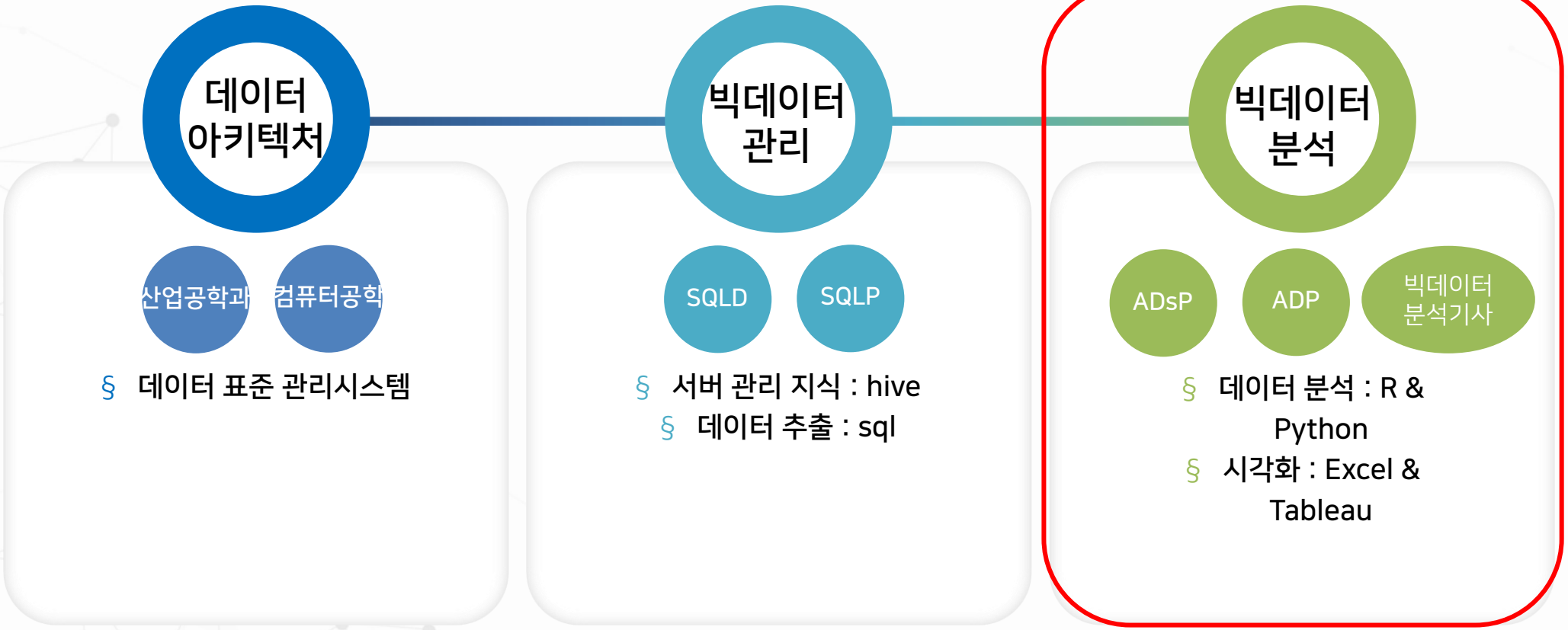
산업공학과 컴퓨터공학

빅데이터 분석

- 빅데이터 분석 기획
- 빅데이터 분석
- 시각화
- 해석

통계학과 수학과

빅데이터 사용 Tool



빅데이터 관련 자격증

데이터 아키텍처

DAsP　DAP

§ DAsP : 30~40%
§ DAP : 10%

빅데이터 관리

SQLD　SQLP

§ SQLD : 30~40%
§ SQLP : 10%

빅데이터 분석

ADsP　ADP　빅데이터 분석기사

§ ADsP : 40%
§ ADP : 2%
§ 빅데이터분석기사 : 20~30%

03 데이터 분석가로 활동하기 위해서!

저의 경우는

- 관련 전공을 수학
- 자격증 획득
- 공모전 참가 및 수상
- 빅데이터 프로젝트 진행
- 관련 직무로 인턴 참가

과연 비전공자가 데이터분석가로 활동할 수 있을까?

03 냉정한 현실

비전공자는

- 전공자에 비해 전공 지식이 부족
- 실무 경험이 부족
- 관련 전공에 비해 기회가 적음

비전공자가 데이터 분석 업무를 직무로 얻기 위해서는 더 많은 노력이 필요하다.

그럼 나는 비전공자인데 왜 빅데이터를 배워야하지? 직무로 갈 수도 없는데?

02 비전공자가 빅데이터를 배워야 하는 이유!

신입 채용

§ IT직군이 아닌 직군에도 디지털 리터러시 역량을 요구

§ 디지털 리터러시 : 디지털을 이해하고 다룰 줄 아는 디지털 활용 능력

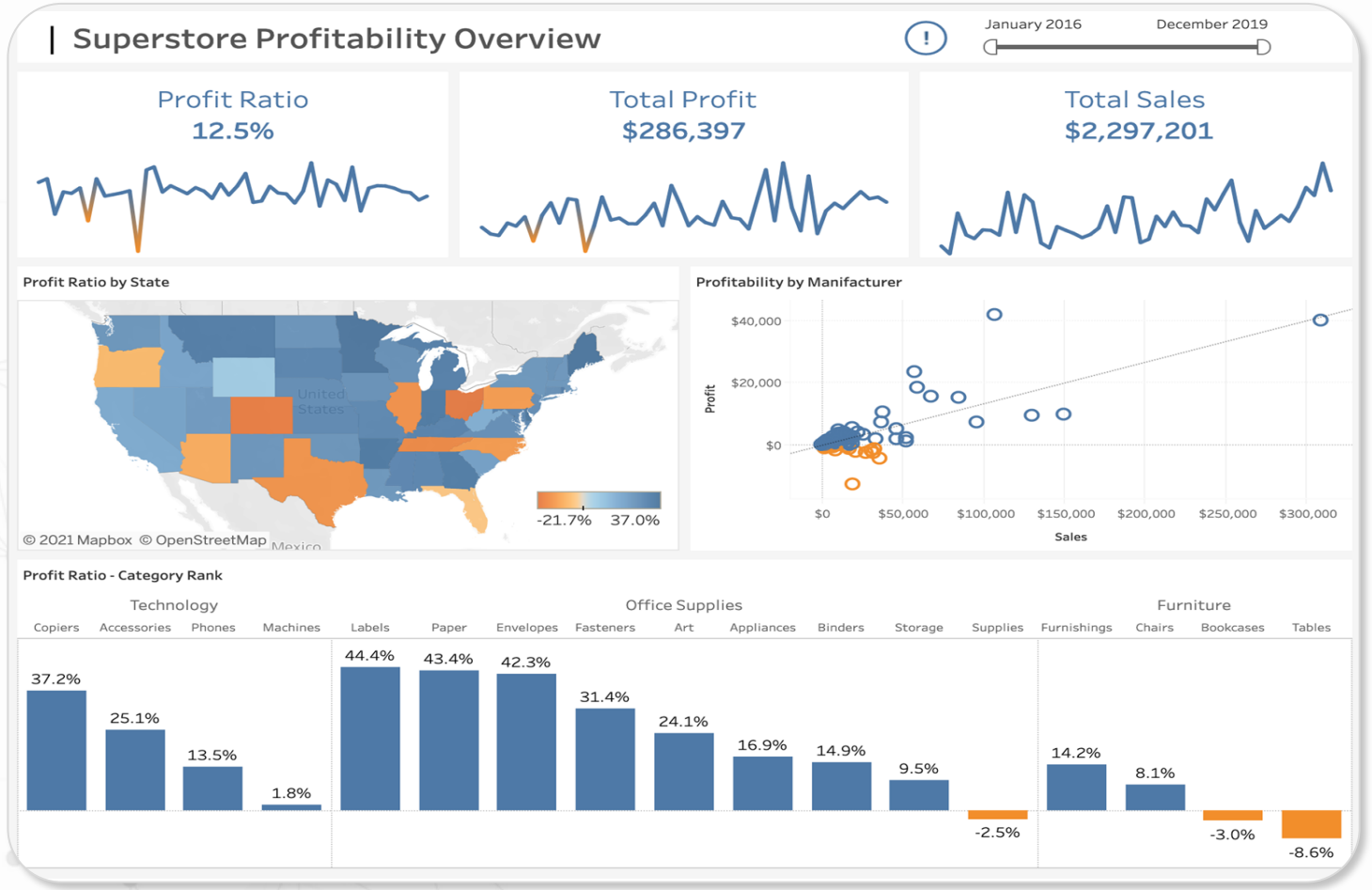
디지털 역량을 IT직군이 아닌 직군에서도 우대한다.

비전공자도 빅데이터를 안다면 우대를 받는다!

Chapter

Ⅳ 디지털 역량 어필 방법

02 자기 영역에서 빅데이터 활용할 수 있는 방안을 찾아라!



매출 관리자

- 어느 매장이 가장 매출이 안좋을까?
- 어느 품목에서 매출이 안좋을까?
- 한달 뒤에 적자가 날 확률이 높은 매장은 어딜까?
- 적자가 나지 않기 위해서는 어떤 조치를 해야 될까?

<도메인 지식>
매출 관리자 > 데이터 분석가

02 자기 영역에서 빅데이터 활용할 수 있는 방안을 찾아라!

인사 관리자

- § 어떤 직원이 가장 성과가 좋을까?
- § 어떤 직원이 승진 대상자지?
- § 해당 직무를 수행하기 위해 가장 필요한 지식은?
- § 어떤 직원이 퇴직을 고민하고 있을까?

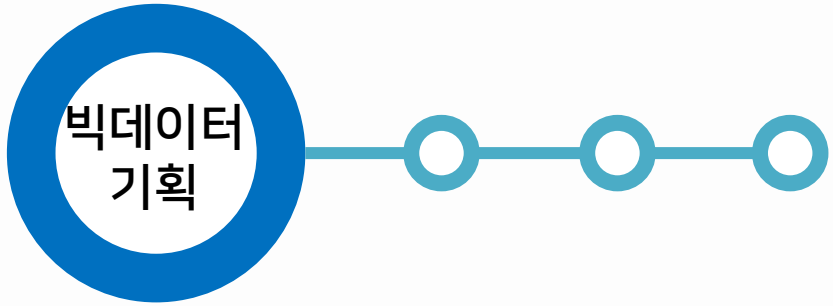
<도메인 지식>
인사 관리자 > 데이터 분석가

적어도 여러분 전공 부분에서 여러분들이 저보다 지식이 많다!

04 이렇게 말해보자!

총무 담당자 신입 채용

<질문>
총무일을 개선할 수 있는 방안이 있나?

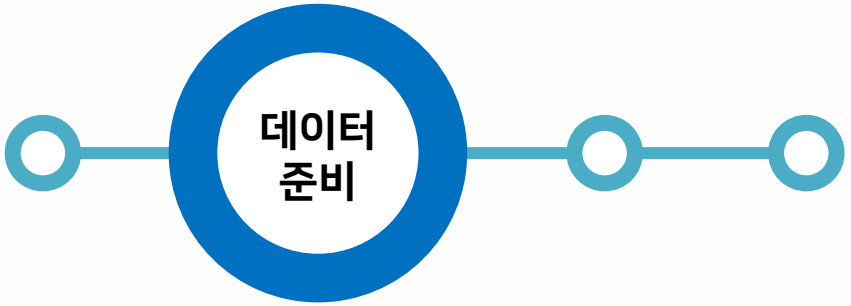


빅데이터 기획	최근에 사무행정자동화(RPA)라는 개념에 대해 알게 되었습니다.
목적	그것을 저희 총무일에 대입해보고 싶습니다.
이유	총무일에는 매월 동일하게 진행되어야 하는 회의, 결산 등의 일이 있습니다.
상세 내용	이렇게 반복적인 업무를 개선하여 효율성을 높이는 내용이 필요합니다.
내용	이 목록을 바탕으로 사무행정자동화를 도입, 운영을 제안합니다.

04 이렇게 말해보자!

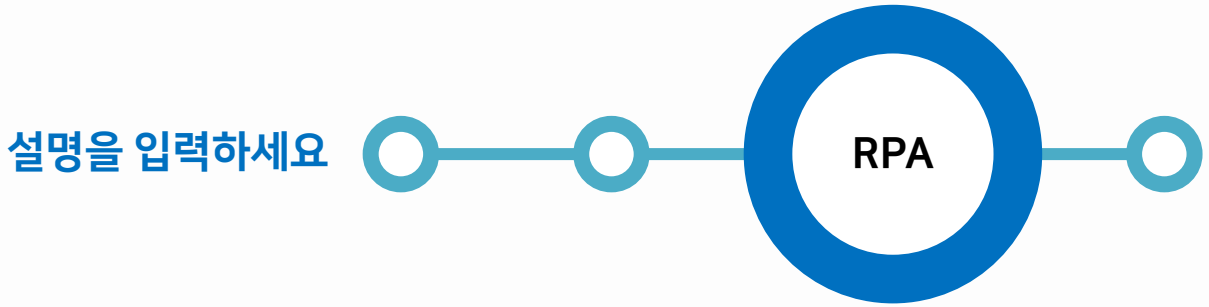
총무 담당자 신입 채용

<질문>
총무일을 개선할 수 있는 방안이 있나?



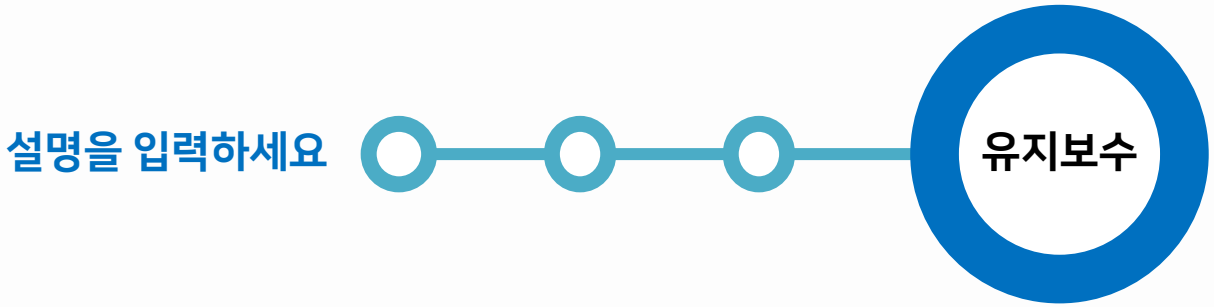
내용	주기적으로 반복되는 일을 조사하겠습니다.
내용	반복되는 일을 목록으로 관리하겠습니다.
내용	반복되는 일의 패턴을 파악하겠습니다.
내용	Python을 활용하겠습니다.

04 이렇게 말해보자!



내용	반복작업의 패턴을 바탕으로 Python 코드를 작성하겠습니다.
내용	Python 코드를 실행하여 자동으로 실행되는 확인해보겠습니다.
내용	실행 속도를 확인하여 우리가 원하는 속도가 나오는지 확인하겠습니다.
내용	오류를 수정하겠습니다.
내용	총무과에 배포하겠습니다.

04 이렇게 말해보자!



내용	새로운 반복작업이 있을 경우 추가로 작업하겠습니다.
내용	패턴이 변경된 반복작업의 경우 업데이트 하겠습니다.
내용	현재 사용되지 않는 작업의 경우 제거하겠습니다.
내용	원하는 속도가 계속해서 나올 수 있도록 관리하겠습니다.
내용	회사내 다양한 팀의 요구를 반영하겠습니다.

04 담당자 반응

Chapter

V

R & Python 설치 방법

05 R & R studio 설치 방법

R & R studio 설치

- https://iamoverthemoon.tistory.com/64
- R 설치방법 URL

R 자체는 사용하기 불편
R studio 추가 설치 진행

05 Python 설치 방법

Python 설치

- https://benn.tistory.com/26
- 아나콘다 다운로드 방법 URL

05 우리가 사용할 플랫폼은?

Google Colab

- Google Colab
- Google에서 제공하는 분석 플랫폼
- R & Python 모두 사용 가능
- 많은 데이터 분석가가 사용하는 플랫폼

05 2회차 강의 진행 미리보기

1 빅데이터 분석과정(이론)

- 빅데이터 기획
- 빅데이터 분석
- 평가
- 해석

2 R 자료형태 및 객체(이론&실습)

- 자료 형태
- 객체
- 실습

3 R 조건문, 반복문, 분기문(이론&실습)

- 조건문
- 반복문
- 분기문

4 R 데이터 가공하기 (이론&실습)

- R 데이터 가공하기
- 표 그려보기

강의을 마치며

궁금한게 있으시면

Name : 민종열 대리

Email : wpdntm3001@naver.com

Hp : 010-5439-5931

감사합니다

한국고용정보원
Korea Employment Information Service