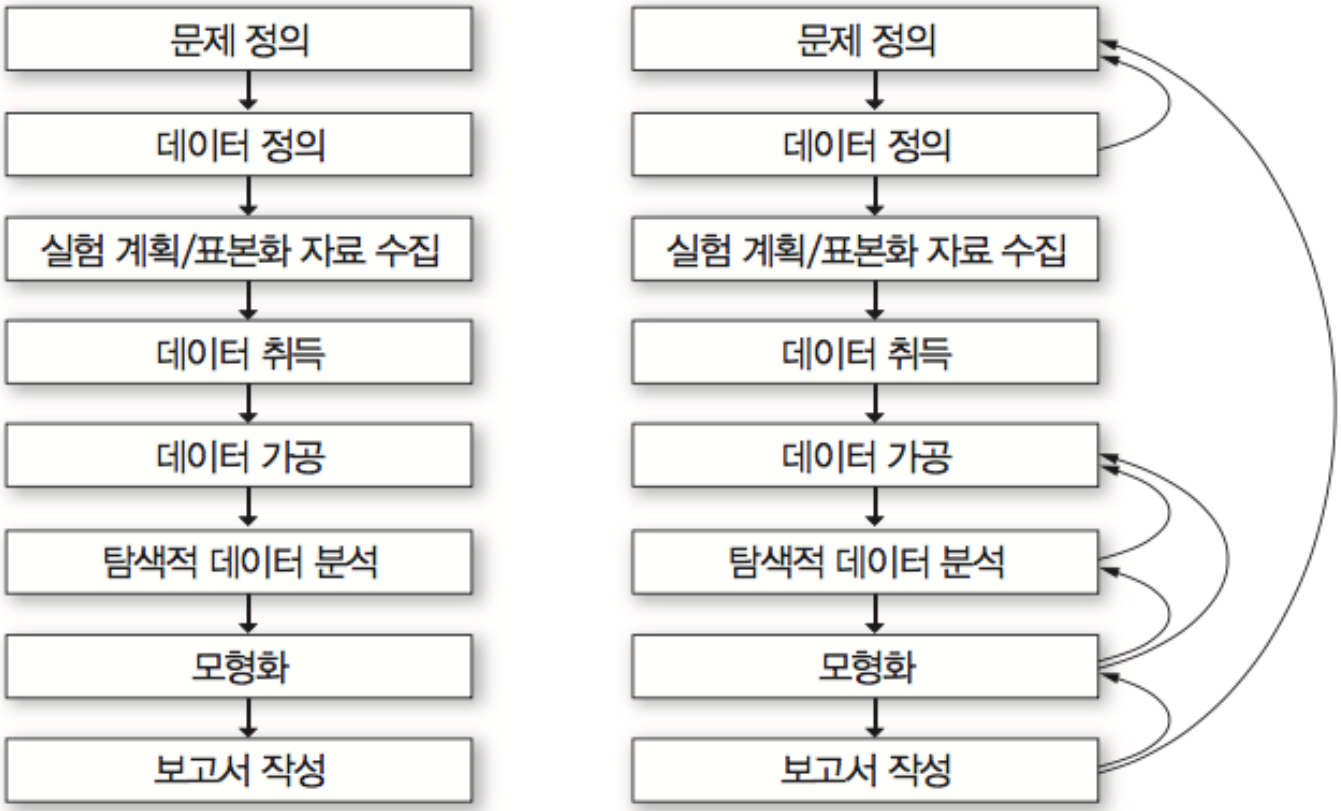
❖ 데이터 과학 프로세스

1. 문제 정의(problem definition) : 현실의 구체적인 문제를 명확하게 표현하고 통계적, 수리적 언어로 '번역'하는 작업
2. 데이터 정의(data definition): 변수(variable), 지표(metric) 등을 정의
3. 실험 계획(design of experiment)/표본화(sampling) : 데이터를 직접 수집해야 하는 경우는 보통 두 가지 목적 중 하나다. 첫째는 어떤 처리의 효과를 알아내기 위한 통제 실험, 둘째는 모집단을 대표하는 표본을 얻기 위한 표본화다. 소스 데이터(source data)가 이미 존재하는 경우에는 불필요
4. 데이터 취득(data acquisition) : 다양한 형태의, 다양한 시스템에 저장된 원데이터를 분석 시스템으로 가져오는 활동
5. 데이터 가공(data processing, data wrangling) : 데이트를 분석하기 적당한 표 형태로 가공하는 작업, 데이터 변환
6. 탐색적 분석과 데이터 시각화(exploratory data analysis, data visualization) : 시각화와 간단한 통계량을 통하여 데이터의 패턴을 발견하고 이상치를 점검하는 분석
7. 모형화(modeling) : 모수 추정, 가설검정등의 활동과 모형분석, 예측분석 등을 포괄
8. 분석 결과 정리(reporting) : 분석 결과를 현실적인 언어로 이해하기 쉽도록 번역해내는 작업



데이터 분석 과정에 대한 이상적 관점(왼쪽)과 현실적 관점(오른쪽)

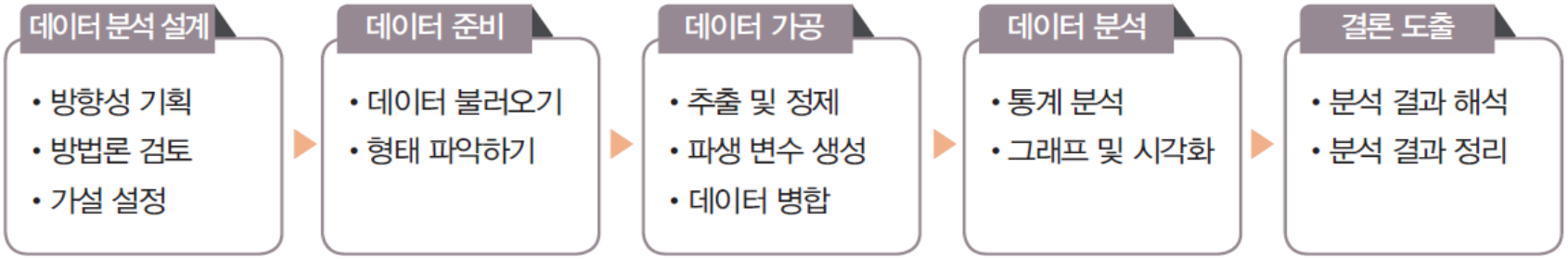
❖ 선형적이면서 직선적인 과정은 현실적이지 않다. 실제 데이터 분석에서는 다음과 같은 경우가 생기게 된다.

- 데이터 수집에서 문제가 생기면 필요한 데이터나 문제를 수정해야 할 수도 있다.
- 탐색적 데이터 분석에서 수집된 데이터의 문제가 발견되기도 한다. 새로 데이터를 수집하거나 문제 가설등을 바꿔야 할 수도 있다.
- 모형화 작업에서 의미 있는 결과를 도출하지 못할 수도 있다. 그러한 경우에도 그 결과 역시 정리하고 알려야 한다.
- 모형화의 결과가 유의미하지 않은 이유를 알아내기 위해 탐색적 데이터 분석을 다시 시행해야 할 수도 있다.
- 일반적으로 모형화 단계에서는 여러 다양한 모형을 시도하게 된다. 모형 결과 자체도 탐색적 데이터 분석을 해야 할 경우가 많다.
- 분석 결과 정리와 공유 후에 여러 피드백을 받게 된다. 이것은 새로운 문제 정의, 데이터 정의 등의 단계로 선순환적으로 이어지게 된다.

❖ 데이터 분석 프로세스를 도식적으로, 선형적으로 이해하는 것은 피해야 할 것이다. 대신에 데이터가 말해주는 내용을 좇아서 능동적으로 적응해나가는, 점진적이면서 순환적(iterative)과정으로 이해하는 것이 더 현실에 가깝다.

❖ 데이터 분석 과정

데이터 분석 설계 -> 데이터 준비 -> 데이터 가공 -> 데이터 분석 -> 결론 도출



1. 데이터 분석 설계

① 분석 주제 구체화, 용어 정리 및 주제 선정

② 브레인스토밍 등을 활용하여 다양한 관점의 가설 설정

③ 가설에 따른 분석 가능 변수 구성

④ 분석 항목 결정

2. 데이터 준비

- 필요한 데이터나 환경에 따라 데이터를 준비하는 방법
  - 필요한 데이터를 찾아 직접 입력하여 생성
  - 기존에 누군가 구성해 둔 데이터를 찾아 활용
- 데이터 형태 파악
  - 구조 및 구성 변수의 형태
  - 데이터 값의 특성 등을 중심으로 파악

3. 데이터 가공 및 통합
   - 원시 데이터를 원하는 형태로 처리
   - 불필요한 변수 제거하여 필요한 변수만 추출
   - 데이터 값에 따라 그룹화

4. 데이터 분석
   - 가공하여 준비한 데이터를 이용하여 다양한 분석
     - 기초 통계량으로 데이터를 파악
     - 다양한 그래프를 통해 분포의 시각화
     - 분석 방법론 적용

5. 결론
   - 다양한 통계량을 통해 가설을 검증
   - 결과를 정리하여 결과를 도출