빅데이터분석 실습

2주 데이터 분석 기획의 이해

데이터 사이언스 전공
담당교수: 곽철완

강의 내용

- 분석 방법론
- 분석 과제 발굴

1. 분석 방법론

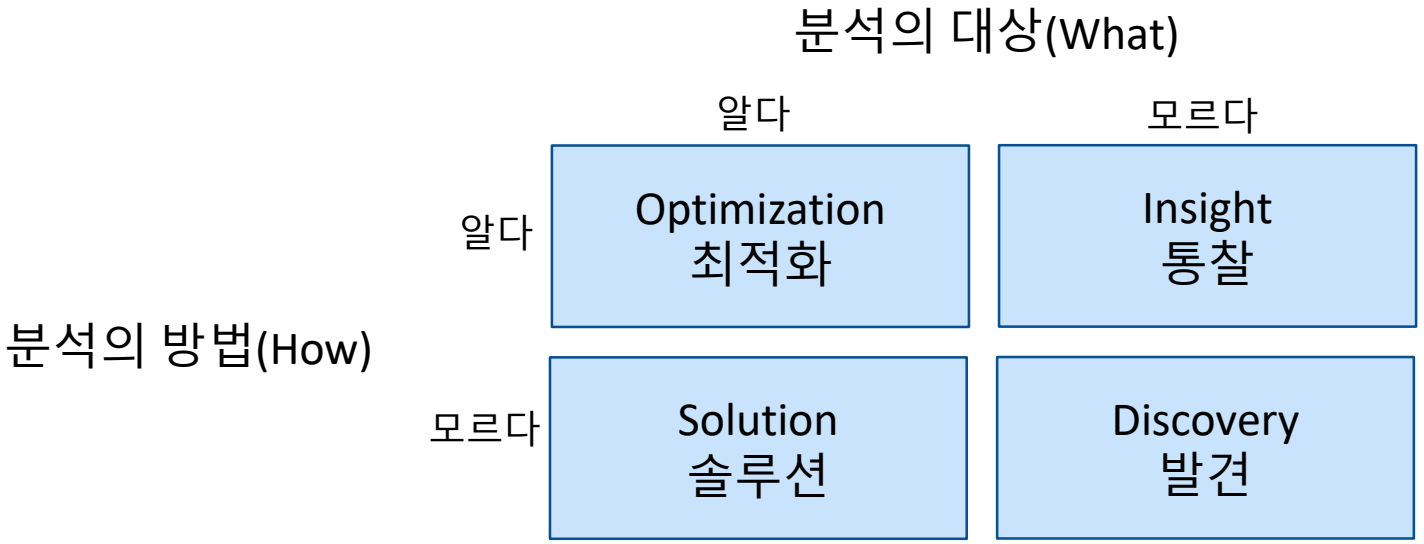
▪ 분석 기획이란?

- 분석에 앞서 과제 정의 및 결과 도출이 가능하도록 방안 계획
- 목표 + 데이터 + 방법

▪ 분석 기획을 위한 3대 역량

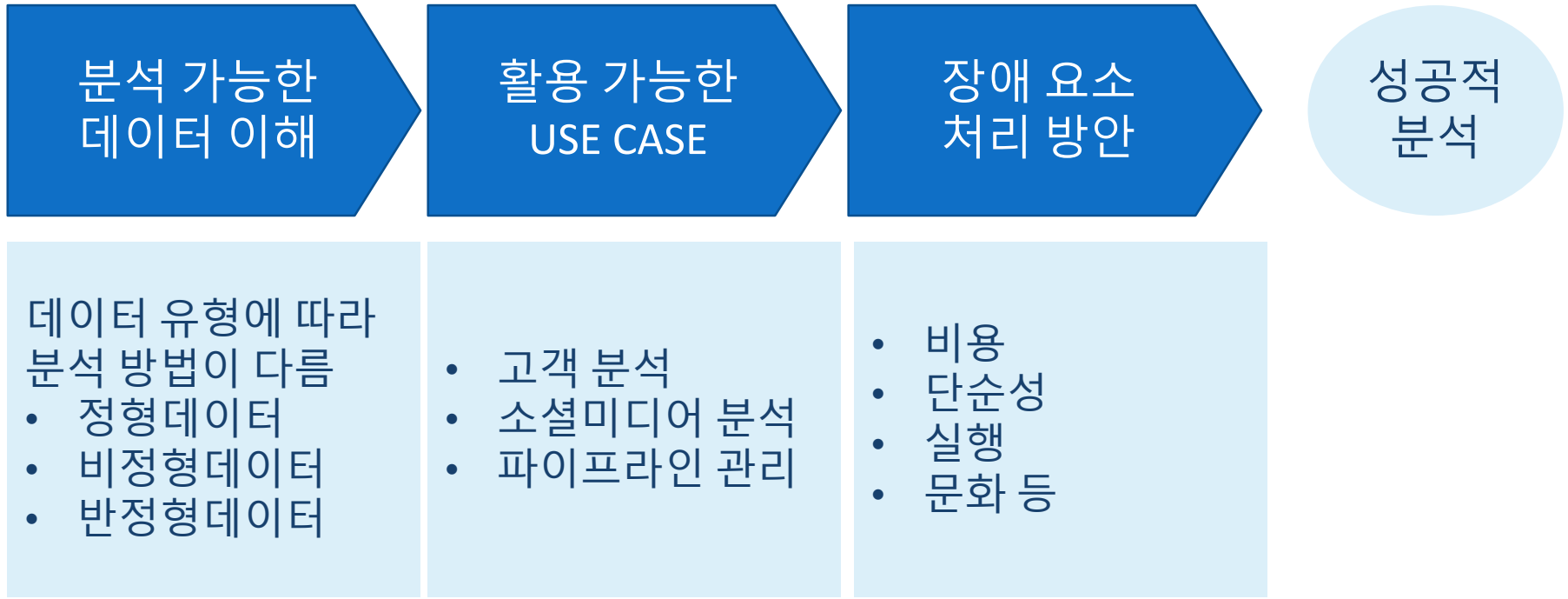
- Domain Knowledge(해당 분야에 대한 전문성 역량)
- Math & Statistics(분석 역량)
- Information Technology(분석 도구에 관한 기술 역량)

▪ 분석 주체 유형 분류



◦ 데이터 분석 방법은 이해하지만, 조직 내 분석 대상을 모르는 경우: 통찰

▪ 분석 기획 시 고려사항


분석 가능한 데이터 이해
활용 가능한 USE CASE
장애 요소 처리 방안
성공적 분석
데이터 유형에 따라 분석 방법이 다름
• 정형데이터
• 비정형데이터
• 반정형데이터
• 고객 분석
• 소셜미디어 분석
• 파이프라인 관리
• 비용
• 단순성
• 실행
• 문화 등

▪ 분석 방법론 틀

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

▪ 분석 방법론의 적용 업무 특성에 따른 모델

- 폭포수 모델: 단계를 순차적으로 진행
- 나선형 모델: 반복을 통해 점증적으로 개발
- 프로토타입 모델: 일부를 우선 개발하여 사용자에게 제공하여 문제점 보완

(1) KDD 분석 방법론

▪ 특징

- 데이터로부터 패턴이나 지식을 발견하기 위한 데이터 마이닝 프로세스
- 머신러닝, 인공지능 등에서 활용

▪ 9개 프로세스

1. 분석 대상 도메인의 이해
2. 분석 대상 데이터셋 선택과 생성
3. 데이터 정제 작업 혹은 전처리 작업
4. 목적에 맞는 피처(변수)를 찾거나 피처 수 축소
5. 목적에 적합한 데이터마이닝 기법 선택
6. 목적에 적합한 알고리즘(예, K-means clustering) 선택
7. 데이터마이닝 실행
8. 데이터마이닝 결과 해석
9. 데이터마이닝에서 발견된 지식 활용

▪ KDD 분석 절차

1) 데이터셋 선택 — Selection
2) 데이터 전처리(데이터 클리닝) — Preprocessing
3) 데이터 변환(학습용 데이터, 평가용 데이터) — Transformation
4) 데이터마이닝(기법 및 알고리즘 선택) — Data Mining
5) 데이터마이닝 결과 평가 — Interpretation/Evaluation

(2) CRISP-DM 분석 방법론

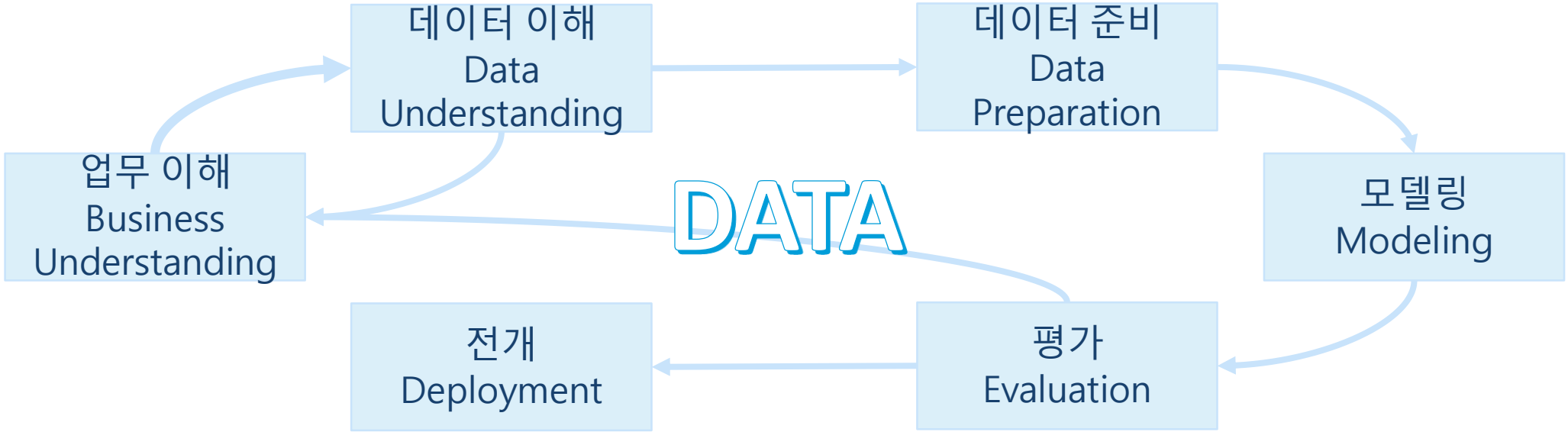
▪ 특징

- CRoss Industrial Standard Process for Data Mining
- 계층적 프로세스 모델로 4개 레벨로 구성

단계(Phases)
일반화 태스크(Generic Tasks)
세분화 태스크(Specialized Tasks): 예, 데이터 클리닝
프로세스 사례(Process instances): 예, 구체적인 실행

▪ CRISP-DM 프로세스

◦ 6개 단계로 구성되어 있고, 각 단계는 한 방향으로 구성되어 있지 않고 단계 간 피드백을 통하여 단계별 완성도를 높임



- 업무이해
  - 업무 목적 파악, 상황 파악, 데이터 마이닝 목표 설정, 프로젝트 계획 수립
- 데이터 이해
  - 초기 데이터 수집, 데이터 기술(설명) 분석, 데이터 탐색, 데이터 품질 확인
- 데이터 준비
  - 분석용 데이터셋 선택, 데이터 정제, 데이터 통합
- 모델링
  - 모델링 기법 선택, 모델의 과적합(overfitting) 문제 확인
- 평가
- 전개

(3) 빅데이터 분석 방법론

▪ 3 계층

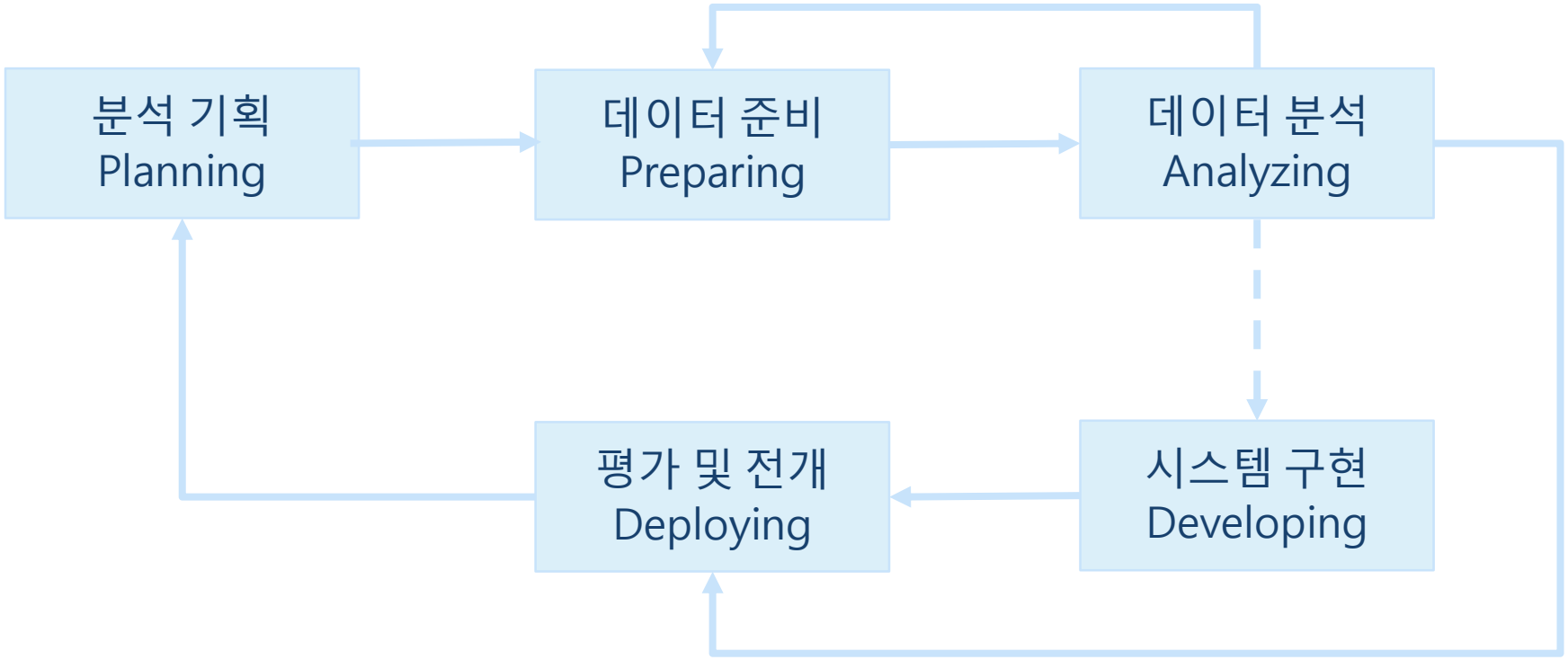
◦ 계층적 프로세스 모델

단계(Phases): 프로세스 그룹을 통해 산출물 생성

태스크(Task): 단계를 구성하는 단위 활동

스텝(Step): input, process & tool, output로 구성된 단위 프로세스

▪ 분석 방법론 5 단계 개요


분석 기획
Planning
데이터 준비
Preparing
데이터 분석
Analyzing
시스템 구현
Developing
평가 및 전개
Deploying

▪ 분석 기획(Planning)

1) 비즈니스 이해 및 범위 설정
   - 비즈니스 이해: 프로젝트 진행 방향 설정
   - 프로젝트 범위 설정: SOW(Statement of Work) 작성

2) 프로젝트 정의 및 계획 수립
   - 모델의 운영 이미지 설계와 모델 평가 기준 설정: WBS(Work Breakdown Structure) 작성

3) 프로젝트 위험 계획 수립

▪ 데이터 준비(Preparing)

1) 필요 데이터 정의
   - 데이터 속성, 데이터 소장처, 시스템 담당자 등

2) 데이터 스토어 설계(데이터 큐레이션 센터)
   - 정형 데이터 스토어, 비정형 데이터 스토어(하둡, NoSQL 등 이용)

3) 데이터 수집 및 정합성 점검
   - 메타데이터, 데이터 사전 작성

▪ 데이터 분석(Analyzing)

1) 분석용 데이터 준비
   - 프로젝트 목표 인식과 분석에 필요한 데이터 범위 확인
   - 분석용 데이터셋 준비

2) 텍스트 분석
   - a bag-of-words: 텍스트 분석의 일반적인 방법으로 출현 단어 수를 이용하여 분석
   - a set of sequences: 텍스트에서 의미를 추출하여 분석(자연어 처리)

3) 탐색적 분석
   - 통계분석, 연관성 분석
   - 데이터 시각화

4) 모델링
   - 머신러닝 기법 이용
   - 예측, 분류, 군집 등의 모델

5) 모델 평가 및 검증

▪ 시스템 구현(Developing)

1) 설계 및 구현
   - 모델링 결과를 시스템으로 구현

2) 시스템 테스트 및 운영
   - 시스템에 구현된 모델을 테스트하여 가동 중인 시스템에 적용

▪ 평가 및 전개(Deploying)

1) 모델 발전 계획 수립

◦ 모델의 생명 주기 설정과 주기적 평가 실시를 통하여 모델 유지 보수 및 재구축 방안 마련

2) 프로젝트 평가 및 보고

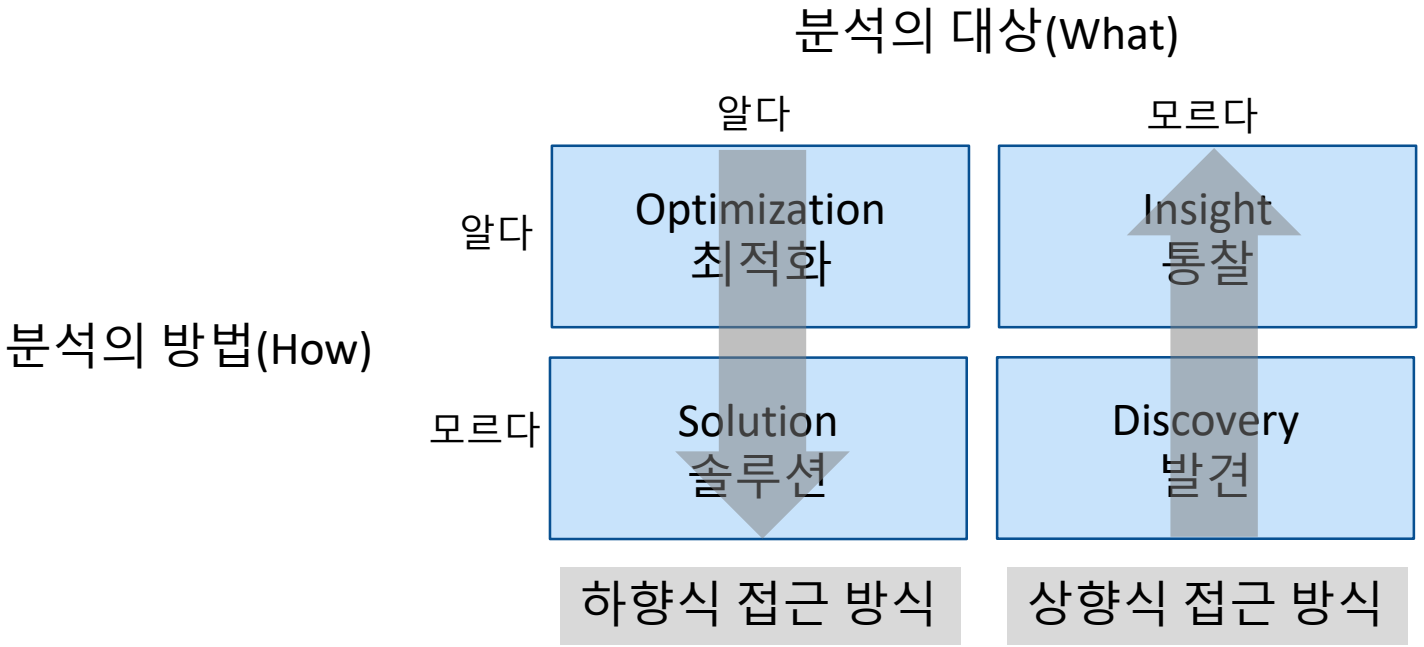
◦ 성과를 정량적, 정성적으로 평가하고, 프로젝트 진행과정에서 산출된 결과물을 지식 자산화

◦ 최종 보고서 작성 및 보고

2. 분석 과제 발굴

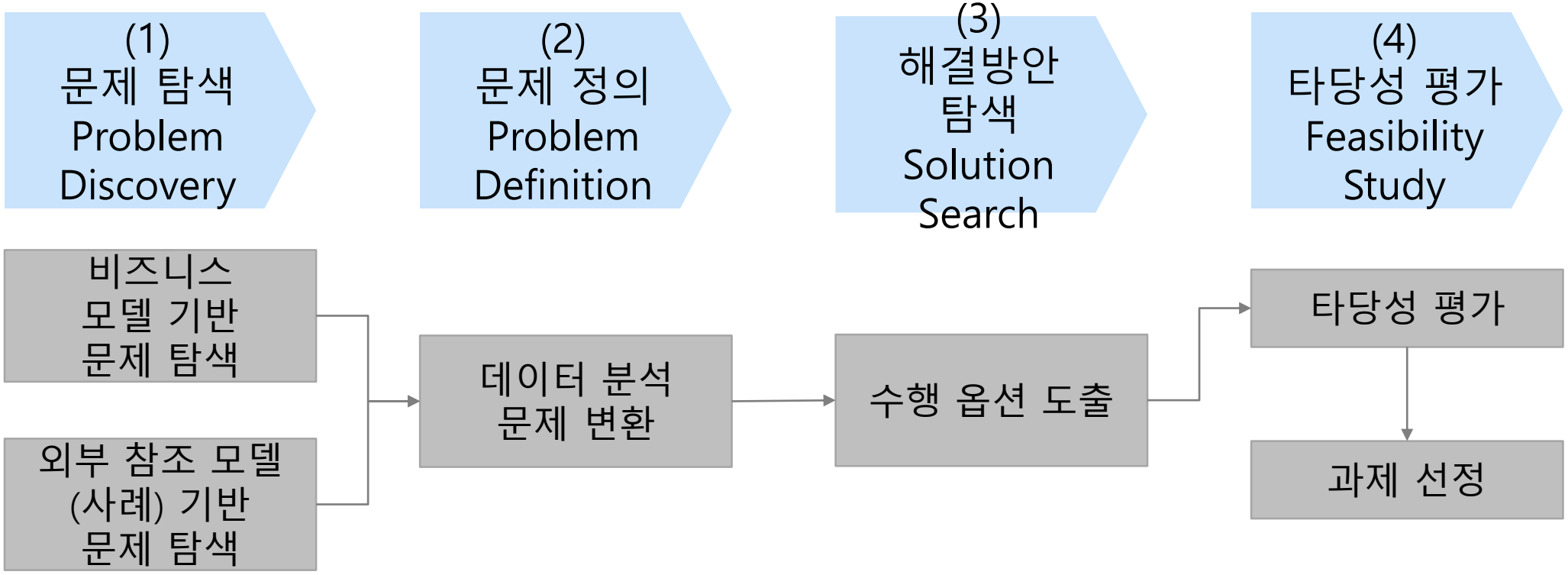
▪ 과제 발굴 접근 방식

- 하향식 접근 방식(Top Down Approach)
- 상향식 접근 방식(Bottom Up Approach): 문제를 정의하고 해결방안 탐색



(1) 하향식 접근법

▪ 단계 개요



1. 문제 탐색 단계

1) 비즈니스 모델 기반 문제 탐색
   - 비즈니스 모델 캔버스를 기반으로 주제 도출
     - 업무 Operation : 내부 프로세스 관련 주제
     - 제품 Production : 제품・서비스 개선 관련 주제
     - 고객 Customer : 고객, 서비스 채널 관점의 주제
     - 규제와 감사 Regulation & Audit : 제품 생산 및 전달 과정의 규제 관점의 주제
     - 지원 인프라 IT & Human Resource : 시스템 영역 및 운영 인력 관점의 주제

① 거시적 관점의 메가트렌드 STEEP

- 사회 영역 Social
  - 노령화, 밀레니엄 세대의 등장, 저출산에 따른 변화

- 기술 영역 Technological
  - 나노 기술, IT 융합 기술, 인공지능, 로봇기술

- 경제 영역 Economic
  - 환율 및 금리 변동, 무역 전쟁

- 환경 영역 Environmental
  - 미세먼지, 탄소배출 규제

- 정치 영역 Political
  - 국내외 정치 변화

② 경쟁자 확대 관점

- 대체재 영역
  - 오프라인 제공 제품이 온라인으로 제공

- 경쟁자 영역
  - 제품 서비스의 주요 경쟁자 동향 파악

- 신규 진입자 영역
  - 향후 시장에 대해 파괴적인 역할을 수행할 신규 진입자 파악

③ 시장 요구(needs) 탐색 관점

- 고객 영역
  - 고객의 구매 동향 및 고객의 요구 사항 파악
- 채널 영역
  - 인터넷 전문 은행 탄생에 따른 변화
- 영향자들 영역
  - M&A 시장 확대에 따른 유사 업종의 신규 기업 인수 기회 탐색

④ 역량의 재해석 관점

- 내부 역량 영역
  - 자사 소유 부동산을 활용한 부가 가치 창출 기회 발굴

- 파트너와 네트워크 영역
  - 수출입, 통관 노하우를 활용한 추가 사업기회 탐색

2) 외부 참조 모델(사례) 기반 문제 탐색

- 분야별 분석 대상 기업 혹은 서비스별로 집단(Pool)을 만들고 여기에서 아이디어를 찾아
- 브레인 스토밍(brain storming)을 통해 주제를 도출하는 방법
- 사전에 많은 사례를 수집해 놓아 후일 활용

3) 분석 Use Case 정의

- 해결해야 할 문제에 대한 상세한 설명과
- 해당 문제를 해결했을 때 효과에 대해 기술
- 향후 이를 통해 데이터 분석 주제를 정하고, 주제의 적합성 평가에 활용

2. 문제 정의 단계

- 식별된 비즈니스 문제를 데이터 분석 문제로 변환하여 정의하는 단계
  - 데이터의 문제란 비즈니스 문제를 해결하기 위해 필요한 데이터 및 기법(how)을 정의하는 것
  - (예) 비즈니스 문제: 고객 이탈

  → 데이터 분석 문제: 고객 이탈에 영향을 미치는 요인을 식별하고, 이탈 가능성을 예측

- 데이터 분석 문제의 정의 및 요구사항: 최종사용자(end user)에서 이루어져야 함

3. 해결방안 탐색 단계

- 데이터 분석 문제 해결을 위한 다양한 방안 모색
  - 기존 시스템의 단순한 보완으로 분석이 가능한지 고려
  - 빅데이터 분석 도구를 통해 체계적이고 심도 있는 방안 고려

4. 타당성 평가 단계

- 경제적 타당성
  - 비용 대비 편익 분석의 관점에서 접근
  - 비용 항목: 데이터, 시스템, 인력, 유지 보수
  - 편익 항목: 실질적인 비용 절감, 추가적 매출과 수익

- 데이터 및 기술적 타당성
  - 데이터 존재 여부, 분석 시스템 환경, 분석 역량
  - 도출된 대안 중 가장 우월한 대안 선택

요약

- 분석방법론
- 분석과제발굴
  - 하향식 접근법

다음 시간

- 분석과제 발굴
  - 상향식 접근법
- 데이터 분석 마스터 플랜