02

빅데이터 활용

학습
내용

01 빅데이터의 활용 사례

02 빅데이터 분석 기획

03 데이터 마이닝의 개요

학습 목표

- 빅데이터의 다양한 활용 사례를 설명할 수 있다.
- 빅데이터 분석 기획의 개념과 분석 방법론을 설명할 수 있다.
- 데이터 마이닝의 개념과 방법을 설명할 수 있다.


생각
해보기
1996년 인공지능 딥블루의 체스 대결
“컴퓨터는 결함 없는 계산에 특화돼 있고, 인간의 뇌는 장기적이고
광범위한 계획 수립 및 다양하고도 일반적인 주제들을
새로운 환경에 접목시키는 것이 탁월하다.”

01

빅데이터의 활용 사례

1) 빅데이터의 활용 분야
2) 빅데이터 활용 사례
3) 빅데이터의 활용 단계

1) 빅데이터의 활용 분야

- 빅데이터의 활용 분야

물류 · 배송 · 운송 산업	물류 · 운송 최적화
소매업	가격 최적화
미디어 엔터테인먼트	지적 자산 관리
제조업	품질 보장 관리
치안 당국	범죄예방 및 수사
보험산업	예측적 피해평가
은행	사기 탐지 및 자금 세탁 탐지
의료서비스산업	환자관리 및 사기 탐지

2) 빅데이터 활용 사례

- 기업의 의사 결정
- 실생활 속의 빅데이터
- 선거 운동
- 온라인 쇼핑
- 의료 분야의 빅데이터

▶ 기업의 의사 결정 - ZARA의 애자일 공급망

항공운송 및 자체 생산 등을 통하여 의류의 디자인 선정, 원자재 구매, 생산, 출고, 진열까지의 공급망 리드타임(Lead-time)을 3주 전후로 줄여 업계의 경쟁 우위를 확보

전 세계 매장의 POS 기기, 온라인 판매, 설문조사, PDA 기기, 의류에 부착된 RFID 등으로부터 수집된 데이터를 데이터 분석 전문가들이 분석함

시장의 변화에 즉각 대응하는
QR 시스템(Quick Response System)을 갖춤

▶ 실생활 속의 빅데이터 - 빅데이터 기반 주차 요금제

미국 LA는 제록스와 빅데이터 분석 알고리즘을 개발
LA 시내 주차장 정보를 실시간으로 수집 · 분석

- ✓ 운전자는 시간 · 날짜 · 계절 · 이벤트 별로 다른 요금을 지불함
- ✓ 빅데이터를 활용한 차등 요금제로 혼잡시간대 주차율을 10% 줄임

시내 공영주차장 주차율은 60%로 감소
주차료 수입은 2.4% 증가 효과

▶ 실생활 속의 빅데이터 - 넷플릭스

1998년 비디오와 DVD 대여 서비스로 시작함

현재 190여 개국에 서비스를 제공하고 전 세계 1억 6,700만 명의 유료 고객을 보유한 세계 최대 동영상 서비스 업체로 성장

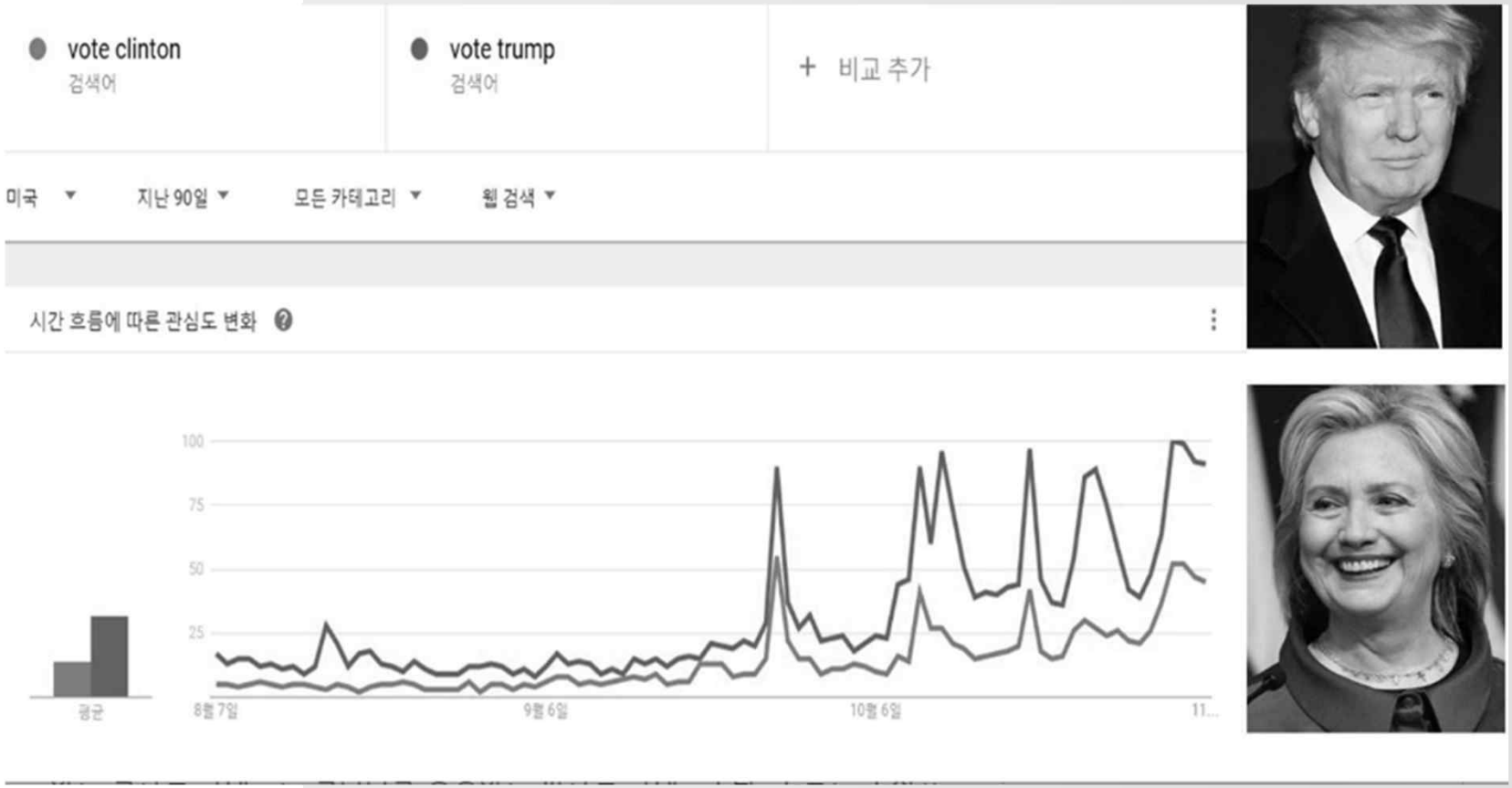
가입자의 시청 습관을 방대하게 수집하여
시청자 선호도를 파악한 후
연출, 배우, 기획, 배급까지 선정

▶ 실생활 속의 빅데이터 - 넷플릭스

사용자 개인 정보가 아닌 영상 사용 패턴에 맞는 취향을 분석하여 추천

- ✓ 2015년부터는 영상 분류별 최적화 기술을 도입
- ✓ 장르에 따라 데이터 양과 전송 속도를 구분하여 서비스

▶선거 운동 - 미국 대선 트럼프의 승리



〈출처 : 우종필 교수"트럼프 승리 적중... 빅데이터는 이미 알았죠', MK, 2016. 11. 10〉

▶ 선거 운동 - 미국 대선 트럼프의 승리

미국 대선이 6개월이 채 남지 않았을 때 트럼프 선거 본부에 합류

✓ 캠브리지 애널리티카는 3가지의 데이터를 이용하여 승리를 이끎

정치 데이터	공개 데이터	퍼스트 파티 데이터
유권자가 그 동안 투표했던 정당 등	소비자 구매 트랜드 데이터, 인구통계 데이터, 지리적 데이터 등	여론 조사, 시장 조사, 모델링 결과에 토대로 계속 증가

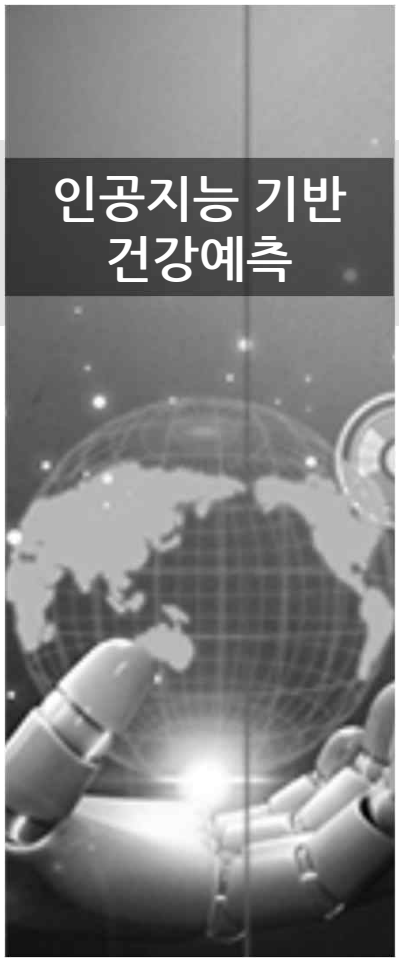
▶온라인 쇼핑 - 아마존 결제 예측 배송

amazon

고객의 주문을 미리 예측해서 가장 가까운 창고에
보관하고 있다가 주문이 접수되는 순간
바로 배송을 시작하는 서비스

- 과거 구매 정보 등을 활용하여 구매 예상 품목을
  고객과 가까운 물류센터로 보냄
- 고객이 물품 주문
- 고객 구입 물품 배송(배송 기간 단축)

▶ 의료 분야의 빅데이터 - 인공지능 기반 건강예측



국민건강보험공단은
인공지능(AI) 기반 건강예측 사업을 추진

- ✓ 지능형 질환 예측모형 개발을 위한 기계학습용 지식베이스를 구축
- ✓ 일반(생애전환기) 건강검진 결과 다면분석 및 판정 알고리즘도 개발해 구현
- ✓ AI 기반 데이터를 건강 iN 홈페이지 및 모바일 앱에 이식해서 대국민 서비스

3) 빅데이터의 활용 단계

01 어떤 일이 있었나?
(What happened)

02 정확히 무엇이 문제인가?
(Where exactly is the problem?)

03 앞으로 예상되는 일은?
(What is happening next?)

04 무엇이 최선의 해결책인가?
(What's the best that can happen?)

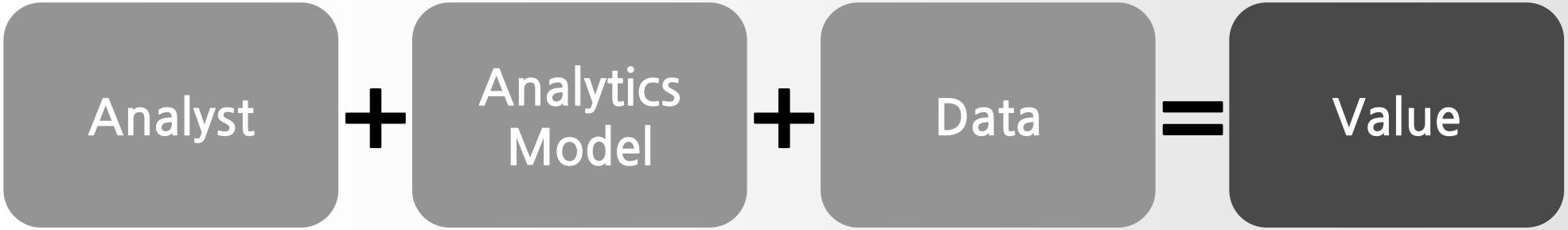
02

빅데이터 분석 기획

1) 분석 기획의 개요
2) 분석 방법론의 개요
3) KDD 분석 방법론
4) CRISP-DM 분석 방법론
5) 빅데이터 분석 방법론

1) 분석 기획의 개요

- 데이터 분석의 요소



데이터와 분석 모델을 활용하여 가치 있는 분석결과를 도출하고 관계자에게 분석 결과를 전달

데이터를 활용하여 가치를 창출할 수 있도록 도와주는 제반의 분석 기법

분석의 근간이 되는 원천 데이터

데이터 분석의 결과 및 목표

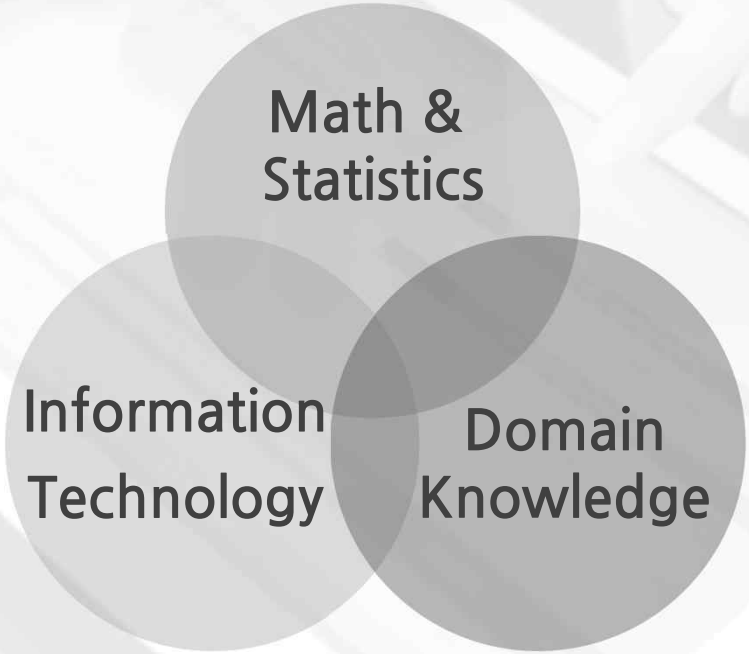
1) 분석 기획의 개요

분석 기획

실제 분석을 수행하기에 앞서 분석을 수행할 과제를 정의하고, 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획하는 일련의 작업

↳ 어떠한 목표(What)를 달성하기 위하여 (Why) 어떠한 데이터를 가지고 어떤 방식으로(How) 수행할 지에 대한 일련의 계획을 수립하는 작업이기 때문에 성공적인 분석 결과를 도출하기 위한 중요한 사전 작전

1) 분석 기획의 개요



데이터 사이언티스트의 역량

해당 문제 영역에 대한 전문성 역량 및
수학/통계학적 지식을 활용한
분석 역량과 분석의 도구인
데이터 및 프로그래밍 기술 역량에
대한 균형 잡힌 시각을 가지고
방향성 및 계획을 수립해야 함

1) 분석 기획의 개요

분석 주제 유형

분석의 방법 (How) \ 분석의 대상(What)	Known	Un-knows
Known	Optimization	Insight
Un-knows	Solution	Discovery

1) 분석 기획의 개요

● 분석 기획 시 고려 사항

가용 데이터(Available Data)

분석을 위한 데이터의 확보가 필수적이며,
데이터 유형에 대한 분석이 먼저 이루어져야 함

적절한 유스케이스(Use Case)

기존에 잘 구현되어 활용되고 있는 유사 분석
시나리오 및 솔루션을 최대한 활용하는 것이 중요

"바퀴를 재발명하지 마라"

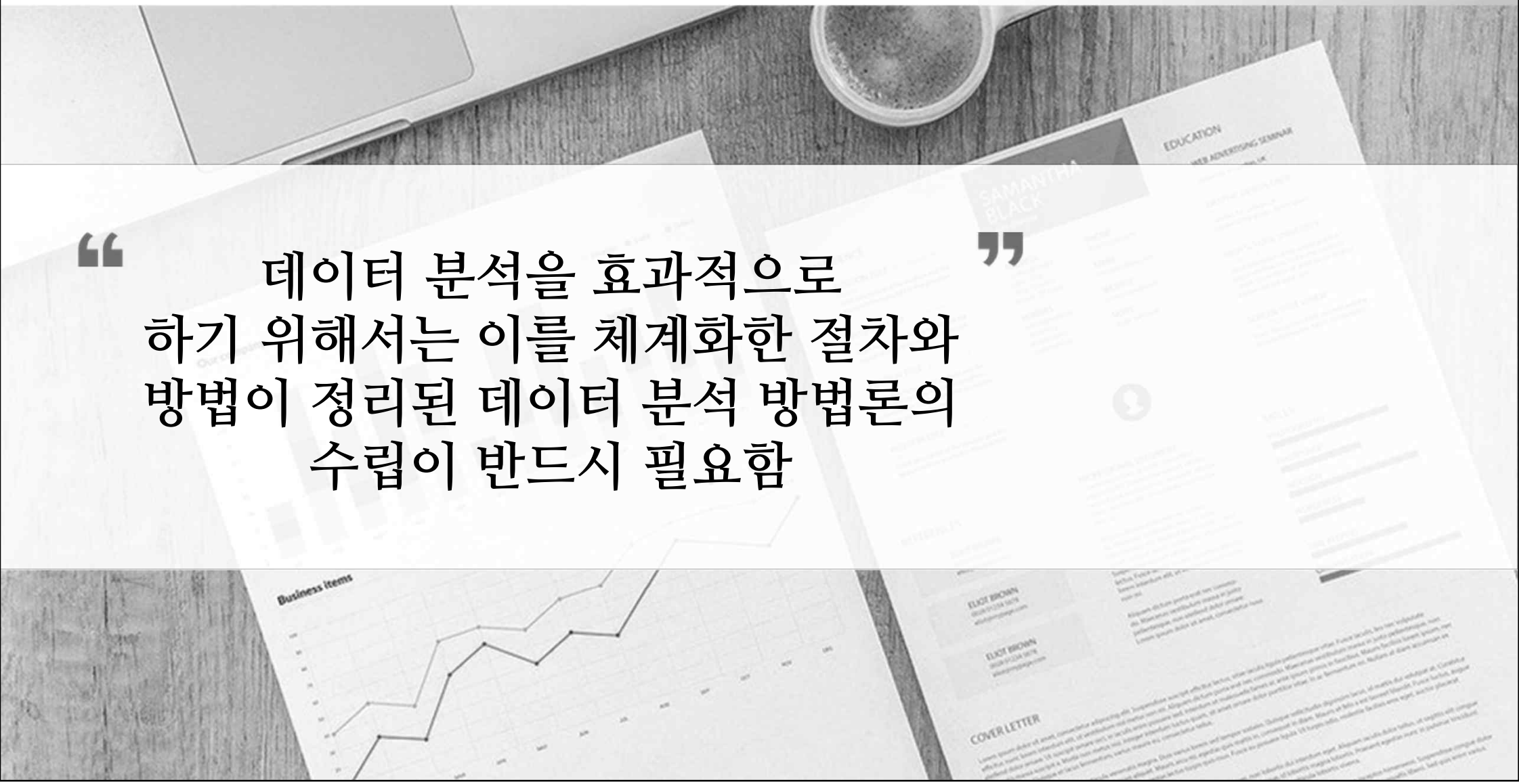
1) 분석 기획의 개요

● 분석 기획 시 고려 사항

분석과제 수행을 위한 장애요소

일회성 분석으로 그치지 않고 조직의 역량으로 내재화하기 위해서는 충분하고 계속적인 교육 및 활용방안 등의 변화 관리가 고려되어야 함

2) 분석 방법론의 종류



“ 데이터 분석을 효과적으로 하기 위해서는 이를 체계화한 절차와 방법이 정리된 데이터 분석 방법론의 수립이 반드시 필요함 ”

2) 분석 방법론의 종류

분석 방법론의 구성

상세한 절차 (Procedures)	방법 (Methods)
도구와 기법 (Tools & Techniques)	템플릿과 산출물 (Templates & Outputs)

2) 분석 방법론의 종류

KDD 분석방법론	• 1996년 Fayyad가 체계적으로 정리한 데이터 마이닝 프로세스 • 데이터에서 패턴을 찾는 과정을 9개의 프로세스로 제시
CRISP-DM 분석 방법론	• 1996년 유럽연합의 ESPRIT에서 있었던 프로젝트에서 시작 • 계층적 프로세스 모델로써 4개 레벨로 구성
빅데이터 분석 방법론	• 계층적 프로세스 모델로서 3계층, 5단계로 구성

3) KDD 분석 방법론

KDD 분석 방법론

KDD(Knowledge Discovery in Databases)는
1996년 Fayyad가 체계적으로 정리한 데이터 마이닝 프로세스

데이터베이스에서 의미 있는 지식을 탐색하는
데이터 마이닝부터 기계학습, 인공지능, 패턴인식,
데이터 시각화 등에서 응용될 수 있는 구조

3) KDD 분석 방법론

데이터셋 선택(Selection)

- 데이터셋 선택 전 분석 대상의 비즈니스 도메인에 대한 이해와 프로젝트 목표 설정이 필수
- 데이터베이스 또는 원시 데이터에서 분석에 필요한 데이터를 선택하는 단계

데이터 전처리(Preprocessing)

- 데이터셋에 포함된 잡음(Noise), 이상치(Outlier), 결측치(Missing Value)를 식별
- 필요시 제거하거나 의미 있는 데이터로 재처리하여 데이터 셋을 정제하는 단계

3) KDD 분석 방법론

데이터 변환(Transformation)

- 분석 목적에 맞게 변수를 생성, 선택하고 데이터의 차원을 축소하여 효율적으로 데이터 마이닝을 할 수 있도록 데이터에 변경하는 단계

데이터 마이닝(Data Mining)

- 분석 목적에 맞게 데이터 마이닝 기법을 선택하고 데이터 마이닝 알고리즘을 선택
- 데이터의 패턴을 찾거나 데이터를 분류 또는 예측 등의 마이닝 작업을 시행

3) KDD 분석 방법론

데이터 마이닝 결과 평가(Interpretation/Evaluation)

- 데이터 마이닝 결과에 대한 해석과 평가, 분석 목적과의 일치성을 확인
- 데이터 마이닝을 통하여 발견된 지식을 업무에 활용하기 위한 방안을 찾는 단계

4) CRISP-DM 분석 방법론

CRISP-DM*은 1996년 유럽연합의 ESPRIT에서 있었던 프로젝트에서 시작됨

CRISP-DM은 계층적 프로세스 모델로써 4개 레벨로 구성함

* Cross-Industry Standard Process for Data Mining

4) CRISP-DM 분석 방법론

업무 이해(Business understanding)

- 비즈니스 관점에서 프로젝트의 목적과 요구사항을 이해하기 위한 단계

데이터 이해(Data understanding)

- 분석을 위한 데이터를 수집하고 데이터 속성을 이해하기 위한 단계

데이터 준비(Data preparation)

- 분석을 위하여 수집된 데이터에서 분석기법에 적합한 데이터를 편성하는 단계

4) CRISP-DM 분석 방법론

모델링(Modeling)

- 다양한 모델링 기법과 알고리즘을 선택하고 모델링 과정에서 사용되는 파라미터를 최적화해 나가는 단계

평가(Evaluation)

- 모델링 결과가 프로젝트 목적에 부합하는지 평가하는 단계

전개(Deployment)

- 모델링과 평가 단계를 통하여 완성된 모델을 실업무에 적용하기 위한 계획을 수립하는 단계

5) 빅데이터 분석 방법론

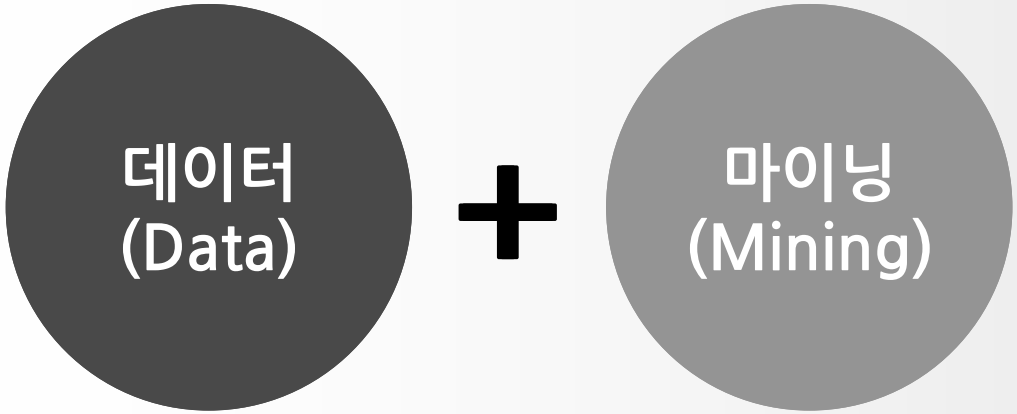
분석 기획 (Planning)	비즈니스 이해 및 범위 설정, 프로젝트 정의 및 계획 수립, 프로젝트 위험계획 수립
데이터 준비 (Preparing)	필요 데이터 정의, 데이터 스토어 설계, 데이터 수집 및 정합성 점검
데이터 분석 (Analyzing)	분석용 데이터 준비, 텍스트 분석, 탐색적 분석, 모델링, 모델 평가 및 검증, 모델 적용 및 운영 방안 수립
시스템 구현 (Developing)	설계 및 구현, 시스템 테스트 및 운영
평가 및 전개 (Deploying)	모델 발전 계획 수립, 프로젝트 평가 및 보고

03

데이터 마이닝의 개요

1) 데이터 마이닝이란?
2) 분류 분석
3) 군집 분석
4) 연관 분석

1) 데이터 마이닝이란?



‘데이터를 채굴한다’라는 의미를 가짐

1) 데이터 마이닝이란?

데이터 마이닝

대용량의 데이터로부터 데이터에 감춰진 관계, 규칙, 패턴 등을 탐색하고 모형화하여 유용한 지식을 추출하는 방법

→ 기업이 보유하고 있는 고객 데이터, 상품 데이터, 거래 데이터 등을 기반으로 데이터 내에 존재하는 지식, 경향, 규칙 등을 발견하여 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하고자 하는 작업

1) 데이터 마이닝이란?

● 데이터 마이닝의 분석 방법

지도 학습 (Supervised Learning)

정답이 주어진 상태에서 학습을 시키는 방법

- 의사 결정 나무
- 인공 신경망
- 회귀 분석
- 로지스틱 회귀분석

비지도 학습 (Unsupervised Learning)

정답이 주어지지 않은 상태에서 학습을 시키는 방법

- 군집 분석
- 연관성 분석

2) 분류 분석

분류 분석

데이터가 어떤 그룹에 속하는지 예측하는데 사용되는 기법

분류

반응 변수가
범주형인 경우

예 카드회사에서 회원들의 가입 정보를 통해 1년 후 신용등급을 알아맞히는 것

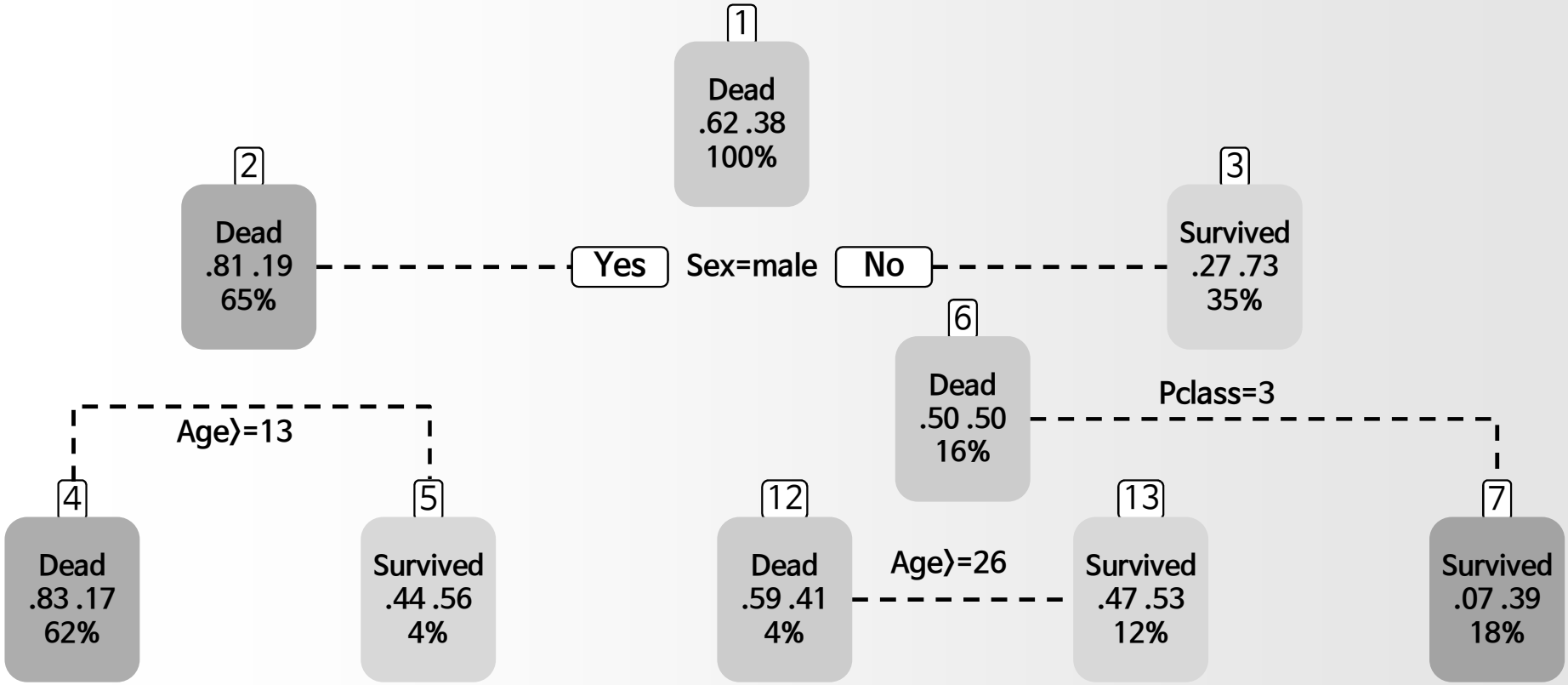
예측

반응 변수가
연속형인 경우

예 카드회사 회원들의 가입정보를 통해 연매출액을 알아맞히는 것

2) 분류 분석

- 의사 결정 나무



3) 군집 분석

군집 분석

각 개체의 유사성을 측정하여 유사한 성격을 가지는 몇 개의 군집으로 집단화하고, 군집들의 특성을 파악하여 군집들 사이의 관계를 분석하는 방법

3) 군집 분석

계층적 군집

가장 유사한 개체를 묶어 나가면서 원하는 개수의 군집을 형성하는 방법

K-평균 군집

군집의 초기값(K개)을 정하고, 각 개체를 가까운 초기값에 할당한 후 각 군집의 평균을 계산하여 초기값을 갱신하고 군집에 할당하는 과정을 반복하는 방법

3) 군집 분석

계층적 군집

K-평균 군집

가장 유사한 개체를 묶어 나가면서 원하는 개수의 군집을 형성하는 방법

계층적 군집의 결과는
덴드로그램(Dendrogram)의 형태로 표현

3) 군집 분석

계층적 군집

K-평균 군집

두 군집간의 거리를 측정하는 방법

최단연결법

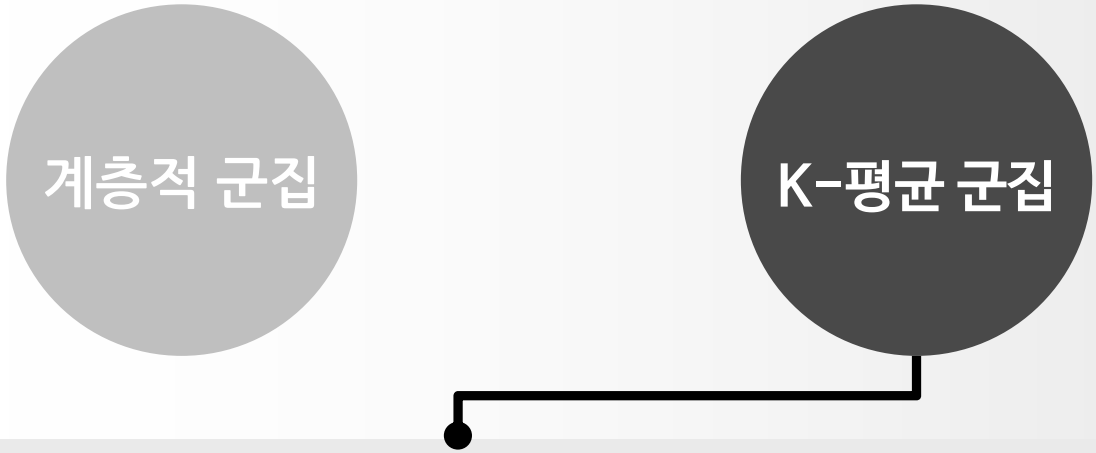
최장연결법

중심연결법

평균연결법

와드연결법

3) 군집 분석



주어진 데이터를 k개의 군집으로 묶는 방법

각 군집과의 거리 차이의 분산을 최소화하는 방식으로 동작함

3) 군집 분석

계층적 군집

K-평균 군집

초기 군집 중심으로 k개의 객체를 임의로 선택

각 개체를 가장 가까운 군집 중심에 할당

각 군집 내의 자료들의 평균을 계산하여 군집의 중심 갱신

군집 중심의 변화가 거의 없을 때까지 단계 반복

4) 연관 분석

연관 분석

'조건-결과' 식으로 표현되는 연관 규칙을 발견해내는 것

장바구니 분석
(Market Basket
Analysis)

4) 연관 분석

가만있자.
일단 기저귀부터 사고,
맥주 코너가
어디더라?

예

마트에서 기저귀를 사는 고객은 맥주를 동시에 구매한다는 연관 규칙을 알아냈다면 기저귀와 맥주를 인접한 진열대에 위치해 놓아 매출 증대가 일어남

4) 연관 분석

- 연관 분석의 측정 지표

지지도 (Support)	신뢰도 (Confidence)	향상도 (Lift)

전체 거래 중에서 품목 A, B가 동시에 포함되는 거래의 비율

$$P(A \cap B) = \frac{A\text{와 } B\text{가 동시에 포함된 거래 수}}{\text{전체 거래 수}}$$

4) 연관 분석

- 연관 분석의 측정 지표

지지도 (Support)	신뢰도 (Confidence)	향상도 (Lift)

품목 A가 포함된 거래 중에서 품목 A, B를 동시에 포함하는 거래일 확률

$$\frac{P(A \cap B)}{P(A)} = \frac{A\text{와 } B\text{가 동시에 포함된 거래 수}}{A\text{를 포함하는 거래 수}}$$

4) 연관 분석

- 연관 분석의 측정 지표

지지도 (Support)	신뢰도 (Confidence)	향상도 (Lift)

품목 B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률

$$\frac{P(A \cap B)}{P(A)P(B)} = \frac{A\text{와 } B\text{를 포함하는 거래 수}}{A\text{를 포함하는 거래 수} \times B\text{를 포함하는 거래 수}}$$

4) 연관 분석

'옥수수차 → 둥굴레차'의 측정지표 구해보기

항목	거래 수
옥수수차	100
둥굴레차	100
율무차	50
{옥수수차, 둥굴레차}	500
{옥수수차, 율무차}	300
{둥굴레차, 율무차}	200
{옥수수차, 둥굴레차, 율무차}	100

지지도

$$P(A \cap B)$$

신뢰도

$$\frac{P(A \cap B)}{P(A)}$$

향상도

$$\frac{P(A \cap B)}{P(A)P(B)}$$

학습 평가

Q1 Q2

Q1

분석 주제 유형에서 문제를 잘 알고 있으면서 기존에 수행하고 있는 방법이 존재하는 경우에 해당하는 유형은?

1. Optimization
2. Solution
3. Discovery
4. Insight

학습 평가

Q1 Q2

Q1

분석 주제 유형에서 문제를 잘 알고 있으면서 기존에 수행하고 있는 방법이 존재하는 경우에 해당하는 유형은?

1. Optimization
2. Solution
3. Discovery
4. Insight

정답 1번

해설 문제를 알고 기존에 수행하고 있는 방법이 존재하는 경우에는 Optimization에 해당합니다.

학습 평가

Q1 Q2

Q2 연관성 분석의 규칙이 얼마나 유의미한지 평가하기 위한 측정 지표가 아닌 것은?

1. 지지도
2. 신뢰도
3. 오분류율
4. 향상도

학습 평가

Q1 Q2

Q2 연관성 분석의 규칙이 얼마나 유의미한지 평가하기 위한 측정 지표가 아닌 것은?

1. 지지도
2. 신뢰도
3. 오분류율 ✓
4. 향상도

정답 3번

해설 연관성 분석의 측정 지표는 지지도, 신뢰도, 향상도입니다.

정리
하기

빅데이터 분석 기획

✓ 빅데이터 분석 기획이란?

- 실제 분석을 수행하기에 앞서 분석을 수행할 과제를 정의하고, 의도했던 결과를 도출할 수 있도록 이를 적절하게 관리할 수 있는 방안을 사전에 계획하는 일련의 작업

✓ 분석 방법론

- 분석 방법론은 상세한 절차(Procedures), 방법(Methods), 도구와 기법(Tools&Techniques), 템플릿과 산출물(Templates&Outputs)로 구성
- KDD 분석 방법론, CRISP-DM 분석 방법론, 빅데이터 분석 방법론 등이 사용됨

정리
하기

데이터 마이닝

- ✓ 데이터 마이닝의 정의
  - 대용량의 데이터로부터 데이터에 감춰진 관계, 규칙, 패턴 등을 탐색하고 모형화하여 유용한 지식을 추출하는 방법

정리
하기

데이터 마이닝

✓ 분석 방법

① 분류 분석

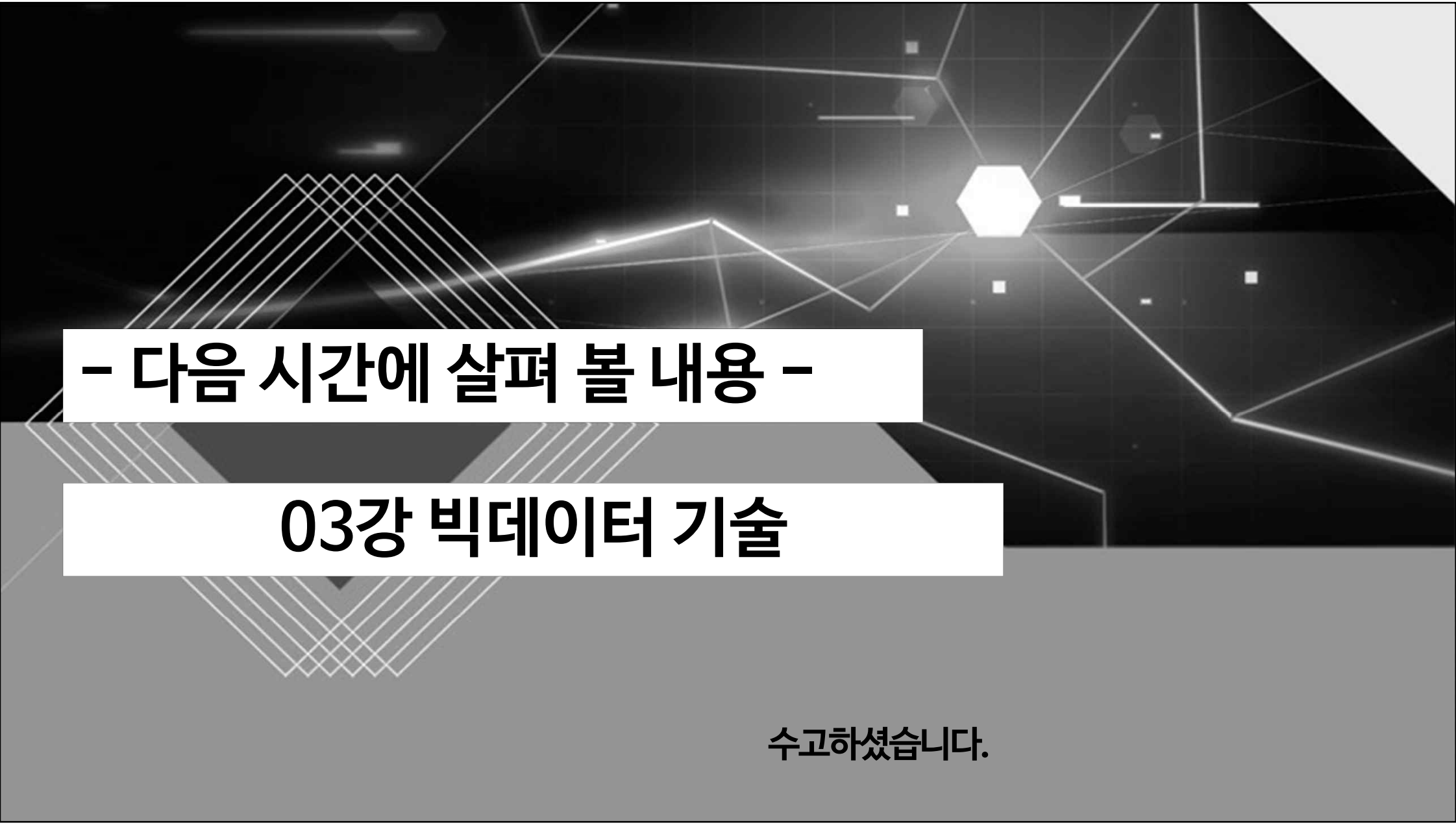
- 데이터가 어떤 그룹에 속하는지 예측하는데 사용되는 기법

② 군집 분석

- 각 개체의 유사성을 측정하여 유사한 성격을 가지는 몇 개의 군집으로 집단화하고, 군집들의 특성을 파악하여 군집들 사이의 관계를 분석하는 방법

③ 연관 분석

- '조건-결과' 식으로 표현되는 연관 규칙을 발견해내는 것으로 장바구니 분석 (Market Basket Analysis)라고 함


- 다음 시간에 살펴 볼 내용 -
03강 빅데이터 기술
수고하셨습니다.