빅데이터분석 실습

3주 데이터 분석 마스터 플랜

데이터 사이언스 전공
담당교수: 곽철완

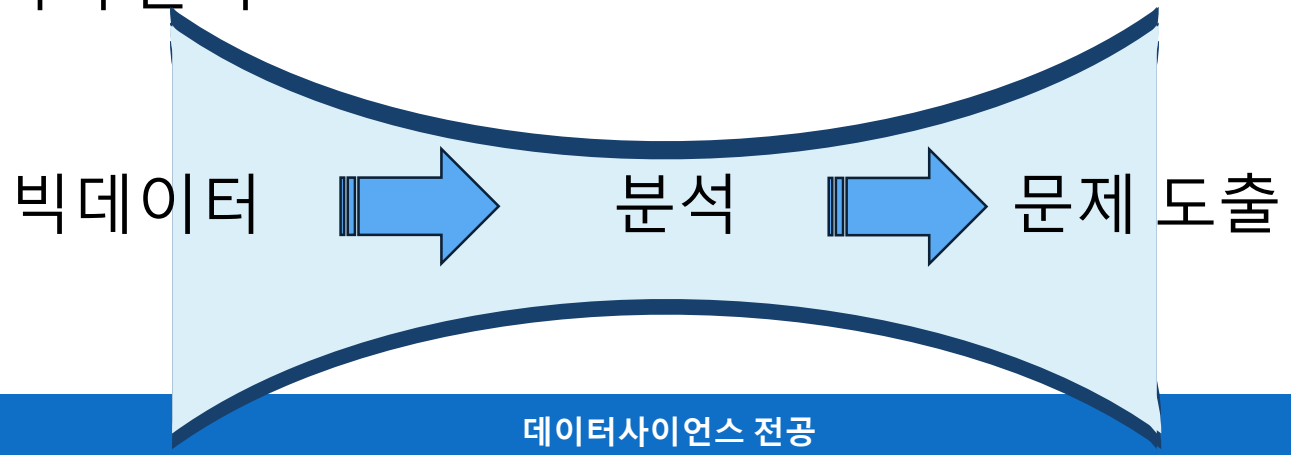
강의 내용

- 상향식 접근법
- 마스터 플랜 수립 프레임 워크
- 분석 거버넌스 체계 수립
- 데이터 분석 기법 개요

(2) 상향식 접근법

▪ 특징

- 다양한 원천 데이터를 대상으로 분석하여 가치 있는 문제 도출
- (예) 제약 회사에서 새로운 의약품 개발
  - 특허기간이 만료된 의약품 약 2천 종류의 데이터를 분석하여, 상호 결합을 통한 새로운 의약품 개발
  - 두 개의 조합이 백만개 이상이기 때문에 효과성 검증을 위해 다양한 기법 적용을 통한 데이터 분석



1. 하향식 접근법의 한계를 극복하기 위한 분석 방법론

- 하향식 접근법의 문제
  - 논리적인 단계별 접근법은 문제의 구조가 분명하고, 해결책 도출을 위해 데이터 분석가 및 의사결정권자가 책임을 가지고 있어, 해결책 도출은 유리
  - 하지만, 새로운 문제 탐색에는 한계가 있어서, 최근 복잡하고 다양한 환경에서 발생하는 문제에는 적합하지 않을 수 있음

- 디자인 사고(Design Thinking) 접근법을 통해 문제점 해결

- 사물을 분석적 관점에서 인식하는 것이 아니라, 있는 그대로 'What' 관점에서 접근

- 데이터 분석은 일반적으로 머신러닝의 비지도 학습(Unsupervised Learning) 적용
  - 데이터 특성 분석을 통하여 데이터로부터 정보를 파악함
  - 군집 분석, 주성분 분석, 장바구니 분석 등

- 지도 학습(Supervised Learning)
  - 사전에 분석한 데이터를 기반으로 새로운 데이터에 대한 예측, 분류 실시

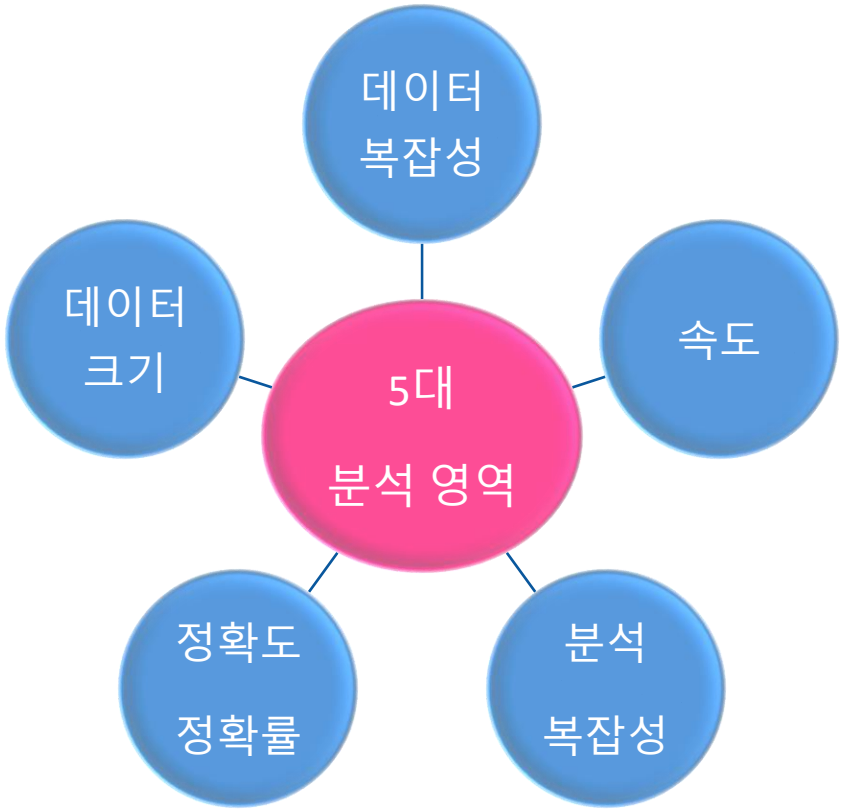
2. 시행착오를 통한 문제 해결

- 프로토타이핑(prototyping) 접근법
  - 먼저 분석을 시도하여 프로토타입 결과를 산출 한 후, 반복적으로 개선하는 방법
  - 프로토타입이 완전하지 못하지만, 신속하게 해결책을 제시한 후, 이를 바탕으로 문제를 좀 더 명확하게 인식하고 필요한 데이터를 식별하여 구체화하는 접근 방식
  - 가설의 생성 → 디자인에 대한 실험 → 실제 환경에서 테스트 → 통찰 도출 및 가설 확인

- 빅데이터 분석 환경에서 프로토타이핑의 필요성
  - 문제의 인식 수준
    - 불명확하거나 이전에 보지 못한 문제에 대해 프로토타입을 통해 문제 이해
  - 필요 데이터 존재 여부의 불확실성
    - 만약 데이터가 존재하지 않는다면, 불가능한 프로젝트 수행의 위험을 사전에 방지
  - 데이터 사용 목적의 가변성
    - 데이터 가치는 가변적이기 때문에 기존 데이터 정의를 재검토하여 데이터의 사용 목적과 범위를 확대

3. 분석 프로젝트(과제) 관리 방안

- 분석 과제 관리를 위한 5대 영역



- Data Size
  - 데이터 양을 고려한 관리 방안

- Data Complexity
  - 비정형 데이터에 대한 분석 모델 선정

- Speed
  - 결과가 도출 되었을 때, 이를 활용하는 측면에서 속도 고려(실시간 수행)

- Analytic Complexity
  - 복잡도와 정확도는 trade off 관계. 해석이 가능하면서 정확도를 올릴 수 있는 모델 파악

- Accuracy & Precision
  - 정확도: 모델값과 실제값 차이
  - 정확률: 모델을 지속적으로 반복했을 때 편차 수준의 일관성

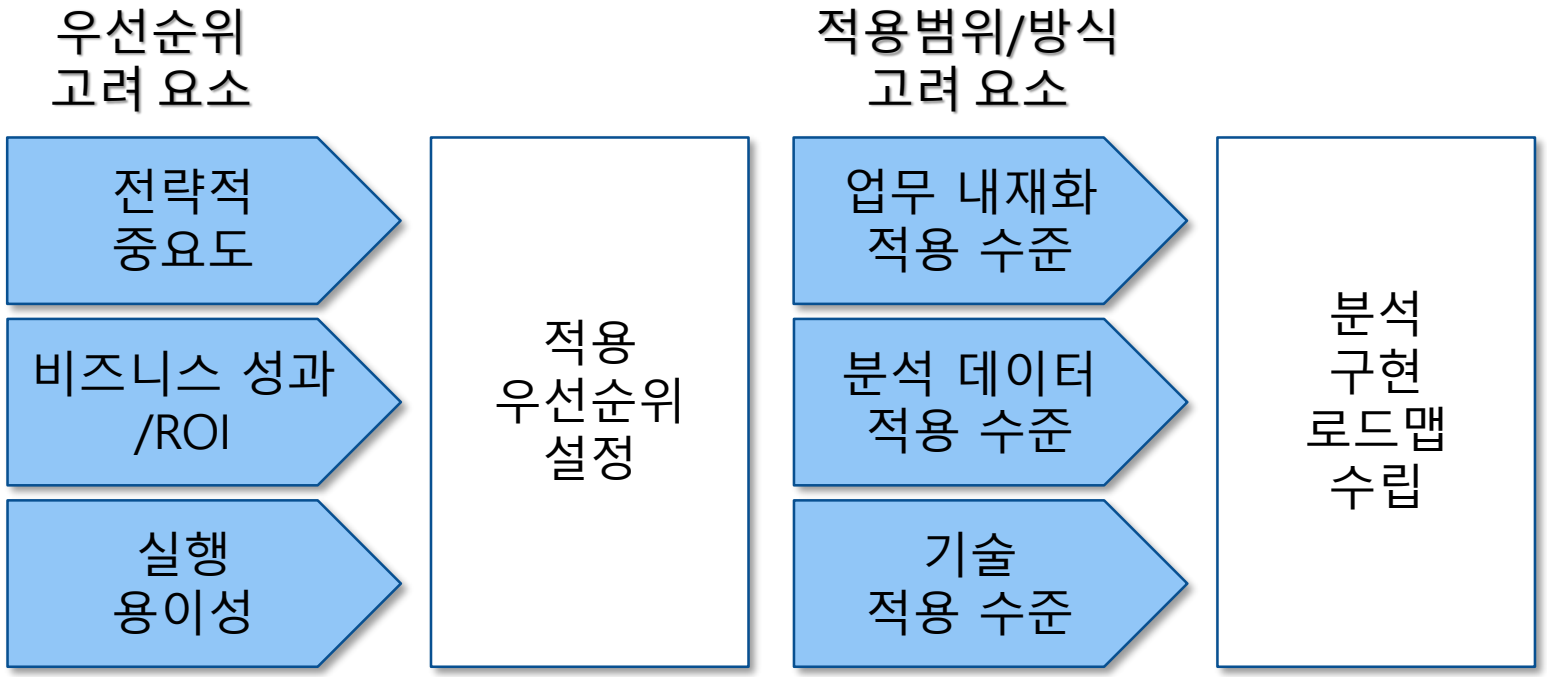
▪ 분석 과제 관리 방안

- 분석 과제 관리 주제
  - 통합(integration), 이해관계자(stakeholder), 범위(scope), 자원(resource), 시간(time), 원가(cost), 리스크(risk), 품질(quality), 조달(procurement), 의사소통(communication)

1. 마스터 플랜 수립 framework

▪ 개요

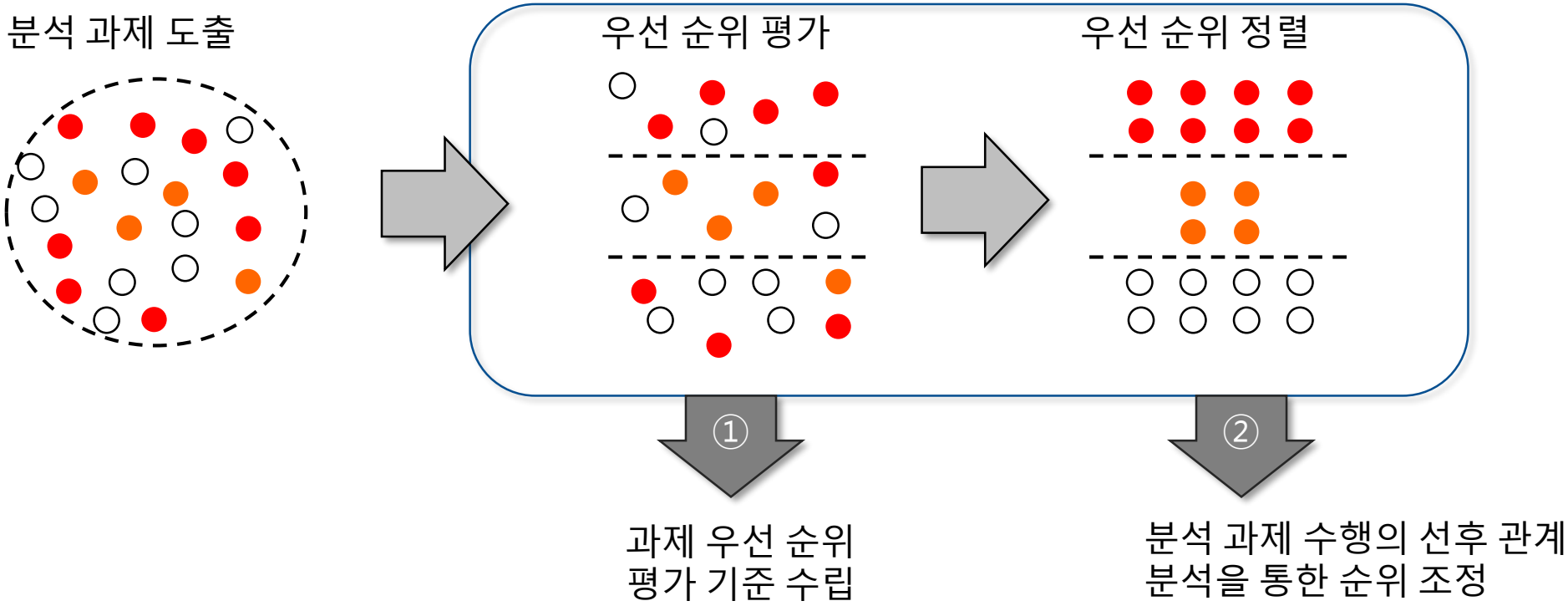
- 데이터 기반 구축을 위해 적용 우선 순위 설정
- 데이터 분석 구현을 위한 로드맵 수립



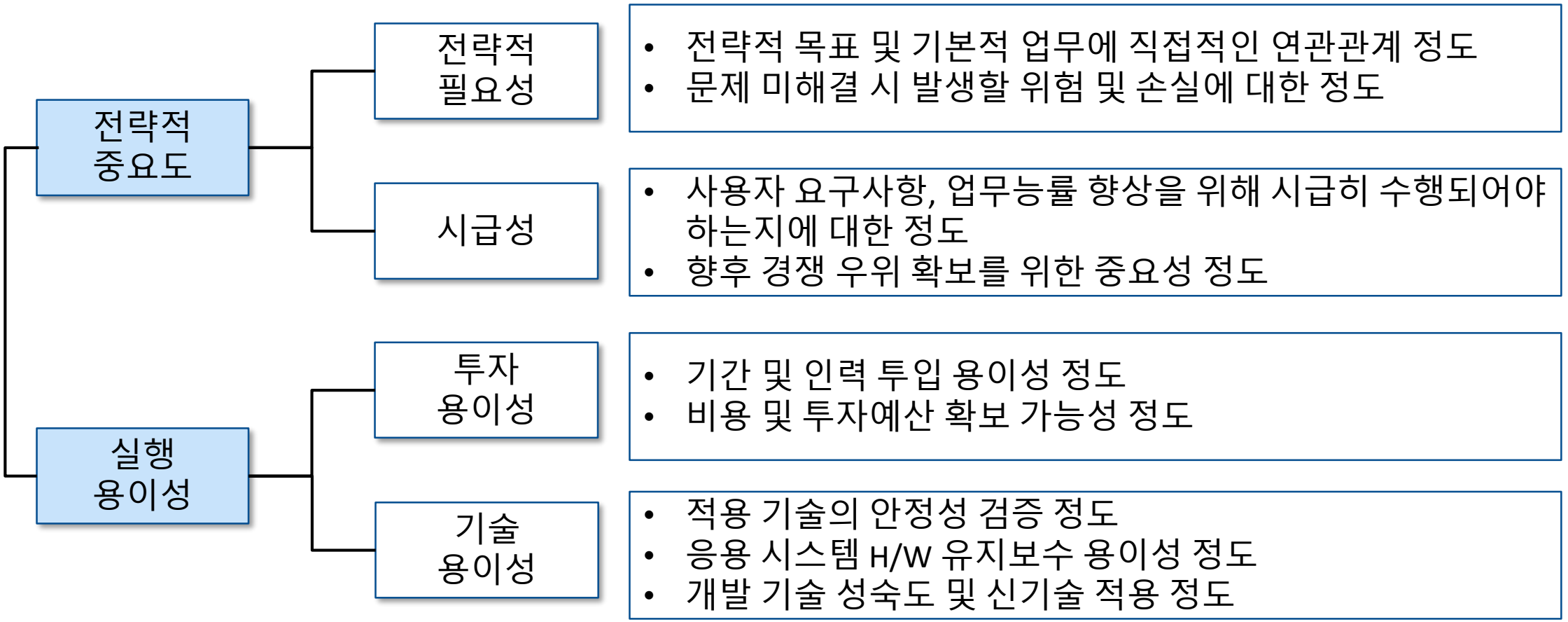
- ISP(Information Strategy Planning)
  - 정보기술 또는 정보시스템을 전략적으로 활용하기 위해
  - 조직 내·외부 환경을 분석하여 기회나 문제점을 도출하고
  - 사용자 요구사항을 분석하여
  - 시스템 구축 우선 순위를 결정하는 중장기 마스터 플랜을 수립하는 절차
- 분석 마스터 플랜
  - 일반적인 ISP 방법론을 활용하되
  - 데이터 분석 기획의 특성을 고려하여 수행하고
  - 기업에 필요한 데이터 분석 과제를 도출한 후
  - 과제의 우선 순위를 결정하고 중·장기 계획 수립

■ 우선순위 평가 방법 및 절차

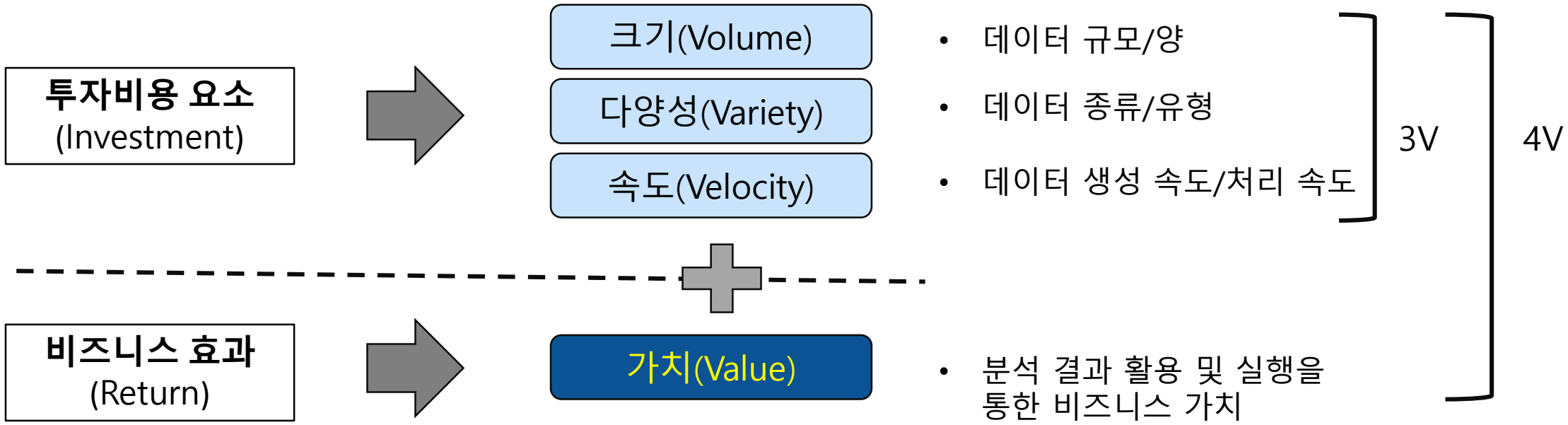
◦ 정의된 데이터 과제의 실행 순서 결정 과정



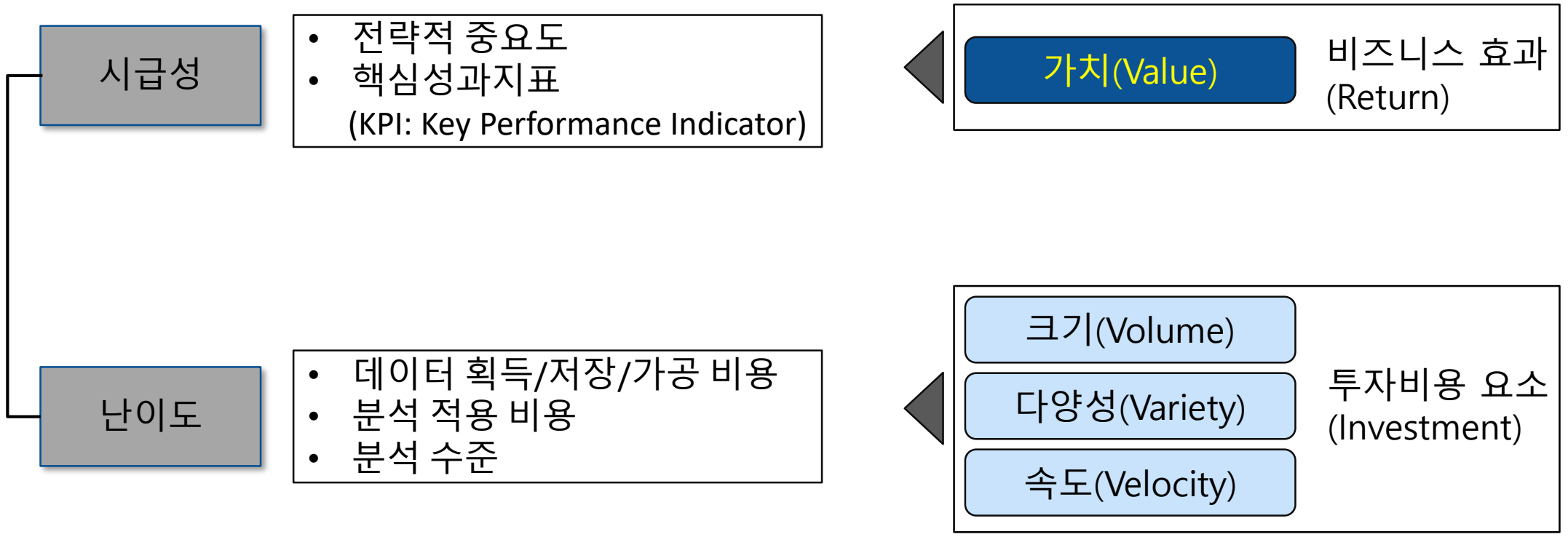
- 우선순위 평가 기준 예시


전략적 중요도
실행 용이성
전략적 필요성
시급성
투자 용이성
기술 용이성
• 전략적 목표 및 기본적 업무에 직접적인 연관관계 정도
• 문제 미해결 시 발생할 위험 및 손실에 대한 정도
• 사용자 요구사항, 업무능률 향상을 위해 시급히 수행되어야 하는지에 대한 정도
• 향후 경쟁 우위 확보를 위한 중요성 정도
• 기간 및 인력 투입 용이성 정도
• 비용 및 투자예산 확보 가능성 정도
• 적용 기술의 안정성 검증 정도
• 응용 시스템 H/W 유지보수 용이성 정도
• 개발 기술 성숙도 및 신기술 적용 정도

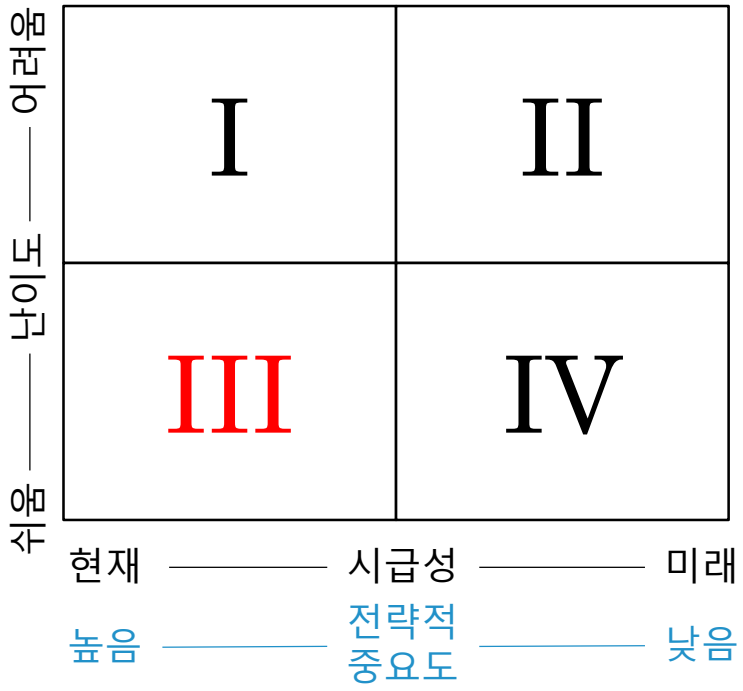
- ROI(Return on investment: 투자 수익율) 관점에서 빅데이터 핵심 특징



- 데이터 분석 과제의 우선순위 평가 기준

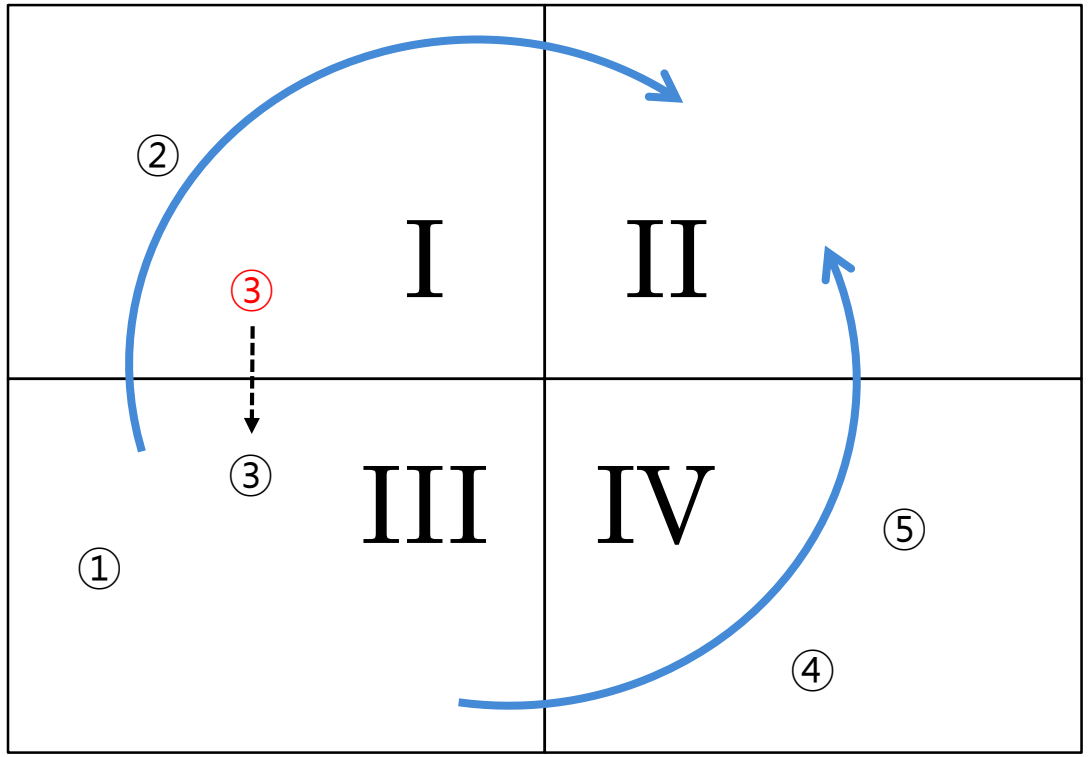


- 포트폴리오 사분면 분석을 통해 과제 우선순위 선정



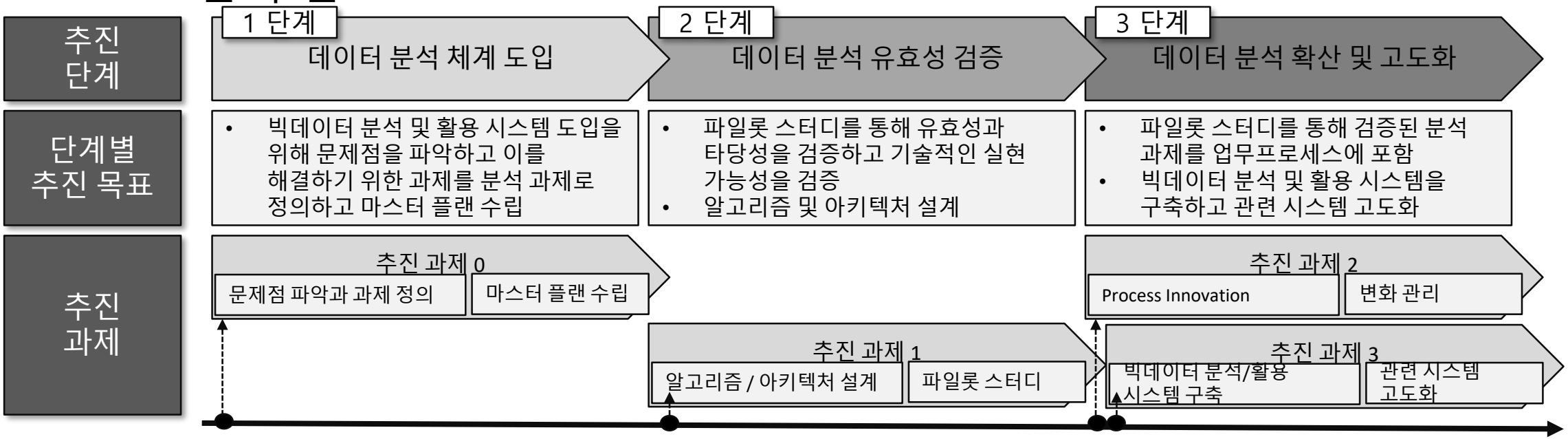
I	• 전략적으로 중요도가 높아 경영에 미치는 영향이 크므로 현재 시급하게 추진이 필요 • 난이도가 높아 현재 수준에서 과제를 곧바로 적용하기 어려움
II	• 현재 시점에서는 전략적 중요가 높지 않지만 중장기적 관점에서 반드시 추진되어야 함 • 분석 과제를 바로 적용하기에 난이도가 높음
III	• 전략적 중요도가 높아 현재 시점에 전략적 가치를 두고 있음 • 과제 추진의 난이도가 어렵지 않아 우선적으로 곧바로 적용 가능할 필요성이 있음
IV	• 전략적 중요도가 높지 않아 중장기적 관점에서 과제 추진이 바람직함 • 과제를 바로 적용하는 것이 어렵지 않음

- 분석 과제 난이도 조율을 통해 분석 적용 우선순위 조정
  - ③번 과제는 I 사분면에 위치해 있으나 데이터 양, 데이터 특성, 분석 범위 등의 난이도 조절을 통해 III 사분면으로 이동하여 우선순위를 조정할 수 있음

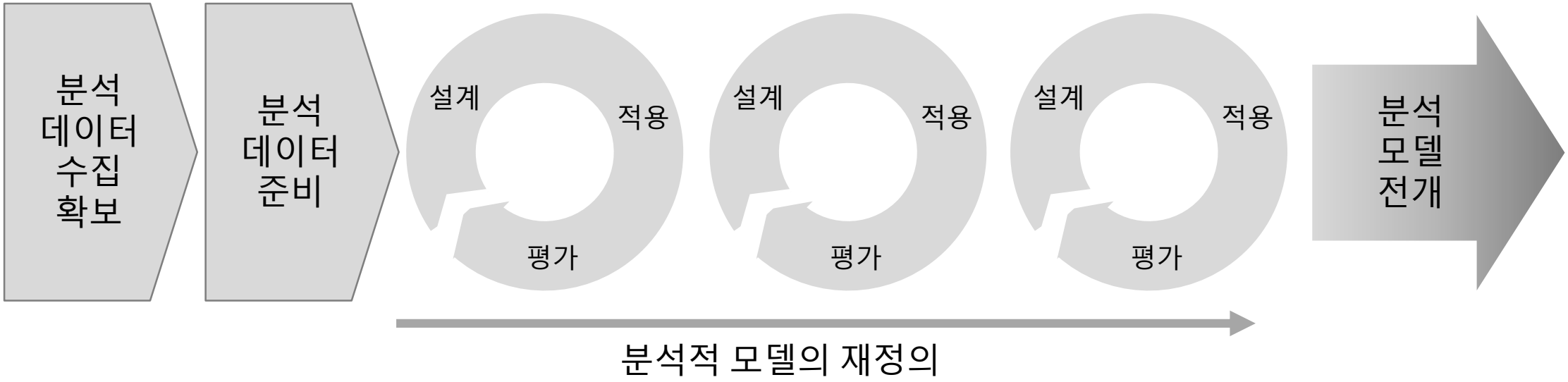


▪ 이행계획 수립

- 단계적 구현 로드맵 수립
  - 포트폴리오 사분면 분석을 통해 과제의 1차적 우선순위 결정
  - 과제별 적용범위 및 방식을 고려하여 최종적인 우선순위 결정 후 단계적 구현 로드맵 수립



- 세부 이행계획 수립
  - 반복적 분석 체계: 모델링 단계를 반복적으로 수행하는 혼합형을 많이 사용

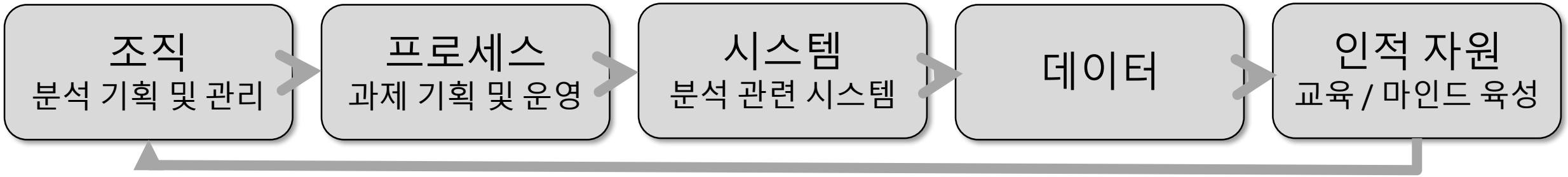


2. 분석 거버넌스 체계 수립

▪ 개요

- 필요성
  - 기업 운영에 데이터 중요성이 강조될 수록 데이터 분석·활용을 위한 체계적인 관리 필요성 증대
  - 어떤 데이터를, 어떤 목적으로, 어떻게 분석에 활용할 것인가 계획
  - 데이터 분석 문화 정착 및 데이터 분석 업무의 지속적인 고도화

- 구성 요소



▪ 데이터 분석 수준 진단

- 목적
  - 데이터 분석 도입 여부와 활용을 위해, 현재 데이터 분석 수준을 진단하여
  - 데이터 분석 유형 및 분석 방향을 결정
- 목표
  - 현재 데이터 분석 수준을 명확하게 이해하여, 미래 목표 수준 정의
  - 경쟁사와 비교하여 우리의 수준을 파악하여, 경쟁력 확보를 위해 선택 및 집중해야 할 영역 선정하고, 개선 방안 도출
- 데이터 분석 수준 진단 기준
  - 분석 준비도(readiness), 분석 성숙도(maturity)

- 분석 준비도

분석 업무 파악

- 발생한 사실 분석 업무
- 예측 분석 업무
- 시뮬레이션 분석 업무
- 최적화 분석 업무
- 분석 업무 정기적 개선

인력 및 조직

- 분석 전문가 담당 직원 존재
- 분석 전문가 교육 훈련 프로그램
- 관리자의 기본적인 분석 능력
- 분석 총괄업무 담당 조직 존재
- 경영진의 분석 업무 이해 능력

분석 기법

- 업무별 적합한 분석기법 사용
- 분석 업무 도입 방법론
- 분석기법 라이브러리
- 분석기법 효과성 평가
- 분석기법 정기적 개선

분석 데이터

- 분석에 필요한 데이터 양
- 데이터의 신뢰성
- 데이터의 적시성
- 비정형 데이터 관리
- 외부 데이터 활용 체계
- 핵심 데이터 관리(Master Data Management)

분석 문화

- 사실에 근거한 의사결정
- 관리자의 데이터 중시 풍토
- 회의 등에서 데이터 활용
- 경영진의 직관보다 데이터 활용
- 데이터 공유 및 협업 문화

IT 인프라

- 운영시스템 데이터 통합
- EAI(Enterprise Application Integration) 운영체계
- 분석 전용 서버 및 스토리지
- 빅데이터 분석 환경
- 데이터 시각화 환경

- 분석 성숙도
  - 평가 도구: CMMI(Capability Maturity Model Integration) 모델
  - 수준 분류

도입 단계	활용 단계	확산 단계	최적화 단계

  - 분석 성숙도 진단 분류

비즈니스 부문
조직역량 부문
IT 부문

◦ 도입 단계

설명	분석을 시작하여 환경과 시스템 구축
비즈니스 부문	• 실적 분석 및 통계 • 정기 보고 수행 • 운영 데이터 기반
조직역량 부문	• 임시 부서에서 수행 • 담당자 역량에 의존
IT 부문	• 데이터 웨어하우스 • 데이터 마트 • ETL(Extract Transform Load)/EAI • OLAP(Online Analytical Processing) 온라인 분석처리

◦ 활용 단계

설명	분석 결과를 실제 업무에 적용
비즈니스 부문	• 미래 결과 예측 • 시뮬레이션 • 운영 데이터 기반
조직역량 부문	• 전문 담당부서에서 수행 • 분석 기법 도입 • 관리자가 분석 수행
IT 부문	• 실시간 대쉬보드 • 통계분석 환경

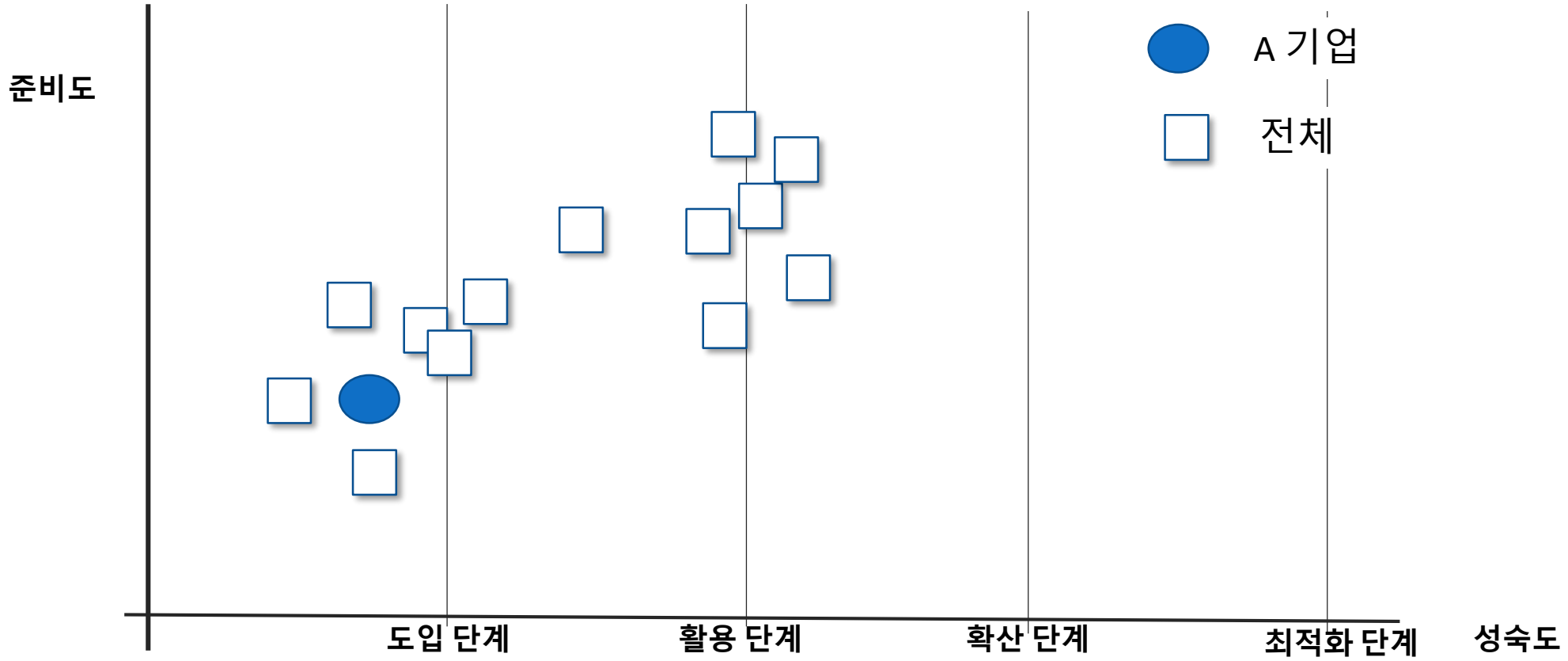
◦ 확산 단계

설명	기업 전체 차원에서 데이터 분석을 관리하고 공유
비즈니스 부문	• 기업 전체 성과 실시간 분석 • 분석 규칙 관리 • 이벤트 관리
조직역량 부문	• 기업 전체 부서 수행 • 분석 COE(Center of Excellence 전문가 조직) 운영 • 데이터 사이언티스트 확보
IT 부문	• 빅데이터 관리 환경 • 시뮬레이션·최적화 • 데이터 시각화 분석 • 분석 전용 서버

◦ 최적화 단계

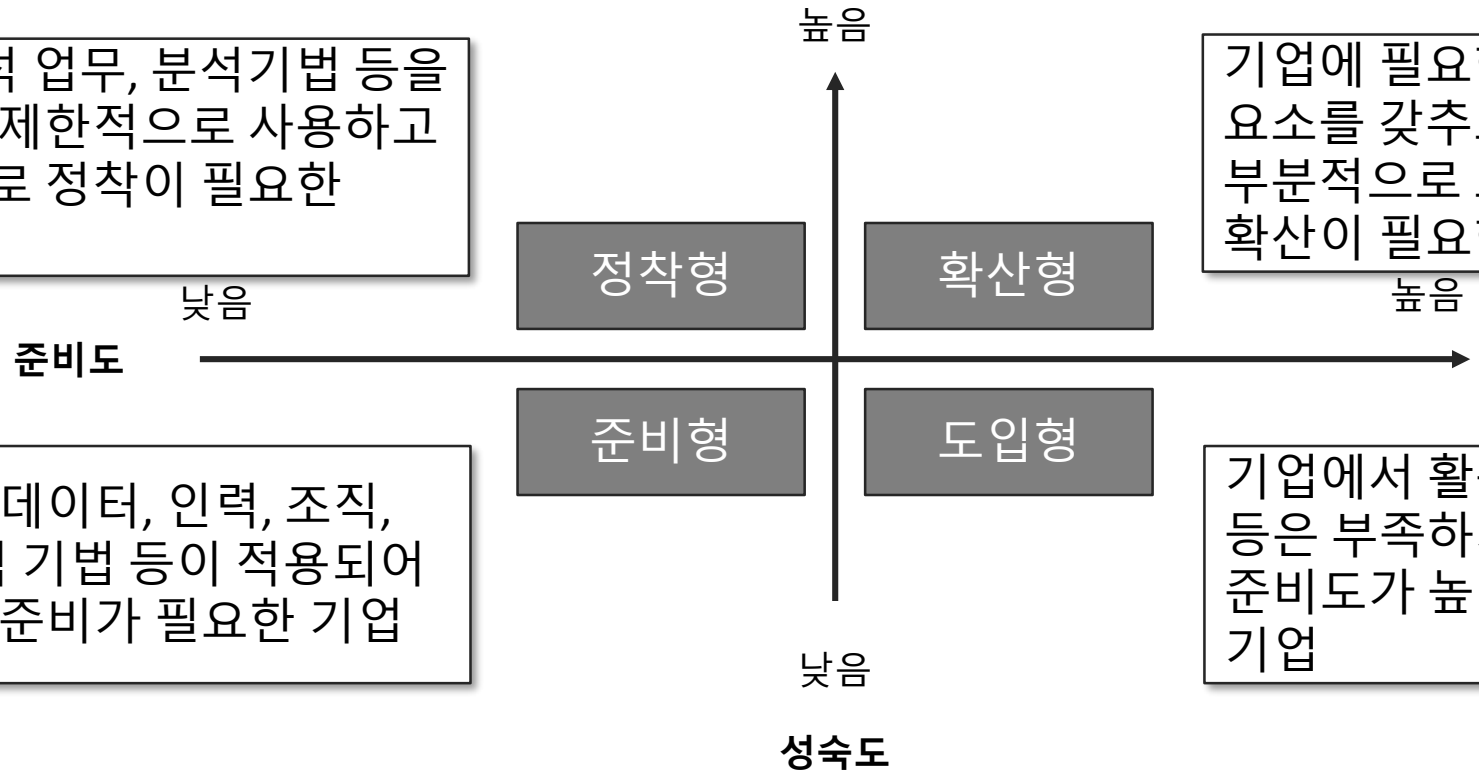
설명	데이터 분석을 진화시켜서 혁신 및 성과 향상에 기여
비즈니스 부문	• 외부 환경 분석 활용 • 최적화 업무 적용 • 실시간 분석 • 비즈니스 모델 진화
조직역량 부문	• 데이터 사이언스 그룹 • 경영진 분석 활용 • 전략 연계
IT 부문	• 분석 협업 환경 • 분석 sandbox(보호되고 안전한 상태에서 다양한 테스트가 가능한 환경) • 프로세스 내재화(internalization) • 빅데이터 분석

- 분석 수준 진단 결과
  - 기업의 현재 분석 수준을 객관적으로 파악



◦ 사분면 분석 결과

조직, 인력, 분석 업무, 분석기법 등을 기업 내부에서 제한적으로 사용하고 있어 일차적으로 정착이 필요한 기업

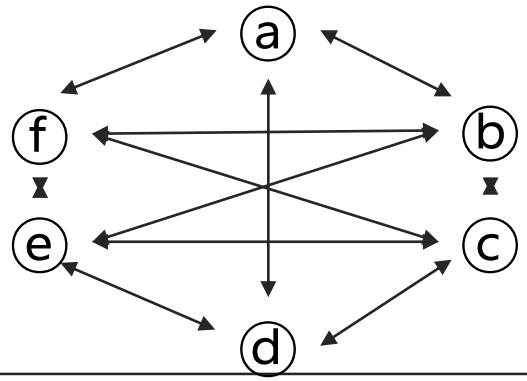
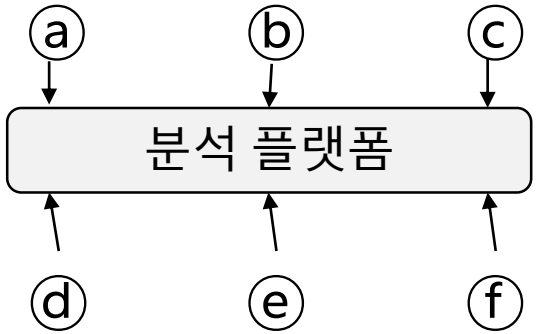


기업에 필요한 6가지 분석 구성 요소를 갖추고 있어, 현재 부분적으로 도입되어 지속적으로 확산이 필요한 기업

기업에 필요한 데이터, 인력, 조직, 분석 업무, 분석 기법 등이 적용되어 있지 않아 사전준비가 필요한 기업

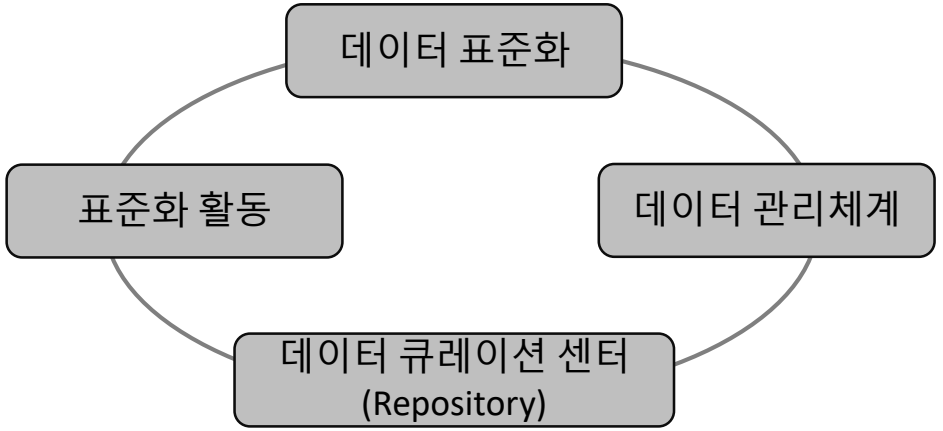
기업에서 활용하는 분석 업무, 기법 등은 부족하지만 적용 조직 등 준비도가 높아 바로 도입할 수 있는 기업

- 분석 지원 인프라 방안 수립
  - 분석 과제 단위별로 별도의 분석 시스템을 구축 대신
  - 장기적, 안정적으로 활용할 수 있는 확장성을 고려한 플랫폼 구조

개별 시스템	플랫폼 구조
	
시스템 간 자체적인 데이터 교환	**분석 플랫폼을 활용한 공동 기능 활용**
시스템 별 독립적인 데이터 관리	중앙집중적 데이터 관리
확장 시 시스템간 인터페이스 폭증	시스템 간 인터페이스 최소화

- 데이터 거버넌스 체계 수립
  - 정의
    - 기업 전체 차원에서 모든 데이터에 대한 정책 및 지침, 표준화, 운영 조직 및 책임 등의 관리체계 수립과 운영을 위한 framework 및 데이터 큐레이션 센터 (repository) 구축
    - 데이터 큐레이트가 데이터 거버넌스의 중요 사항임
  - 특징
    - 데이터의 가용성, 유용성, 통합성, 보안성, 안전성 확보
    - 기업에 따라서 기업 IT 거버넌스 혹은 EA(Enterprise Architecture)의 구성요소로 구축

- 데이터 거버넌스 구성요소
  - 원칙 Principle
    - 데이터 유지 관리를 위한 지침 및 가이드
    - 보안, 품질 기준, 변경 관리
  - 조직 Organization
    - 조직의 역할과 책임
    - 데이터 큐레이터(데이터 관리자, data architect), DB 관리자
  - 프로세스 Process
    - 데이터 관리를 위한 활동과 체계
    - 작업 절차, 모니터링 활동, 측정 활동



- 데이터 거버넌스 체계
  - 데이터 표준화
    - 데이터 표준 용어 설정, Name Rule 수립, 메타데이터 구축, 데이터 사전 구축
  - 데이터 관리 체계
    - 표준 데이터, 메타데이터, 데이터 사전 관리 원칙 수립
  - 데이터 큐레이션 센터 관리
    - 메타데이터 및 표준 데이터 관리를 위한 저장소
  - 표준화 활동
    - 정기적인 모니터링 실시

- 조직 및 인력 방안 수립
  - 데이터 분석 조직 구조 유형

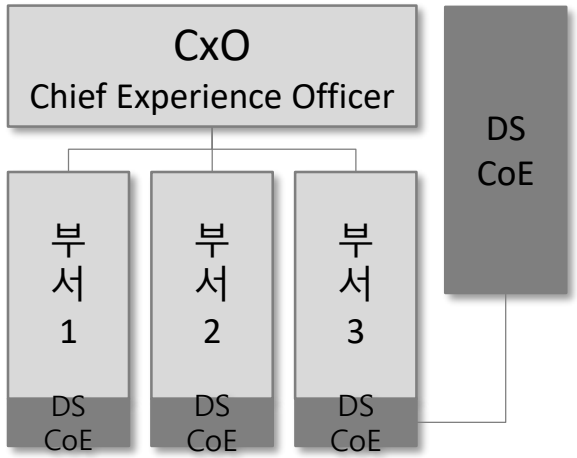
집중 구조

- 전담 분석 조직
- 우선 순위 운영
- 이원화 가능성

기능 구조

- 해당 부서에서 분석 업무 담당
- 기업 전체의 핵심 분석이 어려움

기능 구조



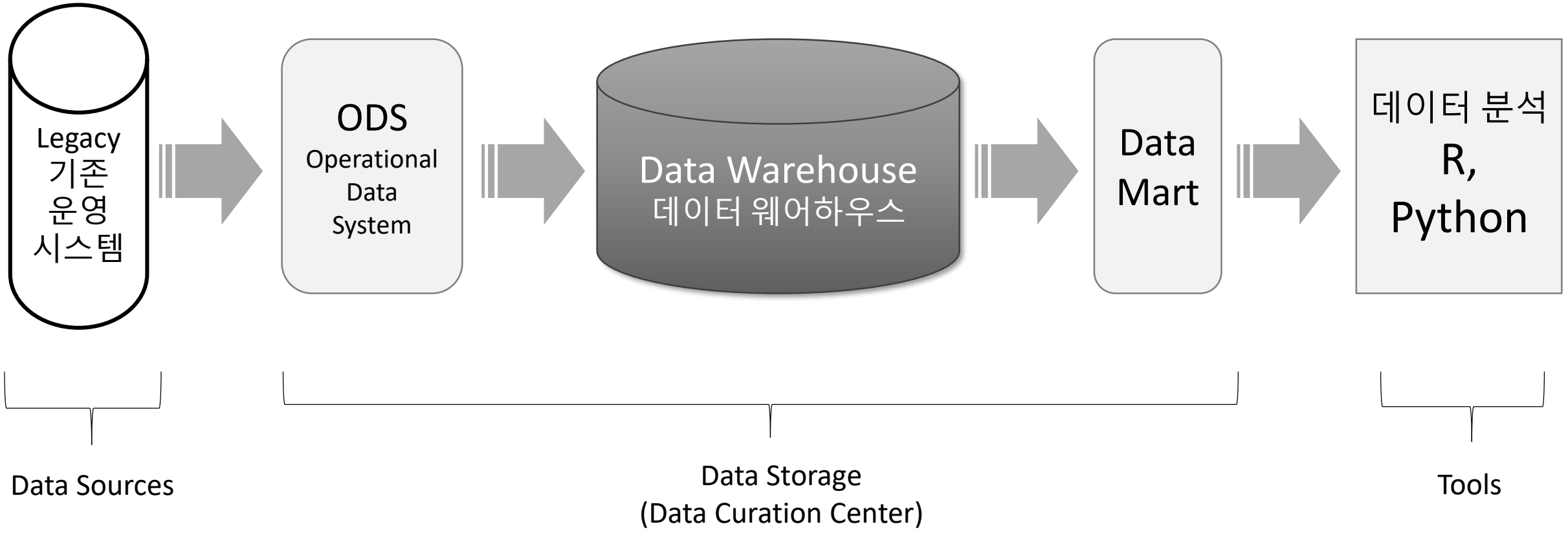
- 분석 조직 인력을 현업 부서에 직접 배치 운영
- 분석 결과를 신속히 처리 가능

※ DS CoE: Data Science Center of Excellence

◦ 데이터 분석 조직의 인력 구성(예시)

3. 데이터 분석 기법 소개

- 데이터 처리



- 단위별 활동
  - 데이터 웨어하우스, 데이터 마트: 분석에 필요한 데이터 제공
  - Legacy 혹은 Operational Data Store: DW에 포함되어 있지 않은 데이터 제공
    - 운영 시스템에서 데이터를 직접 가져와서 사용하면 위험하므로 ODS 이용
  - 데이터 클리닝 작업을 거친 후 분석

▪ 시각화

- 비록 낮은 수준의 분석이지만, 잘 활용하면 효율적
- 대용량의 빅데이터 분석에서 필수적
- 탐색적 분석(특이한 점이나 의미 있는 사실 도출)에서 필수적

다음 시간

- 지도학습 알고리즘